

AI and ML for Chemists

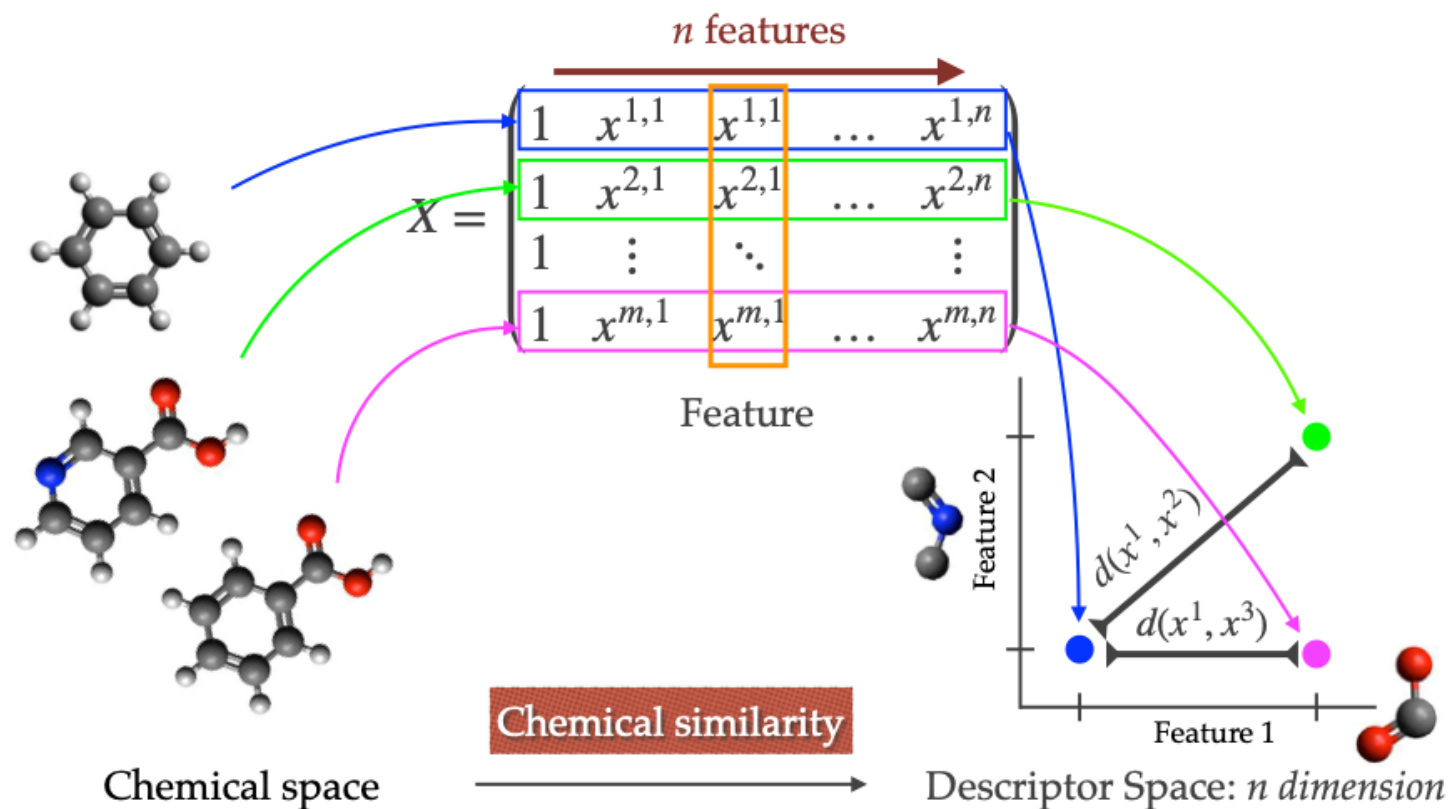
Topic 4.2: Feature engineering

Pre-lecture Explore

Dr. Ganna (Anya) Gryn'ova

Feature importance

refers to techniques that calculate a score for all the input features for a given model. The scores represent the “importance” of each feature. A higher score means that the specific feature will have a larger effect on the model that is being used to predict a certain variable.

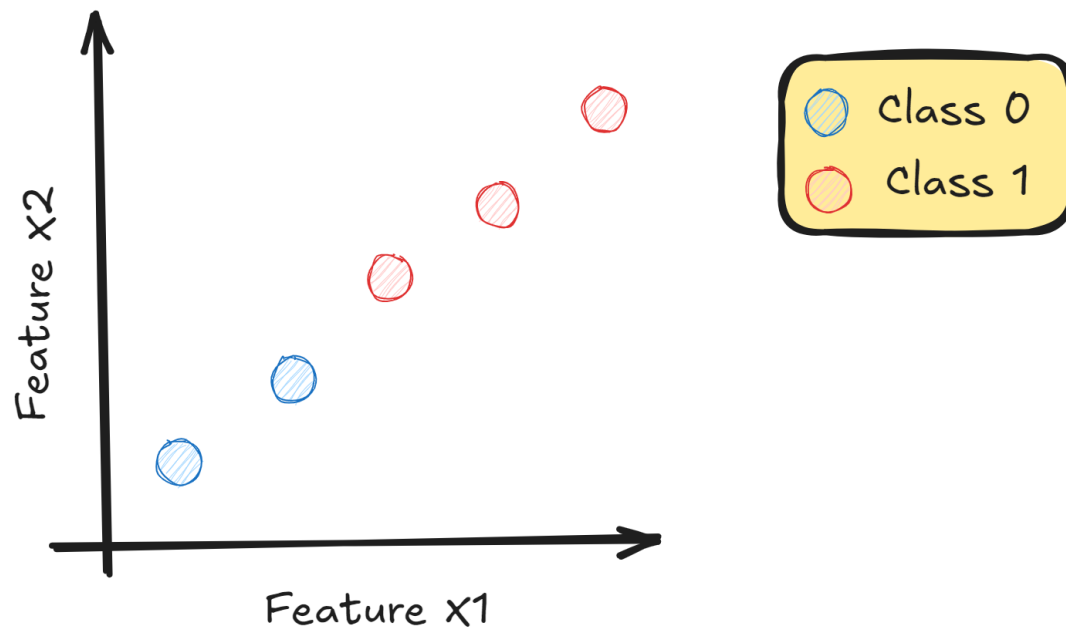


Feature importance

- **Feature selection:** By identifying the most important features, we can select a subset of relevant features for use in building a model, reducing dimensionality and noise in the data, and improving model interpretability.
- **Model interpretability / explainability:** By understanding which features are most important, we can gain insights into the underlying relationships in the data and how the model is making predictions.
- **Model debugging:** If a model is not performing well, feature importance can be used to identify which features may be causing problems and require further investigation.
- **Improving model performance:** By removing less important features, we can improve model performance by reducing overfitting and training time.

Gini importance

also known as Mean Decrease in Impurity (MDI), is used to evaluate how much a feature reduces the impurity of a node in a decision tree. The Gini Impurity of a dataset is a number between 0-0.5, which indicates the likelihood of new, random data being misclassified if it were given a random class label according to the class distribution in the dataset.



Gini Impurity

$$G = 1 - \sum_{k=1}^K p_k^2$$

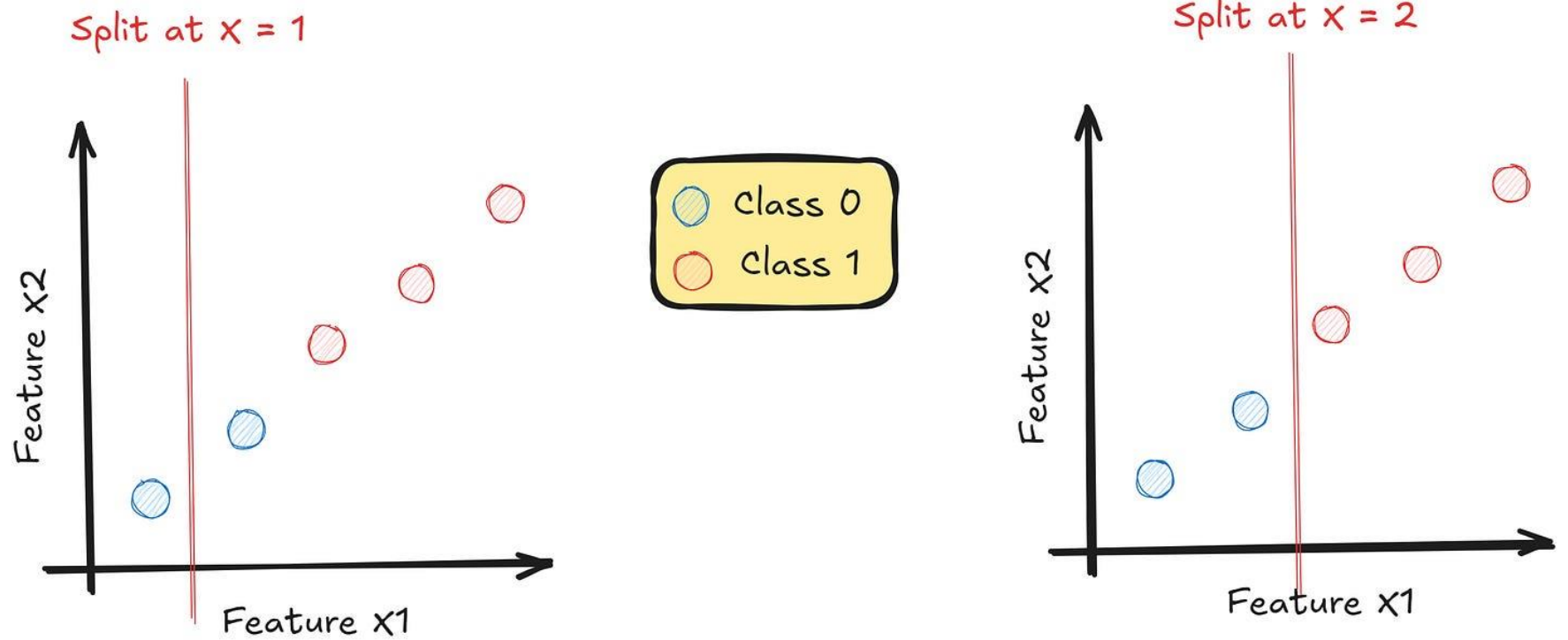
K → The number of classes
↓
Proportion of samples of class k in the node

$$G = 1 - ((2/5)^2 + (3/5)^2)$$

$G = 0.52$

Gini importance

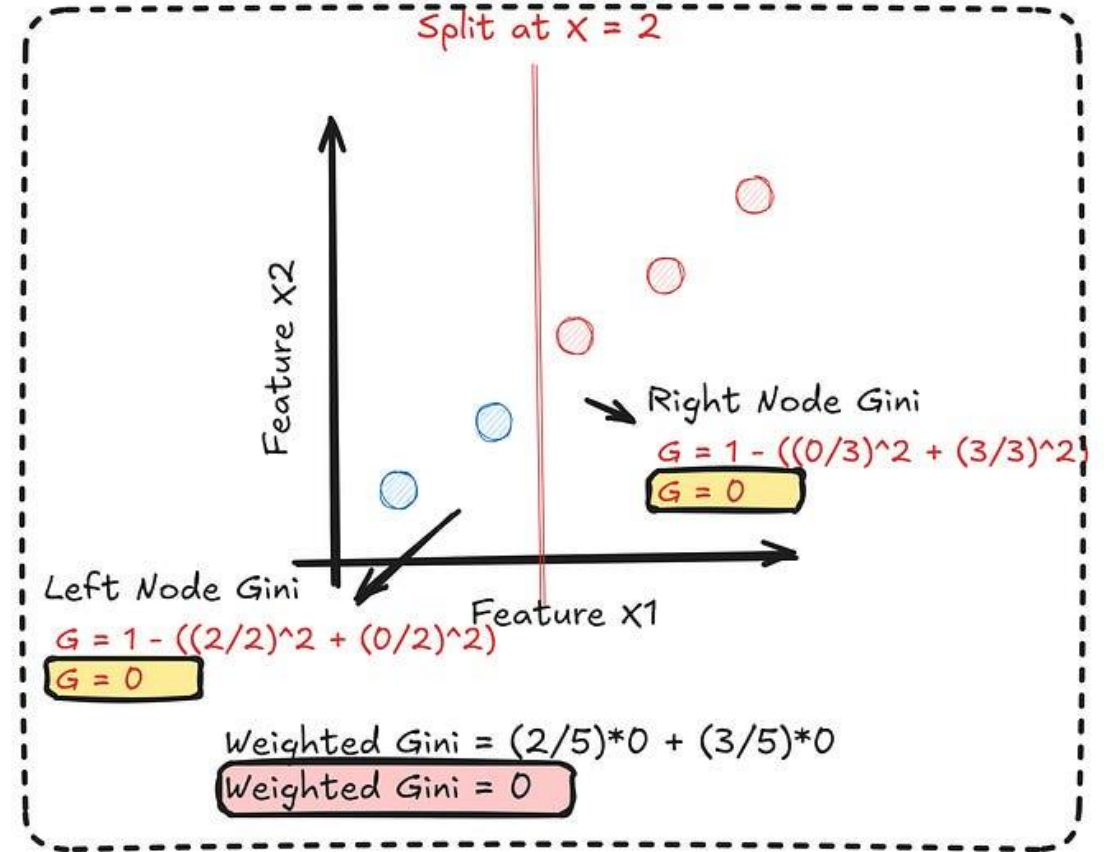
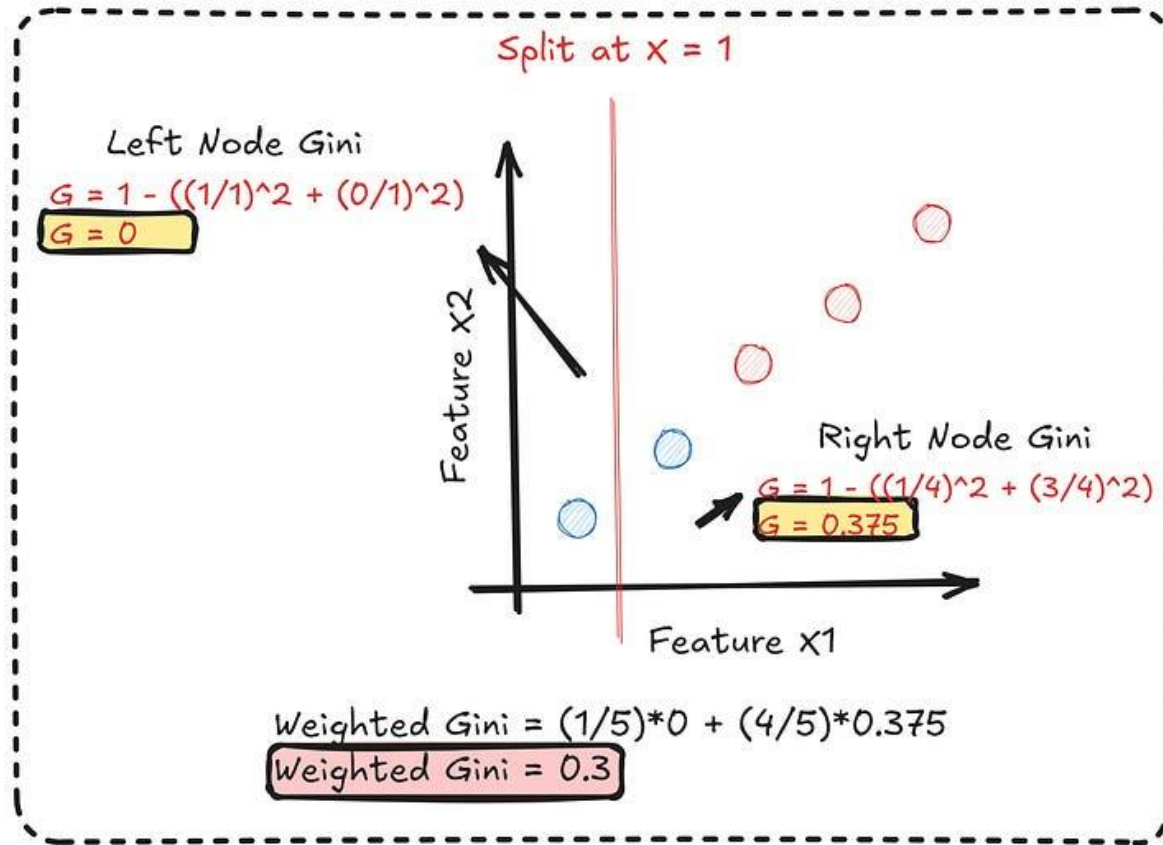
Which split is better?



Weighted Gini Impurity for the split:

$$\frac{n_{\text{left}}}{n_{\text{total}}} G_{\text{left}} + \frac{n_{\text{right}}}{n_{\text{total}}} G_{\text{right}}$$

Gini importance



The best split is the one that has the highest purity, which means we want the lowest Weighted Gini. The example above shows that the split at $X = 2$ is the best.

Gini importance

Reduction in Gini impurity for each feature

$$\Delta G(f_1) = G_{\text{parent}} - \text{Weighted Gini}$$

In this example, G_{parent} is 0.52 and the split-weighted Gini for the best split ($X = 2$) is 0.

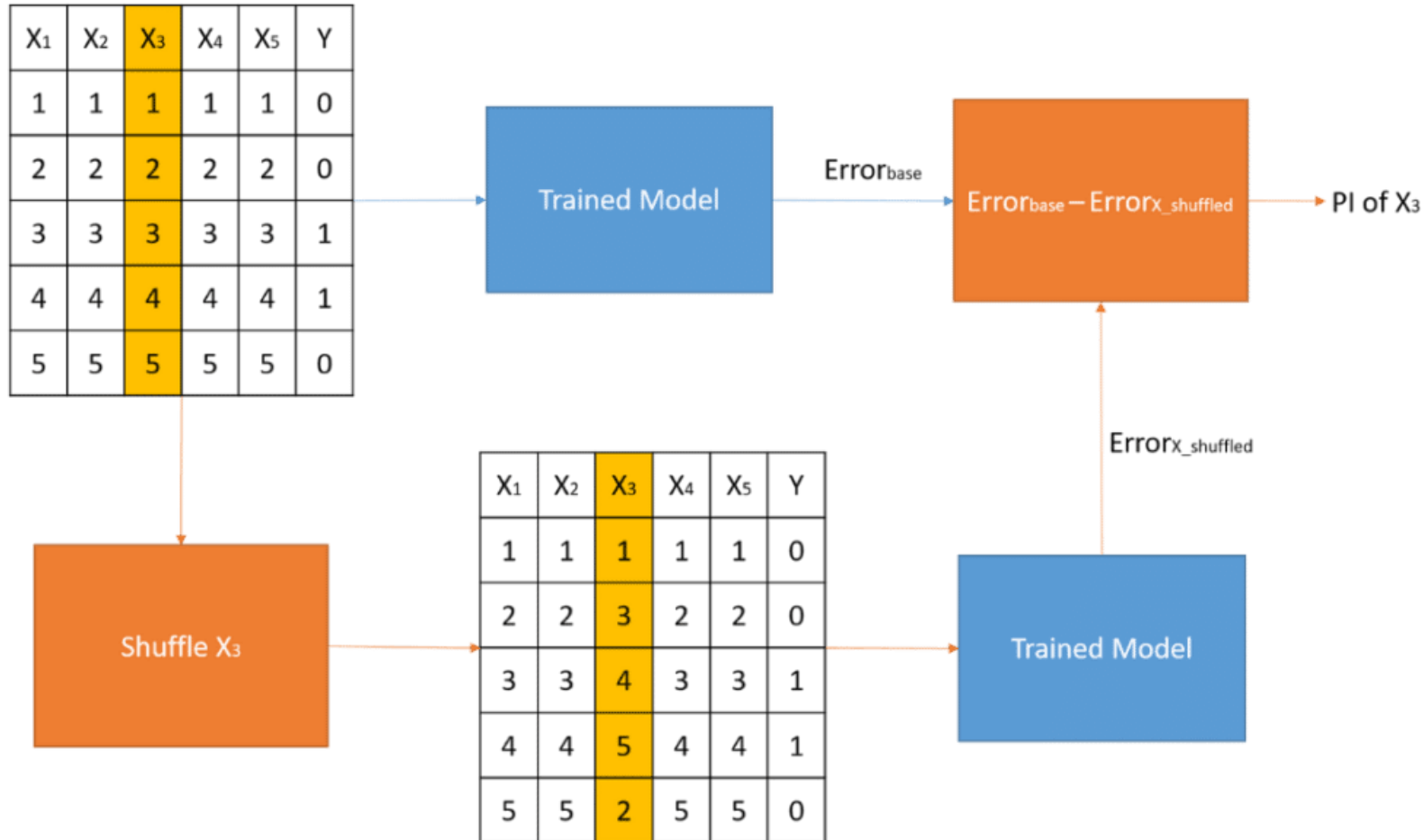
$$\Delta G(f) = 0.52 - 0$$

$$\Delta G(f) = 0.52$$

The sample above shows that the reduction in Gini Impurity is 0.52. That is the **feature importance score**.

Permutation feature importance

measures the increase in the prediction error of the model after we permuted the feature's values, which breaks the relationship between the feature and the true outcome.



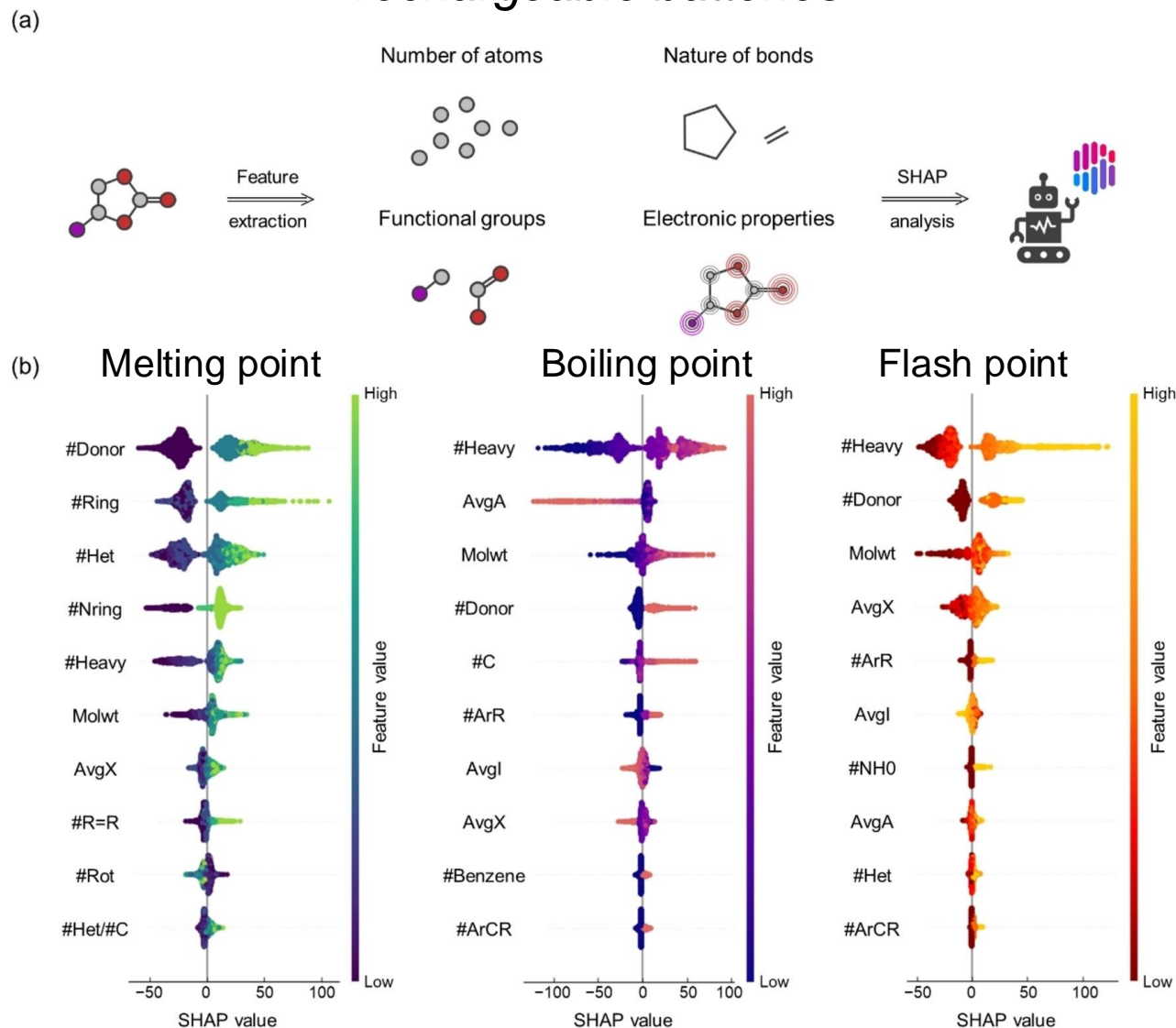
The greater the difference, the more important that feature is.

SHapley Additive exPlanations

(SHAP) values are a way to explain the output of any machine learning model. It uses a game theoretic approach that measures each player's contribution to the final outcome. In machine learning, each feature is assigned an importance value representing its contribution to the model's output.

SHAP values show how each feature affects each final prediction, the significance of each feature compared to others, and the model's reliance on the interaction between features.

Electrolyte molecules for rechargeable batteries



Curse of dimensionality

refers to various phenomena that arise when analysing and organising data in high-dimensional spaces that do not occur in low-dimensional settings such as the three-dimensional physical space of everyday experience.

The curse generally refers to issues that arise when the number of datapoints is small (in a suitably defined sense) relative to the intrinsic dimension of the data. In the realm of machine learning, as the number of features or dimensions in a dataset increases, (1) the amount of data we need to generalise accurately grows exponentially and (2) we are increasing the complexity of our data without necessarily increasing the amount of useful information. It can cause:

- **Data sparsity:** most of the high-dimensional space is empty making clustering and classification tasks challenging.
- Increased **computational** resources and time.
- **Overfitting:** fitting to the noise rather than the underlying pattern.
- **Distances lose meaning --> Performance degradation**, especially for algorithms relying on distance measurements like k-nearest neighbors and kernel ridge regression, can see a drop in performance.
- **Visualisation challenges.**

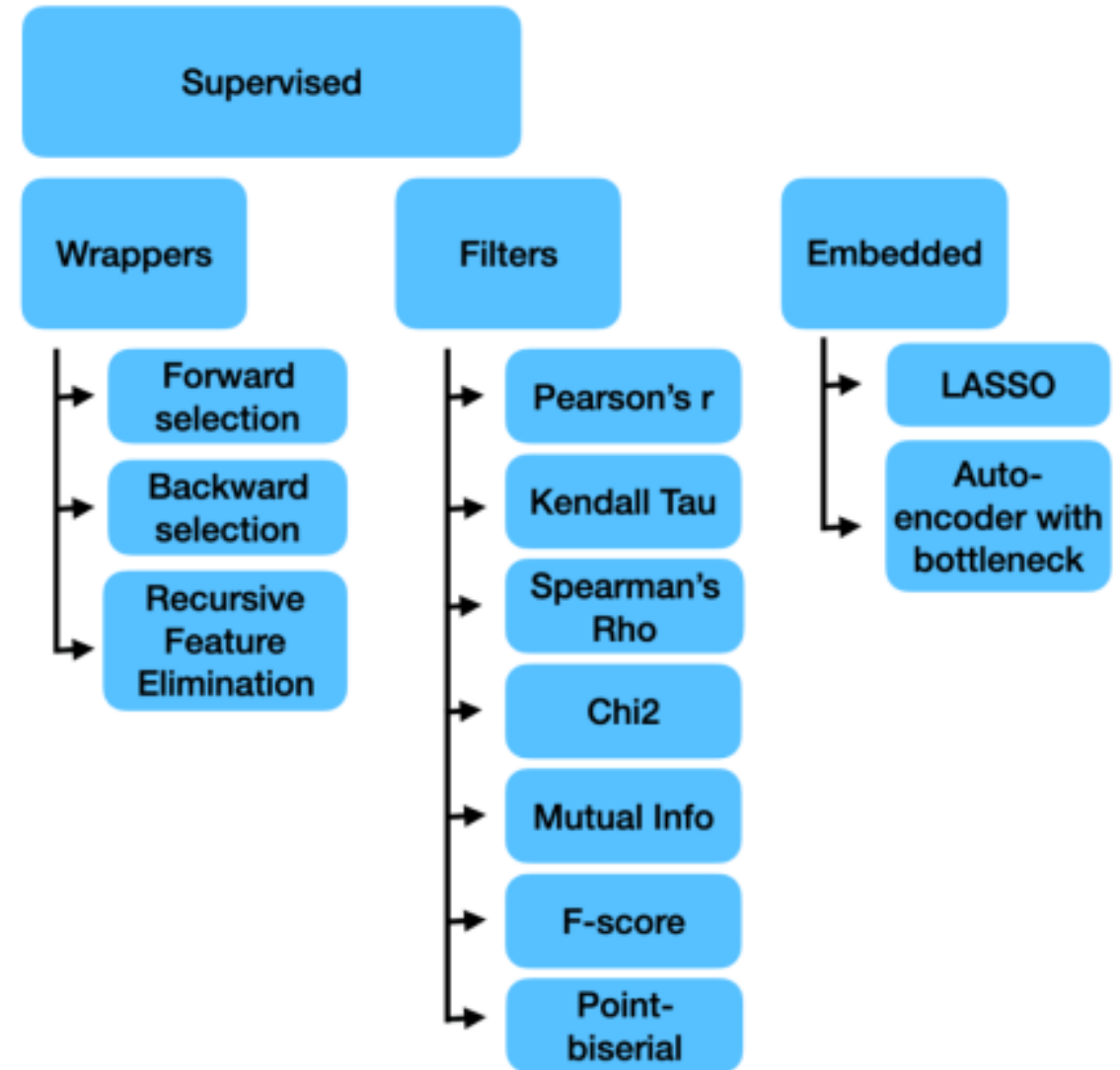
Feature selection

Wrappers: use a model to score different subsets of features to finally select the best one. Each new subset is used to train a model whose performance is then evaluated on a hold-out set. The features subset which yields the best model performance is selected.

Filters: to evaluate the usefulness of each feature, they simply analyse its statistical relation with the model's target, using measures such as correlation or mutual information as a proxy for the model performance metric.

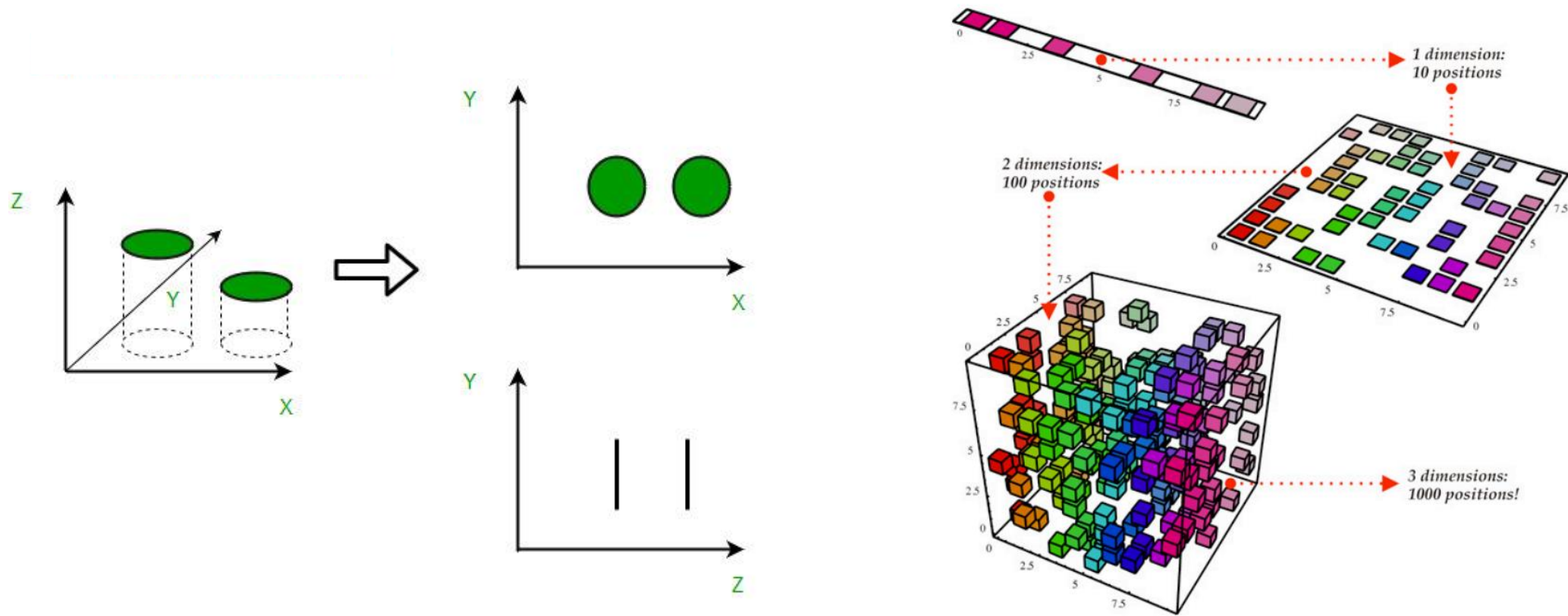
Embedders: embed (incorporate) the selection into the ML model itself, combining speed of the filters with getting the best subset for the particular model (like from a wrapper).

Feature selection methods

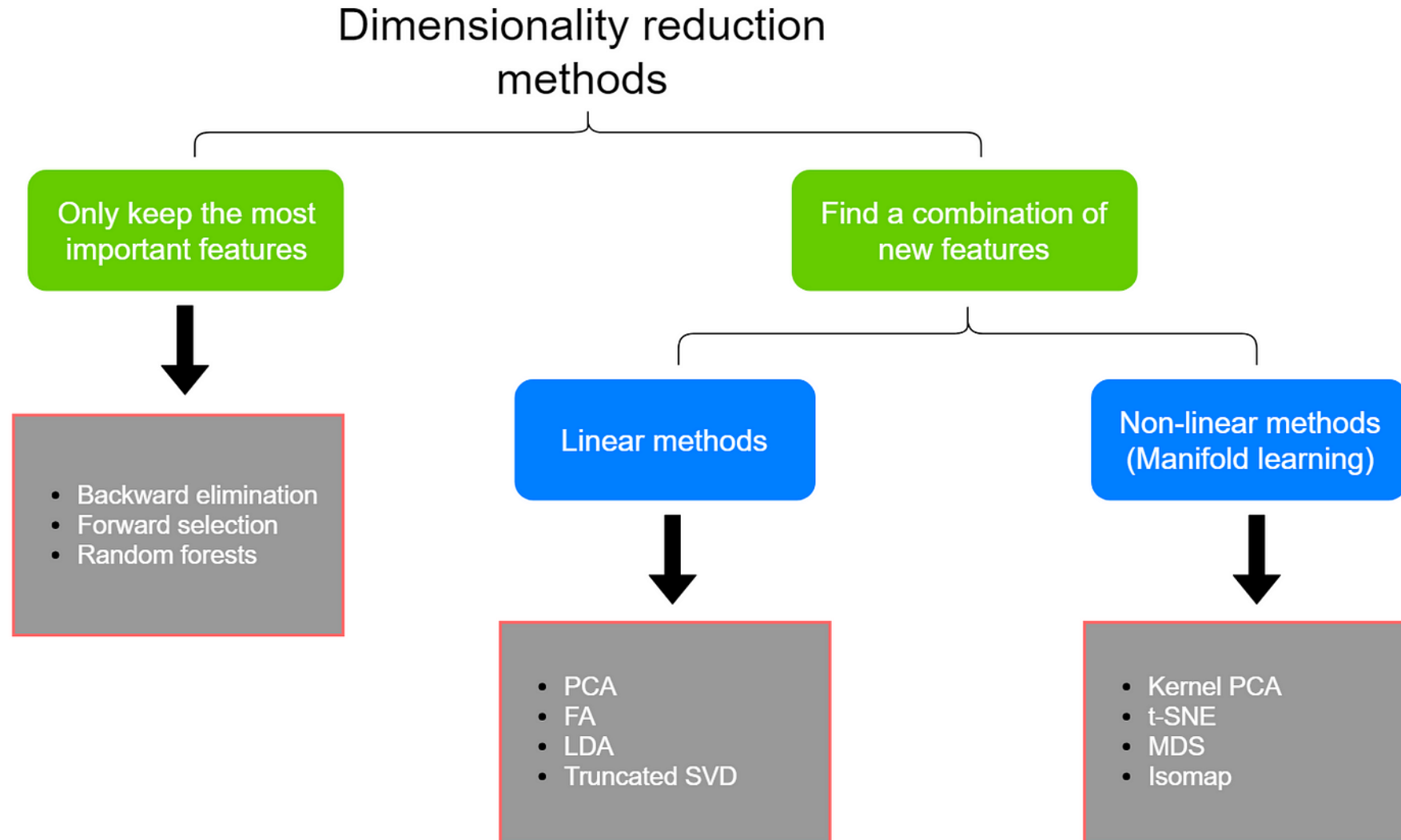


Dimensionality reduction

(feature projection) – a process that reduces the number of random variables under consideration by obtaining a set of principal variables. By reducing the dimensionality, we can retain the most important information in the data while discarding the redundant or less important features.

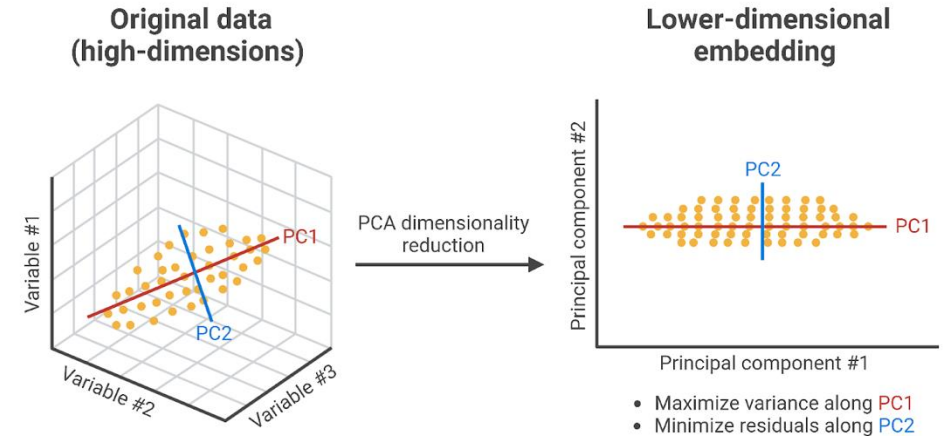


Dimensionality reduction

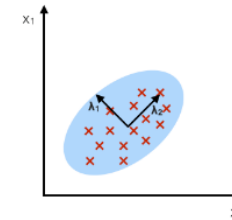


Dimensionality reduction

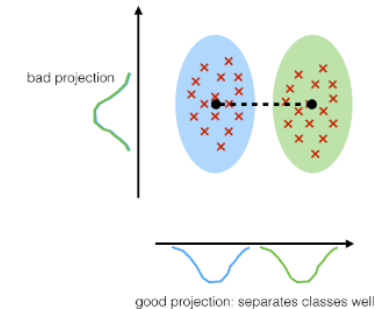
- **Principal Component Analysis (PCA)** is an unsupervised machine learning method that transforms the original variables, correlated with each other, into a new set of variables, which are linear combinations of the original variables. These new variables are called principal components.
- **Linear Discriminant Analysis (LDA)** a supervised machine learning method that is used to separate two groups/classes. LDA focuses on maximising the separability among known categories by creating a new linear axis and projecting the data points on that axis. It's particularly useful for classification tasks.
- **t-Distributed Stochastic Neighbor Embedding (t-SNE)** is an unsupervised machine learning method for non-linear dimensionality reduction that is particularly useful for visualising high-dimensional datasets.



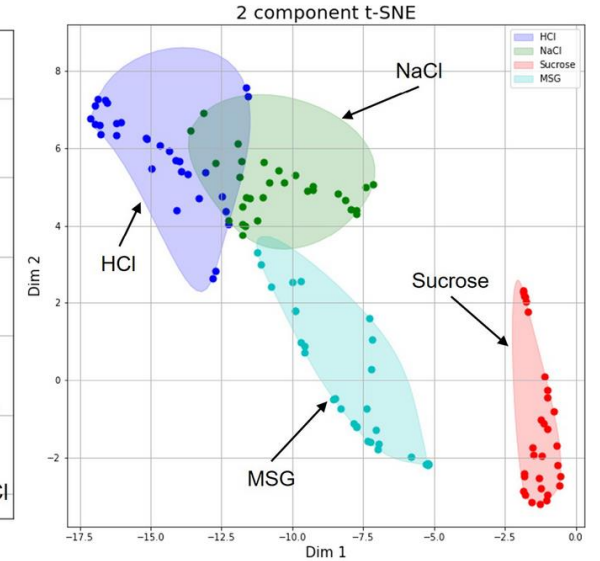
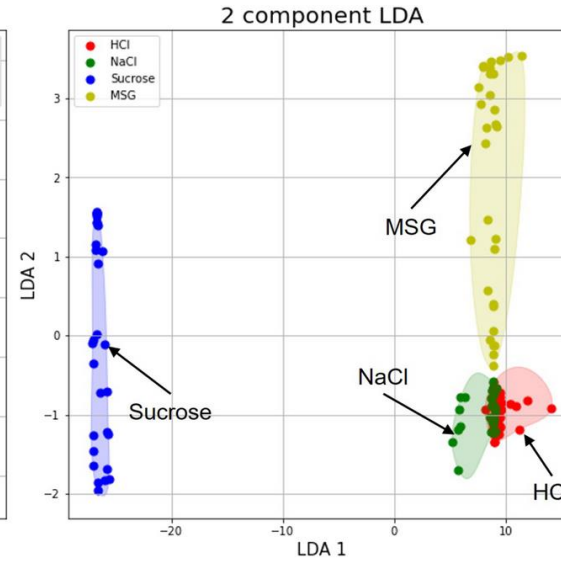
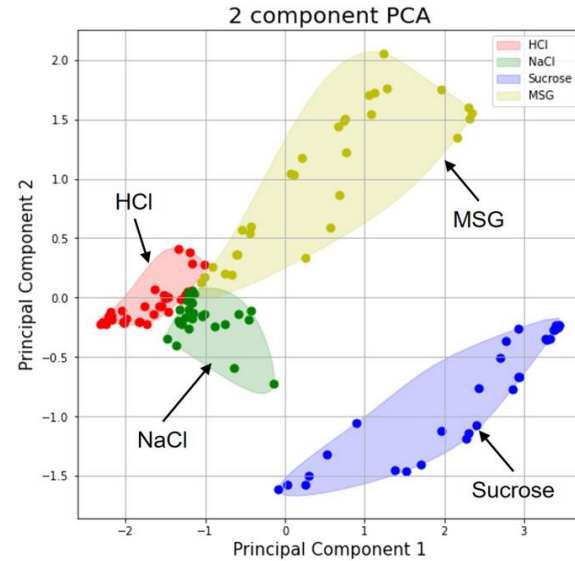
PCA:
component axes that maximize the variance



LDA:
maximizing the component axes for class-separation



Dimensionality reduction



Orange juice adulteration using water dilution and sucrose addition in the range of 0–30% and 0–60%, respectively:

- HCl – acidic
- NaCl – salty
- Sucrose – sweet
- MSG – umami

