

**Accelerated Article Preview**

# The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies

---

Received: 26 November 2024

Kyle Swanson, Wesley Wu, Nash L. Bulaong, John E. Pak & James Zou

Accepted: 22 July 2025

Accelerated Article Preview

Published online: 29 July 2025

Cite this article as: Swanson, K. et al.

The Virtual Lab of AI agents designs  
new SARS-CoV-2 nanobodies. *Nature*  
<https://doi.org/10.1038/s41586-025-09442-9>  
(2025)

This is a PDF file of a peer-reviewed paper that has been accepted for publication. Although unedited, the content has been subjected to preliminary formatting. Nature is providing this early version of the typeset paper as a service to our authors and readers. The text and figures will undergo copyediting and a proof review before the paper is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers apply.

1      The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies

2  
3      Kyle Swanson<sup>1</sup>, Wesley Wu<sup>2</sup>, Nash L. Bulaong<sup>2</sup>, John E. Pak<sup>2,4</sup>, James Zou<sup>1,2,3,4</sup>

4  
5      <sup>1</sup> Department of Computer Science, Stanford University, Stanford, CA, USA

6      <sup>2</sup> Chan Zuckerberg Biohub, San Francisco, CA, USA

7      <sup>3</sup> Department of Biomedical Data Science, Stanford University, Stanford, CA, USA

8      <sup>4</sup> Correspondence: [jamesz@stanford.edu](mailto:jamesz@stanford.edu), [john.pak@czbiohub.org](mailto:john.pak@czbiohub.org)

9      Abstract

10     Science frequently benefits from teams of interdisciplinary researchers<sup>1–3</sup>, but many  
11    scientists do not have easy access to experts from multiple fields<sup>4,5</sup>. While large language  
12    models (LLMs) have shown an impressive ability to aid researchers across diverse  
13    domains, their uses have been largely limited to answering specific scientific questions  
14    rather than performing open-ended research<sup>6–11</sup>. Here, we expand the capabilities of LLMs  
15    for science by introducing the Virtual Lab, an AI-human research collaboration to perform  
16    sophisticated, interdisciplinary science research. The Virtual Lab consists of an LLM  
17    principal investigator agent guiding a team of LLM scientist agents through a series of  
18    research meetings, with a human researcher providing high-level feedback. We apply the  
19    Virtual Lab to design nanobody binders to recent variants of SARS-CoV-2. The Virtual  
20    Lab creates a novel computational nanobody design pipeline that incorporates ESM,  
21    AlphaFold-Multimer, and Rosetta and designs 92 new nanobodies. Experimental validation  
22    reveals a range of functional nanobodies with promising binding profiles across SARS-  
23    CoV-2 variants. In particular, two new nanobodies exhibit improved binding to the recent  
24    JN.1 or KP.3 variants<sup>12,13</sup> while maintaining strong binding to the ancestral viral spike

25 protein, suggesting exciting candidates for further investigation. This demonstrates how the  
26 Virtual Lab can rapidly make an impactful, real-world scientific discovery.

27 Main

28 Interdisciplinary science research is complex, requiring increasingly large teams of researchers  
29 with expertise in diverse fields of science<sup>1–3</sup>. For example, the paper by Jumper et al.<sup>14</sup> that  
30 introduced AlphaFold 2 and later led to the 2024 Nobel Prize in Chemistry<sup>15</sup> included 34  
31 researchers with expertise across computer science, machine learning, bioinformatics, and  
32 structural biology. Building and coordinating large teams of researchers who speak different  
33 scientific languages and have different scientific priorities is challenging<sup>4,5</sup>. Furthermore, it can  
34 be harder for under-resourced groups without connections to many experts across fields to  
35 engage in complex, interdisciplinary science, especially when dedicated interdisciplinary  
36 research funding is lacking<sup>16</sup>.

37 One source of broad scientific knowledge and insights that researchers are now turning to is  
38 large language models (LLMs), such as ChatGPT<sup>17</sup> and Claude<sup>18</sup>. These LLMs have been trained  
39 on vast quantities of text data, including scientific literature, and they are therefore able to aid  
40 researchers in several ways such as by answering science questions, summarizing scientific  
41 papers, and writing scientific code<sup>19</sup>. Several studies have explored the scientific capabilities of  
42 LLMs by measuring their ability to answer scientific questions, and LLMs have shown high  
43 accuracy and can even match or outperform human scientists at these tasks<sup>6–11</sup>.

44 However, answering individual science questions is very different from engaging in  
45 sophisticated research that involves multi-step reasoning across disparate scientific fields with  
46 many unknowns. While some prior work has explored the application of LLMs to research, these

47 studies have often focused on a single scientific domain and have explored a relatively narrow  
48 set of research questions. For example, ChemCrow is a framework that gives GPT-4 access to  
49 chemistry tools and can thus solve components of a chemistry research problem, but it cannot  
50 tackle an open-ended, interdisciplinary research problem<sup>20</sup>. Another framework called  
51 Coscientist includes GPT-4-powered modules such as a planner and a web searcher to handle  
52 several aspects of research<sup>21</sup>. However, Coscientist is primarily applied to relatively standard  
53 chemistry tasks such as chemical synthesis planning as opposed to high-level research design  
54 across disciplines. In contrast, the AI Scientist aims to use LLMs to perform the entire scientific  
55 process from generating a hypothesis to writing code to drafting a paper, but the applications are  
56 limited to narrow subfields of machine learning without real-world experiments or validation<sup>22</sup>.  
57 Si et al.<sup>23</sup> similarly explore the use of LLMs for research idea generation and demonstrate  
58 promising results when comparing LLM research ideas to human research ideas, but the  
59 applications are limited to the field of natural language processing and do not include any  
60 implementation of the research ideas.

61 Here, we introduce the Virtual Lab to overcome these shortcomings via an AI-human research  
62 collaboration that performs interdisciplinary science to investigate broad, complex research  
63 questions. In the Virtual Lab, a human researcher guides a set of interdisciplinary AI agents<sup>24,25</sup>,  
64 such as a biologist or computer scientist, through a set of research meetings that tackle the  
65 different phases of a research project. The AI agents are run by an LLM that powers their  
66 scientific reasoning abilities with instructions that guide each agent's scientific expertise and  
67 interaction with the other agents and the human researcher. The Virtual Lab architecture is  
68 versatile and can potentially be applied to a wide variety of interdisciplinary science research  
69 projects.

70 To demonstrate the abilities of the Virtual Lab, we employ the Virtual Lab to tackle a high-  
71 impact, real-world, open-ended scientific problem: designing new nanobodies that exhibit  
72 binding to the latest variant of SARS-CoV-2. There are myriad ways in which scientists could  
73 attempt to design such nanobodies, so the Virtual Lab must reason across multiple subfields of  
74 biology and computer science to make a series of interrelated decisions about how best to design  
75 these nanobodies. Through a series of meetings, the Virtual Lab develops a novel computational  
76 nanobody design workflow that incorporates the protein language model ESM<sup>26</sup>, the protein  
77 folding model AlphaFold-Multimer<sup>27</sup>, and the computational biology software Rosetta<sup>28</sup> to  
78 mutate existing nanobodies that bind to the receptor binding domain (RBD) of the spike protein  
79 of the Wuhan strain of SARS-CoV-2 into nanobodies that bind to the latest variants of the virus,  
80 where an effective binder is lacking<sup>13</sup>. We experimentally validated 92 mutant nanobodies  
81 designed by the Virtual Lab and found that over 90% of the nanobodies were expressed and  
82 soluble, with two promising candidates showing unique binding profiles to the recent JN.1 and  
83 KP.3 spike RBD variants<sup>12,13</sup>. This outcome illustrates the capability of the Virtual Lab's AI-  
84 human collaboration to execute a complex, interdisciplinary science research project that  
85 translates to a validated result in the real world.

## 86 Virtual Lab architecture

87 We created the Virtual Lab as a collaboration between a human researcher and a team of large  
88 language model (LLM) agents to conduct sophisticated, interdisciplinary research (Fig. 1). The  
89 human researcher provides high-level guidance for the LLM agents while the LLM agents both  
90 decide on general research directions and design solutions to specific research problems. Each  
91 agent is implemented by providing the underlying LLM with a prompt defining the agent, which

92 includes its title, expertise, goal, and role in the research project (see Methods). The human  
93 researcher defines two general agents, a Principal Investigator (PI) and a Scientific Critic, and  
94 the PI agent then automatically creates a set of scientific agents (e.g., an Immunologist)  
95 depending on the scientific topic of interest to the human researcher (Fig. 1a).

96 The Virtual Lab performs research via meetings of two forms: team meetings and individual  
97 meetings (see Methods). In both cases, the human researcher provides an initial agenda to guide  
98 the discussion, and then the agents discuss how to address the agenda. In team meetings (Fig.  
99 1b), all of the agents discuss a broad research question and work together to come up with an  
100 answer. In individual meetings (Fig. 1c), a single scientific agent is given a more specific task to  
101 accomplish, such as writing code for a machine learning model, and the agent either works alone  
102 or in conjunction with the Scientific Critic agent, which provides critical feedback. Both forms of  
103 meetings can be run multiple times in parallel followed by an aggregation meeting to generate  
104 more robust answers (Extended Data Fig. 1). Through a series of team and individual meetings,  
105 the Virtual Lab tackles a complex research project.

## 106 Virtual Lab for nanobody design

107 Given the flexibility of the Virtual Lab architecture, the Virtual Lab can be applied to a wide  
108 variety of interdisciplinary research projects by adapting the agents and the flow of team and  
109 individual meetings to the specific project's goals and constraints. As a demonstration in the  
110 domain of biological research, we applied the Virtual Lab with GPT-4o<sup>29</sup> powering the agents to  
111 design antibodies or nanobodies that can bind to the spike protein of the KP.3 variant of SARS-  
112 CoV-2, which was one of the latest variants at the time of this work<sup>13</sup> (Fig. 2). This is an  
113 important and challenging problem because SARS-CoV-2 is rapidly evolving resistance to

existing antibody/nanobody therapies, so quickly developing new antibody/nanobody therapies that overcome this resistance and bind to the latest variants is crucial to treating those who are infected<sup>30,31</sup>. The Virtual Lab tackles this problem by rapidly creating a computational workflow to design antibodies or nanobodies for the KP.3 variant of SARS-CoV-2, which can then be experimentally validated by human biologists. The Virtual Lab created the computational antibody/nanobody design process in five phases.

1. **Team selection:** An individual meeting with the PI to define a set of scientist agents to work on the project (Fig. 2a).

2. **Project specification:** A team meeting to specify the project direction by deciding on key high-level details (Fig. 2b).

3. **Tools selection:** A team meeting to brainstorm machine learning and/or computational tools for nanobody design (Fig. 2c).

4. **Tools implementation:** A series of individual meetings to implement three components of the nanobody design workflow—ESM, AlphaFold-Multimer, and Rosetta (Fig. 2d). 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.

5. **Workflow design:** An individual meeting with the PI to determine the workflow for applying these computational tools (Fig. 2e).

135 These phases are discussed in more detail in the Methods.

## 136 Computational nanobody design

137 The Virtual Lab built a computational nanobody design workflow that takes existing nanobodies  
138 that bind the Wuhan strain of SARS-CoV-2 and adapts them to bind to the recent KP.3 variant  
139 (Fig. 3a). Specifically, the workflow starts with four nanobodies—Ty1<sup>32</sup>, H11-D4<sup>33</sup>, Nb21<sup>34</sup>, and  
140 VHH-72<sup>35</sup>—and uses three tools—the protein language model ESM, the protein folding model  
141 AlphaFold-Multimer, and the protein modeling software Rosetta—to iteratively introduce point  
142 mutations into those nanobodies to improve their binding to the receptor binding domain (RBD)  
143 of the spike protein from the KP.3 variant of SARS-CoV-2 (see Methods).

144 In the workflow, first, ESM computes the log-likelihood ratio (LLR) of each single point  
145 mutation in the nanobody sequence compared to the input nanobody sequence, with higher ESM  
146 LLRs indicating better (e.g., more stable) nanobodies (see Supplementary Note 1 for a discussion  
147 of the LLR formula created by the Virtual Lab). Then, the top 20 mutant sequences by ESM  
148 LLR are combined with the KP.3 RBD and processed by AlphaFold-Multimer, which predicts  
149 the structure of the complex of the two proteins and computes the interface pLDDT (AF  
150 ipLDDT) as a measure of the confidence of the binding interface between the mutant nanobody  
151 and the spike RBD. Next, those 20 predicted nanobody-spike complexes are fed into Rosetta,  
152 which relaxes their structures and computes the binding energy (RS dG). The ESM LLR, AF  
153 ipLDDT, and RS dG scores are combined into a weighted score using the formula WS = 0.2 \*  
154 (ESM LLR) + 0.5 \* (AF ipLDDT) - 0.3 \* (RS dG). The 20 mutant nanobodies are ranked by WS  
155 and the top five are selected. Those five are then fed back into the pipeline to introduce another  
156 round of point mutations. The process is repeated four times total to introduce up to four

157 mutations. Finally, 23 mutated nanobodies are selected for each of the four starting nanobodies  
158 (92 total) using a modified weighted score  $WS^{WT}$ . This score is the same as the WS except that it  
159 uses a modified ESM LLR<sup>WT</sup>, which is the ratio between the proposed nanobody (with one to  
160 four mutations) and the wild-type nanobody (with zero mutations) rather than the input  
161 nanobody from the previous round (with one less mutation).

162 The successive rounds of optimization improved the quality of the proposed mutant nanobody  
163 sequences according to the three metrics of ESM LLR, AF ipLDDT, and RS dG. Fig. 3b-g show  
164 relevant metrics for Nb21, while results are similar for Ty1 (Extended Data Fig. 2), H11-D4  
165 (Extended Data Fig. 3), and VHH-72 (Extended Data Fig. 4). Extended Data Table 1 shows  
166 scores for the wild-type sequence and some of the mutant sequences that were selected for  
167 experimental validation.

168 In each round, the top sequences selected by ESM LLR had log-likelihood ratios of roughly 1-8,  
169 indicating that each subsequent round of mutation improved the overall quality of the nanobody  
170 compared to the input sequence from the previous round (Fig. 3b). This is according to ESM's  
171 internal understanding of nanobody likelihood, which does not take the antigen (spike protein)  
172 into account but does understand overall nanobody quality. The top mutant nanobody sequences  
173 selected by ESM LLR in each round generally had improved structural complexes with the KP.3  
174 spike protein based on improved AF ipLDDT, improved RS dG, or both (Fig. 3c). The WS  
175 values of the top five sequences at the end of each round improved (Fig. 3d), even when using  
176 the ESM LLR instead of the ESM LLR<sup>WT</sup> that corrects for the effect of multiple mutations and  
177 not just the most recent mutation.

178 After correction, the ESM LLR<sup>WT</sup> for the final selected 23 sequences showed a large  
179 improvement over the wild-type sequence (Fig. 3e). These selected sequences also had improved  
180 AF ipLDDT and RS dG scores compared to the wild-type (Fig. 3f). An example of the  
181 AlphaFold-Multimer predicted structure (with Rosetta relaxation) of a top scoring mutant  
182 nanobody is shown in Fig. 3g. Notably, the final set of 23 selected nanobody sequences includes  
183 sequences with different numbers of mutations (i.e., from different rounds) and with a different  
184 balance of ESM LLR<sup>WT</sup>, AF ipLDDT, and RS dG values, allowing for a diversity of potential  
185 improvements to the wild-type nanobody.

186 Applying this workflow to each of the four starting nanobodies resulted in 92 final selected  
187 sequences (23 per starting nanobody). All 92 mutant nanobodies had a positive ESM LLR,  
188 indicating that ESM preferred the mutant over the wild-type. Among the 92 mutant  
189 nanobodies, 78 (85%) had an AF ipLDDT greater than their respective wild-type  
190 nanobody, and 32 (35%) had an AF ipLDDT  $\geq 80$ , which is in line with the AF  
191 ipLDDT scores of high accuracy AlphaFold–Multimer antibody–antigen structural  
192 models in prior work<sup>36</sup>. Furthermore, 60 (65%) had an RS dG less (better) than  
193 their respective wild-type nanobody, and 23 (25%) of the 92 mutants had an RS  
194 dG  $\leq -50$ , which is in line with strong Rosetta binding energy values of nanobodies  
195 or antibodies in complex with the SARS-CoV-2 receptor binding domain in prior  
196 work<sup>28,37</sup>.

197 Nanobody experimental validation

198 To validate the nanobodies designed by the Virtual Lab, we conducted a set of experiments to  
199 measure their binding to a panel of spike receptor binding domain (RBD) proteins (Extended  
200 Data Fig. 5). We first overexpressed each nanobody in *E. coli*, followed by isolation of soluble  
201 protein from the periplasm. The designed nanobodies show excellent expression, with 38% (35  
202 of 92) of the designs having titers of > 25 mg of soluble, periplasmic nanobody per liter of cell  
203 culture (Fig. 4a, Extended Data Fig. 6) and only 6.5% (6 of 92) of the designs having a titer of <  
204 5 mg/L. Thus, the mutations proposed by the Virtual Lab are well tolerated and do not cause  
205 large-scale misfolding or aggregation of the nanobodies.

206 To determine if the 92 mutant nanobodies—23 each for Ty1, H11-D4, Nb21, and VHH-72—and  
207 the four wild-type nanobodies could bind to the SARS-CoV-2 KP.3 spike RBD, we generated a  
208 spike RBD array that included the KP.3 RBD protein, its closely related parental strain (JN.1  
209 RBD), a closely related variant (KP.2.3 RBD), an early Omicron variant (BA.2 RBD), and the  
210 ancestral strain (Wuhan RBD), which all four wild-type nanobodies show specificity for.

211 Using this RBD array, we first profiled the binding of all 96 nanobodies by indirect ELISA to  
212 each antigen at a nanobody lysate dilution of 1:2 (Fig. 4b). For the H11-D4 and Nb21 series,  
213 binding to Wuhan RBD is overwhelmingly retained in 96% of mutant nanobodies (44 of 46).  
214 Three mutants in the H11-D4 series have high non-specific binding to BSA and all of the RBDs  
215 (Fig. 4b), possibly owing to the Virtual Lab inadvertently introducing an R27C mutation, which  
216 may be leading to disulfide crosslinking. In contrast to the H11-D4 and Nb21 mutants, the Ty1  
217 mutants, overall, exhibit poor binding to Wuhan RBD (10 of 23 mutants). If position 32 of Ty1,  
218 selected by the Virtual Lab as the first residue to mutate for each mutant, is not well tolerated,

219 this could result in the observed poor binding to Wuhan RBD compared to the H11-D4 and Nb21  
220 mutants. Over half of the VHH-72 mutants (13 of 22) retain binding to Wuhan RBD at levels  
221 similar to that observed for the unmutated VHH-72 nanobody. Thus, the Virtual Lab designs are,  
222 overall, well tolerated with respect to preserving their original specificity to Wuhan RBD.

223 Of the 92 Virtual Lab-designed nanobodies, two show promising binding profiles beyond Wuhan  
224 RBD. The first, derived from Nb21, I77V-L59E-Q87A-R37Q (i.e., Nb21 with the mutations  
225 I77V, L59E, Q87A, and R37Q), has positive ELISA binding to JN.1 RBD with no non-specific  
226 binding to MERS-CoV RBD and BSA (Fig. 4b-d). Maximal binding of the purified mutant  
227 nanobody to JN.1 RBD is less than that to Wuhan RBD, with a weaker EC<sub>50</sub> (2.0 ng/mL vs 0.2  
228 ng/mL) revealing that this new binding to JN.1 RBD may be moderate. The wild-type Nb21 has  
229 very low ELISA binding to JN.1 RBD (Fig. 4b), suggesting that the Virtual Lab mutant has  
230 improved upon this existing very weak binding. Interestingly, this mutant also shows some  
231 binding to KP.3 RBD (average intensity = 3.5) compared to the other Nb21 mutants (average  
232 intensity =  $0.06 \pm 0.09$ , n = 22) and the unmutated sequence (average intensity = 0.1) (Fig. 4b).  
233 We further confirmed this KP.3 binding enrichment in separate ELISA experiments. The second,  
234 a Ty1 mutant nanobody (V32F-G59D-N54S-F32S) not only improved binding to Wuhan RBD,  
235 as measured by ELISA, but also gained moderate binding to JN.1 RBD (Fig. 4b-d). In contrast,  
236 we see no evidence for even low levels of unmutated Ty1 nanobody binding to JN.1 RBD.  
237 (Further details in the Supplementary Note 2.)

238 Across the mutant nanobodies, the preserved and improved binding affinities for the Wuhan  
239 RBD (and JN.1 RBD for Nb21) relative to their respective wild-type forms is likely due to the  
240 effect of the Virtual Lab's use of ESM log-likelihoods, which are agnostic to the antigen but  
241 select for evolutionarily favorable nanobody sequences with improved fitness<sup>38</sup>. In contrast, the

242 Ty1 mutant that gained binding affinity for the JN.1 RBD, which is the close ancestor of KP.3  
243 sharing 99.1% identity in the RBD<sup>39</sup> (220 of 222 residues), and the Nb21 mutant that gained  
244 binding affinity for the KP.3 RBD and improved binding affinity for the JN.1 RBD may  
245 demonstrate the effect of the AlphaFold-Multimer and Rosetta scoring, which explicitly aim to  
246 predict binding affinity of the mutant antibody to the KP.3 RBD, and thus by extension the  
247 closely related JN.1 RBD. Through the use of these three tools, the Virtual Lab designed a set of  
248 promising nanobody candidates with potential for further development.

## 249 Analyses of Virtual Lab interactions

250 The Virtual Lab proceeded rapidly through the phases of the nanobody design workflow, with  
251 each meeting (or a set of parallel meetings) only taking the agents about 5-10 minutes (~\$1-2  
252 GPT-4o token cost), for a total of around 1-2 hours (~\$10-20 GPT-4o token cost) to complete all  
253 of the phases of meetings. Factoring in the time to tune the prompts for each phase to elicit the  
254 most productive and relevant discussions and to review and debug the code written by the agents,  
255 the Virtual Lab completed its nanobody workflow design in just a few days, compared to likely  
256 weeks for a human researcher working independently to design and build the same pipeline from  
257 scratch. It then took ~1 week to run the computational nanobody design pipeline (ESM +  
258 AlphaFold-Multimer + Rosetta), followed by ~6 weeks to synthesize the nanobodies and ~2  
259 weeks for binding experiments.

260 Within the Virtual Lab discussions, interestingly, the individual identities of the agents  
261 contributed to a comprehensive, interdisciplinary discussion with each agent providing  
262 perspective based on its specific background (Fig. 5a, additional analysis in Supplementary Note  
263 3). In contrast, a team of generic agents without distinct scientific backgrounds tends to argue

more amongst themselves due to a lack of clearly defined roles, more often leading to suboptimal answers (see ablation experiments in Supplementary Note 4). The human researcher only needs to provide minimal text input, writing just 1,596 words (defined as space-separate tokens) across all phases of the workflow, representing just 1.3% of all the words written by the Virtual Lab (Fig. 5b). In contrast, the LLM agents wrote 122,462 words (98.7% of all words). The ESM, AlphaFold-Multimer, and Rosetta scripts were all written from scratch by the agents while 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.

The Virtual Lab meetings reveal interesting dynamics among the agents that affect the agents' scientific discussions. For example, different agents write different amounts depending on the meeting context (Fig. 5b-e, Extended Data Fig. 7). In team meetings, the PI tends to write the most, which is reasonable given that the PI not only has to synthesize the agent responses after each round of discussion to guide the next round but also has to initiate the meeting and write a summary at the end. The Scientific Critic writes more than the scientist agents since the Scientific Critic must address the limitations of every agents' response while each scientist agent is only concerned with providing its own opinions. The use of parallel meetings (see Methods) and the inclusion of the Scientific Critic were particularly notable since they tended to lead to answers with improved consistency and quality (see Supplementary Note 3).

## Discussion

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,

286 diverse binding profiles across multiple strains of SARS-CoV-2. The human researcher and team  
287 of LLM agents in the Virtual Lab worked together through a series of meetings to rapidly build a  
288 complex nanobody design pipeline that incorporates state-of-the-art machine learning and  
289 computational biology tools. Building this pipeline required knowledge of multiple areas of  
290 science from immunology to protein folding to machine learning and required making decisions  
291 that involved reasoning across many aspects of the project simultaneously. The Virtual Lab  
292 successfully built and ran this nanobody design pipeline, starting with a set of four well-  
293 characterized nanobodies (Ty1, H11-D4, Nb21, and VHH-72) with potency and diverse binding  
294 modes against early variants of SARS-CoV-2<sup>32–35</sup> and developing them into 92 nanobody  
295 candidates for recent variants of SARS-CoV-2 that were experimentally validated by human  
296 researchers. These 92 nanobodies—efficiently selected from the trillions of nanobody sequences  
297 with one to four mutations—include exciting candidates for further development, such as an  
298 Nb21 mutant that enhances binding to the JN.1 RBD and gains binding to the KP.3 RBD and a  
299 Ty1 mutant that gains binding to the JN.1 RBD. This outcome serves as an example of how  
300 human researchers can partner with LLM agents in the Virtual Lab to rapidly achieve a  
301 promising scientific result that can streamline further experiments. Even if the ultimate scientific  
302 decisions of the Virtual Lab agents are similar to those in the scientific literature, the ability of  
303 the agents to quickly adapt those methods to the scientific question at hand shows how LLM  
304 agents can potentially empower human researchers to do complex, interdisciplinary science even  
305 when they do not have access to an expert panel of human scientists.

306 Previous work applying AI to science has generally treated AI methods as tools used by human  
307 researchers, such as AlphaFold to predict protein structures<sup>14</sup> or LLMs to answer scientific  
308 questions<sup>6–11</sup>, with the human researcher making all the high-level research decisions and design

309 choices. In contrast, in the Virtual Lab, human researchers work alongside LLM agents to design  
310 and run a research project. The strength of the Virtual Lab comes from its multi-agent<sup>24,40–42</sup>  
311 architecture, which empowers an AI-human scientific collaboration via a series of meetings  
312 between a human researcher and a team of interdisciplinary LLM agents. The different  
313 backgrounds of the various scientist agents leads to discussions that approach complicated  
314 scientific questions from multiple angles, thereby contributing to comprehensive answers.

315 Furthermore, the PI agent helps guide the discussions, make key decisions, and summarize  
316 conversations for the human researcher, while the Scientific Critic agent pushes the other agents  
317 to improve their answers to maximize the quality of their science. The inclusion of the human  
318 researcher is also vital as it enables the human to provide high-level guidance where the agents  
319 lack relevant context, such as choosing readily available computational tools and introducing  
320 constraints in experimental validation. The team and individual meetings provide two distinct  
321 forums for discussion in the Virtual Lab, enabling high-level conversations about research  
322 directions in the team meetings and low-level implementation of specific solutions in the  
323 individual meetings. Throughout these meetings, the extended conversations between  
324 interdisciplinary agents extracts knowledge and reasoning abilities from the underlying LLM in a  
325 similar way to chain-of-thought prompting<sup>43</sup> but with the added benefit of different agent  
326 perspectives and a human-in-the-loop to guide the conversations.

327 While the Virtual Lab architecture provides useful structure for scientific discussions between  
328 the human researcher and the LLM agents, it has several limitations that are inherent in the  
329 current generation of LLMs. For example, since LLMs are only trained on data up to a certain  
330 date (the “knowledge cutoff”), the agents may not be aware of the most up-to-date scientific  
331 literature and code<sup>44</sup> (e.g., AlphaFold 3<sup>45</sup> instead of AlphaFold-Multimer<sup>27</sup>). However, these

332 issues could be fixed by providing the agents with relevant information and documentation, for  
333 example through retrieval-augmented generation<sup>46,47</sup> or finetuning<sup>48</sup> (see Supplementary Note 5  
334 for an exploration of finetuning agents). Additionally, future work could explore developing  
335 sandboxed environments to allow the agents to independently install computational or AI tools  
336 and then write, debug, and run code that use those tools for a particular application.

337 Another challenge faced by the Virtual Lab, which is also an inherent LLM limitation, is the  
338 need for prompt engineering to obtain useful answers from the LLM agents<sup>49</sup>. Without  
339 appropriate guidance, the LLM agents can give vague answers. This means that the human  
340 researcher may have to iterate on a meeting agenda several times before the Virtual Lab provides  
341 a desirable response (see Supplementary Note 4). Even so, the role of prompt engineering in the  
342 Virtual Lab may shrink as the underlying LLMs are further improved.

343 The current generation of LLMs is also known to sometimes provide incorrect or misleading  
344 answers (often termed hallucinations<sup>50</sup>). In the context of the Virtual Lab, this could mean that  
345 the agents might invent incorrect scientific facts or citations. These limitations could be partly  
346 mitigated through multi-agent interactions such as having the critic question the veracity of  
347 information provided by the other agents or by providing the agents with access to resources  
348 (e.g., the text of scientific papers) to verify their knowledge. It is still important for the human  
349 researcher working with the Virtual Lab to verify key facts and decisions based on trusted  
350 scientific sources.

351 While we applied the Virtual Lab to nanobody design here, the Virtual Lab architecture of LLM  
352 agents and meetings is agnostic to specific research questions or scientific domains. The Virtual  
353 Lab can be implemented with any set of scientist agents and any human researcher, and the

354 conversations in the meetings will naturally adapt based on the human researcher's agenda and  
355 the backgrounds of the agents. Even the underlying LLM that powers the agents could be  
356 exchanged, meaning the Virtual Lab can improve its scientific abilities as LLMs grow more  
357 capable. However, even as the Virtual Lab expands its capabilities, human scientists will still be  
358 vital to guide the AI agents in their choice of scientific questions, methodologies, and analyses to  
359 match the scientific values and interests of the human researchers. While the experimental results  
360 here are limited to the domain of nanobody design, with future work, we envision the Virtual  
361 Lab as a powerful framework for human researchers to engage in interdisciplinary science  
362 research with the help of LLMs.

### 363 References

- 364 1. Porter, A. L. & Rafols, I. Is science becoming more interdisciplinary? Measuring and mapping  
365 six research fields over time. *Scientometrics* **81**, 719–745 (2009).
- 366 2. Sijp, W. Paper authorship goes hyper. *Nature Index* (2018).
- 367 3. Castelvecchi, D. Physics paper sets record with more than 5,000 authors. *Nature*  
368 nature.2015.17567 (2015) doi:10.1038/nature.2015.17567.
- 369 4. Specht, A. & Crowston, K. Interdisciplinary collaboration from diverse science teams can  
370 produce significant outcomes. *PLOS ONE* **17**, e0278043 (2022).
- 371 5. Cohen, J. J. *et al.* Tackling the challenge of interdisciplinary energy research: A research  
372 toolkit. *Energy Res. Soc. Sci.* **74**, 101966 (2021).
- 373 6. Kung, T. H. *et al.* Performance of ChatGPT on USMLE: Potential for AI-assisted medical  
374 education using large language models. *PLOS Digit. Health* **2**, e0000198 (2023).
- 375 7. Singhal, K. *et al.* Large language models encode clinical knowledge. *Nature* **620**, 172–180

- 376 (2023).
- 377 8. Laurent, J. M. *et al.* LAB-Bench: Measuring Capabilities of Language Models for Biology  
378 Research. Preprint at <https://doi.org/10.48550/arXiv.2407.10362> (2024).
- 379 9. Guo, T. *et al.* What can Large Language Models do in chemistry? A comprehensive  
380 benchmark on eight tasks. *Adv. Neural Inf. Process. Syst.* **36**, 59662–59688 (2023).
- 381 10. Sun, L. *et al.* SciEval: A Multi-Level Large Language Model Evaluation Benchmark for  
382 Scientific Research. *Proc. AAAI Conf. Artif. Intell.* **38**, 19053–19061 (2024).
- 383 11. Stribling, D. *et al.* The model student: GPT-4 performance on graduate biomedical  
384 science exams. *Sci. Rep.* **14**, 5670 (2024).
- 385 12. Kaku, Y. *et al.* Virological characteristics of the SARS-CoV-2 JN.1 variant. *Lancet  
386 Infect. Dis.* **24**, e82 (2024).
- 387 13. Kaku, Y. *et al.* Virological characteristics of the SARS-CoV-2 KP.3, LB.1, and KP.2.3  
388 variants. *Lancet Infect. Dis.* **24**, e482–e483 (2024).
- 389 14. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**,  
390 583–589 (2021).
- 391 15. Callaway, E. Chemistry Nobel goes to developers of AlphaFold AI that predicts protein  
392 structures. *Nature* **634**, 525–526 (2024).
- 393 16. Bromham, L., Dinnage, R. & Hua, X. Interdisciplinary research has consistently lower  
394 funding success. *Nature* **534**, 684–687 (2016).
- 395 17. OpenAI *et al.* GPT-4 Technical Report. Preprint at <http://arxiv.org/abs/2303.08774>  
396 (2024).
- 397 18. Anthropic. The Claude 3 Model Family: Opus, Sonnet, Haiku. (2024).
- 398 19. Simon, E., Swanson, K. & Zou, J. Language models for biological research: a primer.

- 399        *Nat. Methods* **21**, 1422–1429 (2024).
- 400      20.     M. Bran, A. *et al.* Augmenting large language models with chemistry tools. *Nat. Mach. Intell.* **6**, 525–535 (2024).
- 401      21.     Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research
- 402      with large language models. *Nature* **624**, 570–578 (2023).
- 403      22.     Lu, C. *et al.* The AI Scientist: Towards Fully Automated Open-Ended Scientific
- 404      Discovery. Preprint at <https://doi.org/10.48550/arXiv.2408.06292> (2024).
- 405      23.     Si, C., Yang, D. & Hashimoto, T. Can LLMs Generate Novel Research Ideas? A Large-
- 406      Scale Human Study with 100+ NLP Researchers. in *The Thirteenth International Conference*
- 407      *on Learning Representations* (2024).
- 408      24.     Wu, Q. *et al.* AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent
- 409      Conversation. Preprint at <https://doi.org/10.48550/arXiv.2308.08155> (2023).
- 410      25.     Gao, S. *et al.* Empowering biomedical discovery with AI agents. *Cell* **187**, 6125–6151
- 411      (2024).
- 412      26.     Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a
- 413      language model. *Science* **379**, 1123–1130 (2023).
- 414      27.     Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. Preprint at
- 415      <https://doi.org/10.1101/2021.10.04.463034> (2021).
- 416      28.     Boorla, V. S. *et al.* De novo design and Rosetta-based assessment of high-affinity
- 417      antibody variable regions (Fv) against the SARS-CoV -2 spike receptor binding domain (
- 418      RBD ). *Proteins Struct. Funct. Bioinforma.* **91**, 196–208 (2023).
- 419      29.     OpenAI *et al.* GPT-4o System Card. Preprint at <http://arxiv.org/abs/2410.21276> (2024).
- 420      30.     Cao, Y. *et al.* Omicron escapes the majority of existing SARS-CoV-2 neutralizing

- 422 antibodies. *Nature* **602**, 657–663 (2022).
- 423 31. Planas, D. *et al.* Considerable escape of SARS-CoV-2 Omicron to antibody  
424 neutralization. *Nature* **602**, 671–675 (2022).
- 425 32. Hanke, L. *et al.* An alpaca nanobody neutralizes SARS-CoV-2 by blocking receptor  
426 interaction. *Nat. Commun.* **11**, 4420 (2020).
- 427 33. Huo, J. *et al.* Neutralizing nanobodies bind SARS-CoV-2 spike RBD and block  
428 interaction with ACE2. *Nat. Struct. Mol. Biol.* **27**, 846–854 (2020).
- 429 34. Xiang, Y. *et al.* Versatile and multivalent nanobodies efficiently neutralize SARS-CoV-2.  
430 *Science* **370**, 1479–1484 (2020).
- 431 35. Wrapp, D. *et al.* Structural Basis for Potent Neutralization of Betacoronaviruses by  
432 Single-Domain Camelid Antibodies. *Cell* **181**, 1004-1015.e15 (2020).
- 433 36. Yin, R. & Pierce, B. G. Evaluation of AlphaFold antibody–antigen modeling with  
434 implications for improving predictive accuracy. *Protein Sci.* **33**, e4865 (2024).
- 435 37. Yang, J. *et al.* Computational design and modeling of nanobodies toward SARS-CoV-2  
436 receptor binding domain. *Chem. Biol. Drug Des.* **98**, 1–18 (2021).
- 437 38. Hie, B. L. *et al.* Efficient evolution of human antibodies from general protein language  
438 models. *Nat. Biotechnol.* **42**, 275–283 (2024).
- 439 39. Planas, D. *et al.* Escape of SARS-CoV-2 Variants KP.1.1, LB.1, and KP3.3 From  
440 Approved Monoclonal Antibodies. *Pathog. Immun.* **10**, 1 (2024).
- 441 40. Chan, C.-M. *et al.* ChatEval: Towards Better LLM-based Evaluators through Multi-  
442 Agent Debate. in (2023).
- 443 41. Liu, Z., Zhang, Y., Li, P., Liu, Y. & Yang, D. A Dynamic LLM-Powered Agent Network  
444 for Task-Oriented Agent Collaboration. Preprint at <https://doi.org/10.48550/arXiv.2310.02170>

- 445 (2024).
- 446 42. Talebirad, Y. & Nadiri, A. Multi-Agent Collaboration: Harnessing the Power of  
447 Intelligent LLM Agents. Preprint at <https://doi.org/10.48550/arXiv.2306.03314> (2023).
- 448 43. Wei, J. *et al.* Chain-of-thought prompting elicits reasoning in large language models. in  
449 *Proceedings of the 36th International Conference on Neural Information Processing Systems*  
450 24824–24837 (Curran Associates Inc., Red Hook, NY, USA, 2024).
- 451 44. Cheng, J. *et al.* Dated Data: Tracing Knowledge Cutoffs in Large Language Models. in  
452 *First Conference on Language Modeling* (2024).
- 453 45. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with  
454 AlphaFold 3. *Nature* **630**, 493–500 (2024).
- 455 46. Lewis, P. *et al.* Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. in  
456 *Advances in Neural Information Processing Systems* vol. 33 9459–9474 (Curran Associates,  
457 Inc., 2020).
- 458 47. Gao, Y. *et al.* Retrieval-Augmented Generation for Large Language Models: A Survey.  
459 Preprint at <https://doi.org/10.48550/arXiv.2312.10997> (2024).
- 460 48. Ding, N. *et al.* Parameter-efficient fine-tuning of large-scale pre-trained language models.  
461 *Nat. Mach. Intell.* **5**, 220–235 (2023).
- 462 49. White, J. *et al.* A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT.  
463 in *Proceedings of the 30th Conference on Pattern Languages of Programs* 1–31 (The Hillside  
464 Group, USA, 2023).
- 465 50. Ji, Z. *et al.* Survey of Hallucination in Natural Language Generation. *ACM Comput Surv*  
466 **55**, 248:1-248:38 (2023).
- 467 51. Meng, E. C. *et al.* UCSF ChimeraX: Tools for structure building and analysis. *Protein*

469 **Figure legends**470 **Fig. 1 | Virtual Lab architecture.** **a**, The workflow for designing agents in the Virtual Lab.

471 Each agent is specified with four criteria: title, expertise, goal, and role. The human researcher in  
472 the Virtual Lab specifies these criteria to define the Principal Investigator (PI) agent and the  
473 Scientific Critic agent. Then, given a short description of the project by the human researcher,  
474 the PI agent automatically creates several scientist agents to work on the project by specifying  
475 their title, expertise, goal, and role, using its own prompt as an example. **b**, The workflow for a  
476 team meeting in the Virtual Lab. The human researcher writes an agenda for the meeting  
477 specifying the topic of discussion. The PI agent begins the meeting by providing initial thoughts  
478 and agenda questions as a guide for the remaining agents. Then, over the course of N rounds of  
479 discussion, each scientist agent provides its response, followed by a critique by the Scientific  
480 Critic agent, with the PI agent then synthesizing the discussion and asking follow-up questions.  
481 Finally, after the N rounds of discussion, the PI agent summarizes the discussion and provides an  
482 answer regarding the meeting agenda. **c**, The workflow for an individual meeting. The human  
483 researcher writes an agenda for the meeting specifying the topic of discussion. Then, the scientist  
484 agent tasked with the individual meeting provides a response to the agenda, which is critiqued by  
485 the Scientific Critic. In each round, the scientist agent improves its answer based on feedback  
486 from the Scientific Critic. Finally, after the N rounds, the scientist agent provides its final,  
487 improved answer.

488 **Fig. 2 | Virtual Lab for nanobody design.** The workflow used to apply the Virtual Lab to489 nanobody design for the latest variant of SARS-CoV-2. **a**, The workflow begins with the human

490 researcher defining the Principal Investigator (PI) and Scientific Critic agents by specifying their  
491 title, expertise, goal, and role. Then, in an individual meeting, the PI agent creates a team of three  
492 scientist agents for the project. **b**, A team meeting discusses the project specification, and the  
493 agents make decisions such as whether to design antibodies or nanobodies. **c**, In another team  
494 meeting, the agents suggest a set of computational tools for nanobody design, including ESM,  
495 AlphaFold-Multimer, and Rosetta. **d**, In a series of individual meetings, the Machine Learning  
496 Specialist and Computational Biologist, with helpful feedback from the Scientific Critic, write  
497 code and subsequently improve that code for the ESM, AlphaFold-Multimer, and Rosetta  
498 components of the nanobody design workflow. **e**, In an individual meeting, the PI agent decides  
499 the workflow for using the three computational tools to design and select mutated nanobody  
500 candidates.

501 **Fig. 3 | Nb21 nanobody analysis.** **a**, Each round of nanobody design begins with ESM  
502 computing log-likelihood ratios (ESM LLR) of single point mutations to the input sequence. For  
503 the top 20 mutant sequences by ESM LLR, AlphaFold-Multimer predicts the structure of the  
504 nanobody and SARS-CoV-2 spike protein and computes the interface pLDDT (AF ipLDDT).  
505 Rosetta relaxes the complex and computes a binding energy (RS dG). The top five mutant  
506 nanobodies are selected via a weighted score (WS) for the next round of optimization. **b-d**,  
507 Evolution of mutant nanobody scores across four rounds of optimization. **b**, The distribution of  
508 ESM LLR values for proposed Nb21 mutant nanobodies across each round of optimization, with  
509 ESM LLR values computed relative to the input nanobody sequence from the previous round.  
510 Shown are the ESM LLR values of the top 20 proposed mutant nanobodies per input nanobody.  
511 **c**, The AF ipLDDT and the RS dG of the top five proposed nanobodies, selected by WS, at the  
512 end of each round of optimization. **d**, The distribution of WS values of the top five proposed

513 nanobodies at the end of each round of optimization. **e-g**, Analysis of the final set of 23 mutant  
514 nanobodies selected across all rounds of optimization. **e**, The distribution of ESM LLR<sup>WT</sup> values  
515 (ESM LLR of the mutant sequence compared to the wild-type sequence) for the selected  
516 nanobodies and the wild-type nanobody. **f**, The AF ipLDDT and RS dG values of the selected  
517 nanobodies and the wild-type nanobody. **g**, The structure (predicted by AlphaFold-Multimer  
518 followed by Rosetta relaxation) of the receptor binding domain of the KP.3 spike protein (cyan)  
519 and the nanobody mutant Nb21 I77V-L59E-Q87A-R37Q (green). Side chains are shown for  
520 interface residues (within 4Å of the opposite chain). Mutant nanobody residues are in pink.  
521 (PyMol 3.1.3, Schrödinger, LLC.)

522 **Fig. 4 | Experimental validation of Virtual Lab nanobodies.** **a**, Histogram of expression levels  
523 across 96 nanobodies. Titer is expressed as milligrams of soluble, periplasmic nanobody per liter  
524 of culture. **b**, ELISA binding profiles of nanobodies to a panel of antigens. For each SARS-CoV-  
525 2 RBD protein and BSA (x-axis), individual spots represent the ELISA binding intensity (y-axis)  
526 of each of the 24 nanobodies. Unmutated nanobodies (H11-D4, Nb21, Ty1, and VHH-72) are  
527 shown in black and nanobodies exhibiting high non-specific binding are shown in shades of red  
528 (light pink, pink, and magenta). The Nb21 mutant (I77V-L59E-Q87A-R37Q) and the Ty1  
529 mutant (V32F-G59D-N54S-F32S) that bind to JN.1 are shown in green. Data are the mean of  
530 two measurements at a nanobody lysate dilution of 1:2. **c**, Comparison of ELISA binding of  
531 mutants and their unmutated sequences. Nb21 mutant I77V-L59E-Q87A-R37Q and Ty1 mutant  
532 V32F544 G59D-N54S-F32S are shown as filled circles and unmutated Nb21 and Ty1  
533 nanobodies are shown as open circles. ELISA binding intensity (y-axis) to Wuhan RBD, JN.1  
534 RBD, MERS CoV RBD, and BSA are shown in blue, red, light teal, and gray, respectively. Data  
535 shown are two biological replicates at a 12-point serial dilution of purified nanobody, fitted to a

536 4-parameter logistic curve. **d**, Location of mutant nanobody mutations. Models of the Nb21  
537 mutant I77V-L59E-Q87A-R37Q and the Ty1 mutant V32F-G59D-N54S-F32S, generated by the  
538 Virtual Lab using AlphaFold-Multimer, are shown in ribbons representation (blue), with the CDR  
539 loops shown in orange. Mutations introduced by the Virtual Lab are shown in pink and in red  
540 circles. Structure images were generated using ChimeraX<sup>51</sup>.

541 **Fig. 5 | Virtual Lab discussion analysis.** **a**, Excerpts from a Virtual Lab team meeting  
542 discussing the nanobody project specification. Each LLM agent addresses the agenda from its  
543 own perspective based on its title, expertise, goal, and role, leading to a comprehensive and  
544 interdisciplinary discussion of the agenda. **b**, The number of words (space-separated tokens)  
545 written by the Virtual Lab (human researcher and each LLM agent) across all phases of the  
546 nanobody design process. **c**, The number of words written by the Virtual Lab in the project  
547 specification phase. **d**, The number of words written by the Virtual Lab in ESM implementation.  
548 **e**, The number of words written by the Virtual Lab in the workflow design phase.

549 **Methods**

550 **Virtual Lab architecture**

551 The following sections describe the architecture of the Virtual Lab in more detail. All prompts  
552 are provided in Supplementary Note 6.

553 **Agents**

554 Each LLM agent in the Virtual Lab is defined with a prompt that specifies four key criteria.

555 1. **Title:** The name of the agent.

556        2. **Expertise:** The scientific expertise the agent has.

557        3. **Goal:** The ultimate goal of the agent in the context of the research project.

558        4. **Role:** The specific role that the agent will play in the research project.

559        The agents of the Virtual Lab are led by an agent called the Principal Investigator (PI). The PI  
560        agent has expertise in artificial intelligence for scientific research with a goal of maximizing the  
561        scientific impact of research and with the role of guiding the research project. The PI agent then  
562        automatically creates a set of scientist agents that are appropriate for the research project based  
563        on a short description of the project written by the human researcher. The PI defines these  
564        scientist agents by specifying each agent's title, expertise, goal, and role, using its own prompt as  
565        an example.

566        In addition to the PI and scientist agents, we find it useful to create an explicit critic agent to  
567        catch errors and oversights from the other agents and to give critical feedback on answers  
568        provided by the other agents<sup>52</sup>. Therefore, a Scientific Critic agent can be added to any team  
569        meeting or individual meeting to provide critical feedback to the other agents.

570        Meetings

571        Interactions in the Virtual Lab happen through meetings, which can either be *team meetings* with  
572        all the agents or *individual meetings* with a single agent (and optionally the critic agent). Both  
573        types of meetings share the following set of inputs that structure the meeting.

574        1. **Agenda:** (required) A description of the scientific topic to be discussed during the  
575        meeting.

578 3. **Agenda rules:** (optional) A set of rules that the agents must follow during the meeting.

579       4. **Summaries:** (optional) Agent-written summaries of previous meetings to provide  
580                   information about previous decisions.

585 The team and individual meetings differ in terms of the agents that participate in the meeting and  
586 the prompts that guide the flow of the meeting.

587 Team meeting

588 In team meetings, all the agents (PI agent, scientist agents, and Scientific Critic agent) participate  
589 in a conversation to address a broad research topic. First, the human researcher writes an agenda  
590 for the team meeting along with any applicable agenda questions and agenda rules. The team  
591 meeting then begins with an automatically constructed prompt then introduces the agents,  
592 agenda, agenda questions (if any), and agenda rules (if any) and describes the flow of the  
593 meeting, which involves multiple rounds of discussion. The PI agent is prompted to start the  
594 discussion by providing their initial thoughts and any guiding questions that they want to ask the  
595 team. Then, each scientist agent and the Scientific Critic agent are prompted one-by-one (in an  
596 order set by the human researcher) to provide their thoughts on the ongoing discussion given

597 everything that has been said by the other agents. At the end of a round of discussion, the PI  
598 agent synthesizes the points raised by each agent, makes decisions based on agent input, and asks  
599 follow-up questions to further the discussion. After N rounds of discussion (with N set by the  
600 human researcher), the PI agent summarizes the discussion for future meetings, provides a  
601 recommendation regarding the agenda, and answers the agenda questions (if any). The human  
602 researcher in the Virtual Lab can then read just this final response by the PI agent, thus  
603 benefiting from the extensive discussions among the LLM agents while only needing to read the  
604 final short response to understand the decisions that were made.

605 Individual meeting

606 In individual meetings, a single agent tackles a specific task that falls within their  
607 area of expertise, optionally with critical feedback provided by the Scientific Critic  
608 agent. To start an individual meeting, the human researcher in the Virtual Lab  
609 selects the agent that will participate. An automatically constructed prompt  
610 introduces the agenda, agenda questions (if any), and agenda rules (if any) and  
611 then immediately asks the agent for a response. If the individual meeting has zero  
612 rounds ( $N=0$ ), then the agent provides a response and the meeting ends. If the  
613 individual meeting includes one or more rounds ( $N \geq 1$ ), then in each round, the  
614 agent provides a response and then the Scientific Critic agent provides critical  
615 feedback to improve the agent's response. After these rounds, the selected agent  
616 responds one more time to provide the final, improved answer.

617 Parallel meetings

618 To improve the expected quality and comprehensiveness of answers for a given meeting, the  
619 same meeting (same agents, same prompts) can be run multiple times in parallel to produce  
620 multiple answers (due to the inherent randomness in responses generated by LLMs). Then, an  
621 individual meeting with the appropriate agent (i.e., the PI agent for team meetings or the relevant  
622 scientist agent for individual meetings) is run to merge the summaries of each of the parallel  
623 meetings into a single answer that incorporates the best elements from each of the parallel  
624 meetings. To boost creativity while producing a consistently high-quality answer, each of the  
625 parallel meetings is run with a higher “creative” temperature of 0.8 while the single merge  
626 meeting is run with a lower “consistent” temperature of 0.2, where temperature is the LLM  
627 parameter that controls the amount of randomness or uncertainty in the generation<sup>53,54</sup>. Parallel  
628 meetings are similar in nature to the method of majority voting from multiple LLM queries<sup>55</sup>, but  
629 the Virtual Lab’s parallel meetings use a more complex and flexible merging of answers via a  
630 meeting with an LLM agent.

631 Virtual Lab for nanobody design

632 We applied the Virtual Lab to nanobody design in five phases using GPT-4o (gpt-4o-2024-08-  
633 06) as the underlying LLM powering the agents.

634 Team selection

635 First, the Virtual Lab used an individual meeting with the PI agent (run five times in parallel  
636 followed by a merge meeting) to create a set of scientist agents for the project. The meeting  
637 agenda contained a background prompt about antibody/nanobody design for the spike protein of

638 the recent KP.3 SARS-CoV-2 variant and a request for the PI agent to select a team of three  
639 scientist agents for the project using the same agent structure (title, expertise, goal, role) as the  
640 PI's own definition. The PI decided to create an Immunologist, a Machine Learning Specialist,  
641 and a Computational Biologist.

642 Project specification

643 Next, in a team meeting, the full team of agents (PI, Immunologist, Machine Learning Specialist,  
644 Computational Biologist, and Scientific Critic) discussed some of the specifics of the project  
645 beyond the general background prompt. This meeting's agenda asked the agents to consider  
646 design choices such as whether to pursue antibodies or nanobodies and whether to modify  
647 existing antibodies/nanobodies or design new ones *de novo*. This team meeting was run in five  
648 parallel iterations followed by an individual merge meeting with the PI agent and the Scientific  
649 Critic to produce the best answer. Below are some of the PI's final merged answers to the agenda  
650 questions.

651 1. Will you design standard antibodies or nanobodies?

652 **Agent Answer:** Nanobodies

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

658        2. Will you design antibodies/nanobodies de novo or will you modify existing  
659                  antibodies/nanobodies (choose only one)?

660        **Agent Answer:** Modify existing nanobodies

661        **Agent Justification:** Modifying existing nanobodies allows us to leverage established  
662                  data, providing a quicker and more reliable path to developing broad-spectrum  
663                  candidates. This approach is supported by the availability of structural data and known  
664                  efficacy, which can accelerate the optimization process. This decision is consistently  
665                  recommended across all summaries, highlighting the efficiency and feasibility of building  
666                  on known frameworks.

667        3. If modifying existing antibodies/nanobodies, which precise antibodies/nanobodies will  
668                  you modify (please list 3-4)?

669        **Agent Answer:** Ty1, H11-D4, Nb21, VHH-72

670        **Agent Justification:** These candidates have demonstrated strong neutralizing capabilities  
671                  against SARS-CoV-2 and have available structural data, making them excellent  
672                  candidates for modification. Ty1 and H11-D4, for example, have shown efficacy in  
673                  neutralizing multiple variants, while Nb21 and VHH-72 offer unique binding properties  
674                  that can be optimized for broader cross-reactivity. Literature supports their efficacy, with  
675                  studies indicating their ability to bind conserved epitopes and maintain stability under  
676                  various conditions.

677        Given these decisions, the following phases proceeded with nanobody design by modifying the  
678                  four nanobodies suggested by the Virtual Lab (Ty1, H11-D4, Nb21, and VHH-72), which are

679 specific to the ancestral Wuhan spike protein, to increase their affinity to the spike protein of the  
680 KP.3 variant of SARS-CoV-2. Furthermore, the Virtual Lab suggested prioritizing “enhancing  
681 interactions with the receptor-binding domain of the spike protein by altering residues that  
682 contribute to binding affinity,” so the Virtual Lab subsequently focused on developing  
683 nanobodies that bind to the receptor binding domain (RBD) of the KP.3 spike protein.

684 Tools selection

685 After specifying the project direction, the Virtual Lab next needed to pick a set of computational  
686 tools to modify the selected nanobodies. To accomplish this, the Virtual Lab ran a team meeting  
687 asking the agents to list several machine learning and/or computational tools that could be used  
688 for nanobody design, with emphasis on pre-trained models for simplicity. Similar to the project  
689 selection meeting, this team meeting was run with five parallel iterations followed by a merge  
690 meeting with the PI and Scientific Critic. The agents decided to use ESM, AlphaFold-Multimer,  
691 and Rosetta as the components of its computational nanobody design workflow.

692 Tools implementation

693 With the project well-specified and a set of computational nanobody tools chosen, the Virtual  
694 Lab then worked on implementing these tools for nanobody design. For each tool, the Virtual  
695 Lab selected the most appropriate scientist agent via an individual meeting with the PI. Then for  
696 each tool, the Virtual Lab ran an individual meeting with the selected scientist agent and the  
697 Scientific Critic (five parallel meetings followed by a merge meeting run by the scientist agent)  
698 to implement the tool. These meetings included a set of agenda rules that specify how code  
699 should be written, e.g., with good documentation and without leaving functions undefined. These  
700 initial implementations contained small errors that needed correction, so the Virtual Lab then ran

701 a single follow-up individual meeting (no parallelization or Scientific Critic) with the scientist  
702 agent to automatically fix all the errors that arose.

703 ESM usage

704 The Machine Learning Specialist agent was responsible for writing a Python script to identify the  
705 most promising point mutations to a nanobody sequence based on the ESM log-likelihood ratio  
706 (LLR) of the mutant sequence compared to the input sequence. The agent wrote a 130-line  
707 Python script with three functions: a main function to run the script, a function to parse  
708 command-line arguments (e.g., the input nanobody sequence), and a function that uses a pre-  
709 trained ESM model to compute log-likelihood ratios for point mutations.

710 AlphaFold-Multimer usage

711 To use AlphaFold-Multimer, the Virtual Lab asked the Computational Biologist agent to write a  
712 Python script that processes a predicted nanobody-spike complex structure from AlphaFold-  
713 Multimer and outputs the interface pLDDT (ipLDDT), which is a measure of the confidence of  
714 the binding interface between the nanobody and the spike protein that has previously been shown  
715 to correlate with antibody-antigen binding affinity<sup>36</sup>. Computing the ipLDDT values across  
716 multiple proposed nanobodies requires reading a PDB file for each predicted nanobody-spike  
717 complex and writing as output a single CSV file with the ipLDDT from every complex. The  
718 Computational Biologist wrote a 144-line Python script with five functions: a main function to  
719 run the whole script, a function to check whether a PDB file contains a protein structure in the  
720 correct format, a function to identify the residues in the interface between the two proteins, a  
721 function to calculate the ipLDDT, and a function to run the ipLDDT calculation on every PDB  
722 file in a directory and save the results to a single CSV file.

723 Rosetta usage

724 The Computational Biologist was also responsible for using Rosetta to calculate nanobody-spike  
725 binding energies as a metric for measuring the quality of each mutated nanobody. Given a PDB  
726 file with a predicted nanobody-spike structure from AlphaFold-Multimer, the Computational  
727 Biologist was asked to write a RosettaScripts XML file to load the PDB file, calculate the  
728 binding energy, and save the binding energy to a Rosetta score file. Additionally, the agent was  
729 asked to write a Python script that loads all the score files in a directory and saves a CSV file  
730 with the binding energy of every nanobody-spike complex.

731 The Computational Biologist wrote a 30-line RosettaScripts XML file that first relaxes the  
732 nanobody-spike structure and then computes the binding energy (dG-separated in Rosetta  
733 terminology) of the interface using the REF15 scoring function. The Computational Biologist  
734 then wrote a 71-line Python script with two functions: a main function to run the whole script  
735 and a function to extract the binding energy score from a given Rosetta score file.

736 Workflow design

737 Finally, the Virtual Lab ran an individual meeting with the PI agent to design a workflow that  
738 uses ESM, AlphaFold-Multimer, and Rosetta to design nanobodies. For each of the four starting  
739 nanobody candidates, the PI agent decided to run ESM to evaluate all possible point mutations  
740 and then to select the top 20 mutations by ESM log-likelihood ratio. Each of these 20 mutant  
741 sequences would then be evaluated by both AlphaFold-Multimer and Rosetta. These 20  
742 sequences would then be ranked and the top five would be selected using the following weighted  
743 score designed by the PI agent:

744  $WS = 0.2 * (ESM\ LLR) + 0.5 * (AF\ ipLDDT) - 0.3 * (RS\ dG)$

745 where WS is the weighted score, ESM LLR is the ESM log-likelihood ratio between the mutated  
746 sequence and the input sequence, AF ipLDDT is the AlphaFold-Multimer ipLDDT binding  
747 interface confidence, and RS dG is the Rosetta dG-separated binding energy value. The PI  
748 correctly uses a negative weight for the Rosetta value since a more negative binding energy is  
749 better. The top five sequences according to WS then serve as the starting sequences for the next  
750 round of mutation, with 4 rounds of mutation in total depending on time constraints and  
751 improvements in the WS across rounds.

## 752 Nanobody design workflow

753 The Virtual Lab ran the nanobody design computational workflow to design improved nanobody  
754 candidates for the KP.3 variant of SARS-CoV-2. The workflow was run independently for each  
755 of the four nanobodies suggested by the agents: Ty1, H11-D4, Nb21, and VHH-72. Below, we  
756 describe the workflow in terms of a single starting nanobody for simplicity.

757 The Virtual Lab workflow began with round 0, which evaluated the wild-type nanobody  
758 sequence without introducing any mutations. ESM LLR was assigned to zero since the wild-type  
759 nanobody sequence was unmodified. Then, the Virtual Lab ran AlphaFold-Multimer (via  
760 LocalColabFold<sup>56</sup> version 1.5.5) on the nanobody sequence and the sequence of the receptor  
761 binding domain (RBD) of the KP.3 spike protein to produce a predicted structure of the complex.  
762 Next, the Virtual Lab computed the AF ipLDDT as a measure of confidence in the binding  
763 interface of the complex. Then, the Virtual Lab ran Rosetta (version 3.14) to relax the complex  
764 and compute the RS dG value as an estimate of the binding energy. Finally, the Virtual Lab  
765 computed the weighted score (WS) of the wild-type nanobody.

766 In round 1, the Virtual Lab ran ESM to calculate the ESM LLR of every possible single point  
767 mutation to the wild-type nanobody. The top 20 mutated sequences by ESM LLR were retained.  
768 For each of these 20 mutated sequences, AlphaFold-Multimer and Rosetta were applied in the  
769 same way as for the wild-type sequence. The Virtual Lab then computed the WS for each of the  
770 20 mutated sequences and selected the top five sequences for the next round. In rounds 2-4, the  
771 Virtual Lab applied the same procedure but now starting with five input sequences to the ESM  
772 LLR script, resulting in 100 top mutated sequences (20 proposed mutant sequences for each of  
773 the five input sequences). These sequences were analyzed by AlphaFold-Multimer and Rosetta,  
774 and the top five of these 100 sequences were selected at the end of each round by their WS.

775 After running all four rounds of mutation, the Virtual Lab needed to select the best mutated  
776 nanobody sequences across all four rounds for experimental validation. Doing so required using  
777 a slight variant of the weighted score (WS). In each round, the WS used the ESM LLR calculated  
778 as a ratio between the proposed mutant sequence and the input sequence for that round (i.e., an  
779 output sequence from the previous round), which differ by a single mutation. However, in order  
780 to fairly select the best sequences across different rounds with different numbers of mutations, an  
781 alternate ESM log-likelihood ratio, the ESM  $\text{LLR}^{\text{WT}}$ , was computed between each proposed  
782 mutant sequence (with one to four mutations) and the wild-type sequence. The Virtual Lab then  
783 scored all mutant nanobody sequences using the  $\text{WS}^{\text{WT}}$ , which is the weighted score calculated  
784 using the ESM  $\text{LLR}^{\text{WT}}$  in place of the ESM LLR. The top 23 mutant sequences were selected for  
785 experimental validation along with the wild-type sequence as a point of reference.

786 Nanobody experimental validation

787 Codon optimized DNA sequences for the SARS-CoV-2 spike RBDs JN.1, KP.3, KP.2.3 and  
788 BA.2<sup>13,57</sup>, modified to include a N-terminal signal peptide (MFVFLVLLPLVSSQ), a C-terminal  
789 glycine/serine linker and 6x his tag, and a stop codon, were synthesized and cloned into pTwist-  
790 CMV-BetaGlobin (Twist Biosciences). For MERS-CoV RBD, the codon optimized DNA  
791 sequence for the RBD was modified to include an N-terminal signal peptide  
792 (MYRMQLLSCIALSLALVTNS), C-terminal glycine/serine linker, 8x his tag, AviTag  
793 sequences, and a stop codon. RBDs were transiently expressed in Expi293 cells (Thermo Fisher  
794 Scientific, not authenticated or tested for mycoplasma contamination), and purified in parallel<sup>58</sup>  
795 by Ni-NTA Excel affinity chromatography followed by desalting into PBS and concentration.  
796 The purification of Wuhan SARS-CoV-2 RBD has been described previously<sup>58</sup>. Codon  
797 optimized DNA sequences for nanobodies, modified to include an N-terminal pelB signal  
798 peptide (MKYLLPTAAAGLLLLAAQPAMA), a C-terminal 6x his tag and a stop codon, were  
799 synthesized and cloned into pET-29b(+) (Twist Biosciences). Nanobodies were expressed in 96-  
800 well and 24-well format in auto-induction media<sup>59</sup>, and periplasmic fractions from 4 mL of cell  
801 culture pellets were released by mild lysis in 400 uL PBS, following methods as described<sup>60</sup>.  
802 Titers of soluble nanobody were estimated from periplasmic fractions by SDS-PAGE  
803 densitometry analysis of nanobody bands using a BSA standard curve. Selected nanobodies were  
804 scaled up at 100 mL in shake flasks and purified from periplasmic fractions by Ni-NTA  
805 chromatography followed by desalting into PBS and concentration.

806 Multiplexed ELISA measurements were performed as generally described<sup>61</sup>. Array patterns  
807 were printed using a sciFLEXARRAYER S12. Each RBD and BSA (negative control) spot was  
808 printed in duplicate, using up to three 200-250 uL drops for each spot, at a source concentration

809 of 50 µg/mL. Unpurified lysates or purified nanobodies were diluted in PBS-T (5% skim milk in  
810 PBS + 0.05% Tween-20), and RBD-bound nanobodies were recognized by anti-Alpaca IgG  
811 VHH secondary antibodies (Jackson ImmunoResearch, 128-065-230 (for H11-D4, Nb21, and  
812 VHH-72 series) and 128-065-232 (for Ty1 series)) at 1:10000 dilution in PBS-T.

## 813 Data availability

814 The computational results of the nanobody design pipeline as well as the experimental ELISA  
815 binding data are available on Zenodo at <https://doi.org/10.5281/zenodo.15331308> (ref.<sup>62</sup>).

## 816 Code availability

817 Code for the Virtual Lab, full discussions by the agents, and computational scores for the  
818 designed nanobodies are available on GitHub at [https://github.com/zou-group/virtual\\_lab](https://github.com/zou-group/virtual_lab) and on  
819 Zenodo at <https://doi.org/10.5281/zenodo.15320491> (ref.<sup>63</sup>).

## 820 Methods References

821 52. Yuksekgonul, M. *et al.* Optimizing generative AI by backpropagating language model  
822 feedback. *Nature* **639**, 609–616 (2025).

823 53. Peeperekorn, M., Kouwenhoven, T., Brown, D. & Jordanous, A. Is Temperature the  
824 Creativity Parameter of Large Language Models? Preprint at  
825 <https://doi.org/10.48550/arXiv.2405.00492> (2024).

826 54. Chen, H. & Ding, N. Probing the “Creativity” of Large Language Models: Can models  
827 produce divergent semantic association? in *Findings of the Association for Computational*

- 828      *Linguistics: EMNLP 2023* (eds. Bouamor, H., Pino, J. & Bali, K.) 12881–12888 (Association  
829      for Computational Linguistics, Singapore, 2023). doi:10.18653/v1/2023.findings-emnlp.858.
- 830      55. Chen, L. *et al.* Are More LLM Calls All You Need? Towards the Scaling Properties of  
831      Compound AI Systems. in *The Thirty-eighth Annual Conference on Neural Information  
832      Processing Systems* (2024).
- 833      56. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat. Methods* **19**,  
834      679–682 (2022).
- 835      57. Kumar, S., Karuppanan, K. & Subramaniam, G. Omicron (BA.1) and sub-variants  
836      (BA.1.1, BA.2, and BA.3) of SARS-CoV-2 spike infectivity and pathogenicity: A  
837      comparative sequence and structural-based computational assessment. *J. Med. Virol.* **94**,  
838      4780–4791 (2022).
- 839      58. Puccinelli, R. R. *et al.* Open-source milligram-scale, four channel, automated protein  
840      purification system. *PLOS ONE* **19**, e0297879 (2024).
- 841      59. Saez, N. J. & Vincentelli, R. High-Throughput Expression Screening and Purification of  
842      Recombinant Proteins in *E. coli*. in *Structural Genomics: General Applications* (ed. Chen, Y.  
843      W.) 33–53 (Humana Press, Totowa, NJ, 2014). doi:10.1007/978-1-62703-691-7\_3.
- 844      60. Pardon, E. *et al.* A general protocol for the generation of Nanobodies for structural  
845      biology. *Nat. Protoc.* **9**, 674–693 (2014).
- 846      61. Byrum, J. R. *et al.* MultiSero: An Open-Source Multiplex-ELISA Platform for Measuring  
847      Antibody Responses to Infection. *Pathogens* **12**, 671 (2023).
- 848      62. Swanson, K., Wu, W., Bulaong, N., Pak, J. & Zou, J. Virtual Lab Data. Zenodo  
849      <https://doi.org/10.5281/zenodo.15331309> (2025).
- 850      63. Swanson, K. Virtual Lab Code. Zenodo <https://doi.org/10.5281/zenodo.15320492> (2025).

851 **Acknowledgments**

852 We thank E. Simon and J. Silberg for their discussions of this work. K.S. acknowledges support  
853 from the Knight-Hennessy Scholarship and the Stanford Bio-X Fellowship. J.Z. is supported by  
854 funding from the Chan Zuckerberg Biohub - San Francisco.

855 **Author contributions**

856 K.S. built the Virtual Lab framework and applied the Virtual Lab to create and run the  
857 computational nanobody design pipeline. W.W, N.L.B, and J.E.P. conducted the nanobody  
858 validation experiments. J.E.P. and J.Z. supervised the work. All authors contributed to the  
859 manuscript.

860 **Competing interests**

861 The authors declare no competing interests.

862 **Additional information**

863 Supplementary Information is available for this paper with Supplementary Notes 1-6.  
864 Correspondence and requests for materials should be addressed to John E. Pak or James Zou.  
865 Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints).

866 Extended data figure legends

867 **Extended Data Table 1 | Nanobody score analysis.** The scores of each wild-type nanobody and  
868 examples of the mutant nanobodies that were selected for experimental validation. ESM LLR<sup>WT</sup>:  
869 ESM log-likelihood ratio between the mutant nanobody sequence and the wild-type sequence.  
870 AF ipLDDT: AlphaFold-Multimer interface pLDDT for the nanobody-spike complex. RS dG:  
871 Rosetta dG-separated binding energy value. WS<sup>WT</sup>: Weighted score combining ESM LLR<sup>WT</sup>, AF  
872 ipLDDT, and RS dG.

873 **Extended Data Fig. 1 | Virtual Lab parallel meetings.** The workflow for parallel meetings in  
874 the Virtual Lab. A set of meetings (team or individual) is run with the same agenda and agents  
875 but with different randomness in the LLM underlying the agents (with a high LLM temperature  
876 to encourage creativity across meetings). The answer from each parallel meeting is then provided  
877 to an agent in an individual meeting (with a low LLM temperature for consistency), and this  
878 agent is asked to merge the best components of the answers from each parallel meeting into a  
879 single optimal answer.

880 **Extended Data Fig. 2 | Ty1 nanobody analysis.** **a-c**, Evolution of mutant nanobody scores  
881 across four rounds of optimization. **a**, The distribution of ESM LLR values for proposed Ty1  
882 mutant nanobodies across each round of optimization, with ESM LLR values computed relative  
883 to the input nanobody sequence from the previous round. Shown are the ESM LLR values of the  
884 top 20 proposed mutant nanobodies per input nanobody. **b**, The AF ipLDDT and the RS dG of  
885 the top five proposed nanobodies, selected by WS, at the end of each round of optimization. **c**,  
886 The distribution of WS values of the top five proposed nanobodies at the end of each round of  
887 optimization. **d-f**, Analysis of the final set of 23 mutant nanobodies selected across all rounds of

optimization. **d**, The distribution of ESM LLR<sup>WT</sup> values (ESM LLR of the mutant sequence compared to the wild-type sequence) for the selected nanobodies and the wild-type nanobody. **e**, The AF ipLDDT and RS dG values of the selected nanobodies and the wild-type nanobody. **f**, The structure (predicted by AlphaFold-Multimer followed by Rosetta relaxation) of the receptor binding domain of the KP.3 spike protein (cyan) and the nanobody mutant Ty1 V32F-G59D-N54S-F32S (green). Side chains are shown for interface residues (within 4Å of the opposite chain). Mutant nanobody residues are in pink. (PyMol 3.1.3, Schrödinger, LLC.)

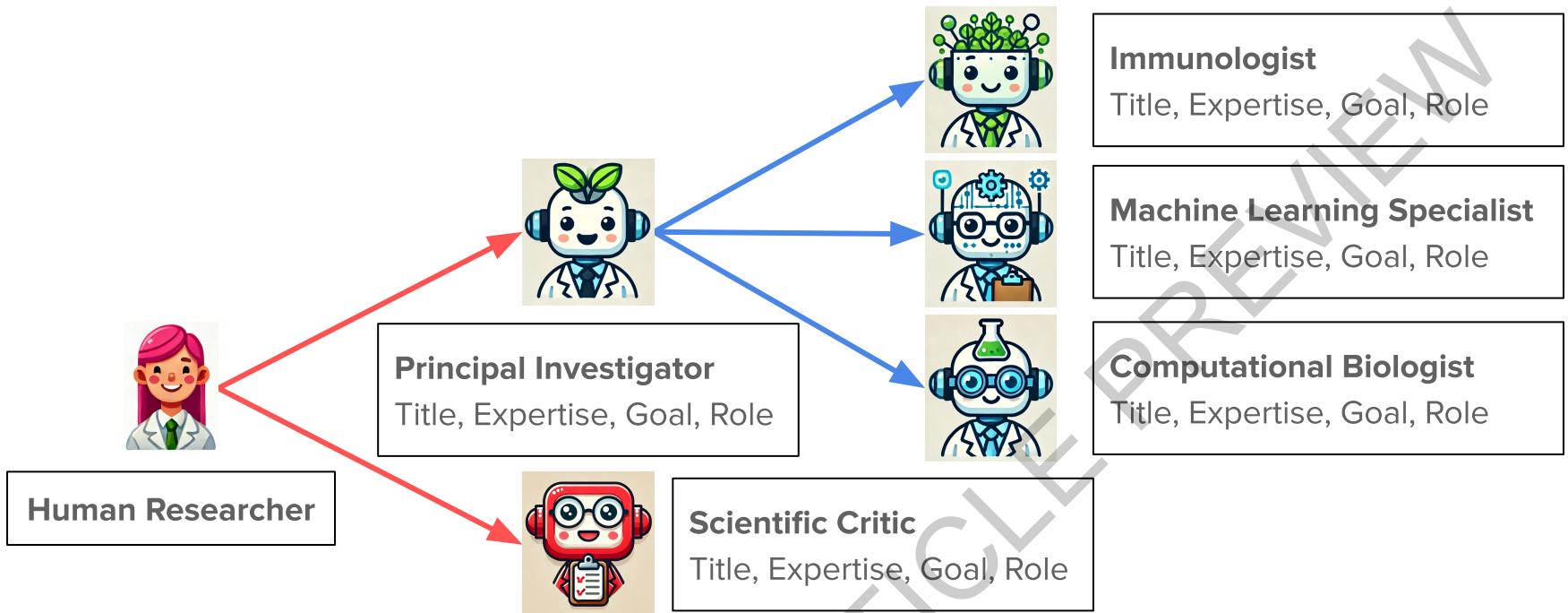
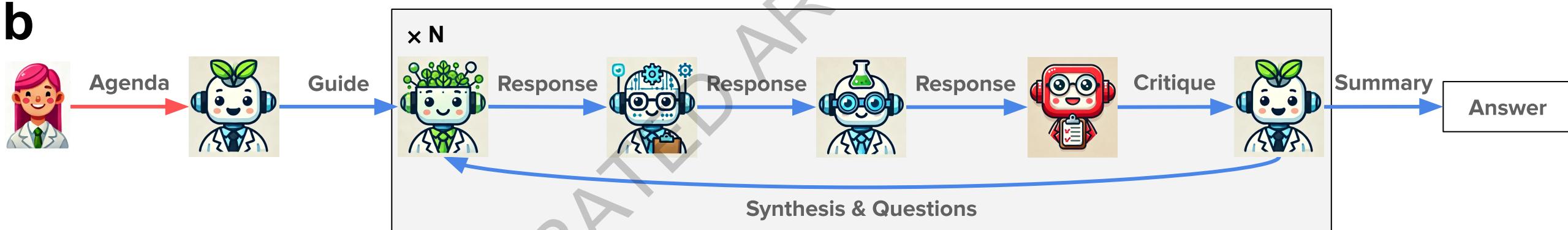
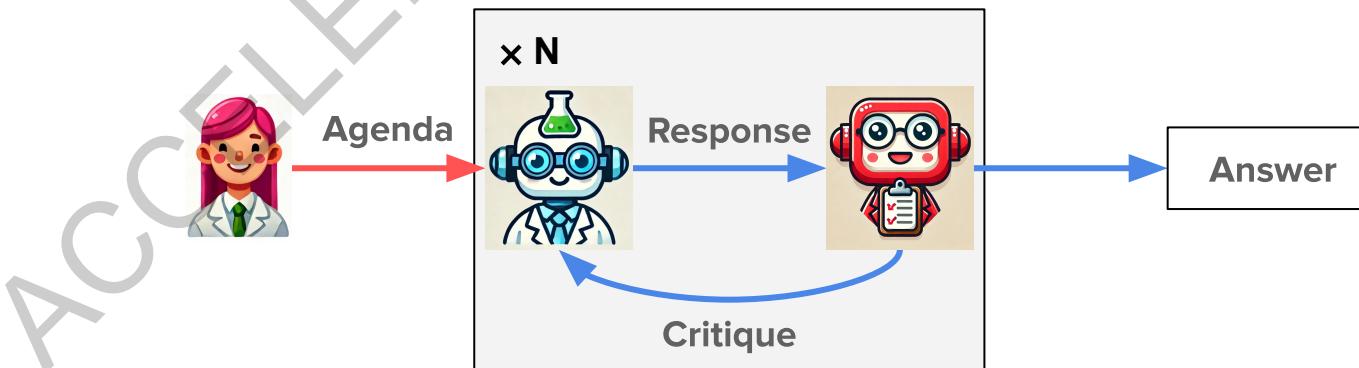
**Extended Data Fig. 3 | H11-D4 nanobody analysis.** **a-c**, Evolution of mutant nanobody scores across four rounds of optimization. **a**, The distribution of ESM LLR values for proposed H11-D4 mutant nanobodies across each round of optimization, with ESM LLR values computed relative to the input nanobody sequence from the previous round. Shown are the ESM LLR values of the top 20 proposed mutant nanobodies per input nanobody. **b**, The AF ipLDDT and the RS dG of the top five proposed nanobodies, selected by WS, at the end of each round of optimization. **c**, The distribution of WS values of the top five proposed nanobodies at the end of each round of optimization. **d-f**, Analysis of the final set of 23 mutant nanobodies selected across all rounds of optimization. **d**, The distribution of ESM LLR<sup>WT</sup> values (ESM LLR of the mutant sequence compared to the wild-type sequence) for the selected nanobodies and the wild-type nanobody. **e**, The AF ipLDDT and RS dG values of the selected nanobodies and the wild-type nanobody. **f**, The structure (predicted by AlphaFold-Multimer followed by Rosetta relaxation) of the receptor binding domain of the KP.3 spike protein (cyan) and the nanobody mutant H11-D4 A14P-Y88V-K74T-R27L (green). Side chains are shown for interface residues (within 4Å of the opposite chain). Mutant nanobody residues are in pink. (PyMol 3.1.3, Schrödinger, LLC.)

910 **Extended Data Fig. 4 | VHH-72 nanobody analysis.** **a-c**, Evolution of mutant nanobody scores  
911 across four rounds of optimization. **a**, The distribution of ESM LLR values for proposed VHH-  
912 72 mutant nanobodies across each round of optimization, with ESM LLR values computed  
913 relative to the input nanobody sequence from the previous round. Shown are the ESM LLR  
914 values of the top 20 proposed mutant nanobodies per input nanobody. **b**, The AF ipLDDT and  
915 the RS dG of the top five proposed nanobodies, selected by WS, at the end of each round of  
916 optimization. **c**, The distribution of WS values of the top five proposed nanobodies at the end of  
917 each round of optimization. **d-f**, Analysis of the final set of 23 mutant nanobodies selected across  
918 all rounds of optimization. **d**, The distribution of ESM LLR<sup>WT</sup> values (ESM LLR of the mutant  
919 sequence compared to the wild-type sequence) for the selected nanobodies and the wild-type  
920 nanobody. **e**, The AF ipLDDT and RS dG values of the selected nanobodies and the wild-type  
921 nanobody. **f**, The structure (predicted by AlphaFold-Multimer followed by Rosetta relaxation) of  
922 the receptor binding domain of the KP.3 spike protein (cyan) and the nanobody mutant VHH-72  
923 R27Y-E31D-F37V-D89E (green). Side chains are shown for interface residues (within 4Å of the  
924 opposite chain). Mutant nanobody residues are in pink. (PyMol 3.1.3, Schrödinger, LLC.)

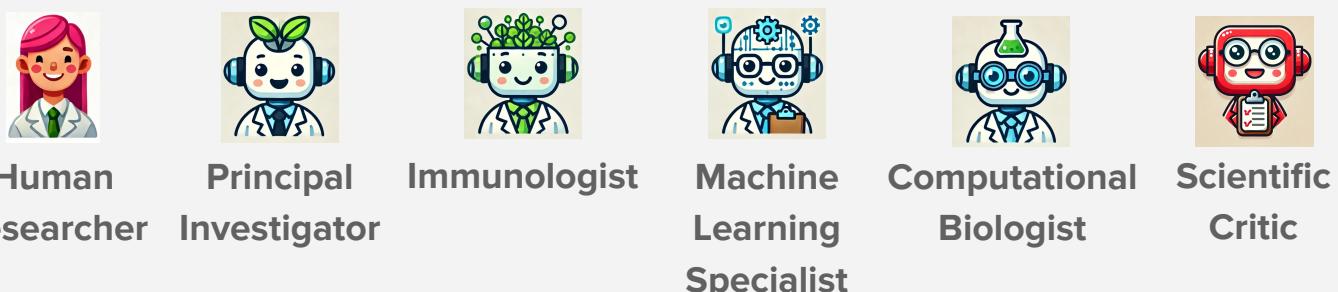
925 **Extended Data Fig. 5 | Workflow for nanobody experimental validation.** The four categories  
926 of experiments (nanobody expression, SARS-CoV-2 spike RBD expression, antigen array  
927 printing, and multiplexed ELISA) are enclosed in boxes. The ribbons representation of a  
928 nanobody (blue) and the RBD (purple) were rendered with ChimeraX<sup>51</sup> from PDB accession  
929 numbers 6XZN and 6M0J, respectively. Unique RBD and control proteins of the array are shown  
930 as colored spots with fiducial markers shown as black spots. Portions of this figure were created  
931 using BioRender.com.

932 **Extended Data Fig. 6 | Nanobody expression.** Periplasmic extracts containing soluble  
933 nanobody were separated by reducing SDS-PAGE and stained with Coomassie blue. An equal  
934 volume of periplasmic extract (8.3 uL) was loaded for each sample. Identifiers for each  
935 nanobody (A1 to H12) are shown, with the 4 unmutated parental nanobodies highlighted in  
936 yellow and the 92 Virtual Lab designs unhighlighted. The expected molecular weight for the  
937 nanobodies (~15 kDa) is enclosed in a red box. Uncropped images of samples analyzed once by  
938 SDS-PAGE are shown.

939 **Extended Data Fig. 7 | Virtual Lab additional discussion analysis.** **a**, The number of words  
940 (space-separated tokens) written by the Virtual Lab (human researcher and each LLM agent) in  
941 the tools selection phase. **b**, The number of words written by the Virtual Lab in AlphaFold  
942 implementation. **c**, The number of words written by the Virtual Lab in Rosetta implementation.  
943 **d**, The number of words written by the Virtual Lab in the team selection phase. **e**, The number of  
944 words written by the Virtual Lab in implementation agent selection.

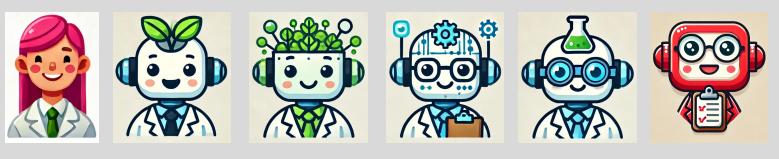
**a****b****c**

### (a) Phase 1: Team selection



### (b) Phase 2: Project specification

#### Team Meeting

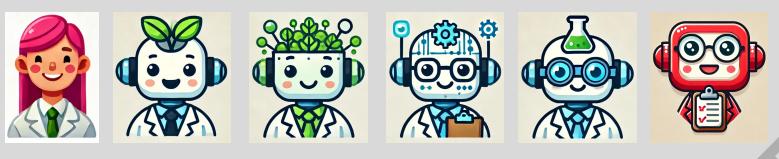


#### Summary

Modify nanobodies Ty1, H11-D4, Nb21, and VHH-72 to improve binding to KP.3.

### (c) Phase 3: Tools selection

#### Team Meeting



#### Summary

ESM, AlphaFold-Multimer, and Rosetta to design improved nanobodies.

### (d) Phase 4: Tools implementation

#### ESM

QVQLVE... → DVQLVE...  
ESM LLR = 3.65

#### Individual Meetings



#### Summary

Python script to compute ESM log-likelihood ratios for single point mutations.

#### AlphaFold-Multimer

DVQLVE... → DVQLVE...  
AF ipLDDT = 76.52

#### Individual Meetings



#### Summary

Python script to extract AlphaFold-Multimer interface pLDDT.

#### Rosetta

RS dG = -37.91

#### Individual Meetings



#### Summary

Python and XML scripts to compute binding energies with Rosetta.

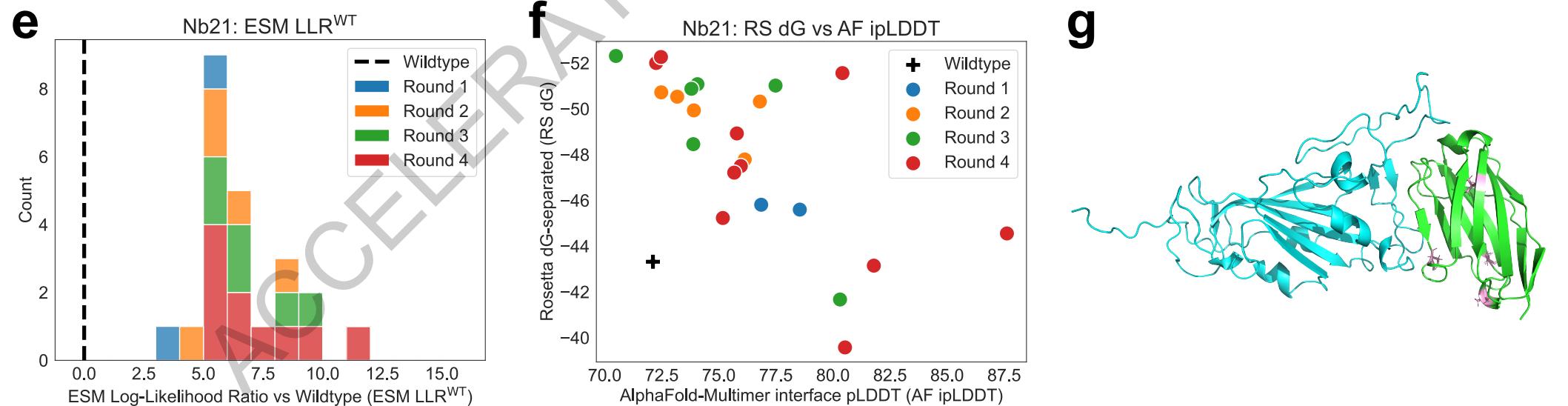
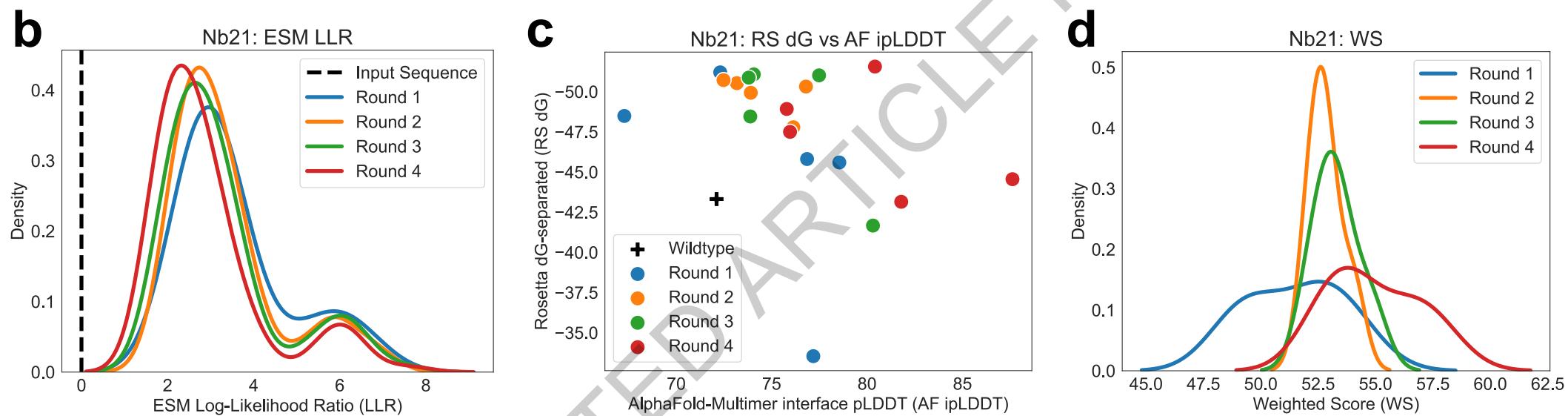
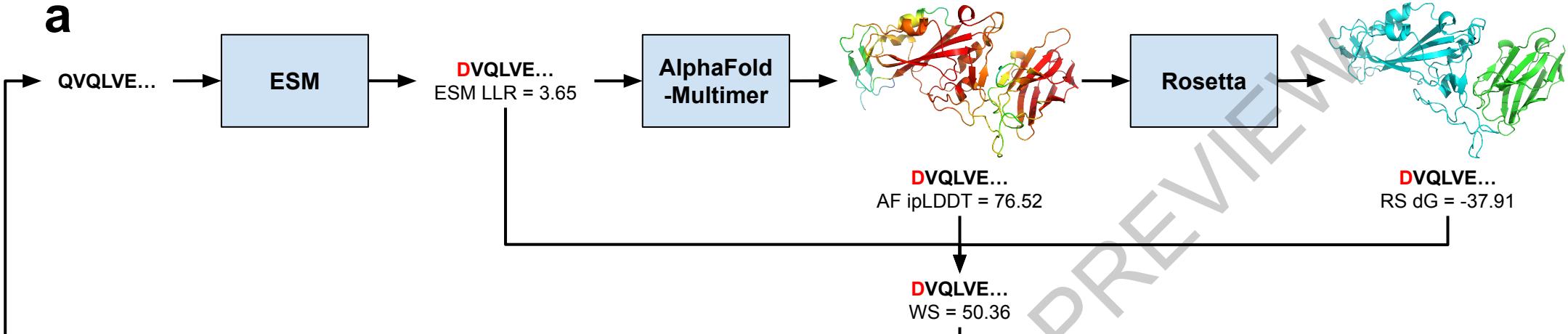
### (e) Phase 5: Workflow design

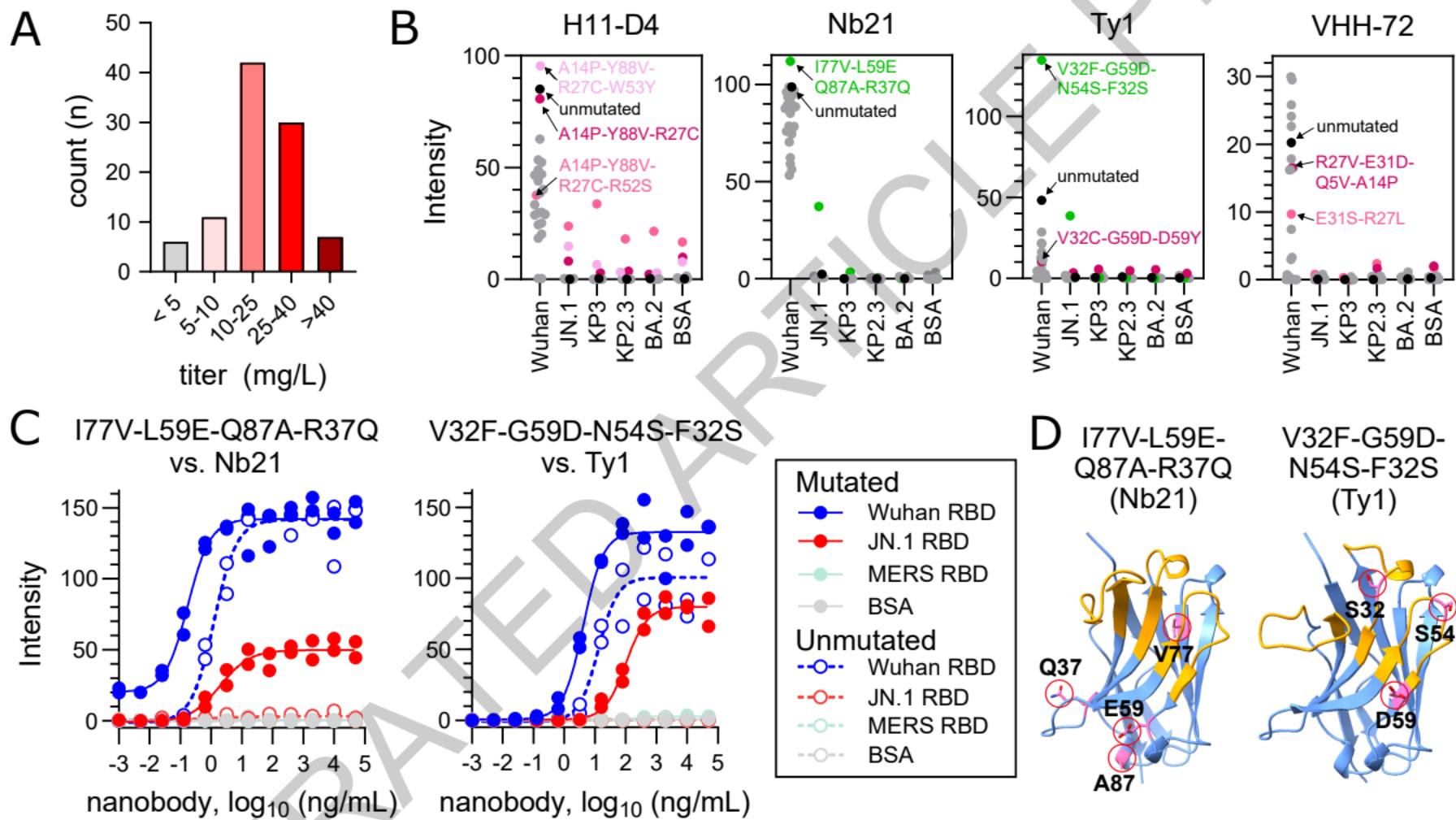
#### Individual Meeting



#### Summary

Tool usage, number of nanobodies to design, and weighted score formula.





**a**

**Principal Investigator:** I want to emphasize the significance of our objective: **to develop effective antibodies or nanobodies** that can target the newest variant of the **SARS-CoV-2 spike protein**, while also maintaining activity against other circulating minor variants and past variants.



**Immunologist:** I recommend **modifying existing nanobodies**. We have a wealth of data on several broadly neutralizing nanobodies that have shown efficacy against diverse SARS-CoV-2 variants. This approach **reduces the risk and time associated with de novo design**, leveraging existing structural and binding data for rapid development.



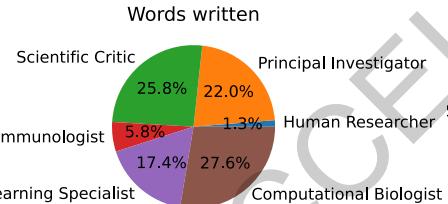
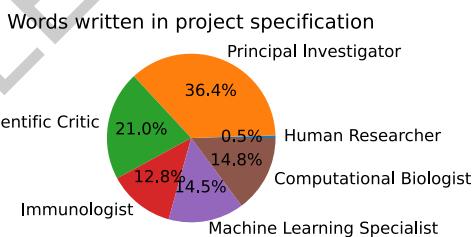
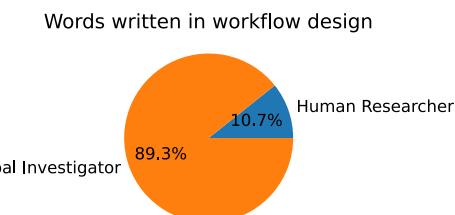
**Machine Learning Specialist:** I agree with the emphasis on nanobodies... Their smaller size also means fewer degrees of freedom, which can **simplify machine learning model complexity**... For modifying existing nanobodies, we can **utilize deep learning models** trained on structural and sequence data to predict changes that enhance binding affinity and cross-reactivity.

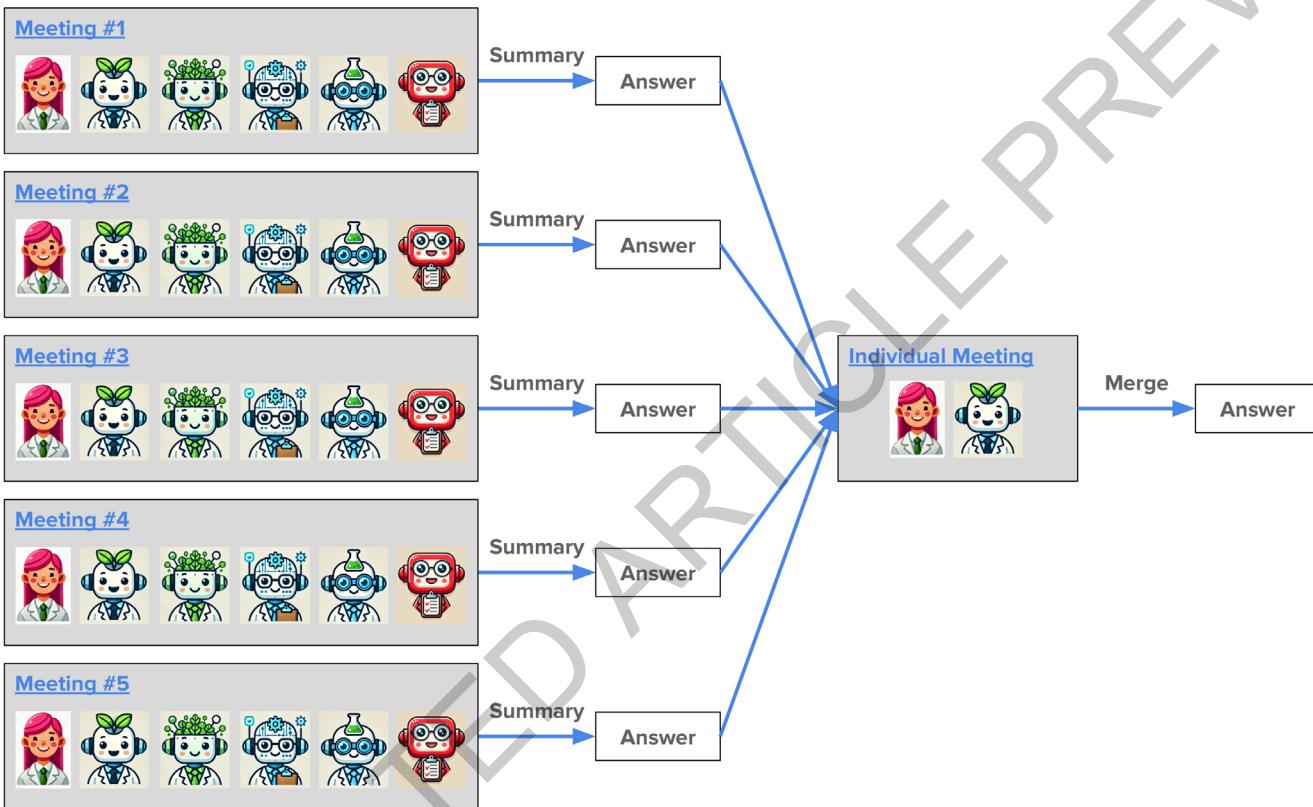


**Computational Biologist:** The machine learning models proposed are excellent, and coupling them with **molecular docking and simulations** can create a robust pipeline. For instance, once a machine learning model suggests modifications, simulations can be used to validate and refine these predictions by assessing the **energetic favorability and structural compatibility** with the spike protein.

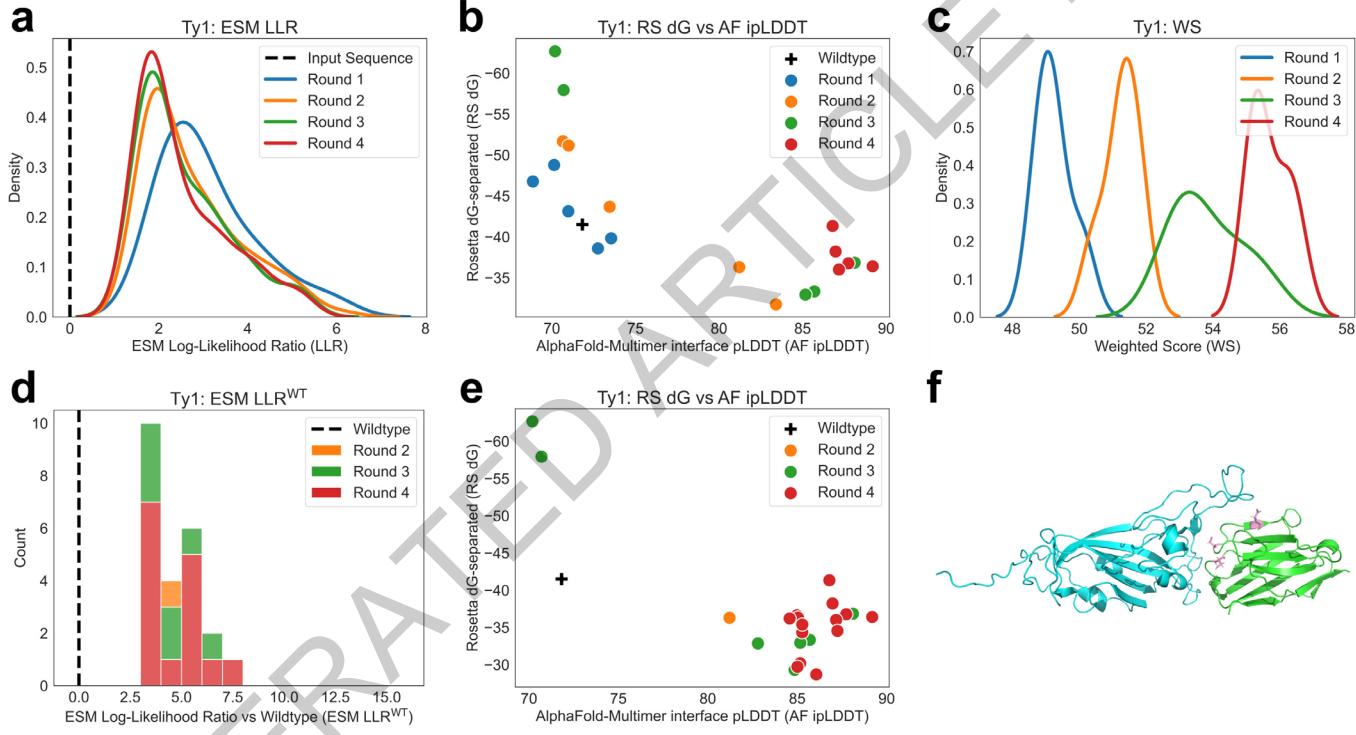


**Scientific Critic:** We must remain **cautious of over-reliance on computational predictions** without thorough cross-validation. Machine learning models can suffer from **bias**, particularly if trained on **limited or non-representative datasets**.

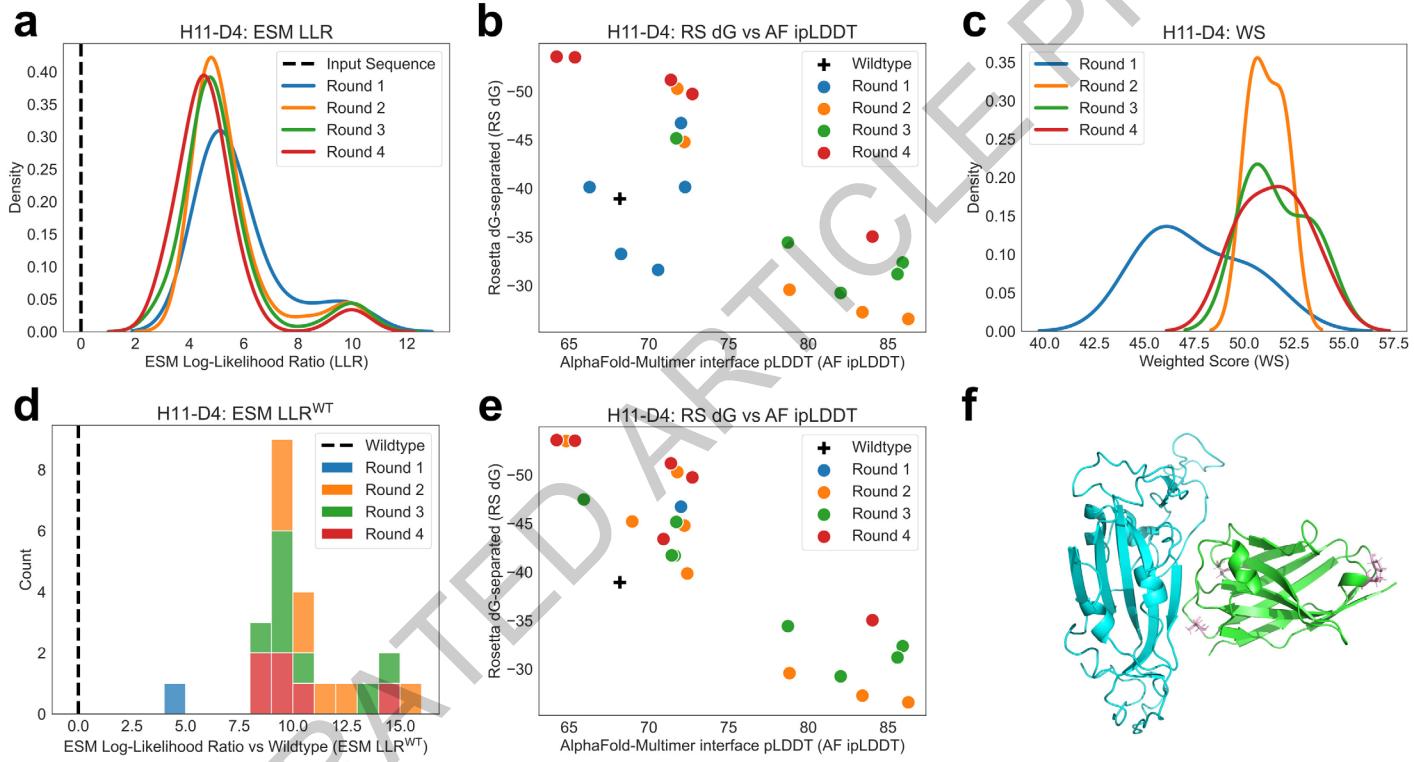
**b****c****d****e**



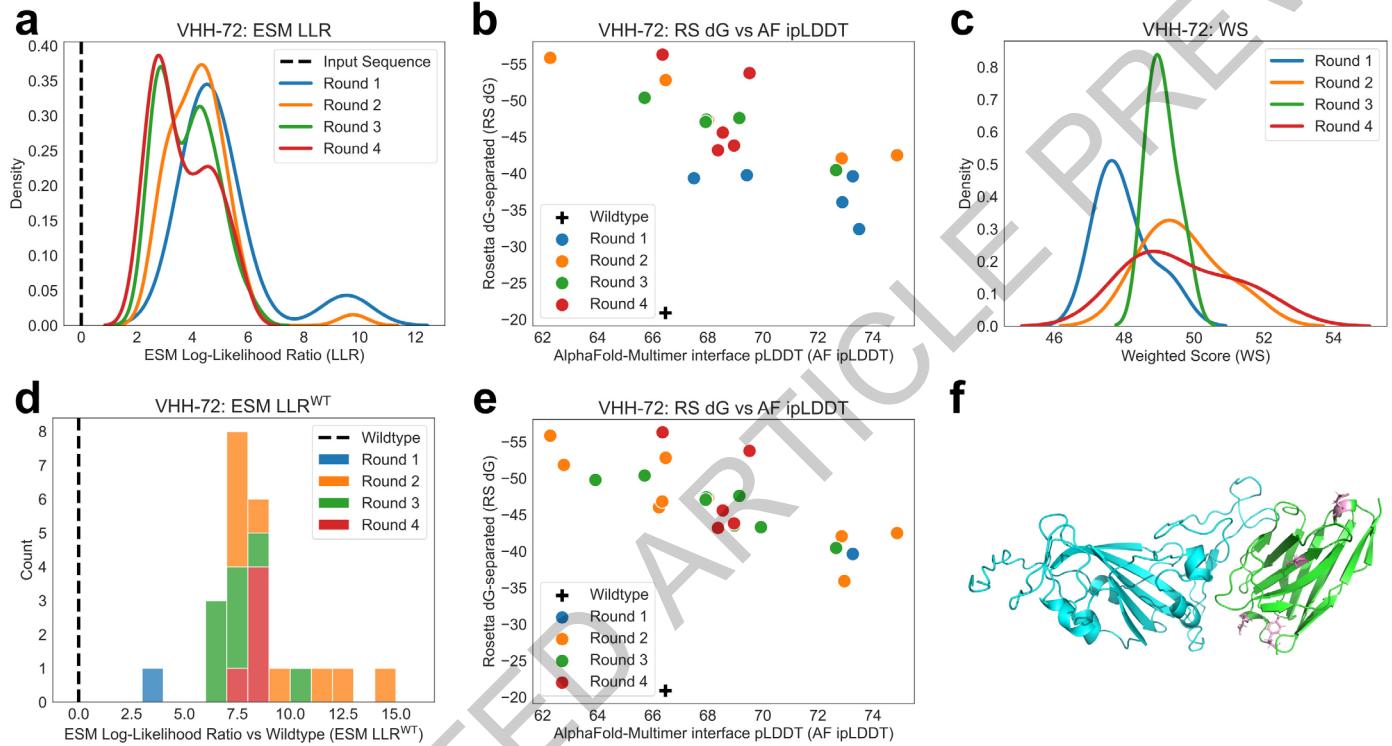
Extended Data Fig. 1



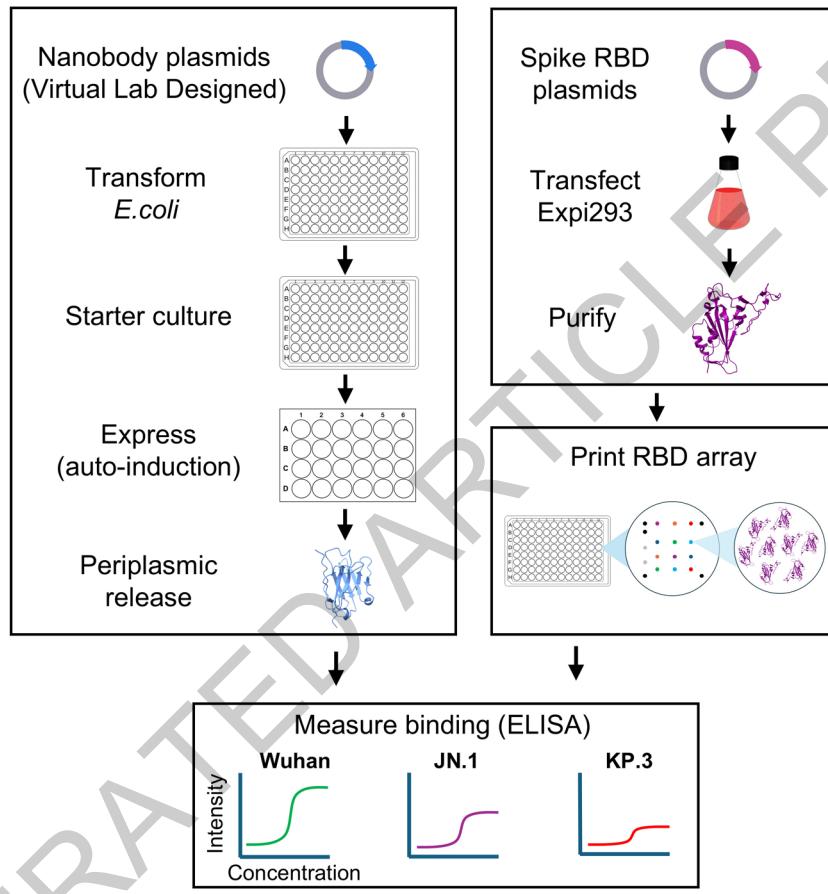
Extended Data Fig. 2



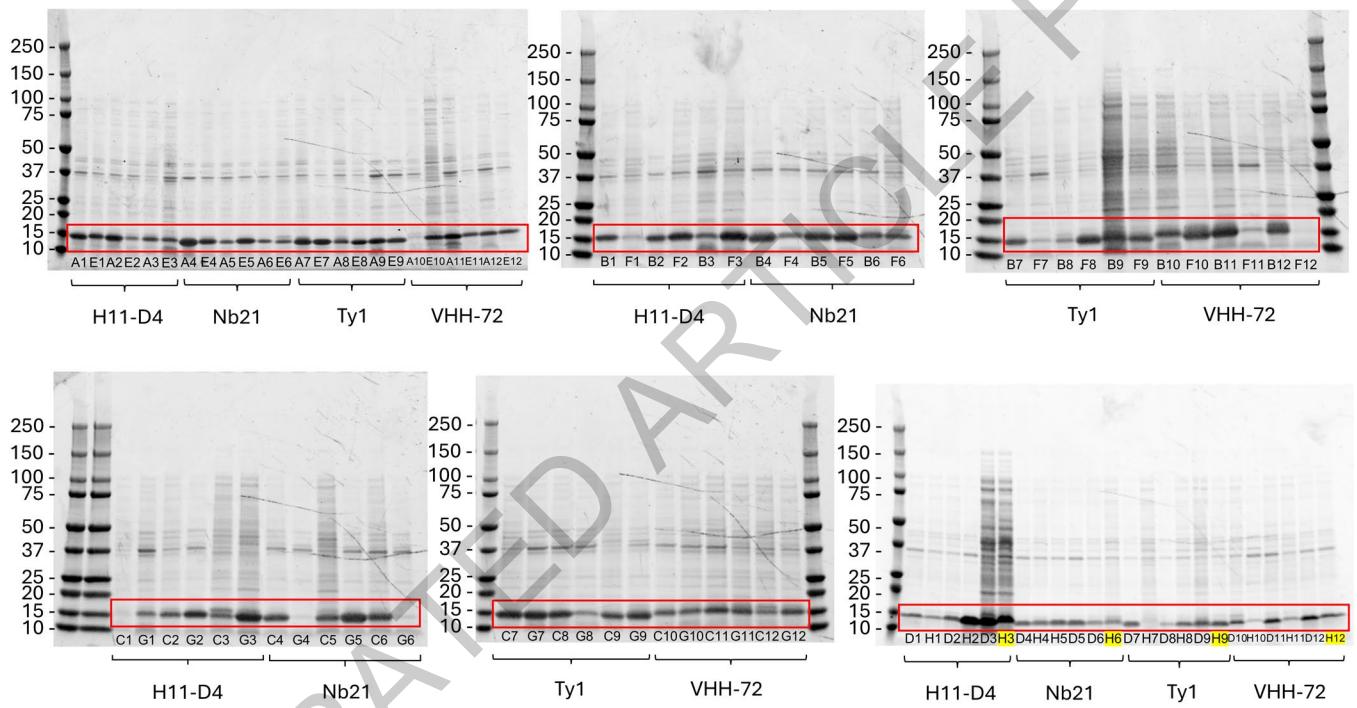
**Extended Data Fig. 3**



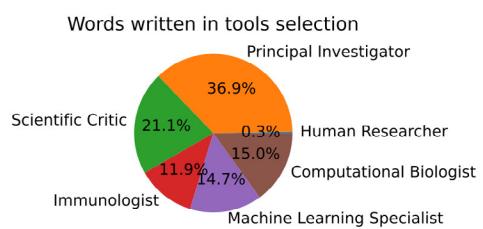
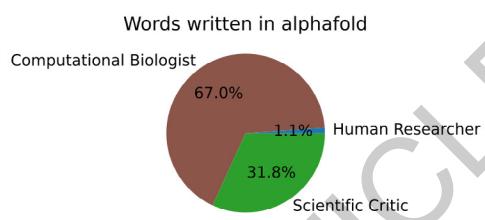
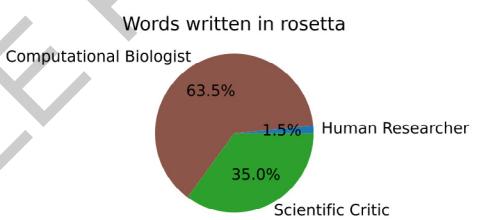
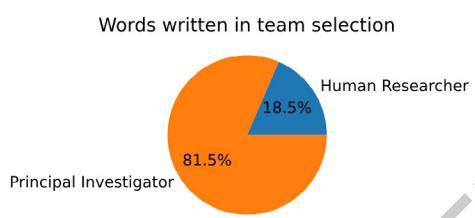
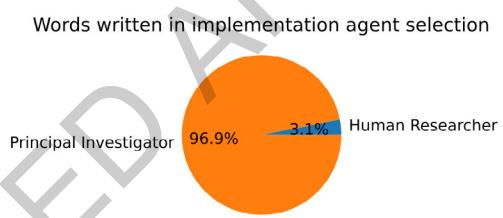
**Extended Data Fig. 4**



Extended Data Fig. 5



**Extended Data Fig. 6**

**a****b****c****d****e**

**Extended Data Fig. 7**

Name	ESM LLR <sup>WT</sup>	AF ipLDDT	RS dG	WS <sup>WT</sup>
Ty1	0.00	71.83	-41.51	48.36
Ty1 V32F-G59D-N54S-F32S	3.51	86.06	-28.69	52.34
H11-D4	0.00	68.18	-38.93	45.77
H11-D4 A14P-Y88V-K74T-R27L	10.67	84.02	-35.04	54.66
Nb21	0.00	72.11	-43.32	49.05
Nb21 I77V-L59E-Q87A-R37Q	7.47	80.41	-51.56	57.17
VHH-72	0.00	66.46	-20.90	39.50
VHH-72 R27Y-E31D-F37V-D89E	8.82	69.51	-53.76	52.65

**Extended Data Table 1**

## Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
  - Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted
  - Give  $P$  values as exact values whenever suitable.*
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

## Data collection

Data was collected using the Python package we created called virtual-lab version 1.1.0, available on GitHub at <https://github.com/zou-group/virtual-lab> and on Zenodo at <https://doi.org/10.5281/zenodo.15320491>. The virtual-lab packaged used ESM2, AlphaFold-Multimer via LocalColabFold 1.5.5, and Rosetta 3.14.

## Data analysis

Data was analyzed using the Python package we created called virtual-lab version 1.1.0, available on GitHub at <https://github.com/zou-group/virtual-lab> and on Zenodo at <https://doi.org/10.5281/zenodo.15320491>. The virtual-lab packaged used ESM2, AlphaFold-Multimer via LocalColabFold 1.5.5, and Rosetta 3.14.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

## Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Text data generated in the study is available at on GitHub at <https://github.com/zou-group/virtual-lab> and on Zenodo at <https://doi.org/10.5281/zenodo.15320491>. The computational results of the nanobody design pipeline as well as the experimental ELISA binding data are available on Zenodo at <https://doi.org/10.5281/zenodo.15331308>.

## Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

Our study does not involve human participants or their data, so this is not applicable.

Reporting on race, ethnicity, or other socially relevant groupings

Our study does not involve human participants or their data, so this is not applicable.

Population characteristics

Our study does not involve human participants or their data, so this is not applicable.

Recruitment

Our study does not involve human participants or their data, so this is not applicable.

Ethics oversight

Our study does not involve human participants or their data, so this is not applicable.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences

Behavioural & social sciences

Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

We tested four wildtype nanobodies and 92 nanobody mutants with an equal number of nanobody mutants (23) for each of the four wildtype nanobodies (96 nanobodies total). Each nanobody was tested using an ELISA binding assay in duplicate. We selected 96 nanobodies in order to fit on a 96-well plate. Since we are not performing statistical comparisons, this sample size was determined to be sufficient to have a high chance of identifying at least one mutant nanobody that was effective against a recent variant of SARS-CoV-2.

Data exclusions

No data was excluded from the analyses.

Replication

ELISA binding assays were run in duplicates and the mean value was reported.

Randomization

Randomization is not applicable to this study because every nanobody was tested in the same manner.

Blinding

Blinding is not applicable to this study since every nanobody was tested in the same manner.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern
<input checked="" type="checkbox"/>	<input type="checkbox"/> Plants

## Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Antibodies

### Antibodies used

Codon optimized DNA sequences for the nanobodies designed in this study were modified to include a N-terminal pelB signal peptide (MKYLLPTAAAGLLLAAQPAMA), a C-terminal 6x his tag, and a stop codon, and they were synthesized and cloned into pET-29b(+) (Twist Biosciences). anti-Alpaca IgG VH1 secondary antibodies (Jackson ImmunoResearch, 128-065-230 (for H11-D4, Nb21, and 822 VHH-72 series) and 128-065-232 (for Ty1 series)) were used for ELISA at 1:10000 dilution in PBS-T.

### Validation

Nanobodies were expressed in 96-well and 24-well format in auto-induction media, and periplasmic fractions from 4 mL of cell culture pellet were released by mild lysis in 400 uL PBS, following methods as described in Pardon, E. et al. A general protocol for the generation of Nanobodies for structural biology. Nat. Protoc. 9, 674–693 (2014). Periplasmic extracts containing soluble nanobody were separated by reducing SDS-PAGE.

## Eukaryotic cell lines

### Policy information about [cell lines and Sex and Gender in Research](#)

#### Cell line source(s)

Expi293 cells (Thermo Fisher Scientific)

#### Authentication

Not authenticated

#### Mycoplasma contamination

Not tested

#### Commonly misidentified lines (See [ICLAC](#) register)

Name any commonly misidentified cell lines used in the study and provide a rationale for their use.

## Plants

### Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

### Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

### Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.