Cell Systems

Transparent Peer Review Record

IgLM: infilling language modeling for antibody sequence design

Richard W. Shuai, Jeffrey A. Ruffolo, Jeffrey J. Gray

Summary

Initial Submission: Received January 09, 2023
Preprint: https://doi.org/10.1101/2021.12.13.472419

Scientific editor: Ernesto Andrianantoandro, Ph.D.

First round of review: Number of reviewers: Three
Three confidential, Zero signed
Revision invited March 12, 2023
Major changes anticipated
Revision received June 14, 2023

Second round of review: Number of reviewers: Three
Three original, Zero new
Three confidential, Zero signed
Accepted October 02, 2023

Editorial decision letter with reviewers' comments, first round of review

Dear Dr. Gray,

I'm enclosing the comments that reviewers made on your paper, which I hope you will find useful and constructive. As you'll see, they express interest in the study, but they also have a number of criticisms and suggestions. Based on these comments, it seems premature to proceed with the paper in its current form; however, if it's possible to address the concerns raised with additional experiments and/or analysis, we'd be interested in considering a revised version of the manuscript.

As a matter of principle, I usually only invite a revision when I'm reasonably certain that the authors' work will align with the reviewers' concerns and produce a publishable manuscript. In the case of this manuscript, the reviewers and I have make-or-break concerns that can be addressed by:

1. Clarifying advance over previous antibody-specific language models, supported by quantitative comparison and additional analysis.
2. Providing a stronger argument for utility of the infilling task, substantiated with quantitative analysis and/or additional data.
3. Providing more methodological detail

The reviewers are generally supportive of the paper and their comments are intended to flesh out the study, both in terms of improving the framing of the study and experiments/analyses to substantiate the main claims. To help guide revision, I've highlighted portions of the reviews that strike me as particularly crucial.

As you address these concerns, it's important that you and I stay on the same page. I'm always happy to talk, either over email or by Zoom, if you'd like feedback about whether your efforts are moving the manuscript in a productive direction. Do note that we generally consider papers through only one major round of revision, so the revised manuscript would be either accepted or rejected based on the next round of comments we receive from the reviewers. If you have any questions or concerns, please let me know. More technical information and advice about resubmission can be found below my signature. Please read it carefully, as it can save substantial time and effort later.

I look forward to seeing your revised manuscript.

All the best,

Ernesto Andrianantoandro, Ph.D.
Scientific Editor, Cell Systems

Reviewers' comments:

Reviewer #1: Shuai et al. present IgLM, an autoregressive transformer model based on the GPT-2 architecture. The preprint has been out for a while- it arguably needs more thinking on novelty and what it offers, especially with models like ProGen-2 (coincidentally authored by one of the authors in this manuscript)!

There are some major issues and minor issues worth addressing. In particular I have some reservations on how the model's generative prowesses are evaluated, and some minor concerns worth mentioning in terms of data prep (which I think the authors may already know about and could just be worth a clarification).

MAJOR
- The section on prompting needs further discussion. Why is it that prompting all of a sudden removes truncated sequences? Why do the authors only use one set of initial residues? e.g. DIQ only for human kappa and not EIV, for example, too.
- The infilling therapeutic antibody CDRs is interesting but there's no mention on whether any of these infilled sequences would retain binding of the antigen. While doing an experiment is ideal, even some docking results to confirm this would be nice (SnugDock, or AF2 multimer, etc)
- The authors should consider benchmarking against something like a T5 for the therapeutic antibody infilling task where spans are predicted in-place, or even using an MLM to re-fill using bidirectional context (e.g. using AbLang)
- The selection of 49 antibodies seems unusual, some of these antibodies (e.g. muromonab) is chimeric/mouse so I would imagine it's trivial to get higher OASis/humanness scores

MINOR
- Authors don't describe use of humanized mice in antibody selection, which has been revolutionary for how we discover antibodies
- Intro mentions diversity of antibodies but does the justification for 10^13 make sense? CDRs inherently have a lot of bias - CDR3 usually ends with ARD and end FDY, so that already just means (for a typical CDR3 of length 10) that there are 20^4 variants - and that's assuming amino acids are uniformly distributed
- The intro could do with citing more recent examples of representation learning models from AbSci (Bachas et al. 2022), sequence generation models such as ProGen2 (Nijkamp et al 2022). Other papers that should be discussed are cases where language models are used to mutate and design antibodies (Hie et al. 2022). AntiBERTa paper seems to not be properly cited, it's cited with the AbLang paper from Olsen et al. Prihoda paper is now out in mAbs which should be updated.
- How are the validation sets created? It could be worth pruning to make sure no study is shared between training and validation sets when pre-training IgLM to avoid perplexities looking lower
- Why are structures predicted with AF2 multimer in the intro / Fig 1 but then IgFold later?
- Why is it that some designs only have CDR3 of length 1? And for these cases, what is the likelihood of

the sequence?
- Authors don't describe how IgLM generates paired sequences - dataset seems to be unpaired and even input seems unpaired?
- Fig 3B it might be good to indicate the length of the native H3
- Fig 4E might be better as a heatmap or some kind of 3D scatterplot where we correlate T and P and have the median change in OASis identity
- Fig 4F the classification of human sequences another good baseline could be using the ANARCI HMM cutoffs? Especially as the authors point out ANARCI's annotation of species isn't perfect

Reviewer #2: Shuai and colleagues here describe a language model for antibody design. They claim that their model can generate full length sequences for various species.

While the work performed is extensive, after reading it, it remains a bit opaque to the reader what the paper has in fact demonstrated. The advance of previous research needs to be much more clear. Also, we would like to see more surprising things than full sequence reconstruction of different animals.

Please see major and minor comments below.

Major comments
The novelty this paper provides is a method that enables generative infilling of antibody sequences of various sizes while conditioning the infilled regions on species type and chain type. I imagine this could enable the generation of some interesting antibody sequences, e.g. a sequence that is part human/part mice/part something else. However, It is not obvious to me what the practical utility of this method is supposed to be. Maybe the authors should include an example where the partial modification of an available antibody sequence would be useful over simply generating sequences completely from scratch, especially if it is not possible to condition the sequences for desirable properties (binding affinity, developability parameters). Even the humanization of a murine antibody is not an option since there is no guarantee that the model will maintain the same binding affinities.

The first half of the results is all about how well the model generates antibody sequences conditioned on species and chain types. But none of this is very surprising, since they were trained with data labelled with that information already (which also makes it not completely self-supervised) and also that information is encoded in non-variable regions, which is not what the infilling task targets anyways.

pg. 3, lines 83-104. It is unclear what the infilling perplexity tells us about the language model's effectiveness. All the reported results are expected (IgLM achieves lower perplexity than IgLM-S, longer sequences have higher perplexity, etc.), and there is no comparison to any other language model to give a reader a good sense on whether IgLM's perplexity scores are particularly good, average, or bad.

pg. 7, lines 228-237. Since IgLM is tasked with only predicting CDR H3 loops while the rest of the antibody is given, is it surprising that it generates more human-like antibodies? Also, what does it mean to

have even more human-like than the parent antibody (line 231), when the parent antibodies have been already optimized to be so?

pg. 7, lines 209-227. All developability measures are done in silico with computational tools. How reliable is it to conclude that IgLM generated sequences with increased developability?
For all sequences generated, it's not entirely completely clear to us how the authors determine whether a given sequence is actually an antibody sequence that belong to a certain species. Can this authors clarify this? How do they quantitatively test for this?

"For mouse 137 and rhesus light chains, IgLM generates the correct species in 34.89% and 88.14% of cases, respectively" → why so low for mouse?

"As temperature increases, sequences for every species begin to diverge from germline, effectively acquiring mutations." Are these mutations nature-like? Meaning, are they inserted according to mutational hotspot motifs?

For the sequences generated, does their V gene distribution follow that of the training data?

Pg. 5, lines 171ff: The therapeutic antibody diversification part is not clear to me: how can the authors be certain that the generated antibodies still bind the same target?

At what temperature threshold do antibodies turn from "being diversified" to "unnatural/random sequences"

Have you tried to classify generated from OAS sequences to measure whether the LM introduced obvious biases?

Minor comments

Page 5, line 57: the (GPT2) addition is not true since GPT2 refers to a transformer pre-trained on text data.

pg. 3, line 73. Can the authors provide a definition of sampling temperature? While they explain what the numbers indicate, it would be good for the reader to have a more precise idea about this measurement.

The authors may want to add this paper to the ref list https://www.nature.com/articles/s41598-021-85274-7

N-Terminal truncations in OAS: why didn't the authors just remove such sequences from the dataset? What is their frequency in the OAS?

There are several typos throughout the text.

Reviewer #3: OVERVIEW
Shuai et al. describe an antibody specific language model, IgLM. The model is trained autoregressively and allows a user to condition on relevant properties and the right-sequence context during autoregressive sequence generation, which would potentially enable improved and more controllable sequence sampling and design. Overall, the I think the work is interesting, but also preliminary. More specifically, I would very much appreciate major revisions in the form of comparisons with previous antibody-specific language models, as well as general protein language models trained on generic protein sequence corpora, before I would be able to make a recommendation for publication. This is important for assessing where and when IgLM provides an advantage over existing tools. Additional details are provided below.

DATA & CODE AVAILABILTY
The code availability and usability are very good. The GitHub describes a very easy-to-use command line implementation. The model availability is also very good, and is easily downloadable/installable. Optionally, it would be nice to also make the processed training data available, or if the dataset size is too large to fit in common repositories like Zenodo or figshare, making the dataset processing scripts available would be nice to have.

REQUIRED MAJOR REVISIONS

1. A major question I had while reading the paper, and which I would like to see addressed, is how well this language model compares to previous antibody language models (including AntiBERTy/a, AbLang, Sapiens) as well as to general protein language models (ESM, ProGen with and without antibody-specific conditioning or training). Foremost, the authors should compare their perplexity results to these language models, and I would be very curious as to how this model compares.

2. Closely related to the above point is that the authors should compare across different language models according to the various downstream metrics described in the paper, such as the coherence of downstream structure prediction, developability, and humanness. There is some comparison to this effect (e.g., involving ProGen2 trained on OAS), but the comparison should be much more exhaustive in order to communicate where this model could or could not be useful.

3. The Methods section is extremely sparse and needs to be improved. The authors need to provide enough experimental detail in the Methods to reproduce all analyses described in the paper. This not only applies to the language model description and training details as currently provided (which are still limited) but also to all of the downstream tasks and additional benchmarking experiments.

4. As an ablation, what happens if the authors sample, e.g., the CDR3 without including the right context in their infilling task? Given that a central claimed improvement is the ability to condition on both the left- and right-sequence context, the actual effect of this would be interesting to see and I was not able to find

an analysis like this in the manuscript.

MINOR REVISIONS, SUGGESTIONS & COMMENTS

1. One thing that is often seen when training on OAS, followed by sample sequences based on high/maximum likelihood, is that the generated sequences reproduce the sequence distribution of the antibody germline. This would potentially make such models useful for germline reversion (though maybe less useful for tasks like affinity maturation). I was wondering if the authors could comment on this in their Discussion. An optional revision is to evaluate the model's ability to identify the germline sequence or to see if there's concordance between maximum likelihood mutations and those recommended by common germline reversion tools.

Authors' response to the reviewers' first round comments

Attached.

Editorial decision letter with reviewers' comments, second round of review

Dear Dr. Gray,

I'm very pleased to let you know that the reviews of your revised manuscript are back, the peer-review process is complete, and only a few minor, editorially-guided changes are needed to move forward towards publication. Pleas note that Reviewer #2 supports publication but did not leave any comments for authors.

In addition to the final comments from the reviewers, I've made some suggestions about your manuscript within the "Editorial Notes" section, below. Please consider my editorial suggestions carefully, ask any questions of me that you need, make all warranted changes, and then upload your final files into Editorial Manager.

I'm looking forward to going through these last steps with you. Although we ask that our editorially-guided changes be your primary focus for the moment, you may wish to consult our FAQ (final formatting checks tab) to make the final steps to publication go more smoothly. More technical information can be found below my signature, and please let me know if you have any questions.

All the best,

Cell Systems

Transparent Peer Review Record

Ernesto Andrianantoandro, Ph.D.
Scientific Editor, Cell Systems

Editorial Notes

Transparent Peer Review: Thank you for electing to make your manuscript's peer review process transparent. As part of our approach to Transparent Peer Review, we ask that you add the following sentence to the end of your abstract: "A record of this paper's Transparent Peer Review process is included in the Supplemental Information." Note that this ***doesn't*** count towards your 150 word total!

Also, if you've deposited your work on a preprint server, that's great! Please drop me a quick email with your preprint's DOI and I'll make sure it's properly credited within your Transparent Peer Review record.

Title:

The title is too generic and gives the impression that generative models have not been used for antibody design before. I suspect it could be more effective. Please include something about how this is different than other generative model based approaches, e.g. the bidirectional context and infilling capacity, since this is the main advance. As you re-consider your title, note that an effective title is easily found on Pubmed and Google. A trick for thinking about titles is this: ask yourself, "How would I structure a Pubmed search to find this paper?" Put that search together and see whether it comes up is good "sister literature" for this work. If it does, feature the search terms in your title. You also may wish to consider that PubMed is sensitive to small differences in search terms. For example, "NF-kappaB" returned ~84k hits as of March, 2018, whereas "NFkappaB" only returned ~8200. Please ensure that your title contains the most effective version of the search terms you feature.

Also, please consider including the name of the model (IgLM) in your title.

Abstract:

Please revise the Abstract to better represent the findings in the manuscript. As with the title, please make sure to be clear about the main advance of the paper and be cognizant of the current literature context. The infilling capability and comparisons with alternative language models need to be better

represented in the Abstract. When revising, please also clarify that the improvements in developability were tested in silico. Please ensure the revised Abstract is 150 words or less.

Manuscript Text:

Please make sure to be transparent about how your data support your claims and be clear about the limits of claims, particularly with regard to "developability" and characteristics that would be most convincingly tested experimentally (folding, binding, biophysical characteristics, etc.). This can be addressed by text changes to be more transparent in the headings, figure legends, abstract, and the main text itself.

For example:

"Infilled loops display improved developability"

To

"Infilled loops display improved developability in silico"

"IgLM generates foldable antibody sequences"

To:

"IgLM generates foldable antibody sequences in silico"

"…that display favorable biophysical properties…"

To

"…that display favorable predicted biophysical properties…"

Or

"…that display favorable biophysical properties in silico…"

Also:

- House style disallows editorializing within the text (e.g. strikingly, surprisingly, importantly, etc.), especially the Results section. These terms are a distraction and they aren't needed—your excellent observations are certainly impactful enough to stand on their own. Please remove these words and others like them. “Notably” is suitably neutral to use once or twice if absolutely necessary.
- Please only use the word "significantly" in the statistical sense.

Figures and Legends:

Please look over your figures keeping the following in mind:

- When color scales are used, please define them, noting units or indicating "arbitrary units," and specify whether the scale is linear or log.
- Please ensure that every time you have used a graph, you have defined "n's" specifically and listed statistical tests within your figure legend.

STAR Methods:

Please convert your Methods section to our STAR Methods format - consult the STAR Methods guidelines for detailed instructions.

Thank you!

Reviewer comments:

Reviewer #1: Thanks for addressing the comments. There's considerable work which has helped to improve the manuscript - I think there are still some bits that could be worth exploration but this will just be going in circles! I think the biggest surprise is that the ProGen2 model performs relatively poorly, despite the added complexity of the model.

Reviewer #3: I would like to thank the authors for their thorough responses to my and the other reviewers'

comments. I am happy to recommend acceptance of the manuscript.

Response to reviewers

We thank the reviewers for their comments and critiques, which have undoubtedly resulted in a stronger manuscript. Below we detail the changes made in response to the reviewer's comments. For convenience, the original comments are included, while our responses (indented) are below. Changes to the manuscript are indicated by red text.

Reviewer 1

Shuai et al. present IgLM, an autoregressive transformer model based on the GPT-2 architecture. The preprint has been out for a while- it arguably needs more thinking on novelty and what it offers, especially with models like ProGen-2 (coincidentally authored by one of the authors in this manuscript)!

There are some major issues and minor issues worth addressing. In particular I have some reservations on how the model's generative prowesses are evaluated, and some minor concerns worth mentioning in terms of data prep (which I think the authors may already know about and could just be worth a clarification).

Major Comments:

The section on prompting needs further discussion. Why is it that prompting all of a sudden removes truncated sequences? Why do the authors only use one set of initial residues? e.g. DIQ only for human kappa and not EIV, for example, too.

> Much of the immune repertoire data prioritizes sequencing of the C-terminal ends. However, because much of the antibody sequence is conserved at the N-terminus, we expected that the model could be encouraged to start at the first residue by providing a short largely-conserved motif, and then generate freely. We have expanded the text describing the justification for using residue prompts, as well as the process for identifying appropriate initial residues in the following text:
>
> However, we observed that the sequences frequently featured N-terminal truncations. These truncations are frequently observed in the OAS database used for training, with over 40% of sequences missing the first fifteen or more residues. For heavy chains, these N-terminal deletions appeared as a left-shoulder in the sequence length distribution (Figure 2B, left) with lengths ranging from 100 to 110 residues. For light chains, we observed a population of truncated chains with lengths between 98 and 102 residues (Figure 2B, right). To address truncation in generated sequences, we utilized a

prompting strategy, wherein we initialize each sequence with a three-residue motif corresponding to the species and chain type tags. The specific initialization sequences were selected according to germline sequences in the IMGT database and are documented in Table S2. For light chains, we identified prompts corresponding to both lambda and kappa classes and divided the generation budget between the two.

The infilling therapeutic antibody CDRs is interesting but there's no mention on whether any of these infilled sequences would retain binding of the antigen. While doing an experiment is ideal, even some docking results to confirm this would be nice (SnugDock, or AF2 multimer, etc)

As the reviewer notes, there is no mention on whether the infilled sequences would retain binding of the antigen. This is because we do not expect IgLM to retain binding of the antigen but rather to act as a generative model that can generate diverse synthetic libraries from an existing antibody while efficiently sampling the natural space of antibodies. We have now expanded the text to clarify IgLM's use case:

IgLM's primary innovation is the ability to generate infilled residue spans at specified positions within the antibody sequence. In contrast to traditional generative language models that only consider preceding the residues, this enables IgLM to generate within the full context of the region to be infilled. IgLM therefore acts as a tool for developing synthetic libraries for large-scale experimental screening by diversifying regions of an existing antibody. Because IgLM is trained on a massive dataset of natural antibodies, it proposes sequences that more efficiently explore the sequence space of natural antibodies, which can reduce the fraction of non-functional antibodies in IgLM-designed libraries compared with randomized synthetic libraries.

Furthermore, to clarify the motive for infilling therapeutic antibody CDRs as a means of library diversification rather than epitope-specific binder design, we have also added the following text to the Therapeutic Antibody Diversification section:

Diversification of antibody CDR loops is a common strategy for antibody discovery or optimization campaigns. Through infilling, IgLM is capable of replacing spans of amino acids within antibody sequences, conditioned on the surrounding context. To demonstrate this functionality, we generated infilled libraries for a set of therapeutic antibodies and evaluated several therapeutically relevant properties. Based on in silico measures of developability and humanness, we show that IgLM proposes libraries containing antibody sequences resembling natural antibodies with controllable diversity, which could then be experimentally screened to discover new high-affinity binders.

The authors should consider benchmarking against something like a T5 for the therapeutic antibody infilling task where spans are predicted in-place, or even using an MLM to re-fill using bidirectional context (e.g. using AbLang)

> We have now conducted an extensive benchmark of the properties of sequences generated by alternative infilling methods, including an OAS-derived baseline and three additional protein language models (ESM-2, AntiBERTy, and ProGen2-OAS). Our results show that all methods are capable of producing sequences with favorable aggregation propensity and solubility, but only antibody-specific language models achieve these properties while retaining high humanness. We describe these results in the following updated text and updated figures:
>
> To contextualize the properties of IgLM-generated infilled libraries, we conducted a benchmark using several alternative protein language models. The benchmark includes ESM-2, a masked language model trained on diverse sequences, AntiBERTy, an antibody-specific masked language model, and ProGen2-OAS, an autoregressive language model trained on antibody sequences. We also compared with a baseline of sequences generated from the OAS data used to train IgLM. Sequences for the OAS baseline, OAS [parent], were generated by sampling from positional amino acid frequencies for loop lengths matching the parent sequence.
>
> For all infilled libraries, we predicted structures with IgFold and computed aggregation propensity, solubility, and humanness for all sequences (Figure S11). To remove length-dependent biases from the evaluation, we compared the developability properties of only loops matching the parent CDR H3 loop length. In general, we found that all methods were able to generate infilled libraries with improved aggregation propensity and solubility relative to the parent sequences (Figure 4F-G). This illustrates the utility of drawing from informed sequence distributions (such as those derived from OAS or learned by language models), rather than randomly mutating sequences as is the norm for library construction. The OAS baseline performed particularly well, indicating that the natural makeup of CDR H3 loops are biophysically well-behaved. However, to produce human-like antibody libraries, we found that antibody-specific language models were significantly more effective than alternative approaches (Figure 4H). Among these models, IgLM produced slightly more human-like sequences than ProGen2-OAS, in accordance with the lower infilling perplexity demonstrated on the heldout set human sequences (Figure S1).

The selection of 49 antibodies seems unusual, some of these antibodies (e.g. muromonab) is chimeric/mouse so I would imagine it's trivial to get higher OASis/humanness scores

> We have now expanded on our motivation for selecting the set of 49 antibodies for our developability benchmarking. These 49 antibodies were selected because they had

experimentally determined structures at the time the set was assembled, and they had been previously used in development of the Therapeutic Antibody Profiler. Although we did not ultimately make use of the experimental structures, future studies may find value in comparing the predictions with ground truth structures. We have provided these justifications in the following updated text:

To evaluate the utility of infilling with IgLM for diversifying antibody sequences, we created infilled libraries for 49 therapeutic antibodies from Thera-SAbDab. These antibodies were selected because they had experimentally determined structures and had been previously evaluated for developability screening (Raybould et al., *PNAS* 2019).

Minor Comments:

Authors don't describe use of humanized mice in antibody selection, which has been revolutionary for how we discover antibodies

As the reviewer notes, transgenic animal systems (particularly humanized mice) have been very successful for antibody discovery. We have now noted this in the introduction, prior to introducing display technologies, which are closer to the focus of this work. The updated text is provided below:

Traditionally, monoclonal antibodies (mAbs) have been obtained using hybridoma technology, which requires the immunization of animals, or transgenic animal systems, which involve integration of human immune loci into alternative species (e.g., mice).

Intro mentions diversity of antibodies but does the justification for 10^13 make sense? CDRs inherently have a lot of bias - CDR3 usually ends with ARD and end FDY, so that already just means (for a typical CDR3 of length 10) that there are 20^4 variants - and that's assuming amino acids are uniformly distributed

As the reviewer's comment highlights, it is difficult to place a bound on the number of feasible antibody sequences. As our initial estimate demonstrates, the combinatorial space of possible sequences over a 10-residue span is quite large (10^20 sequences). However, as we argue in the text, the subspace of natural antibody sequences, which resemble those produced by immune systems, is likely much smaller. Although heuristics like the reviewer notes regarding commonly observed beginning and ending motifs may begin to narrow in on the space of natural antibody sequences, they are not absolute. We have adjusted the following text to clarify this point.

To discover antibodies with high affinity, massive synthetic libraries on the order of 10^10-10^11 variants must be constructed. However, the space of possible synthetic

antibody sequences is very large (diversifying 10 positions of a CDR yields 20^10 ≈ 10^13 possible variants), meaning these approaches still significantly undersample the possible space of sequences. Further, sequences from randomized libraries often contain substantial fractions of non-functional antibodies. These liabilities could be reduced by restricting libraries to sequences that resemble natural antibodies, and are thus more likely to be viable therapeutics.

The intro could do with citing more recent examples of representation learning models from AbSci (Bachas et al. 2022), sequence generation models such as ProGen2 (Nijkamp et al 2022). Other papers that should be discussed are cases where language models are used to mutate and design antibodies (Hie et al. 2022). AntiBERTa paper seems to not be properly cited, it's cited with the AbLang paper from Olsen et al. Prihoda paper is now out in mAbs which should be updated.

We have now updated the citations mentioned by the reviewer in the introduction, as well as provided context for the works from Hie and Bachas, in the following updated text:

For example, the ESM family of models (trained for masked language modeling) have been applied to representation learning, variant effect prediction, and protein structure prediction. Masked language models have also shown promise for optimization and humanization of antibody sequences through suggestion of targeted mutations (Hie et al.).

The Sapiens models were trained on 20M and 19M heavy and light chains respectively, and shown to be effective tools for antibody humanization. Similarly, likelihoods from antibody-specific masked language models have also been used as a proxy for immunogenic risk (or naturalness) (Bachas et al.).

How are the validation sets created? It could be worth pruning to make sure no study is shared between training and validation sets when pre-training IgLM to avoid perplexities looking lower

In the expanded Methods section, we have now provided a more detailed description of the splitting procedure for training and evaluation of IgLM. Due to the highly conserved nature of antibody sequences, and the goal of building a model broadly useful for antibody generation, we have relied on clustering at a fairly high sequence identity threshold to create data splits. The process and reasoning are provided in the following updated text:

To train IgLM, we collected unpaired antibody sequences from the Observed Antibody Space (OAS). OAS is a curated set of over one billion unique antibody sequences compiled from over eighty immune repertoire sequencing studies. After removing

sequences indicated to have potential sequencing errors, we were left with 809M unique antibody sequences. We then clustered these sequences using LinClust at 95% sequence identity, leaving 588M non-redundant sequences. The distribution of sequences corresponding to each species and chain type are documented in Figure 1B and Table S1. The dataset is heavily skewed towards human antibodies, particularly heavy chains, which make up 70% of all sequences.

The highly conserved nature of antibody sequences, which are recombined and mutated from a common set of germline components, makes construction of distinct training and validation sets challenging, as overly aggressive splitting may result in exclusion of entire germline lineages from training. For this work, we held out a random 5% of the clustered sequences as a test set to evaluate model performance. Of the remaining sequences, we randomly selected 558M sequences for training and 1M for validation. This splitting criteria ensures that the model is exposed to all of the available conserved regions of antibody sequences, but can be evaluated on how well it captures mutations to those sequences.

Why are structures predicted with AF2 multimer in the intro / Fig 1 but then IgFold later?

Apriori, it is difficult to estimate what sampling temperatures will produce reasonable sequences from a language model. To set informed boundaries, we carried out an initial analysis using AlphaFold2-Multimer to identify reasonable limits for sampling temperature with IgLM. Based on the results of our preliminary structure prediction analysis, we found that sequences generated by IgLM up to T=1.2 produced confidently predicted structures. To enable the scale of analysis in subsequent experiments, we then used IgFold for high-throughput predictions. We have described this rationale in the following updated text:

In general, IgLM generates sequences with correspondingly confident predicted structures at lower temperatures (up to 1.2), before beginning to degrade in quality at higher temperatures (Figure 1C). For subsequent experiments, we sampled with a maximum temperature of 1.2 to remain within foldable antibody space, and utilized the much faster IgFold model for high-throughput structure predictions.

Why is it that some designs only have CDR3 of length 1? And for these cases, what is the likelihood of the sequence?

As the reviewer's comment suggests, the very short CDR H3 loops generated by IgLM are quite rare (if ever occurring) and are likely not ideal for inclusion in antibody libraries. To assess the favorability of these short loops by IgLM, we computed infilling perplexities for all of the generated loop sequences. In a new supplemental figure, we show that the model scores these short loops very unfavorably, indicating they likely

emerged by chance due to extensive sampling. We have added the following text to present these results:

The median length of infilled loops across antibodies ranges from 11 to 16 residues. IgLM occasionally generated very short CDR H3 loops (fewer than five residues), which were assigned correspondingly low log likelihoods by the model (Figure SX).

Authors don't describe how IgLM generates paired sequences - dataset seems to be unpaired and even input seems unpaired?

As the reviewer notes, IgLM models only single sequences. We have updated the following text to clarify that we generated and subsequently paired heavy and light chains for the purpose of structure prediction.

Heavy and light chain sequences were generated independently of each other, as IgLM only considers single chains. Sequences were then paired according to sampling temperature and their structures predicted using AlphaFold-Multimer.

Fig 3B it might be good to indicate the length of the native H3

We have now updated Figure 2B to indicate the lengths of the parent CDR H3 loops prior to infilling.

Fig 4E might be better as a heatmap or some kind of 3D scatterplot where we correlate T and P and have the median change in OASis identity

We have now updated Figure 4E to show a heatmap as the reviewer suggested.

Fig 4F the classification of human sequences another good baseline could be using the ANARCI HMM cutoffs? Especially as the authors point out ANARCI's annotation of species isn't perfect

Indeed, as the reviewer notes, the HMMs from ANARCI are a strong baseline for human sequence classification. In Figure 4F, one of the baselines adapted from Prihoda et al. (source of all baselines aside from ProGen2 models), is the germline content reported by ANARCI for the best V- and J-gene matches. This method slightly outperforms IgLM and is among the most effective methods for determining whether a sequence is human-like.

Reviewer 2

Shuai and colleagues here describe a language model for antibody design. They claim that their model can generate full length sequences for various species.

While the work performed is extensive, after reading it, it remains a bit opaque to the reader what the paper has in fact demonstrated. The advance of previous research needs to be much more clear. Also, we would like to see more surprising things than full sequence reconstruction of different animals.

Major comments

The novelty this paper provides is a method that enables generative infilling of antibody sequences of various sizes while conditioning the infilled regions on species type and chain type. I imagine this could enable the generation of some interesting antibody sequences, e.g. a sequence that is part human/part mice/part something else. However, It is not obvious to me what the practical utility of this method is supposed to be. Maybe the authors should include an example where the partial modification of an available antibody sequence would be useful over simply generating sequences completely from scratch, especially if it is not possible to condition the sequences for desirable properties (binding affinity, developability parameters). Even the humanization of a murine antibody is not an option since there is no guarantee that the model will maintain the same binding affinities.

> We have now expanded on the motivation for developing IgLM by further discussing the role of antibody sequence libraries in discovery and optimization, and highlighting the potential of generative models of protein sequences to replace existing diversification technologies. This leads into our contrast with existing protein language models, which are poorly suited for diversification of antibody sequences. The updated text is provided below.
>
> Design of antibody libraries typically focuses on diversification of the CDR loop sequences in order to facilitate binding to a diverse set of antigens. Through traditional diversification technologies, many putative antibody sequences can be produced and subjected to experimental screening, enabling the discovery or optimization of specific antibodies. However, such techniques typically produce large fractions on non-viable or poorly behaved sequences, as they are not constrained to the natural space of antibody sequences. Generative models of protein sequences, such as language models, offer an alternative means of efficiently sampling from the natural space of proteins to produce large libraries of sequences. However, existing approaches to protein sequence generation (including antibodies) typically adopt left-to-right decoding strategies. While these models have proven effective for generation of diverse and functional sequences, they are ill-equipped to re-design specific segments of interest within proteins. To

address this limitation, we developed IgLM, an infilling language model for immunoglobulin sequences.

The first half of the results is all about how well the model generates antibody sequences conditioned on species and chain types. But none of this is very surprising, since they were trained with data labelled with that information already (which also makes it not completely self-supervised) and also that information is encoded in non-variable regions, which is not what the infilling task targets anyways.

> Although the model was trained with conditioning information, the considerable imbalance in the training dataset presents challenges for effectively generating under certain conditioning scenarios. Through our analyses, we explore the capabilities of the model in various settings and demonstrate to what extent these data imbalances affect generation fidelity. Further, although prior models such as the original ProGen have been trained with conditioning, ours is the first study to quantitatively assess adherence to conditioning information.

pg. 3, lines 83-104. It is unclear what the infilling perplexity tells us about the language model's effectiveness. All the reported results are expected (IgLM achieves lower perplexity than IgLM-S, longer sequences have higher perplexity, etc.), and there is no comparison to any other language model to give a reader a good sense on whether IgLM's perplexity scores are particularly good, average, or bad.

> We have now expanded our language modeling evaluation to include additional autoregressive language models (ProGen2-base and ProGen2-OAS) and to assess the impact of bidirectional context on IgLM infilling perplexity. These results are described in the followed updated text:
>
> We compared the infilling perplexity of IgLM and IgLM-S given bidirectional context (IgLM [bi] and IgLM-S [bi]) and preceding context only (IgLM [pre] and IgLM-S [pre]) on a heldout test dataset of 30M sequences. We additionally computed infilling perplexity for ProGen2-base and ProGen2-OAS, which only utilize preceding context. Results are tabulated by CDR loop for each method (Figure 1D). As expected, the CDR3 loop, which is the longest and most diverse, has the highest infilling perplexity for all methods. For IgLM, providing bidirectional context yielded reduced complexity, demonstrating that the sequence following CDR loops is important for determining their content. Both ProGen2 models evaluated have 764M parameters, significantly more than the 13M parameters of IgLM. However, with bidirectional context, IgLM is able to better fit the distribution of CDR loops than either model, demonstrating the importance of aligning the model pre-training objective with the downstream task.

In Figure 1D, we also compared IgLM infilling perplexity to methods using only preceding context (IgLM [pre], IgLM-S [pre], ProGen2-base, ProGen2-OAS). For these methods, rather than compute perplexity using our infilling formulation procedure, we instead provide only the amino acid sequence context preceding the span to be predicted. We additionally prepend the appropriate conditioning tokens for each model (i.e. the chain-type and species-of-origin tokens for IgLM, and the 1 character token for the ProGen2 models) prior to inference. We then compute per-token perplexity over the predicted span and the first residue following the span, where the first residue following the span acts as a proxy for the [ANS] token. In this way, we compute infilling perplexity over the same number of tokens with these methods while only providing the preceding amino acid sequence context.

pg. 7, lines 228-237. Since IgLM is tasked with only predicting CDR H3 loops while the rest of the antibody is given, is it surprising that it generates more human-like antibodies? Also, what does it mean to have even more human-like than the parent antibody (line 231), when the parent antibodies have been already optimized to be so?

We have added the following text to the humanness evaluation of IgLM infilled sequences to contextualize the results and clarify how IgLM is able to achieve improvements over the parent sequences:

When compared to their respective parent antibodies, sequences infilled by IgLM were typically more human-like (Figure 4D). This is expected, given that IgLM is trained on natural human antibodies, but not trivial as the parent sequences have been optimized and shown to be safe in humans. To achieve higher humanness, sequences from IgLM must better adhere to the natural distribution of human antibodies than the parent sequences.

pg. 7, lines 209-227. All developability measures are done in silico with computational tools. How reliable is it to conclude that IgLM generated sequences with increased developability? For all sequences generated, it's not entirely completely clear to us how the authors determine whether a given sequence is actually an antibody sequence that belong to a certain species. Can this authors clarify this? How do they quantitatively test for this?

Indeed, all analyses performed on the generated sequences utilized in silico tools. To ground our developability calculations, we have utilized well-established tools, which themselves have been experimentally validated (CamSol and SAP score). However, as reflected in the original publications, neither tool is a perfect reflection of the biophysical properties they aim to compute. For species and chain type classification, we utilized the ANARCI software, which performs germline matching for a given sequence against a database of curated sequences. We have expanded the Methods to include greater detail of both developability calculations and sequence classifications:

To assess the developability and humanness of infilled therapeutic antibody sequences, we utilized a set of in silico tools previously developed for antibodies. Aggregation propensity was calculated based on the predicted Fv structures for each antibody using the Rosetta implementation of the spatial aggregation propensity (SAP) score. Solubility was calculated based on sequence alone, using the public CamSol-Intrinsic web server. To measure humanness (a proxy for immunogenicity), we used the BioPhi OASis identity. OASis identity measures the fraction of 9-mers for a given sequence that have been observed in human repertoires in the OAS database.

To evaluate the adherence of IgLM-generated sequences to provided species and chain type conditioning tags, we used the ANARCI software. ANARCI uses a set of antibody-specific HMMs to compare a given antibody to a database of germline sequences across several species and chain types. To classify the chain type and species for generated sequences, we used the corresponding species and chain type for the top V-gene match returned by ANARCI.

"For mouse 137 and rhesus light chains, IgLM generates the correct species in 34.89% and 88.14% of cases, respectively" → why so low for mouse?

As the reviewer notes, the mixed performance at generating mouse sequences was surprising given that mice are among the best-represented species in the training dataset. We have added the following text to provide some explanation for this result:

For mouse and rhesus light chains, IgLM generates the correct species in 34.89% and 88.14% of cases, respectively (Table S3). The disproportionately low recovery of mouse sequences may be due to inclusion of transgenic mice immune repertoires, which are harvested from mice but consist of human genetic material.

"As temperature increases, sequences for every species begin to diverge from germline, effectively acquiring mutations." Are these mutations nature-like? Meaning, are they inserted according to mutational hotspot motifs?

We have now added a new analysis of the generated full-length sequences, in which we compute Chothia-numbered positional entropies for each species and chain type with increasing temperature. As expected, the majority of mutations appear in the CDR loops, with higher temperature increasing the rate of mutation. We have added the following new text, as well as an additional supplemental figure, presenting these results:

As temperature increases, sequences for every species begin to diverge from germline, effectively acquiring mutations. To evaluate whether these mutations emerge at

biologically relevant positions, we calculated the positional entropy of generated sequences according to the Chothia numbering scheme. As expected, we observe significantly higher entropy in the CDR loops, with temperature further increasing the entropy at these positions (Figure SX).

For the sequences generated, does their V gene distribution follow that of the training data?

We agree with the reviewer that a comparison of V-gene distribution over the training dataset and generated sequences would be an interesting analysis. However, the frequent N-terminal truncations present in the OAS database complicate this analysis on two fronts. First, with large segments of the N-terminal regions of sequences missing in the training dataset, classification of the V-genes may be less accurate. Second, due to these truncations, we needed to use initial-residue prompts to produce full-length sequences. This has the effect of biasing the generated sequences towards particular germline sequences.

Pg. 5, lines 171ff: The therapeutic antibody diversification part is not clear to me: how can the authors be certain that the generated antibodies still bind the same target?

As the reviewer notes, there is no mention on whether the infilled sequences would retain binding of the antigen. This is because we do not expect IgLM to retain binding of the antigen but rather to act as a generative model that can generate diverse synthetic libraries from an existing antibody while efficiently sampling the natural space of antibodies. We have now expanded the text to clarify IgLM's use case:

IgLM's primary innovation is the ability to generate infilled residue spans at specified positions within the antibody sequence. In contrast to traditional generative language models that only consider preceding the residues, this enables IgLM to generate within the full context of the region to be infilled. IgLM therefore acts as a tool for developing synthetic libraries for large-scale experimental screening by diversifying regions of an existing antibody. Because IgLM is trained on a massive dataset of natural antibodies, it proposes sequences that more efficiently explore the sequence space of natural antibodies, which can reduce the fraction of non-functional antibodies in IgLM-designed libraries compared with randomized synthetic libraries.

Furthermore, to clarify the motive for infilling therapeutic antibody CDRs as a means of library diversification rather than epitope-specific binder design, we have also added the following text to the Therapeutic Antibody Diversification section:

Diversification of antibody CDR loops is a common strategy for antibody discovery or optimization campaigns. Through infilling, IgLM is capable of replacing spans of amino acids within antibody sequences, conditioned on the surrounding context. To

demonstrate this functionality, we generated infilled libraries for a set of therapeutic antibodies and evaluated several therapeutically relevant properties. Based on in silico measures of developability and humanness, we show that IgLM proposes libraries containing antibody sequences resembling natural antibodies with controllable diversity, which could then be experimentally screened to discover new high-affinity binders.

At what temperature threshold do antibodies turn from "being diversified" to "unnatural/random sequences"

For this work, we judged the boundary between diversification and randomness according to AlphaFold-Multimer structure predictions. To find this boundary, we conducted an experiment where we gradually increase sampling temperature until sequences were no longer confidently predicted to fold into structures. We have expanded the following text and added a more detailed figure to present these results. We have additionally expanded the Methods section to provide more details on sampling parameters.

In general, IgLM generates sequences with correspondingly confident predicted structures at lower temperatures (up to 1.2), before beginning to degrade in quality at higher temperatures (Figure 1C). For subsequent experiments, we sampled with a maximum temperature of 1.2 to remain within foldable antibody space, and utilized the much faster IgFold model for high-throughput structure predictions.

Have you tried to classify generated from OAS sequences to measure whether the LM introduced obvious biases?

To assess how realistic the infilled loops generated by IgLM are, we have now computed infilling perplexities for all of the sequences produced by IgLM for the set of therapeutic antibodies. These results (presented in a new supplemental figure) show that IgLM favors loop lengths near the natural distribution, though occasionally produces loops on extremes, which are scored unfavorably by the model. These results are described in the following new text:

The median length of infilled loops across antibodies ranges from 11 to 16 residues. IgLM occasionally generated very short CDR H3 loops (fewer than five residues), which were assigned correspondingly low log likelihoods by the model (Figure SX).

Minor comments

Page 5, line 57: the (GPT2) addition is not true since GPT2 refers to a transformer pre-trained on text data.

We have now updated the following text to clarify that our model utilizes the same architecture as GPT2, but is not trained on text:

The IgLM model uses a Transformer decoder architecture based on a modified version of the GPT-2 Transformer as implemented in the HuggingFace Transformers library.

pg. 3, line 73. Can the authors provide a definition of sampling temperature? While they explain what the numbers indicate, it would be good for the reader to have a more precise idea about this measurement.

We have now extended the methods to include further technical details on our sampling procedure for generating sequences. The following added text describes sampling temperature and nucleus sampling probability:

As we sampled sequences under the model, we applied temperature sampling to shape the probability distribution for each token. Applying temperature T corresponds to scaling the logits z from the last layer before applying softmax:

[EQUATION]

where $p(x_i)$ denotes the probability assigned during sampling to token i out of n possible tokens in the vocabulary. Intuitively, sampling with higher temperatures results in more diverse sequences, with the probability distribution across tokens becoming nearly uniform when T is large.

In addition to applying temperature, we also applied nucleus sampling to vary the diversity of sequences generated by IgLM. In nucleus sampling with probability P, the probability distribution during sampling is clipped such that only the smallest set of tokens whose cumulative probability exceeds P are considered during sampling. Intuitively, a lower P restricts sampling to highly probable tokens, which decreases the diversity of sequences while increasing confidence.

The authors may want to add this paper to the ref list
https://www.nature.com/articles/s41598-021-85274-7

We have now updated the introduction to highlight the contribution of this work, which has trained LSTM models on phage display data to optimize antibody sequences.

LSTMs have also been trained on phage display data to aid in discovery of optimized variants (Saka et al.).

N-Terminal truncations in OAS: why didn't the authors just remove such sequences from the dataset? What is their frequency in the OAS?

> The frequency of truncated sequences has been reported to be above 40% missing at least fifteen of the N-terminal residues (Olsen et al., *Bioinformatics Advances* 2022). Because of the pervasiveness of these truncations, it would significantly reduce the amount of data available to train on. Instead, we expected the model could be encouraged to start at the first residue by providing a short largely-conserved motif, and then generate freely. We have expanded the text describing the justification for using residue prompts, as well as the process for identifying appropriate initial residues in the following text:
>
> However, we observed that the sequences frequently featured N-terminal truncations. These truncations are frequently observed in the OAS database used for training, with over 40% of sequences missing the first fifteen or more residues. For heavy chains, these N-terminal deletions appeared as a left-shoulder in the sequence length distribution (Figure 2B, left) with lengths ranging from 100 to 110 residues. For light chains, we observed a population of truncated chains with lengths between 98 and 102 residues (Figure 2B, right). To address truncation in generated sequences, we utilized a prompting strategy, wherein we initialize each sequence with a three-residue motif corresponding to the species and chain type tags. The specific initialization sequences were selected according to germline sequences in the IMGT database and are documented in Table S2. For light chains, we identified prompts corresponding to both lambda and kappa classes and divided the generation budget between the two.

There are several typos throughout the text.

> We thank the reviewer for bringing attention to typos throughout the text, we have now corrected several such issues.

Reviewer 3

Shuai et al. describe an antibody specific language model, IgLM. The model is trained autoregressively and allows a user to condition on relevant properties and the right-sequence context during autoregressive sequence generation, which would potentially enable improved and more controllable sequence sampling and design. Overall, the I think the work is interesting, but also preliminary. More specifically, I would very much appreciate major revisions in the form of comparisons with previous antibody-specific language models, as well as general protein language models trained on generic protein sequence corpora, before I would be able to make a recommendation for publication. This is important for assessing

where and when IgLM provides an advantage over existing tools. Additional details are provided below.

The code availability and usability are very good. The GitHub describes a very easy-to-use command line implementation. The model availability is also very good, and is easily downloadable/installable. Optionally, it would be nice to also make the processed training data available, or if the dataset size is too large to fit in common repositories like Zenodo or figshare, making the dataset processing scripts available would be nice to have.

Major Comments:

A major question I had while reading the paper, and which I would like to see addressed, is how well this language model compares to previous antibody language models (including AntiBERTy/a, AbLang, Sapiens) as well as to general protein language models (ESM, ProGen with and without antibody-specific conditioning or training). Foremost, the authors should compare their perplexity results to these language models, and I would be very curious as to how this model compares.

> We have now expanded our language modeling evaluation to include additional autoregressive language models (ProGen2-base and ProGen2-OAS) and to assess the impact of bidirectional context on IgLM infilling perplexity. Several of the models noted by the reviewer are not amenable to direct comparison with IgLM, as they are trained via masked language modeling, for which perplexity calculation is not possible. Our comparison with ProGen2-base and ProGen2-OAS highlights the value of antibody-specific training, as well as the utility of bidirectional context for CDR loop infilling. These results are described in the followed updated text:
>
> We compared the infilling perplexity of IgLM and IgLM-S given bidirectional context (IgLM [bi] and IgLM-S [bi]) and preceding context only (IgLM [pre] and IgLM-S [pre]) on a heldout test dataset of 30M sequences. We additionally computed infilling perplexity for ProGen2-base and ProGen2-OAS, which only utilize preceding context. Results are tabulated by CDR loop for each method (Figure 1D). As expected, the CDR3 loop, which is the longest and most diverse, has the highest infilling perplexity for all methods. For IgLM, providing bidirectional context yielded reduced complexity, demonstrating that the sequence following CDR loops is important for determining their content. Both ProGen2 models evaluated have 764M parameters, significantly more than the 13M parameters of IgLM. However, with bidirectional context, IgLM is able to better fit the distribution of CDR loops than either model, demonstrating the importance of aligning the model pre-training objective with the downstream task.
>
> In Figure 1D, we also compared IgLM infilling perplexity to methods using only preceding context (IgLM [pre], IgLM-S [pre], ProGen2-base, ProGen2-OAS). For these methods,

rather than compute perplexity using our infilling formulation procedure, we instead provide only the amino acid sequence context preceding the span to be predicted. We additionally prepend the appropriate conditioning tokens for each model (i.e. the chain-type and species-of-origin tokens for IgLM, and the 1 character token for the ProGen2 models) prior to inference. We then compute per-token perplexity over the predicted span and the first residue following the span, where the first residue following the span acts as a proxy for the [ANS] token. In this way, we compute infilling perplexity over the same number of tokens with these methods while only providing the preceding amino acid sequence context.

Closely related to the above point is that the authors should compare across different language models according to the various downstream metrics described in the paper, such as the coherence of downstream structure prediction, developability, and humanness. There is some comparison to this effect (e.g., involving ProGen2 trained on OAS), but the comparison should be much more exhaustive in order to communicate where this model could or could not be useful.

We have now conducted an extensive benchmark of the properties of sequences generated by alternative infilling methods, including an OAS-derived baseline and three additional protein language models (ESM-2, AntiBERTy, and ProGen2-OAS). Our results show that all methods are capable of producing sequences with favorable aggregation propensity and solubility, but only antibody-specific language models achieve these properties while retaining high humanness. We describe these results in the follow updated text and updated figures:

To contextualize the properties of IgLM-generated infilled libraries, we conducted a benchmark using several alternative protein language models. The benchmark includes ESM-2, a masked language model trained on diverse sequences, AntiBERTy, an antibody-specific masked language model, and ProGen2-OAS, an autoregressive language model trained on antibody sequences. We also compared with a baseline of sequences generated from the OAS data used to train IgLM. Sequences for the OAS baseline, OAS [parent], were generated by sampling from positional amino acid frequencies for loop lengths matching the parent sequence.

For all infilled libraries, we predicted structures with IgFold and computed aggregation propensity, solubility, and humanness for all sequences (Figure S11). To remove length-dependent biases from the evaluation, we compared the developability properties of only loops matching the parent CDR H3 loop length. In general, we found that all methods were able to generate infilled libraries with improved aggregation propensity and solubility relative to the parent sequences (Figure 4F-G). This illustrates the utility of drawing from informed sequence distributions (such as those derived from OAS or learned by language models), rather than randomly mutating sequences as is the norm

for library construction. The OAS baseline performed particularly well, indicating that the natural makeup of CDR H3 loops are biophysically well-behaved. However, to produce human-like antibody libraries, we found that antibody-specific language models were significantly more effective than alternative approaches (Figure 4H). Among these models, IgLM produced slightly more human-like sequences than ProGen2-OAS, in accordance with the lower infilling perplexity demonstrated on the heldout set human sequences (Figure S1).

The Methods section is extremely sparse and needs to be improved. The authors need to provide enough experimental detail in the Methods to reproduce all analyses described in the paper. This not only applies to the language model description and training details as currently provided (which are still limited) but also to all of the downstream tasks and additional benchmarking experiments.

We have now significantly expanded the Methods section, including more detailed descriptions of the model, infilling perplexity calculations, generation settings, and sequence characterization.

As an ablation, what happens if the authors sample, e.g., the CDR3 without including the right context in their infilling task? Given that a central claimed improvement is the ability to condition on both the left- and right-sequence context, the actual effect of this would be interesting to see and I was not able to find an analysis like this in the manuscript.

Our expanded language modeling evaluation now includes comparisons for IgLM models using bidirectional context and preceding context only. For IgLM, bidirectional context significantly reduces CDR loop infilling perplexity, even below that of ProGen2-OAS, which has over fifty times more parameters. The results are described in the following updated text:

We compared the infilling perplexity of IgLM and IgLM-S given bidirectional context (IgLM [bi] and IgLM-S [bi]) and preceding context only (IgLM [pre] and IgLM-S [pre]) on a heldout test dataset of 30M sequences. We additionally computed infilling perplexity for ProGen2-base and ProGen2-OAS, which only utilize preceding context. Results are tabulated by CDR loop for each method (Figure 1D). As expected, the CDR3 loop, which is the longest and most diverse, has the highest infilling perplexity for all methods. For IgLM, providing bidirectional context yielded reduced complexity, demonstrating that the sequence following CDR loops is important for determining their content.

In Figure 1D, we also compared IgLM infilling perplexity to methods using only preceding context (IgLM [pre], IgLM-S [pre], ProGen2-base, ProGen2-OAS). For these methods, rather than compute perplexity using our infilling formulation procedure, we instead provide only the amino acid sequence context preceding the span to be predicted. We

additionally prepend the appropriate conditioning tokens for each model (i.e. the chain type and species-of-origin tokens for IgLM, and the 1 character token for the ProGen2 models) prior to inference. We then compute per-token perplexity over the predicted span and the first residue following the span, where the first residue following the span acts as a proxy for the [ANS] token. In this way, we compute infilling perplexity over the same number of tokens with these methods while only providing the preceding amino acid sequence context.

Minor Comments:

One thing that is often seen when training on OAS, followed by sample sequences based on high/maximum likelihood, is that the generated sequences reproduce the sequence distribution of the antibody germline. This would potentially make such models useful for germline reversion (though maybe less useful for tasks like affinity maturation). I was wondering if the authors could comment on this in their Discussion. An optional revision is to evaluate the model's ability to identify the germline sequence or to see if there's concordance between maximum likelihood mutations and those recommended by common germline reversion tools.

As the reviewer notes, IgLM tends to reproduce germline sequences at low temperatures (Figure 2E). We have now expanded on the analysis of this behavior through the following added text and a new supplemental figure:

As temperature increases, sequences for every species begin to diverge from germline, effectively acquiring mutations. To evaluate whether these mutations emerge at biologically relevant positions, we calculated the positional entropy of generated sequences according to the Chothia numbering scheme. As expected, we observe significantly higher entropy in the CDR loops, with temperature further increasing the entropy at these positions (Figure SX).