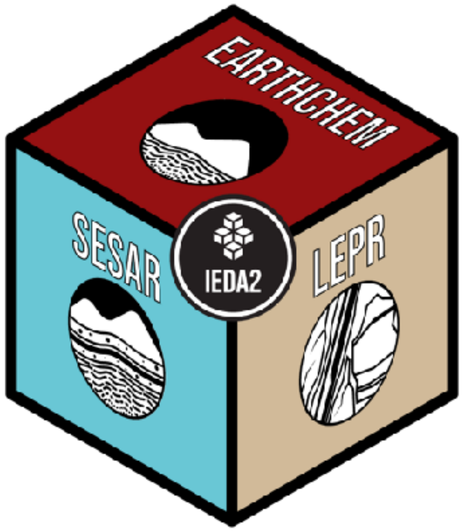
MIN-ML: A Python Package for Exploring Mineral Relations and Classifying Common Igneous Minerals with Machine Learning

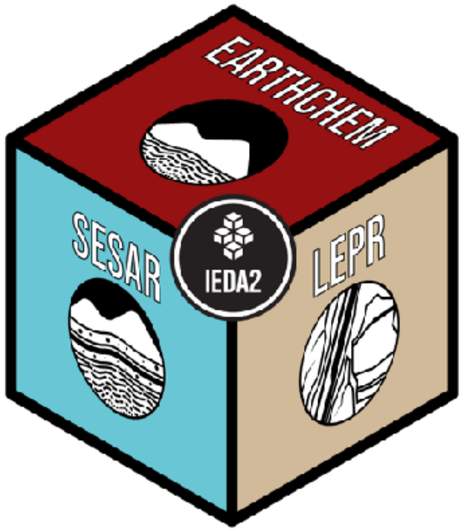
Sarah Shi, Penny Wieser, Norbert Toth, Kerstin Lehnert, Lucia Profeta

MicROCKScopica

Probing igneous processes with predictive machine learning models trained on large mineral databases



Probing igneous processes with predictive machine learning models trained on large mineral databases



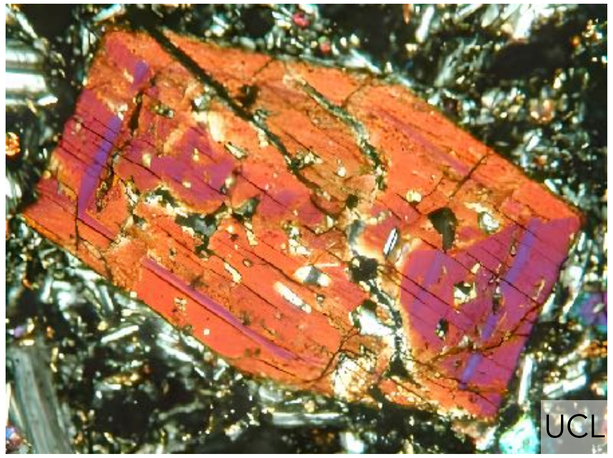
Clinopyroxene Barometry (LEPR)

Petrelli et al., 2021

Jorgenson et al., 2022

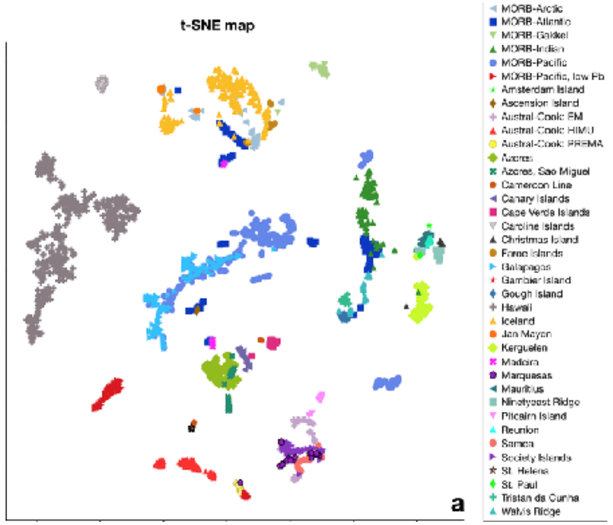
Neave and Putirka, 2017

Putirka, 2008

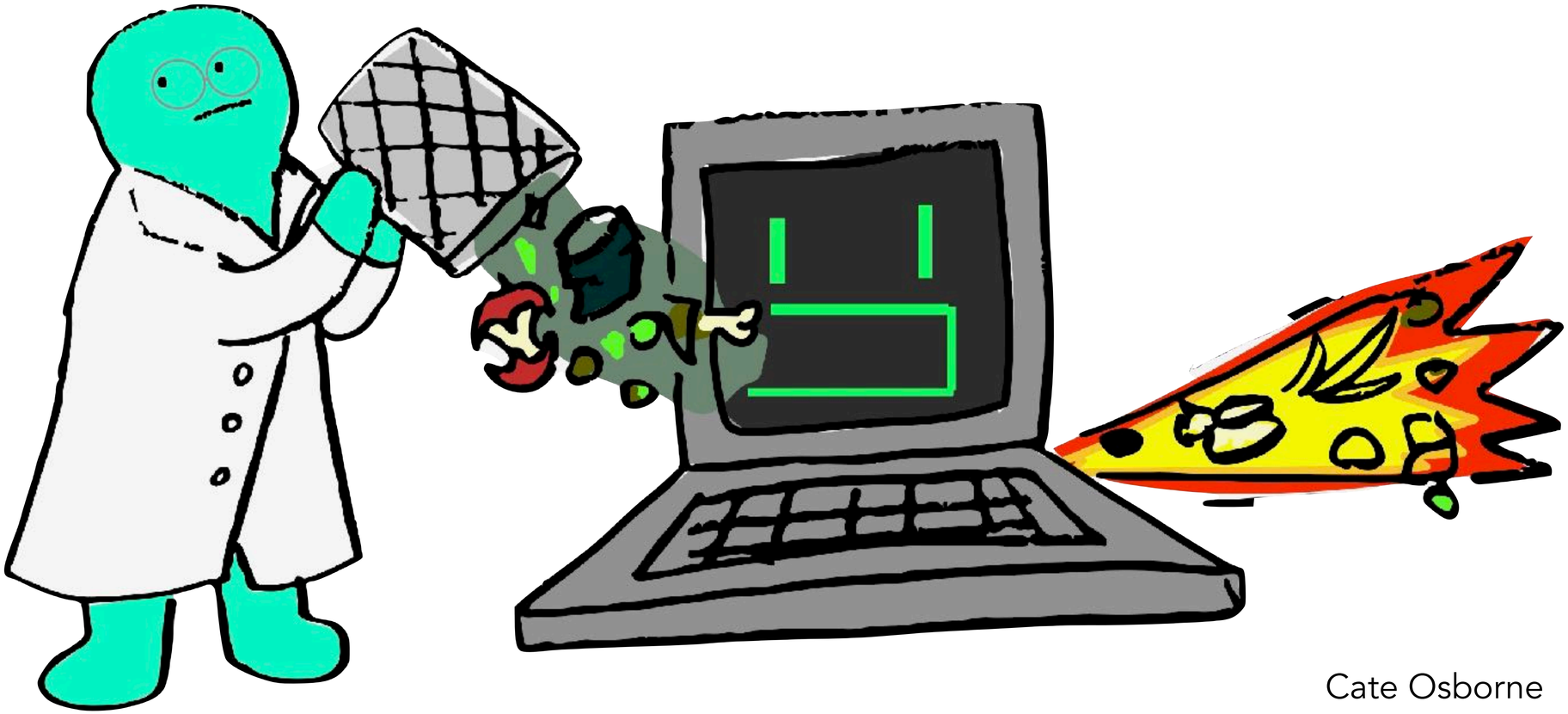


Radiogenic Isotope Heterogeneity in MORB-OIB (GEOROC)

Stracke et al., 2022



`Garbage in, garbage out'



Cate Osborne

`On two occasions I have been asked, "Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?"… I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question.'

Charles Babbage, 1864

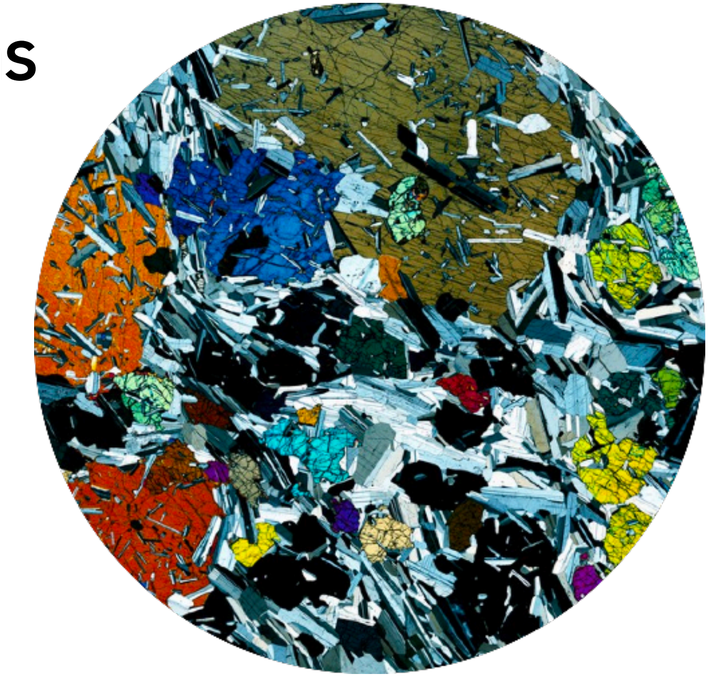
Motivating Question and Scientific Applications

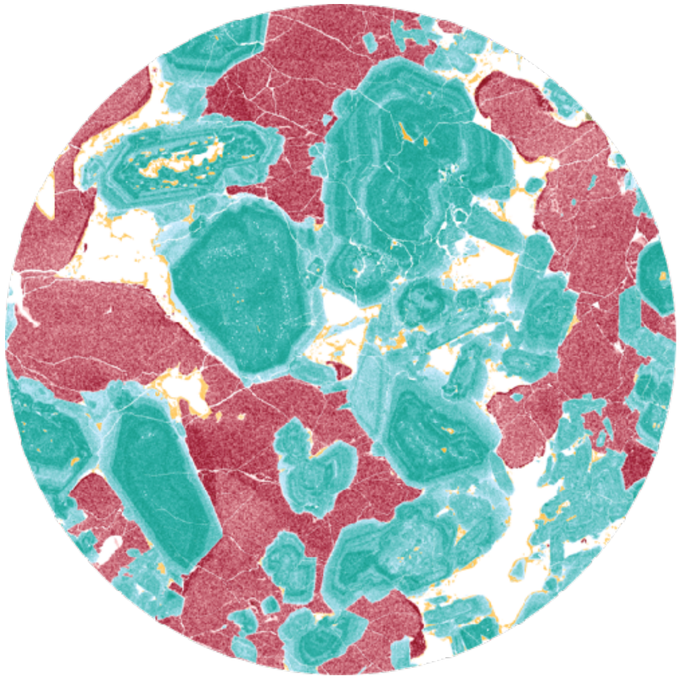
- How can we identify and classify mineral phases from their chemical compositions from electron probe microanalysis (EPMA)?

- Supervised machine learning (provide mineral phase labels from databases and assign labels to unseen data)
- Unsupervised machine learning (do not provide mineral phase labels from databases and find clusters within unseen data)

- How can we use machine learning to perform data validation?

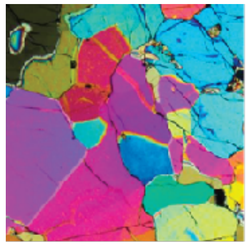
- Curation of databases
- Analytical classifications
  - EPMA - Indeterminate analysis of oxide or mineral
  - EDS - Quantitative mapping of minerals on μm-scale
- Thermobarometry - Misclassified minerals strongly impact calibrations (e.g. clinopyroxene-amphibole confusion in barometry)

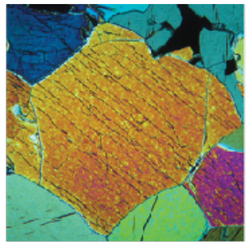


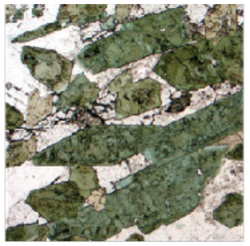


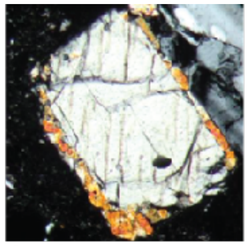
MIN-ML Datasets

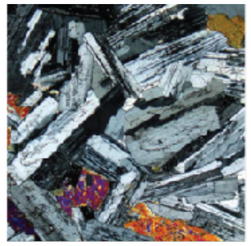
TRAIN (fit model)+VALIDATE (evaluate model while training), 86k

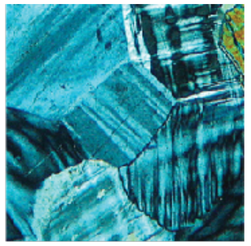

Olivine
21k

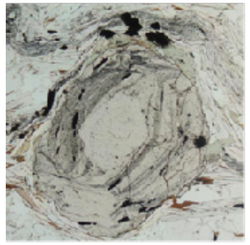

Clinopyroxene
10k

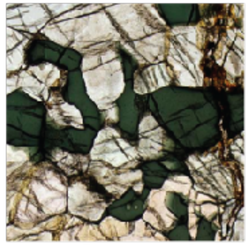

Amphibole
3.2k

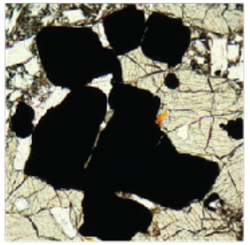

Orthopyroxene
4.5k

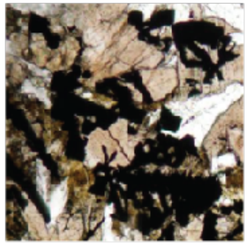

Plagioclase
17k

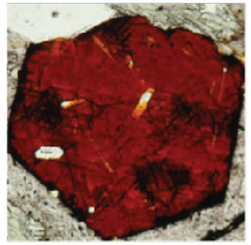

K-Feldspar
5.8k

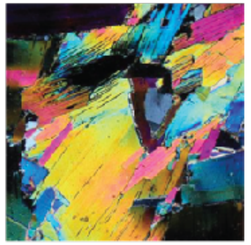

Garnet
2.0k


Spinel
1.2k

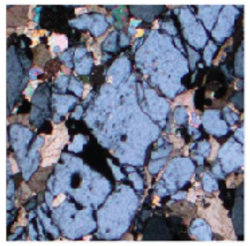

Magnetite
3.6k

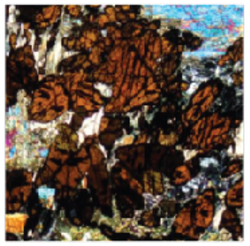

Ilmenite
2.1k

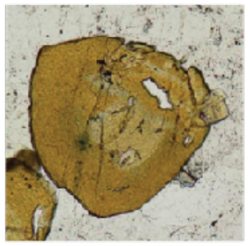

Biotite
3.2k

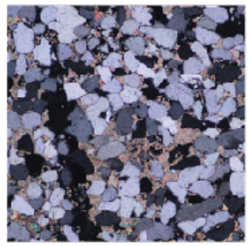

Muscovite
1.0k


Zircon
0.8k


Apatite
1.9k


Rutile
1.8k

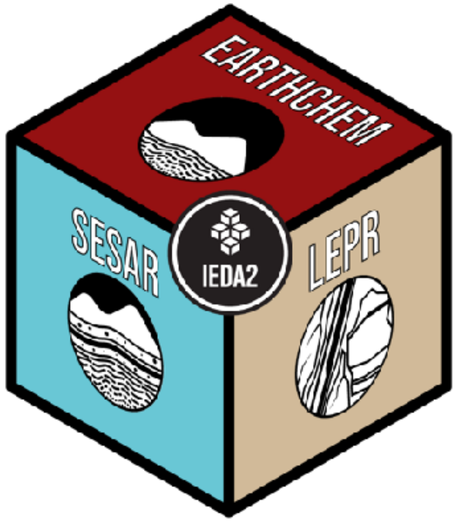

Tourmaline
1.2k


Quartz
40

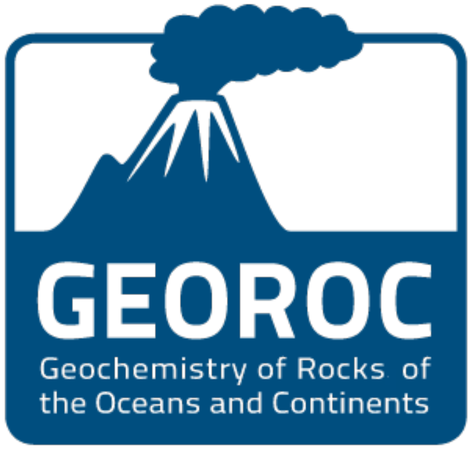
Alex Strekeisen

Compile studies focusing on individual minerals for confidence in phase.
Span tectonic settings and compositions to reflect natural samples.

TEST (evaluate final model)

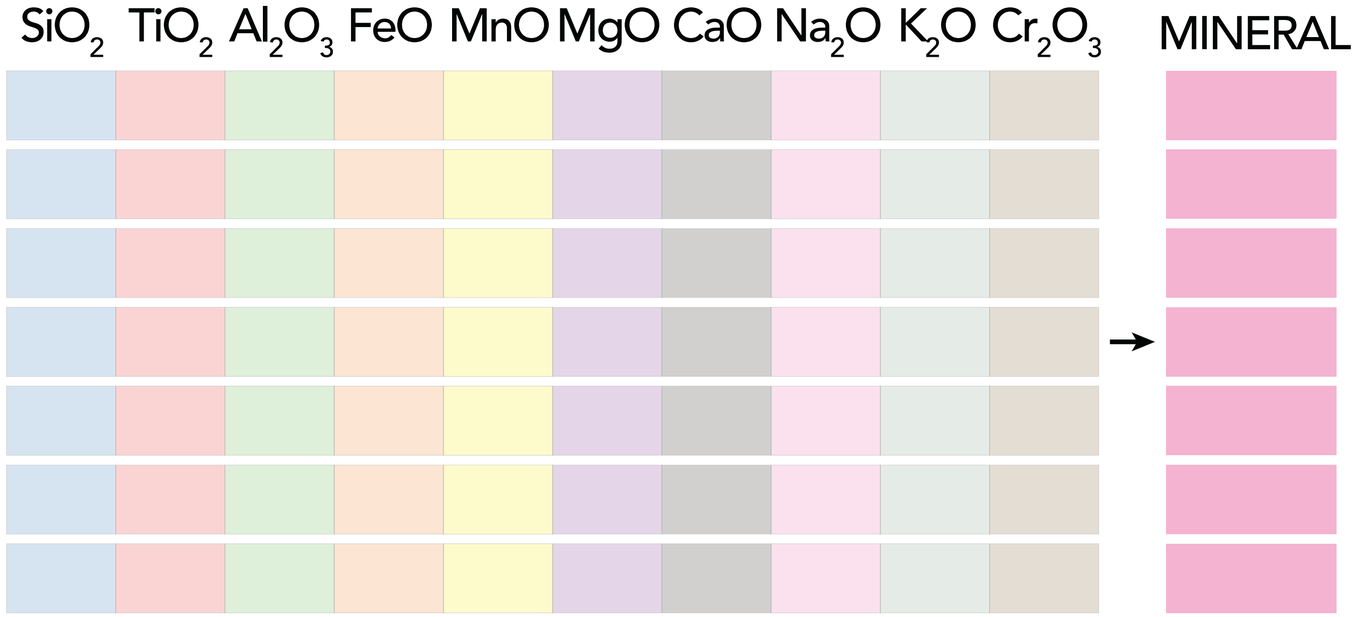


LEPR-12k



GEOROC-727k total

MIN-ML Overview: Supervised

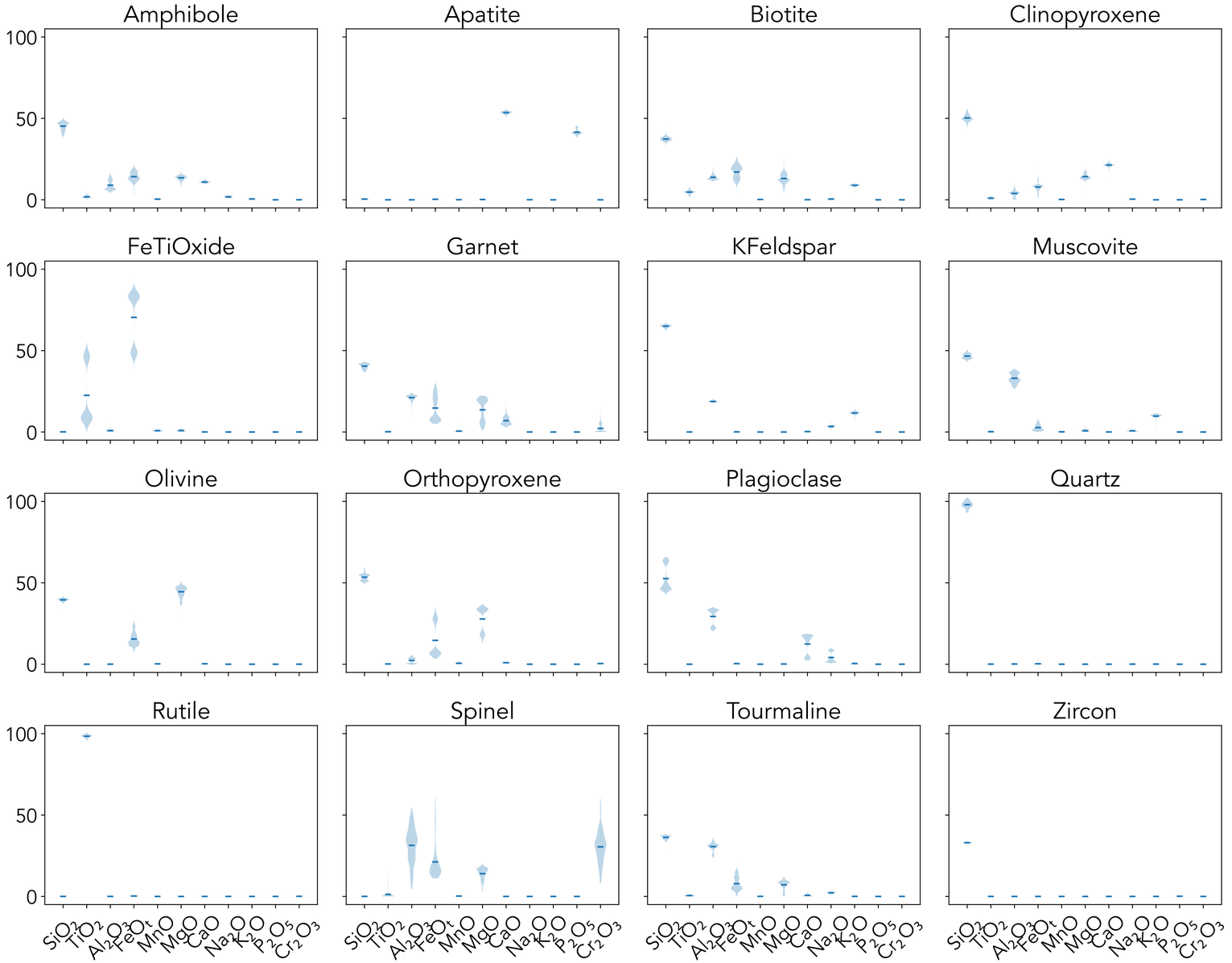


Supervised Learning:

Inputs: 10-D EPMA oxides + **labels**

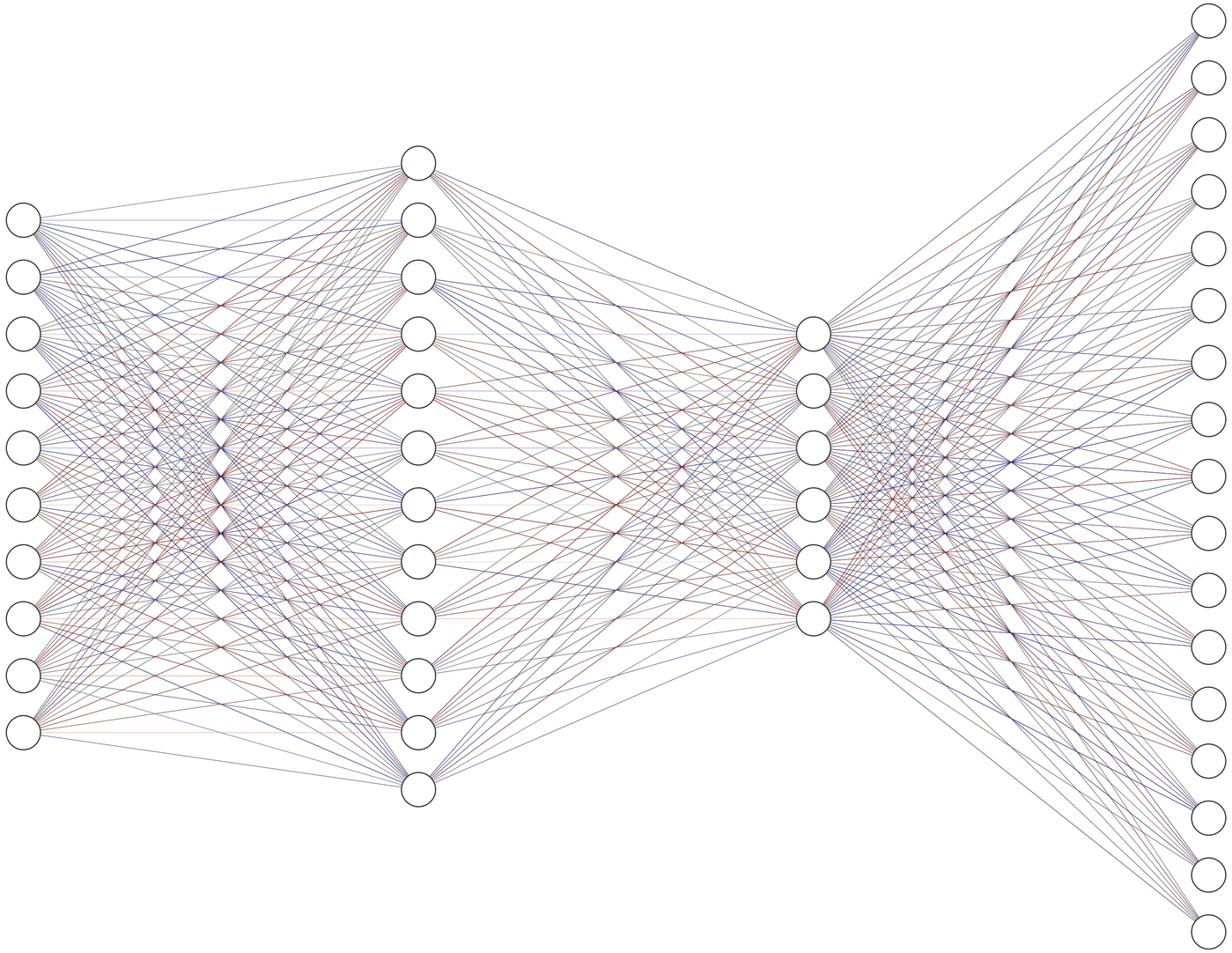
Methods: Neural networks with nested *k*-folds cross validation

Multivariate Normal Mineral Distributions



Most minerals have distinct oxide composition distributions — with the exception of amphibole and clinopyroxene.

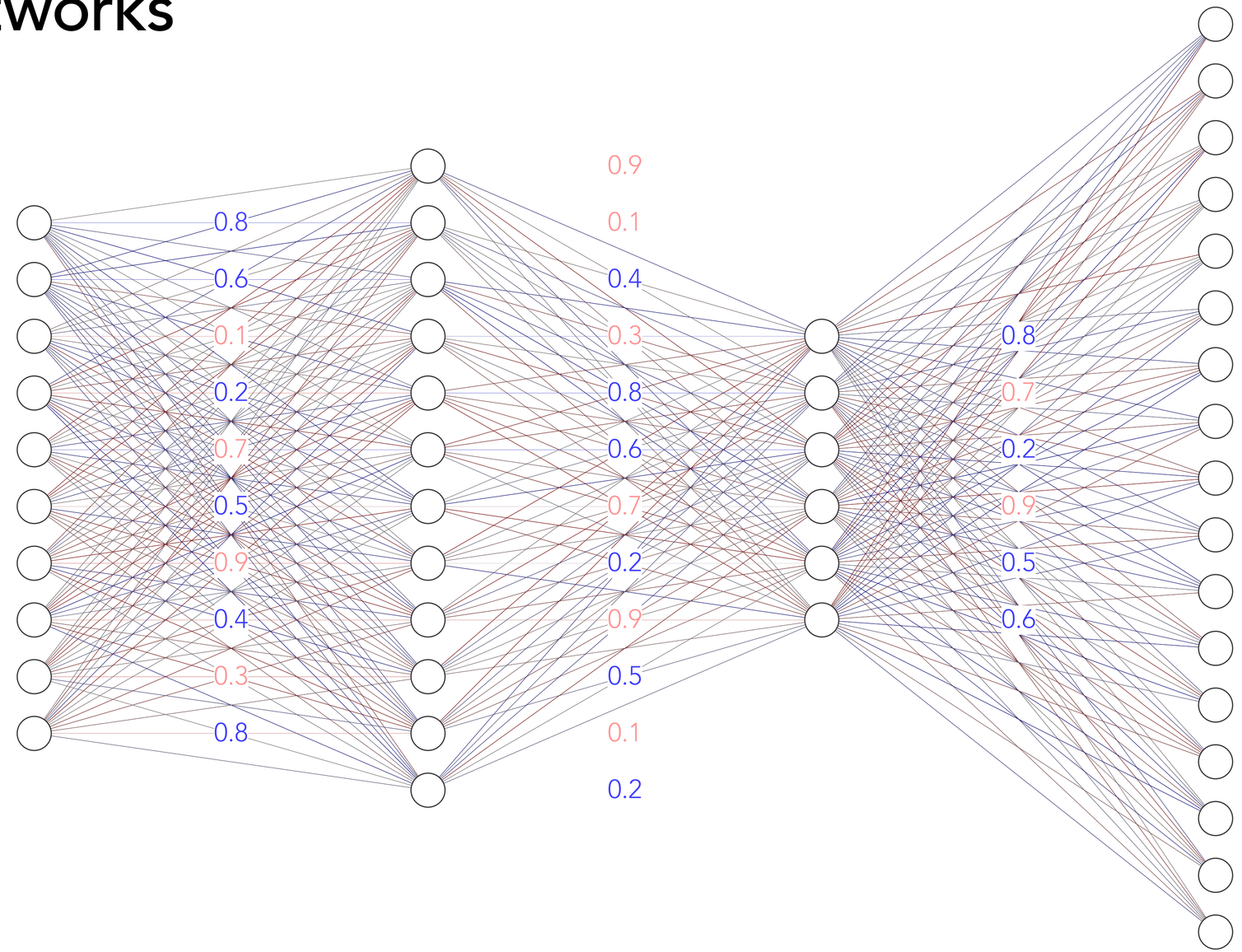
Neural Networks



Input Layer $\in R^{10}$ Hidden Layer $\in R^{n}$ Hidden Layer $\in R^{m}$ Output Layer $\in R^{17}$

Leverage learning performed on training data to make predictions on new unseen data.

Neural Networks



Input Layer
$\in R^{10}$

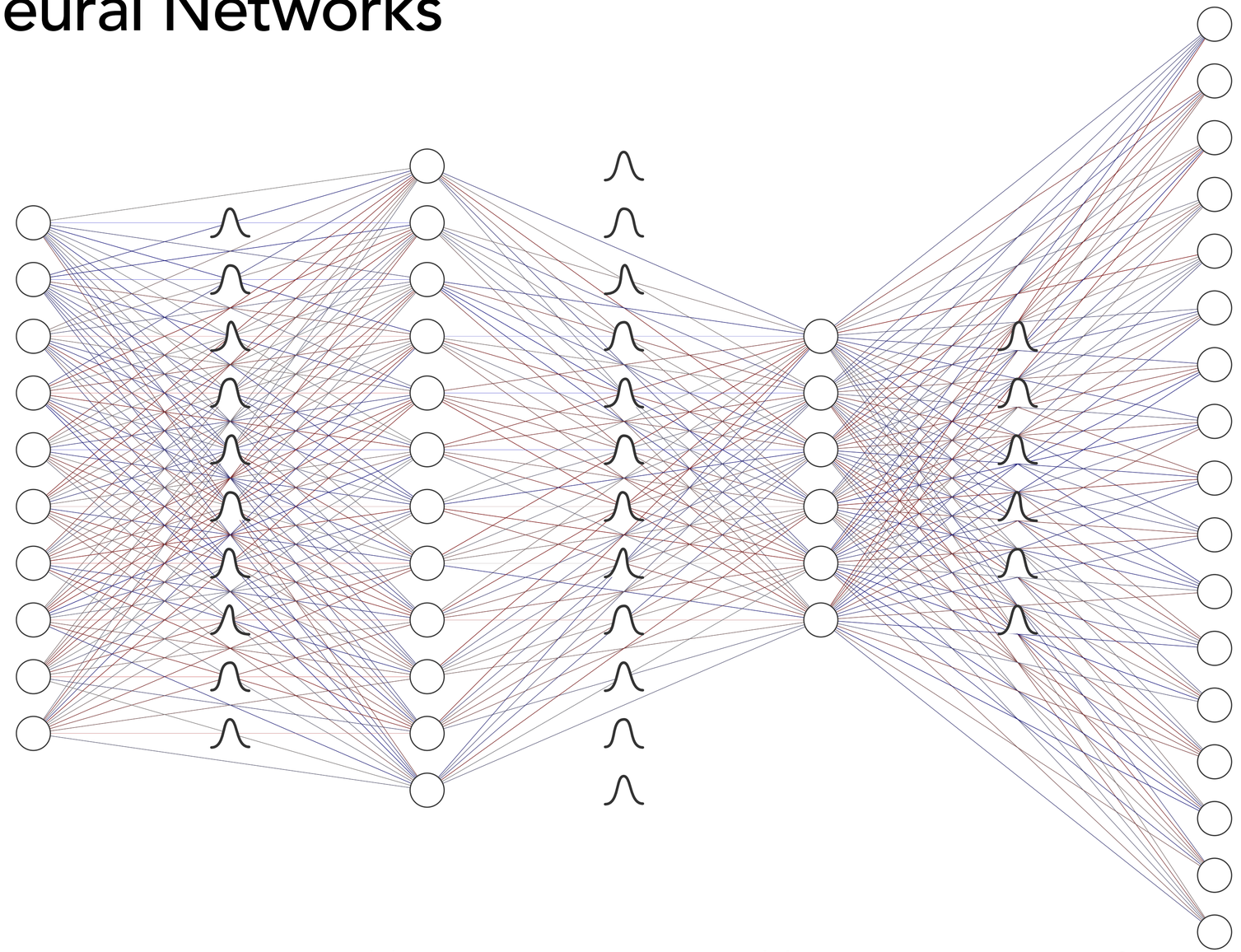
Hidden Layer
$\in R^{n}$

Hidden Layer
$\in R^{m}$

Output Layer
$\in R^{17}$

Leverage learning performed on training data to make predictions on new unseen data.

Bayesian Neural Networks



Input Layer
$\in R^{10}$

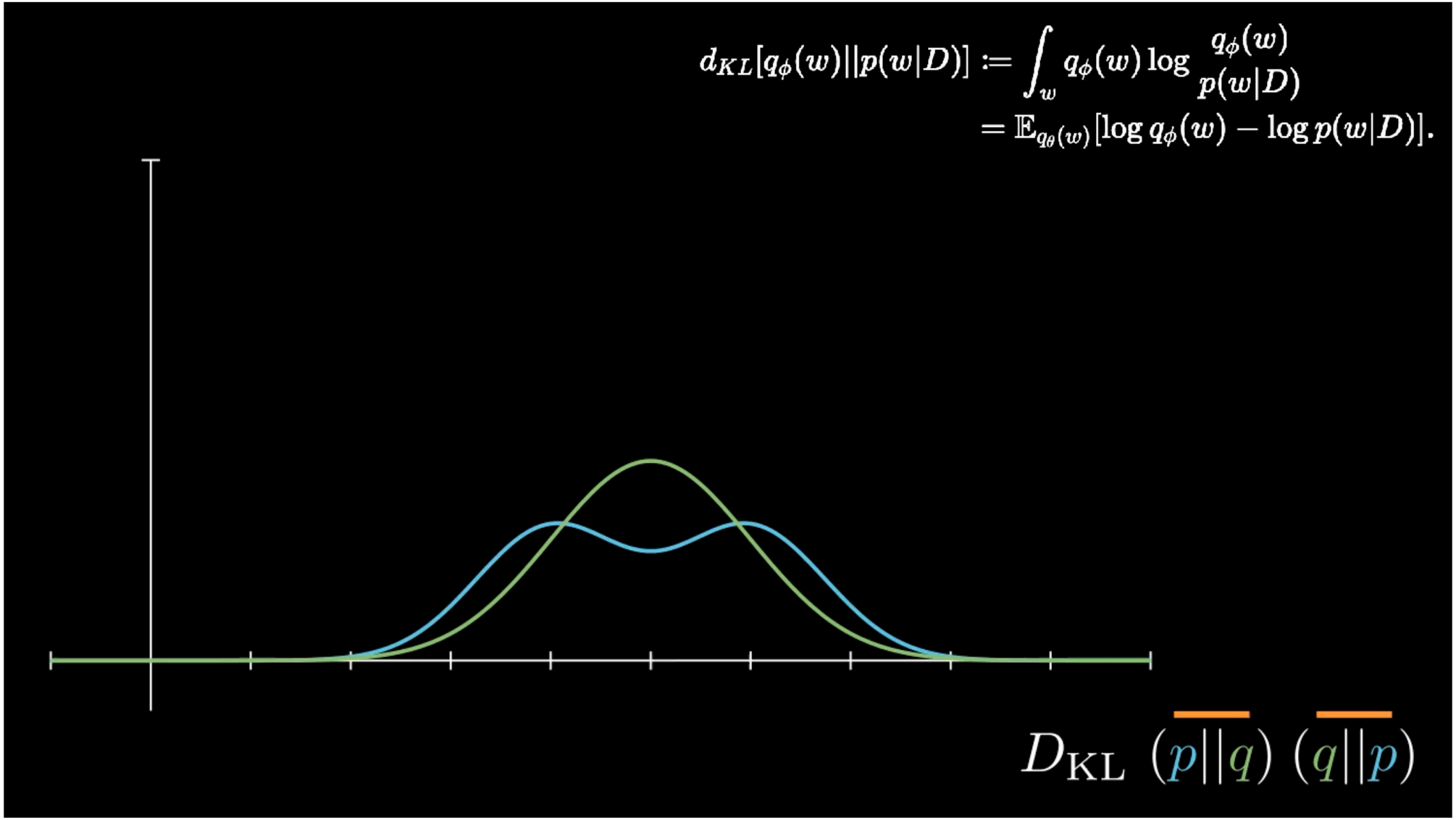
Hidden Layer
$\in R^{n}$

Hidden Layer
$\in R^{m}$

Output Layer
$\in R^{17}$

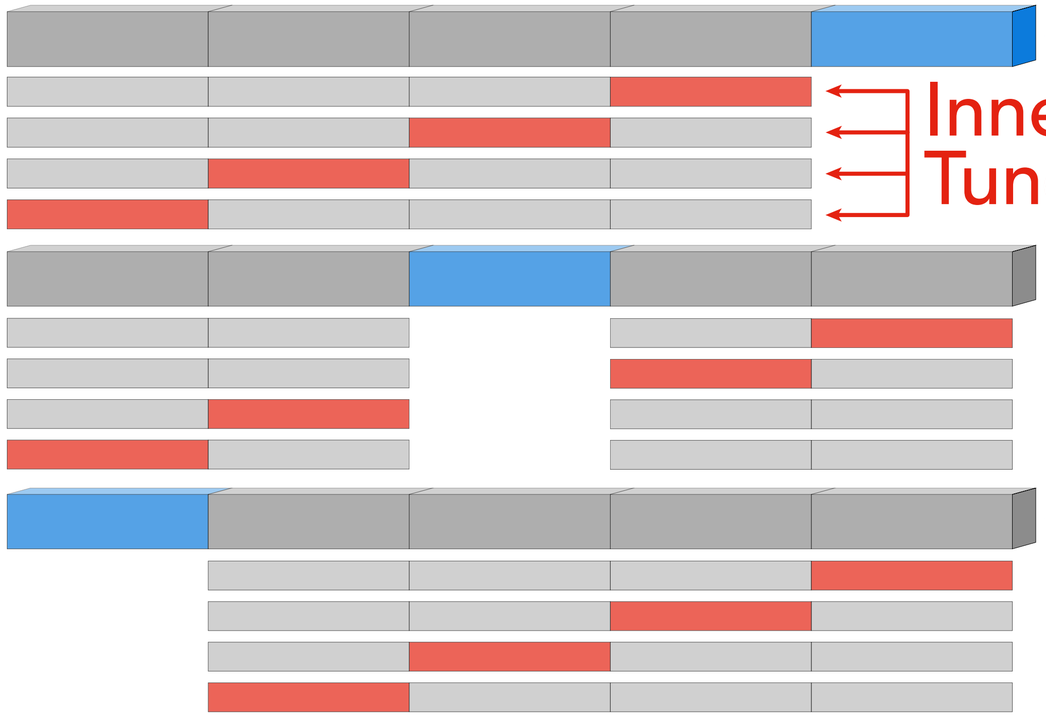
Leverage learning performed on training data to make predictions on new unseen data.

Variational Inference: Can’t calculate? Then approximate!



How do we compute these posterior distributions of weights?

Neural networks with nested *k*-folds cross validation



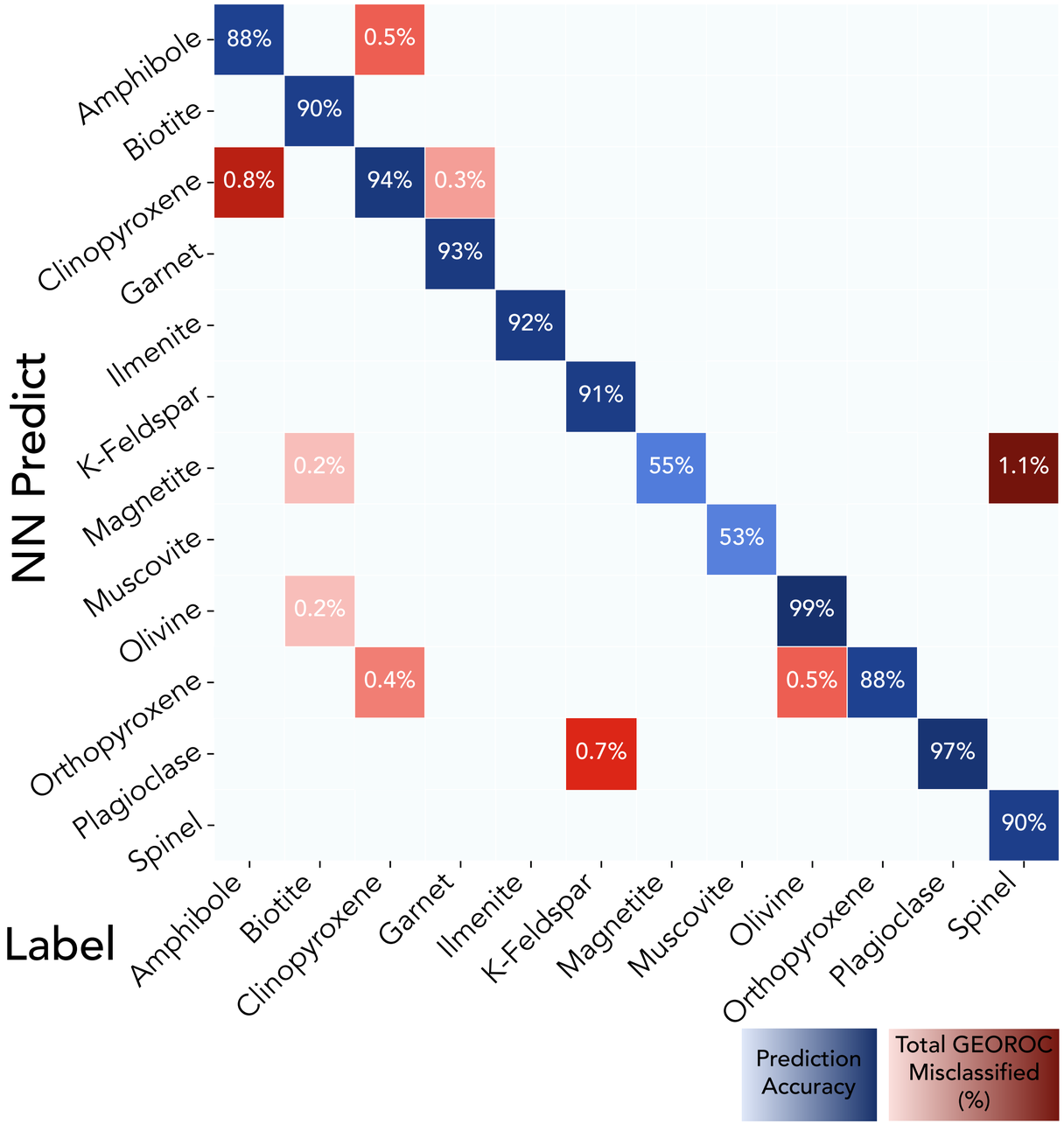
Inner Loops
Tune hyperparameters

Outer Loops
Estimate Error

$$h_\theta(x) = W^{[r]}a^{[r-1]} + b^{[r]}$$
$$a^{[r-1]} = \sigma(W^{[r-1]}a^{[r-2]} + b^{[r-1]})$$
$$\sigma(z) = max(0, z) + 0.02 * min(0, z)$$

Legend:
Outer train
Outer validate
Inner train
Inner validate

Neural Network Results

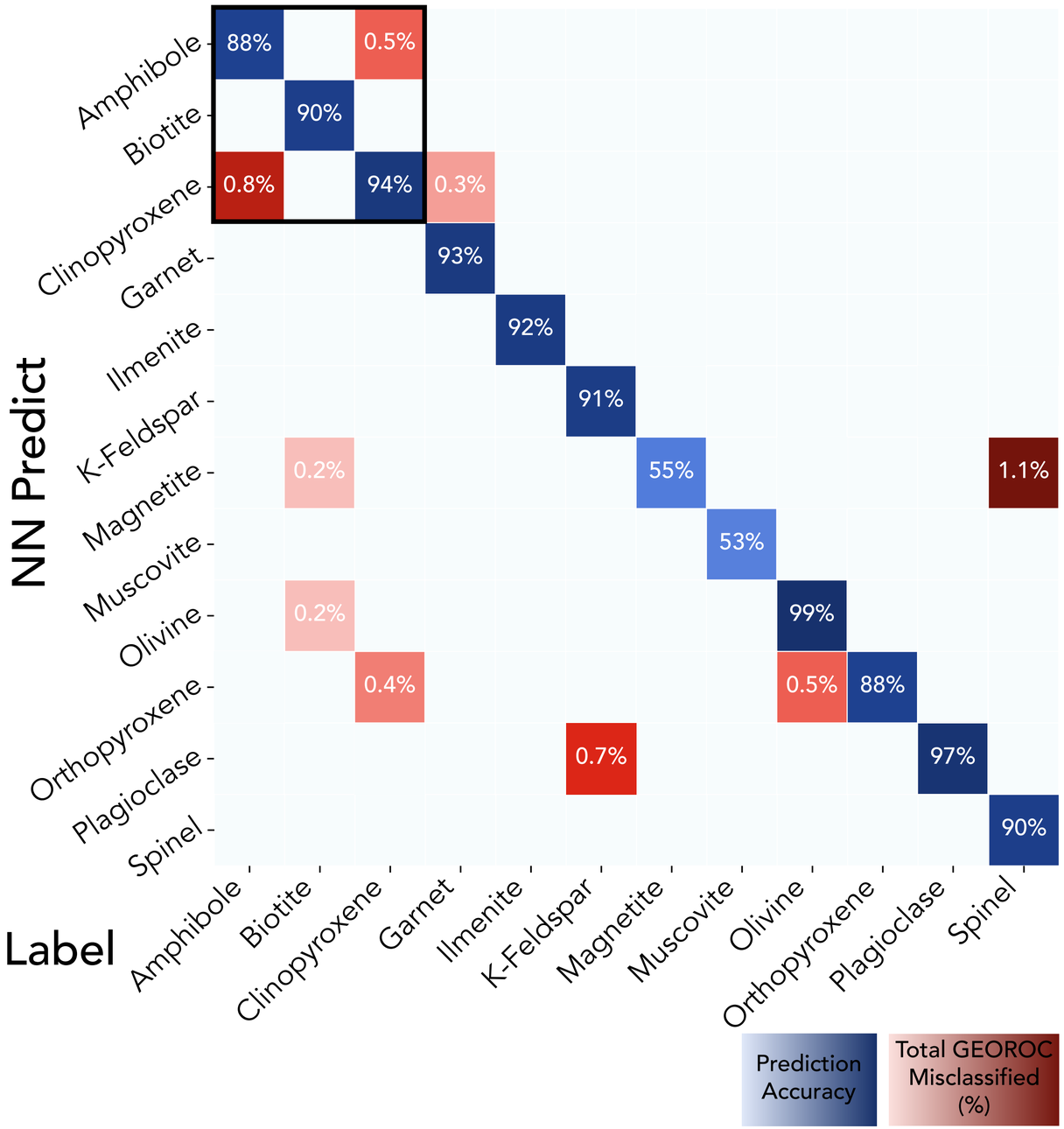


- Nested *k*-folds cross validation of training/ validation neural networks returns accuracy, precision, recall, and f1 scores = 99±1%.
- **GEOROC classification accuracy (727k) = 95%**

How can we use incorrect classifications to better understand the data?

- Take misclassifications or ‘failures’ of the neural network and examine more closely.
- Clinopyroxene and amphibole classifications require closer examination.

Neural Network Results

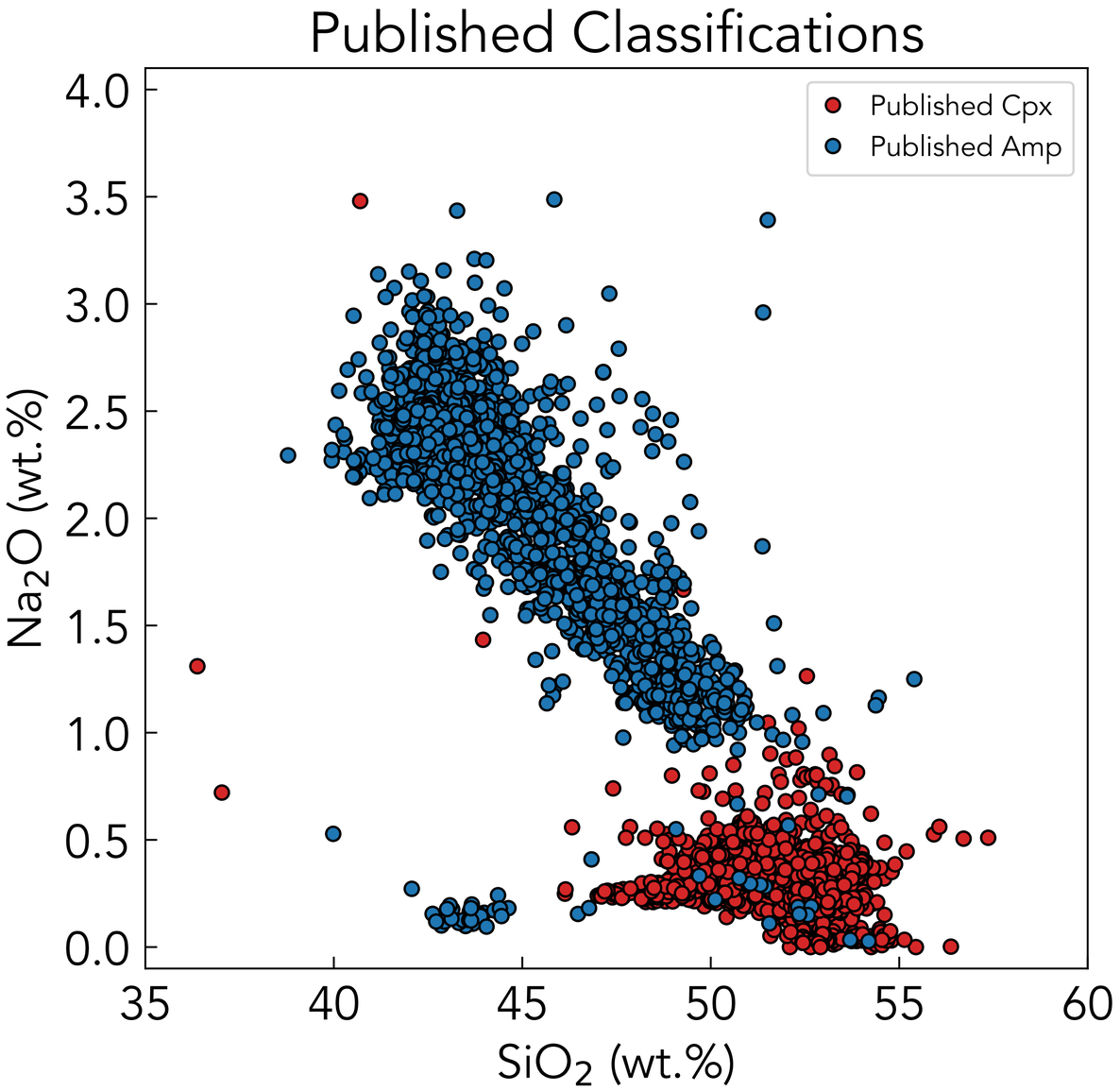


- Nested *k*-folds cross validation of training/ validation neural networks returns accuracy, precision, recall, and f1 scores = 99±1%.
- **GEOROC classification accuracy (727k) = 95%**

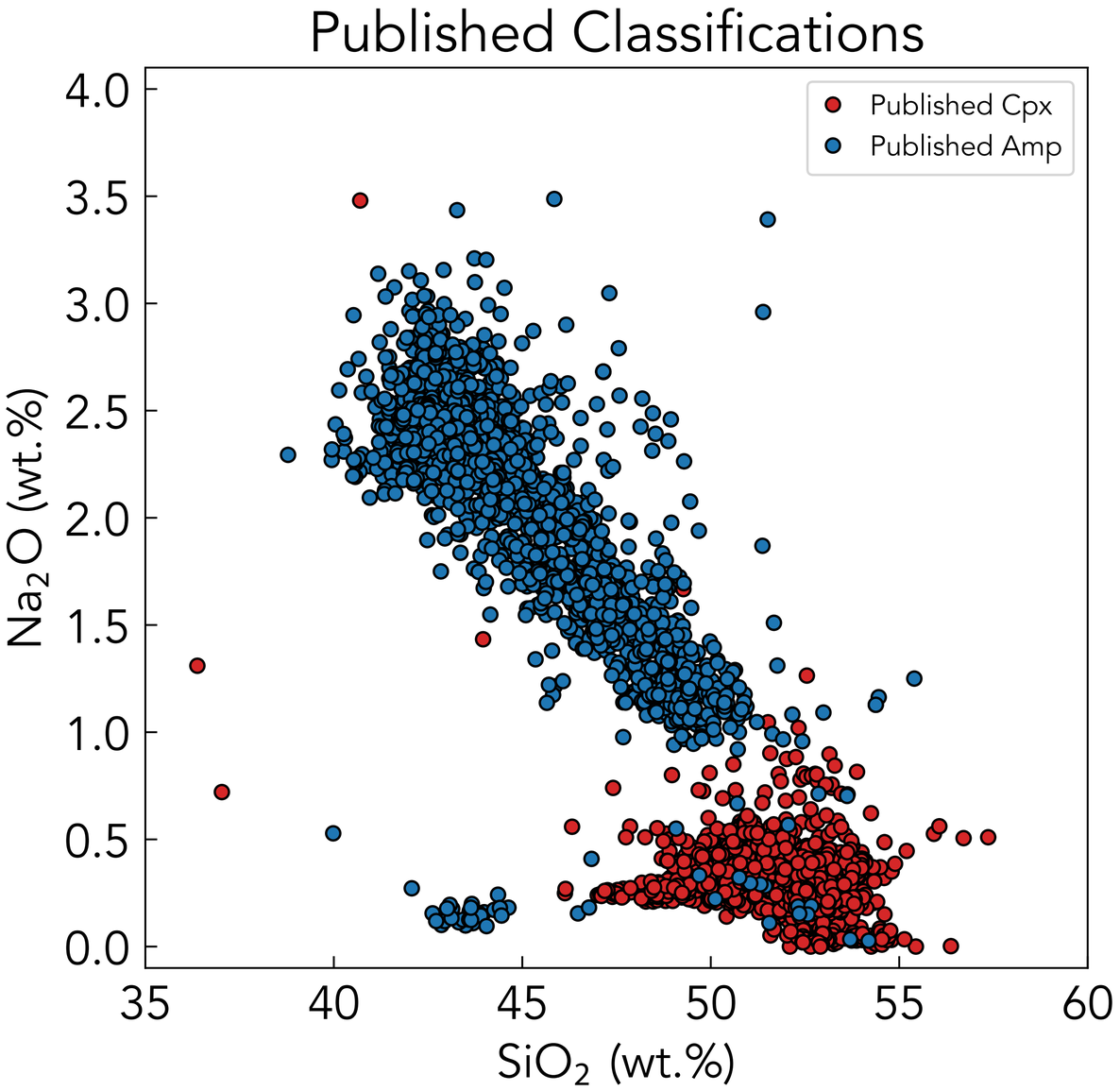
How can we use incorrect classifications to better understand the data?

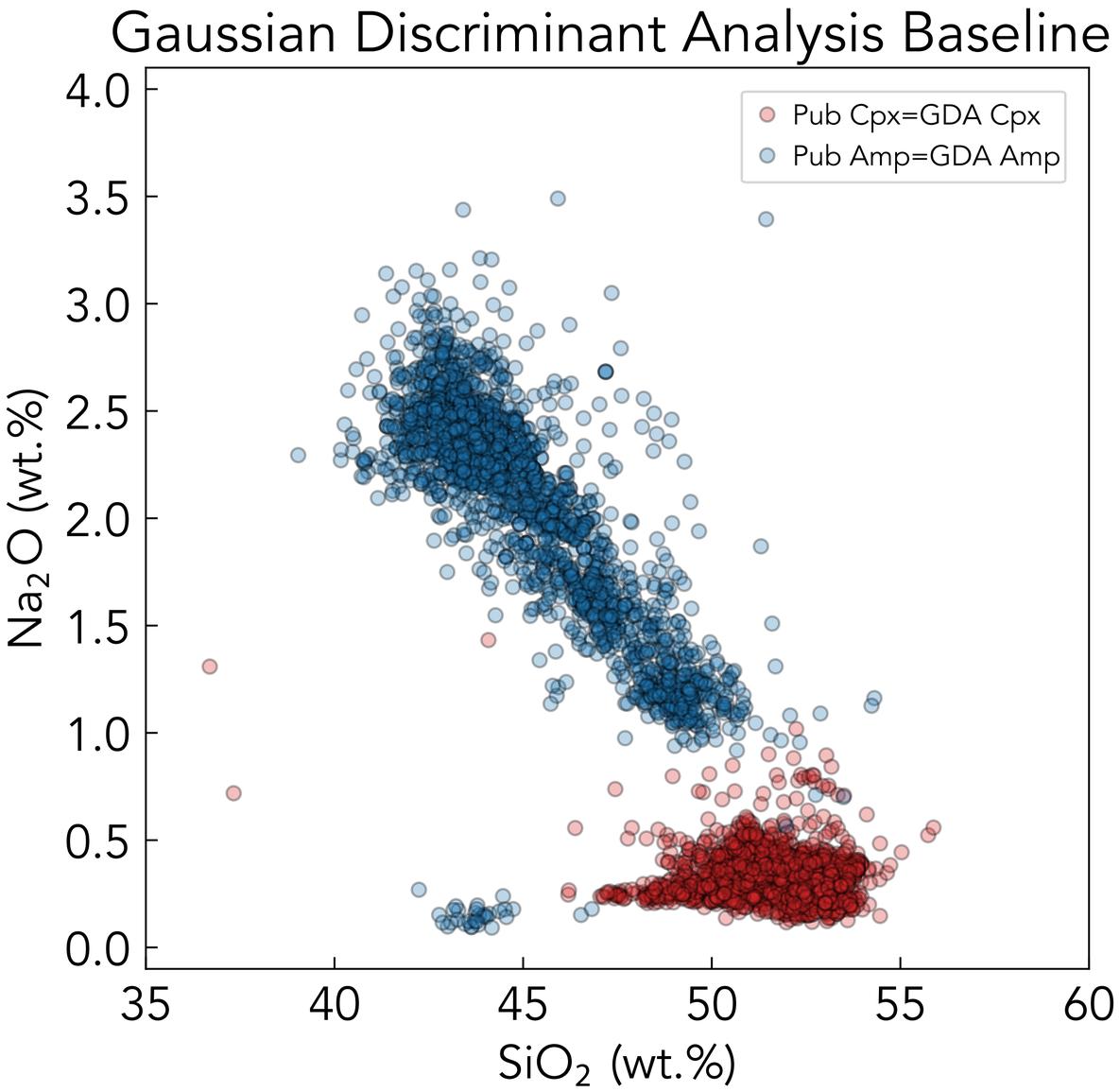
- Take misclassifications or ‘failures’ of the neural network and examine more closely.
- Clinopyroxene and amphibole classifications require closer examination.

Clinopyroxene-Amphibole in the Cascades



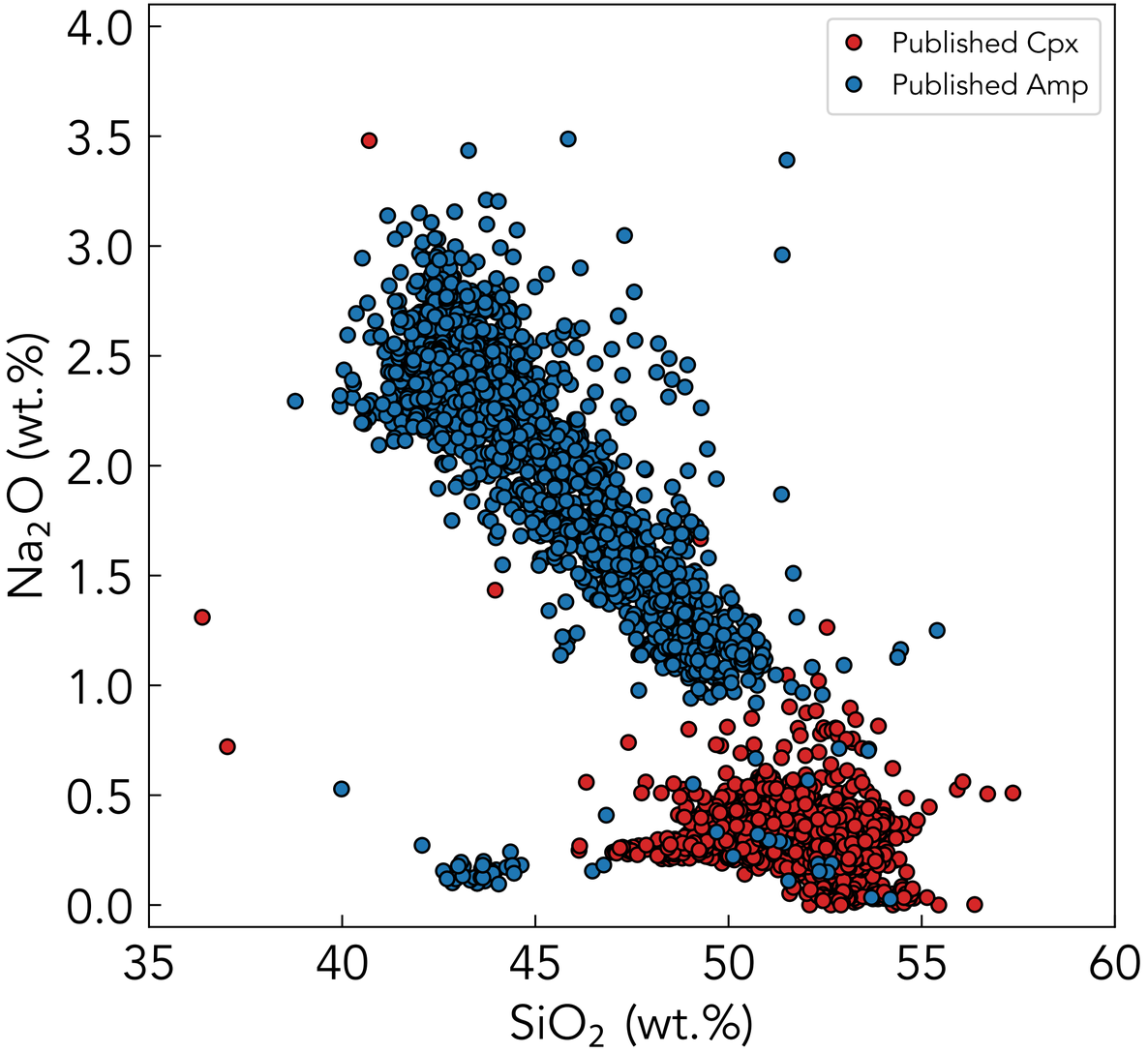
Clinopyroxene-Amphibole in the Cascades: GDA



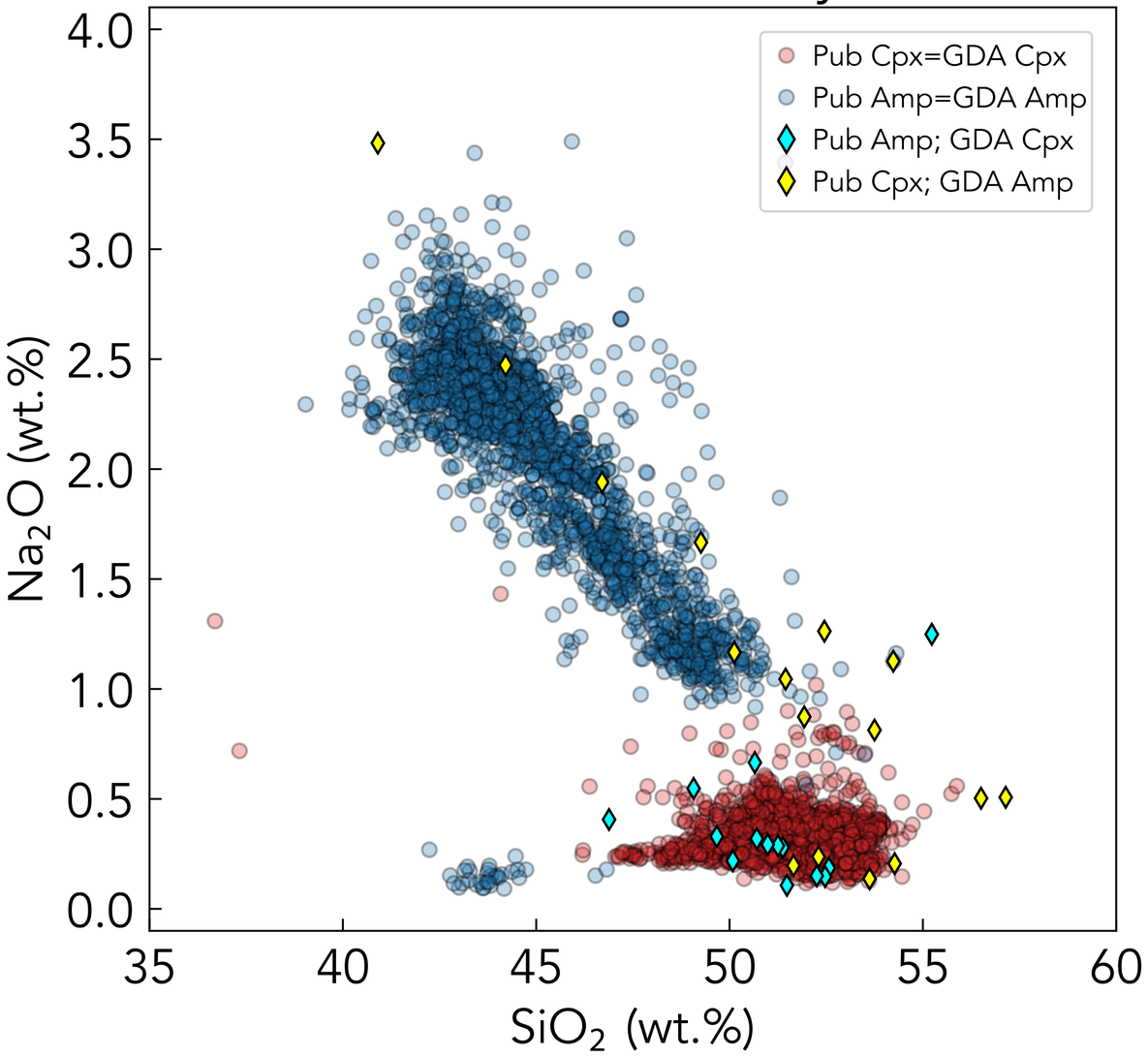


Clinopyroxene-Amphibole in the Cascades

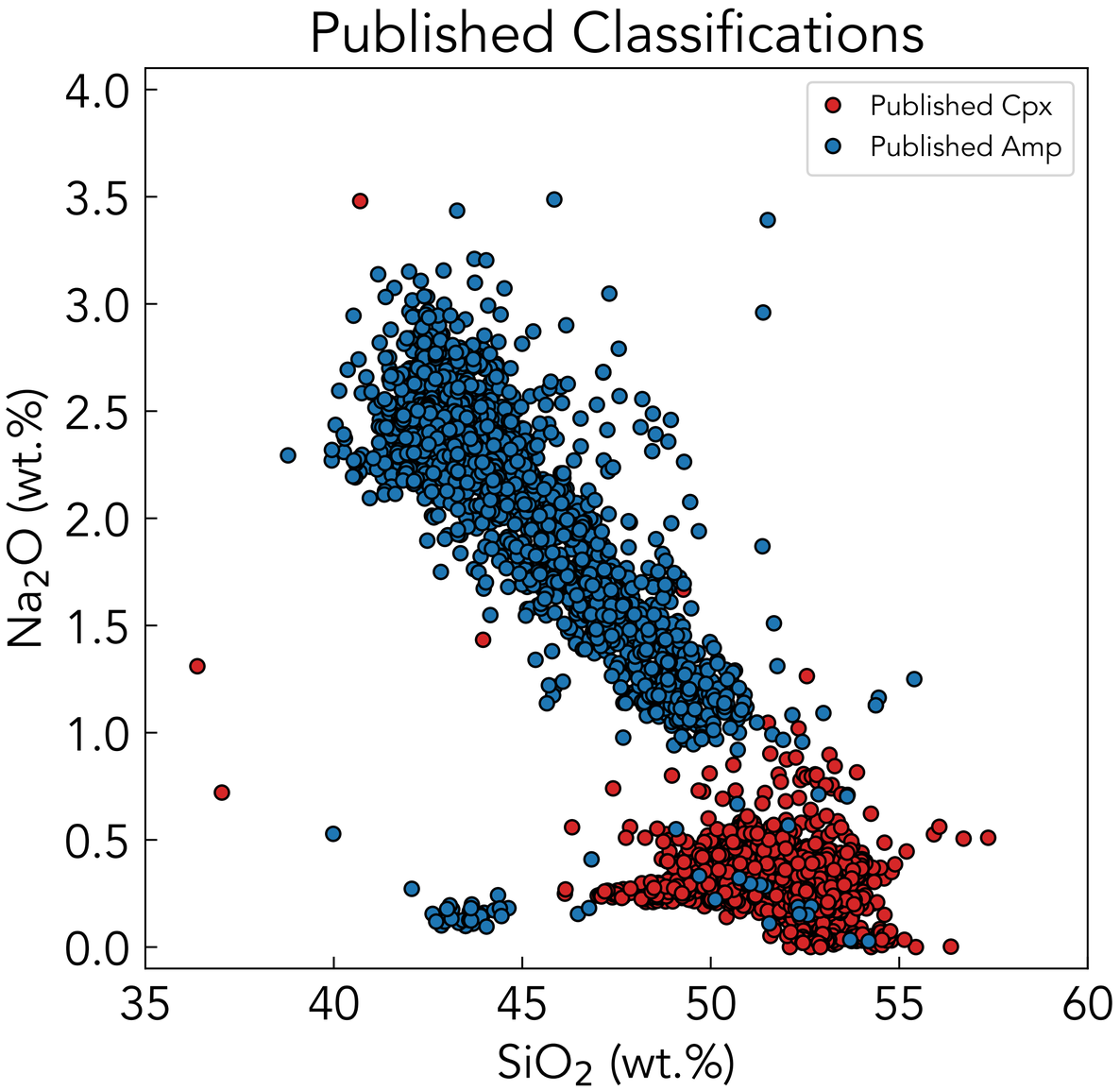
Published Classifications

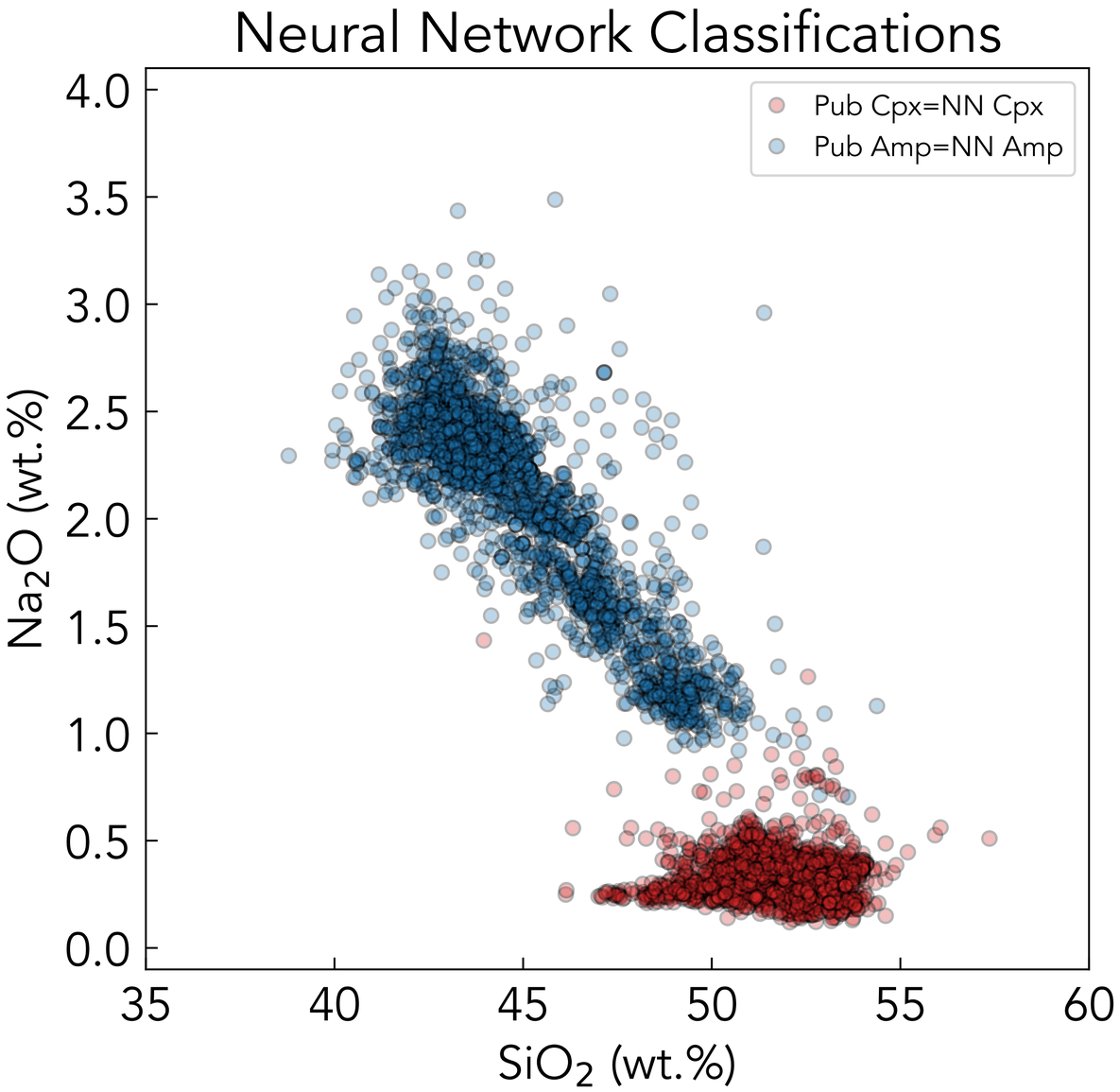


Gaussian Discriminant Analysis Baseline

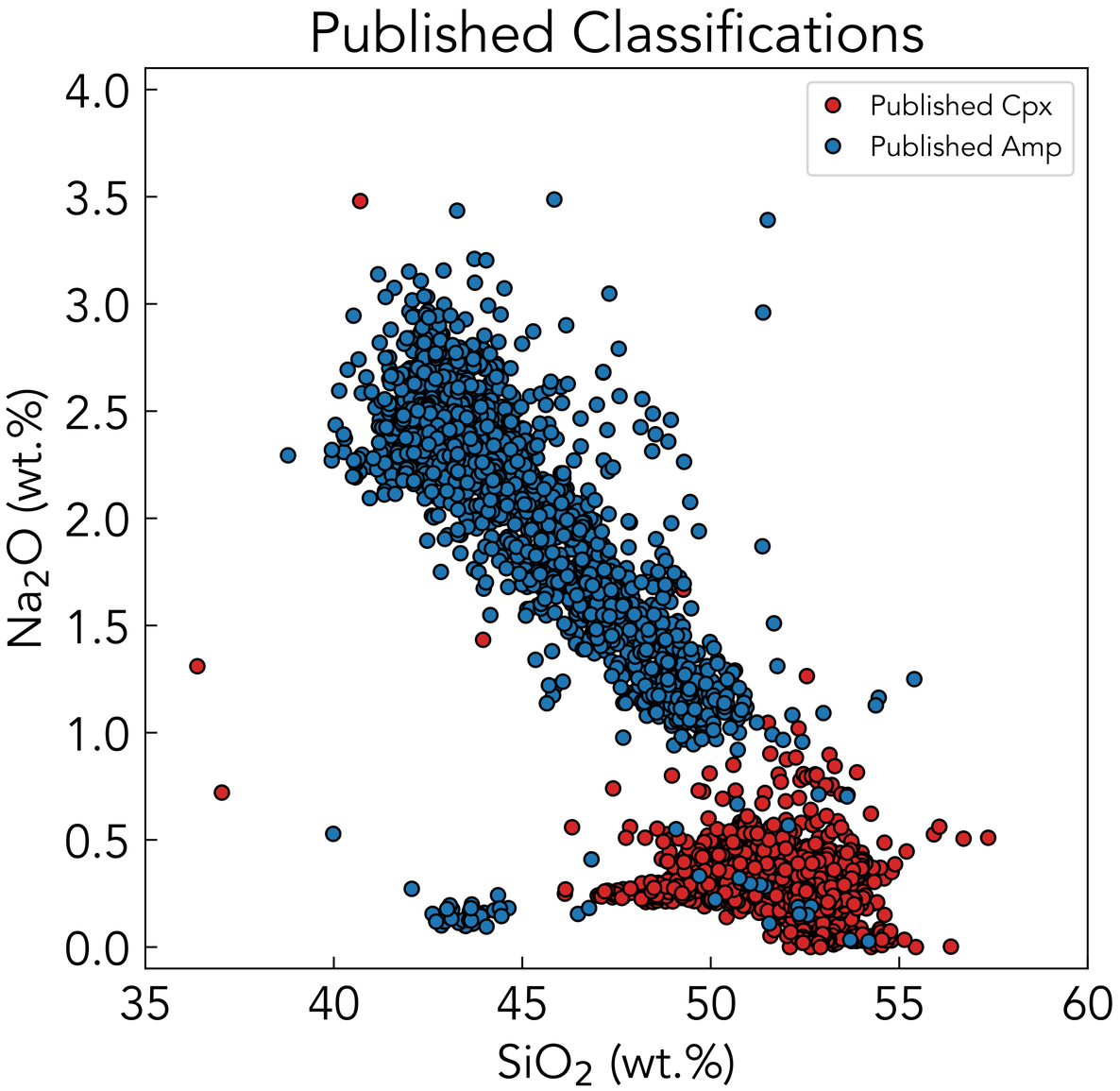


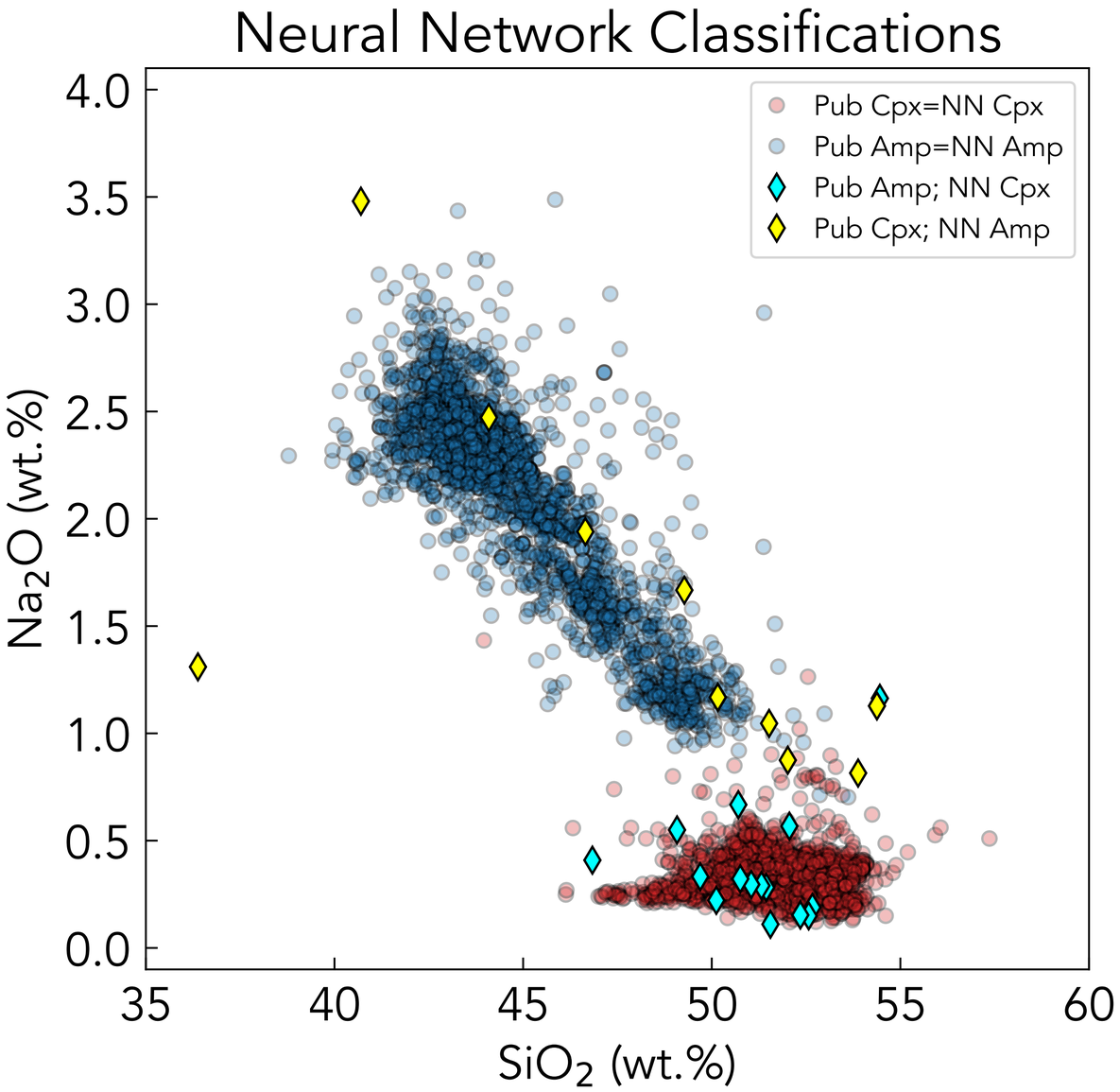
Clinopyroxene-Amphibole in the Cascades



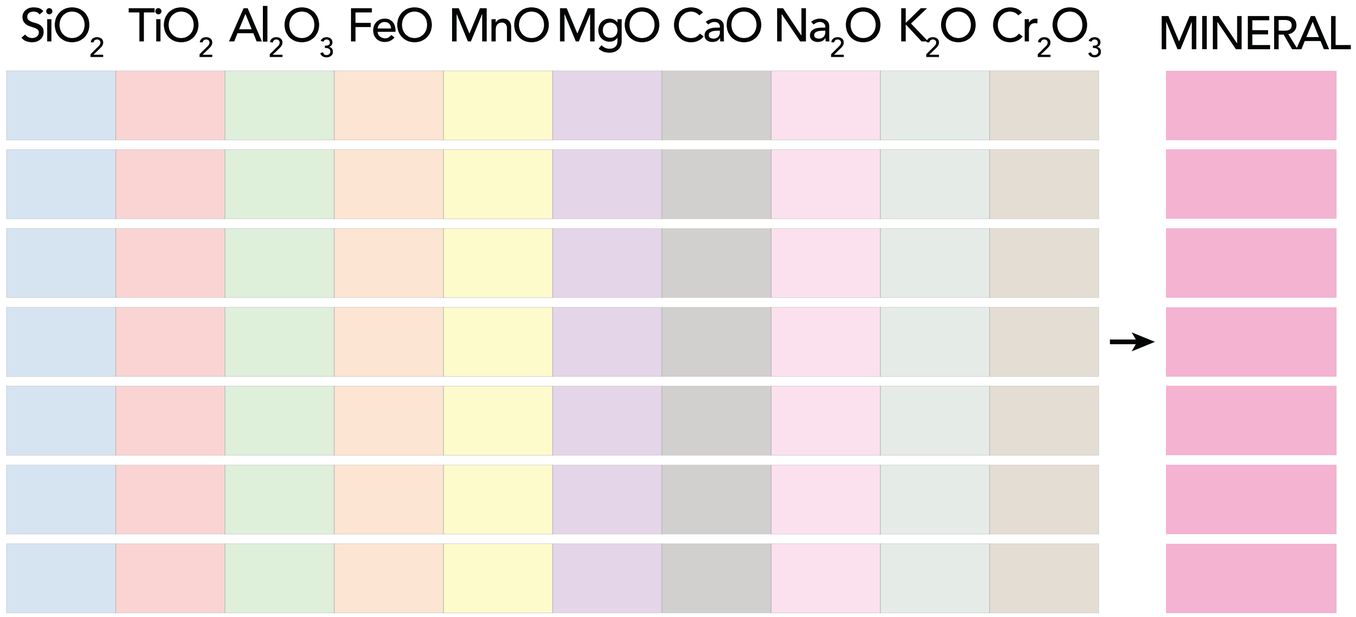


Clinopyroxene-Amphibole in the Cascades





MIN-ML Overview: Unsupervised



Supervised Learning:

Inputs: 10-D EPMA oxides + **labels**

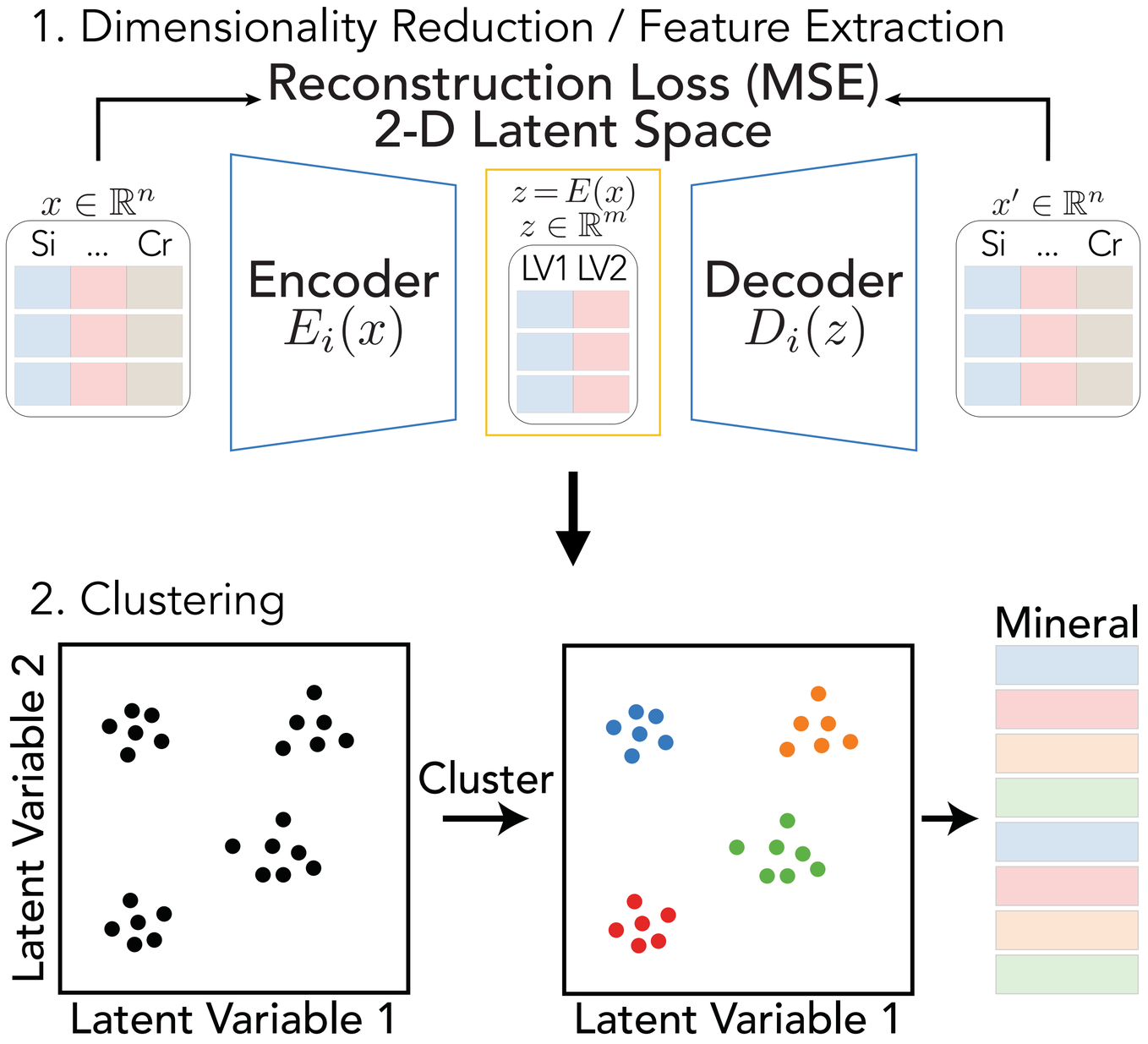
Methods: Neural networks with nested *k*-folds cross validation

Unsupervised Learning:

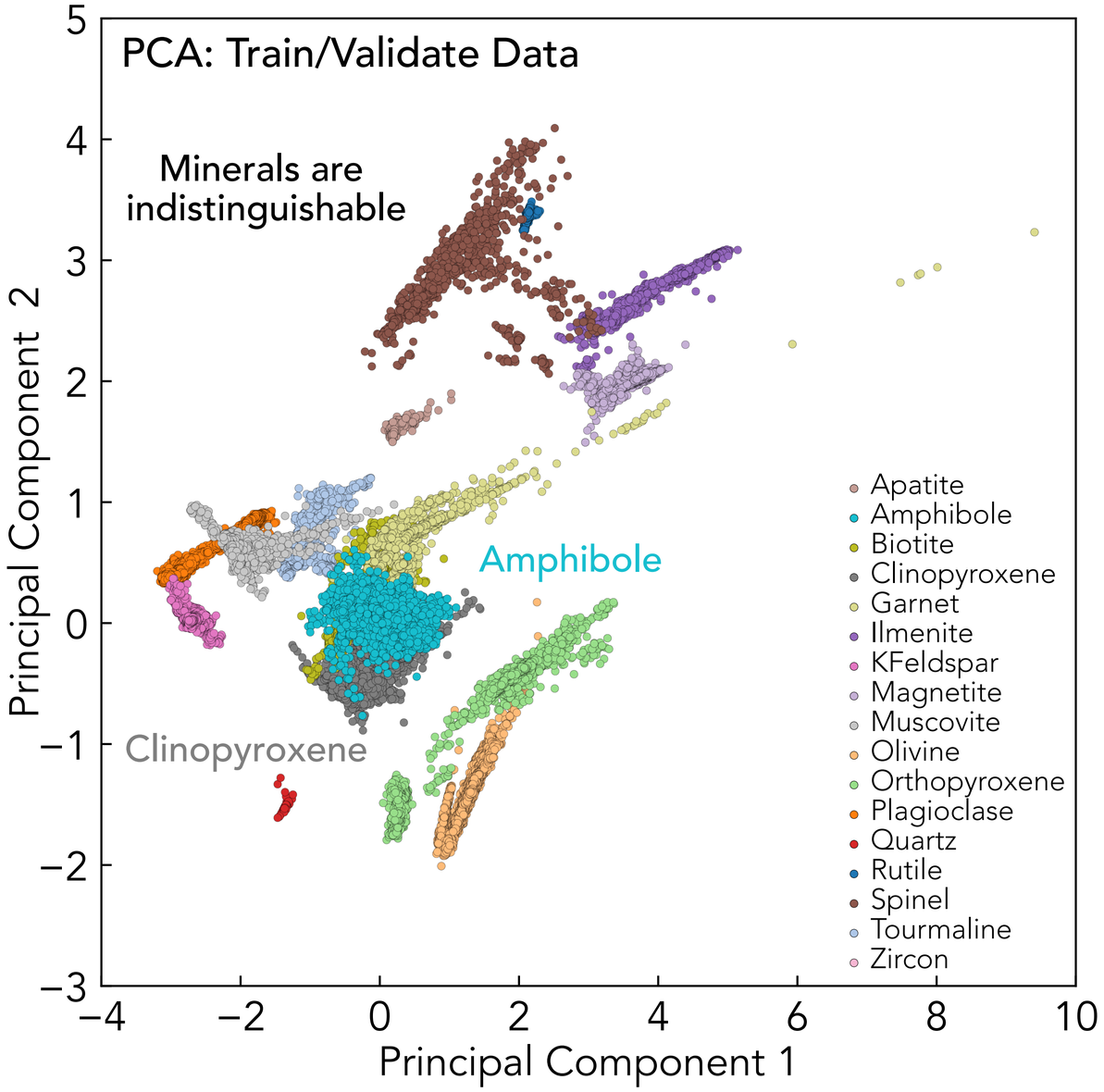
Inputs: 10-D EPMA oxides

Methods: Dimensionality reduction with autoencoder and density-based clustering

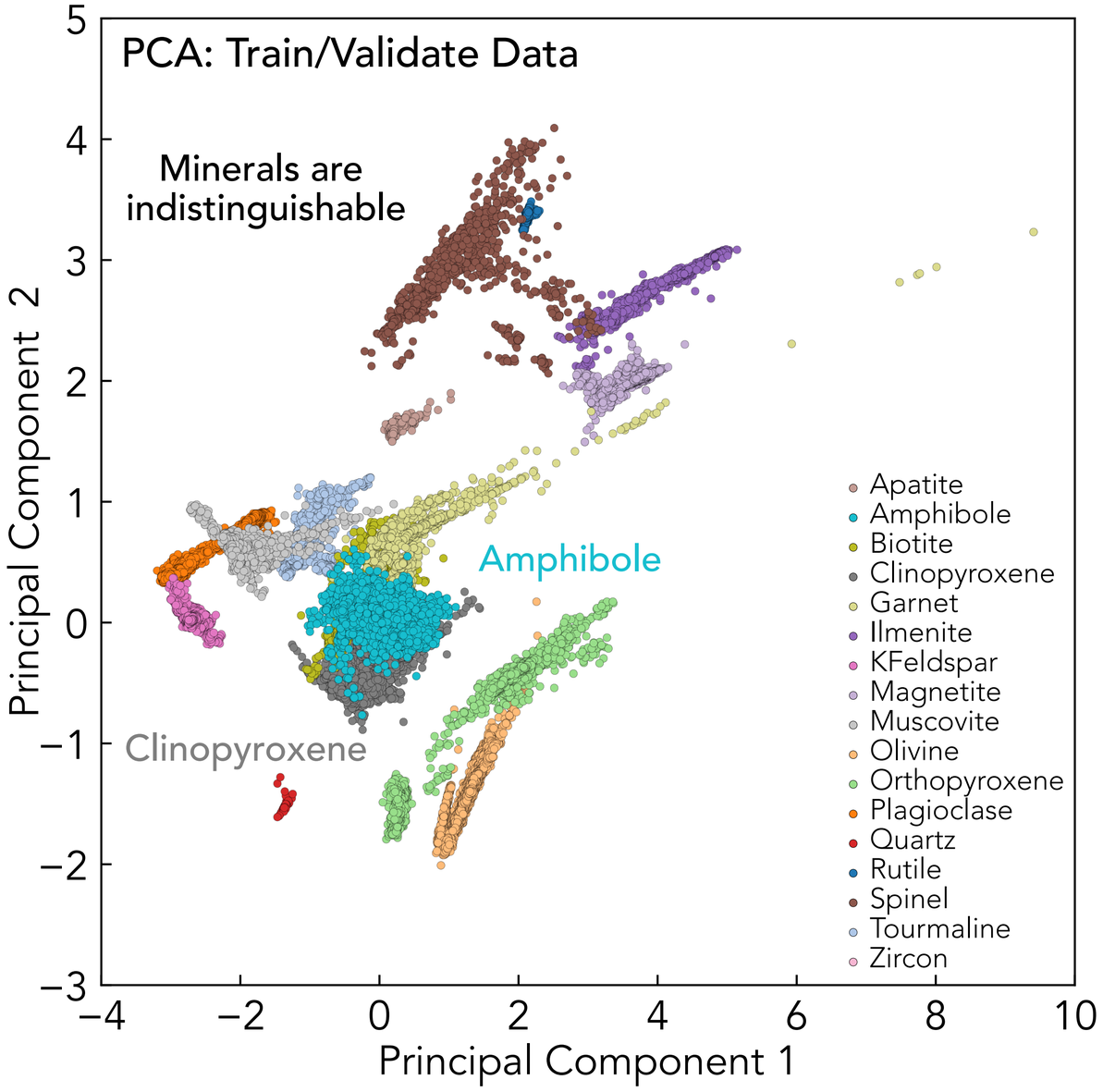
Autoencoder dimensionality reduction+clustering

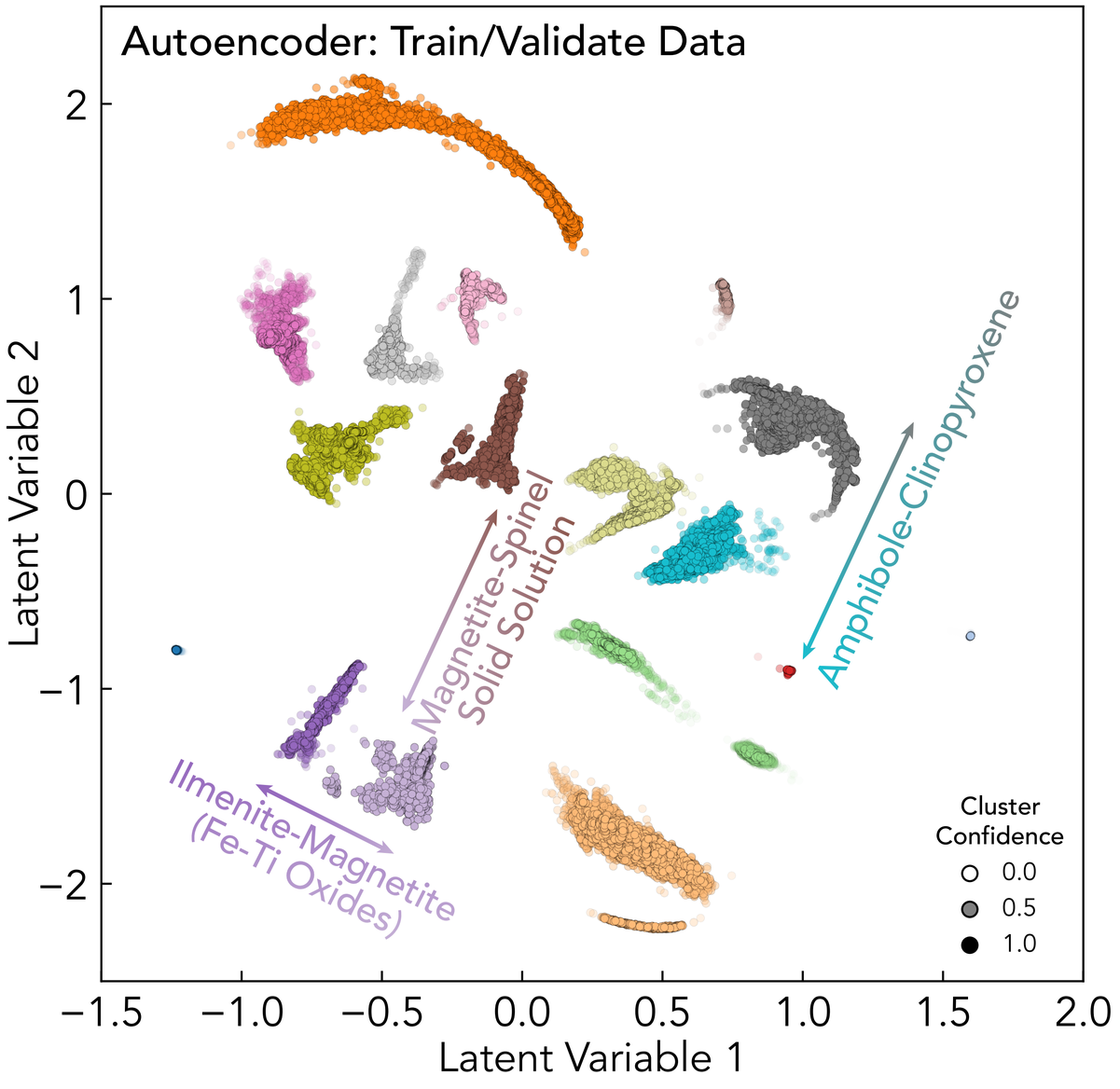


Autoencoder Results

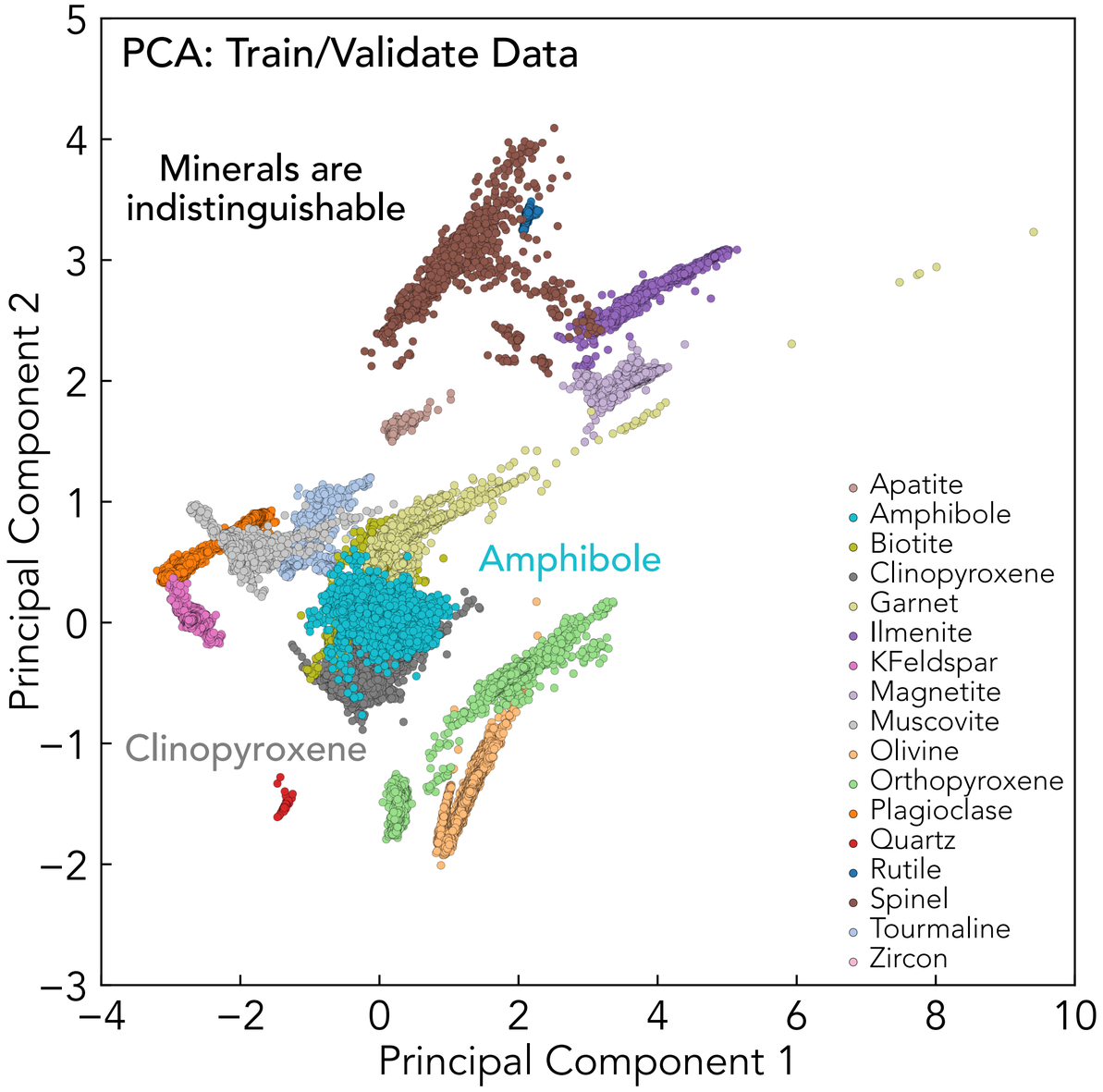


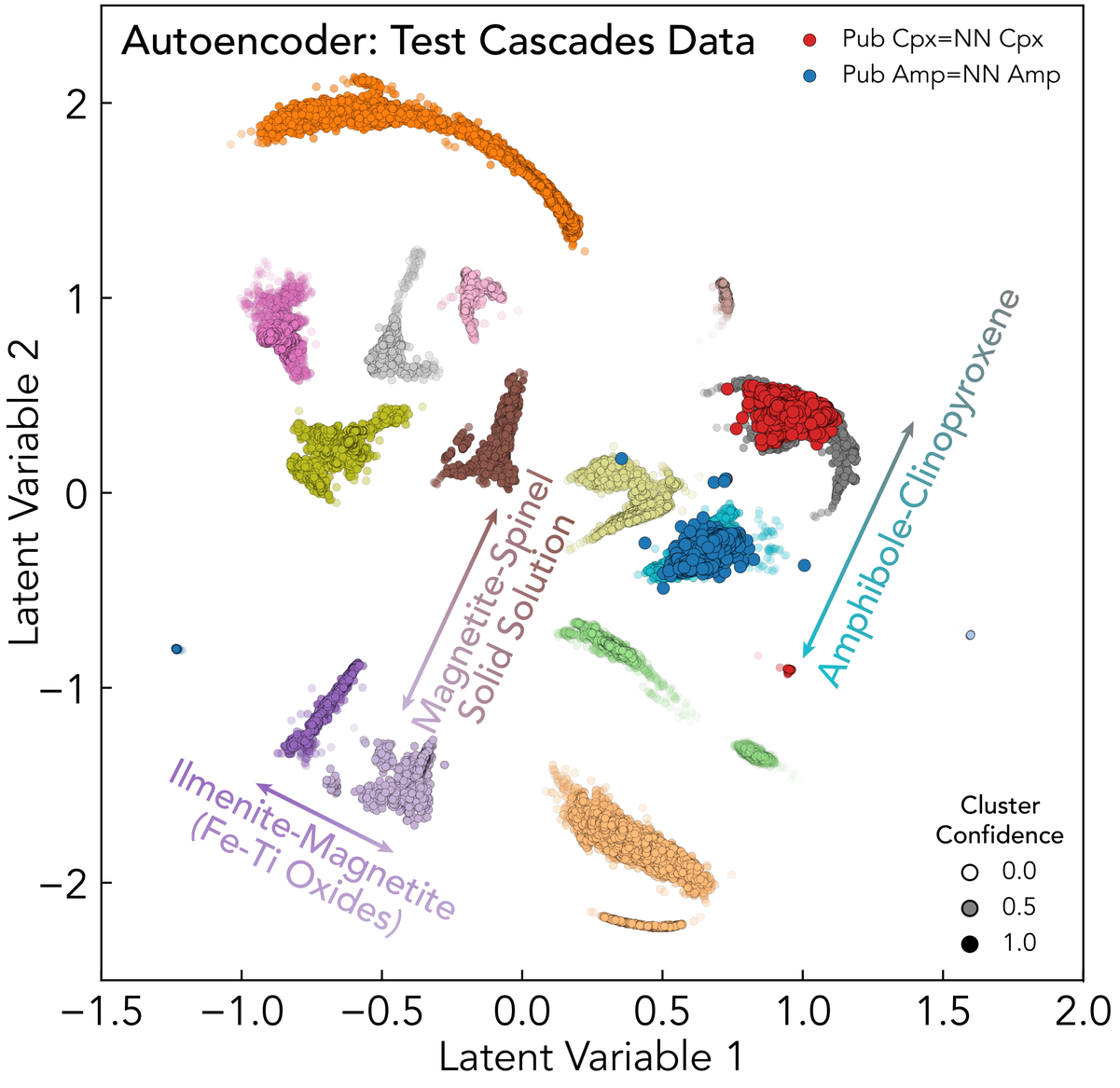
Autoencoder Results



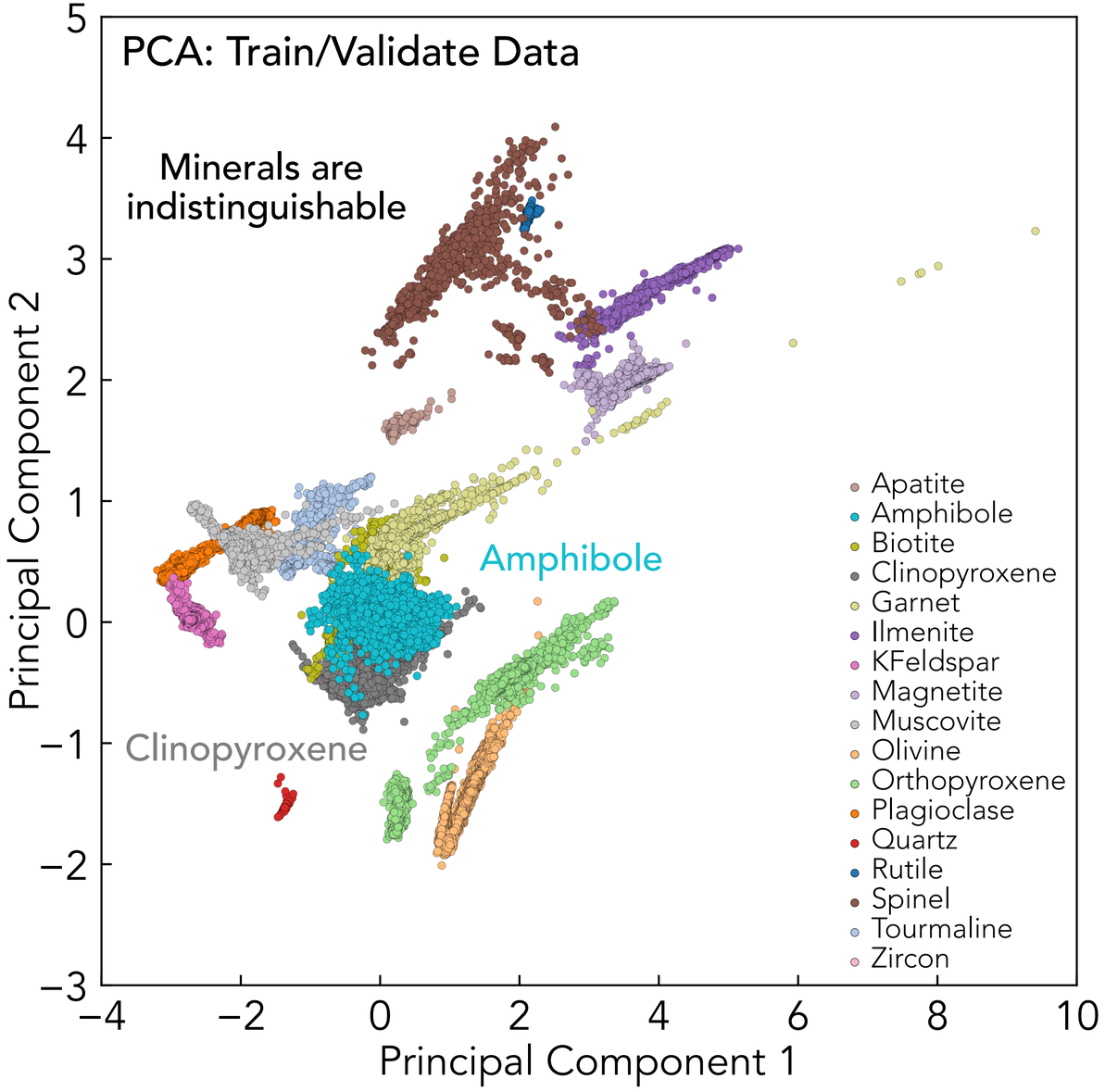


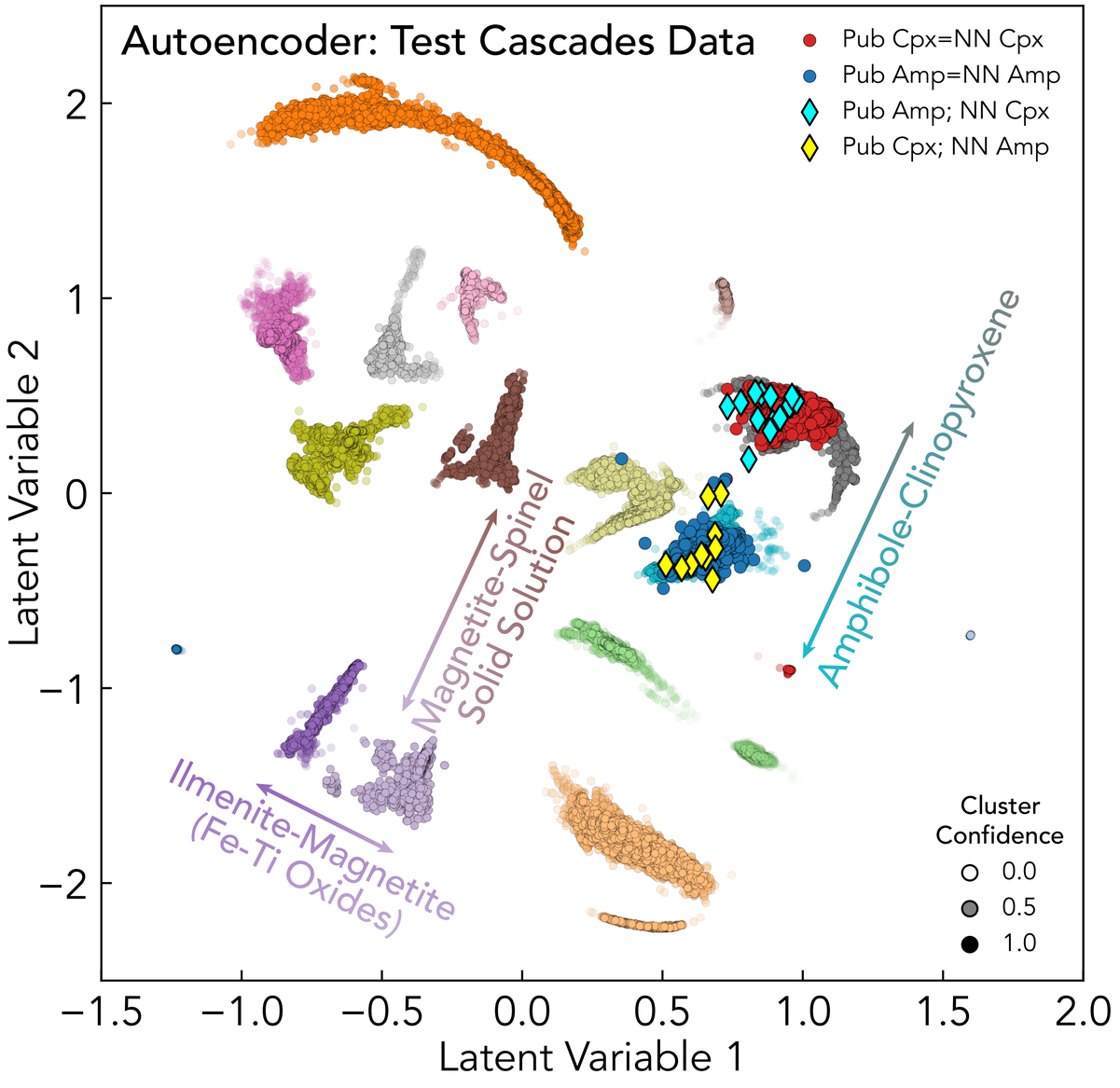
Autoencoder Results





Autoencoder Results





Conclusions

- Supervised neural networks successfully classify most minerals with 95% accuracy
  - `Misclassifications' of neural network show where published classifications may be dubious and can be reëxamined
- Unsupervised autoencoder shows similar trends and offers useful visualizations of latent space

Next Steps

- Analyze spinel-ilmenite-magnetite space
- Determine cutoff uncertainties for results from Bayesian neural network
- Distribute open-source Python package for applying these models to data
- Submit your data to databases to improve these models!