Steering Protein Family Design through Profile Bayesian Flow

Jingjing Gong* Yu Pei* Siyu Long* Yuxuan Song*
Zhe Zhang Wenhao Huang Ziyao Cao
Shuyi Zhang Hao Zhou WeiYing Ma

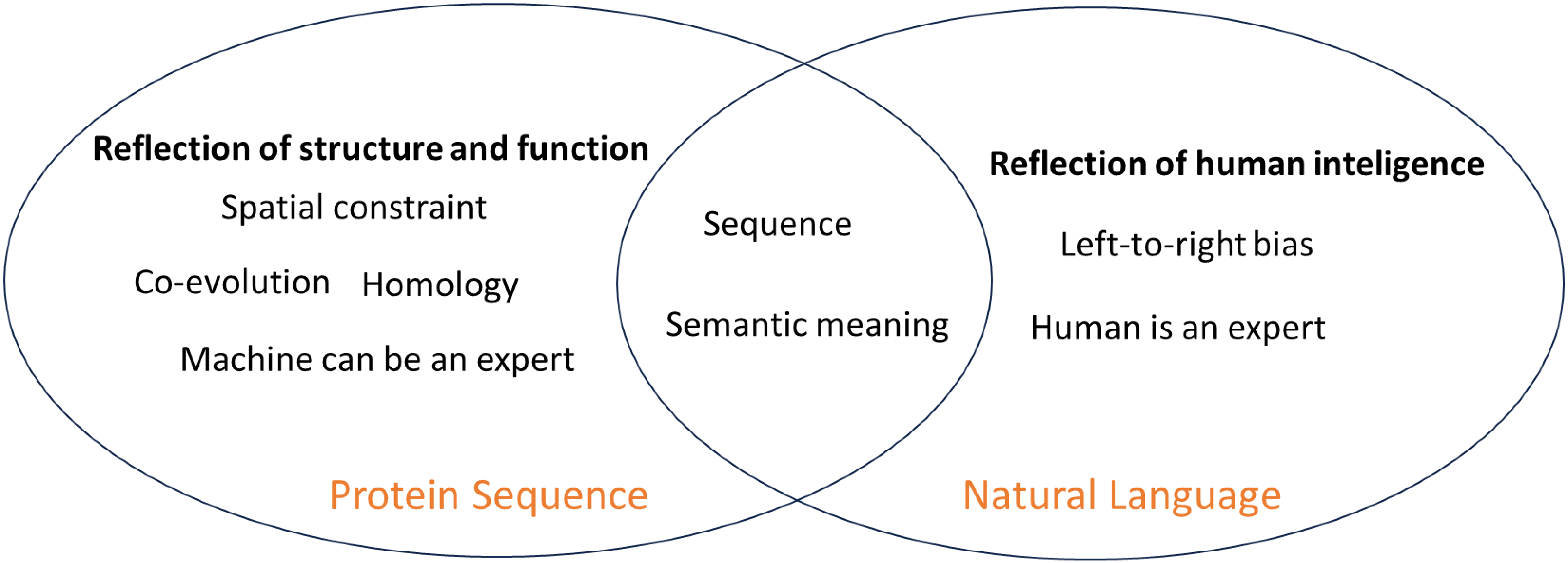
(* denotes equal contribution)

清華大學智能产业研究院
Institute for AI Industry Research, Tsinghua University

Overview

- Some Background
- Profile Bayesian Flow Network
- Experimental Results

Protein Language VS. Natural Language



Reflection of structure and function

Spatial constraint

Co-evolution Homology

Machine can be an expert

Sequence

Semantic meaning

Reflection of human inteligence

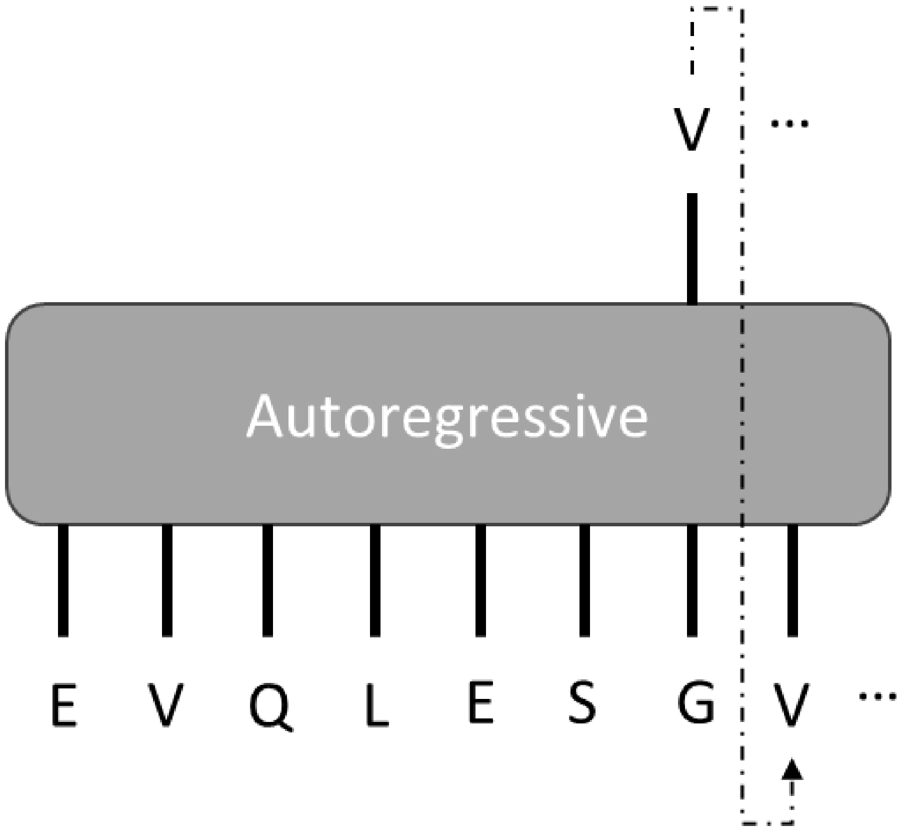
Left-to-right bias

Human is an expert

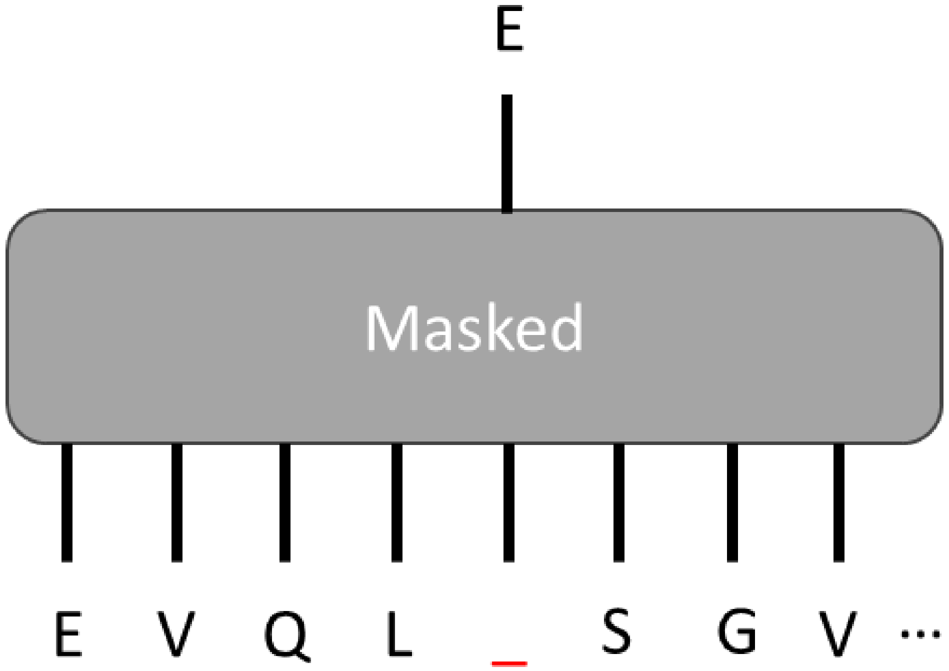
Protein Sequence

Natural Language

Autoregressive VS. Non-autoregressive



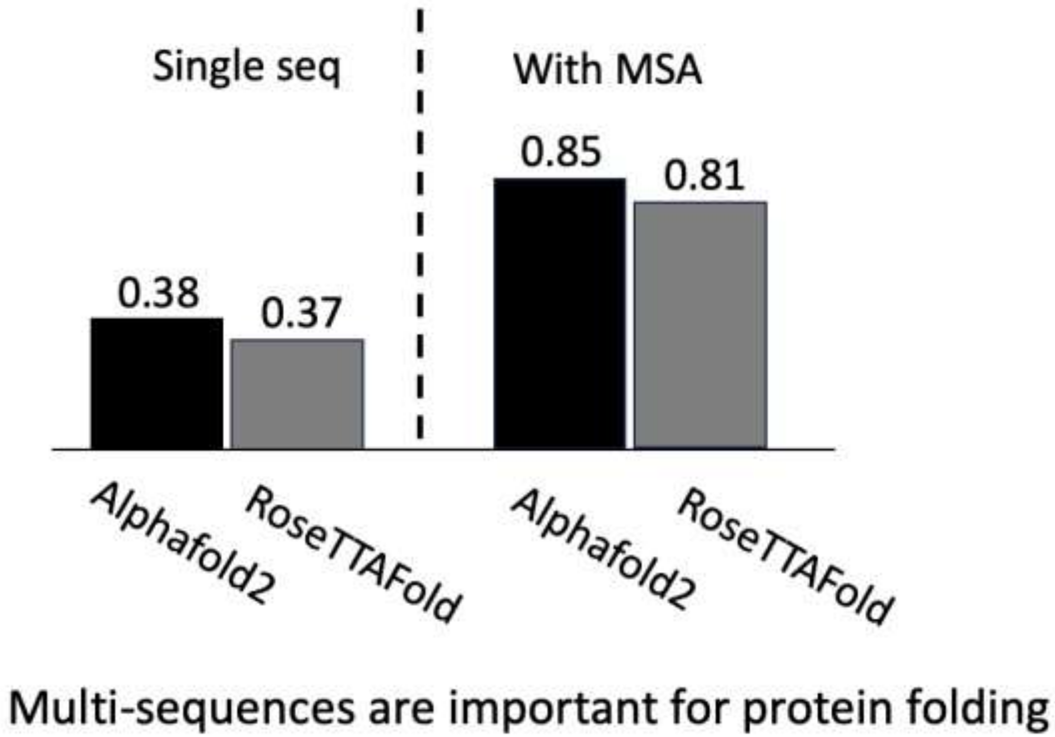
AR Model for protein[1,2]



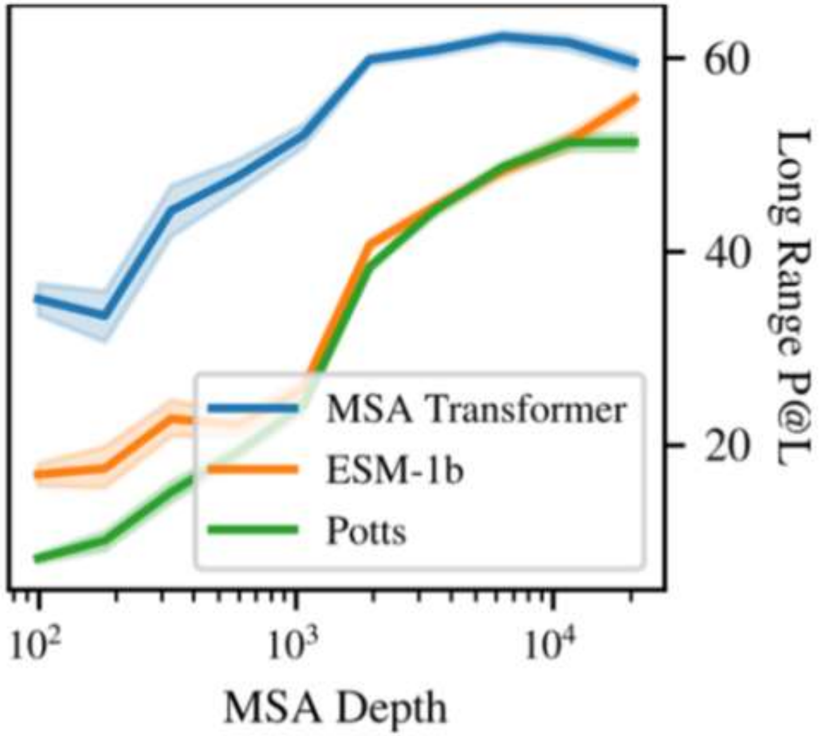
NAR Model for Protein[3,4]

[1] Large language models generate functional protein sequences across diverse families. Madani et al.
[2] ProGen2: Exploring the Boundaries of Protein Language Models. Nijkamp et al.
[3] Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Rives et al.
[4] Evolutionary-scale prediction of atomic level protein structure with a language model. Lin et al.

Co-evolution Information is Critical



Comparison experiments conducted on CASP14, by ESMFold[1]



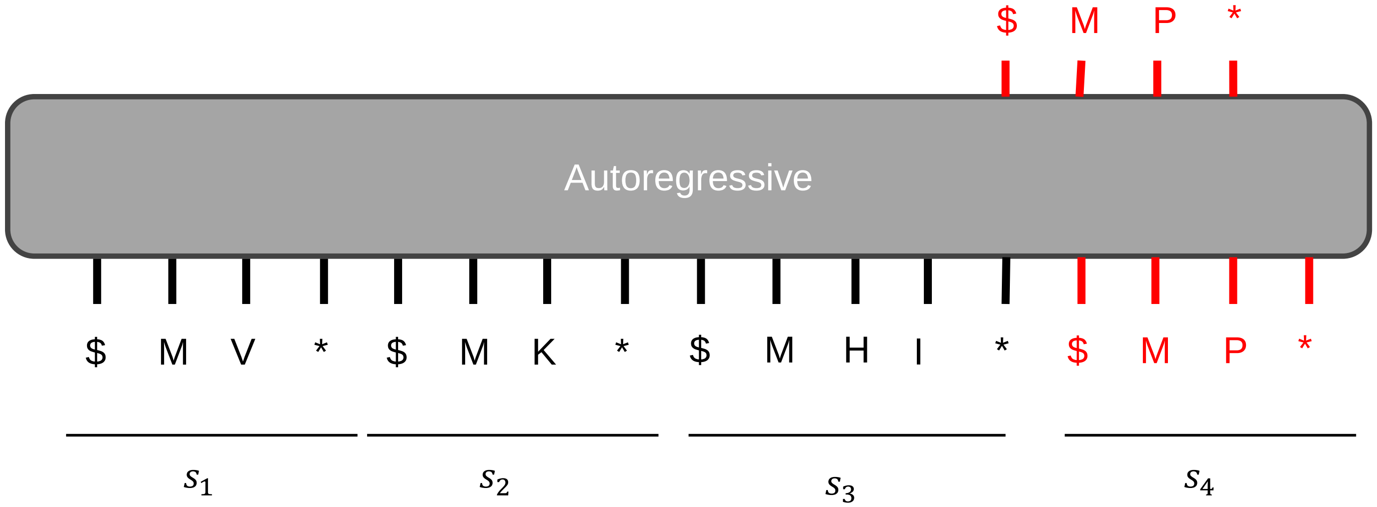
Top-L long-range contact precision comparison on 14,842 proteins conducted by MSA Transformer[2]

[1] Evolutionary-scale prediction of atomic level protein structure with a language model. Lin et al.

[2] MSATransformer. Rao et al.

Multi-Sequence Based Model

AR

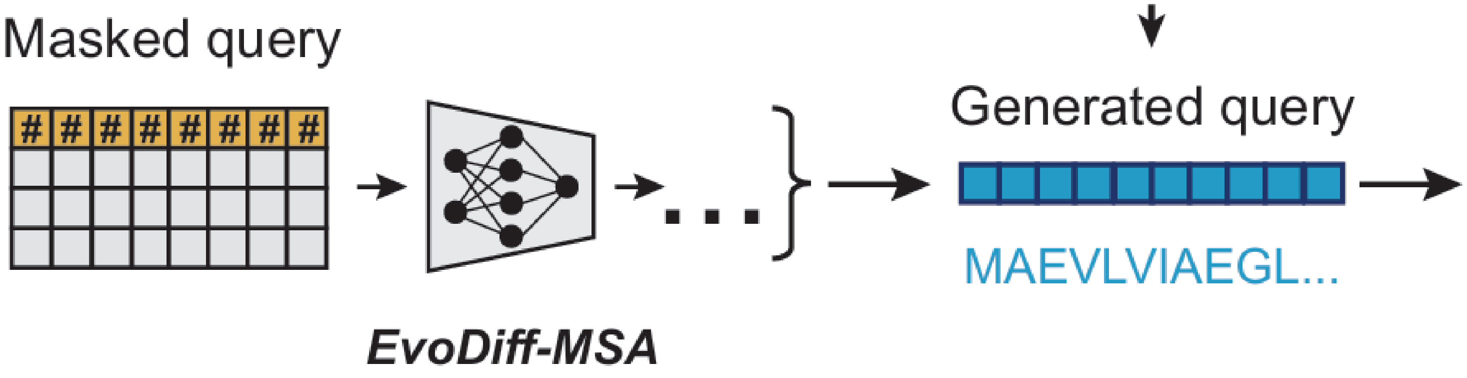


PoET[1], MSAGPT[2]

Multi-seq as Prompt

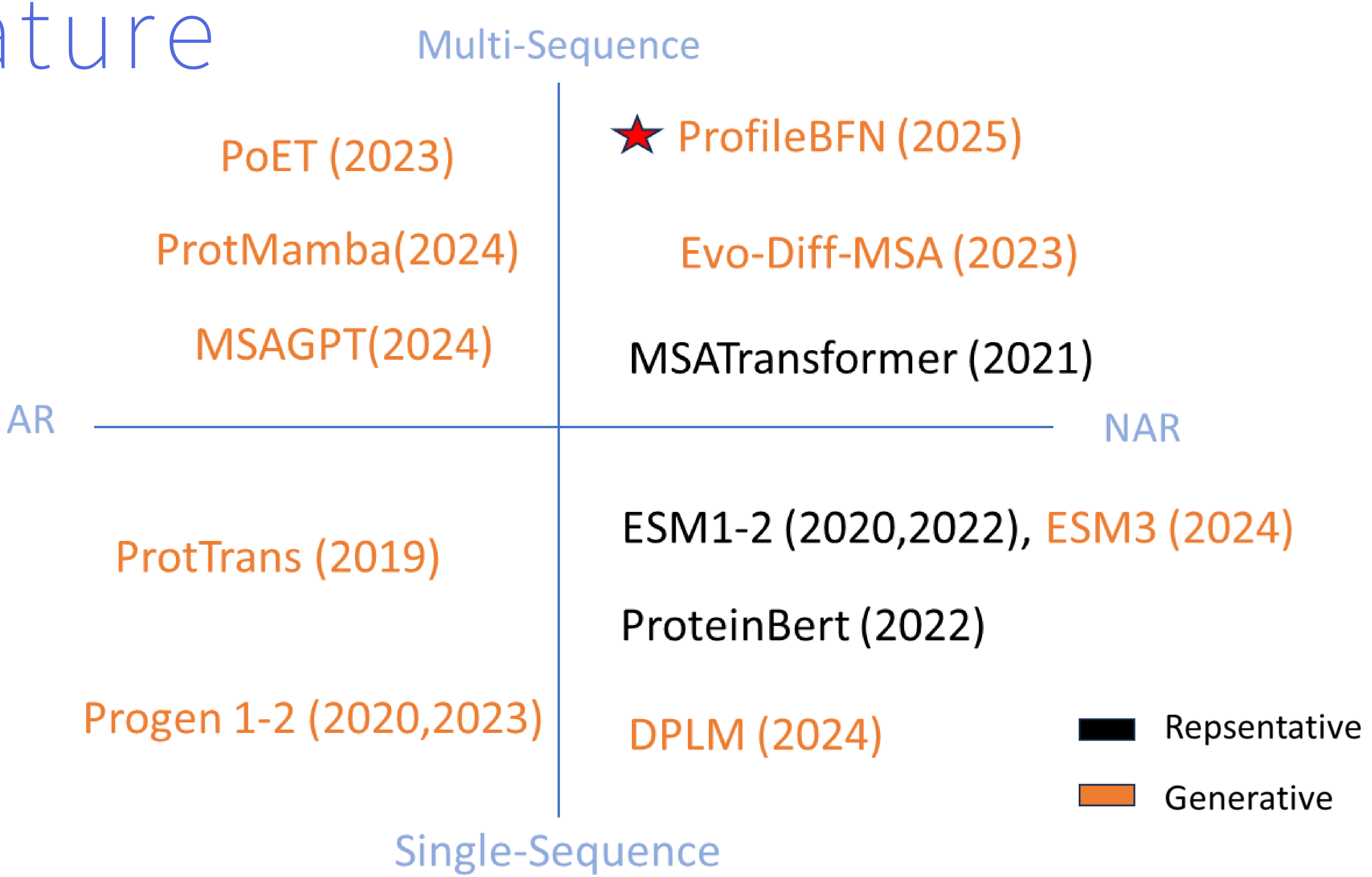
Good quality but inefficient

NAR



MSA as Inputs

Literature Map



Multi-Sequence

PoET (2023)

ProtMamba(2024)

MSAGPT(2024)

★ ProfileBFN (2025)

Evo-Diff-MSA (2023)

MSATransformer (2021)

AR

NAR

ProtTrans (2019)

Progen 1-2 (2020,2023)

ESM1-2 (2020,2022), ESM3 (2024)

ProteinBert (2022)

DPLM (2024)

Single-Sequence

⬛ Repsentative

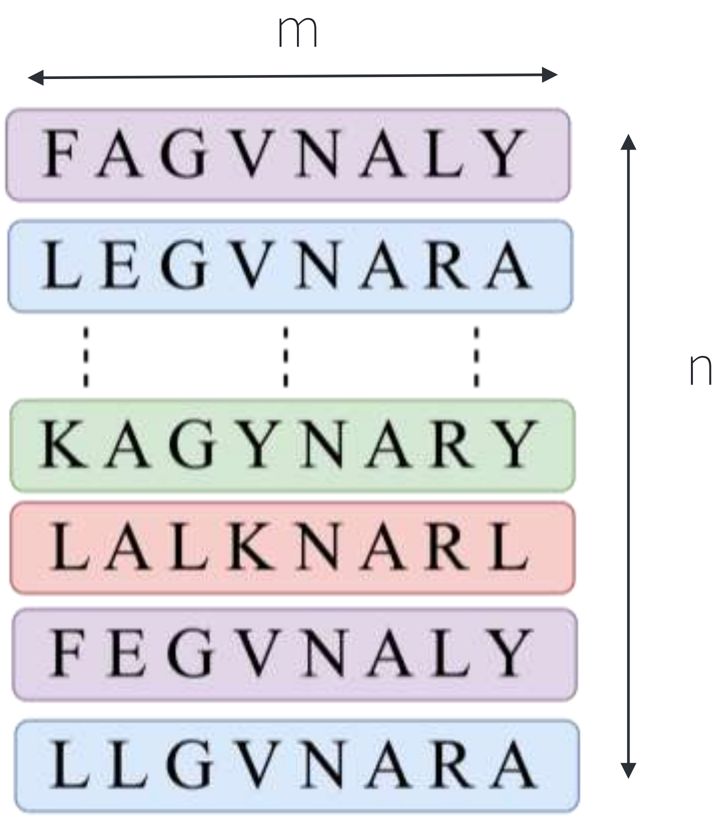
🟧 Generative

Overview

- Some Background
- Profile Bayesian Flow Network
- Experimental Results

What Makes a MSA Profile ?

MSA



FAGVNALY

LEGVNARA

KAGYNARY

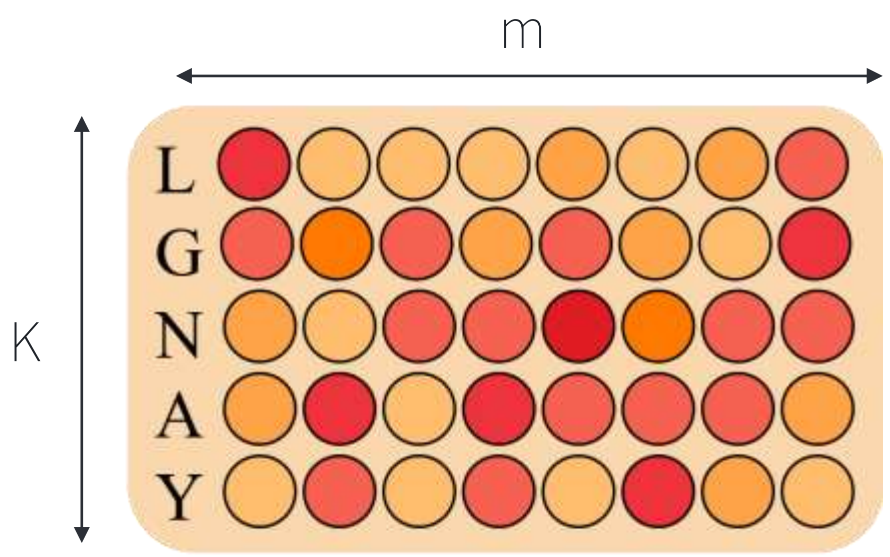
LALKNARL

FEGVNALY

LLGVNARA

$$\boldsymbol{X} \in \{0, \cdots, K\}^{n \times m}$$

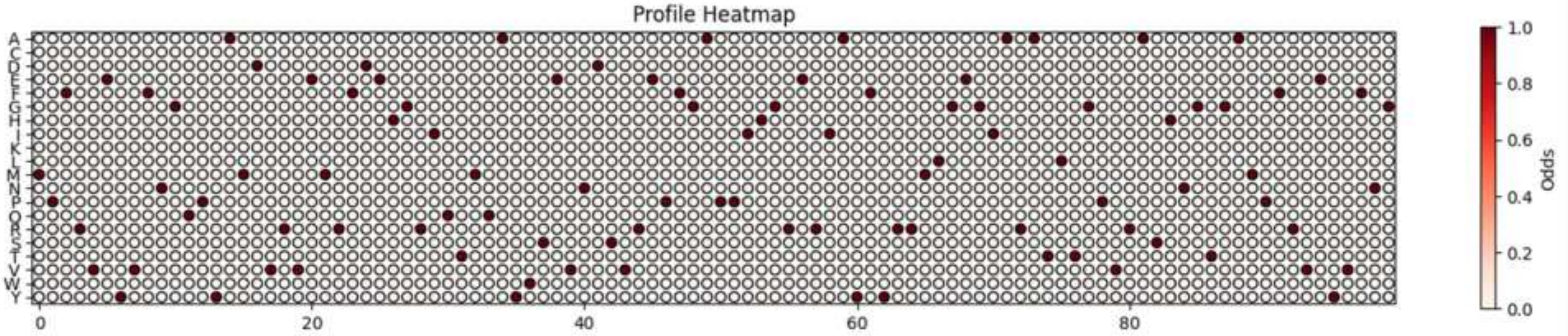
MSA Profile



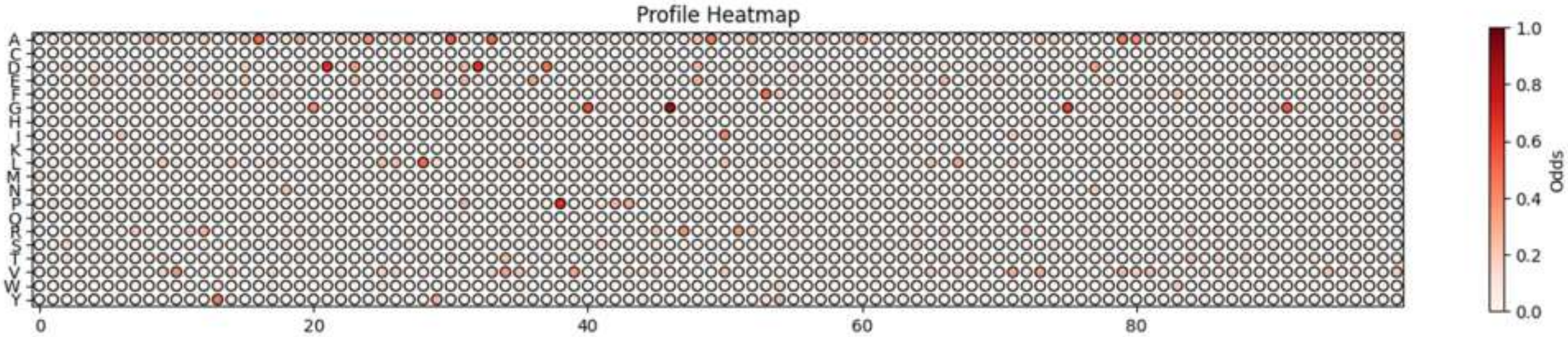
$$\boldsymbol{P}_k^{(i)} = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}_{(\boldsymbol{X}_{ji}=k)}$$

Unified View between Single Sequence and MSA

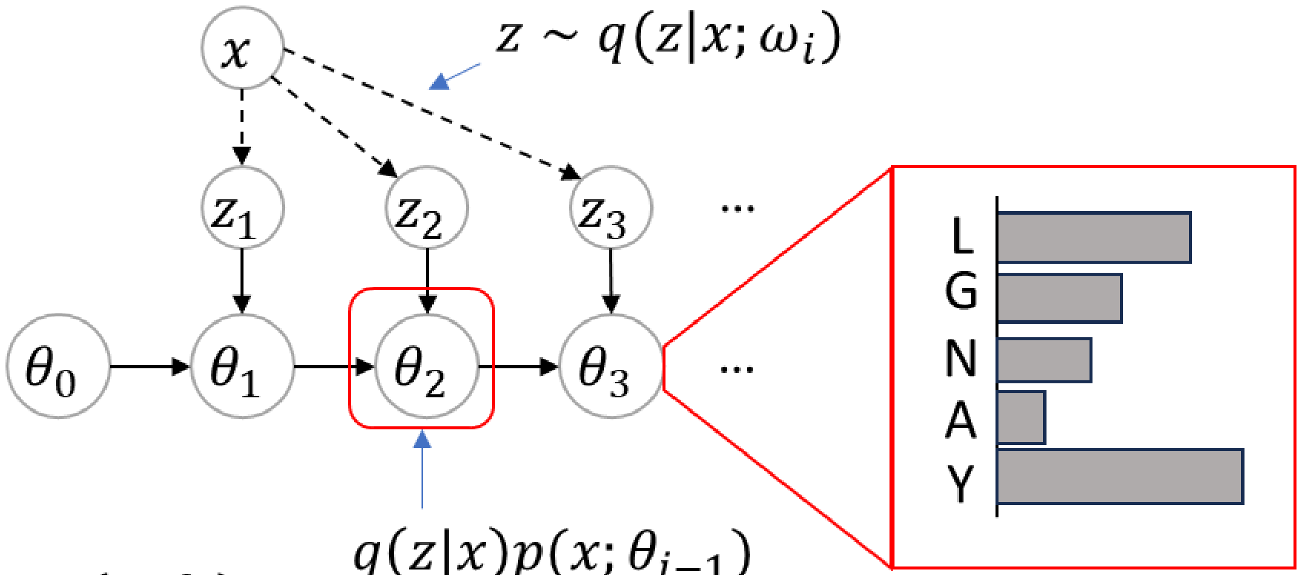
Seq



MSA



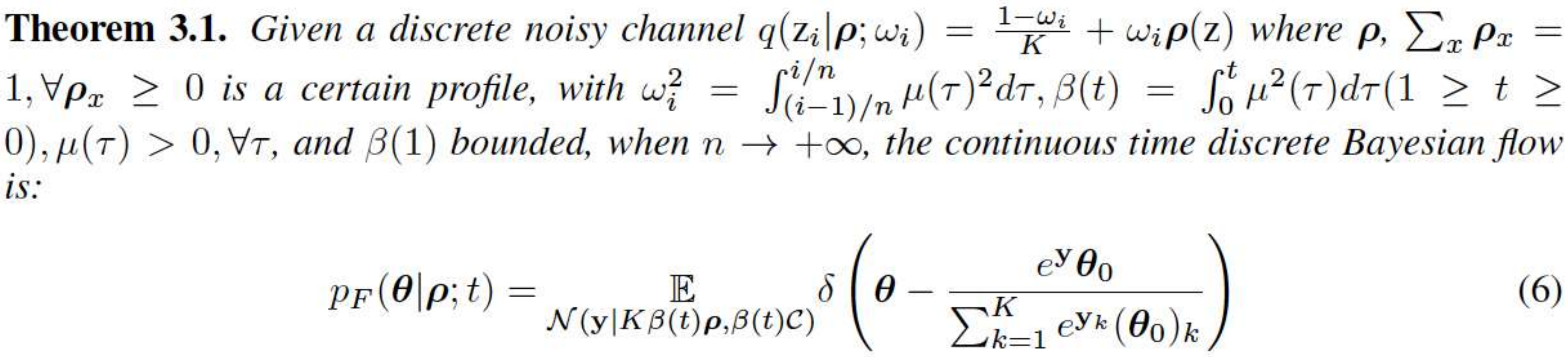
What is the Bayesian Flow?



$$p(x;\theta_i) = \frac{q(z|x)p(x;\theta_{i-1})}{\sum_x q(z|x)p(x;\theta_{i-1})}$$

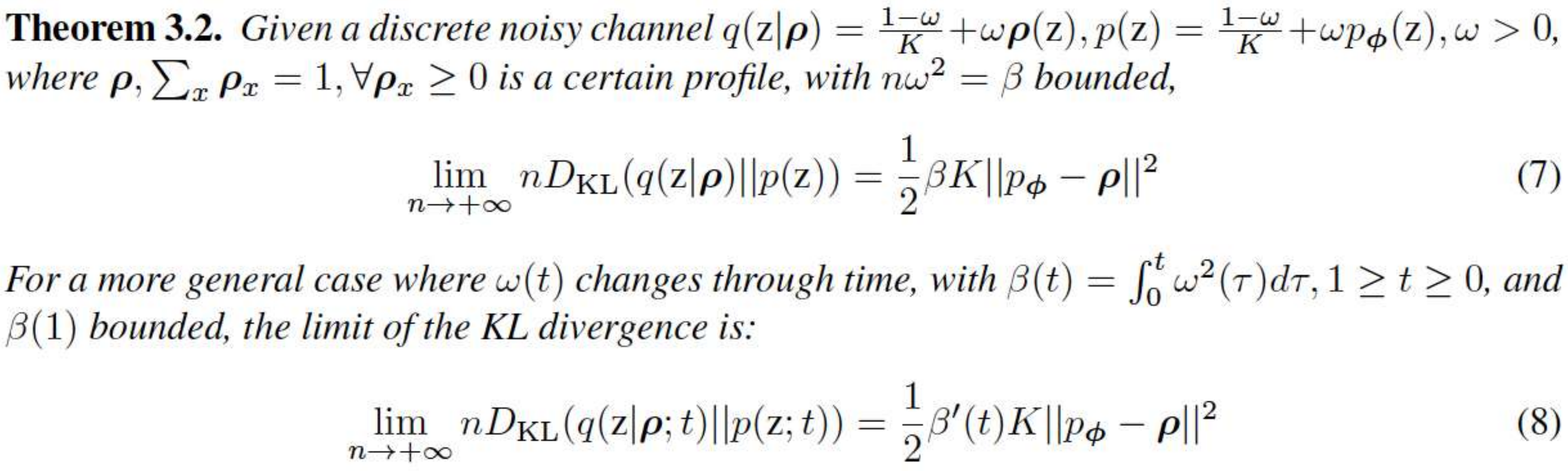
Bayesian update through Bayesian rule

Extension to Profile Data



Theorem 3.1. *Given a discrete noisy channel* $q(\mathrm{z}_i|\boldsymbol{\rho};\omega_i) = \frac{1-\omega_i}{K} + \omega_i\boldsymbol{\rho}(\mathrm{z})$ *where* $\boldsymbol{\rho}$, $\sum_x \boldsymbol{\rho}_x = 1, \forall \boldsymbol{\rho}_x \geq 0$ *is a certain profile, with* $\omega_i^2 = \int_{(i-1)/n}^{i/n} \mu(\tau)^2 d\tau, \beta(t) = \int_0^t \mu^2(\tau) d\tau (1 \geq t \geq 0), \mu(\tau) > 0, \forall \tau$, *and* $\beta(1)$ *bounded, when* $n \to +\infty$, *the continuous time discrete Bayesian flow is:*

$$p_F(\boldsymbol{\theta}|\boldsymbol{\rho};t) = \mathop{\mathbb{E}}_{\mathcal{N}(\mathbf{y}|K\beta(t)\boldsymbol{\rho},\beta(t)\mathcal{C})} \delta\left(\boldsymbol{\theta} - \frac{e^{\mathbf{y}}\boldsymbol{\theta}_0}{\sum_{k=1}^{K} e^{\mathbf{y}_k}(\boldsymbol{\theta}_0)_k}\right) \tag{6}$$



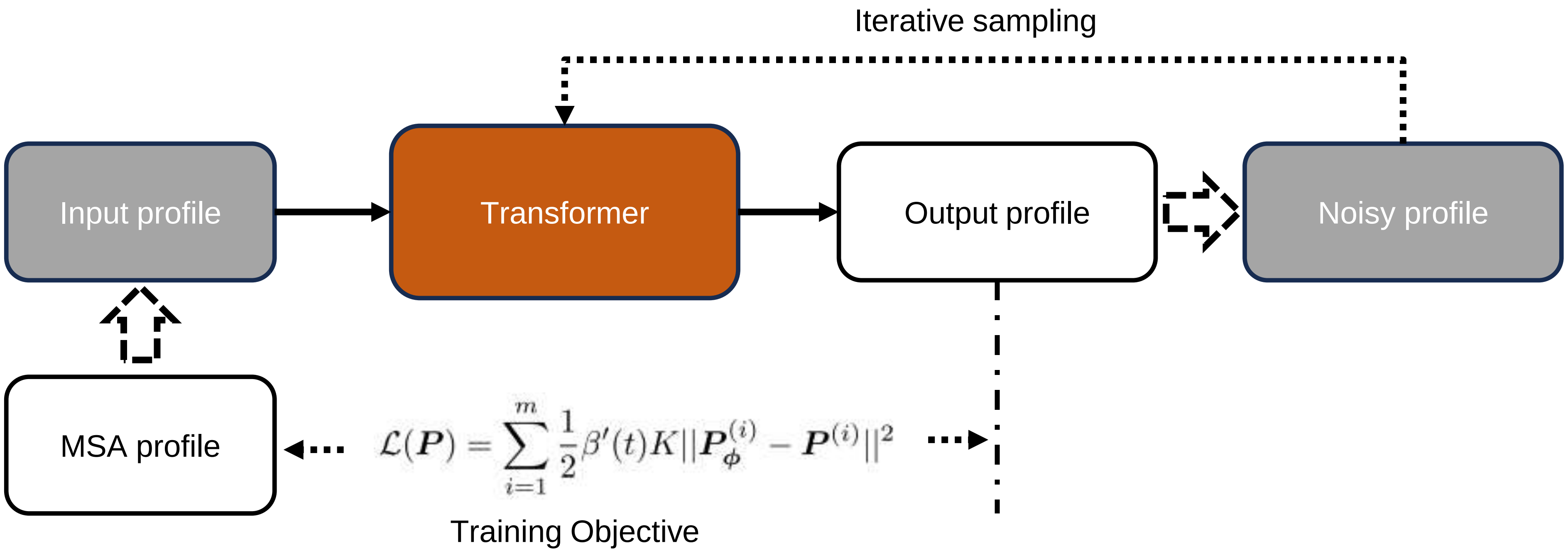
Theorem 3.2. *Given a discrete noisy channel* $q(\mathrm{z}|\boldsymbol{\rho}) = \frac{1-\omega}{K} + \omega\boldsymbol{\rho}(\mathrm{z}), p(\mathrm{z}) = \frac{1-\omega}{K} + \omega p_{\boldsymbol{\phi}}(\mathrm{z}), \omega > 0$, *where* $\boldsymbol{\rho}, \sum_x \boldsymbol{\rho}_x = 1, \forall \boldsymbol{\rho}_x \geq 0$ *is a certain profile, with* $n\omega^2 = \beta$ *bounded,*

$$\lim_{n \to +\infty} nD_{\mathrm{KL}}(q(\mathrm{z}|\boldsymbol{\rho})||p(\mathrm{z})) = \frac{1}{2}\beta K||p_{\boldsymbol{\phi}} - \boldsymbol{\rho}||^2 \tag{7}$$

For a more general case where $\omega(t)$ *changes through time, with* $\beta(t) = \int_0^t \omega^2(\tau) d\tau, 1 \geq t \geq 0$, *and* $\beta(1)$ *bounded, the limit of the KL divergence is:*

$$\lim_{n \to +\infty} nD_{\mathrm{KL}}(q(\mathrm{z}|\boldsymbol{\rho};t)||p(\mathrm{z};t)) = \frac{1}{2}\beta'(t)K||p_{\boldsymbol{\phi}} - \boldsymbol{\rho}||^2 \tag{8}$$

Profile Bayesian Flow



$$\mathcal{L}(\boldsymbol{P}) = \sum_{i=1}^{m} \frac{1}{2}\beta'(t)K||\boldsymbol{P}_{\boldsymbol{\phi}}^{(i)} - \boldsymbol{P}^{(i)}||^2$$

The ⇨ symbol stands for bayesian flow process $p_F(\boldsymbol{\theta}|\boldsymbol{\rho};t) = \mathop{\mathbb{E}}_{\mathcal{N}(\mathbf{y}|K\beta(t)\boldsymbol{\rho},\beta(t)\mathcal{C})} \delta\left(\boldsymbol{\theta} - \frac{e^{\mathbf{y}}\boldsymbol{\theta}_0}{\sum_{k=1}^{K} e^{\mathbf{y}_k}(\boldsymbol{\theta}_0)_k}\right)$

Overview

- Some Background
- Profile Bayesian Flow Network
- Experimental Results

Results: Sampling Efficiency

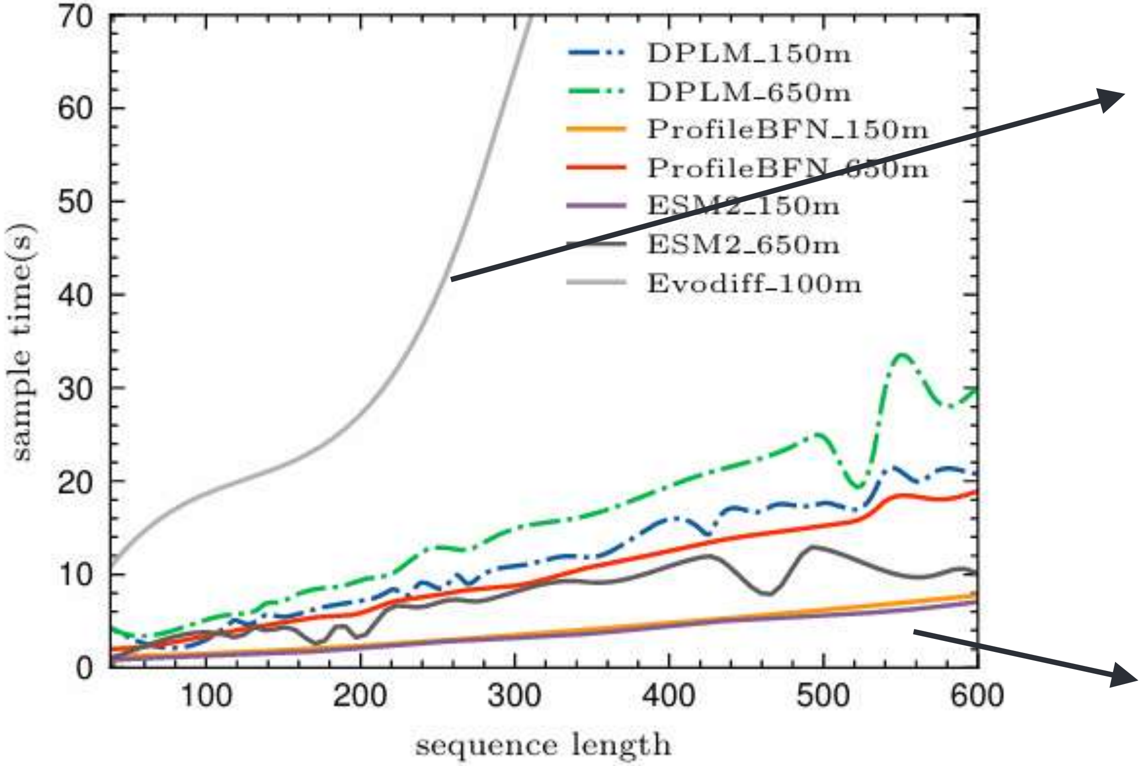
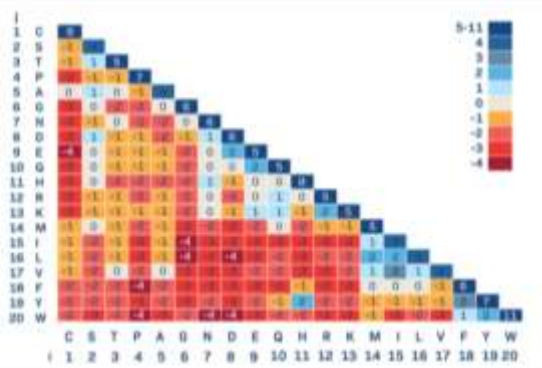


Figure 3: Sampling efficiency comparison. ProfileBFN has a higher sampling efficiency compared to its competitors.

Evodiff sampling time increases rapidly

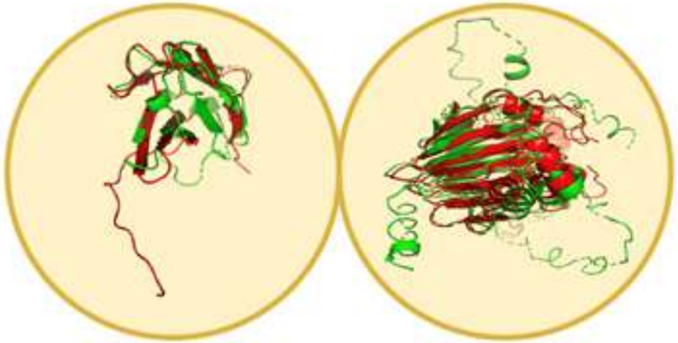
ProfileBFN shares the same complexity with single seq models

Problems in Parameterized Evaluation

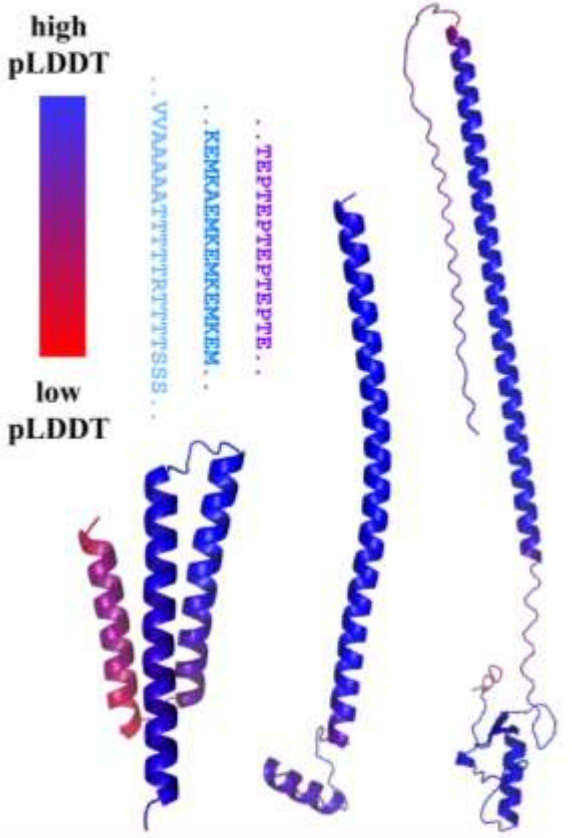


🟥 GT

🟩 Hacked



ID	8SUF	8UAI
Identity	0.40	0.40
pLDDT	75.86	72.77
pTM	0.743	0.769

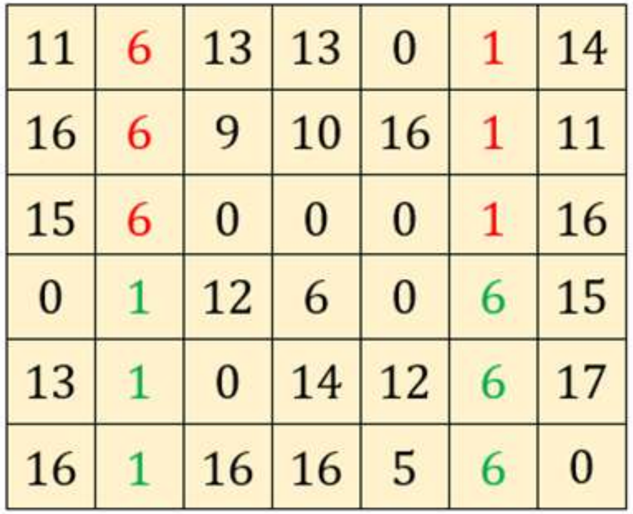


pLDDT	72.34	80.63	81.42

Non-paramerized Evaluation

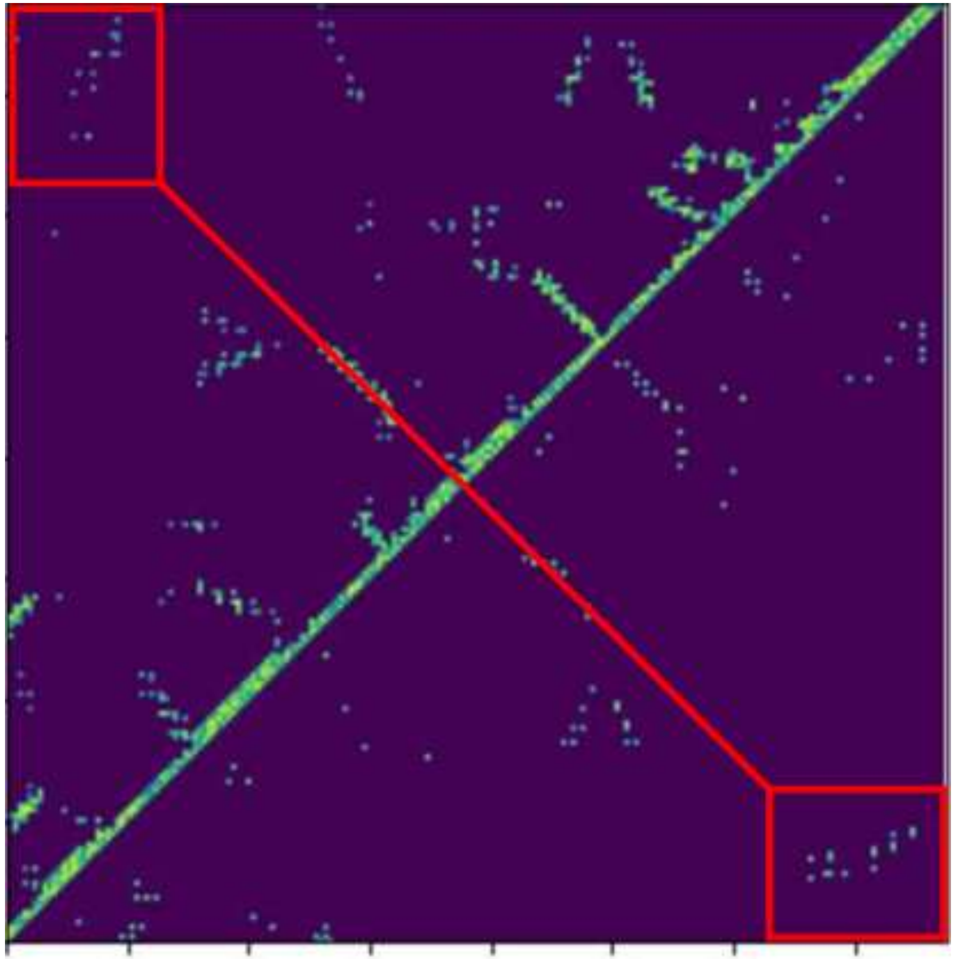
```
MGPP-AQ
SGKLSAM
RG---AS
-AWG-GR
PA-QNGT
SASSFG-
```

→



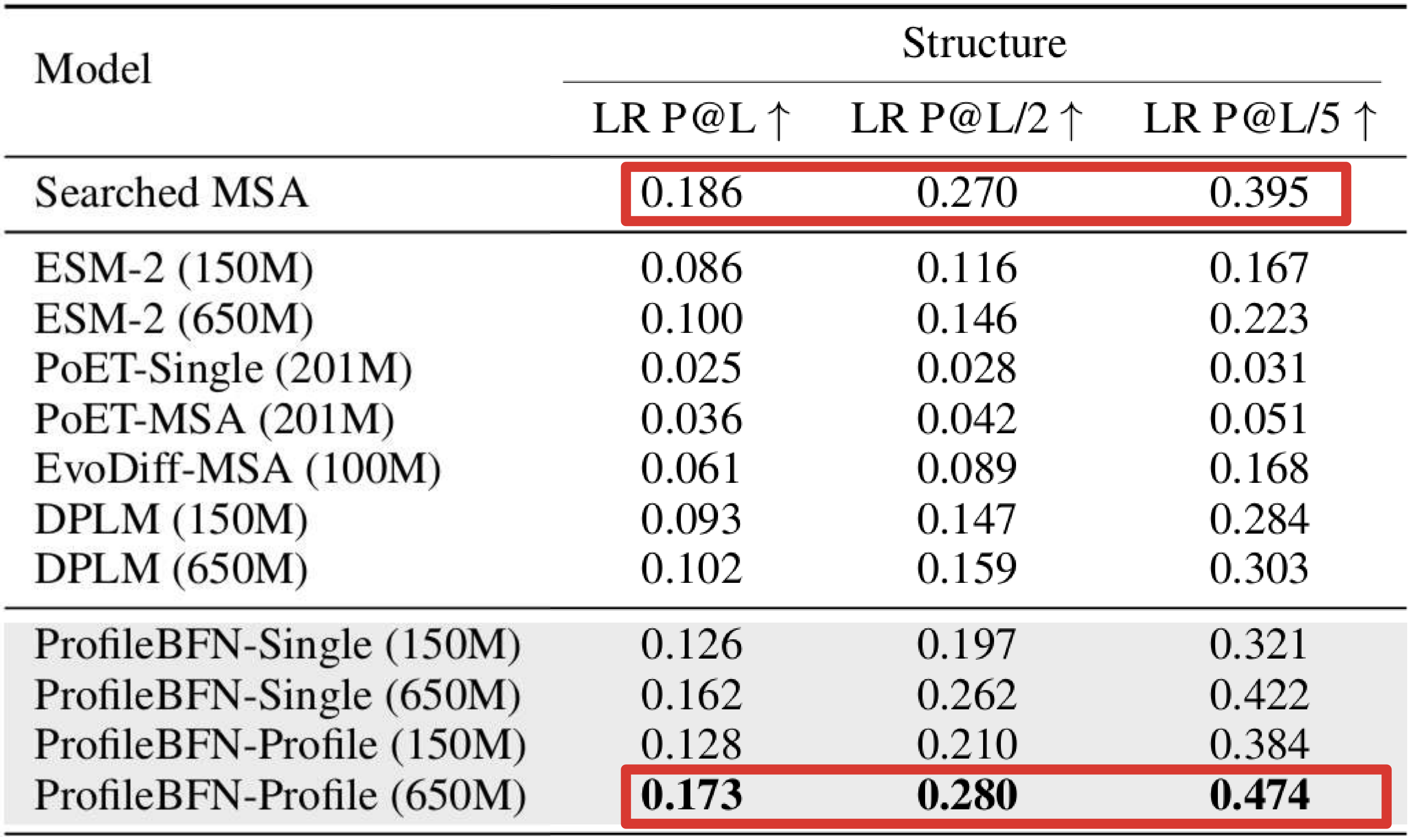
11	6	13	13	0	1	14
16	6	9	10	16	1	11
15	6	0	0	0	1	16
0	1	12	6	0	6	15
13	1	0	14	12	6	17
16	1	16	16	5	6	0

→



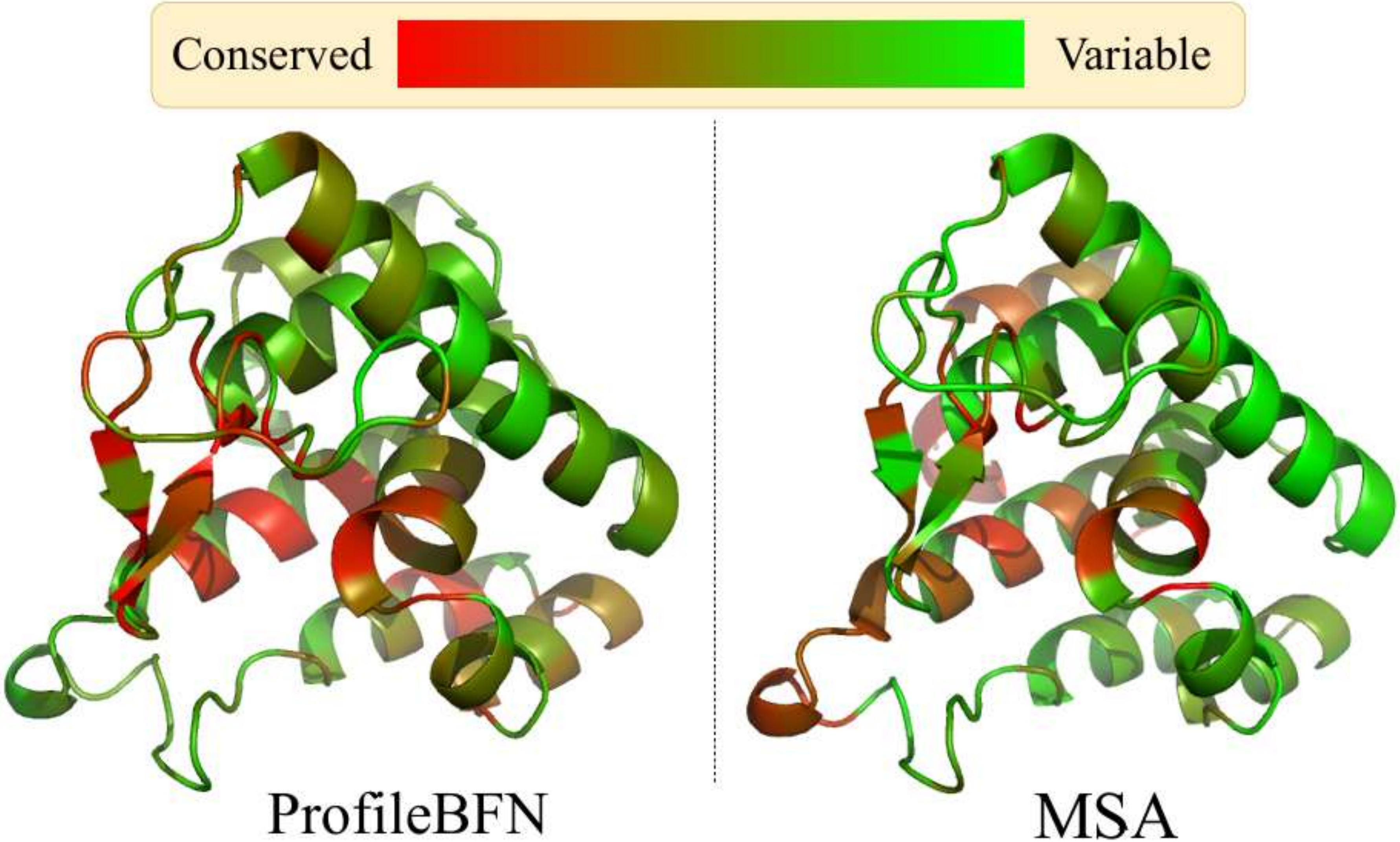
CCMPRED

Results: Profile Better than MSA



Model	Structure		
	LR P@L ↑	LR P@L/2 ↑	LR P@L/5 ↑
Searched MSA	0.186	0.270	0.395
ESM-2 (150M)	0.086	0.116	0.167
ESM-2 (650M)	0.100	0.146	0.223
PoET-Single (201M)	0.025	0.028	0.031
PoET-MSA (201M)	0.036	0.042	0.051
EvoDiff-MSA (100M)	0.061	0.089	0.168
DPLM (150M)	0.093	0.147	0.284
DPLM (650M)	0.102	0.159	0.303
ProfileBFN-Single (150M)	0.126	0.197	0.321
ProfileBFN-Single (650M)	0.162	0.262	0.422
ProfileBFN-Profile (150M)	0.128	0.210	0.384
ProfileBFN-Profile (650M)	**0.173**	**0.280**	**0.474**

Results: Visualization



Summary of Contribution

- Modeling protein in profile space
  rather than sequence space
- Deriving a new kind of BFN
  for family protein design
- Proposing evaluation metric
  that is more convincing

👈 Code

Paper 👉

Thank You!

Welcome to Join Us at Poster Session 15： 00

at #16 for In-depth Discussion!

Thank All Co-Authors’ Hardwork!

清華大學智能产业研究院

Institute for AI Industry Research，Tsinghua University

AlgoMole

GeoBFN

ICLR2024 Oral
Molecular

MolCRAFT

ICML2024 Poster
SBDD

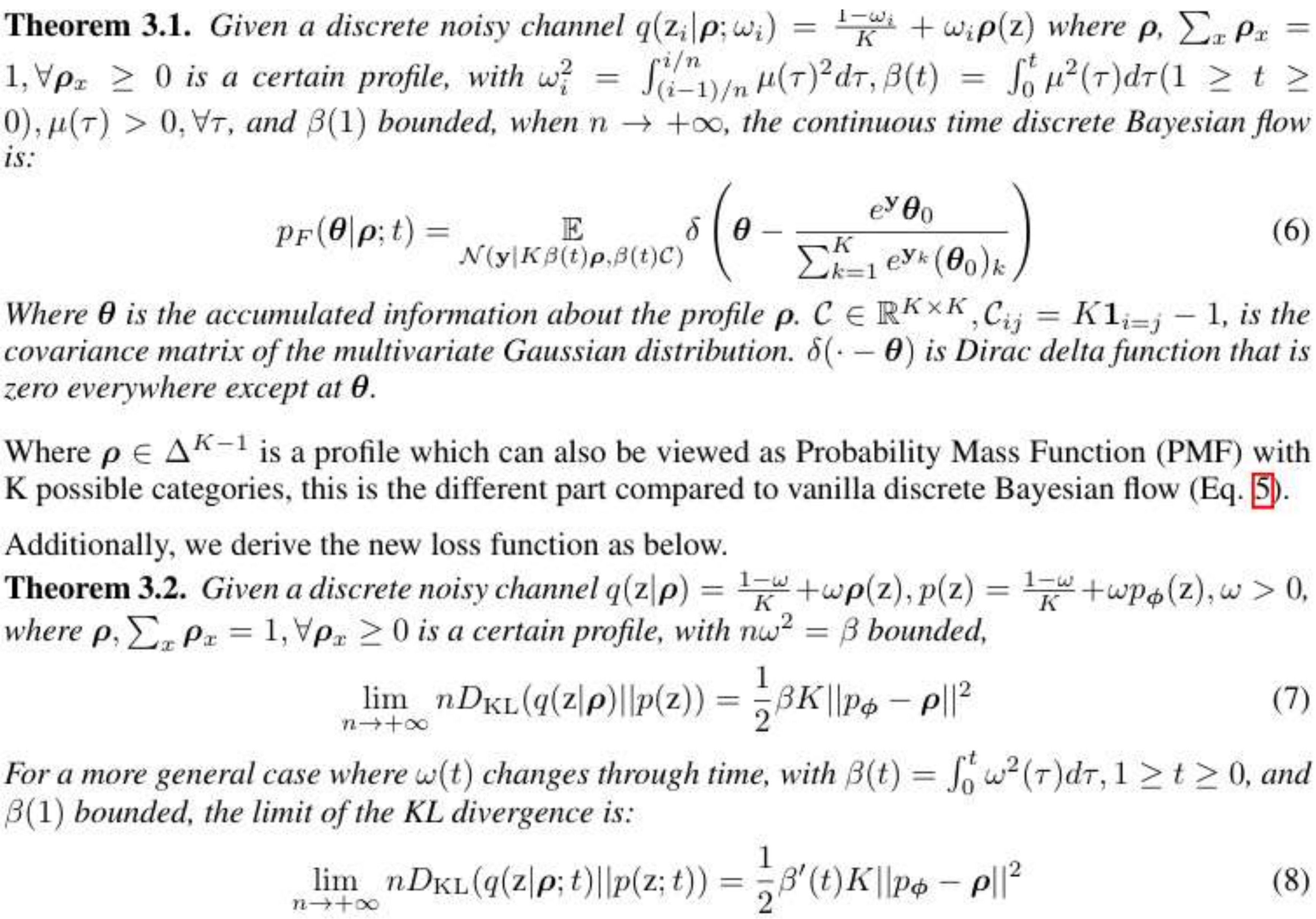
CysBFN

ICLR2025 Spotlight
Material

ProfileBFN

ICLR2025 Oral
Protein Family

Theorems



Theorem 3.1. *Given a discrete noisy channel* $q(\mathrm{z}_i|\boldsymbol{\rho};\omega_i) = \frac{1-\omega_i}{K} + \omega_i\boldsymbol{\rho}(\mathrm{z})$ *where* $\boldsymbol{\rho}$, $\sum_x \boldsymbol{\rho}_x = 1, \forall\boldsymbol{\rho}_x \geq 0$ *is a certain profile, with* $\omega_i^2 = \int_{(i-1)/n}^{i/n} \mu(\tau)^2 d\tau, \beta(t) = \int_0^t \mu^2(\tau)d\tau (1 \geq t \geq 0), \mu(\tau) > 0, \forall\tau$, *and* $\beta(1)$ *bounded, when* $n \to +\infty$, *the continuous time discrete Bayesian flow is:*

$$p_F(\boldsymbol{\theta}|\boldsymbol{\rho};t) = \mathop{\mathbb{E}}_{\mathcal{N}(\mathbf{y}|K\beta(t)\boldsymbol{\rho},\beta(t)\mathcal{C})} \delta\left(\boldsymbol{\theta} - \frac{e^{\mathbf{y}}\boldsymbol{\theta}_0}{\sum_{k=1}^K e^{\mathbf{y}_k}(\boldsymbol{\theta}_0)_k}\right) \quad (6)$$

Where $\boldsymbol{\theta}$ *is the accumulated information about the profile* $\boldsymbol{\rho}$. $\mathcal{C} \in \mathbb{R}^{K\times K}, \mathcal{C}_{ij} = K\mathbf{1}_{i=j} - 1$, *is the covariance matrix of the multivariate Gaussian distribution.* $\delta(\cdot - \boldsymbol{\theta})$ *is Dirac delta function that is zero everywhere except at* $\boldsymbol{\theta}$.

Where $\boldsymbol{\rho} \in \Delta^{K-1}$ is a profile which can also be viewed as Probability Mass Function (PMF) with K possible categories, this is the different part compared to vanilla discrete Bayesian flow (Eq. 5).

Additionally, we derive the new loss function as below.

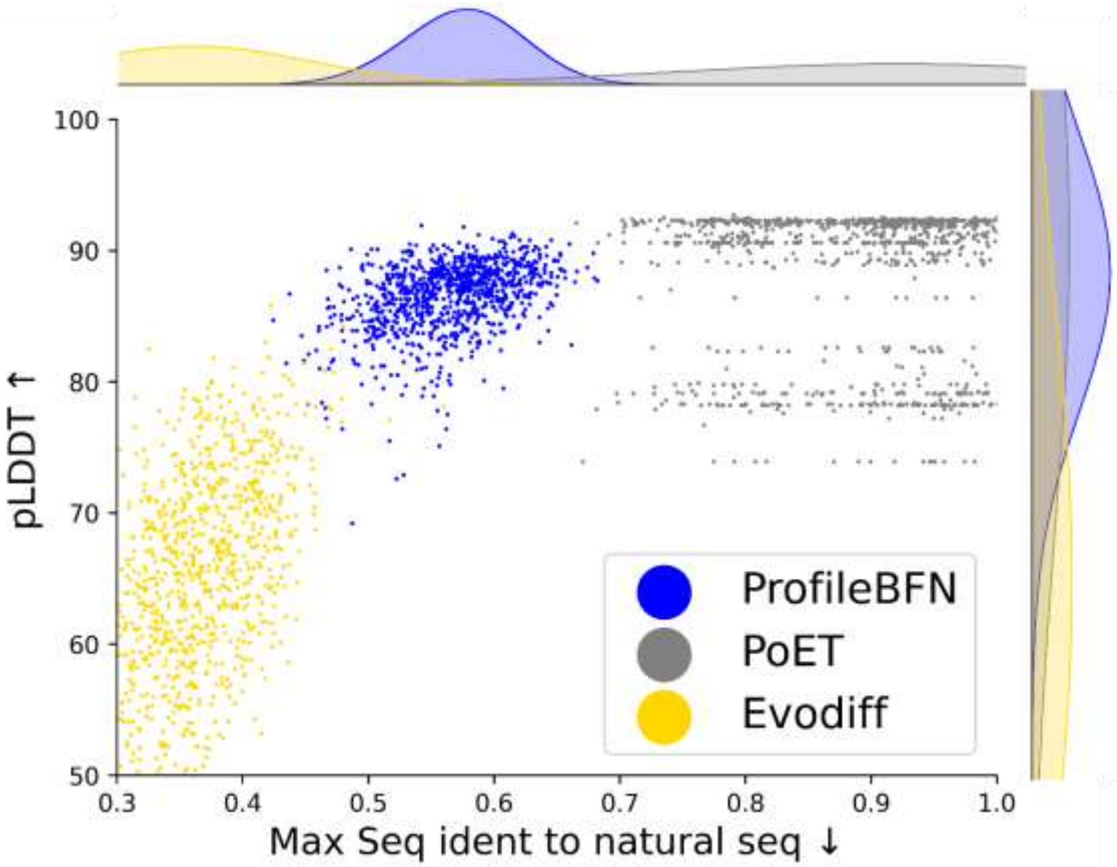
Theorem 3.2. *Given a discrete noisy channel* $q(\mathrm{z}|\boldsymbol{\rho}) = \frac{1-\omega}{K} + \omega\boldsymbol{\rho}(\mathrm{z}), p(\mathrm{z}) = \frac{1-\omega}{K} + \omega p_{\boldsymbol{\phi}}(\mathrm{z}), \omega > 0$, *where* $\boldsymbol{\rho}, \sum_x \boldsymbol{\rho}_x = 1, \forall\boldsymbol{\rho}_x \geq 0$ *is a certain profile, with* $n\omega^2 = \beta$ *bounded,*

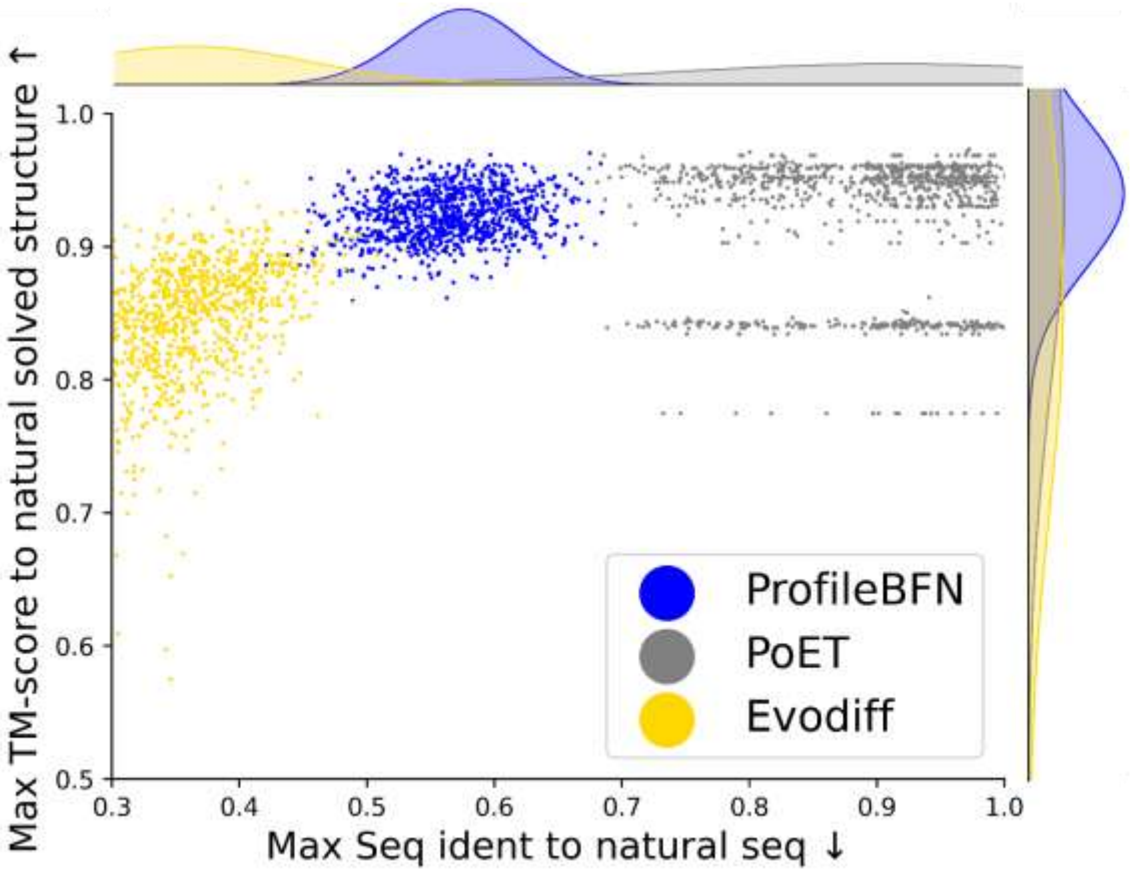
$$\lim_{n\to+\infty} nD_{\mathrm{KL}}(q(\mathrm{z}|\boldsymbol{\rho})||p(\mathrm{z})) = \frac{1}{2}\beta K||p_{\boldsymbol{\phi}} - \boldsymbol{\rho}||^2 \quad (7)$$

For a more general case where $\omega(t)$ *changes through time, with* $\beta(t) = \int_0^t \omega^2(\tau)d\tau, 1 \geq t \geq 0$, *and* $\beta(1)$ *bounded, the limit of the KL divergence is:*

$$\lim_{n\to+\infty} nD_{\mathrm{KL}}(q(\mathrm{z}|\boldsymbol{\rho};t)||p(\mathrm{z};t)) = \frac{1}{2}\beta'(t)K||p_{\boldsymbol{\phi}} - \boldsymbol{\rho}||^2 \quad (8)$$

Paramerized Results





Applications: Enzyme

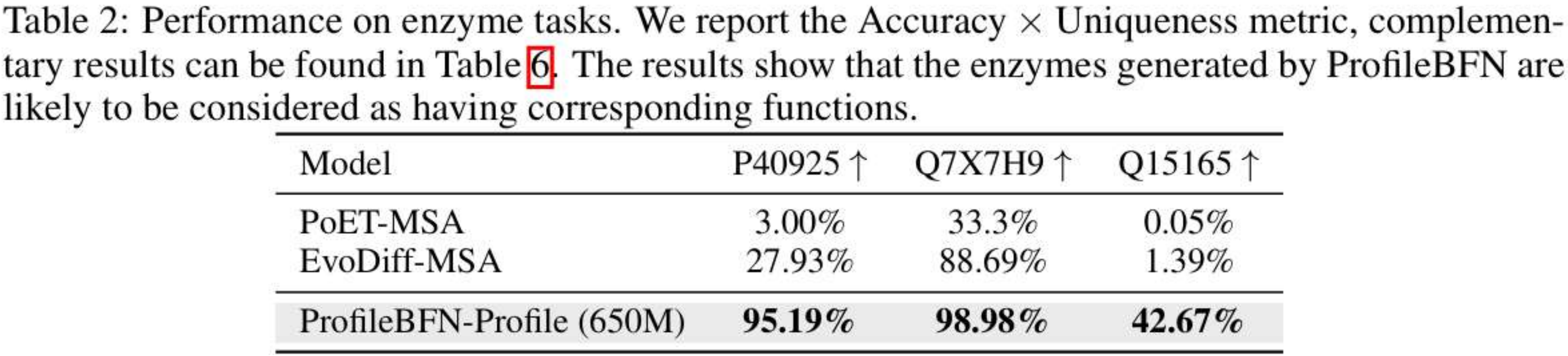
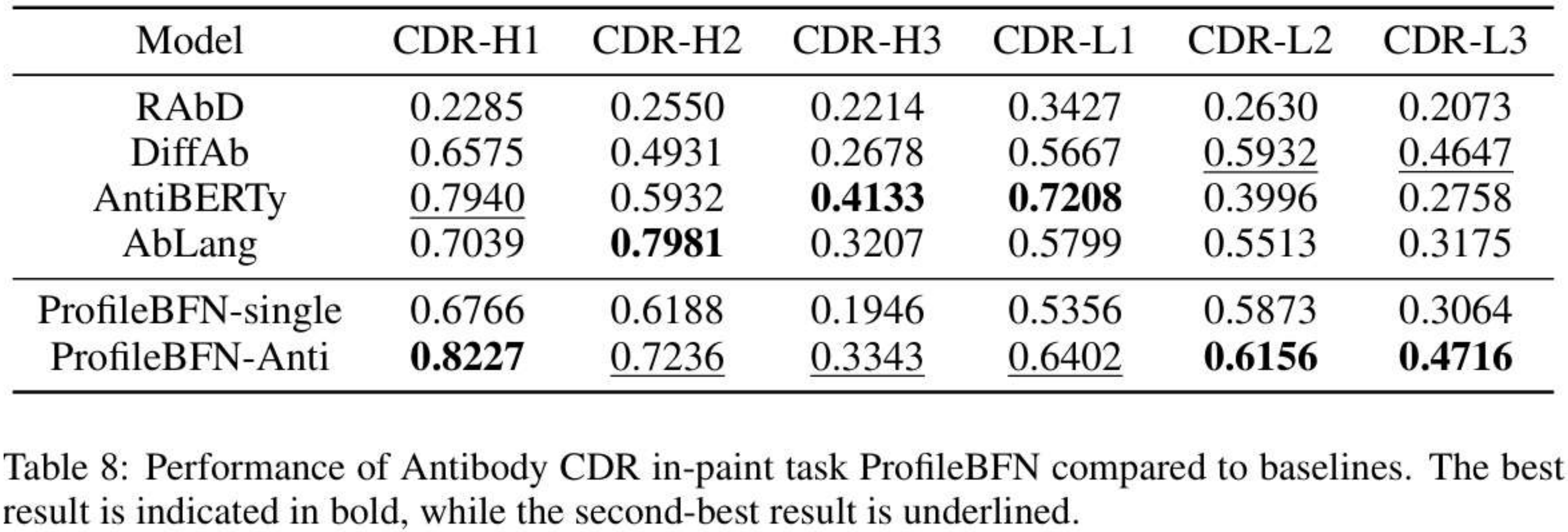


Table 2: Performance on enzyme tasks. We report the Accuracy × Uniqueness metric, complementary results can be found in Table 6. The results show that the enzymes generated by ProfileBFN are likely to be considered as having corresponding functions.

Model	P40925 ↑	Q7X7H9 ↑	Q15165 ↑
PoET-MSA	3.00%	33.3%	0.05%
EvoDiff-MSA	27.93%	88.69%	1.39%
ProfileBFN-Profile (650M)	**95.19%**	**98.98%**	**42.67%**

Applications: Antibody



Model	CDR-H1	CDR-H2	CDR-H3	CDR-L1	CDR-L2	CDR-L3
RAbD	0.2285	0.2550	0.2214	0.3427	0.2630	0.2073
DiffAb	0.6575	0.4931	0.2678	0.5667	0.5932	0.4647
AntiBERTy	0.7940	0.5932	**0.4133**	**0.7208**	0.3996	0.2758
AbLang	0.7039	**0.7981**	0.3207	0.5799	0.5513	0.3175
ProfileBFN-single	0.6766	0.6188	0.1946	0.5356	0.5873	0.3064
ProfileBFN-Anti	**0.8227**	0.7236	0.3343	0.6402	**0.6156**	**0.4716**

Table 8: Performance of Antibody CDR in-paint task ProfileBFN compared to baselines. The best result is indicated in bold, while the second-best result is underlined.

Applications: Folding

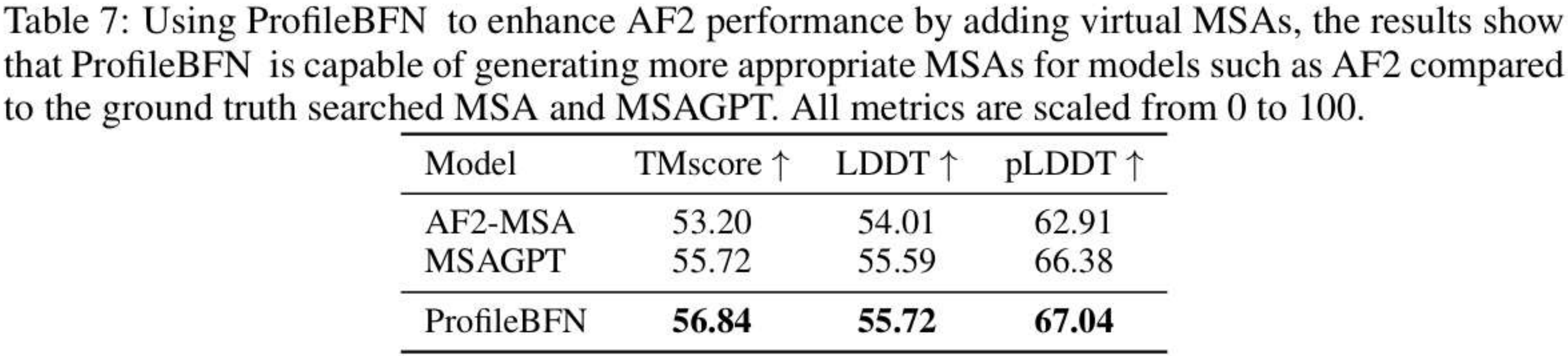
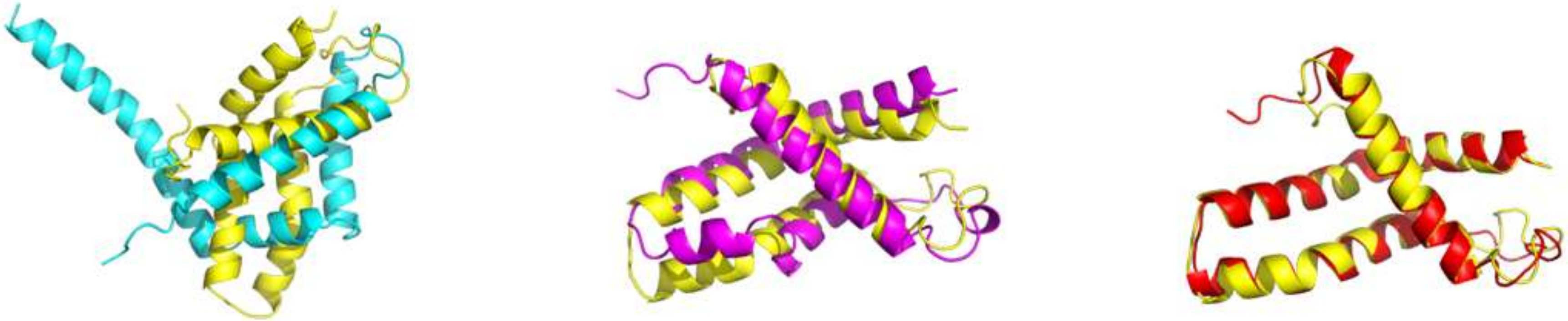


Table 7: Using ProfileBFN to enhance AF2 performance by adding virtual MSAs, the results show that ProfileBFN is capable of generating more appropriate MSAs for models such as AF2 compared to the ground truth searched MSA and MSAGPT. All metrics are scaled from 0 to 100.

Model	TMscore ↑	LDDT ↑	pLDDT ↑
AF2-MSA	53.20	54.01	62.91
MSAGPT	55.72	55.59	66.38
ProfileBFN	**56.84**	**55.72**	**67.04**

T1033



TM-score = 0.385
pLDDT = 52.39

TM-score = 0.617
pLDDT = 53.10

TM-score = 0.842
pLDDT = 76.63

Representation Learning

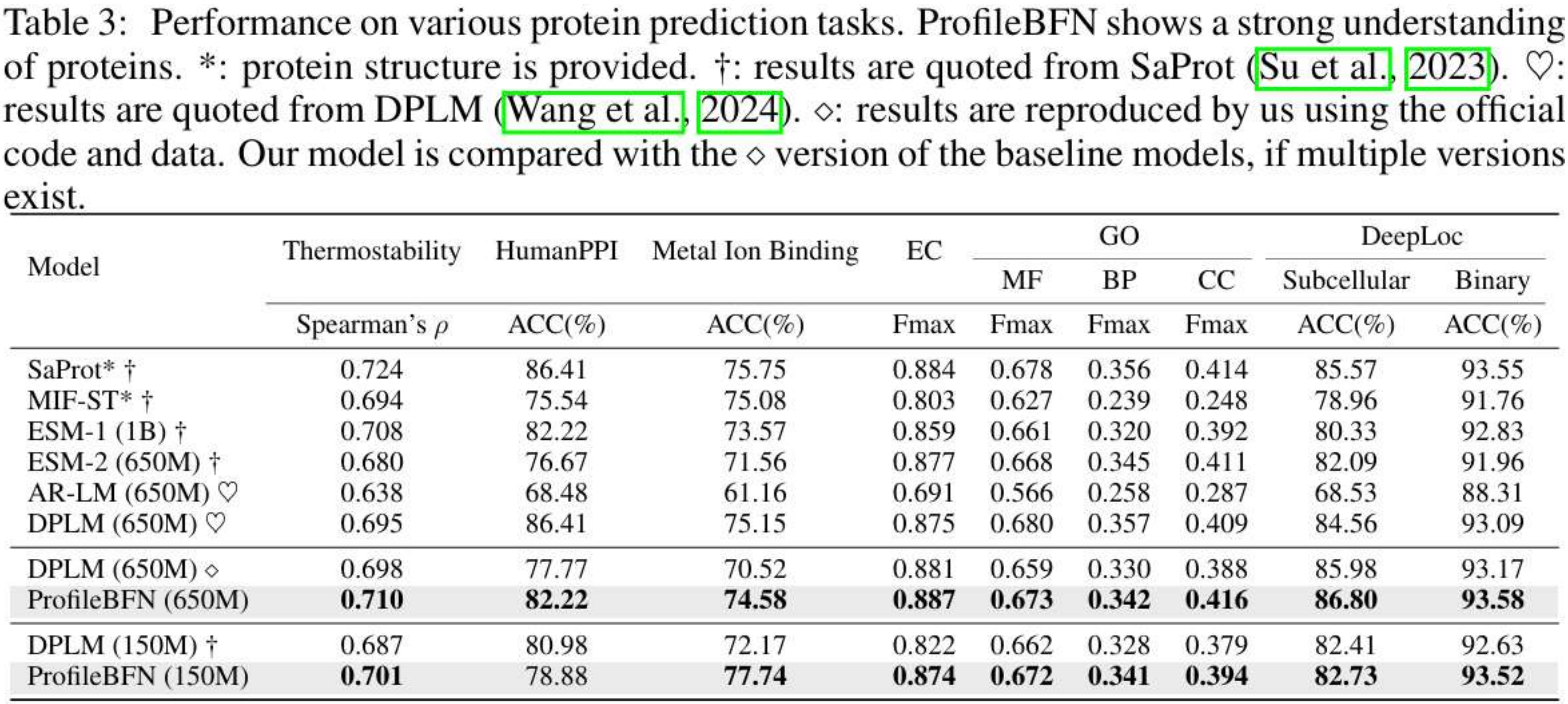
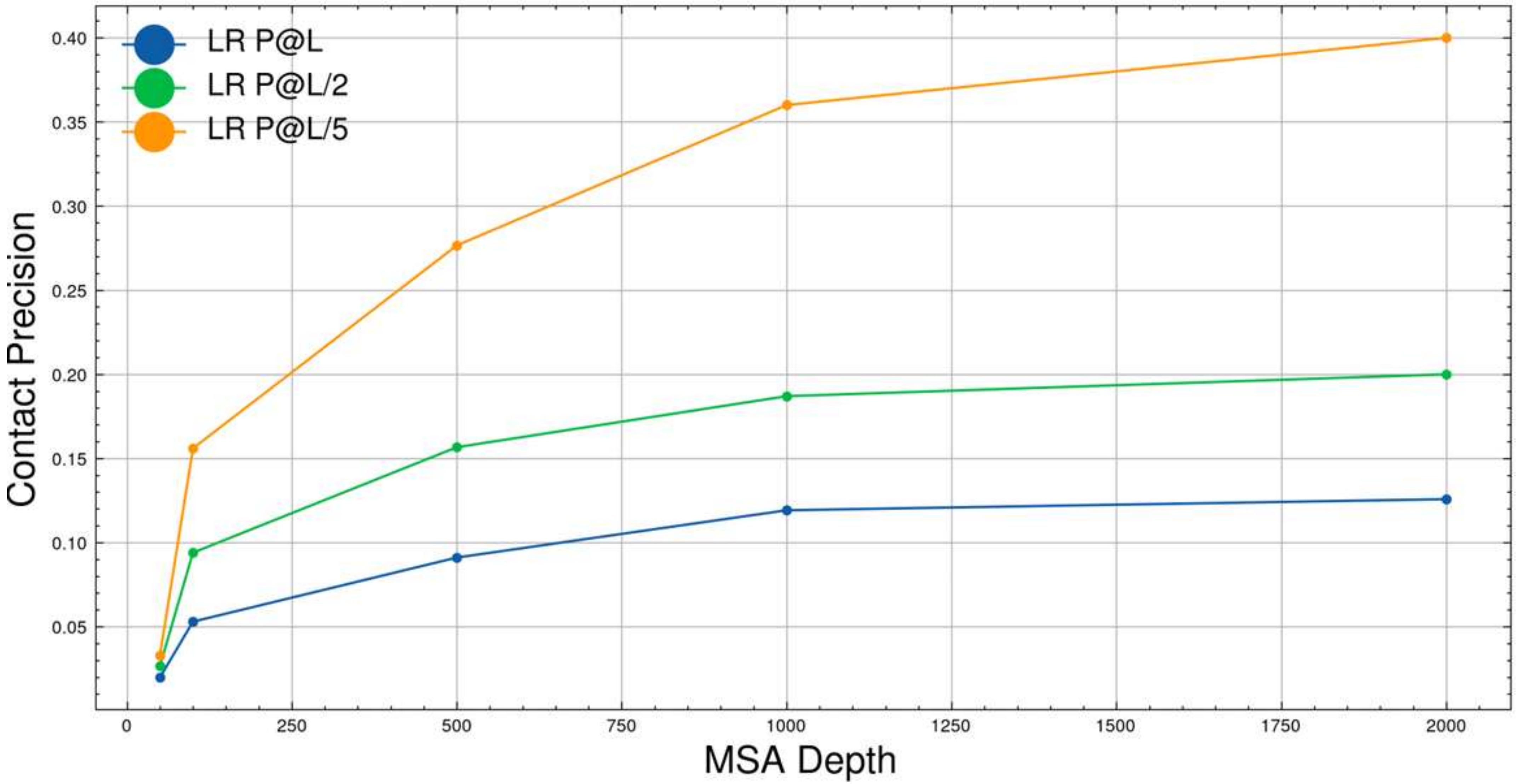


Table 3: Performance on various protein prediction tasks. ProfileBFN shows a strong understanding of proteins. *: protein structure is provided. †: results are quoted from SaProt (Su et al., 2023). ♡: results are quoted from DPLM (Wang et al., 2024). ◇: results are reproduced by us using the official code and data. Our model is compared with the ◇ version of the baseline models, if multiple versions exist.

Model	Thermostability	HumanPPI	Metal Ion Binding	EC	GO			DeepLoc	
					MF	BP	CC	Subcellular	Binary
	Spearman's ρ	ACC(%)	ACC(%)	Fmax	Fmax	Fmax	Fmax	ACC(%)	ACC(%)
SaProt* †	0.724	86.41	75.75	0.884	0.678	0.356	0.414	85.57	93.55
MIF-ST* †	0.694	75.54	75.08	0.803	0.627	0.239	0.248	78.96	91.76
ESM-1 (1B) †	0.708	82.22	73.57	0.859	0.661	0.320	0.392	80.33	92.83
ESM-2 (650M) †	0.680	76.67	71.56	0.877	0.668	0.345	0.411	82.09	91.96
AR-LM (650M) ♡	0.638	68.48	61.16	0.691	0.566	0.258	0.287	68.53	88.31
DPLM (650M) ♡	0.695	86.41	75.15	0.875	0.680	0.357	0.409	84.56	93.09
DPLM (650M) ◇	0.698	77.77	70.52	0.881	0.659	0.330	0.388	85.98	93.17
ProfileBFN (650M)	**0.710**	**82.22**	**74.58**	**0.887**	**0.673**	**0.342**	**0.416**	**86.80**	**93.58**
DPLM (150M) †	0.687	80.98	72.17	0.822	0.662	0.328	0.379	82.41	92.63
ProfileBFN (150M)	**0.701**	78.88	**77.74**	**0.874**	**0.672**	**0.341**	**0.394**	**82.73**	**93.52**

MSA Depth


LR P@L
LR P@L/2
LR P@L/5
Contact Precision
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0
250
500
750
1000
1250
1500
1750
2000
MSA Depth