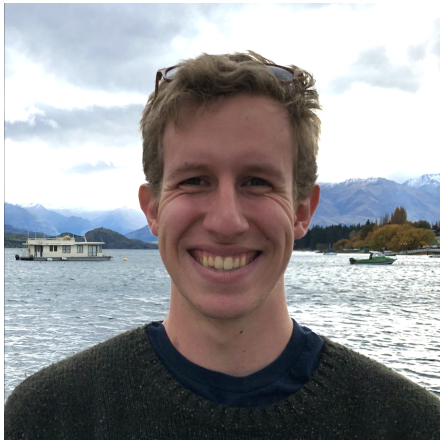
MINdicator

A Novel Application of Machine Learning to a New SEM Silicate Database

Benjamin Parfitt
Reiform
ben@reiform.com

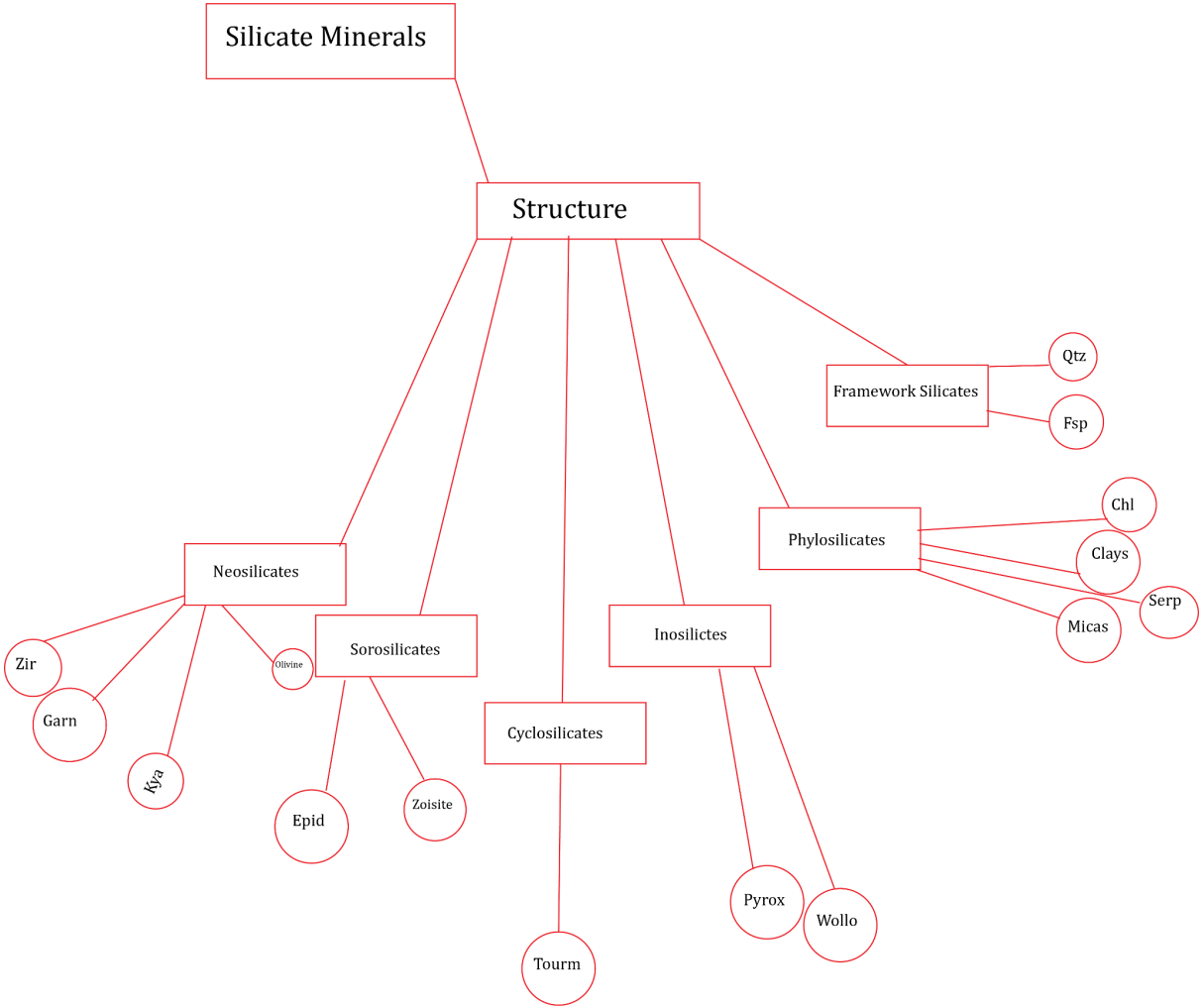
Robert Welch
Harvard University
Earth and Planetary Sciences
rwelch@g.harvard.edu

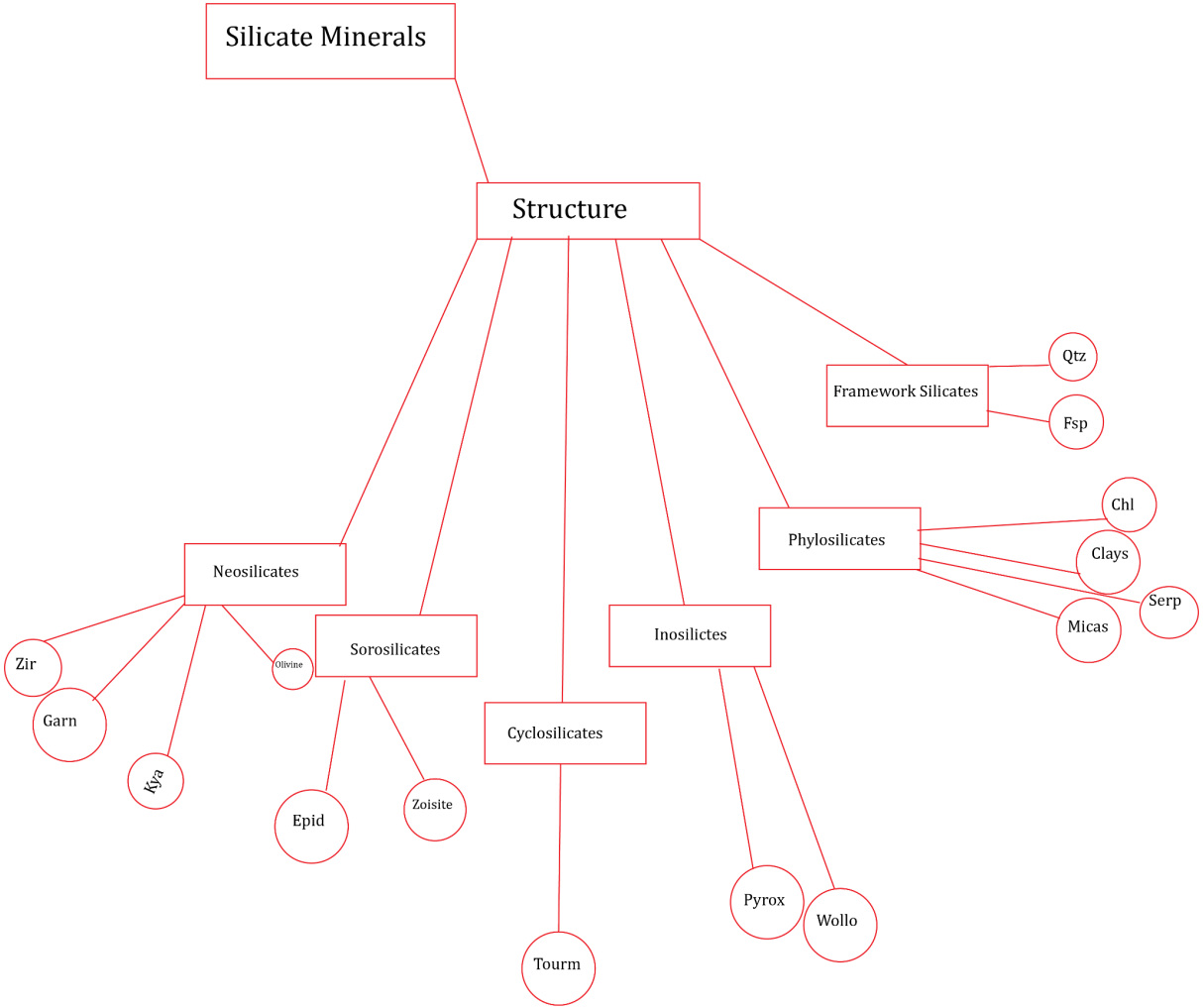


Benjamin Parfitt is a machine learning engineer and cofounder at Reiform, a startup focused on increasing the use of modern computational techniques in the sciences. Ben earned his B.A. in computer science and mathematics at Hamilton College (2019).

Robert Welch is a PhD student at Harvard University, focusing on structural geology as a part of the structure and earth resources group. Robert earned his B.A. in the geosciences at Hamilton College (2020).

Why Silicate Minerals?

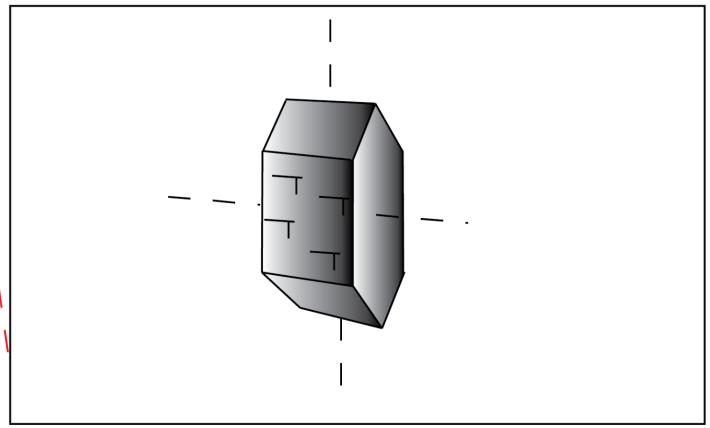


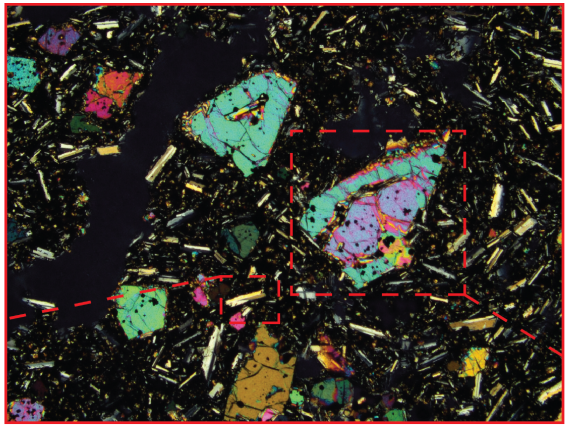


Why Silicate Minerals?

Silicate Minerals make up over 90% of minerals in the Earth's Crust

The Problem Space

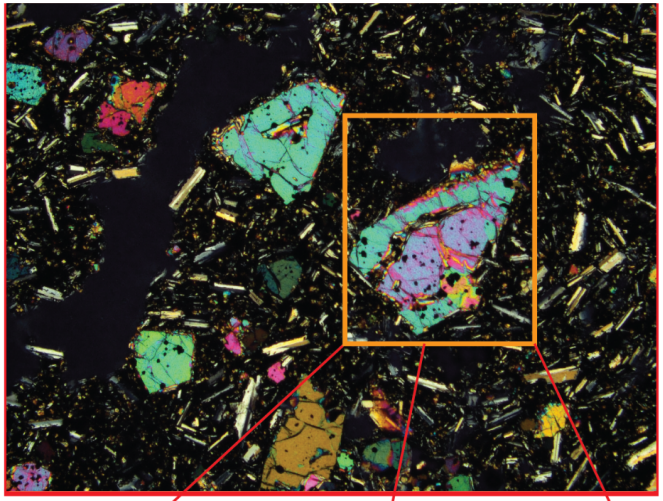




How to Identify a Mineral:

- Luster
- Specific Gravity
- Cleavage
- Hardness
- Symmetry
- Color
- Etc..

The Problem Space

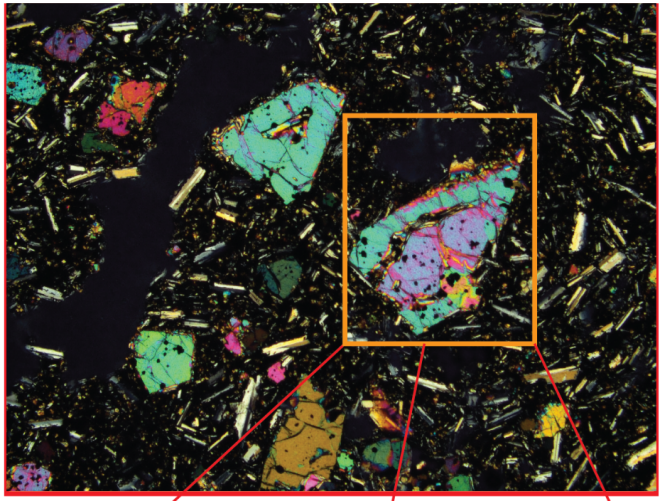


X-ray Fluoresence

Reported as
multi-spectral data

Reported as Percent
Weight Oxide

The Problem Space



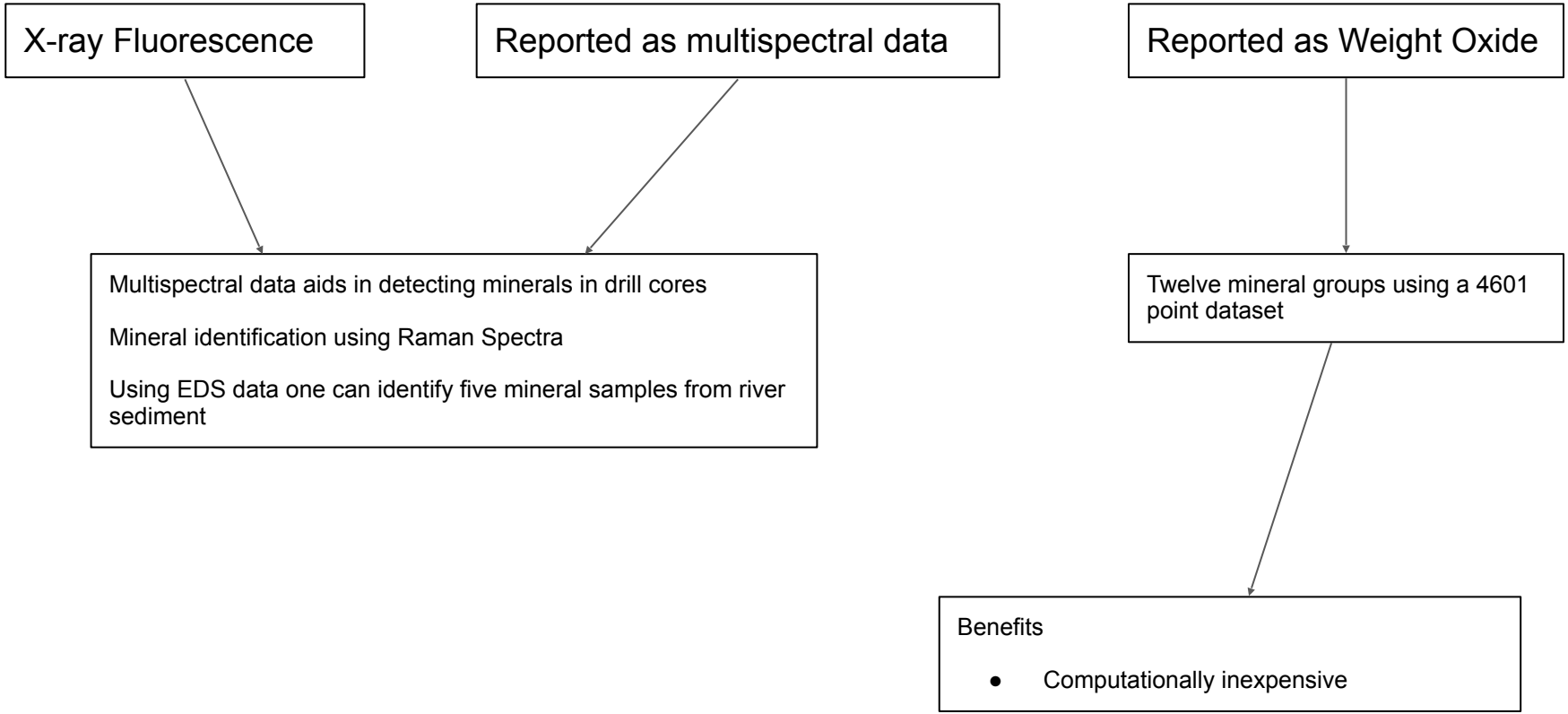
How can we identify a mineral quickly with just chemical data?

X-ray Fluoresence

Reported as multi-spectral data

Reported as Percent Weight Oxide

The Gap in Literature



The New Dataset

Earthchem

An NSF funded geochemical database

Additional Clay Mineral Data

Ross and Hendricks, 1943

Srodon et al., 2009

Our Dataset

- 6 Structural Groups
-20 Groups
-17 Subgroups

Dividing into Train, Validation, and Test sets

- Train: 75%
- Validation: 10%
- Test: 15%

Initial Approach

Initial Approach

- Train four different models for each task
  - Tasks:
    - Structural Family Identification
    - Mineral Group Identification
    - Mineral Subgroup Identification
  - Models:
    - Decision Tree
    - K-Nearest-Neighbors
    - Extremely Randomized Trees
    - Support Vector Machine

Initial Approach

- Train four different models for each task
  - Tasks:
    - Structural Family Identification
    - Mineral Group Identification
    - Mineral Subgroup Identification
  - Models:
    - Decision Tree
    - K-Nearest-Neighbors
    - Extremely Randomized Trees
    - Support Vector Machine
- Each of the models was validated using the validation set

Initial Approach

- Train four different models for each task
  - Tasks:
    - Structural Family Identification
    - Mineral Group Identification
    - Mineral Subgroup Identification
  - Models:
    - Decision Tree
    - K-Nearest-Neighbors
    - Extremely Randomized Trees
    - Support Vector Machine
- Each of the models was validated using the validation set
- The F1-score metric was used for comparisons

Initial Approach

- Train four different models for each task
  - Tasks:
    - Structural Family Identification
    - Mineral Group Identification
    - Mineral Subgroup Identification
  - Models:
    - Decision Tree
    - K-Nearest-Neighbors
    - Extremely Randomized Trees
    - Support Vector Machine
- Each of the models was validated using the validation set
- The F1-score metric was used for comparisons
- K-Nearest-Neighbors was best for each task

Results

Precision = True Positive/(True Positive + False Positive)
Recall = True Positive/(True Positive + False Negative)
F1 Score = 2 x (precision * recall) / (precision + recall)

Results

Precision = True Positive/(True Positive + False Positive)
Recall = True Positive/(True Positive + False Negative)
F1 Score = 2 x (precision * recall) / (precision + recall)

TABLE III
THE PRECISION, RECALL, AND F1-SCORE FOR THE BEST CLASSIFIER FOR EACH TASK FROM SUBGROUP, GROUP, AND STRUCTURE, AS INDICATED IN TABLE II ON THE TEST DATASET.

Metric	Subgroup (KNN)	Group (KNN)	Structure (KNN)
Precision	95.327	93.295	97.845
Recall	88.551	92.186	98.994
F1 Score	90.764	92.332	98.404

The New Technique

The New Technique

Take advantage of the explicit subdivisions that already exist

The New Technique

Take advantage of the explicit subdivisions that already exist

- Rather than force the model to learn the difference between all mineral groups at once, give them extra information regarding the structural family

The New Technique

Take advantage of the explicit subdivisions that already exist

- Rather than force the model to learn the difference between all mineral groups at once, give them extra information regarding the structural family
- Avoid a decision tree

The New Technique

Take advantage of the explicit subdivisions that already exist

- Rather than force the model to learn the difference between all mineral groups at once, give them extra information regarding the structural family
- Avoid a decision tree
- Process:
  a. Choose a high performing structure classifier S
  b. For each data point (vector) v, compute the probability vector $P_S(v)$
  c. Set v_1 to be equal to v concatenated with $P_S(v)$
  d. Let D be the set of all such v_1
  e. Train four new classifiers to identify the mineral group using the vectors from D
  f. Validate the new classifiers, and evaluated the best classifier on the Test set

Results (New Process)

TABLE VI

THE RESULTS OF EVALUATION ON THE VALIDATION DATA AFTER TRAINING THE SUBGROUP AND GROUP CLASSIFIERS ON THE DATASET AUGMENTED WITH THE HIGHEST TOP-3 RECALL FROM THE PREVIOUS CLASSIFIER. "CHANGE" INDICATES THE CHANGE IN ACCURACY FROM THE CLASSIFIERS TRAINED WITH THE ORIGINAL DATA TO THOSE TRAINED WITH THE AUGMENTED DATA. THE BEST RESULTS ARE IN **BOLD**.

	SubGroup (%)		Group (%)	
ML Algorithm	Macro F1	Change	Macro F1	Change
DecisionTree	89.286	0.858	88.860	−2.046
KNN	**91.107**	0.023	**92.249**	−0.029
ExtraTree	86.888	0.370	88.765	0.813
SVM	90.255	**3.411**	91.778	**3.358**

Results (New Process)

TABLE VI

THE RESULTS OF EVALUATION ON THE VALIDATION DATA AFTER TRAINING THE SUBGROUP AND GROUP CLASSIFIERS ON THE DATASET AUGMENTED WITH THE HIGHEST TOP-3 RECALL FROM THE PREVIOUS CLASSIFIER. "CHANGE" INDICATES THE CHANGE IN ACCURACY FROM THE CLASSIFIERS TRAINED WITH THE ORIGINAL DATA TO THOSE TRAINED WITH THE AUGMENTED DATA. THE BEST RESULTS ARE IN **BOLD**.

	SubGroup (%)		Group (%)	
ML Algorithm	Macro F1	Change	Macro F1	Change
DecisionTree	89.286	0.858	88.860	−2.046
KNN	**91.107**	0.023	**92.249**	−0.029
ExtraTree	86.888	0.370	88.765	0.813
SVM	90.255	**3.411**	91.778	**3.358**

TABLE VII

THE RESULTS OF RUNNING THE BEST CLASSIFIERS (AS INDICATED IN TABLE VI) ON THE TEST DATASETS AFTER AUGMENTING THEM WITH THE PROBABILITIES FROM THE CLASSIFIERS WITH THE HIGHEST TOP-3 RECALL. AVERAGE PRECISION, RECALL, AND F1-SCORE OVER ALL CLASSES ARE REPORTED.

	SubGroup (%)		Group (%)	
Metric	KNN	Change	KNN	Change
Precision	95.204	−0.123	92.057	−1.237
Recall	88.546	−0.005	90.462	−1.724
F1 Score	90.705	−0.060	90.302	−2.030

Results (Deeper Dive)

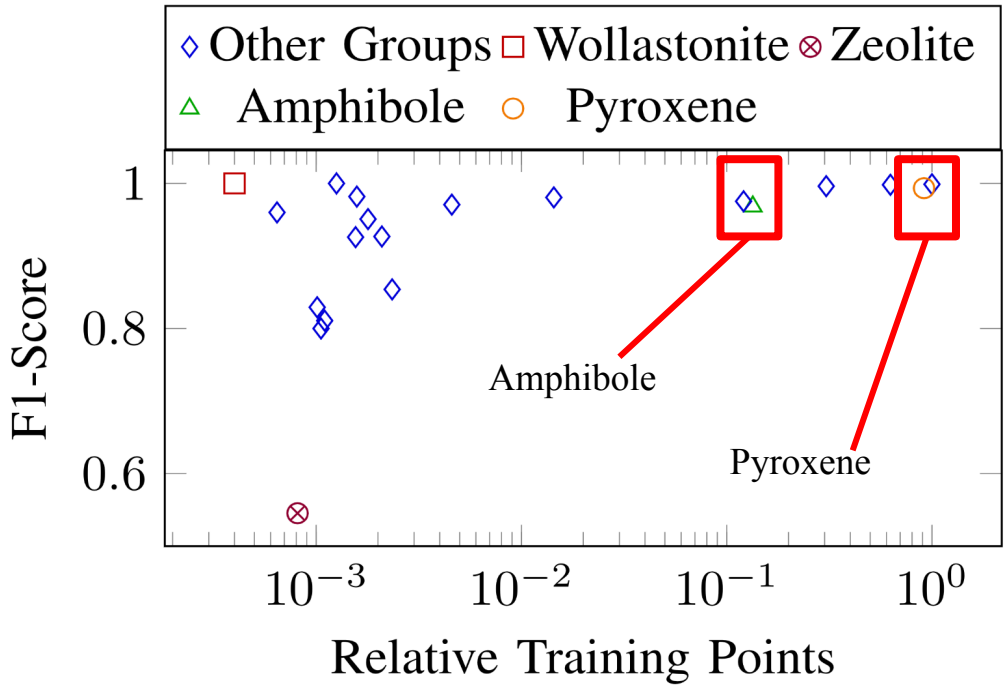
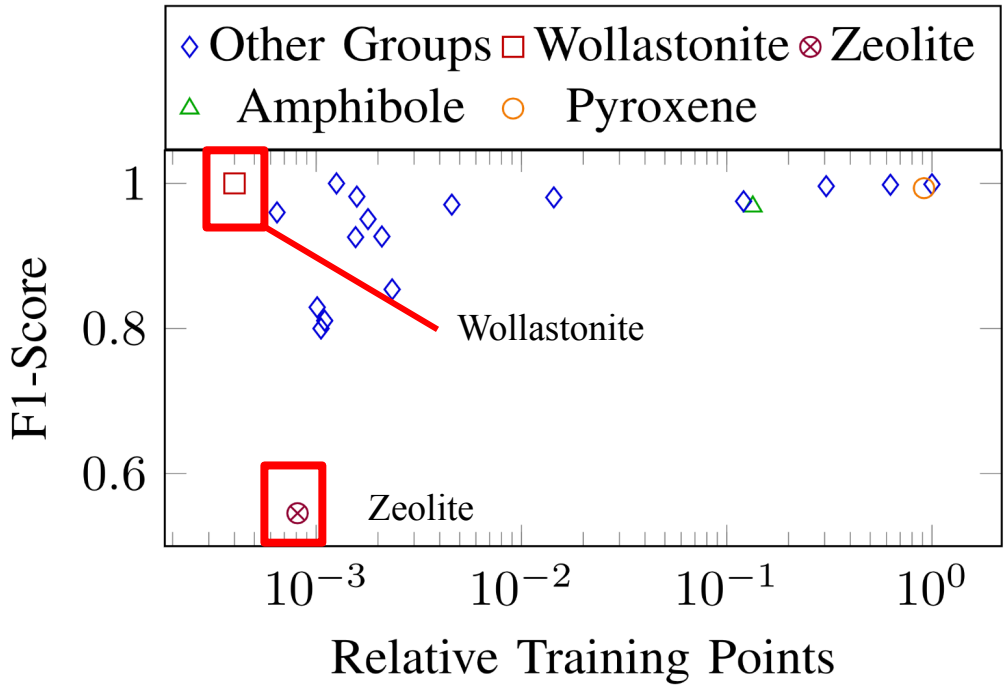


Fig. 1. The relative number of training data points per each group class (points in class/max(points in classes)) versus the F1-score of the best model for the group task (from Table I) on test data.

High Relative Training Point Takeaways:

- Both Amphibole and Pyroxene have a different structure but similar chemistry.
- Both minerals have high accuracy
- A decrease in accuracy is not solely due to similar chemistry

Results (Deeper Dive)

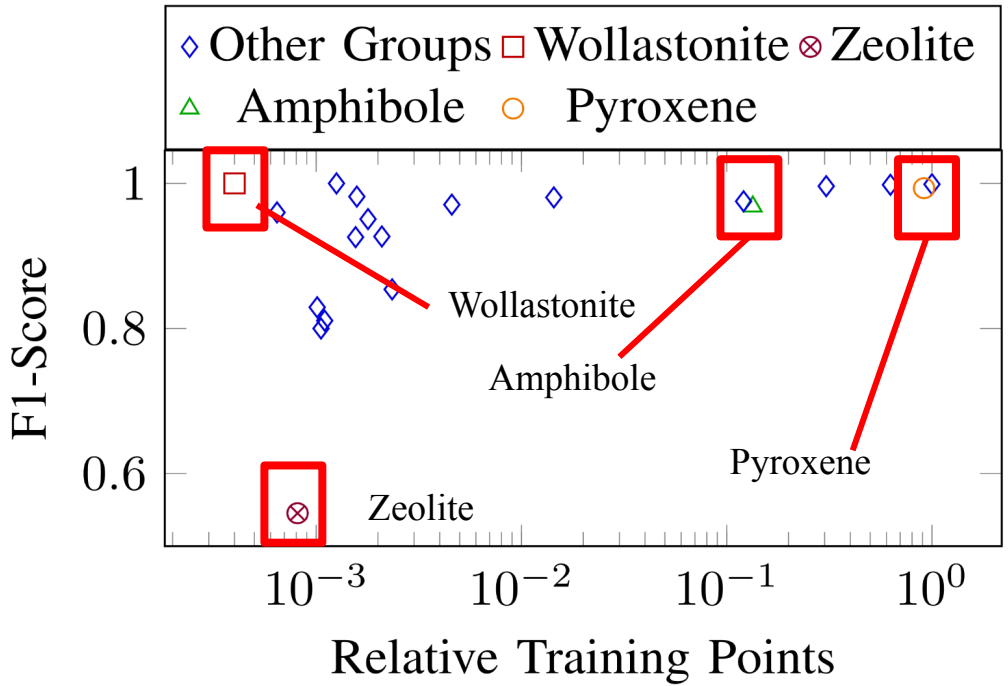


Low Relative Training Point Takeaways:

- Both Zeolite and Wollastonite have a different structure and chemistry.
- Both minerals have different accuracies.
- A decrease in accuracy is due to uniqueness

Fig. 1. The relative number of training data points per each group class (points in class/max(points in classes)) versus the F1-score of the best model for the group task (from Table I) on test data.

Results (Deeper Dive)



Takeaways:

- A decrease in accuracy is not solely due to similar chemistry.
- A decrease in accuracy is due to uniqueness (small relative training points).
- You can avoid accuracy decreases in non-unique samples if you have enough training points.

Fig. 1. The relative number of training data points per each group class (points in class/max(points in classes)) versus the F1-score of the best model for the group task (from Table I) on test data.

Results (Deeper Dive)

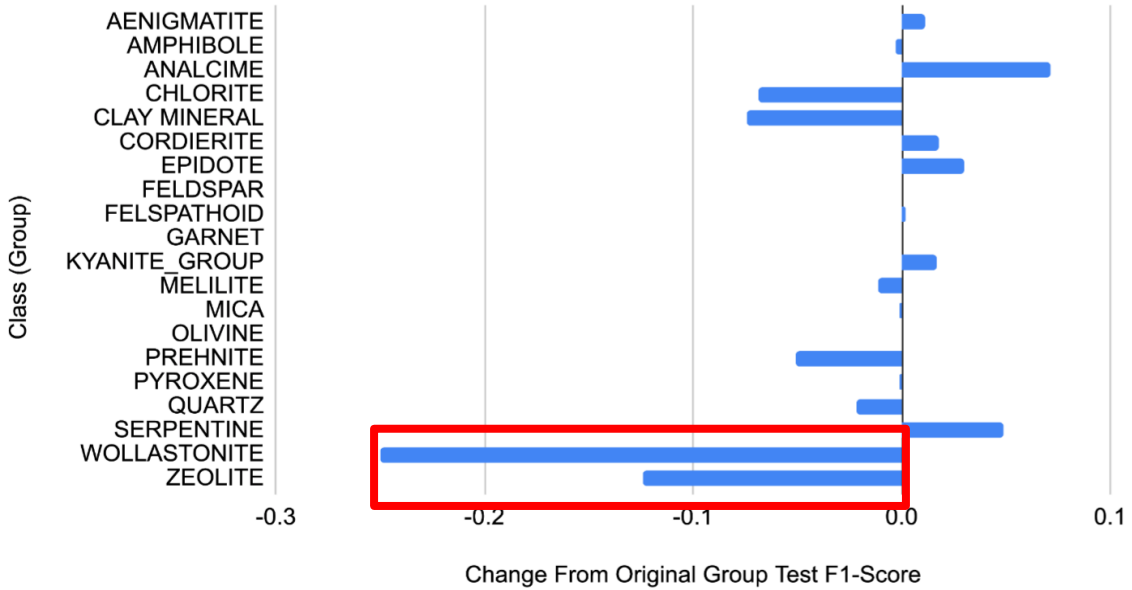


Fig. 2. The difference in classification F1-score from the original Group classifier to the Group classifier with structure data. Higher indicates better performance from the Group classifier with structure data.

Ensemble Learning Takeaways:

- The groups that experience a decrease in accuracy are those with a low count of relative training points
- Ensemble learning decreases the uniqueness of minerals which does not increase the accuracy of those with a low count of relative training points.

Future Work

Future Work

- More Ensemble Techniques

Future Work

- More Ensemble Techniques
- Dimensionality Reduction
  - Principal Component Analysis
  - Linear Discriminant Analysis
  - Simpler Techniques

Future Work

- More Ensemble Techniques
- Dimensionality Reduction
  - Principal Component Analysis
  - Linear Discriminant Analysis
  - Simpler Techniques
- More Datasets (currently obtaining more clay mineral data)

Future Work

- More Ensemble Techniques
- Dimensionality Reduction
  - Principal Component Analysis
  - Linear Discriminant Analysis
  - Simpler Techniques
- More Datasets (currently obtaining more clay mineral data)
- Synthetic Data

Accessing The Models (Openly)

Accessing The Models (Openly)

Use of the models is available at mindicator.reiform.com for free

Accessing The Models (Openly)

Use of the models is available at mindicator.reiform.com for free

- Currently only the three models for the original method, not the new technique

Accessing The Models (Openly)

Use of the models is available at mindicator.reiform.com for free

- Currently only the three models for the original method, not the new technique
- More coming soon

Accessing The Models (Openly)

Use of the models is available at mindicator.reiform.com for free

- Currently only the three models for the original method, not the new technique
- More coming soon
- Benefits:
  - No downloads required
  - No programming knowledge required
  - We do **not** steal your data :-)

Accessing The Models (Openly)

Use of the models is available at mindicator.reiform.com for free

- Currently only the three models for the original method, not the new technique
- More coming soon
- Benefits:
  - No downloads required
  - No programming knowledge required
  - We do **not** steal your data :-)

All code will be posted as well for download if interested