Argonne NATIONAL LABORATORY

Targeting SARS-CoV-2 with AI- and HPC-enabled Lead Generation

Yadu Babuji, Ben Blaiszik, Tom Brettin, Kyle Chard, Ryan Chard, Austin Clyde, Ian Foster, Zhi Hong, Shantenu Jha, Zhuozhao Li, Xuefeng Liu, Arvind Ramanathan, Yi Ren, Nicholaus Saint, Marcus Schwarting, Rick Stevens, Hubertus van Dam, Rick Wagner, Logan Ward

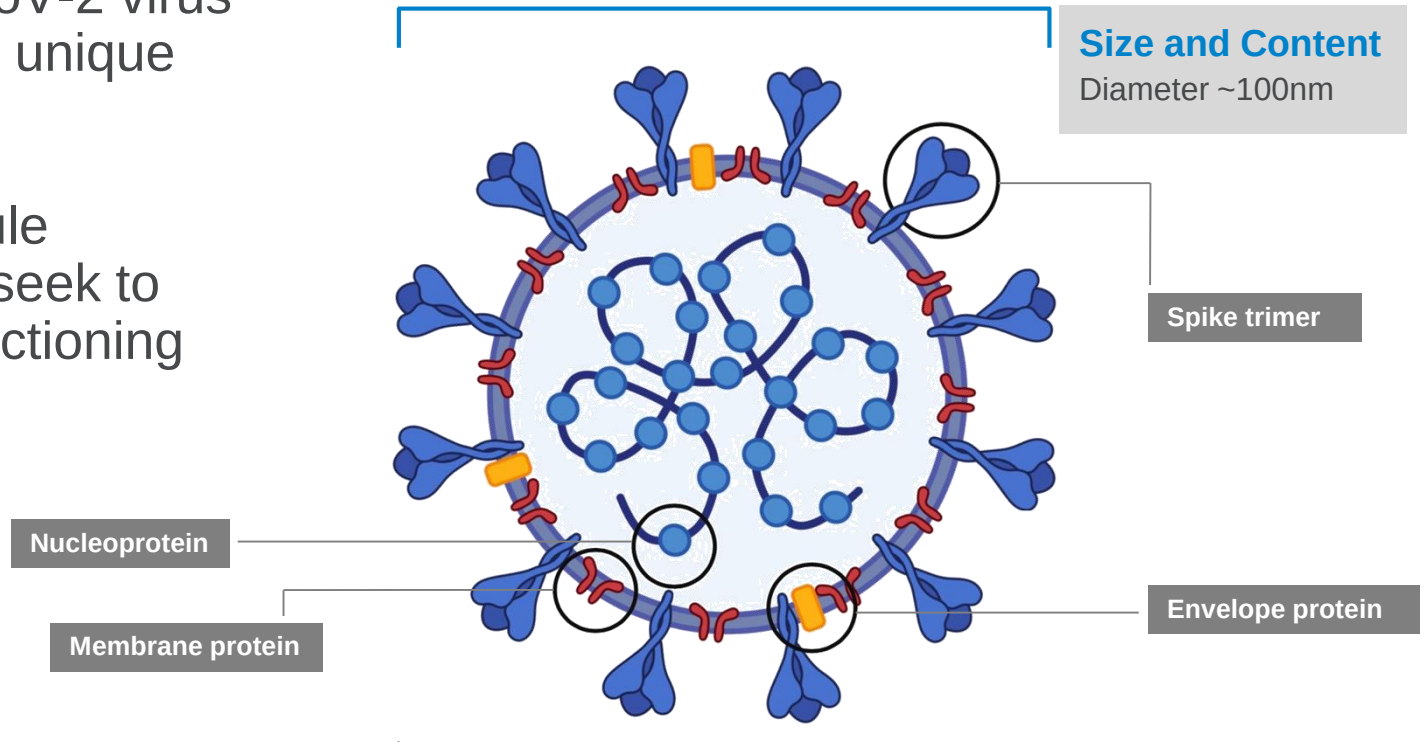
U.S. DEPARTMENT OF ENERGY Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.

globus labs

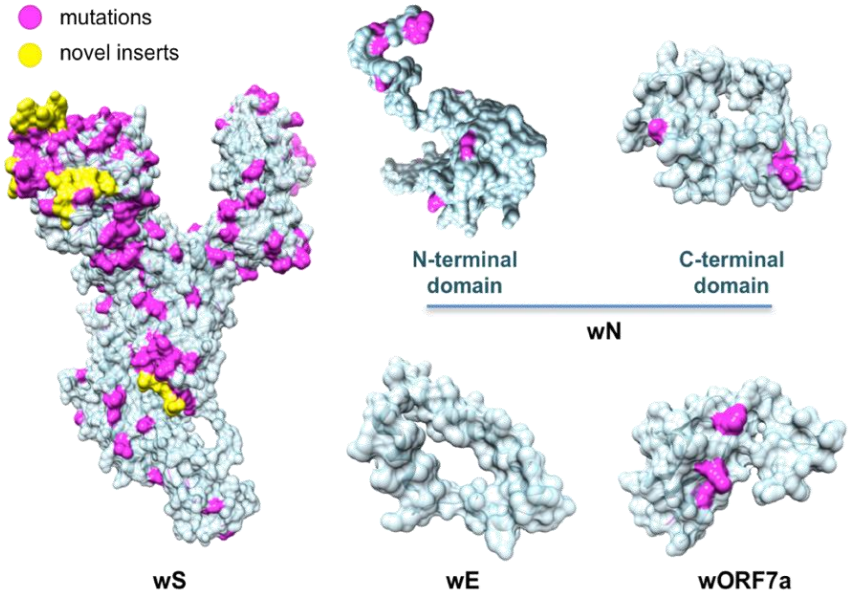
Small-molecule therapeutics against COVID

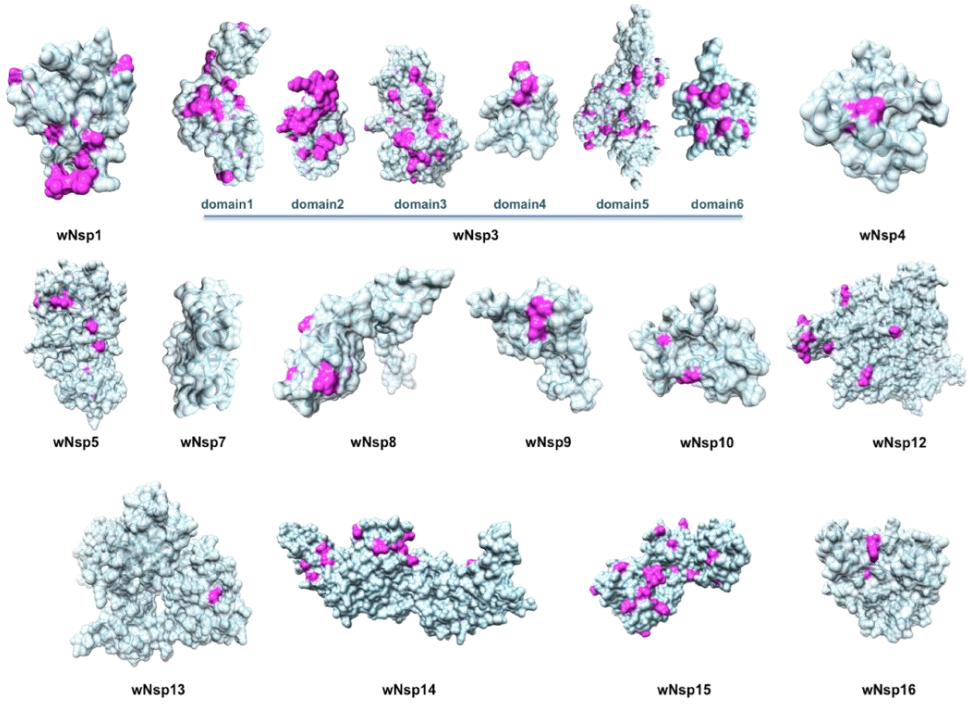
The SARS-CoV-2 virus comprises 28 unique proteins

Small-molecule therapeutics seek to disrupt its functioning



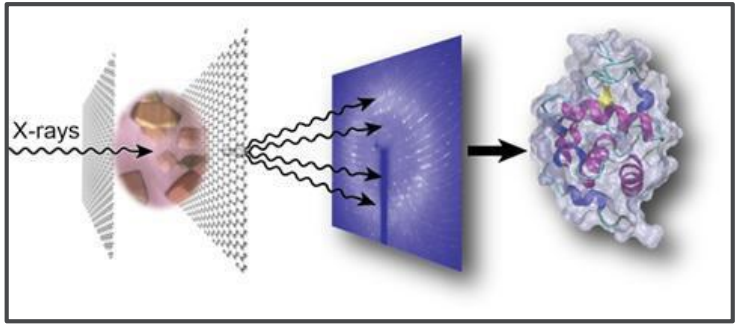
Structures exist for most of the key proteins



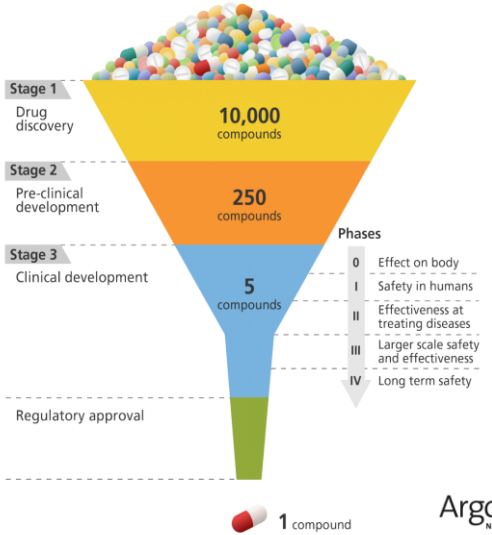


Leveraging Argonne resources in the fight

Serial X-Ray Crystallography (SSX)
Solving complex COVID protein structures with high-throughput imaging at near room temperature (structures provide insight into drug and vaccine development)



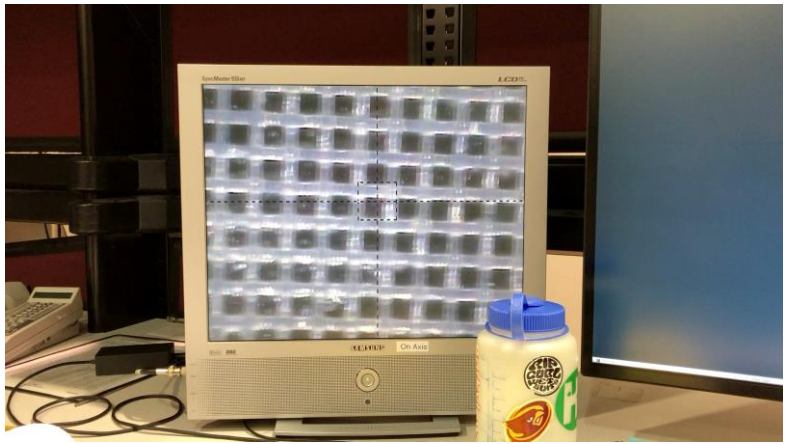
Extreme-scale machine learning
Applying machine learning models to screen potential therapeutics based on molecule databases and computed molecular features



Enabling Serial Crystallography (SSX) at Scale

Connecting light sources and leadership computing facilities to enable new science

- Perform serial imaging of chips with thousands of embedded protein crystals
- Analyze batches of images as collected
- Report statistics and and summary images during experiment
- Return crystal structure to scientist

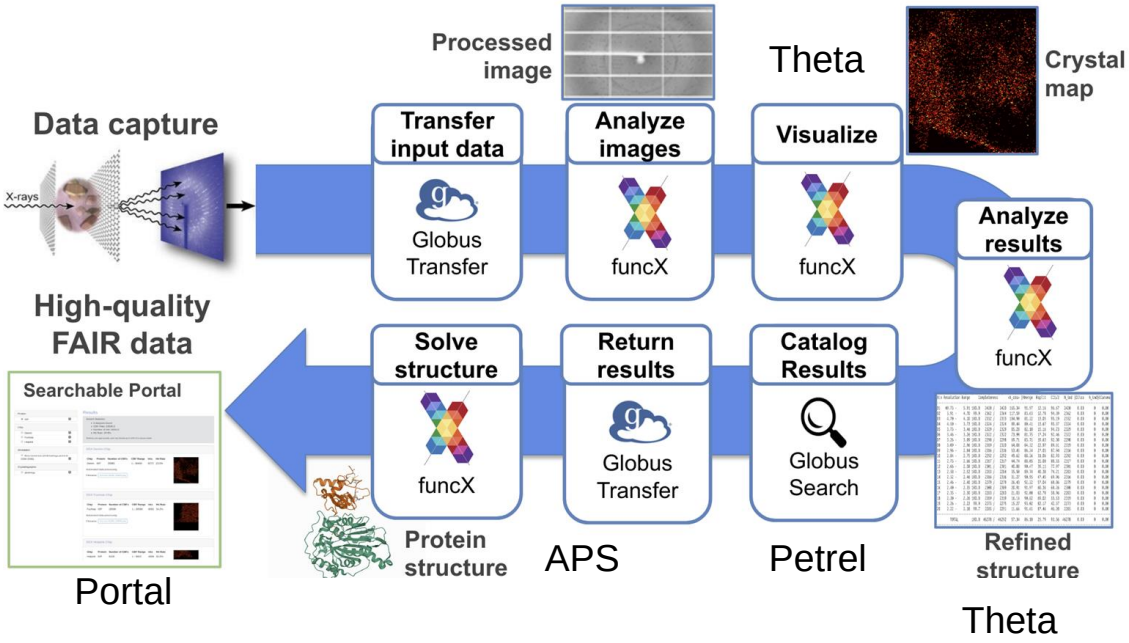


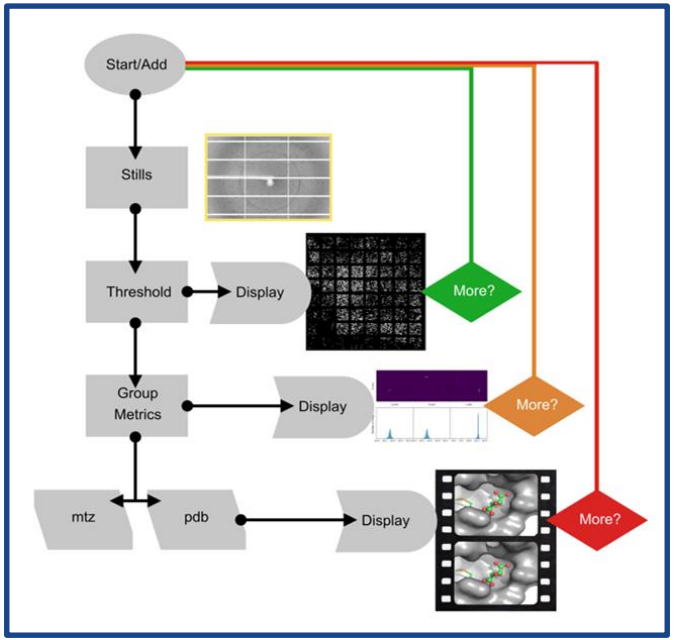
Ryan Chard et al.

Automating and scaling the analysis of SSX data

Globus Automate flow to batch files, move data to ALCF, perform analysis using funcX/Parsl, and catalog results

Integrates with APS DM to trigger flow, ALCF resources for computing, and ALCF portal for monitoring experiments and reprocessing data

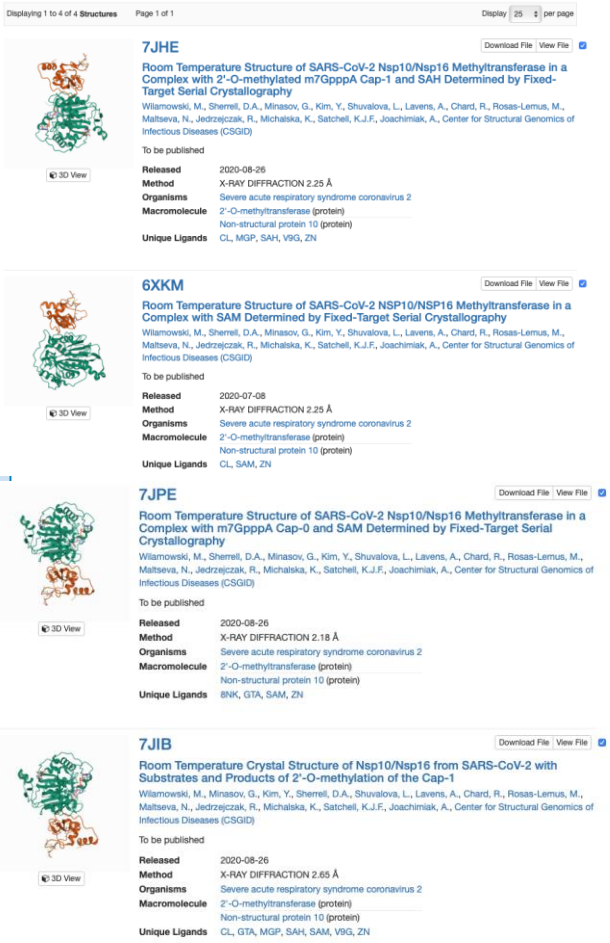




Understanding COVID protein structure with SSX

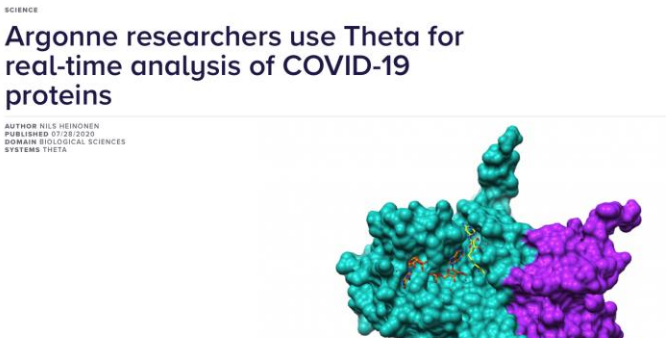
ALCF + APS capabilities were used to determine the room temperature structure of 2 viral surface proteins

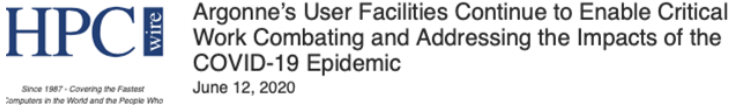
4 structures are now available in PDB



"These data services have taken the time to solve a structure from weeks to days and now to hours"

Darren Sherrell, SBC beamline scientist APS Sector 19





Challenges screening potential candidates

- Many molecules (>10^9 drug-like molecules in collected databases)
- Testing in the wet lab is very expensive, clinical trials even more so
- Protein docking simulations are computationally expensive

⟶ apply machine learning methods to predict which molecules have a high likelihood of docking

AI and supercomputers can accelerate drug discovery

CHEMICAL LIBRARY DATABASE

4B known molecules

Enamine

DRUGBANK

BindingDB GDB

eMolecules

cureFFI MOSES

ZINC15

LINCS

SureChEMBL

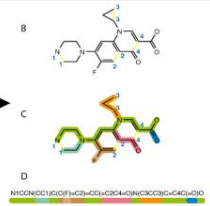
PubChem

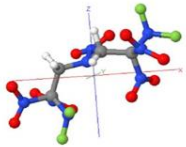
AND MORE

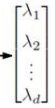
CANONICALIZATION

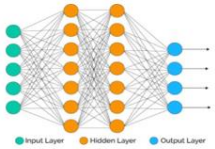
COMPUTE FEATURES

DEEP LEARNING FILTERING



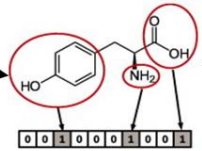


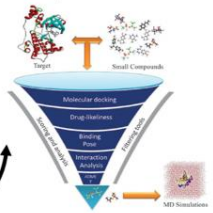




FINGERPRINTING

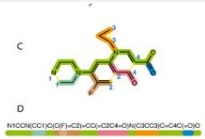
SIMILARITY SEARCH

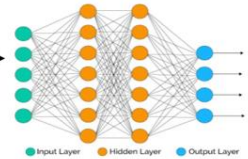




GENERATE IMAGES

CNN FILTERING





COMPUTING RESOURCES

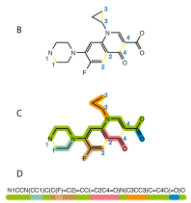
Parsl-computed features for AI-based drug screening

23 input datasets, **4.2B molecules**, **60 TB** of molecular features and representations

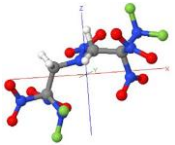
Parsl processing pipelines used ~2M core hours on ALCF Theta, TACC Frontera, OLCF Summit

1. Convert each molecule to a **canonical SMILES**
2. For each molecule, compute:
   a. ~1800 2D and 3D **molecular descriptors** using Mordred
   b. **Molecular fingerprints** encoding structure
   c. **2D images** of the molecular structure

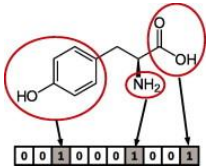
Computed data provide **crucial input features to AI models** for predicting molecular properties such as docking scores and toxicity



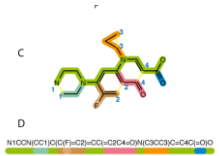
Canonical SMILES
23 CSV files with 4.2B molecules



Mordred Descriptors
420,130 CSV files, 48.70TB



Molecular Fingerprints
4,221 CSV files with base64 encoded fingerprints, 578.27GB



2D images
420,707 Pickle GZ files, 11.48 TB