LEMON-ZEST
Evolution-Informed Tokenization for Efficient Protein Language Modeling

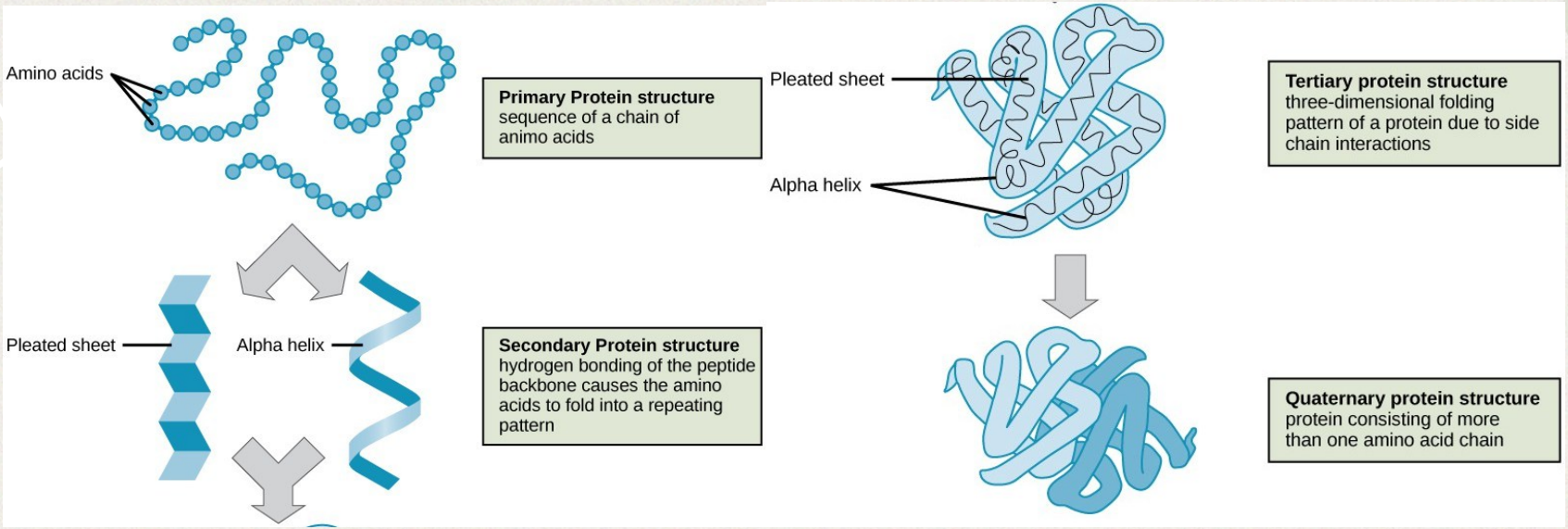
Biswajit Banerjee[1] • Claudia Alvarez-Carreño[2] • Anton S. Petrov[1]

[1]Georgia Institute of Technology • [2]University College London

01

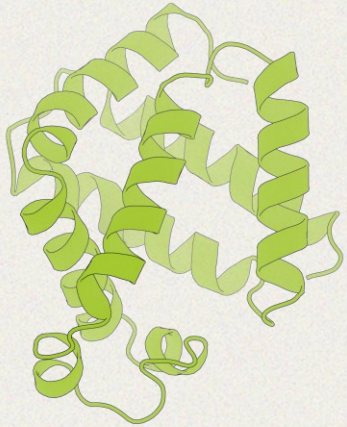
Protein Nomenclature

Introduction

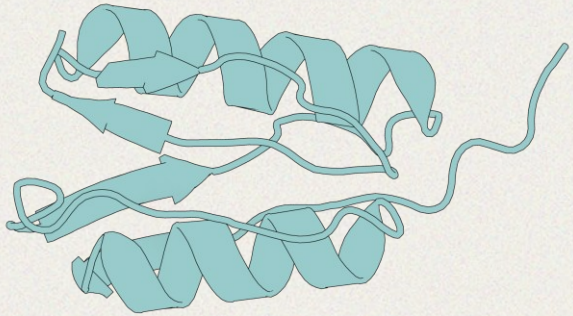


https://courses.lumenlearning.com/suny-dutchess-introbio2/chapter/proteins/

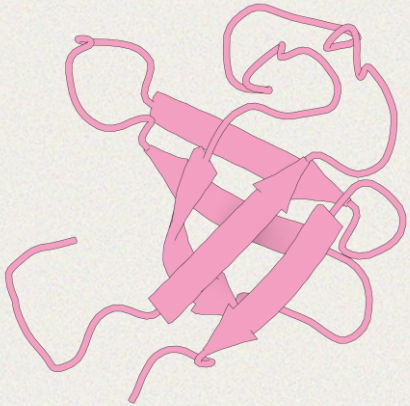
Fold Classification



Alpha Array



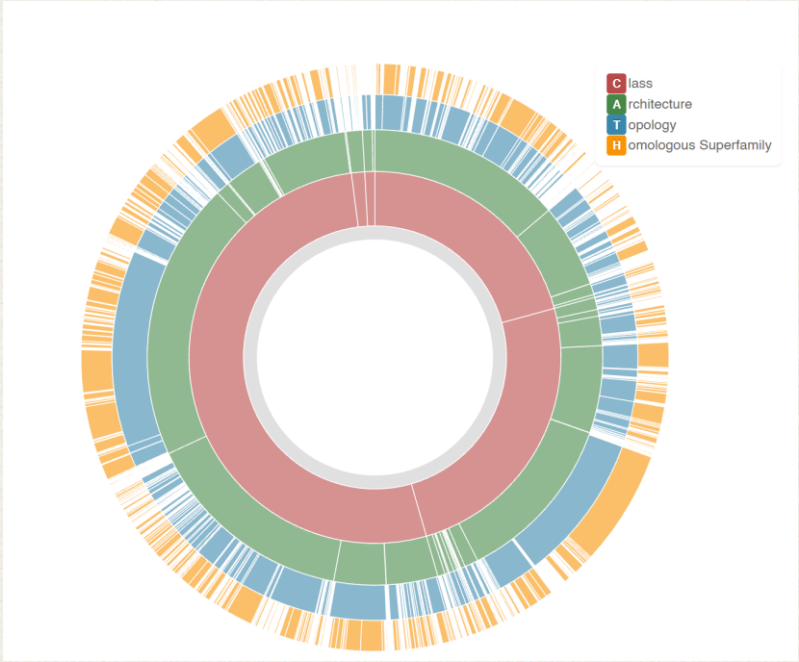
Alpha Beta 3-layerd sandwich



Beta Barrels

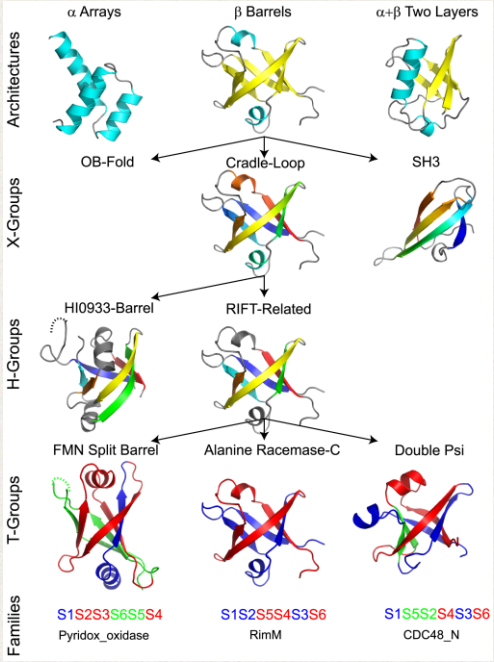
https://3dproteinimaging.com

Fold Classification Systems



Classes, **A**rchitectures, **T**opologies and **H**omologous Superfamilies.

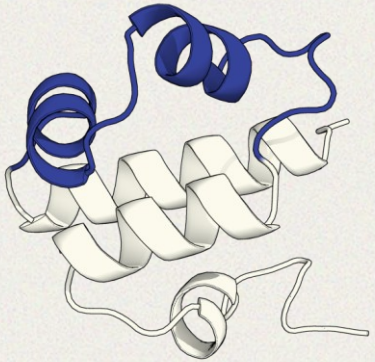
https://www.cathdb.info/browse/sunburst



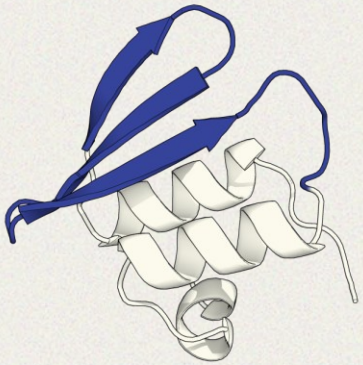
Evolutionary **C**lassification **O**f Protein **D**omains

Alignment and Homology

```
3b7hA00       TDGEFVSEHLMELITQQNLTINRVATLAGLNQSTV----.-NAMFEGRSKRPTITTIRKVCGTLGISVHDFFD
Consensus     ~~~d~~~~Ri~eL~~~rn~TiNrLAtlaGl~qSTi----.-~~~~~~~~t~~PK~~TLr~ic~~LgiSv~dFld
              .-..-|..||+++.++|++++...+...-=|-.|+       ..+++.+.+.-|.-|++++|..||++-.||.+
Consensus     ~~~~Ef~rrl~rl~~~rG~~~~~~~~rGKGSH~~l~~g~.r~Tvv~~~k~El~~Gt~~amLkqLGlt~~dl~~
1whzA00       PRPEEVARKLRRLG--------FVERMAKGGHRLYTHPDgRIVVVPFHSGELPKGTFKRILRDAGLTEEEFHN
Confidence    34456789999999999999998888774444444       4567777888899999999999999875
```



3b7hA00
1.12.260.40



1whzA00
3.30.920.30

02

Amino Acids Vocabulary

Amino Acids Mutability

- 20 major Amino acids
- Divided into 6 major categories, in order – Special, Hydrophobic, Polar, Basic, Aromatic, Acidic.
- The Substitution matrix shows the likeliness of AA substitution in folds.
- Using dynamic programming the sequences can be aligned
- Hence multiple sequences can fold to similar structure

	C	S	T	A	G	P	D	E	Q	N	H	R	K	M	I	L	V	W	Y	F	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
A	0	1	0	4																	A
G	-3	0	-2	0	6																G
P	-3	-1	-1	-1	-2	7															P
D	-3	0	-1	-2	-1	-1	6														D
E	-4	0	-1	-1	-2	-1	2	5													E
Q	-3	0	-1	-1	-2	-1	0	2	5												Q
N	-3	1	0	-2	0	-2	1	0	0	6											N
H	-3	-1	-2	-2	-2	-2	-1	0	0	1	8										H
R	-3	-1	-1	-1	-2	-2	-2	0	1	0	0	5									R
K	-3	0	-1	-1	-2	-1	-1	1	1	0	-1	2	5								K
M	-1	-1	-1	-1	-3	-2	-3	-2	0	-2	-2	-1	-1	5							M
I	-1	-2	-1	-1	-4	-3	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-1	-4	-3	-4	-3	-2	-3	-3	-2	-2	2	2	4					L
V	-1	-2	0	0	-3	-2	-3	-2	-2	-3	-3	-3	-2	1	3	1	4				V
W	-2	-3	-2	-3	-2	-4	-4	-3	-2	-4	-2	-3	-3	-1	-3	-2	-3	11			W
Y	-2	-2	-2	-2	-3	-3	-3	-2	-1	-2	2	-2	-2	-1	-1	-1	-1	2	7		Y
F	-2	-2	-2	-2	-3	-4	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	1	3	6	F
	C	S	T	A	G	P	D	E	Q	N	H	R	K	M	I	L	V	W	Y	F	

https://en.wikipedia.org/wiki/BLOSUM

Existing major PLMs use AA Level encoding

- Context size is limited to protein AA length
- Scales Poorly with the protein length
- At AA level sequences are noisy
- PLMs must re-learn amino acids biology

How much fold level information can be uncovered by sequence alone ?

LEMON demonstrate that biology can be encoded at tokenization level and efficiently recover fold level information.

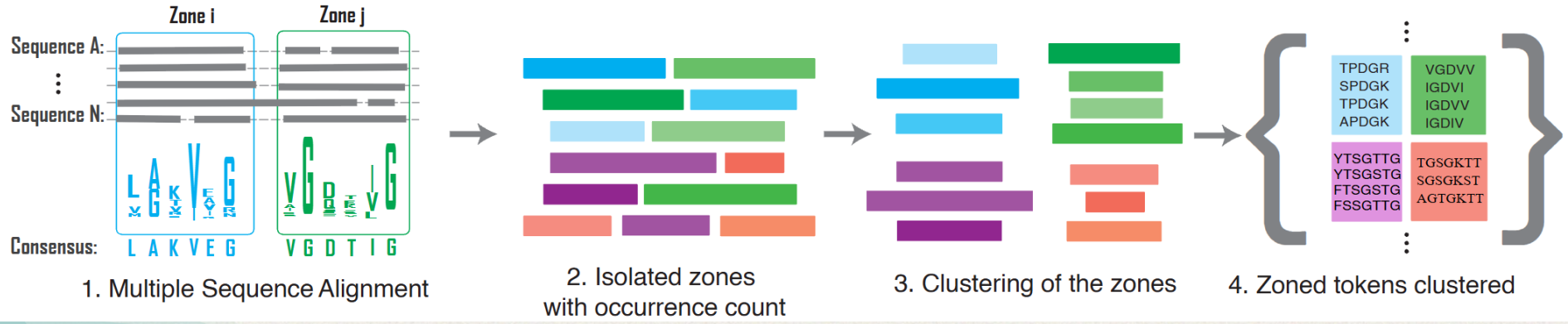
Opening a new direction for efficient, biologically-grounded protein representation learning.

03

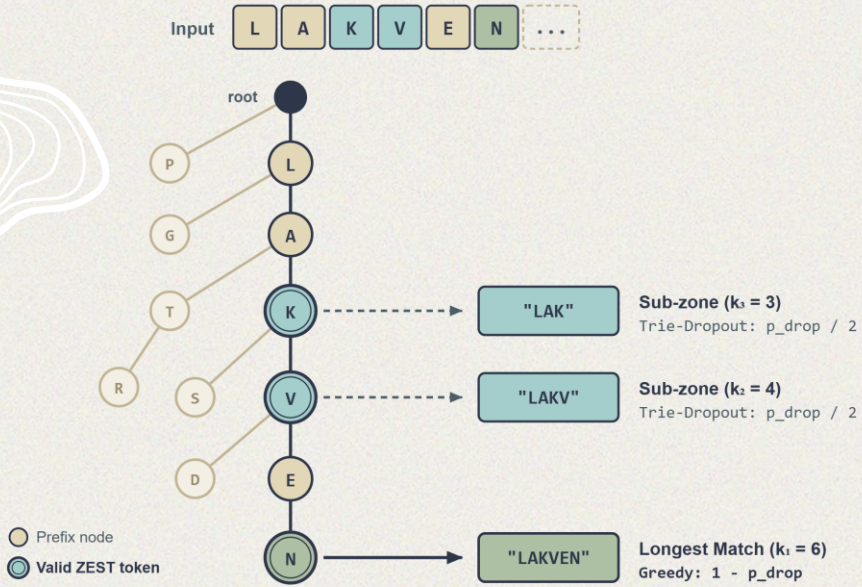
Methods

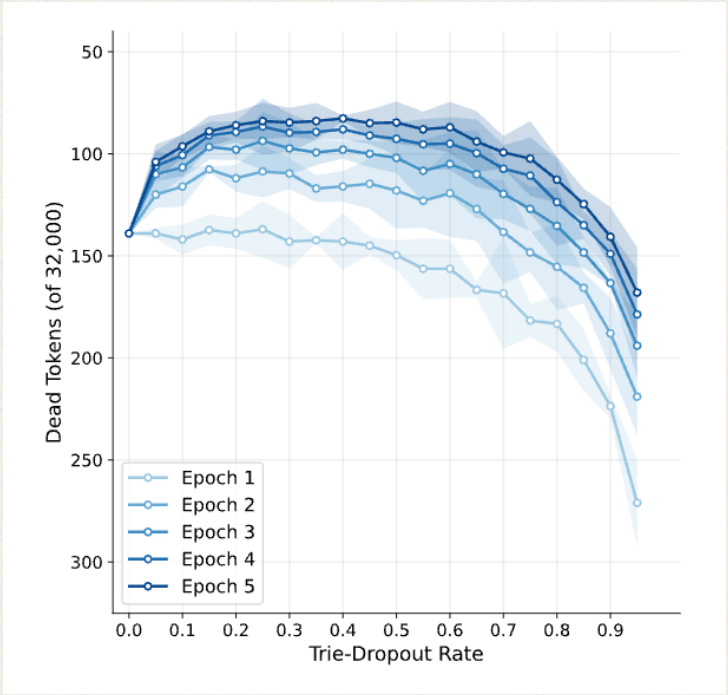
Zoned **E**ncoding of **S**equence **T**raits

```
uL2(SH3)    [1vy4:BD] 135:--FI-PVGTVVHAVELEP :149  159:--AGTSAQIQGRE-G----------DYVILRLPS----------GELRKVHGE--
uL24(SH3)   [1vy4:BY]   6:--HV-KKGDTVLVASGK- :19    21:--KGRVGKVKEVL-PK----------KYAVIVEGVN---IVKKAVRVSP-KYPQGGFIE---KEAPLHAS--
bL19(SH3)   [1vy4:BT]  19:LPEF-RPGDTVRVSYKVK :35    41:RIQDFEGIVIRIR-RN--------GFN-TTFTVRKVSYG----------VGVERIFPLH--
aL14(SH3)   [4v6u:B5]   3:--AI-DVGRIAVVIAGRR :17    18:--AGQKVVVVDII-D----------K--NFVLVTGAG--L----------N----------KVKRRRMNI---
aL21(SH3)   [4v6u:BR]  32:LQEF-EVGQRVHIVIEPS :48    59:RFHGRTGTVVGKR-G----------EAYIVEIKDGS----------KVKTLFIHPV--
aS4(SH3)    [4v6u:AE] 179:---F-EKGAYVFVTQG-- :190  193:--VARKGRIVEIK-RFP------MGWP-DVVTIEDEE----------GELFDTLKE--
bEF--P(SH3) [1ueb:A]    6:--DL-RPGTKVKMD---- :16    17:---GGLWECVEYQ-HQ---KLGRGG---AKVVAKFKN--LET----------GATVERTFN--
aIF5A(SH3)  [1iz6:A]   10:-SKL-KPGRYIIID---- :21    22:---DEPCRIVNIT-VS--SPGKHGS---AKARIEAVG----IFD----------GKVRSIVKP--
```



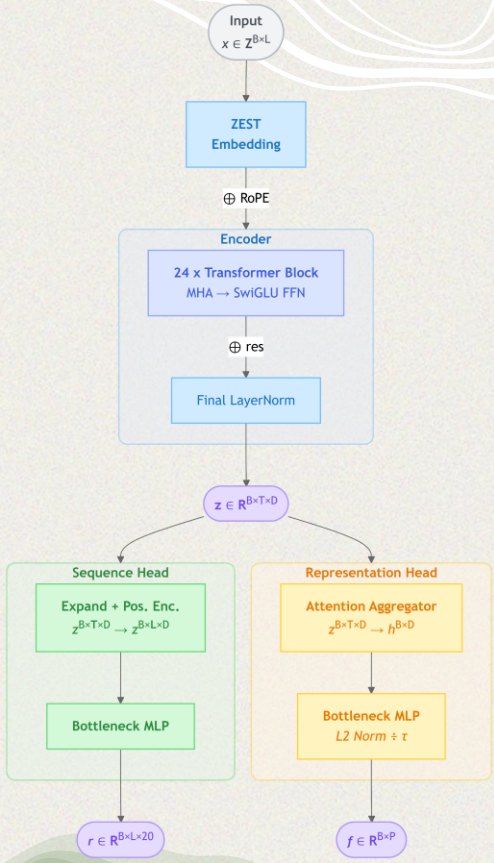
Trie-Drop out





Layered **E**xtraction of **M**olecular **O**rdering from **N**ature

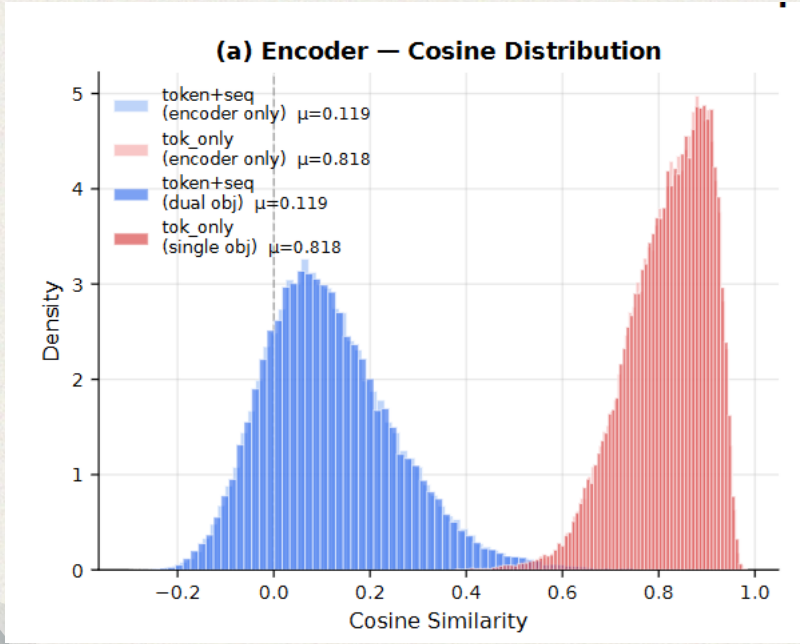
- Tokenizer compresses L to T
- Encoder enriches representation to (T x D)
- Sequence head Expands residue level to L
- Representation head projects to a P-dimentional space



04

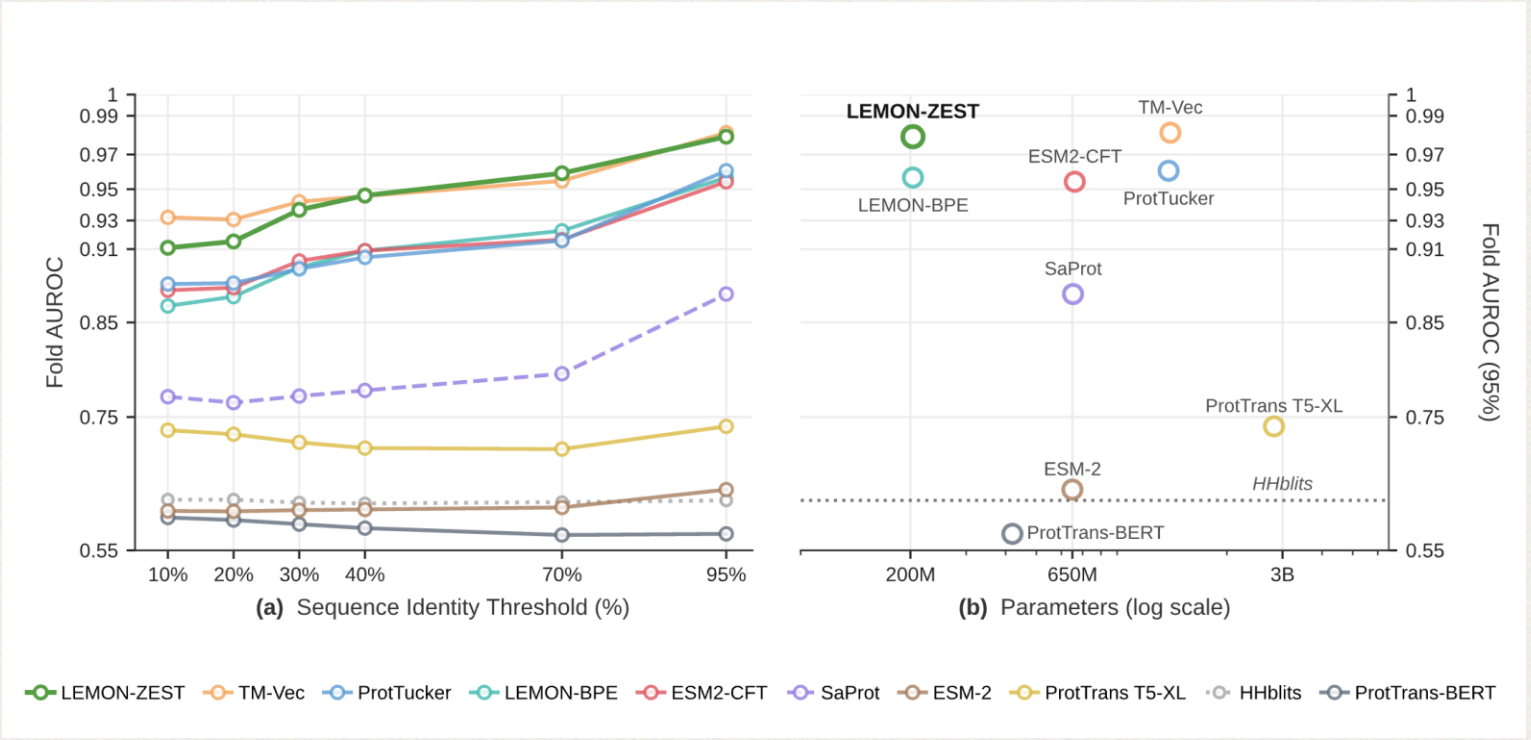
Results

Effect of Dual Head pretraining



Model	Level	Isotropy Metrics			Fold / Arch	
		$\bar{\cos}\downarrow$	$d_{\text{eff}}\uparrow$	$k_{90}\uparrow$	AUROC ↑	mAP ↑
Finetuned						
LEMON	Encoder	+0.119	43.0	183	0.758	0.173
	Contrastive-head	+0.019	25.4	29	**0.875**	**0.316**
Tok-only	Encoder	+0.818	19.5	119	0.678	0.117
	Contrastive-head	+0.229	8.6	9	0.825	0.192

Across Sequence Identity



Perfomance across length

Model	Length Bins	Mean(ms)	P95(ms)	Peak mem(GB)	GFLOPs
LEMON (204M)	100 – 499	1.80 ± 0.0	1.90 ± 0.0	0.44	24
	500 – 999	2.00 ± 0.0	2.10 ± 0.0	0.47	73
	1000 – 1999	2.80 ± 0.0	3.00 ± 0.0	0.53	159
ESM2 (150M)	100 – 499	4.00 ± 0.1	5.60 ± 0.1	0.49	72
	500 – 999	16.5 ± 0.1	18.6 ± 0.1	1.00	231
ESM2 (650M)	100 – 499	7.40 ± 0.1	10.5 ± 0.1	1.51	311
	500 – 999	26.8 ± 0.2	29.9 ± 0.1	2.07	972
ProstT5 (1.2B)	100 – 499	13.8 ± 0.2	18.3 ± 0.2	3.16	431
	500 – 999	50.8 ± 0.3	57.5 ± 0.1	5.48	1350
	1000 – 1999	60.5 ± 0.0	60.7 ± 0.0	5.63	2990

ESM2 models has a context window of 1022 residues

05

Discussion

Limitations

Finetuned with single domains

- TED domains are chopped from full proteins
- Behaviour on multi-domain proteins are not tested

Clustered Vocabulary

- Trades fine-grained detail for fold-level robustness.
- Subsequent experiments demonstrate AA-identities are preserved

Limited Scale

- Limited access to GPU hours
- The effectes of scaling is limited to 200M

Summary of Contributions

Protein Vocabulary

A natively many to one protein vocabulary.

Trie-Dropout

A novel mechanism to add dropout to protein sequences and have some test time gains

LEMON

A 200M compact model trained on the vocabulary having both fold level and residue level representation

Extensive Benchmarking

We show the perfomance across different classification systems and some zero shot benchmarks.

THANK YOU

Reach out for more information
sumon@gatech.edu