4.2 Billion compound “feature database” to support machine learning in drug discovery

Rick Stevens
Argonne National Laboratory
The University of Chicago

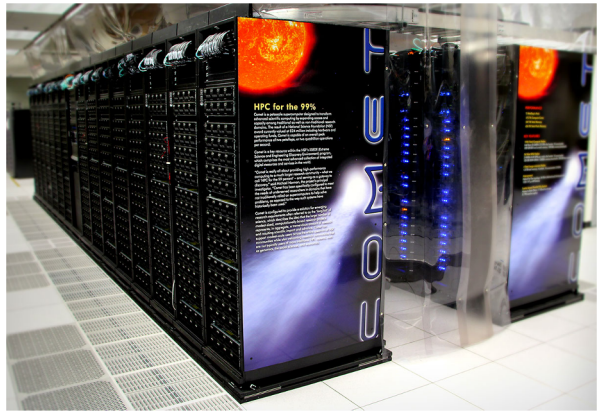
Crescat scientia; vita excolatur

COVID-19 and SARS-COV-2 Acknowledgements

Rick Stevens – Argonne National Laboratory – University of Chicago

Ian Foster, Arvind Ramanathan, Tom Brettin, Austin Clyde, Kyle Chard, Maulik Shukla, Jonathan Ozik, Chick Macal**, Yadu Babuji, Carlos Olivares,** Alex Partin, Gyorgy Babnigg**,** Xuefeng Liu, **Ben Blaiszik**, **Ryan Chard**, **Zhi Hong, Zhuozhag Li, Marcus Schwarting**, **Logan Ward**, Anna Woodward, Shantenu Jha, Kerstin Van Dam, Frank Alexander, Misha Salim, Matteo Turilli, Andre Merzky, Mike Papka, Katherine Riley, Venkat Vishwanath, Carlos Simmerling, **Hubertus Van Dam**, Gina Tourassi, Dan Stanzione, Mike Norman, and many others > 200

Computing at Argonne, Oak Ridge, TACC, SDSC, LRZ



How to dock billions of molecules to find drug candidates?

How to dock billions of molecules to find drug candidates?

How to do it hundreds of times?

Why we built this database

- In February at the beginning of the pandemic Argonne launched an effort to adapt our Cancer drug screening pipelines for COVID-19
- We had the idea that we wanted to expand the reach of virtual screening to essentially all the publicly available compound datasets
- A conventional virtual screening approach was not going to be scalable to the 100 or so receptor models we had created for COVID-19 drug targets
- Hybrid AI/HPC approach was needed to give us the throughput to virtually dock 400-800 Billion compounds (4B x 100 receptors x FRED/AutoDock)
- **Basic Plan of Attack:** Bootstrap with docking ~1-10M compounds per receptor, train DNN models with those results, infer on the large-scale data base, capture high scoring hits, redock those top hits, proceed to the rest of the pipeline (MD, QM, FE, exp screens, SAR, etc. etc.)

Which features to generate for ML?

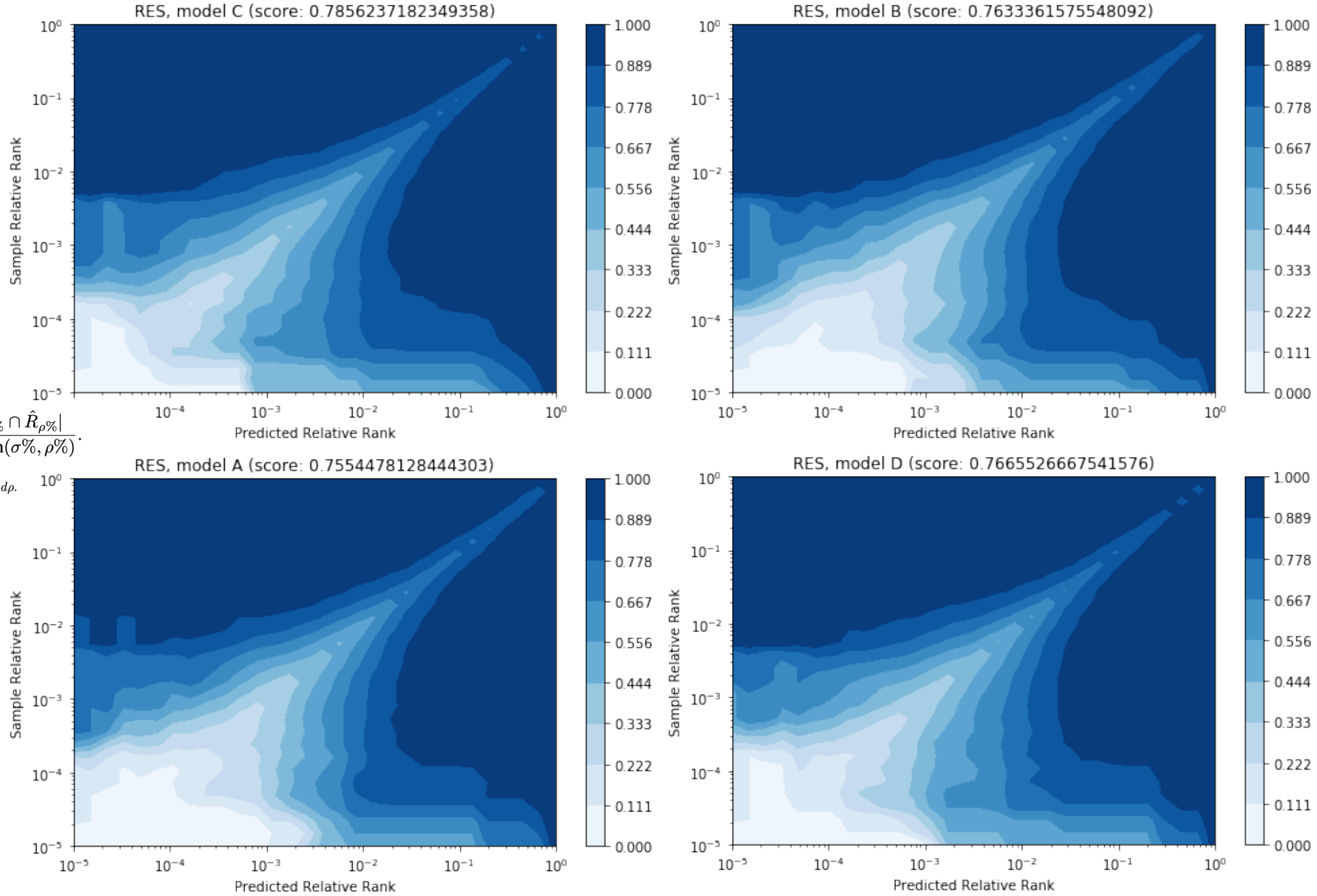
- Descriptors, Fingerprints, Graphs, Images?, voxels?, text?, Strings?, Etc.
- We had been comparing descriptors, fingerprints and graphs over the last few years for our drug response work in Cancer and AMR
- Images had been explored over the last year in our group for Cancer
- Images were looking interesting since you can bring power of CNNs and many tools from computer vision
- We settled on SMILES, Descriptors, Fingerprints and 2D Images as the base datatypes which can also easily support graphs on the fly

Images

Vector

Graphs

SMILES



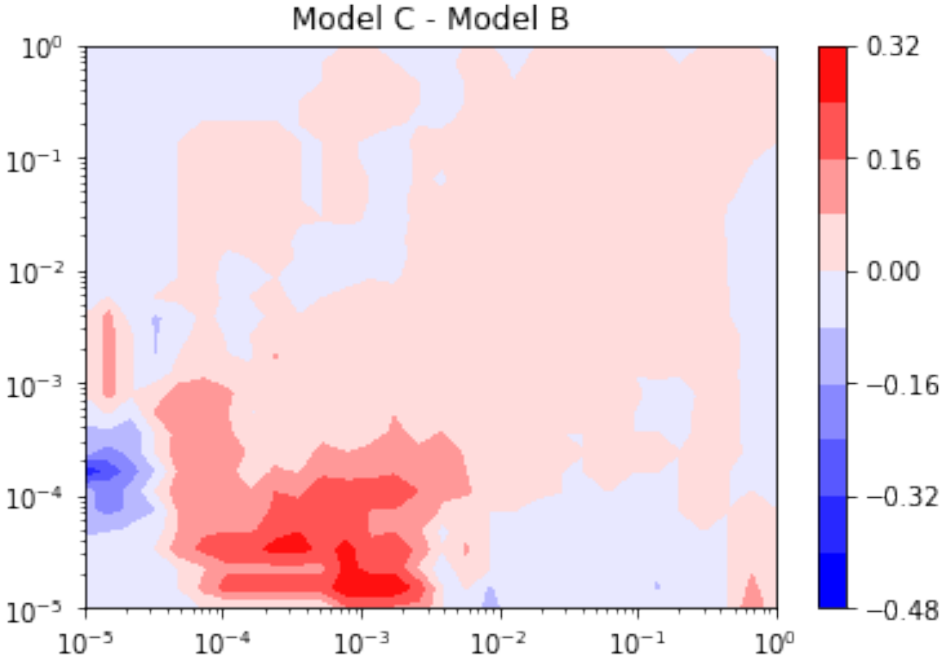
$$\text{Enrichment}(R, \hat{R}, \sigma, \rho) = \frac{|R_{\sigma\%} \cap \hat{R}_{\rho\%}|}{|R| \min(\sigma\%, \rho\%)}.$$

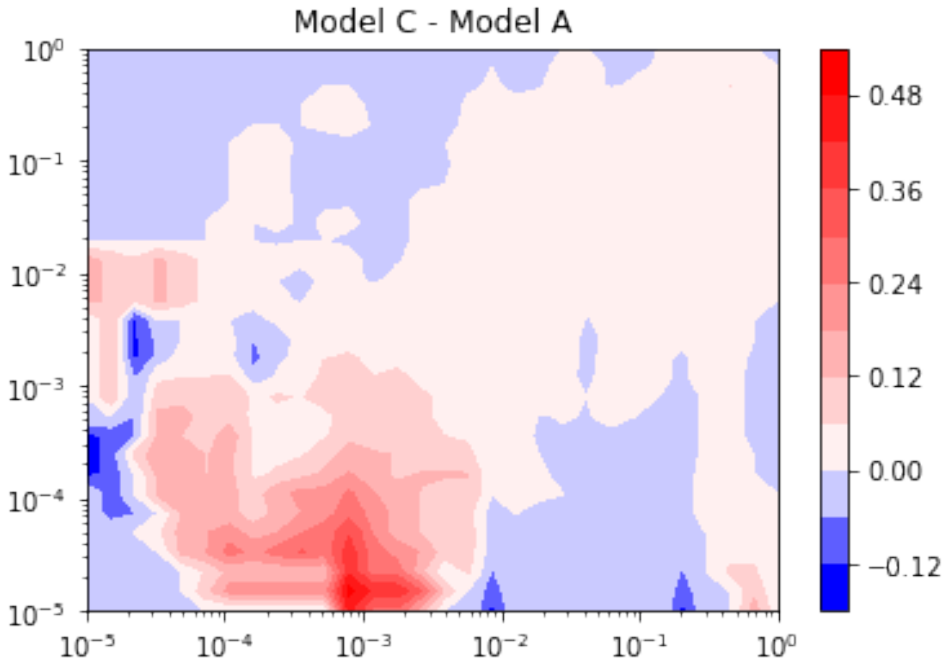
$$\text{RES Score} = \int_0^1 \int_0^1 \text{Enrichment}(R, \hat{R}, \sigma, \rho) d\sigma d\rho.$$

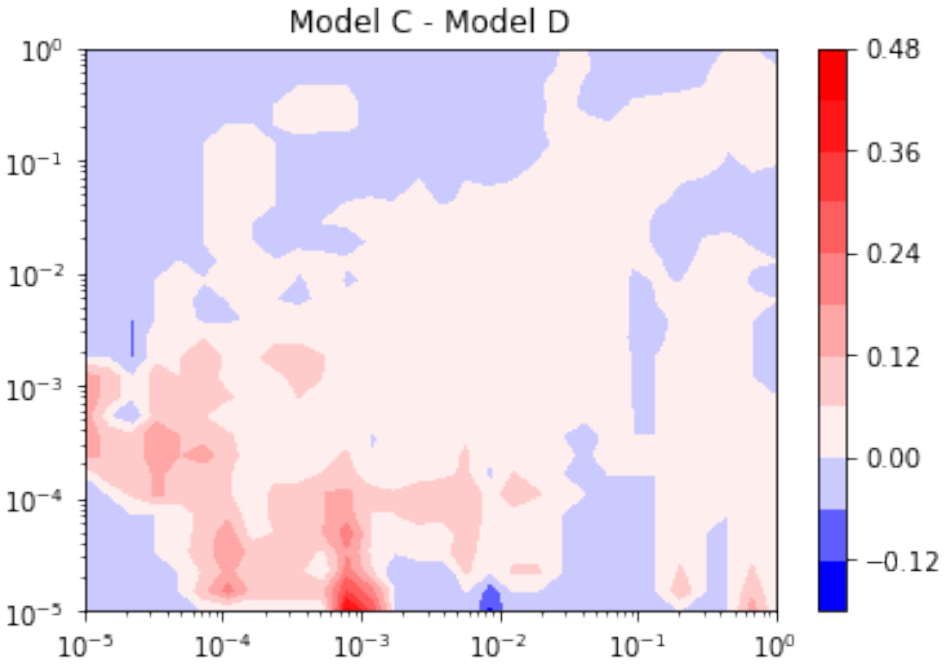
Regression Enrichment Surface (RES): https://arxiv.org/pdf/2006.01171.pdf

Images versus *X*

Images appear to do better overall in the regions of the regression enrichment surface we care about







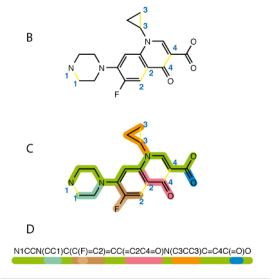
HPC-computed features for AI-based drug screening

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

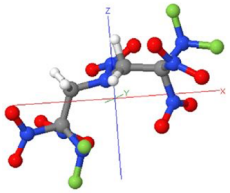
Data processing pipeline used ~2M core hours on ALCF Theta, TACC Frontera, OLCF Summit

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

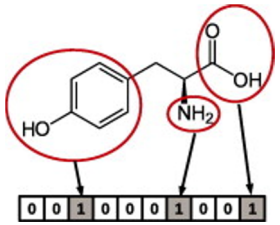
Computed data provide **crucial input features to AI models** for predicting molecular properties such as docking scores and toxicity



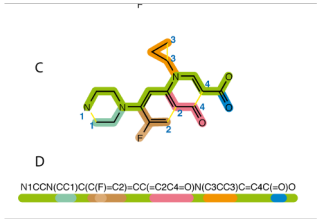
Canonical SMILES
23 CSV files with 4.2B molecules



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

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

Source datasets

Collect and process a broad set of molecular datasets from >20 distributed locations

Chemical Databases

Enamine
DRUGBANK
BindingDB GDB
eMolecules
cureFFI MOSES
ZINC15
LINCS
SureChEMBL
PubChem
And more…

Key	Canonical SMILES	Fingerprints	Descriptors	Images
BDB	1 file; 132 MB	2 files; 302 MB	link	link
CAS	1 file; 4 MB	1 file; 9 MB	5 files; 350 MB	5 files; 2 GB
DBK	1 file; 650 KB	2 files; 2 MB	1 file; 71 MB	1 file; 476 MB
DCL	1 file; 260 KB	1 file; 576 KB	1 file; 29 MB	1 file; 196 MB
DUD	1 file; 7 MB	2 files; 32 MB	10 files; 745 MB	10 files; 4 GB
E15	1 file; 1 GB	16 files; 2 GB	1,555 files; 116 GB	1,555 files; 765 GB
EDB	1 file; 18 MB	1 file; 44 MB	28 files; 2 GB	32 files; 15 GB
EMO	1 file; 1 GB	26 files; 3 GB	2,232 files; 165 GB	2,595 files; 1 TB
ENA	1 file; 85 GB	1,212 files; 191 GB	120,692 files; 8 TB	link
FFI	1 file; 117 KB	1 file; 239 KB	1 file; 11 MB	1 file; 73 MB
G13	1 file; 36 GB	978 files; 101 GB	97,739 files; 7 TB	link
G17	1 file; 2 GB	51 files; 5 GB	5,000 files; 371 GB	5,000 files; 2 TB
HOP	1 file; 37 KB	1 file; 68 KB	1 file; 2 MB	1 file; 17 MB
LIT	link	link	link	link
MCU	link	link	link	link
MOS	1 file; 93 MB	2 files; 250 MB	194 files; 14 GB	123 files; 60 GB
PCH	1 file; 6 GB	98 files; 14 GB	9,755 files; 725 GB	9,754 files; 4 TB
QM9	1 file; 3 MB	1 file; 10 MB	14 files; 993 MB	14 files; 6 GB
REP	1 file; 623 KB	1 file; 1 MB	1 file; 46 MB	1 file; 307 MB
SAV	1 file; 28 GB	266 files; 55 GB	28,323 files; 2 TB	26,282 files; 12 TB
SUR	1 file; 1 GB	18 files; 2 GB	1,792 files; 132 GB	link
ZIN	1 file; 85 GB	1,226 files; 193 GB	147,132 files; 10 TB	link
Total	**21 files; 248 GB**	**3,907 files; 571 GB**	**414 K files; 30 TB**	**412 K files; 49 TB**

The Databases were captured In March/April 2020

If sufficient external demand, we will update the collections early next year

We strongly encourage anyone using this data base to dereference through the source databases

We have since added Mcule ULTIMATE and are in the process of updating for ZINC20

Source database quality varies considerably ☺

We included some non-drug like sources for QC

We have drug like filters on our pipelines using a variety of tools PAINS etc.

Key	Description and link	# Molecules	% Uniq
BDB	The Binding Database	1,813,538	20.4
CAS	CAS COVID-19 Antiviral Candidate Compounds	49,437	55.5
CHM	CheMBL db of bioactive mols with drug-like properties	1,940,732	
DBK	Drugbank	10,095	76.1
DCL	DrugCentral Online Drug Compendium	3,981	2.4
DUD	DUDE database of useful decoys	99,782	99.7
E15	15.5M-molecule subset of ENA	15,547,091	99.7
EDB	DrugBank plus Enamine Hit Locator Library 2018	310,782	61.2
EMO	eMolecules	25,946,988	93.9
ENA	Enamine REAL Database	1,211,723,723	85.9
FFI	CureFFI FDA-approved drugs and CNS drugs	1,497	12.2
G13	GDB-13 small organic molecules up to 13 atoms	977,468,301	99.5
G17	GDB-17-Set up to 17 atom extension of GDB-13	50,000,000	100.0
HOP	Harvard Organic Photovoltaic Dataset	350	83.7
LIT	COVID-relevant small mols extracted from literature	803	
MOS	Molecular Sets (MOSES)	1,936,962	81.3
MCU	MCULE compound database	45,472,755	
PCH	PubChem	97,545,266	48.5
QM9	QM9 subset of GDB-17	133,885	84.0
REP	Repurposing-related drug/tool compounds	13,553	0.0
SAV	Synthetically Accessible Virtual Inventory (SAVI)	283,194,309	99.8
SUR	SureChEMBL dataset of molecules from patents	17,915,384	9.8
ZIN	ZINC15	1,475,804,828	85.1
Total		**4,206,934,042**	

Step 1: Convert to canonical SMILES

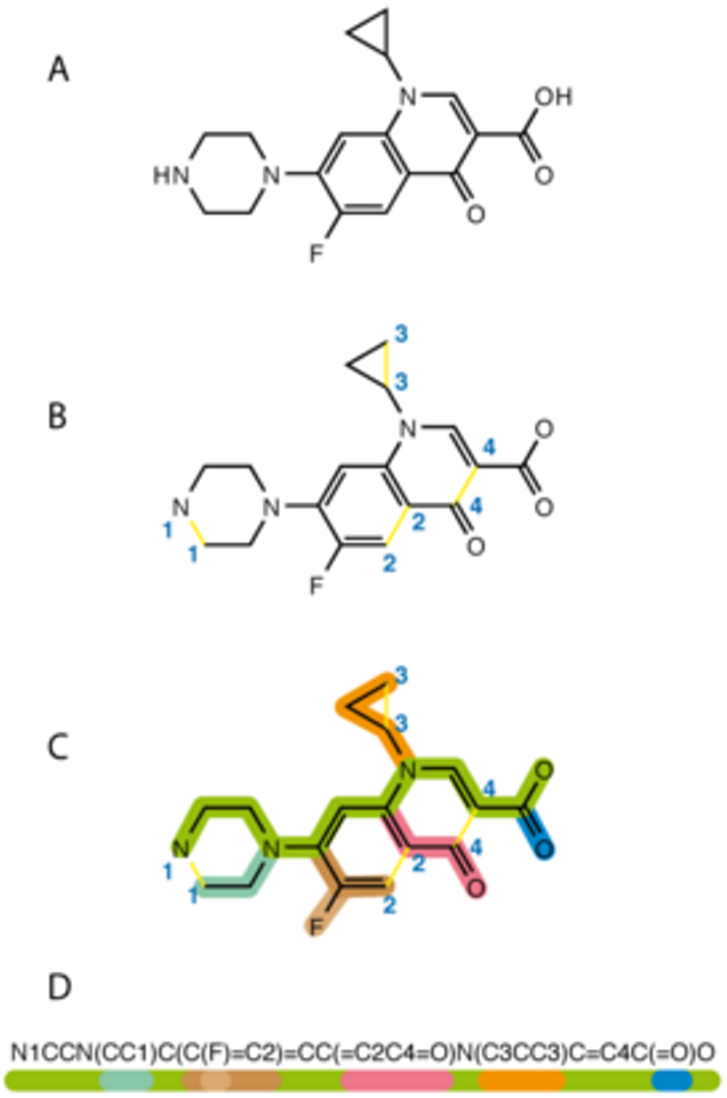
Simplified molecular-input line-entry system (SMILES) is a specification in the form of a line notation for describing the structure of chemical species using short ASCII strings.

Unfortunately, SMILES may be represented differently for the same molecule and are not optimized for matching.

BrC1=C2C3=C4C(C=C5C=NC(S3)=C45)=C2SC=C1

Brc1ccsc2c1c1sc3c4c1c2cc4cn3

During this step we also computing InchiKeys and internal IDs



Step 2: Compute molecular descriptors

Mordred Descriptors vs Dragon7 vs others

- For machine learning inference we want to identify features of molecules that define structure, properties, etc.
- Molecular descriptors: “final result of a logical and mathematical procedure, which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment”
- Thousands of molecular descriptors have been developed,
  - E.g., the number of carbon atoms; molecular weight; predictive values of LogP; properties calculated from 2D structures (e.g., Eccentric Connectivity Index) and 3D structures (e.g., charged partial surface area); and properties based on quantum mechanics

Load a Dataset from CSV

```
data_file = 'FFI-desc.csv'
data = pd.read_csv(data_file)

headers = pd.read_csv('descriptor_headers.csv')
headers = [h.strip() for h in list(headers.columns)]
headers = ['dataset','name','smiles'] + headers

data.columns = headers
data.head()
```

	dataset	name	smiles	ABC	ABCGG	nAcid	nBase	SpAbs_A	SpM
0	FFI	Abiraterone	CC(=O)O[C@H]1CC[C@]2(C(=CCC3C2CC[C@]2(C3CC=C2c...	23.503382	17.343279	0.0	0.0	38.038227	2.6
1	FFI	Acamprosate	CC(=O)NCCCS(=O)(=O)O	7.766604	7.585127	1.0	0.0	11.571944	2.1
2	FFI	Acarbose	OC[C@H]1O[C@H](O[C@@H]2[C@@H](CO)O[C@H]([C@@H]...	33.790264	26.700638	0.0	1.0	55.609055	2.48
3	FFI	Acebutolol	CCCC(=O)Nc1ccc(c(c1)C(=O)C)OCC(CNC(C)C)O	17.546020	14.908962	0.0	1.0	28.601421	2.35
4	FFI	Acenocoumarol	CC(=O)CC(c1c(=O)oc2c(c1O)cccc2)c1ccc(cc1)[N+](...	20.172247	16.404648	0.0	0.0	32.642193	2.48

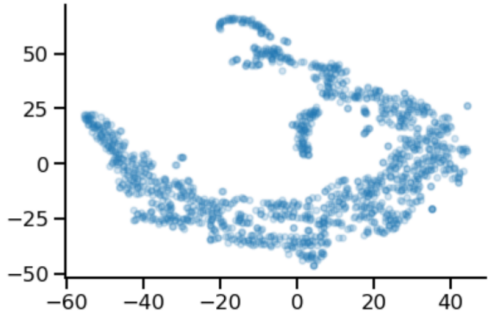
5 rows × 1829 columns

Display a TSNE Embedding of the Space

```
X = [descriptors for smiles, (name, descriptors) in data.items()]

X_embedded = TSNE(n_components=2).fit_transform(df_desc.values)
```

```
plt.scatter(X_embedded[:,0],
            X_embedded[:,1],
            alpha=0.2,
            s=25)
sns.despine()
```



Step 3: Compute fingerprints for similarity search

- Having identified potential candidate molecules it is desirable to also explore similar molecules
- Not easy to compare billions of SMILES.. need more efficient representation
- Molecular fingerprints encode the structure of a molecule.
  - The most commonly as a series of binary digits (bits) that represent the presence or absence of particular substructures in the molecule.
  - Can then compute distance between these fingerprints
- Two programs
  1. create fingerprints for all SMILES
  2. search for most similar fingerprints for a set of SMILES

```python
import pandas as pd

data_file = 'FFI-0-1000000.csv'
df = pd.read_csv(data_file,
                 names=["dataset","name","smiles","fingerprint"])
df.head()
```

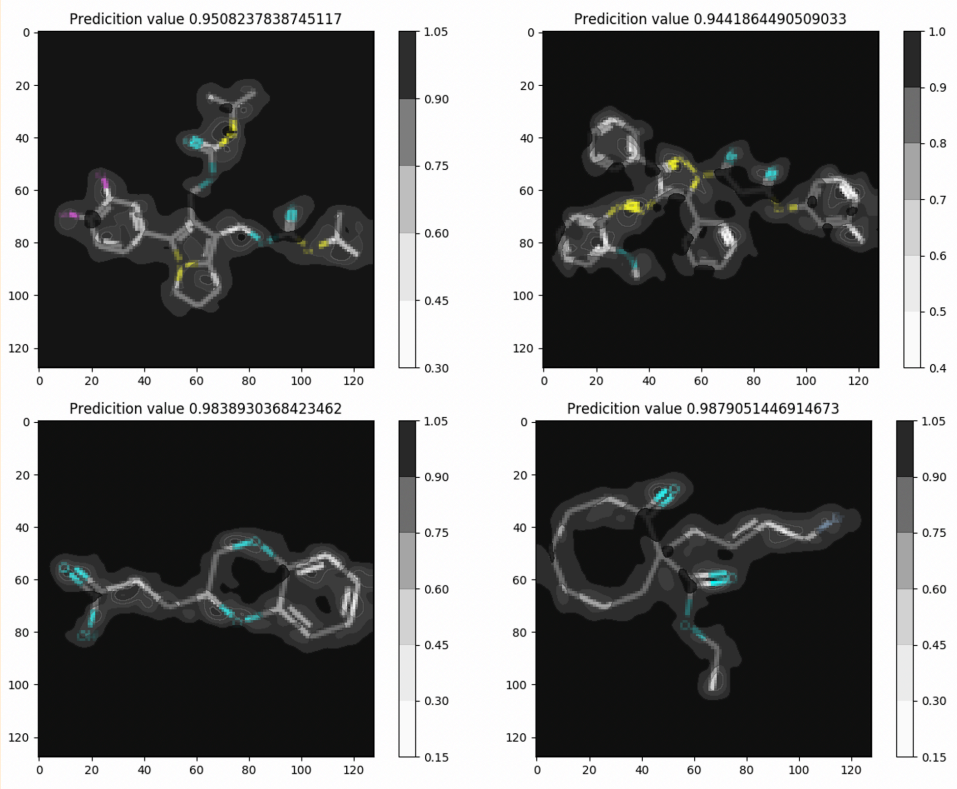
	dataset	name	smiles	fingerprint
0	FFI	Abacavir	NC1=NC2=C(N=CN2[C@@H]2C[C@H](CO)C=C2)C(NC2CC2)=N1	4P///wAIAAAuAAAAEhhyXoI2TOg8fNAO9GBi7AoEphhkFI...
1	FFI	Abiraterone	CC(=O)O[C@H]1CC[C@]2(C)C3CC[C@@]4(C)C(CC=C4C4=...	4P///wAIAAA2AAAADhDMUigGWDIEsjZ0hBwCDoguNlhQjF...
2	FFI	Acamprosate	CC(=O)NCCCS(O)(=O)=O	4P///wAIAAAWAAAAoOhe0AkA3HIgfpZ5AGROTmy2gQCMDQ...
3	FFI	Acarbose	C[C@H]1O[C@H](O[C@@H]2[C@@H](CO)O[C@H](O[C@@H]...	4P///wAIAAAyAAAAiBaaCiwULpDgLnA6cEJaTAwUEKg6Ig...
4	FFI	Acebutolol	CCCC(=O)NC1=CC(C(C)=O)=C(OCC(O)CNC(C)C)C=C1	4P///wAIAAAvAAAAAk5MGi5kHAxGbhQdAD5K0lYOSKwWGE...

Step 4. Generate Molecular Image Files

128 x 128 pixels (color)

Images were created for each molecule to simplify application of deep learning methods

Attention Maps on Images in Drug Response



Load a Dataset from Pickle

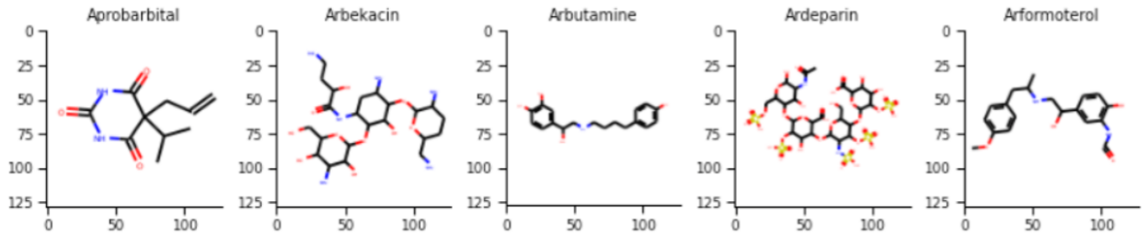
```
data_file = 'FFI-img.pkl'
df = pd.DataFrame(pd.read_pickle(data_file),
                  columns=['dataset','name', 'SMILES','image'])
df.head()
```

	dataset	name	SMILES	image
0	FFI	Abacavir	NC1=NC2=C(N=CN2[C@@H]2C[C@H](CO)C=C2)C(NC2CC2)=N1	<PIL.PngImagePlugin.PngImageFile image mode=RG...
1	FFI	Abiraterone	CC(=O)O[C@H]1CC[C@]2(C)C3CC[C@@]4(C)C(CC=C4C4=...	<PIL.PngImagePlugin.PngImageFile image mode=RG...
2	FFI	Acamprosate	CC(=O)NCCCS(O)(=O)=O	<PIL.PngImagePlugin.PngImageFile image mode=RG...
3	FFI	Acarbose	C[C@H]1O[C@H](O[C@@H]2[C@@H](CO)O[C@H](O[C@@H]...	<PIL.PngImagePlugin.PngImageFile image mode=RG...
4	FFI	Acebutolol	CCCC(=O)NC1=CC(C(C)=O)=C(OCC(O)CNC(C)C)C=C1	<PIL.PngImagePlugin.PngImageFile image mode=RG...

Visualize Some Molecules

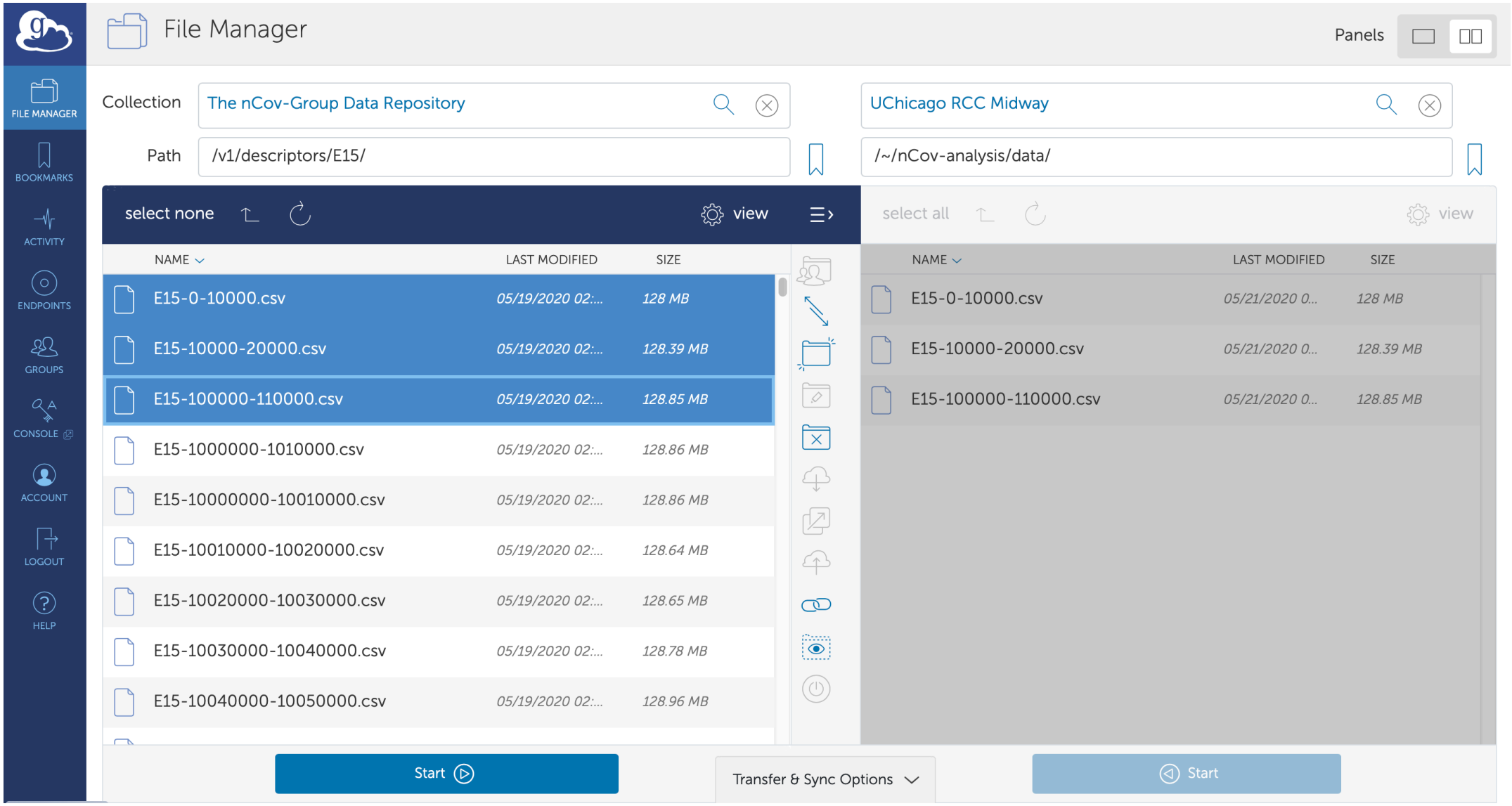
```
n_mols, offset = (5, 100)

fig, axs = plt.subplots(1, n_mols, figsize=(10,4), constrained_layout=True)
for i, (index, mols) in enumerate(df[offset:offset+n_mols].iterrows()):
    axs[i].set_title(mols[1])
    axs[i].imshow(mols[3])
    sns.despine()
```



Accessing the Data Via Globus

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



ECP CANDLE Team Prepared 60TB of data for 4.2B molecules for SARS-CoV-2 Research

arXiv:2006.02431 (q-bio)

[Submitted on 28 May 2020]

Targeting SARS-CoV-2 with AI- and HPC-enabled Lead Generation: A First Data Release

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

Researchers across the globe are seeking to rapidly repurpose existing drugs or discover new drugs to counter the the novel coronavirus disease (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). One promising approach is to train machine learning (ML) and artificial intelligence (AI) tools to screen large numbers of small molecules. As a contribution to that effort, we are aggregating numerous small molecules from a variety of sources, using high-performance computing (HPC) to computer diverse properties of those molecules, using the computed properties to train ML/AI models, and then using the resulting models for screening. In this first data release, we make available 23 datasets collected from community sources representing over 4.2 B molecules enriched with pre-computed: 1) molecular fingerprints to aid similarity searches, 2) 2D images of molecules to enable exploration and application of image-based deep learning methods, and 3) 2D and 3D molecular descriptors to speed development of machine learning models. This data release encompasses structural information on the 4.2 B molecules and 60 TB of pre-computed data. Future releases will expand the data to include more detailed molecular simulations, computed models, and other products.

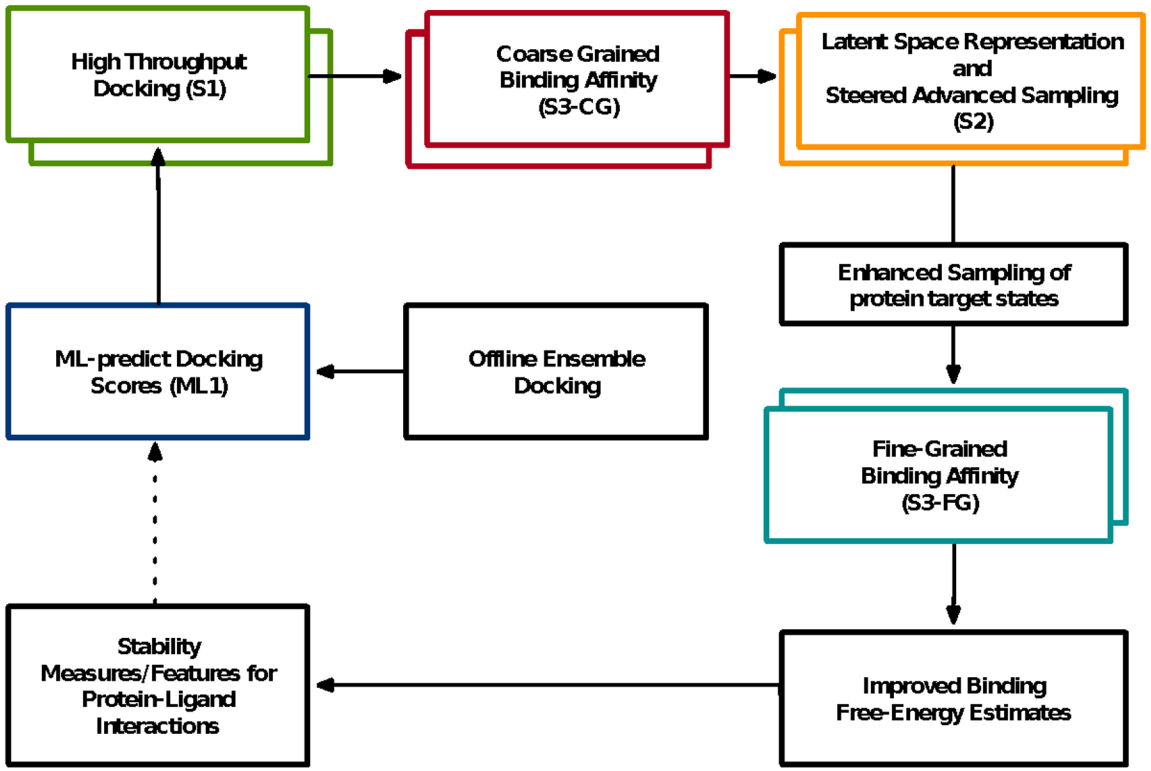
How are we accessing the data?

- We have the data in several forms
  - SQL database with various keys
  - Pickle files for Python loading (tfrecords, etc.)
  - CSVs for general computing
- We have built a few “quick and dirty” query tools
  - Fingerprint searching
  - Neural Fingerprint searching
  - Image similarity searching
- We tend to use subsets with tools like FastROCS on our GPU clusters for structural/shape/color searches

Scalable AI-HPC Infrastructure for Drug Screening

Multi-stage campaign employed to select promising drug candidates:

- **Stage-1**: High-throughput ensemble docking to identify small molecules ("hits")
- **Stage-2**: AI-driven Molecular Dynamics for modeling specific binding regions and understanding mechanistic changes involving drugs
- **Stage-3**: Binding Free Energy calculations of promising leads ("Hit-to-Lead")

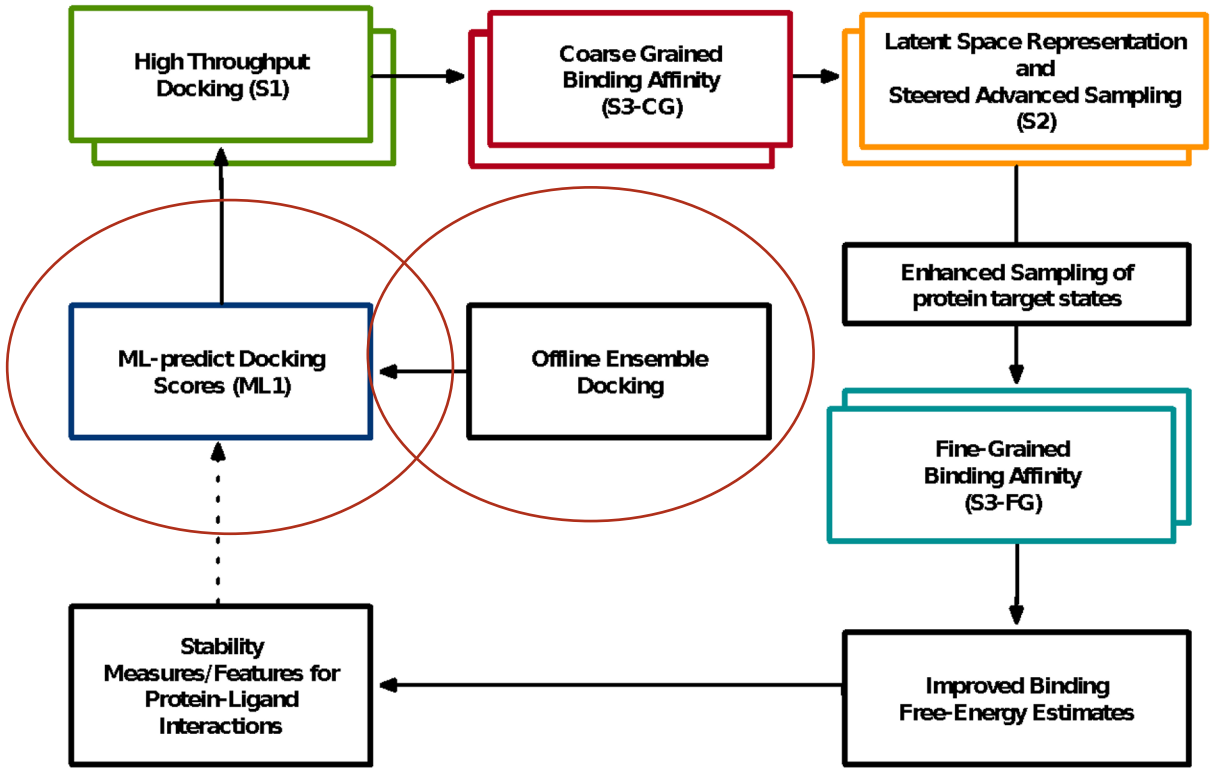


Scalable AI-HPC Infrastructure for COVID-19

Stage-0 (Offline) Building Surrogate

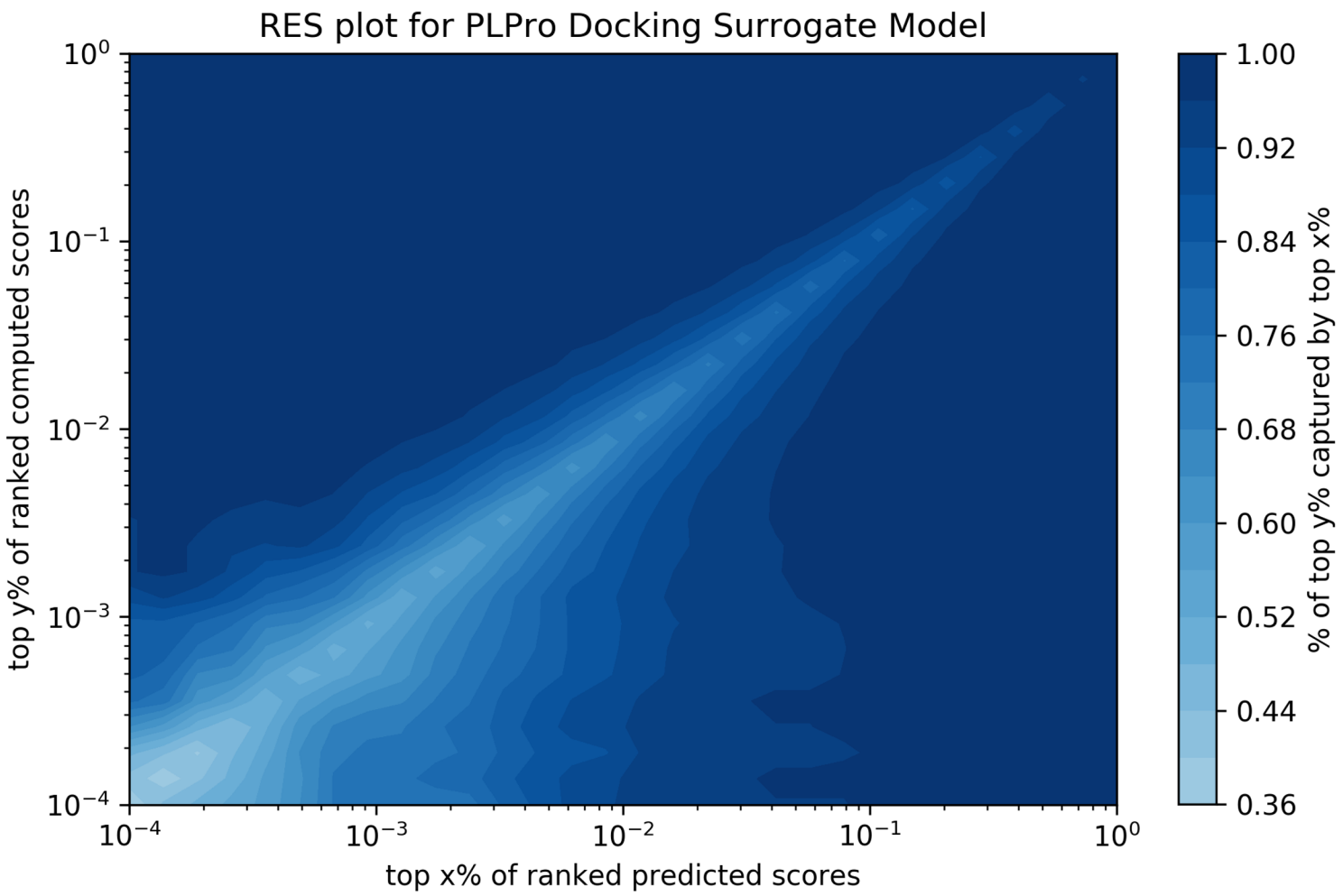
- ~O(10^8) docking hits to generate enough diversity data to use to train & create a surrogate model (ML1)
- ML1 Surrogate: A “resnet-50” deep neural network
- Transforms image representations of ligand molecules into a docking score

https://arxiv.org/abs/2010.06574



S1 ⟹ ML1 (50,000x performance improvement on single x86)

Impacting SARS-CoV-2 Medical Therapeutics



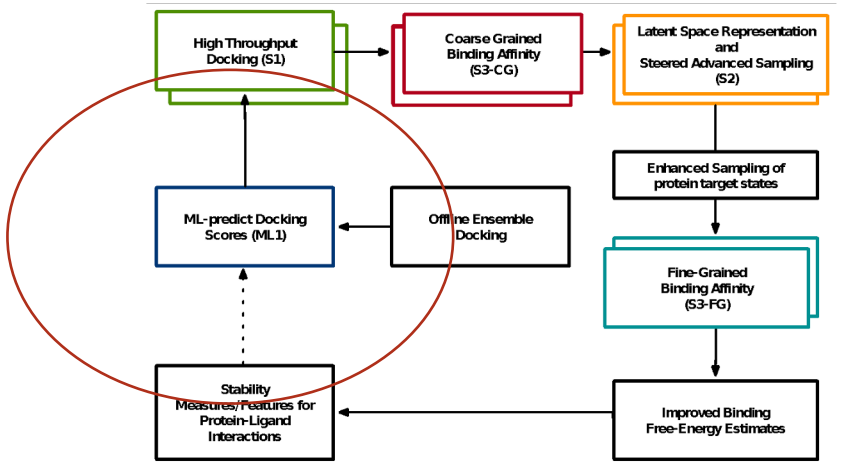


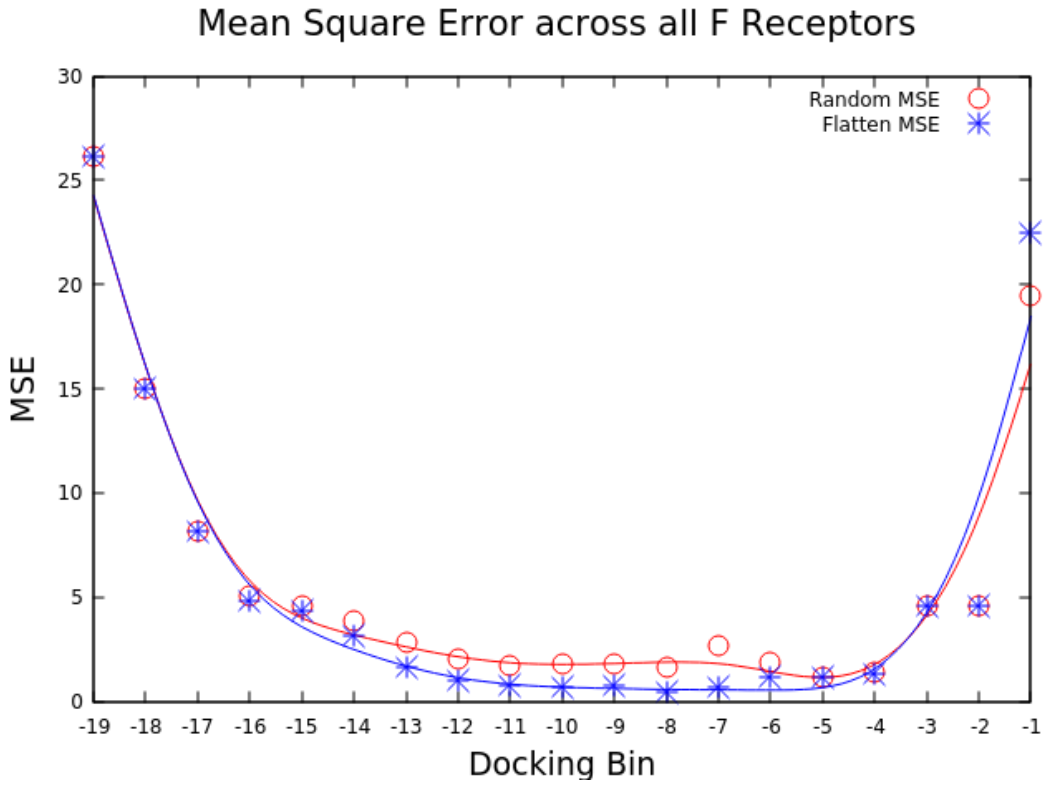
Table 3: Throughput and performance measured as peak flop per second (mixed precision, measured over short but time interval) per Summit node (6 NVIDIA V100 GPU).

Comp.	#GPUs	Tflop/s	Throughput
ML1	1536	753.9	319674 ligands/s
S1	6000	112.5	14252 ligands/s
S3–CG	6000	277.9	2000 ligand/s
S3–FG	6000	732.4	200 ligand/s

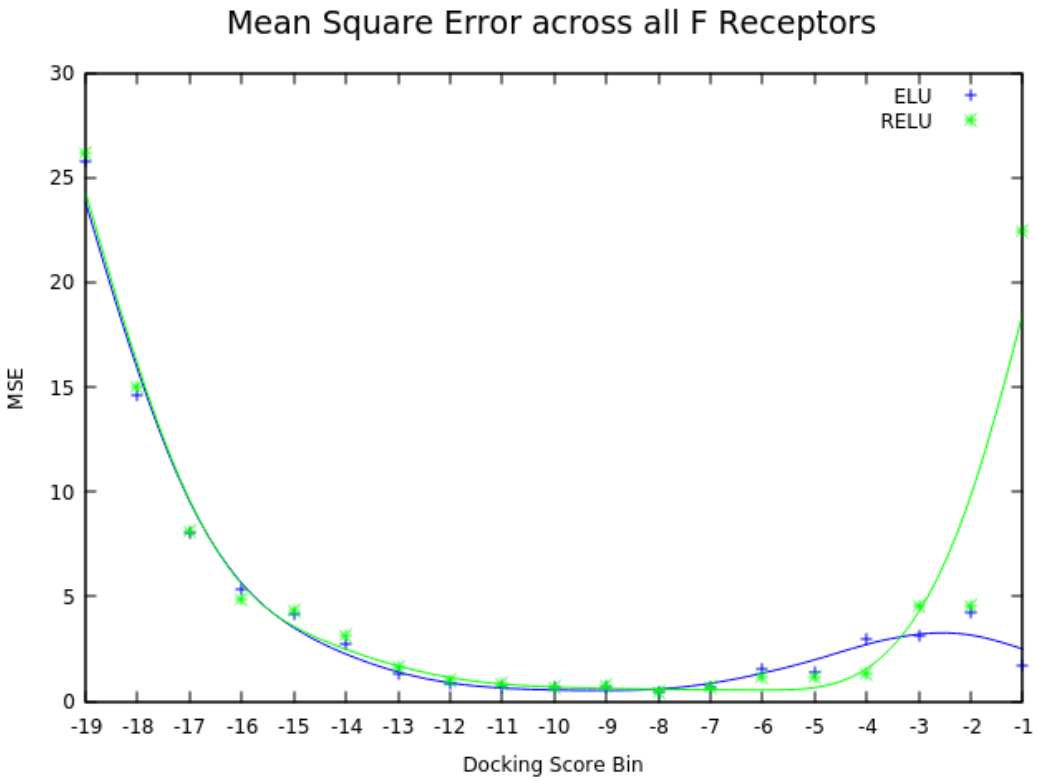
Regression Analysis Surface (RES): https://arxiv.org/pdf/2006.01171.pdf

Error analysis for ML inferencing (ML1)

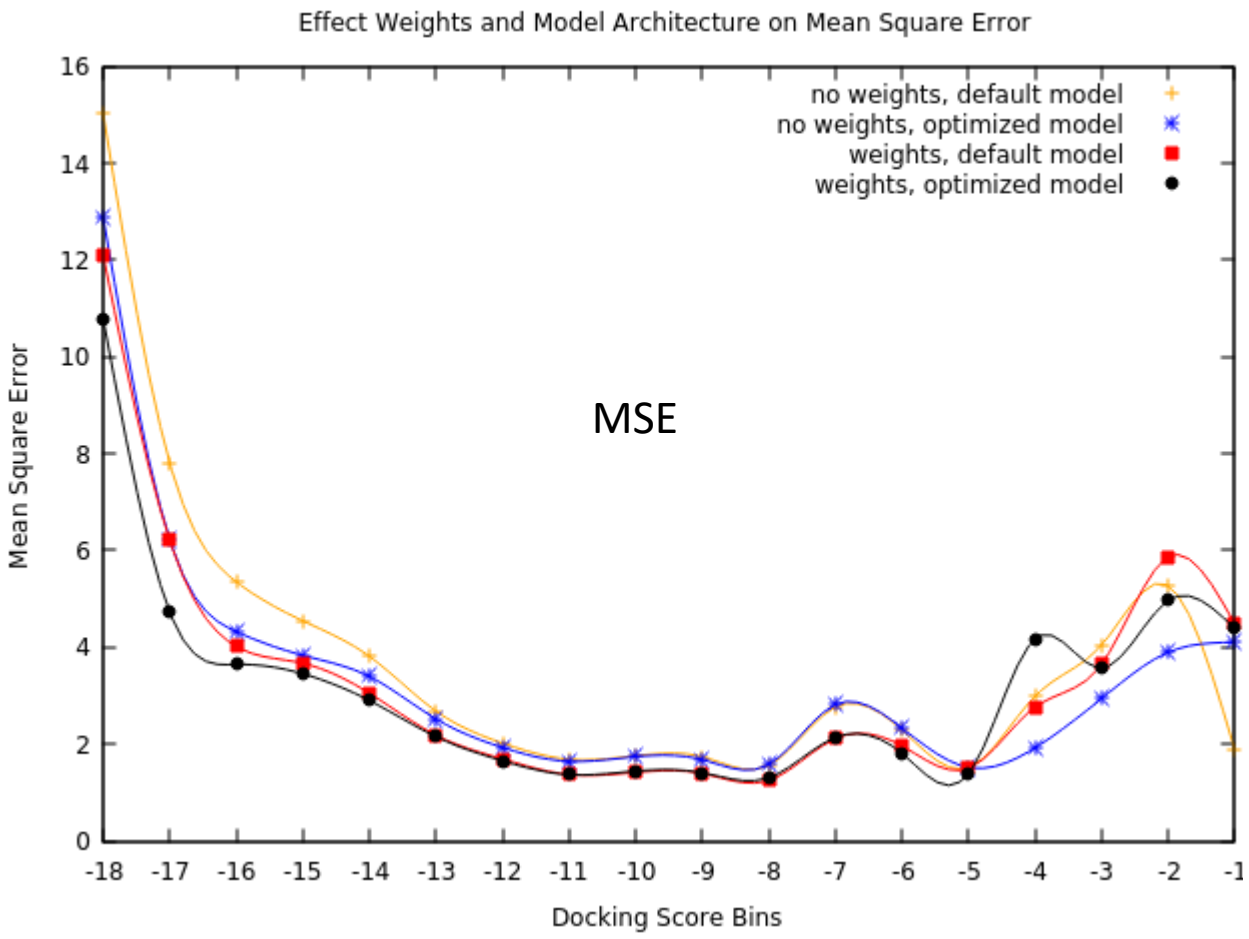
MSE across all F receptors for different sampling techniques (random vs flatten)

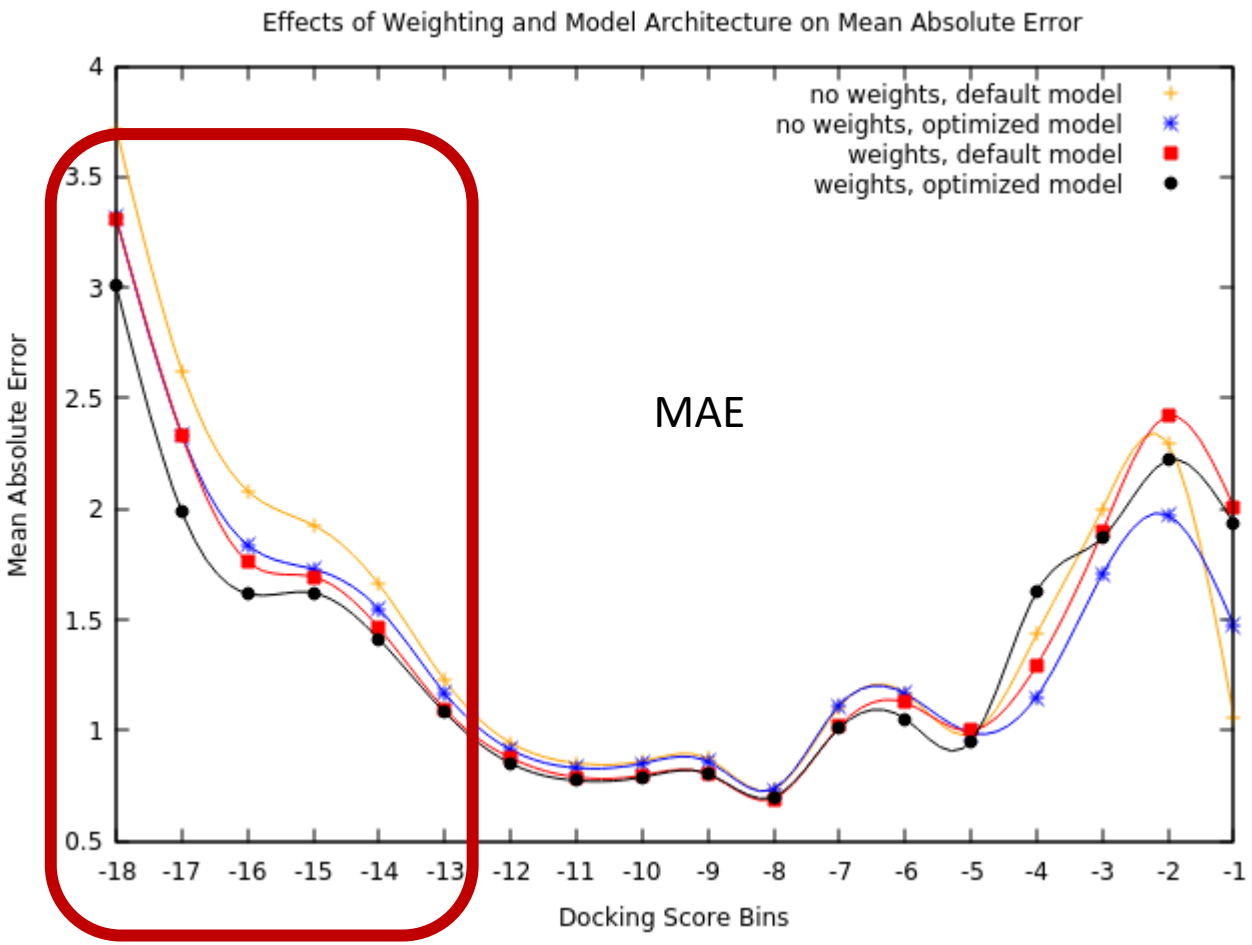


MSE across all F receptors for different activations (Elu vs Relu)



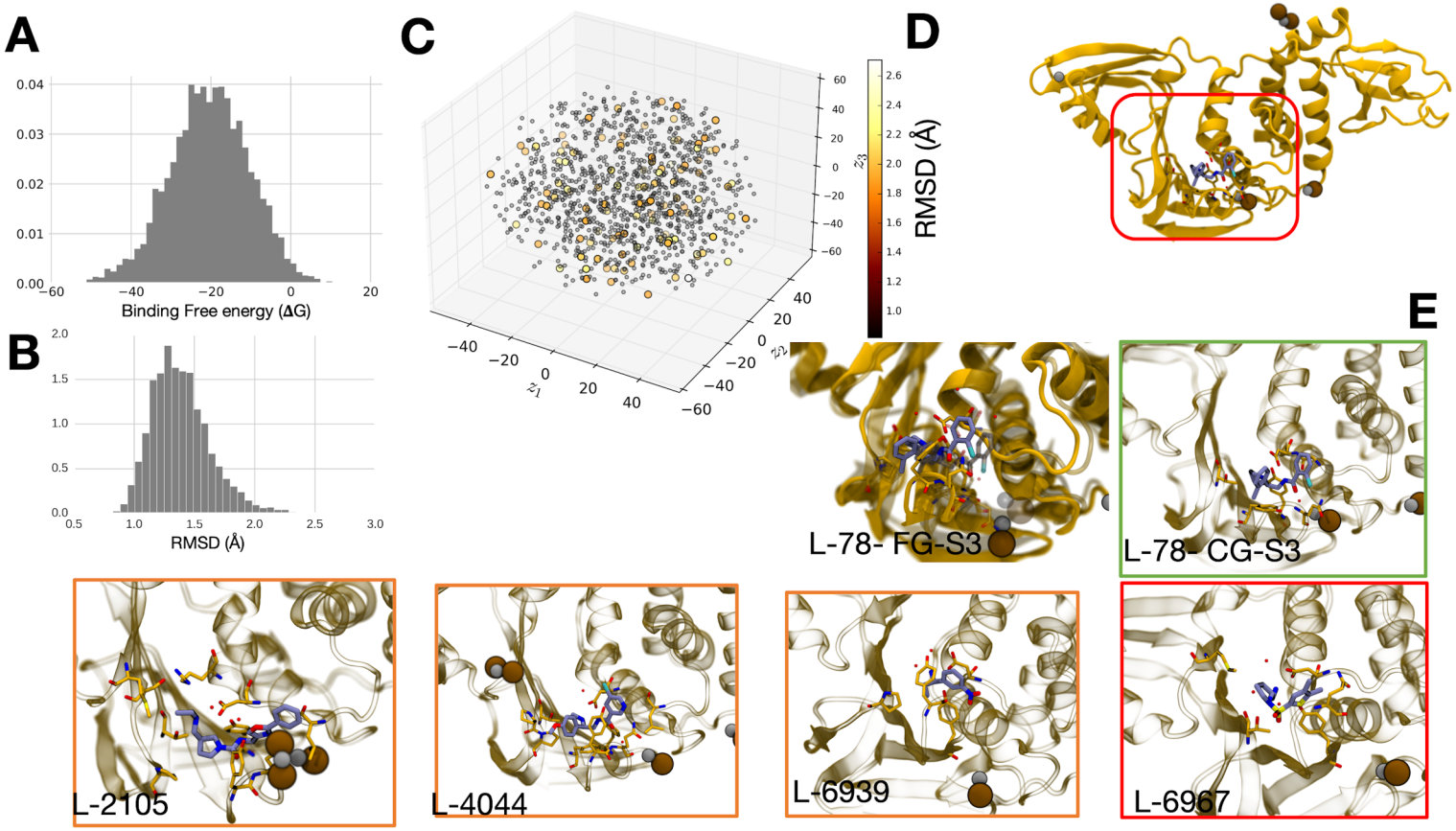
Sample Weighting and Model HPO





model architectures and sample weighting vs no sample weighting

Impacting SARS-CoV-2 Medical Therapeutics



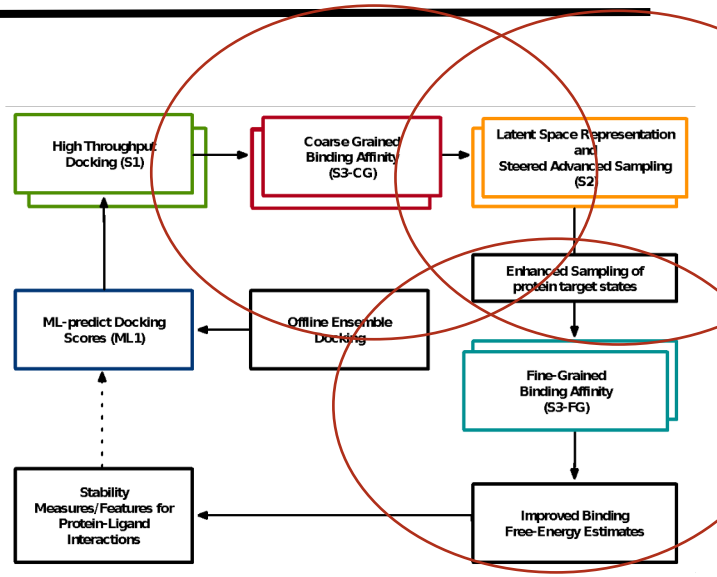


Table 3: Throughput and performance measured as peak flop per second (mixed precision, measured over short but time interval) per Summit node (6 NVIDIA V100 GPU).

Comp.	#GPUs	Tflop/s	Throughput
ML1	1536	753.9	319674 ligands/s
S1	6000	112.5	14252 ligands/s
S3–CG	6000	277.9	2000 ligand/s
S3–FG	6000	732.4	200 ligand/s

IMPECCABLE: Integrated Modeling PipelinE for COVID Cure by Assessing Better LEads

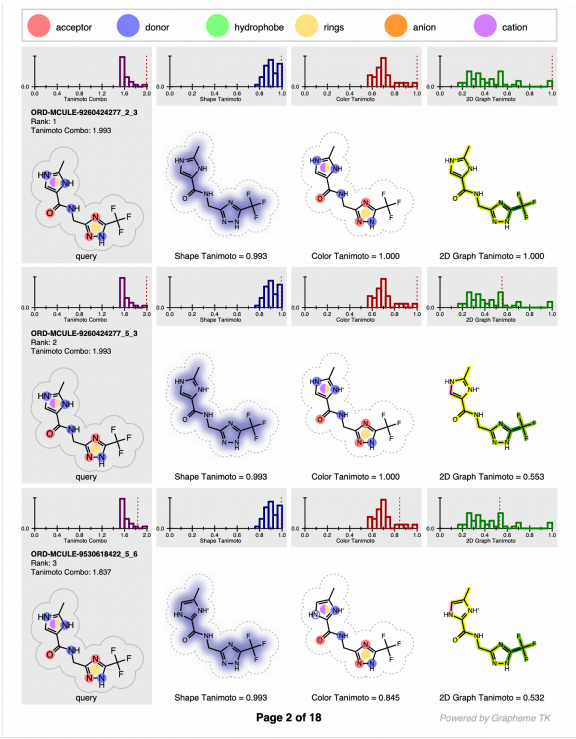
https://arxiv.org/abs/2010.06574

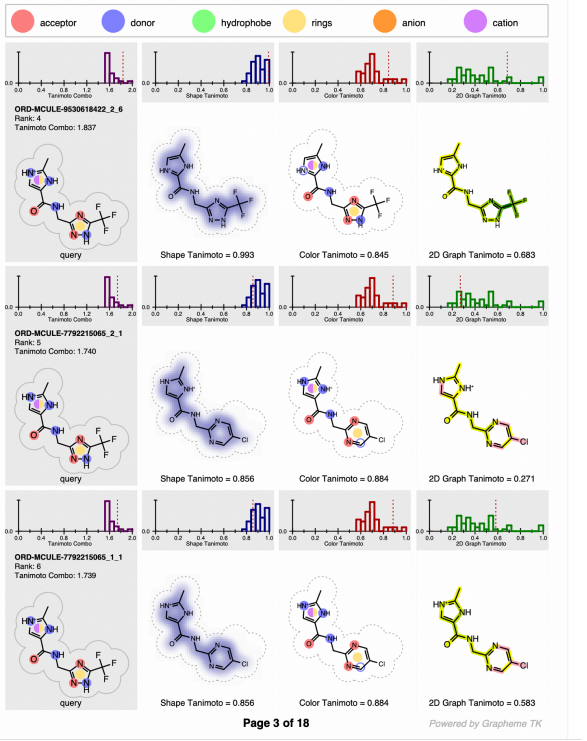
Impacting SARS-CoV-2 Medical Therapeutics

- **Scale of Operation:**
  - ~10^{10} docking calculations using OpenEye and Autodock in ratio 10:1
  - ~10^{12} docking surrogate (ML inference) docking scores (**3.6 hours per 4.2B**)
  - Thousands of ML-driven MD calculations over multiple platforms (Summit, Lassen, …)
  - 5 x 10^4 Binding Free Energy Calculations across machines
  - 2.5 x 10^6 node-hours (equal to ~25 days of 100% of Summit)
    - Assuming 5-year lifetime of Summit at $500M $6M cost of computing!
- For S1, we estimate 1.25 x 10^6 node hours (lower bound)
  - Peak Performance: ~4000 nodes for docking studies on Frontera (06 Sep 2020),
- Robust and Extensible Computational Infrastructure and Capabilities
  - Campaign operation over multiple heterogeneous resources
  - Extending computational infrastructure to covalent inhibitors of cysteine proteases

Current Status

- The DOE nine lab NVBL group has completed our first rounds of virtual screening
- We experimentally screened 1200 compounds on 10 COVID-19 drug targets
- We have 63 hits that are in downstream experimental pipelines
- 600 additional molecules are entering antiviral assays to bracket the initial hits
- The Compound database is now being used to support a variety of projects in Cancer and Machine Learning





Special Thanks

- Ian Foster and the team at Globus Labs, UChicago and Argonne
- Arvind Ramanathan, Austin Clyde, Tom Brettin at Argonne
- Shantenu Jha, Huub VanDamn and team at Rutgers/Brookhaven
- Peter Coveney at University College London

- And the the NVBL team at Argonne, Oak Ridge, Brookhaven, Livermore, Los Alamos, Pacific Northwest, Sandia, Berkeley, SLAC

COVID-19 and SARS-COV-2 Acknowledgements

Rick Stevens – Argonne National Laboratory – University of Chicago

Ian Foster, Arvind Ramanathan, Tom Brettin, Austin Clyde, Kyle Chard, Maulik Shukla, Jonathan Ozik, Chick Macal**, Yadu Babuji, Carlos Olivares,** Alex Partin, Gyorgy Babnigg**,** Xuefeng Liu, **Ben Blaiszik**, **Ryan Chard**, **Zhi Hong, Zhuozhag Li, Marcus Schwarting**, **Logan Ward**, Anna Woodward, Shantenu Jha, Kerstin Van Dam, Frank Alexander, Misha Salim, Matteo Turilli, Andre Merzky, Mike Papka, Katherine Riley, Venkat Vishwanath, Carlos Simmerling, **Hubertus Van Dam**, Gina Tourassi, Dan Stanzione, Mike Norman, and many others > 200