Supporting Information

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

Samantha Tetef[1], Niranjan Govind[2], Gerald Seidler[1,*]

[1]Department of Physics, University of Washington, Seattle WA 98195, USA

[2]Physical and Computational Sciences Directorate, Pacific Northwest National Laboratory, Richland, Washington 99352, USA

(*) corresponding author: seidler@uw.edu

Table of Contents

Explanation of VAE loss	2
Increasing latent space dimension	4
Hyperparameter tuning	5
FastICA, FA, and NMF	6
Figure S1: Loss plotted against number of epochs	7
Figure S2: Reconstructed VtC-XES spectra	8
Figure S3: Schemes 1 and 2 confusion matrices	9
Figure S4: Scheme 3 confusion matrix: VtC-XES	10
Figure S5: Scheme 3 confusion matrix: XANES	11
Figure S6: Dimensionally reduced spaces	12
Figure S7: Oxidation KNN: VtC-XES	13
Figure S8: Oxidation KNN: XANES	14
Figure S9: Sulfur Type KNN: VtC-XES	15
Figure S10: Sulfur Type KNN: XANES	16
Figure S11: Aromaticity KNN: VtC-XES	17
Figure S12: Aromaticity KNN: XANES	18
Figure S13: Accuracy for increasing latent dimension	19
SI References	20

Explanation of VAE loss

In Fig. S1 we show the special loss, or objective function, used in VAEs as a function of training epoch for both the training and validation data sets for the XANES and VtC-XES datasets. This special loss is defined as the mean of the reconstruction loss (binary cross entropy) and the Kullback-Leibler (KL) divergence. KL divergence ensures the VAE is fully utilizing the latent space by penalizing lost information. In general, it is given by

$$D_{KL}\left[P(z|x)||Q(z||x)\right] = E[\, logP(z|x) - logQ(z|x)\,]$$

$$\equiv \sum P(z\,|\,x) log \frac{P(z\,|\,x)}{Q(z\,|\,x)}$$

where P is the probability distribution and Q is the approximation of P. Thus, KL divergence identifies how much information it lost using the approximation Q. In a VAE objective function, z is the latent space representation of our data x, Q is the encoder, and P is the decoder. Thus, the KL divergence is

$$-\frac{1}{2}\sum 1 + log\sigma_z^2 - \mu_z^2 - \sigma_z^2$$

where log var(z) and mean(z) are the two parallel latent space layers of the VAE. Moreover, KL divergence encourages the latent space to be centered around zero with normal variance and is therefore regularized. For an in depth derivation of VAE objective function, see Rocchetto et al. [1]

A plot such as Fig. S1 is a useful heuristic for understanding training convergence and for evaluating the degree of overfitting or underfitting. To be specific, starting with the XANES, the losses plateau at about 20 epochs and the validation loss does not increase. This indicates that the resulting neural network is generalizable and is not overfitting and thus is likely to have high utility, i.e., it has not overfit such that it cannot address spectra outside the training data set but also enough detail has been encoded that most useful information has likely been incorporated. The VtC-XES shows a similar plateau in the VAE losses, which indicates this model is not overfitting as well.

Increasing latent space dimension

Much as with PCA, where one must wisely choose the number of PCs to get a good representation of the training data set, we are also free to modify the dimension of the latent space for the VAE. This is investigated in Fig. S2 where representative XES spectra (one from each type) are compared to the corresponding decoded spectra as a function of the dimension of the latent space, starting with two dimensions on the left and proceeding to 50 dimensions at the right. Increasing the latent space dimension up to 50 dimensions does not drastically change the accuracy of the decoded spectra, as the most distinct features are obtained just from a two-dimensional latent space. Hence, for the VTC-XES for this broad collection of sulphorganics, a two-dimensional representation VAE is enough to capture the most distinct spectral features, giving a dramatically effective encoding and dimensionality reduction.

Hyperparameter tuning

Hyperparameters for all machine learning methods were selected using multiple validation sets separated from within the entire training set. The VAE was limited to one or two hidden layers for each the encoder and decoder with layer dimension sizes constrained to powers of two between 32 and 1024. The ANN classifier was also limited to one or two hidden layers with dimensions constrained to powers of two between 32 and 1024. Dropout was also constrained to be between 5% and 20% and implemented to encourage generalizability. The t-SNE perplexity value was selected from values between 5 and 50, where the smallest perplexity value was chosen such that (a) there did not appear to be spurious or artificial clusters and (b) yielded consistent embeddings upon recalculation, indicating a global minimum was reached. The k nearest neighbor hyperparameter for KNN was then selected from a neighborhood of values around the t-SNE perplexity value, since both approximately represent cluster or group size. This was determined to be between 10 and 30. All other hyperparameters not specified in the manuscript were set to default values.

FastICA, FA, and NMF

The three supplemental linear dimensionality reduction methods included in this study are Fast Independent Component Analysis (FastICA)[2], Factor Analysis (FA)[3, 4], and Non-negative Matrix Factorization (NMF)[5]. FastICA is an implementation of independent component analysis, which is a generalization of PCA. Often, independent component analysis is used to separate out independent signals, or components, contributing to data. However, because its aim is to calculated independent components, there is not a clear statistical method to reduce dimension, such as maximizing the explained variance as with PCA. Factor analysis (FA) is similar to PCA as well, except it calculates the eigenvalue decomposition on the reduced correlation matrix instead of the full correlation matrix. This analysis on the reduced correlation matrix helps identify latent, or hidden, features in the data, i.e., variables that cause correlated features in the original dataset. However, it has been shown that if the number of included datapoints is large enough (about 40), PCA and FA have similar results [6]. Finally, non-negative matrix factorization (NMF) is another linear dimensionality reduction algorithm that assumes the data is (as the name suggests) non-negative, which is true for both XANES and XES spectra. NMF calculates two non-negative matrices whose product reproduces the original dataset. This encourages factors to be positive and thus more physically interpretable.

Figure S1: Loss plotted against number of epochs

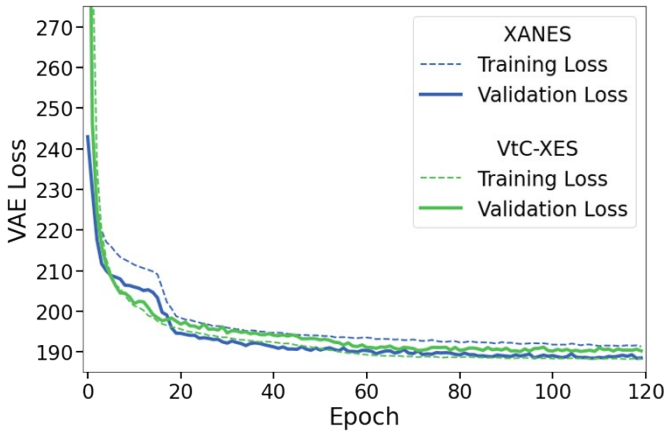

Fig. S1. Loss plotted against number of epochs for the VAE model for both the XANES data (blue) and the VtC-XES data (green).

Figure S2: Reconstructed VtC-XES spectra

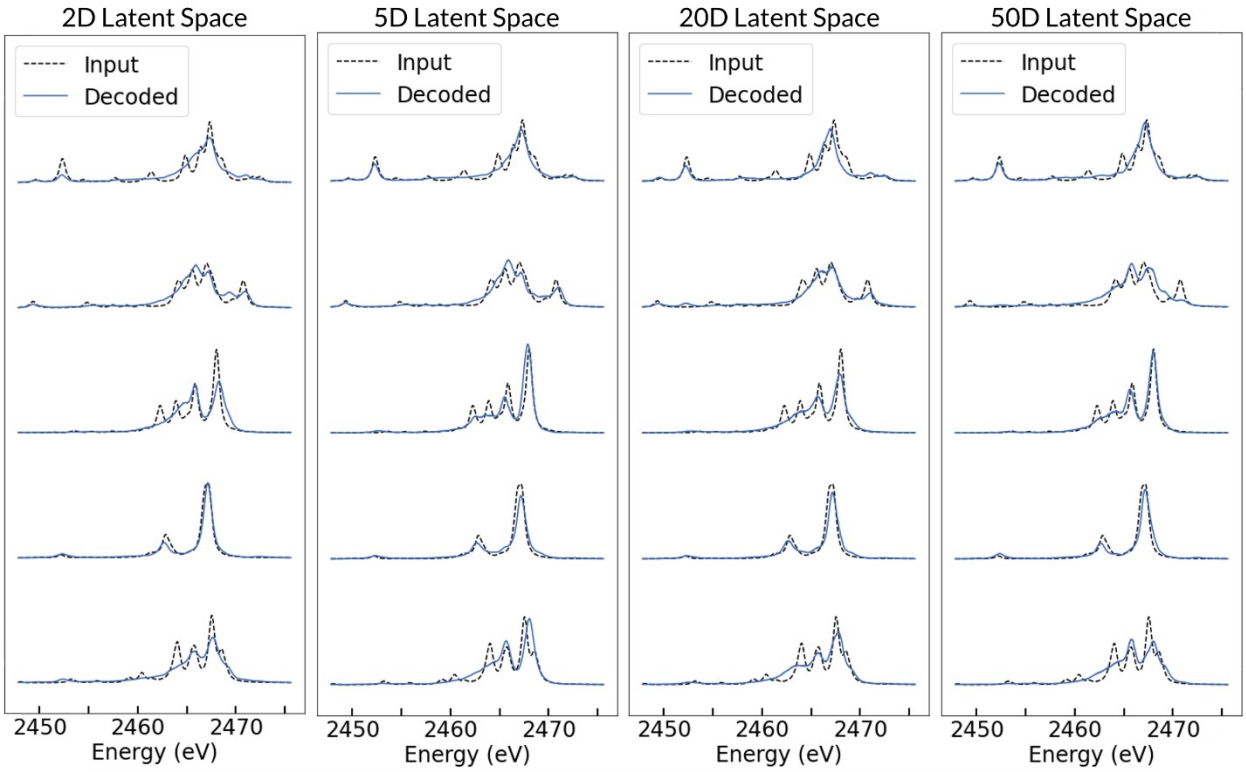

Fig. S2. Reconstructed VtC-XES spectra with increasing latent space dimension.

Figure S3: Schemes 1 and 2 confusion matrices

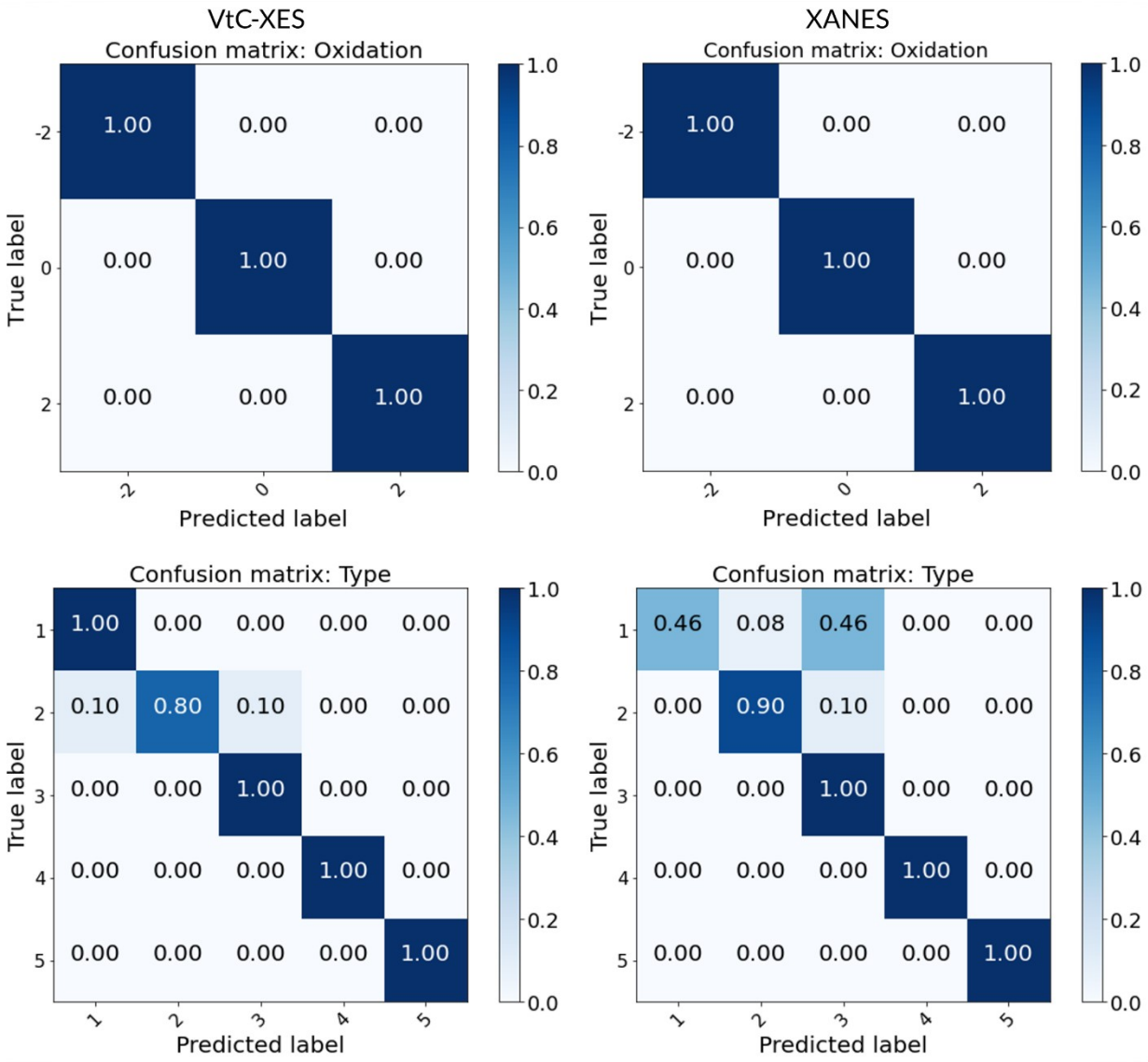

Fig. S3. Classification via NN: confusion matrices for XES and XANES for both categorization schemes: 1) oxidation and 2) bond type.

Figure S4: Scheme 3 confusion matrix: VtC-XES

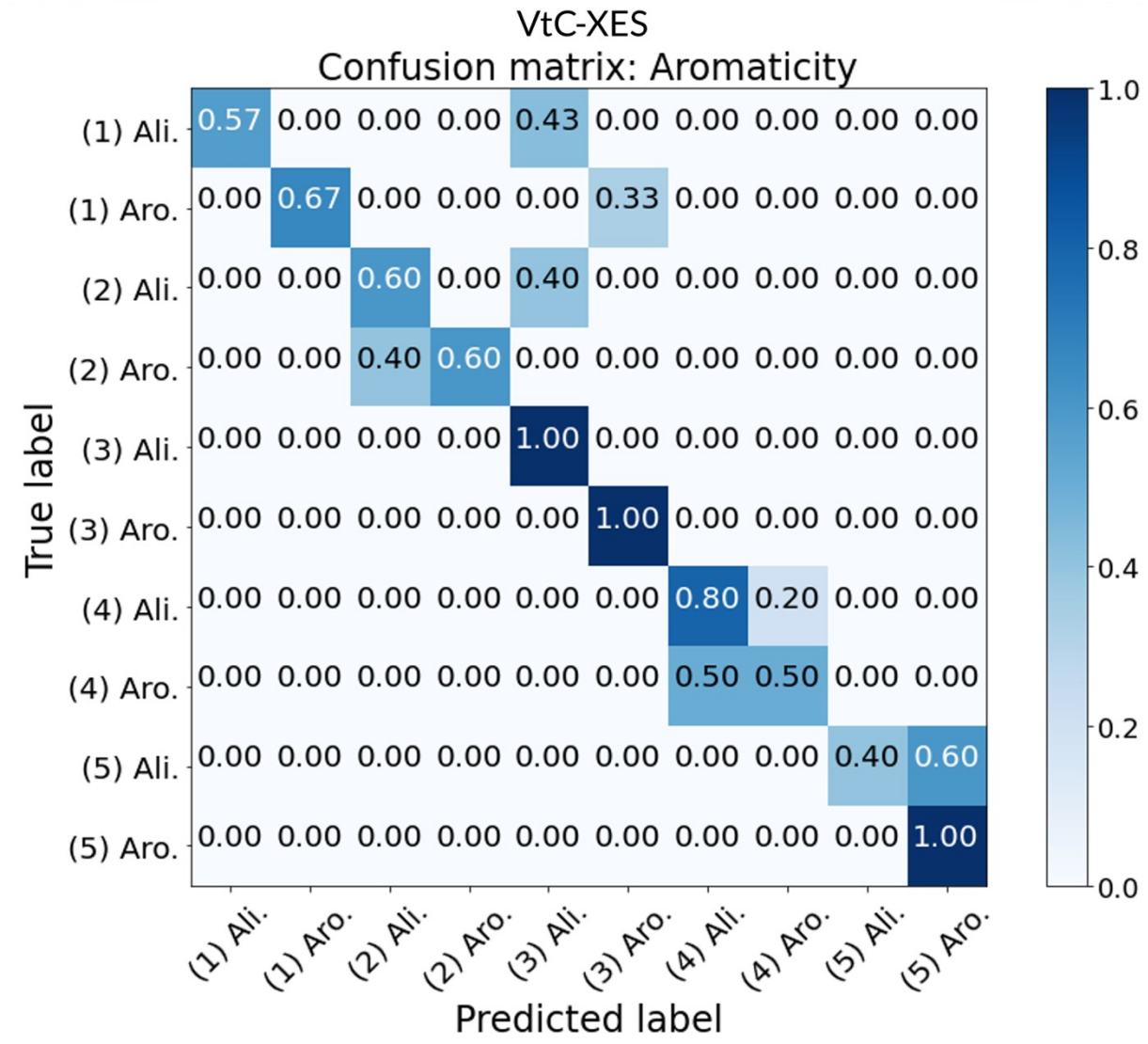

Fig. S4. Classification via NN: confusion matrices for VtC-XES for classification of aromatic versus aliphatic compounds within Types 1 to 5.

Figure S5: Scheme 3 confusion matrix: XANES

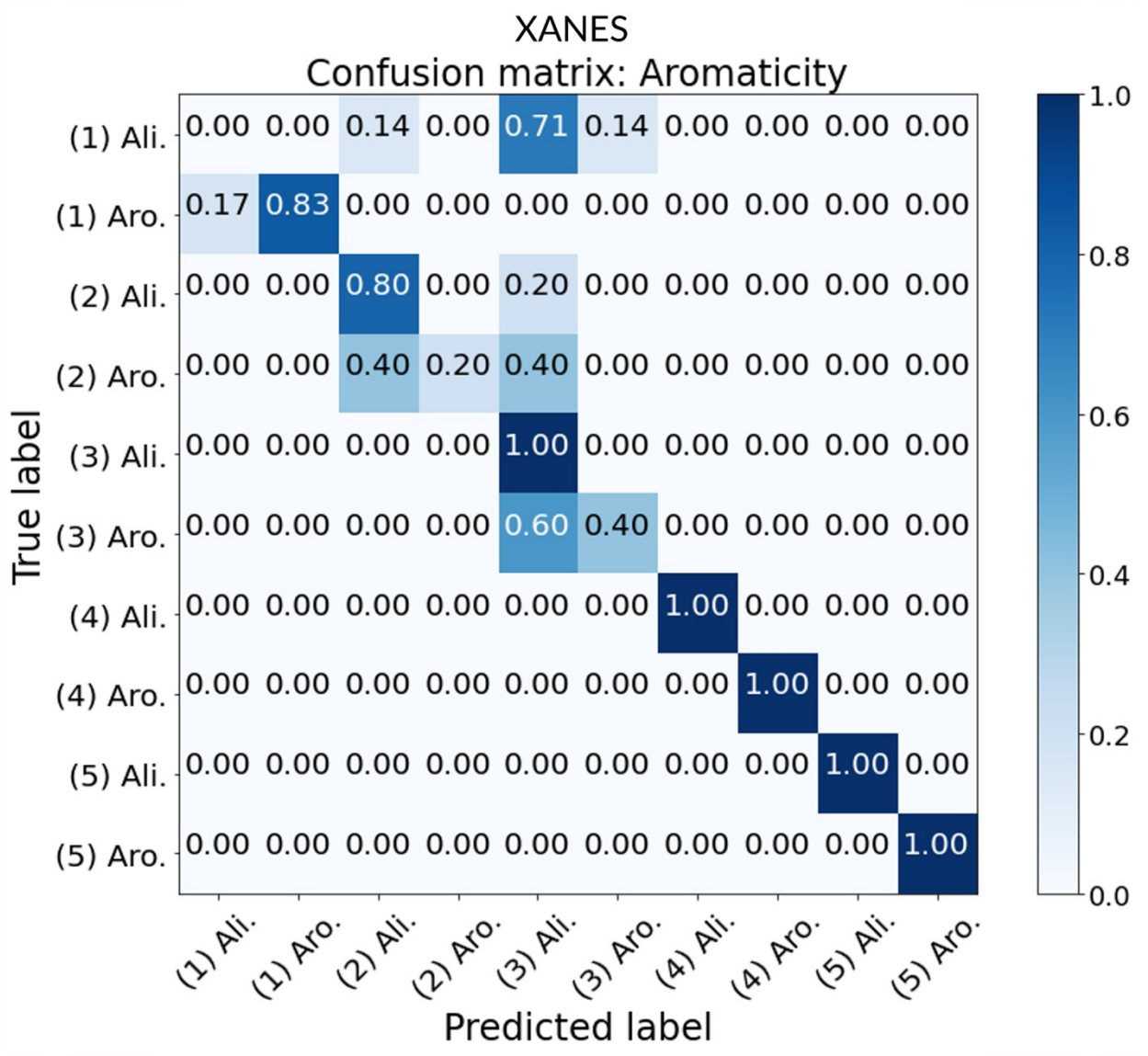

Fig. S5. Classification via NN: confusion matrices for XANES for classification of aromatic versus aliphatic compounds within Types 1 to 5.

Figure S6: Dimensionally reduced spaces

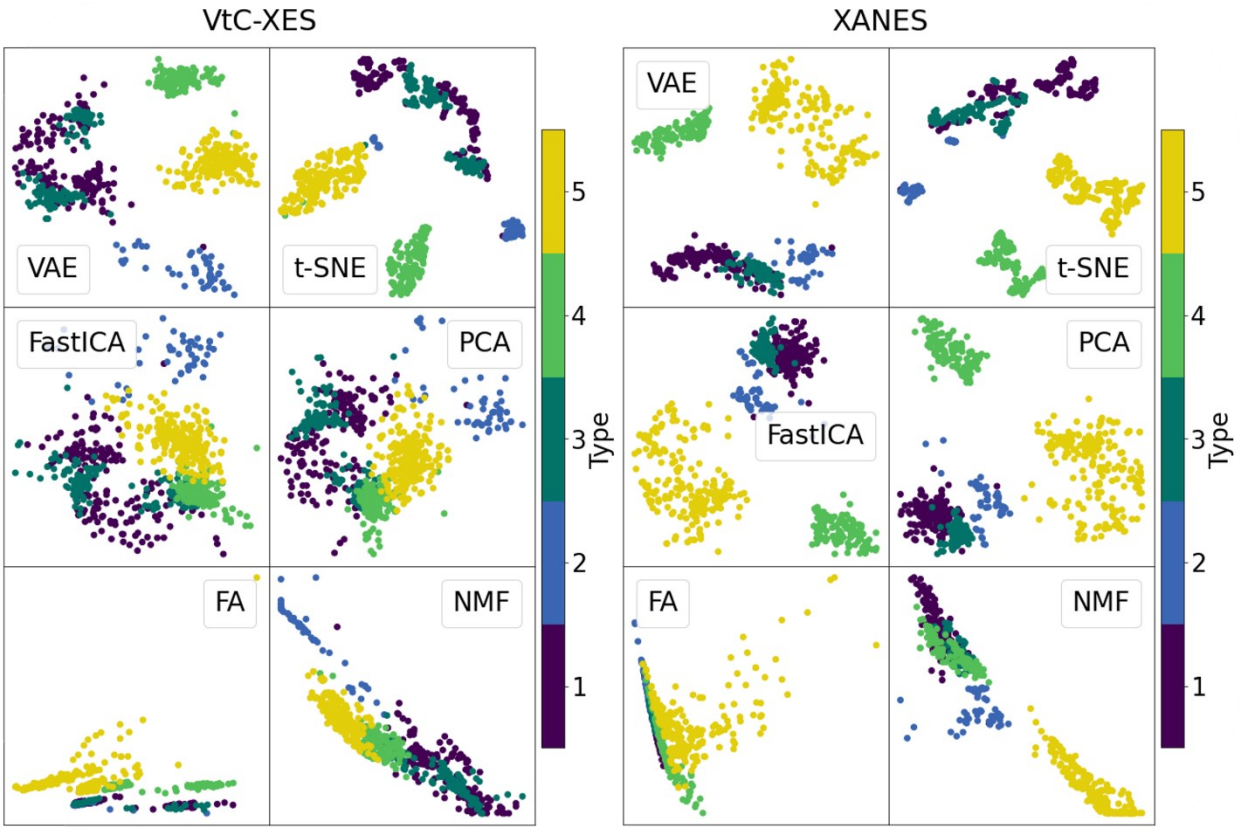

Fig. S6. Unsupervised dimension reduction: VAE, t-SNE, FastICA, PCA, FA, and NMF for VtC-XES (left) and XANES (right), color-coded by sulfur bonding Type.

Figure S7: Oxidation KNN: VtC-XES

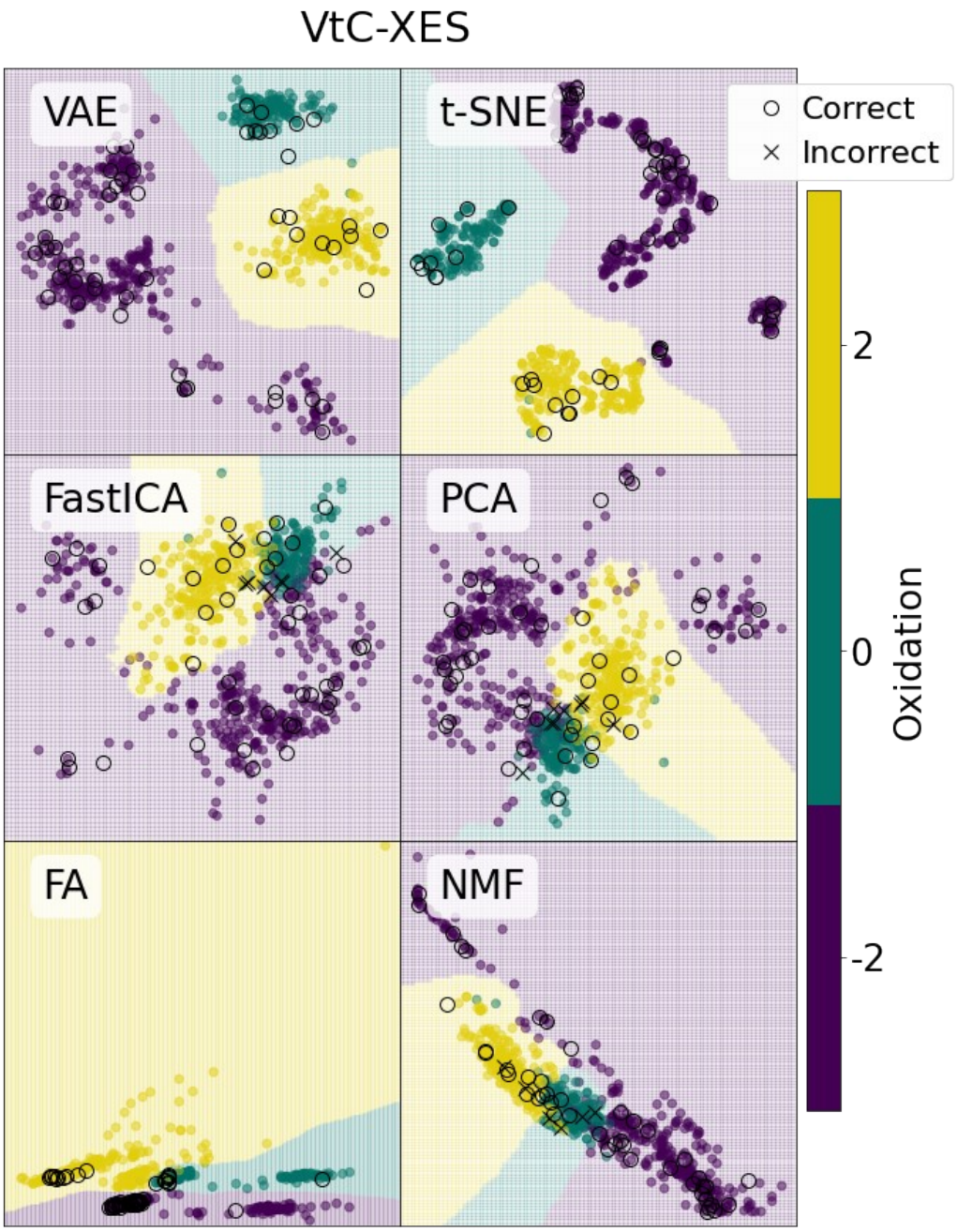

Fig. S7. KNN classification for Oxidation for VtC-XES.

Figure S8: Oxidation KNN: XANES

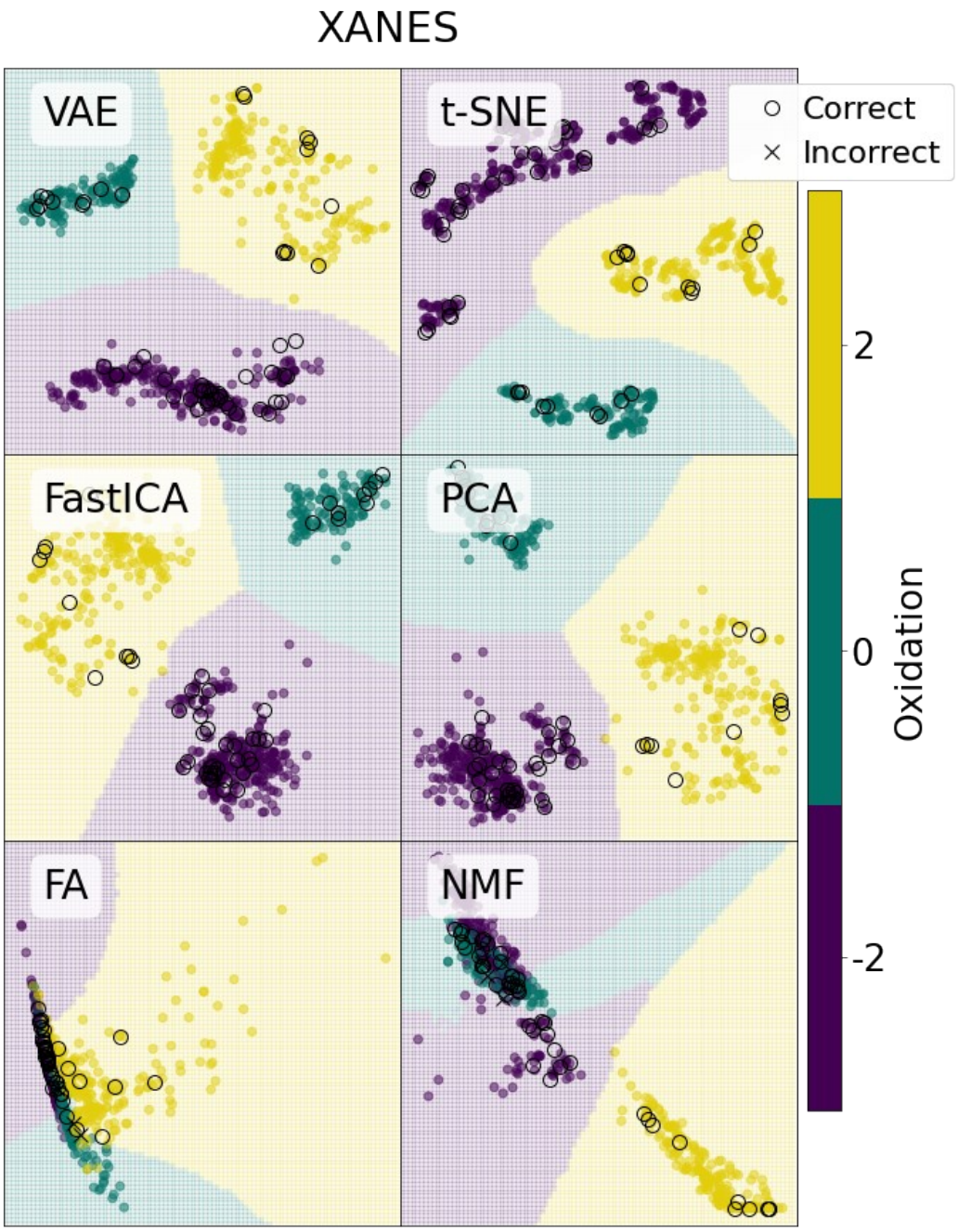

Fig. S8. KNN classification for Oxidation for XANES.

Figure S9: Sulfur Type KNN: VtC-XES

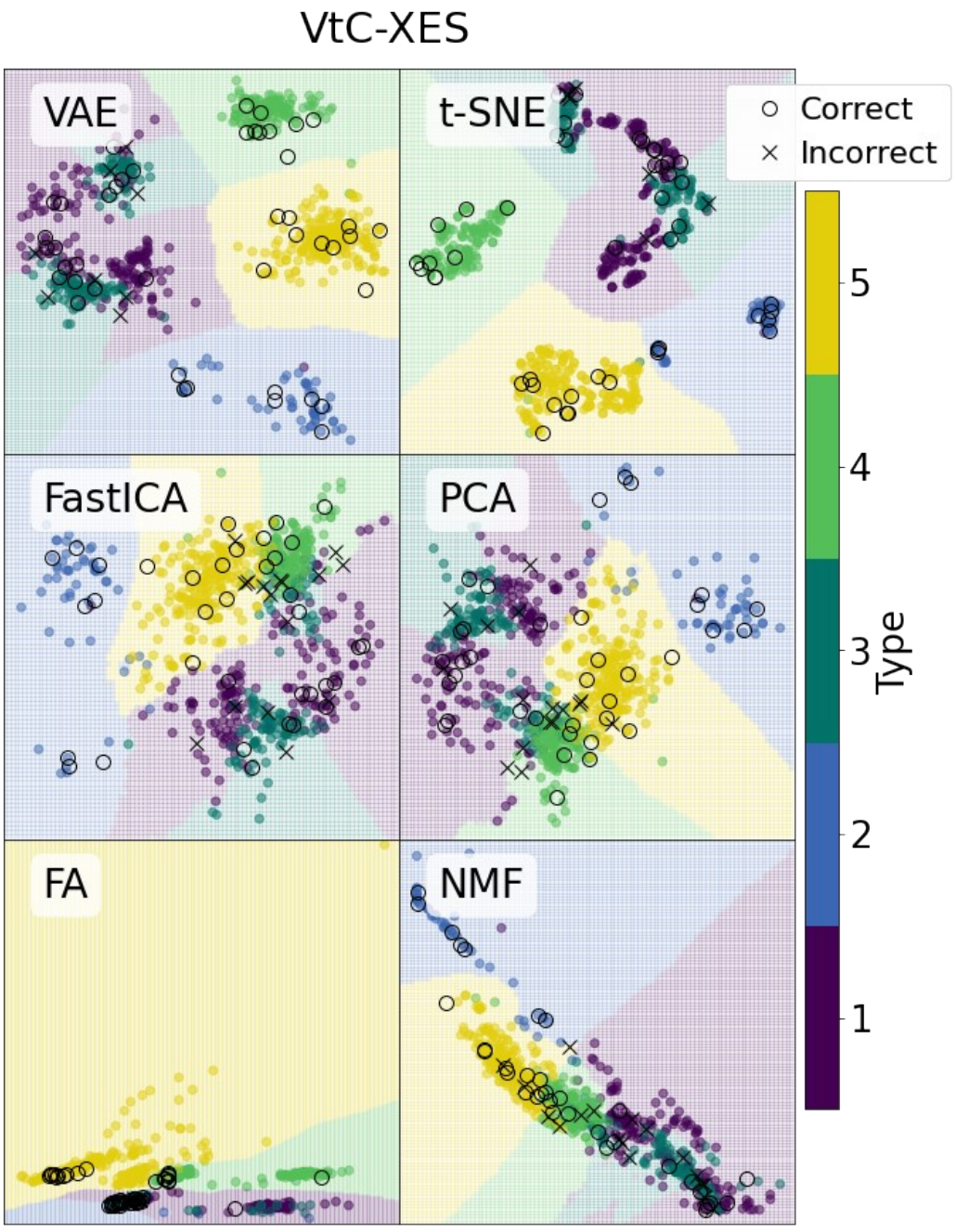

Fig. S9. KNN classification for sulfur bond Type on VAE, t-SNE, FastICA, PCA, FA, and NMF for VtC-XES.

Figure S10: Sulfur Type KNN: XANES

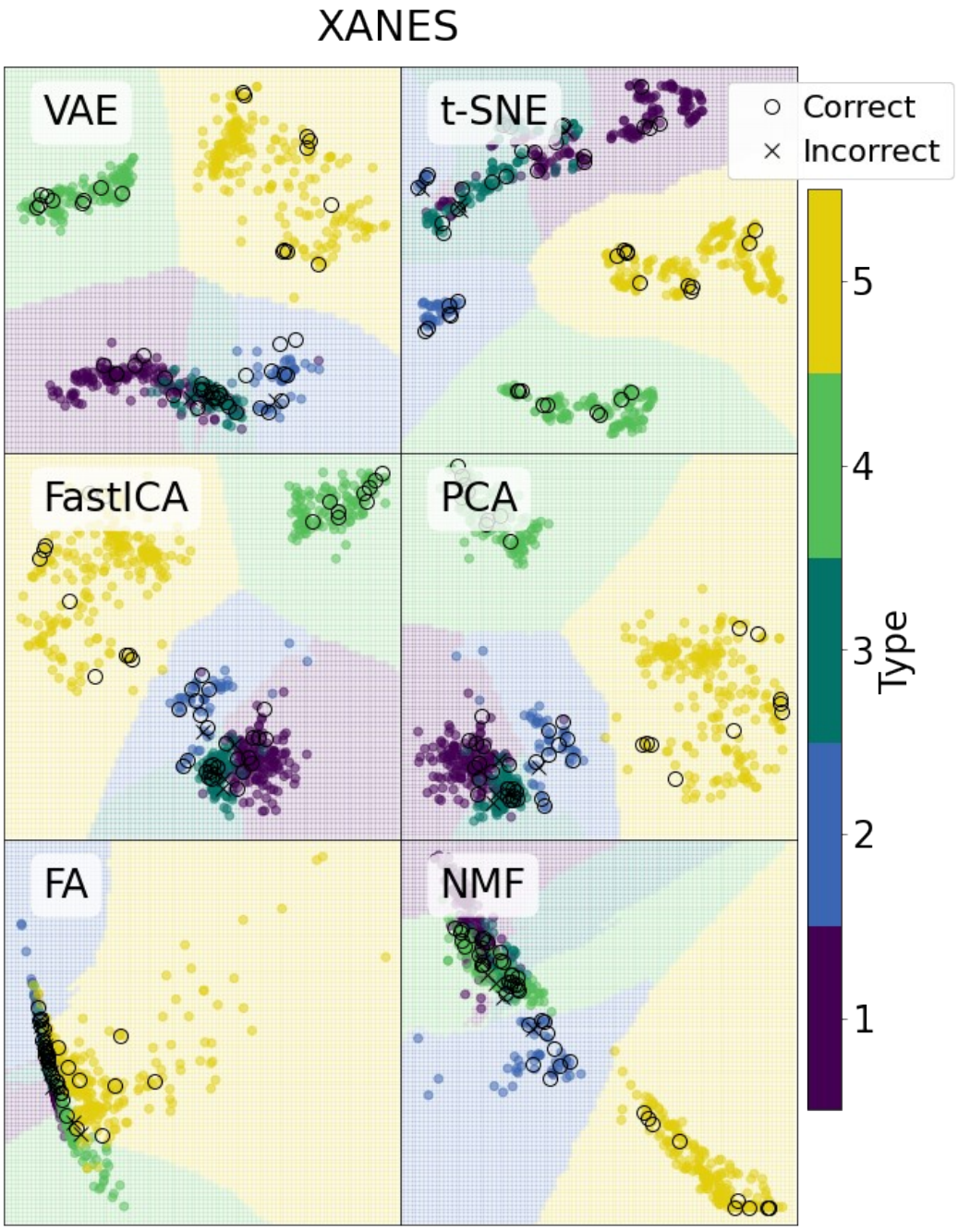

Fig. S10. KNN classification for sulfur bond Type on VAE, t-SNE, FastICA, PCA, FA, and NMF for XANES.

Figure S11: Aromaticity KNN: VtC-XES

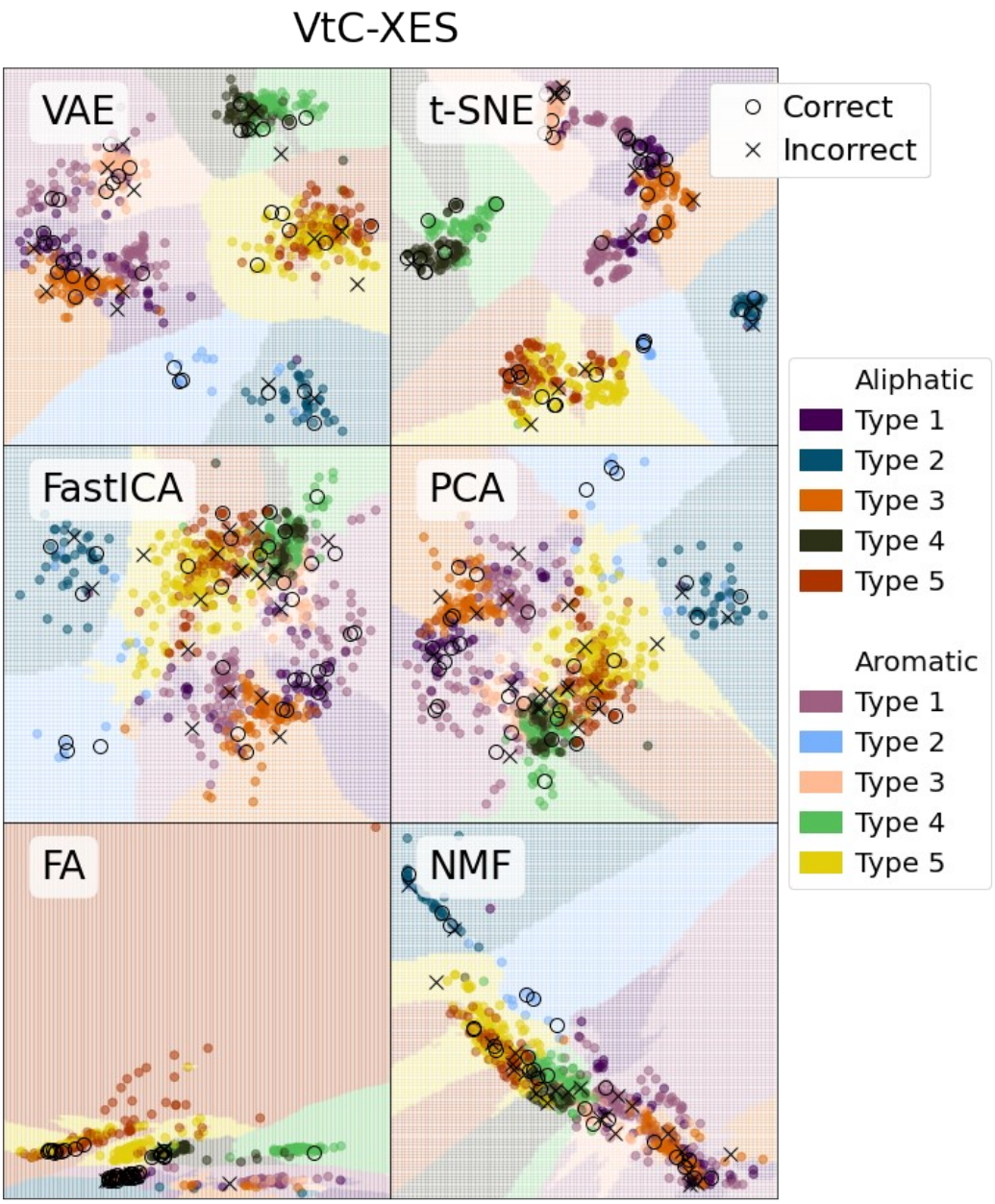

Fig. S11. KNN classification for Aromaticity for VtC-XES.

Figure S12: Aromaticity KNN: XANES

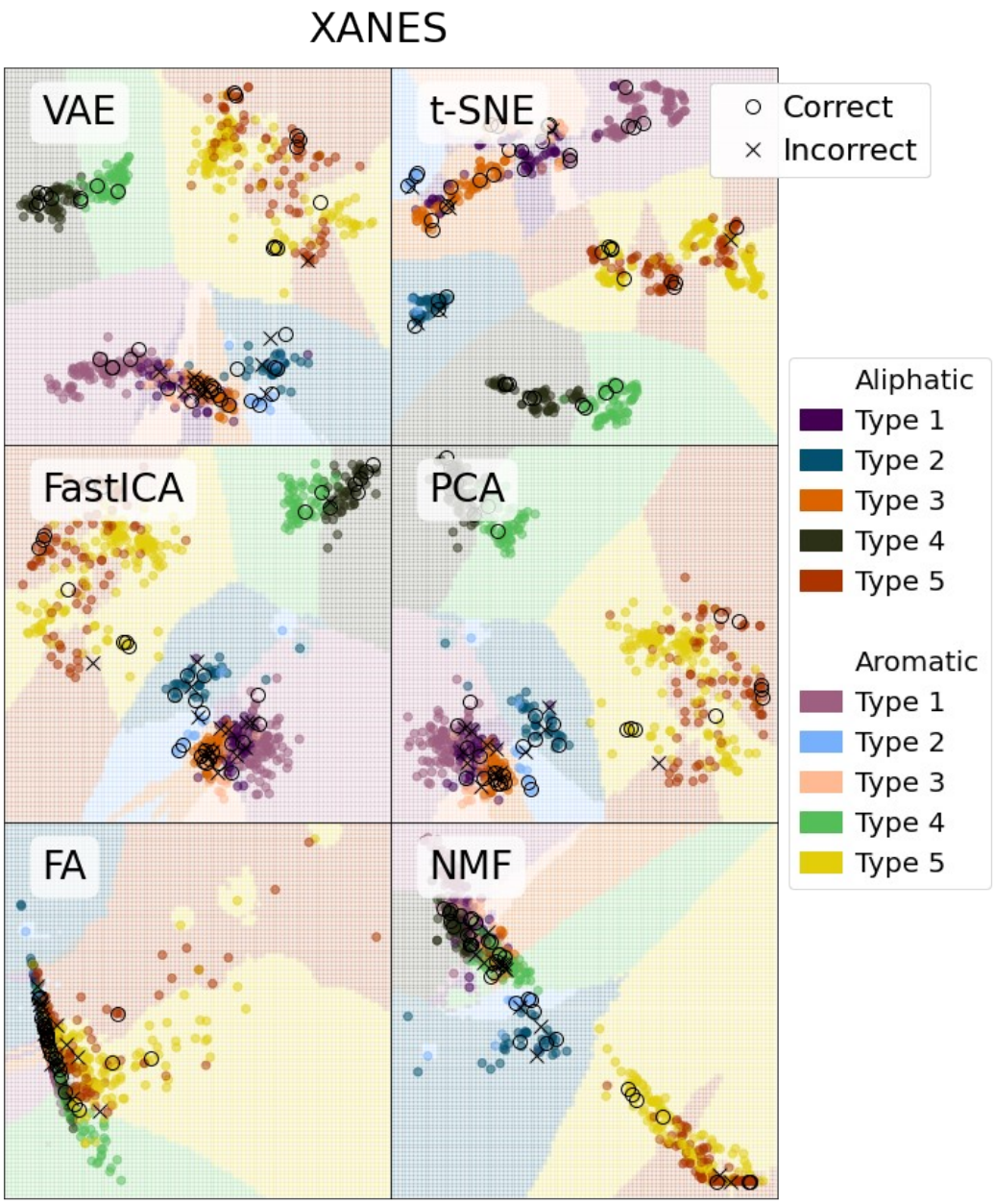

Fig. S12. KNN classification for Aromaticity for XANES.

Figure S13: Accuracy for increasing latent dimension

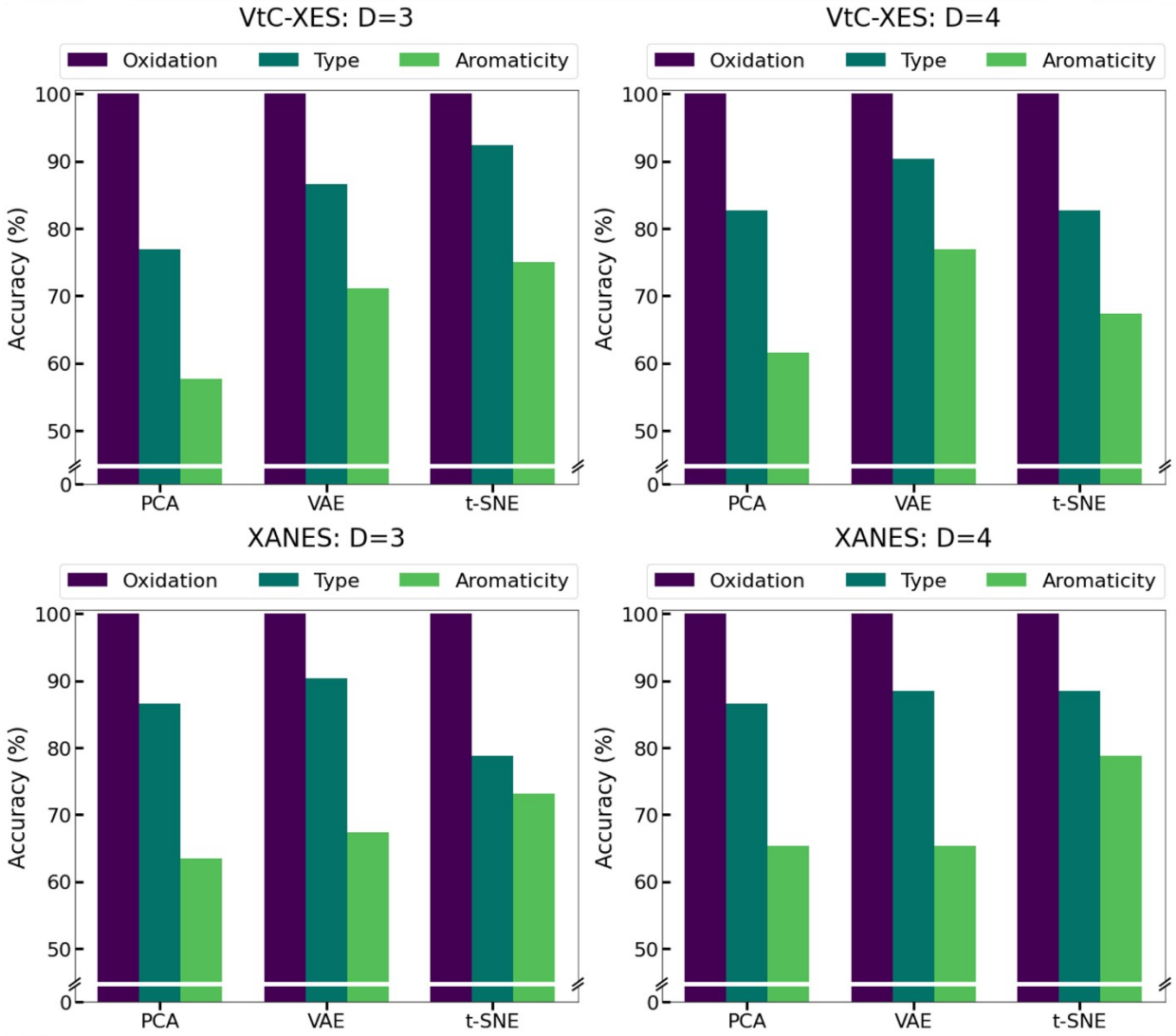

Fig. S13. Accuracy of KNN classification schemes on the PCA, VAE, and t-SNE reduced spaces for VtC-XES (top) and XANES (bottom) while increasing the latent or embedding dimension, D.

SI References

1. A. Rocchetto, E. Grant, S. Strelchuk, G. Carleo and S. Severini, *npj Quantum Information*, 2018, **4**, 28.
2. A. Hyvärinen and E. Oja, *Neural Networks*, 2000, **13**, 411-430.
3. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
4. D. Barber, *Bayesian Reasoning and Machine Learning*, Cambridge University Press, 2012.
5. D. D. Lee and H. S. Seung, *Nature*, 1999, **401**, 788-791.
6. S. C. Snook and R. L. Gorsuch, *Psychological Bulletin*, 1989, **106**, 148–154.