ICLR

On Pre-trained Language Models for Antibody

Danqing Wang, Fei Ye, Zhou Hao
ByteDance Research, Shanghai, China
University of California, Santa Barbara
Institute for AI Industry Research, Tsinghua University

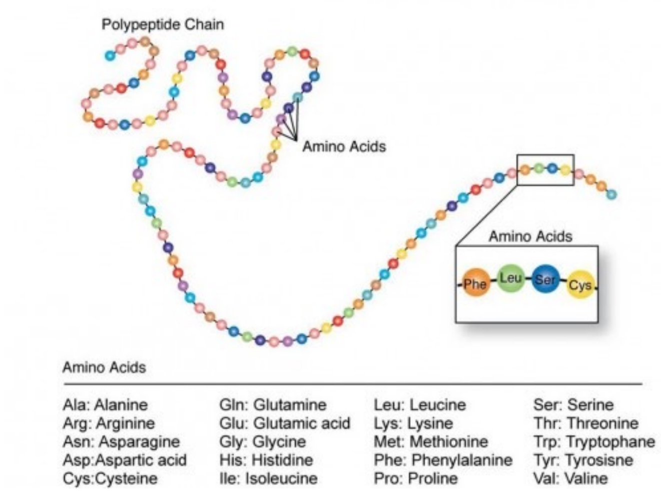
ByteDance AI Lab
字节跳动人工智能实验室

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

Protein & Antibody

❖ Protein

- Sequence composed of 20 amino acids



Amino Acids

Ala: Alanine
Arg: Arginine
Asn: Asparagine
Asp:Aspartic acid
Cys:Cysteine

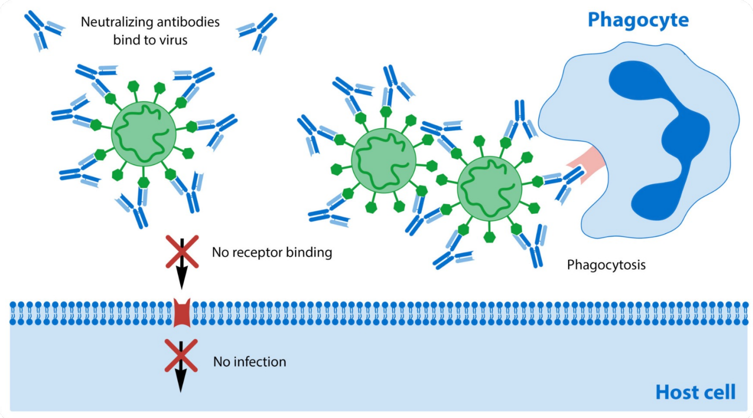
Gln: Glutamine
Glu: Glutamic acid
Gly: Glycine
His: Histidine
Ile: Isoleucine

Leu: Leucine
Lys: Lysine
Met: Methionine
Phe: Phenylalanine
Pro: Proline

Ser: Serine
Thr: Threonine
Trp: Tryptophane
Tyr: Tyrosisne
Val: Valine

❖ Antibody

- one type of therapeutic protein
- Y-shape to bind with virus



How to represent biological sequences?

Pretrained Language Models demonstrate remarkable achievements

- ❖ Pretrained Protein Language Models (PPLMs)
  - ➢ ESM (Rives et al., 2021)
  - ➢ MSA-Transformer (Rao et al., 2021)
  - ➢ ProtTrans (Elnaggar et al., 2021)
- ❖ Pretrained Antibody Language Models (PALMs)
  - ➢ Ablang (Olsen et al., 2022b)
  - ➢ AntiBERTy (Ruffolo et al., 2021)

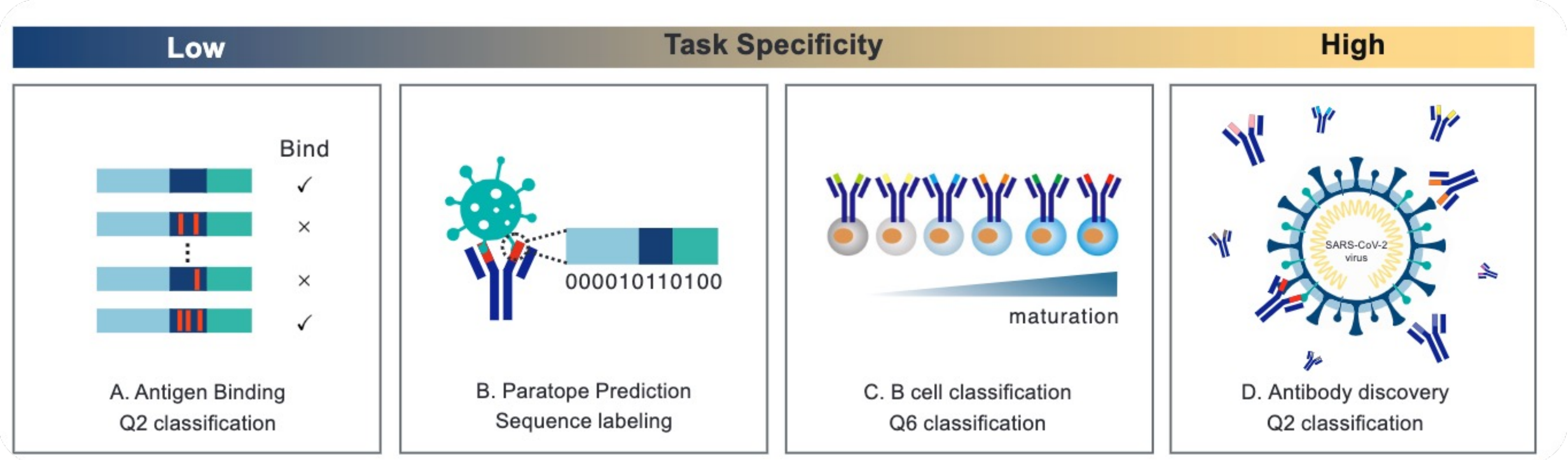
Q1: Can PPLMs directly be used for antibody tasks?

Q2: Are current PALMs highly related to real-world antibody discovery?

First, Real-world Antibody Discovery Tasks

=> a standard evaluation for antibody

- ❖ AnTibody Understanding Evaluation (ATUE)



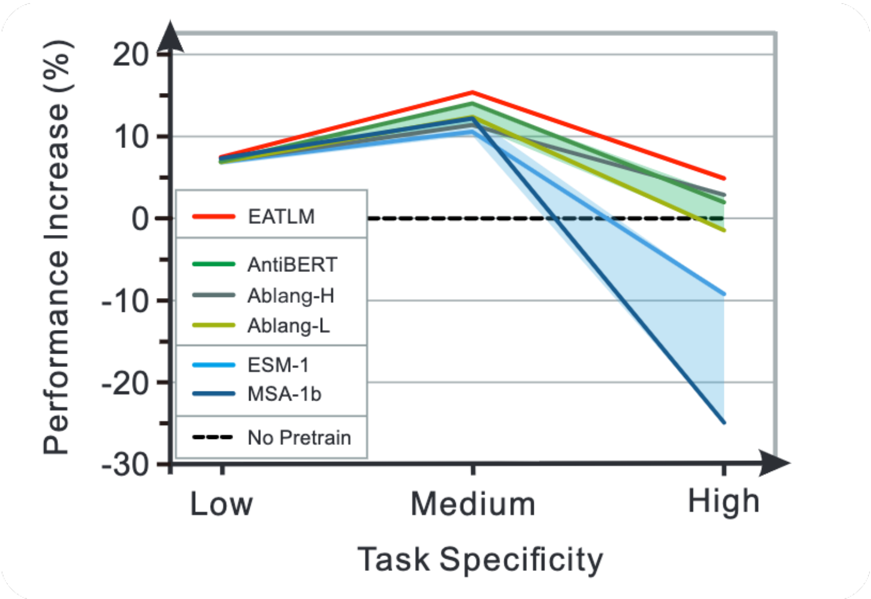
Key Observations on ATUE

- Low antibody-specificity
  - PPLMs perform similarly to PALMs
- Medium specificity
  - PALMs > PPLMs
- High specificity
  - PALMs are not enough

PPLMs can only solve low specificity tasks

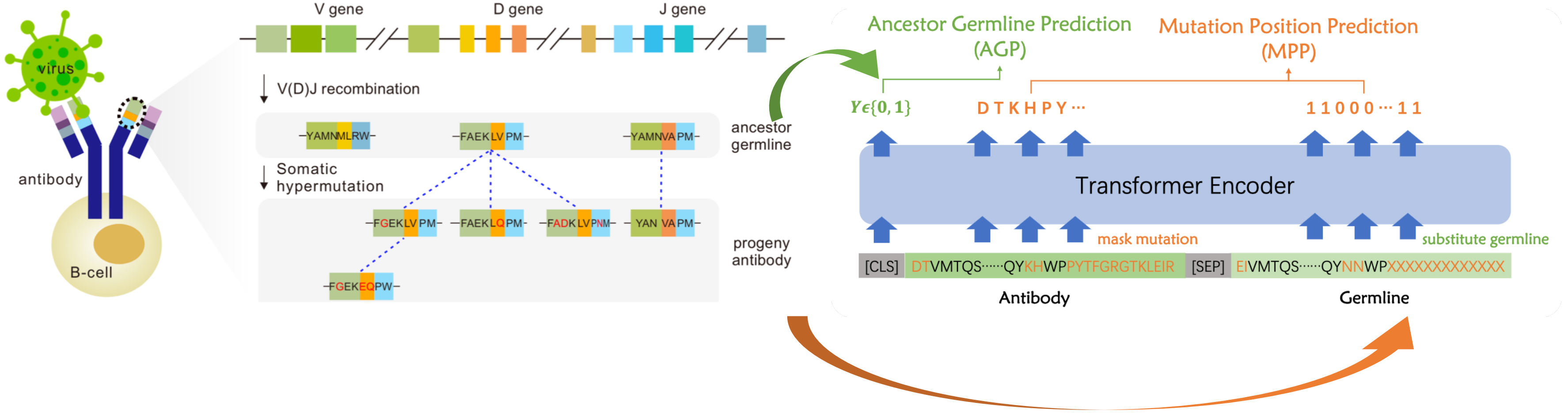
Current PALMs are not good antibody discovers

Opposite evolution mechanism



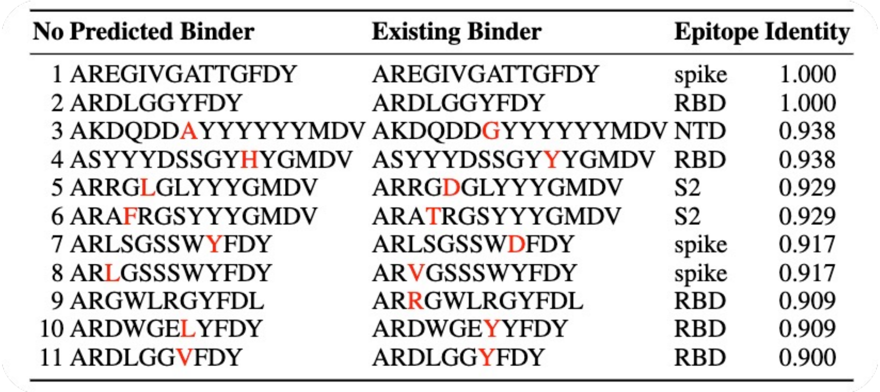
Secret of Antibody Evolution

- EvoluTion-aware AnTibody Language Model (EATLM)
  - Incorporate antibody evolution into pretraining

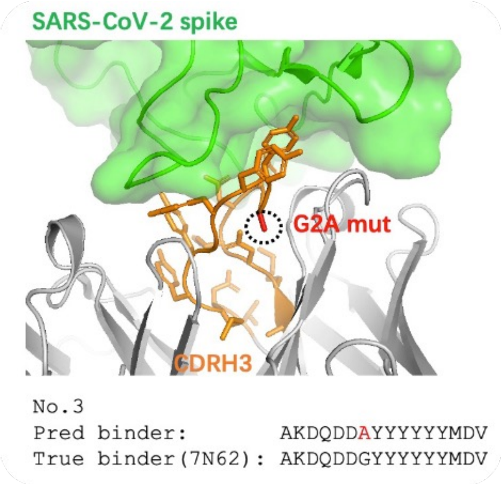


Accelerate Real-world Antibody Discovery

- Promising Antibody Binders for SARS-CoV-2



No	Predicted Binder	Existing Binder	Epitope	Identity
1	AREGIVGATTGFDY	AREGIVGATTGFDY	spike	1.000
2	ARDLGGYFDY	ARDLGGYFDY	RBD	1.000
3	AKDQDDAYYYYYYMDV	AKDQDDGYYYYYYMDV	NTD	0.938
4	ASYYYDSSGYHYGMDV	ASYYYDSSGYYYGMDV	RBD	0.938
5	ARRGLGLYYYGMDV	ARRGDGLYYYGMDV	S2	0.929
6	ARAFRGSYYYGMDV	ARATRGSYYYGMDV	S2	0.929
7	ARLSGSSWYFDY	ARLSGSSWDFDY	spike	0.917
8	ARLGSSSWYFDY	ARVGSSSWYFDY	spike	0.917
9	ARGWLRGYFDL	ARRGWLRGYFDL	RBD	0.909
10	ARDWGELYFDY	ARDWGEYYFDY	RBD	0.909
11	ARDLGGVFDY	ARDLGGYFDY	RBD	0.900



```
No.3
Pred binder:          AKDQDDAYYYYYYMDV
True binder(7N62):    AKDQDDGYYYYYYMDV
```

Take Away

- ❖ Present PPLMs struggle with antibody specificity tasks.
- ❖ By integrating the antibody evolution process, the pretraining can more accurately capture specificity.
- ❖ EATLM successfully identifies multiple promising SARS-Cov-2 binders.

Thanks for listening!

Code

Paper