

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

<https://doi.org/10.1038/s41586-025-09442-9>

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

Received: 26 November 2024

Accepted: 22 July 2025

Published online: 29 July 2025

 Check for updates

Science frequently benefits from teams of interdisciplinary researchers<sup>1–3</sup>, but many scientists do not have easy access to experts from multiple fields<sup>4,5</sup>. Although large language models (LLMs) have shown an impressive ability to aid researchers across diverse domains, their uses have been largely limited to answering specific scientific questions rather than performing open-ended research<sup>6–11</sup>. Here we expand the capabilities of LLMs for science by introducing the Virtual Lab, an artificial intelligence (AI)–human research collaboration to perform sophisticated, interdisciplinary science research. The Virtual Lab consists of an LLM Principal Investigator agent guiding a team of LLM scientist agents through a series of research meetings, with a human researcher providing high-level feedback. We applied the Virtual Lab to design nanobody binders to recent variants of SARS-CoV-2. The Virtual Lab created a novel computational nanobody design pipeline that incorporates the protein language model ESM, the protein folding model AlphaFold-Multimer and the computational biology software Rosetta and designed 92 new nanobodies. Experimental validation reveals a range of functional nanobodies with promising binding profiles across SARS-CoV-2 variants. In particular, two new nanobodies exhibit improved binding to the recent JN.1 or KP.3 variants<sup>12,13</sup> while maintaining strong binding to the ancestral viral spike protein, suggesting that these are suitable candidates for further investigation. This work demonstrates how the Virtual Lab can rapidly make an impactful, real-world scientific discovery.

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

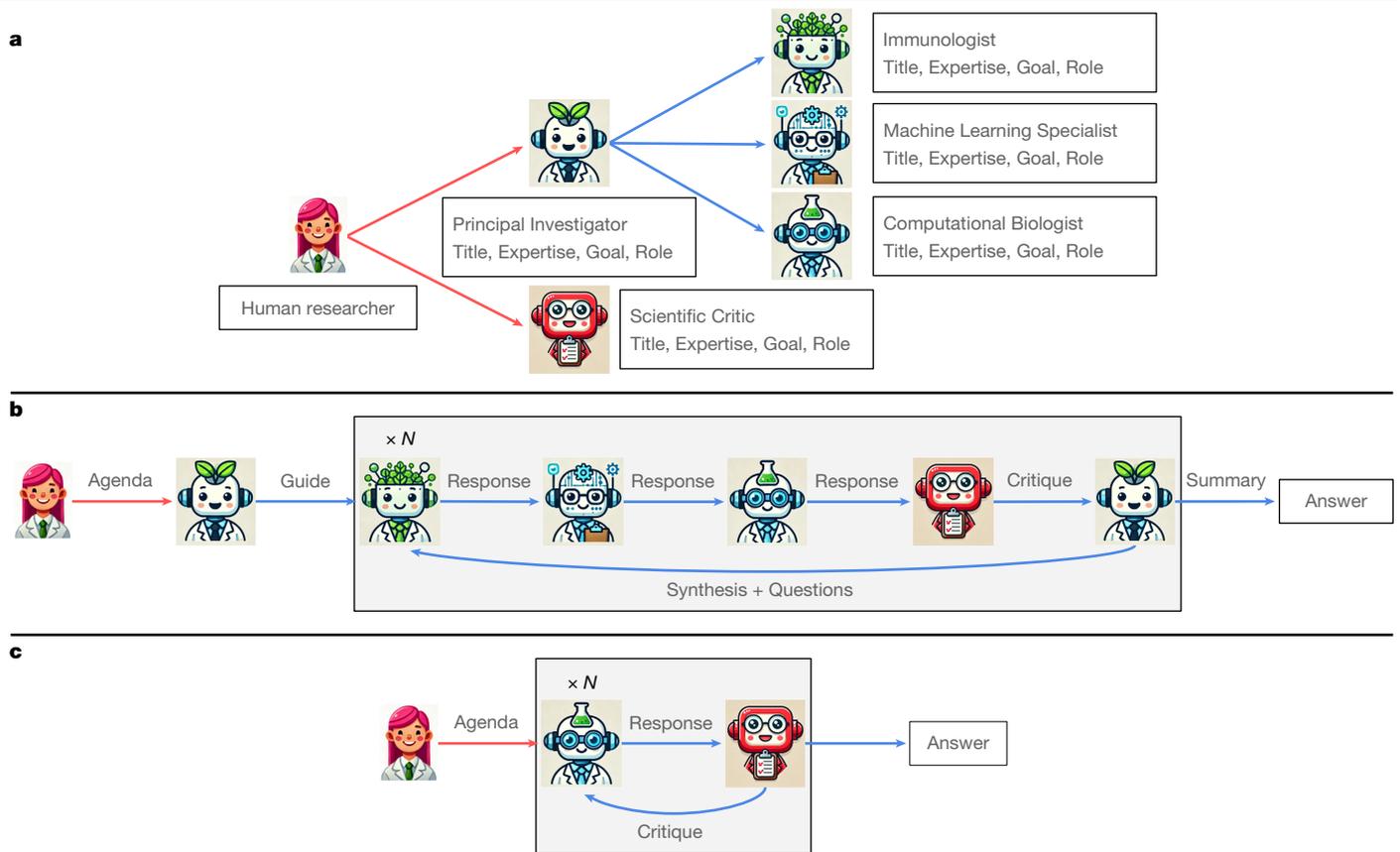
One source of broad scientific knowledge and insights that researchers are now turning to is LLMs such as ChatGPT<sup>17</sup> and Claude<sup>18</sup>. These LLMs have been trained on vast quantities of text data, including scientific literature, and they are therefore able to aid researchers in several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

However, answering individual science questions is very different from engaging in sophisticated research that involves multi-step reasoning across disparate scientific fields with many unknowns. Although

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

Here we introduce the Virtual Lab to overcome these shortcomings via an AI–human research collaboration that performs interdisciplinary science to investigate broad, complex research questions. In the Virtual Lab, a human researcher guides a set of interdisciplinary

<sup>1</sup>Department of Computer Science, Stanford University, Stanford, CA, USA. <sup>2</sup>Chan Zuckerberg Biohub, San Francisco, CA, USA. <sup>3</sup>Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. ✉e-mail: [john.pak@czbiohub.org](mailto:john.pak@czbiohub.org); [jamesz@stanford.edu](mailto:jamesz@stanford.edu)



**Fig. 1 | The Virtual Lab architecture.** **a**, The workflow for designing agents in the Virtual Lab. Each agent is specified with four criteria: Title, Expertise, Goal and Role. The human researcher in the Virtual Lab specifies these criteria to define the PI agent and the Scientific Critic agent. Then, given a short description of the project by the human researcher, the PI agent automatically creates several scientist agents to work on the project by specifying their Title, Expertise, Goal and Role, using its own prompt as an example. **b**, The workflow for a team meeting in the Virtual Lab. The human researcher writes an agenda for the meeting, specifying the topic of discussion. The PI agent begins the meeting by providing initial thoughts and agenda questions as a guide for the remaining agents. Then, over the course of  $N$  rounds of discussion, each scientist agent

provides its response, followed by a critique by the Scientific Critic agent, with the PI agent then synthesizing the discussion and asking follow-up questions. Finally, after the  $N$  rounds of discussion, the PI agent summarizes the discussion and provides an answer regarding the meeting agenda. **c**, The workflow for an individual meeting. The human researcher writes an agenda for the meeting specifying the topic of discussion. Then, the scientist agent tasked with the individual meeting provides a response to the agenda, which is critiqued by the Scientific Critic. In each round, the scientist agent improves its answer based on feedback from the Scientific Critic. Finally, after the  $N$  rounds, the scientist agent provides its final, improved answer.

AI agents<sup>24,25</sup>, such as a biologist or computer scientist, through a set of research meetings that tackle the different phases of a research project. The AI agents are run by an LLM that powers their scientific reasoning abilities with instructions that guide each agent's scientific expertise and interaction with the other agents and the human researcher. The Virtual Lab architecture is versatile and can potentially be applied to a wide variety of interdisciplinary science research projects.

To demonstrate the abilities of the Virtual Lab, we use it to tackle a high-impact, real-world, open-ended scientific problem: designing new nanobodies that exhibit binding to the latest variant of SARS-CoV-2. There are many ways in which scientists could attempt to design such nanobodies, so the Virtual Lab must reason across multiple subfields of biology and computer science to make a series of interrelated decisions about how to best design these nanobodies. Through a series of meetings, the Virtual Lab develops a novel computational nanobody design workflow that incorporates the protein language model ESM<sup>26</sup>, the protein folding model AlphaFold-Multimer<sup>27</sup> and the computational biology software Rosetta<sup>28</sup> to mutate existing nanobodies that bind to the receptor-binding domain (RBD) of the spike protein of the ancestral (Wuhan) strain of SARS-CoV-2 to create nanobodies that bind to the latest variants of the virus, for which an effective binder is lacking<sup>13</sup>.

We experimentally validated 92 mutant nanobodies designed by the Virtual Lab, finding that more than 90% of the nanobodies were expressed and soluble, and that two promising candidates showed unique binding profiles to the recent JN.1 and KP.3 spike RBD variants<sup>12,13</sup>. This outcome illustrates the capability of the Virtual Lab's AI-human collaboration to execute a complex, interdisciplinary science research project that translates to a validated result in the real world.

### Virtual Lab architecture

We created the Virtual Lab as a collaboration between a human researcher and a team of LLM agents to conduct sophisticated, interdisciplinary research (Fig. 1). The human researcher provides high-level guidance for the LLM agents, whereas the LLM agents both decide on general research directions and design solutions to specific research problems. Each agent is implemented by providing the underlying LLM with a prompt defining the agent, which includes its title, expertise, goal and role in the research project (Methods). The human researcher defines two general agents, a Principal Investigator (PI) and a Scientific Critic, and the PI agent then automatically creates a set of scientific agents (for example, an immunologist) depending on the scientific topic of interest to the human researcher (Fig. 1a).

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

### Virtual Lab for nanobody design

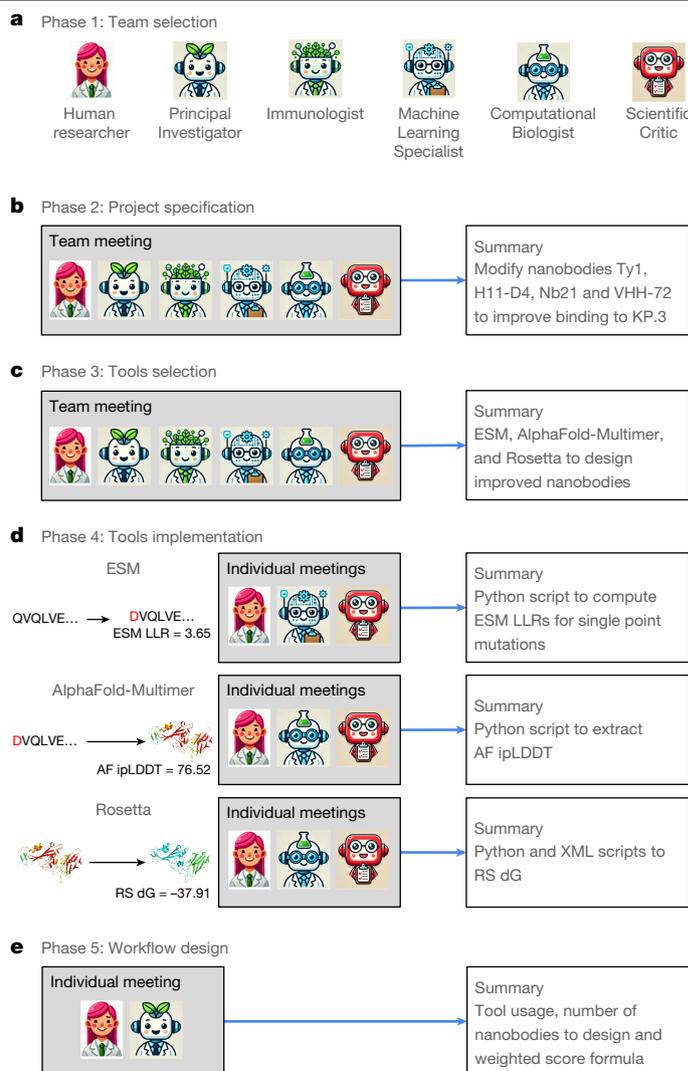
Given the flexibility of the Virtual Lab architecture, it can be applied to a wide variety of interdisciplinary research projects by adapting the agents and the flow of team and individual meetings to the goals and constraints of the specific project. As a demonstration in the domain of biological research, we applied the Virtual Lab with GPT-4o<sup>29</sup> powering the agents to design antibodies or nanobodies that can bind to the spike protein of the KP.3 variant of SARS-CoV-2, which was one of the latest emerging variants at the time of this work<sup>13</sup> (Fig. 2). This is an important and challenging problem because SARS-CoV-2 is rapidly evolving resistance to existing antibody and nanobody therapies, so quickly developing new antibody or 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, an individual meeting with the PI to decide which scientist agent implements each component. Then, for each component, an 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). These phases are discussed in more detail in the Methods.

### Computational nanobody design

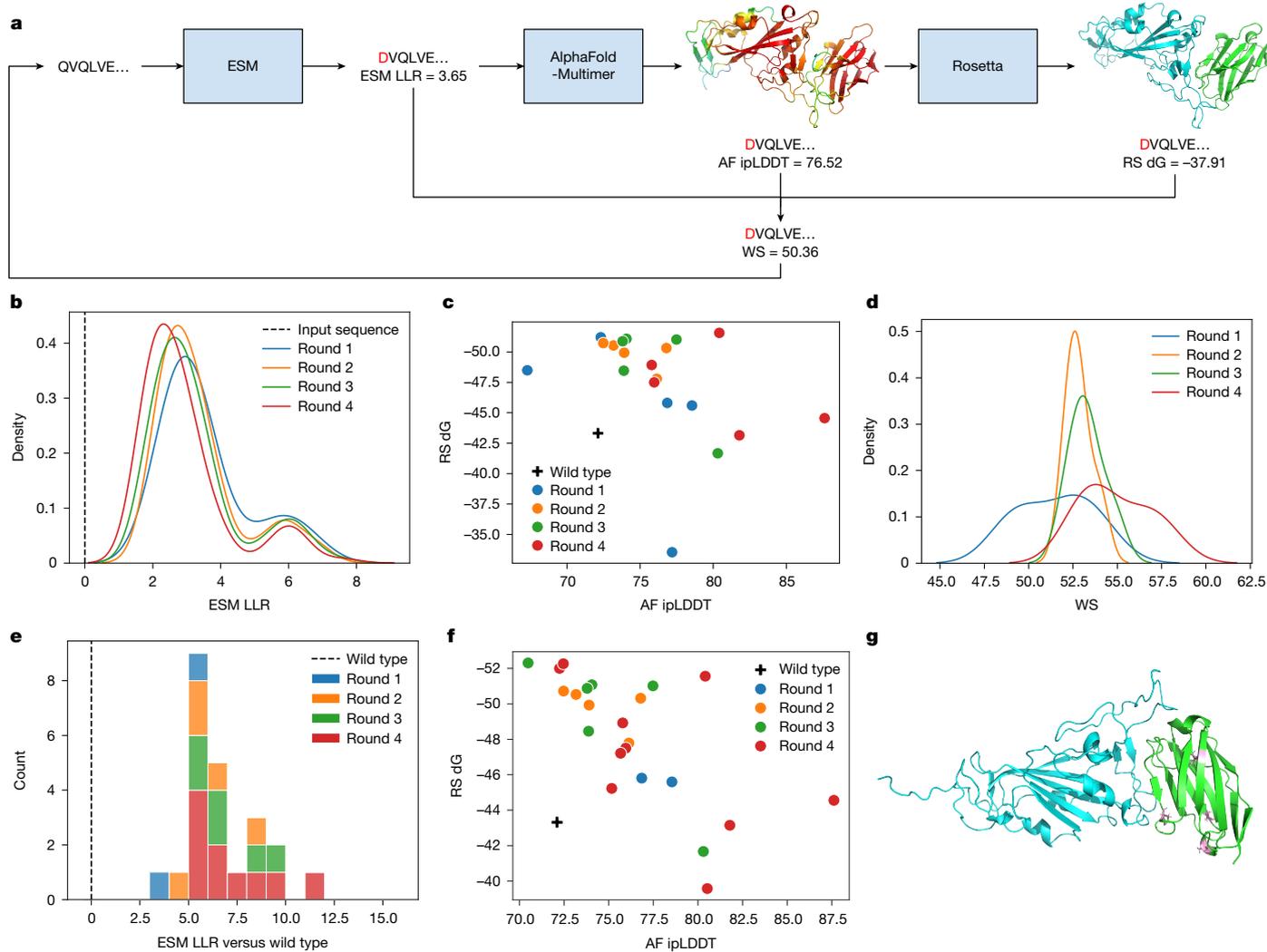
The Virtual Lab built a computational nanobody design workflow that takes existing nanobodies that bind the Wuhan strain of SARS-CoV-2 and adapts them to bind to the recent KP.3 variant (Fig. 3a). Specifically, the workflow starts with four nanobodies—Ty1<sup>32</sup>, H11-D4<sup>33</sup>, Nb21<sup>34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

In the workflow, first, ESM computes the log-likelihood ratio (LLR) of each single point mutation in the nanobody sequence compared



**Fig. 2 | Virtual Lab for nanobody design.** The workflow used to apply the Virtual Lab to nanobody design for the KP.3 variant of SARS-CoV-2. **a**, The workflow begins with the human researcher defining the PI and Scientific Critic agents by specifying their Title, Expertise, Goal and Role. Then, in an individual meeting, the PI agent creates a team of three scientist agents for the project. **b**, A team meeting discusses the project specification, and the agents make decisions such as whether to design antibodies or nanobodies. **c**, In another team meeting, the agents suggest a set of computational tools for nanobody design, including ESM, AlphaFold-Multimer and Rosetta. **d**, In a series of individual meetings, the Machine Learning Specialist and Computational Biologist, with helpful feedback from the Scientific Critic, write code and subsequently improve that code for the ESM, AlphaFold-Multimer and Rosetta components of the nanobody design workflow. **e**, In an individual meeting, the PI agent decides the workflow for using the three computational tools to design and select mutated nanobody candidates.

with the input nanobody sequence, with higher ESM-computed LLRs (ESM LLRs) indicating better (that is, more stable) nanobodies (see Supplementary Note 1 for a discussion of the LLR formula created by the Virtual Lab). Then, the top 20 mutant sequences by ESM LLR are combined with the KP.3 RBD and processed by AlphaFold-Multimer, which predicts the structure of the complex of the two proteins and computes the interface predicted local distance difference test (AF ipLDDT) as a measure of the confidence of the binding interface between the mutant nanobody and the spike RBD. Next, those 20 predicted nanobody-spike complexes are fed into Rosetta, which relaxes their structures and computes the binding energy (RS dG). The ESM LLR, AF ipLDDT and



**Fig. 3 | Nb21 nanobody analysis.** **a**, Each round of nanobody design begins with ESM computing LLRs of single point mutations to the input sequence. For the top 20 mutant sequences by ESM LLR, AlphaFold-Multimer predicts the structure of the nanobody and SARS-CoV-2 spike protein and computes the AF ipLDDT. Rosetta relaxes the complex and computes the RS dG. The top five mutant nanobodies are selected via a WS for the next round of optimization. **b–d**, Evolution of mutant nanobody scores across four rounds of optimization. **b**, The distribution of ESM LLR values for proposed Nb21 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. **c**, The AF ipLDDT and the RS dG of the top five proposed nanobodies, selected by WS,

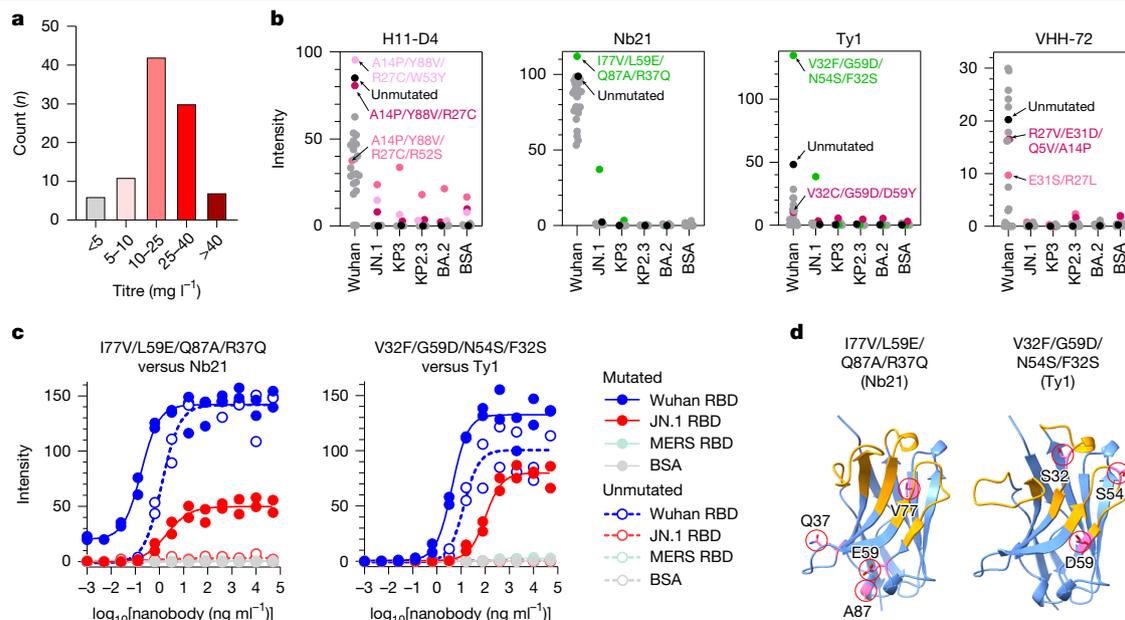
at the end of each round of optimization. **d**, The distribution of WS values of the top five proposed nanobodies at the end of each round of optimization. **e–g**, Analysis of the final set of 23 mutant nanobodies selected across all rounds of optimization. **e**, The distribution of ESM LLR<sup>WT</sup> values for the selected nanobodies and the wild-type nanobody. **f**, The AF ipLDDT and RS dG values of the selected nanobodies and the wild-type nanobody. **g**, The structure (predicted by AlphaFold-Multimer followed by Rosetta relaxation) of the RBD of the KP.3 spike protein (cyan) and the nanobody mutant Nb21 I77V-L59E-Q87A-R37Q (green). Side chains are shown for interface residues (residues within 4 Å of the opposite chain). Mutant nanobody residues are in pink. Structure images were generated in PyMol 3.1.3.

RSdG scores are combined into a weighted score (WS) using the formula  $WS = 0.2 \times (\text{ESM LLR}) + 0.5 \times (\text{AF ipLDDT}) - 0.3 \times (\text{RS dG})$ . The 20 mutant nanobodies are ranked by WS and the top five are selected. Those five are then fed back into the pipeline to introduce another round of point mutations. The process is repeated four times total to introduce up to four mutations. Finally, 23 mutated nanobodies are selected for each of the four starting nanobodies (92 total) using a modified weighted score,  $WS^{\text{WT}}$ . This score is the same as the WS except that it uses a modified ESM LLR<sup>WT</sup>, which is the ratio between the proposed nanobody (with one to four mutations) and the wild-type nanobody (with zero mutations) rather than the input nanobody from the previous round (with one less mutation).

The successive rounds of optimization improved the quality of the proposed mutant nanobody sequences according to the three metrics

of ESM LLR, AF ipLDDT and RS dG. Figure 3b–g show relevant metrics for Nb21, and similar results were obtained for Ty1 (Extended Data Fig. 2), H11-D4 (Extended Data Fig. 3) and VHH-72 (Extended Data Fig. 4). Extended Data Table 1 shows scores for the wild-type sequence and some of the mutant sequences that were selected for experimental validation.

In each round, the top sequences selected by ESM LLR had LLR values in the range 1–8, indicating that each subsequent round of mutation improved the overall quality of the nanobody compared with the input sequence from the previous round (Fig. 3b). This is according to ESM's internal understanding of nanobody likelihood, which does not take the antigen (spike protein) into account but does understand overall nanobody quality. The top mutant nanobody sequences selected by ESM LLR in each round generally had improved structural complexes



**Fig. 4 | Experimental validation of Virtual Lab nanobodies.** **a**, Histogram of expression levels across 96 nanobodies. Titre is expressed as milligrams of soluble, periplasmic nanobody per litre of culture. **b**, ELISA binding profiles of nanobodies to a panel of antigens. For each SARS-CoV-2 RBD protein and BSA, individual spots represent the ELISA binding intensity of each of the 24 nanobodies. Unmutated nanobodies (H11-D4, Nb21, Ty1 and VHH-72) are shown in black and nanobodies exhibiting high non-specific binding are shown in shades of red (light pink, pink and magenta). The Nb21 mutant (177V/L59E/Q87A/R37Q) and the Ty1 mutant (V32F/G59D/N54S/F32S) that bind to JN.1 are shown in green. Data are the mean of 2 measurements at a nanobody lysate

dilution of 1:2. **c**, Comparison of ELISA binding of mutants and their unmutated sequences. Data shown are 2 biological replicates at a 12-point serial dilution of purified nanobody, fitted to a 4-parameter logistic curve. **d**, Location of mutant nanobody mutations. Models of Nb21(177V/L59E/Q87A/R37Q) and Ty1(V32F/G59D/N54S/F32S) generated by the Virtual Lab using AlphaFold-Multimer, are shown in ribbon representation (blue), with the complementarity-determining region loops shown in orange. Mutations introduced by the Virtual Lab are shown in pink and in red circles. Structure images were generated using ChimeraX<sup>51</sup>.

with the KP.3 spike protein, based on improved AF ipLDDT, improved RS dG or both (Fig. 3c). The WS values of the top five sequences at the end of each round improved (Fig. 3d), even when using the ESM LLR instead of the ESM LLR<sup>WT</sup> that corrects for the effect of multiple mutations and not just the most recent mutation.

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

Applying this workflow to each of the 4 starting nanobodies resulted in 92 final selected sequences (23 per starting nanobody). All 92 mutant nanobodies had a positive ESM LLR, indicating that ESM preferred the mutant over the wild type. Among the 92 mutant nanobodies, 78 (85%) had an AF ipLDDT greater than their respective wild-type nanobody, and 32 (35%) had an AF ipLDDT  $\geq 80$ , which is in line with the AF ipLDDT scores of high-accuracy AlphaFold-Multimer antibody-antigen structural models<sup>36</sup>. Furthermore, 60 (65%) had an RS dG lower (better) than their respective wild-type nanobody, and 23 (25%) of the 92 mutants had an RS dG  $\leq -50$ , which is in line with strong Rosetta binding energy values of nanobodies or antibodies in complex with the SARS-CoV-2 RBD<sup>28,37</sup>.

### Experimental validation of nanobodies

To validate the nanobodies designed by the Virtual Lab, we conducted a set of experiments to measure their binding to a panel of spike RBD

proteins (Extended Data Fig. 5). We first overexpressed each nanobody in *Escherichia coli* and isolated soluble protein from the periplasm. The designed nanobodies show excellent expression, with 38% (35 out of 92) of the designs having titres of more than 25 mg of soluble, periplasmic nanobody per litre of cell culture (Fig. 4a and Extended Data Fig. 6) and only 6.5% (6 out of 92) of the designs having a titre of less than 5 mg l<sup>-1</sup>. Thus, the mutations proposed by the Virtual Lab are well tolerated and do not cause large-scale misfolding or aggregation of the nanobodies.

To determine whether the 92 mutant nanobodies—23 each for Ty1, H11-D4, Nb21 and VHH-72—and the 4 wild-type nanobodies could bind to the SARS-CoV-2 KP.3 spike RBD, we generated a spike RBD array that included the KP.3 RBD protein, its closely related parental strain (JN.1 RBD), a closely related variant (KP.2.3 RBD), an early Omicron variant (BA.2 RBD) and the ancestral strain (Wuhan RBD), which all four wild-type nanobodies show specificity for.

Using this RBD array, we first profiled the binding of all 96 nanobodies by indirect enzyme-linked immunosorbent assay (ELISA) to each antigen at a nanobody lysate dilution of 1:2 (Fig. 4b). For the H11-D4 and Nb21 series, binding to Wuhan RBD is overwhelmingly retained in 96% of mutant nanobodies (44 out of 46). Three mutants in the H11-D4 series have high non-specific binding to bovine serum albumin (BSA) and all of the RBDs (Fig. 4b), possibly owing to the Virtual Lab inadvertently introducing an R27C mutation, which may be leading to disulfide crosslinking. In contrast to the H11-D4 and Nb21 mutants, the Ty1 mutants, overall, exhibit poor binding to Wuhan RBD (10 out of 23 mutants). If position 32 of Ty1, selected by the Virtual Lab as the first residue to mutate for each mutant, is not well tolerated, this could result in the observed poor binding to Wuhan RBD compared to the H11-D4 and Nb21 mutants. More than half of the VHH-72 mutants (13 out of 22) retain binding to Wuhan RBD at levels similar to that observed for the unmutated VHH-72 nanobody. Thus, the Virtual Lab designs

are, overall, well tolerated with respect to preserving their original specificity to Wuhan RBD.

Of the 92 Virtual Lab-designed nanobodies, two show promising binding profiles beyond that of Wuhan RBD. The first, derived from Nb21(I77V/L59E/Q87A/R37Q) (that is, Nb21 with the mutations I77V, L59E, Q87A and R37Q), shows binding to JN.1 RBD in ELISAs with no non-specific binding to the Middle East respiratory syndrome coronavirus (MERS-CoV) RBD and BSA (Fig. 4b–d). Maximal binding of the purified mutant nanobody to JN.1 RBD is less than that to Wuhan RBD, with a weaker half-maximal effective concentration ( $EC_{50}$ ) ( $2.0 \text{ ng ml}^{-1}$  versus  $0.2 \text{ ng ml}^{-1}$ ) revealing that this new binding to JN.1 RBD may be moderate. The wild-type Nb21 has very low ELISA binding to JN.1 RBD (Fig. 4b), suggesting that the Virtual Lab mutant has improved upon this existing very weak binding. Of note, this mutant also shows increased binding to KP.3 RBD (average intensity = 3.5) compared with the other Nb21 mutants (average intensity =  $0.06 \pm 0.09$ ,  $n = 22$ ) and the unmutated sequence (average intensity = 0.1) (Fig. 4b). We further confirmed this KP.3 binding enrichment in separate ELISA experiments. The second, a Ty1 mutant nanobody (V32F/G59D/N54S/F32S) not only improved binding to Wuhan RBD, as measured by ELISA, but also gained moderate binding to JN.1 RBD (Fig. 4b–d). By contrast, we see no evidence for even low levels of unmutated Ty1 nanobody binding to JN.1 RBD (further details in the Supplementary Note 2).

Across the mutant nanobodies, the preserved and improved binding affinities for the Wuhan RBD (and JN.1 RBD for Nb21) relative to their respective wild-type forms is likely to be due to the effect of the Virtual Lab's use of ESM log likelihoods, which are agnostic to the antigen but select for evolutionarily favourable nanobody sequences with improved fitness<sup>38</sup>. By contrast, the Ty1 mutant that gained binding affinity for the JN.1 RBD, which is the close ancestor of KP.3 sharing 99.1% identity in the RBD<sup>39</sup> (220 out of 222 residues), and the Nb21 mutant that gained binding affinity for the KP.3 RBD and improved binding affinity for the JN.1 RBD may demonstrate the effect of the AlphaFold-Multimer and Rosetta scoring, which explicitly aim to predict binding affinity of the mutant antibody to the KP.3 RBD, and thus by extension the closely related JN.1 RBD. Through the use of these three tools, the Virtual Lab designed a set of promising nanobody candidates with potential for further development.

## Analyses of Virtual Lab interactions

The Virtual Lab proceeded rapidly through the phases of the nanobody design workflow, with each meeting (or a set of parallel meetings) only taking the agents about 5–10 min (approximately US\$1–2 GPT-4o token cost), for a total of around 1–2 h (US\$10–20 GPT-4o token cost) to complete all of the phases of meetings. Factoring in the time to tune the prompts for each phase to elicit the most productive and relevant discussions and to review and debug the code written by the agents, the Virtual Lab completed its nanobody workflow design in just a few days, compared with an estimated period of several weeks for a human researcher working independently to design and build the same pipeline from scratch. It then took about a week to run the computational nanobody design pipeline (ESM, AlphaFold-Multimer and Rosetta), followed by around six weeks to synthesize the nanobodies and two weeks for binding experiments.

Within the Virtual Lab discussions, the individual identities of the agents contributed to a comprehensive, interdisciplinary discussion with each agent providing a perspective based on their specific background (Fig. 5a; additional analysis in Supplementary Note 3). By contrast, a team of generic agents without distinct scientific backgrounds tends to argue more among themselves owing 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-separated tokens) across all phases of the workflow,

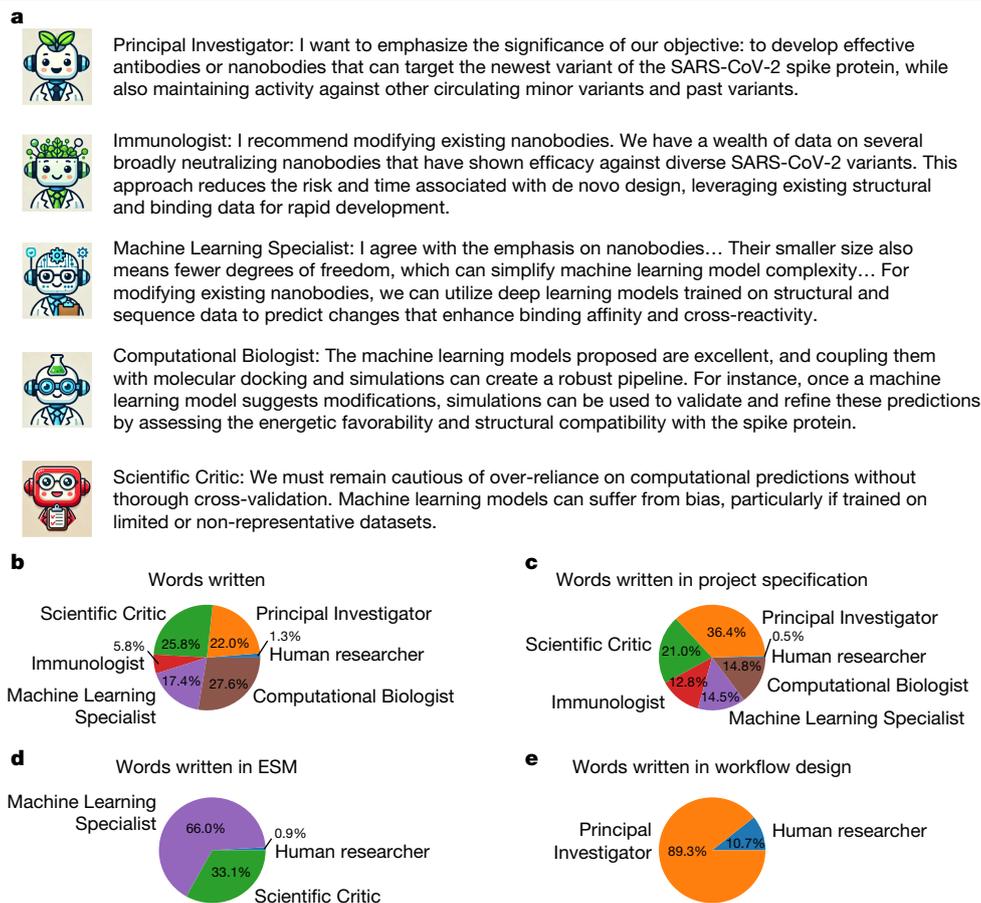
representing just 1.3% of all the words written by the Virtual Lab (Fig. 5b). By contrast, the LLM agents wrote 122,462 words (98.7% of all words). The ESM, AlphaFold-Multimer and Rosetta scripts were all written from scratch by the agents, and some complementary data wrangling and job scheduling scripts 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 their scientific discussions. For example, different agents write different amounts depending on the meeting context (Fig. 5b–e and 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, because the Scientific Critic must address the limitations of every agents' response, whereas each scientist agent is only concerned with providing its own opinions. The use of parallel meetings (Methods) and the inclusion of the Scientific Critic were particularly notable, because they tended to lead to answers with improved consistency and quality (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, diverse binding profiles across multiple strains of SARS-CoV-2. The human researcher and team of LLM agents in the Virtual Lab worked together through a series of meetings to rapidly build a complex nanobody design pipeline that incorporates state-of-the-art machine learning and computational biology tools. Building this pipeline required knowledge of multiple areas of science from immunology to protein folding to machine learning and required making decisions that involved reasoning across many aspects of the project simultaneously. The Virtual Lab successfully built and ran this nanobody design pipeline, starting with a set of four well-characterized nanobodies (Ty1, H11-D4, Nb21 and VHH-72) with potency and diverse binding modes against early variants of SARS-CoV-2<sup>32–35</sup> and developing them into 92 nanobody candidates for recent variants of SARS-CoV-2 that were experimentally validated by human researchers. These 92 nanobodies—efficiently selected from the trillions of nanobody sequences with 1–4 mutations—include candidates for further development, such as a Nb21 mutant that enhances binding to the JN.1 RBD and gains binding to the KP.3 RBD and a Ty1 mutant that gains binding to the JN.1 RBD. This outcome serves as an example of how human researchers can partner with LLM agents in the Virtual Lab to rapidly achieve a promising scientific result that can streamline further experiments. Even if the ultimate scientific decisions of the Virtual Lab agents are similar to those in the scientific literature, the ability of the agents to quickly adapt those methods to the scientific question at hand shows how LLM agents can potentially empower human researchers to do complex, interdisciplinary science even when they do not have access to an expert panel of human scientists.

Previous work applying AI to science has generally treated AI methods as tools used by human researchers, such as AlphaFold to predict protein structures<sup>14</sup> or LLMs to answer scientific questions<sup>6–11</sup>, with the human researcher making all the high-level research decisions and design choices. By contrast, in the Virtual Lab, human researchers work alongside LLM agents to design and run a research project. The strength of the Virtual Lab comes from its multi-agent<sup>24,40–42</sup> architecture, which empowers an AI–human scientific collaboration through a series of meetings between a human researcher and a team of interdisciplinary LLM agents. The different backgrounds of the various scientist agents leads to discussions that approach complicated scientific questions



**Fig. 5 | Virtual Lab discussion analysis.** **a**, Excerpts from a Virtual Lab team meeting discussing the nanobody project specification. Each LLM agent addresses the agenda from its own perspective based on its Title, Expertise, Goal and Role, leading to a comprehensive and interdisciplinary discussion of the agenda. **b**, The number of words (space-separated tokens) written by the

Virtual Lab (human researcher and each LLM agent) across all phases of the nanobody design process. **c**, The number of words written by the Virtual Lab in the project specification phase. **d**, The number of words written by the Virtual Lab in ESM implementation. **e**, The number of words written by the Virtual Lab in the workflow design phase.

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

Although the Virtual Lab architecture provides useful structure for scientific discussions between the human researcher and the LLM agents, it has several limitations that are inherent in the current generation of LLMs. For example, because LLMs are only trained on data up to a certain date (the ‘knowledge cutoff’), the agents may not be aware of the most up-to-date scientific literature and code<sup>44</sup> (for example, knowledge about AlphaFold 3<sup>45</sup> versus AlphaFold-Multimer<sup>27</sup>). However, these issues could be fixed by providing the agents with relevant information and documentation, for example through retrieval-augmented

generation<sup>46,47</sup> or finetuning<sup>48</sup> (see Supplementary Note 5 for an exploration of finetuning agents). Additionally, future work could explore developing sandboxed environments to enable the agents to independently install computational or AI tools and then write, debug and run code that uses those tools for a particular application.

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

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

Although we applied the Virtual Lab to nanobody design here, the Virtual Lab architecture of LLM agents and meetings is agnostic to specific research questions or scientific domains. The Virtual Lab can be

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

## Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-025-09442-9>.

- Porter, A. L. & Rafols, I. Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics* **81**, 719–745 (2009).
- Sijp, W. Paper authorship goes hyper. *Nature Index* [www.nature.com/nature-index/news/paper-authorship-goes-hyper](http://www.nature.com/nature-index/news/paper-authorship-goes-hyper) (2018).
- Castelvecchi, D. Physics paper sets record with more than 5,000 authors. *Nature* <https://doi.org/10.1038/nature.2015.17567> (2015).
- Specht, A. & Crowston, K. Interdisciplinary collaboration from diverse science teams can produce significant outcomes. *PLoS ONE* **17**, e0278043 (2022).
- Cohen, J. J. et al. Tackling the challenge of interdisciplinary energy research: a research toolkit. *Energy Res. Soc. Sci.* **74**, 101966 (2021).
- Kung, T. H. et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit. Health* **2**, e0000198 (2023).
- Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
- Laurent, J. M. et al. LAB-Bench: measuring capabilities of language models for biology research. Preprint at <https://doi.org/10.48550/arXiv.2407.10362> (2024).
- Guo, T. et al. What can large language models do in chemistry? A comprehensive benchmark on eight tasks. *Adv. Neural Inf. Process. Syst.* **36**, 59662–59688 (2023).
- Sun, L. et al. SciEval: a multi-level large language model evaluation benchmark for scientific research. *Proc. AAAI Conf. Artif. Intell.* **38**, 19053–19061 (2024).
- Stribling, D. et al. The model student: GPT-4 performance on graduate biomedical science exams. *Sci. Rep.* **14**, 5670 (2024).
- Kaku, Y. et al. Virological characteristics of the SARS-CoV-2 JN.1 variant. *Lancet Infect. Dis.* **24**, e82 (2024).
- Kaku, Y. et al. Virological characteristics of the SARS-CoV-2 KP.3, LB.1, and KP.2.3 variants. *Lancet Infect. Dis.* **24**, e482–e483 (2024).
- Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
- Callaway, E. Chemistry Nobel goes to developers of AlphaFold AI that predicts protein structures. *Nature* **634**, 525–526 (2024).
- Bromham, L., Dinnage, R. & Hua, X. Interdisciplinary research has consistently lower funding success. *Nature* **534**, 684–687 (2016).
- OpenAI et al. GPT-4 Technical Report. Preprint at <https://doi.org/10.48550/arXiv.2303.08774> (2024).
- Anthropic. The Claude 3 Model Family: Opus, Sonnet, Haiku (Anthropic, 2024).
- Simon, E., Swanson, K. & Zou, J. Language models for biological research: a primer. *Nat. Methods* **21**, 1422–1429 (2024).
- M. Bran, A. et al. Augmenting large language models with chemistry tools. *Nat. Mach. Intell.* **6**, 525–535 (2024).
- Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570–578 (2023).
- Lu, C. et al. The AI scientist: towards fully automated open-ended scientific discovery. Preprint at <https://doi.org/10.48550/arXiv.2408.06292> (2024).
- Si, C., Yang, D. & Hashimoto, T. Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers. in *13th Int. Conf. Learn. Represent.* <https://openreview.net/pdf?id=M23dTGWCZy> (ICLR, 2025).
- Wu, Q. et al. AutoGen: enabling next-gen LLM applications via multi-agent conversation. In *1st Conf. Lang. Model.* <https://openreview.net/forum?id=BAakY1hNKS> (COLM, 2024).
- Gao, S. et al. Empowering biomedical discovery with AI agents. *Cell* **187**, 6125–6151 (2024).
- Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
- Evans, R. et al. Protein complex prediction with AlphaFold-Multimer. Preprint at <https://doi.org/10.1101/2021.10.04.463034> (2021).
- Boorla, V. S. et al. De novo design and Rosetta-based assessment of high-affinity antibody variable regions (Fv) against the SARS-CoV-2 spike receptor binding domain (RBD). *Proteins Struct. Funct. Bioinformatics* **91**, 196–208 (2023).
- OpenAI et al. GPT-4o System Card. Preprint at <https://doi.org/10.48550/arXiv.2410.21276> (2024).
- Cao, Y. et al. Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. *Nature* **602**, 657–663 (2022).
- Planas, D. et al. Considerable escape of SARS-CoV-2 Omicron to antibody neutralization. *Nature* **602**, 671–675 (2022).
- Hanke, L. et al. An alpaca nanobody neutralizes SARS-CoV-2 by blocking receptor interaction. *Nat. Commun.* **11**, 4420 (2020).
- Huo, J. et al. Neutralizing nanobodies bind SARS-CoV-2 spike RBD and block interaction with ACE2. *Nat. Struct. Mol. Biol.* **27**, 846–854 (2020).
- Xiang, Y. et al. Versatile and multivalent nanobodies efficiently neutralize SARS-CoV-2. *Science* **370**, 1479–1484 (2020).
- Wrapp, D. et al. Structural basis for potent neutralization of betacoronaviruses by single-domain camelid antibodies. *Cell* **181**, 1004–1015.e15 (2020).
- Yin, R. & Pierce, B. G. Evaluation of AlphaFold antibody–antigen modeling with implications for improving predictive accuracy. *Protein Sci.* **33**, e4865 (2024).
- Yang, J. et al. Computational design and modeling of nanobodies toward SARS-CoV-2 receptor binding domain. *Chem. Biol. Drug Des.* **98**, 1–18 (2021).
- Hie, B. L. et al. Efficient evolution of human antibodies from general protein language models. *Nat. Biotechnol.* **42**, 275–283 (2024).
- Planas, D. et al. Escape of SARS-CoV-2 variants KP.1.1, LB.1, and KP.3.3 from approved monoclonal antibodies. *Pathog. Immun.* **10**, 1 (2024).
- Chan, C.-M. et al. ChatEval: towards better LLM-based evaluators through multi-agent debate. In *12th Int. Conf. Learn. Represent.* <https://openreview.net/forum?id=FQepisCUWu> (ICLR, 2024).
- Liu, Z., Zhang, Y., Li, P., Liu, Y. & Yang, D. A dynamic LLM-powered agent network for task-oriented agent collaboration. In *1st Conf. on Lang. Model.* <https://openreview.net/forum?id=XIIOWp1XA9> (COLM, 2024).
- Talebirad, Y. & Nadiri, A. Multi-agent collaboration: harnessing the power of intelligent LLM agents. Preprint at <https://doi.org/10.48550/arXiv.2306.03314> (2023).
- Wei, J. et al. Chain-of-thought prompting elicits reasoning in large language models. In *Proc. 36th International Conference on Neural Information Processing Systems* 24824–24837 (Curran Associates, 2024).
- Cheng, J. et al. Dated data: tracing knowledge cutoffs in large language models. In *1st Conf. Lang. Model.* <https://openreview.net/forum?id=wS7PxDjy6m> (COLM, 2024).
- Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
- Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv. Neural Inf. Process. Syst.* **33**, 9459–9474 (2020).
- Gao, Y. et al. Retrieval-augmented generation for large language models: a survey. Preprint at <https://doi.org/10.48550/arXiv.2312.10997> (2024).
- Ding, N. et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nat. Mach. Intell.* **5**, 220–235 (2023).
- White, J. et al. A prompt pattern catalog to enhance prompt engineering with ChatGPT. In *Proc. 30th Conference on Pattern Languages of Programs* 1–31 (Hillside Group, 2023).
- Ji, Z. et al. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **55**, 1–38 (2023).
- Meng, E. C. et al. UCSF ChimeraX: tools for structure building and analysis. *Protein Sci.* **32**, e4792 (2023).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2025

# Article

## Methods

### Virtual Lab architecture

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

**Agents.** Each LLM agent in the Virtual Lab is defined with a prompt that specifies four key criteria: (1) Title, the name of the agent; (2) Expertise, the scientific expertise the agent has; (3) Goal, the ultimate goal of the agent in the context of the research project; and (4) Role, the specific role that the agent will play in the research project.

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

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

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

1. **Agenda:** (required) a description of the scientific topic to be discussed during the meeting.
2. **Agenda questions:** (optional) a set of questions that the agents must answer by the end of the meeting.
3. **Agenda rules:** (optional) a set of rules that the agents must follow during the meeting.
4. **Summaries:** (optional) agent-written summaries of previous meetings to provide information about previous decisions.
5. **Contexts:** (optional) additional information (for example, scientific papers) for the agents to take into consideration.
6. **Rounds:** (required) the number of rounds (typically  $N = 3$ ) of discussion among the agents.

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

**Team meeting.** In team meetings, all the agents (PI agent, scientist agents and Scientific Critic agent) participate in a conversation to address a broad research topic. First, the human researcher writes an agenda for the team meeting along with any applicable agenda questions and agenda rules. The team meeting then begins with an automatically constructed prompt that introduces the agents, agenda, agenda questions (if any), and agenda rules (if any) and describes the flow of the meeting, which involves multiple rounds of discussion. The PI agent is prompted to start the discussion by providing their initial thoughts and any guiding questions that they want to ask the team. Then, each scientist agent and the Scientific Critic agent are prompted one-by-one (in an order set by the human researcher) to provide their thoughts on the ongoing discussion given everything that has been said by the other agents. At the end of a round of discussion, the PI agent synthesizes the points raised by each agent, makes decisions based on agent input, and asks follow-up questions to further the discussion. After  $N$  rounds of discussion (with  $N$  set by the human researcher), the PI agent summarizes the discussion for future meetings, provides

a recommendation regarding the agenda, and answers the agenda questions (if any). The human researcher in the Virtual Lab can then read just this final response by the PI agent, thus benefiting from the extensive discussions among the LLM agents while only needing to read the final short response to understand the decisions that were made.

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

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

### Virtual Lab for nanobody design

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

**Team selection.** First, the Virtual Lab used an individual meeting with the PI agent (run five times in parallel followed by a merge meeting) to create a set of scientist agents for the project. The meeting agenda contained a background prompt about antibody/nanobody design for the spike protein of the recent KP.3 SARS-CoV-2 variant and a request for the PI agent to select a team of three scientist agents for the project using the same agent structure (Title, Expertise, Goal, Role) as the PI's own definition. The PI decided to create an Immunologist, a Machine Learning Specialist and a Computational Biologist.

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

1. Will you design standard antibodies or nanobodies?

Agent answer: Nanobodies

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

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

Agent answer: Modify existing nanobodies

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

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

Agent answer: Ty1, H11-D4, Nb21, VHH-72

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

Given these decisions, the following phases proceeded with nanobody design by modifying the four nanobodies suggested by the Virtual Lab (Ty1, H11-D4, Nb21 and VHH-72), which are specific to the ancestral Wuhan spike protein, to increase their affinity to the spike protein of the KP.3 variant of SARS-CoV-2. Furthermore, the Virtual Lab suggested prioritizing “enhancing interactions with the RBD of the spike protein by altering residues that contribute to binding affinity”, so the Virtual Lab subsequently focused on developing nanobodies that bind to the RBD of the KP.3 spike protein.

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

**Tools implementation.** With the project well-specified and a set of computational nanobody tools chosen, the Virtual Lab then worked on implementing these tools for nanobody design. For each tool, the Virtual Lab selected the most appropriate scientist agent via an individual meeting with the PI. Then for each tool, the Virtual Lab ran an individual meeting with the selected scientist agent and the Scientific Critic (five parallel meetings followed by a merge meeting run by the scientist agent) to implement the tool. These meetings included a set of agenda rules that specify how code should be written—for example, with good documentation and without leaving functions undefined. These initial implementations contained small errors that needed correction, so the Virtual Lab then ran a single follow-up individual meeting (no parallelization or Scientific Critic) with the scientist agent to automatically fix all the errors that arose.

*ESM usage.* The Machine Learning Specialist agent was responsible for writing a Python script to identify the most promising point mutations

to a nanobody sequence based on the ESM LLR of the mutant sequence compared to the input sequence. The agent wrote a 130-line Python script with three functions: a main function to run the script, a function to parse command-line arguments (for example, the input nanobody sequence), and a function that uses a pre-trained ESM model to compute LLRs for point mutations.

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

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

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

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

$$WS = 0.2 \times (\text{ESM LLR}) + 0.5 \times (\text{AF ipLDDT}) - 0.3 \times (\text{RS dG})$$

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

#### Nanobody design workflow

The Virtual Lab ran the nanobody design computational workflow to design improved nanobody candidates for the KP.3 variant of SARS-CoV-2. The workflow was run independently for each of the four nanobodies suggested by the agents: Ty1, H11-D4, Nb21 and VHH-72.

# Article

Below, we describe the workflow in terms of a single starting nanobody for simplicity.

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

In round 1, the Virtual Lab ran ESM to calculate the ESM LLR of every possible single point mutation to the wild-type nanobody. The top 20 mutated sequences by ESM LLR were retained. For each of these 20 mutated sequences, AlphaFold-Multimer and Rosetta were applied in the same way as for the wild-type sequence. The Virtual Lab then computed the WS for each of the 20 mutated sequences and selected the top 5 sequences for the next round. In rounds 2–4, the Virtual Lab applied the same procedure but now starting with 5 input sequences to the ESM LLR script, resulting in 100 top mutated sequences (20 proposed mutant sequences for each of the 5 input sequences). These sequences were analysed by AlphaFold-Multimer and Rosetta, and the top 5 of these 100 sequences were selected at the end of each round by their WS.

After running all four rounds of mutation, the Virtual Lab needed to select the best mutated nanobody sequences across all four rounds for experimental validation. Doing so required using a slight variant of the WS. In each round, the WS used the ESM LLR calculated as a ratio between the proposed mutant sequence and the input sequence for that round (that is, an output sequence from the previous round), which differ by a single mutation. However, in order to fairly select the best sequences across different rounds with different numbers of mutations, an alternate ESM LLR, the ESM LLR<sup>WT</sup>, was computed between each proposed mutant sequence (with one to four mutations) and the wild-type sequence. The Virtual Lab then scored all mutant nanobody sequences using the WS<sup>WT</sup>, which is the weighted score calculated using the ESM LLR<sup>WT</sup> in place of the ESM LLR. The top 23 mutant sequences were selected for experimental validation along with the wild-type sequence as a point of reference.

## Nanobody experimental validation

Codon-optimized DNA sequences for the SARS-CoV-2 spike RBDs from JN.1, KP.3, KP.2.3 and BA.2<sup>13,57</sup>, modified to include a N-terminal signal peptide (MFVFLVLLPLVSSQ), a C-terminal Gly-Ser linker and 6× His tag and a stop codon, were synthesized and cloned into pTwist-CMV-BetaGlobin (Twist Biosciences). For the MERS-CoV RBD, the codon-optimized DNA sequence for the RBD was modified to include an N-terminal signal peptide (MYRMQLLSIALSLALVTNS), C-terminal Gly-Ser linker, 8× His tag, AviTag sequences and a stop codon. RBDs were transiently expressed in Expi293 cells (Thermo Fisher Scientific, not authenticated or tested for mycoplasma contamination), and purified in parallel<sup>58</sup> by Ni-NTA Excel affinity chromatography followed by desalting into PBS and concentration. The purification of Wuhan SARS-CoV-2 RBD has been described previously<sup>58</sup>. Codon-optimized DNA sequences for nanobodies, modified to include an N-terminal pelB signal peptide (MKYLLP TAAAGLLLLAAQPAMA), a C-terminal 6× His tag and a stop codon, were synthesized and cloned into pET-29b(+) (Twist Biosciences). Nanobodies were expressed in 96-well and 24-well format in auto-induction media<sup>59</sup>, and periplasmic fractions from 4 ml of cell culture pellets were released by mild lysis in 400 μl PBS, following methods as described<sup>60</sup>. Titres of soluble nanobody were estimated from periplasmic fractions by SDS-PAGE densitometry analysis of nanobody bands using a BSA standard curve. Selected nanobodies were scaled up at 100 ml in shake flasks and purified from periplasmic fractions by Ni-NTA chromatography followed by desalting into PBS and concentration.

Multiplexed ELISA measurements were performed as generally described<sup>61</sup>. Array patterns were printed using a sciFLEXARRAYER S12. Each RBD and BSA (negative control) spot was printed in duplicate, using up to three 200–250 pl drops for each spot, at a source concentration of 50 μg ml<sup>-1</sup>. Unpurified lysates or purified nanobodies were diluted in PBS-T (5% skim milk in PBS + 0.05% Tween-20), and RBD-bound nanobodies were recognized by anti-Alpaca IgG VHH secondary antibodies (Jackson ImmunoResearch, 128-065-230 (for H11-D4, Nb21, and VHH-72 series) and 128-065-232 (for Ty1 series)) at 1:10,000 dilution in PBS-T.

## Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

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

## Code availability

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

52. Yuksekgonul, M. et al. Optimizing generative AI by backpropagating language model feedback. *Nature* **639**, 609–616 (2025).
53. Peeperkorn, M., Kouwenhoven, T., Brown, D. & Jordanous, A. Is temperature the creativity parameter of large language models? In *15th Int. Conf. Comput. Creativity* (Association for Computational Creativity, 2024).
54. Chen, H. & Ding, N. Probing the “creativity” of large language models: can models produce divergent semantic association? In *Findings of the Association for Computational Linguistics: EMNLP 2023* (eds Bouamor, H., Pino, J. & Bali, K.) 12881–12888 (Association for Computational Linguistics, 2023).
55. Chen, L. et al. Are more LLM calls all you need? Towards the scaling properties of compound AI systems. In *38th Annual Conference on Neural Information Processing Systems* (NeurIPS, 2024).
56. Mirdita, M. et al. ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
57. Kumar, S., Karuppanan, K. & Subramaniam, G. Omicron (BA.1) and sub-variants (BA.1.1, BA.2, and BA.3) of SARS-CoV-2 spike infectivity and pathogenicity: A comparative sequence and structural-based computational assessment. *J. Med. Virol.* **94**, 4780–4791 (2022).
58. Puccinelli, R. R. et al. Open-source milligram-scale, four channel, automated protein purification system. *PLoS ONE* **19**, e0297879 (2024).
59. Saez, N. J. & Vincentelli, R. in *Structural Genomics: General Applications* (ed. Chen, Y. W.) 33–53 (Humana Press, 2014).
60. Pardon, E. et al. A general protocol for the generation of Nanobodies for structural biology. *Nat. Protoc.* **9**, 674–693 (2014).
61. Byrum, J. R. et al. MultiSero: an open-source multiplex-ELISA platform for measuring antibody responses to infection. *Pathogens* **12**, 671 (2023).
62. Swanson, K., Wu, W., Bulaong, N., Pak, J. & Zou, J. Virtual Lab Data. *Zenodo* <https://doi.org/10.5281/zenodo.15331309> (2025).
63. Swanson, K. Virtual Lab Code. *Zenodo* <https://doi.org/10.5281/zenodo.15320492> (2025).

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

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

**Competing interests** The authors declare no competing interests.

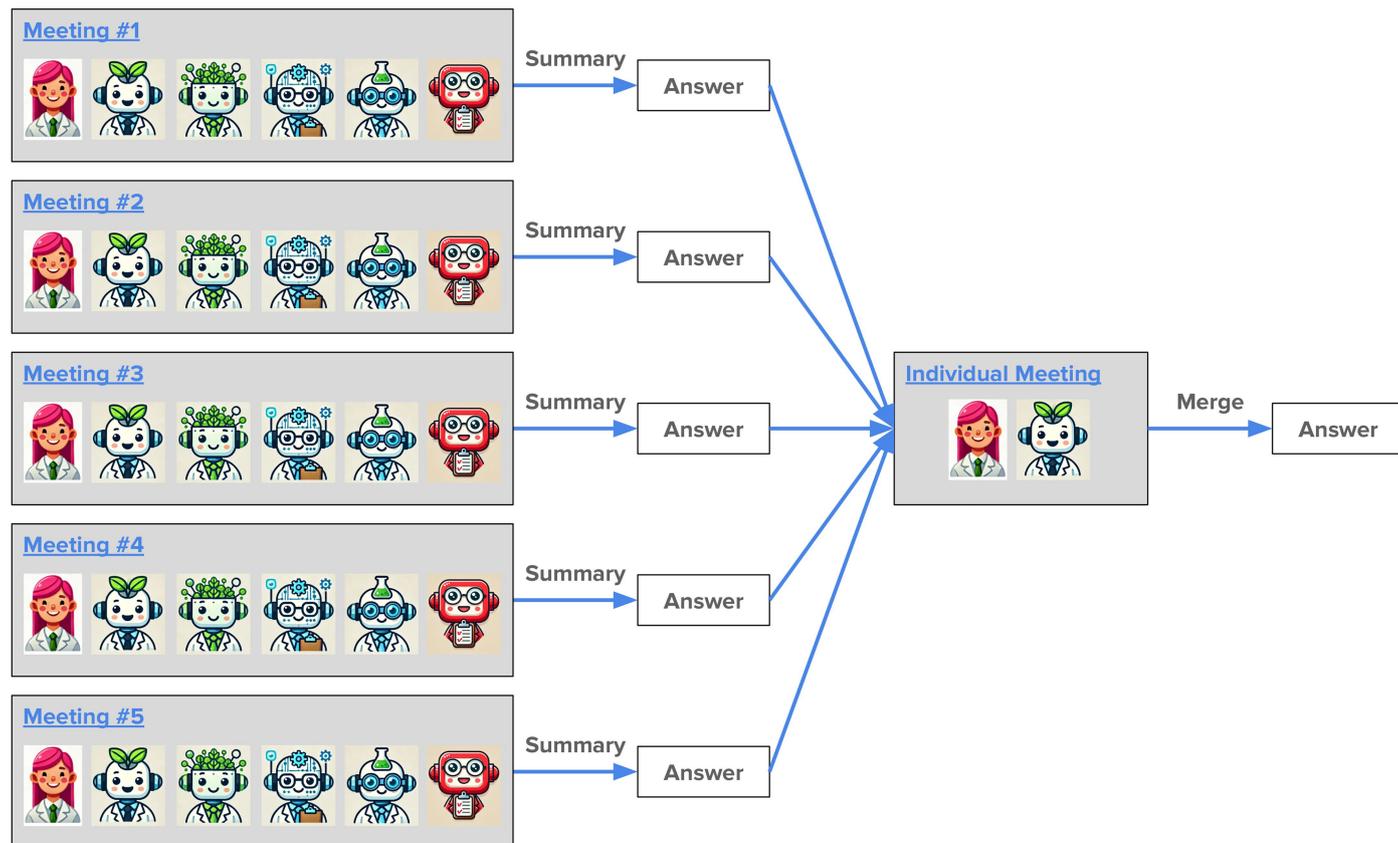
## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41586-025-09442-9>.

**Correspondence and requests for materials** should be addressed to John E. Pak or James Zou.

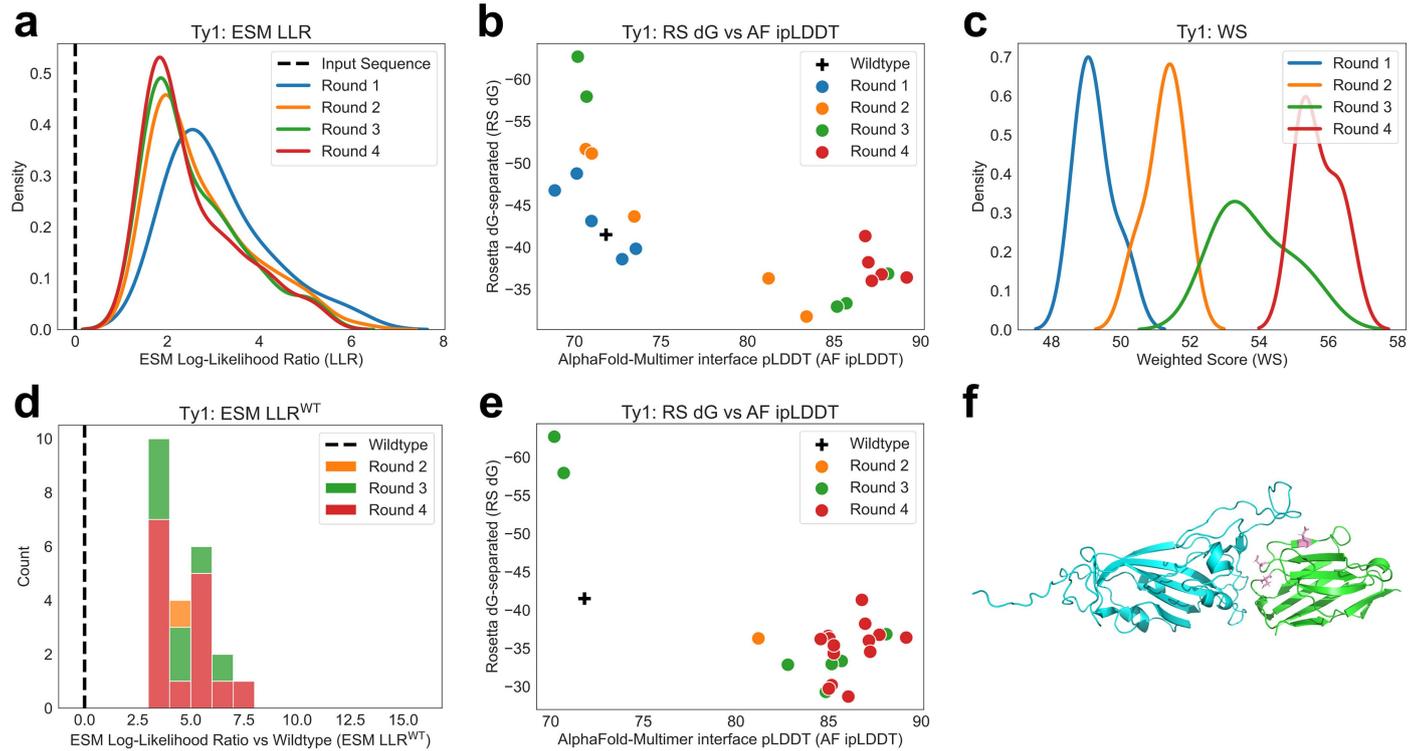
**Peer review information** *Nature* thanks Bryan Briney, Olivier Elemento, Eric Topol and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



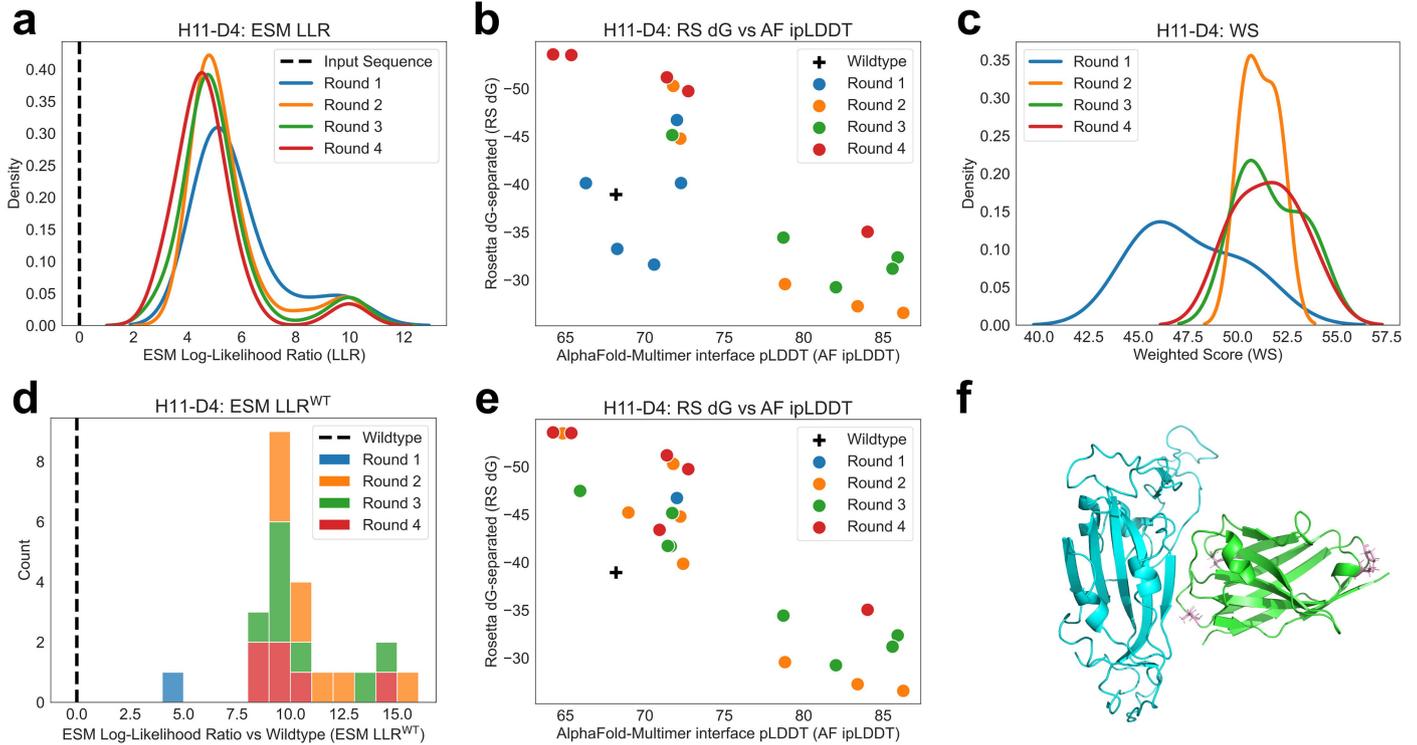
**Extended Data Fig. 1 | Virtual Lab parallel meetings.** The workflow for parallel meetings in the Virtual Lab. A set of meetings (team or individual) is run with the same agenda and agents but with different randomness in the LLM underlying the agents (with a high LLM temperature to encourage creativity across meetings).

The answer from each parallel meeting is then provided to an agent in an individual meeting (with a low LLM temperature for consistency), and this agent is asked to merge the best components of the answers from each parallel meeting into a single optimal answer.



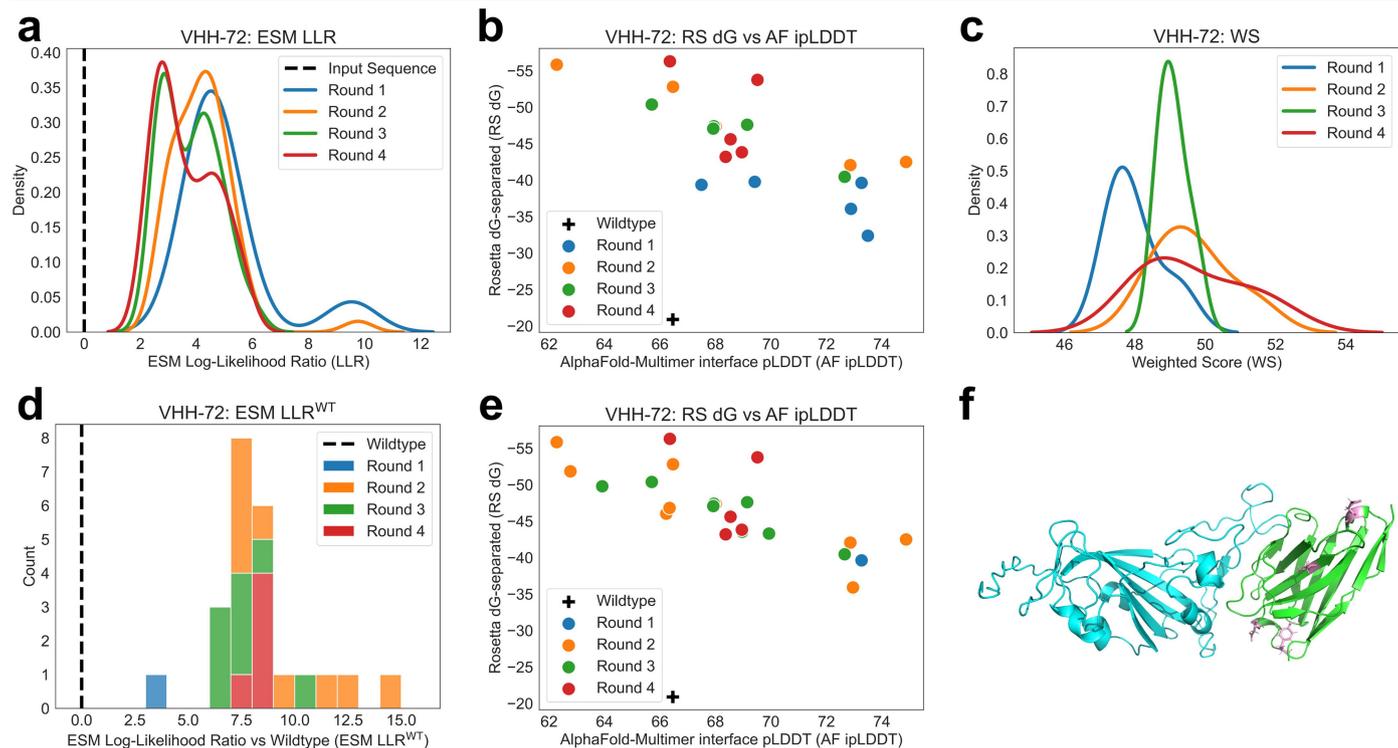
**Extended Data Fig. 2 | Ty1 nanobody analysis.** **a-c**, Evolution of mutant nanobody scores across four rounds of optimization. **a**, The distribution of ESM LLR values for proposed Ty1 mutant nanobodies across each round of optimization, with ESM LLR values computed relative to the input nanobody 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 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.).



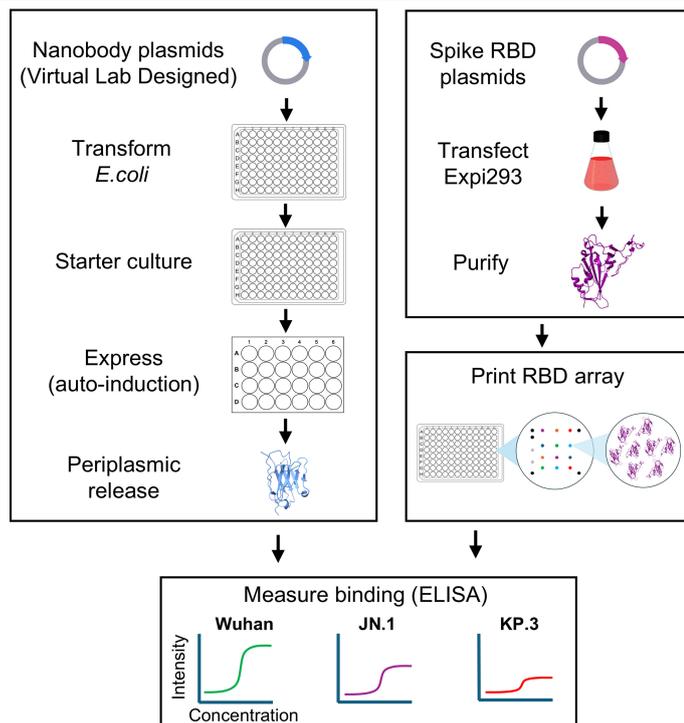
**Extended Data Fig. 3 | H11-D4 nanobody analysis.** **a-c**, Evolution of mutant nanobody scores across four rounds of optimization. **a**, The distribution of ESM LLR values for proposed H11-D4 mutant nanobodies across each round of optimization, with ESM LLR values computed relative to the input nanobody 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 4 Å of the opposite chain). Mutant nanobody residues are in pink. (PyMol 3.1.3, Schrödinger, LLC.).



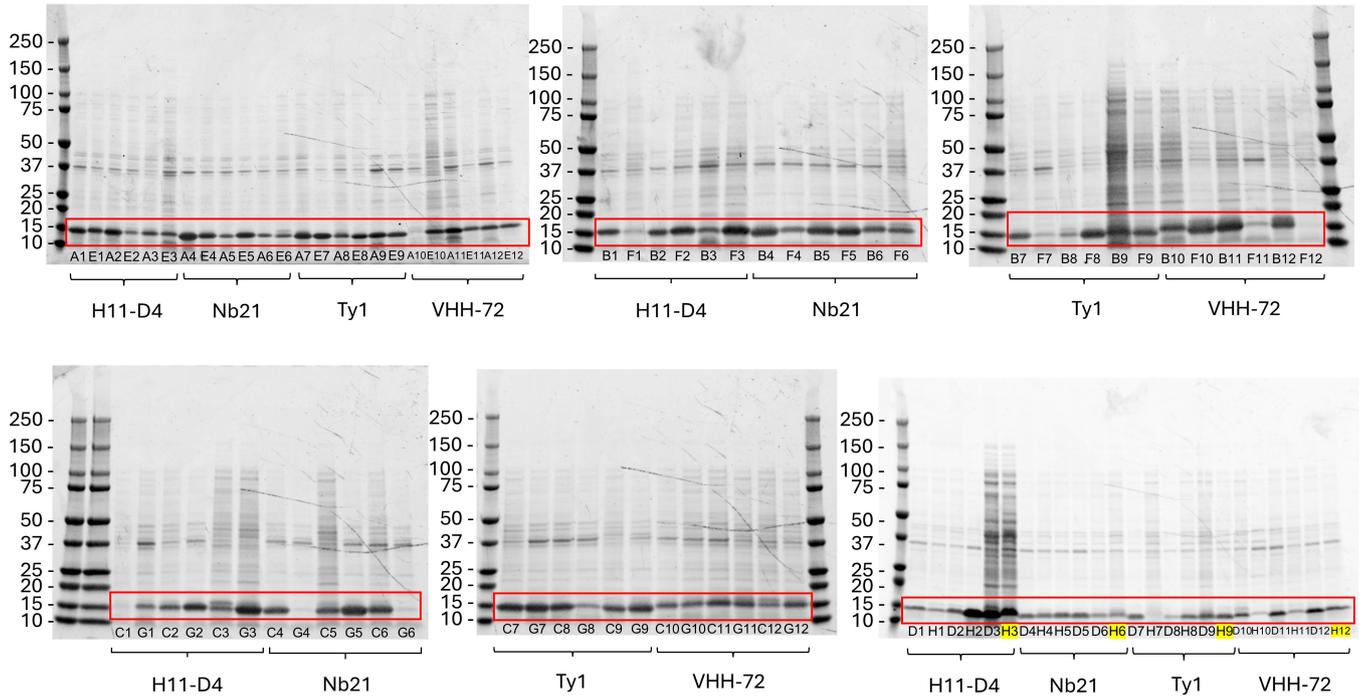
**Extended Data Fig. 4 | VHH-72 nanobody analysis.** **a-c**, Evolution of mutant nanobody scores across four rounds of optimization. **a**, The distribution of ESM LLR values for proposed VHH-72 mutant nanobodies across each round of optimization, with ESM LLR values computed relative to the input nanobody 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 VHH-72 R27Y-E31D-F37V-D89E (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.).

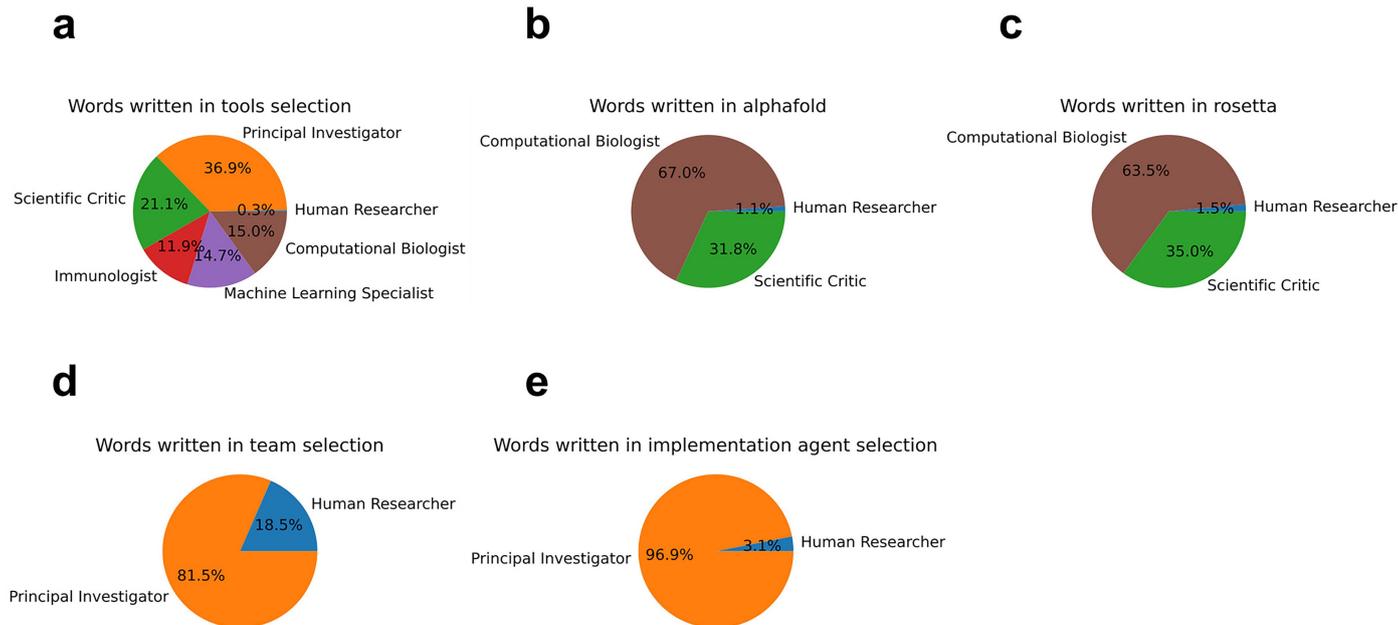


**Extended Data Fig. 5 | Workflow for nanobody experimental validation.**

The four categories of experiments (nanobody expression, SARS-CoV-2 spike RBD expression, antigen array printing, and multiplexed ELISA) are enclosed in boxes. The ribbons representation of a nanobody (blue) and the RBD (purple) were rendered with ChimeraX<sup>51</sup> from PDB accession numbers 6XZN and 6M0J, respectively. Unique RBD and control proteins of the array are shown as colored spots with fiducial markers shown as black spots. Portions of this figure were created in BioRender. Bulaong, N. (2025) <https://BioRender.com/6du2yu4>.



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



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

of words written by the Virtual Lab in Rosetta implementation. **d**, The number of words written by the Virtual Lab in the team selection phase. **e**, The number of words written by the Virtual Lab in implementation agent selection.

# Article

## Extended Data Table 1 | Nanobody score analysis

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

The scores of each wild-type nanobody and examples of the mutant nanobodies that were selected for experimental validation. ESM LLR<sup>WT</sup>: ESM log-likelihood ratio between the mutant nanobody sequence and the wild-type sequence. AF ipLDDT: AlphaFold-Multimer interface pLDDT for the nanobody-spike complex. RS dG: Rosetta dG-separated binding energy value. WS<sup>WT</sup>: Weighted score combining ESM LLR<sup>WT</sup>, AF ipLDDT, and RS dG.

## 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  |
|-------------------------------------|--|
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided<br><i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of all covariates tested  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> 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) |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. $F$ , $t$ , $r$ ) with confidence intervals, effect sizes, degrees of freedom and $P$ value noted<br><i>Give <math>P</math> values as exact values whenever suitable.</i>                            |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> 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 <a href="https://github.com/zou-group/virtual-lab">https://github.com/zou-group/virtual-lab</a> and on Zenodo at <a href="https://doi.org/10.5281/zenodo.15320491">https://doi.org/10.5281/zenodo.15320491</a> . 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 <a href="https://github.com/zou-group/virtual-lab">https://github.com/zou-group/virtual-lab</a> and on Zenodo at <a href="https://doi.org/10.5281/zenodo.15320491">https://doi.org/10.5281/zenodo.15320491</a> . 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	<input type="text" value="Our study does not involve human participants or their data, so this is not applicable."/>
Reporting on race, ethnicity, or other socially relevant groupings	<input type="text" value="Our study does not involve human participants or their data, so this is not applicable."/>
Population characteristics	<input type="text" value="Our study does not involve human participants or their data, so this is not applicable."/>
Recruitment	<input type="text" value="Our study does not involve human participants or their data, so this is not applicable."/>
Ethics oversight	<input type="text" value="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	<input type="text" value="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	<input type="text" value="No data was excluded from the analyses."/>
Replication	<input type="text" value="ELISA binding assays were run in duplicates and the mean value was reported."/>
Randomization	<input type="text" value="Randomization is not applicable to this study because every nanobody was tested in the same manner."/>
Blinding	<input type="text" value="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 &amp; experimental systems

## Methods

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

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 (MKYLLPTAAAGLLLLAAQPAMA), a C-terminal 6x his tag, and a stop codon, and they were synthesized and cloned into pET-29b(+) (Twist Biosciences). anti-Alpaca IgG VHH 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 $\mu$ L 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 <a href="#">ICLAC</a> 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.