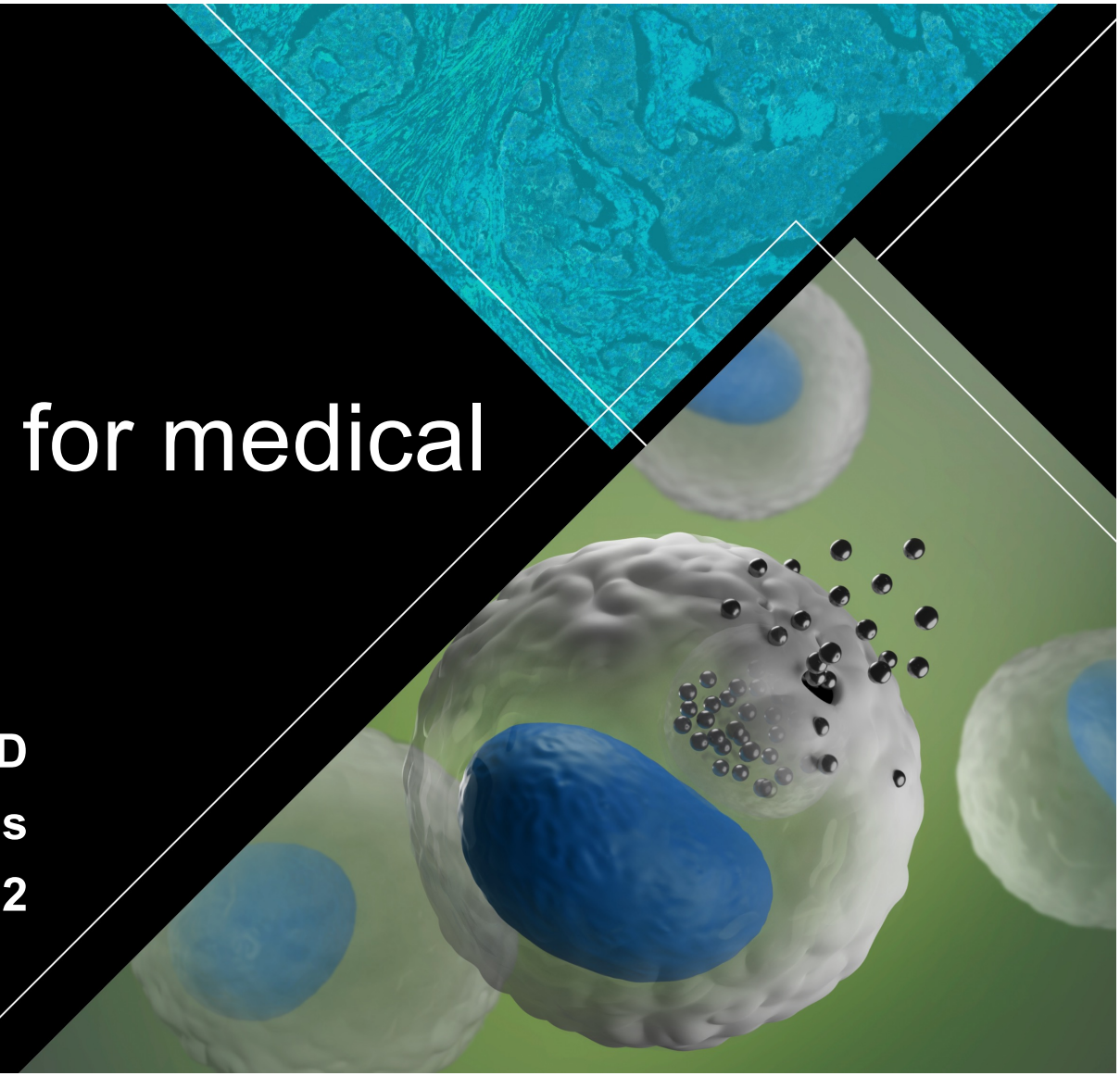
Frederick National Laboratory for Cancer Research

IMPROVE-ing AI for medical applications

Ryan Weil, PhD

Cancer Data Science Initiatives

October 27, 2022



SPONSORED BY THE NATIONAL CANCER INSTITUTE

NCI-DOE Collaboration

Strategic interagency collaboration created in 2016 to accelerate cancer research using emerging exascale computing capabilities.

Argonne NATIONAL LABORATORY

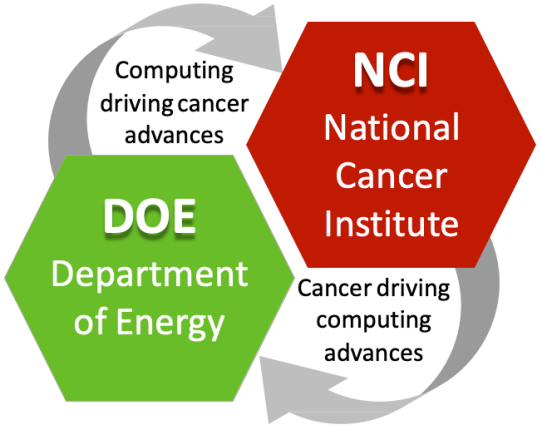
Frederick National Laboratory for Cancer Research

NIH NATIONAL CANCER INSTITUTE

Los Alamos NATIONAL LABORATORY

Lawrence Livermore National Laboratory

OAK RIDGE National Laboratory



AI *is* going to the clinic!

- ***Software as a medical device*** **has an expected compounded annual growth rate (CAGR) of 21.9% during 2020–2027 and will be a >$85B a year area.**
  - ◇Many of these products are expected to have an AI component
  - ◇FDA with other regulatory agencies has released Good AI/ML practice guidance similar to GMP/GCP*.
- ***AI as a biomarker/companion diagnostic*****–especially with mIoT devices–is already being considered for the treatment selection**
- **There are many challenges to pervasive AI at the edge in medical applications.**

*https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

Barriers to adoption of AI at the edge

Soft problems

- **Liability/Insurance**
  - The problem is as hard or harder than liability for self driving.
- **Data**
  - For most applications in medical AI the training data sets are minuscule compared to the complexity of the data.
- **Control**
  - The mindset has to change in care giver community and patients.
- **Cost**
  - Will the payers determine the value of the prediction is worth the computational and development cost.
  - Are the data required as inputs (tests, sensors, etc) in line with the value of the prediction.

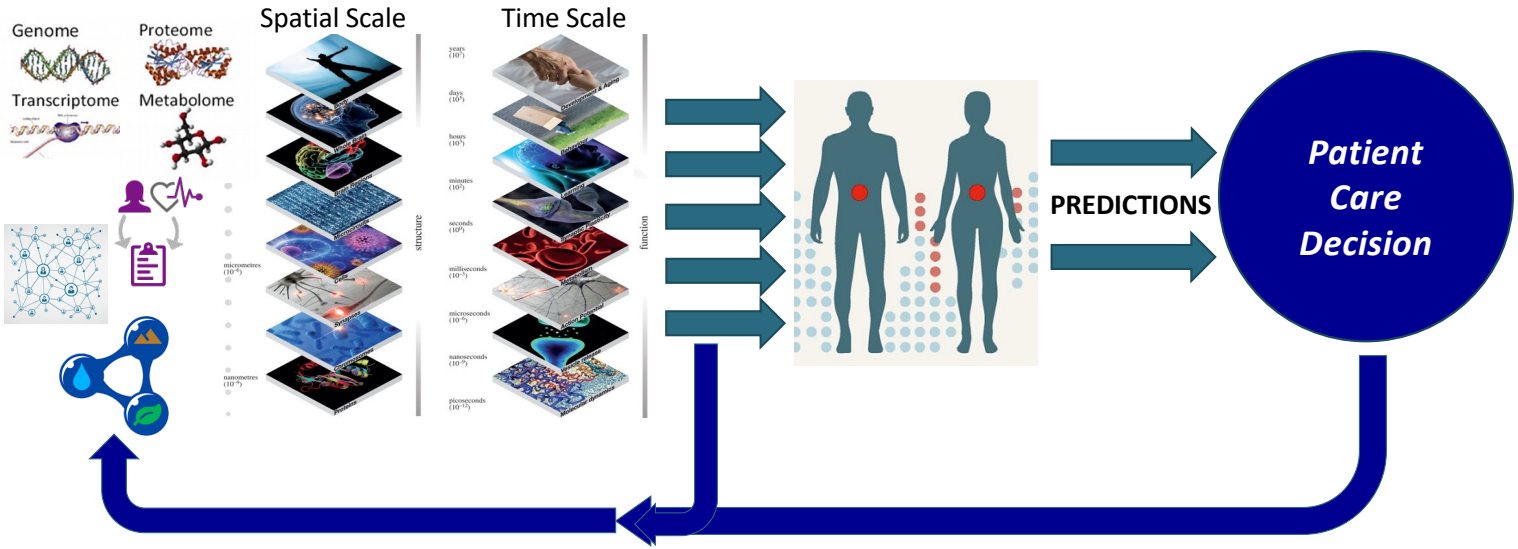
The biosciences are behind physics

- **Physics informed AI is maturing**
  - It is relatively straight forward for an AI to rederive Bernoulli's equations from lots of training data.
  - Data for physical/mechanical systems are far more readily available.

- **Biomedically informed AI is in its infancy**
  - Fitness watches can identify arrythmia and insulin pumps can predict dosing, but those use limited data and have a singular output as well as manual override.
  - AI for pathology (whole slide imaging) is maturing and Paige.ai has approved products for **aiding** pathologists in diagnosing cancers.
  - Digital twins are being developed but there is a lot of work to be done.

Digital Twin for Predictive Oncology

Patient-tailored models incorporating multi-omic, clinical, environmental and social data that can evaluate and predict the most effective prevention and therapeutic plans



Hernandez-Boussard, T. *et al.* Digital twins for predictive oncology will be a paradigm shift for precision cancer care. *Nat Med* **27,** 2065–2066 (2021). https://doi.org/10.1038/s41591-021-01558-5

Barriers: Data acquisition and use

- **IRBs**
  - IRBs are not generally well setup to assess the *informational risk* to patients.
- **Consent for altruistic reasons is no longer a viable model**
  - Leads to bias since there are significant differences based on race, gender, age, condition, etc.
  - This leads to greater benefits for people that are well represented in the training data.
- **HIPAA**
  - Data sharing laws were not set up for supporting medical AI at scale.
  - Statistical and safe harbor deidentification don't stand up to analysis at scale.

This slide attempts to summarize a talk by Barbara Evans at U of F on Sustainable AI in Clinical Care

Regulatory considerations

- **Depending on their purpose AI models can be considered *in vitro* diagnostic devices and therefore subject to 21 CFR 812 (Investigational Device Exemptions) and require a 510k and post market surveillance in order to be marketed.**

- **Significant risk means an investigational device that:**
  - (1) Is intended as an implant and presents a potential for serious risk to the health, safety, or welfare of a subject;
  - (2) Is purported or represented to be for a use in supporting or sustaining human life and presents a potential for serious risk to the health, safety, or welfare of a subject;
  - ***(3) Is for a use of substantial importance in diagnosing, curing, mitigating, or treating disease, or otherwise preventing impairment of human health and presents a potential for serious risk to the health, safety, or welfare of a subject; or***
  - ***(4) Otherwise presents a potential for serious risk to the health, safety, or welfare of a subject.***

(21 CFR 812.3(m)

More things to consider from 21 CFR 812

- **Is deidentified data still considered human subjects?**
  - ***Subject*** means a human who participates in an investigation, either as an individual on whom or on whose specimen an investigational device is used or as a control. A subject may be in normal health or may have a medical condition or disease. (21 CFR 812.3(p)

- **Error handling is higher impact**
  - ***Unanticipated adverse device effect*** means any serious adverse effect on health or safety or any life-threatening problem or death caused by, or associated with, a device, if that effect, problem, or death was not previously identified in nature, severity, or degree of incidence in the investigational plan or application (including a supplementary plan or application), or any other unanticipated serious problem associated with a device that relates to the rights, safety, or welfare of subjects. (21 CFR 812.3(s)

The more immediate problem

- **Approval for medical devices normally requires proving effectiveness and superiority (or at least non-inferiority).**

- **How is this possible when there is:**

  No standard benchmark *for testing or training*

  No standard performance measures *for evaluation*

  No clear consensus approach or method

  No clear human baseline *to compare with*

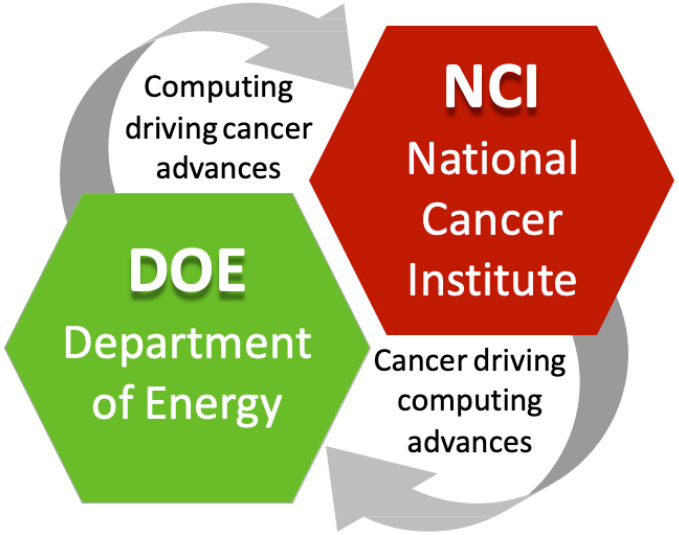
There is an opportunity to close some of the gaps discussed above.

IMPROVE: Innovative Methodologies and New Data for Predictive Oncology Model Evaluation

Rick Stevens
Argonne National Laboratory/University of Chicago

Jeff Hildesheim
Division of Cancer Biology
National Cancer Institute

Ryan Weil
Frederick National Laboratory for Cancer Research



Anticipated *Impacts* of IMPROVE

Closing Gaps in the development and application of deep learning models for predictive modeling of therapeutic response, including:

- **Well-curated, clinically relevant, standardized training and testing datasets**
- **Standardized, easily-applicable workflow (including software pipeline, performance metrics, data, etc.) for evaluating and comparing prediction models to drive model improvement**
- **Understanding the model attributes related to predictive power, interpretability, and uncertainty quantification (including errors and failure to predict and how this is handled)**
- **Engaging the community for expert opinions and collaborations on developing a model evaluation framework and generating benchmark data**

Creating Approaches for evaluating and improving modeling that are intended to be generalizable to deep learning models in other domains in NCI and DOE

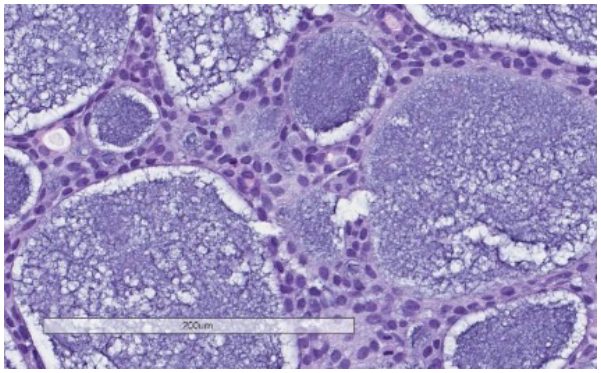
- ◇ E.g., materials design, HPC surrogates

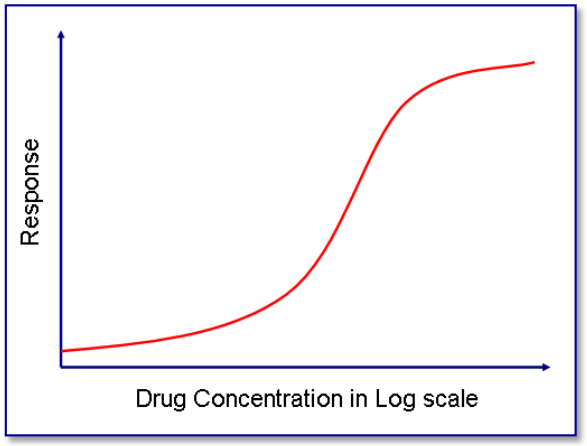
Generating New Hypotheses and identifying previous hidden cancer types and treatment targets.

Aims of IMPROVE Project

- **Two related aims with the shared goal of IMPROVING deep learning models for predicting drug response in tumors:**
  - **Aim 1: IMPROVE *Models*:** Development of semi-automatic protocols for **comparing deep learning models** and **identifying model attributes** that contribute to prediction performance.
  - **Aim 2: IMPROVE *Data*:** Development of protocols for **specifying drug screening experiments and generating new data**

Data Driven Modeling of Cancer Drug Response



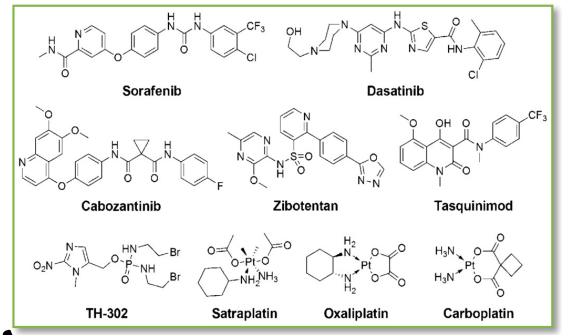


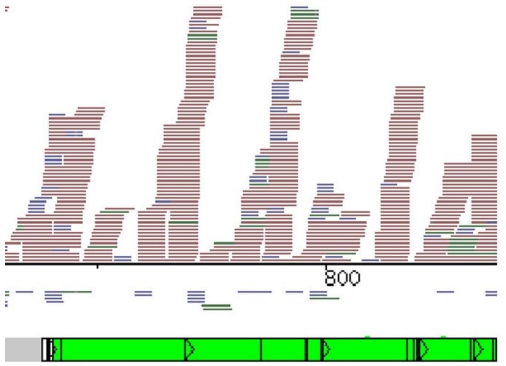
Drug (s)
descriptors
fingerprints
structures
SMILES
dose

$$R = f(T, D, [PI^*])$$

IC50
AUC
GI50
% growth
Z-score
Response

gene expression levels
SNPs
protein abundance
microRNA
methylation
Tumor





[PI] (patient /treatment history, etc.)*

Components that define the predictive power of a model

How much of the predictive power of a given model is due to the **structure and nature of the model** itself vs. the **quality and coverage of the data** the model is trained and tested on?

Our approach focuses on addressing **two key bottlenecks:**

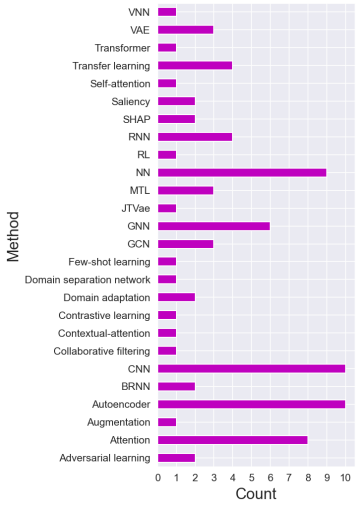
1. **Changing the model structure and tuning hyperparameters**
2. **Improving the datasets used for training and testing**

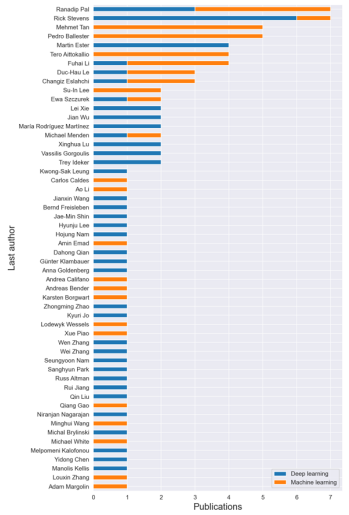
With broad community involvement to improve transferability

IMPROVE Aim 1: Evaluation and Comparison of Drug Response Prediction Models

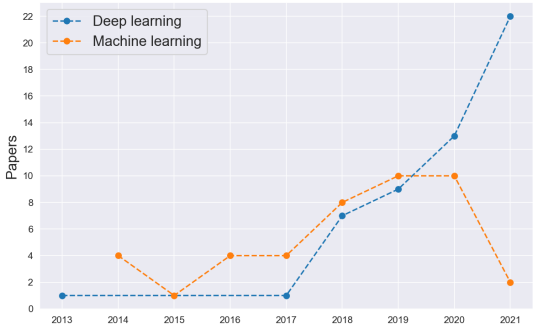
Comprehensive literature survey to collect information about research groups and models (ongoing task)

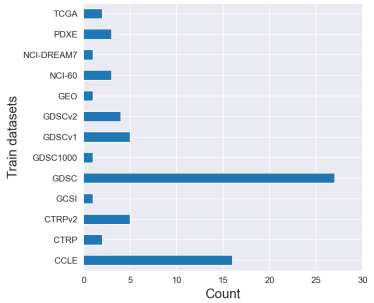
- **> 100 papers about machine/deep learning drug response prediction**
- **Categorize models according multiple criteria to select representative ones for comparison study**
  - Model architecture and technique
  - Functionality, e.g. transfer learning, interpretability and uncertainty quantification
  - Code availability and documentation
  - Training and validation data

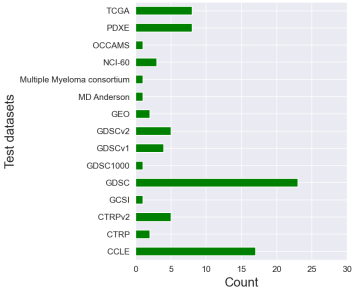




Adapt and modify code to train and test the models, and conduct <u>reproducibility analysis</u>







There's a lot already out there

First name	Last name	Institute	Expertise	Email	Link	Recommendation	First author	Paper	Year	Publisher	Primary outputs	Code	Framework	Drug set	Model name	Cancer features	Drug features	Train bio model	Test bio model	C/W/CW	DL methods	ML methods	Baselines	UQ	XAI
Anna	Goldenberg	Associate	Anna	anna.go	http://	Yes, it would be great to have at least one of them as	L. Rampášek	Dr.VAE:	2019	Bioinformatics	AAC	Github	PyTorch	Single	Dr.VAE	Gene expression	None	CTRPv2	CTRPv2	W	VAE	None	RidgeLR, RF,		
Bernd	Freisleben						S. Daoud	Q-Rank:	2020	IEEE Journal of Biomedical and	GI50	Github	R			Gene expression	PaDEL fingerprints	NCI-DREAM7			RL	None	IntegratedMR		
Changiz	Eslahchi	Professor,	He is focussing	ch-	http://	Maybe. He is working on various aspects of cancer	A. Emdadi	Auto-	2021	BMC Bioinformatics	IC50	Github	NumPy		Auto-HMM-LMF						Autoencoder	HMM, Matrix			
Dahong	Qian	Professor in		dahong	https/	No.	Z. Zuo	SWnet: a	2021	BMC Bioinformatics	IC50	Github	PyTorch	Single	SWnet	Gene expression,	Morgan	CCLE, GDSC	CCLE, GDSC	W	GNN, Saliency	None	CDRscan,		
Duc-Hau	Le	Associate	He is working	hauldh	https/	Maybe. He has started working on drug sensitivity	T. Nguyen	Graph	2020	IEEE/ACM Trans Comput Biol	AUC, IC50	Github	PyTorch	Single	GraphDRP	CNA	Molecular graphs	GDSC	GDSC	W	Attention,	None	tCNNS		Yes
Ewa	Szczurek	Assistant	Her main	szczure	https/	No. Her new group has recently started working on	K. Koras	Interpret	2021	Scientific Reports	AUC, IC50	Github	PyTorch	Single	DEERS	mRNA expression,	Kinase inhibition	GDSC	GDSC	W	Autoencoder	None	ElasticNet,		Yes
Fuhai	Li	Assistant	Fuhai Li is	fuhai.li	https/	Yes. Although his group is not big nor very established,	T. Zhang	Synergist	2021	Methods Mol Biol		No	NA	Combo	AuDNNsynergy						Autoencoder	None	DeepSynergy,		
Günter	Klambauer	Assistant	Tenure-Track	klamba	https/	Maybe. Although he only has one paper that applies DL	K. Preuer	DeepSyn	2018	Bioinformatics		Github	TF1	Combo	DeepSynergy	Gene expression	Morgan				NN	None	ElasticNet,		
Hojung	Nam	Associate	Professor since	hjnam	https/	Maybe	I. Jin	HiDRA:	2021	Journal of Chemical Information	IC50	Github	TF1	Single	HiDRA	Gene expression	Morgan	GDSC1000	CCLE, GDSC1, GDSC2,	CW	Attention,	None	CDRscan,		Yes
Hyunju	Lee	Professor,	Computationa	hyunjul	https/	No, drug response prediction is not her main research	S. Park	Super.FE	2021	BMC Bioinformatics	IC50, binary	Github	PyTorch	Single	Super.FELT	Copy number, Gene	None	GDSC	CCLE, CTRP, PDXE, TCGA	CW	Autoencoder	None	AutoBorutaR		
Jae-Min	Shin	Yongin in silico				No. They have not done any other research related to	Y. Chang	Cancer	2018	Scientific Reports	IC50	No	TF1	Single	CDRscan	Mutation	PaDEL descriptors	CCLP, GDSC	CCLP, GDSC		CNN	None	RF, SVM		
Jian	Wu	Zhenjian	Artificial	wujian	https/		Y. Zhu	TGSA -	2021	Bioinformatics	IC50	Github	PyTorch	Single	TGSA	CNV, Gene	Molecular graphs	GDSCv2	GDSCv2	W	Attention,		CDRscan,		
Jianxin	Wang	Professor,	His main focus	jxwang	http://	No. He has not done any other work related to drug	M. Li	DeepDSC	2019	IEEE/ACM Transaction of	IC50	No	TF1	Single	DeepDSC	Gene expression	Morgan	CCLE, GDSC	CCLE, GDSC	W	Autoencoder	None	DNN, KBMF,		
Kwong-Sak	Leung	Professor,	His main focus	ksleung	http://	Maybe. He is working on drug-drug interaction, drug	Pengfei Liu	Improvin	2019	BMC Bioinformatics	IC50	Github	TF1 (w/o Keras)	Single	tCNNS	Mutation	SMILES	GDSC	GDSC	W	CNN	None			
Kyuri	Jo	Assistant	She has	kyurijo	https/	Maybe. She has two papers about drug response	S. Kim	Graph	2021	Mathematics	IC50	Github	TF1 (w/o Keras)	Single	DrugGCN	Gene expression	Graph structure	GDSC	GDSC	W	GCN	None	Bagging		
Lei	Xie	Professor,	He has a	Lei.Xie	http://	Yes. His group has 2 papers of drug response prediction	Di He	Robust	2021	bioRxiv		Github	PyTorch	Single	CODE-AE						Domain	None	Simpler		
Lei	Xie	Professor,	He has a	Lei.Xie	http://	Yes. His group has multiple papers of drug response	Q.Liu	TranSyne	2021	PLoS Computational Biology		Github	PyTorch	Combo	TransSynergy	Gene dependency,	Network				Attention,	None	DeepSynergy		Yes
Manolis	Kellis	MIT, USA				It seems that these are students.	H. Dong	Variation	2021	ICLR 2021	IC50	Github	PyTorch	Single	GENEVAE	Gene expression	SMILES	CCLE, GDSC	CCLE, GDSC	W	JTVae, VAE	None	MLP, SVR		
Maria	Martinez	IBM Research,	Maria joined	MRM@	https/		A. Oskooei	PaccMan	2018	ICML	IC50	Yes	TF1/PyTorch	Single	PaccMann	Gene expression	Morgan	GDSC	GDSC	W	Attention,	None	Simplifie		
Maria	Martinez	IBM Research,	Maria joined	MRM@	https/	Maybe. She has two pulications about drug response	M. Manica	Toward	2019	Molecular Pharmaceutics	IC50	Yes	TF1/PyTorch	Single	MCA	Gene expression	Morgan	GDSC	GDSC	W	Attention,	None	Simplifie		
Martin	Ester	Professor of	His research	ester@	http://	Yes. Martin Ester is continuously developing novel deep	H. Sharifi-Noghabi	MOLI:	2019	Bioinformatics	IC50, binary	Github	PyTorch	Single	MOLI	Gene expression	None	GDSCv1	TCGA, PDXE	CW	Autoencoder	None	DNN		
Martin	Ester	Professor of	His research	ester@	http://	Yes. Martin Ester is continuously developing novel deep	H. Sharifi-Noghabi	AITL:	2020	Bioinformatics	IC50, binary	Github	PyTorch	Single	AITL	Copy number, Gene	None	GDSCv1	TCGA, PDXE	CW	Adversarial	None	ADDA, MOLI,		
Martin	Ester	Professor of	His research	ester@	http://	Yes. Martin Ester is continuously developing novel deep	H. Sharifi-Noghabi	Out-of-	2021	Nature Machine Intelligence	AAC	Github	PyTorch	Single	Velodrome	Gene expression	None	CTRPv2, GDSCv2	CTRPv2, GDSCv2, GEO,	CW	Transfer	None	DeepAll-ERM,		
Martin	Ester	Professor of	His research	ester@	http://	Yes. Martin Ester is continuously developing novel deep	O. Snow	Interpret	2021	KDD	AAC	Github	TF2	Single	BDKANN+	Gene expression	None	GDSCv2, CTRPv2	GDSCv2, CTRPv2	CW	SHAP	None			Yes
Melpomeni	Kalofonou	Research Fellow,	Her current	m.kalof	https/	No. She has not done any other work related to drug	Z. Zhao	A	2019	IEEE Biomedical Circuits and	ActArea	No	NA	Single		Gene expression	None	CCLE	CCLE	W	CNN, RNN	None	ElasticNet,		
Michael	Menden	Junior Group	He is a Junior	michael	https/	Maybe. Michael P. Menden is a backup candidate for	M. Menden	Machine	2013	Plos ONE	IC50	No	Java	Single		Copy number,	PaDEL descriptors	GDSC	GDSC	W	NN	None			
Michal	Brylinski	Associate	His main focus	mbrylin	https/	No. He has not done any other work related to drug	M. Singha	GraphGR	2020	bioRxiv	IC50 to binary	Github	PyTorch	Single	GraphGR,			CCLE	CCLE		GNN	None			
Niranjan	Nagarajan	Senior Group	Focuses on	nagaraj	http://	No, drug response prediction is not her main research	RP da Silva	TUGDA:	2021	Bioinformatics	IC50, binary	Github	PyTorch	Single	TUGDA			GDSC	GDSC, PDXE, TCGA	CW	Autoencoder,	None	ElasticNet,	Yes	Yes
Qin	Liu	Professor,	Her main	qin.liu	https/	No. She has not done any other work related to drug	Q. Li	Predictio	2020	IEEE Annual Computing and	IC50	No	TF	Single	MFNN	Gene expression		GDSC	GDSC	W	CNN, RNN	None	RF, SVM		
Ranadip	Pal	Professor,	Machine	ranadip	https/	Yes. His group is continously working on drug	K. Matlock	Investiga	2018	BMC Bioinformatics		No	NA								NN	RF, KNN			
Ranadip	Pal	Professor,	Machine	ranadip	https/	Yes. His group is continously working on drug	SR Dhruba	Applicati	2018	BMC Bioinformatics	AUC	Github	Matlab			Gene expression	None	CCLE, GDSC	CCLE, GDSC	CW	Transfer	None	Different		
Ranadip	Pal	Professor,	Machine	ranadip	https/	Yes. His group is continously working on drug	O. Bazgir	Represen	2020	Nature Communications	IC50	Github	TF	Single	REFINED	Gene expression	Descriptors	GDSC, NCI-60	GDSC, NCI-60	W	CNN	None	DNN,		
Rick	Stevens	ANL					F. Xia	Predictin	2018	BMC Bioinformatics															
Rick	Stevens	ANL	Machine	stevens		Yes. His group is continously working on drug resp	A. Clyde	A	2020	arXiv	AUC	Github	PyTorch	Single	None	Gene expression,	Descriptors	CCLE, GDSC, NCI60	CCLE, GDSC, NCI60	W	Attention, NN	None	None		
Rick	Stevens	ANL					Y. Zhu	Ensemble	2020	Scientific Reports	AUC	No	TF1	Single		Gene expression	Descriptors	CCLE, GCSI, CTRP,	CCLE, GCSI	CW	Transfer	LightGBM	LightGBM,		
Rick	Stevens	ANL					A. Partin	Learning	2021	BMC Bioinformatics	AUC	Github	TF2	Single	None	Gene expression	Descriptors	CTRPv2, GDSC1,	CTRPv2, GDSC1, GDSC2,	W	NN	None	LightGBM		
Rick	Stevens	ANL					F. Xia	A cross-	2021	Briefings in Bioinformatics				UNO											
Rick	Stevens	ANL					Y. Zhu	Converti	2021	Scientific Reports	AUC	Github	TF1	Single	IGTD	Gene expression	Descriptors	CTRP, GDSC	CTRP, GDSC		CNN	None	LightGBM,		
Rui	Jiang	Beijing National	Computationa	ruijiang	http://	No. Drug response prediction is not the main research	Q. Liu	DeepCDR	2020	Bioinformatics	IC50, binary	Github	TF1, DeepChem	Single	DeepCDR	DNA methylation,	Molecualr graph	CCLE, GDSC	TCGA	CW	CNN, GCN	None	CDRscan,		
Russ	Altman	Stanford, USA		russ.alt	https/		Y. Jiang	DrugOrc	2021	bioArxiv	IC50, ActArea	Github	PyTorch	Single	DrugOrchestra	Gene expression	Pre-trained	GDSC, CCLE, PDXE	GDSC, CCLE, PDXE	CW	TML	None	RF, ST-NN,		
Sanghyun	Park	Yonsei	Research	sanghy	http://	No. Drug response prediction is not the main research	J. Choi	RefDNN:	2020	Scientific Reports	IC50 to binary	Github	TF1 (w/o Keras)		RefDNN	Gene expression	Structure	CCLE, GDSC	CCLE, GDSC	W	NN	ElasticNet	CaDRRe5,		
Seungyoon	Nam	Gachon		nams@	https/	No. Drug response prediction is not the main research	M. Joo	A Deep	2019	International Journal of	IC50	Github	TF		DeepIC50	Mutation	Fingerprints	CCLE, GDSC	CCLE, GDSC	W	CNN	None	Ridge		
Trey	Ideker	Professor in the	Has a large	tideker	https/	Yes. He has a large group working on a variety of	B. Kuenzi	Predictin	2020	Cancer Cell	AUC	Github	PyTorch	Single, Combo	DrugCell (VNN)	Mutations	Morgan	CTRPv2, GDSCv1,	CTRPv2, GDSCv1, PDXE	CW	VNN	None	DNN		Yes
Trey	Ideker	Professor in the	Has a large	tideker	https/	Yes. He has a large group working on a variety of	J. Ma	Few-shot	2021	Nature Cancer	AUC	Github	PyTorch	Single	TCRP	Gene expression,	None	CCLE, GDSCv1, PDXE	CCLE, GDSCv1, PDXE	CW	Few-shot		KNN, LR, RF		
Vassilis	Gorgoulis	Professor,	His main	vgorg@	https/	Maybe. His big group is involved in diverse research	K. Vougas	Deep	2017	bioRxiv		Github	R			Gene expression	None				NN				
Vassilis	Gorgoulis	Professor,	His main	vgorg@	https/	Maybe. His big group is involved in diverse research	T. Sakellaropoulo	A Deep	2019	Cell Reports	IC50	Github	R	Single		Gene expression	None	GDSC	OCCAMS, MD Anderson,		NN	None	ElasticNet, RF		
Wei	Zhang	Assistant	His research	wzhang	https/	No. His lab is relatively new and has not done research	KT Ahmed	Network-	2020	BMC Medical Genomics	AUC, ED50	Github	PyTorch	Single		Gene expression	None	NSCLC	NSCLC		GNN	None	ElasticNet,		
Wen	Zhang	Huazhong		zhangw	http://		X. Liu	GraphCD	2021	Briefings in Bioinformatics	IC50 to binary	Github	PyTorch		GraphCDR	DNA methylation	Molecular graphs	GDSC	CCLE, GDSC	CW	Attention,		DeepCDR,		
Xinghua	Lu	Professor of	Interests in	xinghua	https/	Yes. Xinghua Lu has two publications of using deep	M. Ding	Precision	2018	Molecular Cancer Research		No	Matlab								Autoencoder	None			
Xinghua	Lu	Professor of	Interests in	xinghua	https/	Yes. Xinghua Lu has two publications of using deep	Y. Tao	Predictin	2020	Proceedings of Machine Learning	ActArea to binary	Github	PyTorch	Single	CADRE	Gene expression	Drug target	CCLE, GDSC	CCLE, GDSC	W	Collaborative	None	Versions of		
Yidong	Chen	Professor,	Specializes in	cheny8	https/	Yes. He has a large group working on many	YC Chiu	Predictin	2019	BMC Medical Genomics	IC50	No	TF	Single	DeepDR	Gene expression,	None	CCLE, TCGA			Autoencoder	None	Liner		
Zhongming	Zhao	Chair Professor	He has a large	Zhong	https/	Maybe. His group has one paper of drug response	P. Jia	Deep	2021	Nature Communications	IC50, ActArea	Github	TF and R		VAEN	Gene expression	None	CCLE, GDSC, TCGA	CCLE, GDSC, TCGA	CW	VAE	ElasticNet			
Jian	Wu						R. Feng	AGMI:	2021	ArXiv	IC50		PyTorch	Single	AGMI										
Andrea	Califano	Professor, Depart	Use systems bio	ac2248@	http://	Maybe. Andrea Califano's group is working on drug resp	M. Bansal	A commu	2014	Nature Biotechnology	IC20	Yes	DREAM	Pair	DREAM-methods	Gene expression	None	CCL (DREAM)	CCL (DREAM)		None	DREAM-method	DREAM-methods		
Andreas	Bender	Director of Digita	He is applying a	ab454@	https/	No. Andreas Bender is mainly working on target selectio	I. Cortés-Ciriano	Improved	2016	Bioinformatics		No	Sklearn	Single			Fingerprints (Morg	NCI-60	NCI-60	W	None	RF, SVM			
Changiz	Eslahchi	Professor, Shahid	He is focussing	ch-eslah	http://e	Maybe. He is working on various aspects of cancer resear	A. Emdadi	DSPLMF:	2020	Frontiers in Genetics					DSPLMF						None	Logistic Matrix Factorization			
Changiz	Eslahchi	Professor, Shahid	He is focussing	ch-eslah	http://e	Maybe. He is working on various aspects of cancer resear	FA Moughari	ADRML: a	2020	Scientific Reports	IC50	Github	NumPy	Single	ADRML						None	Manifold Learning			
Duc-Hau	Le	Associate Profess	He is working o	hauldhu	https/	Maybe. He has started working on drug sensitivity predic	DH Le	Drug Resp	2018	Journal of Molecular Biology	IC50	No	NA	Single	GloNetDRP			CCLE, GDSC	CCLE, GDSC	W	None	Network based	KBMTL		
Duc-Hau	Le	Associate Profess	He is working o	hauldhu	https/	Maybe. He has started working on drug sensitivity predic	G. Nguyen	An investi	2020	IEEE International Conference on Knowledge and Syster	No	NA	Single								None	Regression and classification models			
Ewa	Szczurek	Assistant Profess	Her main resear	szczurek	https/	No. Her new group has recently started working on drug	K. Koras	Feature se	2020	Scientific Reports	AUC	Yes	Sklearn	Single				GDSC	GDSC	W	None				
Fuhai	Li	Assistant Profess	Fuhai Li is focus	fuhai.li@	https/	Yes. Although his group is not big nor very established, h	R. KE	Integrativ	2017	AMIA Jt Summits Transl Sci Proc				Combo	SynGeNet						None	Genomic interaction based network approach			
Fuhai	Li	Assistant Profess	Fuhai Li is focus	fuhai.li@	https/	Yes. Although his group is not big nor very established, h	J. Xu	Diffusion	2018	Pac Symp Biocomput		Upon request		Combo							None	Disease signaling network based approach			
Fuhai	Li	Assistant Profess	Fuhai Li is focus	fuhai.li@	https/	Yes. Although his group is not big nor very established, h	KE Regan-Fendt	Synergy fr	2019	NPJ Syst Biol Appl		Yes	R	Combo	SynGeNet						None	genomic interaction based network approach			

IMPROVE Aim 2: Data Generation to Evaluate and Improve Drug Response Models

- Aim 2 will design and execute high-throughput experiments to generate new data aiming at evaluating and improving drug response prediction models
- Data will include RNA-seq and DNA-seq data of cancer models and drug response data with multiple doses and replicates, and potentially other clinically relevant tests.
- Cancer models can be patient-derived organoids (PDOs), xenograft organoids (PDXOs), and primary cell lines (PDCs), which are better representations of patient tumors than immortalized cancer cell lines
- In addition to data generation, we will continuously curate and standardize new drug screening/response data from the public domain

Anticipated Resources Developed

Software

- **A pipeline enabling the evaluation of new prediction models and comparison with existing state-of-the-art models; standardized evaluation metrics and scenarios will be implemented.**
  - GitHub link: https://github.com/JDACS4C-IMPROVE
  - Multiple prediction performance metrics and functional metrics, e.g., interpretability and uncertainty quantification
  - Multiple cross-validation scenarios

Models

- **Existing state-of-the-art drug response prediction models included in the pipeline that can run in batch mode that have been curated/validated and publicly available for easy adoption by the cancer research community.**
- **Improved prediction models through transfer learning on newly generated/curated data**

Anticipated Translational Goals

Benchmark Data

- **Newly generated drug screening data on PDOs, PDXOs, or PDCs.**
- **Newly curated, standardized, and aggregated drug screening/response data on cell lines, PDOs, PDXs, and patients**

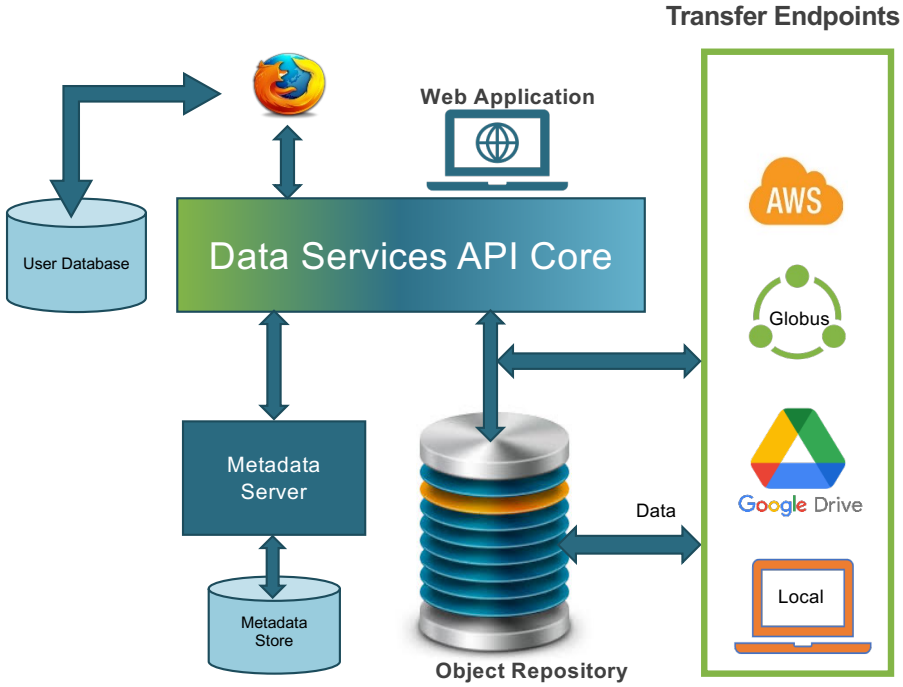
Advancing the state of the art

- **Systematic errors in the ability of AI to predict outcomes/treatments can indicate novel subtypes and highlight previously unappreciated therapeutic targets.**
- **Potential help move from stage/grade classification to classification based on treatment classes and the likelihood of a favorable outcome.**
- **Aiding researchers in knowing, which models are believable and how they can be applied in real-world situations.**
- **Providing a systematic measurement of the value of each type of test/data in relation to cost and patient impact.**

Everything Needs to be Open

- The IMPROVE *framework, our model analysis results, any improved models and all the data* produced will be open source and available to the whole community
- IMPROVE *will hold development hackathons that will be open and an annual meeting* that will be open to the community for participation
- IMPROVE *will work with the ecosystem to advocate for open models, open data, and open-source* enabling replication of modeling results

Predictive Oncology Model & Data Clearinghouse (MoDaC)



- Clearinghouse for annotated mathematical models and datasets from NCI collaborations
- Public facing web interface and RESTful APIs for submitting data.
- Metadata based search capability for locating models and datasets. Browsing and filtering support.
- Models and datasets can be staged in restricted access mode until ready for sharing.
- Multiple endpoint types supported for data transfer.
- DOI Support
  - Global identifier per asset.
  - Shareable link for citations.

https://modac.cancer.gov

Acknowledgments

ANL

Rick Stevens

Tom Brettin

Yitan Zhu

And the whole DOE team

NCI

Emily Greenspan

Jeff Hildesheim

FNL

Eric Stahlberg

Amar Khalsa & the resource transfer team

Lynn Borkon & Petrina Hollingsworth

All of CDSI

For questions or to collaborate please email me
Ryan.weil@nih.gov

Backup slides