Supporting Material

Filter feature selection for unsupervised clustering of designer drugs using DFT simulated IR spectra data

Kedan He*
Department of Physical Sciences, School of Arts and Sciences, Eastern Connecticut State University, Willimantic, Connecticut 06226, United States.

Table of Contents

Appendix 0. GitHub repository..................................................................................................................S2
Appendix A. DFT IR Spectrum Simulation using Lorentzian Broadening.................................................S2
Appendix B. Total IR spectra using Boltzmann factor distribution..........................................................S3
Appendix C. hierarchical clustering analysis and comparison.................................................................S3
Appendix 0. GitHub repository

For additional files and python codes used in this study, please go to the following GitHub repository: https://github.com/nina23bom/IR_NPS_Feature-selection_Clustering

Appendix A. DFT IR Spectrum Simulation using Lorentzian Broadening

The spectrum simulation python code reads molecule’s DFT calculated vibrational spectrum (normal mode frequency $X$, IR or Raman intensity $Y$) and generate the simulated experimental spectrum.

The peaks of the simulated spectrum are widened using the probability density function of the Lorentzian distribution $^1$:

$$ f(x) = \frac{1}{\pi \gamma [1 + \left(\frac{x - x_0}{\gamma}\right)^2]} = \frac{1}{\pi \gamma} \left[\frac{\gamma^2}{(x - x_0)^2 + \gamma^2}\right], \quad (1) $$

where $x_0$ is the frequency $X$ from the input file, and $\gamma$ is the scale parameter which specifies the half-width at half-maximum (HWHM).

In a distribution, full width at half maximum $(FWHM)^2$ is the difference between the two values of the independent variable at which the dependent variable is equal to half of its maximum value. In other words, it is the width of a spectrum curve measured between those points on the $y$-axis which are half the maximum amplitude. Half width at half maximum $(HWHM)$ is half of the FWHM if the function is symmetric, hence:

$$ \gamma = \frac{FWHM}{2}, \quad (2) $$

Single ideal peak $(X, Y)$ is widened by

$$ y = Y f(x), \quad (3) $$

However, when a customized FWHM is used, $f(x)$ changes accordingly and the simulated peak height $y$ is not equal to the given peak height $Y$ anymore. In order to keep the original peak height unchanged, when $x = x_0 = X$, it should force the $f(x) = 1$, then we have:

$$ f(x) = \frac{1}{\pi \gamma [1 + \left(\frac{x - x_0}{\gamma}\right)^2]} = \frac{1}{\pi \gamma} = \frac{y}{\pi FWHM} = 1, \quad (4) $$

$$ FWHM = \frac{2}{\pi \gamma}, \quad (5) $$

and we get $FWHM_0 = 2/\pi$. If the $FWHM$ is changed, $f(x)$ should be scaled by $FWHM/FWHM_0$

The equation for the simulated spectrum is the sum of all the simulated peaks

$$ y_{sim} = \sum y \quad (6) $$
NumPy is used to create the array for plotting the figure, Matplotlib to draw the figure.
Appendix B. Total IR spectra using Boltzmann factor distribution

The relative populations of the low-energy conformers are computed through the probabilities defined in Equation (7):

\[ P_i(T) = \frac{e^{-\frac{\Delta G_i}{k_B T}}}{\sum_{j=1}^{M} e^{-\frac{\Delta G_j}{k_B T}}} \]

where \( P_i \) is the probability of state \( i \), \( k_B \) is the Boltzmann constant, \( T \) is the temperature in K, and \( \Delta G_i \) is the Gibbs free energy of the \( i \)th conformer.

The total IR spectra is constructed as a weighted average according to the relative population probabilities computed in Equation (7). The temperature is assumed to be 300 K. The Gibbs free energy is obtained from frequency calculation using Gaussian.

Table S1. Example of weigh calculated for the lowest-energy conformers of amphetamine:

| Conformer | \( \Delta G \) (Hartree) | \( \Delta G_i - \Delta G_0 \) (kJ/mol) | Exp(-deltaG/RT) | \( \chi \) Mole | \% Pop |
|-----------|----------------|-------------------------------|----------------|--------------|--------|
| G0        | -405.4899     | 0.0000                        | 1.0000         | 0.4657       | 46.5658 |
| G1        | -405.4894     | 1.3049                        | 0.5927         | 0.2760       | 27.5977 |
| G2        | -405.4887     | 3.1480                        | 0.2831         | 0.1318       | 13.1816 |
| G3        | -405.4885     | 3.8910                        | 0.2102         | 0.0979       | 9.7859  |
| G4        | -405.4869     | 7.9920                        | 0.0406         | 0.0189       | 1.8905  |
| G5        | -405.4863     | 9.6303                        | 0.0210         | 0.0098       | 0.9802  |
| **Sum =** |                |                               | 2.1475         | 1.0000       | 100.00  |

Appendix C. Hierarchical clustering analysis and comparison

Two quantitative measures of chemical structural similarity were calculated for class label validation and comparison. The substructural key-based 2D molecular fingerprint MACCS encodes the absence (0) and presence (1) of a specific chemical group in each bit position. The binary Tanimoto coefficient \( T_C \) is used for the binary molecular fingerprint and calculated as follows:

\[ T_C = \frac{c}{a + b - c} \]

where “c” is the number of bits common to the two fingerprints and “a” and “b” denote the number of bits set in each of the two fingerprints. The maximum common substructure (MCS) similarity \( T_{MCS} \) was also calculated only on the matched heavy atoms as described:

\[ T_{MCS} = \frac{N_C}{N_A + N_B - N_C} \]
where $N_C$ is the number of matched heavy atoms in MCS of molecule A and B, $N_A$ and $N_B$ are the number of heavy atoms in molecule A and B, respectively. Both were calculated using the MACCSkeys and rdFMCS modules implemented in RDKit software.\(^6\)

**REFERENCES**

1. Johnson, N. L.; Kotz, S.; Balakrishnan, N., Continuous Univariate Distributions. New York: Wiley: 1994; Vol. Volume 1.
2. Weik, M. H., full-width at half-maximum. In *Computer Science and Communications Dictionary*, Springer US: Boston, MA, 2001; pp 661-661.
3. McQuarrie, D. A. Statistical Mechanics. University Science Books: 2000; Vol. 55.
4. Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Model.* 2002, 42 (6), 1273-1280.
5. Zhang, B.; Vogt, M.; Maggiora, G. M.; Bajorath, J. Design of chemical space networks using a Tanimoto similarity variant based upon maximum common substructures. *J. Comput.-Aided Mol. Des.* 2015, 29 (10), 937-950.
6. *RDKit: Open-source cheminformatics; http://www.rdkit.org*. 