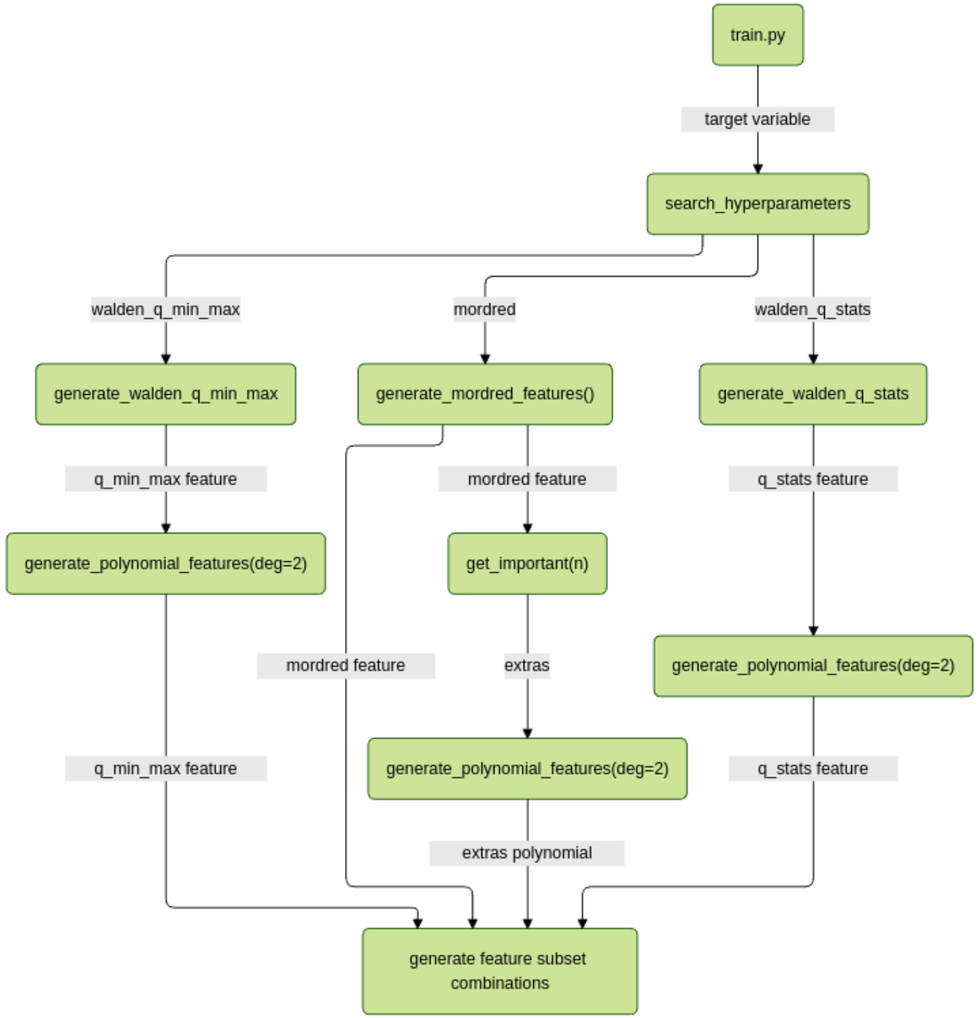
GENERALIZATION IN HIGH-D MOLECULE SPACE

CONTENTS

1. POLYNOMIAL & WALDEN FEATURE GENERATION



XGBOOST

Interactive booster tree (1/73)

XGBOOST TREE 0

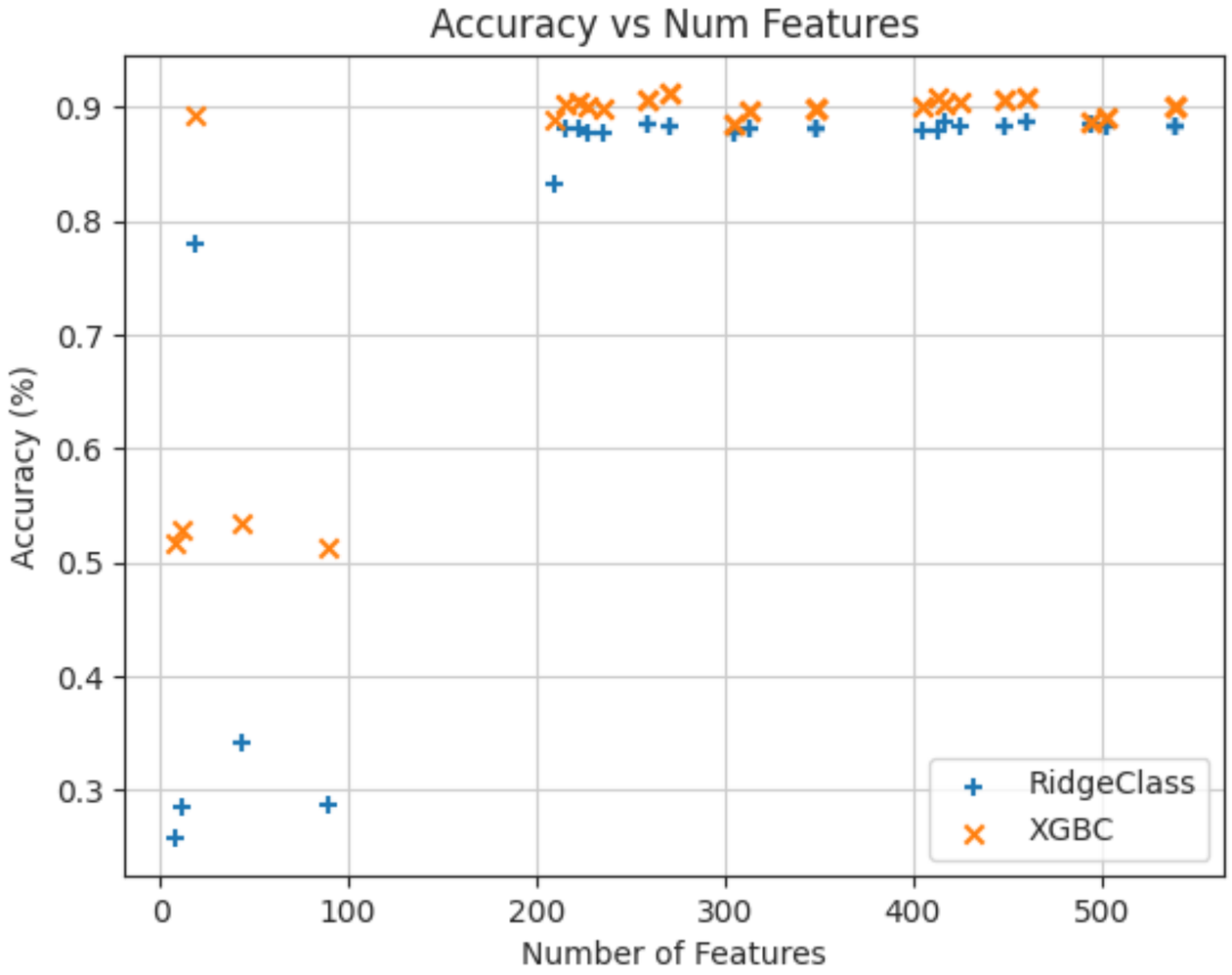
	Tree	Node	ID	Feature	Split	Yes	No	Missing	Gain	Cover	Category
0	0	0	0-0	peoe_vsa3	26.602526	0-1	0-2	0-2	223.890350	925.333252	NaN
1	0	1	0-1	maxpartialcharge	0.261492	0-3	0-4	0-4	139.774567	833.777710	NaN
2	0	2	0-2	minpartialcharge	-0.479242	0-5	0-6	0-6	14.453644	91.555550	NaN
3	0	3	0-3	maxpartialcharge	0.122978	0-7	0-8	0-8	31.111345	183.555542	NaN
4	0	4	0-4	maxabspartialcharge	0.464026	0-9	0-10	0-10	30.901764	650.222168	NaN
...	...	...	...	...	...	...	...	...	...	...	...
68	0	68	0-68	Leaf	NaN	NaN	NaN	NaN	-0.171784	25.777777	NaN
69	0	69	0-69	Leaf	NaN	NaN	NaN	NaN	-0.219193	502.666626	NaN
70	0	70	0-70	Leaf	NaN	NaN	NaN	NaN	-0.024324	3.111111	NaN
71	0	71	0-71	Leaf	NaN	NaN	NaN	NaN	0.128571	1.333333	NaN
72	0	72	0-72	Leaf	NaN	NaN	NaN	NaN	0.440586	78.666664	NaN

73 rows × 11 columns

XGBOOST NODE STATUS

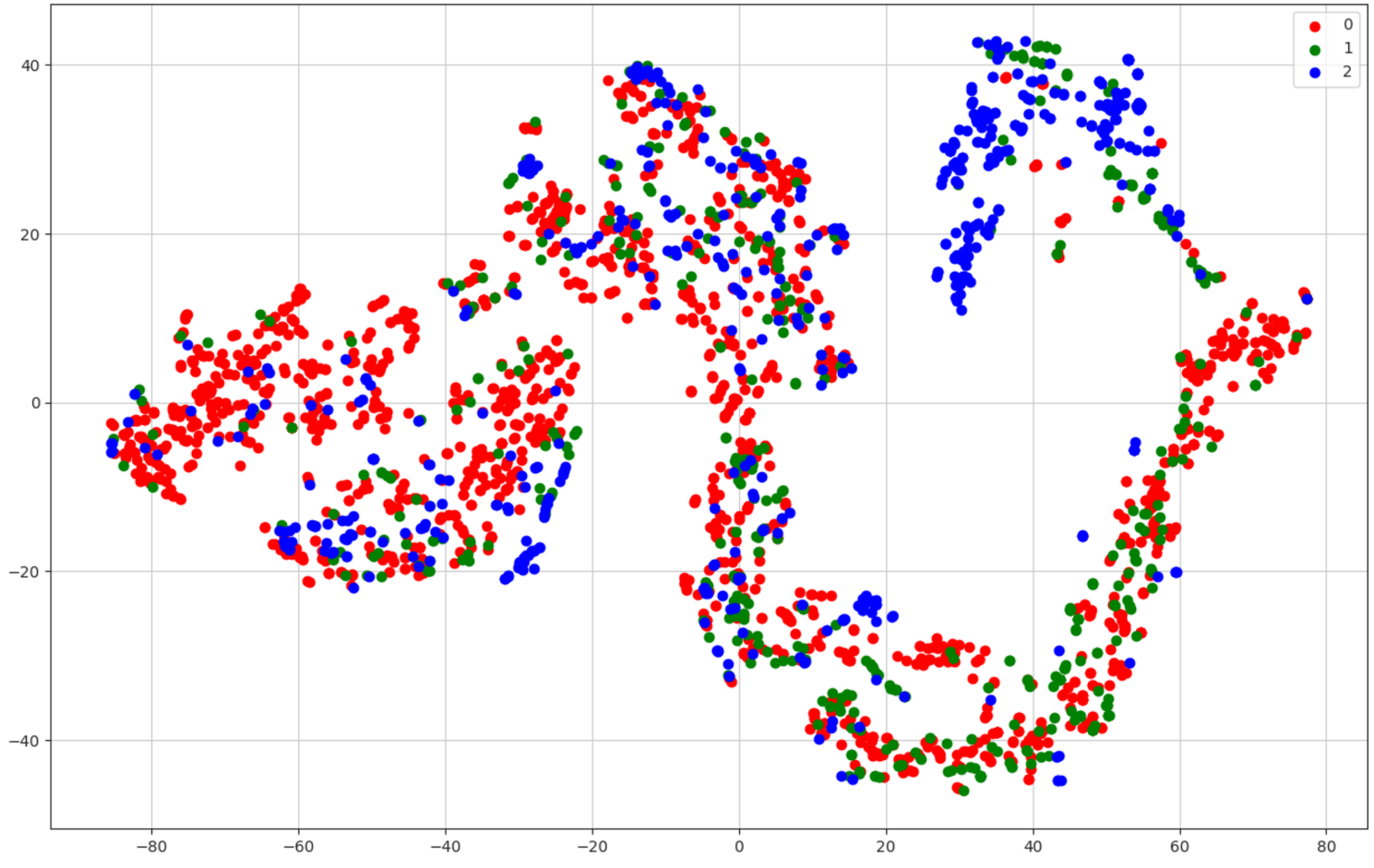
	Tree	Node	ID	Feature	Split	Yes	No	Missing	Gain	Cover	Category
count	73.0	73.000000	73	73	36.000000	36	36	36	7.300000e+01	73.000000	0.0
unique	NaN	NaN	73	27	NaN	36	36	36	NaN	NaN	NaN
top	NaN	NaN	0-72	Leaf	NaN	0-1	0-2	0-2	NaN	NaN	NaN
freq	NaN	NaN	1	37	NaN	1	1	1	NaN	NaN	NaN
mean	0.0	36.000000	NaN	NaN	5.078631	NaN	NaN	NaN	8.726143e+00	88.298319	NaN
std	0.0	21.217131	NaN	NaN	9.427902	NaN	NaN	NaN	3.096659e+01	191.931901	NaN
min	0.0	0.000000	NaN	NaN	-4.317507	NaN	NaN	NaN	-2.191926e-01	1.333333	NaN
25%	0.0	18.000000	NaN	NaN	0.169036	NaN	NaN	NaN	-7.663455e-09	3.111111	NaN
50%	0.0	36.000000	NaN	NaN	1.839286	NaN	NaN	NaN	3.557739e-01	16.444443	NaN
75%	0.0	54.000000	NaN	NaN	5.460695	NaN	NaN	NaN	5.472839e+00	67.111107	NaN
max	0.0	72.000000	NaN	NaN	40.459999	NaN	NaN	NaN	2.238904e+02	925.333252	NaN

RIDGE POLYNOMIAL MODEL VIABLE IN HIGH-D



T-SNE - 270D -> 2D

Explorable 2D projection of molecule space? Predict and display accuracy/confidence "volume" as area? Predict unlabeled molecules as part of confidence volume map?


0
1
2
40
20
0
−20
−40
−80
−60
−40
−20
0
20
40
60
80

GENERALIZATION

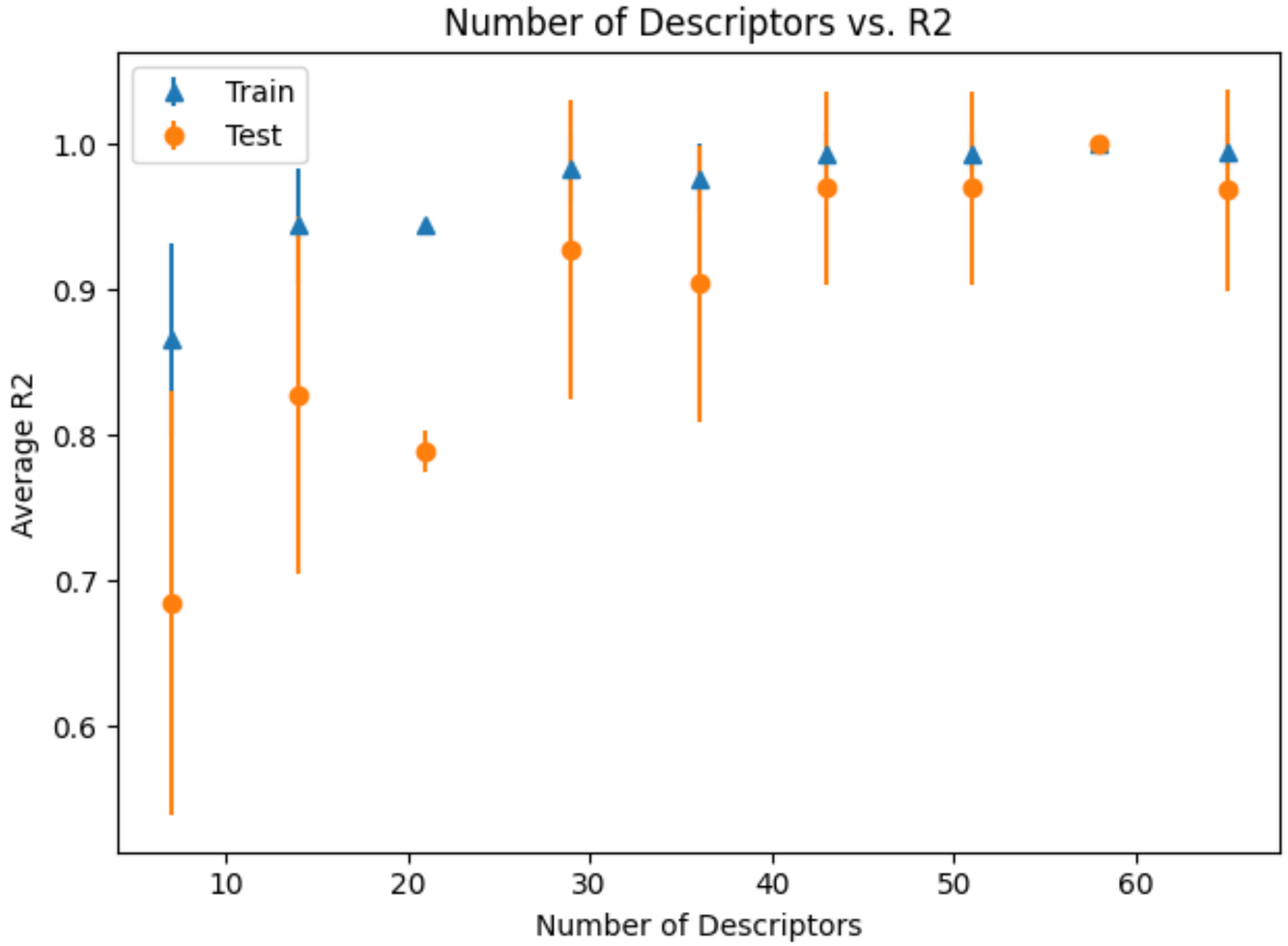
Is it possible to generalize to **more** unseen data?

Can we measure how much better we are doing?

2. CURSE OF DIMENSIONALITY:

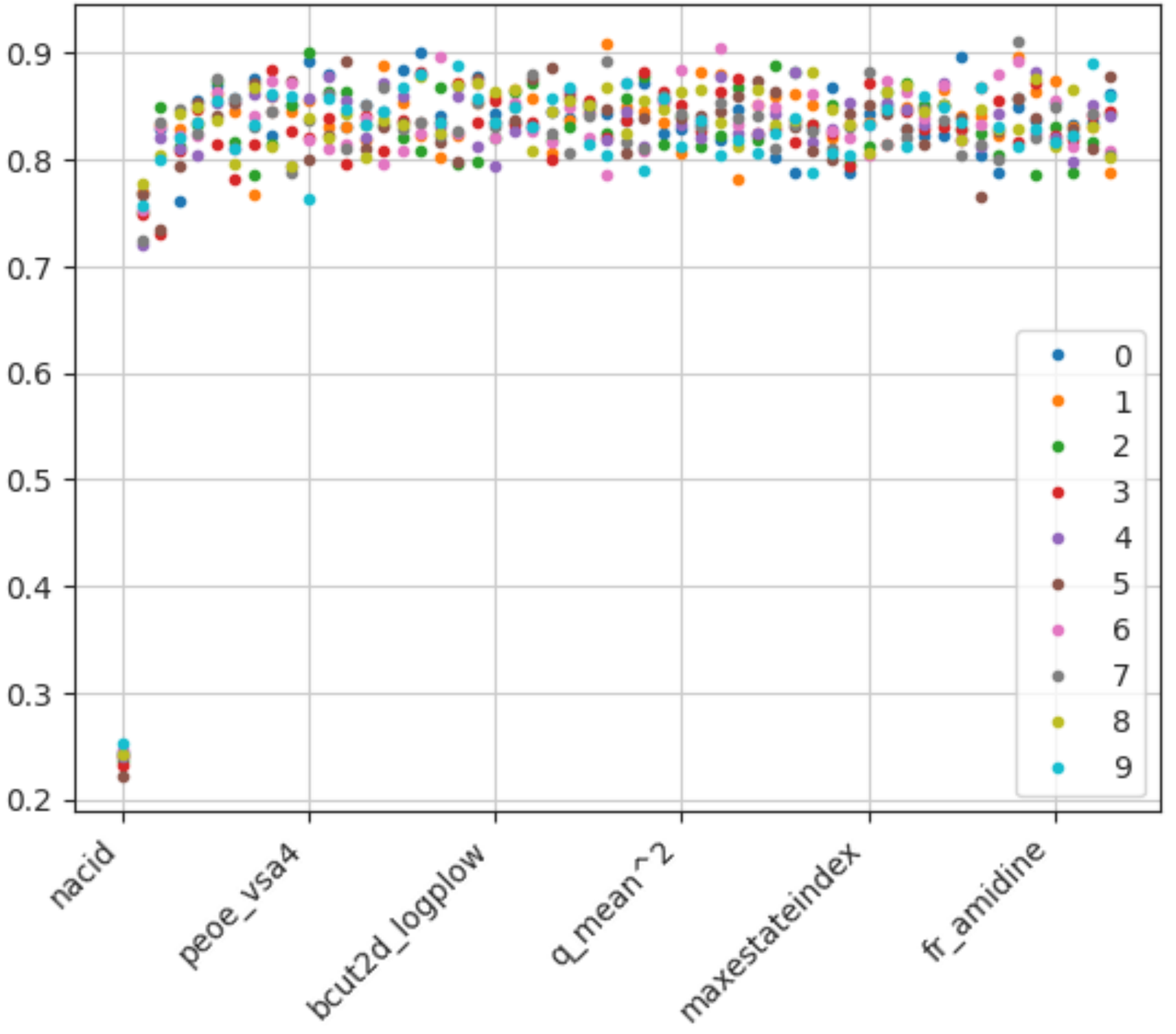
- Scott, Krishna research
- Everything in molecule space is far apart
- **65%** of molecules are singletons
- Impractical to create combinatoric features
- "Library of Mendel" vast and vanishing differences between high-D vectors
- Physics/chemistry creates latent patterns in high dimensional space difficult to predict

SCOTT'S SINGLETON OVERFITTING PLOT

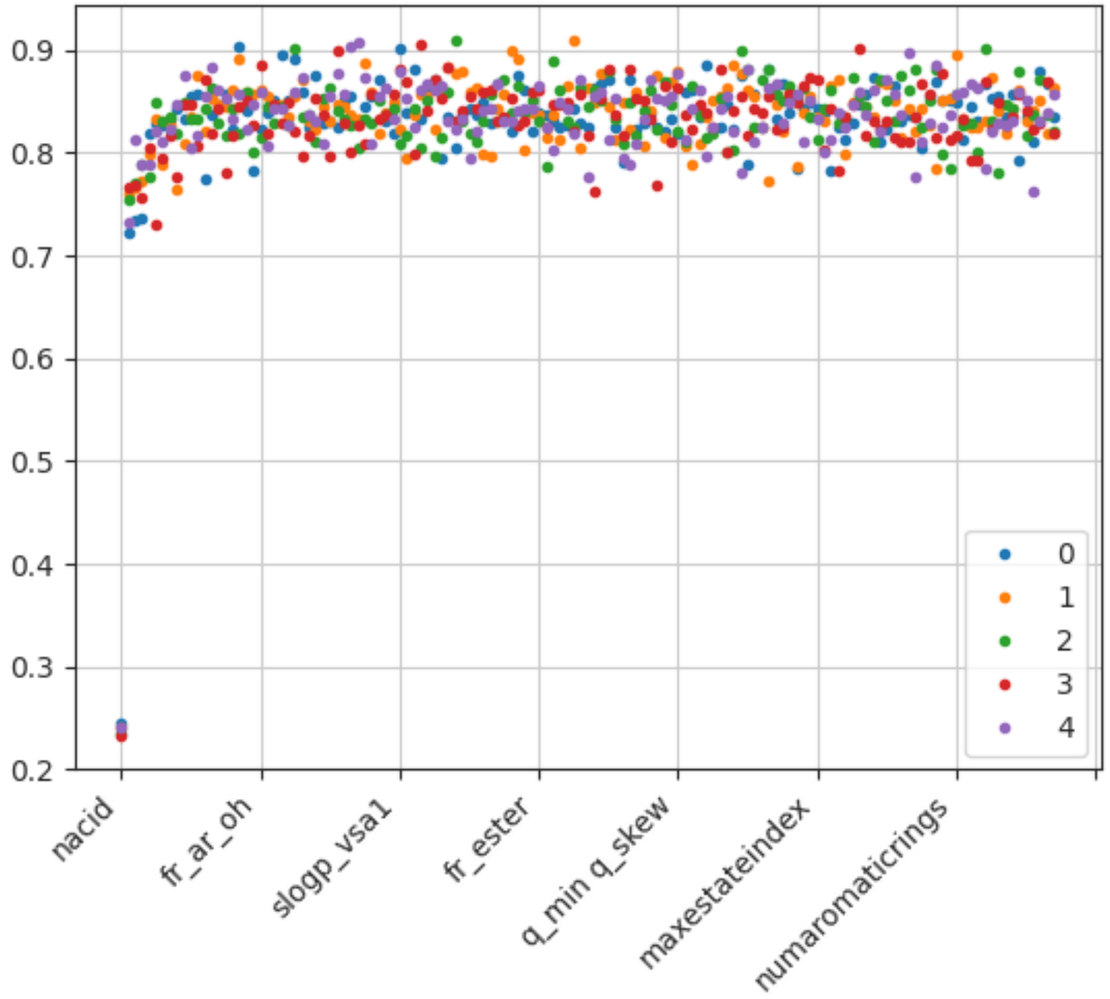


KRISHNA ELBOW PLOT - MACRO F1 SCORE VS NUM FEATURES

(XGBC, 10-fold shuffled CV)


0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
nacid
peoe_vsa4
bcut2d_logplow
q_mean^2
maxestateindex
fr_amidine
0
1
2
3
4
5
6
7
8
9

HIGH VARIANCE - 10% SAMPLE CV



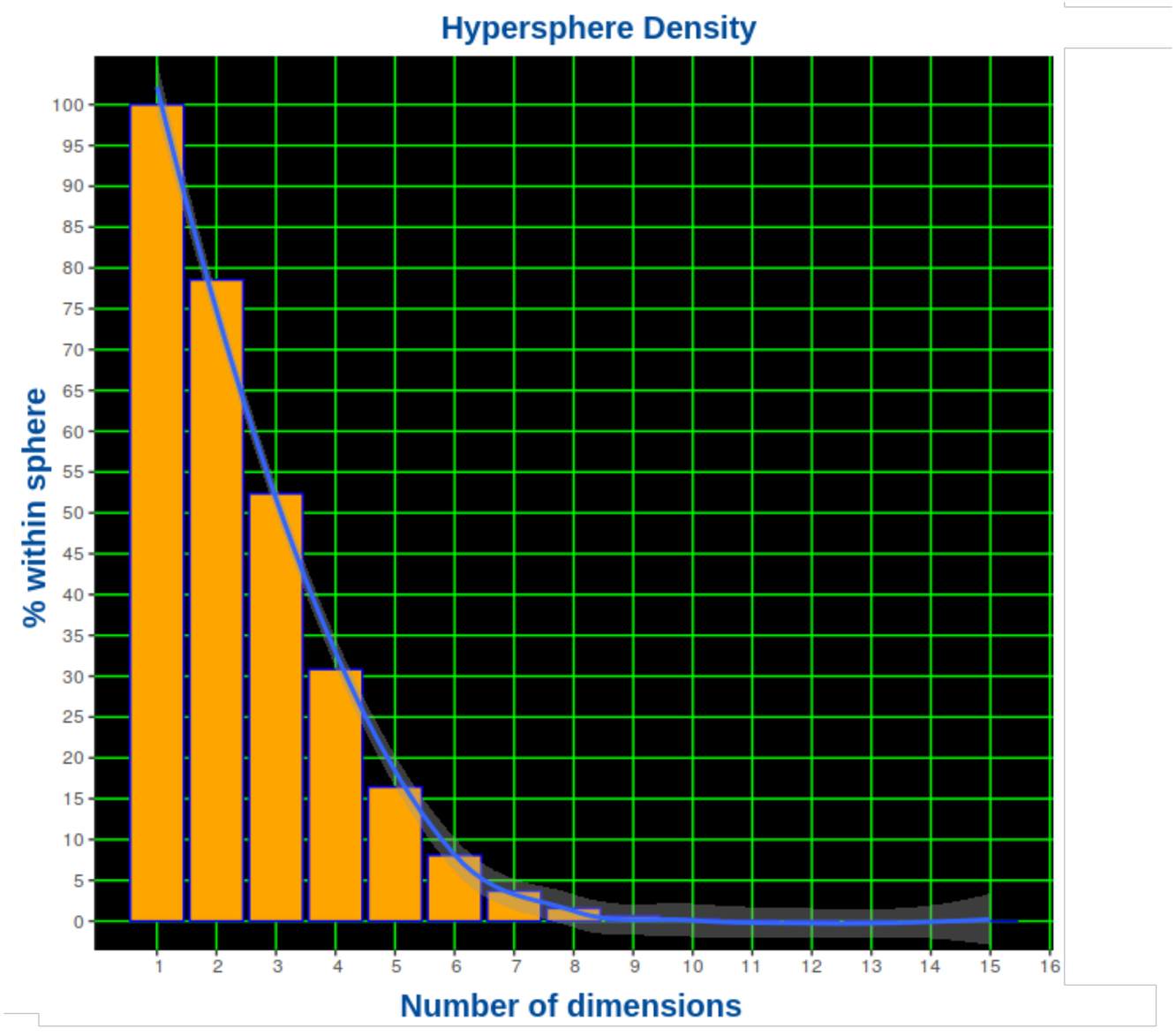
FEATURE SELECTION

Is it possible to anticipate important features for unseen data?

- Will zero-importance variables ever become important?
- Does it matter? (model has zero weights for them)

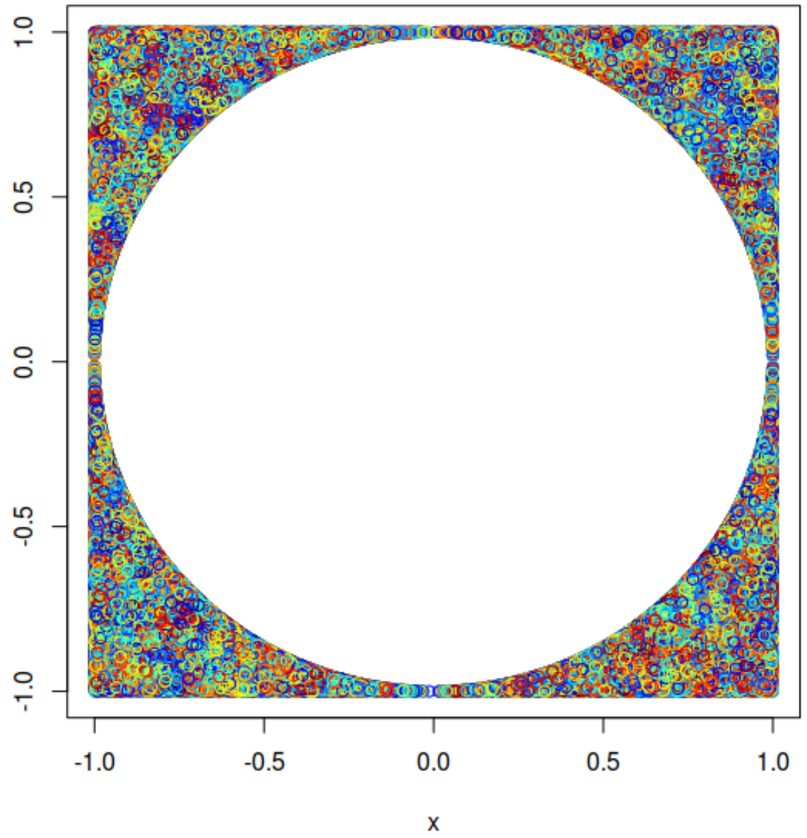
CURSE OF DIMENSIONALITY

Even 15 dimensions is problematic


Hypersphere Density
% within sphere
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Number of dimensions

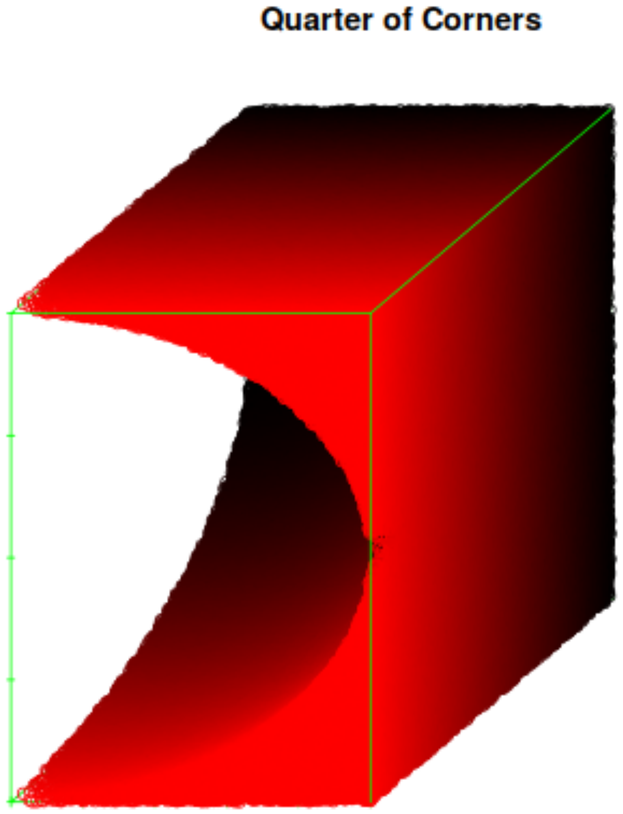
RANDOM 2-D VECTORS

78% of random vectors "on target" (reasonable distance from closest peer)



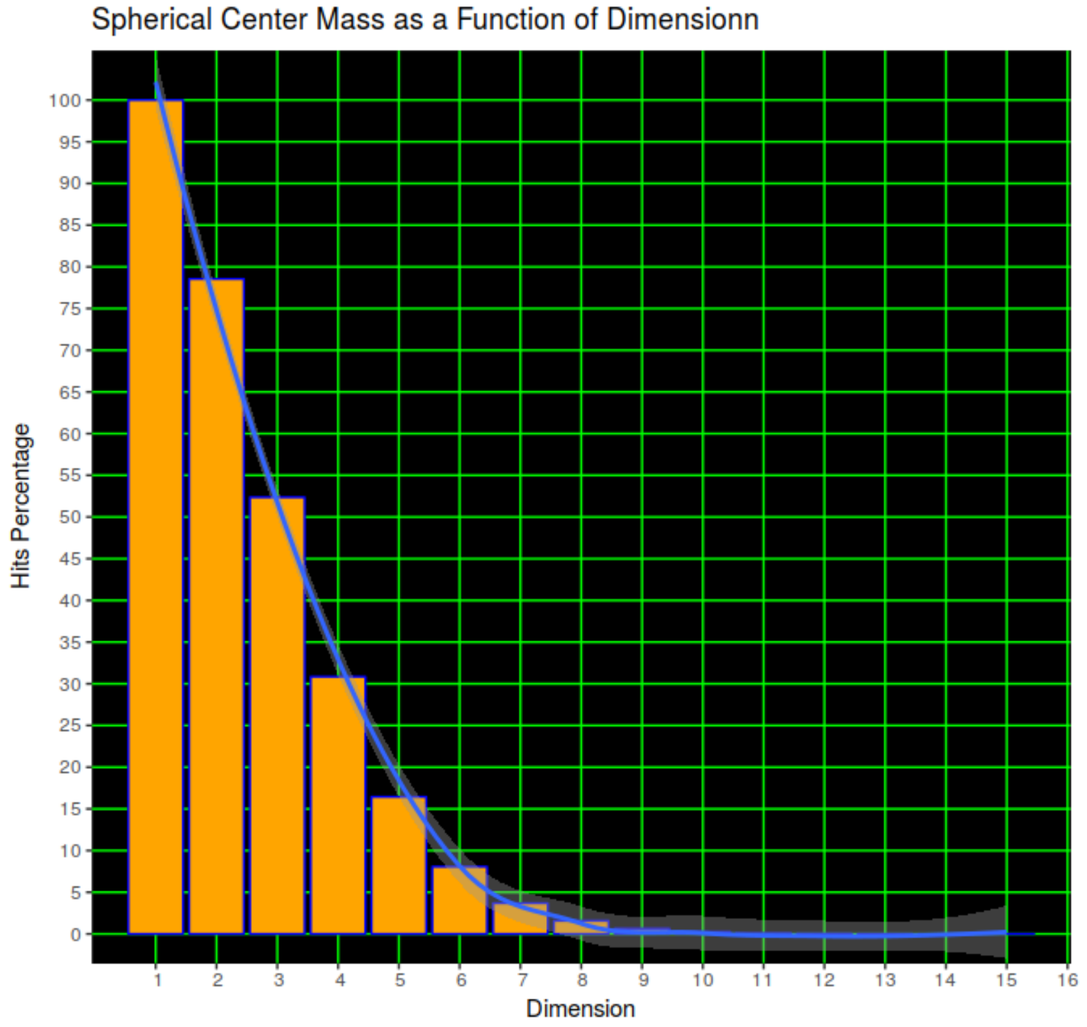
RANDOM 3-D VECTORS

48 % of 3D random vectors "on target"



RANDOM VECTORS IN HIGH-D SPACE

Vanishingly small number of vectors "on target" in 12+ dimensional space. 35% of data "on target" for our 270-D features (mostly of these from public data?).



3. POSSIBLE IMPROVEMENTS

- Metrics
- Cross validation
- Accuracy
- Confidence
- Infrastructure

METRICS

What you want to measure:

- Measure chemist satisfaction
- Measure business performance (hit/miss percents)
- Prediction confidence profile
- Show progress (no glass ceiling) in announcements & internal reports
- Improve promotion reliability

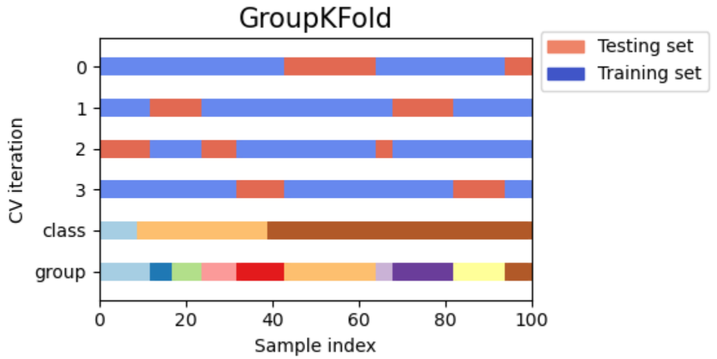
METRIC - SAMPLE WEIGHTS

Sample weights to improve both training and validation

- Time series recency
- "Active validation" (important compounds flagged by chemists)
- Batch assay target std deviation
- Improve confidence estimates

CROSS VALIDATION

Grouped K-Fold CV



ACCURACY VOLUME

- Measure learned manifold volume (accuracy time volume)
- Generalization to training set neighbors
- Better predictions on future compounds
- Better predictions on *subsets* of future compounds

CONFIDENCE

- Lower confidence on false positives
- Lower confidence on false negatives
- Increase confidence on false
- Add uniqueness/isolation feature (distance to closest training example)

INFRASTRUCTURE

Improved DX (developer experience) accelerates progress.

- Deploy joblibs directly+immediately
- Latency
- Reliability

Could chemists deploy new models with a web form?

MODEL IMPROVEMENT

- Feature selection
- Active learning
- Feature extraction

FEATURE SELECTION

- Aggregate importances across models and folds
- RL or baysean feature selection model (train the trainer)
- Improve model evaluation metrics
- Chemist-suggested feature combinations
- Chemist-suggested new features
- Gradient decent (HyperOpt) on each added interaction/polynomial features
- Include Bhanushee fingerprints among feature subsets

FEATURE SELECTION -- ACTIVE LEARNING

Feedback from chemists to improve model and metrics.

- Target molecule space regions of interest to chemists
- RLHF of feature selection model (train the trainer)
- Model to predict ambiguity (std dev in batch assays)

FEATURE EXTRACTION/EMBEDDING:

- Transfer learning: sharing feature embeddings across target classes
- Neural networks: CNN, RNN, transformer, GNN
- Smiles feature embedding (variational autoencoder CNN/RNN)
- Steered embeddings and distance metrics (polynomial model all targets)
- Examine XGBoost decision trees to design interaction and poly features

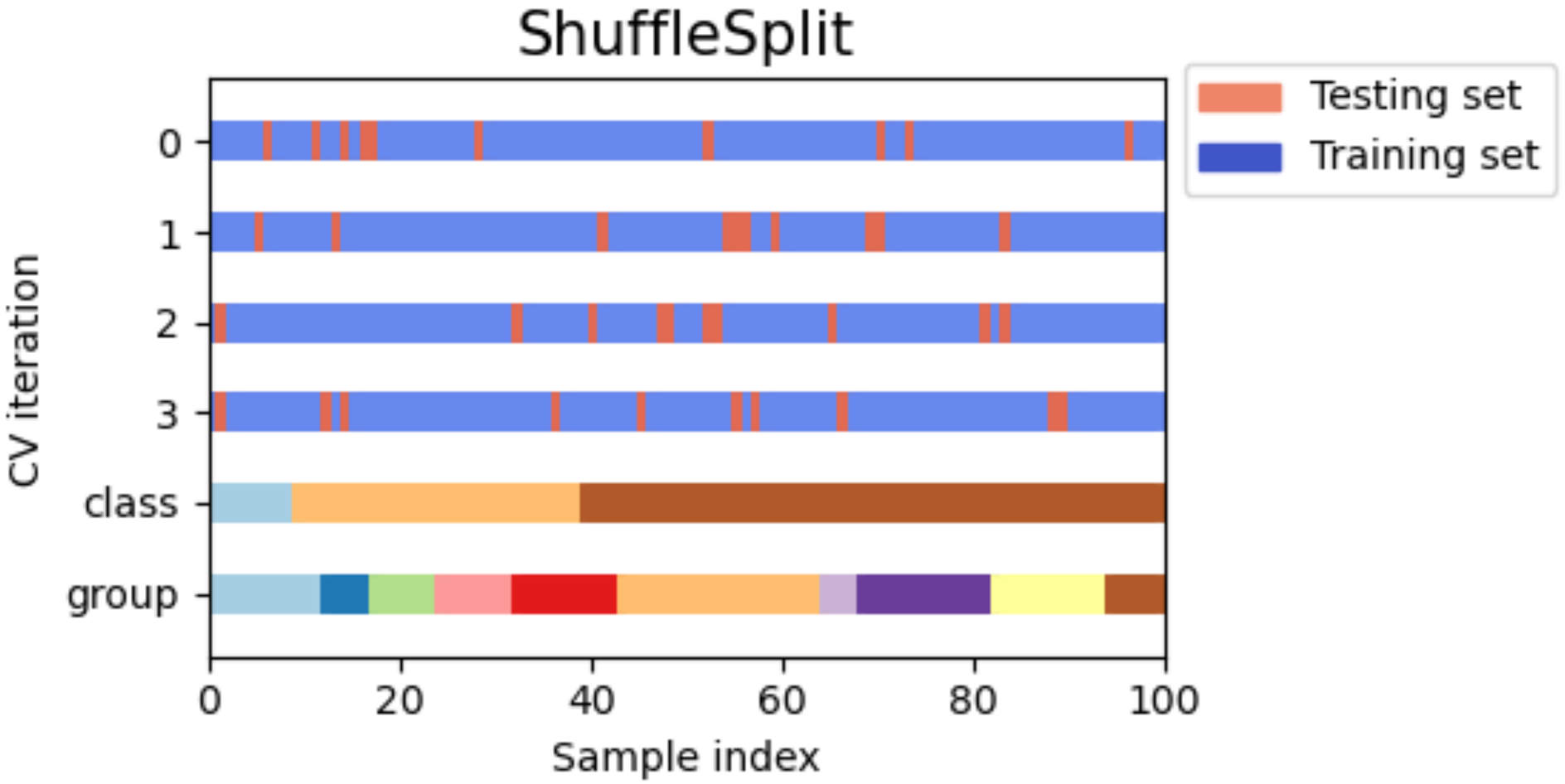
METRICS FOR FEATURE SELECTION

Reduce noise in promotion decision metrics and Krishna curve with CV

- 10 fold CV
- Leave-one-singleton-out
- CV on clusters by project
- CV on clusters by distance (Tanimoto, Levenstein, cosine, Euclidean, Manhattan)
- CV on time series

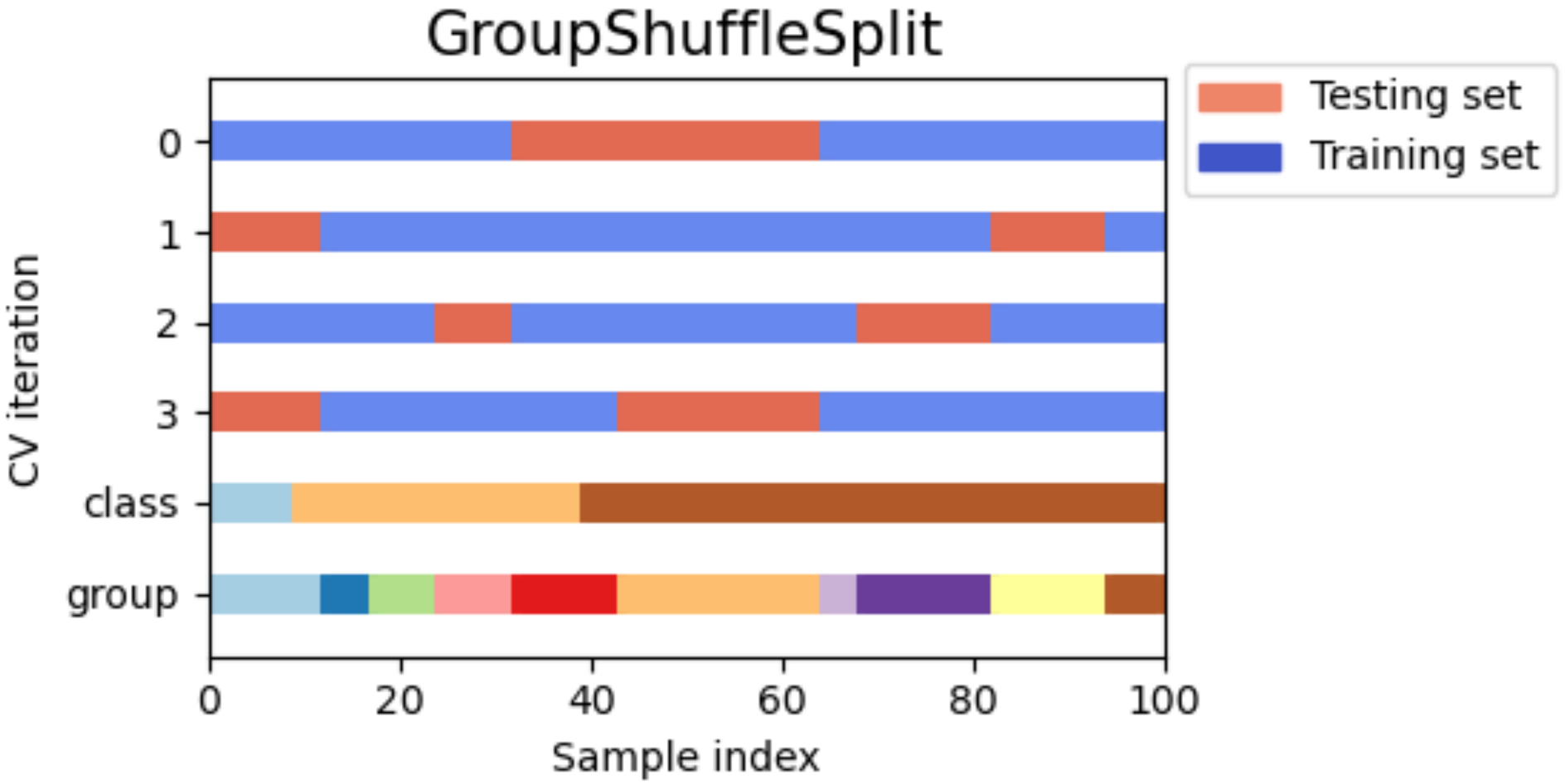
BASELINE

Currently doing 1 20% split. Minimal baseline -- 5-10 random split CV.

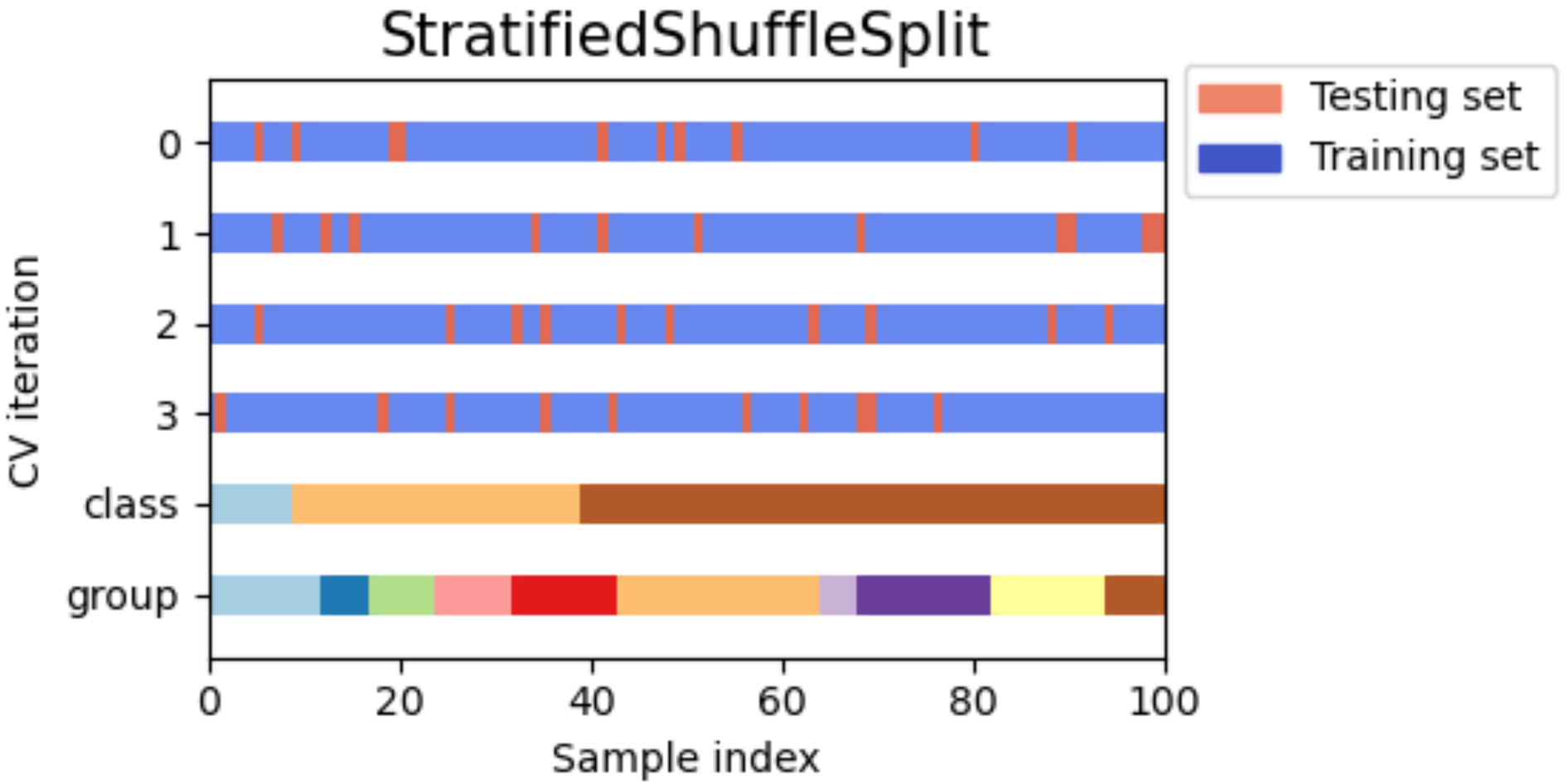


OPTION 1. GROUPED SHUFFLE SPLIT

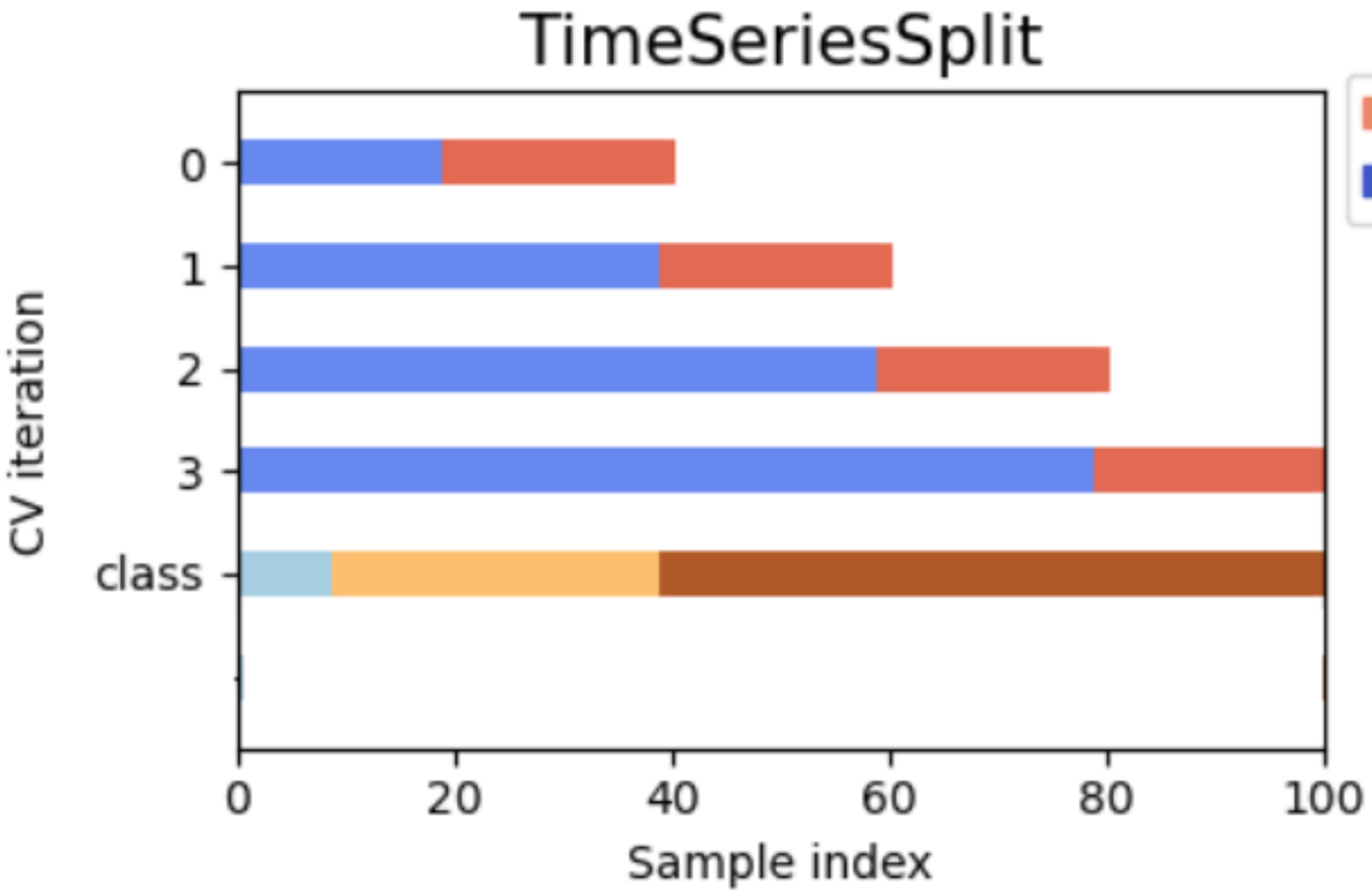
Group according to desired and undesired class.



OPTION 2. STRATIFIED SHUFFLE SPLIT



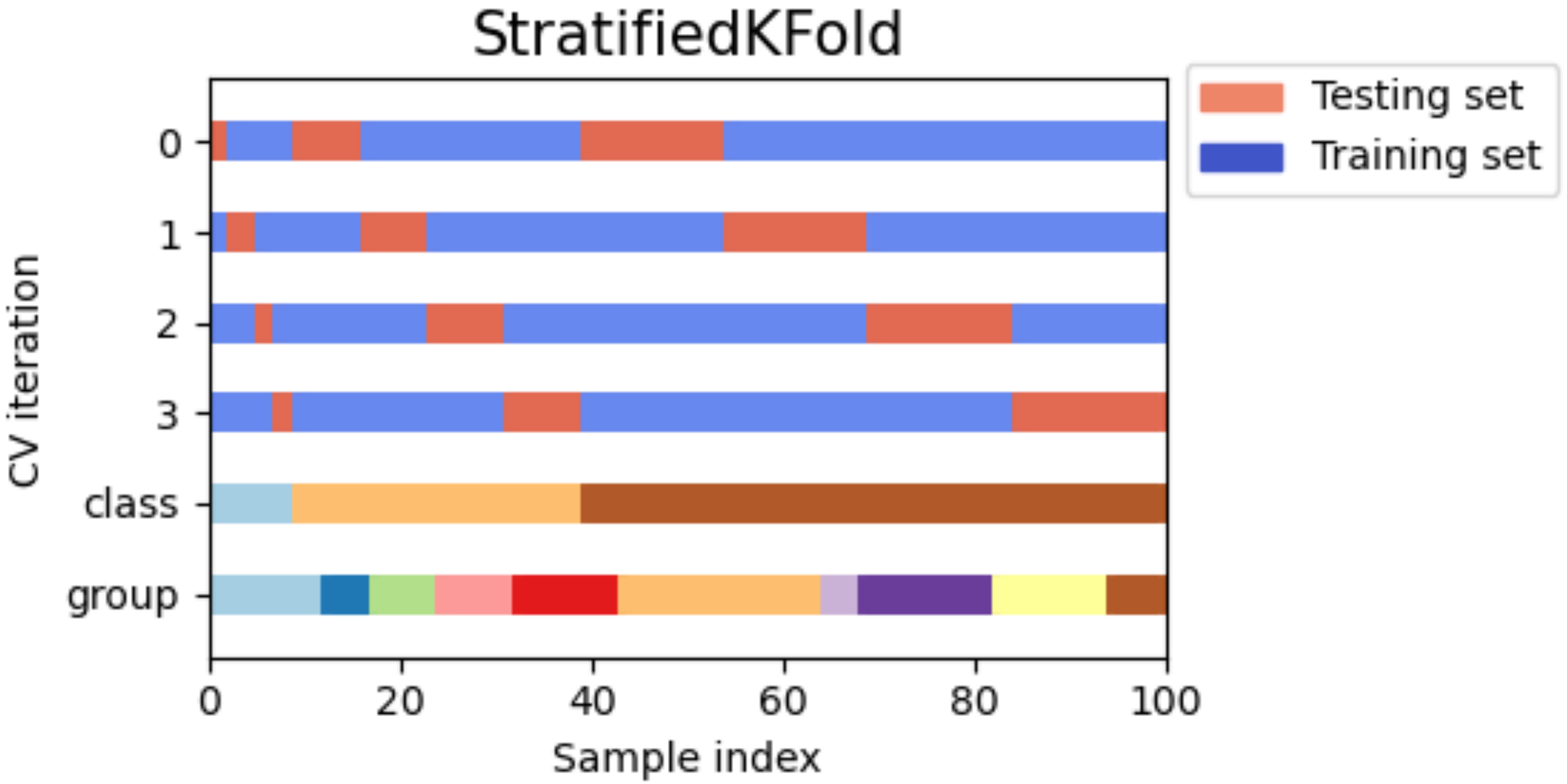
OPTION 3. TIME SERIES SPLIT



4 splits
3 classes
0 groups

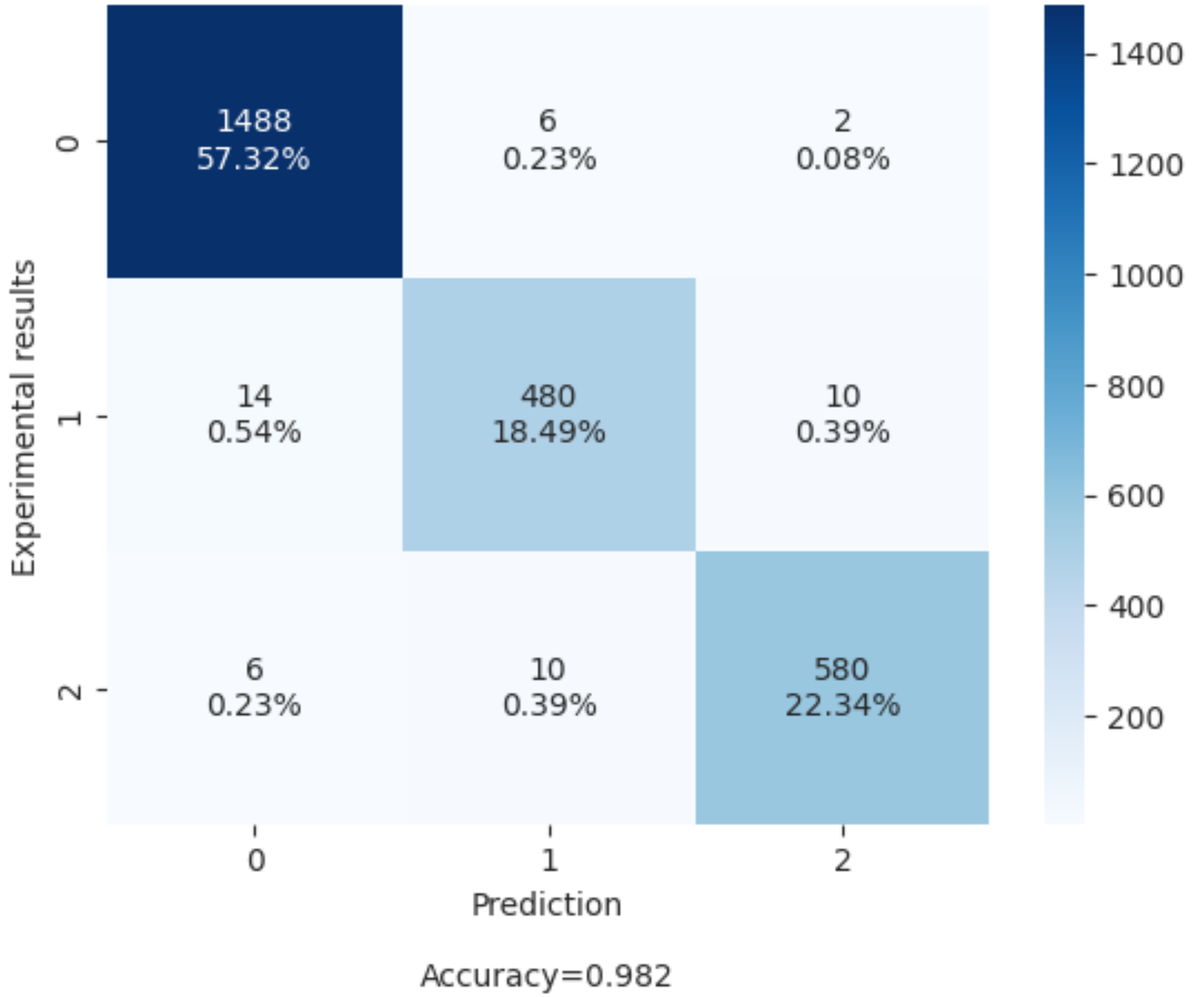
STRATIFIED K-FOLD

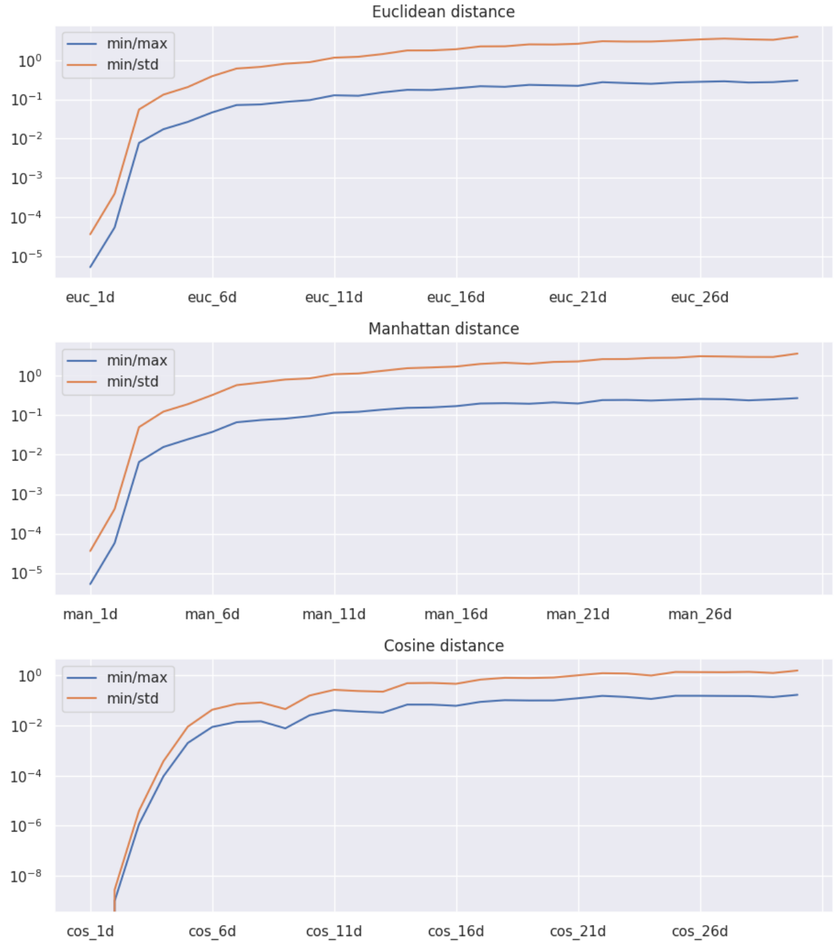
Folding won't detect anomalous (unlucky) splits.

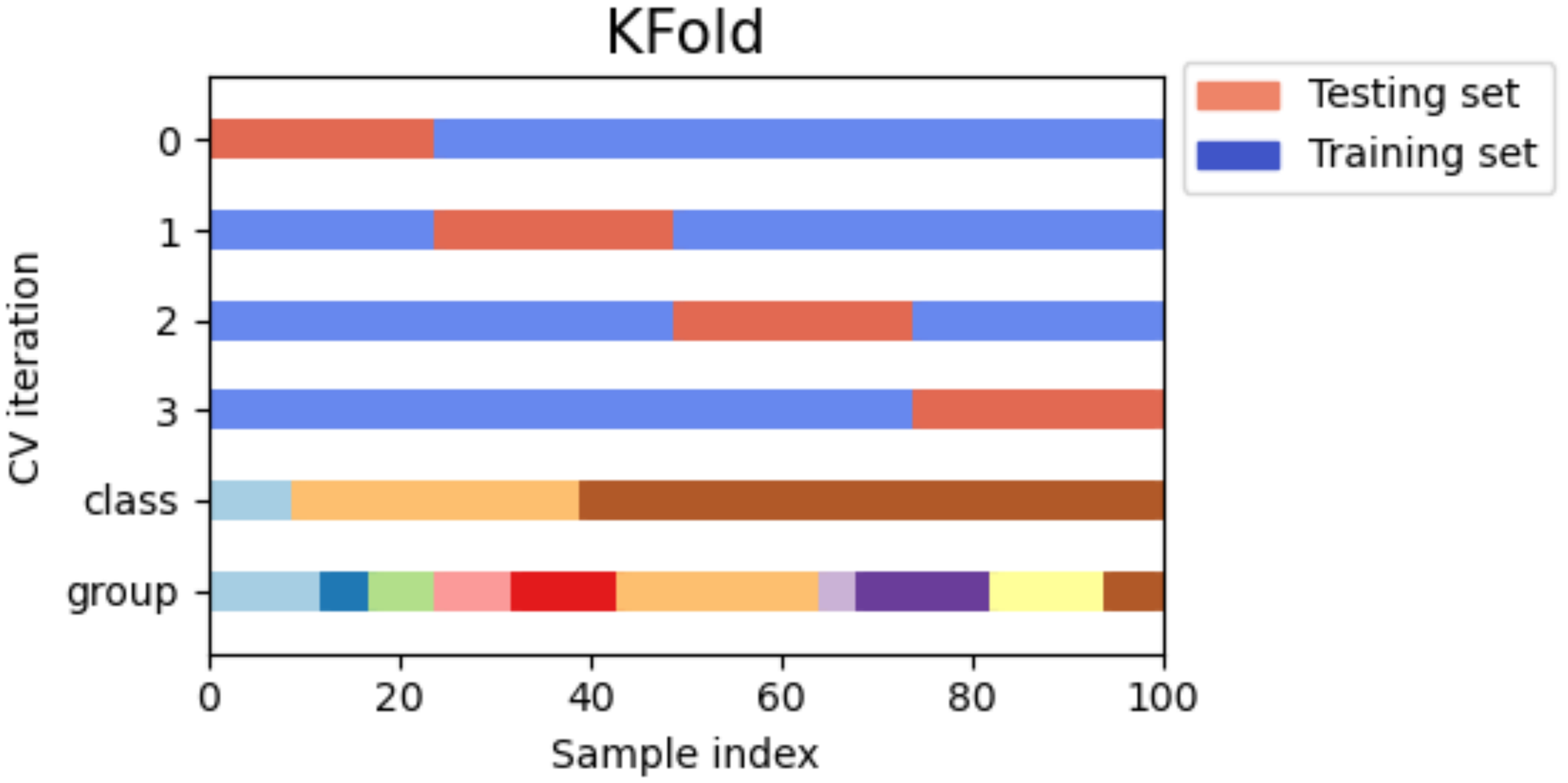


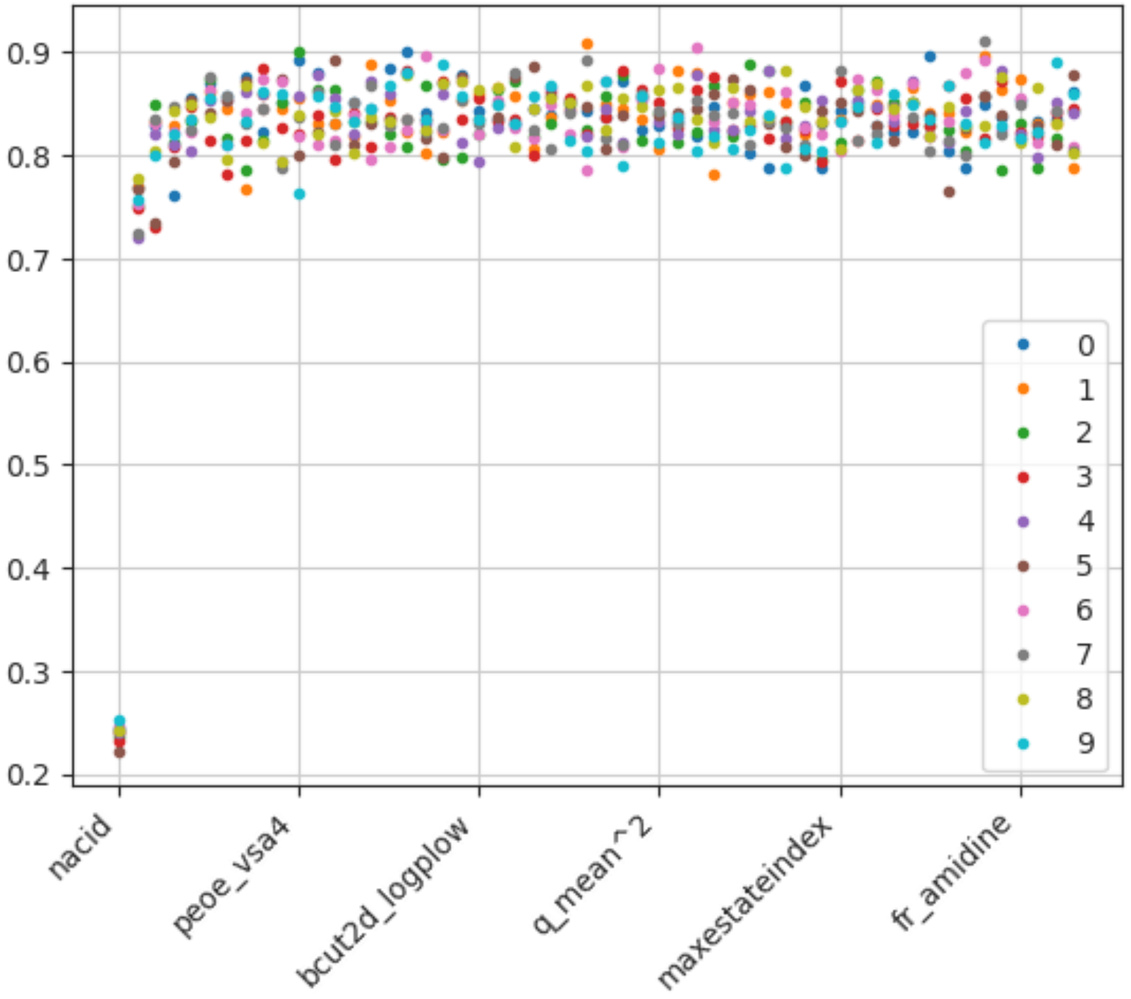
PKA A1 WALDEN

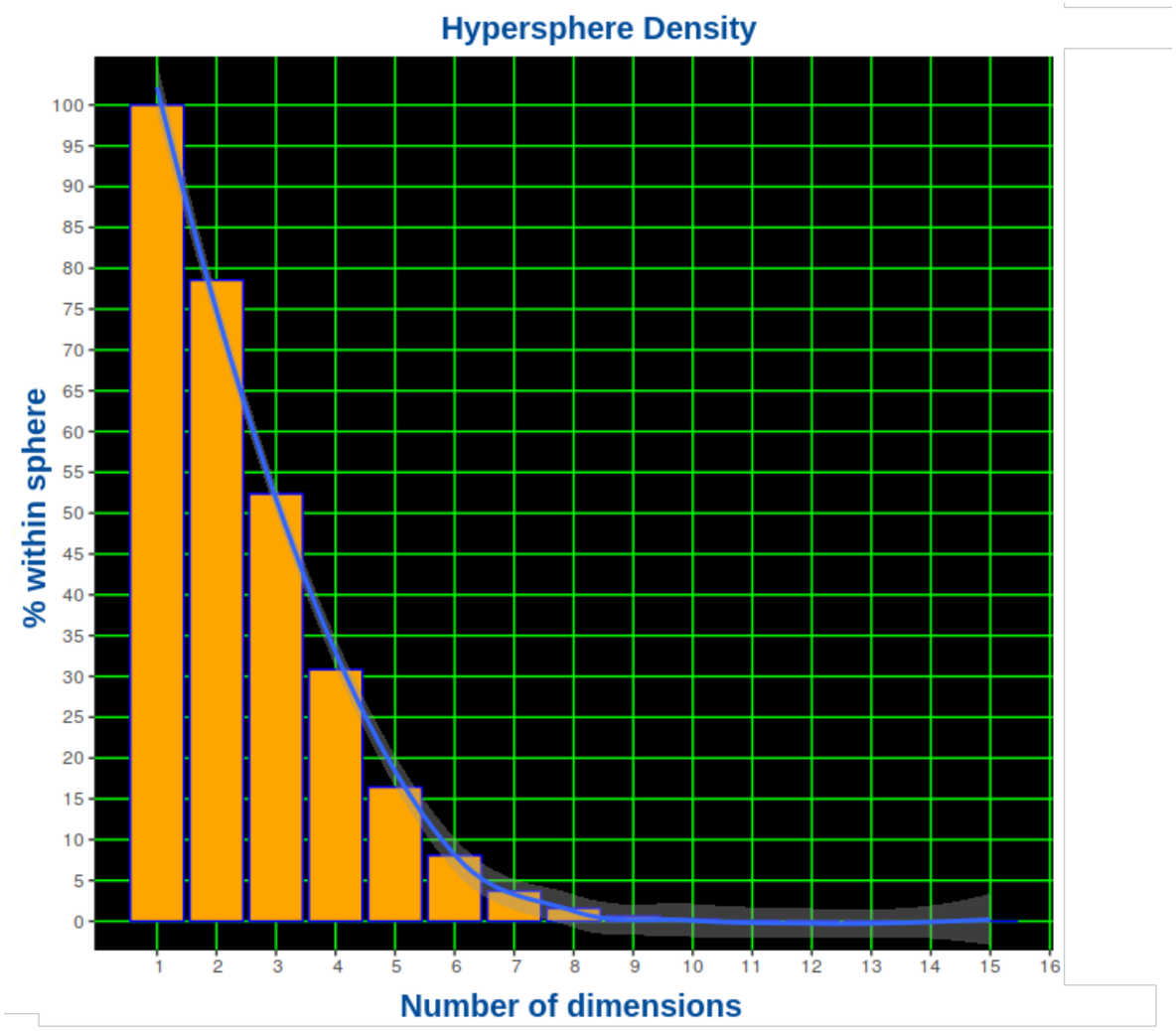
Confusion Matrix

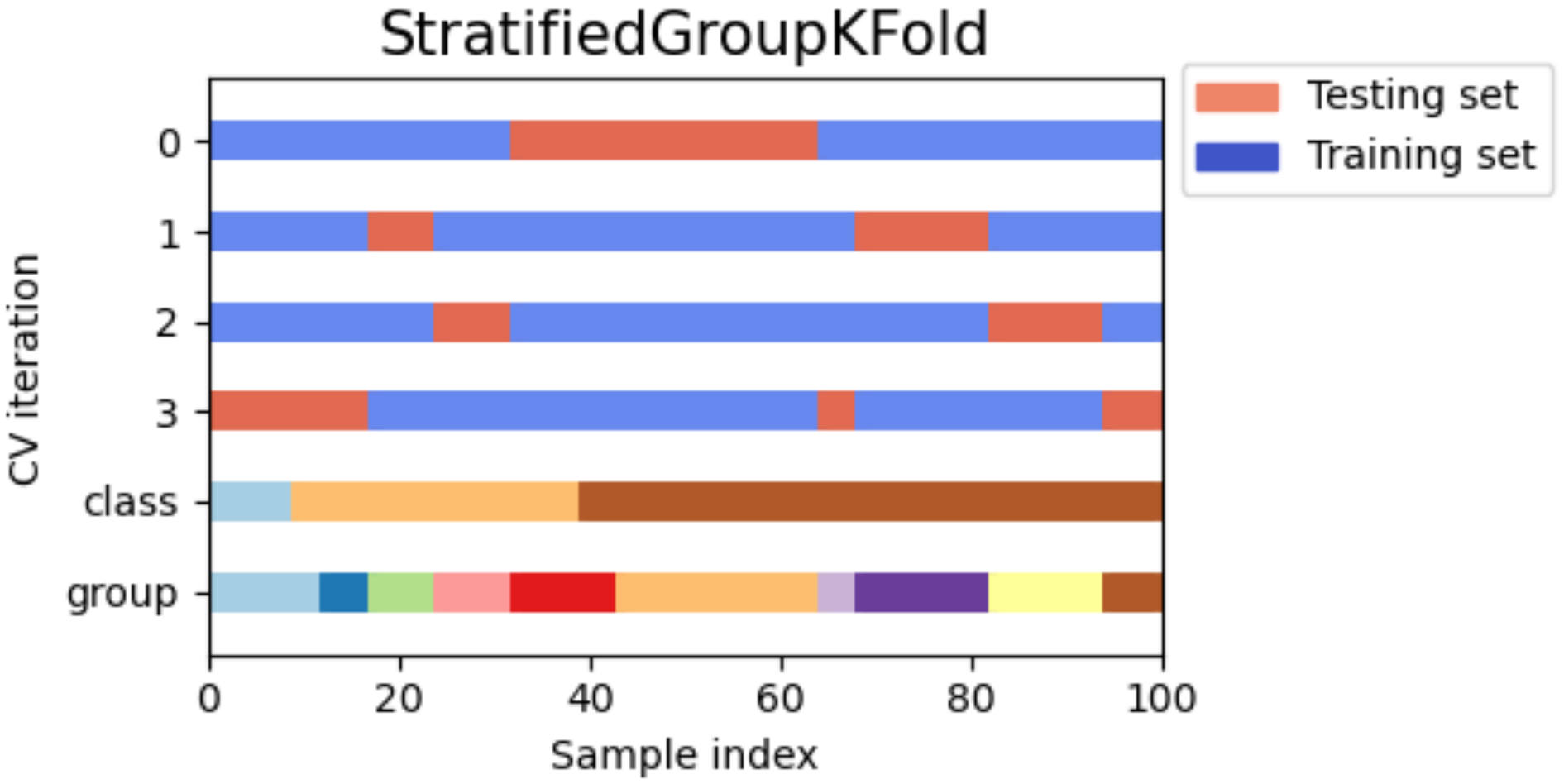



Euclidean distance
min/max
min/std
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
euc_1d
euc_6d
euc_11d
euc_16d
euc_21d
euc_26d
Manhattan distance
min/max
min/std
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
man_1d
man_6d
man_11d
man_16d
man_21d
man_26d
Cosine distance
min/max
min/std
10^{0}
10^{-2}
10^{-4}
10^{-6}
10^{-8}
cos_1d
cos_6d
cos_11d
cos_16d
cos_21d
cos_26d


KFold
Testing set
Training set
CV iteration
0
1
2
3
class
group
0
20
40
60
80
100
Sample index


0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
nacid
peoe_vsa4
bcut2d_logplow
q_mean^2
maxestateindex
fr_amidine
0
1
2
3
4
5
6
7
8
9


Hypersphere Density
% within sphere
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Number of dimensions


StratifiedGroupKFold
Testing set
Training set
CV iteration
0
1
2
3
class
group
0
20
40
60
80
100
Sample index

- [^1]: grid vs random hyperparmeter search - https://i0.wp.com/neptune.ai/wp-content/uploads/2022/10/grid_random.png?ssl=1
- [^2]: IDY-1609 - https://utiliware.atlassian.net/browse/IDY-1609
- [^3]: Scott's report on overfitting and feature selection - https://utiliware.atlassian.net/wiki/spaces/IDB/pages/1291976705/Overfitting+Clustering+and+Test+Set+Selection