AIntibody: An experimentally-validated *in silico* antibody discovery design challenge

M. Frank Erasmus[1*], Laura Spector[1], Katheryn Perea-Schmittle[1], Fortunato Ferrara[1], Roberto DiNiro[1], Christina Pettus[1], Luis Antonio Rodriguez Carnero[2], Jianquan Li[1], Alberto Cisneros[1], Adeline Fanni[1], Thomas J. Pohl[1], David Knight[1], Zachary Cruz[1], Esteban Molina[1], Wei Wang[3], Peter M. Tessier[4], Crystal Richardson[5], Laure Turner[5], Sumit Kumar[5], Daniel Bedinger[6], Pietro Sormanni[7], Monica L. Fernández-Quintero[8], Andrew B. Ward[8], Johannes R. Loeffler[8], Olivia M. Swanson[8], Charlotte M. Deane[9] , Matthew I. J. Raybould[9], Andreas Evers[10], Carolin Sellmann[10], Sharrol Bachas[11], Jeff Ruffolo[12], Horacio Nastri[13], Karthik Ramesh[13], Jesper Sørenson[14], Rebecca Croasdale-Wood[15], Oliver Hijano[16], Melody Shahsavarian[16], Yu Qiu[16], Paolo Marcatili[17], Erik Vernet[17], Rahmad Akbar[17], Simon Friedenson[18], Tzvika Hartman[19], Vinodh babu Kurella[20], Shipra Malhotra[20], Satyendra Kumar[20], Patrick Kidger[21], Juan C. Almagro[22], Eric Furfine[23], Marty Stanton[23], Santiago David Villalba[24], Florian Tomszak[24], Molly Dovner[25], Sara D'Angelo[1], Andrew R. M. Bradbury[1*]

[1]Specifica LLC, a Q[2] Solutions Company, Santa Fe, USA; [2]New Mexico Consortium, Los Alamos, USA; [3]University of California, San Diego, USA; [4]University of Michigan, Ann Arbor, USA; [5]GENEWIZ from Azenta Life Sciences, South Plainfield, USA; [6]Carterra®, Salt Lake City, USA; [7]University of Cambridge, Cambridge, United Kingdom; [8]Scripps Health, San Diego CA; [9]University of Oxford, Oxford, United Kingdom; [10]Merck KGaA, Darmstadt, Germany; [11]Evolutionary Scale, San Francisco, USA; [12]Profluent Bio, Berkeley, USA; [13]Incyte, Inc., Wilmington, USA; [14]OpenEye Scientific, Cadence Molecular Sciences, Santa Fe, USA. [15]Astrazeneca, Cambridge, United Kingdom; [16]Sanofi, Inc., Cambridge, USA; [17]Novo Nordisk, A/S, Bagsværd, Denmark; [18]Alloy Therapeutics, Waltham, USA & Basel, Switzerland; [19]Biolojic Design, Rehovot, Israel; [20]Takeda, Inc., Cambridge, USA; [21]Cradle Bio, Zürich, Switzerland; [22]GlobalBio, Inc., Cambridge, MA, USA; [23]Mosaic Biosciences, Boulder, CO, USA.; [24]Bayer AG Leverkusen, Germany; [25]Sapidyne, Inc., Boise, USA

*To whom correspondence should be directed.

Abstract [Scope of Competition]

Herewith we announce the **AI**ntibody competition to validate the performance of computational models, including those based on artificial intelligence, to generate antibody candidates. Teams are invited to submit antibody sequences in response to the challenges described below. Submitted sequences will be synthesized as proteins and independently benchmarked in a wet lab. The long-term goal is to produce an ongoing challenge series analogous to the protein structural prediction competition, CASP, which famously led to the recognized triumph of AlphaFold in 2018. As the competition evolves, challenges are anticipated to become progressively more difficult, and each will be followed by a peer-reviewed manuscript to provide the community with clarity on the results, highlighting areas for improvement and satisfactory capabilities.

The Importance of Benchmarking

Science is frequently subject to the Gartner hype cycle[1]: emergent technologies spark intense initial enthusiasm with the recruitment of dedicated scientists. As limitations are recognized, disillusionment often sets in: some scientists turn away, disappointed in the inability of the new technology to deliver on initial promise, while others persevere and further develop the technology. While the value (or not) of a new technology usually becomes clear with time, appropriate benchmarks can be invaluable in highlighting strengths and areas for improvement, substantially speeding up technology maturation. A particular challenge in computational engineering and AI/ML [artificial intelligence/machine learning] is that such benchmarks and best practices are uncommon in the literature making it particularly hard for non-experts to assess the impact and performance of these methods. While multiple papers have highlighted best practices and evaluation guidelines[2-4], the true test for such methods is ultimately prospective performance, which requires experimental testing.

In the 1990's several groups were attempting to predict the structure of proteins from their amino acid sequences, and the success and value of different models was assessed *ad hoc*. A biannual competition

known as CASP (Critical Assessment of Structure Prediction)[5] was established in 1994 to compare the performance of various algorithms. In this competition, teams that determine protein structures using experimental methods (predominantly X-ray crystallography and NMR) withhold their structures, supplying protein sequences and other pertinent details to the modeling community, via the predictioncenter.org website. Groups participate as expert teams, utilizing a combination of human expertise and computational methods, or as fully automated methods conducted on servers. All predicted structures are then compared to the previously determined experimental structure by an independent panel, blinded to the authors of the various predicted models under assessment. Submission deadlines for models are set at three weeks for expert groups and three days for servers, with each team allowed to provide up to five models per target, and the primary model from each group being given priority in the independent assessment. Widely viewed as the "protein structure prediction world championship", there were incremental improvements in structural predictions until AlphaFold[6] won dramatically in 2018[7] and 2020[8]. Although AlphaFold has not competed since, all methods these days are inspired by the AlphaFold architecture, demonstrating conclusively the power of deep learning algorithms for protein structure prediction, and the value of CASP in conclusively demonstrating AlphaFold's superiority. Notably other competitions aimed particularly at predicting antibody structure from sequence[9,10] were also initiated but have not been held since 2014, demonstrating the broad efforts and inherent challenges in sustaining such initiatives.

Fast forward thirty years, and the reasons for initiating the CASP competition – a desire to understand the capabilities of new technologies – can now be found in the application of AI/ML and other *in silico* approaches to the development and improvement of proteins[11], particularly antibody therapeutics[12-14]. Numerous companies and academic groups claim AI/ML solutions to affinity maturation, antibody developability and *de novo* antibody and library design. However, significant challenges lie in understanding the value of these algorithms: how they differ from one another in performance, how the quality of predicted antibodies compares to existing experimental practices, how effective they are in generating antibodies recognizing particular epitopes (including against experimentally more challenging targets including those with membrane, glycan, or flexible components), how generalizable they are, and particularly whether they are able to provide antibodies with the desired properties more rapidly than experimental approaches. Most results are based on retrospective studies (i.e. without conducting new experiments) and without making data accessible. This imposes severe limitations in the ability of the broader scientific community to validate such claims and therefore the real-world impact of these novel methods. Machine learning methods can perform very differently - both, better or worse - in prospective studies compared to retrospective ones, as has been demonstrated, for example, on docking into AlphaFold2 generated structures[15]. For this reason, prospective evaluations of new problems are essential to assess real-world performance. These difficulties emphasize the importance of a public and prospective benchmarking effort to provide a quantifiable and unbiased assessment of these techniques.

Introduction to the Competition

Here we propose the launch of an AI/ML benchmarking exercise (named **AI**ntibody and accessed via the eponymous **AI**ntibody.org website) that will present a series of escalating AI/ML challenges in antibody discovery, calibrated according to the success of the outcome of previous competitions. The challenges in this first **AI**ntibody competition will be based on experimental datasets generated by some of the authors and made available to participants and the broader scientific community. As the responses are designed to test the present immediate value of AI/ML algorithms, participants will have fourteen days from data access to provide sequence solutions. However, the structure of the supplied data can be found in **Table 1-2** allowing participants ample time to prepare in advance. This will prevent experimental validation, which will be carried out independently and conducted blindly by third parties.

In addition to CASP, this competition has similarities to CACHE (Critical Assessment of Computational Hit-finding Experiments)[16], D3R (Drug Design Data Resource) Grand Challenge[17], and SAMPL (Statistical Assessment of Modeling of Proteins and Ligands)[18]. CACHE is an extensive public-private partnership benchmarking initiative recently established to enable the development of computational methods for drug hit-finding and addressing the problem that no "algorithm can currently select, design or rank potent drug-like small-molecule protein binders consistently". The D3R Grand Challenge focuses on (small molecule) ligand-protein pose and binding affinity prediction, while SAMPL conducts blind prediction challenges in computational drug discovery, with an emphasis on binding modes, affinities, and physical properties for small molecules. However, unlike these small molecule competitions, this contest is focused on antibodies and utilizes different organizing principles. Given the urgent need for benchmarking in AI/ML mediated antibody discovery, this first competition is an unfunded *ad hoc* contest based on datasets generated by a single party against the RBD of SARS-CoV-2, with participants paying a significantly discounted fee [$120/antibody] to cover the costs of gene synthesis and expression of each submitted antibody sequence. We anticipate the *ad hoc* nature of this first competition will provide deeper understanding and time to organize future contests by more formal partnerships or societies, involving steering committees comprising academic and industry scientists, ideally with external funding.

Testing antibodies for biological activity is as complex as the testing of small drugs. However, the challenges presented here in this inaugural competition are relatively straightforward, involving only affinity and developability (**Table 3**), the subjects of many AI/ML state-of-the-art performance claims and among the most important determinants of biological activity. Results will be published blinded. Although not obligatory, participants will be invited to be co-authors, akin to the approach taken recently when diverse antibodies against SARS-CoV-2 were analyzed[19]. Participants will be able to see the range of results and their ranking within that range, but individual performances will remain masked. For each challenge described below, the focus lies in engineering *only the CDRs* (IMGT definition for all but LCDR2, for which Kabat is used). For example, frameworks of the antibodies provided in the datasets should not be modified within the context of the challenge. While *in vivo* affinity maturation is focused on CDRs, framework mutations are often also introduced. The challenges introduced here limit available diversity space to CDRs, providing a direct comparison to experimental methods[20]. However, future challenges may allow the introduction of framework mutations.

This benchmarking exercise will answer the question: what are AI/ML capabilities as applied to antibody discovery at this moment. This competition will provide insight to aid in the establishment of realistic timelines for future utility of AI/ML in antibody discovery.

Competition Details and Guidelines

The challenges proposed in this inaugural competition are based on unpublished SARS-CoV-2 next-generation sequencing (NGS) datasets from which some antibodies were characterized[21,22]. Given the vast amount of additional public data available for SARS-CoV-2 binding antibodies (e.g., COVIC[19] & Cov-AbDab[23]), the following should provide the best possible scenario for AI/ML task success.

Once designed/identified sequences have been uploaded, Azenta will synthesize genes for up to 1000 antibodies, express and purify them, carry out size exclusion chromatography and provide coded antibodies to other partners. Antibody affinities will be assessed by Carterra using surface plasmon resonance (Carterra LSA-XT) and Mosaic will assess developability (HIC HPLC, BVP ELISA, AC-SINS, Tm and Tagg). Bio-Techne will provide the target. These assays will ensure standardized conditions and unbiased head-to-head comparison of predicted sequences.

Competition #1 – In Silico* Antibody Affinity Maturation*: Participants will be provided with experimental datasets derived from the affinity maturation of an antibody recognizing the RBD of SARS-CoV-2 (**Table 1**). Each dataset comprises three NGS sub-datasets of CDR sequences (LCDR1+2; LCDR3 and HCDR1+2 in which the remaining CDRs are constant) generated during phase 1 of a previously described experimental affinity maturation method[20] (**Figure 1**), in which phase 2 involves combining all phase 1 outputs and experimentally selecting for higher affinity. Each antibody population has diversity only in the indicated CDRs (**Figure 1a**), the remaining CDRs being parental, and has been displayed on yeast and sorted for target binding (**Figure 1b**). While each population binds the target more tightly than the parental, individual NGS sequences have not been assessed for their ability to encode antibodies with improved binding activity and *may include PCR or sequencing errors*. While amino acid sequences of phase 2 characterized antibodies with their affinities have been determined (**Figure 1c-d**), this data will not be provided, in order that computational methods are given the opportunity to generate the same sequences.

The computational goal is to design antibodies with improved affinity for the RBD of SARS-CoV-2 using the NGS datasets that also exhibit favorable developability properties. Designs should only be applied to HCDR1-2 and LCDR1-3 and not frameworks or HCDR3. The blinded assessment will determine the affinities of designed antibodies and how well they compare to those obtained experimentally in phase 2. While the experimental affinity maturation was carried out on scFvs displayed on yeast, antibodies were, and will be, tested as full length IgGs. Results will be compared to the affinities of experimentally derived sequences generated by combining the three phase 1 outputs and selecting from the corresponding combinatorial library.

***Competition #2 – In Silico* Affinity Rank Prediction for Antibody Discovery**: Participants will be provided with an NGS dataset of a single selection output recognizing the RBD of SARS-CoV-2, clustered by HCDR3 sequence[22] using a previously published library[24], which comprises natural CDRs embedded within well-behaved therapeutic scaffolds. While not all individual NGS sequences have been assessed for their ability to encode antibodies with binding activity, and therefore *may include PCR or sequencing errors*, those sequences encoding antibodies (as IgG) demonstrated to bind the target will be identified/provided, with their corresponding affinities. Furthermore, the relative frequency of different sequences within the clusters will be provided, see **Table 2** for representative dataset.

The computational goal is to identify those sequences within the existing NGS dataset that encode the highest affinity antibodies (that have not already had their affinities determined) in the two largest HCDR3 clusters, by largest number of VL+VH sequences, from experimental bin group #1 [i.e., 28F and 27F] and the largest cluster from experimental bin group #2 [i.e., 47F] (**Table 4**). Results will be compared to the affinities of those antibodies that were experimentally derived and for which sequences were provided.

***Competition #3* – NGS Inspired Computational Antibody Design:** Given the same NGS output in **competition #2**:

The AI/ML goal is to generate *out-of-library* sequences of antibodies binding the same target with as high affinities as possible that also exhibit favorable developability properties, using the provided NGS datasets described here and any other useful publicly available data. Only CDRs should be designed, and frameworks should remain unmodified (see **Table 2** for the format). Such out-of-library sequences should not be present within the NGS dataset itself, nor be derived from any previous independent experiments.

Results will be compared to the affinities of those antibodies that were experimentally derived and for which sequences were provided.

While each of these challenges have high affinity as an endpoint, developability will also be assessed to ensure affinity is not generated at the expense of developability. Antibodies will be judged as passing or failing on each of the five developability assays described above, and antibodies will be required to pass three or more assays.

Follow-up Results Publication and Participation Requirements

Participants agree to make the *winning* algorithm (from each of the three competitions) publicly available, in the form of a paper describing the algorithmic details. This may be done anonymously as part of the follow-up manuscript from the competition organizers, or (to avoid autoplagiarism) as part of a separately-authored standalone manuscript. In line with Nature's code release policy[25], authors are encouraged but not required to additionally open-source their code. All participants (anonymously, but regardless of winning position) also agree to provide at least a single-sentence description of their methods (e.g., 'protein language model based on ESM2 and fine-tuned on antibody database XYZ'). By contributing, participants agree to this requirement for inclusion in the follow-on manuscript.

How to Participate?

To participate in **AI**ntibody, please follow registration instructions at **AI**ntibody.org. Participants may compete in one or more challenges. Registration will open on the publication date [TBD publication date] and close thirty days later or as soon as the total number of proposed sequences to be submitted reaches 1000. The datasets will be released to participants thirty days after publication and to the community upon publication of the results paper.

Future Competitions

We anticipate this to be the first of a series of **AI**ntibody competitions, each comprising a set of challenges of increasing complexity that will provide opportunities for scientists to test, compare and improve their AI/ML models. Future contests are expected to evolve in complexity depending upon the outcome of this and other competitions, in parallel with the development of AI/ML use in antibody discovery and design. In follow-on competitions anticipated below, the term “antibodies” refers to VHH’s or antibodies, each of which will be separate competitions. To maintain standardization between competitions, challenges 2 & 3 will be retained using a different target, and NGS selection output, each time:

1. Given an antibody sequence with known affinity to a target, generate candidates with improved affinities.
2. Given an antibody sequence with known affinity to a target, with poor developability properties (e.g., thermal stability, polyreactivity), generate candidates with equal or better affinities that lack the poor developability issues.
3. Given the sequence of a target with known structure, and structurally similar targets in the PDB, generate specific antibodies binding to the target.
4. Given the sequence of a target with unknown structure, but similar targets in the PDB, generate specific antibodies binding to that target.
5. Challenges 2-4 above but applied to a specific epitope.
6. Given a set of antibody sequences known to bind distinct epitopes on a given target, predict which antibodies bind to different epitopes.
7. Given a set of antibody sequences known to bind distinct epitopes on a given target, predict which antibodies will allow sandwich binding.

8. Given the sequence of a target, and a specific epitope within that target, generate *de novo* antibodies binding to that specific epitope.
9. Given a set of diverse antibodies, predict their epitopes.
10. Given the sequence of a target with known structure, generate antibodies binding to that target.
11. Given an antibody sequence to a known human target epitope, create antibody designs that are cross-reactive to cynomolgus monkey and mouse with epitopes identical or near parental human epitope.
12. Create a *de novo* or rationally guided antibody designs which do not bind any known targets (isotype controls).

Acknowledgements

We thank Leonard Wossnig of LabGenius for his generous insights and contributions to the manuscript. We also thank Bio-Techne for their gracious supply of the RBD antigen for this Benchmarking Study.

References

1 Linden, A. & Fenn, J. Understanding Gartner's hype cycles. *Strategic Analysis Report Nº R-20-1971. Gartner, Inc* **88**, 1423 (2003).

2 Lewis Chinery, A. M. H., Brij Bhushan Mehta, Rahmad Akbar, Puneet Rawat, Andrei Slabodkin, Khang Le Quy, Fridtjof Lund-Johansen, Victor Greiff, Jeliazko R. Jeliazkov, Charlotte M. Deane. Baselining the Buzz Trastuzumab-HER2 Affinity, and Beyond. *bioRxiv* (2024). https://doi.org:https://doi.org/10.1101/2024.03.26.586756

3 Wossnig, L., Furtmann, N., Buchanan, A., Kumar, S. & Greiff, V. Best practices for machine learning in antibody discovery and development. *Drug Discov Today* **29**, 104025 (2024). https://doi.org:10.1016/j.drudis.2024.104025

4 Bender, A. *et al.* Evaluation guidelines for machine learning tools in the chemical sciences. *Nat Rev Chem* **6**, 428-442 (2022). https://doi.org:10.1038/s41570-022-00391-9

5 Moult, J. The current state of the art in protein structure prediction. *Curr Opin Biotechnol* **7**, 422-427 (1996). https://doi.org:10.1016/s0958-1669(96)80118-2

6 Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706-710 (2020). https://doi.org:10.1038/s41586-019-1923-7

7 Senior, A. W. *et al.* Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins* **87**, 1141-1148 (2019). https://doi.org:10.1002/prot.25834

8 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021). https://doi.org:10.1038/s41586-021-03819-2

9 Almagro, J. C. *et al.* Antibody modeling assessment. *Proteins* **79**, 3050-3066 (2011). https://doi.org:10.1002/prot.23130

10 Almagro, J. C. *et al.* Second antibody modeling assessment (AMA-II). *Proteins* **82**, 1553-1562 (2014). https://doi.org:10.1002/prot.24567

11 Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089-1100 (2023). https://doi.org:10.1038/s41586-023-06415-8

12 Hie, B. L. *et al.* Efficient evolution of human antibodies from general protein language models. *Nat Biotechnol* **42**, 275-283 (2024). https://doi.org:10.1038/s41587-023-01763-2

13 Harvey, E. P. *et al.* An in silico method to assess antibody fragment polyreactivity. *Nat Commun* **13**, 7554 (2022). https://doi.org:10.1038/s41467-022-35276-4
14 Li, L. *et al.* Machine learning optimization of candidate antibody yields highly diverse sub-nanomolar affinity antibody libraries. *Nat Commun* **14**, 3454 (2023). https://doi.org:10.1038/s41467-023-39022-2
15 Lyu, J. *et al.* AlphaFold2 structures template ligand discovery. *bioRxiv* (2024). https://doi.org:10.1101/2023.12.20.572662
16 Ackloo, S. *et al.* CACHE (Critical Assessment of Computational Hit-finding Experiments): A public-private partnership benchmarking initiative to enable the development of computational methods for hit-finding. *Nat Rev Chem* **6**, 287-295 (2022). https://doi.org:10.1038/s41570-022-00363-z
17 Gathiaka, S. *et al.* D3R grand challenge 2015: Evaluation of protein-ligand pose and affinity predictions. *J Comput Aided Mol Des* **30**, 651-668 (2016). https://doi.org:10.1007/s10822-016-9946-8
18 Guthrie, J. P. A blind challenge for computational solvation free energies: introduction and overview. *J Phys Chem B* **113**, 4501-4507 (2009). https://doi.org:10.1021/jp806724u
19 Hastie, K. M. *et al.* Defining variant-resistant epitopes targeted by SARS-CoV-2 antibodies: A global consortium study. *Science* **374**, 472-478 (2021). https://doi.org:10.1126/science.abh2315
20 Teixeira, A. A. R. *et al.* Simultaneous affinity maturation and developability enhancement using natural liability-free CDRs. *MAbs* **14**, 2115200 (2022). https://doi.org:10.1080/19420862.2022.2115200
21 Ferrara, F. *et al.* A pandemic-enabled comparison of discovery platforms demonstrates a naive antibody library can match the best immune-sourced antibodies. *Nat Commun* **13**, 462 (2022). https://doi.org:10.1038/s41467-021-27799-z
22 Erasmus, M. F. *et al.* Insights into next generation sequencing guided antibody selection strategies. *Sci Rep* **13**, 18370 (2023). https://doi.org:10.1038/s41598-023-45538-w
23 Raybould, M. I. J., Kovaltsuk, A., Marks, C. & Deane, C. M. CoV-AbDab: the coronavirus antibody database. *Bioinformatics* **37**, 734-735 (2021). https://doi.org:10.1093/bioinformatics/btaa739
24 Azevedo Reis Teixeira, A. *et al.* Drug-like antibodies with high affinity, diversity and developability directly from next-generation antibody libraries. *MAbs* **13**, 1980942 (2021). https://doi.org:10.1080/19420862.2021.1980942
25 *Reporting standards and availability of data, materials, code and protocols*, <https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards#availability-of-computer-code>