Experiments in Deep Learning for RNA Secondary Structure Prediction

Ivo L. Hofacker

Institute for Theoretical Chemistry
and
Research Group Bioinformatics and Computational Biology
University of Vienna

Benasque, August 2022

tbi

Data is the stumbling block for deep learning in computational biology

Deep neural networks (DNNs) are data hungry because of the huge number of trainable parameters

- We may not have enough data for training complex DNNs
- Available data are heavily biased towards a few model systems

Using synthetic data

If we can artificially generate synthetic data we can:

- Generate and train on arbitrarily large unbiased data sets
- Find out what is easy / hard / impossible to learn by DNNs
- Identify which network architectures work best
- Ask how much data is needed for training
- Test how biases in training effect the DNNs ability to generalize
- Pre-train networks before final training on real data

Synthetic data need not be accurate – they only need to reflect the complexity of the real problem!

RNA secondary structure

Problems with RNA secondary structure prediction by energy minimization:

- Parameters from limited number of experiments
- Ignores pseudo-knots, ignores tertiary structure
- Ignores non-nearest neighbor effects
- Poor treatment of non-canonical base pairs
- Poor performance on long-range base pairs
- ...

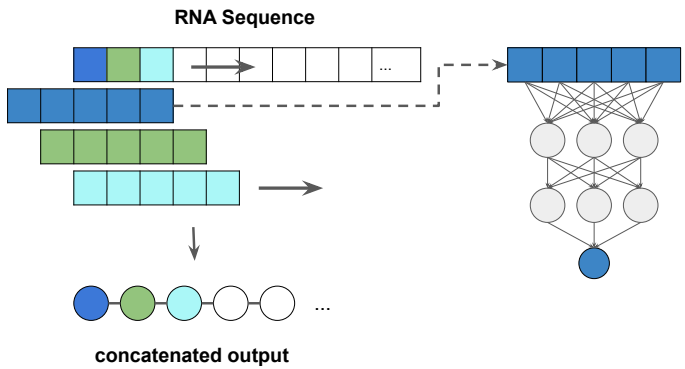
Still good enough for tests with synthetic data

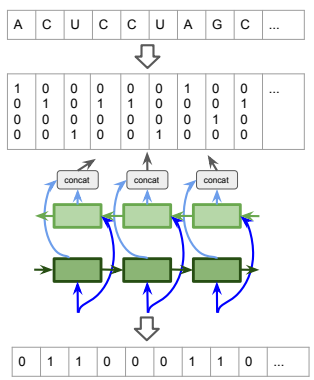
If DNNs can not be trained to emulate RNAfold, then they won't be able to solve the "real" RNA structure prediction problem.

A simplified problem: Predicting pairedness

Try to predict which nucleotides are paired / unpaired

- Much smaller solution space → should be easier
- Similar to protein secondary structure prediction
- Tested several networks:
  - fully connected feed-forward network on a sliding window
  - 1D-convolutional network on sliding window
  - bi-directional long-term short-term memory (BLSTM) network
- train on 80 000 random sequences with RNAfold structures
  test on 20 000 independent random sequences





Predicting pairedness

Modeltype	Parameters	Epochs	Accuracy	F1	MCC
BLSTM	1 Layer, 40 Neurons	43	0.667	0.594	0.166
	1 Layer, 80 Neurons	27	0.664	0.589	0.168
	3 Layers, 40 Neurons	38	0.676	0.609	0.207
FCFF	Window 15	89	0.654	0.559	0.120
	Window 35	94	0.659	0.559	0.118
	Window 71	59	0.661	0.569	0.118
1D-CNN	Window 15	67	0.660	0.588	0.156
	Window 35	65	0.666	0.586	0.166
	Window 71	30	0.668	0.580	0.170

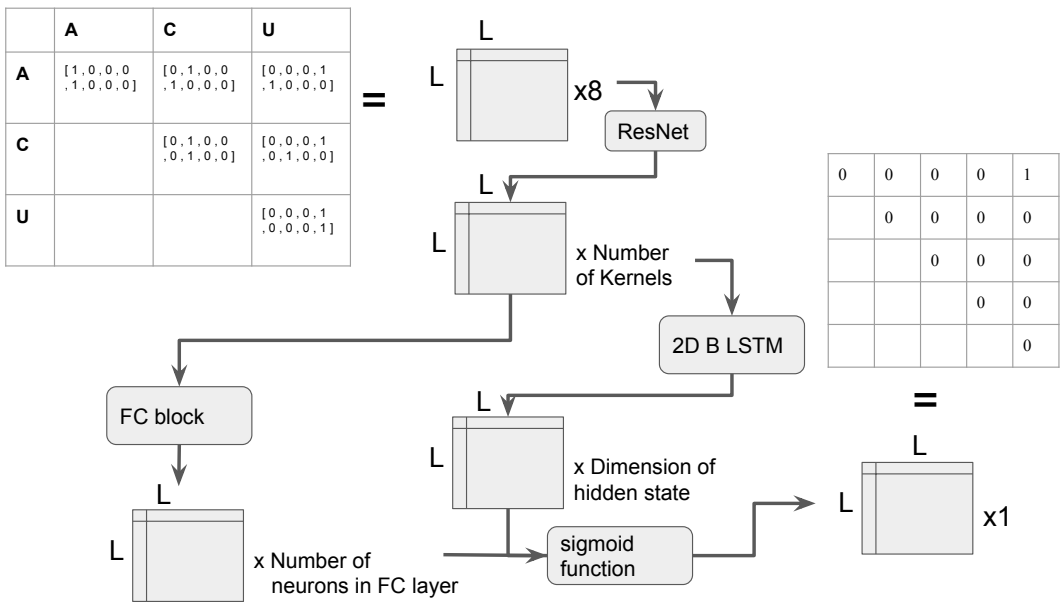
BLSTM slightly better, but poor performance by all architectures

Probable reason: Predicting pairedness is not simpler than predicting the full secondary structure right

Predicting pair matrices

Approach chosen by most recent structure prediction DNNs

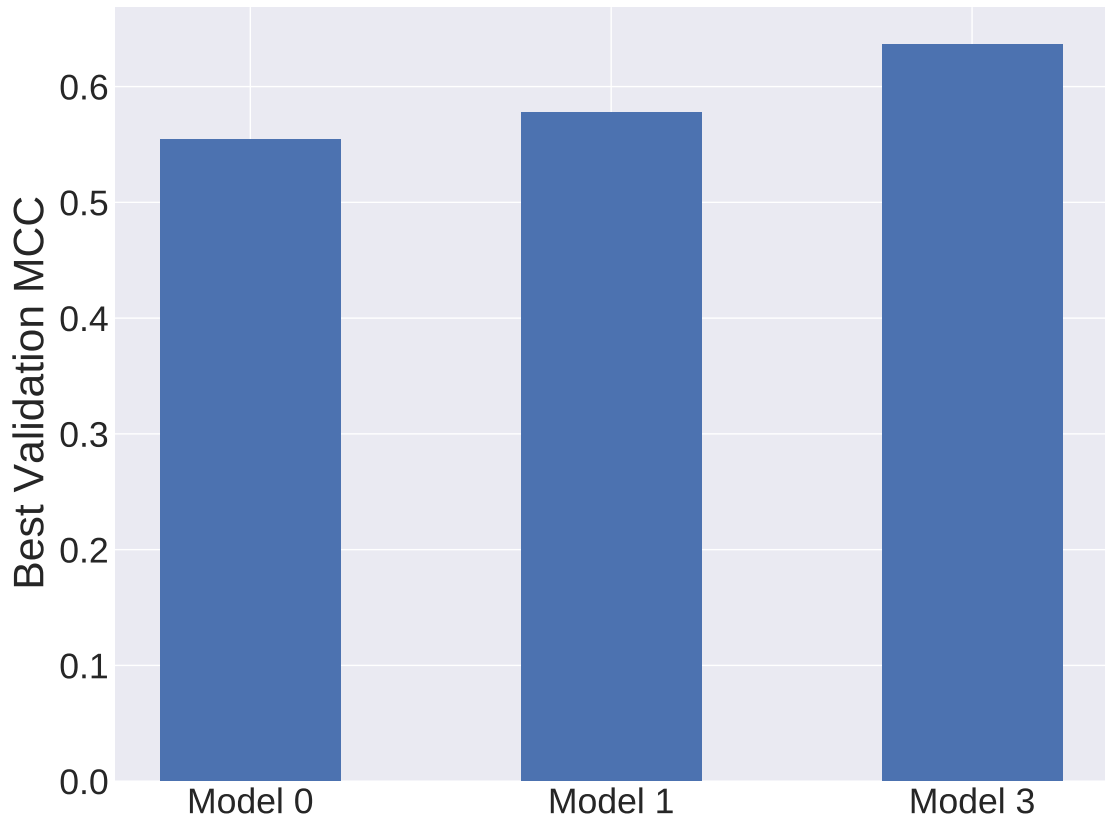
- Represent sequence by an $n \times n$ matrix of all possible pairs
- Use 2D convolutional networks on these matrices
- We chose the SPOT-RNA architecture [1]
- Minimal post-processing: values > 0.5 represent base pair, greedily remove base triples and pseudo-knots.



[1] Singh et al., Nat. Commun., 2019

Model Performance

Training and validation sets with fixed length $n = 70$



What features are easy / hard to learn?

Do the structures predicted by the DNN look statistically similar to the RNAfold ground truth?

Relative frequency of base pair types)							
model / length	GC	CG	AU	UA	GU	UG	NC
VRNA / 70	0.257	0.262	0.169	0.170	0.071	0.071	0.00
DNN / 70	0.258	0.260	0.170	0.172	0.070	0.070	$9.63 \cdot 10^{-5}$
VRNA / 100	0.262	0.255	0.173	0.170	0.068	0.071	0.00
DNN / 100	0.257	0.252	0.177	0.175	0.068	0.070	$2.30 \cdot 10^{-5}$

Learning the base pair frequencies is no problem at all

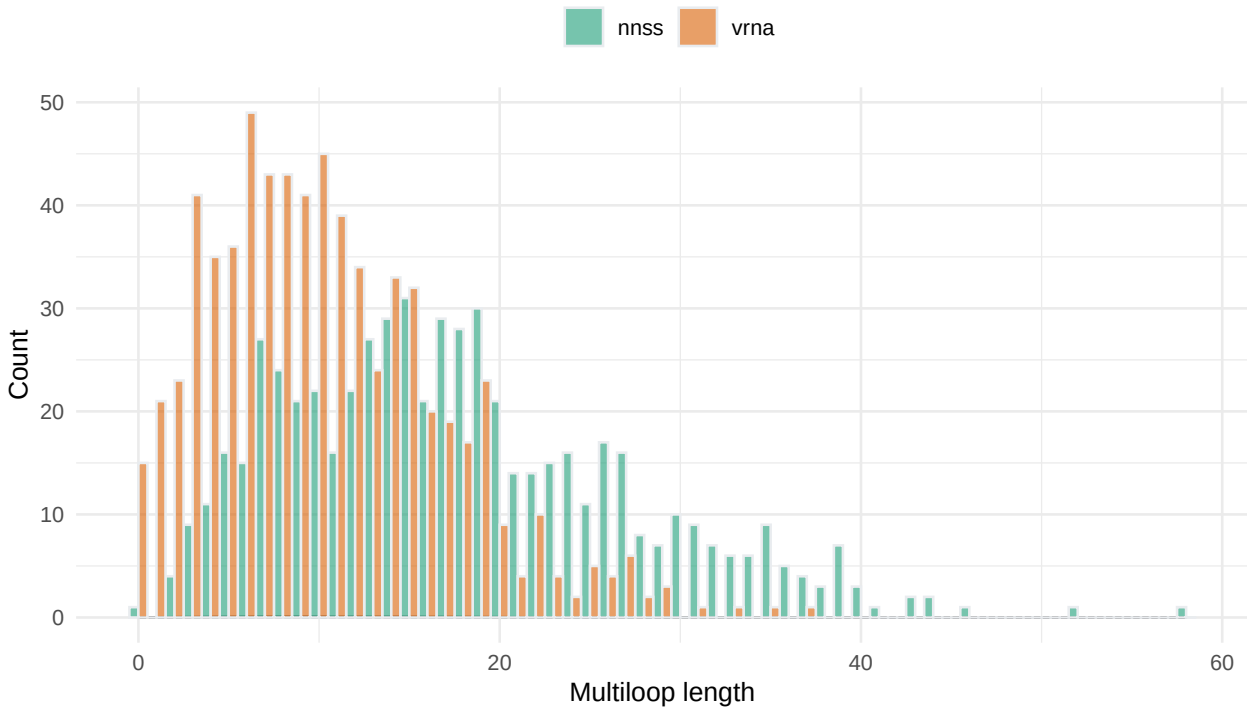
What features are easy / hard to learn?

How about different loop types?

Frequency of bases in context						
model / length	paired	EL	BL	HL	IL	ML
VRNA / 70	0.508	0.176	0.033	0.156	0.114	0.014
DNN / 70	0.445	0.222	0.027	0.161	0.127	0.019
VRNA / 100	0.541	0.123	0.031	0.143	0.126	0.035
DNN / 100	0.433	0.185	0.030	0.146	0.152	0.053
Average number of structural element						
model / length	helix	EL	BL	HL	IL	ML
VRNA / 70	4.825	0.992	1.112	1.754	1.841	0.118
DNN / 70	4.354	0.993	0.840	1.730	1.686	0.098
VRNA / 100	7.132	0.991	1.586	2.314	2.889	0.343
DNN / 100	6.146	0.991	1.080	2.135	2.632	0.299

- Frequency and size of hairpin and interior loops match very well
- Network produces fewer, but larger multi-loops
- Number of base pairs does not match

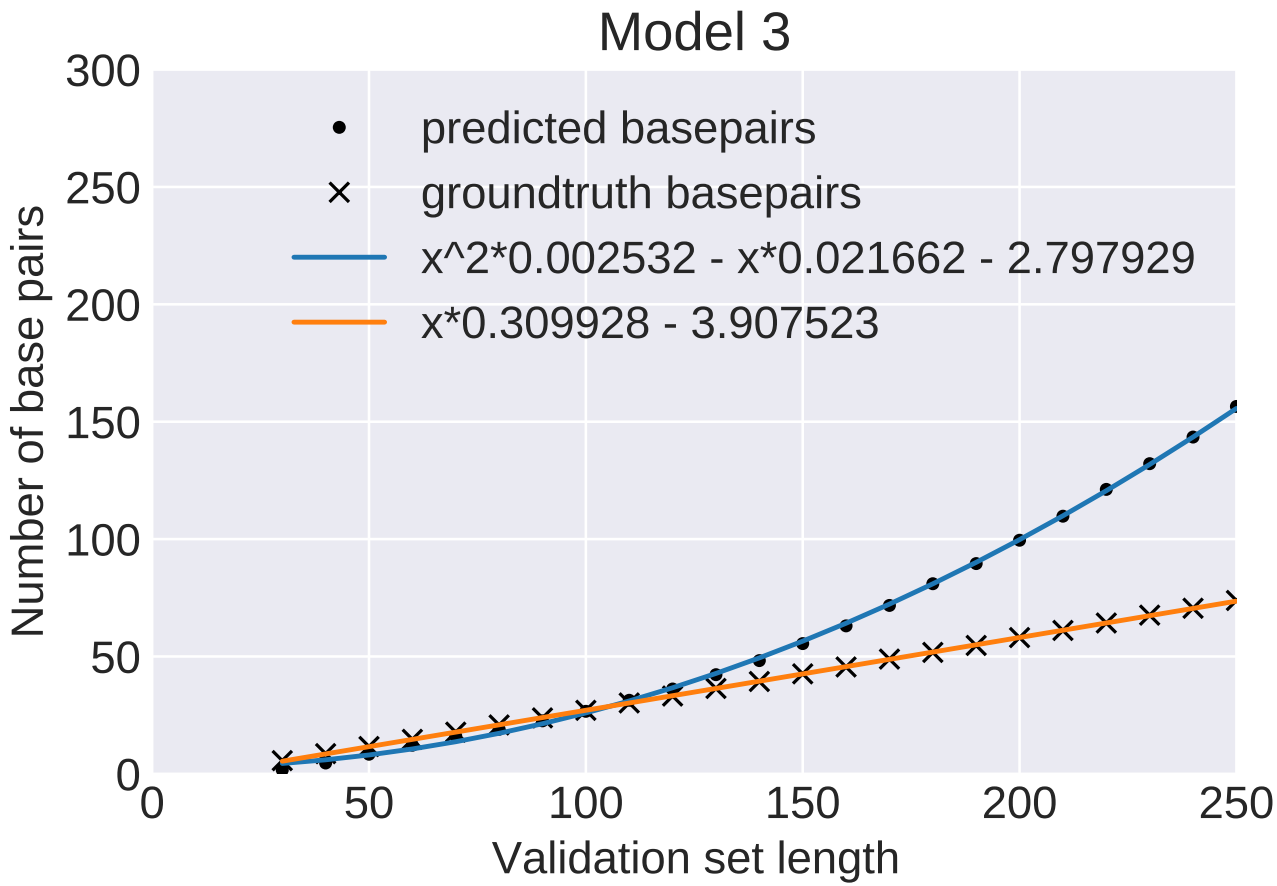
Multi loop length



Pseudo-knots and base triples

- DNNs have no problem predicting pseudo-knots and base triples – should be an advantage
- But can they learn, to predict the right amount?
- Our ground truth has no PKs, no base triples
- At length 100:
  Almost 50% (975/2000) of structures contain a PK
  75% (1512/2000) contain multi-pairs

What features are easy / hard to learn?



- An RNA structure can contain $< \frac{n}{2}$ pairs
- DNNs working on pair matrices naturally predict a quadratic growth

What features are easy / hard to learn?

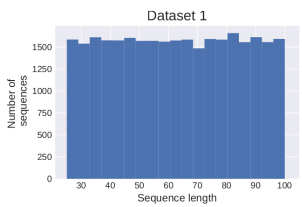
Convolutional networks always focus on local features of the matrix
Therefore:

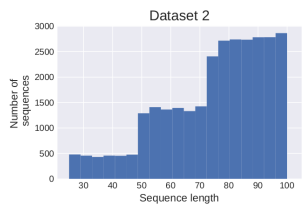
- *local* features (e.g. interior loops) are easiest to get right
- larger features (e.g. multi loops) are harder
- *global* properties (total number of pairs) are hardest to learn

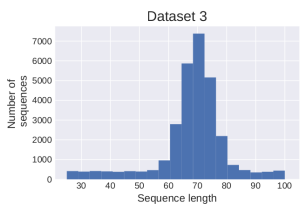
Biases in the training set

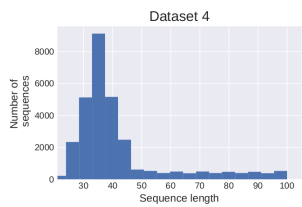
How do biases in the training set affect performance?

Example different length distributions:







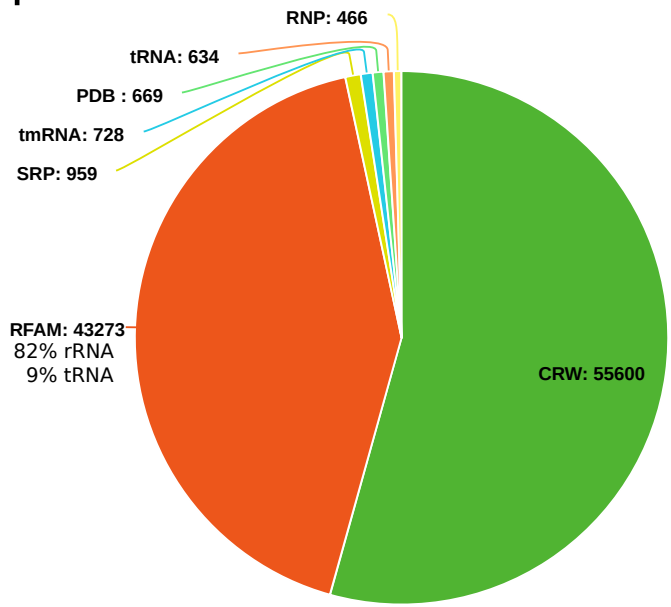


Validation	Training			
	1	2	3	4
1	0.64	0.61	0.64	0.63
2	0.59	0.58	0.60	0.57
3	0.61	0.59	0.62	0.59
4	0.71	0.68	0.70	0.75
train	0.72	0.66	0.71	0.87

The bpRNA dataset

Do we have large enough data sets? Use bpRNA[2]!

- over 100 000 sequence/structure pairs
- Collected from 7 databases
- CRW: SSU, LSU, and 5S rRNAs
- RFAM: 2588 families
  but 82% rRNA, 9% tRNA



→ Many sequences, but low structural diversity!

How does this effect prediction accuracy on RNA not from these families?

[2] Danaee et al., NAR 2018

The bpRNAinv data set

How can we mimic training data with low structural diversity using synthetic data?

- Take all structures from bpRNA
- Remove pseudo-knots, restrict to $n \leq 120$, ≤ 6 pairs in PKs
- For each structure design a sequence using RNAinverse
- Resulting data set has same structure distribution as bpRNA
- Sequences are completely unrelated to each other

After training with these data set, test performance on two test sets:

1. A test set produced in the same way by RNAinverse
   unrelated sequences, but same structure bias
2. Di-nucleotide shuffling of training sequences
   same sequence composition, but unrelated structures

The bpRNAinv data set

How can we mimic training data with low structural diversity using synthetic data?

- Take all structures from bpRNA
- Remove pseudo-knots, restrict to $n \leq 120$, ≤ 6 pairs in PKs
- For each structure design a sequence using RNAinverse
- Resulting data set has same structure distribution as bpRNA
- Sequences are completely unrelated to each other

After training with these data set, test performance on two test sets:

1. A test set produced in the same way by RNAinverse
   unrelated sequences, but same structure bias
2. Di-nucleotide shuffling of training sequences
   same sequence composition, but unrelated structures

The bpRNAinv data set

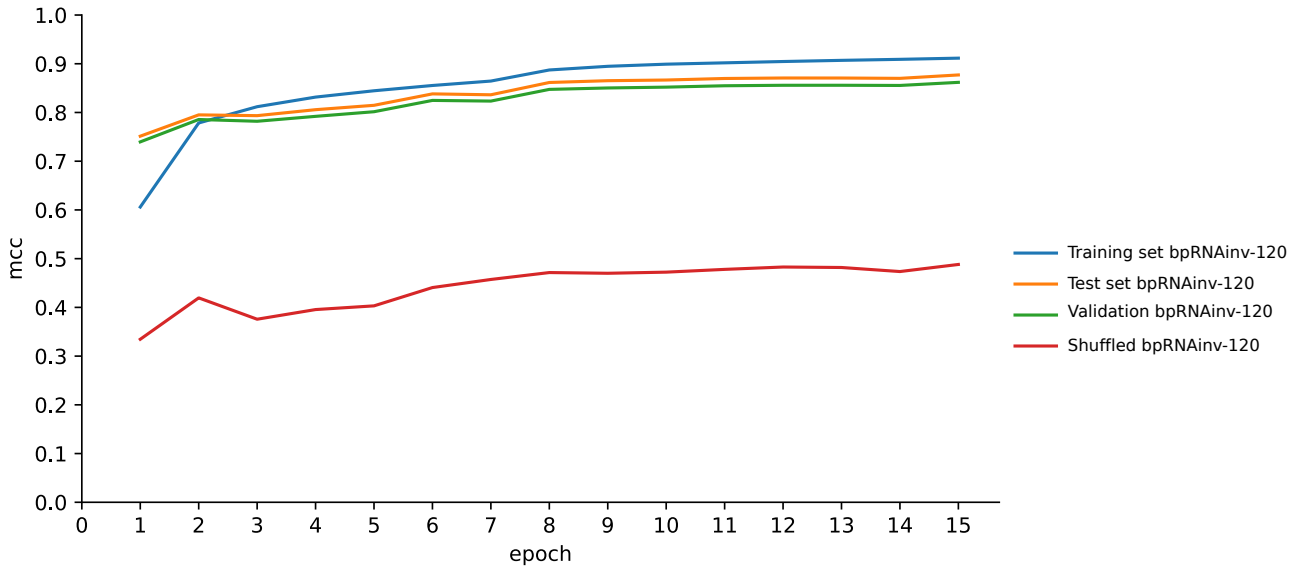
How can we mimic training data with low structural diversity using synthetic data?

- Take all structures from bpRNA
- Remove pseudo-knots, restrict to $n \leq 120$, ≤ 6 pairs in PKs
- For each structure design a sequence using RNAinverse
- Resulting data set has same structure distribution as bpRNA
- Sequences are completely unrelated to each other

After training with these data set, test performance on two test sets:

1. A test set produced in the same way by RNAinverse
   unrelated sequences, but same structure bias
2. Di-nucleotide shuffling of training sequences
   same sequence composition, but unrelated structures

Performance with the bpRNAinv data set



- Performance on the inverse folded test set almost as good as training set
- Poor performance when structures are dis-similar to training structures

Take home lessons

- Synthetic data allow to test the capabilities of DNNs with full control over biases
- DNNs for RNA secondary structure prediction can easily learn about local structure features, but struggle with non-local or global features
- Current architectures generalize well to novel *sequences* as long as structures are covered by the training set
- Poor generalization to RNA with novel *structures*

Acknowledgements

- **Julia Wieland**
- **Stefan Badelt**
- Christoph Flamm
- Michael T. Wolfinger
- Ronny Lorenz

More at: Flamm et al., *Frontiers in Bioinformatics* (2022).
doi: 10.3389/fbinf.2022.835422