Machine learning-guided antibody optimization of a SARS-CoV-2 binding antibody enabled by AlphaSeq

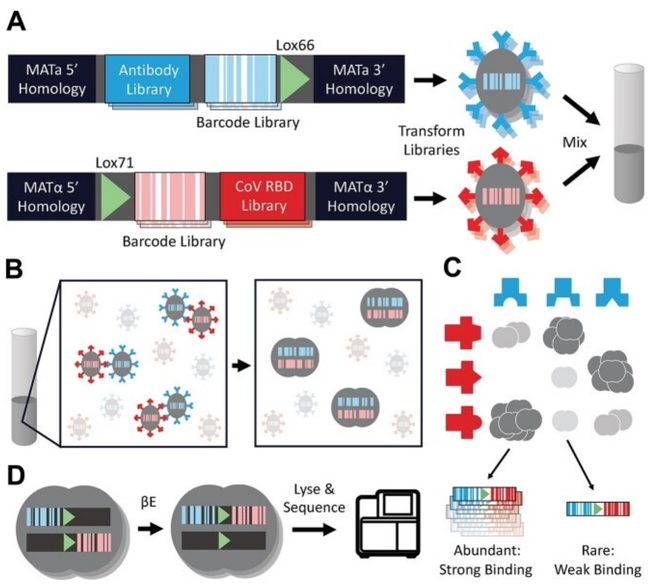
Emily Engelhart[1]; Christof Angermueller[3]; Ben Jester[2]; Zelda Mariet[3]; Ryan Emerson[1]; Davis Goodnight[1]; Kyle Minch[1]; Kim Wellman[1]; Randolph Lopez[1]; Mike Frumkin[3]; Lucy Colwell[3]; David Younger[1]

[1]A-Alpha Bio, Inc. Seattle, WA; [2]Lumen Bioscience, Inc. Seattle, WA; [3]Google Research

Abstract

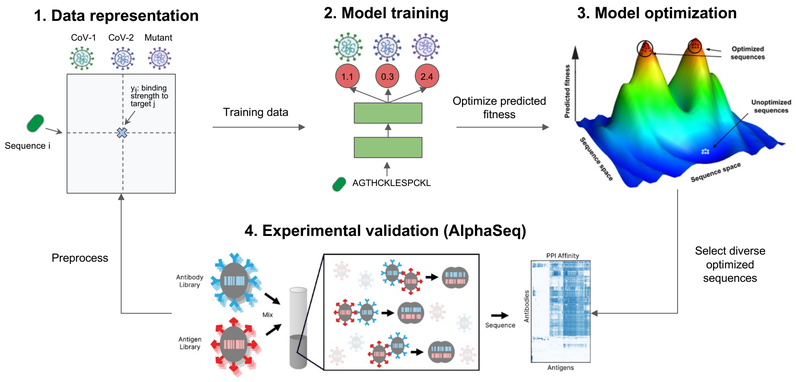
Rapid discovery and development of therapeutic antibodies to new pathogens and pathogen variants is crucial for combatting infectious diseases. An off-the-shelf computational model to generate evolved candidates from a starting antibody would save months by eliminating wet-lab antibody discovery. In collaboration with Google Research and Lumen Bioscience, we developed a combined experimental and machine learning (ML) framework for optimizing a well-described anti-SARS-CoV antibody (VHH-72) to recognize diverse variants of SARS-CoV-2. We leverage the AlphaSeq platform, using yeast synthetic biology to measure millions of protein interactions on a library-on-library scale, to train and validate an ML model for antibody sequence proposals. Starting from a parental antibody with moderate affinity to SARS-CoV-2, we improve its affinity by more than 50-fold while maintaining cross-reactivity to SARS-CoV. We confirm that >90% of top hits have improved affinity by biolayer interferometry (BLI) and therapeutic potency by pseudovirus neutralization. Top hits are sequence diverse, containing up to 8 mutations from the parental sequence, and have a range of recombinant expression yields. Furthermore, we observe that many top hits show significant cross-reactivity to the SARS-CoV-2 Omicron variant, even though it was not included in the original training data. After 3 iterations of AlphaSeq + ML, model generated sequences outperform non-ML baseline sequences (combinatorial recombination of the best performing sequences) by up to 4-fold.

AlphaSeq Platform



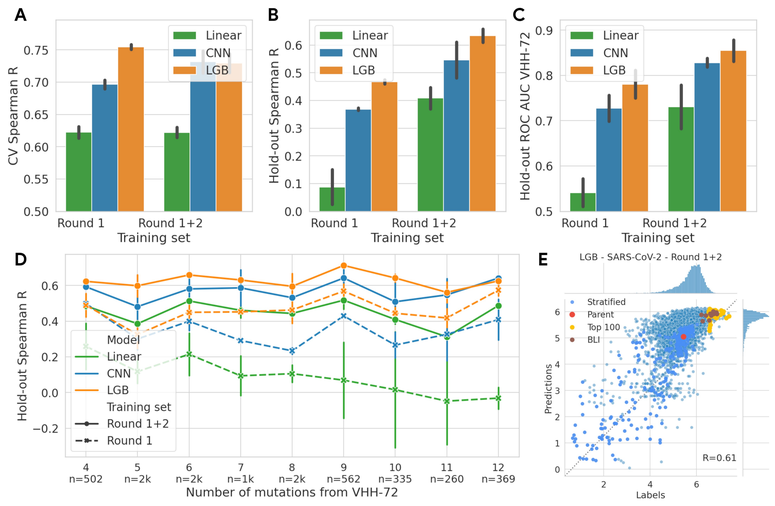
AlphaSeq uses synthetic biology and next generation sequencing to measure protein-protein interactions at a library-on-library scale. **(A)** Two libraries are built that each contain protein sequences for display on the yeast cell surface, randomized DNA barcodes, and a recombination site. **(B)** Libraries are mixed in liquid culture and interactions between surface displayed proteins drives agglutination and cell fusion. **(C)** The number of fused cells with a given protein pair is dependent on protein interaction strength. **(D)** Recombination is induced with β-estradiol to consolidate DNA barcodes. Cells are then lysed and sequenced to count the abundance of each barcode pair and determine all protein interaction strengths.

Iterative Machine Learning Guided Antibody Optimization



An initial dataset of VHH sequences and experimental binding measurements against CoV variants is the input for ML model training. The model consists of two components, a regressor and classifier, and predicts binding strengths given input sequences. The classifier predicts whether an input VHH sequence binds measurably to a specific SARS-CoV target, with 1 indicating “yes” and 0 indicating “no.” The regressor predicts a quantitative binding strength. The model output is the product of the classifier and regressor predictions. Model optimization is performed *in silico* by searching for sequences with high predicted affinities. A diverse subset consisting of 10,000s of optimized sequences are selected and validated experimentally using the AlphaSeq platform. The resulting experimental measurements are preprocessed and used as additional training data for subsequent rounds of optimization.

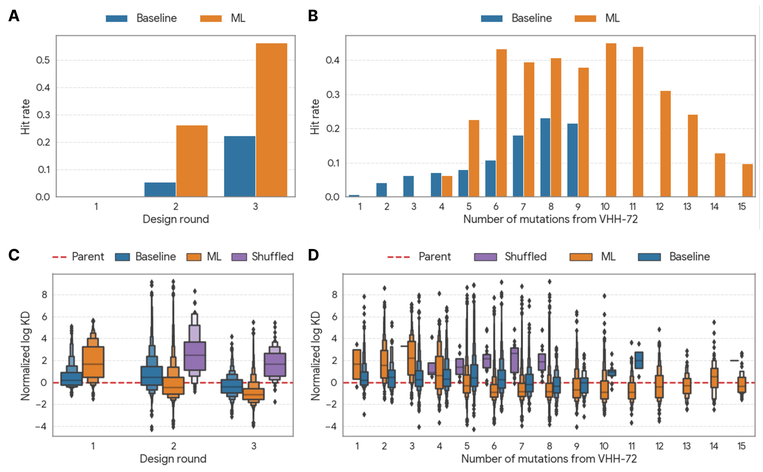
Validation: ML Models



(A) 5-fold cross-validation performance quantified by Spearman's R of models trained on *only sequences of round 1* or *sequences of both round 1 and round 2*. Due to the high performance of the LGB model, only that model was used to design sequences of round 2. LGB and CNN models were used to design sequences of round 3. Error bars show the variation over the eight CoV targets that were assayed in both round 1 and round 2. **(B)** Spearman’s R, computed on hold-out sequences of round 3, of models trained on *only sequences of round 1* or *sequences of both round 1 and round 2*. **(C)** Hold-out ROC AUC score to correctly classify sequences that bind stronger than the parent sequence. **(D)** Hold-out Spearman's R stratified by the number of mutations from VHH-72. **(E)** Scatterplot of model predictions vs. ground truth labels for SARS-CoV-2 of the LGB model trained on round 1 and round 2. Each dot corresponds to one hold-sequence assayed in round 3 that was not used for model training. Labels and predictions correspond to binding affinities (higher is stronger) obtained by normalizing AlphaSeq log KD values and computing the difference from the maximum observed log KD.

Validation: Designed Sequences

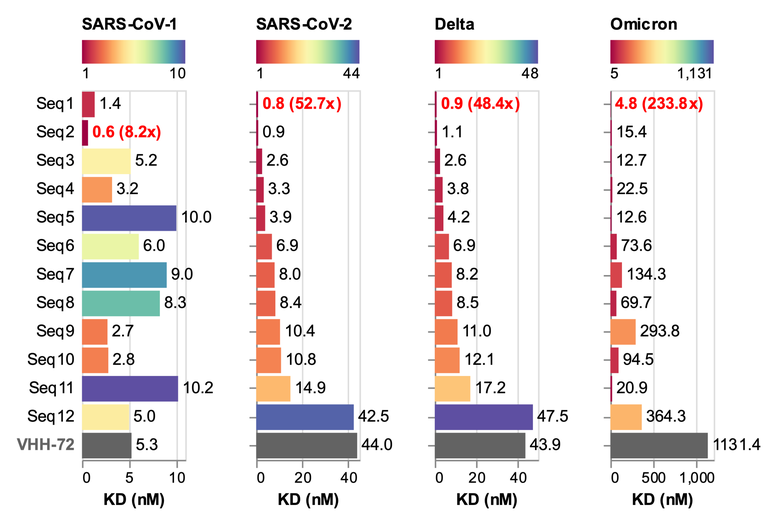
Using ML-guided design, we identified sequences with up to 15 mutations from parental VHH-72 with a significantly higher success rate compared to human-guided sequence design.



(A) Hit rate, the fraction of sequences that improve affinity compared to parental by at least one standard deviation, is higher with ML design and **(B)** enables sampling of a larger sequence space. ML-designed sequences result in **(C)** stronger (lower KD) binders and **(D)** enable a greater mutational distance from VHH-72 compared with human-designed sequences, as measured by AlphaSeq.

BLI Measurements of Top Hits

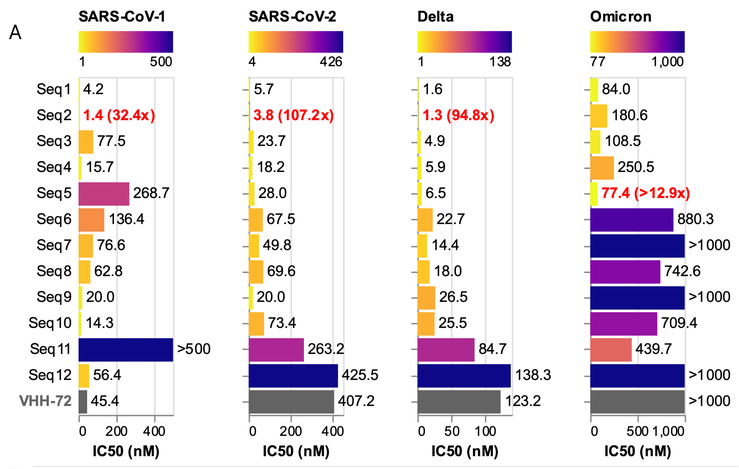
From 12 ML derived variants, 11 bind stronger to SARS-CoV-2 and maintain binding to SARS-CoV-1. Additionally, all tested variants gain cross-reactivity to the Omicron variant.



Biolayer interferometry (BLI) affinity measurements of 12 top VHH hits and VHH-72 against SARS-CoV-1 RBD, SARS-CoV-2 RBD, SARS-CoV-2 Delta RBD, and SARS-CoV-2 Omicron RBD.

Pseudovirus Neutralization Measurements of Top Hits

From 12 ML derived variants, 11 have improved SARS-CoV-2 neutralization and retain neutralization of the Delta variant. Additionally, 5 of the top hits gain neutralization of the Omicron variant, which is undetectable for parental VHH-72.



(A) Pseudovirus neutralization potency (IC50) for the top VHH hits and parental VHH-72 against the RBDs of SARS-CoV-1, SARS-CoV-2, SARS-CoV-2 Delta, and SARS-CoV-2 Omicron. **(B)** Neutralization curves of the top three ML-designed sequences (Seq1-3) and parental VHH-72.

Conclusions

After 3 rounds of ML-guided optimization, ML-generated sequences outperformed a non-ML baseline approach by up to 4-fold. We validated that >90% of top hits have improved binding by BLI and increased potency by pseudovirus neutralization to SARS-CoV-2, our primary target sequence. We also observed cross-reactivity to Delta and Omicron variants, which were not included in the original training set. The top VHH hits have an average of 6 mutations and up to 8 mutations from parental VHH-72, which suggests that ML-based optimization is capable to generating higher-order mutational variants and allow for exploration of a diverse sequence space.

Next Steps: Transfer Learning

A-Alpha Bio’s ML-guided antibody optimization workflow allows for a rapid design/test/iterate workflow on 10,000s of antibody variants against 10s of targets. This approach generates many sequence-diverse antibody variants with the desired affinity, specificity, and developability properties.

We have recently demonstrated that using a model pre-trained on antibody-antigen binding data, even from an unrelated target, can significantly improve binding predictions when starting a new antibody optimization campaign. We expect that our predictive power will continue to improve as we grow our PPI database (currently >300M PPIs) and implement more powerful pre-trained models.