Exploring WORD2VEC models for Capturing the Similarity of Codon Embeddings

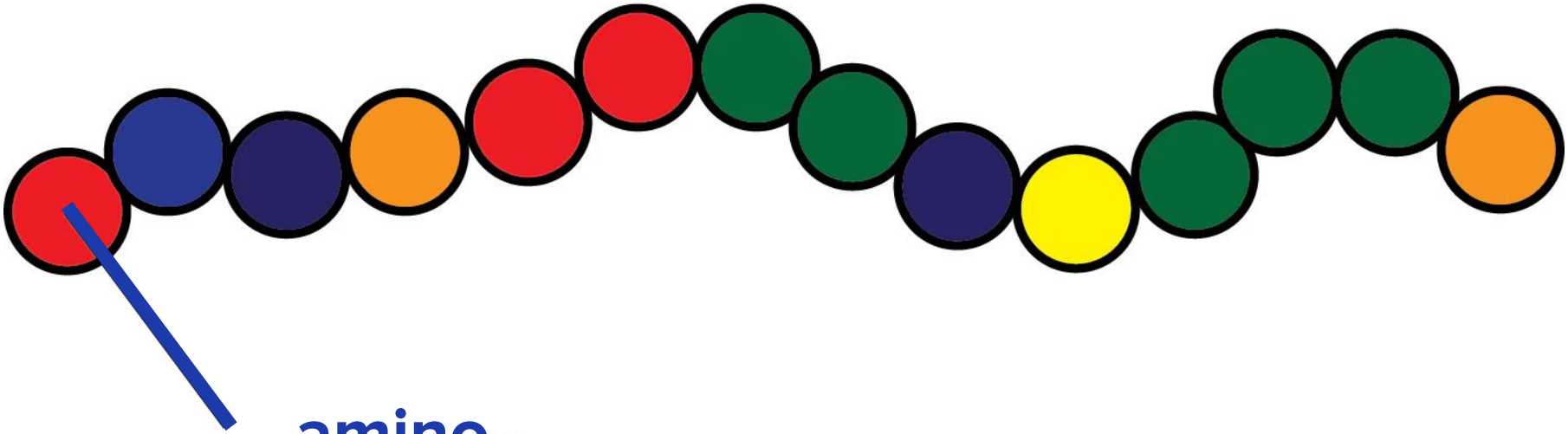
Anđa Denić, Jelena Pejić, Aleksandar Trokicić

University of Niš
Faculty of Sciences and Mathematics

Biology introduction

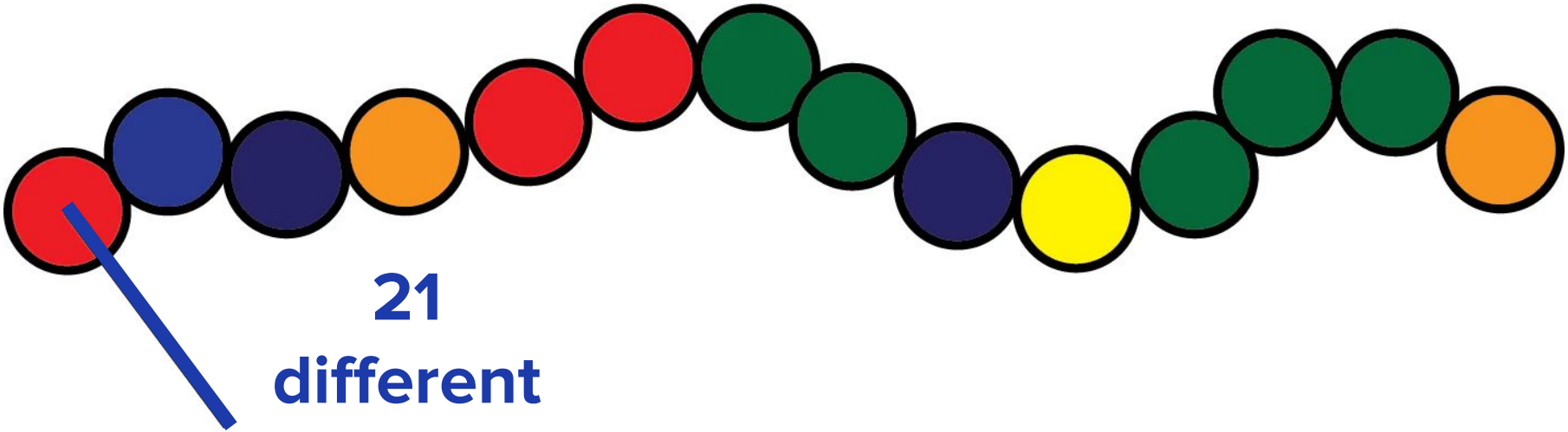
proteins

Protein



amino acid

Protein



21 different amino acids

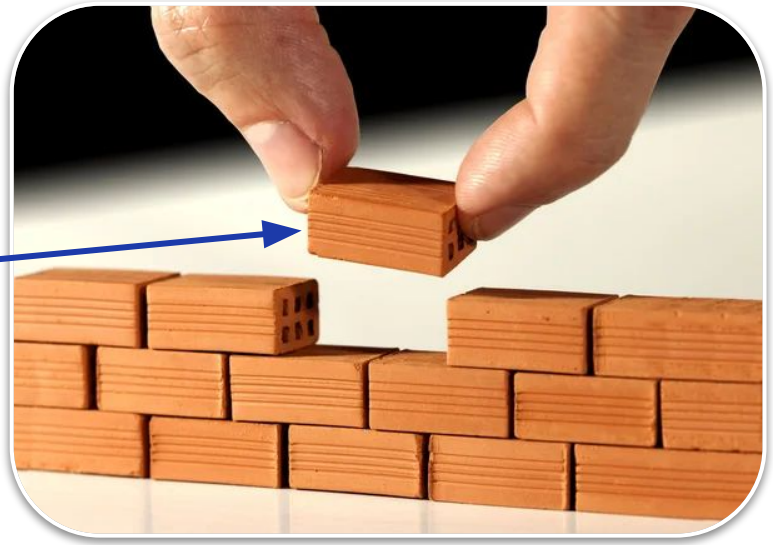
Biology introduction

TGTTCGAGCTCTTCAAACCGG

Gene

code

protein



Biology introduction

codon

Biology introduction

One codon code one amino acid.

Biology introduction

4 x 4 x 4 = 64 different codons
that code
21 amino acids

Codon table

Second letter

First letter	U		C		A		G		Third letter
U	UUU	phenylalanine (Phe)	UCU	serine (Ser)	UAU	tyrosine (Tyr)	UGU	cysteine (Cys)	U
	UUC		UCC		UAC		UGC		C
	UUA	leucine (Leu)	UCA		UAA	STOP	UGA	STOP	A
	UUG		UCG		UAG		UGG	tryptophan (Trp)	G
C	CUU	leucine (Leu)	CCU	proline (Pro)	CAU	histidine (His)	CGU	arginine (Arg)	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	glutamine (Gln)	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	isoleucine (Ile)	ACU	threonine (Thr)	AAU	asparagine (Asn)	AGU	serine (Ser)	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	lysine (Lys)	AGA	arginine (Arg)	A
	AUG	methionine (Met)	ACG		AAG		AGG		G
G	GUU	valine (Val)	GCU	alanine (Ala)	GAU	aspartic acid (Asp)	GGU	glycine (Gly)	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	glutamic acid (Glu)	GGA		A
	GUG		GCG		GAG		GGG		G

Biology introduction

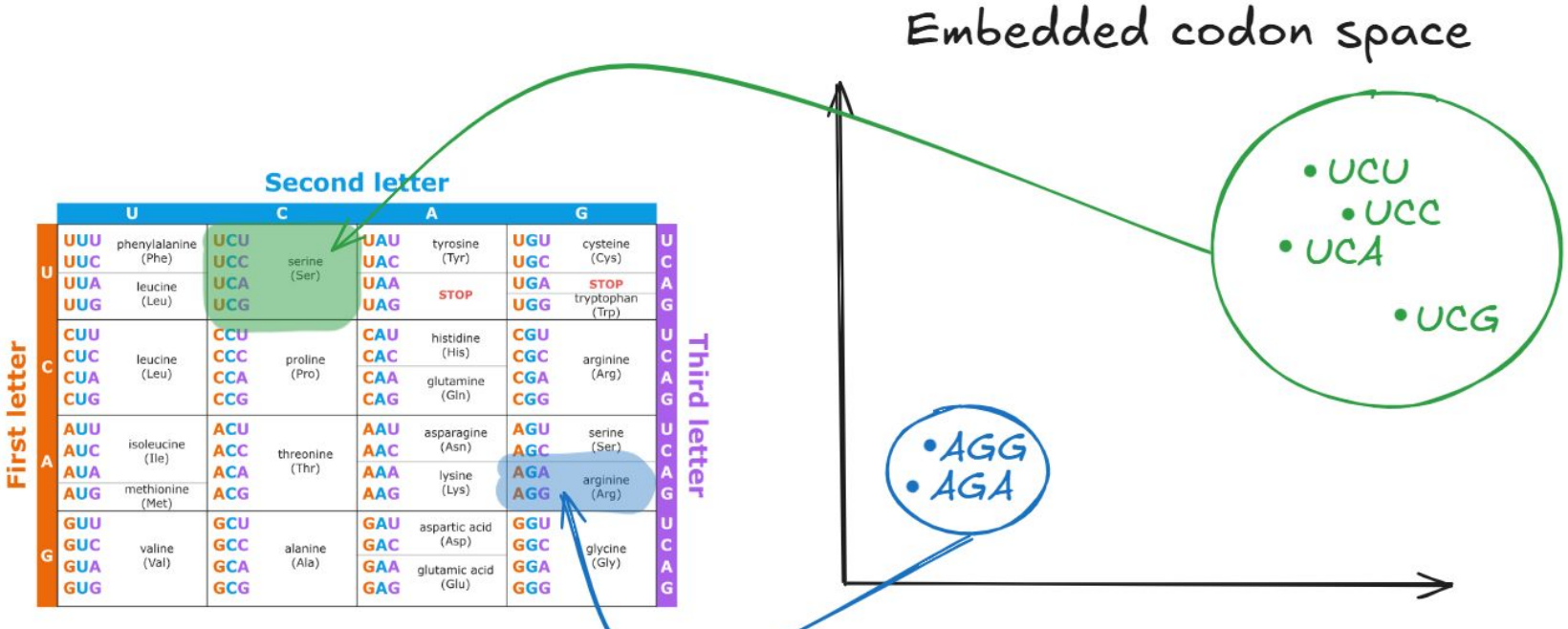
In nature
64 codons
are clustered in
21 functional groups.

Project goal

Find functional codon embedding space.

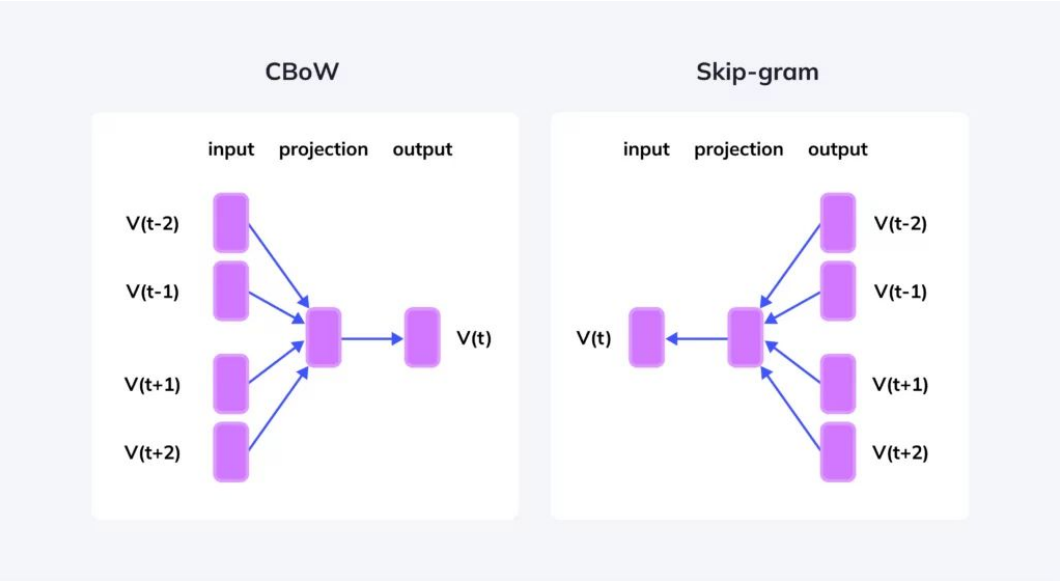
Project goal

Find functional codon embedding space.



Word2vec

Popular NLP model that learns a vector representation of words using its context.



Context in NLP

Context of a **word ‘day’**:

Today is a **beautiful** **and** **sunny** **day** **in** **Kragujevac**, **the** **fourth** largest city in Serbia.

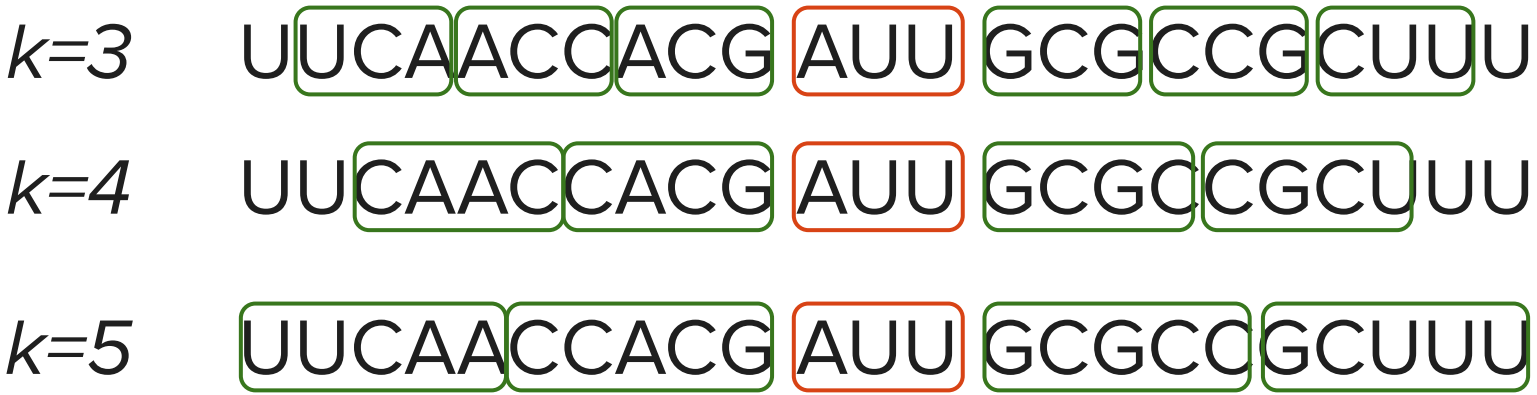
Context in bioinformatics?

Context of a **codon 'AUU'**?

UUCAACCACG AUU GCGCCGCUUU

Context in bioinformatics?

Context of a **codon 'AUU'**?



Context in bioinformatics?

Context of a **codon 'AUU'**?

m=1 UUCAACCACG AUU GCGCCGCUUU

m=3 UUCAACCACG AUU GCGCCGCUUU

Context in bioinformatics?

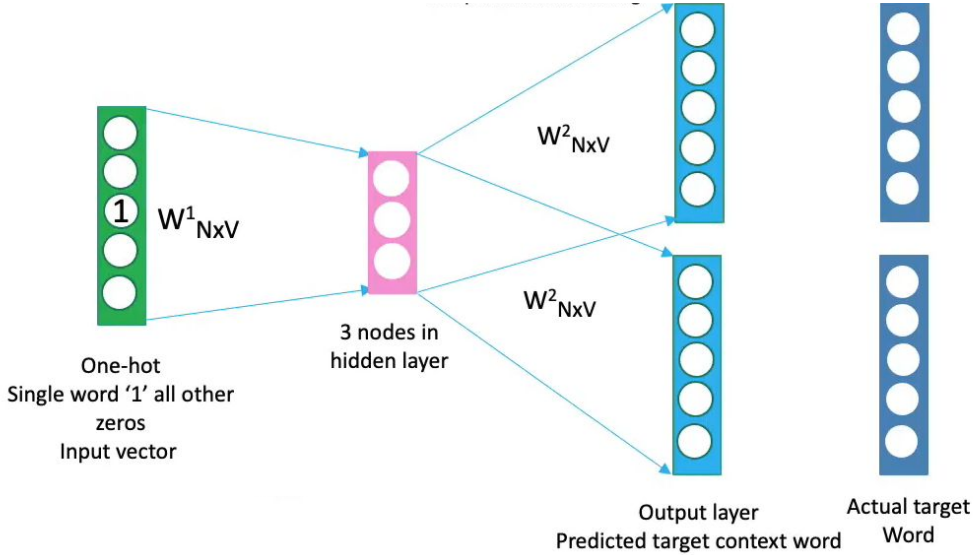
Context of a **codon 'AUU'**?

What is the best context for a codon?

We trained 60 different skip gram word2vec models

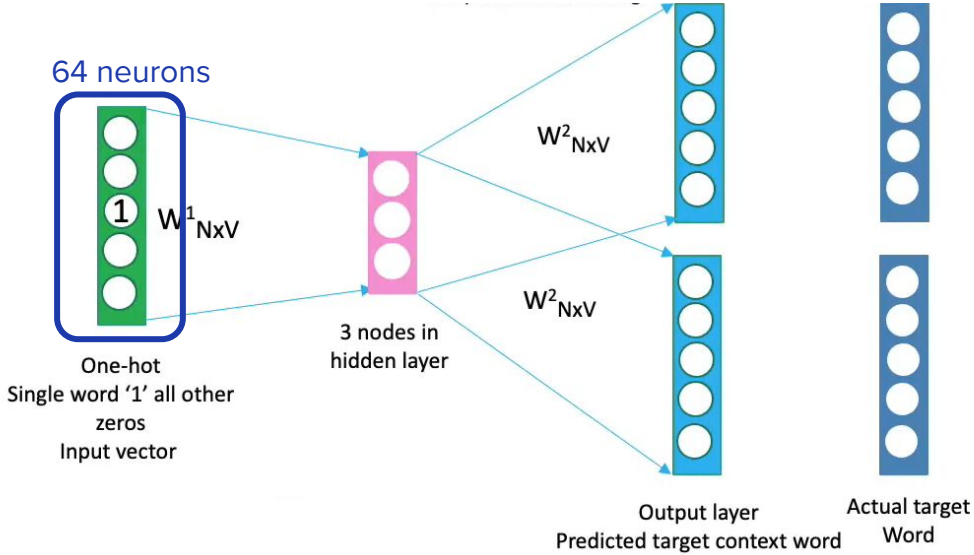
Skip gram word2vec architecture

Skip gram is just a feed forward network with one hidden layer.



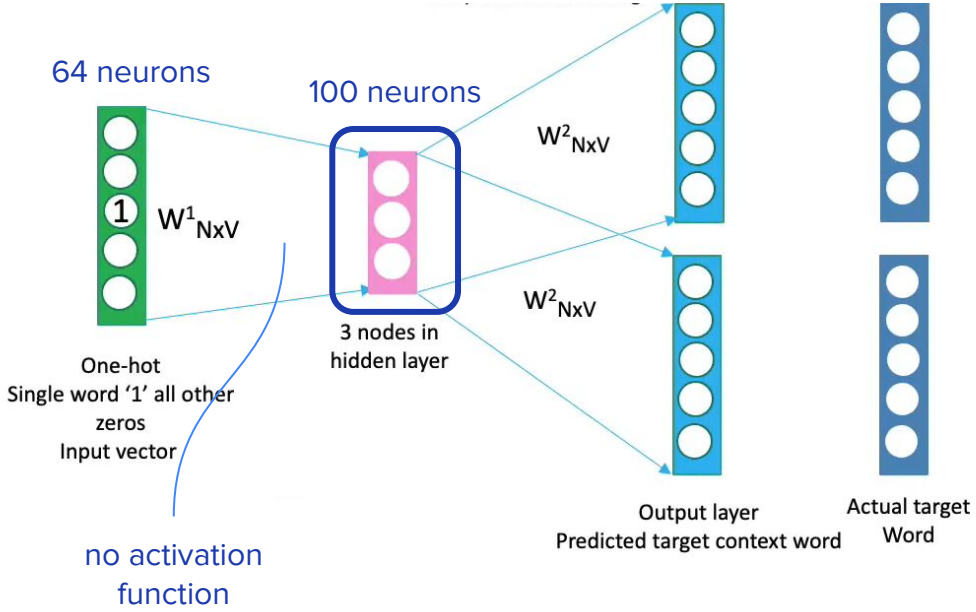
Skip gram word2vec architecture

Skip gram is just a feed forward network with one hidden layer.



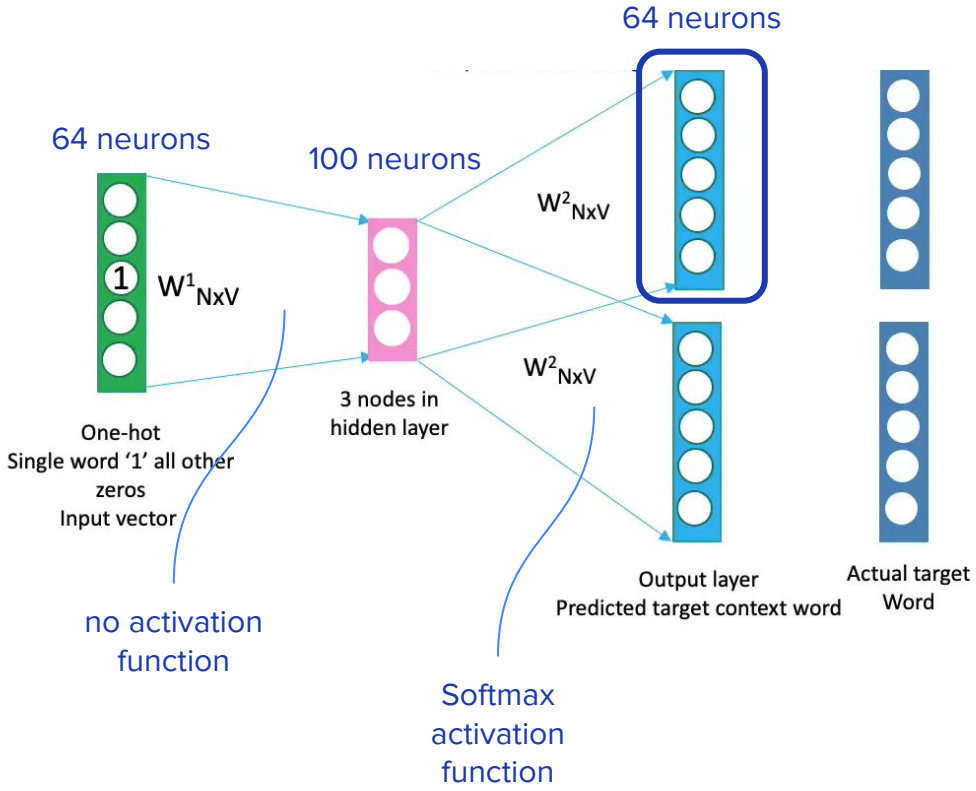
Skip gram word2vec architecture

Skip gram is just a feed forward network with one hidden layer.



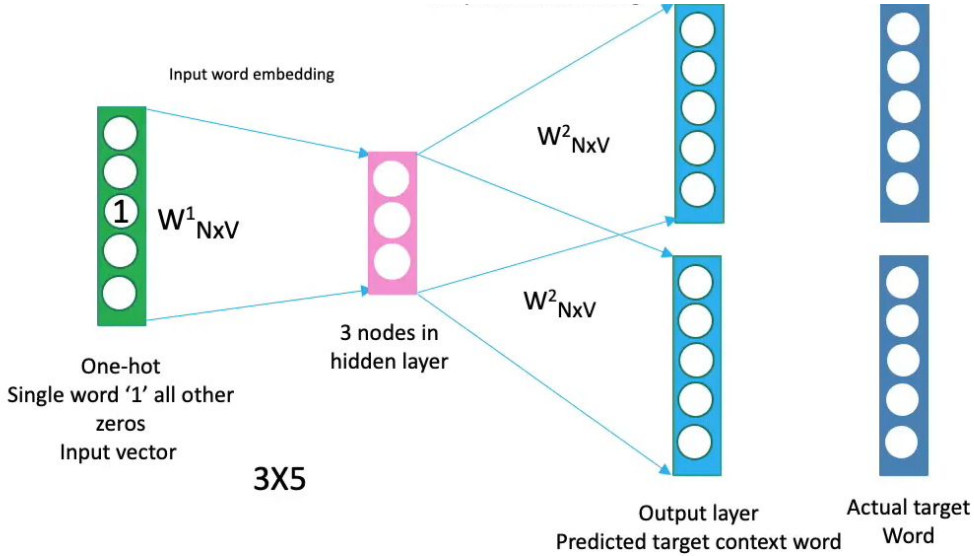
Skip gram word2vec architecture

Skip gram is just a feed forward network with one hidden layer.



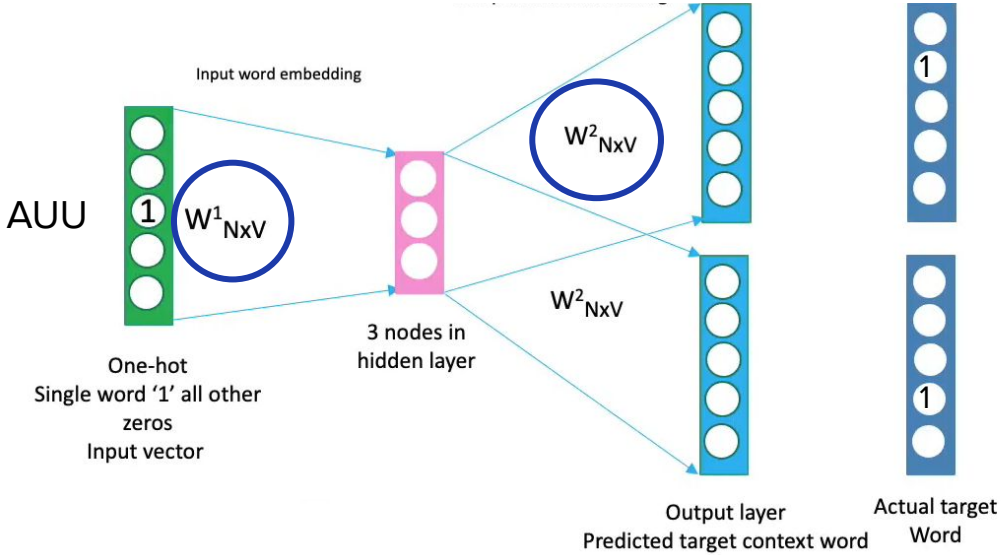
Training

Model learns to predict context codons.

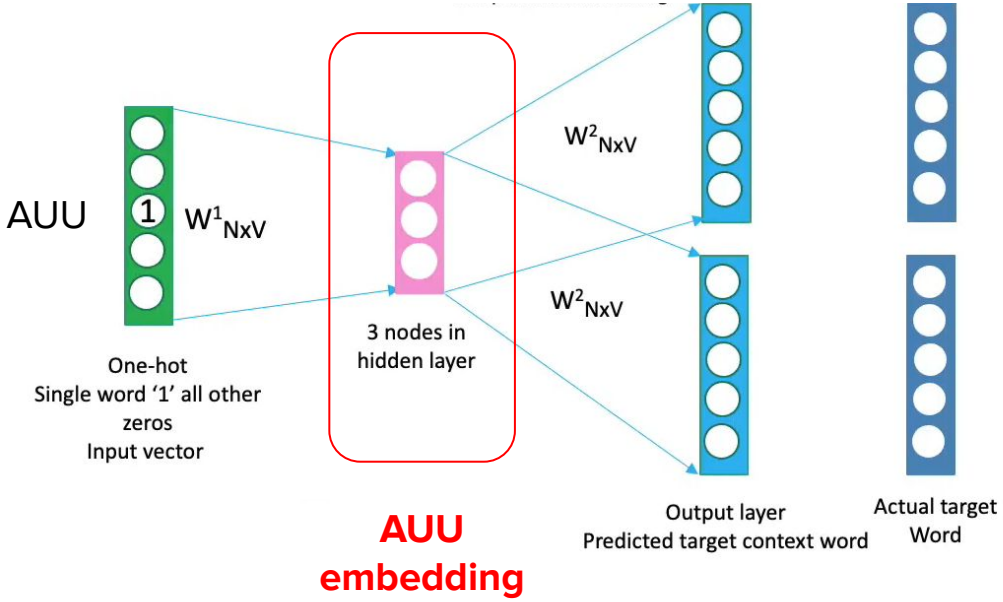


Training

We train the network to learn weight matrices W1 and W2.



Inference
Codon Embedding



Evaluation

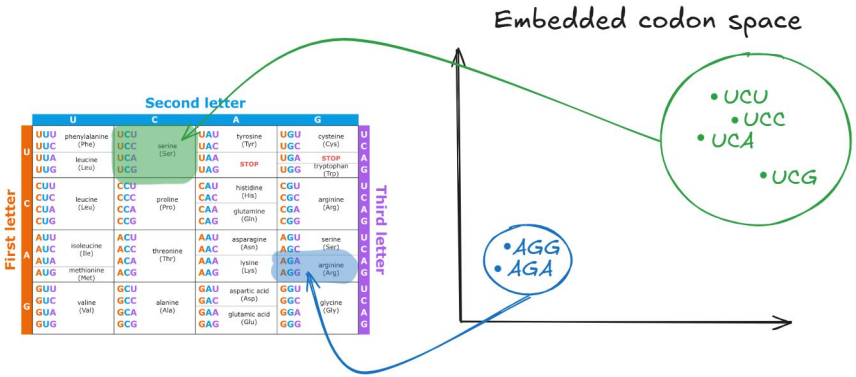
Cosine similarity function measures similarity between codon embeddings.

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Evaluation

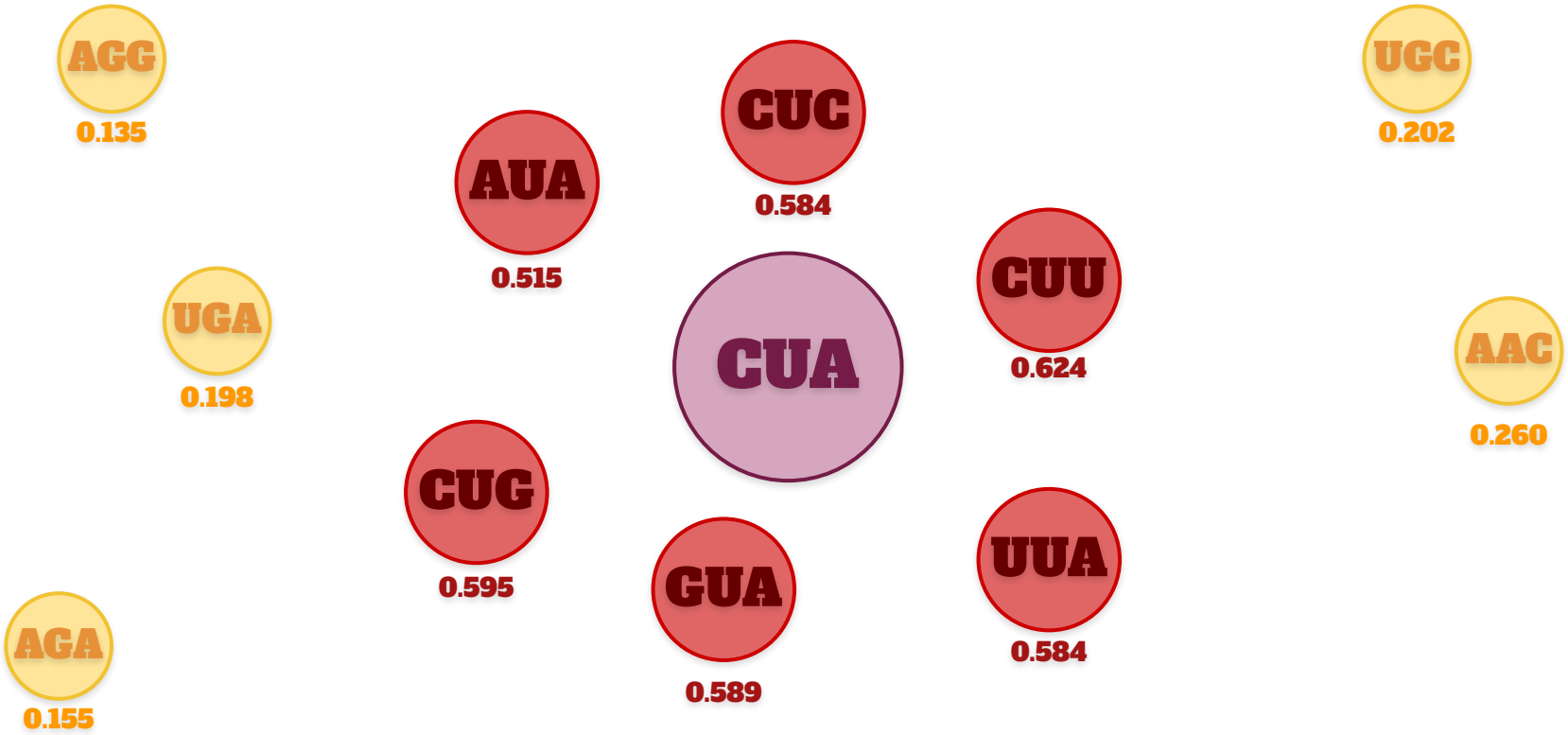
The goal is for codons from the same functional group to have high cosine similarity.

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

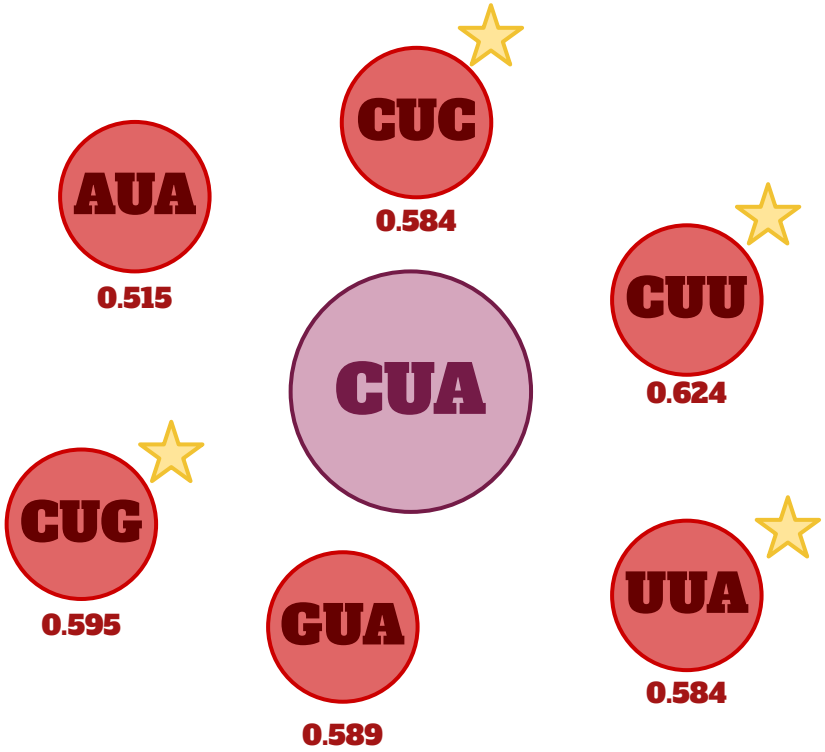


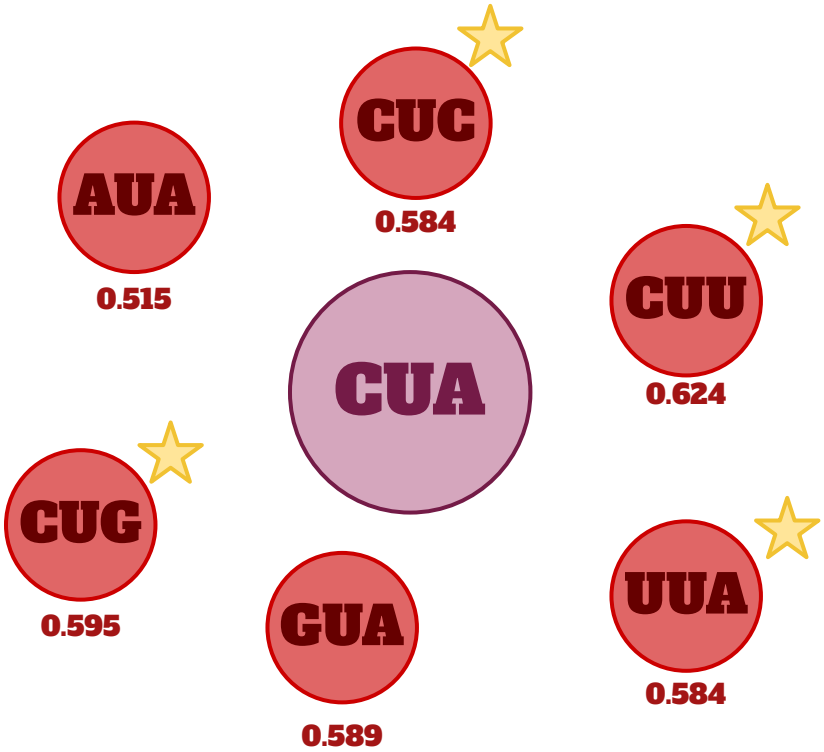
How to measure the score of the trained model?

Score of the model on a codon



$$\text{Recall} = \frac{\#\text{ codons from the same group in top 6 neighbors}}{\text{total number of codons in the functional group}}$$


AUA
0.515
CUC
0.584
CUU
0.624
CUA
CUG
0.595
GUA
0.589
UUA
0.584



$$\text{Recall ('CUA')} = \frac{4}{5} = 80\%$$

Score of a model = average recall over all codons

Results

- 3-mers as neighbors
- overlapping windows
- about 5 neighbors

perform the best.

	Overlapping					Non-overlapping				
	m=3	m=5	m=10	m=20	m=40	m=3	m=5	m=10	m=20	m=40
Vibrio cholerae	84.38%	84.38%	84.38%	83.28%	80.94%	45.52%	37.66%	40.52%	34.48%	35.37%
Escherichia coli	84.38%	84.38%	82.81%	81.66%	79.27%	41.46%	42.66%	42.97%	39.27%	35.78%

Conclusion

AI (word2vec) can learn the code written in genes! 🎉

Conclusion

AI (word2vec) can learn the code written in genes! 🎉

(At least, it can vectorise codons in such a way that codon embeddings are ‘close’ if their biochemical function is similar.)

Future work

🧬 Train models on different organism's genome
🧬 The best context may vary between organisms

Potential applications

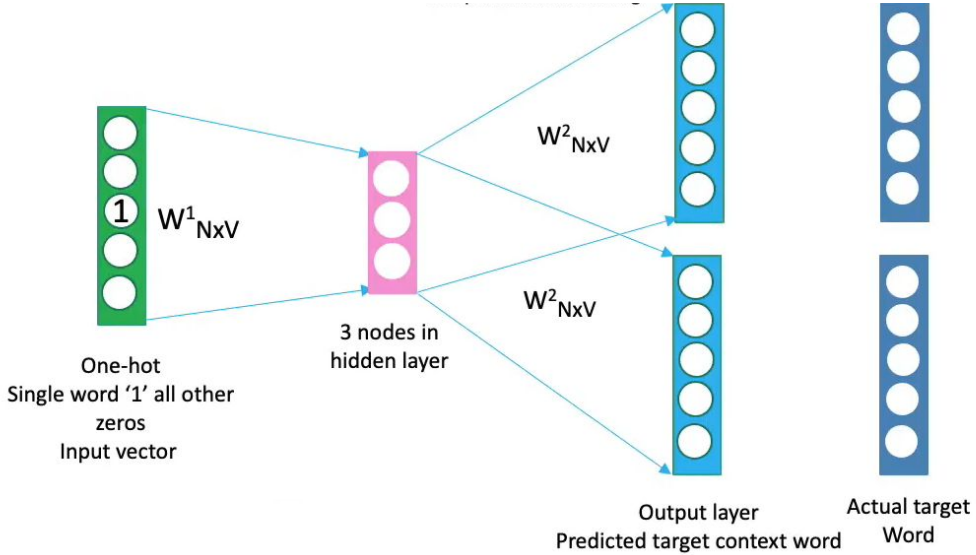
🧬 Gene function prediction

🧬 Detection of start and end of a gene

Thank you

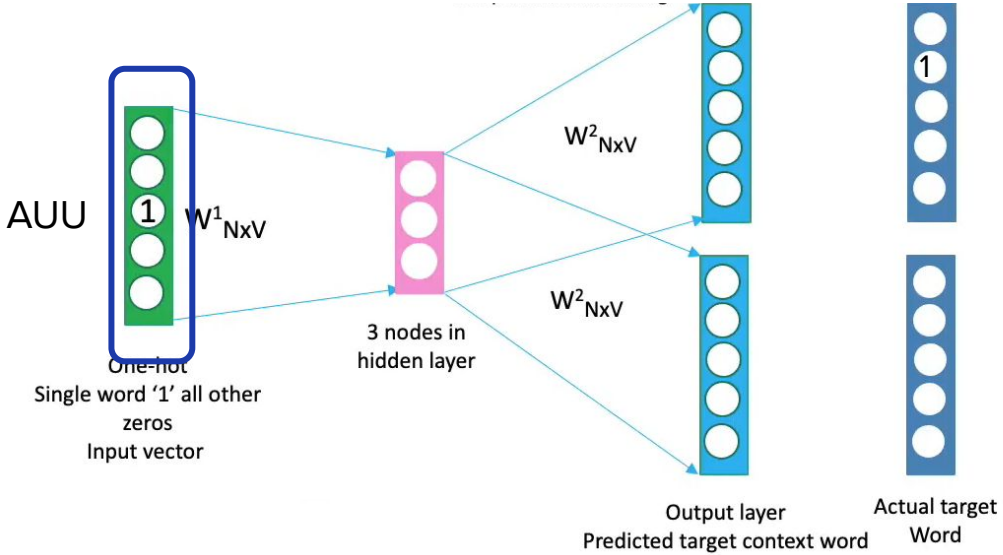
Training example

UUCAACCACG AUU GCGCCGCUUU

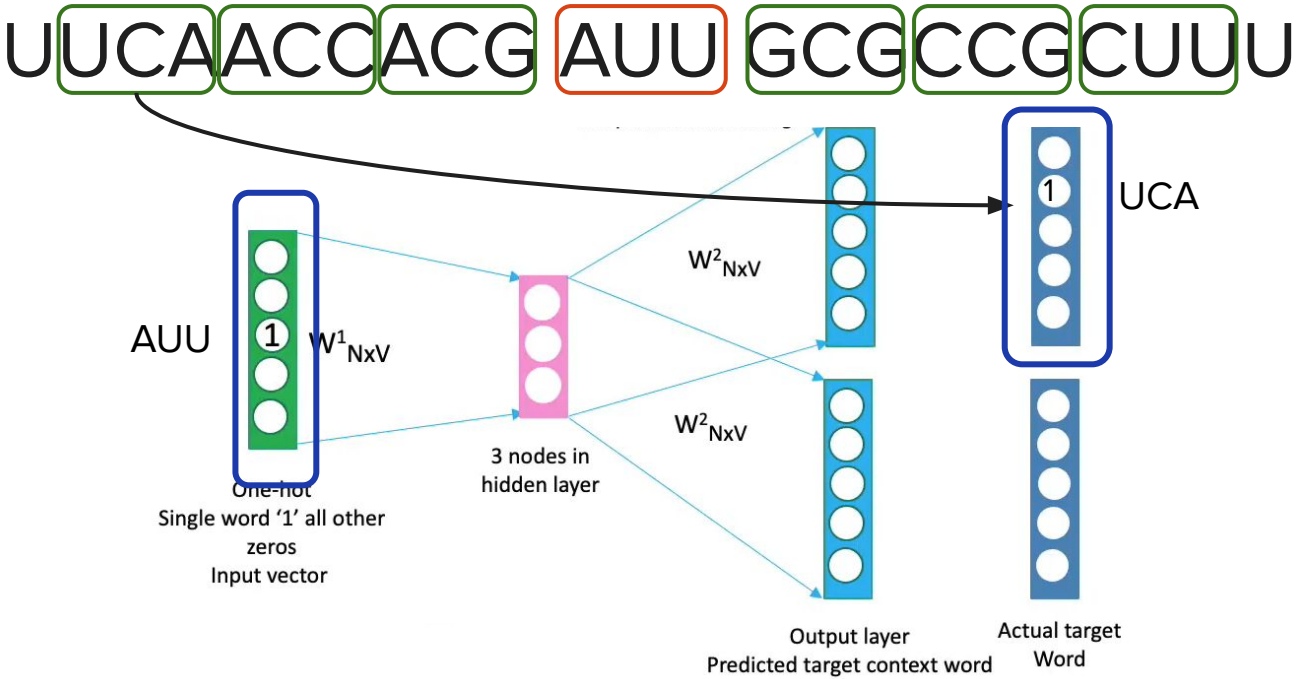


Training example

UUCAACCACG AUU GCGCCGCUUU



Training example



Training example