Machine learning for RNA (secondary) structure prediction

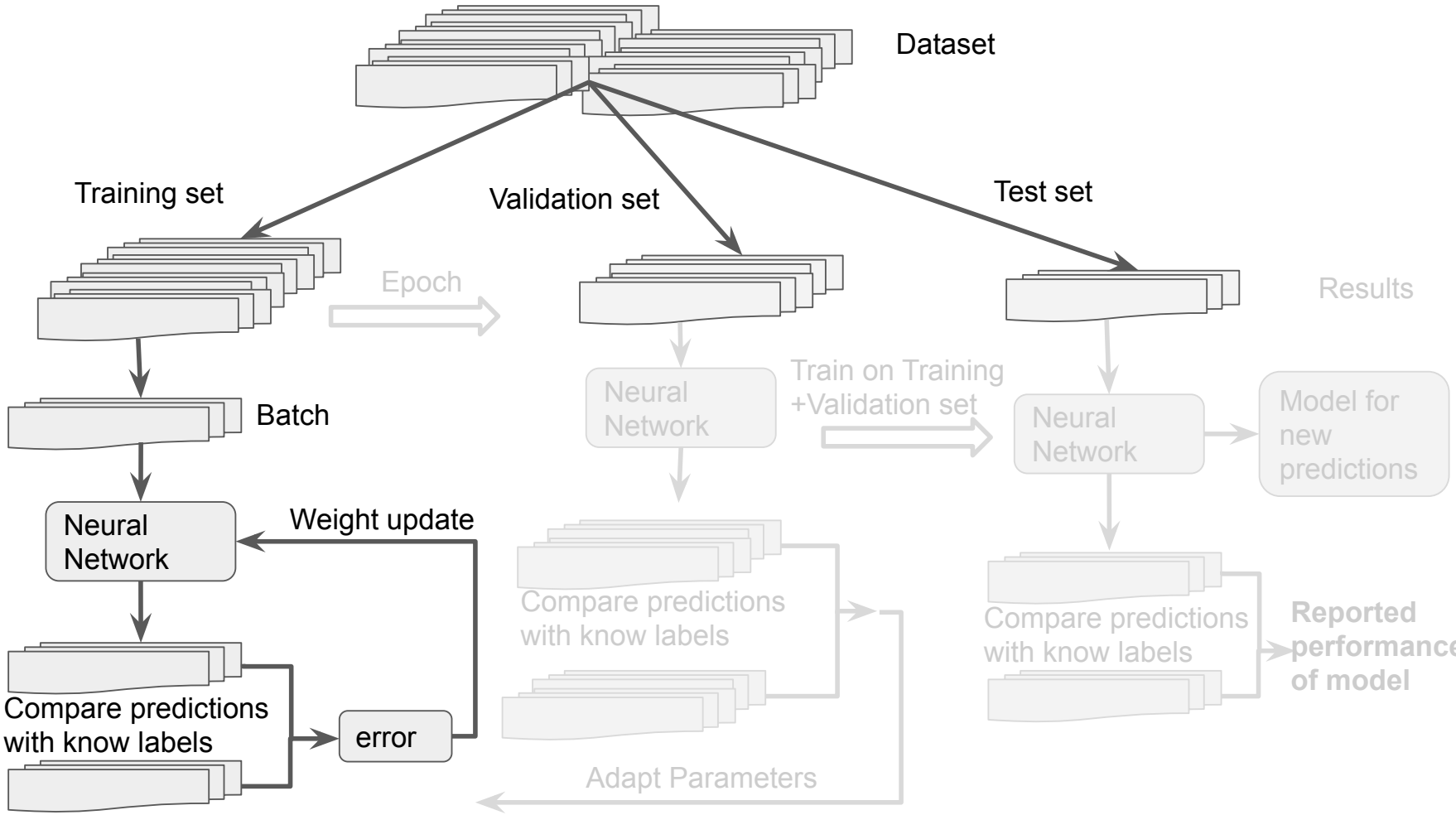
Julia Wielach

12th February 2020

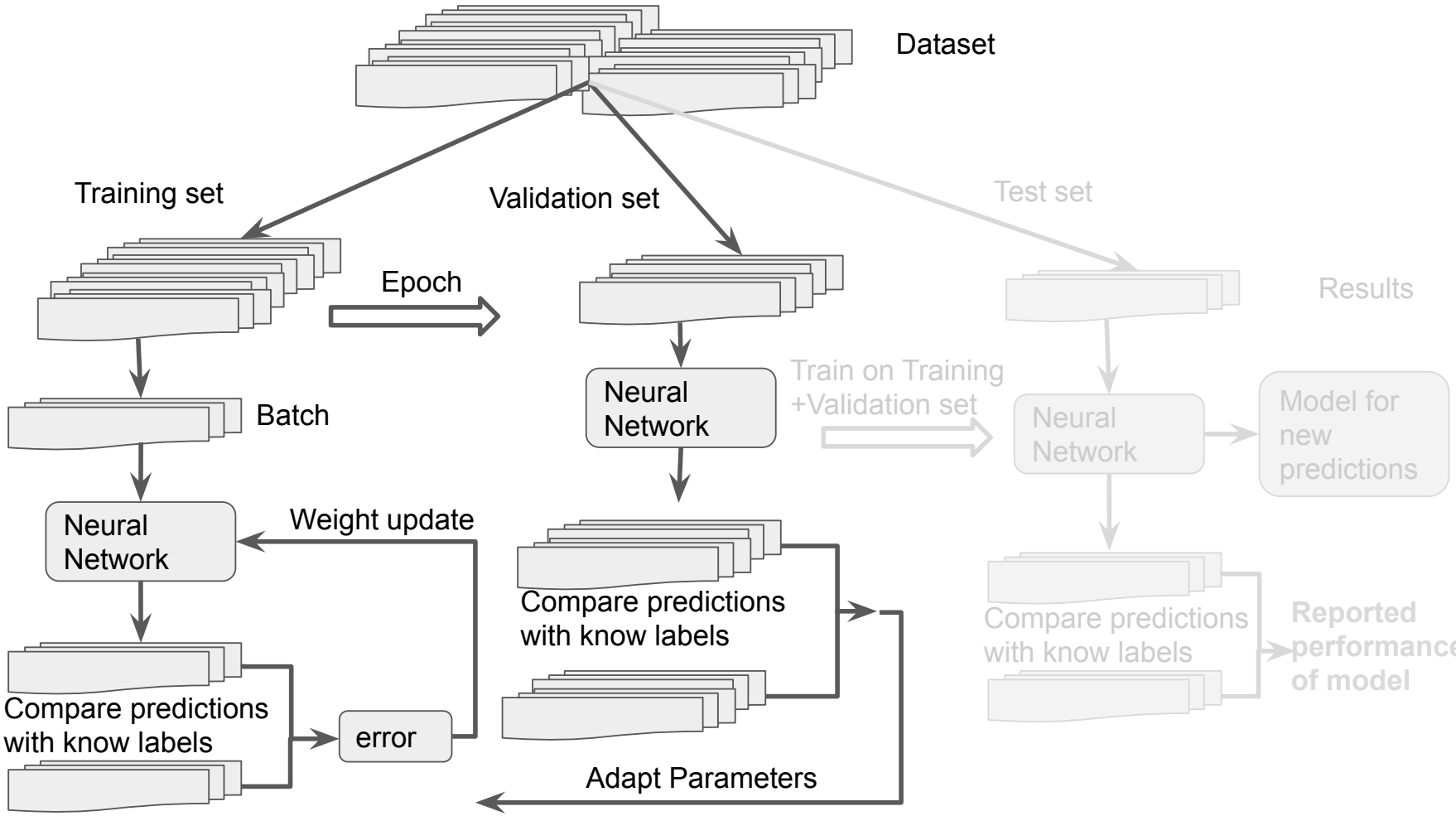
Overview

- Machine Learning
  - Training Process
  - Network Types
- ML for RNA structure prediction
  - Introduction
  - Aim of Project
  - Input
  - Output
  - Networks

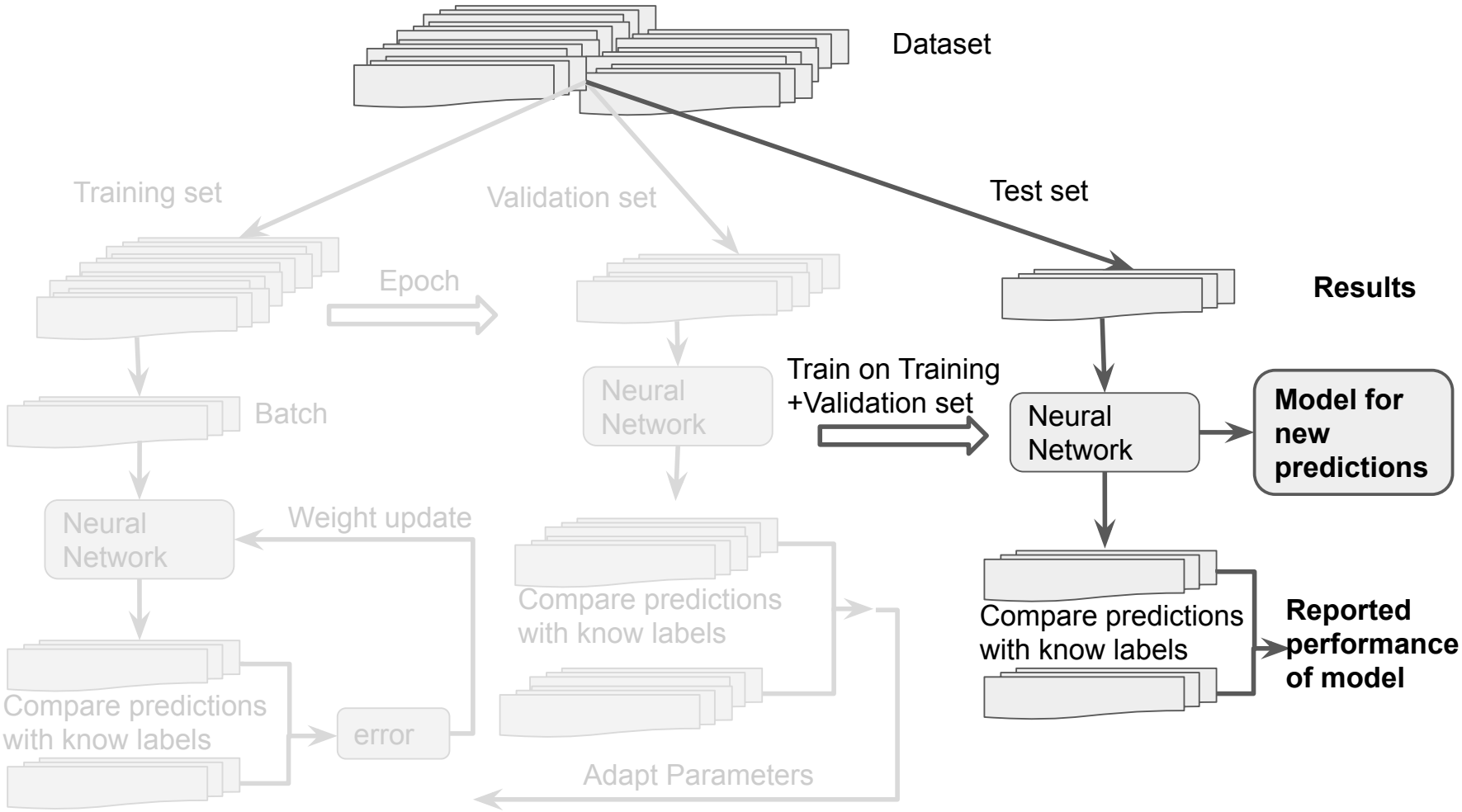
Training


Dataset
Training set
Validation set
Test set
Epoch
Results
Batch
Neural Network
Train on Training +Validation set
Neural Network
Model for new predictions
Neural Network
Weight update
Compare predictions with know labels
error
Compare predictions with know labels
Compare predictions with know labels
Reported performance of model
Adapt Parameters

Validation


Dataset
Training set
Validation set
Test set
Epoch
Batch
Neural Network
Weight update
Compare predictions with know labels
error
Adapt Parameters
Train on Training +Validation set
Results
Model for new predictions
Reported performance of model

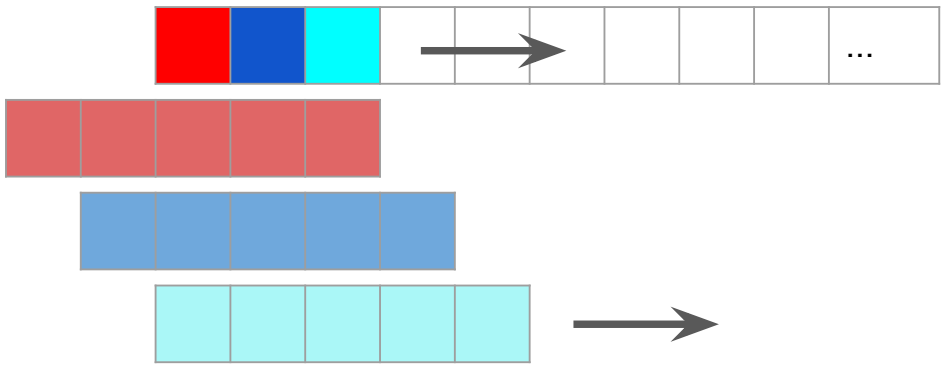
Testing

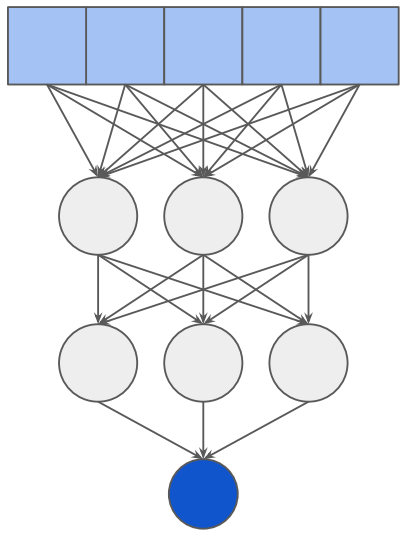

Dataset
Training set
Validation set
Test set
Epoch
Results
Batch
Neural Network
Train on Training +Validation set
Neural Network
Model for new predictions
Weight update
Neural Network
Compare predictions with know labels
Compare predictions with know labels
Reported performance of model
Compare predictions with know labels
error
Adapt Parameters

Sliding Window

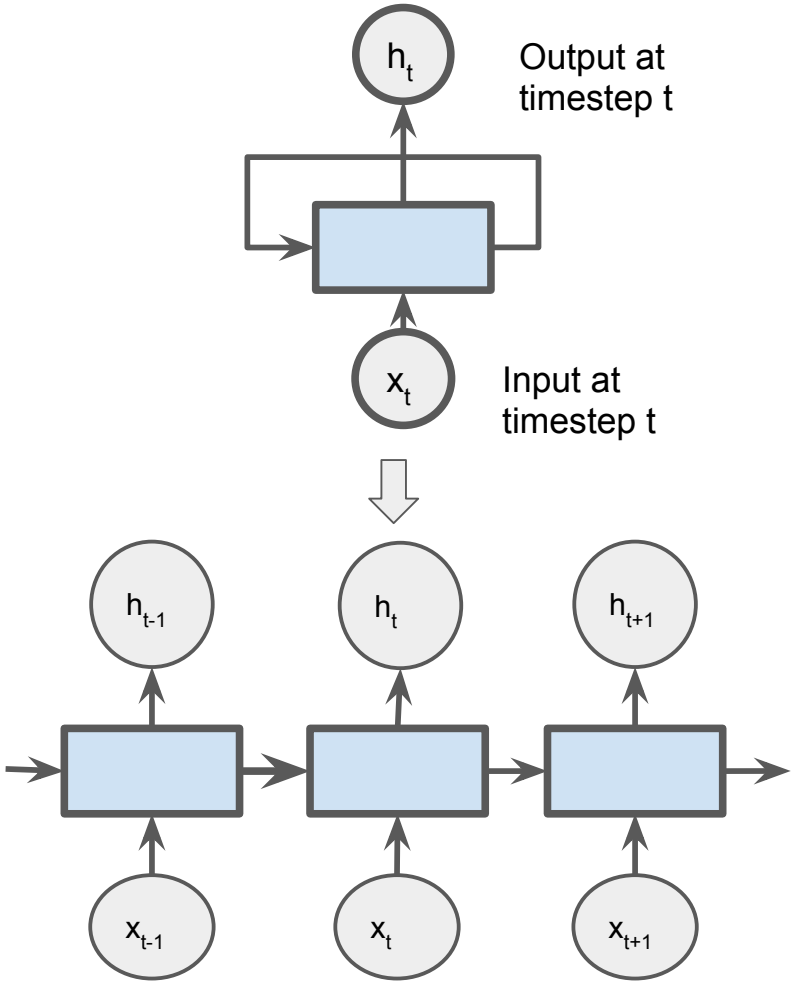
Benchmark

Only local dependencies (within window)



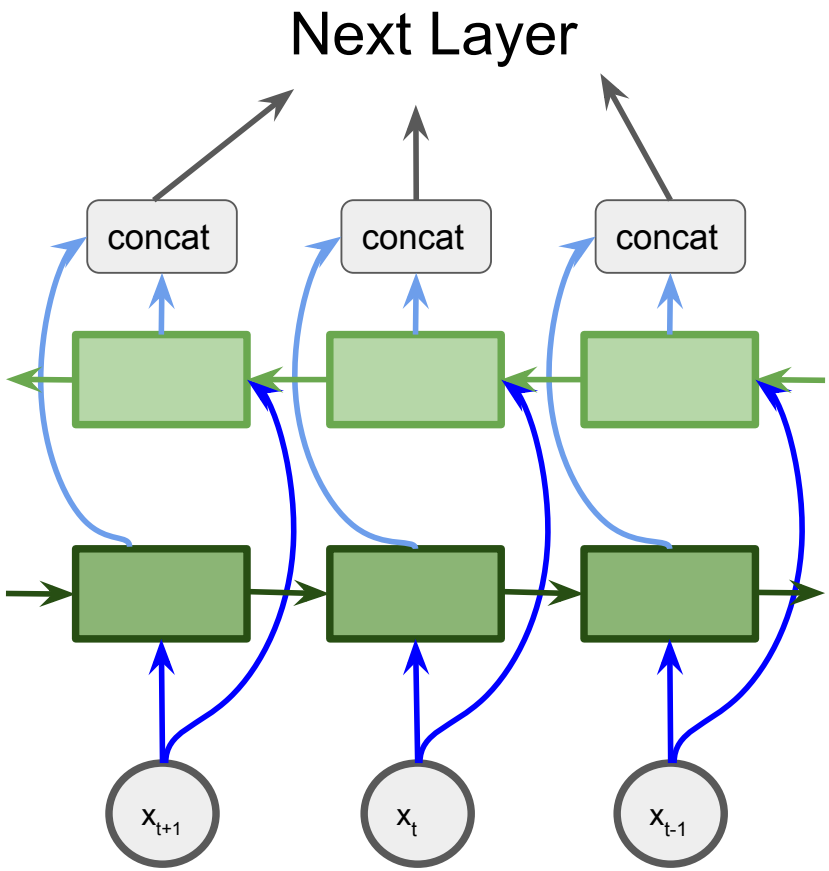


Recurrent Neural Network (RNN)



Graphic based on colah's blog

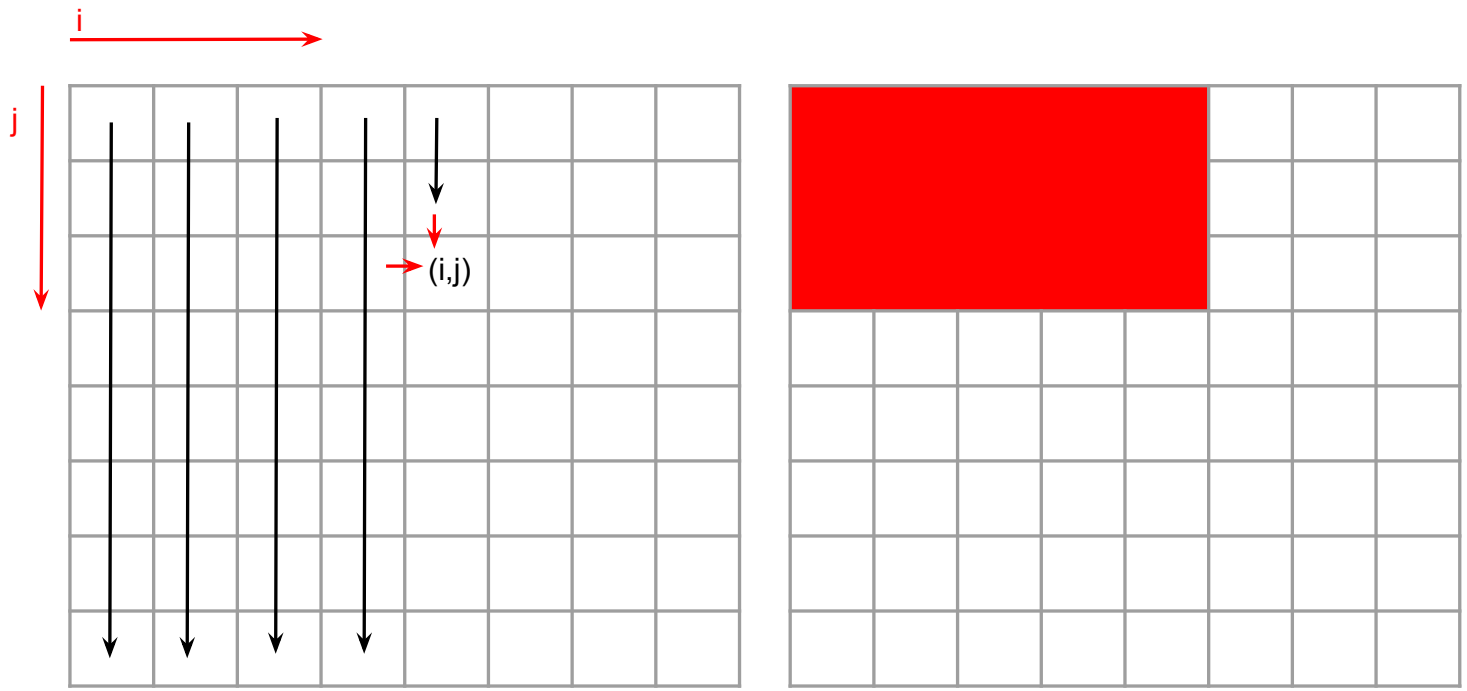
Bidirectional RNN



Schuster, M. & Paliwal, K. K.(1997)

Multidimensional RNN (2D)

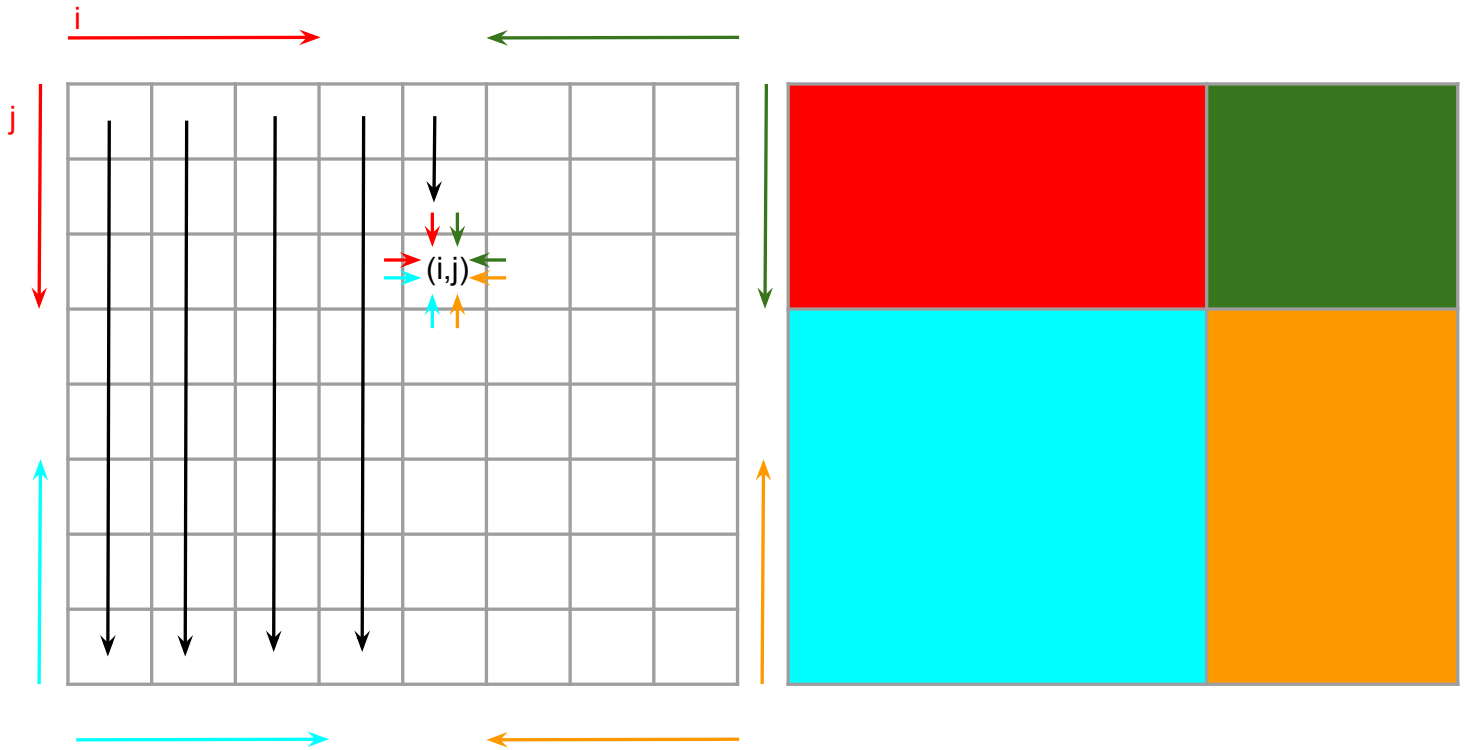
Input from two directions instead of one
(i,j) only reached after (i,j-1) and (i-1,j)



Graves A., Fernández S., Schmidhuber J. (2007)

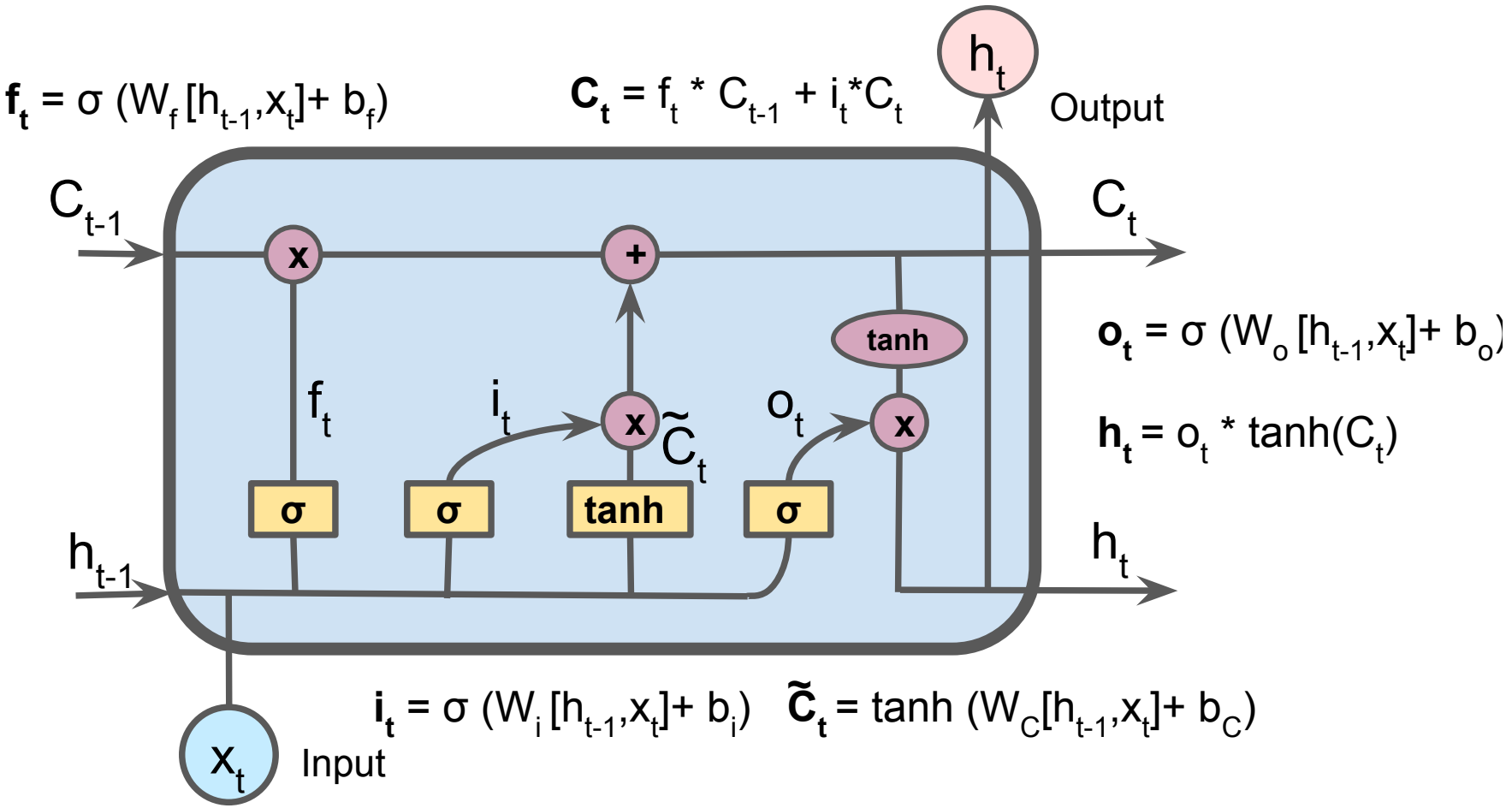
Multidimensional RNN (2D)

At every point context from 4 directions



Graves A., Fernández S., Schmidhuber J. (2007)

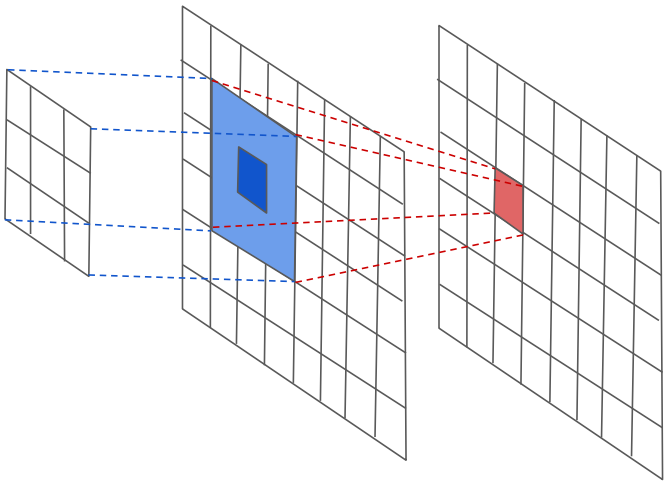
Long Short Term Memory (LSTM) Network

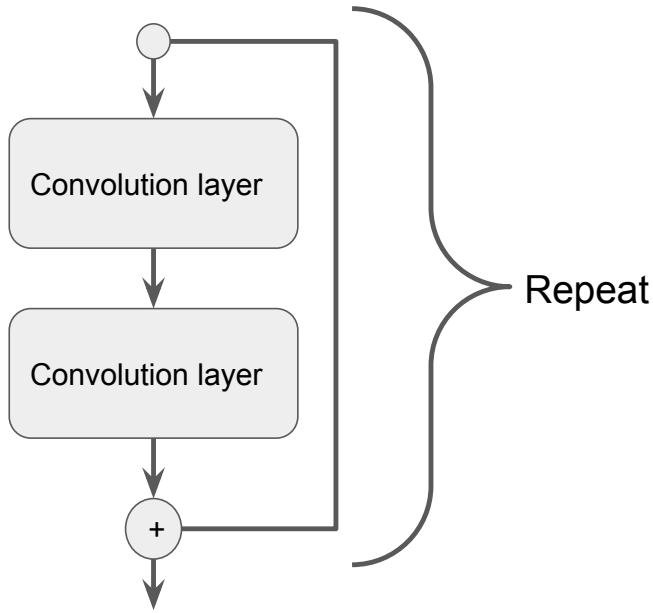


Hochreiter, S.& Schmidhuber, J.(1997)
Graphic based on colah's blog

Convolution NNs & ResNets

Capture local surroundings
Problems with very deep networks $\rightarrow$ performance decrease
Skip Connections

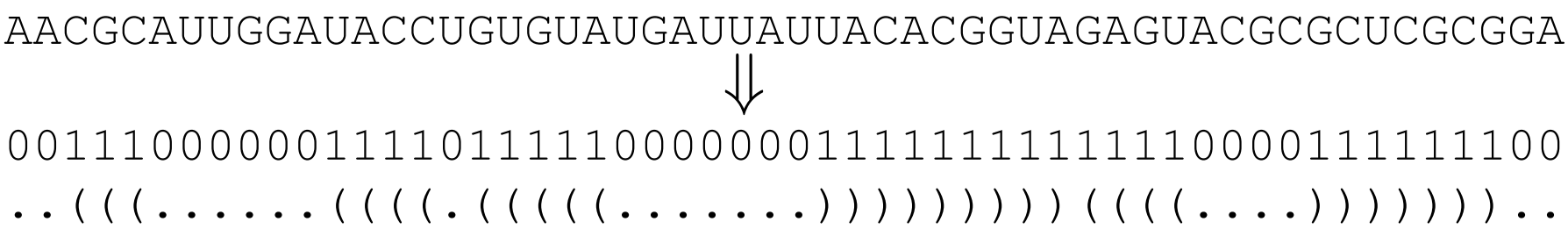




He, K. & Zhang, X. (2015)
CNN graphic based on River Trail Documentation

ML for RNA structure

What?



Why?

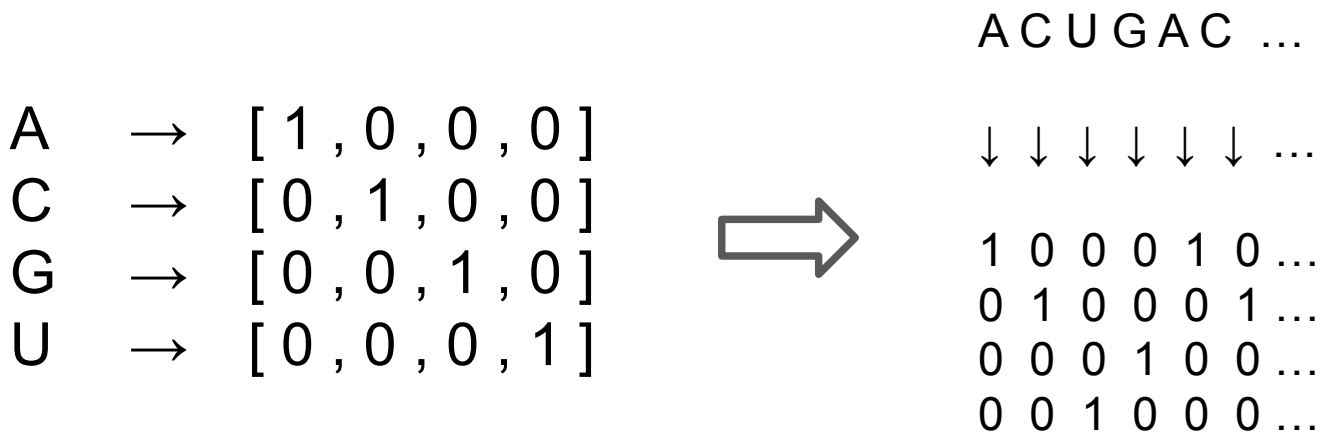
- Used for similar tasks
- Possibly capture complicated dependencies
- Already used in publications
- No performance comparisons

Current situation + project

- First Networks published
  - Very good performances reported
  - No comparisons between new models
  - Possible bias introduced by datasets
  - Predictions specific for different types of RNAs
- Tests with artificial random RNA sequences predicted by RNAfold
  - Similar performance reachable ?
  - Dataset size dependency
  - Capacity dependency

Input

One hot encoding



Input

Concatenation of one hot encodings
Matrix of size LxLx8

	A	**C**	**U**	**...**
A	[1,0,0,0,1,0,0,0]	[0,1,0,0,1,0,0,0]	[0,0,0,1,1,0,0,0]	
C		[0,1,0,0,0,1,0,0]	[0,0,0,1,0,1,0,0]	
U			[0,0,0,1,0,0,0,1]	
...				

Singh, J., Hanson, J., Paliwal, K. et al. (2019)

Input

Matrix with value "rating" basepair
Matrix of size LxLx1

	A	C	G	U
A	0	0	0	2 +3 * b
C		0	3 +2 * b	0
G			0	0.8
U				0

Zhang H, Zhang C, Li Z, Li C, Wei X, Zhang B and Liu Y (2019)

Output

- Problem of incorrect secondary structure

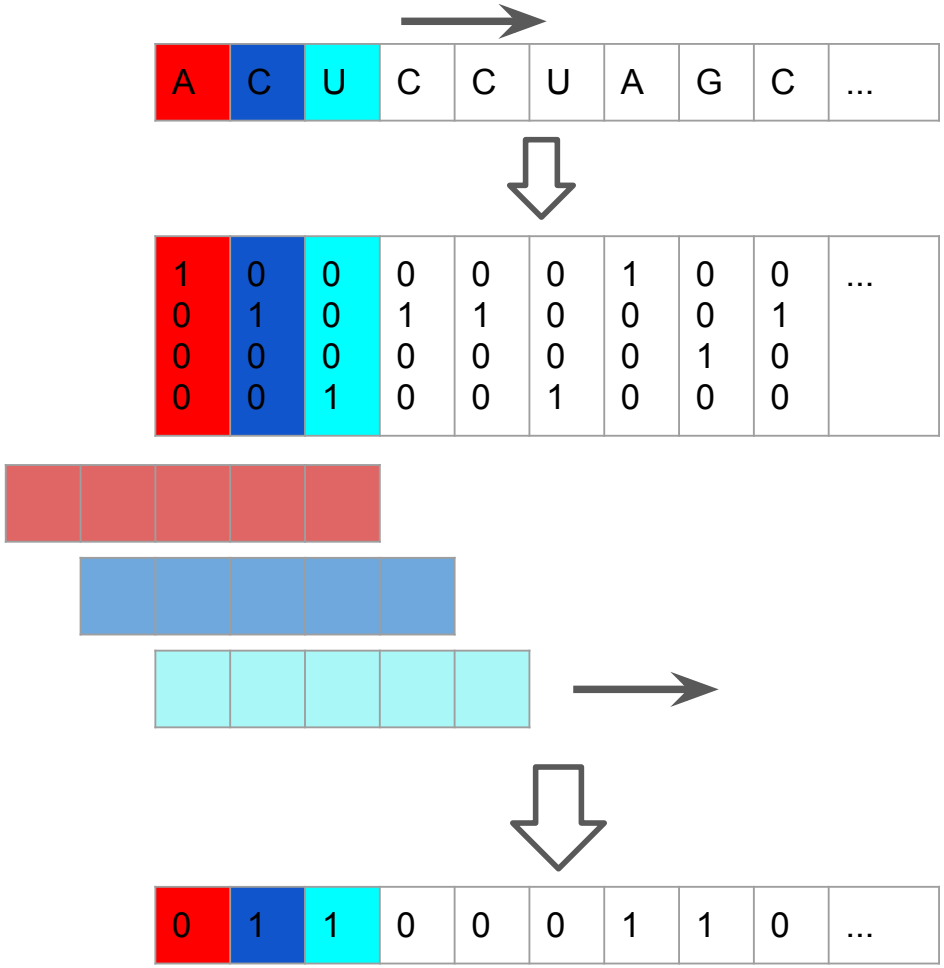
  Incorrect pairing partners

  Opening and closing brackets $\rightarrow$ different number

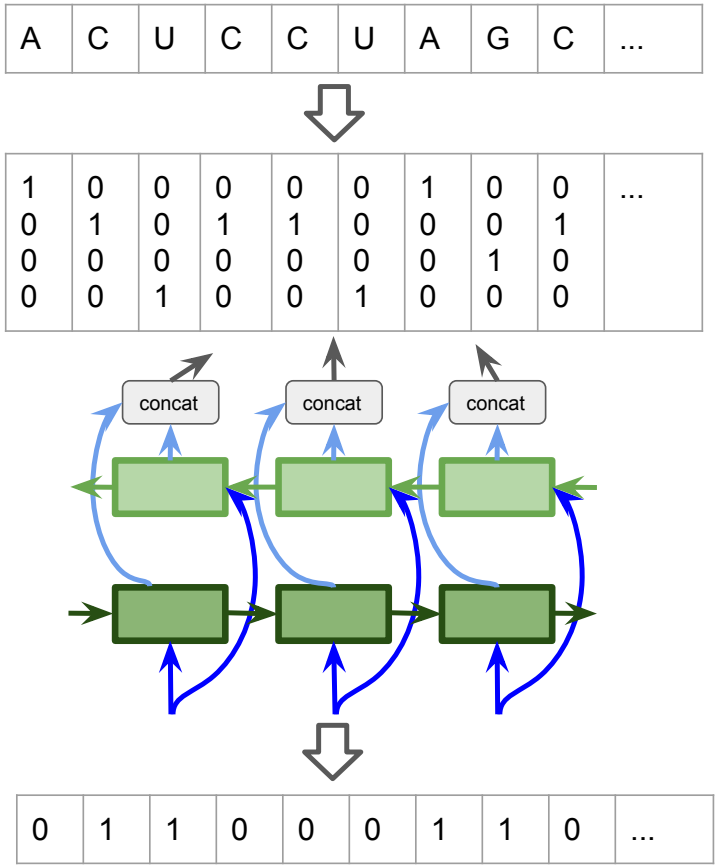
  Separate label for every base $\rightarrow$ conflicting basepairs
  eg. i+j, j+k

- Classifiy additional interactions (Pseudoknots, Triple and Non-canonical basepairs)
- Additional Corrections
- Simplification (1/0 for Paired/Unpaired)

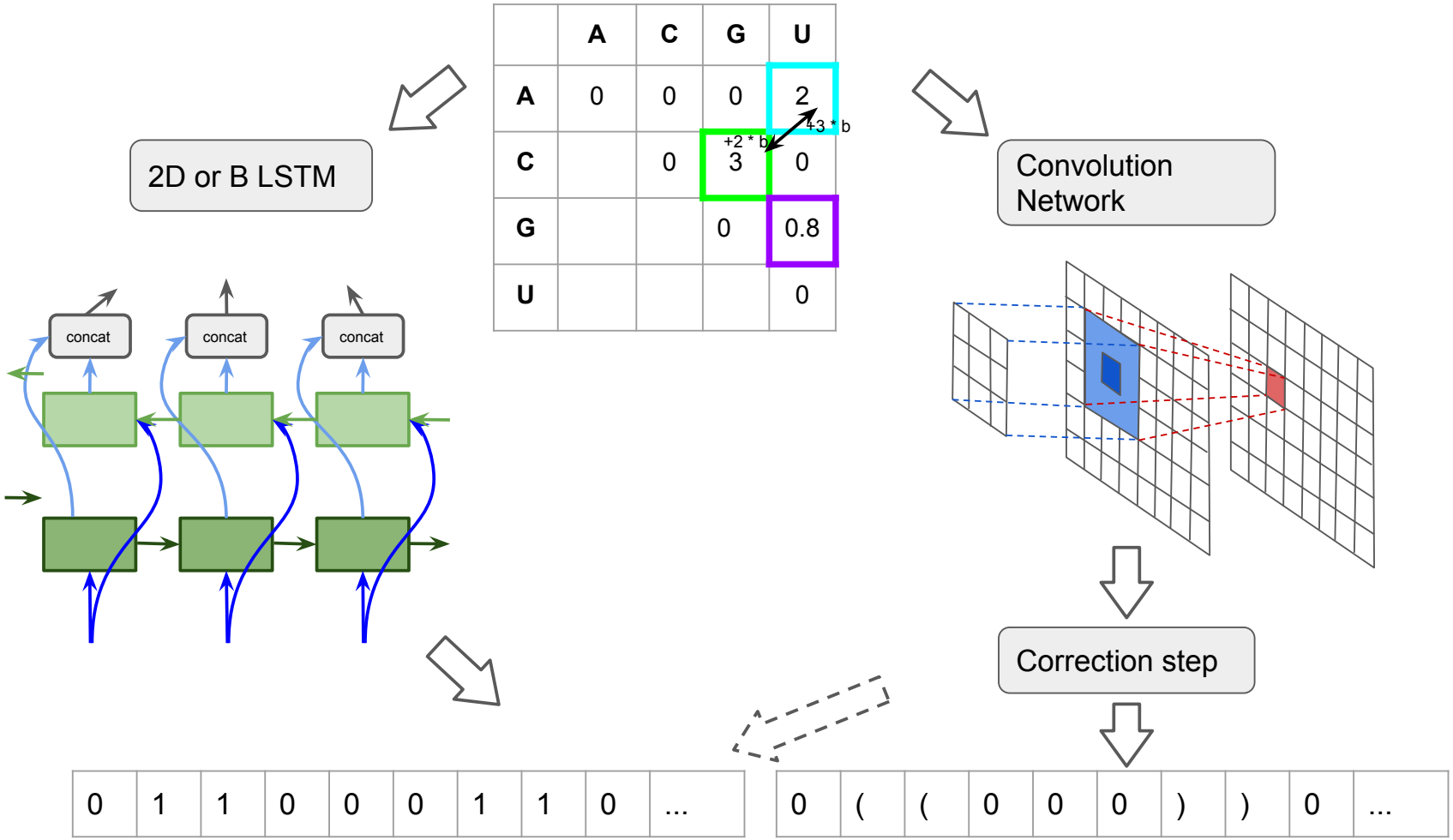
Sliding Window



B-LSTM


A C U C C U A G C ...
1 0 0 0 0 0 1 0 0 ...
0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 ...
concat
concat
concat
0 1 1 0 0 0 1 1 0 ...

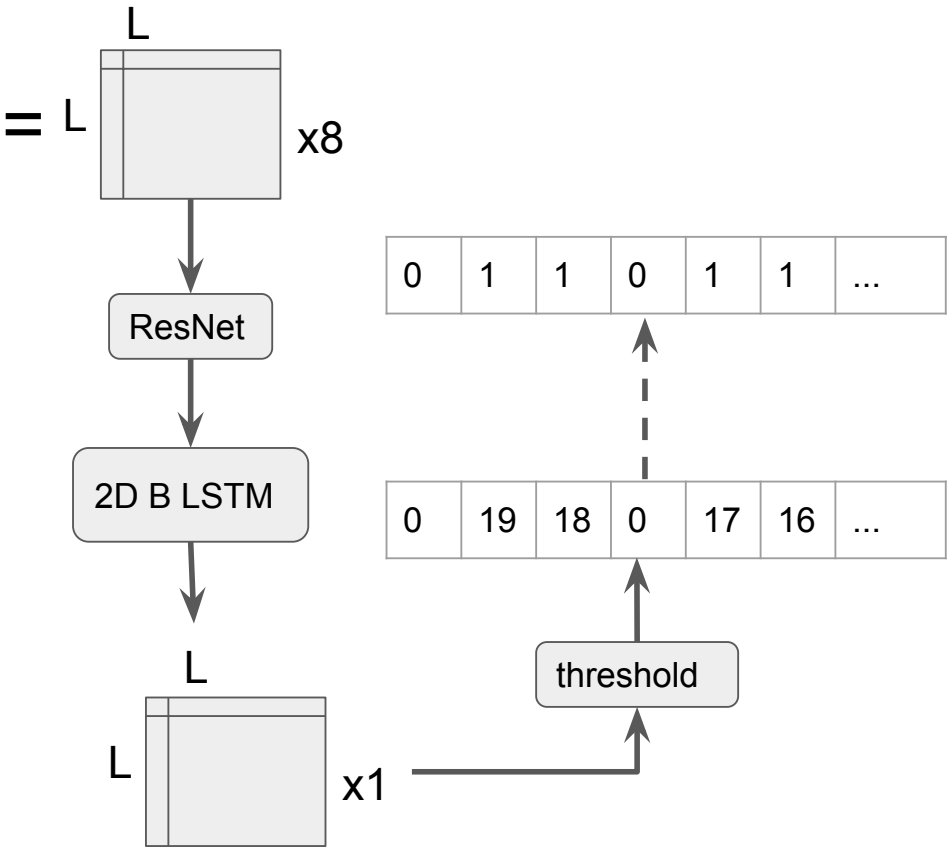
LSTM / CNN



CNN from Zhang H, Zhang C, Li Z, Li C, Wei X, Zhang B and Liu Y (2019)

ResNet & 2D BLSTM

	A	C	U	...
A	[1,0,0,0,1,0,0,0]	[0,1,0,0,1,0,0,0]	[0,0,0,1,1,0,0,0]	
C		[0,1,0,0,0,1,0,0]	[0,0,0,1,0,1,0,0]	
U			[0,0,0,1,0,0,0,1]	
...				

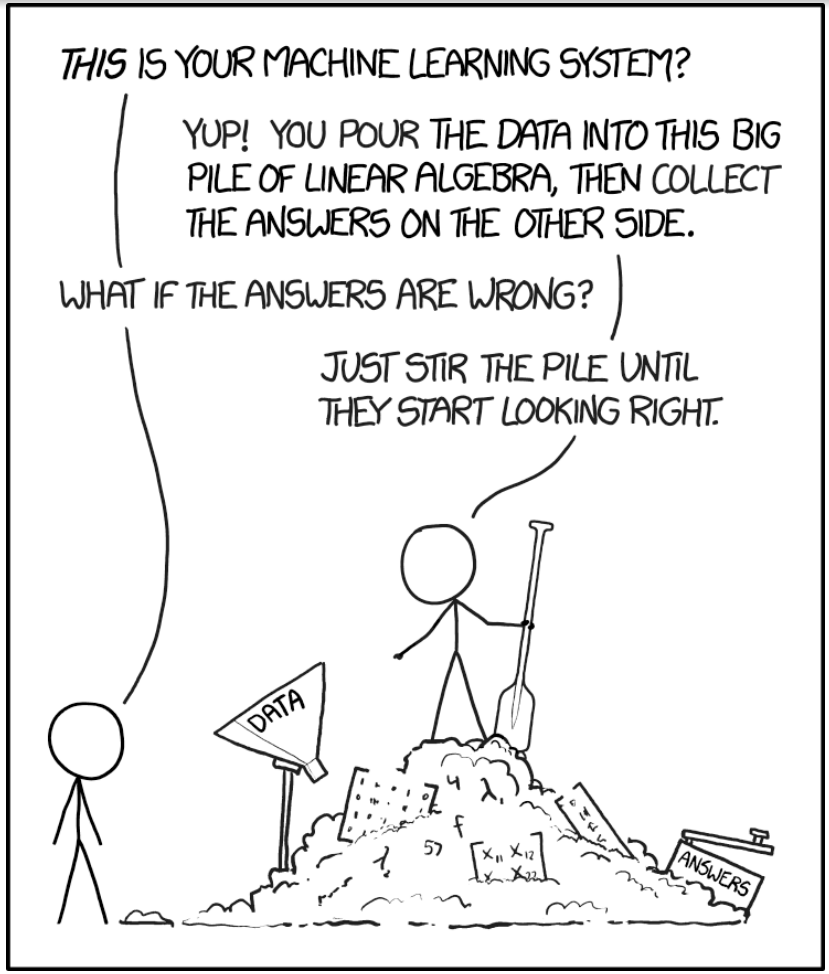


Implementation based on Singh, J., Hanson, J., Paliwal, K. et al. (2019)

Acknowledgements:

- Christoph Flamm
- Ivo Hofacker
- Michael Wolfinger

- Gregor Entzian
- Irene Katharina Beckmann
- Maria Waldl



https://xkcd.com/1838/

TBI

universität wien

Supplement 1 (Metrics)

	Actually Positive	**Actually negative**	
Predicted Positive	True positive (T_P)	False positive (F_P)	Precision/Positive Predictive Value: T_P / T_P+F_P
Predicted Negative	False negative (F_N)	True negative (T_N)	Negative Predictive Value: T_N / T_N+F_N
	Sensitivity: T_P / T_P+F_N	Specificity: T_N / T_N+F_P	

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N}$$

$$MCC = \frac{T_P \times T_N - F_P \times F_N}{\sqrt{(T_P + F_P)(T_P + F_N)(T_N + F_P)(T_N + F_N)}}$$