JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

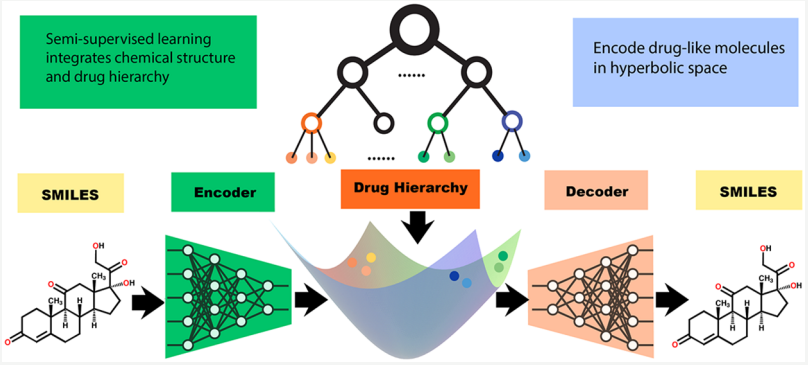
pubs.acs.org/jcim Article

Semi-supervised Hierarchical Drug Embedding in Hyperbolic Space

Ke Yu,* Shyam Visweswaran,* and Kayhan Batmanghelich*

Cite This: J. Chem. Inf. Model. 2020, 60, 5647−5657 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Learning accurate drug representations is essential for tasks such as computational drug repositioning and prediction of drug side effects. 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. Both quantitative and qualitative results support that the learned drug embedding can accurately reproduce the chemical structure and recapitulate the hierarchical relations among drugs. Furthermore, our approach can infer the pharmacological properties of novel molecules by retrieving similar drugs from the embedding space. We demonstrate that our drug embedding can predict new uses and discover new side effects of existing drugs. We show that it significantly outperforms comparison methods in both tasks.

INTRODUCTION

The study of drug representation provides the foundation for a variety of applications in computational pharmacology, such as computational drug repositioning and prediction of drug side effects. 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 for drug repositioning typically aim to identify mechanisms of action that are shared among drugs that imply that the drugs may also share therapeutic purposes.[2] However, such methods are limited when prior knowledge of drugs is 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. Side effects of drugs are undesirable effects that may cause harm to individuals and may even cause death. Computational methods for predicting drug side effects often integrate several drug features from heterogeneous data sources (e.g., chemical, biological, and therapeutic properties).[3] However, the utility of drug hierarchy in learning drug representation has not yet been explored. A drug hierarchy encodes a broad spectrum of known drug relations. For example, a widely used drug hierarchy, the Anatomical Therapeutic Chemical (ATC) classification system, groups drugs that have similar mechanisms of action and therapeutic, pharmacological, and chemical characteristics.

Representing the chemical structure of drug-like molecules has received substantial attention recently.[4] This approach focuses on learning representations that can be used to identify promising molecules that satisfy specified properties.[5−7]

Special Issue: Generative Models for Molecular Design

Received: June 15, 2020
Published: November 3, 2020

https://dx.doi.org/10.1021/acs.jcim.0c00681
J. Chem. Inf. Model. 2020, 60, 5647−5657

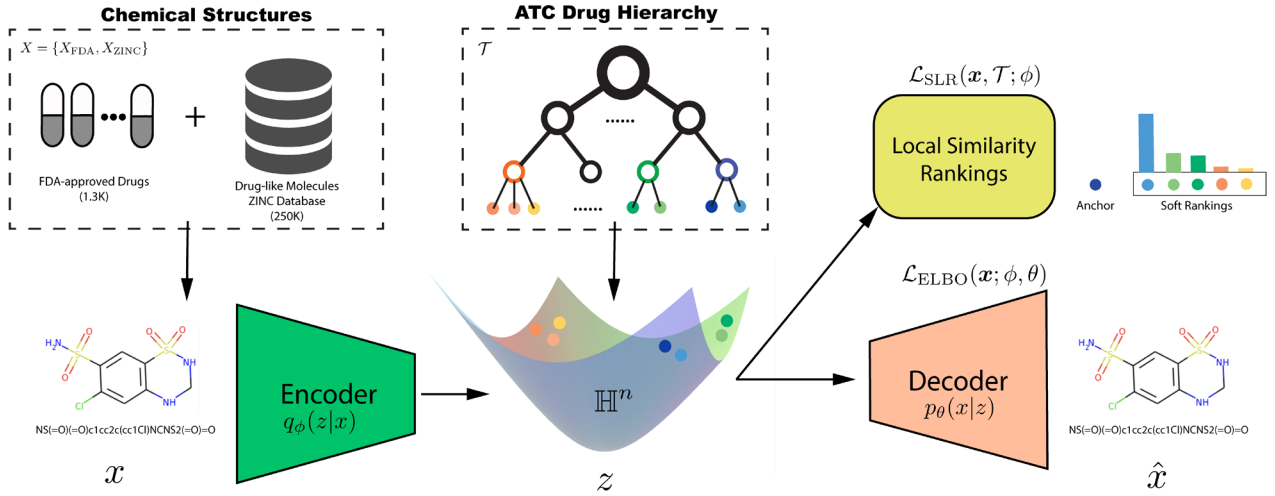


Figure 1. Schematic diagram of the proposed drug embedding method. Our semi-supervised learning approach integrates the chemical structures of a small number of FDA-approved drug molecules (X_{FDA}) and a larger number of drug-like molecules (X_{ZINC}) drawn from the ZINC database. We use VAE to encode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

Typically, a large set of drug-like molecules is encoded in a latent space, which is then coupled with a predictive model. However, this approach does not directly incorporate prior knowledge about existing drugs. In another approach, knowledge about existing drugs is leveraged to predict hitherto unknown properties of drugs. Such knowledge-based methods view every drug as a node in graph and predict linkages where the linkage may indicate a new use,[8] a side effect,[9] or an adverse drug−drug interaction.[10] However, such an approach is limited to the drugs available in the knowledge database and learns task-specific representations that may not transfer well to additional tasks. Our method merges the two approaches described above by combining chemical structure representation learning with known knowledge of drugs to learn useful and generalizable drug representation.

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 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)[11] structure representation. We obtain drug similarity relationships from the ATC drug hierarchy that hierarchically groups drugs by the system of action, therapeutic intent, and pharmacological and chemical characteristics. We use the hyperbolic space for the embedding since it is amenable for learning continuous concept hierarchies.[12−15]

We formulate the learning of the drug embedding as a Variational Auto-Encoder (VAE) where the codes ($\boldsymbol{z}$) reside in hyperbolic space. More specifically, we adopt a variant that replaces the prior normal distribution in the VAE with the so-called wrapped normal distribution in the Lorentz model of hyperbolic space. To integrate the hierarchical relationships from the drug hierarchy, we use a loss function that enforces the pairwise hyperbolic distance between drugs to be consistent with pairwise shortest path lengths in the ATC hierarchy.

We evaluate the effects of ATC knowledge and the hyperbolic space independently on the quality of drug embeddings in their ability to accurately capture the chemical structure and preserve the ATC hierarchy in the latent space. Our experiments show that the relationships entailed by the ATC hierarchy are an effective inductive bias for learning the chemical features, and the hyperbolic space is superior to the Euclidean space in representing the top levels of the ATC hierarchy, i.e., the anatomical groups and therapeutic groups. We also evaluate the efficacy of our embedding for drug repositioning and for predicting side effects on two publicly available data sets. The results show that our embedding performs better than comparison methods for these two computational pharmacology tasks.

BACKGROUND

Substantial research has been done in the past few years in applying machine learning to drug discovery and related tasks.[16−18] Successful applications in drug discovery include target identification and validation,[19−22] compound design with desirable properties,[7,23] prediction of drug toxicity,[24] and prediction of biomarkers of clinical end points.[25−27] Machine learning has also been applied to computational pharmacology tasks, such as drug repositioning,[8] prediction of side effects,[28] and prediction of adverse drug−drug interactions.[10] However, many of these methods developed for computational pharmacology tasks represent a drug as a node in a graph and ignore the rich information in the chemical structure of the drug. Moreover, such approaches cannot be readily applied to a new drug that is not in the data set. Our method can be viewed as knowledge representation learning that integrates information from (1) a large corpus of drug-like molecules that is used in drug discovery and (2) an expert-curated drug hierarchy that contains rich information about known drugs. To the best of our knowledge, our method is unique in that it enables localizing novel molecules in the context of the clinically approved drugs.

Molecular featurization methods can be divided into the following groups: (1) methods that extract expert-crafted descriptors from molecular structures, (2) methods that map molecular structures to bit strings, such as extended-connectivity

https://dx.doi.org/10.1021/acs.jcim.0c00681
J. Chem. Inf. Model. 2020, 60, 5647−5657

fingerprints (ECFP),[29] and (3) recent deep learning based methods. The deep learning-based methods can be further categorized into two groups. The first group consists of methods that encode the molecular formula as a string of characters and use a variant of recurrent neural network (RNN)[7,30,31] to extract features, and the second group contains methods that represent as undirected graphs where nodes denote atoms and edges denote bonds.[32,33] Each group has advantages and disadvantages.[34] Our method belongs to the first group of deep learning-based methods. However, our framework is quite general in that the encoder−decoder can be replaced with a graph-neural encoder−decoder if needed.

There are several methods to quantify the similarity (or distance) of molecular representations. For fingerprint-based similarity calculations, Tanimoto index, Dice index, Cosine coefficient, and Soergel distance have been identified as excellent metrics.[35,36] For deep generative methods[7,37] involving encoding the molecules as continuous vectors, Euclidean distance is the most popular metric to assess molecular similarity in the latent space. Our method uses the hyperbolic distance in the latent space as the similarity metric.

Embedding hierarchical relations in a latent space has been an active area of research.[15,38] Hyperbolic space is an appealing choice for embedding a hierarchy because it can represent tree-like structures with arbitrarily low distortion.[14] There are several equivalent geometric models[39] of hyperbolic space. Many applications of hyperbolic space to machine learning[12,13,15] have adopted the Poincaré ball model. However, as proposed in ref 40, the Lorentz model allows for a more efficient closed-form computation of geodesics and avoids numerical instabilities that arise from the Poincaré distance. A more recent study[41] introduced the *wrapped normal distribution* in the Lorentz model. To the best of our knowledge, our method is the first hyperbolic VAE that can induce hierarchical structure from pairwise similarity measurements in a latent space.

METHODS

Learning Chemical Grammar Using VAE. We use a VAE to encode the chemical structure of drug-like molecules. More specifically, we model a molecule as a random variable generated by encoding a SMILES string into a code ($\boldsymbol{z}$), which is then decoded back to a reconstruction of the input by passing through a decoder $p_\theta(\boldsymbol{x}|\boldsymbol{z})$. Finding the optimal θ using maximum likelihood requires computing the so-called evidence function, $\log p_\theta(\boldsymbol{x})$, which is difficult to compute since it entails integrating over $\boldsymbol{z}$.

Instead of directly maximizing likelihood, variational Bayes maximizes the variational evidence lower bound (ELBO).[42] The ELBO is given by

$$\log p_\theta(\boldsymbol{x}) \geq \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z})) \tag{1}$$

where the first term after the inequality is the reconstruction term, the second term is the regularization term, and $\mathbb{E}$ and D_{KL} denote the expectation and Kullback−Leibler (KL) divergence, respectively. The global optimal $q(\boldsymbol{z}|\boldsymbol{x})$ is achieved when $q(\boldsymbol{z}|\boldsymbol{x}) = p(\boldsymbol{z}|\boldsymbol{x})$; the variational distribution approximates the posterior distribution. In order to control the relative effect of KL divergence[43] we adopt β-VAE,[44] a more general form of VAE that applies a scaling hyperparameter β to the D_{KL} term in the ELBO. We employ the RNN[45] architecture for both the encoder and the decoder networks, in order to perform sequence-to-sequence learning on SMILES strings.

In the classic VAE,[46] the prior $p(\boldsymbol{z})$ is modeled with the standard normal distribution, the encoder $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is modeled by a Gaussian distribution $\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_\phi, \Sigma_\phi)$, and the first term in the ELBO is estimated using a Monte Carlo estimator

$$\mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x}_i)}[\log p_\theta(\boldsymbol{x}_i|\boldsymbol{z})] \approx \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(\boldsymbol{x}_i|g_\phi(\boldsymbol{\epsilon}^{(l)}, \boldsymbol{x}_i)) \tag{2}$$

where $g_\phi(\boldsymbol{\epsilon}^{(l)}, \boldsymbol{x}_i) = \boldsymbol{\mu}_\phi^{(i)} + \boldsymbol{\sigma}_\phi^{(i)} \odot \boldsymbol{\epsilon}^{(l)}$, $\boldsymbol{\epsilon}^{(l)} \sim \mathcal{N}(0, \mathbf{I})$ is the reparameterization trick, and L is the number of samples per data point. To extend VAE from a flat Euclidean space to a curved manifold, the Gaussian distribution needs to be extended to the hyperbolic space.

Wrapped Normal. Intuitively, hyperbolic space can be viewed as a continuous version of tree because its volume and surface area grow exponentially with the radius. Compared to Euclidean space, the hyperbolic space better captures the hierarchical characteristic of trees. In this paper, we employ a specific model of the hyperbolic space, namely, the Lorentz (Minkowski/Hyperboloid) model. The Lorentz model $\mathbb{H}^n$ of n-dimensional hyperbolic space is defined as

$$\mathbb{H}^n = \{\boldsymbol{z} \in \mathbb{R}^{n+1} : \langle \boldsymbol{z}, \boldsymbol{z}\rangle_{\mathcal{L}} = -1, z_0 > 0\}, \tag{3}$$

and $\langle \boldsymbol{z}, \boldsymbol{z}'\rangle_{\mathcal{L}} = -z_0 z_0' + \sum_{i=1}^{n} z_i z_i'$ is the so-called *Lorentzian inner product*, which is also the metric tensor of the hyperbolic space. We adopt the so-called wrapped normal distribution proposed by Nagano et al., 2019,[41] which we denote by $\mathcal{N}_{\mathbb{H}}^{\mathrm{W}}(\boldsymbol{z}|\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{z} \in \mathbb{H}^n$, and $\boldsymbol{\mu}$ is the hyperbolic mean. The sampling strategy can be summarized in three steps as illustrated in Figure 2. First, we define a Gaussian random variable,

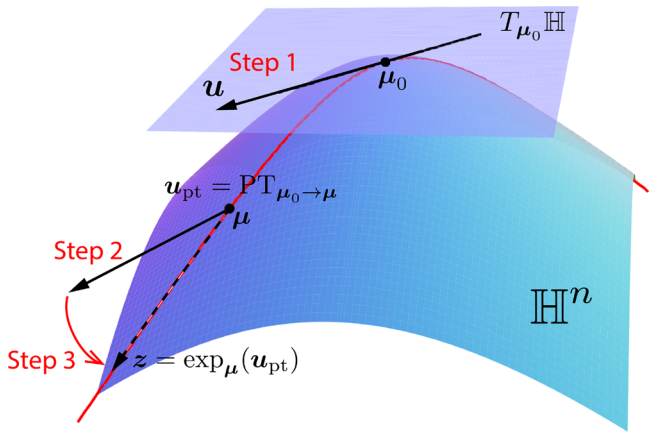


Figure 2. Steps of sampling from the wrapped normal distribution.

$\boldsymbol{u} \sim \mathcal{N}(\mathbf{0}, \Sigma)$, on the tangent space (see Supporting Information eq S3) at the origin of the hyperbolic space, $\boldsymbol{u} \in T_{\boldsymbol{\mu}_0}\mathbb{H}$, and then, we *parallel transport* (see Supporting Information eq S8), $\boldsymbol{u}_{\mathrm{pt}} = \mathrm{PT}_{\boldsymbol{\mu}_0\to\boldsymbol{\mu}}(\boldsymbol{u})$, the random vector to another tangent space at a desired location $\boldsymbol{\mu}$, $\boldsymbol{u}_{\mathrm{pt}} \in T_{\boldsymbol{\mu}}\mathbb{H}$. The *parallel transport* translates a vector from $T_{\boldsymbol{\mu}_0}\mathbb{H}$ to $T_{\boldsymbol{\mu}}\mathbb{H}$ along the geodesic (see Supporting Information eq S2) between $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}$ without changing its metric tensor. Finally, we map the transported vector into hyperbolic space via the *exponential map* (see Supporting Information eq S4), $\boldsymbol{z} = \exp_{\boldsymbol{\mu}}(\boldsymbol{u}_{\mathrm{pt}})$. Importantly, this sampling scheme is sequentially norm-preserving, i.e., $\|\boldsymbol{u}\|_2 =$

https://dx.doi.org/10.1021/acs.jcim.0c00681
J. Chem. Inf. Model. 2020, 60, 5647−5657

$\|\boldsymbol{u}\|_{\mathcal{L}} = \|\boldsymbol{u}_{\text{pt}}\|_{\mathcal{L}} = d_l(\boldsymbol{z}, \boldsymbol{\mu})$, where $d_l(\boldsymbol{z}, \boldsymbol{\mu})$ denotes the hyperbolic distance between $\boldsymbol{z}$ and $\boldsymbol{\mu}$ on the Lorentz manifold.

The reparameterization trick used in the VAE needs to be modified since the algebraic addition of coordinates of two points on a manifold does not necessarily reside on the manifold. The composition of these two operations, $\exp_{\boldsymbol{\mu}}(\text{PT}_{\boldsymbol{\mu}_0 \to \boldsymbol{\mu}}(\boldsymbol{u}))$, can be viewed as the reparametrization trick in the hyperbolic VAE. The inside operation shifts the tangent space from $\boldsymbol{\mu}_0$ to $\boldsymbol{\mu}$ analogous to the addition operation of the classic reparameterization trick. The $\exp_{\boldsymbol{\mu}}$ projects the shifted vector to the manifold. Therefore, we sample $\boldsymbol{z}_i^{(l)} \sim q_\phi(\boldsymbol{z}|\boldsymbol{x}_i)$ using

$$\boldsymbol{z}_i^{(l)} = g_\phi(\boldsymbol{u}^{(l)}, \boldsymbol{\mu}_i) = \exp_{\boldsymbol{\mu}_i}(\text{PT}_{\boldsymbol{\mu}_0 \to \boldsymbol{\mu}_i}(\boldsymbol{u}^{(l)})) \tag{4}$$

where $\boldsymbol{u}^{(l)} \sim \mathcal{N}(\boldsymbol{0}, \Sigma)$, and l denotes the index of sample. 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 $\mathcal{N}_{\mathbb{H}}^{\text{W}}(\boldsymbol{z}|\boldsymbol{\mu}, \Sigma)$.

KL Divergence. To compute the KL divergence, we need to evaluate the probability density of the wrapped normal. The wrapped normal distribution can be viewed as change of variable from a normal distribution via the eq 4. Applying the change of variable, we obtain

$$\log q_\phi(\boldsymbol{z}_i^{(l)}|\boldsymbol{x}_i) = \log \mathcal{N}(g_\phi^{-1}(\boldsymbol{z}_i^{(l)}, \boldsymbol{\mu}_i); \boldsymbol{0}, \Sigma) - \log \det\left(\frac{\partial g_\phi(\boldsymbol{u}^{(l)}, \boldsymbol{\mu}_i)}{\partial \boldsymbol{u}^{(l)}}\right) \tag{5}$$

The inverse operation $g_\phi^{-1}(\boldsymbol{z}_i^{(l)}, \boldsymbol{\mu}_i)$ simply maps $\boldsymbol{z}_i^{(l)}$ back to $\boldsymbol{u}^{(l)}$ by applying the *logarithmic map* (see Supporting Information eq S5) and the *inverse parallel transport* (see Supporting Information eq S9). We compute the second log-determinant term following the derivation in Nagano et al., 2019.[41]

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 the drug hierarchy which defines similarity or dissimilarity between drugs at various levels. In this section, we incorporate the ATC hierarchy into our model. Note that the terminal nodes of the ATC hierarchy are drugs that have SMILES string representations, while the internal nodes of the ATC hierarchy are drug classes, e.g., beta blocking agents. Inspired by concept embedding in hyperbolic space,[40] we incorporate the ATC hierarchy 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 the ATC hierarchy $\mathcal{T}$, 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

$$p(\boldsymbol{x}_i, \boldsymbol{x}_j; \phi) = \frac{\exp(-d_l(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j))}{\sum_{k \in \mathcal{D}(i,j)} \exp(-d_l(\boldsymbol{\mu}_i, \boldsymbol{\mu}_k))} \tag{6}$$

where $\boldsymbol{\mu}_i$ is the hyperbolic mean of $q_\phi(\boldsymbol{z}|\boldsymbol{x}_i) = \mathcal{N}_{\mathbb{H}}^{\text{W}}(\boldsymbol{z}|\boldsymbol{\mu}_i, \Sigma_i)$, and $d_l(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)$ is the hyperbolic distance between $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$. The likelihood function of the soft local rankings per $\boldsymbol{x}_i \in X_{\text{FDA}}$ is given by

$$\mathcal{L}_{\text{SLR}}(\boldsymbol{x}_i, \mathcal{T}; \phi) = \sum_j \log p(\boldsymbol{x}_i, \boldsymbol{x}_j; \phi) \tag{7}$$

where $\boldsymbol{x}_j \in \{X_{\text{FDA}} - \boldsymbol{x}_i\}$.

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 results of our experiments (see Supporting Information Figure S2 for more information) 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 an 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. *Formulation.* We employ a semi-supervised learning approach that combines a small number of drugs X_{FDA} with a larger number of drug-like molecules X_{ZINC}. The supervised learning task is to maximize the likelihood of the soft local rankings with respect to the ATC hierarchy $\mathcal{T}$. The unsupervised learning task is to maximize the ELBO of the marginal likelihood of the chemical structures of drugs and drug-like molecules $X = \{X_{\text{ZINC}}, X_{\text{FDA}}\}$. We then formulate the drug embedding problem as

$$\underset{\phi,\theta}{\text{argmax}}(\mathcal{L}_{\beta\text{-ELBO}}(\boldsymbol{x}; \phi, \theta) + c \cdot \mathcal{L}_{\text{SLR}}(\boldsymbol{x}, \mathcal{T}; \phi)) \tag{8}$$

where $c = 1$ when $\boldsymbol{x} \in X_{\text{FDA}}$, $c = 0$ when $\boldsymbol{x} \in X_{\text{ZINC}}$, and $|X_{\text{ZINC}}| \gg |X_{\text{FDA}}|$. 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.

Training. In practice, the learning procedure for the parameters ϕ, θ is summarized as

$$\underset{\phi,\theta}{\text{argmax}} \frac{1}{|X|} \sum_{\boldsymbol{x}_i \in X} (\log p_\theta(\boldsymbol{x}_i|\boldsymbol{z}_i) - \beta \cdot \tilde{D}_{\text{KL}}(q_\phi(\boldsymbol{z}_i|\boldsymbol{x}_i) \| p(\boldsymbol{z}_i))) + \gamma \cdot \frac{1}{|X_{\text{FDA}}|} \sum_{\boldsymbol{x}_i \in X_{\text{FDA}}} \tilde{\mathcal{L}}_{\text{SLR}}(\boldsymbol{x}_i, \mathcal{T}; \phi) \tag{9}$$

where $\boldsymbol{z}_i$ is a single sample in hyperbolic space, $p(\boldsymbol{z})$ is a wrapped normal distribution, and β and γ are scaling hyperparameters governing the relative weights of KL divergence and soft local ranking loss during training. Parameters are estimated using mini-batch gradient descent, and gradients are straightforward to compute using the hyperbolic reparametrization trick eq 4. See Algorithm 1 in the Supporting Information for the algorithmic description of the method. For details of model architectures, training settings, and implementation details, please refer to the Supporting Information. All code for the hyperbolic drug embedding is available in our GitHub repository: https://github.com/batmanlab/drugEmbedding.

RESULTS

In this section, we first describe the data sets used in our experiments, and then, we perform two sets of experiments to evaluate different components of our model: effect of the ATC information in preserving hierarchical relations among drugs and importance of hyperbolic space as the coding space. Finally, we study the efficacy of hyperbolic embeddings for drug repositioning and discovering side effects of drugs.

https://dx.doi.org/10.1021/acs.jcim.0c00681
J. Chem. Inf. Model. 2020, 60, 5647−5657

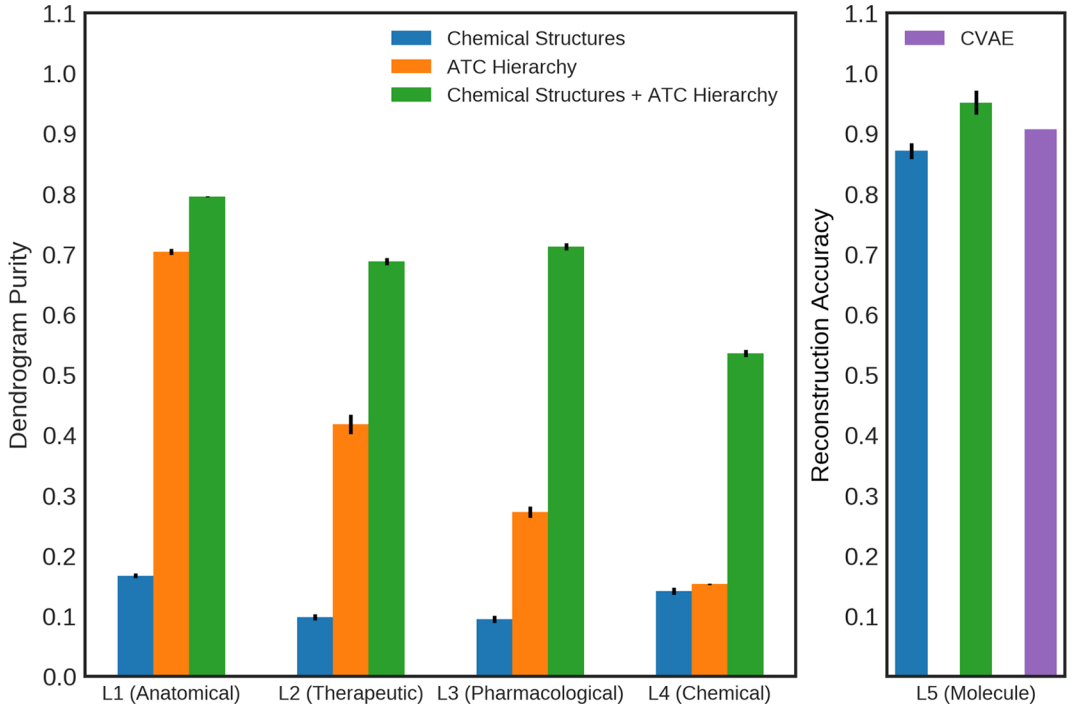


Figure 3. Effect of knowledge sources on the accuracy of recapitulating the ATC hierarchy and the reconstruction of the chemical structure. Results obtained using the embedding from chemical structures alone are shown in blue, results obtained using the embedding from the ATC hierarchy alone are shown in orange, and results obtained using the embedding from both sources of knowledge are shown in green. The baseline result taken from the CVAE[7] is shown in purple. The left panel shows the dendrogram purity (DP) at ATC levels 1, 2, 3, and 4. The right panel shows the reconstruction accuracy of the chemical structures. CVAE uses the Euclidean latent space, and all the other results are from the Lorentz model with dimension size of 64. The results were obtained by averaging three independent runs, and the error bars denote standard deviations.

Data Sets. *Chemical Structures.* We obtained SMILES strings of 1,365 FDA-approved drugs that were curated by ref 7. We obtained SMILES strings of 250,000 drug-like molecules that were extracted at random by ref 7 from the ZINC[47] database that contains a curated collection of >200 M commercially available chemicals. 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.

ATC. The ATC classification system was created by the World Health Organization (WHO)[48] that leverages the location of action, therapeutic, pharmacological, and chemical properties of drugs to group them hierarchically. Traversing from the top to the bottom of the hierarchy, the ATC system groups drugs according to the anatomical organ on which they act (level 1), therapeutic intent (level 2), pharmacological properties (level 3), and chemical characteristics (level 4). A drug that has several uses appears in several places in the ATC hierarchy. 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 represent active chemical substance (level 5). Of the 1,365 drugs, 1,055 were mapped to 1,355 terminal nodes at level 5 in the ATC tree.

SIDER. The Side Effect Resource (SIDER) database[49] contains 5,868 distinct side effects and 1,427 drugs for which one or more side effects have been documented. We obtained the SIDER data set in DeepChem,[50] which has grouped side effects into 27 classes based on the anatomical organ that is affected by the side effect.

RepoDB. RepoDB[51] is a benchmark data set that contains information on drug repositioning. It contains a curated set of drug repositioning successes and failures where each success or failure is a drug-indication pair where indication refers to a specific condition that the drug is used to treat. After mapping to the FDA-approved drugs, we obtained 4,738 successful and 2,576 failed drug-indication pairs.

Evaluation of Drug Embeddings. We assess the quality of hyperbolic embeddings in their ability to capture the chemical structure accurately as well as preserve relationships faithfully as entailed by the ATC hierarchy. To learn embeddings, we randomly split the chemical structures data set into training, validation, and test sets in the proportions 90%:5%:5%. The validation set is used to determine the best-fit model.

Metrics. We evaluate the embeddings in their ability to recapitulate the ATC hierarchy by applying agglomerative hierarchical clustering to the embeddings. We compare the embedding-induced hierarchy to the ATC hierarchy using dendrogram purity.[52] The dendrogram purity (DP) of a hierarchy $\tilde{\mathcal{T}}$ that is obtained from a set of drug embeddings $\{\boldsymbol{\mu}_i\}$ is computed as

$$\mathrm{DP}(\tilde{\mathcal{T}}) = \frac{1}{|\mathcal{W}^{\star}|} \sum_{\boldsymbol{\mu}_i,\boldsymbol{\mu}_j \in \mathcal{W}^{\star}} \mathrm{pur}(\mathrm{lvs}(\mathrm{LCA}(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)), C^{\star}(\boldsymbol{\mu}_i)) \tag{10}$$

where $C^{\star}(\boldsymbol{\mu}_i)$ is the (ground-truth) cluster that the drug $\boldsymbol{x}_i$ belongs to in the ATC $\mathcal{T}$, $\mathcal{W}^{\star}$ is the set of unordered pairs of drugs that belong to the same cluster, $\mathrm{LCA}(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)$ is a function that gives the lowest common ancestor of $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$ in $\tilde{\mathcal{T}}$, $\mathrm{lvs}(n)$ is the set of descendant leaves for any internal node n in $\tilde{\mathcal{T}}$, and $\mathrm{pur}(S_1, S_2) = |S_1 \cap S_2|/|S_1|$. Intuitively, DP measures the average purity of the lowest common ancestors of pairs of drugs that belong to the same ATC cluster. Note that $\mathrm{DP}(\tilde{\mathcal{T}})$ is a holistic

https://dx.doi.org/10.1021/acs.jcim.0c00681
J. Chem. Inf. Model. 2020, 60, 5647−5657

Table 1. Effect of Hyperbolic Space on the Accuracy of Recapitulating the ATC Hierarchy and the Reconstruction of the Chemical Structure[a]

			latent space dimension				
metric	ATC level	geometry	2	4	8	32	64
DP	L1 (anatomical)	Euclidean	$0.690_{\pm 0.013}$	$0.721_{\pm 0.030}$	$0.748_{\pm 0.029}$	$0.774_{\pm 0.005}$	$0.775_{\pm 0.008}$
		Lorentz	$\mathbf{0.757_{\pm .006}}$	$\mathbf{0.761_{\pm .014}}$	$\mathbf{0.771_{\pm .006}}$	$\mathbf{0.790_{\pm .003}}$	$\mathbf{0.795_{\pm .001}}$
DP	L2 (therapeutic)	Euclidean	$0.488_{\pm 0.023}$	$0.626_{\pm 0.008}$	$0.655_{\pm 0.017}$	$0.681_{\pm 0.003}$	$0.688_{\pm 0.003}$
		Lorentz	$\mathbf{0.617_{\pm .007}}$	$\mathbf{0.643_{\pm .015}}$	$\mathbf{0.666_{\pm .020}}$	$\mathbf{0.684_{\pm .007}}$	$\mathbf{0.690_{\pm .006}}$
DP	L3 (pharmacological)	Euclidean	$0.384_{\pm 0.027}$	$0.601_{\pm 0.018}$	$0.668_{\pm 0.026}$	$\mathbf{0.715_{\pm .006}}$	$\mathbf{0.725_{\pm .006}}$
		Lorentz	$\mathbf{0.577_{\pm .023}}$	$\mathbf{0.641_{\pm .018}}$	$0.668_{\pm 0.018}$	$0.696_{\pm 0.009}$	$0.714_{\pm 0.001}$
DP	L4 (chemical)	Euclidean	$0.238_{\pm 0.022}$	$0.402_{\pm 0.017}$	$0.454_{\pm 0.017}$	$\mathbf{0.597_{\pm .008}}$	$\mathbf{0.625_{\pm .006}}$
		Lorentz	$\mathbf{0.334_{\pm .007}}$	$\mathbf{0.441_{\pm .010}}$	$\mathbf{0.457_{\pm .013}}$	$0.517_{\pm 0.006}$	$0.528_{\pm 0.006}$
RA	L5 (molecule)	Euclidean	$0.004_{\pm 0.004}$	$\mathbf{0.020_{\pm .001}}$	$\mathbf{0.353_{\pm .031}}$	$0.904_{\pm 0.036}$	$0.951_{\pm 0.016}$
		Lorentz	$\mathbf{0.005_{\pm .005}}$	$0.012_{\pm 0.008}$	$0.309_{\pm 0.042}$	$\mathbf{0.906_{\pm .036}}$	$0.951_{\pm 0.020}$

[a]Dendrogram purity (DP) values of drug hierarchies and reconstruction accuracy (RA) values of SMILES obtained using embeddings in hyperbolic space (Lorentz model) and in Euclidean space. Values are shown at different ATC levels for both manifold geometries. The values are the average of three independent runs with standard deviations. A boldface value indicates that the corresponding manifold geometry has a higher value.

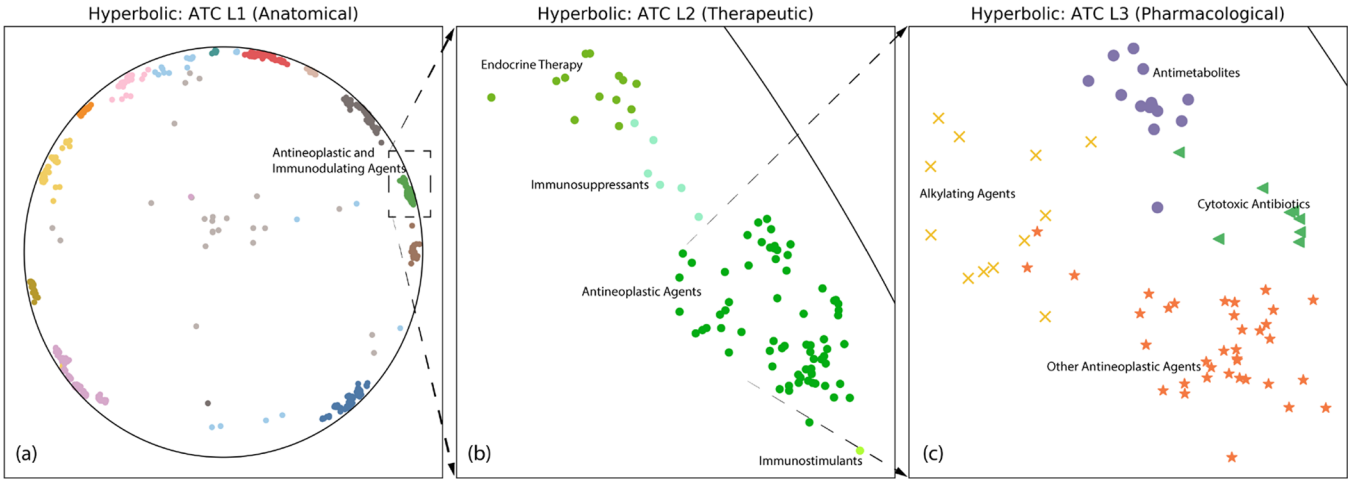


Figure 4. Visualization of hyperbolic drug embedding in a two-dimensional Poincaré disk that shows drugs with colored symbols. In panel (a) drugs that belong to the same group at ATC level 1 are denoted by circles of the same color. Panel (b) shows drugs of one group from ATC level 1 namely, "Antineoplastic 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.

measure of the complete ATC hierarchy that includes drugs in the training, validation, and test sets.

We also evaluate how well the embeddings are decoded to the original SMILES strings. Following ref 7, we evaluate the reconstruction accuracy as the proportion of successful decoding of latent representation after 200 attempts for 1,000 molecules randomly chosen from the test set.

Effect of Knowledge Source. We evaluate DP and reconstruction accuracy of embeddings obtained from a single source of knowledge that includes (1) chemical structures only by maximizing $\mathcal{L}_{\beta\text{-ELBO}}(\boldsymbol{x}; \phi, \theta)$ using the entire X and (2) ATC hierarchy only by maximizing $\mathcal{L}_{\text{SLR}}(\boldsymbol{x}, \mathcal{T}; \phi)$ using X_{FDA}. We compare them to the embedding that is obtained from both chemical structures and ATC hierarchy.

The left panel in Figure 3 shows the DP at different ATC levels. The embedding obtained from both sources of knowledge has substantially better DP than embeddings derived from only one source of knowledge. The improvement in DP is large at ATC levels 3 and 4 that cluster drugs by chemical structure. This result provides support that information learned from the task of SMILES reconstruction can help inform the task of drug clustering.

The right panel in Figure 3 shows the reconstruction accuracy of embeddings. The embedding obtained from both sources of knowledge has better performance on molecule reconstruction than the embedding using only chemical structures. This result suggests that the ATC hierarchy is an appropriate inductive bias for the task of decoding SMILES. Since drugs are grouped by their chemical characteristics at the lower ATC levels (3 and 4), minimizing the local ranking loss helps cluster drugs with similar chemical structures in the latent space and may create a smoother latent space that is suitable for decoding. Compared with the baseline model CVAE,[7] which uses the Euclidean space, the Lorentz embedding obtained from both sources of knowledge has superior reconstruction accuracy.

Effect of Hyperbolic Space. We compare embeddings from the Lorentz model with embeddings from the Euclidean model. The results in Table 1 show that overall the Lorentz embeddings have higher DP values and outperform the Euclidean embeddings. In low dimensional spaces (dimension size of two to four), the Lorentz model produces higher-quality

https://dx.doi.org/10.1021/acs.jcim.0c00681
J. Chem. Inf. Model. 2020, 60, 5647−5657

embeddings across all ATC levels, suggesting that hyperbolic space has superior capacity at the same dimension. In addition, the Lorentz model shows consistently higher DP values at ATC level 1, suggesting that it is superior to its Euclidean counterpart in recapitulating the global aspects of the hierarchy. For local aspects of the ATC hierarchy (levels 3 and 4), the improvement from hyperbolic representation decreases as the latent dimensionality increases. Euclidean latent space with dimension sizes of 32 and 64 performed better at ATC levels 3 and 4, suggesting that the Lorentz model with high dimensions might be overfitted at the lower levels of the ATC hierarchy. Besides, the DP results may be less reliable at ATC levels 3 and 4 due to the smaller sample sizes of the clusters. For the reconstruction of SMILES strings, the Lorentz and the Euclidean models show comparable accuracy. We chose representations in the Lorentz space with the latent dimension of 64 in the following experiments because it provides the highest reconstruction accuracy and highest DP values at ATC levels 1 and 2.

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 ref 40. In Figure 4(a), we observe that most of the drugs are placed near the boundary of the Poincaré disk and form tight clusters that correspond to drug groups at ATC level 1. The hyperbolic embedding exhibits a clear hierarchical structure where the clusters at the boundary can be interpreted as distinct subtrees with the root of the tree positioned at the origin. A small number of drugs (gray 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 4(b) and (c) demonstrate that embedding in hyperbolic space can effectively induce a multilevel tree. More specifically, in Figure 4(b), we zoom into a level 1 group called "Antineoplastic and Immunodulating Agents" and show that the members in this group form clusters that correspond to level 2 groups. We further zoom into a level 2 group called "Antineoplastic Agents" (see Figure 4(c)) and demonstrate that members in this group form clusters that correspond to level 3 groups. This example demonstrates that the embedding retains the hierarchical structure to the deepest levels.

Summary. The preceding results show that best performing embedding is obtained with the Lorentz model with dimension of size 64, when both the chemical structures and the ATC hierarchy are leveraged. We refer to this embedding as the Lorentz Drug Embedding (LDE) and its Euclidean counterpart as the Euclidean Drug Embedding (EDE) in the following experiments.

Evaluation of Drug Repositioning. Drug repositioning is the discovery of new uses, called indications, for approved drugs. Compared to *de novo* drug discovery that takes an enormous amount of time, money, and effort, drug repositioning is more efficient since it takes advantage of drugs that are already approved. We evaluate LDE for drug repositioning by deriving kNN models to discriminate between approved and unapproved drug-indication pairs in the repoDB data set. 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. 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 $\{X_{\mathrm{kNN}}\}$ from the training set in the latent space. We apply majority voting to the retrieved drug-indication pairs in $\{X_{\mathrm{kNN}}\}$ to predict the status of each indication associated with $\boldsymbol{x}_i$. For indications of $\boldsymbol{x}_i$ that do not exist in $\{X_{\mathrm{kNN}}\}$, we assume that it has equal probability of being either being successfully approved or not.

Table 2 shows an example of the drug esomeprazole as the query drug for which we want to predict new indications.

Table 2. Example of Drug Repositioning Prediction for Esomeprazole Using kNN (k = 3)[a]

	query drug	retrieved drugs		
FDA status	esomeprazole	omeprazole	rabeprazole	famotidine
approved	erosive esophagitis	√		√
	zollinger-ellison syndrome	√	√	√
	peptic esophagitis	√		√
	gastresophageal reflux disease	√	√	√
	peptic ulcer	√	√	
unapproved	nausea	×		
	laryngeal diseases	×		
	cystic fibrosis	×		

[a]The first two columns show the ground-truth status of indications associated with esomeprazole. In the third column, a check mark represents one approved vote from a retrieved drug, a cross mark represents one unapproved vote from a retrieved drug, and no mark represents that the status of corresponding indication is unknown.

Esomeprazole was first approved by the FDA in 2001 and thus is not in the training set. The three most similar drugs to esomeprazole in the latent space are omeprazole, rabeprazole, and famotidine, which were approved by the FDA in 1989, 1999, and 1986, respectively. Table 2 shows that, based on the status of indications associated with the retrieved drugs, we successfully predicted all uses of esomeprazole that have been approved by the FDA. Moreover, we observe that esomeprazole is not likely to be approved for nausea, laryngeal diseases, and cystic fibrosis based on the failed approval of omeprazole for these indications. Figure 5 shows the two-dimensional molecular structures of esomeprzole and its three nearest neighbors.

Because we are not aware of any other approach developed on the repoDB data set with the same chronological split, we compare the performance of LDE for drug repositioning using kNN, for each k in [3, 5, 7, 9, 11], to the following baselines: (1) kNN on RDKit-calculated descriptors, (2) kNN on Morgan (ECFP) fingerprints (bit vector of size 2048), (3) kNN on count-based Morgan fingerprints, and (4) kNN on Lorentz drug embedding without ATC information. We use the Tanimoto coefficient as the similarity metric for fingerprints-based representations, Euclidean distance as the similarity metric for RDKit-calculated descriptors, and hyperbolic distance as the similarity metric for LDE. Performance is evaluated using area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC). Figure 6 shows that the LDE with ATC information, i.e., pairwise similarity between drugs, outperforms other drug representations by a large margin. Averaging across different k values, the LDE with ATC information surpasses Morgan (ECFP) fingerprints, the second best representation, by 12% (AUROC) and 15.8% (AUPRC). Compared to LDE without ATC information,

https://dx.doi.org/10.1021/acs.jcim.0c00681
J. Chem. Inf. Model. 2020, 60, 5647−5657

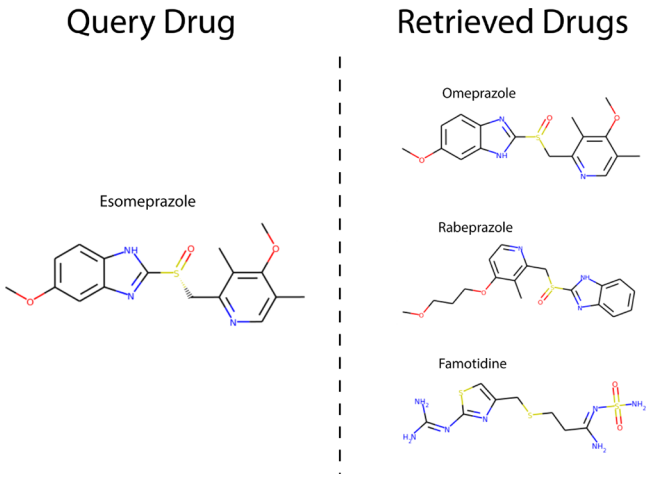


Figure 5. Molecular structures of esomeprazole and its three nearest neighbors retrieved using kNN. Among the retrieved drugs, omeprazole is closely related to esomprazole in chemical structure, and rabeprazole shares a substructure with esomprazole. Although famotidine is structurally different, it belongs to the same pharmacological group as omeprazole and rabeprazole in the ATC hierarchy.

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 is likely driven by the drug−drug similarity that is encoded in the ATC hierarchy.

We also compare the performance of drug repositioning using kNN (k = 11) between LDE and EDE. Figure 7 shows that the LDE substantially outperforms the EDE in low dimensional spaces (2 and 4) and has comparable AUC scores in a high dimensional space (64). This result shows that the hyperbolic space has superior capacity than the Euclidean space when the input hierarchy is relatively large for the latent space in which it is embedded. It is an appealing property as the drug hierarchy is expected to grow as new drugs are approved in the future.

Evaluation of Side Effect Predictions. Side effects are unwanted reactions to drugs, and they occur commonly. Often, not all side effects of a drug are known at the time it is approved for medical use. Thus, it is of critical importance to identify side effects of approved drugs. We applied LDE to predict side effects and compared its performance to several state of the art drug representations for predicting side effects.[50] We apply the side effect prediction methods to predict the presence or absence of side effects in each of the 27 classes of drugs as defined in the SIDER database. We perform three independent runs with different random seeds. In each run, we randomly split the SIDER database into training, validation, and test sets in the proportions 80%:10%:10%. We use mean AUROC as the evaluation metric.

The comparison drug representations include (1) graph-based representations including a Weave and Graph Convolutional (GC) network that represent each molecule as an undirected graph, (2) Fingerprint (ECFP) representation that is a fixed length binary encoding of topological characteristics of the molecule, and (3) our drug embeddings LDE and EDE. For LDE and EDE, we use random forest (RF) classifiers to predict side effects. For ECFP, we use influence relevance voting (IRV), a refined kNN classifier, and random forest classifiers to predict side effects. We use the Tanimoto coefficient[53] as the similarity metric for ECFP+IRV. For comparison, we use the results for Weave, GC, ECFP+IRV, and ECFP+RF from Wu et al.[50] since their experimental settings are the same as our settings.

Figure 8 shows that LDE has a significantly better performance (P-value < 0.05) in predicting side effects compared to both graph-based and ECFP representations. Compared to EDE, LDE has a slightly better, but not significantly better, performance. This result shows that incorporating the drug hierarchy in the hyperbolic embedding improves accuracy of predicting side effects.

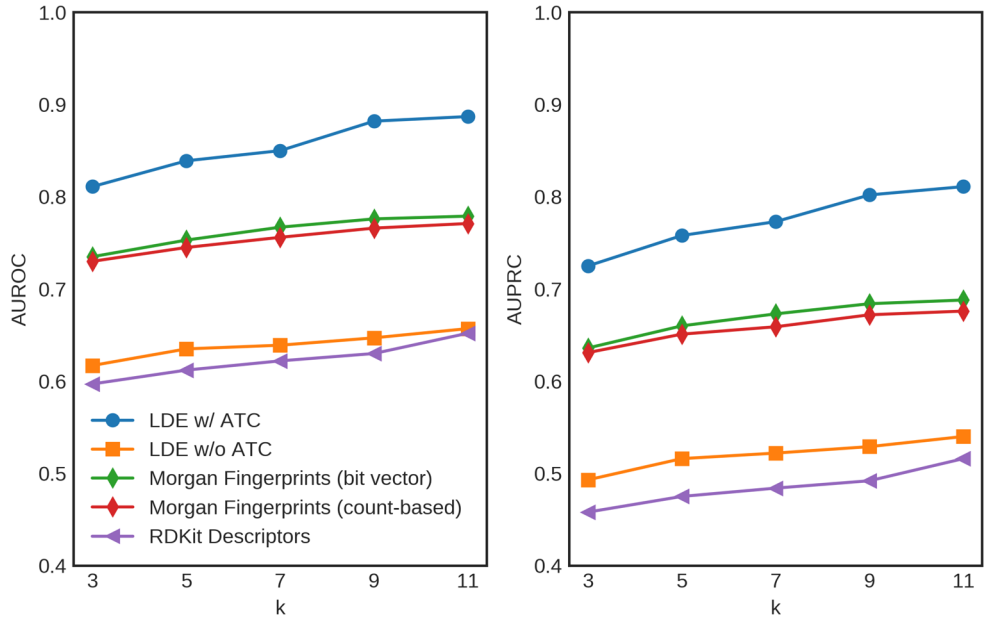


Figure 6. Comparison of representations for drug repositioning prediction using kNN ($k \in [3, 5, 7, 9, 11]$). The left panel shows AUROC scores, and the right panel shows AUPRC scores.

https://dx.doi.org/10.1021/acs.jcim.0c00681
J. Chem. Inf. Model. 2020, 60, 5647−5657

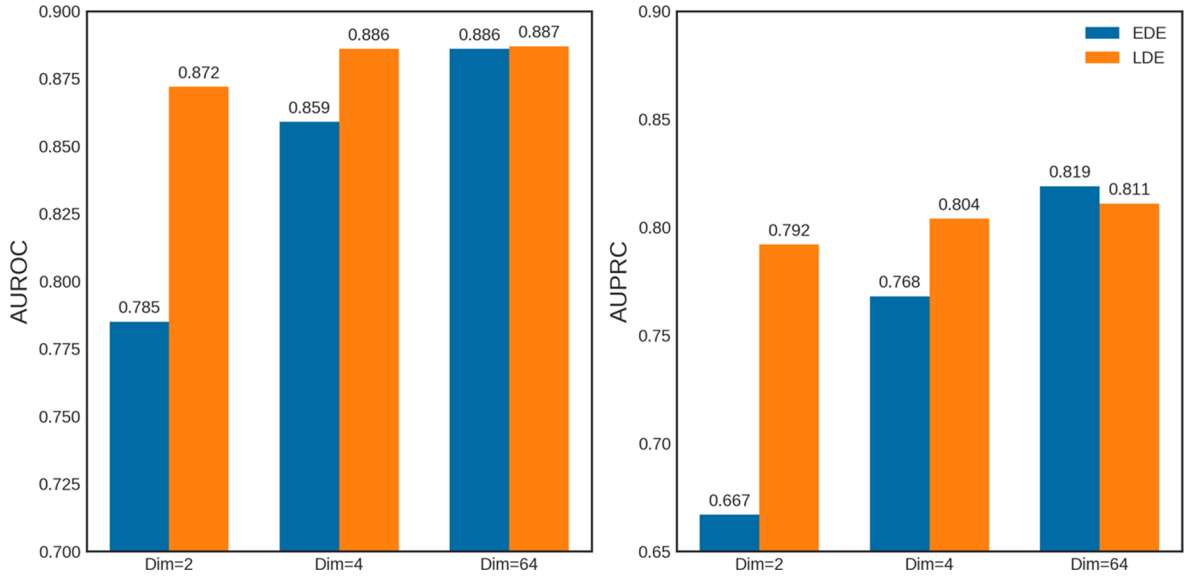


Figure 7. Comparison of LDE to EDE for drug repositioning prediction using kNN ($k = 11$) at different latent dimensions (2, 4, and 64). The left panel shows AUROC scores, and the right panel shows AUPRC scores.

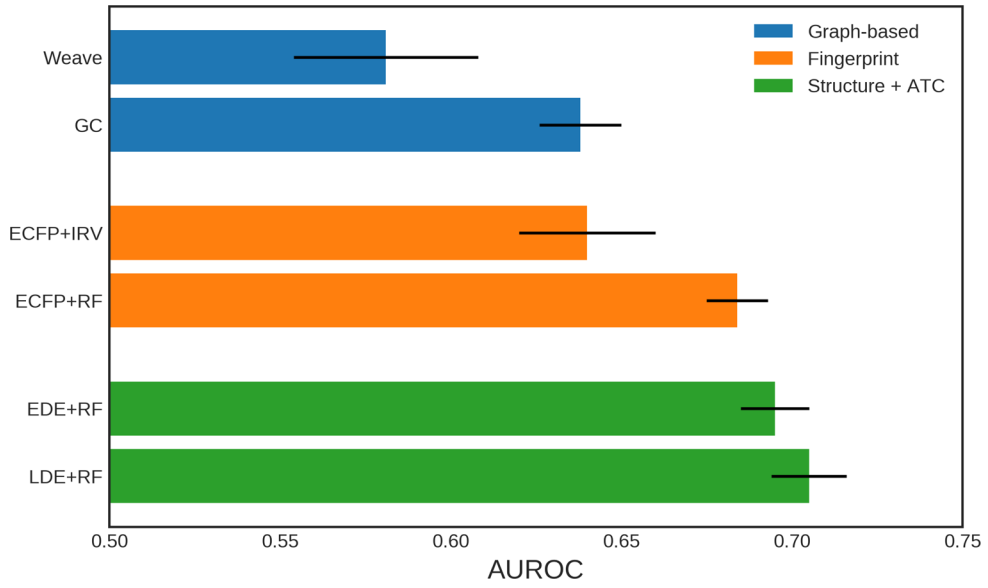


Figure 8. Prediction of side effects using three different representations of molecules: (1) Graph-based (Weave, GC), (2) Fingerprint (ECFP), and (3) our drug embeddings (EDE, LDE). The AUROC values are the average of three independent runs, and the error bars denote standard deviations.

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 empirically that our embedding recapitulates the hierarchical relationships in the ATC hierarchy and can accurately reproduce the chemical structure. Our results support that learning chemical structure can help preserve the ATC hierarchy in the latent space and vice versa. We further showed that the embedding can be used for drug repositioning and to discover new side effects.

There are several directions for future work. We plan to investigate the utility of integrating additional types of biomedical knowledge, such as drug-target interaction information, into the model. Our approach is general and can easily incorporate additional sources of knowledge. Besides, as our framework is built on a probabilistic generative model, we plan to investigate its utility for drug discovery, for example, searching new molecules that are similar to the FDA-approved drugs in a desired pharmacological class.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.0c00681.

Code and data sets, algorithmic description of proposed method, details of implementation, additional ablation studies on hyperparameter tuning, effect of sampling strategies, qualitative evaluation of Euclidean 2D drug

https://dx.doi.org/10.1021/acs.jcim.0c00681
J. Chem. Inf. Model. 2020, 60, 5647−5657

embeddings, evaluation on additional benchmark data sets (Tox21, PDBbind), and background on hyperbolic geometry (PDF)

AUTHOR INFORMATION

Corresponding Authors

Ke Yu − *Intelligent Systems Program, School of Computing and Information, University of Pittsburgh, Pittsburgh, Pennsylvania 15206, United States*; orcid.org/0000-0001-9882-5729; Email: yu.ke@pitt.edu

Shyam Visweswaran − *Intelligent Systems Program, School of Computing and Information and Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania 15206, United States*; Email: shv3@pitt.edu

Kayhan Batmanghelich − *Intelligent Systems Program, School of Computing and Information and Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania 15206, United States*; Email: kayhan@pitt.edu

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.0c00681

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The research reported in this publication was supported by the National Institutes of Health under award numbers 1R01HL141813-01 and R01LM012095, the National Science Foundation under award number 1839332 Tripod+X, the SAP SE, and a Provost Fellowship in Intelligent Systems at the University of Pittsburgh (awarded to K.Y.). We thank the support of NVIDIA Corporation for the donation of the Titan X Pascal GPU that was used for this research. This work used the Bridges system, which is supported by NSF award number TG-ASC170024, at the Pittsburgh Supercomputing Center.

REFERENCES

(1) Nosengo, N. New tricks for old drugs. *Nature* **2016**, *534*, 314−316.

(2) Pushpakom, S.; Iorio, F.; Eyers, P. A.; Escott, K. J.; Hopper, S.; Wells, A.; Doig, A.; Guilliams, T.; Latimer, J.; McNamee, C.; Norris, A.; Sanseau, P.; Cavalla, D.; Pirmohamed, M. Drug repurposing: progress, challenges and recommendations. *Nat. Rev. Drug Discovery* **2019**, *18*, 41−58.

(3) Liu, M.; Wu, Y.; Chen, Y.; Sun, J.; Zhao, Z.; Chen, X.; Matheny, M. E.; Xu, H. Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. *J. Am. Med. Inform. Assoc.* **2012**, *19*, e28−e35.

(4) Walters, W. P.; Murcko, M. Assessing the impact of generative AI on medicinal chemistry. *Nat. Biotechnol.* **2020**, *38*, 143−145.

(5) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370−3388.

(6) Segler, M. H.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* **2018**, *4*, 120−131.

(7) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **2018**, *4*, 268−276.

(8) Yu, L.; Ma, X.; Zhang, L.; Zhang, J.; Gao, L. Prediction of new drug indications based on clinical data and network modularity. *Sci. Rep.* **2016**, *6*, 32530.

(9) Timilsina, M.; Tandan, M.; d'Aquin, M.; Yang, H. Discovering Links Between Side Effects and Drugs Using a Diffusion Based Method. *Sci. Rep.* **2019**, *9*, 10436.

(10) Zitnik, M.; Agrawal, M.; Leskovec, J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* **2018**, *34*, i457−i466.

(11) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **1988**, *28*, 31−36.

(12) Nickel, M.; Kiela, D. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems* **2017**, 6338−6347.

(13) Mathieu, E.; Le Lan, C.; Maddison, C. J.; Tomioka, R.; Teh, Y. W. Continuous Hierarchical Representations with Poincaré Variational Auto-Encoders. *Advances in neural information processing systems* **2019**, 12544−12555.

(14) De Sa, C.; Gu, A.; Ré, C.; Sala, F. Representation tradeoffs for hyperbolic embeddings. *J. Mach. Learn. Res.* **2018**, *80*, 4460.

(15) Monath, N.; Zaheer, M.; Silva, D.; McCallum, A.; Ahmed, A. Gradient-based Hierarchical Clustering using Continuous Representations of Trees in Hyperbolic Space. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* **2019**, 714−722.

(16) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547−555.

(17) Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discovery* **2019**, *18*, 463−477.

(18) Ekins, S.; Puhl, A. C.; Zorn, K. M.; Lane, T. R.; Russo, D. P.; Klein, J. J.; Hickey, A. J.; Clark, A. M. Exploiting machine learning for end-to-end drug discovery and development. *Nat. Mater.* **2019**, *18*, 435.

(19) Jeon, J.; Nim, S.; Teyra, J.; Datti, A.; Wrana, J. L.; Sidhu, S. S.; Moffat, J.; Kim, P. M. A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening. *Genome Med.* **2014**, *6*, 57.

(20) Ferrero, E.; Dunham, I.; Sanseau, P. In silico prediction of novel therapeutic targets using gene−disease association data. *J. Transl. Med.* **2017**, *15*, 182.

(21) Rouillard, A. D.; Hurle, M. R.; Agarwal, P. Systematic interrogation of diverse Omic data reveals interpretable, robust, and generalizable transcriptomic features of clinically successful therapeutic targets. *PLoS Comput. Biol.* **2018**, *14*, e1006142.

(22) Li, H.; Sze, K.-H.; Lu, G.; Ballester, P. J. Machine-learning scoring functions for structure-based drug lead optimization. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2020**, *10*, e1465.

(23) Winter, R.; Montanari, F.; Steffen, A.; Briem, H.; Noé, F.; Clevert, D.-A. Efficient multi-objective molecular optimization in a continuous latent space. *Chem. Sci.* **2019**, *10*, 8016−8024.

(24) Unterthiner, T.; Mayr, A.; Klambauer, G.; Hochreiter, S. Toxicity prediction using deep learning. 2015, arXiv preprint. arXiv:1503.01445. https://arxiv.org/abs/1503.01445 (accessed 2020-10-30).

(25) Li, B.; Shin, H.; Gulbekyan, G.; Pustovalova, O.; Nikolsky, Y.; Hope, A.; Bessarabova, M.; Schu, M.; Kolpakova-Hart, E.; Merberg, D.; et al. Development of a drug-response modeling framework to identify cell line derived translational biomarkers that can predict treatment outcome to erlotinib or sorafenib. *PLoS One* **2015**, *10*, e0130700.

(26) van Gool, A. J.; Bietrix, F.; Caldenhoven, E.; Zatloukal, K.; Scherer, A.; Litton, J.-E.; Meijer, G.; Blomberg, N.; Smith, A.; Mons, B.; et al. Bridging the translational innovation gap through good biomarker practice. *Nat. Rev. Drug Discovery* **2017**, *16*, 587−588.

(27) Kraus, V. B. Biomarkers as drug development tools: discovery, validation, qualification and use. *Nat. Rev. Rheumatol.* **2018**, *14*, 354−362.

(28) Kwak, H.; Lee, M.; Yoon, S.; Chang, J.; Park, S.; Jung, K. Drug-Disease Graph: Predicting Adverse Drug Reaction Signals via Graph Neural Network with Clinical Data. *Pacific-Asia Conference on Knowledge Discovery and Data Mining* **2020**, *12085*, 633−644.
(29) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(30) Kusner, M. J.; Paige, B.; Hernández-Lobato, J. M. Grammar variational autoencoder. *Proceedings of the 34th International Conference on Machine Learning-Volume 70* **2017**, 1945−1954.
(31) Gupta, A.; Müller, A. T.; Huisman, B. J.; Fuchs, J. A.; Schneider, P.; Schneider, G. Generative recurrent networks for de novo drug design. *Mol. Inf.* **2018**, *37*, 1700111.
(32) Simonovsky, M.; Komodakis, N. Graphvae: Towards generation of small graphs using variational autoencoders. *International Conference on Artificial Neural Networks* **2018**, *11139*, 412−422.
(33) Liu, Q.; Allamanis, M.; Brockschmidt, M.; Gaunt, A. Constrained graph variational autoencoders for molecule design. *Advances in neural information processing systems* **2018**, 7795−7804.
(34) Alperstein, Z.; Cherkasov, A.; Rolfe, J. T. All smiles variational autoencoder. 2019, arXiv preprint. arXiv:1905.13343. https://arxiv.org/abs/1905.13343 (accessed 2020-10-30).
(35) Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminf.* **2015**, *7*, 20.
(36) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046−1053.
(37) Samanta, S.; O'Hagan, S.; Swainston, N.; Roberts, T. J.; Kell, D. B. VAE-Sim: a novel molecular similarity measure based on a variational autoencoder. *Molecules* **2020**, *25*, 3446.
(38) Goyal, P.; Hu, Z.; Liang, X.; Wang, C.; Xing, E. P. Nonparametric variational auto-encoders for hierarchical representation learning. *Proceedings of the IEEE International Conference on Computer Vision* **2017**, 5094−5102.
(39) Helgason, S. *Differential geometry and symmetric spaces*; American Mathematical Soc.: 2001; Vol. *341*.
(40) Nickel, M.; Kiela, D. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. 2018, arXiv preprint. arXiv:1806.03417. https://arxiv.org/abs/1806.03417 (accessed 2020-10-30).
(41) Nagano, Y.; Yamaguchi, S.; Fujita, Y.; Koyama, M. A Wrapped Normal Distribution on Hyperbolic Space for Gradient-Based Learning. *International Conference on Machine Learning* **2019**, 4693−4702.
(42) Jordan, M. I.; Ghahramani, Z.; Jaakkola, T. S.; Saul, L. K. An introduction to variational methods for graphical models. *Mach. Learn.* **1999**, *37*, 183−233.
(43) He, J.; Spokoyny, D.; Neubig, G.; Berg-Kirkpatrick, T. Lagging inference networks and posterior collapse in variational autoencoders. 2019, arXiv preprint. arXiv:1901.05534. https://arxiv.org/abs/1901.05534 (accessed 2020-10-30).
(44) Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; Lerchner, A. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *ICLR* **2017**, *2*, 6.
(45) Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. 2014, arXiv preprint. arXiv:1406.1078. https://arxiv.org/abs/1406.1078 (accessed 2020-10-30).
(46) Kingma, D. P.; Welling, M. Auto-encoding variational bayes. 2013, arXiv preprint. arXiv:1312.6114. https://arxiv.org/abs/1312.6114 (accessed 2020-10-30).
(47) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: a free tool to discover chemistry for biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757−1768.
(48) Organization, W. H. *WHO Collaborating Centre for Drug Statistics Methodology, ATC classification index with DDDs*; World Health Organization Collaborating Centre for Drug Statistics Methodology: 2014.
(49) Kuhn, M.; Letunic, I.; Jensen, L. J.; Bork, P. The SIDER database of drugs and side effects. *Nucleic Acids Res.* **2016**, *44*, D1075−D1079.
(50) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **2018**, *9*, 513−530.
(51) Brown, A. S.; Patel, C. J. A standard database for drug repositioning. *Sci. Data* **2017**, *4*, 170029.
(52) Heller, K. A.; Ghahramani, Z. Bayesian hierarchical clustering. *Proceedings of the 22nd international conference on Machine learning* **2005**, 297−304.
(53) Butina, D. Unsupervised data base clustering based on daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets. *J. Chem. Inf. Model.* **1999**, *39*, 747−750.

https://dx.doi.org/10.1021/acs.jcim.0c00681
J. Chem. Inf. Model. 2020, 60, 5647−5657