Unsupervised Clustering on Drug Molecule Representations

Lisa Jin
jin9@llnl.gov

Amanda Minnich
minnich2@llnl.gov

Why perform unsupervised clustering?

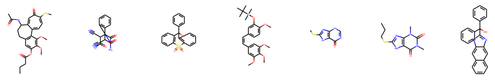
The goal of unsupervised clustering is to summarize the similarity between data points. This is critical in the drug discovery setting, where new drug molecules must be understood in the context of an existing chemical space (i.e., *diversity*). We can leverage this knowledge in downstream tasks, such as generating drug candidates that are both feasible and able to satisfy a desired set of properties.

Without ground truth labels, or even the number of clusters k that may exist, we rely on hierarchical clustering. Given this method, there still remain several choices:

1. What is the best **representation** (e.g., fingerprints, descriptors) of the molecules?
2. Where should one **truncate** a resulting dendrogram for clustering?
3. How can one **evaluate** the quality of a set of cluster labels?

Datasets and molecular representations

To answer the preceding questions, we use subsets of three drug molecule studies: CTRP [2], GDSC [5], NCI60 [1]. Designed to study the potential anticancer effects of small molecules, they contain tumor response values for human cancer cell lines.



Example molecules from the {CTRP, GDSC, NCI60} datasets, which are of sizes {479, 234, 1006}.

Hierarchical, or *agglomerative*, clustering is bottom-up approach that recursively clusters a set of points until a single cluster remains. As clustering involves minimizing intra-cluster distance, it is necessary to define the notion of distance.

Choosing an appropriate distance metric, however, depends on the representation of the data points. In the case of drug molecules, we consider the following.

Name	Definition	Distance	Length
Morgan fingerprints	Bit vectors based on molecular structure	Tanimoto	1024
Dragon7 descriptors	Float vectors of chemical/physical features	Cosine	5269

Codebase access: `git clone -b jin9-dev ssh://git@cz-bitbucket.llnl.gov:7999/biom/molresp.git`

References

[1] Michael C Alley et al. "Feasibility of drug screening with panels of human tumor cell lines using a microculture tetrazolium assay". In: *Cancer Research* 48.3 (1988), pp. 589–601.

[2] Amrita Basu et al. "An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules". In: *Cell* 154.5 (2013), pp. 1151–1161.

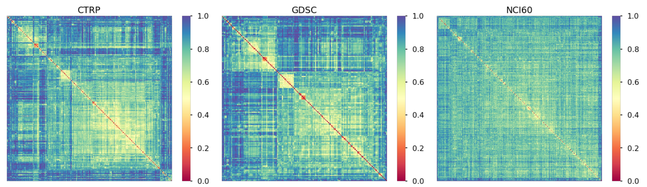
[3] Laurens van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE". In: *Journal of Machine Learning Research* 9.Nov (2008), pp. 2579–2605.

[4] Leland McInnes and John Healy. "UMAP: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426* (2018).

[5] Wanjuan Yang et al. "Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells". In: *Nucleic Acids Research* 41.D1 (2012), pp. D955–D961.

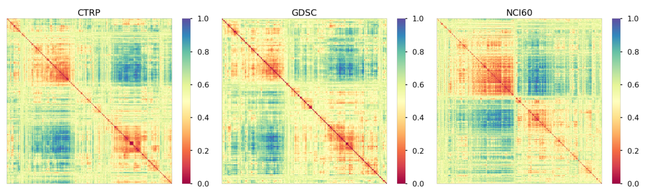
Distance matrix indicators of feature quality

To explore the relative clustering 'potential' for the two molecular representations, we first plot heatmaps of each of their distance matrices. Ideally, the distance matrices will initially contain submatrices which have lower internal distances.



Distance heatmaps (Morgan).

The Dragon7 descriptors (below) reveal a more apparent inherent structure compared to the more uniform distance distribution present in the Morgan fingerprints (above).

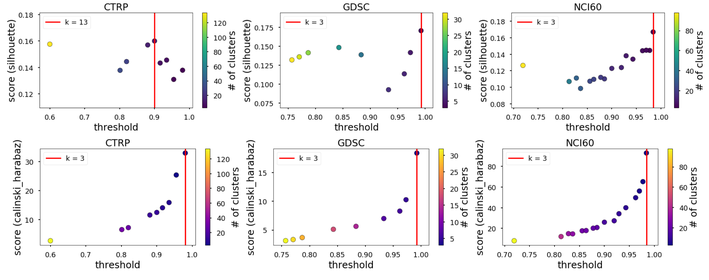


Distance heatmaps (Dragon7).

Evaluation metric for dendrogram truncation

For a given molecule representation, how can we decide *where* to truncate its clustering dendrogram and *how* to evaluate cluster quality? Clearly, the former question depends on how we answer the latter. We consider the following evaluation metrics.

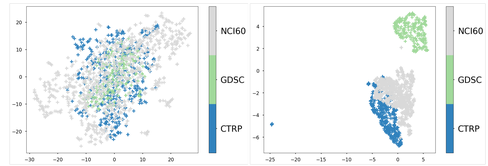
- **Silhouette coefficient**: difference between each cluster's mean distance to nearest neighbor cluster and its intra-cluster distance.
- **Calinski-Harabaz index**: ratio of inter- and intra-cluster dispersion.



Evaluation scores over thresholds for Dragon7. Vertical red lines are optimal truncation thresholds.

Comparing dimensionality reduction methods

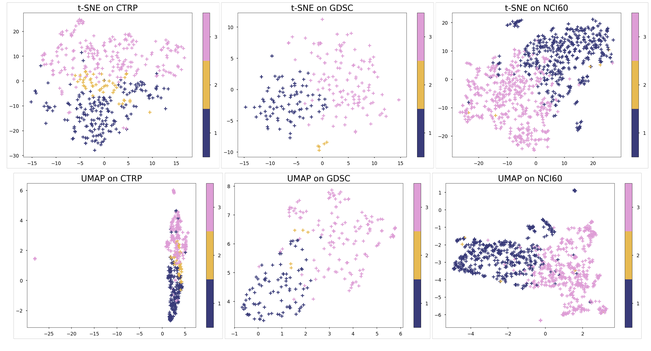
After arriving at a set of cluster labels, we can validate the cluster quality by using dimensionality reduction techniques like t-SNE [3] and UMAP [4] to project the data in 2D. While t-SNE is the current state-of-the-art, it scales with both n and m for a dataset $X \in \mathbb{R}^{n \times m}$. UMAP scales primarily with n and is more efficient.



t-SNE (left) and UMAP (right) projections of pooled datasets.

Visualizing clusters in 2D projections

We compare the techniques by projecting the optimal clusters found for each dataset below. For both techniques, the cluster labels appear separable in the projected space.



t-SNE (top) and UMAP (bottom) projections per dataset.

Summary and future applications

Due to the exploratory nature of hierarchical clustering, we present a pipeline that:

- Compares feature representations.
- Tunes the threshold (quantitative) for dendrogram truncation.
- Visualizes (qualitative) of computed clusters.

These methods help answer questions posed in the introduction. The clusters found could also serve as a reference for understanding how new molecules align with existing data. In addition, the labels could guide interpretation of a supervised model's results.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.