# Dimension Reduction of Molecular Fingerprints Using Nonnegative Matrix Factorization

Takaaki Ohnishi

Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan, ohnishi.takaaki@i.u-tokyo.ac.jp

In order to quantify the similarity between molecules, molecular fingerprints[1] are widly used in cheminformatics for virtual screening, compound classification, prediction of molecular properties and retrosynthesis planning[2]. Molecules are characterized by fingerprints(vectors), which are sparse strings representations of molecular structures in which each position indicates the occurrence count of specific structural features(descriptors) within the molecule of interest. Here, each molecule is represented as a single point in a hyper dimensional space determined by its descriptors. Tanimoto coefficient is most commonly used as similarity measure between two fingerprints represented as two vectors. In many cases in analysis, the similarity is calculated between a reference fingerprint and many query fingerprints. Because it requires a huge number of descriptors to characterize the molecular structures precisely, fingerprint space are usually high-dimensional and hence it takes a lot of time to calculate the similality when applied to very large databases. In order to remove the curse of dimensionality, the projection of the space that maps high-dimensional fingerprint space to a lower dimensional chemical space is required. Herein an application of Nonnegative Matrix Factorization(NMF) is introduced that is a powerful technique for better data representation or dimensionality reduction[3, 4]. NMF is strictly focused on nonnegative data, decomposes the given data into factors of nonnegative values, and thus provides more interpretable results. It is widely used in machine learning, signal processing, and text mining. Despite this, there has been little analysis of molecular fingerprints using NMF. Proposed method is tested using large database. Results show the improvements in comparison with principal components analysis and multidimesional scaling.

## Acknowledgements

This work was supported by JSPS Grant-in-Aid for Scientific Research on Innovative Areas: 17H06468.

## References

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

[2] Coley, C. W., Rogers, L., Green, W. H., & Jensen, K. F. (2017). Computer-assisted retrosynthesis bases on molecular similarity. *ACS central science*, 3(12), 1237-1245.

[3] Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755), 788.

[4] Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. *In Advances in neural information processing systems*, 556-562.