Unsupervised Machine Learning for Unbiased Chemical Classification in X-ray Absorption and Emission Spectroscopies

Samantha Tetef, Niranjan Govind, Gerald T. Seidler

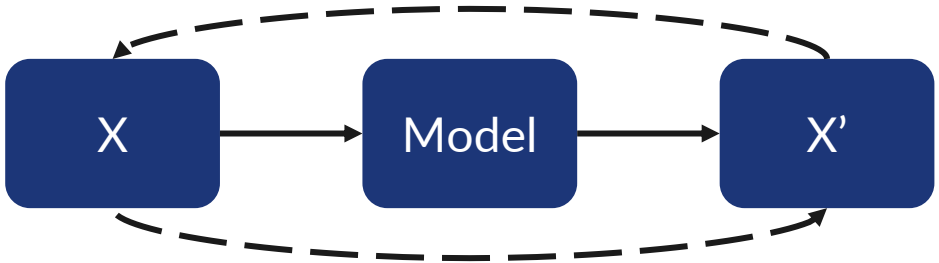
DIRECT
Data Intensive Research
Enabling Clean Technologies

Question:
What is the chemically-relevant information in two related X-ray spectroscopy methods – VtC-XES (emission) and XANES (absorption)?

Technique:
We can use unsupervised machine learning (ML) to allow spectral fingerprints emerge from the data and use it to generate unbiased chemical classes

I picked 2D to see what was going on

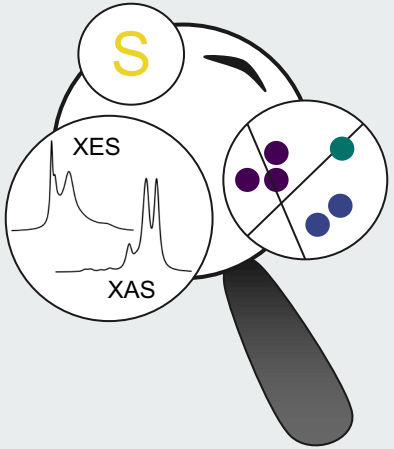
High dimensional input

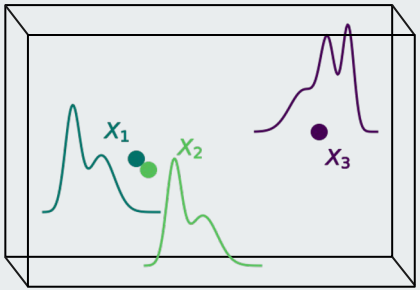


Low dimensional output

Outline

1. Background on spectroscopy methods:
   Absorption (XAS) vs Emission/ Fluorescence (XES)
2. Background on dimensionality reduction (DR) algorithms
3. Interpreting the reduced space via cluster similarities
   a. Expected & unexpected fingerprints
4. Comparing DR algorithms using supervised ML on identifying the fingerprints found above

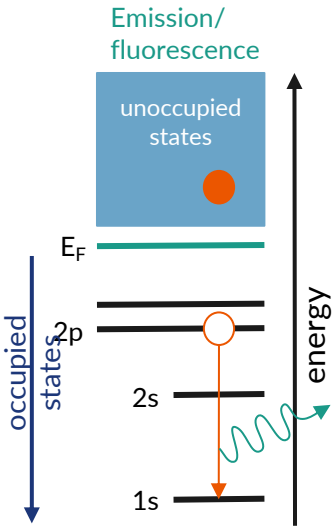




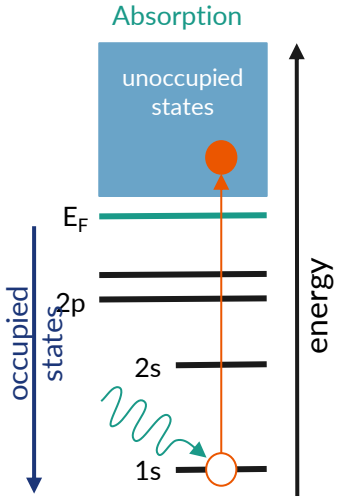
VtC-XES versus XANES

Are they indeed complementary?

Valence to Core X-ray **Emission** Spectroscopy



X-ray **Absorption** Near-Edge Fine Structure

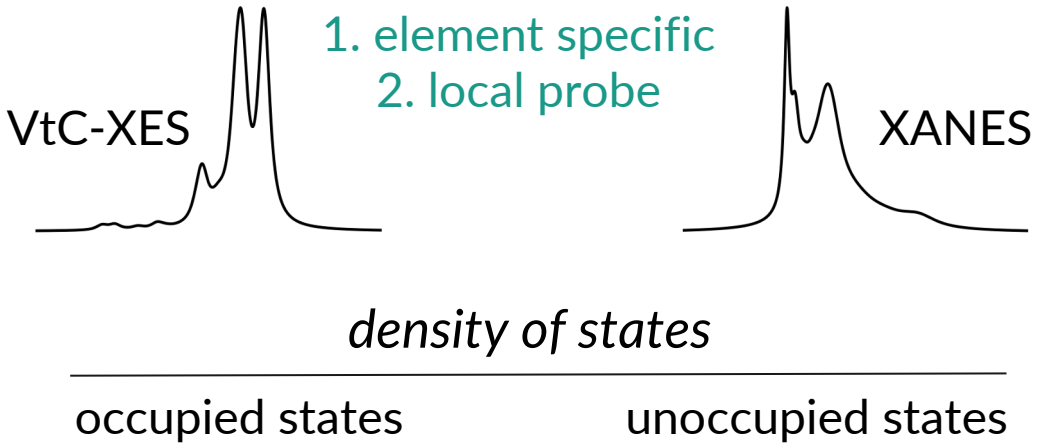


VtC-XES versus XANES

Are they indeed complementary?

Valence to Core X-ray **Emission** Spectroscopy

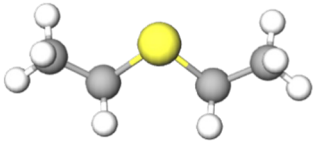
X-ray **Absorption** Near-Edge Fine Structure



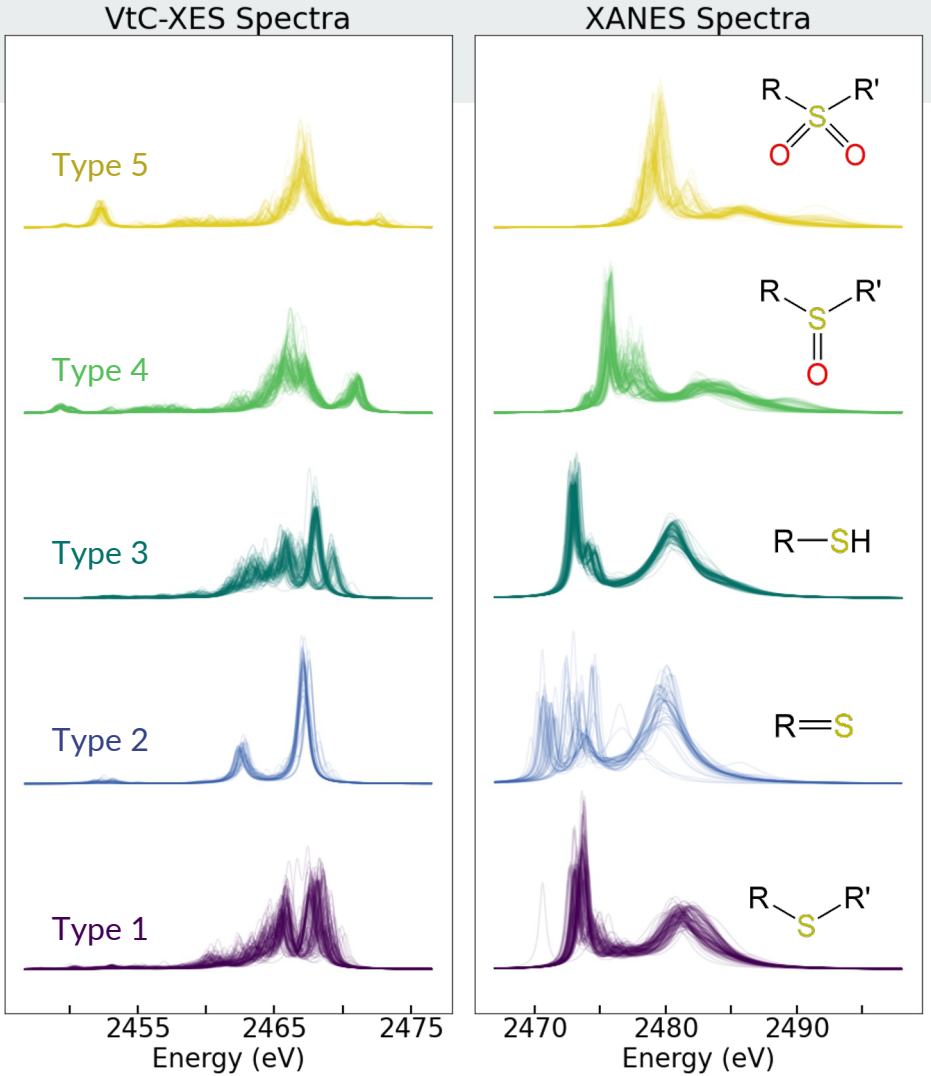
Hypothesis: The properties encoded in VtC-XES and XANES will be similar, but their sensitivities will be different.

The Dataset

Obtain structures



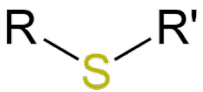
TD-DFT calculation



Sulforganic classes from the literature

Type 1

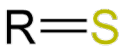
C-S-C bonds
sulfides



1. conjugated
2. not conjugated

Type 2

C=S bonds
thiocarbonyls



1. conjugated
2. not conjugated

Type 3

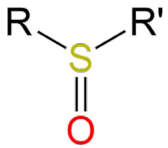
C-SH bonds
thiols/mercaptans

1. conjugated
2. not conjugated

Oxidation:
-2

Type 4

S=O bonds
sulfoxides

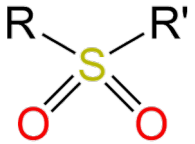


1. conjugated
2. not conjugated

Oxidation:
0

Type 5

(S=O)2 bonds
sulfones

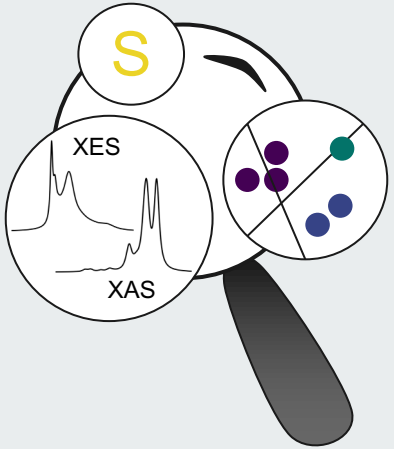


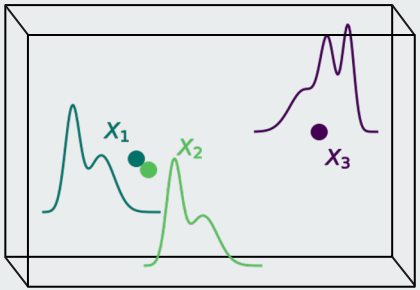
1. sulfone
2. sulfonamide

Oxidation:
+2

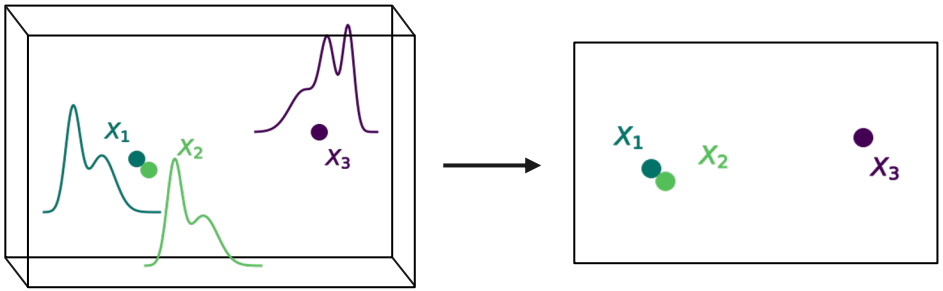
Outline

1. Background on spectroscopy methods:
   Absorption (XAS) vs Emission/ Fluorescence (XES)
2. **Background on dimensionality reduction (DR) algorithms**
3. Interpreting the reduced space via cluster similarities
   a. Expected & unexpected fingerprints
4. Comparing DR algorithms using supervised ML on identifying the fingerprints found above

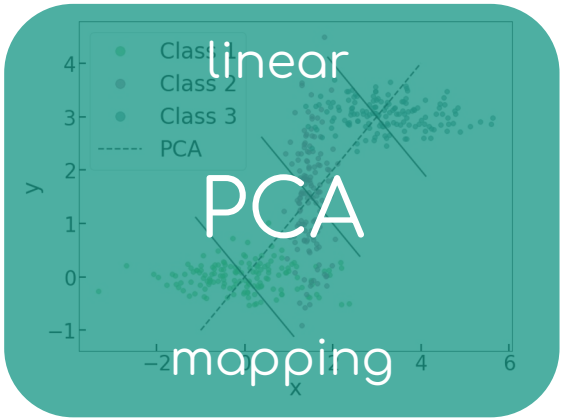


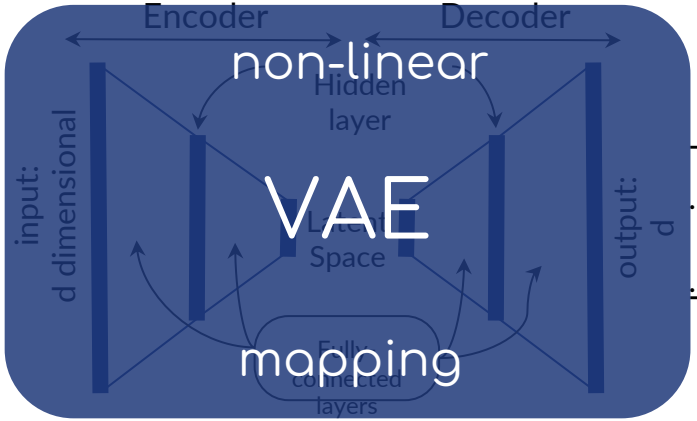


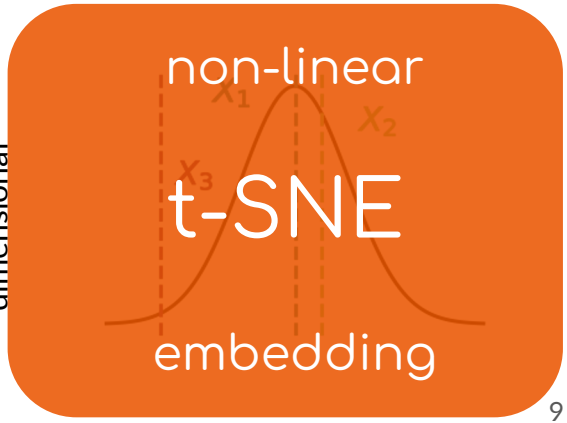
Unsupervised ML extracts features by grouping similar data



more rigid → *fewer constraints*

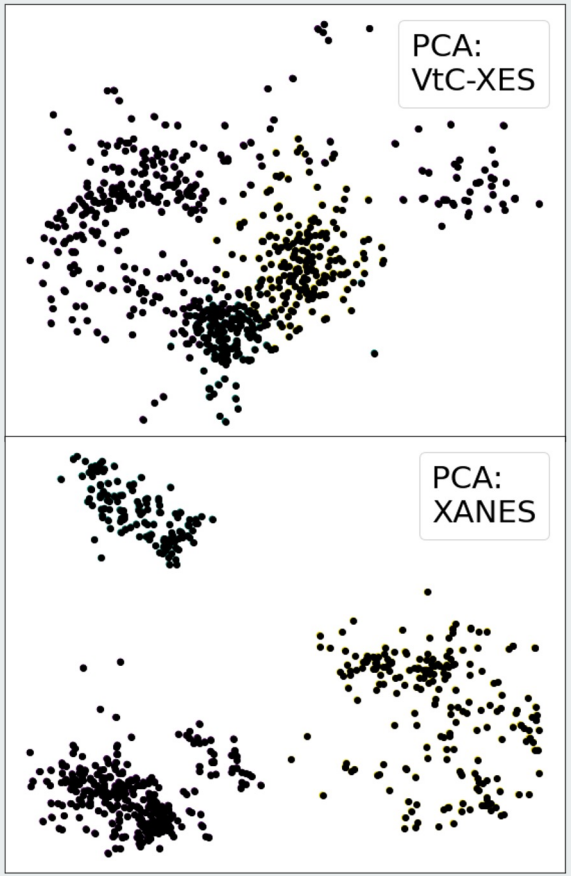


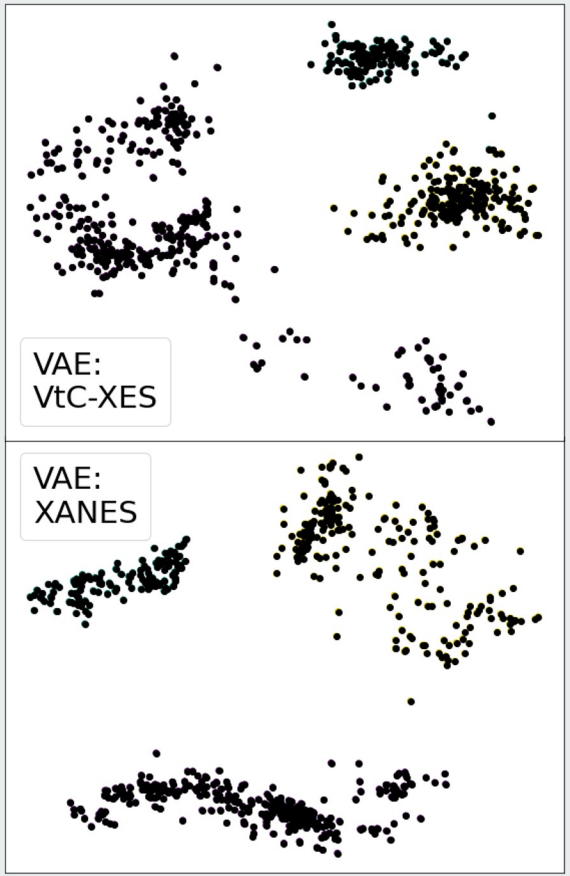


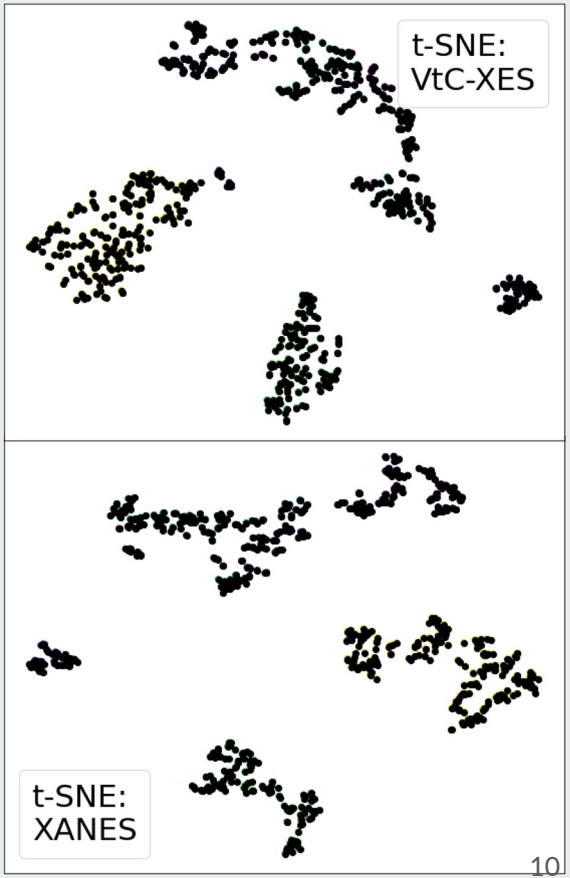


Clustering → spectral similarities → important chemical properties

PCA VS **VAE** VS **t-SNE**

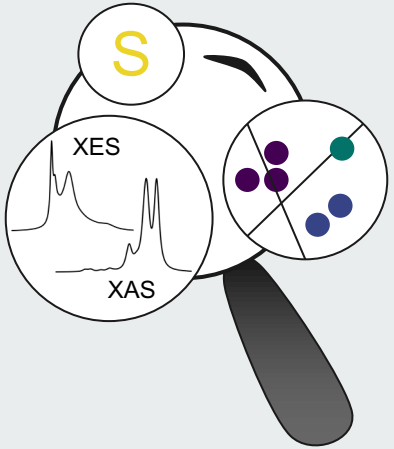


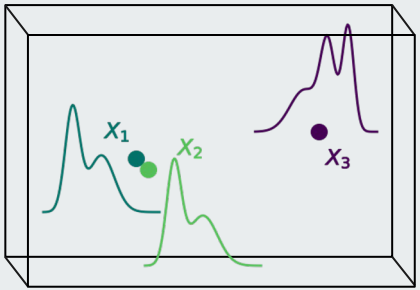




Outline

1. Background on spectroscopy methods:
   Absorption (XAS) vs Emission/ Fluorescence (XES)
2. Background on dimensionality reduction (DR) algorithms
3. **Interpreting the reduced space via cluster similarities**
   a. **Expected & unexpected fingerprints**
4. Comparing DR algorithms using supervised ML on identifying the fingerprints found above

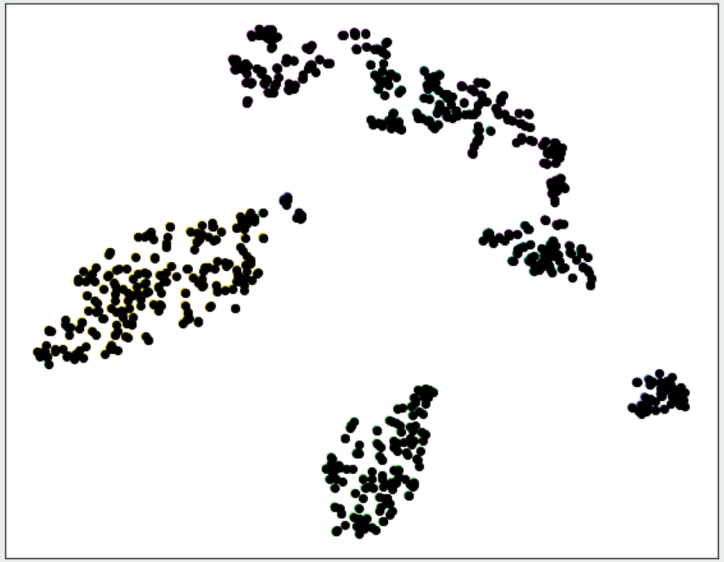




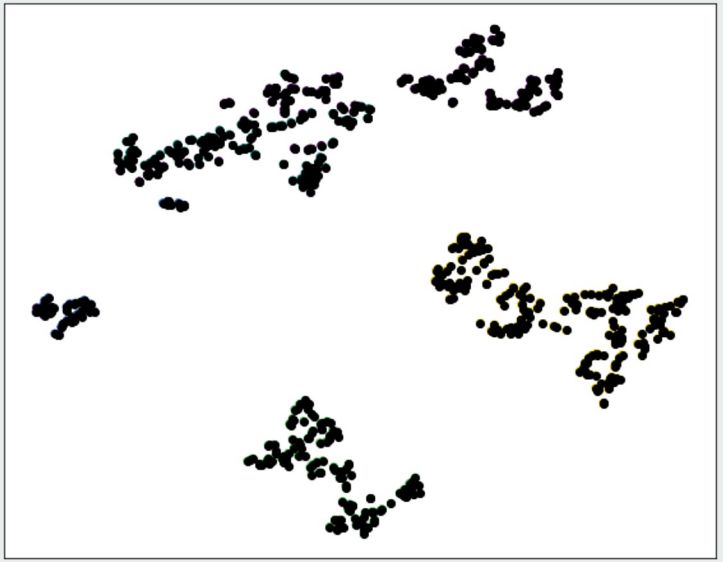
What do the clusters indicate about encoded information?

t-SNE reduced space

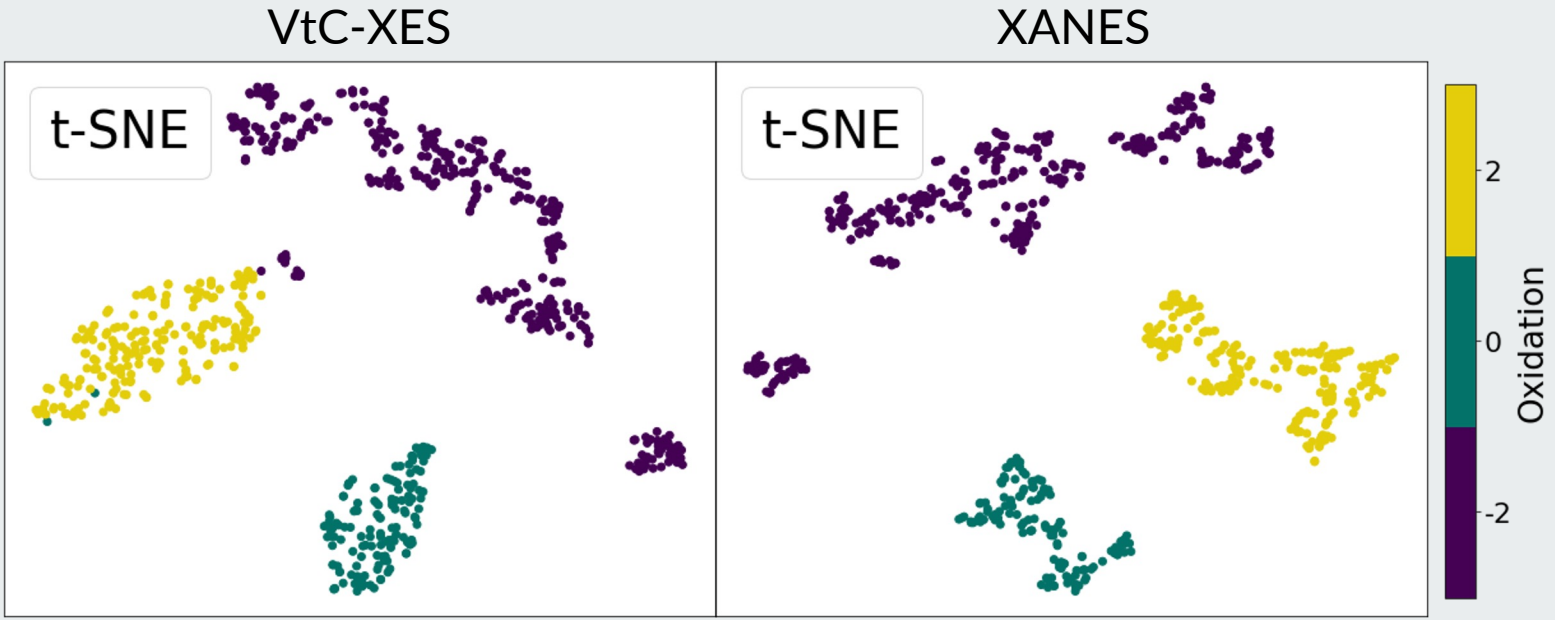
VtC-XES



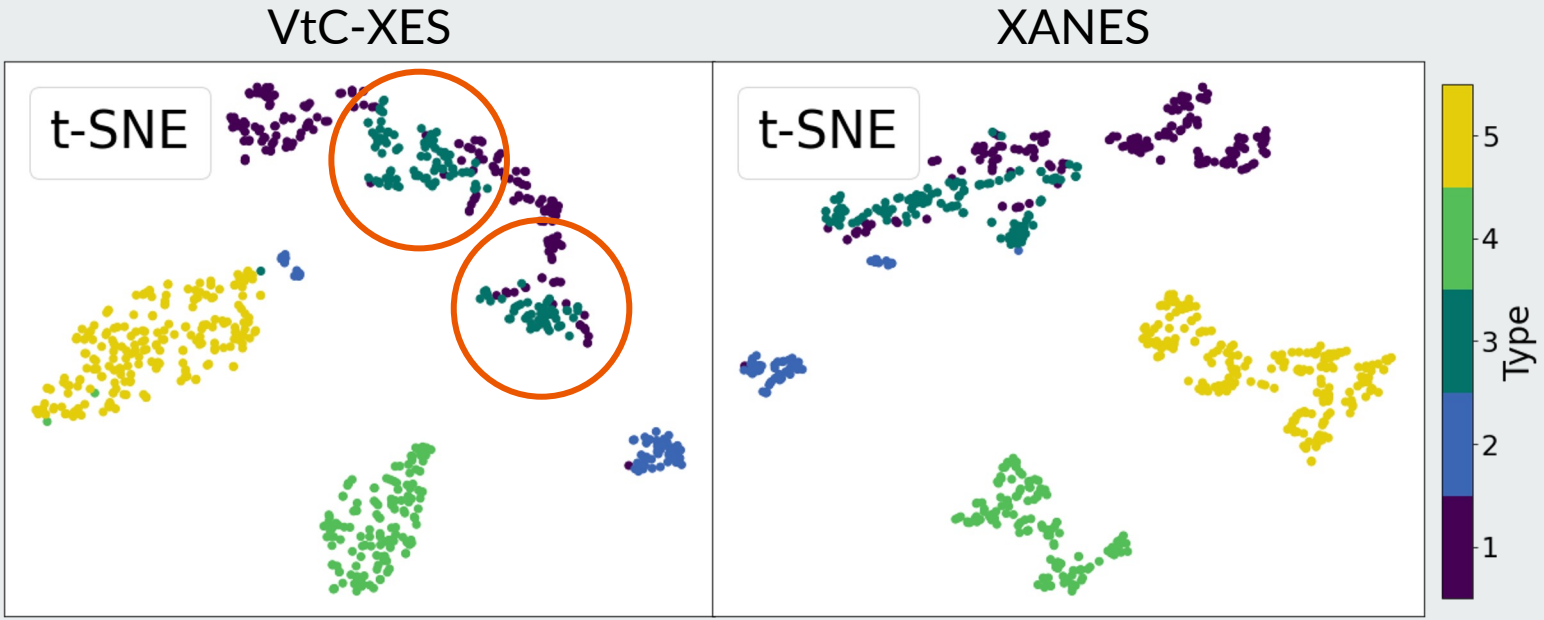
XANES



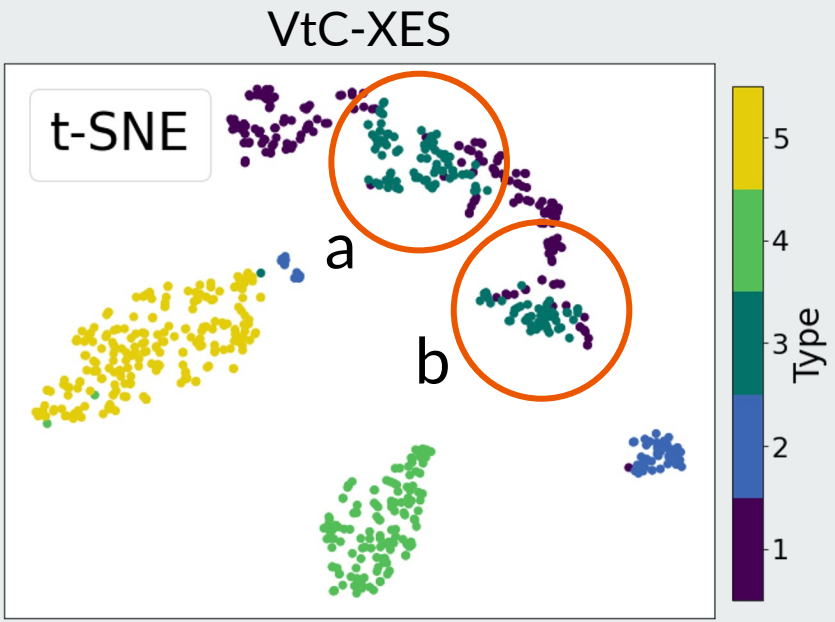
Oxidation state is the best distinguishing property



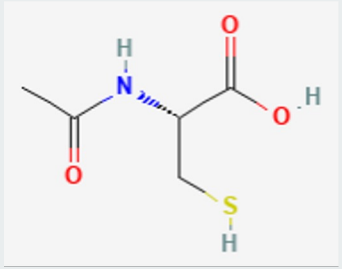
“Type” is also a good distinguishing property

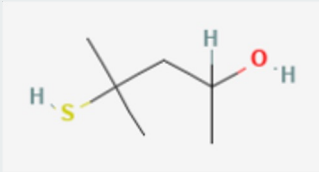


“Type” is also a good distinguishing property

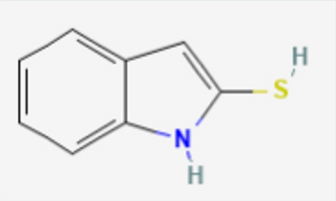


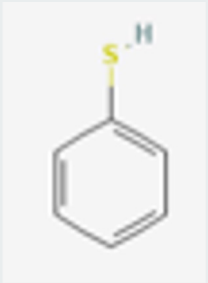
a





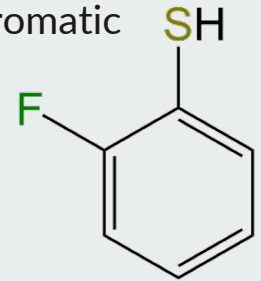
b



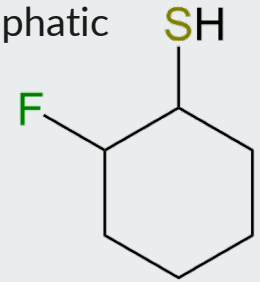


Delocalized = aromatic vs localized = aliphatic

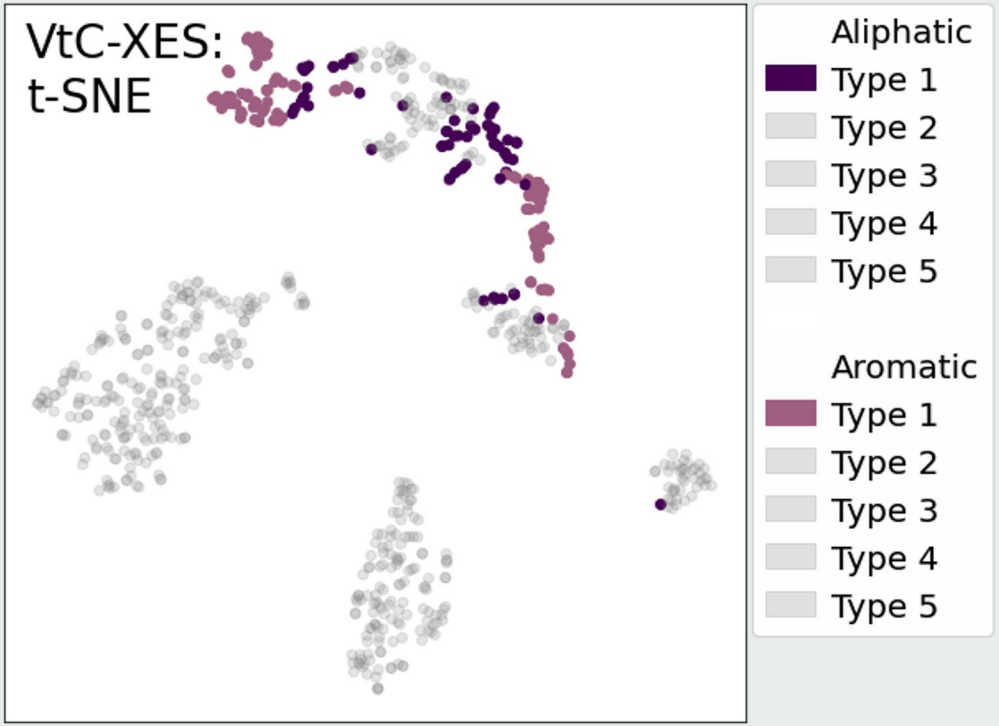
Aromatic

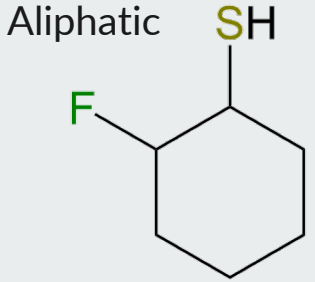


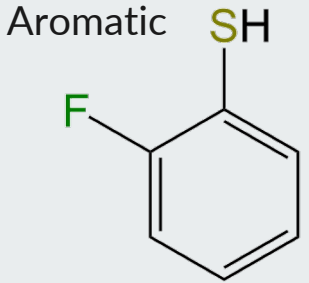
Aliphatic



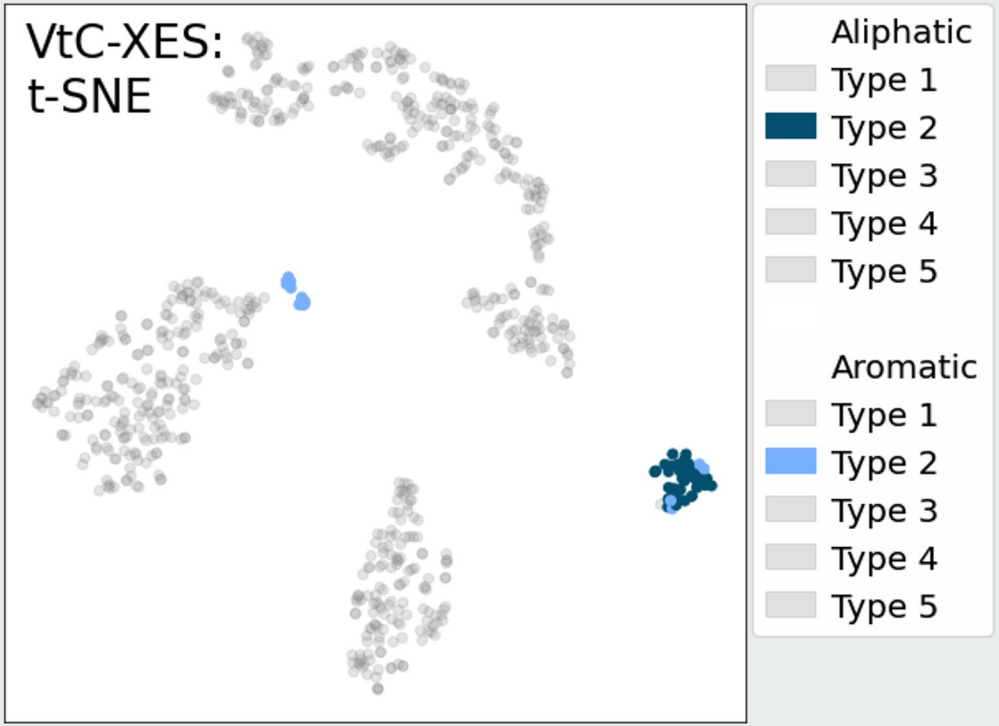
Delocalized = aromatic vs localized = aliphatic

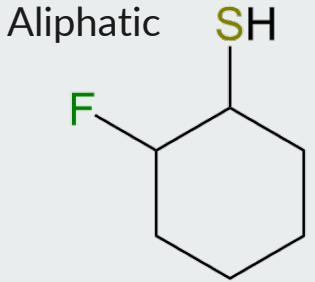


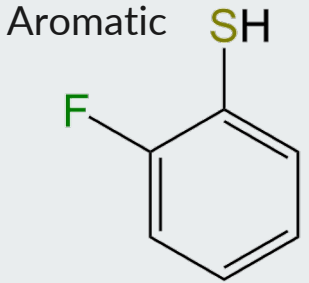




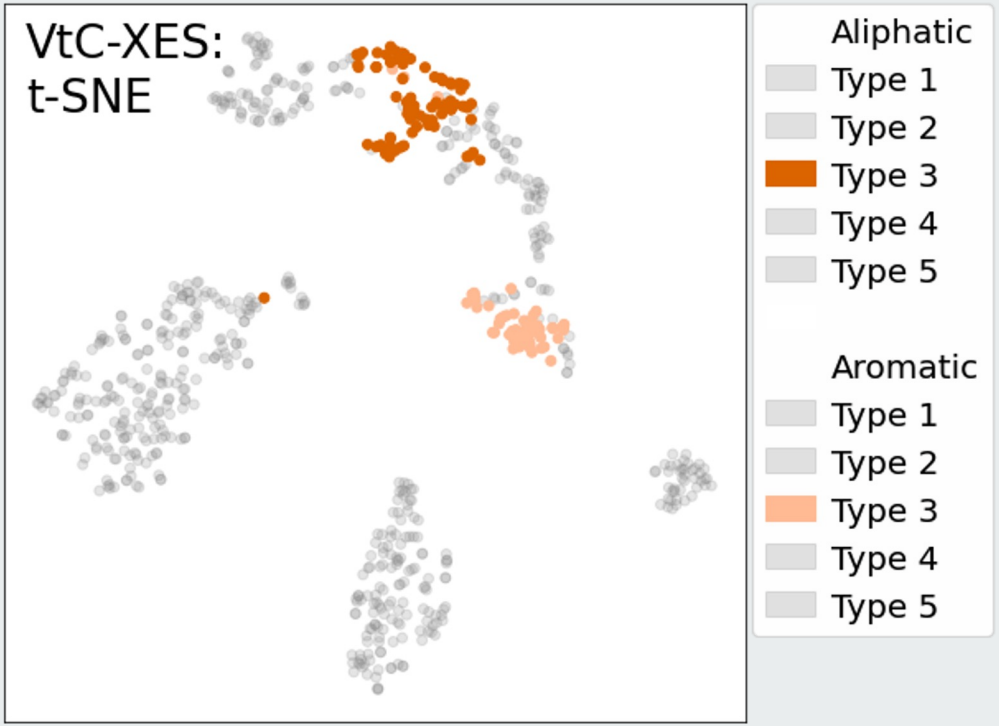
Delocalized = aromatic vs localized = aliphatic

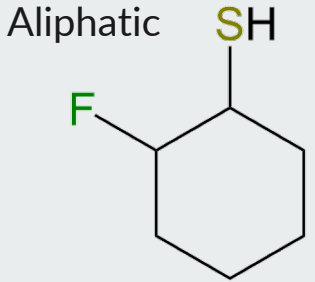


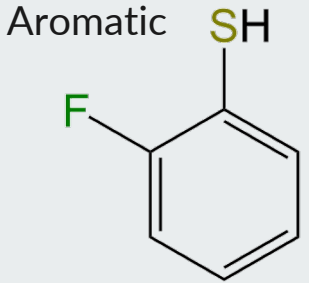




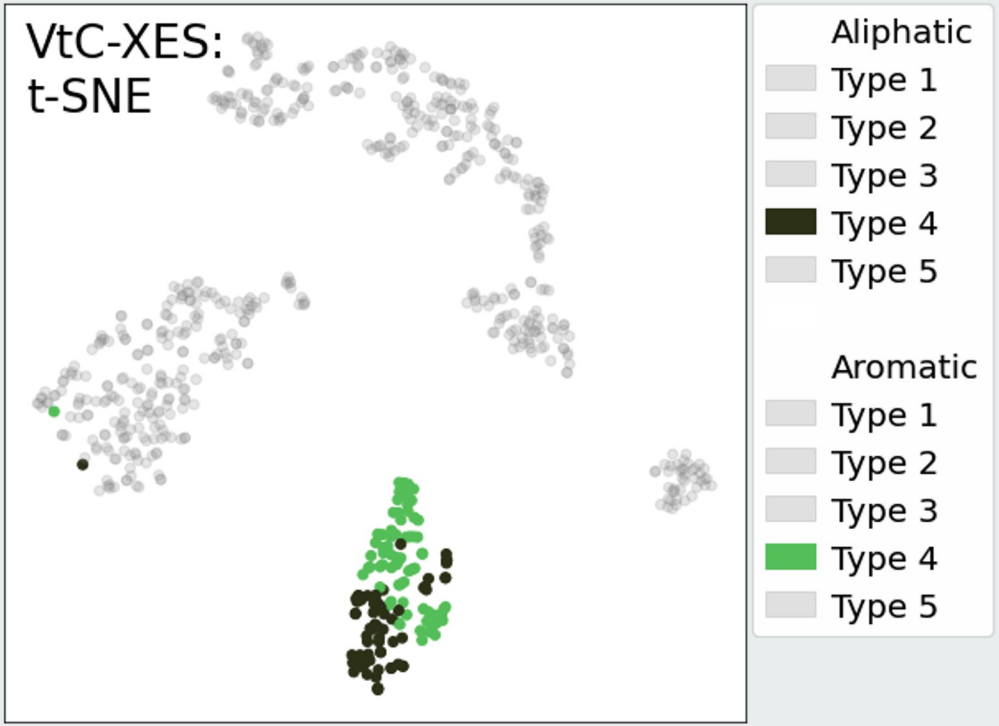
Delocalized = aromatic vs localized = aliphatic

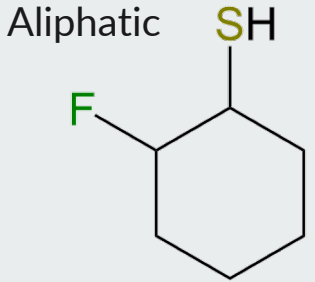


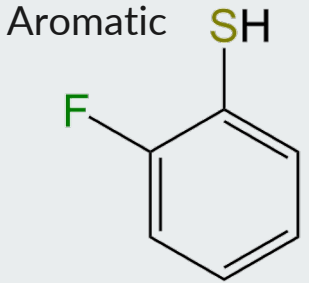




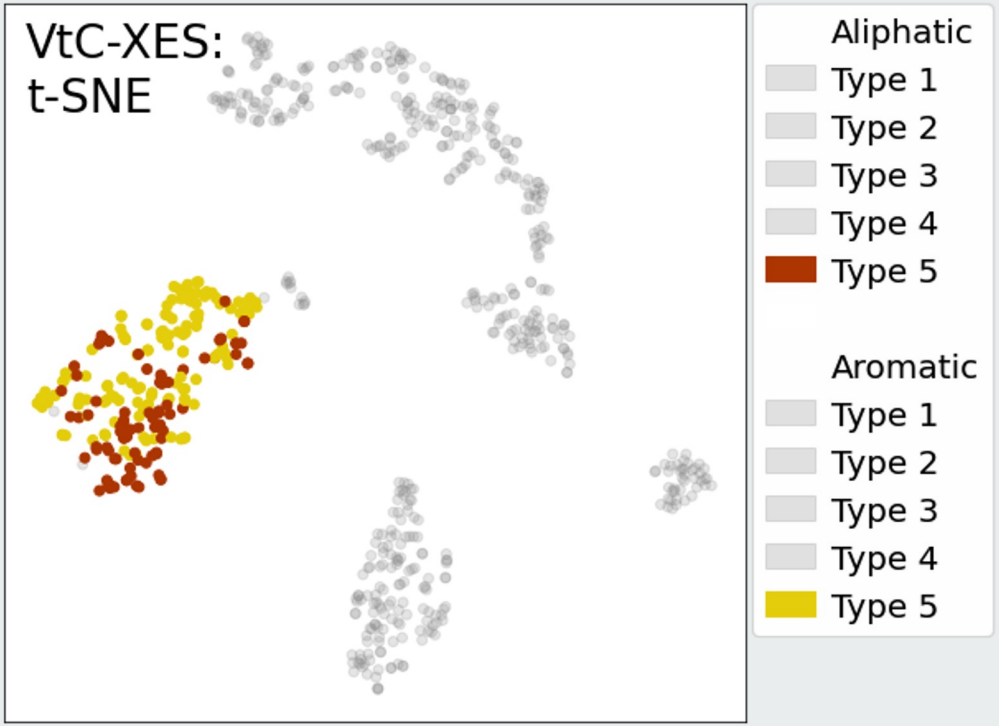
Delocalized = aromatic vs localized = aliphatic

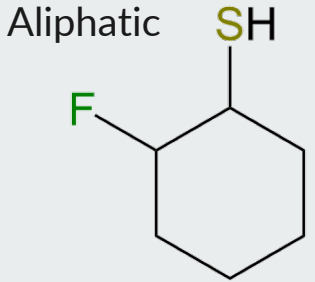


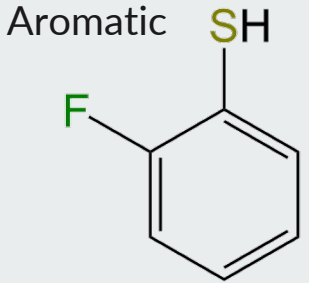


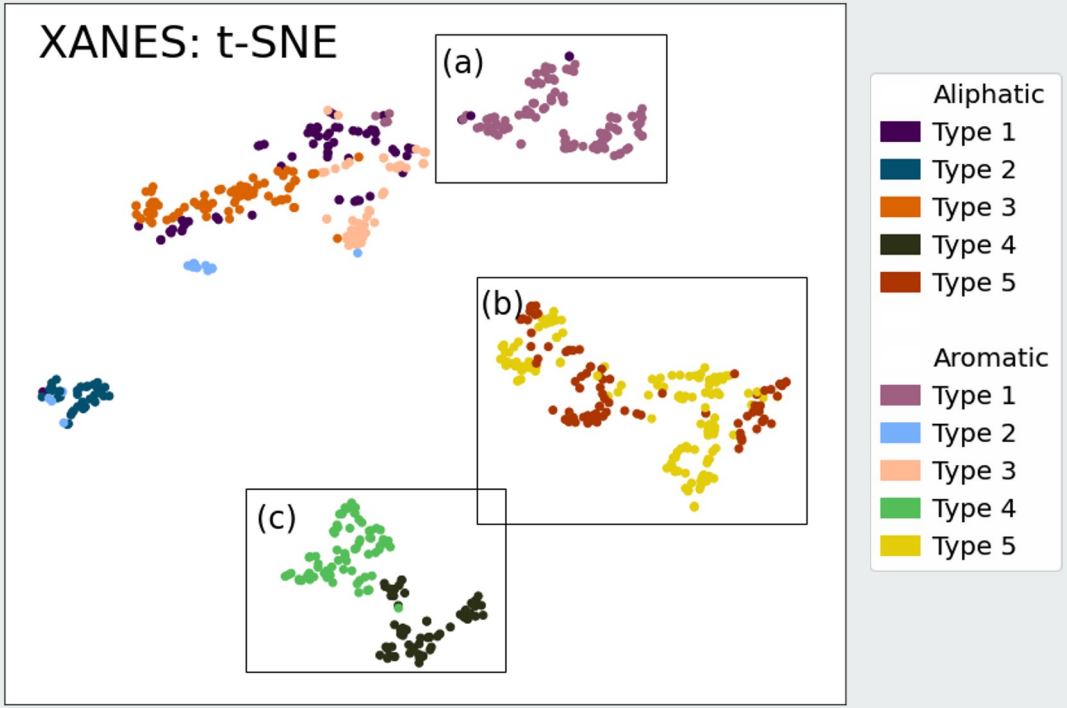


Delocalized = aromatic vs localized = aliphatic

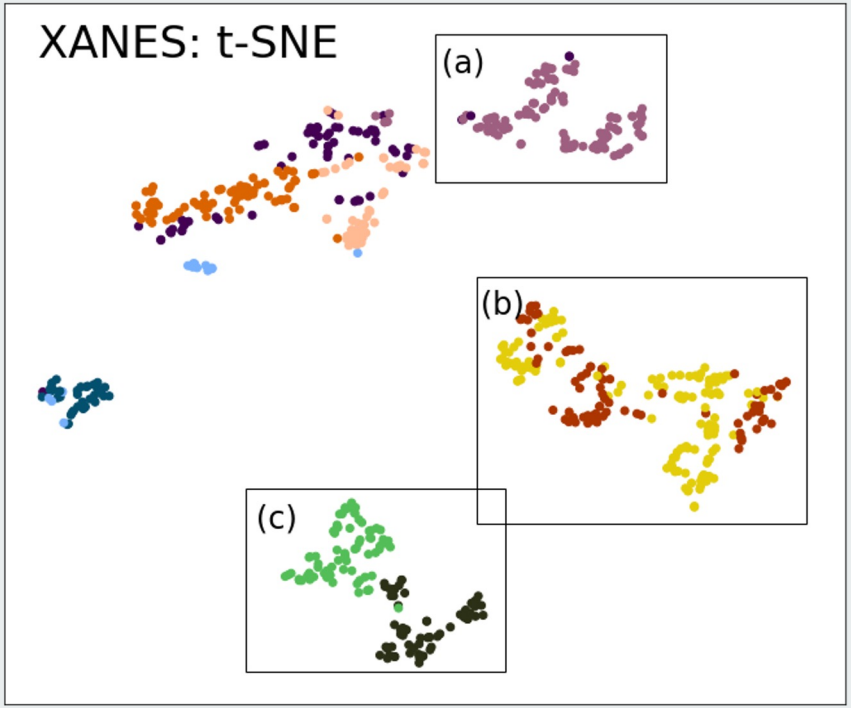


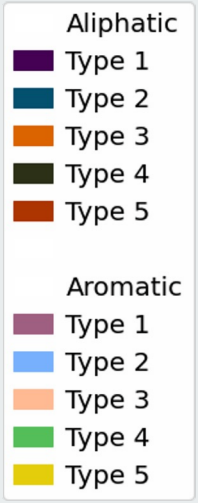


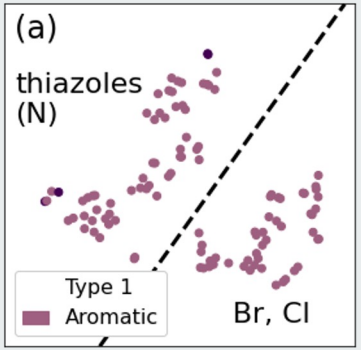




A closer look at the sub-clusters in the t-SNE of the XANES spectra...

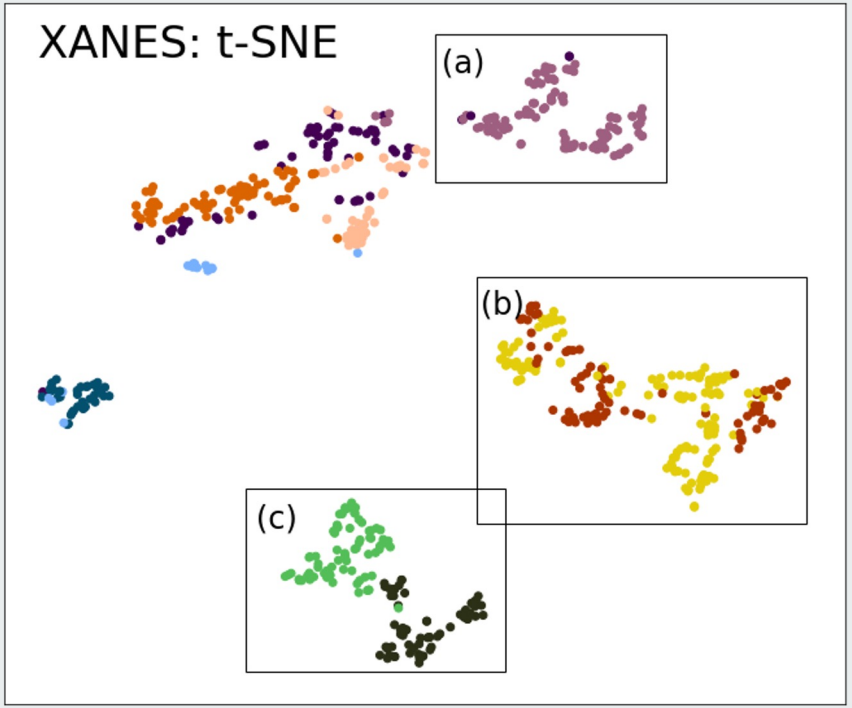


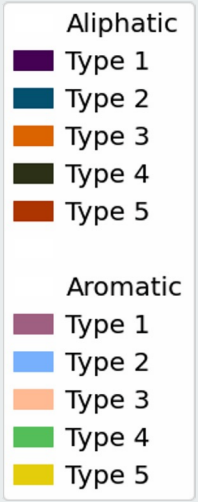


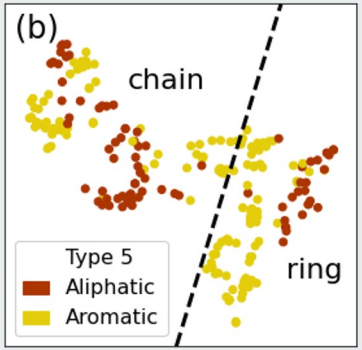


(a) Type 1: aromatic

Separated based on inclusion of Cl or Br vs N

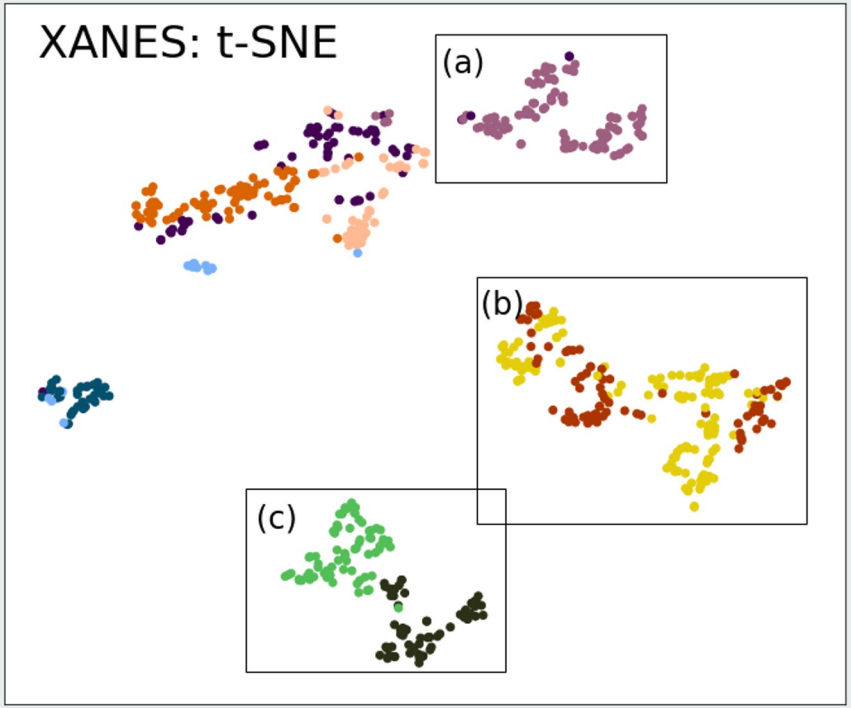


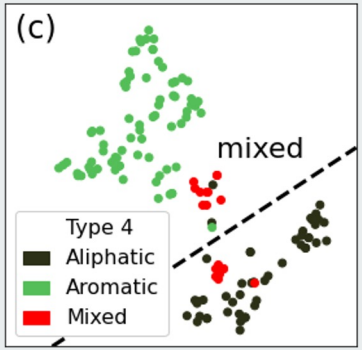




(b) Type 5: aliphatic

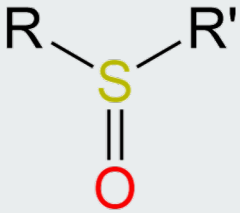
Separated based on chain vs ring





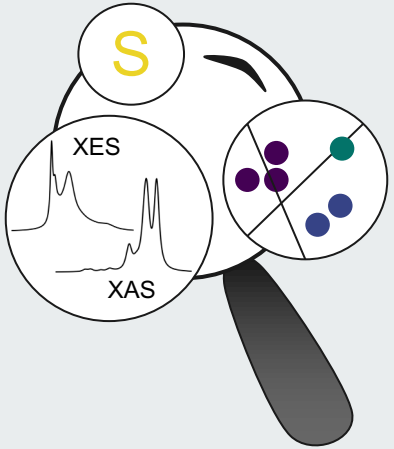
(c) Type 4: all

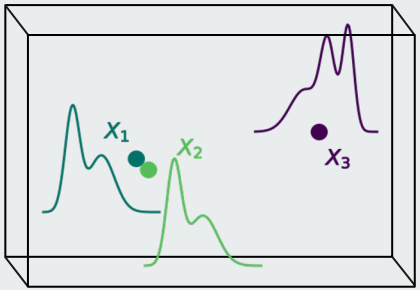
Separated by one R group aromatic and the other aliphatic
→ form bridge



Outline

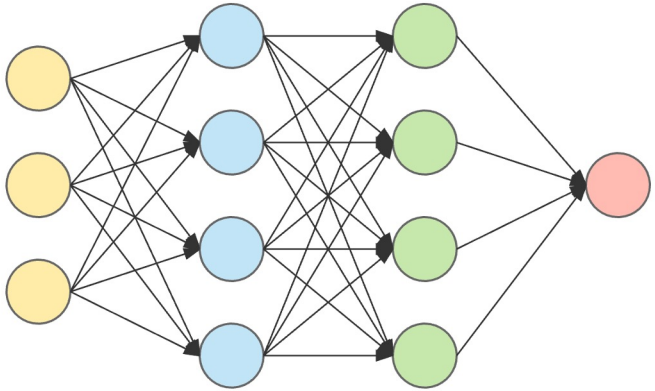
1. Background on spectroscopy methods:
   Absorption (XAS) vs Emission/ Fluorescence (XES)
2. Background on dimensionality reduction (DR) algorithms
3. Interpreting the reduced space via cluster similarities
   a. Expected & unexpected fingerprints
4. **Comparing DR algorithms using supervised ML on identifying the fingerprints found above**



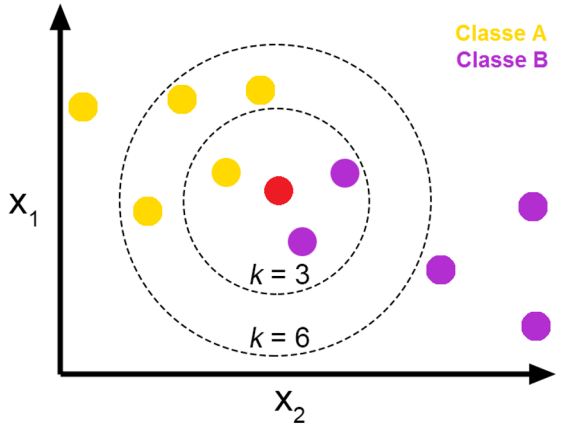


Quantitative comparison PCA, VAE, & t-SNE using supervised ML

1. Neural network trained directly on spectra



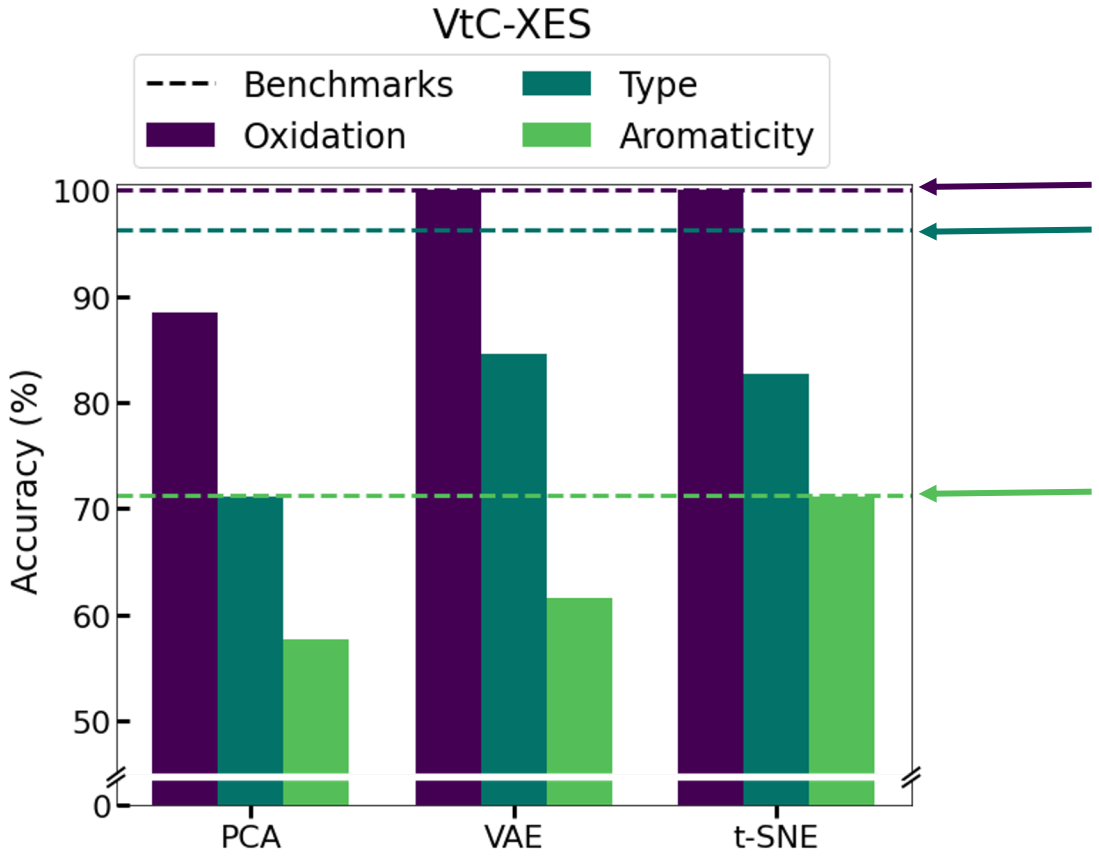
2. K-nearest neighbors (KNN) on reduced spaces



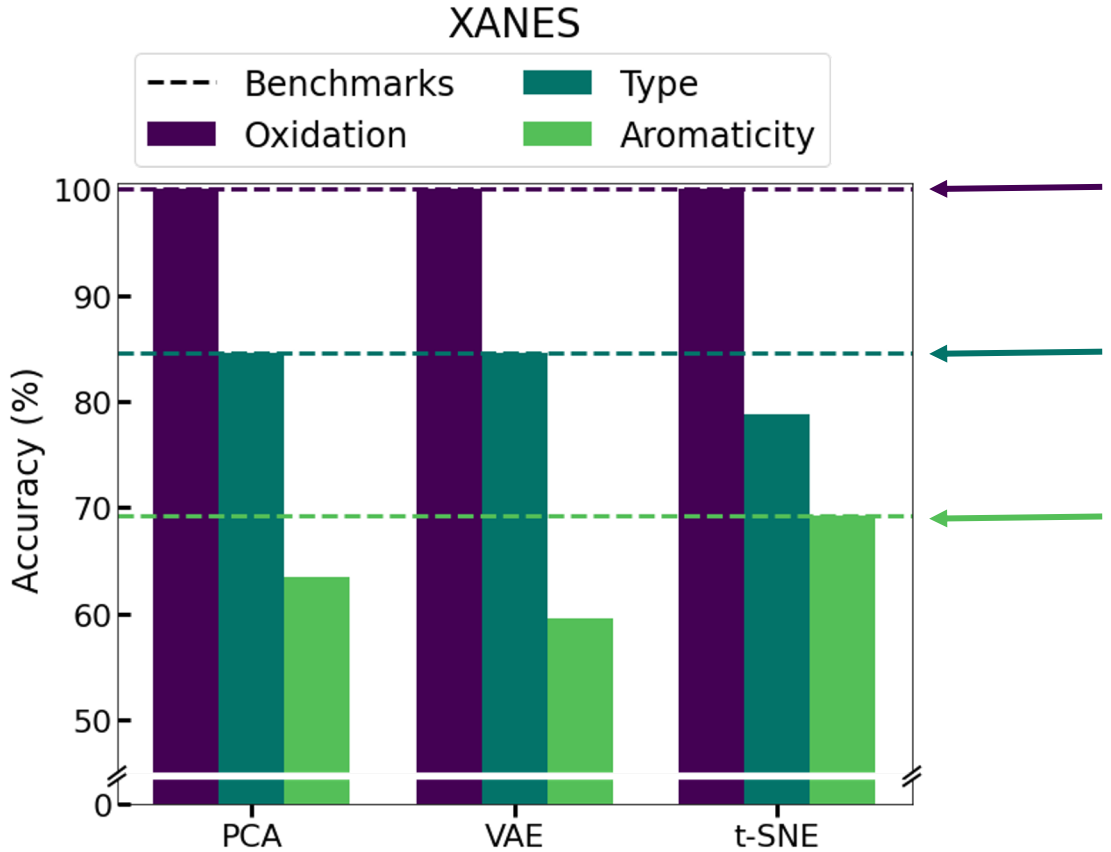
3 classification schemes

Oxidation Type Aromaticity

Quantitative comparison: PCA, VAE, vs t-SNE



Quantitative comparison: PCA, VAE, vs t-SNE



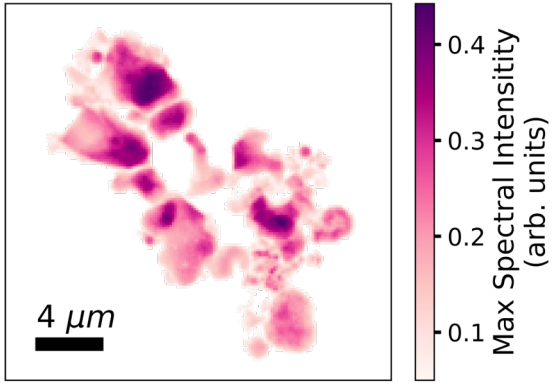
Conclusions

1. Representing spectra in just **two dimensions** captured chemical properties like oxidation, bonding environment, and the aromaticity of R groups

1. VtC-XES and XANES encoded similar information, but VtC-XES had surprising sensitivity to aromaticity

1. Reducing the constraints on the DR algorithm allowed for better clustering and thus easier interpretation and identification of chemical classes

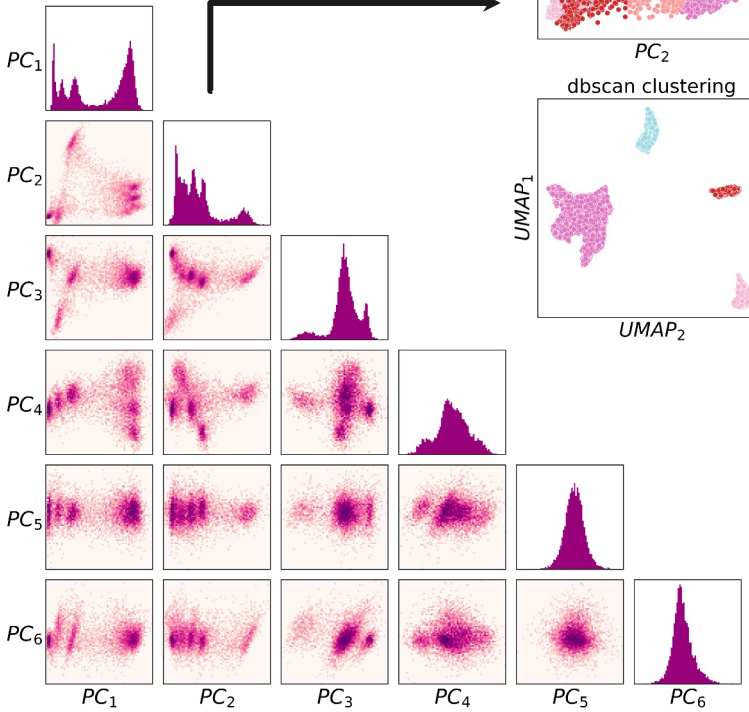
Other applications where UML seeds SML

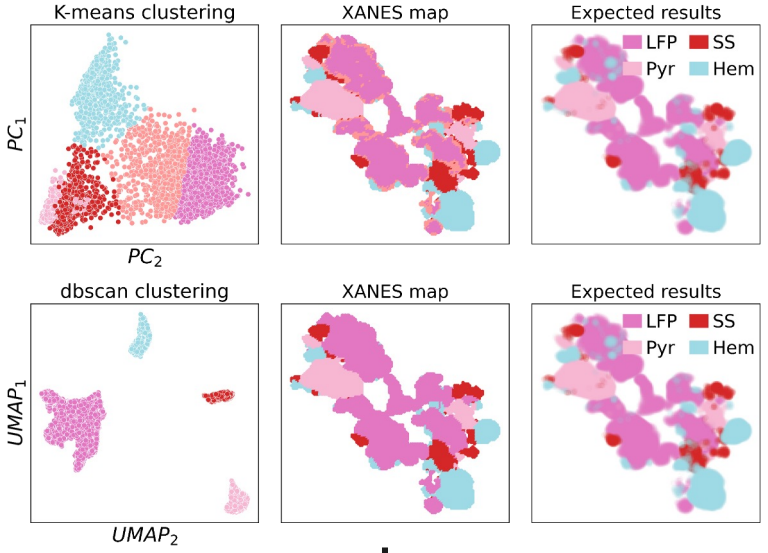


Every pixel (150 nm) had an associated XANES spectrum (25,000 total) and we want to identify spatial phases

Avoid pixel-by-pixel analysis using DR + clustering

Non-linear DR easier to interpret (and less dependent on number of dimensions kept)





Regression on each cluster