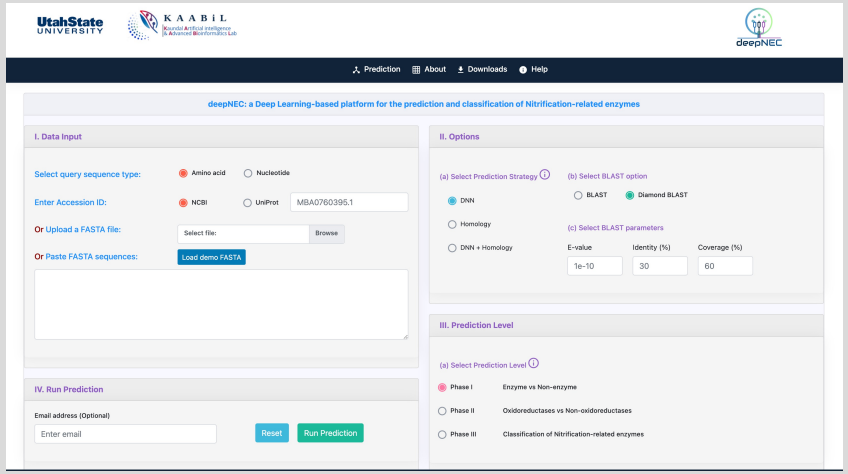
deepNEC: a novel alignment-free tool for the characterization of nitrification-related enzymes using deep learning, a step towards comprehensive understanding of the nitrogen cycle

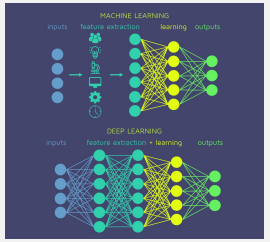
Naveen Duhan

SRS STUDENT RESEARCH SYMPOSIUM 2021

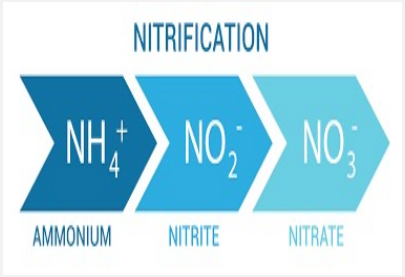
Discipline: Life Sciences
Session 12 PM to 1 PM
Thursday April 14, 2021

deepNEC





Machine learning based nitrification enzymes prediction.



Thirteen nitrification-related enzyme classes.

URL: http://bioinfo.usu.edu/deepNEC/

What is Nitrification

- Nitrification is a process of nitrogen compound oxidation (effectively, loss of electrons from the nitrogen atom to the oxygen atoms) and is catalyzed step-wise by a series of enzymes.

$$2\,NH_4^+ + 3\,O_2 \longrightarrow 2\,NO_2^- + 4\,H^+ + 2\,H_2O$$ (*Nitrosomonas*, *Comammox*)

$$2\,NO_2^- + O_2 \longrightarrow 2\,NO_3^-$$ (*Nitrobacter*, *Nitrospira*, *Comammox*)

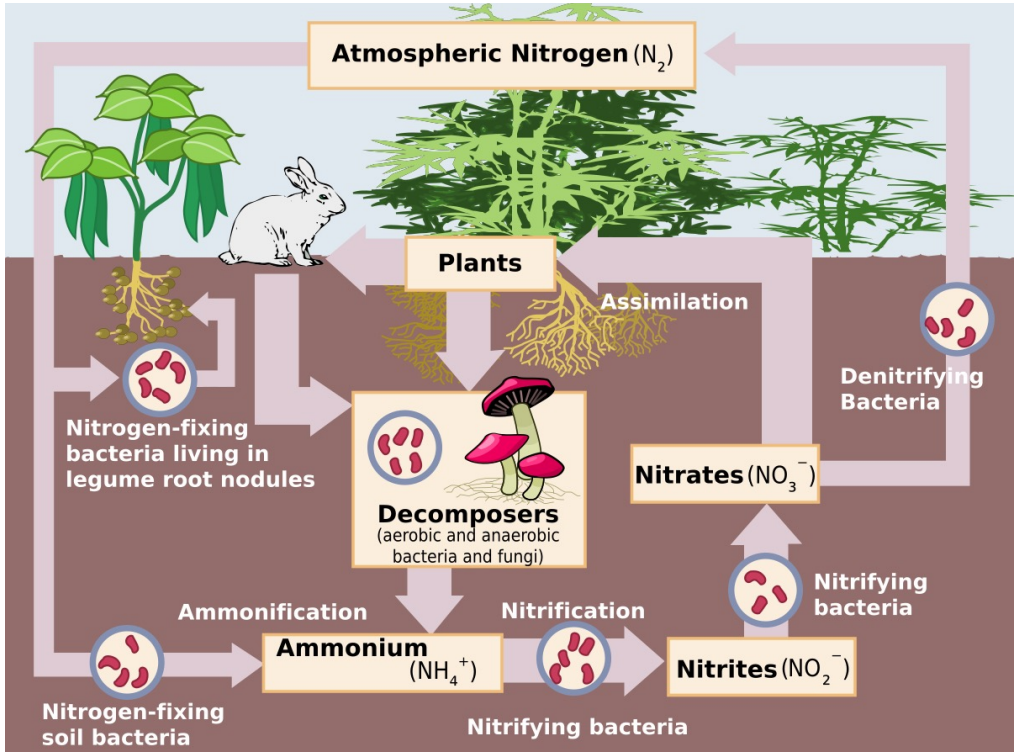
OR

$$NH_3 + O_2 \longrightarrow NO_2^- + 3\,H^+ + 2\,e^-$$

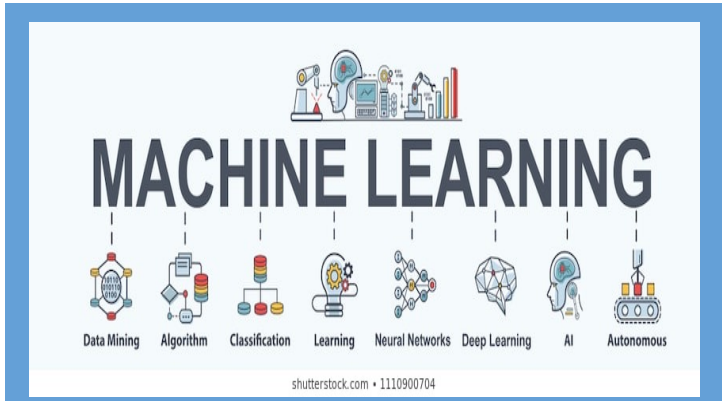
$$NO_2^- + H_2O \longrightarrow NO_3^- + 2\,H^+ + 2\,e^-$$

Why nitrification is important?

- Nitrification is an important microbial two-step transformation in the nitrogen cycle
- In agricultural systems, nitrification is the dominant N-flow system with NO3- making up more than 95 percent of the total N-flow
- Enzyme like ammonia monooxygenase, hydroxylamine dehydrogenase, nitrite oxidoreductase, etc., play a vital role in nitrification process
- Only microorganism performed nitrification



Why Machine learning?



- The easiest and most effective way to functionally annotate microbial genome is employing experimental methods such as enzymatic assays
- Common assumption that similar protein sequences tend to have similar functions
- Homology based methods are widely used to decipher function of an enzyme but fails when there is no significant similarity found
- Most substantially methods is to extract features from protein sequences and train and classify using machine learning approaches

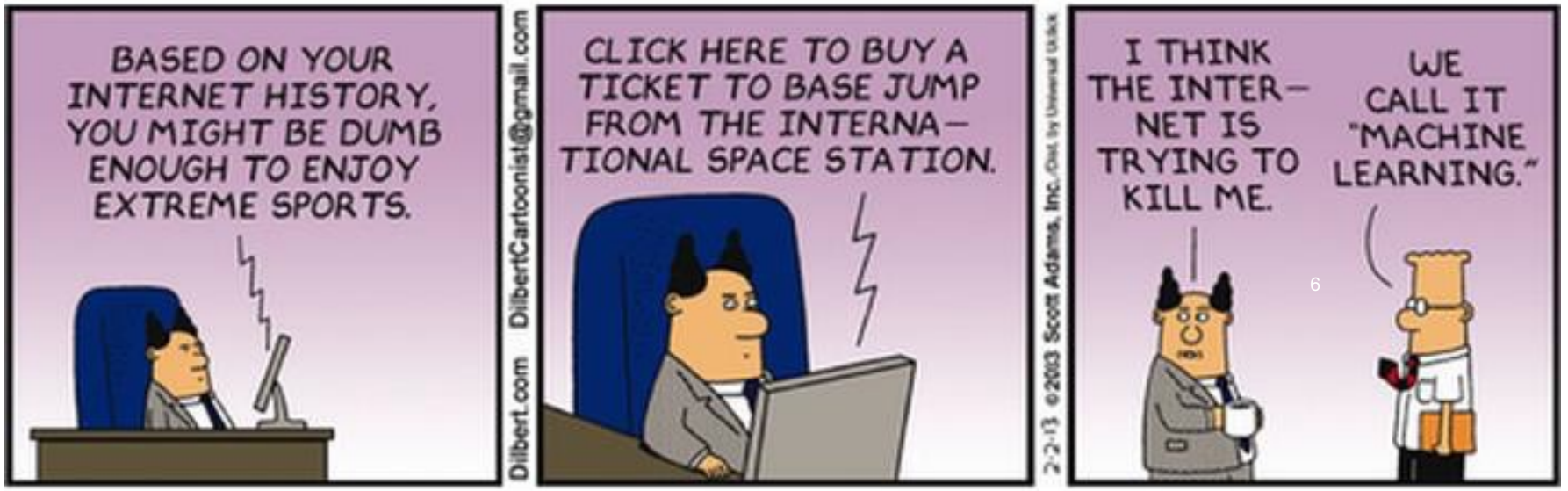
Machine Learning

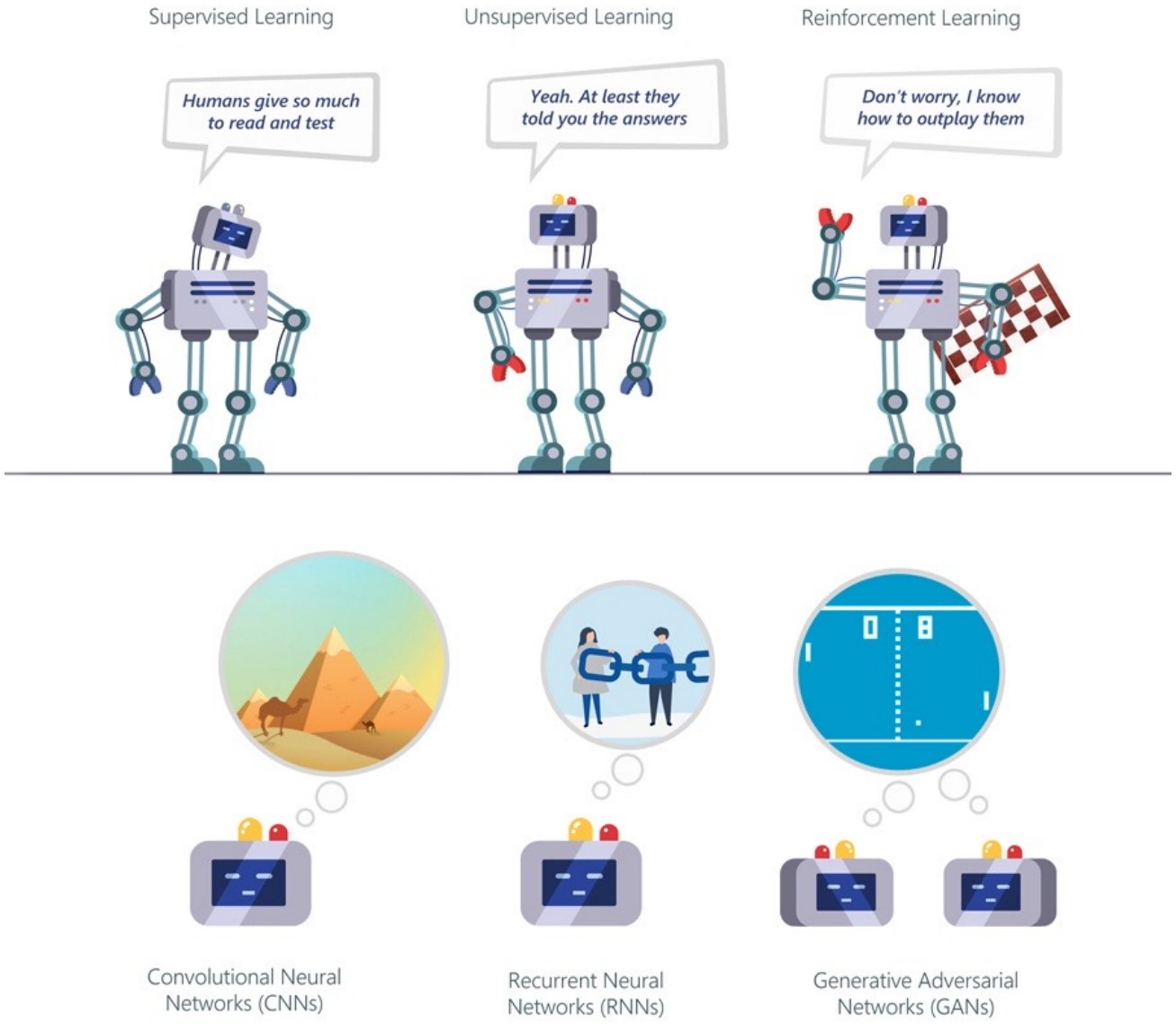
Learning from ‘examples’

It is hard for people to explicitly write the 'rules' for making decisions

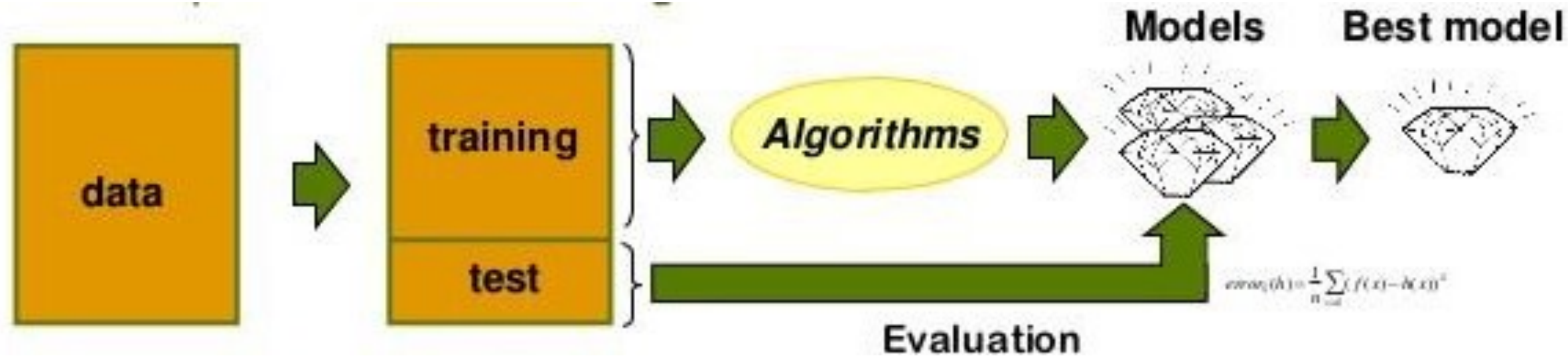
The solution is dependent on lots of complex cases

We don't have the expertise to fully write 'the rules' but we have lots of examples




Supervised Learning
Unsupervised Learning
Reinforcement Learning
Humans give so much to read and test
Yeah. At least they told you the answers
Don't worry, I know how to outplay them
Convolutional Neural Networks (CNNs)
Recurrent Neural Networks (RNNs)
Generative Adversarial Networks (GANs)

Model Quality



Biological Data

Sequence data

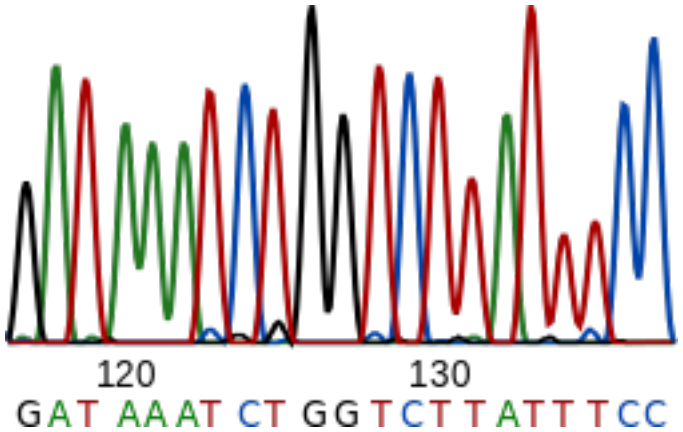
- Protein/DNA sequences

Matrix data

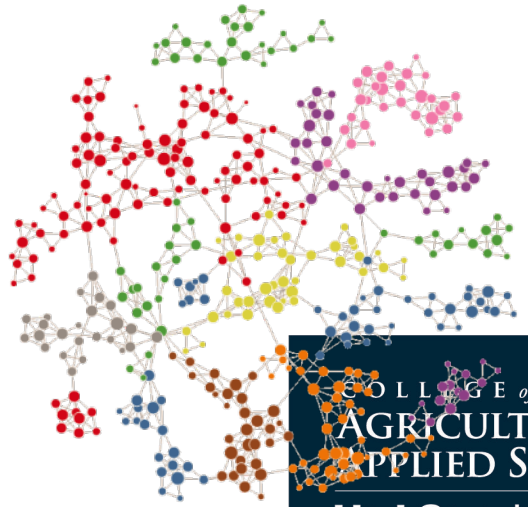
- Gene expression

Network data

- Molecular network

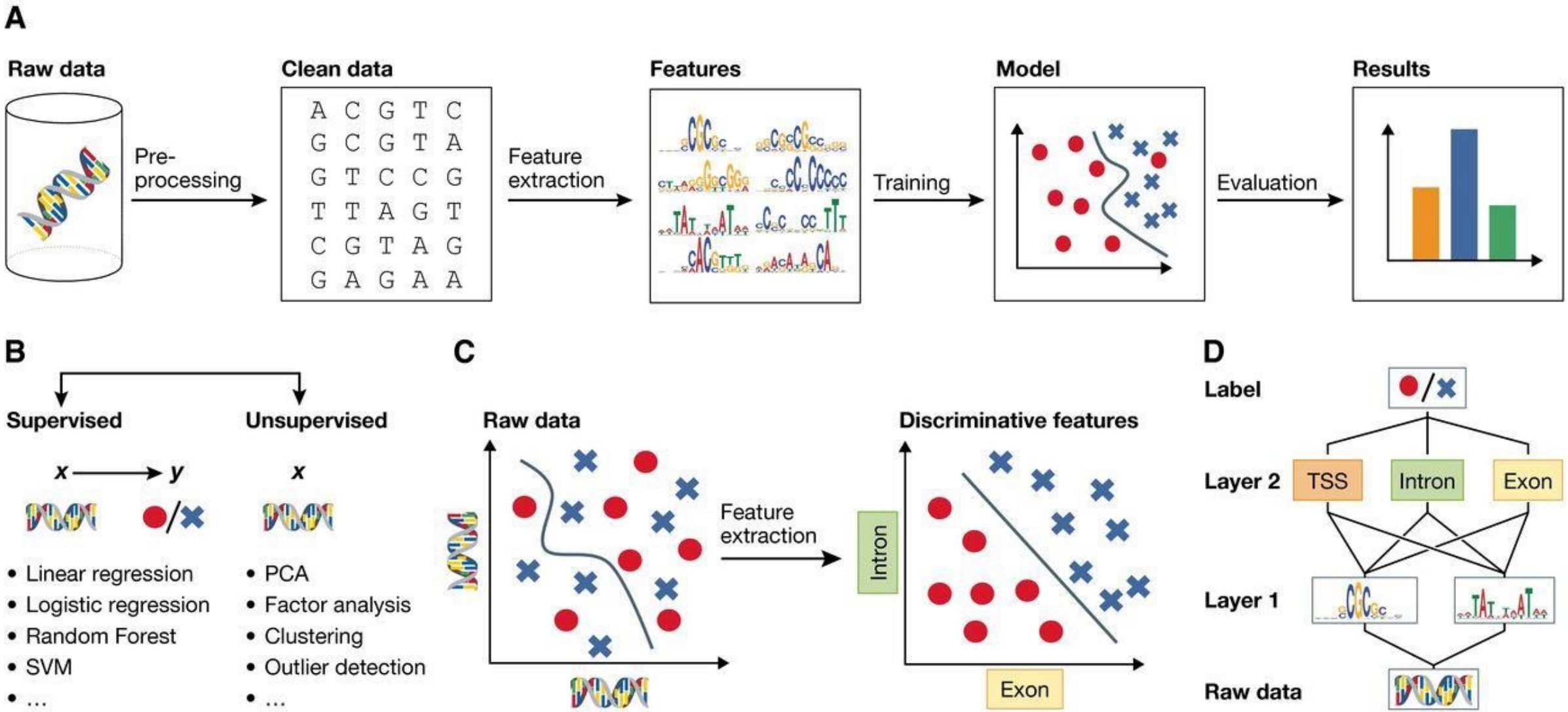


Heterogeneous data



College of Agriculture and Applied Sciences
UtahStateUniversity

Deep learning with Biological data



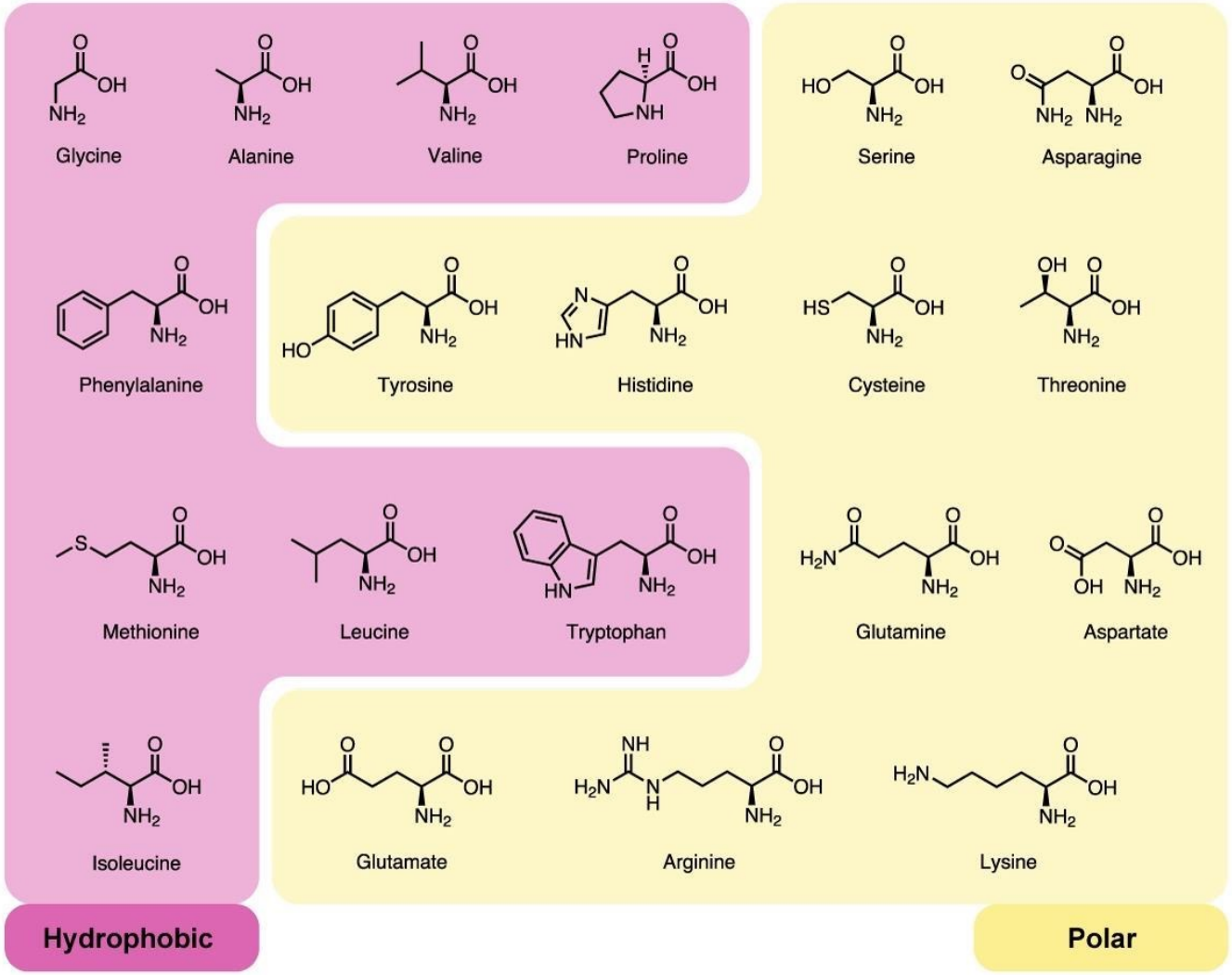
Data Collection

- SWISS-PROT (released on Feb 13, 2019) was downloaded.
- Sequences were separated in Enzymes and Non-enzymes.
- To remove redundancy bias sequence with 40 % similarity were removed
- 28,287 Enzyme were left.
- 28,287 Non-enzyme sequences were taken randomly.

Nitrification enzyme classes

Nitrification-related Enzyme	Training	Independent test
Nitrate reductase (NADH) [EC:1.7.1.1]	2453	100
Nitrate reductase [NAD(P)H] [EC:1.7.1.2]	6814	100
Nitrate reductase (NADPH) [EC:1.7.1.3]	1998	100
Nitric oxide reductase [NAD(P)+, nitrous oxide-forming] [EC:1.7.1.14]	5088	100
Nitrite reductase (NO-forming) [EC:1.7.2.1]	29812	100
Nitric oxide reductase (cytochrome c) [EC:1.7.2.5]	5446	100
Hydroxylamine dehydrogenase [EC:1.7.2.6]	2024	100
Hydrazine synthase [EC:1.7.2.7]	3265	100
Hydrazine dehydrogenase [EC:1.7.2.8]	2720	100
Nitrate reductase (quinone) [EC:1.7.5.1]	19110	100
Ferredoxin-nitrate reductase [EC:1.7.7.2]	5525	100
nitrate reductase (cytochrome) [EC:1.9.6.1]	5362	100
Ammonia monooxygenase [EC:1.14.99.39]	10087	100
Non-nitrification Enzyme	8234	100

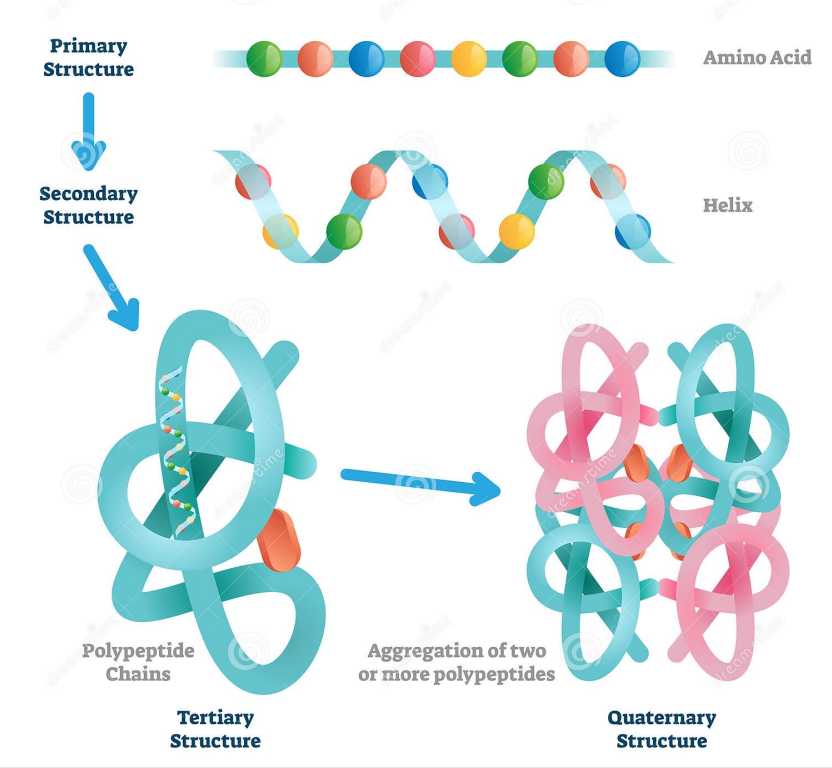
Amino Acids Structure



A A Abbrevation

Amino Acid	Three-Letter Abbreviation	One-Letter Abbreviation
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartate	Asp	D
Cysteine	Cys	C
Glutamate	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Preprocessing of data



- PROTEIN SEQUENCE IS OF VARIABLE LENGTH
- FEATURE GENERATION FROM PROTEIN SEQUENCES
- CONVERT SEQUENCES TO SAME LENGTH VECTOR
- CREATE LABELS AND FEATURES
- MAKE TEST AND TRAIN DATASETS

Manual One-Hot Encoder

- To preserve the original sequence information, we created manual one-hot encoding of the input sequences.
- This encoding uses one 1 and twenty-one 0s to represent each amino acid.
- For example,

```
if (inp == 'Q'):  _res+= '100000000000000000000'
if (inp == 'S'):  _res+= '010000000000000000000'
```

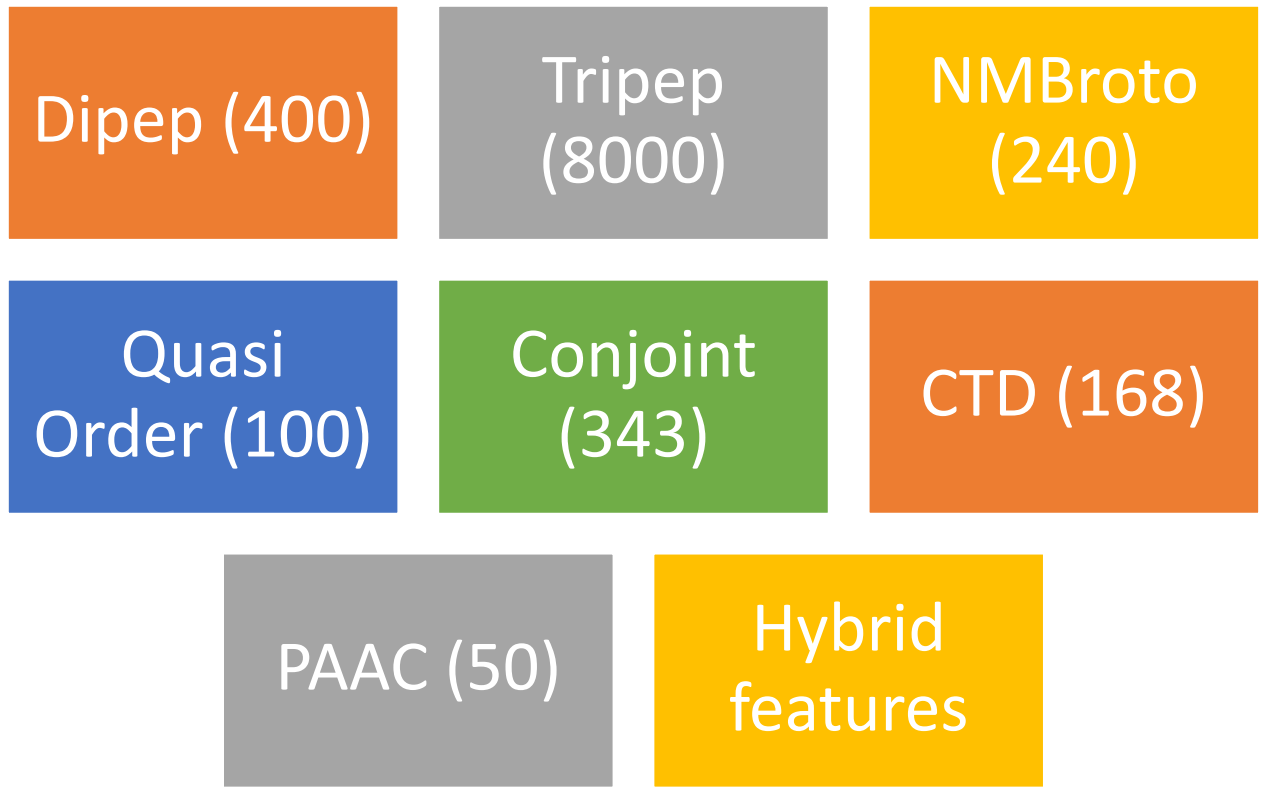
- L by 22 matrix was produced with each row representing a specific spot and each column representing the appearance of certain amino acid.

Amino acid composition: (Vector size 20)

- The amino acid composition describes the fraction of each type of the amino acid in a protein sequence. The fraction is calculated as:
- $f(r) = \frac{N_r}{N} \quad r = 1,2,\dots,20$
- Where N_r is the number of the amino acid type r and N is the length of the sequence.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Seq1	10.02	9.7	2.95	6.96	1.79	6.22	2.74	7.38	2	2.85	9.28	4.43	2.64	3.59	3.69	6.12	5.7	1.58	3.48	6.86
Seq2	9.87	9.99	3.53	5.17	2.23	5.29	1.88	8.58	1.06	2.82	4.47	3.76	2.35	4	8.11	8.58	6.46	1.41	2.82	7.64
Seq3	2.99	4.18	3.28	8.06	1.19	6.87	2.99	5.97	4.18	6.87	9.85	6.27	2.39	5.37	4.48	5.97	3.88	2.69	6.27	6.27
Seq4	9.28	9.28	1.62	7.19	2.55	3.71	1.86	7.89	2.55	2.09	10.21	6.03	1.86	2.78	7.42	6.96	5.8	1.86	1.86	7.19
Seq5	4.56	9.13	2.9	6.22	1.24	7.05	3.32	3.32	3.32	8.3	8.3	5.81	1.66	7.05	4.56	4.98	8.3	0.83	1.66	7.47
Seq6	7.33	6.67	8	2.67	4	4	0.67	10	4	4.67	6.67	3.33	3.33	4.67	4	8	6.67	2.67	2	6.67
Seq7	7.99	6.61	3.31	7.99	1.65	5.51	3.03	6.61	3.31	3.03	9.92	3.03	1.65	2.75	4.68	8.54	5.23	3.86	2.48	8.82
Seq8	7.44	6.34	2.48	6.61	2.48	6.89	2.75	6.06	2.48	2.75	11.02	7.16	4.41	2.75	4.96	6.89	5.51	0.83	2.75	7.44
Seq9	4.23	4.93	9.86	4.23	0.7	3.52	4.93	4.93	2.82	11.27	7.04	6.34	1.41	5.63	2.11	9.15	4.93	1.41	4.93	5.63
Seq10	7.46	3.5	5.06	5.34	0.18	7.73	5.06	6.35	2.76	5.62	9.48	5.52	1.66	4.42	4.14	5.89	7	1.57	5.06	6.17

Protein Features



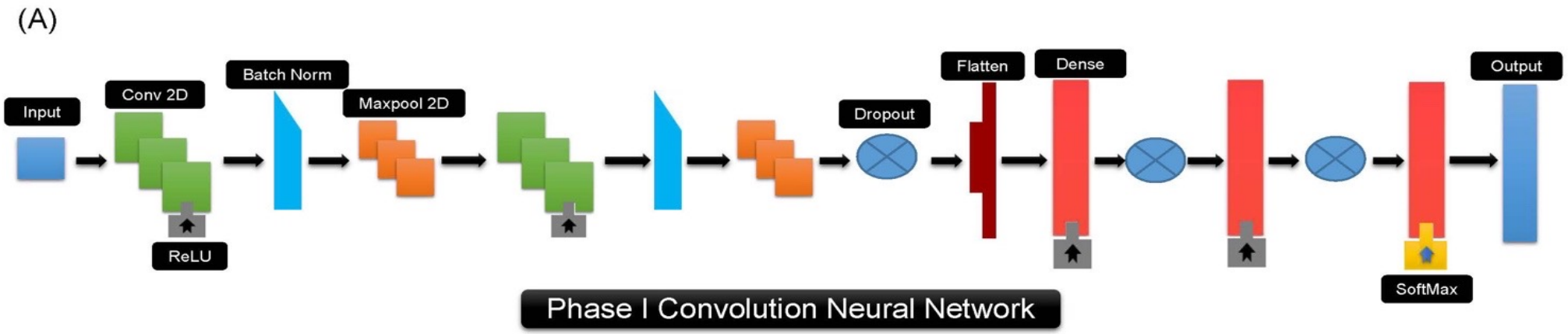
Separate Training for these models:

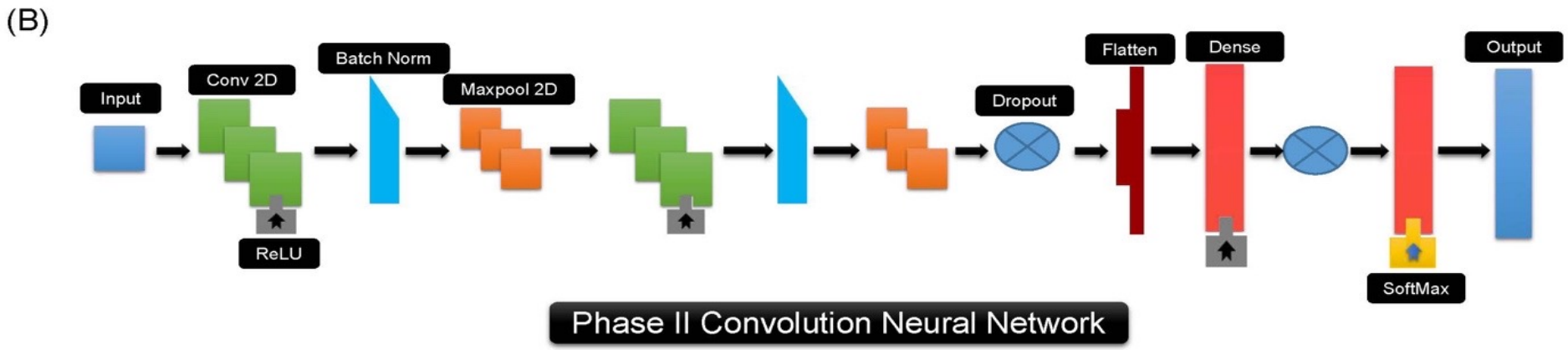
Prediction system works in 2 phases

Enzyme/Non-enzyme model

Nitrification enzyme class model

Model Training Architecture





deepNEC

Statistics measures used to calculate accuracy

Metrics	Formula
True Positives (TP)	True enzymes
True Negatives (TN)	True Non-enzymes
False Positives (FP)	False enzymes
False Negatives (FN)	False non-enzymes
Sensitivity	TPR = TP / (TP + FN)
Specificity	SPC = TN / (FP + TN)
Positive Predictive Value (Precision)	PPV = TP / (TP + FP)
Accuracy	ACC = (TP + TN) / (TP + TN + FP + FN)
F1 Score	F1 = 2TP / (2TP + FP + FN)
Matthews Correlation Coefficient	MCC = (TP x TN – FP x FN) / (sqrt((TP + FP) x (TP + FN) x (TN + FP) x (TN + FN)))

Phase I statistics

Enzyme

Non-enzyme

Metrics	Training average 10-FOLD	Independent testing
Sensitivity	95.76%	94.47%
Specificity	95.64%	92.40%
Precision	95.64%	92.55%
Accuracy	95.70%	93.43%
F1-score	95.70%	93.50%
MCC	0.914	0.868

Phase II statistics: Training | Nitrification-related Enzymes

Nitrification Enzyme	Sensitivity	Specificity	Precision	Accuracy	F1score	MCC
EC:1.7.1.1	71.11	99.85	92.02	99.20	80.14	0.807
EC:1.7.1.2	92.78	98.81	83.97	98.43	88.15	0.877
EC:1.7.1.3	66.58	99.62	76.87	99.01	71.21	0.711
EC:1.7.1.14	98.20	98.85	80.88	98.82	88.69	0.885
EC:1.7.2.1	99.96	99.98	99.95	99.98	99.96	0.999
EC:1.7.2.5	78.15	99.90	97.58	98.80	86.78	0.867
EC:1.7.2.6	96.44	99.97	98.50	99.91	97.45	0.974
EC:1.7.2.7	99.49	99.94	98.18	99.93	98.83	0.988
EC:1.7.2.8	97.41	99.94	97.54	99.87	97.47	0.974
EC:1.7.5.1	99.65	99.85	99.31	99.82	99.48	0.994
EC:1.7.7.2	99.37	99.94	98.83	99.91	99.10	0.991
EC:1.9.6.1	99.72	99.98	99.53	99.96	99.63	0.996
EC:1.14.99.39	100	100	99.99	100	100	1.00
Non-nitrification	99.23	99.94	99.28	99.89	99.25	0.992

Phase II statistics : Independent testing

Nitrification Enzyme	Sensitivity	Specificity	Precision	Accuracy	F1score	MCC
EC:1.7.1.1	70.00	99.42	90.91	97.15	79.10	0.784
EC:1.7.1.2	87.00	95.75	63.04	95.08	73.11	0.716
EC:1.7.1.3	68.00	98.92	83.95	96.54	75.14	0.738
EC:1.7.1.14	99.00	98.33	83.19	98.38	90.41	0.899
EC:1.7.2.1	100.00	100.00	100.00	100.00	100.00	1.000
EC:1.7.2.5	80.00	99.92	98.77	98.38	88.40	0.881
EC:1.7.2.6	94.00	99.75	96.91	99.31	95.43	0.951
EC:1.7.2.7	97.00	99.75	97.00	99.54	97.00	0.968
EC:1.7.2.8	95.00	99.33	92.23	99.00	93.60	0.931
EC:1.7.5.1	100.00	99.42	93.46	99.46	96.62	0.964
EC:1.7.7.2	98.00	100.00	100.00	99.85	98.99	0.989
EC:1.9.6.1	99.00	100.00	100.00	99.92	99.50	0.995
EC:1.14.99.39	100.00	100.00	100.00	100.00	100.00	1.000
Non-nitrification	98.00	99.61	95.14	99.50	96.55	0.963

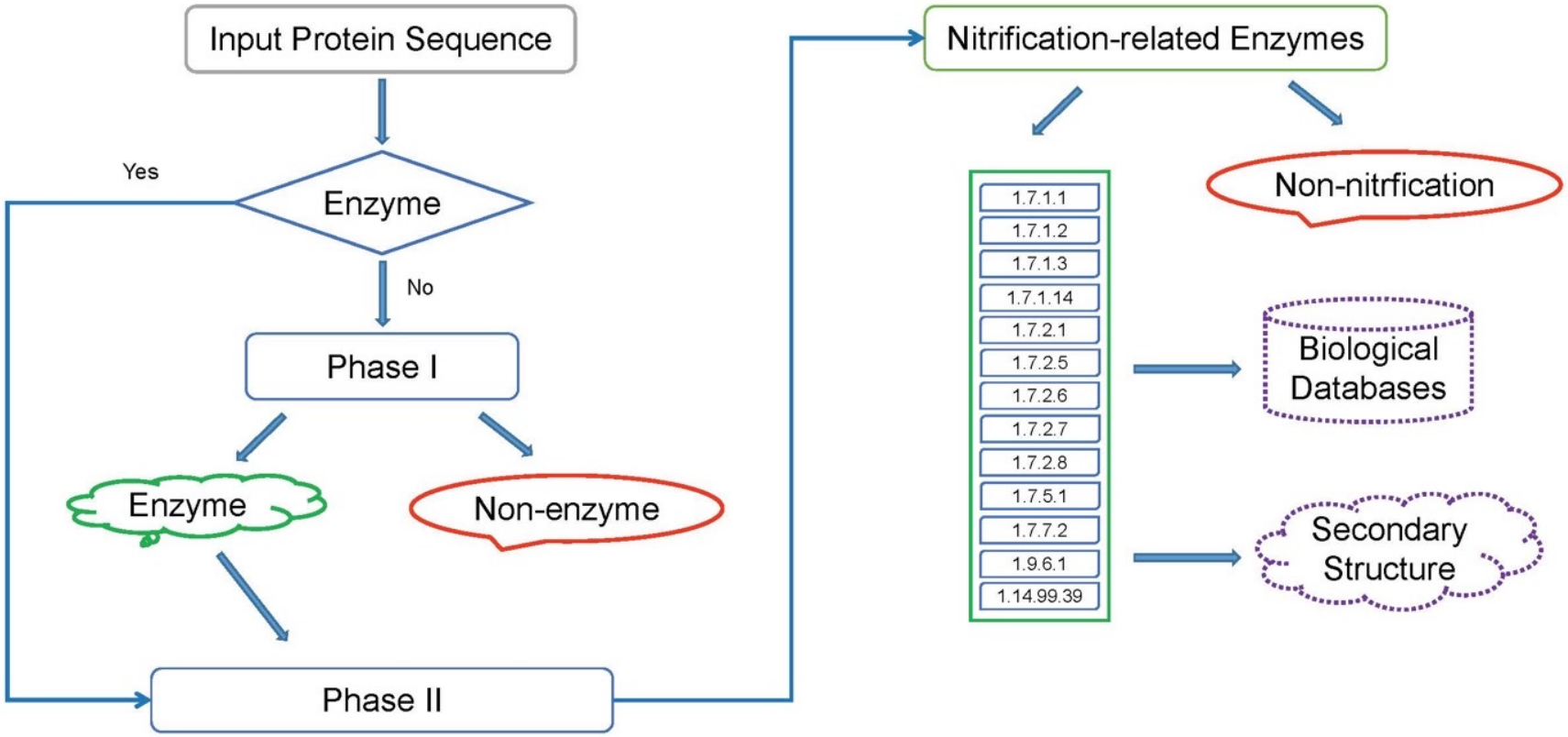
Comparison with other tools

Metrics	deepNEC	ECPred[1]	DeepEC[2]
Sensitivity	96	96	76
Specificity	88	56	72
Precision	88.89	68.57	73.08
Accuracy	92	76	74
F1-score	92.31	80	74.51
MCC	0.843	0.567	0.480

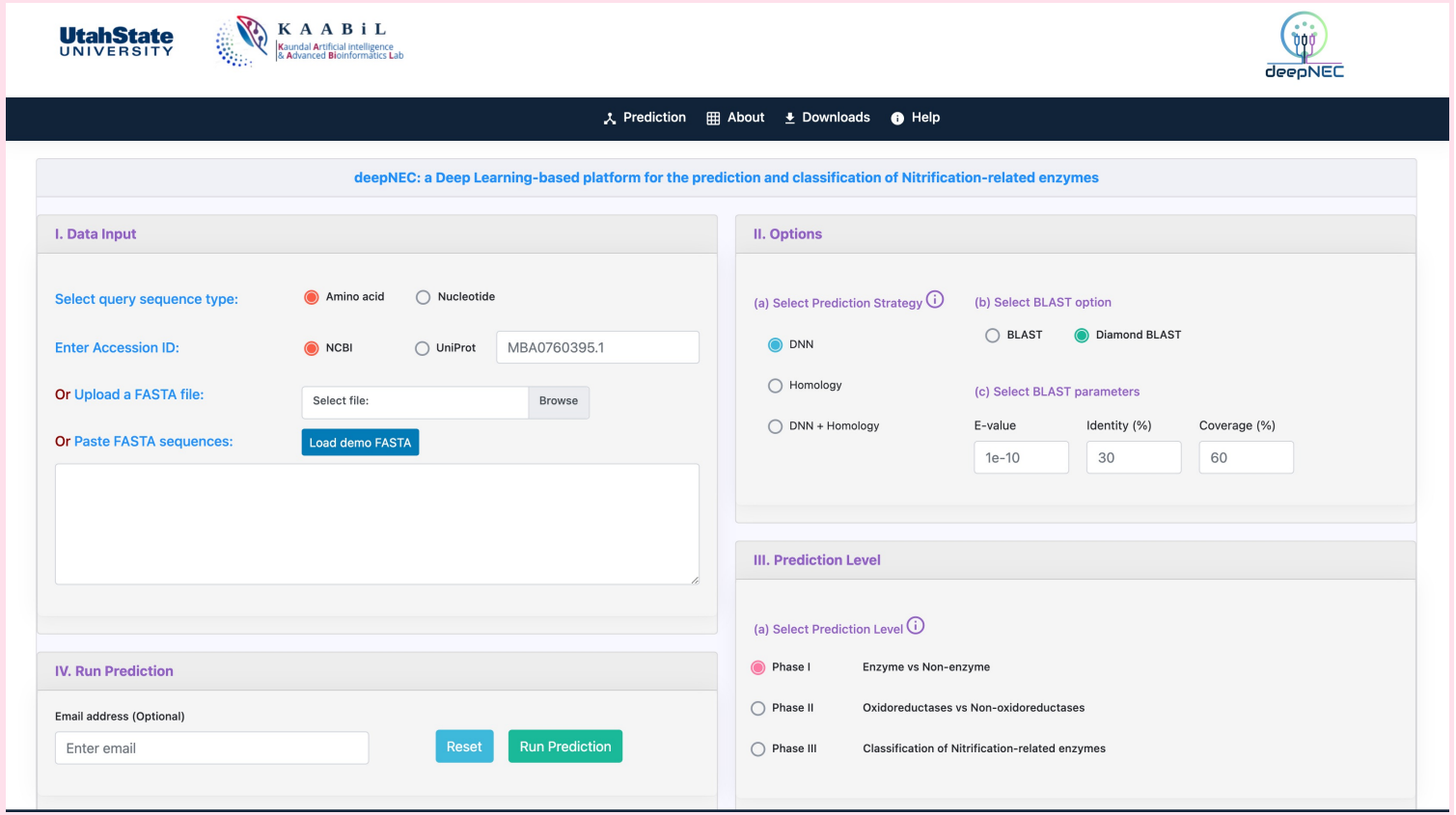
[1]Dalkiran, A., Rifaioglu, A.S., Martin, M.J. *et al.* ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. *BMC Bioinformatics* **19,** 334 (2018).

[2]Ryu, J. Y., Kim, H. U. & Lee, S. Y. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. *Proc. Natl. Acad. Sci.* **116**, 13996 LP – 14001 (2019).

deepNEC webserver workflow



Application of deepNEC tool



Novel functional genes involved in nitrification

Agriculture

Microbial biotechnology

Wastewater Treatment and Reuse

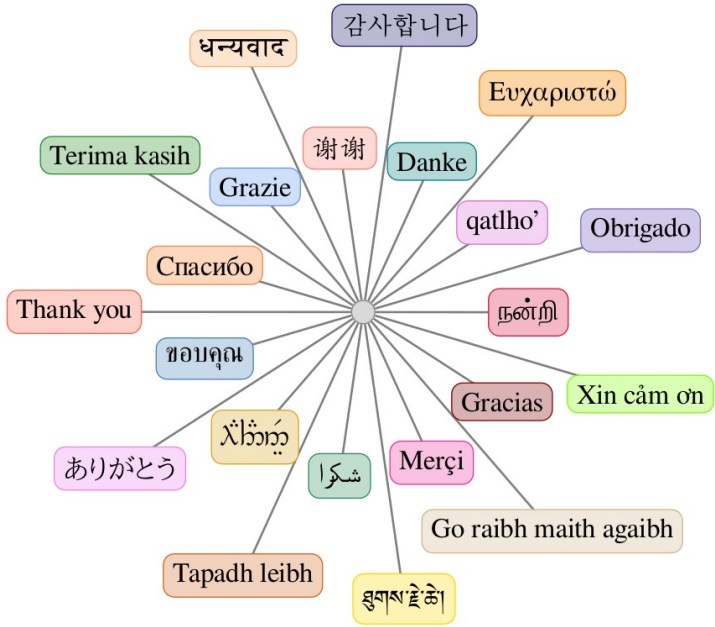
Contact

Naveen Duhan
Utah State University
naveen.duhan@usu.edu

Rakesh Kaundal
Utah State University
rakundal@usu.edu

For More Information:

http://bioinfo.usu.edu/



UtahState University