insitro

Machine Learning for Unraveling the Complex Biology of CNS Disease

Daphne Koller, insitro Founder & CEO

At the convergence of human biology and machine learning lies a healthier you.

Meet insitro.

Genetics Suggest Causality

17 Million variants across recent cohort of 123k individuals

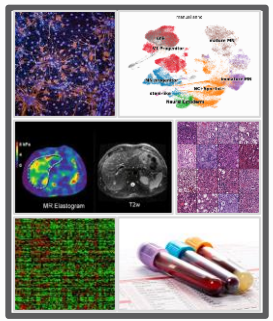
Associating phenotype with genotype suggests causality

Many variants with **small effect-sizes**

Rich Phenotypes to Bridge Genetics and Clinical Outcomes

17 Million variants across recent cohort of 123k individuals is not sufficiently powered

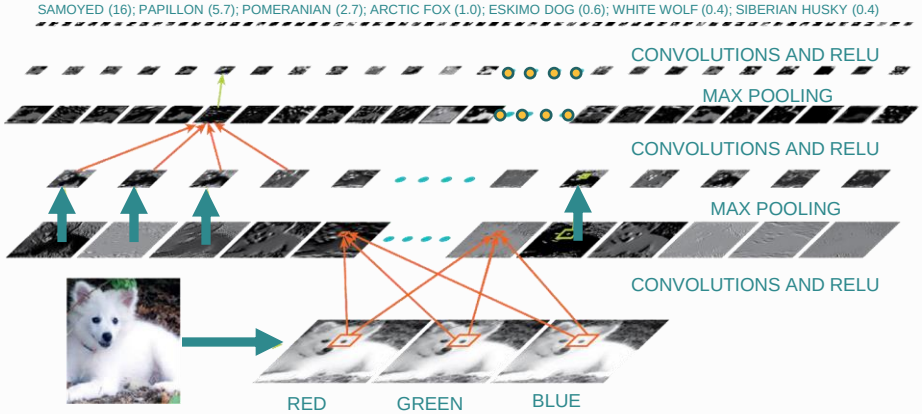
High-content biological data offers the potential to get closer to the causal biology



Machine Learning

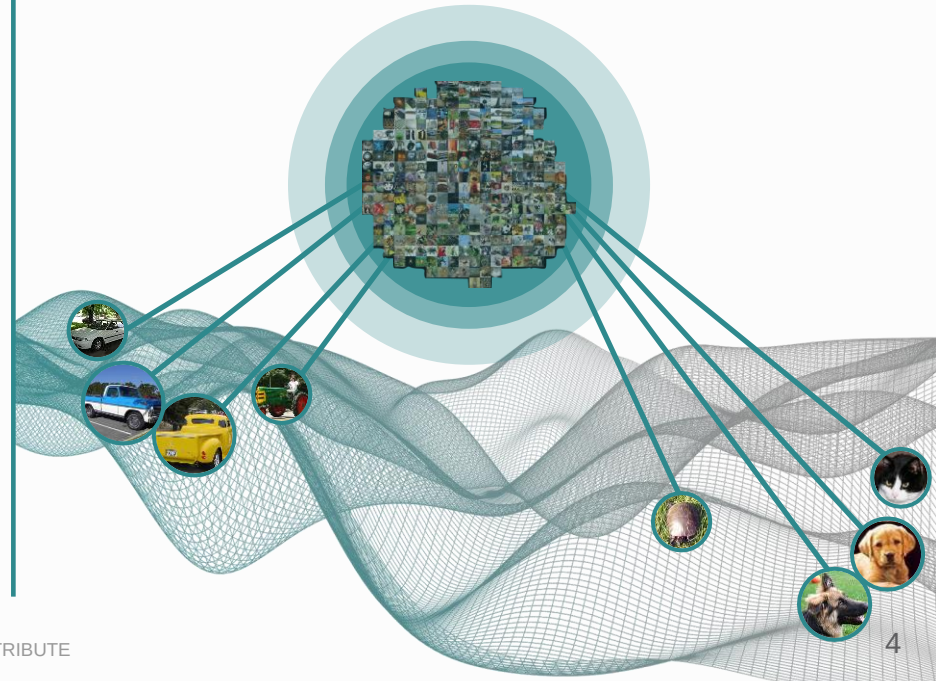
ML models can construct novel, important features

- Traditional ML models required manually defined features
- People are not very good at defining predictive features
- ML models trained end-to-end can discover features beyond human abilities

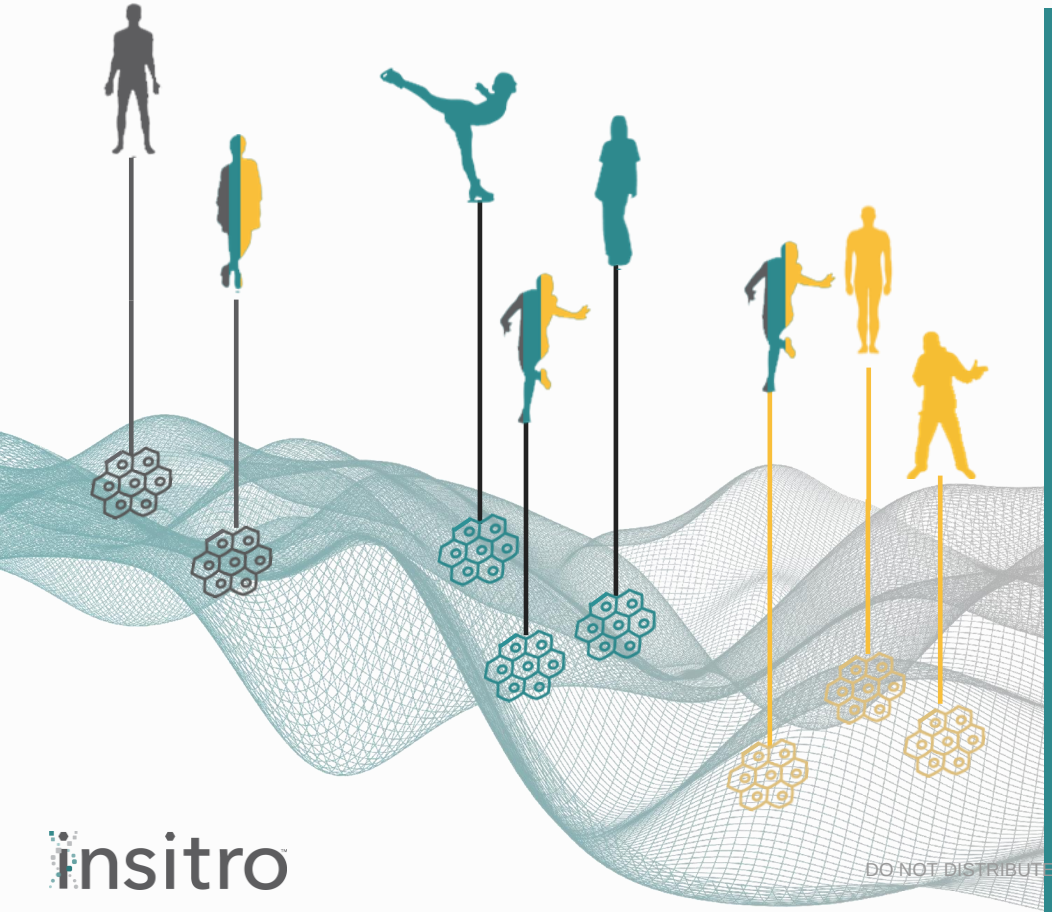


… which induce a new representation of the data

- ML features define a low-dimensional manifold in which the data are embedded
- Manifold induces a task-relevant distance function where semantically related objects are adjacent



Interrogating Disease for Better Medicines

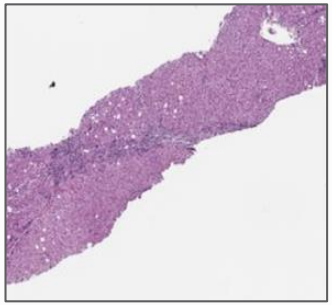


Our Approach

- *High-content data from human cohorts and genetically diverse iPSCs empowers deeper understanding of disease biology*
- *Strong genetic drivers of disease provide "positive samples" of disease state for use by machine learning*
- *Dense biomarker data reveals underlying heterogeneity in patient population*
- *Discovery of patient segments and causal drivers of disease enables identification and de-risking of therapeutic targets and interventions*

insitro

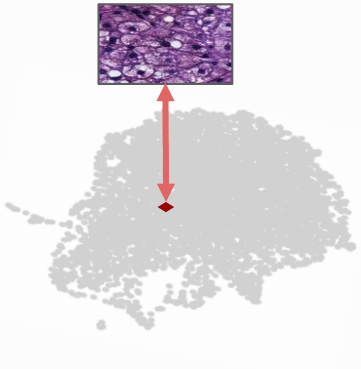
ML-enabled statistical genetics



Data
(e.g., histology)

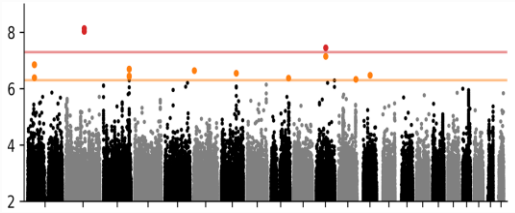
ML to extract features

Embedding



Embedding
(a type of feature vector)

Downstream analyses
(e.g., GWAS)



Novel insights into disease
(e.g., genetic drivers)

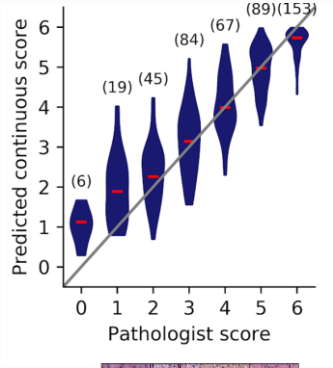
We leverage ML and high-content data to identify, de-risk, and prosecute new genetic insights, therapeutic targets, patient populations, and uses for existing molecules

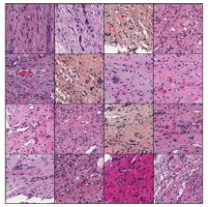
Pathologist-level Prediction Performance

The ML model was trained on 5 Gilead NASH clinical trials using 4,178 slides* and evaluated on a held-out set of 463 slides from different patients and clinical centers

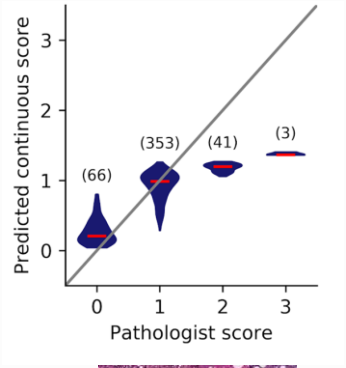
ρ = Spearman correlation

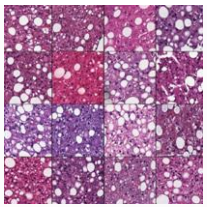
Fibrosis (Ishak) score (ρ = 0.90)



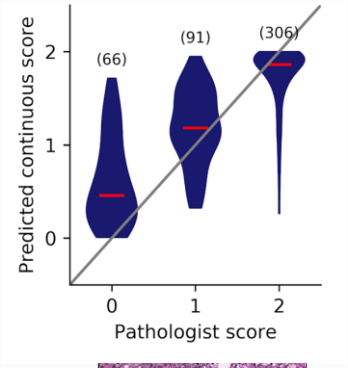


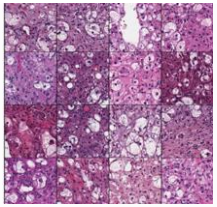
Steatosis (ρ = 0.70)



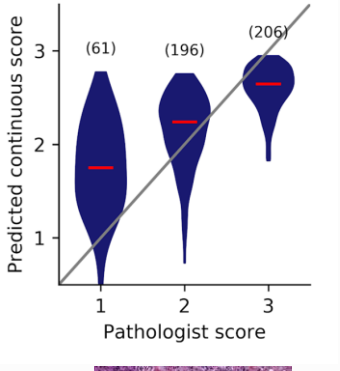


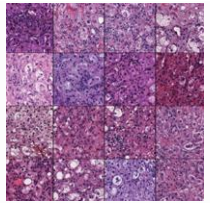
Hepatocyte ballooning (ρ = 0.71)





Lobular inflammation (ρ = 0.67)

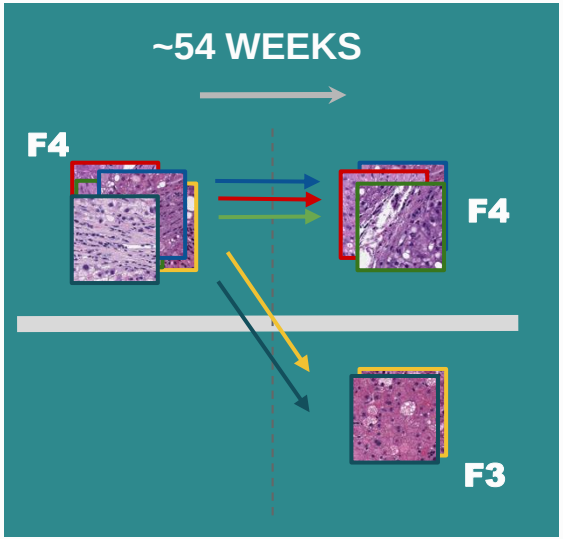




P. Casale et. al., Oral presentation at EASL, April 2020; *Resulting aggregate dataset contains ~2 million tiles

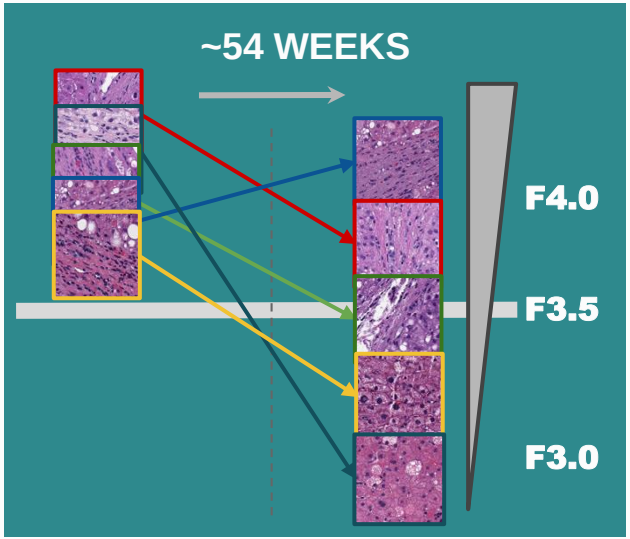
Example: Novel Genetic Drivers of Fibrosis Progression

What *pathologist* sees...



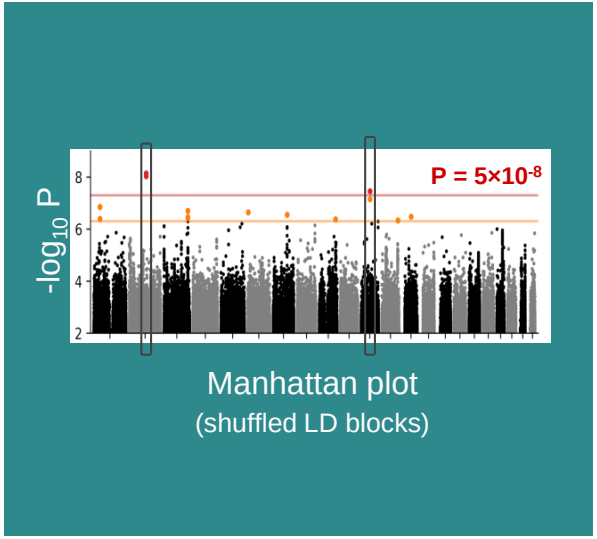
- Ordinal scores only
- Binary progression / regression

What *insitro* sees...



- Greater resolution
- Quantitative estimate of progression/regression

What we uncovered…



- Two novel, genome-wide significant variants
- Compelling biology
- Eminently prosecutable

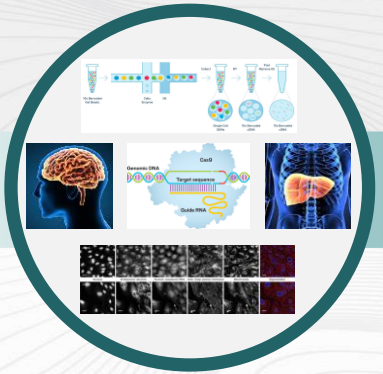
ML-enabled Disease Models

Analyze human data

Derive insights from genetic, phenotypic, and clinical data

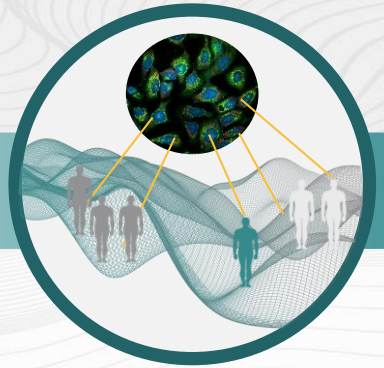
Build *in vitro* systems

Use iPSCs and biology at scale to generate multi-modal data



Build ML models

Use ML to build phenotypic manifolds from massive datasets



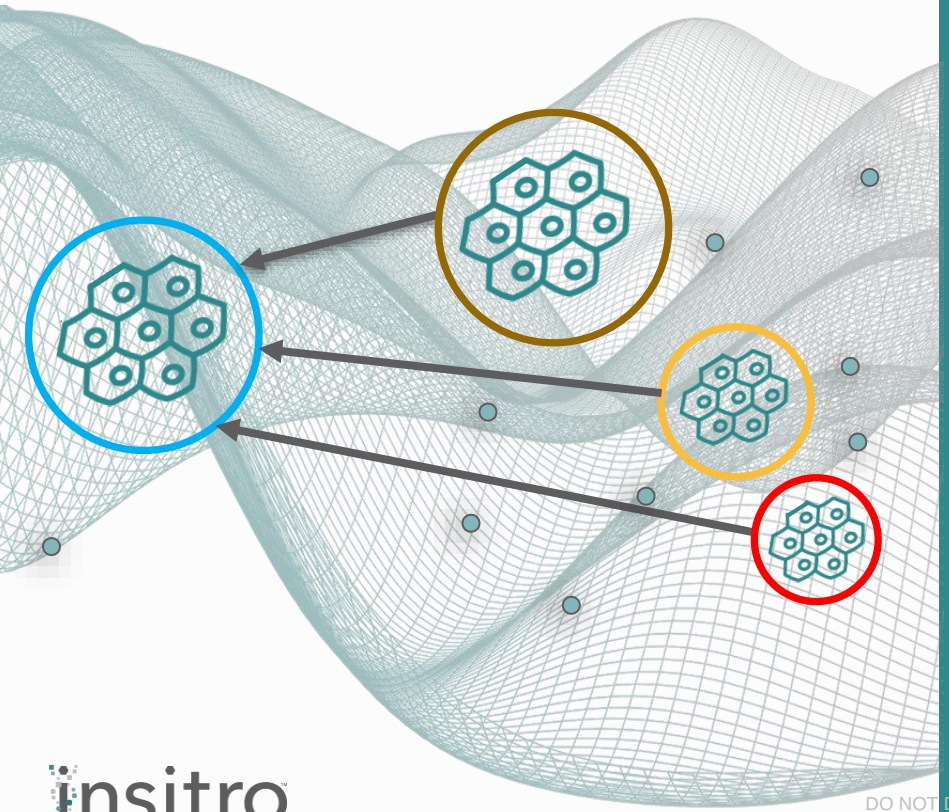
From this we get...

- Insights on disease
- Novel genetic drivers
- Screening platform

Predictive Disease Models

Leading edge technology stack integrating iPSC technology, automation, and compute to generate at-scale data that spans genetic diversity

Screening Using the Phenotypic Manifold



Identifying Relevant Interventions

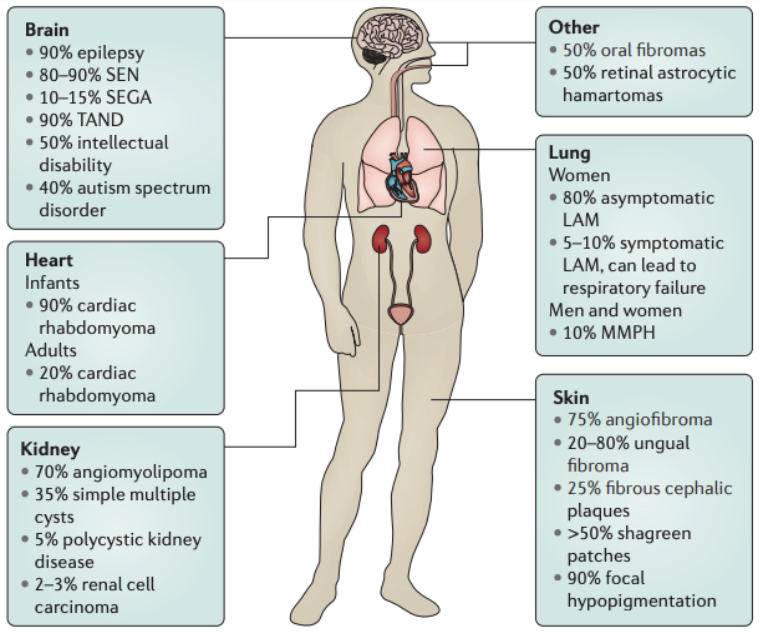
- ***Genetic interventions*:** *starting point for target-based drug discovery*
- ***Chemical interventions:** existing molecules, evaluating new hits*
- ***Integration with therapeutics design** allowing ML-enabled "phenotypic SAR"*

* Has the potential to uncover genetic modifiers – high-value novel targets, hard to find in other ways

insitro

Introduction – Tuberous Sclerosis (TSC)

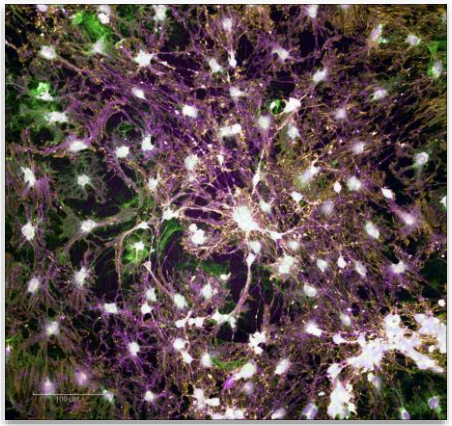
- TSC caused by LoF mutations in *TSC1* or *TSC2,* leading to overactivation of the mTOR pathway
- Benign tumors present in skin, brain, and kidneys
- Neurologic impairment and refractory epilepsy are most common cause of morbidity
- Rapamycin approved for adjunctive treatment for partial onset seizures
  - ~50% with intractable epilepsy show no benefit
  - Issues with BBB penetration and tolerability

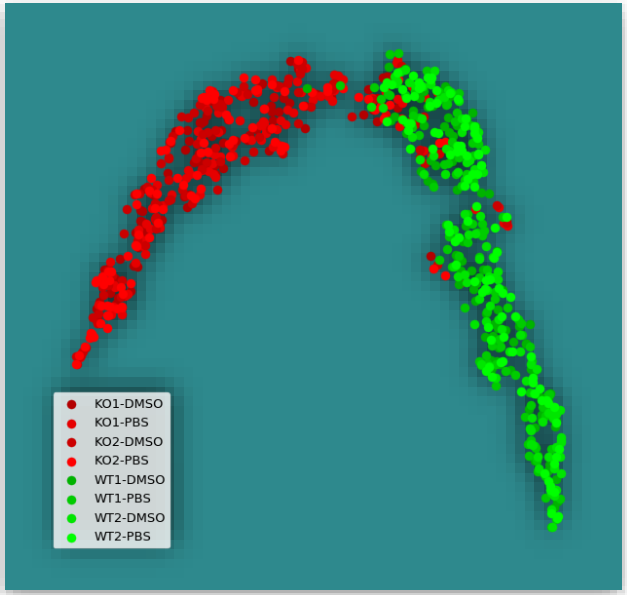


Theilmann *et al.* (2020) *Neuropharmacology* 180:108297; Henske *et al.* (2016) *Nat Rev Dis Primers.* 2:16035.

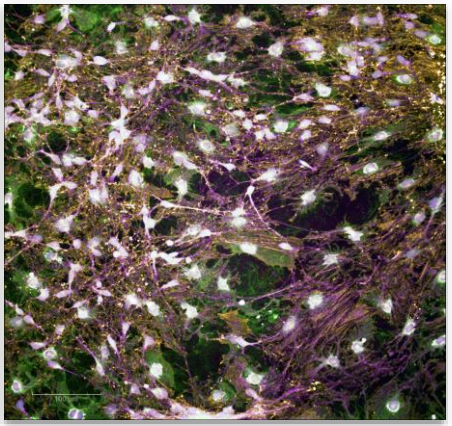
Example: Disease Phenotypes in iNeurons

TSC1$^{-/-}$





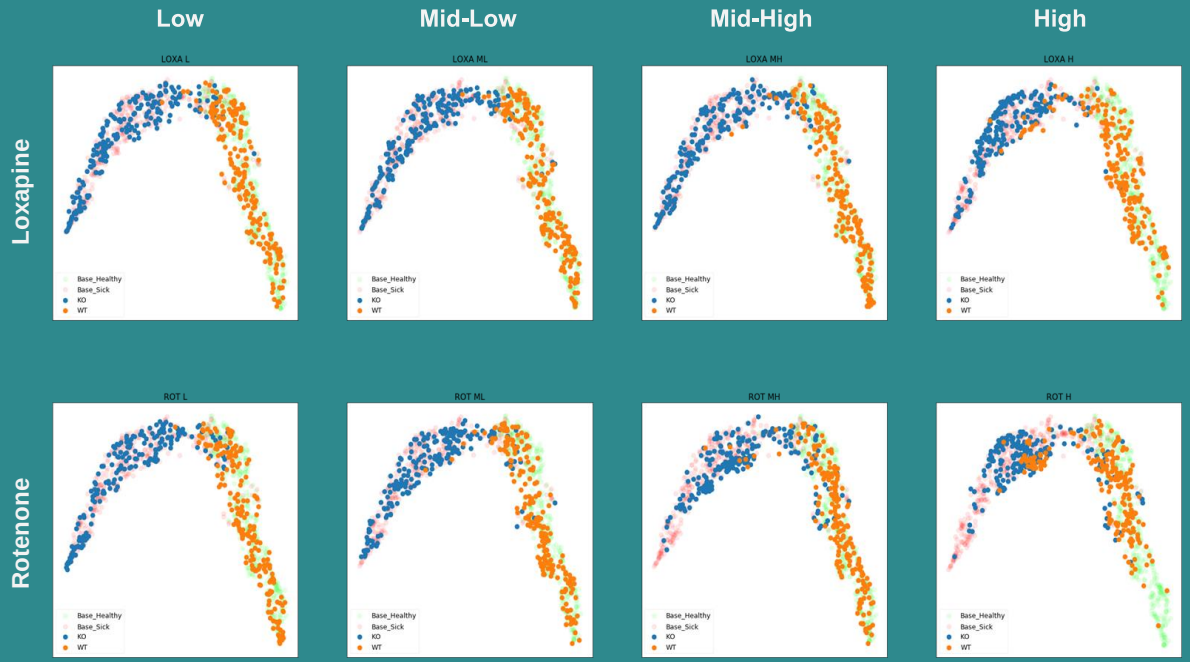
WT

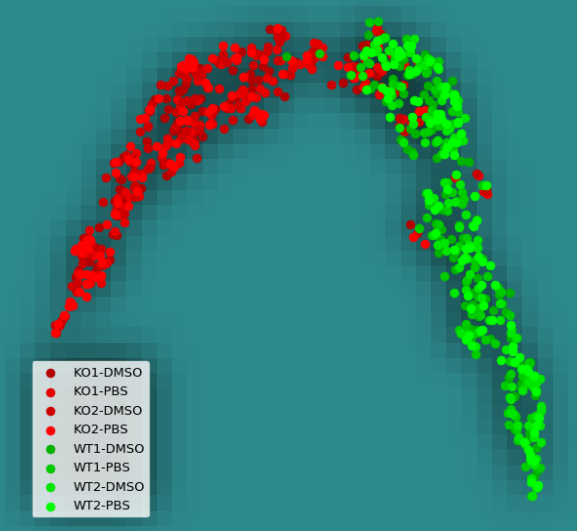


Hoechst	Concanavalin A	Wheat Germ Ag./Phalloidin	Mitotracker Red

iPSC-derived disease models with genetic engineering allow data generation at a scale that enabled machine learning for sick vs. healthy phenotypes

Negative Controls Do Not Revert Disease Phenotype



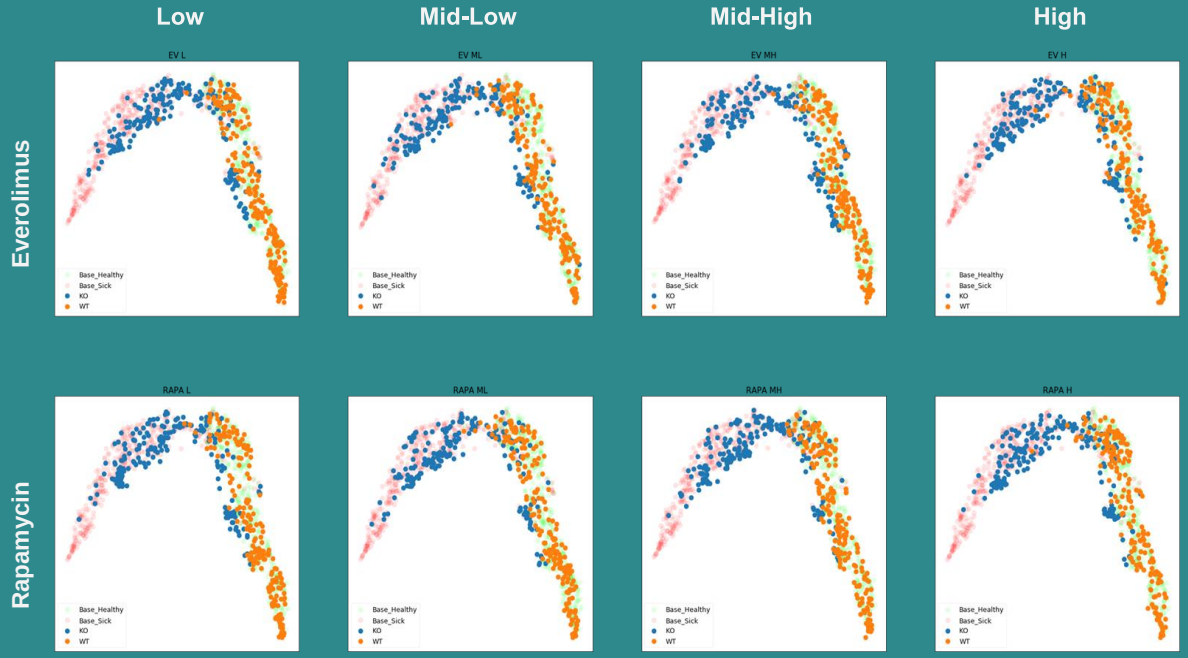


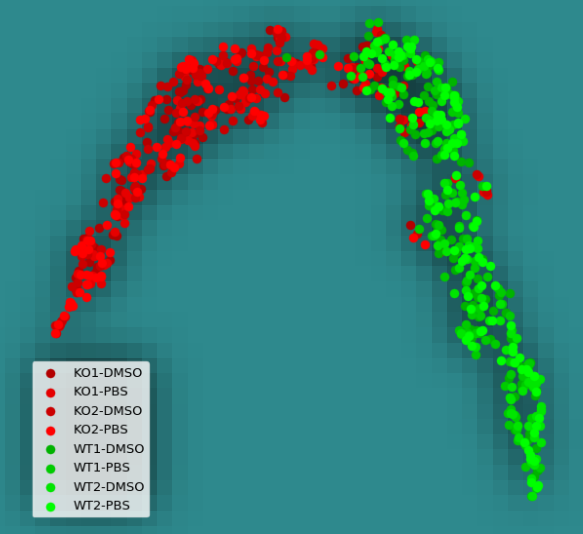
EV, RAPA - known effective compounds
ROT, LOXA - orthogonal compounds

Knockout
Wildtype

Knockout + drug
Wildtype + drug

Positive Controls Show Disease Reversion





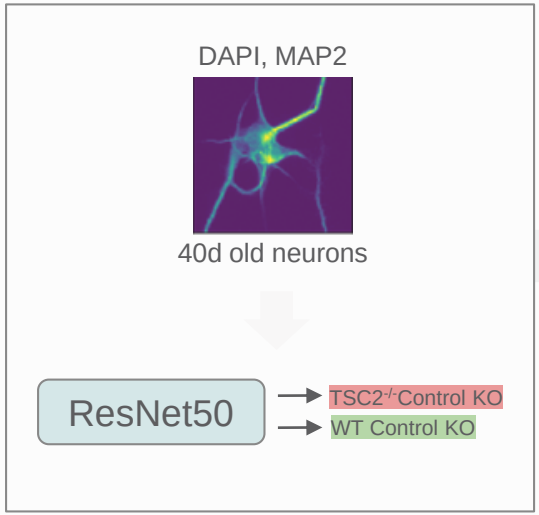
EV, RAPA - known effective compounds
ROT, LOXA - orthogonal compounds

Knockout
Wildtype

Knockout + drug
Wildtype + drug

Morphology Model Predicts Rapamycin Treated and X KO Cells as "Healthy"

1) Train model to classify cells sick ($TSC2^{-/-}$) or healthy (WT) given DAPI and MAP2 staining



2) Test perturbations: Are cells WT-like?

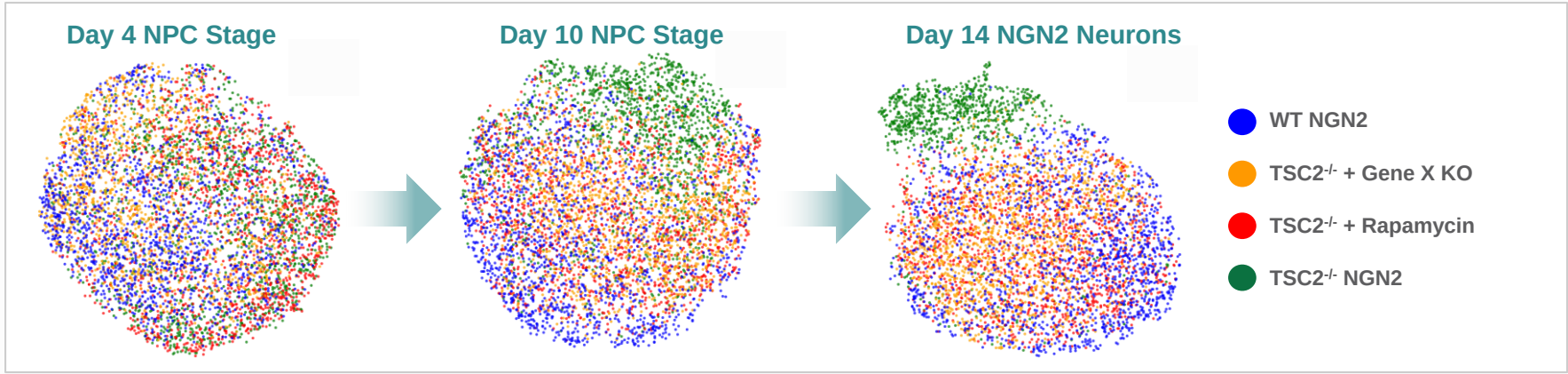
Perturbation	% Predicted as WT Control KO
$TSC2^{-/-}$ +DMSO ctrl	8
$TSC2^{-/-}$ +Rapamycin	91.7
$TSC2^{-/-}$ parental line	7.8
$TSC2^{-/-}$ X KO*	62.3*
$TSC2^{-/-}$ Y KO	19.3
$TSC2^{-/-}$- Z KO	16.4
WT +DMSO ctrl	88.7
WT +Rapamycin	95.2
WT parental line	71.3
WT Control KO	89.6
WT X KO	70.0
WT Y KO	76.1
WT Z KO	70.1

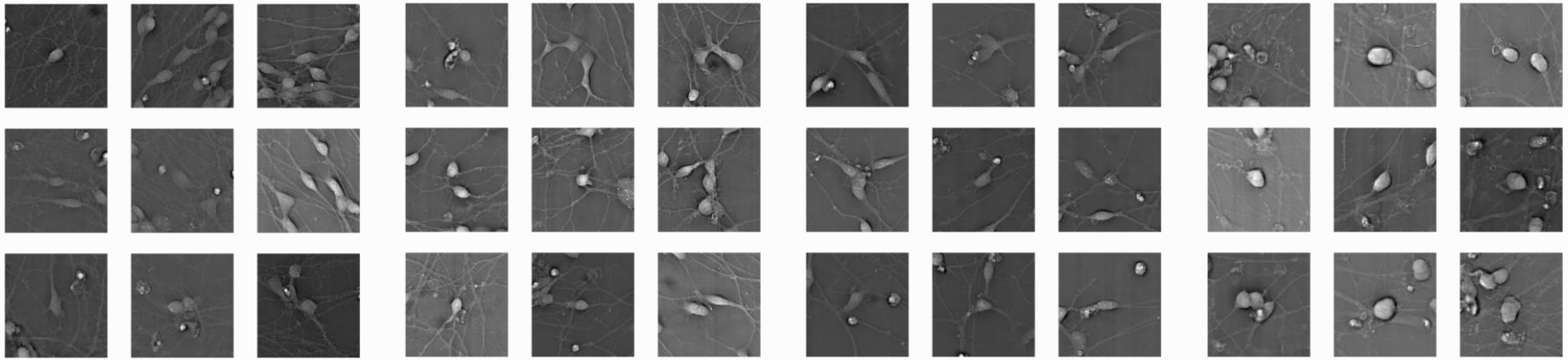
*X KO: 70% efficiency

Model Predictions

- WT perturbations are mostly classified as WT ("healthy")
- $TSC2^{-/-}$ are mostly classified as $TSC2^{-/-}$ ("sick")
- **Rapamycin** treated and **X** KO "sick" cells mostly classified as "healthy"
- **Y** and **Z** KO "sick" cells mostly classified as "sick", consistent with pS6 results
- Some evidence of morphological effects due to **X** KO

Phenotypic Changes and Reversion Visible via Morphology

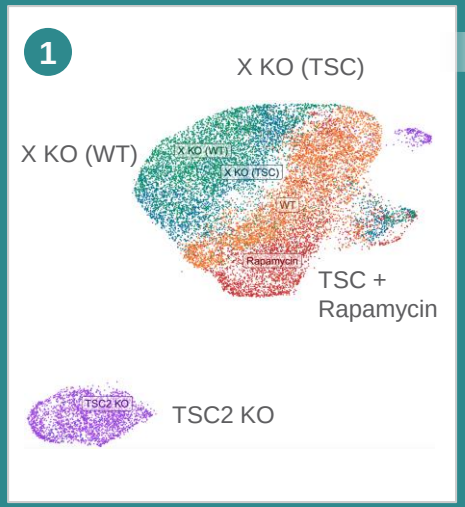




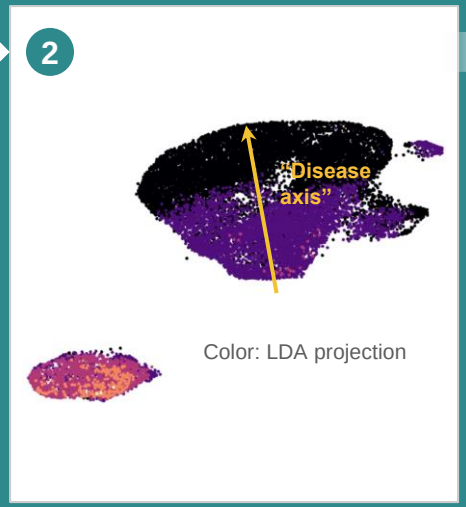
Day 14: Wildtype | Day 14: TSC$^{-/-}$ + Rapamycin | Day 14: TSC$^{-/-}$ + GeneX KO | Day 14: TSC$^{-/-}$

Analysis of scRNA-seq Validates Rapamycin and X as Effective but Distinct Reversions

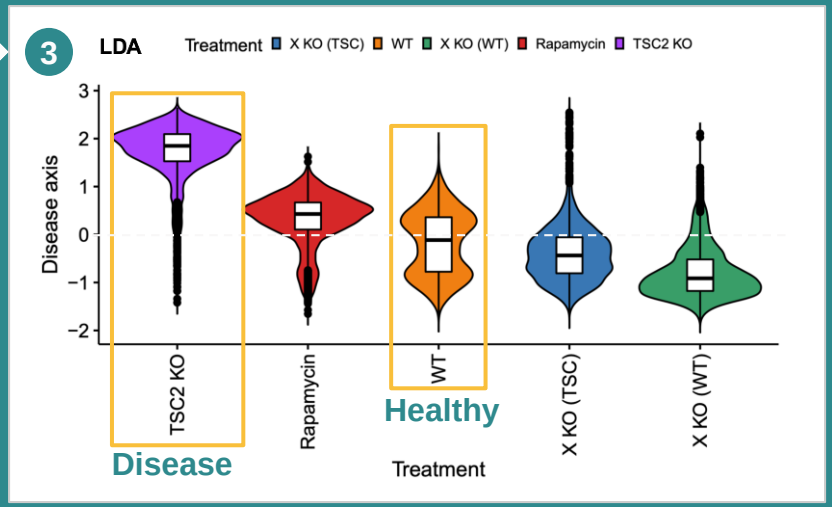
Unbiased Manifold Construction Using ACTIONet



Identifying a "Disease Axis" Using LDA of TSC$^{-/-}$ and WT

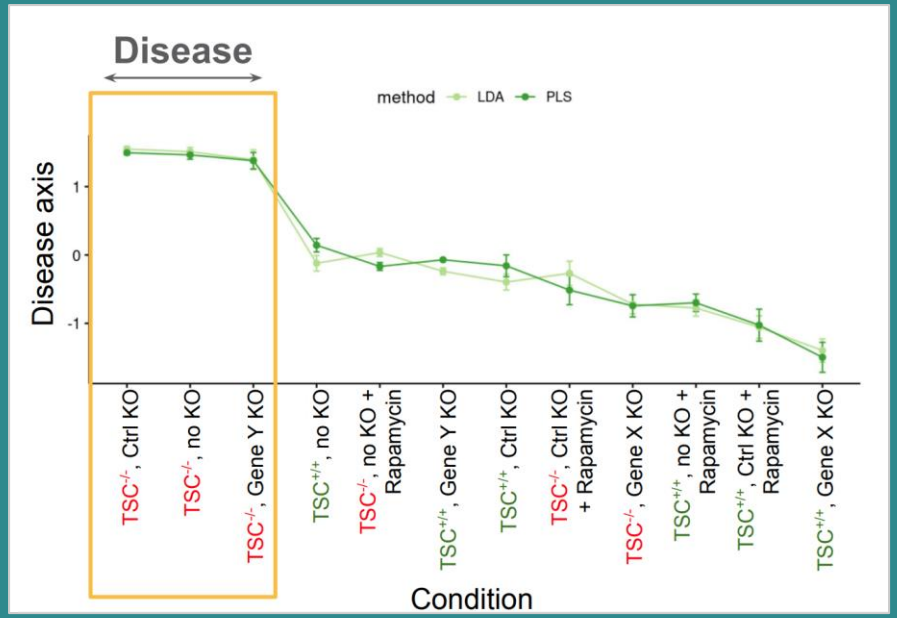


Projection of Cells onto the "Disease Axis"

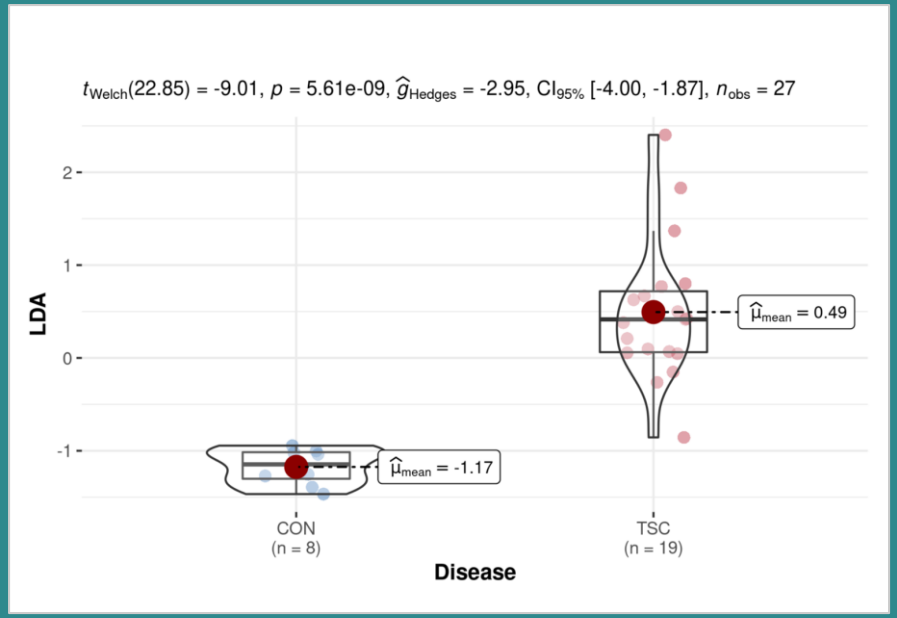


- Both X KO and Rapamycin show reversal characteristics
- Neither X KO or Rapamycin revert cells to exactly the healthy state

Validation of LDA Projections in Bulk RNA-seq Datasets



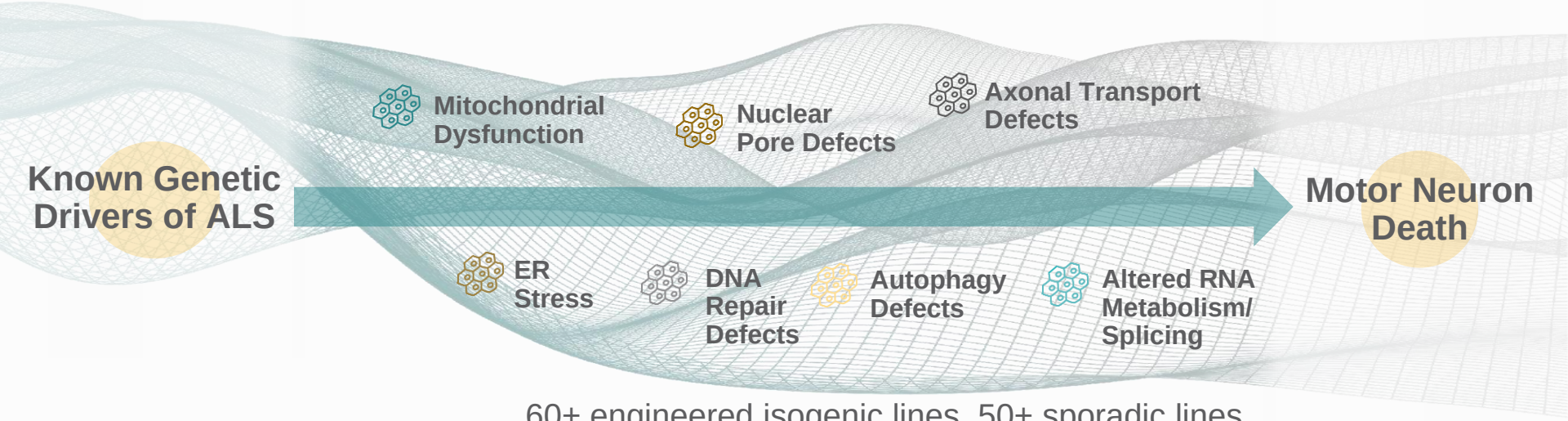
Ordering of Bulk (in-house) Samples Based on their LDA Projections Uncovers Ordering of Samples



Projection of *in-vivo* Samples Using LDA Gene Loadings Separates Based on Disease Phenotype

Goal: Map the Phenotypic Manifold for CNS Disease

Example: ALS Phenotypic Landscape



60+ engineered isogenic lines, 50+ sporadic lines
automated differentiation; dense, multi-modal phenotyping

Identify conserved pathophysiology *across heterogeneous genetic causes* to discern coherent, responsive patient populations and discover high-impact genetic modifiers of disease