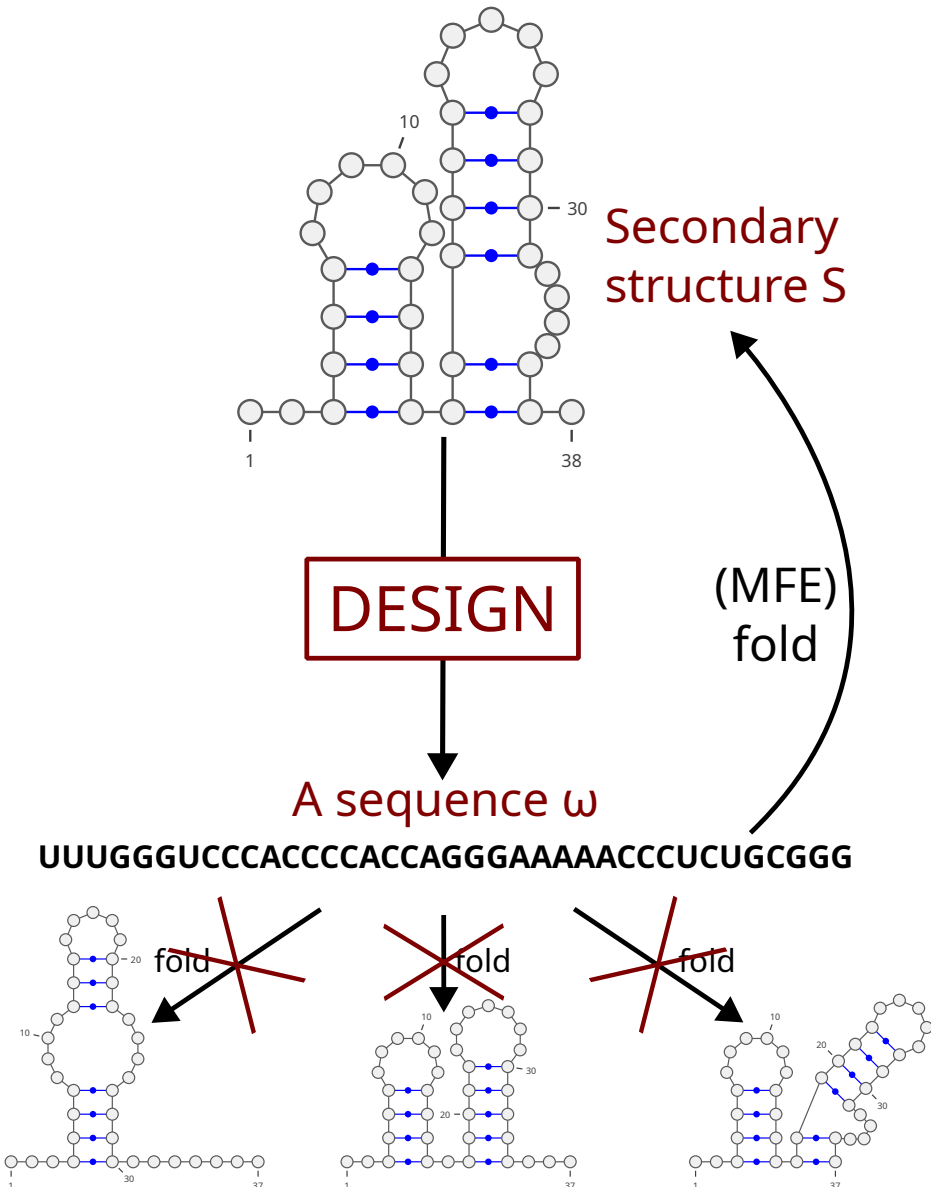
Old dog, new tricks: Exact seeding strategy improves RNA design performances

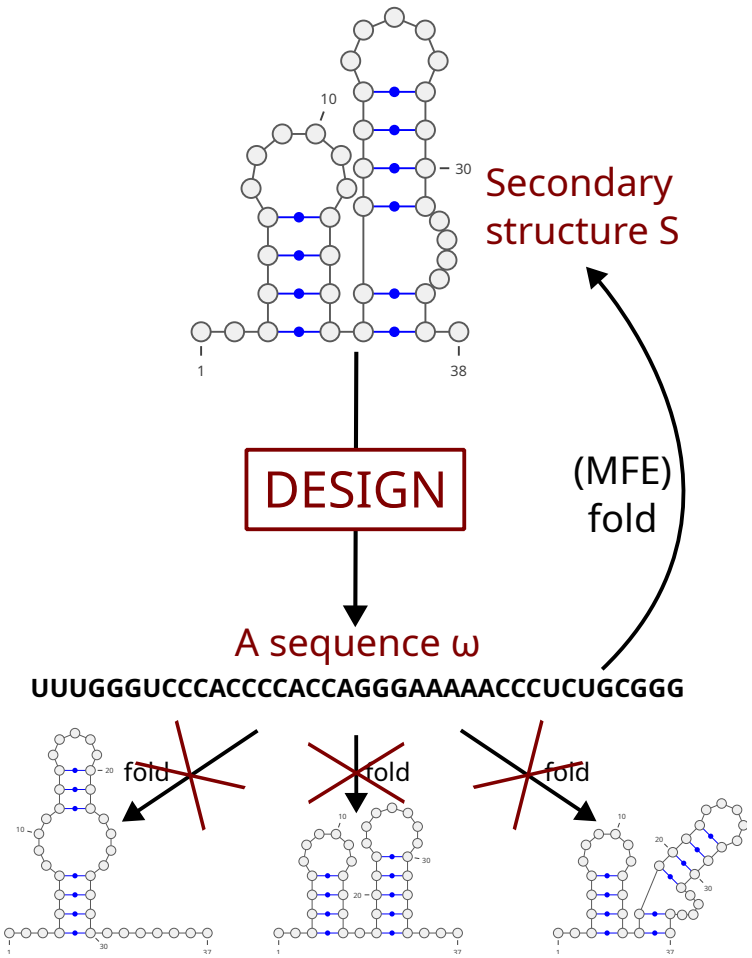
Théo Boury[1], Leonhard Sidl[2,3], Ivo L. Hofacker[2,3], Yann Ponty[1], and Hua-Ting Yao[2]

1, Laboratoire d'Informatique de l'Ecole Polytechnique (LIX; UMR 7161), Institut Polytechnique de Paris, France
2, Department of Theoretical Chemistry, University of Vienna, 1090 Vienna, Austria
3, Faculty of Computer Science, Research Group Bioinformatics and Computational Biology, University of Vienna, 1090 Vienna, Austria

Motivation: RNA Design



Studied problem: RNA Inverse Folding



Problem 1 (RNA Inverse Folding):
Input: A secondary structure S (pseudoknot-free).
Output: An RNA sequence ω with $\forall S' \neq S, \Delta G(\omega, S) < \Delta G(\omega, S')$

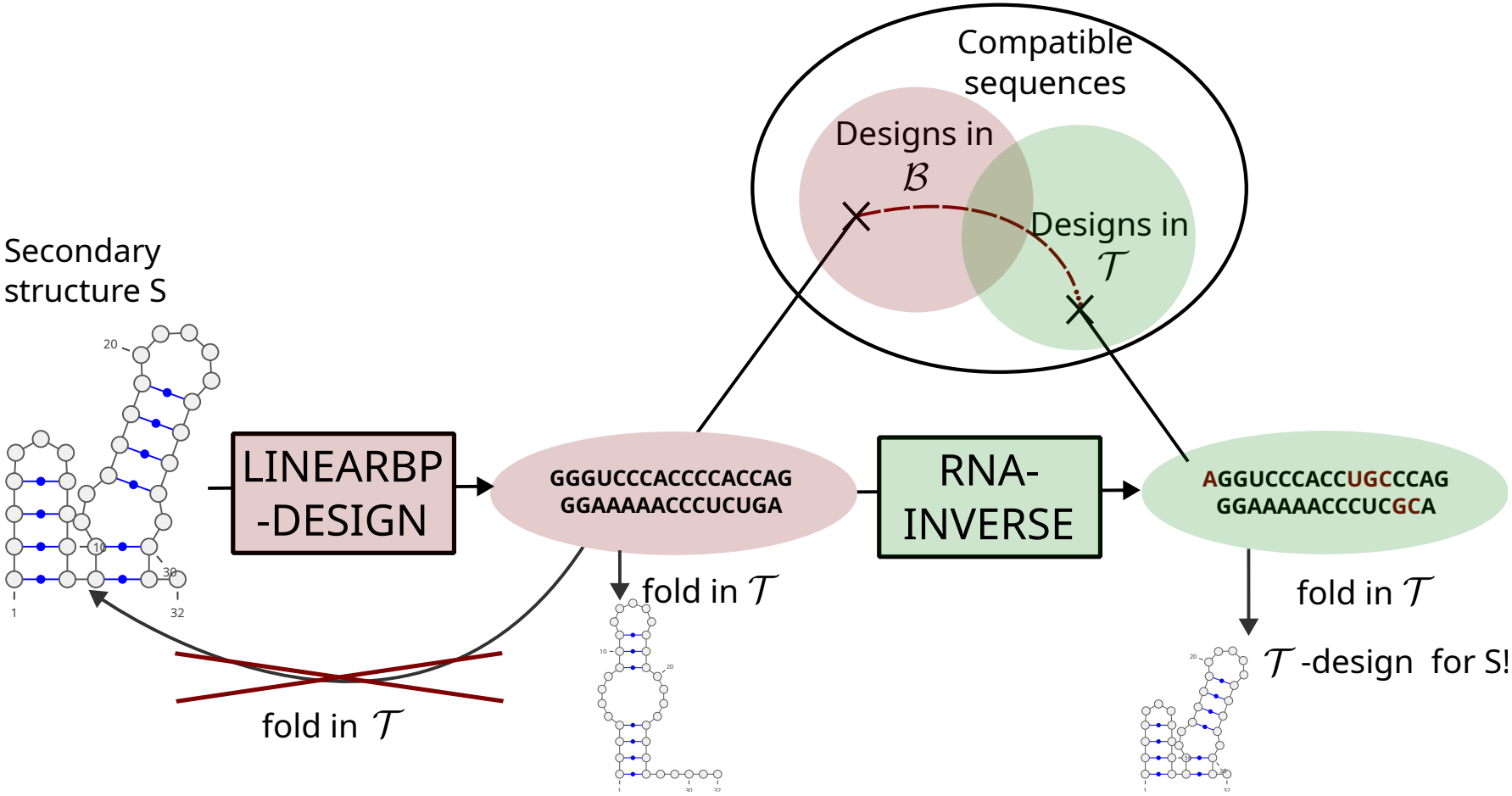
where $\Delta G(\omega, S)$ is the free-energy of ω in:

1. Base pairs maximization model $\mathcal{B}$

[Bonnet *et al*, RECOMB 2018]

2. Turner nearest-neighbor model $\mathcal{T}$

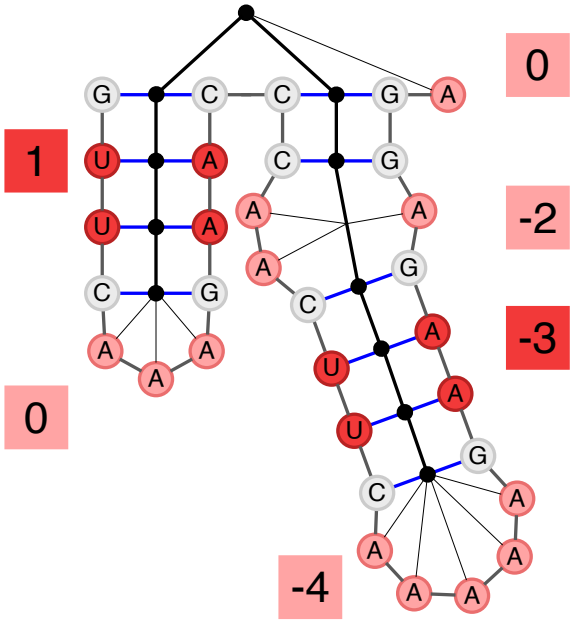
Our result



- A new design tool based on `RNAinverse` + `LinearBPDesign` seeds.
- $\mathcal{B}$-designs are good proxies for $\mathcal{T}$-designs.
- Seeding matters: time and diversity improvements.

Solve RNA Design in $\mathcal{B}$: `LinearBPDesign` [Boury et al, 2024]

2-separated sequence

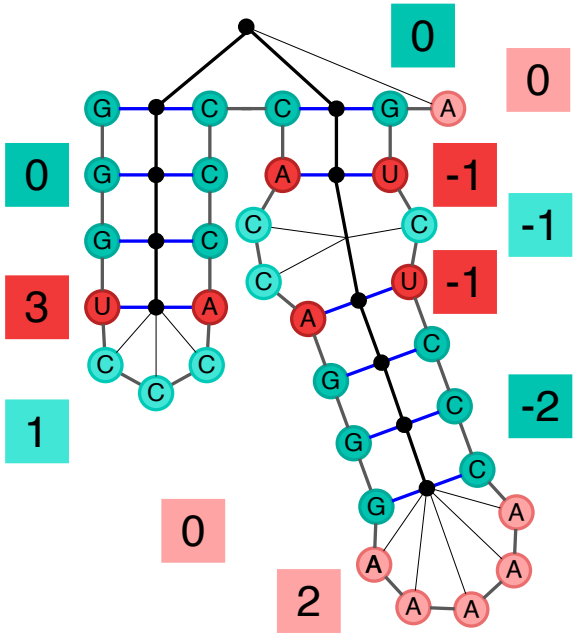


GUUCAAAGAACCCAAC
UUCAAAAAGAAGAGGA

X_U and X_A
can only interact if
$(X_U \bmod 2) = (X_A \bmod 2)$

Adapting toward $\mathcal{T}$: Biseparated sequences [This paper!]

(2, 2)-biseparated sequence



GGGUCCCACCCCACCA
GGGAAAAACCCUCUGA

X_U and X_A can only interact if $(X_U \bmod 2) = (X_A \bmod 2)$

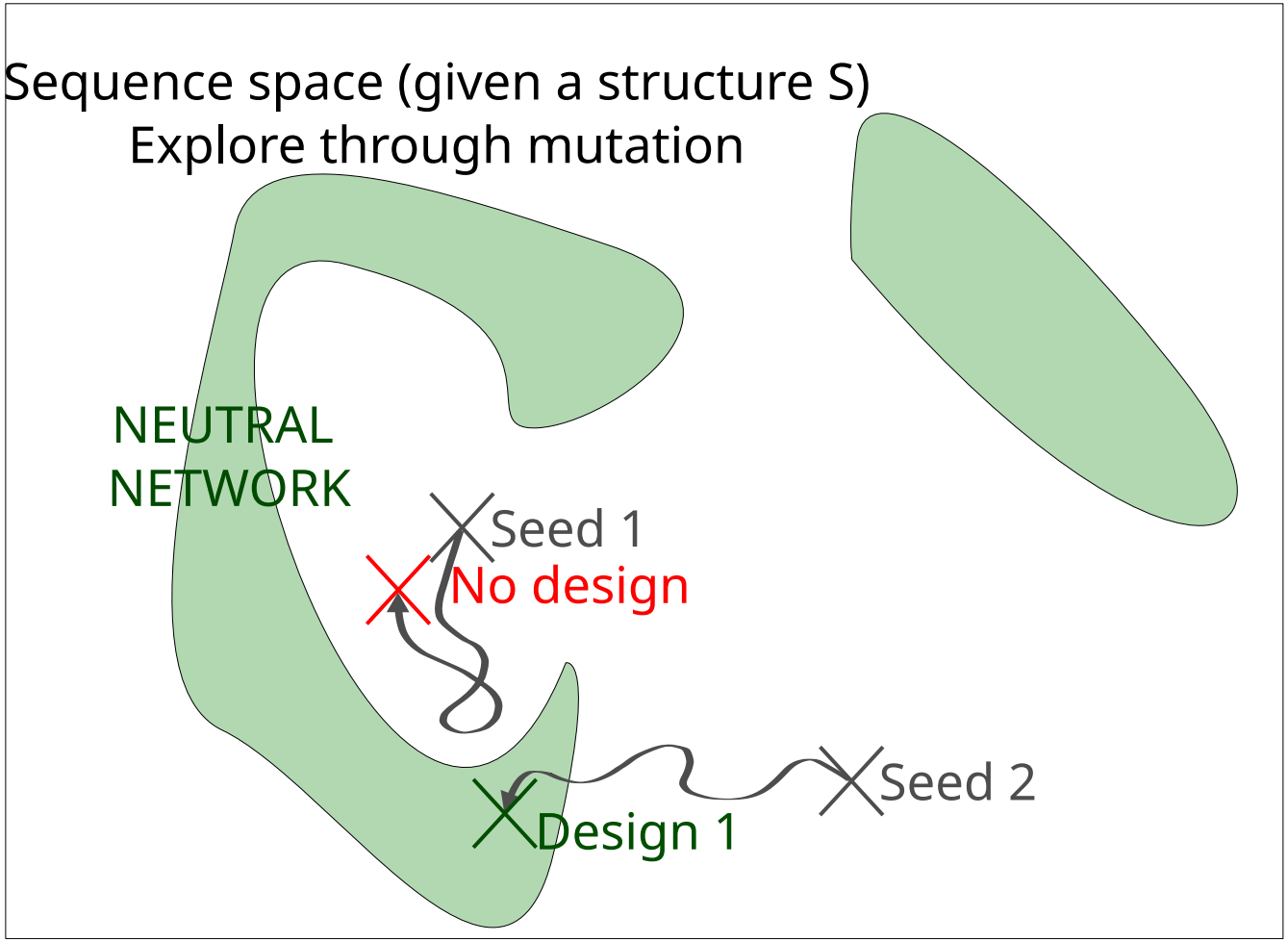
X_G and X_C can only interact if $(X_G \bmod 2) = (X_C \bmod 2)$

Biseparated sequences are computed in linear time

$$\mathrm{p}^{(\xi_{L_A},\xi_{L_C})}_{v\to\mu,(\ell_A,\ell_C)} = \begin{cases} \mathbb{1}_{(l\in\xi_{L_A})\wedge(\mu=\mathrm{A})} + \mathbb{1}_{(l\in\xi_{L_C})\wedge(\mu=\mathrm{C})} & \text{if } v \text{ is leaf} \\ 0 & \text{if } \ell\in\xi_{L_A} \\ & \text{and } \mu\in\{\mathrm{AU},\mathrm{UA}\} \\ 0 & \text{if } \ell\in\xi_{L_C} \\ & \text{and } \mu\in\{\mathrm{GC},\mathrm{CG}\} \\ 1 & \text{if children}(v)=\varnothing \\ \displaystyle\sum_{\substack{\mu' \text{ "proper" assignment} \\ \text{children}(v) \\ \to \Sigma^2\cup\{\mathrm{A},\mathrm{C}\}}} \prod_{\substack{v_i \\ \in\text{children}(v)}} \mathrm{p}^{(\xi_{L_A},\xi_{L_C})}_{v_i\to\mu'(v_i),(\ell'_A,\ell'_C)} & \text{otherwise} \end{cases}$$

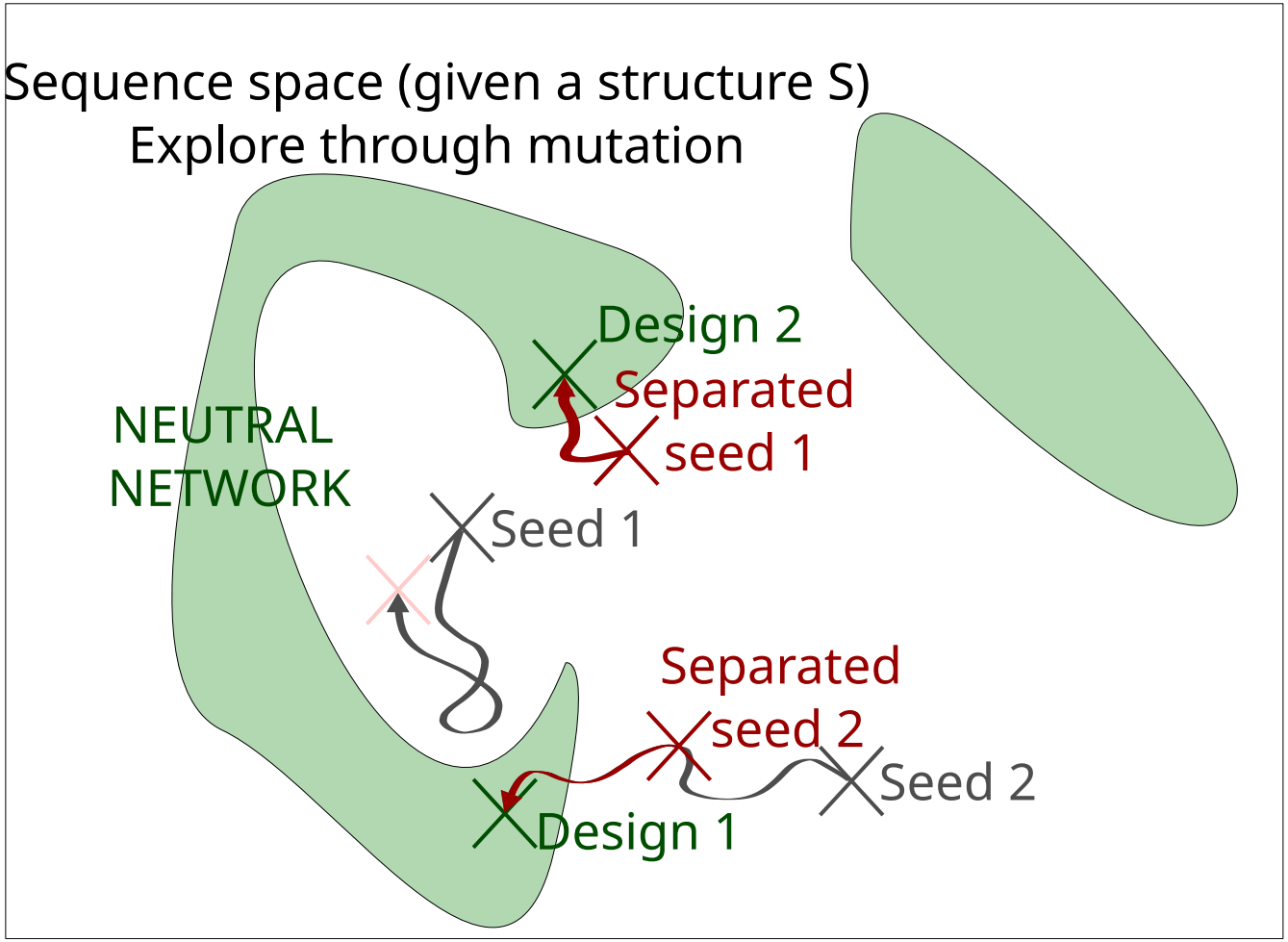
- Find (bi)separated sequences for "most" RNA structures in $O(n)$!

The dog: RNAinverse, a heuristic "simple" tool for $\mathcal{T}$



- RNAinverse was firstly introduced to reach the neutral network by random mutations from uniformly sampled seeds. [Hofacker et al, 1994 (... before me!)]
- Does not run too far from the seeds: good to study them!

The trick: interface RNAinverse with (bi)separated seeds



Questions:

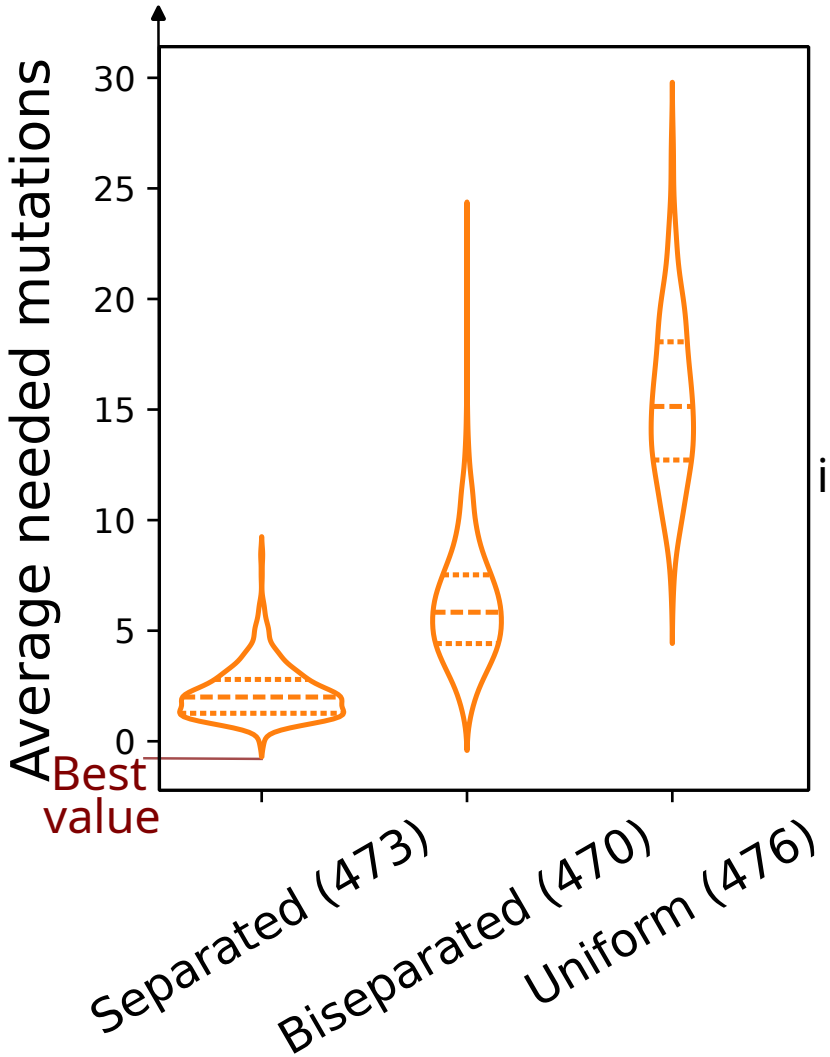
1. Are some (bi)separated seeds directly in the neutral network?
2. Otherwise, how quickly do we reach the neutral network?
3. How much are we covering the neutral network?

1. (Bi)separated sequences are mostly $\mathcal{T}$-designs

Seeds	Number of MFE solved struct. (/2000)	Number of random solved struct. (/1000)	**Total (/3000)**
Uniform	1065	15	**1080**
Separated	1531	952	**2483**
Biseparated	1524	979	**2527**
Uniform ($>$ 300 seqs)	1	0	**1**
Separated ($>$ 300 seqs)	1525	400	**1925**
Biseparated ($>$ 300 seqs)	1392	317	**1709**

- Benchmark: 1000 Random structures + 2000 MFE structures.
- A reasonnable amount of $\mathcal{B}$-designs are $\mathcal{T}$-designs with no use of `RNAinverse`.

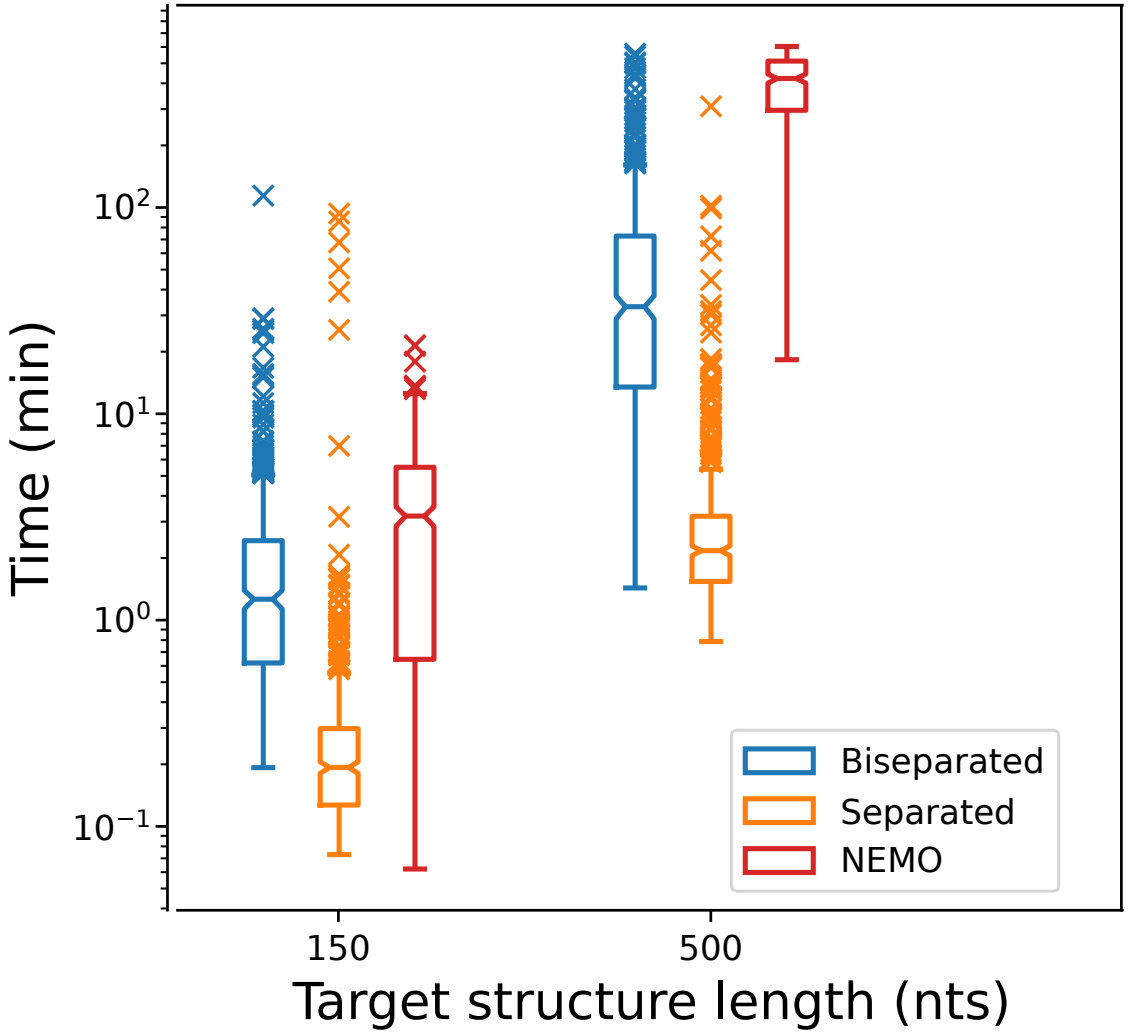
2. $\mathcal{B}$-designs reach a $\mathcal{T}$-designs in a few mutations



(Number of solved instances in parenthesis)

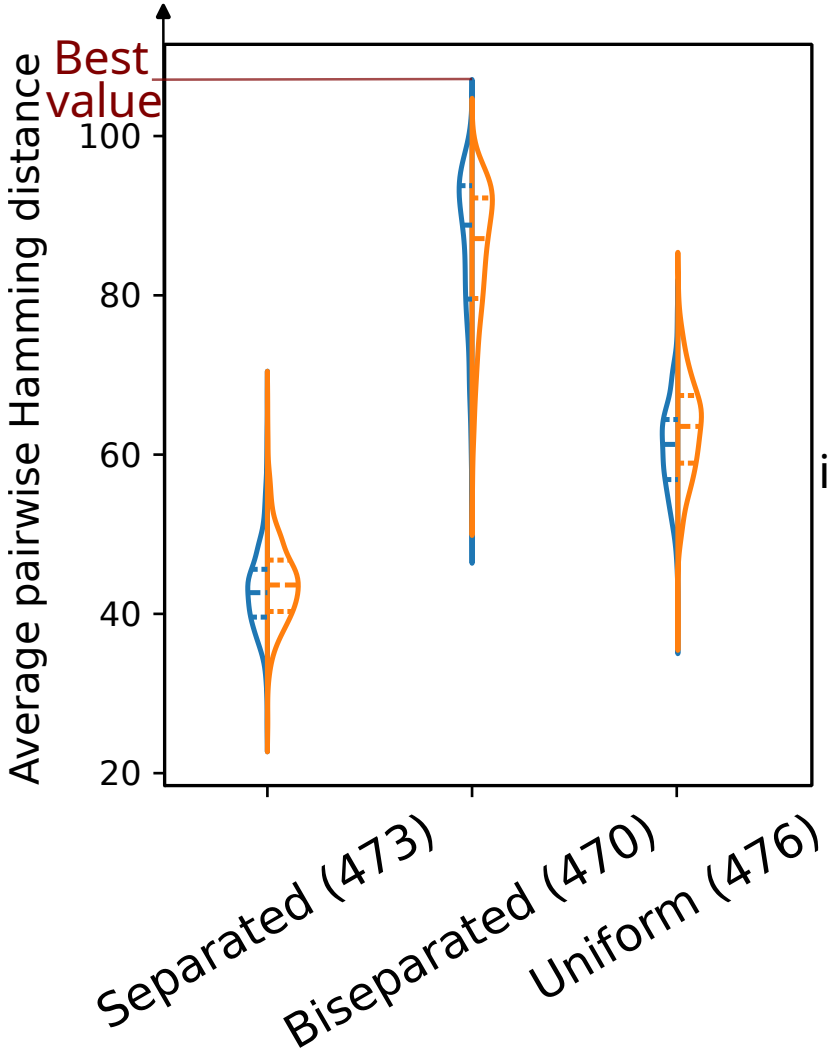
- We sample 100 sequences for each 476 "hard" structure ($\mathcal{B}$-design $\not\to$ $\mathcal{T}$-design).
- Biseparated and Separated seeds are close to be $\mathcal{T}$-designs

2. Our time computations are competitive



- Time benefit from linear time + proximity to the neutral network.
- NEMO solved all structures with at least one solution but is more time-consuming.

3. Biseparated sequences enable diversity!

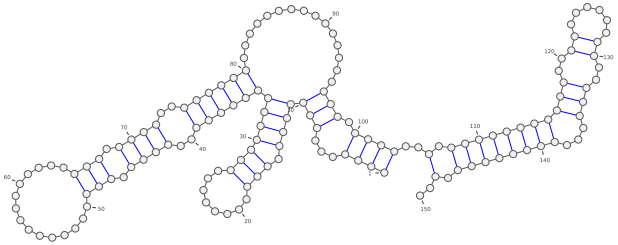


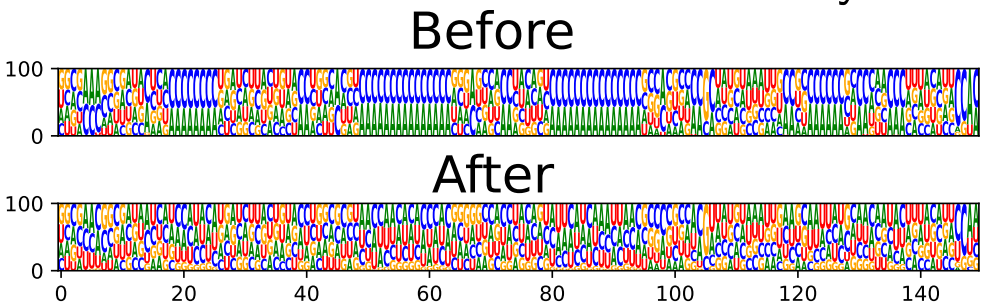
(Number of solved instances in parenthesis)

- Same 100 sequences for each 476 "hard" structure as before.
- More likely to cover the neutral network with biseparated seeds!

“Almost” conclusion

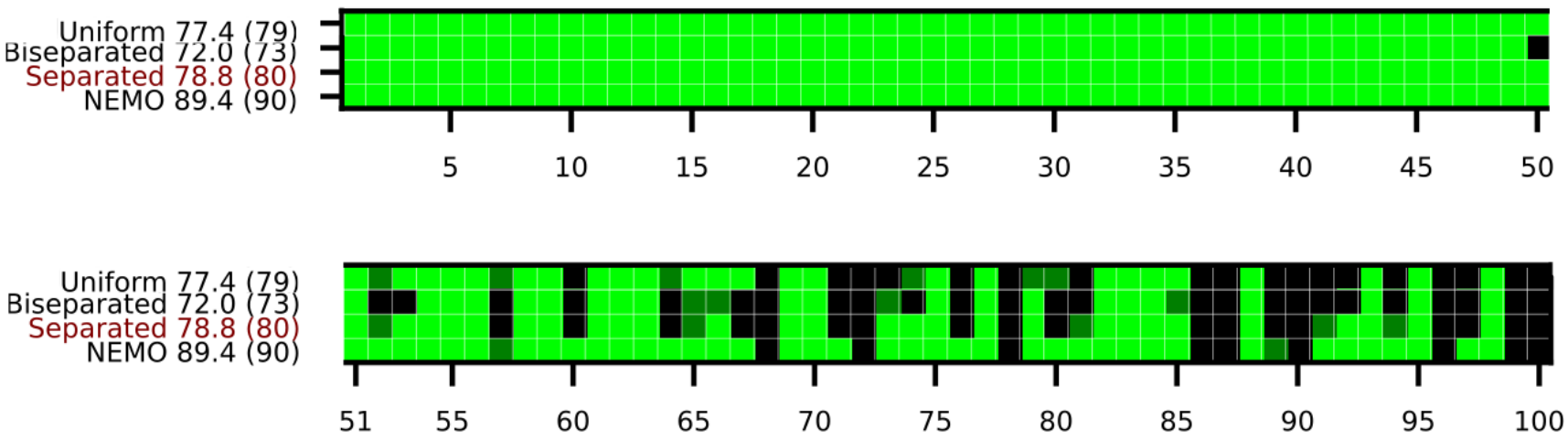
- We revisit `RNAinverse`: old but gold with (bi)separated seeds.
- **Diversity** matters: biseparable seeds are quickly computed and varied.
- We can walk in the neutral network to increase even more diversity:





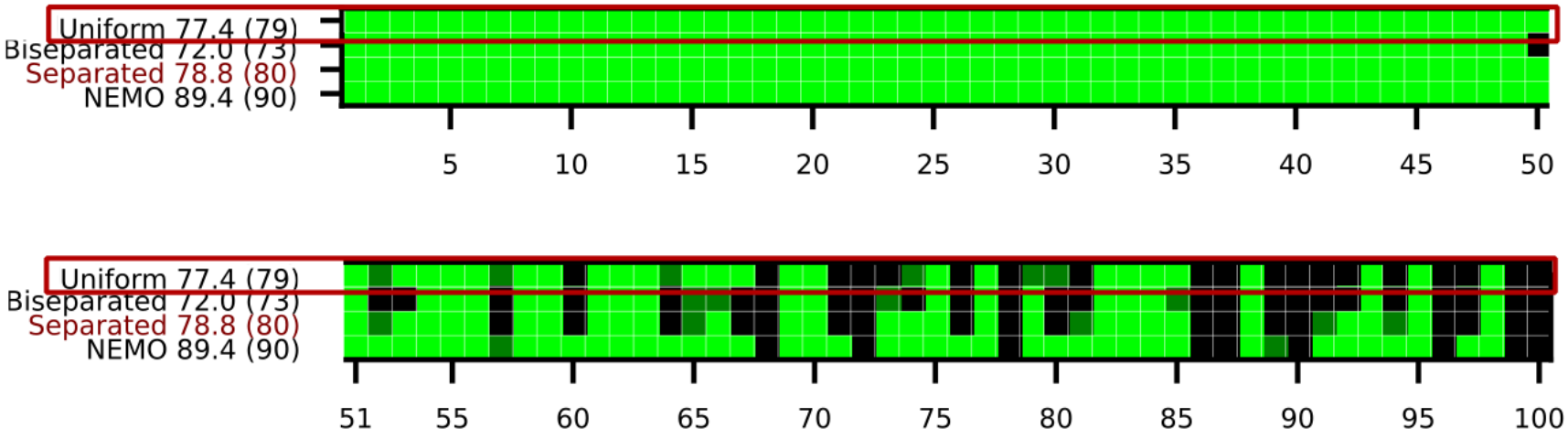
- **Perspective:** Mix “positive” design and “negative” seeds, more comparisons in the paper!

What about the EterRNA 100 benchmark?



- Old dog: `RNAinverse` with no (bi)separated seeds (53)

AAA... is it really about our result?



- Forced As at unpaired positions is all you need to be competitive with the state of the art on `EterRNA` 100!!!

Final thought

We need more benchmarks and evaluations for design!

- We need new benchmarks of objectively hard synthetic structures.
- We should go beyond the "inverse optimization" problem: getting just one highly constraint solution is not enough. (Diversity, GC-content, etc)

Acknowledgements

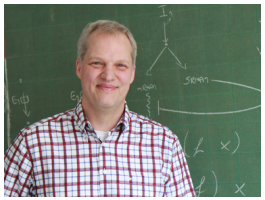
Thanks to...

AMIBio

Yann Ponty

tbi

Hua-Ting Yao



Ivo L. Hofacker

Leonhard Sidl

Special thanks to:

Laurent Bulteau

Sebastian Will

Vladimir Reinharz

FWF

You!

Link to the paper

E4H

EUR BERTIP
(ANR 18EURE0002)
plan France 2030

anr

ENS DE LYON