Creating a platform for rapid computational antibody design via machine learning, HPC, and laboratory experimentation

Thomas Desautels
LLNL

LLNL: Daniel Faissol, Adam Zemla, Ed Lau, Fangqiang Zhu, John Goforth, Denis Vashchenko, Mary Silva, Rebecca Haluska, Drew Bennett, Emilia Grzesiak, Alexander Ladd, Brent Segelke, Feliza Bourguet, Victoria Lao, Monica Borucki, Dina Weilhammer, Jacky Lo, Nicole Collette, Magdalena Franco, Kathryn Arrildt
Sandia NL: Brooke Harmon, Oscar Negrete, Max Stefan

LLNL ML4I Workshop
August 12, 2021

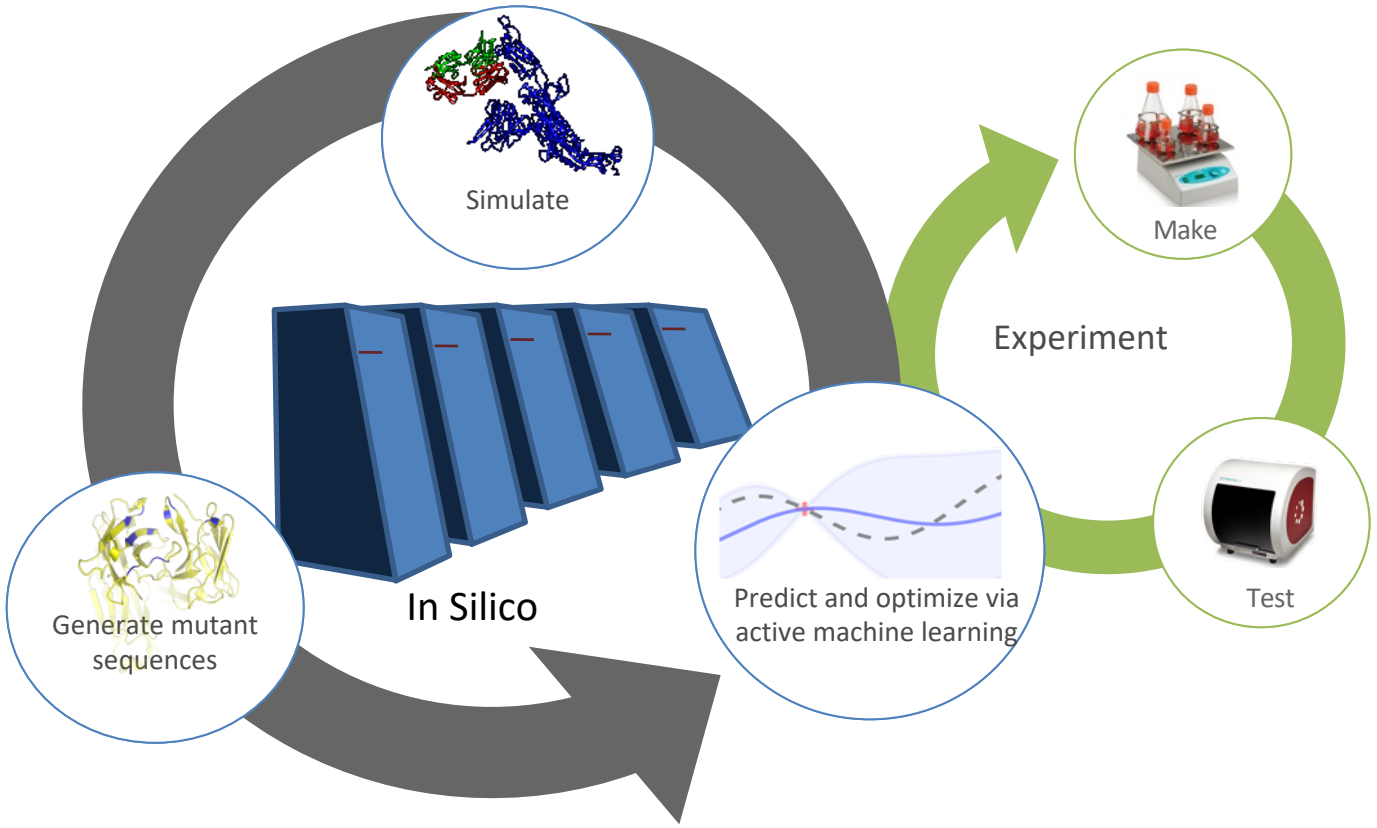
LLNL-PRES-825249
This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344 and was supported by the LLNL-LDRD Program under Project No. 20-ERD-032. Lawrence Livermore National Security, LLC

Lawrence Livermore National Laboratory

The problem: how can you *rapidly* respond to a new pathogen?

- Premises:
  - New pathogens can emerge with little warning
  - The immune system may need assistance to effectively counter a new pathogen
  - Vaccine antigens and therapeutic antibodies are the most important protein design targets
    - Basically the only things that have worked at all for COVID
- Ordinarily, vaccines and therapeutic antibodies take years or decades to reach market
- **In the long-term, we want a system for scalable, high-confidence, *in silico* design that could accelerate delivery of a countermeasure that is (1) effective, (2) manufacturable, and (3) safe.**
- In a familiar LLNL plan, do design & certification as much as possible in the computer
  - Critically, this can enable *preemptive* design against emerging virus variants or novel members of families of pathogens

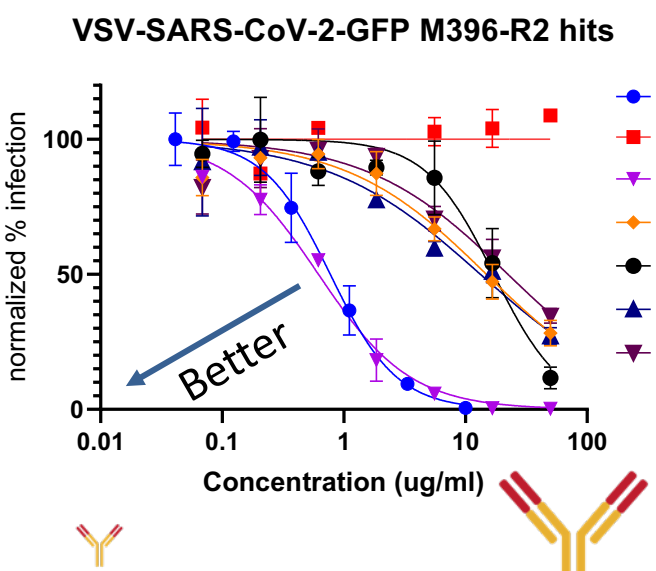
Our approach to countermeasure design combines simulation and ML-driven decision-making with laboratory experimentation



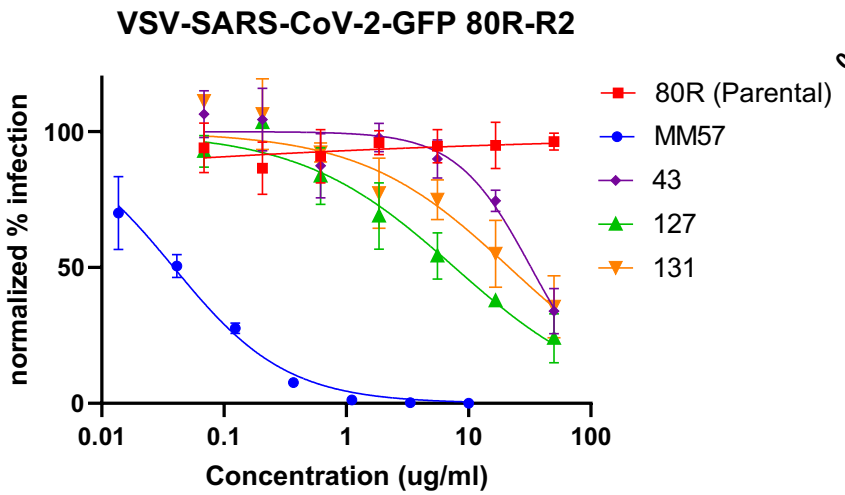
- Select & simulate computationally until promising candidates are found
- Send best candidates for laboratory testing
- If necessary, re-design from most promising candidates identified in the laboratory

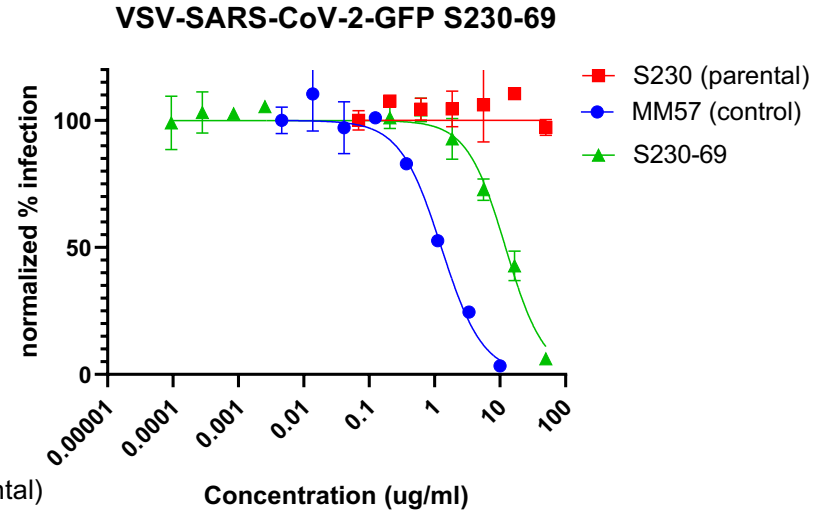
We've *executed and validated* rapid antibody *design* against SARS-CoV-2: novel to our knowledge

- From Jan 2020 to present, designed several neutralizing antibodies for SARS-CoV-2

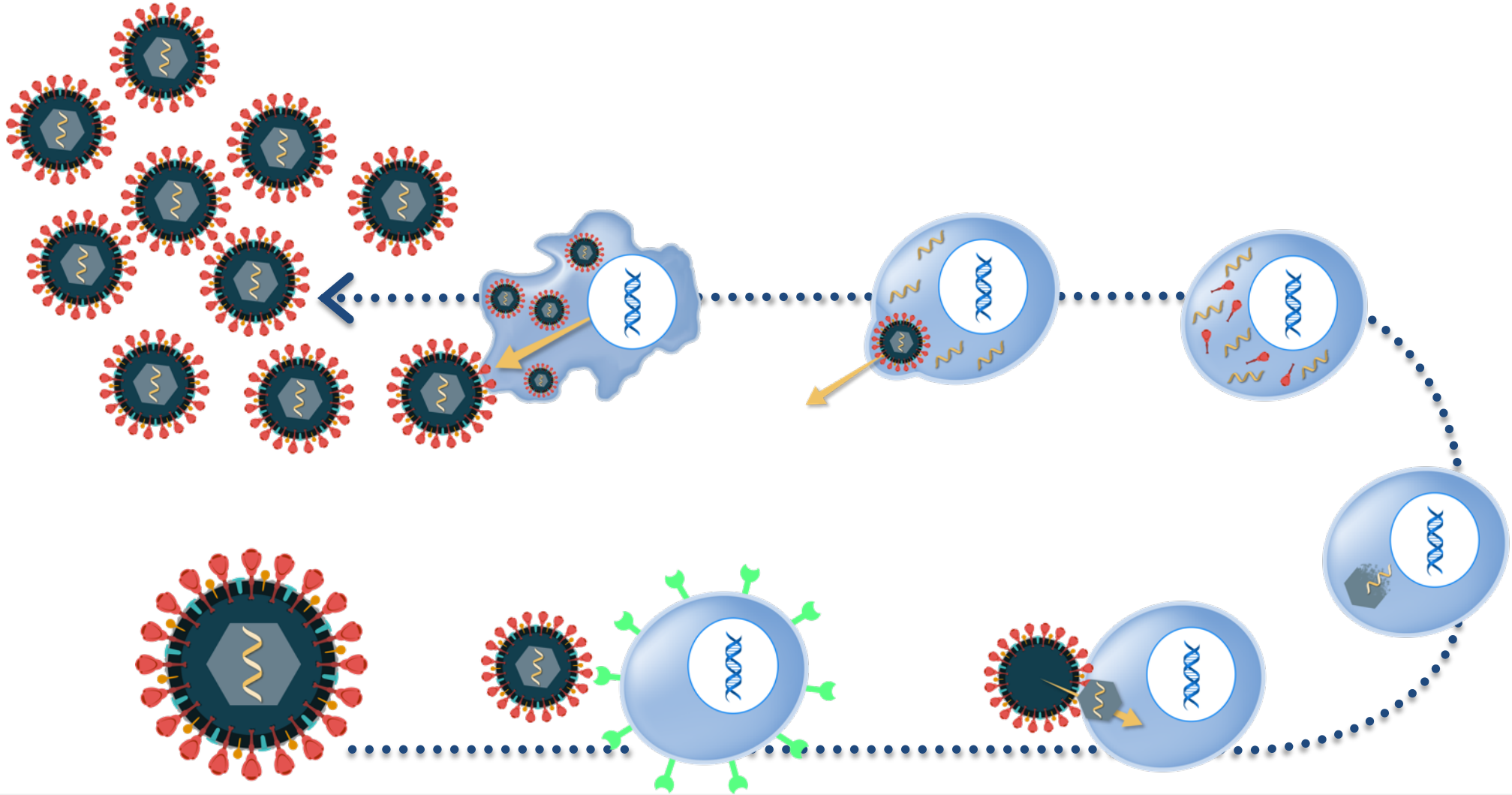


Our best antibody

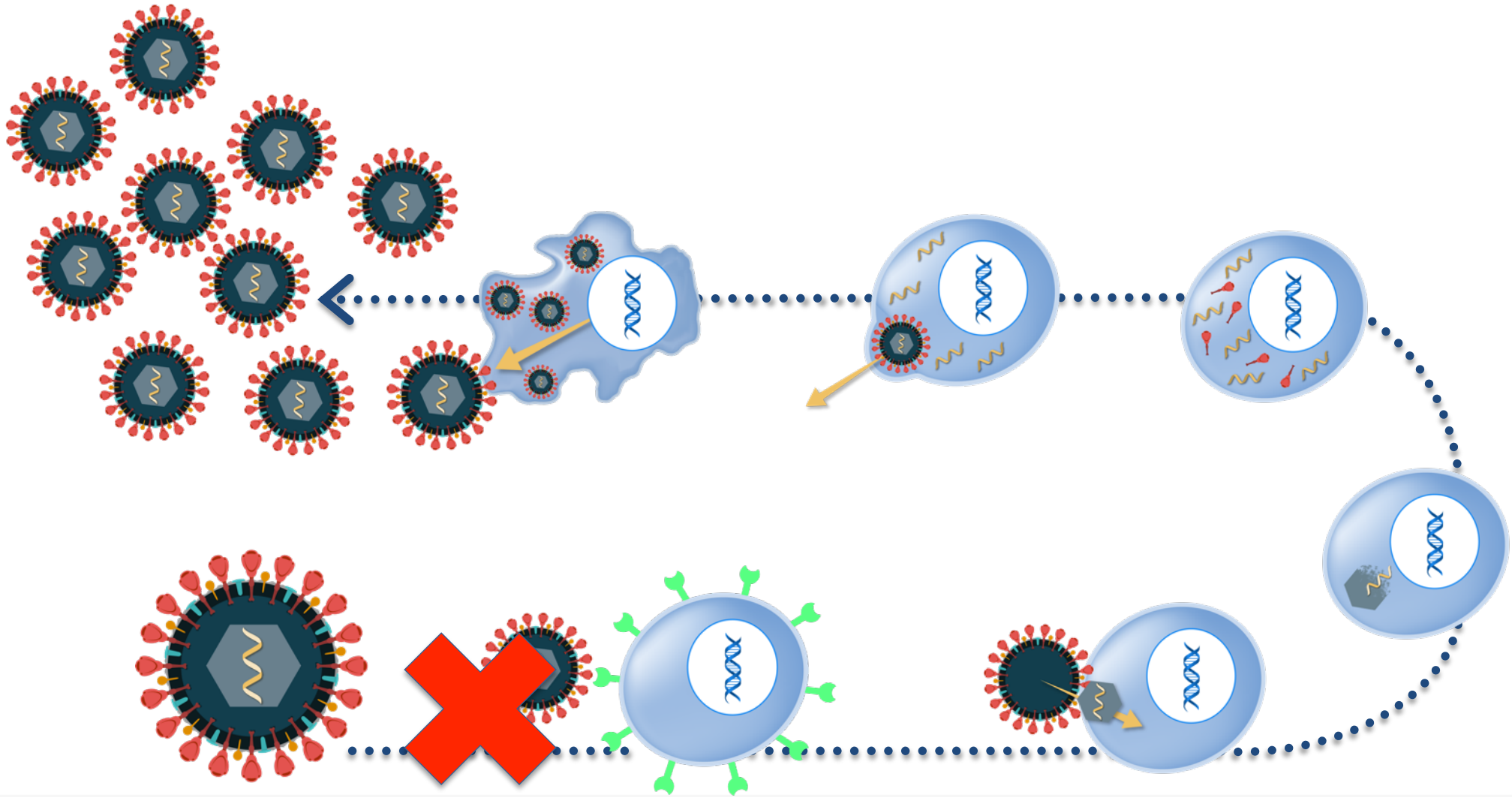




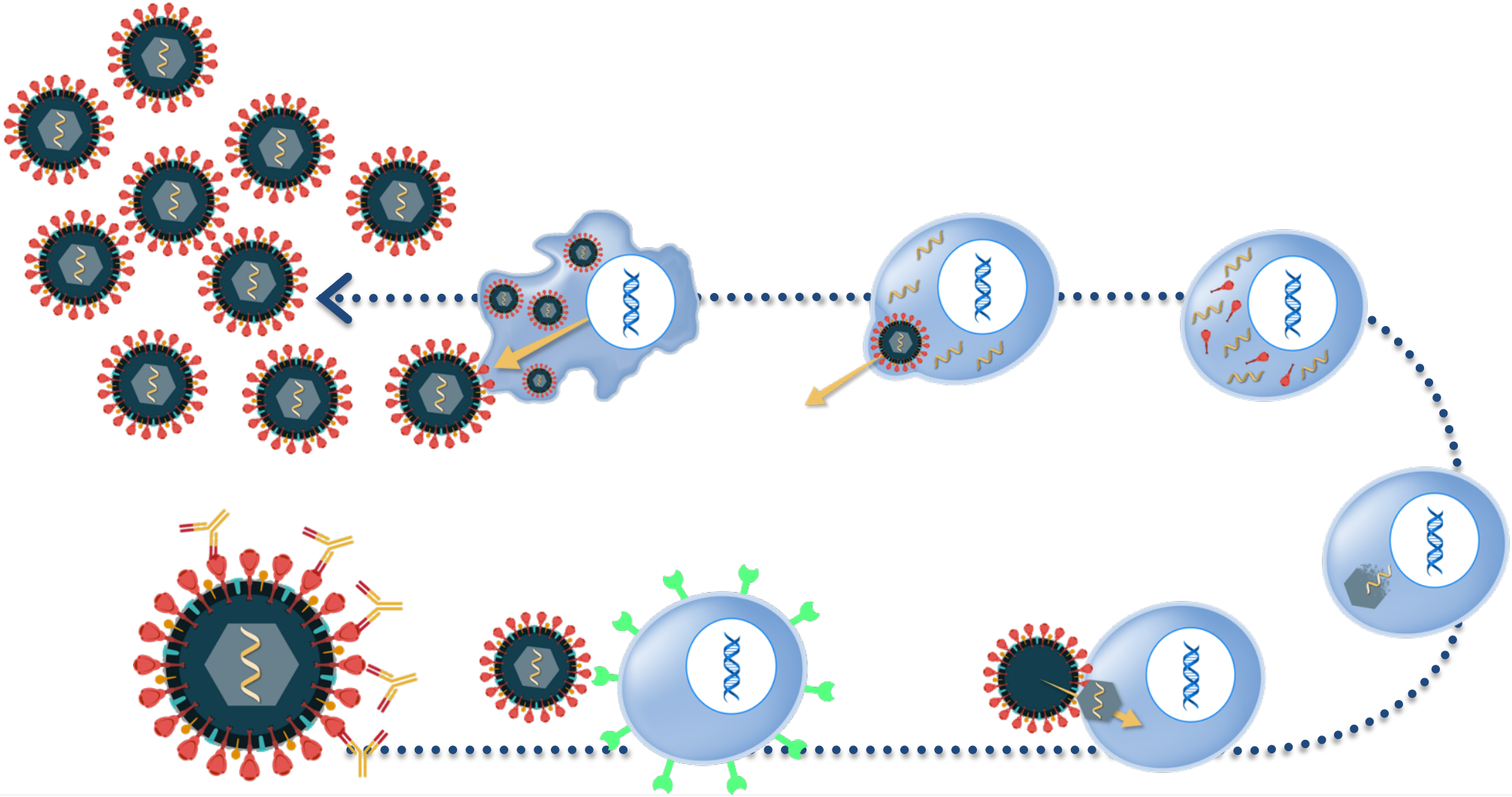
Viruses reproduce by entering and hijacking host cells



If we could stop viral entry, we could stop the viral cycle

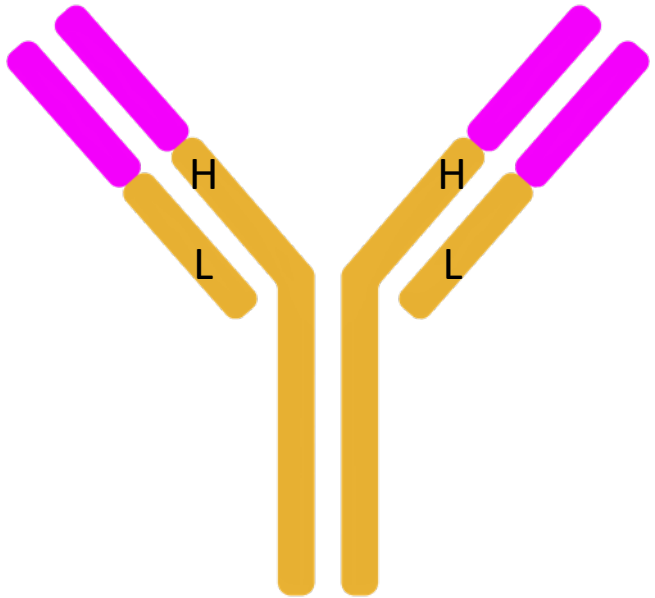


Neutralizing antibodies can stop viral entry



Proteins are described by their amino acid sequence: Antibody design becomes finding a suitable sequence

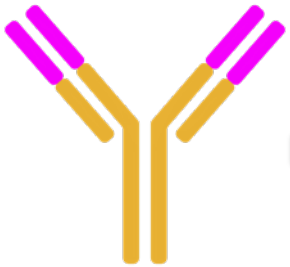
> m396 heavy chain
QVQLQQSGAEVKKPGSSVKVSCKASGGTFS
SYTISWVRQAPGQGLEWMGGITPILGIANY
AQKFQGRVTITTDESTSTAYMELSSLRSEDTA
VYYCARDTVMGGMDVWGQGTTVTVSSAS
TKGPSVFPLAPSSKSTSGGTSALGCLVKDYFP
EPVTVSWNSGALTSGVHTFPAVLQSSGLYSLS
SVVTVPSSSLGTQTYICNVNHKPSNTKVDKK
VEPKSCDKTSPLFVHHHHHHGDYKD
DDDKG



> m396 light chain
SYELTQPPSVSVAPGKTARITCGGNNIGSKSV
HWYQQKPGQAPVLVVYDDSDRPSGIPERFS
GSNSGNTATLTISRVEAGDEADYYCQVWDSS
SDYVFGTGTKVTVLGQPKANPTVTLFPPSSE
EFQANKATLVCLISDFYPGAVTVAWKADGSP
VKAGVETTKPSKQSNNKYAASSYLSLTPEQW
KSHRSYSCQVTHEGSTVEKTVAPTECS

m396 neutralizes SARS-CoV-1, but not SARS-CoV-2; **can its sequence be modified to bind a target antigen and neutralize a new virus?**

The design space is vastly larger than what we can simulate or test



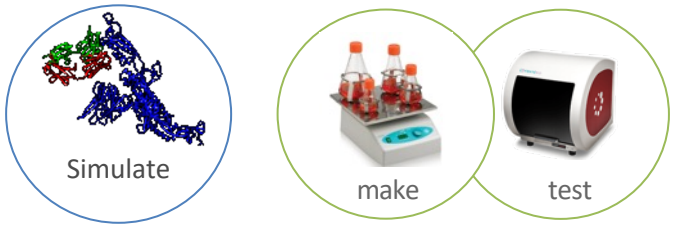
CoV-1 +
changes
~10^{30}

Computer
Simulations
1,000,000

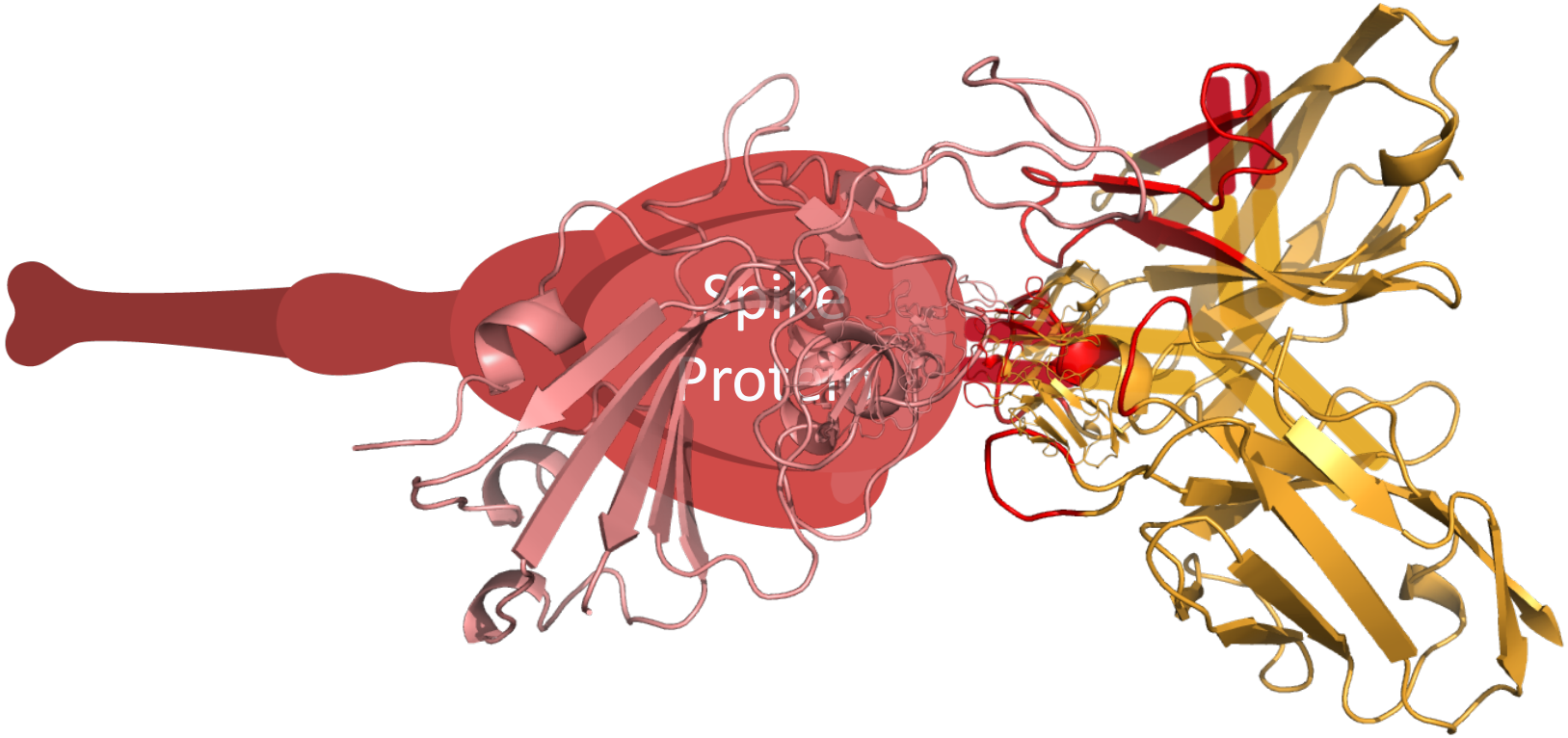
Laboratory
Experiments
100-1,000

CoV-2
Need just one!

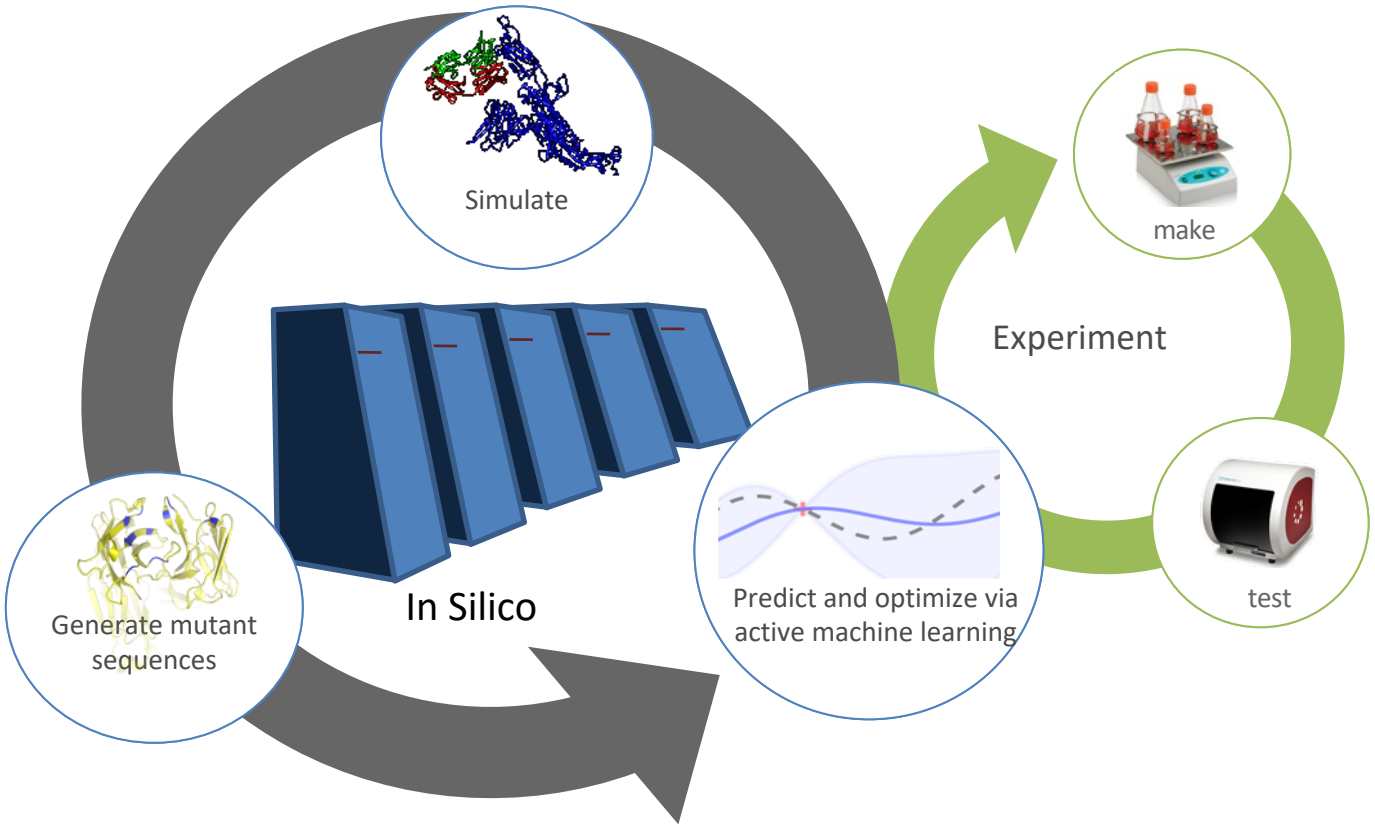
Strong binding is our main target; neutralization objective *may* follow



- In simulation and in the laboratory, we can ask questions like:
  - How strongly does the antibody bind its target? *dG* (binding free energy) or K_D (rate const.)
  - How does this change as we mutate the antibody? *ddG* (mutational change in dG)



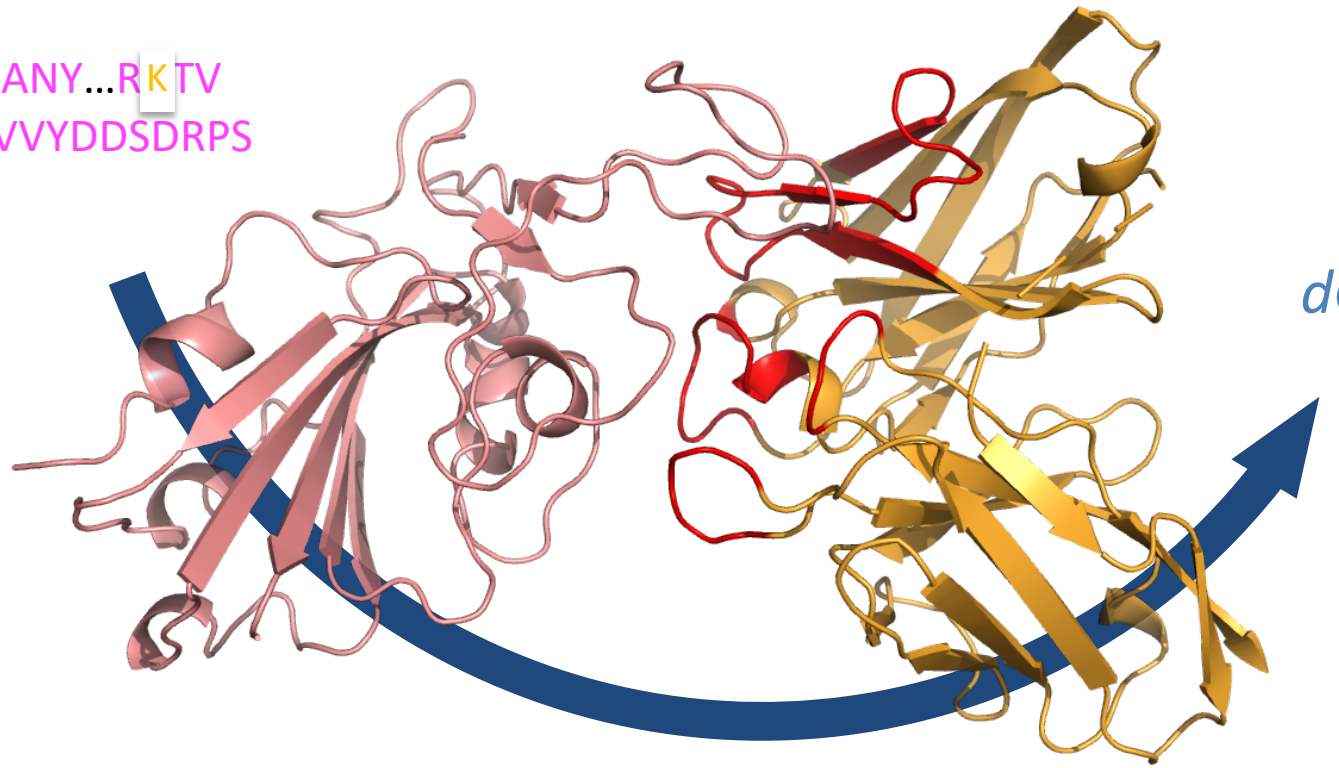
Platform software and active machine learning support these simulation and experimental tools



Pose the design problem as active learning

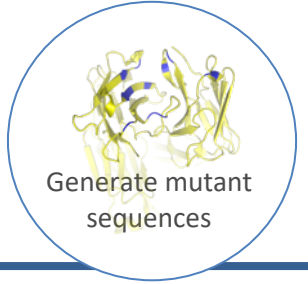
- Improve the antibody sequence by iteratively selecting antibodies from a discrete set and evaluating them

> m396 mutable residues
…GTFSSYTIS…WMGG S PILGIANY…R K TV
MGGMDV…/…NIGSKSVH…LVVYDDSDRPS
…QVWDSSSDY



dG, *ddG*, or K_D

Enumerate many antibody designs

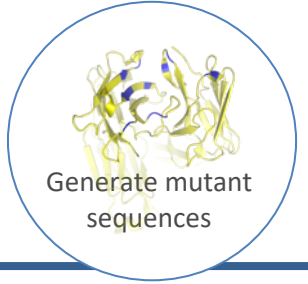


> m396 mutable residues
…GTFSSYTIS…WMGG S PILGIANY…R K TV
MGGMDV…/…NIGSKSVH…LVVYDDSDRPS
…QVWDSSSDY

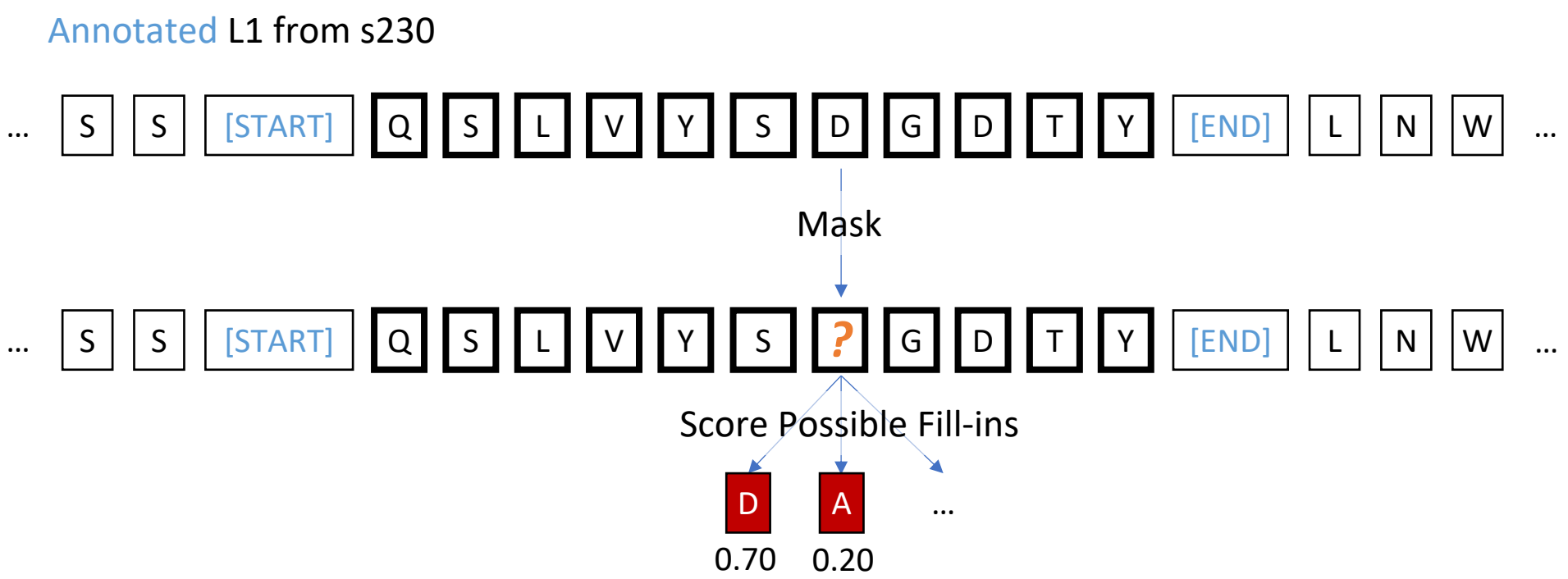
- Generators for novel sequences have so far been mostly tabular
  - Based on frequency of “typical” mutational “swaps”
  - OR based on expensive, high-fidelity calculations of single changes to template antibody in hypothesized complex with SARS-CoV-2 spike.

- This works all right, but can lead you to unrealistic sequence designs
  - Downstream problems in manufacturability, etc. are major concerns

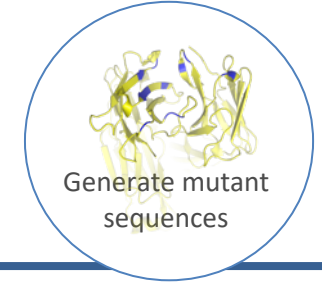
More realistic antibody sequences via language modeling



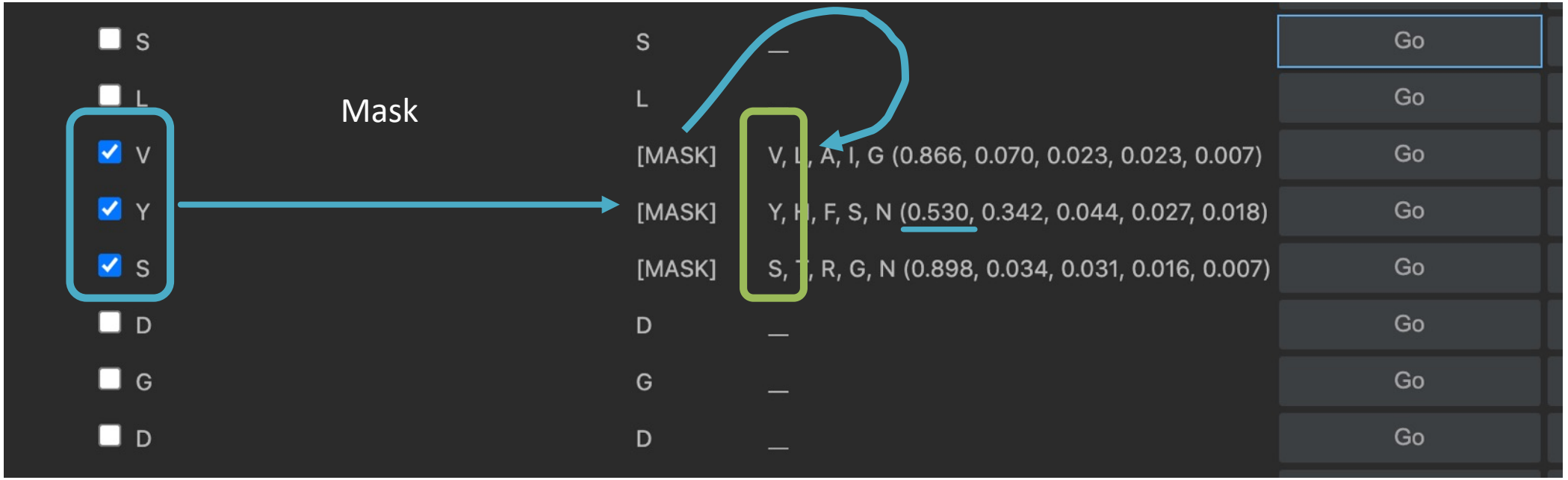
- Use a transformer model to learn to fill “masked” amino acids in the antibody sequence



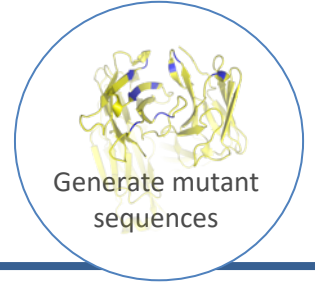
Our models learn to produce reasonable antibodies



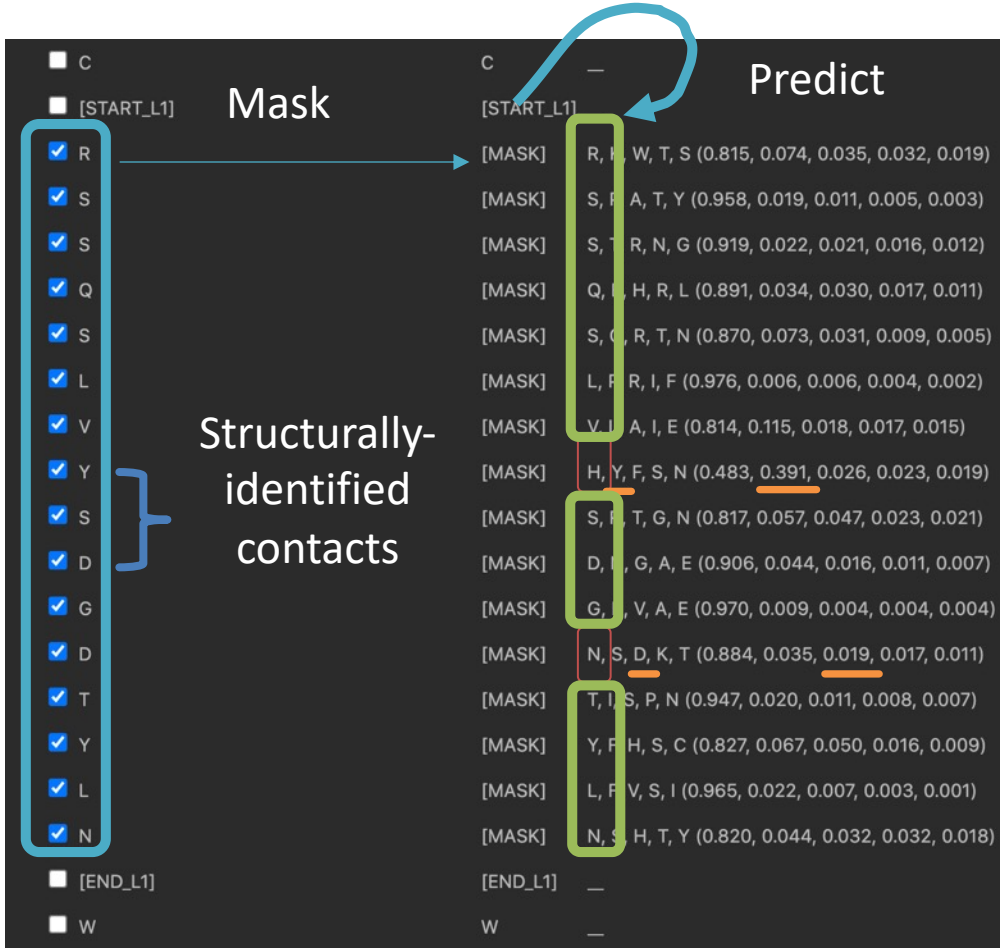
mask and predict 3 central amino acids of s230's L1 "loop"



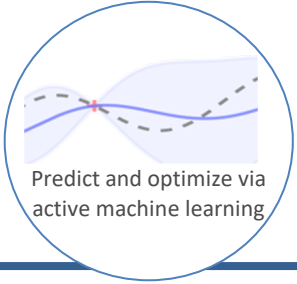
Our models learn to produce reasonable antibodies

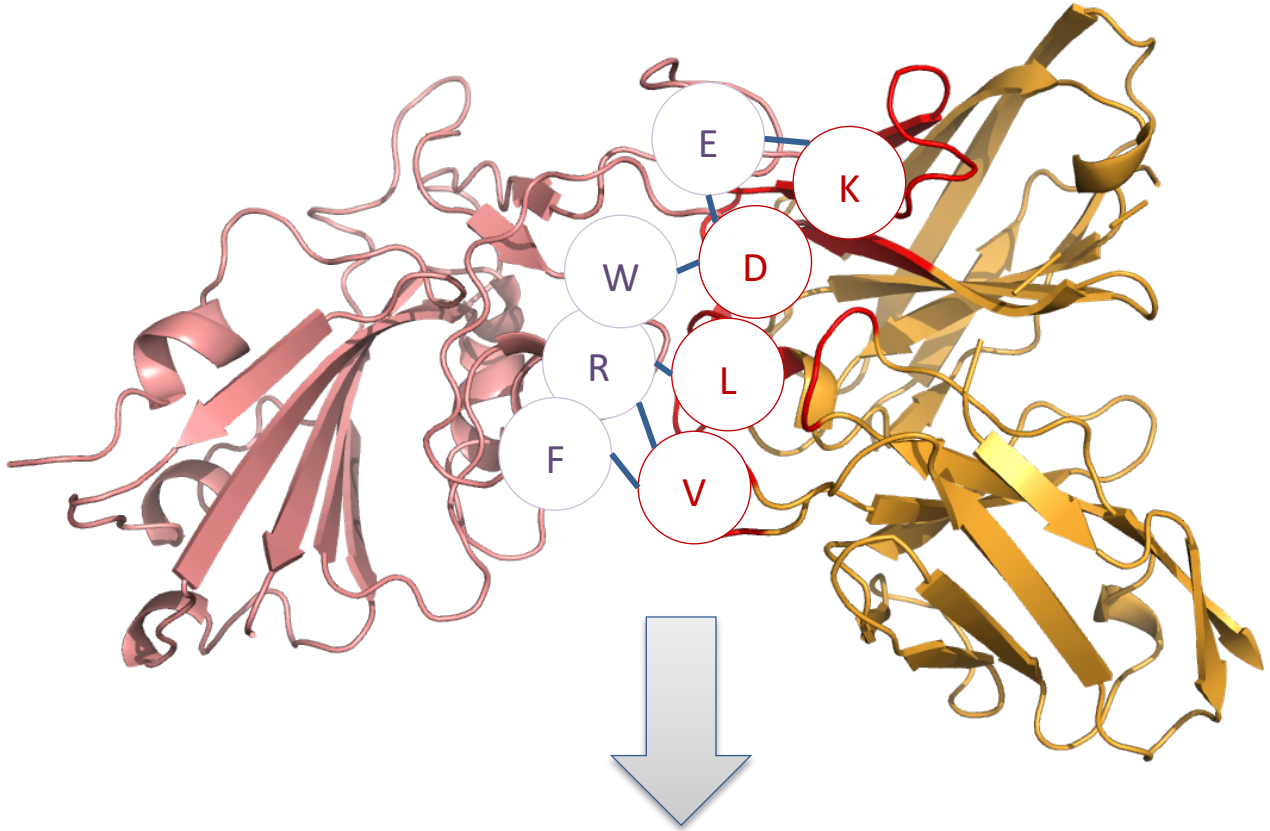


mask and predict all 16 amino acids of s230's L1 "loop"



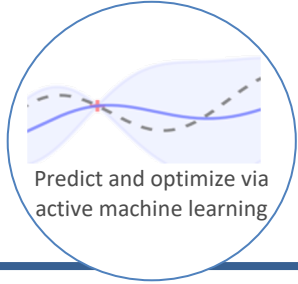
To predict how an antibody sequence will bind, we use a structure-based representation of the interactions





x = [0, 1, 0, 2, 0, 0, 1, …]
Vector of interaction type counts

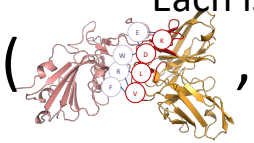
Represented in feature space, binding free energy estimates feed into a multi-fidelity Gaussian process model



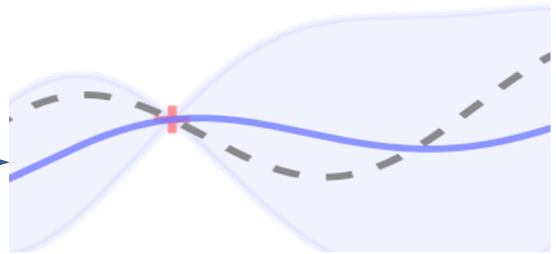
Executed studies

Data: $\{(x, i, y)_1 , \ldots (x, i, y)_n \}$

Each is a tuple:

( , FoldX, -4.3)

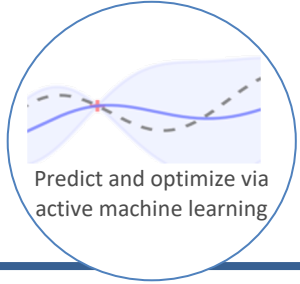
Observations



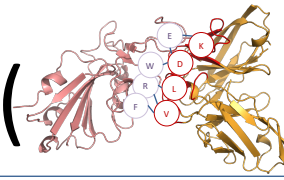
Gaussian process model:

$f : (x, i) \rightarrow \mu(x, i), \sigma(x, i)$

The next set of simulations is selected via Bayesian optimization using the Gaussian process model



Decision set: PROPOSED studies

( , FoldX,__)

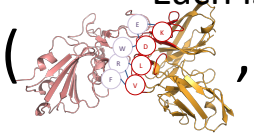
Selection:

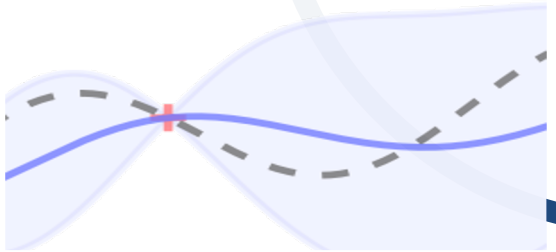
$(x, i)_t = argmax\ s$

Executed studies

Data: {(x, i, y)$_1$, … (x, i, y)$_n$ }

Each is a tuple:

( , FoldX, -4.3)

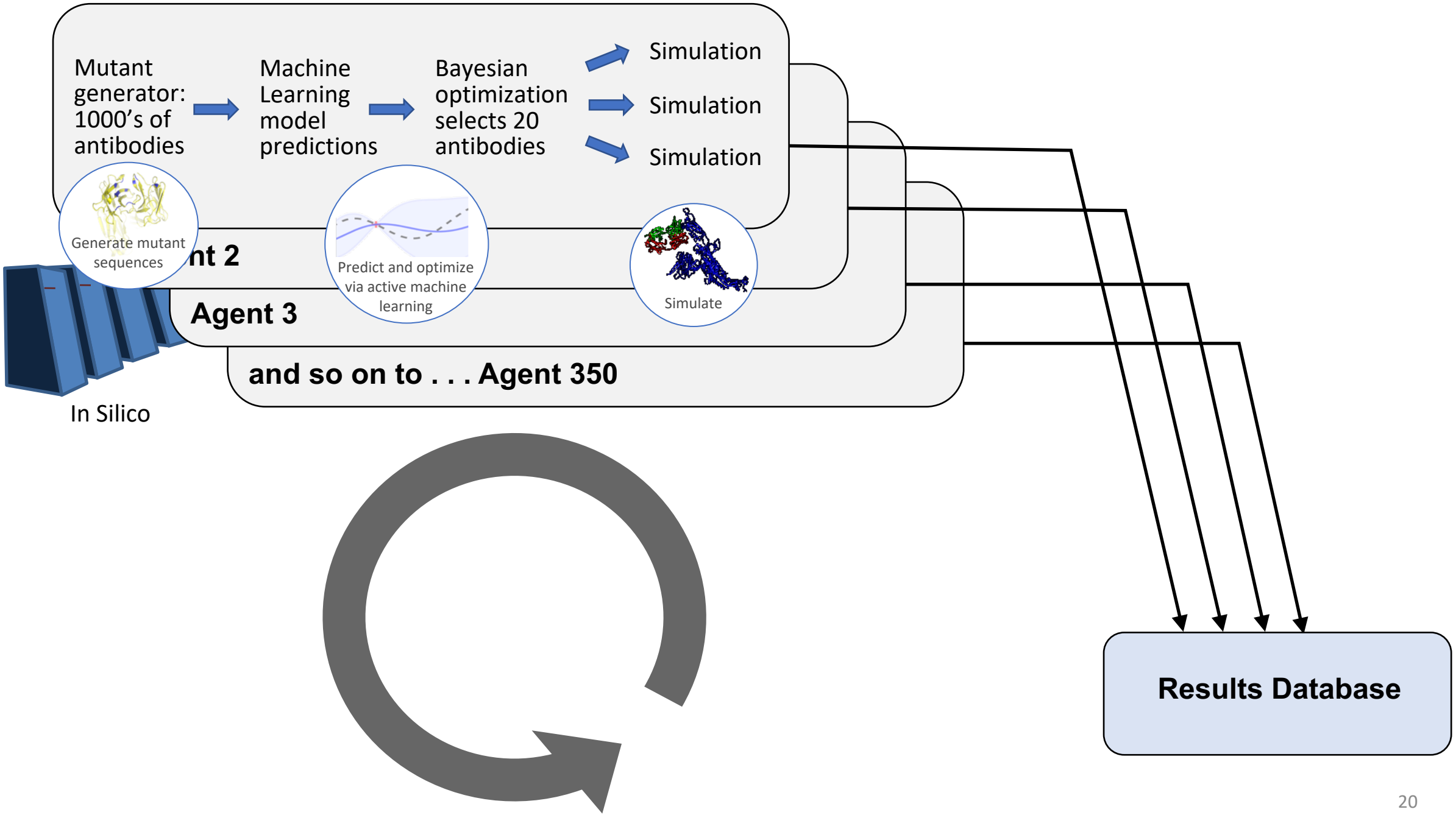


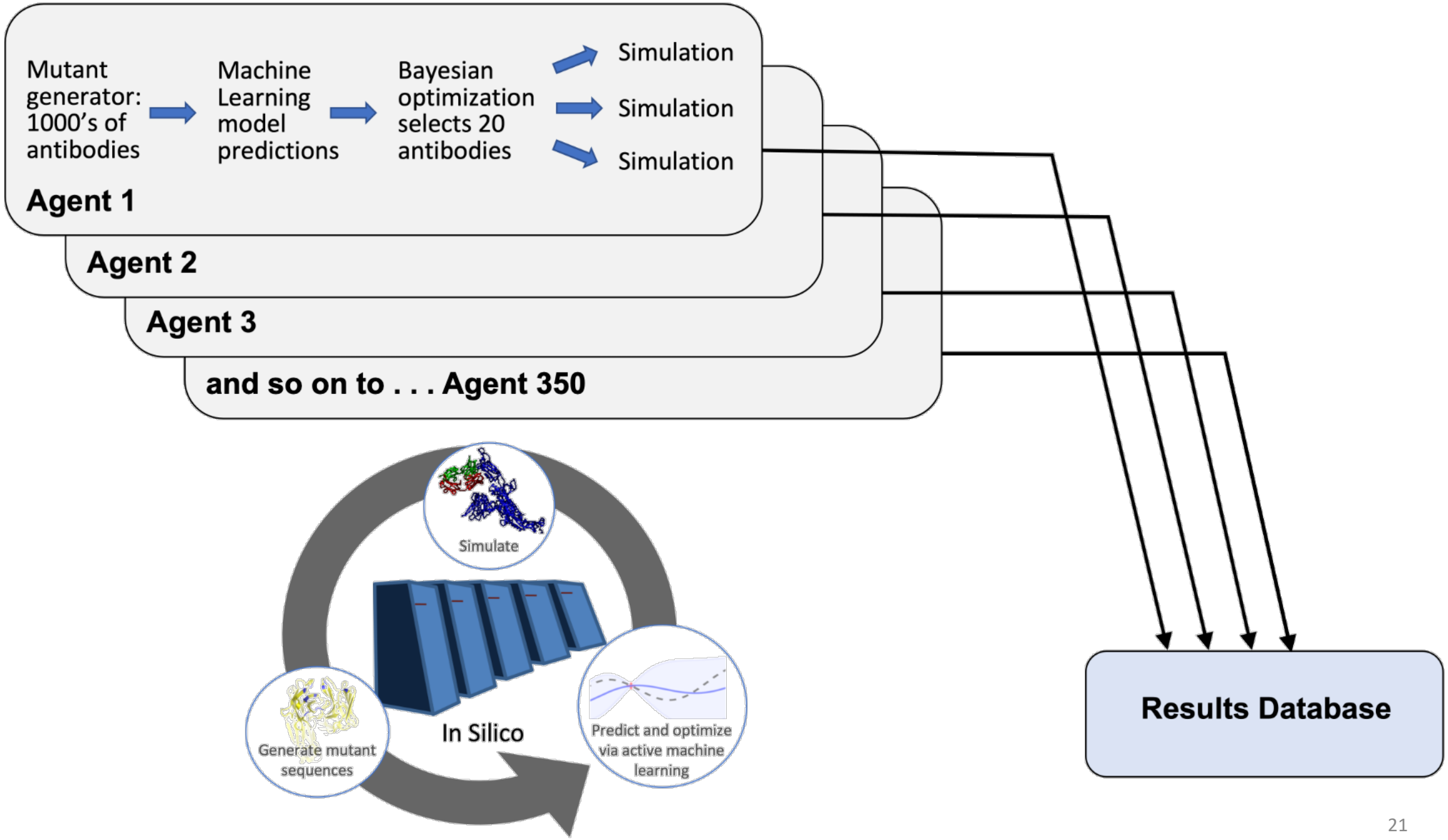
Gaussian process model:

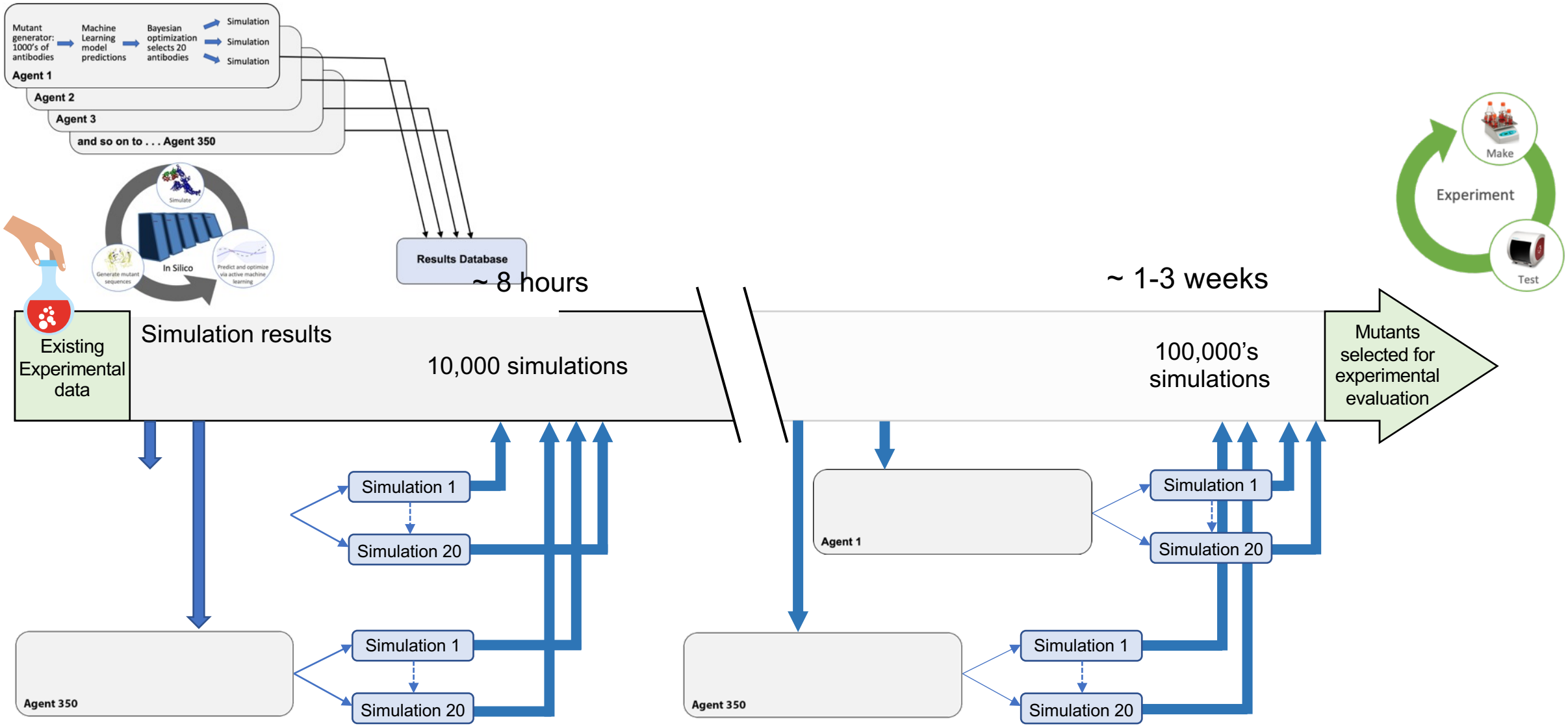
$f: (x, i) \rightarrow \mu(x, i), \sigma(x, i)$

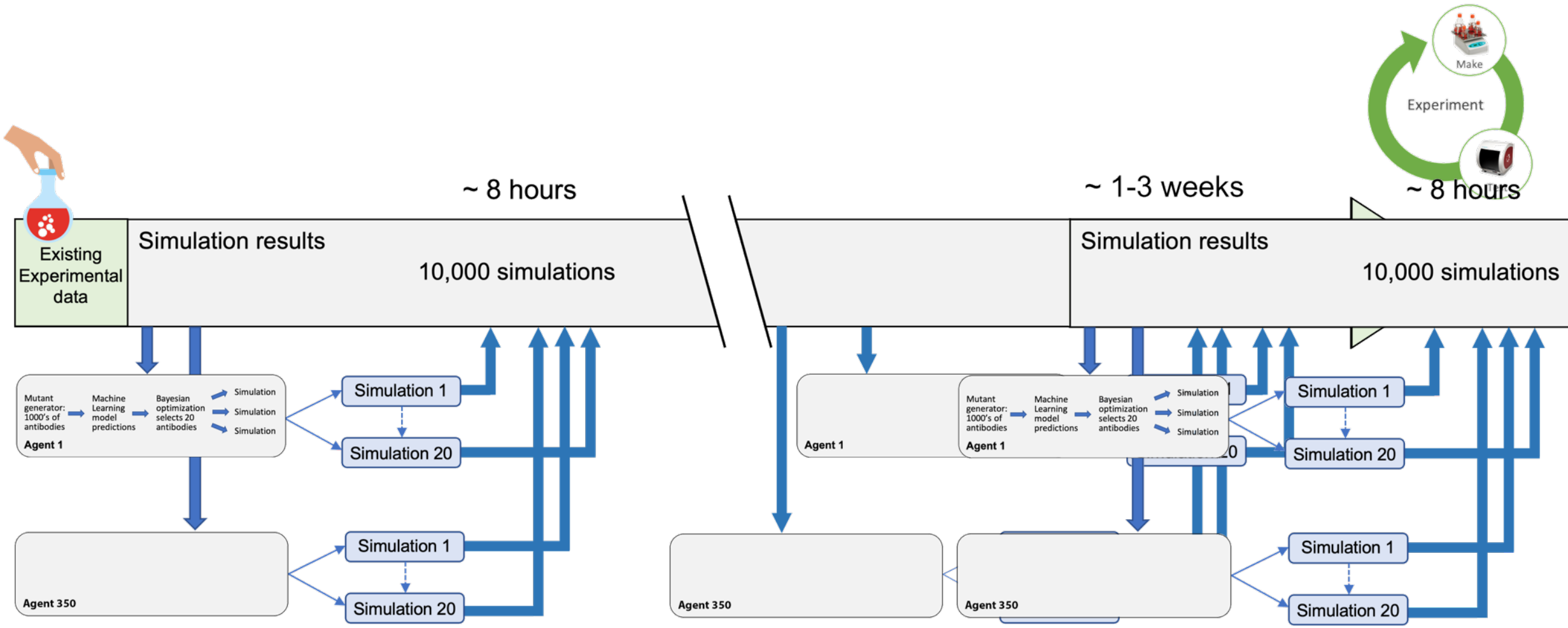
Score:

$s(\mu(x, i), \sigma(x, i))$


Mutant generator: 1000's of antibodies
Machine Learning model predictions
Bayesian optimization selects 20 antibodies
Simulation
Simulation
Simulation
Generate mutant sequences
Predict and optimize via active machine learning
Simulate
nt 2
Agent 3
and so on to . . . Agent 350
In Silico
Results Database

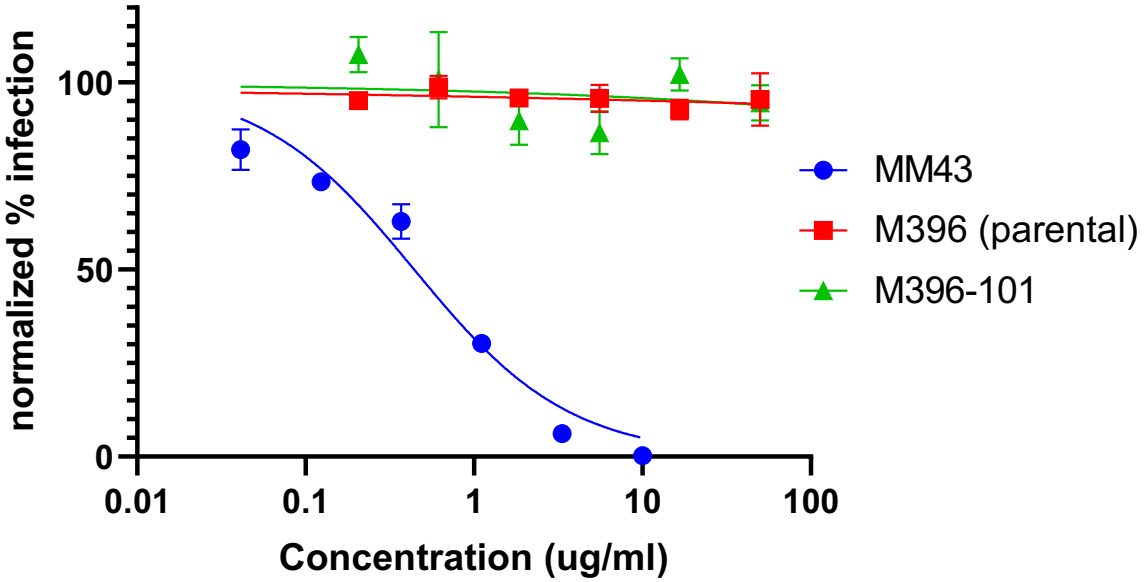

Mutant generator: 1000's of antibodies
Machine Learning model predictions
Bayesian optimization selects 20 antibodies
Simulation
Simulation
Simulation
Agent 1
Agent 2
Agent 3
and so on to . . . Agent 350
Simulate
Generate mutant sequences
In Silico
Predict and optimize via active machine learning
Results Database


Mutant generator: 1000's of antibodies
Machine Learning model predictions
Bayesian optimization selects 20 antibodies
Simulation
Simulation
Simulation
Agent 1
Agent 2
Agent 3
and so on to . . . Agent 350
Simulate
In Silico
Generate mutant sequences
Predict and optimize via active machine learning
Results Database
Make
Experiment
Test
~ 8 hours
~ 1-3 weeks
Existing Experimental data
Simulation results
10,000 simulations
100,000's simulations
Mutants selected for experimental evaluation
Simulation 1
Simulation 20
Agent 1
Agent 350


Make
Experiment
~ 8 hours
~ 1-3 weeks
~ 8 hours
Existing Experimental data
Simulation results
10,000 simulations
Simulation results
10,000 simulations
Mutant generator: 1000's of antibodies
Machine Learning model predictions
Bayesian optimization selects 20 antibodies
Simulation
Simulation
Simulation
Agent 1
Simulation 1
Simulation 20
Agent 350
Simulation 1
Simulation 20
Agent 1
Mutant generator: 1000's of antibodies
Machine Learning model predictions
Bayesian optimization selects 20 antibodies
Simulation
Simulation
Simulation
Agent 1
Simulation 1
Simulation 20
Agent 350
Agent 350
Simulation 1
Simulation 20

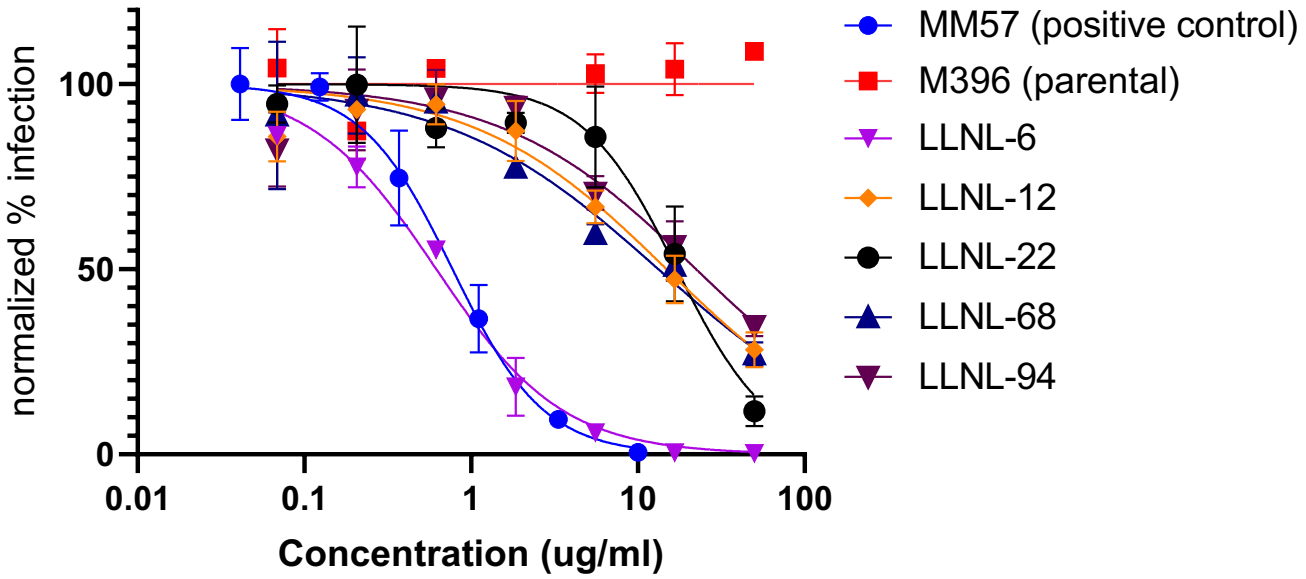
Several of our m396-derived antibodies inhibit VSV-SARS-CoV-2 virus

mAb101-Iteration 1
rVSV-SARS-CoV-2 S-GFP



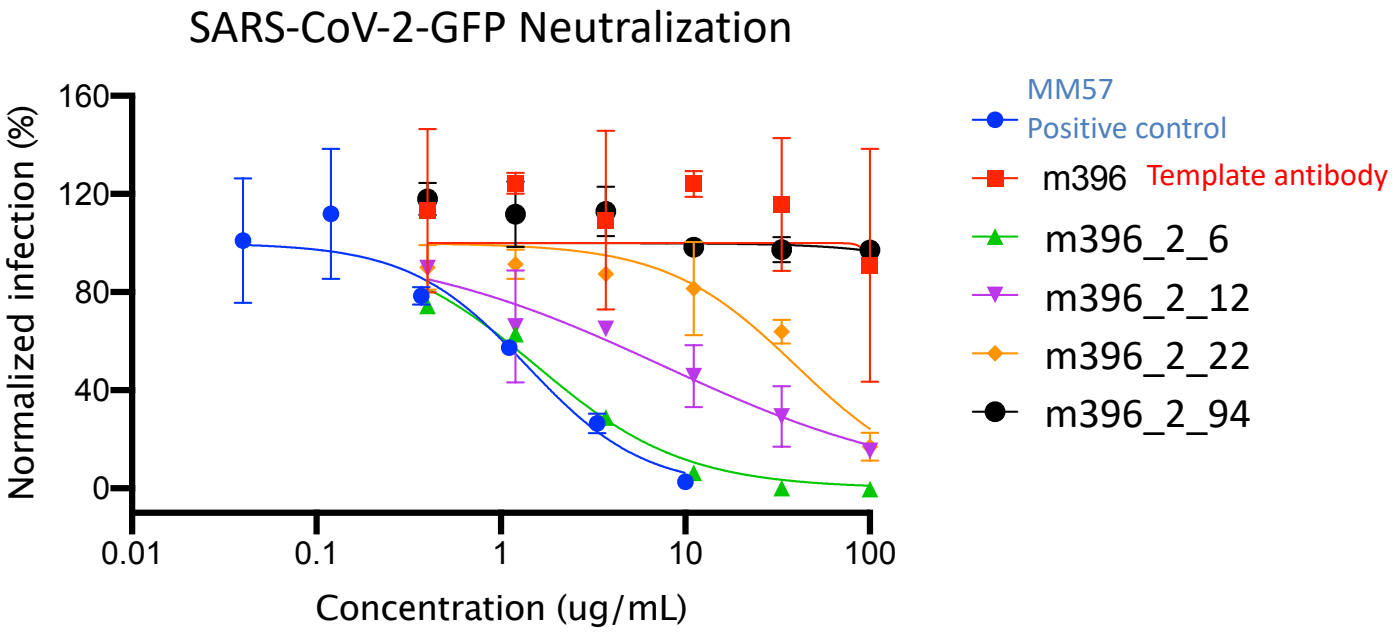
Design Iteration 2

VSV-SARS-CoV-2-GFP M396-R2 hits



Recall:
down and left is more virus neutralized needing less antibody = better

We confirmed these m396-derived antibodies neutralize authentic SARS-CoV-2 virus in our BSL-3 facility



This work is the product of a growing multidisciplinary team

- LLNL:
  Daniel Faissol, Adam Zemla, Ed Lau, Fangqiang Zhu, John Goforth, Denis Vashchenko, Mary Silva, Rebecca Haluska, Claudio Santiago, Sam Nguyen, Drew Bennett, Emilia Grzesiak, Brent Segelke, Feliza Bourguet, Victoria Lao, Monica Borucki, Dina Weilhammer, Jacky Lo, Nicole Collette, Kathryn Arrildt, and Magdalena Franco (now ThermoFisher)

- Sandia NL:
  Brooke Harmon, Oscar Negrete, Max Stefan

- Generous computer time and support from LC!
  - Catalyst, early access to Mammoth
  - Workflow enablement (database) and Sina (database interface) groups are critical to our ongoing success

PyTorch, GPyTorch, BioPython
Maestro, Sina, Improv
FoldX, RosettaFlex