

The Virtual Lab of AI Agents Designs New SARS-CoV-2 Nanobodies

Corresponding Author: Professor James Zou

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)

This is the first report using multiple agents with domain expertise, powered by LLMs and AI tools, to conduct a scientific endeavor with purportedly minimal human supervision. The in silico discovery of two potent nanobodies vs relevant SARS-CoV-2 variants was validated by subsequent lab studies by the team of human authors. This represents a very different and creative way to achieve human-A.I. symbiosis, as compared with other ideas such as "lab-in-the-loop", the A.I. Scientist, or Co-Scientist, by maximizing the input of a team of agent A.I.s, exemplified by the nanobody design pipeline, which would be deemed as a very complex task.

Specific points and questions:

1. The actual role of the human researcher remains fuzzy and is important to further clarify. At various points in the paper, it is noted the human chose the computational tools, the constraints in experimental validation, noted mistakes and fixed code, helped with prompt engineering, wrote 18.5% of the words for the virtual team selection, and "wrote the agenda for the meeting specifying the topic of discussion." A dedicated section should be added to the paper to fully understand the extent of autonomy of the multi-agent team. The words ascribed to human researcher oversight and input as "relatively minimal input" are too subjective.
2. The timeline, number of team and individual meetings should be plainly spelled out. The brief 5-10 minute meetings, often in parallel, seem rather limited vs the complex task that was assigned.
3. The input of retrieval augmentation generation is alluded to but needs to be more specific.
4. Was this the first project introduced to multi-agents by the authors? Were there any prior attempts? The potential zero-shot nature of success would be noteworthy.
5. Ironically, without updated A.I. tools such as AlphaFold3 the project came up with designs for nanobodies directed vs the current circulating SARS-CoV-2 variants (KP.3, JN.1). The limitation, as acknowledged, of using somewhat dated models didn't seem to affect the relevant design.
6. How was the Scientific Critic agent trained?
7. Please define "LLM temperature" as presented in Supplemental Figure 1.

Review by Eric Topol, Scripps Research

Referee #2

(Remarks to the Author)

This work established a so-called "Virtual Lab", where the lab members are large language model (LLM) agents with different scientific backgrounds. Research in the Virtual Lab is conducted through a series of team meetings between the LLM agents, as well as a limited amount of human input. Each agent represents an expert for a specific scientific field like Machine Learning, Immunology, a PI and a Scientific critic. Agents are transformed to experts via prompting. Subsequently, the agents engage in scientific discussion (guided by the PI agent) to solve an open challenge. To showcase its power, the authors tasked the Virtual Lab with designing novel nanobody binders to contemporaneous variants of SARS-CoV-2. The authors demonstrate that the computational pipeline proposed by the Virtual Lab successfully utilized sophisticated tools in computational protein engineering and machine learning such as ESM, AlphaFold-Multimer, and Rosetta, to design two nanobodies able to bind the recent JN.1 or KP.3 variants of SARS-CoV-2, while maintaining binding to the ancestral SARS-

CoV-2 spike protein. In doing so, this work proposes a novel potential framework which leverages LLMs for the design and execution of scientific research. Overall, the work is creative and novel, and it is particularly impressive that the Virtual Lab was able to successfully build and run the nanobody development pipeline with such a little amount of human input, raising the prospect of democratizing sophisticated computational pipelines for laboratories which may not have the expertise internally. However, the authors' case for the effectiveness/added-value of the Virtual Lab could be made more compelling by comparing their approach with more naïve one(s) to illustrate its effectiveness compared to some reference point.

Comments:

1. Conceptually, the Virtual Lab approach is very compelling. However, for the present use-case (designing SARS-CoV-2 nanobodies) it is unclear how well the Virtual Lab pipeline works compared to existing approaches. The authors should consider comparing the Virtual Lab pipeline with a more conventional approach that uses the same or similar computational tools. One potential comparison could be to use the approach from Hie...Kim. Nature Biotechnology 2023, so use ESM to suggest nanobody mutations over the course of the same number of rounds. If there is a similar (or better) success rate using this previously published approach, it may suggest that the conversations between LLM agents that define the Virtual Lab framework don't necessarily improve experimental outcomes, but rather serve to automate and synchronize existing computational tools.

2. How important is it for the LLM agents to have different backgrounds? My understanding is that all agents are instances of the same underlying GPT-4 model. Are the different specialists (e.g., Immunologist, Machine Learning Specialist) uniquely fine-tuned or prompted in a way that makes them substantially different from each other? It would be valuable to explicitly demonstrate the added value of agent specialization. For example, could an alternative version of the nanobody design pipeline be designed using random specialists or all PIs? This comparison could help clarify whether the "expertise" of the agents leads to meaningful improvements. Such a comparison could perhaps be conducted in silico.

3. It would be helpful to include some author commentary on whether the base antibodies/nanobodies selected by the models (Ty1, H11-D4, Nb21, and VHH-72) were truly good starting points, or if there might have been better alternatives. Currently, evaluation of these initial choices is done entirely by GPT-4, and while its reasoning seems sensible, LLMs are known to justify and display excessive confidence in their answers even if they are incorrect or suboptimal.

One example of this is in the section on Project Specification, Point 1: choosing nanobodies over conventional antibodies. The AI agent chooses nanobodies because:

"Nanobodies offer superior stability, tissue penetration, and ease of production, making them ideal for targeting conserved and cryptic epitopes on the SARS-CoV-2 spike protein. This choice is consistently supported across all summaries, emphasizing their advantages in accessing unique epitopes and potential for broad-spectrum activity."

There are many reasons to disagree with this statement depending on context, for example Nanobodies would not be the better choice for a therapeutic drug, where there are nearly 100 standard antibodies clinically approved and only a single nanobody. Similarly, nanobodies have a much higher risk for immunogenicity.

Overall, greater discussion (by the authors) of the study design choices made by the Virtual Lab would provide valuable context on whether the choices were well-founded.

4. It would be helpful to include additional data on the specificity of the top nanobody candidates. While the current data showing minimal binding to BSA is reassuring, including another control such as a non-SARS-CoV-2 viral glycoprotein would strengthen the conclusion that the nanobodies are indeed specific.

5. Could the authors clarify how the nanobody expression levels (very low, low, average, high, and very high) are defined?

6. The Discussion makes a bold claim that, "the Virtual Lab represents a shift from AI as a tool to AI as a partner for science research". However, it could be argued that a single human conversing with an LLM agent like GPT-4 to design and execute a research plan (which is now commonplace) already represents a form of partnership. The authors should consider softening this language.

7. The authors should consider increasing the text size in the figures, as many of the plots currently have text that is small and difficult to read.

8. The architecture currently relies solely on prompt engineering without mechanisms for weight adjustment or further tuning. The authors could consider incorporating methods that allow the agents to acquire domain specific knowledge. This would address limitations in handling complex problems where established tools are unavailable and mitigate the risk of generating oversimplified or inadequate solutions. While the authors reference RAG in the limitations, elaborating further on how this or similar strategies could be integrated into the system would strengthen the discussion.

9. LLMs have inherent limitations in recognizing the boundaries of their own knowledge, often leading to unviable or misleading suggestions (see point 3 above for a specific example). This issue could derail the research process or compromise result quality. The authors should explicitly include this limitation in the paper and discuss potential mitigation strategies to enhance the reliability of the approach.

10. The developed work does not currently benchmark its performance against state-of-the-art tools, which makes it difficult to assess the approach's competitiveness or effectiveness. Including benchmarks in future studies would provide a clearer understanding of how the proposed system performs relative to existing methodologies and strengthen its validation.

11. The authors do not provide sufficient information on the number or nature of prompt adjustments made to achieve the presented results. Examples of these adjustments, including both successful and unsuccessful prompts, would improve reproducibility. This would also help the community gain deeper insights into the robustness of the system and enable the development of more effective prompt strategies. Also, the authors could explore integrating a chain-of-thought reasoning approach in future work to improve the explainability of their outputs and address some of the inherent LLMs limitations.

12. The study does not include ablation experiments, such as comparisons with and without prompting or with and without scientific critic. Including these experiments would help isolate the contributions of individual components and provide clarity on their impact.

Referee #3

(Remarks to the Author)
SUMMARY

This manuscript describes the development of a novel AI-driven research tool called the Virtual Lab, which uses a team of large language models (LLMs) with different scientific “backgrounds” to conduct interdisciplinary research. This is innovative approach is very much in line with cutting edge AI research generally (2025 has been described as the Year of Agents), so this agent-based approach to scientific research is both creative and timely. The Virtual Lab was tasked with creating novel antibodies or nanobodies against recent variants of SARS-CoV-2. Their approach was to engineer existing nanobodies with affinity for ancestral variants in order to confer variant binding. The Virtual Lab designed a computational workflow based on ESM, AlphaFold-Multimer and Rosetta to select productive mutations in four existing nanobodies. Recombinant expression and functional testing showed that most engineered nanobodies expressed and were soluble and uncovered two nanobodies able to bind the JN.1 variant of SARS-CoV-2.

The scientific approach devised by the Virtual Lab (in silico directed evolution of existing nanobodies using a pipeline of existing modeling tools) is not particularly novel; indeed, a very similar approach was used in Hie et al in 2023. I don't necessarily view this negatively — on the contrary, I find the fact that a Virtual Lab devises at an experimental approach that mirrors those produced by fully human research groups to be a significant achievement. The use of a slightly modified method for computing log likelihood (compared to Hie et al, which used a similar workflow for antibody engineering, suggests at least some element of novelty to the Virtual Lab's approach.

Overall, this manuscript is extremely innovative, both technically and conceptually, and is certain to be of very high impact. While there are several issues that should be addressed, I am strongly in favor of publishing a revised manuscript in Nature.

MAJOR CONCERNS

- In the introduction, the authors mention limitations of other AI-focused research tools, and a recurring theme is that none are able to generalize effectively across diverse scientific disciplines. The Virtual Lab includes “people” with different expertise, but it appears that each of these experts requires substantial domain knowledge to instantiate, at least for the first time. For example, it appears that the code for each of the three primary computational tools (ESM, Rosetta, and AlphaFold) required a non-trivial amount of human intervention, as the initial LLM-produced code had multiple errors. Would providing the computationally-focused agents with a sandboxed development environment allow them to do such error correction on their own by iteratively running the generated code, identifying errors, and fixing them?

- While all of the prompts are provided, it would also be helpful to see at least a summarization of the resulting meetings. For example, it would be interesting to see how the agents arrive at their weighting scheme for ESM LLR, Rosetta dG, and AF ipLDDT.

- The human-in-the-loop organizational structure of the Virtual Lab seems designed to address perceived limitations in truly novel thinking and out-of-distribution creativity by the current generation of LLMs (although new thinking models may show improvements in this area). It would be interesting to include a bit of discussion from the authors about how they envision

this requirement changing in the future — do they view human involvement as a temporary workaround that could potentially be removed entirely at some point in the future (with the virtual PI autonomously identifying high impact scientific questions and formulating an initial hypothesis), or whether human feedback will likely always be required albeit with more open-ended instructions.

MINOR ISSUES

- The figure panel labels in Fig 1 (and data labels generally in many other figures) are very small — increasing the size would greatly aid legibility. On the other hand, I really like the avatars for the Virtual Lab members, which makes it clear and easy to understand the structure and activities of the Virtual Lab.
- When assessing engineered nanobody binding, the initial ELISA screen was described as performed using “a single nanobody dilution (0.5X)”, but it is not clear what the 0.5X is relative to — a 1:2 dilution of some unknown nanobody concentration, half the antigen concentration, or some other value? An absolute concentration, or additional information about what the relative concentration is referring to, should be provided.
- More detail about what the various expression levels (very low, low, average, high, and very high) in Fig 5 correspond to in terms of actual concentration —relative to culture volume — should be provided.
- The inclusion of an unpaired Cys in several of the engineered nanobodies would seem to be something that the ESM log-likelihood should have flagged as highly disfavored. Is there some other reason why the weighted score would rank these variants so highly, like the ipLDDT or dG values being very favorable? Are there any obvious adjustments to the WS formula that could mitigate this?
- In Fig 5C, either actual concentration values (rather than dilution) should be provided, or the starting concentration prior to dilution should be provided.
- The potential use of different LLMs for each persona is discussed, but I did not notice any detail about the specific model used for each Virtual Lab agent. That would be helpful information — was the same model used for all agents, or were specialized models used depending on the agent’s role?

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)

The authors have provided a very responsive revised paper that has been strengthened and now includes pivotal details to make the case for use of agents A.I. compelling in the future.

Referee #2

(Remarks to the Author)

The authors have improved the manuscript and addressed all of my comments. I appreciate the addition of information on prompt engineering and ablation experiments. The manuscript demonstrates an effective framework for what will be a new paradigm shift of collaboration and execution of research with AI LLM that provide more independent agency. It will especially support interdisciplinary research and open up expertise to people who are not specialists in every domain.

Referee #3

(Remarks to the Author)

The revised manuscript adequately addresses my concerns. I recommend moving forward with publication.

Referee #4

(Remarks to the Author)

The manuscript by James Zou et al presents Virtual Lab, a multi-agent framework in which a human researcher delegates a scientific project to a team of role-specialised GPT-4o-powered agents (Principal Investigator, three domain scientists, and a Scientific Critic). The human provides high-level agendas; the agents then hold “team” and “individual” meetings (each executed five times in parallel and later merged) to debate strategy, write code, and critique one another. Applied to nanobody engineering against the KP.3/JN.1 SARS-CoV-2 variant, the human and agents generated a comprehensive analysis pipeline (ESM → AlphaFold-Multimer → Rosetta pipeline), and produced 92 mutant nanobodies. Experimental testing confirmed two binders with retained or gained activity against contemporary RBDs. This specific task produced

122,462 words of discussion and code, of which only 1,596 words (1.3 %) were produced by the human user. The work exemplifies a fascinating collaborative model in which the human retains strategic oversight but offloads detailed implementation to an autonomous AI team. As

Strengths

- The study is one of the first to demonstrate that a structured LLM-powered agent hierarchy can plan, code, and execute a non-trivial protein-engineering pipeline under human oversight but with overall limited human intervention.
- The use of virtual "meetings" represents an interesting paradigm for human-AI collaboration. Agendas authored by the researcher steer the agents, the agents contribute new ideas, eg, weighing scores to rank mutations, while the Critic agent raises methodological concerns and spots coding errors; this mirrors real laboratory supervision and likely mitigates errors and hallucinations. The added benefit is that meeting transcripts expose interdisciplinary reasoning, enabling audit of agent decisions and revealing actual contribution profiles (e.g., PI 36 %, Critic 21 %)
- The Parallel-meeting merge strategy is also interesting. Replicating meetings five times reduces LLM stochasticity and encourages consensus answers, an interesting variant of majority voting
- Two experimentally validated nanobodies with activity against emergent variants attest that the pipeline can yield biologically useful molecules

Weaknesses and opportunities for improvement:

- Claims of generality and domain-agnosticism remain speculative in my view; all data arise from one nanobody-design case, expanding on a paper from another Stanford group. Extending the framework to at least one unrelated discipline (e.g. materials discovery) or subdiscipline (e.g. cell therapy) would substantiate generality. I recognize that this is extra work, and that instead the authors may want to tone down their domain-agnostic claims.
- The study's central claims, superior coordination, reduced human labour and domain generality, are supported almost exclusively by qualitative narratives. Although a "generic-agent" ablation is provided, its impact is discussed anecdotally; same for distinct prompts. There are no numerical comparisons of hit-rate, runtime, token cost, API expenditure, or actual human labor across configurations. Rigorous, task-level and global performance metrics are essential to assess whether the proposed architecture materially outperforms (i) a single well-prompted GPT-4o agent, (ii) a generic multi-agent team without roles, (iii) an expert human baseline or (iv) any other variants. It is for example difficult to assess whether removing the critic agent, changing the number of parallel meetings, etc actually decrease the chances of discovering strong binders. Likewise, I recognize that this is a lot of extra work, but it would turn the work into a rigorous scientific study as opposed to a very cool story with potential but unproven broader impact.
- Two binders out of 92 designs (~2 %) are reported in isolation; benchmarking against existing computational-antibody pipelines (e.g., the ESM-guided approach of Hie et al., 2024) would clarify relative efficiency and success probability. At first glance, Hie et al report hit rates 14–71 % in round 1 and improvements in all seven antibodies.
- Word counts alone do not capture the real supervisory burden. A timeline of human actions (prompt crafting, code review, troubleshooting, experiment planning) and their duration would make the "minimal supervision" claim more concrete.

(Remarks on code availability)

Although the authors have deposited their code on github, reproducing the pipeline remains non-trivial. Several critical steps invoke external (ColabFold/AlphaFold-Multimer, Rosetta 2024.08) and require manual environment configuration. Providing a Docker/Singularity image or scripted setup would substantially lower the barrier to adoption/testing.

Version 2:

Reviewer comments:

Referee #4

(Remarks to the Author)

The authors have provided a detailed response to my initial critiques. The revised manuscript is substantially strengthened, and they have appropriately addressed the major concerns raised in the first round of review.

Specifically, the authors have corrected the primary weakness of the original manuscript by performing three new sets of

quantitative experiments. These additions provide rigorous, data-driven support for the central claims of the paper. They added:

- A scaled-up comparison of scientist vs. generic agents demonstrating a statistically significant performance benefit for the specialized agent architecture.
- A blinded human expert evaluation across 30 research questions that provides compelling evidence that the Virtual Lab's structured, multi-agent framework produces higher-quality scientific output than either a single agent or a team of generic agents.
- A quantitative analysis of the code generated that shows that the agent-critic interaction significantly improves code quality.

In addition, the authors' rebuttal regarding the experimental hit rate is well-reasoned. They successfully argue that their multi-tool pipeline (ESM, AlphaFold-Multimer, Rosetta) is tailored to a more complex problem than direct comparisons would suggest and provide new data showing that simpler metrics like ESM scores alone do not correlate with binding, justifying their more sophisticated approach.

The authors have also appropriately toned down their claims of domain-agnosticism and provided a more concrete estimation of the human labor involved, making the "minimal supervision" claim more credible.

In my opinion, the authors have been very diligent in their revisions. They have transformed the manuscript from a "very cool story" into a rigorous scientific study with strong, quantifiable evidence supporting its conclusions. The manuscript is much improved, and I recommend its publication.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

We thank all the reviewers for their helpful comments, which have really helped us to improve the manuscript. We have carefully revised the paper to incorporate the suggestions.

Referee 1

This is the first report using multiple agents with domain expertise, powered by LLMs and AI tools, to conduct a scientific endeavor with purportedly minimal human supervision. The *in silico* discovery of two potent nanobodies vs relevant SARS-CoV-2 variants was validated by subsequent lab studies by the team of human authors. This represents a very different and creative way to achieve human-A.I. symbiosis, as compared with other ideas such as "lab-in-the-loop", the A.I. Scientist, or Co-Scientist, by maximizing the input of a team of agent A.I.s, exemplified by the nanobody design pipeline, which would be deemed as a very complex task.

Thank you very much for your encouraging comments and helpful suggestions! We really appreciate it.

Specific points and questions:

1. The actual role of the human researcher remains fuzzy and is important to further clarify. At various points in the paper, it is noted the human chose the computational tools, the constraints in experimental validation, noted mistakes and fixed code, helped with prompt engineering, wrote 18.5% of the words for the virtual team selection, and "wrote the agenda for the meeting specifying the topic of discussion." A dedicated section should be added to the paper to fully understand the extent of autonomy of the multi-agent team. The words ascribed to human researcher oversight and input as "relatively minimal input" are too subjective.

Thank you for this suggestion! We have added more information to the section "Analyses of Virtual Lab interactions" to further clarify the role of the human researcher. We clarified that "the entirety of the text written by the human researcher is under the 'Nanobody design prompts' subsection of the Methods section." Furthermore, we specified that "all the Python scripts and the one XML script for using ESM, AlphaFold-Multimer, and Rosetta...were written from scratch by the agents. Some complementary data wrangling and job scheduling scripts were written by the human researcher to handle the specific conditions of our compute infrastructure. All scripts were run by the human researcher in accordance with the decisions made by the agents."

2. The timeline, number of team and individual meetings should be plainly spelled out. The brief 5-10 minute meetings, often in parallel, seem rather limited vs the complex task that was assigned.

Thank you for this comment. We have clarified the precise set of meetings in the "Overview" subsection under the "Virtual Lab for nanobody design" section. Specifically, for the tools implementation phase, which involves multiple meetings, we specified: "First, one individual meeting with the PI to decide which scientist agent implements which component. Then, for each component, one individual meeting with the selected scientist agent and the Scientific Critic to write the code for that component followed by one (ESM and AlphaFold-Multimer) or

two (Rosetta) individual meetings with the same scientist agent (no Scientific Critic) to correct errors in the code.” Additionally, we clarified the usages of parallel meetings: “Note that all meetings (team and individual) except for the code correction meetings are run in parallel five times followed by a merge meeting.” In our open source code, we provide the exact parameters controlling the setup of each meeting:

https://github.com/zou-group/virtual-lab/blob/main/nanobody_design/run_nanobody_design.ipynb.

Regarding the brevity of the meetings, we agree that 5-10 minute meetings sound short given the complexity of nanobody design for SARS-CoV-2. However, we would like to emphasize that this is 5-10 minutes of LLM time, not human time, which translates into tens of pages of text written by the LLM agents that could take human scientists multiple hours to generate. Based on the length and content of the agent discussion (e.g., see this example transcript of an LLM discussion

https://github.com/zou-group/virtual-lab/blob/main/nanobody_design/discussions/project_specification/discussion_1.md), we believe that these meetings do capture much of the complexity of this research project. Furthermore, the AI agents conduct five parallel meetings for each decision, which are then summarized by the PI agent, enabling the Virtual Lab to explore a broad range of options.

3. The input of retrieval augmentation generation is alluded to but needs to be more specific.

Thank you for this suggestion. We mention retrieval-augmented generation (RAG) in the Discussion section as a method that could be used to give the agents access to recent knowledge, but we did not use RAG in our initial experiments. For this revision, we have conducted additional experiments quantifying the benefit of retrieval. We show that allowing agents to retrieve and read relevant papers further enhances its knowledge and performance, particularly for fast-developing topics such as new SARS-CoV-2 variants (Supplementary Information section titled “Finetuning agents”).

4. Was this the first project introduced to multi-agents by the authors? Were there any prior attempts? The potential zero-shot nature of success would be noteworthy.

Yes, this was the first project introduced to our multi-agent Virtual Lab framework. We appreciate this question and agree that the zero-shot success is a good demonstration of the abilities of the Virtual Lab.

5. Ironically, without updated A.I. tools such as AlphaFold3 the project came up with designs for nanobodies directed vs the current circulating SARS-CoV-2 variants (KP.3, JN.1). The limitation, as acknowledged, of using somewhat dated models didn't seem to affect the relevant design.

Thank you for this comment. We agree that it is interesting that the Virtual Lab was able to develop nanobody binders for recent variants of SARS-CoV-2 even without using the most up-to-date versions of AI tools. This suggests that tools selected by the agents, such as

AlphaFold-Multimer and Rosetta, are reasonably performant for nanobody design. We expect that as computational biology tools continue to improve, the Virtual Lab agents will also benefit.

6. How was the Scientific Critic agent trained?

Thank you for this question. The Scientific Critic, like the other agents, is instantiated by prompting the underlying LLM (in our case, GPT-4o). The prompt for the Scientific Critic (which also appears in the Methods section) is as follows: “You are a Scientific Critic. Your expertise is in providing critical feedback for scientific research. Your goal is to ensure that proposed research projects and implementations are rigorous, detailed, feasible, and scientifically sound. Your role is to provide critical feedback to identify and correct all errors and demand that scientific answers that are maximally complete and detailed but simple and not overly complex.” Other work has also shown that prompting LLMs to provide critical feedback is sufficient to improve the answers of LLM agents [1].

[1] 1. Yuksekgonul, M. et al. TextGrad: Automatic ‘Differentiation’ via Text. Preprint at <https://doi.org/10.48550/arXiv.2406.07496> (2024).

7. Please define “LLM temperature” as presented in Supplemental Figure 1.

We added a definition of LLM temperature where it is first mentioned in the “Parallel meetings” section: “temperature is the LLM parameter that controls the amount of randomness or uncertainty in the generation.”

Review by Eric Topol, Scripps Research

Referee 2

This work established a so-called “Virtual Lab”, where the lab members are large language model (LLM) agents with different scientific backgrounds. Research in the Virtual Lab is conducted through a series of team meetings between the LLM agents, as well as a limited amount of human input. Each agent represents an expert for a specific scientific field like Machine Learning, Immunology, a PI and a Scientific critic. Agents are transformed to experts via prompting. Subsequently, the agents engage in scientific discussion (guided by the PI agent) to solve an open challenge. To showcase its power, the authors tasked the Virtual Lab with designing novel nanobody binders to contemporaneous variants of SARS-CoV-2. The authors demonstrate that the computational pipeline proposed by the Virtual Lab successfully utilized sophisticated tools in computational protein engineering and machine learning such as ESM, AlphaFold-Multimer, and Rosetta, to design two nanobodies able to bind the recent JN.1 or KP.3 variants of SARS-CoV-2, while maintaining binding to the ancestral SARS-CoV-2 spike protein. In doing so, this work proposes a novel potential framework which leverages LLMs for the design and execution of scientific research. Overall, the work is creative and novel, and it is particularly impressive that the Virtual Lab was able to successfully build and run the nanobody development pipeline with such a little amount of human input, raising the prospect of democratizing sophisticated computational pipelines for laboratories which may not have the expertise internally. However, the authors’ case for the effectiveness/added-value of the Virtual Lab could be made more compelling by comparing their approach with more naïve one(s) to illustrate its effectiveness compared to some reference point.

Thank you very much for your encouraging feedback and thoughtful suggestions. We agree with you that more comparisons would further strengthen the paper and have added several new experiments to the paper.

Comments:

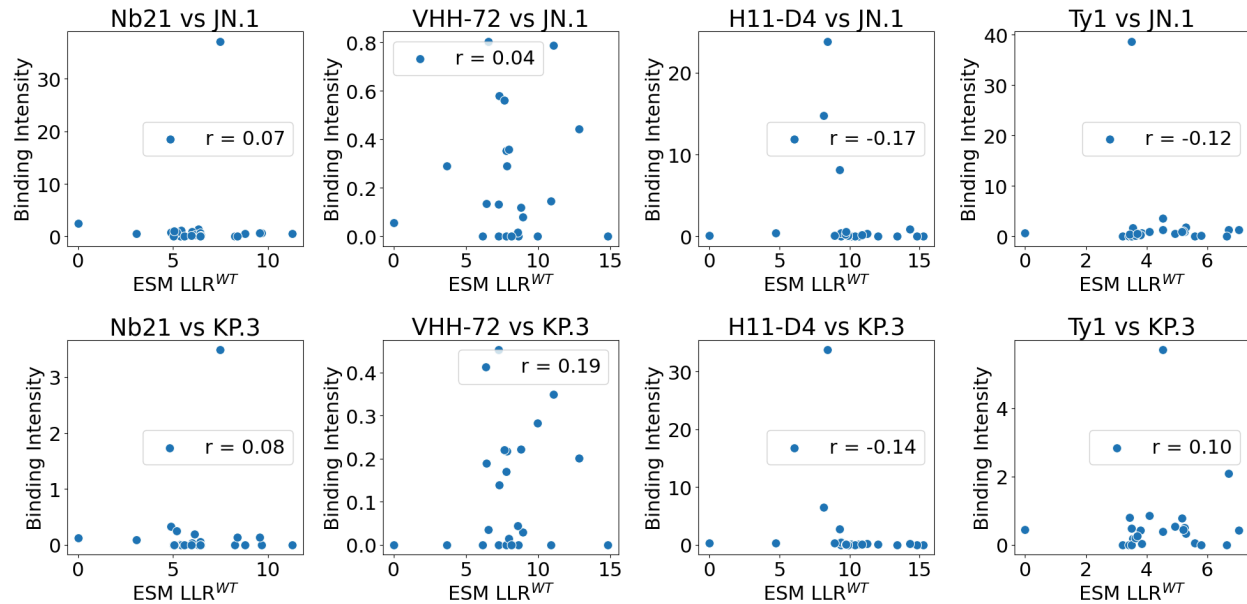
1. Conceptually, the Virtual Lab approach is very compelling. However, for the present use-case (designing SARS-CoV-2 nanobodies) it is unclear how well the Virtual Lab pipeline works compared to existing approaches. The authors should consider comparing the Virtual Lab pipeline with a more conventional approach that uses the same or similar computational tools. One potential comparison could be to use the approach from Hie...Kim. Nature Biotechnology 2023, so use ESM to suggest nanobody mutations over the course of the same number of rounds. If there is a similar (or better) success rate using this previously published approach, it may suggest that the conversations between LLM agents that define the Virtual Lab framework don’t necessarily improve experimental outcomes, but rather serve to automate and synchronize existing computational tools.

We appreciate this valuable suggestion to compare to similar computational approaches for antibody/nanobody design. As discussed in the “ESM usage” section, we note some similarities and differences in the ESM-based methodology developed by the Virtual Lab agents and methods in the literature such as the Hie et al paper. Notably, the Virtual Lab extends beyond this previous work by incorporating two additional tools: AlphaFold-Multimer and Rosetta. ESM as used in both the Hie et al paper and by the Virtual Lab only evaluates proteins like antibodies

or nanobodies in isolation and not in the context of a particular target protein like the SARS-CoV-2 spike protein. Thus, these forms of ESM-based methods only evaluate overall antibody/nanobody quality (e.g., stability) but cannot evaluate their specific binding potential to a given target protein. Underscoring this limitation, across our experiments on 92 generated nanobodies, the ESM scores have very low correlation with ELISA binding intensity to either JN.1 or KP.3 (Response Figure 1 below). To address this problem, the Virtual Lab incorporates AlphaFold-Multimer and Rosetta, both of which model the nanobodies in complex with the SARS-CoV-2 spike protein, thereby providing a computational estimate of binding.

Notably, when analyzing our two mutant nanobody hits, the ESM score alone cannot explain why those mutant nanobodies are hits and the others are not. The successful Ty1 mutant (V32F-G59D-N54S-F32S) had an ESM LLR^{WT} score of only 3.5 compared to the best Ty1 mutant ESM LLR^{WT} score of 7.0, making it 19th out of the 23 generated Ty1 mutants. In contrast, the successful Ty1 mutant's AF ipLDDT score was 8th highest out of 23, indicating that it was more promising according to a target-specific scoring mechanism. The previous approach using ESM alone would have likely missed this successful candidate. Similarly, the successful Nb21 mutant (I77V-L59E-Q87A-R37Q) had an ESM LLR^{WT} score of only 7.5 compared to the best Nb21 mutant ESM LLR^{WT} score of 11.3, making it 7th out of 23 generated Nb21 mutants. However, it was 4th out of 23 in terms of both AF ipLDDT and RS dG scores. ESM score alone would imply that that several other Ty1 and Nb21 mutants with higher ESM scores than the hits should also have been successful, but this was not the case, so the AF ipLDDT and RS dG scores helped introduce target-specific scores that elevated the rank of the successful mutant nanobodies.

That said, we ultimately believe that the primary goal of the Virtual Lab, as you nicely stated, is to automate the usage of existing powerful computational tools for complex scientific problems like nanobody design for SARS-CoV-2 rather than to create the best new tool. This is a problem that would normally require a large team of interdisciplinary scientific experts but now can be tackled to a similar degree of success by a small team of human researchers augmented by the AI agents of the Virtual Lab. To help clarify this point, we added the following to the discussion: "Even if the ultimate scientific decisions of the Virtual Lab agents are similar to those in the scientific literature, the ability of the agents to quickly adapt those methods to the scientific question at hand can empower human researchers to do complex, interdisciplinary science even when they do not have access to an expert panel of human scientists."



Response Figure 1: ELISA binding intensity versus ESM log-likelihood scores for designed nanobodies from all four parental nanobodies (Nb21, VHH-72, H11-D4 and Ty1) and both JN.1 and KP.3 variants. The Pearson correlation values are low, indicating minimal correlation between ESM score and experimental binding intensity.

2. How important is it for the LLM agents to have different backgrounds? My understanding is that all agents are instances of the same underlying GPT-4 model. Are the different specialists (e.g., Immunologist, Machine Learning Specialist) uniquely fine-tuned or prompted in a way that makes them substantially different from each other? It would be valuable to explicitly demonstrate the added value of agent specialization. For example, could an alternative version of the nanobody design pipeline be designed using random specialists or all PIs? This comparison could help clarify whether the “expertise” of the agents leads to meaningful improvements. Such a comparison could perhaps be conducted in silico.

Thank you for the great suggestion! Yes, all agents are instances of the same underlying GPT-4o model. The specialized agents are defined by instruction prompts written by the PI agent (the precise prompts are provided in the “Nanobody design prompts” subsection of the Methods section).

To demonstrate the efficacy of using specialized agents, we ran a new experiment comparing our team of specialized agents to a team of generic agents with prompts like the following.

Assistant: You are an Assistant. Your expertise is in helping people with their problems. Your goal is to help people with their problems. Your role is to help people with their problems.

We ran both teams using the same agenda and agenda questions and the same number of agents but with either specialized or generic agents. The results are in the new “Generic vs specialized agent” subsection of the “Ablation experiments” section of the Supplementary

Information. Notably, the specialized team comes up with a more consistent and feasible approach of modifying existing nanobodies while the generic team arrives at a less feasible approach of simultaneously modifying existing nanobodies and designing de novo nanobodies, which is more challenging.

On a similar note, we added another example of the value of having an agent act as a critic and provide feedback to the scientist agents (see the new “Single versus multi agent” subsection of the “Ablation experiments” section of the Supplementary Information.) In this example, the Computational Biologist writes code to extract the AlphaFold-Multimer interface pLDDT value from a PDB file, but the agent makes a mistake and calculates the pLDDT of the whole complex instead of just the interface. The critic agent notices this mistake and provides feedback that allows the Computational Biologist to correct its error and accurately calculate the interface pLDDT.

Finally, we also investigated creating more powerful agents via finetuning. To that end, we created a new finetuning pipeline that allows each specialized agent to select a set of scientific papers for a given topic (e.g., AlphaFold-Multimer) that it wants to finetune itself on. We then finetuned the agents and ran evaluations to show that the agents gained relevant knowledge through this procedure. More details and results are in the new “Finetuning agents” section of the Supplementary Information.

3. It would be helpful to include some author commentary on whether the base antibodies/nanobodies selected by the models (Ty1, H11-D4, Nb21, and VHH-72) were truly good starting points, or if there might have been better alternatives. Currently, evaluation of these initial choices is done entirely by GPT-4, and while its reasoning seems sensible, LLMs are known to justify and display excessive confidence in their answers even if they are incorrect or suboptimal.

One example of this is in the section on Project Specification, Point 1: choosing nanobodies over conventional antibodies. The AI agent chooses nanobodies because: “Nanobodies offer superior stability, tissue penetration, and ease of production, making them ideal for targeting conserved and cryptic epitopes on the SARS-CoV-2 spike protein. This choice is consistently supported across all summaries, emphasizing their advantages in accessing unique epitopes and potential for broad-spectrum activity.” There are many reasons to disagree with this statement depending on context, for example Nanobodies would not be the better choice for a therapeutic drug, where there are nearly 100 standard antibodies clinically approved and only a single nanobody. Similarly, nanobodies have a much higher risk for immunogenicity.

Overall, greater discussion (by the authors) of the study design choices made by the Virtual Lab would provide valuable context on whether the choices were well-founded.

Thank you for bringing up this important consideration. We agree that nanobodies might not be an appropriate choice in all circumstances. However, in our situation, our goal is more limited

since we are focused on a proof-of-concept where we develop nanobody binders for SARS-CoV-2 as opposed to fully fledged nanobody therapies. Given this context, we believe the design choices of the Virtual Lab are reasonable and are well-justified. In particular, in terms of the choice of starting nanobodies, the four nanobodies selected by the Virtual Lab agents seem reasonable. The four selected nanobodies (Ty1, H11-D4, Nb21, and VHH-72) share certain similarities: all were reported in 2020 in some of the earliest and most highly cited papers reporting SARS-CoV-2 nanobodies (Xiang, 2020 Science; Huo, 2020 Nat SMB; Mcinerney, 2020 Nat Comm; McLellan, 2020 Cell), were shown to potently neutralize pseudovirus, and had available structural characterization of their binding to SARS-CoV-2 Spike/RBD. The nanobodies have a variety of binding modes to Spike/RBD and proposed molecular mechanisms of neutralization, ranging from direct competition for ACE2 binding by Nb21 to broadly neutralizing disruption of RBD dynamics by VHH-72. Importantly, the LLMs avoided suggesting poorly characterized nanobodies as starting points. We updated the Discussion section to mention this: “The Virtual Lab successfully built and ran this nanobody design pipeline, starting with a set of four well-characterized nanobodies (Ty1, H11-D4, Nb21, and VHH-72) with potency and diverse binding modes against early variants of SARS-CoV-2 and developing them into with 92 nanobody candidates for recent variants of SARS-CoV-2 that were experimentally validated by human researchers.”

4. It would be helpful to include additional data on the specificity of the top nanobody candidates. While the current data showing minimal binding to BSA is reassuring, including another control such as a non-SARS-CoV-2 viral glycoprotein would strengthen the conclusion that the nanobodies are indeed specific.

Thank you for this good suggestion! In addition to showing no binding to BSA, we conducted new experiments showing that the top mutated nanobody candidates, and their unmutated sequences, do not bind to MERS-CoV RBD at purified nanobody concentrations up to 50 µg/mL (Figure 5C, reproduced below), reinforcing our conclusions that these nanobody candidates are indeed specific for SARS-CoV-2 RBD. MERS-CoV RBD is a suitable additional control for specificity as it is a viral glycoprotein that shares very low amino acid sequence identity (~21%) with SARS-CoV-2 RBD.

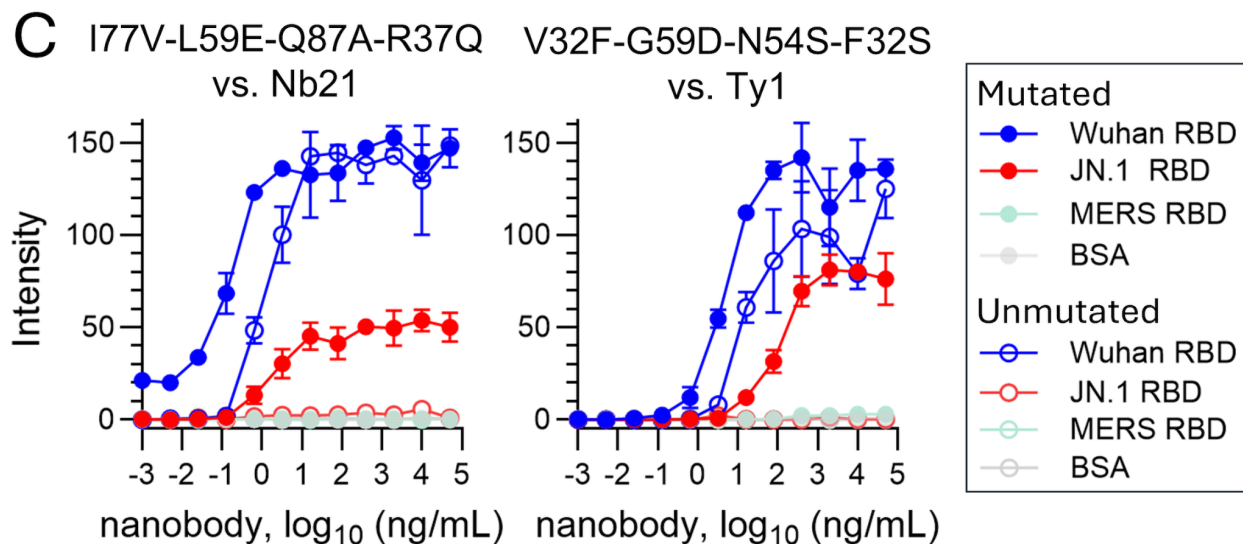


Fig. 5C: Comparison of ELISA binding of mutants and their unmutated sequences. Nb21 mutant I77V-L59E-Q87A-R37Q and Ty1 mutant V32F-G59D-N54S-F32S are shown as filled circles and unmutated Nb21 and Ty1 nanobodies are shown as open circles. ELISA binding intensity (y-axis) to Wuhan RBD, JN.1 RBD, MERS-CoV RBD, and BSA are shown in blue, red, light teal, and gray, respectively. Data are the mean of two measurements at a 12-point serial dilution of purified nanobody.

5. Could the authors clarify how the nanobody expression levels (very low, low, average, high, and very high) are defined?

Thank you. For defining expression levels of the nanobodies, we have replaced the qualitative descriptions (very low, low, average, high and very high) with a titer, expressed as milligrams of soluble nanobody per liter of cell culture (Figure 5A, reproduced below).

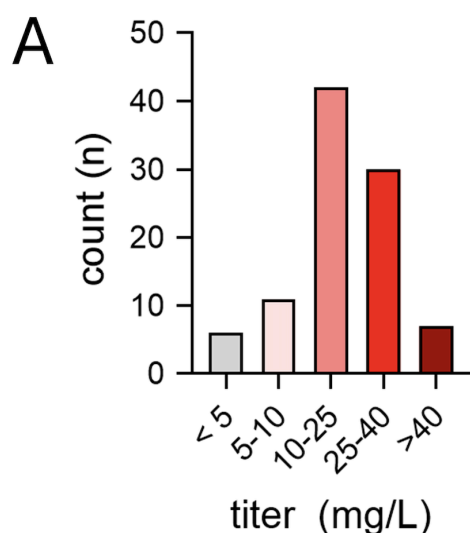


Fig. 5A: Histogram of nanobody expression levels. Titer is expressed as milligrams of soluble, periplasmic nanobody per liter of culture.

6. The Discussion makes a bold claim that, “the Virtual Lab represents a shift from AI as a tool to AI as a partner for science research”. However, it could be argued that a single human conversing with an LLM agent like GPT-4 to design and execute a research plan (which is now commonplace) already represents a form of partnership. The authors should consider softening this language.

Thank you for this suggestion. We have removed this claim and changed those sentences to the following: “Previous work applying AI to science has generally treated AI methods as tools used by human researchers, such as AlphaFold to predict protein structures or LLMs to answer scientific questions, with the human researcher making all the high-level research decisions and design choices. In contrast, in the Virtual Lab, human researchers work alongside LLM agents to design and run a research project.”

7. The authors should consider increasing the text size in the figures, as many of the plots currently have text that is small and difficult to read.

Thank you for this suggestion. We have increased the font size on the figures.

8. The architecture currently relies solely on prompt engineering without mechanisms for weight adjustment or further tuning. The authors could consider incorporating methods that allow the agents to acquire domain specific knowledge. This would address limitations in handling complex problems where established tools are unavailable and mitigate the risk of generating oversimplified or inadequate solutions. While the authors reference RAG in the limitations, elaborating further on how this or similar strategies could be integrated into the system would strengthen the discussion.

Thank you for this suggestion. We agree and we have done extensive work building and evaluating a finetuning pipeline to further improve the domain-specific knowledge of the Virtual Lab agents. In short, the agents are given the ability to search for and select their own set of scientific papers (from PubMed Central) that they would like to finetune themselves on for a given topic (e.g., AlphaFold-Multimer). This technique is especially useful for ingesting new knowledge beyond the knowledge cutoff of the underlying LLM (October 2023 for the GPT-4o model we used), such as new information about the JN.1 and KP.3 variants of SARS-CoV-2. Details on our finetuning method and evaluation are in the new “Finetuning agents” section of the Supplementary Information. Finetuning improves the knowledge of the agents as evaluated by their ability to answer multiple choice questions as illustrated in Supplementary Table 1 (reproduced below). This is especially pronounced for information about the SARS-CoV-2 variants KP.3 and JN.1, which were not prevalent at the time of the GPT-4o and GPT-4o-mini knowledge cutoff in October 2023.

Supplementary Table 1 | Finetuning Q&A evaluation. The accuracy of specialized and finetuned agents powered by GPT-4o-mini or GPT-4o on up to 100 question-and-answer pairs based on summaries of papers from the six finetuning topics.

Topic	Specialized (GPT-4o-mini)	Finetuned (GPT-4o-mini)	Specialized (GPT-4o)	Finetuned (GPT-4o)
Nanobodies	83	89	90	90
SARS-CoV-2 spike protein	86	85	92	92
SARS-CoV-2 variants KP.3 and JN.1	66.67	75.56	76.67	83.33
ESM	89	92	92	93
AlphaFold-Multi mer	86	89	92	93
Rosetta	80	83	85	87

9. LLMs have inherent limitations in recognizing the boundaries of their own knowledge, often leading to unviable or misleading suggestions (see point 3 above for a specific example). This issue could derail the research process or compromise result quality. The authors should explicitly include this limitation in the paper and discuss potential mitigation strategies to enhance the reliability of the approach.

Thank you for bringing this up. We agree that this is an important topic. We added the following paragraph to the Discussion section to address this point.

“The current generation of LLMs is also known to sometimes provide incorrect or misleading answers (often termed hallucinations). In the context of the Virtual Lab, this could mean that the agents might invent incorrect scientific facts or provide fake citations in their discussions. These limitations could be partly mitigated through multi-agent interactions such as having the critic question the veracity of information provided by the other agents, aggregating across multiple parallel meetings, or providing the agents with access to resources (e.g., the text of scientific papers) to verify their knowledge. However, even in those situations, it is still incumbent upon the human researcher working with the Virtual Lab to verify all facts and decisions based on trusted scientific sources to ensure the validity of the decisions of the agents.”

10. The developed work does not currently benchmark its performance against state-of-the-art tools, which makes it difficult to assess the approach's competitiveness or effectiveness. Including benchmarks in future studies would provide a clearer understanding of how the proposed system performs relative to existing methodologies and strengthen its validation.

Thank you for noting this. In terms of comparing the Virtual Lab's nanobody design pipeline (ESM + AlphaFold-Multimer + Rosetta) to other approaches, as with your first question, we would like to emphasize that the goal of the Virtual Lab is to rapidly apply existing computational tools to the scientific problem at hand, not necessarily to build a better set of tools. In our response to comment #1 above, we also discuss possible benefits of the pipeline created by the Virtual Lab compared to the previous approach which relied on using just ESM. In terms of analyzing the effectiveness of the Virtual Lab framework itself, we performed a series of new experiments evaluating the components of the Virtual Lab to demonstrate superiority to simpler LLM methods. In the new "Ablation experiments" section of the Supplementary Information, we compare our specialized agents to a set of generic agents and our multi agent conversations to single agent conversations, illustrating benefits of our design. This complements our existing discussion of the effect of multi agent interactions and parallel meetings in the section "Analyses of Virtual Lab interactions." Our new finetuning method and results (see "Finetuning agents" section of the Supplementary Information) also quantify improvements of further enhanced science agents compared to a standard LLM.

11. The authors do not provide sufficient information on the number or nature of prompt adjustments made to achieve the presented results. Examples of these adjustments, including both successful and unsuccessful prompts, would improve reproducibility. This would also help the community gain deeper insights into the robustness of the system and enable the development of more effective prompt strategies. Also, the authors could explore integrating a chain-of-thought reasoning approach in future work to improve the explainability of their outputs and address some of the inherent LLMs limitations.

Thank you for suggesting this. We agree that more information on prompt engineering would be valuable, so we included a new set of prompt engineering experiments in the new "Ablation experiments" section of the Supplementary Information. There, we focused on the tools selection team meeting and showed four variants of the original agenda and agenda questions and discussed the elements of the prompt that the Virtual Lab agents are sensitive to and those that they are robust to. We hope that this will help the community better understand prompt engineering in the context of multi-agent systems like the Virtual Lab. To promote transparency and reproducibility, we have made the Virtual Lab fully open source and have shared all agent prompts and meeting logs on our GitHub site.

We also agree that chain-of-thought is a valuable approach and would be worth including. To a large degree, the agents engage in reasoning similar to chain-of-thought in that they take turns addressing the agenda of the meeting from different perspectives one-by-one. However, each agent itself could also potentially benefit from chain-of-thought reasoning or internal reasoning such as with an OpenAI o1 or o3 model. We plan to explore these ideas in future work to further empower the Virtual Lab agents.

12. The study does not include ablation experiments, such as comparisons with and without prompting or with and without scientific critic. Including these experiments would help isolate the contributions of individual components and provide clarity on their impact.

Thank you for this suggestion. We have added a new “Ablation experiments” section to the Supplementary Information with new experiments comparing our specialized agents to generic agents, comparing single agent responses to agent + critic interactions, and showing the effect of prompt engineering. We hope that these experiments and the discussion of their results help elucidate the effect of the different components of the Virtual Lab.

Referee #3 (Remarks to the Author):

SUMMARY

This manuscript describes the development of a novel AI-driven research tool called the Virtual Lab, which uses a team of large language models (LLMs) with different scientific “backgrounds” to conduct interdisciplinary research. This innovative approach is very much in line with cutting edge AI research generally (2025 has been described as the Year of Agents), so this agent-based approach to scientific research is both creative and timely. The Virtual Lab was tasked with creating novel antibodies or nanobodies against recent variants of SARS-CoV-2. Their approach was to engineer existing nanobodies with affinity for ancestral variants in order to confer variant binding. The Virtual Lab designed a computational workflow based on ESM, AlphaFold-Multimer and Rosetta to select productive mutations in four existing nanobodies. Recombinant expression and functional testing showed that most engineered nanobodies expressed and were soluble and uncovered two nanobodies able to bind the JN.1 variant of SARS-CoV-2.

The scientific approach devised by the Virtual Lab (in silico directed evolution of existing nanobodies using a pipeline of existing modeling tools) is not particularly novel; indeed, a very similar approach was used in Hie et al in 2023. I don’t necessarily view this negatively — on the contrary, I find the fact that a Virtual Lab devises an experimental approach that mirrors those produced by fully human research groups to be a significant achievement. The use of a slightly modified method for computing log likelihood (compared to Hie et al, which used a similar workflow for antibody engineering, suggests at least some element of novelty to the Virtual Lab’s approach.

Overall, this manuscript is extremely innovative, both technically and conceptually, and is certain to be of very high impact. While there are several issues that should be addressed, I am strongly in favor of publishing a revised manuscript in Nature.

Thank you very much for your encouraging feedback and for your thoughtful suggestions and ideas! We have revised the paper and performed additional experiments based on your feedback.

MAJOR CONCERNS

- In the introduction, the authors mention limitations of other AI-focused research tools, and a recurring theme is that none are able to generalize effectively across diverse scientific disciplines. The Virtual Lab includes “people” with different expertise, but it appears that each of these experts requires substantial domain knowledge to instantiate, at least for the first time. For example, it appears that the code for each of the three primary computational tools (ESM, Rosetta, and AlphaFold) required a non-trivial amount of human intervention, as the initial LLM-produced code had multiple errors. Would providing the computationally-focused agents with a sandboxed development environment allow them to do such error correction on their own by iteratively running the generated code, identifying errors, and fixing them?

Thank you for the suggestion. We agree that this is a fantastic idea and we have been thinking about similar ideas since it would be a great way to allow the agents to do their own error correction. However, given the significant engineering and research work that would be involved in building such a sandboxed environment and developing agents that can make good use of it, we plan to explore this for future work. We added this idea as future work in the Discussion section: “Additionally, for issues related to coding in particular, future work could explore developing sandboxed environments to allow the agents to independently install computational or AI tools and then write, debug, and run code that use those tools for a particular application.”

- While all of the prompts are provided, it would also be helpful to see at least a summarization of the resulting meetings. For example, it would be interesting to see how the agents arrive at their weighting scheme for ESM LLR, Rosetta dG, and AF ipLDDT.

We agree that it’s helpful to see the Virtual Lab discussions to better understand how the agents came to their answers. To that end, we have made the full transcript of all Virtual Lab meetings available publicly on our GitHub

(https://github.com/zou-group/virtual-lab/tree/main/nanobody_design/discussions). For the specific discussion on the weighting scheme, here is one of the five parallel individual meetings with the PI to discuss the scheme:

https://github.com/zou-group/virtual-lab/blob/main/nanobody_design/discussions/workflow_design/discussion_1.md. This is one of the shorter discussions since it only involves the PI. For an

example of a longer discussion, please see a team meeting from the Project specification phase, where high-level decisions like antibodies vs nanobodies are made:

https://github.com/zou-group/virtual-lab/blob/main/nanobody_design/discussions/project_specification/discussion_1.md. The PI-written summary is at the end of that discussion.

- The human-in-the-loop organizational structure of the Virtual Lab seems designed to address perceived limitations in truly novel thinking and out-of-distribution creativity by the current generation of LLMs (although new thinking models may show improvements in this area). It would be interesting to include a bit of discussion from the authors about how they envision this requirement changing in the future — do they view human involvement as a temporary workaround that could potentially be removed entirely at some point in the future (with the virtual

PI autonomously identifying high impact scientific questions and formulating an initial hypothesis), or whether human feedback will likely always be required albeit with more open-ended instructions.

This is a great question. It's challenging to predict the ultimate autonomous capabilities of AI agents in the future, but likely some amount of human supervision will still be needed to guide the AI agents toward solving scientific problems that are most valued by human society, which may change over time. The idea of high impact science is ultimately somewhat subjective depending on what humans believe is most interesting about the scientific world or what discoveries would most benefit humans, so humans will be important to make those kinds of decisions even if AI agents are capable of doing much of the scientific work to determine the answer to the scientific questions prioritized by humans. Moreover, in the medium term, we would still need human researchers to carry out most wet lab validations and thus human input on experimental constraints and considerations will likely be important. We added some thoughts on this question in the Discussion section: "However, even as the Virtual Lab expands its capabilities, human scientists will still be vital to guide the AI agents in their choice of scientific questions, methodologies, and analyses to match the scientific values and interests of the human researchers."

MINOR ISSUES

- The figure panel labels in Fig 1 (and data labels generally in many other figures) are very small — increasing the size would greatly aid legibility. On the other hand, I really like the avatars for the Virtual Lab members, which makes it clear and easy to understand the structure and activities of the Virtual Lab.

Thank you for the good suggestion! We have increased the font size on the figures.

- When assessing engineered nanobody binding, the initial ELISA screen was described as performed using "a single nanobody dilution (0.5X)", but it is not clear what the 0.5X is relative to — a 1:2 dilution of some unknown nanobody concentration, half the antigen concentration, or some other value? An absolute concentration, or additional information about what the relative concentration is referring to, should be provided.

Thank you for this suggestion. We have clarified in the section "Nanobody experimental validation" that the initial ELISA screen of all 96 nanobodies was performed using a 1:2 dilution of nanobody lysate. The top nanobody candidates were subsequently evaluated as purified nanobodies at defined concentration (ug/mL) (see below).

- More detail about what the various expression levels (very low, low, average, high, and very high) in Fig 5 correspond to in terms of actual concentration —relative to culture volume — should be provided.

Thank you. For nanobody expression levels, we have replaced the qualitative descriptions (very low, low, average, high, and very high) with a titer, expressed as milligrams of soluble nanobody per liter of cell culture. (See “Nanobody experimental validation” section and Figure 5A, reproduced below.)

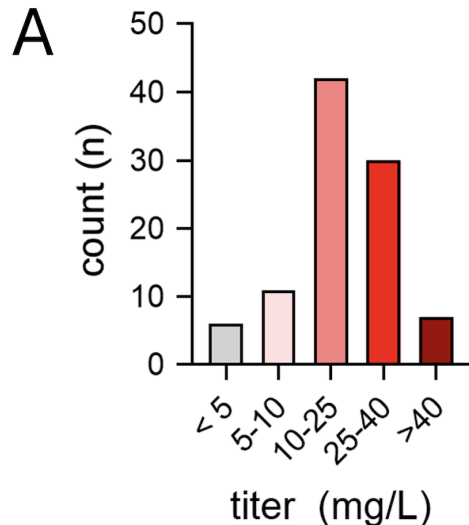


Fig. 5A: Histogram of nanobody expression levels. Titer is expressed as milligrams of soluble, periplasmic nanobody per liter of culture.

- The inclusion of an unpaired Cys in several of the engineered nanobodies would seem to be something that the ESM log-likelihood should have flagged as highly disfavored. Is there some other reason why the weighted score would rank these variants so highly, like the ipLDDT or dG values being very favorable? Are there any obvious adjustments to the WS formula that could mitigate this?

Thank you for bringing up this question! We thought it would be interesting to ask the Virtual Lab agents for their opinion, so we ran five parallel team meetings with the following agenda and agenda questions. Below, we show excerpts from the final summary of the meetings written by the Principal Investigator agent. We agree with the Principal Investigator’s analysis based on the fact that the mutated nanobodies with added Cys residues tended to have lower ESM LLR scores but medium to high AF ipLDDT and RS dG scores, so adjusting the weights of AF ipLDDT and RS dG relative to ESM LLR or adding a specific unpaired Cys penalty could address this issue.

Agenda (human researcher)

Your team has designed 92 mutated nanobodies (23 each for the wild-type nanobodies H11-D4, Nb21, Ty1, and VHH-72) to improve their binding to the KP.3 variant of the SARS-CoV-2 spike protein receptor binding domain (RBD). Each nanobody has 1-4 mutations relative to the wild-type nanobody. Your team used ESM log-likelihood ratios (ESM LLR) to score the nanobody mutations independent of the antigen, AlphaFold-Multimer to predict the structure of

the mutated nanobody in complex with the KP.3 RBD and compute the interface pLDDT (AF ipLDDT) as a metric of binding confidence, and Rosetta to calculate the binding energy of the mutated nanobody in complex with the KP.3 RBD (RS dG) based on the AlphaFold-Multimer predicted structure followed by a Rosetta relaxation. You have ranked the mutant nanobodies and selected the top ones using a weighted score of $WS = 0.2 * (ESM \text{ LLR}) + 0.5 * (AF \text{ ipLDDT}) - 0.3 * (RS \text{ dG})$.

An astute reviewer gave the following feedback: "The inclusion of an unpaired Cys in several of the engineered nanobodies would seem to be something that the ESM log-likelihood should have flagged as highly disfavored. Is there some other reason why the weighted score would rank these variants so highly, like the ipLDDT or dG values being very favorable? Are there any obvious adjustments to the WS formula that could mitigate this?"

Upon examining the 92 mutated nanobodies, your team found the following:

Nb21 and VHH-72 have no Cys changes (no additions or removals).

Among the 23 Ty21 mutants, two have an added Cys. The Ty1-V32C-G59H-H59Y mutant has an ESM LLR rank of 11 / 23, an AF ipLDDT rank of 22 / 23, and an RS DG rank of 1 / 23 while Ty1-V32C-G59D-D59Y mutant has an ESM LLR rank of 12 / 23, an AF ipLDDT rank of 23 / 23, and an RS DG rank of 2 / 23.

Among the H11-D4 mutants, three have an added Cys. The H11-D4-A14P-Y88V-R27C mutant has an ESM LLR rank of 19 / 23, an AF ipLDDT rank of 3 / 23, and an RS DG rank of 19 / 23 while the H11-D4-A14P-Y88V-R27C-W53Y mutant has an ESM LLR rank of 22 / 23, an AF ipLDDT rank of 8 / 23, and an RS DG rank of 6 / 23 and the H11-D4-A14P-Y88V-R27C-R52S mutant has an ESM LLR rank of 21 / 23, an AF ipLDDT rank of 18 / 23, and an RS DG rank of 12 / 23.

Based on the above information, how would you address the reviewer's questions?

Agenda Questions (human researcher)

1. Is there some other reason why the weighted score would rank these variants so highly, like the ipLDDT or dG values being very favorable?
2. Are there any obvious adjustments to the WS formula that could mitigate this?

Answer Summary (Principal Investigator, Excerpts)

1. Is there some other reason why the weighted score would rank these variants so highly, like the ipLDDT or dG values being very favorable?

Yes, the high ipLDDT and RS dG values are likely contributing significantly to the high rankings of these variants, despite the presence of unpaired cysteines. The ipLDDT and RS dG metrics

are heavily weighted in the current WS formula, reflecting strong structural confidence and favorable binding energies. These metrics can overshadow sequence-based concerns, such as the presence of unpaired cysteines, leading to high rankings for variants that might otherwise be considered structurally unstable. The current weighting prioritizes structural predictions and binding affinity, which are critical for initial assessments of nanobody efficacy.

2. Are there any obvious adjustments to the WS formula that could mitigate this?

Yes, introducing a penalty for unpaired cysteines in the WS formula could mitigate this issue. Incorporating a penalty for unpaired cysteines will help balance the scoring system by accounting for potential structural instability and immunogenicity concerns. This adjustment ensures that the scoring reflects both binding affinity and structural integrity, providing a more accurate assessment of nanobody viability.

Implementation Details:

- Penalty Quantification: Implement a penalty of 5-10% reduction in the WS score per unpaired cysteine. This range is based on preliminary retrospective analysis of similar structural disruptions in protein engineering literature, which suggests that unpaired cysteines can significantly impact protein stability.

- Metric Weighting Adjustments: Consider adjusting the weightings of ipLDDT and RS dG in the WS formula to ensure that the introduction of the penalty does not disproportionately affect the overall scoring. This adjustment will be based on the outcomes of the penalty implementation and empirical validations.

- In Fig 5C, either actual concentration values (rather than dilution) should be provided, or the starting concentration prior to dilution should be provided.

Thank you for this suggestion. For these top nanobody candidates shown in Figure 5C (reproduced below), we have reperformed binding titrations at absolute concentrations of purified nanobodies (in ug/mL). The Methods section has been expanded to include this information.

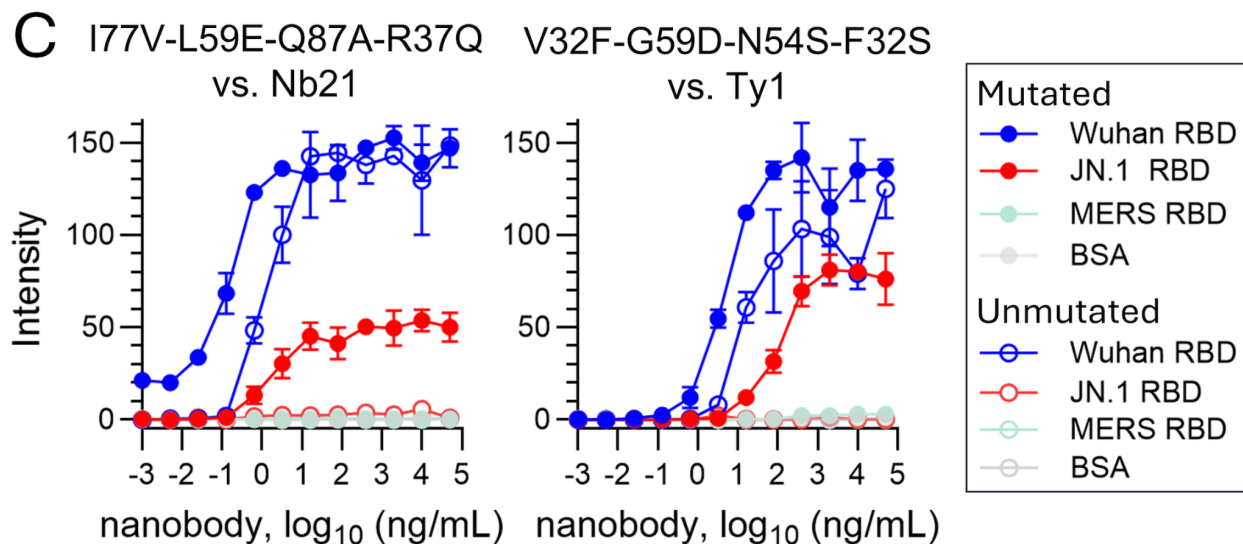


Fig. 5C: Comparison of ELISA binding of mutants and their unmutated sequences. Nb21 mutant I77V-L59E-Q87A-R37Q and Ty1 mutant V32F-G59D-N54S-F32S are shown as filled circles and unmutated Nb21 and Ty1 nanobodies are shown as open circles. ELISA binding intensity (y-axis) to Wuhan RBD, JN.1 RBD, MERS-CoV RBD, and BSA are shown in blue, red, light teal, and gray, respectively. Data are the mean of two measurements at a 12-point serial dilution of purified nanobody.

- The potential use of different LLMs for each persona is discussed, but I did not notice any detail about the specific model used for each Virtual Lab agent. That would be helpful information — was the same model used for all agents, or were specialized models used depending on the agent's role?

Thank you for the question. As noted in the “Overview” subsection of the section “Virtual Lab for nanobody design”, we used GPT-4o: “we applied the Virtual Lab with GPT-4o powering the agents.” We have now further clarified that we specifically used gpt-4o-2024-08-06. This same model was used for all agents. Specialization of the agents was achieved through prompting. As discussed in the new Supplementary Information, the agents could be further specialized and improved through methods such as finetuning on domain-specific knowledge.

Referee #4 (Remarks to the Author):

The manuscript by James Zou et al presents Virtual Lab, a multi-agent framework in which a human researcher delegates a scientific project to a team of role-specialised GPT-4o-powered agents (Principal Investigator, three domain scientists, and a Scientific Critic). The human provides high-level agendas; the agents then hold “team” and “individual” meetings (each executed five times in parallel and later merged) to debate strategy, write code, and critique one another. Applied to nanobody engineering against the KP.3/JN.1 SARS-CoV-2 variant, the human and agents generated a comprehensive analysis pipeline (ESM → AlphaFold-Multimer → Rosetta pipeline), and produced 92 mutant nanobodies. Experimental testing confirmed two binders with retained or gained activity against contemporary RBDs. This specific task produced 122,462 words of discussion and code, of which only 1,596 words (1.3 %) were produced by the human user. The work exemplifies a fascinating collaborative model in which the human retains strategic oversight but offloads detailed implementation to an autonomous AI team.

As Strengths

- The study is one of the first to demonstrate that a structured LLM-powered agent hierarchy can plan, code, and execute a non-trivial protein-engineering pipeline under human oversight but with overall limited human intervention.
- The use of virtual "meetings" represents an interesting paradigm for human-AI collaboration. Agendas authored by the researcher steer the agents, the agents contribute new ideas, eg, weighing scores to rank mutations, while the Critic agent raises methodological concerns and spots coding errors; this mirrors real laboratory supervision and likely mitigates errors and hallucinations. The added benefit is that meeting transcripts expose interdisciplinary reasoning, enabling audit of agent decisions and revealing actual contribution profiles (e.g., PI 36 %, Critic 21 %)
- The Parallel-meeting merge strategy is also interesting. Replicating meetings five times reduces LLM stochasticity and encourages consensus answers, an interesting variant of majority voting
- Two experimentally validated nanobodies with activity against emergent variants attest that the pipeline can yield biologically useful molecules

Thank you very much for your insightful feedback! We truly appreciate your helpful suggestions, which have enabled us to further strengthen this paper.

Weaknesses and opportunities for improvement:

- Claims of generality and domain-agnosticism remain speculative in my view; all data arise from one nanobody-design case, expanding on a paper from another Stanford group. Extending the framework to at least one unrelated discipline (e.g. materials discovery) or subdiscipline (e.g.

cell therapy) would substantiate generality. I recognize that this is extra work, and that instead the authors may want to tone down their domain-agnostic claims.

Thank you for this suggestion. Since completing another full research project with the Virtual Lab followed by experimental validation would be outside the scope of this paper, we have modified the text to tone down the domain-agnostic claims.

In the abstract, we made the following changes.

“This demonstrates the ability of the Virtual Lab to rapidly make **an** impactful, real-world scientific discovery.”

In the introduction, we made the following changes.

“The Virtual Lab architecture is versatile and can **potentially** be applied to a variety of interdisciplinary science research projects.”

“This **outcome illustrates** the capability of the Virtual Lab’s AI-human collaboration to **execute a** complex, interdisciplinary science research **project** that translates to **a** validated result in the real world.”

In the discussion, we made the following changes.

“The Virtual Lab achieved its goal of engaging in **a** sophisticated, interdisciplinary science research **project**, as demonstrated by its design of nanobodies with experimentally validated, diverse binding profiles across multiple strains of SARS-CoV-2.”

“**This outcome serves as an example of how** human researchers can partner with LLM agents **in the Virtual Lab** to rapidly achieve **a** promising scientific result that can streamline further experiments. Even if the ultimate scientific decisions of the Virtual Lab agents are similar to those in the scientific literature, the ability of the agents to quickly adapt those methods to the scientific question at hand **shows how LLM agents can potentially** empower human researchers to do complex, interdisciplinary science even when they do not have access to an expert panel of human scientists.”

“**While the experimental results here are limited to the domain of nanobody design, with future work,** we envision the Virtual Lab as a powerful framework for human researchers to engage in interdisciplinary science research with the help of LLMs.”

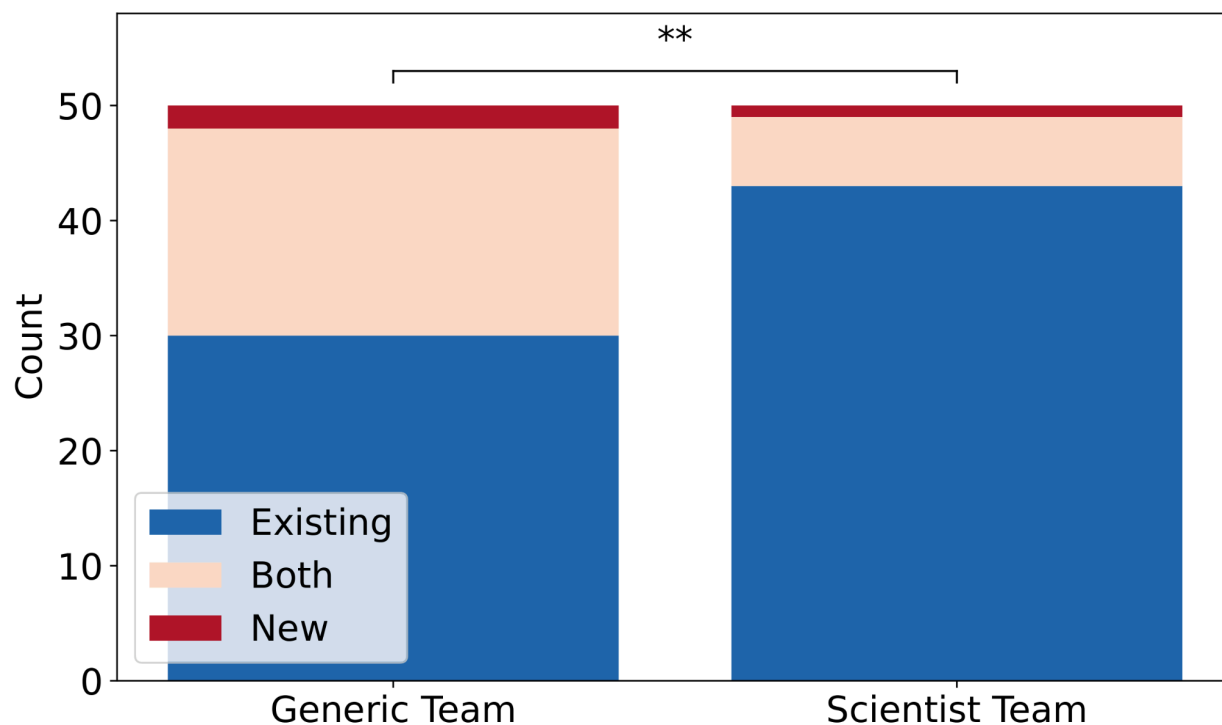
- The study’s central claims, superior coordination, reduced human labour and domain generality, are supported almost exclusively by qualitative narratives. Although a “generic-agent” ablation is provided, its impact is discussed anecdotally; same for distinct prompts. There are no numerical comparisons of hit-rate, runtime, token cost, API expenditure, or actual human labor across configurations. Rigorous, task-level and global performance metrics are essential to

assess whether the proposed architecture materially outperforms (i) a single well-prompted GPT-4o agent, (ii) a generic multi-agent team without roles, (iii) an expert human baseline or (iv) any other variants. It is for example difficult to assess whether removing the critic agent, changing the number of parallel meetings, etc actually decrease the chances of discovering strong binders. Likewise, I recognize that this is a lot of extra work, but it would turn the work into a rigorous scientific study as opposed to a very cool story with potential but unproven broader impact.

Thank you for suggesting a more quantitative evaluation of the impact of the Virtual Lab framework. We agree that this would improve the rigor of our paper. To address runtime, token cost, and human labor, we included these details in the “Analyses of Virtual Lab interactions” section (please see our response to your comment on word counts below). To address the remaining questions about the effect of the Virtual Lab framework of scientist agents, we performed three new sets of quantitative experiments and analyses.

First, we scaled up our current analysis of the effect of generic agents versus scientist agents in the project specification meeting. Previously, we only performed a qualitative analysis on five discussions from a team of generic agents and five discussions from a team of scientist agents. We have now scaled this up to 50 discussions from each team, and we analyzed the decisions of each team type. Please see the section below for details, which has been added to the Supplementary Information.

To better quantify the difference between generic agents and scientist agents, we ran the project specification team meeting 50 times each with the generic agent team and the scientist agent team, and we counted the teams’ decisions on protein type (antibody or nanobody) and design approach (de novo design or modifying existing antibodies/nanobodies or a hybrid approach). In all cases except for one, the generic agents and scientist agents decided to pursue nanobody design. Interestingly, echoing the qualitative results above, the generic agents suggest the less efficient hybrid or de novo design approaches more often than the scientist agents ($p < 0.01$), which tend to prefer the more efficient method of modifying existing antibodies/nanobodies (Supplementary Fig. 1). The human protein engineering experts (J.E.P. and W.W) also preferred the modification approach suggested by the scientist agents.



Supplementary Fig. 1 | Generic agents and scientist agents select the design approach.

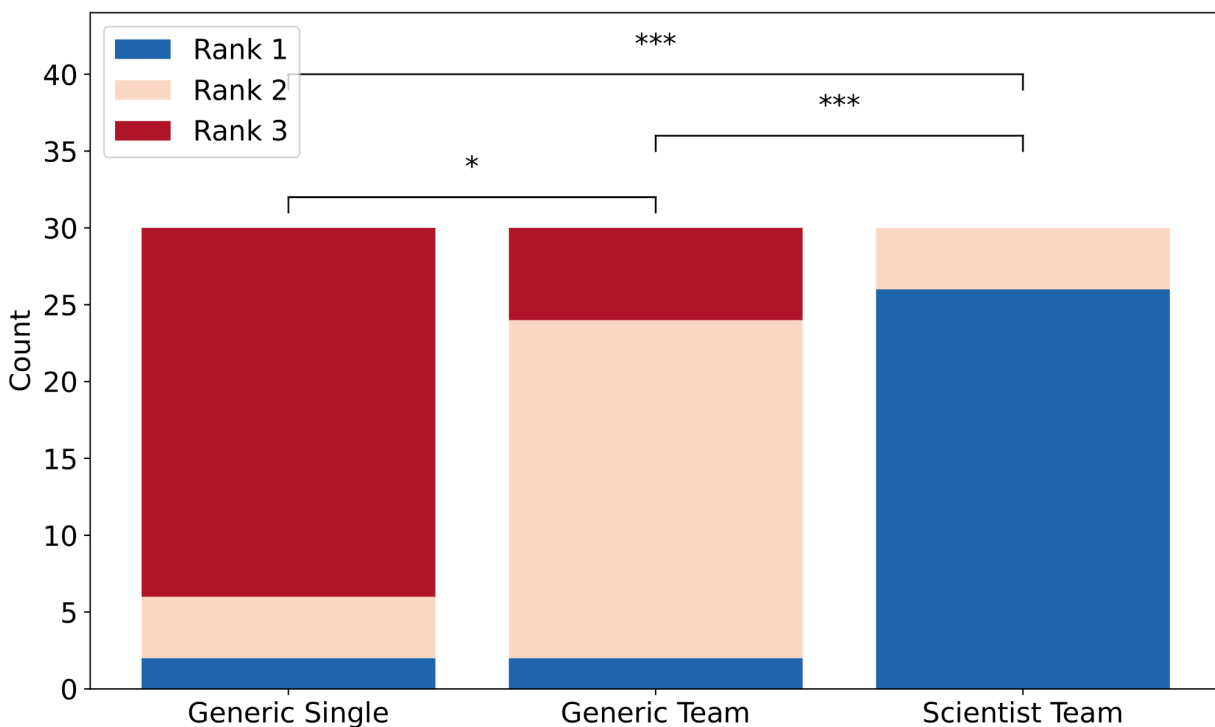
The number of times a team of generic agents and a team of scientist agents chose de novo design, modifying existing antibodies/nanobodies, or a hybrid approach at the end of 50 identical project specification team meetings. $p = 0.006$ (** indicates $p < 0.01$) using Fisher's exact test to compare modifying existing antibodies/nanobodies to de novo design or a hybrid approach (combined).

Second, we designed and conducted a new human evaluation to compare the effect of a single generic agent, a team of generic agents, and a team of scientist agents. Please see the section below for details, which has been added to the Supplementary Information.

As another evaluation of the Virtual Lab framework's effect, we conducted a blinded human expert evaluation comparing discussions involving a single generic agent, a generic agent team, and a scientist agent team (i.e., the Virtual Lab framework). With the help of GPT-4o, we crafted 30 open-ended research questions, 10 related to protein engineering and 20 related to computational biology. For each question, we ran each of the three frameworks (single generic agent, generic agent team, and scientist agent team) on the same question with a single meeting and with no prompt engineering. For the scientist agent team, we first gave the Principal Investigator the research question and asked it to define a set of scientist agents for the discussion. Then, we used those scientist agents in a team meeting to discuss the question. For the generic agent team, we created a team of generic agents (named Assistant 1 to Assistant N) with the same number of agents as the scientist team and ran a team meeting with them. For the single generic agent, we created one generic agent and ran an individual meeting with one round.

After running these meetings, we blinded the discussions by replacing the names of the scientist agents (e.g., Computational Biologist) with the names of the generic agents (e.g., Assistant 1) so that the names of the agents cannot be used to distinguish between the generic agent team and the scientist agent team, which have the same number of agents. We also randomized the order of the responses to avoid bias. Next, we asked one protein engineering expert to review the blinded discussions of the 10 protein engineering questions and another computational biology expert to review the blinded discussions of the 20 computational biology questions. For each question, they read all three discussions and ranked them from 1 (best) to 3 (worst) in terms of scientific accuracy, creativity, and comprehensiveness.

Supplementary Fig. 2 shows the result of the human evaluation. The scientist agent team (i.e., Virtual Lab framework) significantly outperformed both the generic agent team and the single generic agent in terms of human preference ($p < 0.001$). The discussions of the scientist agent team tended to be more comprehensive and nuanced than those with generic agents. Furthermore, the generic agent team significantly outperformed the single generic agent, adding evidence that multi-agent systems can provide higher quality scientific insights and decisions compared to single agent systems. This is consistent with similar findings on the benefits of multi-agent systems in other domains^{62,63}.



Supplementary Fig. 2 | Human evaluation of generic agents versus scientist agents. The results of a blinded human evaluation of 30 open-ended research questions (10 protein engineering, 20 computational biology), in which human reviewers ranked the discussions of a single generic agent, a generic agent team, and a scientist agent team. All significance tests are two-sided Wilcoxon signed-rank tests with a Bonferroni correction. * indicates $p < 0.05$, **

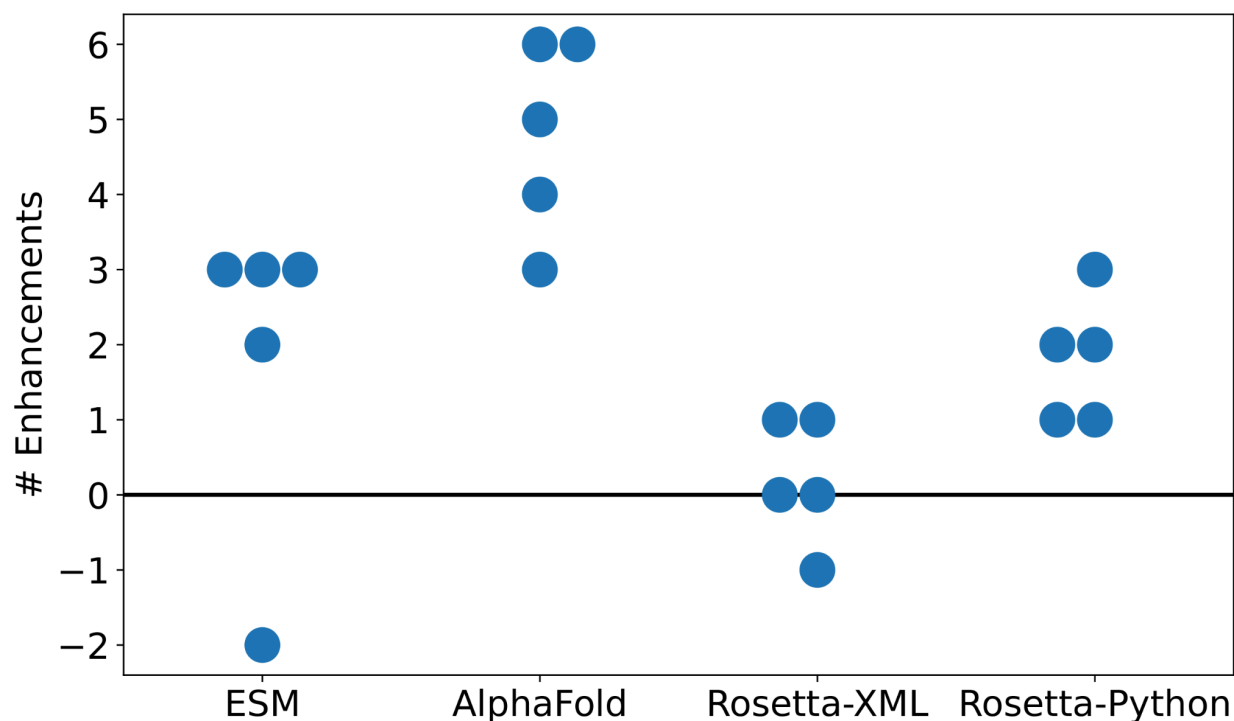
indicates $p < 0.01$, *** indicates $p < 0.001$. Generic Single vs Generic Team: $p = 0.0196$, Generic Single vs Scientist Team: $p = 3.88 \times 10^{-6}$, Generic Team vs Scientist Team: $p = 1.28 \times 10^{-5}$.

62. Du, Y., Li, S., Torralba, A., Tenenbaum, J. B. & Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. in *Proceedings of the 41st International Conference on Machine Learning* vol. 235 11733–11763 (JMLR.org, Vienna, Austria, 2024).
63. Wang, J., Wang, J., Athiwaratkun, B., Zhang, C. & Zou, J. Mixture-of-Agents Enhances Large Language Model Capabilities. in *The Thirteenth International Conference on Learning Representations* (2024).

Third, we performed a human code review of all code written by the Virtual Lab agents during the nanobody design workflow to assess the effect of a multi-agent system compared to single agents. Please see the section below for details, which has been added to the Supplementary Information.

To quantify the effect of the agent-critic interactions across the different phases of the Virtual Lab nanobody design process, we looked at every script written by the Virtual Lab throughout these phases and compared the original script written by a single scientist agent to the final script written after the interaction between the scientist agent and the critic agent. We performed a manual code review and counted the number of enhancements introduced via this multi-agent interaction, where an enhancement is a major positive change to the code (e.g., correcting a bug or improving error handling as opposed to simply renaming a variable or adding a comment). Major negative changes are counted as negative enhancements. We reviewed the Python scripts written for ESM, AlphaFold-Multimer, and Rosetta as well as the XML scripts written for Rosetta. For each script type (e.g., ESM), we reviewed five individual pairs of the original single agent script and the final multi agent script, one from each of the five parallel meetings.

As seen in Supplementary Fig. 3, the multi-agent interaction produced major enhancements in 16 of the 20 scripts, with no changes or just 1-2 negative changes in the other four scripts. Notably, the most complex script (AlphaFold, which involves PDB file parsing, 3D geometry calculations, and confidence score extraction) benefits the most from the multi-agent interaction, while the simplest script (Rosetta-XML, which is just a few lines of Rosetta script tags) has the fewest enhancements, likely because it has the least room to improve. Across all 20 scripts, the number of positive enhancements is statistically significant ($p = 0.0005$ using a one-sided Wilcoxon signed-rank test), demonstrating that multi-agent interactions have a strong tendency to improve code quality.



Supplementary Fig. 3 | Code enhancements via multi-agent interactions. The number of significant enhancements (positive if beneficial, negative if detrimental) introduced in the final code produced by a multi-agent interaction with a scientist agent and a critic agent compared to the original code written by the single scientist agent. For each script type (e.g., ESM), five pairs of multi agent and single agent scripts are compared, one from each of the five parallel meetings.

- Two binders out of 92 designs (~2 %) are reported in isolation; benchmarking against existing computational-antibody pipelines (e.g., the ESM-guided approach of Hie et al., 2024) would clarify relative efficiency and success probability. At first glance, Hie et al report hit rates 14–71 % in round 1 and improvements in all seven antibodies.

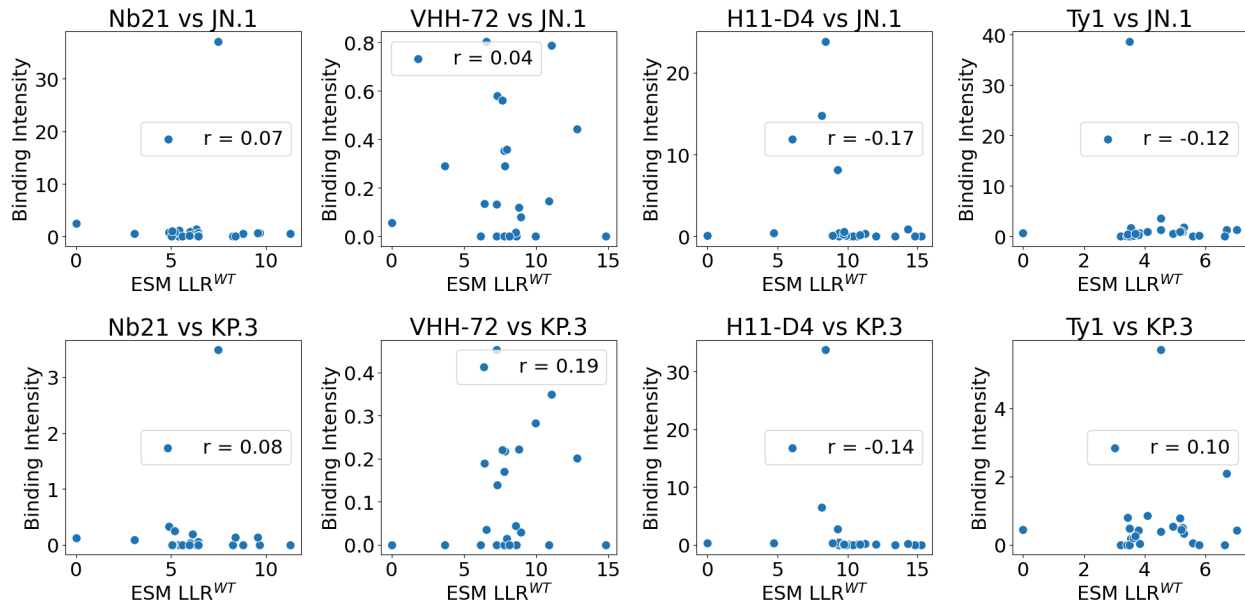
Thank you for suggesting this comparison. It is difficult to directly compare with the hit rate of Hie et al given the differences in targets. We expect lower hit rates than those reported in Hie et al, due to the increased difficulty of our task - designing binders to novel SARS-CoV-2 variants – compared to their focus on improving affinity for a well-studied antigen. Instead, the focus of our work is to demonstrate the ability of AI agents in the Virtual Lab to perform open-ended research and design reasonable computational pipelines achieving the desirable outcome with a relatively high level of autonomy, rather than necessarily outperforming the best pipeline handcrafted by human experts for a specific objective.

In the “ESM usage” section, we discuss similarities and differences in the computational methodology developed by the Virtual Lab agents and methods in the literature such as the Hie et al paper. Notably, the Virtual Lab extends beyond this previous work by incorporating two additional tools: AlphaFold-Multimer and Rosetta. ESM as used in both the Hie et al paper and

by the Virtual Lab only evaluates proteins like antibodies or nanobodies in isolation and not in the context of a particular target protein like the SARS-CoV-2 spike protein. Thus, these forms of ESM-based methods only evaluate overall antibody/nanobody quality (e.g., stability) but cannot evaluate their specific binding potential to a given target protein. Underscoring this limitation, across our experiments on 92 generated nanobodies, the ESM scores have very low correlation with ELISA binding intensity to either JN.1 or KP.3 (Response Figure 1 below). To address this problem, the Virtual Lab incorporates AlphaFold-Multimer and Rosetta, both of which model the nanobodies in complex with the SARS-CoV-2 spike protein, thereby providing a computational estimate of binding.

Notably, when analyzing our two mutant nanobody hits, the ESM score alone cannot explain why those mutant nanobodies are hits and the others are not. The successful Ty1 mutant (V32F-G59D-N54S-F32S) had an ESM LLR^{WT} score of only 3.5 compared to the best Ty1 mutant ESM LLR^{WT} score of 7.0, making it 19th out of the 23 generated Ty1 mutants. In contrast, the successful Ty1 mutant's AF ipLDDT score was 8th highest out of 23, indicating that it was more promising according to a target-specific scoring mechanism. The previous approach using ESM alone would have likely missed this successful candidate. Similarly, the successful Nb21 mutant (I77V-L59E-Q87A-R37Q) had an ESM LLR^{WT} score of only 7.5 compared to the best Nb21 mutant ESM LLR^{WT} score of 11.3. However, it was 4th out of 23 in terms of both AF ipLDDT and RS dG scores. ESM score alone would imply that several other Ty1 and Nb21 mutants with higher ESM scores than the hits should also have been successful, but this was not the case, so the AF ipLDDT and RS dG scores helped introduce target-specific scores that elevated the rank of the successful mutant nanobodies.

That said, we believe that the primary benefit of the Virtual Lab is to automate the usage of existing powerful computational tools for complex scientific problems like nanobody design for SARS-CoV-2 rather than necessarily creating the best new tool by itself. This is a problem that would normally require a large team of interdisciplinary scientific experts but now can be tackled to a similar degree of success by a small team of human researchers augmented by the AI agents of the Virtual Lab.



Response Figure 1: ELISA binding intensity versus ESM log-likelihood scores for designed nanobodies from all four parental nanobodies (Nb21, VHH-72, H11-D4 and Ty1) and both JN.1 and KP.3 variants. The Pearson correlation values are low, indicating minimal correlation between ESM score and experimental binding intensity.

- Word counts alone do not capture the real supervisory burden. A timeline of human actions (prompt crafting, code review, troubleshooting, experiment planning) and their duration would make the “minimal supervision” claim more concrete.

Thank you for suggesting this timeline. To better illustrate the supervisory burden beyond word counts, we added the following to the “Analyses of Virtual Lab interactions” section.

“The Virtual Lab proceeded rapidly through the phases of the nanobody design workflow, with each meeting (or a set of parallel meetings) only taking the agents about 5-10 minutes (~\$1-2 GPT-4o token cost), for a total of around 1-2 hours (~\$10-20 GPT-4o token cost) to complete all of the phases of meetings.”

“Factoring in the time to tune the prompts for each phase to elicit the most productive and relevant discussions and to review and debug the code written by the agents, the Virtual Lab completed its nanobody workflow design in just a few days, compared to likely weeks for a human researcher working independently to design and build the same pipeline from scratch. It then took ~1 week to run the computational nanobody design pipeline (ESM + AlphaFold-Multimer + Rosetta), followed by ~6 weeks to synthesize the nanobodies and ~2 weeks for binding experiments.”

We provided our best estimate since it’s challenging to know the counterfactual of exactly how long it would have taken the same human researcher to design and build the pipeline without the help of the LLM agents, now that the Virtual Lab has already released the solution.

Referee #4 (Remarks on code availability):

Although the authors have deposited their code on github, reproducing the pipeline remains non-trivial. Several critical steps invoke external (ColabFold/AlphaFold-Multimer, Rosetta 2024.08) and require manual environment configuration. Providing a Docker/Singularity image or scripted setup would substantially lower the barrier to adoption/testing.

Thank you for reviewing our code and for suggesting the use of Docker/Singularity. We agree that a low barrier to entry is ideal. To this end, we have not only released all of our code on GitHub, we have also made the Virtual Lab available via the Python Package Index (PyPI) for single-line installation (e.g., `pip install virtual-lab` for general usage or `pip install virtual-lab[nanobody-design]` to install the specific tools like ESM and ColabFold for running the nanobody design pipeline). While we would like to make the nanobody design pipeline available on Docker or Singularity, the pipeline uses Rosetta, whose license does not permit its distribution in a Docker/Singularity image (please see <https://rosettacommons.org/software/licensing-faq/>). However, academic and commercial licenses for Rosetta are available for most users. We have documented our installation instructions (https://github.com/zou-group/virtual-lab/blob/main/nanobody_design/README.md) and all of the steps of the pipeline in our GitHub repo (https://github.com/zou-group/virtual-lab/blob/main/nanobody_design/workflow.md) in detail to make it easy for readers to reproduce our work and build upon it simply by following the provided scripts.

Referee #4 (Remarks to the Author):

The authors have provided a detailed response to my initial critiques. The revised manuscript is substantially strengthened, and they have appropriately addressed the major concerns raised in the first round of review.

Specifically, the authors have corrected the primary weakness of the original manuscript by performing three new sets of quantitative experiments. These additions provide rigorous, data-driven support for the central claims of the paper. They added:

- A scaled-up comparison of scientist vs. generic agents demonstrating a statistically significant performance benefit for the specialized agent architecture.
- A blinded human expert evaluation across 30 research questions that provides compelling evidence that the Virtual Lab's structured, multi-agent framework produces higher-quality scientific output than either a single agent or a team of generic agents.
- A quantitative analysis of the code generated that shows that the agent-critic interaction significantly improves code quality.

In addition, the authors' rebuttal regarding the experimental hit rate is well-reasoned. They successfully argue that their multi-tool pipeline (ESM, AlphaFold-Multimer, Rosetta) is tailored to a more complex problem than direct comparisons would suggest and provide new data showing that simpler metrics like ESM scores alone do not correlate with binding, justifying their more sophisticated approach.

The authors have also appropriately toned down their claims of domain-agnosticism and provided a more concrete estimation of the human labor involved, making the "minimal supervision" claim more credible.

In my opinion, the authors have been very diligent in their revisions. They have transformed the manuscript from a "very cool story" into a rigorous scientific study with strong, quantifiable evidence supporting its conclusions. The manuscript is much improved, and I recommend its publication.

Thank you very much for your feedback! We really appreciate your thoughtful comments and support!