Deep Learning for Protein Classification

BY AAROHI CHOPRA

BASED ON: PROTEIN FAMILY CLASSIFICATION WITH NEURAL NETWORKS

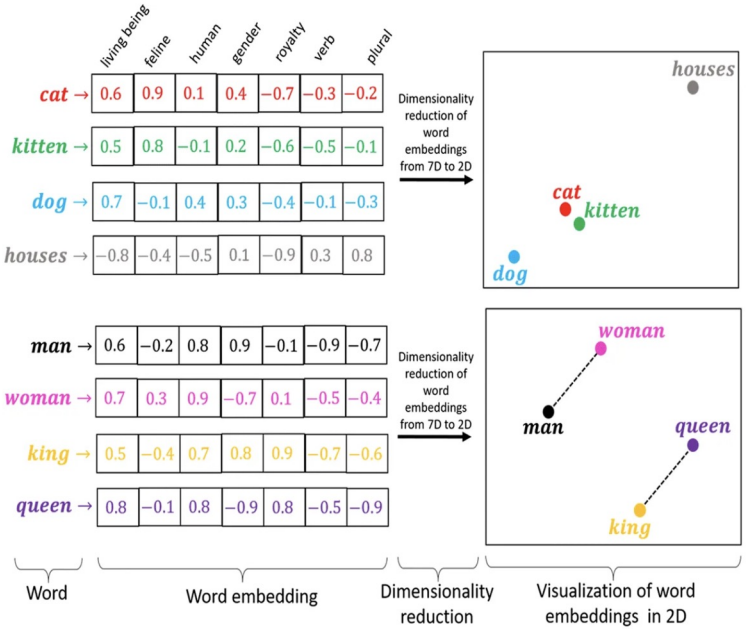
Contents

You can visit this page for my notes

https://docs.google.com/document/d/130JtG5lVGhnKk_ozJXrYRrJh7h4O7wEbo6jP-y3XS9g/edit?usp=sharing

Brief Background

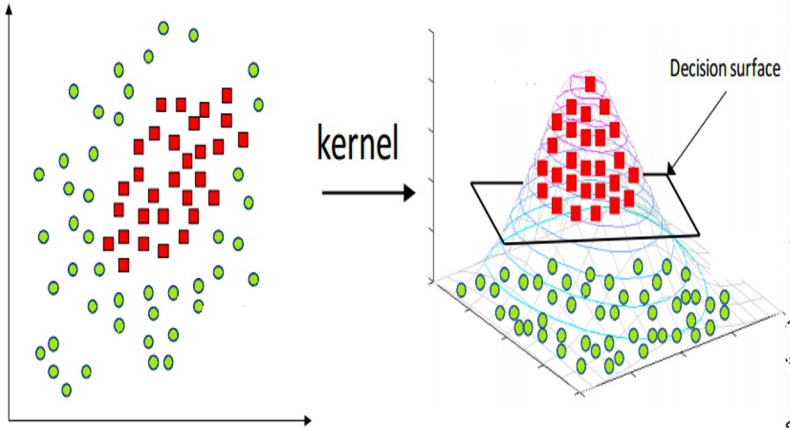
EMBEDDINGS



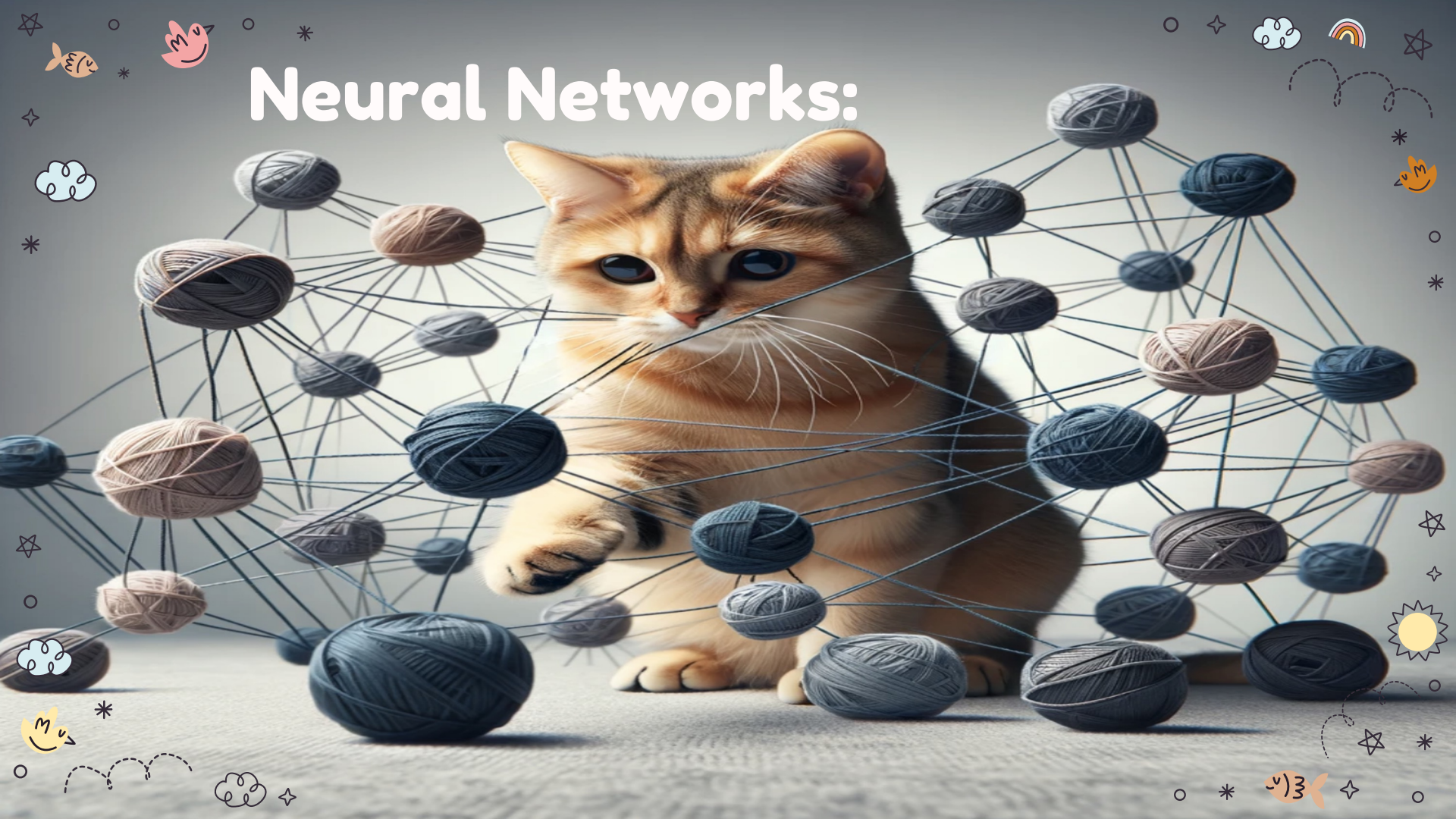
WORD2VEC VS GLOVE

SVM: the base

- THE GENERAL IDEA: FIND A HYPERPLANE IN N DIMENSIONAL SPACE THAT CLASSIFIES DATA POINTS, ALONG WITH HAVING MAXIMAL MARGINS
- SUPPORT VECTORS?
- WHAT IF DATA IS NONLINEAR? USE KERNEL FUNCTION TO MAP DATA TO HIGHER DIM WHERE IT IS SEPARABLE



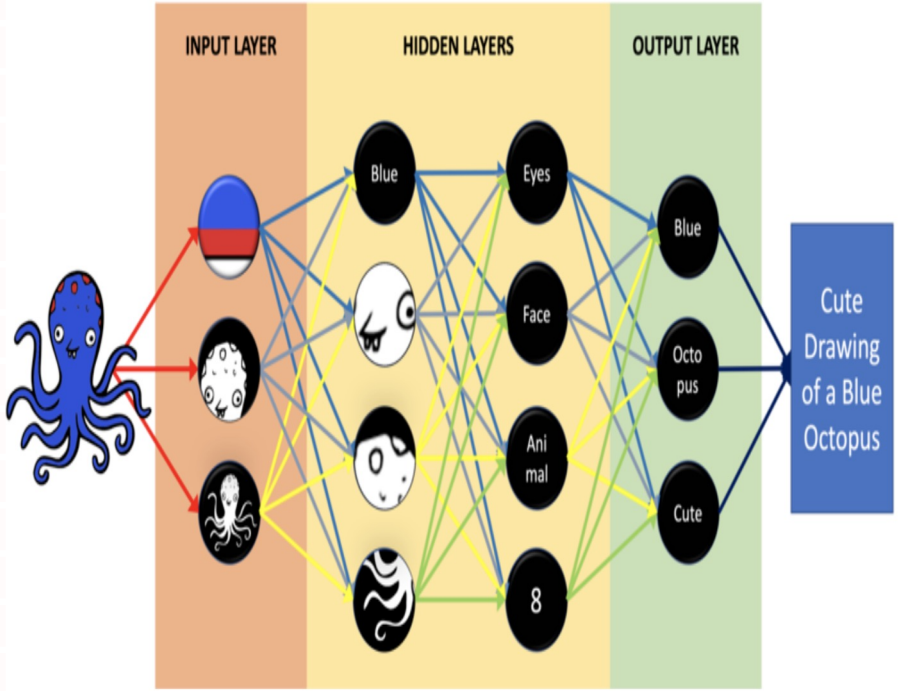
Neural Networks:



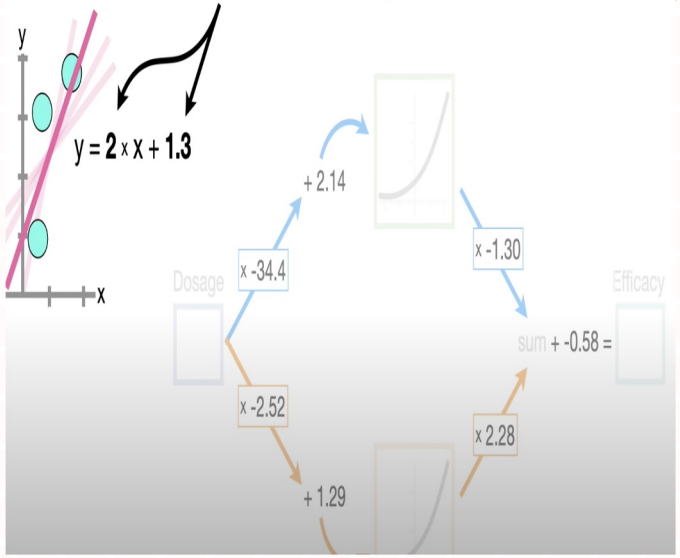
NN:

BASIC IDEA:

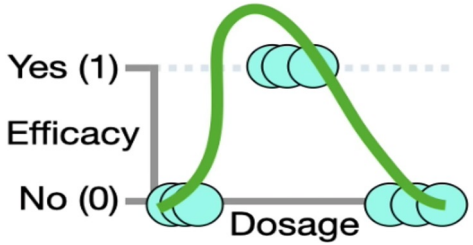
- ARTIFICIAL NEURAL NETWORKS ARE COMPOSED OF LAYERS OF NODE
- A NODE REPRESENTS NEURON IN THE BRAIN
- THE FIRST LAYER IS THE INPUT LAYER, FOLLOWED BY HIDDEN LAYERS FINALLY THE OUTPUT LAYER
- EACH NODE IN THE NEURAL NET PERFORMS SOME SORT OF CALCULATION, WHICH IS PASSED ON TO OTHER NODES DEEPER IN THE NETWORK



A BIT MORE DETAIL:



1. NEURAL NETWORK CAN FIT THE GREEN SWIGGLE!!!!
2. PARAMETERS ESTIMATES ARE ANALOGOUS TO THE SLOPE AND INTERCEPT VALUE THAT WE SOLVE WHEN WE FIT A STRAIGHT LINE TO THE DATA
3. WE USE ACTIVATION FUNCTIONS ALONG WITH PARAMETERS WHICH ARE EVENTUALLY ADDED TO MAKE THE GREEN SWIGGLE
4. THE MAGIC IS IN ADJUSTING THE WEIGHTS.



GRUs

- RNNS

  MEMORY RETENTION

  CHALLENGES VANISHING GRADIENT EXPLODING GRADIENT
- GATES

  DECIDE WHETHER INFORMATION SHOULD BE PASSED ALONG

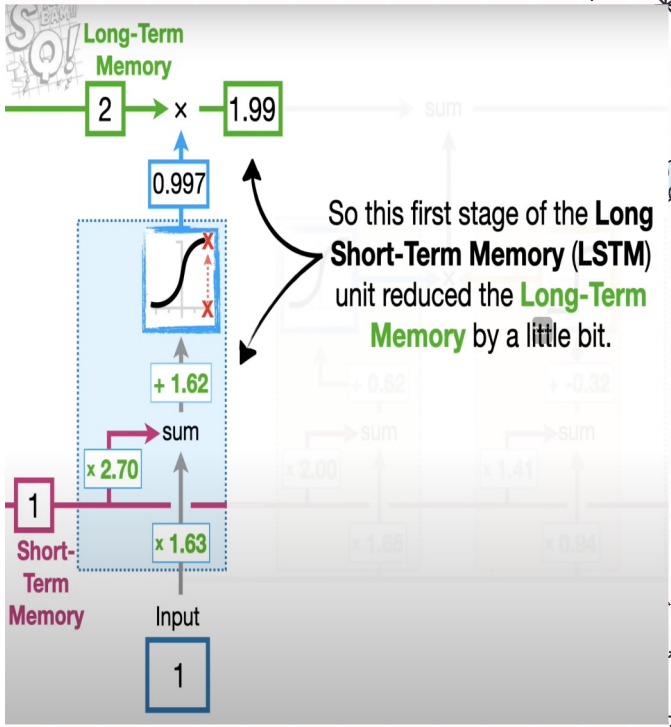
  WHAT SHOULD BE CONTINUED TO REMEMBER
- TYPES OF GATES:

  RESET GATE: DETERMINES HOW TO COMBINE THE NEW INPUT WITH THE PREVIOUS MEMORY.

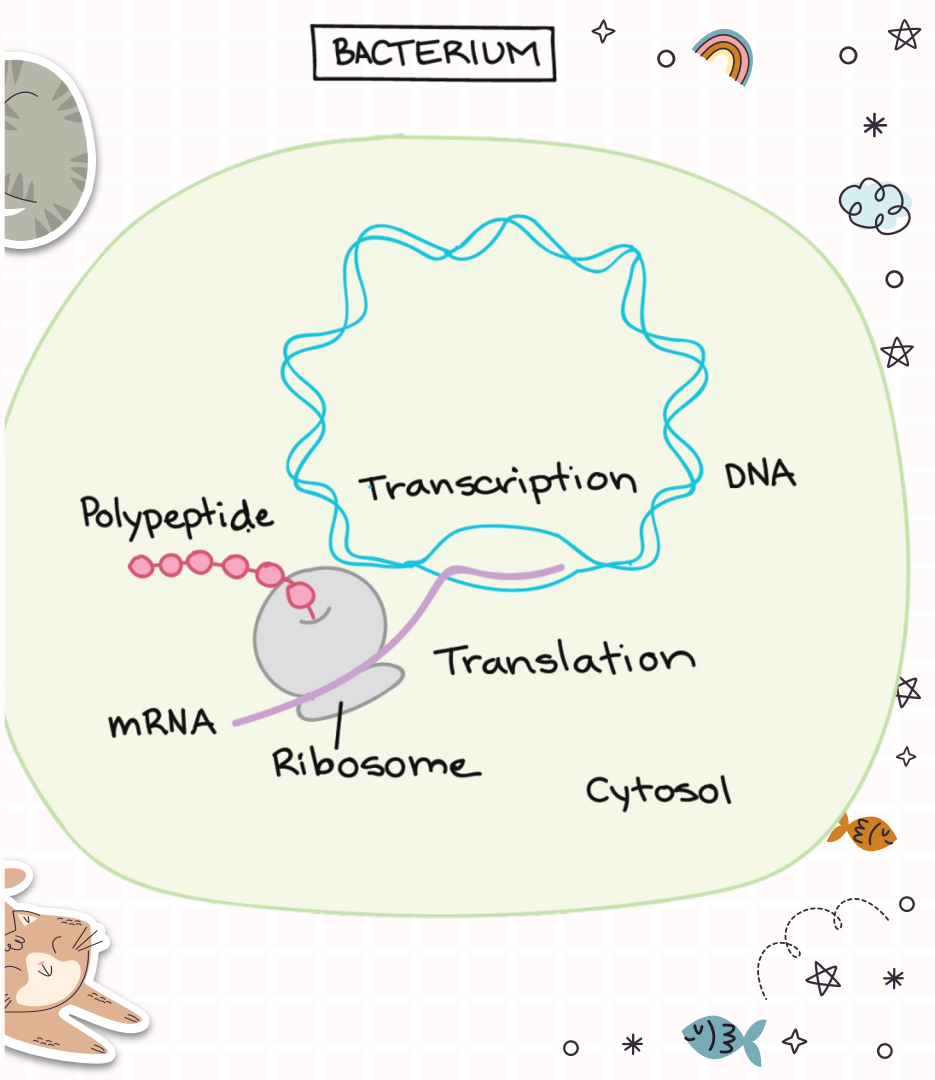
  UPDATE GATE: DECIDES HOW MUCH OF THE PAST INFORMATION NEEDS TO BE PASSED ALONG TO THE FUTURE.

LSTMs

- HANDLING THE SAME 2 PROBLEMS THAT WE FACED EARLIER WITH RNNS
- MAIN IDEA ??
- NOTE: IT USES SIGMOID ACTIVATION FUNCTION
- HOW DO WE AVOID THE ISSUES?
- ACHIEVED NEAR PERFECT ACCURACY ON THE TRAINING SET,
- DECREASED THE HIDDEN STATE SIZE AS A WAY TO GENERALIZE THE MODEL FURTHER.
- ATTEMPTED TO DECREASE THE DROPOUT PROBABILITY INSTEAD, WHICH SHOWED ONLY A SLIGHT IMPROVEMENT.



Journey to Proteomics: Decoding genes to protein families



Quick Recap:

Journey to Proteomics, decoding genes to protein families

DNA:

- INFORMATION TEMPLATE
- FUNCTIONAL POLYPEPTIDE CHAIN
- DUPLICATION
- PROMOTER REGION
- TRANSCRIPTION

RNA:

- RNA POLYMERASE
- UNWINDING
- ADDITION OF COMPLEMENTARY BASES ALONG WITH URACIL(T->U)
- REGULATORY RNA
- TRANSFER RNA
- RIBOSOME BINDING SITE
- PREFACE AND OUTRO(5' UTR 3'UTR)
- CREATING OF MATURE RNA

PROTEIN:

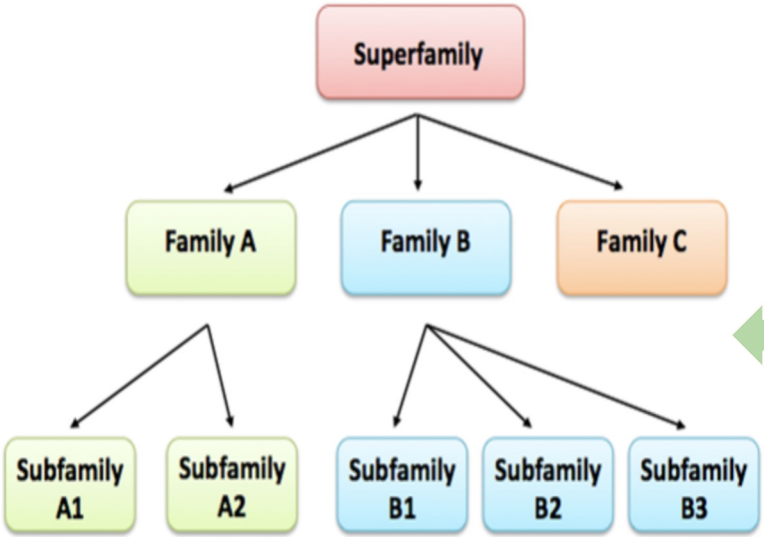
- TRANSLATION
- ELONGATION
- TERMINATION
- PROTEIN FOLDING
- POST TRANSLATIONAL MODIFICATION
- PROTEOMICS: BEYOND TRANSLATION

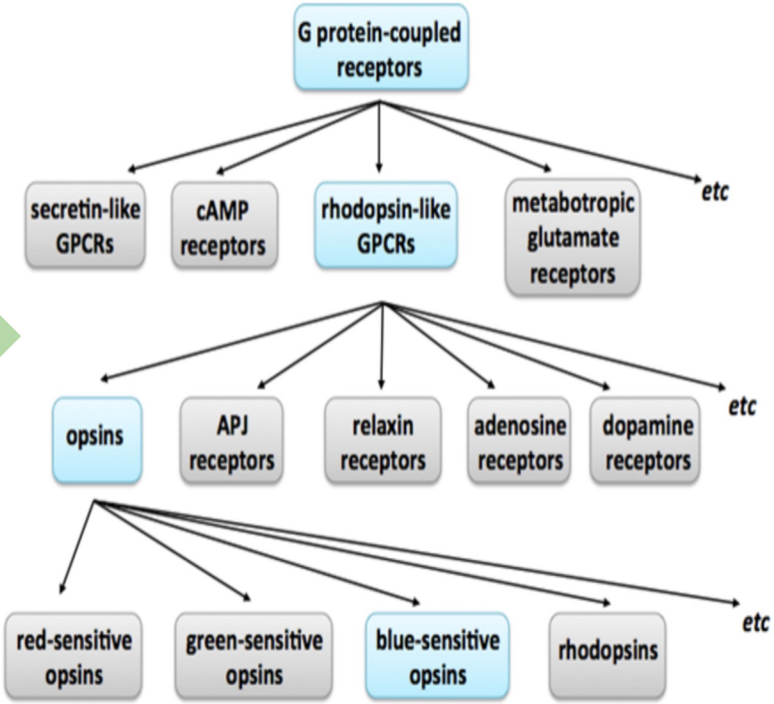
Few terms

1. PROMOTER REGION
2. RNA POLYMERASE
3. PRE/PRIMARY RNA
4. REGULATORY RNA
5. TRANSFER RNA
6. ANTICODON
7. RIBOSOME BINDING SITE
8. POLY A TAIL
9. 5’ UTR
10. 3’UTR
11. CHAPERONE PROTEINS

Family and domain of proteins

1. FAMILIES
2. STRUCTURE
3. EVOLUTION
4. CLASSIFICATION


Superfamily
Family A
Family B
Family C
Subfamily A1
Subfamily A2
Subfamily B1
Subfamily B2
Subfamily B3


G protein-coupled receptors
secretin-like GPCRs
cAMP receptors
rhodopsin-like GPCRs
metabotropic glutamate receptors
etc
opsins
APJ receptors
relaxin receptors
adenosine receptors
dopamine receptors
etc
red-sensitive opsins
green-sensitive opsins
blue-sensitive opsins
rhodopsins
etc

Dataset

Database: UniProt

ORGANISM: FELIS CATUS

1. ONLY SEQUENCES MORE THAN 200
2. DOWNLOADED THEM
3. ANNOTATED THEM WITH FAMILY NAME USING THE FAMILY FIELD(HTTPS://WWW.UNIPROT.ORG/HELP/QUERY-FIELDS) IN THE DATABASE

Total # of sequences	550,960
Sequence length	10-35000
Total # of families	10345
Total # families with > 200 sequences	589
# of sequences used for classification	317460

Table 1: Uniprot dataset of annotated protein sequences.

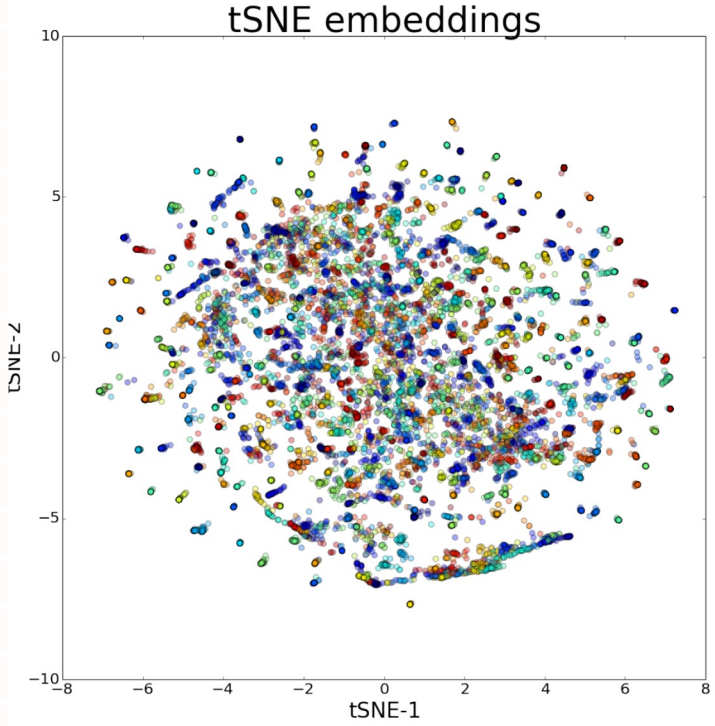
```
MAFSAEDVLKEYDRRRRMEALLLSLYYP...   Pox_VLTF3
MSIIGATRLQNDKSDTYSAGPCYAGGCS...   Pox_G9-A16
MQNPLPEVMSPEHDKRTTTPMSKEANKF...        US22
```

Awesome words

RESULTS:

Model	# hidden units	lr	l2reg	dropout	Val. F1	Test F1
SVM	-	-	-	-	-	0.87876
LSTM	100	0.01	0	0.85	0.926192	0.92515
LSTM	50	0.007	0	0.85	0.890175	0.888509
LSTM	100	0.01	0	0.70	0.925665	0.922225
biLSTM	100	0.007	0	0.9	0.928740	0.927899
GRU	**100**	**0.01**	**0**	**0.8**	**0.953141**	**0.948452**
CNN	384	0.001	0	0.5	0.9006	0.897853
CNN	384	0.001	0.0001	0.5	0.934	0.934

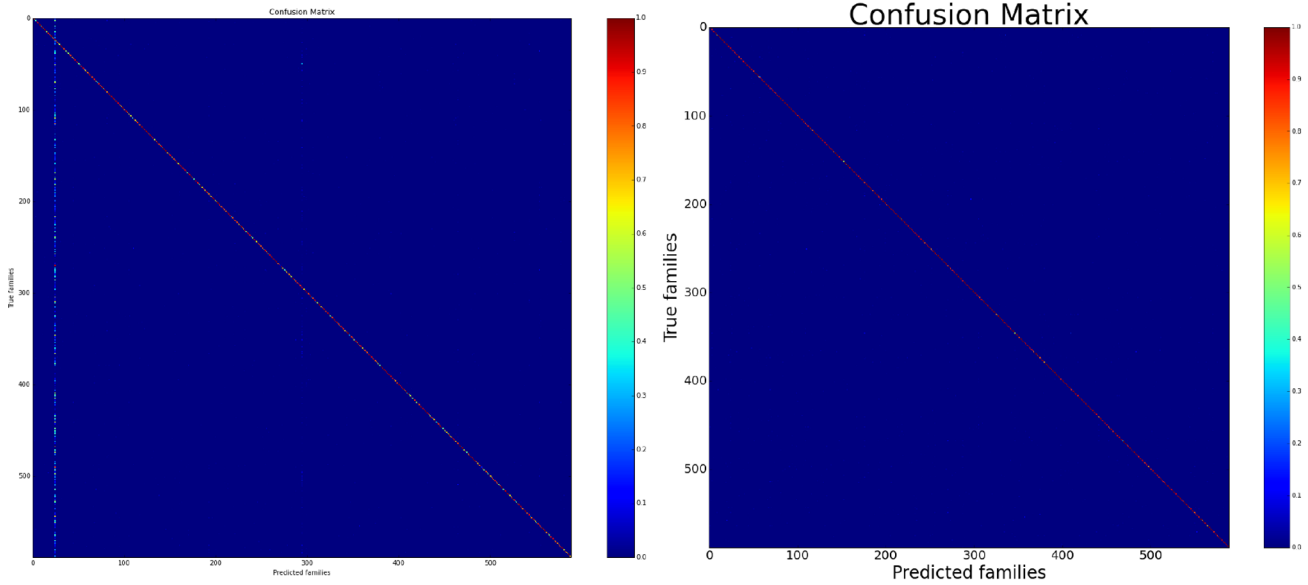
Table 2: Test F1 scores for different models and hyperparameters



(a) GRU tSNE embeddings (color coded)

WHAT IS TSNE EMBEDDINGS?

CONCLUSION:
SIMILAR PROTEIN EMBEDDINGS SHOULD HAVE BEEN GROUPED TOGETHER BUT WE DON'T SEE ANY PATTERN HERE



(a) SVM confusion matrix

(b) GRU confusion matrix

Figure 6: Classification Performance SVM vs GRU

FUTURE RESEARCH:

1. GRUS PERFORMED 7% HIGHER THAN SVMS, WHY?
   SVMS TRAINED MANY SINGLE CLASS CLASSIFIER AND THE MULTICLASS PROTEINS BEING MORE COMPLEX PERFORMED WELL WITH A NN ARCHITECTURE
2. EXPERIMENTING WITH DIFFERENT LENGTHS OF N-GRAM USED TO TRAIN THE GLOVE EMBEDDINGS CAN BE INTERESTING
3. THE REPRESENTATIONS DID NOT TAKE THE BIOLOGICAL/SEMANTIC MEANING INTO ACCOUNT, FURTHER INVESTIGATION ON THAT ASPECT IS REQUIRED

Thanks!

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon** and infographics & images by **Freepik**