Argonne NATIONAL LABORATORY

MACHINE LEARNING FOR NEXT-GENERATION SEQUENCING AND DRUG RESPONSE PREDICTION

ARVIND RAMANATHAN (ON BEHALF OF THE PILOTS 1, 2 & 3 TEAMS)

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

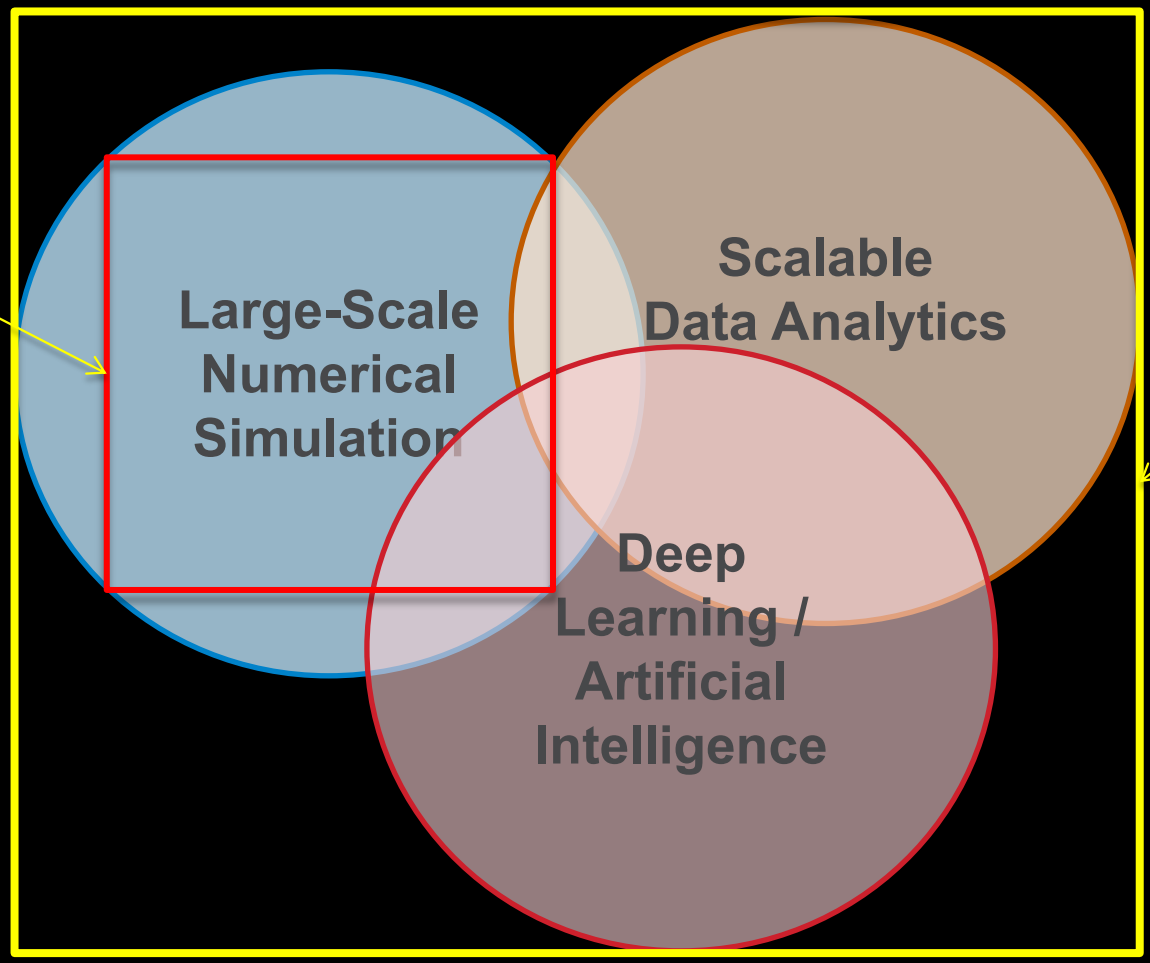
U.S. DEPARTMENT OF ENERGY Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.

Oct 21, 2019

DOE-NCI PARTNERSHIP: ENABLE THE MOST CHALLENGING MACHINE LEARNING PROBLEMS IN CANCER RESEARCH TO RUN ON THE MOST CAPABLE SUPERCOMPUTERS IN THE DOE

CANDLE: Cancer Deep Learning Environment

Traditional HPC Systems



PRIMER ON DOE SUPERCOMPUTING FACILITIES

Argonne NATIONAL LABORATORY

OAK RIDGE National Laboratory

The Largest Concentration of Scientific Computing in One Organization on the Planet

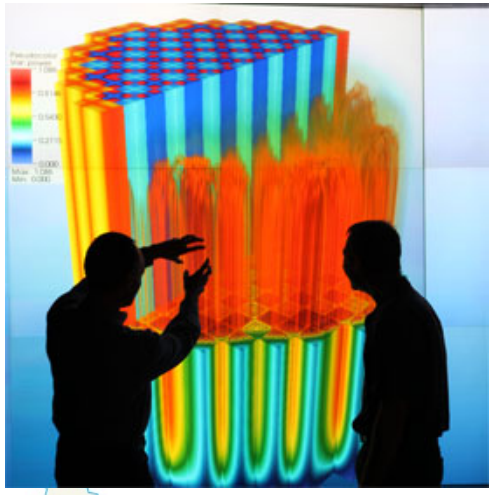
Los Alamos NATIONAL LABORATORY

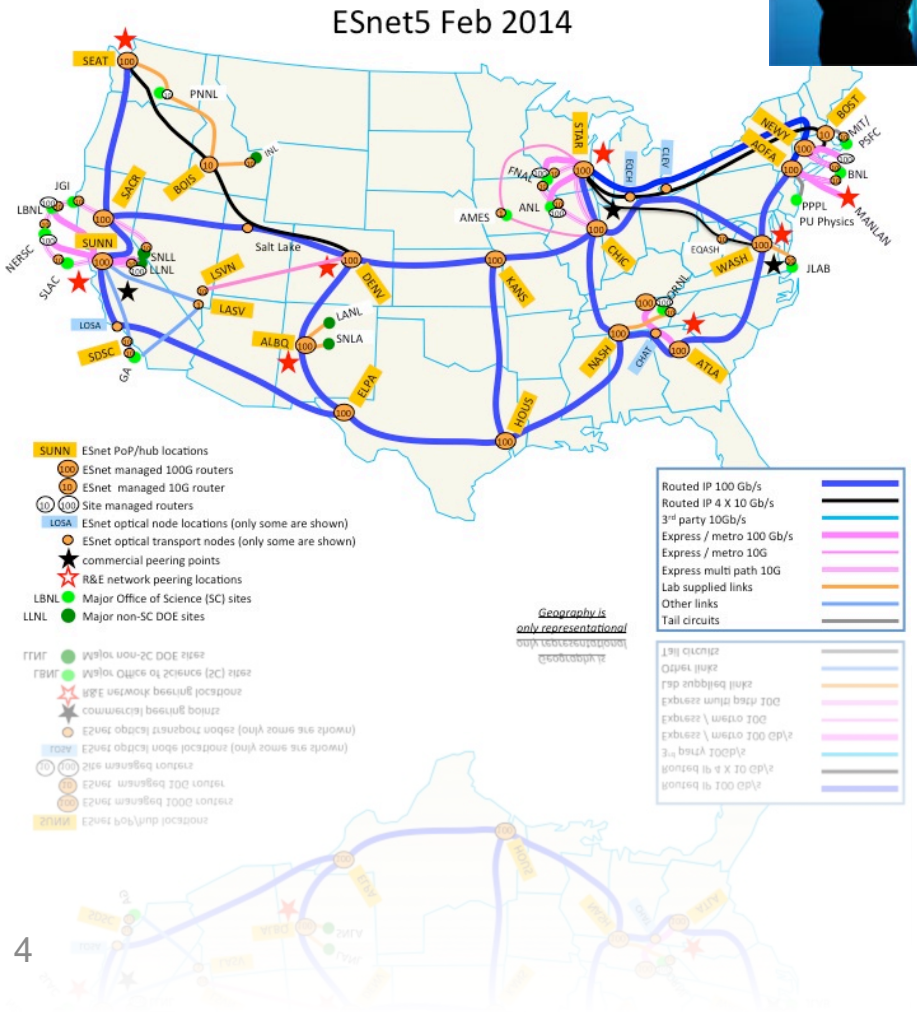
Lawrence Livermore National Laboratory

LEADERSHIP IS NOT ONLY ABOUT COMPUTING

- Hundreds of Petabytes Storage Systems
- Large-scale data analysis and visualization
- World leading network interconnecting facilities

  (100 Gb/s ⟹ 1 Tb/s)
- DOE invests more than $1Billion/yr in the computing capabilities a the laboratories



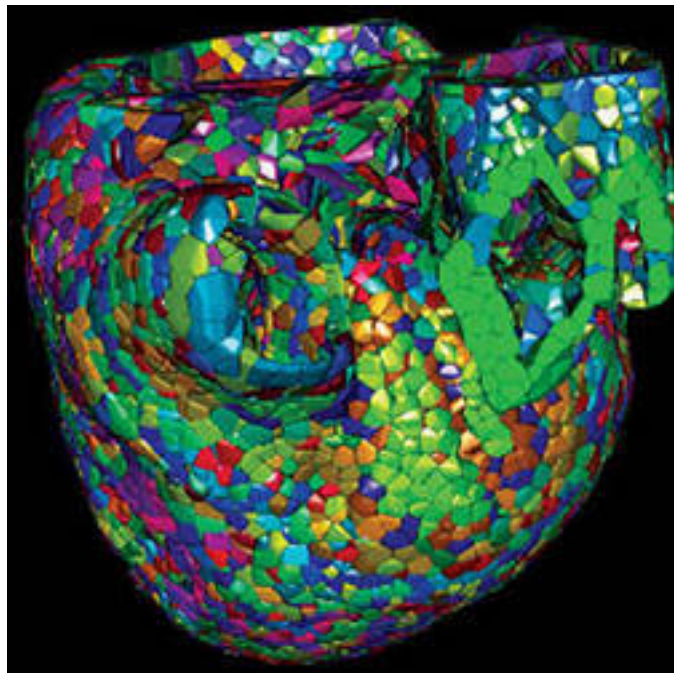


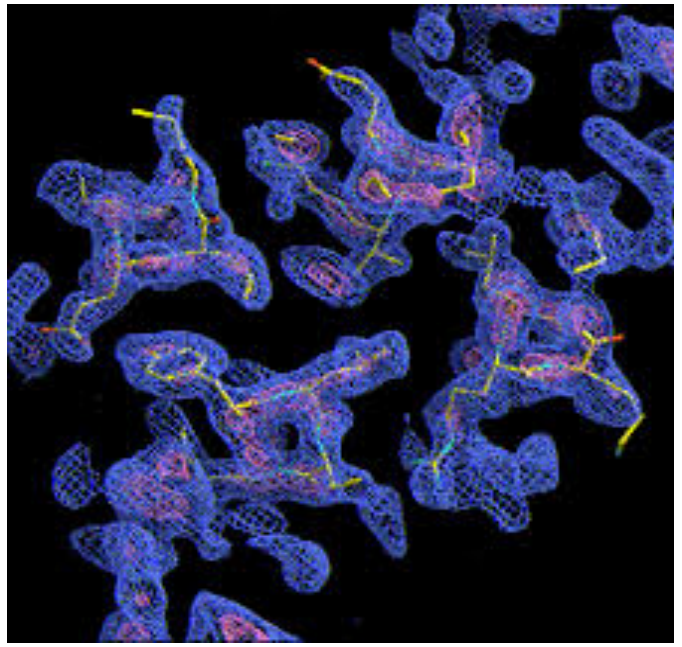
MATHEMATICS AND COMPUTER SCIENCE

- > 1000 Computer Scientists, Mathematicians and Statisticians at the laboratories
- Expertise in
  - Modeling and Simulation of Complex Phenomena
  - Mathematical Techniques
  - Software for Scientific Computing
  - Parallel Computing
  - Machine Learning
  - Data Analysis
  - Data Mining
  - Uncertainty Quantification
  - Verification and Validation
  - Software Engineering

WORLD LEADING COMPUTATIONAL SCIENCE

- Hundreds of Computational “X” Scientists at each Laboratory
- Groups, Codes and Tools Spanning Many Disciplines Relevant to Precision Medicine
- Comparative Genomics and Systems Biology, Biophysics, Microbiology, Proteomics, Mesoscale Modeling, Text and Image Analysis, Data Modeling and Data Integration, Predictive Modeling





OUTLINE

- DOE-NCI Pilot Projects
  - Machine learning projects and their focus areas
  - Machine learning for Next Generation Sequencing and Drug Discovery

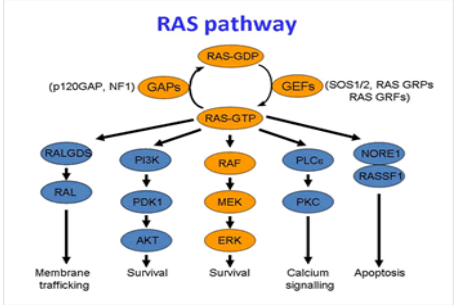
- Overview of example models and their initial results

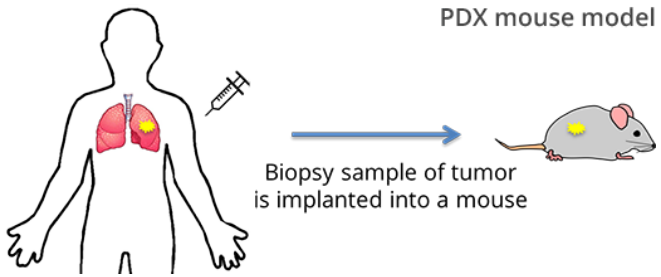
- Overview of CANDLE Technology Stack
  - Example workflows implemented on CANDLE
  - Hyperparameter optimization
  - What this workshop is all about?

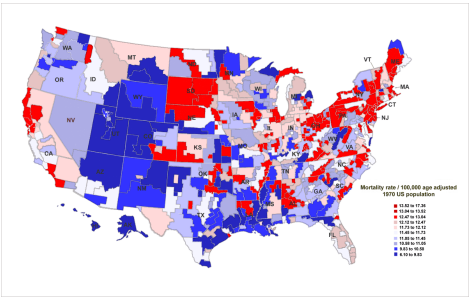
DOE-NCI PILOTS

NCI-DOE PARTNERSHIP WILL EXTEND THE FRONTIERS OF PRECISION ONCOLOGY (THREE PROJECTS)

- Cancer Biology
  - **Molecular Scale Modeling of RAS Pathways**
  - Unsupervised Learning and Mechanistic models
  - Mechanism understanding and Drug Targets
- Pre-clinical Models
  - **Cellular Scale PDX and Cell Lines**
  - ML, Experimental Design, Hybrid Models
  - Prediction of Drug Response
- Cancer Surveillance
  - **Population Scale Analysis**
  - Natural Language and Machine Learning
  - Agent Based Modeling of Cancer Patient Trajectories

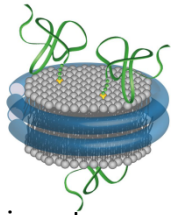




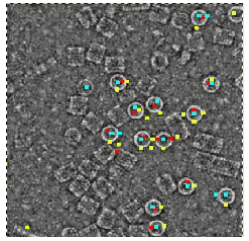


PILOT 2: RAS PROTEINS IN MEMBRANES

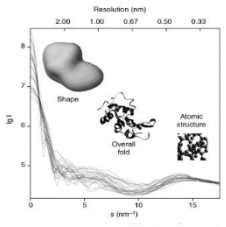
RAS activation experiments at NCI/FNL



Experiments on nanodisc



CryoEM imaging



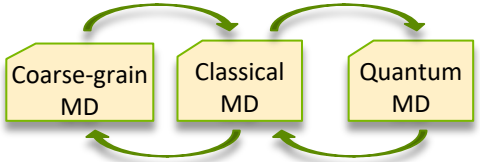
X-ray/neutron scattering

Multi-modal experimental data, image reconstruction, analytics

Protein structure databases

New adaptive sampling molecular dynamics simulation codes

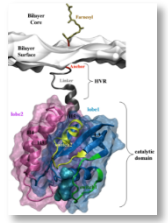
Adaptive time stepping



Adaptive spatial resolution

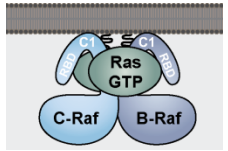
High-fidelity subgrid modeling

Predictive simulation and analysis of RAS activation



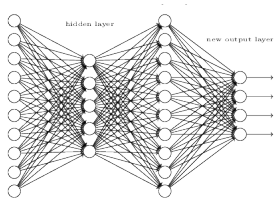
Granular RAS membrane interaction simulations

Atomic resolution sim of RAS-RAF interaction

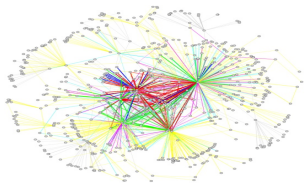


Inhibitor target discovery

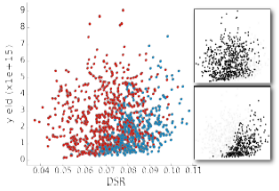
Machine learning guided dynamic validation



Unsupervised deep feature learning

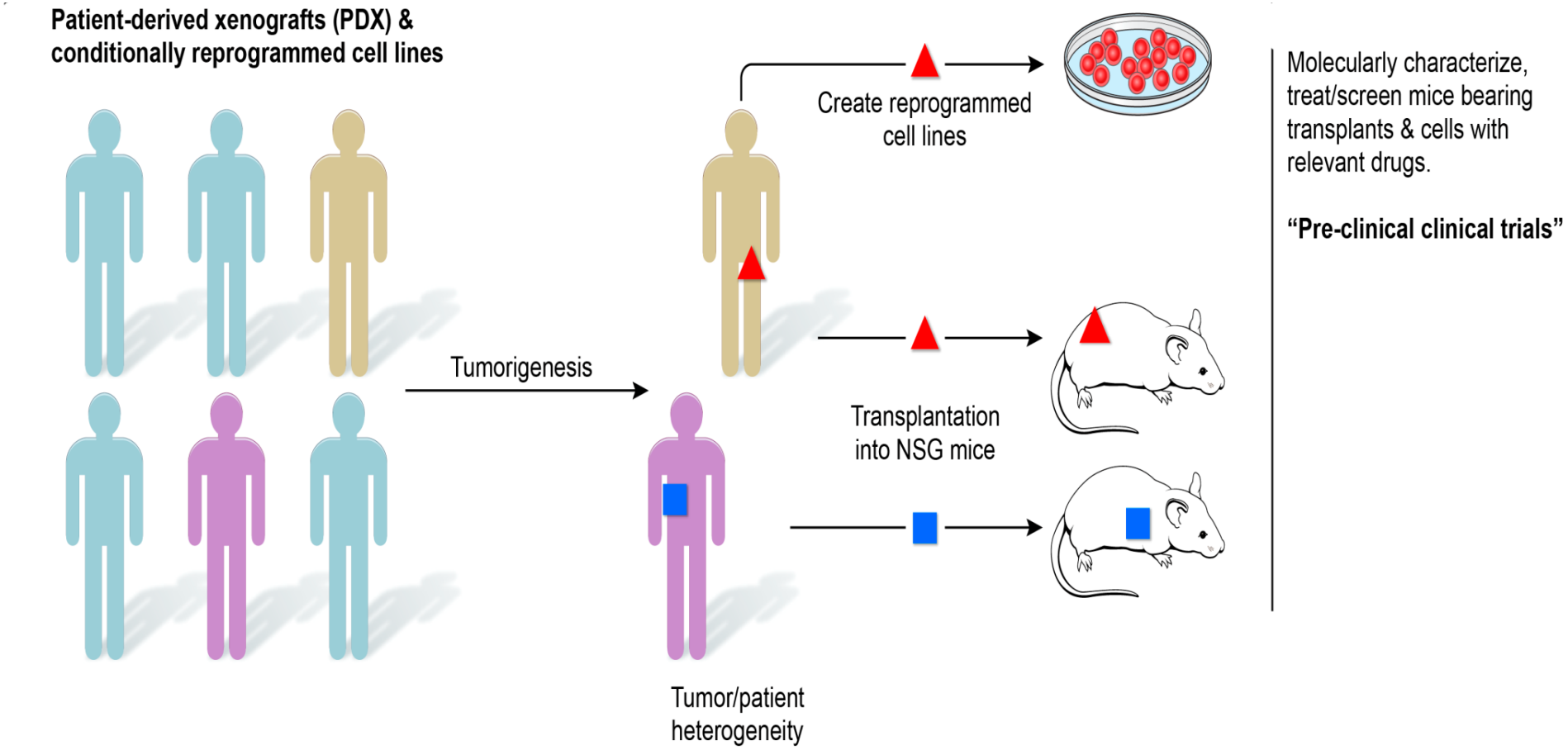


Mechanistic network models



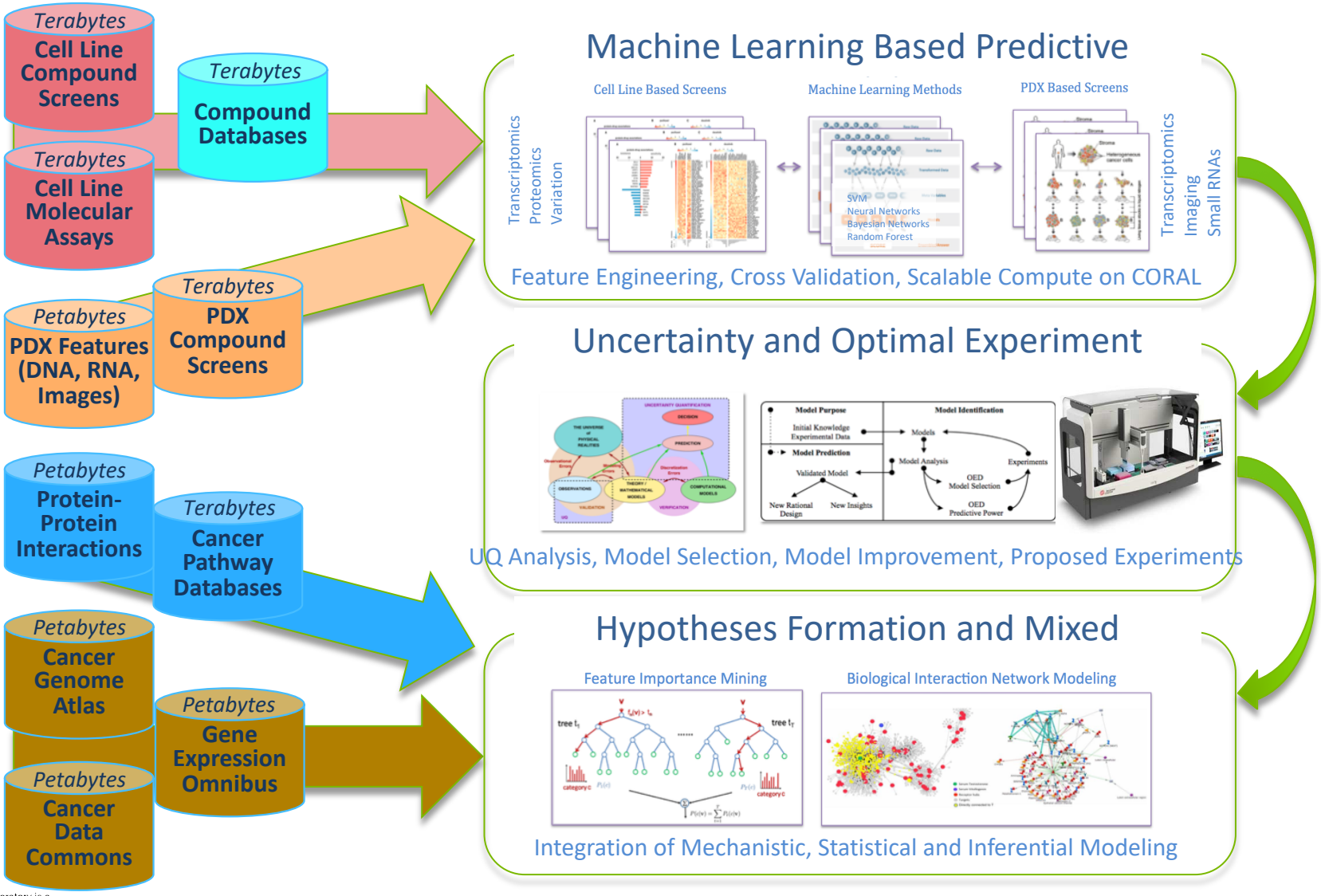
Uncertainty quantification

PILOT 1: PATIENT DERIVED XENOGRAFT MODELS

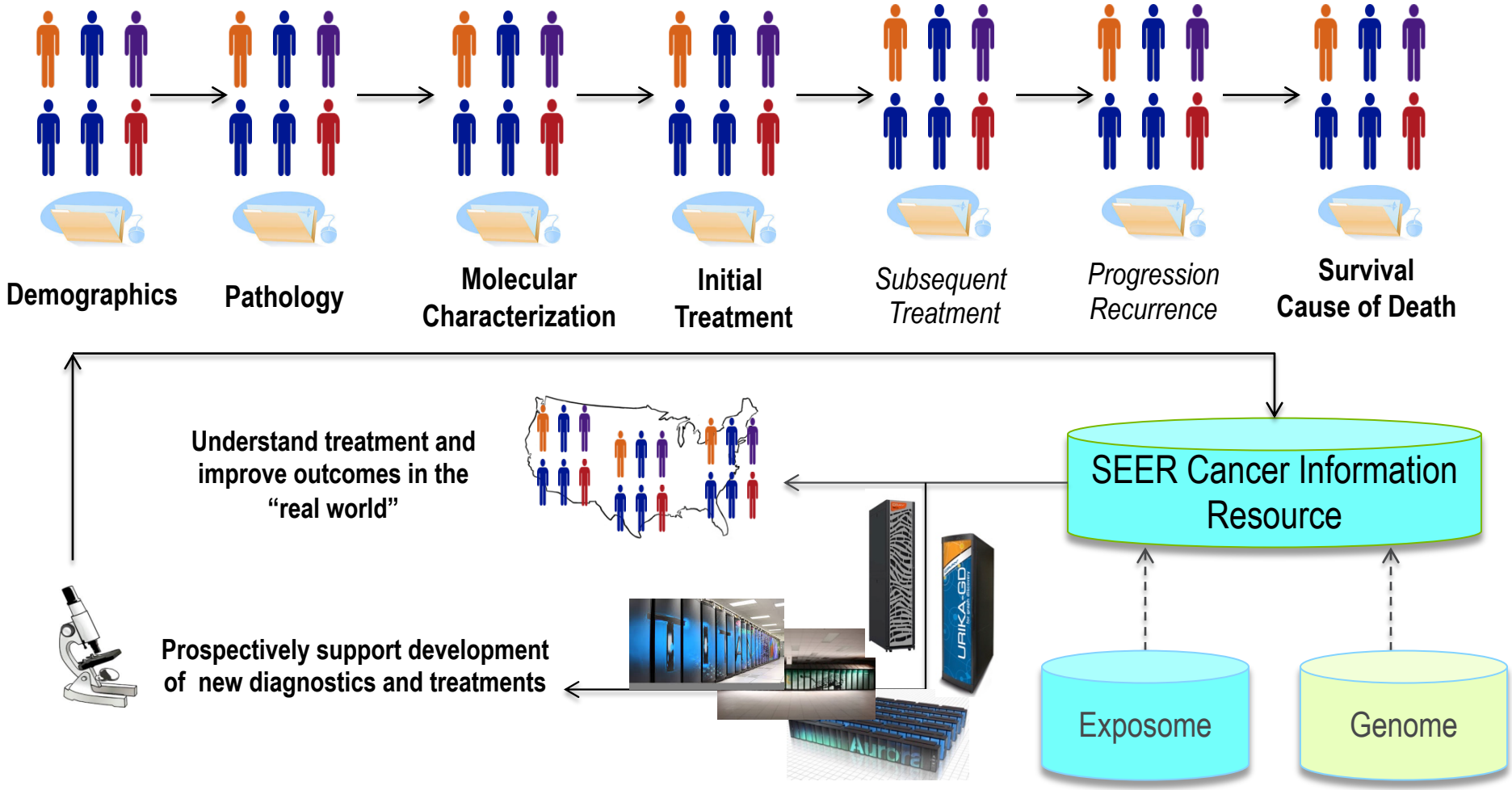


Nature Rev. Clin. Oncol. 11: 649-662, 2014.

PILOT 1: PREDICTIVE MODELS FOR PRE-CLINICAL SCREENING

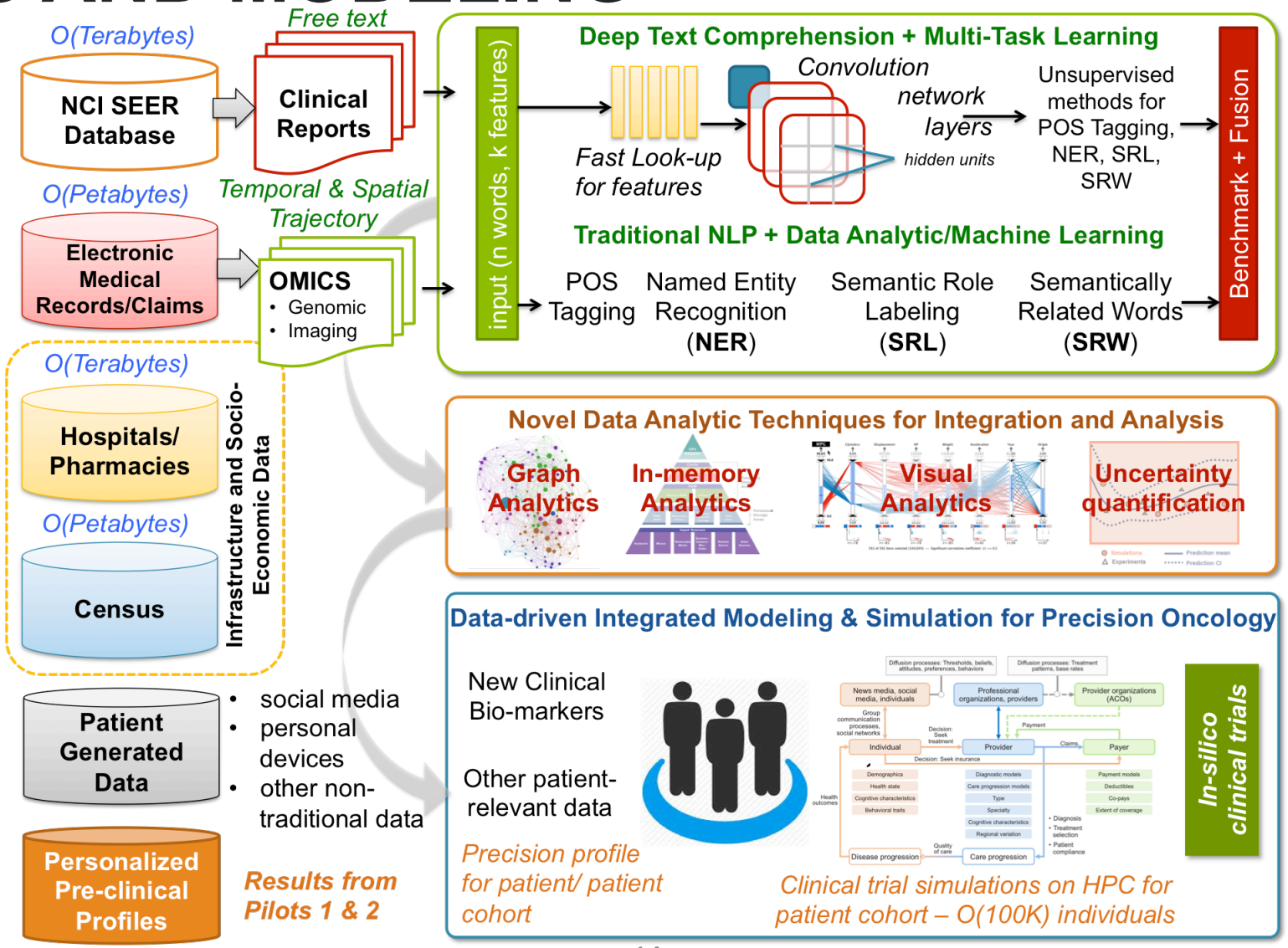


PILOT 3: AI TO SUPPORT NATIONAL CANCER SURVEILLANCE



Improve the effectiveness of cancer treatment in the "real world" through computing

PILOT 3: POPULATION INFORMATION INTEGRATION, ANALYSIS AND MODELING

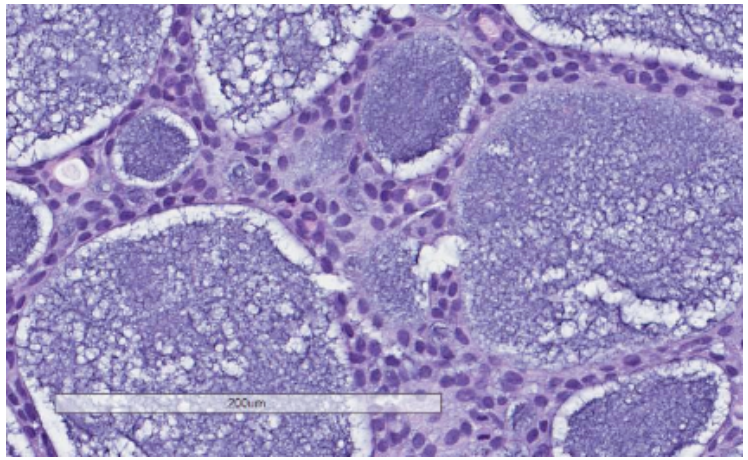


OVERVIEW OF MACHINE LEARNING CHALLENGES IN DOE-NCI PILOTS

PILOT 1: OVERARCHING MODELING GOAL

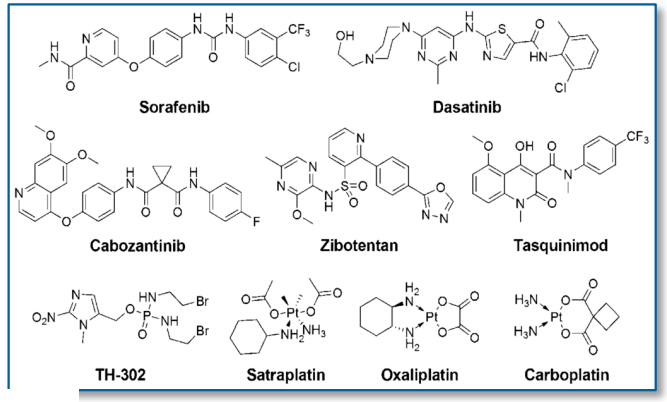
A single model trained on data from many cancer samples, many drugs that can predict drug response across wide range of tumors and drug combinations

MODELING CANCER DRUG RESPONSE

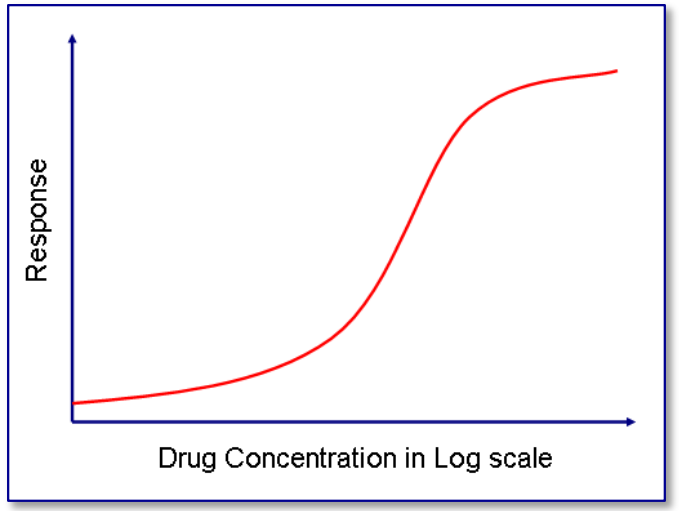


Drug (s)

- descriptors
- fingerprints
- structures
- SMILES
- dose



$$\mathcal{R} = f(\mathcal{T}, \mathcal{D}_1, \mathcal{D}_2)$$

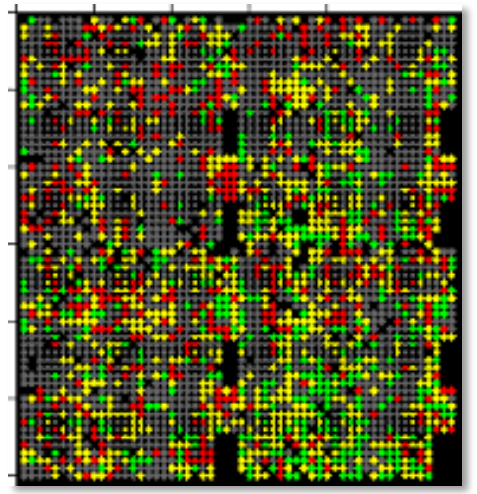


Response

- IC50
- AUC
- GI50
- % growth
- Z-score

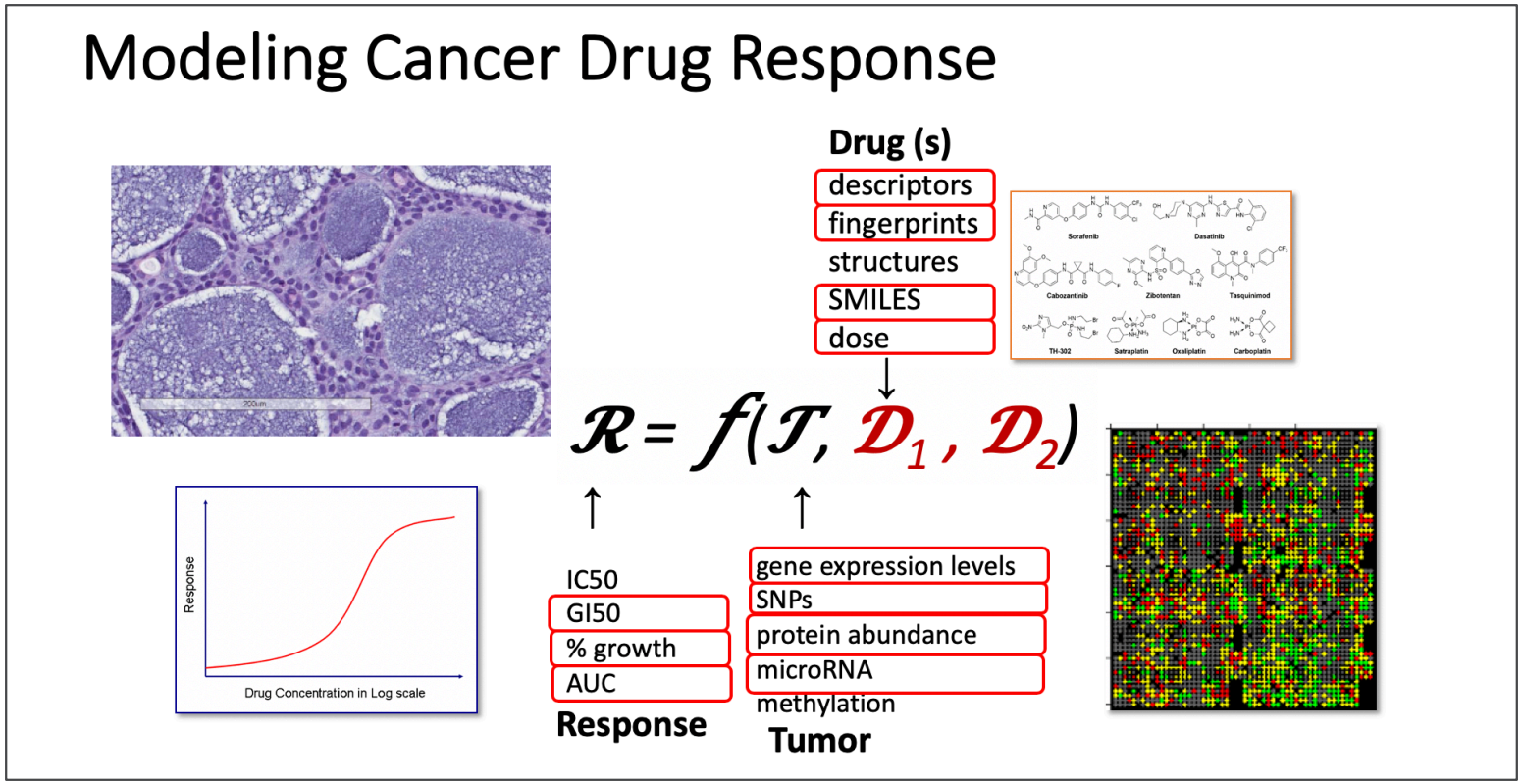
Tumor

- gene expression levels
- SNPs
- protein abundance
- microRNA
- methylation

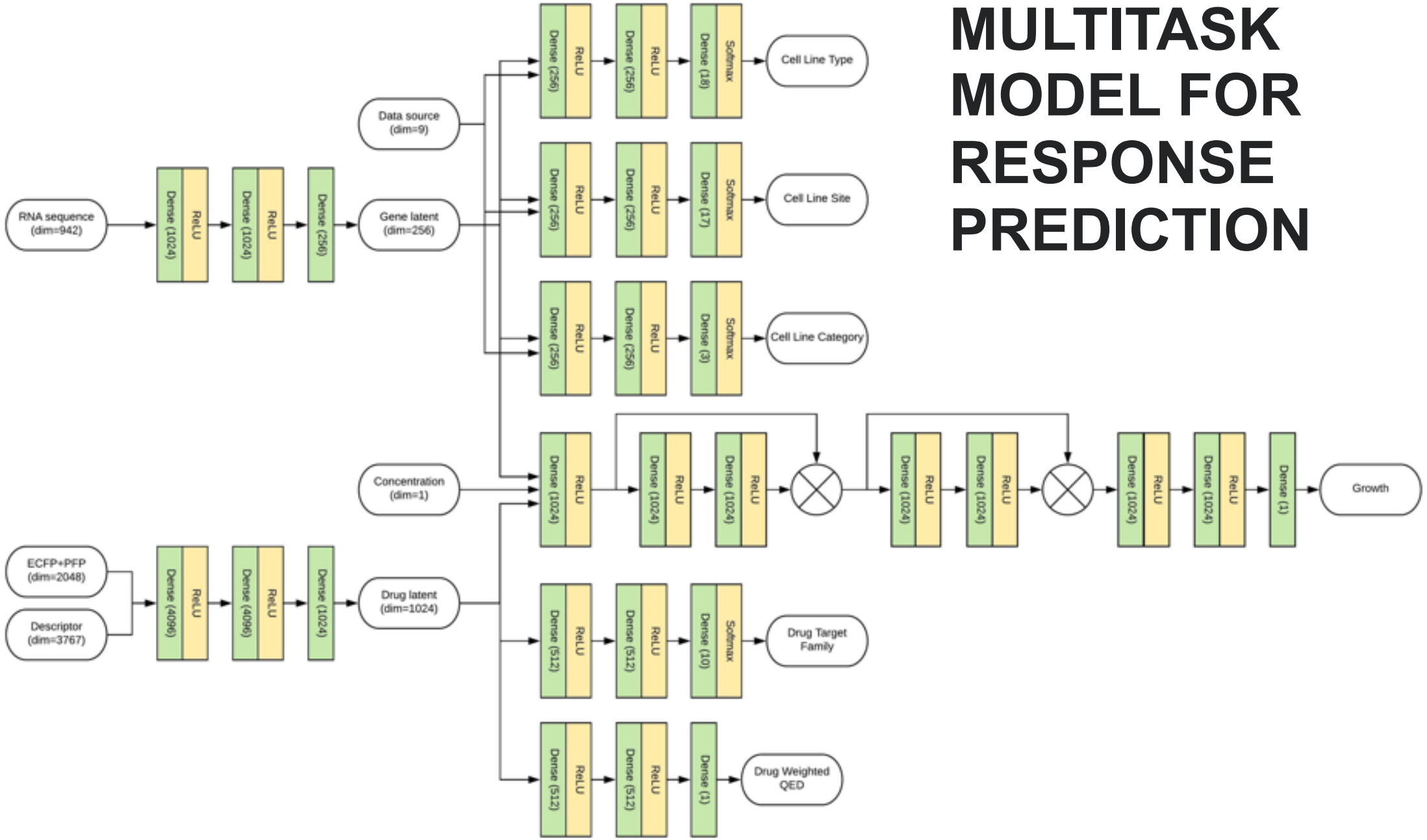


WHAT FEATURES TO USE FOR DRUGS AND TUMORS?

- Tumors
  - RNAseq
  - SNPs/CNVs
  - Protein Abundance
- Drugs
  - Descriptors
  - Structures
  - SMILES
- Vector Embeddings
  - AE/VAE
  - Expression
  - SMILES



DEEP MULTITASK MODEL FOR RESPONSE PREDICTION



CAN WE BUILD MODELS THAT ARE PREDICTIVE OF DRUG RESPONSE?

Dose Independent, Top 6. Top21, cancers, Attention MLP (Means from 10-fold CV)

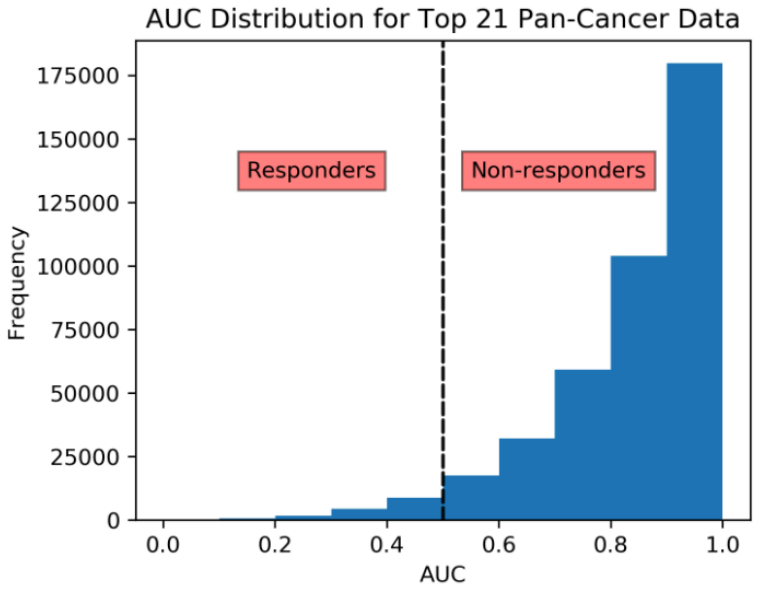
Top 6 Cancer Types

Precision	Recall	f1-score	
0.917	0.790	0.837	Mordred, Lincs1000 (bin.3)
0.933	0.853	0.882	Dragon7, Lincs1000 (bin.3)
0.933	0.855	0.884	Dragon7, Lincs1000 (bin.1)

Top 21 Cancer Types

Precision	Recall	f1-score	
0.95	**0.927**	**0.935**	Dragon7, Lincs1000 (bin.3)

(~6,200 features)

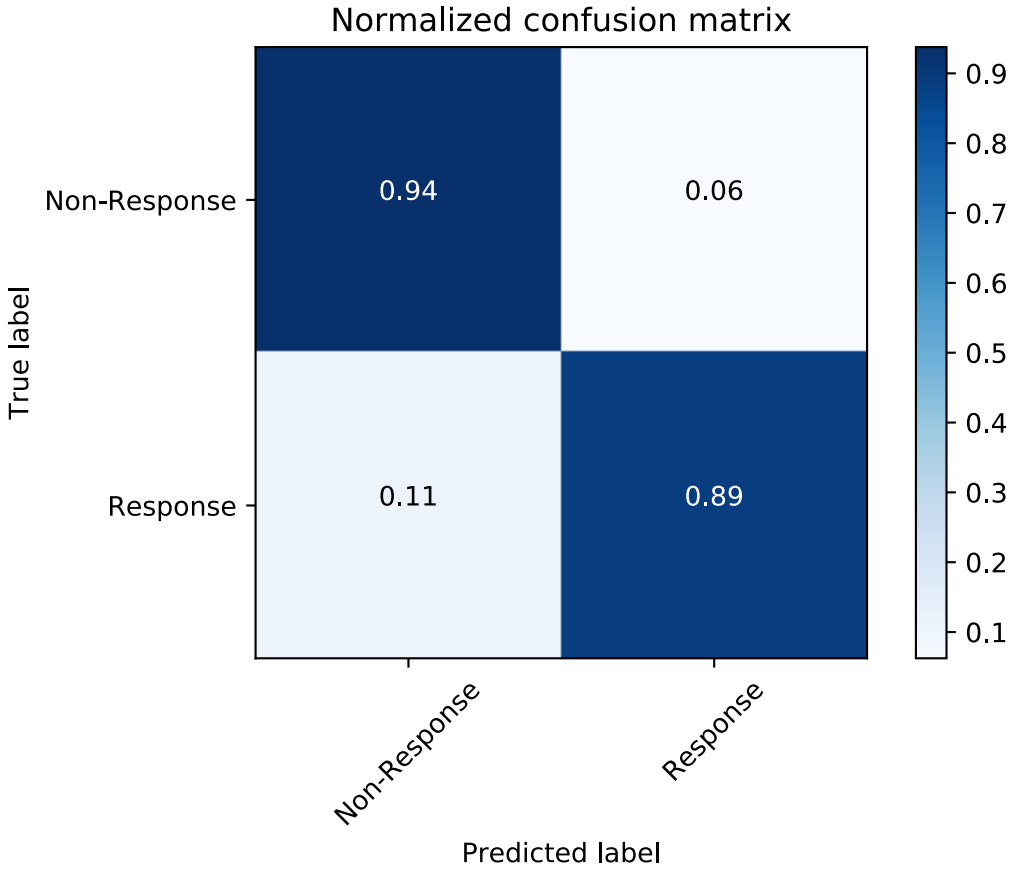


SINGLE DRUG RESPONSE

Drug	R^2	MAE	AUC	Accuracy
Afatinib	0.4369	0.0737	0.9248	**0.9679**
Bortezomib	0.3871	0.0752	**0.9429**	0.9569
Docetaxel	**0.5748**	0.1154	0.9158	0.8853
Doxorubicin	0.3749	0.1103	0.7794	0.7105
Etoposide	0.3787	0.1108	0.8855	0.8768
GDC-0941	0.3294	0.0744	0.6924	0.9478
Navitoclax	0.4329	0.0982	0.9035	0.9295
Paclitaxel	0.5299	0.1285	0.8471	0.7626
Selumetinib	0.2944	0.1056	0.8831	0.9115
SN-38	0.3415	0.1150	0.8269	0.8361
Temsirolimus	0.2048	0.1136	0.7406	0.8912
Tipifarnib	0.3187	0.1115	0.8474	0.8981
Vinorelbine	0.1407	0.1289	0.7605	0.8367
Vorinostat	0.4041	**0.0627**	0.9134	0.9532
mean	**0.3678**	**0.1017**	**0.8474**	**0.8832**

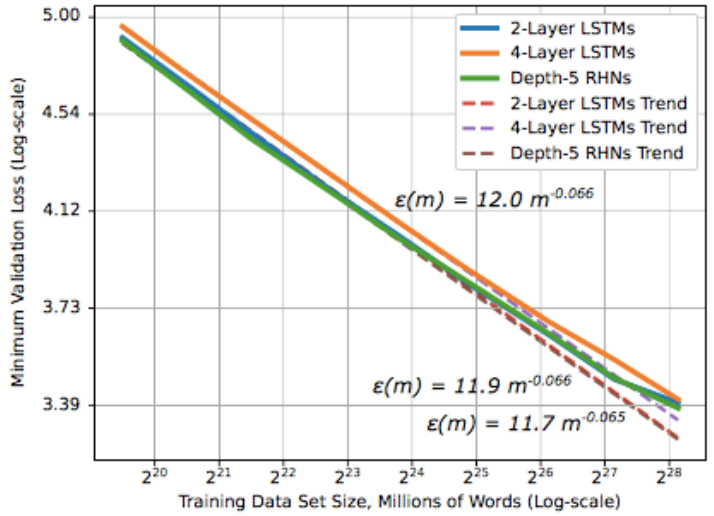
Multi-Drug “Pan cancer”

Top 21 Cancer Types in MD DI formulation



Models are best of RF, LGB, GB, LR, etc.; features are RNAseq and D7 descriptors

LEARNING CURVE POWER LAW



NLP Learning Curves

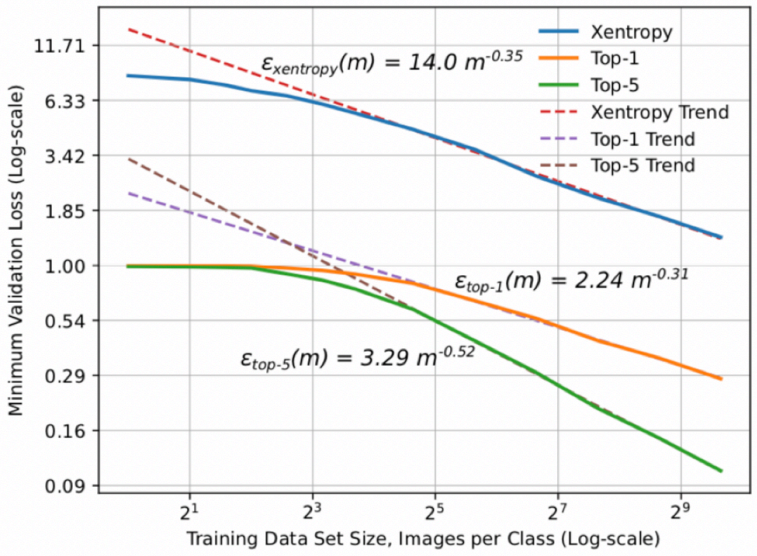
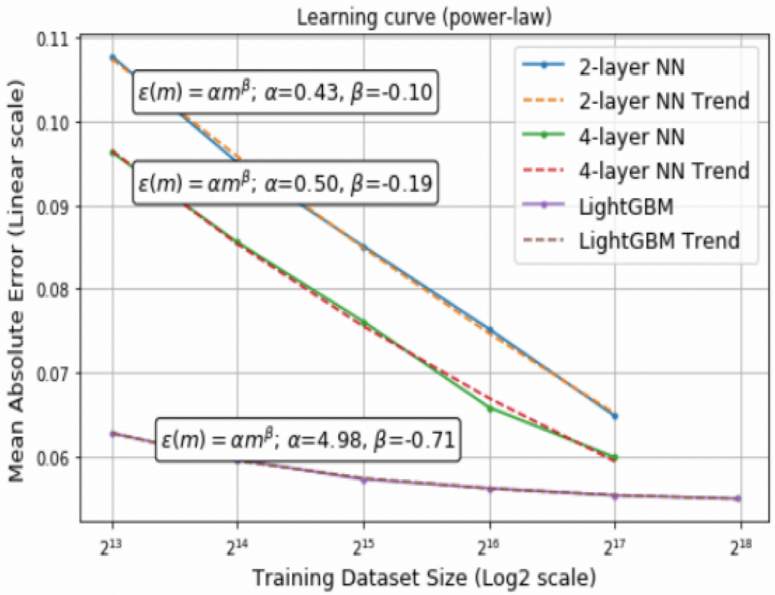


Image Classification



Top6 Cancer Response

It seems that the advent of models that beat the power-law exponent — that get **more data efficient as they learn** — might be an important empirical milestone on that path.

https://arxiv.org/pdf/1712.00409.pdf

CAN WE BUILD MODELS THAT GENERALIZE ACROSS STUDIES?

UNO-MT

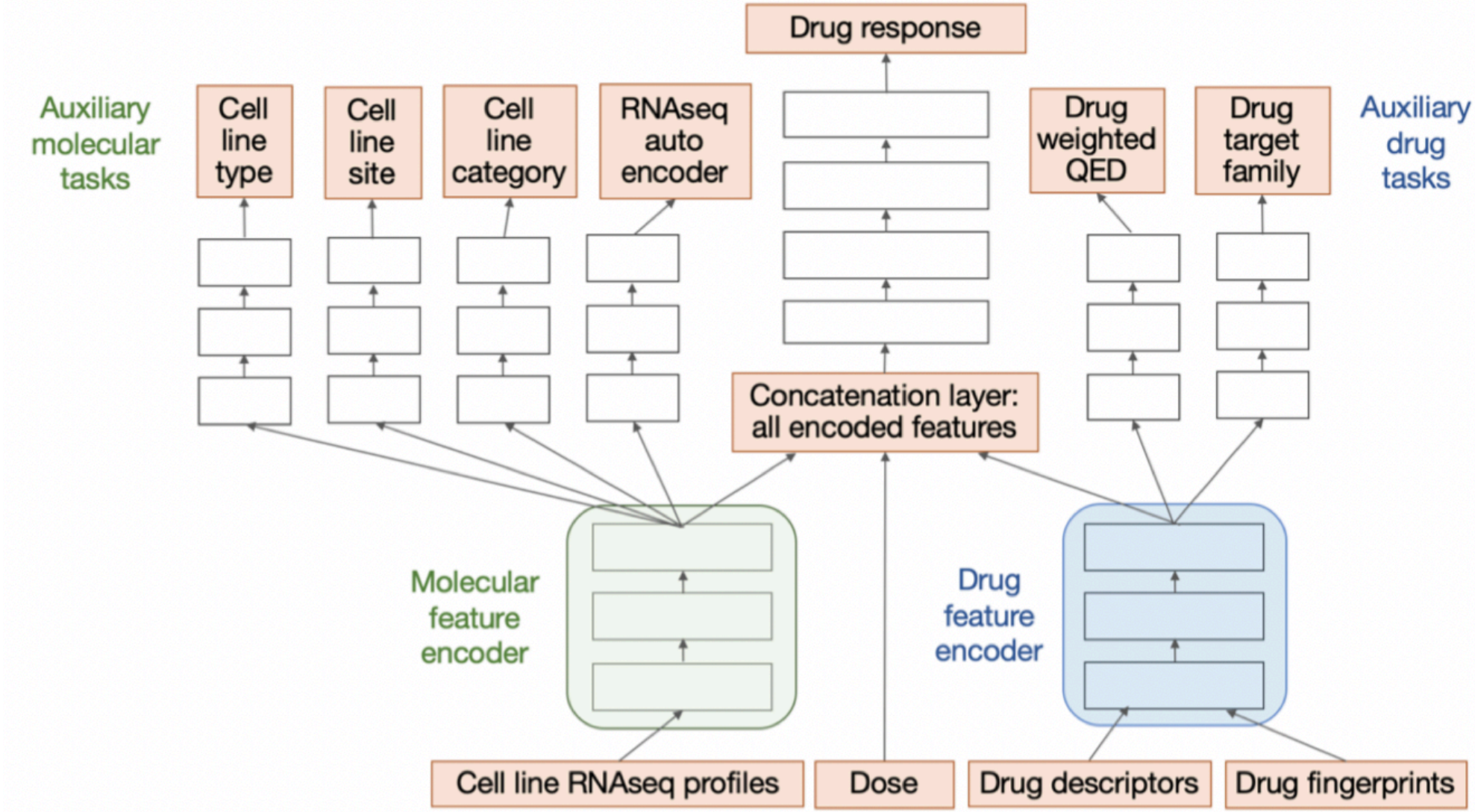


Table 2: Baseline cross study validation results with Random Forest

Training Set	Testing Set: NCI60	CTRP	GDSC	CCLE	gCSI
NCI60	$R^2 = 0.45$ MAE = 30.4	$R^2 = 0.23$ MAE = 34.6	$R^2 = 0.15$ MAE = 37.3	$R^2 = 0.29$ MAE = 34.3	$R^2 = 0.14$ MAE = 54.0
CTRP	$R^2 = 0.41$ MAE = 31.7	$R^2 = 0.30$ MAE = 35.0	$R^2 = 0.15$ MAE = 37.4	$R^2 = 0.45$ MAE = 29.0	$R^2 = 0.17$ MAE = 39.6
GDSC	$R^2 = 0.33$ MAE = 36.0	$R^2 = 0.14$ MAE = 41.5	$R^2 = 0.13$ MAE = 40.4	$R^2 = 0.17$ MAE = 42.4	$R^2 = 0.08$ MAE = 43.0
CCLE	$R^2 = 0.12$ MAE = 42.6	$R^2 = -0.03$ MAE = 48.9	$R^2 = -0.11$ MAE = 47.1	$R^2 = 0.17$ MAE = 42.4	$R^2 = 0.32$ MAE = 38.5
gCSI	$R^2 = -0.38$ MAE = 55.0	$R^2 = -0.51$ MAE = 59.0	$R^2 = -0.59$ MAE = 58.7	$R^2 = -0.09$ MAE = 48.6	$R^2 = 0.25$ MAE = 39.9

UnoMT Multitask Deep Learning Cross-Study

Best out of Study $R^2 = 0.61$

Table 6. Best cross study validation results with a 3-task UnoMT

		Testing set							
		NCI60	CTRP	GDSC	CCLE	gCSI	N/T Cat Acc	Site Acc	Type Acc
Training set	NCI60	R2 = 0.81 MAE = 17.1	**R2 = 0.38** **MAE = 32.2**	**R2 = 0.24** **MAE = 35.3**	**R2 = 0.48** **MAE = 33.4**	**R2 = 0.46** **MAE = 33.4**	**99.43%**	**96.75%**	**96.97%**
	CTRP	**R2 = 0.44** **MAE = 29.8**	R2 = 0.68 MAE = 22.7	**R2 = 0.23** **MAE = 34.4**	**R2 = 0.61** **MAE = 28.3**	**R2 = 0.60** **MAE = 28.5**	**99.56%**	96.62%	96.58%
	GDSC	R2 = 0.32 MAE = 34.0	**R2 = 0.25** **MAE = 36.7**	R2 = 0.53 MAE = 27.2	R2 = 0.50 MAE = 32.6	**R2 = 0.60** **MAE = 29.2**	**99.43%**	**96.93%**	**96.97%**
	CCLE	**R2 = 0.27** **MAE = 36.9**	**R2 = 0.20** **MAE = 39.2**	**R2 = 0.11** **MAE = 38.9**	R2 = 0.68 MAE = 25.4	R2 = 0.39 MAE = 34.2	**99.12%**	**96.36%**	**96.36%**
	gCSI	R2 = 0.00 MAE = 44.9	**R2 = 0.11** **MAE = 43.1**	R2 = 0.05 MAE = 42.8	**R2 = 0.33** **MAE = 40.6**	R2 = 0.80 MAE = 192	**99.43%**	**96.84%**	**96.62%**

MAE = Mean Absolute Error (in percent growth)

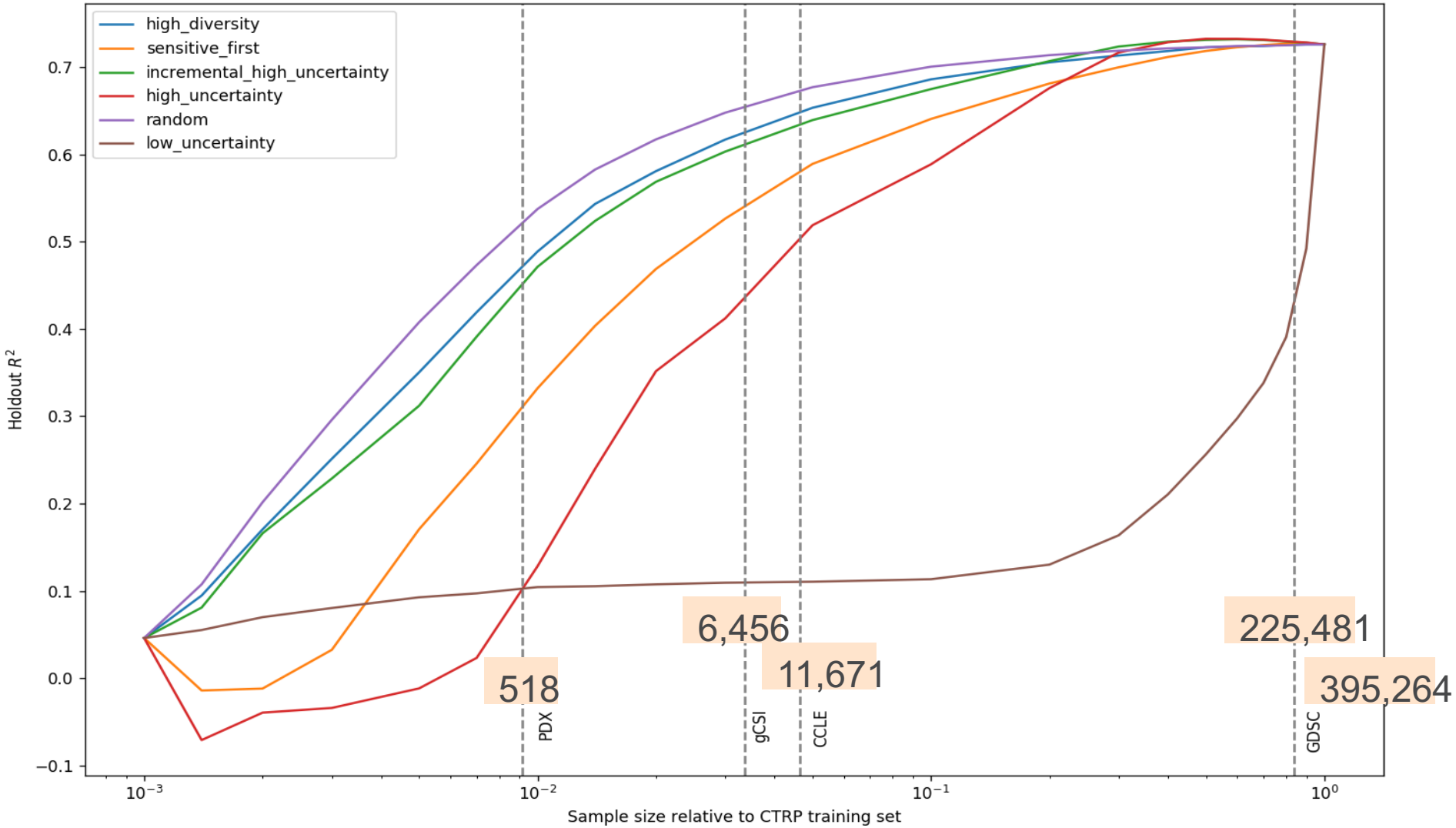
Comparison on PDX Prediction Performance With and Without Transfer Learning

Analysis name	R^2	P-value (R^2)	Spearman rank correlation coefficient	P-value (Spearman rank correlation coefficient)
PDX-Only	0.064(0.031)		0.372(0.022)	
CCLE-TL	0.042(0.016)	8.01E-02	0.355(0.013)	7.28E-02
gCSI-TL	0.100(0.016)	8.29E-03	0.389(0.017)	7.55E-02
NCI60-TL	0.102(0.013)	5.16E-03	0.407(0.016)	1.43E-03
CTRP-TL	0.092(0.019)	3.35E-02	0.415(0.013)	1.51E-04
GDSC-TL	0.110(0.017)	1.50E-03	0.419(0.013)	7.22E-05

PDX-only is the analysis without transfer learning. -TL in analysis name indicates transfer learning from a CCL dataset.

- Mean (standard deviation) of prediction performance is evaluated through 10 times of 10-fold cross-validations on PDXs
- Four out of the five transfer learning analyses show a prediction performance statistically significantly better than that of PDX-only analysis, evaluated by the p-value of t-test ≤ 0.05

ACTIVE LEARNING SIMULATION



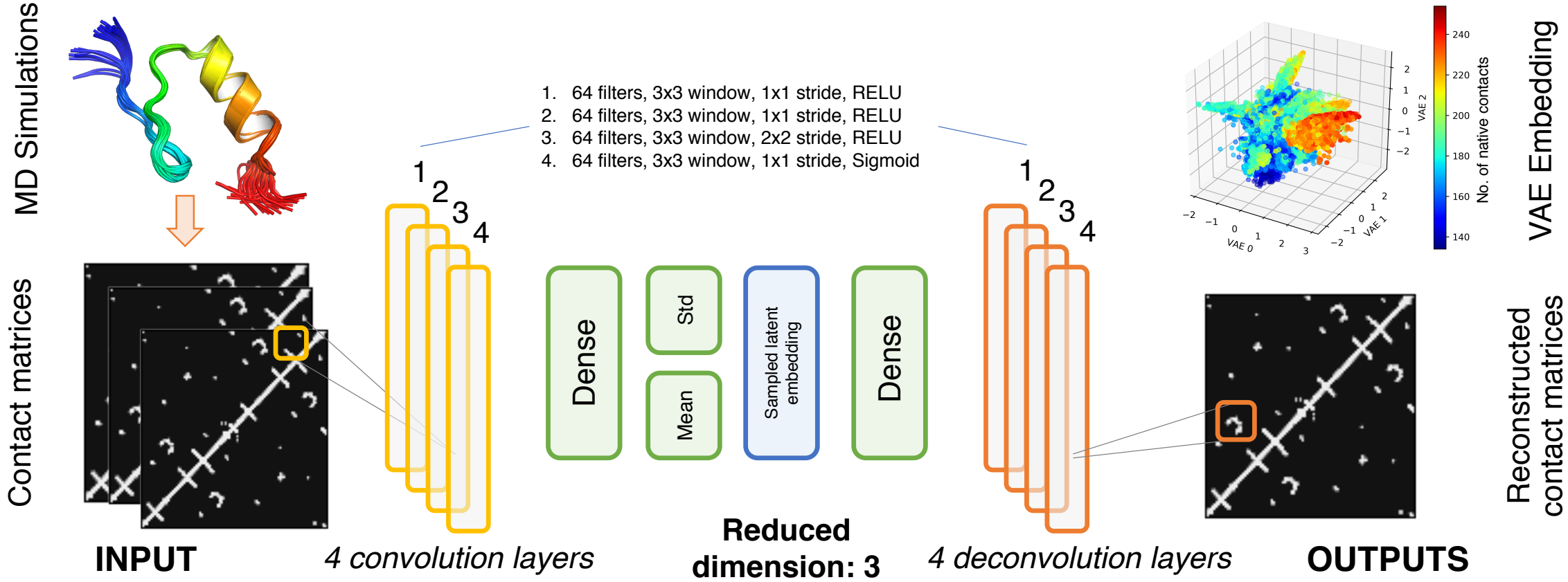
SUMMARY

- A suite of deep learning models that have been applied to drug response prediction:
  - DL models show better predictive power
  - More data → more predictive power!
  - An active learning simulation demonstrates how much data we may ultimately need to have a single model that works across different types of cancers and different drugs
- Uncertainty quantification across models (although not discussed)
- Consistent evaluation across multiple datasets and prediction tasks

PILOT 2: OVERARCHING MODELING GOAL

Build unsupervised machine learning models to potentially steer molecular dynamics simulations towards “interesting states”

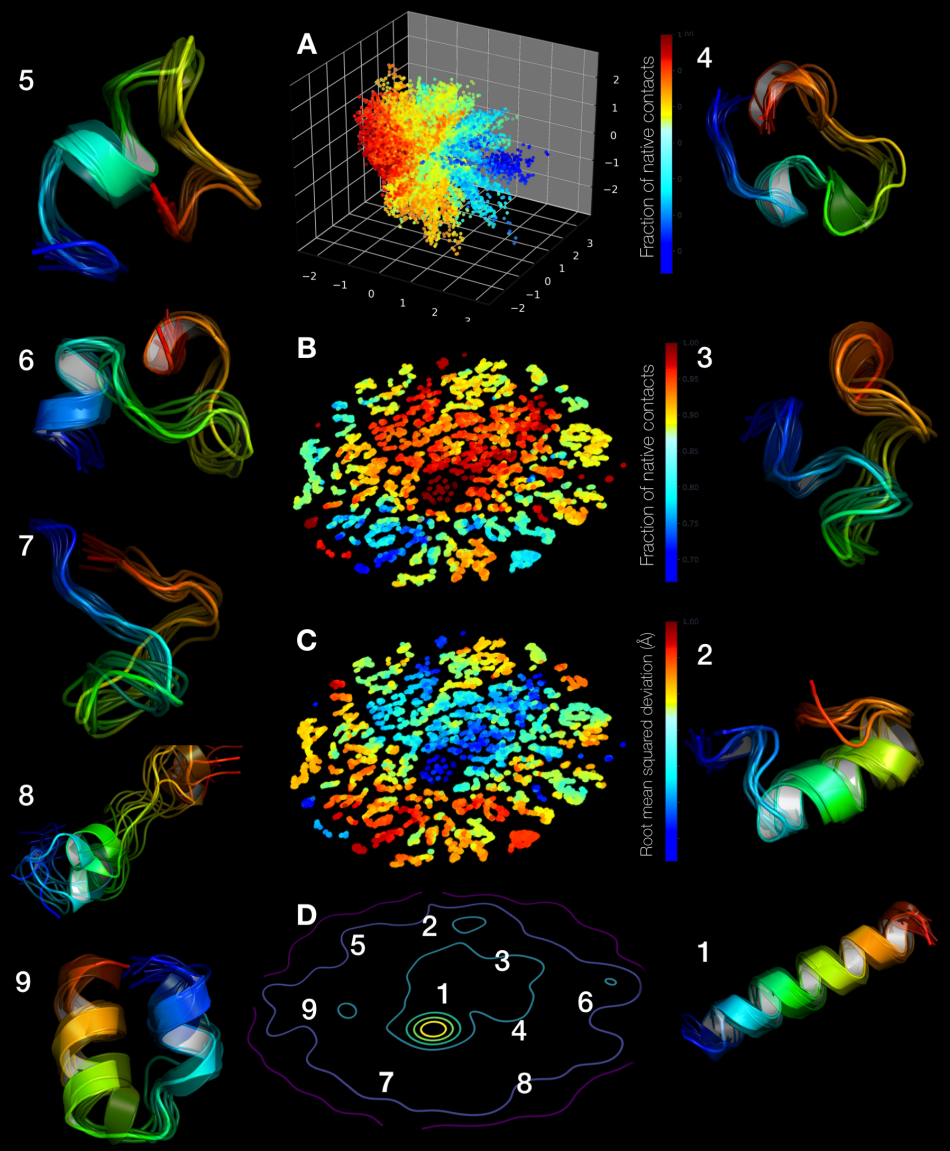
A VARIATIONAL APPROACH TO ENCODE PROTEIN FOLDING WITH CONVOLUTIONAL AUTO-ENCODERS (CVAE)

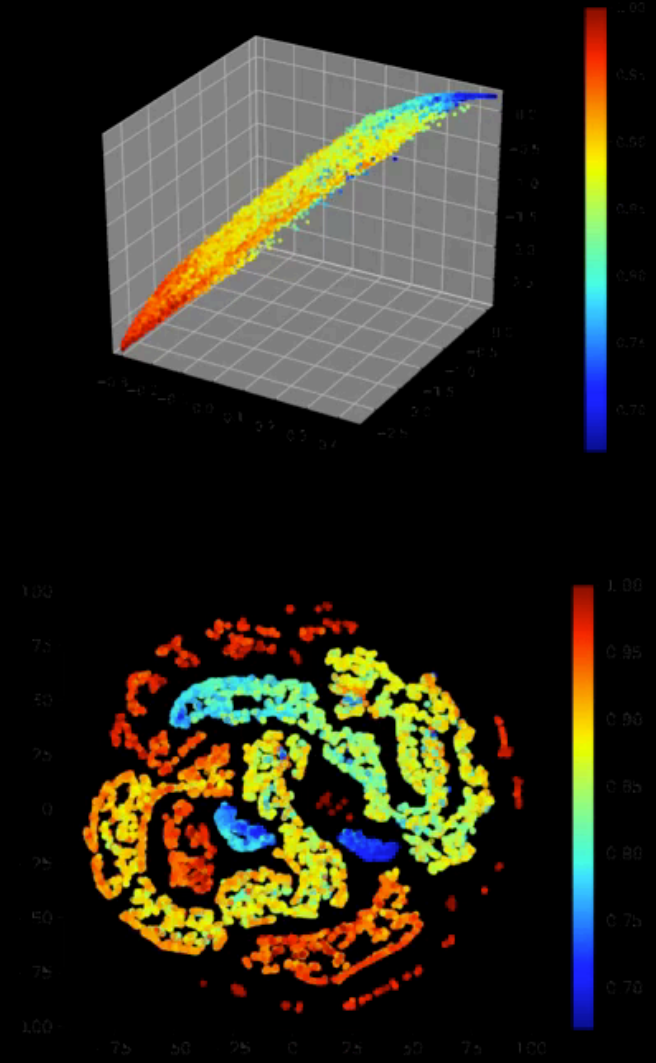


D. Bhowmik, M.T. Young, S. Gao, A. Ramanathan, BMC Bioinformatics (2019)
Source code: http://ramanathanlab.org

Related work:
Hernandez 17 arXiv,
Doerr 17 arXiv

CVAE REVEALS ”METASTABLE STATES” IN PROTEIN FOLDING…

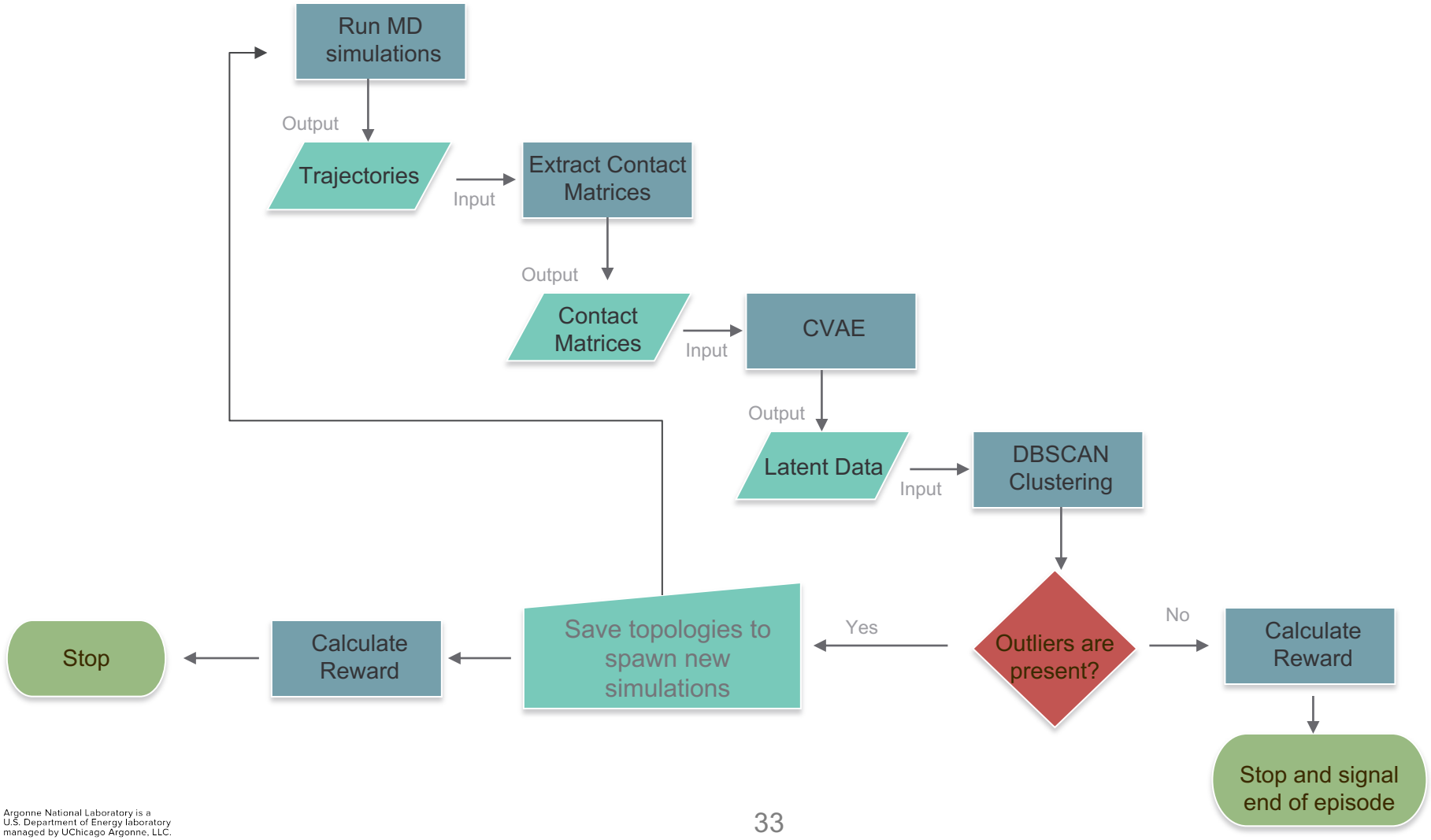




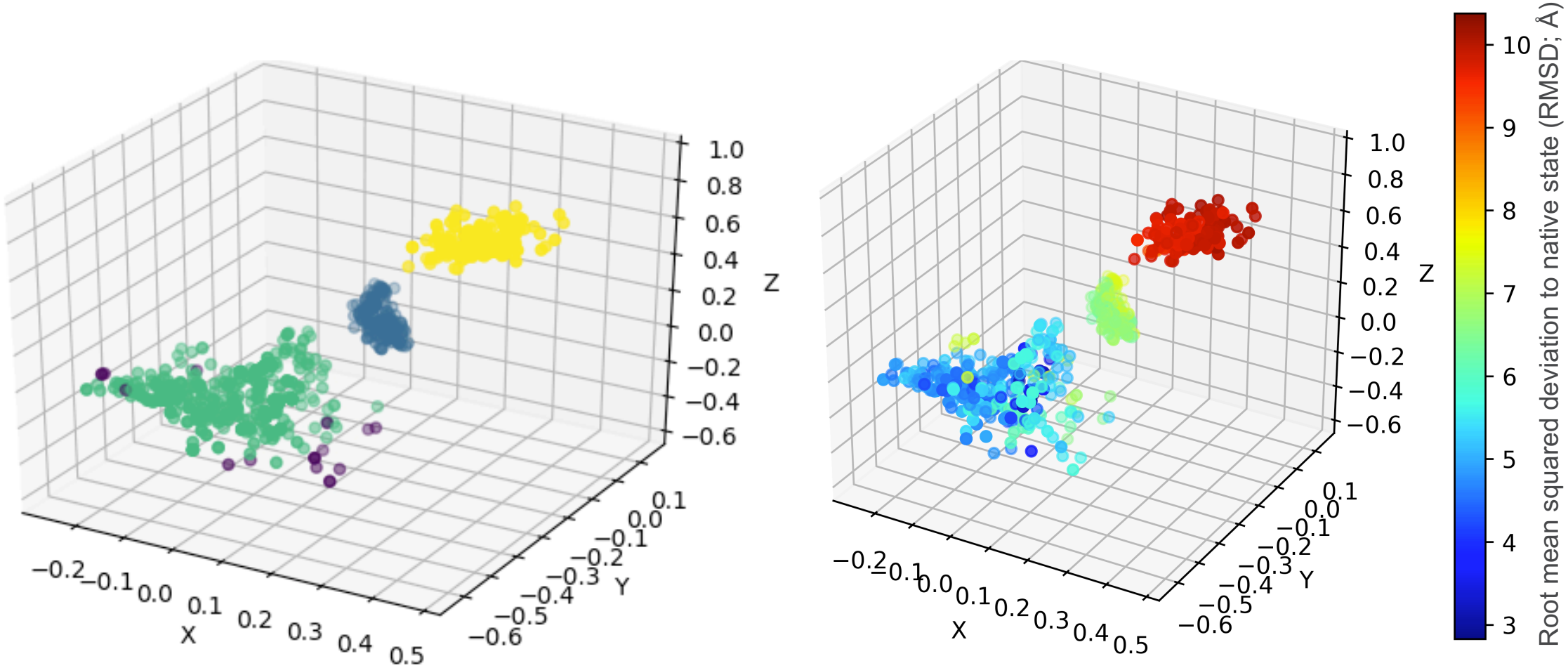
MSM Builder Datasets, Pande group

WHERE TO SAMPLE NEXT?

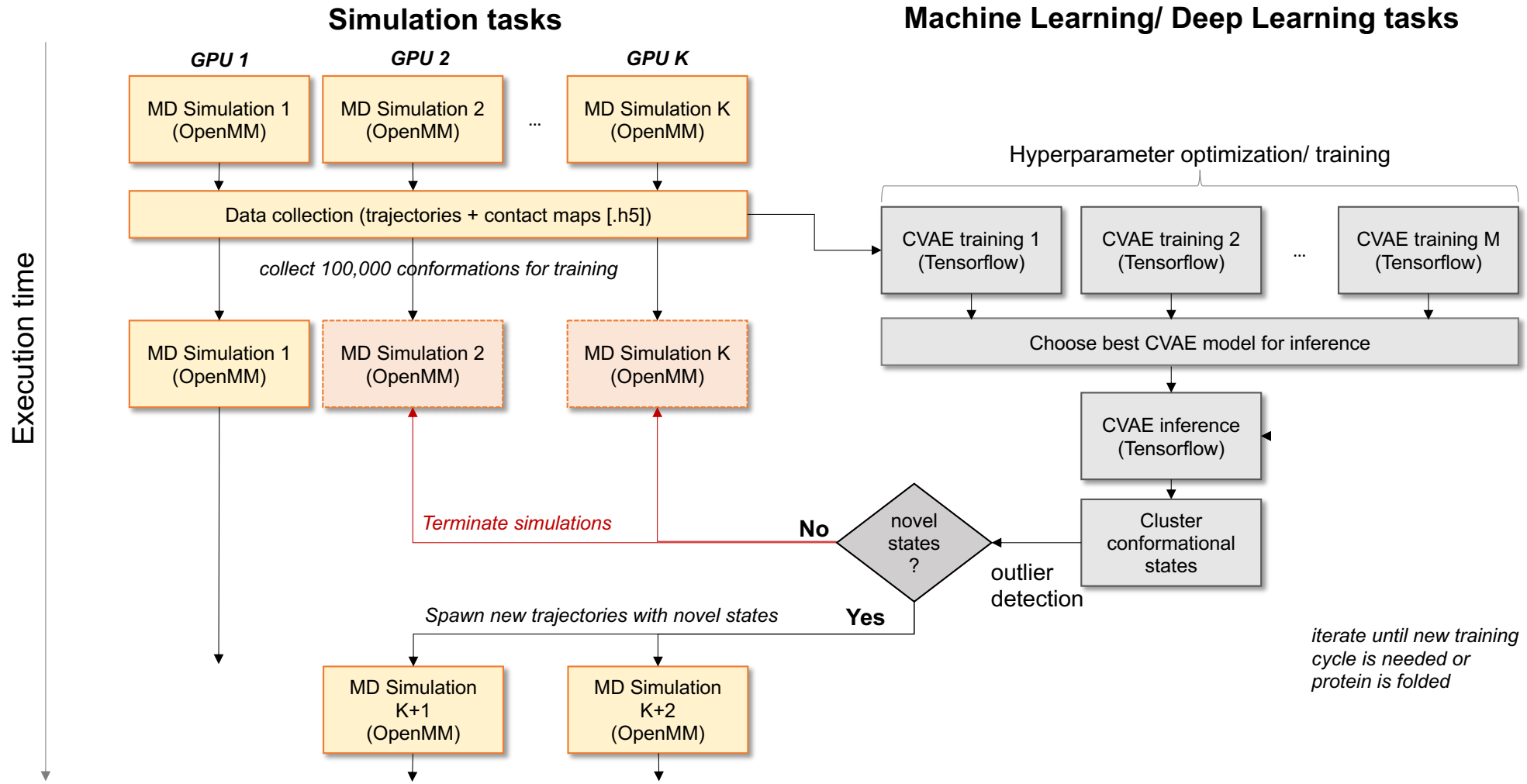
IDEA: CONVERT FROM ‘TRAINING’ MODE TO ‘INFERENCE’ MODE…



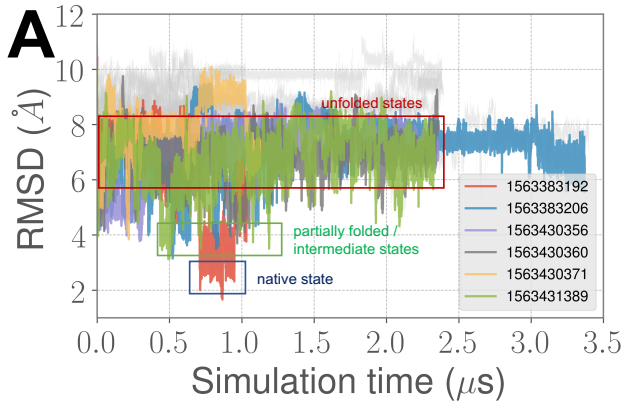
NOVEL DATA POINTS IN THE LATENT SPACE ENABLE SAMPLING FOLDED STATES

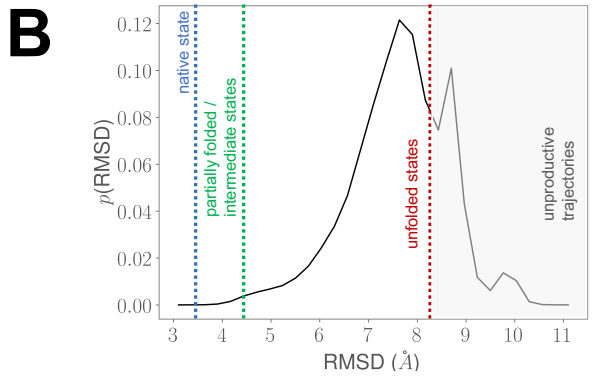


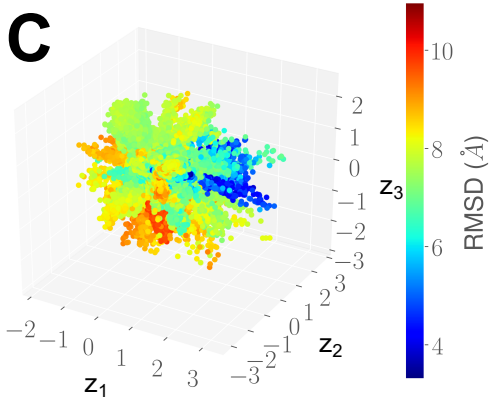
PUTTING TOGETHER A SCALABLE WORKFLOW

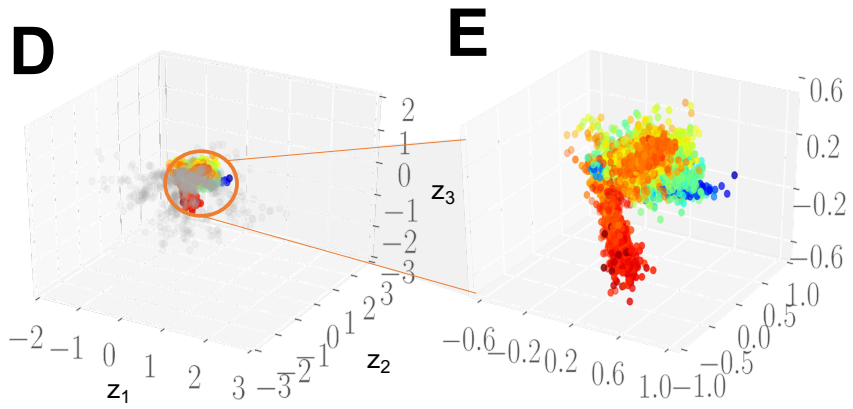


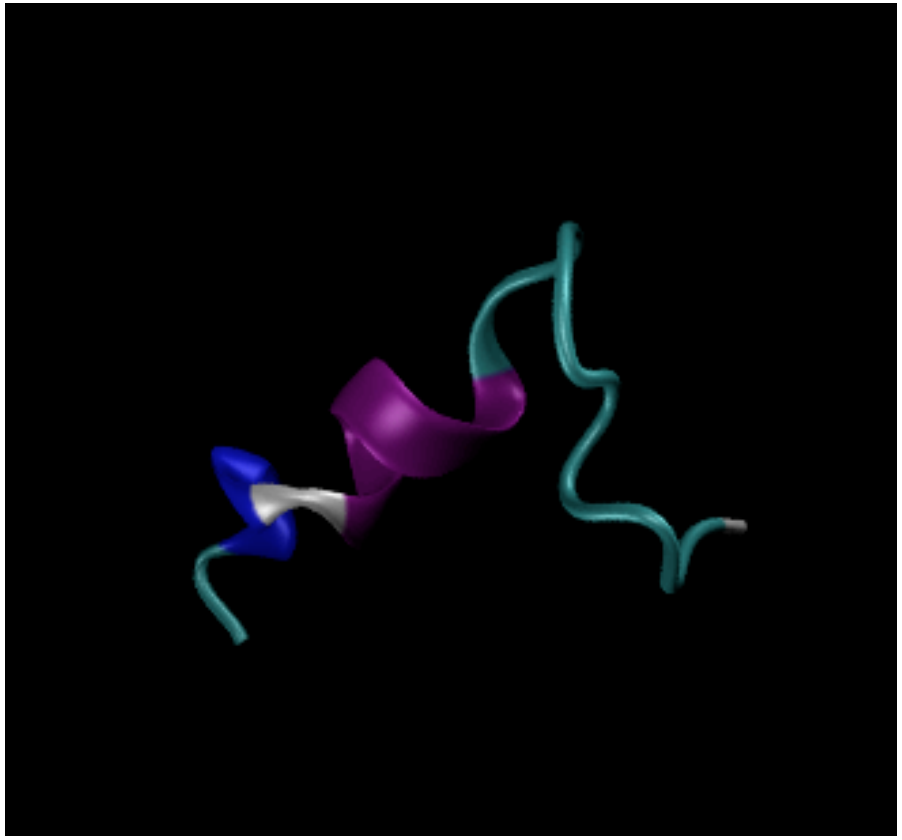
YES, WE CAN FOLD A PROTEIN… [CASE 1: FS-PEPTIDE]

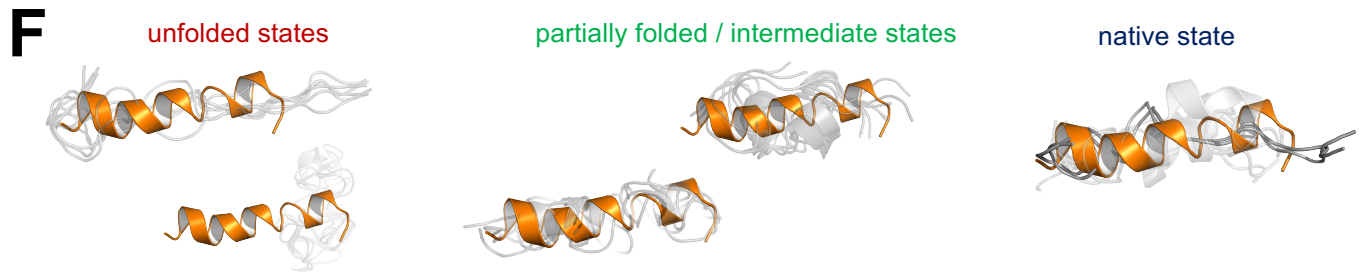












VISUALIZATION CAPABILITIES: INTERACTING WITH INSIGHTS FROM DEEP LEARNING APPROACHES

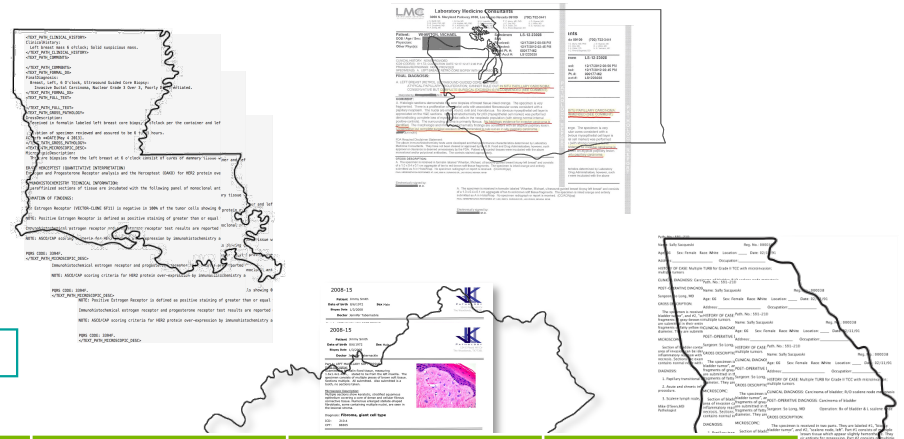
SUMMARY

- Demonstration that deep learning interleaved with MD simulations can lead to productive trajectories:
  - protein folding is one example
  - refining MD simulations in the context of experimental data
- Scaling issues with AI/DL integrated simulation workflows need new ways to think about performance:
  - key challenge emerges from training times of AI/DL are ‘on par’ with simulation timescale
  - Effective performance metric: ratio of the time taken to solution (e.g., achieving RMSD of 0.3 Å to the native state) of application with and without learning
- New hardware/software needs for AI/DL coupled MD workflows:
  - Streaming analytics

CANCER PATHOLOGY REPORT PROCESSING PIPELINE

NCI SEER Database

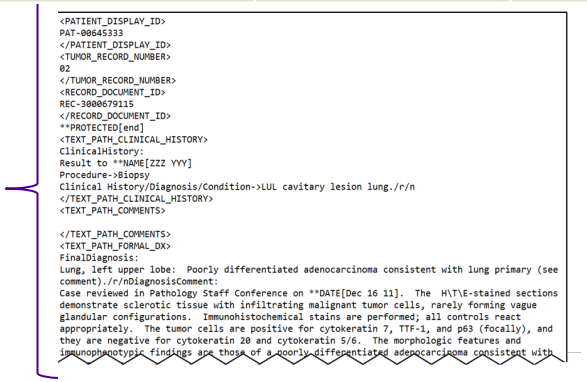
Integration with structured data from Electronic medical records for patients



Registry	PatientID	Record No.	Tumor No.	Primary Site	Source Section	Relevant text	Primary site category	Primary site code
KY	114431		3	Breast	Final diagnosis	Mammary carcinoma	Breast	C50.9 Breast,NOS
KY	118420		5	Breast	Final diagnosis	BREAST PRIMARY	BREAST	C50.9 Breast, NOS
SE	0084621	500713999	01	Lung	Final diagnosis	Lung, right lower lobe	lung	C34.3 lower lobe, lung

Patient → **Pathologist**

Diagnosis by a pathologist analyzing tissue specimen from patient



Certified Tumor Registrar

CTR at a cancer registry reviews complete patient medical record + path report

NCI-SEER IS A PRIMARY DATA SOURCE… NEED TO MODERNIZE

- **NEED**
  - Abstracting structured data from free-text pathology reports is critical for the national cancer surveillance program
- **CHALLENGE**
  - Manual abstraction is time-consuming, costly, and not scalable
- **GOAL**
  - Develop a scalable framework for automated information extraction from pathology reports

```
<TEXT_PATH_CLINICAL_HISTORY>
ClinicalHistory:
  Left breast mass 6 o?clock; Solid suspicious mass.
</TEXT_PATH_CLINICAL_HISTORY>
<TEXT_PATH_COMMENTS>

</TEXT_PATH_COMMENTS>
<TEXT_PATH_FORMAL_DX>
FinalDiagnosis:
  Breast, Left, 6 O'clock, Ultrasound Guided Core Biopsy:
    Invasive Ductal Carcinoma, Nuclear Grade 3 Over 3, Poorly Differentiated.
</TEXT_PATH_FORMAL_DX>
<TEXT_PATH_FULL_TEXT>

</TEXT_PATH_FULL_TEXT>
<TEXT_PATH_GROSS_PATHOLOGY>
GrossDescription:
  Received in formalin labeled left breast core biopsy 6 o?clock per the container and lef

  Fixation of specimen reviewed and assured to be 6 to 48 hours.
AC:lefb **DATE[May 4 2013].
</TEXT_PATH_GROSS_PATHOLOGY>
<TEXT_PATH_MICROSCOPIC_DESC>
MicroscopicDescription:
  The core biopsies from the left breast at 6 o'clock consist of cores of mammary tissue w

ER/PR HERCEPTEST (QUANTITATIVE INTERPRETATION)
Estrogen and Progesterone Receptor analysis and the Herceptest (DAKO) for HER2 protein ove

IMMUNOHISTOCHEMISTRY TECHNICAL INFORMATION:
Deparaffinized sections of tissue are incubated with the following panel of monoclonal ant

SUMMATION OF FINDINGS:

The Estrogen Receptor (VECTOR-CLONE 6F11) is negative in 100% of the tumor cells showing 0

NOTE: Positive Estrogen Receptor is defined as positive staining of greater than or equal

Immunohistochemical estrogen receptor and progesterone receptor test results are reported

NOTE: ASCO/CAP scoring criteria for HER2 protein over-expression by immunohistochemistry a


PQRS CODE: 3394F.
</TEXT_PATH_MICROSCOPIC_DESC>
```

DATASETS USED FOR PRELIMINARY RESEARCH

STUDY 1: Limited dataset of de-identified breast and lung cancer electronic pathology (e-path) reports from 5 different SEER registries

~2,500 breast and lung cancer de-identified e-path reports

Partially annotated for **subsite, laterality, grade, behavior**

STUDY 2: Large dataset of e-path reports from Louisiana Tumor Registry housed at the PHI enclave within ORNL

~267,000 reports from Louisiana Tumor Registry (2004-2017)

Gold standard for **site, laterality, grade, behavior, histology** derived from consolidated "Cancer/Tumor/Case" (CTC) records

EXPERIMENTAL PIPELINE

Data Pre-processing

- Duplicate records
- Non-contradicting labels
- Incorrect organ annotations
- Small sample sizes
- Corpus curation

Feature Representation

- TF-IDF
- Bag-of-graphs
- RAKE
- CHUNK
- GLOVE

Training Protocols

With over sampling

Without over sampling

Rule-based Systems (RL)

- Contextualize (keywords for topics of interest)
- Term identification
- Classification

Machine Learning (ML)

- Naïve Bayes (NB)
- Logistic Regression (LR)
- Random Forest (RF)
- Support Vector Machines (SVM)
- Extreme gradient boosting tree (Xgboost)

Deep Learning (DL)

- Convolutional neural nets (CNN)
- Hierarchical Attention nets (HAN)
- Multi-task Deep neural net (MT-DNN)

Performance Metrics

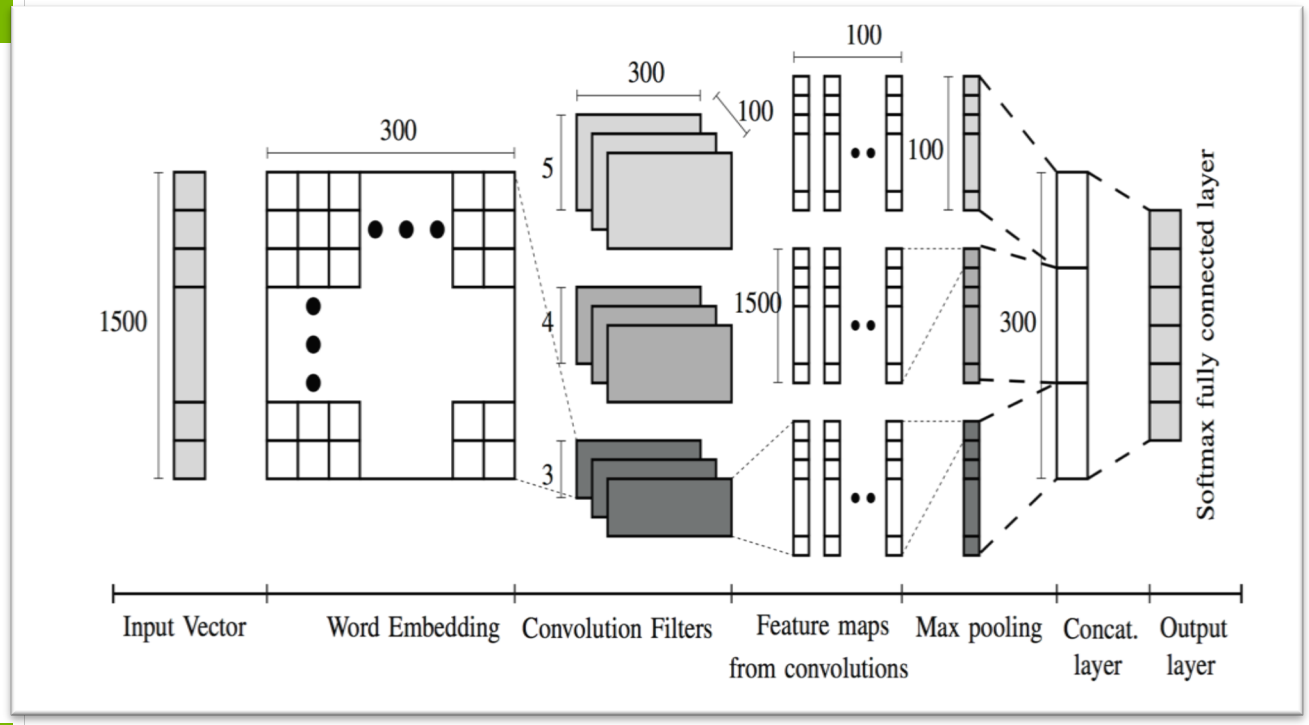
- Precision (positive predictive value) / Recall (sensitivity) / F1 per class
- Macro / Micro scores (aggregate performance over all)

Validation Strategies

- K-fold cross validation (K-fold)
- Leave-one-registry out (LORO)
- Leave-one-case-out per registry (LOO_R)

A ‘GENTLE’ INTRODUCTION TO CONVOLUTIONAL NETS (CNN) FOR TEXT

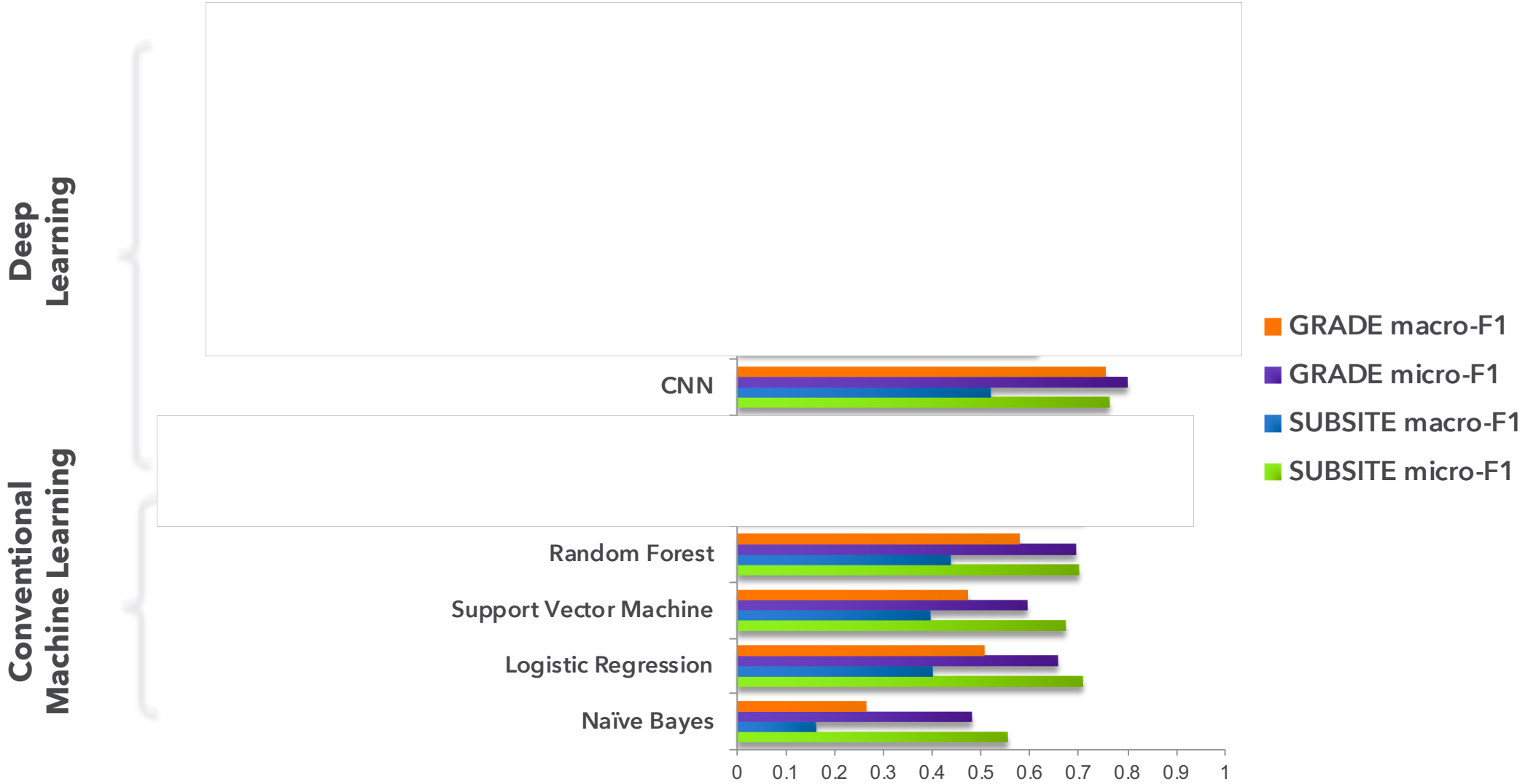
Given a document represented as a collection of words, how do we extract features automatically?



- Text is presented in the form of a document matrix – a sequence of word embedding vectors
- Multiple convolutional filters capture context along a document:
  - Word lengths {3,4,5} are used to “slide” along the entire length
- Network learns to select context features in via max pooling
- Selected features are concatenated and fed though a fully connected layer where regularization occurs
- Output is finally a softmax classifier

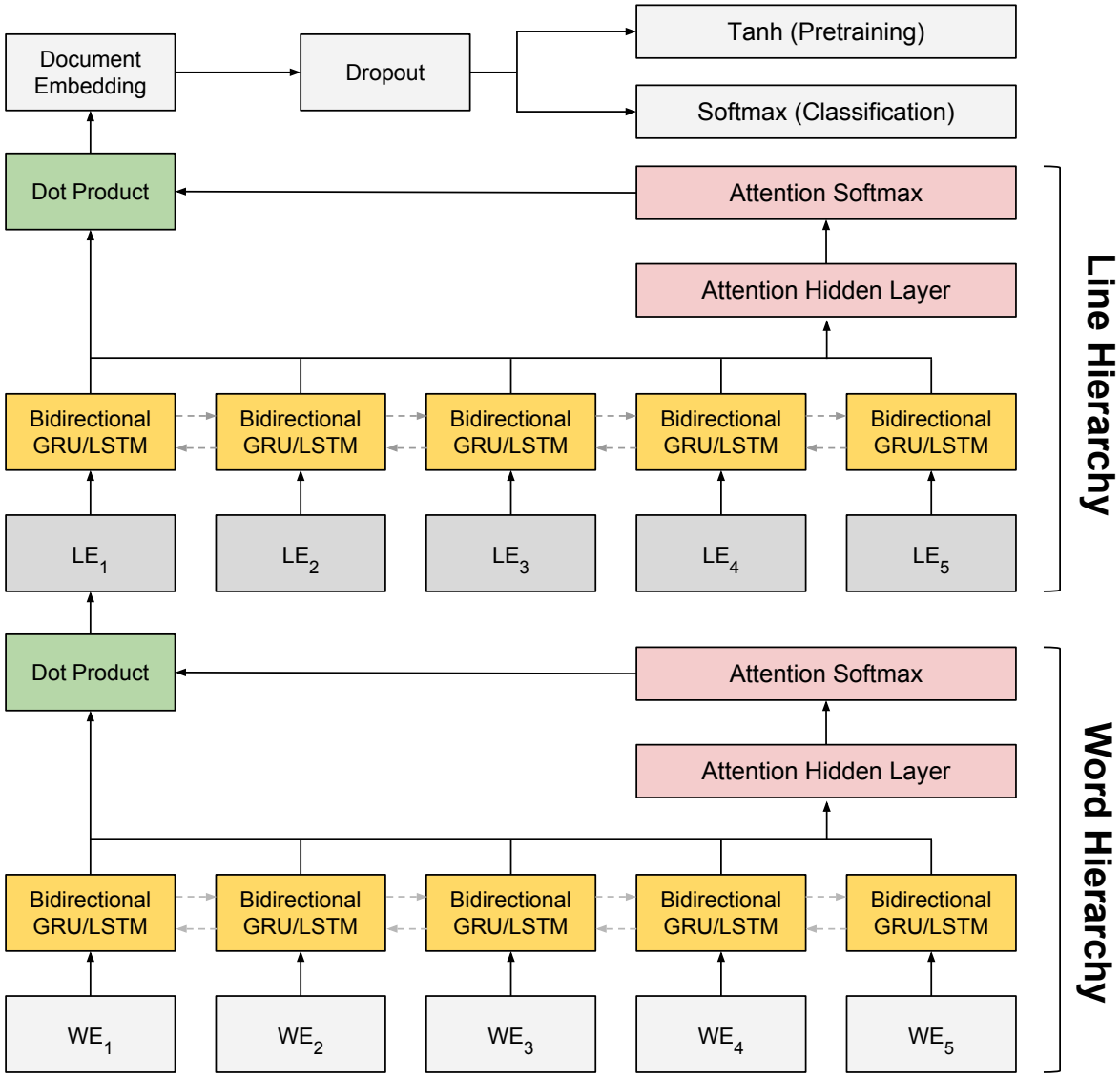
“Deep Learning for Automated Extraction of Primary Sites from Cancer Pathology Reports,” IEEE Journal of Biomedical and Health Informatics [January 2018]

CNNS PERFORM BETTER IN BASIC INFORMATION EXTRACTION TASKS COMPARED TO CONVENTIONAL ML APPROACHES



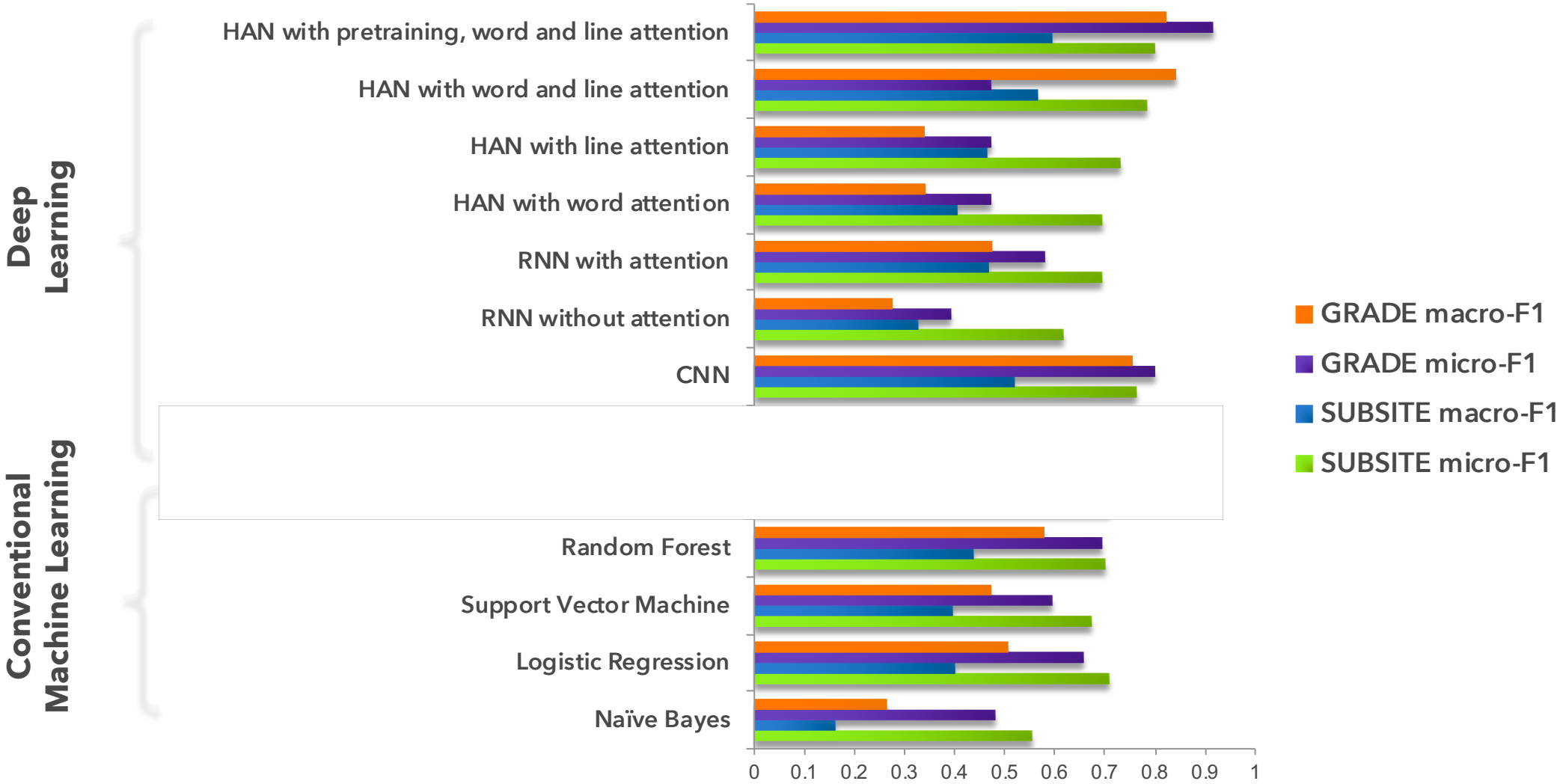
LAYERING AN RNN WITH ATTENTION… HIERARCHICAL ATTENTION NETWORK (HAN)

- Word level embedding:
  - capture important words in a sentence
  - ***Output***: sentence embedding weighted based on word occurrence/ co-occurrence most relevant for classification task
- Sentence level embedding:
  - capture important sentences within a document
  - **Output**: weighted sentence embedding based on relevance for classification task
- Final document embedding is fed into classification



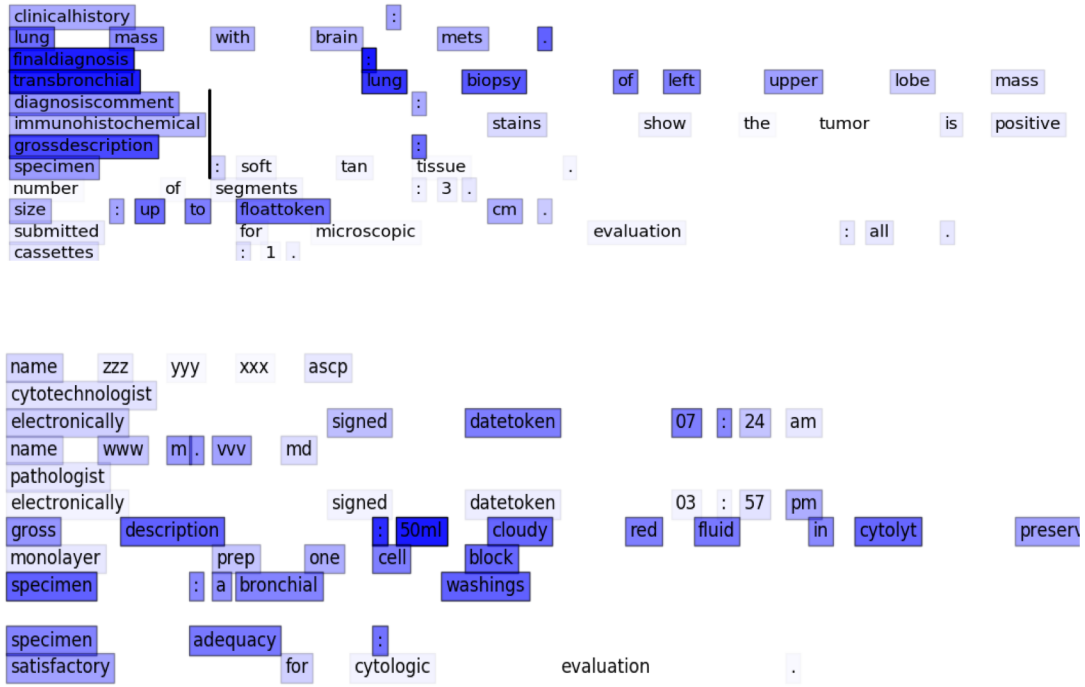
Hierarchical Attention Networks for Information Extraction from Cancer Pathology Reports," Journal of American Medical Informatics Association [appeared online, Nov 2017]

HAN PERFORMS BETTER IN BASIC INFORMATION EXTRACTION TASKS COMPARED TO CONVENTIONAL ML APPROACHES



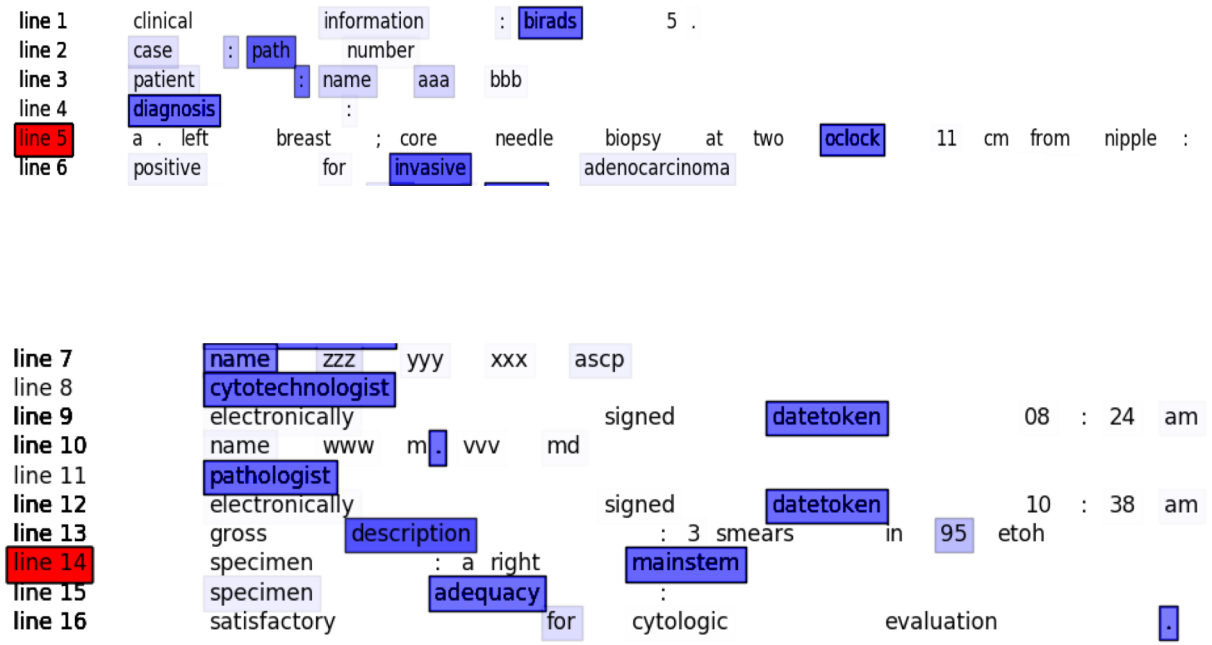
INTERPRETING WHAT CNNS AND HANS LEARNED FROM EPATH REPORTS

CNN



CNNs blindly associate context with importance based on how often words occur in its neighborhood. Moving along a row, these words may not always capture the required clinical context.

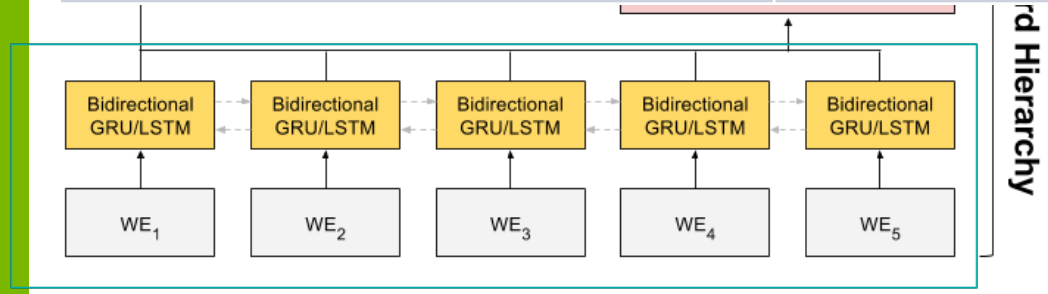
HAN



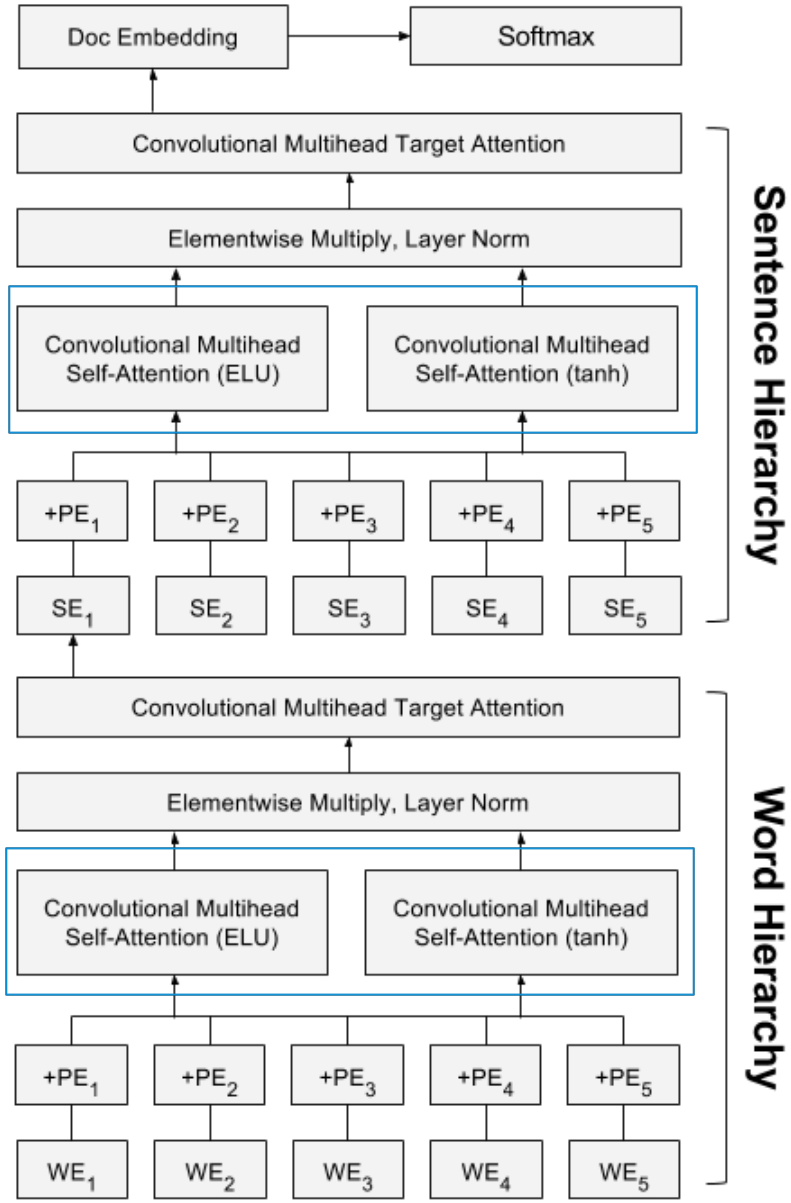
HANs interpret context based on most important words in a sentence → sentences → document. Neighboring words/sentences provide overall importance.

HAN IS SLOW: TWEAKING THE NETWORK TO ACCELERATE TRAINING

	Pubmed
Naïve Bayes	76.63 --, 0.2s
Logistic Regression	76.46 --, 15s
CNN Baseline	77.25 13ms, 1hr
Hierarchical Attention Network	78.45 111ms, 9hr
Hierarchical Convolutional Attention Network	78.14 35ms, 3hr



Computationally expensive!!!

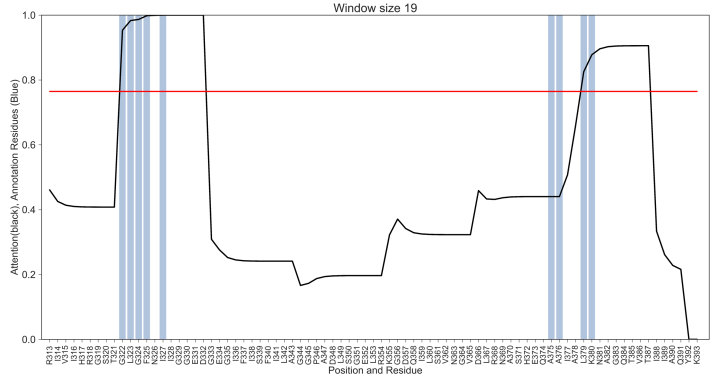


Gao, S., Ramanathan, A., in review (ACL)

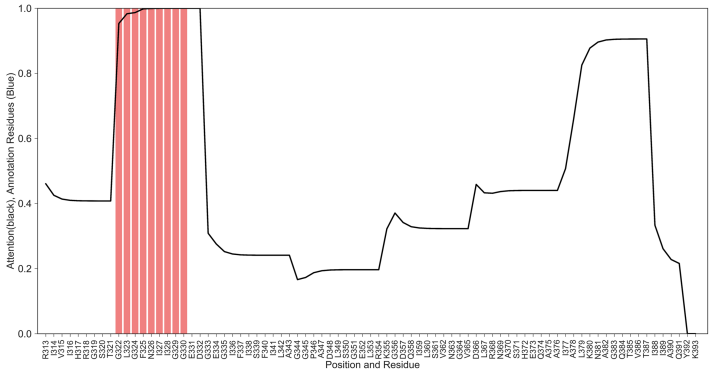
CAN THE H(C)AN BE USED ON OTHER TYPES OF DATA? E.G., PROTEIN ALIGNMENTS TO UNDERSTAND CO-EVOLUTIONARY MODULES

- Predict “hotspots” across protein sequence databases

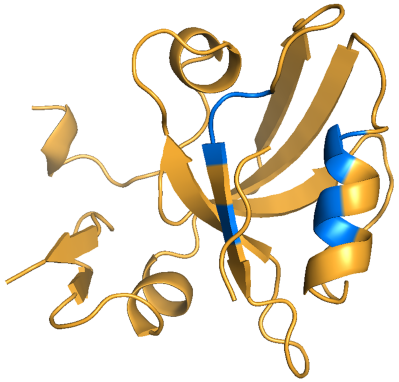
A Attention results for PDZ (PSD-95) for window size 9 and overalp 1



B Highest attention window for PDZ (PSD-95) for window size 9 and overalp 1



C Spatial location of annotated residues (blue)



D Spatial location of attention residues(red)

Protein Family	AUC (sequences)	F1 (sequences)	SCA AUC score	SCA F1 score
Cadherin	**0.568**	**0.817**	0.546	0.670
PDZ (NCBI)	**0.715**	**0.840**	0.520	0.753
PDZ (PFAM)	**0.660**	**0.827**	0.520	0.753
Tau	**0.555**	**0.643**	0.393	0.502
HSP70	0.510	**0.771**	**0.553**	0.709

Catanho, M., Gao, S., Ramanathan, A., Coleman, T. P., 2018 (submitted)

SUMMARY

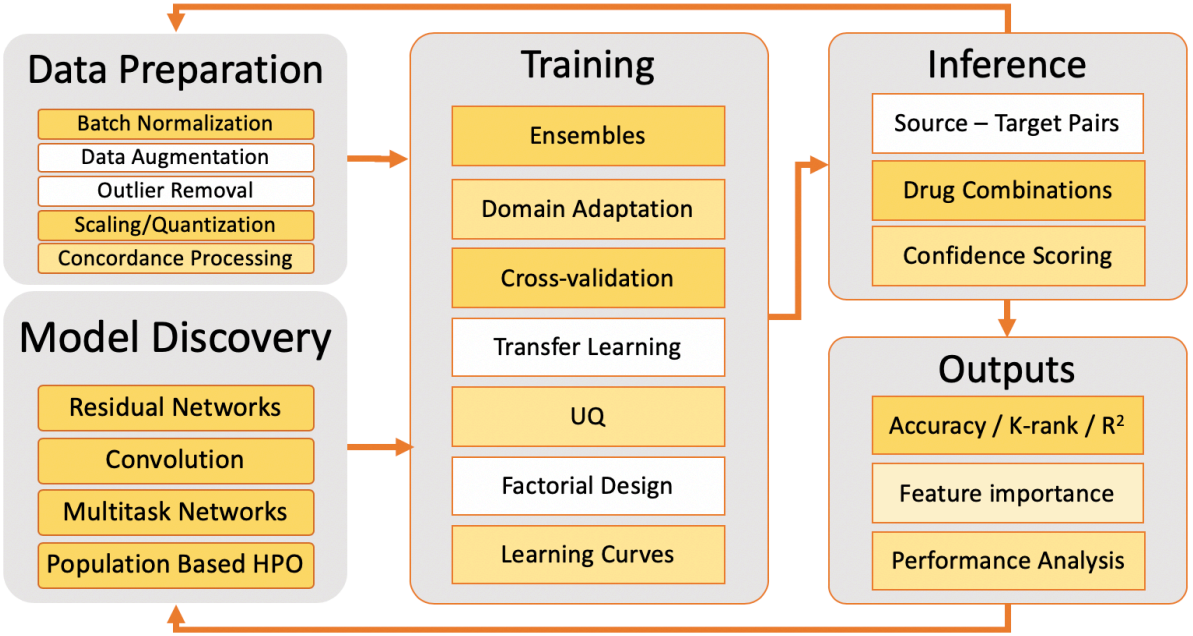
- Deep learning shows promise for automated information extraction from unstructured pathology reports to increase efficiency, data quality, and timeliness of cancer surveillance.
  - Cross-registry performance was robust across all tasks.
- Current DL NLP Work:
  - reportability de novo metastasis / recurrence
  - Privacy preserving sharing of DL NLP models

CANCER DISTRIBUTED (DEEP) LEARNING ENVIRONMENT (CANDLE)
EXASCALE COMPUTING PROJECT

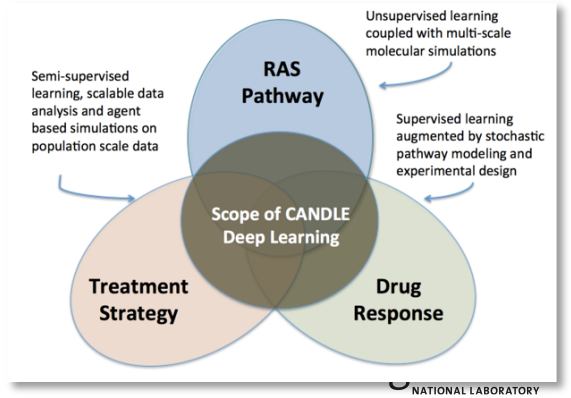
CANDLE: EXASCALE DEEP LEARNING TOOLS

Deep Learning Needs Exascale

- Automated model discovery
- Hyper parameter optimization
- Uncertainty quantification
- Flexible ensembles
- Cross-Study model transfer
- Data augmentation
- Synthetic data generation
- Reinforcement learning



https://github.com/ECP-CANDLE

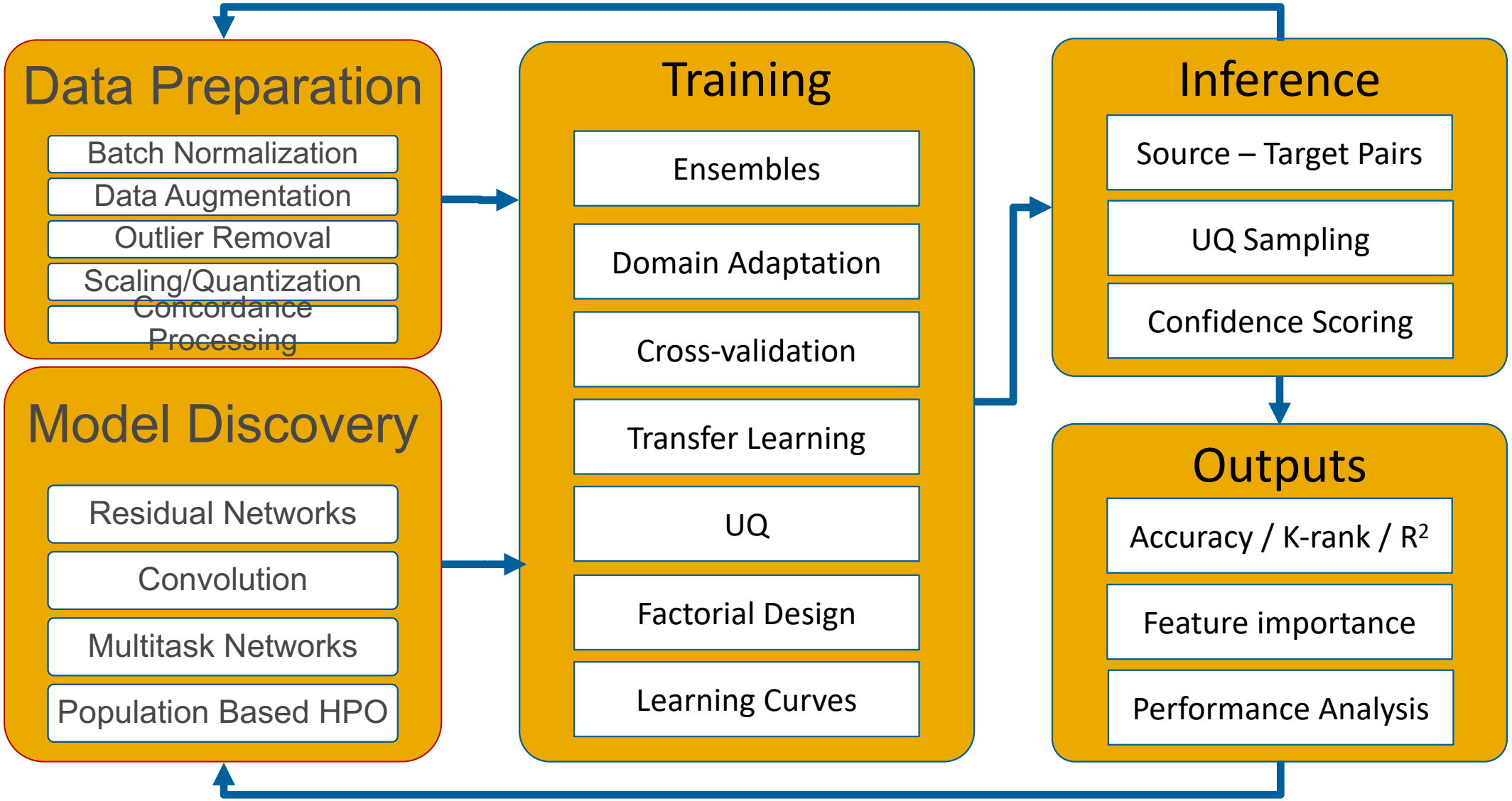


CANDLE PROJECT

- **ECP-CANDLE GitHub Organization:**
- https://github.com/ECP-CANDLE

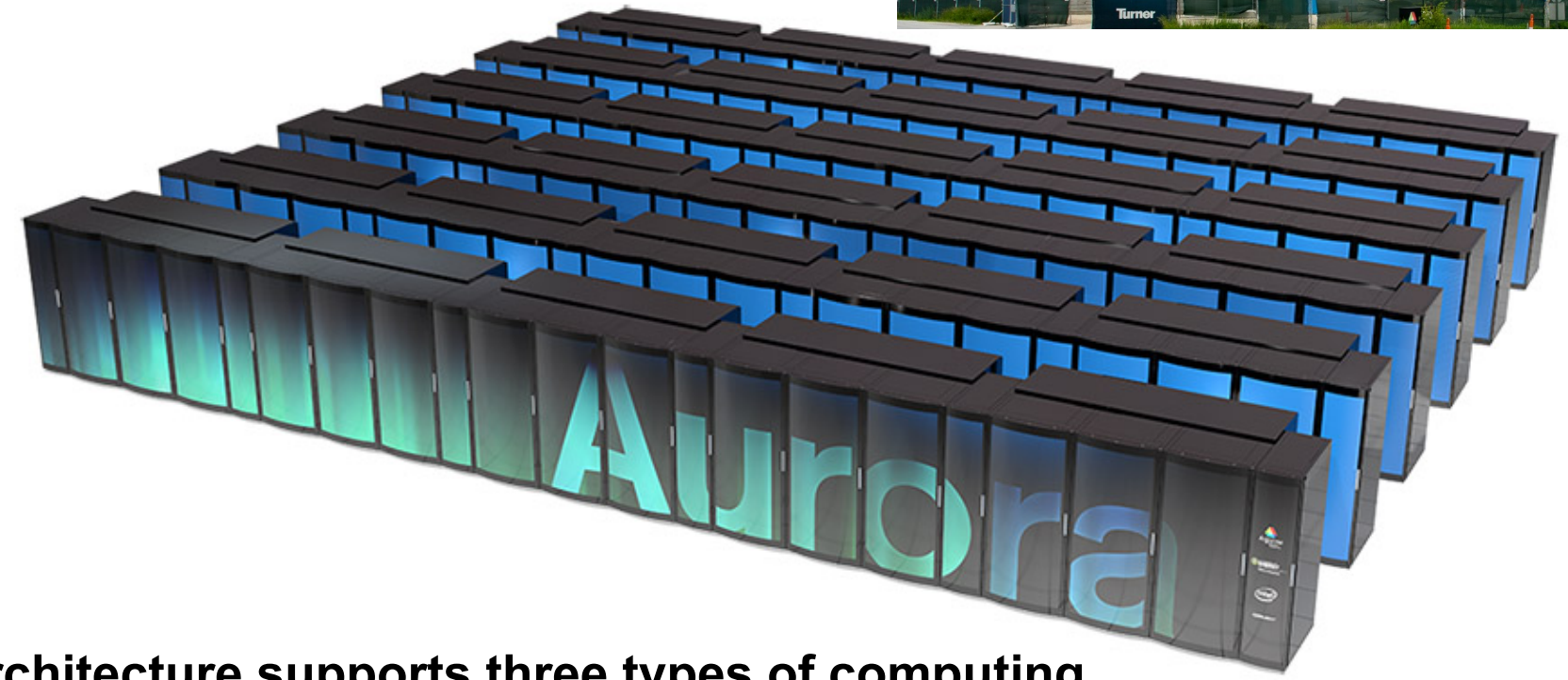
- **CANDLE Python Library** – make it easy to run on DOE Big Machines, scale for HPO, UQ, Ensembles, Data Management, Logging, Analysis
- **CANDLE Benchmarks** – exemplar codes/models and data representing the three primary challenge problems
- **Runtime Software** – Supervisor, Reporters, Data Management, Run Data Base
- **Tutorials** – Well documented examples for engaging the community
- **Contributed Codes** – Examples outside of Cancer, including Climate Research, Materials Science, Imaging, Brain Injury
- **Frameworks** – Leverage of TensorFlow, Keras, Horovod, PyTorch, etc.
- **LL Libraries** – CuDNN, MKL, etc. (tuned to DOE machines)

SCOPE OF CANDLE WORKFLOWS



AURORA: HPC AND AI

> ExaFlops/s for HPC
>> Exaops/s for AI



Architecture supports three types of computing

- Large-scale Simulation (PDEs, traditional HPC)
- Data Intensive Applications (scalable science pipelines)
- Deep Learning and Emerging Science AI (training and inferencing)

EXASCALE MACHINE TARGETS IN 2021/2022

- Aurora and Frontier are similar machines in that

1. Both are GPU accelerated x86 based nodes
2. ~10,000 nodes each with CPUs + GPUs
3. >> 10,000 GPUs (DP > 1 EF, HP > 10 EF)
4. Big Memories, including NVM and solid state storage
5. Lots of I/O bandwidth but < than the typical GB/GPU noticed by NVIDIA as sweet spot
6. Caching data will be important for DL training
7. Framework optimization for each flavor of GPU will be important (AMD vs Intel)
8. Both will have Cray OS environment, support for containers etc.
9. CANDLE is targeting both platforms

DEEP LEARNING USE CASES ON EXASCALE PLATFORMS

- Contrary to expectations it will be rare to run a single deep learning training model on the full system
- Individual Cancer problems as hard as they are are not (currently) big enough to efficiently use the full machine
- So while some problems will use pipelining, model parallelism, data parallelism to use perhaps 10% of the machine on one problem, the bulk of the use cases are for some type of ensemble
- This is fine as we have more than enough volume to keep an Exascale system busy

CURRENT PLATFORMS FOR HYPERPARAMETER OPTIMIZATION RELY ON SEQUENTIAL OPTIMIZATION TECHNIQUES

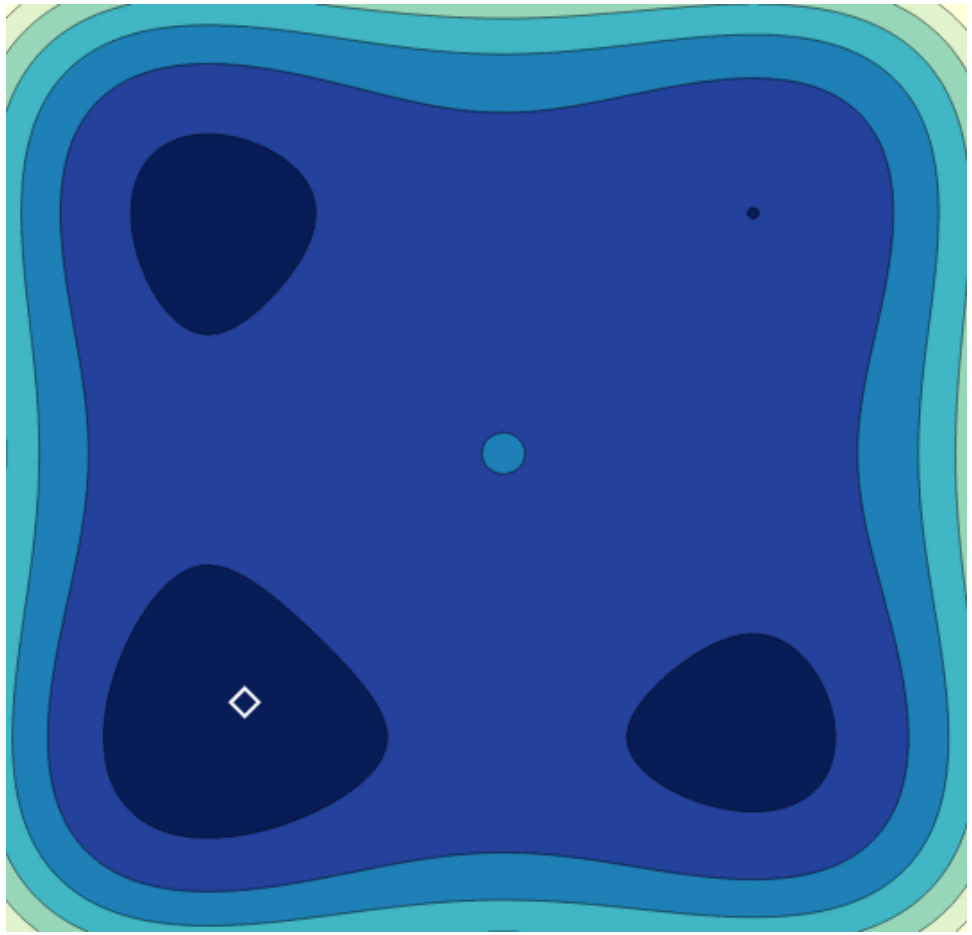
- Bayesian optimization, Bandit optimiza
  search processes
- Exponential scaling:
  - The number of samples required to
    optimization procedure is scales ex
    dimensions, as in 2^D, where D is nu
  - Forgotten in the recent excitement

HyperSpace: Distributed Parallel Bayesian Optimization

- Hyperspace, instead seeks to focus on the search space:
  - Parallelism to exploit the statistical structure of the search space
  - Reveal partial dependencies across parameter spaces
- Build many surrogate functions in parallel
- {Prayer}!

N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger. Gaussian pr
abs/0912.3995, 2009.
S. Grunewalder, J.-Y. Audibert, M. Opper, and J. Shawe-Taylor. Regre
Thirteenth International Conference on Artificial Intelligence and Statis
2010.

HYPERSPACE: PARALLEL EXPLORATION OF LARGE SEARCH SPACES



1. Define the bounds of each hyperparameter search space.

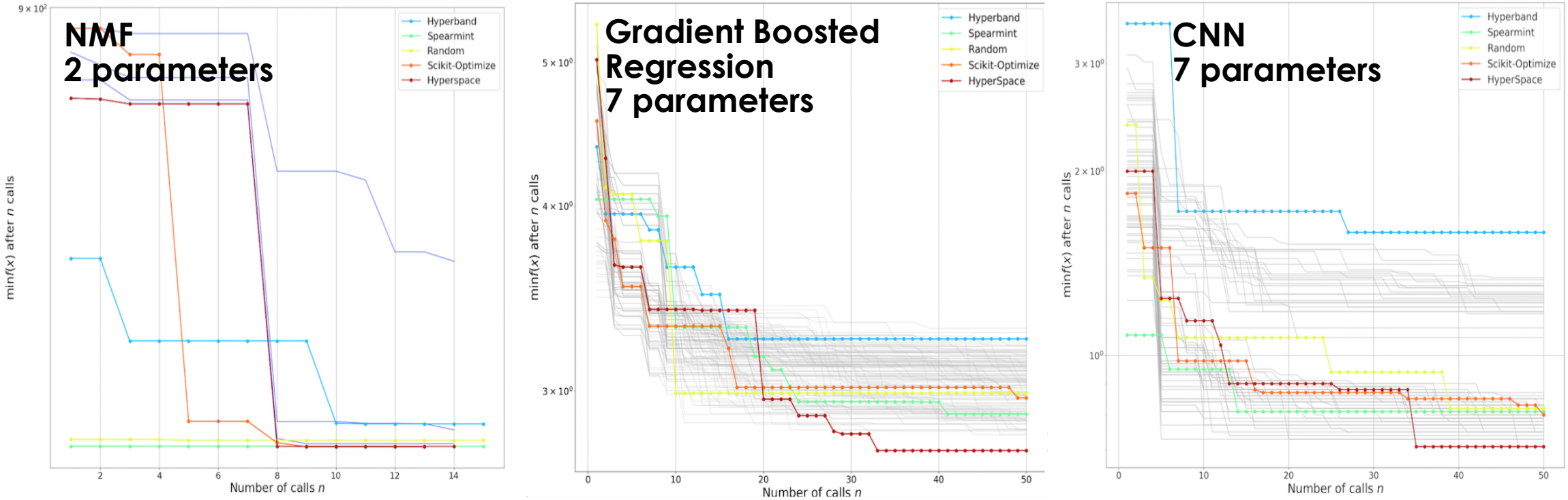
2. Divide each search space bound into two nearly equal sub-bounds with overlap ϕ, where $\{\phi \in \mathbb{R} \mid 0 \leq \phi \leq 1\}$.

3. Create all possible combinations of hyperparameter sub-bounds to form 2^D search spaces (hyperspaces) where D is the number of model hyperparameters.

4. Run Bayesian optimization over each hyperspace in parallel

M. Todd Young, J. D. Hinkle, R. Kannan, A. Ramanathan, HyperSpace: Massively Parallel Bayesian Optimization, Workshop on High Performance Machine Learning, 2018, Lyon, France

HYPERSPACE CAN OPTIMIZE MANY DIFFERENT ML/DL

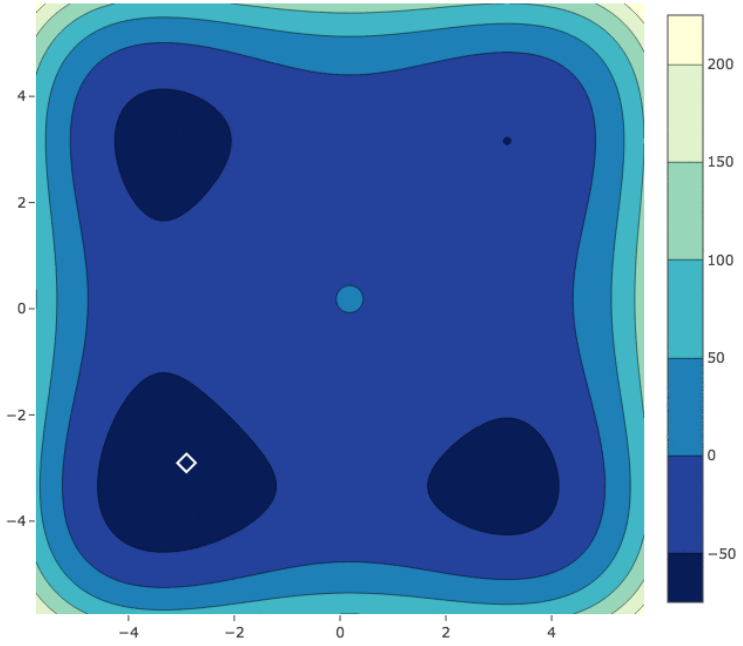


HyperSpace can effectively scale across supercomputing resources to reveal how models perform under many unique hyperparameter configurations.

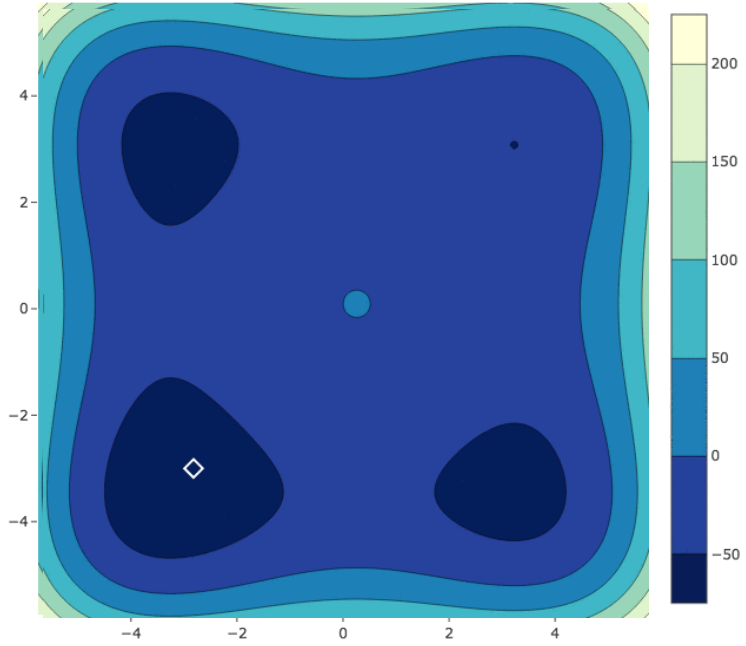
- Discovers regions in the hyperparameter search space where models perform well *and* where they perform poorly
- Finds families of solutions where various settings of hyperparameters perform equally well
- Opens the possibility of meta learning for hyperparameter optimization (future direction)

M. Todd Young, J. D. Hinkle, R. Kannan, A. Ramanathan, HyperSpace: Massively Parallel Bayesian Optimization, Workshop on High Performance Machine Learning, 2018, Lyon, France

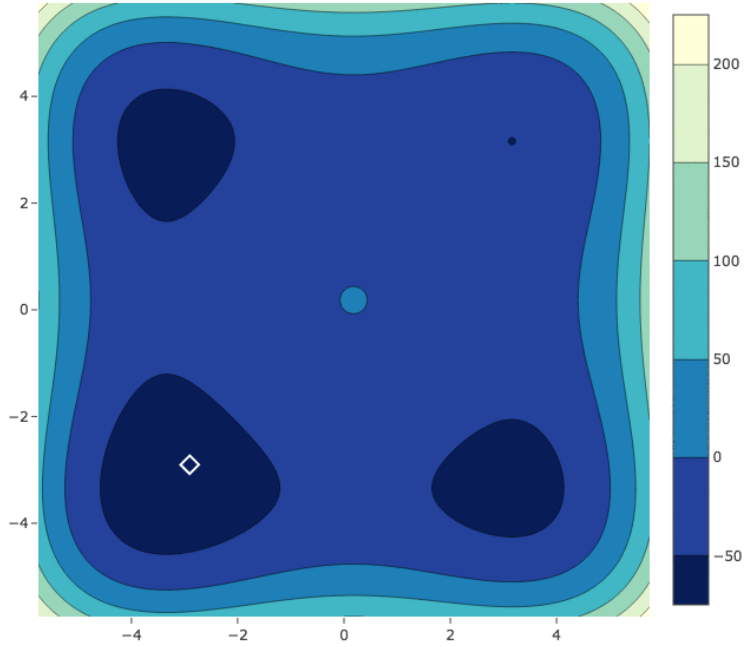
PARALLEL EXPLORATION OF LARGE SEARCH SPACES WORKS BETTER THAN RANDOM/ SEQUENTIAL BASED OPTIMIZATION



HyperSpace



Random search



SMBO

HOW ARE WE USING LARGE-SCALE COMPUTING?

- **Deep Sweeps on Features/Feature Combinations**
  - Recently ran 16K model jobs on Summit (Pilot1)
- **Hyperparameter Optimization (full machine runs)**
  - Tuning model settings (Big runs on Cori, Theta, Summit, Titan)
- **Neural Architecture Search (Model Discovery)**
  - Big runs on Theta (SC19 Paper)
- **Hierarchical "LOOCV" Cross Validation Study (Exascale CP)**
  - Bayesian approach to online learning (accelerated convergence)
- **Data Augmentation and Generative Networks**
  - Exploring strategies for "Low Data" learning
- **Uncertainty Quantification**
  - Bootstrapping, parameter sweeps
- **Data Scaling Studies (learning curve estimates)**
  - Accuracy and Error as a function of data scale

BYOM! BRING YOUR OWN MODELS …

- CANDLE Hackathons:
  - Nov 11-15 at Argonne
- Goals:
  - enable one to build CANDLE compliant code for your models
  - test runs on Theta (current supercomputer @ Argonne), Summit (ORNL), and other test platforms
  - have fun!
- What to bring?
  - bring your models in either Keras/Tensorflow, Pytorch (less supported currently but can be built and supported)
- We are always looking for examples other than cancer datasets!
  - Imaging, NGS, pharmacogenomics, neuroscience, structural biology, etc.

THANK YOU!!!
QUESTIONS/COMMENTS?