Natural Language Processing Techniques for Identifying Bacteriocins

Master thesis, Lasse Buur Rasmussen

Supervised by Asker Daniel Brejnrod and Mani Arumugam

UNIVERSITY OF COPENHAGEN

Background

- Antibiotics
  - Antibiotic-resistant bacteria
  - Long-lasting alterations of gut microbiome

Background

- Antibiotics
  - Antibiotic-resistant bacteria
  - Long-lasting alterations of gut microbiome
- Bacteriocins
  - Small antimicrobial peptides ribosomally synthesized by bacteria

Background

- Antibiotics
  - Antibiotic-resistant bacteria
  - Long-lasting alterations of gut microbiome
- Bacteriocins
  - Small antimicrobial peptides ribosomally synthesized by bacteria
- Hamid et al. neural models & Word2Vec embedding
  - High performance - challenge traditional bioinformatic classification methods.

Background

- Antibiotics
  - Antibiotic-resistant bacteria
  - Long-lasting alterations of gut microbiome
- Bacteriocins
  - Small antimicrobial peptides ribosomally synthesized by bacteria
- Hamid et al. neural models & Word2Vec embedding
  - High performance - challenge traditional bioinformatic classification methods.
- ELMo advantages
  - Word2Vec underlying assumption – words single context.
  - Multiple representations each word.

Main findings

- Increased accuracy with ELMo embedding.
  - Hamid et al. test accuracy 86.0%.
  - Our classifier test accuracy 94.8%.
- Found 40 putative bacteriocins.

Contents

Contents

Encoding Biological Sequences

- Why?
  - Machines like numbers
  - Quantative rather than categorical

Encodings used

Encoding	Equation	Dimensions
Atchley clust.	$100C$	100D
Atchley	$97KM \cdot 15AF$	1455D
One-hot	$99AA \cdot 22LE$	2178D
Reduced alphabet	$99AA \cdot 11LE$	1089D
Word2Vec clust.	$100C$	100D
Word2Vec	$97KM \cdot 200W2V$	19400D
ELMo summed	$99AA \cdot 1024ELS$	101376D
ELMo	$99AA \cdot 3072EL$	304128D

Table 1: Dimensionality of the different encodings with a sequence set of maximum length 99. Legend: C = cluster, D = dimensions, KM = k-mer, AF = Atchley factors, AA = Amino acids, LE = Letters, $W2V$ = Word2Vec dimensions, ELS = Summed ELMo dimensions, EL = ELMo dimensions.

- Sample dimensionality for 99 amino acids long sequence

Encodings used

Encoding	Equation	Dimensions
Atchley clust.	$100C$	100D
Atchley	$97KM \cdot 15AF$	1455D
One-hot	$99AA \cdot 22LE$	2178D
Reduced alphabet	$99AA \cdot 11LE$	1089D
Word2Vec clust.	$100C$	100D
Word2Vec	$97KM \cdot 200W2V$	19400D
ELMo summed	$99AA \cdot 1024ELS$	101376D
ELMo	$99AA \cdot 3072EL$	304128D

Table 1: Dimensionality of the different encodings with a sequence set of maximum length 99. Legend: C = cluster, D = dimensions, KM = k-mer, AF = Atchley factors, AA = Amino acids, LE = Letters, $W2V$ = Word2Vec dimensions, ELS = Summed ELMo dimensions, EL = ELMo dimensions.

- Sample dimensionality for 99 amino acids long sequence

Sequences with different lengths

```
MNTKMME
KYYGNGVSCNKKGCSVD
MSRYTGPSWKQSRRLGLSLTGTGKEL
MANHSSAKKVVRQTVKRTLIN
GSRYLCTPGSCW
```

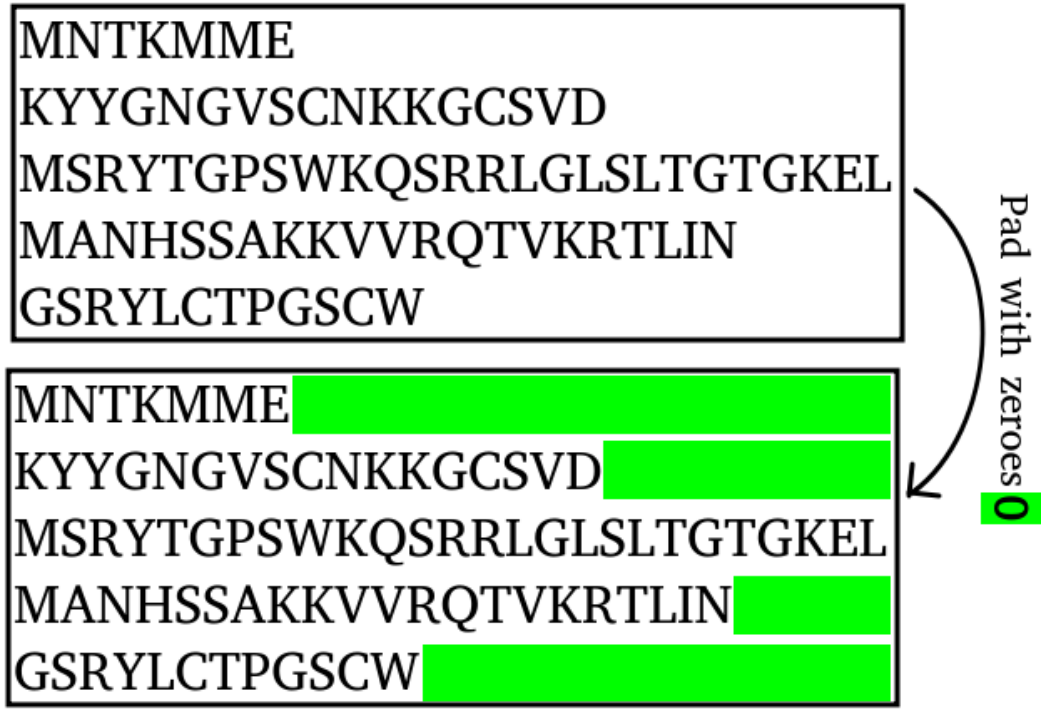
Encodings used

Encoding	Equation	Dimensions
Atchley clust.	$100C$	100D
Atchley	$97KM \cdot 15AF$	1455D
One-hot	$99AA \cdot 22LE$	2178D
Reduced alphabet	$99AA \cdot 11LE$	1089D
Word2Vec clust.	$100C$	100D
Word2Vec	$97KM \cdot 200W2V$	19400D
ELMo summed	$99AA \cdot 1024ELS$	101376D
ELMo	$99AA \cdot 3072EL$	304128D

Table 1: Dimensionality of the different encodings with a sequence set of maximum length 99. Legend: C = cluster, D = dimensions, KM = k-mer, AF = Atchley factors, AA = Amino acids, LE = Letters, $W2V$ = Word2Vec dimensions, ELS = Summed ELMo dimensions, EL = ELMo dimensions.

- Sample dimensionality for 99 amino acids long sequence

Sequences with different lengths



Encodings used

Encoding	Equation	Dimensions
Atchley clust.	$100C$	100D
Atchley	$97KM \cdot 15AF$	1455D
One-hot	$99AA \cdot 22LE$	2178D
Reduced alphabet	$99AA \cdot 11LE$	1089D
Word2Vec clust.	$100C$	100D
Word2Vec	$97KM \cdot 200W2V$	19400D
ELMo summed	$99AA \cdot 1024ELS$	101376D
ELMo	$99AA \cdot 3072EL$	304128D

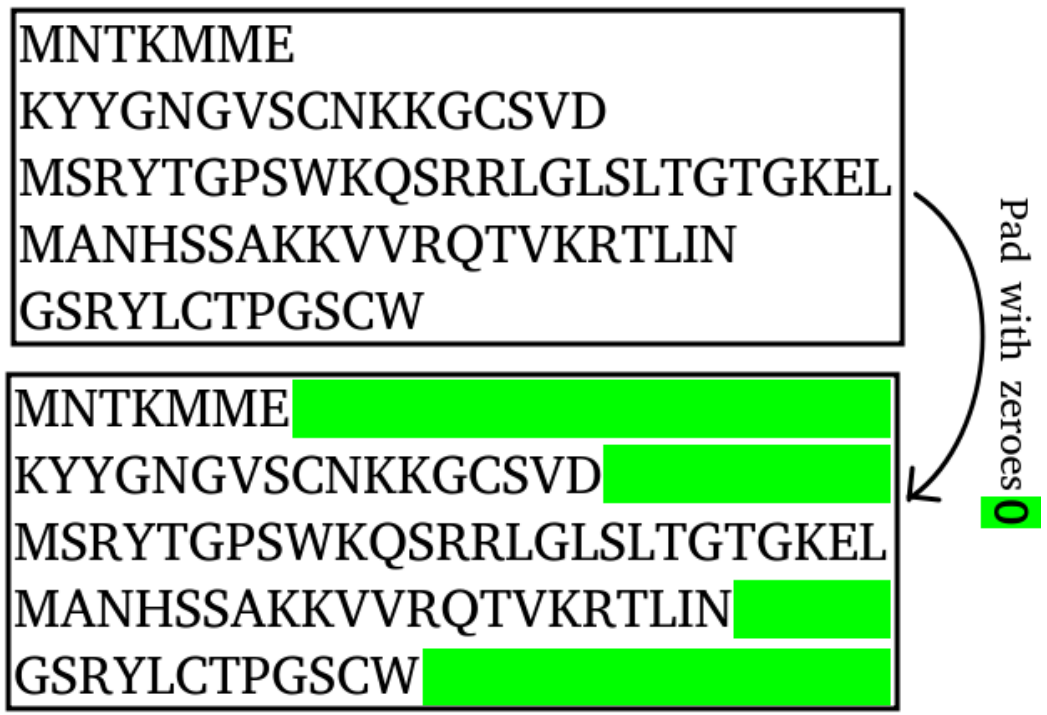
Table 1: Dimensionality of the different encodings with a sequence set of maximum length 99. Legend: C = cluster, D = dimensions, KM = k-mer, AF = Atchley factors, AA = Amino acids, LE = Letters, $W2V$ = Word2Vec dimensions, ELS = Summed ELMo dimensions, EL = ELMo dimensions.

Sequences containing X

```
VAPFPEQFLX
ISLEICXIFHDN
```

- Sample dimensionality for 99 amino acids long sequence

Sequences with different lengths

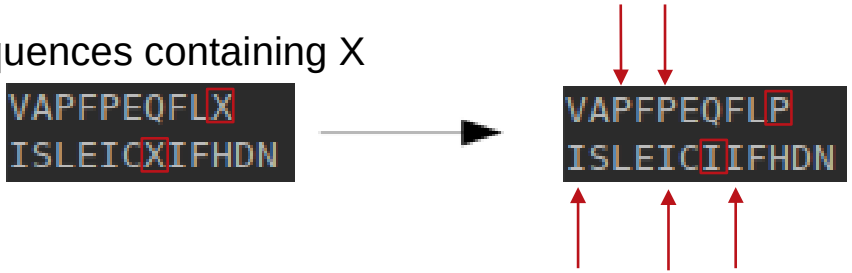


Encodings used

Encoding	Equation	Dimensions
Atchley clust.	$100C$	100D
Atchley	$97KM \cdot 15AF$	1455D
One-hot	$99AA \cdot 22LE$	2178D
Reduced alphabet	$99AA \cdot 11LE$	1089D
Word2Vec clust.	$100C$	100D
Word2Vec	$97KM \cdot 200W2V$	19400D
ELMo summed	$99AA \cdot 1024ELS$	101376D
ELMo	$99AA \cdot 3072EL$	304128D

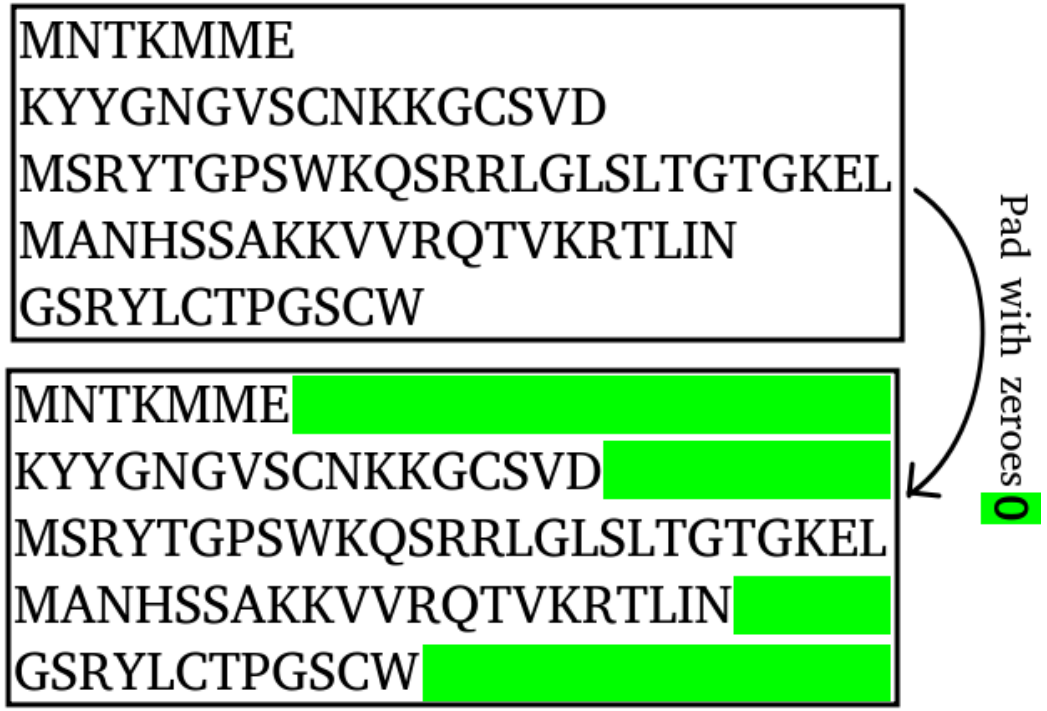
Table 1: Dimensionality of the different encodings with a sequence set of maximum length 99. Legend: C = cluster, D = dimensions, KM = k-mer, AF = Atchley factors, AA = Amino acids, LE = Letters, $W2V$ = Word2Vec dimensions, ELS = Summed ELMo dimensions, EL = ELMo dimensions.

Sequences containing X



- Sample dimensionality for 99 amino acids long sequence

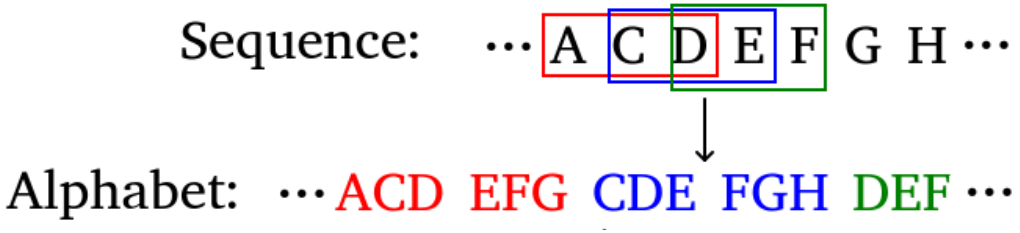
Sequences with different lengths



Amino Acids to K-Mers

Encoding	Equation	Dimensions
Atchley clust.	$100C$	100D
Atchley	$97KM \cdot 15AF$	1455D
One-hot	$99AA \cdot 22LE$	2178D
Reduced alphabet	$99AA \cdot 11LE$	1089D
Word2Vec clust.	$100C$	100D
Word2Vec	$97KM \cdot 200W2V$	19400D
ELMo summed	$99AA \cdot 1024ELS$	101376D
ELMo	$99AA \cdot 3072EL$	304128D

Table 1: Dimensionality of the different encodings with a sequence set of maximum length 99. Legend: C = cluster, D = dimensions, KM = k-mer, AF = Atchley factors, AA = Amino acids, LE = Letters, $W2V$ = Word2Vec dimensions, ELS = Summed ELMo dimensions, EL = ELMo dimensions.



One-Hot Encoding

Encoding	Equation	Dimensions
Atchley clust.	$100C$	100D
Atchley	$97KM \cdot 15AF$	1455D
One-hot	$99AA \cdot 22LE$	2178D
Reduced alphabet	$99AA \cdot 11LE$	1089D
Word2Vec clust.	$100C$	100D
Word2Vec	$97KM \cdot 200W2V$	19400D
ELMo summed	$99AA \cdot 1024ELS$	101376D
ELMo	$99AA \cdot 3072EL$	304128D

Table 1: Dimensionality of the different encodings with a sequence set of maximum length 99. Legend: C = cluster, D = dimensions, KM = k-mer, AF = Atchley factors, AA = Amino acids, LE = Letters, $W2V$ = Word2Vec dimensions, ELS = Summed ELMo dimensions, EL = ELMo dimensions.

One-Hot Encoding

Encoding	Equation	Dimensions
Atchley clust.	$100C$	100D
Atchley	$97KM \cdot 15AF$	1455D
One-hot	$99AA \cdot 22LE$	2178D
Reduced alphabet	$99AA \cdot 11LE$	1089D
Word2Vec clust.	$100C$	100D
Word2Vec	$97KM \cdot 200W2V$	19400D
ELMo summed	$99AA \cdot 1024ELS$	101376D
ELMo	$99AA \cdot 3072EL$	304128D

Table 1: Dimensionality of the different encodings with a sequence set of maximum length 99. Legend: C = cluster, D = dimensions, KM = k-mer, AF = Atchley factors, AA = Amino acids, LE = Letters, $W2V$ = Word2Vec dimensions, ELS = Summed ELMo dimensions, EL = ELMo dimensions.

$$A = \begin{array}{c} \\ G \\ H \\ \vdots \\ V \\ M \end{array} \begin{array}{c} \begin{array}{cccccccccccccccccccc} A & R & N & D & C & E & Q & G & H & I & L & K & M & F & P & S & T & W & Y & V \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

Encodings used

Encoding	Equation	Dimensions
Atchley clust.	$100C$	100D
Atchley	$97KM \cdot 15AF$	1455D
One-hot	$99AA \cdot 22LE$	2178D
Reduced alphabet	$99AA \cdot 11LE$	1089D
Word2Vec clust.	$100C$	100D
Word2Vec	$97KM \cdot 200W2V$	19400D
ELMo summed	$99AA \cdot 1024ELS$	101376D
ELMo	$99AA \cdot 3072EL$	304128D

Table 1: Dimensionality of the different encodings with a sequence set of maximum length 99. Legend: C = cluster, D = dimensions, KM = k-mer, AF = Atchley factors, AA = Amino acids, LE = Letters, $W2V$ = Word2Vec dimensions, ELS = Summed ELMo dimensions, EL = ELMo dimensions.

Encodings used

Encoding	Equation	Dimensions
Atchley clust.	$100C$	100D
Atchley	$97KM \cdot 15AF$	1455D
One-hot	$99AA \cdot 22LE$	2178D
Reduced alphabet	$99AA \cdot 11LE$	1089D
Word2Vec clust.	$100C$	100D
Word2Vec	$97KM \cdot 200W2V$	19400D
ELMo summed	$99AA \cdot 1024ELS$	101376D
ELMo	$99AA \cdot 3072EL$	304128D

Table 1: Dimensionality of the different encodings with a sequence set of maximum length 99. Legend: C = cluster, D = dimensions, KM = k-mer, AF = Atchley factors, AA = Amino acids, LE = Letters, $W2V$ = Word2Vec dimensions, ELS = Summed ELMo dimensions, EL = ELMo dimensions.

- Polarity
- Secondary structure
- Molecular volume
- Codon diversity
- Electrostatic charge

Atchley Factors Encoding

Example with 2 sequences, k-mers with k=3 and 3 clusters

1 Replace Xs with most frequent AA in window of size 14

```
VAPFPEQFLX
ISLEICXIFHDN
```

→

```
VAPFPEQFLQ
ISLEICSIFHDN
```

2 Unique k-mers from data set with atchley factors

	0	1	2	3	4	...	10	11	12	13	14
APF	-0.591	-1.302	-0.733	1.570	-0.146	...	-1.006	-0.590	1.891	-0.397	0.412
CSI	-1.343	0.465	-0.862	-1.020	-0.255	...	-1.239	-0.547	2.131	0.393	0.816
EIC	1.357	-1.453	1.477	0.113	-0.837	...	-1.343	0.465	-0.862	-1.020	-0.255
EQF	1.357	-1.453	1.477	0.113	-0.837	...	-1.006	-0.590	1.891	-0.397	0.412
FHD	-1.006	-0.590	1.891	-0.397	0.412	...	1.050	0.302	-3.656	-0.259	-3.242
FLQ	-1.006	-0.590	1.891	-0.397	0.412	...	0.931	-0.179	-3.005	-0.503	-1.853
FPE	-1.006	-0.590	1.891	-0.397	0.412	...	1.357	-1.453	1.477	0.113	-0.837
HDN	0.336	-0.417	-1.673	-1.474	-0.078	...	0.945	0.828	1.299	-0.169	0.933
ICS	-1.239	-0.547	2.131	0.393	0.816	...	-0.228	1.399	-4.760	0.670	-2.647
IFH	-1.239	-0.547	2.131	0.393	0.816	...	0.336	-0.417	-1.673	-1.474	-0.078
ISL	-1.239	-0.547	2.131	0.393	0.816	...	-1.019	-0.987	-1.505	1.266	-0.912
LEI	-1.019	-0.987	-1.505	1.266	-0.912	...	-1.239	-0.547	2.131	0.393	0.816
PEQ	0.189	2.081	-1.628	0.421	-1.392	...	0.931	-0.179	-3.005	-0.503	-1.853
PFP	0.189	2.081	-1.628	0.421	-1.392	...	0.189	2.081	-1.628	0.421	-1.392
QFL	0.931	-0.179	-3.005	-0.503	-1.853	...	-1.019	-0.987	-1.505	1.266	-0.912
SIF	-0.228	1.399	-4.760	0.670	-2.647	...	-1.006	-0.590	1.891	-0.397	0.412
SLE	-0.228	1.399	-4.760	0.670	-2.647	...	1.357	-1.453	1.477	0.113	-0.837
VAP	-1.337	-0.279	-0.544	1.242	-1.262	...	0.189	2.081	-1.628	0.421	-1.392

[18 rows x 15 columns]

3 Assign each k-mer to cluster

```
Out[20]:
APF    0
CSI    0
EIC    1
EQF    0
FHD    1
FLQ    1
FPE    0
HDN    0
ICS    1
IFH    1
ISL    0
LEI    2
PEQ    2
PFP    2
QFL    2
SIF    2
SLE    2
VAP    1
dtype: int32
```

4 Calculate fraction of k-mers belonging to each cluster per sequence

	0	1	2
VAPFPEQFLQ	0.125	0.375	0.5
ISLEICSIFHDN	0.400	0.300	0.3

Reduced Alphabet Encoding

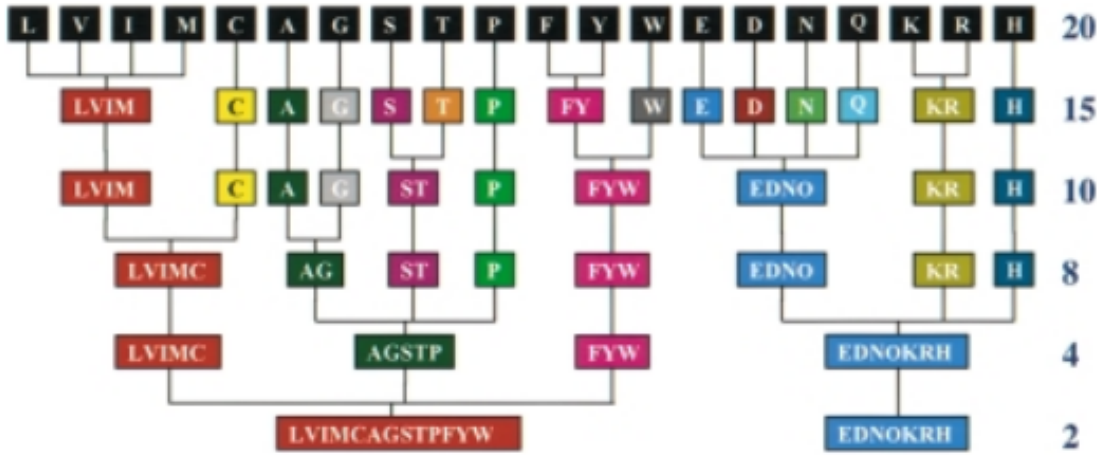
Encoding	Equation	Dimensions
Atchley clust.	$100C$	100D
Atchley	$97KM \cdot 15AF$	1455D
One-hot	$99AA \cdot 22LE$	2178D
Reduced alphabet	$99AA \cdot 11LE$	1089D
Word2Vec clust.	$100C$	100D
Word2Vec	$97KM \cdot 200W2V$	19400D
ELMo summed	$99AA \cdot 1024ELS$	101376D
ELMo	$99AA \cdot 3072EL$	304128D

Table 1: Dimensionality of the different encodings with a sequence set of maximum length 99. Legend: C = cluster, D = dimensions, KM = k-mer, AF = Atchley factors, AA = Amino acids, LE = Letters, $W2V$ = Word2Vec dimensions, ELS = Summed ELMo dimensions, EL = ELMo dimensions.

Reduced Alphabet Encoding

Encoding	Equation	Dimensions
Atchley clust.	$100C$	100D
Atchley	$97KM \cdot 15AF$	1455D
One-hot	$99AA \cdot 22LE$	2178D
Reduced alphabet	$99AA \cdot 11LE$	1089D
Word2Vec clust.	$100C$	100D
Word2Vec	$97KM \cdot 200W2V$	19400D
ELMo summed	$99AA \cdot 1024ELS$	101376D
ELMo	$99AA \cdot 3072EL$	304128D

Table 1: Dimensionality of the different encodings with a sequence set of maximum length 99. Legend: C = cluster, D = dimensions, KM = k-mer, AF = Atchley factors, AA = Amino acids, LE = Letters, $W2V$ = Word2Vec dimensions, ELS = Summed ELMo dimensions, EL = ELMo dimensions.



Word2Vec countries and capital cities

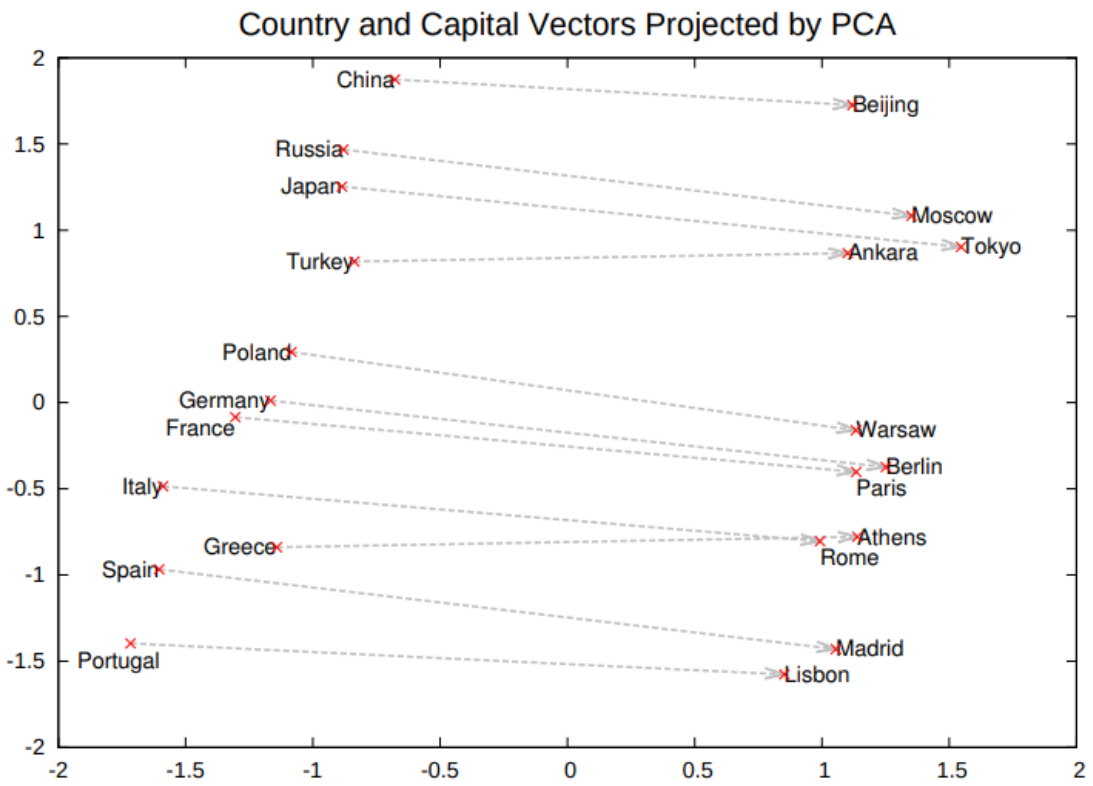
Encoding	Equation	Dimensions
Atchley clust.	$100C$	100D
Atchley	$97KM \cdot 15AF$	1455D
One-hot	$99AA \cdot 22LE$	2178D
Reduced alphabet	$99AA \cdot 11LE$	1089D
Word2Vec clust.	$100C$	100D
Word2Vec	$97KM \cdot 200W2V$	19400D
ELMo summed	$99AA \cdot 1024ELS$	101376D
ELMo	$99AA \cdot 3072EL$	304128D

Table 1: Dimensionality of the different encodings with a sequence set of maximum length 99. Legend: C = cluster, D = dimensions, KM = k-mer, AF = Atchley factors, AA = Amino acids, LE = Letters, $W2V$ = Word2Vec dimensions, ELS = Summed ELMo dimensions, EL = ELMo dimensions.

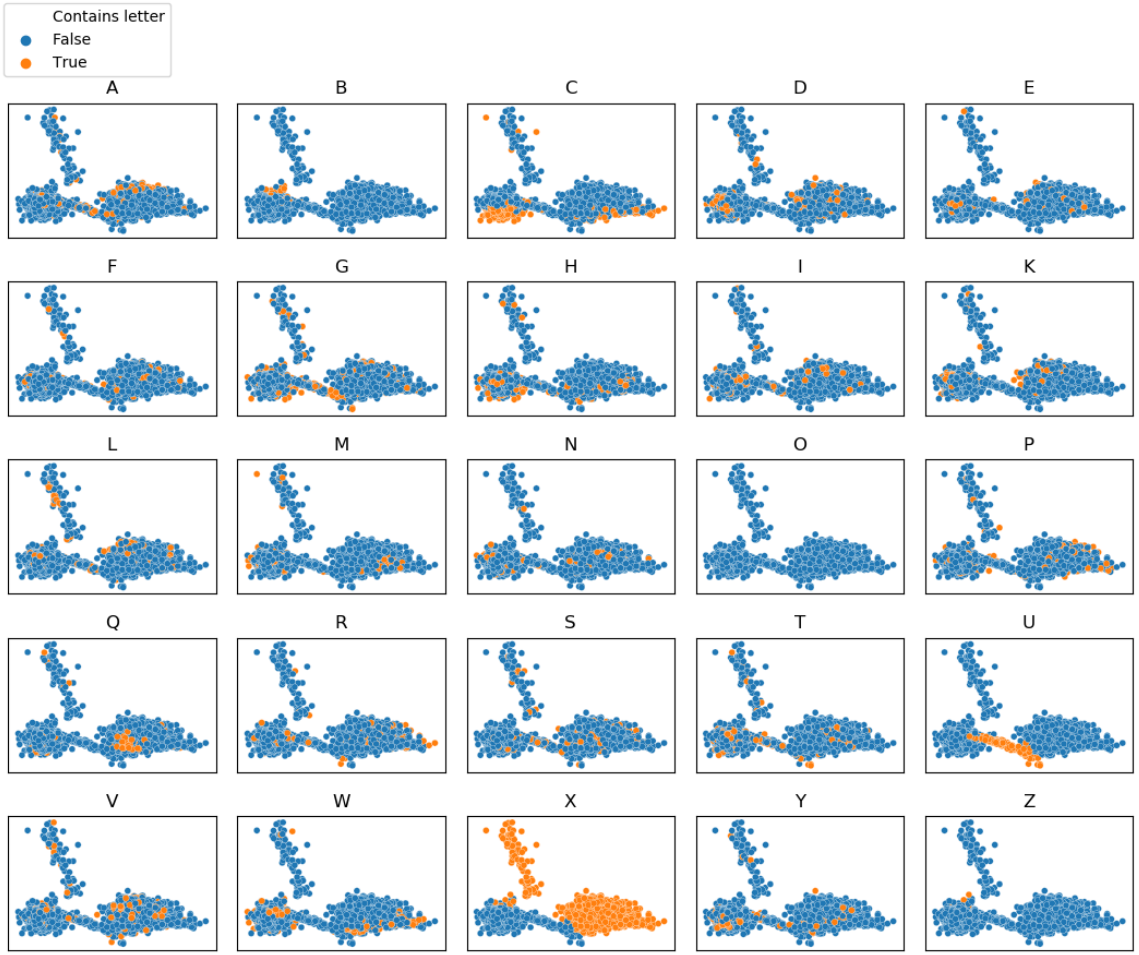
Word2Vec countries and capital cities

Encoding	Equation	Dimensions
Atchley clust.	$100C$	100D
Atchley	$97KM \cdot 15AF$	1455D
One-hot	$99AA \cdot 22LE$	2178D
Reduced alphabet	$99AA \cdot 11LE$	1089D
Word2Vec clust.	$100C$	100D
Word2Vec	$97KM \cdot 200W2V$	19400D
ELMo summed	$99AA \cdot 1024ELS$	101376D
ELMo	$99AA \cdot 3072EL$	304128D

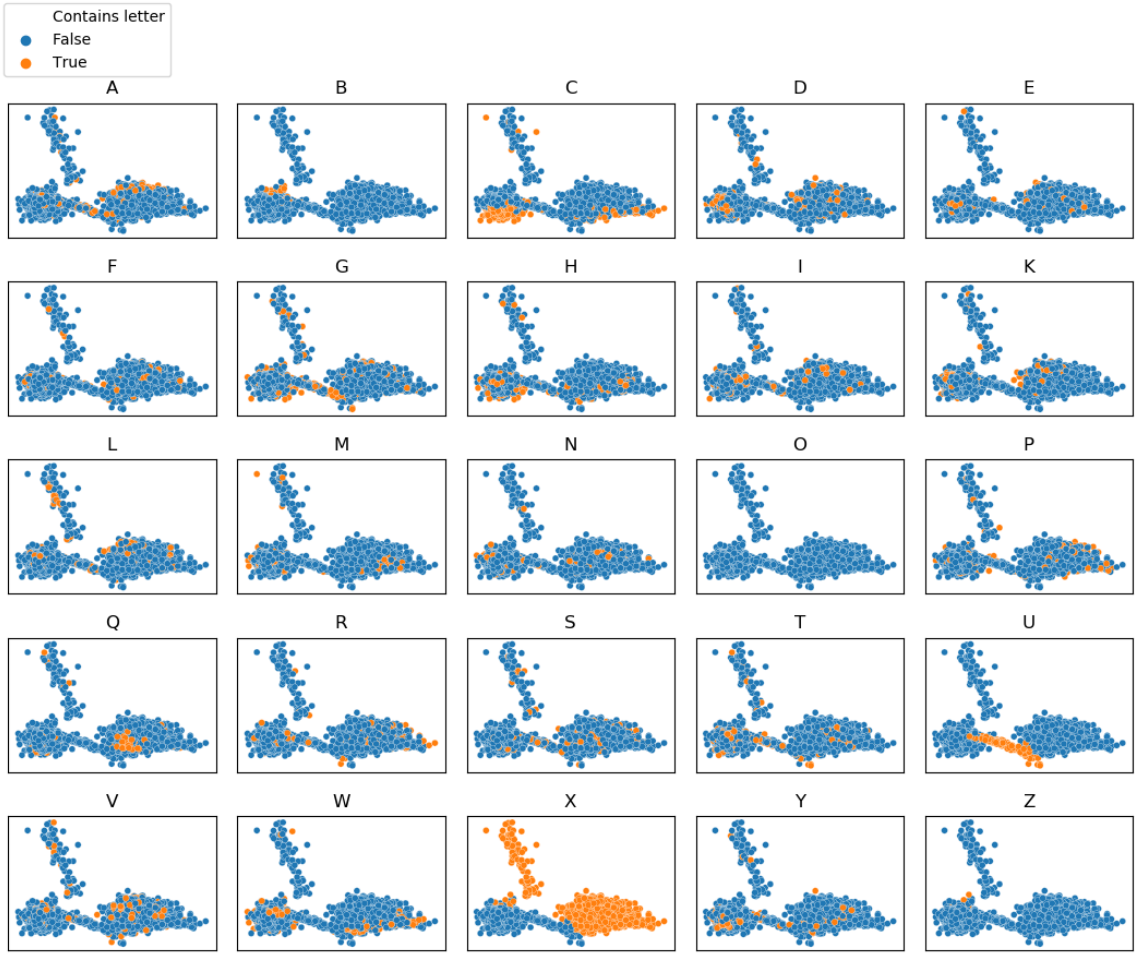
Table 1: Dimensionality of the different encodings with a sequence set of maximum length 99. Legend: C = cluster, D = dimensions, KM = k-mer, AF = Atchley factors, AA = Amino acids, LE = Letters, $W2V$ = Word2Vec dimensions, ELS = Summed ELMo dimensions, EL = ELMo dimensions.

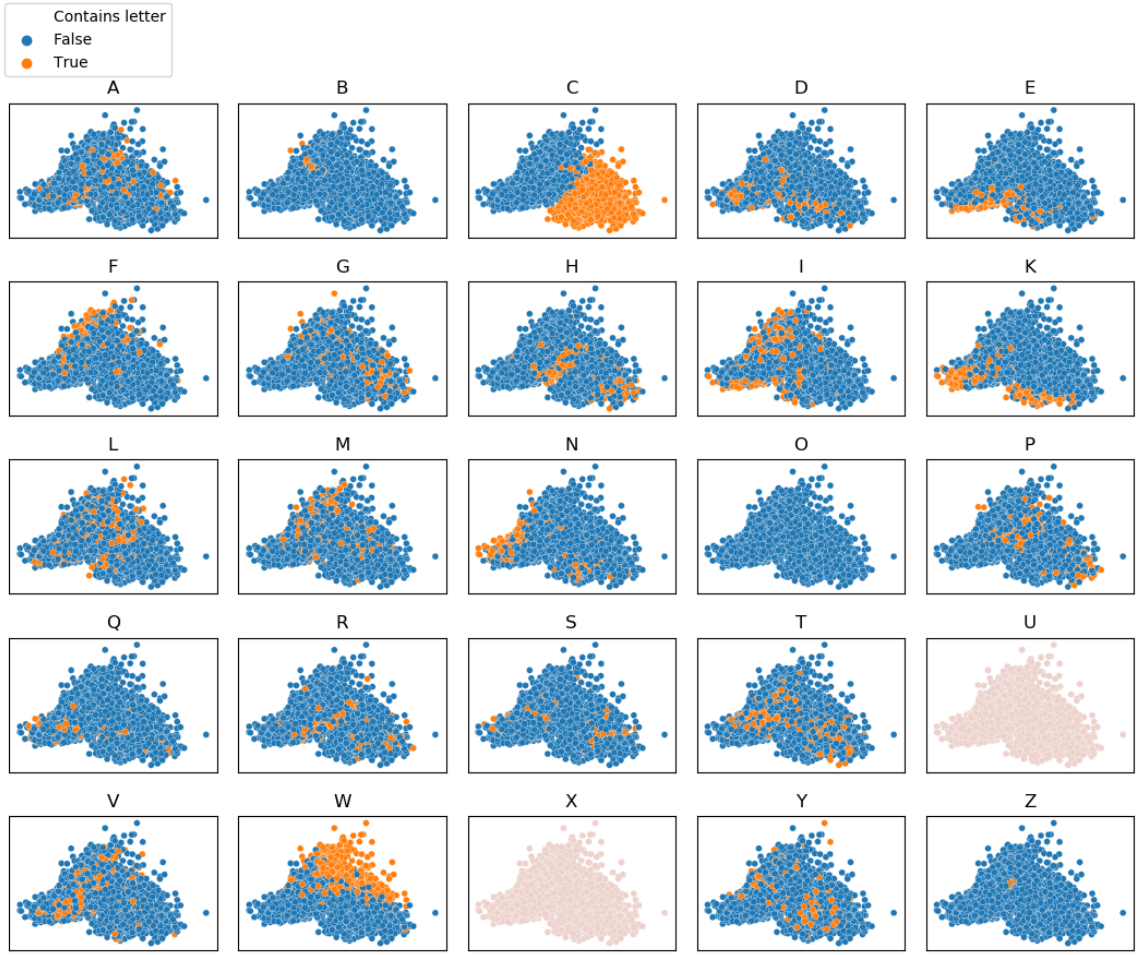


Word2Vec 3-mer PCA projections

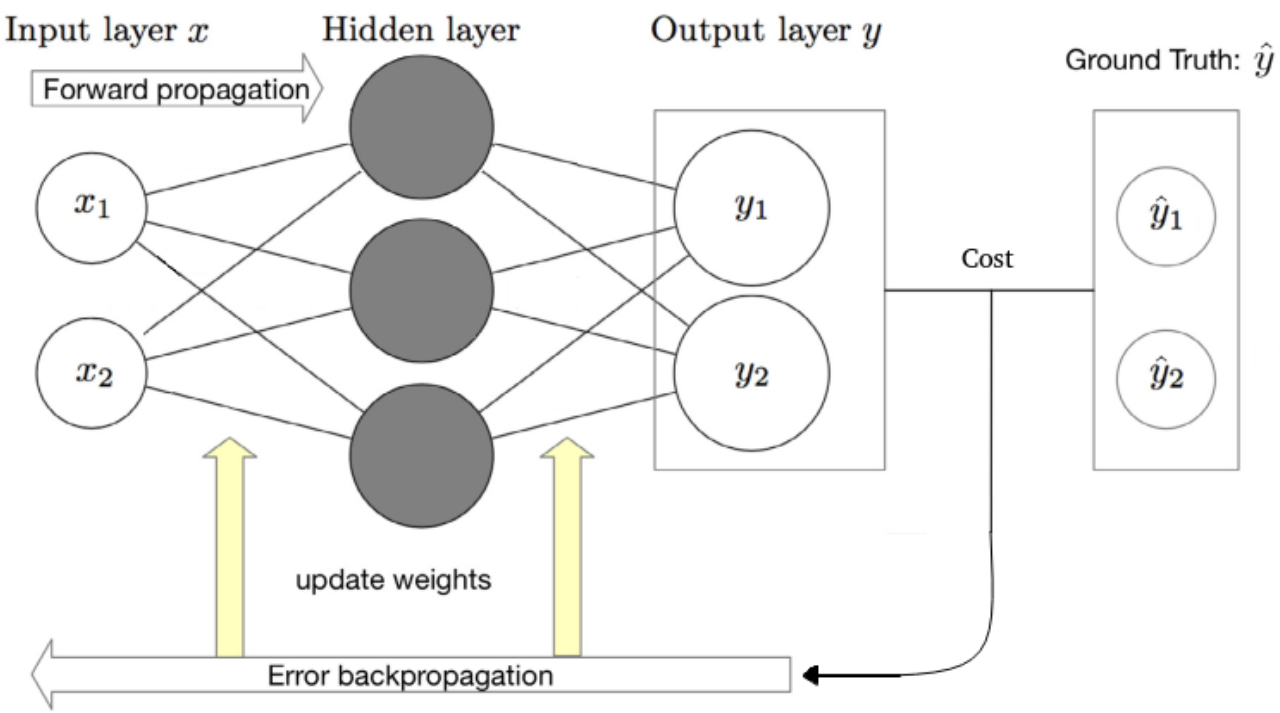


Word2Vec 3-mer PCA projections

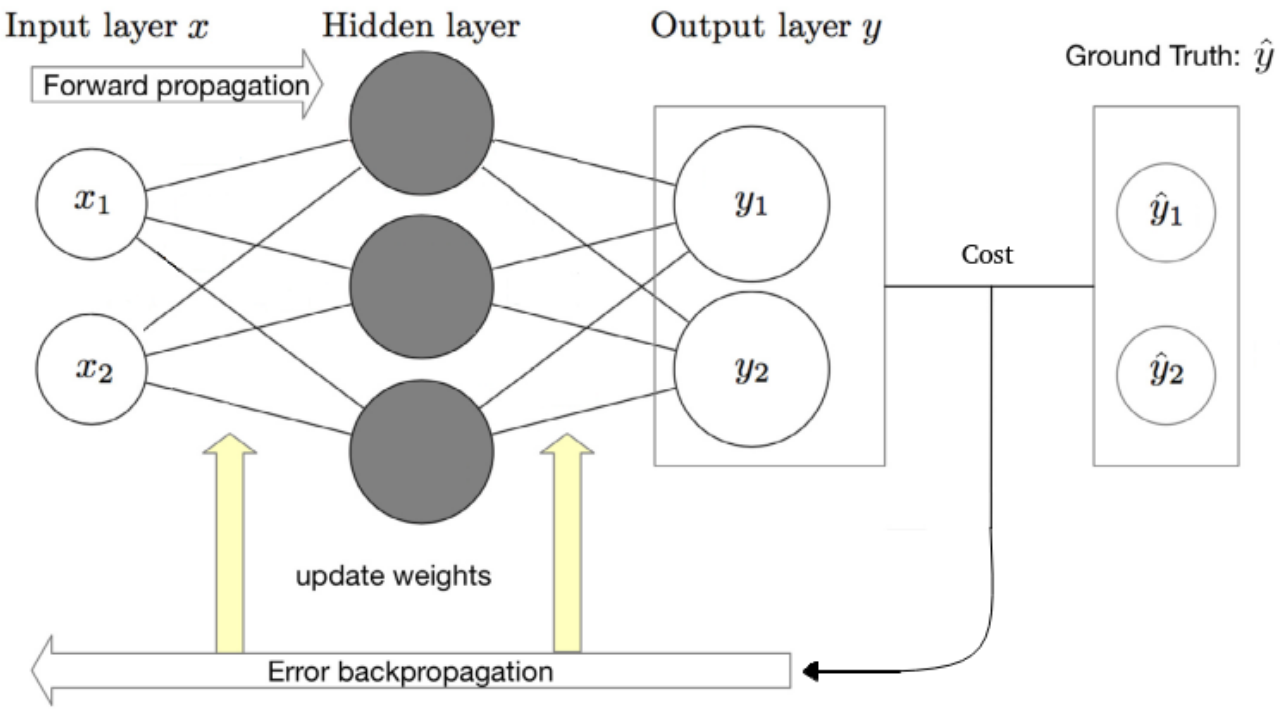




Deep Neural Network

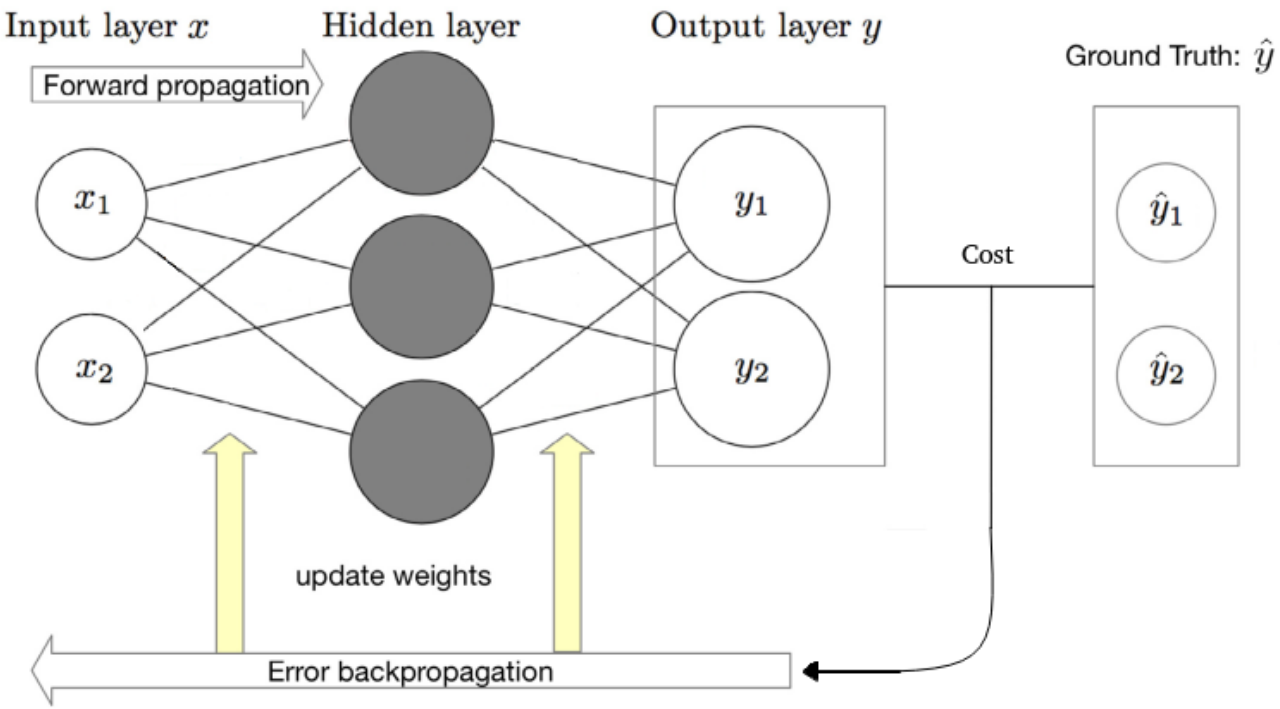


~~Deep~~ Neural Network



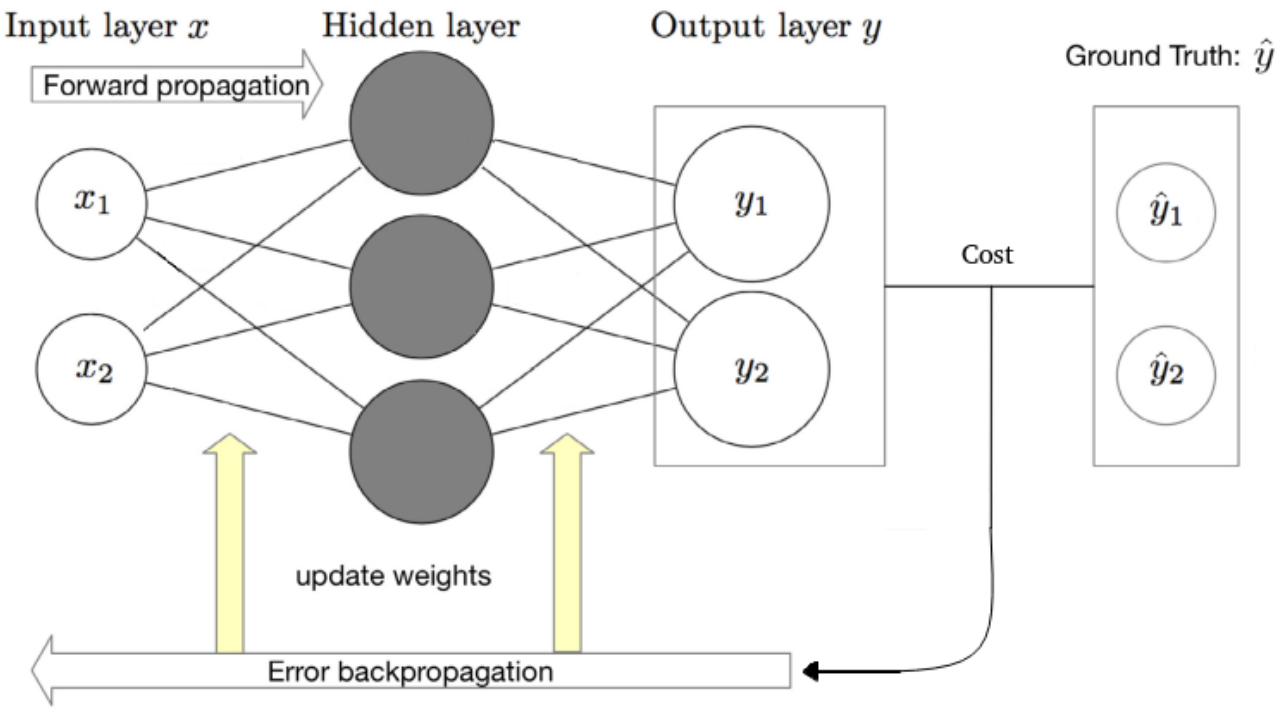
~~Deep~~ Neural Network

- Multiply input with weights and add biases



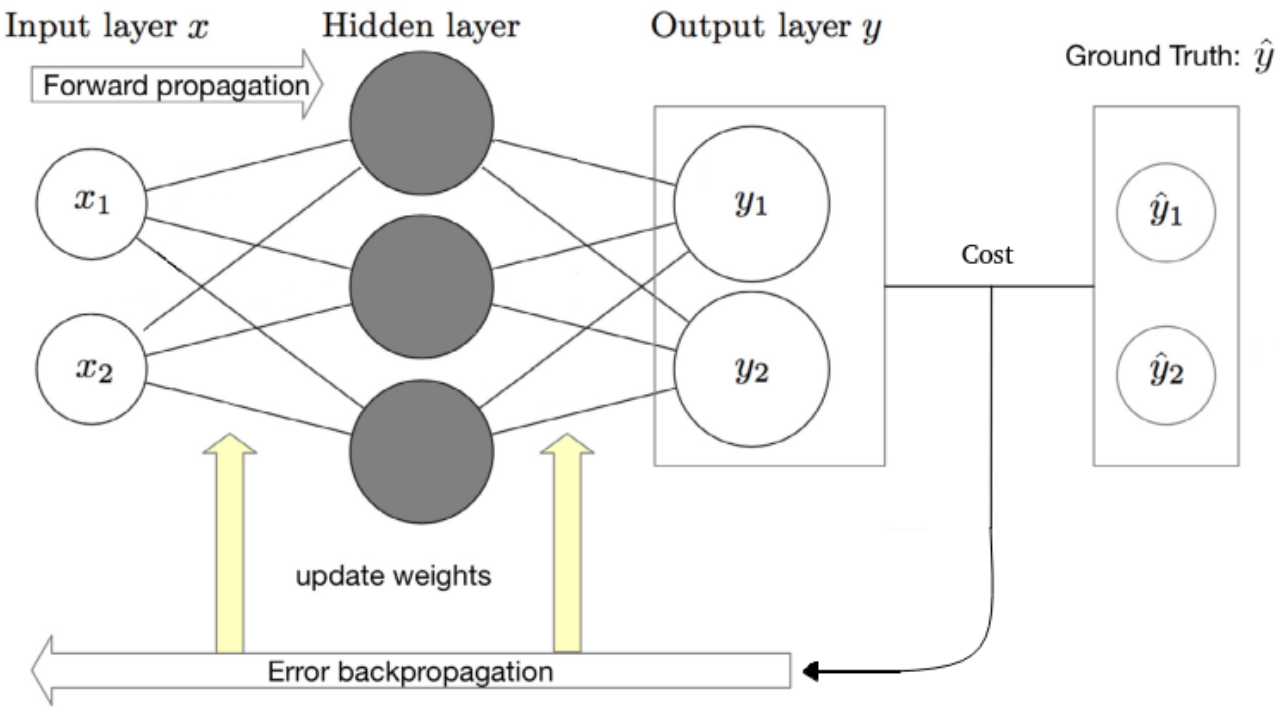
~~Deep~~ Neural Network

- Multiply input with weights and add biases
- Sum in next node



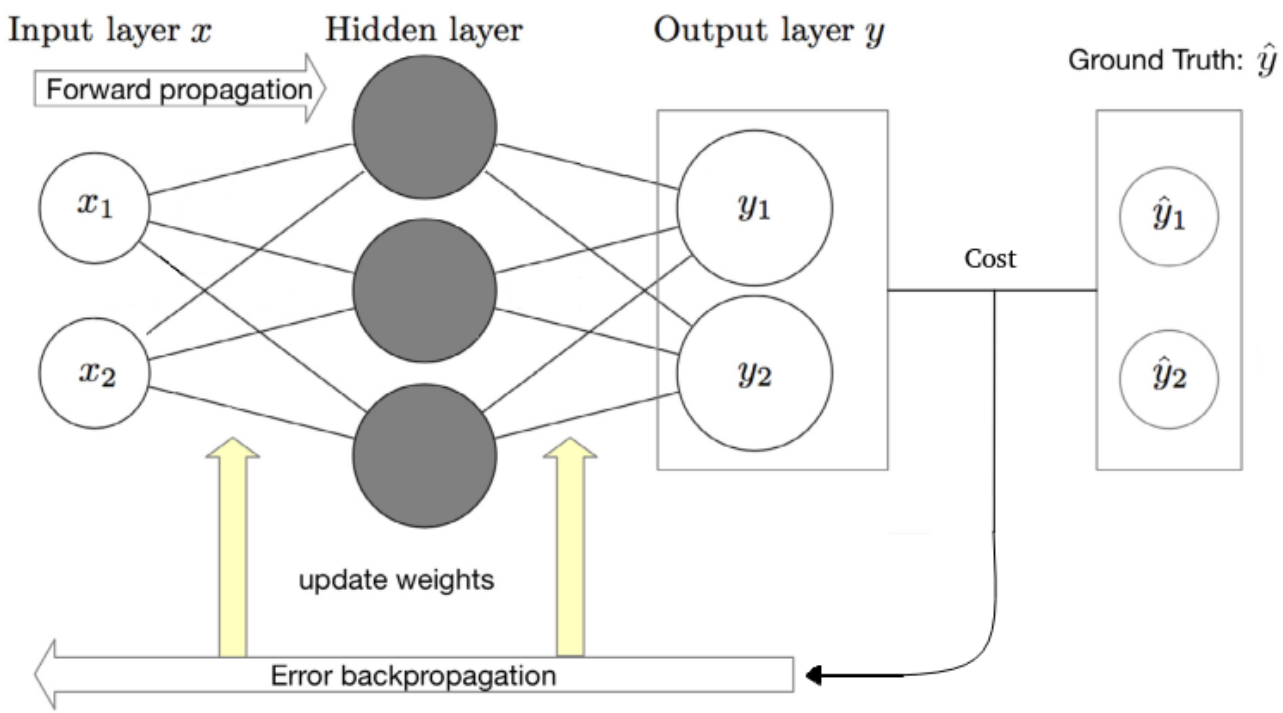
~~Deep~~ Neural Network

- Multiply input with weights and add biases
- Sum in next node
- Activate node



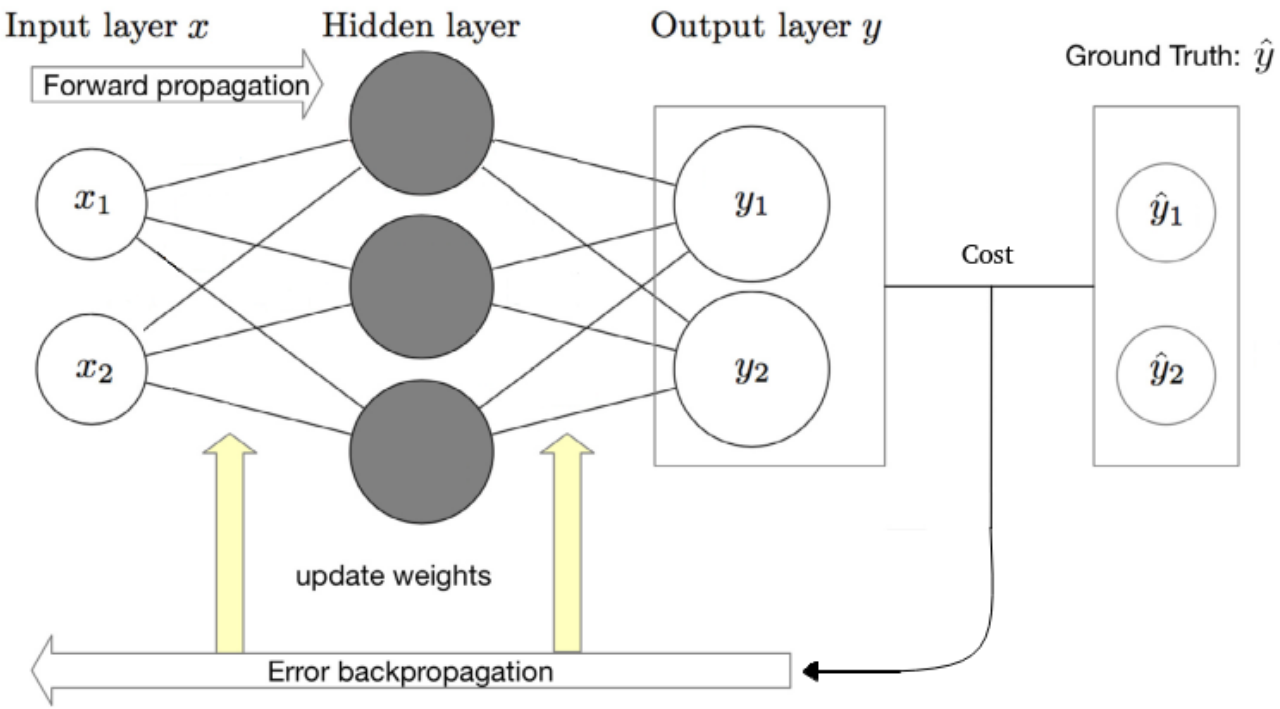
~~Deep~~ Neural Network

- Multiply input with weights and add biases
- Sum in next node
- Activate node
- Gradient of cost function with respect to weights and biases



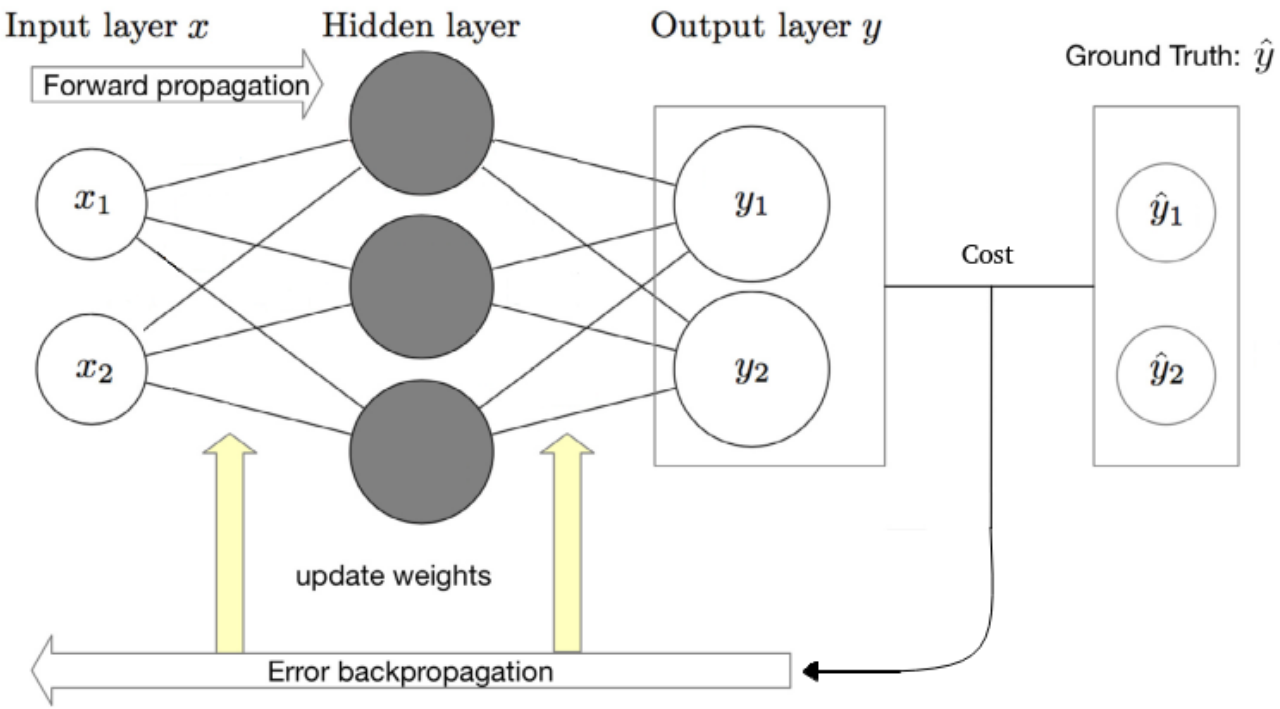
~~Deep~~ Neural Network

- Multiply input with weights and add biases
- Sum in next node
- Activate node
- Gradient of cost function with respect to weights and biases
- Back-propagate through network

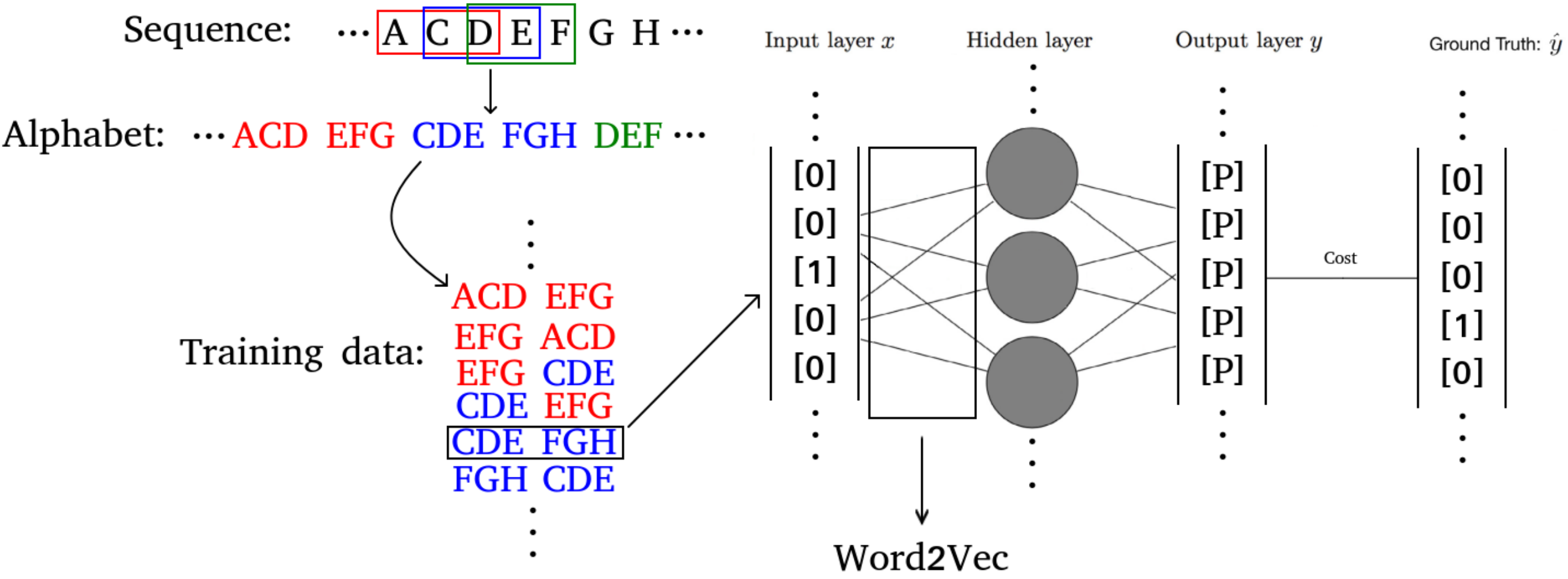


~~Deep~~ Neural Network

- Multiply input with weights and add biases
- Sum in next node
- Activate node
- Gradient of cost function with respect to weights and biases
- Back-propagate through network
- Step in the opposite direction of the gradient to minimize cost



The Word2Vec architecture



ELMo Embedding

Encoding	Equation	Dimensions
Atchley clust.	$100C$	100D
Atchley	$97KM \cdot 15AF$	1455D
One-hot	$99AA \cdot 22LE$	2178D
Reduced alphabet	$99AA \cdot 11LE$	1089D
Word2Vec clust.	$100C$	100D
Word2Vec	$97KM \cdot 200W2V$	19400D
ELMo summed	$99AA \cdot 1024ELS$	101376D
ELMo	$99AA \cdot 3072EL$	304128D

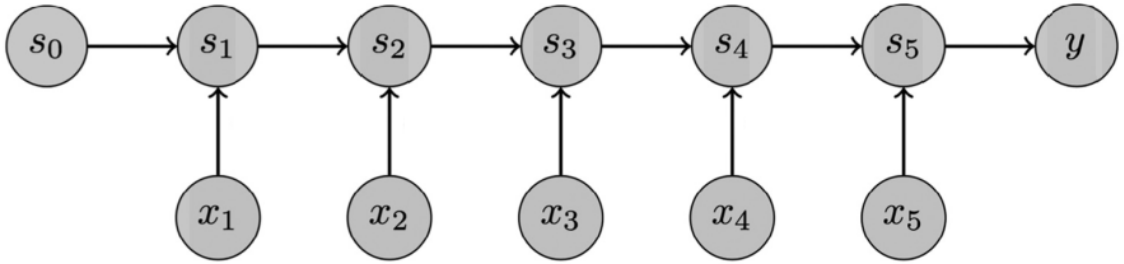
Table 1: Dimensionality of the different encodings with a sequence set of maximum length 99. Legend: C = cluster, D = dimensions, KM = k-mer, AF = Atchley factors, AA = Amino acids, LE = Letters, $W2V$ = Word2Vec dimensions, ELS = Summed ELMo dimensions, EL = ELMo dimensions.

ELMo Embedding

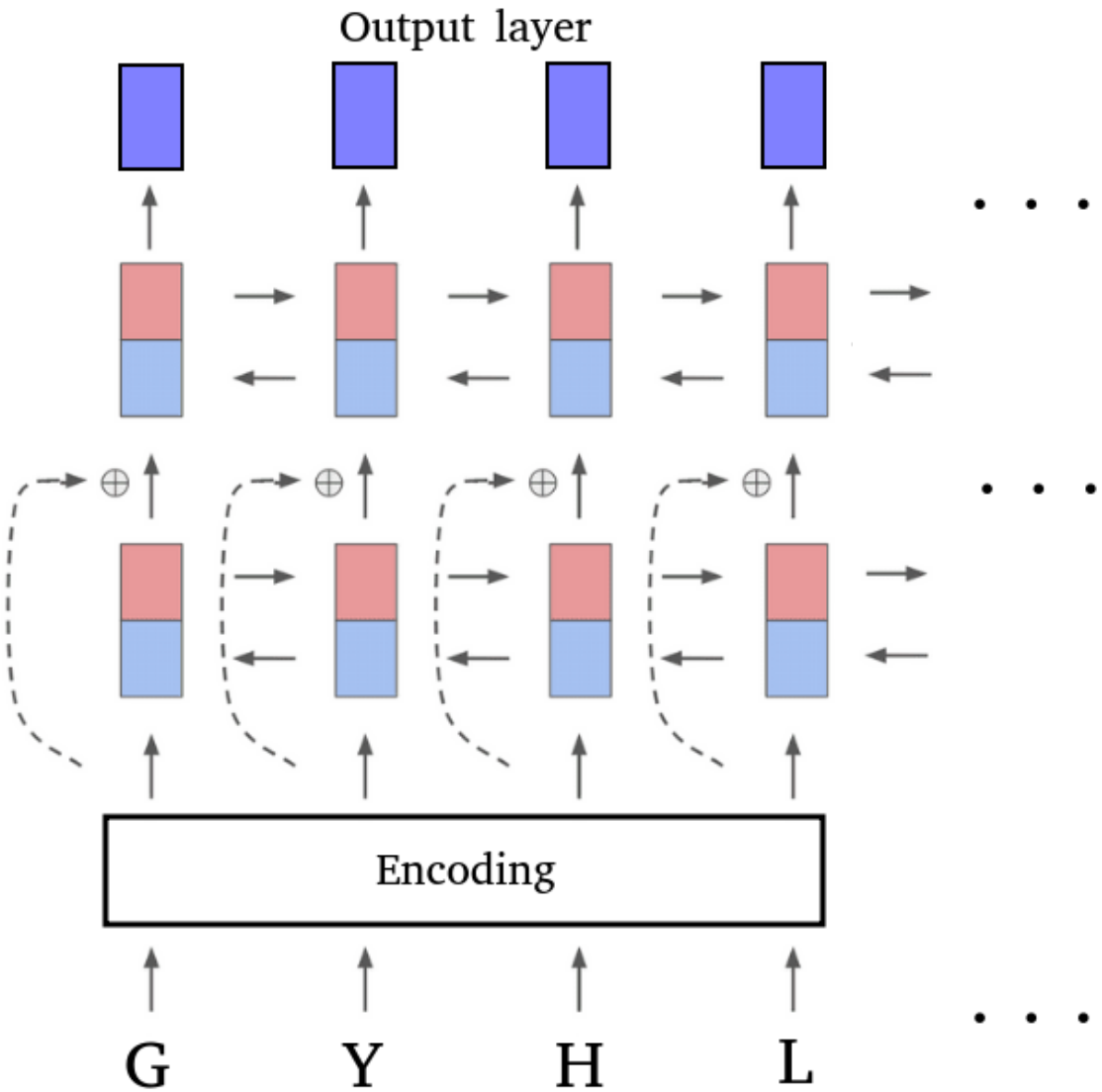
Encoding	Equation	Dimensions
Atchley clust.	$100C$	100D
Atchley	$97KM \cdot 15AF$	1455D
One-hot	$99AA \cdot 22LE$	2178D
Reduced alphabet	$99AA \cdot 11LE$	1089D
Word2Vec clust.	$100C$	100D
Word2Vec	$97KM \cdot 200W2V$	19400D
ELMo summed	$99AA \cdot 1024ELS$	101376D
ELMo	$99AA \cdot 3072EL$	304128D

Table 1: Dimensionality of the different encodings with a sequence set of maximum length 99. Legend: C = cluster, D = dimensions, KM = k-mer, AF = Atchley factors, AA = Amino acids, LE = Letters, $W2V$ = Word2Vec dimensions, ELS = Summed ELMo dimensions, EL = ELMo dimensions.

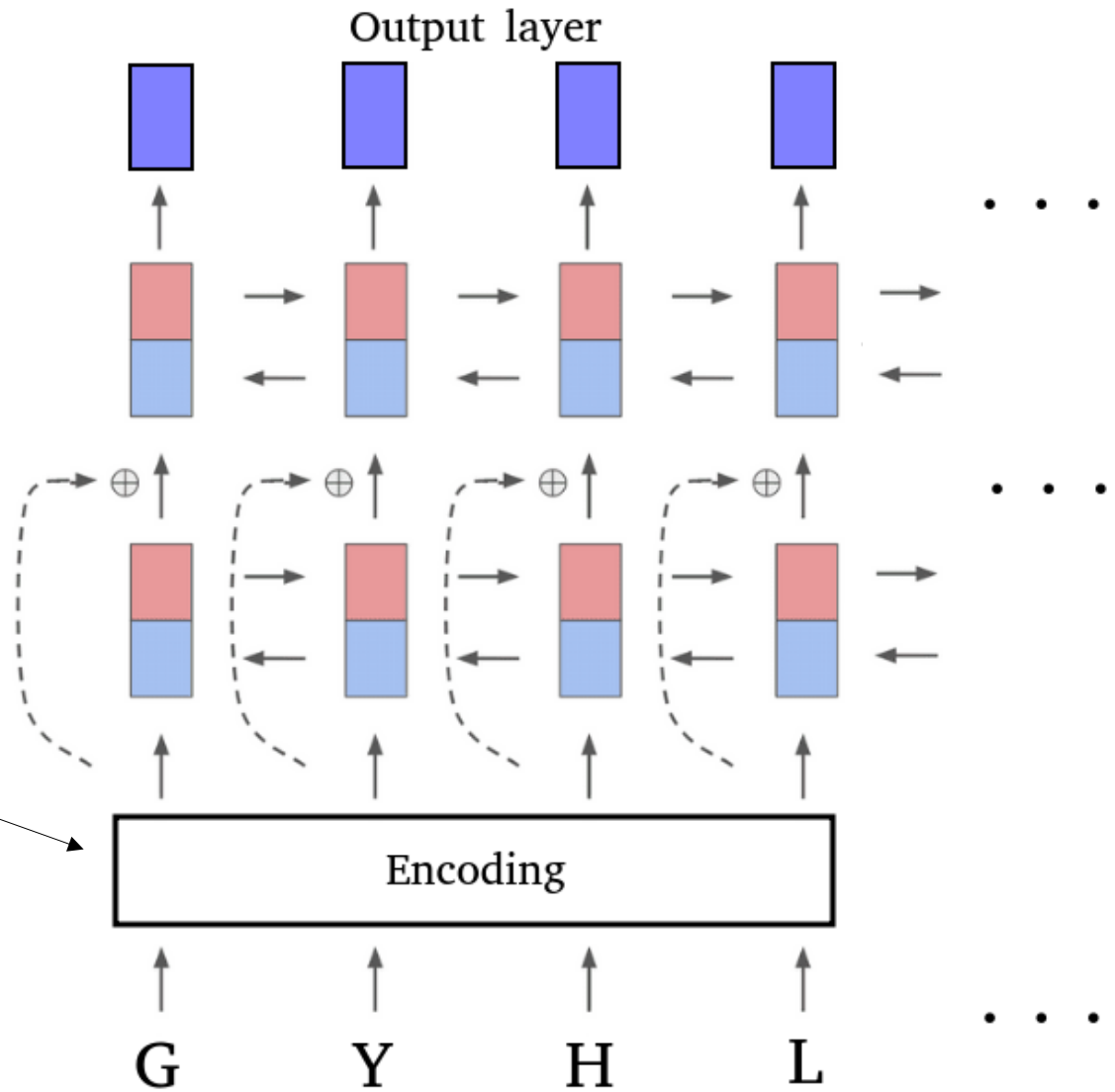
- Recurrent Neural Network



ELMo architecture

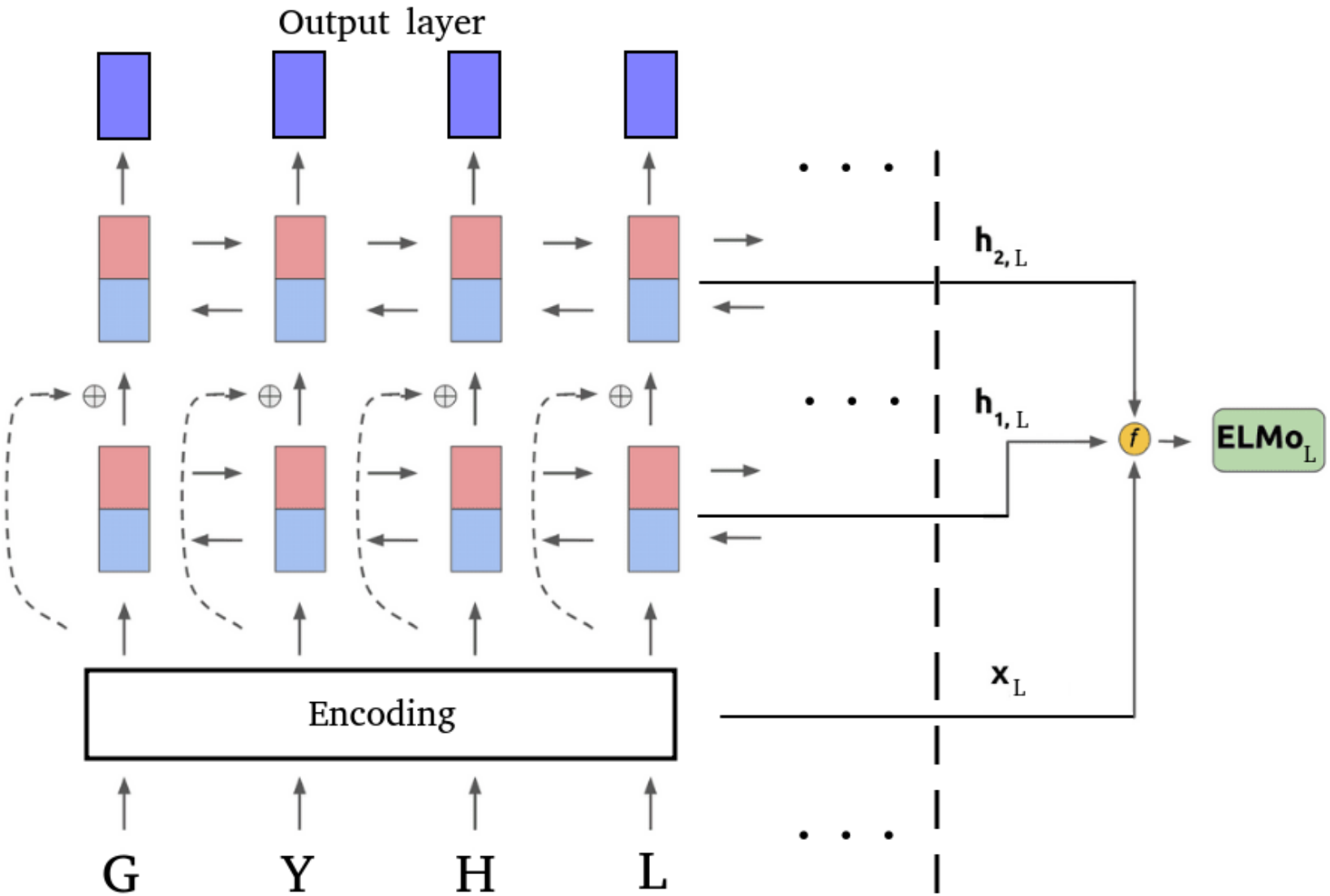


ELMo architecture

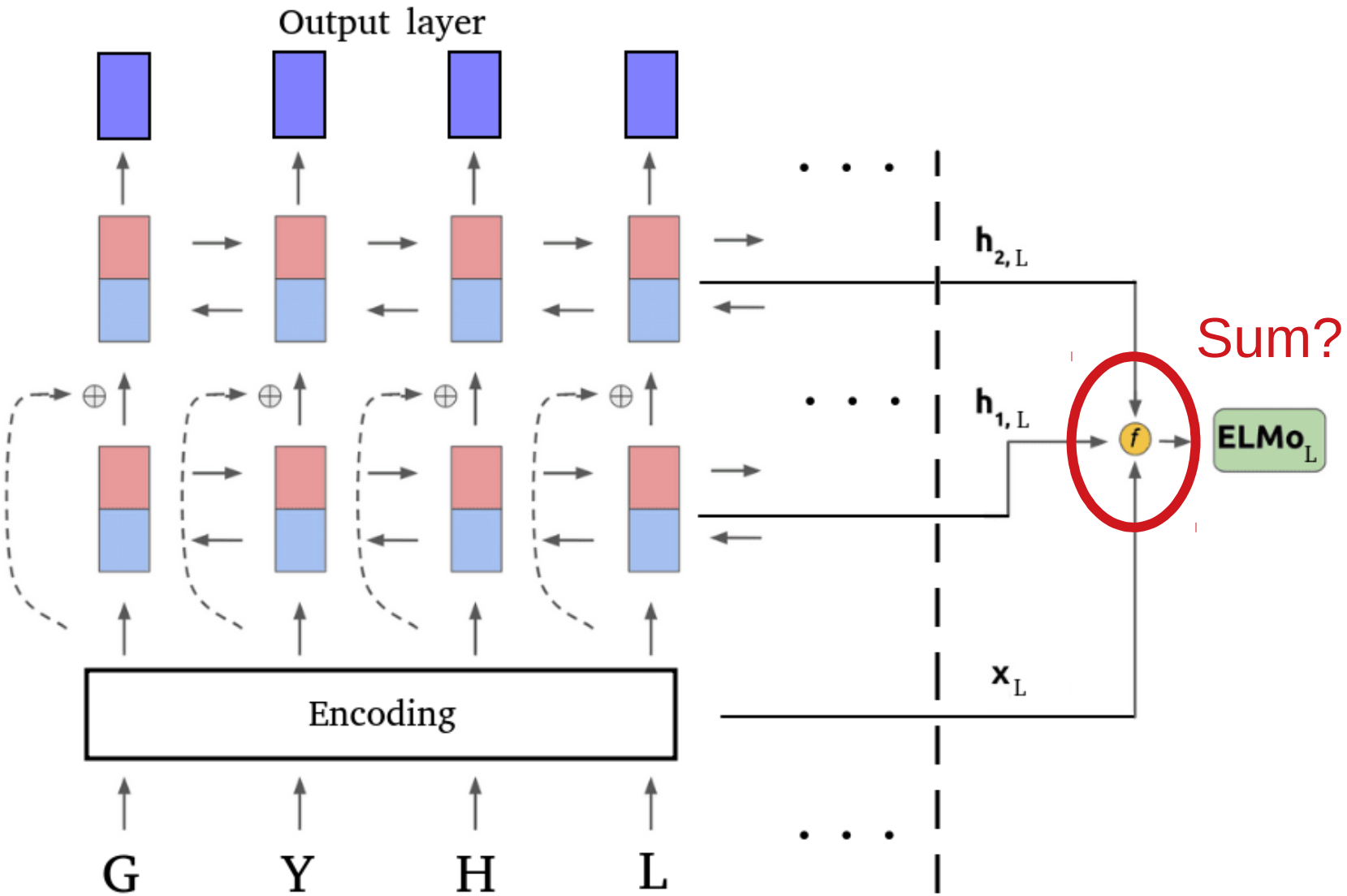


Authors picked convolutional network

ELMo architecture



ELMo architecture



Embedding models		
Name	Source	Trained on
Word2Vec	Md-Nafiz Hamid and Iddo Friedberg. Identifying antimicrobial peptides using word embedding with deep recurrent neural networks. Bioinformatics, 35(12):2009–2016, jun 2019.	Uniprot/TrEMBL database
ELMo	Michael Heinzinger, Ahmed Elnaggar, Yu Wang, Christian Dallago, Dmitrii Nechaev, Florian Matthes, and Burkhard Rost. Modeling the Language of Life – Deep Learning Protein Sequences, 2019.	UniRef50

Contents

Encoding selection

- Similar data set
- Linear support vector machine
- Cross validation procedure with # folds = 10
- Goal: Select encodings to move forward with

Data sets		
Name	Source	# data points
GLY	UniProt GO term: "Glycolytic-process"; taxonomy: Bacteria; length: 14-99 amino acids	1954
LIP	UniProt GO term: "Lipid-A-biosynthetic-process"; taxonomy: Bacteria; length: 14-99 amino acids	1142
UNI	UniProt keywords: "not Antibiotic", "not Antimicrobial", "not Plasmid"; taxonomy: Bacteria; length: 10-359 amino acids	1003
BAC	CAMP (anything containing "bacteriocin"), BAGEL, Bactibase; length: 10-359 amino acids	1003

Training data					
Name	Positive	Negative	Pos. Neg. Ratio	# training points	# test points
UniProt	LIP	GLY	0.66:1	2476	620
Bacteriocin	BAC	UNI	1:1	1604	402

Data sets		
Name	Source	# data points
GLY	UniProt GO term: "Glycolytic-process"; taxonomy: Bacteria; length: 14-99 amino acids	1954
LIP	UniProt GO term: "Lipid-A-biosynthetic-process"; taxonomy: Bacteria; length: 14-99 amino acids	1142
UNI	UniProt keywords: "not Antibiotic", "not Antimicrobial", "not Plasmid"; taxonomy: Bacteria; length: 10-359 amino acids	1003
BAC	CAMP (anything containing "bacteriocin"), BAGEL, Bactibase; length: 10-359 amino acids	1003

Training data					
Name	Positive	Negative	Pos. Neg. Ratio	# training points	# test points
UniProt	LIP	GLY	0.66:1	2476	620
Bacteriocin	BAC	UNI	1:1	1604	402

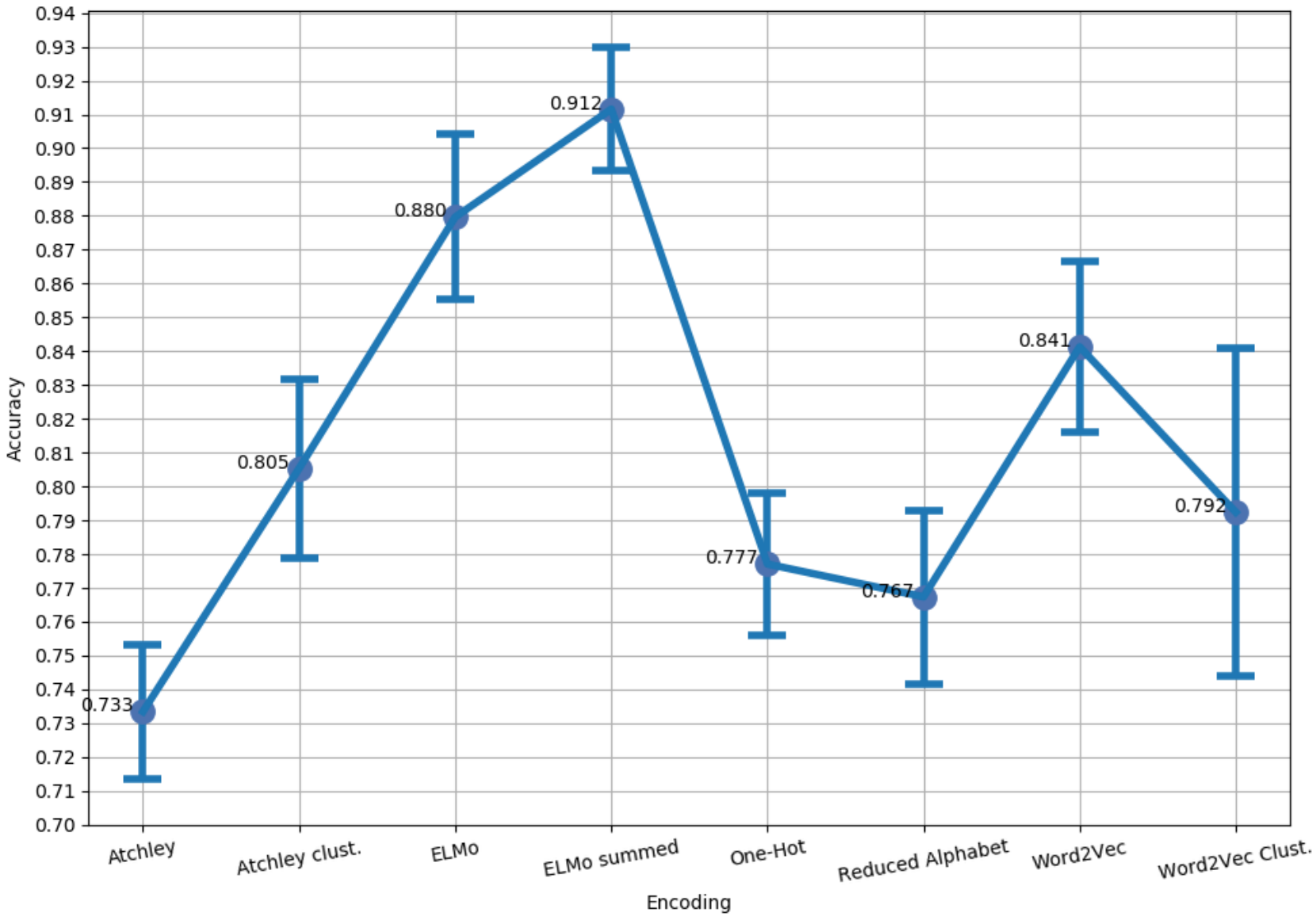
Data sets		
Name	Source	# data points
GLY	UniProt GO term: "Glycolytic-process"; taxonomy: Bacteria; length: 14-99 amino acids	1954
LIP	UniProt GO term: "Lipid-A-biosynthetic-process"; taxonomy: Bacteria; length: 14-99 amino acids	1142
UNI	UniProt keywords: "not Antibiotic", "not Antimicrobial", "not Plasmid"; taxonomy: Bacteria; length: 10-359 amino acids	1003
BAC	CAMP (anything containing "bacteriocin"), BAGEL, Bactibase; length: 10-359 amino acids	1003

Training data					
Name	Positive	Negative	Pos. Neg. Ratio	# training points	# test points
UniProt	LIP	GLY	0.66:1	2476	620
Bacteriocin	BAC	UNI	1:1	1604	402

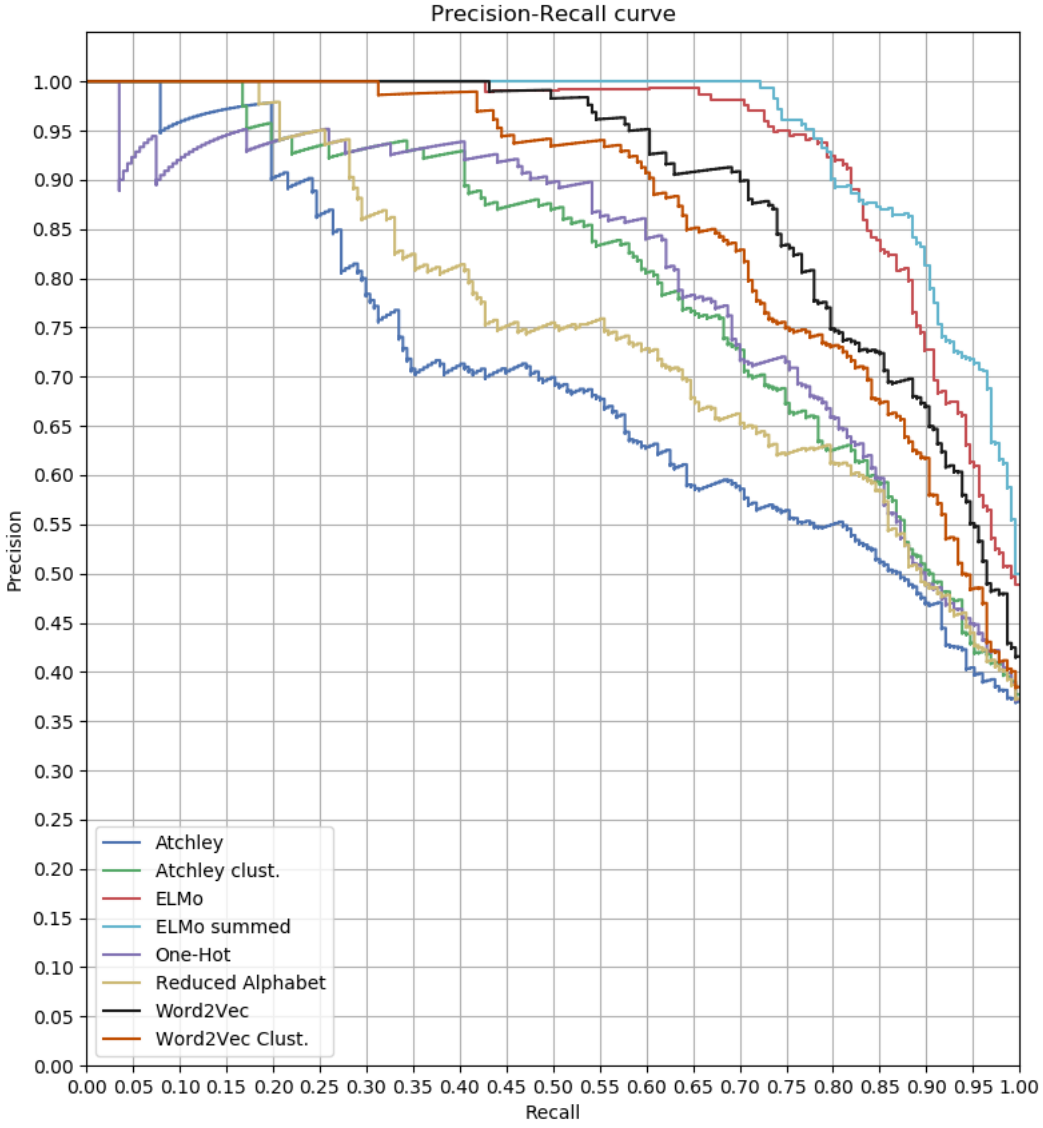
Data sets		
Name	Source	# data points
GLY	UniProt GO term: "Glycolytic-process"; taxonomy: Bacteria; length: 14-99 amino acids	1954
LIP	UniProt GO term: "Lipid-A-biosynthetic-process"; taxonomy: Bacteria; length: 14-99 amino acids	1142
UNI	UniProt keywords: "not Antibiotic", "not Antimicrobial", "not Plasmid"; taxonomy: Bacteria; length: 10-359 amino acids	1003
BAC	CAMP (anything containing "bacteriocin"), BAGEL, Bactibase; length: 10-359 amino acids	1003

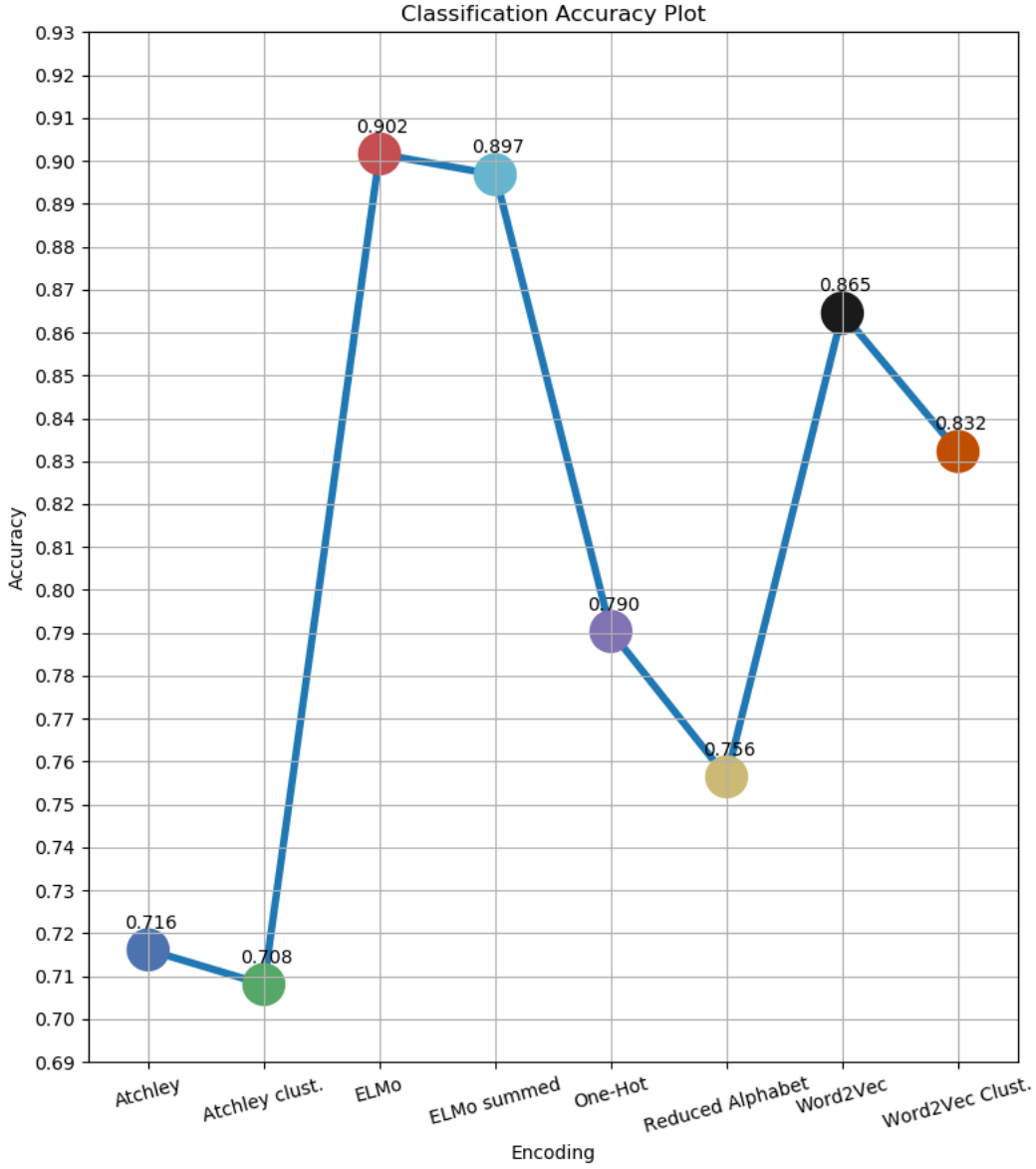
Training data					
Name	Positive	Negative	Pos. Neg. Ratio	# training points	# test points
UniProt	LIP	GLY	0.66:1	2476	620
Bacteriocin	BAC	UNI	1:1	1604	402

UniProt data set cross validation accuracy

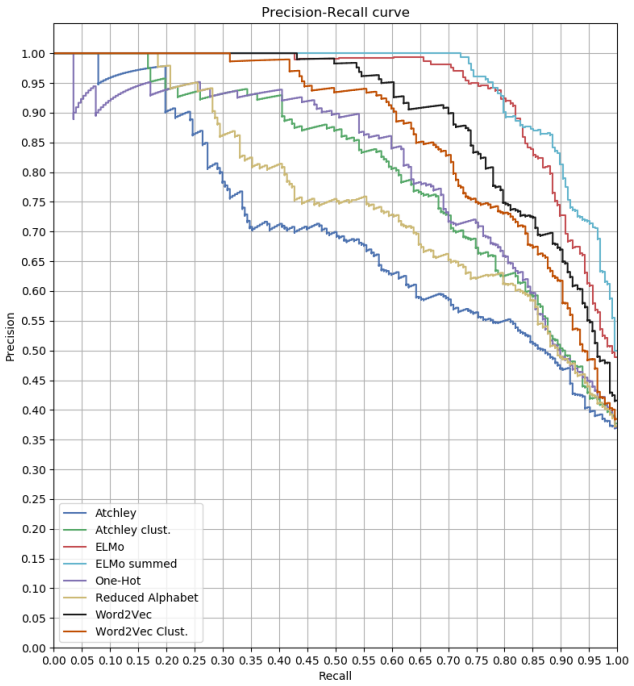


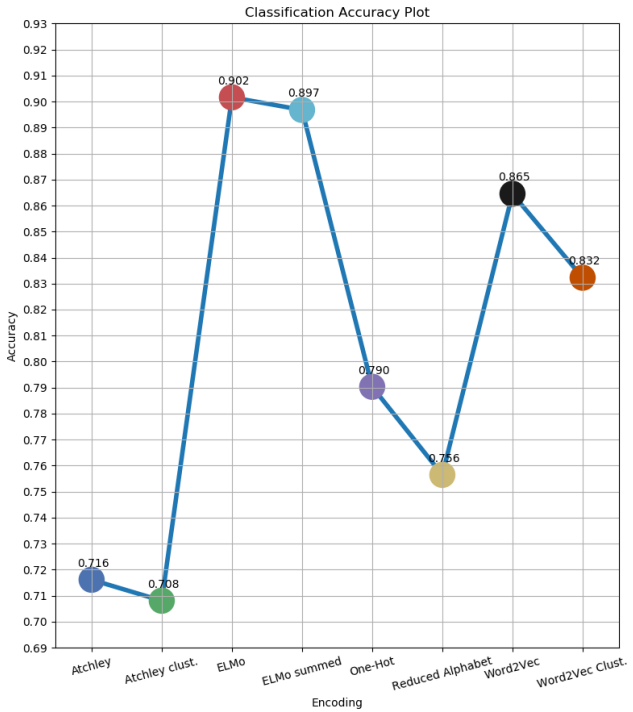
UniProt data set test accuracy

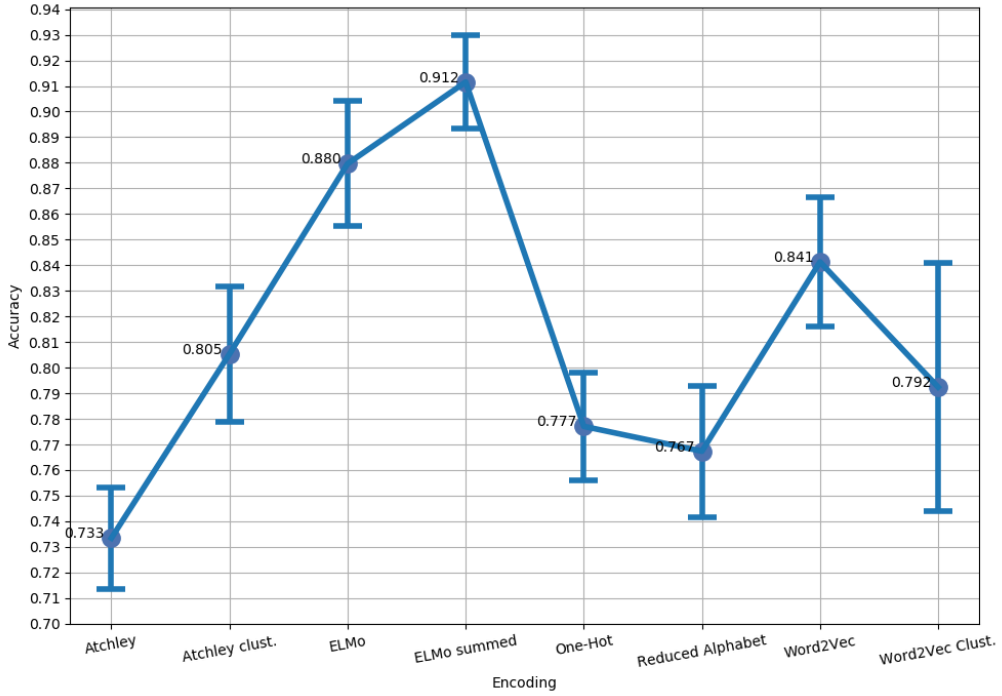




UniProt data set accuracy



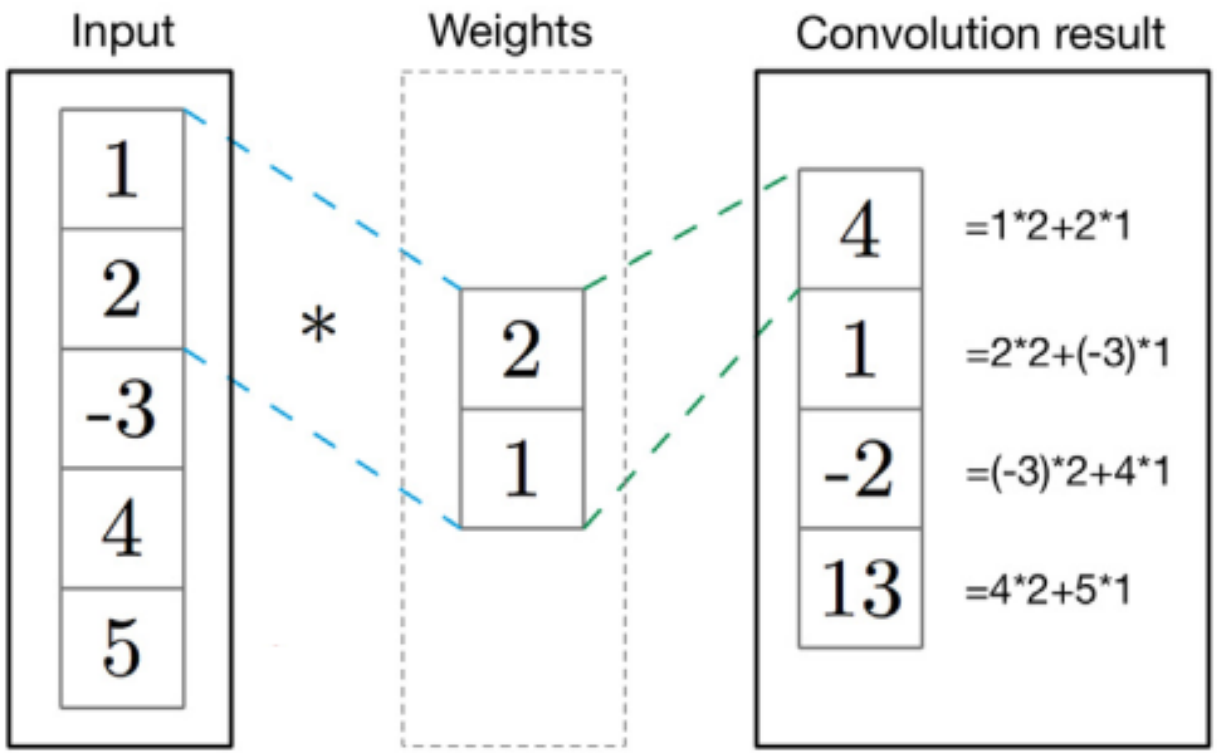




- Word embedding encodings best performance
- Summing ELMo similar to not summing
- Clustering Word2Vec decreases performance

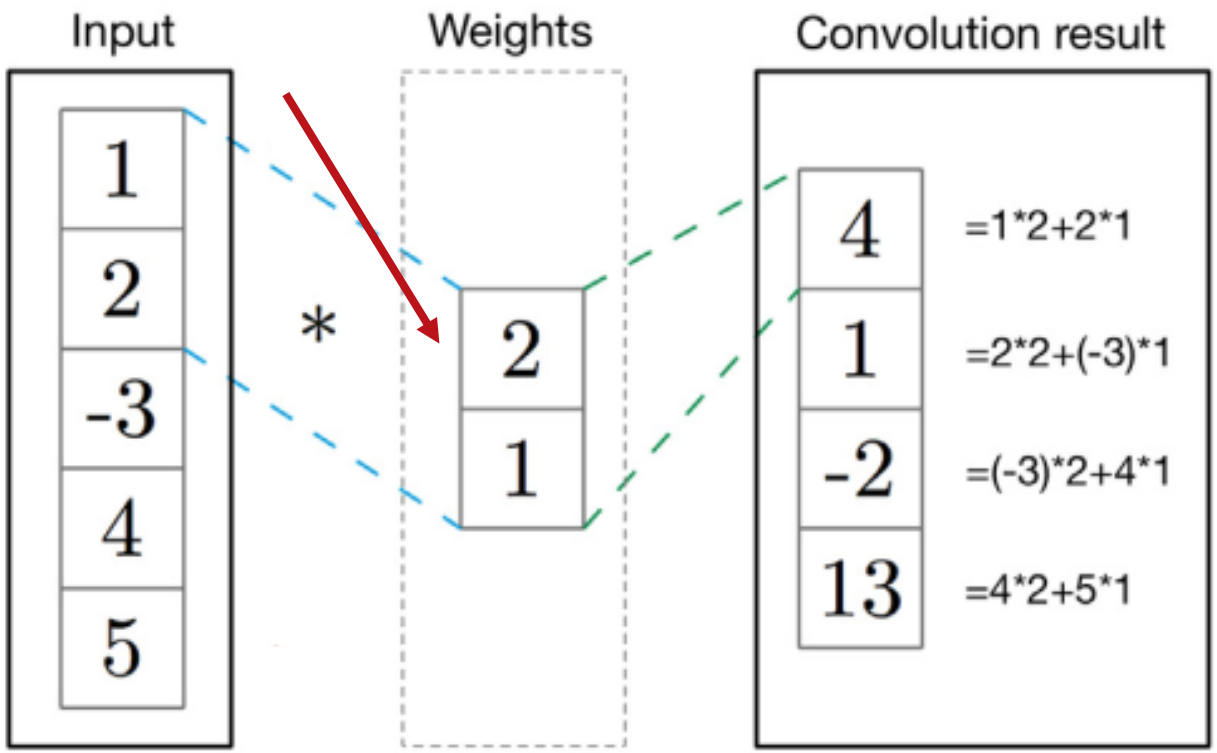
Contents

Convolutional neural network



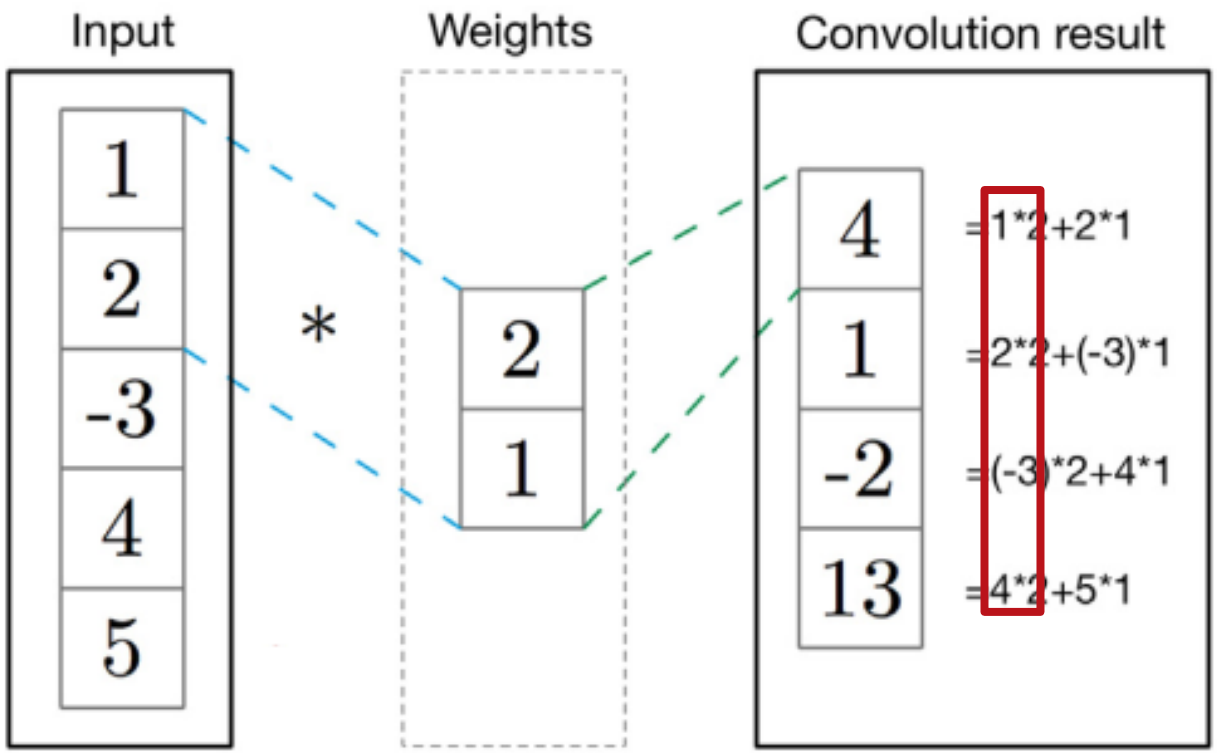
- Single filter

Convolutional neural network



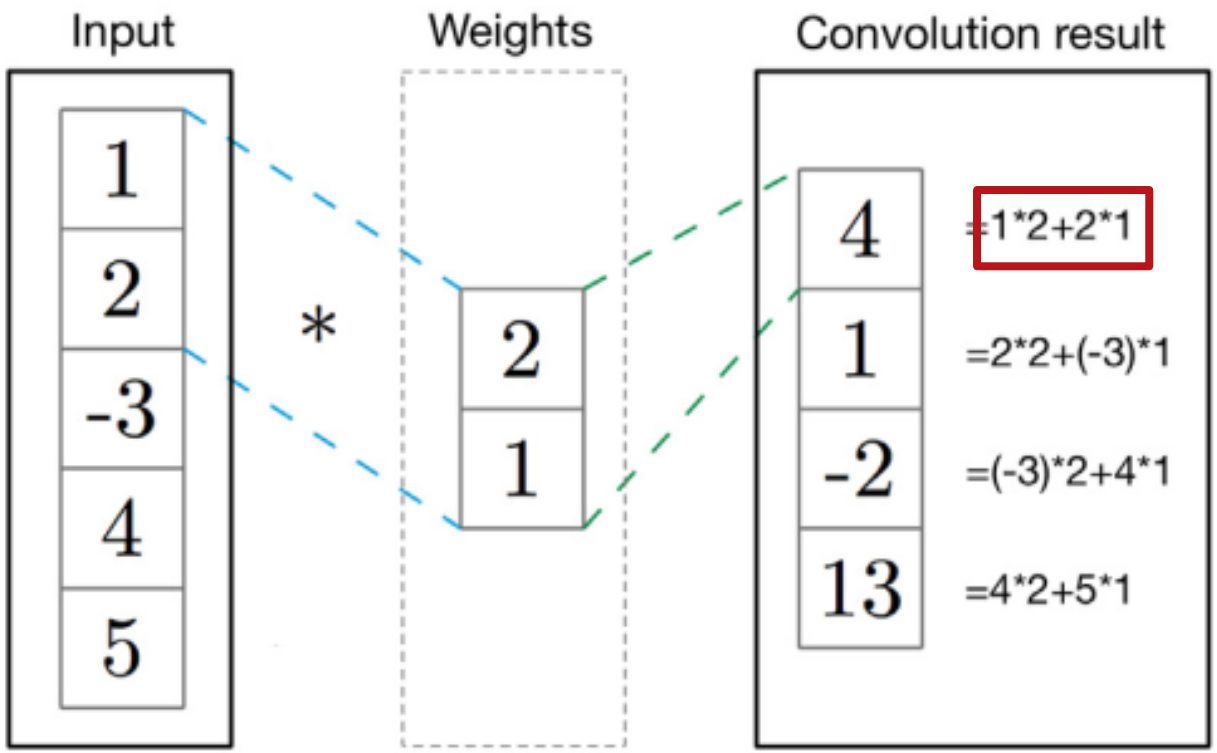
- Single filter
- Kernel size 2

Convolutional neural network



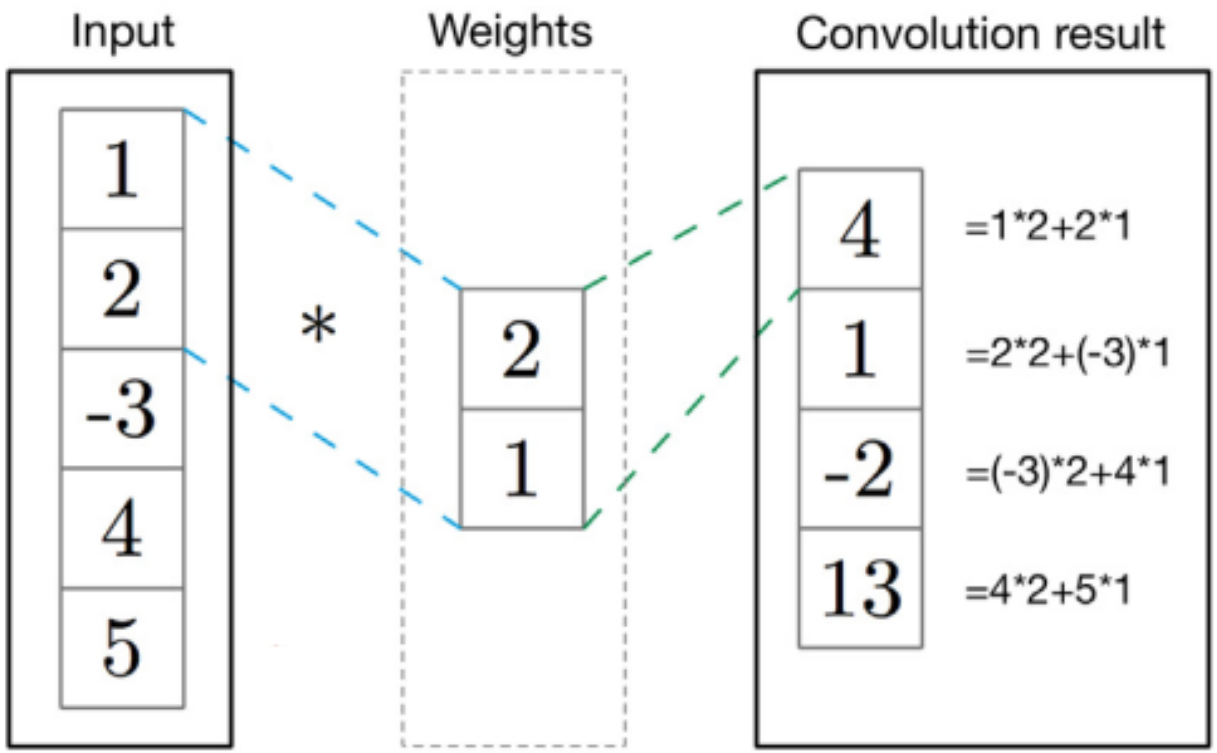
- Single filter
- Kernel size 2
- Stride of 1

Convolutional neural network



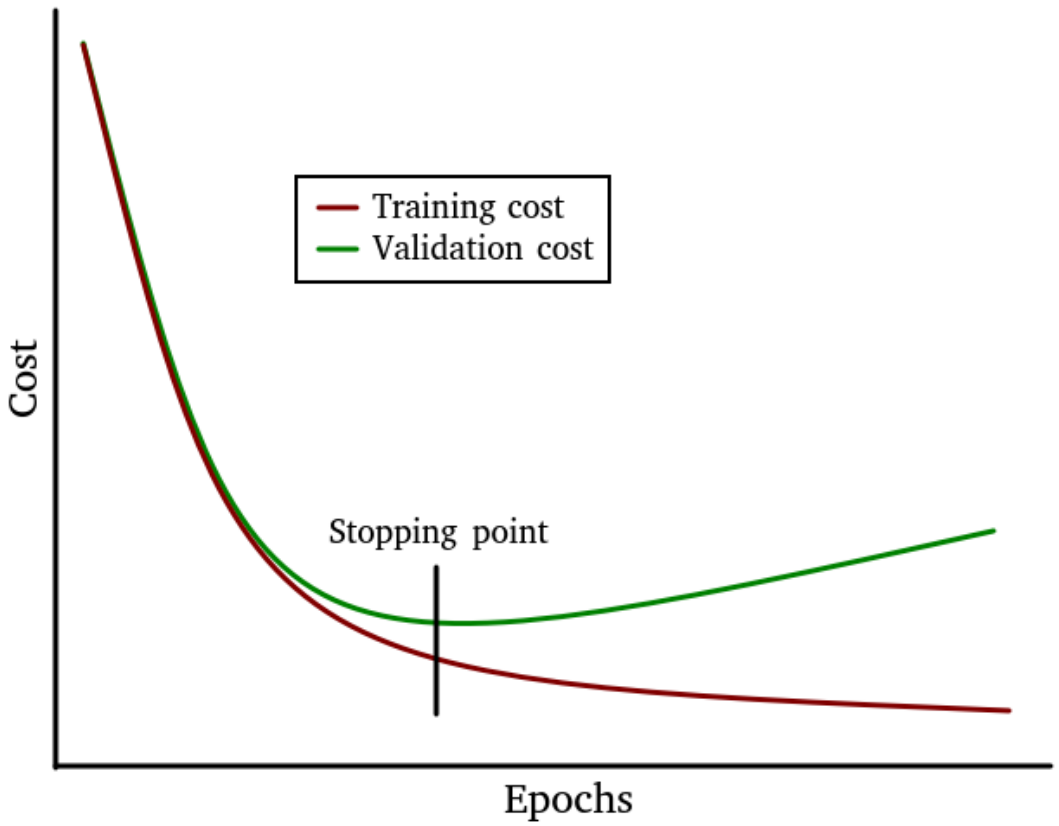
- Single filter
- Kernel size 2
- Stride of 1
- Inner matrix product

Convolutional neural network



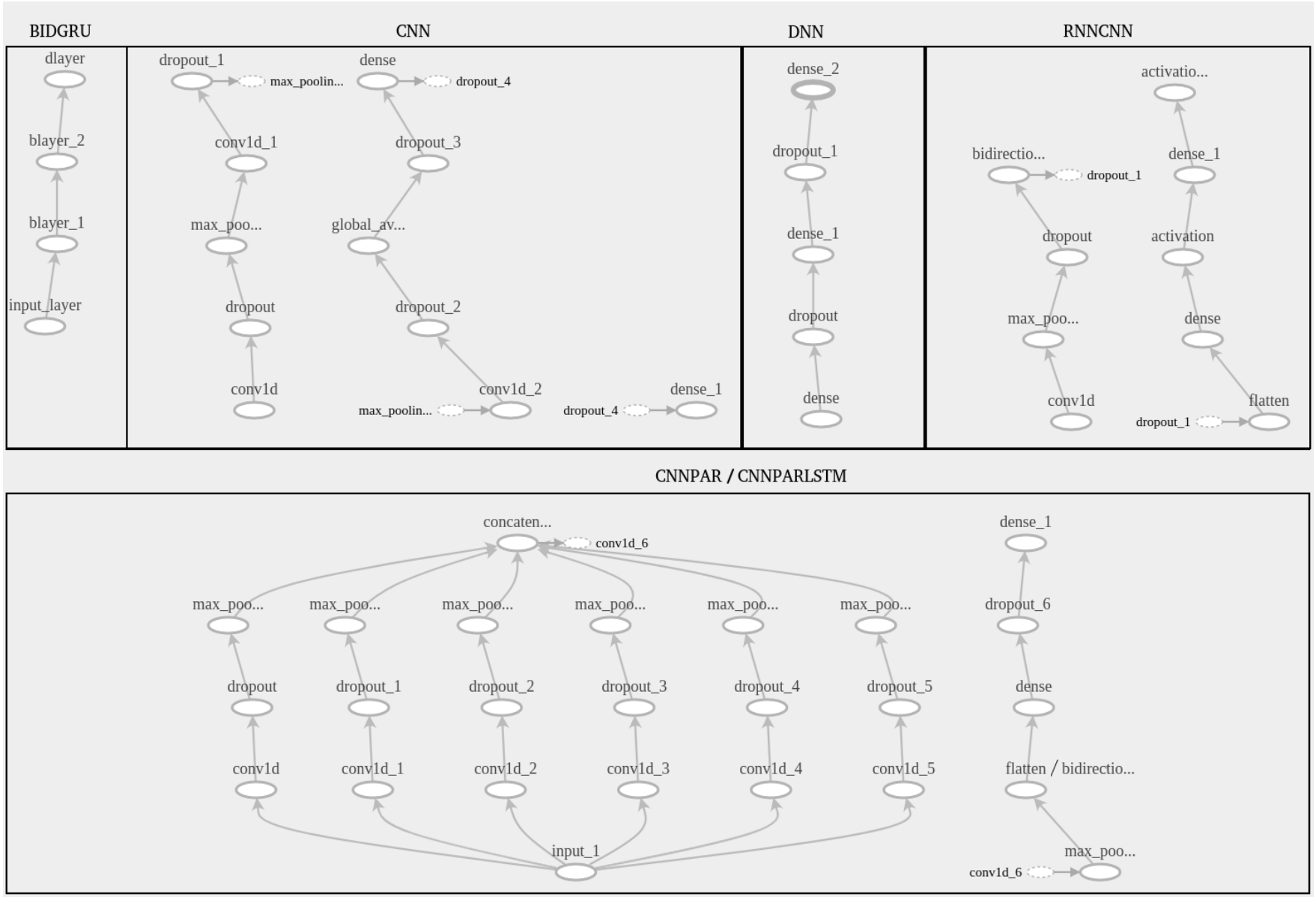
- Single filter
- Kernel size 2
- Stride of 1
- Inner matrix product
- “Compresses” information

Preventing overfitting



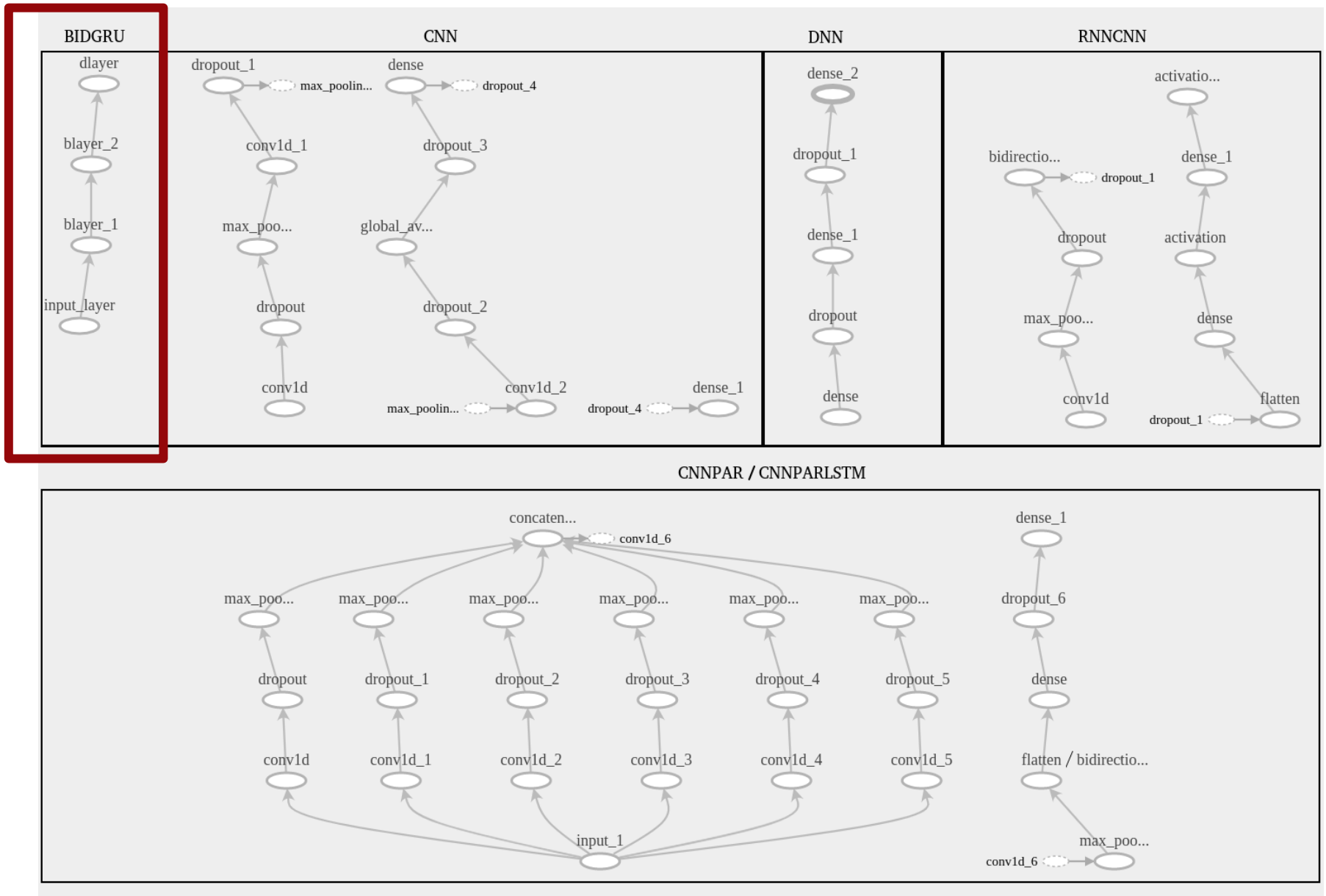
- Early stopping
- Dropout
- Regularization

NN architectures

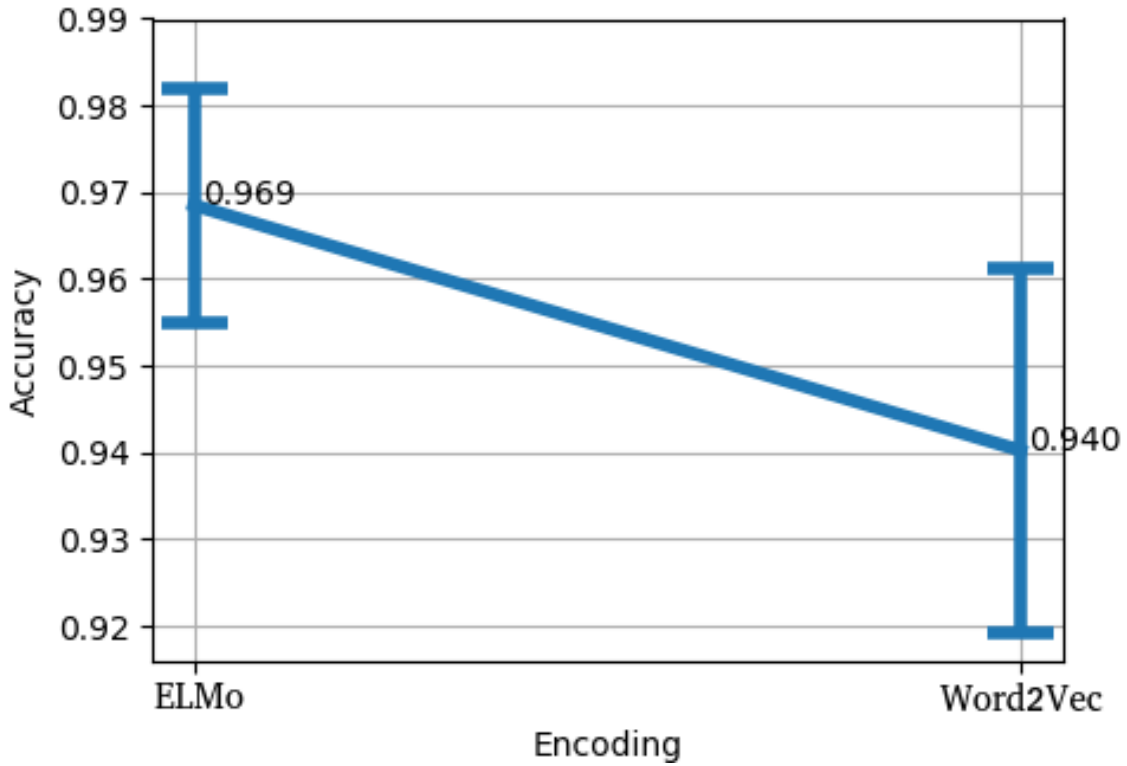


NN architectures

Hamid et al



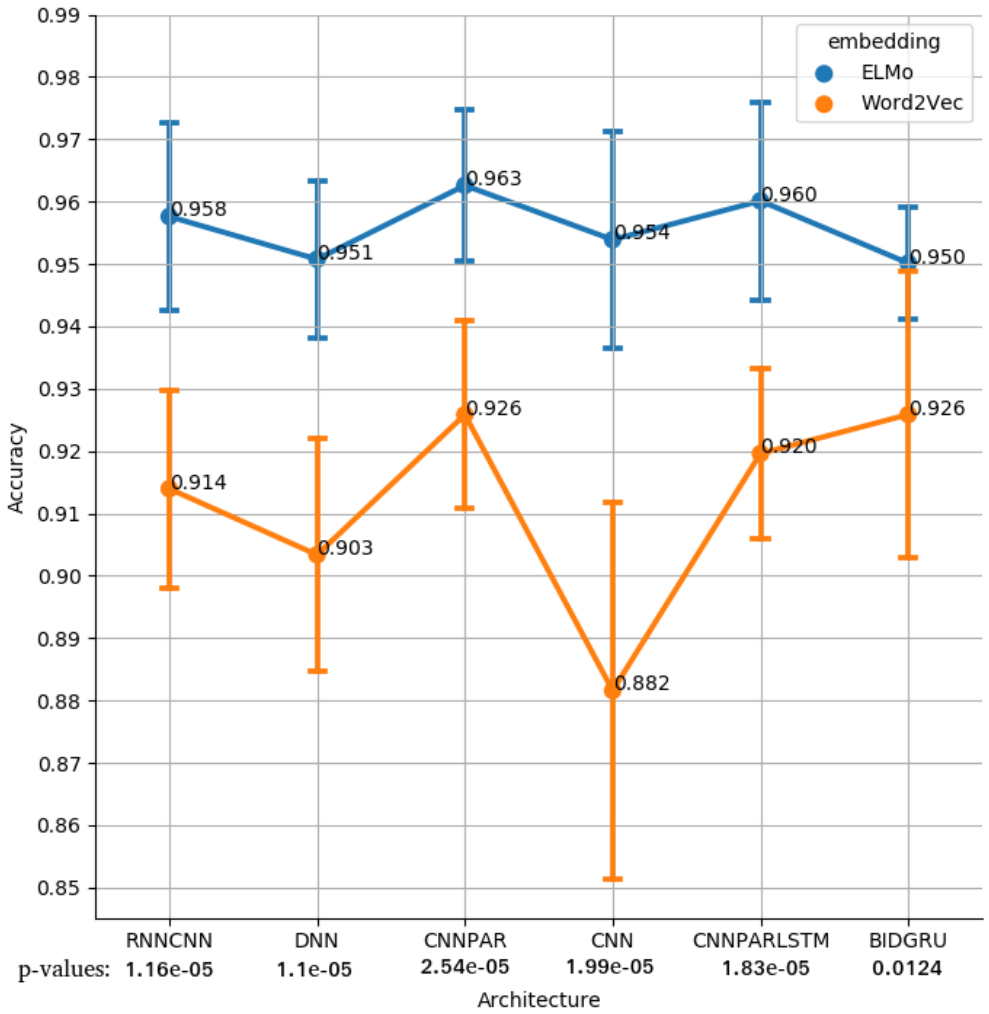
BIDGRU UniProt cross validation accuracy



Data sets		
Name	Source	# data points
GLY	UniProt GO term: "Glycolytic-process"; taxonomy: Bacteria; length: 14-99 amino acids	1954
LIP	UniProt GO term: "Lipid-A-biosynthetic-process"; taxonomy: Bacteria; length: 14-99 amino acids	1142
UNI	UniProt keywords: "not Antibiotic", "not Antimicrobial", "not Plasmid"; taxonomy: Bacteria; length: 10-359 amino acids	1003
BAC	CAMP (anything containing "bacteriocin"), BAGEL, Bactibase; length: 10-359 amino acids	1003

Training data					
Name	Positive	Negative	Pos. Neg. Ratio	# training points	# test points
UniProt	LIP	GLY	0.66:1	2476	620
Bacteriocin	BAC	UNI	1:1	1604	402

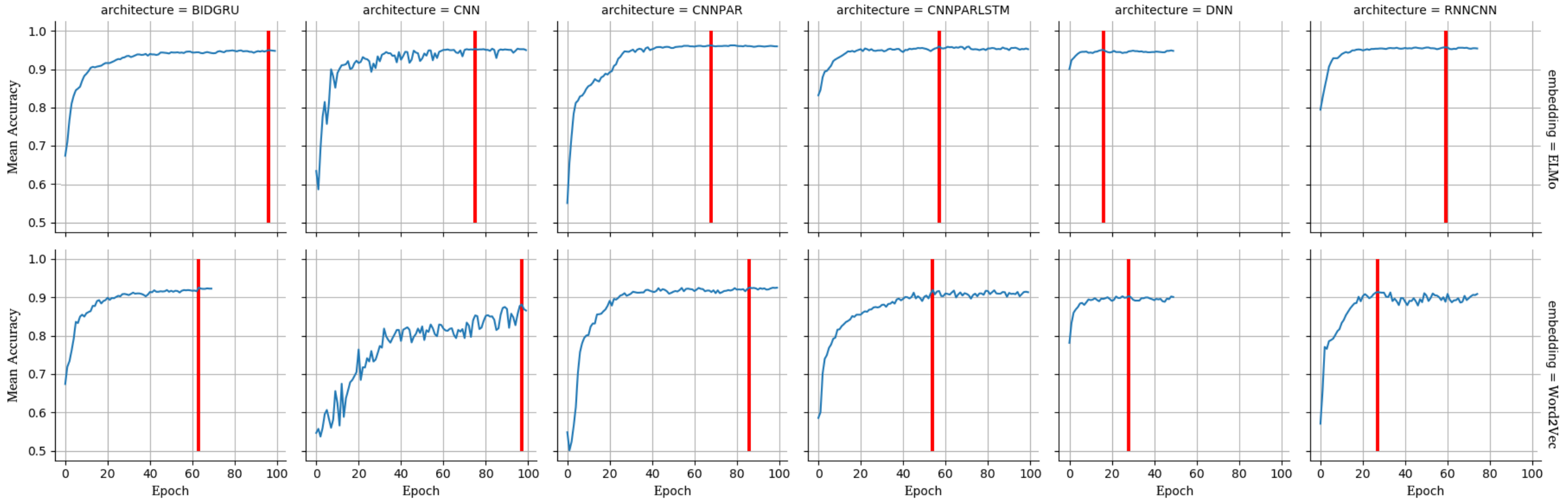
Bacteriocin data set cross validation accuracy



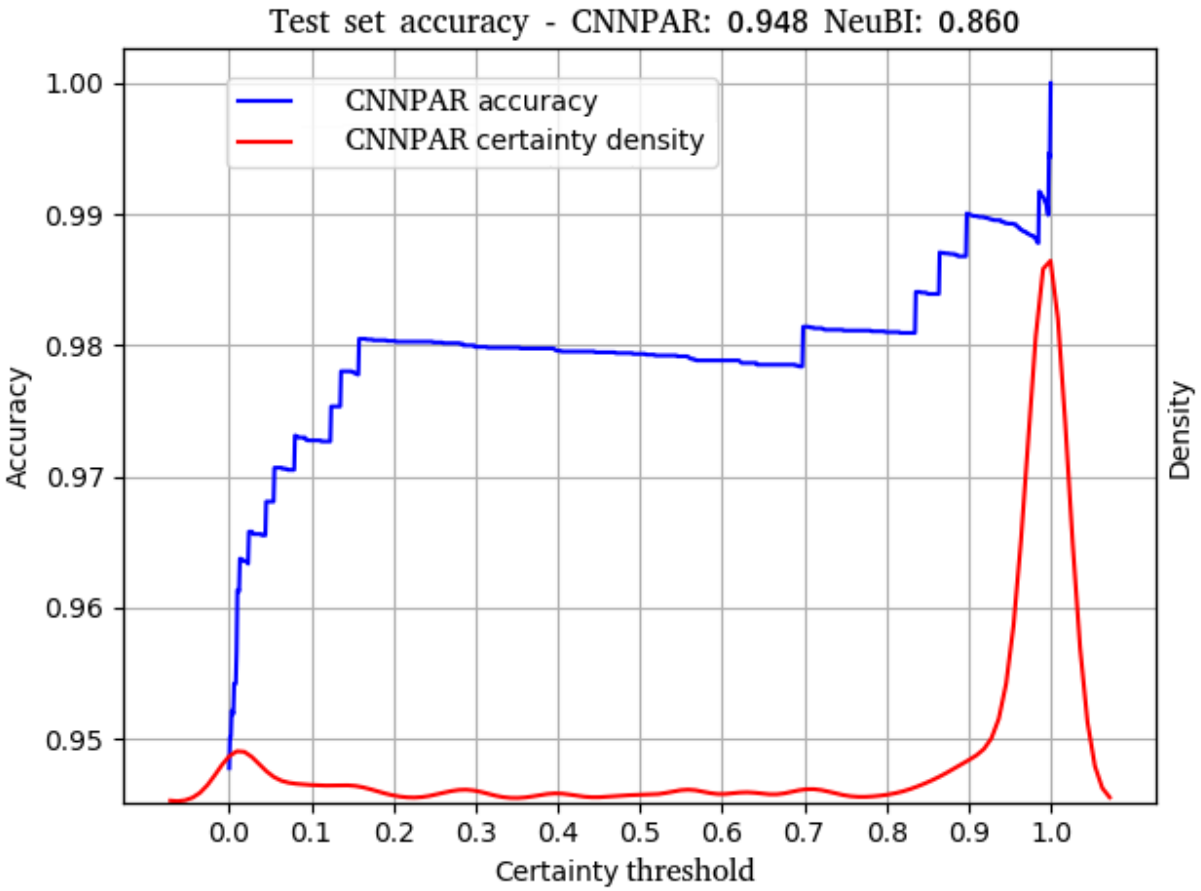
Data sets		
Name	Source	# data points
GLY	UniProt GO term: "Glycolytic-process"; taxonomy: Bacteria; length: 14-99 amino acids	1954
LIP	UniProt GO term: "Lipid-A-biosynthetic-process"; taxonomy: Bacteria; length: 14-99 amino acids	1142
UNI	UniProt keywords: "not Antibiotic", "not Antimicrobial", "not Plasmid"; taxonomy: Bacteria; length: 10-359 amino acids	1003
BAC	CAMP (anything containing "bacteriocin"), BAGEL, Bactibase; length: 10-359 amino acids	1003

Training data					
Name	Positive	Negative	Pos. Neg. Ratio	# training points	# test points
UniProt	LIP	GLY	0.66:1	2476	620
Bacteriocin	BAC	UNI	1:1	1604	402

Mean accuracy per epoch



CNNPAR bacteriocin data set test accuracy



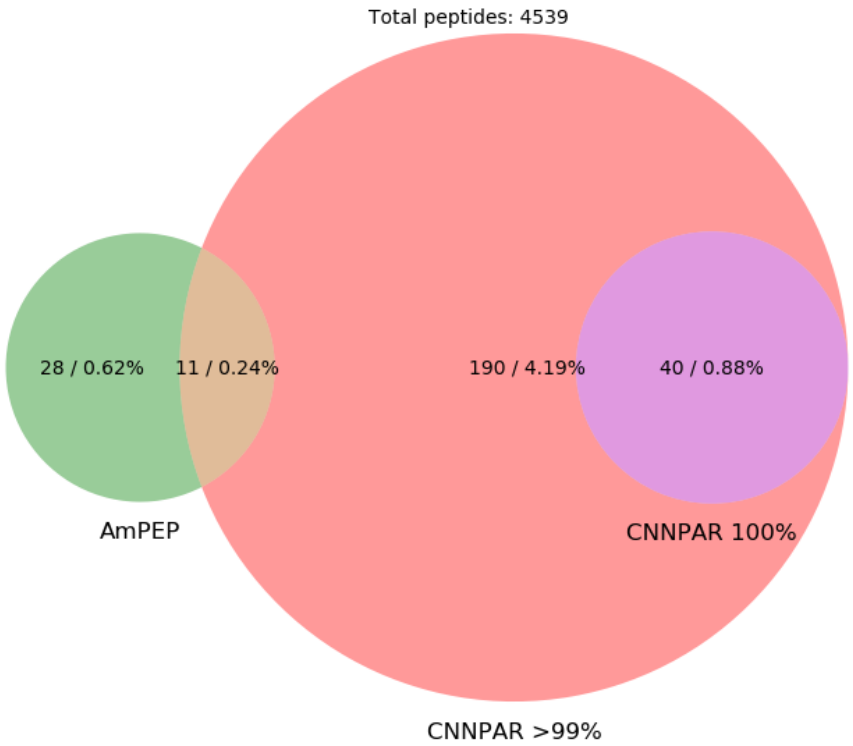
- CNNPAR accuracy of 94.8%
  - > 98% with certainty > 0.8
- NeuBI accuracy of 86.0%

Contents

Application on Sberro's small protein families

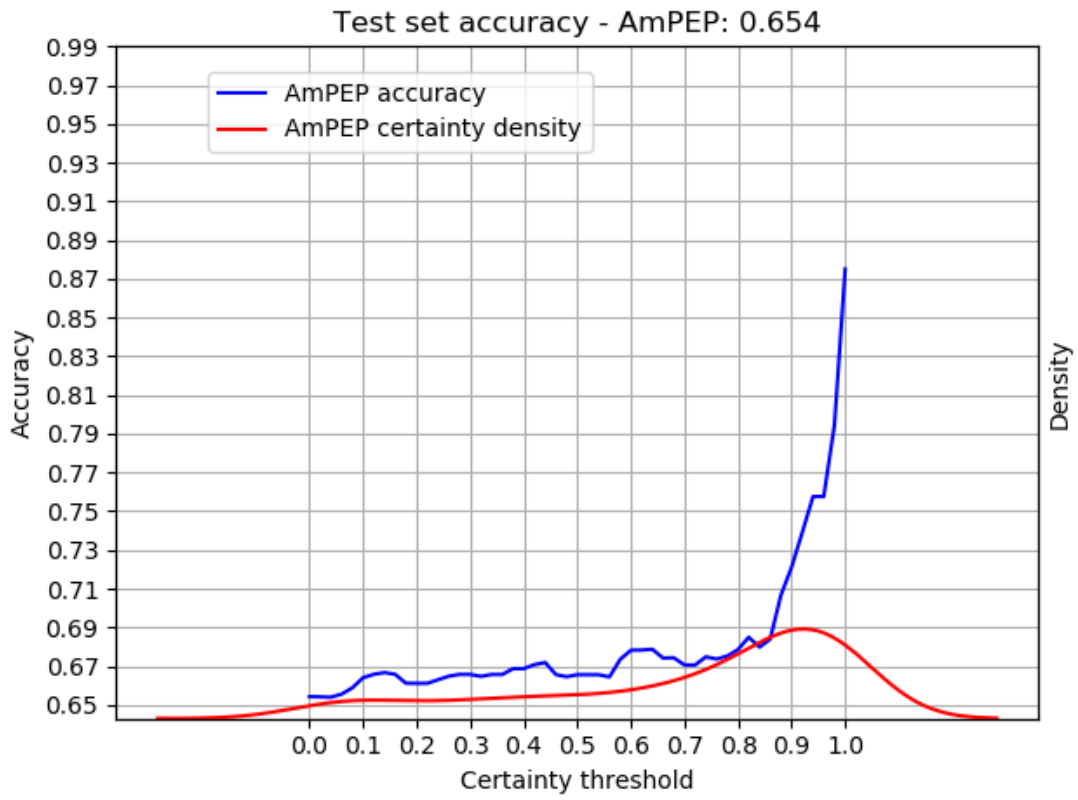
- Human microbiome samples
- ~4500 small protein families < 50 AAs long
- 39 AMPs identified with AmPEP

Application on Sberro's small protein families

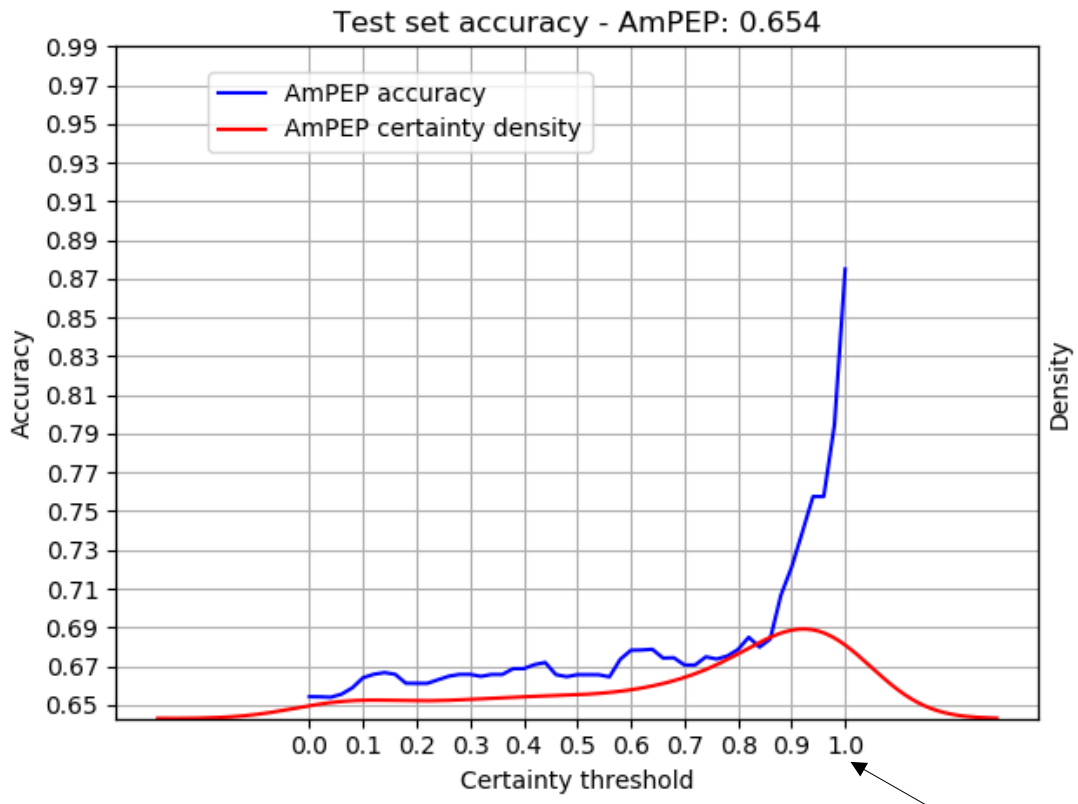


- More bacteriocins than AMPs
  - Broader class classifier harder to train?

AmPEP bacteriocin data set test accuracy

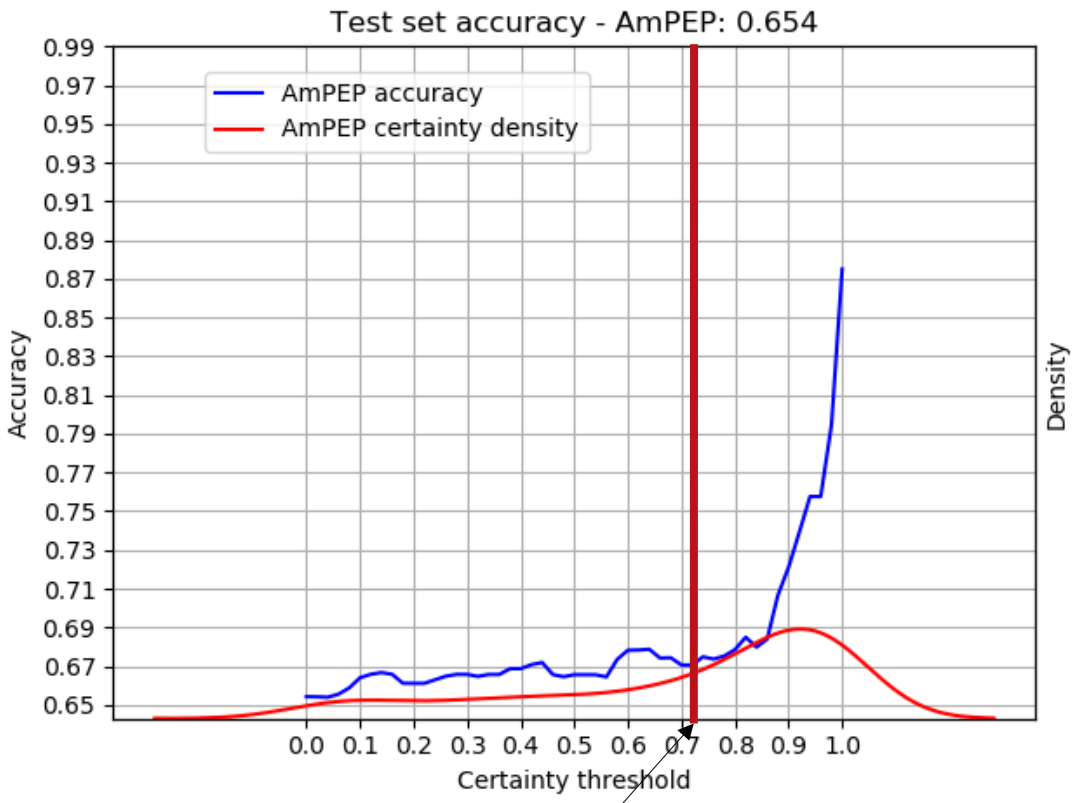


AmPEP bacteriocin data set test accuracy



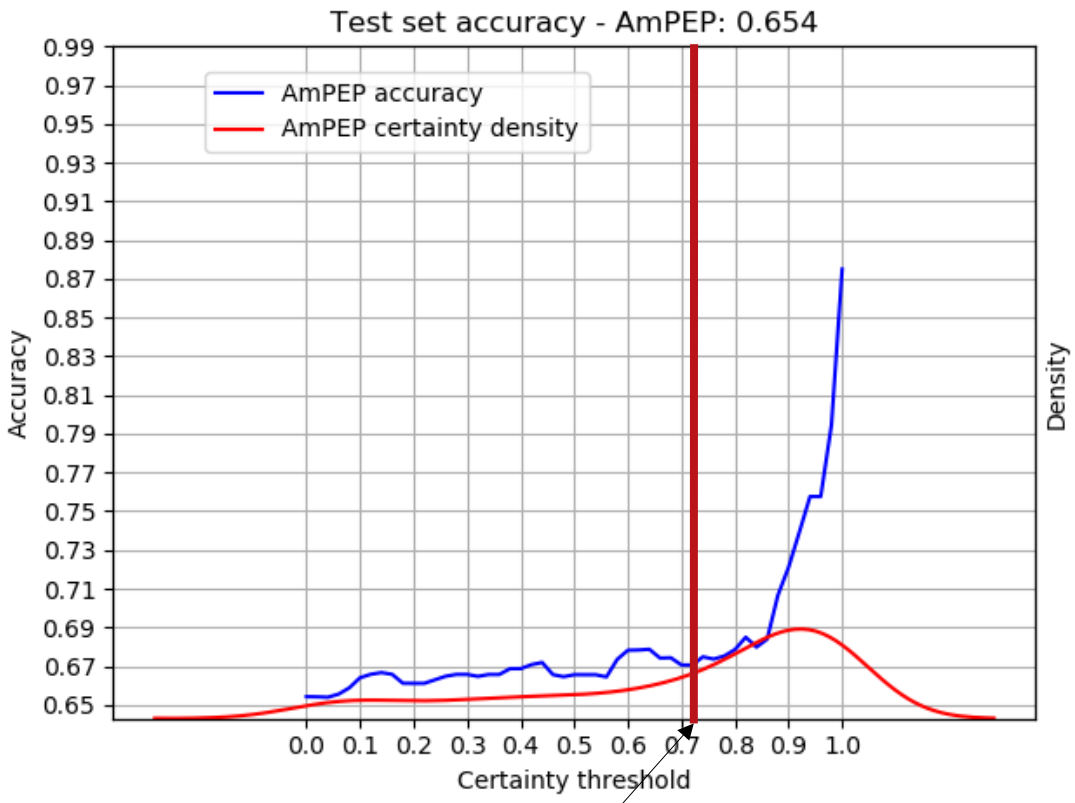
Normalized certainty

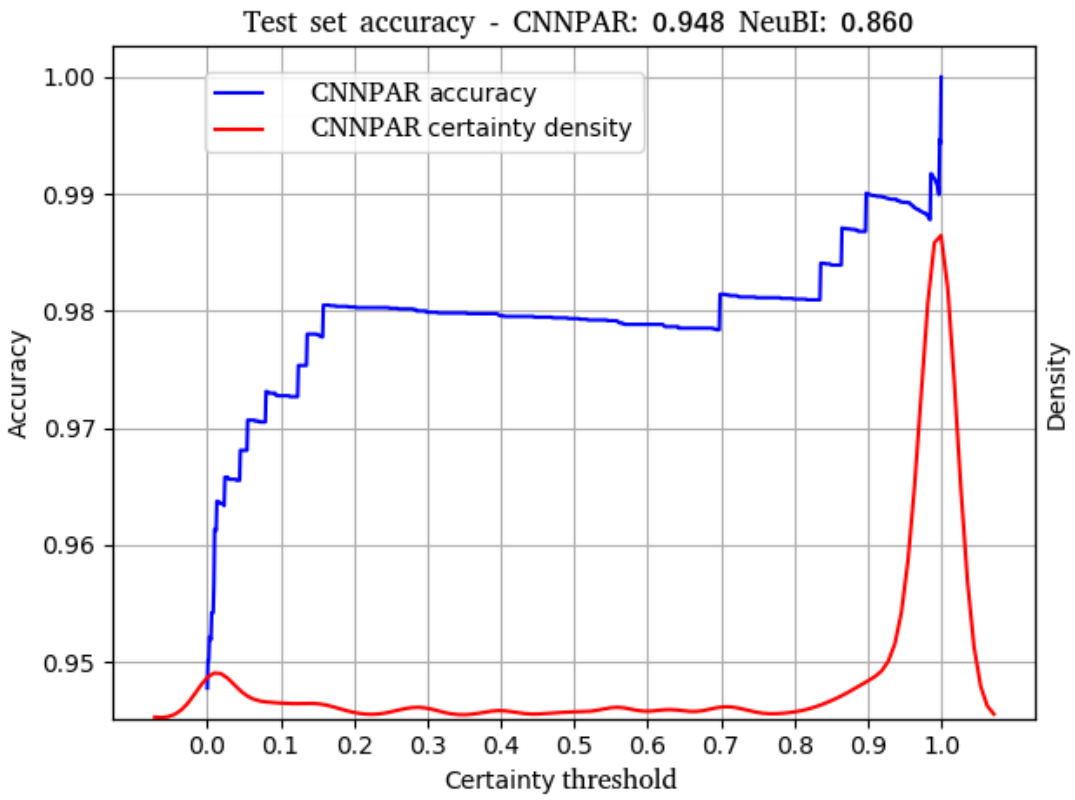
AmPEP bacteriocin data set test accuracy



Maximum certainty observed in Sberro

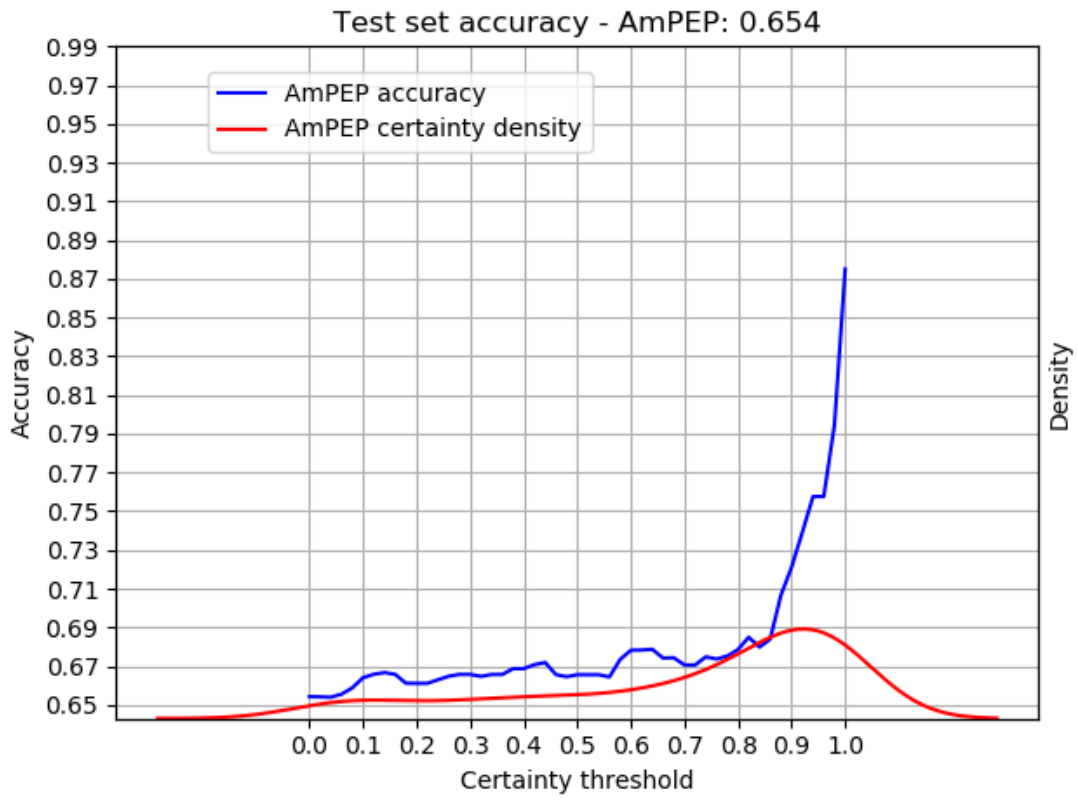
AmPEP bacteriocin data set test accuracy





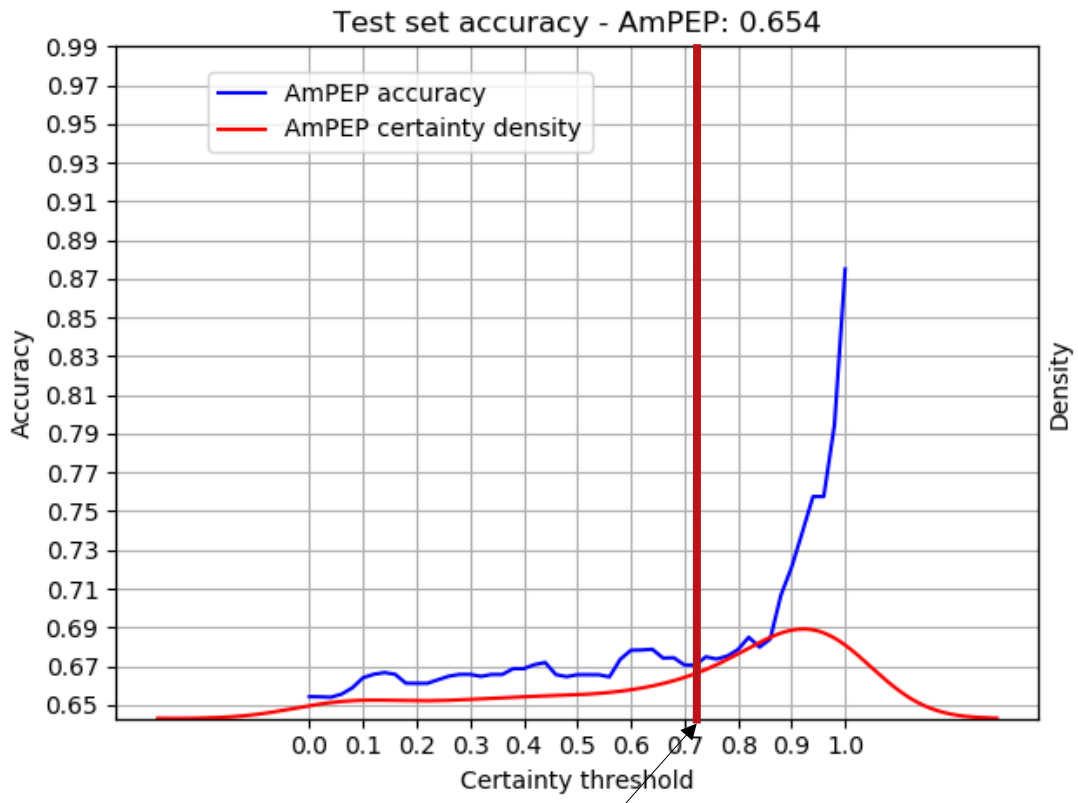
Maximum certainty observed in Sberro

AmPEP bacteriocin data set test accuracy



- AmPEP < ~67% accuracy
- Conclusion: AmPEP is not good for identifying bacteriocins

AmPEP bacteriocin data set test accuracy



Maximum certainty observed in Sberro

- AmPEP < ~67% accuracy
- Conclusion: AmPEP is not good for identifying bacteriocins

Main findings

- Increased accuracy with ELMo embedding.
  - Hamid et al. test accuracy 86.0%.
  - Our classifier test accuracy 94.8%.
- Found 40 putative bacteriocins.

The End

Distribution of ribosomal binding site fractions