Evaluating the Biological Realism of IgLM-Generated Antibody Sequences

Natalia Rincon, Kevin Wiehe and Jeffrey J. Gray
Biomedical Engineering, Whiting School of Engineering, The Johns Hopkins University

Antibody sequences are highly diverse and current SOTA Ab language models do not fully capture known human biology.

Advances in language models have opened new avenues for the design of synthetic antibodies but it remains unknown whether generated sequences exhibit biological realism. We assess the sequence diversity and humanness of antibody sequences generated by the Immunoglobulin Language Model (IgLM), focusing on adherence to known immunological evolution mechanisms such as proper V(D)J recombination and somatic hypermutation. This study highlights current limitations and strengths of IgLM in antibody sequence generation.

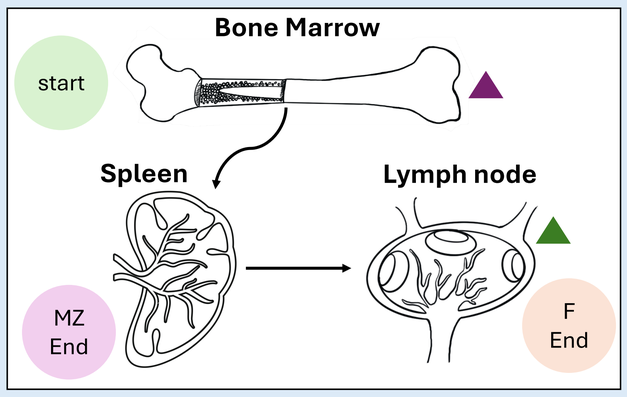
Current limitations include:

- Germline bias in generated sequences
- No disease specificity
- Insufficient metrics to evaluate the biological quality of generated sequences

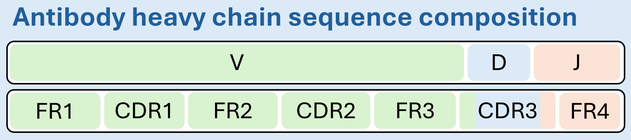
Four main biological processes for antibody diversity

1. **Combinatorial diversity of different V(D)J gene segments** ▲
2. Junctional diversity at the splice junctions (insertions/deletions)
3. Combinatorial diversity of different heavy and light chain pairs
4. **Somatic hypermutation (SHM)** ▲

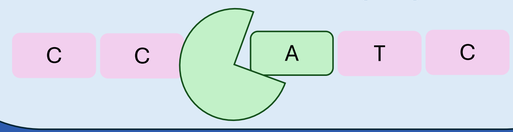
B cells undergo V(D)J recombination from germline genes to achieve diverse antibody sequences.



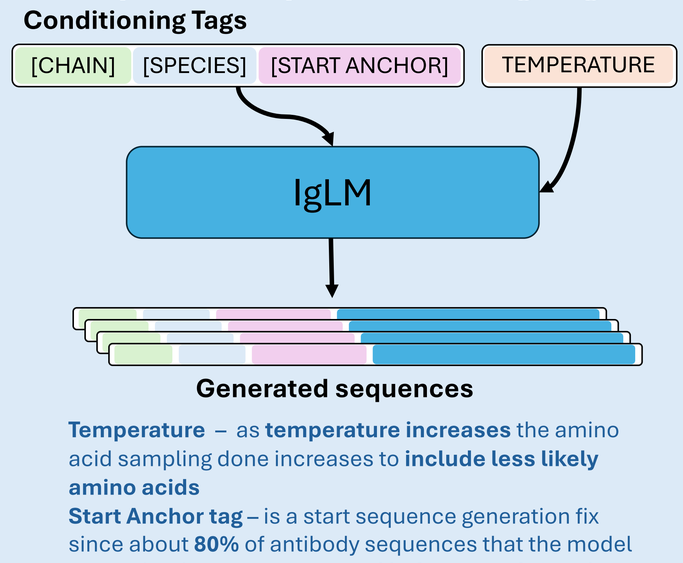
Antibody heavy chain sequence composition



Activation induced deaminase (AID) drives SHM



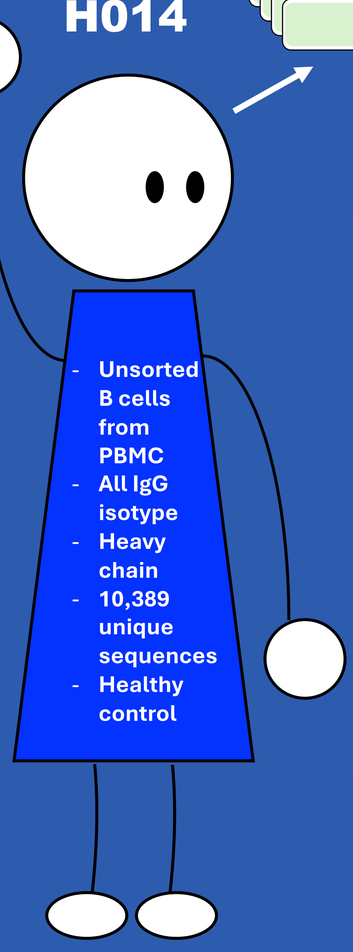
Immunoglobulin language model, IgLM, generates antibody sequences by conditioning tags.

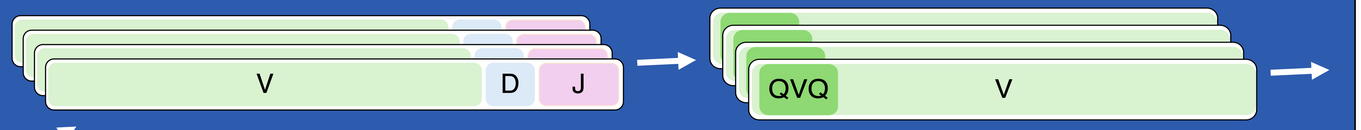


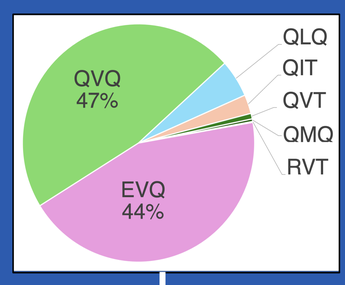
Temperature – as **temperature increases** the amino acid sampling done increases to **include less likely amino acids**
Start Anchor tag – is a start sequence generation fix since about **80%** of antibody sequences that the model was trained on were **truncated in framework 1 region**

IgLM generates sequences autoregressively and the start anchor tag defines which V gene will be used for the output antibody sequence.

EVQLVESGGGLVQPGGSLRLSCAASGFTFSSYAMHWVRQAPGKGLEYVSAISSNGGSTYYANSVKGRFTISRDNSKNTLYL
QMGSLRAEDMAVYYCAREVYSSGSWDYFDYWGQGTLVTVSS







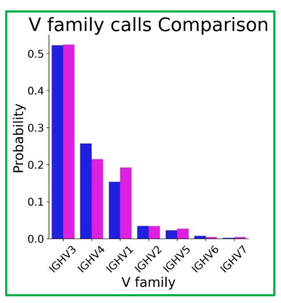
IgLM

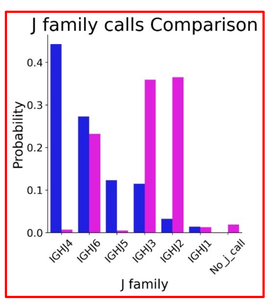
QVQ	EVQ	QLQ	QIT	QVT	QMQ	RVT
471	439	51	26	8	4	1

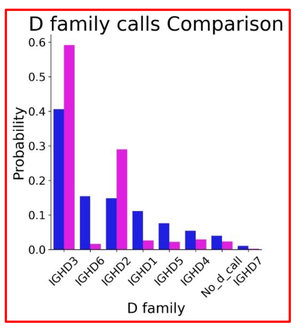
Synthetic H014

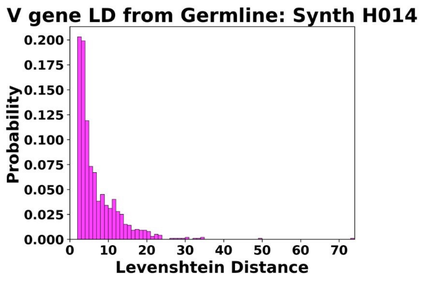
- IgLM temp. = 1
- Varying start anchor
- Heavy chain only
- 1,000 unique sequences

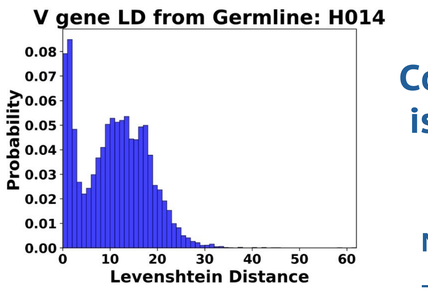
Synthetic repertoire H014 matches real repertoire H014 only in germline V gene family usage.









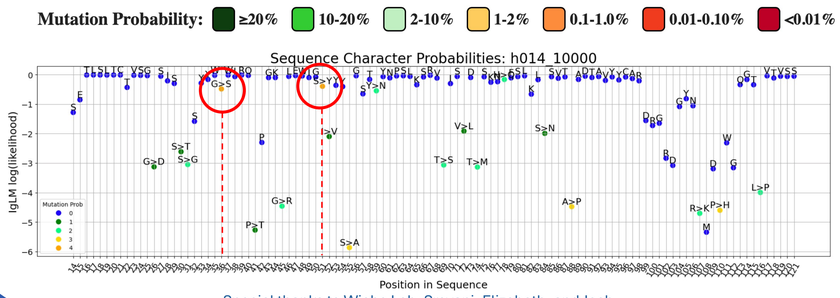


Comparison against mechanistic tools is necessary to explain discrepancies found when comparing sequence distributions.

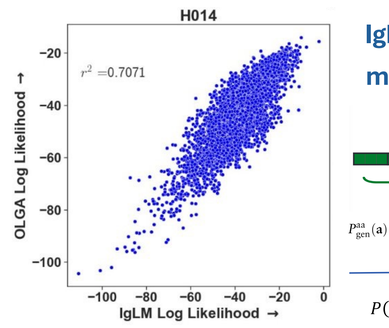
Mechanistic tools

- ARMADiLLO[5] is based on observed AID activity statistics
- OLGA[2] is based on observed V(D)J statistics

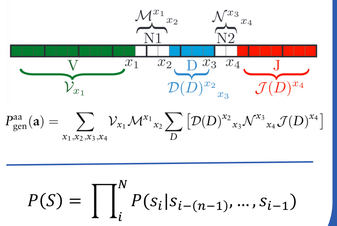
IgLM has high confidence in some mutations found to be improbable by ARMADiLLO.



$$\hat{P}(X_{i_{U\to Y}} | UCA,\ t) = \frac{1}{N}\sum_{j=0}^{N} 1(X_{ij} = Y)$$



IgLM log-likelihoods are correlated with mechanistic tool OLGA.

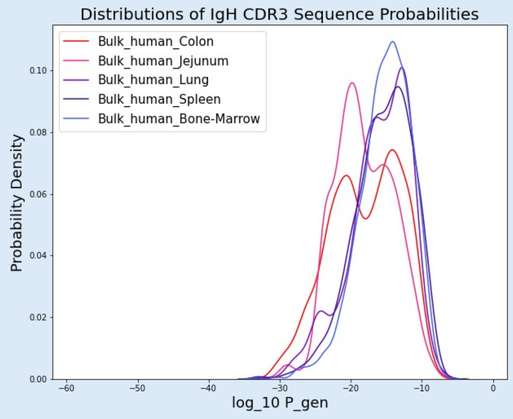


$$P_{\text{gen}}^{\text{aa}}(\mathbf{a}) = \sum_{x_1,x_2,x_3,x_4} \mathcal{V}_{x_1}\mathcal{M}^{x_1}{}_{x_2}\sum_{D}\left[\mathcal{D}(D)^{x_2}{}_{x_3}\mathcal{N}^{x_3}{}_{x_4}\mathcal{J}(D)^{x_4}\right]$$

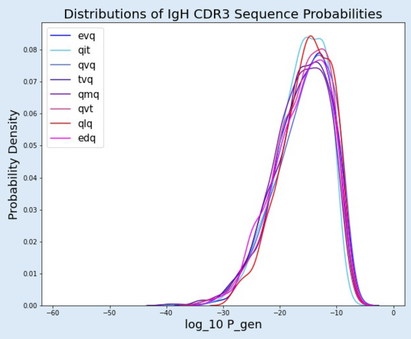
$$P(S) = \prod_{i}^{N} P(s_i | s_{i-(n-1)}, \ldots, s_{i-1})$$

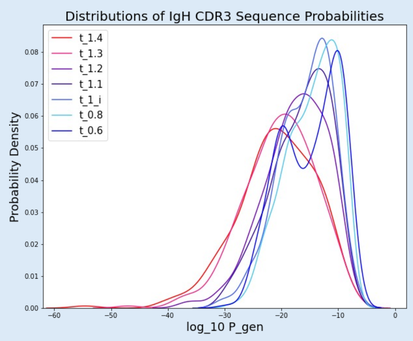
Antibody repertoires extracted from different tissues from a single individual have varying probability distributions for CDR3 sequence probability.

- Data is unsorted B cells extracted from **Subject 149 in Meng 2017 study**
- Probability of generating a CDR3 sequence computed with **OLGA[2]**
- CDR3 extracted from each sequence using **AbNumber[4]**



Varying IgLM generation temperature is one way to imitate tissue-specific CDR3 distribution changes seen in real data.





Default IgLM temperature = 1
Varying start anchor prompt

Recommended start anchor = EVQ
Varying temperature

Bringing domain expertise in protein language models to improve sequence generation

We can design better tools for antibody design by quantifying the quality of synthetic antibody sequences generated by SOTA generative language models. In this study, I highlighted limitations in IgLM's gene usage, bias towards germline, lack of diversity and limitations of its confidence metric. I also show how we can imitate real repertoires when using more advanced and fine-tuned parameters.

Future directions

- Test more known human biology such as junctional diversity and combinatorial diversity
- Test other DL-based antibody generation tools such as AbLang, AntiBERTy and IgT5
- Develop a more tunable model – able to specify sequence generation for a given B cell type, tissue type, and/or disease

Special thanks to Wiehe Lab: Sravani, Elizabeth, and Josh

References:
1. Shuai et al. IgLM: Infilling language modeling for antibody sequence design, 2023
2. Sethna et al. OLGA: fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs, 2019
3. Meng et al. An atlas of B-cell clonal distribution in the human body, 2017
4. Dunbar et al. ANARCI: antigen receptor numbering and receptor classification, 2015
5. Wiehe et al. Functional Relevance of Improbable mutations for HIV Broadly Neutralizing Antibody Development, 2018
6. Bernades et al. Longitudinal Multi-omics Analyses Identify Responses of Megakaryocytes, Erythriod Cells and Plasmablasts as Hallmarks of Severe COVID-19, 2020