



HAL
open science

MetGem Software for the Generation of Molecular Networks Based on the t-SNE Algorithm

Nicolas Elie, Florent Olivon, Gwendal Grelier, Fanny Roussi, Marc Litaudon,
David Touboul

► **To cite this version:**

Nicolas Elie, Florent Olivon, Gwendal Grelier, Fanny Roussi, Marc Litaudon, et al.. MetGem Software for the Generation of Molecular Networks Based on the t-SNE Algorithm. 12èmes journées du RFMF (Réseau Français de Métabolomique et Fluxomique) 2019, May 2019, Clermont-Ferrand, France. 2018. hal-04427165

HAL Id: hal-04427165

<https://hal.science/hal-04427165v1>

Submitted on 30 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

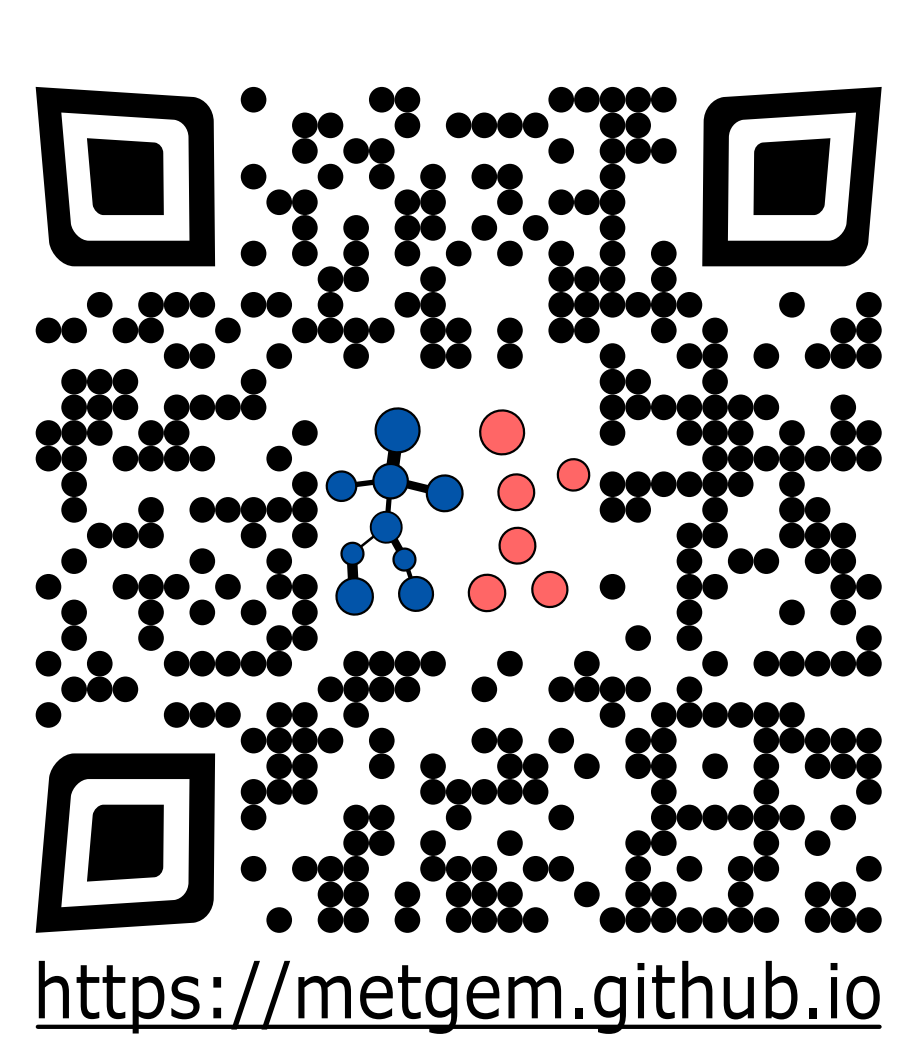
MetGem Software for the Generation of Molecular Networks Based on the t-SNE Algorithm



Nicolas Elie, Florent Olivon, Gwendal Grelier, Fanny Roussi, Marc Litaudon and David Touboul

Institut de Chimie des Substances Naturelles, CNRS UPR 2301, Université Paris-Sud, Université Paris-Saclay, Gif-sur-Yvette, France.

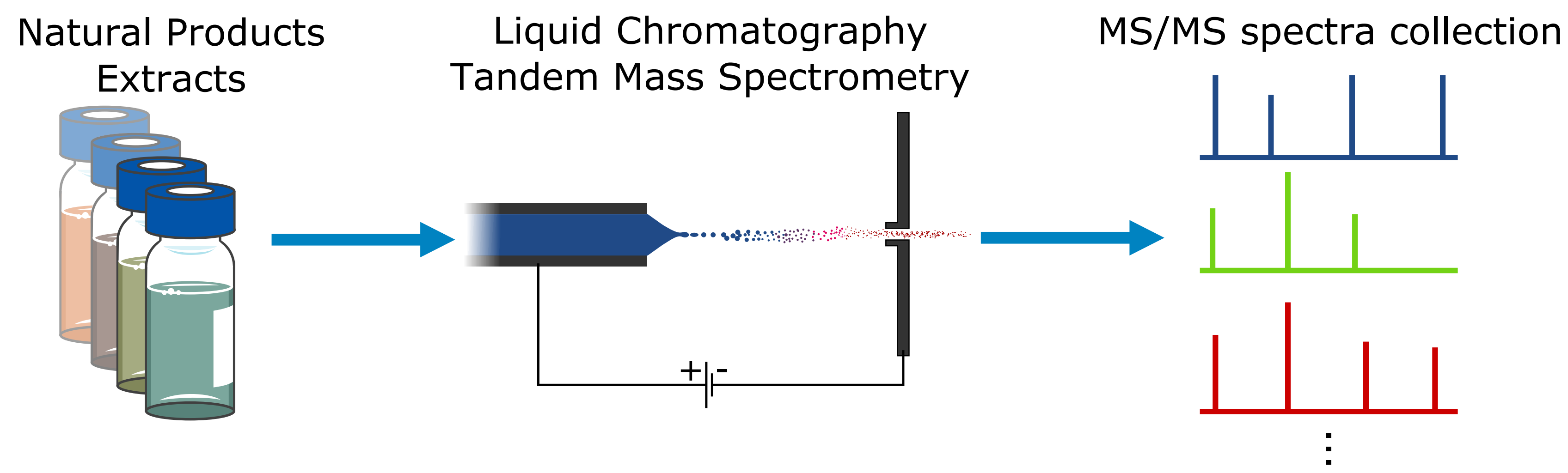
F. Olivon, N. Elie *et al.*, *Anal. Chem.*, **2018**, 90 (23), pp 13900–13908, DOI: 10.1021/acs.analchem.8b03099



<https://metgem.github.io>

Download it from here

Molecular Networks



Spectra are compared by pairs

Nodes repulse each other like charged particles

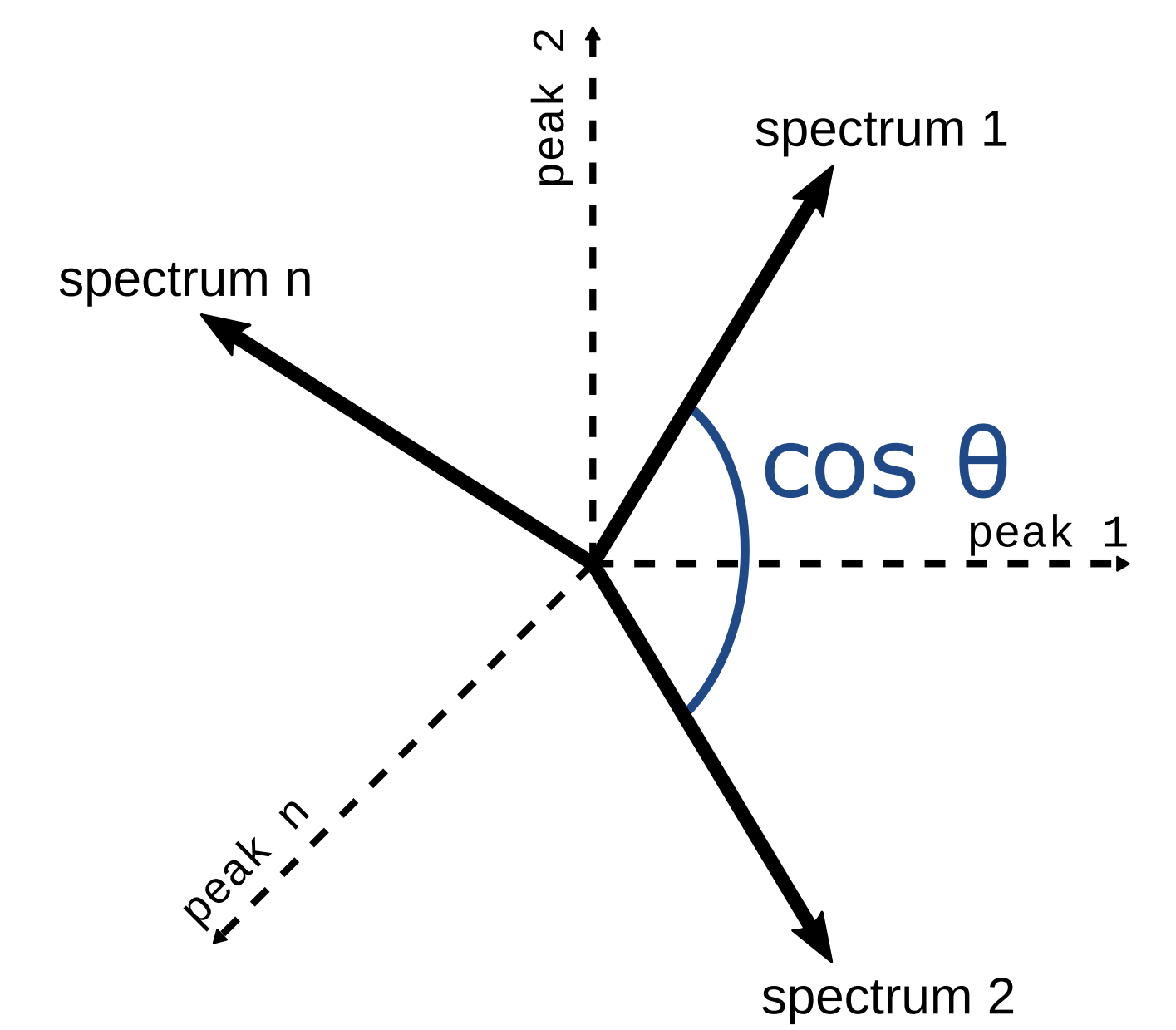
Edges act as spring attracting nodes with stiffness proportional to cosine scores

Because of cosine threshold and topK algorithm, some links are cut, leading to information lost



M. Wang *et al.*, *Nature Biotechnol.*, **2016**, 34, 828-837

Cosine scores



$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

unrelated score
 $\cos \theta \approx 0$

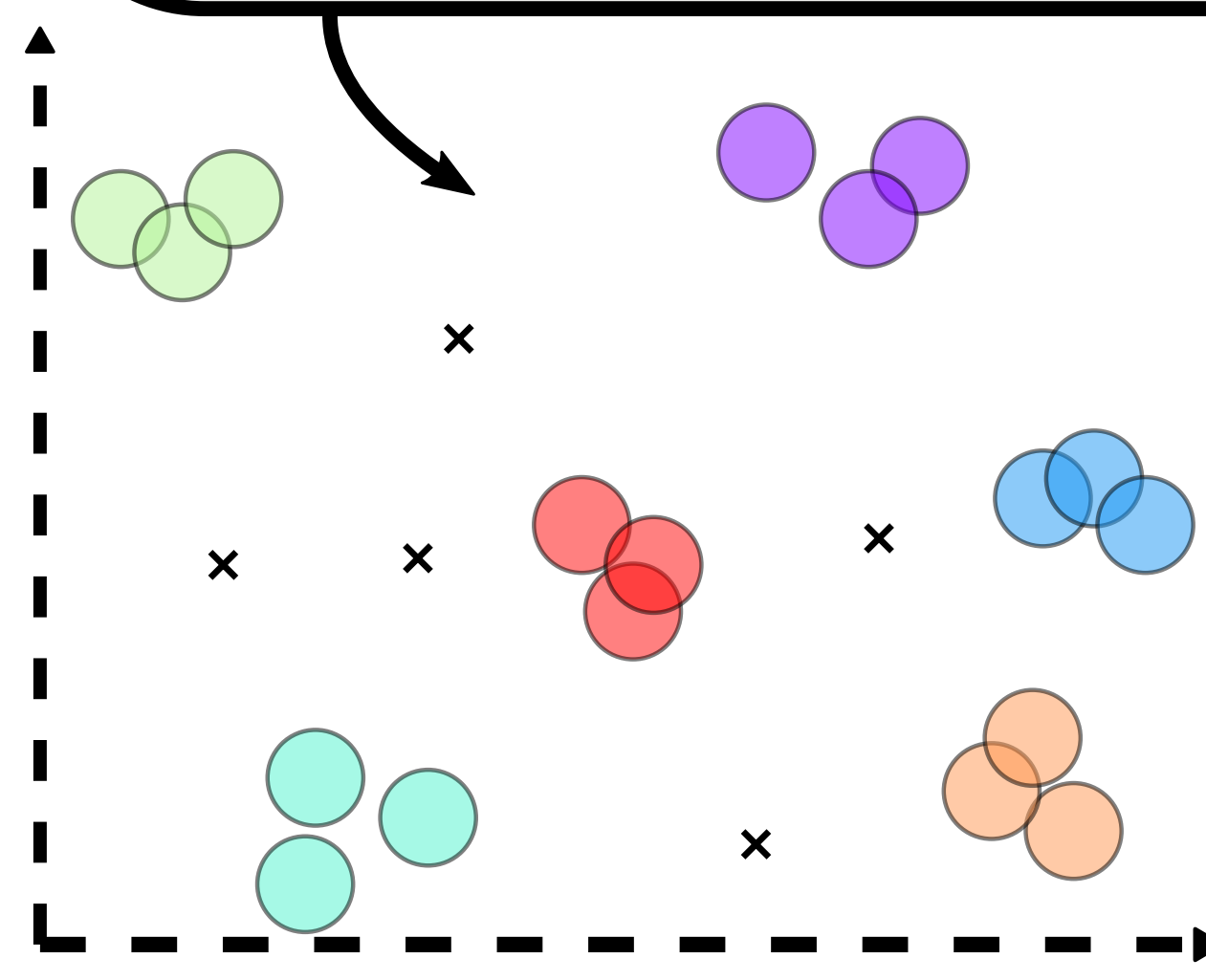
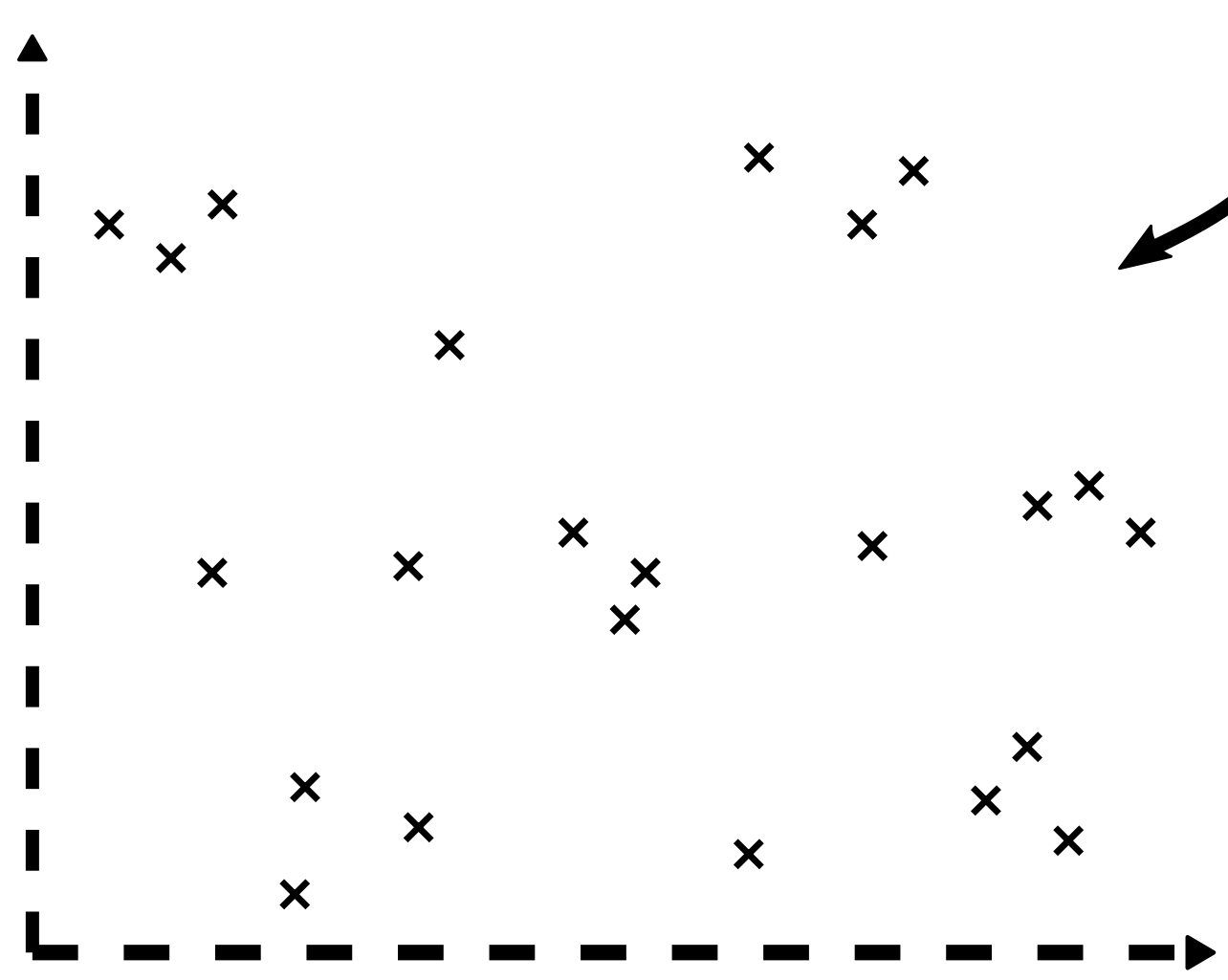
similar score
 $\cos \theta \approx 1$

t-SNE

Non-linear projection of a n -dimensional space into a 2D space with priority to local distances over global distances.

How to represent these data in a space of lower dimension?

We are lucky, our data is 2D so we can plot it but real world data is usually way larger...



PCA

t-SNE

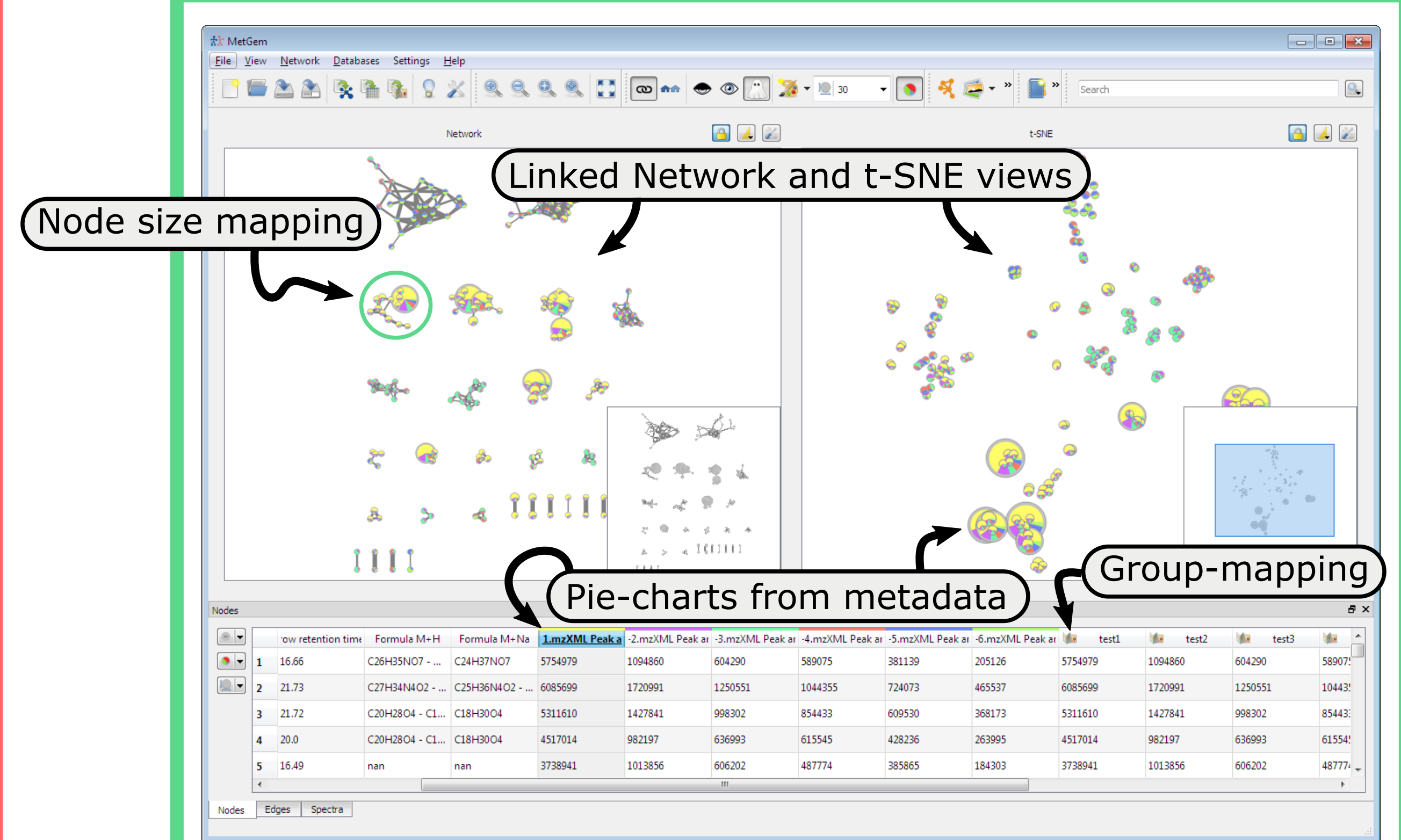
The common solution give bad results

t-SNE is doing better. It can even get improved if we exclude isolated points

Intercluster distances are not meaningful!

L. Van der Maaten *et al.*, *J. of M. Learning Research*, **2008**, 9, 2579-2605
StatQuest: t-SNE, Clearly Explained: youtu.be/NEaUSP4YerM

Key features



- ✓ Quickly and easily tune parameters
- ✓ Fast generation of molecular networks
- ✓ No internet connection required
- ✓ Read data from: MGF files exported by **MZmine 2**
MSP files (**NIST** file format)
- ✓ Use databases from different sources
 - MS-Dial
 - GNPS
 - ISDB
- ✓ Plugins can be easily created to add new databases
- ✓ One-click import of user databases
- ✓ Export both views to cytoscape
- ✓ Cross-platform
 - Linux
 - macOS

Free software: contributions, suggestions are welcome!



nicolas.elie@cnsr.fr, david.touboul@cnsr.fr

<https://github.com/metgem/metgem>

Source code repository