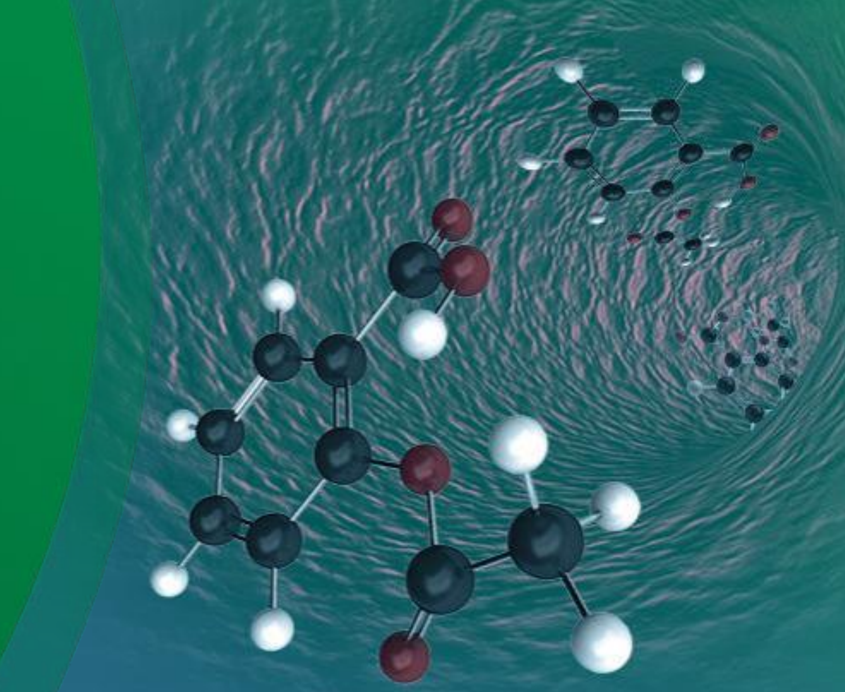


Protein Identification using Machine Learning

Jenny Dunstan

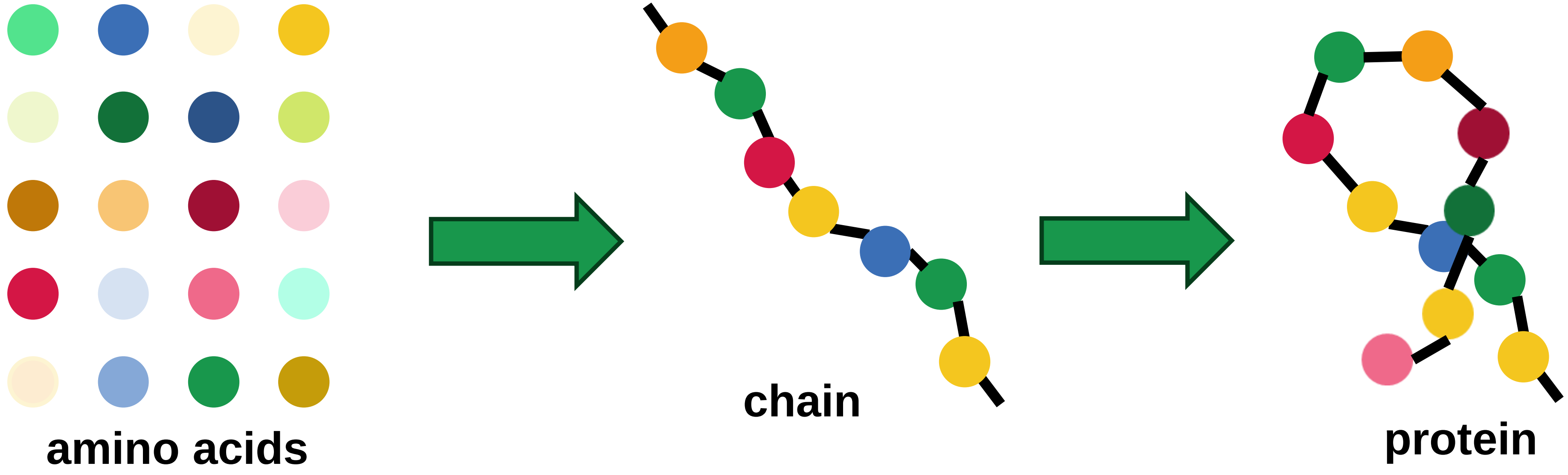
Supervisor: Bikash Bhandari

proteinID

EMBL-EBI

Overview

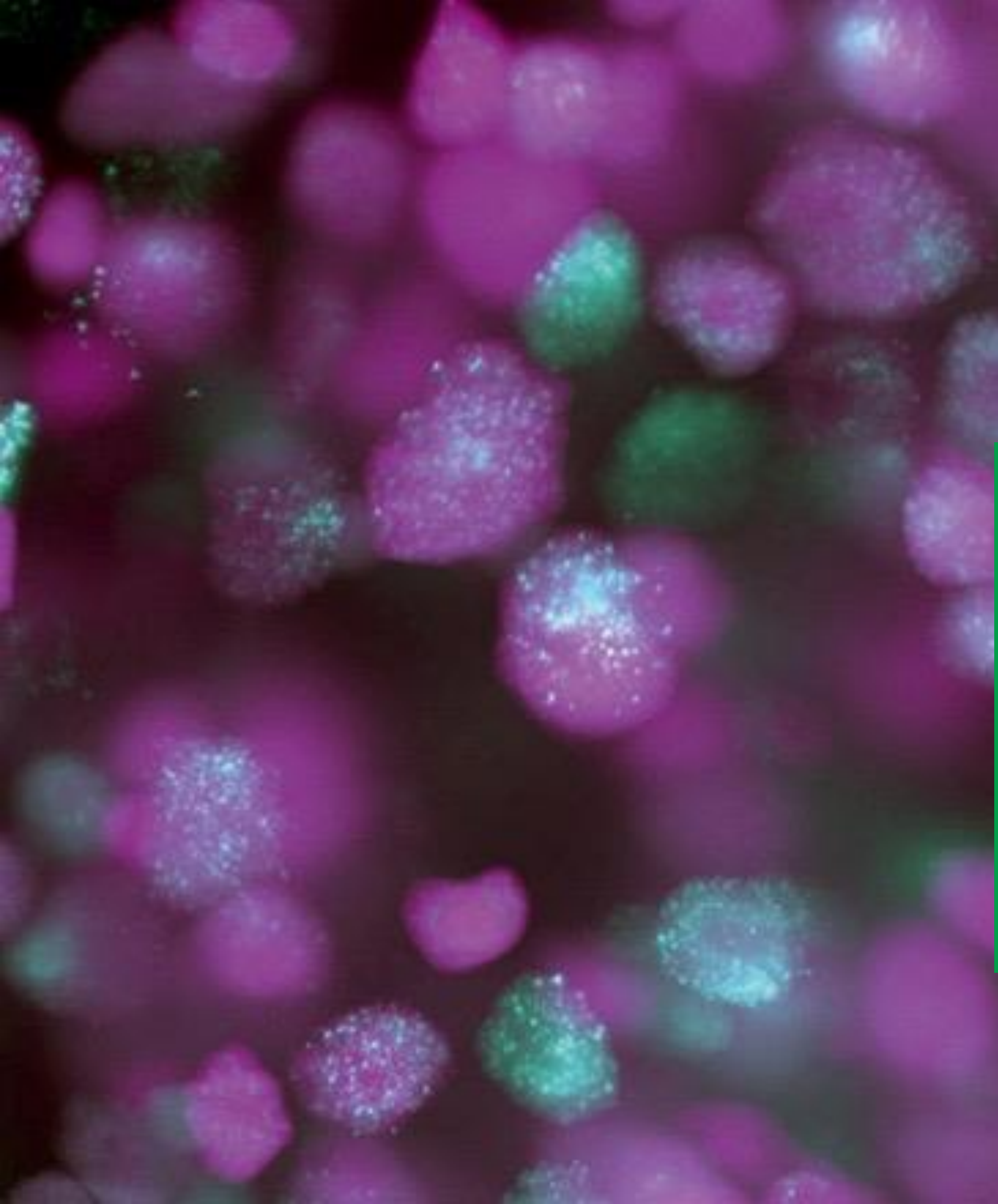
- What is a protein?
- Aims of the project



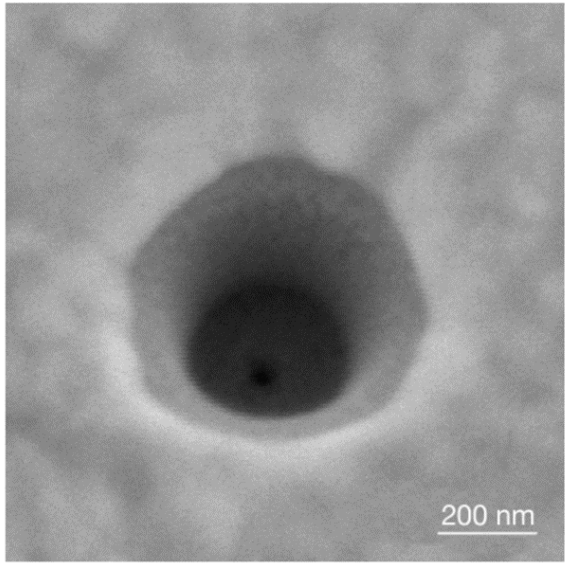
amino acids

chain

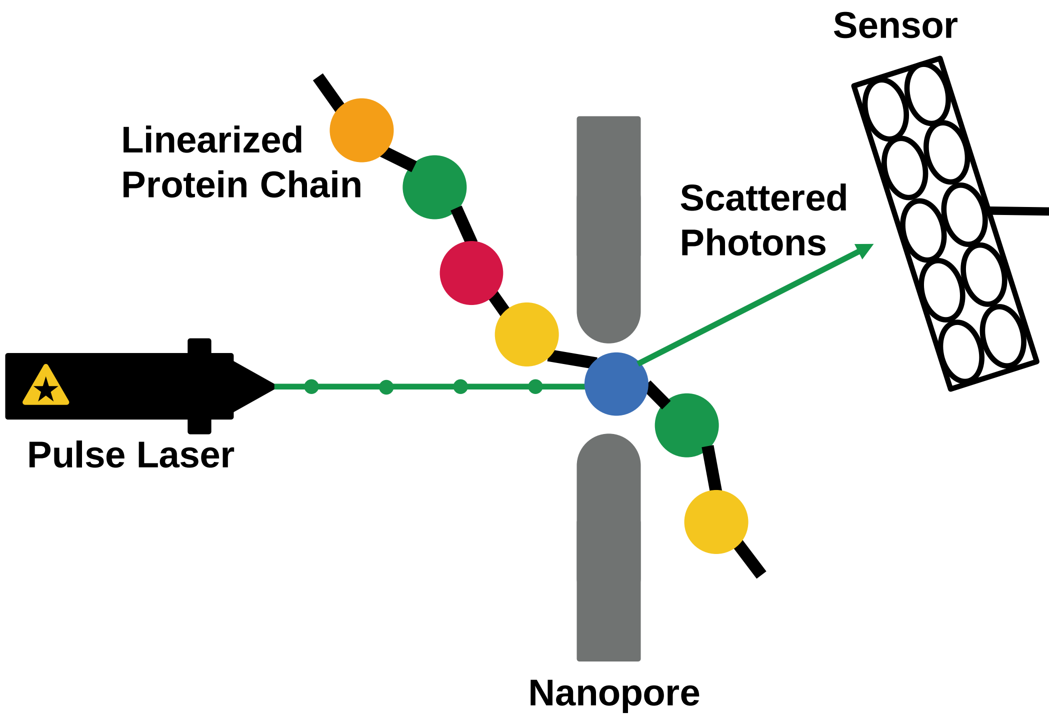
protein

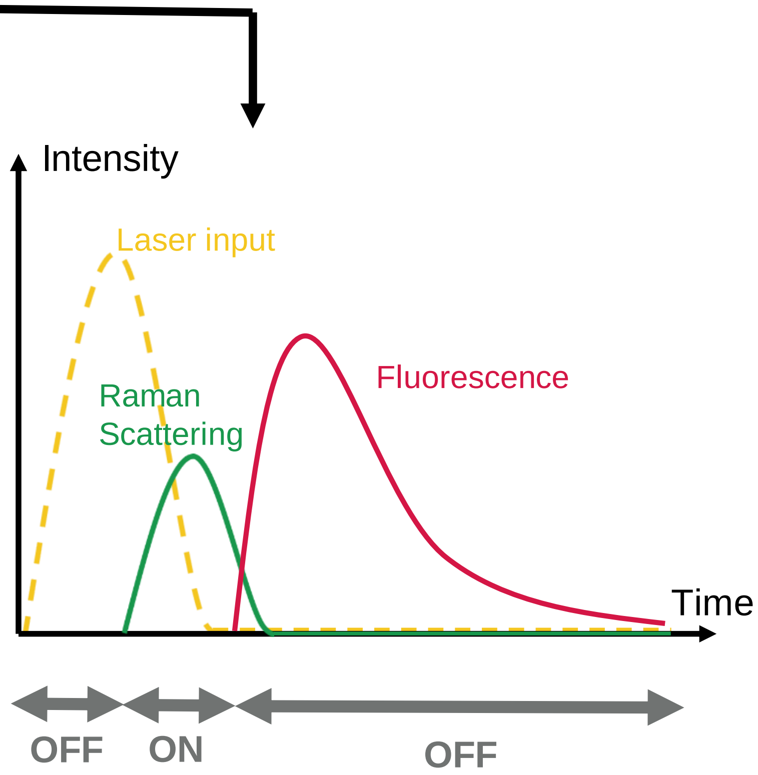


Nanopore Sensor



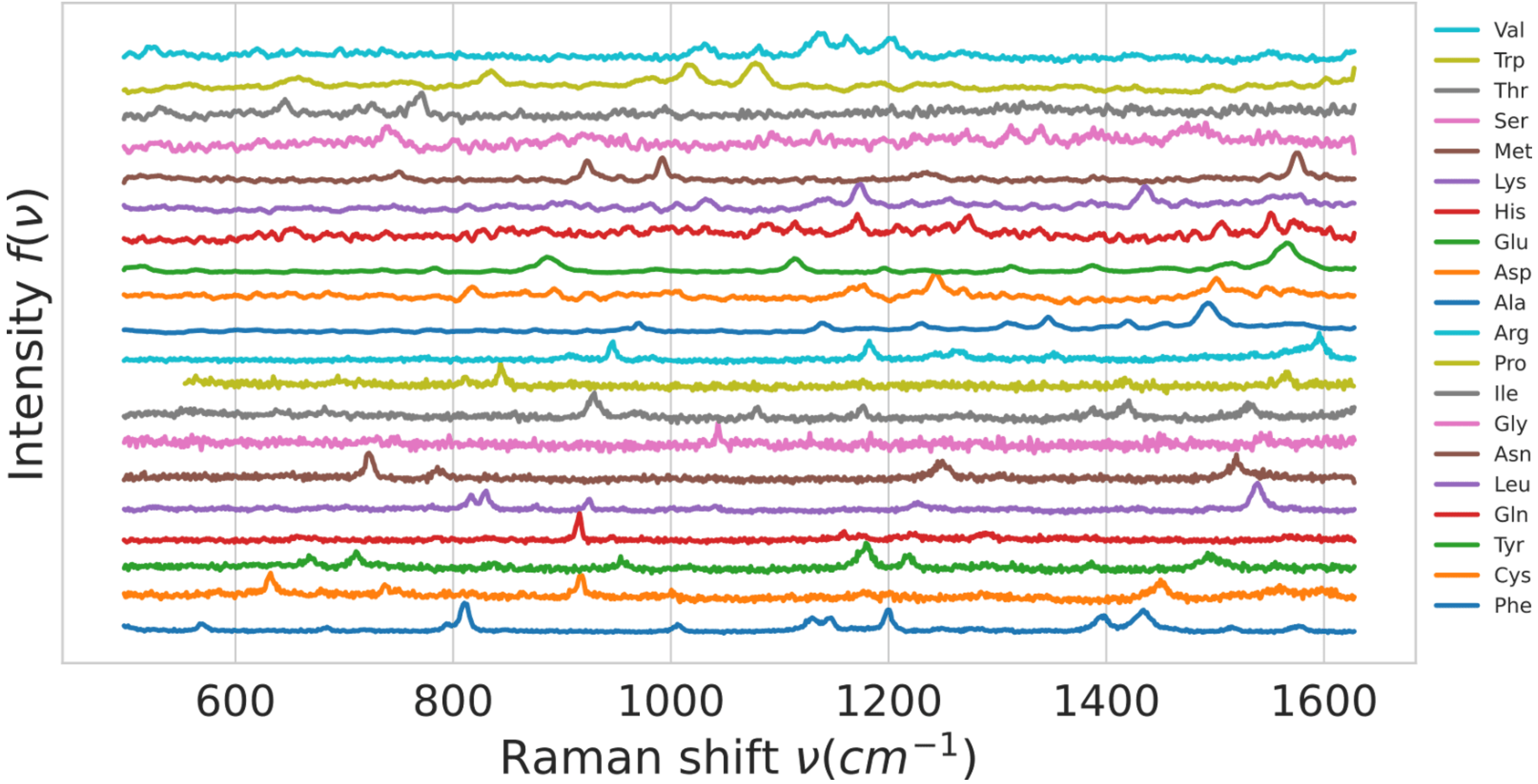
How The Nanopore Sensor Works





t = time amino acid is in nanopore (translocation time)
T = time sensor is switched on for

Output From The Sensor



t = time amino acid is in nanopore (translocation time)
T = time sensor is switched on for

EMBL-EBI

Generating Synthetic Data

Sample from these distributions

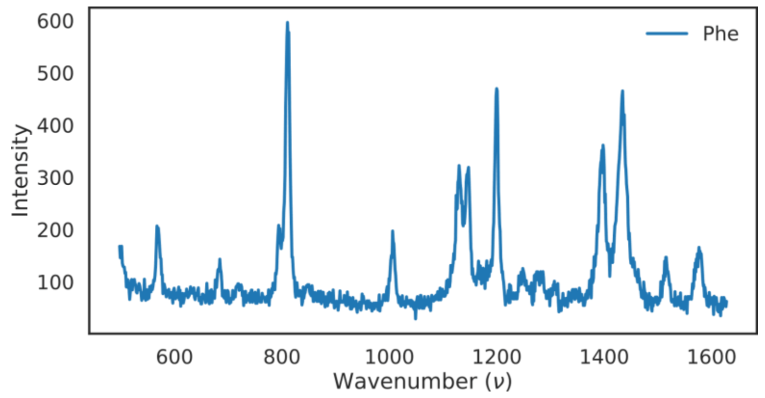
Adjustments:

1. Non-uniform t
2. Different emission amounts
3. Detector Bands

Database of ~19000 human protein sequences

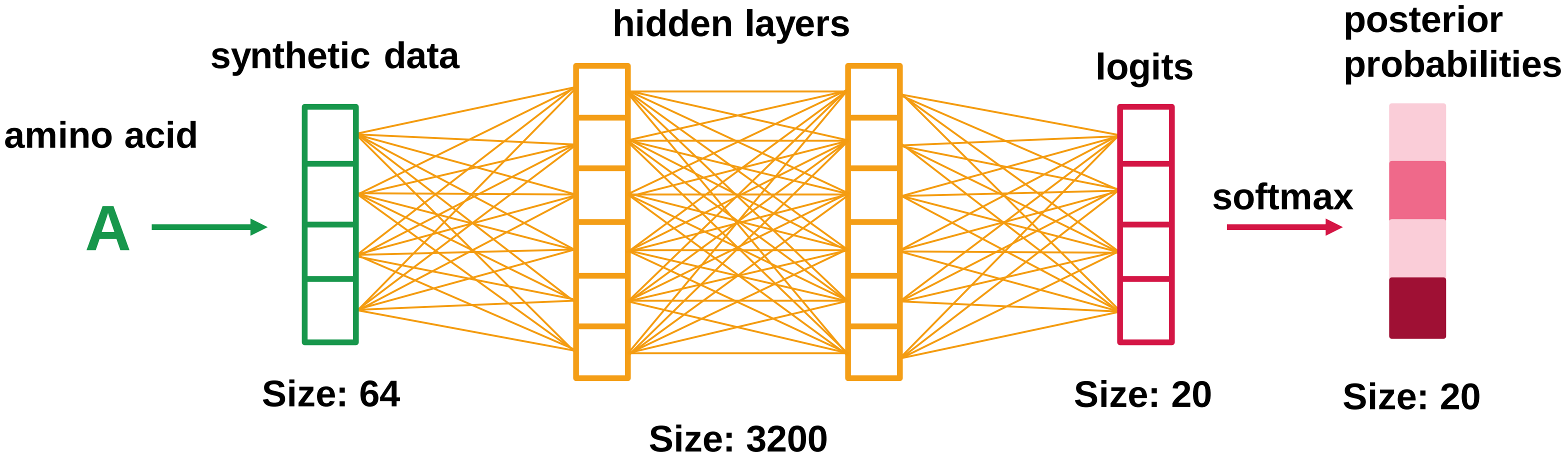
Use 100 amino-acid-length fragments

Single Amino Acid Method



Machine Learning Model for Individual Amino Acids

- Classification of the signals from the 20 different amino acids using a fully connected neural network:



- The predicted acid is one with maximum posterior probability:

posterior probabilities:	0.40	0.15	0.00	0.10	0.05	0.00	0.05	0.00	0.05	0.15	0.05	0.00
possible amino acids:	A	B	C	D	E	F	G	H	I	J	K	L

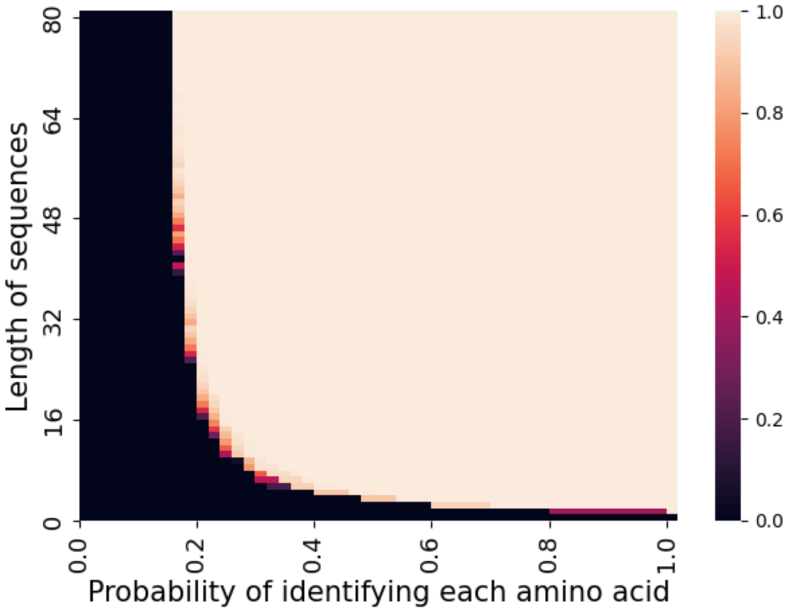
⇨ A

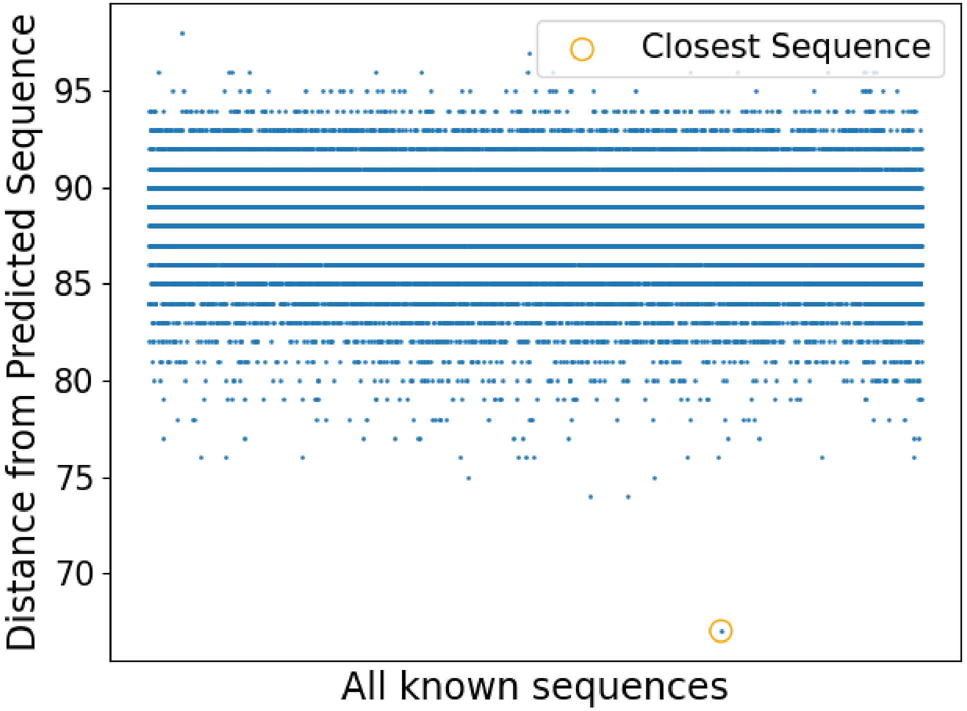
Database Lookup

- Generate a predicted sequence using the machine learning model.

- Compare to a database of known protein sequences

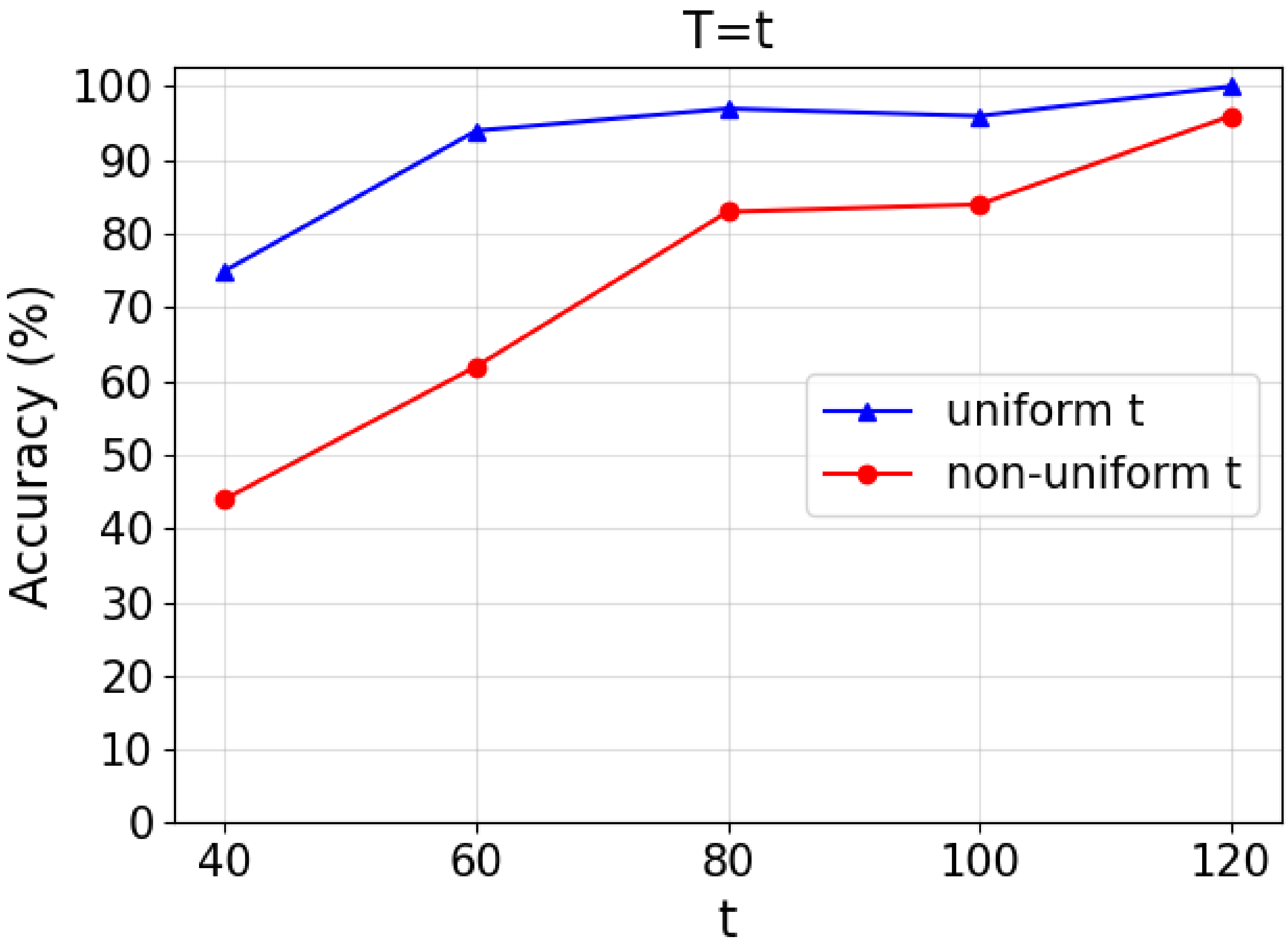
- Probability required



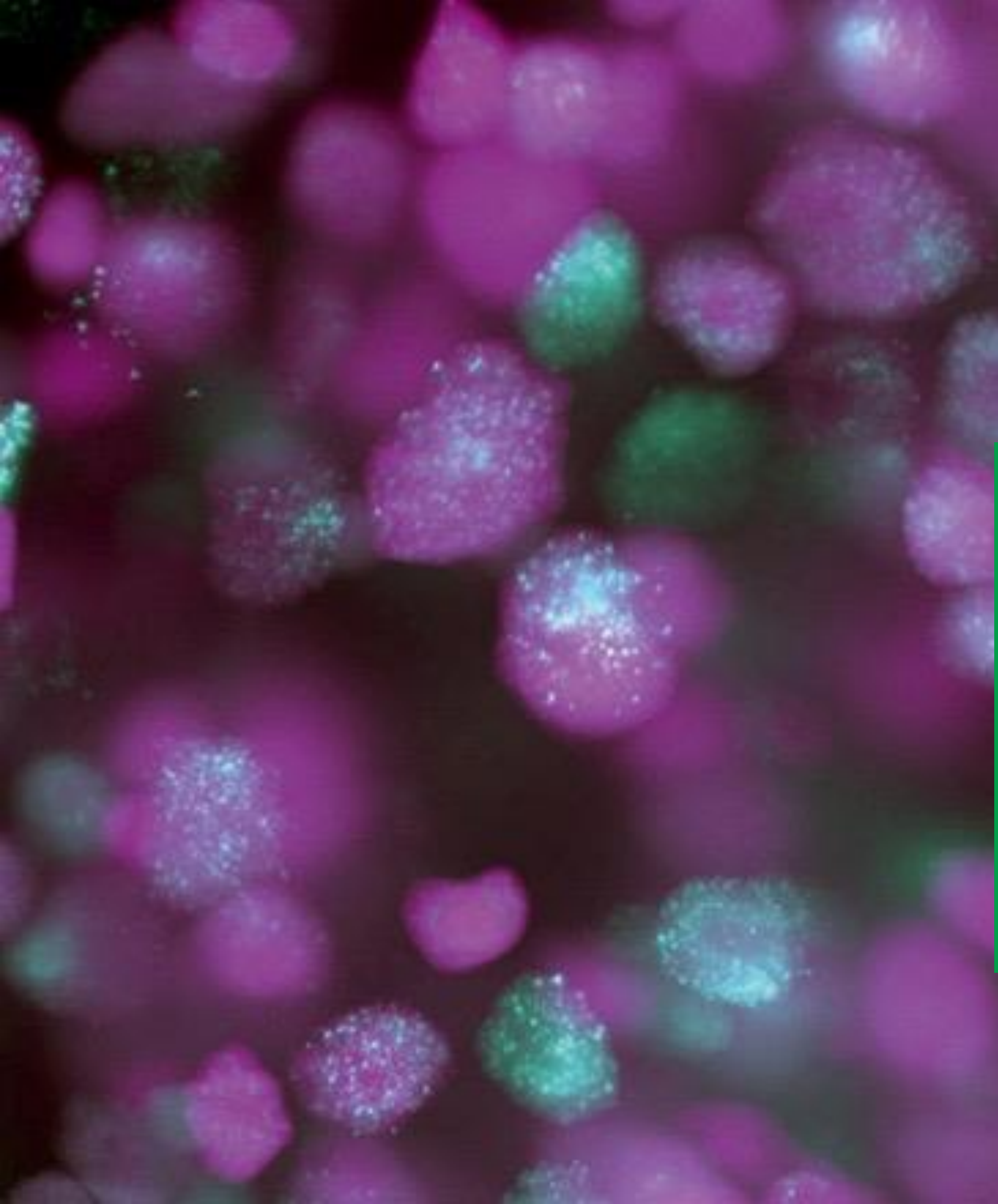


EMBL-EBI

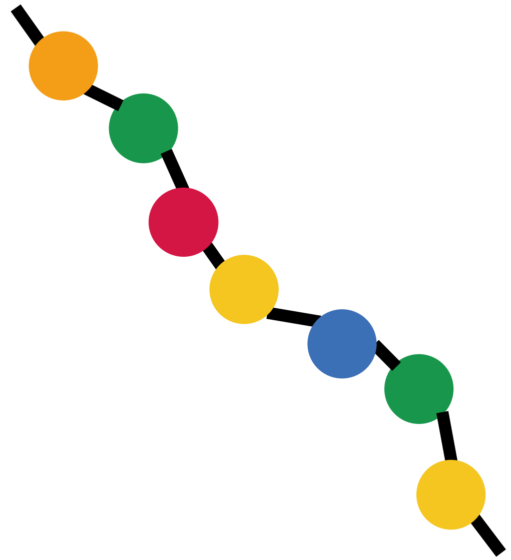
Results



t = time amino acid is in nanopore (translocation time)
T = time sensor is switched on for



Whole Sequence Method



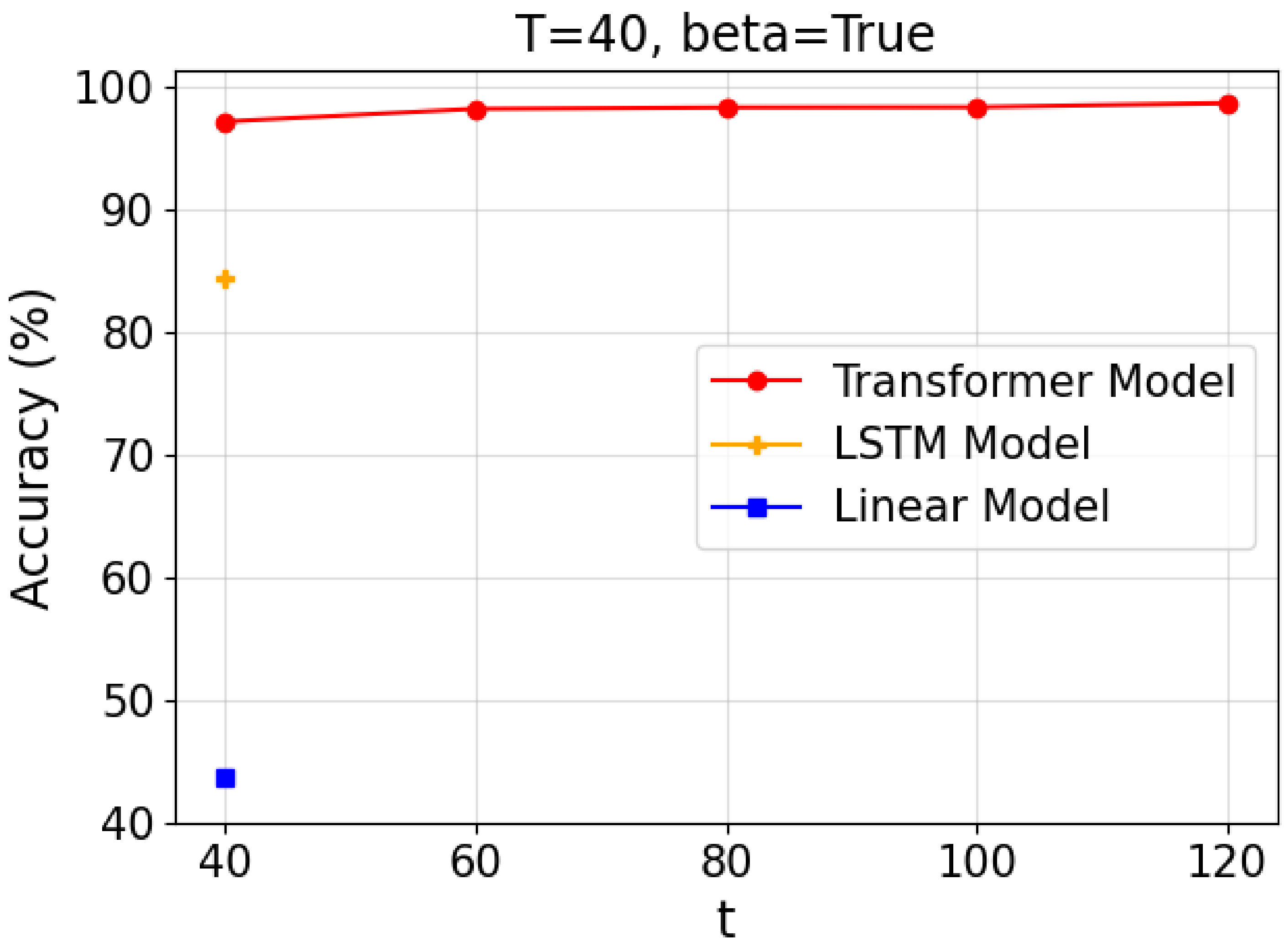
Machine Learning Model for Full Sequences

- We now have a classification model on all 19,200 different sequences (of length 100).
- 5 training data and 1 testing data per sequence
- Results for $T = t = 40$:

Model Type	Accuracy (%)
Linear Neural Network	43.8
LSTM Model	84.3
Vision Transformer	97.2

t = time amino acid is in nanopore (translocation time)
T = time sensor is switched on for

Results

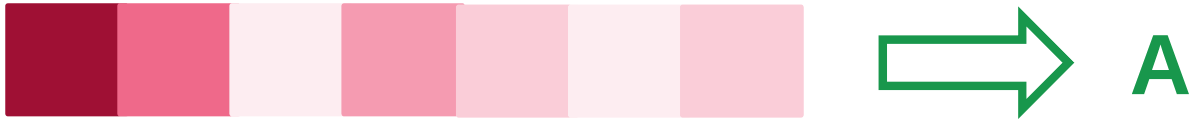


t = time amino acid is in nanopore (translocation time)
T = time sensor is switched on for

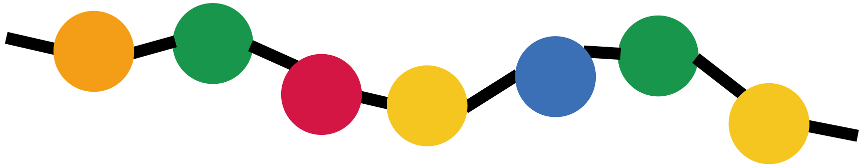
Conclusions and Further Work

Conclusions

Individual Acid Method



Full Sequence Method



Advantages:

- Applicable to sequencing unknown proteins

Disadvantages:

- Two sources of uncertainty (Neural Network and Database Lookup)

Advantages:

- Increased accuracy
- Easier to extend to non-uniform t

Disadvantages:

- Model must be retrained on each database

EMBL-EBI

Further Work

- Improve the accuracy of the models
- Translocation time, t, is not known
- Full length protein sequences
- Insertions/Deletions