Evaluating chemical descriptors with the loss-data framework

Walid Ahmad
Reverie Labs
walid@reverielabs.com

Abstract

Selecting chemical representations for machine learning models is a challenging task, one which is subject to trial and error. Extended connectivity fingerprints (ECFPs), for example, are a classic featurization technique used widely in the prediction of properties of molecules, which effectively encodes molecular substructures in a bit vector. However, it is not obvious when to use vanilla ECFPs instead of alternative featurizations. We propose using recent progress in the field of representation learning to evaluate and improve the quality of chemical descriptors. Specifically, we show that using the loss-data framework with surplus description length and ε sample complexity can provide insight into which descriptors are appropriate for specific tasks. We find that applying simple dimensionality reduction techniques such as UMAP and PCA can improve the quality of ECFP descriptors for certain datasets.

1 Introduction

Machine learning (ML) approaches have seen a recent surge in cheminformatics because of their utility throughout different stages of the drug discovery pipeline. In particular, supervised machine learning has shown promise for predicting molecular characteristics, such as ADMET properties and binding affinities [11]. However, developing effective ML approaches for supervised tasks requires not only a good model, but also a good representation, or featurization, of chemical compounds. The quality of the featurization affects how well, and how quickly, a machine learning model can learn from the data. Which representation to use, and when, is an open question, particularly in small molecule design problems, where a wide variety of representations are available [15]. In this work, we investigate the effectiveness of common, one-dimensional featurizations of molecules and demonstrate how to assess different descriptors.

1.1 Molecular descriptors

Molecular descriptors are operations that transform a symbolic representation of a molecule into a vectorized representation. The vectorized representations can be used as inputs to algorithms, such as ML models, to assist in a variety of downstream tasks, including modeling quantitative structure activity relationships (QSAR) for virtual screens [3, 7, 10].

Simple computed properties A basic class of molecular descriptors includes those constructed using predicted physiochemical properties or experimental measurements. Popular measurements used in these representations include molecular weight, logP, number of hydrogen bond donors, number of hydrogen bond acceptors, polar surface area, and more. The values of interest can be concatenated into a vector and used as a molecular representation. Models trained with such representations may be largely invariant to the underlying molecular structures.

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

Extended connectivity fingerprints Extended connectivity fingerprints (ECFPs) are one of the most widely used molecular descriptors. ECFPs are circular fingerprints that utilize a variant of the Morgan algorithm [13] to quantize neighborhoods around individual atoms to detect the presence of substructures. The molecule is represented as a bit-vector where on-bits indicate the presence of a particular substructure. ECFPs can contain varying levels of granularity, based on the number of bits that are used [14].

ML-based descriptors ML models can be used to construct chemical descriptors by extracting the outputs at a given layer in the model. Encoder-decoders are the canonical architecture for constructing descriptors in this way. An encoder network compresses an input representation of a molecule into a latent vector, from which a decoder reconstructs a molecular representation. The input and output representation can be the same (e.g. SMILES strings [6]) or different (e.g. SMILES to InChI [17]).

1.2 Dimensionality reduction techniques

Principal component analysis (PCA) PCA [9] is a linear dimensionality reduction technique which projects data into into a lower-dimensional space using singular value decomposition (SVD). Principal components are ordered eigenvectors of the covariance matrix, and thus maximize the variance of the projected data.

Uniform manifold approximation and projection (UMAP) UMAP [12] is a relatively recent dimensionality reduction technique which claims to preserve both the local and global structure of the data. It has received attention in bioinformatics, for example, for its utility in visualizing single-cell data [2].

1.3 Evaluating representations

The increased interest in representation and self-supervised learning in the machine learning community has given rise to a family of techniques for analyzing the quality of representations. Notions of robustness for representations normally entail evaluating downstream performance on some relevant task, using simple models that are quick to train – these models are sometimes referred to as “probes" in the literature. Probes can be linear [1, 5] or nonlinear [4].

Nonetheless, the question of representation evaluation is complicated. A particular representation may work reasonably well in certain data regimes, and for certain tasks, but not for others. In Whitney et al. [16], the authors provide an overview of current state-of-the-art methods for evaluating representations, and identify the *loss-data framework* as a useful analytical tool. The loss-data curve, which plots validation loss versus training set size, elucidates how well probes can learn with different numbers of examples. This is in contrast to a traditional loss-curve, which holds the number of examples static. Whitney et al. [16] also propose two new, robust methods for representation evaluation based on the loss-data curve, which we use here.

Surplus description length Surplus description length (SDL) is a measure of the extra entropy needed to encode data from a data generating distribution $\mathcal{D}$ using a representation ϕ. On a dataset with i points, the SDL is

$$m_{\text{SDL}}(\phi, \mathcal{D}, \mathcal{A}) = \sum_{i=1}^{N} \left[L(\mathcal{A}_{\phi}, i) - \varepsilon\right]_{+}, \tag{1}$$

where $\mathcal{A}$ is the probe algorithm, L is the expected loss, and ε is a success criterion, i.e. a loss tolerance for which a model is considered successful at the task.

ϵ sample complexity (SC) ϵSC measures the smallest number of samples needed for a probe $\mathcal{A}$ to achieve a loss value of ε on the dataset,

$$m_{\epsilon\text{SC}}(\phi, \mathcal{D}, \mathcal{A}) = \min\{n \in \text{N} : L(\mathcal{A}_{\phi}, n) \leq \varepsilon\}. \tag{2}$$

Both of these methods involve selecting a loss threshold, ε, which corresponds to a line on the y-axis of the loss-data curve (see Fig. 1a or 2a for an example). εSC measures the number of data points needed to obtain ε loss, and SDL integrates the loss curve above ε.

2 Experiments

We conduct an expository analysis of chemical descriptors using SDL and εSC, by considering a subset of the MoleculeNet [18] benchmark datasets: the Tox21 and Lipophilicity datasets. For each task in each dataset, we sample $n \leq N$ datapoints, where n is chosen linearly from $[0, N]$. Every sample is used to train a nonlinear probe with a 90/10 training/validation split. The probe is a simple feed-forward neural network with 2 hidden layers of size 512 each. The validation loss at each sampled n is used to compute SDL and εSC. We repeat the process with 4 different random seeds. Two ε values are selected by taking 1.5x and 2.0x the lowest loss value found (l_{min}) for each set of descriptors, to serve as proxies for success criteria.

The descriptors evaluated include ECFP descriptors of varying lengths: 1024, 2048, and 4096. For each length, we also fit and apply PCA and UMAP to project the fingerprints into a space of dimension $d \in \{2, 16, 128\}$. Lastly, we also evaluate Continuous Data-Driven Descriptors from Winter et al. [17], which are 512-dimensional latent vectors from a SMILES autoencoder architecture, trained on ZINC12 [8]. For brevity, the results shown only include ECFP descriptors of size 2048 and associated dimensionality techniques. We observed qualitatively similar results in using ECFP descriptors of different lengths. No dimensionality reduction techniques are applied to the CDDD descriptors.

3 Results

The loss-data framework proves useful in identifying which representations are appropriate for each task. For the Tox21 dataset, where UMAP and PCA projections visually indicate separation between active and inactive compounds, those representations are found to be the most expressive. Conversely, for the Lipophilicity dataset, the transformations show no benefit over ECFP, and the ML-based CDDD descriptors are the most expressive.

3.1 Tox21

Tox21 contains qualitative (binary classification) toxicity measurements from 12 biological targets including nuclear receptor signaling and stress response pathways. SDL and εSC are computed using log-loss for each descriptor. The average ranks for each type of descriptor by method are shown in Table 1. For these Tox21 tasks, the UMAP projections are by-and-large the best representations, followed by the PCA projections. We also note that the ML-based descriptors (CDDD) are not objectively better than the ECFP descriptors.

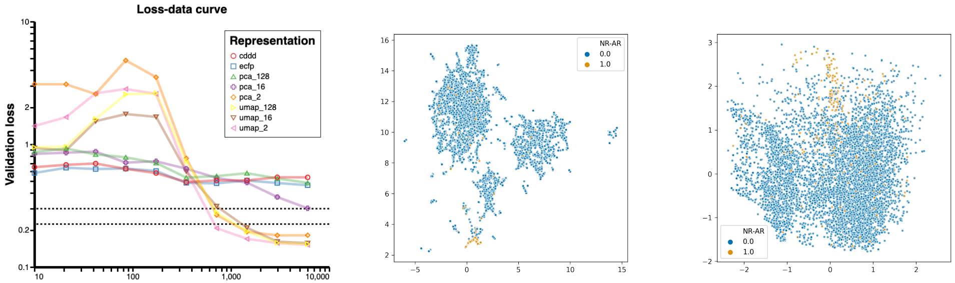
Method	CDDD	ECFP	PCA$_{128}$	PCA$_{16}$	PCA$_{2}$	UMAP$_{128}$	UMAP$_{16}$	UMAP$_{2}$
SDL, $\varepsilon = 1.5 \times l_{min}$	6.75 ± 1.29	6.75 ± 0.87	7.25 ± 0.62	5.25 ± 0.45	3.5 ± 1.00	2.08 ± 0.67	$\mathbf{1.92 \pm 1.08}$	2.50 ± 1.09
SDL, $\varepsilon = 2.0 \times l_{min}$	6.25 ± 2.14	6.50 ± 1.00	7.17 ± 0.72	4.92 ± 1.09	4.08 ± 1.73	2.33 ± 0.98	$\mathbf{1.92 \pm 1.08}$	2.83 ± 1.47
εSC, $\varepsilon = 1.5 \times l_{min}$	6.46 ± 0.14	6.46 ± 0.14	6.46 ± 0.14	6.46 ± 0.14	2.63 ± 1.23	2.79 ± 0.50	3.21 ± 0.58	$\mathbf{1.54 \pm 0.58}$
εSC, $\varepsilon = 2.0 \times l_{min}$	6.50 ± 0.00	6.50 ± 0.00	6.50 ± 0.00	6.50 ± 0.00	2.33 ± 0.83	2.71 ± 0.54	3.21 ± 0.58	$\mathbf{1.75 \pm 0.69}$

Table 1: Average rank for each representation, across the 12 classification tasks in Tox21.

As an illustrative example, we show the loss-data curve for the androgen receptor (NR-AR) task in Fig. 1a, and the two-dimensional UMAP and PCA transformations (Fig. 1b and Fig. 1c). The two-dimensional embeddings separate many of the active compounds from the inactives, reflecting that SDL εSC have selected reasonable representations.

3.2 Lipophilicity

The Lipophilicity dataset contains regression values for experimental measurements of the octanol/water distribution coefficient (logD). We again evaluate all descriptors, using root-mean-squared-error (RMSE) as the loss function, and plot the loss-data curve in Fig. 2a. Here, the CDDD descriptors are the best representation, followed by ECFP. Only the high-dimensional PCA projection is competitive with the ECFP descriptors. In this case, the two-dimensional embeddings (Fig. 2b and Fig. 2c) are not informative, reflecting the results captured by the loss-data curve.

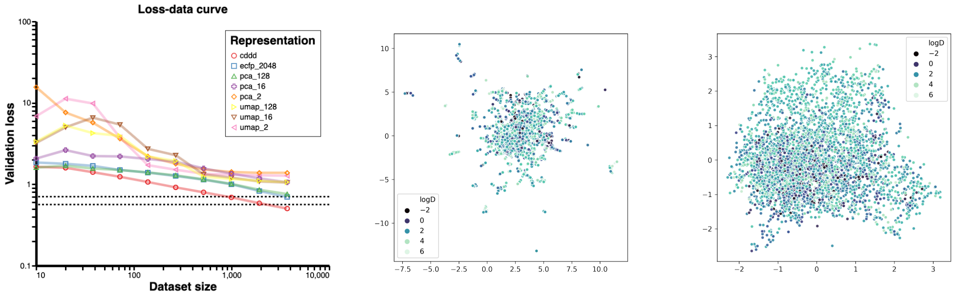


(a) Loss-data curve.

(b) UMAP projection.

(c) PCA projection.

Figure 1: The Tox21 NR-AR task. (a) The loss-data curve shows that UMAP transformations are found to be the best representations based on SDL and εSC, followed by PCA. (b) UMAP groups active compounds separately from inactives. (c) PCA similarly embeds active compounds close to one another.



(a) Loss-data curve.

(b) UMAP projection.

(c) PCA projection.

Figure 2: The logD task associated with the Lipophilicity dataset. (a) The loss-data curve shows that the CDDD descriptors are the best representations based on SDL and εSC, followed by ECFP. (b) The UMAP embedding shows no separation by experimental value. (c) The PCA embedding is also uninformative for experimental logD.

4 Discussion

In this work, we consider the problem of evaluating arbitrary chemical descriptors using tools from representation learning. We show the utility of the loss-data framework with SDL and εSC, by examining ECFP descriptors, PCA and UMAP transformations, and an example of ML-based descriptors.

The loss-data framework can also be used to understand which descriptors work well in low-data regimes – a common problem that cheminformatics practitioners face. Further work in this domain can also help inform which descriptors to use for different families of tasks.

Identifying which representations are useful, and when, is of practical importance. ECFPs, in particular suffer from the curse of dimensionality; even though longer bit-vector representations are desirable for their granularity, they are very sparse. This makes it both more difficult for machine learning models to learn a generalized notion of chemical structure, and more computationally intensive. Evaluating representations via probes can help identify when alternative descriptors or computationally efficient transformations such as UMAP or PCA will be beneficial.

Introducing this type of representation evaluation as a first-step in QSAR modeling pipelines can help reduce training time, increase interpretability, and performance.

References

[1] G. Alain and Y. Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

[2] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell. Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, 37(1):38–44, 2019.

[3] A. Cereto-Massagué, M. J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé, and G. Pujadas. Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58–63, 2015.

[4] A. Conneau, G. Kruszewski, G. Lample, L. Barrault, and M. Baroni. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*, 2018.

[5] A. Ettinger, A. Elgohary, and P. Resnik. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139, 2016.

[6] 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 central science*, 4(2):268–276, 2018.

[7] G. Hu, G. Kuang, W. Xiao, W. Li, G. Liu, and Y. Tang. Performance evaluation of 2d fingerprint and 3d shape similarity methods in virtual screening. *Journal of chemical information and modeling*, 52(5):1103–1113, 2012.

[8] J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad, and R. G. Coleman. Zinc: a free tool to discover chemistry for biology. *Journal of chemical information and modeling*, 52(7): 1757–1768, 2012.

[9] I. T. Jolliffe. Principal components in regression analysis. In *Principal component analysis*, pages 129–155. Springer, 1986.

[10] S. Kwon, H. Bae, J. Jo, and S. Yoon. Comprehensive ensemble in qsar prediction for drug discovery. *BMC bioinformatics*, 20(1):521, 2019.

[11] H. Li, C. Yap, C. Ung, Y. Xue, Z. Li, L. Han, H. Lin, and Y. Z. Chen. Machine learning approaches for predicting compounds that interact with therapeutic and admet related proteins. *Journal of pharmaceutical sciences*, 96(11):2838–2860, 2007.

[12] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[13] H. L. Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5(2): 107–113, 1965.

[14] D. Rogers and M. Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

[15] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, et al. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 18(6):463–477, 2019.

[16] W. F. Whitney, M. J. Song, D. Brandfonbrener, J. Altosaar, and K. Cho. Evaluating representations by the complexity of learning low-loss predictors. *arXiv preprint arXiv:2009.07368*, 2020.

[17] R. Winter, F. Montanari, F. Noé, and D.-A. Clevert. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chemical science*, 10 (6):1692–1701, 2019.

[18] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2): 513–530, 2018.