HPC User Forum
May 2021

ExaLearn 2021

Francis J. Alexander

Brookhaven National Laboratory

BROOKHAVEN NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

@BrookhavenLab

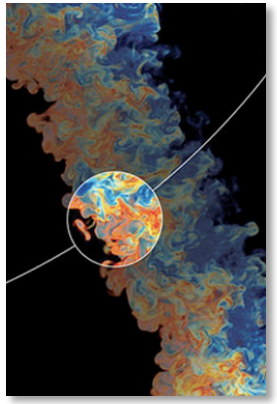
AD subprojects target national problems in DOE mission areas

National security

Next-generation, **stockpile stewardship** codes

Reentry-vehicle-environment simulation

Multi-physics science simulations **of high-energy density physics** conditions



Energy security

Turbine **wind plant** efficiency

Design and commercialization of **SMR**s

Nuclear fission and fusion reactor **materials design**

Subsurface use for **carbon capture**, petroleum extraction, waste disposal

High-efficiency, low-emission **combustion engine** and gas turbine design

Scale up of **clean fossil fuel** combustion

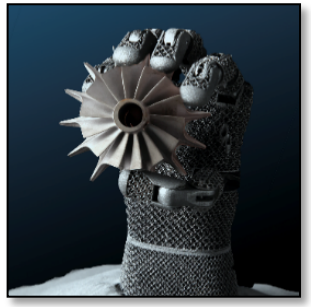
Biofuel catalyst design

Economic security

Additive manufacturing of qualifiable metal parts

Reliable and efficient planning of the **power grid**

Seismic hazard risk assessment



Scientific discovery

Cosmological probe of the standard model of particle physics

Validate fundamental laws of nature

Plasma wakefield accelerator design

Light source-enabled **analysis of protein and molecular structure** and design

Find, predict, and control materials and properties

Predict and control **magnetically confined fusion plasmas**

Demystify **origin of chemical elements**

Earth system

Accurate regional impact assessments in **Earth system models**

Stress-resistant crop analysis and catalytic conversion of **biomass-derived alcohols**

Metagenomics for analysis of biogeochemical cycles, climate change, environmental remediation

Health care

Accelerate and translate **cancer research** (partnership with NIH)

This is a diverse portfolio of applications!

ExaLearn: Co-design Center for Exascale Machine Learning Technologies

Project PI: Francis J. Alexander, Brookhaven Lab

Partner PIs and Institutions:

- Ian Foster, ANL
- Aric Hagberg, LANL
- Peter Nugent, LBNL
- Brian Van Essen, LLNL
- David Womble, ORNL
- James A. Ang, PNNL
- Michael Wolf, SNL

EXASCALE COMPUTING PROJECT

National Nuclear Security Administration

U.S. DEPARTMENT OF ENERGY | Office of Science

Overarching Goals for ExaLearn

- Provide exascale ML software for use by:
  - ECP Applications Projects
  - Other ECP Co-design Centers
  - DOE Experimental Facilities
  - DOE Leadership Class Computing Facilities
- Establish multidisciplinary collaborations in learning technologies that cross-cut ECP projects:
  - AD projects that share an interest in ML methods
  - ST projects
  - HI/PathForward projects

Guiding Principles

- ExaLearn produces a **Software Toolset** that:
  - Is applicable to multiple problems within the DOE mission
  - Has a line-of-sight to exascale computing, e.g., uses exascale platforms directly or provides essential components to an exascale workflow
  - Does not replicate capabilities easily obtainable from existing, widely available packages
  - Builds in domain knowledge where possible (not often done industry, although efforts beginning at IBM, GE, etc.); “physics”-based ML and AI are recurring themes
  - Quantifies uncertainty in a predictive capacity
  - Is both interpretable and reproducible
  - Is based on mathematically well-grounded methods
    - For example, some nice theory now for GANs, but more work needs to be done.

Application Priorities Determine Machine Learning Methods

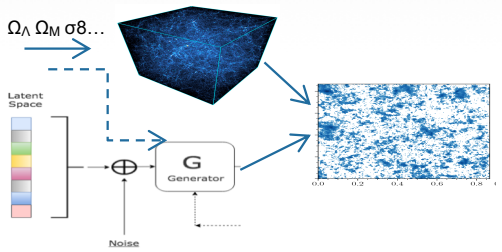
ExaLearn focuses on employing the "right tool for the job"

- Deep Learning (CNN, RNN, etc.)
- Ensemble Methods and Random Forest Methods
- Reinforcement Learning
- Kernel Methods
- Tensor Methods
- Graph-Based Learning
- Transformer Based Methods
- Large-scale System Integration (combining traditional HPC workloads with machine learning)

ExaLearn Application Pillars

Surrogates

- ML-created models
- Faster and/or higher fidelity models
- Generative networks
- Using ML to replace complicated physics
- Cosmology



Control

- ML-controlled experiments
- Efficient exploration of complex space
- Reinforcement Learning
- Use RL agent to control light source experiments
- Temperature control for Block Co-Polymer (BCP) experiments

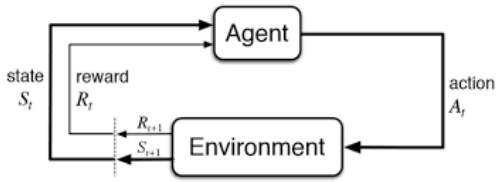
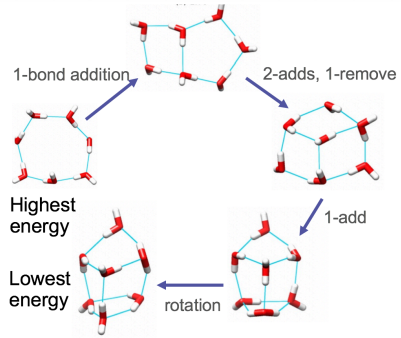


Image courtesy Sutton, Barto, Reinforcement Learning 2017

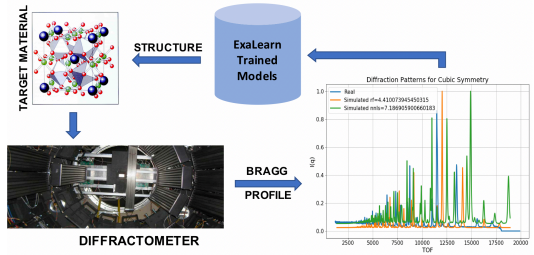
Design

- ML-created physical structures
- Optimized proposal for desired behavior of structure within complex design space
- Graph-Convnets
- Use Graph-CNN to propose new structures that respect chemistry
- Molecular Design

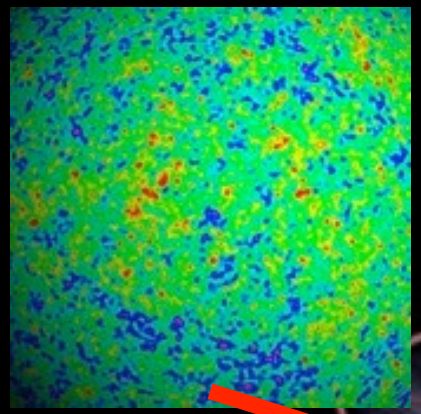


Inverse

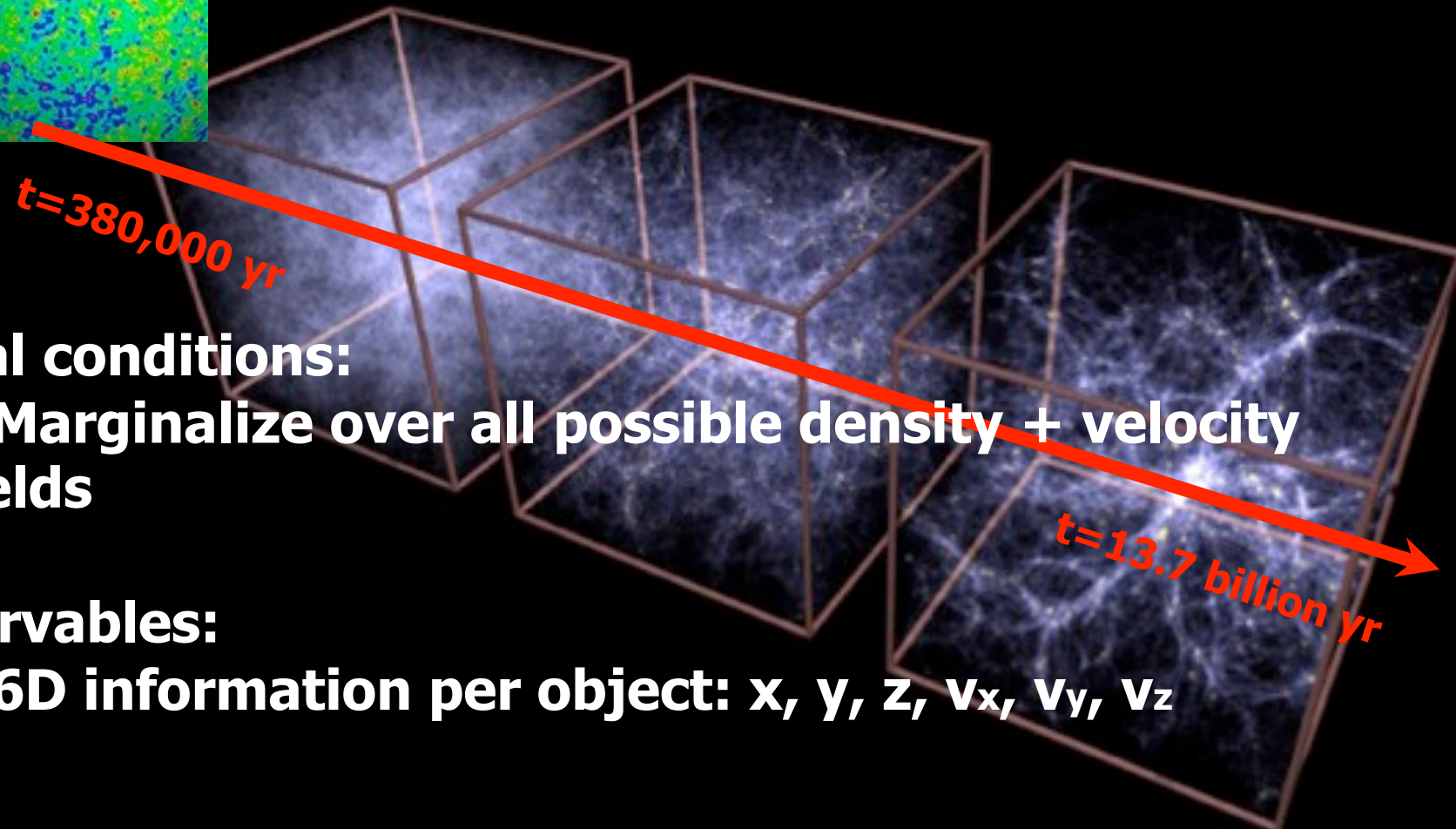
- ML projection from observation to original form
- Back-out complex input structure from observed data
- Regression models
- Predicting crystal structure from light source imaging
- Material structure from neutron scattering



Fitting the Universe



$\mathcal{L}$(Our Universe| initial conditions, forces)



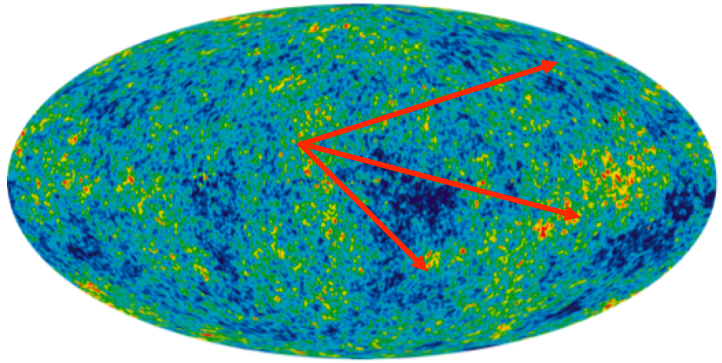
Initial conditions:

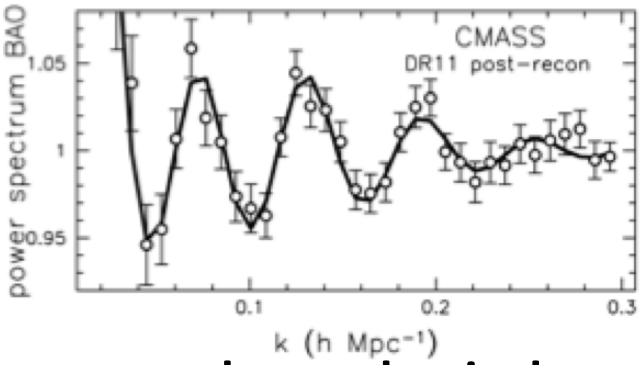
Marginalize over all possible density + velocity fields

Observables:

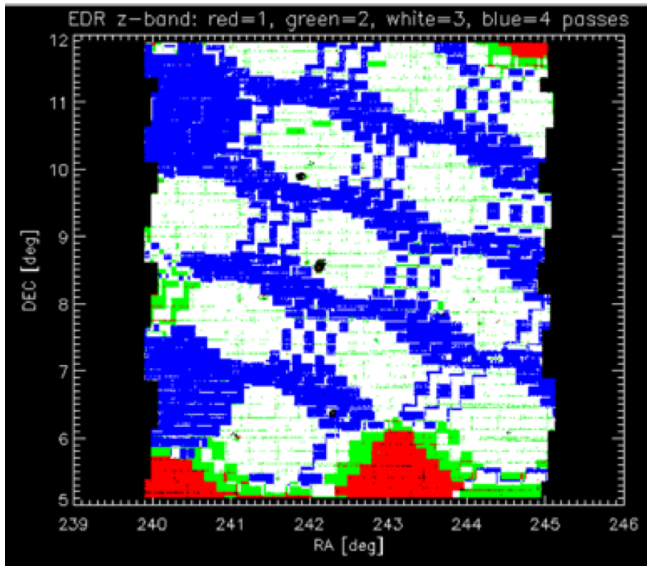
6D information per object: x, y, z, v_x, v_y, v_z

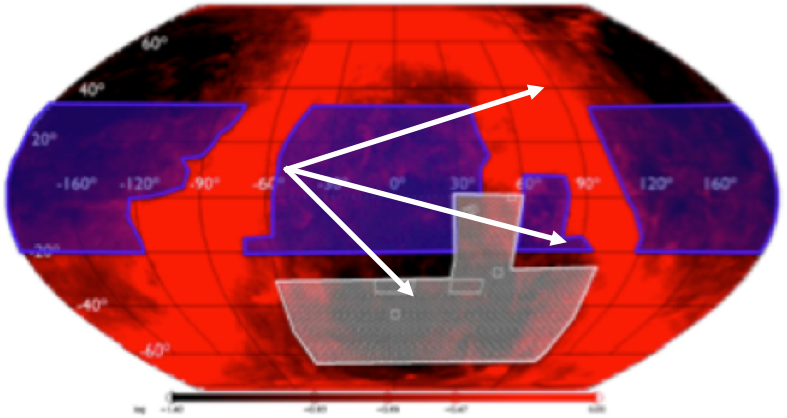
Why do we need simulations?



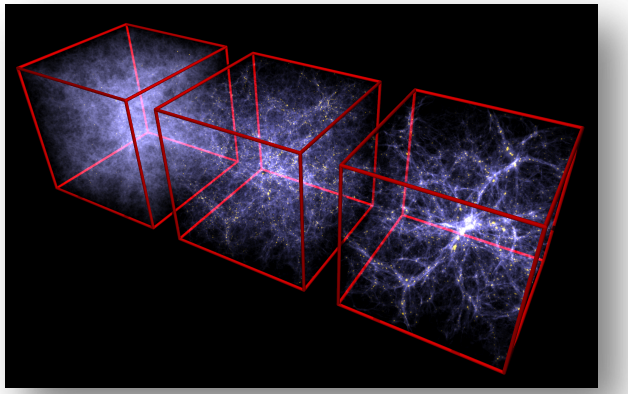


Two point correlation functions are a way to measure cosmology: what is the power spectrum of distances between every galaxy and every other galaxy as a function of time (redshift)? It requires a deep understanding of galaxy selection, completeness and the systematics of each.





Surrogates: Realistic Simulations on the Cheap

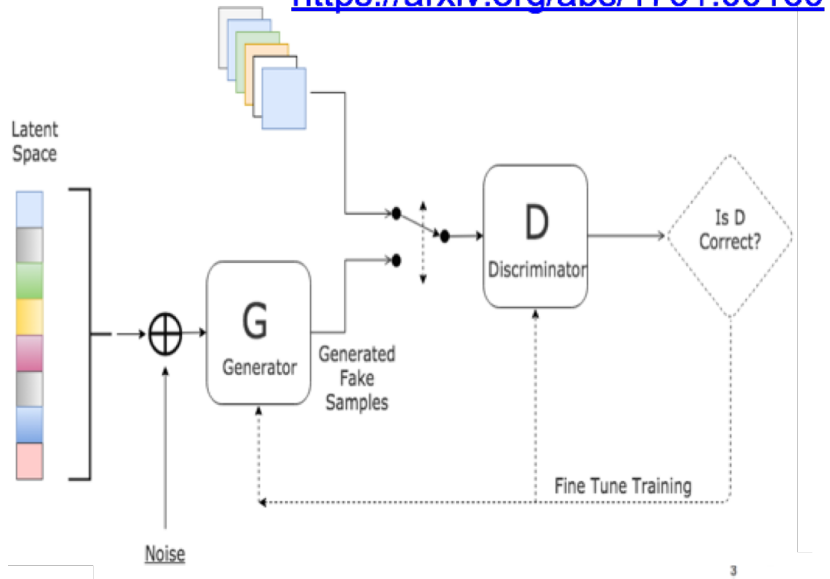


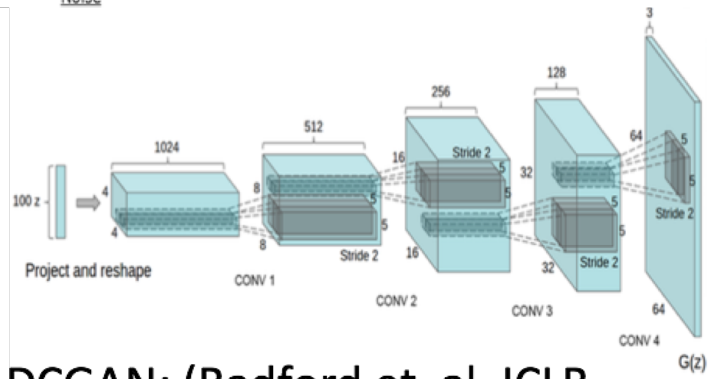
- **Challenge and Importance**: Many DOE simulation efforts could benefit from having realistic surrogate models in place of computationally expensive simulations. These can be used to quickly flesh out parameter space, help with real-time decision making and experimental design, and determine the best areas to perform additional simulations. We are targeting large-scale structure simulations of the universe. As the field is well developed, the scale can easily be ramped up to an exascale ML challenge, and the field is robust enough to explore systematics at the sub-percent level.
- **ML impact**: Neural-networks-based generative models can make reliable surrogate models of expensive simulations for data augmentation purposes. Such surrogate models can be used to aid in cosmological analysis to reduce systematic uncertainties in observations.
- **Timeliness**: The ExaSky application project is producing the largest LSS simulations now, the DESI experiment starts next year, and LSST takes its first science images in 2021.
- **Urgency**: All cosmological measurements today are limited by systematics, not statistics. To reduce these uncertainties and make the most of these future experiments, thousands (if not millions) of exascale-sized simulations will need to be carried out. Surrogate models are a viable path forward to achieve this goal—but only if their limitations are fully understood.
- **Benefit to ECP-Large DOE Experiments**: Once demonstrated, this software framework can be easily adapted to other fields and simulation areas, such as combustion.

Deep Learning to the Rescue?

- Jointly optimize Discriminator (D) and Generator (G) NNs
  - ✦G architecture like decoder in ConvAE
  - ✦Loss for G/D in opposition
- On 'natural images' GANs can be unstable, our problems have advantages:
  - ✦underlying physics structure
  - ✦existing, labeled simulation samples
  - ✦metrics to evaluate
- Build on industry research – e.g. convolutional DCGAN

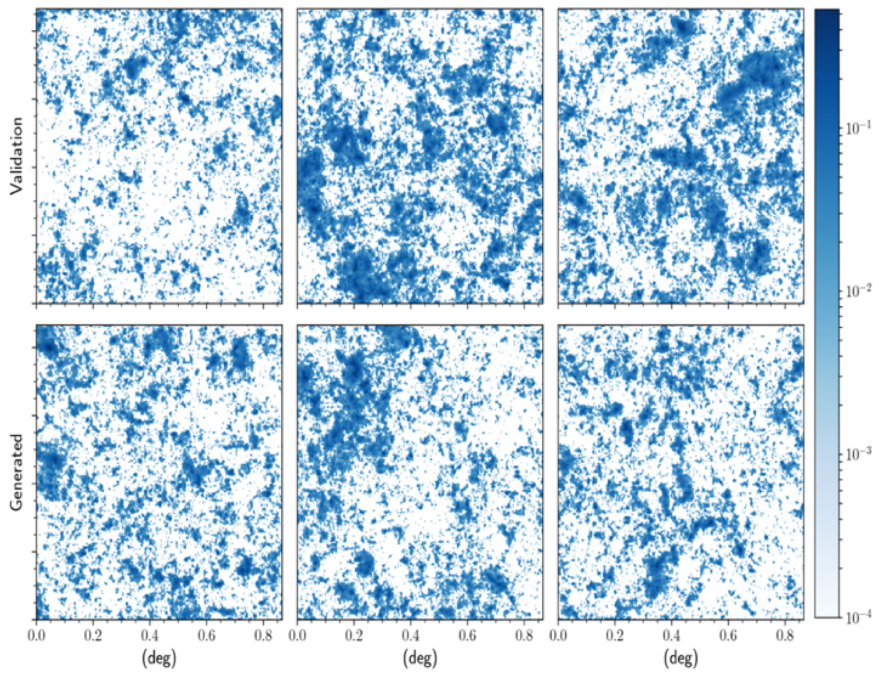
Goodfellow *et al.*
https://arxiv.org/abs/1406.2661
https://arxiv.org/abs/1701.00160

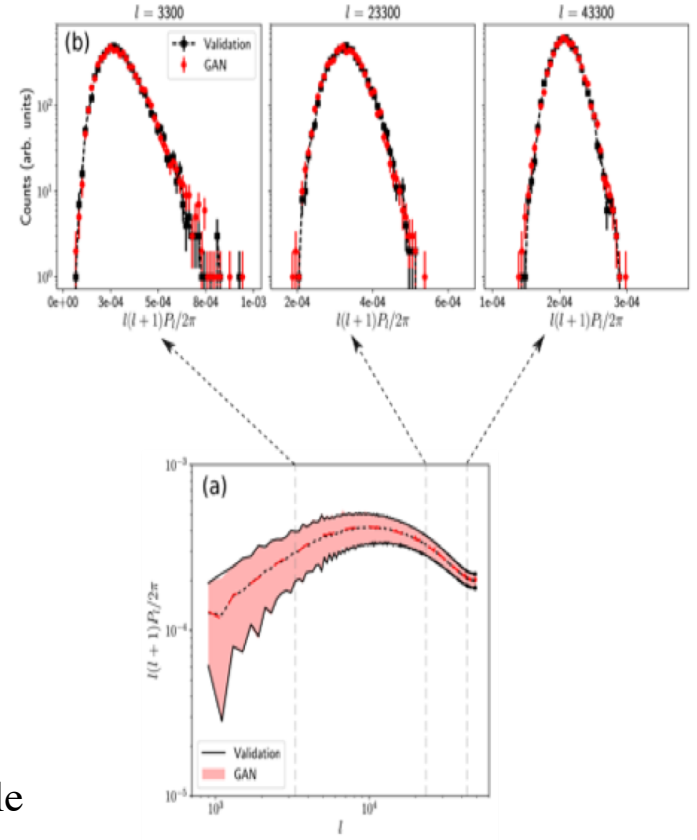




DCGAN: (Radford et. al. ICLR 2016)https://arxiv.org/abs/1511.06434

CosmoGAN

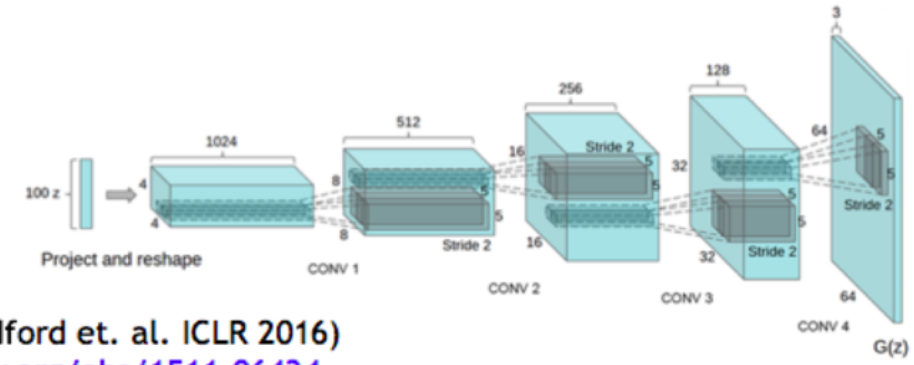




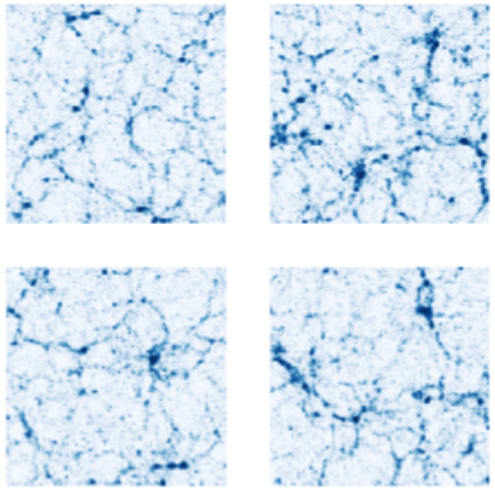
- Calculate power spectrum for generated images and validation sample
- Excellent agreement (K-S p_value > 0.995 for 246/248 moments)
- GAN not explicitly trained to reproduce these distributions
- *Also higher-order Minkowski functionals are reproduced*

GANs for cosmology

- Building on success of CosmoGAN "1.0" (weak lensing convergence maps), use the widely-used DCGAN network architecture
  - Simple CNN setup works well for scaling up network size and parallelizing the model for large inputs



DCGAN: (Radford et. al. ICLR 2016)
https://arxiv.org/abs/1511.06434

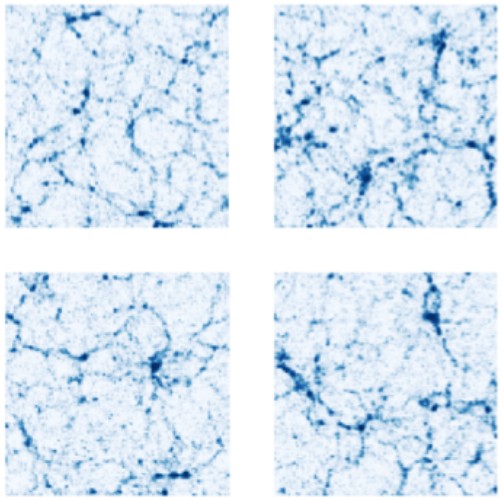


Looking beyond the good press…

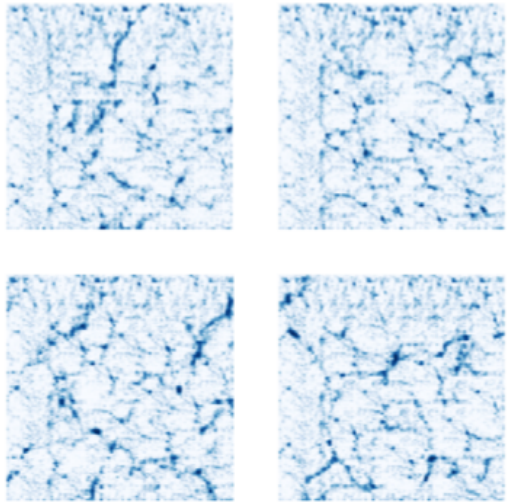
GANs for cosmology: pitfalls & challenges

- GANs can achieve high sample quality but are notoriously brittle
  - As G gets closer to the real data manifold, gradients from D are unbounded
  - Loss functions do not strongly correlate with sample quality, **mode collapse** is common

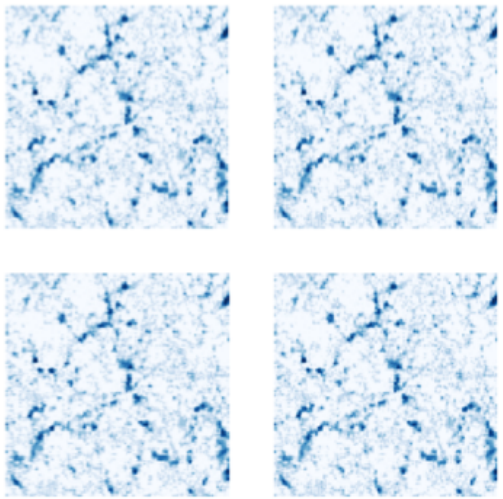
Real samples



Partial mode collapse



Complete mode collapse



Making things Stable

GAN regularizers to improve stability

Mitigate instability by **adding regularization terms** to the objective functions

From "non-science DL" literature (dataset agnostic):

- Gradient penalties ("R1 regularization") $R_1(\psi) := \frac{\gamma}{2} \mathrm{E}_{p_{\mathcal{D}}(x)} \left[\|\nabla D_\psi(x)\|^2 \right]$
  - DL theory: R1 stabilizes training dynamics close to Nash equilibrium
- Feature matching $||\mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}} \mathbf{f}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{z}}(\boldsymbol{z})} \mathbf{f}(G(\boldsymbol{z}))||_2^2$
  - Penalize generator to match the statistics of the intermediate feature maps in the discriminator

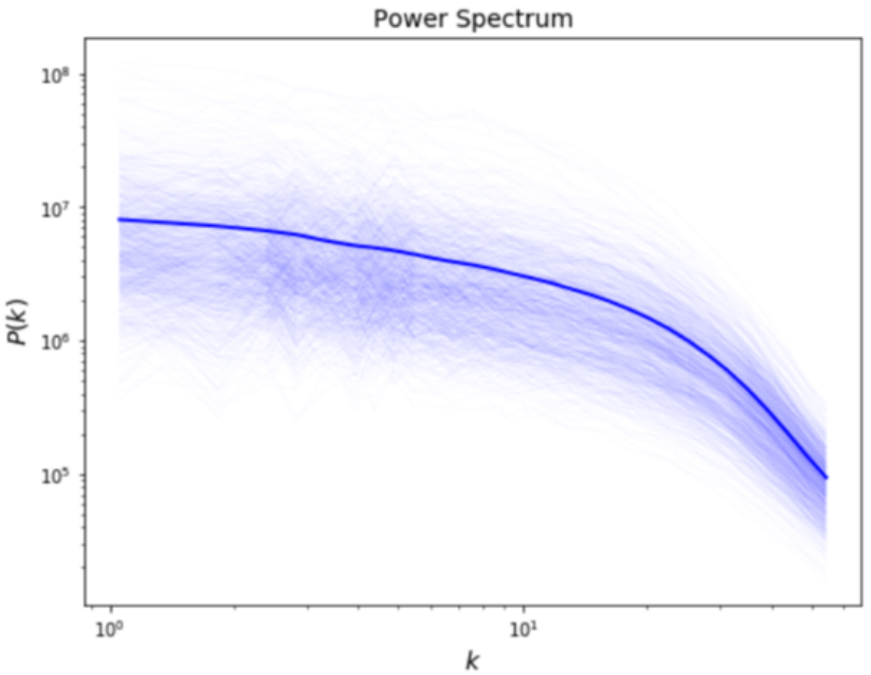
Adding Physics to the GANs

Physically-motivated constraints

- Additional loss term(s) to push generator towards generating samples with a **realistic power spectrum** *P(k)*, one of the target statistics
  - Define $\mathcal{L}_{\text{spec}}$ such that the mean and variance of *P(k)*, per *k* bin, matches expected distribution

$$\mathcal{L}_{\text{spec}} = \log ||Q(P_{\text{generated}}(k)) - Q(P_{\text{target}}(k))||_2^2$$

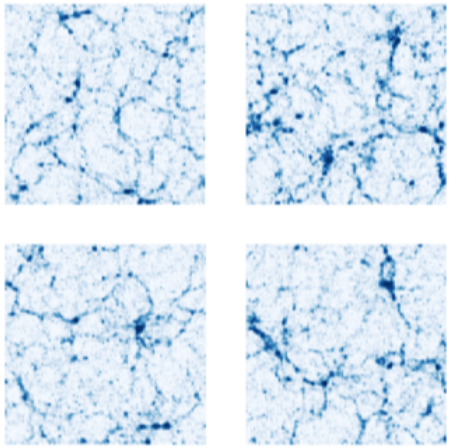
  - Backprop through power spectrum computation (2D FFTs + binning over $|k|$) so generator gets useful gradients

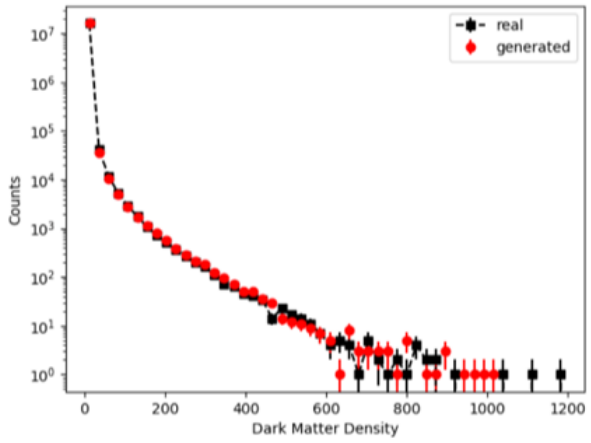


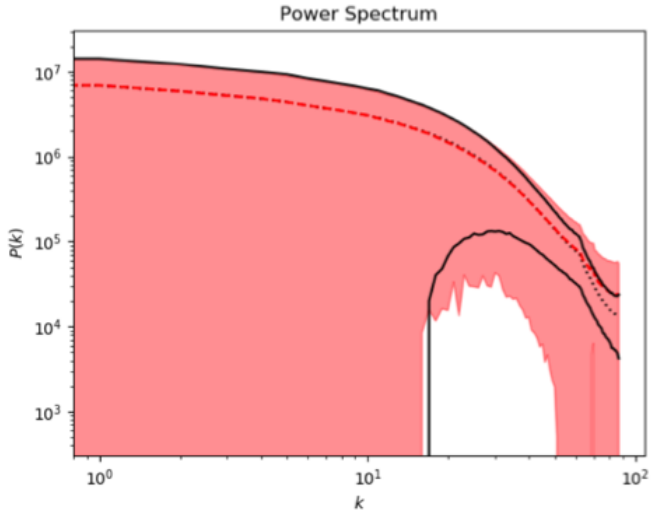
Current 2-D State-of-the-Art

Results: ML regularizers + *P(k)* constraint

- With constraints in place, generated samples tightly match target distributions for summary statistics (mass density histogram, power spectrum)



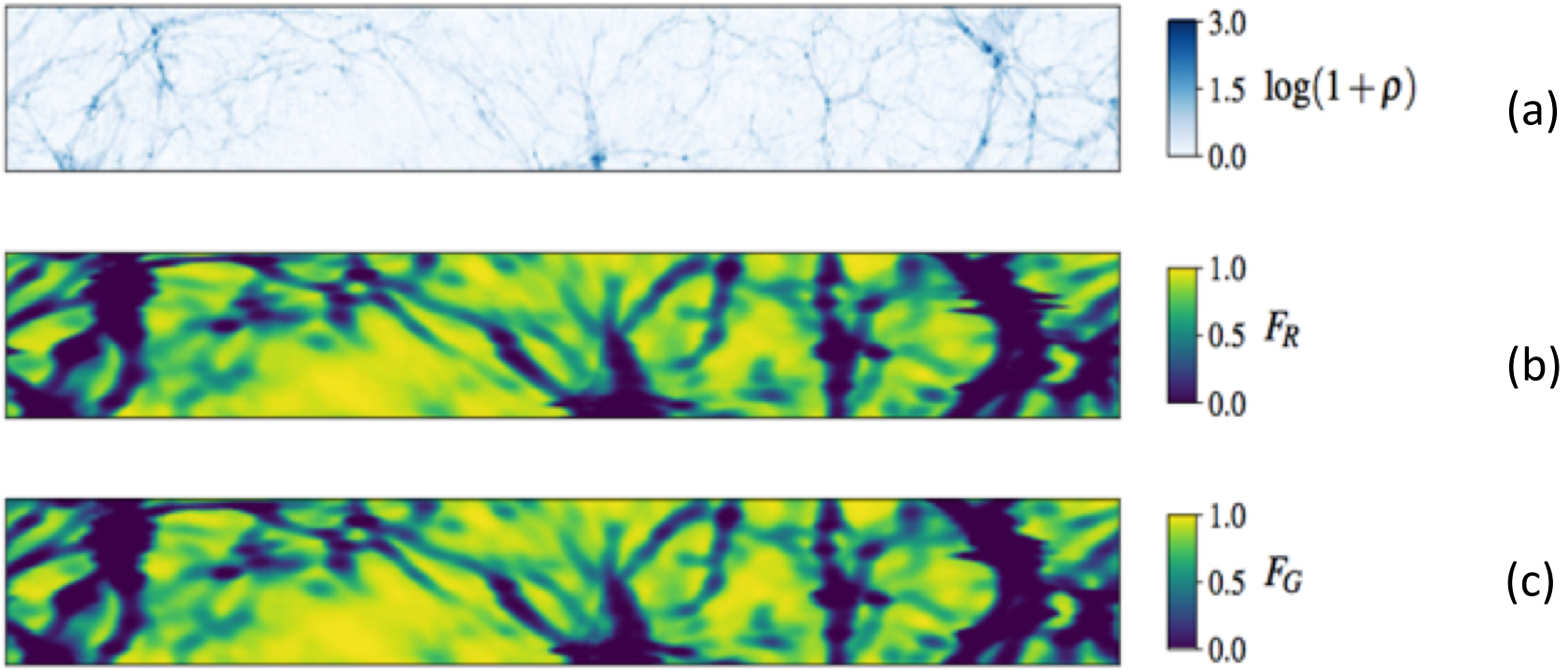




- Now, onto **scaling up** these techniques for larger sample sizes, in 3D, etc

LBANN to the rescue!

VAE for simple to full-physics models to observations

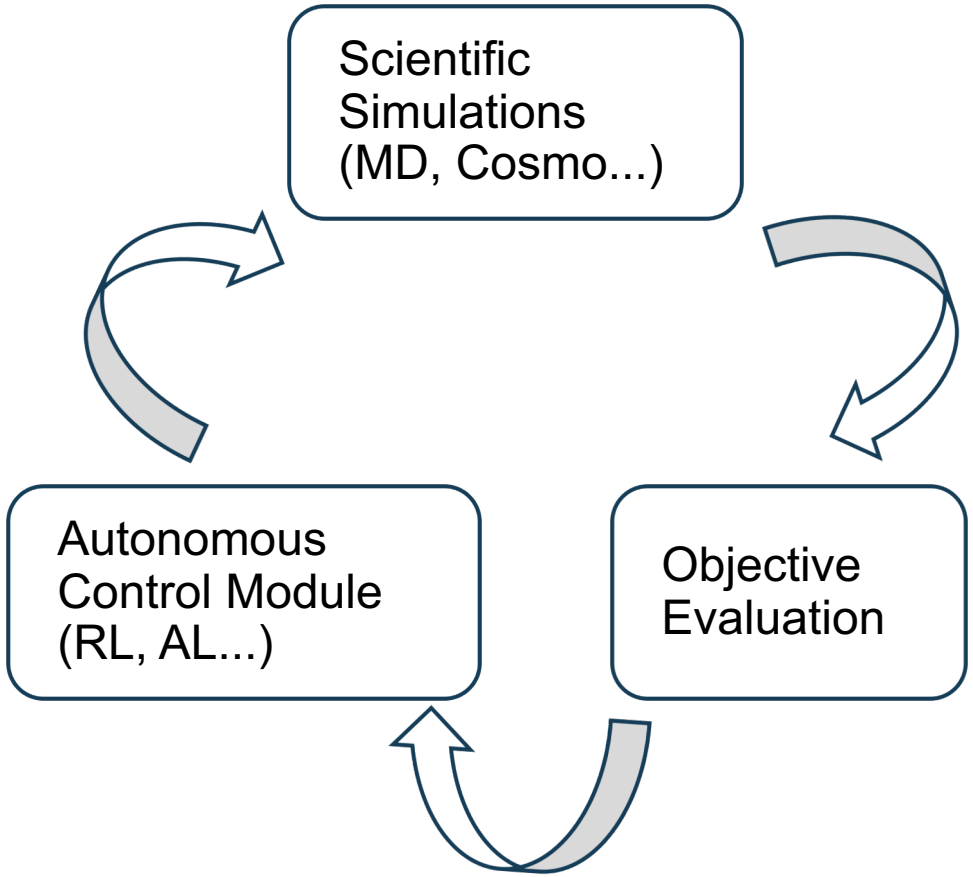


Visualization of the pipeline output. A 3D dark matter distribution (of which a 2D slice is shown in panel (a)) is the principal input to the workflow, which tries to produce the corresponding Ly-alpha flux field FR (b). The prediction FG is shown in panel (c). Generally, structures at both large and small scales, as well as the distortions that warp them in redshift space, are captured well.

Nyx with only gravity and particle and Nyx with full hydro and gas physics.

Objective-Driven Experimental Design

- **Motivation**: At large scales, simulation “efficiency” runs risk of being lower without careful steering. Or “wrong” “suboptimal” calculations may be performed without careful analysis of results generated.
- **Definition**: ODED is an ML-enabled autonomous system to design and execute computations.
- **End Results**:
  - To reach the objective earlier (with less computation)
  - To produce a better result (within the computation budget)
- **How It works**:
  - Iterative learning, simulation, and objective evaluation
  - Autonomous Control Module could be Reinforcement Learning, Active Learning, or other types of modules.



Future Vision: Integrate the Four Machine Learning Types in ExaLearn for a Single Application

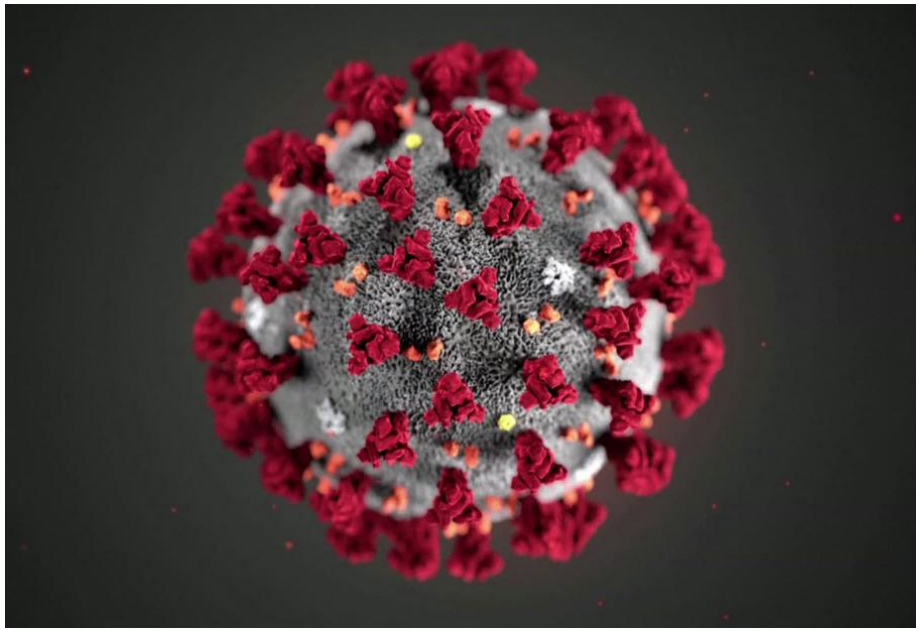
Example: Tokamak Plasma Fusion

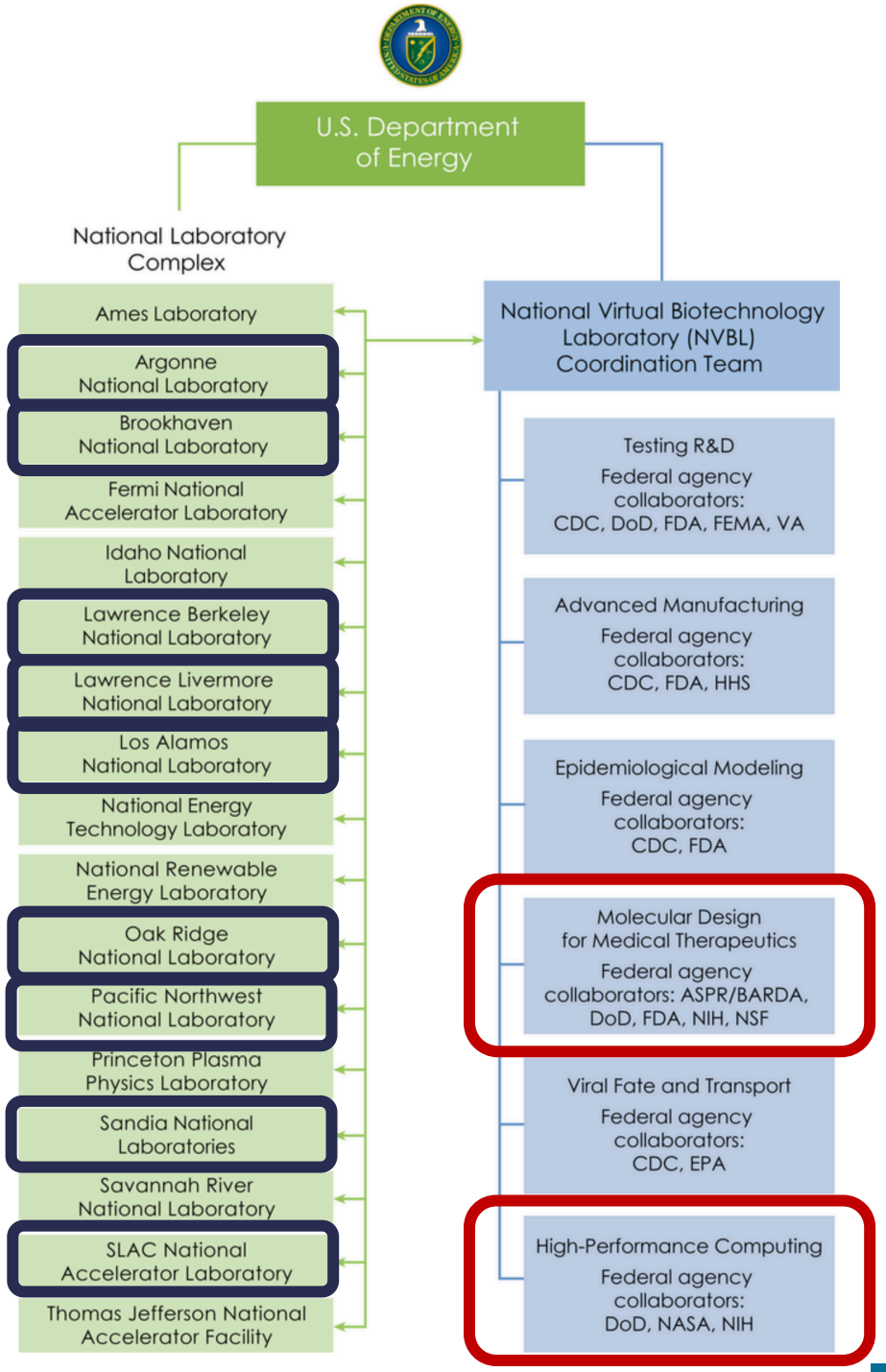
- Generate goal-driven surrogate models for dynamic processes to replace expensive whole device simulations (WDM)
- Use these surrogates to generate training data for a RL-based real-time controller
- Apply pipeline from the inverse problems pillar to predictions plasma equilibria configurations in tokamaks and stellarators
- Apply tools from the design pillar to optimize tokamak design and control policy

- EXAWIND: Exascale Predictive Wind Plant Flow Physics Modeling
- Combustion-Pele: Transforming Combustion Science and Technology with Exascale Simulations
- ExaSMR: Coupled Monte Carlo Neutronics and Fluid Flow Simulation of Small Modular Reactors
- MFIX-Exa: Performance Prediction of Multiphase Energy Conversion Device
- WDMApp: High-Fidelity Whole Device Modeling of Magnetically Confined Fusion Plasmas
- WarpX: Exascale Modeling of Advanced Particle Accelerators

National Virtual Biotechnology Lab (NVBL)

- Aid U.S. policymakers in responding to the COVID-19 pandemic with epidemiological information for decision making
- Accelerate production of critical medical supplies across the nation
- Supercomputing and artificial intelligence for design of targeted therapeutics
- Leverage chemical testing, analysis and biology within DOE to facilitate new antigen and antibody testing





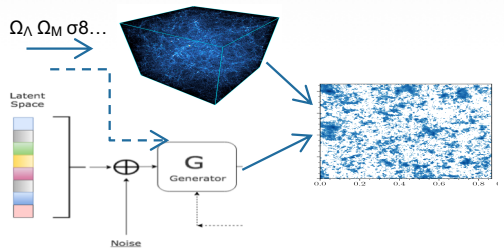
Nine Laboratory Collaboration on HPC and AI for Molecular Design

- Targeting > 12 Virus Proteins and 3 Human Proteins implicated in Virus entry/replication
- ECP Team Created a Library of > 4B molecules prepared for docking and machine learning and made available to the research community [arXiv:2006.02431]
- Over 60 Receptor models (models of binding sites) developed to drive docking Hybrid Docking/ML protocol used to virtually screen more than 4 Billion compounds for each target
- Leveraging structures determined at Argonne's APS
- Targets aiming at blocking viral entry, viral replication, virus maturation
- Thousands of compounds identified that score higher that the best drug repurposing candidates
- 1200 top hit molecules for Top 7 drug targets are in various stages of being assayed
- This work has inspired new AI research integrated reinforcement learning and physics-based modeling

ExaLearn Application Pillars: Q1: Original Target Apps and Pivot to COVID-19

Surrogates

- ML-created models
- Faster and/or higher fidelity models
- Generative networks
- Use ML to replace complicated physics, learned low fidelity to high fidelity mapping,
- Cosmology
- Fast, Accurate Surrogates to replace large-scale epidemiological simulations



Control

- ML-controlled experiments
- Efficient exploration of complex action space
- Reinforcement Learning
- Use RL agent to control light source experiments
- Temperature control for phase ordering dynamics:
- Using RL to design optimal non-pharmaceutical interventions

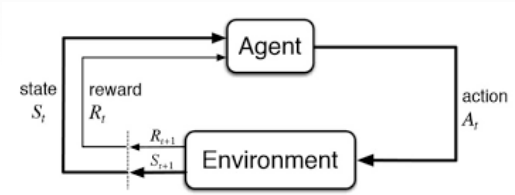
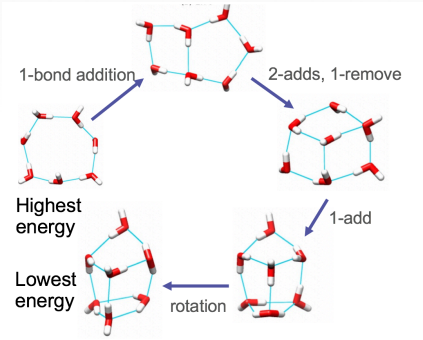


Image courtesy Sutton, Barto, Reinforcement Learning 2017

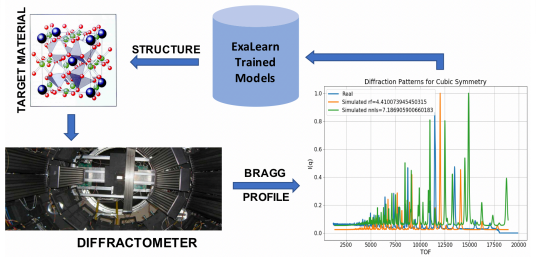
Design

- ML-created physical structures
- Optimized proposal for desired behavior of structure within complex design space
- Graph-Convnets
- Use Graph NN's and Deep RL to propose new structures that respect chemistry
- Molecular Generator and Design for COVID-19 Therapeutics



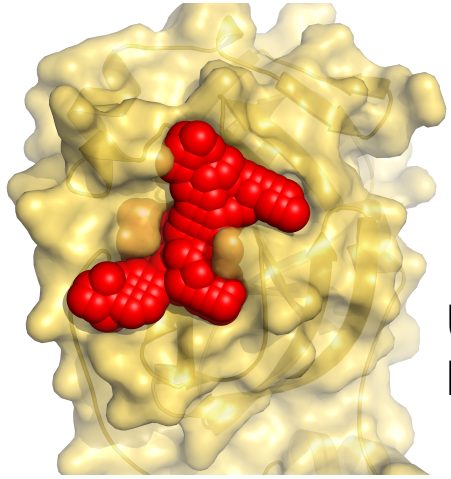
Inverse

- ML projection from observation to original form
- Back-out complex input structure from observed data
- Regression models
- Predicting crystal structure from light source imaging
- Material structure from neutron scattering
- ML framework for protein structure prediction from SNS experimental data.



ExaLearn: Scaling up training of molecular generator to accelerate the discovery of small molecules for COVID-19

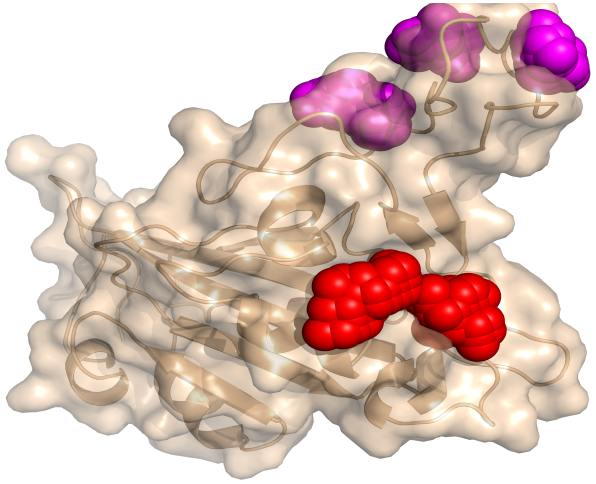
- 4 protein targets from 2 SARS-CoV-2 proteins
- Screening millions of compounds to find drug candidates against SARS-CoV-2 proteins



Main protease

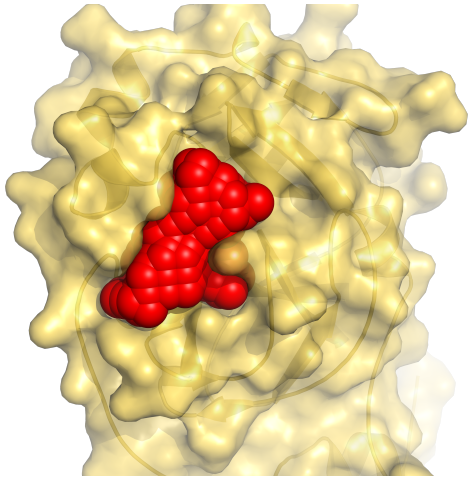
U of Shanghai
PDB ID: 6LU7

Viral spike protein
bound with
human ACE2



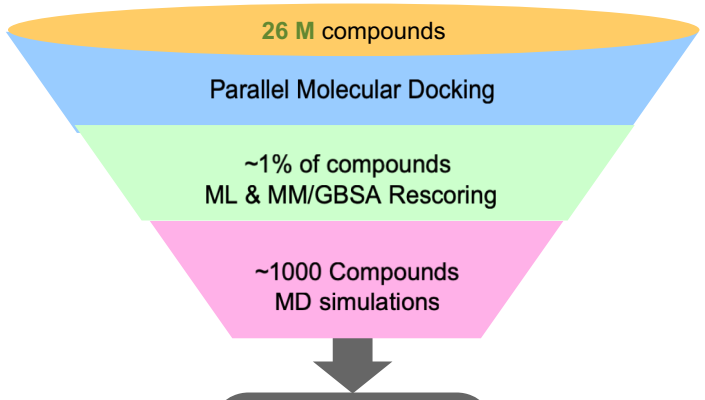
Stops viral entry

PDB ID: 6Y84
Diamond, UK
Pre-publication



Stops viral replication

ATOM design tools will be applied to propose new improved molecular structures



Lead optimization

Machine Learning(ML)/Docking Pipeline

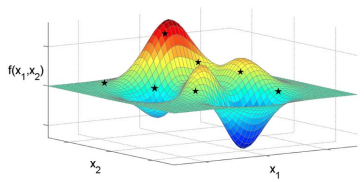
- Efficacy
- Safety
- PK
- Developability

Working Compound Library

Design Criteria

Decoder

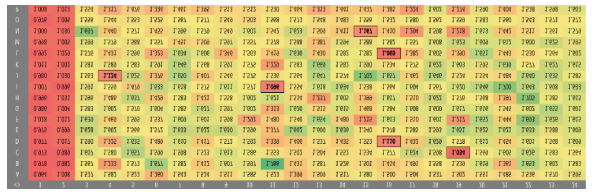
Molecular Optimizer



Encoder

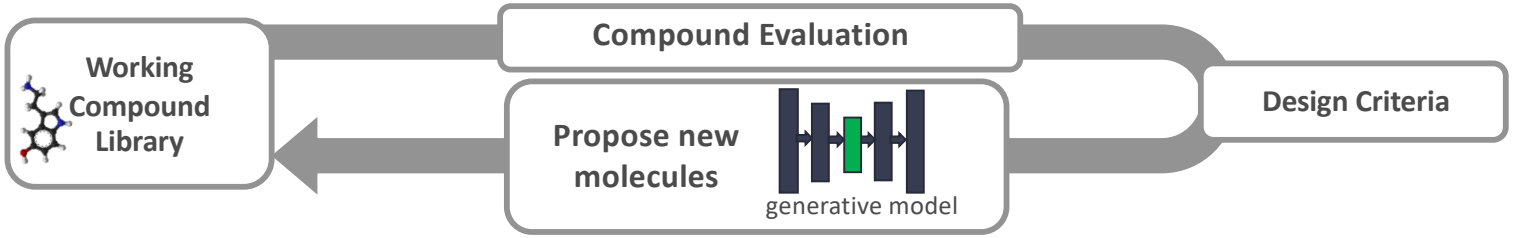
Generative Network Trained on potent molecules

Preparing to order first 100 molecules for experimental testing



Experimental binding assay readout

Developed a novel neural network architecture for generating small molecules

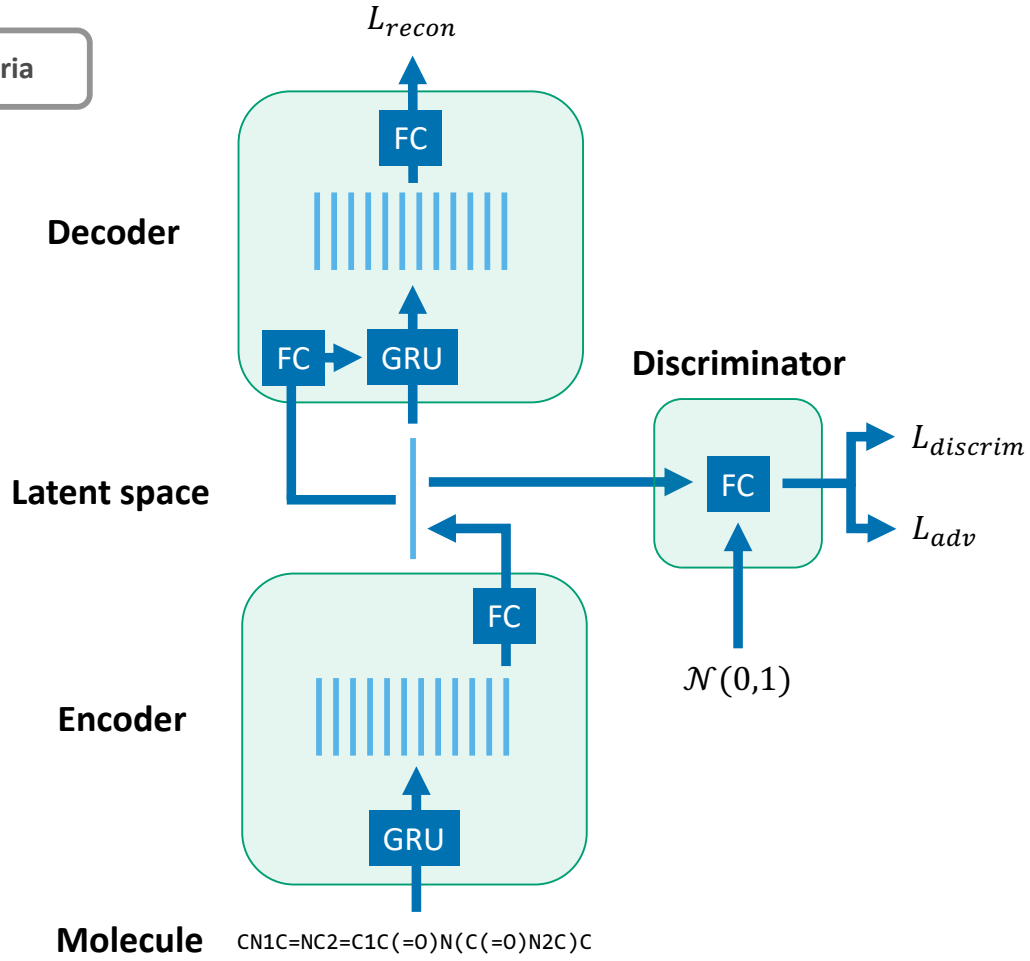


Task:

- Input lead compound as SMILE string
- Encode into latent space
- Move in latent space to optimize for desired properties
- Project from latent space back to new compound (SMILE string)

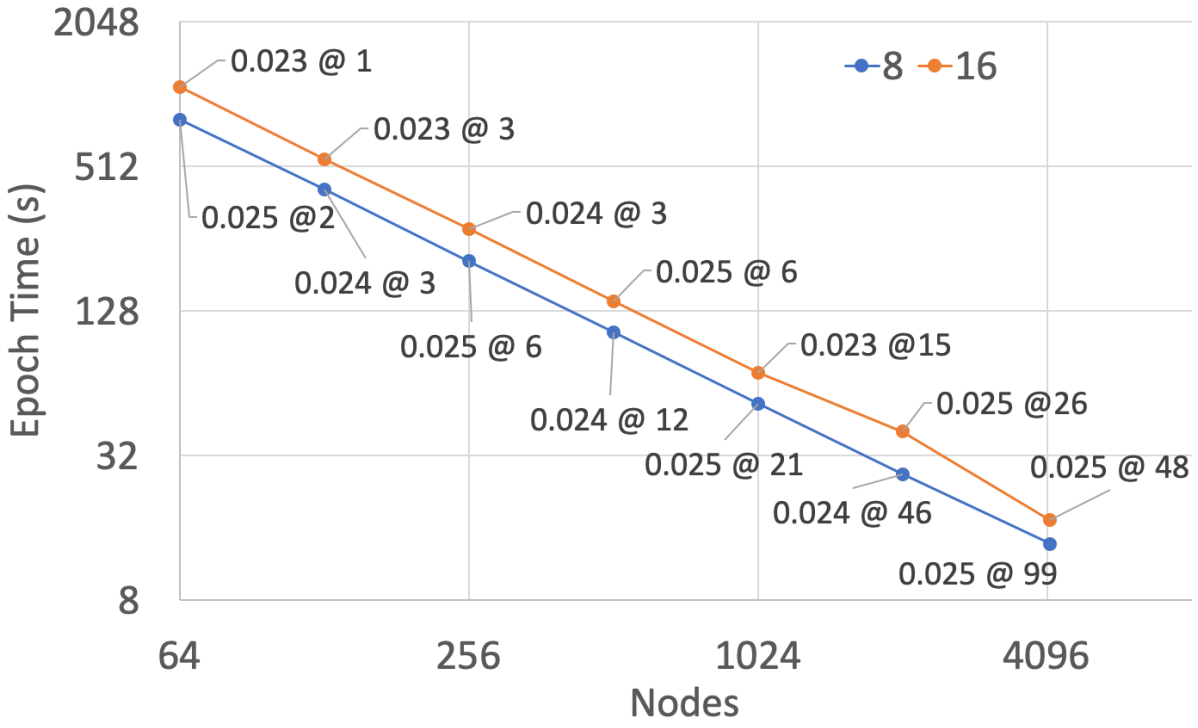
Novel architecture for molecule generation:

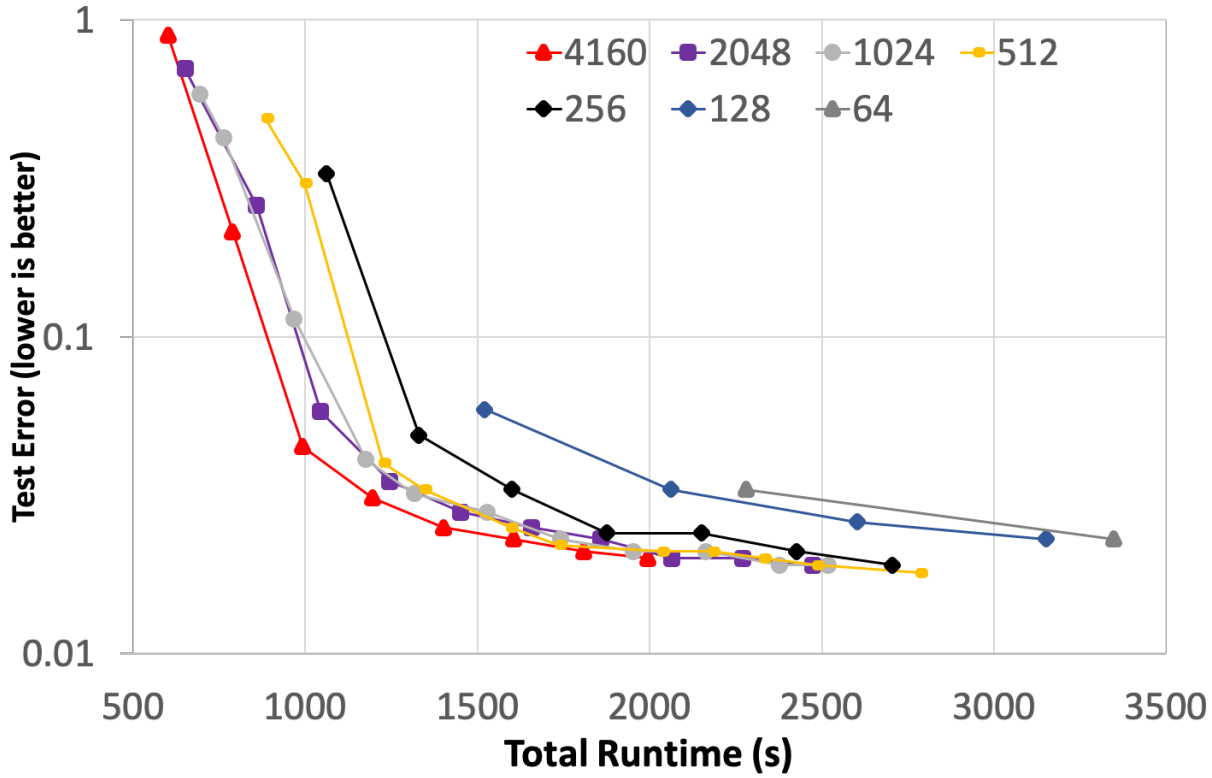
- character-Wasserstein AutoEncoder (cWAE)



LBANN-enabled training of models at scales previously unobtainable

- Optimized data ingestion scaled training to 1.613B compound data set
- Asynchronous LTFB algorithm enables scaling without loss of model quality





ExaLearn investments enabled Gordon Bell submission for HPC deep learning training at scale for COVID-19 research

- Capabilities developed in ExaLearn will enable future research for deep learning at scale.
  - Model architecture search and hyperparameter exploration
  - Rapid model retraining on new data sets

GPUs/trainer	Trainers	Epoch time	PFLOPS
16	1040	17.2 s	**253.3**
8	2080	13.7 s	**318.0**

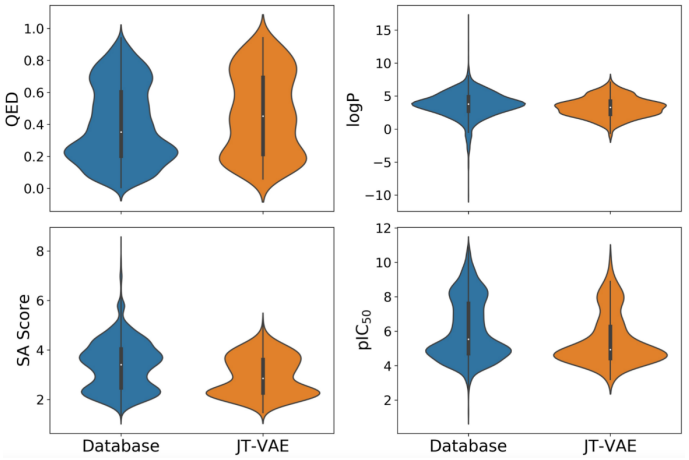
Table 6. Peak performance training with 4,160 nodes on Sierra.

- Deep Learning at scale has center-wide impact ➔ half-precision TensorCores lead to dramatic power swings
  - Periodic 2-3 MW swings caused concern from power company – frequent 200 KW swings cause center concerns.
  - Total of ~266,240 node hours over one weekend without LBANN software fault
  - Asynchronous learning algorithm minimized center-wide power swings

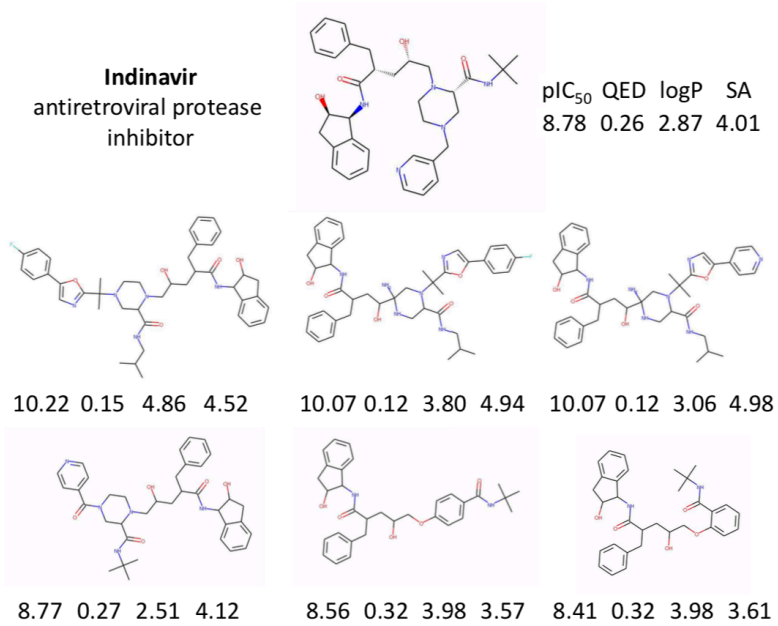
Scientific impact

- Our COVID-19 effort led us to two major findings:
  - **What is the best tool for your need?**: How do different generative models compare for different drug discovery tasks (repurposing vs. new drug)?
  - **Going beyond "Cool"**: AI systems generate too many molecules, more than we can validate. Where do these modeling efforts need to advance for AI-driven molecule design to be trustworthy?

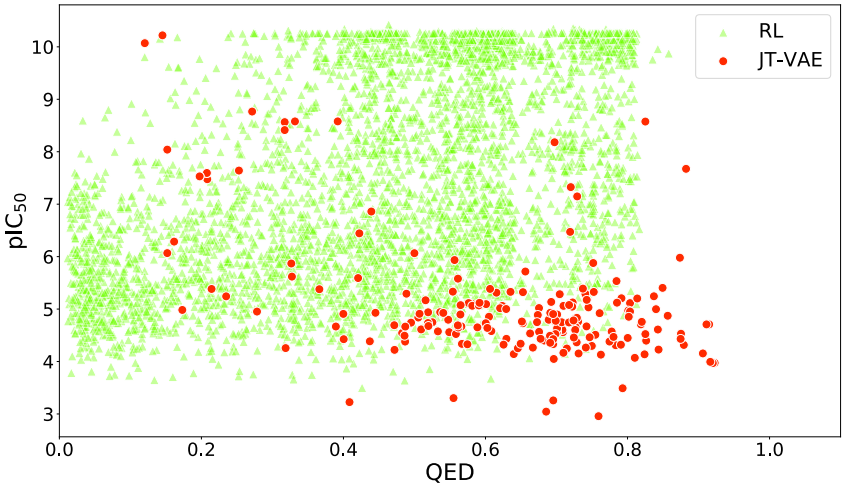
Comparison of VAE vs. Deep RL



If you want a molecule that is close to ones existing in your database, use VAE.



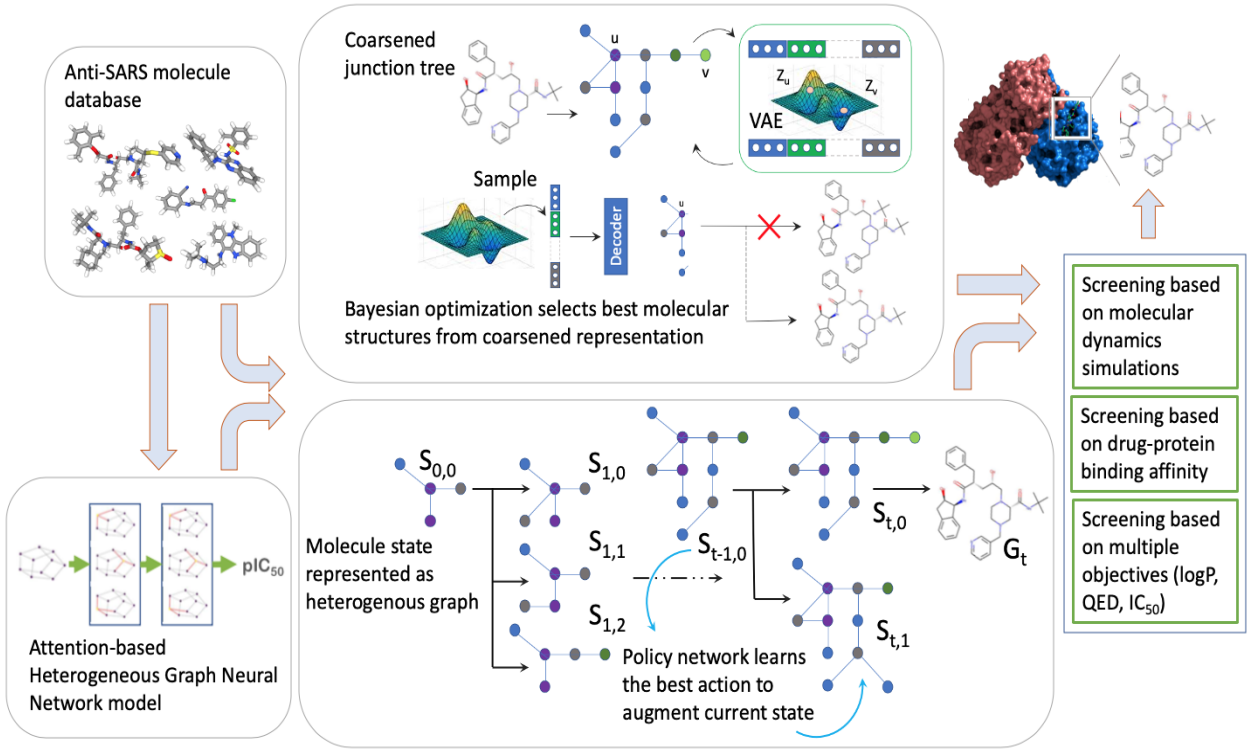
One of our top molecules (generated by JT-VAE) was a match to a widely researched COVID-19 therapeutic.



Deep RL is a promising approach to find novel candidates that we will miss if just "searching where the light is."

Molecular design accomplishment

- Generate new molecules that could inhibit coronavirus proteins
  - **Objective:** Maximize pIC_{50} model trained on BindingDB data for SARS/MERS
  - **Search space:** >>10^{12} potential molecular materials
- Adapted two state-of-the-art molecule generation approaches for molecule design
- Interfacing with NVBL drug screening team to screen candidates for SARS-CoV-2
- Multi-objective optimization produced multiple drug molecules with high binding affinity with SARS-CoV2-3CL protease



Team: Sutanay Choudhury (PNNL), Logan Ward (ANL), Jenna Pope (PNNL), Neeraj Kumar (PNNL), Ganesh Sivaraman (ANL), Marcus Emory Schwarting (ANL)

These investments have broad impact across ECP and DOE

- Dramatically improved our ability to design robust and diverse small molecules that are optimized for specific design criteria:
  - ExaLearn electrolytes/catalysts design
  - CANDLE and ATOM projects on precision medicine: cancer and COVID-19
  - Candidate molecules to tie into COPA simulations
  - Possible outputs into material design
- **Scalability and Performance Cross-Cut**: Ability to rapid train and retrain neural network architectures using leadership-class computing systems will have broad impact across the field of scientific machine learning.
- **Design Pillar**: The COVID pivot positions us better to produce an AI-driven Molecule Design capability in DOE.
- **Inverse Pillar**: The data set will enable training of machine learning models for determination of key parameters, such as radius of gyration, number/fraction of native contacts, etc., of the molecular structure of a COVID-19 virus.
- **Control Pillar**: Investment led to improvements within the scalability of the EXARL framework.

Questions?