ICML
International Conference
On Machine Learning

Reprogramming Pretrained Language Models for Antibody Sequence Infilling

Igor Melnyk[1], Vijil Chenthamarakshan[1], Pin-Yu Chen[1], Payel Das[1], Amit Dhurandhar[1], Inkit Padhi[1], Devleena Das[2]
[1]IBM Research, [2]Georgia Institute of Technology

IBM Research

Georgia Tech

Introduction

- Antibody Design
  - Essential in treatment of cancer, infectious and other diseases
  - Antigen high specificity results in less adverse effects during treatment
  - Complementarity Determining Region (CDR) crucial for antigen recognition and binding
- Challenges
  - Design and tailoring of CDR
  - Need for sequence and structural diversity in designed CDRs
  - Limited training data
- Our Work
  - ReprogBert for protein sequence infilling
  - Model Reprogramming English LLM for the task of CDR design
  - Diverse generated sequences while maintaining protein structural integrity
  - Efficient performance in data-scarce domains

ReprogBert

- Proposed System
  - Protein sequence infilling inspired by masked language modeling
  - Design CDR by infilling, guided by the rest of protein sequence
  - Model reprogramming repurposes English LLM to protein domain
  - Sequence-only method, protein structure information is not used
  - Based on `base-bert-uncased` from HuggingFace
- Model Reprogramming
  - Protein sequence (target domain), with $|V_t| = 30$ tokens

$$x_t = \langle a_1, a_2, \ldots, a_n \rangle$$

  - Language sequence (source domain), with $|V_s| = 30522$ tokens

$$y_s = \langle w_1, w_2, \ldots, w_n \rangle$$

  - Mappings: target to source and reverse

$$f_\theta : x_t \rightarrow x_s \qquad g_\gamma : y_s \rightarrow y_t$$

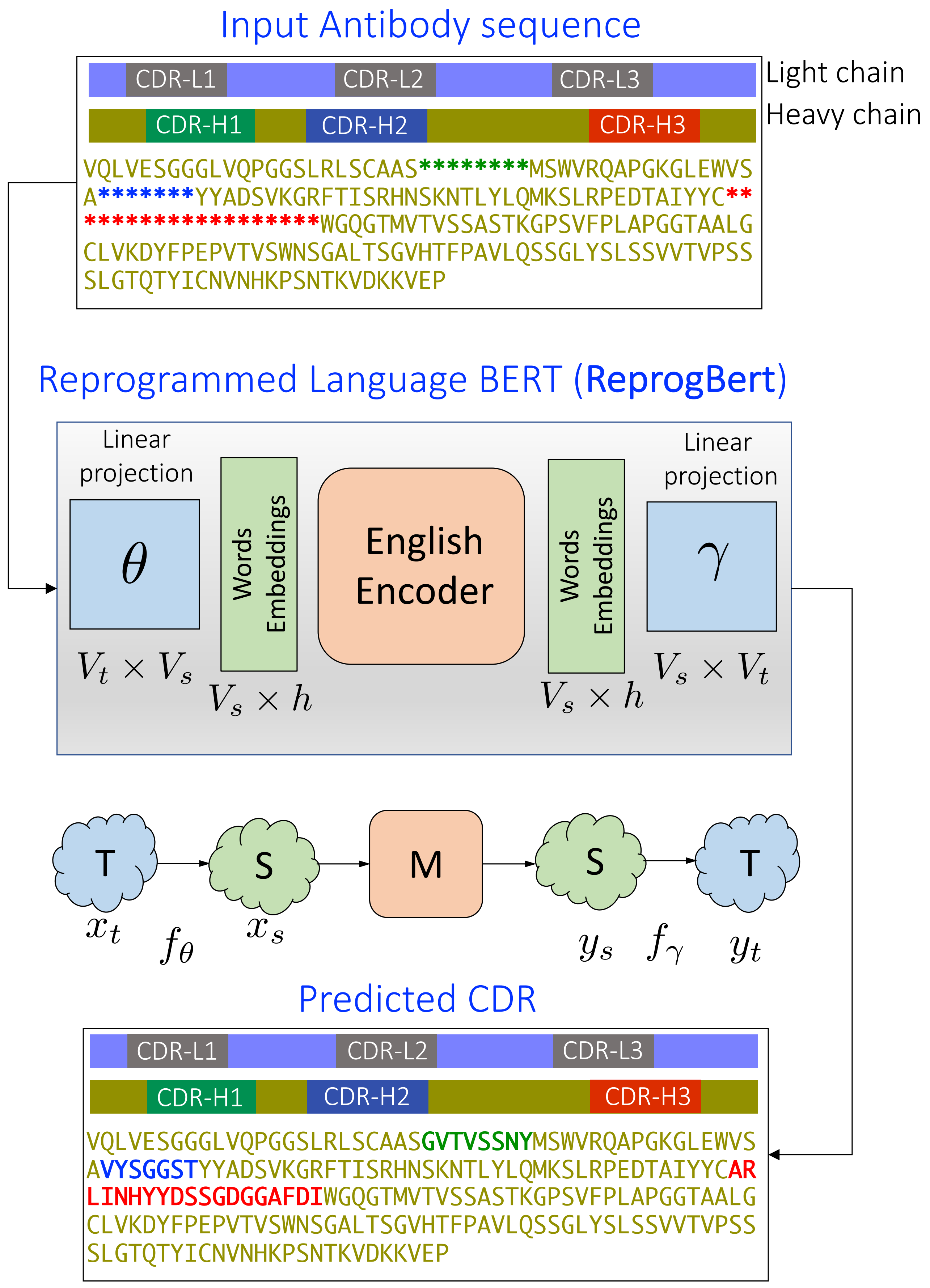
  - Constrain the maps to be linear

$$x_s = x_t\theta \qquad y_t = y_s\gamma \qquad \begin{array}{l}\theta \in \mathbb{R}^{|V_t|\times|V_s|}\\ \gamma \in \mathbb{R}^{|V_s|\times|V_t|}\end{array}$$

  - Example

$$x_t \in \mathbb{R}^{n\times|V_t|} \xrightarrow{\theta \in \mathbb{R}^{|V_t|\times|V_s|}} x_s \in \mathbb{R}^{n\times|V_s|} \xrightarrow{E \in \mathbb{R}^{|V_s|\times d}} x_s^E = x_sE$$

- Training: Only θ and γ are learned, all other model parameters fixed



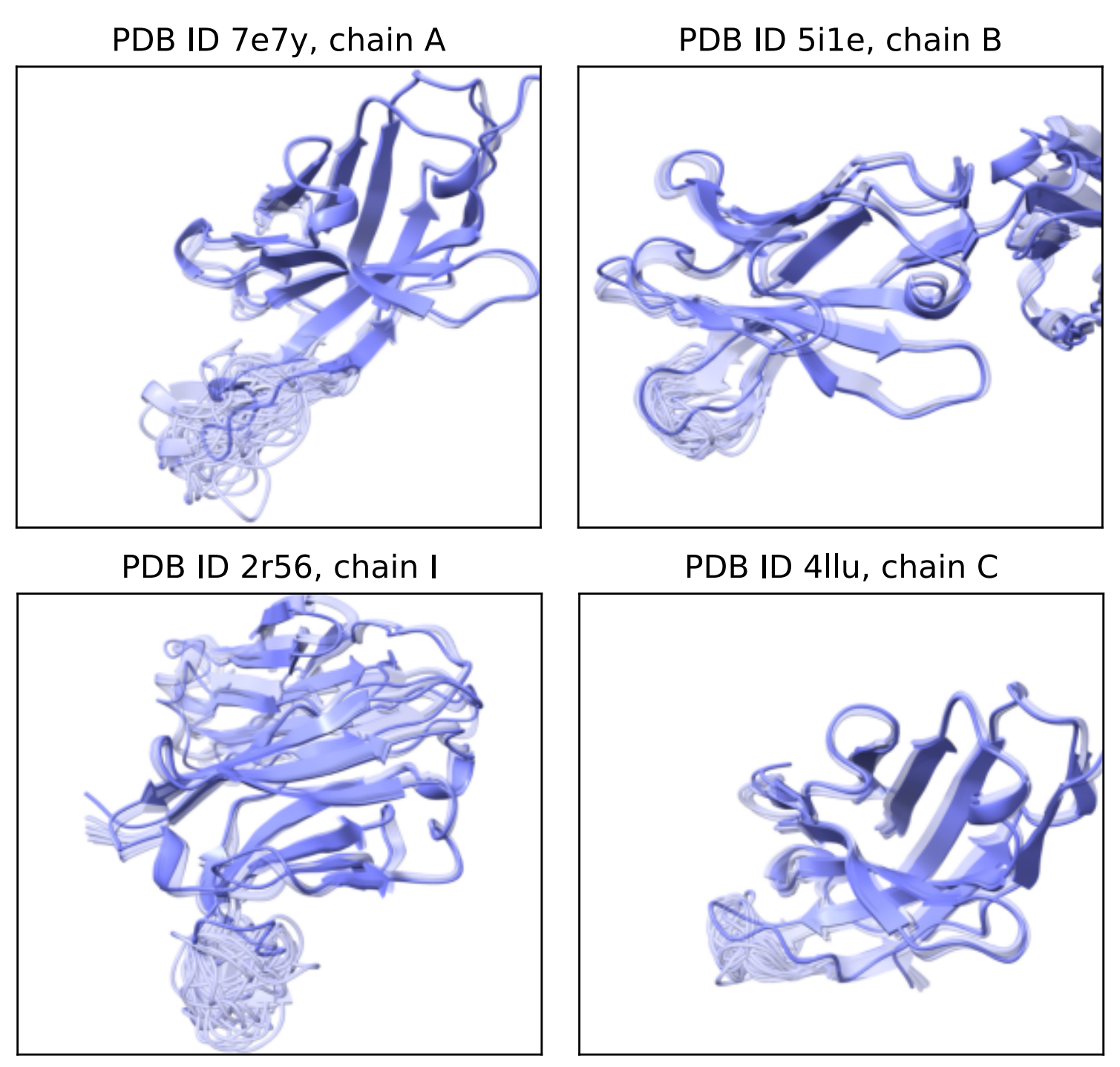
Input Antibody sequence

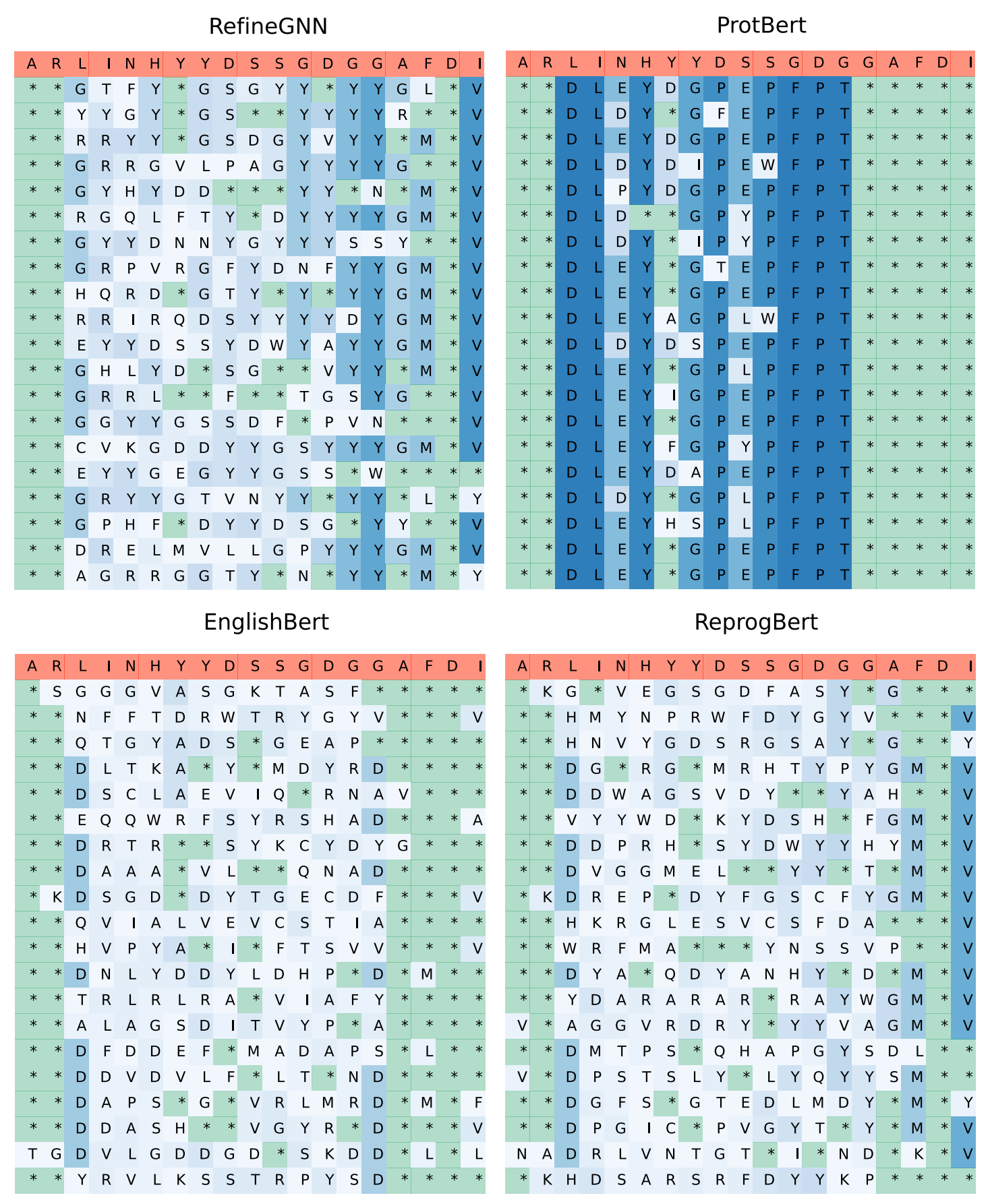
```
VQLVESGGGLVQPGGSLRLSCAAS********MSWVRQAPGKGLEWVS
A*******YYADSVKGRFTISRHNSKNTLYLQMKSLRPEDTAIYYC**
*****************WGQGTMVTVSSASTKGPSVFPLAPGGTAALG
CLVKDYFPEPVTVSWNSGALTSGVHTFPAVLQSSGLYSLSSVVTVPSS
SLGTQTYICNVNHKPSNTKVDKKVEP
```

Reprogrammed Language BERT (ReprogBert)

Predicted CDR

```
VQLVESGGGLVQPGGSLRLSCAASGVTVSSNYMSWVRQAPGKGLEWVS
AVYSGGSTYYADSVKGRFTISRHNSKNTLYLQMKSLRPEDTAIYYCAR
LINHYYDSSGDGGAFDIWGQGTMVTVSSASTKGPSVFPLAPGGTAALG
CLVKDYFPEPVTVSWNSGALTSGVHTFPAVLQSSGLYSLSSVVTVPSS
SLGTQTYICNVNHKPSNTKVDKKVEP
```





Code: github.com/IBM/ReprogBert

Paper: arxiv.org/abs//2210.07144

Experiments

- Baselines
  - LSTM *Saka et al., 2021 and Akbar et al., 2022*
    - sequence-only model, smaller capacity, single attention layer
  - AR-GNN *Jin et al., 2021*
    - autoregressive graph neural network, a sequence and structure-based model
  - RefineGNN *Jin et al., 2021*
    - designs protein sequence and 3D structure of CDR together as graphs
  - AbLang *Tobias H. Olsen & Deane, 2022*
    - LM trained on the antibody sequences to restore missing residues
  - Our proposed baselines:
    - ProtBert *Elnaggar et al., 2020*
      - specialized protein BERT model, pretrained on millions of protein sequences
    - EnglishBert
      - out-of-domain token embeddings replaced with in-domain AA embeddings
- Structural Antibody Database (SabDab)
  - Dataset statistics

CDR	Train	Validation	Test	Average CDR length	Average CDR diversity
CDR-H1	4050	359	326	8.1	60.8
CDR-H2	3876	483	376	7.9	68.2
CDR-H3	3896	403	437	14.5	76.9

  - Infilling results

	SabDab CDR-H3								
	PPL	PPL-ProGen	RMSD	RMSD-AF	RMSD-IF	TM-AF	TM-IF	AAR	DIV
LSTM	9.20	–	–	–	–	–	–	–	–
AR-GNN	9.44	–	3.63	–	–	–	–	–	–
Refine-GNN	8.38	7.2	2.50	5.62	3.43	85.0	94.0	28.2	25.7
AbLang	–	–	–	–	–	–	–	22.0	71.3
ProtBert	–	6.8	–	5.40	3.39	85.2	94.0	41.5	14.5
EnglishBert	–	5.9	–	5.53	3.26	84.9	94.0	35.6	59.8
ReprogBert	–	5.4	–	5.54	3.44	85.1	94.0	32.6	67.4

- Coronavirus Antibody Database (CoV-AbDab)
  - Dataset statistics

Dataset	CDR	Train	Validation	Test	Average CDR length
CoV-AbDab	CDR-H3	2282	291	291	15.7

  - SARS-CoV2 virus neutralization

	Neutralization Score	
Model	CoV-AbDab	CoV-AbDab + SabDab
Original	–	69.3
LSTM	–	72.0
AR-GNN	–	70.4
Refine-GNN	–	75.2
ProtBert	72.7	74.7
EnglishBert	70.5	71.0
ReprogBert	75.6	76.7