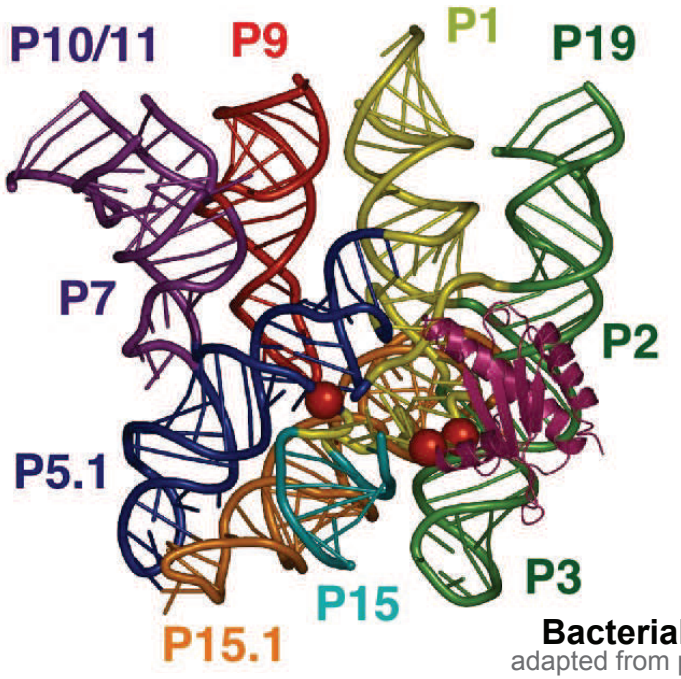
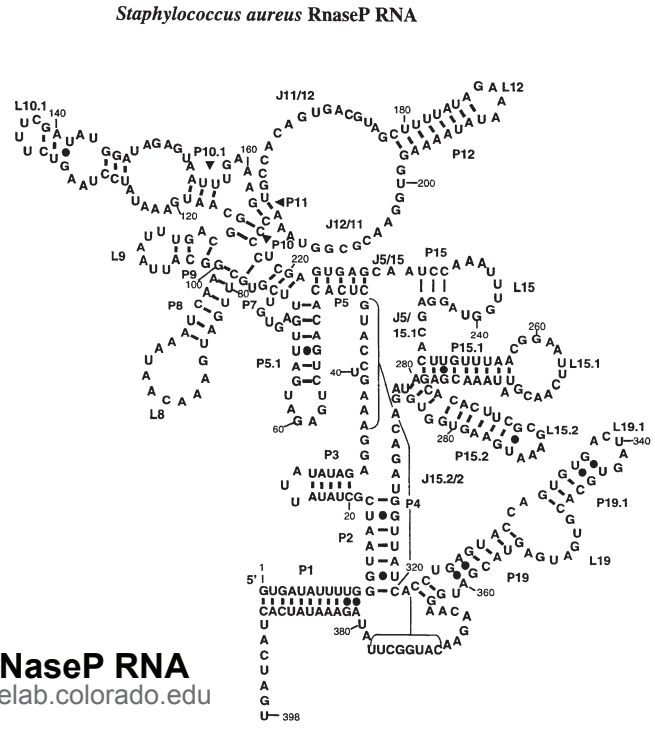
Minimal-AI:
Unsupervised learning of the rules of RNA base pairing

What does it take to learn the rules of RNA base pairing?





Bacterial RNaseP RNA
adapted from pacelab.colorado.edu

helical structures

canonical base pairs
A:U G:C G:U
U:A C:G U:G

A lot less than you may think

Jay Pratap, Ryan Krueger, ER; Commun. Biol., 2026

To learn the rules of RNA base pairing

Deep models with **millions** of parameters

Minimal **model**?

Trained on **100 thousands** RNA seq/structures

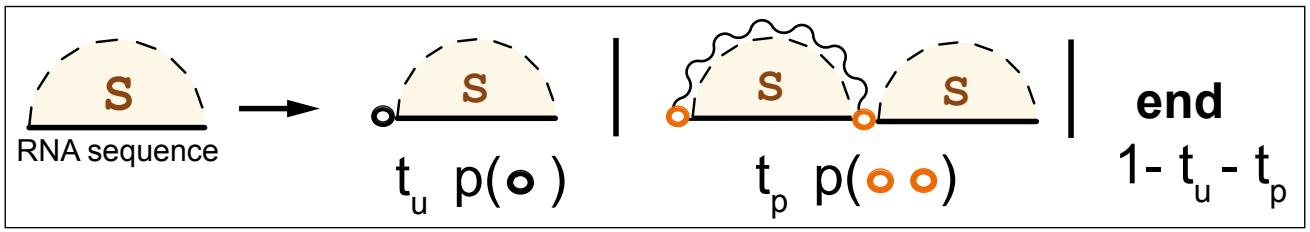
Minimal **inputs**?

Minimal **model**

Thermodynamic models	Probabilistic models
`Nearest Neighbors NN` `ViennaRNA, RNAfold`	**Stochastic Context-free grammars (SCFGs)**
minimum free energy	maximum probability
hard to train	easy to train (from structures)
$\sim$ 2k parameters, fixed	flexible (from 20 to thousands)
	TORNADO, NN SCFG mimics

G5 Grammar (Ivo Hofacker)

```
S -> o S
S -> o S o S
S -> end
```



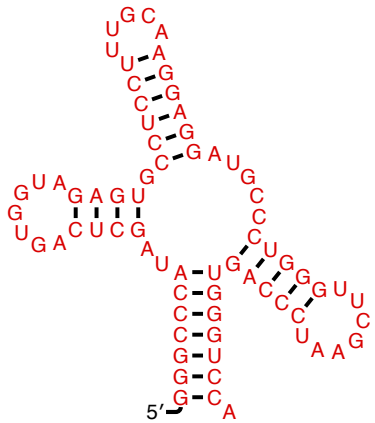
Eddy&Durbin, Sakakibara&Haussler, 1994
Dowell & Eddy, 2004

Minimal training **inputs**

RNA secondary structures?

structural information

X-ray, NMR, CryoEM



		A	C	G	U
5'	A	0	0	0	1
	C	0	0	10	0
	G	0	5	0	2
	U	4	0	0	0

3'

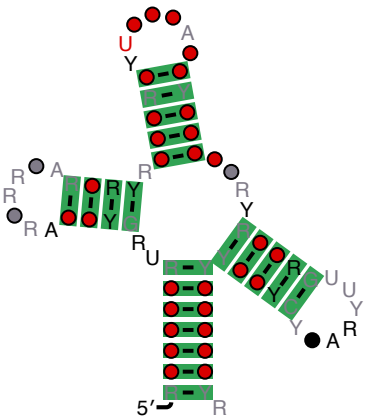
maximum likelihood training

Contrafold, Do *et al.*, Bioinf., 2006
TORNADO, Rivas&Eddy, RNA, 2012

RNA alignments?

evolutionary information

genomes, homology



Infernal,
Eddy, Kolbe, Nawrocki, 2013

R-scape, Rivas,Clements,Eddy, Nat Meth, 2017

RNA sequences only??

5'-GGGCCCAUAGCUCAGUGGUAGAGUGCCUCCUUUGCAAGGAGGAUGCCCUGGGUUCGAAUCCCAGUGGGUCCA

Pratap, Krueger, Rivas, Comms Biol, 2026

The not-deep-at-all model

G6 Grammar

Start

S -> L S

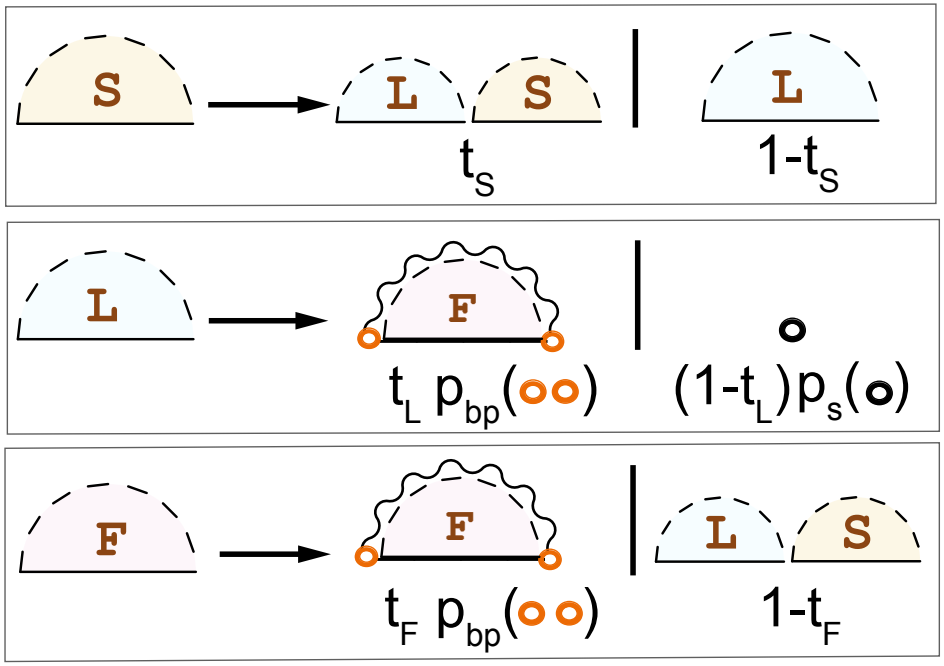
S -> L

L -> o F o new helix starts, adds first pair

L -> o a single residue

F -> o F o helix adds a base pair

F -> L S helix ends



$0 \le t_S, t_L, t_F \le 1$

o an RNA residue

o o an RNA base pair

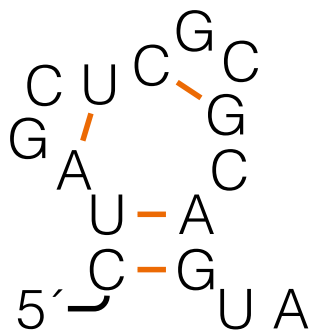
S,L,F non-terminals (NT) transform following one of 2 rules

Knudsen & Hein, NAR, 2003
Dowell & Eddy, BMC Bioinf, 2004

21 free parameters

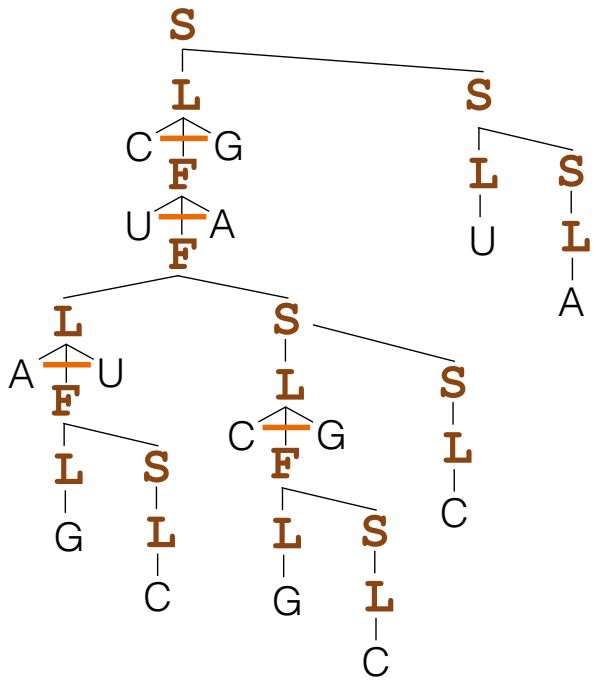
$T_6 = [t_s, t_L, t_F, P_s(a), P_{bp}(a,b)]$

Start			
S -> L S	t_s	
S -> L	$1-t_s$	
L -> o F o	t_L	helix starts
L -> o	$1-t_L$	
F -> o F o	t_F	helix continues
F -> L S	$1-t_F$	helix ends



structure2

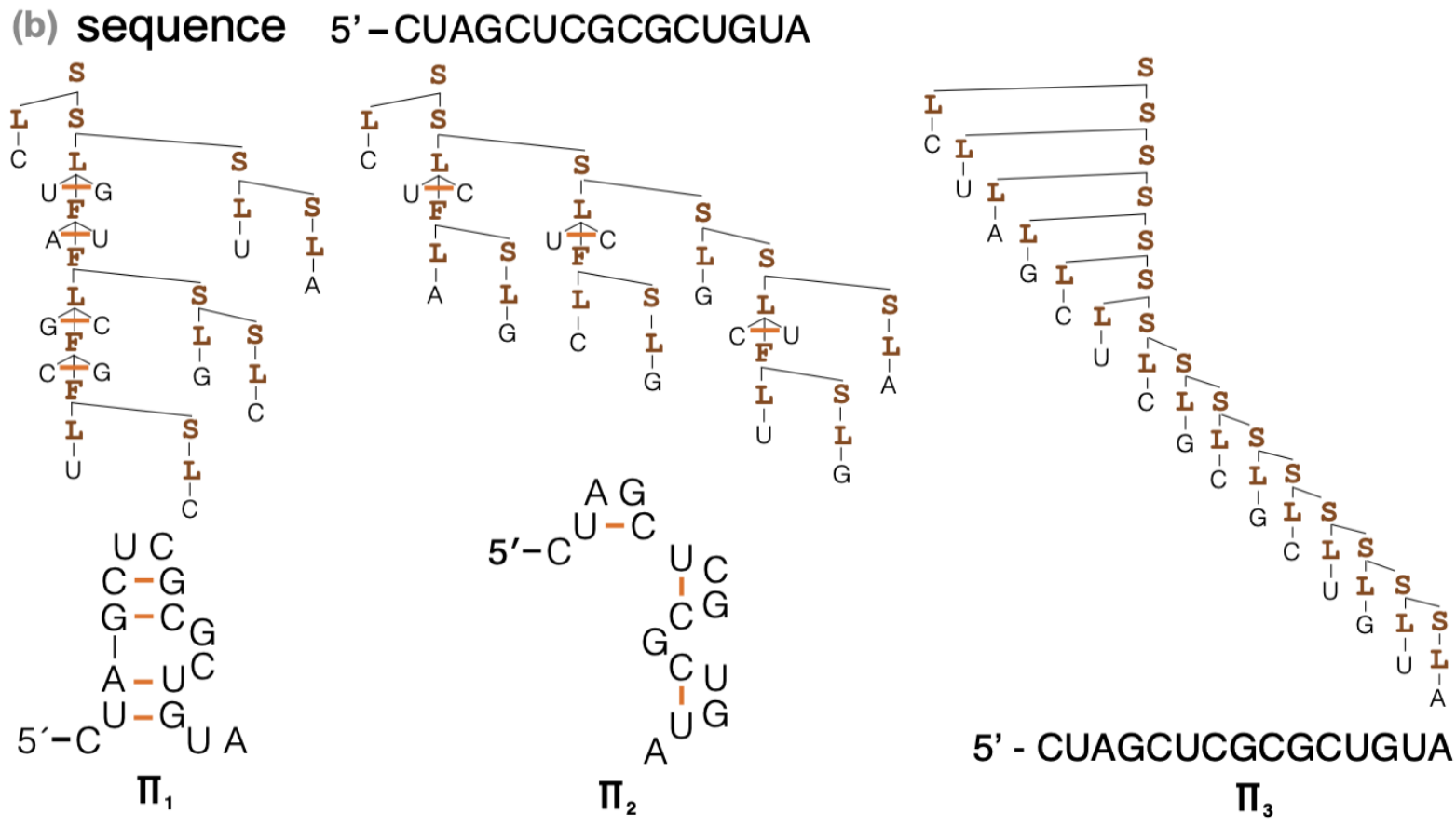
5’ - CUAGCUCGCGCAGUA

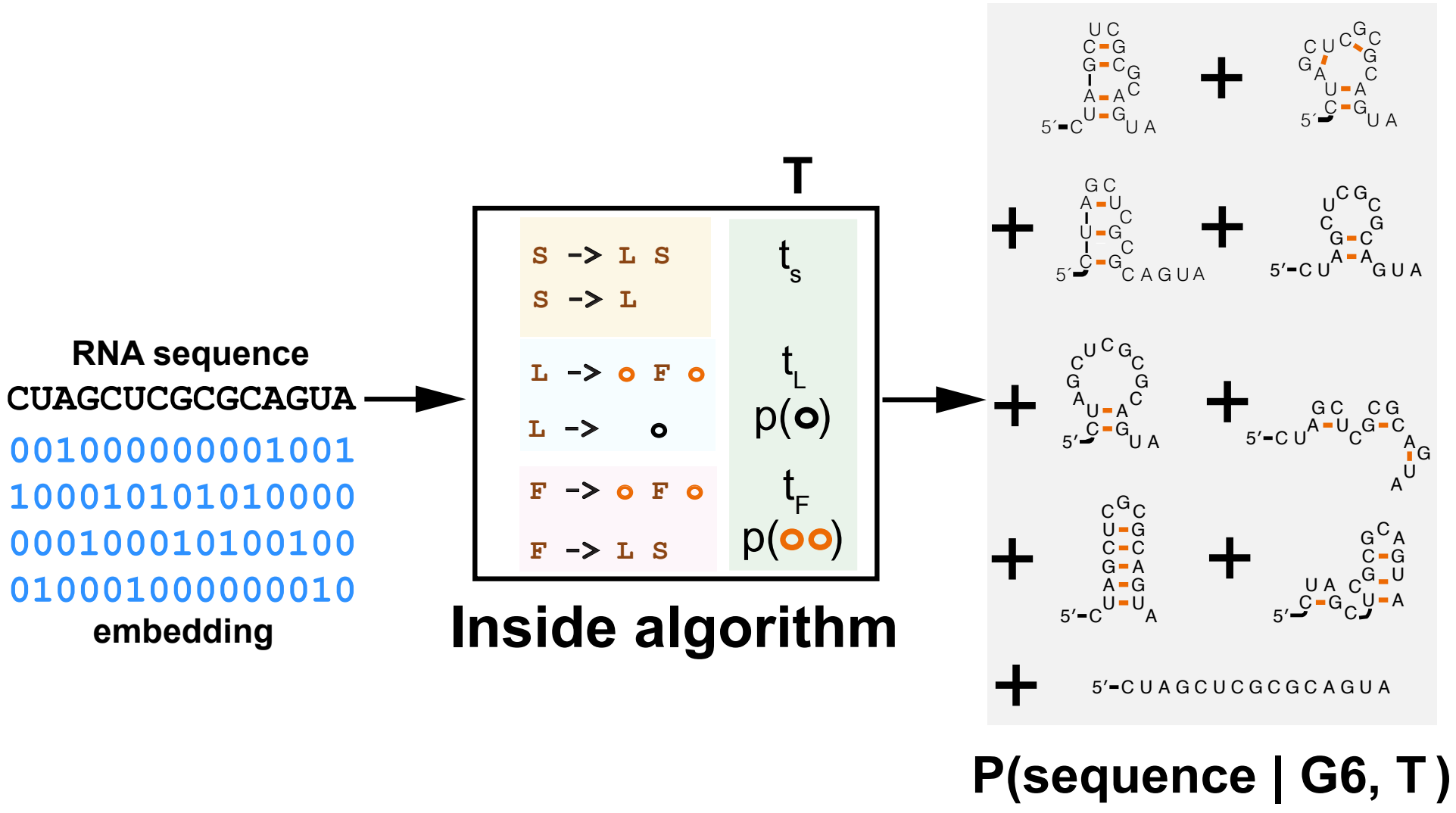


P(seq, structure2 | G6, T_6)

For this minimal experiment, we selected the G6 grammar [9], which has only 21 free probability parameters. The G6 model, first introduced in the method PFOLD [17], is described in Figure 1. The G6 grammar provides rules to generate sequences. It's built-in assumptions (implicit inductive biases) are:

- Residues can be either uncorrelated (unpaired) or pairwise correlated (paired).
- The ratio of unpaired to paired residues in a sequence is unconstrained.
- Arbitrary number of unpaired residues can occur interspersed in between any base pair.
- A residue cannot pair with more than one other residue.
- Paired residues can be disjoint or nested but they cannot cross.
- All unpaired residues are generated by the same rule.
- There are two distinct rules for paired residues, allowing (but not forcing) the model to distinguish terminal (or disjoint) base pairs from other base pairs that are nested together forming helices.


(b) sequence 5’ – CUAGCUCGCGCUGUA
5´–C
5’–C
5’ - CUAGCUCGCGCUGUA
π1
π2
π3


RNA sequence
CUAGCUCGCGCAGUA
001000000001001
100010101010000
000100010100100
010001000000010
embedding
T
S -> L S
S -> L
L -> o F o
L -> o
F -> o F o
F -> L S
t_s
t_L
p(o)
t_F
p(oo)
Inside algorithm
5′-CUAGCUCGCGCAGUA
P(sequence | G6, T)

Thermodynamic models

Discriminative models

$P(\pi \mid seq) = \frac{e^{-\Delta G(\pi, seq)}}{Z}$

MacCaskill:
$Z(seq) = \sum_{\pi} e^{-\Delta G(\pi, seq)}$

Probabilistic models

Generative models

$P(seq, \pi \mid SCFG)$

Inside: $P(seq \mid SCFG) = \sum_{\pi} P(seq, \pi \mid SCFG)$

Maximize: $P(dataset \mid SCFG, T) = \prod_{seq \in dataset} P(seq \mid SCFG, T)$

Minimize:
$-\log P(dataset \mid SCFG, T) = -\sum_{seq \in dataset} \log P(seq \mid SCFG, T)$

The Loss

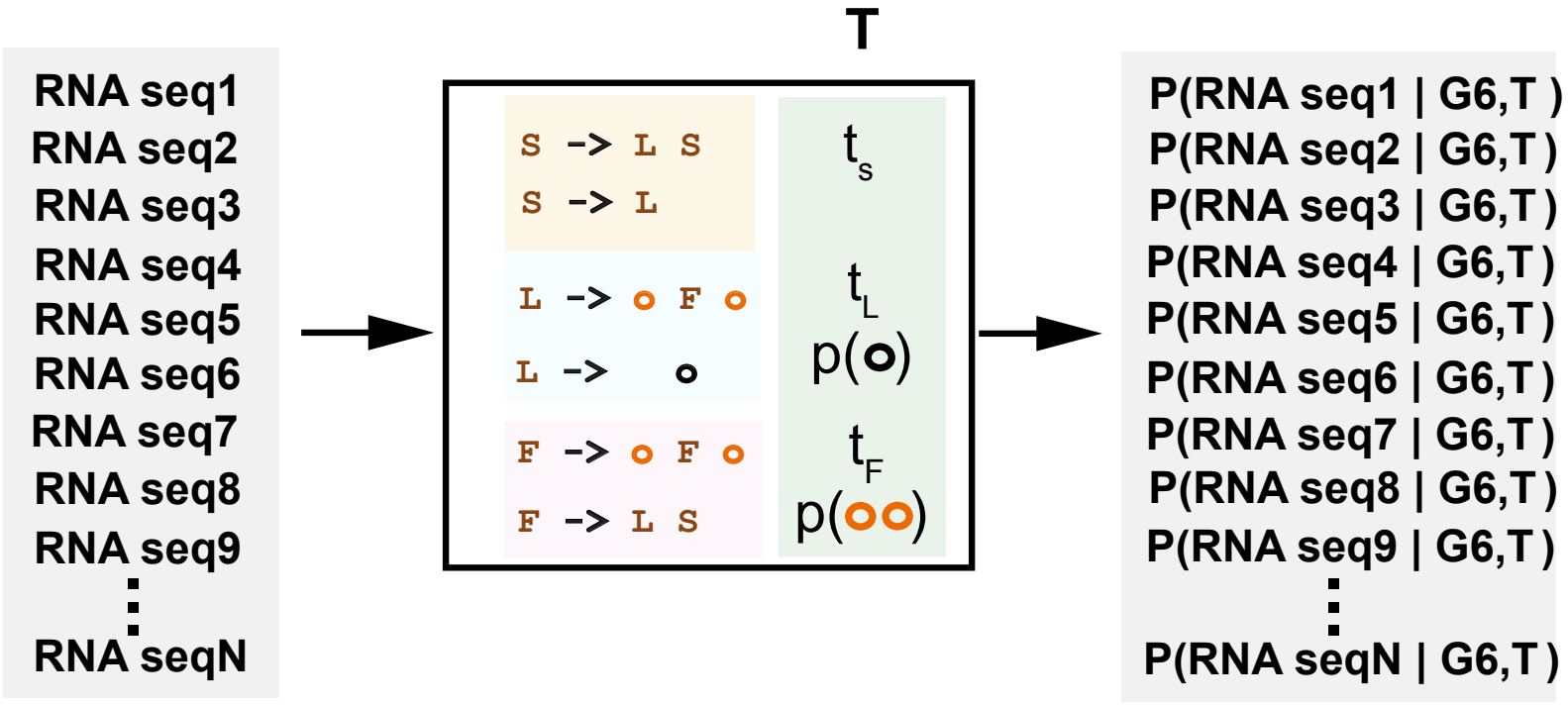
$$\begin{aligned}
\mathbf{min}_T[-\log P(dataset \mid SCFG, T)] &= \mathbf{min}_T[-\sum_{seq \in dataset} \log P(seq \mid SCFG, T)] \\
&= \mathbf{min}_T[-\sum_{seq} P_{data}(seq) \log P(seq \mid SCFG, T)] \\
&= \mathbf{min}_T[\sum_{seq} P_{data}(seq) \log \frac{P_{data}(seq)}{P(seq \mid SCFG, T)}] \\
&= \mathbf{min}_T \, KL(P_{data}(seq) || P(seq \mid SCFG, T))
\end{aligned}$$

Maximizing (in T) $P(seq \mid SCFG, T)$

is equivalent to

Minimizing the Kullback-Leibler (KL) divergence between $P_{data}(seq)$ and $P(seq \mid SCFG, T)$

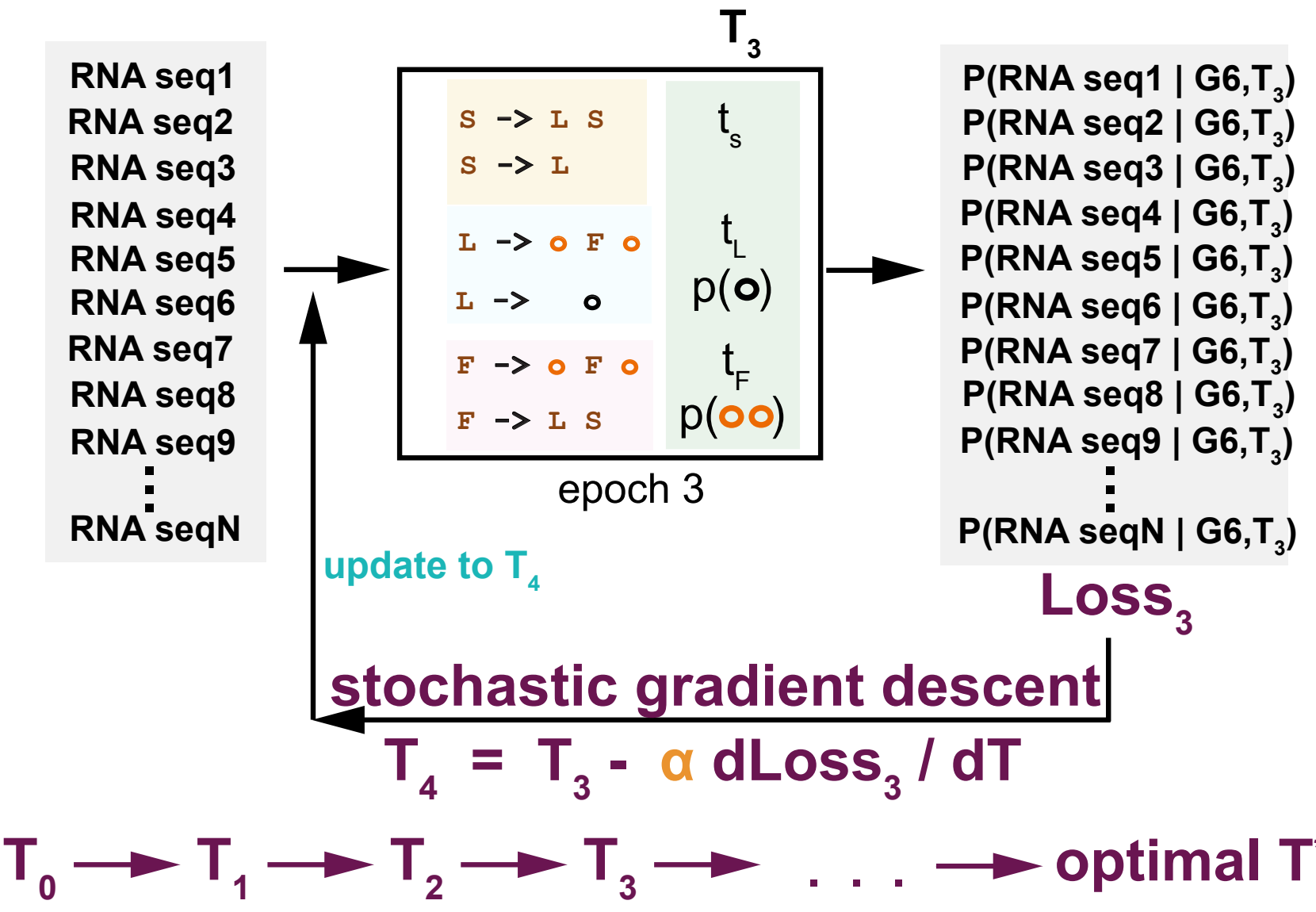
maximize the probability of the training sequences



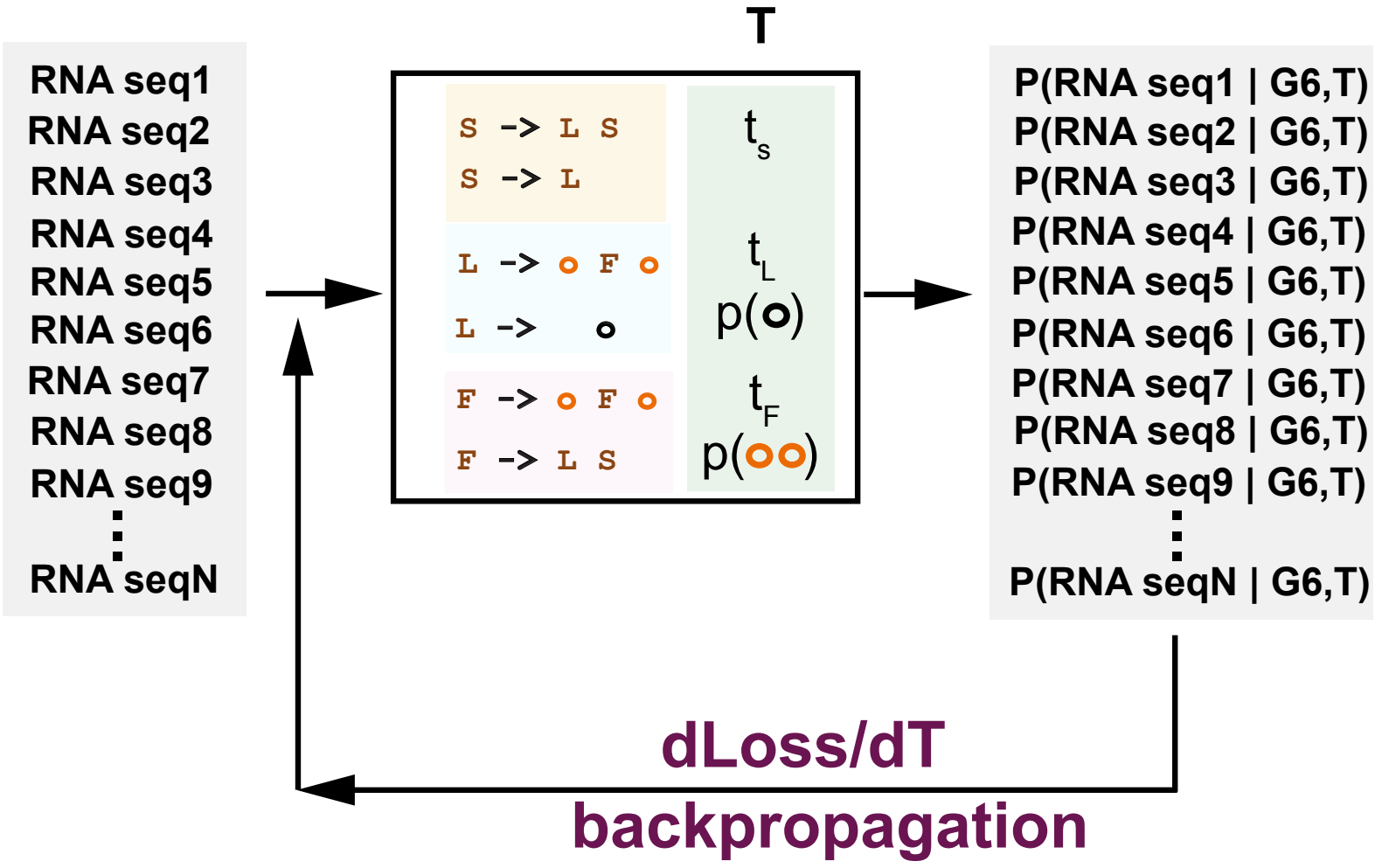
T that minimizes

$$\text{Loss} = -\sum_i \log P(\text{seq}_i \mid \text{G6, T})$$

maximize the probability of the training sequences

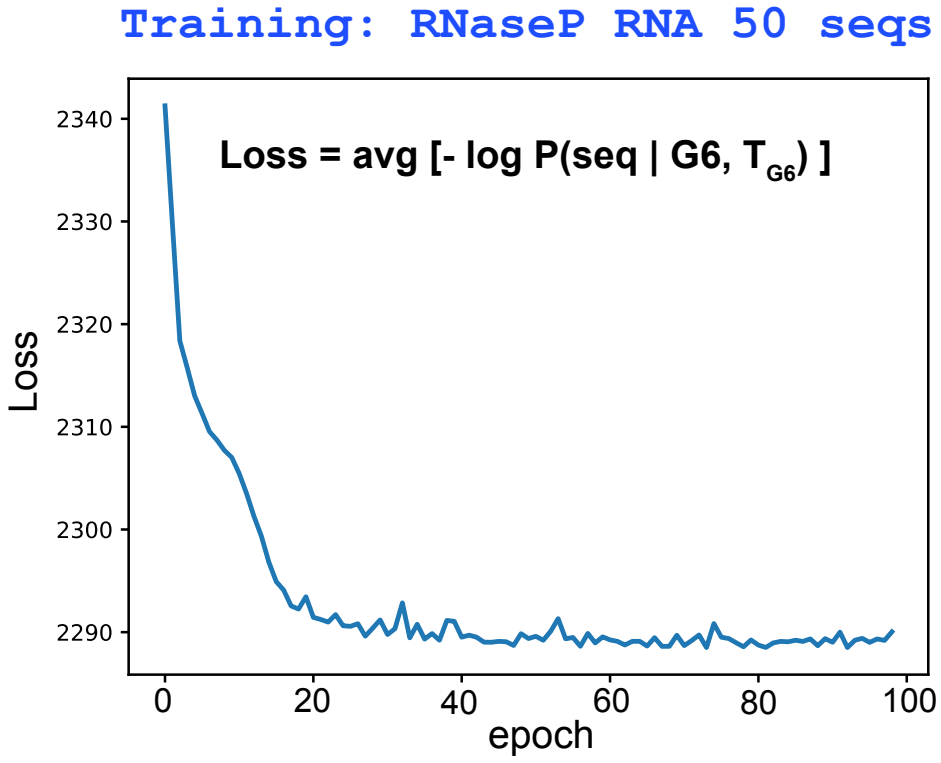


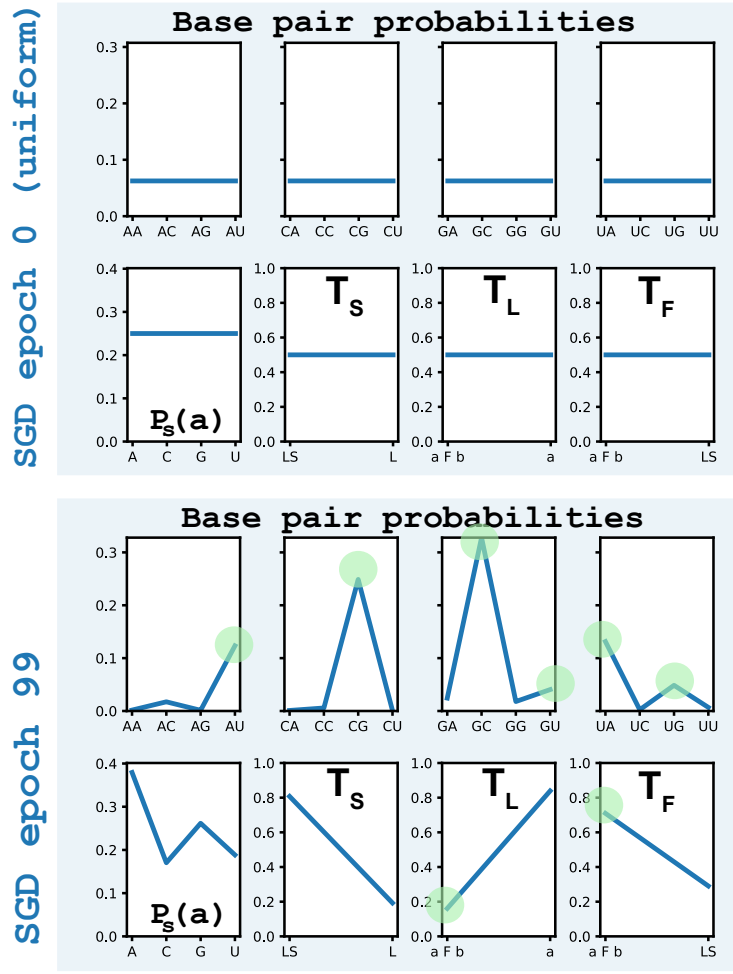
maximize the probability of the training sequences



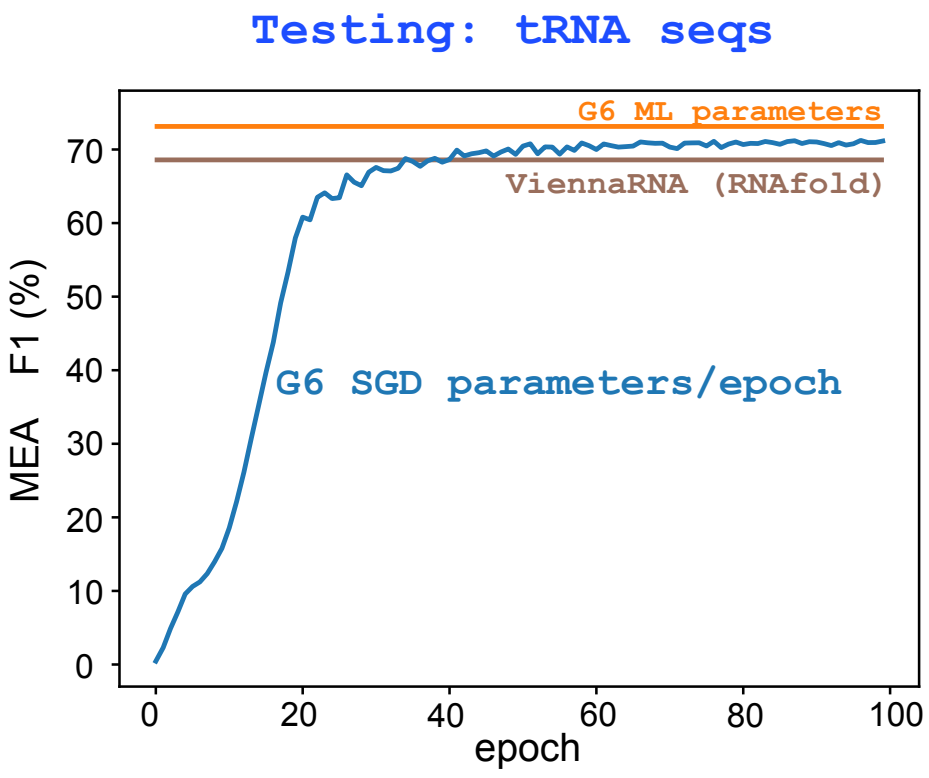
Python/JAX automatic differentiation

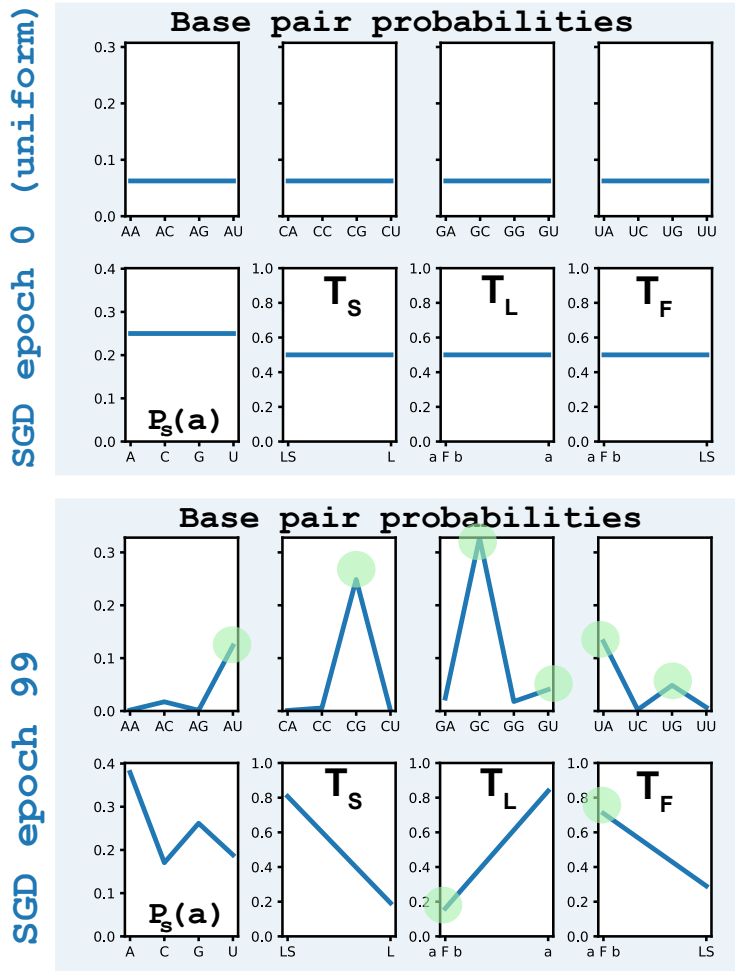
Learns the rule of RNA base pairing





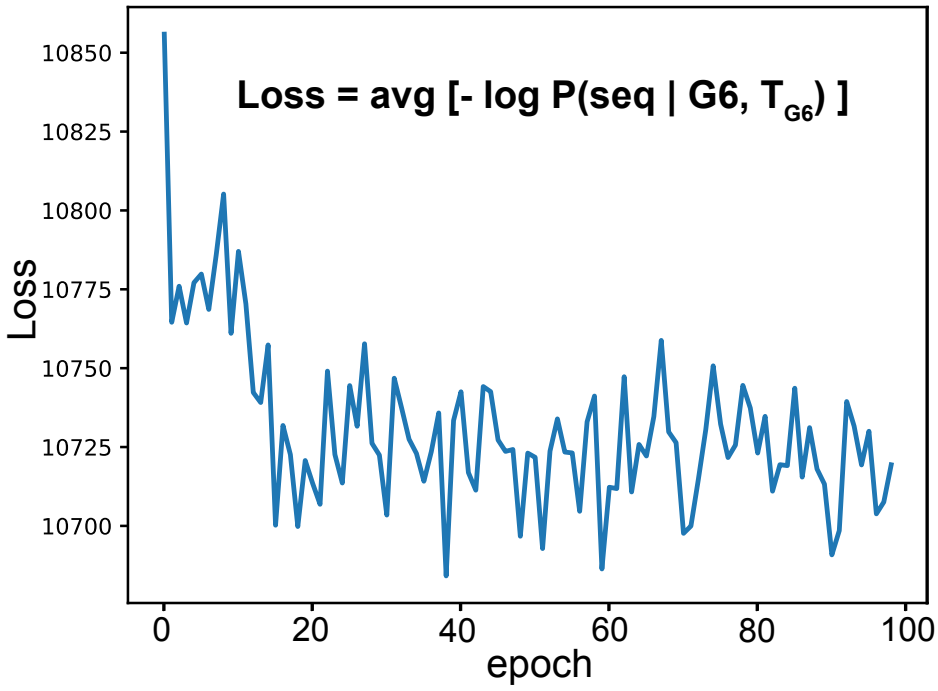
Learns the rule of RNA base pairing



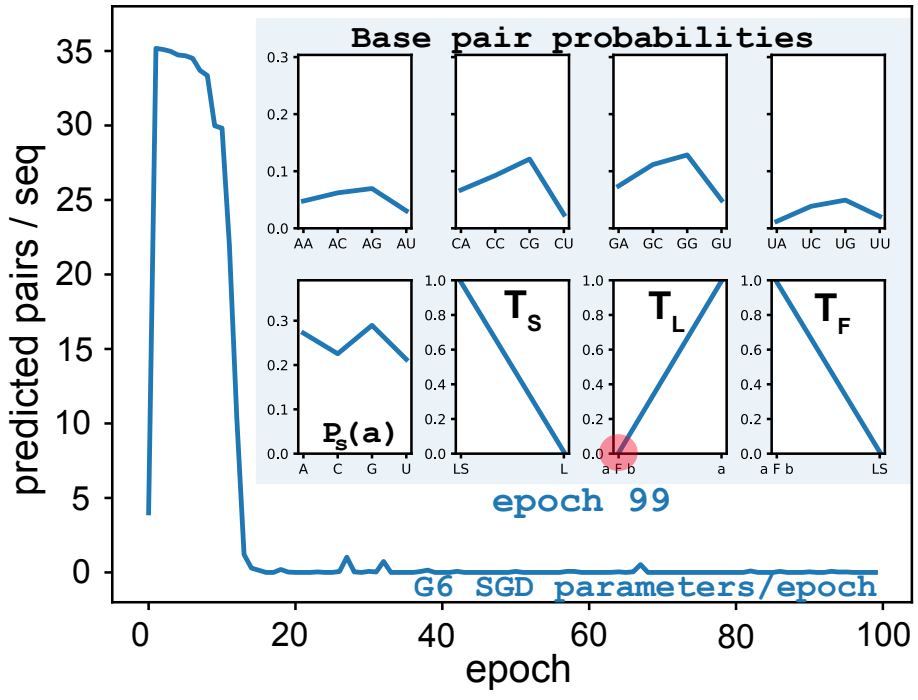


Training on random (shuffled) sequences

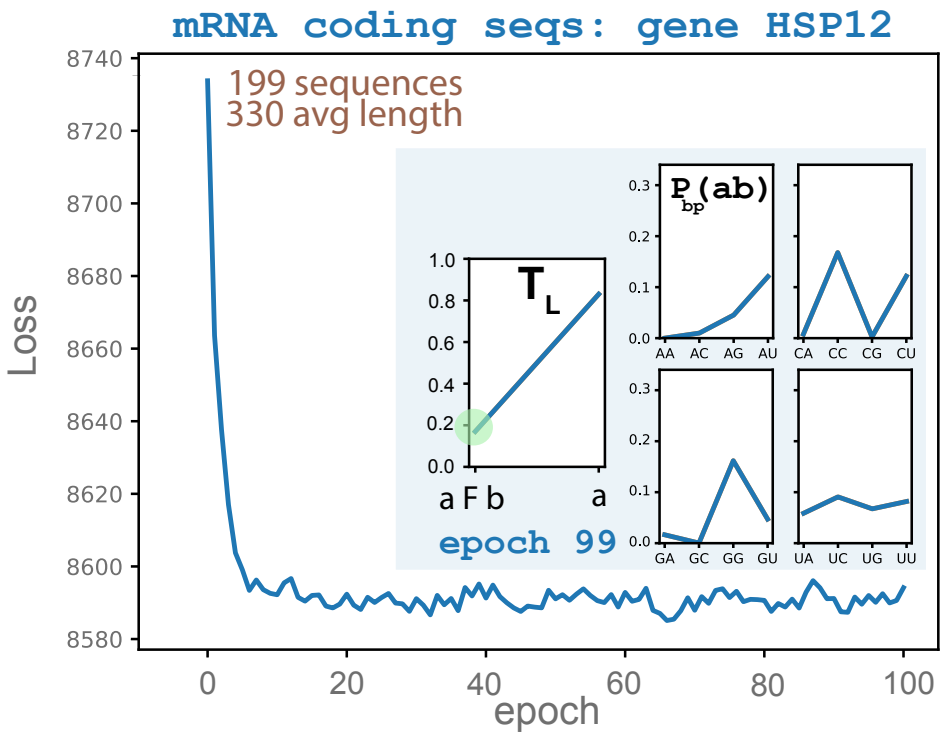
Training:shuffled RNaseP 225 seqs

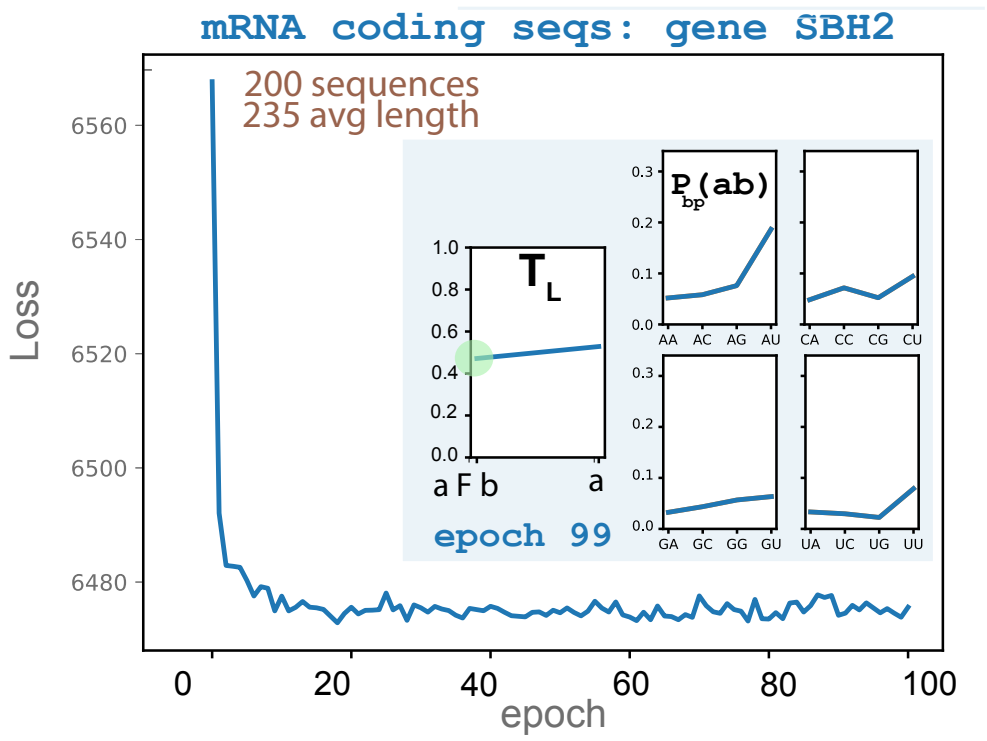


Testing:tRNAs



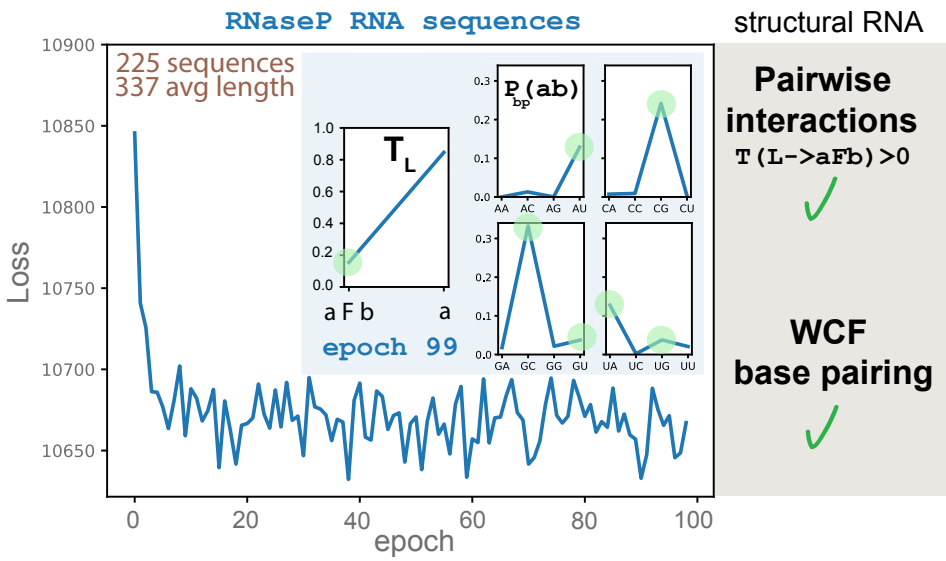
Training on mRNAs

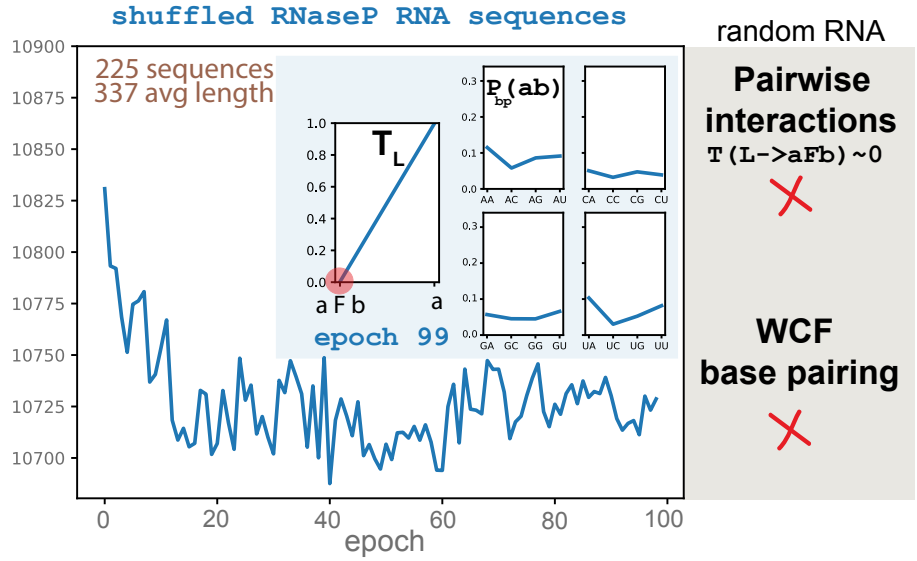


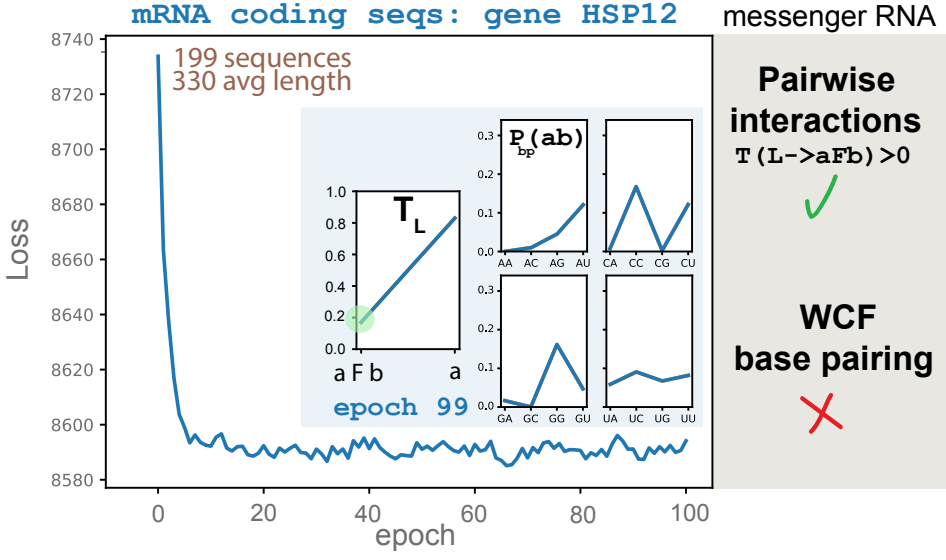


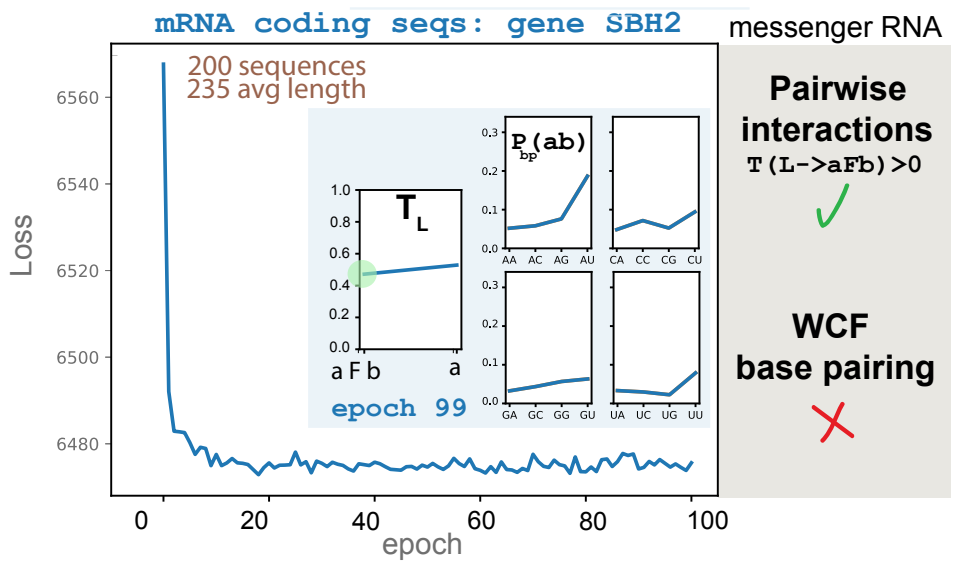
not canonical RNA base pairs

different for different mRNAs


RNaseP RNA sequences
structural RNA
225 sequences
337 avg length
Loss
epoch
epoch 99
Pairwise interactions
T(L->aFb)>0
WCF base pairing


shuffled RNaseP RNA sequences
random RNA
225 sequences
337 avg length
Loss
epoch
epoch 99
Pairwise interactions
T(L->aFb)~0
WCF base pairing


mRNA coding seqs: gene HSP12
messenger RNA
199 sequences
330 avg length
Loss
epoch
epoch 99
Pairwise interactions
T(L->aFb)>0
WCF base pairing


mRNA coding seqs: gene SBH2
messenger RNA
200 sequences
235 avg length
Loss
epoch
epoch 99
Pairwise interactions
T(L->aFb)>0
WCF base pairing

Are we ditching deep learning?
absolutely not!

We love:

working on fast hardware: **GPUs, TPUs**

computational frameworks to train parameters: **PyTorch, TensorFlow**

Easy automatic way to calculate gradients: **Backpropagation, JAX**

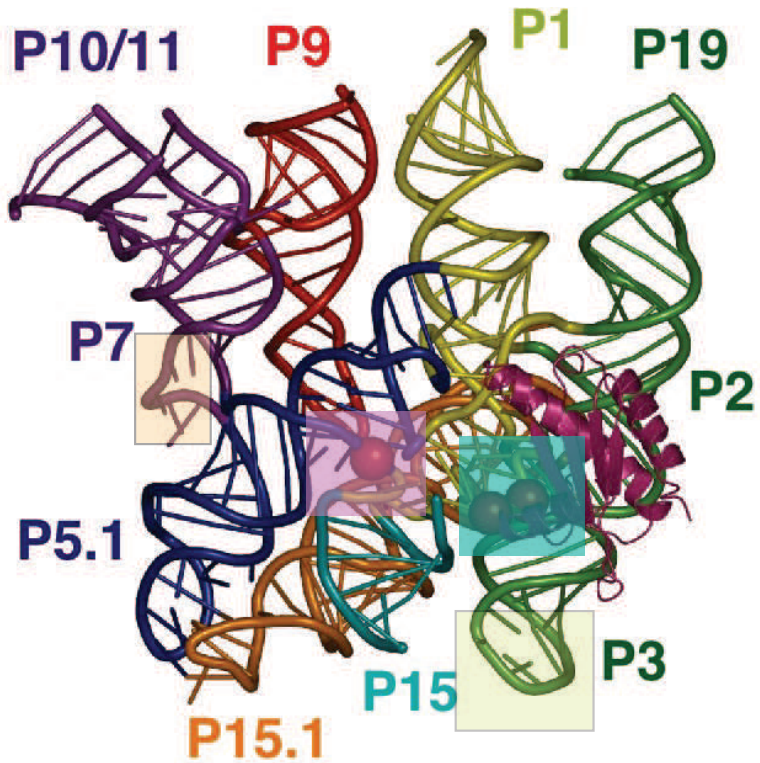
"room" to be surprised by what we can learn: **more parameters**

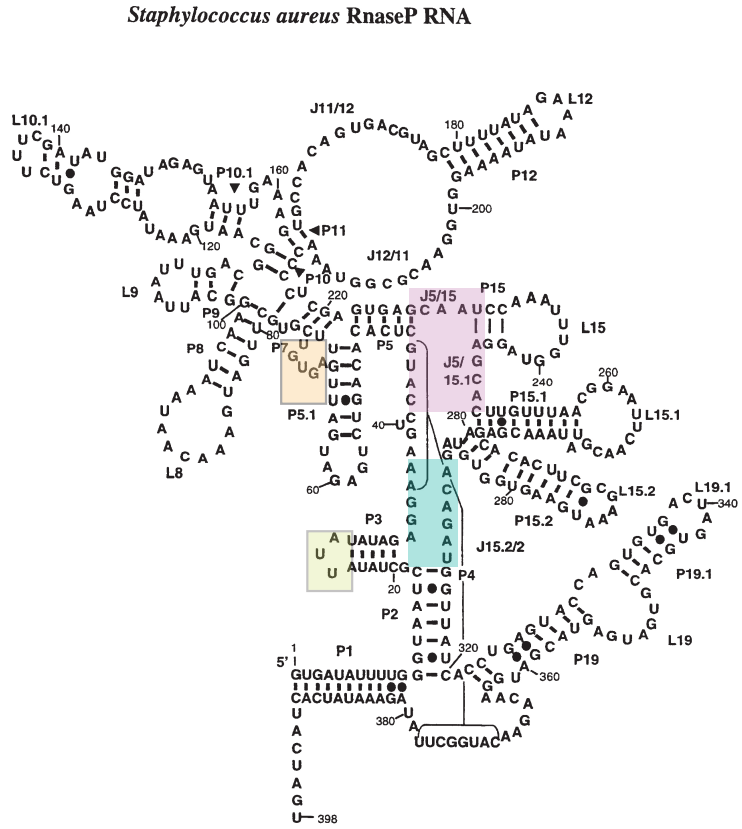
Learning from data w/o previous assumptions

There is more to RNA structure than helices of canonical base pairs

loops

Important for molecule binding

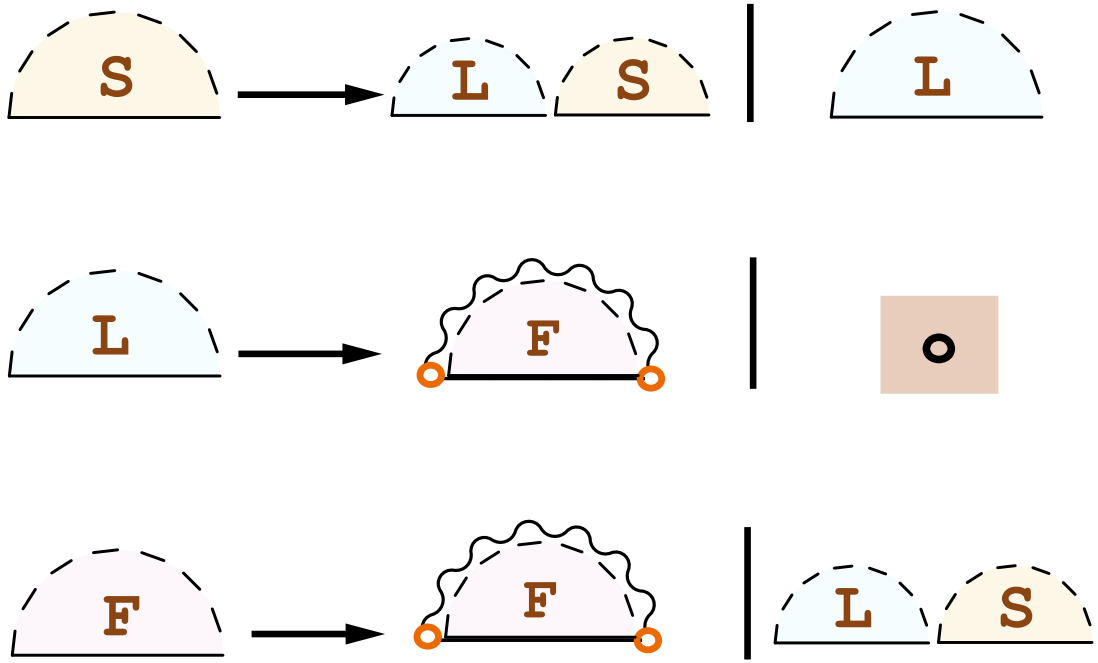




Bacterial RNaseP RNA

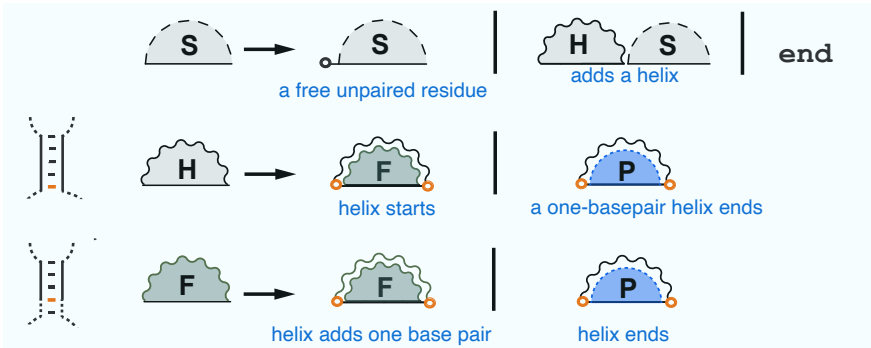
adapted from pacelab.colorado.edu

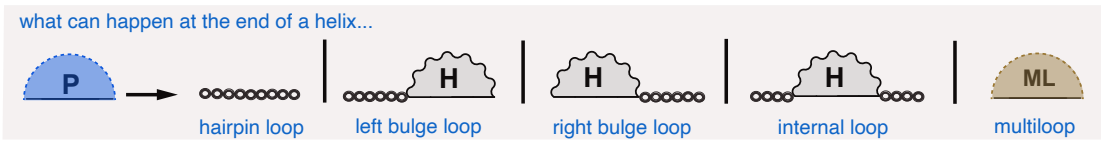
G6 Grammar

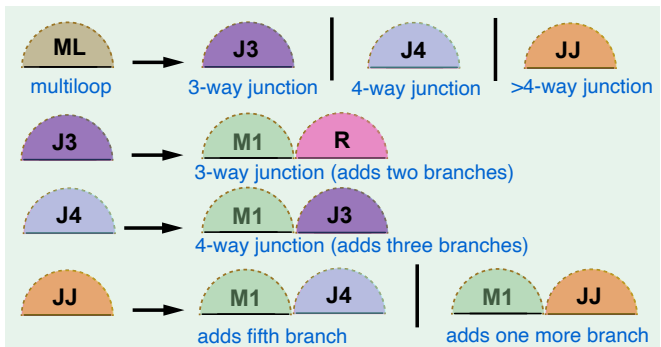


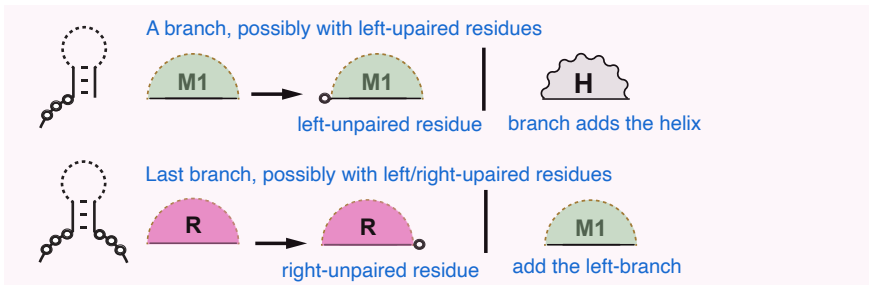
all unpaired residues are treated equally

RNA Basic Grammar (RBGJ3J4)









o an RNA residue, not forming any base pairing

oooo a set of contiguous unpaired RNA residues

o o an RNA base pair

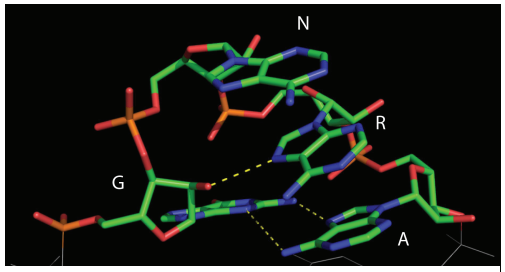
S, H, F, P, ML, M, J3, J4, JJ, M1, R

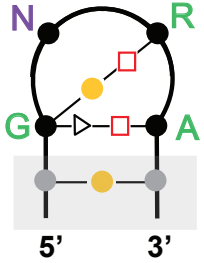
non-terminals that transform following one of the allowed rules

Rivas, PLOSCB, 2020

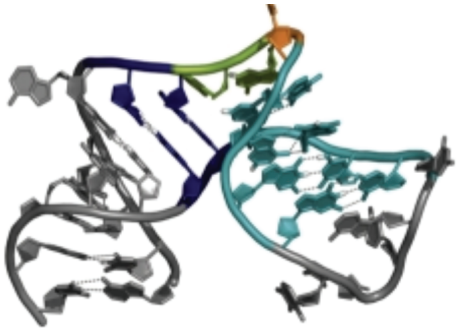
3D motifs in loops

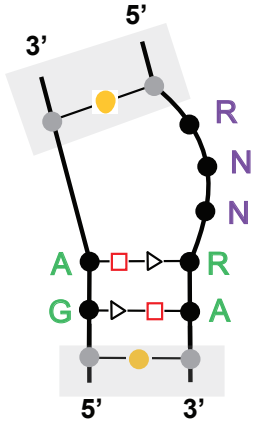
GNRA tetraloop



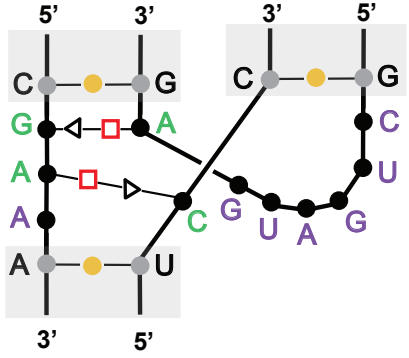


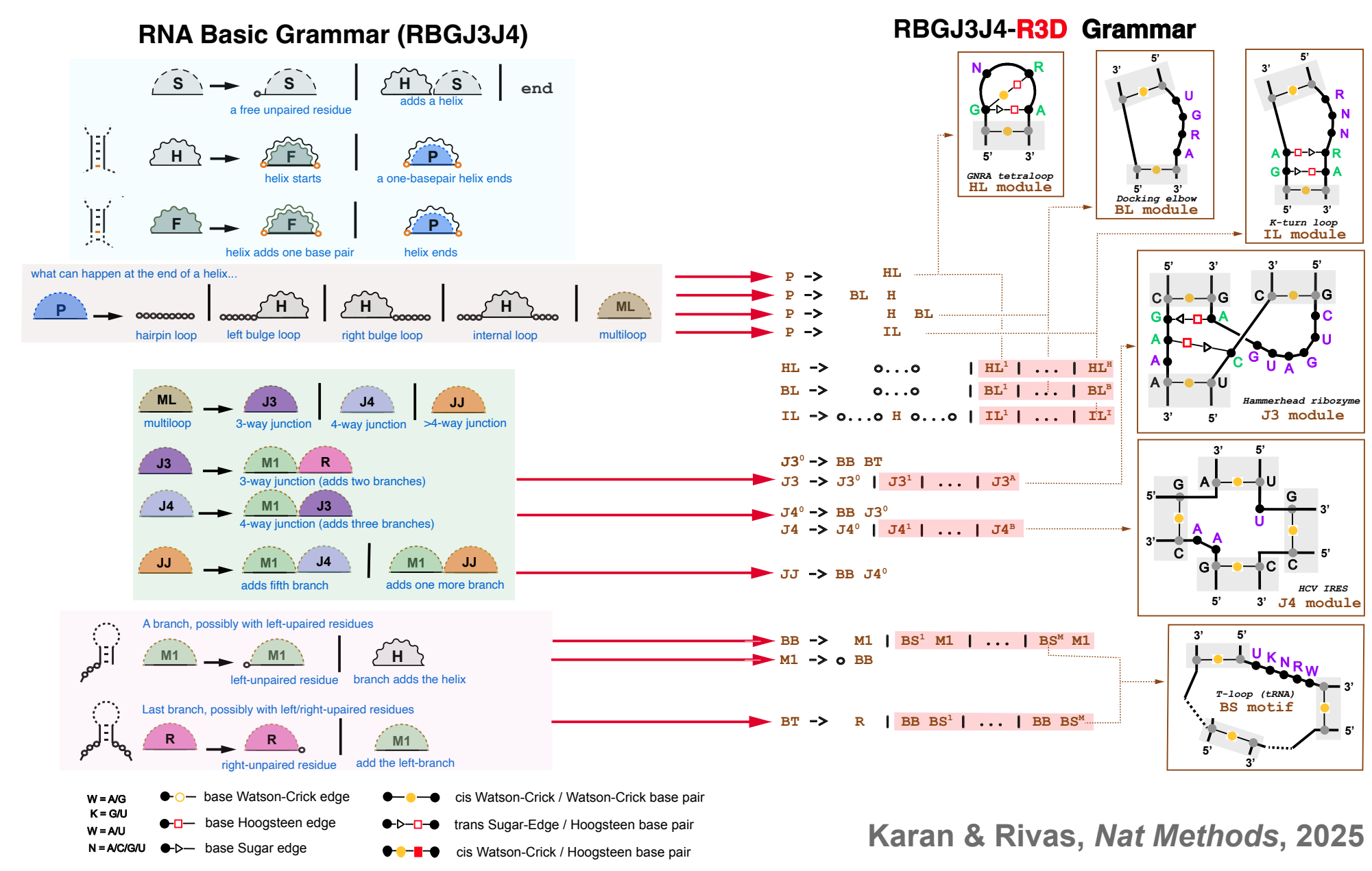
K Turn





Hammerhead ribozyme J3





Karan & Rivas, *Nat Methods*, 2025

RNA 3D Motifs in the Literature

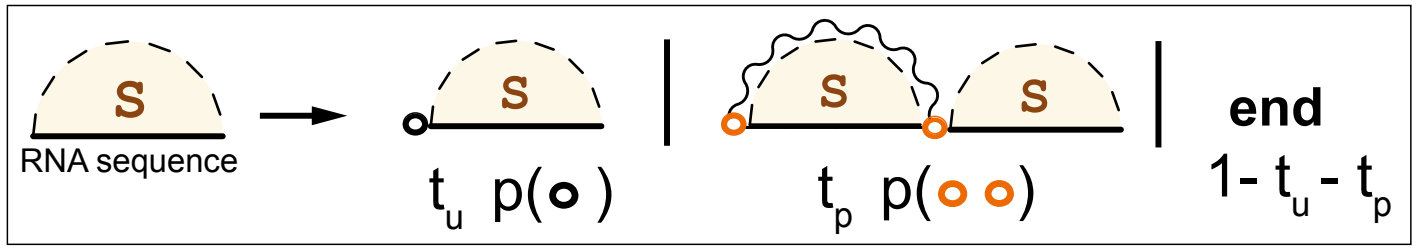
Type	Motif name	PDB	Rfam homology	CaCoFold-R3D	Other Methods	Motif Model
HL	GNRA-tetraloop	7EI6	synthetic	170	–	Hendrix *et al.* , 2005 [17]
HL	U-turn	4WFL	RF00017_Metazoan_SRP	yes	–	Kempf *et al.* , 2014 [22]
HL	UNCG-tetraloop	1F7Y	synthetic	48	–	Hendrix *et al.* , 2005 [17]
HL	ANYA-tetraloop	6MSF	synthetic	18	–	Hendrix *et al.* , 2005 [17]
HL	CUYG-tetraloop	2L6I	synthetic	20	–	Hendrix *et al.* , 2005 [17]
HL	YGNN-tetraloop	1AFX	synthetic	13	–	Hendrix *et al.* , 2005 [17]
HL	GANC-tetraloop	3BWP	synthetic	15	–	Keating *et al.* , 2009 [21]
HL	UNAC-tetraloop	4A4R	synthetic	26	–	Zhao *et al.* , 2012 [99]
HL	T-loop-tetraloop	7SAM	RF01085_TLS-PK4	yes	–	RMfam [36]
HL	L8_RNaseP_bact_a	1U9S	RF00010_RNaseP_bact_a	yes	–	Das & Baker, 2010 [20]
HL	Sarcin-Ricin_loop	1SCL	RF02540_LSU_rRNA_archaea	yes	BayesPairing2	Szewczak *et al.*, 1993 [100]
HL	CsrA_binding	2MF0	RF00018_CsrB	yes	–	Duss et al. , 2014 [23]
HL	GAGUA_pentaloop_sars	1XJR	RF00164_Coronavirus s2m motif	yes	–	Das & Baker, 2010 [20]
IL	Kink-turn_SAM_riboswitch	2GIS	RF00162_SAM_riboswitch	yes	BayesPairing2	Das & Baker, 2010 [20]
IL	Kink-turn_rRNA	1JJ2	RF02540_LSU_rRNA_archaeal	yes	RMDetect	Das & Baker, 2010 [20]
IL	LoopE	354D	RF00001_5S_rRNA	yes	–	Das & Baker, 2010 [20]
IL	GAAA_Tetraloop-receptor	1GID	RF00028_Group I intron	yes	–	Cate *et al.* , 1996 [8]
IL	C-loop	1S72	RF02541_LSU_rRNA_bacterial	yes	RMDetect	Lescoute *et al.* , 2005 [84]
IL	G-Bulge	3DIL	RF00168_Lysine riboswitch	yes	RMDetect	Serganov *et al.* , 2008 [87]
IL	G-Bulge_Das	1Q9A	RF02541_Bacterial LSU_RNA	yes	–	Das & Baker, 2010 [20]
IL	tandem GA	1SA9	RF00177_SSU_rRNA_bacteria	yes	RMDetect	Cruz & Westhof, 2011 [3]
IL	Twist up	1J5E	RF02541_LSU_rRNA_bacterial	no , 9	–	Zhong & Zhang, 2011 [101]
IL	UAA/GAN	6AHR	RF00009_RNaseP_nuc	yes	–	Lee *et al.* , 2006 [102]
IL	J4a/4b	2QBZ	RF00380_M-box_riboswitch	yes	–	Das & Baker, 2010 [20]
IL	J4/5	2R8S	RF00028_Group I intron	no , 6	–	Das & Baker, 2010 [20]
IL	GUA/GG_RRE	1CSL	RF00055_HIV-1 RRE	too conserved, 10	–	Das & Baker, 2010 [20]
IL	SRP RNA DOMAIN IV	1CQ5	RF00017_Metazoan_SRP	yes	–	Schmitz *et al.* , 1999 [65]
IL	Hook_turn_GAAC/CUA	1MHK	synthetic	32	–	Das & Baker, 2010 [20]
IL	Docking elbow-IL	3Q1Q	RF00011_RNaseP_bact_b	yes	–	Lehmann *et al.* , 2012 [83]
IL	pK-turn	3OK7	RF00010_RNase P RNA	no , 45	–	Meyer *et al.* , 2012 [89]
J3	Purine_riboswitch_J3	2EEW	RF00167_Purine riboswitch	yes	–	Das & Baker, 2010 [20]
J3	hammerhead_CUGA/A/C	359D	RF02275_Hammerhead_HH9	yes	–	Das & Baker, 2010 [20]
J3	J3_typeA	4RZD	RF02680_PreQ1-III	yes	–	Lescoute & Westhoff, 2006 [103]
J3	J3_typeB	1J5E	RF00177_SSU_rRNA_bacteria	yes	–	Lescoute & Westhoff, 2006 [103]
J3	J3_typeC	4LX6	RF03165_2dG-II riboswitch	yes	–	Lescoute & Westhoff , 2006 [103]
J3	J3_TPP	2GDI	RF00059_TPP riboswitch	yes	–	Serganov *et al.* , 2006[62]
J3	J3_groupII	5J01	RF02001_group-II-D1D4-3	yes	–	this work
J4	HCV_IRES_AA/0/U/0	1KH6	RF00061_IRES_HCV	yes	–	Das & Baker, 2010 [20]
J4	J4_U1RNA	3PGW	RF00003_U1 RNA	yes	–	Weber et al. , 2010 [104]
BS	T-loop	1EHZ	RF00005_tRNA	yes	–	Hendrix *et al.* , 2005 [17]
BS	LoopE-a	4FRN	RF01689_AdoCbl-II	yes	–	this work
BS	LoopE-b	4FRN	RF01689_AdoCbl-II	yes	–	this work
BS	CRC_binding	–	RF00195_RsmY	yes	–	Sonnleitner *et al.* , 2009 [105]
PK		5DDP	RF01739_Glutamine riboswitch	yes	–	Ren *et al.* , 2015 [64]

Unsupervised Learning of RNA 3D Motifs

Can we learn the 3D motifs from RNA sequences alone?

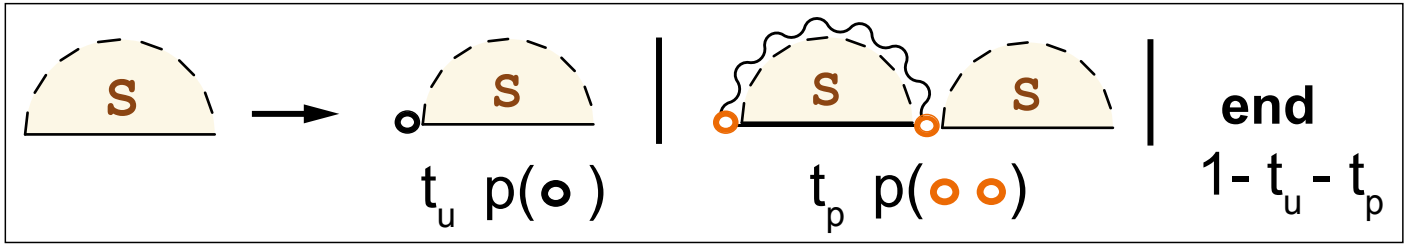
G5 SCFG

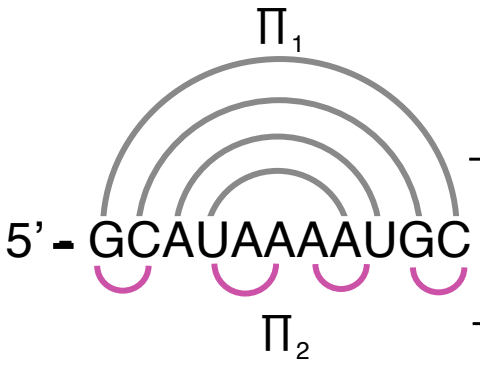
```
S -> o S
S -> o S o S
S -> end
```



G5 Grammar

```
S -> o S
S -> o S o S
S -> end
```





$P(\Pi_1)$ → $t_p^4 = 0.0067$

=

$P(\Pi_2)$ → $t_p^4 = 0.0067$

G5:

G6:

$t_L \times t_F^3 = 0.0519$

>

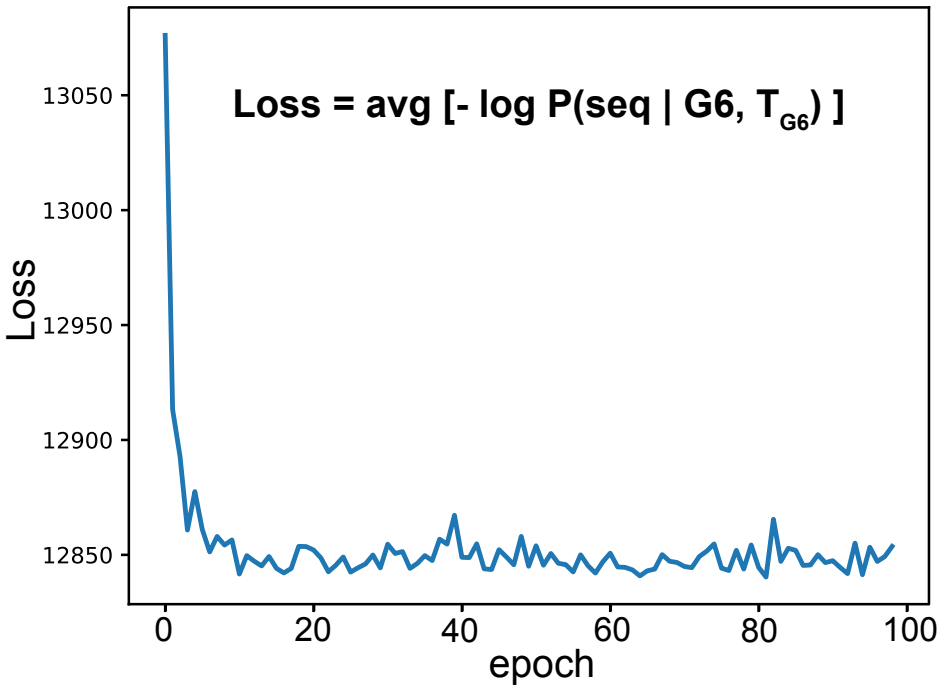
$t_L^4 = 0.0006$

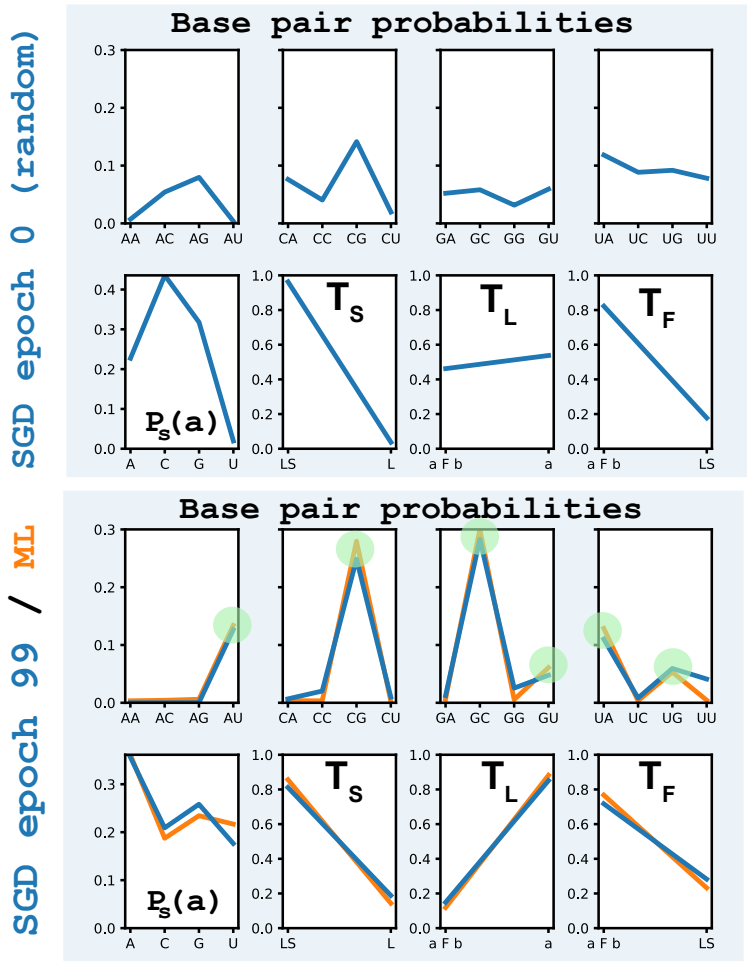
G5 cannot favor helix stacking

Learns the rule of RNA base pairing

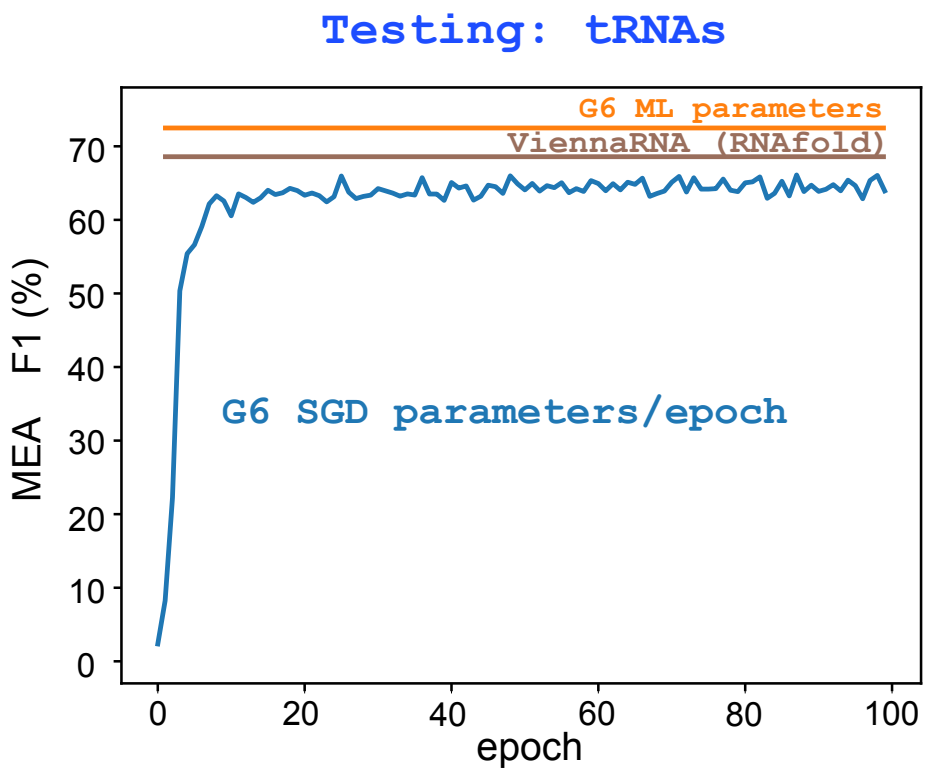
Training: 7 RNA Fams 400 seqs

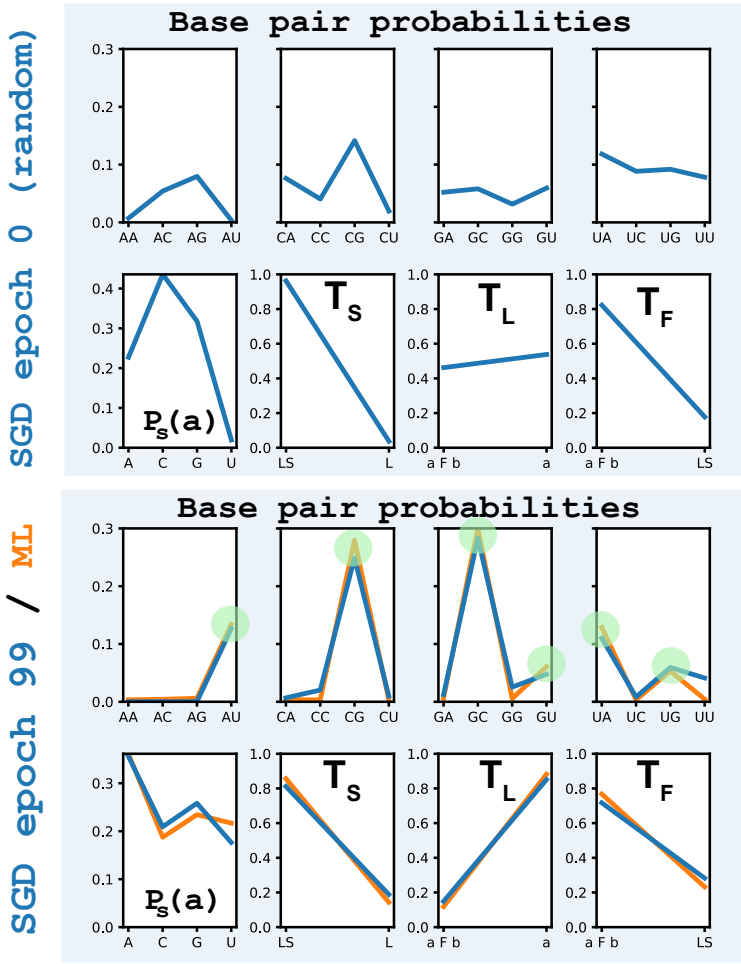
RNaseP RNA, SRP RNA, 5S rRNA, telom RNA, tmRNA, group I, and group II introns





Learns the rule of RNA base pairing





To learn the rules of RNA base pairing

Deep models with **millions** of parameters

Minimal **model**

21 free parameters

Trained on **100 thousands** RNA seq/structures

Minimal **inputs**

25-50 RNA sequences only

Merging SCFGs with standard deep-learning methods:

Stochastic Gradient Descent

Backpropagation by automatic differentiation

A loss that optimizes the marginal probability of the sequences (all structures integrated)

Loss is calculated exactly with Inside algorithm, (no need for "evidence lower bound")

SCFGs directly interpretable: a structural RNA genefinder