Low Dimensional Chemical Space via Manifold Learning

Nathaniel Charest*, Charles Lowe*, Christian Ramsland*, Todd Martin* and Antony Williams*

*Center for Computational Toxicology & Exposure, Office of Research & Development, U.S. Environmental Protection Agency, Research Triangle Park, NC, USA

The views expressed here do not necessarily reflect US Environmental Protection Agency policy or position.

ORCID: 0000-0003-4252-0365
Nathaniel Charest I Charest.Nathaniel@epa.gov I 919-541-2532

Overview

Chemical modeling requires a quantitative definition of chemical space. A quantitative chemical space encodes patterns within chemical data, and structure-activity modeling (SAR) works to relate those patterns to chemical activities and properties of interest. A common assumption of the SAR community is that, while lower dimensional chemical spaces are desirable, it often takes many independent quantities to appropriate describe the necessary chemical information to predictively capture activities of interest.

A low-dimensional chemical space is desirable because it is easier to visualize, simpler to interpret, and can be used to generate insights that enable greater scientific understanding. Because of the elegance of low-dimensional chemical spaces, methods for reducing many independent chemical descriptors to few are sought after.

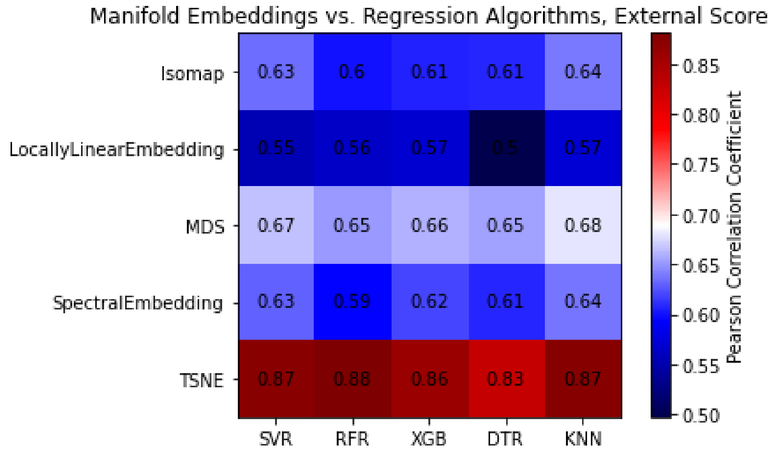
Machine learning offers the concept of manifolds to facilitate these efforts. It often takes many independent variable to capture qSAR, however applying so-called “manifold learning” algorithms to large amounts of data gives us insight into whether we can automatically detect lower-dimensional projections of chemical data in such a way that preserves the information necessary to accurately predict chemical properties through quantitative structure-activity methodology.

Does The Projection Preserve Information?

We tested the ability of five common manifold learning methods to preserve enough structural detail for regression algorithms to learn and predict the water solubility of compounds. The manifold learner reduces a 16-dimensional chemical space to 2 dimensions. Regression algorithms are then trained on and compared between both projections of chemical space.

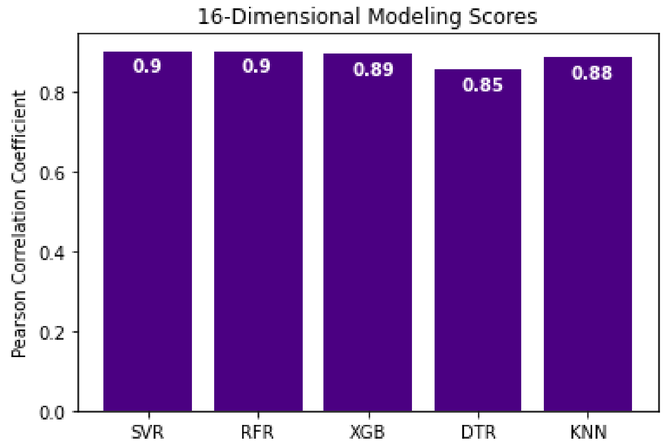
Manifold Learning Algorithms	Regression Algorithm
t-Distributed Stochastic Neighbor Embedding[1]	SVR – Support Vector Regression
Multidimensional Scaling[2]	RFR – Random Forest Regression
Isomaps[3]	XGB – Extreme Gradient Boosted Trees
Spectral Embedding[4]	DTR – Decision Tree Regression
Locally Linear Embeddings[5]	KNN – k-Nearest Neighbor Regression

Results

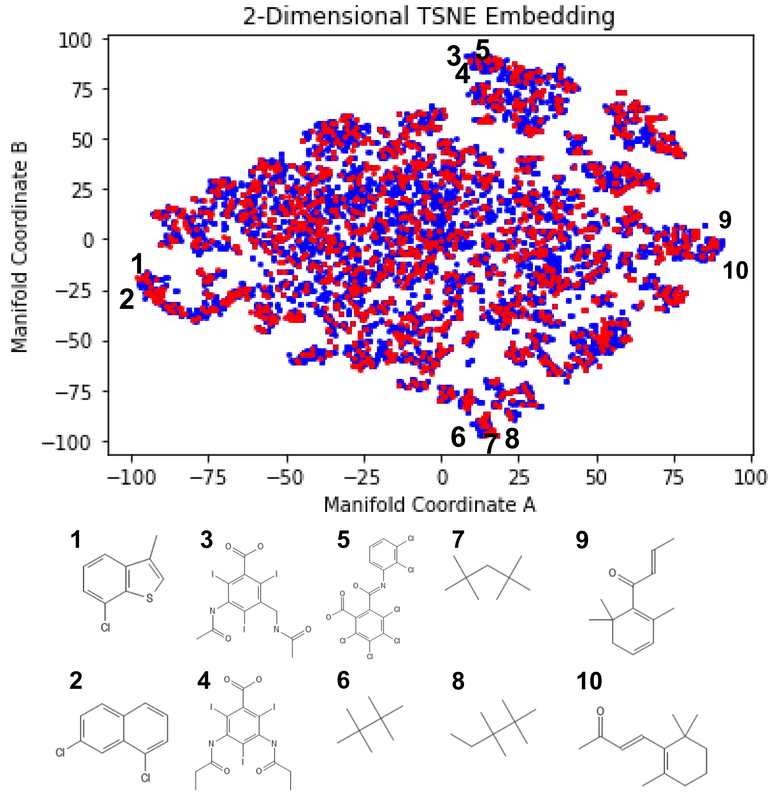


t-Distributed Stochastic Neighbor Embedding (TSNE) produces the best embedding of chemical space according to these results. The 2-dimensional space from t-SNE projection produces modeling scores comparable with the original 16-dimensional space.

This likely arises from the fact many shallow learners take advantage of local similarities between compounds that TSNE explicitly utilizes when projecting the manifolds. This preserves patterns that are useful for predicting the water solubility of compounds.



Visualizing Chemical Space



References

1) van der Maaten, L.J.P.; Hinton, G.E. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9:2579-2605, 2008.
2) Kruskal, J.; Nonmetric multidimensional scaling: a numerical method. Psychometrika 29 ,1964
3) Tenenbaum, J.B.; De Silva, V.; & Langford, J.C. A global geometric framework for nonlinear dimensionality reduction. Science 290 (5500)
4) On Spectral Clustering: Analysis and an algorithm, 2001 Andrew Y. Ng, Michael I. Jordan, Yair Weiss
5) Roweis, S. & Saul, L. Nonlinear dimensionality reduction by locally linear embedding. Science 290:2323 (2000).