TARGETING COVID-19 PROTEOME WITH AI & MULTISCALE SIMULATIONS

ARVIND RAMANATHAN (ON BEHALF OF THE COVID-19 TEAM – ANL + BNL)
Data Science & Learning Division, Computing, Environment and Life Sciences, Argonne National Laboratory, Lemont, IL 60439
CASE, University of Chicago
http://ramanathanlab.org
ramanathana@anl.gov

Mar 18, 2020
INTRODUCTION TO COVID-19 AND SARS-COV-2

- Observed first in Wuhan (Dec 2019)
  - Quickly spread to the province of Hubei and then onto the world
- Spreads via close contact or through respiratory particles
- Virus is larger and far more stable than its counterparts (SARS and MERS)
  - can live on surfaces for a while
- Need a comprehensive strategy to identify small molecules (or other therapeutic strategies) to treat infection

USING AI/ML TO DISCOVER DRUGS THAT CAN TARGET SARS-COV-2 PROTEOME

0

Docking @ Scale across supercomputing facilities

Literature (PubMed)

Drug Databases (ENAMIN, DRUGBANK, ZINC, etc.)

1

APS

PDB (Structure)

COVID-19 specific targets

10,000,000,000 compounds screened with AI models

Top 2.5%

250,000,000 poses docked

Top 2.5%

6,250,000 systems build and minimized

Top 2.5%

156,250 systems simulated

(That’s about 12H on 1024 summit nodes)

2

Physics based models

AI-based models

AI-based consensus ranking/ scoring

Multiscale molecular simulations (DeepDriveMD)

Ranked hits/ compounds
FIRST RELEASE OF 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

https://2019-ncovgroup.github.io/data/
THE COVID’19 DATA PIPELINE:
USING AI AND SUPERCOMPUTERS TO ACCELERATE DRUG DEVELOPMENT

CHEMICAL LIBRARY DATABASE

4B known molecules

CANONICALIZATION

COMPUTE FEATURES

DEEP LEARNING FILTERING

FINGERPRINTING

SIMILARITY SEARCH

GENERATE IMAGES

CNN FILTERING

COMPUTING RESOURCES

cureFFI MOSES
ZINC15 LINCS
SureChEMBL PubChem
AND MORE
NATURAL LANGUAGE PROCESSING: DATASET AND CODE

Manual Extraction:
- Engaged Argonne CELS admin staff to extract small molecules from key SARS/SARS-CoV-2/MERS papers
- Extracted >800 molecules, structures

Automated Extraction:
- Labeled relevant small molecules in their natural language context in CORD-19 papers
- Built named deep-learning entity recognition (NER) models to extract drug references from entire corpus (>24k full text articles)

Lit - A Collection of Literature Extracted Small Molecules to Speed Identification of COVID-19 Therapeutics Dataset
https://doi.org/10.26311/lit
Yadu Babuji, Ben Blaiszik, Kyle Chard, Ryan Chard, Ian Foster, India Gordon, Zhi Hong, Kasia Karbarz, Zhuozhao Li, Linda Novak, Susan Sarvey, Marcus Schwarting, Julie Smagacz, Logan Ward & Monica Orozco White
Dataset published 2020 via Materials Data Facility

Drug NER Model (SpaCy)
hongzhichiicago/drug_ner_spacy


Code, training data: https://github.com/globus-labs/covid-nlp
USING AI/ML TO DISCOVER DRUGS THAT CAN TARGET SARS-COV-2 PROTEOME

0

Docking @ Scale across supercomputing facilities

Drug Databases (ENAMIN, DRUGBANK, ZINC, etc.)

Literature (PubMed)

1

APS

PDB (Structure)

COVID-19 specific targets

10,000,000,000 compounds screened with AI models

Top 2.5%

250,000,000 poses docked

Top 2.5%

6,250,000 systems build and minimized

Top. 2.5%

156,250 systems simulated

(that’s about 12H on 1024 summit nodes)

Ranked hits/ compounds

2

Physics based models

AI-based models

AI-based consensus ranking/ scoring

Multiscale molecular simulations (DeepDriveMD)
**DEEPDRIVEMD: DL DRIVEN ADAPTIVE ENSEMBLES MD**

**Simulation tasks**

- **GPU 1**
  - MD Simulation 1 (OpenMM)
  - Data collection (trajectories + contact maps [.h5])
  - collect 100,000 conformations for training

- **GPU 2**
  - MD Simulation 2 (OpenMM)

- **GPU K**
  - MD Simulation K (OpenMM)

**Machine Learning/ Deep Learning tasks**

- CVAE training 1 (Tensorflow)
- CVAE training 2 (Tensorflow)
- CVAE training M (Tensorflow)

Choose best CVAE model for inference

- CVAE inference (Tensorflow)

Cluster conformational states

- **Learning phase**
  - Hyperparameter optimization/ training
  - Terminate simulations
  - Spawn new trajectories with novel states

- **Inference phase**
  - novel states?
  - outliner detection
  - iterate until new training cycle is needed or protein is folded

**CANDLE infrastructure**

Collaboration with Shantenu Jha (Rutgers/ Brookhaven) and RADICAL team

H. Ma, et al, ParCO, 2019
H. Ma, et al, Workshop on Deep Learning on Supercomputers, 2019
DEEPDRIVEMD OVERVIEW: INTERLEAVE SIMULATIONS AND ANALYTICS ADAPTIVELY FOR REDUCING COMPUTING OVERHEADS

- Generate ensemble of simulations in parallel as opposed to one realization of process
  - Statistical approach: $O(10^6 - 10^8)$!
- Ensemble methods necessary, not sufficient!
  - Adaptive Ensembles: Intermediate data, determines next stages
- Adaptivity: How, What
  - Internal data: Simulation generated data used to determine “optimal” adaptation

"Big iron" Dedicated analytics clusters

“Big iron” Traditional Compute + Simulations

Unsustainable at Exascale

- Data movement bottlenecks
- Parallel analytics bottlenecks

Proposed Interleaving Analytics + Simulations

- High performance framework to monitor & analyze simulations as they are running with little/no modification to simulation software
- Demonstrate on protein folding, but generalize framework for broad applicability

- Generate ensemble of simulations in parallel as opposed to one realization of process
  - Statistical approach: $O(10^6 - 10^8)$!
- Ensemble methods necessary, not sufficient!
  - Adaptive Ensembles: Intermediate data, determines next stages
- Adaptivity: How, What
  - Internal data: Simulation generated data used to determine “optimal” adaptation
A VARIATIONAL APPROACH TO ENCODE PROTEIN FOLDING WITH CONVOLUTIONAL AUTO-ENCODERS

Contact matrices
MD Simulations

INPUT
4 convolution layers

1. 64 filters, 3x3 window, 1x1 stride, RELU
2. 64 filters, 3x3 window, 1x1 stride, RELU
3. 64 filters, 3x3 window, 2x2 stride, RELU
4. 64 filters, 3x3 window, 1x1 stride, Sigmoid

Reduced dimension: 3

VAE Embedding

OUTPUTS
Reconstructed contact matrices

No. of native contacts

Sampled latent embedding

Mean

Std

1 2 3 4

1 2 3 4

4 deconvolution layers

Related work:
Hernandez 17 arXiv,
Doerr 17 arXiv

INPUT

OUTPUTS
DEEP CLUSTERING OF PROTEIN FOLDING SIMULATIONS

- Convolutional Variational Auto Encoders (CVAE)
  - Low dimensional representations of states from simulation trajectories.
  - CVAE can transfer learned features to reveal novel states across simulations

- Integrating Bayesian learning to support uncertainty in sampling novel states
  - HPC Challenge (1): DL approaches to achieve near real-time training & prediction!
  - HPC Challenge (2): Hyperparameter optimization (while model is training)!

LARGER NUMBER OF SIMULATIONS IMPROVES FOLDING EFFECTIVENESS (HENCE SAMPLING)

<table>
<thead>
<tr>
<th>System</th>
<th>Total no. of simulations</th>
<th>Total simulation time (us)</th>
<th>First, subsequent simulations</th>
<th>Iterations</th>
<th>Min. RMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fs-peptide</td>
<td>840</td>
<td>18.2</td>
<td>100, 10</td>
<td>7</td>
<td>0.29</td>
</tr>
<tr>
<td>BBA (FSD-EY)</td>
<td>1200</td>
<td>22.8</td>
<td>100, 10</td>
<td>10</td>
<td>1.8</td>
</tr>
<tr>
<td>VHP</td>
<td>1200</td>
<td>22.8</td>
<td>100, 10</td>
<td>10</td>
<td>3.83</td>
</tr>
</tbody>
</table>

**RMSD**

6-7 Å

8-9 Å

8-10 Å

RMSD=9-10 Å

10-12 Å

RMSD=10-12 Å
ITERATIVE EXPLORATION OF STATES WITH DEEP LEARNING PROVIDES ACCESS TO FOLDED STATES

Training phase – no adaptive sampling
DEEPDRIVEMD SHOWS AT LEAST AN ORDER OF MAGNITUDE EFFICIENT SAMPLING COMPARED TO TRADITIONAL APPROACHES

- including the data from the “learning phase”: one order of magnitude improvement in sampling:
  - Distinct “cross-over” after training where sampling is accelerated significantly after learning/estimating the conformational states

- not including the data from “learning phase”: At least two orders of magnitude improvement in sampling:
  - If Anton trajectories take $O(\text{microsecond})$ to sample a particular state, DeepDriveMD samples it in $O(100 \text{ ns})$
  - For BBA, 98% sampled states are observed within 10 microseconds!
USING FULLY CONVOLUTIONAL VAE TO IDENTIFY CONFORMATIONAL STATES IN SPIKE PROTEIN SIMULATIONS

- Modification of the VAE architecture to accommodate larger systems (E.g. Spike protein – 1.5 million atoms)
- Model parallel example:
  - encoder and decoder on individual GPUs
  - implemented with Pytorch
- Can improve performance with layer-wise adaptive rescaling
- Joint work with Alex Brace (Argonne intern), Abe Stern (NVIDIA), Anda Trifan (CSGF), Rommie Amaro (UCSD), Carlos Simmerling (Stony Brook University)

Total no. of parameters – 1.14 billion!

<table>
<thead>
<tr>
<th>No. GPUs (V100)</th>
<th>Memory</th>
<th>Time per batch (8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20213/32510 MiB</td>
<td>7.561</td>
</tr>
<tr>
<td>2</td>
<td>9947/32510 MiB (Encoder) 12987/32510 MiB (Decoder)</td>
<td>7.481</td>
</tr>
</tbody>
</table>
USING AI/ML TO DISCOVER DRUGS THAT CAN TARGET SARS-COV-2 PROTEOMEN

0

- APS
- PDB (Structure)
- COVID-19 specific targets
- Docking @ Scale across supercomputing facilities
- Literature (PubMed)
- Drug Databases (ENAMIN, DRUGBANK, ZINC, etc.)

1

- 10,000,000,000 compounds screened with AI models
- Top 2.5%
- 250,000,000 poses docked
- Top 2.5%
- 6,250,000 systems build and minimized
- Top 2.5%
- 156,250 systems simulated
  (that’s about 12H on 1024 summit nodes)

2

- Physics based models
- AI-based models
- AI-based consensus ranking/ scoring
- Multiscale molecular simulations (DeepDriveMD)

Ranked hits/ compounds
REINFORCEMENT LEARNING DRIVEN MD

- Motivation: physics-based models are guided by an action space determined by AI
- Can we expand the compound space explored using RL?
- For SARS-CoV-2 proteome:
  - relevant for specific mutations compared to other CoV proteins
  - suggest repurposing based on shape/structural complementarity
Fragment growth with an expert docking policy

Joint work with A. Clyde, UChicago
Action space

MMGBSA

Sampling
FUTURE WORK / OUTLOOK

Conformational landscapes of proteins:

- **Sampling remains challenging**: are there techniques that can aid accurate biophysical characterization of protein conformational landscapes?
- **Deep learning / AI techniques show promise**: are they learning biophysical characteristics that can be used to guide simulations?
- **Protein interactions need “context”**: are there multi-scale methods to integrate information across experiments, simulations and theory?

AI/ML coupled to simulations (challenges)

- Improvement in additional AI/ML models
- Active learning approaches for docking ligands
- Runtime systems are unprepared for such use cases where AI/ML systems drive simulations:
  - improving exchange of data with concurrently running models
  - tracking datasets as simulations are running (online/ in situ training)
FUNDING AND ACKNOWLEDGEMENTS

- Everyone in the team (all ~300)
- Computing support:
  - ALCF, OLCF
  - TACC, SDSC, IU
  - HPC Consortium
- Funding acknowledgement:
  - DOE National Virtual Biotechnology Laboratory (NVBL)
  - Argonne internal funding (LDRD)
  - DOE Exascale computing project (Cancer Deep Learning Environment)

Simulations driven by AI depict how the CoV-2 spike protein attaches to the human ACE2 receptor protein.
THANK YOU!
(RAMANATHANA@ANL.GOV)