Machine Learning: A New Approach To Drug Discovery

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

August 2020
Amazing advances in medicine

Vaccines
Biologics
Cancer immunotherapies
Genetically-targeted therapies
Cell therapies
Low R&D Success & High Costs

"Eroom’s Law": Exponential Decrease in R&D Productivity

This dynamic is driven by a failure to predict downstream outcomes

Big data in biology: Convergence of revolutions

Advances in cell biology and bioengineering allow the creation of immense amounts of data

- iPSC derived models
- Genetic perturbation
- High content phenotyping (e.g., microscopy)
- Automation and microfluidics

Advances in ML can analyze immense amounts of data and leverage them to make predictions

- Preference recommendation
- Image captioning
- Video activity recognition
- Natural language translation
- Speech understanding
- ...

ML is capable of making sense of an immense amount of high-content biological data, most of which is too high-dimensional for humans to interpret
End-to-end models to uncover features

SAMOYED (16); PAPILLON (5.7); POMERANIAN (2.7); ARCTIC FOX (1.0); ESKIMO DOG (0.6); WHITE WOLF (0.4); SIBERIAN HUSKY (0.4)

Machine learning uncovers structure in data
Machine learning uncovers structure in data
What will an intervention do when administered to a person?
Humans as a model for humans
Growing amount of genetic data

Cumulative Number of Human Genomes

1st SANGER
IHGSC et al.
Venter et al.

1st PERSONAL GENOME
Levy et al.

1,000 GENOMES

TODAY
Growth of Genomic Data
Moore’s Law

Source: www.genome.gov/sequencingcosts/
Growing amount of phenotypic data

- Lifestyle, Medical History, Sociodemographic
- Physical Measures
- Environmental Measures
- Urinary Biomarkers
- Genetics Data via the EGA
- Cognitive Function & Hearing Test
- Health Outcome Data
- Genotyping
- Web-based Questionnaire Data
- Physical Activity Monitor
- Imaging

UK Biobank Data
(data collection on 500,000 people over 30 years)
Genetics suggests causality

17 Million variants across recent cohort of 123k individuals
Growing number of genetic associations

Many variants with small effect-sizes

Cumulative SNPs

Source: Visscher et al., AJHG, 2017.
Disentangling mechanism using cell biology

17 Million variants across recent cohort of 123k individuals

High-content biological data gets us closer to the causal biology
The insitro Approach

Build in vitro systems
Use iPSCs and biology at scale to generate multi-dimensional data in optimized format for ML

Build ML models
Use ML to build phenotypic manifolds from huge amounts of *in vitro* model system data

Analyze existing human data
Derive insights from genetic, phenotypic, and clinical data that can guide *in vitro* disease modeling

Generate unique insights on:
- Clinical strategies
- Novel targets
- New uses for existing molecules

Predictive Disease Models
Machine Learning powered alignment of human and *in vitro* data to create more predictive disease models
The cellular phenotypic manifold
ML analysis of Clinical Data to Unravel Disease Processes and Genetic Architecture
NASH disease progression

Normal
- Metabolic syndrome
- Type II diabetes
- Obesity

Hepatocytes
- Small lipid droplets

Steatosis

NASH
- Fibrosis (F1, F2)
- Hepatocyte ballooning
- Inflammation

Cirrhosis
- Advanced Fibrosis

insitro
ML analysis from NASH histopathology images

H&E stained whole slide image

Pathology scores

<table>
<thead>
<tr>
<th></th>
<th>Steatosis</th>
<th>Lob. Infl.</th>
<th>Ballooning</th>
<th>Fibrosis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

Goal: Extract more information from histopathology data

- Better diagnostic tool
- Insight on disease processes and pathophysiology
- Quantitative endophenotypes that reveal novel genetic associations

# samples = 4,641
# tiles ~ 2 millions
ML analysis from NASH histopathology images

H&E stained whole slide image

Pathology scores

<table>
<thead>
<tr>
<th></th>
<th>Steatosis</th>
<th>Lob. Infl.</th>
<th>Ballooning</th>
<th>Fibrosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

Tile features

- $z_1$
- $z_2$
- $z_3$
- $z_4$
- $z_5$

Attention mechanism automatically selects most relevant tiles for each score

Biopsy-level features

# samples = 4,641
# tiles ~ 2 millions

Casale, Bereket, et al., EASL 2020
Disentangling clinical phenotypes

Attention model automatically disentangles the different axes of the disease

t-SNE map of tile representations

Casale, Bereket, et al., EASL 2020
Predictive accuracy

Cross validation performed on a disjoint set of clinical centers

Pathologist calls require trichrome; ML uses only H&E

\( \rho = \text{Spearman correlation} \)

Cross-pathologist Spearman’s rank correlation is generally \( \sim 0.7 \)

Casale, Bereket, et al., EASL 2020
ML phenotypes better capture disease state

- CNN scores improve predictions over pathologist scores for the majority of 60 NASH-related serum biomarkers (7 significant, adjusted P < 0.05)

- CNN scores identified 131 genes significantly associated with fibrosis progression in liver RNA-seq, as compared to 1 gene when using the conventional pathologist scores

Casale, Bereket, et al., EASL 2020
CNN scores enable GWAS of disease progression

- We performed a genome-wide association study of disease progression in STELLAR patients (n~550)
- This analysis yielded 2 novel & compelling genome-wide significant genes and several suggestive loci
- No significant locus was retrieved when conducting the same analysis based on pathologist scores
Summary: ML analysis of clinical data

- ML-based analysis of histopath data predicts NASH histological scores with similar accuracy to cross-pathologist agreement solely from biopsy-level labels (no pixel-level annotations)

- The learned histological scores provide greater resolution, empowering association analyses with clinical labels and genetic association studies

- This work provides a proof of principle for the ability of machine learning to empower downstream association analyses, potentially leading to the identification of candidate targets and the discovery of new disease biology
Predictive Cellular Models for Human Clinical Outcome
A Cellular Model for NASH

NASH tiles

non-NASH tiles

A Cellular Model for NASH

Training set

Validation set

A Cellular Model for NASH

Highest ranked tiles by class

NASH

Non-NASH

Tuberous Sclerosis Complex

Tuberous Sclerosis Complex (TSC) is caused by a heterozygous germline mutation in the TSC1 or TSC2 gene.

TSC presents in 1 in 6,000 live births.

Patients experience growths throughout the body, including the skin, lungs, kidneys, and heart.

The greatest burden to quality of life is caused by hamartoma formation in the brain.

Severe, often intractable epilepsy occurs in 70%-80% of patients.

TSC is also strongly linked to intellectual disability and autism.

CNS iPSC models: TSC2^- neuronal phenotypes

- TSC2^- iPSC-derived neurons show a strong cellular phenotype in specialized assays
- That phenotype is reverted using drugs that target mTORC1
- This provides us with ground truth and a “positive control” for reversion of disease phenotype
Identifying disease phenotypes

- Use of an untargeted (generic CellPaint) assay makes phenotypes hard to discern
- Experiment was cut short due to COVID-19 lab shutdown, before strong disease phenotypes had a chance to manifest (40 day maturation cut down to 15)
- **Our ML models reliably distinguished healthy and sick phenotypes invisible to experts**

Max Salick, Adam Riesselman, et al., unpublished.
Identifying disease reversion
Identifying disease reversion

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

Max Salick, Adam Riesselman, et al., unpublished.
Identifying disease reversion

Low | Mid-Low | Mid-High | High
---|---|---|---
Everolimus

Rapamycin

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

Max Salick, Adam Riesselman, et al., unpublished.

- Knockout
- Wildtype
- Knockout + drug
- Wildtype + drug
A cellular model for high-penetrance SCZ mutations

Perfect prediction of TSC2 mutant neurons matches expectations due to extremely strong neuronal phenotype.

Overlap of TCF4 with negative control suggests a more developmental role of TCF4 leading to Pitt-Hopkins.

Grouping of GeneX clones, and separation from controls, indicates a new ML-identified schizophrenia phenotype.

Max Salick, Adam Riesselman, et al., unpublished.
Summary: ML analysis of clinical data

- Cellular models show subtle, disease-related phenotypes even in generic assays, and even when those phenotypes cannot be detected by humans.

- Cellular manifolds induced via machine learning can provide insights on interventions that might revert the disease.

- This is a proof-of-concept for our broader goal of building iPSC-derived models for human disease and identifying novel interventions that might modulate human clinical outcomes.
Epochs of science

1800s

Physics

Chemistry

1900s

Computing

1950s

Data

1990s

Biology
The next epoch

Digital Biology

2020s