



HAL
open science

Réseau moléculaire massif pour l'annotation en lipidomique

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

► **To cite this version:**

Nicolas Elie, Romain Magny, Nicolas Auzeil, Olivier Laprévotte, David Touboul. Réseau moléculaire massif pour l'annotation en lipidomique. Journées du Réseau Francophone de Métabolomique et Fluxomique (RFMF 2021), Nov 2021, Aussois, France. hal-04427181

HAL Id: hal-04427181

<https://hal.science/hal-04427181>

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.

Réseau moléculaire massif pour l'annotation en lipidomique



Nicolas Elie¹, Romain Magny^{2,3}, Nicolas Auzeil³, Olivier Laprèvote^{3,4} et David Touboul¹



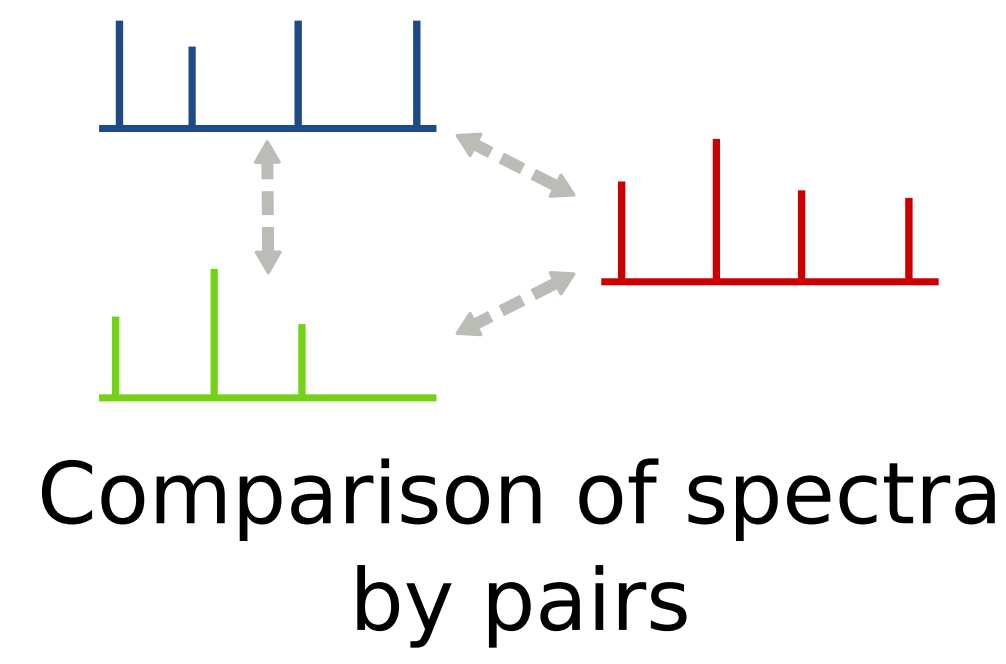
1. Institut de Chimie des Substances Naturelles, CNRS UPR 2301, Université Paris-Sud, Université Paris-Saclay, Gif-sur-Yvette, France
2. Sorbonne Université UM80, INSERM UMR 968, CNRS UMR 7210, Institut de la Vision, IHU ForeSight, 75006 Paris, France
3. UMR CNRS 8038 CiToM, Chimie Toxicologie Analytique 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

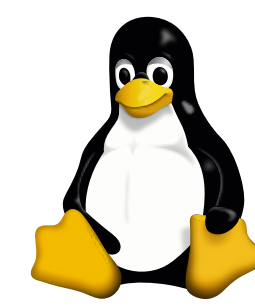
~5 × 10⁵ *in silico* lipids spectra

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

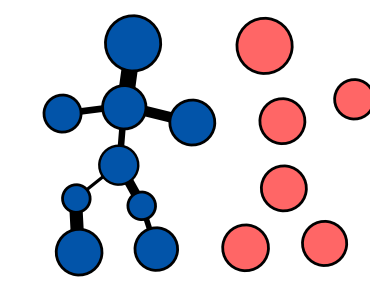
t-SNE parameters



t-SNE projection



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 similarity scores in sparse matrices (Compressed Sparse Row format)

		indices				
indptr	0	0	0	3	0	4
	1	0	0	5	7	0
	2	0	0	0	0	0
	3	0	0	0	0	0
	4	0	2	6	0	0

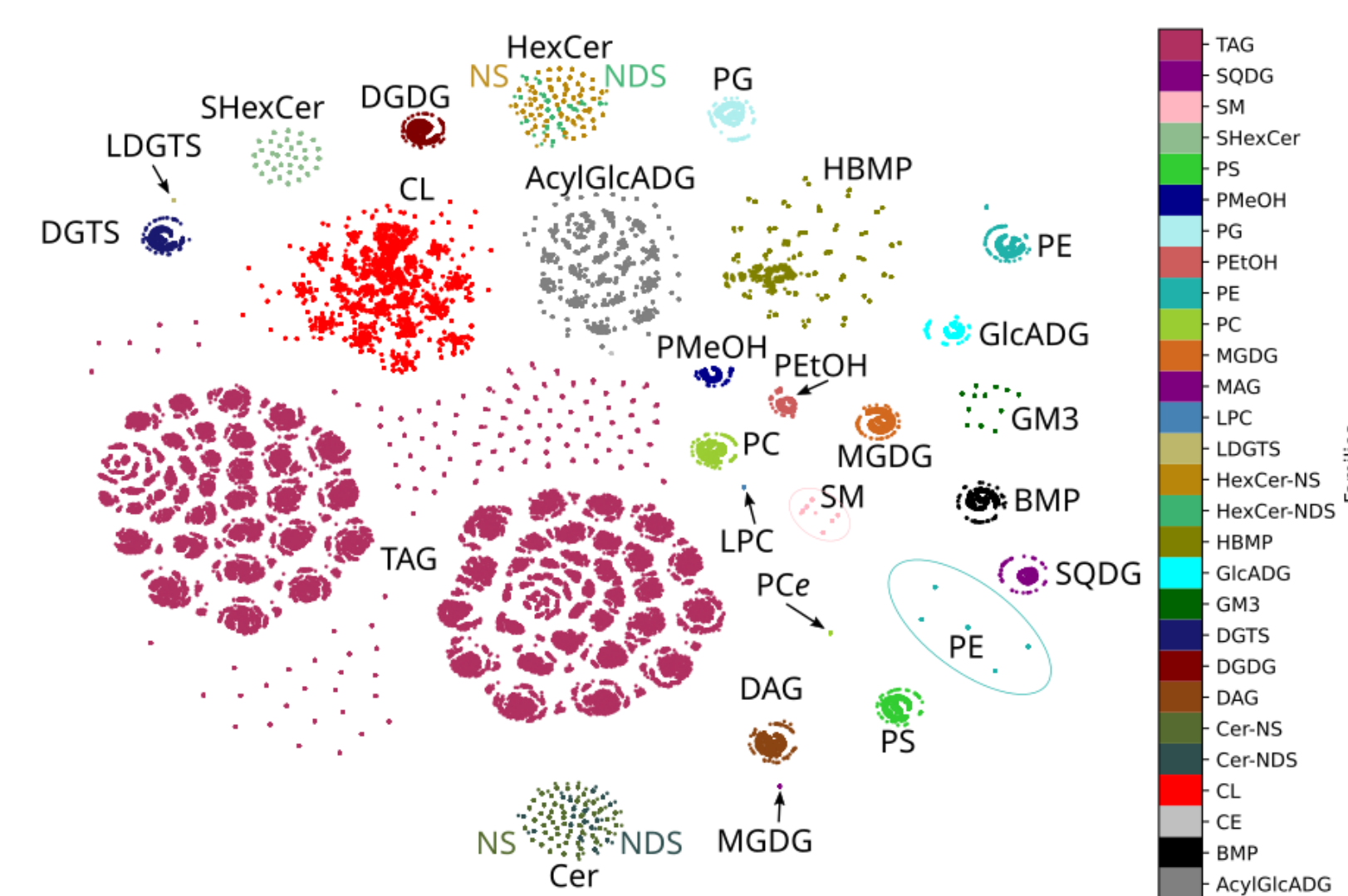
indices	0	2	4	4	6	
indptr	2	4	2	3	1	2
data	3	4	5	7	2	6

Families

Precursors

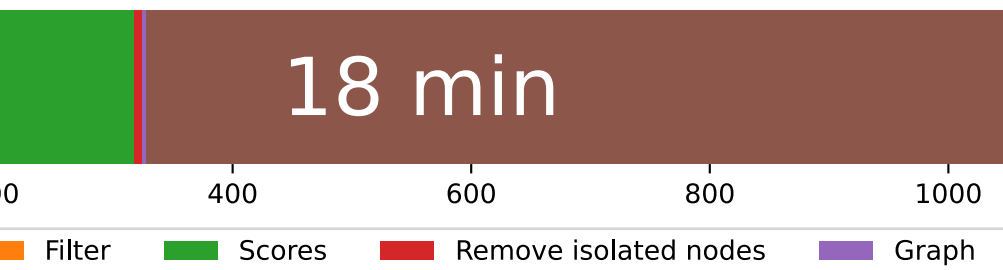
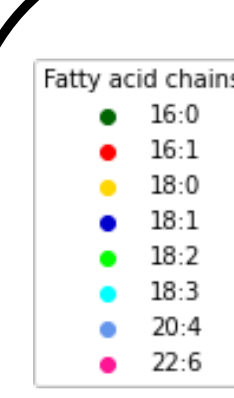
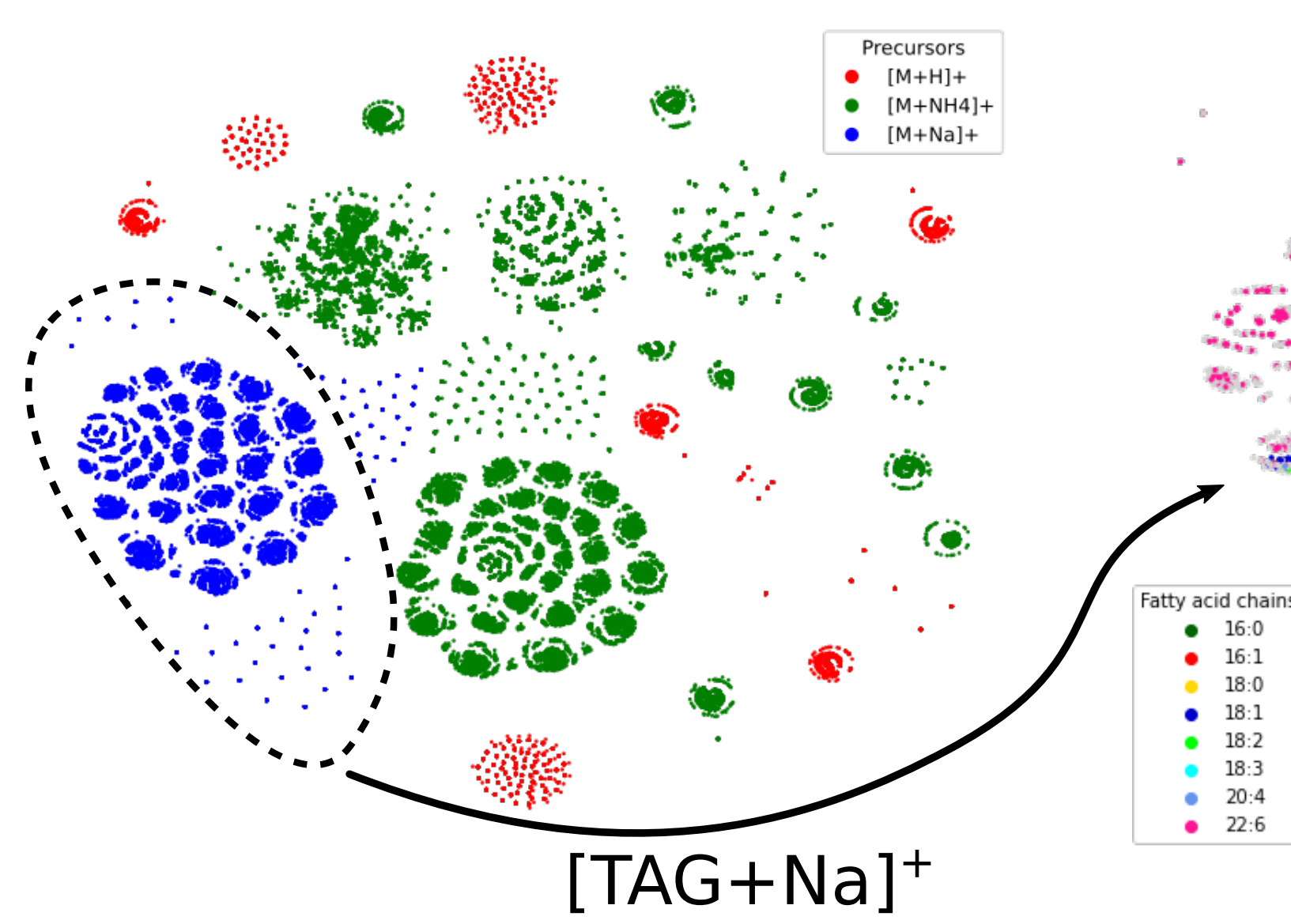
Fatty acid chains

Positive



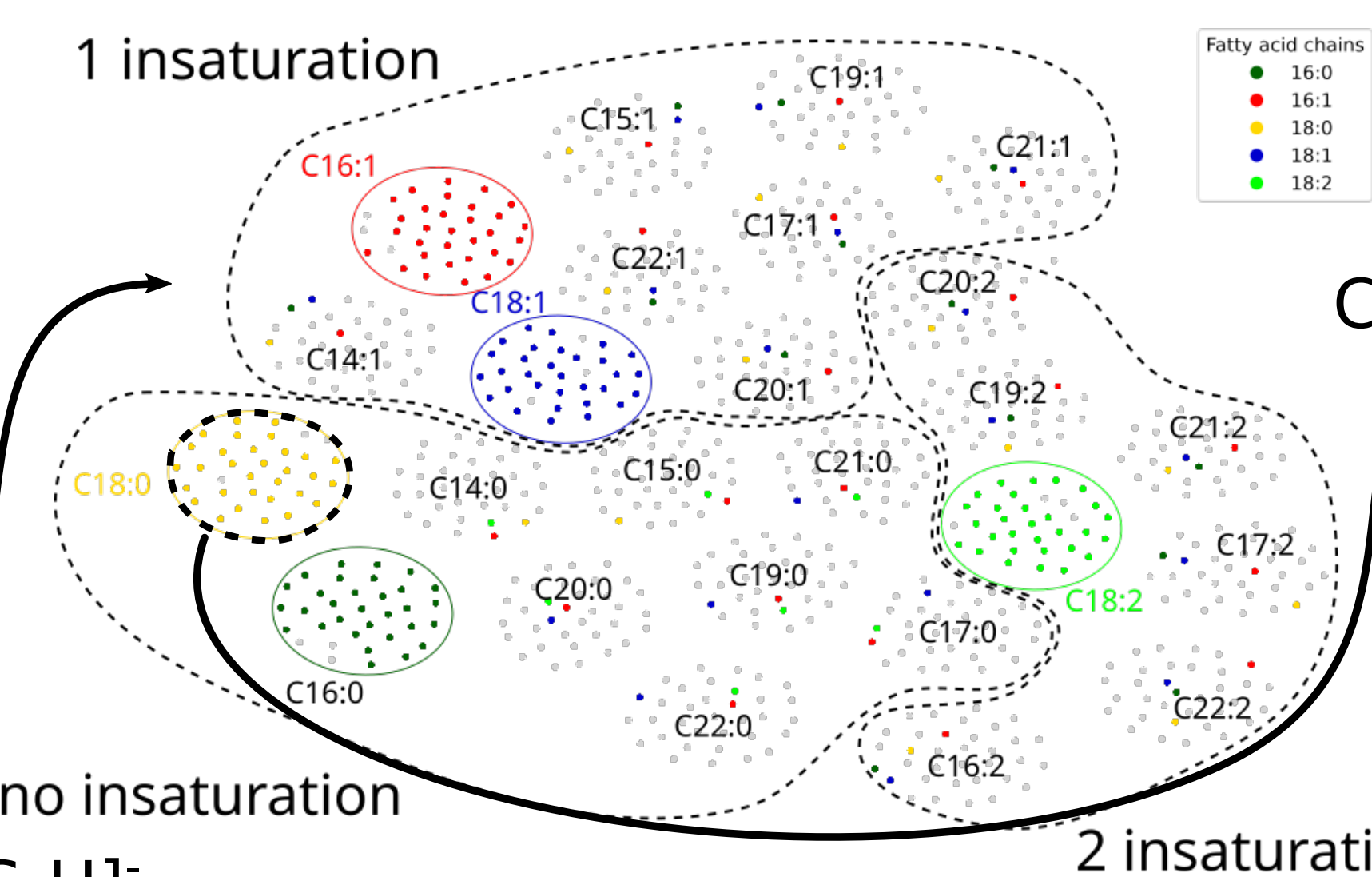
1.4 × 10⁵ spectra

28 lipid families



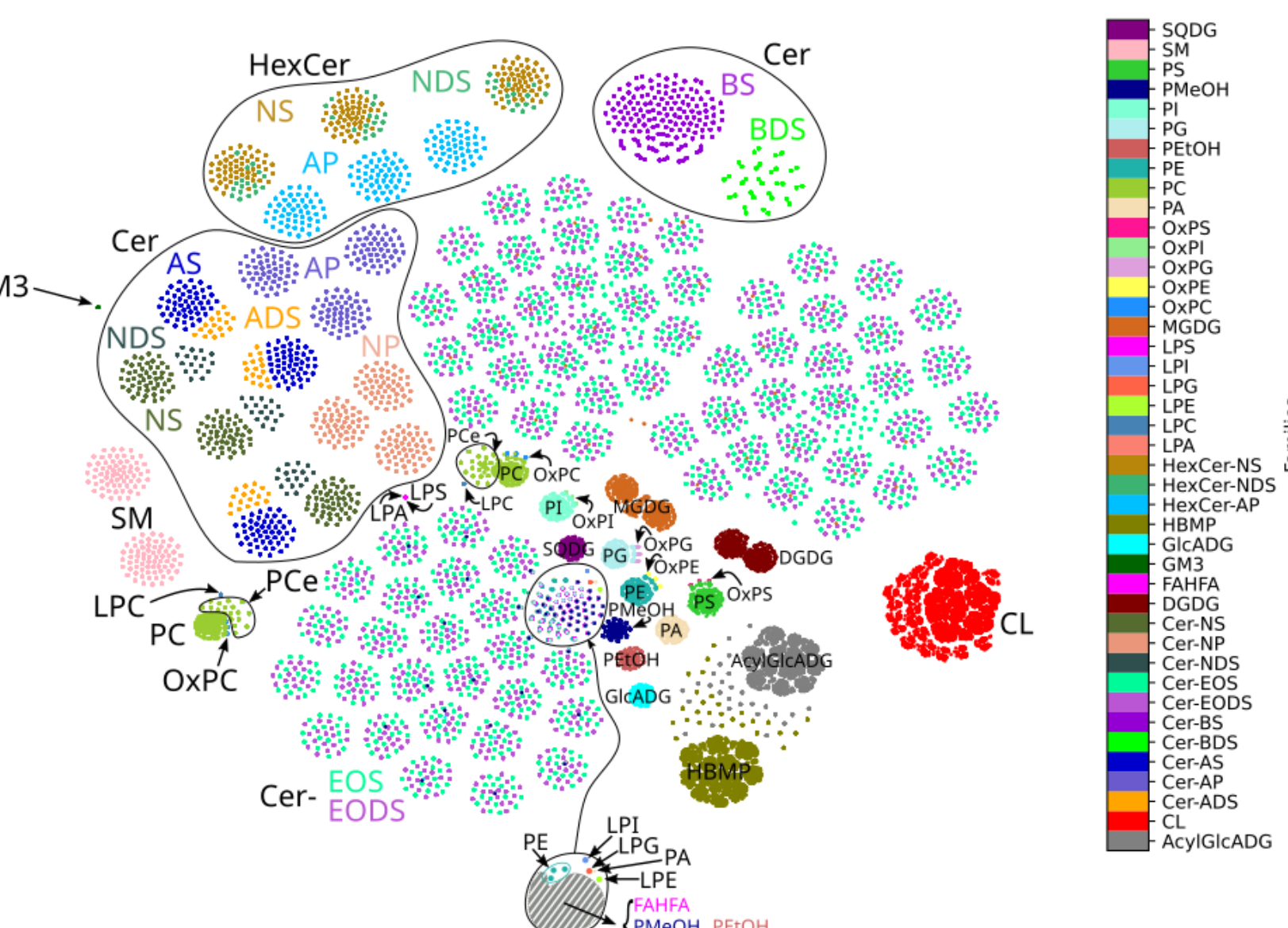
Read spectra Filter Scores Remove isolated nodes Graph t-SNE

1 insaturation



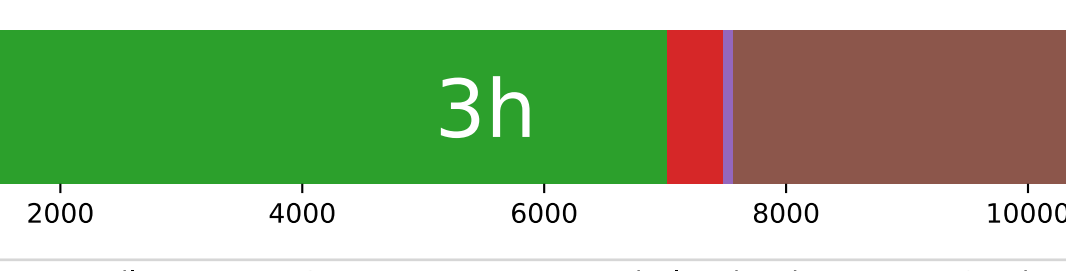
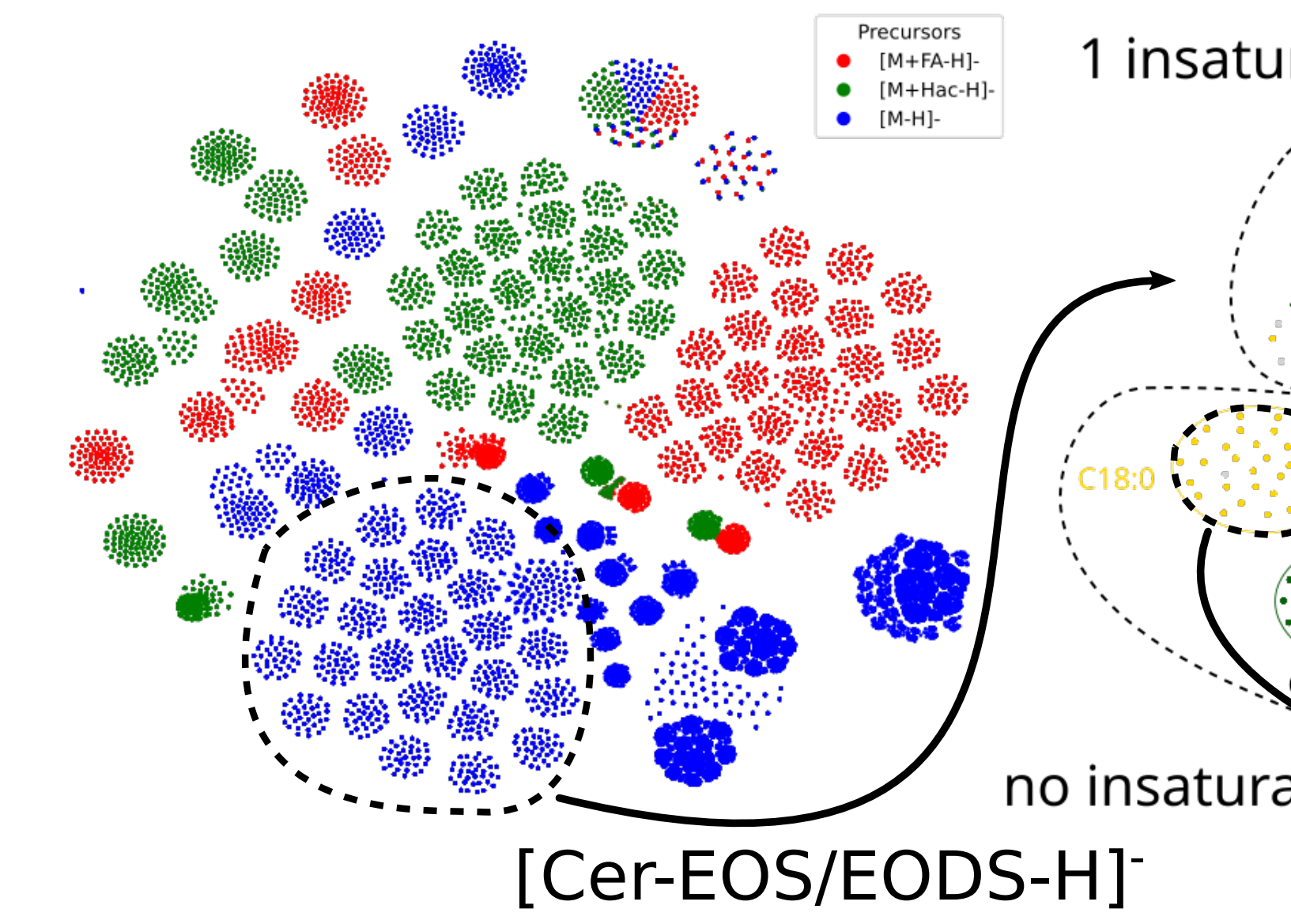
Dense matrix: 76 GiB
Sparse matrix: 3 GiB

Negative



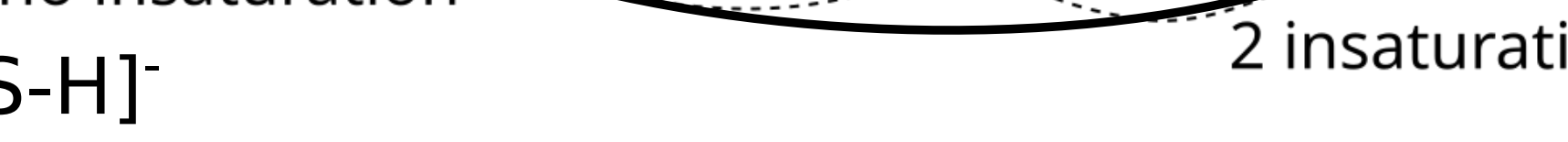
3.4 × 10⁵ spectra

42 lipid families



Read spectra Filter Scores Remove isolated nodes Graph t-SNE

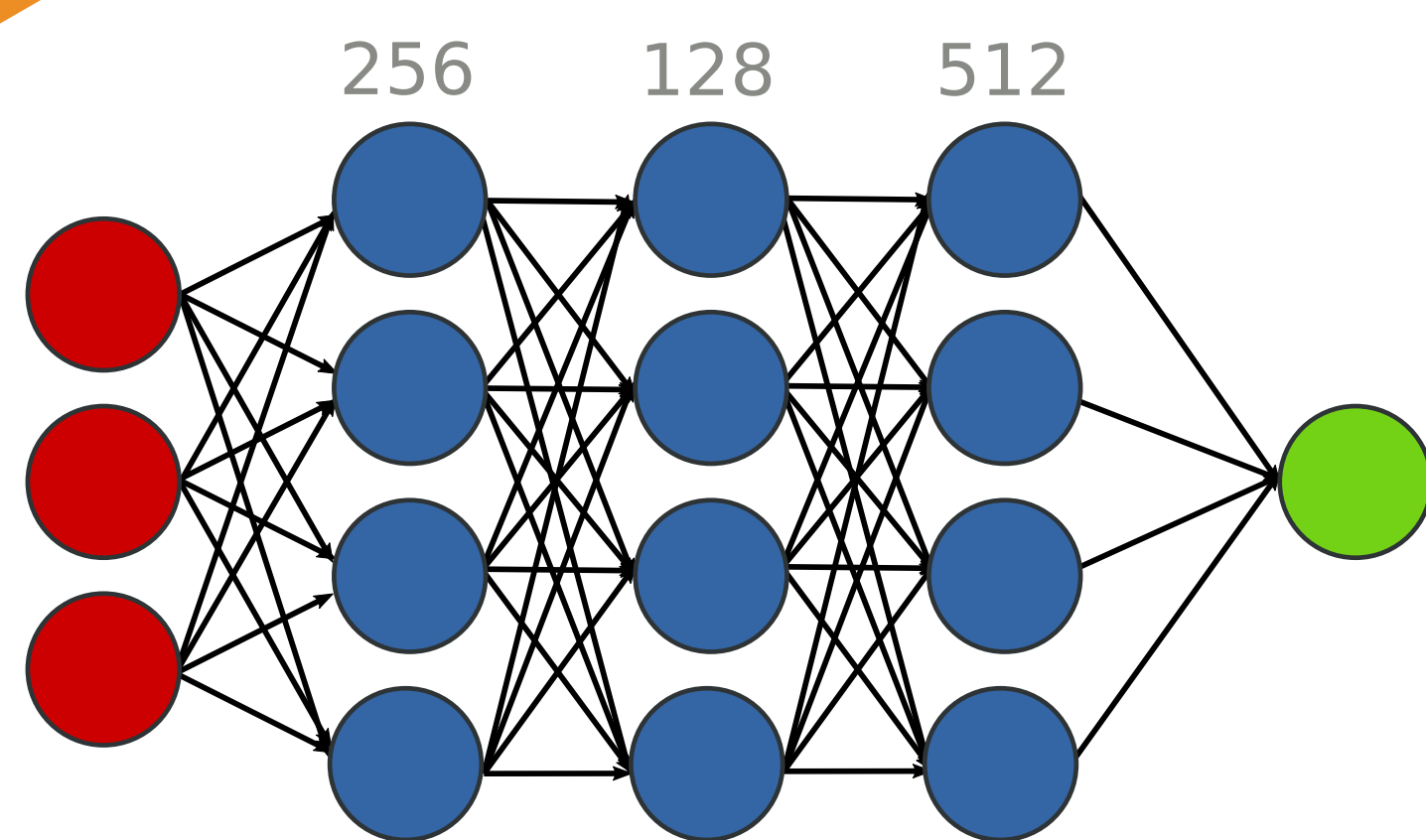
no insaturation



Dense matrix: 440 GiB
Sparse matrix: 4 GiB

t-SNE projection of test set

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%

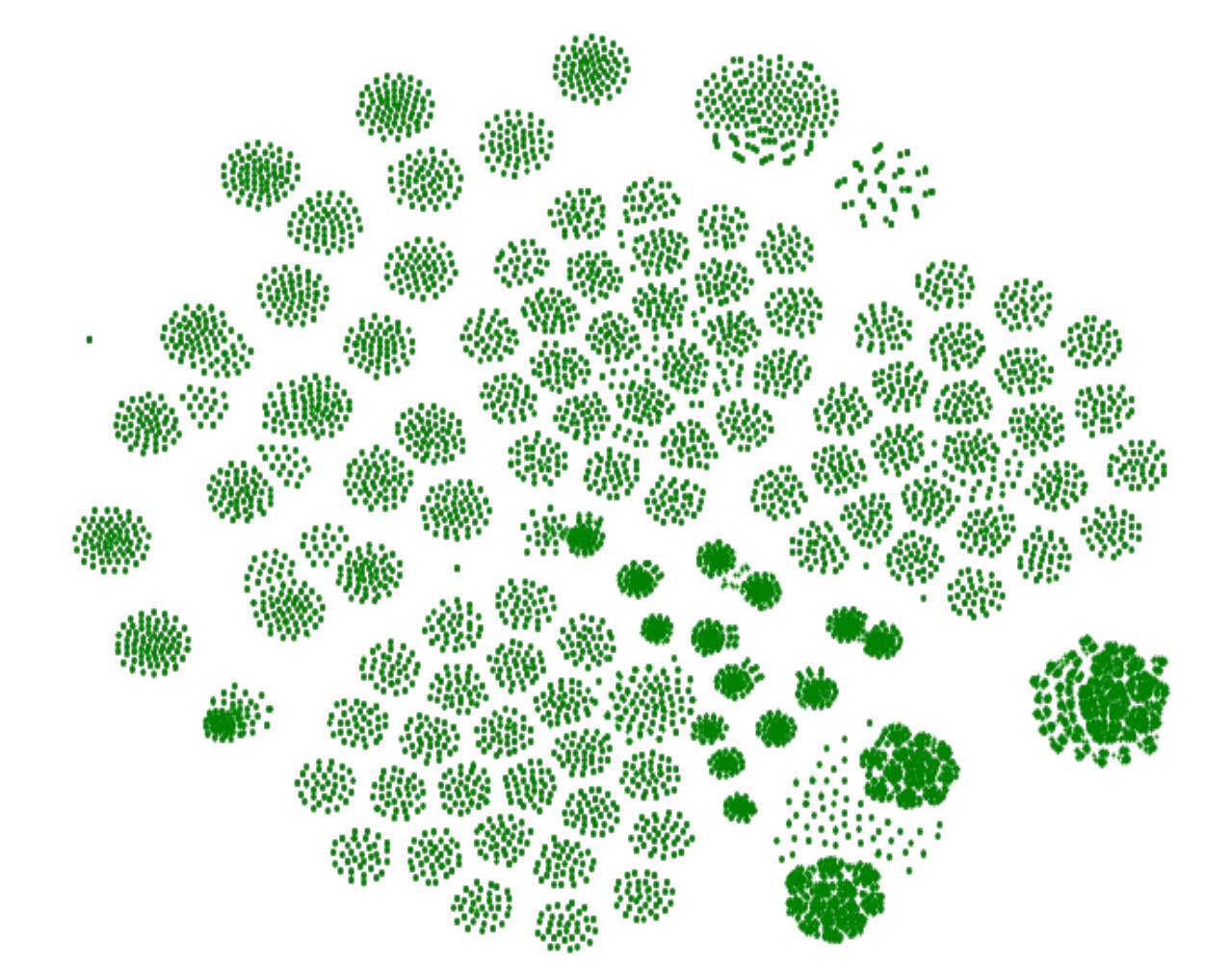
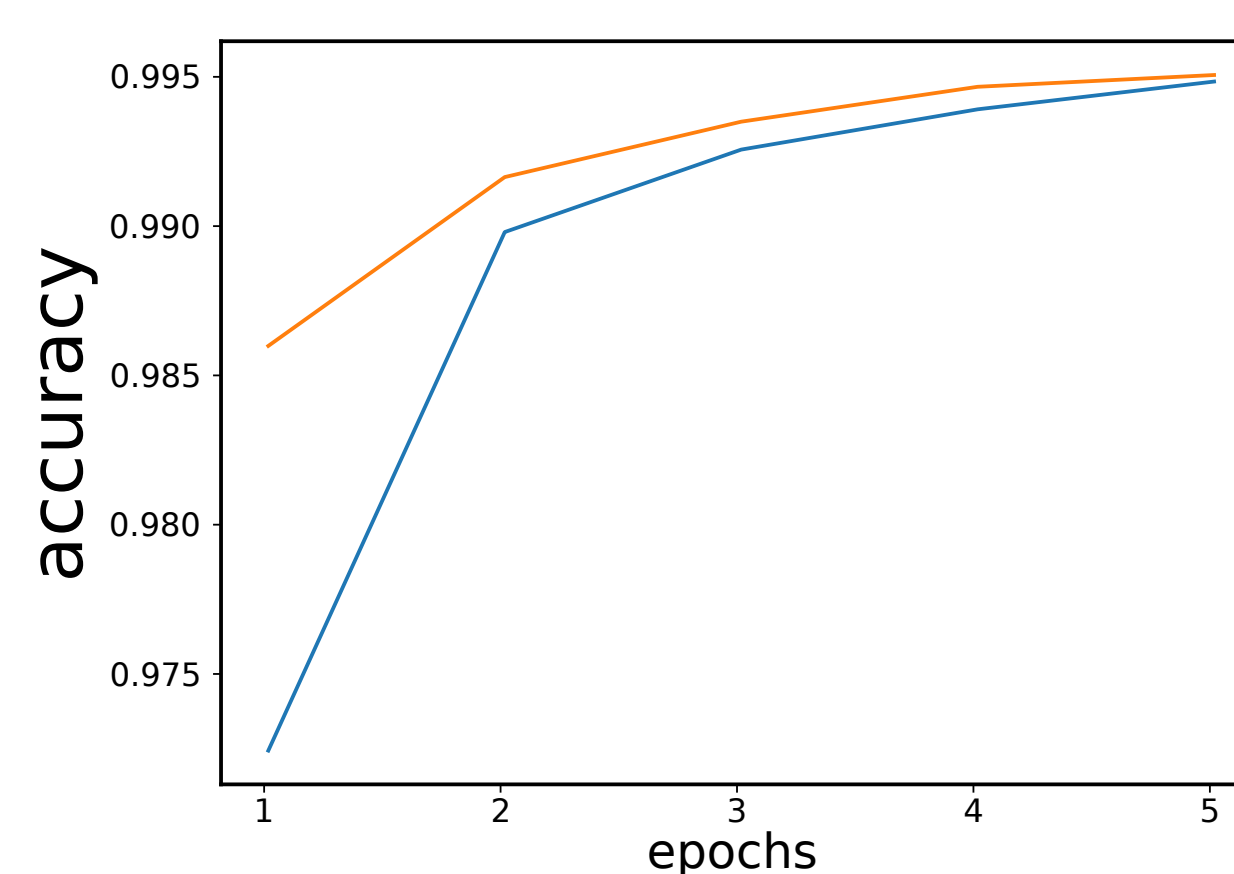
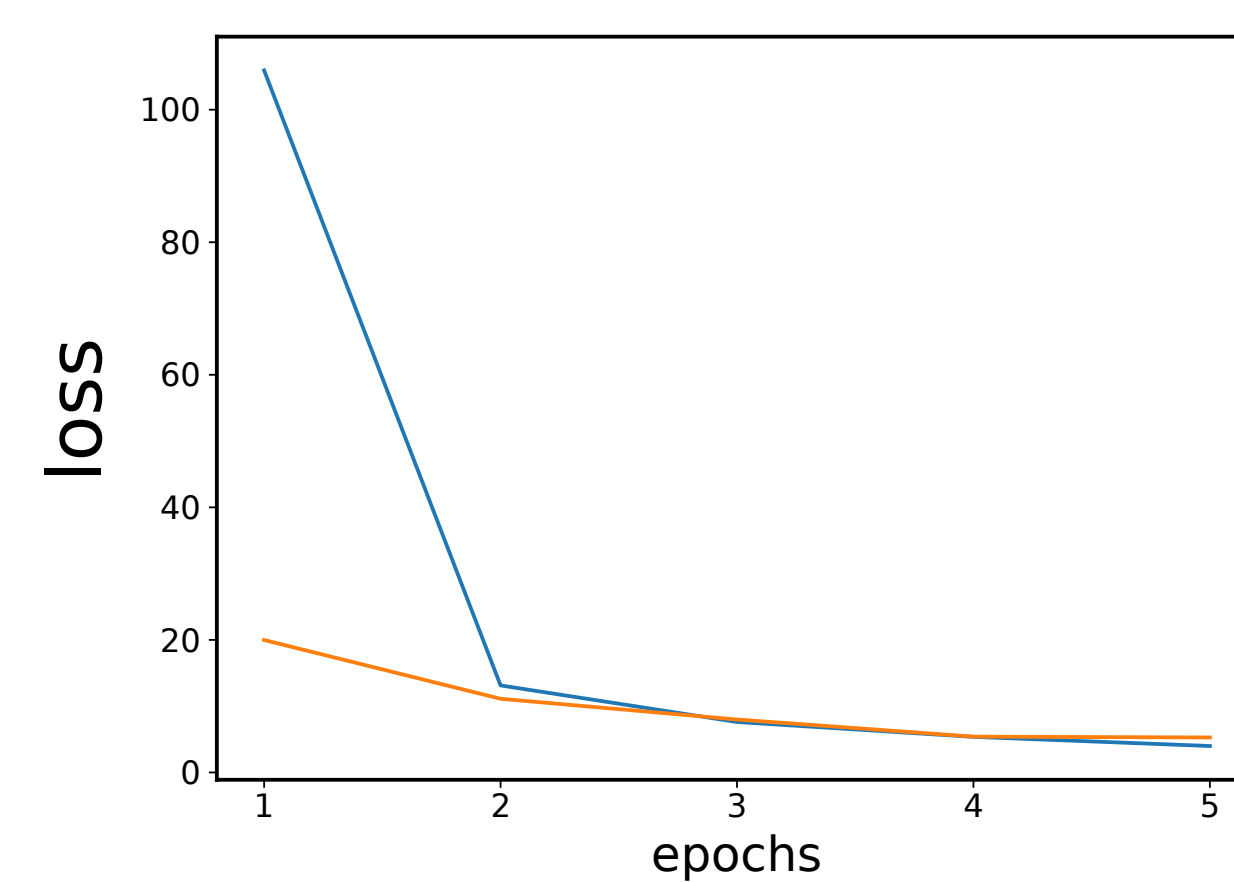
Validation set 20%

Accuracy on validation



99.45%

Parameter	Value
Loss	mse
Metrics	accuracy
# epochs	5



Model predictions

