National Institute of Allergy and Infectious Diseases

2019 Health Innovations Conference

Using deep learning to improve antimicrobial peptide recognition

Tuesday, 19 March 2019

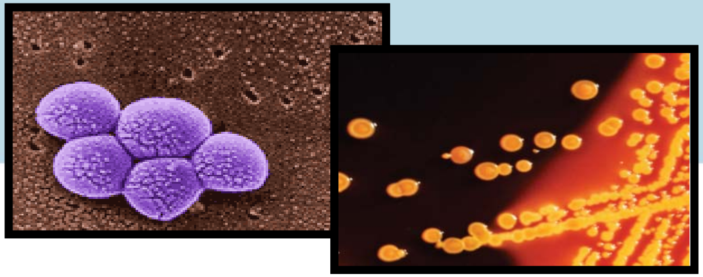
NIAID

National Institute of Allergy and Infectious Diseases

Daniel Veltri, Ph.D.

Motivation

- Reports of antibiotic resistance have *increased* despite a *slowdown* in new antibiotics coming to market in recent decades,
- The U.S. Center for Disease Control reports over 2 million infections and 23,000 deaths each year due to antibiotic-resistant bacteria and fungi in the U.S. [1],
- The WHO has put out numerous reports warning of the risks of resistant bacteria to hospitals around the world.



Methicillin-resistant *Staphylococcus aureus* (left)
Carbapenem-resistant Enterobacteriaceae (right)

The evolving threat
of antimicrobial resistance

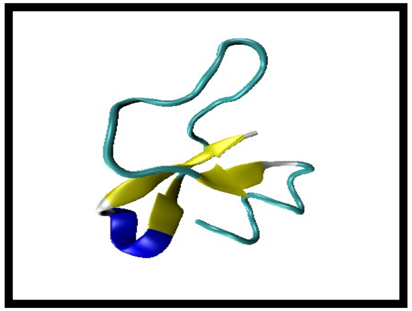
Options for action

Executive Summary

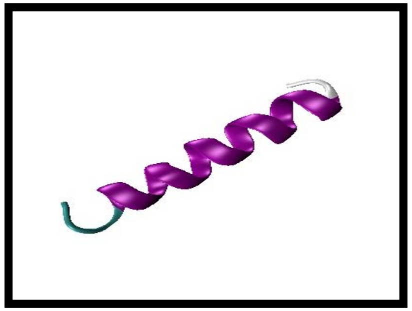
World Health
Organization

[1] CDC: https://www.cdc.gov/drugresistance/about.html

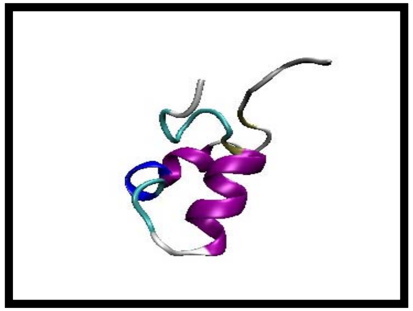
Antimicrobial Peptides (AMPs)



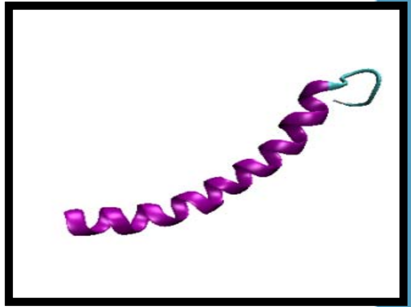
Beta Defensin 1
Homo sapiens
PDB: 1IJV



Magainin 2
Xenopus laevis
PDB: 2MAG



Aurelin
Aurelia aurita
PDB: 2LG4



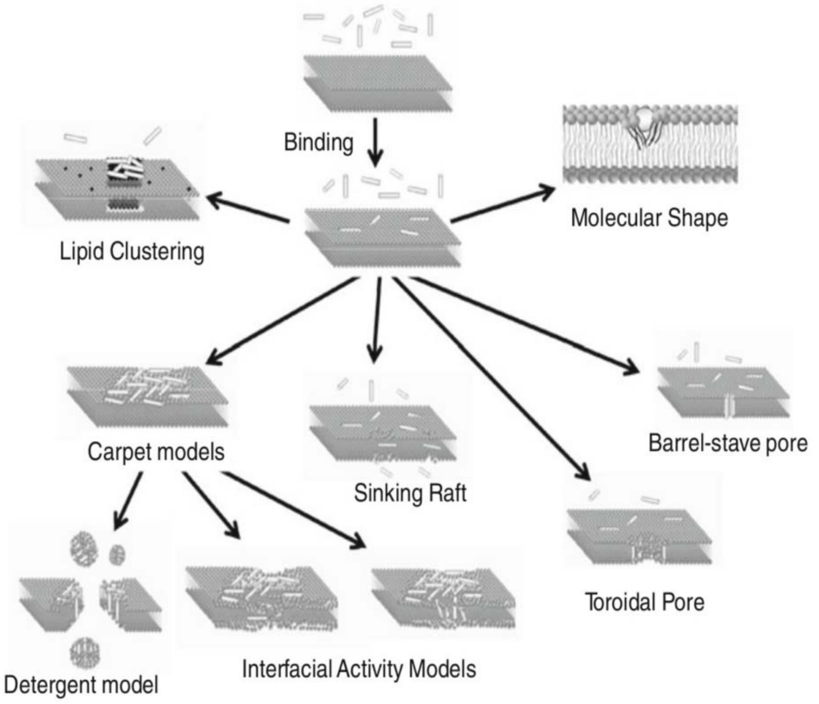
Cathelicidin LL-37
Homo sapiens
PDB: 2K6O

- One promising area for new antibiotic research has been natural AMPs- short peptides with innate antibacterial activity found across all phyla,
- To date, efforts to design and/or modify AMPs have had limited success in delivering new drugs to market.

NIH National Institute of Allergy and Infectious Diseases

AMPs are Complicated!

- Amino acid (AA) physicochemical properties are important for AMP activity (*charge, hydrophobic etc.*),
- AMPs are highly diverse, both in sequence and killing mechanism,
- We still do not know *exactly* how physicochemica properties relate to
  AMP activity- knowledge needed
  to *guide AMP design.*



Some proposed AMP attack mechanisms

Figure Source: Wimley (2011) *J. of Mem. Bio.* 239(1):27-34.

Prior AMP Classification Work

- Most work to date has focused on AMP recognition- taking query peptide sequences and assigning *AMP* or *non-AMP* labels,
- Top techniques report accuracies in the high 80 to mid 90% range,
- Approaches often pair physicochemical properties with sliding window averages or machine learning algorithms like artificial neural networks (ANN), support vector machines (SVM), etc.,
- A major issue in the field is that *few groups make their code or complete data sets available*. This makes it difficult to perform reliable comparisons as a “gold standard” benchmark data set is not currently available.

National Institute of
Allergy and
Infectious Diseases

Prior AMP Classification Performance

Group	Algorithm	Performance
	ANN	[illegible]
Xiao et al. (2013)	Fuzzy K-Nearest Neighbor	0.84
Meher et al. (2017)	SVM	0.84

Matthew’s Correlation Coefficient (MCC):

$$MCC = \frac{(TP \text{ x } TN) - (FN \text{ x } FP)}{\sqrt{(TP + FN)\text{x}(TN + FP)\text{x}(TP + FP)\text{x}(TN + FN)}}$$

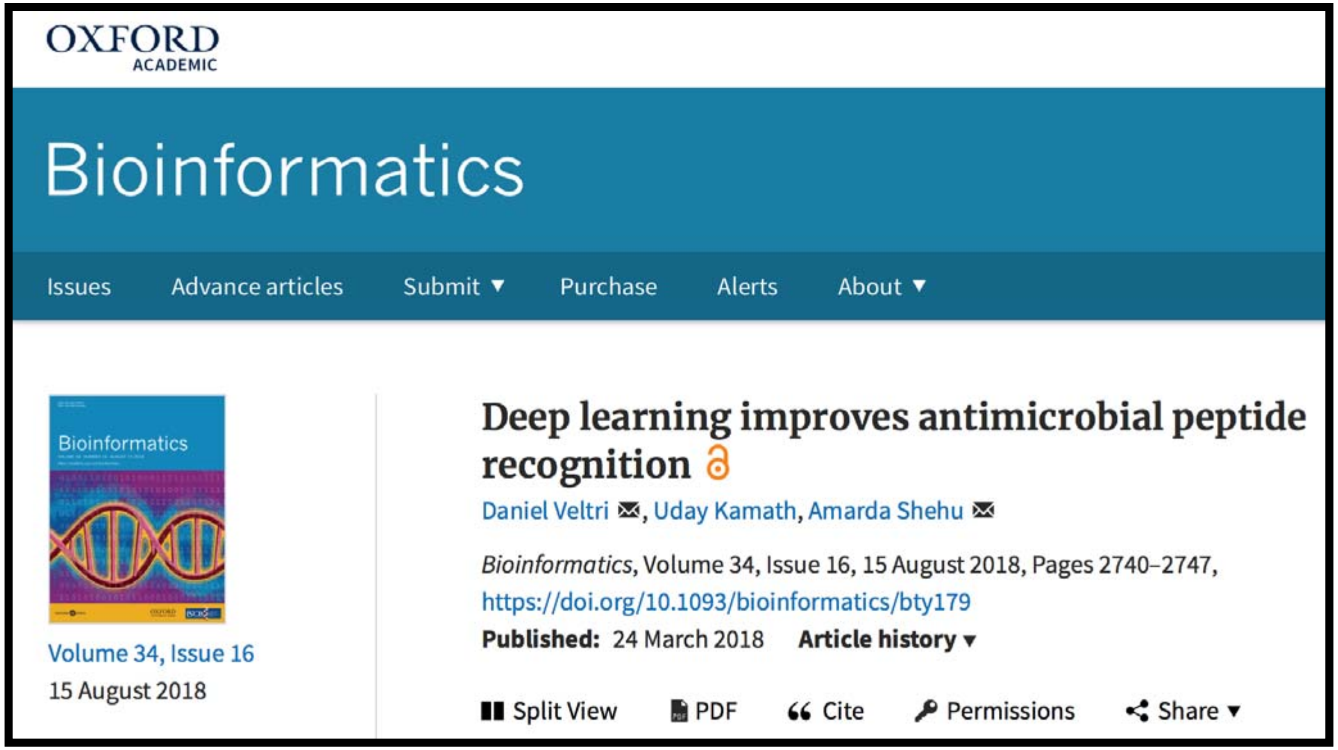
MCC values range from -1 to 1, with 1 denoting perfect classification performance.

TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative

National Institute of Allergy and Infectious Diseases

† “Random Forest” is trademarked and licensed to Salford Systems (San Diego, CA)

Using Deep Learning for AMP Classification



Bioinformatics

Deep learning improves antimicrobial peptide recognition

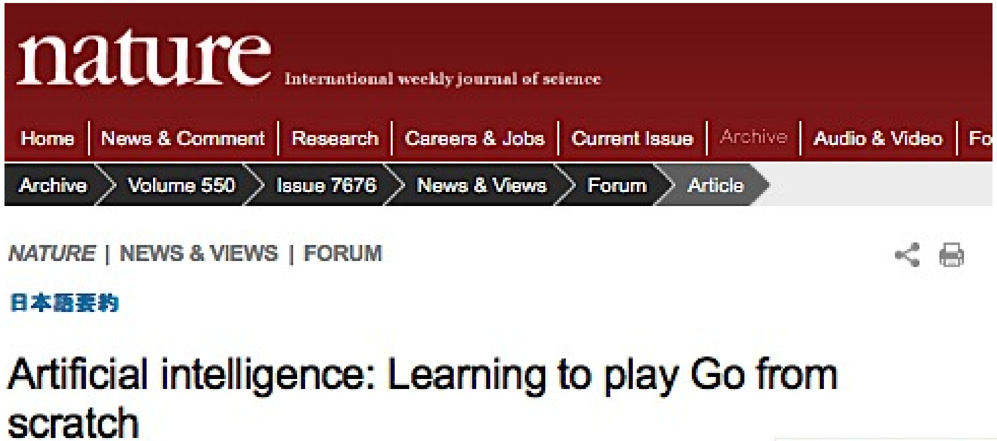
Daniel Veltri, Uday Kamath, Amarda Shehu

Bioinformatics, Volume 34, Issue 16, 15 August 2018, Pages 2740–2747, https://doi.org/10.1093/bioinformatics/bty179

Published: 24 March 2018 **Article history** ▾

Volume 34, Issue 16
15 August 2018

Deep Neural Networks (DNN) in the News...



nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | Fo
Archive > Volume 550 > Issue 7676 > News & Views > Forum > Article

NATURE | NEWS & VIEWS | FORUM

日本語要約

Artificial intelligence: Learning to play Go from scratch

Satinder Singh, Andy Okun & Andrew Jackson

Affiliations | Corresponding authors

Nature **550**, 336–337 (19 October 2017) | doi:10.1038/550336a
Published online 18 October 2017

In March 2016, the artificial-intelligence program AlphaGo defeated a world Go champion, Lee Sedol.

Lee Jin-Man/AP/Rex/Shutterstock

Go board image from Wikimedia Foundation

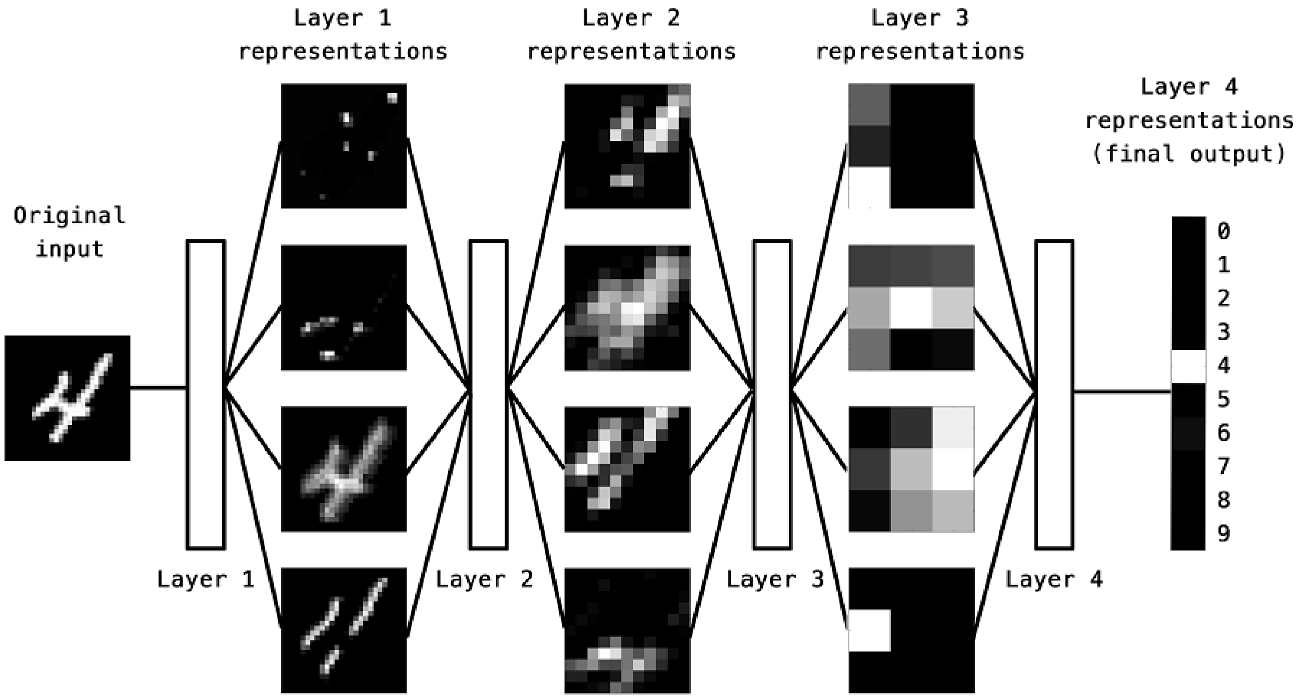
National Institute of Allergy and Infectious Diseases

Deep Learning Packages

So many flavors to choose from…

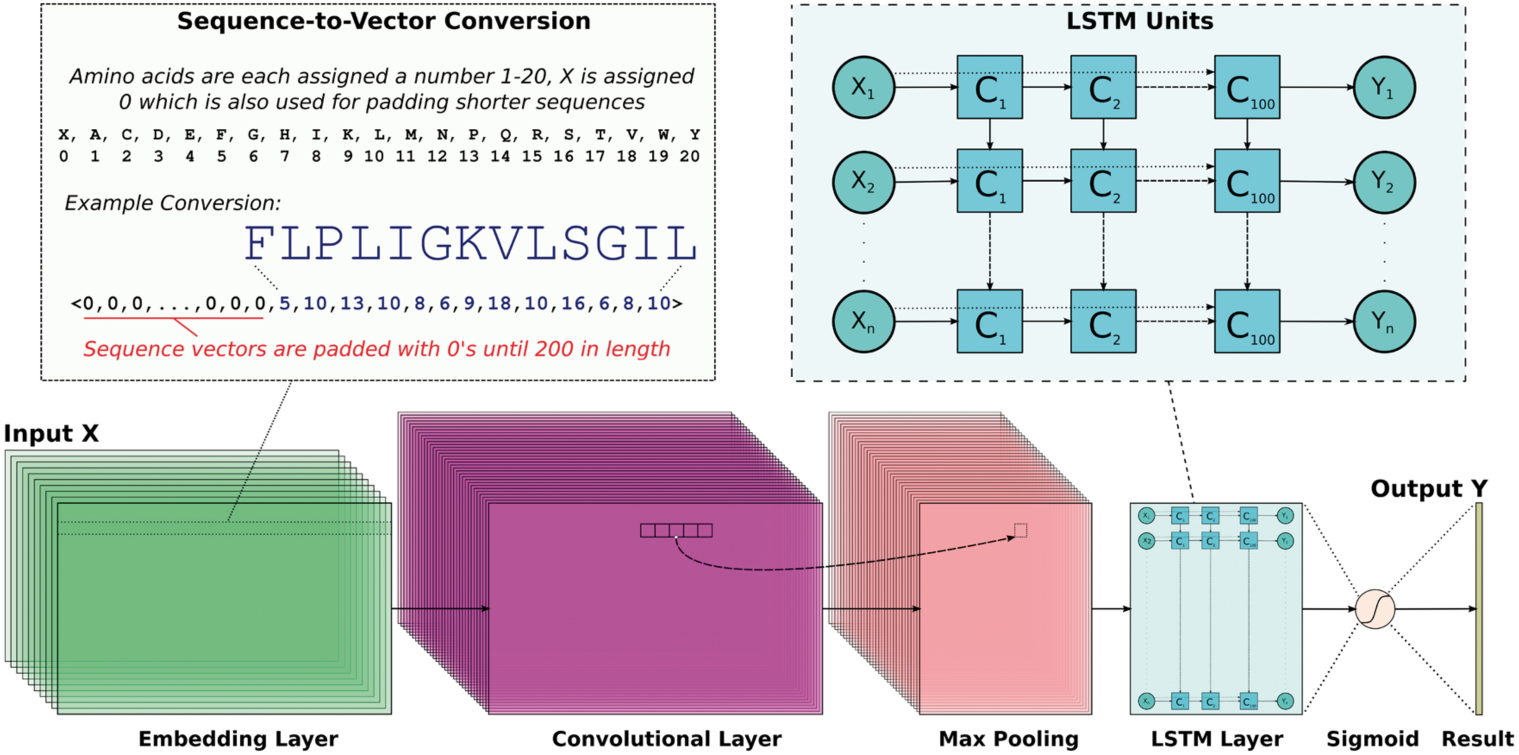
K Keras
TensorFlow
Caffe
opennn neural networks
Microsoft Cognitive Toolkit
aka.ms/CognitiveToolkit
Microsoft
torch
theano
H_2O

Deep Neural Networks Have Multiple Layers



Source: Chollet and Allaire “Deep Learning with R” pp.9, 2018.

Our Model Architecture



Sequence-to-Vector Conversion

Amino acids are each assigned a number 1-20, X is assigned 0 which is also used for padding shorter sequences

X	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Example Conversion:

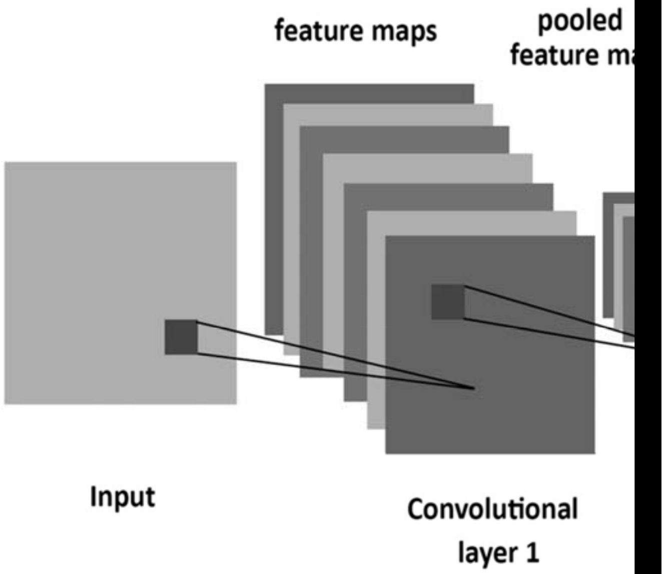
FLPLIGKVLSGIL

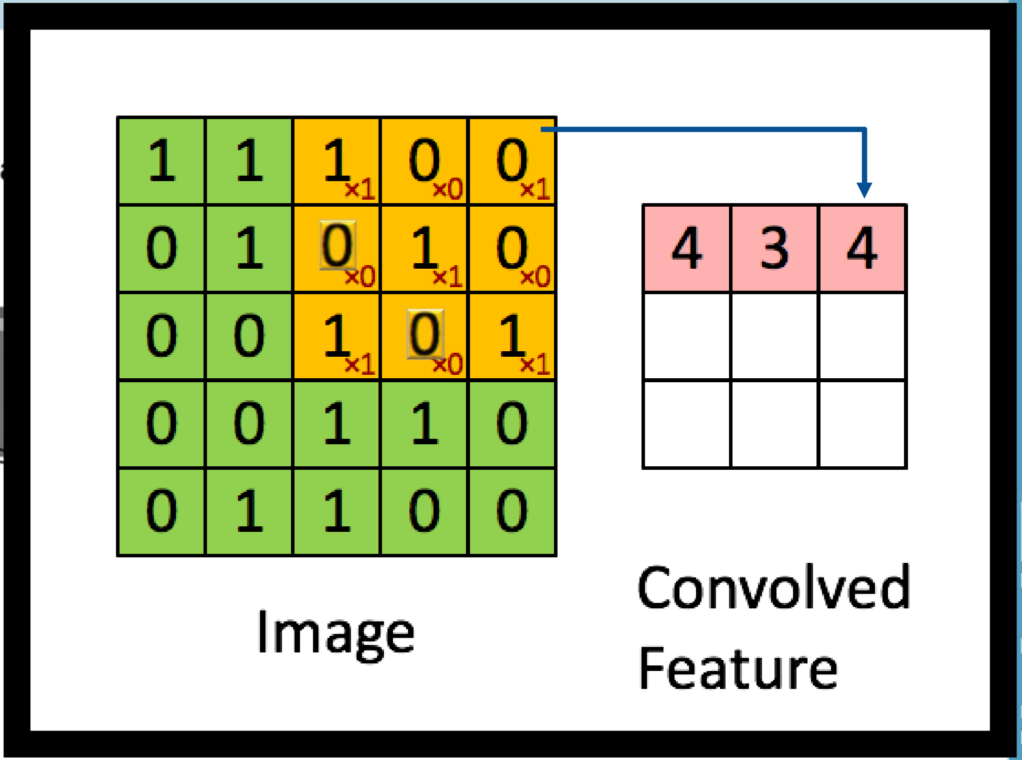
<0,0,0,...,0,0,0,5,10,13,10,8,6,9,18,10,16,6,8,10>

Sequence vectors are padded with 0's until 200 in length

National Institute of Allergy and Infectious Diseases

Convolutional Layers





1	1	$1_{\times1}$	$0_{\times0}$	$0_{\times1}$
0	1	$0_{\times0}$	$1_{\times1}$	$0_{\times0}$
0	0	$1_{\times1}$	$0_{\times0}$	$1_{\times1}$
0	0	1	1	0
0	1	1	0	0

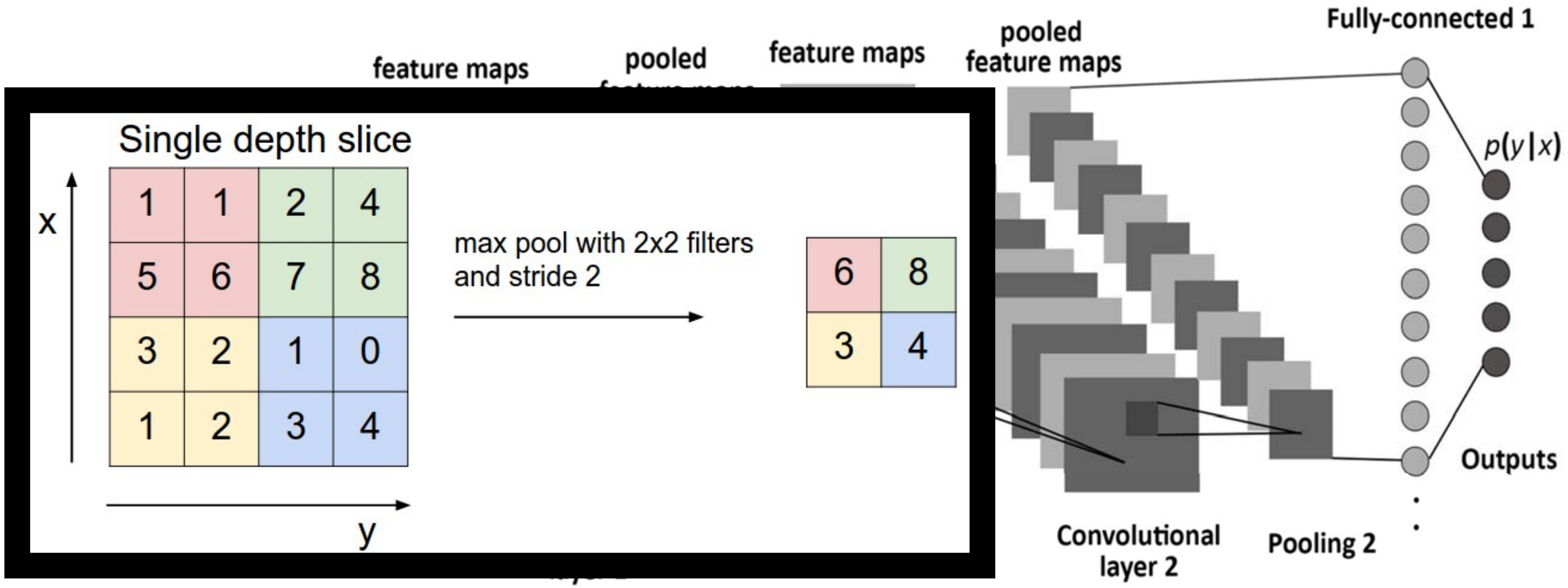
Image

4	3	4

Convolved Feature

Figure from: towardsdatascience.com

Pooling Layer

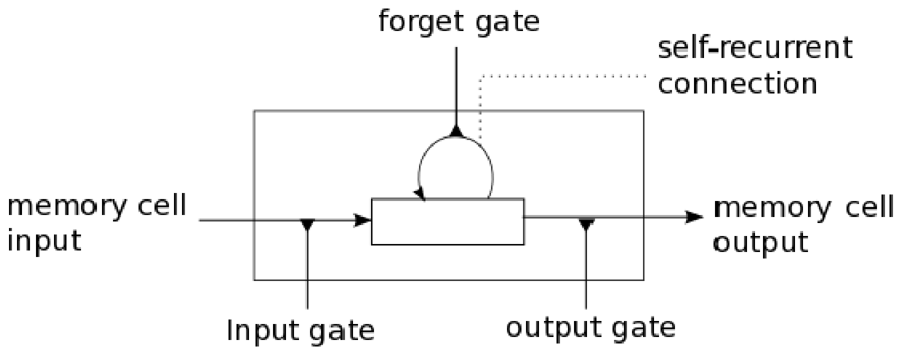


1	1	2	4
5	6	7	8
3	2	1	0
1	2	3	4

6	8
3	4

Figure from: towardsdatascience.com

Long Short-Term Memory (LSTM)



Direction of Reading →

Ignore!

. . . TCCGCGATCGTTCGGGTGGCCTTTAATATTATGTGCGCGTTAGCTGGTCACGCG

Recognize Pattern!

Original LSTM Paper: Hochreiter and Schmidhuber (1997) Long short-term memory.

Figure: *deeplearning.net*

Data Set Construction

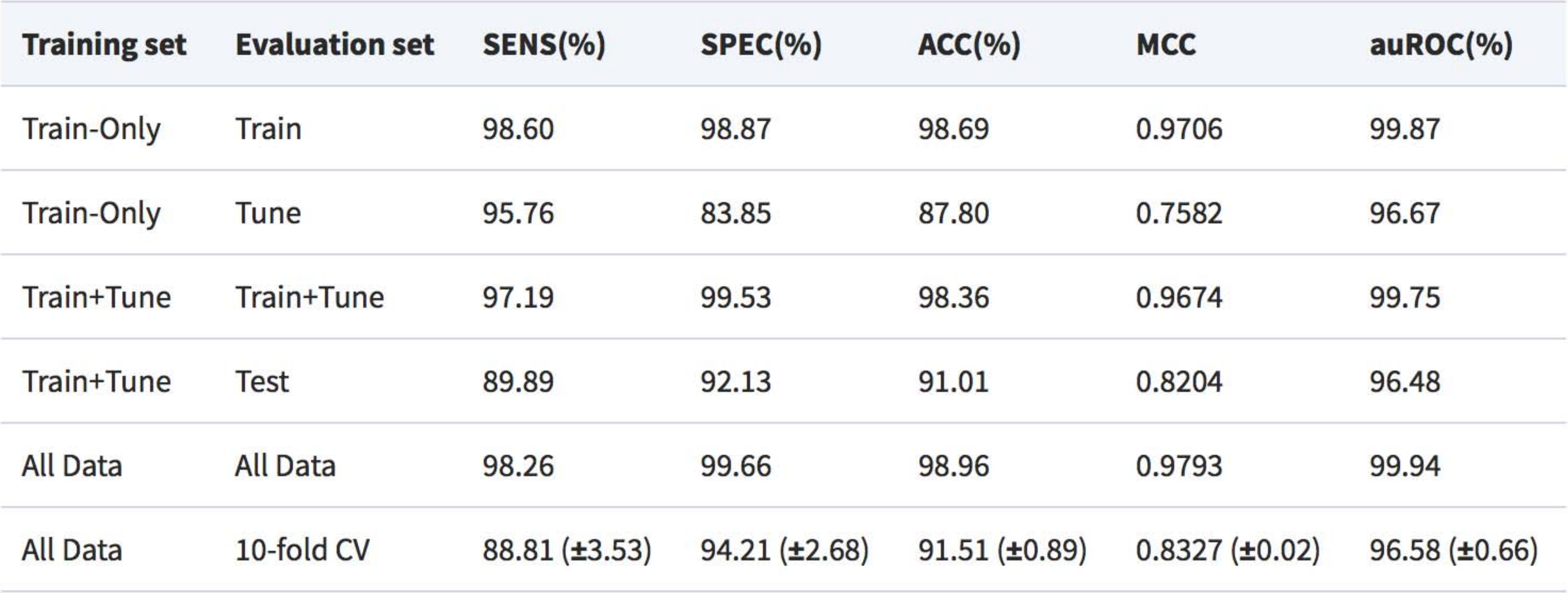
- AMPs were taken from the Antimicrobial Peptide Database vr3 (aps.unmc.edu/AP). Removed any <10 AA in length or sharing ≥90% sequence identity,

- Non-AMPs taken from UniProt using keyword filtering. Removed any <10 AA in length or sharing ≥40% sequence identity,

- Randomly selected even number of AMPs and Non-AMPs for each partition: 712 Training, 354 Tuning, and 712 Testing.

National Institute of
Allergy and
Infectious Diseases

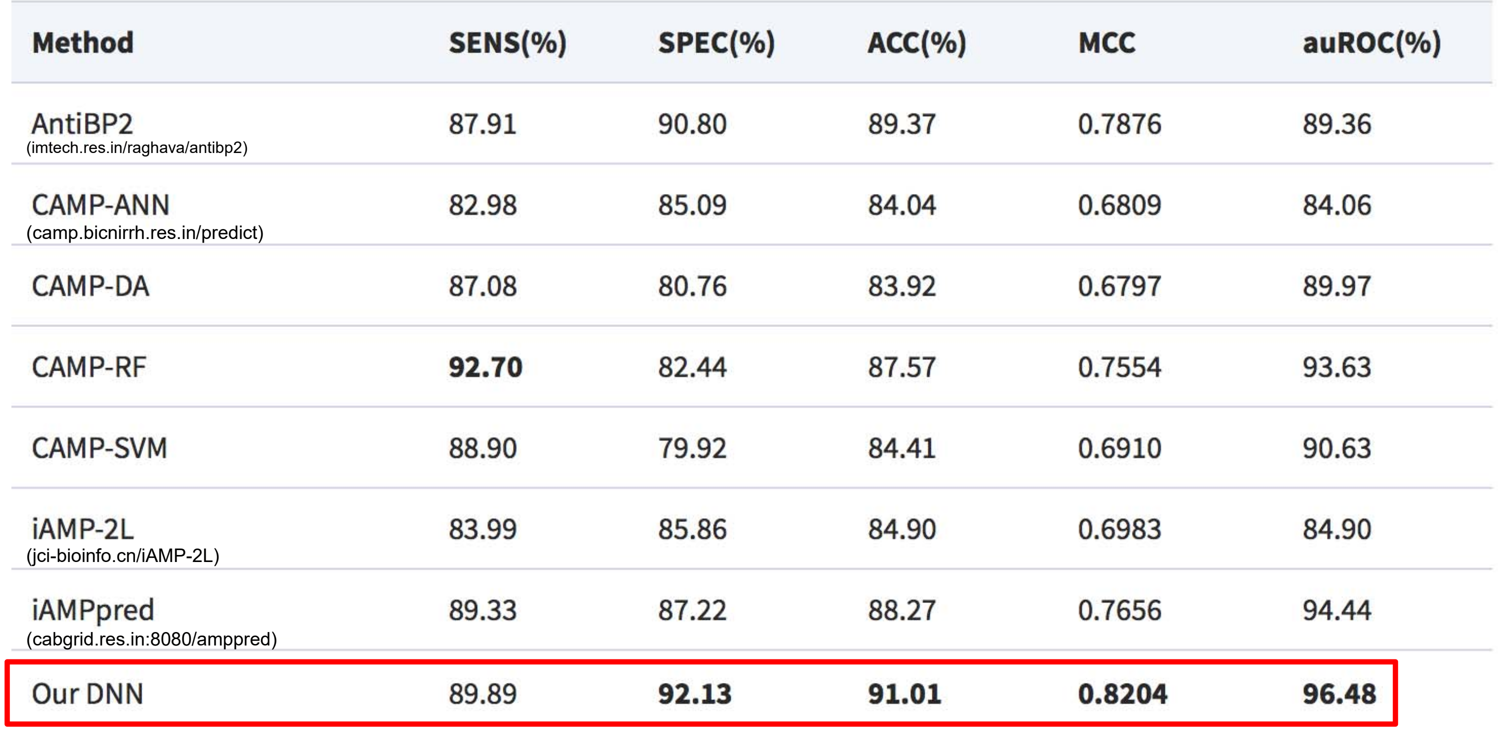
Model Training and Testing Performance



Training set	Evaluation set	SENS(%)	SPEC(%)	ACC(%)	MCC	auROC(%)
Train-Only	Train	98.60	98.87	98.69	0.9706	99.87
Train-Only	Tune	95.76	83.85	87.80	0.7582	96.67
Train+Tune	Train+Tune	97.19	99.53	98.36	0.9674	99.75
Train+Tune	Test	89.89	92.13	91.01	0.8204	96.48
All Data	All Data	98.26	99.66	98.96	0.9793	99.94
All Data	10-fold CV	88.81 (±3.53)	94.21 (±2.68)	91.51 (±0.89)	0.8327 (±0.02)	96.58 (±0.66)

A Head-to-Head AMP Server Comparison

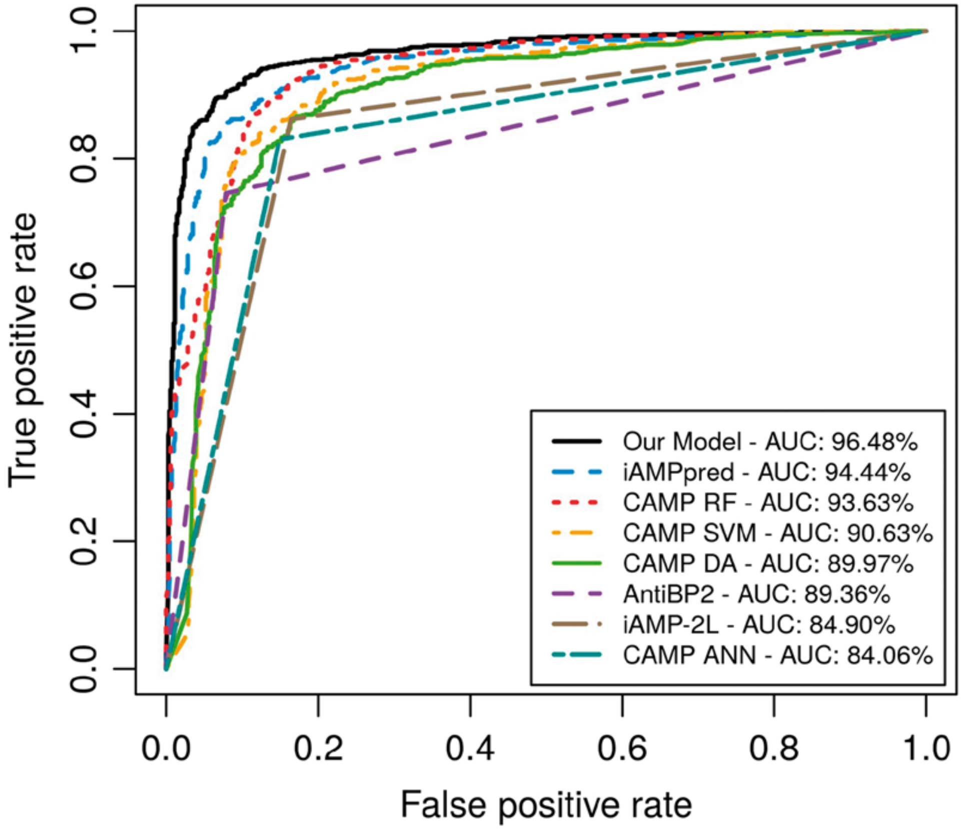
Classification performance on our testing data set



Method	SENS(%)	SPEC(%)	ACC(%)	MCC	auROC(%)
AntiBP2 (imtech.res.in/raghava/antibp2)	87.91	90.80	89.37	0.7876	89.36
CAMP-ANN (camp.bicnirrh.res.in/predict)	82.98	85.09	84.04	0.6809	84.06
CAMP-DA	87.08	80.76	83.92	0.6797	89.97
CAMP-RF	**92.70**	82.44	87.57	0.7554	93.63
CAMP-SVM	88.90	79.92	84.41	0.6910	90.63
iAMP-2L (jci-bioinfo.cn/iAMP-2L)	83.99	85.86	84.90	0.6983	84.90
iAMPpred (cabgrid.res.in:8080/amppred)	89.33	87.22	88.27	0.7656	94.44
Our DNN	89.89	**92.13**	**91.01**	**0.8204**	**96.48**

National Institute of Allergy and Infectious Diseases

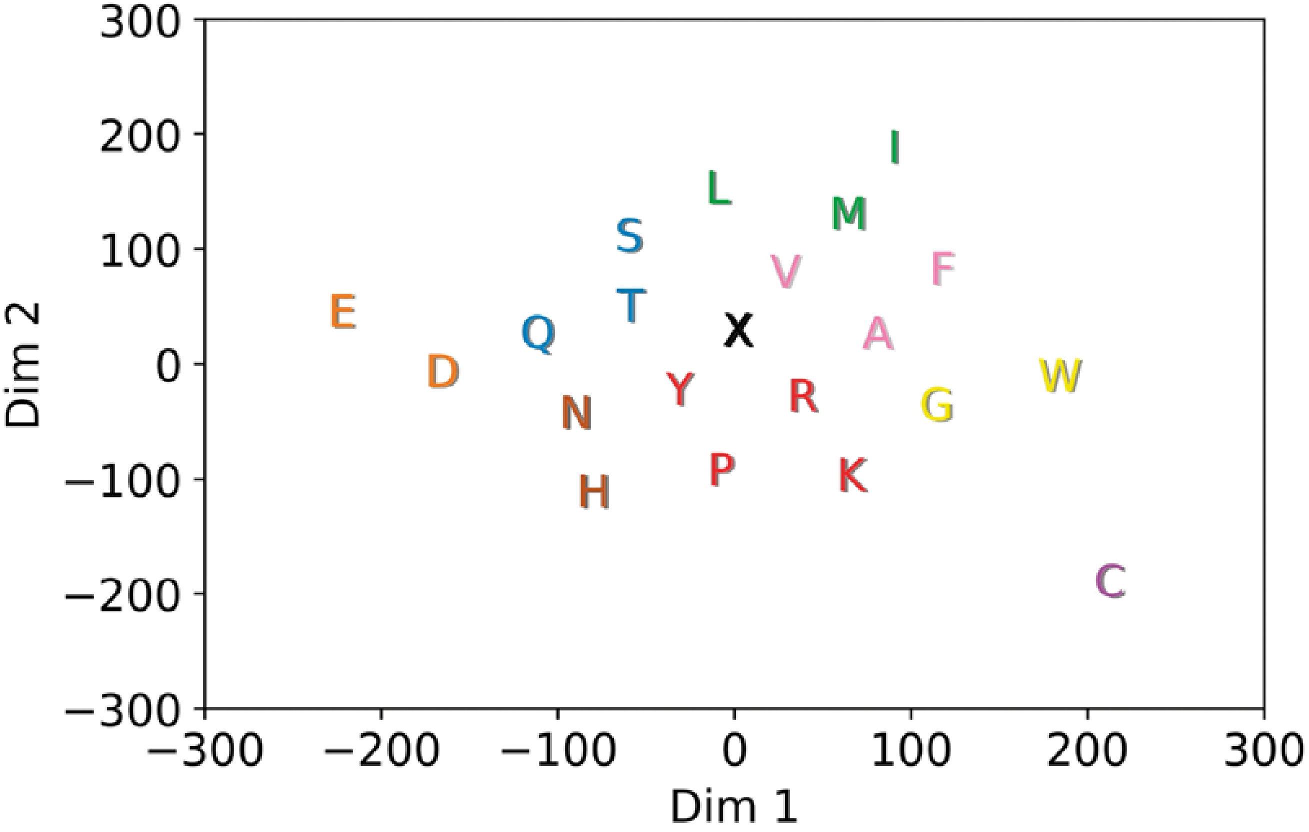
AMP Server Comparison ROC Curve



National Institute of Allergy and Infectious Diseases

Embedding Vector of Amino Acids

A 2D t-SNE [1] projection of the 128 dim. AA embedding vectors. K-means (*k=9*) used to select clusters



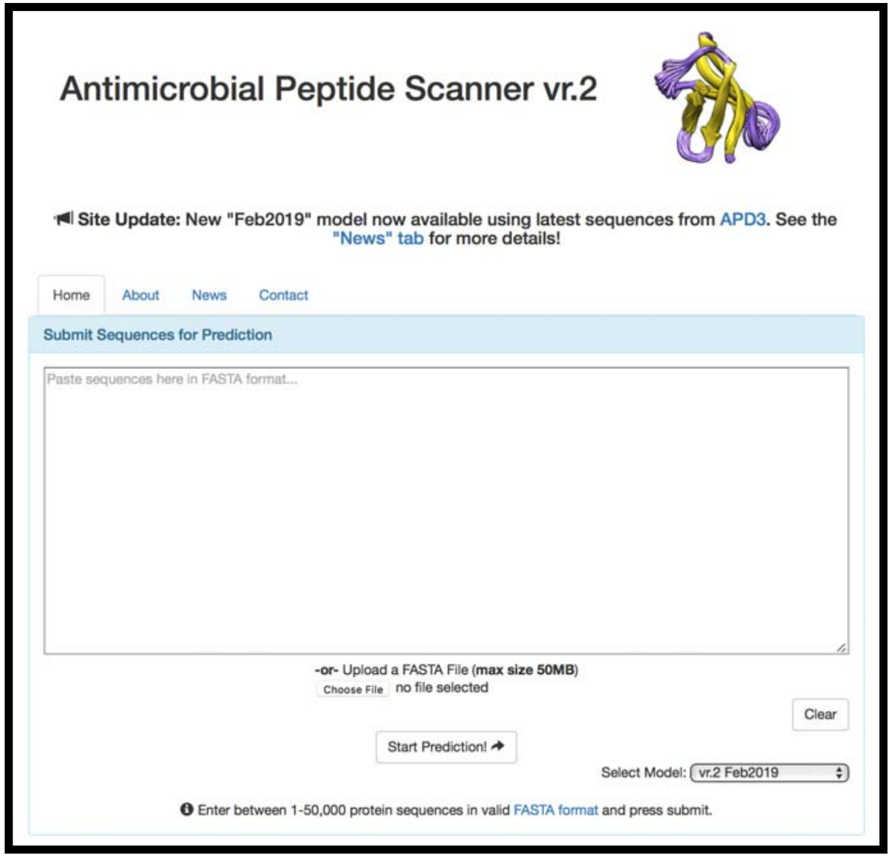
[1] Van der Maaten et al.
J. Machine Learn. Res., 2008

National Institute of Allergy and Infectious Diseases

AMP Scanner vr.2 Website

Feel free to try our methods out at:
www.ampscanner.com

Data sets are available to download and contact information if you would like the code from me!



Antimicrobial Peptide Scanner vr.2

Site Update: New "Feb2019" model now available using latest sequences from APD3. See the "News" tab for more details!

Submit Sequences for Prediction

Paste sequences here in FASTA format...

-or- Upload a FASTA File (max size 50MB)

no file selected

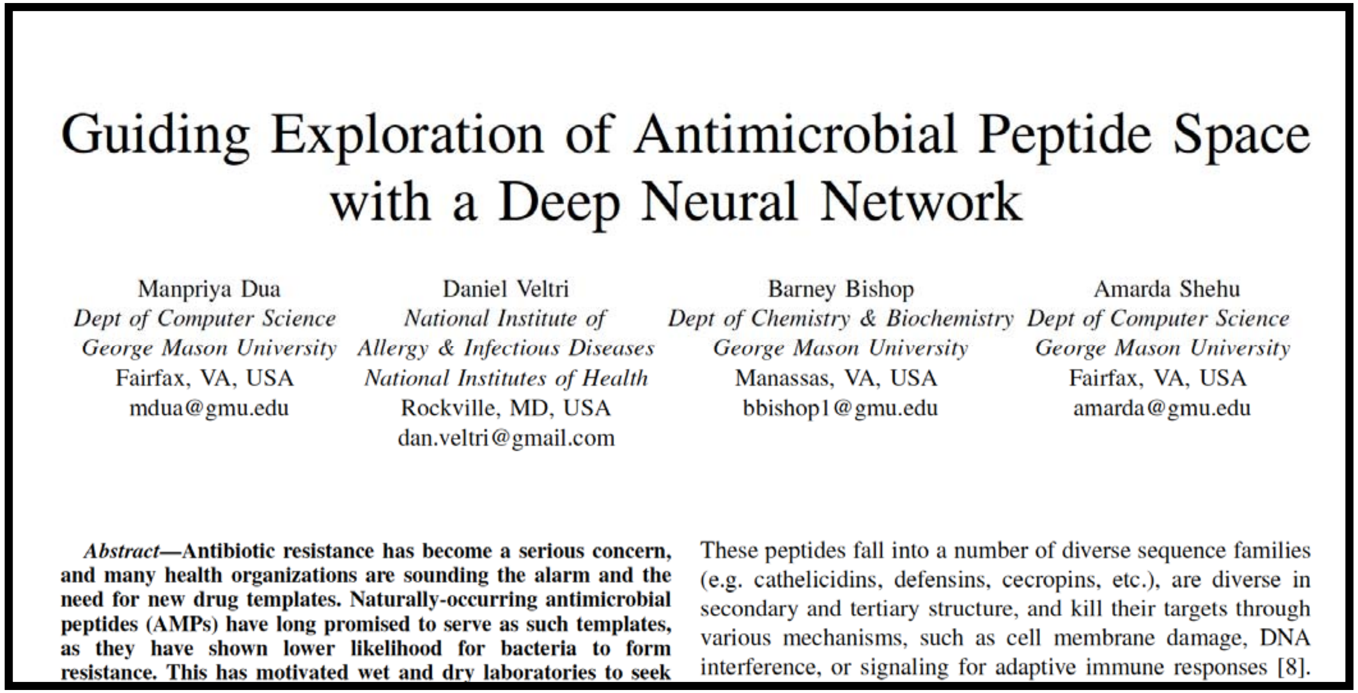
Select Model: vr.2 Feb2019

Enter between 1-50,000 protein sequences in valid FASTA format and press submit.

NIH National Institute of Allergy and Infectious Diseases

NIAID

Building a Generative Model for AMP-like Sequences



Guiding Exploration of Antimicrobial Peptide Space with a Deep Neural Network

Manpriya Dua
Dept of Computer Science
George Mason University
Fairfax, VA, USA
mdua@gmu.edu

Daniel Veltri
National Institute of
Allergy & Infectious Diseases
National Institutes of Health
Rockville, MD, USA
dan.veltri@gmail.com

Barney Bishop
Dept of Chemistry & Biochemistry
George Mason University
Manassas, VA, USA
bbishop1@gmu.edu

Amarda Shehu
Dept of Computer Science
George Mason University
Fairfax, VA, USA
amarda@gmu.edu

***Abstract*—Antibiotic resistance has become a serious concern, and many health organizations are sounding the alarm and the need for new drug templates. Naturally-occurring antimicrobial peptides (AMPs) have long promised to serve as such templates, as they have shown lower likelihood for bacteria to form resistance. This has motivated wet and dry laboratories to seek**

These peptides fall into a number of diverse sequence families (e.g. cathelicidins, defensins, cecropins, etc.), are diverse in secondary and tertiary structure, and kill their targets through various mechanisms, such as cell membrane damage, DNA interference, or signaling for adaptive immune responses [8].

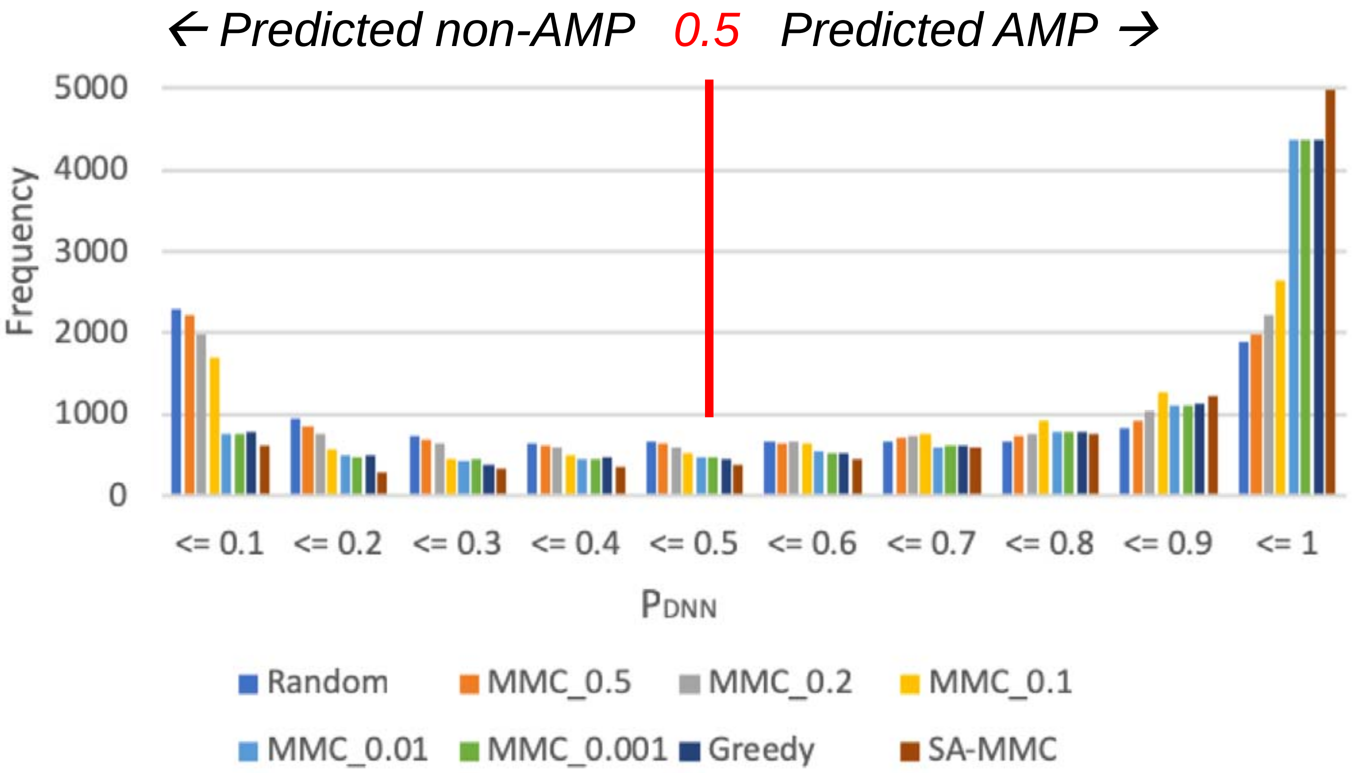
4 Different Sampling Methods

- ***RANDOM-*** randomly select AAs and a total peptide length (**L**) from a population of training AMPs (baseline method),
- ***GREEDY-*** Perform *RANDOM*, then select (**L+1**) AA's to substitute with changes that improve AMP probability (use prior deep learning model to judge),
- ***Metropolis Monte Carlo (MMC)-*** Perform *GREEDY* but, with a small probability, accepted *worse* AA changes. Temperature (**T**) parameter decides how often we do this (higher **T** → more changes → more diverse sequences),
- **Simulated Annealing (SA-MMC)**- Similar to MMC above but starts with a high **T** to start more diverse and gradually lowers **T** over time to become greedier.

NIH National Institute of Allergy and Infectious Diseases

Distribution of *All* Generated Peptides

The simulated annealing (SA-MCC) method performs best- it generates the most sequences predicted to be antimicrobial



National Institute of Allergy and Infectious Diseases

Future Directions

- Now that we can generate and evaluate AMP sequences, can we use *adversarial learning* to build improved AMP classifiers?

- More work needs to be done predicting how AMPs may work against *specific bacteria* of medical interest. Can we do better at predicting MIC, EC50 etc.?

National Institute of Allergy and Infectious Diseases

Collaborators

Amarda Shehu, Ph.D.
George Mason University
Computer Science Dept.

Uday Kamath, Ph.D.
Digital Reasoning Systems, Inc.

Barney Bishop, Ph.D.
George Mason University
Chemistry Dept.

Manpriya Dua, M.S.
George Mason University
Computer Science Dept.

Special thanks to members of NIAID BCBB, the Shehu lab and Jianlin Cheng (U. Missouri) for their helpful feedback and suggestions.

NIH National Institute of Allergy and Infectious Diseases

Thank you for listening!

Questions?

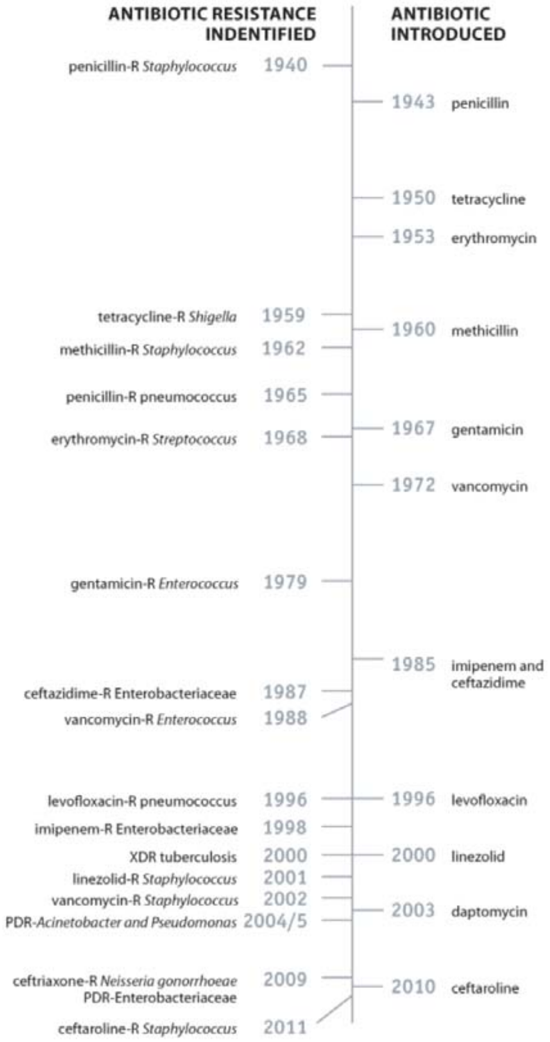
NIH National Institute of Allergy and Infectious Diseases

NIAID

Extra Slides

National Institute of
Allergy and
Infectious Diseases

Antimicrobial Resistance Rates



ANTIBIOTIC RESISTANCE INDENTIFIED

- penicillin-R *Staphylococcus* 1940
- tetracycline-R *Shigella* 1959
- methicillin-R *Staphylococcus* 1962
- penicillin-R pneumococcus 1965
- erythromycin-R *Streptococcus* 1968
- gentamicin-R *Enterococcus* 1979
- ceftazidime-R Enterobacteriaceae 1987
- vancomycin-R *Enterococcus* 1988
- levofloxacin-R pneumococcus 1996
- imipenem-R Enterobacteriaceae 1998
- XDR tuberculosis 2000
- linezolid-R *Staphylococcus* 2001
- vancomycin-R *Staphylococcus* 2002
- PDR-*Acinetobacter and Pseudomonas* 2004/5
- ceftriaxone-R *Neisseria gonorrhoeae* 2009
- PDR-Enterobacteriaceae
- ceftaroline-R *Staphylococcus* 2011

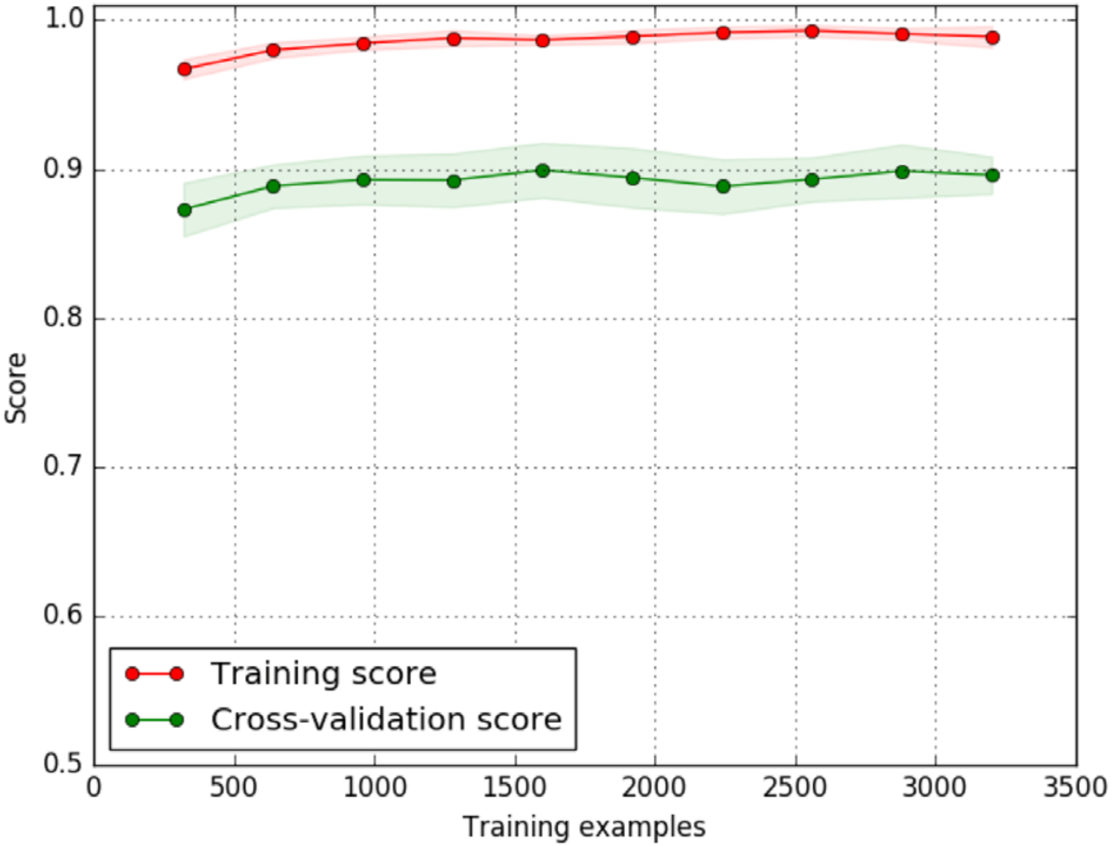
ANTIBIOTIC INTRODUCED

- 1943 penicillin
- 1950 tetracycline
- 1953 erythromycin
- 1960 methicillin
- 1967 gentamicin
- 1972 vancomycin
- 1985 imipenem and ceftazidime
- 1996 levofloxacin
- 2000 linezolid
- 2003 daptomycin
- 2010 ceftaroline

Source: US Center for Disease Control
https://www.cdc.gov/drugresistance/about.html

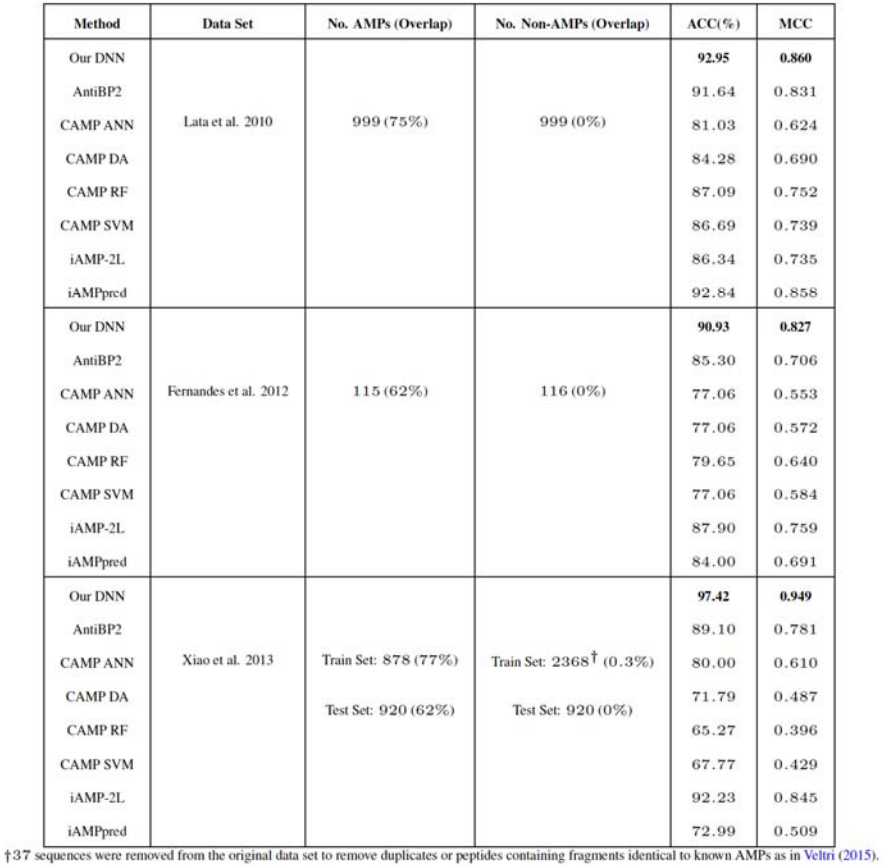
National Institute of
Allergy and
Infectious Diseases

Learning Curves (10-Fold CV)



National Institute of Allergy and Infectious Diases

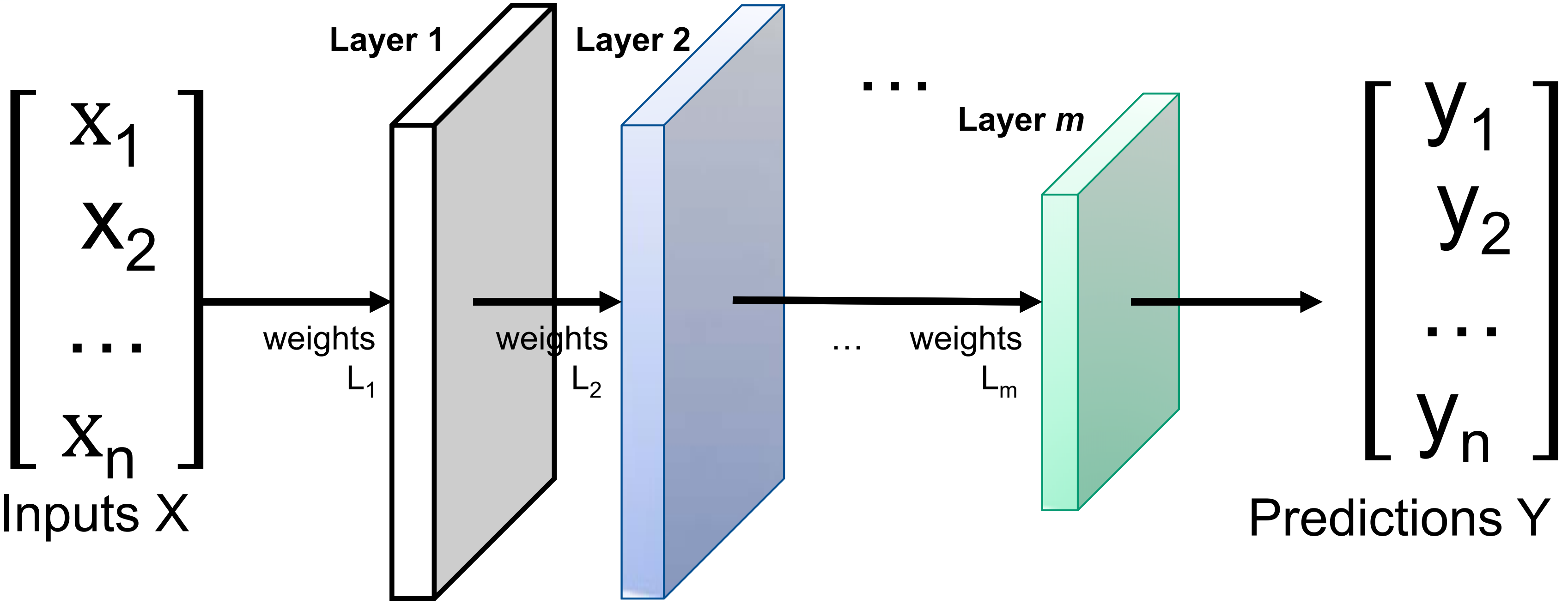
Performance Comparison on Other AMP Data Sets



Method	Data Set	No. AMPs (Overlap)	No. Non-AMPs (Overlap)	ACC(%)	MCC
Our DNN	Lata et al. 2010	999 (75%)	999 (0%)	**92.95**	**0.860**
AntiBP2				91.64	0.831
CAMP ANN				81.03	0.624
CAMP DA				84.28	0.690
CAMP RF				87.09	0.752
CAMP SVM				86.69	0.739
iAMP-2L				86.34	0.735
iAMPpred				92.84	0.858
Our DNN	Fernandes et al. 2012	115 (62%)	116 (0%)	**90.93**	**0.827**
AntiBP2				85.30	0.706
CAMP ANN				77.06	0.553
CAMP DA				77.06	0.572
CAMP RF				79.65	0.640
CAMP SVM				77.06	0.584
iAMP-2L				87.90	0.759
iAMPpred				84.00	0.691
Our DNN	Xiao et al. 2013	Train Set: 878 (77%) Test Set: 920 (62%)	Train Set: 2368† (0.3%) Test Set: 920 (0%)	**97.42**	**0.949**
AntiBP2				89.10	0.781
CAMP ANN				80.00	0.610
CAMP DA				71.79	0.487
CAMP RF				65.27	0.396
CAMP SVM				67.77	0.429
iAMP-2L				92.23	0.845
iAMPpred				72.99	0.509

†37 sequences were removed from the original data set to remove duplicates or peptides containing fragments identical to known AMPs as in Veltri (2015).

Weights and Layers

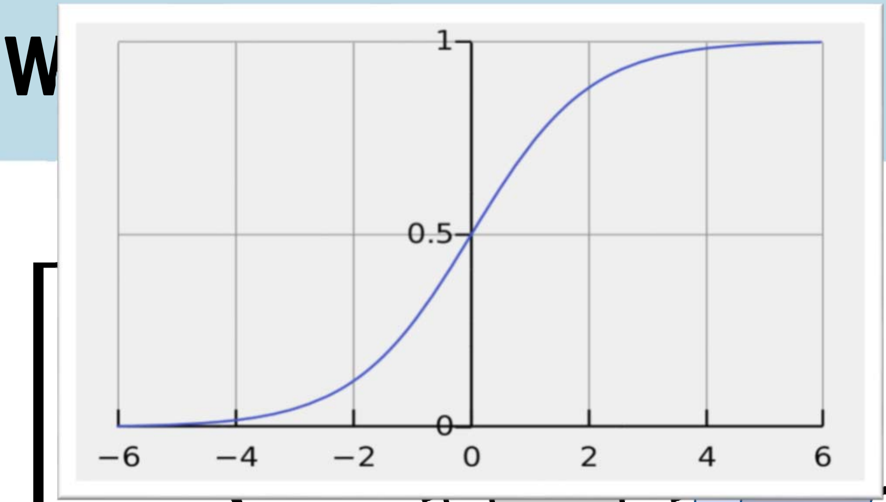

Layer 1
Layer 2
...
Layer *m*
$\begin{bmatrix} x_1 \\ x_2 \\ \ldots \\ x_n \end{bmatrix}$
Inputs X
weights L_1
weights L_2
...
weights L_m
$\begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix}$
Predictions Y

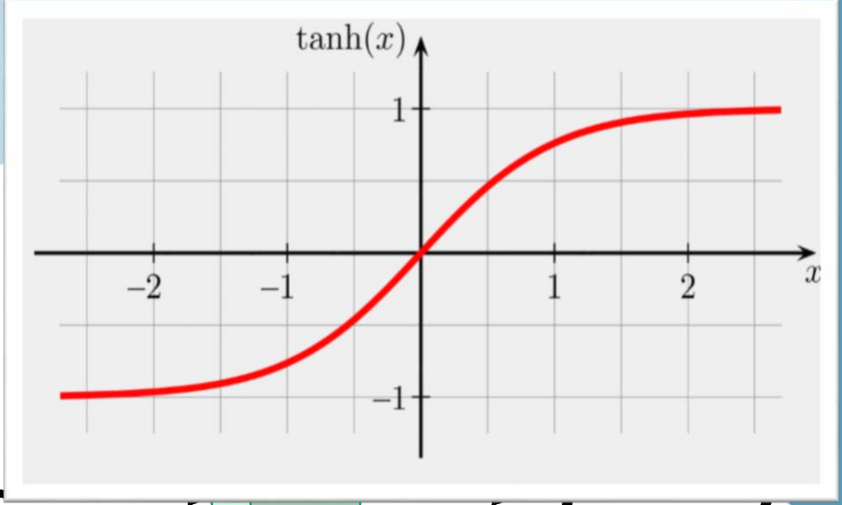
Often weights are randomly initialized and layer outputs are often “activated” using functions to force numbers in a certain range. Typical examples include: ***sigmoid***, ***tanh***, and **rectifier linear unit** (ReLU) functions.

NIH National Institute of Allergy and Infectious Diseases

NIAID

W

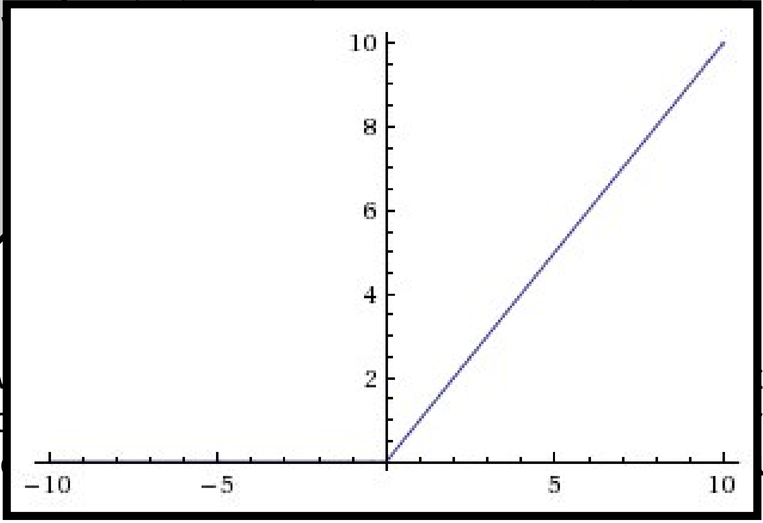




… x_n

Inputs X

weights L_1



… y_n

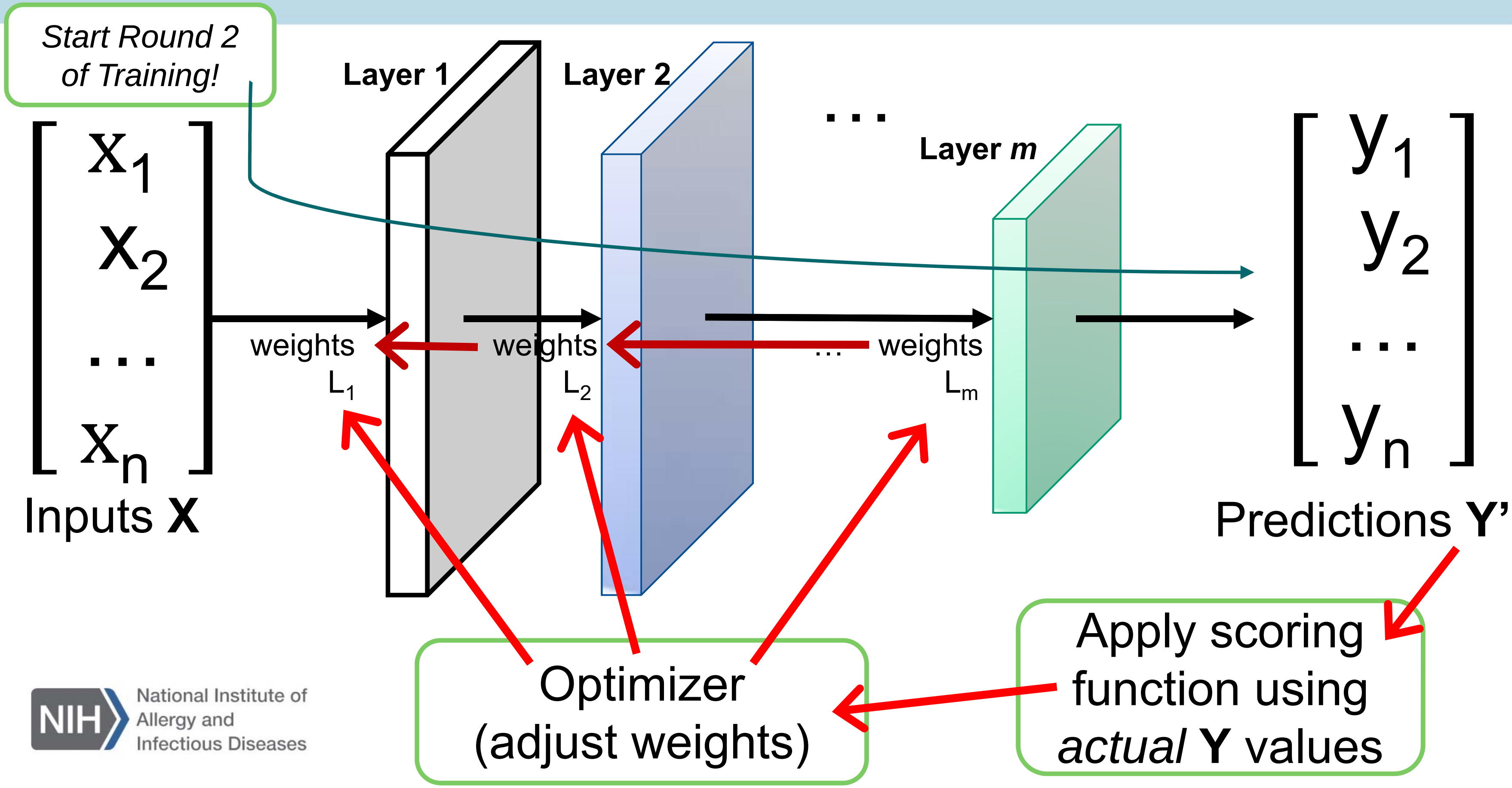
Predictions Y

Often v s are often
“activated ange. Typical
exampl **unit** (ReLU)

National Institute of
Allergy and
Infectious Diseases

NIAID

Weights and Layers – Optimizing the Network



Start Round 2 of Training!

Optimizer (adjust weights)

Apply scoring function using *actual* **Y** values

Optimizing via Backpropagation

The Secret Sauce of Deep Neural Networks (DNNs)

- How do DNNs learn so well? The key is they compute answers across layers in a ***forward pass*** and then to use a ***backwards pass*** to optimize the weights. This way **all** layers are updated each round (sometimes called an 'epoch') of training!
- How does this backward pass work? **The chain rule** (remember from calculus?) – where we can calculate the derivative (the slope or rate of change) from *two or more* functions.

You might have seen this written as: $(f \circ g)' = (f' \circ g) \cdot g'$

or maybe like this: $\frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx}$ or maybe like this: $F'(x) = f'\big(g(x)\big)g'(x)$

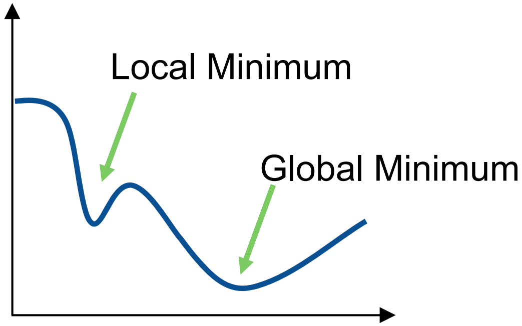
The takeaway: we can calculate the derivative using multiple functions at the same time!

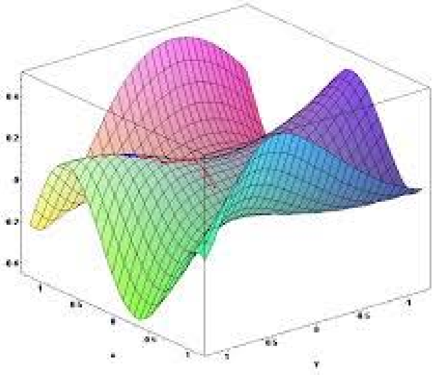
Optimizing via Backpropagation

The Secret Sauce of Deep Neural Networks (DNNs)

- For our DNNs we are calculating the ***gradient*** (a vector of derivatives) to account for the change across the network based on the forward pass results.
- Given a function $f(x)$ where x's are our **training** inputs- the gradient forms a vector: $\nabla f(x) =$

$$[\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}] = [y,x]$$





NIH National Institute of Allergy and Infectious Diseases

NIAID

Some Backpropagation Intuition

Lets look at multiplication: $\boldsymbol{f(x,y) = xy} \rightarrow \frac{\partial f}{\partial x} = y \quad \frac{\partial f}{\partial y} = x$

If x = 4 and y = −3 $\quad f(x,y) = -12 \quad \frac{\partial f}{\partial x} = -3 \quad \frac{\partial f}{\partial y} = 4$

Lets look at basic addition: $\boldsymbol{f(x,y) = x + y} \rightarrow \boxed{???}$

What happens to each function if we change x … or change y?

Some Backpropagation Intuition

Lets look at ***multiple*** functions:

$$f(x, y, z) = (x + y)\, z$$

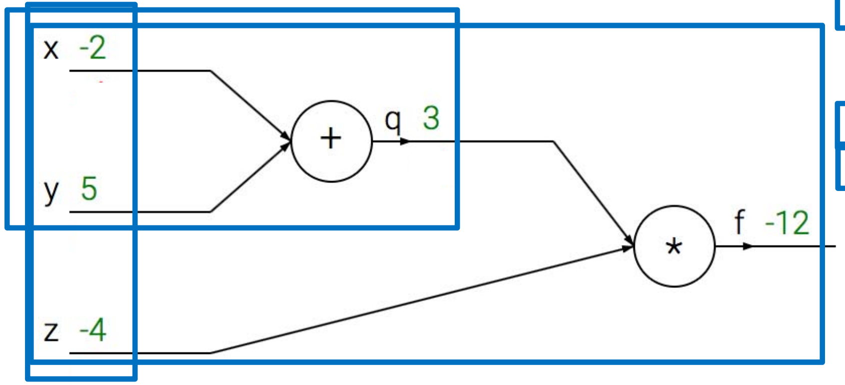
We can rewrite this as: $q = x + y \;\; and\; f = qz$

SO $\frac{\partial f}{\partial q}$ **= z ,** $\frac{\partial f}{\partial z}$ **= q** … for $(x + y)$ as we saw before: $\frac{\partial f}{\partial x}$ **= 1 ,** $\frac{\partial f}{\partial y}$ **= 1**

The **chain rule** says **multiply:** $\frac{\partial f}{\partial x} = \frac{\partial f}{\partial q} \cdot \frac{\partial q}{\partial x}$

Lets look at this with code and a visual representation!

Backpropagation Example



```
# set some inputs
x = -2; y = 5; z = -4

# perform the forward pass
q = x + y # q becomes 3
f = q * z # f becomes -12

# perform the backward pass (backpropagation)
# first backprop through f = q * z
dfdq = z # df/dq = z, so gradient on q becomes -4
dfdz = q # df/dz = q, so gradient on z becomes 3

# now backprop through q = x + y
dfdx = 1.0 * dfdq # dq/dx = 1  chain rule!
dfdy = 1.0 * dfdq # dq/dy = 1
```

*Forward pass is **green**. Backward pass is **red**.*

The big takeaway: The final gradient [$\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$, $\frac{\partial f}{\partial z}$] tells us how sensitive our function f is to the variables x , y , and z.