RNA 2D Structure Prediction: Integrating Shape/DMS-Mapseq and AI

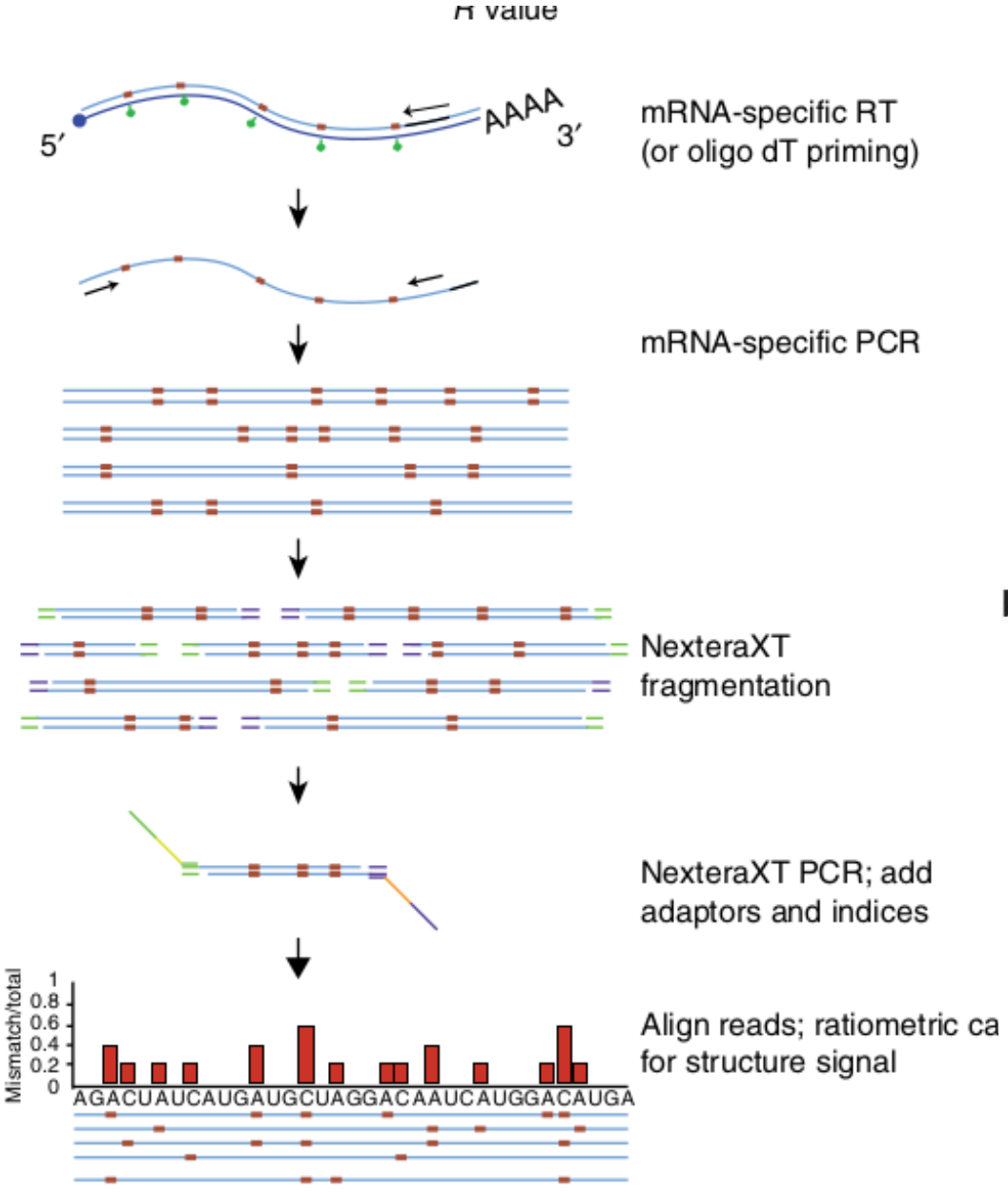
Rolf Backofen

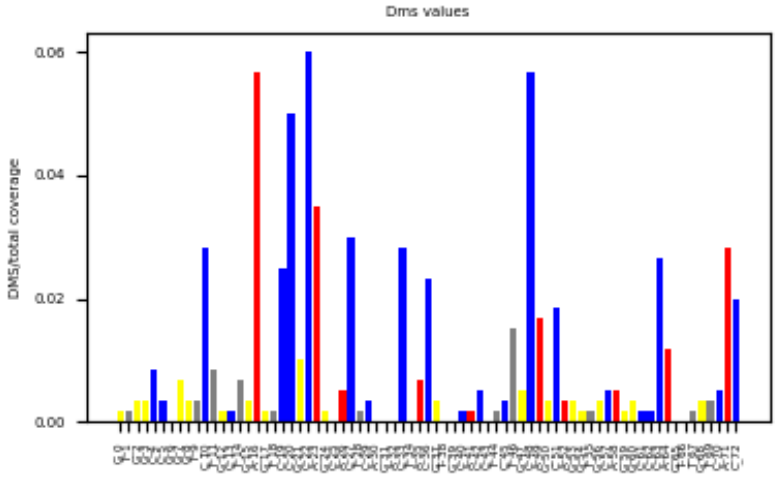
Lehrstuhl für Bioinformatik
Albert-Ludwigs-Universität

7. Juli 2026

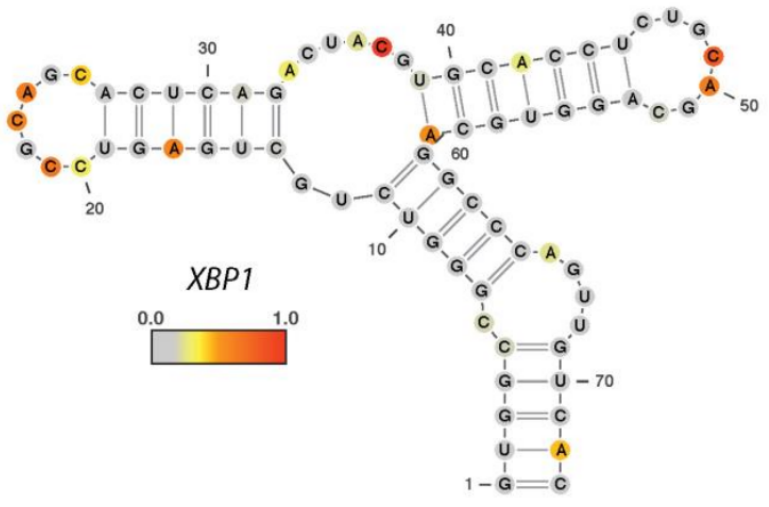
Part 1: RNA Secondary Structure via integrating DMS-MapSeq and OOPS-seq

Structural informed folding





b

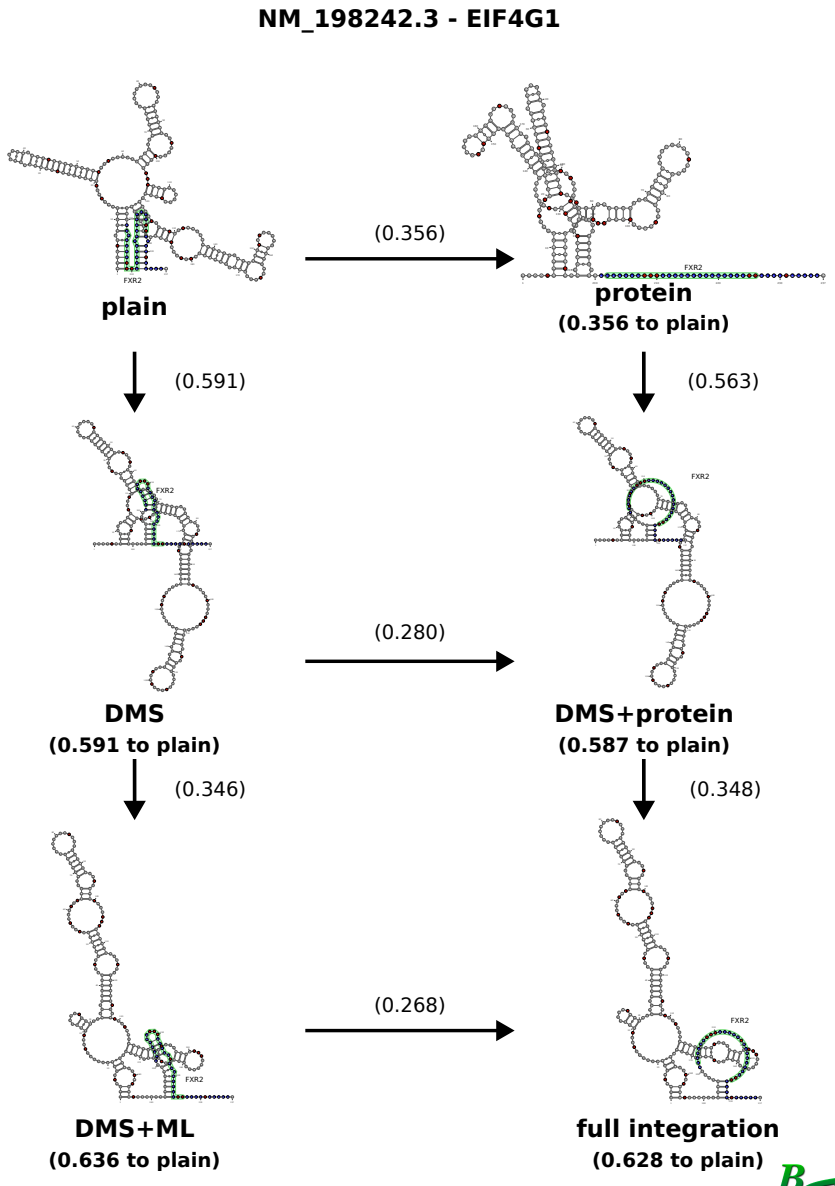


Example: EIF4G1 UTR

- EIF4G1-5' UTR struc1ture ensemble as determined by INSIGHT-RNA

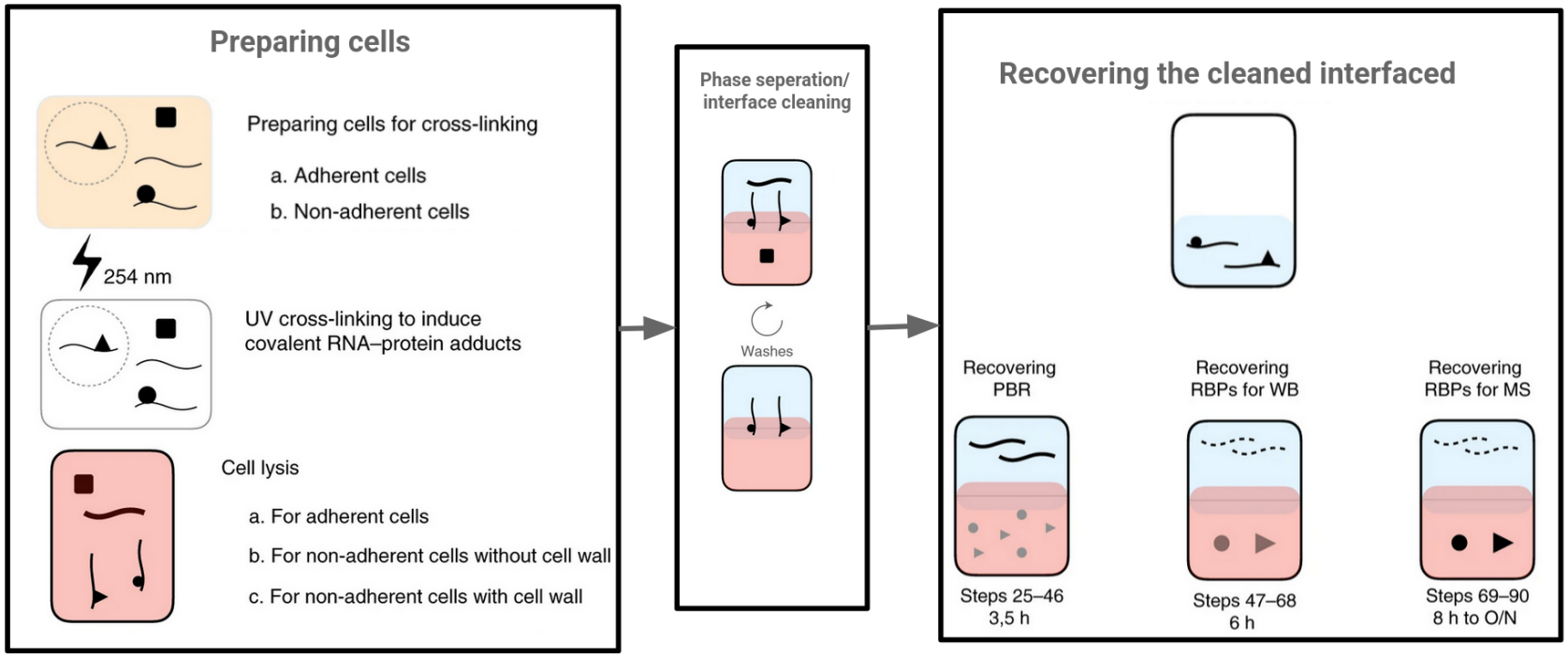
- shown for:
  - unconstrained (plain, *left-top*)
  - *in vivo* DMS-MaPseq-informed (DMS)
  - DMS-MaPseq-informed + machine learning-refined (DMS+ML)
  - OOPS-seq/eCLIP-informed (protein)
  - DMS-MaPseq and protein interaction-informed (DMS+protein)
  - integrating all sources (full integration)

- Jensen-Shannon divergence (JSD) of the structure ensembles to indicate the magnitude of the changes

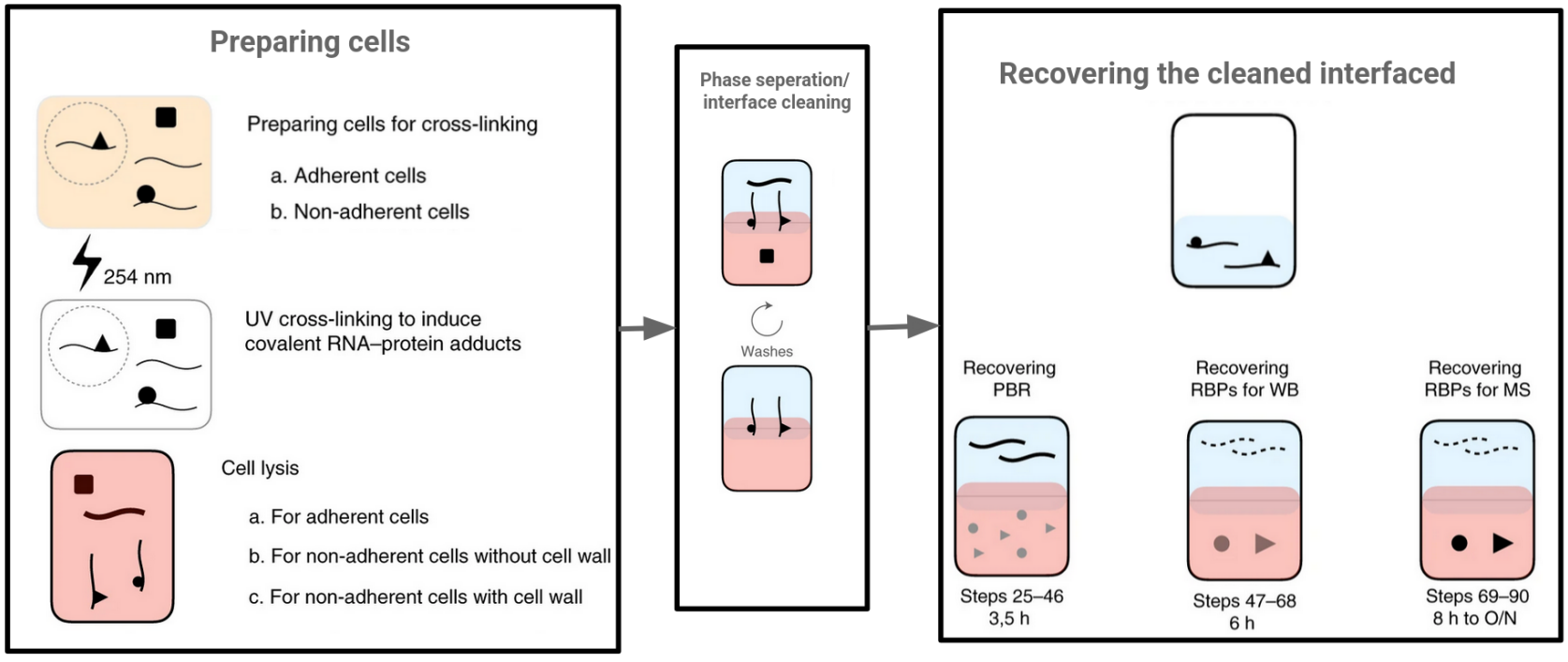


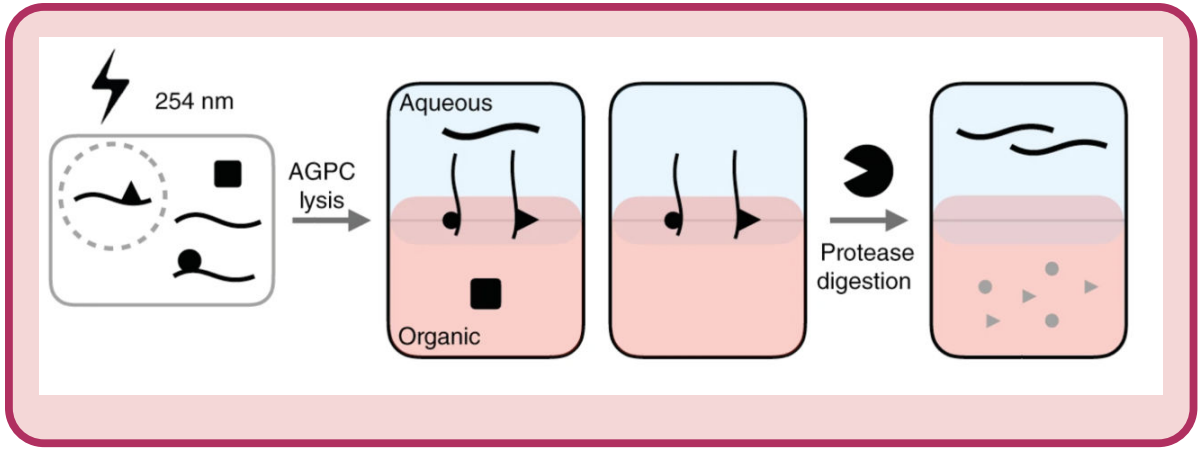
Orthogonal Organic Phase Seperation Sequencing (OOPS-Seq) based analysis

Basic OOPS-seq protocol



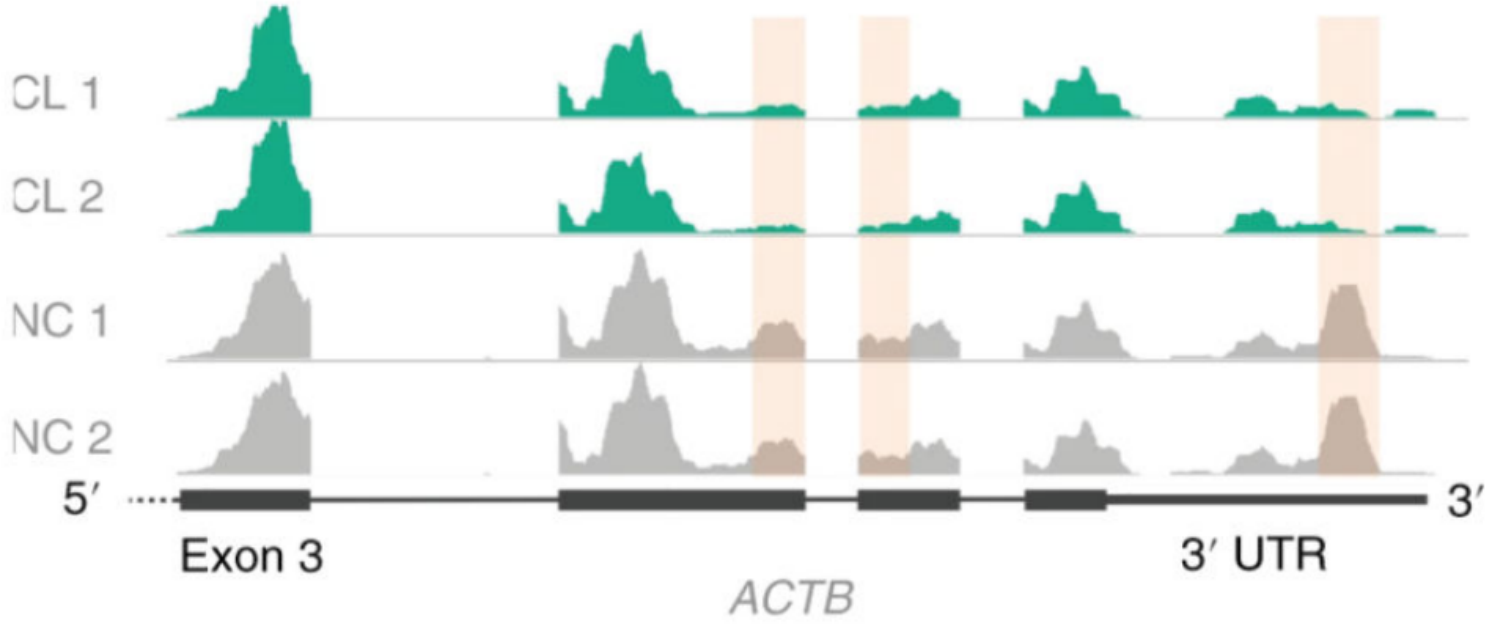
Basic OOPS-seq protocol





Nat Biotechnol. 2019, 1, 169–178. doi:10.1038/s41587-018-0001-2

Basic OOPS-seq protocol (Example)

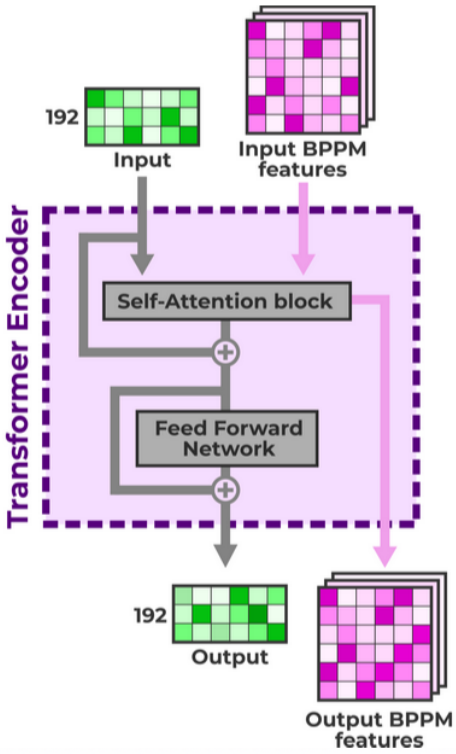


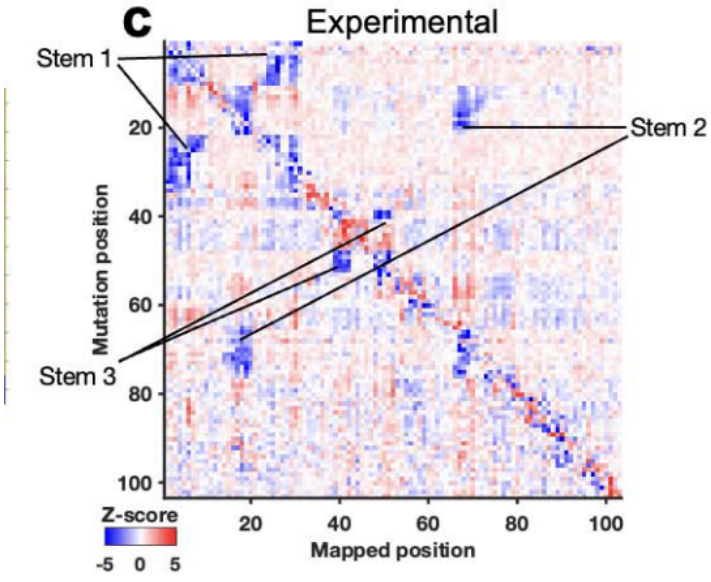
- Example of an OOPS-seq profile for ACTB: comparison between 2 cross linked and 2 non-cross linked replicates
  - region of significant difference in the 3' UTR are highlighted
  - protein binding leads to less reads in cross-linked replicates.

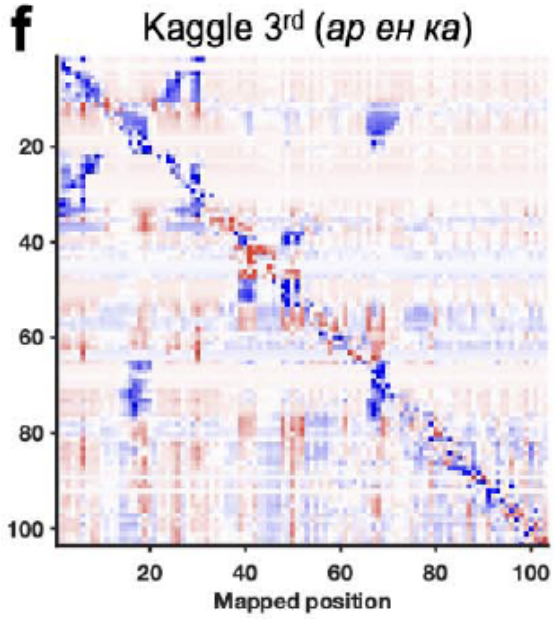
Nat Biotechnol. 2019, 1, 169–178. doi:10.1038/s41587-018-0001-2 .

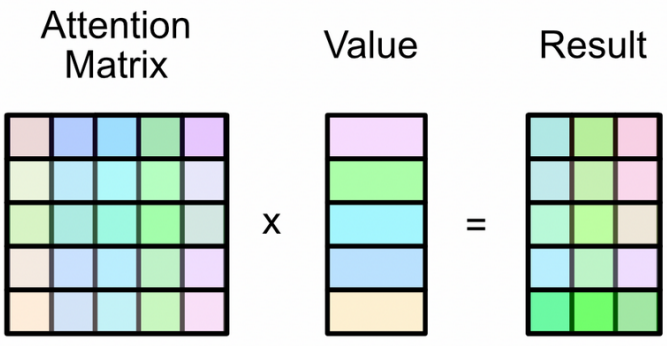
In silico prediction and incorporation of information sources

Transformer architecture





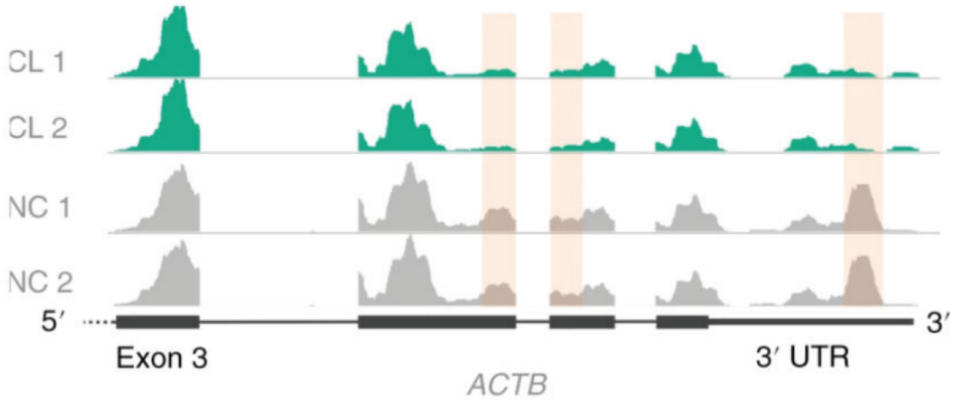




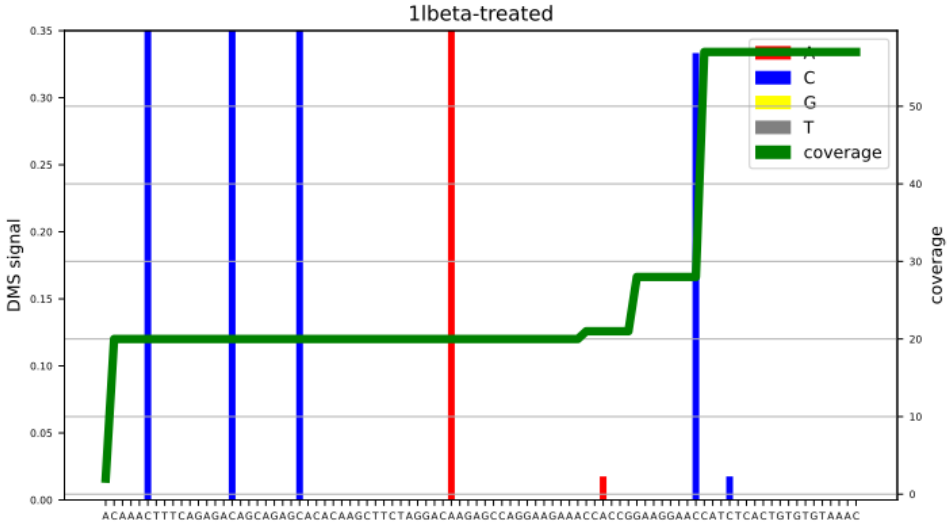
- Self-attention mechanisms are basic matrix multiplications determining the importance of each token to every other
- The dot product $Q \times K$ compares the query vector (Q) of each token with all other token's key vectors (K) multiplication, giving the attention matrix.

bioRxiv preprint https://doi.org/10.1101/2024.02.24.581671

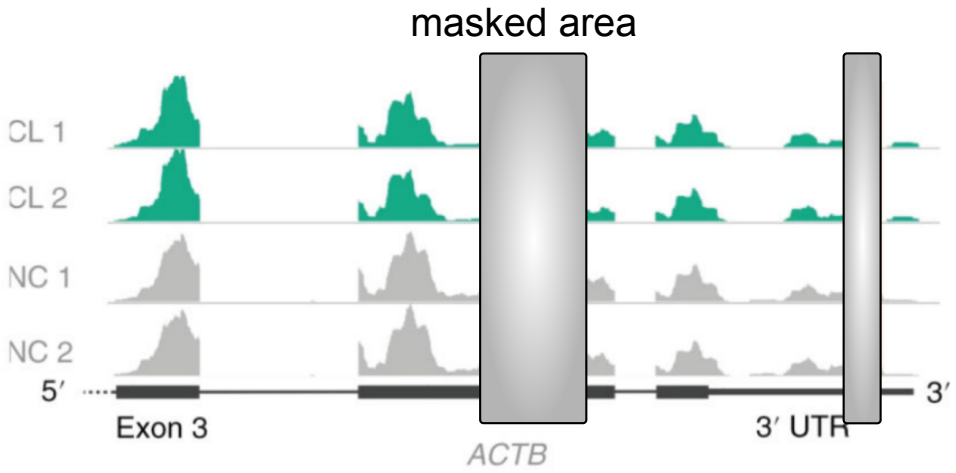
Adapting the Ribonanza Model for in-vivo datasets



- negative correlation between OOPS-seq data and DMS-MapSeq data
- occupied areas are less accessible for DMS
- mask areas that are covered by OOPS-seq areas
- mask areas with low coverage due to lower correlation

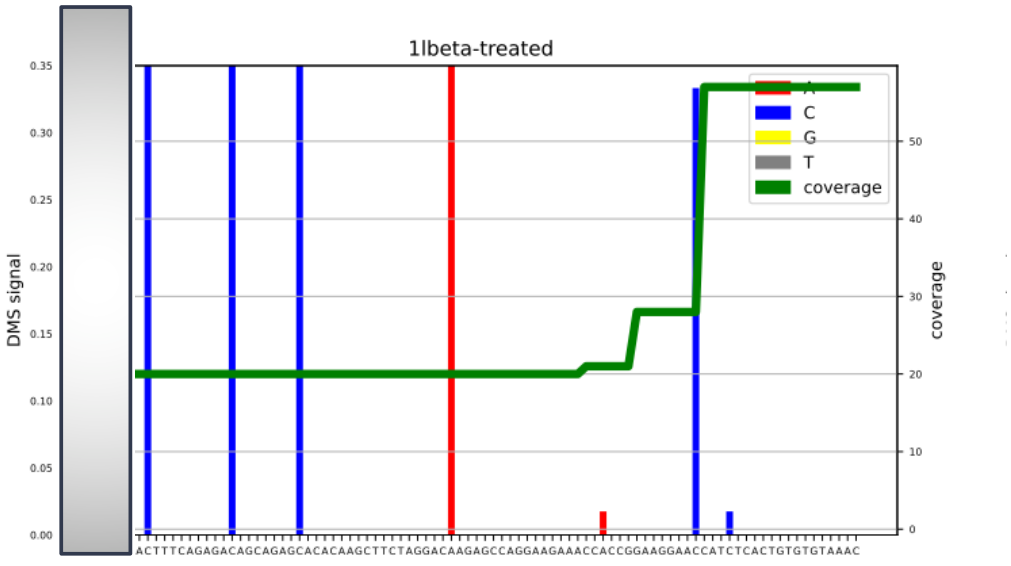


Adapting the Ribonanza Model for in-vivo datasets

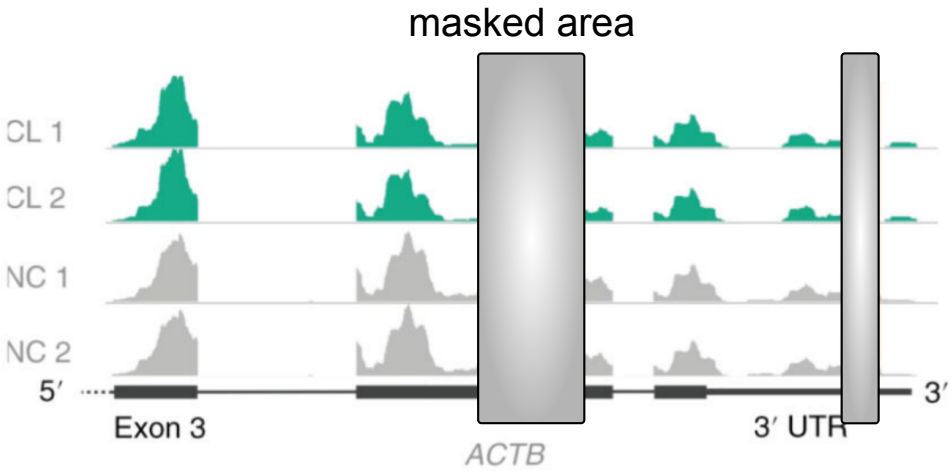


- negative correlation between OOPS-seq data and DMS-MapSeq data
- occupied areas are less accessible for DMS
- mask areas that are covered by OOPS-seq areas
- mask areas with low coverage due to lower correlation

masked area

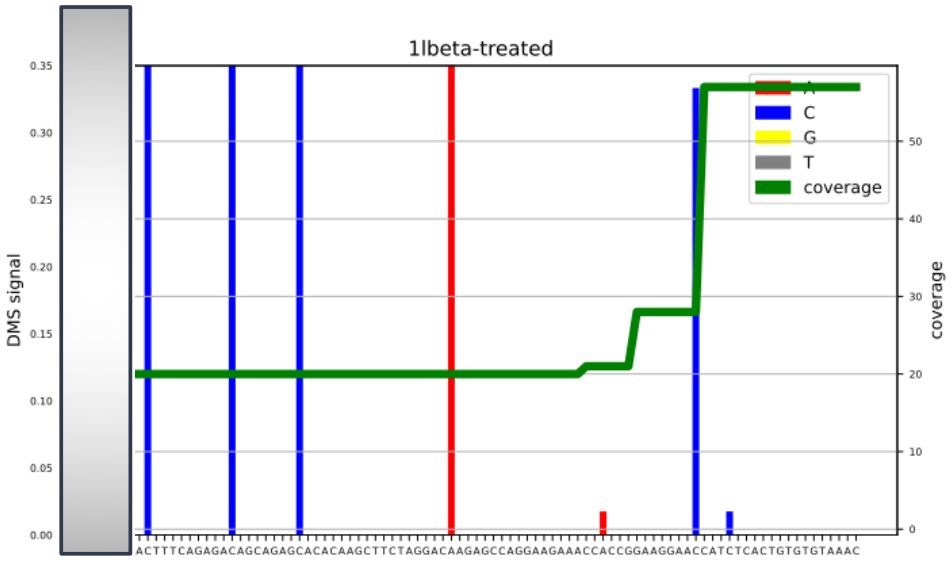


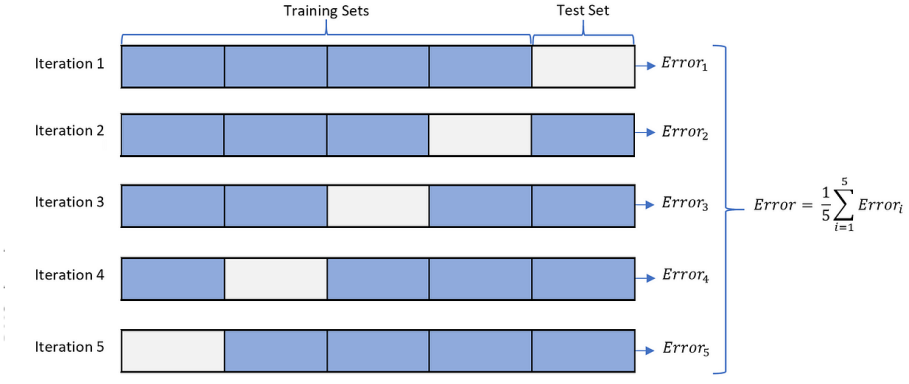
Adapting the Ribonanza Model for in-vivo datasets



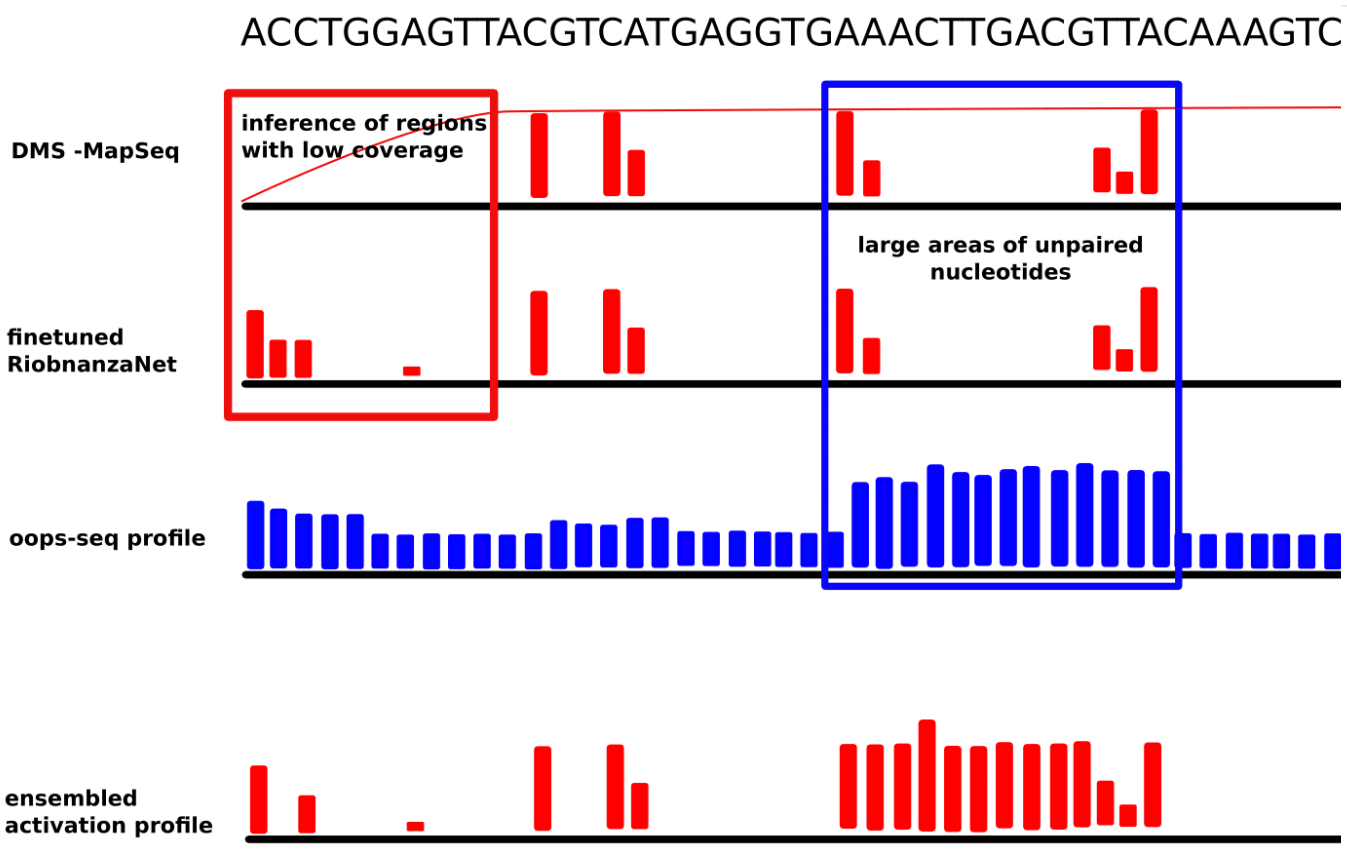
- negative correlation between OOPS-seq data and DMS-MapSeq data
- occupied areas are less accessible for DMS
- mask areas that are covered by OOPS-seq areas
- mask areas with low coverage due to lower correlation

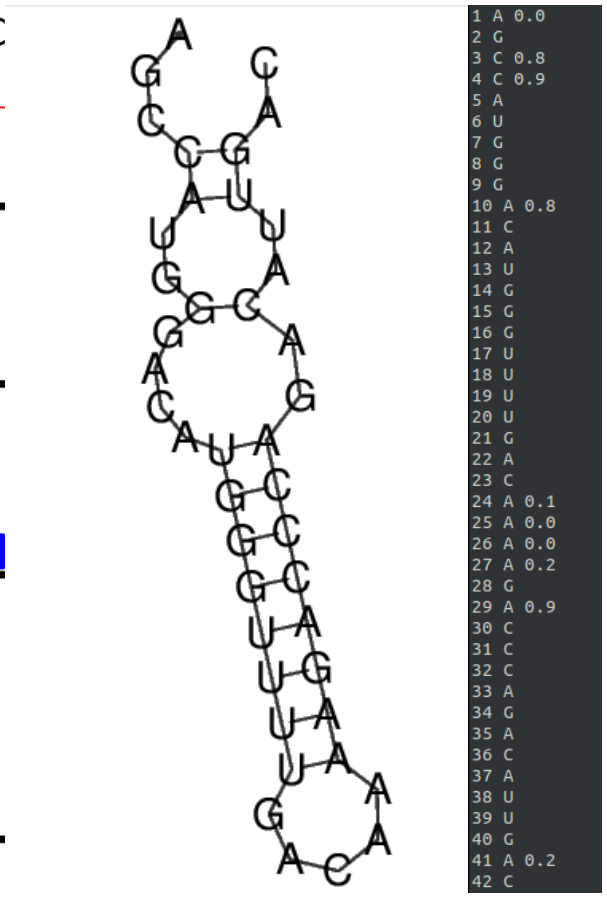
masked area



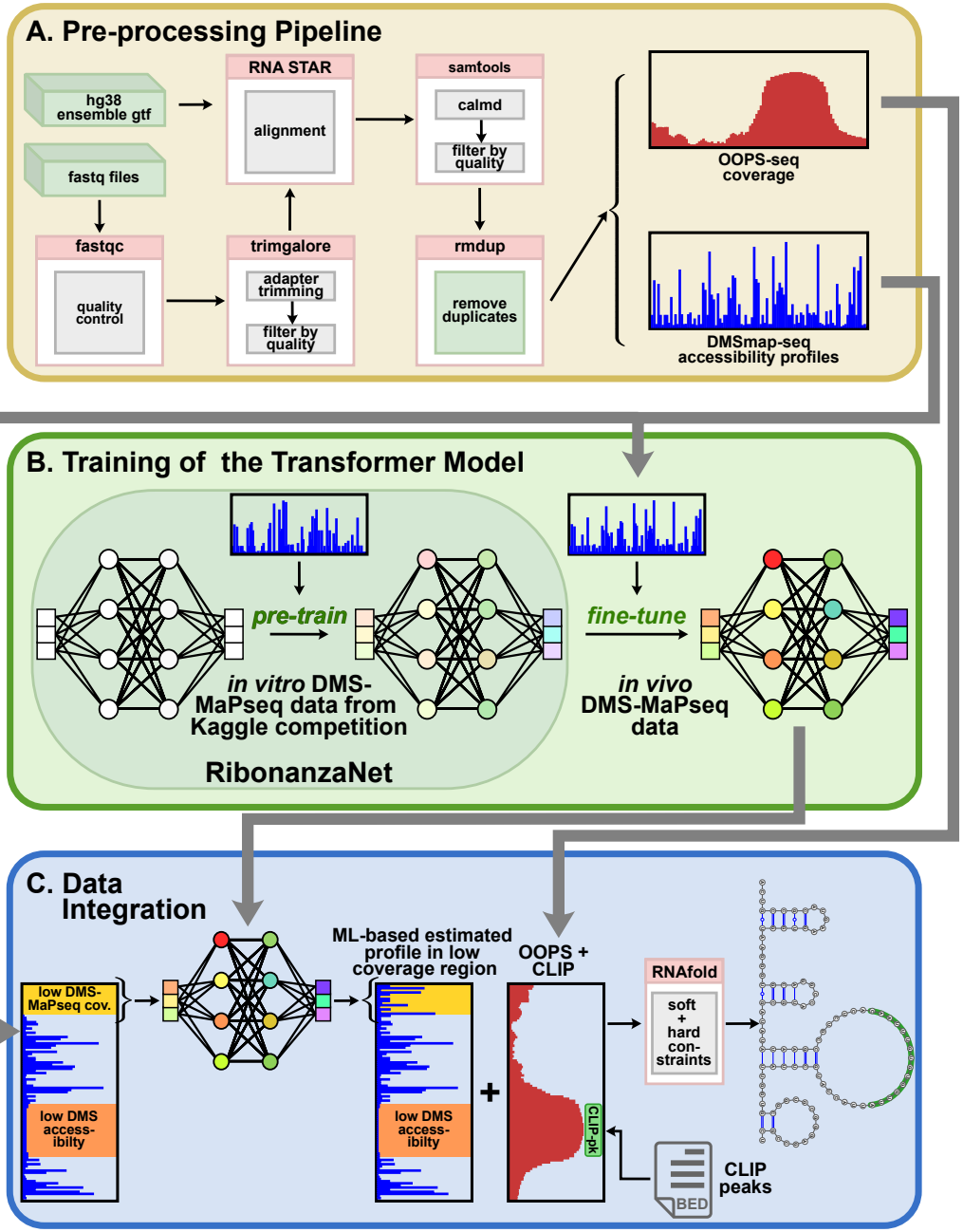


Final integration of data sources

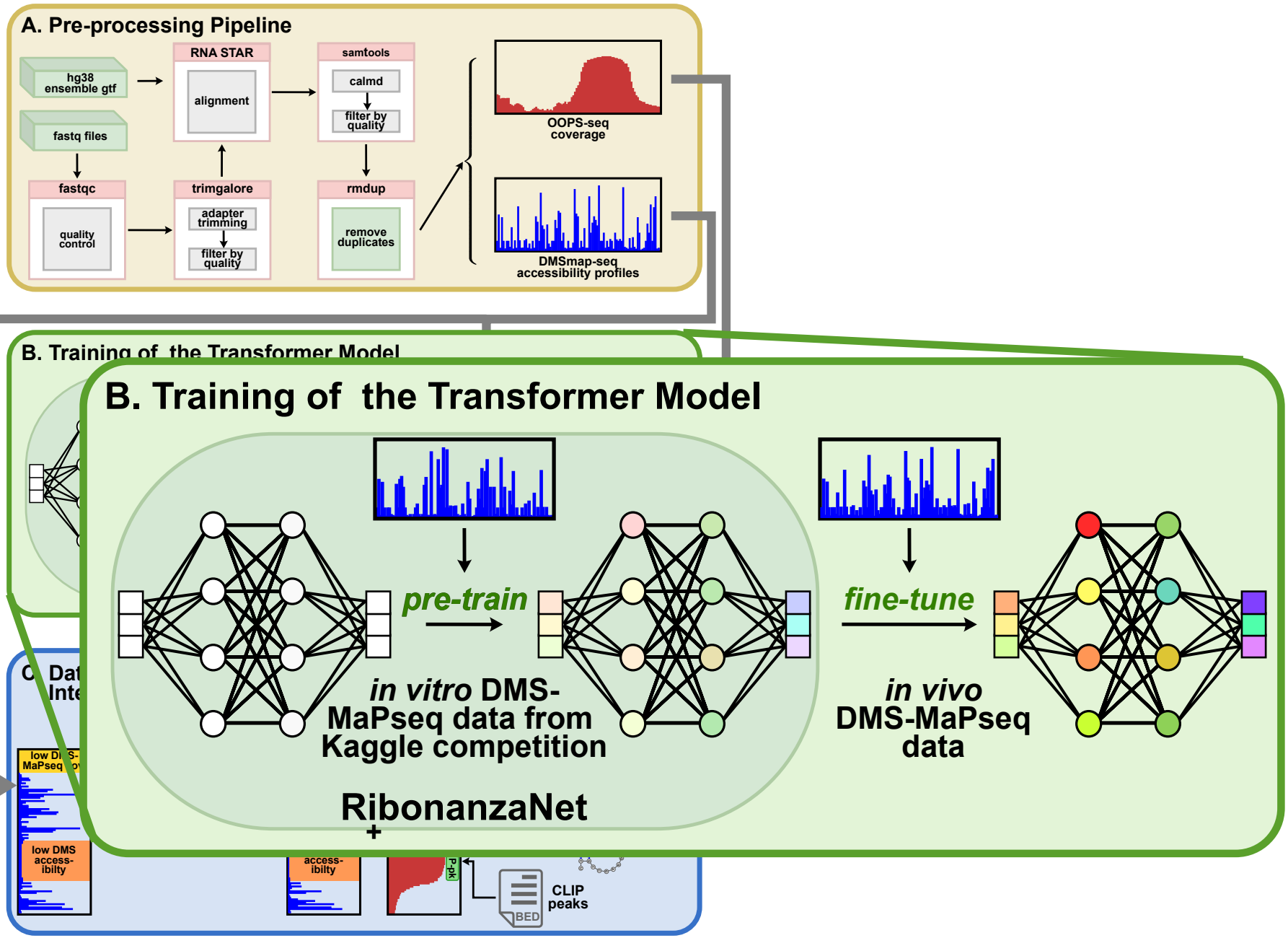




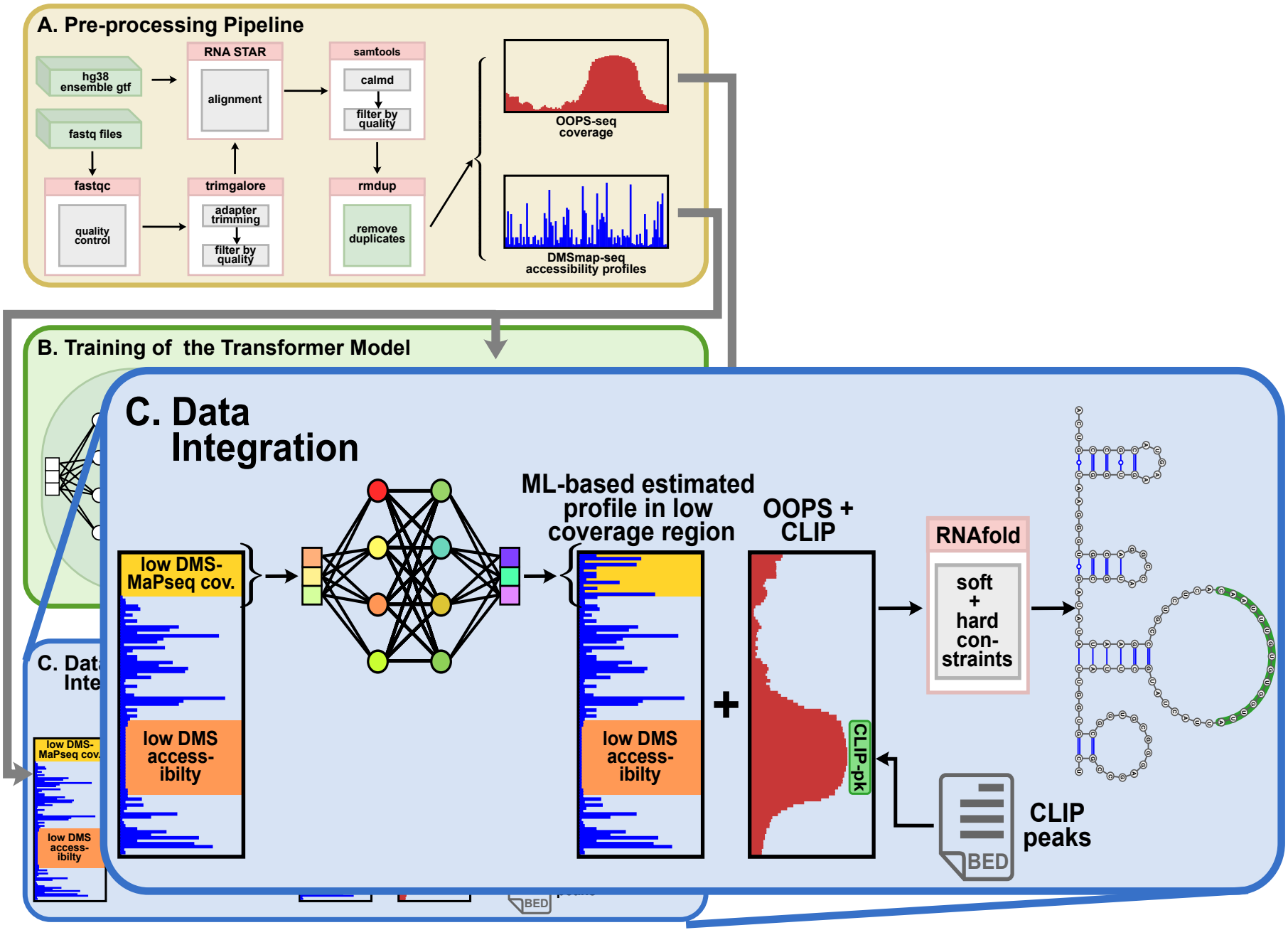
Summary of Workflow



Summary of Workflow



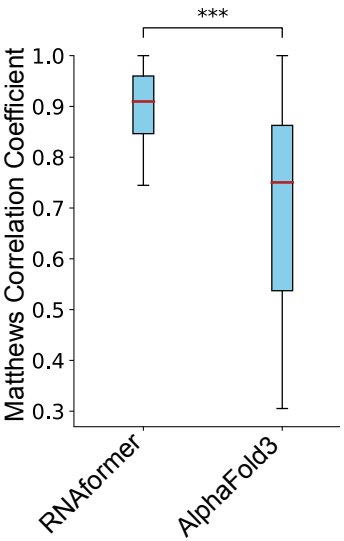
Summary of Workflow

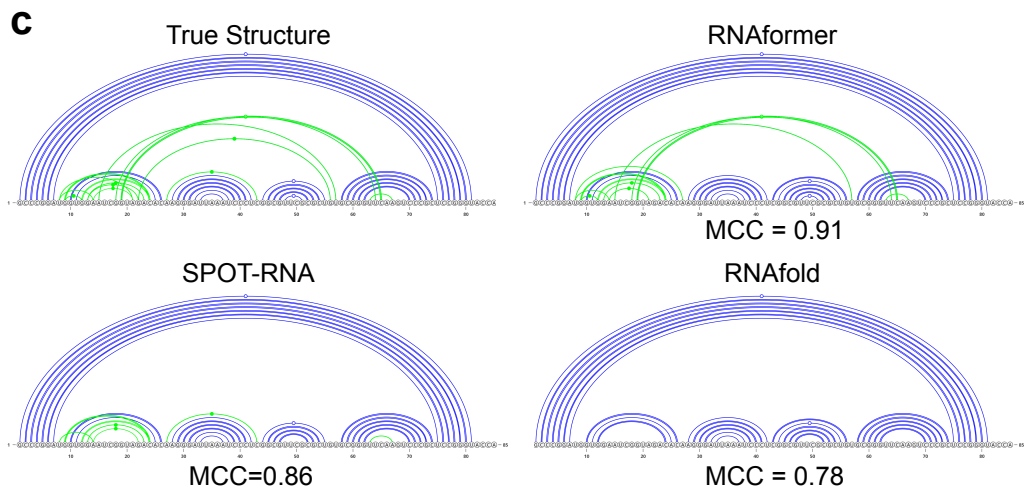


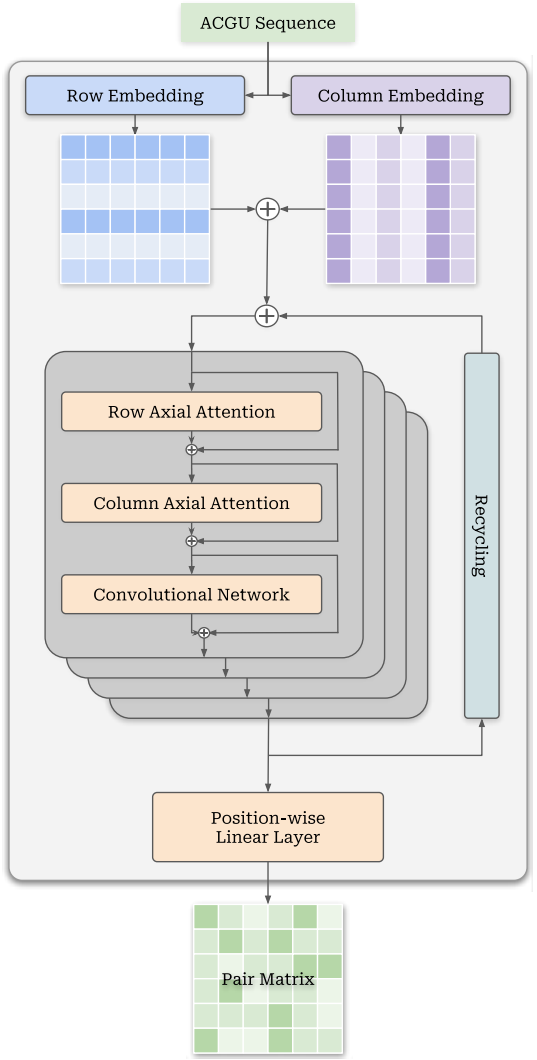
Part 2: AI-based 2D prediction from sequence (RNAformer)

RNAformer architecture

- Uses axial attention (long-range dependencies) to reduce computational costs followed by convolutional network (local dependency/motifs)
- output: base-pair matrix with **crossing** base pairs.
- careful design to reducing data homology between training and test data for both sequence and structure
- It works

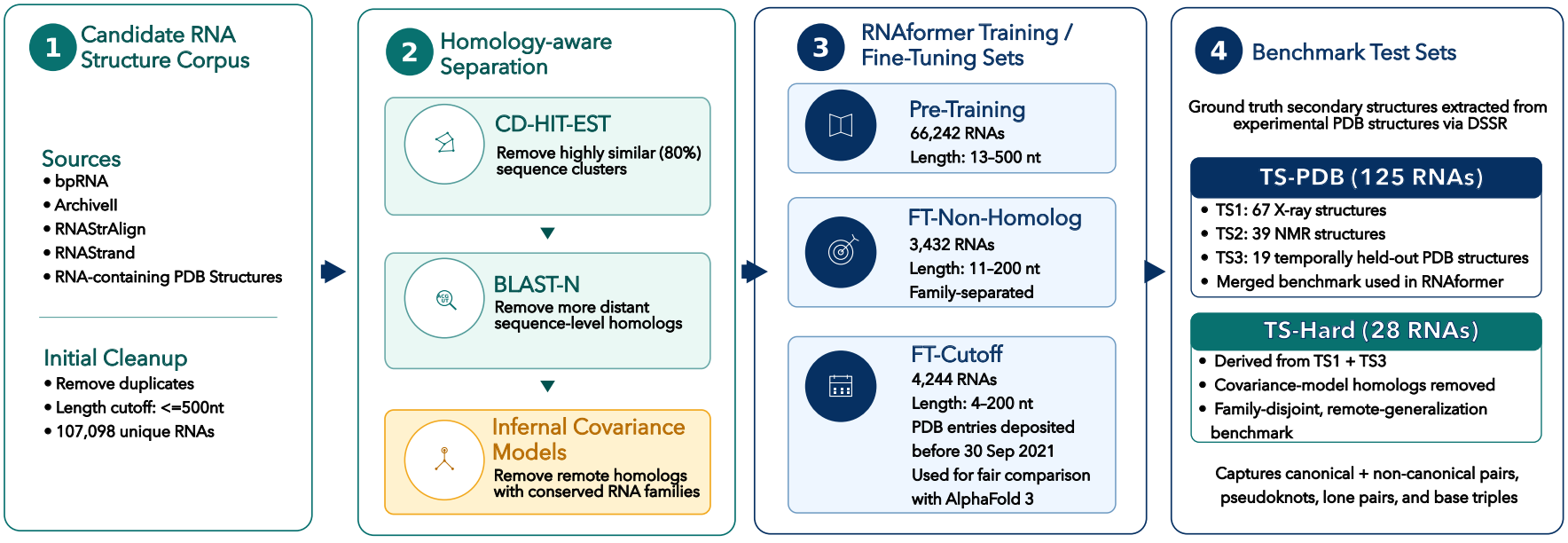






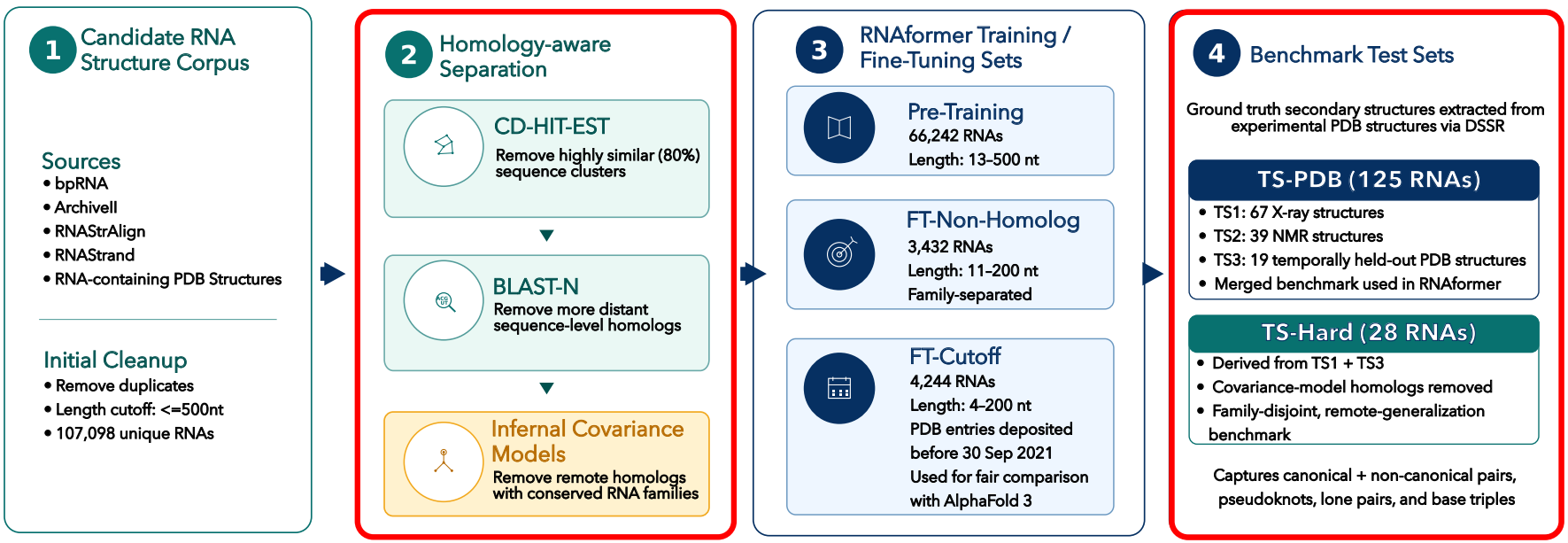
RNAformer's homology-aware Data Pipeline

- whole data pipeline:



RNAformer's homology-aware Data Pipeline

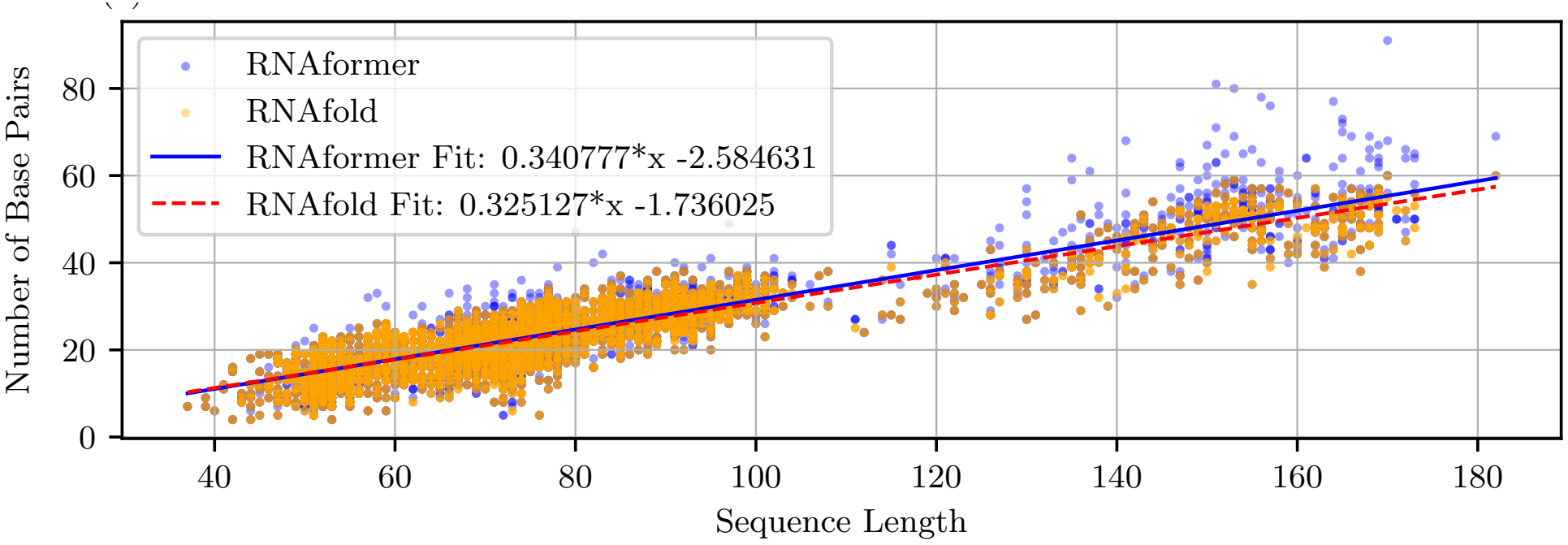
- whole data pipeline:



RNAformer extends identity level similarity separation at the sequence level with BLAST-N and uses Infernal to build covariance models and remove structures with conserved RNA families at the structural level

Answering some of the RNA communkty criticism

- One criticism was (correctly) that the homology problem was overseen → see last slide2
- Another one (correct) was that the number of predicted base-pairs predicted grew for ML-models quadratically in sequence length



- **AND:** we used it successfully to improve RNA 3D structure prediction

Why is RNA 3D Prediction Hard?

The amount of available 3D Structures is a problem.

Table 1. Numbers of all PDB-released structures (*) and residues in X-ray and cryo-EM structures (**) with high resolution (≤2.0 Å) over decades. In the first column, amino acids are abbreviated as AAs, and nucleotides as nts

Open in new tab

	≤1980	1981–1990	1991–2000	2001–2010	2011–2022	Total	% of the total
Proteins (*)	78	634	12 121	43 205	108 677	164 715	91.57
AAs ≤2.0 Å (**)	5050	45 236	1 609 401	11 390 238	28 513 777	41 563 702	99.78
RNA (*)	2	23	306	1392	4488	6211	3.45
RNA nts ≤2.0 Å (**)	0	0	1270	5974	26 921	34 165	0.08
DNA (*)	1	91	1061	2009	5800	8962	4.98
DNA nts ≤2.0 Å	0	238	5430	15 730	38 107	59 505	0.14

Schneider et al., 2025

But also. . .

"[...] The efficiency of 3D RNA structure prediction is likely to be improved using information from multiple sequence alignments (MSA) [...] Such a strategy is also applied in AlphaFold"

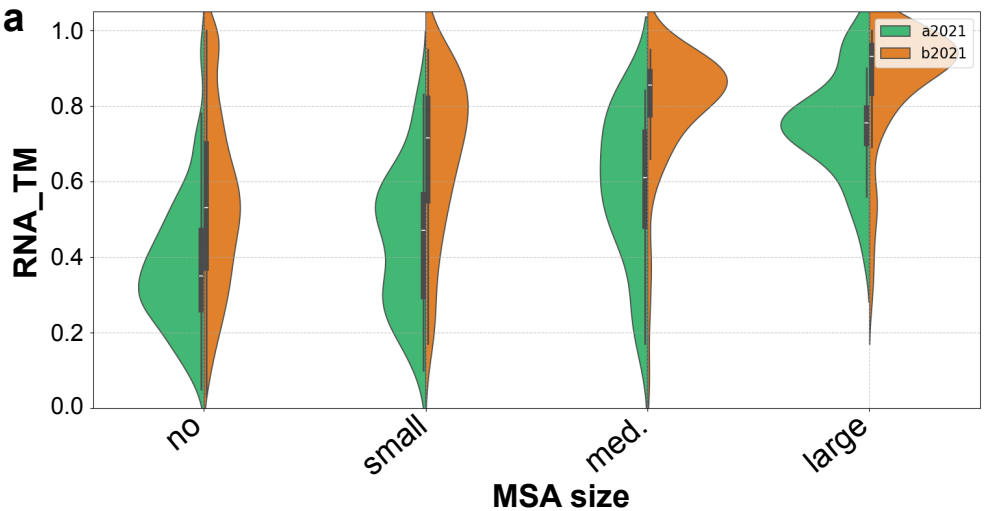
"[...] Rfam [...] contains far less data than Pfam."

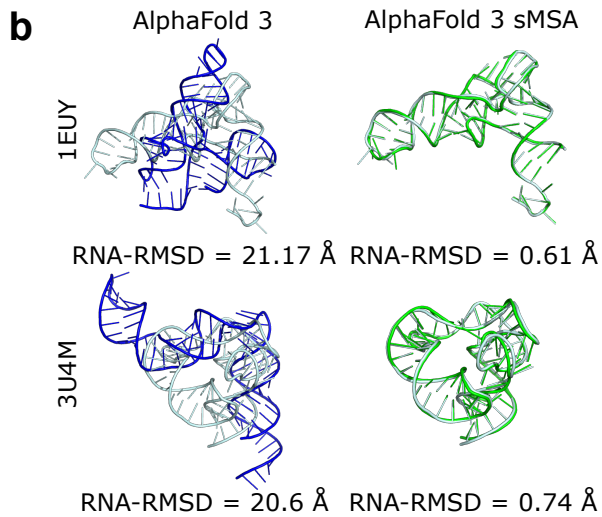
"[...] RNA alignments are on average smaller than protein alignments."

"[...] Rfam alignments have several global biases"

RNAformer-based MSA do improve AF3 prediction

- We use RNAformer to generate synthetic RNA MSA combined with light-weight structure-aware mutation scheme
- It improves AF3 prediction especially for cases where AF3 does not find sufficient members for MSA.





Take-Home Messages

- For mRNAs, we need to assess mRNA structure and protein-RNA binding
- for each, there are separated protocols

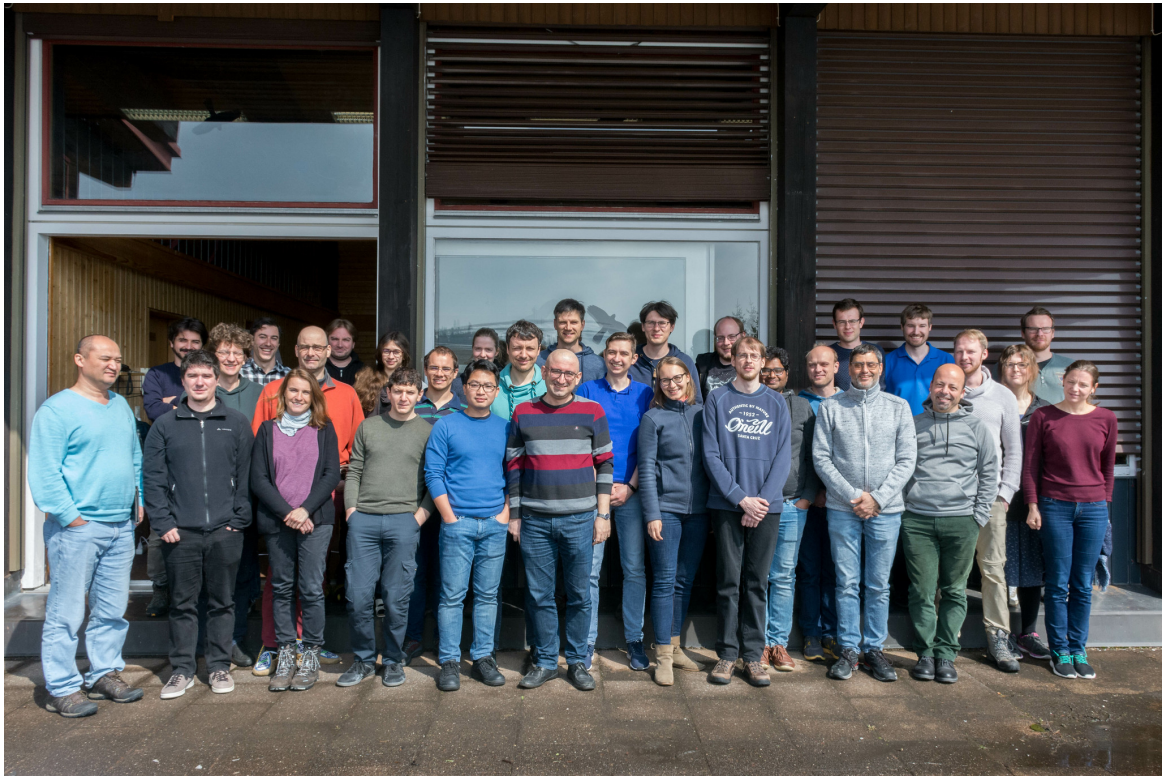
  CLIP-seq (single RBP) and OOPS-seq for protein-RNA interactions
  DMS-seq for mRNA structure

- however: each of them suffers from a low coverage/ missing data.
- Solution: Train specific AI models for each data type
  - CLIP-seq models are very advanced and have proven quality
  - DMS-seq and OOPS-seq are newer and more challenging
  - **BUT** important part is data integration.

- RNAformer
  - can predict all types of 2D structure
  - trained with a homology-aware pipeline, solve some AI-issues
  - 2D-predictions are sufficient to recreate most evolutionary information
  - Thus, synthetic, RNAformer-based MSA enhances AF3 in RNA 3D prediction

Acknowledgements

My Group



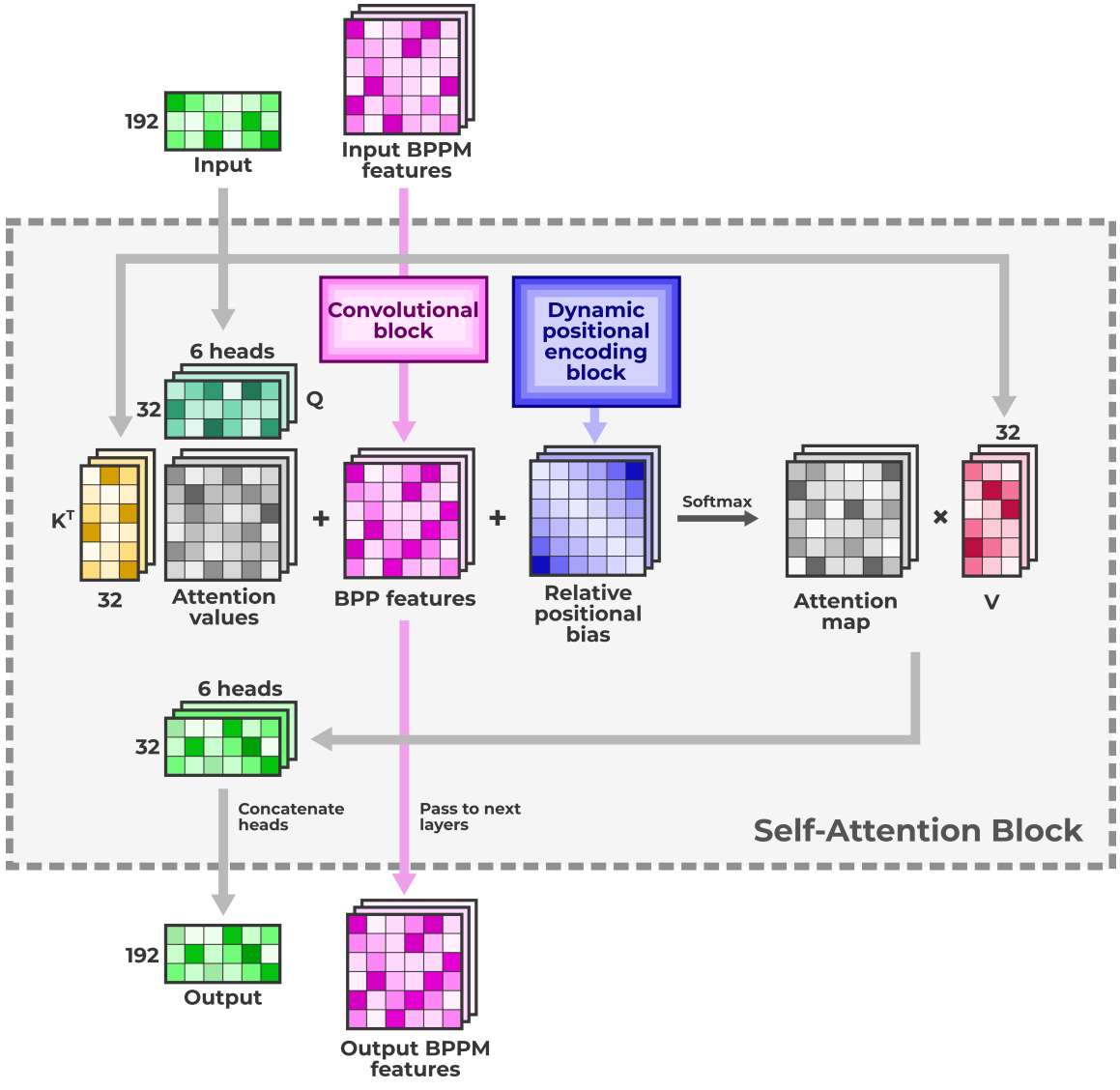
Collaboratoring groups

- Omer Salem Alkhnbashi
- Asifa Akthar
- Maria Stella Carro
- Arul M. Chinnaiyan
- Jan Gorodkin
- Markus Landthaler
- Matthias W. Hentze
- Julian König
- Cenk Sahinalp
- Tobias Schmid
- Gian Gaetano Tartaglia

DFG
Bundesministerium für Bildung und Forschung
Baden-Württemberg Stiftung
WIR STIFTEN ZUKUNFT
de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

Thank you.

In silico prediction of DMS values



- after attention values for each head are calculated, we add BPPM features updated by the 'Convolutional block'
- number of channels corresponding to the number of heads in the SA block

bioRxiv preprint https://doi.org/10.1101/2024.02.24.581671