ProtGPT2

Dominik Grabarczyk

Parallels between Natural language and Proteins

NATURAL LANGUAGES	PROTEINS
Letters	Amino Acids
↓	↓
Words	Domains
↓	↓
Sentences	Proteins
↓	↓
Meaning	Function

Parallels between Natural language and Proteins

NATURAL LANGUAGES

A, B, C, D, E

↓

I, To Be, Student

↓

I am a student

↓

Description of occupation

PROTEINS

g, a, v, l, l, t, s

↓

lqvgqvelgg, agslqp

↓

lqvgqvelgggpgagslqpl

↓

Regulation of carbohydrate metabolism

Question

CAN NLP DO THIS?

What is ProtGPT2

- Protein LLM
- Transformer decoder
- Using many common practices from NLP
  - Decoding strategies
  - Tokenisation

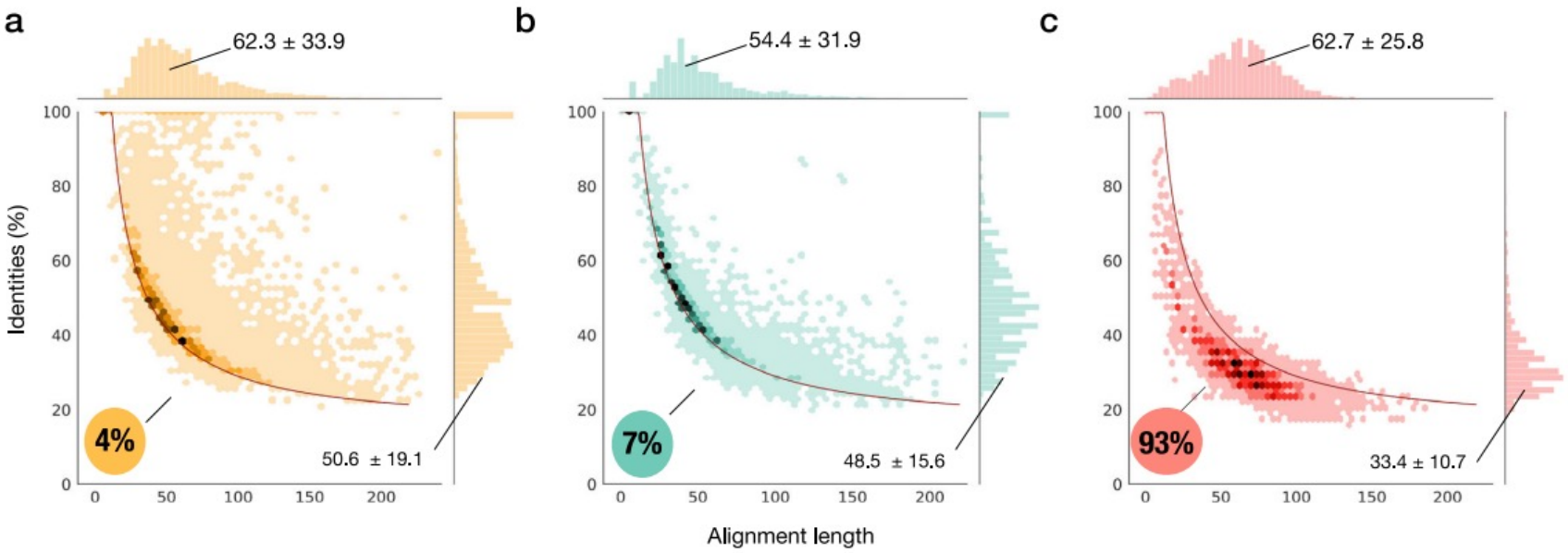
Did they succeed

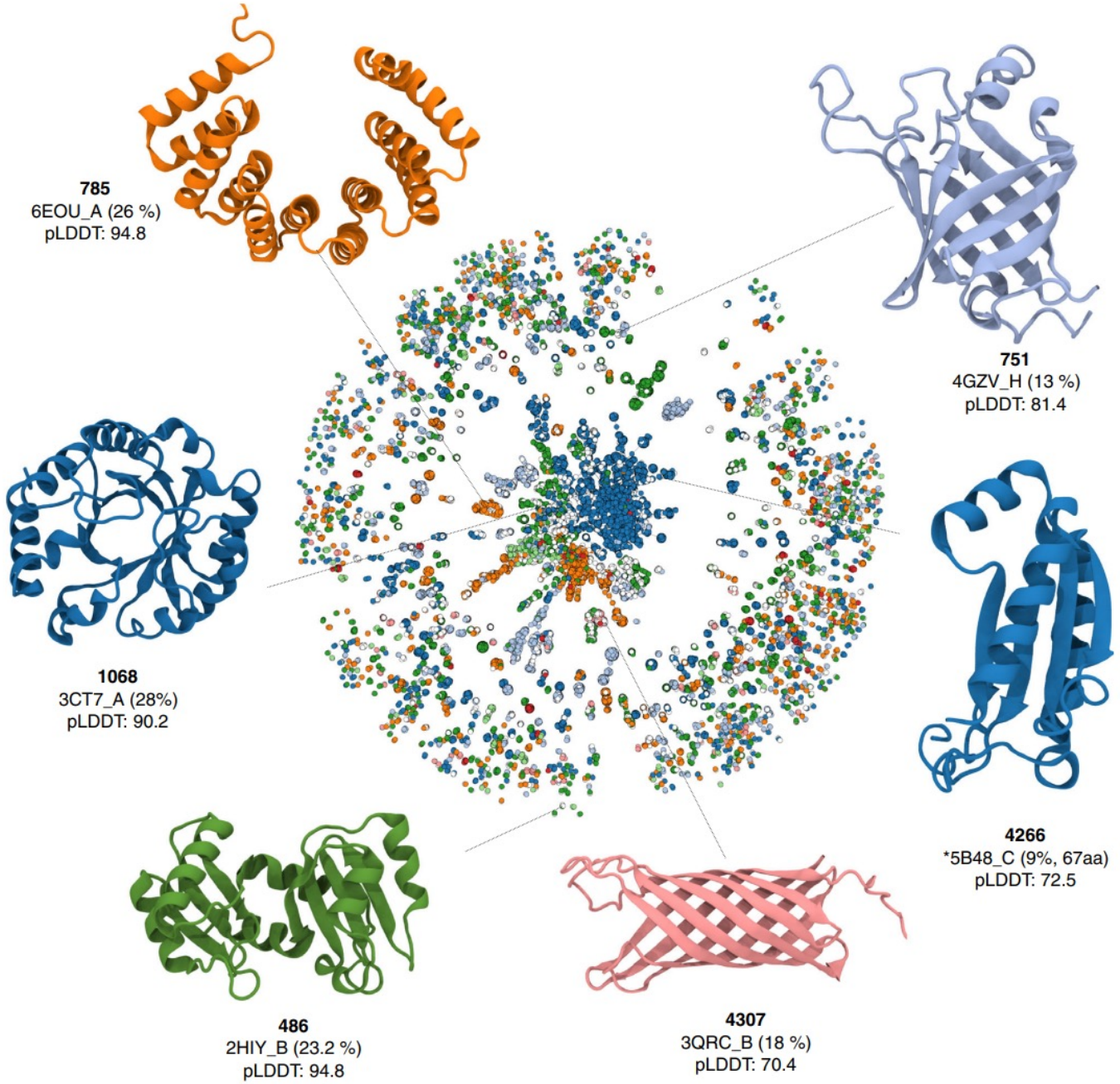
- Depends on your criteria?
  - No “unified” benchmark
  - No downstream tasks
  - BUT results are interesting

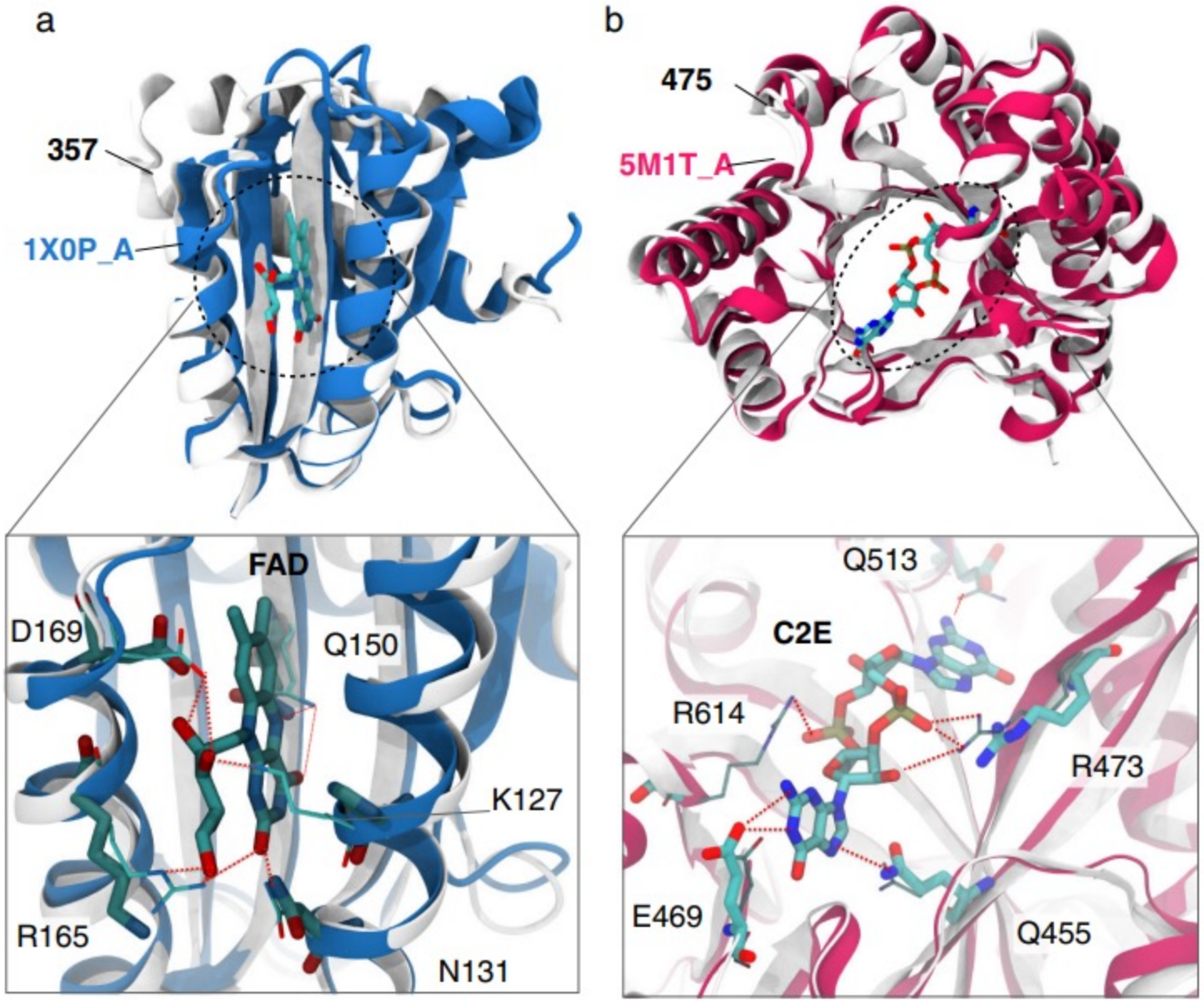
Table 1 | Disorder and secondary structure predictions of the natural and ProtGPT2 dataset

	Natural dataset	ProtGPT2 dataset
IUPred3 (globular domains)	88.40%	87.59%
Ordered content	79.71%	82.59%
Alpha-helical content	45.19%	48.64%
Beta-sheet content	41.87%	39.70%
Coil content	12.93%	11.66%

(*n* = 10,000 independent sequences/dataset).




785
6EOU_A (26 %)
pLDDT: 94.8
751
4GZV_H (13 %)
pLDDT: 81.4
1068
3CT7_A (28%)
pLDDT: 90.2
4266
*5B48_C (9%, 67aa)
pLDDT: 72.5
486
2HIY_B (23.2 %)
pLDDT: 94.8
4307
3QRC_B (18 %)
pLDDT: 70.4
a: All alpha proteins
b: All beta proteins
c: (α/β) proteins
d: (α + β) proteins
e: Multi-domain proteins
f: Membrane/cell surface proteins
g: Small proteins
ProtGPT2

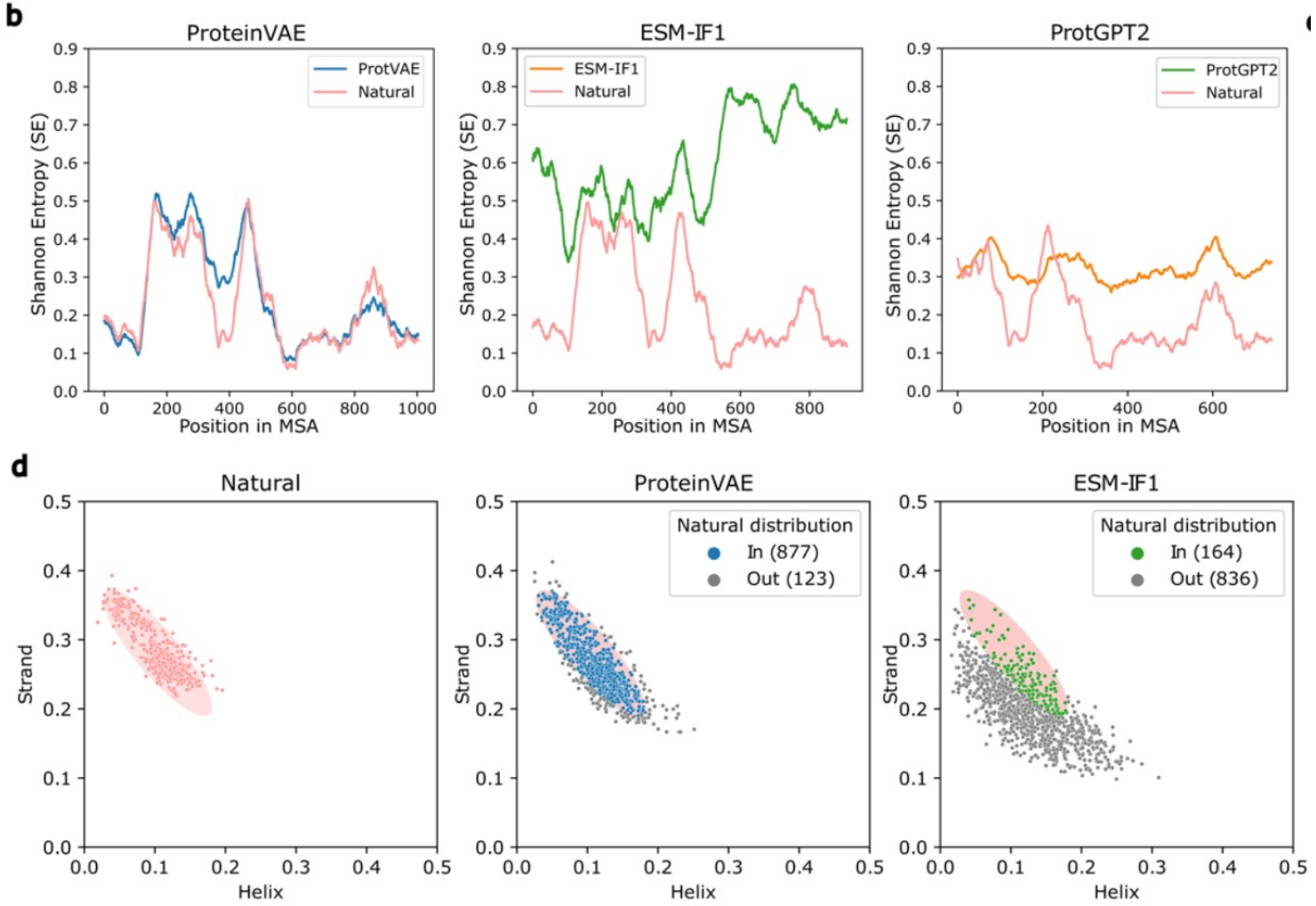

a
357
1X0P_A
FAD
D169
Q150
K127
R165
N131
b
475
5M1T_A
Q513
C2E
R614
R473
E469
Q455

What's the point

- Therapeutics development
  - Vaccines
  - Personalised medicine
- Structure and function prediction
- Genetic modification

What's next?

- Different architectures
- Robust and standardised validation methods
- Development of real-world applications



Lyu et al. 2023

Thank you