CSpace: Interactive Exploration of Chemical Spaces

Ryan Jenkins

December 18, 2019

Abstract

Analysis of similarity of novel chemicals to chemicals with known physiological effects and known mechanisms of action plays an important role in drug discovery, and studying relationships between known chemicals can yield significant insights into the relationship between chemical structure and interaction. CSpace is an interactive tool for visualizing and exploring "chemical spaces", embeddings of sets of chemicals into low dimensional spaces under some similarity metric.

1 Introduction

While it is widely understood that similarity of chemical structure does not always equate to similarity of effect or interaction, the notion of chemical similarity has nevertheless been successfully studied and employed in domains like drug discovery [1]. The notion of chemical similarity or chemical “distance” naturally gives rise to a notion of chemical “space”, an abstract space wherein points represent possible chemical structures or some facet of chemical structure.

CSpace is a tool for interactively visualizing chemical spaces in three dimensions with the goal of providing insight into the organization of large groups of chemicals and the relationship between chemical structure and effect.

2 CSpace Concepts/Structure

The most basic record in CSpace is a chemical. Chemical structure is stored internally as a Simplified Molecular Input Line Entry System(SMILES) string [2]. Chemicals records also contain various optional metadata including the IUPAC name and PubChem CID. Each chemical may have one or more tags which have no semantic meaning within the system but are used during presentation, e.g. caffeine can be tagged as a purine and as a CNS stimulant.

CSpace currently supports importing chemicals from structure-data files(SDF) which wrap a collection of data in the MDL molfile format. PubChem offers a wealth of chemical information retrievable in the SDF format.

Chemicals are aggregated into “chemical sets”, typically by taking the union of chemicals holding any of a some set of tags, e.g. a chemical set of the nitrogenous bases may be formed from the from purine and pyrimidine tags.

A chemical set may have multiple “facets”. A facet is an embedding of the chemicals in a set to a 3 dimensional euclidean space. The identify of a facet is formed by its chemical set, a similarity metric for use between chemicals, and an embedding strategy. Chemical set facets are the central object which CSpace visualizes.

3 Algorithms Used

Creating chemical set facets is a central task for CSpace. The supported similarity metrics and embedding algorithms on a chemical set are briefly described below.

3.1 Similarity Metrics

Currently two similarity metrics are supported: RDKit and Gobbi-Poppinger substructure fingerprints.

The RDKit fingerprinting algorithm [3] is a modified formulation of the Daylight chemical fingerprinting algorithm [4]. Like the Daylight algorithm, it operates by enumerating and counting paths through the atom/bond graph of a molecule, by default considering all paths between 1 and 7 bonds in length. The identity of a bond path (for the purposes of counting) consists of the atomic weight of the atoms participating in each bond and the aromaticity of the bond. These bond path histograms are then hashed to produce a fixed length bit vector that summarizes the structure of the chemical.

The Gobbi-Poppinger fingerprinting algorithm [5] operates on a similar premise to the RDKit algorithm, however rather than enumerating and counting all possible bond paths within some range of sizes, it is instead equipped with a fixed library of substructure patterns which are identified and counted in molecules. This produces a smaller histogram of substructures, however the structures are selected to be chemically meaningful pharmacophores. These histograms are hashed in a similar manner to the RDKit approach and also yield a fixed length bit vector.

After chemicals fingerprints are computed, a distance matrix D is computed:

$$D_{i,j} = 1 - T(i, j) \tag{1}$$

Where $T(i, j)$ is the Tanimoto coefficient between molecules i and j in the chemical set.

3.2 Embedding Approaches

CSpace currently supports a total of 4 embedding techniques. They are: metric and non-metric variants of SMACOF, ISOMAP, as well as t-distributed Stochastic Neighbor Embedding(t-SNE).

The goal of all five methods is to create an n-dimensional (3, in our case) set of points from a dissimilarity matrix where inter-point distances are in some sense preserved. Metric SMACOF achieves this by minimizing the difference in inter-point distances (strain) in the embedding with respect to the original distance matrix. The non-metric variant of this technique attempts to maintain the relative ordering for pair-wise distances rather than globally minimizing stress (e.g. if $D_{i,j} < D_{i,k}$ then for an embedded point E_i it holds that $||E_i - E_j|| < ||E_i - E_k||$ for all i, j, k).

t-SNE is a non-linear embedding technique that balances minimization of strain in near-by clusters of points and inter-cluster distances [6].

ISOMAP constructs a K-NN graph of the points in a chemical set with edge weights equal to inter-point distances. It then reconstructs the distance matrix such that the distance between each pair of points is equal to the minimal cost path between the points in the constructed graph and then performs

MDS/SMACOF on the new distance matrix to arrive at an embedding [7]. ISOMAP leverages the assumption that points lie on a low dimensionality manifold and that minimal paths through the K-NN graph approximate geodesics along this manifold.

4 Examples

As there is no singularly correct formulation of chemical similarity, accordingly there is no singular objective measure of the correctness of a CSpace embedding. One useful property of an embedding would be that similarity of effect in humans between two chemicals corresponds to proximity in an embedded space, although there are many more potentially useful sorts of similarities that proximity in a chemical space might signal (e.g. proximate chemicals have a similar method of synthesis, or similar levels of toxicity regardless of other effects). In this section we will look at examples of embeddings generated by CSpace and make the case for their general plausibility. To this end we will look at two chemical sets collected from PubChem.

4.1 Classes of Chemicals Containing Frequent Recreational Drugs

This dataset of 1030 compounds consists of 469 benzodiazepines, 251 amphetamines, 136 opiates, and 184 ergolines. Note that these four classes were selected because they contain chemicals frequently used recreationally, however membership in the class is defined chemically. For example 3-[5-(3-Nitro-phenyl)-furan-2-yl]-2-(piperidine-1-carbonyl)-acrylonitrile is categorized as an opiate under the Anatomical Therapeutic Chemical Classification System but is not commonly synthesized or used recreationally.

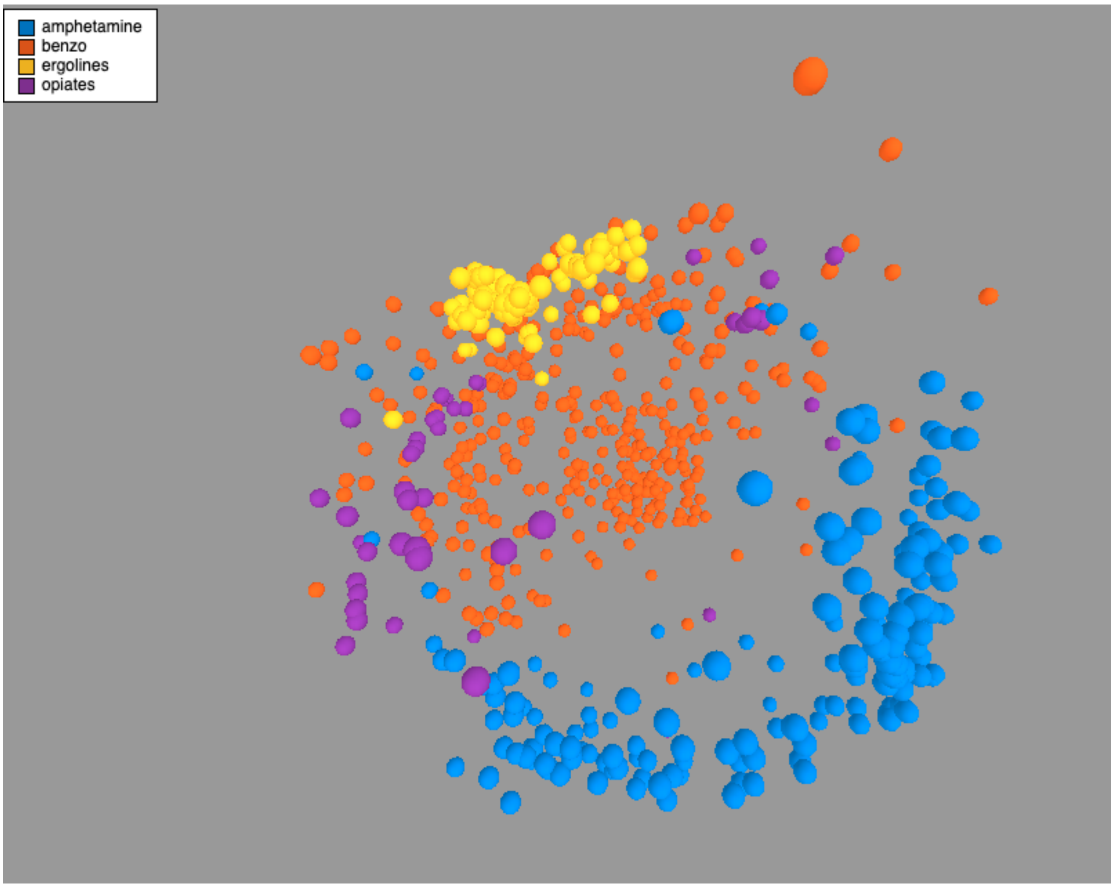


Figure 1: Recreational classes under RDK fingerprinting and metric SMACOF

Figure 1 shows the SMACOF embedding of Tanimoto/RDKit fingerprints for this dataset. Here we see good separation between the amphetamines, benzodiazepines, and ergolines. The opiates group is a little more spread out but generally closer to the benzodiazepine group. Interestingly the ergoline group, which contains stimulants, is also positioned closer to both benzodiazepines and opiates than amphetamines. A likely explanation for this is that these three classes of compounds are all more complex and generally heavier than amphetamines.

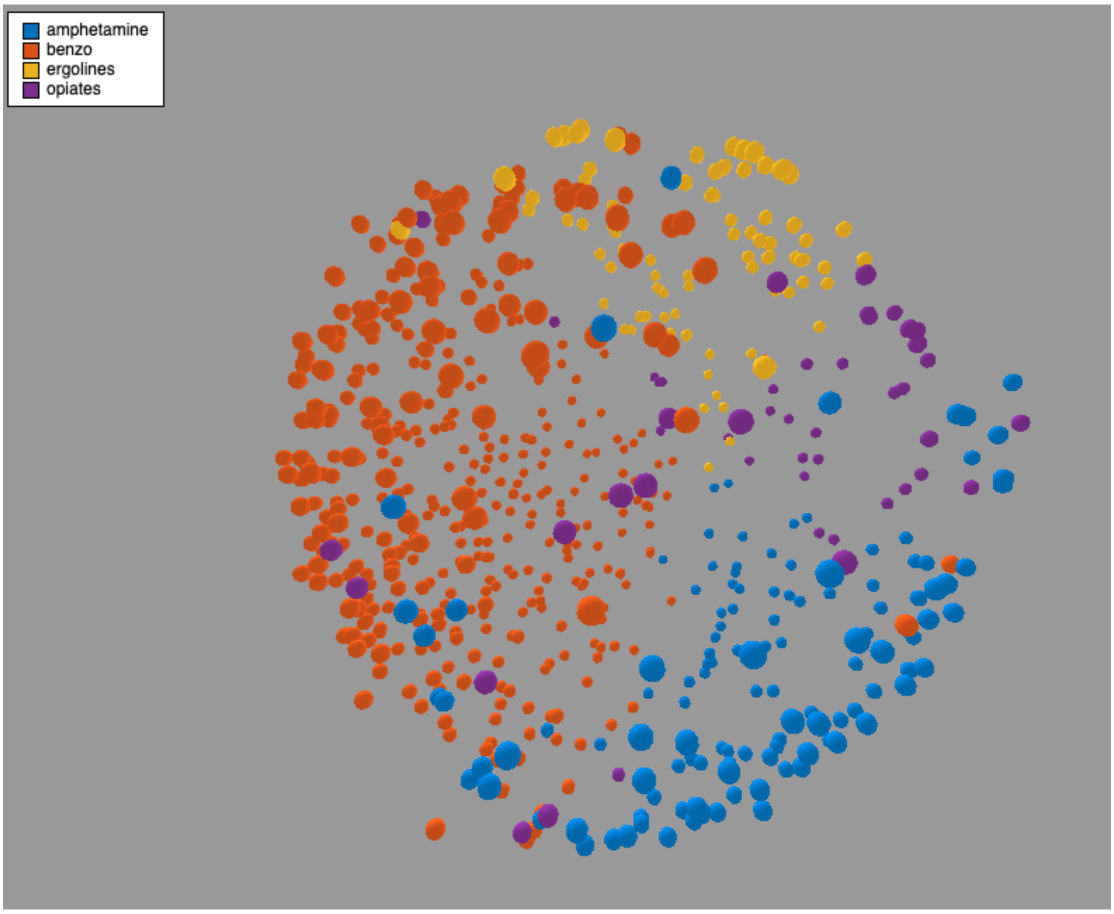
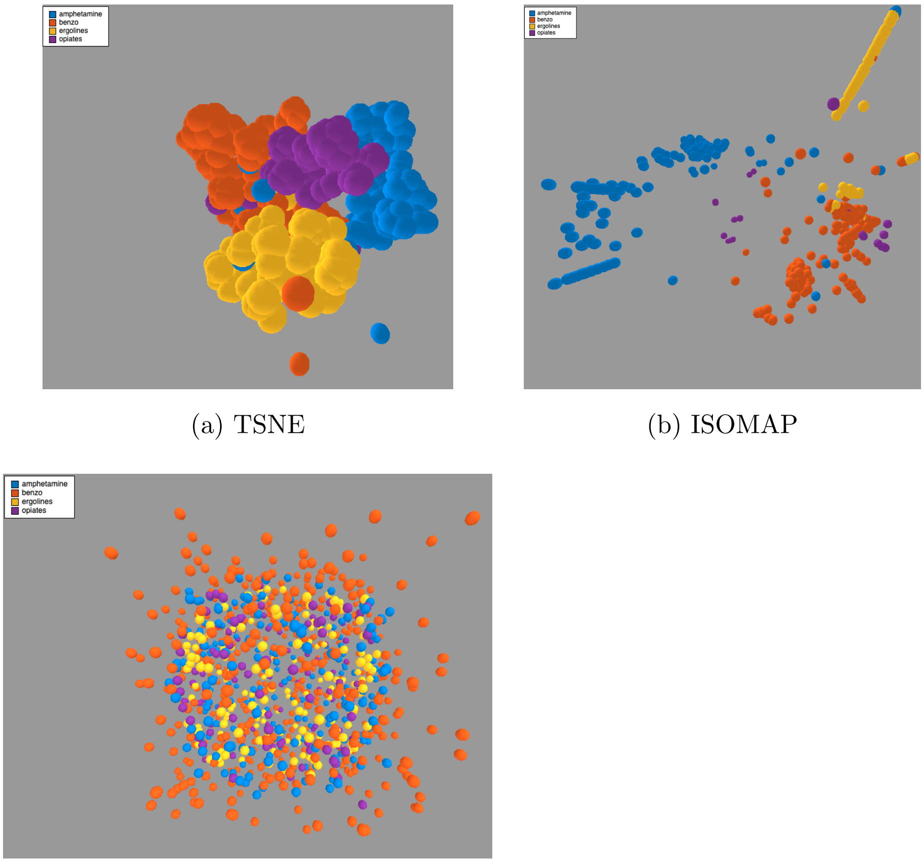


Figure 2: Recreational classes under Gobbi-Poppinger and metric SMACOF

The Gobbi-Poppinger/Tanimoto similarity measure produces similar results to RDK fingerprinting although it differentiates members of each class over more space while maintaining fairly clean separation, as seen in figure 2.



(a) TSNE

(b) ISOMAP

(c) Non-metric SMACOF

Figure 3: Four embedding strategies for the recreational dataset

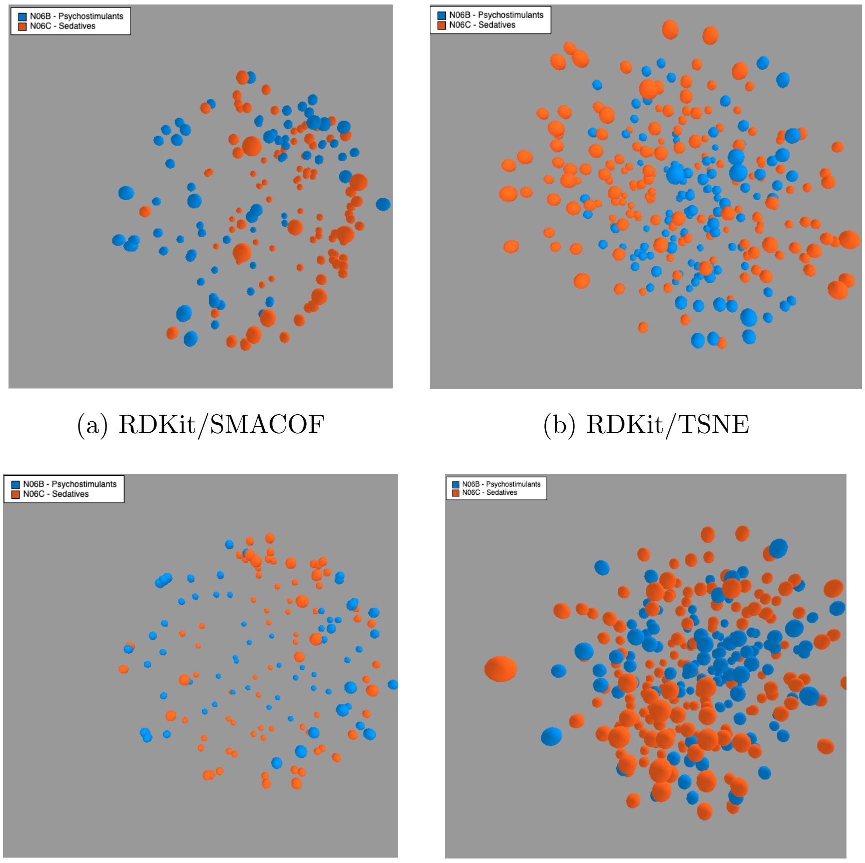
Figure 3 shows the three remaining embedding techniques all with the RDKit fingerprinting similarity metric. TSNE produces tighter cluster of chemicals with a few more wide outliers than SMACOF. Notably it groups the opiates class into a much tighter cluster than SMACOF. I speculate that this is because opiates are the smallest class in the dataset, global embeddings like SMACOF can minimize global stress while still distorting the distance relationships between members of a small class, whereas TSNE prioritizes maintaining local distance relationships.

ISOMAP produces relatively good separation of classes, although again, less so for opiates. It tends to produce "spurs", chemicals extending linearly in a given direction under embedding, most noticeable with ergolines.

Nnotice that non-metric variant of SMACOF shows almost no class separation, the reason for this is not understood at this point.

4.2 Clinical Sedatives and Psychostimulants

This set of 289 compounds consists of 120 chemicals identified as psychostimulants and 169 sedatives, corresponding to the N06B and N05C categories of the Anatomical Therapeutic Chemical Classification System. Unlike the previous dataset, these classes are only defined by their therapeutic uses and are therefore more structurally diverse.



(a) RDKit/SMACOF

(b) RDKit/TSNE

(c) Gobbi/SMACOF

(d) Gobbi/TSNE

Figure 4: Three embeddings of the sedatives and psychostimulants chemical set

As we can see from figure 4, this dataset yields only weak separation between classes for all similarity metrics and embeddings. TSNE remains prone to producing outlying points. This is reasonable given the character of the data as being categorized by therapeutic use rather than chemical structure.

5 Conclusion

The CSpace application allows for visualization of chemical spaces. The algorithms currently employed seem to model structural similarity well and can produce interesting visualizations however this does not translate to *in vivo* effect particularly well.

References

[1] N. Nikolova and J. Jaworska, "Approaches to measure chemical similarity– a review," *QSAR & Combinatorial Science*, vol. 22, pp. 1006–1026, Dec. 2003.

[2] D. Weininger, "SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of Chemical Information and Modeling*, vol. 28, pp. 31–36, Feb. 1988.

[3] G. Landrum, "The rdkit book - rdk fingerprints," 2007.

[4] "Daylight theory manual," May 2019.

[5] A. Gobbi and D. Poppinger, "Genetic optimization of combinatorial libraries," *Biotechnology and Bioengineering*, vol. 61, no. 1, pp. 47–54, 1998.

[6] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[7] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.