UNIVERSITY OF CAMBRIDGE

💣 gRNAde
Geometric deep learning for 3D RNA inverse design

Chaitanya K. Joshi, Arian R. Jamasb, Ramon Viñas, Charles Harris, Simon Mathis, Pietro Liò

Computational Biology Workshop, International Conference on Machine Learning, 2023

Forthcoming book chapter in *Methods in Molecular Biology (RNA Design: Methods and Protocols)*

📄 **Preprint** (not up to date): https://arxiv.org/abs/2305.14749

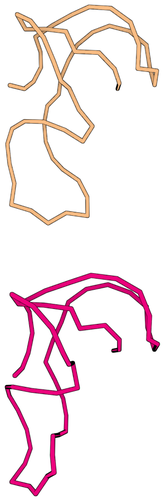
💻 **Codebase**: github.com/chaitjo/geometric-rna-design

Executive summary

Fixed backbone(s) inverse design of RNA sequence

RNA Conformational Ensemble

Extract Backbones



Set of Backbone Geometric Graphs

Multi-state Graph Neural Network Encoder

Equivariant to:

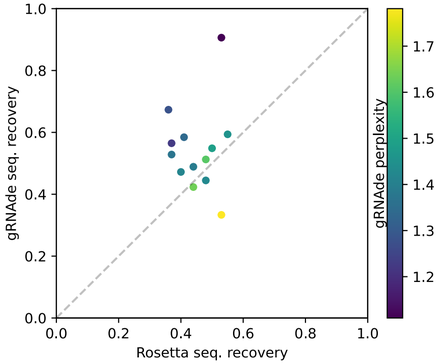
- 3D roto-translations
- node order
- conformation order

Fixed backbone re-design

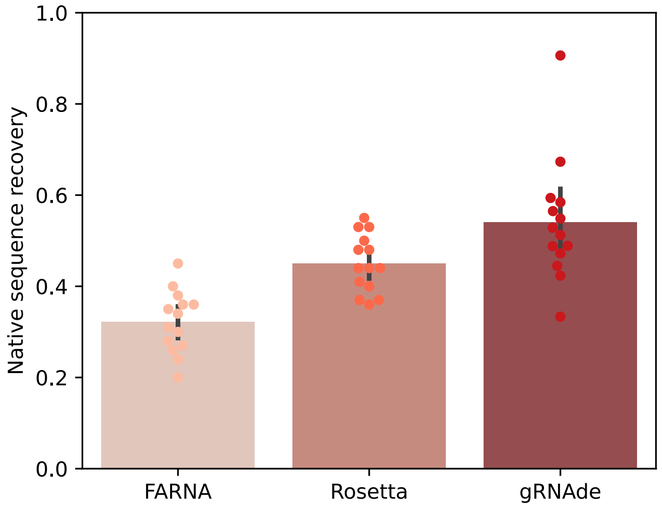
Sequence Decoder

GAGCGU...

RNA Sequence

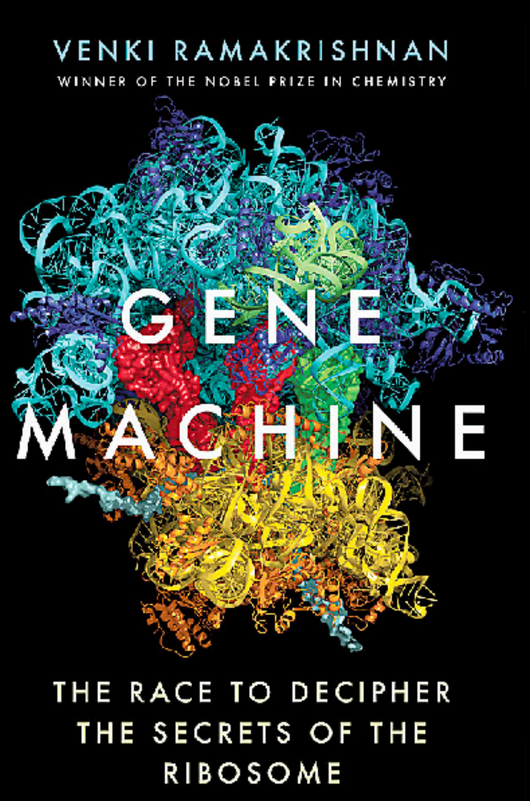


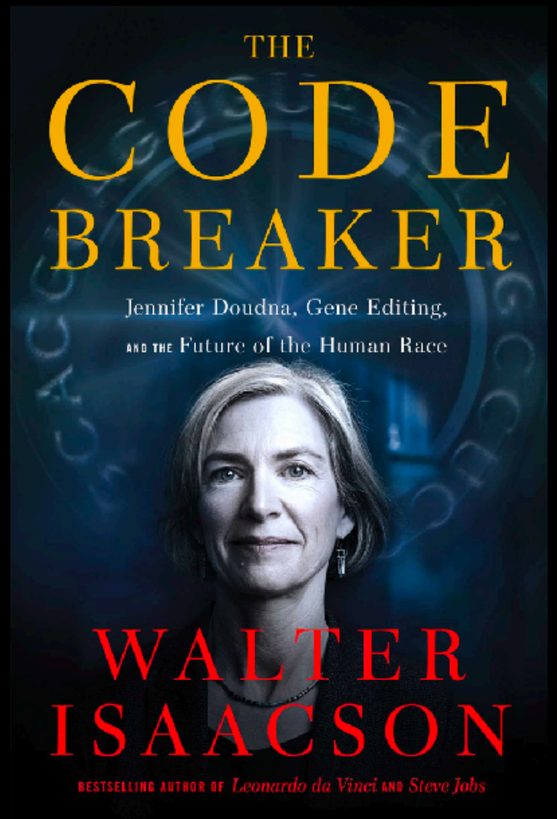
gRNAde improves **sequence recovery** and **speed** compared to Rosetta fixed-backbone RNA design

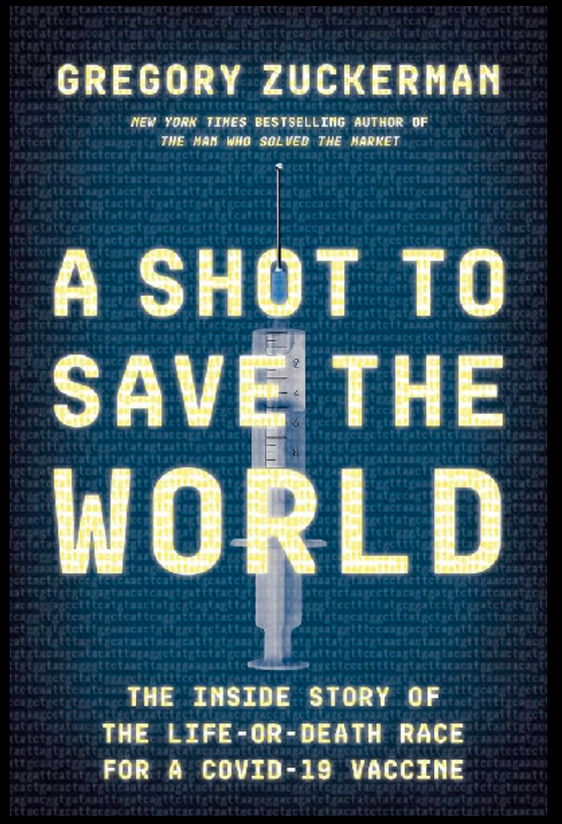


ProteinMPNN-analogue for RNA. Open-source and ready to use on GitHub. gRNAde **101** tutorial: github.com/chaitjo/geometric-rna-design/blob/main/tutorial/tutorial.ipynb

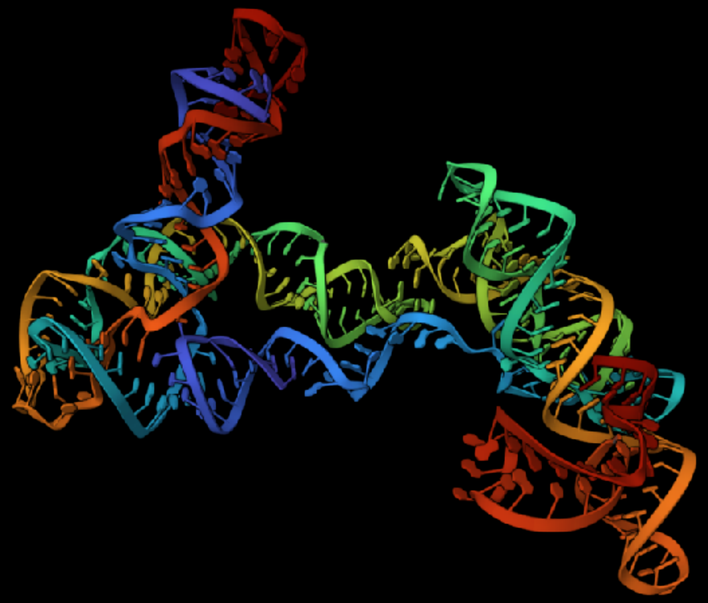
RNA at the forefront of biotechnology



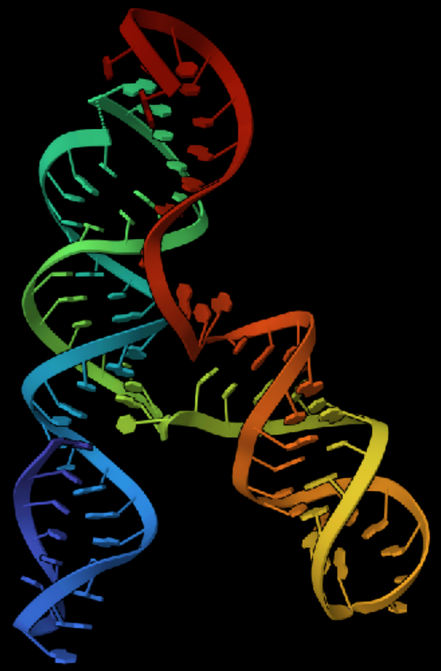




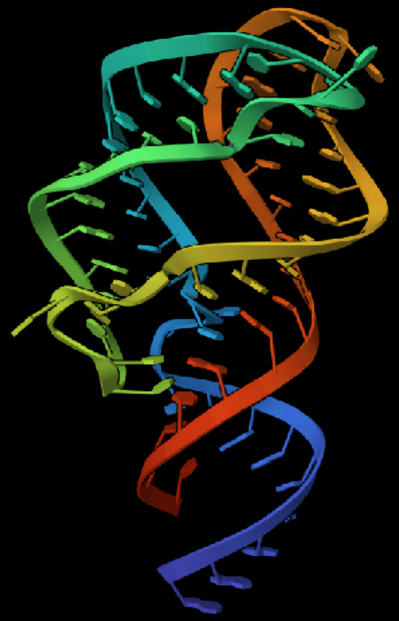
And many RNA are structured



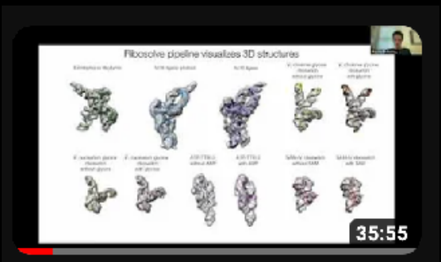
RNA polymerase
ribozyme
8T2P
McRae et al.



SARS-CoV-2
frameshift
element
6XRZ
Zhang et al.



Adenine
riboswitch
aptamer
5E54
Stagno et al.

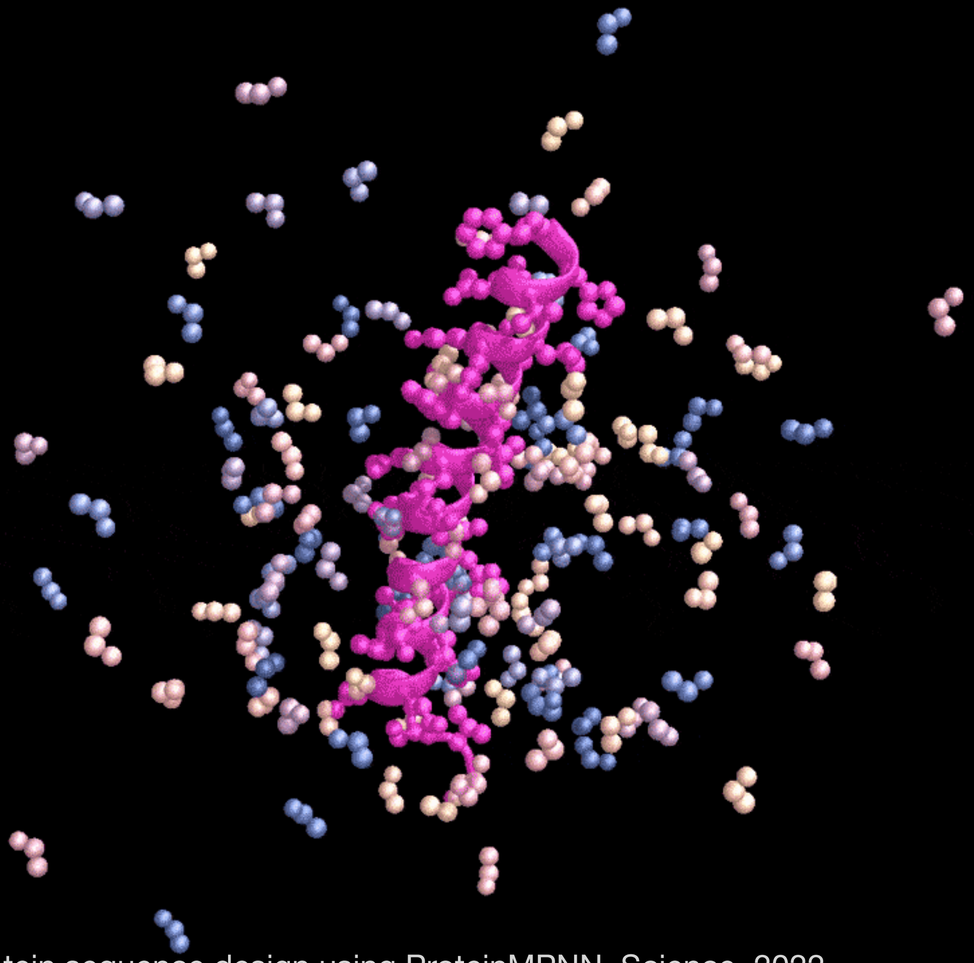


NGBS2022 Talk 10: RNA modelling and design - Rhiju Das

466 views • 4 months ago

Meanwhile

3D deep learning for protein design is starting to work



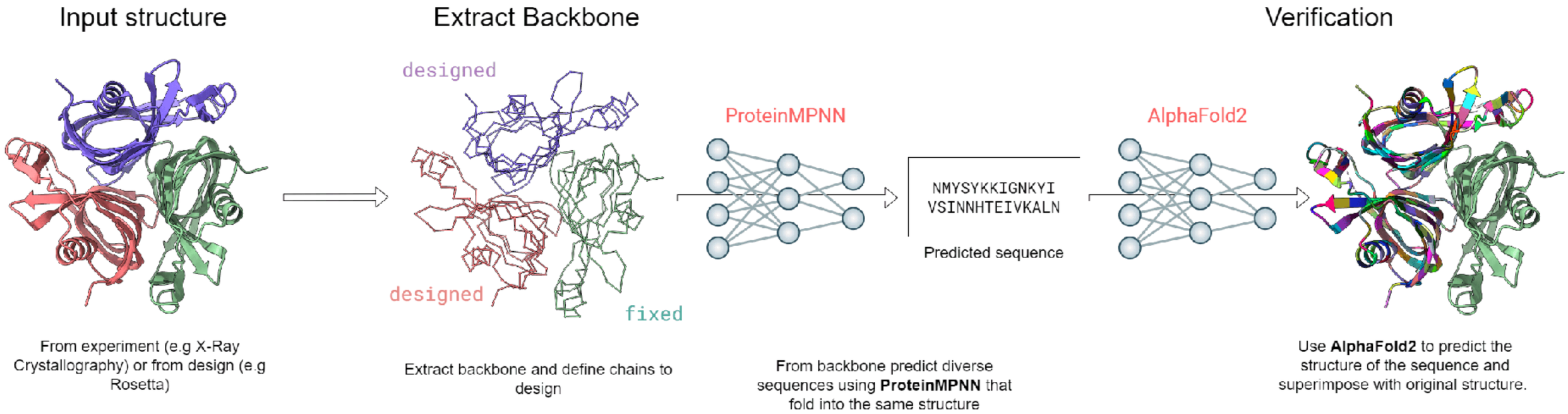
What about RNA?

Dauparas et al. Robust deep learning–based protein sequence design using ProteinMPNN. Science. 2022.
Watson, Juergens et al. De novo design of protein structure and function with RFdiffusion. Nature. 2023.

‘Generative AI’
is starting to work for protein design

Structure-based protein design workflow

Assumption: Structure → Function

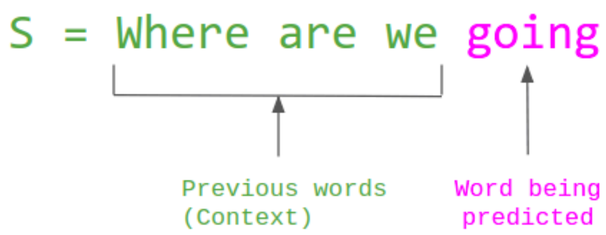


Not shown: protein Language Models (purely sequence-based)

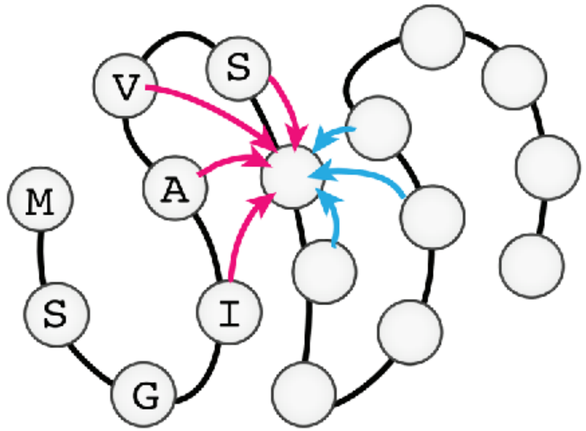
Dauparas et al. Robust deep learning–based protein sequence design using ProteinMPNN. Science. 2022.
Figure: Simon Duerr

Analogy to ChatGPT

Natural language models



$$P(S) = P(\text{Where}) \times P(\text{are} \mid \text{Where}) \times P(\text{we} \mid \text{Where are}) \times P(\text{going} \mid \text{Where are we})$$



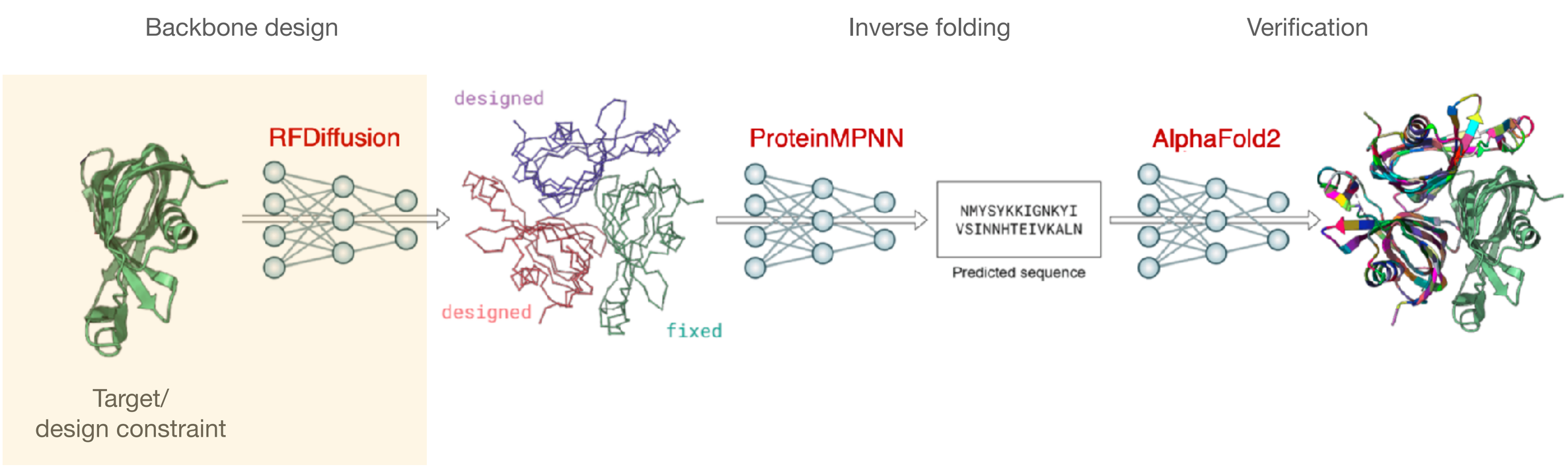
Sequence generation: **Language model**

Sequence generation conditioned on structure: **ProteinMPNN** (inverse folding)

Ingraham et al. Generative Models for Graph-Based Protein Design. NeurIPS, 2019.

De-novo protein design workflow

Starting from scratch

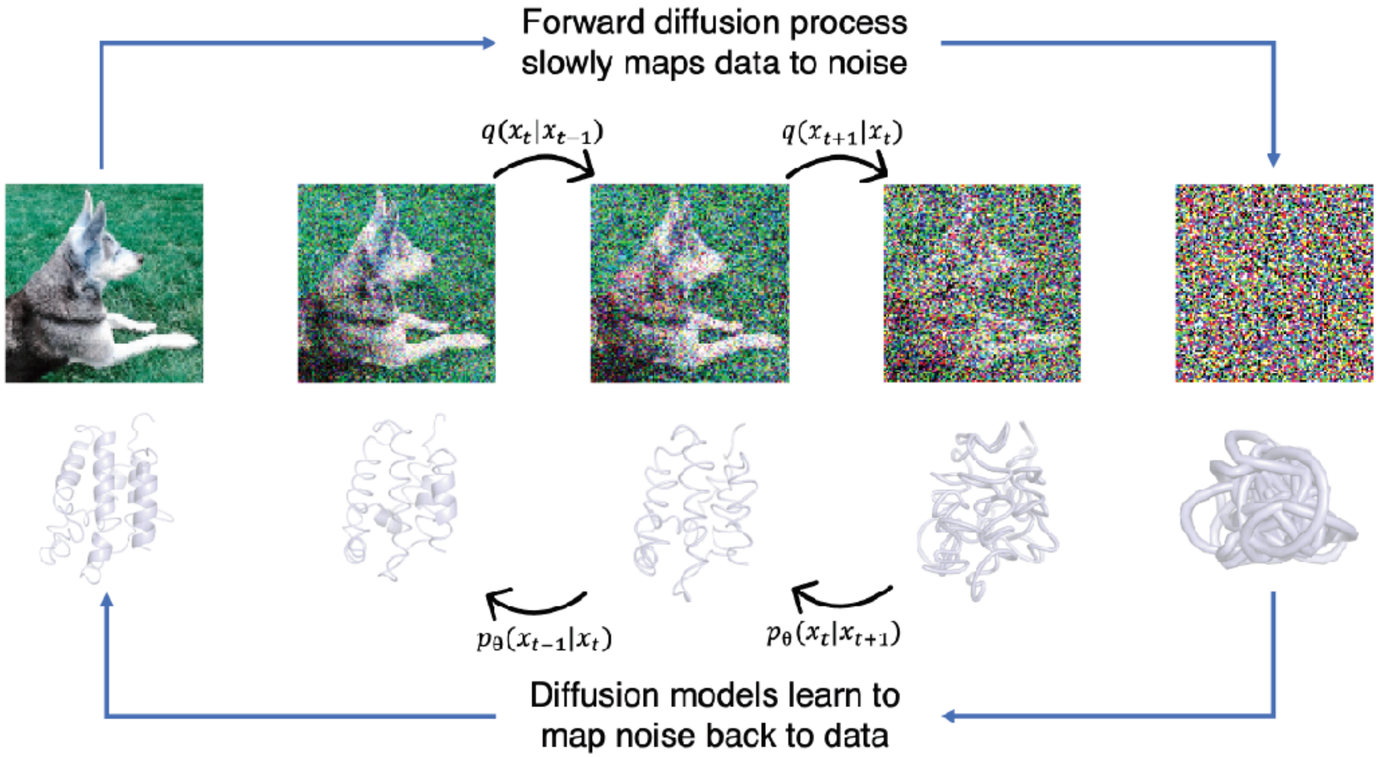


Watson, Juergens et al. De novo design of protein structure and function with RFdiffusion. Nature. 2023.

Analogy to DALL-E

DALL·E

Image generation models

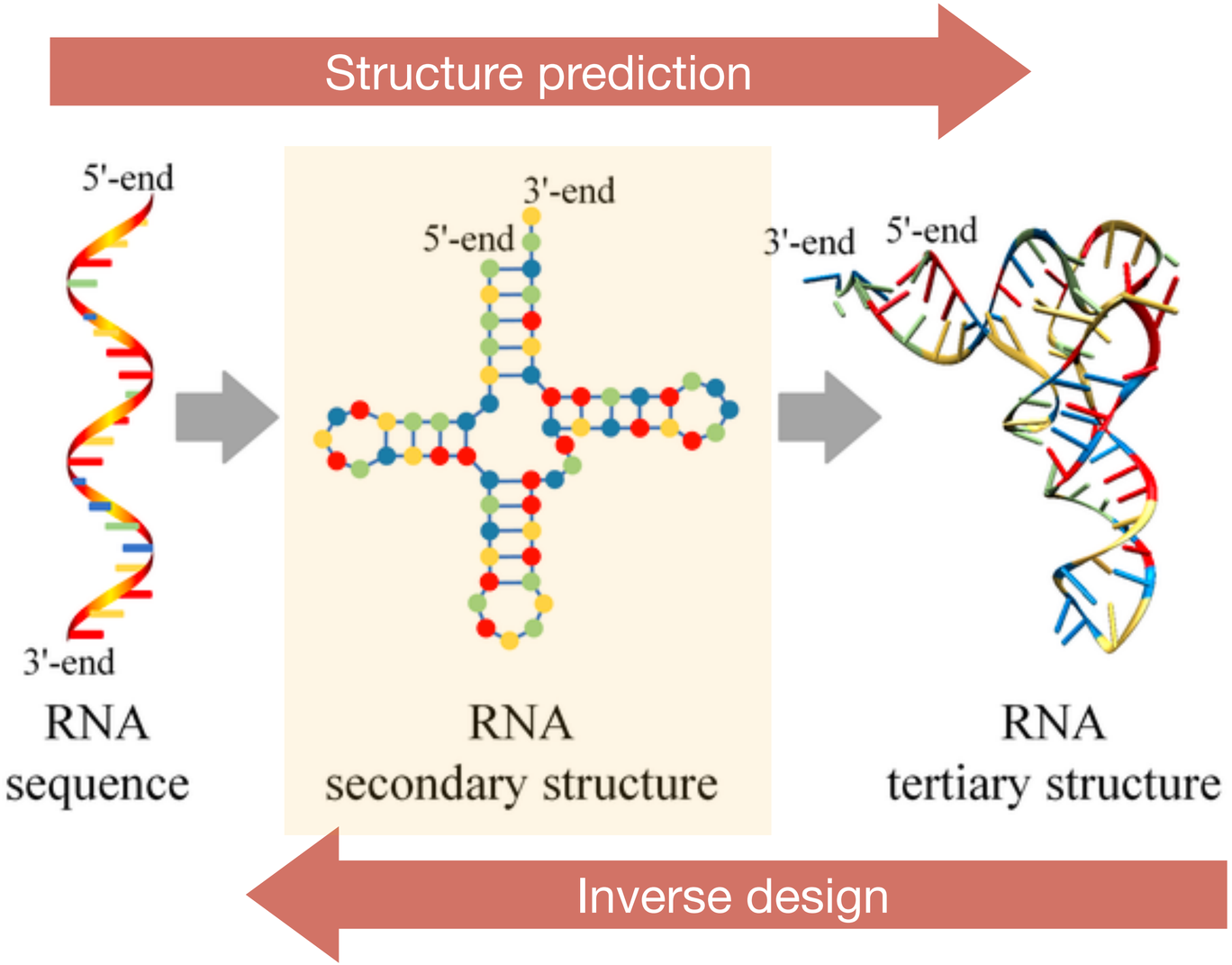


Backbone design: **RFdiffusion**

What about RNA?

RNA structure modelling and design

Emphasis on secondary structure



Relatively fewer tools for 3D design

Potential application: aptamers, riboswitches, ribozymes

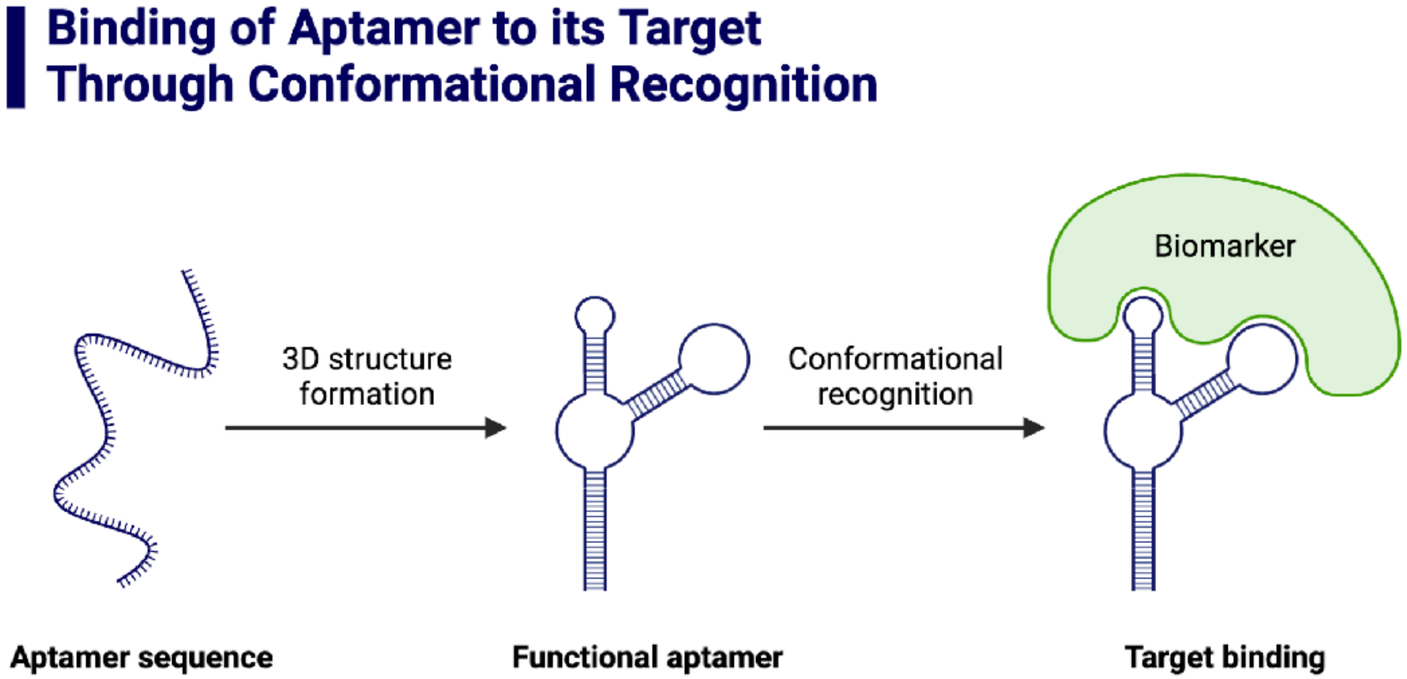
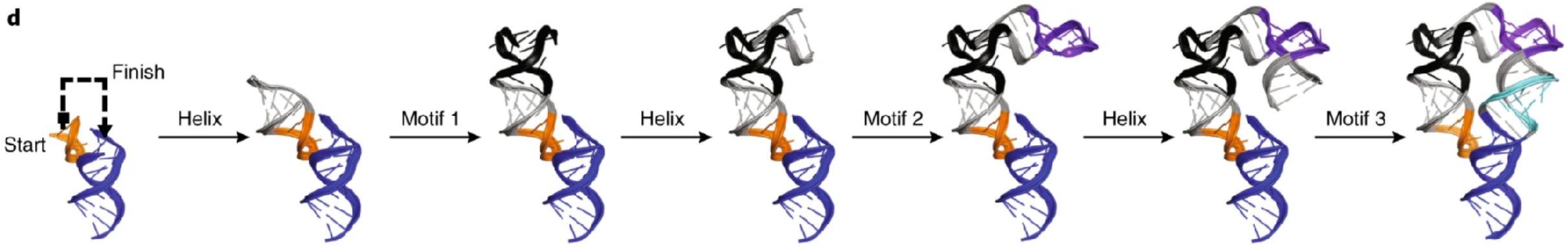
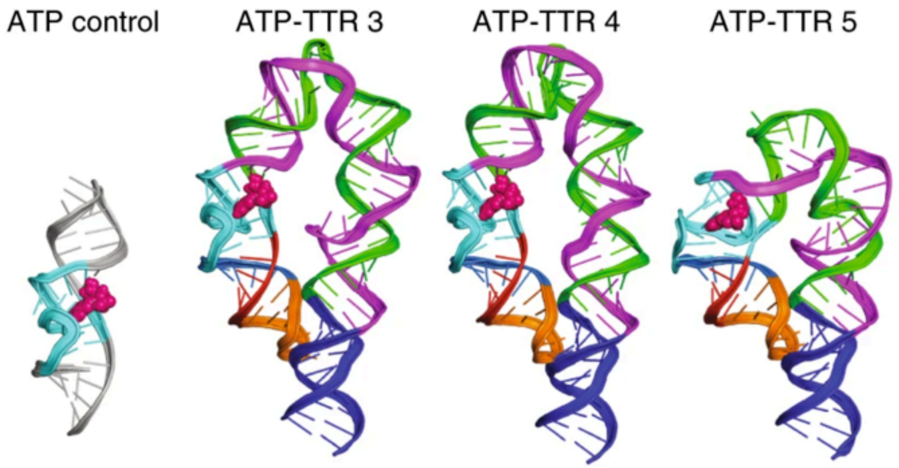


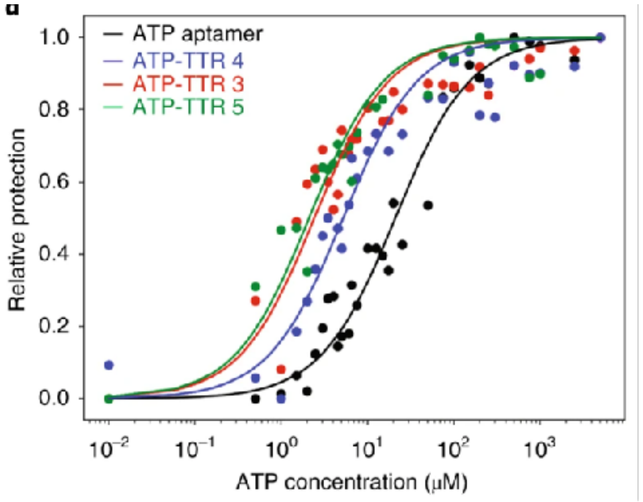
Figure: biorender

RNAMake

Uses classical algorithms for alignment between RNA motifs



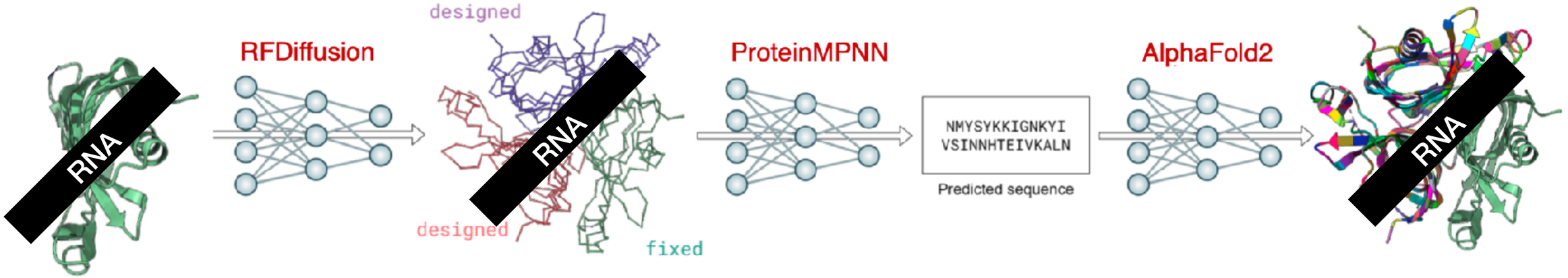




Yesselman et al. Computational Design of Three-dimensional RNA Structure and Function. Nature Nanotechnology, 2020.

Deep learning toolkit for RNA design

…work in progress



Nothing public using DL

RNAMake (non-DL)

💣 **gRNAde**

This talk!

RF-NA, RhoFold, etc.

Several teams working on this.

Not shown: RNA Language Models — Several teams working on this.

Towards deep learning:
What data exists?

Geometric Deep Learning for RNA

Main challenge: paucity of 3D structural data

“trained with only 18 known RNA structures”

> **Geometric deep learning of RNA structure.** *Science, 2021.*
>
> Raphael JL Townshend, Stephan Eismann, Andrew M Watkins, Ramya Rangan, Maria Karelina, Rhiju Das, and Ron O Dror.

“trained on 2,986 RNA chains, non-redundant to 122 test RNAs”

> **Integrating end-to-end learning with deep geometrical potentials for ab initio RNA structure prediction.** *Nature Communications, 2023.*
>
> Yang Li, Chengxin Zhang, Chenjie Feng, Robin Pearce, Peter L. Freddolino, Yang Zhang.

All RNA structures in the PDB

RNAsolo: cleaned, PDB-derived RNA 3D structures

	Solo RNAs	RNAs from protein–RNA complexes	RNAs from DNA–RNA hybrids	All RNAs
X-ray	1454	6439	91	7984
NMR	573	146	28	747
Electron microscopy	73	4104	0	4177
Multi-method	1	5	0	6
Total	2101	10694	119	12914
Total (today)	2387	13218	136	15741 (13870 ≤3.5Å)

RNAPOLIS Adamczyk et al. RNAsolo: a repository of clean, experimentally determined RNA 3D structures. Bioinformatics. 2022.

All RNA structures in the PDB

RNAsolo: cleaned, PDB-derived RNA 3D structures

	Solo RNAs	RNAs from protein–RNA complexes	RNAs from DNA–RNA hybrids	All RNAs
Total (today)	2387	13218	136	15741

3825 equivalence classes

vs.

ProteinMPNN, RFdiffusion: entire PDB
208,659 proteins ≤3.5Å → **25,361 clusters** at 30% seq.id.

One order of magnitude more proteins!

Should we just wait?

Not necessarily…

Other successful (in-silico) tools were trained on carefully chosen subsets:

- Chroma: 28819 structures ≤2.6Å
- Genie: 8766 domains
- FrameFlow: 3938 domains

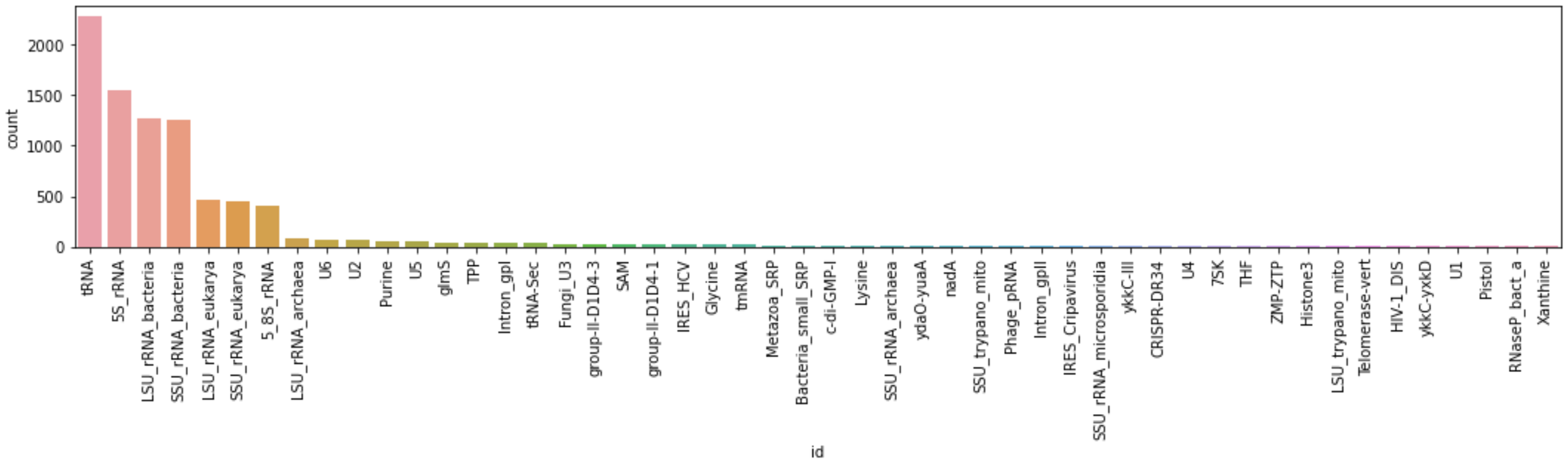
"...achieve similar in-silico performance to RFdiffusion with a quarter of the parameters – an important consideration…models are often run tens of thousands of times…"

– *Winnifrith et al. 2023.*

Winnifrith, Outeiral, Hie. Generative artificial intelligence for de novo protein design. 2023.

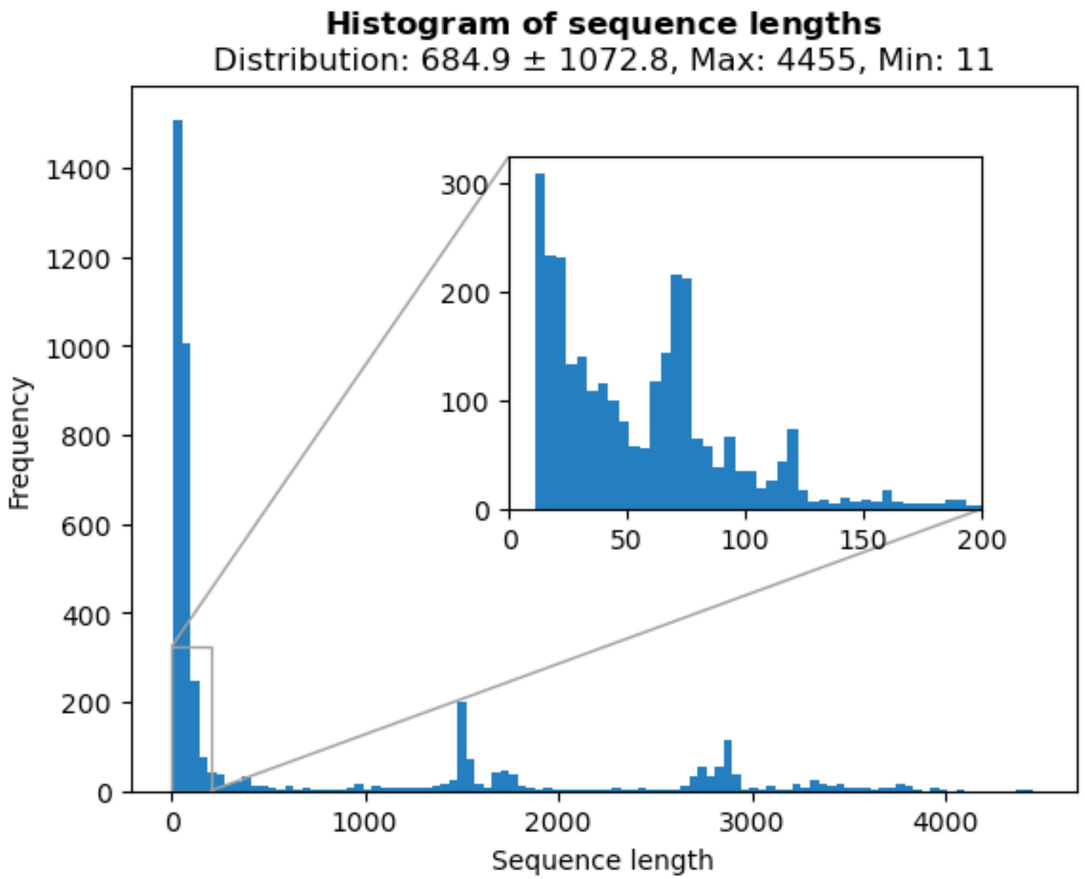
RFam families in the PDB

Majority from protein-RNA complexes, tRNAs, ribosomal RNAs



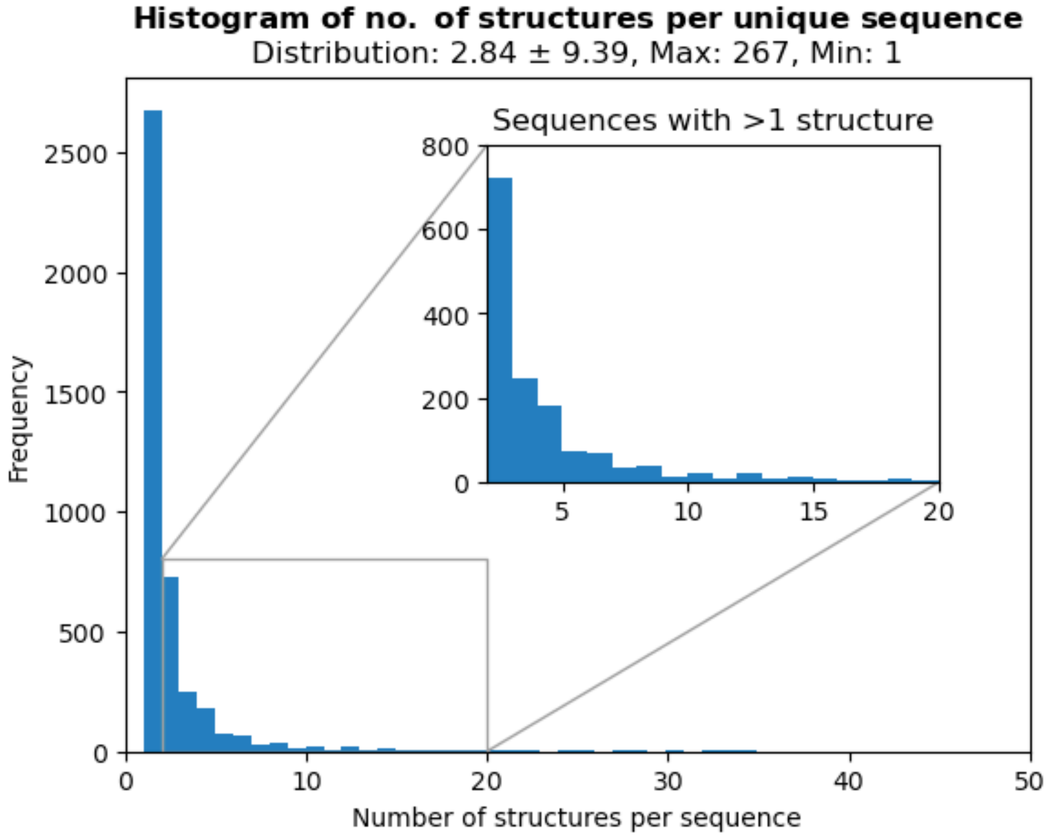
Distribution of sequence lengths

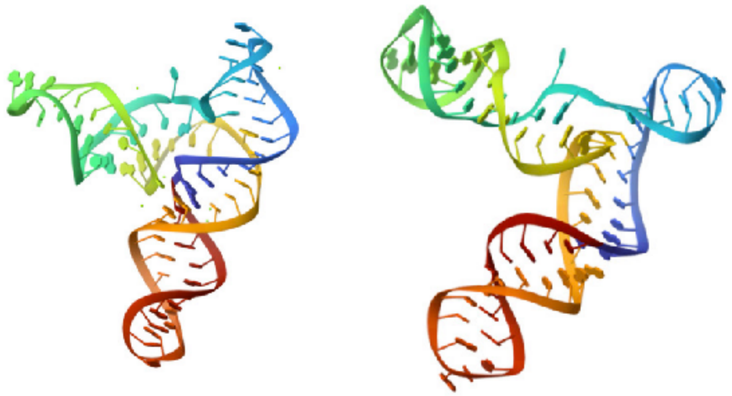
Mostly shorter than 500 nucleotides



RNA adopt multiple conformations

Critical for functionality & perhaps interesting for design

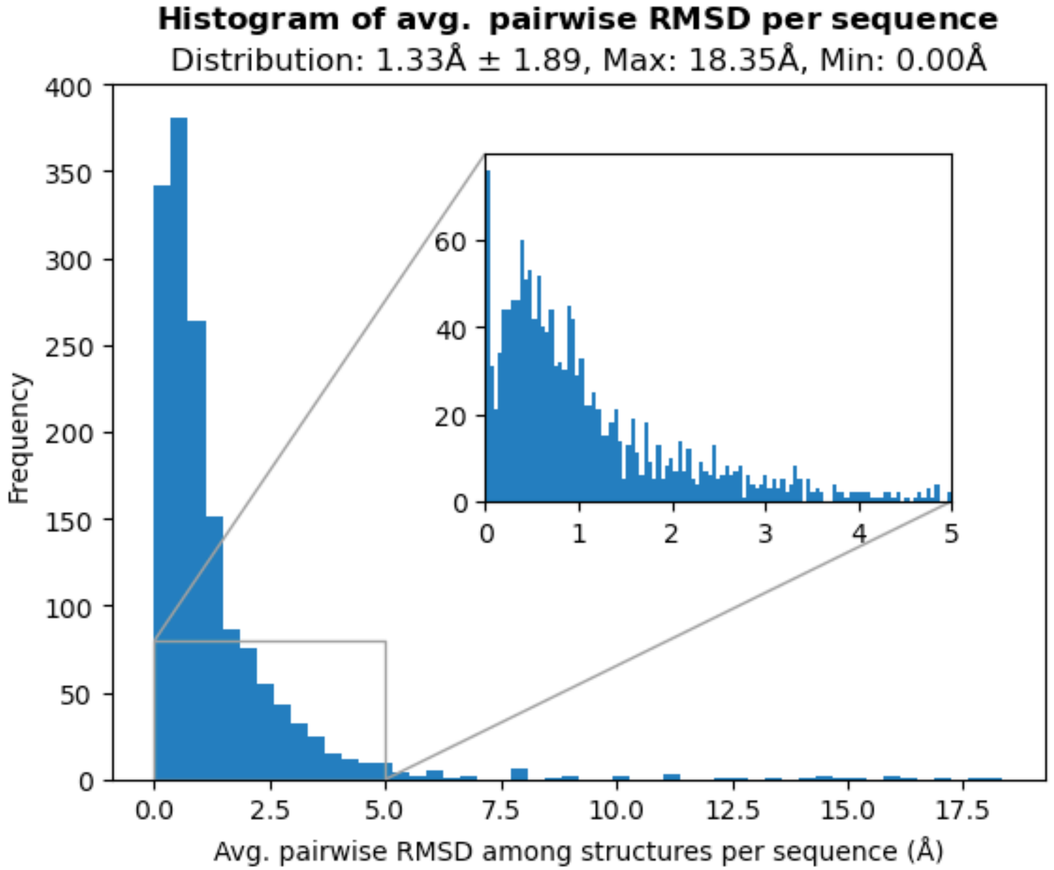


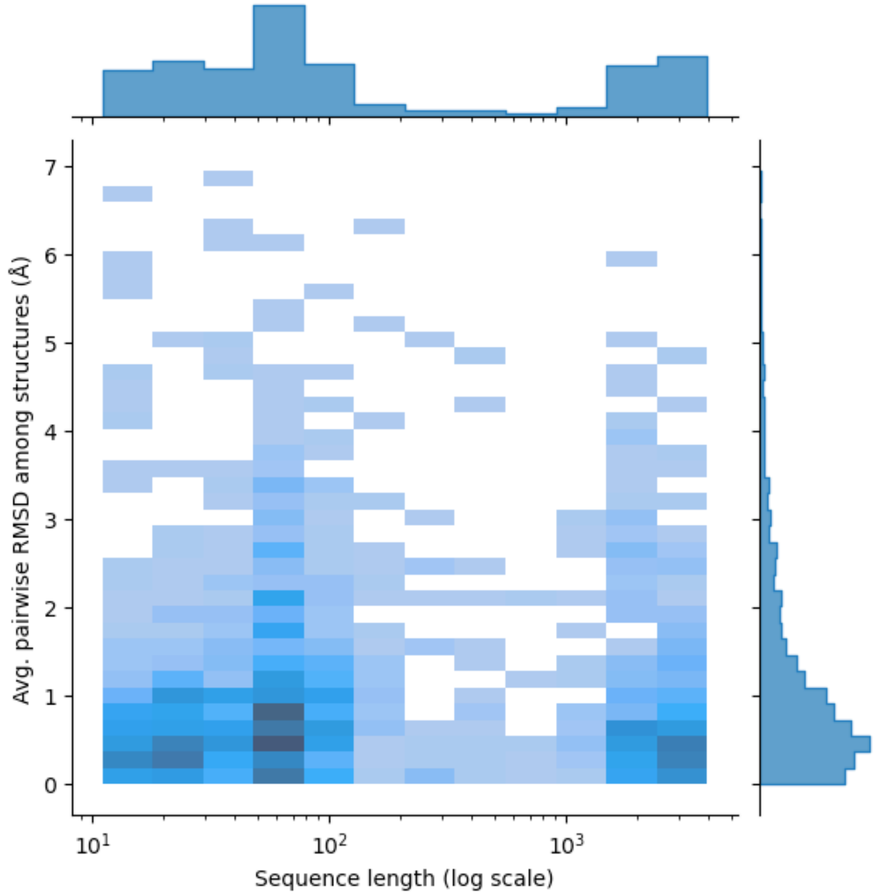


L1 ligase ribozyme
(PDB 2OIU)

RNA adopt multiple conformations

Same sequence can have very different structures

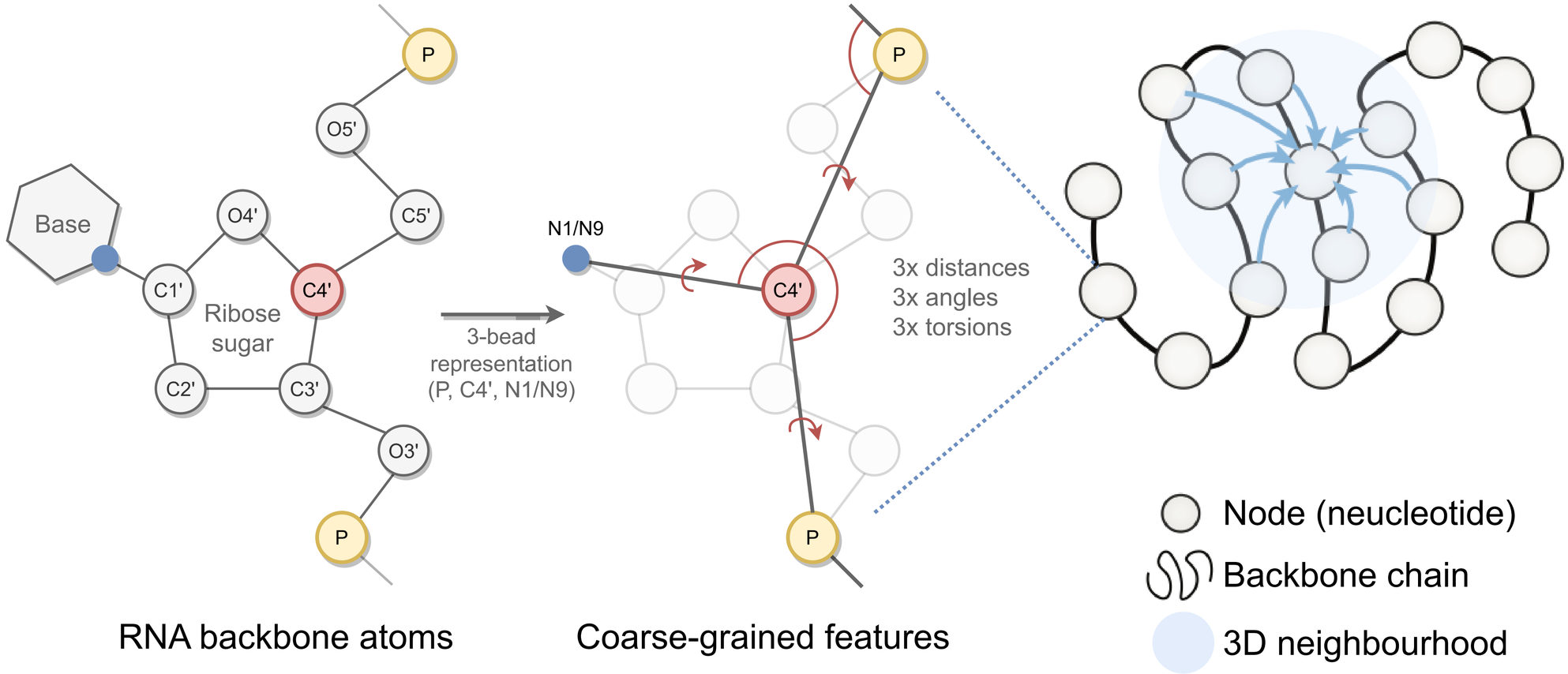




The gRNAde pipeline for RNA inverse folding

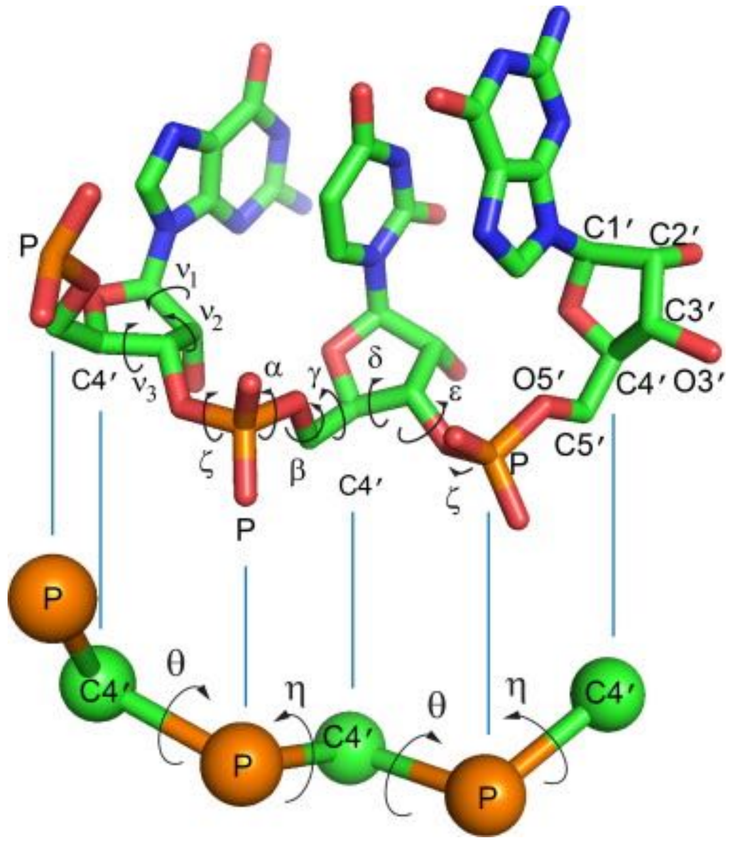
Fixed backbone re-design

Input: PDB file(s) → geometric graph in 3D



Why the 3-bead representation?

P, C4', N1 (pyrimidine) or N9 (purine)

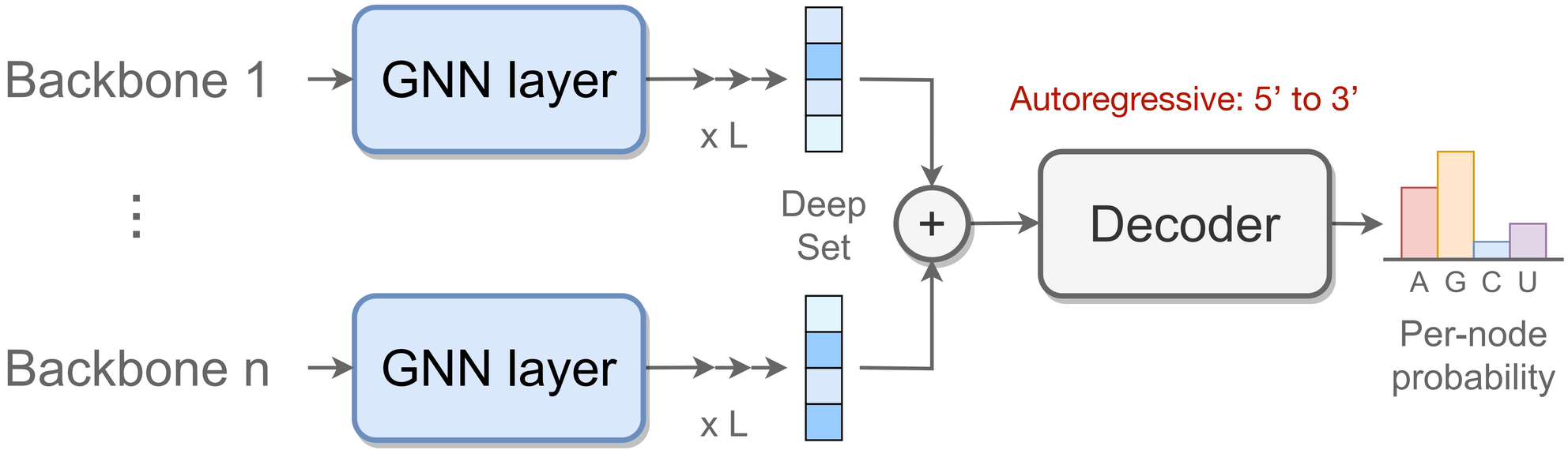


"The pseudotorsional descriptors η and θ, together with sugar pucker, ***are sufficient to describe the RNA backbone conformation fully*** *in most cases."*

Coarse-grained modeling of RNA 3D structure. Dawson et al. Methods. 2016.

gRNAde model architecture

One or more featurized graphs → per-node probability over 4 bases

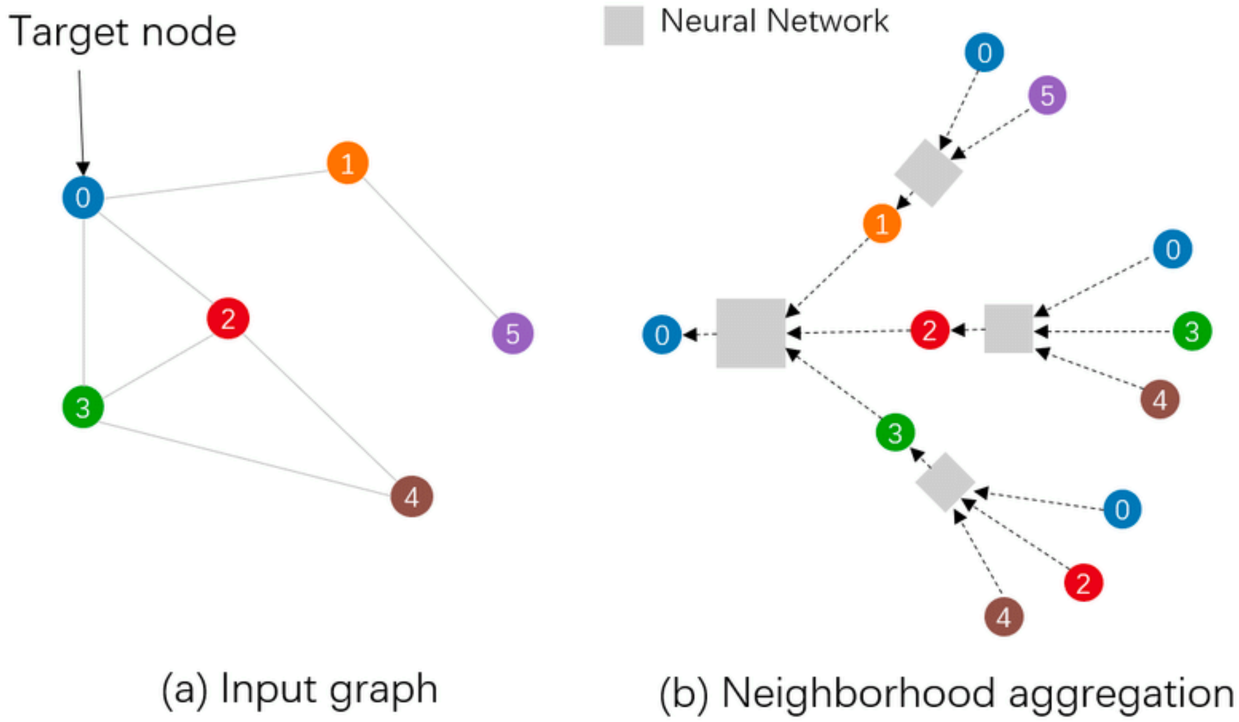


Can be sampled from to design new sequences

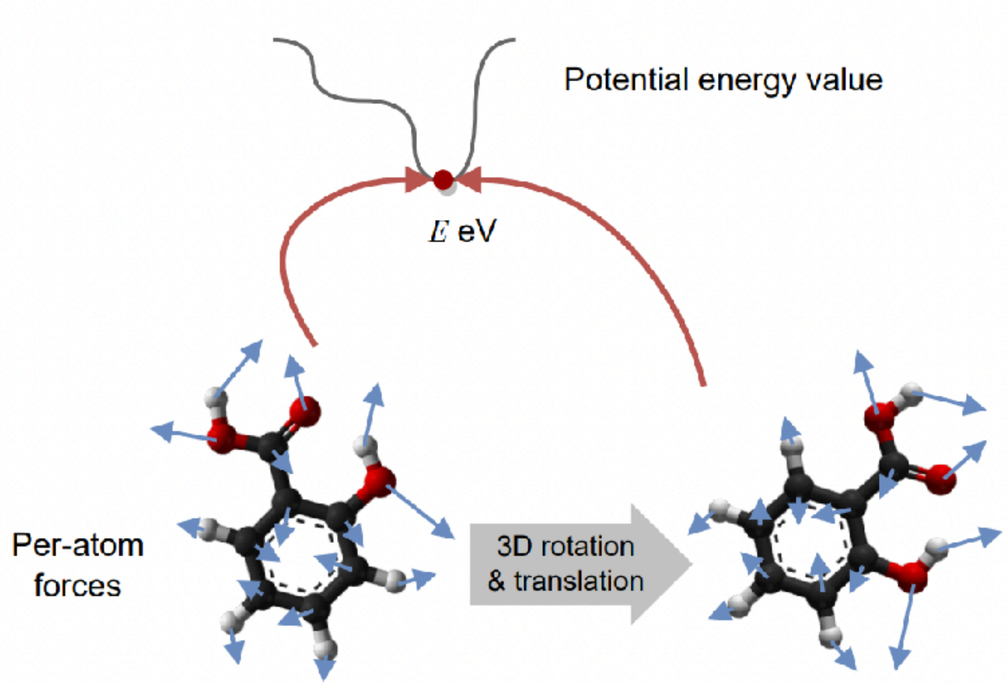
Jing et al. Learning from Protein Structure with Geometric Vector Perceptrons. ICLR 2020.
Zaheer et al. Deep sets. NeurIPS 2017.

Graph Neural Networks for 3D structure

Learn to propagate information along the graph



Account for 3D symmetries

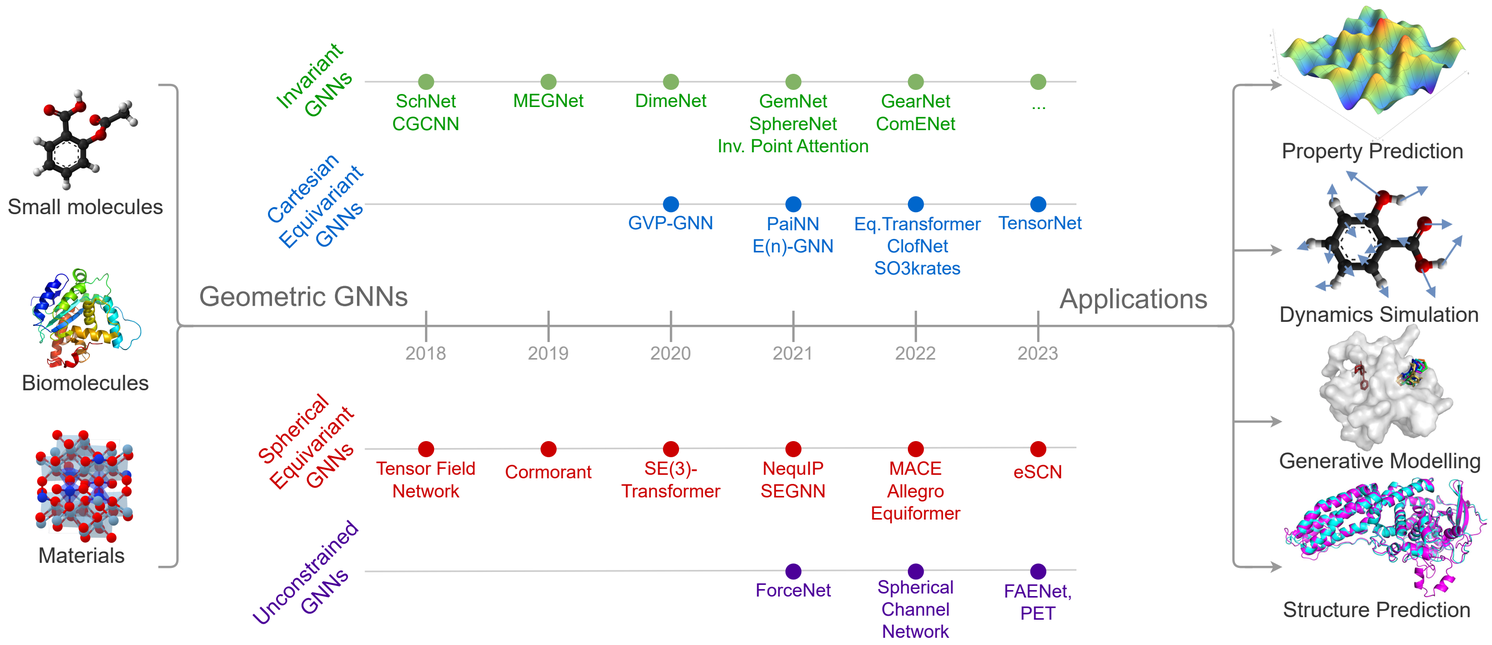


Where to start with GNNs for biomolecules?

A Hitchhiker's Guide to Geometric GNNs for 3D Atomic Systems

Alexandre Duval[*,1,2] **Simon V. Mathis**[*,3] **Chaitanya K. Joshi**[*,3] **Victor Schmidt**[*,1,4]
Santiago Miret[5] **Fragkiskos D. Malliaros**[2] **Taco Cohen**[6]
Pietro Liò[3] **Yoshua Bengio**[1,4] **Michael Bronstein**[7]

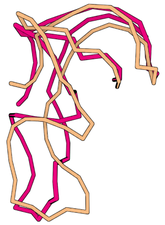
[1]Mila [2]Université Paris-Saclay [3]University of Cambridge [4]Université de Montréal
[5]Intel Labs [6]Qualcomm AI Research [7]University of Oxford
[*]Equal first authors.



What is a good designs?

In-silico evaluation metrics to prioritise designs

Not shown:
Perplexity
model's guess of
P(seq|struct)

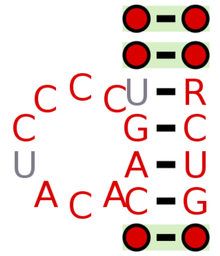


Backbone graph(s)

gRNAde

GAGCGU...

Sampled sequences



Predicted secondary structures

EternaFold

True sequence

Sequence recovery

s_1
⋮
s_{16}

True secondary structure

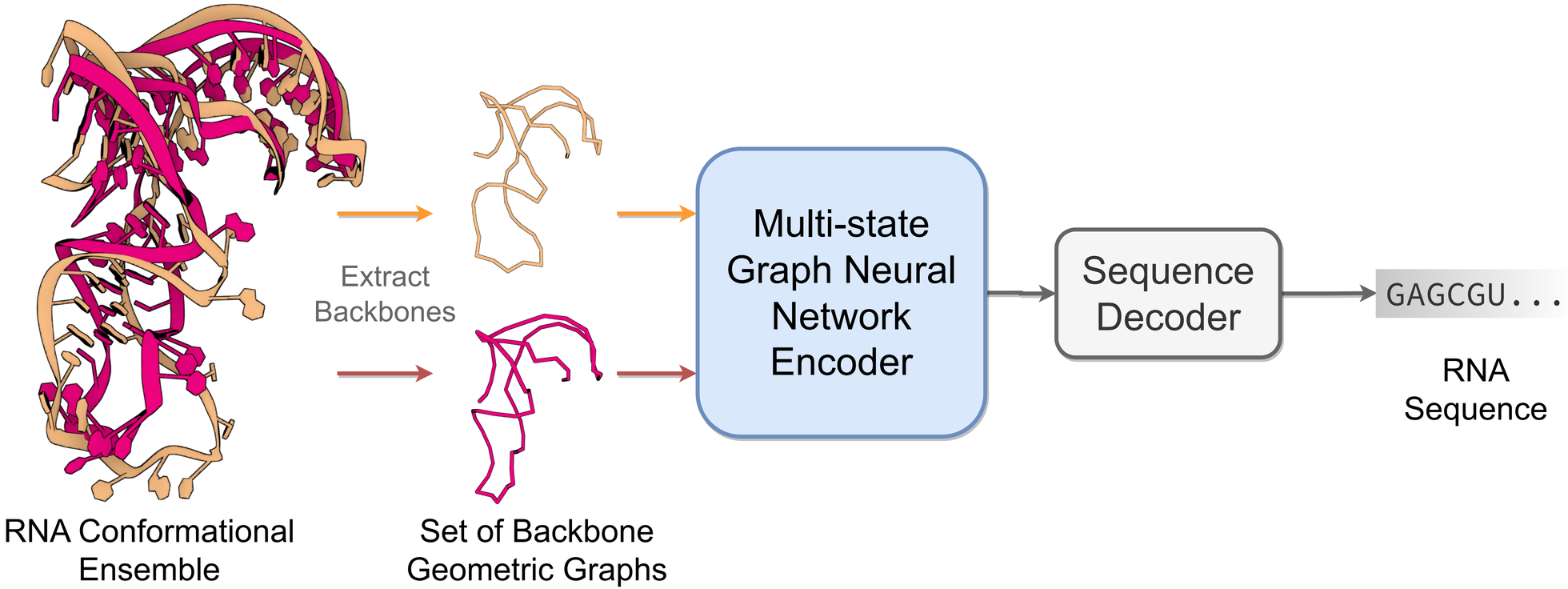
Secondary structure self-consistency score

ss_1
⋮
ss_{16}

What can we do with gRNAde?

Fixed backbone re-design

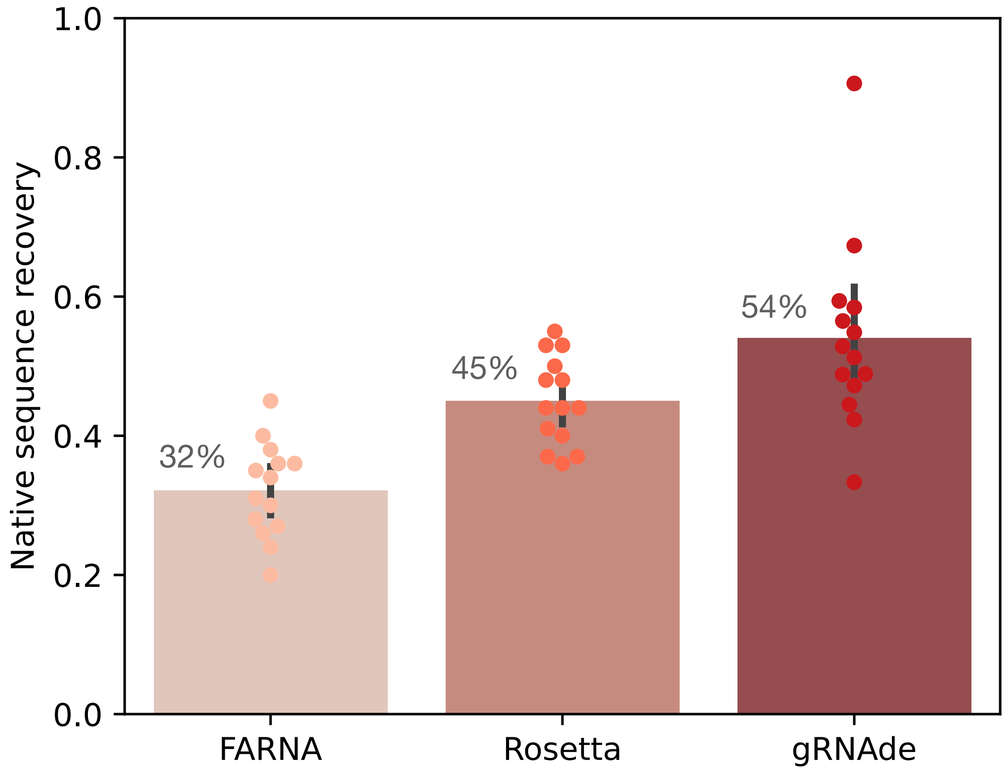
Input: native PDB file → Output: designed sequences



Benchmarking single-state design

Re-design 14 RNAs of interest from the PDB by Das et al.

Improved sequence recovery



Faster inference speed

- gRNAde: under 1 second for 100s of nts.
- Rosetta: order of hours…

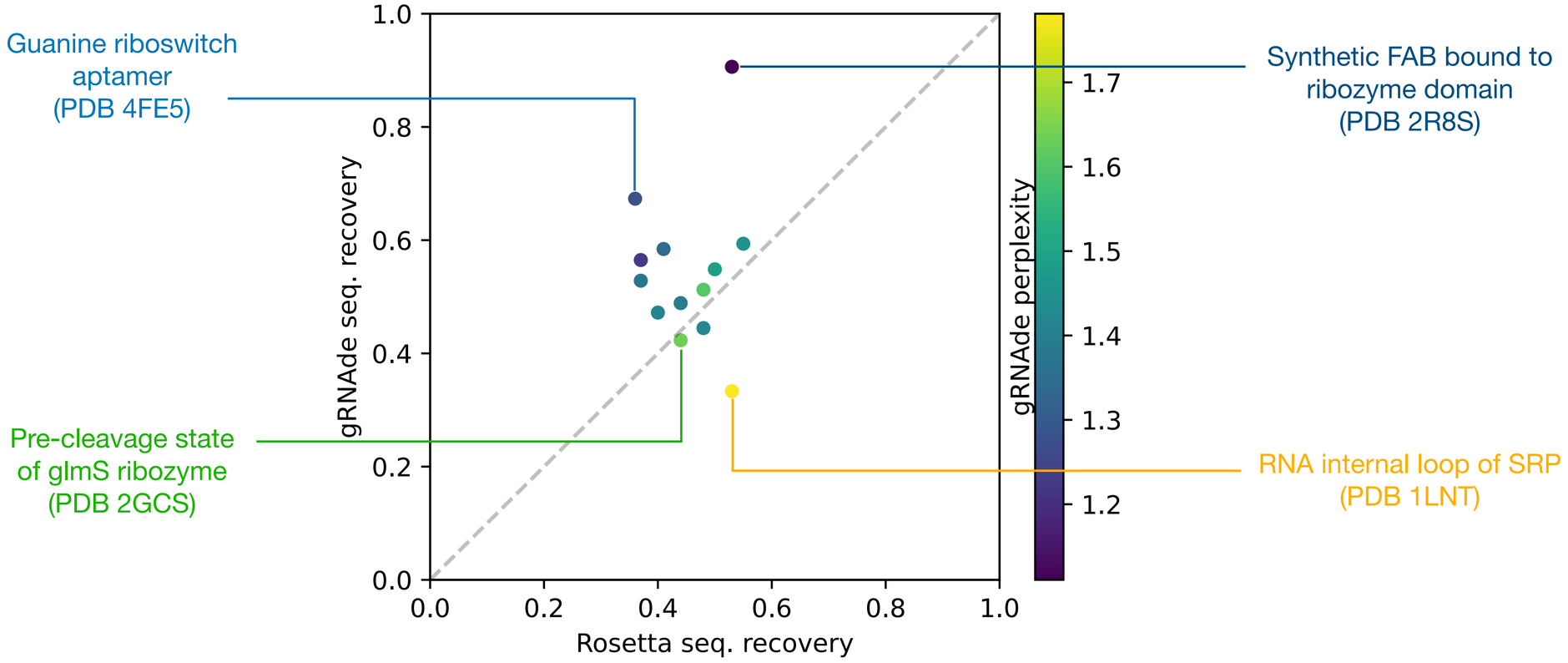
Rosetta documentation:
“runs on RNA backbones longer than ~ten nucleotides take many minutes or hours”

Tried to evaluate for generalisation:
Excluded all 14 RNAs and structurally identical RNAs (TM-score threshold 0.45) from training data.

Das, Karanicolas, Baker. Atomic accuracy in predicting and designing noncanonical RNA structure. Nature Methods, 2010.

Perplexity correlates well with recovery

Indicator of model's confidence in its own prediction

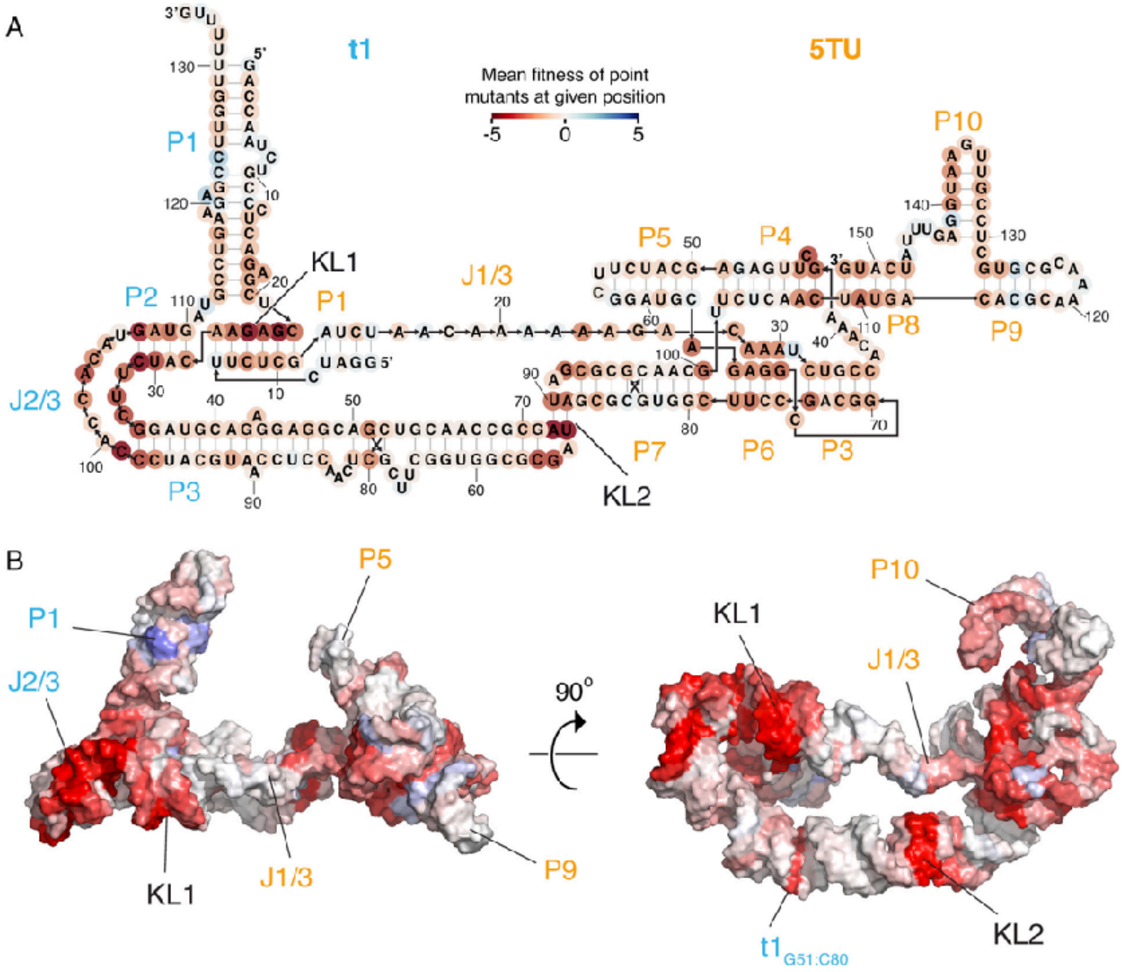


Could perplexity be correlated with fitness/function, too?

Can gRNAde understand RNA fitness landscapes? A retrospective analysis on an RNA Polymerase Ribozyme (McRae et al., PNAS 2024)

Structure + Functional landscape

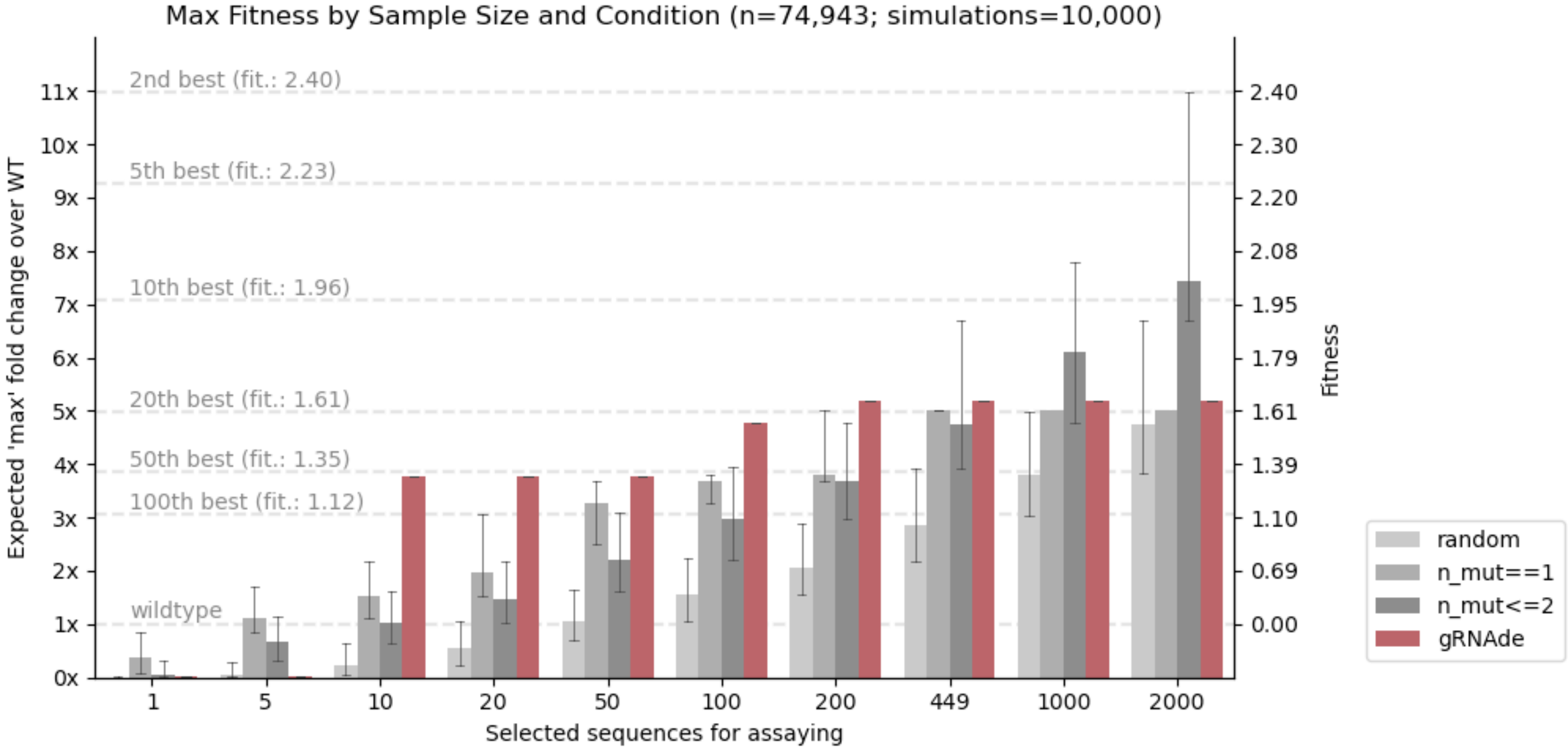
Allows retrospectively analysis of gRNAde for RNA engineering



- **Cryo-EM structure** at 5A resolution (not in gRNAde's training set).
- **70,000+ data points** of (mutant sequence, fitness).
- **gRNAde's perplexity**: likelihood of sequence folding into given backbone; can be used as an unsupervised ranker of mutants for a given structure.
- Latent features can be used for finetuning (supervised learning), too.

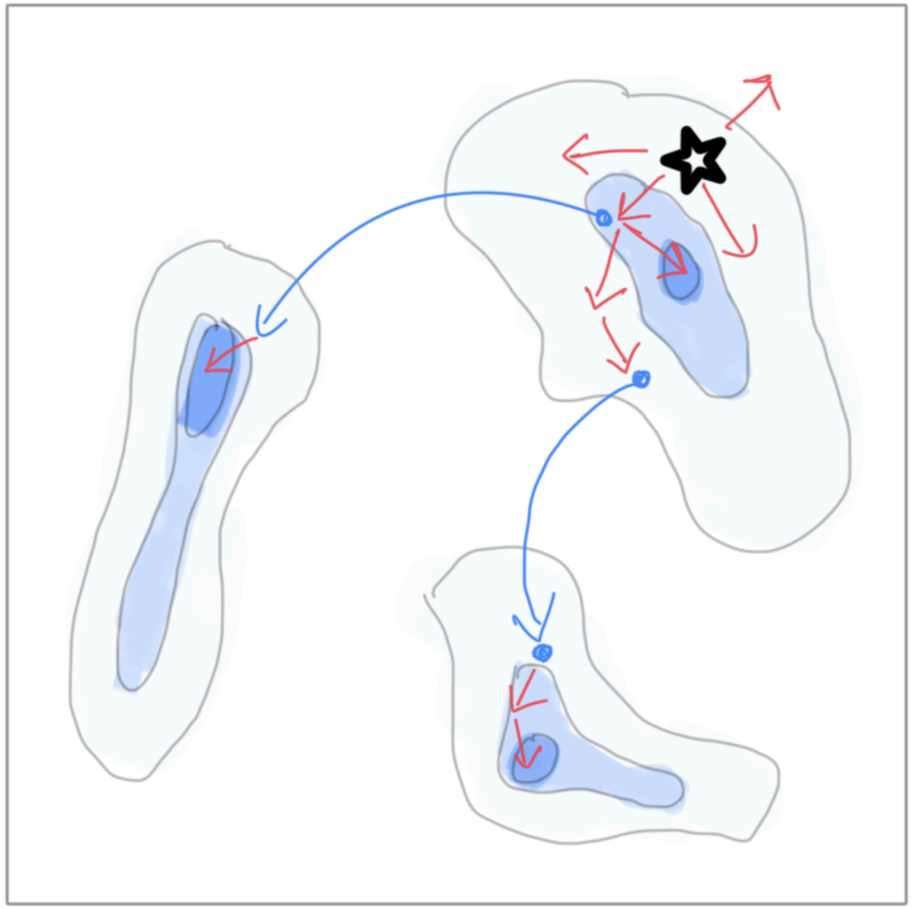
McRae et al. Cryo-EM structure and functional landscape of an RNA polymerase ribozyme. PNAS, 2024.

Unsupervised learning of Ribozyme fitness



ML-augmented RNA engineering

Evolution: local exploration, gRNAde: global jumps in sequence space



★ WILDTYPE

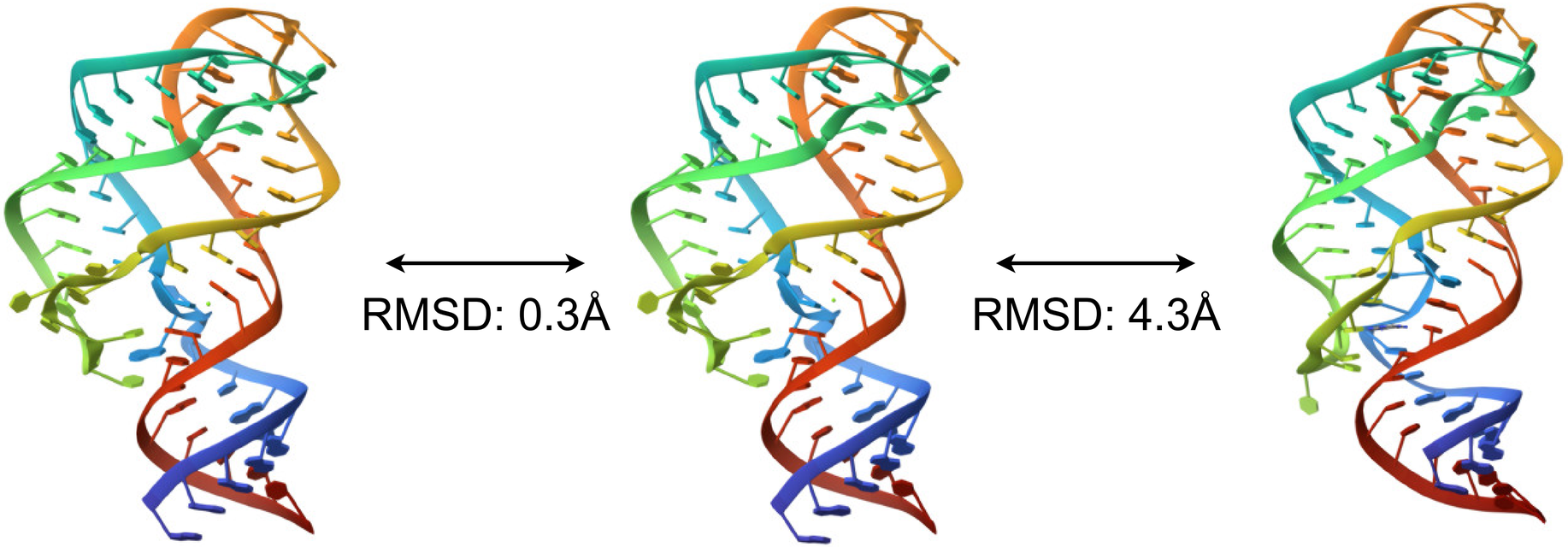
→ EVOLUTION

→ ML MODEL
(eg. gRNAde)

Multi-state RNA design

Explicitly designing conformational ensembles

Single-state design can be ambiguous



5E54: Apo

5SWD: Intermediate

5SWE: Holo

Stagno et al. Structures of riboswitch RNA reaction states by mix-and-inject XFEL serial crystallography. Nature, 2017.
Hoetzel, Suess. Structural changes in aptamers are essential for synthetic riboswitch engineering. Journal of Molecular Biology, 2022.
Ken et al. RNA conformational propensities determine cellular activity. Nature, 2023.

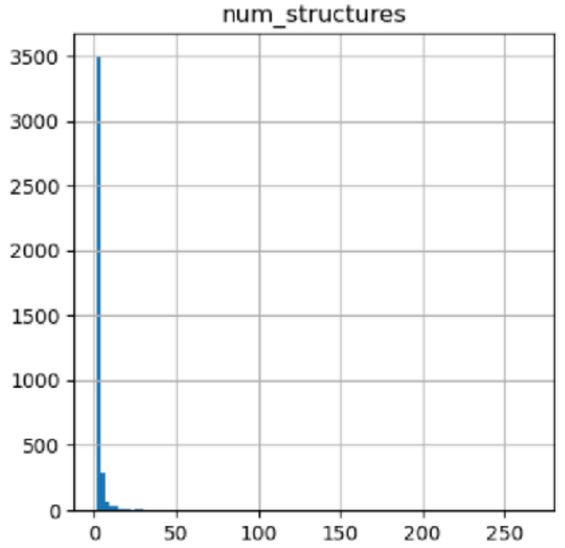
Benchmarking multi-state design

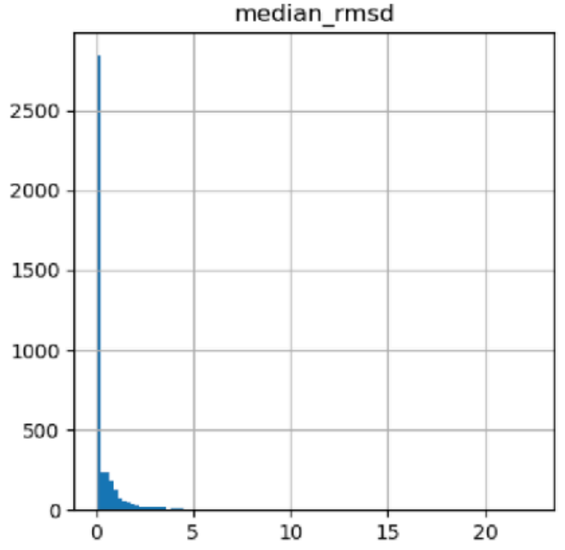
Creating a challenging set of structurally flexible RNAs

1. **Cluster RNAsolo** based on **structural similarity** — US-align with TM-score threshold 0.45.
2. Order clusters based on **median intra-sequence RMSD** among available structures in the cluster.
3. Training, validation, and test splits become progressively more flexible.
   - **Top 100 samples** from clusters with highest intra-seq. RMSD — test set.
   - **Next 100 samples** from clusters with highest intra-seq. RMSD — validation set.
   - Very large (> 1000 nts) RNAs — training set.
4. If any samples were not assigned clusters, append them to the training set.

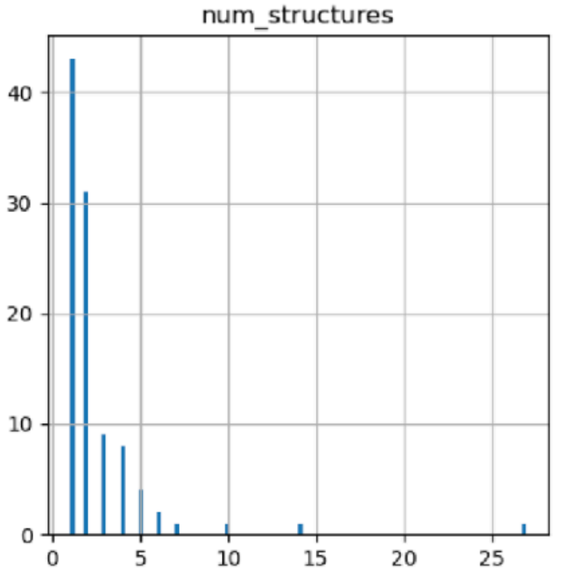
Test/validation set: **100 RNAs each**, training set: **~4000 RNAs**.

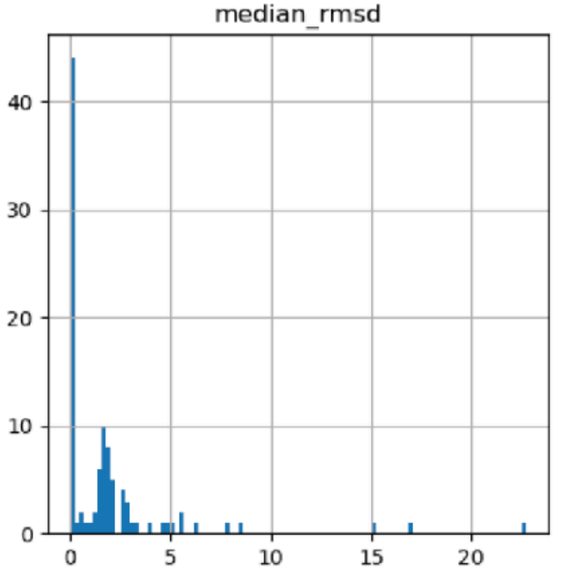
Split: train
Average median RMSD: 0.38 +- 0.99
Median number of structures: 1.00



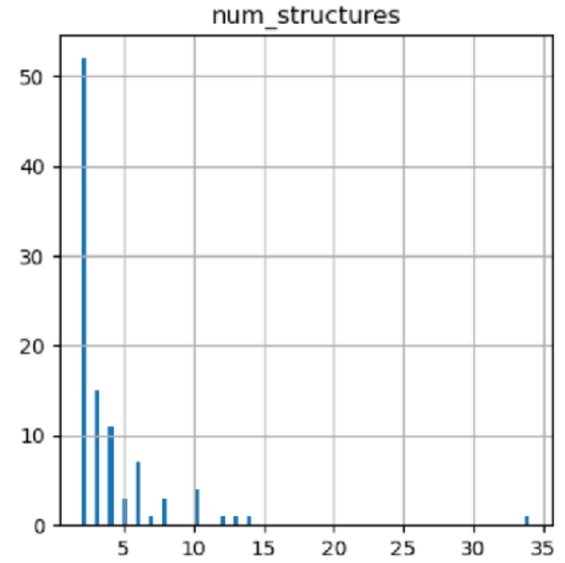


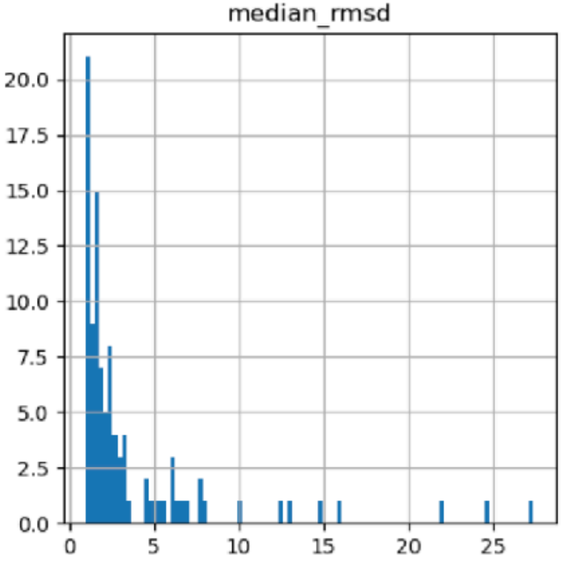
Split: val
Average median RMSD: 1.89 +- 3.42
Median number of structures: 2.00





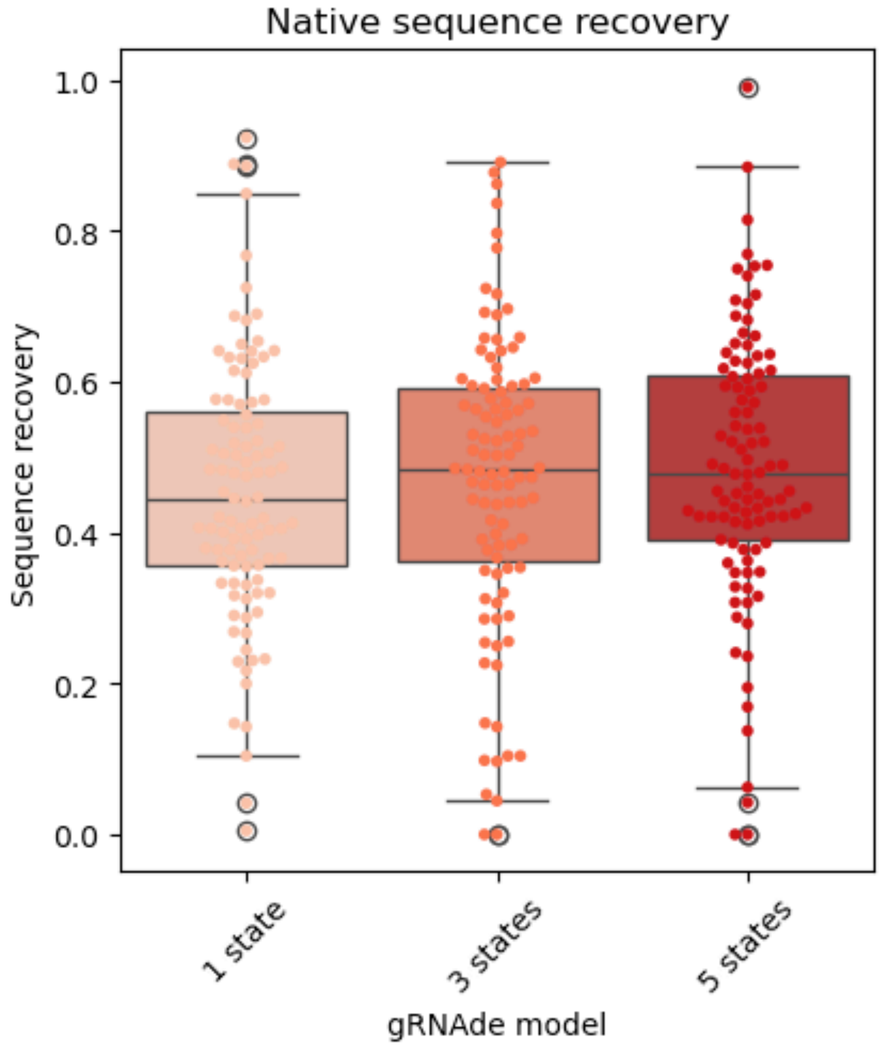
Split: test
Average median RMSD: 3.72 +- 4.74
Median number of structures: 2.00

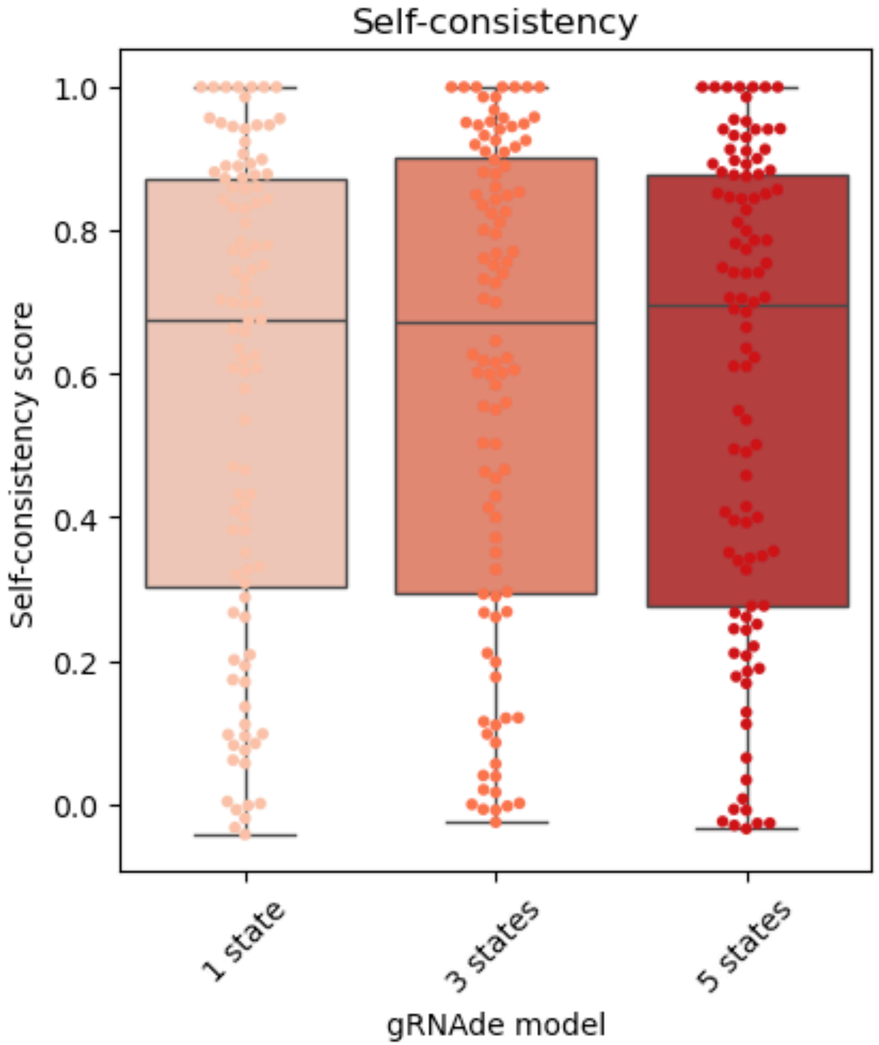




Multi-state models slightly improve recovery

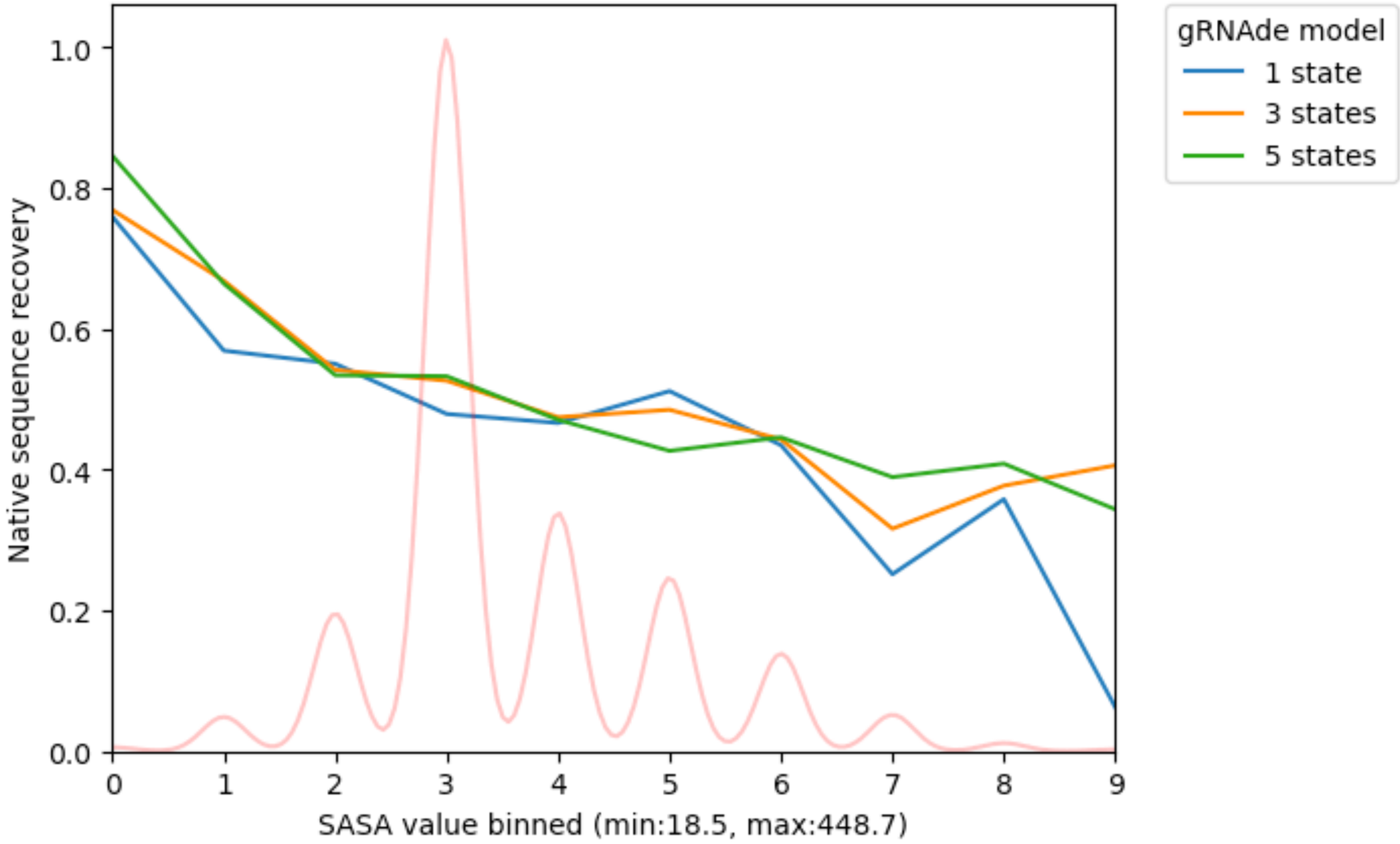
Room for improvement in designing models and evaluation





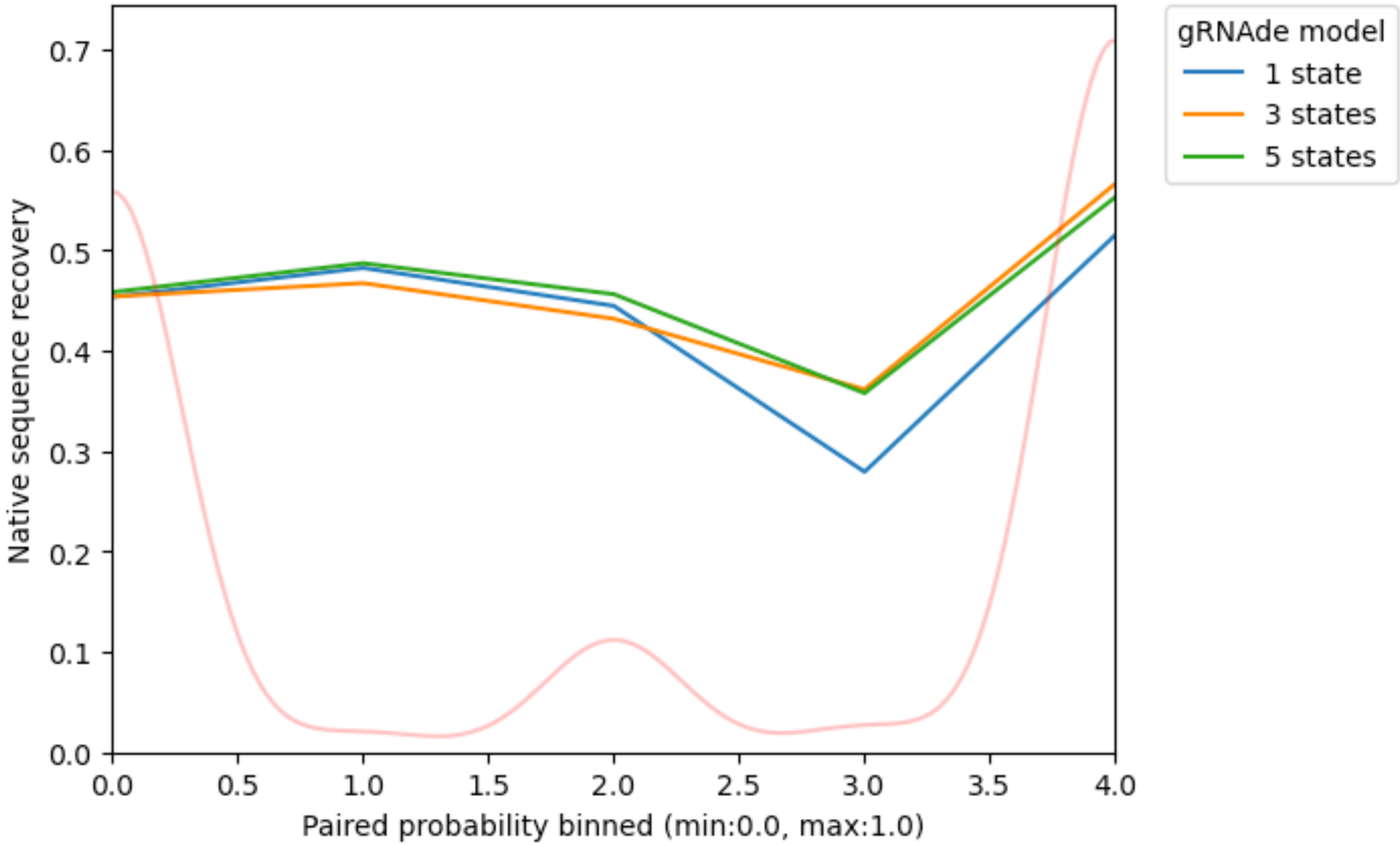
Surface vs. core nucleotides

Multi-state models show improved recovery on surface



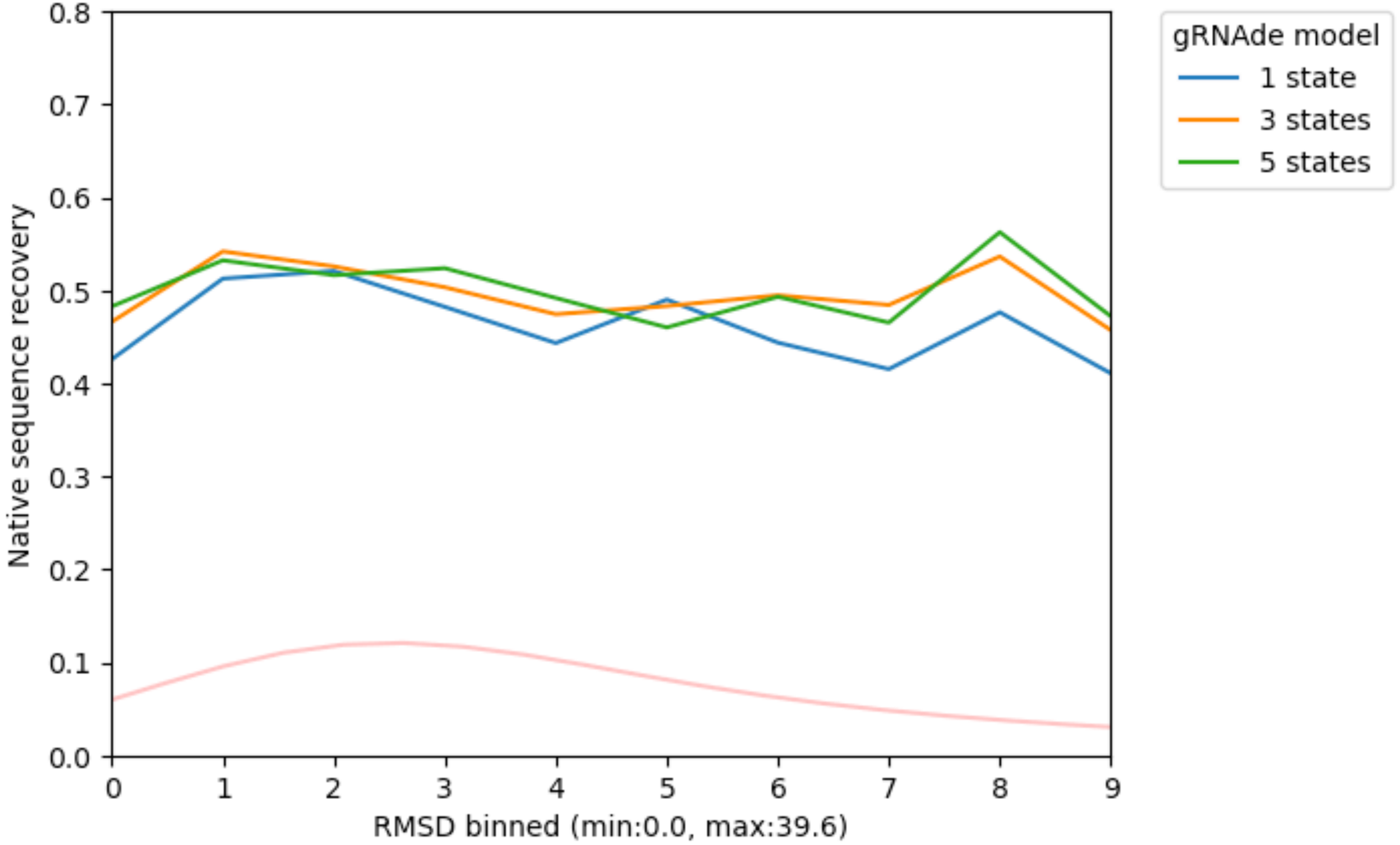
Paired vs. unpaired nucleotides

Multi-state models recover paired positions better



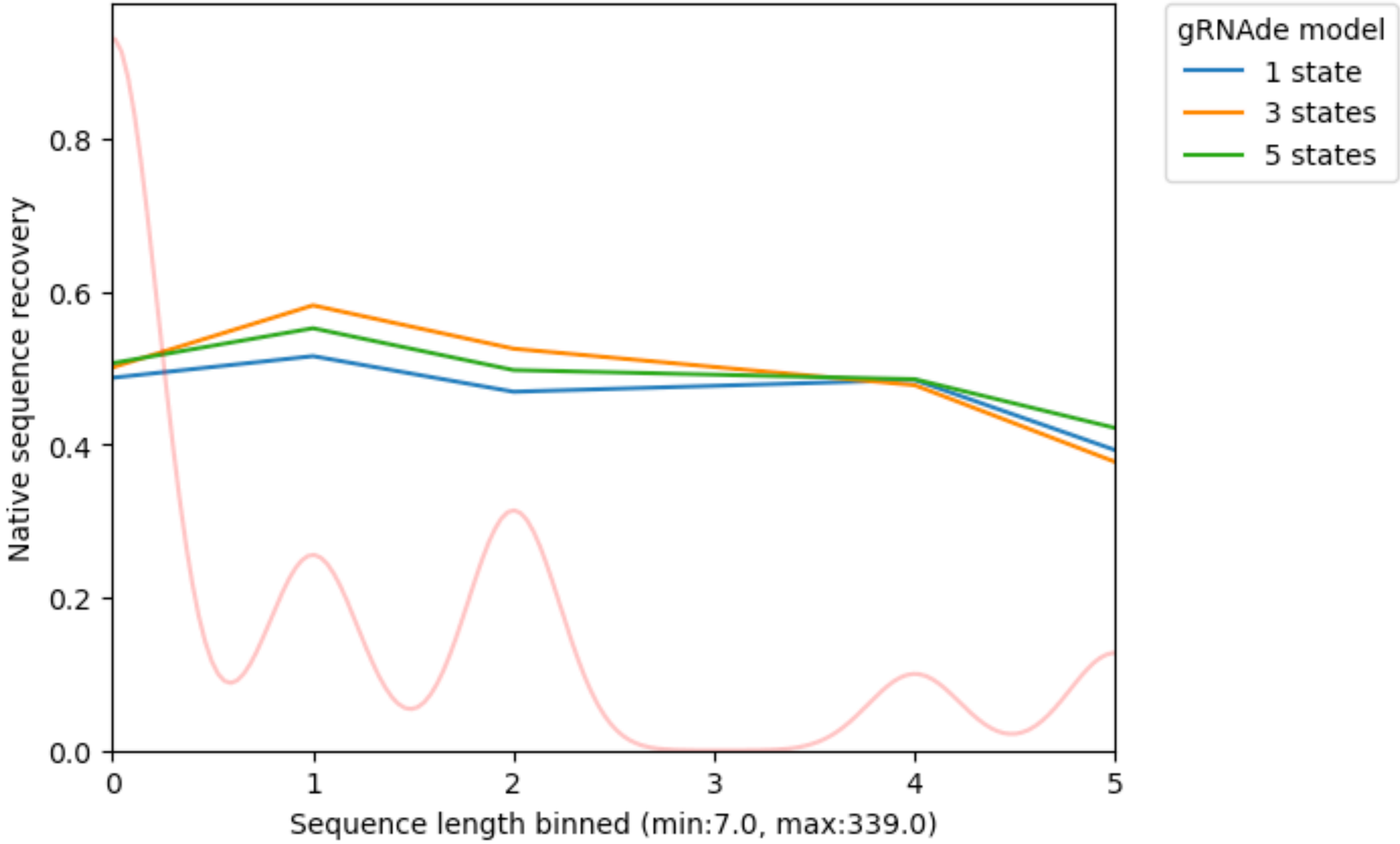
Highly variably located nucleotides

Multi-state models show improved recovery in variable regions



Nucleotides in longer sequences

Advantages of multi-state models for medium length sequences



Limitations & Future Work

Things we are thinking about

Application

- How to chose the number of states? (What's the design scenario?)
- How to prioritise amongst designed sequences?
- Wet lab validation?

Methods

- Support for multiple interacting RNA chains, or accounting for interactions with ligands.
- Support partial re-design, negative design against undesired conformations.
- Improved architectures and benchmarking of multi-state design.

Resources

- Open-source code and checkpoints: github.com/chaitjo/geometric-rna-design
- Tutorial available + forthcoming book chapter in *Methods in Molecular Biology*.

Thank you for listening! Questions?

Email: chaitanya.joshi@cl.cam.ac.uk, **Website:** chaitjo.com

Thank you to:

Pietro Liò, Arian Jamasb, Ramon Viñas, Charles Harris, Simon Mathis, and my labmates at Cambridge

Roger Foo (NUS, Singapore)

Phil Holliger (MRC LMB)

Alex Borodavka (Cambridge Biochemistry)

Janusz Bujnicki (IIMCB, Warsaw)