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

Thomas Desautels
Staff Scientist, LLNL

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

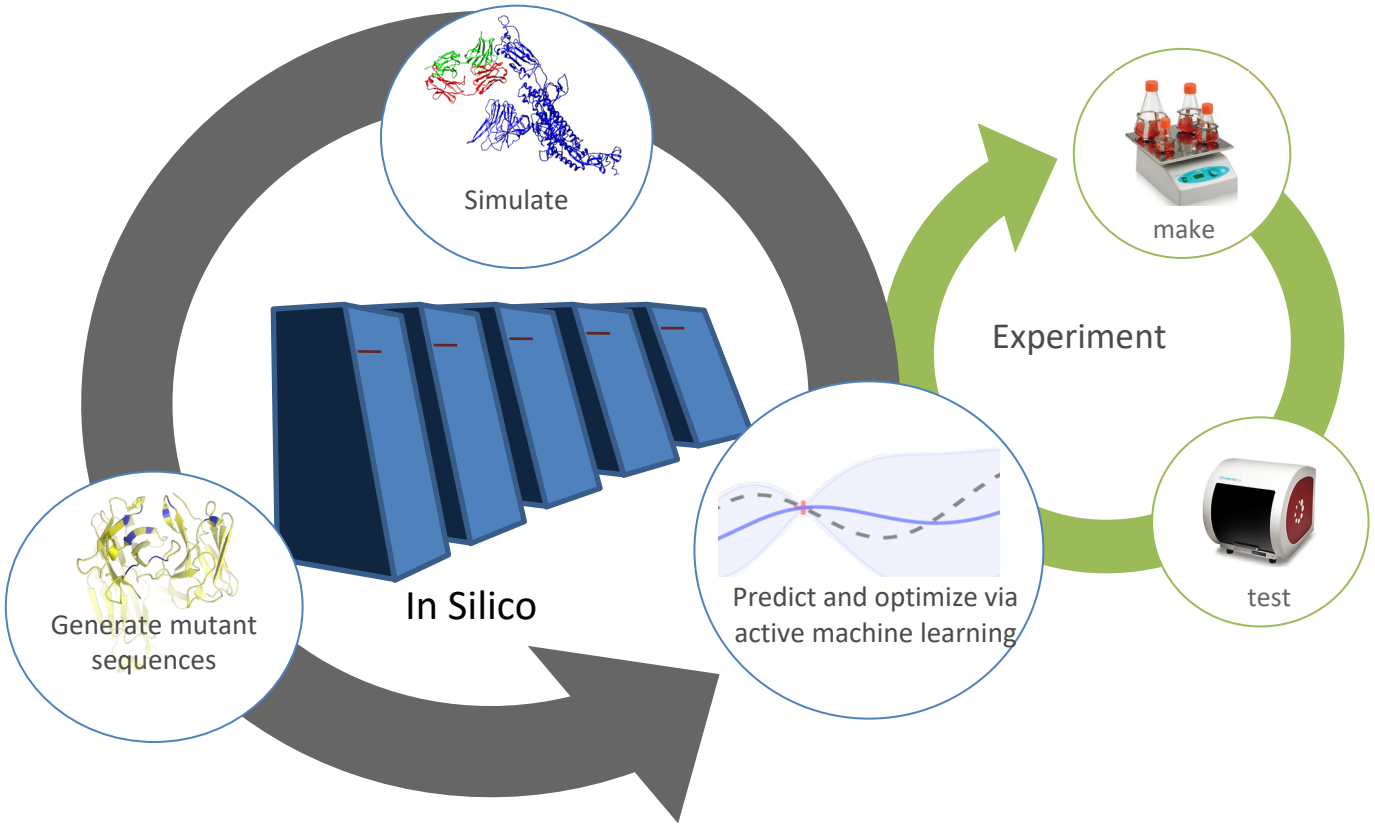
LLNL Data Science Institute Workshop:
AI in Healthcare

March 25, 2021

LLNL-PRES-820717
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

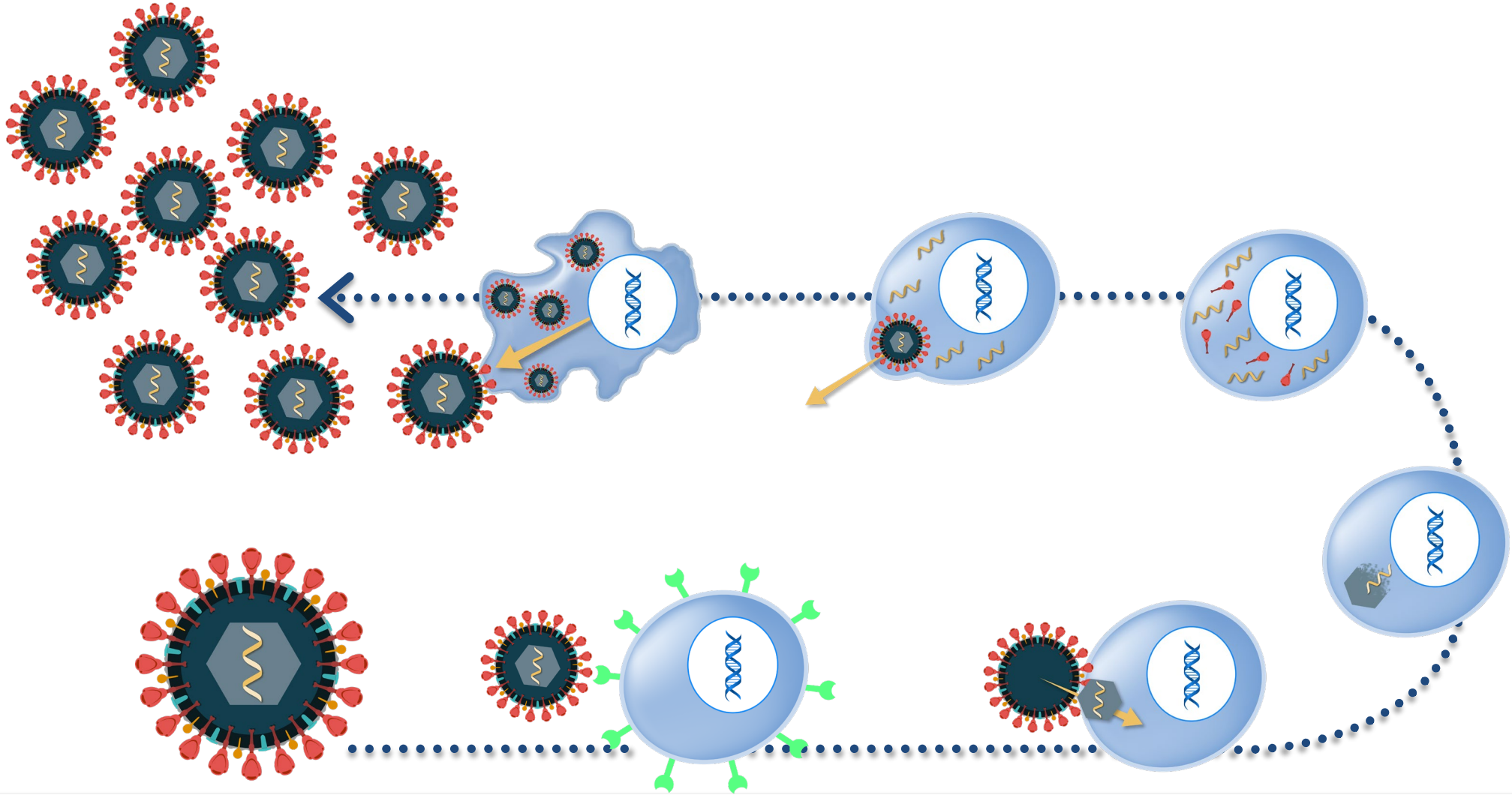
Lawrence Livermore National Laboratory

LLNL is developing a machine-learning-driven platform for rapid, rational, design of therapeutic antibodies and vaccine antigens

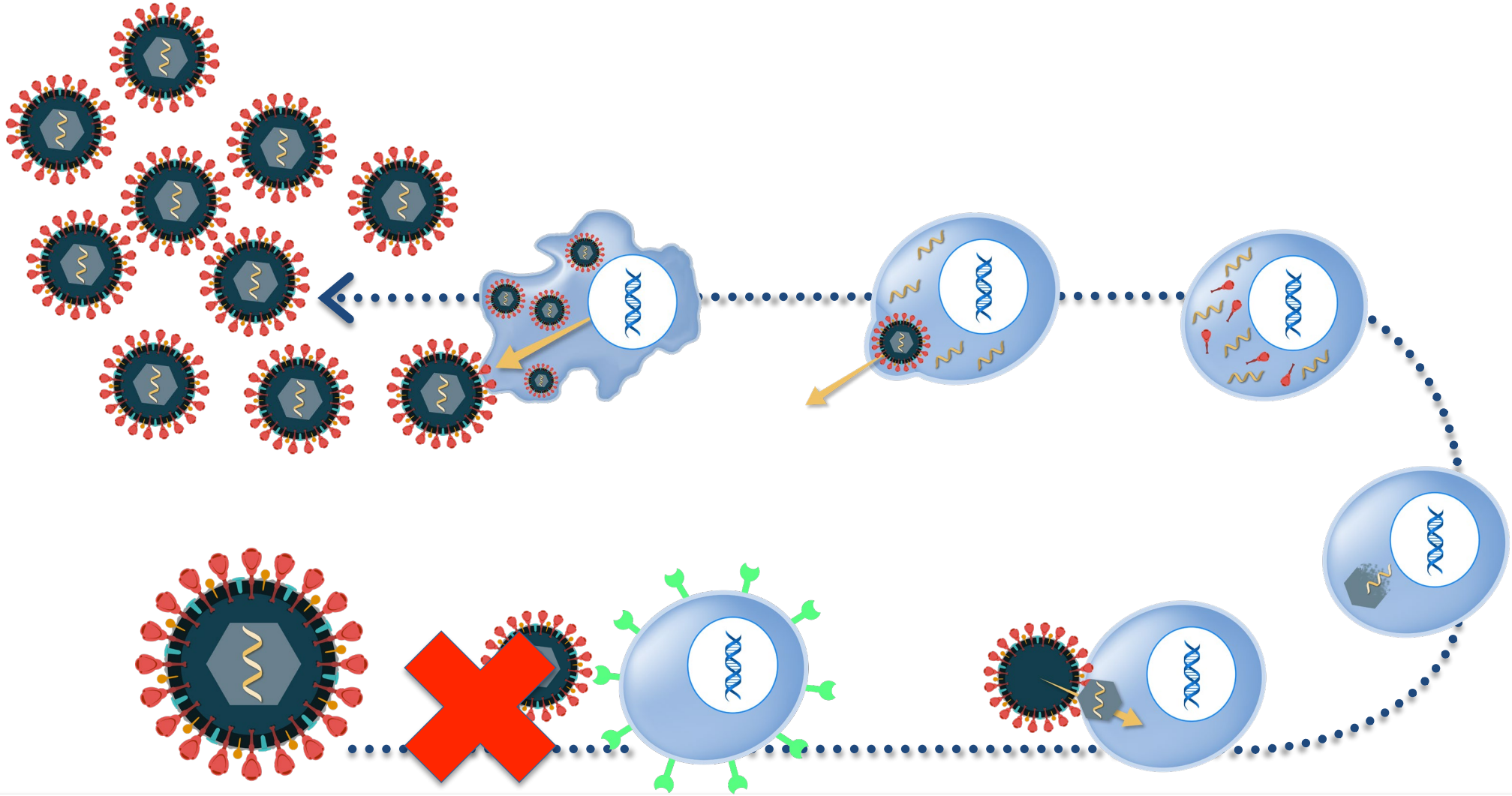


- Our platform is a novel autonomous, ML-driven system that performs in-silico biomolecular design at scale using HPC
- Active learning and Bayesian optimization approach harnesses predictions from molecular simulations and bioinformatic predictors together with high-value, directed experiments
  - > 1 million simulations performed w/ 3 million core hours on HPC to date
- Approach does not require starting from survivor serum sample isolation or library screening
- Demonstrated feasibility in pilot with a major pharma company by re-designing antigens for multiple antibody targets.
- Efforts currently focused on designing monoclonal antibodies against SARS-CoV-2

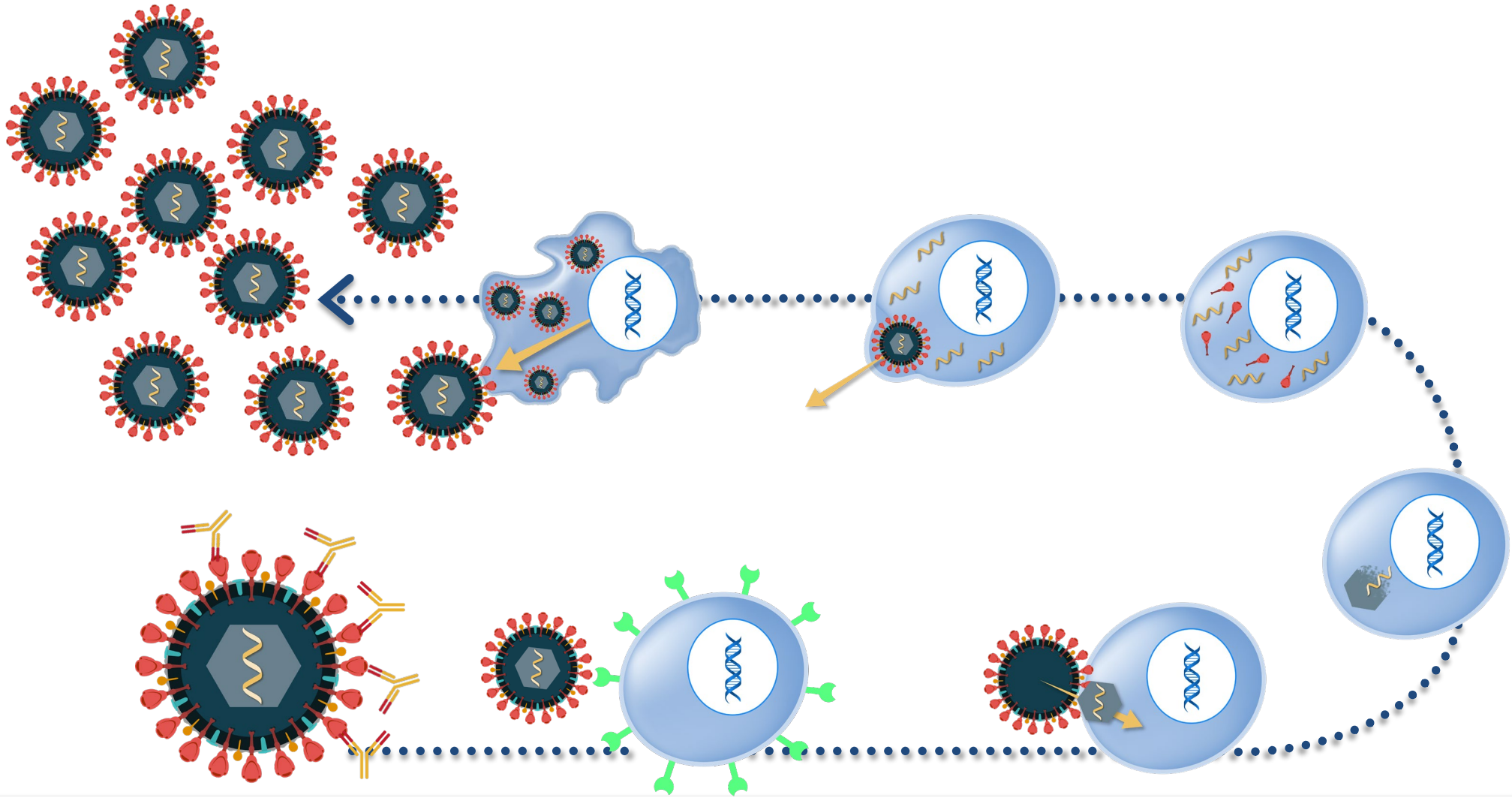
Viruses reproduce by entering and hijacking host cells



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

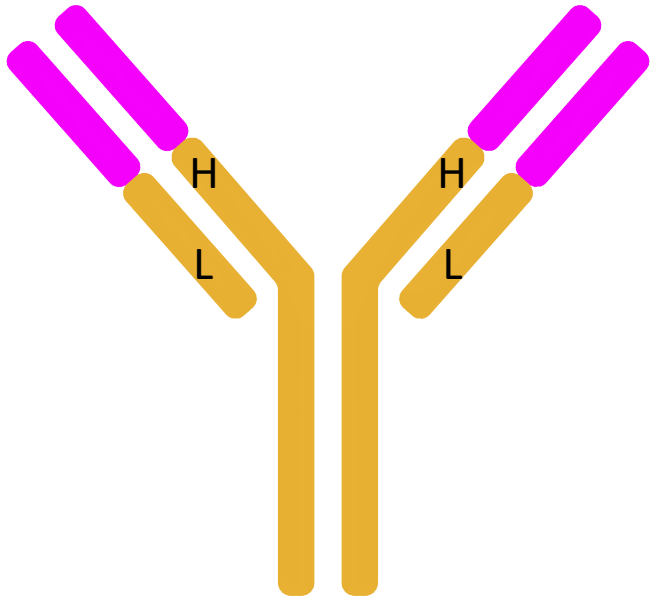


Neutralizing antibodies can stop viral entry



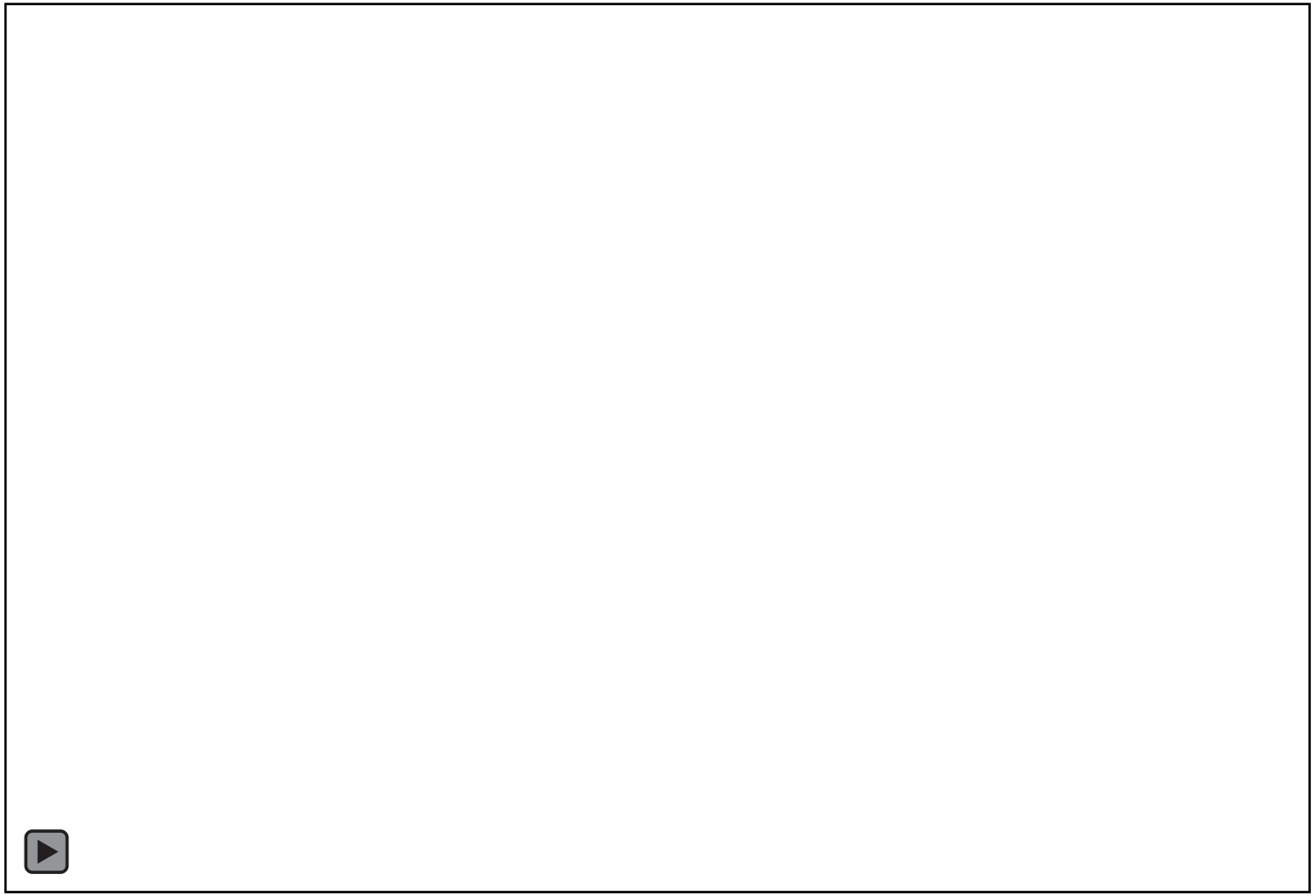
Antibodies are human proteins that specifically and sensitively recognize pathogen targets

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

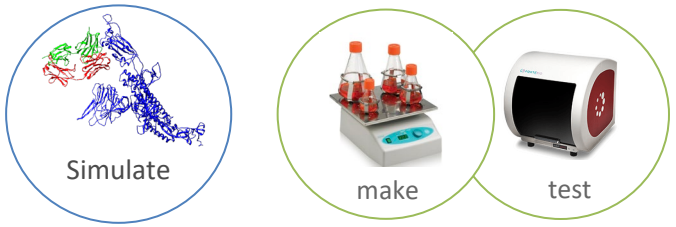


> m396 light chain
SYELTQPPSVSVAPGKTARITCGGNNIGSKSV
HWYQQKPGQAPVLVVYDDSDRPSGIPERFS
GSNSGNTATLTISRVEAGDEADYYCQVWDSS
SDYVFGTGTKVTVLGQPKANPTVTLFPPSSE
EFQANKATLVCLISDFYPGAVTVAWKADGSP
VKAGVETTKPSKQSNNKYAASSYLSLTPEQW
KSHRSYSCQVTHEGSTVEKTVAPTECS

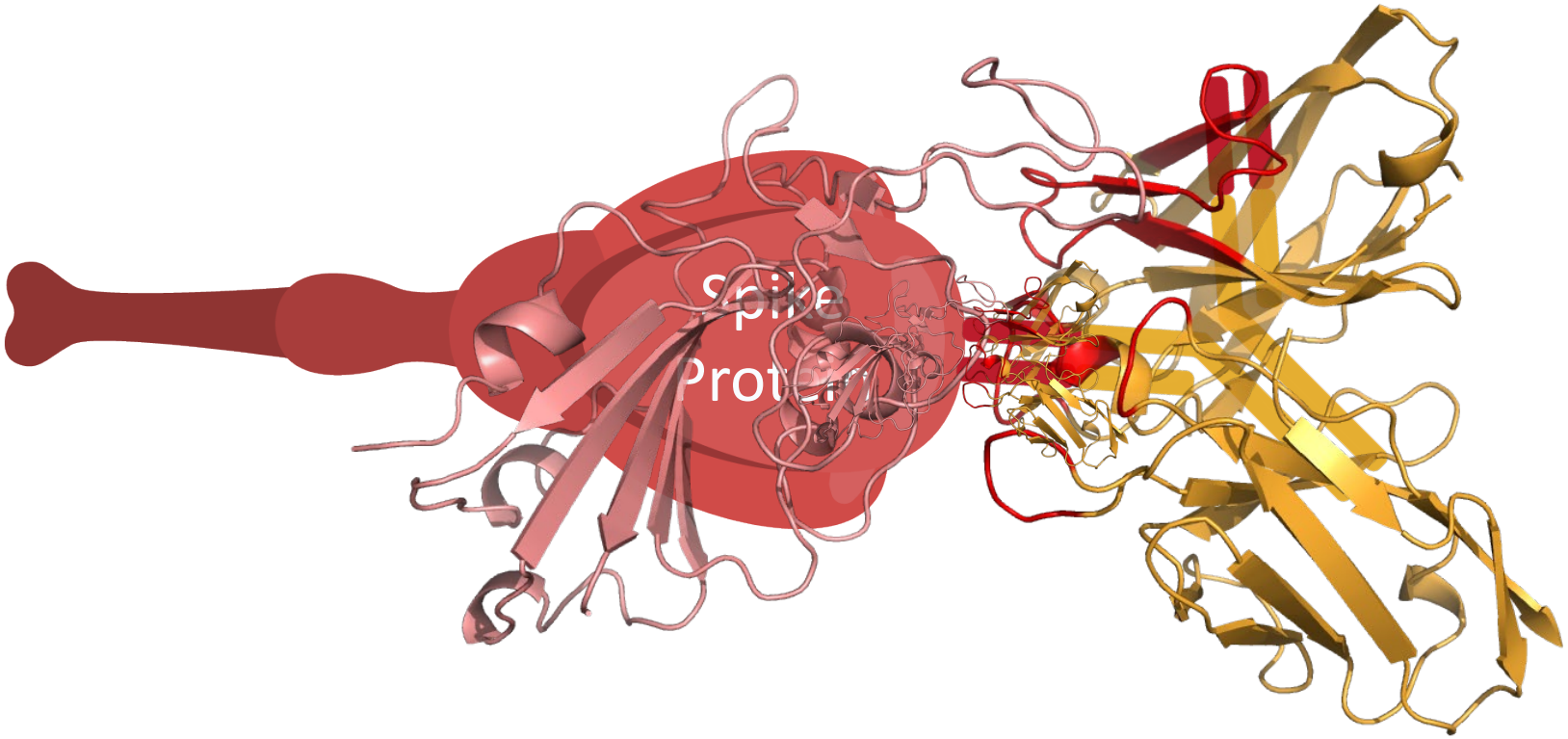
Sequence encodes 3D structure; structure mediates function



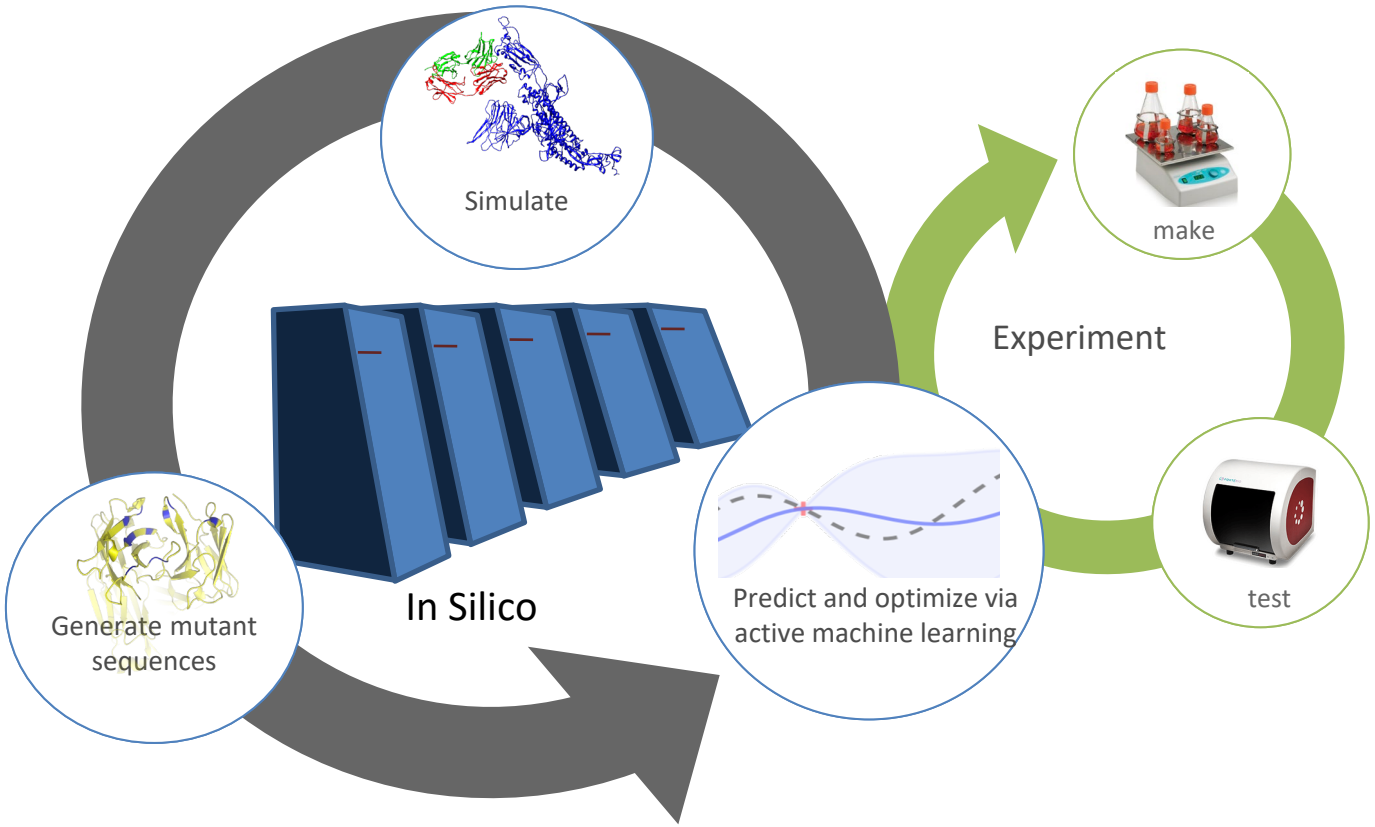
Strong binding is our main target



- 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)



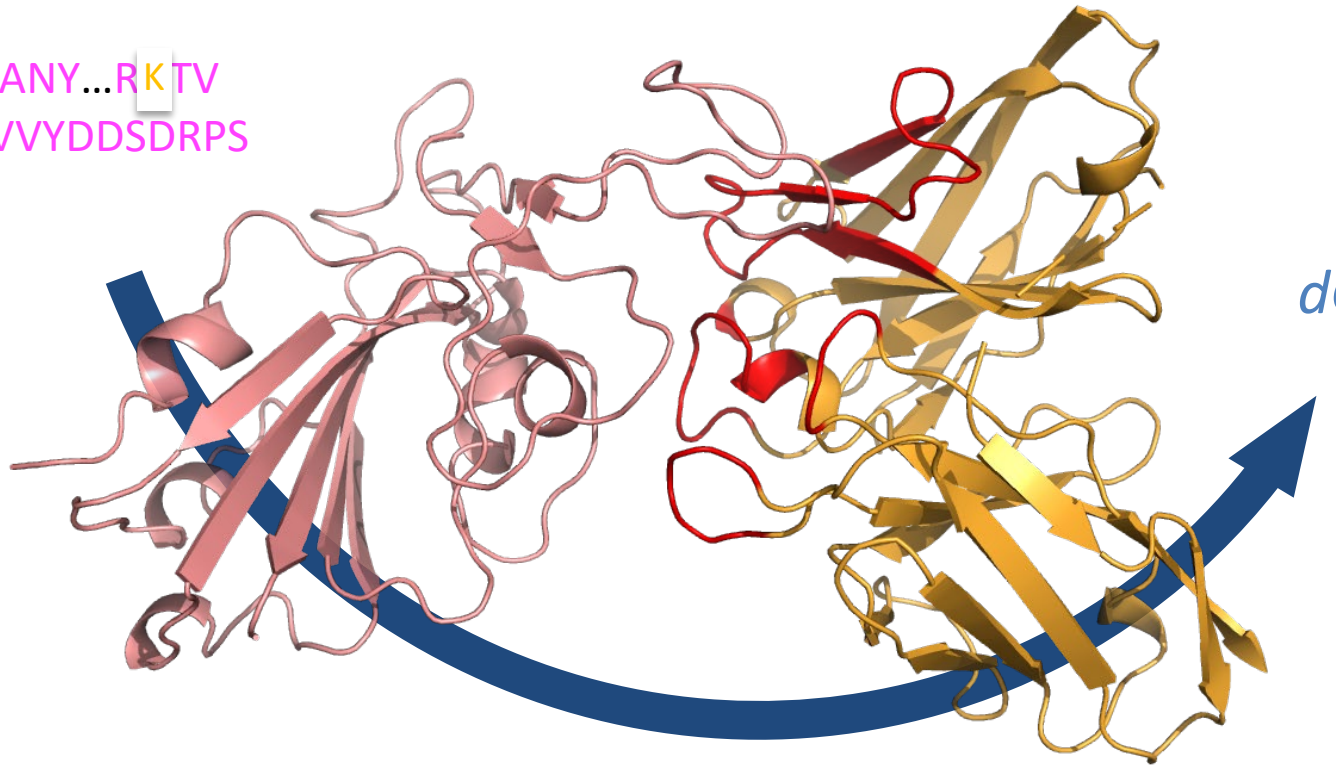
Platform software and active machine learning support these simulation and experimental tools



Pose this problem as active learning

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

> m396 mutable residues
…GTFSSYTIS…WMGGSPILGIANY…RKTV
MGGMDV…/…NIGSKSVH…LVVYDDSDRPS
…QVWDSSSDY

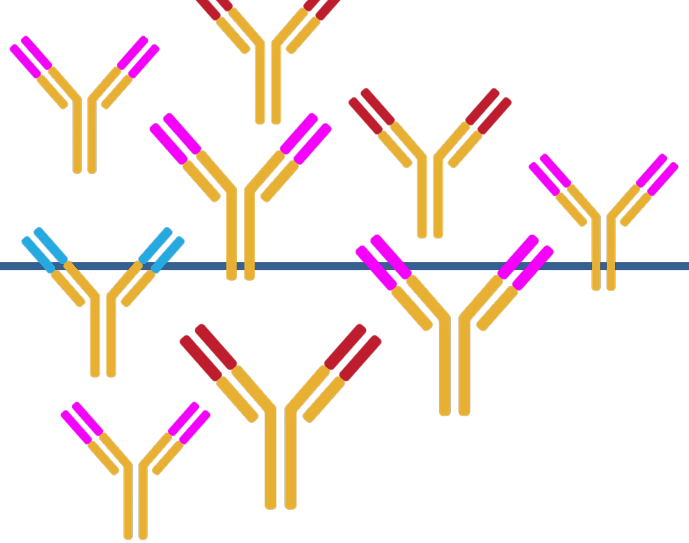


dG, *ddG*, or K_D

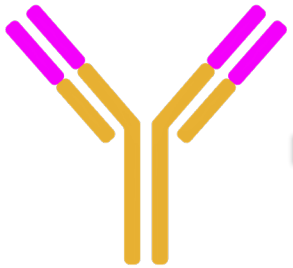
How do you get started?

- *"De novo"* antibody design is a major challenge for the field.

- In the last year, we've started from template antibodies that neutralize the closely related SARS-CoV-1 virus (early 2000's), but *don't* neutralize SARS-CoV-2.
- Likely several mutational steps away from any related, SARS-CoV-2 neutralizers, impractical to search this in the lab
- **HOWEVER,** we have good reason to believe that the interaction we're trying to "restore" should be neutralizing for SARS-CoV-1 too.



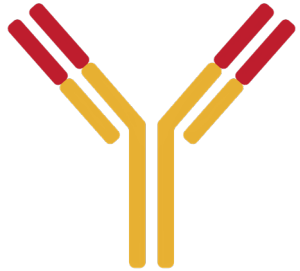
The design space is vastly larger than what we can simulate or test



CoV-1 +
changes
~10^{30}

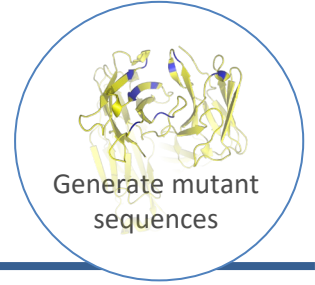
Computer
Simulations
1,000,000

Laboratory
Experiments
100-1,000



CoV-2
Need just one!

Enumerate many antibody designs

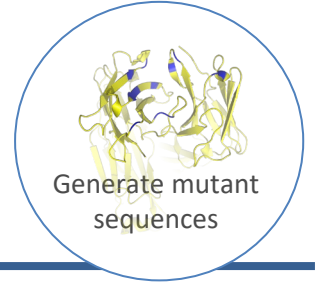


> m396 mutable residues
…GTFSSYTIS…WMGG S PILGIANY…R K TV
MGGMDV…/…NIGSKSVH…LVVYDDSDRPS
…QVWDSSSDY

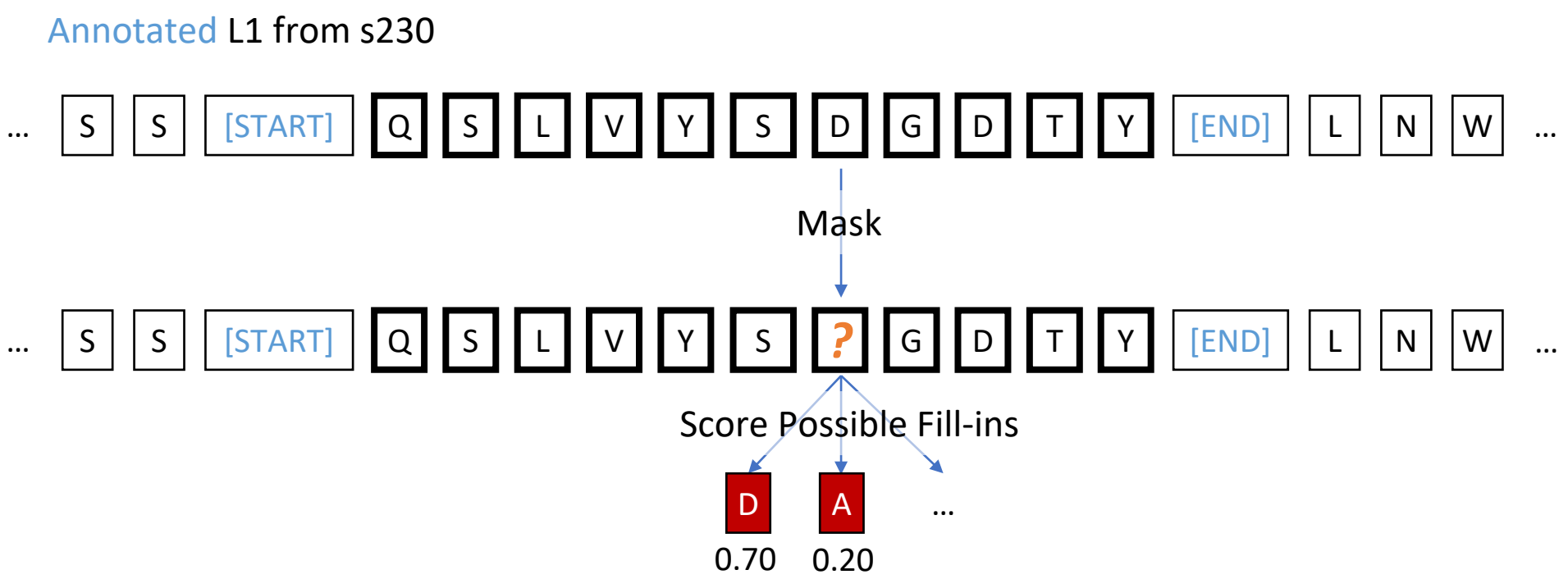
- Generators for novel sequences have so far been mostly tabular
  - Based on frequency of “typical” mutational “swaps”
  - OR based on expensive, high-fidelity calculations of single changes to template antibody in hypothesized complex with SARS-CoV-2 spike.

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

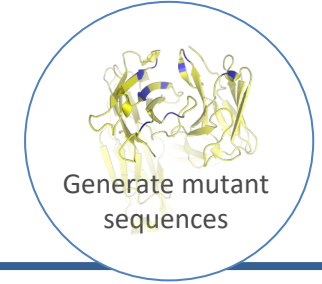
More realistic antibody sequences via language modeling



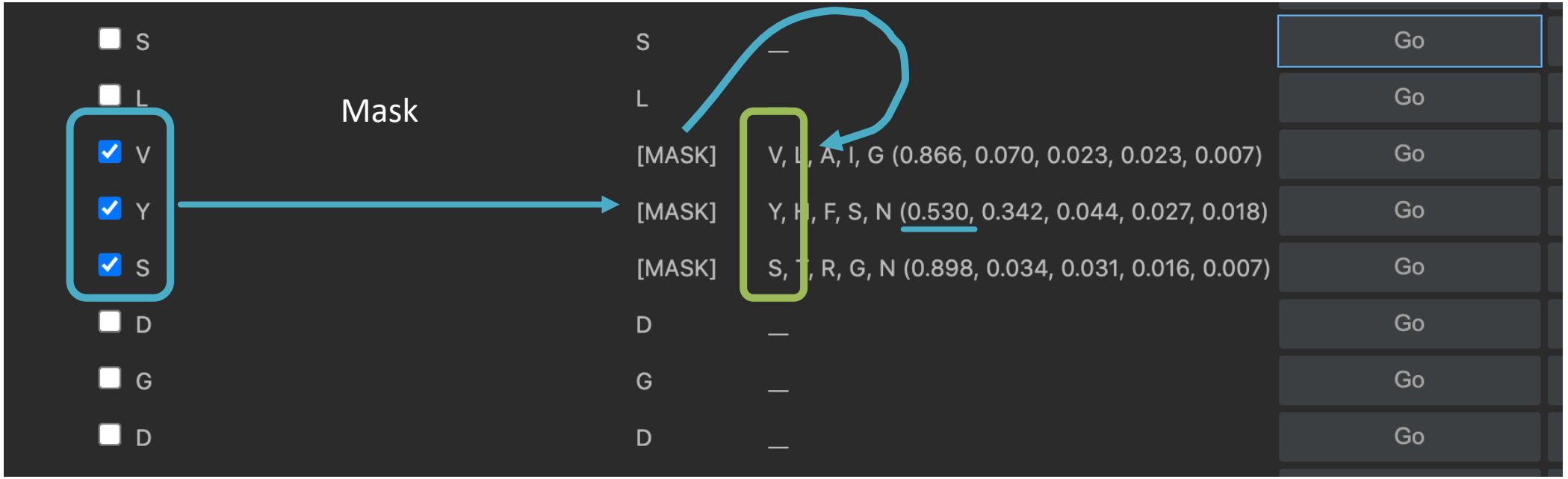
- Use a transformer model to learn to fill “masked” amino acids in the antibody sequence



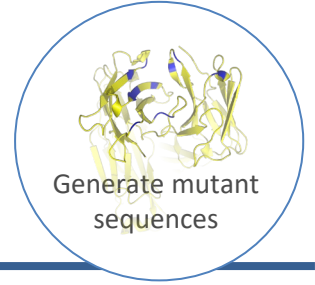
Our models learn to produce reasonable antibodies



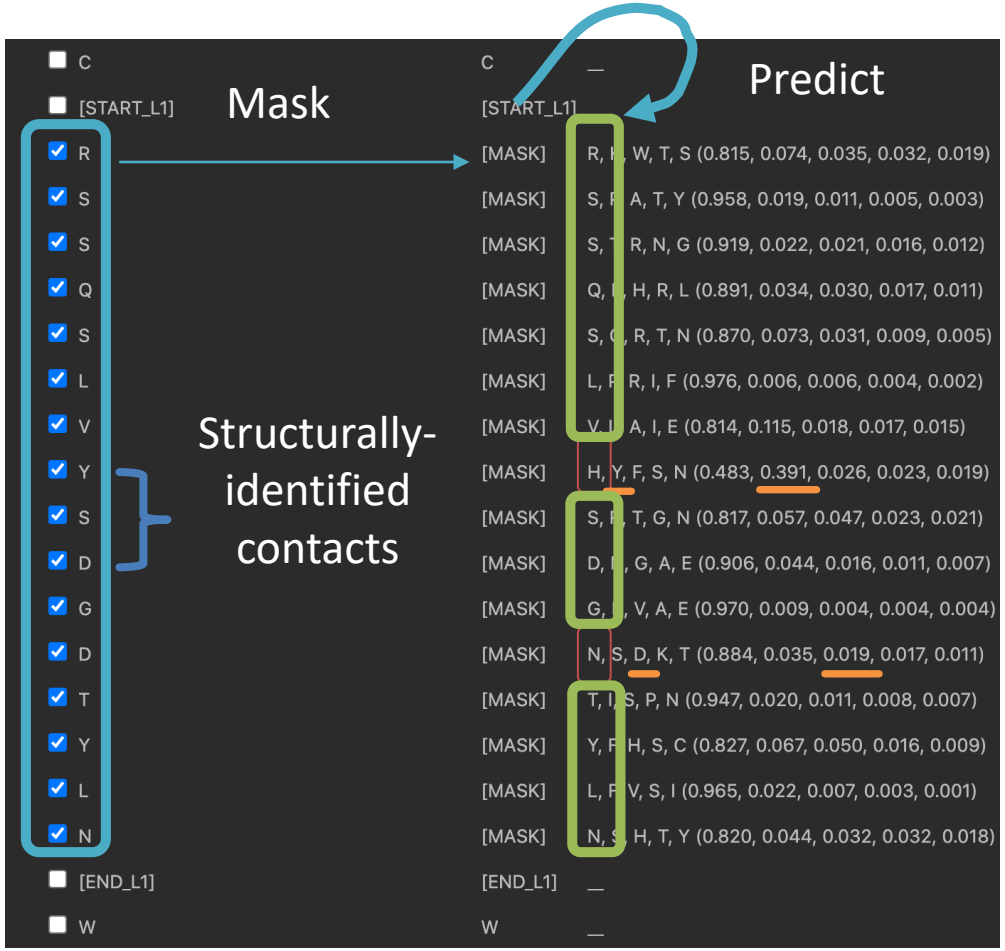
mask and predict 3 central amino acids of s230's L1 "loop"



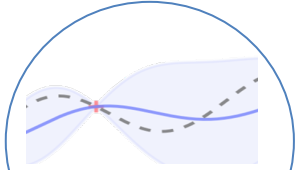
Our models learn to produce reasonable antibodies

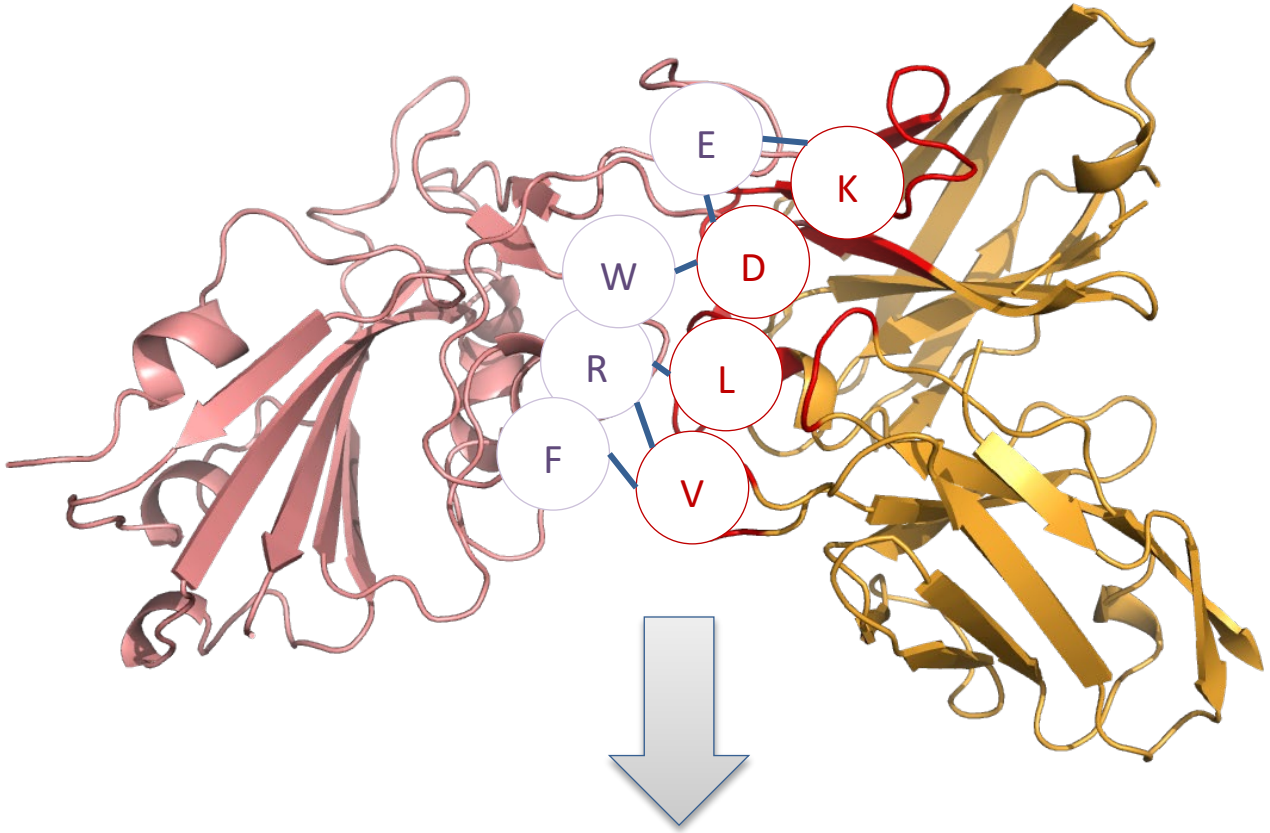


mask and predict all 16 amino acids of s230's L1 "loop"



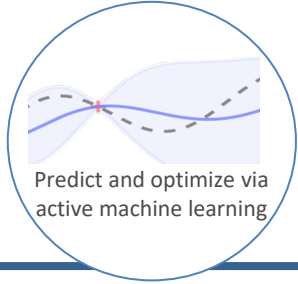
To predict how an antibody sequence will bind, we use a structure-based representation of the interactions





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

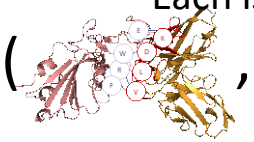
Represented in feature space, binding free energy estimates feed into a multi-fidelity Gaussian process model



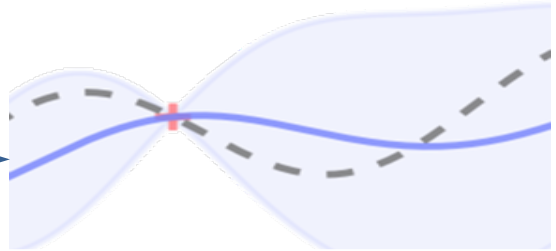
Executed studies

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

Each is a tuple:

( , FoldX, -4.3)

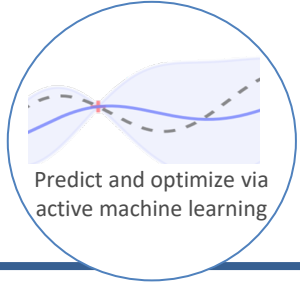
Observations



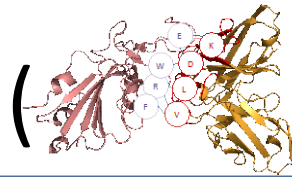
Gaussian process model:

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

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



Decision set: PROPOSED studies

( , FoldX,__)

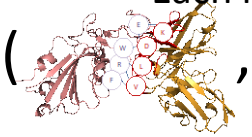
Selection:

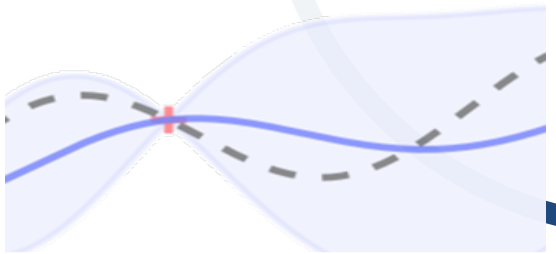
$(x, i)_t = argmax\ s$

Executed studies

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

Each is a tuple:

( , FoldX, -4.3)

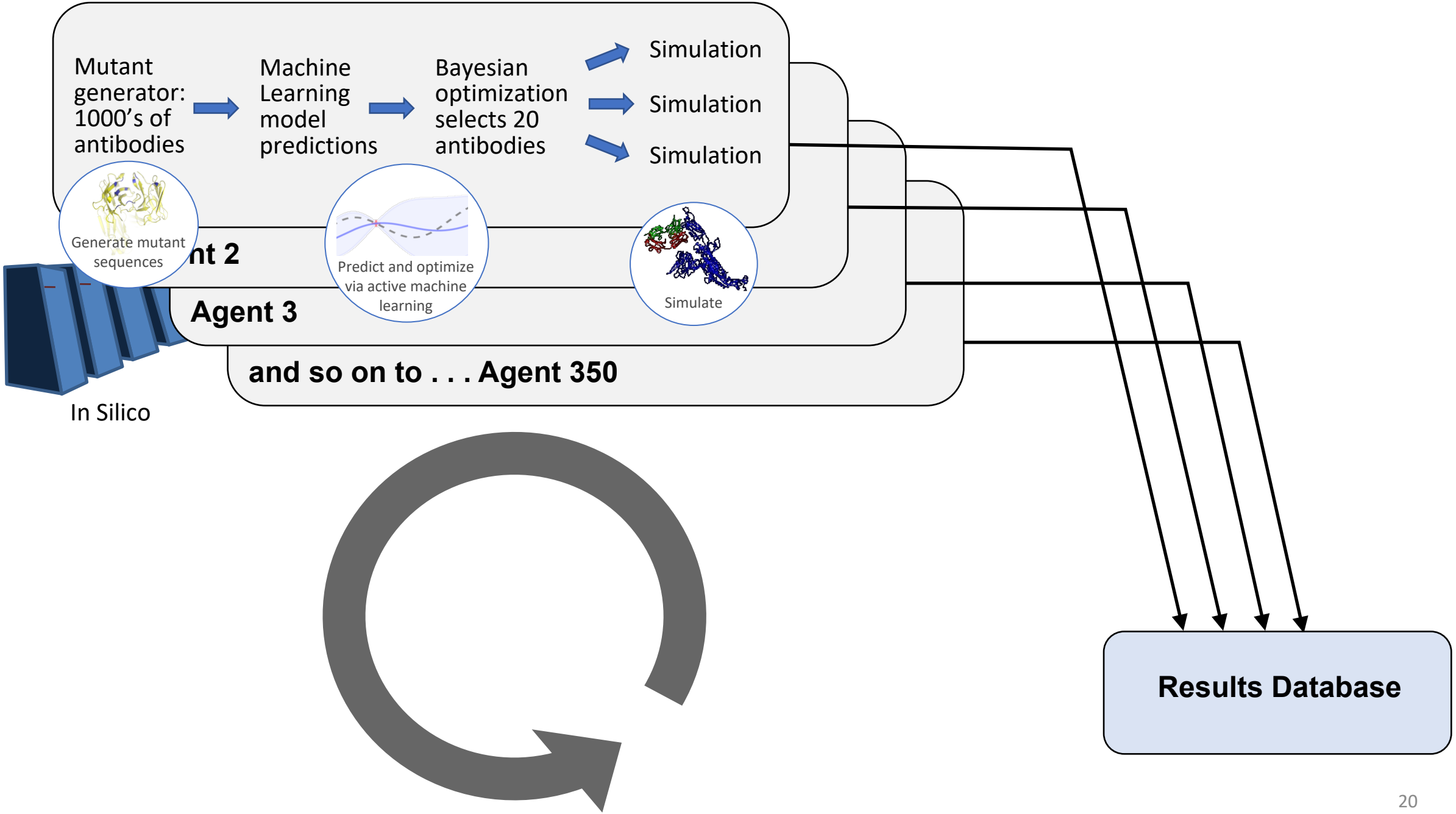


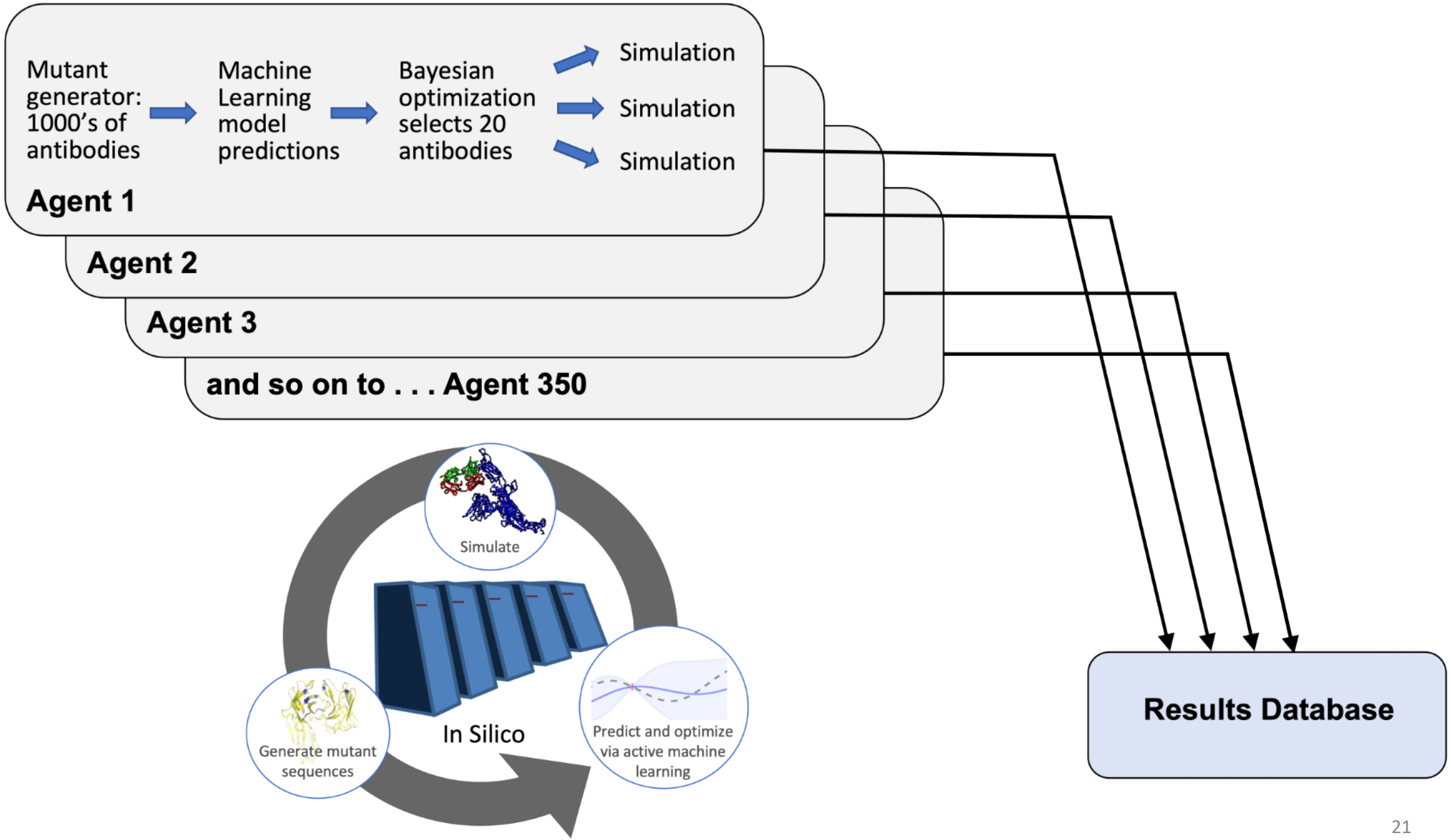
Gaussian process model:

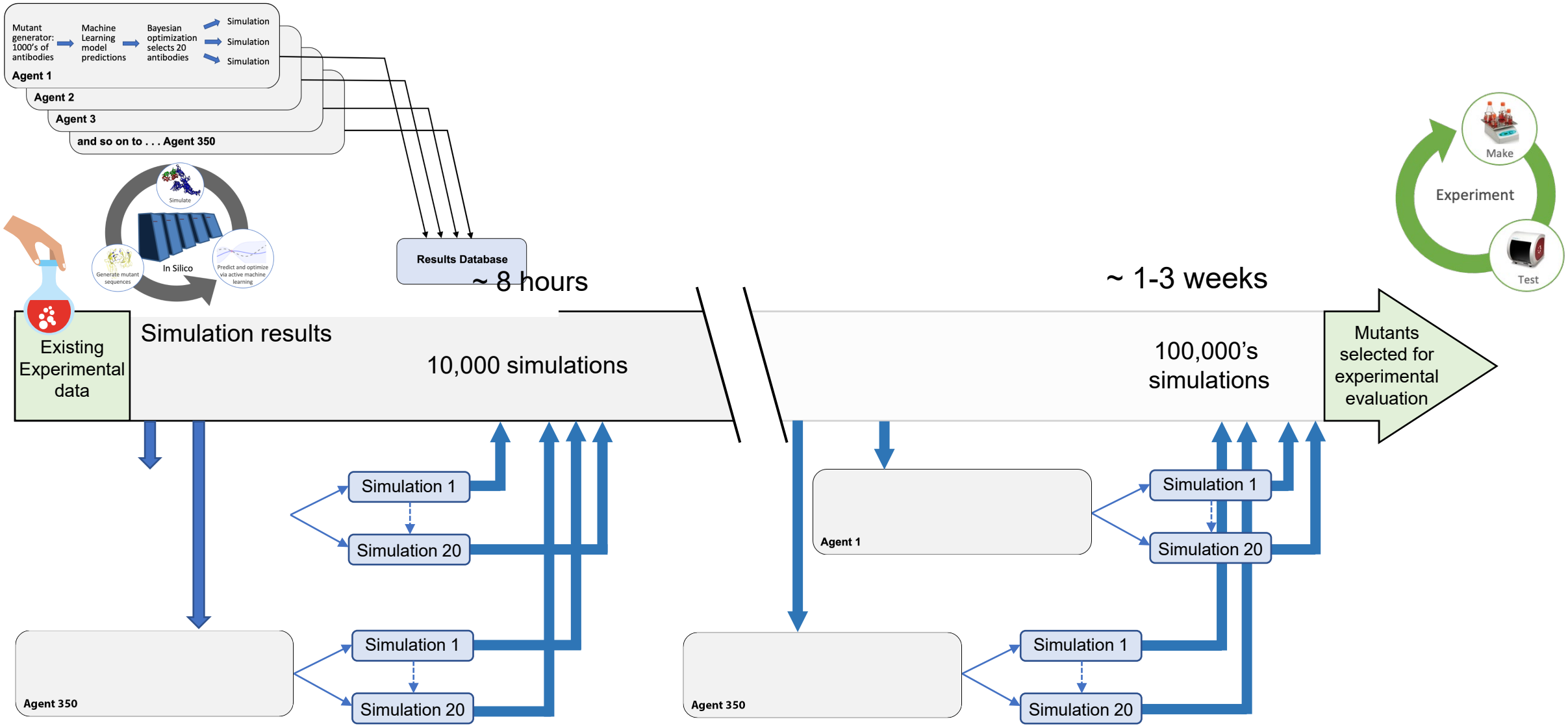
$f: (x, i) \rightarrow \mu(x, i), \sigma(x, i)$

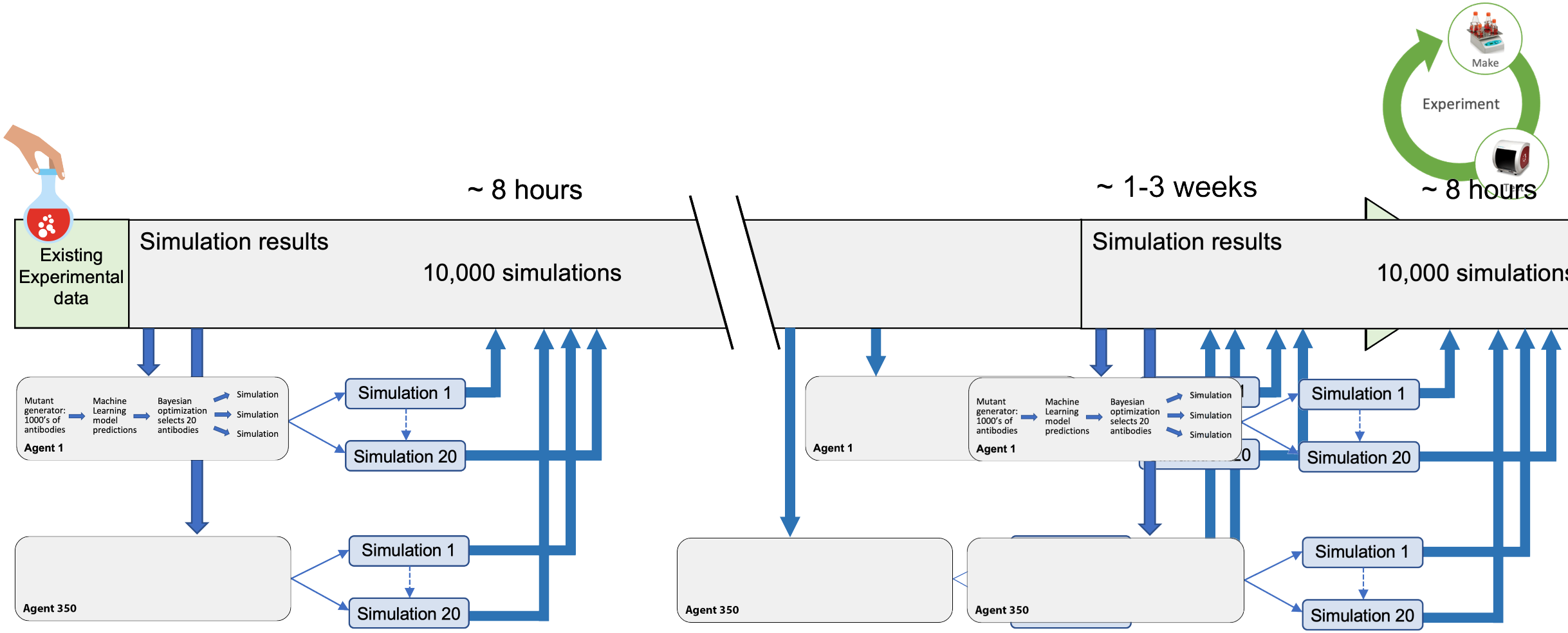
Score:

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

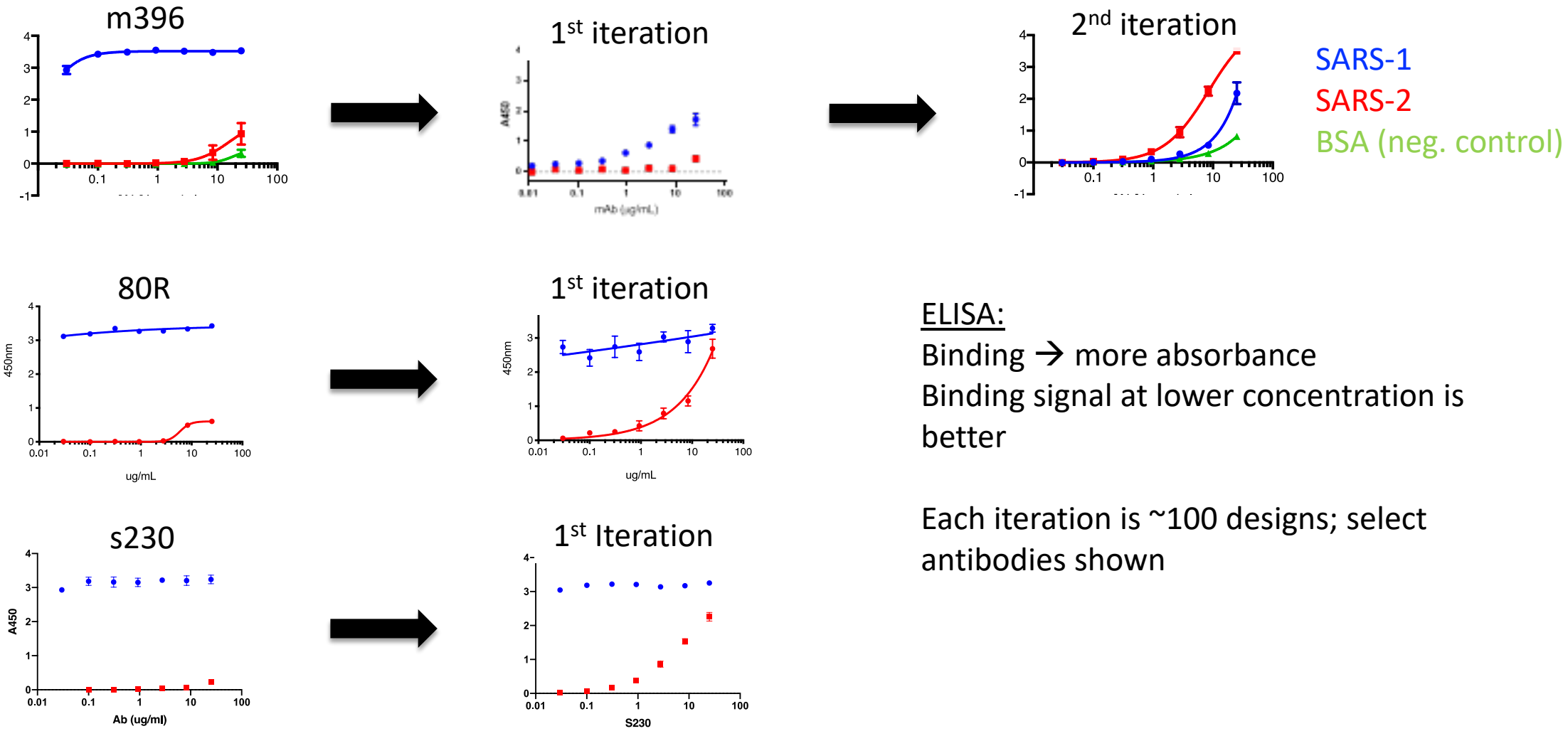

Mutant generator: 1000's of antibodies
Machine Learning model predictions
Bayesian optimization selects 20 antibodies
Simulation
Simulation
Simulation
Generate mutant sequences
Predict and optimize via active machine learning
Simulate
nt 2
Agent 3
and so on to . . . Agent 350
In Silico
Results Database


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

For two systems, we designed binders without having received *any* experimental feedback



ELISA:

Binding → more absorbance

Binding signal at lower concentration is better

Each iteration is ~100 designs; select antibodies shown

This work is the product of a multidisciplinary team

- LLNL:

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

- Sandia NL:

  Brooke Harmon, Oscar Negrete, Max Stefan

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