Réseau moléculaire massif pour l'annotation en lipidomique

Nicolas Elie, Romain Magny, Nicolas Auzeil, Olivier Laprévote, David Touboul

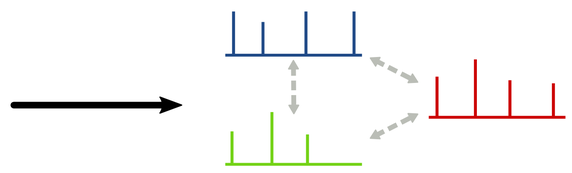
Réseau moléculaire massif pour l'annotation en lipidomique

Nicolas Elie[1], Romain Magny[2,3], Nicolas Auzeil[3], Olivier Laprévote[3,4] et David Touboul[1]

1. Institut de Chimie des Substances Naturelles, CNRS UPR 2301, Université Paris-Sud, Université Paris-Saclay, Gif-sur-Yvette, France
2. Sorbonne Université UM80, INSERM UMR 968, CNRS UMR 7210, Institut de la Vision, IHU ForeSight, 75006 Paris, France
3. UMR CNRS 8038 CiTCoM, Chimie Toxicologie Analytique et Cellulaire, Université de Paris, Faculté de Pharmacie, 75006 Paris, France
4. Hôpital Européen Georges Pompidou, AP-HP, Service de Biochimie, 75006 Paris, France

Lipidblast

~5× 10^5 *in silico* lipids spectra



Comparison of spectra by pairs

t-SNE projection

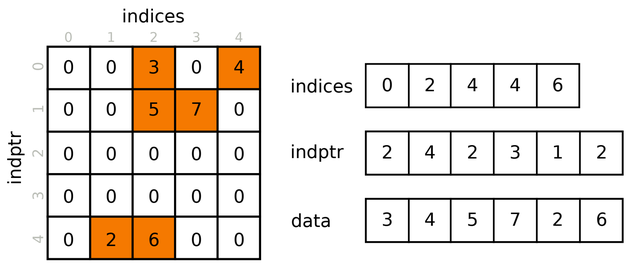
Polarity	Positive	Negative
Perplexity	100	200
Learning Rate	200	30000
Early exaggeration	12	12
# iterations	1000	1000

t-SNE parameters

Operating System: KUbuntu 20.04
Processors: 32 × Intel® Xeon® CPU E5-2650 0 @ 2.00GHz
Memory: **128 GiB** + 184 GiB of swap

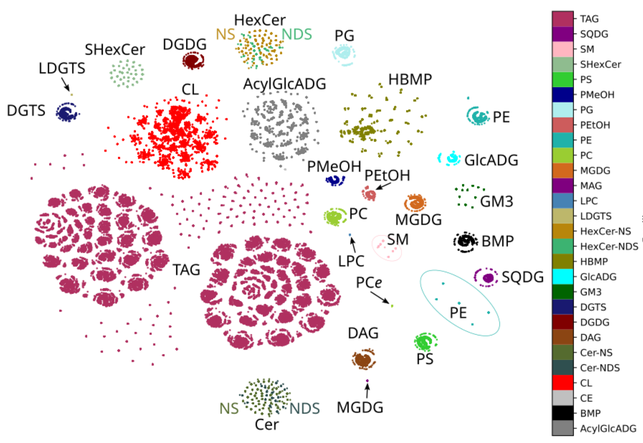
MetGem

To handle such a large dataset, libmetgem library has been updated to store similiarity scores in sparse matrices (Compressed Sparse Row format)

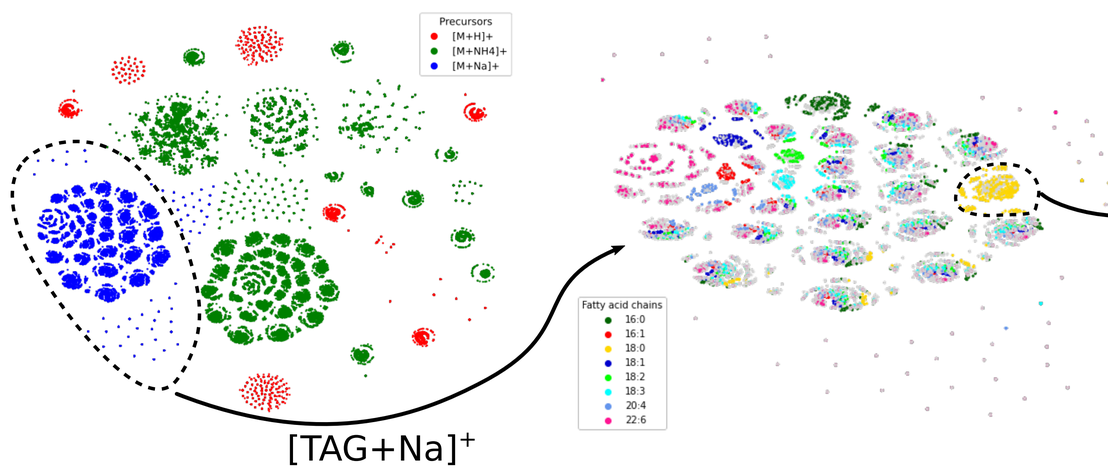


Positive

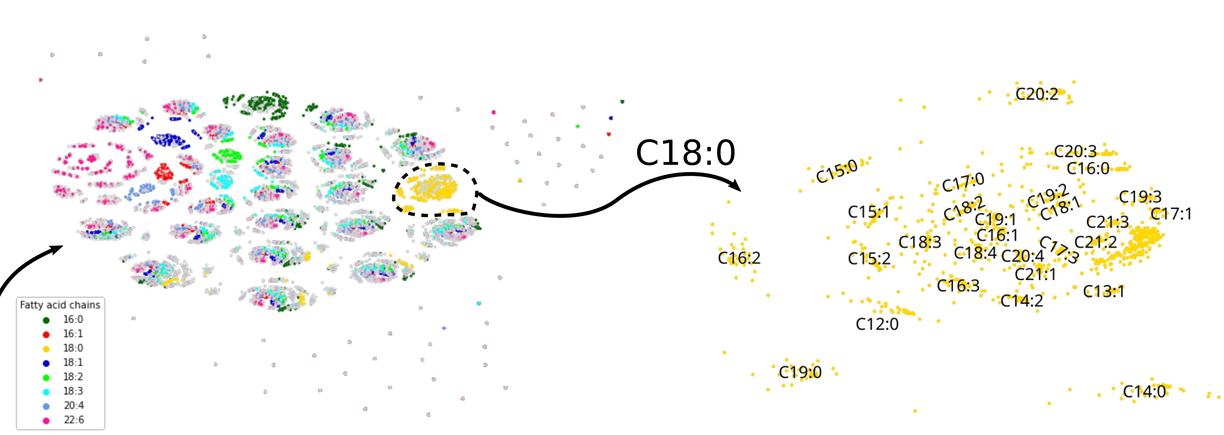
Families



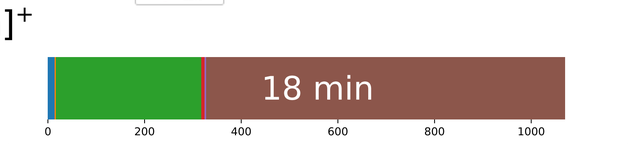
Precursors



Fatty acid chains

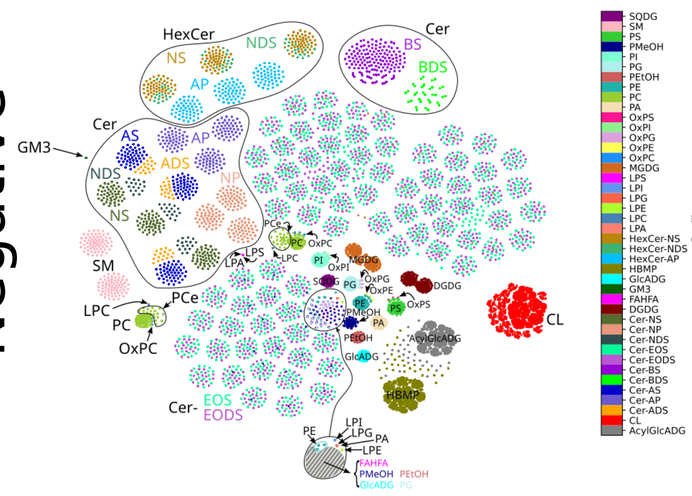


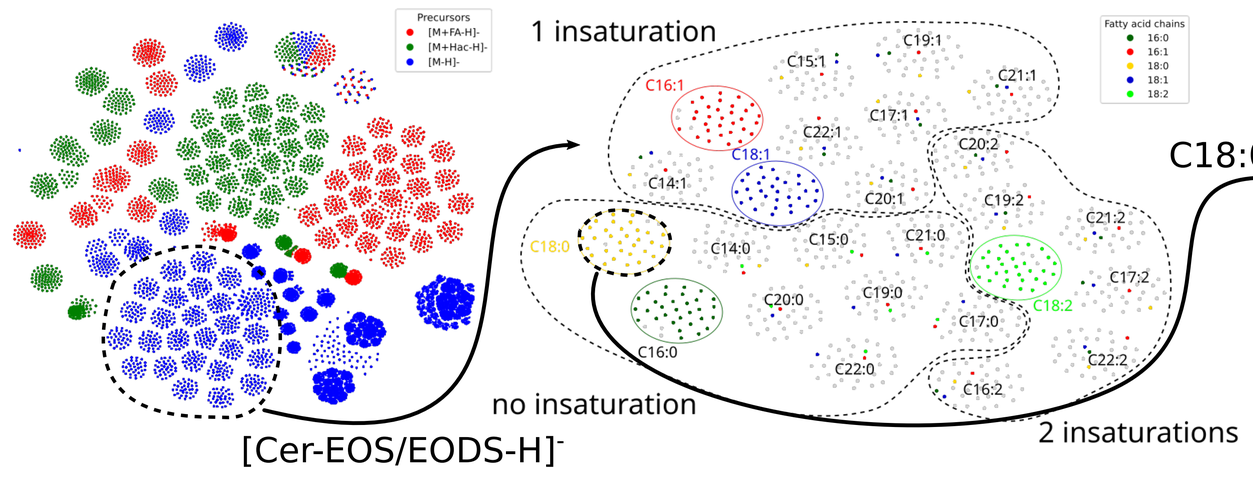
1.4×10^5 spectra 28 lipid families



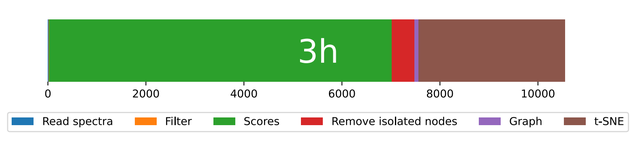
Dense matrix: 76 GiB
Sparse matrix: 3 GiB

Negative



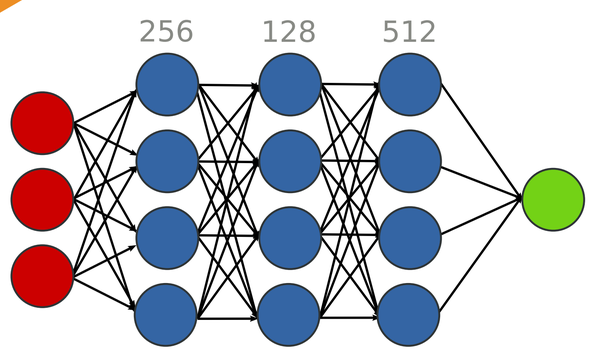


3.4×10^5 spectra 42 lipid families



Dense matrix: 440 GiB
Sparse matrix: 4 GiB

Could we use Deep Learning to predict position of new points?



MultiLayer Perceptron

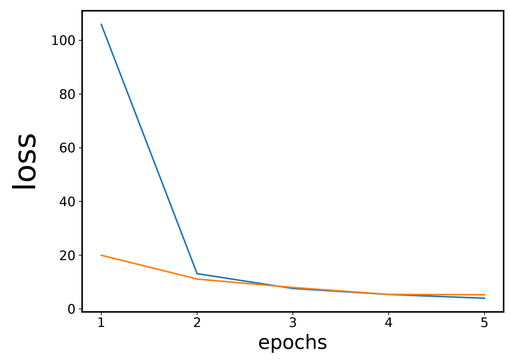
Dataset was split in 3 subsets

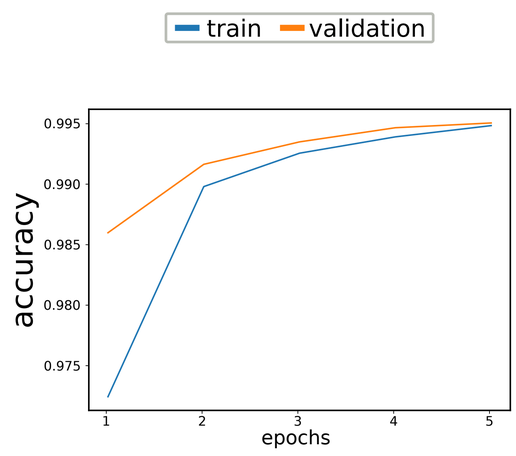
Training set 60 %
Test set 20 %
Validation set 20 %

Accuracy on validation

99.45%

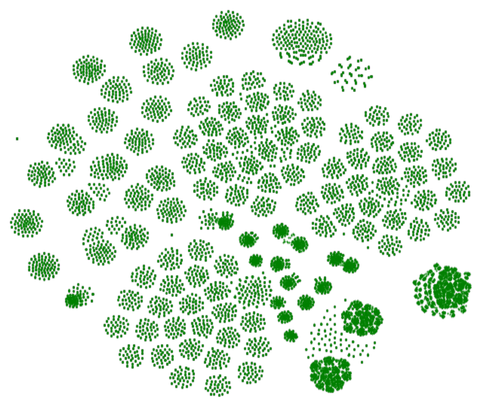
Parameter	Value
Loss	mse
Metrics	accuracy
# epochs	5





train validation

t-SNE projection of test set



Model predictions