RNA bioinformatics: Still combinatorial in 2023?

Yann Ponty

AMIBio@LIX, Institut Polytechnique de Paris
UMR 7161 CNRS & École Polytechnique

Who am I?

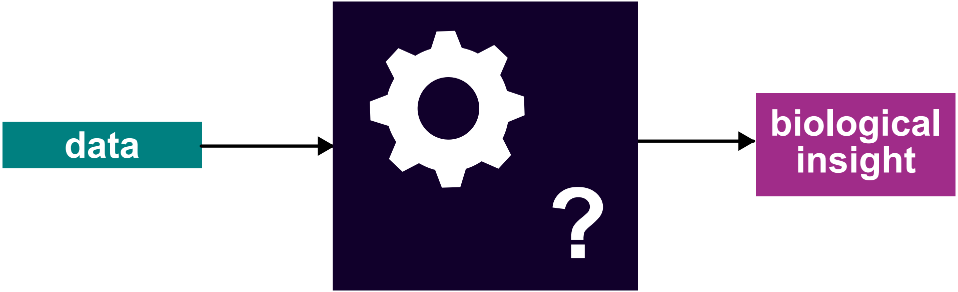
- Initial background in Computer Science
- Dabbled in Theoretical Comp Sci/Discrete Maths (random gen, disc algos)
- Contributing to RNA structural/omics bioinfo
- Cultural shock getting into Bioinformatics

  Old enough to remember the first "AI Bioinfo winter" (SVMs)

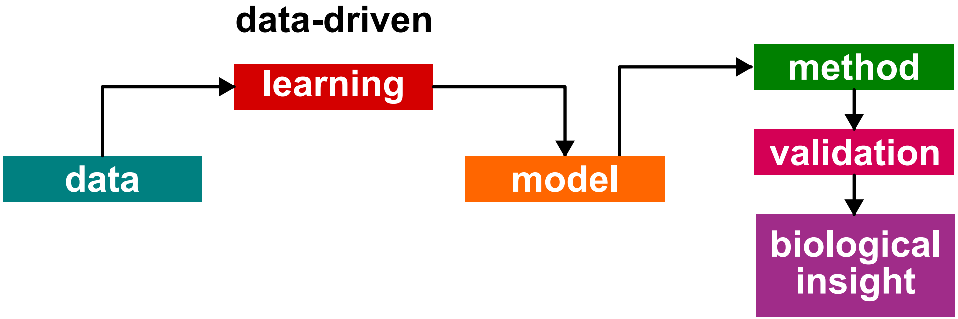
Strong interest in defining/enforcing scientific standards

- Associate editor@OUP Bioinformatics
- Proceedings chair for ISMB/ECCB 2023 (with Sushmita Roy)
- President for committee rewarding best French PhD in Computer Science

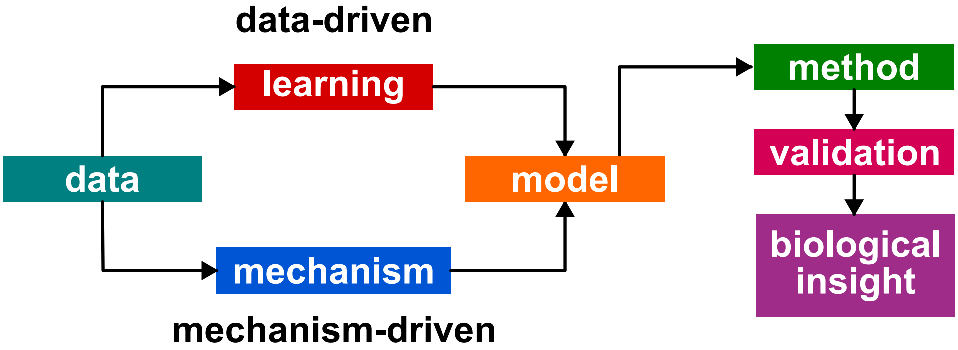
A personal take on predictive Bioinformatics



A personal take on predictive Bioinformatics



A personal take on predictive Bioinformatics



Method dev. as a modeling discipline:

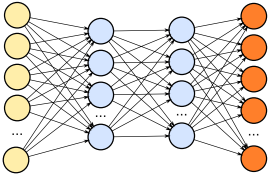
Mechanism-driven model + Exact/deterministic algorithms

→ Performance as (in)validation of model

Machine Learning (ML): The beauty...

Machine Learning as a tool for scientific discovery

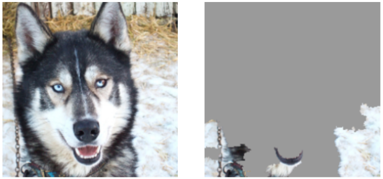
- Great promises
- Self-improving methods
- Generates/prioritizes hypotheses
- Available workforce (ubiquitous in curriculums)
- Highly promoted/funded by research institutions and glamorous journals. . .



Machine Learning (ML): The beauty... and the beast

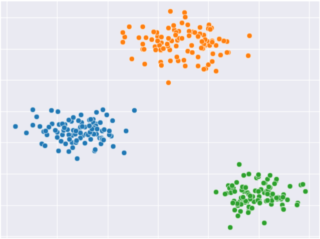
Multiple (potential) pitfalls for ML in Bio*:

- Tricky evaluation (data leakage) → Extrapolation/generalization???
- Reproducibility issues (code/datasets availability, stability, retraining)
- Fishing expeditions/storytelling, selective reporting
- Educational deadend?
- Future(?) ecological disaster? Random blue checkmarks AI zealots on Twitter (grumble. . .)



(a) Husky classified as wolf (b) Explanation

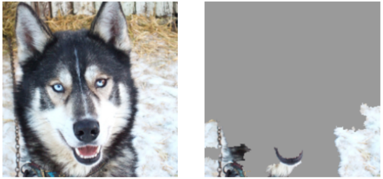
[Ribeiro et al, KDD'16]



Machine Learning (ML): The beauty... and the beast

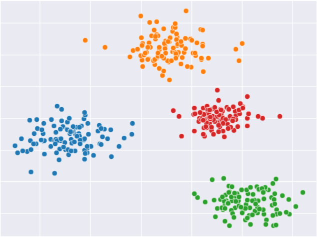
Multiple (potential) pitfalls for ML in Bio*:

- Tricky evaluation (data leakage) → Extrapolation/generalization???
- Reproducibility issues (code/datasets availability, stability, retraining)
- Fishing expeditions/storytelling, selective reporting
- Educational deadend?
- Future(?) ecological disaster? Random blue checkmarks AI zealots on Twitter (grumble. . .)



(a) Husky classified as wolf (b) Explanation

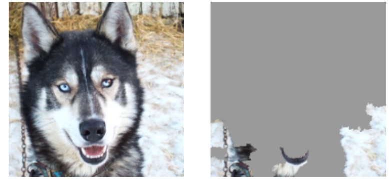
[Ribeiro et al, KDD'16]



Machine Learning (ML): The beauty... and the beast

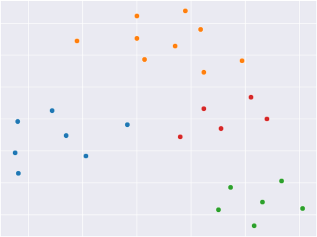
Multiple (potential) pitfalls for ML in Bio*:

- Tricky evaluation (data leakage) → Extrapolation/generalization???
- Reproducibility issues (code/datasets availability, stability, retraining)
- Fishing expeditions/storytelling, selective reporting
- Educational deadend?
- Future(?) ecological disaster? Random blue checkmarks AI zealots on Twitter (grumble. . .)



(a) Husky classified as wolf (b) Explanation

[Ribeiro et al, KDD'16]



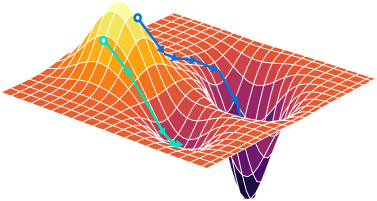
Machine Learning (ML): The beauty... and the beast

Multiple (potential) pitfalls for ML in Bio*:

- Tricky evaluation (data leakage) → Extrapolation/generalization???
- Reproducibility issues (code/datasets availability, stability, retraining)
- Fishing expeditions/storytelling, selective reporting
- Educational deadend?
- Future(?) ecological disaster? Random blue checkmarks AI zealots on Twitter (grumble. . .)

Available upon request

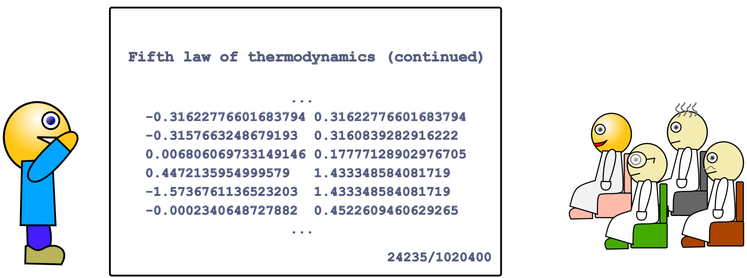
aka iff I'm in a good mood,
PhD/postdoc still in lab, HDDs haven't burned,
pharma hasn't expressed interest in data...



Machine Learning (ML): The beauty... and the beast

Multiple (potential) pitfalls for ML in Bio*:

- Tricky evaluation (data leakage) → Extrapolation/generalization???
- Reproducibility issues (code/datasets availability, stability, retraining)
- Fishing expeditions/storytelling, selective reporting
- Educational deadend?
- Future(?) ecological disaster? Random blue checkmarks AI zealots on Twitter (grumble. . .)

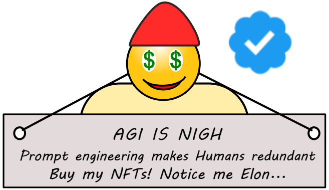


```
Fifth law of thermodynamics (continued)

                    ...
 -0.31622776601683794 0.31622776601683794
 -0.3157663248679193  0.3160839282916222
 0.006806069733149146 0.17777128902976705
 0.4472135954999579   1.433348584081719
 -1.5736761136523203  1.433348584081719
 -0.0002340648727882  0.4522609460629265
                    ...

                              24235/1020400
```

Machine Learning (ML): The beauty... and the beast

Multiple (potential) pitfalls for ML in Bio*:

- Tricky evaluation (data leakage) → Extrapolation/generalization???
- Reproducibility issues (code/datasets availability, stability, retraining)
- Fishing expeditions/storytelling, selective reporting
- Educational deadend?
- Future(?) ecological disaster? Random blue checkmarks AI zealots on Twitter (grumble. . .)

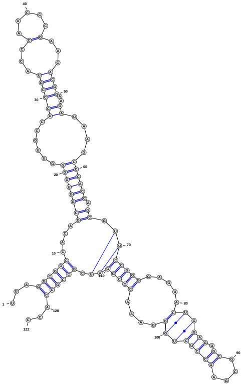

AGI IS NIGH
Prompt engineering makes Humans redundant
Buy my NFTs! Notice me Elon...

RNA structure(s)

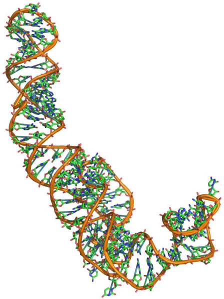
RNA = Linear Polymer = Nucleotides sequence $w \in \{A, C, G, U\}^\star$

```
UUAGGCGGCCACAGC
GGUGGGGUUGCCUCC
CGUACCCAUCCCGAA
CACGGAAGAUAAGCC
CACCAGCGUUCCGGG
GAGUACUGGAGUGCG
CGAGCCUCUGGGAAA
CCCGGUUCGCCGCCA
CC
```

Primary struct.



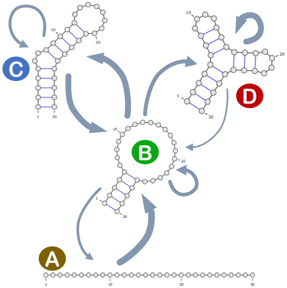
Secondary (2D) struct.



Tertiary ($\approx$ 3D) struct.

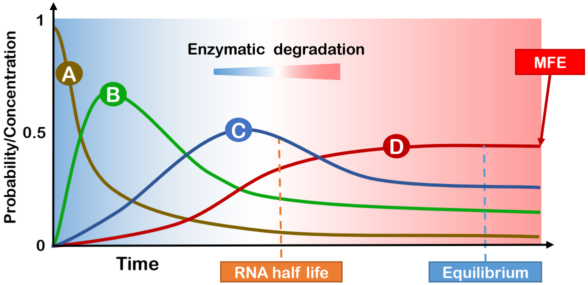
5s rRNA (PDBID: 1K73:B)

Paradigms in RNA structural bioinformatics



A – Kinetic Landscape

Continuous-time Markov chain

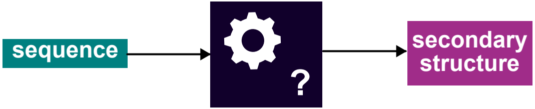


B – Evolution of concentrations

Free-energy $E : \Sigma^{\star} \times \mathcal{S} \to \mathbb{R}$, at Boltzmann equilibrium $\mathbb{P}(S \mid w) \propto e^{-E(w,S)/RT}$

- Minimum Free-Energy (MFE): Functional structure = Most stable/probable
- Partition function: Equilibrium properties of Boltzmann ensemble
- Kinetics: Finite-time evolution of concentrations/probabilities

A crowded ML field for RNA 2D prediction



Method	Output	PKs?	Architecture	Availability
CONTRAfold	Pairwise contacts	No	CLLM	Code+weights+webserver
EternaFold	Pairwise contacts	No	CLLM	Code+weights+webserver
DMfold	DBN	Yes	bi-LSTM	Code only
RNA-state-inf	Binary paired/unpaired	N/A	bi-LSTM	Code only
SPOT-RNA2	Pairwise contacts	Yes	CNN	Code+weights+webserver
CROSS	Binary paired/unpaired	N/A	CNN-like	Webserver
RPRes	Binary paired/unpaired	N/A	bi-LSTM+CNN	Code only
2dRNA	Pairwise contacts	Yes	bi-LSTM+CNN	Webserver
2dRNA-LD	Pairwise contacts	Yes	bi-LSTM+CNN	Webserver
SPOT-RNA	Pairwise contacts	Yes	CNN+bi-LSTM	Code+weights+webserver
MXfold2	Pseudo-dG	No	CNN+bi-LSTM	Code+weights+webserver
CNNFold	Pairwise contacts	Yes	CNN(NxN input)	Code+weights
UFold	Pairwise contacts	Yes	CNN(NxN input)	Code+weights+webserver
CDPfold	DBN	No	CNN(N×Ninput)	Code
E2Efold	Pairwise contacts	Yes	Transformer+CNN	Code+weights
ATTfold	Pairwise contacts	Yes	Transformer+CNN	No

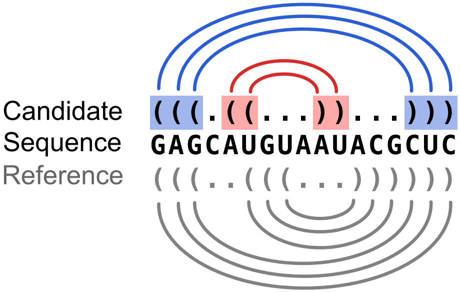
[Wu *et al*, Briefings in Bioinfo 2023]

Performances of 2D structure prediction

RNAStrand benchmark

[Adronescu *et al*, BMC Bioinf 2008]

Method	F_1
RNAfold 1.8.5	0.737
UNAfold 3.8	0.725
RNAstructure 5.7	0.744



Candidate (((.((...))...)))

Sequence GAGCAUGUAAUACGCUC

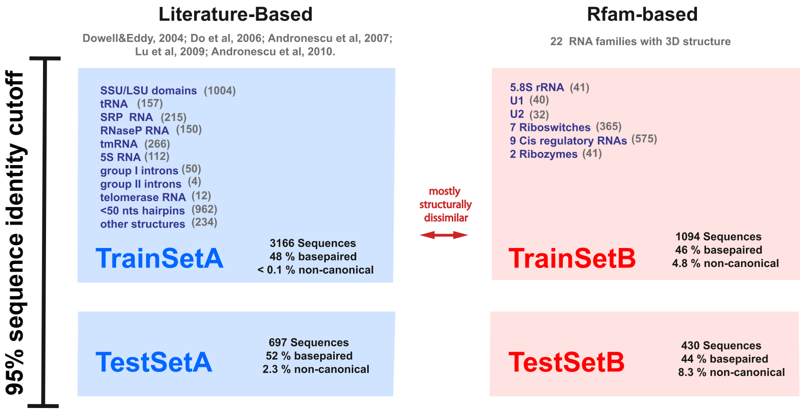
Reference (((..((()...))))))

Sens = 3/6 = 0.5 PPV = 3/5 = 0.6

F_1 = 0.545...

$$F_1\text{-score} = \frac{2 \times \text{PPV} \times \text{Sens}}{\text{PPV} + \text{Sens}}$$

The TORNADO dataset



95% sequence identity cutoff

Literature-Based

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

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)

TrainSetA

3166 Sequences
48 % basepaired
< 0.1 % non-canonical

TestSetA

697 Sequences
52 % basepaired
2.3 % non-canonical

mostly structurally dissimilar

Rfam-based

22 RNA families with 3D structure

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

TrainSetB

1094 Sequences
46 % basepaired
4.8 % non-canonical

TestSetB

430 Sequences
44 % basepaired
8.3 % non-canonical

[Rivas *et al*, RNA 2012]

TrainSetA vs TestSetA: 95% sim. cutoff → Learn *k*-mer to template association

(May happen even for extreme cutoffs)

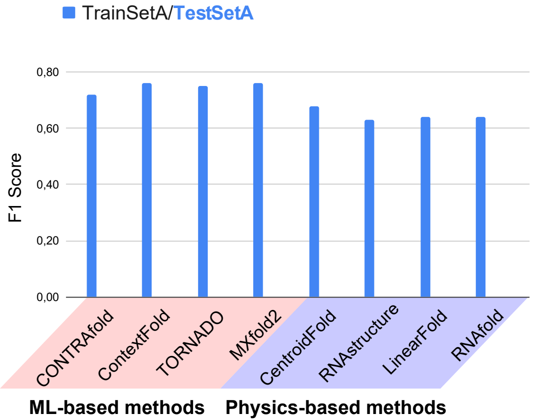
TrainSetA vs TestSetB: Rewards learning structurally generalizable models

Performances of 2D structure prediction

RNAStrand benchmark

[Adronescu *et al*, BMC Bioinf 2008]

Method	F_1
RNAfold 1.8.5	0.737
UNAfold 3.8	0.725
RNAstructure 5.7	0.744



[Sato *et al*, Nature Comm 2021]

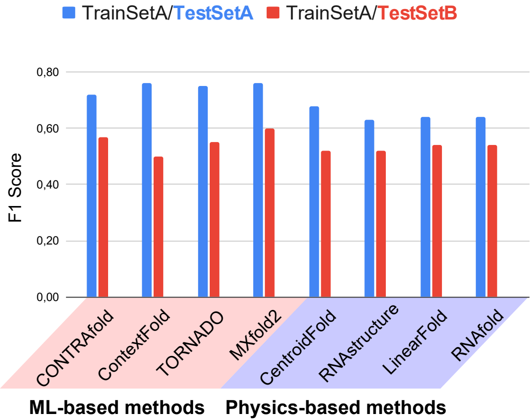
$$\text{F}_1\text{-score} = \frac{2 \times \text{PPV} \times \text{Sens}}{\text{PPV} + \text{Sens}}$$

Performances of 2D structure prediction

RNAStrand benchmark

[Adronescu *et al*, BMC Bioinf 2008]

Method	F_1
RNAfold 1.8.5	0.737
UNAfold 3.8	0.725
RNAstructure 5.7	0.744

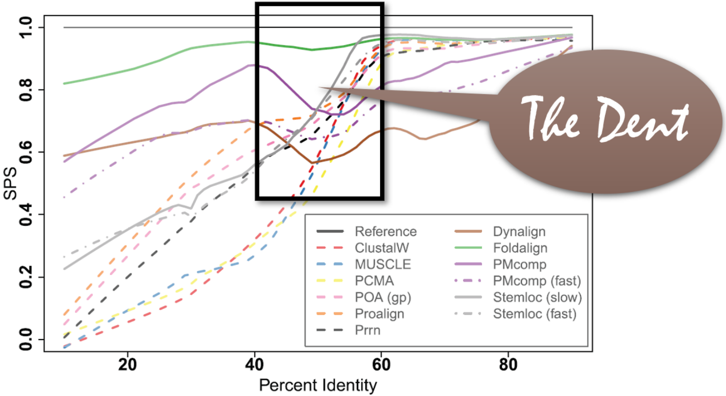


[Sato *et al*, Nature Comm 2021]

$$F_1\text{-score} = \frac{2 \times \text{PPV} \times \text{Sens}}{\text{PPV} + \text{Sens}}$$

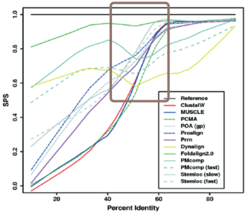
Biased benchmarks: precedent in comparative folding/alignment

Bralibase: Benchmark for comp. RNA folding [Gardner,Wilm & Washietl, NAR 2005]

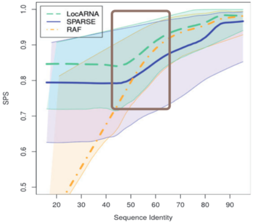


[Löwes *et al*, Briefings in Bioinfo 2016]

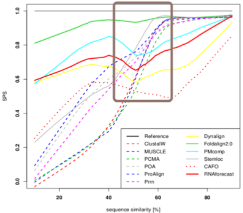
Biased benchmarks: precedent in comparative folding/alignment



[Gardner *et al*, NAR 2005]



[Will *et al*, Bioinformatics 2015]



[Höchsmann *et al*, Unpublished]

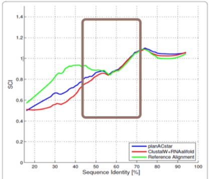
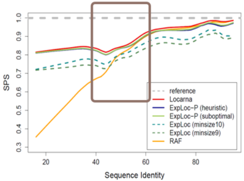
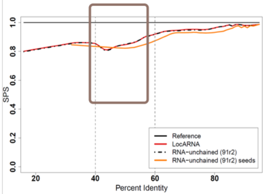


Figure 4 Correlation between gap content and SCI scores.

[Bremges *et al*, BMC Bioinfo, 2010]



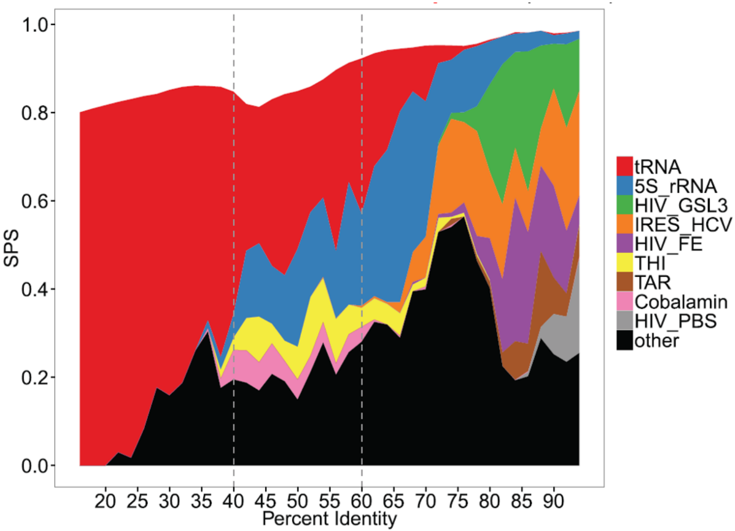
[Schmiedl *et al*, RECOMB 2012]



[Bourgeade *et al*, J Comp Biol, 2015]

[Löwes *et al*, Briefings in Bioinfo 2016]

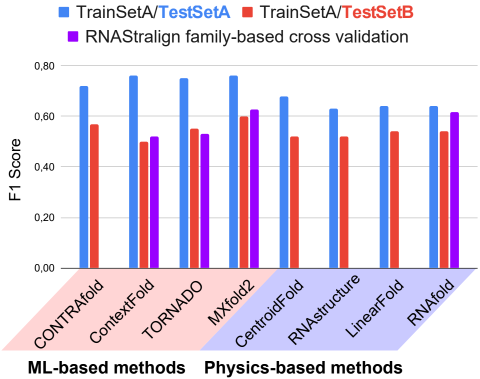
Biased benchmarks: precedent in comparative folding/alignment



[Löwes *et al*, Briefings in Bioinfo 2016]

The (nc)RNA datasphere

- ► 34M sequences, inc 22M presumably structured (RNACentral)
- ► 4000+ functional ncRNA families (RFAM)
- ► 250-300 non-redundant 3D models (PDB)



[Sato *et al*, Nature Comm 2021]

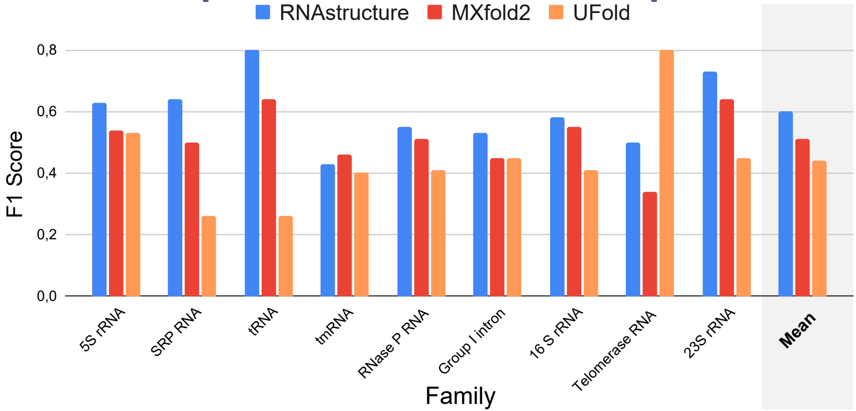
Existing methods trained on datasets:

- ► highly-redundant sequence-wise
- ► low-diversity structure-wise

Do ML methods generalize to new structures?
(Do ML perfs translate into *new* biological insight?)

Generalization to new families/structures remains problematic

[Szikszai *et al*, Bioinformatics 2022]



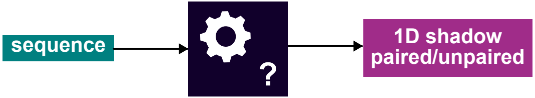
Family-fold cross-validation on ArchiveII dataset [Sloma & Mathews, RNA 2016]

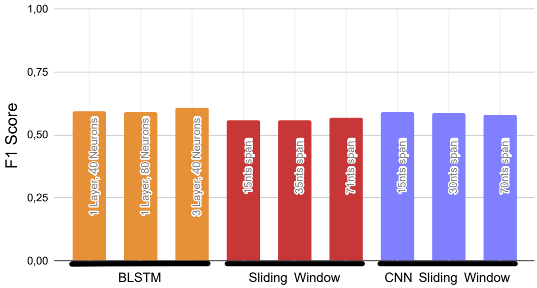
3974 RNAs of length 77-438 (large rRNAs split into smaller domains)

What if you had access to (unbounded) additional data?

Idea: Assess NN models’ capacity to emulate RNAfold on random sequences

[Flamm *et al*, Frontiers in Bioinfo 2022]



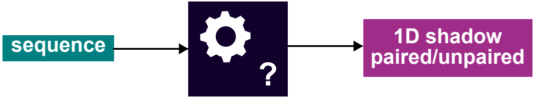


Perfs *plateau* at 80k seq/structs (70nts)

What if you had access to (unbounded) additional data?

Idea: Assess NN models’ capacity to emulate RNAfold on random sequences

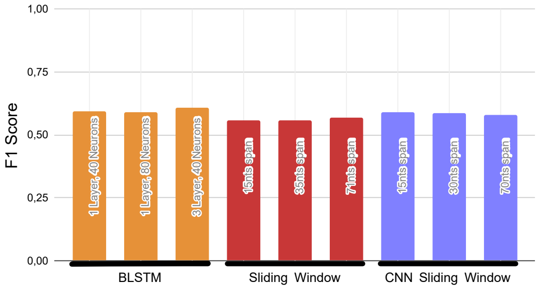
[Flamm *et al*, Frontiers in Bioinfo 2022]



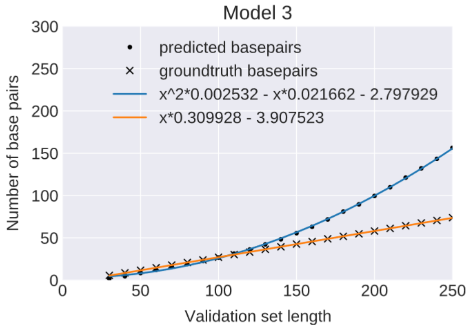
sequence

?

1D shadow paired/unpaired

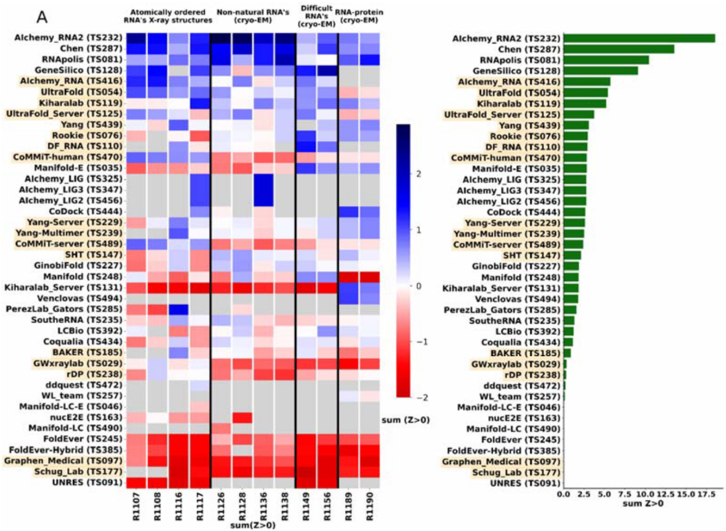


Perfs *plateau* at 80k seq/structs (70nts)



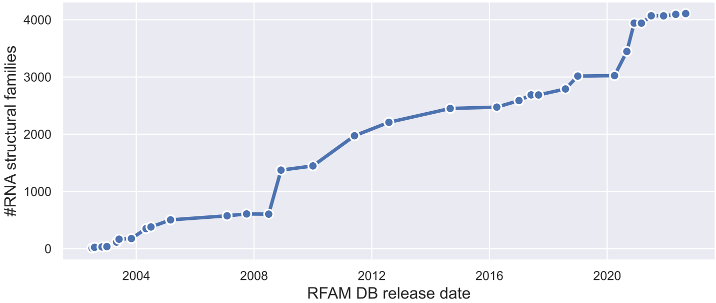
Popular CNN predicts $\Theta(n^2)$ BPs!

RNA 3D structure: No AlphaFold moment at CASP15



[Das *et al*, under review]

Conclusions and musings



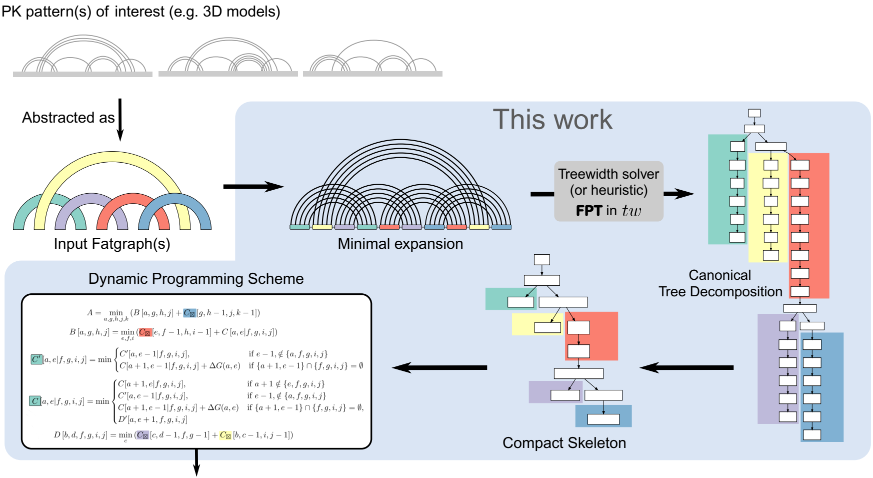
- Still a need for improved RNA prediction (possibly ML-based)
- Purely combinatorial methods still ± state-of-the-art for new families...
- Hybrid approaches *à la* MxFold2: Best of both worlds?
- Assessing intrinsic limits of architectures: RNAFold as surrogate model

Conclusions and musings

So what's special about RNA?

- Modular but combinatorial structure
- New folds being routinely discovered (+ can be designed)
- Relatively scarce 3D data
- Opportunity: Tons of probing data (ML)
- Potential of LLMs/transformers (incoming)
- Pseudoknots-ready algorithms

Automated derivation of folding algorithms inc. pseudoknots



$$A = \min_{a,g,h,j,k} \left(B\,[a,g,h,j] + C_{\boxtimes}[g,h-1,j,k-1]\right)$$

$$B\,[a,g,h,j] = \min_{e,f,i} \left(C_{\boxtimes}[e,f-1,h,i-1] + C\,[a,e|f,g,i,j]\right)$$

$$\boxed{C'}[a,e|f,g,i,j] = \min \begin{cases} C'[a,e-1|f,g,i,j], & \text{if } e-1, \notin \{a,f,g,i,j\} \\ C[a+1,e-1|f,g,i,j] + \Delta G(a,e) & \text{if } \{a+1,e-1\} \cap \{f,g,i,j\} = \emptyset \end{cases}$$

$$\boxed{C}[a,e|f,g,i,j] = \min \begin{cases} C[a+1,e|f,g,i,j], & \text{if } a+1 \notin \{e,f,g,i,j\} \\ C'[a,e-1|f,g,i,j], & \text{if } e-1, \notin \{a,f,g,i,j\} \\ C[a+1,e-1|f,g,i,j] + \Delta G(a,e) & \text{if } \{a+1,e-1\} \cap \{f,g,i,j\} = \emptyset, \\ D'[a,e+1,f,g,i,j] & \end{cases}$$

$$D\,[b,d,f,g,i,j] = \min_{c} \left(C_{\boxtimes}\,[c,d-1,f,g-1] + C_{\boxtimes}\,[b,c-1,i,j-1]\right)$$

General algorithm for Fatgraph MFE folding problem
Correct for any input sequence (+ `C/C++` Code generation)

Complexity: $O(n^{tw})$ or $O(n^{tw+1})$ for simple energy models
$O(n^{tw+1})$ for full Turner model

[Marchand *et al*, WABI 2023]

Conclusions and musings

(RNA) community needs to enforce stricter standards for ML publications:

- Enforce datasets and source code availability
  [Szikszai *et al*, Bioinfo'22] found 4/8 recent DL methods non-functional
- Realistic retraining mandatory
  Precondition for self-improvement, benchmarking of novel methods
- Consider mechanistic and ML methods as largely incomparable
- Better datasets/benchmarks needed, but perhaps not sufficient
- Sequence-based leakage should be systematically investigated

What are ze questions?

Many thanks to:

- Ze SFBI for putting ze session together
- Ze ISCB (Diane and Steven)
- Ze whole proceedings program committee (Sushmita + 20^{+} ACs + 200^{+} members)
- Ze (nc)RNA bioinfo community for being generally awesome
- Ze AMIBio team at Ecole Polytechnique (Sebastian and Sarah)
- You for letting me indulge in zis – typically French – rant

Time for ~~pitchforks and tomatoes~~ questions?

What are ze questions?

Many thanks to:

- Ze SFBI for putting ze session together
- Ze ISCB (Diane and Steven)
- Ze whole proceedings program committee (Sushmita + 20^+ ACs + 200^+ members)
- Ze (nc)RNA bioinfo community for being generally awesome
- Ze AMIBio team at Ecole Polytechnique (Sebastian and Sarah)
- You for letting me indulge in zis – typically French – rant

Time for ~~pitchforks and tomatoes~~ questions?

What are ze questions?

Many thanks to:

- Ze SFBI for putting ze session together
- Ze ISCB (Diane and Steven)
- Ze whole proceedings program committee (Sushmita + 20^+ ACs + 200^+ members)
- Ze (nc)RNA bioinfo community for being generally awesome
- Ze AMIBio team at Ecole Polytechnique (Sebastian and Sarah)
- You for letting me indulge in zis – typically French – rant

Time for ~~pitchforks and tomatoes~~ questions?

What are ze questions?

Many thanks to:

- Ze SFBI for putting ze session together
- Ze ISCB (Diane and Steven)
- Ze whole proceedings program committee (Sushmita + 20^+ ACs + 200^+ members)
- Ze (nc)RNA bioinfo community for being generally awesome
- Ze AMIBio team at Ecole Polytechnique (Sebastian and Sarah)
- You for letting me indulge in zis – typically French – rant

Time for ~~pitchforks and tomatoes~~ questions?

What are ze questions?

Many thanks to:

- Ze SFBI for putting ze session together
- Ze ISCB (Diane and Steven)
- Ze whole proceedings program committee (Sushmita + 20^+ ACs + 200^+ members)
- Ze (nc)RNA bioinfo community for being generally awesome
- Ze AMIBio team at Ecole Polytechnique (Sebastian and Sarah)
- You for letting me indulge in zis – typically French – rant

Time for ~~pitchforks and tomatoes~~ questions?

What are ze questions?

Many thanks to:

- Ze SFBI for putting ze session together
- Ze ISCB (Diane and Steven)
- Ze whole proceedings program committee (Sushmita + 20^+ ACs + 200^+ members)
- Ze (nc)RNA bioinfo community for being generally awesome
- Ze AMIBio team at Ecole Polytechnique (Sebastian and Sarah)
- You for letting me indulge in zis – typically French – rant

Time for ~~pitchforks and tomatoes~~ questions?