CSI Support for the DOE COVID-19 Response

Kerstin Kleese van Dam
Computational Science Initiative

BROOKHAVEN NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

@BrookhavenLab

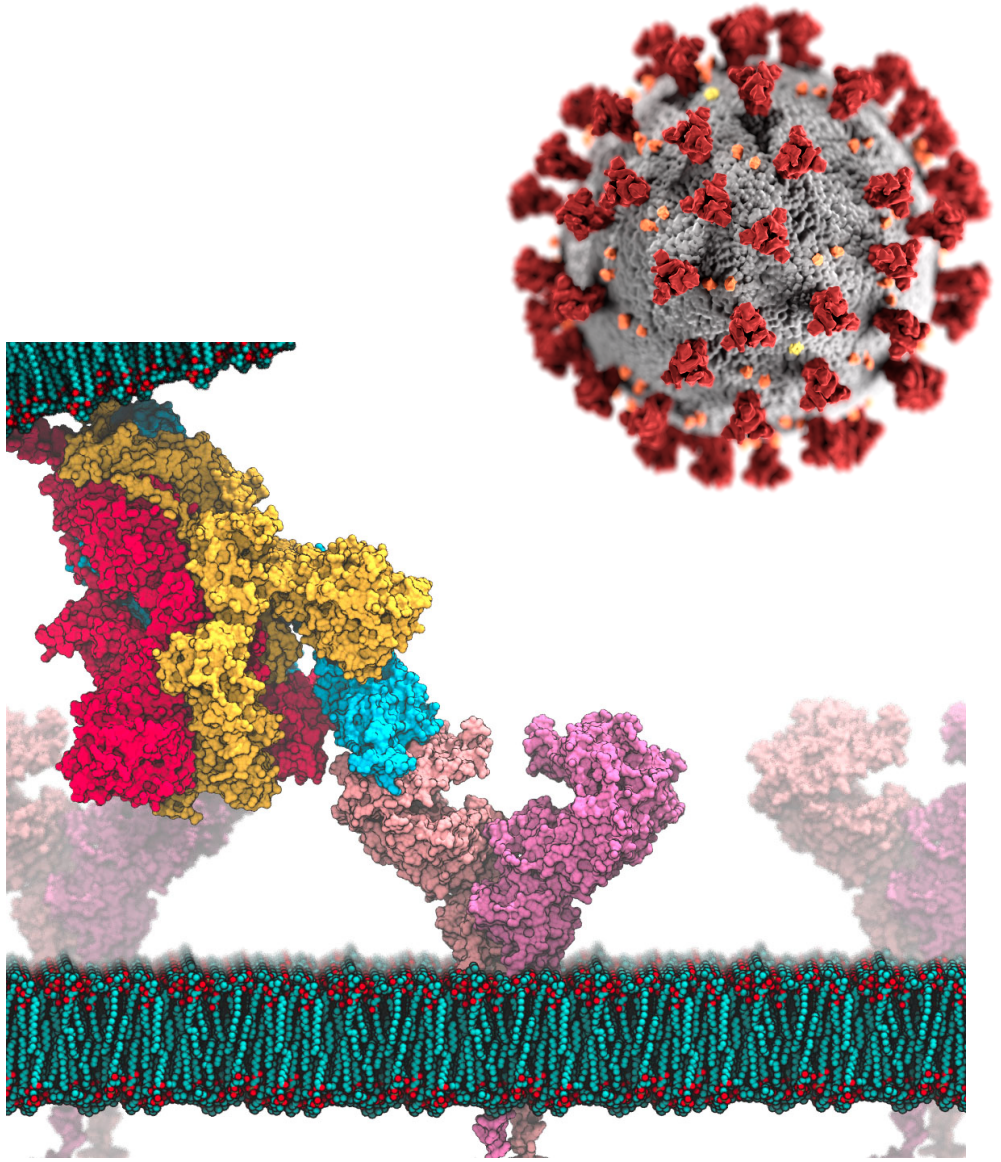
Areas of Engagement

- The Computational Science Initiative (CSI) has been involved in the DOE COVID-19 response since early March.
- Four Distinct Areas:
  - Computational Drug Discovery
  - Literature Evaluation
  - Epidemiology
  - COVID-19 Archive

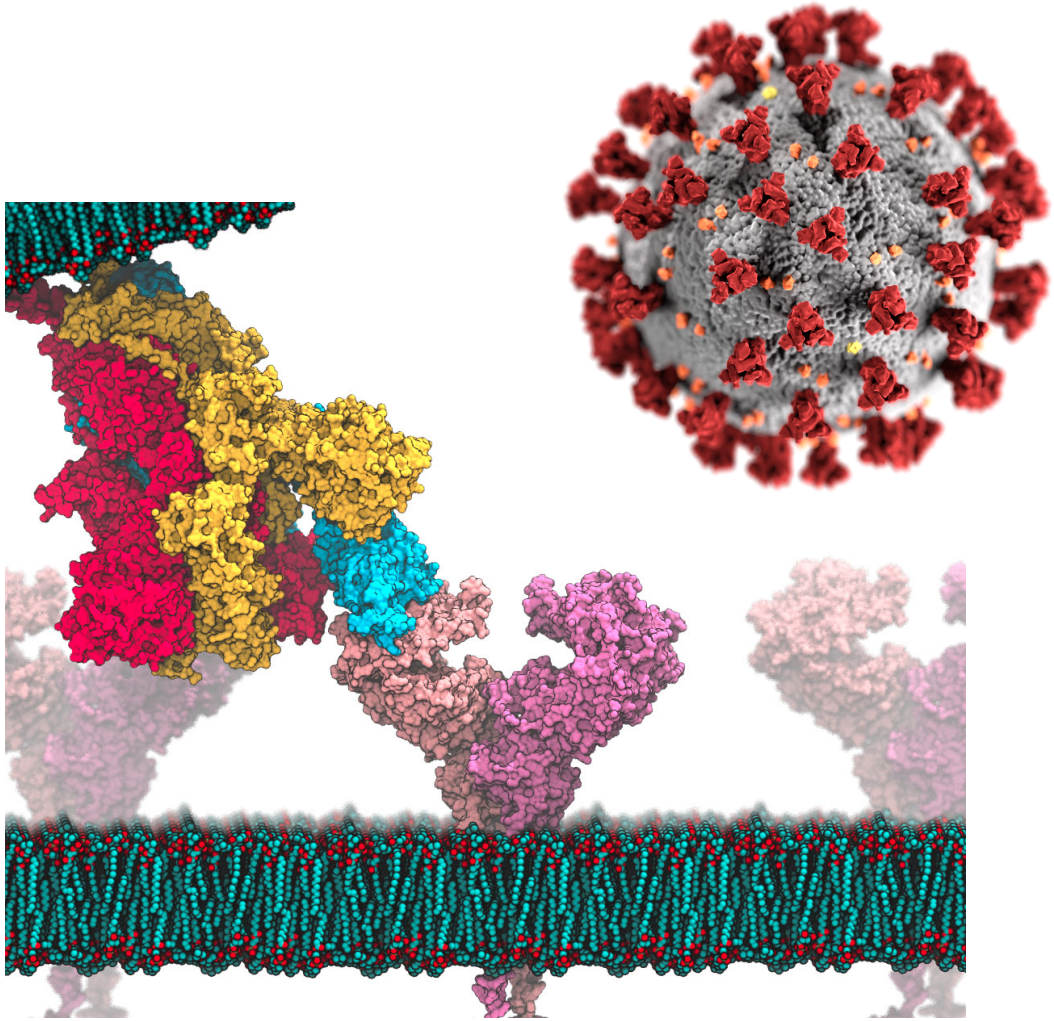
Computational Drug Discovery

Drug Discovery

- Critical task is to find a drug to combat the virus - stopping it from infecting the body or stopping it from multiplying.
- To date, we have identified **4 billion drug compounds** that could potentially combat the virus.
- We have identified **68 possible targets on the virus** where the drugs could attach.
- There is no time to test all of these **68 x 4 billion** possibilities experimentally.
- BNL and partners have developed computational methods to test all possible options and assess which ones are most likely to succeed.



Computational Drug Design

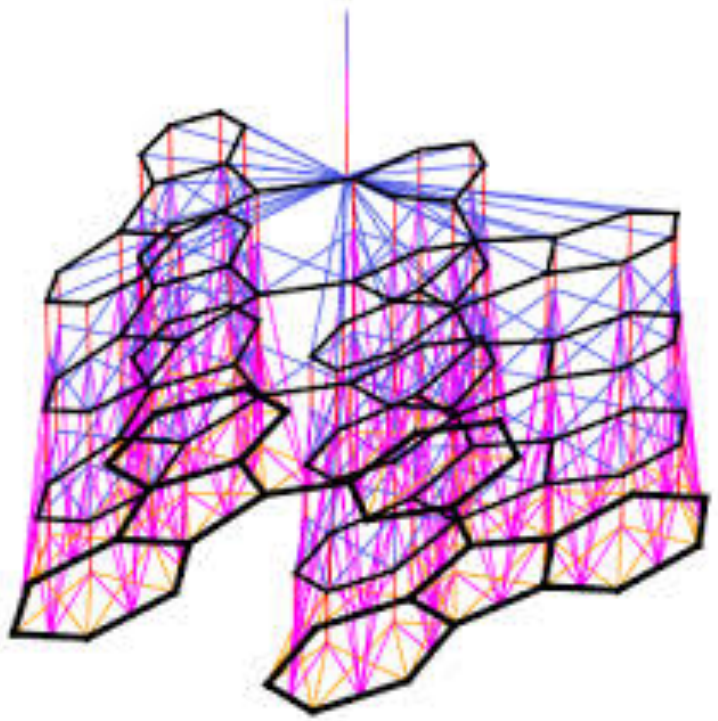


- Using different methods to test if drug compounds would attach to the virus:

- Molecular modeling (see right) can take days on large scale computers - highly accurate
- Biological Docking programs - can take minutes to hours - quite accurate
- AI based methods - often less then a second - 5 Million compounds/hour - basic accuracy
- **We use different methods at different stages to verify leads**

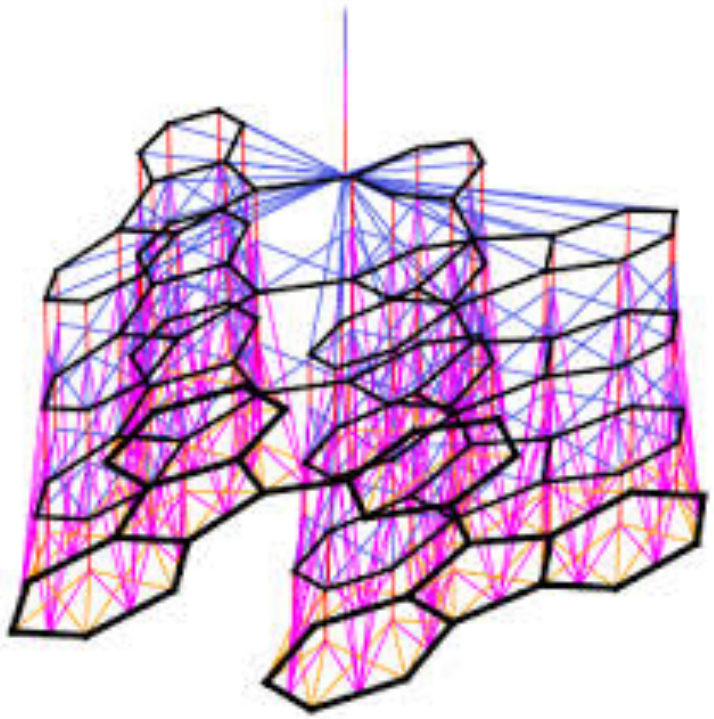
AI - Artificial Intelligence - Based Drug Design

- You can represent molecule as a graph with three types of encoded features: Atom-features, Bond-features, and Graph.
- Put emphasize on the graph components that are important
- **Reinforcement Learning** (RL): Machine Learning method where positive outcomes are rewarded, or **reinforced**, in the model during training.
- RL models can start from an empty graph, and add one atom or a bond at a time.
- Our RL model can generate molecules for certain targeted properties that makes them more likely to succeed against the virus



Molecular Fingerprint from: Convolutional Networks on Graphs for Learning Molecular Fingerprints. David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre Rafael Gomez-Bombarelli, Timothy Hirzel, Alan Aspuru-Guzik, Ryan P. Adams, Harvard University

Artificial Intelligence and Machine Learning Research and Development

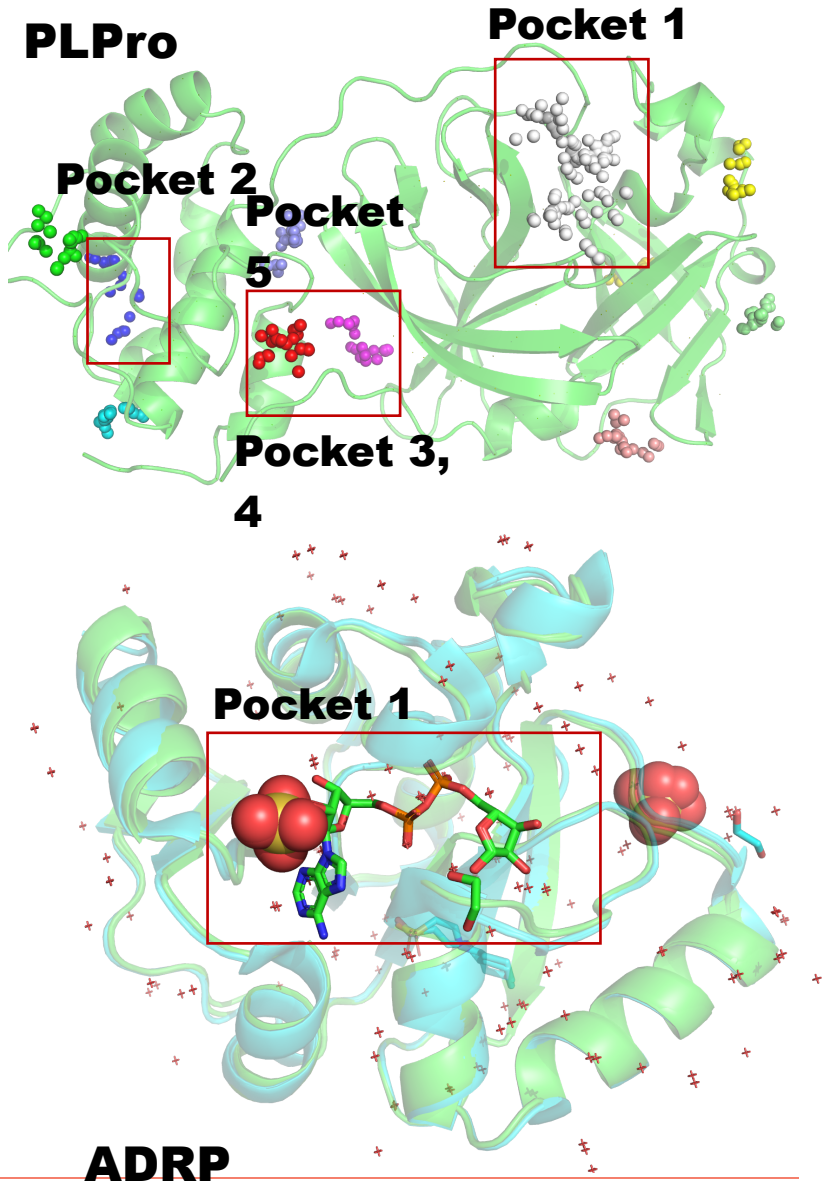


Molecular Fingerprint from: Convolutional Networks on Graphs for Learning Molecular Fingerprints. David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre Rafael Gomez-Bombarelli, Timothy Hirzel, Alan Aspuru-Guzik, Ryan P. Adams, Harvard University

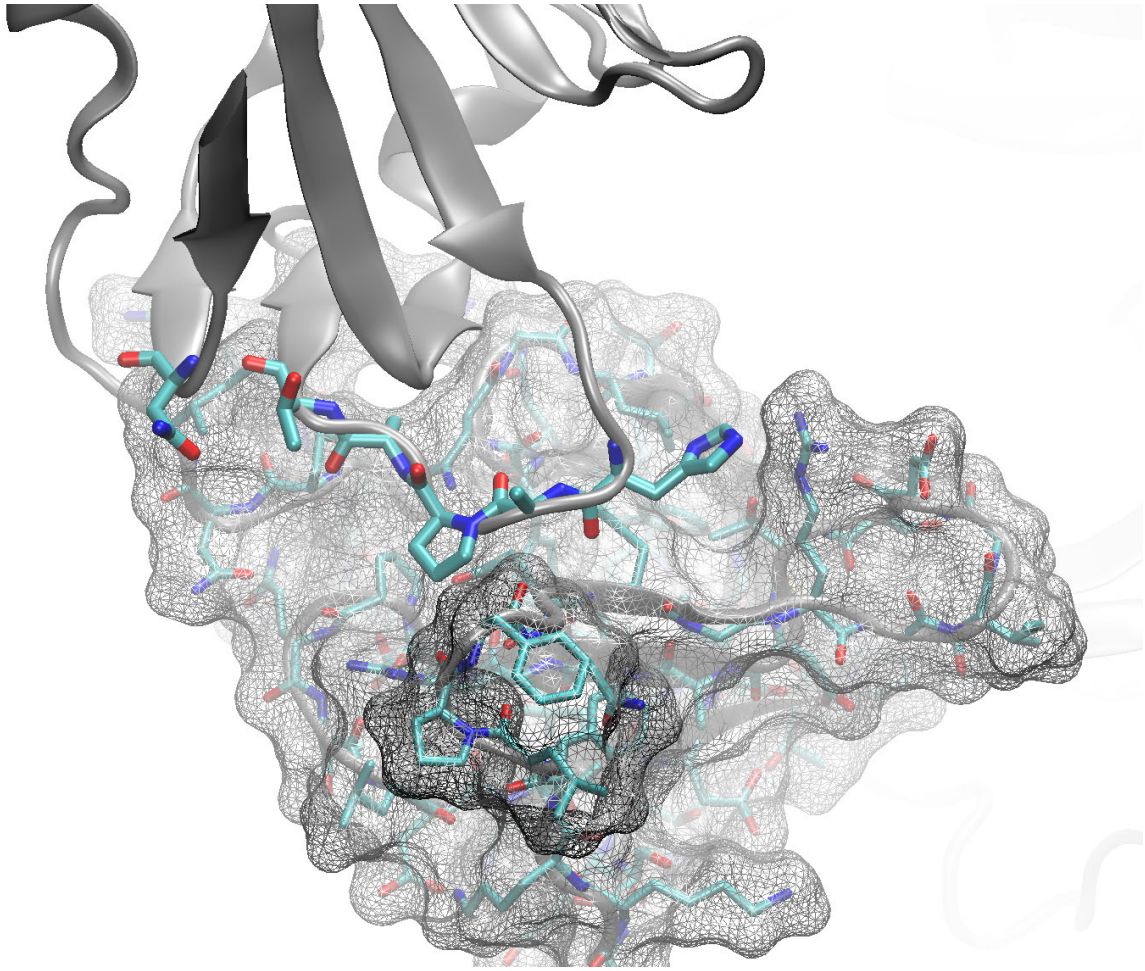
- **Neural Fingerprints** (NFP):
  - Taking the Molecular Graph we can put emphasize on the graph components that are important
  - We then Ggenerate a vector from the graph to **create a "fingerprint." representing the key features of the compound**
  - The 'fingerprint' is then fed to an AI model to predict the likelihood that is will bind to the virus or the host.
  - **We can currently scan 5M Molecules/hour and are working to reach 10M/hour**

Docking Studies

- **We validate the top results of the AI studies with ‘Docking Studies**
- Drug design is based on Lock & Key model
  - The protein pockets are the lock.
  - A drug candidate is the key.
- Using the biological and chemical knowledge of the molecule and the part of the virus it may attach to, we determine how likely and strongly such a connection might be.
- At BNL we evaluated **300K potential compounds against 25 possible virus targets**, with two different software packages
- Using various BNL computers in CSI, NPP and at the NSLS II



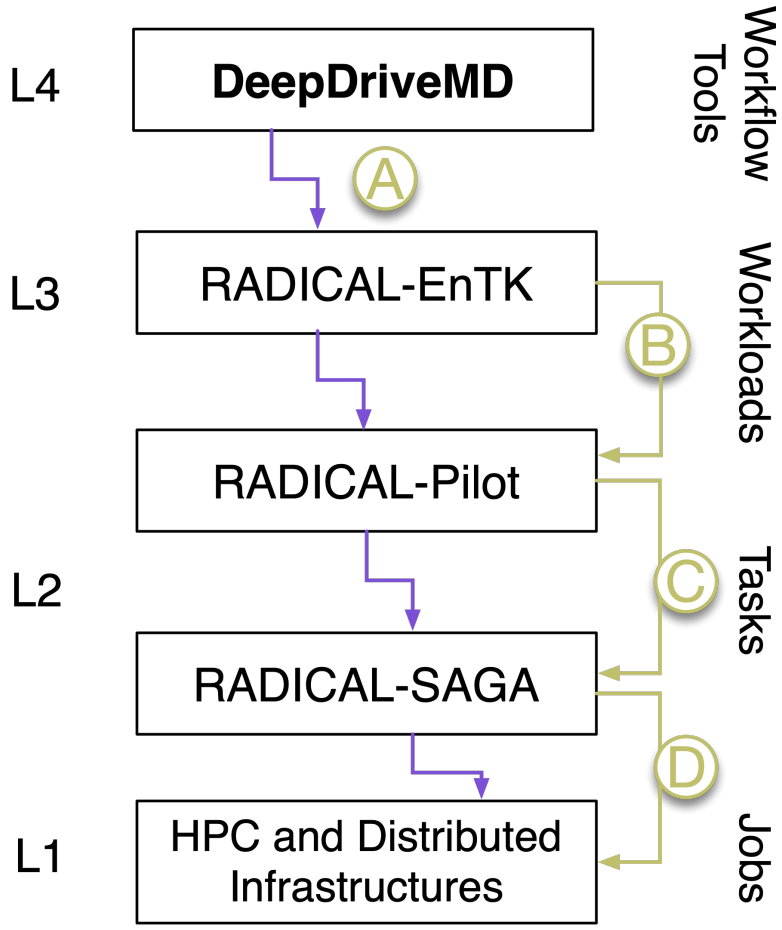
Detailed Molecular Modeling



- Build accurate models of both part of the virus including one pocket and the host
- Observe how the virus infects the host
- Study how this process could be interrupted and at which stage
- Test highly ranked drug compounds to see if they have the desired impact
- We found a pocket where the right drug could allow the immune system to gain access to the virus

Analyzing Billions of Options

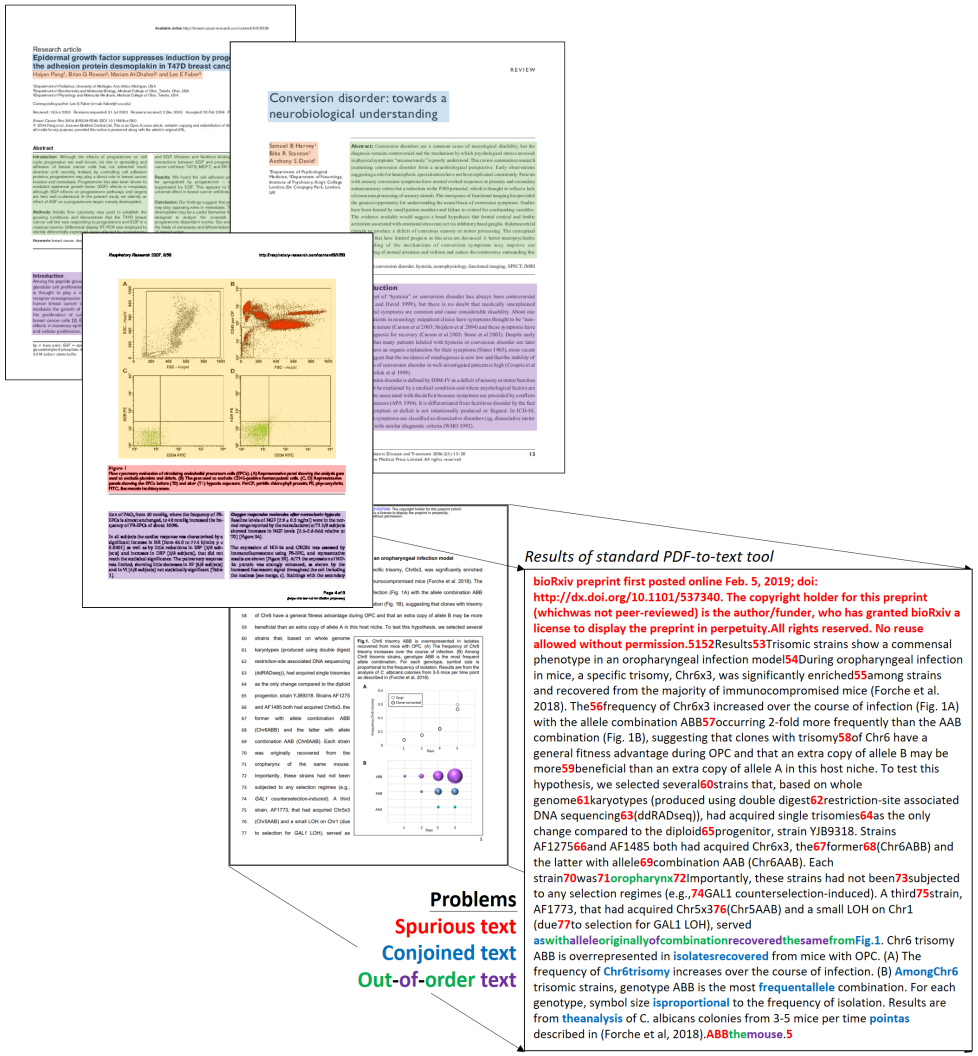
- To analyze so many options in a short time with the available **compute resources** requires skill and expert computational help.
- Optimized Resource Provision for Computational Workflows: assuring the codes have the **compute** and **storage** they need when they need it.
- Optimized workflows can now analyze **5M Molecules /per hour** (SMILES are identifiers for drug-like molecules).
- **Identified 30+ lead molecules** that have been submitted to experimental and medical experts for further analysis.



Literature Service

BNL CSI

Searching the Literature for the Researchers



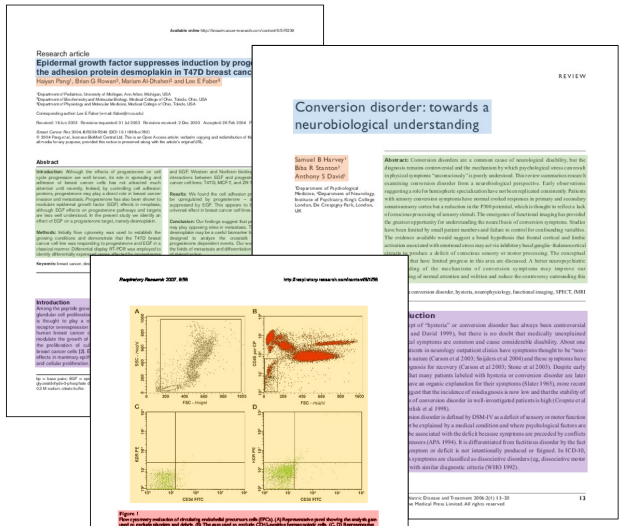
Results of standard PDF-to-text tool

bioRxiv preprint first posted online Feb. 5, 2019; doi: http://dx.doi.org/10.1101/537340. The copyright holder for this preprint (whichwas not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.All rights reserved. No reuse allowed without permission.5152Results53Trisomic strains show a commensal phenotype in an oropharyngeal infection model54During oropharyngeal infection in mice, a specific trisomy, Chr6x3, was significantly enriched55among strains and recovered from the majority of immunocompromised mice (Forche et al. 2018). The56frequency of Chr6x3 increased over the course of infection (Fig. 1A) with the allele combination ABB57occurring 2-fold more frequently than the AAB combination (Fig. 1B), suggesting that clones with trisomy58of Chr6 have a general fitness advantage during OPC and that an extra copy of allele B may be more59beneficial than an extra copy of allele A in this host niche. To test this hypothesis, we selected several60strains that, based on whole genome61karyotypes (produced using double digest62restriction-site associated DNA sequencing63(ddRADseq)), had acquired single trisomies64as the only change compared to the diploid65progenitor, strain YJB9318. Strains AF127566and AF1485 both had acquired Chr6x3, the67former68(Chr6ABB) and the latter with allele69combination AAB (Chr6AAB). Each strain70was71oropharynx72Importantly, these strains had not been73subjected to any selection regimes (e.g.,74GAL1 counterselection-induced). A third75strain, AF1773, that had acquired Chr5x376(Chr5AAB) and a small LOH on Chr1 (due77to selection for GAL1 LOH), served aswithalleleoriginallyofcombinationrecoveredthesamefromFig.1. Chr6 trisomy ABB is overrepresented in isolatesrecovered from mice with OPC. (A) The frequency of Chr6trisomy increases over the course of infection. (B) AmongChr6 trisomic strains, genotype ABB is the most frequentallele combination. For each genotype, symbol size isproportional to the frequency of isolation. Results are from theanalysis of C. albicans colonies from 3-5 mice per time pointas described in (Forche et al, 2018).ABBthemouse.5

- Since the start of the first infections, more than **70,000 research papers** have been published about COVID-19.
- Researchers around the world are looking for a new drug, treatments for infected patients, and a way to predict how the pandemic will develop.
- Having the latest information and insights at our fingertips can accelerate everyone's progress, point to new avenues, and avoid wasting time on areas that are not promising.
- However, no one can read so many papers, up to **several thousands of new articles per week**.

Searching the Literature for the Researchers

- BNL developed a **literature service** using **Artificial Intelligence** that:
  - Annotates the existing articles with standardized key words
  - Allows the scientists to find not only the relevant articles, but the part that is of most interest to them.
- We can find relevant text, images, and tables in the publications and show a prioritized list to the researchers.



Results of standard PDF-to-text tool

bioRxiv preprint first posted online Feb. 5, 2019; doi: http://dx.doi.org/10.1101/537340. The copyright holder for this preprint (whichwas not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.All rights reserved. No reuse allowed without permission.5152Results53Trisomic strains show a commensal phenotype in an oropharyngeal infection model54During oropharyngeal infection in mice, a specific trisomy, Chr6x3, was significantly enriched55among strains and recovered from the majority of immunocompromised mice (Forche et al. 2018). The56frequency of Chr6x3 increased over the course of infection (Fig. 1A) with the allele combination ABB57occurring 2-fold more frequently than the AAB combination (Fig. 1B), suggesting that clones with trisomy58of Chr6 have a general fitness advantage during OPC and that an extra copy of allele B may be more59beneficial than an extra copy of allele A in this host niche. To test this hypothesis, we selected several60strains that, based on whole genome61karyotypes (produced using double digest62restriction-site associated DNA sequencing63(ddRADseq)), had acquired single trisomies64as the only change compared to the diploid65progenitor, strain YJB9318. Strains AF127566and AF1485 both had acquired Chr6x3, the67former68(Chr6ABB) and the latter with allele69combination AAB (Chr6AAB). Each strain70was71oropharynx72Importantly, these strains had not been73subjected to any selection regimes (e.g.,74GAL1 counterselection-induced). A third75strain, AF1773, that had acquired Chr5x376(Chr5AAB) and a small LOH on Chr1 (due77to selection for GAL1 LOH), served **aswithalleleoriginallyofcombinationrecoveredthesamefromFig.1**. Chr6 trisomy ABB is overrepresented in **isolatesrecovered** from mice with OPC. (A) The frequency of **Chr6trisomy** increases over the course of infection. (B) **AmongChr6** trisomic strains, genotype ABB is the most **frequentallele** combination. For each genotype, symbol size **isproportional** to the frequency of isolation. Results are from **theanalysis** of C. albicans colonies from 3-5 mice per time **pointas** described in (Forche et al, 2018).**ABBthemouse.5**

Problems
Spurious text
Conjoined text
Out-of-order text

Search:

protease

searchKey.

Max number of articles:

10

numberOfArticles.

Including source:

PubChem (Related chemicals) PubMed (Related articles on PubMed) Annotating the abstracts

Selected: []

Search

Clear

Results

Related Articles

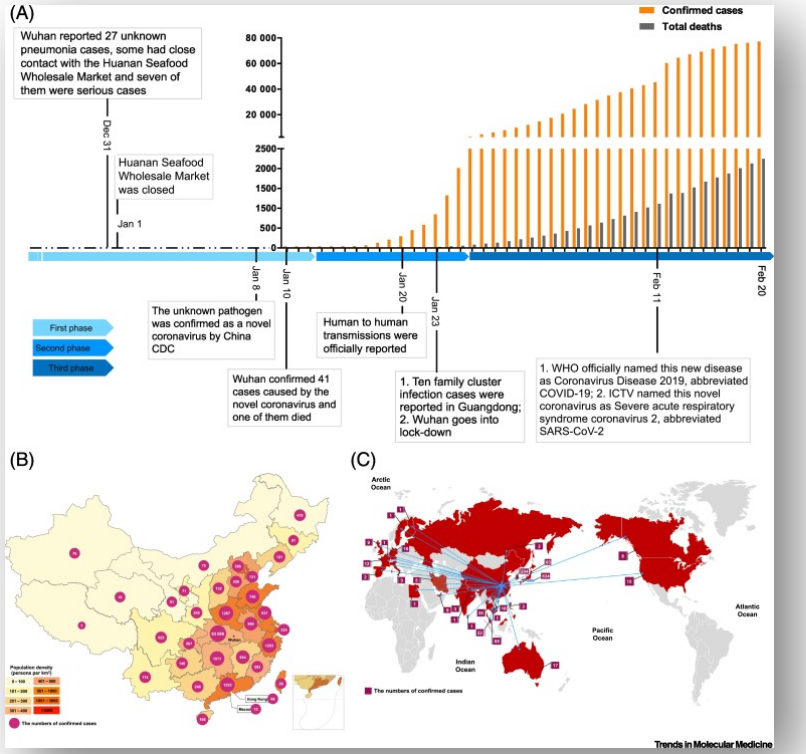
Publisher	Journal	Title	Date	Authors	Abstract	DOI
PMC	Class 3.4-6 Hydrolases,	SARS coronavirus	2013	Schomburg, Dietmar;	EC number 3.4.22.69 Recommended name SARS coronavirus main proteinase Synonyms 3C-like protease <2,3> [9,16,38,49,51] 3CL protease <2> [14,48]	10.1007/978-3-642-36260-6_3

Epidemiology

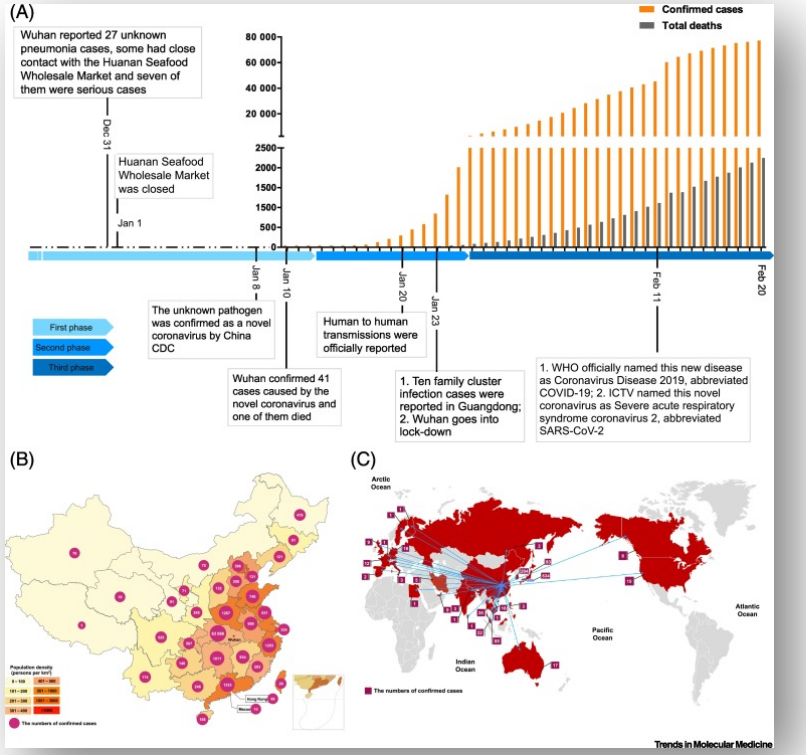
BNL CSI

Epidemiology

- Epidemiology models aim to predict how the virus will affect the population. For example:
  - How many will get infected?
  - How many will need hospital treatment?
  - How many will die?
- These questions need to be answered in the context of the circumstances in a particular area. For example:
  - What protective measures are in place and when?
  - Are people adhering to restrictions?
  - How many people travel through the region?
  - What is the general health of the population?



Epidemiology



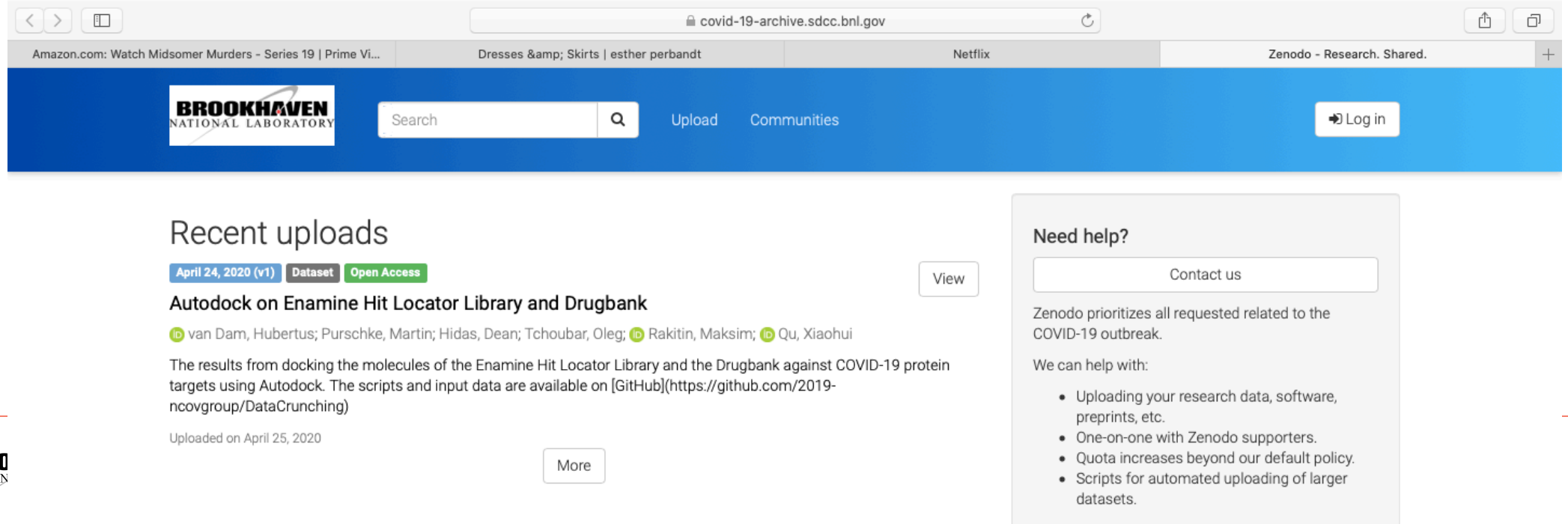
- Running these complex models is very time consuming, and each model as to be run many, many times to account for different scenarios and changes as time goes on.
- Using Artificial Intelligence, BNL and its partners are working on creating faster models that would allow us to study more scenarios quicker. This will help us to gain more confidence in the results.
- In addition, we have developed an AI method that will help us to determine which scenarios we should run to gain the most insights

COVID-19 Archive

BNL CSI and SDCC

COVID-19 Data Archive

- When COVID-19 hit the world, it was not the first time a pandemic had struck and been investigated by scientists - SARS, MERS went before
- Building on such knowledge can tremendously accelerate work on a new virus - data & tools
- However, the results of prior research where distributed across the world in many, often inaccessible places. It took scientists months to unearth key information.
- BNL took the stance that this should not happen again at least for US DOE research and established a central archive for results



Thank you to all of the staff who made this possible:
Shantenu Jha, Shinjae Yoo, Frank Alexander, Hubertus van Dam,
Sam Chen, Carlos Soto, Gilchan Park, Ai Kagawa, Ray Ren, Sean McCorkle,
Li Tan, Barbara Chapman, Vivek Kale, Uma Ganapathy, Carlos Gamboa

Carlos Simmerling and students (SBU)

BNL volunteers for docking studies from NSLS II, RHIC, and CFN