

第5回バイオ・生成AI研究会

「深層学習とシミュレーションを組み合わせたRNAの立体構造の精密化」

東京科学大学 佐藤研究室 築山 翔

1. 自身の研究について

「深層学習とシミュレーションを組み合わせたRNAの立体構造予測」

2. RNAの言語モデルについて

RNA-FM, RCLM

3. 実働部隊として

1. 自身の研究について

「深層学習とシミュレーションを組み合わせたRNAの立体構造予測」

2. RNAの言語モデルについて

RNA-FM, RCLM

3. 実働部隊として

RNA 3D structure prediction

Background

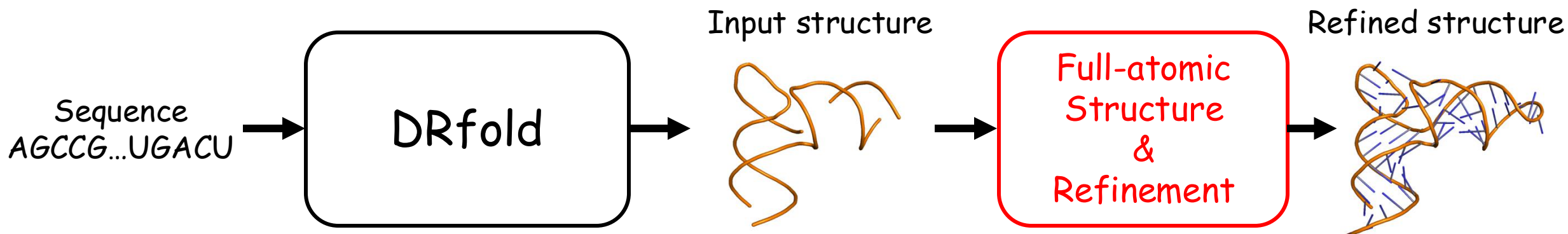
- Molecule 3D structure is predicted by computational approaches.
- The amount of data is limited (about 100 times less than that for proteins).
- Structure flexibility is high.

Purpose

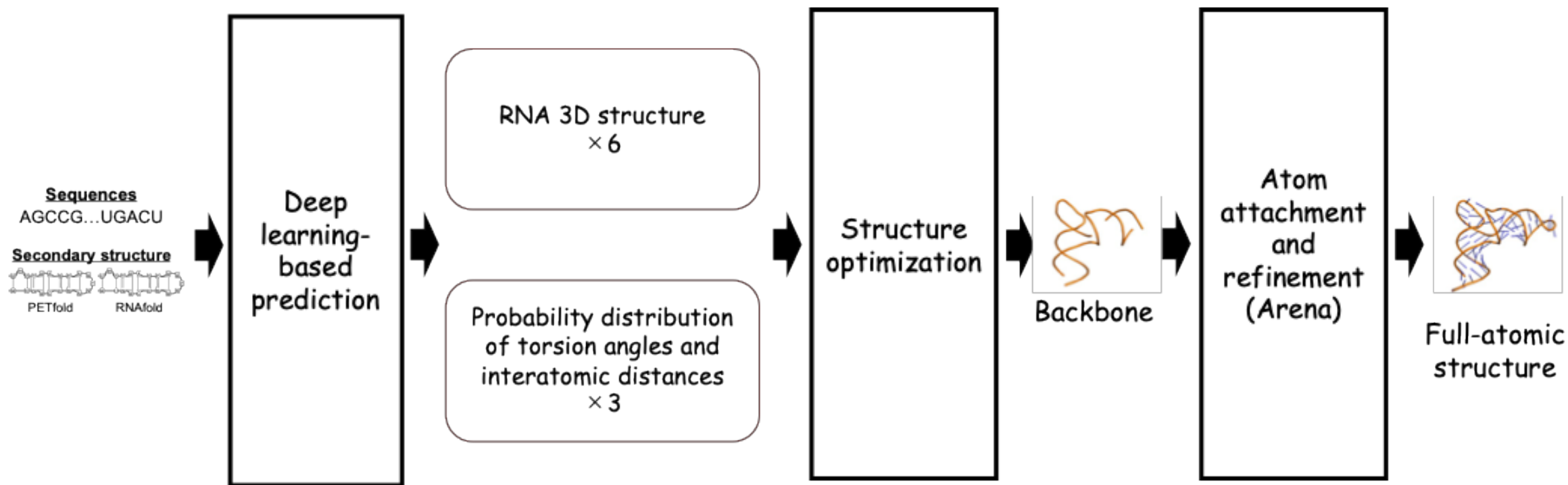
- Working with Yang Zhang's lab at NUS on RNA 3D structure prediction.
- Poor performance of local structure prediction in DRfold
- Development of a new refinement method



Yang Zhang lab.



- DRfold is an RNA 3D structure prediction method by deep-learning and L-BFGS-based optimization
- The inputs are a sequence and two secondary structures
RNAfold, PETfold



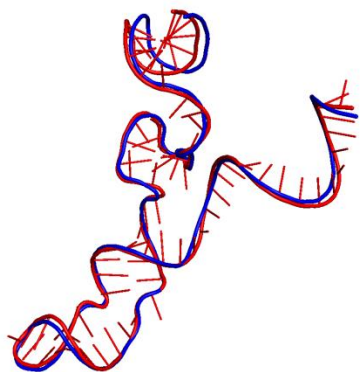
Yang Li et al. "Integrating end-to-end learning with deep geometrical potentials for ab initio RNA structure prediction." *Nature Communications* 14, no. 1 (2023): 5745.

Perry ZR, Pyle AM, Zhang C. "Arena: rapid and accurate reconstruction of full atomic RNA structures from coarse-grained models." *J Mol Biol.* 2023 Sep 15;435(18).

DRfold vs AlphaFold3

- 48 structures are used for the evaluation

	RMSD	TM-score	Clashscore	RMSD (lengths)	RMSD (angles)	INF
DRfold	10.6	0.408	50.8	0.0443	6.29	0.705
AlphaFold3	11.9	0.388	12.4	0.0179	1.19	0.709

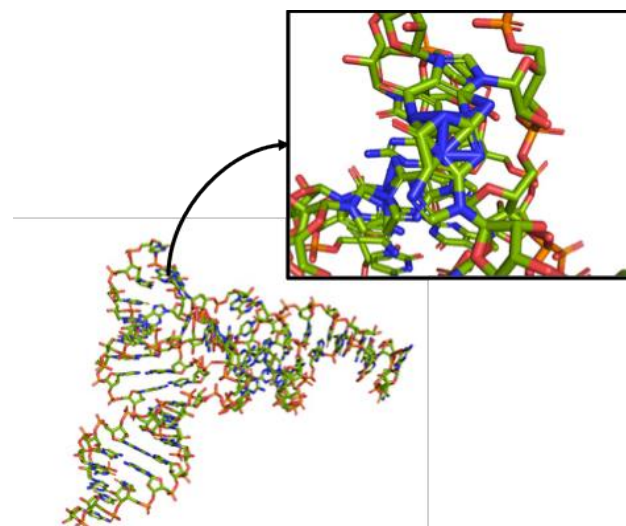


DRfold
TM-score: 0.713
RMSD: 2.127

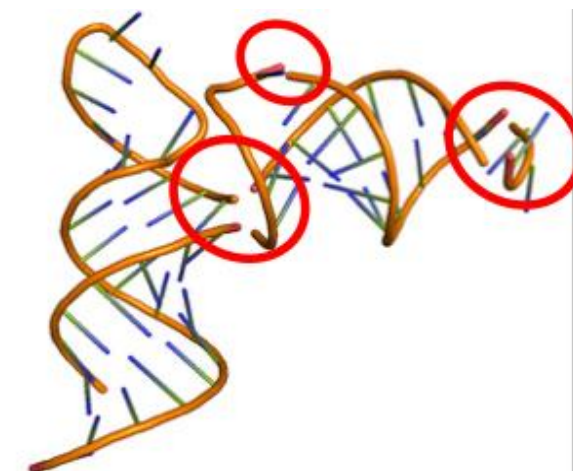


AlphaFold3
TM-score: 0.2689
RMSD: 14.744

Red: Experimental structure, Blue: DRfold, Green: AlphaFold3



Serious atomic clashes



Incorrect geometries

Development of new refinement method

1. 自身の研究について

「深層学習とシミュレーションを組み合わせたRNAの立体構造予測」

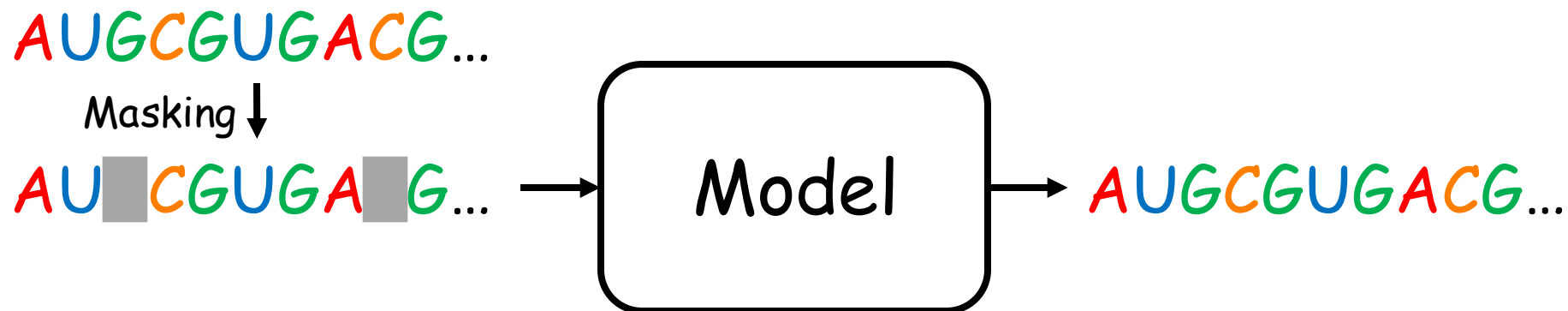
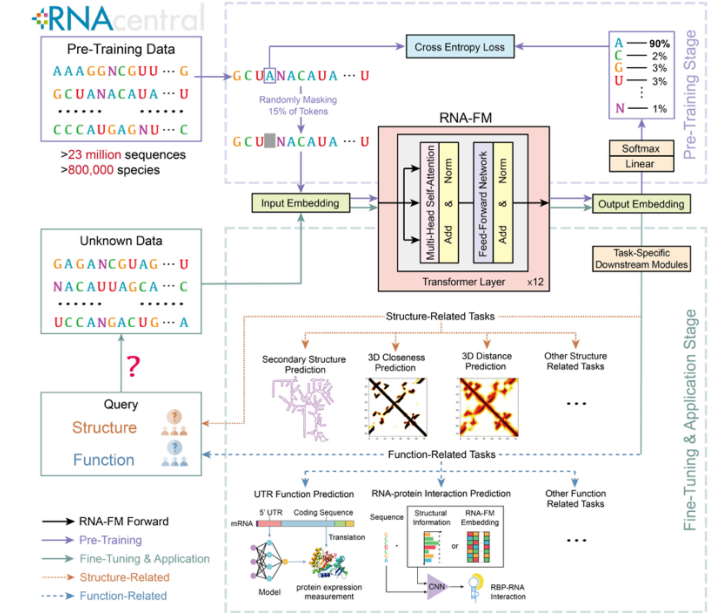
2. RNAの言語モデルについて

RNA-FM, RCLM

3. 実働部隊として

RNA language model (RNA-FM)

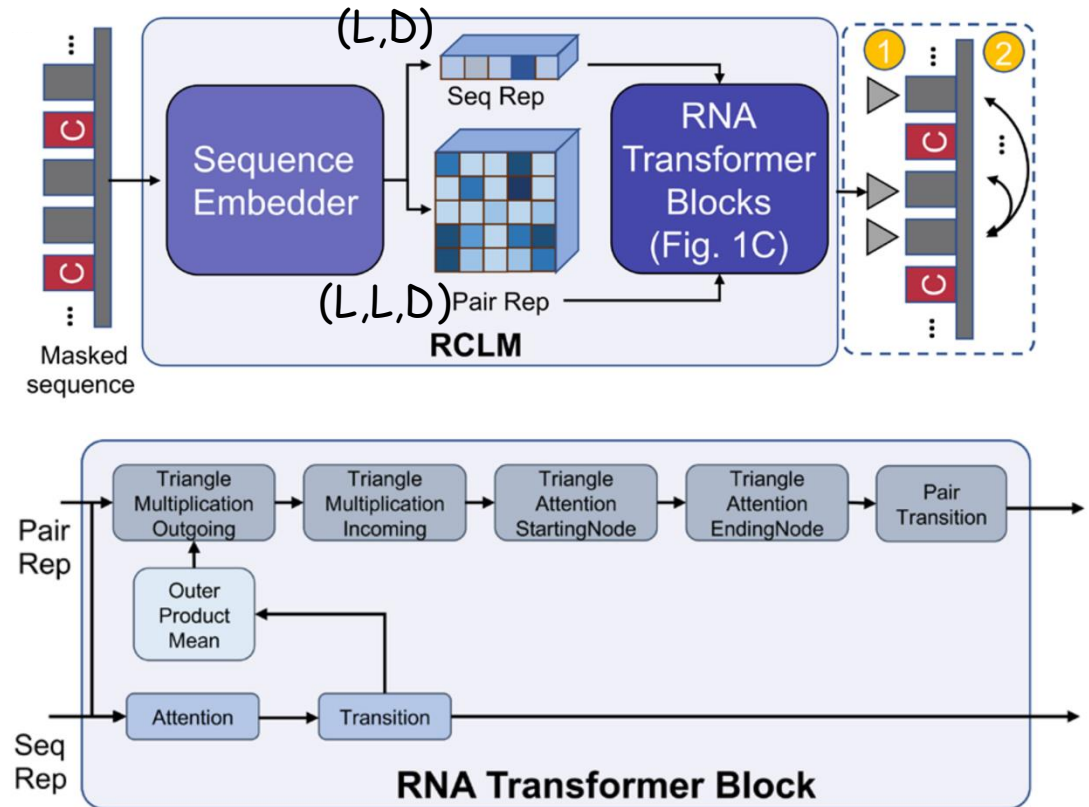
- BERT-based RNA language model.
- 23 million ncRNA sequences
- Training with masked language modeling
- The usefulness was demonstrated in RNA secondary structure prediction and distance map prediction.



Chen, J. et al., Interpretable RNA Foundation Model from Unannotated Data for Highly Accurate RNA Structure and Function Predictions. Eprint arXiv:2204.00300 (2022)

RNA language model (DRfold2; RCLM)

- DRfold2: RNA 3D structure prediction.
- RCLM generates a pair representation as well as a sequential representation
- The neural network consists of components such as triangle multiplication outgoing and triangle multiplication incoming.
- 30 million ncRNA sequences
- Training with masked language modeling
- To consider both individual tokens and interactions between pairs of masked positions,



$$L_{RCLM} = - \left(\underbrace{\sum_{i \in \mathcal{M}} \log P(X_i = x_i)}_{\text{Single nucleotide likelihood}} + \underbrace{\sum_{i, j: i \in \mathcal{M} \text{ OR } j \in \mathcal{M}} \log P(X_i = x_i, X_j = x_j)}_{\text{Pairwise composite likelihood}} \right)$$

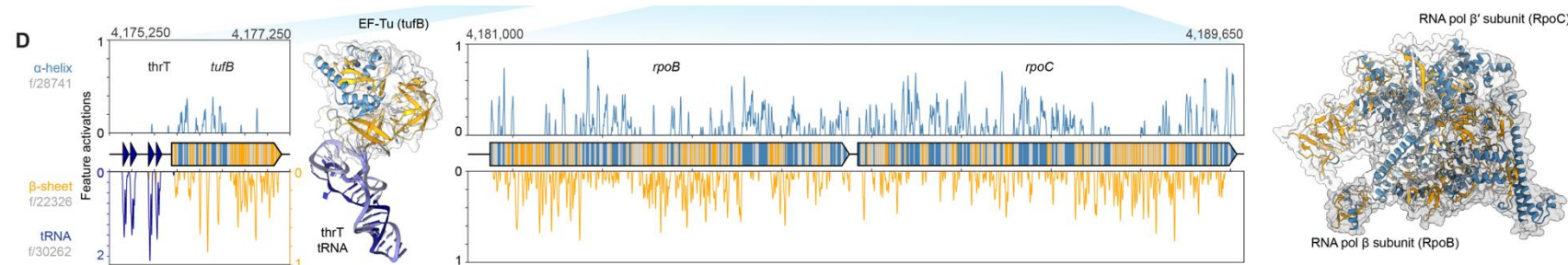
Single nucleotide likelihood Pairwise composite likelihood

RNA language model

RNA type	RNA language models
ncRNA	RNA-FM, RCLM, RNABERT, RiNALMo, RNA-MSM, AIDO.RNA, ERNIE-RNA, GenerRNA, RFamLlama, RNA-km, RNAErnie etc.
Coding RNA, 5'UTR, 3'UTR,	IsoCLR, SpliceBERT, Orthrus, LoRNA CodonBERT, G4mer, HELM UTR-LM 3UTRBERT etc.

1. 立体構造と特徴表現の関係性を調べる

- 構造情報とDNA言語モデル由来の特徴の関係性を解析
- Evo2: 2次構造 (α -helix, β -sheet) と対応する特徴が含まれている
- タンパク質構造レベルの高次な情報を捉えることを実証



Garyk B. et al, *Genome modeling and design across all domains of life with Evo 2*.
Eprint bioRxiv 2025.02.18.638918 (2025).

2. DNA言語モデルをDownstream taskに活用する

- タンパク質・RNA 2次/3次構造予測
- RNAの安定性、翻訳効率との関係