RNA secondary structure prediction beyond thermodynamics

A four-ingredient unifying perspective

E. R. & Sean Eddy

Single-sequence secondary structure prediction:

hermodynamic Models	Mfold/UNAFold	Zuker, Stiegler & Sankoff (1981-)
	ViennaRNA	Hofacker & Stadler (1994-)
	RNAstructure	Mathews (1999-)

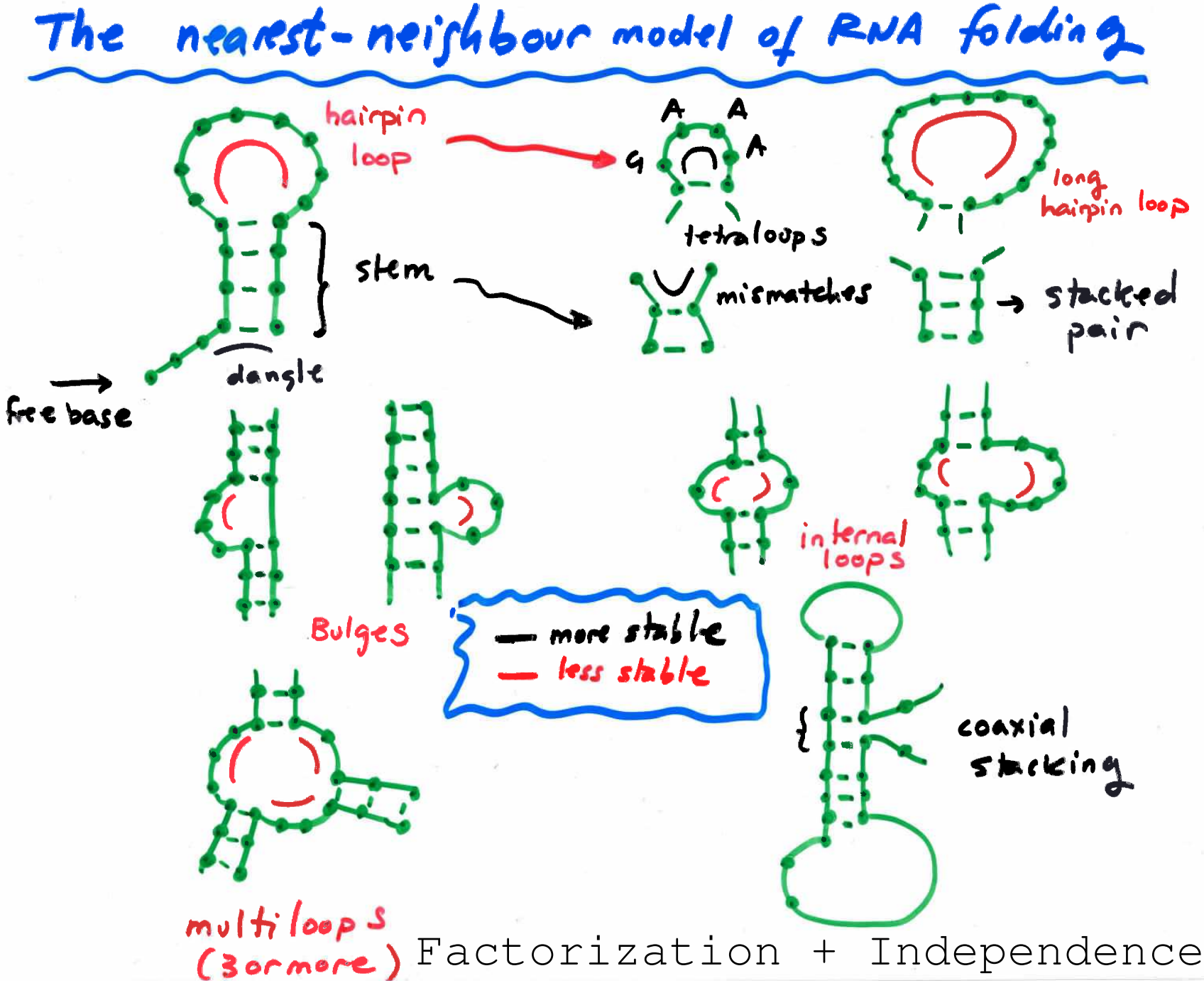
lumans do it by comparative analysis (C. Woese, R. Gutell)

ccurate - requires a lot of sequences

ovariation is a statistical signal, not thermodynamic

se of probabilistic models for RNA secondary structure prediction

Features of RNA secondary structure

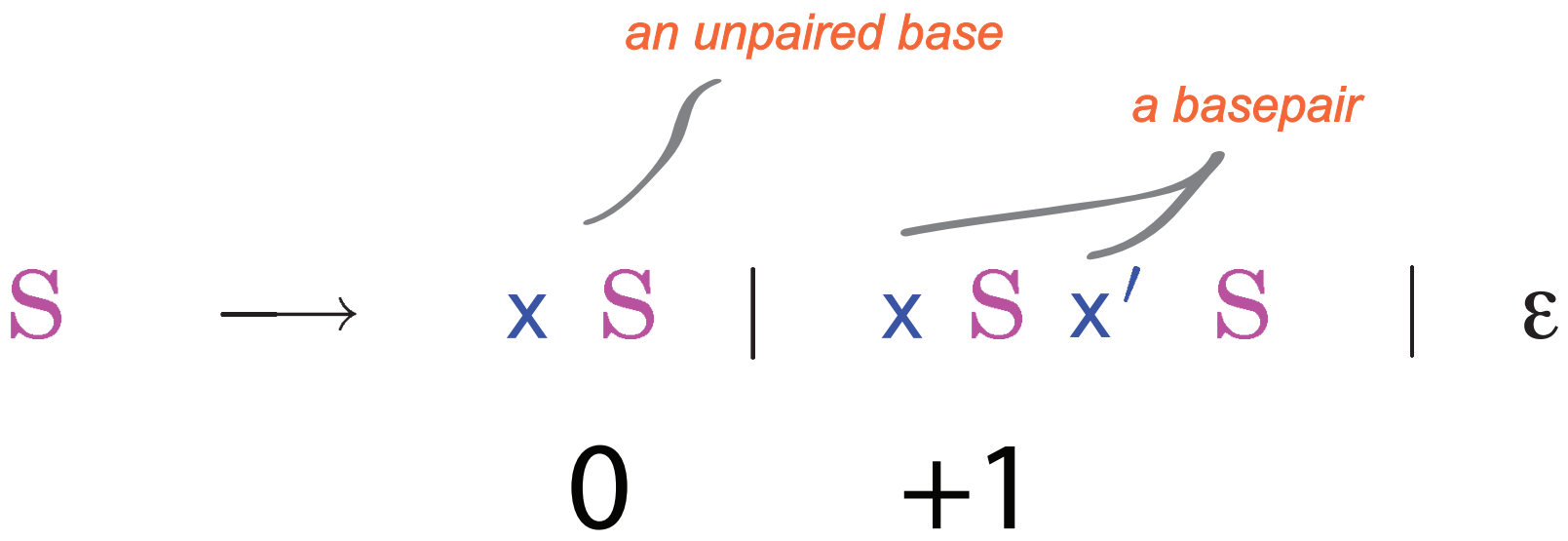


Factorization + Independence

Context-free grammars

for RNA secondary structure

=SIMPLEST



S = nonterminal

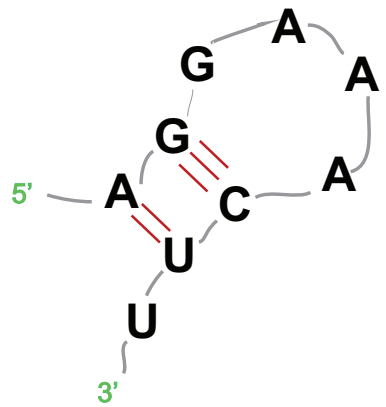
x , x′ = terminals (A,C,U, or G)

ε = empty string

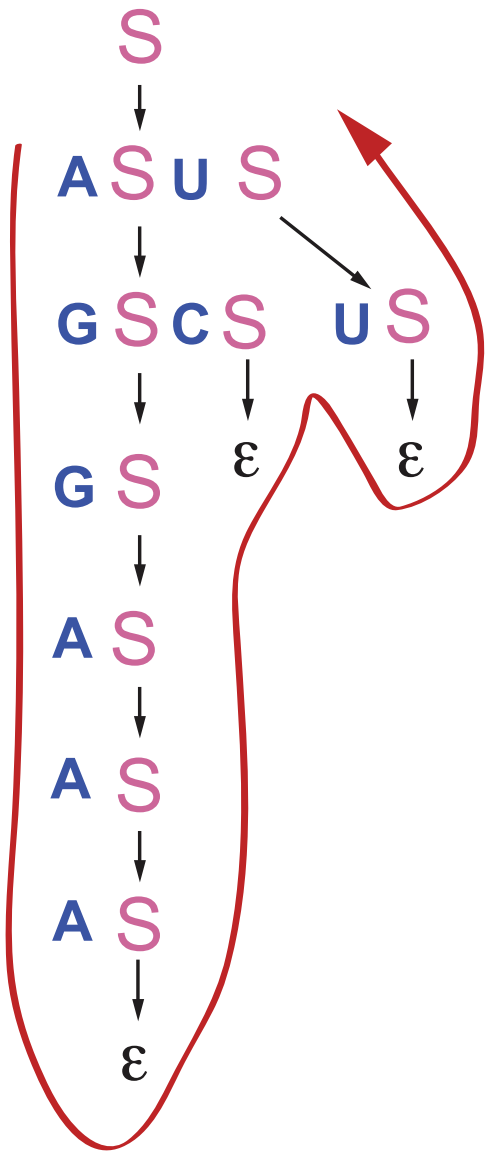
Nussinov, 1979

Nussinov structure generation

$$S \longrightarrow x\,S \mid x\,S\,x'\,S \mid \varepsilon$$



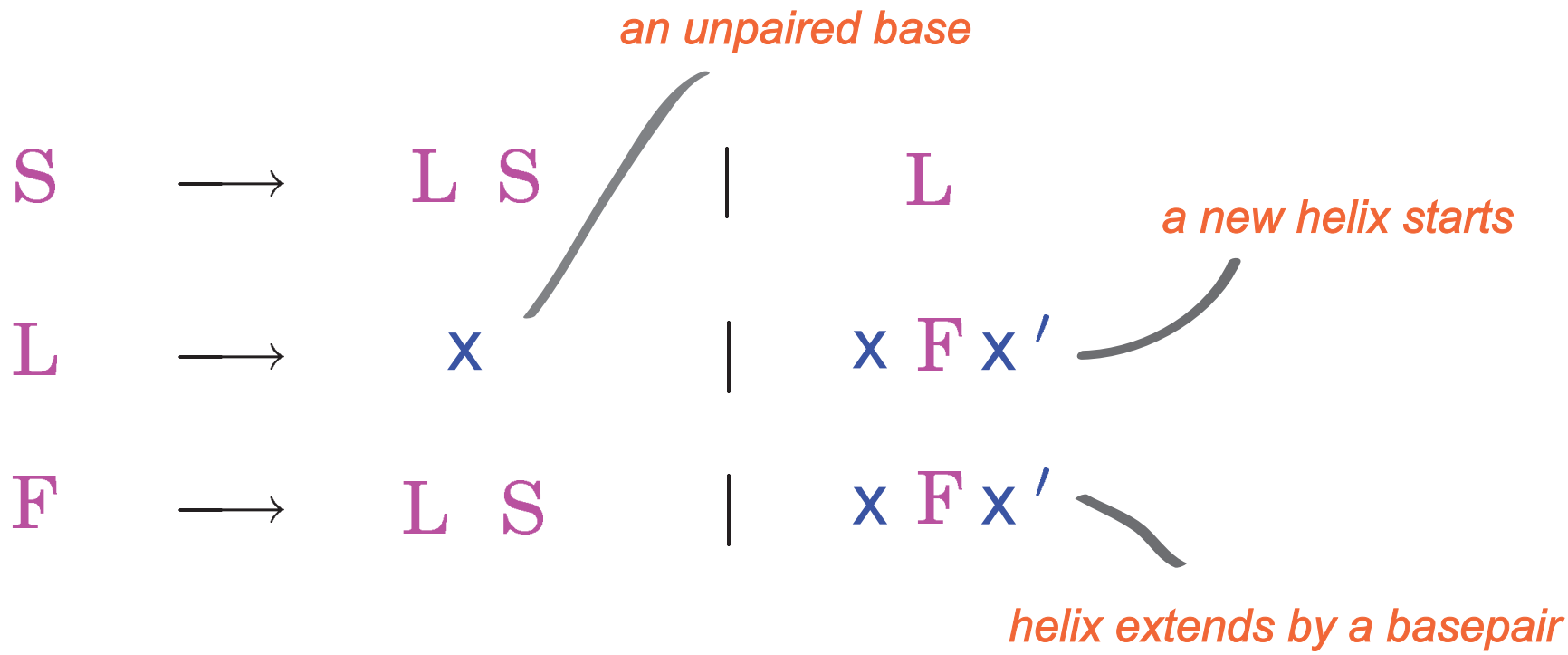
5' - A G G A A A C U U - 3'



Context-free grammars

for RNA secondary structure

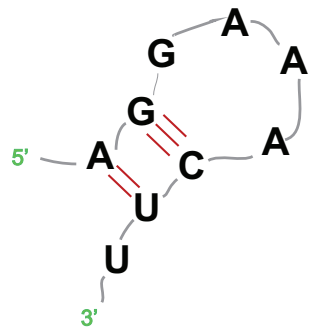
G6 [Pfold]



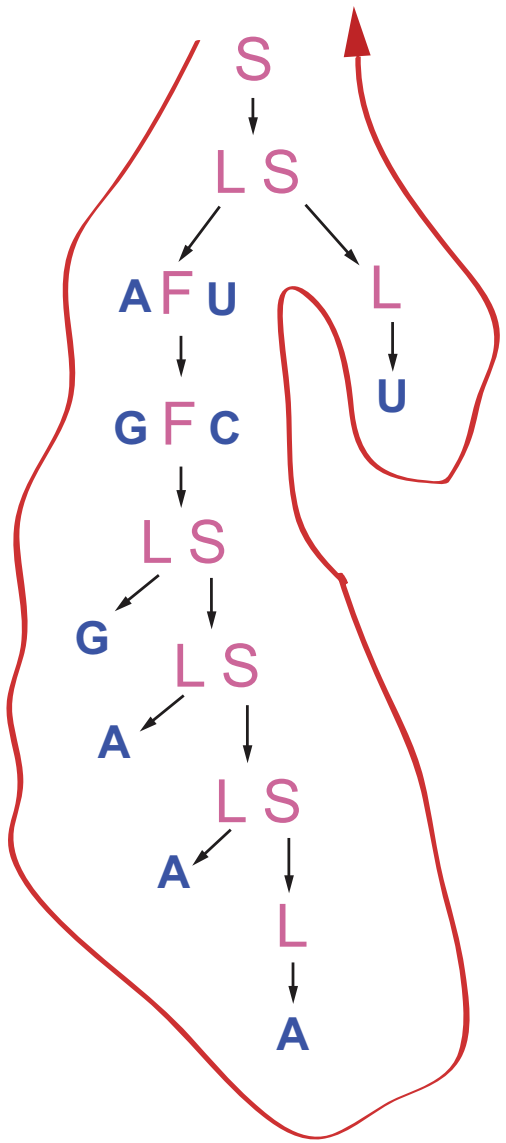
Hein & Knudsen, 1999
Dowell & Eddy, 2004

G6 structure generation

S	$\longrightarrow$	L S	L
L	$\longrightarrow$	x	x F x′
F	$\longrightarrow$	L S	x F x′



5' - A G G A A A C U U - 3'



A "basic" complex grammar

$$S \longrightarrow a\ S \mid F0\ S \mid \epsilon$$

a new helix starts

$$F0 \longrightarrow a\ F5\ a'$$

$$F5 \longrightarrow a\ F5\ a'$$

helix extends by a basepair

$$F5 \longrightarrow a\ P\ a'$$

a helix ends

inside a helix....

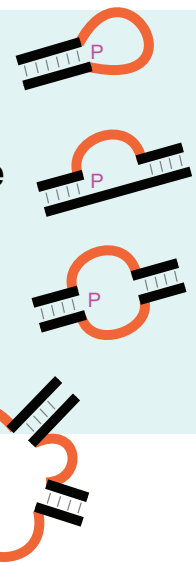
$$P \longrightarrow a_1 \ldots a_n$$

$$P \longrightarrow a_1 \ldots a_n\ F0$$

$$P \longrightarrow F0\ a_1 \ldots a_n$$

$$P \longrightarrow a_1 \ldots a_n\ F0\ a_{n+1} \ldots a_m$$

$$P \longrightarrow M1\ M$$



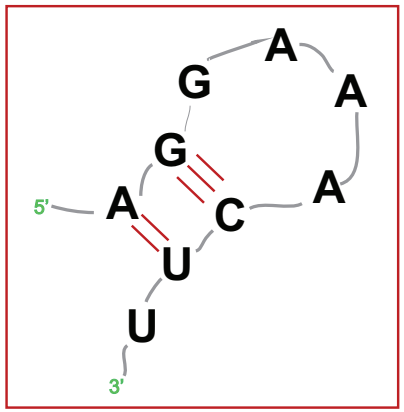
$$M \longrightarrow M1\ M \mid R$$

$$M1 \longrightarrow a\ M1 \mid F0$$

$$R \longrightarrow R\ a \mid M1$$

Scoring Schemes

Thermodynamic versus Statistical



Thermodynamic

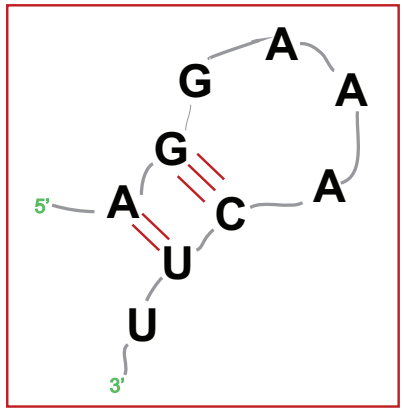
ΔG (Kcal/mol)

er, SantaLucia, Freier, Zuker, 1987)

dangles off A═U	= - 0.16
═C stacked on A═U	= - 3.41
rpin loop of 4 nts	= + 9.09
AA stability bonus	= - 3.97
al ΔFree Energy	= + 1.55 Kcal/mol

Scoring Schemes

Thermodynamic versus Statistical



Thermodynamic

ΔG (Kcal/mol)

(Turner, SantaLucia, Freier, Zuker, 1987)

U dangles off A═U = - 0.16

G≡C stacked on A═U = - 3.41

hairpin loop of 4 nts = + 9.09

GAAA stability bonus = - 3.97

Total ΔFree Energy = + 1.55 Kcal/mol

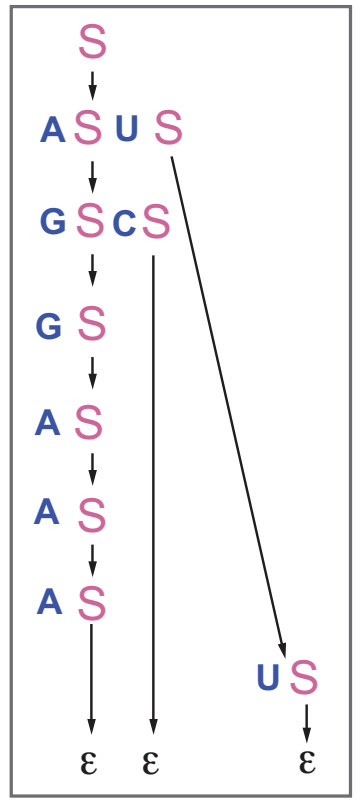
Free Energies ≈ log Probabilities

Statistical

$S \rightarrow x\,S \mid x\,S\,x'\,S \mid \varepsilon$

transition probabilities t_1 (0.5) t_2 (0.24) t_3 (0.26)

emission probabilities P(a)=0.25 P(AU or GC or GU)=1/6



+1	A S U S	log t_2 + log P(AU)
+1	G S C S	log t_2 + log P(GC)
+0	G S	log t_1 + log P(G)
+0	A S	log t_1 + log P(A)
+0	A S	log t_1 + log P(A)
+0	A S	log t_1 + log P(A)
+0	U S	log t_1 + log P(U)
+0 +0	ε ε ε	3 * log t_3

+2 log P(total) = ∑(all the terms)

= - 23.5 (6 *10^{-11})

Going beyond thermodynamic models

One **complicated** thermodynamic model to several **simple** probabilistic models

Thermodynamic models outperform Probabilistic models

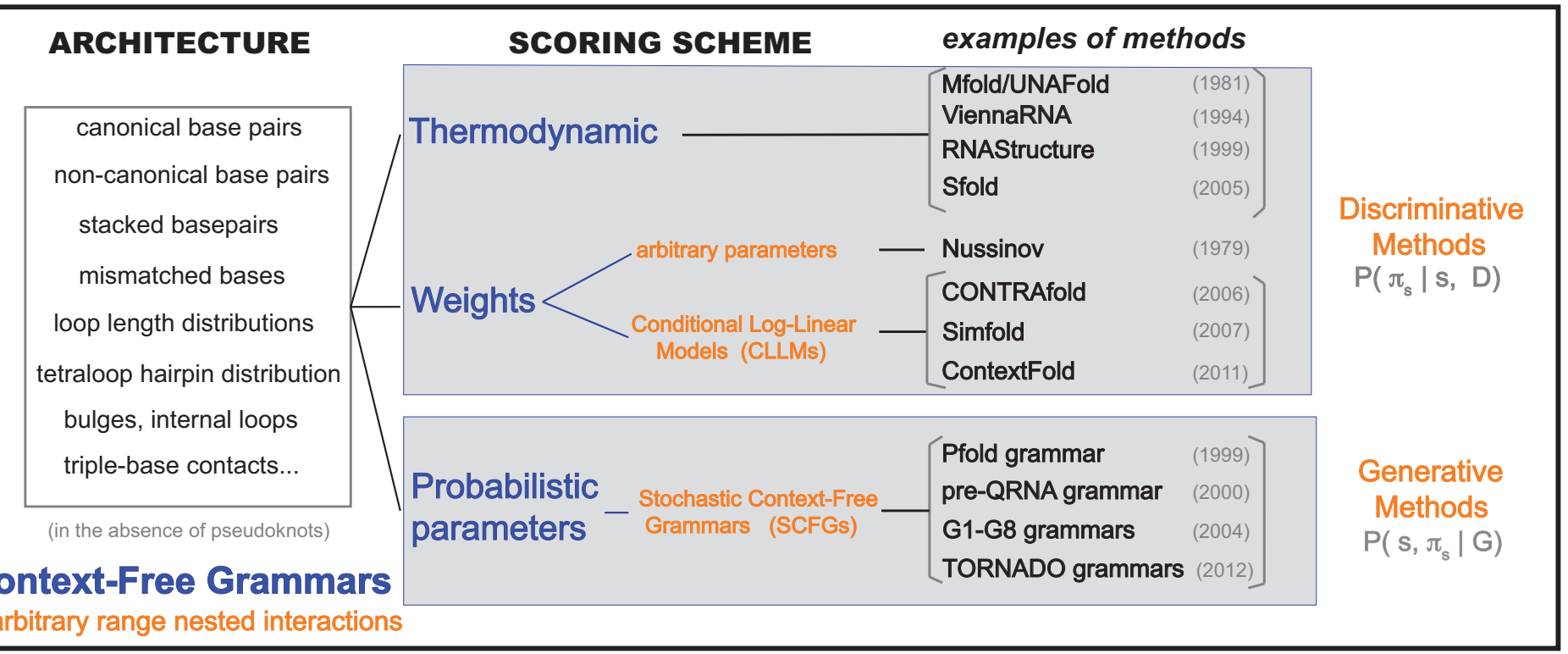
Grammar	Parameters	Folding Accuracy	Scoring Scheme
G6	21	48 %	probabilistic
ViennaRNA	∼14,000	54 %	Thermodynamic

Still performance is poor

Previous literature claimed: probabilistic models are **too constrained and cannot** implement all the complexities of the thermodynamic models. Need to move to other type of statistical methods.

Disentangle

architecture from scoring scheme



Why Statistical Models?

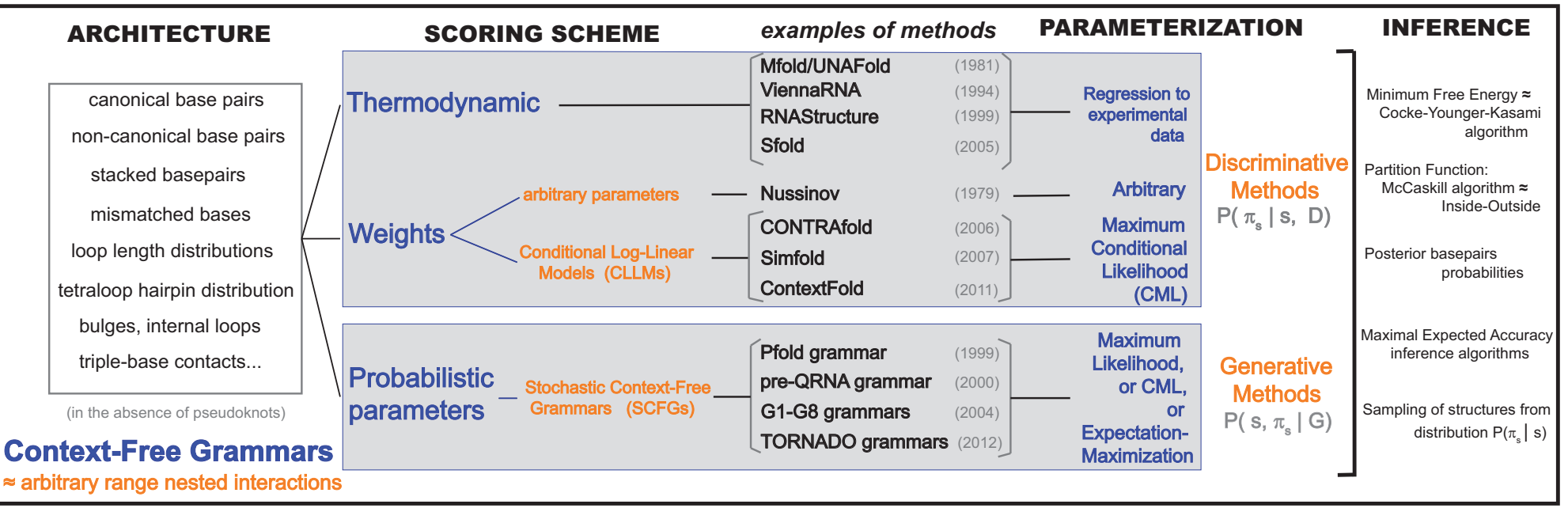
specifically with probabilistic parameters

Statistical models learn parameters from known RNA structures which is an **ever-growing** source of information versus the **slowly-produced** thermodynamic parameters.

Advantage of statistical **probabilistic** models:

- Easily Trainable
- Generative
- Optimal comparison of alternative hypotheses (Neyman & Pearson '33)
- Easy integration of complementary sources of information

THE 4 INGREDIENTS



PARAMETERIZATION

specific for the different scoring schemes

INFERENCE

formally identical algorithms for all scoring schemes

INFERENCE ALGORITHMS

PROBABILISTIC	THERMODYNAMIC
CYK algorithm	Minimum Free Energy
Inside algorithm	McCaskill algorithm

Outside algorithm

Posterior decoding algorithms

Maximum Expected Accuracy inference algorithm

Sampling of structures from the distribution $P(\pi|seq)$

Same algorithms for all scoring systems

L^3 in time

L^2 in memory

TORNADO

tool to swap ingredients

Architecture

Scoring scheme

Parameterization

Folding algorithm

Relative performance of probabilistic versus thermodynamic scores?

old architecture fixed, vary scoring schemes

Contribution of the different elements of RNA 2D structure?

ary architecture, hold everything else fixed

Existing complex grammars

have created TORNADO "emulations" of the state of the art
RNA models that exist to date.

ViennaRNA
thermodynamic

ViennaRNA-G
TORNADO grammar

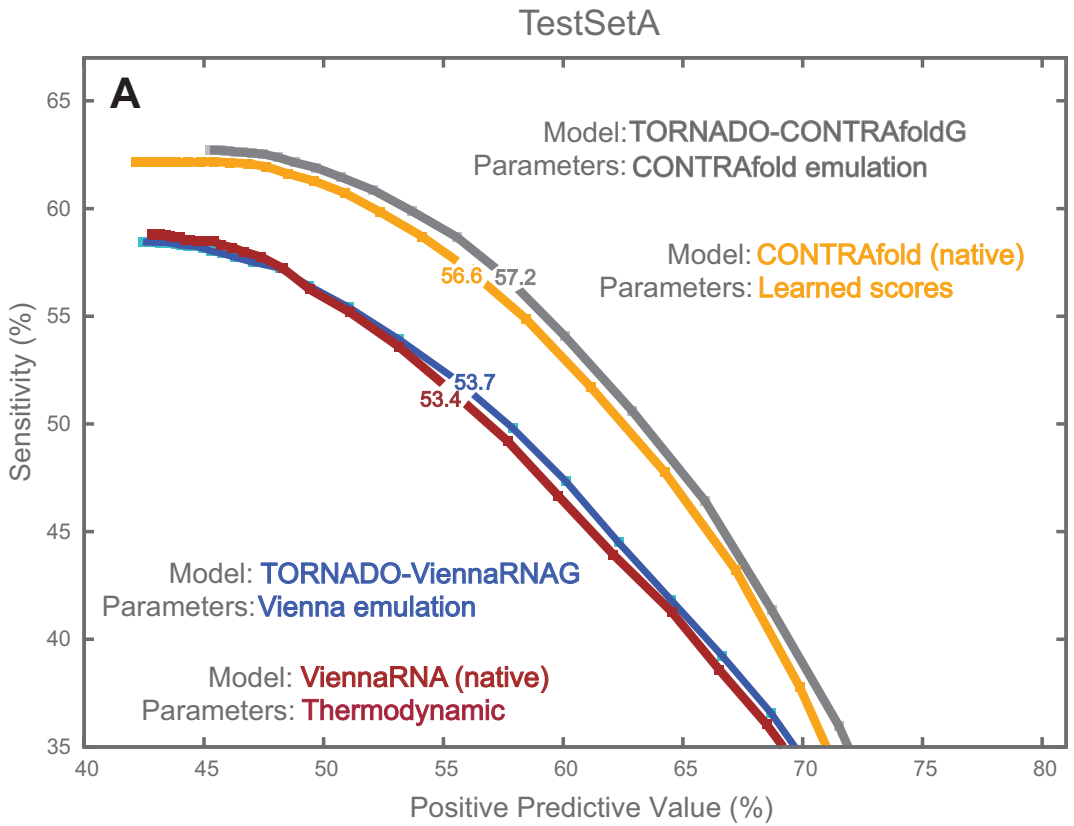
14,000 parameters

CONTRAFold
learned parameters

CONTRAFold-G
TORNADO grammar

1,500 parameters

Context-free grammars reproduce the complexity of the thermodynamic nearest-neighbor model



ensitivity = fractions of true basepairs predicted correctly

ositive Predictive Value = fraction of predicted basepairs that are correct

Probabilistic Complex Grammars

What happens if now one turns the parameters of these models into **probabilities** trained using known RNA structures?

Parameterization

Training and test sets

Literature-Based

Dowell&Eddy, 2004; Do et al, 2006; Andronescu et al, 2007; Lu et al, 2009; Andronescu et al, 2010.

TrainSetA

3166 Sequences
48 % basepaired
< 0.1 % non-canonical

- SSU/LSU domains (1004)
- tRNA (157)
- SRP RNA (215)
- RNaseP RNA (150)
- tmRNA (266)
- 5S RNA (112)
- group I introns (50)
- group II introns (4)
- telomerase RNA (12)
- <50 nts hairpins (962)
- other structures (234)

TestSetA

697 Sequences
52 % basepaired
2.3 % non-canonical

- SSU/LSU domains (135)
- tRNA (140)
- SRP RNA (31)
- RNaseP RNA (29)
- tmRNA (63)
- 5S RNA (50)
- group I introns (28)
- group II introns (4)
- telomerase RNA (30)
- <50 nts hairpins (179)
- other structures (8)

structurally dissimilar

Rfam-based

22 RNA families with 3D structure

TrainSetB

1094 Sequences
46 % basepaired
4.8 % non-canonical

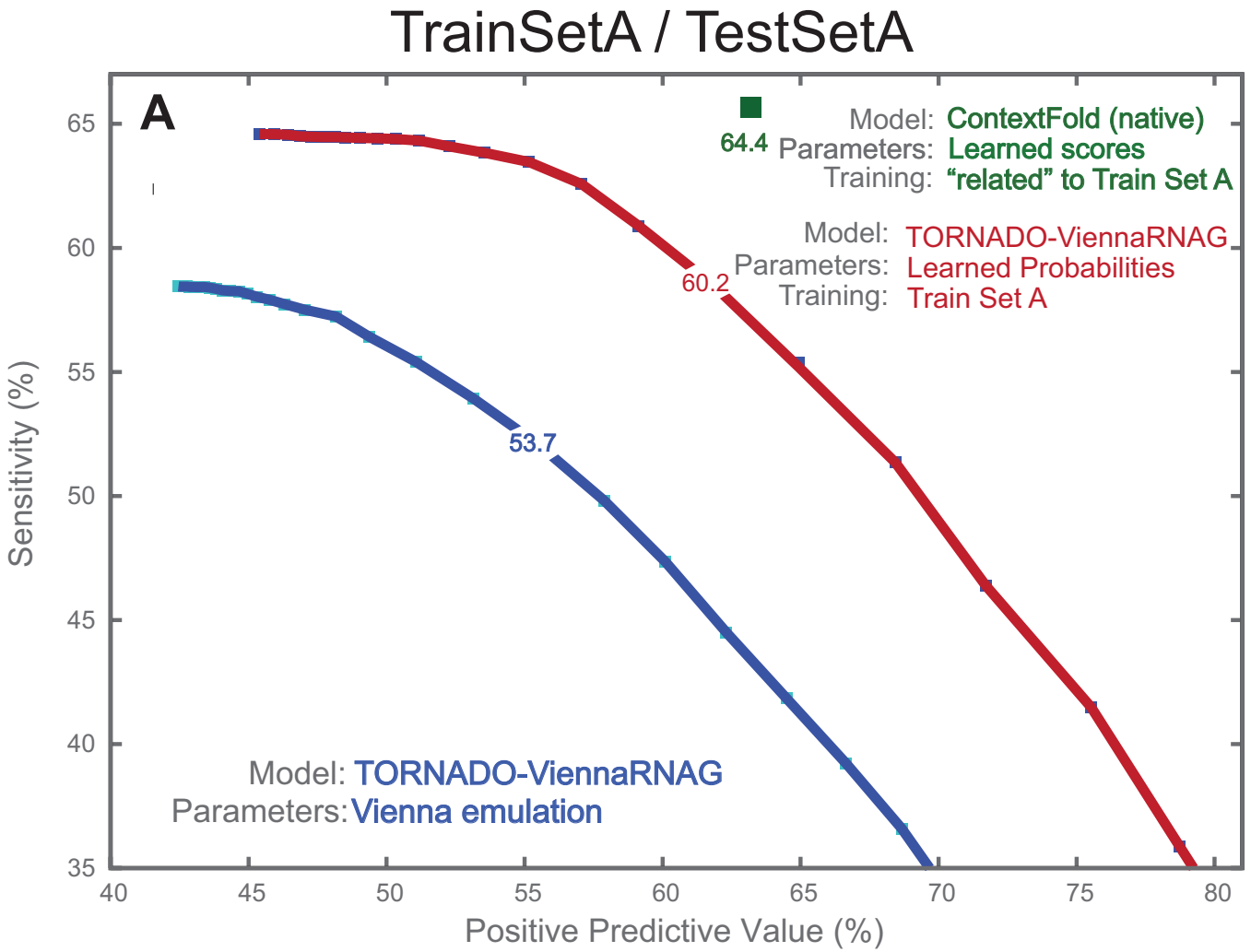
- 5.8S rRNA (41)
- U1 (40)
- U2 (32)
- 7 Riboswitches (365)
- 9 Cis regulatory RNAs (575)
- 2 Ribozymes (41)

TestSetB

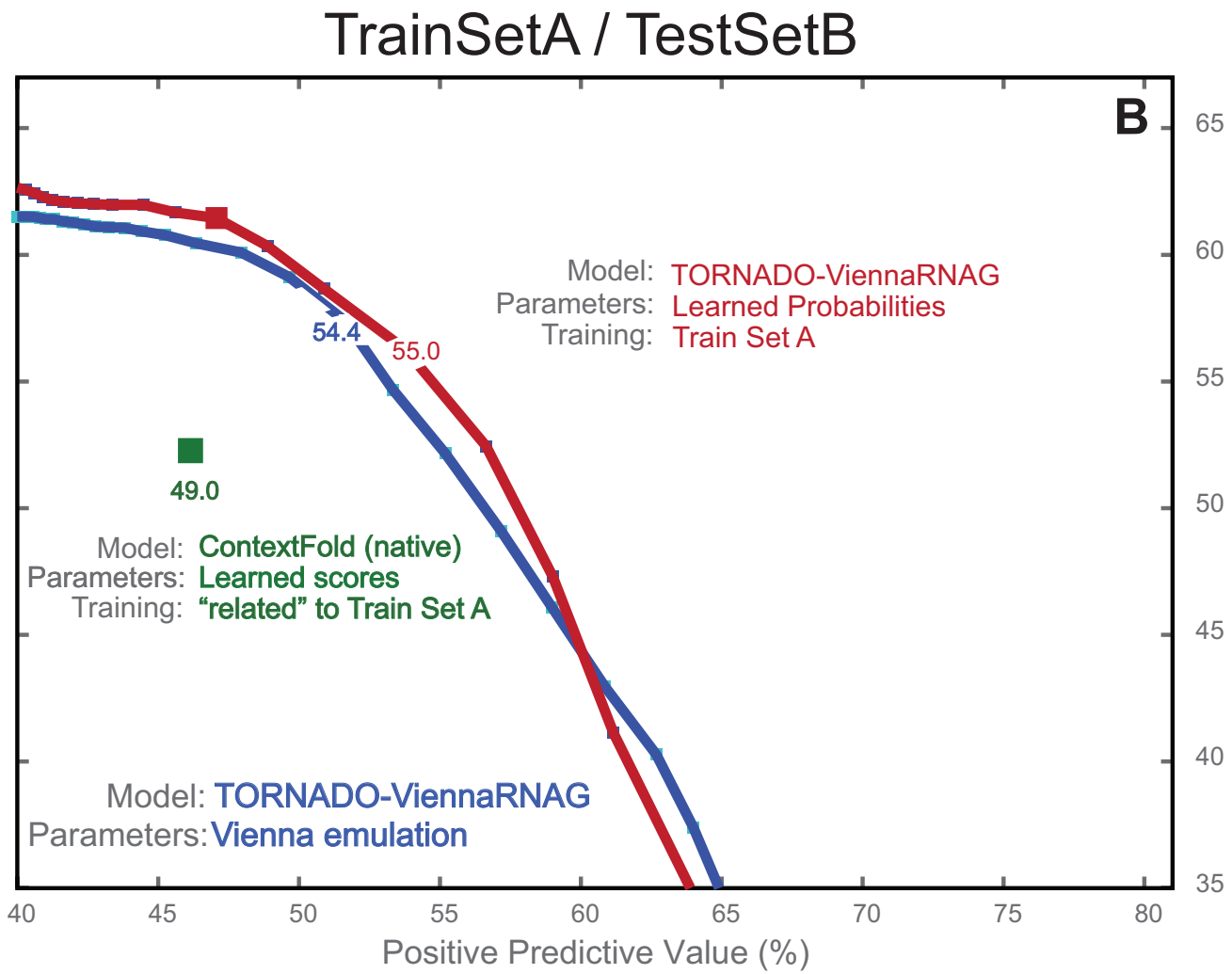
430 Sequences
44 % basepaired
8.3 % non-canonical

- 5.8S rRNA (14)
- U1 (18)
- U2 (45)
- 7 Riboswitches (233)
- 9 Cis regulatory RNAs (116)
- 2 Ribozymes (3)
- bacteriophage pRNA (1)

Benchmark
on same structures



Benchmark
on different structures



Need to train in one type of structures and test on a completely different set

Folding Accuracy

METHOD	ARCHITECTURE # free tied parameters (6 bps)	(16 bps)	SCORING SCHEME	PARAMETERIZATION	FOLDING METHOD	BENCHMARK set best F (%) TestSetA	TestSetB
6	11	21	probabilistic	maximum likelihood	c-MEA	49.1	47.5
asic grammar	532	572	probabilistic	maximum likelihood	c-MEA	56.9	56.5
ONTRAfold v2.02	~300	—	weights	maximum cond. likelihood	c-MEA	57.2	57.9
ONTRAfoldG	1,278	5448	probabilistic	maximum likelihood	c-MEA	58.3	58.6
NAFold-3.8	~3,500	—	thermodynamic	fit to exp. data	CYK	51.0	51.3
mfold BL	~3,500	—	weights	maximum cond. likelihood	CYK	56.5	55.3
NAstructure v5.2	12,700	—	thermodynamic	fit to exp. data	GCE	53.5	53.8
ennaRNA v1.8.4	12,700	—	thermodynamic	fit to exp. data	GCE	53.7	54.3
ennaRNAG	14,307	90,497	probabilistic	maximum likelihood	c-MEA	60.2	59.4
ennaRNAG_plus	14,557	91,997	probabilistic	maximum likelihood	c-MEA	60.5	59.5
ontextFold v1.00	205,000	—	weights	maximum cond. likelihood	CYK	64.4	49.0

Contrafold 58%

Thermodynamic 54%

SCFGs 60%

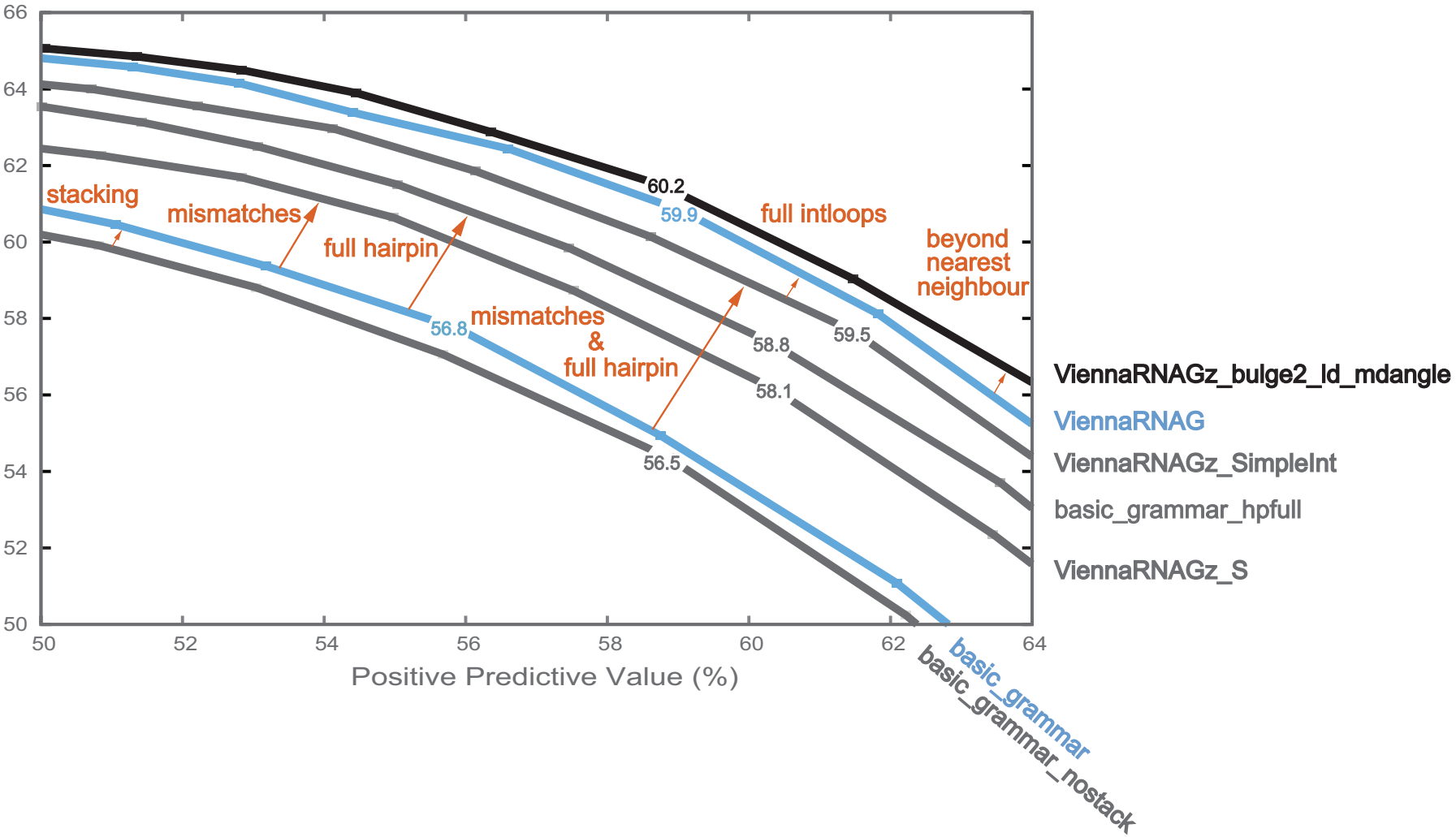
Probabilistc methods implemented in TORNADO

A gradation of SCFGs

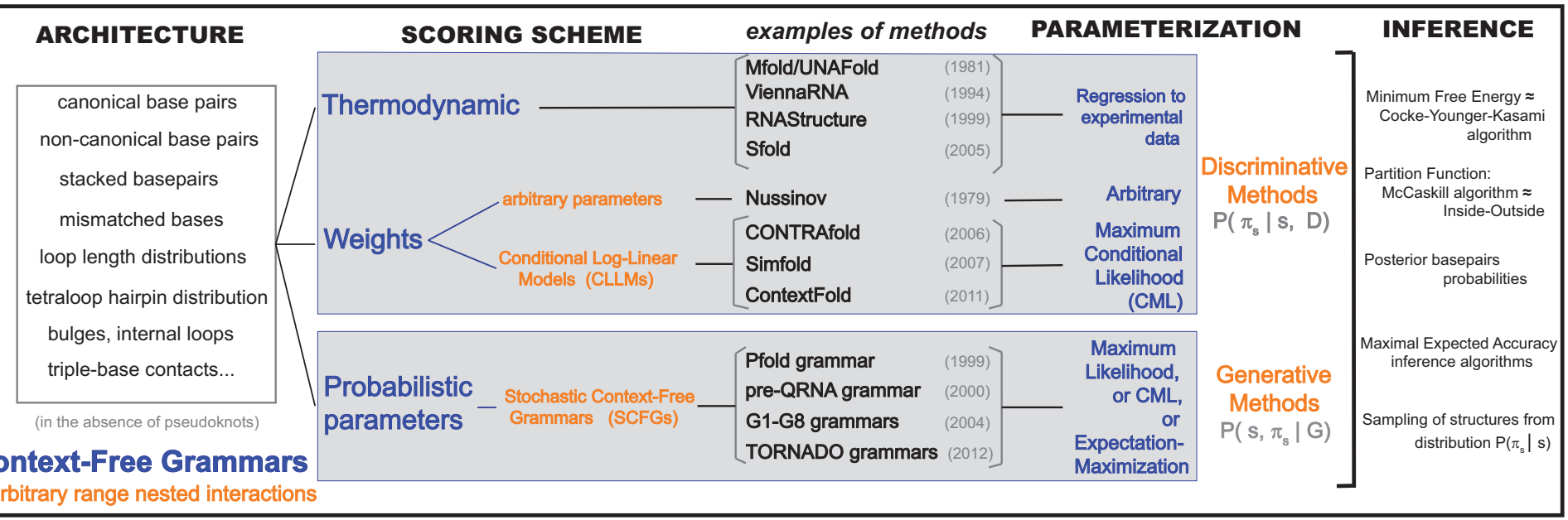
exploring different structural features

Grammar	Total Free Tied Parameters 4x4 bps	6 bps	Remarks
g6	21	11	Pfold grammar
g6s	261	41	Pfold + stacking
g6_stem	294	74	Pfold + stacking + helix length dist.
basic_grammar_nostack	572	532	loop length dist.
basic_grammar	1,022	582	loop length dist + stacking.
basic_grammar_dangle	1,143	643	basic_grammar + dangles
ViennaRNAGz_S	1,862	892	ViennaRNAGz_SimpleInt without tetraloops
CONTRAfoldGS	2,101	811	CONTRAfoldG with simpler 1nt bulges
basic_grammar_hpfull	5,342	2,202	basic_grammar + hairpin tetraloops + hairpin closing mismatches
CONTRAfoldG	5,448	1,278	CONTRAfold emulation
ViennaRNAGz_SimpleInt	6,105	2,495	ViennaRNAG minus 2x2,2x1 Internal loops
ViennaRNAGz_nostack	90,497	14,257	ViennaRNAG minus stacking
ViennaRNAG	90,947	14,307	ViennaRNA emulation
ViennaRNAGz_stem	90,980	14,340	ViennaRNAG+ stem length dist.
ViennaRNAGz_bulge2	91,670	14,400	ViennaRNAG+ explicit 1,2 bulges
ViennaRNAGz_ld	91,012	14,374	ViennaRNAG+ all emissions by length dist
ViennaRNAGz_mangle	91,187	14,397	ViennaRNAG+ multiloop mismatches
ViennaRNAGz_bulge2_ld_mdangle	91,977	14,557	ViennaRNAG+ explicit 1,2 bulges + + all length dist + multiloop mismatches

tacking less important than expected



The four ingredients of single-sequence RNA secondary structure prediction



TORNADO allows us to select all four ingredients independently

(1) Probabilistic models can describe the complex features of the nearest-neighbor model of RNA folding and more

(2) Probabilistic models are easily trainable on known data, while they use identical algorithms for inference

(3) Probabilistic models (trained properly) perform as well or better. Performance ceiling.

Complex model easily overfit

Complex RNA models learn many features of the structures

Adequate benchmarking requires to use **training sets** with different folds than those in **testing sets**

Last Thought

All methods, thermodynamic or statistic (both probabilistic and weighted), can be formaly expressed as context-free grammars, thus the architecture is independent of the scoring scheme.

RNA Folding Servers to also include probabilistic scorings?