Supplementary Information: Addressing the antibody germline bias and its effect on language models for improved antibody design

Tobias H. Olsen[1,2], **Iain H. Moal**[1] **and Charlotte M. Deane**[2]

[1] GSK Medicines Research Centre, GSK, Stevenage SG1 2NY, United Kingdom
[2] Department of Statistics, University of Oxford, Oxford OX1 3LB, United Kingdom

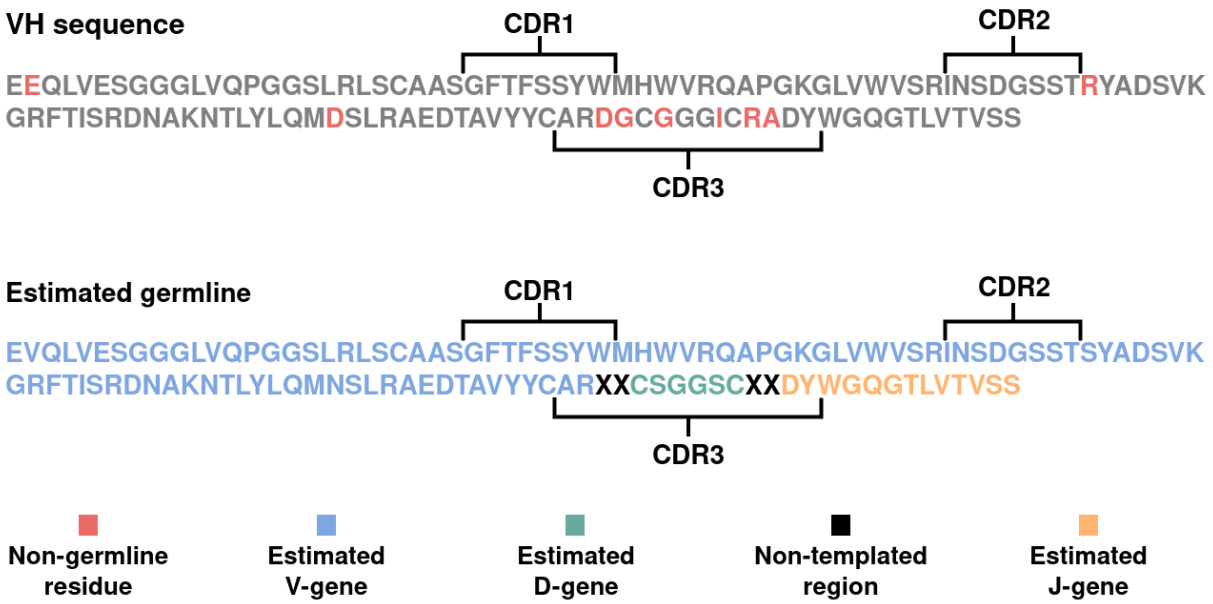

Figure S1: Example of a VH sequence compared to its germline sequence estimated by IgBLAST [39]. Residues are labelled as non-germline (NGL) residues (red) when different from the estimated germline or part of a non-templated region.

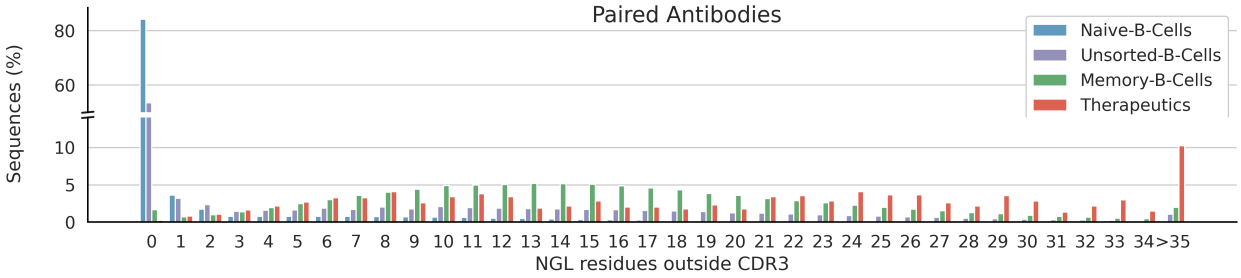

Figure S2: Distribution of NGL residues per VH-VL domain by source. Naive B-cell derived antibodies predominantly lack NGL residues, while memory B-cell derived antibodies display an average of ∼15.3. Therapeutic antibodies exhibit an average of ∼20.3 NGL residues.

	Non-germline residues			
	Heavy		Light	
	FWR	CDR1/2	FWR	CDR1/2
ESM-2	241.7	155.3	158.3	116.8
AntiBERTy	595.2	360.4	848.1	398.5
AbLang-1	31.8	19.8	55.1	26.4
Ab-Unpaired	361.8	245.4	229.4	113.4
Ab-Paired	197.4	110.6	293.5	117.0
Ab-FL	22.7	23.7	24.5	22.7
Ab-ModMask	27.0	28.2	30.0	26.1
Ab-FT	26.9	28.6	29.0	27.3
AbLang-2	21.5	21.4	23.7	20.6

Table S1: Comparison of perplexity computed when predicting NGL residues which has been reverted to the germline in the input. Comparison is between the general protein language model (LM) ESM-2 [18], the antibody-specific LMs AntiBERTy [22] and AbLang-1 [24], and our new selection of antibody-specific LMs (see Methods 2.4). Although AbLang-2 shows the best performance, its perplexity is only slightly better than random, highlighting the need for further work to enable LMs to suggest mutations away from a know germline.

	Non-germline residues			
	Heavy		Light	
	FWR	CDR1/2	FWR	CDR1/2
ESM-2	3.0	3.3	2.5	2.0
AntiBERTy	2.8	2.2	3.3	2.2
AbLang-1	25.5	19.2	39.7	18.1
Ab-Unpaired	2.9	2.2	3.1	2.3
Ab-Paired	4.7	3.5	6.5	3.7
Ab-FL	2.8	2.8	3.1	2.7
Ab-ModMask	2.5	2.3	2.7	2.3
Ab-FT	2.6	2.5	2.8	2.4
AbLang-2	2.9	2.9	2.9	2.6

Table S2: Comparison of perplexity computed when predicting unmasked NGL residues. Comparison is between the general protein language model (LM) ESM-2 [18], the antibody-specific LMs AntiBERTy [22] and AbLang-1 [24], and our new selection of antibody-specific LMs (see Methods 2.4). With the exception of AbLang-1, all models predict unmasked NGL residues with similar low perplexity. This is likely because AbLang-1 was trained on a significantly reduced set of antibody sequences, having therefore seen fewer NGL residues during training.

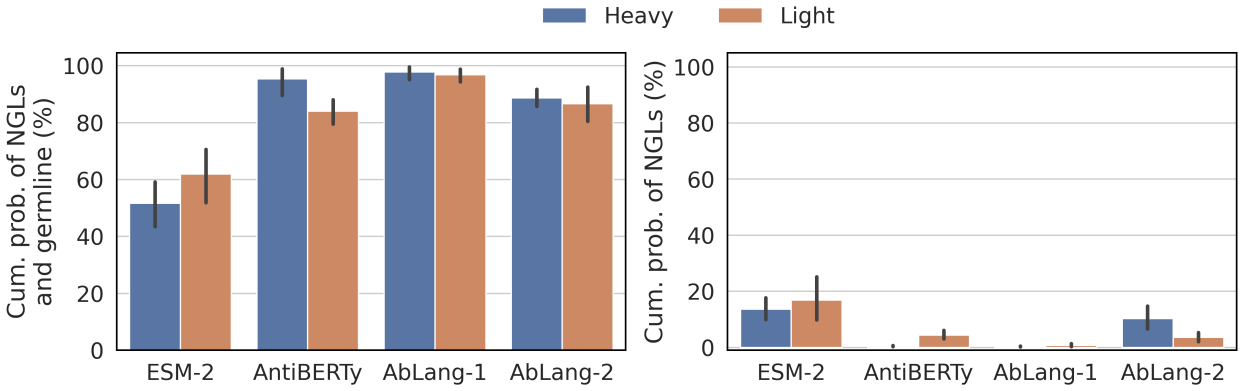

Figure S3: Comparison of cumulative probabilities of valid residues for the general protein language model (LM) ESM-2 [18] and the antibody-specific LMs AntiBERTy [22], AbLang-1 [24] and AbLang-2. Clonotypes were formed by grouping antibodies by source, V/J genes, and identical CDR3s. The strict clonotyping yielded 39 and 13 sites with two known NGL residues outside of the CDR3 in VHs and VLs, respectively, and a single site with three in a VH. The cumulative probabilities for known NGL residues and the germline for AntiBERTy, AbLang-1, and AbLang-2 are >80%. ESM-2 presents 52% and 62% for the VHs and VLs. The cumulative probabilities for known NGL residues for AntiBERTy and AbLang-1 show <1% for the VH, while ESM-2 and AbLang-2 display 13.6% and 10.3%. For the VL, values are 4.4% and 0.7% for AntiBERTy and AbLang-1, and 16.9% and 3.6% for ESM-2 and AbLang-2.