Annual Summary of 2021

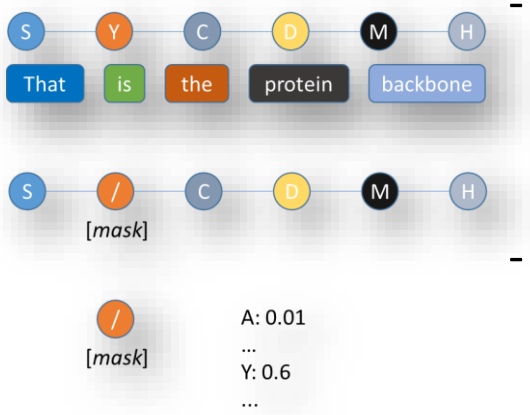
Yijia Xiao

3/22/2022

Protein-LM Large-scale protein language models (KDD2021 workshop)

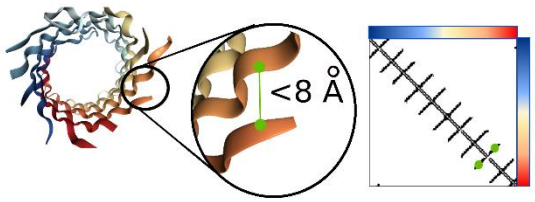
Motivation

- Language and protein are similar
  - Governed by intrinsic law
  - Linguistics / life code
- Success in large NLP models
  - Bert / GPT3 / Switch Transformer
- Apply *language model* to *protein*!



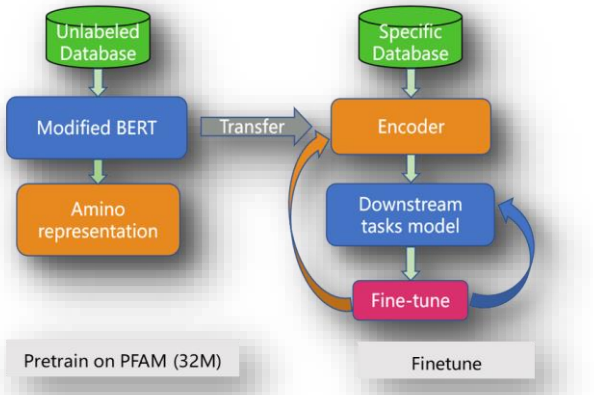
Related works

- TAPE (UC Berkeley)
  - Evaluating Protein Transfer Learning with TAPE
  - A series of downstream tasks
- ESM (Facebook)
  - Evolutionary Scale Modeling
  - Pretrained language models with 1 billion parameters



Methods

- Pre-train
  - Capture universal protein representation from unlabeled data
  - Masked language model: mask & predict multiple positions
- Fine-tune
  - Supervised on labeled data
  - Transfer to downstream tasks
- Alleviate need of annotation



Results

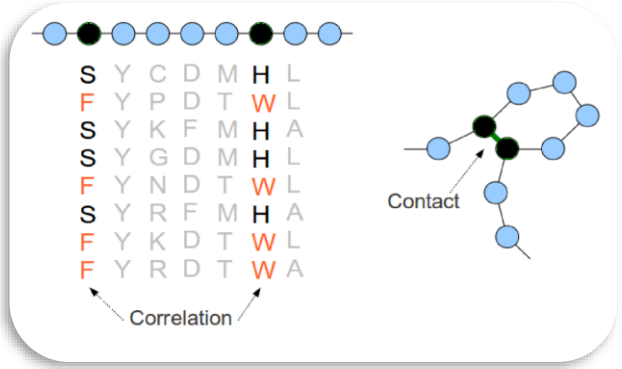
- Improved performance on 4 downstream tasks
- Paper (4 citations); Github (60+ stars)

Task	Metric	TAPE	Protein-LM (3B)
contact pred	P@L/5	0.36	**0.75**
remote homology	Accuracy	0.21	**0.30**
2-nd structure	Accuracy	0.73	**0.79**
fluorescence	Spearman'r	0.68	0.68
stability	Spearman'r	0.73	**0.79**

Megatron-MSA Cracking the Grammar of Protein with MSA

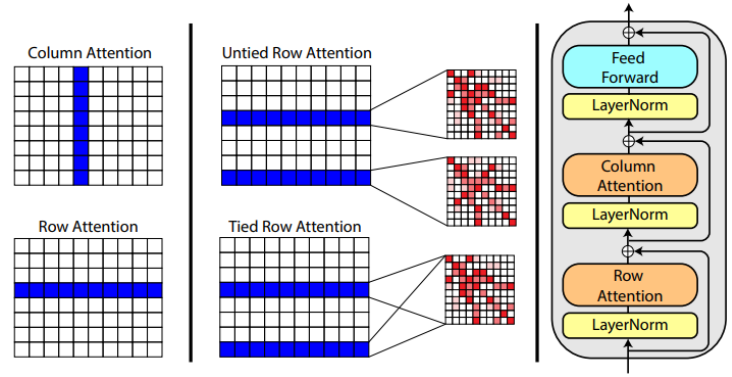
Motivation

- From individual to multiple
  - MSA: multiple sequence
  - Co-evolutionary info
- Sequence variation
  - Spatial proximity
- Model alignments jointly!



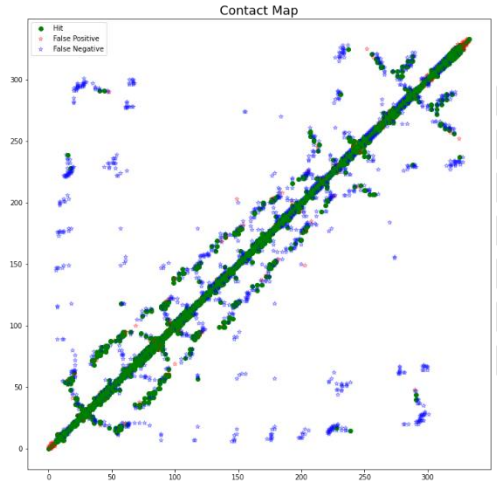
Related works

- MSA Transformer (Facebook)
  - Axial attention
  - Interleave row & col
  - Residual dependency



Methods

- Data preparation
  - HHBlits for constructing MSA database (1.5M samples)
- Training framework
  - Megatron-LM (NVIDIA, efficient LM training)
- Contact map
  - An intermediate representation between secondary and tertiary structures
  - Attention map ($L \times L$) is a good proxy of contact map



Results

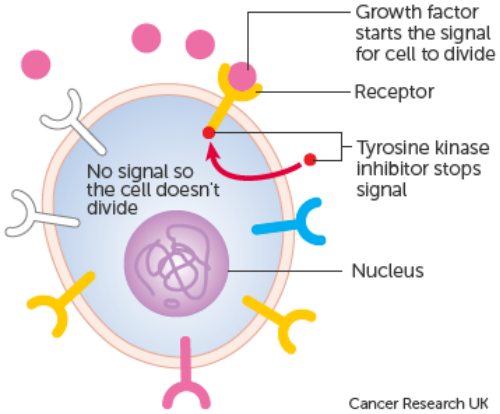
- Improved CC@CAMEO

		CAMEO		
	Model	P@L	P@L/2	P@L/5
Baseline in MSA Transformer	ProtTrans-T5	25.3	\	42.6
	ESM-1b	30.7	\	52.3
	Potts	23.9	\	42.7
	Facebook baseline	43	51.3	59.6
Protein-MSA (ours)	Protein-MSA-1B	**46**	**54.7**	**63.1**

SPLD-ExtraTrees Predicting kinase inhibitor resistance (Briefings in Bioinformatics)

Motivation

- Protein mutations are common
  - Cause drug resistance
- Diseases and targeted curations
  - Suppress growth & division signal
  - The affinity changes ΔΔG: Δ of ΔG
- *Resistance* analysis via affinity!



Problem formulation

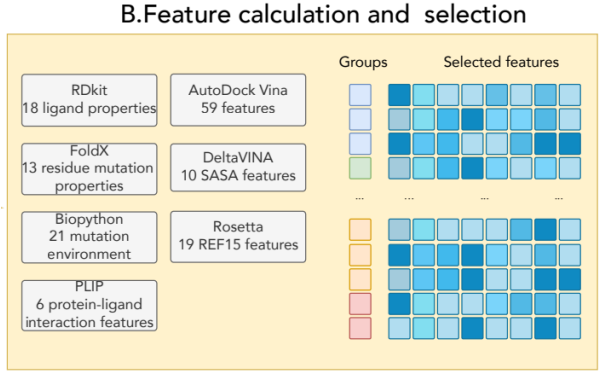
- Dataset: molecular features → ΔΔG (n samples)

$$\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n},\ \boldsymbol{x}_i \in \mathbb{R}^m, y_i \in \mathbb{R}$$

- ΔΔG

  Mutation → protein stability

- f_β : *Extra-Trees* model



Methods

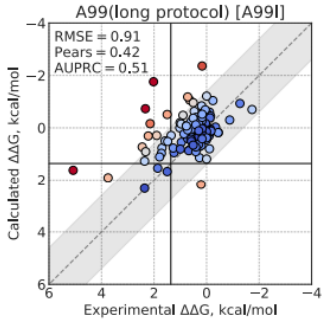
- Self-paced learning
  - Learn step by step, starting with easy samples
  - Latent variable $\boldsymbol{v} = [v_1, \cdots, v_n]$

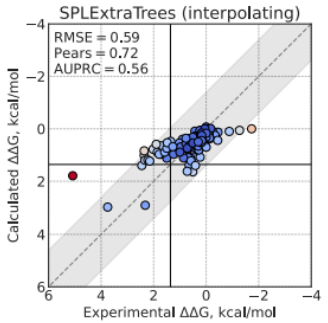
$$\min_{\boldsymbol{\beta}, \boldsymbol{v} \in [0,1]^n} \mathbf{E}(\boldsymbol{\beta}, \boldsymbol{v}; \lambda) = \sum_{i=1}^{n} v_i L\big(y_i, f(\boldsymbol{x}_i, \boldsymbol{\beta})\big) - \lambda \sum_{i=1}^{n} v_i$$

- Diversity reward
  - b protein groups $\quad \boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \in \mathbb{R}^{m \times n} \rightarrow \boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(b)}$

$$\boldsymbol{v} = \left[\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(b)}\right] \quad \boldsymbol{v}^{(j)} = \left(v_1^{(j)}, \ldots, v_{n_j}^{(j)}\right)^T \in [0,1]^{n_j}$$

$$\min_{\boldsymbol{\beta}} \min_{\boldsymbol{\beta}, \boldsymbol{v} \in [0,1]^n} \mathbb{E}(\boldsymbol{\beta}, \boldsymbol{v}; \lambda, \gamma) = \sum_{i=1}^{n} v_i L\big(y_i, f(\boldsymbol{x}_i, \boldsymbol{\beta})\big) - \lambda \sum_{i=1}^{n} v_i - \gamma \parallel \boldsymbol{v} \parallel_{2,1}$$





Results (Paper)

Table 2: Summary of the computational methods used, their calculation costs and performance. Mean prediction performance x_{lower}^{upper} over 20 repetitions are reported. The best results are highlighted in **bold**.

Abbreviation	Method		Force field or scoring function	Approximate cost per ΔΔG estimate		Performance		
				Hardware	Compute hours	RMSE (kcal/mol)	Pearson	AUPRC
A99[1]	Molecular Dynamics		Amber99sb*-ILDN and GAFF v2.1	10 CPU cores and 1 GPU	59	$0.91^{1.05}_{0.77}$	$0.44^{0.59}_{0.20}$	$\mathbf{0.56^{0.77}_{0.32}}$
A99*l*[1]	Molecular Dynamics		Amber99sb*-ILDN and GAFF v2.1	10 CPU cores and 1 GPU	98	$0.91^{1.09}_{0.74}$	$0.42^{0.59}_{0.20}$	$0.51^{0.75}_{0.27}$
REF15[2]	Rosetta		REF15	1 CPU core	32	$0.72^{0.83}_{0.60}$	$0.67^{0.81}_{0.45}$	$0.53^{0.74}_{0.29}$
ExtraTrees*[3]	ML	Scenario 1	n/a	1 CPU core	0.02	$0.87^{1.06}_{0.68}$	$0.12^{0.29}_{-0.04}$	$0.20^{0.39}_{0.10}$
SPLExtraTrees	ML		n/a	1 CPU core	0.02	$0.75^{0.77}_{0.75}$	$0.50^{0.54}_{0.38}$	$0.48^{0.52}_{0.34}$
SPLDExtraTrees	ML		n/a	1 CPU core	0.02	$0.73^{0.74}_{0.72}$	$0.54^{0.56}_{0.47}$	$0.50^{0.52}_{0.43}$
ExtraTrees	ML	Scenario 2	n/a	1 CPU core	0.02	$0.81^{0.89}_{0.66}$	$0.34^{0.54}_{0.22}$	$0.35^{0.47}_{0.22}$
SPLExtraTrees	ML		n/a	1 CPU core	0.02	$0.73^{0.80}_{0.53}$	$0.53^{0.65}_{0.38}$	$0.46^{0.57}_{0.35}$
SPLDExtraTrees	ML		n/a	1 CPU core	0.02	$0.70^{0.76}_{0.57}$	$0.60^{0.68}_{0.49}$	$0.55^{0.72}_{0.42}$
ExtraTrees*[3]	ML	Scenario 3	n/a	1 CPU core	0.02	$0.68^{0.80}_{0.55}$	$0.57^{0.72}_{0.34}$	$0.47^{0.68}_{0.25}$
SPLExtraTrees	ML		n/a	1 CPU core	0.02	$0.59^{0.61}_{0.58}$	$0.72^{0.73}_{0.70}$	$\mathbf{0.56^{0.57}_{0.51}}$
SPLDExtraTrees	ML		n/a	1 CPU core	0.02	$\mathbf{0.58^{0.59}_{0.57}}$	$\mathbf{0.74^{0.75}_{0.72}}$	$\mathbf{0.56^{0.60}_{0.52}}$

[1] Data for the molecular dynamic simulations with the A99 and A99*l* force field are obtained from the work in [4].
[2] Data for the Rosetta REF15 scoring function are obtained from the work in [4].
[3] Data for the ExtraTrees* are obtained from the work in [4].

Thanks for your time!