BIOAI

Reimagining AI in Biopharma

Tom Chittenden, PhD, DPhil, PStat
Chief Scientific Officer, President of R&D, SAB Chairman
Lecturer on Pediatrics, BCH & Harvard Medical School

The Lab of the Future Congress USA
10 March 2023

AI/SciML Platform for Drug Discovery and Drug Development

Generative AI/MLmodels for **cell differentiation** and unsupervised clinical & phenotype data integration and **phenomapping**

Functional enrichment & advanced **Natural language processing (NLP)** of published research literature

Domain adaptation and **transfer learning** methods for cross-species and/or cross-cohort signal acquisition

Novel NLP embeddings for advanced feature learning/engineering

Novel causal inference and **qML models**

Advanced deep convolutional **neural network** topologies

Discovery

- Analysis of **Single Cell data**
- Analysis of **multi-omic data**
- Class assignment for omics & phenotypic patient cohort data
- **Drug target discovery**

Pre-Clinical

- Biological interpretation of AI/MLfindings

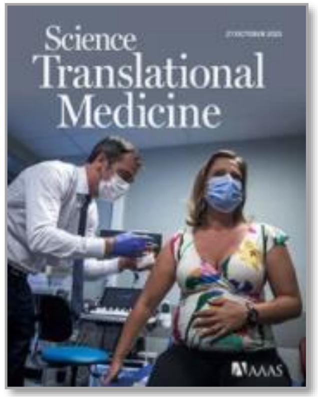
Clinical

- **Drug repositioning**
- **Analysis of RWE data**
- **MOA**
- **Drug efficacy assessment in clinical trials**

Commercial

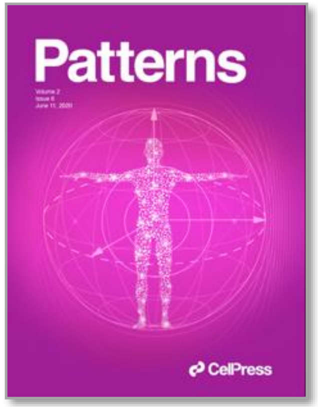
- Digital pathology, medical image analysis for **radiogenomics** & patient inclusion/ exclusion criteria for clinical trials

Scientific Impact in the Biomedical Sciences



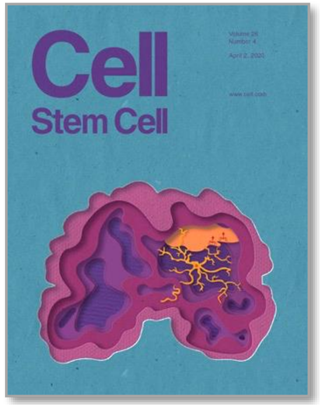
Identification of driver genes for critical forms of COVID-19 in a deeply phenotyped young patient cohort link

Featured on Cover
Accompanying Editorial



Quantum processor-inspired machine learning in the biomedical sciences link

Featured at NHLBI

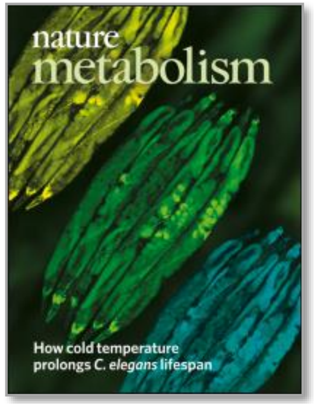


Smooth muscle cell reprogramming in aortic aneurysms link

Accompanying Editorial

Chronic mTOR activation induces a degradative smooth muscle cell phenotype link

Accompanying Editorial



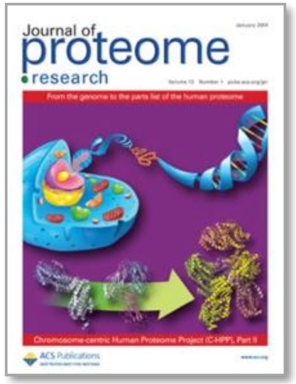
Endothelial TGF-β signaling drives vascular inflammation and atherosclerosis link

Accompanying Editorial

Endothelial ERK1/2 signaling maintains integrity of the quiescent endothelium link

Featured on Cover

Enhancing retrosynthetic reaction prediction with Deep Learning Using multiscale reaction classification link



Proteomic analysis and identification of cellular interactors of the giant ubiquitin ligase HERC2 link

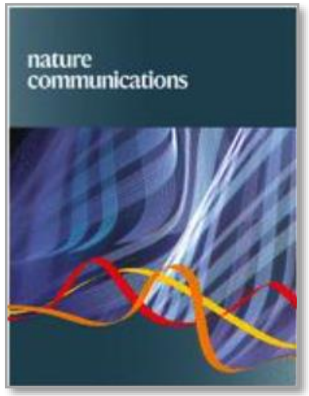
Featured on Cover
Accompanying Editorial

Somatic mutation in single human neurons tracks developmental and transcriptional history link

FGF-dependent metabolic control of vascular development link

Accompanying Editorial

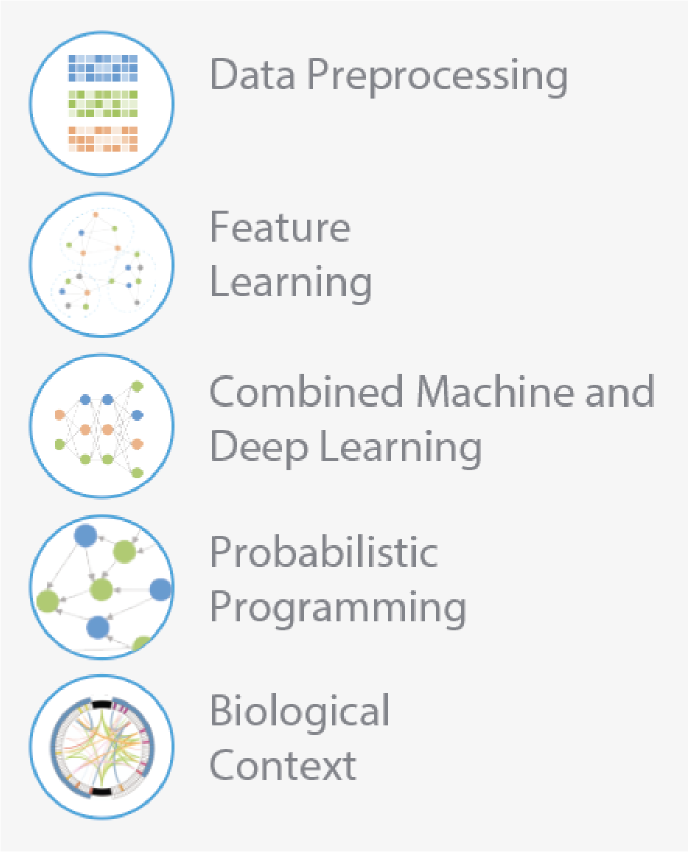
Aging & neurodegeneration are associated with increased mutations in single human neurons link



Shear-induced Notch-Cx37-p27 axis arrests endothelial cell cycle to enable arterial specification link

Ensemble AI/SciML and **Precision Medicine**

The computational power of modern AI/ML technology is well-positioned to uncover new and actionable insights from the exponentially growing pool of biological data.



Data Preprocessing

Feature Learning

Combined Machine and Deep Learning

Probabilistic Programming

Biological Context

FEATURE LEARNING

The intelligent simplification of high-dimensional multi-omic data without loss of information

ENSEMBLE MACHINE & DEEP LEARNING

Intelligent algorithms capable of self-optimization to achieve incredible accuracy with complex, layered data

CAUSAL INFERENCE

Specialized statistical learning models capable of elucidating casual dependencies within biological data

NATURAL LANGUAGE PROCESSING

Intelligent scanning of sentence syntax to understand and validate findings in context, at scale

The combination of several AI/SciML methods create a proprietary ensemble AI/SciML strategy capable of revealing novel patterns and causal dependencies in disparate and varied biological data.

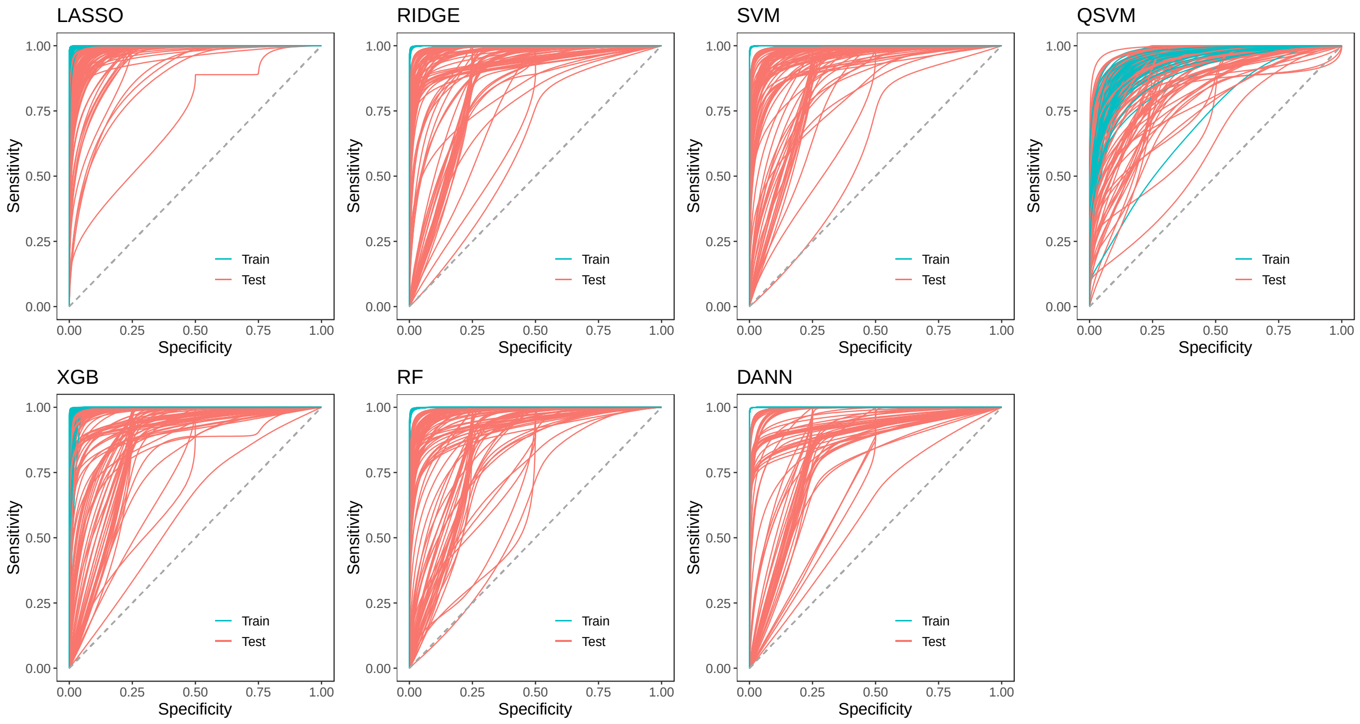
COVID-19 Severity Study: Ensemble AI/SciML/qML approach to compare critical vs non-critical patients

- Unique study design focused on adults younger than **50 without co-morbidities or age-related decline**
- During height of first wave (before corticosteroid use), our collaborators at Strasbourg recruited COVID-19 patients with ARDS requiring mechanical ventilation and matched patients who were sick enough to be hospitalized but required oxygen without assisted ventilation

Group	Description	Whole blood RNA-Seq	Plasma Proteomics	PBMC Proteomics	WGS
Critical	Severe COVID patients with ARDS requiring ventilators	**46**	45	34	47
Non-critical	Control COVID patients hospitalized and given oxygen but not ICU care	**23**	23	21	25
COVID negative controls	Healthy controls from blood donors at same hospital	22	22	22	22

Carapito, R *et al*., *Science Translational Medicine* 2021

Cohort 1 RNA-seq: Critical(ARDS) vs non-critical groups with ensemble strategy comprising 7 different AI/ML algorithms including Quantum SVM (69 patients total) – Leads to stable gene feature sets to allow for subsequent causal inference (other methods produce variable gene lists)



	LASSO	Ridge	SVM	qSVM
Accuracy (Train/Test)	0.9991 ± 0.0004 / 0.9677 ± 0.0050	1.0000 ± 0.0000 / 0.9169 ± 0.0072	1.0000 ± 0.0000 / 0.9223 ± 0.0075	0.9245 ± 0.0028 / 0.8677 ± 0.0121
Balanced Acc. (Train/Test)	0.9987 ± 0.0006 / 0.9503 ± 0.0078	1.0000 ± 0.0000 / 0.8990 ± 0.0094	1.0000 ± 0.0000 / 0.9068 ± 0.0092	0.9189 ± 0.0039 / 0.8607 ± 0.0118
AUROC (Train/Test)	**1.0000 ± 0.0000 / 0.9908 ± 0.0036**	**1.0000 ± 0.0000 / 0.9547 ± 0.0075**	**1.0000 ± 0.0000 / 0.9633 ± 0.0070**	**0.9667 ± 0.0029 / 0.9386 ± 0.0081**
F1 (Train/Test)	0.9993 ± 0.0003 / 0.9780 ± 0.0034	1.0000 ± 0.0000 / 0.9404 ± 0.0052	1.0000 ± 0.0000 / 0.9487 ± 0.0049	0.9426 ± 0.0020 / 0.9095 ± 0.0071
MCC (Train/Test)	0.9980 ± 0.0009 / 0.9251 ± 0.0118	1.0000 ± 0.0000 / 0.8128 ± 0.0169	1.0000 ± 0.0000 / 0.8364 ± 0.0161	0.8339 ± 0.0065 / 0.7398 ± 0.0198

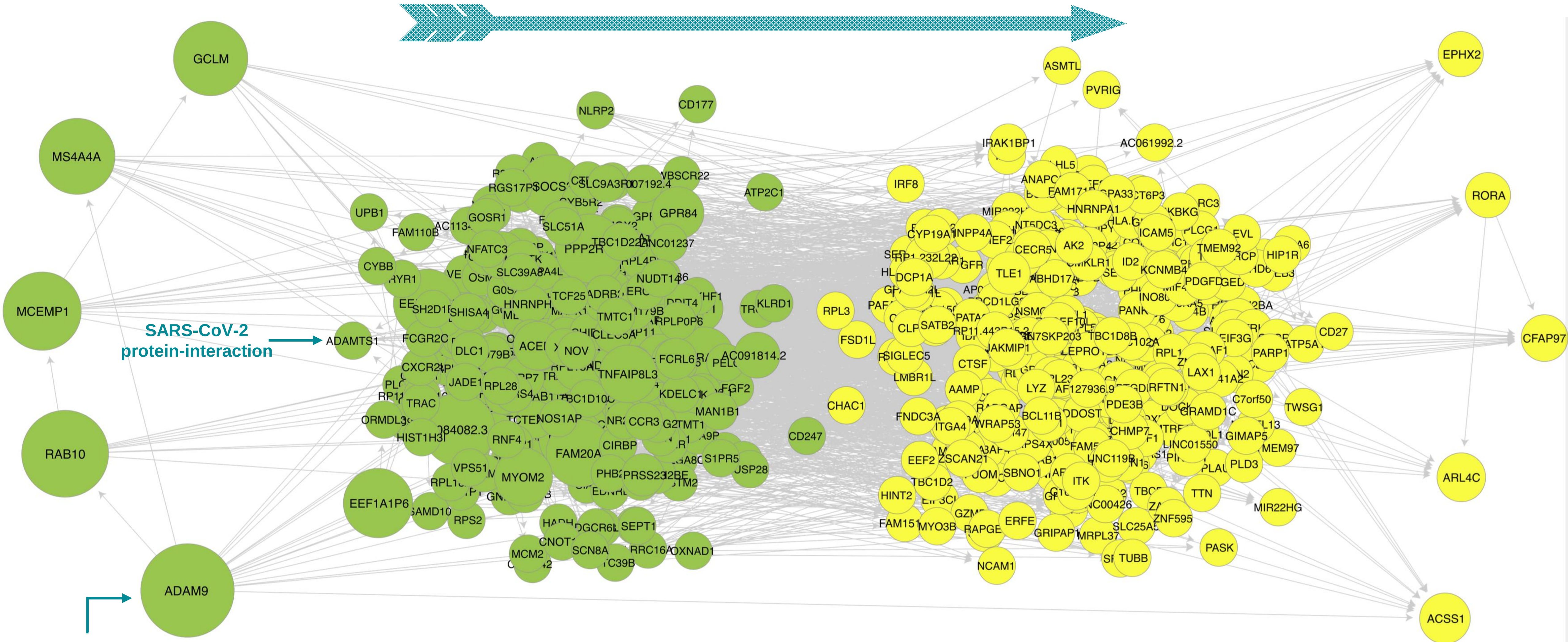
	XGB	RF	DANN
Accuracy (Train/Test)	0.9952 ± 0.0008 / 0.9146 ± 0.0076	1.0000 ± 0.0000 / 0.9254 ± 0.0072	1.0000 ± 0.0000 / 0.9131 ± 0.0083
Balanced Acc. (Train/Test)	0.9930 ± 0.0012 / 0.8932 ± 0.0100	1.0000 ± 0.0000 / 0.9072 ± 0.0094	1.0000 ± 0.0000 / 0.9032 ± 0.0097
AUROC (Train/Test)	**0.9999 ± 0.0000 / 0.9443 ± 0.0079**	**1.0000 ± 0.0000 / 0.9360 ± 0.0091**	**1.0000 ± 0.0000 / 0.9435 ± 0.0081**
F1 (Train/Test)	0.9964 ± 0.0006 / 0.9391 ± 0.0054	1.0000 ± 0.0000 / 0.9467 ± 0.0052	1.0000 ± 0.0000 / 0.9359 ± 0.0062
MCC (Train/Test)	0.9893 ± 0.0018 / 0.8061 ± 0.0181	1.0000 ± 0.0000 / 0.8308 ± 0.0168	1.0000 ± 0.0000 / 0.8091 ± 0.0185

High accuracies with AUROC of test sets from 0.94 to 0.99

Carapito, R *et al.*, *Science Translational Medicine* 2021

Integrated AI/SciML and Structural Causal Modeling on top of 600 RNA-seq genes from critical (ARDS) vs non-critical groups

Putative flow of Causal Dependency



SARS-CoV-2 protein-interaction

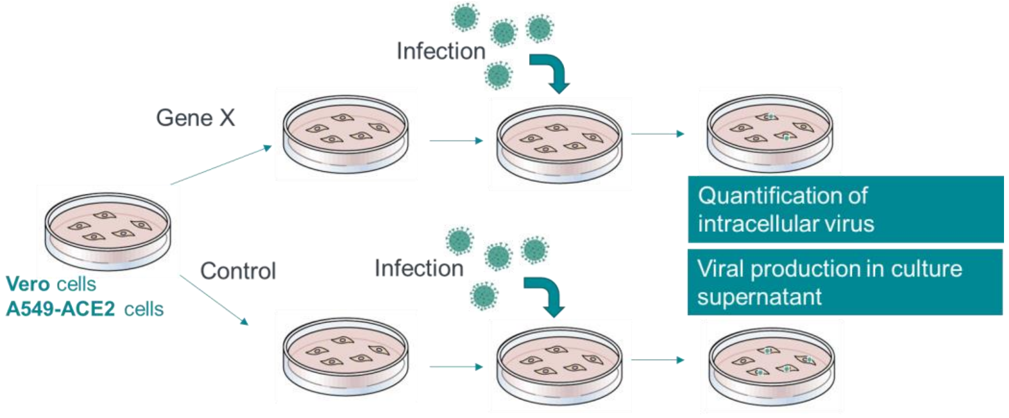
SARS-CoV-2 protein-interaction

Carapito *et al.*, *Science Translational Medicine* 2021

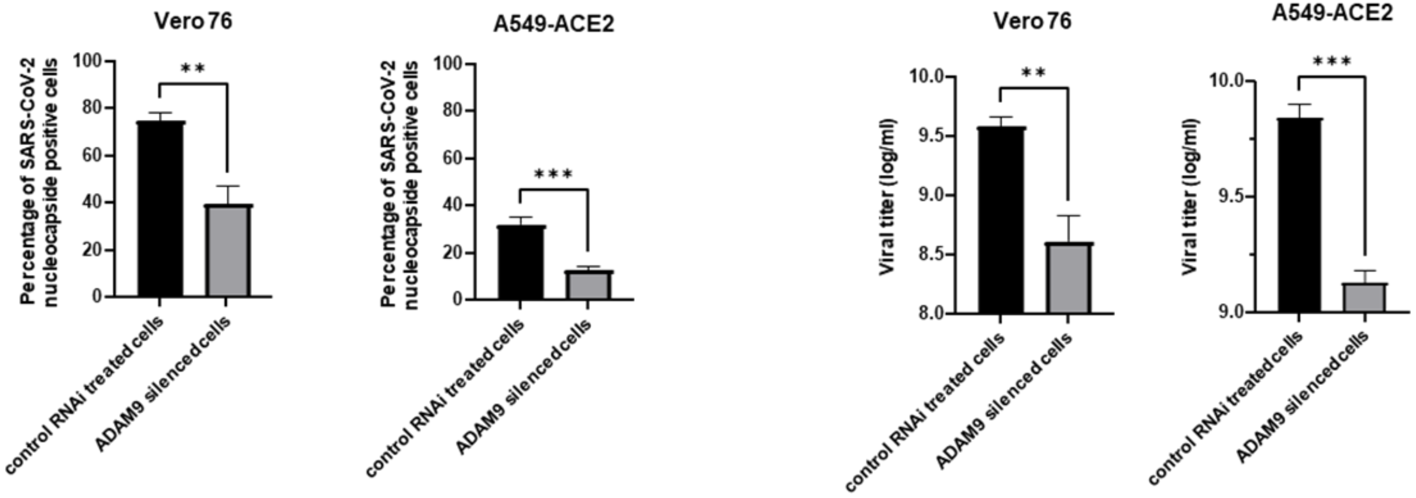
COVID-19 project – Experimental Validation of AI/SciML Findings (effects of ADAM9 silencing on viral uptake and replication)

- Silencing Gene X inhibits SARS-CoV-2 uptake and replication in human lung epithelial cells: ***strong empirical evidence of molecular mechanism of action*** **driving complex disease etiology**

- Actively pursuing a possible **first AI/ML-based repositioning** of a drug currently in Phase III trials ADAM9+ tumors

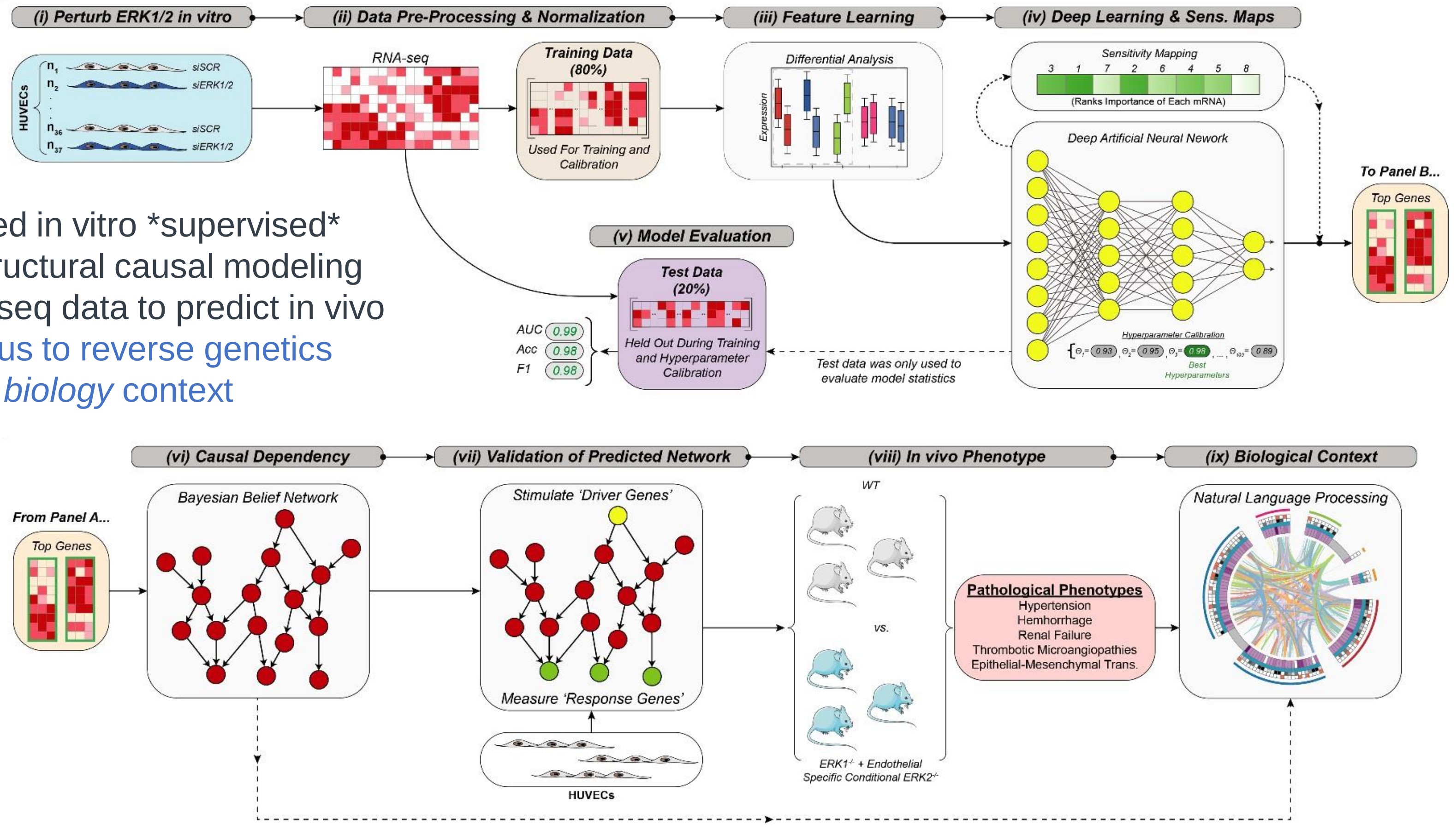


Carapito, R *et al*., *Science Translational Medicine* 2021



In Silico Phenotype Projection: Identifying Causal Drivers of CVD (Hypertension, Vascular Hemorrhage, and Renal Failure)

(Ricard et al., *JEM* 2019)



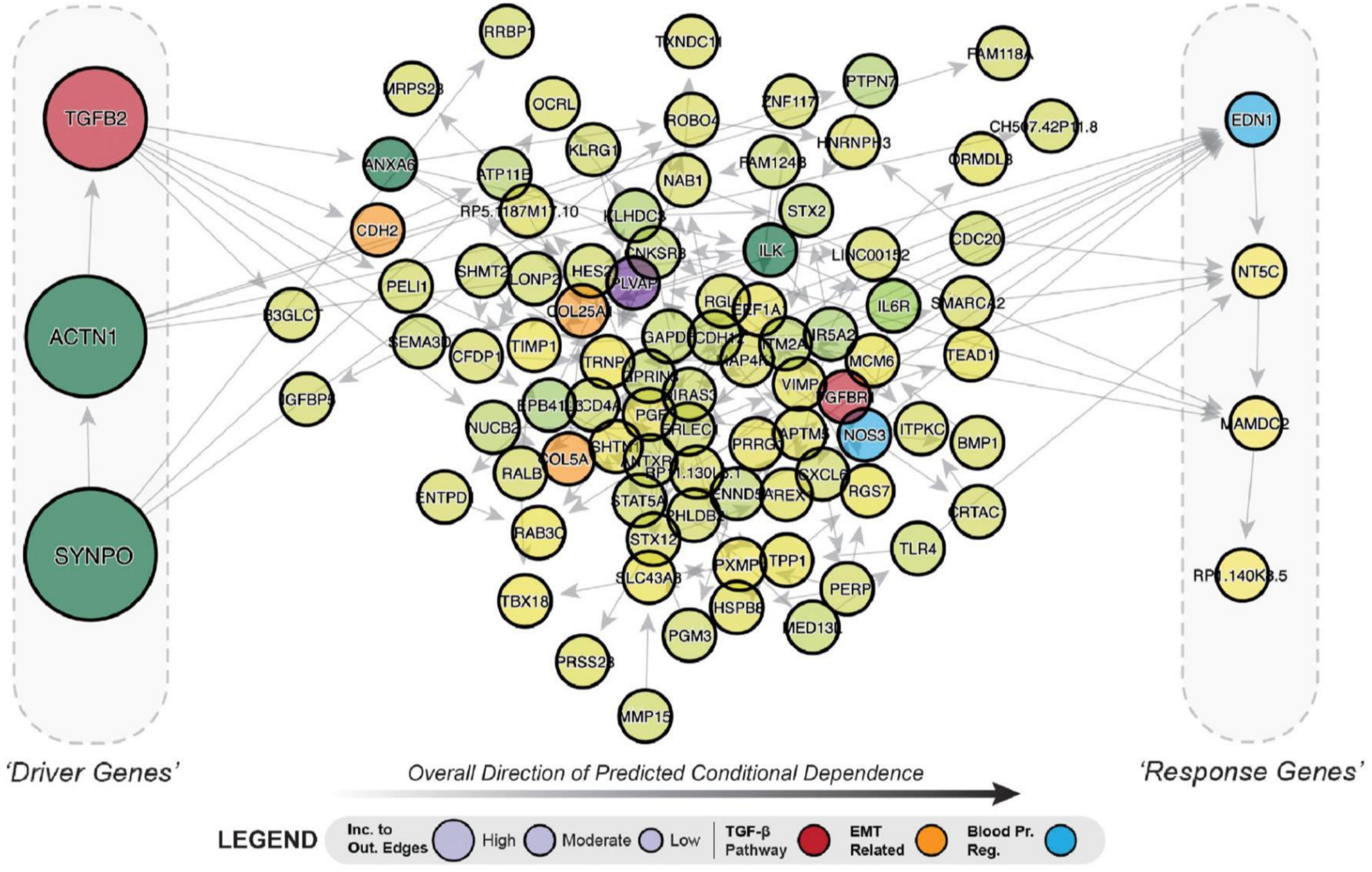
Example of integrated in vitro *supervised* classification and structural causal modeling (SCM) of bulk RNA-seq data to predict in vivo phenotype. Analogous to reverse genetics with added *systems biology* context

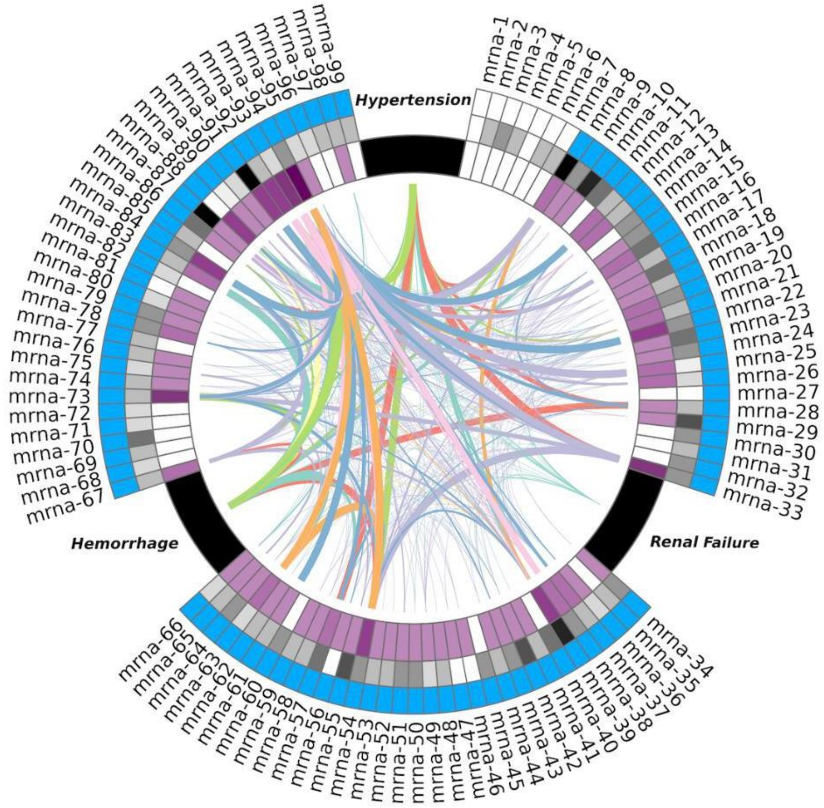
Research Collaboration with Yale Cardiovascular Research Center
Deep Learning, Structural Causal Modeling, and NLP of Bulk and Single Cell RNA-seq Data

In Silico Phenotype Projection: Identifying Causal Drivers of CVD (Hypertension, Vascular Hemorrhage, and Renal Failure)

Example of causal structural modeling (SCM) of bulk RNA-seq analysis, which is downstream *supervised* bulk RNA-seq analysis on pervious slide to derive putative causal drivers of disease etiology, phenotype, etc.

(Ricard *et al*., *JEM* 2019)



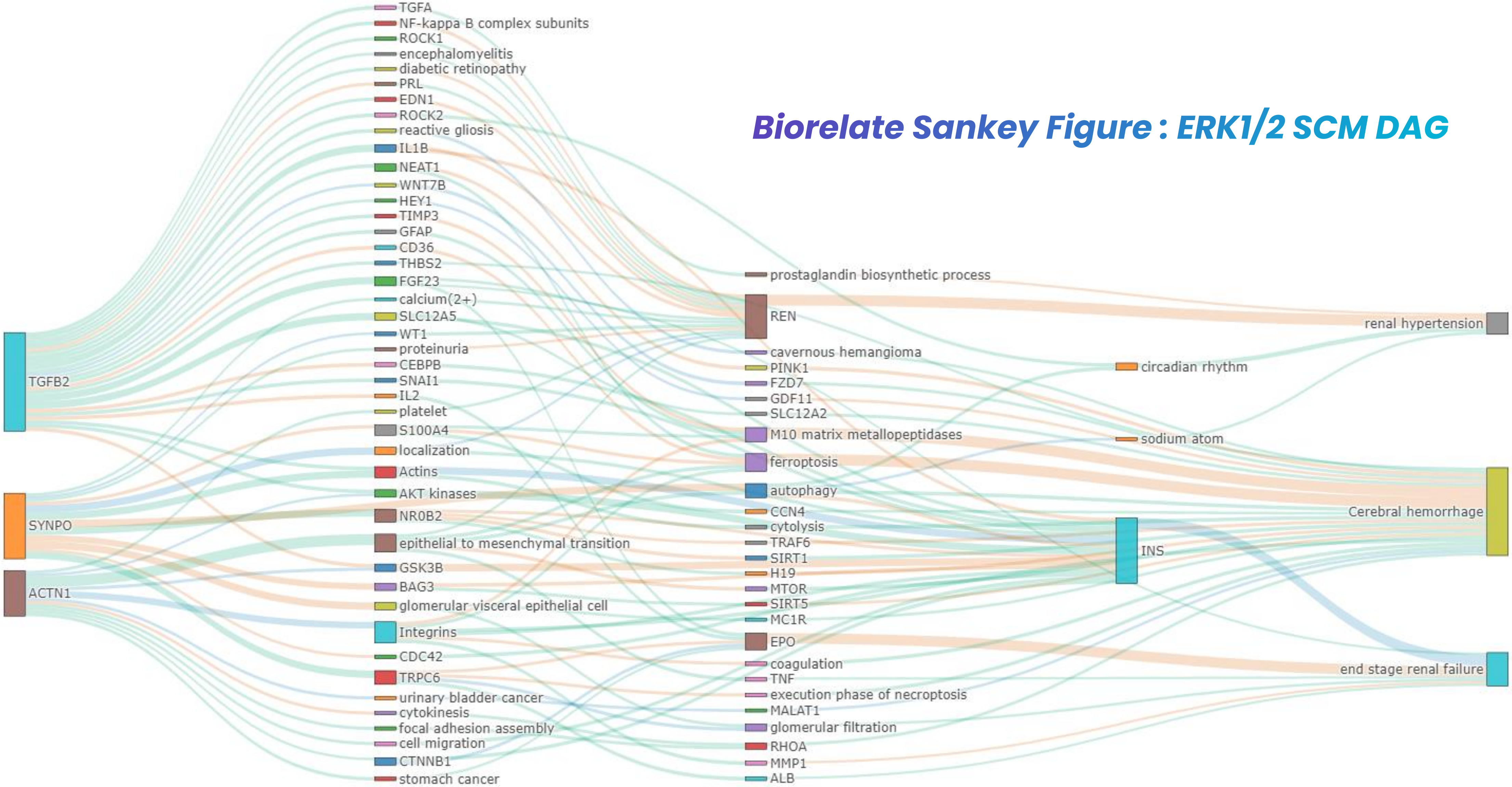


Research Collaboration with Yale Cardiovascular Research Center

Deep Learning, Structural Causal Modeling, and NLP of Bulk and Single Cell RNA-seq Data

In Silico Phenotype Projection: Identifying Causal Drivers of CVD (Hypertension, Vascular Hemorrhage, and Renal Failure)

Biorelate Sankey Figure : ERK1/2 SCM DAG



Research Collaboration with Yale Cardiovascular Research Center

Deep Learning, Structural Causal Modeling, and NLP of Bulk and Single Cell RNA-seq Data

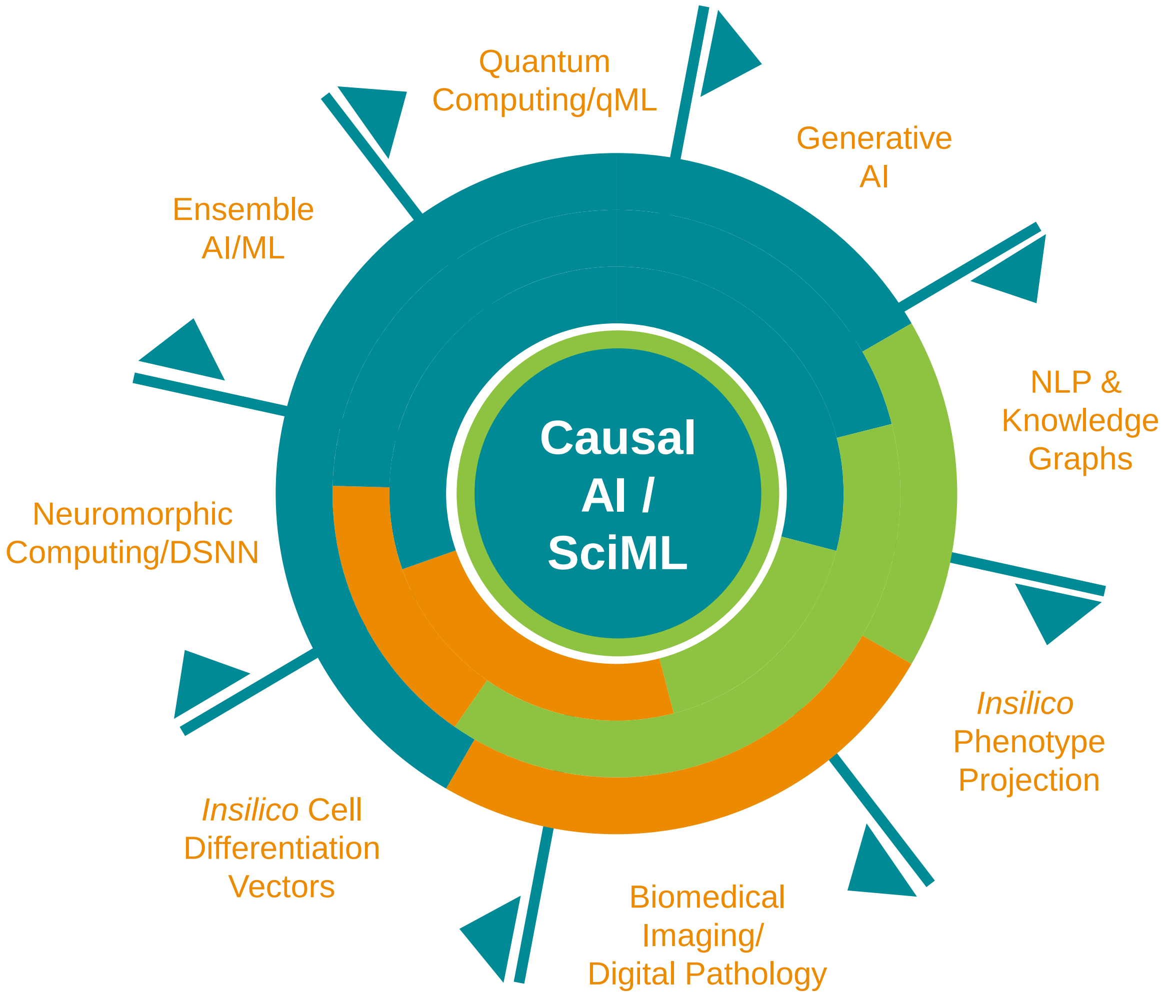
A Novel AI/SciML R&D Innovation Engine

Inputs

Disease-related questions

- NASH
- Alzheimer's
- Oncology
- Lupus
- COVID-19
- Cardiovascular
- Biology of Aging
- Rare Diseases

• • •



Causal AI / SciML

Quantum Computing/qML

Generative AI

NLP & Knowledge Graphs

Insilico Phenotype Projection

Biomedical Imaging/ Digital Pathology

Insilico Cell Differentiation Vectors

Neuromorphic Computing/DSNN

Ensemble AI/ML

Outputs

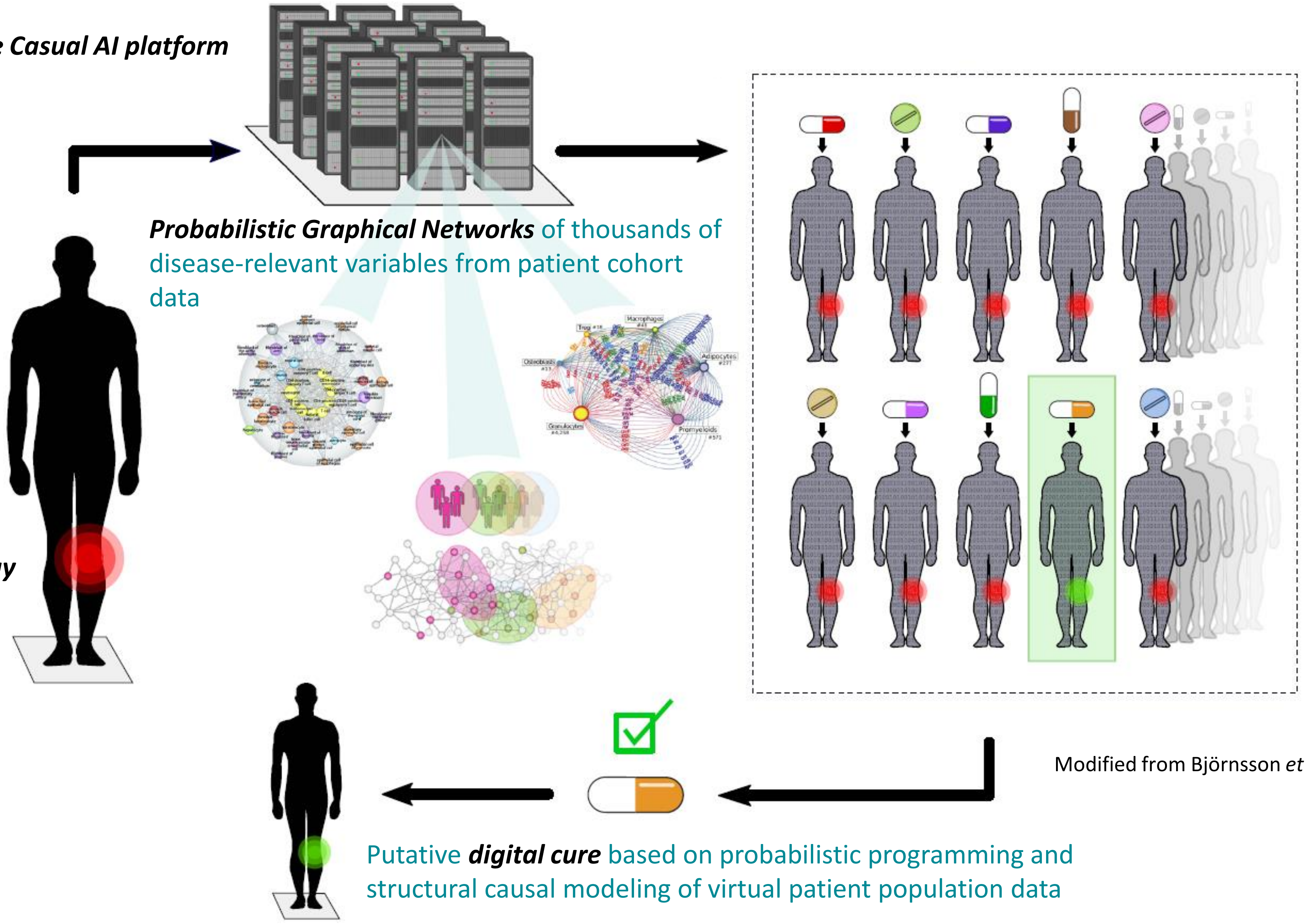
- Iterative Innovation
- Novel Insights
- True Discovery

Innovation Accelerator

An opportunity to fail early / fast

Causal AI and Digital Twin Technology Roadmap:

The future of drug target discovery and drug development



World's first Exascale Casual AI platform
(Oracle)

Probabilistic Graphical Networks of thousands of disease-relevant variables from patient cohort data

Pathology

Engineer Digital twins from patient data to create ***virtual patient populations*** for *in silico* drug efficacy assessment for thousands of therapeutics

Modified from Björnsson *et al*. *Genome Medicine* (2020)

Putative ***digital cure*** based on probabilistic programming and structural causal modeling of virtual patient population data

Acknowledgements

BioAI Health

- Alan Jerusalmi
- Michael James
- Krishna Bairavi
- Ayushi Shah
- Mike Bonham
- John Cogswell

Queen Mary University of London

- Sir Mark Caulfield
- Andrew Livingston
- Greg Slabaugh

Elucidata

- Abhishek Jha
- Devika Benoy

Biorelate

- Mark Hughes

Oracle

- Alison Derbenwick Miller
- Ryan Payton

Genuity Science

- Jeff Gulcher
- Hannes Smarason
- Judith Conroy
- Amy Swearingen

Genuity AI Research Institute

- Sweta Bajaj
- Jose Malagon Lopez
- Richard Li
- Allison McLean
- Raony Guimaraes
- Tiago Magalhaes
- Chris Hoefler

Boston Children's Hospital
Harvard Medical School

- Chris Walsh

Cardiovascular Research Center
Yale University Medical School

- Mike Simons

University of Southern California

- Daniel Lidar
- Razieh Mohseninia

University of Tennessee Health Science Center

- Hao Chen

University of Strasbourg Medical Center

- Seiamak Bahram
- Raphael Carapito