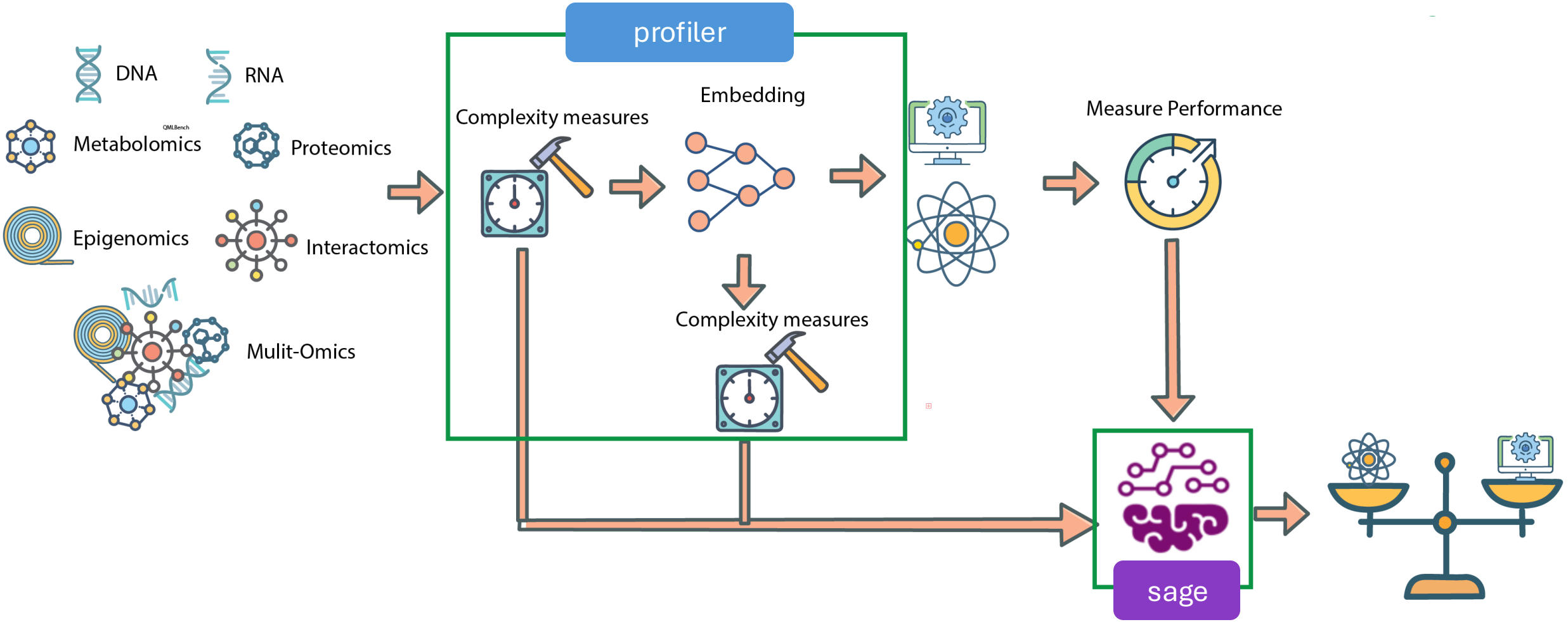
qml4omics: A Quantum-Classical Machine Learning Benchmarking tool for multi-omics data

qml4omics


DNA
RNA
Metabolomics
Proteomics
Epigenomics
Interactomics
Mulit-Omics
profiler
Complexity measures
Embedding
Complexity measures
Measure Performance
sage

qml4omics: Data Complexities

Dimensional

- Intrinsic Dimension (Rank)
- Manifold (Fractal Dimension)
- Volume
- Effective rank
- Eigenspectra

Distributional

- Kurtosis & Skewness
- Mutual Information
- Sparsity
- Entropy
- Condition Number

Geometric

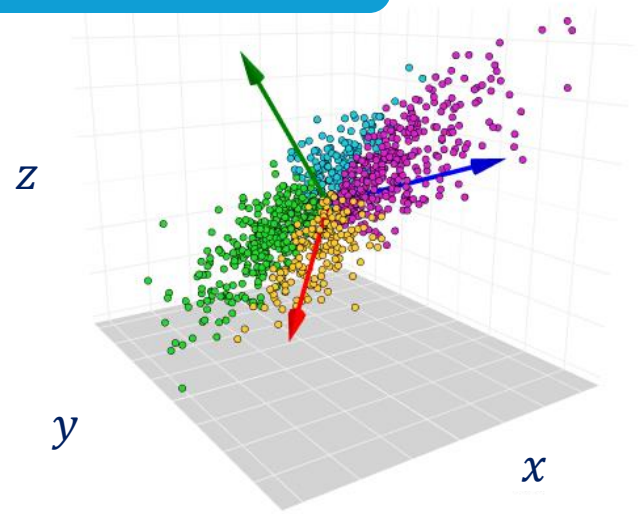
- Manifolds
- Clusters
- Density
- Topological Data Analysis
- Graph-based measures

Sampling

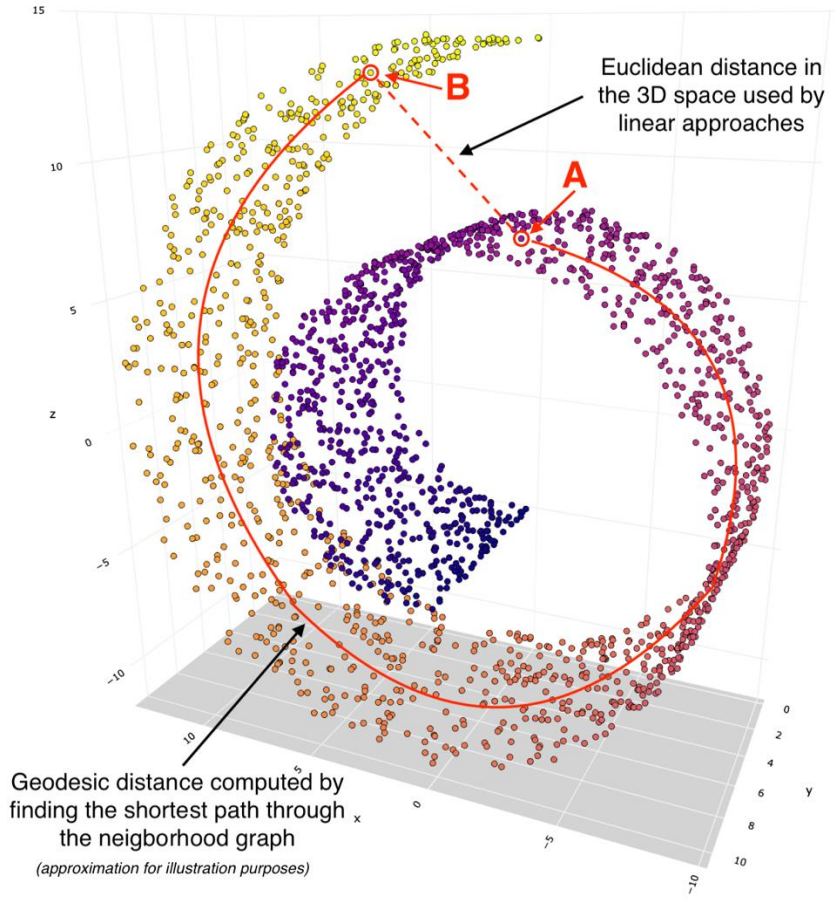
- Class imbalance ratio
- Class overlap measures
- Margin of separation between classes
- Sampling density variation

qml4omics: Embeddings

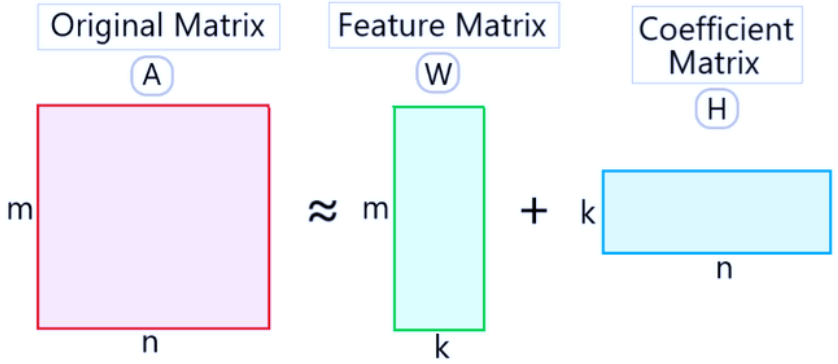
Principal Component Analysis (PCA)



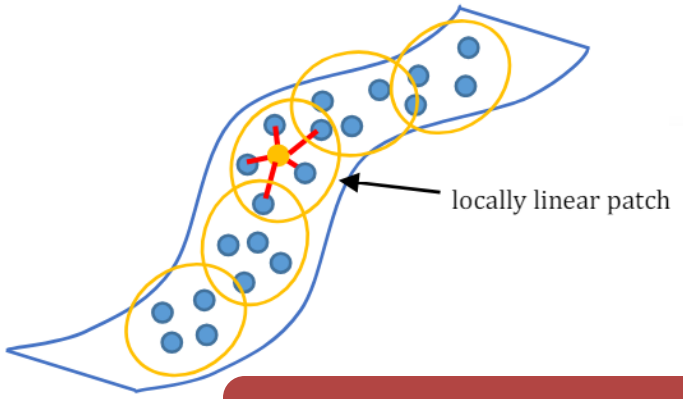
Spectral



Isomap

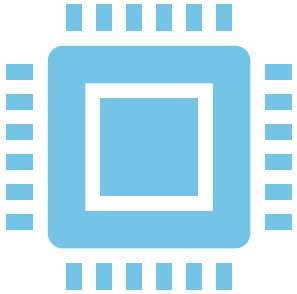


Non-negative Matrix Factorization (NMF)

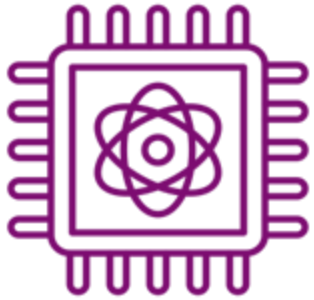


Locally Linear Embedding (LLE)

qml4omics: Models



- Logistic Regression
- Support Vector Classifiers
- Naïve Bayes
- Random Forest
- XGBoost
- Multi-layer Perceptron

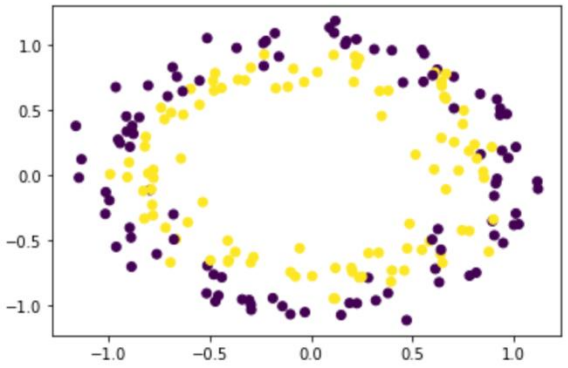


- Quantum Kernel Estimation
- Projected Quantum Kernel
- Quantum Support Vector Classifiers
- Variational Quantum Classifier / Quantum Neural Networks

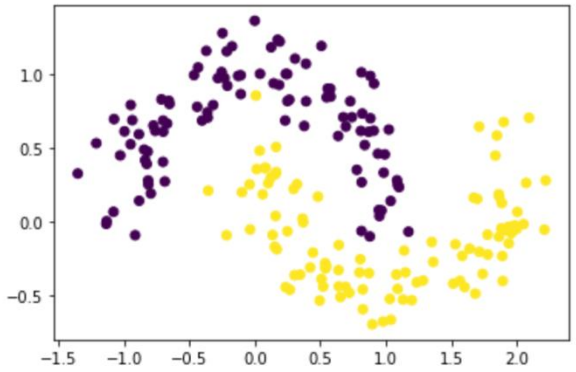
Artificial Data Generation

- To diversify datasets, we developed functions to generate artificial data based on user-defined combinations of data characteristic.
- These modules generate blobs, moons, circles, spheres, spirals.
- Located inside /data/artificial_datasets/make_X

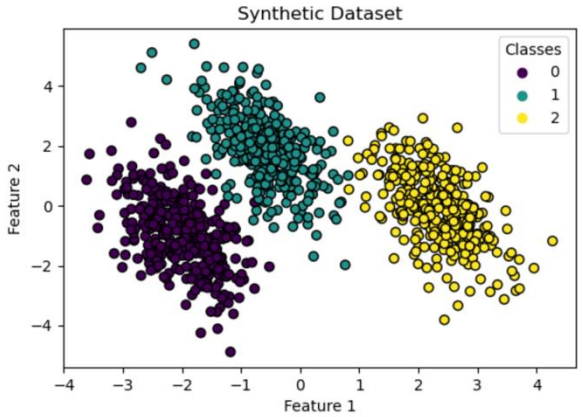
blobs



Circles



Half moons

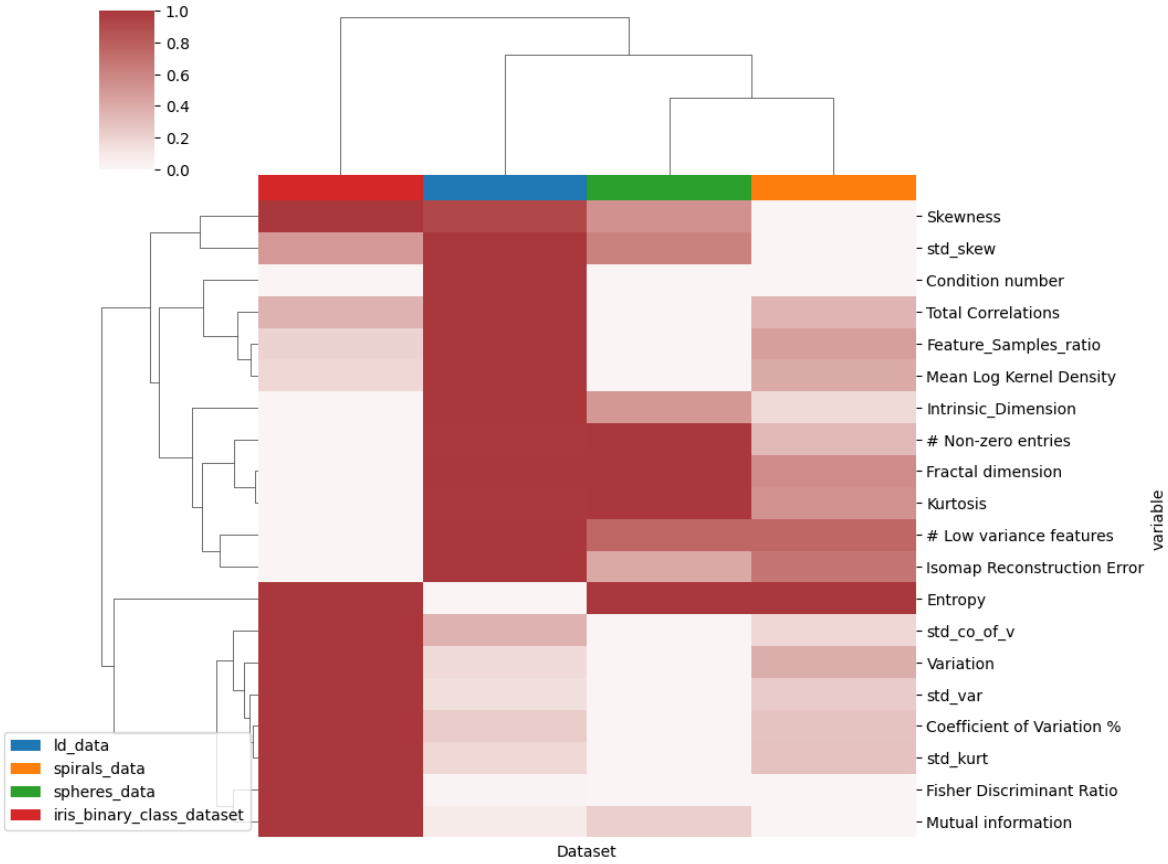


Blobs

Understanding the analyses

Hierarchical clustering heat maps

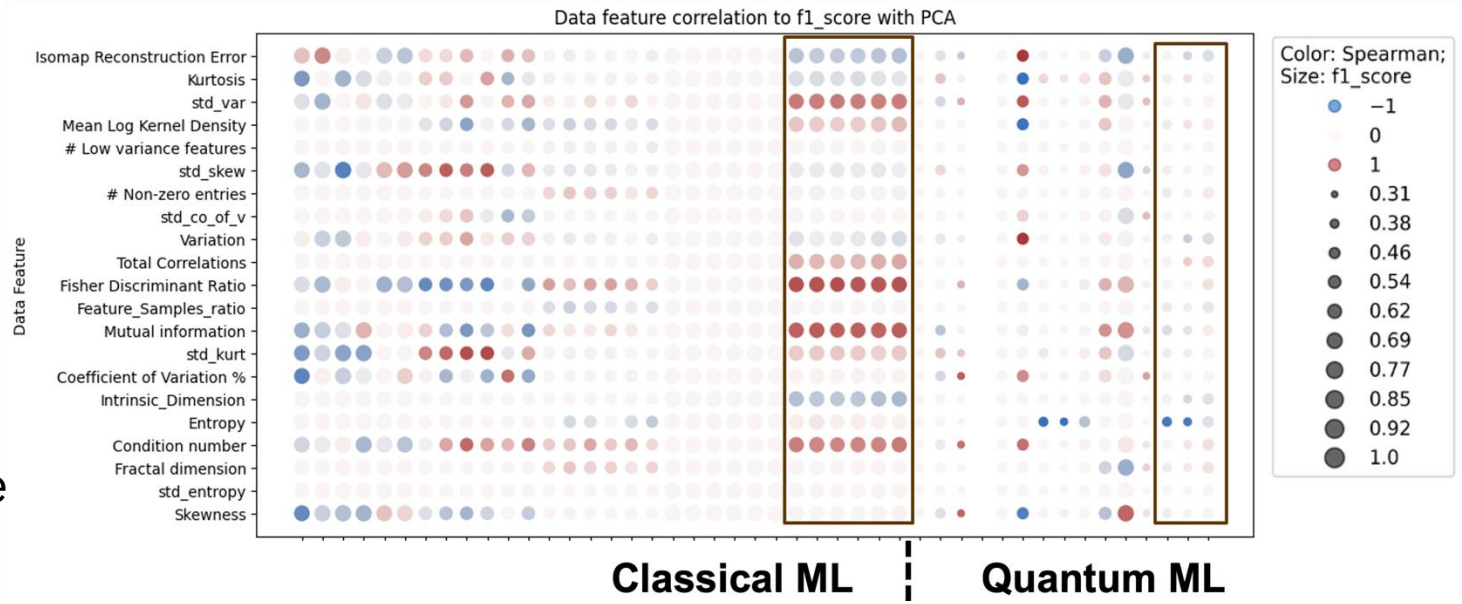
- What it's doing here:
  - Complexity measure range is normalized between 0 and 1.
  - Euclidean distance is calculated between columns and rows, clustering together those with the shortest distance → similar intensities for complexity measures.
  - The dendrogram branches create a pairing hierarchy.
  - Outlier has longest branch.
  - Helps answer:
    Is there some structure or pattern in my data?



Understanding the analyses

Spearman Rank Correlations

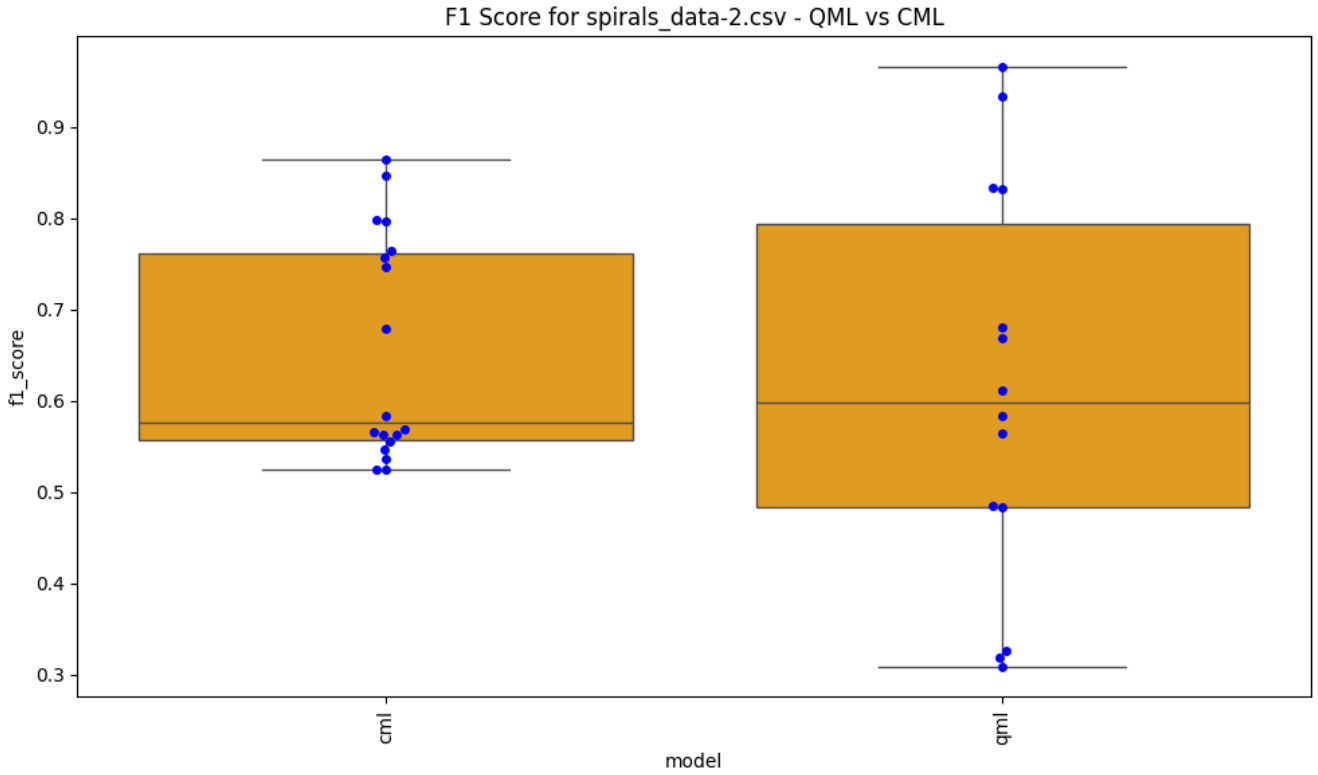
- What it's doing here:
  - Correlates data complexity measure to model performance (F1-score)
  - Red = positive correlated
  - Blue = anti-correlated
  - Size of sphere = magnitude of F1-score
  - Helps answer:
    What complexity measures influence your model score the most?



Understanding the analyses

Box-and-whisker plots

- What it's doing here:
  - Plots distribution of median F1-scores per datasets, across all splits of data, per model
  - Top and bottom of box = upper and lower quartiles (Q3 and Q1)
  - Whiskers denote range in F1- scores
  - Helps answer:
    What is the locality, spread, and skewness groups in my data (F1-scores) based on their quartiles?



Understanding QSage

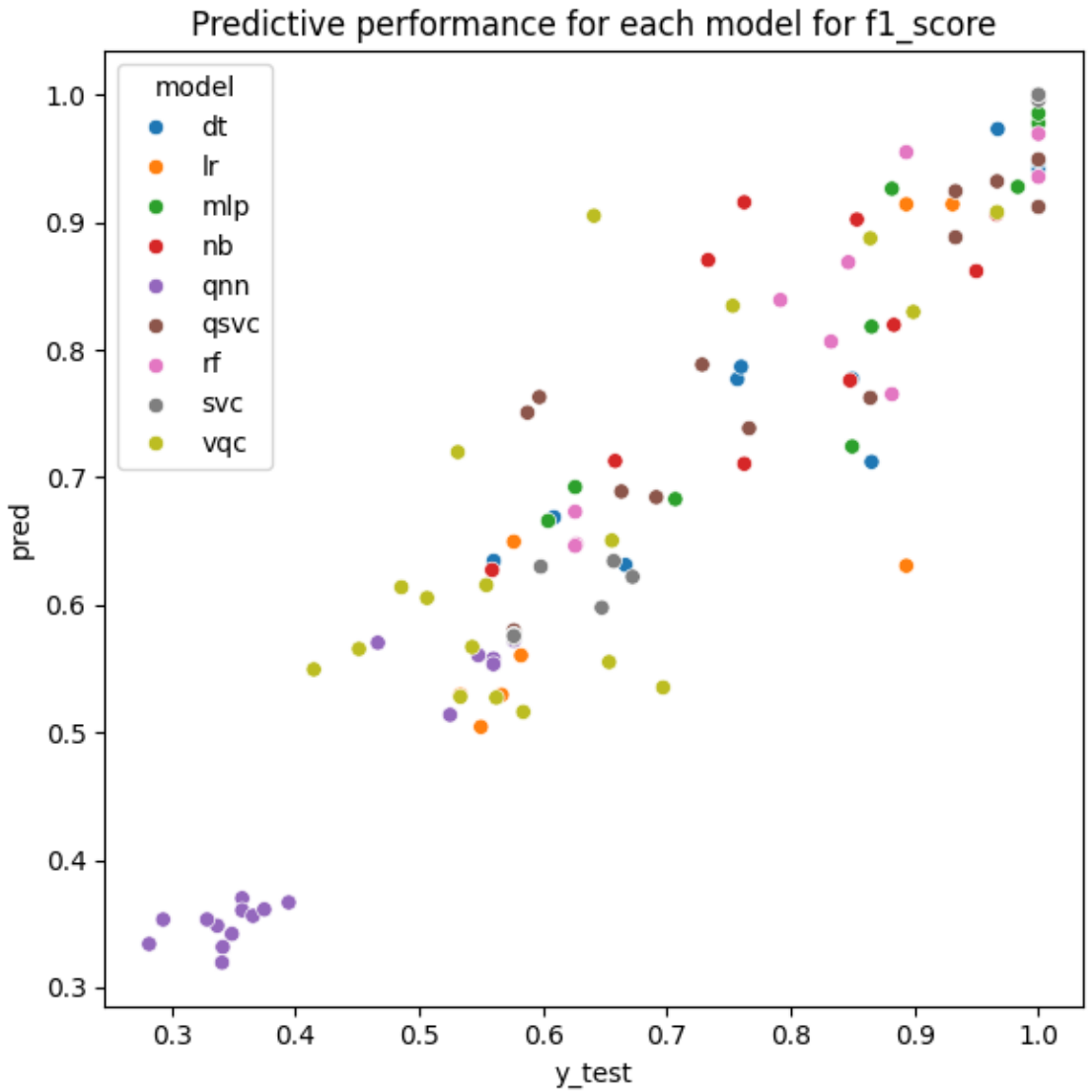
So, what is the big picture?

What do we do with all this stuff?

- *What if I were to tell you, what ML method to use, just by looking at your data?*

- We trained a new model on all of these correlations --> QSage

Predicts F1, AUC, and accuracy beforehand --> no need to run all model!

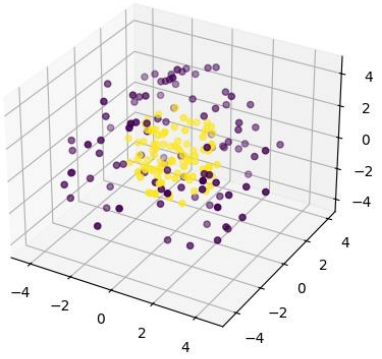


Examples

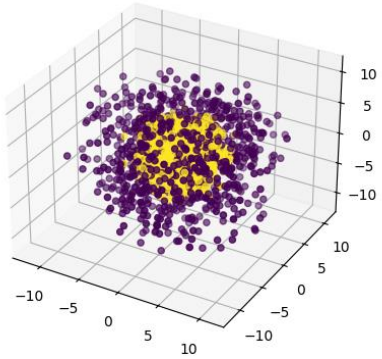
Examples: geometric shapes

- Let's look at higher dimensional artificial, geometric data sets (3D and beyond).
- Task – generate QML and CML models for these and compare performance.
- This data is periodic – can QML do well with these

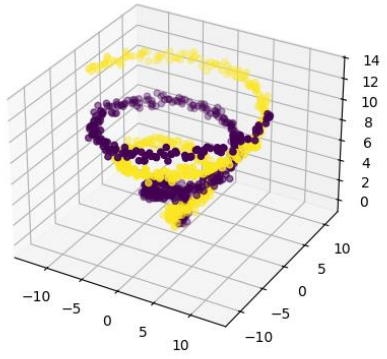
5 qubits/features



10 qubits/features

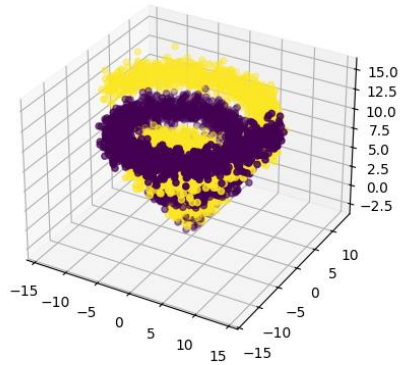


Low dimension and # of samples



3 qubits/features

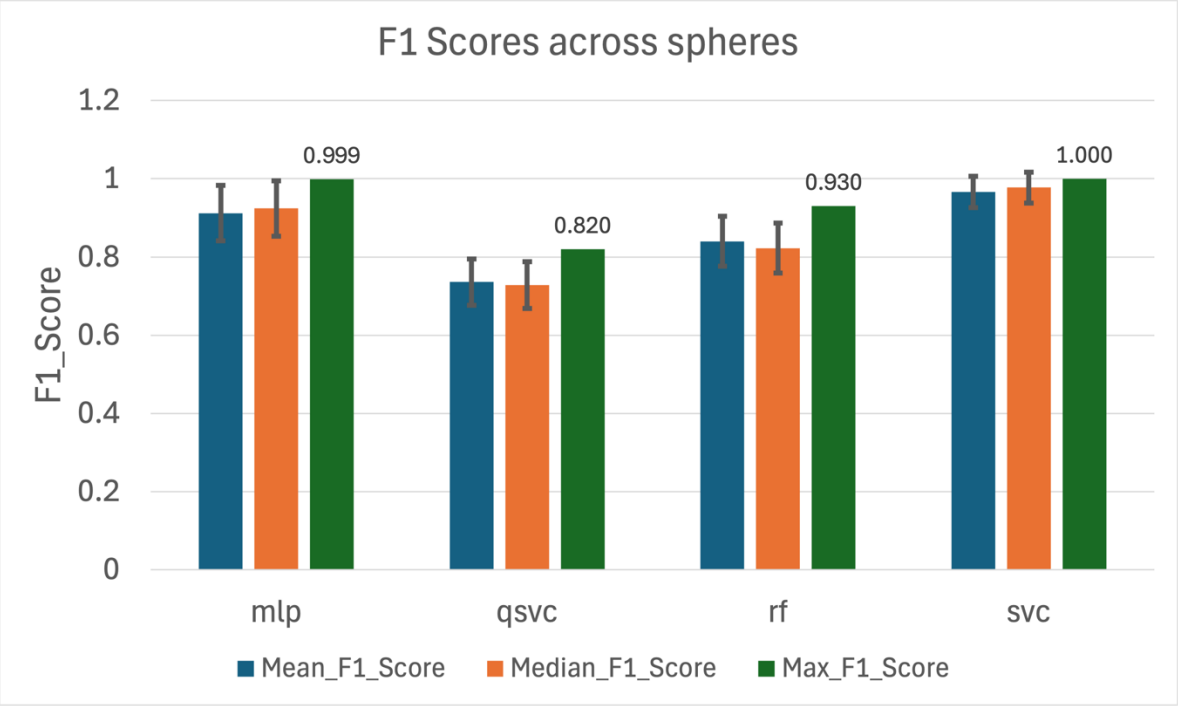
High dimension and # of samples

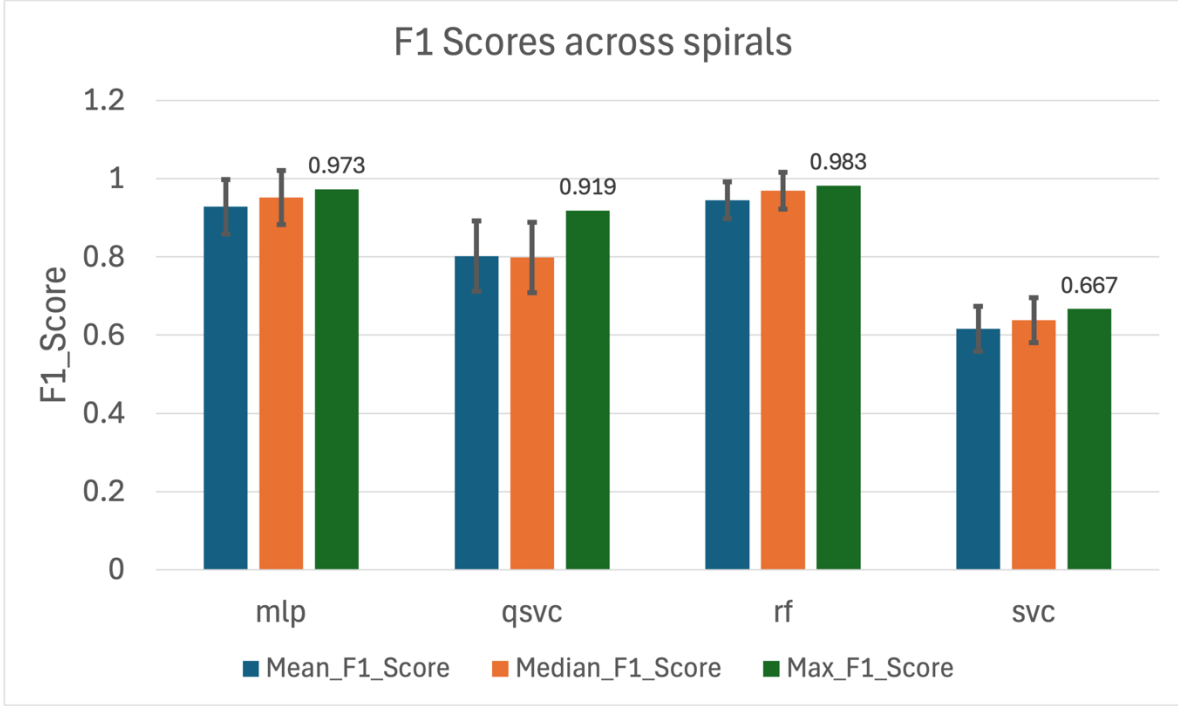


12 qubits/features

Examples: spheres and spirals

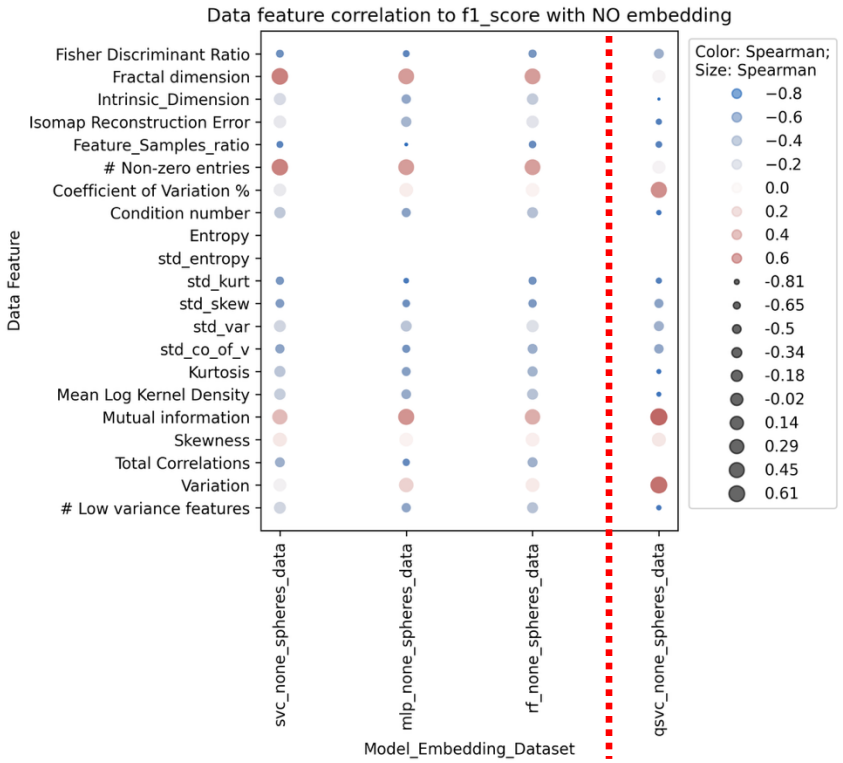
- Spirals seem QML friendly.
- SVC>QSVC with spheres, but it flips to QSVC>SVC with spirals
- RF improves with spheres, MLP is consistent across both



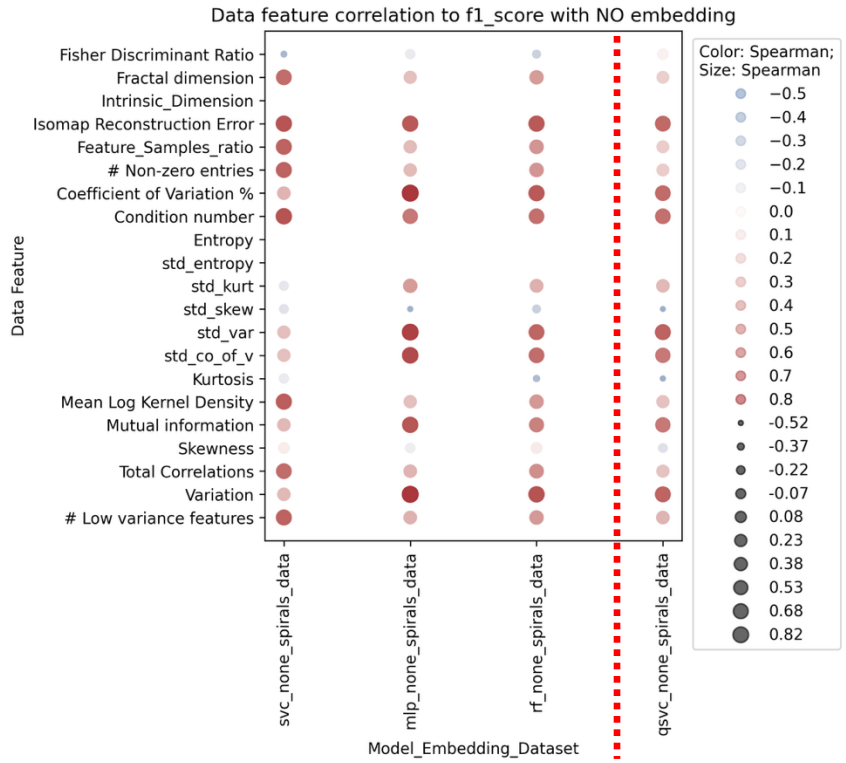


Examples: spheres and spirals

- Spheres: clear switch with Intrinsic dimension, Coeff of variance, and total correlations
- Spirals: correlation type switches (red vs blue) for Fischer Discrimination Ratio (measures imbalance) between CML and QSVC



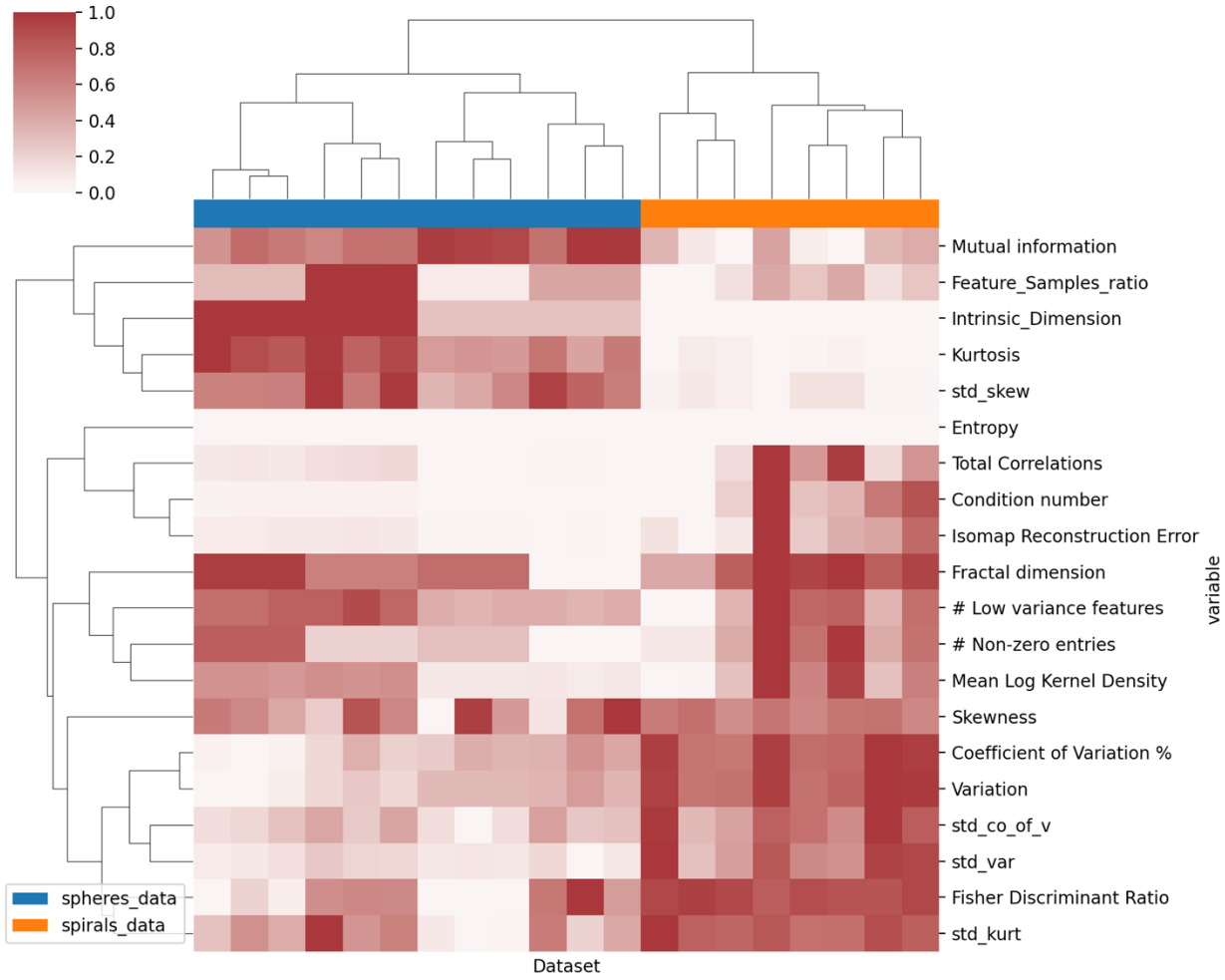
Spheres



Spirals

Examples: spheres and spirals

- Remember: on average QSVC>SVC with spirals.
- So, what is it about the spirals?
- There is a rather obvious disparity in a few areas



Let’s take it for a ride

So, let's try it out!

1) Go over the `config.yaml` file and learn how to change parameters

2) Activate your environment
   - If you haven't done so, set up your environment now following the instructions in the README.md

3) Run the main code. In your terminal, type:
   ```
   python qml4omics-profiler.py --config-name=config.yaml
   ```

4) Wait and watch the progress outputs being printed out.

5) We'll analyze the results and run your own data in the afternoon 😎.