

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





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



CONTRACTOR AND A CONTRACT A CONTRACTACT A CONTRACT A CONTRACTACT A CONTRACTACT A CONTRAC



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/ParsI, and catalog results

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



U.S. DEPARTMENT OF U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.

With Andrzej Joachimiak, Darren Sherrell et al. APS Sector 19

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



Displaying 1 to 4 of 4 Structures Page 1 of 1

7JHE

Download File View File

Display 25 \$ per page

Download File Maw File

Download File View File



6XKM



7JPE

7JIB

Room Temperature Structure of SARS-CoV-2 Nsp10/Nsp16 Methyltransferase in a Complex with m7GpppA Cap-0 and SAM Determined by Fixed-Target Serial Crystallography Wiamowski, M., Shemel, D.A., Minasov, G., Kim, Y., Shuvalova, L., Lavens, A., Chard, R., Rosas-Lemus, M.,

Maltseva, N., Jedrzejczak, R., Michalska, K., Satchell, K.J.F., Joachimiak, A., Center for Structural Genomics of Infectious Diseases (CSGID)

```
To be published
Released 2020-08-26
Method X-RAY DIFFRACTION 2.18 Å
Organisms Severe acute repiratory syndrome coronavirus 2
Macromolecule 2-O-methytransferase (protein)
Non-structural protein 10 (protein)
Unique Ligands 8HK, GTA, SAM, ZN
```



\$13D View

Room Temperature Crystal Structure of Nsp10/Nsp18 from SARS-CoV-2 with Substrates and Products of 2'-O-methylation of the Cap-1 Witamowski, M., Minazov, G., Kim, Y., Sherrell, D.A., Shuxalova, L., Lavens, A., Chard, R., Rosas-Lemus, M., Mattavas, N., Jedzrejczak, R., Michalska, K., Satchell, K.J.F., Joachimiak, A., Center for Structural Genomics of Inflectious Diseases (CSGID) To be published

To be published
Released
2020-08-26
Wethod
X:PAY DIFFRACTION 2.65 Å
Organisms
Severe acute respiratory syndrome coronavirus 2
Macromolecule
C:-O-methytimatenae (protein)
Non-structural protein 10 (protein)
Unique Ligands
C:, (CTX, MGP, SAK, SAM, V96, ZNi

"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

Argonne researchers use Theta for real-time analysis of COVID-19 proteins





SCIENCE

UTHOR NILS HEINONER

PUBLISHED 07/28/2020 DOMAIN BIOLOGICAL SCIENCES SYSTEMS THETA

> Argonne's User Facilities Continue to Enable Critical Work Combating and Addressing the Impacts of the COVID-19 Epidemic

Since 1987 - Covering the Fastest Computers in the World and the People Who

Argonne



Challenges screening potential candidates

- Many molecules (>10⁹ 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



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



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 Al 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



https://2019-ncovgroup.github.io/data/

