A multi-view generative model for molecular representation improves prediction tasks

Jonathan Yin
Yale University
New Haven, CT 06520
jonathan.yin@yale.edu

Hattie Chung
Broad Institute
Cambridge, MA 02142
hchung@broadinstitute.org

Aviv Regev
Klarman Cell Observatory, Broad Institute
Cambridge, MA 02142
aregev@broadinstitute.org

Abstract

Unsupervised generative models have been a popular approach to representing molecules. These models extract salient molecular features to create compact vectors that can be used for downstream prediction tasks. However, current generative models for molecules rely mostly on structural features and do not fully capture global biochemical features. Here, we propose a multi-view generative model that integrates low-level structural features with global chemical properties to create a more holistic molecular representation. In proof-of-concept analyses, compared to purely structural latent representations, multi-view latent representations improve model accuracy on various tasks when used as input to feed-forward prediction networks. For some tasks, simple models trained on multi-view representations perform comparably to more complex supervised methods. Multi-view representations are an attractive method to improve representations in an unsupervised manner, and could be useful for prediction tasks, particularly in contexts where data is limited.

1 Introduction

1.1 Deep learning models for molecular representation

Recent advances in machine learning have revolutionized molecular representations, specifically by identifying features that are important for function. Supervised deep learning models in particular have enjoyed large success for molecular property prediction [1,2]. These models learn molecular features that are highly attuned to a specific function, but these learned features may not generalize well to other tasks. Furthermore, they typically require significant quantities of labeled training data as they are trained on raw, low-level representations. Thus, developing a more abstract, meaningful molecule-intrinsic representation could facilitate training and lead to more accurate prediction across a wide array of tasks.

In recent years, variational autoencoders (VAEs) have become popular for creating broadly-tuned unsupervised molecular representations [3,4,5,6]. Since the features encoded in the learned representation are not function-specific, the latent spaces created by VAEs could be suitable to use as
Figure 1: Architecture of multi-view two-tower VAE. For property prediction, molecules are first encoded into their latent molecular representation and then passed as input to a separate prediction model.

1.2 Multi-view representation fusion

Different views of the same data—e.g. RNA and protein, or audio and video, or image and text—often contain complementary information. Multi-view representation fusion, a form of multi-view representation learning, seeks to integrate data across multiple modalities into a single, more comprehensive representation [8, 9, 10]. This joint representation then facilitates the development of prediction models by enabling useful information to be readily extracted. One example application of fusion-based multi-view learning is a bimodal deep autoencoder that combines audio and video data into a shared representation [11].

Here, we leverage multi-view representation fusion to improve molecular representations in unsupervised generative models so that they are more useful for property predictions and de novo design.

2 Two-tower molecule VAE for multi-view molecular representation

We present a bimodal “two-tower” VAE model that simultaneously encodes low-level structural information and global-level molecular features that reflect emergent properties of a molecule (Figure 1). By integrating these two distinct but complementary views, we aim to create a richer shared representation that can better capture the functional properties of a molecule based on its structure.

Our two-tower models build on two published single-tower VAEs, a character VAE (CharVAE) [4] and a grammar VAE (GVAE) [6]. Both of these published models use a string-based structural representation known as the simplified molecular-input line-entry system (SMILES) [12] as training input. The first tower of our two-tower VAE mimics these single-tower VAEs to capture local structural features. The second tower encodes global chemical properties that are challenging to capture at a local level. To ensure that our second tower encoded properties applicable to various learning tasks, we used a set of 200 RDKit features [13] similar to the approach taken by a supervised model Chemprop [1]. These features include the number of aliphatic carbocycles or the number of radical electrons in a molecule.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Description</th>
<th>No. of tasks</th>
<th>Task type</th>
<th>No. of compounds</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESOL</td>
<td>Water solubility</td>
<td>1</td>
<td>Regression</td>
<td>1,128</td>
<td>RMSE</td>
</tr>
<tr>
<td>FreeSolv</td>
<td>Hydration free energy in water</td>
<td>1</td>
<td>Regression</td>
<td>642</td>
<td>RMSE</td>
</tr>
<tr>
<td>Lipophilicity</td>
<td>Octanol/water distribution coefficients</td>
<td>1</td>
<td>Regression</td>
<td>4,200</td>
<td>RMSE</td>
</tr>
<tr>
<td>HIV</td>
<td>Inhibition of HIV replication</td>
<td>1</td>
<td>Classification</td>
<td>41,127</td>
<td>ROC-AUC</td>
</tr>
<tr>
<td>BACE</td>
<td>Inhibition of human (\beta)-secretase 1</td>
<td>1</td>
<td>Classification</td>
<td>1,513</td>
<td>ROC-AUC</td>
</tr>
<tr>
<td>BBBP</td>
<td>Toxicity</td>
<td>1</td>
<td>Classification</td>
<td>2,039</td>
<td>ROC-AUC</td>
</tr>
<tr>
<td>Tox21</td>
<td>Toxicity</td>
<td>12</td>
<td>Classification</td>
<td>7,831</td>
<td>ROC-AUC</td>
</tr>
<tr>
<td>Clintox</td>
<td>Toxicity</td>
<td>2</td>
<td>Classification</td>
<td>1,478</td>
<td>ROC-AUC</td>
</tr>
<tr>
<td>SIDER</td>
<td>Side effects of drugs</td>
<td>27</td>
<td>Classification</td>
<td>1,427</td>
<td>ROC-AUC</td>
</tr>
</tbody>
</table>

To construct a latent representation of the combined modalities, inputs for each tower are first individually sent through several encoder layers before being passed simultaneously through a single encoder. This produces the shared latent representation that is used as input for property prediction (see 3.1 Experiments). Decoding is conducted similar to encoding; vectors in the latent space are first passed through a single decoder and then through two separate decoders corresponding to the structural input and RDKit features.

We trained our VAE model on 250,000 drug-like molecules from the ZINC molecule dataset \([14]\), with 5,000 randomly selected molecules held out for testing (see 3.2 Experiments). Molecular features for the second tower were generated using DescriptaStorus \([1, 2]\).

### 3 Experiments

#### 3.1 Property prediction using multi-view latent representations

We used 9 publicly available datasets from MoleculeNet \([15]\) to evaluate how well the multi-view latent space from our VAEs could predict molecular properties (Table 1). For each dataset, we mapped the input molecules to their latent vectors using the encoder portion of each VAE. We then fed these embeddings into simple feed-forward networks (FFNs) to predict the desired output properties. Latent spaces resulting in higher prediction accuracy were interpreted as better molecular representations since the same FFN was used for task prediction across all input representations.

Across all 9 data sets, the FFNs trained on the latent representations of two-tower models outperformed those trained on the latent representations of single-tower models. For some datasets (BBBP, ESOL), the basic FFNs trained on the unsupervised two-tower encodings even performed comparably to Chemprop \([1, 2]\), a state-of-the-art supervised model (Figure 2). This indicates that the inclusion of a second tower improved the latent space’s utility for property prediction.

#### 3.2 Molecule reconstruction and validity

Next, we verified that the addition of the second tower did not negatively affect the ability to reconstruct molecules from their latent representations. We determined reconstruction accuracy and validity as previously documented \([5]\). Each molecule was encoded and decoded 10 times, and the proportion of the 100 decoded molecules identical to the input molecule was reported. We computed validity by sampling 1,000 latent vectors from the prior distribution \(N(0, I)\) and decoding each of these vectors 100 times. Our two-tower model achieved higher reconstruction accuracy than the single-tower counterparts, and maintained prior validities comparable to single-tower models (Table 2).

\(^1\) with hyperparameters matching those used in this study, primarily an increase in latent space size
Figure 2: Prediction accuracies of feed-forward networks (FFNs) trained using various latent representations as input: structural representations (light blue/orange), multi-view representations (dark blue/orange). Prediction accuracy from Chemprop, a supervised model, is shown for reference (green bars). **Left:** classification datasets from MoleculeNet on x-axis, ROC-AUC on y-axis. **Right:** regression datasets from MoleculeNet on x-axis, root mean squared error (RMSE) shown on y-axis.

Table 2: Reconstruction accuracy and sample validity results.

<table>
<thead>
<tr>
<th>Method</th>
<th>Reconstruction</th>
<th>Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>CharVAE</td>
<td>45.6%</td>
<td>0.1%</td>
</tr>
<tr>
<td>Published CharVAE</td>
<td>44.6%</td>
<td>0.7%</td>
</tr>
<tr>
<td>Two-tower CharVAE</td>
<td>69.3%</td>
<td>0.2%</td>
</tr>
<tr>
<td>GVAE</td>
<td>52.7%</td>
<td>5.1%</td>
</tr>
<tr>
<td>Published GVAE</td>
<td>53.7%</td>
<td>7.2%</td>
</tr>
<tr>
<td>Two-tower GVAE</td>
<td>61.8%</td>
<td>4.9%</td>
</tr>
</tbody>
</table>

Figure 3: **Left:** random molecules sampled from prior distribution \(N(0, I)\) using the two-tower GVAE model. **Right:** visualizing the local neighborhood of the two-tower GVAE model, starting from the molecule in the center (outlined in black).
We also qualitatively examined the latent space of the two-tower GVAE for continuity in the neighborhood of a molecule (Figure 3). Following a previously published approach [6], we generated 2 random orthogonal unit vectors in latent space and moved in combinations of these directions to create a grid of latent vectors that were then decoded into molecules. Despite training the model on both structural and chemical features, we still observe smooth transitions between molecules in the grid, often by a single atom at a time.

4 Discussion

We present a multi-view generative model for molecular representation using a two-tower VAE. We show that the resulting latent representation improves utility in predicting multiple independent molecular properties compared to single-tower VAEs and may also offer a more interpretable latent space. In the future, we aim to explore the generative capabilities of our two-tower model using Bayesian optimization, and implement a two-tower approach built with graphical approaches to molecular representation.

Author Contributions

H.C. and J.Y. conceived and designed the study. J.Y. developed and implemented the analyses with supervision from H.C. A.R. provided project oversight. J.Y., H.C., and A.R. wrote the manuscript.

Corresponding authors

Correspondence to H.C. (hchung@broadinstitute.org) and A.R. (aregev@broadinstitute.org)

Acknowledgements

We would like to thank Michael Truell and John Ingraham for helpful discussions.

Competing Interests Statement

A.R. is a founder and equity holder of Celsius Therapeutics, an equity holder in Immunitas Therapeutics, and until August 31, 2020 was an SAB member of Syros Pharmaceuticals, Neogene Therapeutics, Asimov and ThermoFisher Scientific. From August 1, 2020, A.R. is an employee of Genentech.

References


