

- How may we create a machine learning classifier that can tell apart two groups of proteins?
- How may we improve upon the performance of a machine learning classifier that does not perform so well?
- How may we check for overfitting in a trained machine learning classifier?
- How may we pick the most relevant data points that are related to the class of a sample?
- Can we also train a deep learning predictor to tell proteins apart?
- Understand how protein sequences can become readable to machine learning predictors
- Practise machine learning optimisation techniques including GridSearchCV and dimensionality reduction
- Check for overfitted data by testing with a totally naive dataset
- Practise prediction by applying a deep learning model to the problem and evaluating its performance

Separating Mouse and Human Antibody Sequences using Protein Encoding and Machine Learning Classifiers

Keypoints

- Protein sequences must be numerically encoded to be readable by machine learning algorithms.
- It is sometimes necessary to experiment with different machine learning classifiers or hyperparameterisation techniques, to achieve the best prediction results.
- Check the performance of your model with a “held back” dataset which was not included in the training set.