Small molecule antiviral discovery for SARS-CoV-2

Jonathan Allen, Ph.D.
Informatics Scientist
allen99@lnl.gov

LLNL-PRES-820754
This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Lawrence Livermore National Laboratory

DISCLAIMER

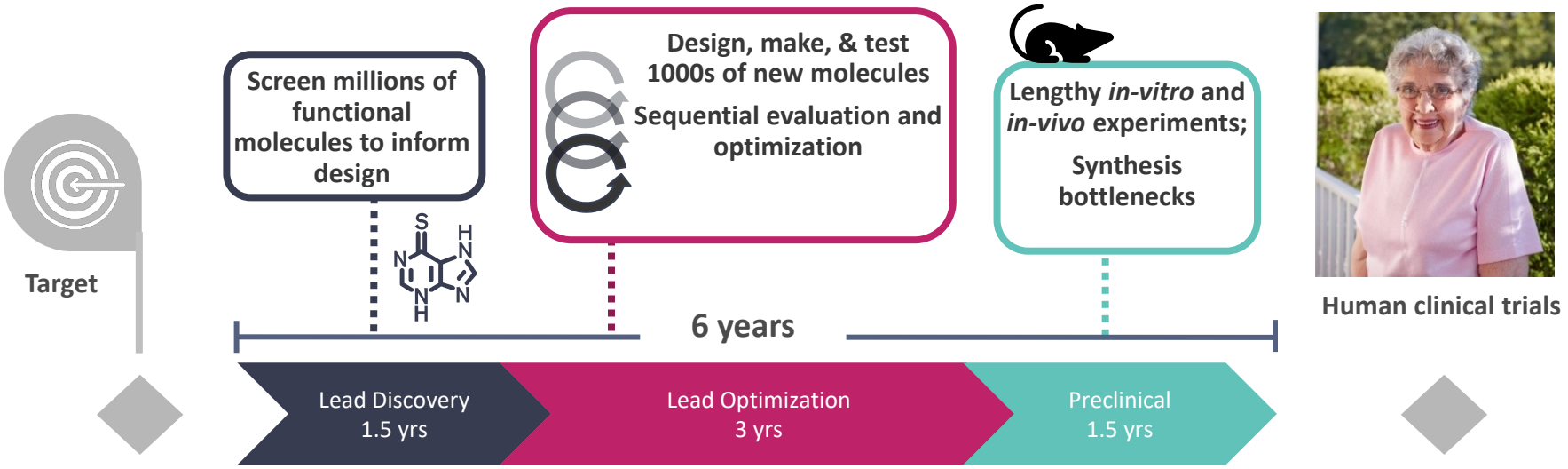
LLNL-PRES-820754

Rapid response to a disease outbreak is a national security priority

- Pathogen outbreaks cause massive disruptions to society and pandemics pose an existential threat to national security
  - Bacterial attack (Ba) 2001;
  - SARS 2003 : Cost $40 billion ; Mortality : 813 ; Fatality rate: 9.6%
  - “Swine flue” (H1N1) 2009;
  - Ebola 2014 : Cost $53 billion ; Mortality : 11,325 ; Fatality rate: 40%
  - Zika 2016 : Cost $20 billion ; Mortality: ~245 (1/2018) Fatality rate: 8.3-10.5%
  - SARS-CoV-2: 500K+ deaths and growing (3/2021)
- We can reduce cost and save lives for milder outbreaks while increasing preparedness for a deadly pandemic
- Immediate actions needed to respond to a new pathogen:
  1. Detection and diagnostics to initiate public health response
  2. Identify therapeutic targets
     - existing treatments
     - propose novel treatments
     - vaccine development

Current drug discovery: long, costly, high failure

Is there a better way to get medicines to patients?



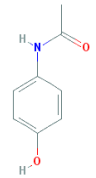
- **33% of total cost of medicine development**
- **Clinical success only ~12%, indicating poor translation in patients**

Source: http://www.nature.com/nrd/journal/v9/n3/pdf/nrd3078.pdf

ATOM Consortium enabling novel therapeutics from validated targets

Target identification for antiviral drug development is a challenging problem

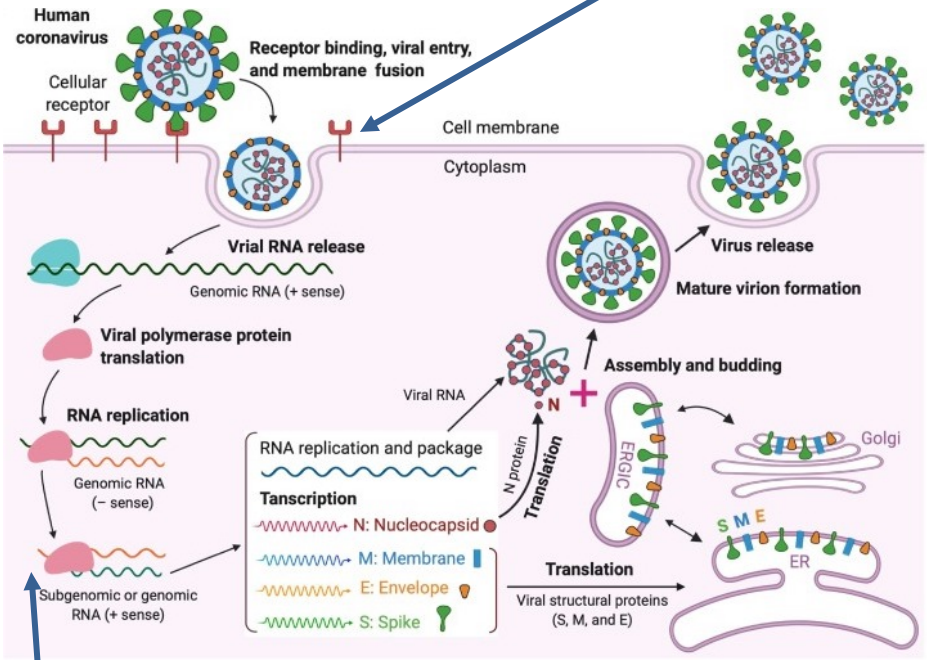
Drug types



Small molecule

Peptide

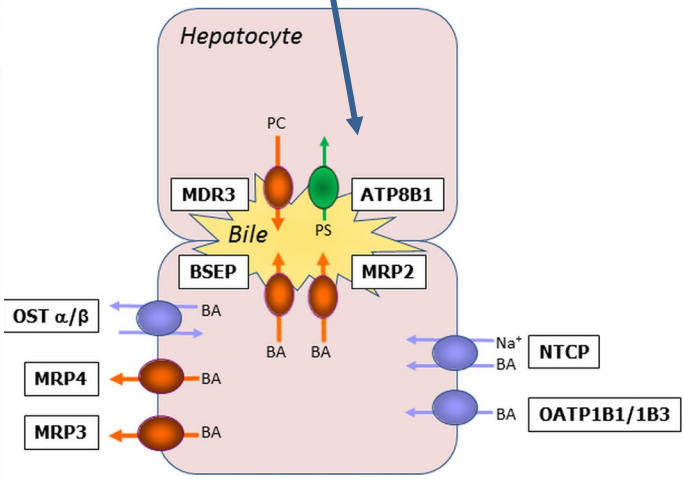
Human proteins to target



Jiang et al., 2020

Viral proteins to target

Human proteins to avoid

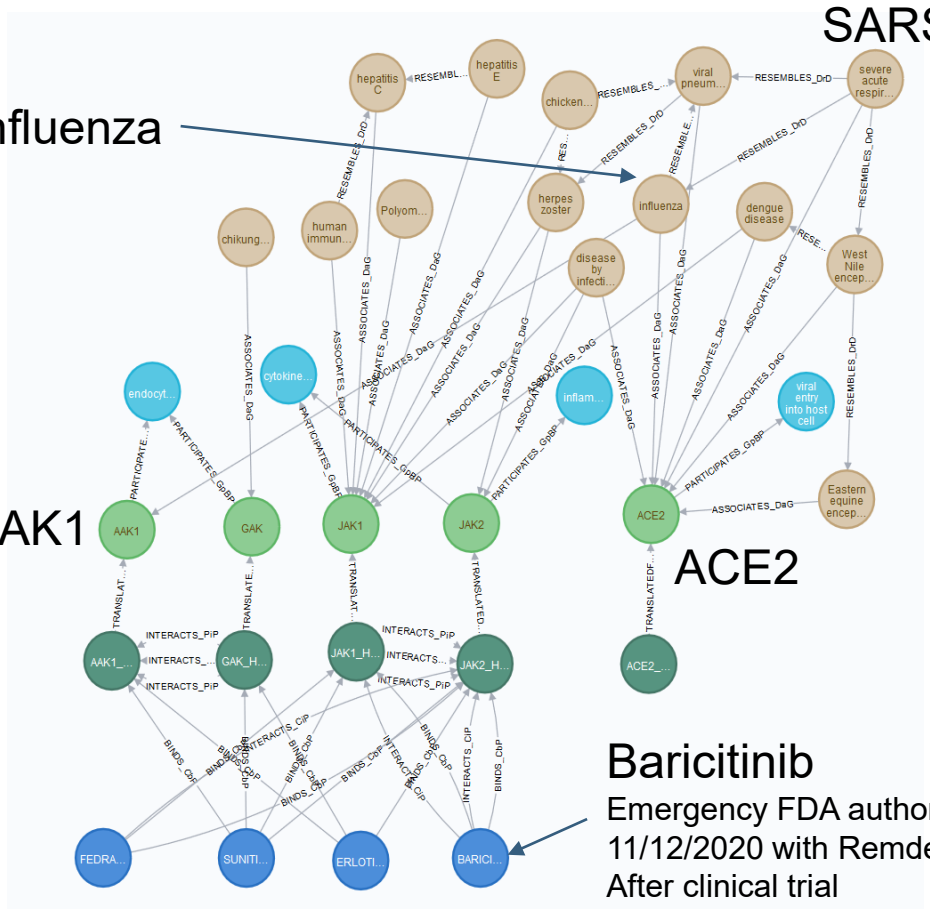


Kenna et al., 2018

1. Main protease (NSP5)
2. RdRp (NSP12)
3. Papain-like protease (NSP3)
4. Spike
5. Orf3a
6. TMPRSS (human)

Goal: design new potent viral inhibitors with no adverse health reactions

Biological knowledge networks identified drug repurposing candidates early in SARS-CoV-2 outbreak



Two targets stand out in graph:
ACE2 ← SARS
AAK1 ← Influenza ← SARS

- clathrin-dependent endocytosis

Potential COVID-19 therapies and associated targets

Beige=infectious disease type
Blue=biological process
Light green=protein
Dark green=gene
Blue=drug

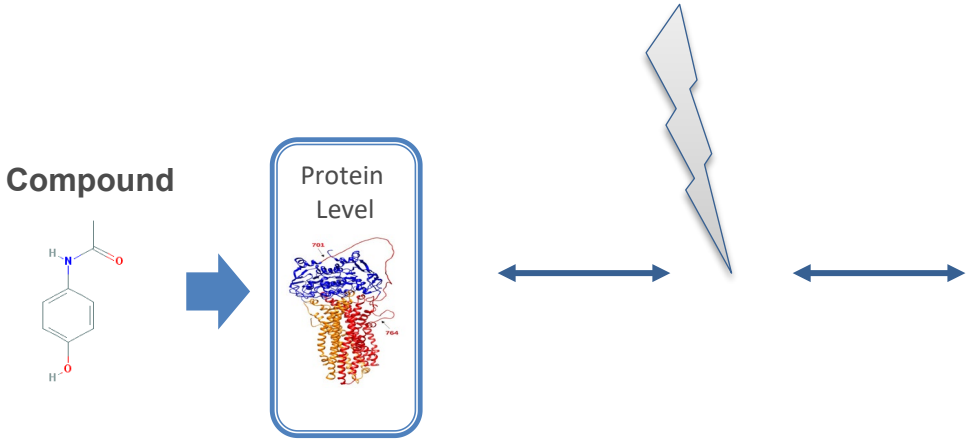
Recreation of BenevolentAI report from Feb. 4 2020
Using UCSF SPOKE graph

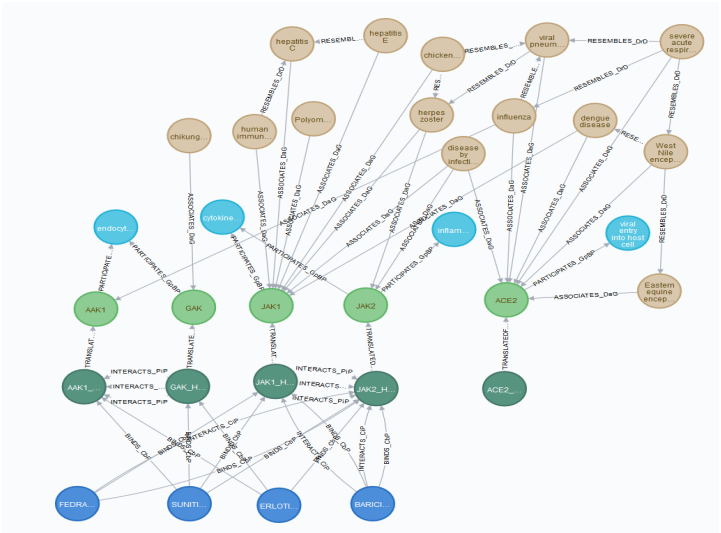
AAK1 and ACE2 are identified as potential targets

Knowledge networks can be expanded to include molecular interaction predictions

Gap between
Lower level molecular
Interactions and
Higher order functions

Knowledge networks

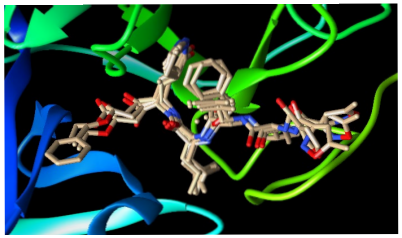




For a new pathogen: knowledge network must consider new compounds and new targets not in the database

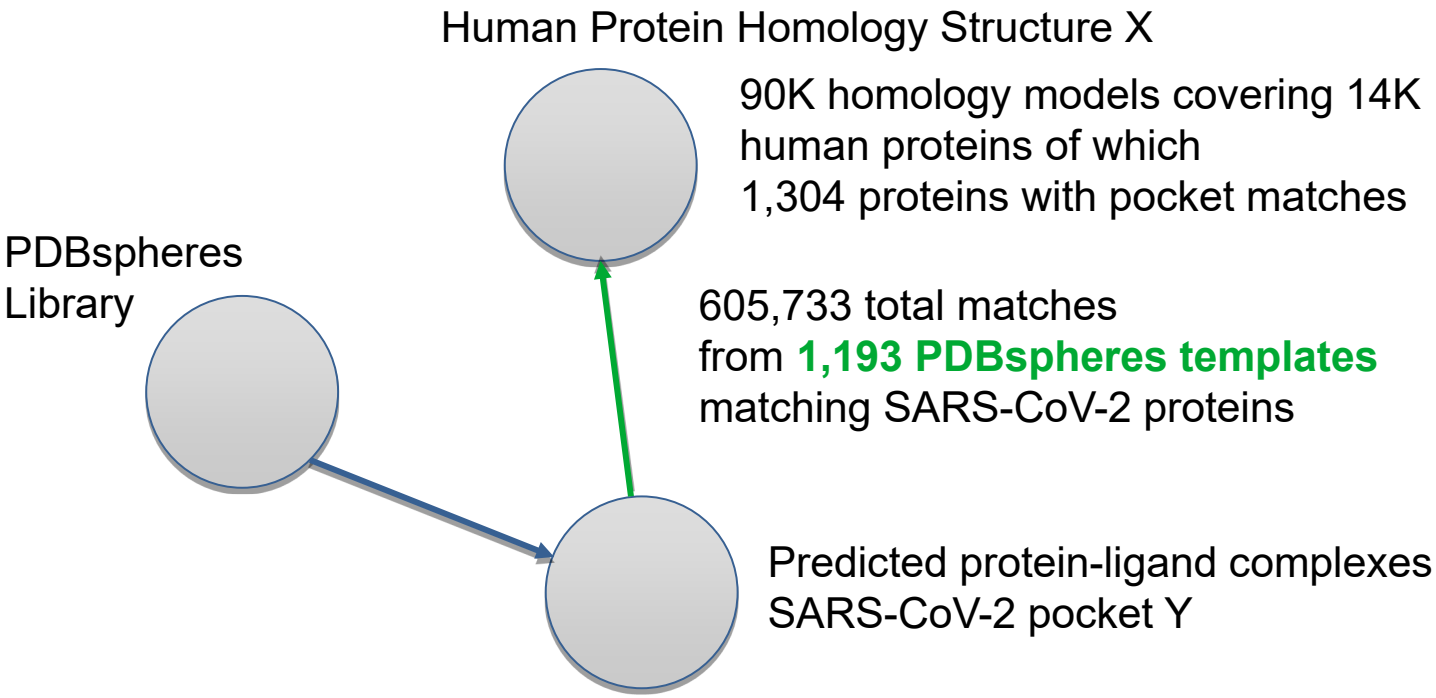
PDBspheres method description

- **The PDBsheres method has three main components:**



  - **Constructed PDBspheres library of binding site templates**
    - currently it contains 1,838,709 compound binding site models
    - Currently it contains 63,204 peptide binding site models
  - **Search system to detect pockets in proteins**
    - LGA program is used to perform all structure similarity searches
    - set of pocket candidates to test can be exhaustive
    - (i.e. all 1.9 M pockets from the library)
    - or, it can be preselected based on specific targeted ligands
    - or, it can be preselected based on sequence similarity between
    - query protein and protein-pockets from the PDBspheres library
  - **Metrics to assess confidence in detected pocket**
    - LGA/GDT (Global Distance Test) metric is used to assess similarities on Calpha level
    - GDC (Global Distance Calculations) allows evaluation on ALL atoms (including side-chains) level

PDBspheres search procedure to detect nCoV - human pocket matches



List of 12 nCoV proteins with similar pockets identified in 1,304 human proteins:

nCoV_nsp3
nCoV_nsp5
nCoV_nsp7
nCoV_nsp8
nCoV_nsp12
nCoV_nsp13
nCoV_nsp14
nCoV_nsp15
nCoV_nsp16
nCoV_Spike
nCoV_E
nCoV_ORF3a

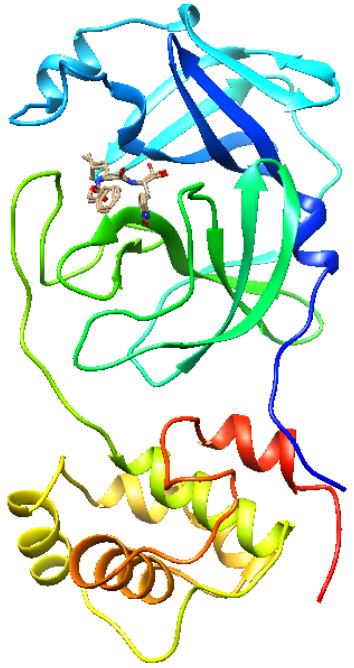
(medium confidence;15-20-60)

Modeling the virus expands drug development targets and may be needed to treat new viruses

- 19 of 25 SARS-CoV-2 proteins structurally modeled with confidence

Models publicly released:
https://covid19drugscreen.llnl.gov/homology_models

- PDBSpheres: software for automated pocket identification templates for ~14K human proteins and SARS-CoV-2
  - Human pocket matching – 2.1M total template matches
  - SARS-Cov-2 pocket matching – 908 ligands matched to 6,961 templates from 2,681 library templates
- Future work: explicit binding affinity assignment



Two protease inhibitors verified in literature/crystal structure for main protease

General principles for target ranking

- Evidence from previous drug development efforts focusing on target
- Information on chemotypes and interaction mechanisms to jump start drug design
- Understanding of mechanism of action
- Limited off-target interaction with disease and tissue relevant proteins

Ranking of viral proteins for off-target interactions

- Definitions:

Set of viral proteins: $V = \{v_1, v_2, v_3, \dots, v_{13}\}$

Set of human proteins: H $= \{h_1, h_2, h_3, \dots, h_{4162}\}$

Set of all ligands: L $= \{l_1, l_2, l_3, \dots, l_{908}\}$

Set of ligands that bind to viral protein v_i: $\mathrm{O_i} = \{\text{all } l_k \text{ that bind to } v_i\}$

Set of ligands that bind to human protein $\mathrm{h_i}$: $T_i = \{\text{all } l_k \text{ that bind to } h_i\}$

q := index for interactions with viral protein

r := index for interactions with human protein

$N_{v_i l_k}$:= number of interactions between viral protein v_i and ligand l_k

$N_{h_j l_k}$:= number of interactions between human protein h_j and ligand l_k

Indicator function: $\mathbf{1}_{O_i T_j}(l_k) = \begin{cases} 1 & \text{if } l_k \in O_i \wedge l_k \in T_j \\ 0 & \text{otherwise} \end{cases}$

CLAIM: Lower score corresponds to less evidence of off-target interactions

- Overlap score:

$$\text{numerator}_{v_i} = \sum_{j=1}^{4162} \sum_{k=1}^{908} \left(\frac{\sum_{q=1}^{N_{v_i l_k}} GDC_{v_i l_k}^{(q)}}{N_{v_i l_k}} + \frac{\sum_{r=1}^{N_{h_j l_k}} GDC_{h_j l_k}^{(r)}}{N_{h_j l_k}} \right) \mathbf{1}_{O_i T_j}(l_k)$$

Ranking of viral proteins for off-target interactions

The strength of off-target interactions from each ligand that binds to the viral protein

Viral protein	Off-target score
nsp9	0.23
E	0.84
nsp15	0.87
nsp14	1.62
S	4.30
nsp16	4.77
nsp3	5.59
nsp13	5.70
nsp7	6.09
nsp8	8.44
nsp5	8.54
ORF3a	14.27
nsp12	56.43

The strength of off-target interactions from each residue in the viral protein

Viral protein	Off-target score
nsp9	1.07
nsp15	1.18
nsp14	1.81
S	3.77
nsp13	5.71
E	7.37
nsp16	7.41
nsp3	15.97
nsp5	26.91
ORF3a	56.85
nsp8	58.92
nsp7	68.32
nsp12	152.68

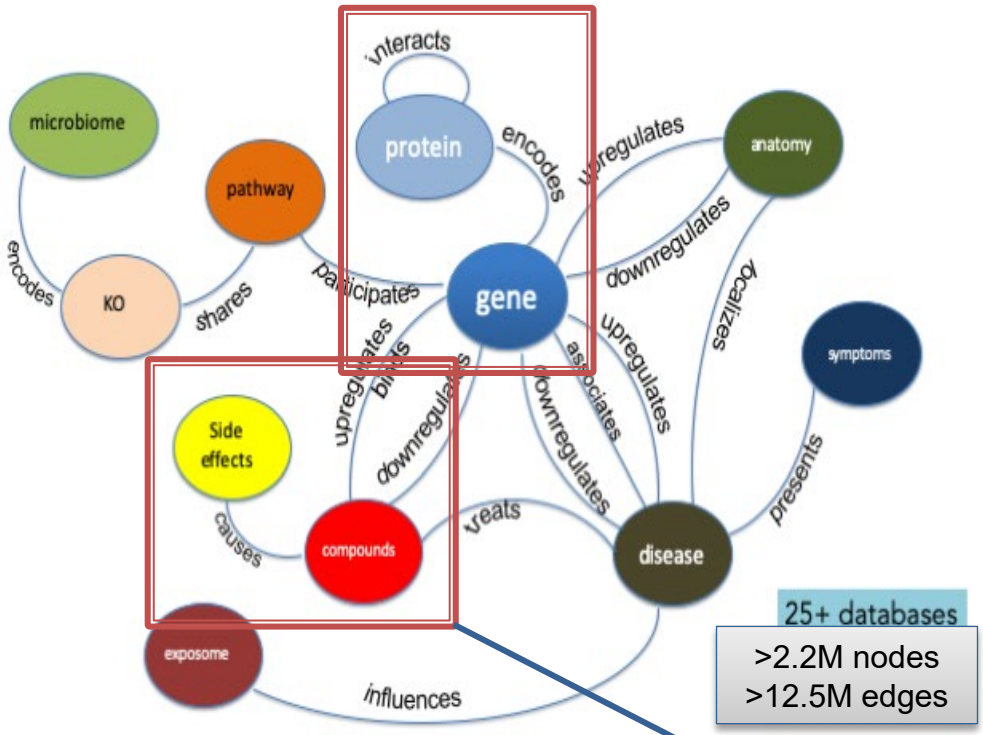
Spike (S)
Papain-like protease (NSP3)
Main protease (NSP5)
Orf3a
RdRp (NSP12)

Spearman rank correlation coefficient = 0.868

Integration of computational modeling data with a biological knowledge graph improves drug target identification

Ligand-protein pocket similarity between SARS-CoV-2 spike protein and human proteins

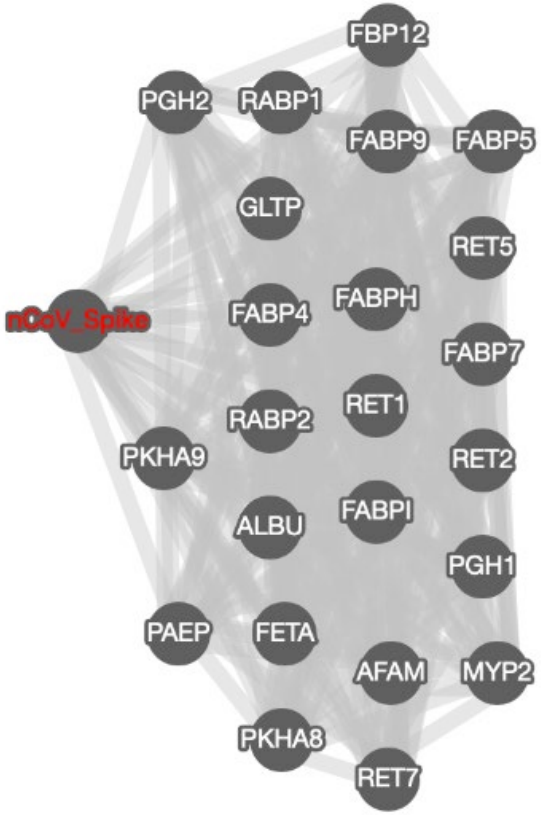
Viral protein	Off-target score
nsp9	0.23
E	0.84
nsp15	0.87
nsp14	1.62
S	4.30
nsp16	4.77
nsp3	5.59
nsp13	5.70
nsp7	6.09
nsp8	8.44
nsp5	8.54
ORF3a	14.27
nsp12	56.43



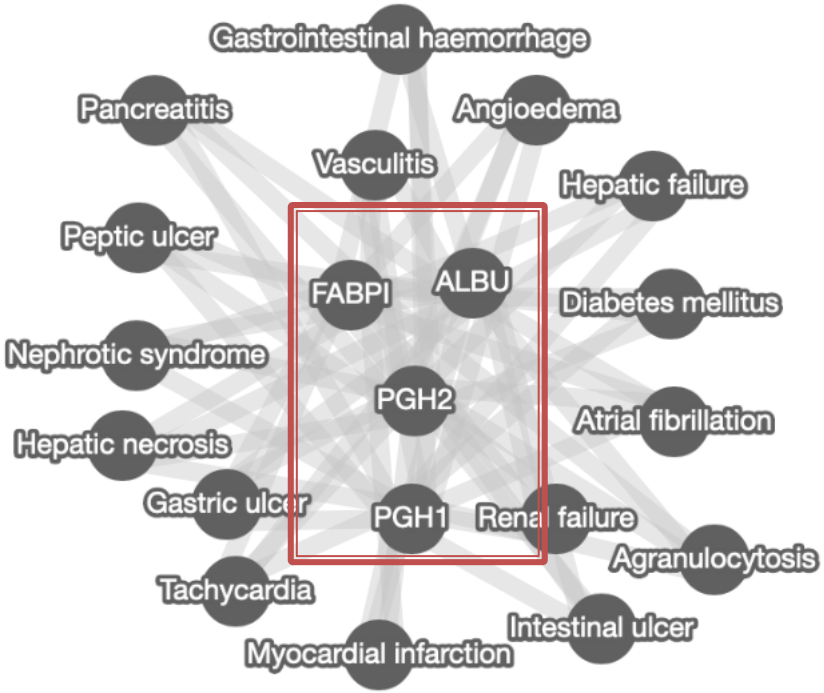
$$\sum_{j=1}^{4162}\sum_{k=1}^{908}\left(\frac{\sum_{q=1}^{N_{v_i l_k}} GDC_{v_i l_k}^{(q)}}{N_{v_i l_k}} + \frac{\sum_{r=1}^{N_{h_j l_k}} GDC_{h_j l_k}^{(r)}}{N_{h_j l_k}}\right) \text{weight } \mathbf{1}_{O_i T_j}(l_k)$$

Protein targets linked to safety concerns can be used to prioritize target list

Ligand-protein pocket similarity network



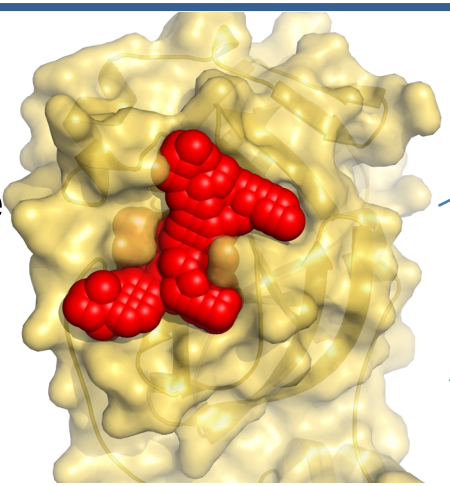
Adverse drug reactions for related human proteins

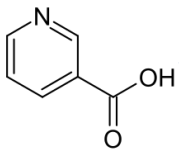


Four proteins associated with compounds with side effects

Using the main protease as a SARS-CoV-2 antiviral target

SARS-CoV-2
Main protease



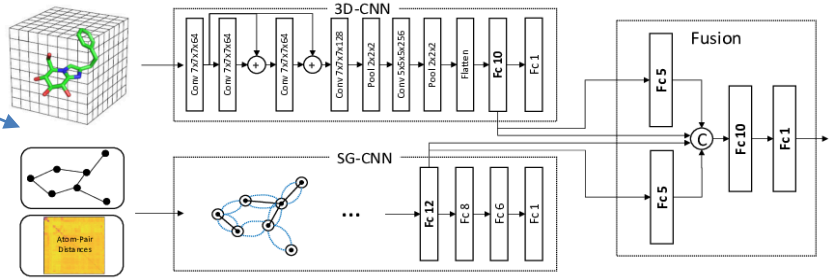


Traditional docking

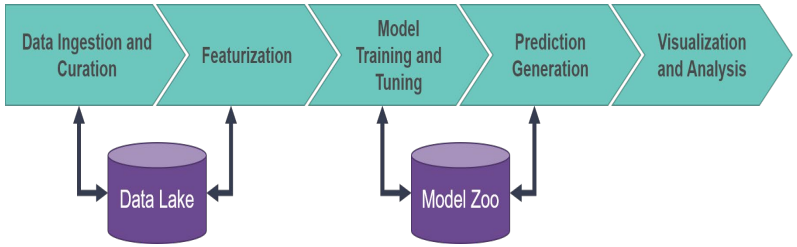
Physics based re-scoring

https://github.com/XiaohuaZhangLLNL/conveyorlc

Atom and 3D structure-based deep learning

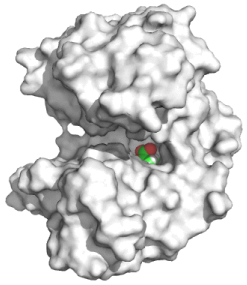


Target specific trained model pre-trained on structure model first



Calculated protein interactions with new molecules expand relationships in the knowledge graph

- Vina – speed=***moderate*** fast (1-2 minutes)
- MM/GBSA – speed=moderate (62 minutes)
- Implicit solvent MD = slower (7.2 hrs/GPU)
- Explicit solvent MD = slower (at least 7.2 hrs)



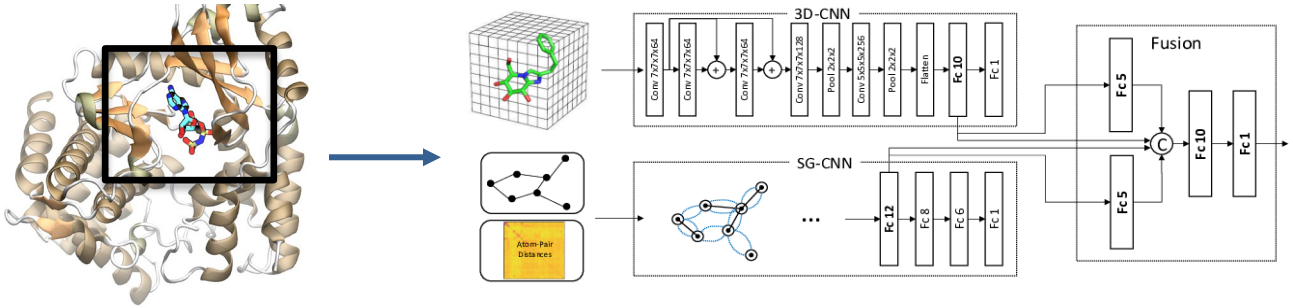
Physics based protein-ligand binding affinity does not scale to modeling billions of interactions

Two machine learning strategies currently employed

- Generate target specific scoring data using MM/GBSA
  - Use ML model to learn scoring function
    - Pros: Develop a faster scoring function that could match MM/GBSA accuracy
    - Cons: MM/GBSA scores still have limitations in accuracy

- Use 3D structure based spatial information to learn across multiple targets
  - Pros: Train on experimental binding data, readily applies to any new target (within reason-relative to training data)
  - Cons: Requires some 3D structure of the protein and a pocket

Fusion models for Atomic and molecular STructures (FAST)

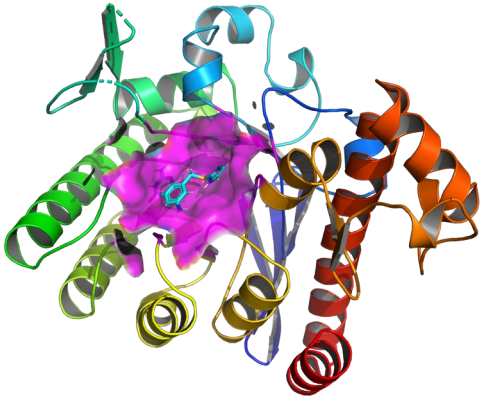


- 3D-CNNs have been used by numerous teams starting with AtomNet in 2015. (AtomWise)
- 3D Spatial Graphs were introduced with PotentialNet in 2018. (Genesis Therapeutics)
- No publications comparing the approaches directly
- Our results suggest potential benefits for combining two approaches

Open Source: https://github.com/llnl/fast

Jones, D., Kim, H et al., 2021 JCIM (accepted)

Extract atomic features that generalize across multiple targets

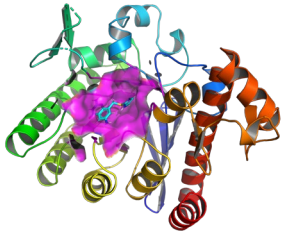


- Element type: one-hot encoding of B, C, N, O, P, S, Se, halogen or metal
- Atom hybridization (1, 2, or 3)
- Number of heavy atom bonds (i.e., heavy valence)
- Number of bonds with other heteroatoms
- Structural properties: bit vector (1 where present) encoding of hydrophobic, aromatic, acceptor, donor, ring
- Partial charge
- Molecule type to indicate protein atom versus ligand atom (-1 for protein, 1 for ligand)
- Van der Waals radius

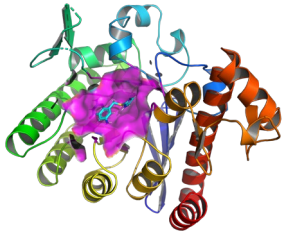
Combining representations improves prediction accuracy

Models trained on a dataset called 2016 version of PDBBind http://www.pdbbind.org.cn/

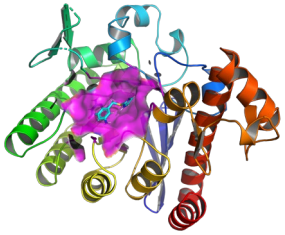
Ligand A
Protein A + Ki



Ligand B
Protein B + Ki

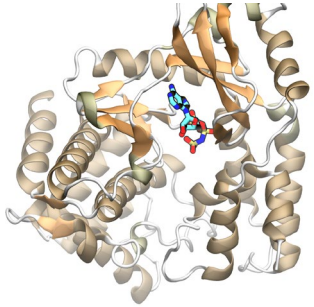


Ligand C
Protein C + Ki



…..

Created a special hold out set – structures taken from 2019 with a detailed analysis to find structurally novel pockets and novel ligands – 222 complexes.



Combining representations improves prediction accuracy

Traditional “test” set from 2016

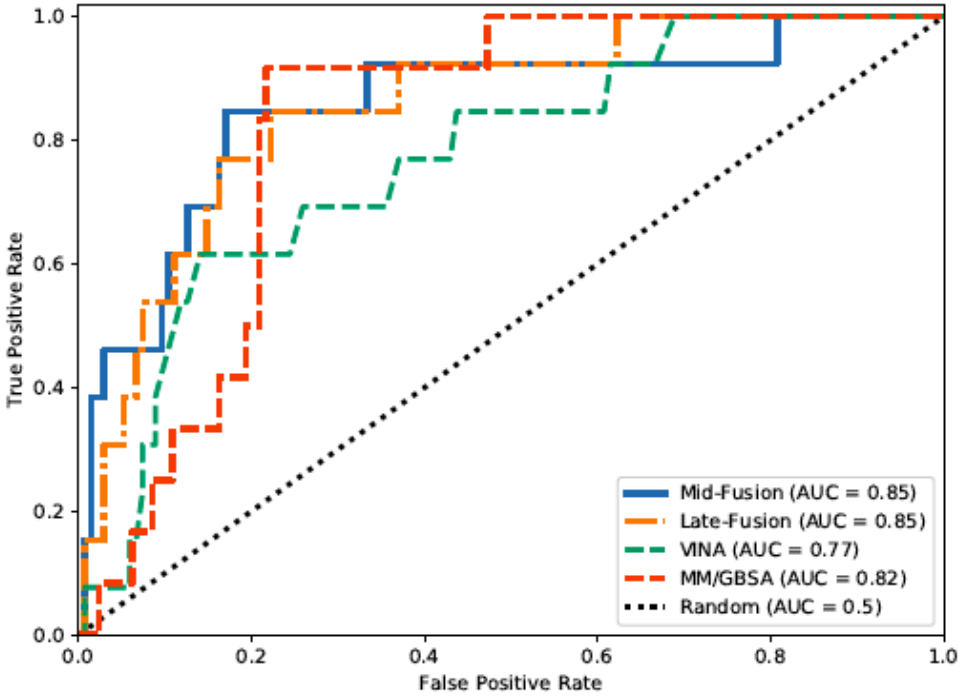
Model	r^2	*Pearson r*	*Spearman r*	*MAE*	*RMSE*
SG-CNN (R)	.424	.666	.647	1.321	1.650
SG-CNN (G)	.519	.747	.746	1.194	1.508
SG-CNN (R + G)	.600	.782	.766	1.084	1.375
3D-CNN (R)	.523	.723	.716	1.164	1.501
3D-CNN (G)	.420	.649	.658	1.294	1.655
3D-CNN (R + G)	.397	.677	.657	1.334	1.688
Late Fusion	.628	.808	.803	1.044	1.326
Mid-level Fusion	**.638**	.810	.807	**1.019**	1.308
Pafnucy[21]	-	.78	-	1.13	1.42
KDeep[12]	-	**.82**	**.82**	-	**1.27**
Fusion (Ligand only)	-0.916	.485	.492	2.495	3.008
Fusion (Pocket only)	-2.380	.501	.485	3.485	3.995

Method	*Pearson r*	*Spearman r*	*MAE*	*RMSE*
Vina	.599	.605	-	-
MM/GBSA	.647	.649	-	-
Mid-level Fusion	**.803**	**.797**	**1.035**	**1.327**

Our challenging hold out set from 2019

Model	*Pearson r*	*Spearman r*	*MAE*	*RMSE*
SG-CNN	.515	.511	1.152	1.450
3D-CNN	.427	.406	1.211	1.488
Late Fusion	.539	.525	**1.062**	**1.326**
Mid-level Fusion	**.545**	**.532**	1.074	1.338
KDeep[12]	.487	.478	1.135	1.424
Pafnucy[21]	.528	.528	1.106	1.381

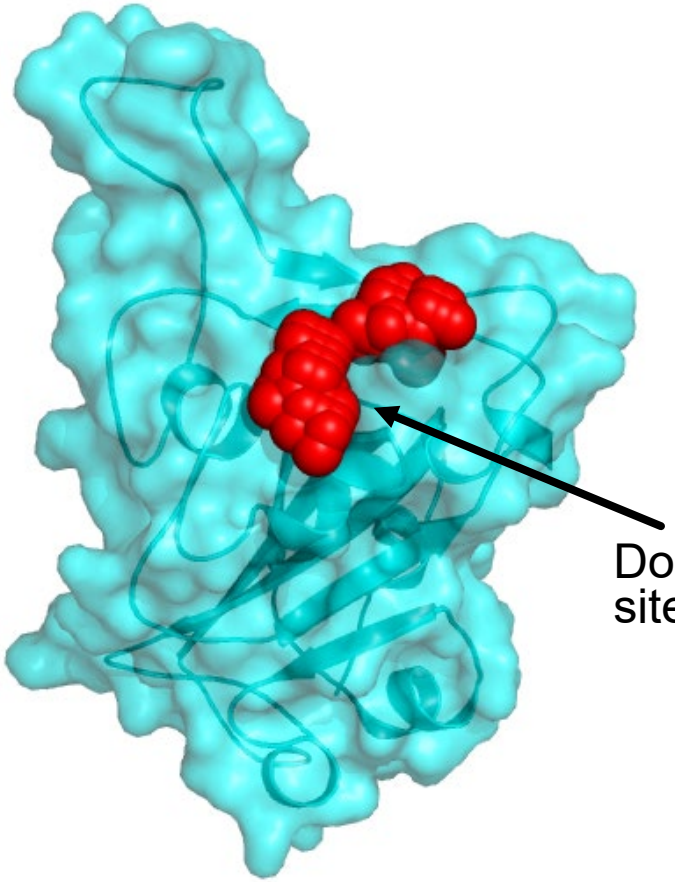
Combining representations improves prediction accuracy



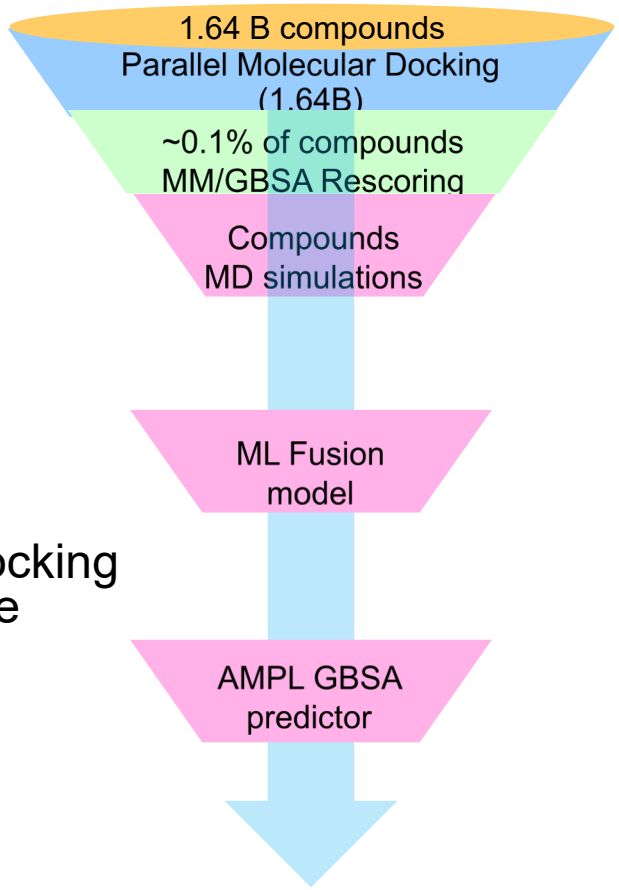
Preliminary observations indicate Fusion model provides a more scalable alternative or compliment to more expensive scoring functions

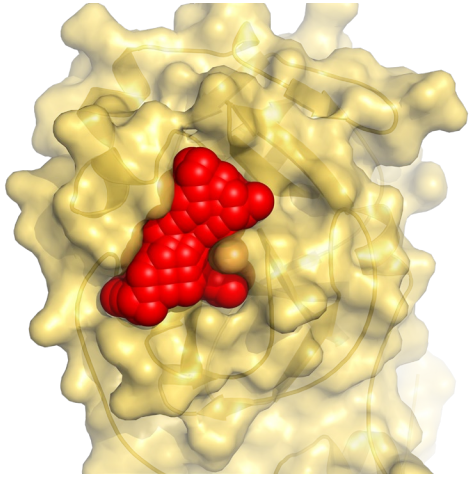
Fusion model scores 10,000 poses per second (with 6 compute nodes)

Identified 10 lead compounds as potential spike or Mpro antivirals with 5 compounds interfering with cell infection



Receptor binding domain of SARS-CoV-2 spike protein

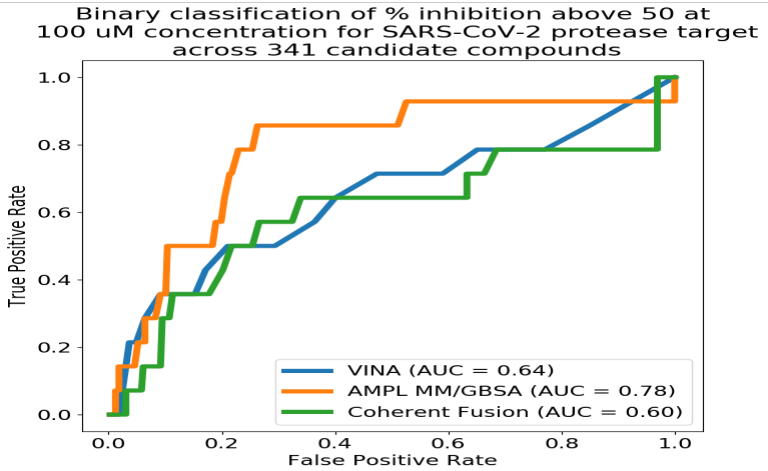


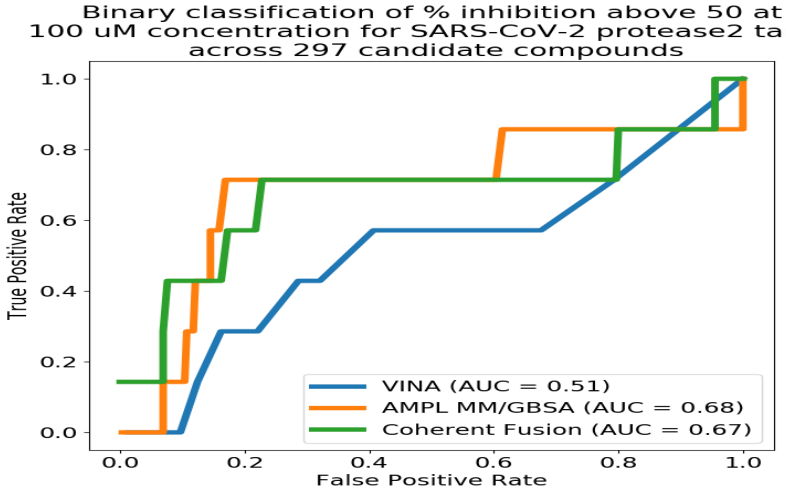


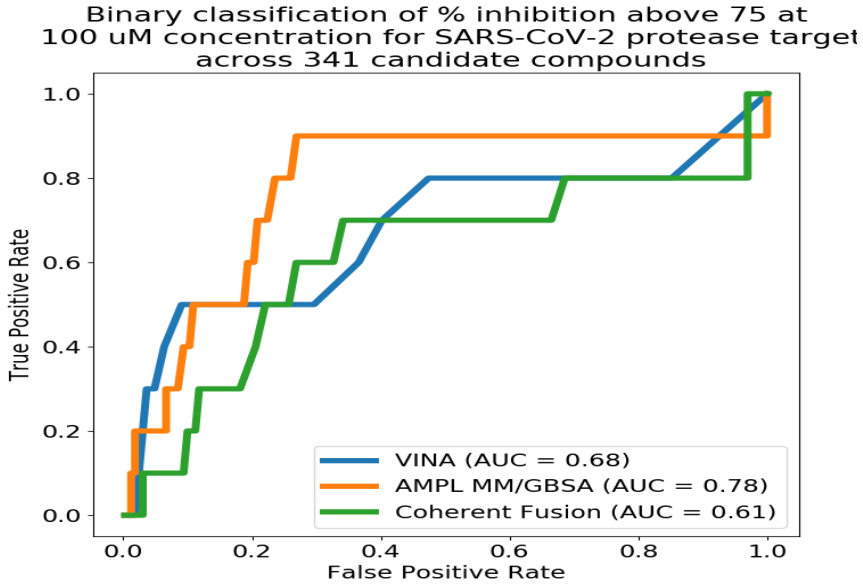
Portion of active site
In main protease (Mpro)

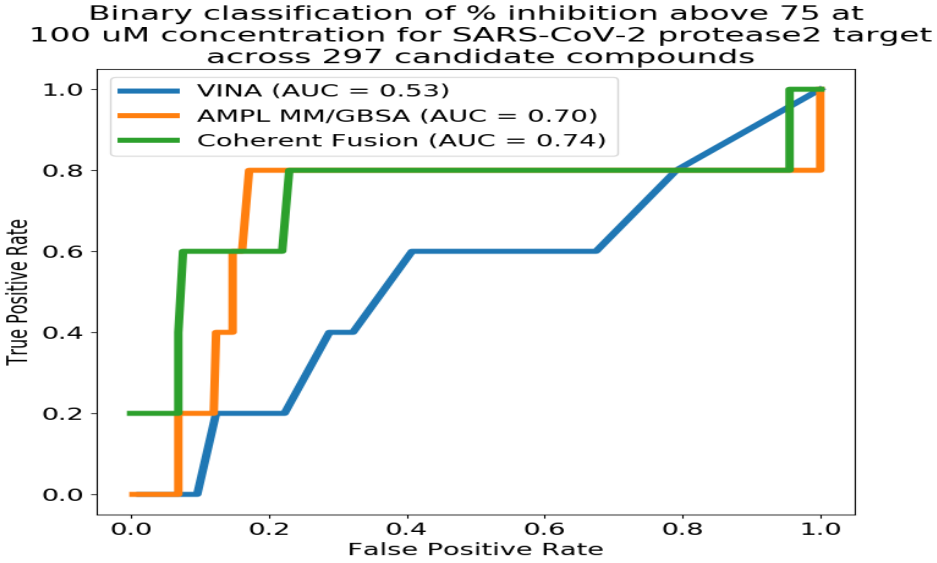
LDRD 20-ERD-065 "Identifying Potential Antiviral Small Molecules against COVID-19 "
DOE NVBL BER "Molecular design and analysis to inform therapeutics related to COVID-19"

Predictive value of different scoring methods compared against initial experimental screens

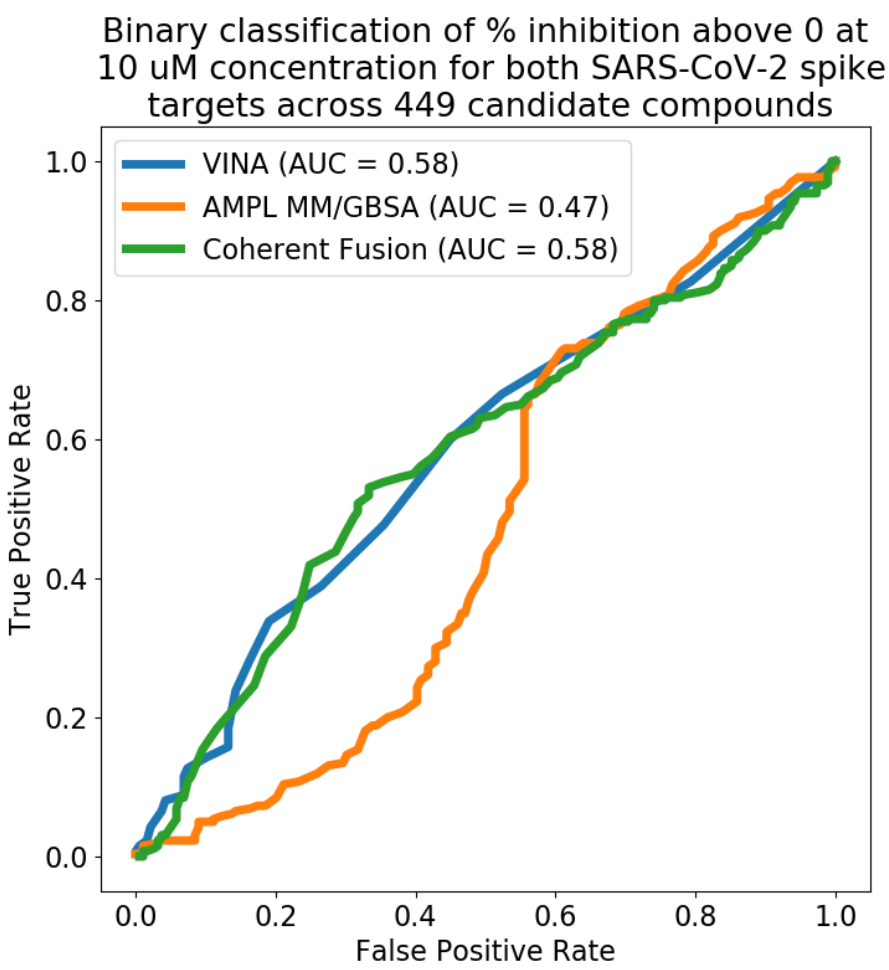


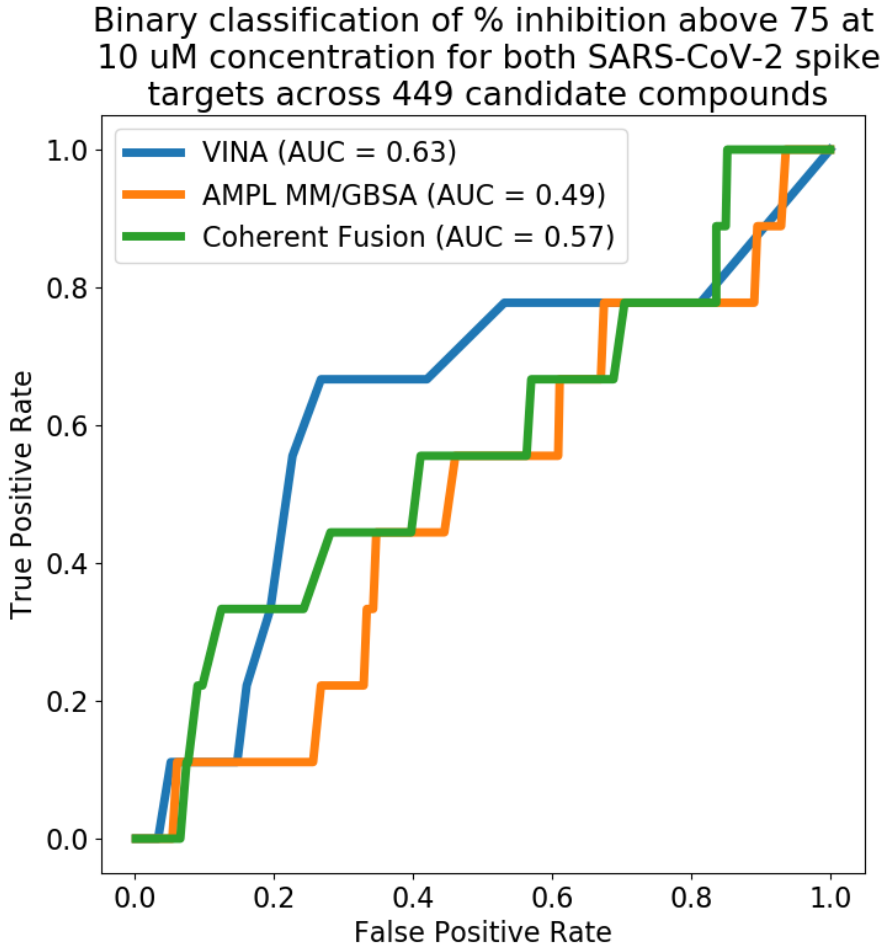






Predictive value of different scoring methods compared against initial experimental screens

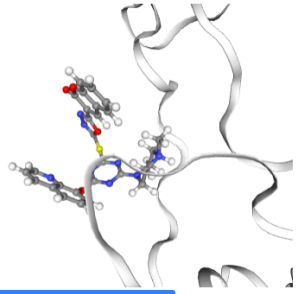




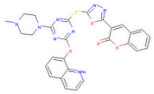
Search Calculation By Protein, Compound, Type, and Score Threshold using Web Server

Calculations: spike and CHEMBL2238342

Calculations Information



Protein: spike

Compound:  CHEMBL2238342

Download PDB File Shown

Download PDB Files For All Calculations

Top Compound Rank (Top 10 in Bold)

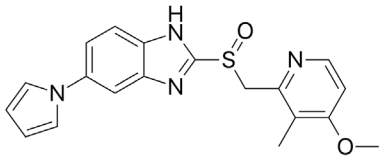
top_100_any_compound_3D-CNN	**4**	**top_100_group_3D-CNN**	**4**

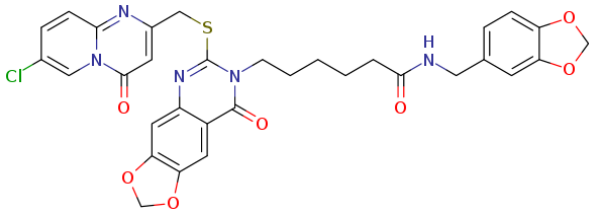
Calculations Poses & Scores

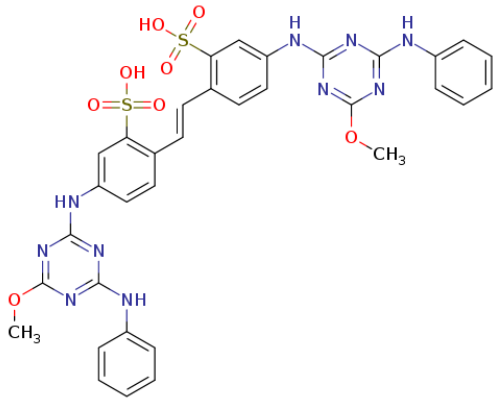
Calculation Type	Scores (Top Score in Bold)	
	Pose	**Score**
GBSA	1	**-44.97**

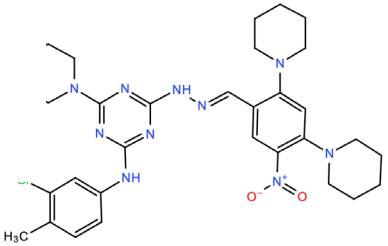
https://covid19drugscreen.llnl.gov/

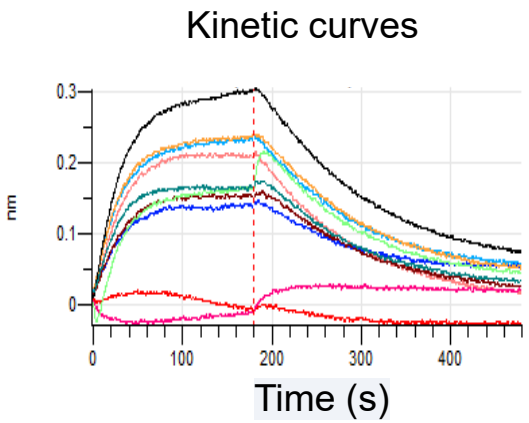
Identified 10 lead compounds as potential spike or Mpro antivirals with 5 compounds interfering with cell infection











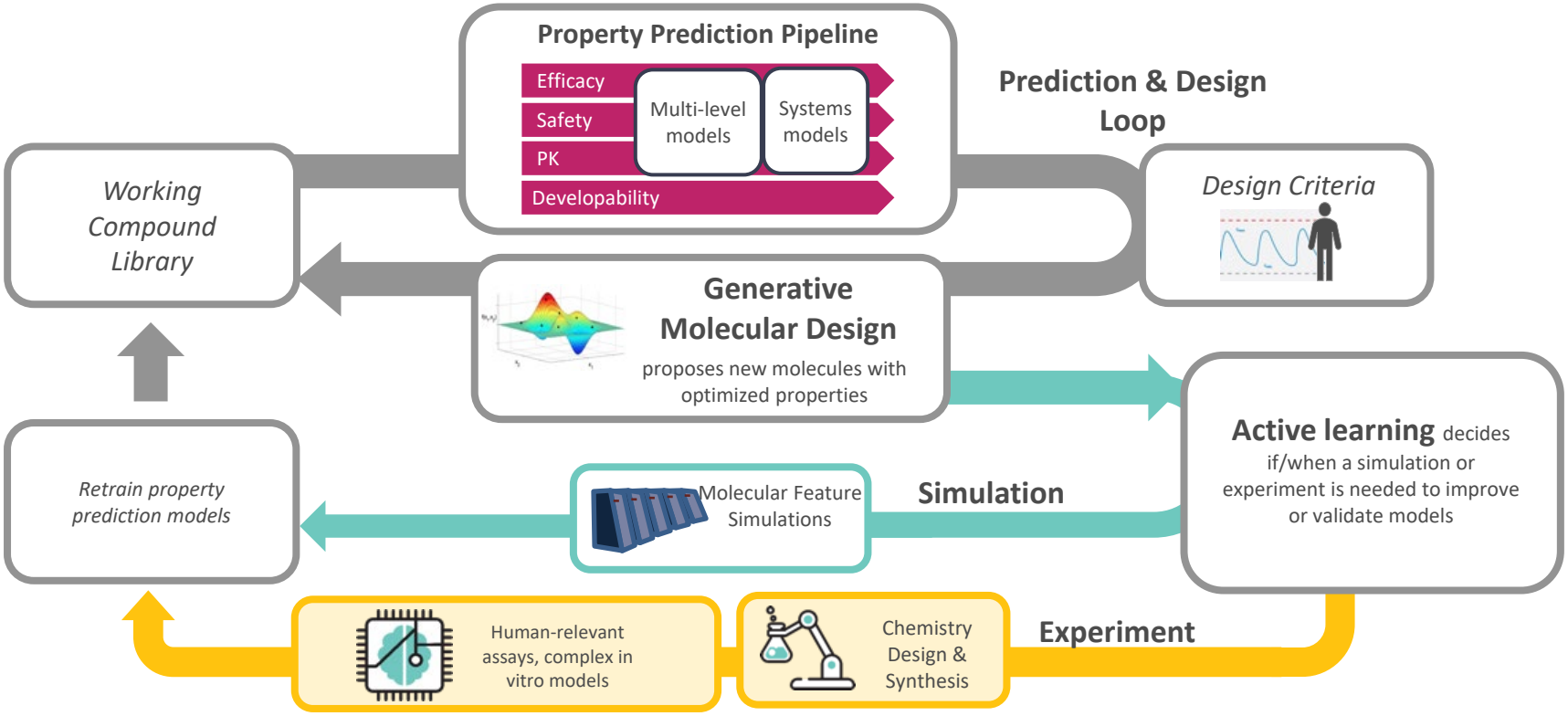
645 compounds screened for Mpro
575 compounds screened for spike
71 common compounds screened both

Five promising (4 main protease and 1 spike) compounds, as identified by *in vitro* experiments, have shown to strongly inhibit in the live virus assay.

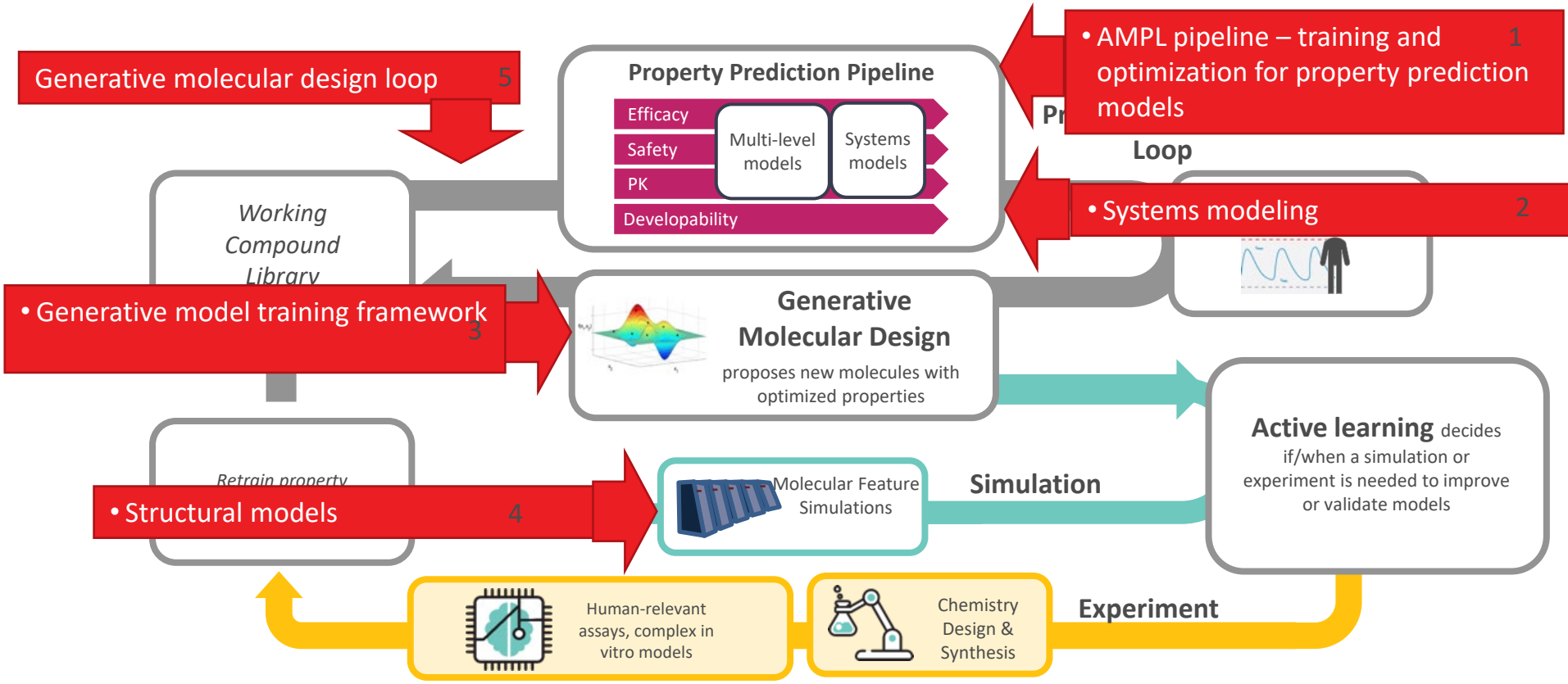
"Discovery of Small-molecule Inhibitors of SARS-CoV-2 Proteins Using a Computational and Experimental Pipeline". Lau et al., 2021 (In submission)

LDRD 20-ERD-065 "Identifying Potential Antiviral Small Molecules against COVID-19 "
DOE NVBL BER "Molecular design and analysis to inform therapeutics related to COVID-19"

The ATOM Platform
Active Learning Drug Discovery Framework



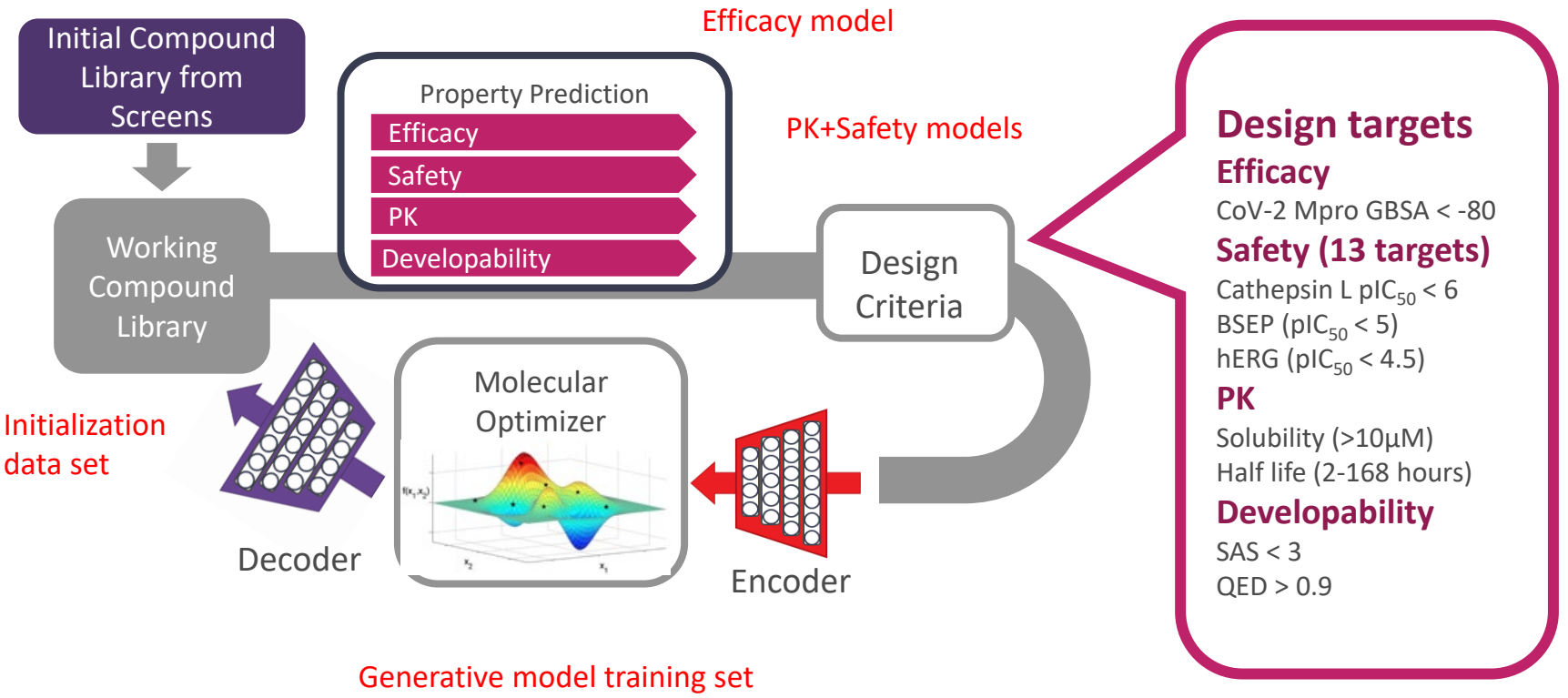
The ATOM Platform
Active Learning Drug Discovery Framework



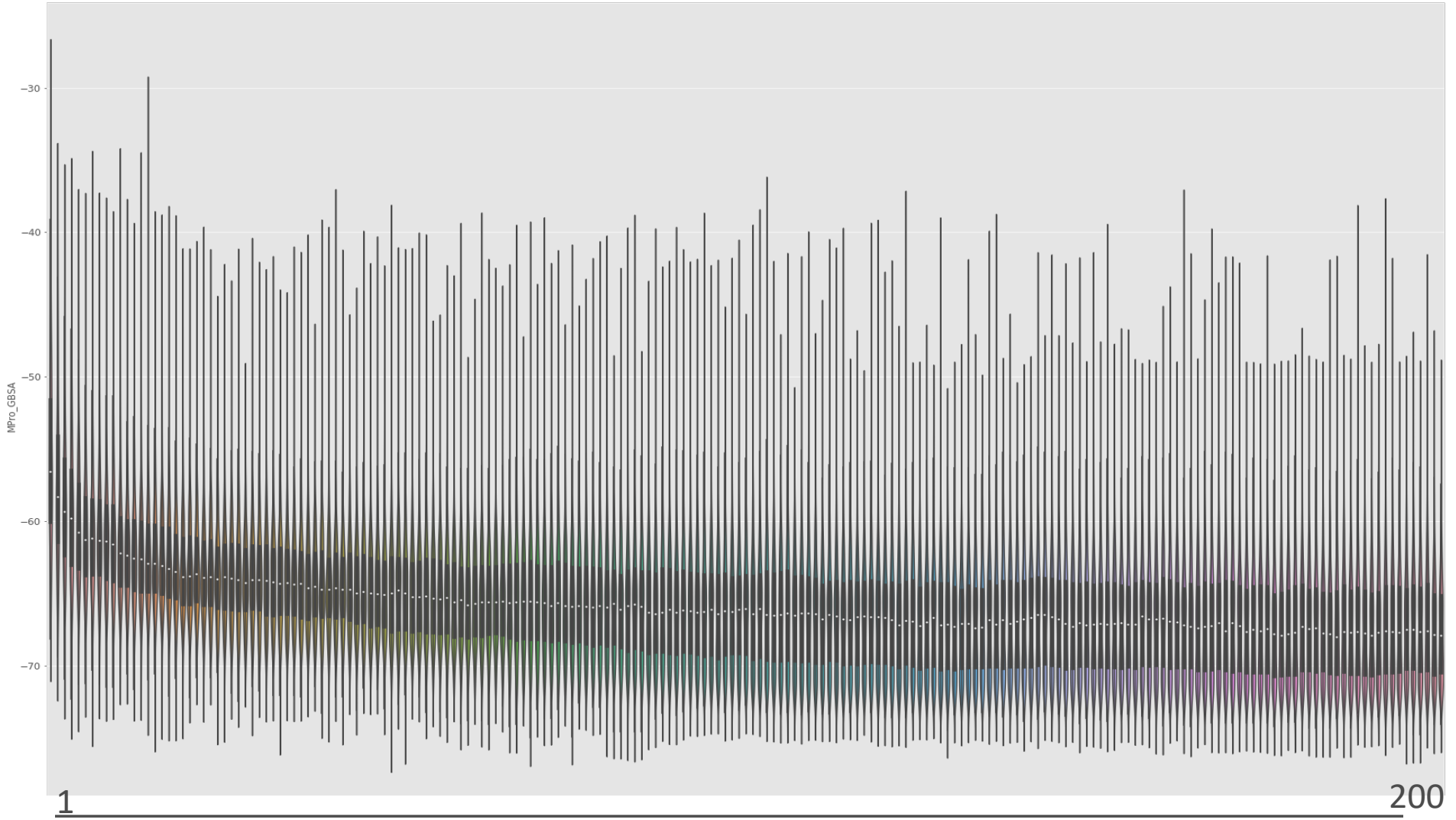
ATOM

Targeting GMD platform for opensource release after partner testing (currently Frederick National Lab, LBNL, Purdue)

Design of the SARS-CoV-2 main protease inhibitor



Optimizer finds new molecules with predicted increase in Mpro binding



ATOM

Next steps

- Begin compound synthesis and testing of new molecules
  - Measure activity inhibition of SARS-CoV-2 Mpro
  - Measure activity against additional off-target safety and PK parameters.
- Re-tune machine learning models using initial experimental feedback
- Improve prediction of model uncertainty to better inform active learning
- Improve training of generative model
- Increase scaling of design optimization loop

Acknowledgement

Target identification: Adam Zemla (Structural modeling), Jeffrey Drocco (Structural modeling and analysis), Sarah Sandholtz (Structural modeling and analysis), Marisa Torres (Graph database and analysis), Mary Silva (Graph database and analysis), Monica Borucki (Virology)Thanks to UCSF SPOKE Team (Sharat Israni, Sergio Baranzini)

Small molecule data science: Kevin McLoughlin, Stewart He, Garrett Stevenson, Hyojin Kim, Derek Jones, Marisa Torres, Aiden Epstein, Adam Zemla

Molecular modeling: Xiaohua Zhang, Brian Bennion, Dan Kirshner, Sergio Wong, Drew Bennett, Felice Lightstone

Experimental team: Oscar Negrete (Sandia), Sean Lund (Sandia) ; Brent Segalke, Feliza Bourguet, Jacky Lo, Deepa Murugesh, Ed Saada

Funding: LLNL-LDRD, DOE-NVBL, DOE-ATOM, DoD-DTRA

QUESTIONS?

Lawrence Livermore
National Laboratory