Benchmarking deep generative models for diverse antibody sequence design

Igor Melnyk, Payel Das,
Vijil Chenthamarakshan, Aurelie Lozano
IBM Research AI

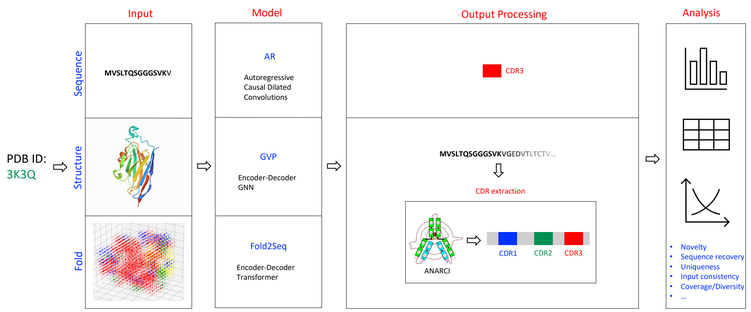
Introduction

- Antibody design plays a key role in research, diagnostics and therapeutics
- Designing functional sequences typically has combinatorial complexity
- Need to impose sequence and structural constraints

- Antigen binding specificity largely determined by CDR
- Among CDRs, CDR3 contributes most sequence and length diversity
- Sampling diverse CDR3s is the main focus of many antibody design methods

- We benchmark three recent deep generative models:
- **AR** – autoregressive approach uses causal dilated convolutions for input prefix sequence to generate CDR3 subsequence
- **GVP** – encoder-decoder GNN that represents input structure information that is autoregressively decoded into protein sequence
- **Fold2Seq** – encoder-decoder Transformer that embeds fuzzy input fold information in joint sequence-fold space which is then decoded into protein sequence

System Overview



- We used Chain A PDB ID 3k3Q (llama nanobody) as input to all three methods (AR, GVP, Fold2Seq)
- For GVP and Fold2Seq, the generated sequence is analyzed by ANARCI to extract CDR3
- For AR, we considered extra filter to exclude sequences not ending with beta-strand of nanobody template
- Generated CDR3s are then analyzed for different properties

Results

Sequence Recovery and NLL of generated CDRs

Model	Seq Recovery Rate (%)	NLL
Fold2Seq	30.711	2.572
GVP	40.131	2.987
AR	48.865	0.375
Natural	–	0.371
Synthetic	–	4.912
NGS	–	5.102

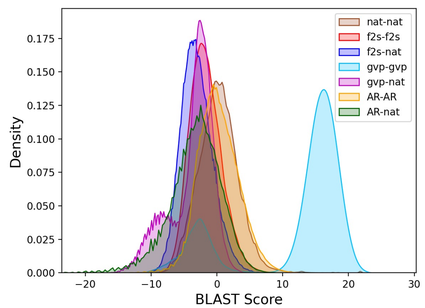
- Natural – natural llama library
- Synthetic – synthetic library
- NGS – next-generation sequencing library
- All methods have SRR > 30%, implying fold consistency
- GVP is more accurate than Fold2Seq at recovery, while Fold2Seq has lower NLL, indicative of functional fitness
- AR has highest recovery rate

Uniqueness and novelty of CDRs

		Fold2Seq	GVP	AR unfiltered	AR filtered	Natural Llama
CDR3	Uniqueness	100	88.33	87.57	13.85	100
	Novelty	43.36	52.71	11.92	8.97	52.64
CDR2	Uniqueness	100	9.15	–	–	100
	Novelty	58.70	9.15	–	–	83.83
CDR1	Uniqueness	92.49	56.20	–	–	100
	Novelty	60.75	51.99	–	–	83.37

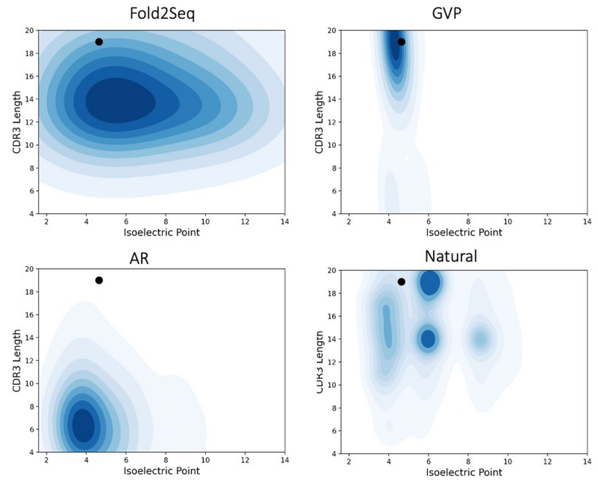
- AR filtered – filtering based on final beta-strand
- AR unfiltered – no filtering applied
- Fold2Seq outperforms AR and GVP in terms of uniqueness
- GVP generates more novel CDR3s, while Fold2Seq is better at CDR1&2

Kernel Density Estimate for pairwise similarity



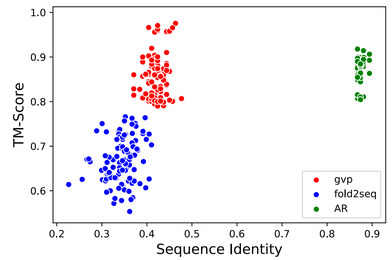
- E.g., f2s-f2s – self-similarity, f2s-nat – similarity to natural sequences
- Fold2Seq sequences are more diverse
- GVP generates sequences which are similar to each other

Density plot for isoelectric point and CDR3 length



- Black dot – ground truth CDR3
- Fold2Seq produces significant coverage of the natural sequence
- GVP generates sequences close to the input PDB ID (limited diversity)
- AR tends to generate short sequences

Sequence identity vs structure similarity



- GVP sequences exhibit higher TM-score than Fold2Seq
- Fold2Seq shows greater sequence diversity with structural consistency
- AR shows high sequences identity and TM score since only small CDR part is generated, the rest is copied

References

[AR] Jung-Eun Shin et al (2021) Protein design and variant prediction using autoregressive generative models. Nature Communications

[GVP] Bowen Jing at al (2021) Learning from protein structure with geometric vector perceptrons. ICLR

[Fold2Seq] Yue Cao et al (2021) Fold2seq: A joint sequence (1d)-fold (3d) embedding-based generative model for protein design. ICML