Deep Generative Framework with Physical/Stochastic Models in RNA Sequence Design

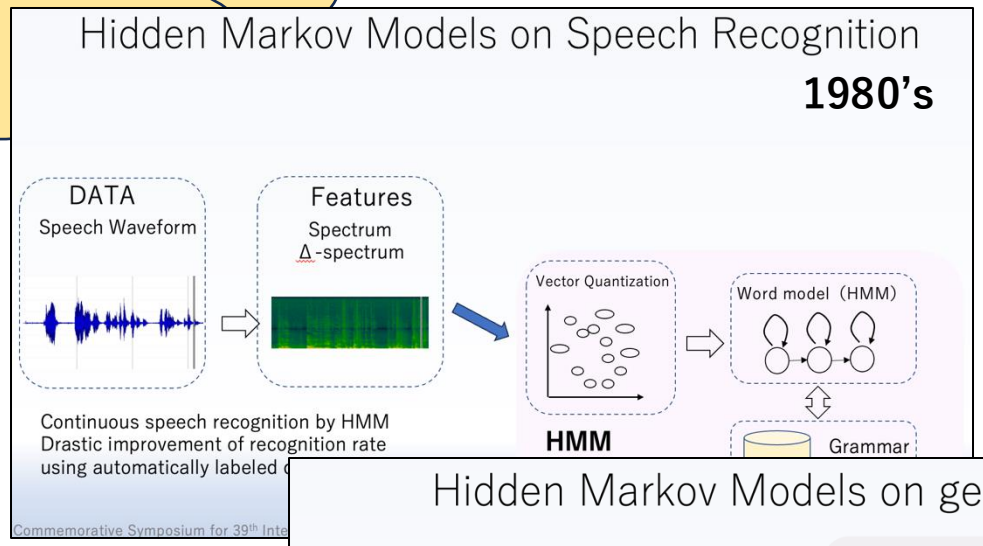
浅井 潔

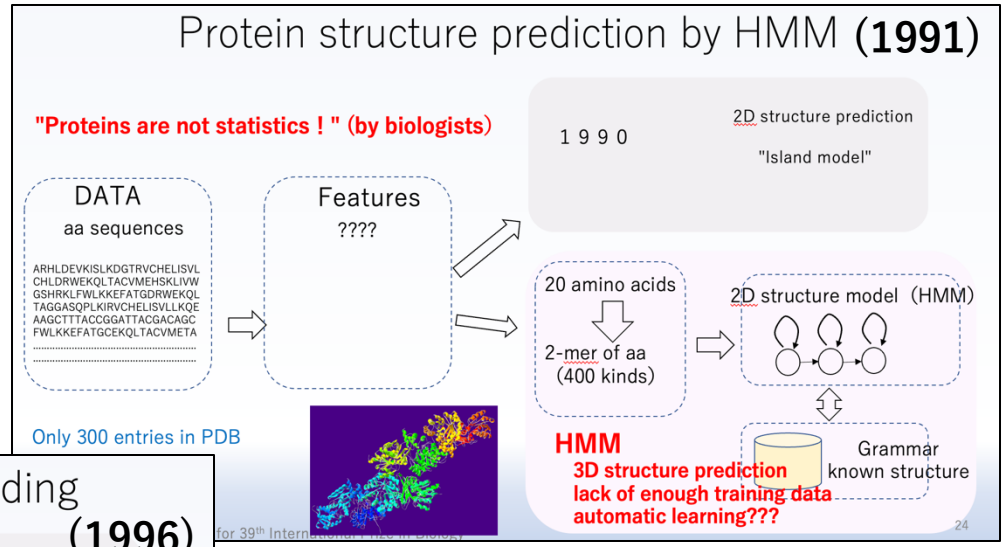
東京大学　大学院新領域創成科学研究科
メディカル情報生命科学専攻
生命データサイエンスセンター
理学部 生物情報科学科

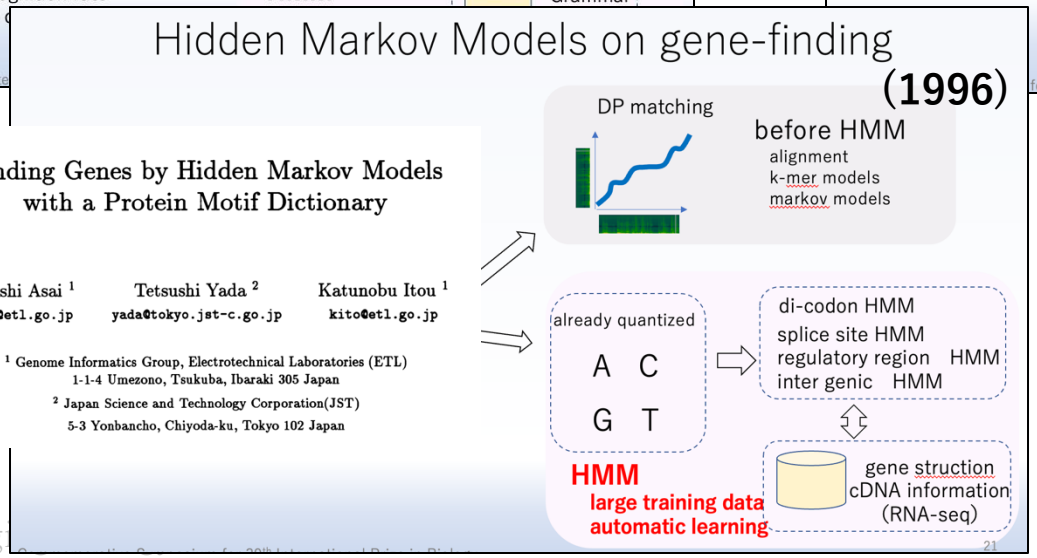
I love stochastic LANGUAGE models

ChatGPT

2020's







AlphaFold

2020's

連続音声認識技術のタンパク構造予測システムへの応用
Continuous Speech Recognition Techniques for Protein Structure Prediction Systems

◎浅井 潔[1]、速水 悟[1]、鬼塚 健太郎[2]、半田剣一[1]
Kiyoshi ASAI[1], Satoru HAYAMIZU[1],
Kentaro ONIZUKA[2], and Ken-ichi HANDA

1 通産省工業技術院電子技術総合研究所（ＥＴＬ）
Electrotechnical Laboratory
1-1-4 Umezono, Tsukuba Ibaraki, 305 Japan
E-mail: asai@etl.go.jp FAX: +81-298-585939
2 （財）新世代コンピュータ技術開発機構（ＩＣＯＴ）
Institute for New Generation Computer Technology

Genome Informatics 1992 Volume 3, 97-100.

https://doi.org/10.11234/gi1990.3.97

【要旨】

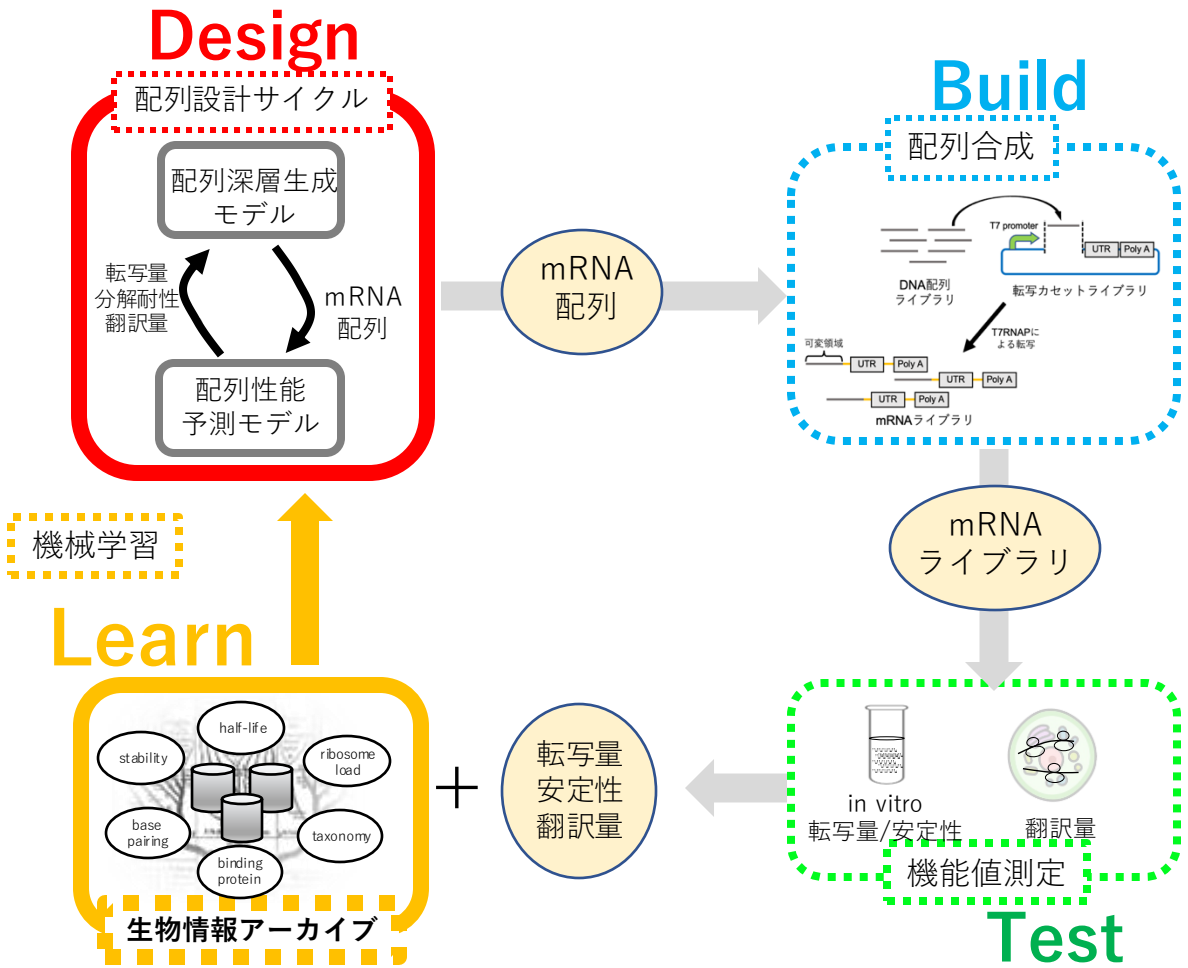
連続音声認識技術をタンパク質の構造予測の問題に適用することにより、タンパク質の局所構造と全体の立体構造を統一して扱う枠組をについて述べる。これはタンパク質データベースから得られる統計情報と、生物化学的知識を、確率モデルと文法的規則によって記述し、構文解析などの手法によって統合しようというものである。もちろん、対象である人の声とタンパク質とはあまり似ていない。似ているのは、どちらも階層構造を持っていることである。音声の場合は音素 → 単語 → 文節 → 文 → 意味、タンパク質の場合は、１次構造 → ２次構造 → 超２次構造 → 立体構造 → 機能と言う具合である。紹介する構造予測統合システムは、このような階層構造を意識して、それらを統合的に解析しながら組み立てる枠組である。

生物情報アーカイブを活用した
深層生成モデルによるmRNA最適設計技術

最新の深層学習では、
大量データで事前学習したモデルを
少量データで追加学習して高性能を実現

膨大な生物情報は公共データに蓄積
「生物情報アーカイブ」を活用

深層生成モデルによるmRNA最適設計技術
生物情報アーカイブを活用した深層生成モデル
配列生成と性能予測を繰り返す配列設計サイクル
実験量を大幅に削減したDBTL サ イクル



配列深層生成モデルの開発

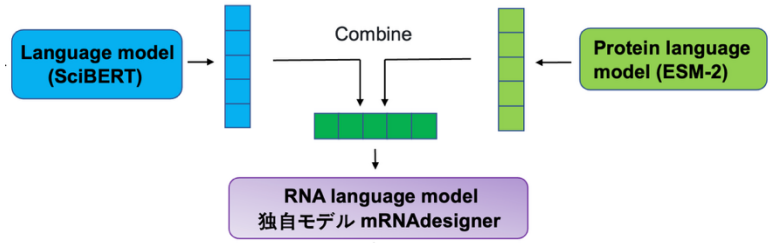
（齋藤裕＠北里大）

言語モデルによるmRNA配列生成

自然言語処理は文字列処理

生物配列にも適用可能

BLMとLLMを結合



（佐藤健吾＠東京科学大）

拡散モデルによるmRNA配列生成

ノイズ付加された情報から
もとの情報を復元するプロセスを学習

画像生成AIに使われている拡散モデルを応用して「好ましい」 mRNA 配列を生成する

（浅井研）

物理／確率モデル＋生成モデル

入力微分によるmRNA配列最適化

学習習済み深層学習モデルを入力データで微分
逆誤差伝播法が適応可能

入力微分は材料・分子設計等にも使われている

確率文法＋VAEによるRNA配列設計

確率文法と深層学習を結合

Optimizing **Parameters** in Computing (Prediction) System

Input (data) x → Computation $f(\cdot|\theta)$ → Output (**prediction**) $\hat{y} = f(x|\theta)$

Parameter

Gradient decent on parameters

$$\theta^{(t+1)} \longleftarrow \theta^{(t)} - \eta \frac{\partial \mathcal{L}}{\partial \theta}$$

loss function $\mathcal{L}(y, \hat{y})$

if $\mathcal{L}(y, \hat{y}) = (y - \hat{y})^2$

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial \mathcal{L}}{\partial f(x,\theta)} \frac{\partial f(x,\theta)}{\partial \theta} = 2\{\hat{y} - y\}\partial_\theta f(x,\theta)$$

$$\partial_\theta \equiv \frac{\partial}{\partial \theta}$$

Optimizing **Parameters** in Artificial **Neural Networks**

Input (data) x → Computation $f(\cdot|\theta)$ (Parameter) → Output (**prediction**) $\hat{y} = f(x|\theta)$

Gradient decent on parameters

$$\theta^{(t+1)} \longleftarrow \theta^{(t)} - \eta \frac{\partial \mathcal{L}}{\partial \theta}$$

Back-propagation

$$\delta_j^r \equiv \frac{\partial \mathcal{L}}{\partial a_j^r} = \sum_k \frac{\partial z_j^r}{\partial a_j^r} \frac{\partial a_k^{r+1}}{\partial z_j^r} \frac{\partial \mathcal{L}}{\partial a_k^{r+1}}$$

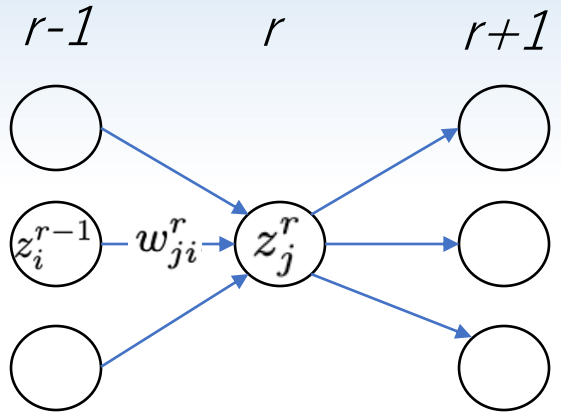
$$= h'(a_j^r) \sum_k w_{kj}^{r+1} \delta_k^{r+1}$$

$$\delta_j^L = \frac{\partial \mathcal{L}}{\partial a_j^L} = \frac{\partial \mathcal{L}(\hat{y} - y)}{\partial \hat{y}}$$

$$\frac{\partial \mathcal{L}}{\partial w_{ji}^r} = \frac{\partial \mathcal{L}}{\partial a_j^r} \frac{\partial a_j^r}{\partial w_{ji}^r} = \frac{\partial \mathcal{L}}{\partial a_j^r} z_i^{r-1} = \delta_j^r z_i^{r-1}$$

Artificial neural networks

$$w_{ji}^{(t+1)} \longleftarrow w_{ji}^{(t)} - \eta \frac{\partial \mathcal{L}}{\partial w_{ji}}$$



$$a_j^r = \sum_i w_{ji}^r z_i^{r-1}$$

Why '**derivative**' in RNA Sequence Design?

Optimizing **Inputs** in Computing (Prediction) System

Input (data) x → Computation $f(\cdot|\theta)$ (Parameter) → Output (**prediction**) $\hat{y} = f(x|\theta)$

Design Problem

Gradient decent on parameters

$$\theta^{(t+1)} \longleftarrow \theta^{(t)} - \eta \frac{\partial \mathcal{L}}{\partial \theta}$$

Gradient decent on inputs

$$x^{(t+1)} \longleftarrow x^{(t)} - \eta \frac{\partial \mathcal{L}}{\partial x}$$

Fixing parameters already optimized

Optimizing **Inputs** in Artificial Neural Networks

Input (data) → Computation → Output (**prediction**)

$$x \xrightarrow{f(\cdot|\theta)} \hat{y} = f(x|\theta)$$

Parameter

Back-propagation

$$\delta_j^r \equiv \frac{\partial \mathcal{L}}{\partial a_j^r} = \sum_k \frac{\partial z_j^r}{\partial a_j^r}\frac{\partial a_k^{r+1}}{\partial z_j^r}\frac{\partial \mathcal{L}}{\partial a_k^{r+1}}$$

$$= h'(a_j^r)\sum_k w_{kj}^{r+1}\delta_k^{r+1}$$

$$\delta_j^L = \frac{\partial \mathcal{L}}{\partial a_j^L} = \frac{\partial \mathcal{L}(\hat{y}-y)}{\partial \hat{y}}$$

$$\frac{\partial \mathcal{L}}{\partial w_{ji}^r} = \frac{\partial \mathcal{L}}{\partial a_j^r}\frac{\partial a_j^r}{\partial w_{ji}^r} = \frac{\partial \mathcal{L}}{\partial a_j^r} z_i^{r-1} = \delta_j^r z_i^{r-1}$$

$$w_{ji}^{(t+1)} \longleftarrow w_{ji}^{(t)} + \eta\frac{\partial \mathcal{L}}{\partial w_{ji}}$$

$$\frac{\partial \mathcal{L}}{\partial w_{ji}^r} = \frac{\partial \mathcal{L}}{\partial a_j^r}\frac{\partial a_j^r}{\partial w_{ji}^r} = \frac{\partial \mathcal{L}}{\partial a_j^r} z_i^{r-1} = \delta_j^r z_i^{r-1}$$

$$x^{(t+1)} \longleftarrow x^{(t)} + \eta\frac{\partial \mathcal{L}}{\partial x}$$

$$\frac{\partial \mathcal{L}}{\partial x_i} = \frac{\partial \mathcal{L}}{\partial a_j^1}\frac{\partial a_j^1}{\partial x_i} = \sum_j \delta_j^1 \partial_{x_i} \sum_i w_{ji}^1 x_i = \sum_j \delta_j^1 w_{ji}^1$$

Optimizing **Inputs** in Artificial Neural Networks

Input (data) x → Computation $f(\cdot|\theta)$ → Output (**prediction**) $\hat{y} = f(x|\theta)$

Parameter

1-hot vectors?

	A	T	C	C	G	A	T	C	T	A	A
a	1	0	0	0	0	1	0	0	0	1	1
c	0	0	1	1	0	0	0	1	0	0	0
g	0	0	0	0	1	0	1	0	0	0	0
t	0	1	0	0	0	0	0	0	1	0	0

$$w_{ji}^{(t+1)} \longleftarrow w_{ji}^{(t)} + \eta \frac{\partial \mathcal{L}}{\partial w_{ji}}$$

$$\frac{\partial \mathcal{L}}{\partial w_{ji}^r} = \frac{\partial \mathcal{L}}{\partial a_j^r} \frac{\partial a_j^r}{\partial w_{ji}^r} = \frac{\partial \mathcal{L}}{\partial a_j^r} z_i^{r-1} = \delta_j^r z_i^{r-1}$$

$$x^{(t+1)} \longleftarrow x^{(t)} + \eta \frac{\partial \mathcal{L}}{\partial x}$$

$$\frac{\partial \mathcal{L}}{\partial x_i} = \frac{\partial \mathcal{L}}{\partial a_j^1} \frac{\partial a_j^1}{\partial x_i} = \sum_j \delta_j^1 \partial_{x_i} \sum_i w_{ji}^1 x_i = \sum_j \delta_j^1 w_{ji}^1$$

Optimizing **Inputs** in Artificial Neural Networks

Input (data) → Computation → Output (**prediction**)

$$x \longrightarrow f(\cdot|\theta) \longrightarrow \hat{y} = f(x|\theta)$$

Parameter

1-hot vectors?

a	0.35	0.25	0.31	0.35
c	0.28	0.38	0.32	0.28
g	0.24	0.14	0.21	0.24
t	0.13	0.23	0.16	0.13

During/after the iteration,
The values are no longer'1-hot'

$$w_{ji}^{(t+1)} \longleftarrow w_{ji}^{(t)} + \eta \frac{\partial \mathcal{L}}{\partial w_{ji}}$$

$$\frac{\partial \mathcal{L}}{\partial w_{ji}^r} = \frac{\partial \mathcal{L}}{\partial a_j^r} \frac{\partial a_j^r}{\partial w_{ji}^r} = \frac{\partial \mathcal{L}}{\partial a_j^r} z_i^{r-1} = \delta_j^r z_i^{r-1}$$

$$x^{(t+1)} \longleftarrow x^{(t)} + \eta \frac{\partial \mathcal{L}}{\partial x}$$

$$\frac{\partial \mathcal{L}}{\partial x_i} = \frac{\partial \mathcal{L}}{\partial a_j^1} \frac{\partial a_j^1}{\partial x_i} = \sum_j \delta_j^1 \partial_{x_i} \sum_i w_{ji}^1 x_i = \sum_j \delta_j^1 w_{ji}^1$$

配列微分による配列最適化・設計

Linder J & Seelig G, Fast activation maximization for molecular sequence design.

BMC Bioinformatics 22, 510 (2021)

Gradient Ascent

$$\boldsymbol{x}^{(t+1)} \leftarrow \boldsymbol{x}^{(t)} + \eta \cdot \nabla_{\boldsymbol{x}} \mathcal{P}(\boldsymbol{x})$$

1-hot vectors

$$\boldsymbol{x} \in \{0,1\}^{N \times M}$$

nucleotide logits

$$\sigma(\boldsymbol{l})_{ij} = \frac{e^{\boldsymbol{l}_{ij}}}{\sum_{k=1}^{4} e^{\boldsymbol{l}_{ik}}}$$

gradient

$$\frac{\partial \sigma(\boldsymbol{l})_{ij}}{\partial \boldsymbol{l}_{ik}} = \sigma(\boldsymbol{l})_{ik} \cdot \left(\mathbb{1}_{(j=k)} - \sigma(\boldsymbol{l})_{ij}\right)$$

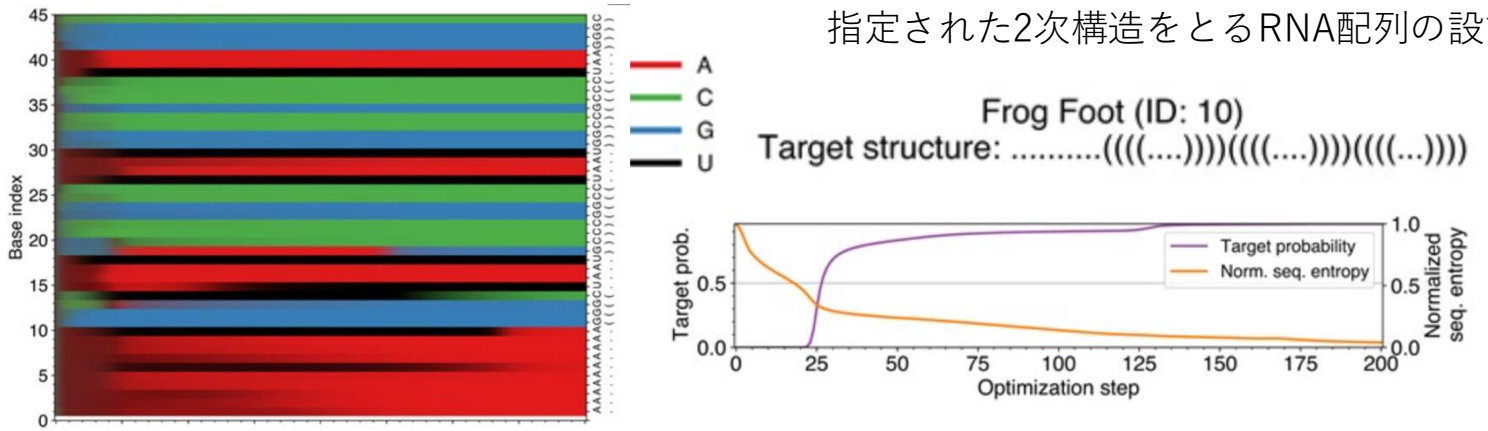
Matthies MC et al., **Differentiable partition function** calculation for RNA. ***Nucleic Acids Research,*** 52:3, e14 (2024)

逆フォールディング問題

Structure-sequence partition function

$$Z_{\mathrm{ss}}(x) = \sum_{s \in S} \sum_{q \in Q} x_q \cdot e^{-\beta E_{q,s}}$$

指定された2次構造をとるRNA配列の設計



Probability distribution and derivative partition function of RNA 2D structures

seconary, not 2 dimensional

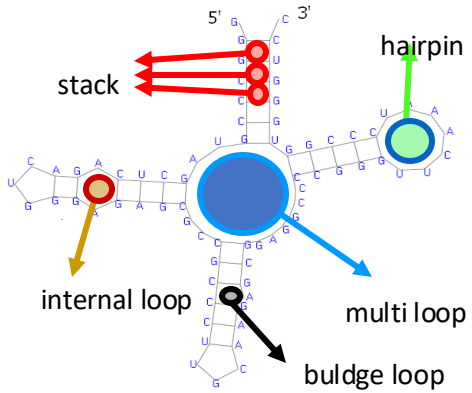
Probability of 2D structure $\boldsymbol{\sigma}^*$ for RNA sequence $\boldsymbol{x}$ is:

$$Prob(\sigma^*|x) = \frac{1}{Z(x)} e^{-E(\sigma^*,x)/kT}$$

McCaskill DP algorithm is mathematically differentiable

The **partition function** (分配関数)

$$Z(x) = \sum_{\xi \in \Omega} \exp \frac{-E(\xi, x)}{RT}$$



cf. CFG

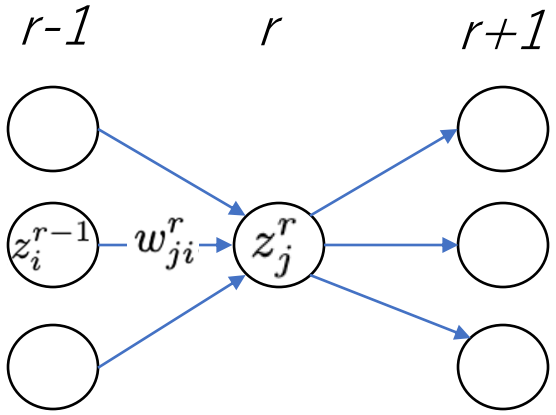
$$\begin{aligned}
Z_{i,j} &= 1.0 + \sum_{i \le k < j} Z_{i,k} Z^1_{k+1,j} \\
Z^1_{i,j} &= \sum_{i < k \le j} Z^b_{i,k} \\
Z^b_{i,j} &= e^{-\beta f_1(i,j)} \\
&\quad + \sum_{i<k<\ell<j} Z^b_{k,\ell} e^{-\beta f_2(i,j,k,\ell)} \\
&\quad + \sum_{i+1<k<j} Z^m_{i+1,k-1} Z^{m1}_{k,j-1} e^{-\beta f_3(i,j)} \\
Z^m_{i,j} &= \sum_{i \le k < j} \left[e^{-\beta f_4(i,k-1)} + Z^m_{i,k-1} \right] Z^{m1}_{k,j} e^{-\beta f_5} \\
Z^{m1}_{i,j} &= \sum_{i \le k < j} Z^b_{i,k} e^{-\beta f_4(k+1,j)}
\end{aligned}$$

Mathematical Differentiation

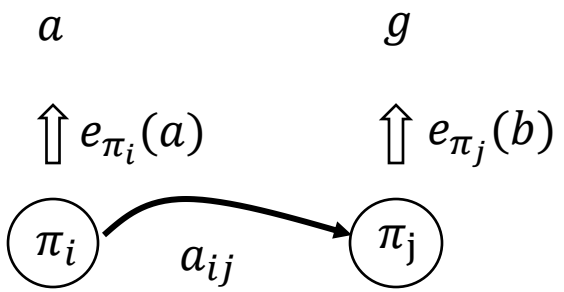
Back propagation for gradient decent is applicable

$$\begin{aligned}
\partial Z_{i,j} &= \sum_{i \le k < j} \partial Z_{i,k} Z^1_{k+1,j} + Z_{i,k} \partial Z^1_{k+1,j} \\
\partial Z^1_{i,j} &= \sum_{i < k \le j} \partial Z^b_{i,k} \\
\partial Z^b_{i,j} &= \partial e^{-\beta f_1(i,j)} \\
&\quad + \sum_{i<k<\ell<j} \left[\partial Z^b_{k,\ell} e^{-\beta f_2(i,j,k,\ell)} + Z^b_{k,\ell} \partial e^{-\beta f_2(i,j,k,\ell)} \right] \\
&\quad + \sum_{i+1<k<j} \Big[\partial Z^m_{i+1,k-1} Z^{m1}_{k,j-1} e^{-\beta f_3(i,j)} \\
&\quad + Z^m_{i+1,k-1} \partial Z^{m1}_{k,j-1} e^{-\beta f_3(i,j)} + Z^m_{i+1,k-1} Z^{m1}_{k,j-1} \partial e^{-\beta f_3(i,j)} \Big] \\
\partial Z^m_{i,j} &= \sum_{i \le k < j} \left[\partial e^{-\beta f_4(i,k-1)} + \partial Z^m_{i,k-1} \right] Z^{m1}_{k,j} e^{-\beta f_5} \\
&\quad + \sum_{i \le k < j} \left[e^{-\beta f_4(i,k-1)} + Z^m_{i,k-1} \right] \partial Z^{m1}_{k,j} e^{-\beta f_5} \\
&\quad + \sum_{i \le k < j} \left[e^{-\beta f_4(i,k-1)} + Z^m_{i,k-1} \right] Z^{m1}_{k,j} \partial e^{-\beta f_5} \\
\partial Z^{m1}_{i,j} &= \sum_{i \le k < j} \left[\partial Z^b_{i,k} \partial e^{-\beta f_4(k+1,j)} + Z^b_{i,k} \partial e^{-\beta f_4(k+1,j)} \right]
\end{aligned}$$

Iterations of Neural Network and HMM



Neural Network



HMM

Total probability of the sequence x calculated by forward algorithm

$$a_j^r = \sum_i w_{ji}^r z_i^{r-1}$$

$$z_j^r = h(a_j^r)$$

Same shape of formula, substituting $\boldsymbol{h(x) = x}$

$$f_j^t = \sum_i f_i^{t-1} w_{ji}^t$$

$$w_{ji}^t \equiv a_{ij} e_j(x_t)$$

Initialization:

$$f_0^0 = 1, \quad f_j^0 = 0 (j \neq 0)$$

Iteration:

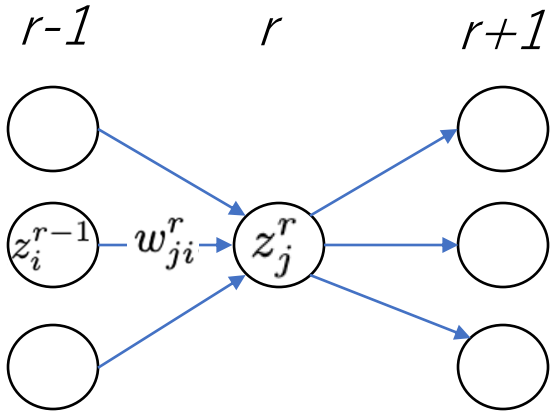
for $t = 0, \ldots, T-1$

$$f_j^{t+1} = \sum_i f_i^t a_{ij} e_j(x_{t+1})$$

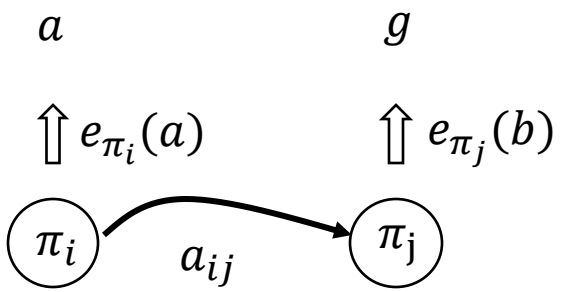
Termination:

$$p(x) = \sum_j f_k^L$$

Iterations of Neural Network and HMM



Neural Network



HMM

Back-propagation

$$\delta_j^r \equiv \frac{\partial \mathcal{L}}{\partial a_j^r} = \sum_k \frac{\partial z_j^r}{\partial a_j^r} \frac{\partial a_k^{r+1}}{\partial z_j^r} \frac{\partial \mathcal{L}}{\partial a_k^{r+1}}$$

$$= h'(a_j^r) \sum_k w_{kj}^{r+1} \delta_k^{r+1}$$

$$\delta_j^L = \frac{\partial \mathcal{L}}{\partial a_j^L} = \frac{\partial \mathcal{L}(\hat{y} - y)}{\partial \hat{y}}$$

$$f_j^{t+1} = \sum_i f_i^t a_{ij} e_j(x_{t+1})$$

$$f_j^t = \sum_i f_i^{t-1} w_{ji}^t \qquad w_{ji}^t \equiv a_{ij} e_j(x_t)$$

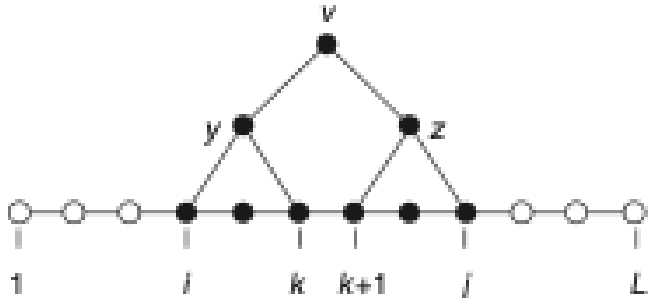
$$\delta_j^t \equiv \frac{\partial p(x)}{\partial f_j^t} = \sum_k \frac{\partial p(x)}{\partial f_k^{t+1}} \frac{\partial f_k^{t+1}}{\partial f_j^r} = \sum_k \delta_k^{t+1} w_{kj}^{t+1}$$

Backward algorithm of HMM is a backpropagation of forward algorithm

Jason Eisner. Inside-Outside and Forward-Backward Algorithms Are Just Backprop (tutorial paper). In Proceedings of the Workshop on Structured Prediction for NLP, pages 1–17

Iterations of SCFG and HMM

SCFG（確率文脈自由文法）



$$\alpha(i,j,v) = \sum_{k=i}^{j-1} \sum_{y,z \in \mathcal{C}} \alpha(i,k,y)\alpha(k+1,j,z)t(v \to yz)$$

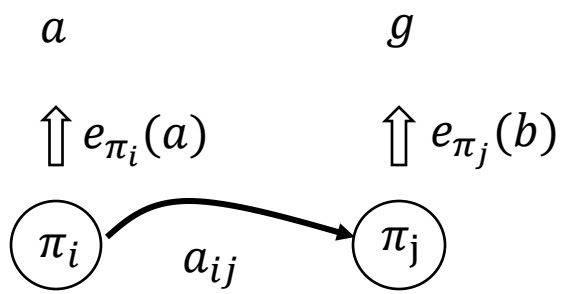
$$\beta(i,j,v) \equiv \frac{\partial P(x|v \to x_i,\ldots,x_j)}{\partial \alpha(i,j,v)}$$

$$= \sum_{k=j+1}^{L} \sum_{y \in \mathcal{M}} \left[\frac{\partial P(x|v \to x_i,\ldots,x_j)}{\partial \alpha(i,k,y)} \frac{\partial \alpha(i,k,y)}{\partial \alpha(i,j,v)} \right]$$

$$+ \sum_{k=1}^{i-1} \sum_{y \in \mathcal{M}} \left[\frac{\partial P(x|v \to x_i,\ldots,x_j)}{\partial \alpha(i,j,y)} \frac{\partial \alpha(k,j,y)}{\partial \alpha(i,j,v)} \right]$$

$$= \sum_{k=j+1}^{L} \sum_{y \in \mathcal{M}} \beta(i,k,y) \sum_{z \in \mathcal{M}} \alpha(j+1,k,z)t(y \to vz)$$

$$+ \sum_{k=1}^{i-1} \sum_{y \in \mathcal{M}} \beta(k,j,y) \sum_{z \in \mathcal{M}} \alpha(k,i-1,z)t(y \to zv)$$



HMM

$$f_j^{t+1} = \sum_i f_i^t a_{ij} e_j(x_{t+1})$$

$$f_j^t = \sum_i f_i^{t-1} w_{ji}^t \qquad w_{ji}^t \equiv a_{ij} e_j(x_t)$$

$$\delta_j^t \equiv \frac{\partial p(x)}{\partial f_j^t} = \sum_k \frac{\partial p(x)}{\partial f_k^{t+1}} \frac{\partial f_k^{t+1}}{\partial f_j^r} = \sum_k \delta_k^{t+1} w_{kj}^{t+1}$$

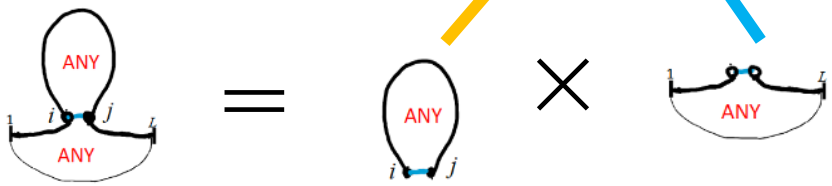
Backward algorithm of HMM is a backpropagation of forward algorithm

Jason Eisner. Inside-Outside and Forward-Backward Algorithms Are Just Backprop (tutorial paper). In Proceedings of the Workshop on Structured Prediction for NLP, pages 1–17

Base Pairing Probability (BPP)

$$P_{ij}^{(\mathrm{bp})} = \frac{1}{Z(X)} Z_{ij}^{b} W_{ij}^{b}$$

$$P_{ij}^{(\mathrm{bp})} \equiv Prob((i,j) \in \sigma^{*}|x) = \sum_{\sigma|(i,j)\in\sigma^{*}} P(\sigma|x)$$



Differentiable !

$$Z_{i,j}^{b} = e^{-\beta f_1(i,j)} + \sum_{i<k<\ell<j} Z_{k,\ell}^{b} e^{-\beta f_2(i,j,k,\ell)} + \sum_{i+1<k<j} Z_{i+1,k-1}^{m} Z_{k,j-1}^{m1} e^{-\beta f_3(i,j)}$$

$$Z_{i,j}^{m} = \sum_{i\le k<j} \left[e^{-\beta f_4(i,k-1)} + Z_{i,k-1}^{m} \right] Z_{k,j}^{m1} e^{-\beta f_5}$$

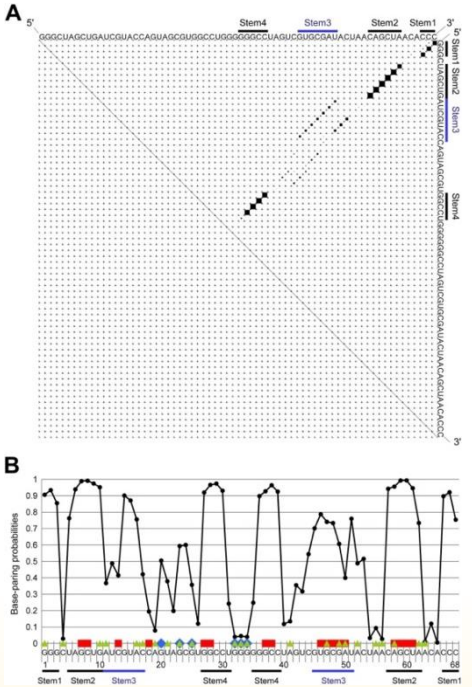
$$Z_{i,j}^{m1} = \sum_{i\le k<j} Z_{i,k}^{b} e^{-\beta f_4(k+1,j)}$$

Inside partition function

$$W_{ij}^{b} = Z_{1,i-1} Z_{j+1,L} + \sum_{k<i<j<\ell} W_{k,\ell}^{b} e^{-\beta f_2(k,i,j,\ell)} + \sum_{k<i<j<\ell} W_{k,\ell}^{b} e^{-\beta f_3} \left[\begin{array}{l} Z_{k+1,i-1}^{m} e^{-\beta f_1(j,\ell)} \\ + Z_{j+1,\ell-1}^{m} e^{-\beta f_1(k,i)} \\ + Z_{k+1,i-1}^{m} Z_{j+1,\ell-1}^{m} \end{array} \right]$$

Outside partition function

Base-Pairing Probability



Optimizing **Parameters** in Computing (Prediction) System

Input (data) x → Computation $f(\cdot|\theta)$ → Output (**prediction**) $\hat{y} = f(x|\theta)$

Parameter

Gradient decent on parameters

$$\theta^{(t+1)} \longleftarrow \theta^{(t)} - \eta \frac{\partial \mathcal{L}}{\partial \theta}$$

loss function $\mathcal{L}(y, \hat{y})$

if $\mathcal{L}(y, \hat{y}) = (y - \hat{y})^2$

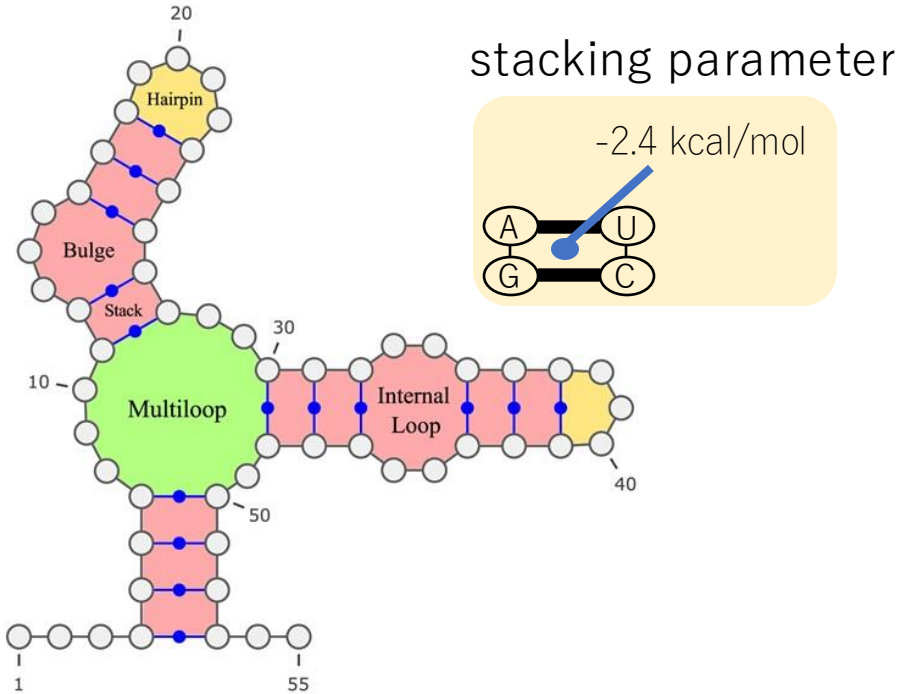
$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial \mathcal{L}}{\partial f(x,\theta)} \frac{\partial f(x,\theta)}{\partial \theta} = 2\{\hat{y} - y\}\partial_\theta f(x,\theta)$$

$$\partial_\theta \equiv \frac{\partial}{\partial \theta}$$

Estimation of energy parameters of RNA 2D structures

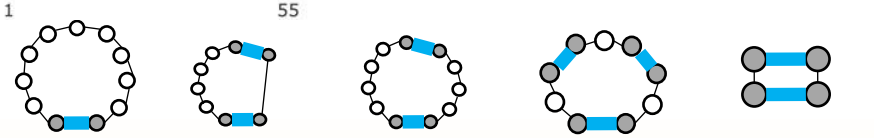
Nearest neighbor energy model of RNA 2D structure and determination of the energy parameters

Nearest neighbor energy parameters



stacking parameter

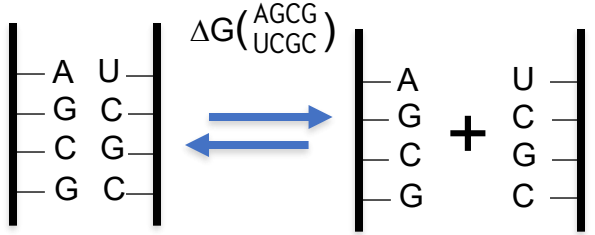
-2.4 kcal/mol



hairpin loop bulge loop Internal loop multi loop stacking

Circle: nucleotide, Blue line: base pair interaction, Black line: phosphate backbone

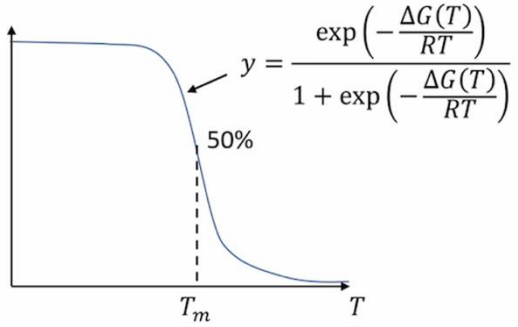
Determination of energy parameters



hybridizing sequences

no	sequences[a]	experimental[b] $-\Delta G°_{37}$ (kcal/mol)	$-\Delta H°$ (kcal/mol)	$-\Delta S°$ (eu)	T_M[d] (°C)
		Two-State Sequences Used in Re			
1	CCGG	4.55	34.21	95.6	27.2
2	CGCG	3.66	33.31	95.6	19.3
3	GCGC	4.61	30.48	83.4	26.6
4	GGCC	5.37	35.79	98.1	34.3
5	ACGCA/	4.97	45.40	130.4	29.4
6	AGCGA/	5.05	46.31	133.0	30.2
7	CACAG/	4.70	40.20	115.4	24.5
8	GCACG/	6.17	45.31	126.2	37.5
9	GCUCG/	6.14	43.38	120.1	37.2
10	ACCGGUp	8.51	59.78	164.5	53.9
11	AGCGCU	7.99	50.07	135.7	52.0
12	AGGCCU	8.36	48.19	128.4	55.3
13	CACGUG	6.59	50.31	141.0	42.8
14	CAGCUGp	6.68	51.55	144.7	43.1
15	CCAUGG	7.30	56.93	159.9	46.4
16	CCGCGG	9.84	60.79	164.3	59.8
17	CCUAGG	7.80	54.10	149.1	50.0
18	CGCGCGp	9.12	54.51	146.4	57.8
19	CGGCCGp	9.90	54.12	142.6	63.2
20	CUGCAGp	7.11	55.41	155.7	45.3
21	GACGUC	7.35	58.06	163.5	46.2
22	GAGAGA/	6.95	62.05	178.1	40.6
23	GAGCUC	7.98	62.30	175.3	48.7
24	GAGGAG/	8.50	55.70	152.2	50.9
25	GCAACG/	7.01	50.57	140.5	42.6
26	GCAUCG/	7.26	51.89	143.9	44.1
27	GCAUGC	7.38	62.34	177.2	45.7
28	GCCGCG/	10.88	59.69	157.4	63.9
29	GCCGGCp	11.20	62.72	166.0	67.2
30	GCGCCG/	10.91	57.85	151.3	65.2
31	GCGCGCp	10.62	65.98	178.5	62.1
32	GCGCGG/	11.38	71.16	192.7	61.9
33	GCGGCG/	10.40	58.50	155.0	61.8
34	GCGUCG/	8.76	52.38	140.6	53.7
35	GCUACG/	7.56	58.02	162.7	45.0
36	GCUAGC	7.92	59.13	165.1	49.3
37	GGAUCC	7.44	53.70	149.1	47.6
38	GGCGCCp	11.33	67.78	182.0	65.2
39	GGCGCG/	10.78	63.53	170.1	61.6
40	GGUACC	7.35	54.90	153.4	46.6
41	GUCGAC	7.09	53.63	150.1	45.3
42	GUGCAC	7.65	59.61	167.5	47.7
43	GUGGUG/	7.67	48.84	132.7	47.4
44	GUGUCG/	7.18	50.88	140.9	43.7
45	UCAUGA	4.31	41.89	121.2	27.2
46	UCCGGAp	7.79	51.92	142.3	50.1

UV absorbance melting curve analysis



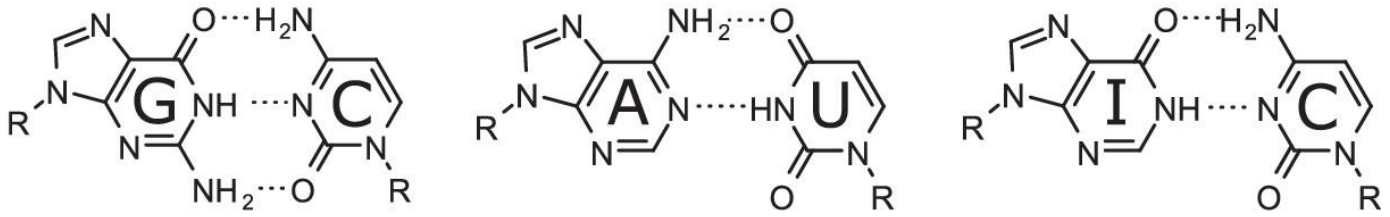
System of linear equations

$$\Delta G\left(\begin{smallmatrix}AGCG\\UCGC\end{smallmatrix}\right) \simeq \Delta G\left(\begin{smallmatrix}3'A\\5'U\end{smallmatrix}\right) + \Delta G\left(\begin{smallmatrix}AG\\UC\end{smallmatrix}\right) + \Delta G\left(\begin{smallmatrix}GC\\CG\end{smallmatrix}\right) + \Delta G\left(\begin{smallmatrix}CG\\GC\end{smallmatrix}\right) + \Delta G\left(\begin{smallmatrix}G&5'\\C&3'\end{smallmatrix}\right)$$

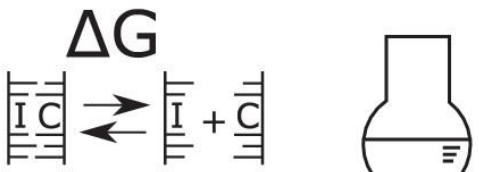
Experiments are costly and time consuming

Marco C Matthies, Ryan Krueger, Andrew E Torda, Max Ward, Differentiable partition function calculation for RNA, *Nucleic Acids Research*, Volume 52, Issue 3, 9 February 2024, Page e14,

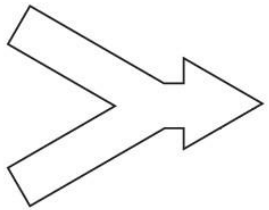
Free-Energy Calculation of Ribonucleic Inosines and Its Application to Nearest-Neighbor Parameters



UV melting experiments



MD simulations

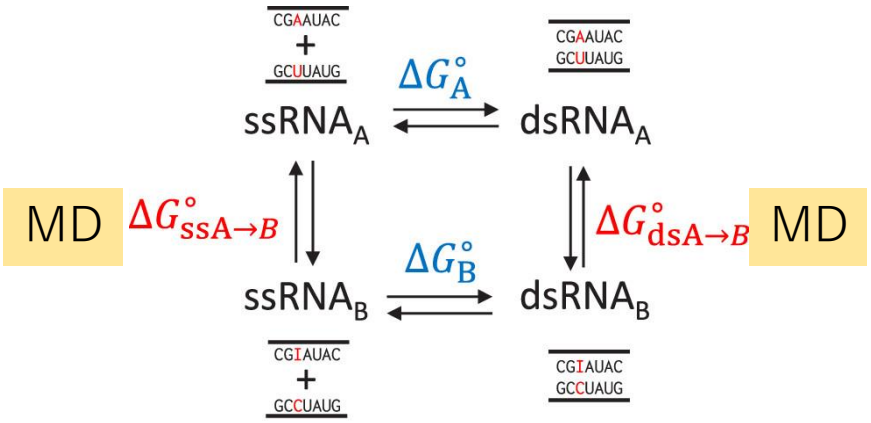


NN parameters

$\left|\begin{matrix}GC\\IC\end{matrix}\right| = x.xx$

$\left|\begin{matrix}AU\\IC\end{matrix}\right| = x.xx$

...



$$\Delta\Delta G^\circ_{AB} = \Delta G^\circ_{dsA\to B} - \Delta G^\circ_{ssA\to B} = \Delta G^\circ_B - \Delta G^\circ_A$$

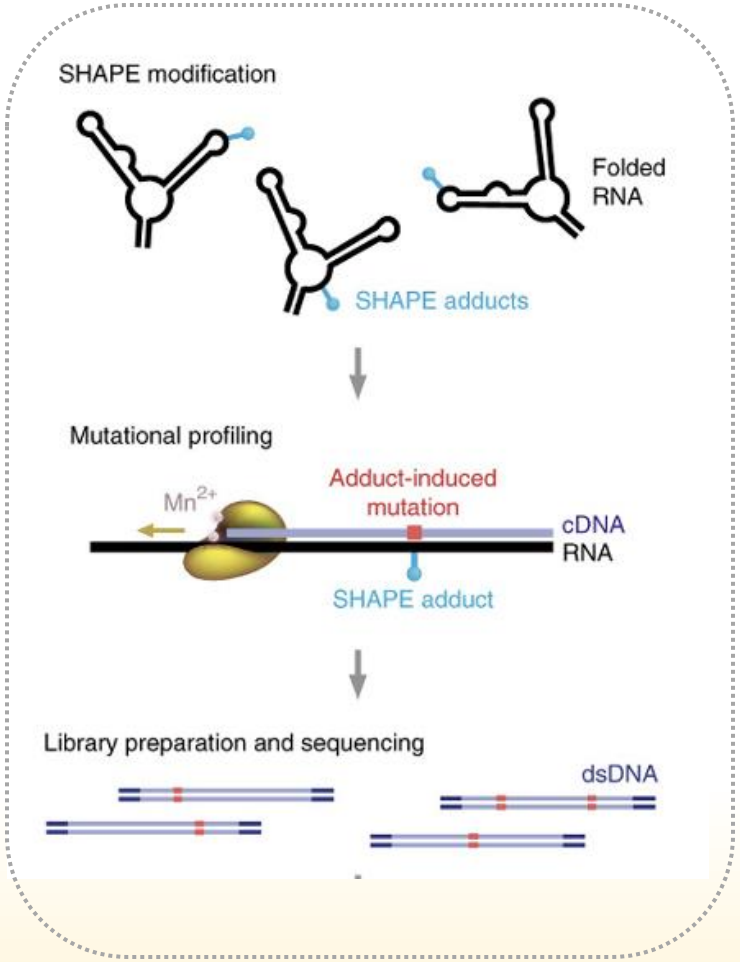
Combination of MD and experiments
can ***reduce*** the experimental ***costs***
for parameter determination,
but still time consuming!

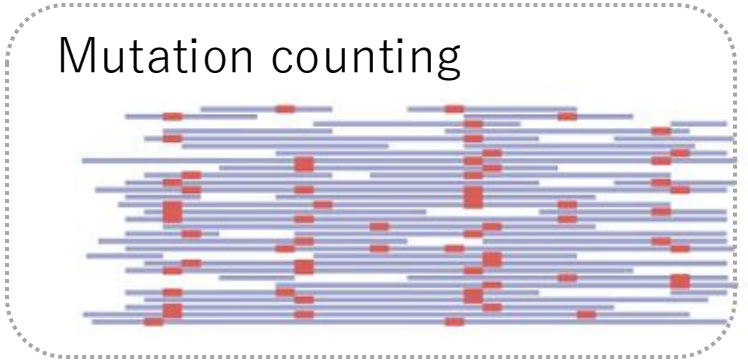
Identified energy parameters: Inosine & m^{6}A

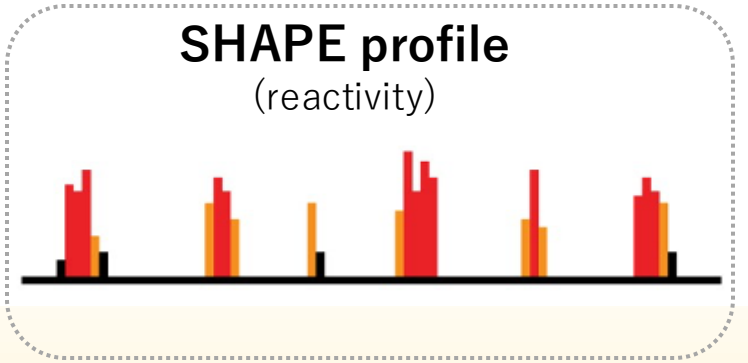
Sakuraba S et al., J. Chem. Theory Comput. 2020, 16, 9, 5923–5935. DOI: (10.1021/acs.jctc.0c00270)

Chemical probing for RNA 2D structure analysis

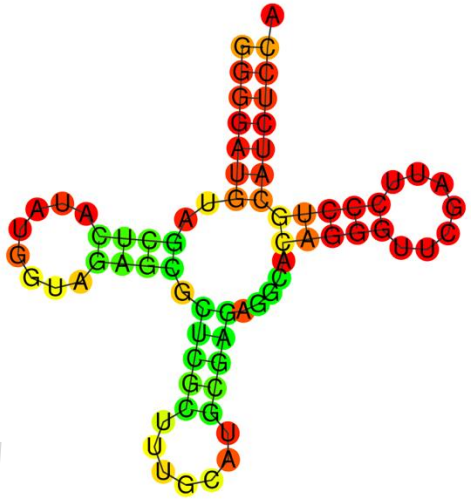
SHAPE-MaP







\+

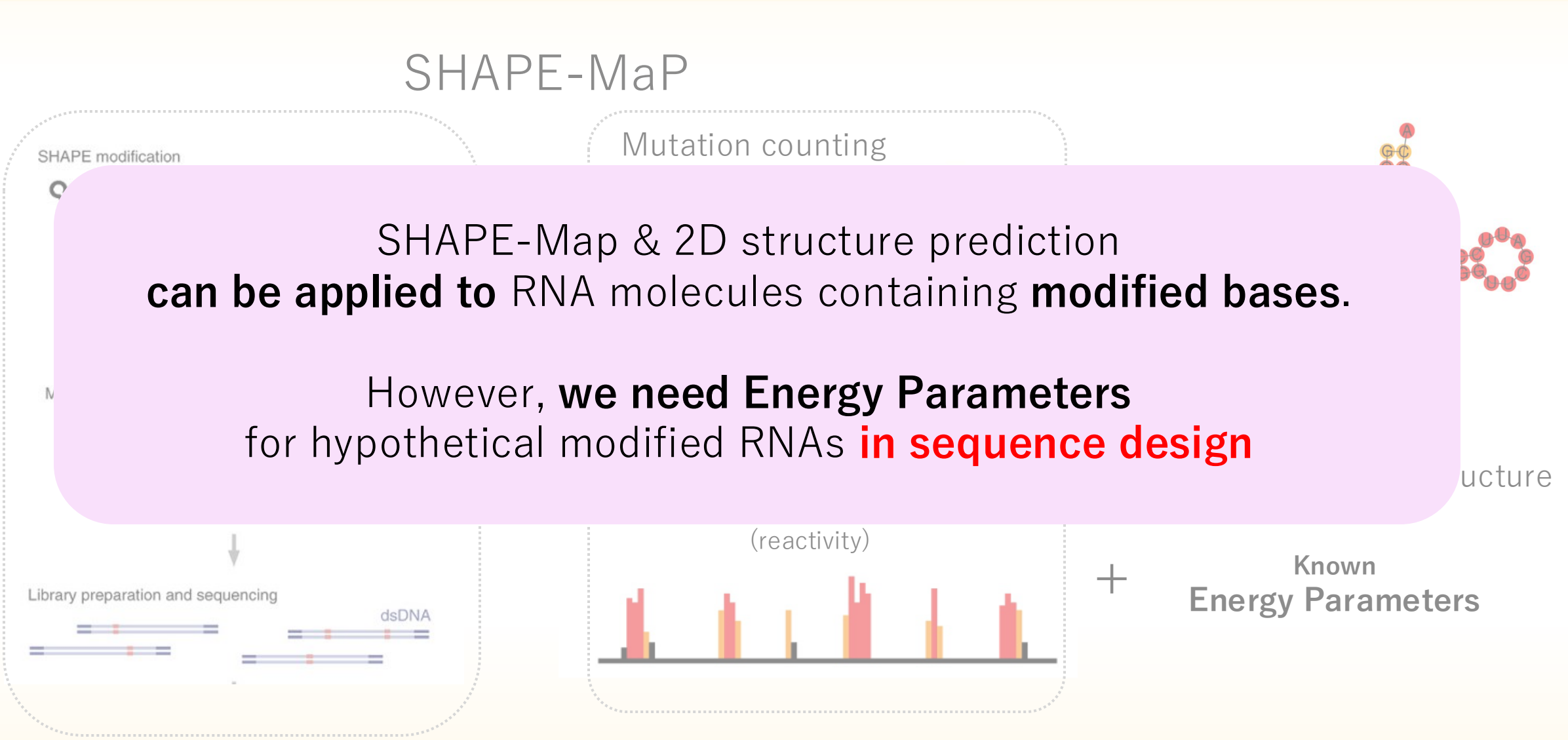


2D structure

Known
Energy Parameters

Siegfried, N., Busan, S., Rice, G. *et al.* RNA motif discovery by SHAPE and mutational profiling (**SHAPE-MaP**). *Nat Methods* **11**, 959–965 (2014).

Chemical probing for RNA 2D structure analysis



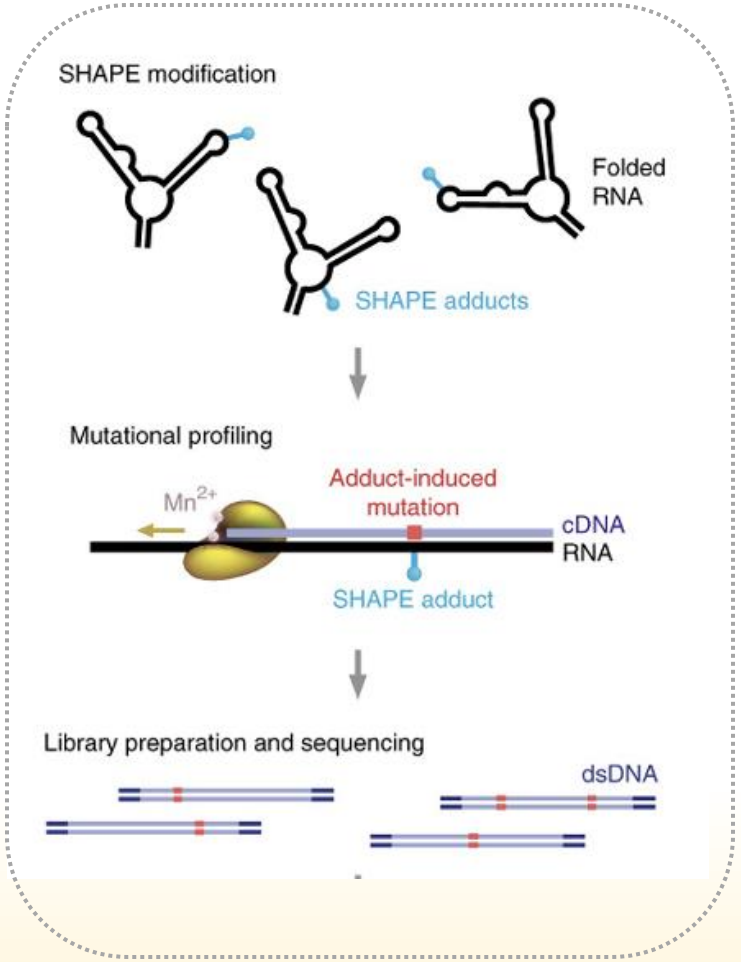
SHAPE-Map & 2D structure prediction
can be applied to RNA molecules containing **modified bases.**

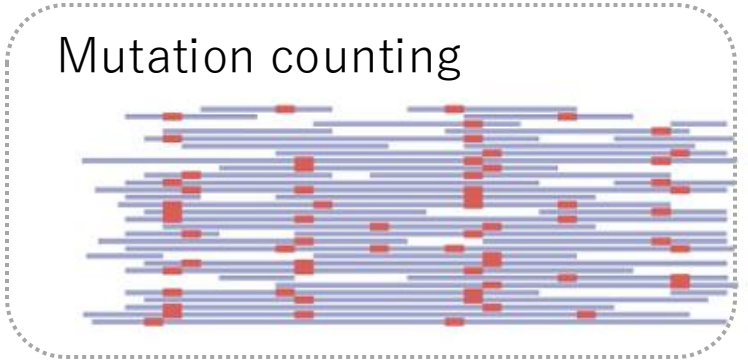
However, **we need Energy Parameters**
for hypothetical modified RNAs **in sequence design**

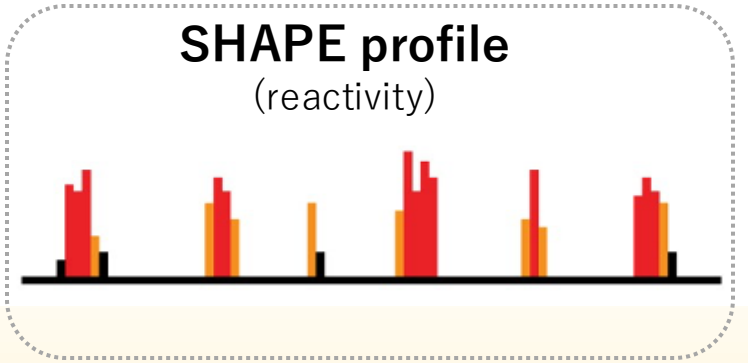
Siegfried, N., Busan, S., Rice, G. *et al.* RNA motif discovery by SHAPE and mutational profiling (**SHAPE-MaP**). *Nat Methods* **11**, 959–965 (2014).

Chemical probing for RNA 2D structure analysis

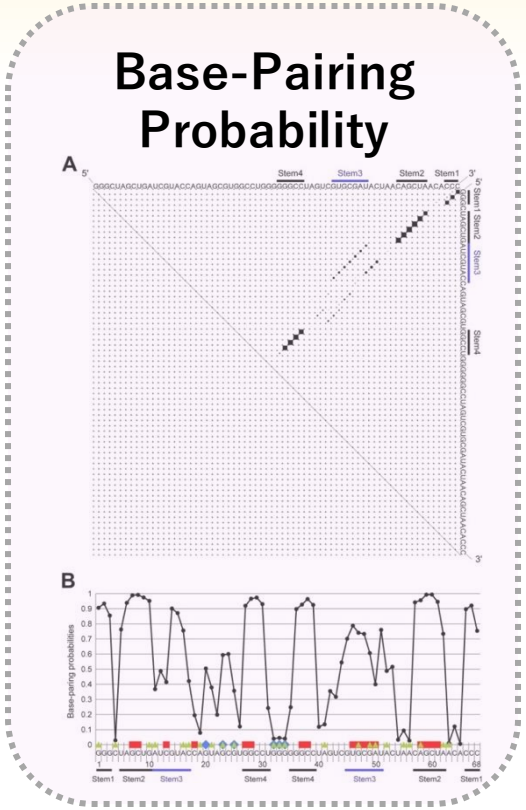
SHAPE-MaP







+



Known
Energy Parameters

Siegfried, N., Busan, S., Rice, G. *et al.* RNA motif discovery by SHAPE and mutational profiling (**SHAPE-MaP**). *Nat Methods* **11**, 959–965 (2014).

A method for estimating Energy Parameters of RNAs by Differentiating Base-Pairing Probabilities

Yamamura K et al., in revision

RNA sequences containing m6A

McCaskill Algorithm

Calculated BPPs

Derivative

SHAPE-MaP based BPPs

$$\partial Z_{i,j} = \sum_{i<h<j} \partial Z_{i,h} Z^1_{h+1,j} + \sum_{i<h<j} Z_{i,h} \quad \partial Z^1_{h+1,j}$$

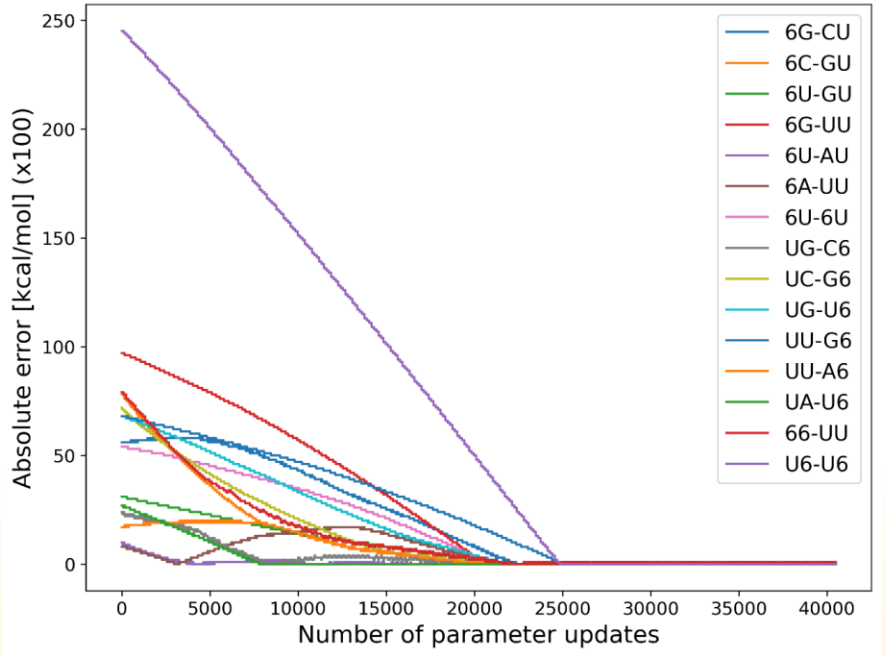
$$\partial Z^1_{i,j} = \sum_{i<h\leq j} \partial Z^b_{i,h}$$

$$\partial Z^b_{i,j} = -\beta\ \partial f_1(i,j)\, e^{-\beta f_1(i,j)} + \sum_{i<h<j-1} \sum_{h<l<j} \left(\partial Z^b_{h,l}\, e^{-\beta f_2(i,j,h,l)} + Z^b_{h,l}\,(-\beta)\,\partial f_2(i,j,h,l) e^{-\beta f_2(i,j,h,l)}\right) + \sum_{i<h<j} \left(\partial Z^m_{i+1,h-1} Z^{m1}_{h,j-1}\, e^{-\beta f_3} + Z^m_{i+1,h-1}\, \partial Z^{m1}_{h,j-1}\, e^{-\beta f_3} + Z^m_{i+1,h-1} Z^{m1}_{h,j-1}\,(-\beta)\ \partial f_3\, e^{-\beta f_3}\right)$$

$$\partial Z^m_{i,j} = \sum_{i<h<j} \left[\left(-\beta\,\partial f_4 e^{-\beta f_4} + \partial Z^m_{i,h-1}\right) Z^{m1}_{h,j} e^{-\beta f_5} + \left(e^{-\beta f_4} + Z^m_{i,h-1}\right) \partial Z^{m1}_{h,j} e^{-\beta f_5} + \left(e^{-\beta f_4} + Z^m_{i,h-1}\right) Z^{m1}_{h,j} (-\beta)\,\partial f_5 e^{-\beta f_5}\right]$$

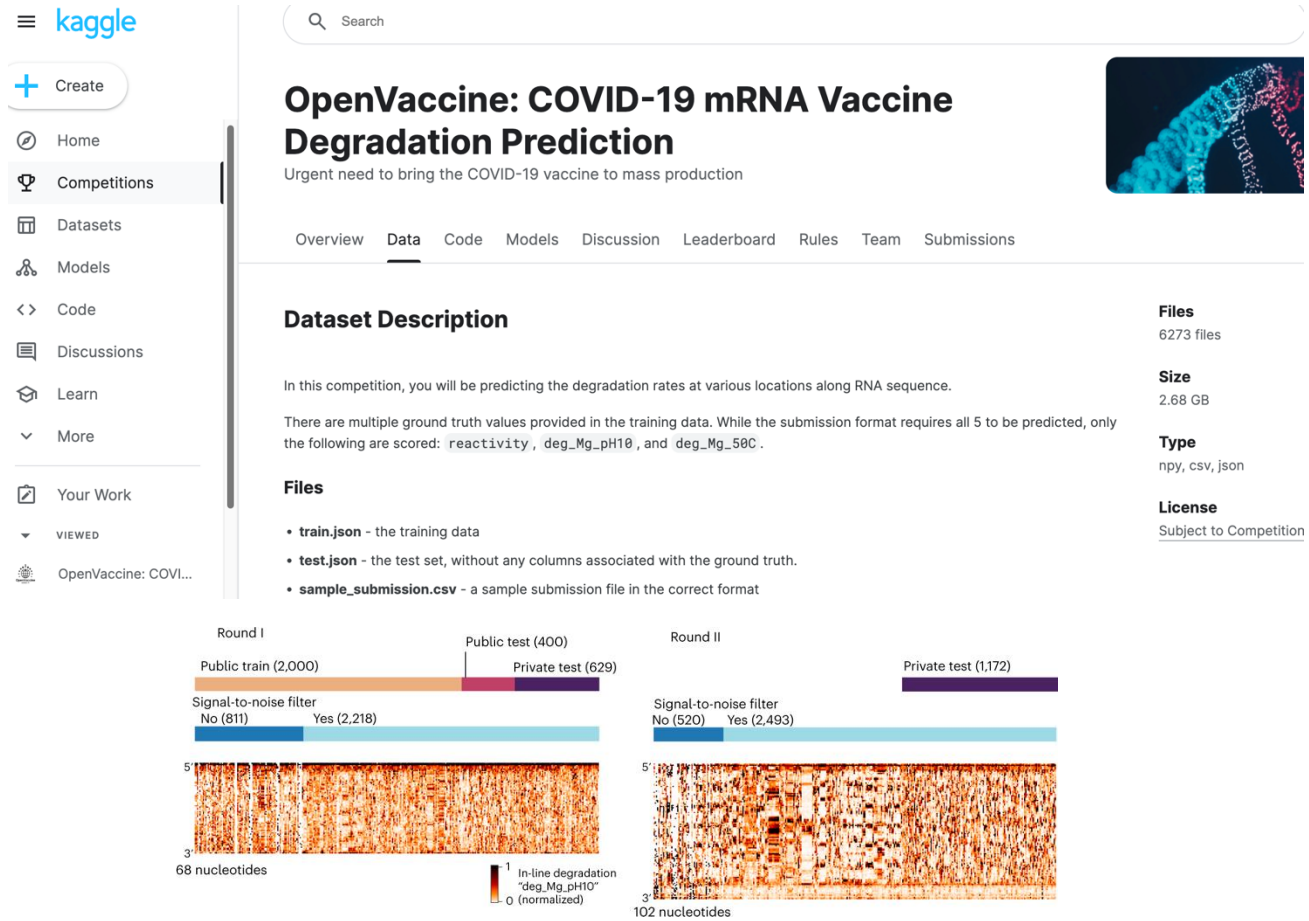
$$\partial Z^{m1}_{i,j} = \sum_{i<h<j} \left((-\beta)\,\partial f_4\, e^{-\beta f_4} Z^b_{i,h} + \partial Z^b_{i,h}\, e^{-\beta f_4}\right)$$

Convergence in gradient decent optimization of m6A stacking parameters



RNA sequence design

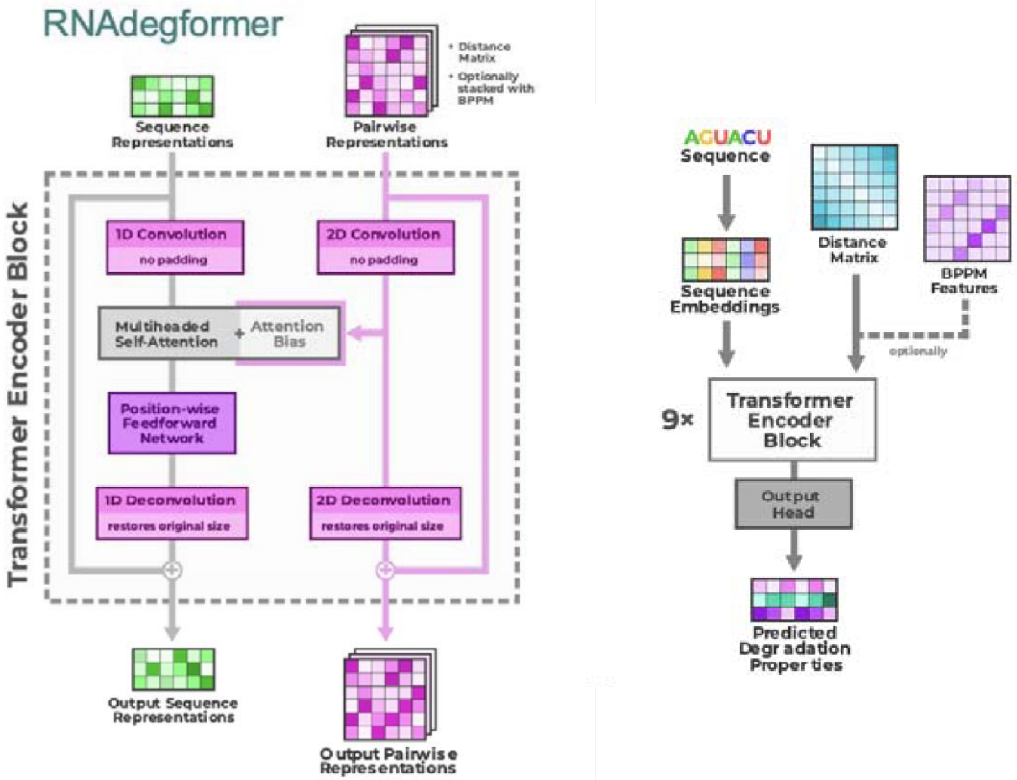
Predicting RNA degradation Kaggle competition



- **Kaggle Competitions** provide an open platform for everyone to compete by applying solutions to real-world problems.
- The **OpenVaccine competition** specifically aims to predict the degradation rates of RNA sequences to stabilize mRNA vaccines.
- Both the data and solutions from this competition are valuable for our RNA degradation prediction tasks.

[1] https://www.kaggle.com/competitions/stanford-covid-vaccine/overview
[2] Wayment-Steele H K, Kladwang W, Watkins A M, et al. Deep learning models for predicting RNA degradation via dual crowdsourcing[J]. Nature Machine Intelligence, 2022, 4(12): 1174-1184.

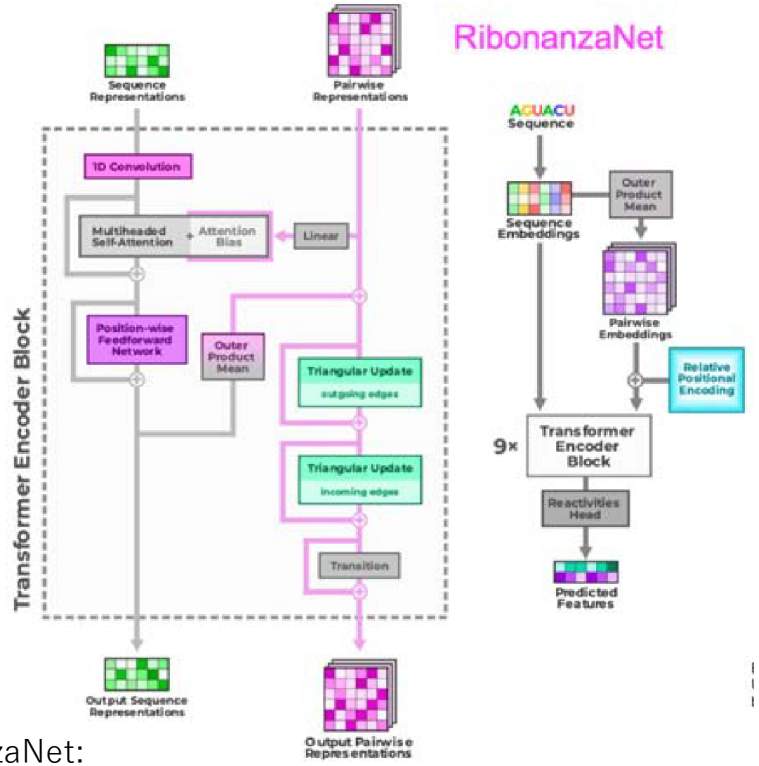
Predicting RNA degradation: Previous studies



RNAdegformer:

- One of best solution in **OpenVaccine.**
- Use base pair probability matrix, and distance matrix on secondary structure to build better self attention of RNA sequence.

He S, Gao B, Sabnis R, et al. RNAdegformer: accurate prediction of mRNA degradation at nucleotide resolution with deep learning[J]. Briefings in Bioinformatics, 2023, 24(1): bbac581.



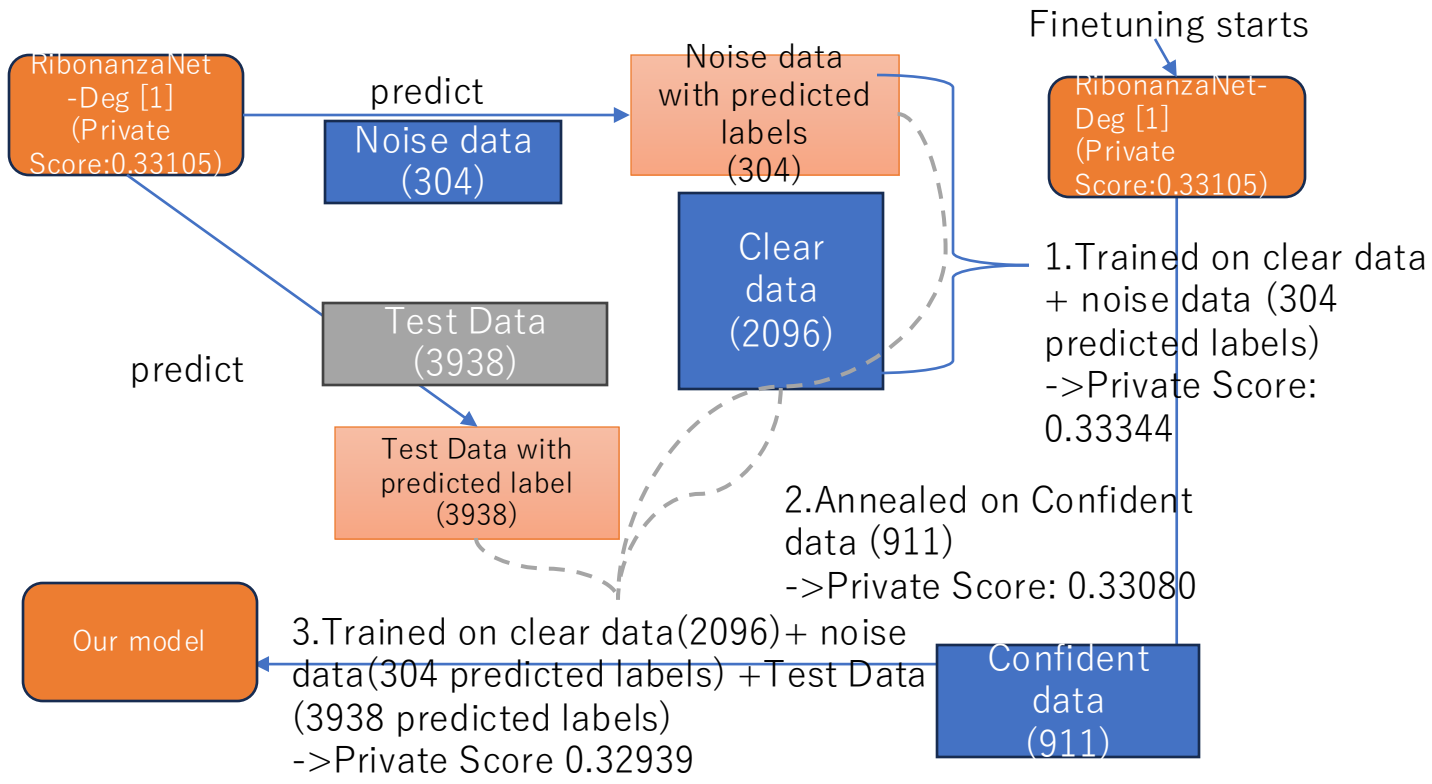
RibonanzaNet:

- An improved version of RNAdegformer, but only trained for **reactivity data**.
- After fine-tuning in OpenVaccine data, it becomes the most accuracy method for **degradation prediction**.

He S, Huang R, Townley J, et al. Ribonanza: deep learning of RNA structure through dual crowdsourcing[J]. bioRxiv, 2024: 2024.02. 24.581671.

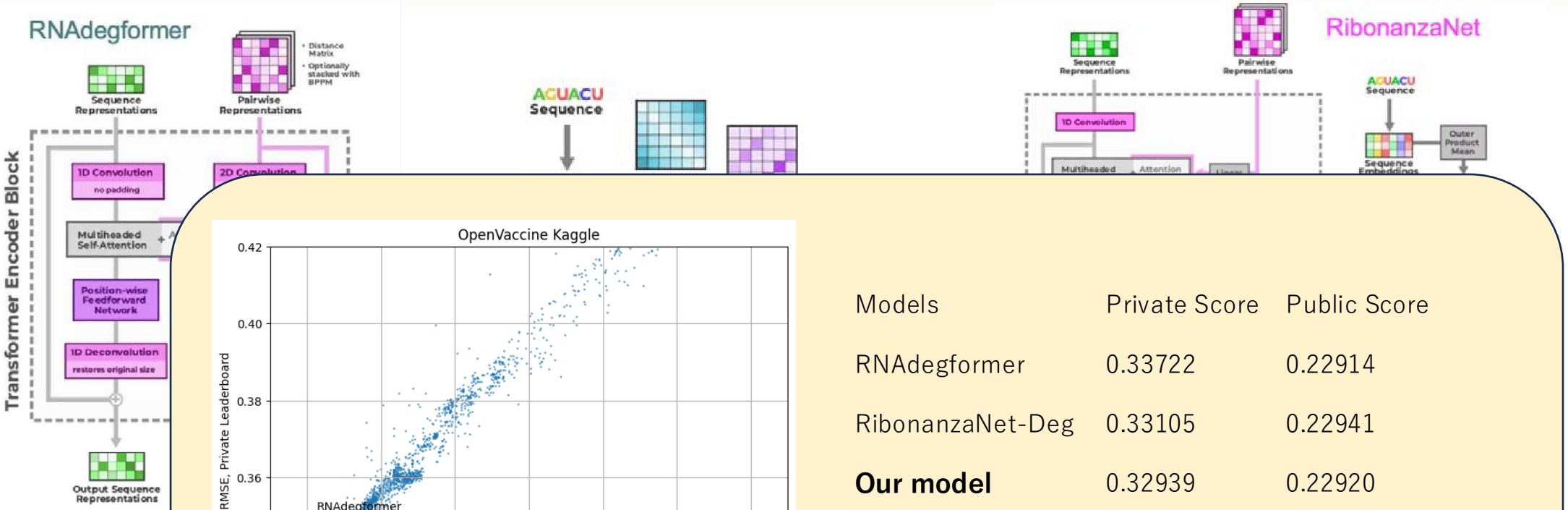
Our semi-supervised finetuning process

Finetune with state-of-the-art models: RibonanzaNet-Deg



[1] He S, Huang R, Townley J, et al. Ribonanza: deep learning of RNA structure through dual crowdsourcing[J]. bioRxiv, 2024: 2024.02. 24.581671.

Predicting RNA degradation by deep learning



Models	Private Score	Public Score
RNAdegformer	0.33722	0.22914
RibonanzaNet-Deg	0.33105	0.22941
Our model	0.32939	0.22920

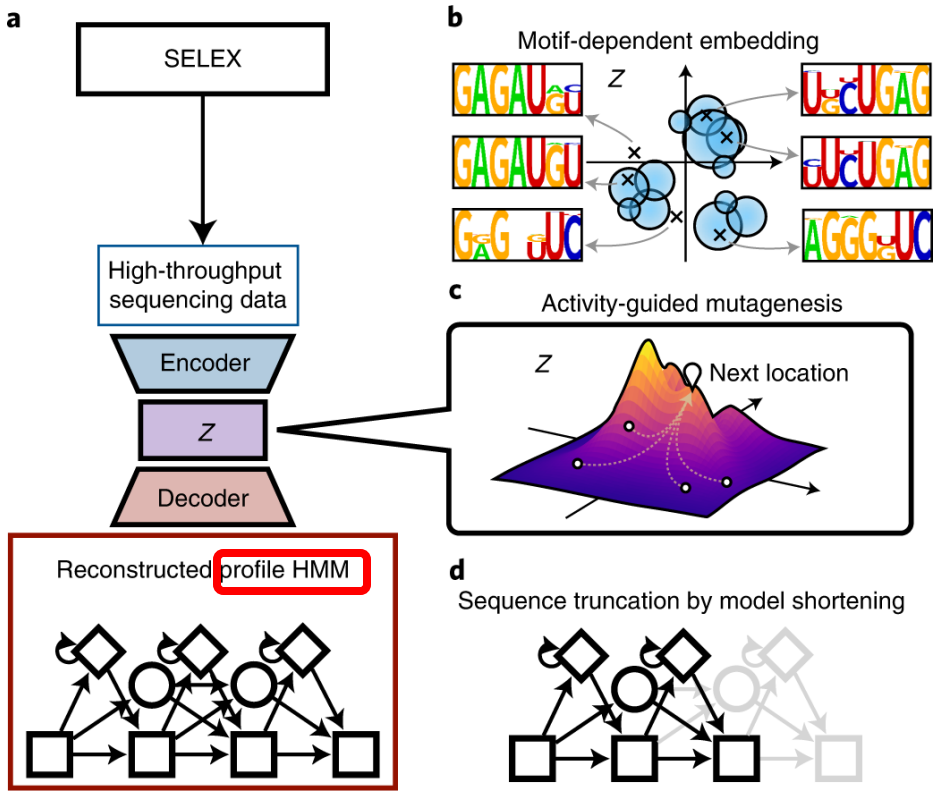
RNAdegformer:
on secondary str
er, but
thod.

He S, Gao B, Sabnis
mRNA degradation at
Briefings in Bioinformatics, 2023, 24(1): bbac581.

of RNA structure
through dual crowdsourcing[J]. bioRxiv, 2024: 2024.02. 24.581671.

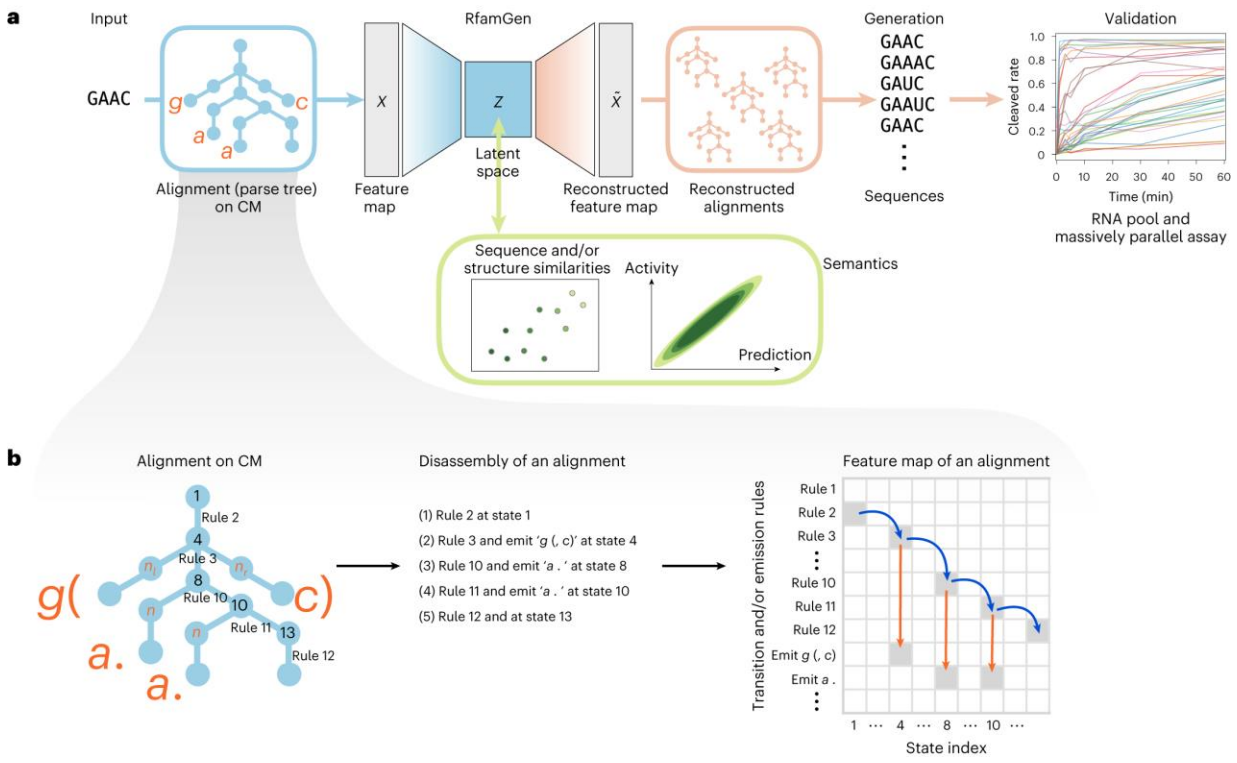
確率生成モデル（確率文法）と深層学習の結合

RaptGen for RNA aptamer



Iwano, N., Adachi, T., Aoki, K. *et al.* Generative aptamer discovery using RaptGen. *Nat Comput Sci* **2**, 378–386 (2022).

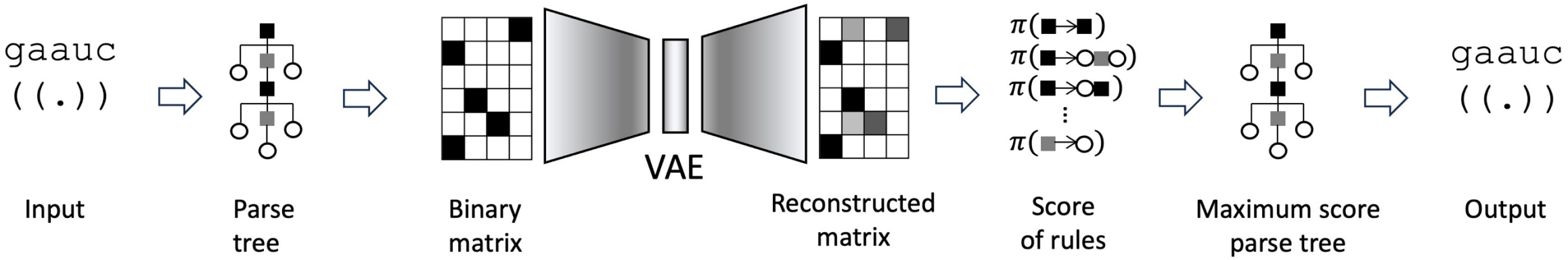
RfamGen: SCFG + VAE



Sumi, S., Hamada, M. & Saito, H. Deep generative design of RNA family sequences. Nat Methods 21, 435–443 (2024)

Deep generative model of RNAs based on variational autoencoder with context free grammar

RNA配列とその2次構造に基づいて構文解析木を作り、バイナリ行列に変換してVAEへの入力とする
VAEはバイナリ行列を復元するように学習する
VAEからサンプリングして生成した（連続値）行列からSCFGを作る
SCFGの最大スコアの構文解析木から配列と二次構造を生成する



Terai G and Asai K, Bioinformatics, accepted

近似文法エンコーディング

- G4 grammar (Dowell and Eddy, 2004, *BMC bioinformatics*, **5**:71)

$$S \rightarrow nS|T|\varepsilon$$
$$T \rightarrow Tn|nS\hat{n}|TnS\hat{n}$$

$$S \rightarrow nS|T|\varepsilon$$
$$T \rightarrow Tn|U|TU$$
$$U \rightarrow nS\hat{n}$$

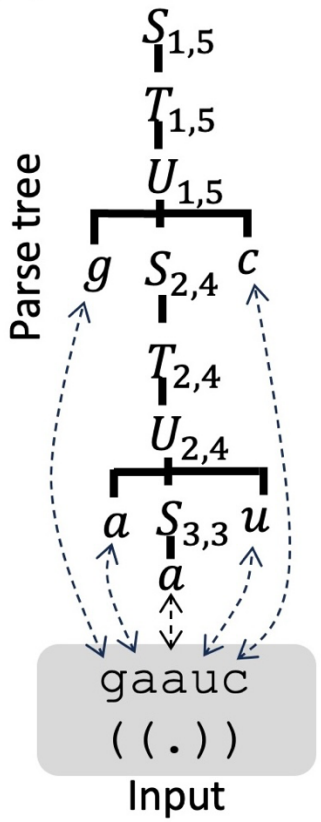
最終的に用いた文法

$$S_{i,j} \rightarrow nS_{i+1,j}|T_{i,j}$$
$$T_{i,j} \rightarrow T_{i,j-1}n|U_{i,j}|T_{i,k}U_{k+1,j}$$
$$U_{i,j} \rightarrow nS_{i,j}\hat{n}$$
$$S_{i,j} \rightarrow n$$

Uの導入で、塩基対の出力がUに集約されて、
プログラムの実装がシンプルになる。

Deep generative model of RNAs based on variational autoencoder with context free grammar

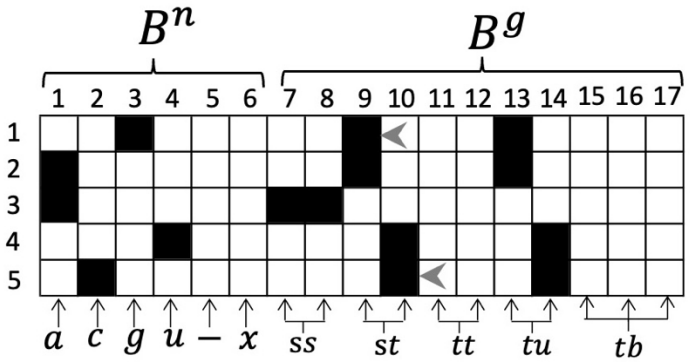
(a)



(b)

st: $S_{1,5} \rightarrow T_{1,5}$
tu: $T_{1,5} \rightarrow U_{1,5}$
us: $U_{1,5} \rightarrow gS_{2,4}c$
st: $S_{2,4} \rightarrow T_{2,4}$
tu: $T_{2,4} \rightarrow U_{2,4}$
us: $U_{2,4} \rightarrow aS_{3,3}u$
st: $S_{3,3} \rightarrow a$

(c)



$$M^S_{i,j} = \max \begin{bmatrix} M^S_{i+1,j} \cdot \max_{n \in N} \{\pi(ss_{i,j,n})\} \\ M^T_{i,j} \cdot \pi(st_{i,j}) \end{bmatrix}$$

$$M^T_{i,j} = \max \begin{bmatrix} M^T_{i,j-1} \cdot \max_{n \in N} \{\pi(tt_{i,j,n})\} \\ M^U_{i,j} \cdot \pi(tu_{i,j}) \\ \max_{i<k<j} \{M^T_{i,k} \cdot M^U_{k+1,j} \cdot \pi(tb_{i,j,k})\} \end{bmatrix}$$

$$M^U_{i,j} = \max_{n\hat{n} \in P} \left[M^S_{i+1,j-1} \cdot \pi(us_{i,j,n,\hat{n}}) \right]$$

Terai G and Asai K, Bioinformatics, accepted

生成したRNAの品質（bit score）

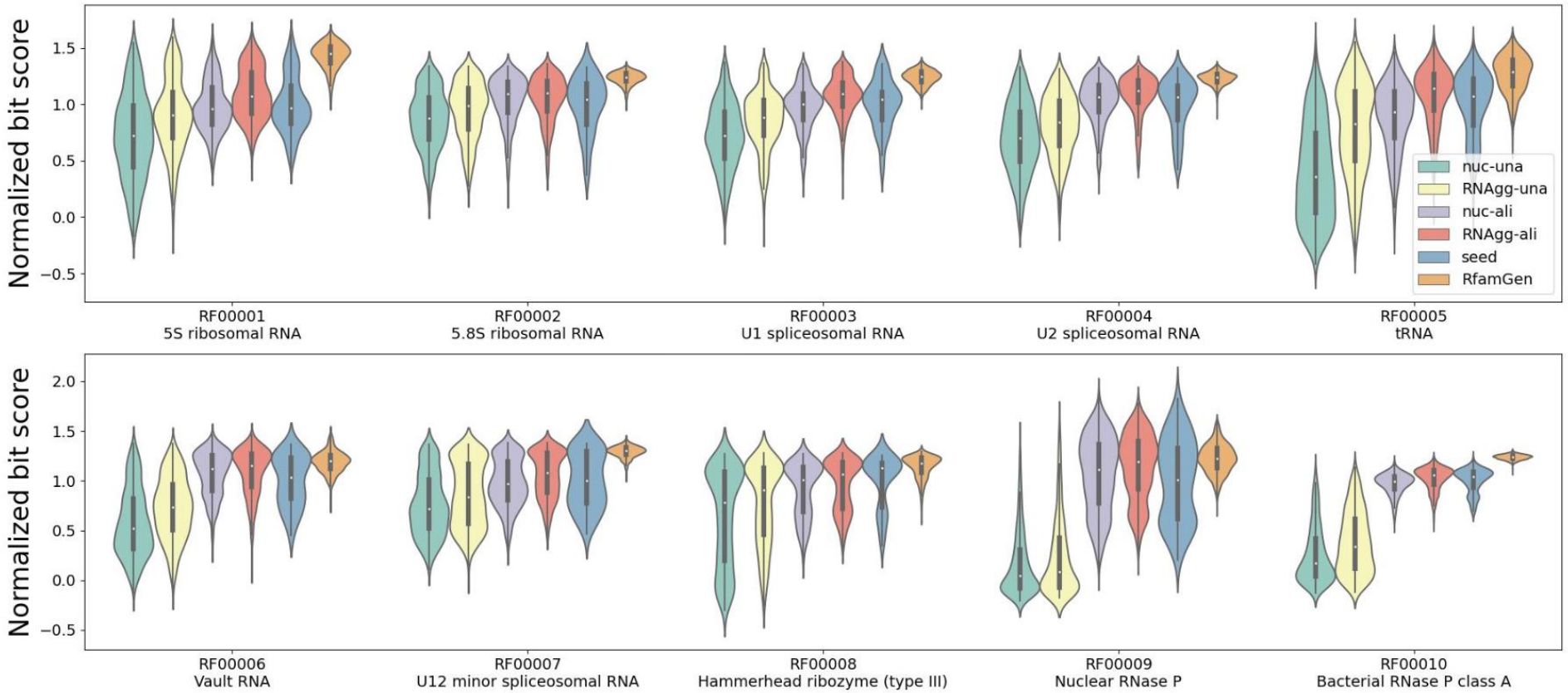
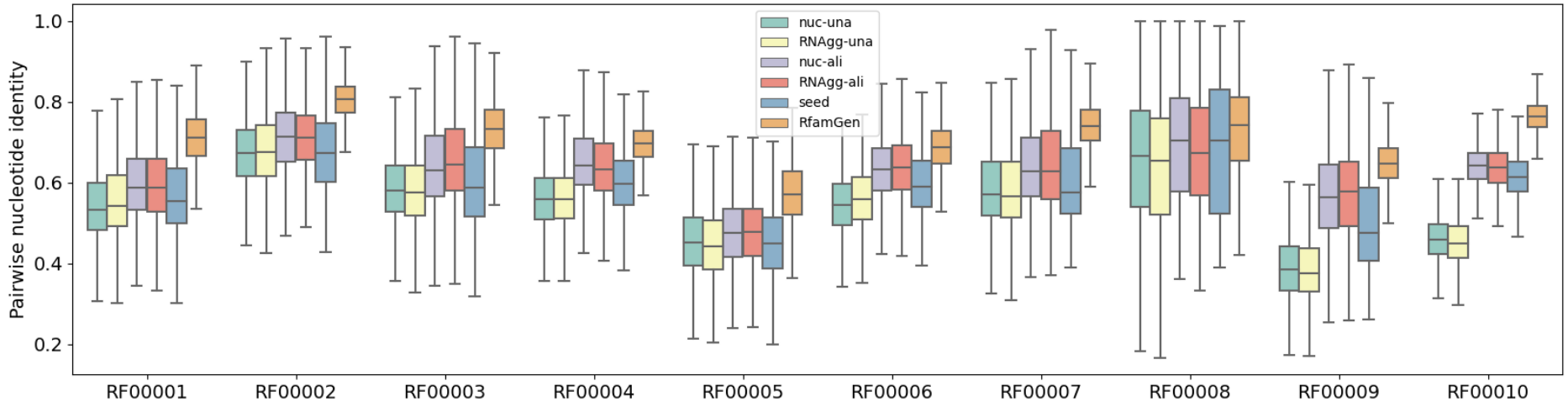


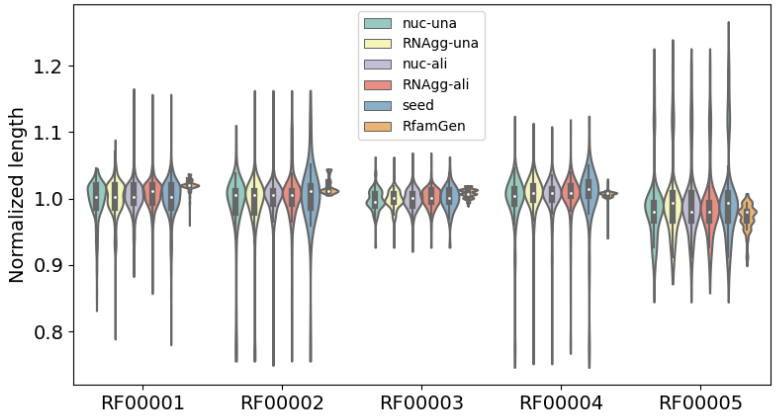
Fig. 3: Normalized bit score of generated RNAs for 10 RNA families. The bit scores are normalized so that the mean bit scores of Rfam seed sequences is 1.

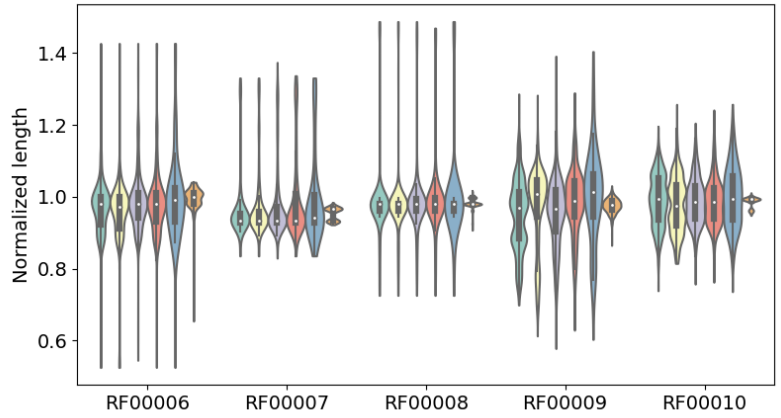
生成したRNAの多様性

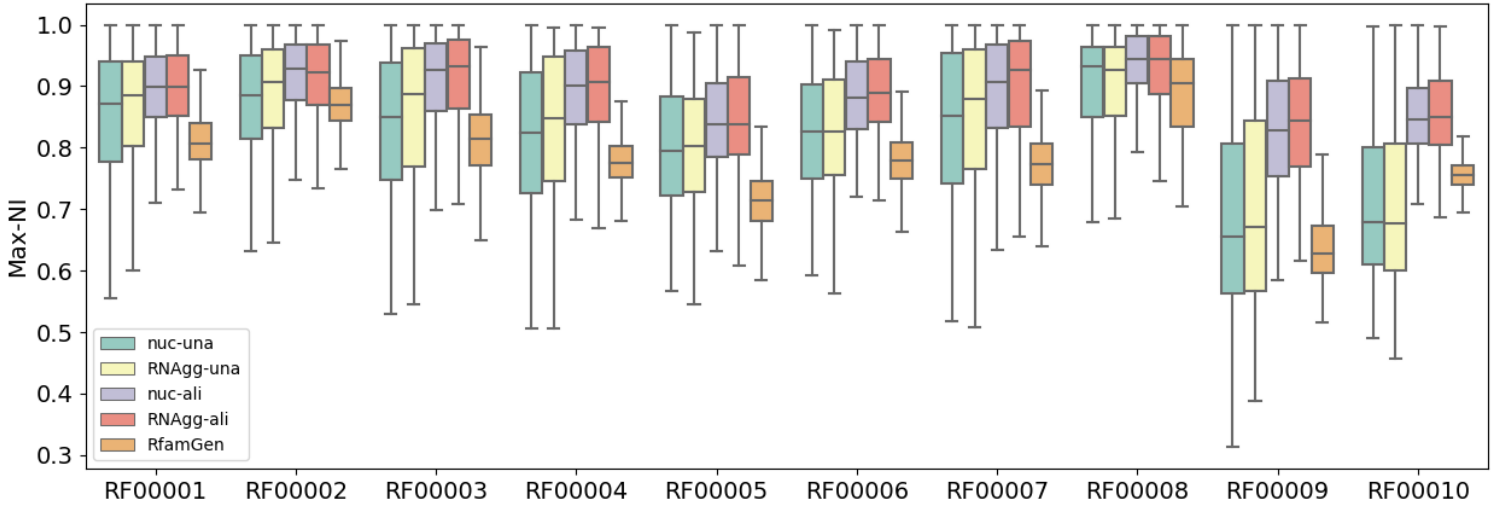
- ３つの指標を使って評価:
  - Pairwise nucleotide identity between generated sequences
  - Length distribution
  - Maximum nucleotide identity to the training sequences
    - 学習データとの類似性



長さ分布、学習データとの類似性

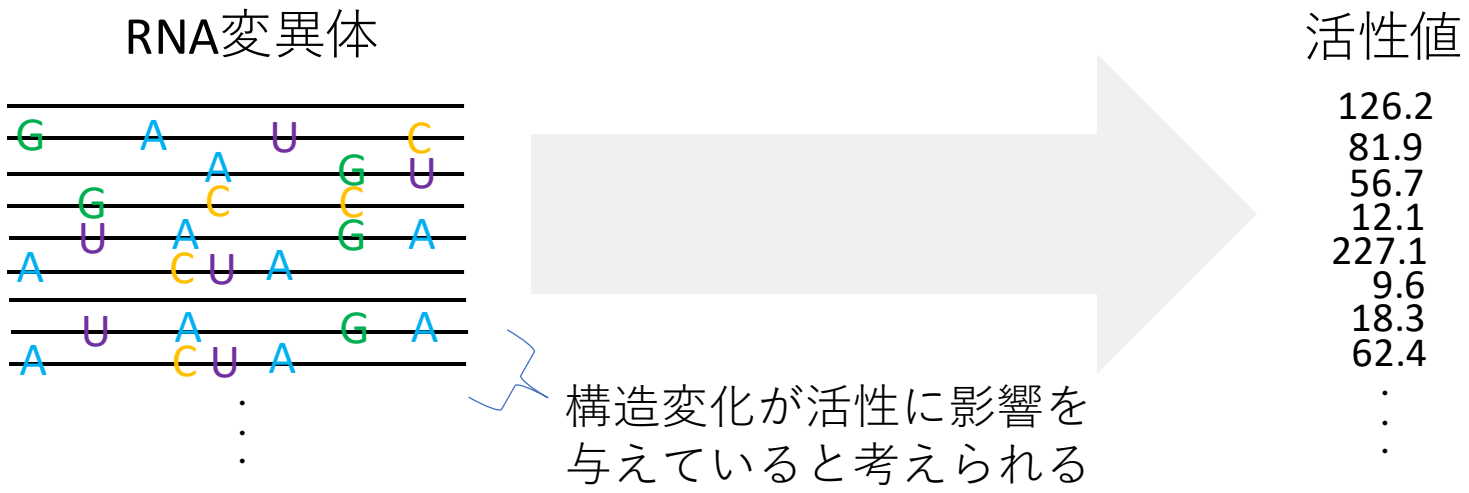




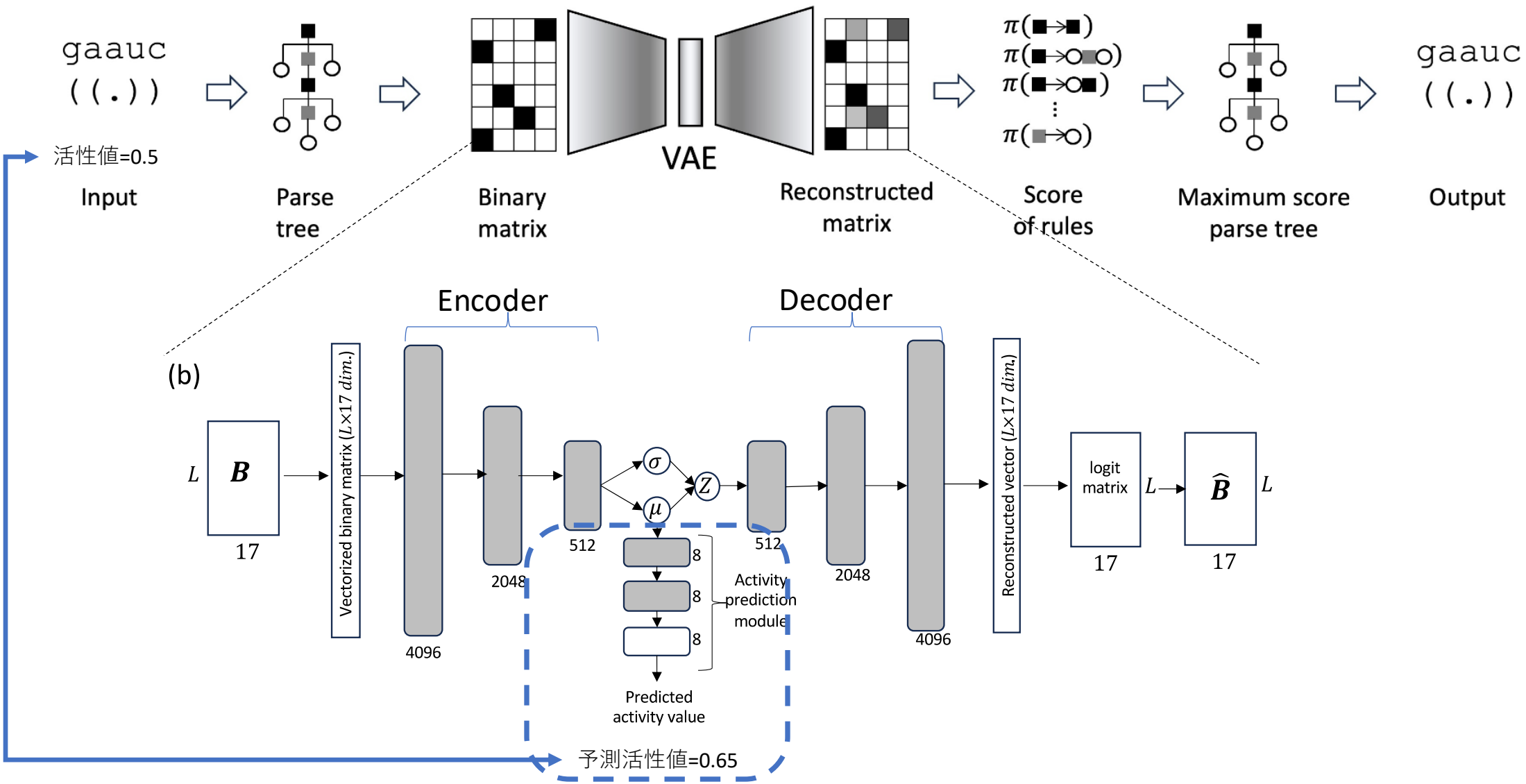


活性値を考慮したRNAの生成

- 「活性値」はRNAの機能の強さを表す数値
  - RNA Aptamer -> 結合強度
  - Ribozyme -> 自己切断活性
  - snoRNA -> 修飾効率
  - …



活性値を考慮するときのVAEアーキテクチャ



Aptazyme dataset Kobori et al. [2017].

- 4^7 aptazyme変異体配列（長さ 160 nt.）

可変領域

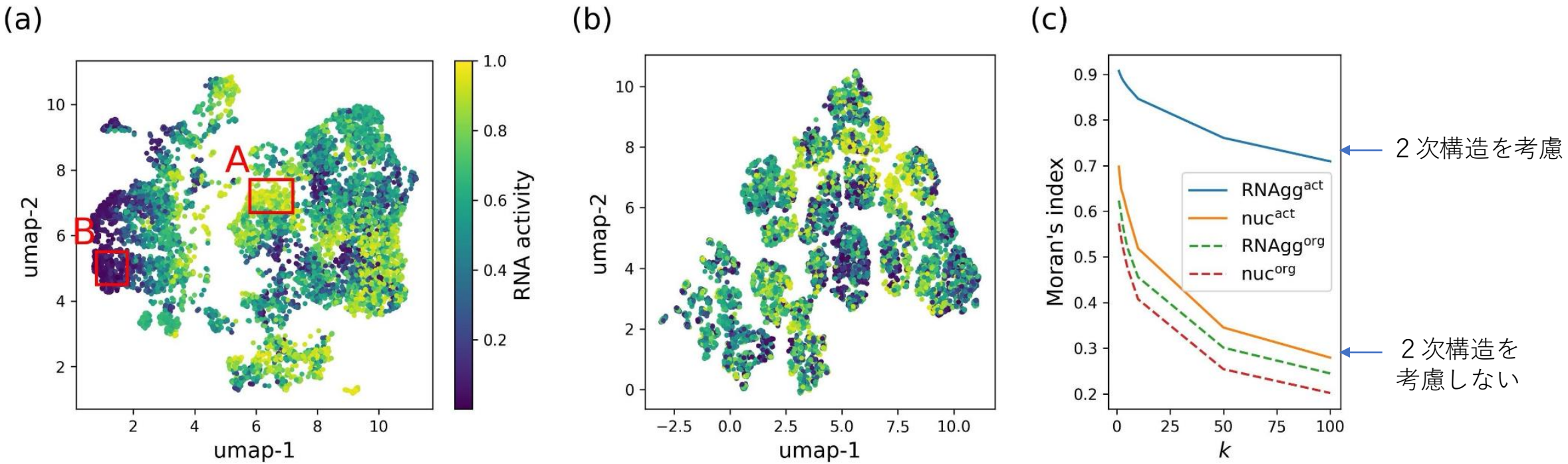
5’-
TAATACGACTCACTATAGGGTCG**NNNNNNN**ATAATCGCGTGGATATGGCACGCAAGTTTCTACC
GGGCACCGTAAATGTCCGACTGGAGCCGTTCGGGCGGCTATAAACAGACCTCAGGCCCGAAG
CGTGGCGGCACCTGCCG
-3’

可変領域に全ての可能な変異を導入し、NGSを使って
すべてのaptazyme変異体の自己切断活性を測定

- MXfold2で２次構造を予測し、
学習データの一部として用いた

- データセットをランダムに２分割し、
学習データ/テストデータとして用いた

Aptazyme dataset Kobori et al. [2017].



2 次構造を考慮

2 次構造を考慮しない

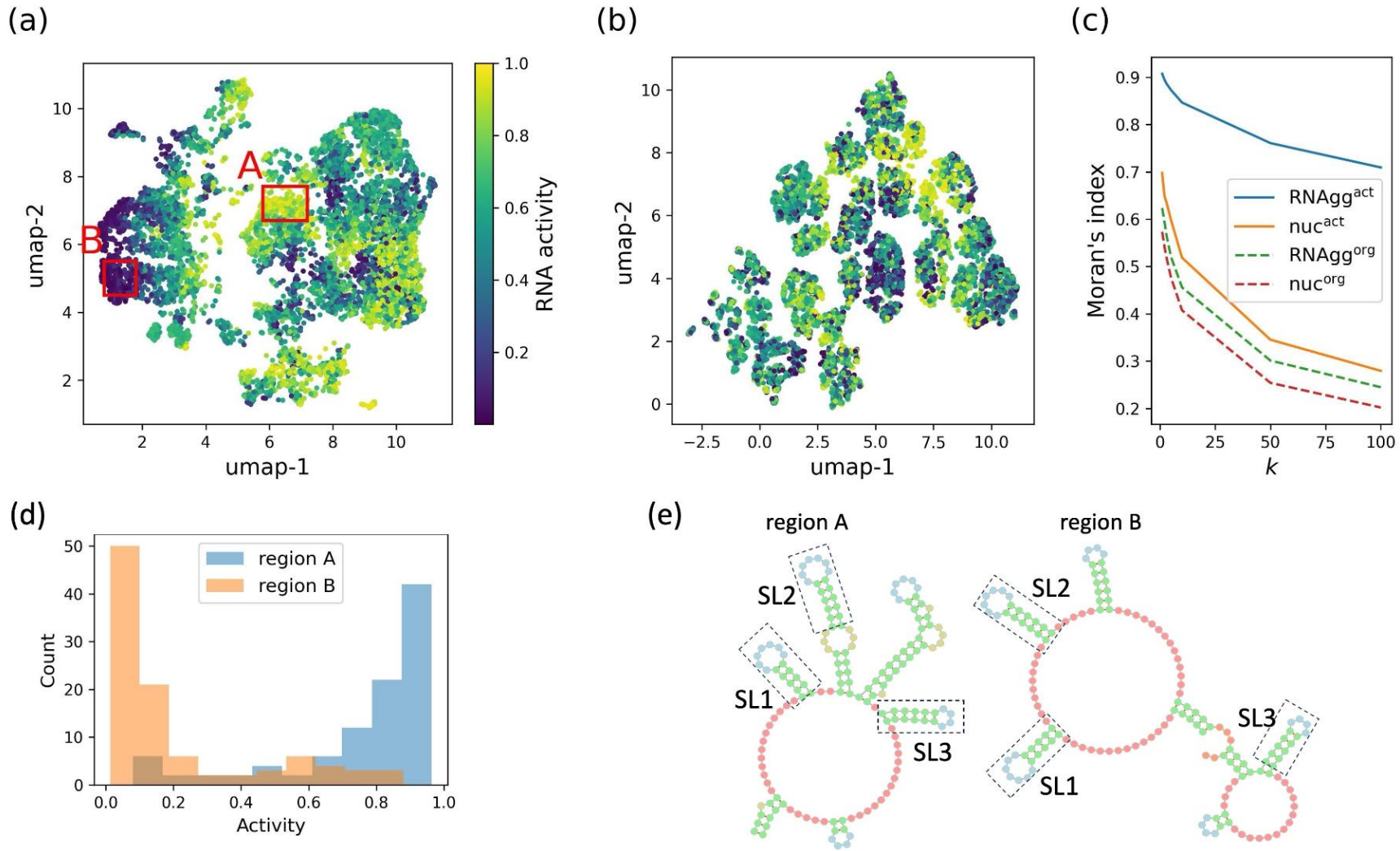


Fig. 4: Relationship between aptazyme and self-cleavage activity in latent space. (a) UMAP visualization of aptazymes in the latent space generated by the $RNAgg^{act}$ model. The color scale indicates the self-cleavage activity of each aptazyme. (b) UMAP visualization of aptazymes in the latent space generated by the nuc^{act} model, with the same color scale as in (a). (c) Moran's I for various methods, showing how activity values cluster spatially. The x-axis represents k, which is the number of neighbors included in the calculation. (d) Distributions of self-cleavage activity values for aptazymes generated from regions A and B, as highlighted in (a). (e) Comparison of the consensus secondary structures of aptazymes generated from regions A and B. Dashed boxes highlight common stem loops.

Remarks and discussion

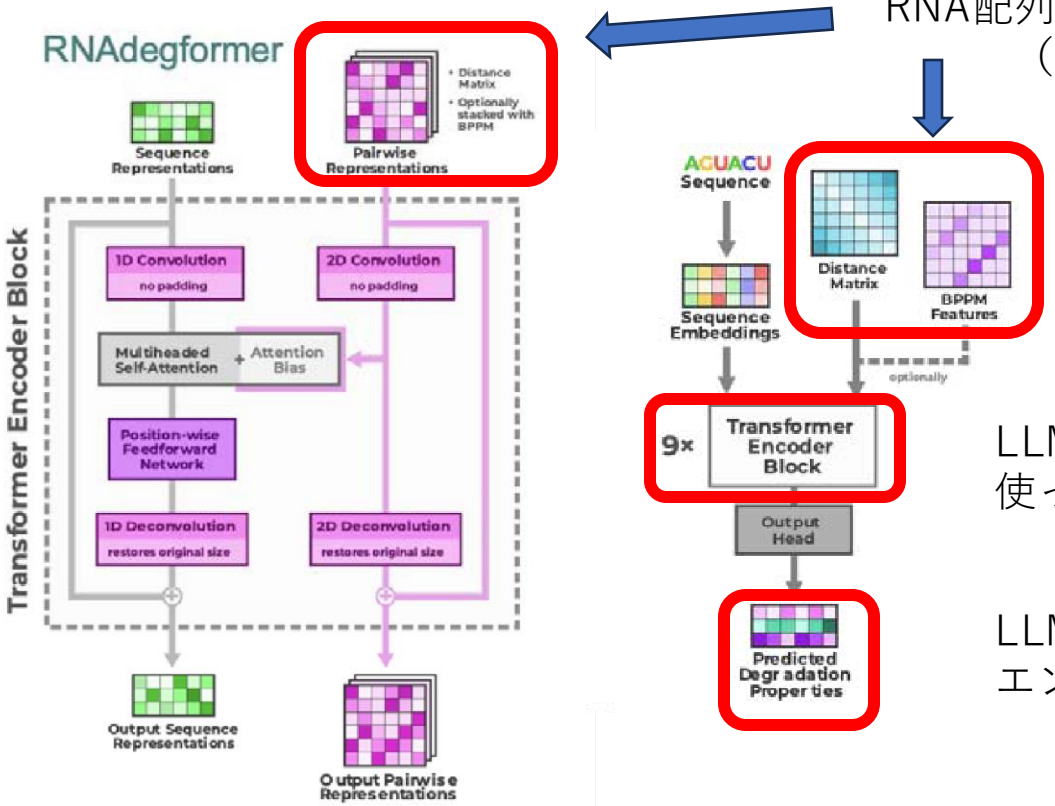
- Artificial Neural Networks and DP algorithms of RNA are DIFFERENTIABLE

- Gradient decent optimization of parameters are applicable to:
  - Artificial neural networks (i.e. deep learning)
  - DP algorithms of stochastic/physical models (i.e. McCaskill algorithm for RNA)

- We can optimize parameters for computation/prediction systems:
  - Energy parameters of RNA 2D structures (MDs: McCaskill)
  - Degradation/Stability/Accessibility/Translation Efficiency (NNs)

- Gradient decent in inputs as a method for input design
  - Sequences are discrete, but there are tricks to overcome the problem
  - Possibility of combination of physical/stochastic models

Combination of physical/stochastic models and NN



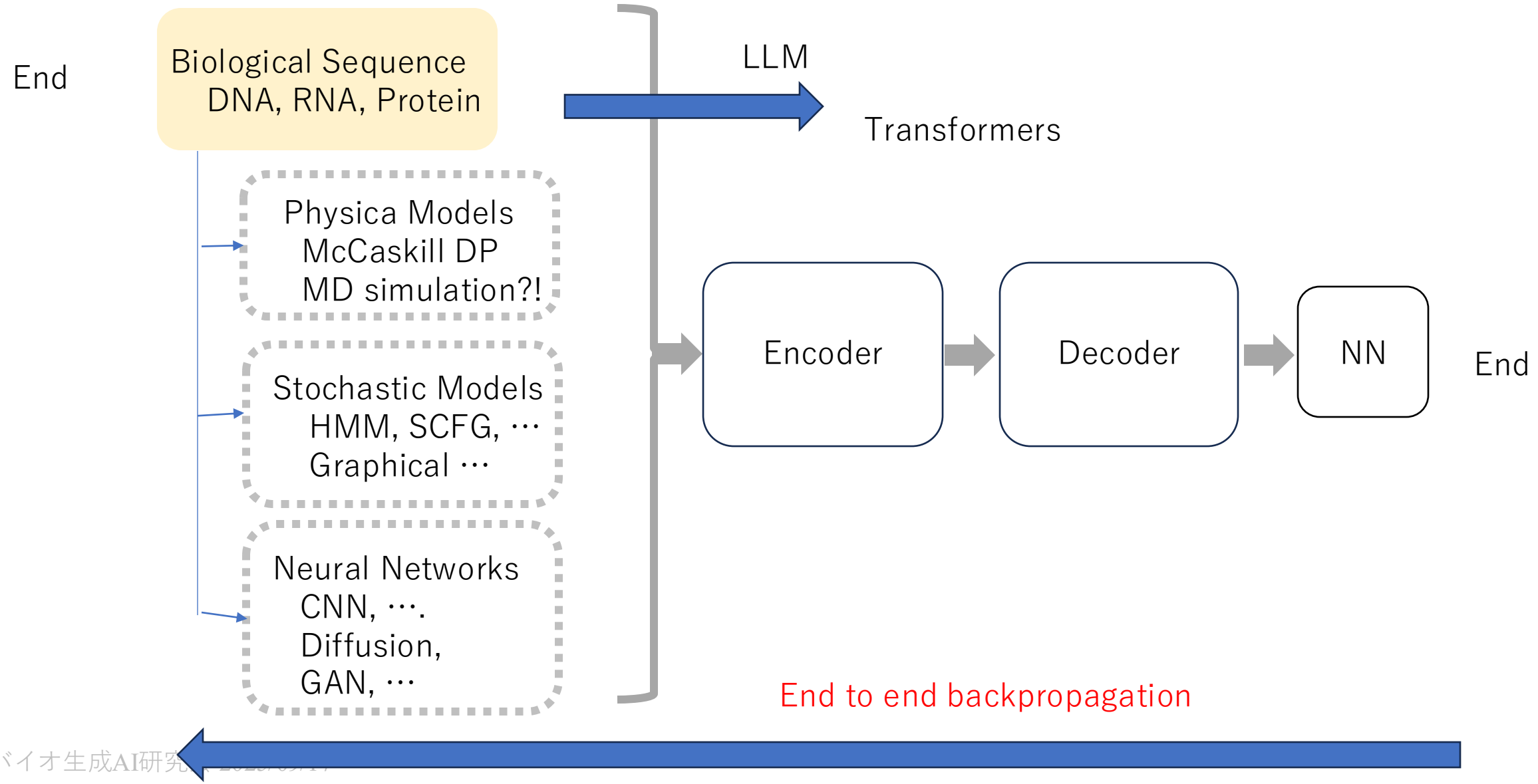
RNA配列に加えて、構造情報を取り込めるようにエンコードしている
（２D畳み込み、塩基対確率、、、、）

LLMで使われるTransformerをエンコーダとして使っている。

LLMで使われるTransformerをエンコーダとして使っている。

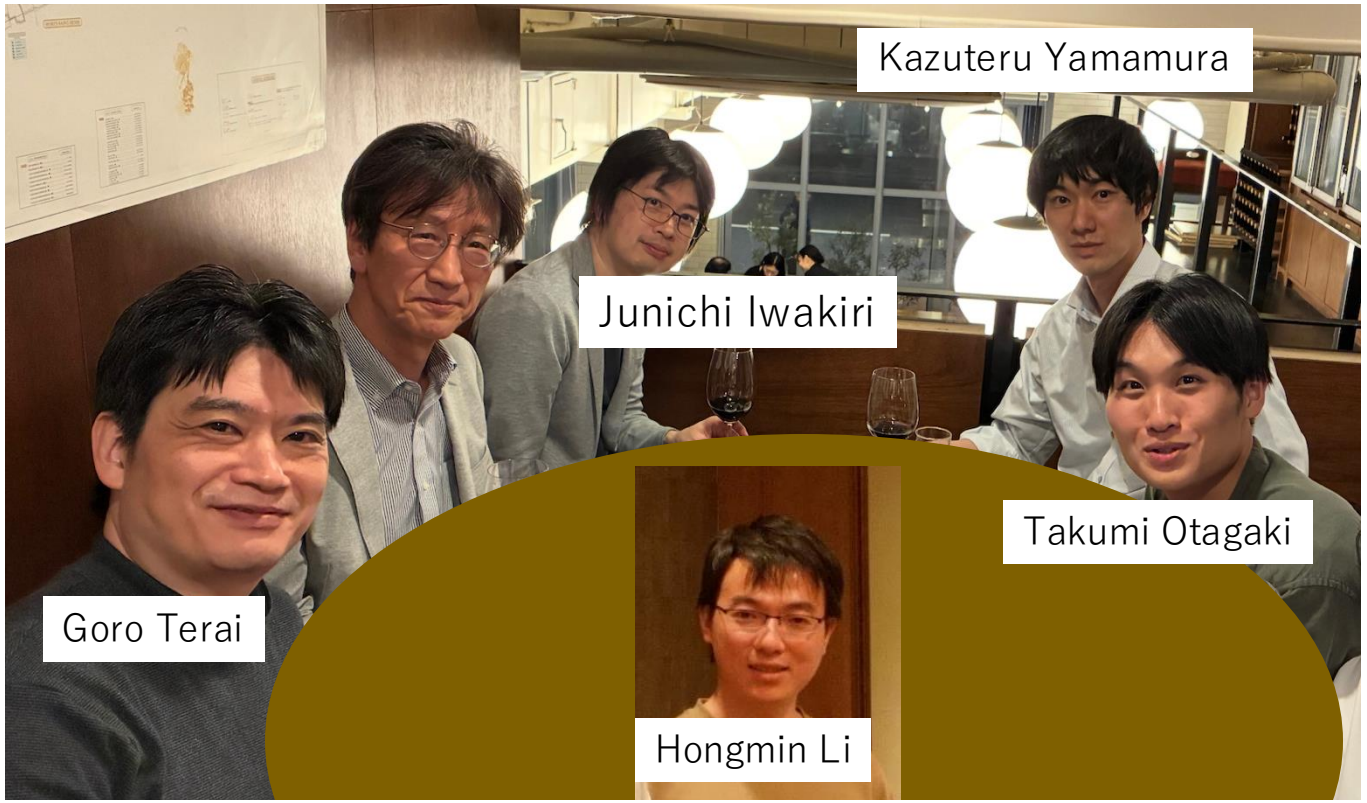
He S, Gao B, Sabnis R, et al. RNAdegformer: accurate prediction of mRNA degradation at nucleotide resolution with deep learning[J]. Briefings in Bioinformatics, 2023, 24(1): bbac581.

Combination of physical/stochastic models with LLM



Acknowledgments

Lab members



Collaborators

Kengo Sato	Sci. Inst. Tokyo
Yutaka Saito	Kitazato Univ.
Bien Bien	Kitazato Univ.
Hiroshi Abe	Nagoya Univ.
Fumitaka Hashiya	Nagoya Univ.
Hiroshi Ueda	Univ. of Tokyo
Tetsuro Hirose	Osaka Univ.
Mikiko Siomi	Univ. of Tokyo

& Lab members of above PIs

Special Thanks to participants of
RNA Dojo 2023, 2024
Benasque RNA 2024