HIGH THROUGHPUT, HIGH FIDELITY DATA GENERATION TO ENABLE AI LEAD OPTIMIZATION

Sharrol Bachas*, Goran Rakocevic*, David Spencer, Robel Haile, Anand Sastry, John Sutton, George Kasun, Vincent Blay, Christa Kohnert, Cailen McCloskey, Edriss Yassine, Borka Medjo, Nebojsa Tijanic, Shaheed Abdulhaqq, Randal Olson, Jennifer Stanton, Bailey White, Rebecca Viazzo, Rebecca Consbruck, Hayley Carter, Chelsea Chung, Brea Luton, Nic Diaz, Curtis Orona, Kaitlin Witherell, Alexander Brown, Amber Brown, Joshua Meier, Matthew Weinstock, Gregory Hannum, Ariel Schwartz, Miles Gander, Roberto Spreafico

Traditional antibody optimization approaches involve screening a small subset of the available sequence space, often resulting in drug candidates with suboptimal binding affinity, developability or immunogenicity. Based on two distinct antibodies, we demonstrate that deep contextual language models trained on high-throughput affinity data can quantitatively predict binding of unseen antibody sequence variants. These variants span a K_D range of three orders of magnitude over a large mutational space. Our models reveal strong epistatic effects, which highlight the need for intelligent screening approaches. In addition, we introduce the modeling of "Naturalness", a metric that scores antibody variants for similarity to natural immunoglobulins. We show that Naturalness score is associated with measures of drug developability and immunogenicity, and that it can be optimized alongside binding affinity using a genetic algorithm. This approach promises to accelerate and improve antibody engineering, and may increase the success rate in developing novel antibody and related drug candidates.

Read the full manuscript:

www.biorxiv.org/content/10.1101/2022.08.16.504181v1

FAB CDR LIBRARIES ARE TRANSFORMED INTO SOLUPRO™ *E. coli*

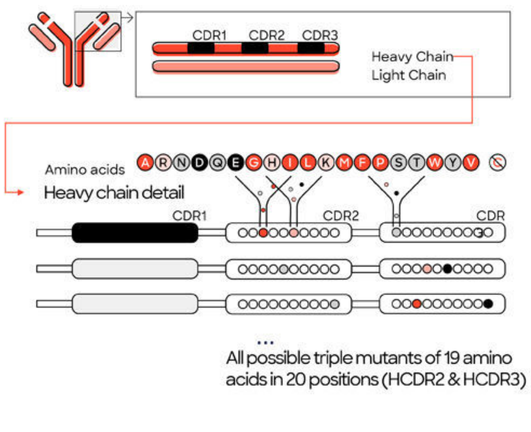
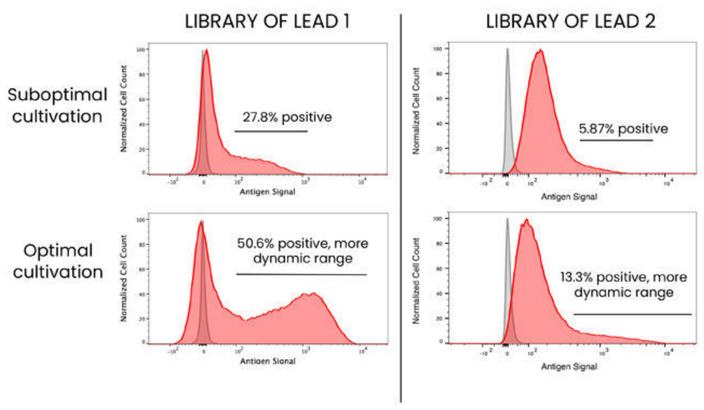


Illustration of the combinatorial mutagenesis strategy of a trastuzumab fab library composed of up to triple mutants in 20 positions (10 in CDRH2, 10 in CDRH3).

50k measured variants using ACE (activity-specific cell enrichment, a flow cytometry based assay) corresponds to 0.71% of the triple mutant combinatorial space. ACE data is used to train deep contextual language models that further interrogate the sequence space.

CULTIVATION MATRIX IDENTIFIES OPTIMAL GROWTH CONDITIONS

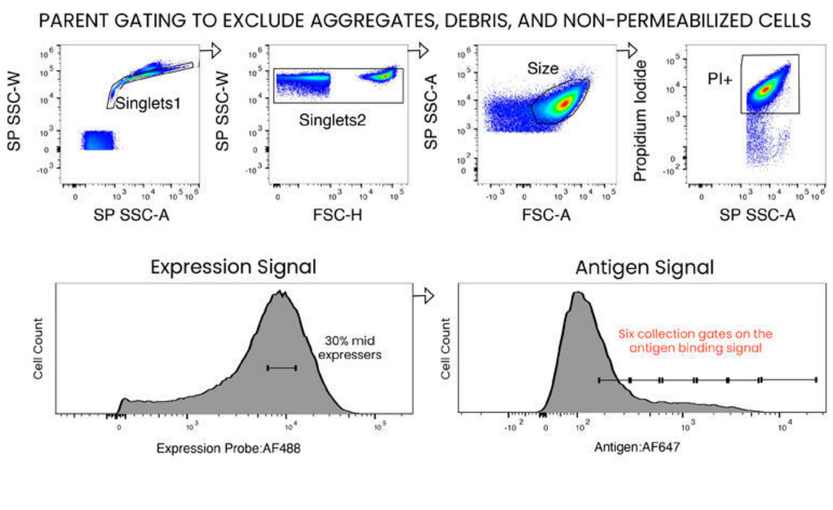
Factors such growth media and cultivation time impact the proportion of correctly folded and expressed fabs. Various seed and induction conditions are tested for libraries off each new drug lead. Higher quality protein increases the dynamic range of the binding signal and thus overall precision of the ACE assay. Red histograms indicate the library antigen binding signal and grey overlays are isotype controls measured in ACE. All libraries are sequenced to confirm improvement in ACE signal is not attributable to preferential growth of high affinity variants (not shown).



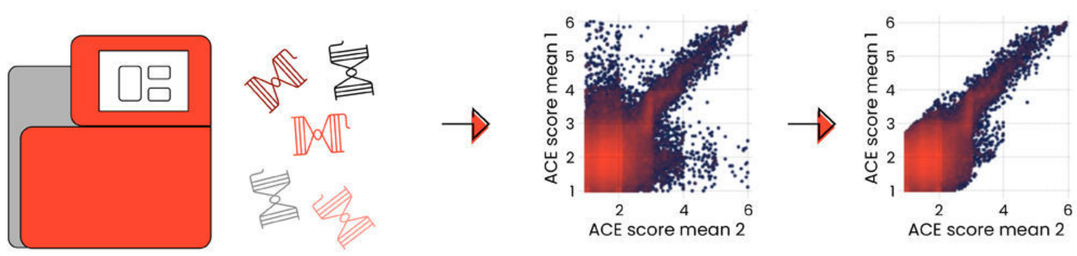
QUANTITATIVE AFFINITY ACE (qaACE) GENERATES LARGE TRAINING DATASETS (>50K VARIANTS)

Fab libraries expressed intracellularly in SoluPro™ *E. coli* → Cells are fixed, permeabilized, and stained for antigen binding and fab titer → The library is screened by flow cytometry to identify binding affinity of each library variant

A.



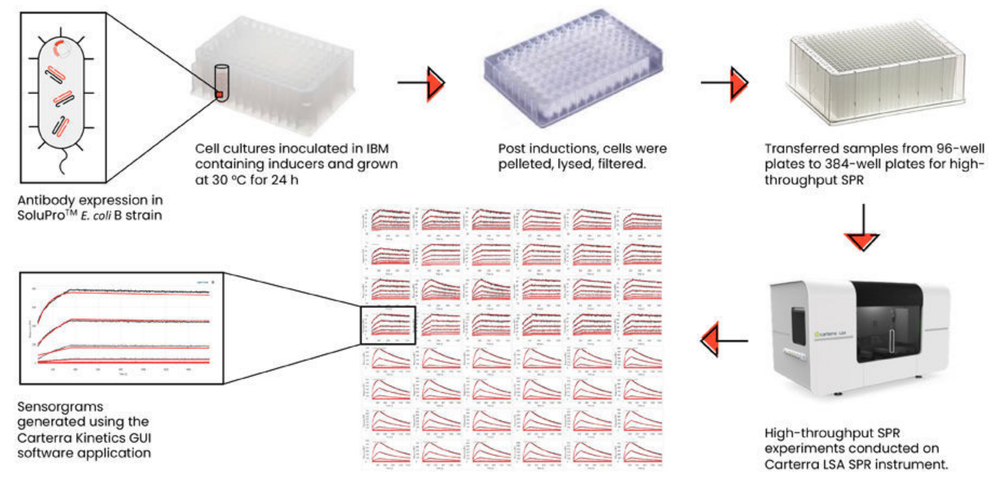
B.



The flow cytometry gating scheme is shown in (A). After parent gating to reduce aggregates, debris, and non-permeabilized cells, bias to antigen binding signal from expression variability is controlled through an additional parent gate on the 30% mid expressers. Six collection gates are then used to bin evenly across the log range of the antigen signal. (B) After sorting, unique molecular identifiers are added to flank the CDR region. Collected material is then amplified and sequenced. Read counts weighted by distribution in sort gates are used to assign ACE scores to each variant, and variants measured consistently between multiple sort replicates are retained.

SPR USED TO VALIDATE QAACE DATASETS AND MODEL PREDICTIONS

A.

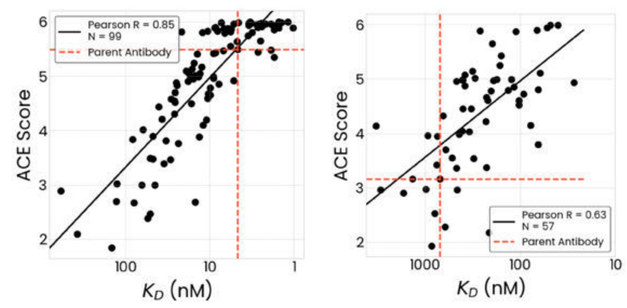


(A) SPR workflow. A subset of ACE-measured variants are ordered as gBlocks, transformed and cultured in SoluPro™ *E. coli*, and culture lysates screened via SPR by immobilizing fab on the biosensors.

Standard SPR runs measure binding traces for up to 1,500 variants, each measured at 4 concentrations in duplicate. qua confirm that ACE scores correlate with SPR measured equilibrium constants (K_D). Following model training on ACE scores, a subsequent round of SPR is performed to confirm model predictions.

(B) Correlation between SPR measured KDs and ACE scores for two representative lead molecules.

B.

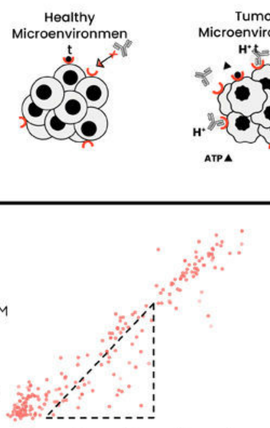


Lead Molecule vs Antigen 1 | Lead Molecule vs Antigen 2

ACE SCREENING FOR CONDITIONAL BIOLOGICS

ACE datasets can be generated for multiple parameters, such as pH dependent binding (right) or binding affinity against multiple antigens (not shown). These paired datasets enable modeling for biologics co-optimized for the desired traits.

Affinity constants (K_D) from a subset of trastuzumab fab variants screened at neutral and low pH in ACE are shown. Variants with preferential binding in low pH are indicated in the dashed triangle, and the equilibrium constants for lead conditional biologics are shown in the table.



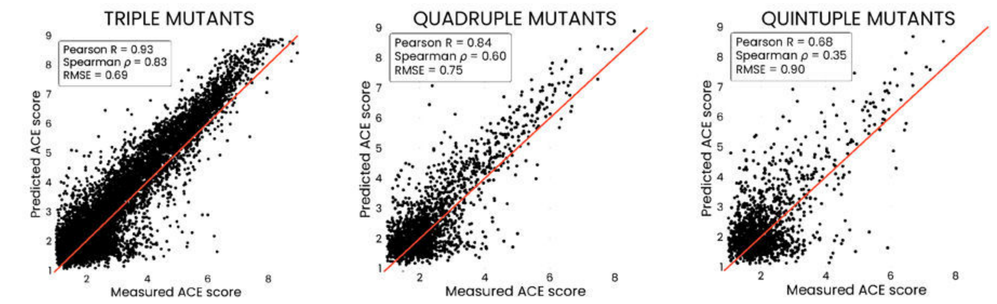
pH dependent binders for the acidic tumor microenvironment

Current Opinion in Biotechnology 2022, 78: 102809

Attribute	K_D pH 5.8	K_D pH 7.4	Fold-change
Hit 1	0.7 nM	12.6 nM	18
Hit 2	6 nM	63 nM	10.5
Hit 3	0.7 nM	6.8 nM	9.7
Hit 4	0.7 nM	12.6 nM	18

LEAD OPTIMIZATION THROUGH ARTIFICIAL INTELLIGENCE PREDICTIONS OF BINDING AFFINITY AND NATURALNESS SCORE

A.

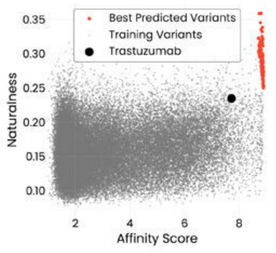


Search space = 6.7 Million
Excellent predictive performance

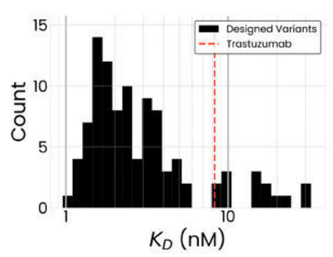
Search space = 515 Million
Performance is maintained

Search space = 30 Billion
Performance begins to decline

B.



C.



(A) Model predictive performance for triple, quadruple, and quintuple mutants after training on the 50k triple mutant trastuzumab dataset.

(B) Model predictions for variants co-optimized for high affinity and maximum "naturalness," a measure of similarity to immunoglobulins occurring in the OAS database and associated with immunogenicity and developability.

(C) 85% of the top 100 predicted "natural" variants bind tighter to HER2 than parental trastuzumab

KEY POINTS

Deep contextual language models trained on high-throughput affinity data can quantitatively predict binding of unseen antibody sequence variants.

Models can co-optimize for multiple parameters, such as affinity and naturalness.

Models can be tuned to optimize for specific target affinity for one or more targets

Absci's ACE technology can generate large (>50k), high-quality training datasets to support AI lead optimization. This end-to-end technology involves rational library design, expression and cultivation in SoluPro™ *E. coli*, ACE high-throughput screening, and medium-throughput validation (100s) using SPR.

Training datasets can be generated for multiple parameters, such as pH-dependent binding or affinity to multiple antigens.

PARTNER WITH US

Leverage our AI lead optimization technology to enable your projects.

©2023 Absci Corp. All rights reserved.