WYSS INSTITUTE

# Mathematical Representations
# For Biological Systems

Sahil Loomba

17th August, 2018

*Hansjörg Wyss Institute for Biologically Inspired Engineering at Harvard University*

# Overview

- Brief Perspective on Modeling in Biology

# Overview

- Brief Perspective on Modeling in Biology
- Systems Biology: Discovery of biological mechanisms
    - Markov Interaction Network: Uncertainty meets Knowledge
    - Hypothesis discovery as arbitrary probabilistic queries

# Overview

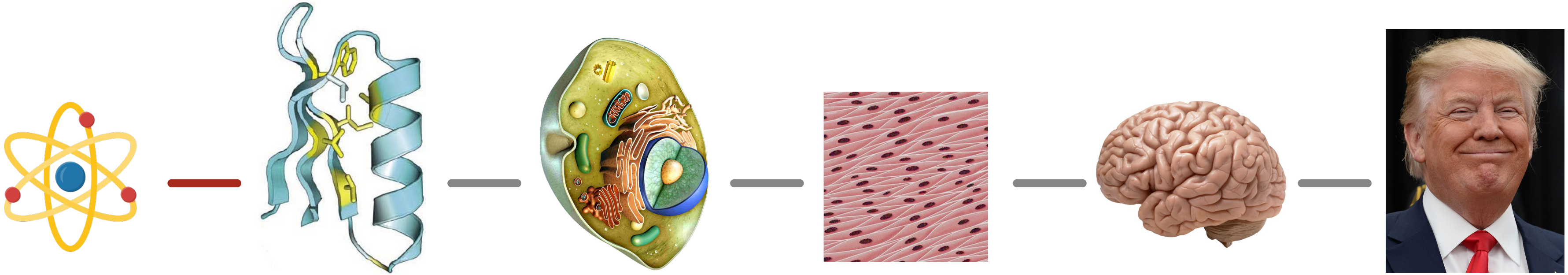
- Brief Perspective on Modeling in Biology
- Systems Biology: Discovery of biological mechanisms
  - Markov Interaction Network: Uncertainty meets Knowledge
  - Hypothesis discovery as arbitrary probabilistic queries
- Synthetic Biology: Design of bioengineered parts
  - Language embedding models to represent biomolecules
  - Design challenges as arbitrary downstream machine learning

# Overview

- Brief Perspective on Modeling in Biology
- Systems Biology: Discovery of biological mechanisms
  - Markov Interaction Network: Uncertainty meets Knowledge
  - Hypothesis discovery as arbitrary probabilistic queries
- Synthetic Biology: Design of bioengineered parts
  - Language embedding models to represent biomolecules
  - Design challenges as arbitrary downstream machine learning
- Closing the loop on iterative discovery and design
  - Organism-on-Chip for high-throughput drug discovery

# Biology Exhibits Hierarchical Compositionality

## Principle of Abstraction

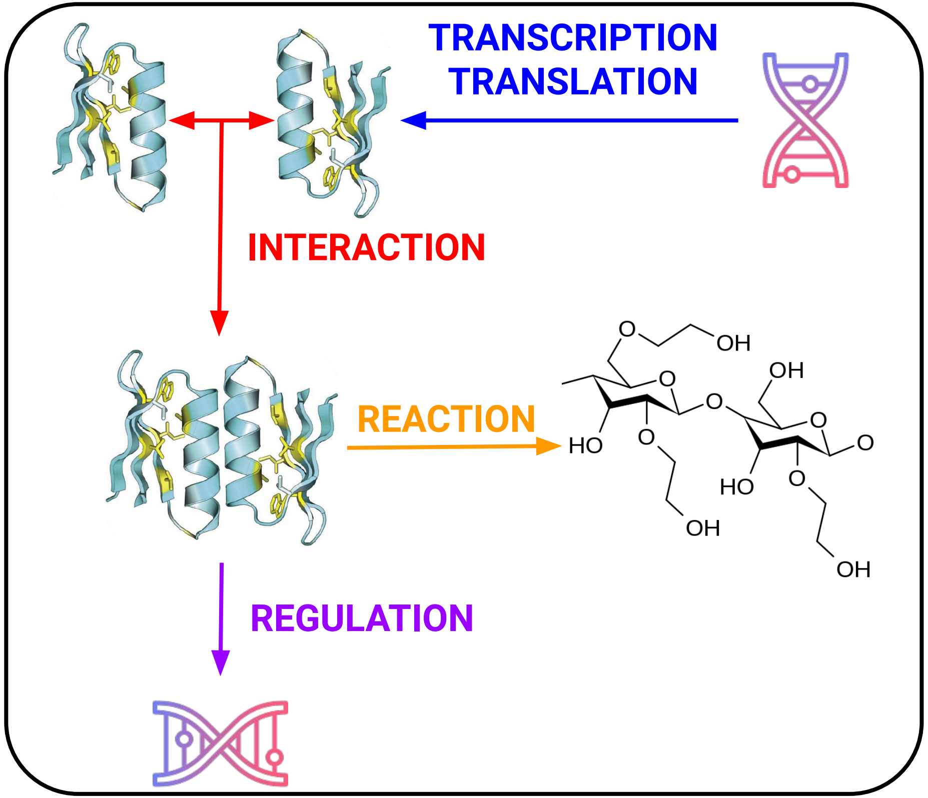


| ATOMS | BIOMOLECULES | CELLS | TISSUES | ORGANS | ORGANISM |
|---|---|---|---|---|---|
| Mass Spectrum, Molecular Properties | Structure, Transcriptomics, Proteomics, Metabolomics | Cellular Phenotype | Images (Histology), EEG | Images (MRI), fMRI | Organismal Phenotype |

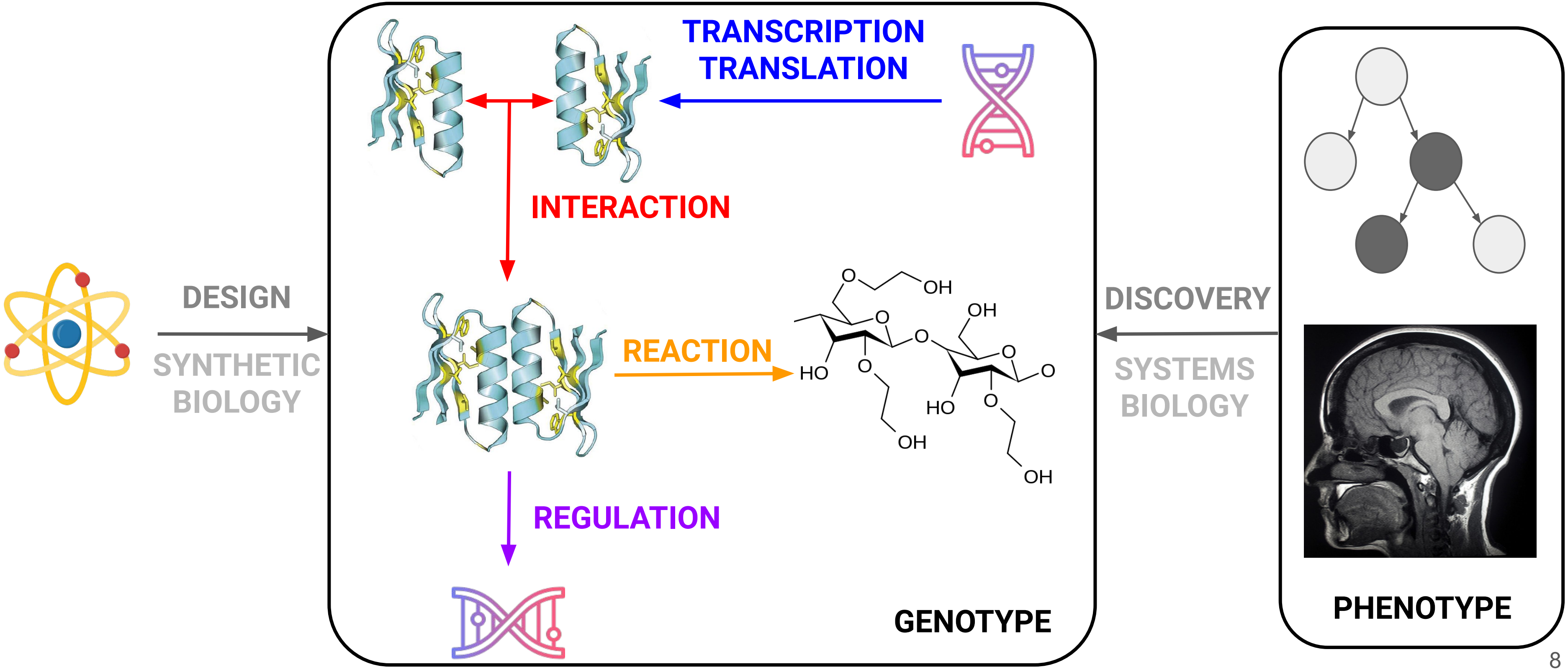
$\xrightarrow{\quad f \quad}$

$\xleftarrow{\quad f^{-1} \quad}$

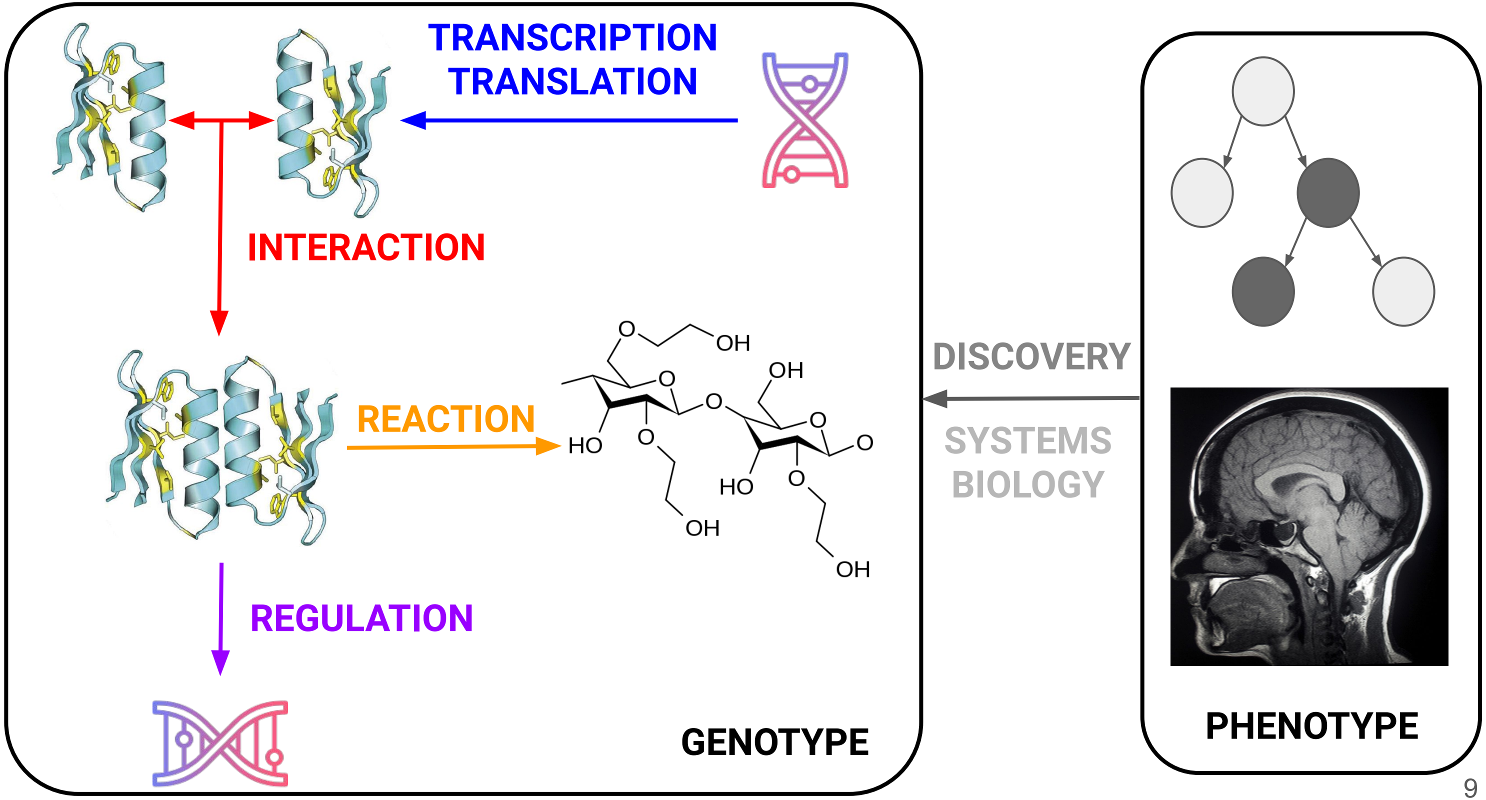
# Biology Exhibits Hierarchical Compositionality



# Biology Exhibits Hierarchical Compositionality



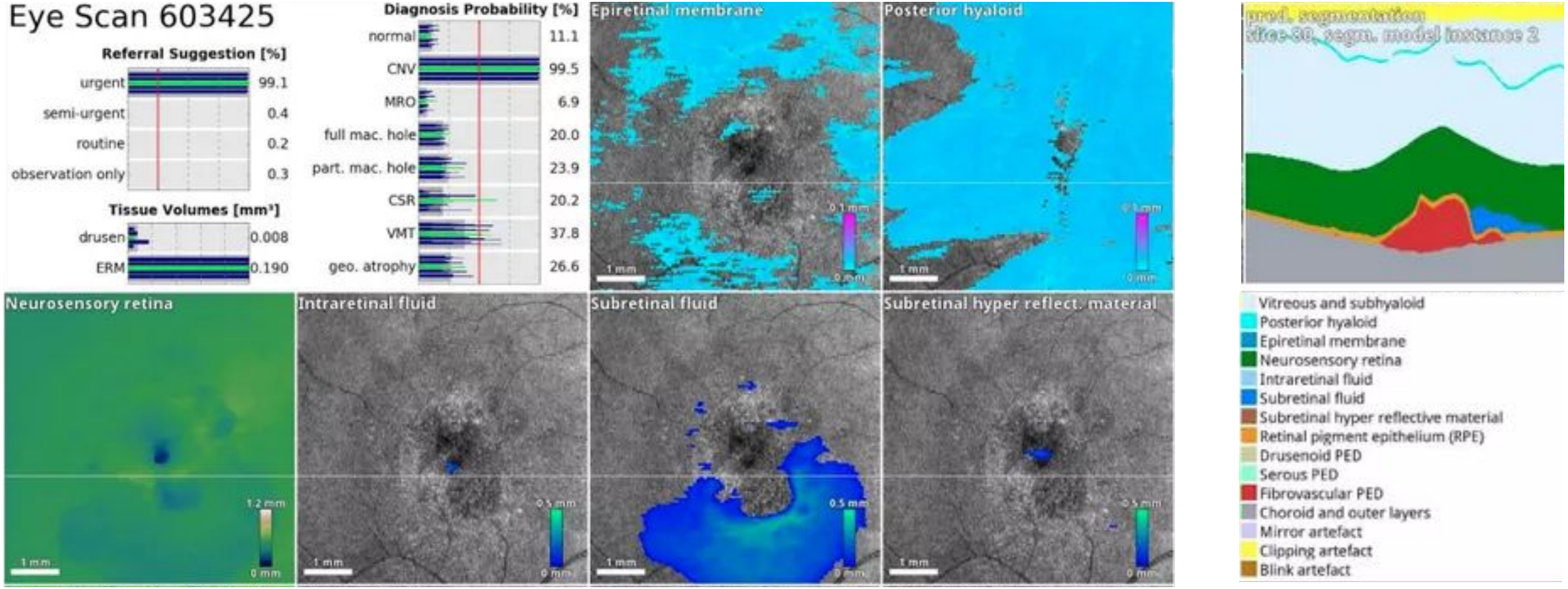
# Problems of Discovery in Systems Biology



# DeepMind's AI can detect over 50 eye diseases as accurately as a doctor

*The system analyzes 3D scans of the retina and could help speed up diagnoses in hospitals*

By James Vincent | @jjvincent | Aug 13, 2018, 11:01am EDT



# Biology: A Modeling Perspective

**A Typical "Machine Learning" Problem**

- Clearly defined input and output
- Large amount of data (usually cheap to acquire)
- Availability of labels: supervised
- Problems in medical biology at **phenotype level**

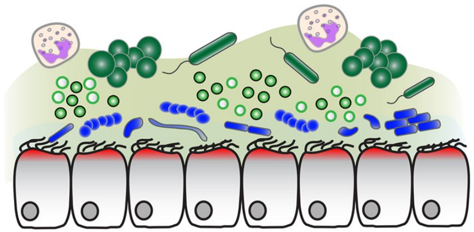
**Atypical "Machine Learning" Problem**

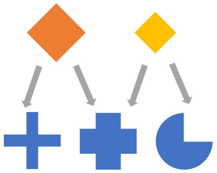
- Arbitrary query of interest
- Small amount of data (usually expensive to acquire)
- Few to no labels: semi-supervised
- Problems in systems and synthetic biology at **genotype level**

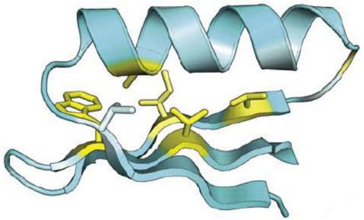
Wyss Institute

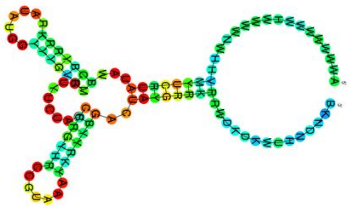
# Work at Wyss: A Biological Perspective

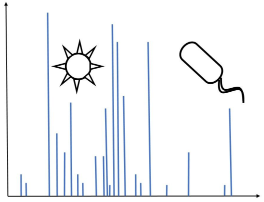
- Discovery (Systems Biology)
  - Mechanism of tolerance to pathogens
  - Countermeasures to induce tolerance
- Design (Synthetic Biology)
  - Protein Stability Problem
  - Riboswitch Design Challenge
- Diagnostics (Medical Biology)
  - Predicting breathing severity in Asthma
  - Mass Spec for Pathogen Detection







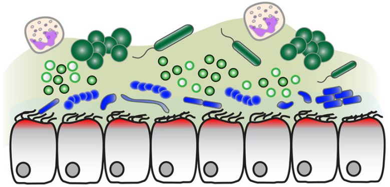


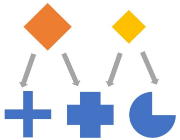



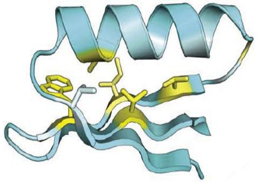
# Work at Wyss: A Modeling Perspective

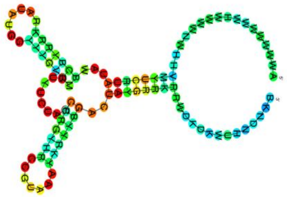
## A Typical "Machine Learning" Problem

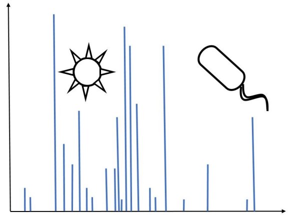

## Atypical "Machine Learning" Problem



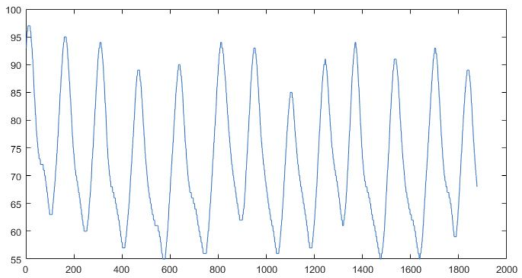




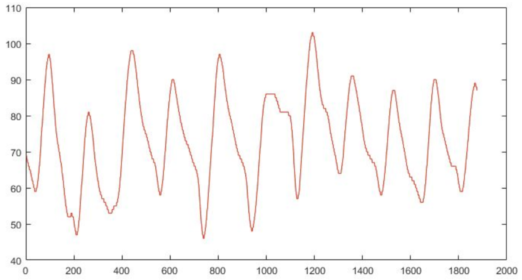




# Typical Machine Learning | case-in-point
## *Predicting Breathing Severity in Asthma*

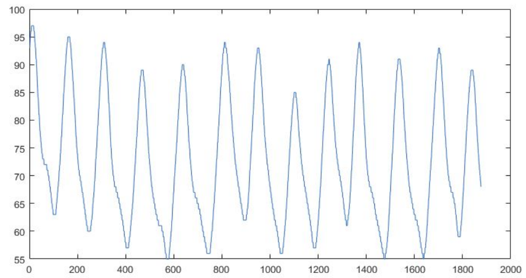



**LOW SEVERITY**

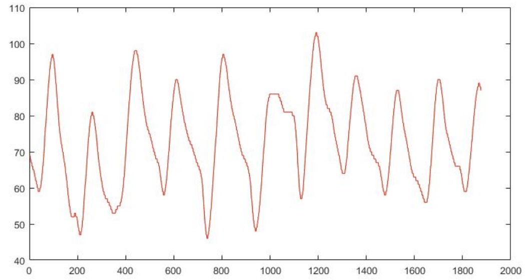



**HIGH SEVERITY**

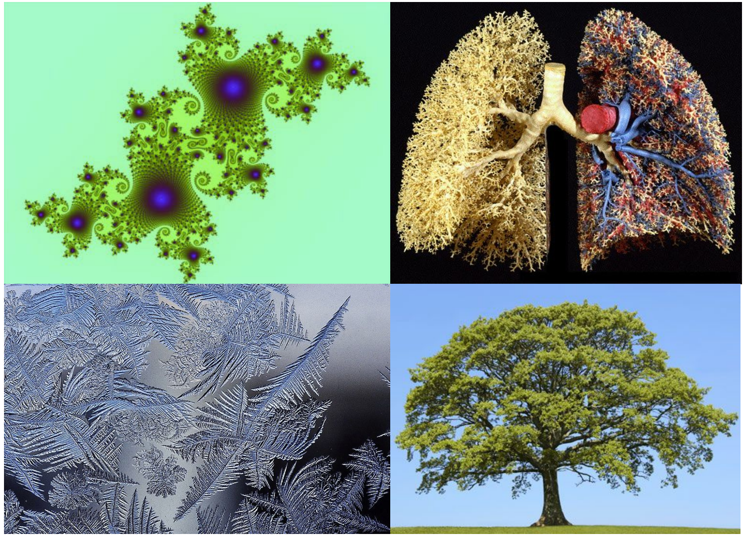
# Typical Machine Learning | case-in-point
*Predicting Breathing Severity in Asthma*



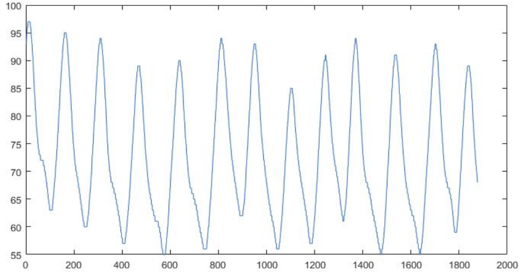

**LOW SEVERITY**



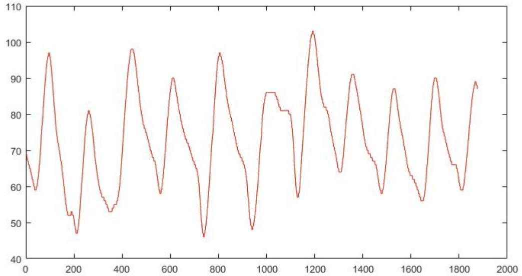

**HIGH SEVERITY**



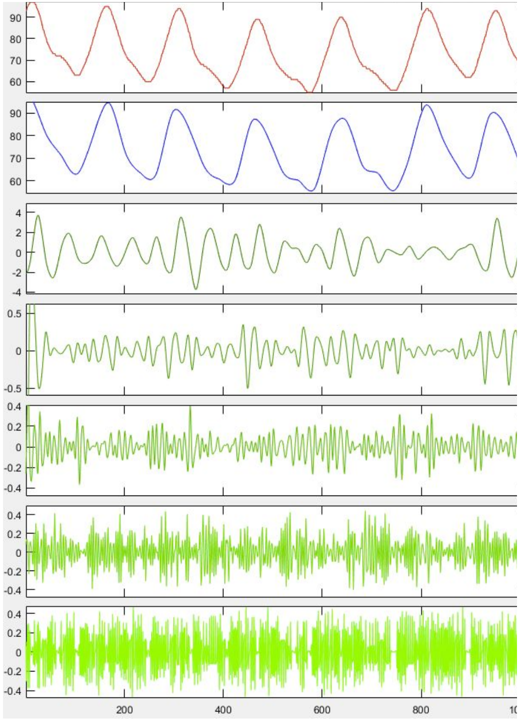
# Typical Machine Learning | case-in-point
## *Predicting Breathing Severity in Asthma*




**LOW SEVERITY**



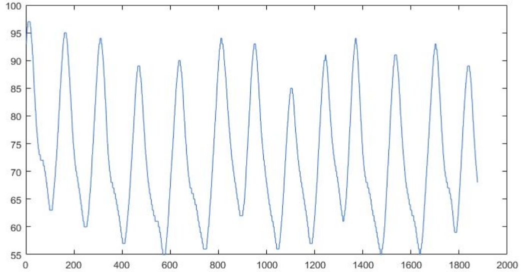

**HIGH SEVERITY**



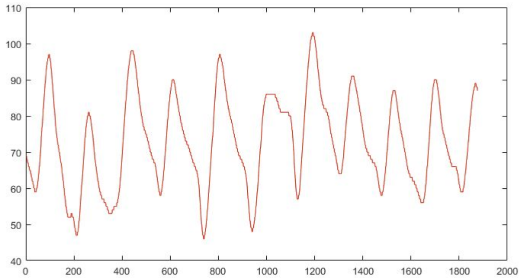
WYSS INSTITUTE

# Typical Machine Learning | case-in-point

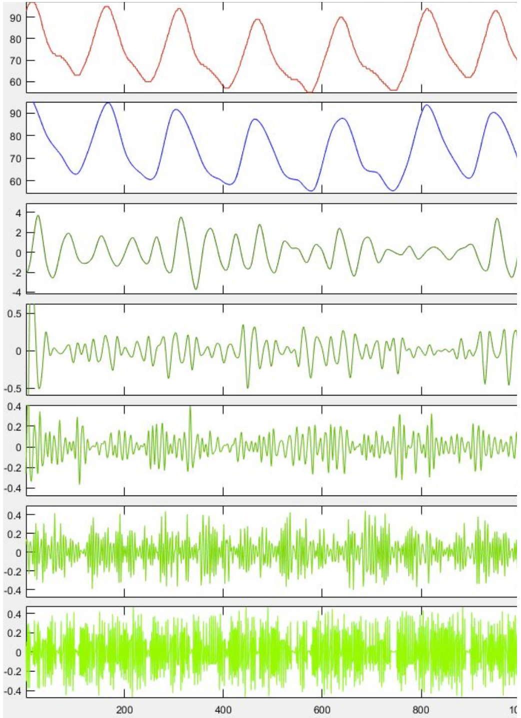
## *Predicting Breathing Severity in Asthma*



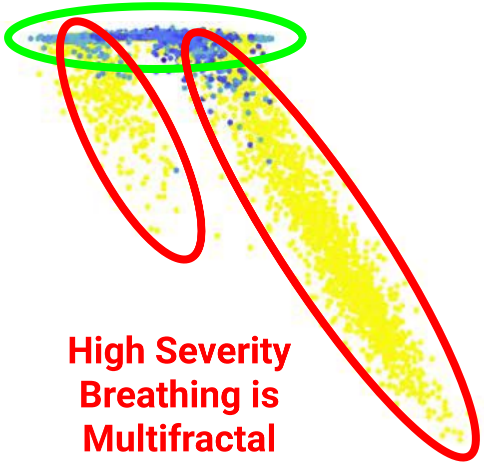

**LOW SEVERITY**




**HIGH SEVERITY**



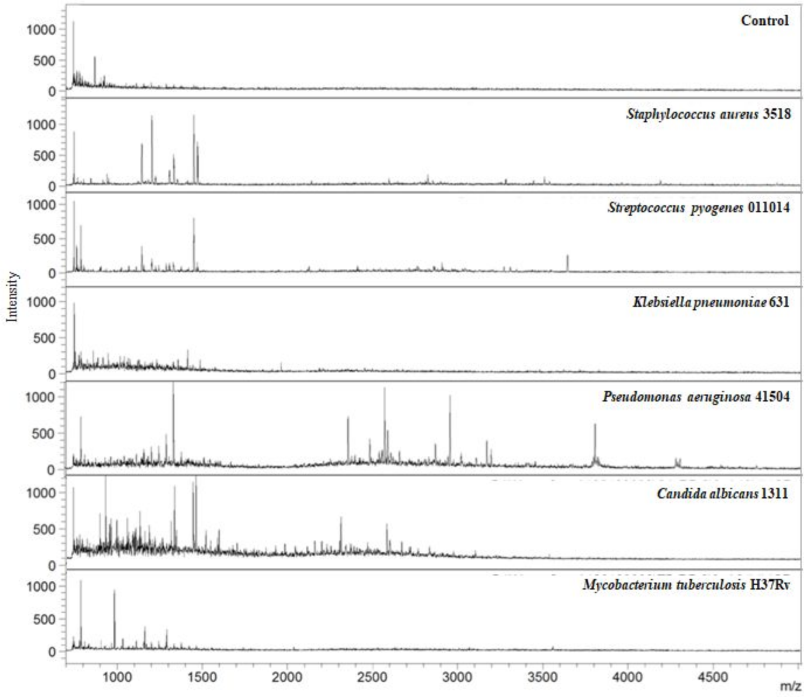

Low Severity Breathing is Monofractal



High Severity Breathing is Multifractal

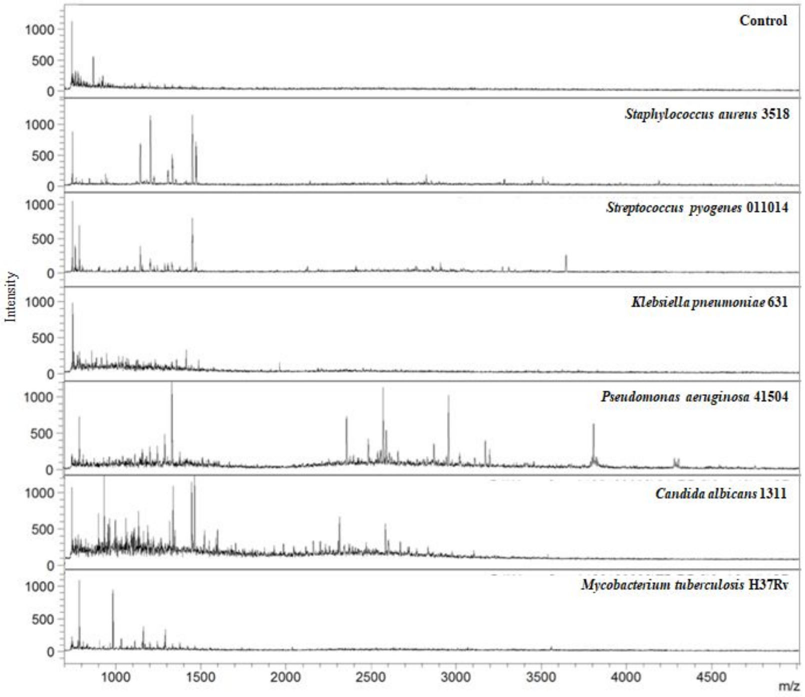
**Simple ML model on this 2D space gives prediction accuracies of upto 97%**

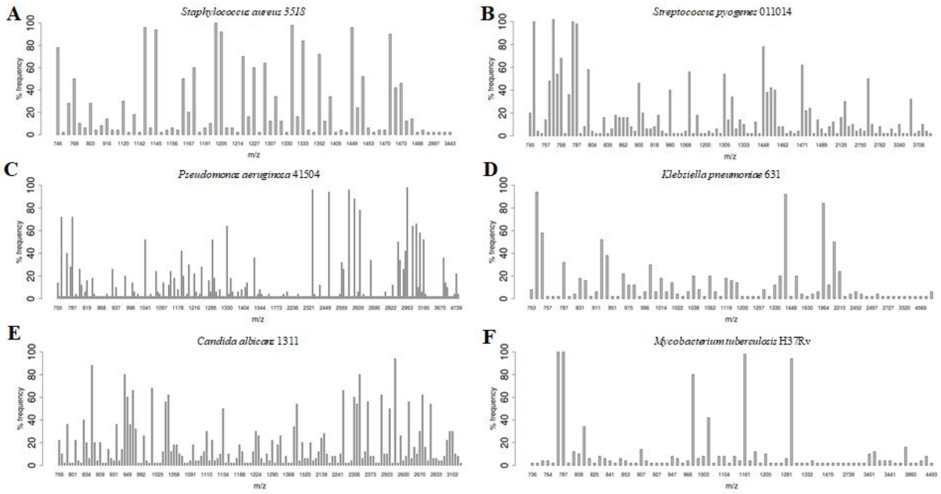
# Atypical Machine Learning | case-in-point
## *MALDI-TOF MS for Pathogen Detection*




**Shannon Duffy**

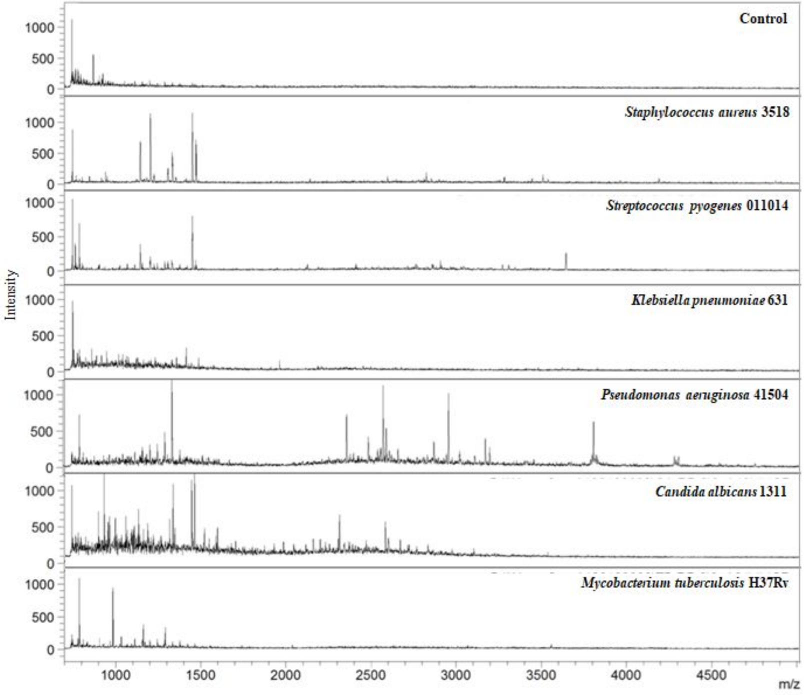
# Atypical Machine Learning | case-in-point
## *MALDI-TOF MS for Pathogen Detection*

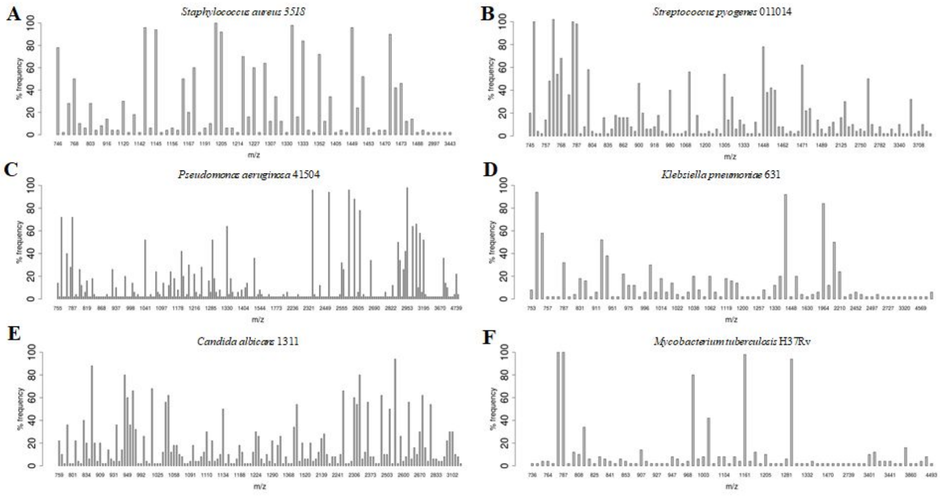



**Shannon Duffy**

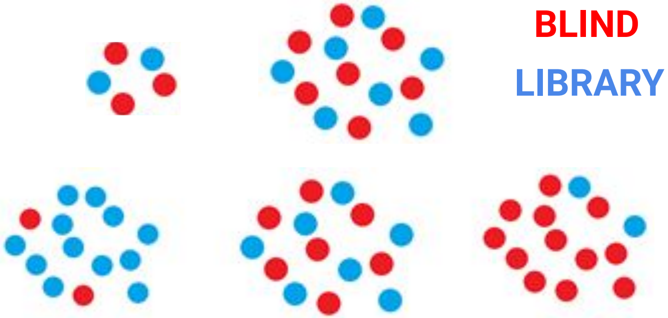
# Atypical Machine Learning | case-in-point
## *MALDI-TOF MS for Pathogen Detection*




**Cluster Frequency Weighting**

**Cluster Proportion Weighting**



**BLIND**

**LIBRARY**

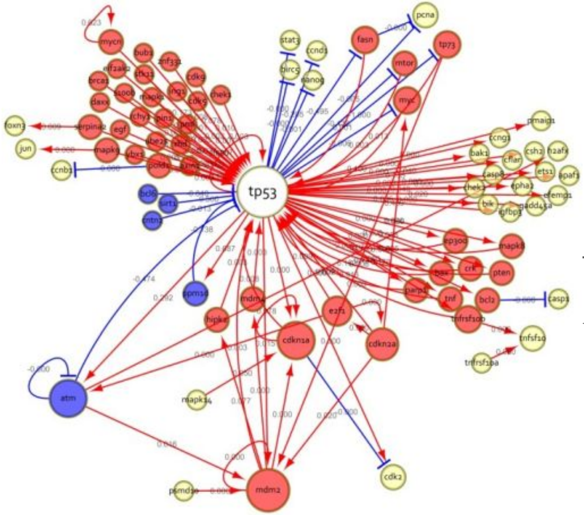
**Shannon Duffy**

**Accuracy of 100% on a Probabilistic Model with 12 blind samples**

# Core Tenets for Mathematical Representations in Biology

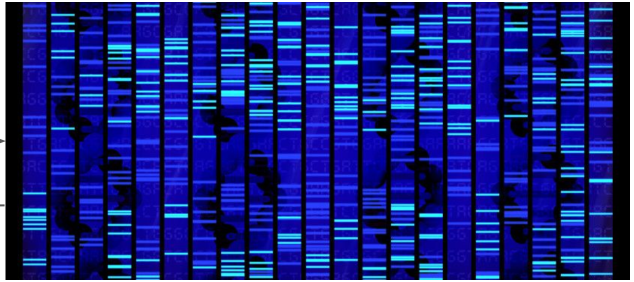
- Biology is <u>uncertain</u> and stochastic
- Biology is <u>not discrete</u>
- Biology is <u>complex</u>, but has some core underlying "latent" principles that govern "observed" complex phenomena

- Explicitly incorporate uncertainty through <u>probabilistic</u> modeling
- Develop "soft", <u>continuous</u> and distributed representations
- Include domain <u>knowledge</u> to define structure of core variables that generate evidence



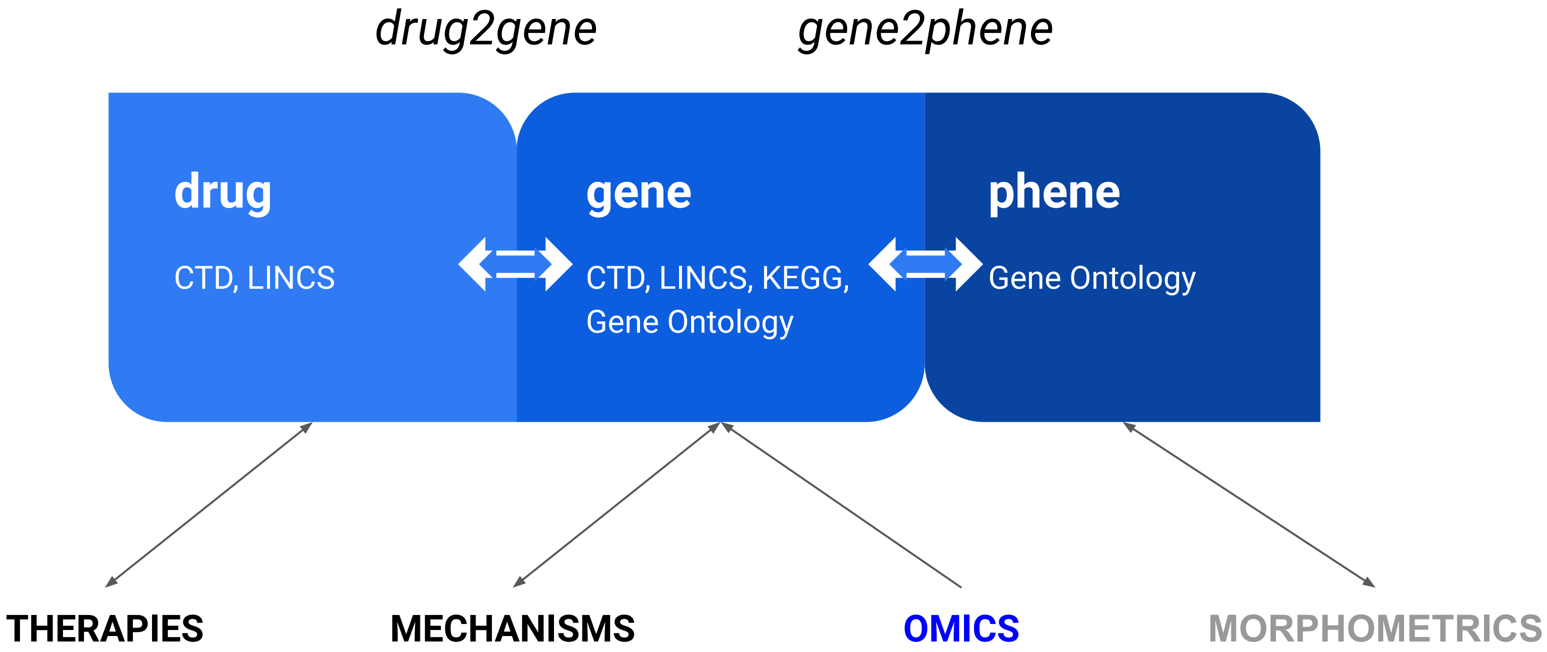

$f$

$f^{-1}$



# Uncertainty Meets Knowledge

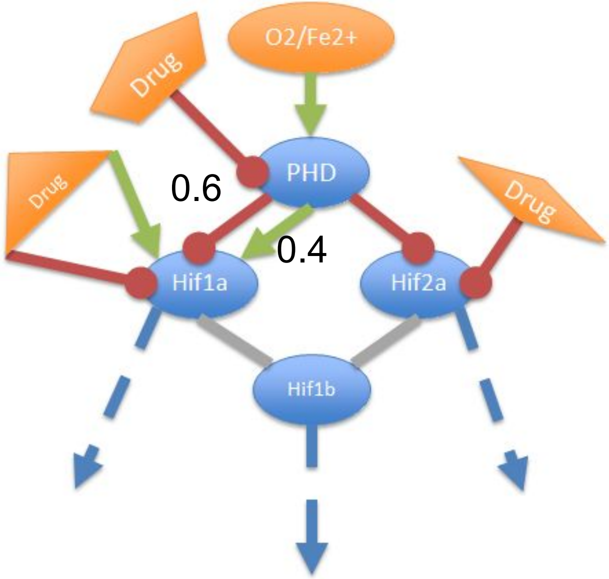
## With Great Knowledge Comes Great Power



# Uncertainty Meets Knowledge: *NeMoCAD*

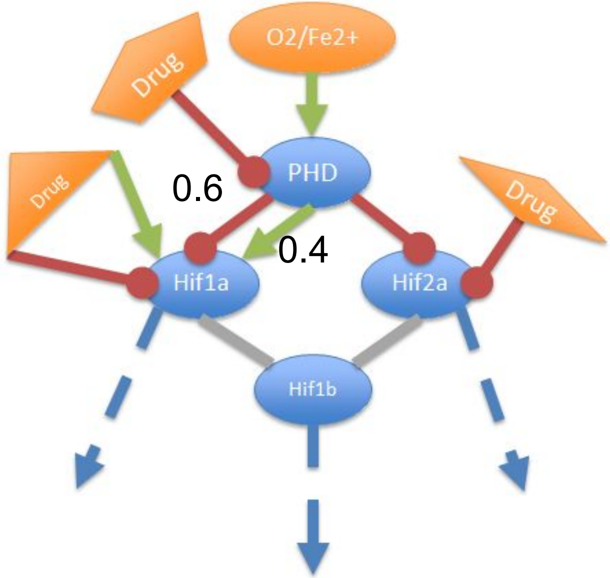
## Network Model for Causality Aware Discovery

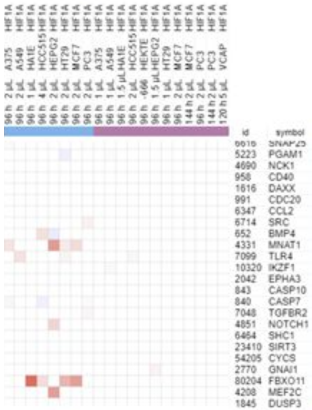
- **Network** of interactions between genes encode routes of causal mechanistic influence in a biological system
- **Probabilistic** weights obtain from gene knockout and/or overexpression data encode weight of causal interactions to form a Markov Network
- Arbitrary mechanistic queries can be turned into corresponding probabilistic inference queries for **discovery**
- Adding **drugs** as nodes of the network also allows us to discover new drug **therapies**

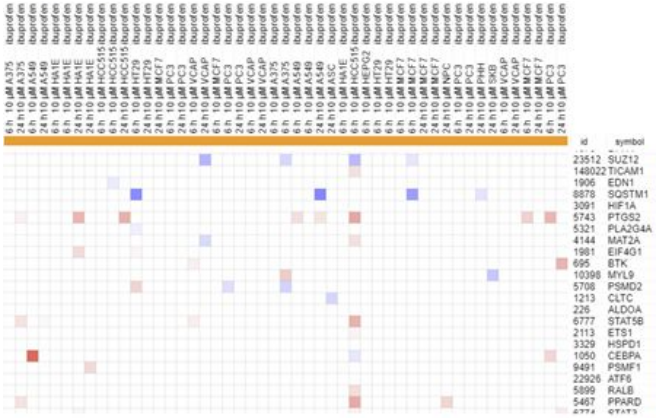


# Uncertainty Meets Knowledge: *NeMoCAD*

## Network Model for Causality Aware Discovery

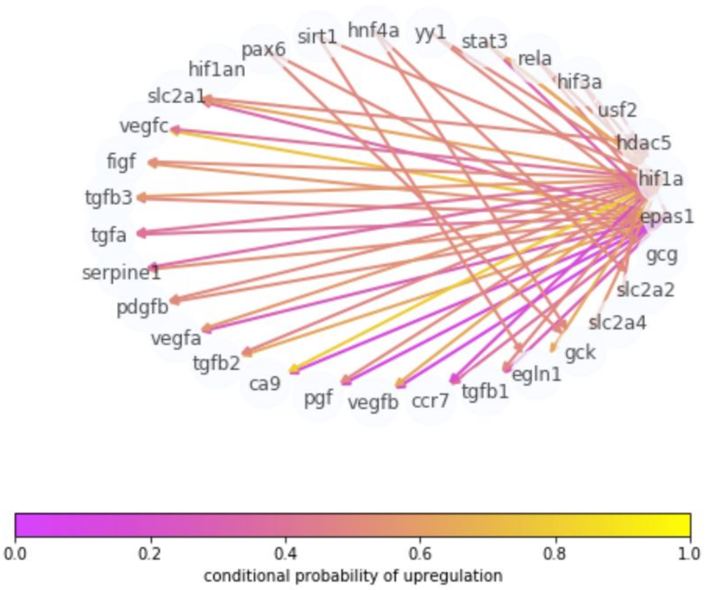




# Uncertainty Meets Knowledge: *NeMoCAD*

## Drug Discovery | gene2drug




**MARKOV INTERACTION NETWORK**

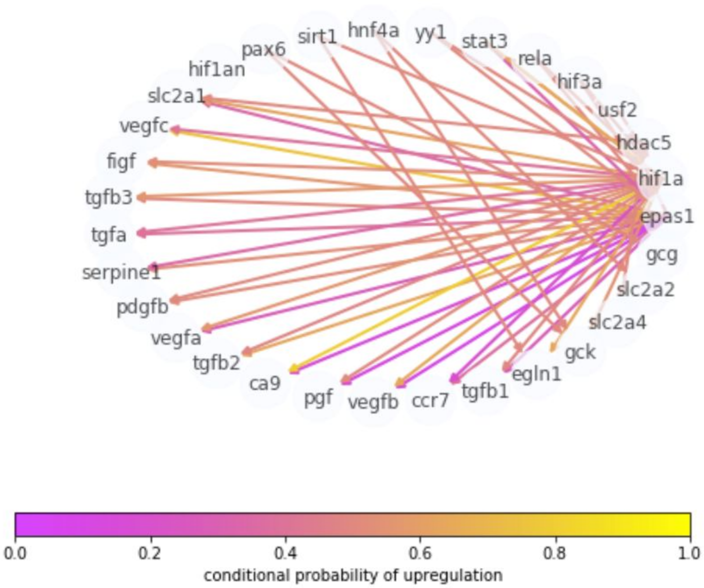
```
{'hif1a':1, 'epas1':0}
```

**LOOPY BELIEF PROPAGATION**

WYSS INSTITUTE

# Uncertainty Meets Knowledge: *NeMoCAD*

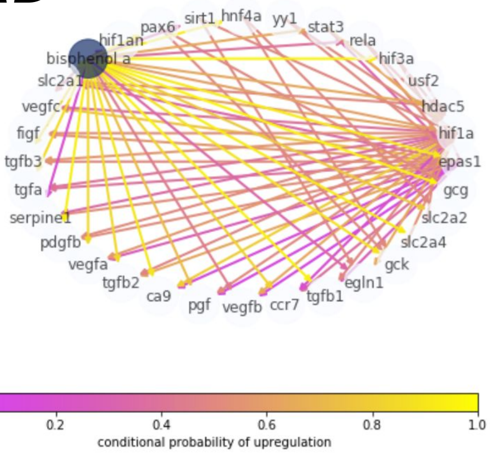
## Drug Discovery | gene2drug

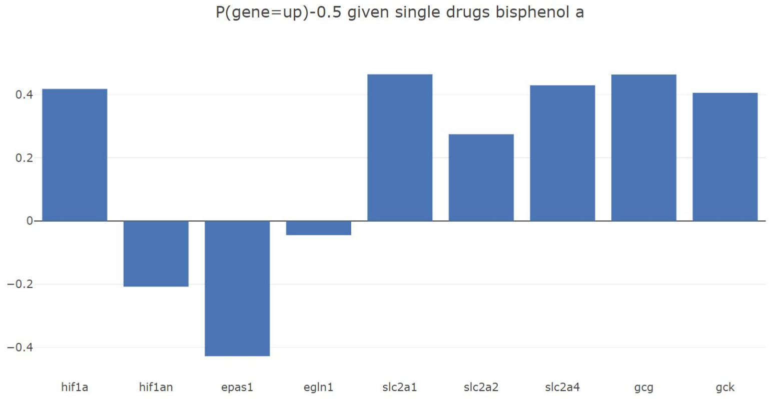



**MARKOV INTERACTION NETWORK**

```
{'hif1a':1, 'epas1':0}
```

**LOOPY BELIEF PROPAGATION**



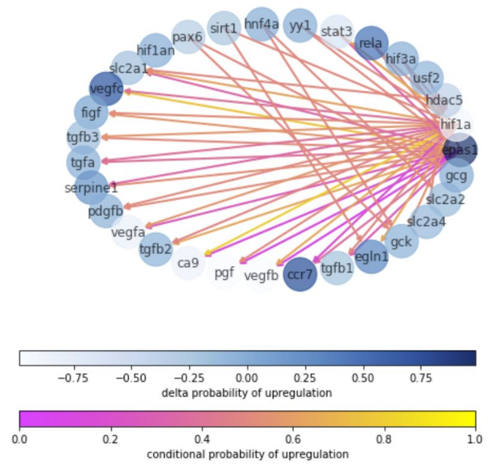



# Uncertainty Meets Knowledge: *NeMoCAD*

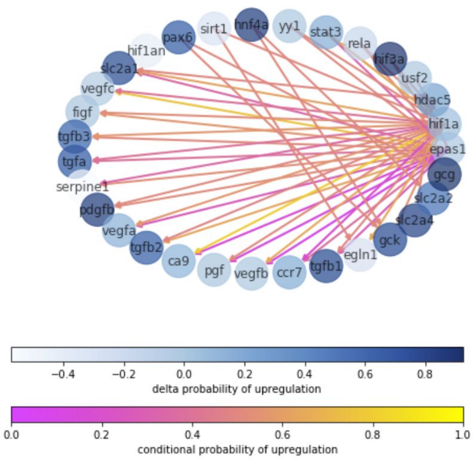
## Drug Combination Investigations | gene2drug

```
{'hif1a':1, 'epas1':0}
```

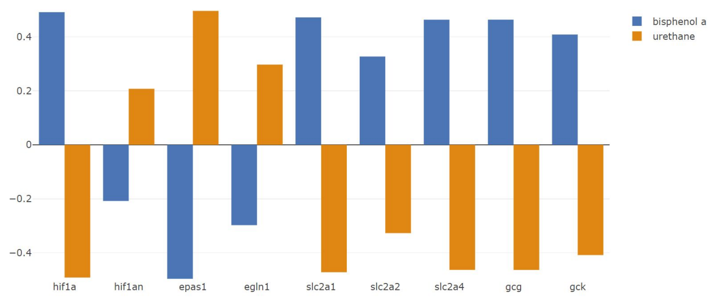
combo bisphenol a, urethane versus only bisphenol a



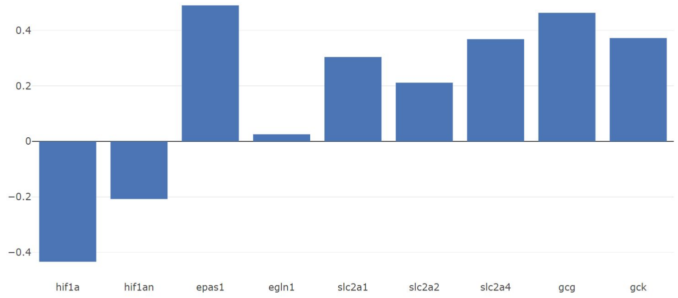

combo bisphenol a, urethane versus only urethane




P(gene=up)-0.5 given single drugs bisphenol a, urethane

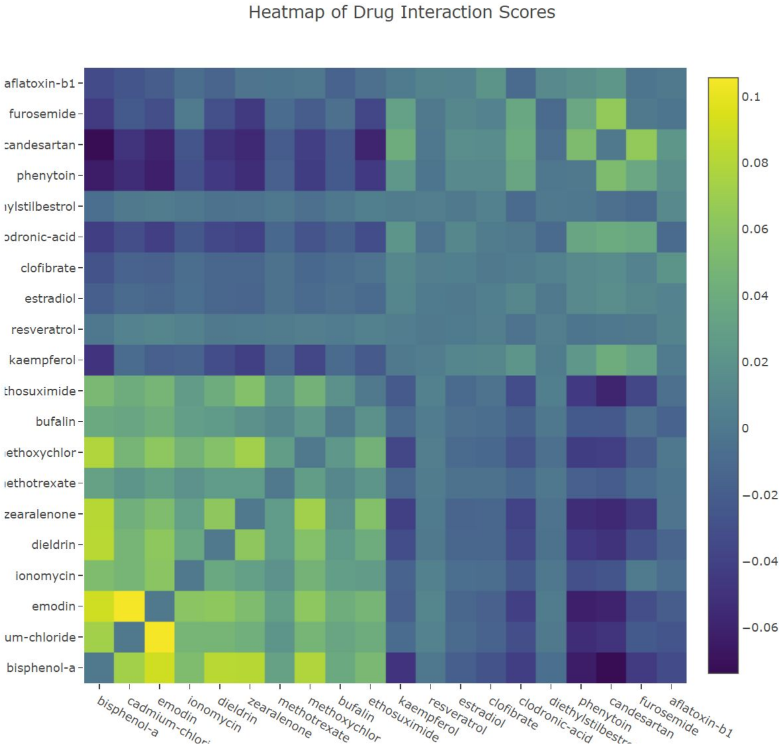



P(gene=up)-0.5 given combo drugs bisphenol a, urethane

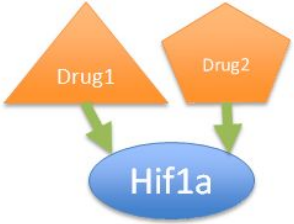


# Uncertainty Meets Knowledge: *NeMoCAD*

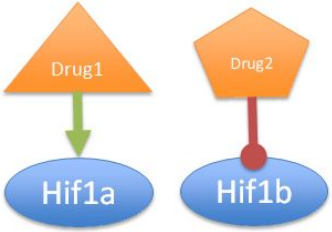
## Drug Synergy Contextualized by Gene Space | drug2drug



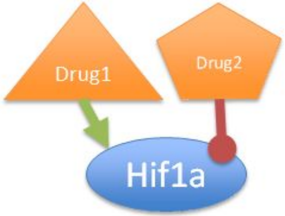

*Synergy*




*Orthogonality*



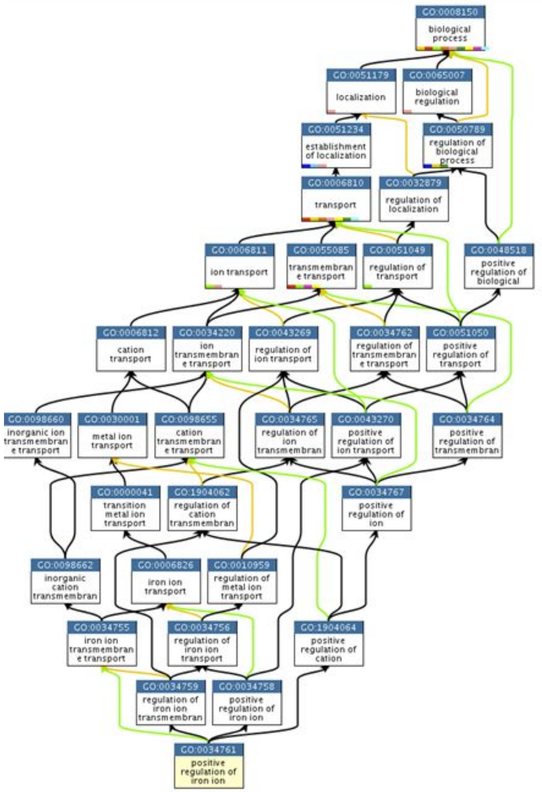

*Antagonism*



WYSS INSTITUTE

# Uncertainty Meets Knowledge: *NeMoCAD*

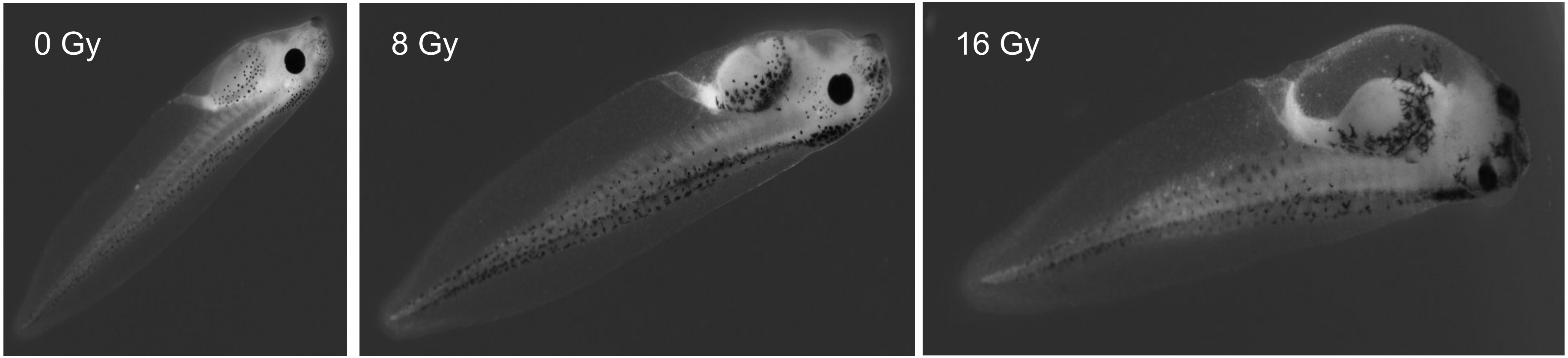
## Incorporating Gene Ontology | gene2phene




```
trnt1     GO:0003723,GO:0006396,GO:0016779
nr5a2     GO:0003700,GO:0003707,GO:0005634,
tbx1
0006351,GO:0007275,GO:0045944,GO:0050793,
tbx1.L
0006351,GO:0007275,GO:0045944,GO:0050793,
nr1d1     GO:0003700,GO:0003707,GO:0005634,
nucb1     GO:0005509
nsa2      GO:0000460,GO:0000470,GO:0005730,
csnk1a1 GO:0004674,GO:0005524
csnk1a1.L     GO:0000777,GO:0004674,GO:0005
):0045104,GO:0051301
hoxc6     GO:0003700,GO:0005634,GO:0007275,
sorbs2    GO:0007015,GO:0007155,GO:0016477
apln      GO:0001664,GO:0005179,GO:0005615,
fzd4      GO:0004930,GO:0007275,GO:0016021,
suclg1    GO:0000166,GO:0004775,GO:0004776,
sostdc1 GO:0005615
tcf3      GO:0000790,GO:0001710,GO:0001714,
atat1
0005874,GO:0005905,GO:0005925,GO:0019799,
rab28     GO:0003924,GO:0005525,GO:0005622,
rhob      GO:0003924,GO:0005525,GO:0005622,
zbtb16    GO:0003676,GO:0046872
fzd9
0016021,GO:0016055,GO:0017147,GO:0035567,
egfl8     GO:0005509
ddx39b    GO:0003676,GO:0005524
fgg GO:0005102,GO:0005577,GO:0007165,GO:0
fgg.L     GO:0005102,GO:0005577,GO:0007165,
rassf2    GO:0007165,GO:0042981
nr1d2     GO:0003700,GO:0005634,GO:0006351,
mapk6     GO:0004707,GO:0005524,GO:0005622,
crx
0007275,GO:0009952,GO:0042706,GO:0043565,
cga GO:0005179,GO:0005576
rxrg      GO:0003700,GO:0003707,GO:0005634,
hoxb2.S GO:0003700,GO:0005634,GO:0007275,
rora      GO:0004879,GO:0005634,GO:0006351,
```

WYSS INSTITUTE

# Uncertainty Meets Knowledge: *NeMoCAD*
# Transcriptomics to Therapy and Phenotype | gene2drug+phene

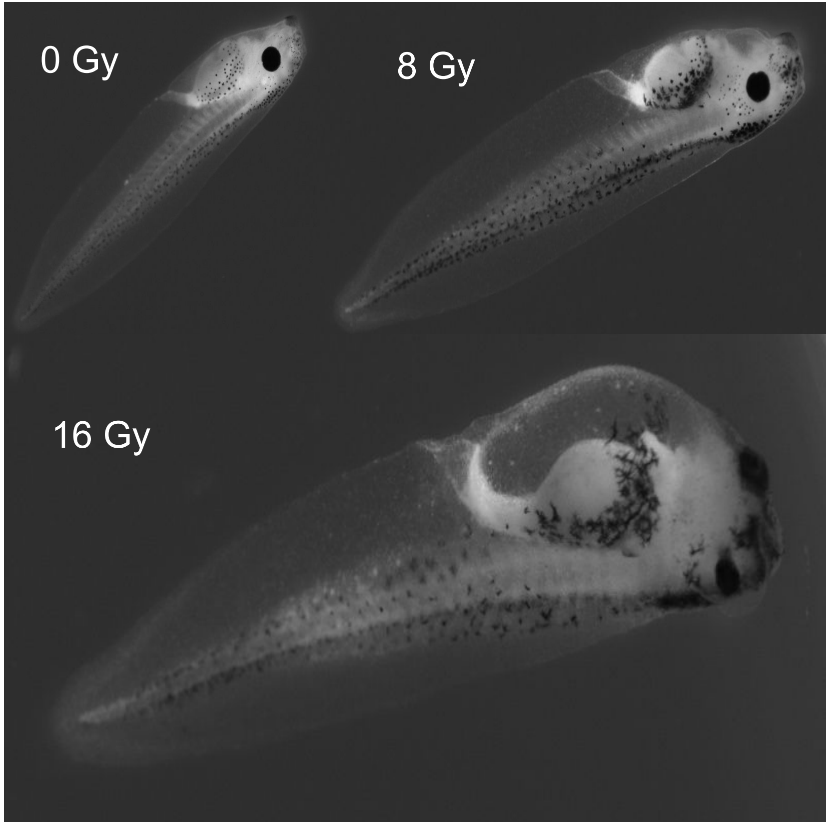



**Therapies to counter effects of Radiation**

- Pick a condition to antagonize: 8 Gy → 16 Gy
- Run differential gene expression analysis → Input fold-changes + p-values into probabilistic model *NeMoCAD* to obtain drug therapies
- Enrich for phenotypes using Gene Ontology

**Vishal Keshari**

# Uncertainty Meets Knowledge: *NeMoCAD*
# Transcriptomics to Therapy and Phenotype | gene2drug+phene




**Top enriched phenotypes: *gene2phene***

1. Cholesterol monooxygenase activity
2. Helicase activity (important for DNA repair), regulation of response to DNA damage and integrity
3. Muscle cell fate specification, visceral muscle development

**Top drugs selected: *gene2drug***

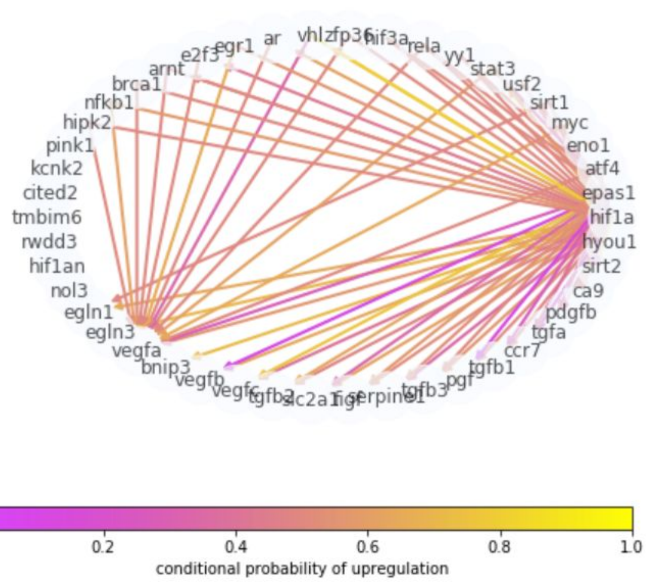
1. Dexrazoxane, a chemotherapy protective drug that is shown to reduce tissue damage.
2. Mercaptopurine, an immunosuppressive chemotherapy drug used to treat acute lymphatic leukemia
3. Atropine, an involuntary nervous system blocker
4. Ifosfamide, a chemotherapy drug used in treating multiple cancers
5. Pregnenolone, an endogenous steroid that is a precursor to most other steroid hormones

# Uncertainty Meets Knowledge: *NeMoCAD*

## drug+phene2gene

```
hypoxia_drugphene2gene = nemo.query_phenet_given_drugs('hypoxia', ['bisphenol a', 'urethane'], combo_func=16)
```

```
found 18 GO terms of interest for query hypoxia
found 16 genes of interest for query hypoxia
Formed common network with 46 nodes
Formed common network with 48 nodes
gene network
```



WYSS INSTITUTE

# Uncertainty Meets Knowledge: *NeMoCAD*

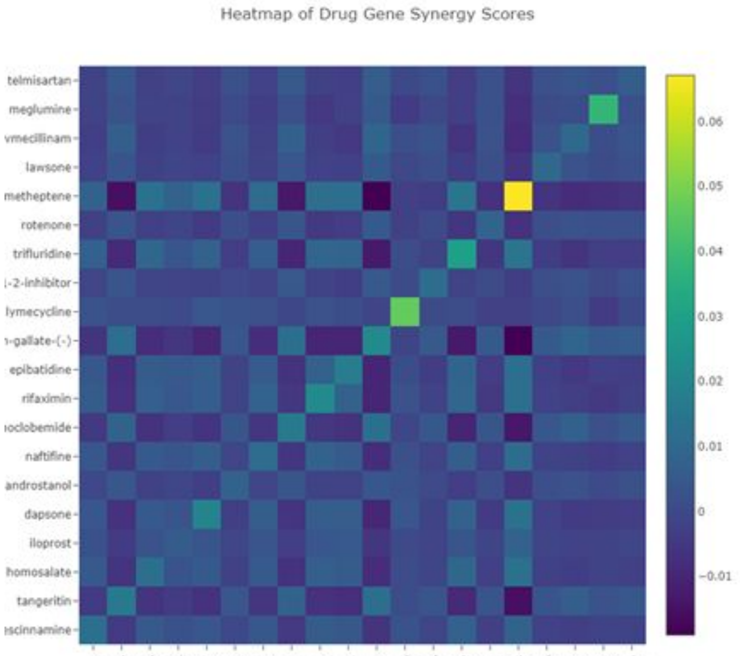
## phene2drug

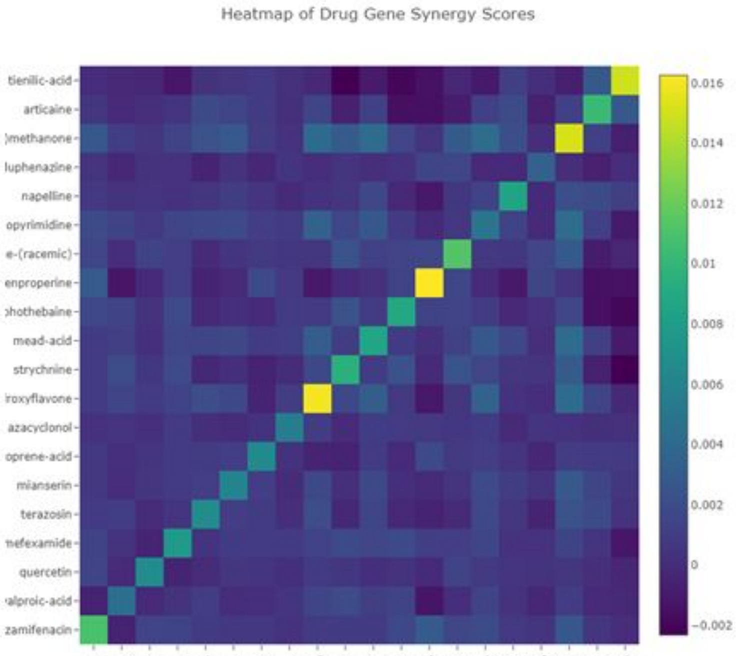
```
drugs_eye = nemo.query_drugs_given_phene('eye development', normalize='mean')
```

```
Found 15 relevant GO terms that map to query eye development
```

```
drugs_iron = nemo.query_drugs_given_phene('iron', normalize='mean')
```

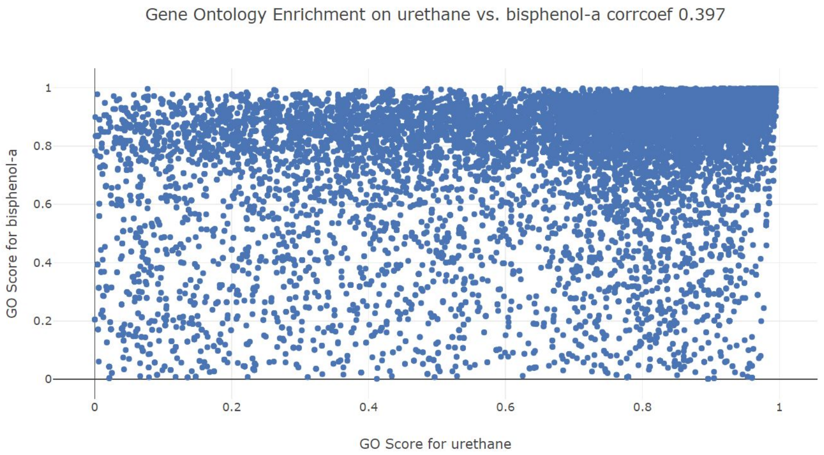
```
Found 62 relevant GO terms that map to query iron
```

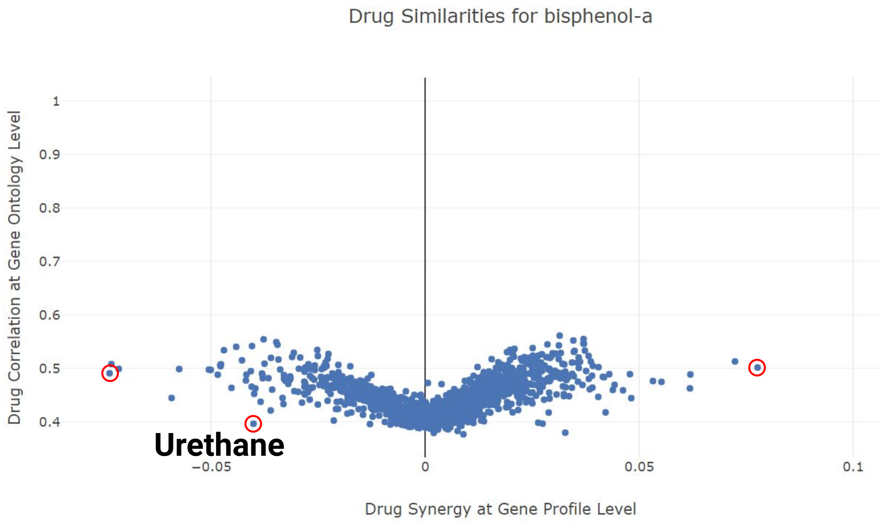


# Uncertainty Meets Knowledge: *NeMoCAD*

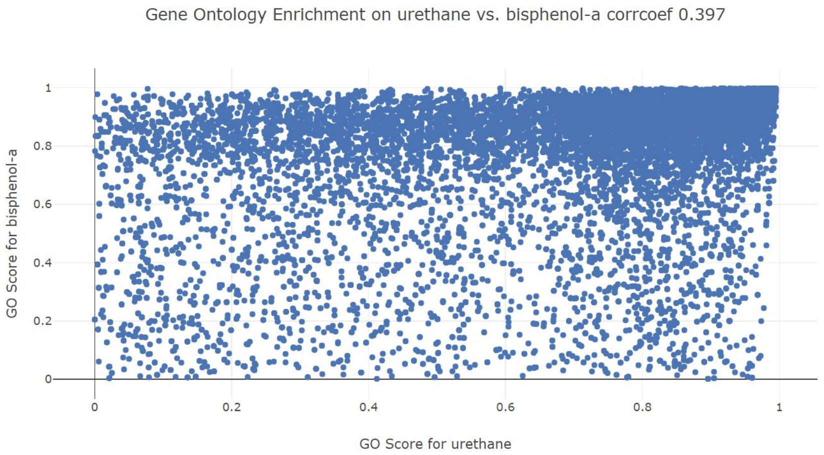
## drug2phene

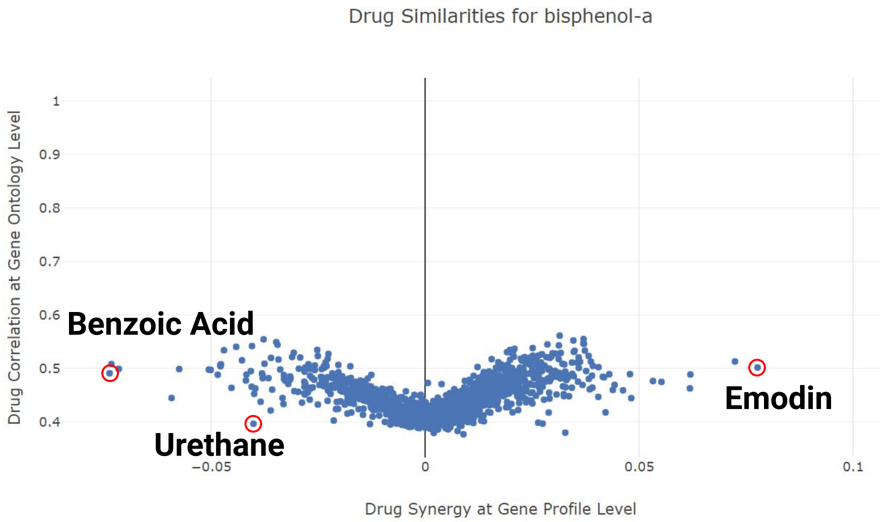


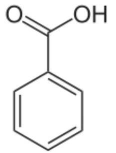



# Uncertainty Meets Knowledge: *NeMoCAD*

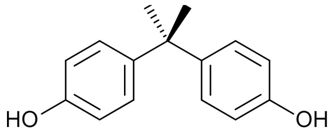
## drug2phene



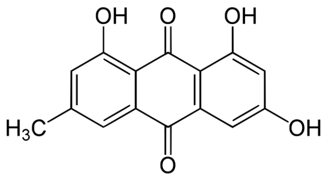




**Benzoic Acid**

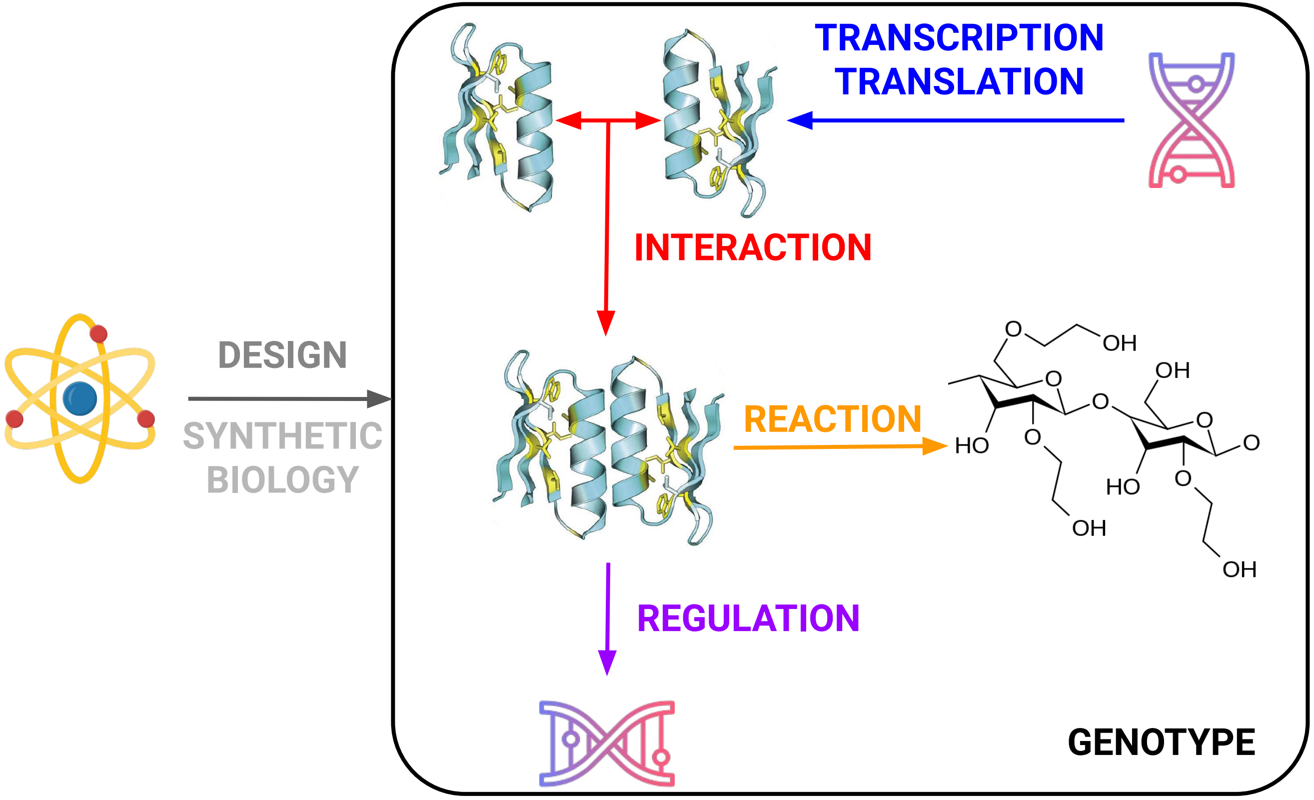



**Bisphenol A**




**Emodin**

# Problems of Design in Synthetic Biology



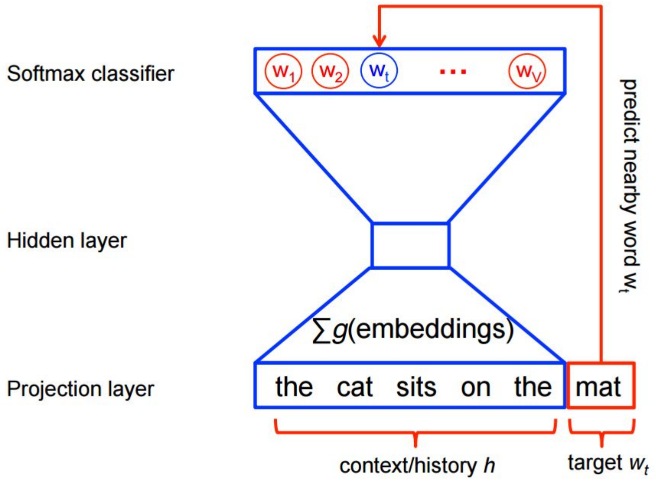
WYSS INSTITUTE

# Representing Biomolecules as Symbolic Sequences

- Proteins are simply Amino Acid sequences: `VPLLGLY…`
- Genes/mRNAs are simply Nucleotide sequences: `AATCGGTA…`

# Representing Biomolecules as Symbolic Sequences

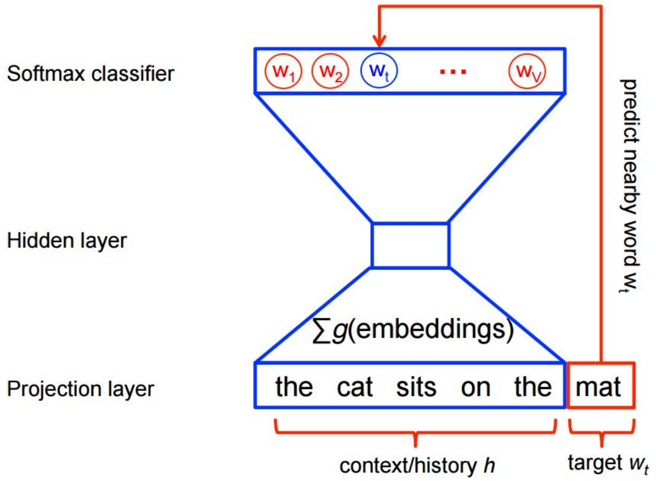
- Proteins are simply Amino Acid sequences: `VPLLGLY…`
- Genes/mRNAs are simply Nucleotide sequences: `AATCGGTA…`
- Using <u>language</u> models for "representing" arbitrary biomolecules as mathematical entities
    - AA : Sequences = Words : Sentences

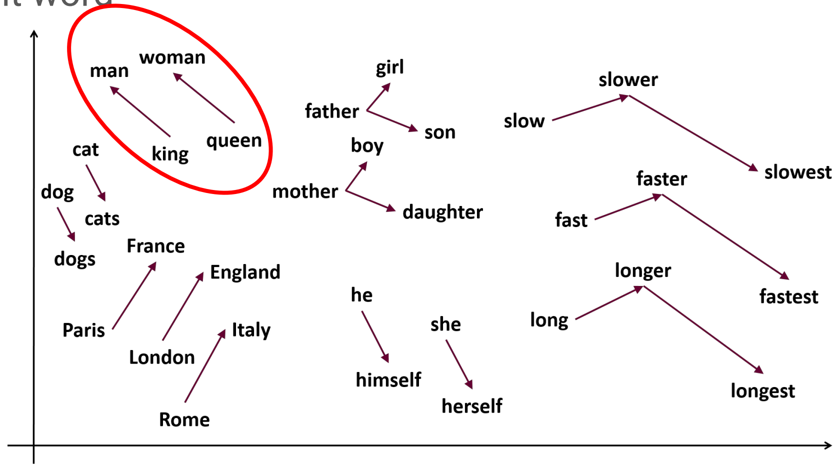



**word2vec, Mikolov et al. (2013)**

# Representing Biomolecules as Symbolic Sequences

- Proteins are simply Amino Acid sequences: `VPLLGLY…`
- Genes/mRNAs are simply Nucleotide sequences: `AATCGGTA…`
- Using language models for “representing” arbitrary biomolecules as mathematical entities
    - AA : Sequences = Words : Sentences
- One simplifying assumption: local neighborhood decides global high-level properties
- Unsupervised: predict context words from current word



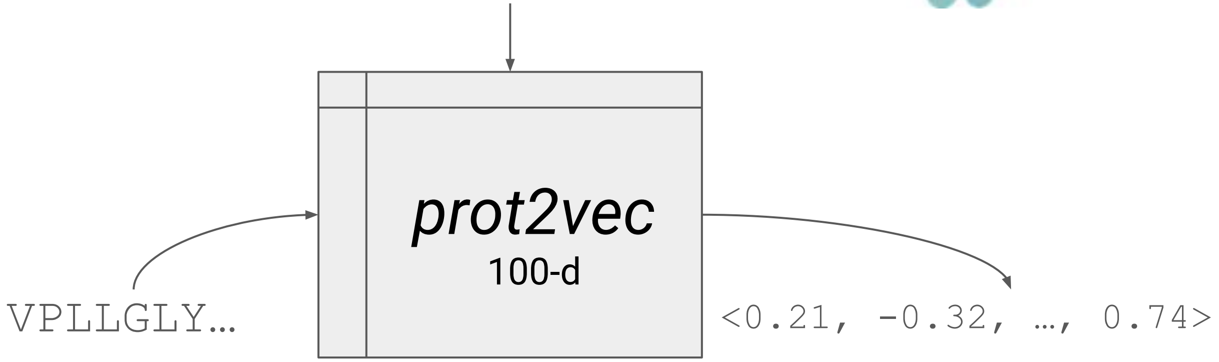




**word2vec, Mikolov et al. (2013)**

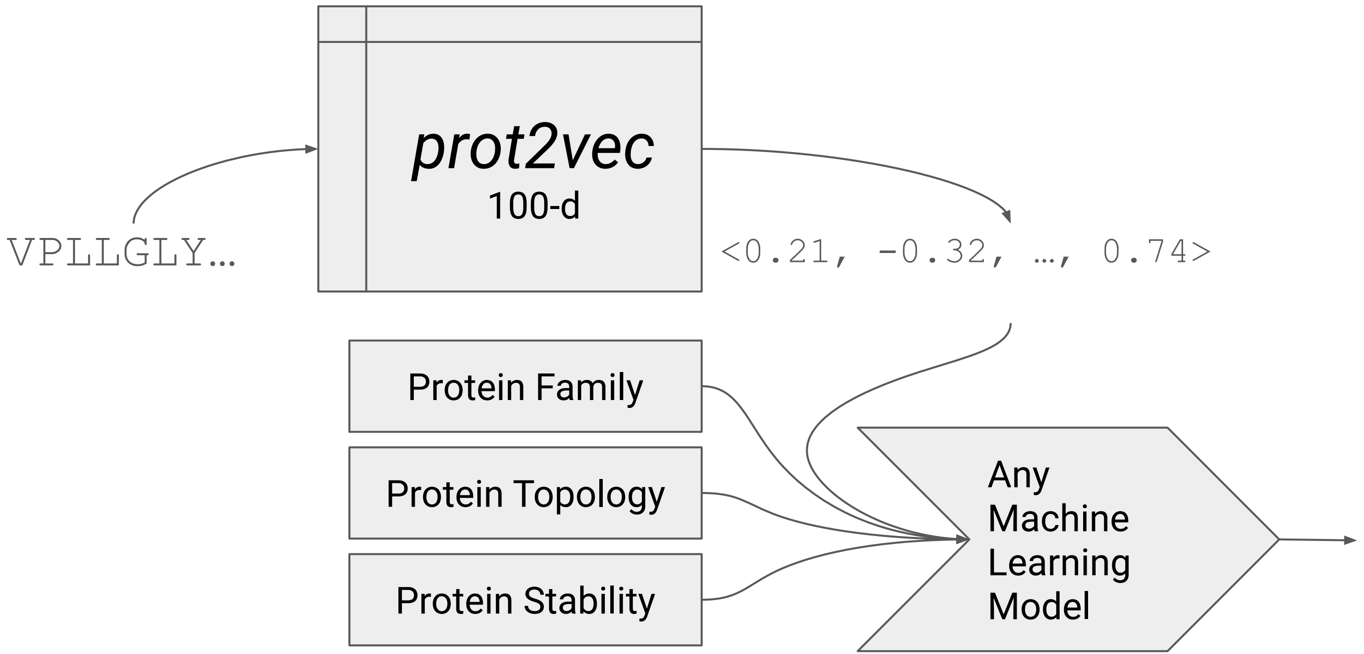
Wyss Institute

# Representing Biomolecules as Symbolic Sequences
## *prot2vec*: a Learnt Space of Proteins

93,588 AA sequences from *Homo sapien* Proteome

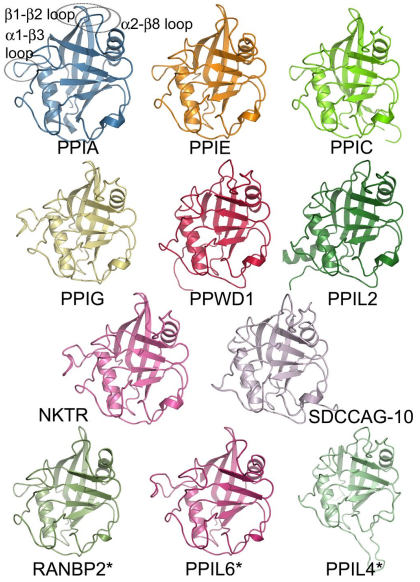


# Representing Biomolecules as Symbolic Sequences
## *prot2vec*: a Learnt Space of Proteins



WYSS INSTITUTE

# Representing Biomolecules as Symbolic Sequences
## *prot2vec* predicts Protein Families




*prot2vec*
100-d

<-0.12, 0.78, …, 0.56>

324,017 AA sequences from SwissProt across 7027 protein families

Protein Family

Protein Topology

Protein Stability

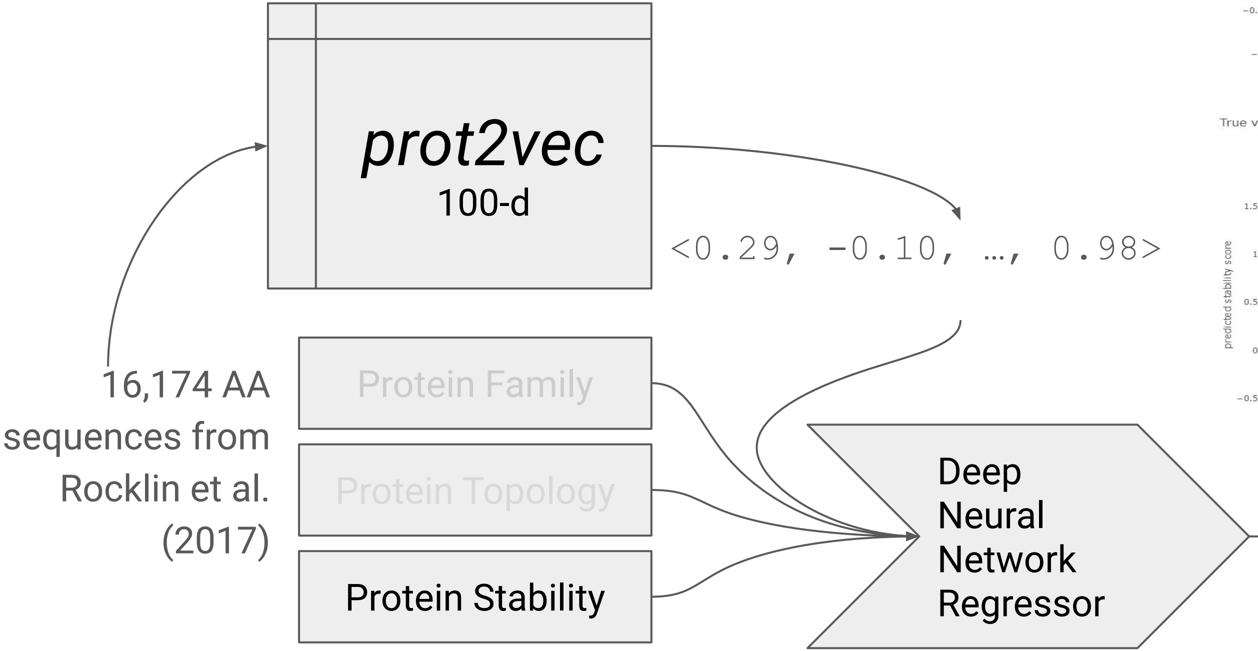
Nearest Neighbor Classifier

→ 73.2% accuracy

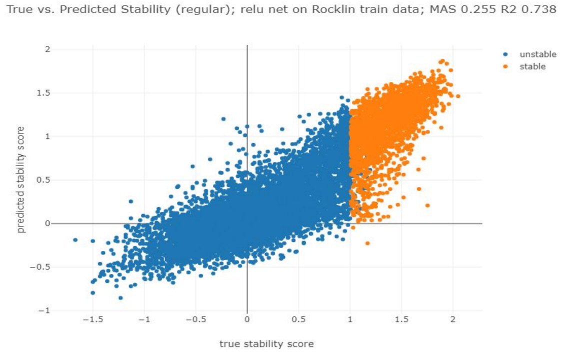
**Insight:** proteins close in *sequence* space → close in *prot2vec* space → close in *function* space

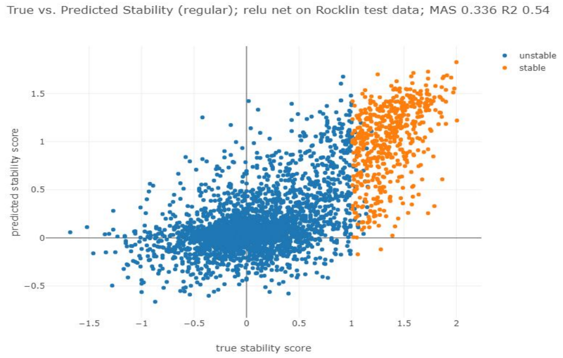
WYSS INSTITUTE

# Representing Biomolecules as Symbolic Sequences
## *prot2vec* estimates Protein Stability

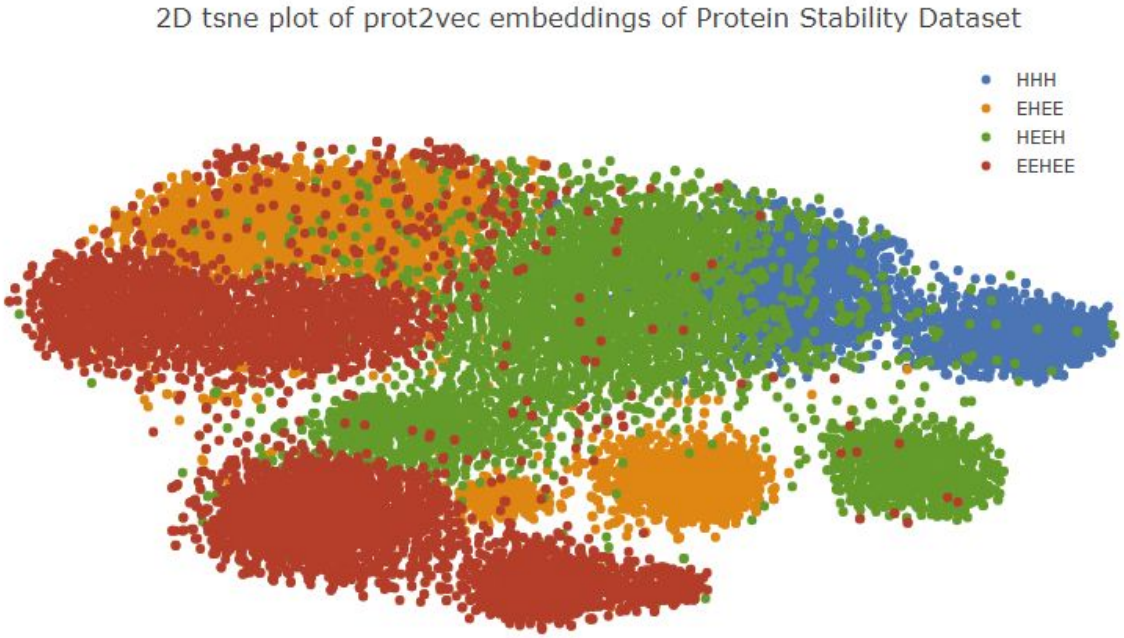



0.54 $R^2$ value



# Representing Biomolecules as Symbolic Sequences
## *prot2vec* captures Protein Topologies



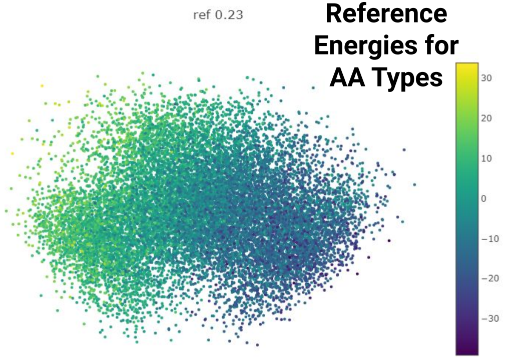

**Insight:** proteins close in *sequence* space → close in *prot2vec* space → close in *topology* space

**Experimental Data from Rocklin et al. (2017)**

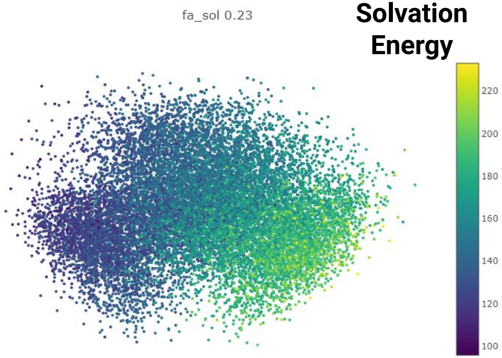
# Representing Biomolecules as Symbolic Sequences
# *prot2vec* correlates to Biophysical Parameters

**Question:** What do these 100 dimensions mean? Are they arbitrary?

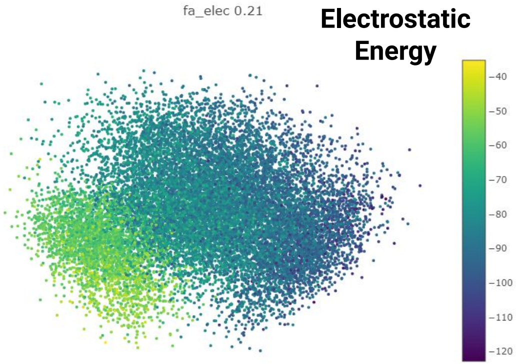
**Reference Energies for AA Types**



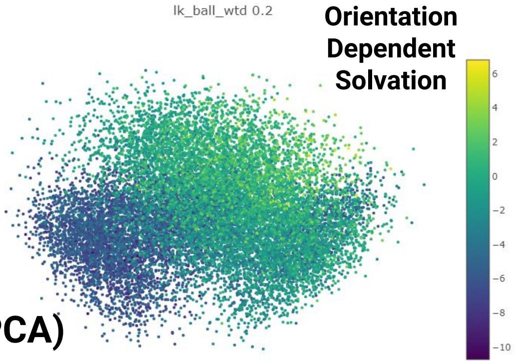

**Solvation Energy**



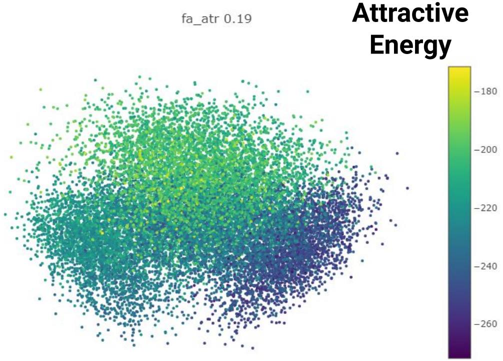

**Electrostatic Energy**



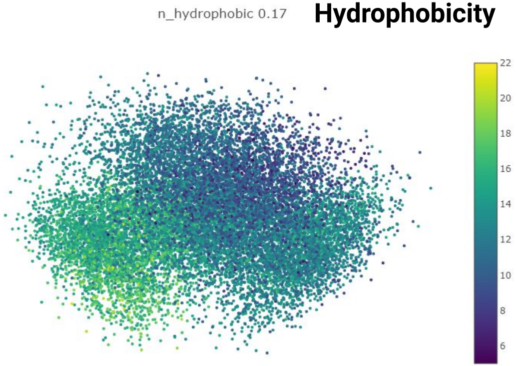

**Orientation Dependent Solvation**




**Attractive Energy**



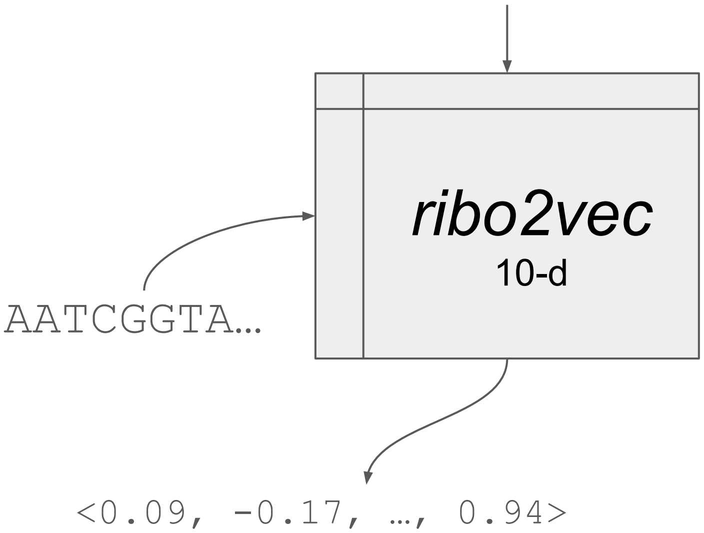

**Hydrophobicity**




**(PCA)**

# Representing Biomolecules as Symbolic Sequences
# *ribo2vec*: a Learnt Space of Riboswitches

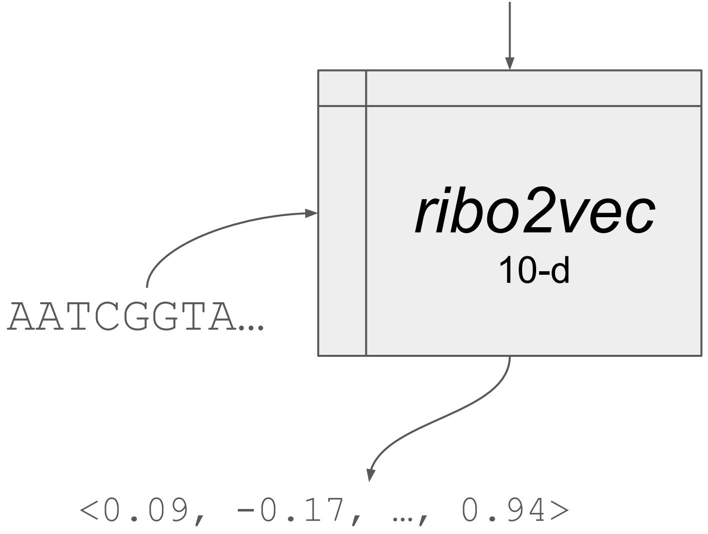
49,159 mRNA sequences of natural Riboswitches

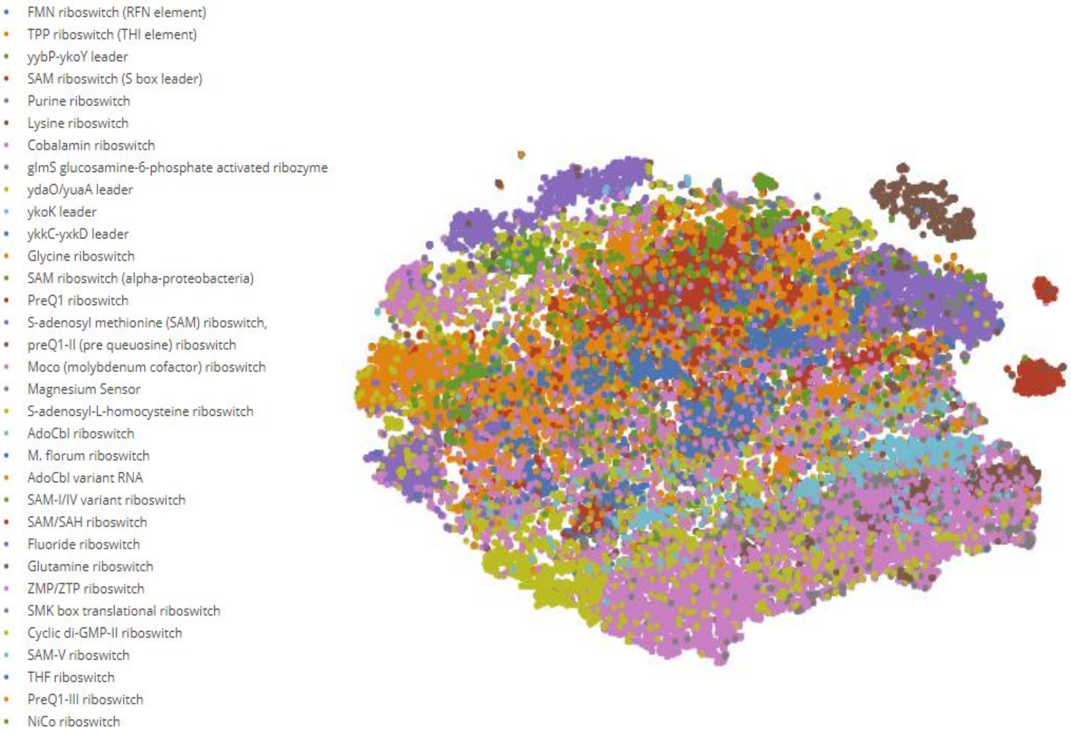


# Representing Biomolecules as Symbolic Sequences
# *ribo2vec*: a Learnt Space of Riboswitches

Rfam

49,159 mRNA sequences of natural Riboswitches

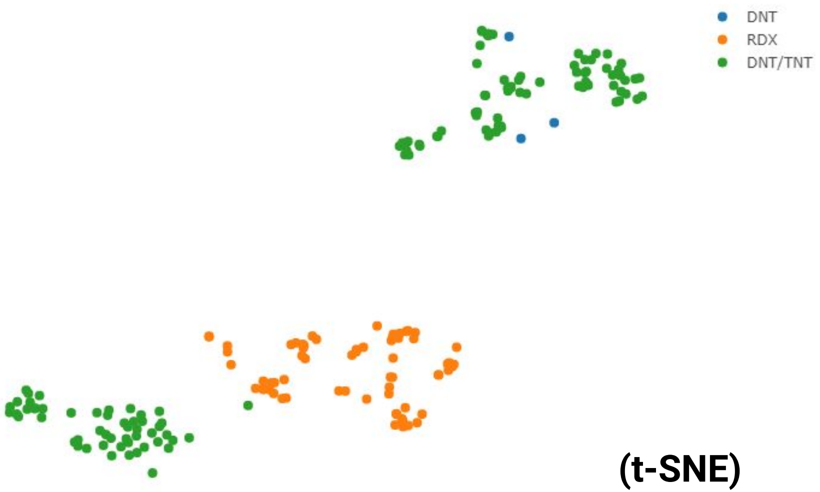


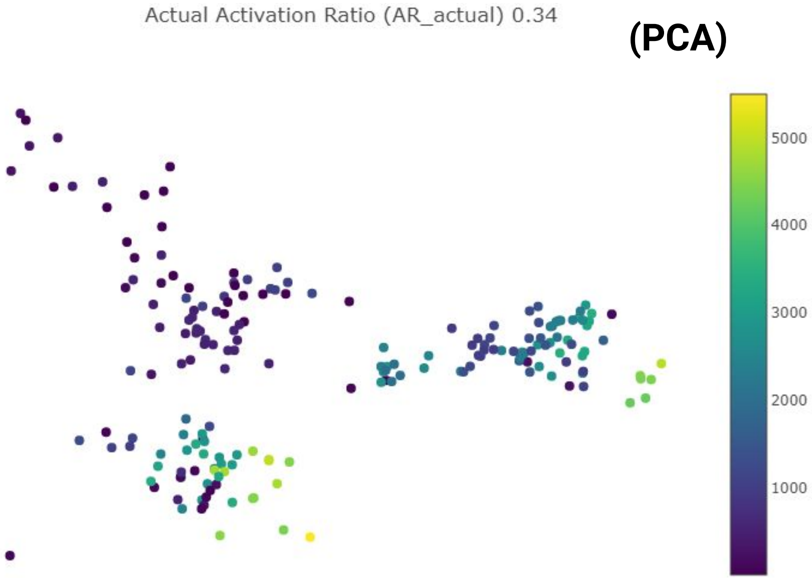
# Representing Biomolecules as Symbolic Sequences
# *ribo2vec* correlates to Biophysical Parameters

**Question:** What do these 10 dimensions mean? Are they arbitrary?

Visualize 192 mRNA design sequences to detect DNT/TNT




**(t-SNE)**



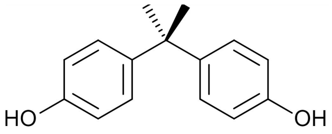

**(PCA)**

**Insight:** some "arbitrary" dimensions of *ribo2vec* correlate with experimental activation ratio

**Experimental Data from Howard Salis**

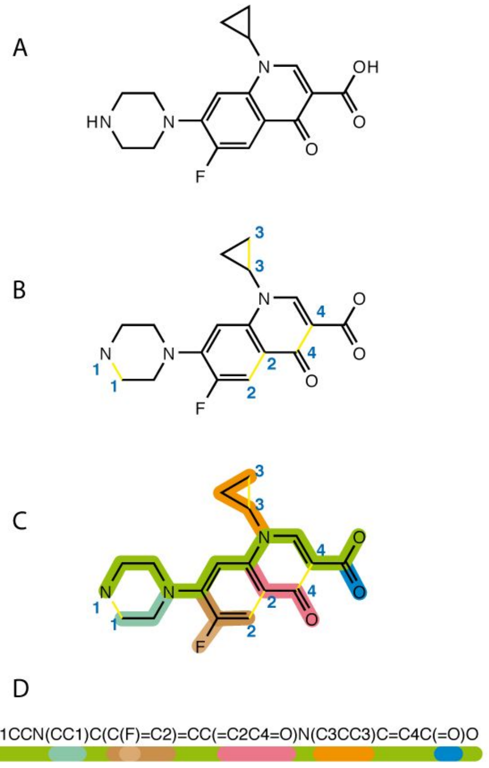
# Representing Molecules as Symbolic Sequences

- **Problem:** arbitrary metabolites, small molecules, drugs, ligands and carbohydrates are NOT simple linear sequences
- **Solution:** Simplified Molecular-Input Line-Entry System (SMILES) representation




**Bisphenol A**

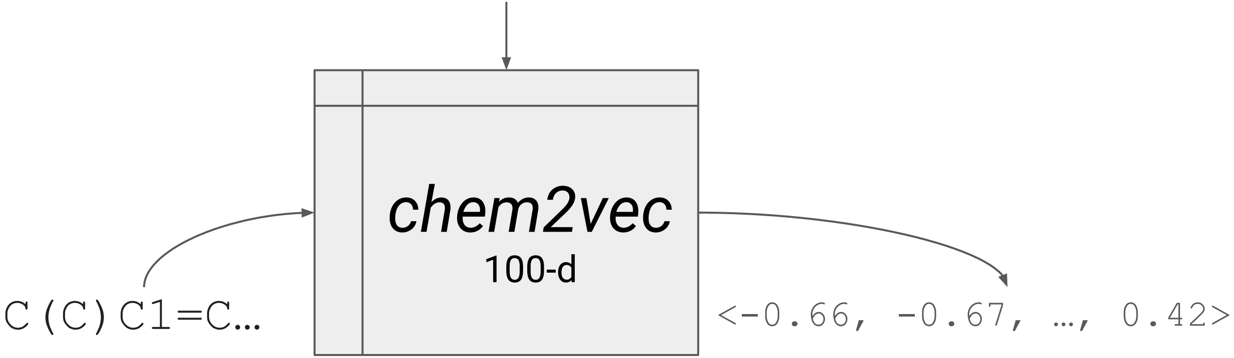
`CC(C)(C1=CC=C(C=C1)O)C2=CC=C(C=C2)O`




**Ciprofloxacin**

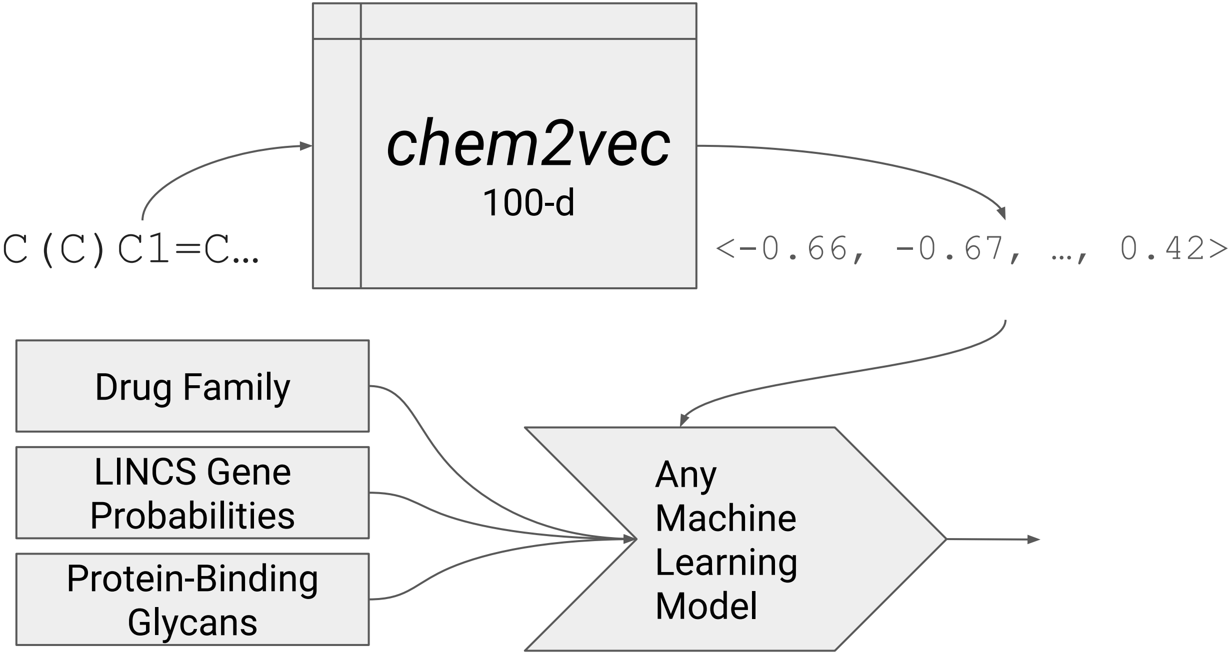
# Representing Molecules as Symbolic Sequences
## *chem2vec*: a Learnt Space of Chemicals

First 1 million chemicals from PubChem



WYSS INSTITUTE

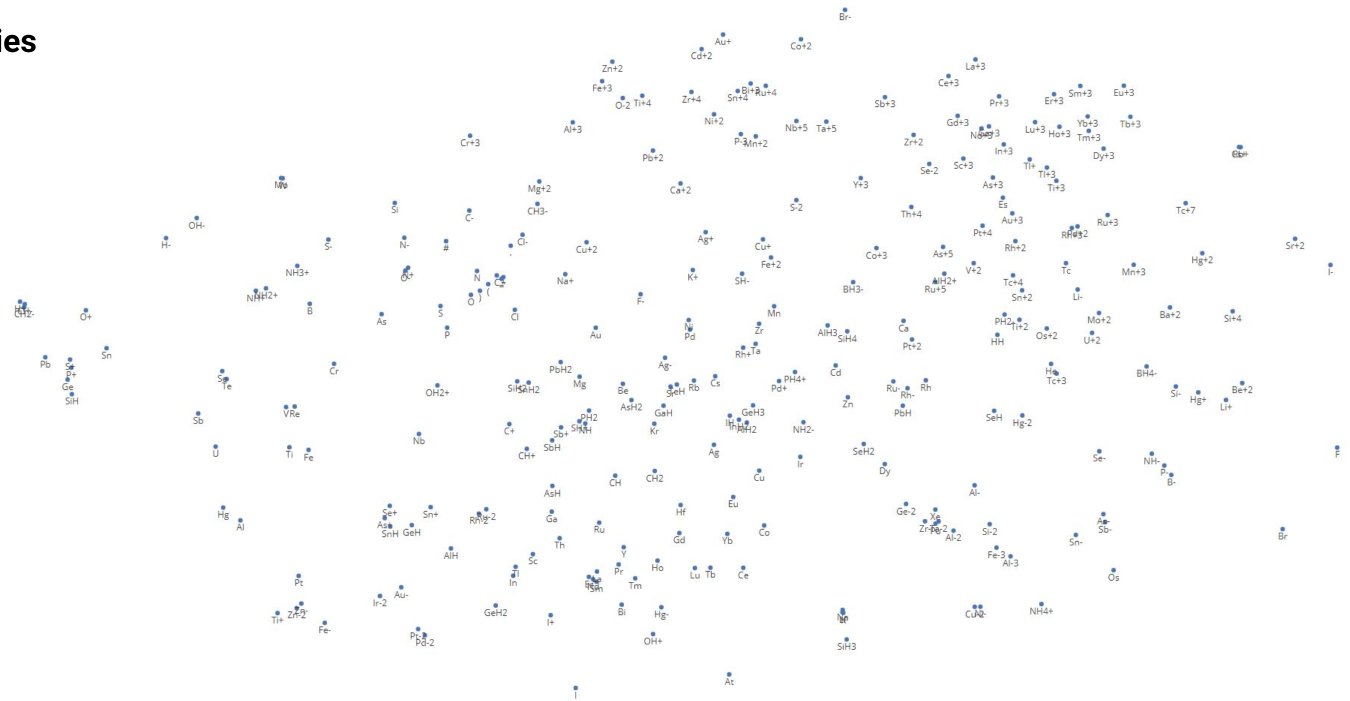
# Representing Molecules as Symbolic Sequences
# *chem2vec*: a Learnt Space of Chemicals



# Representing Molecules as Symbolic Sequences
# *chem2vec* encodes Chemical Valencies

**Visualizing just “atomized” chemical species**

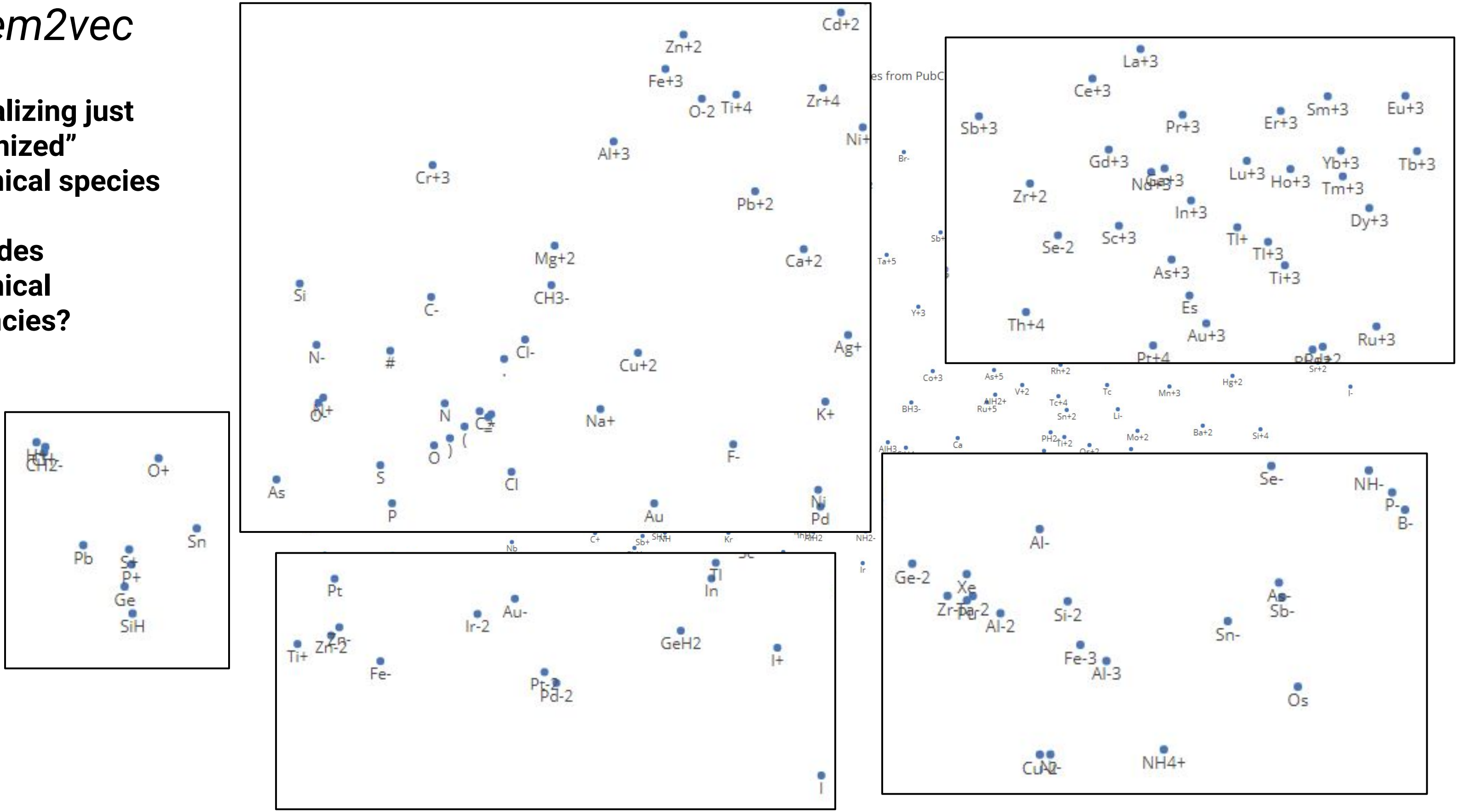
t-SNE Visualization of chem2vec Embedding of Chemical Species from PubChem



# Representing Molecules as Symbolic Sequences
*chem2vec*

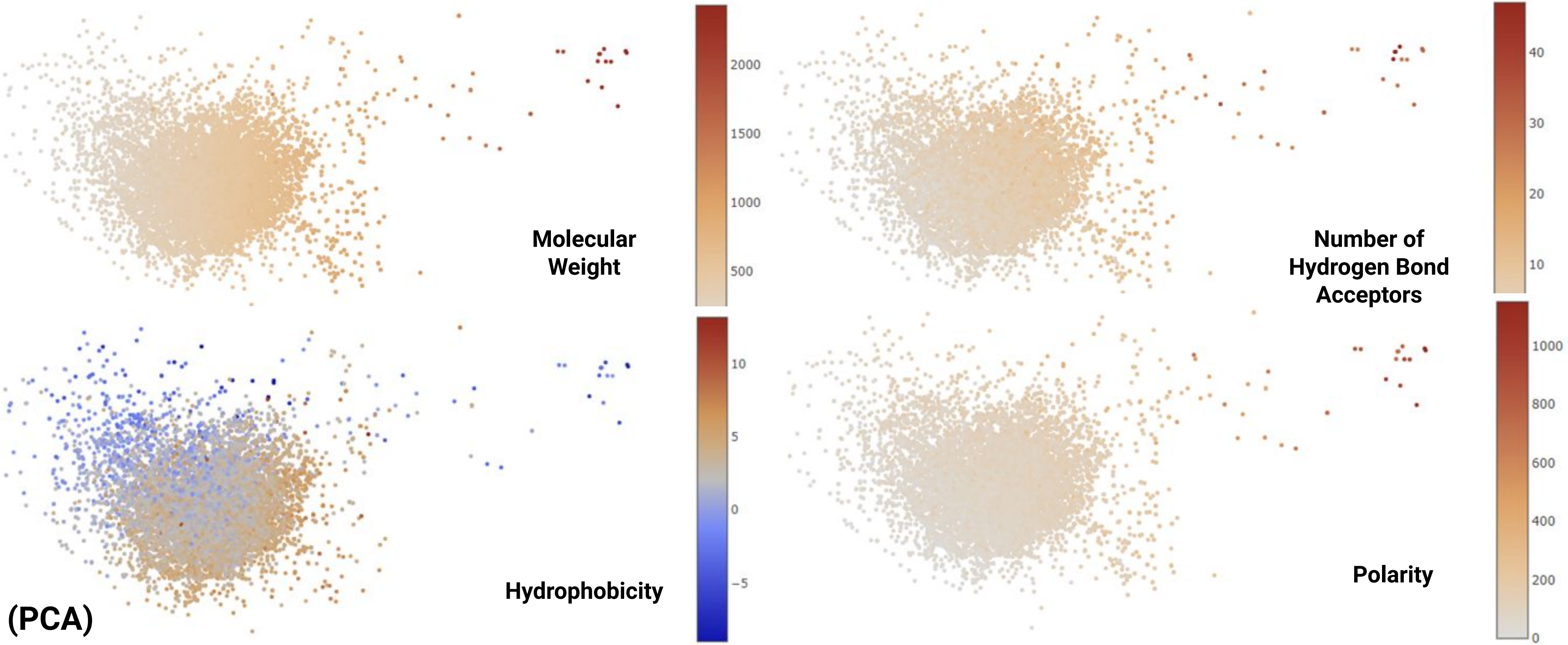
**Visualizing just “atomized” chemical species**

**Encodes chemical valencies?**



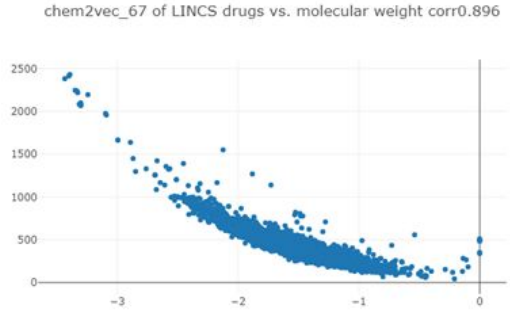
# Representing Molecules as Symbolic Sequences
# *chem2vec* correlates to Molecular Properties

**Question:** What do these 100 dimensions mean? Are they arbitrary?

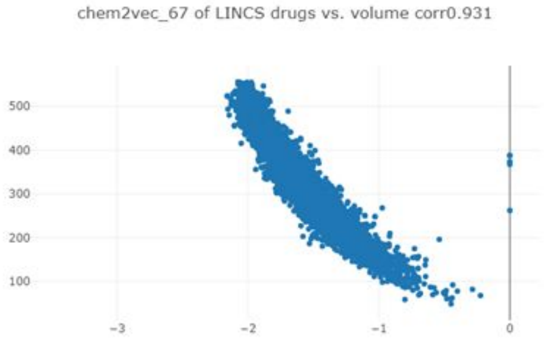


# Representing Molecules as Symbolic Sequences
# *chem2vec* correlates to Molecular Properties

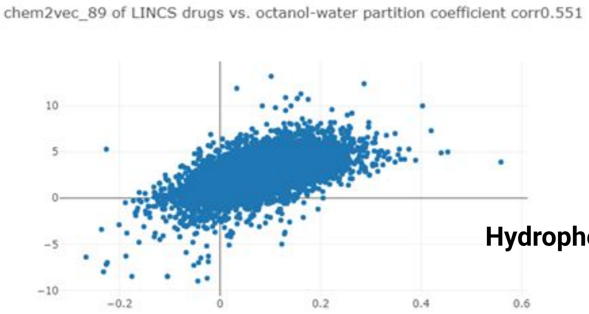
Linear Correlation between molecular properties and *chem2vec* dimensions



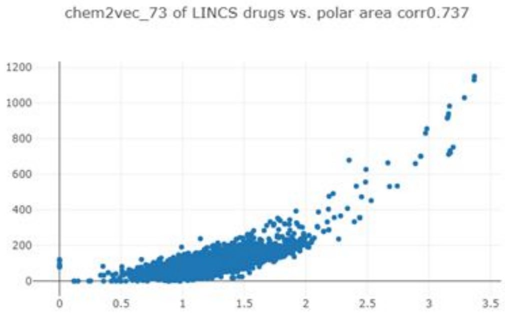

**Molecular Weight**




**Volume**



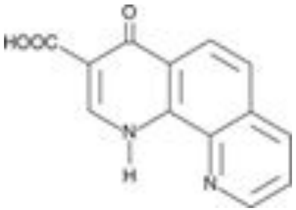

**Hydrophobicity**



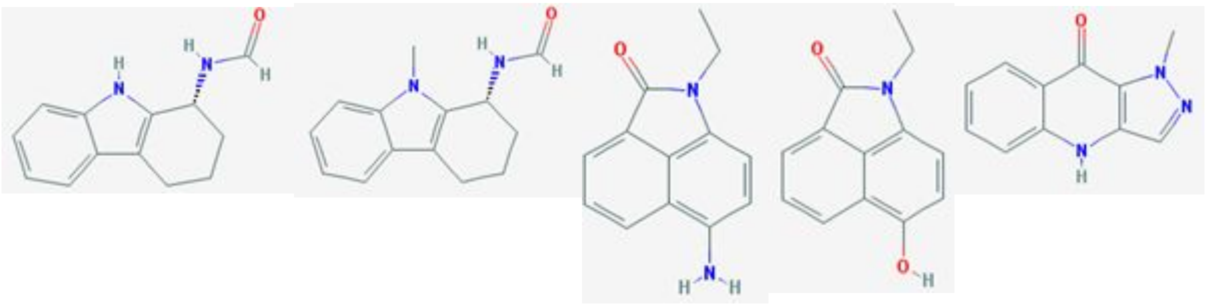

**Polarity**

# Representing Molecules as Symbolic Sequences
# *chem2vec* serves as a drug query engine

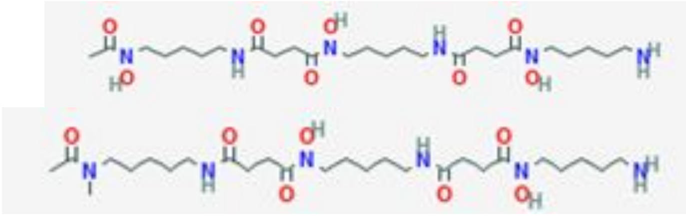
Say we have drug candidates we find improving tolerance , and we wish to explore drugs “similar” to them in the functional space, but potentially better in PK/PD and toxicity

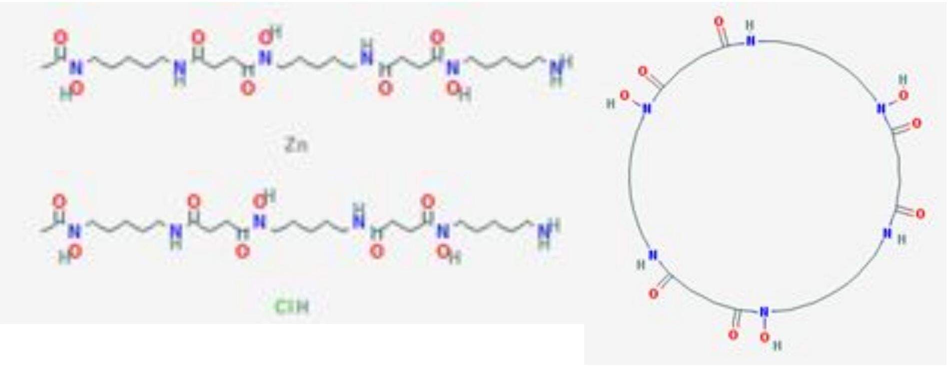


**“Miracle” Drug for Xenopus**

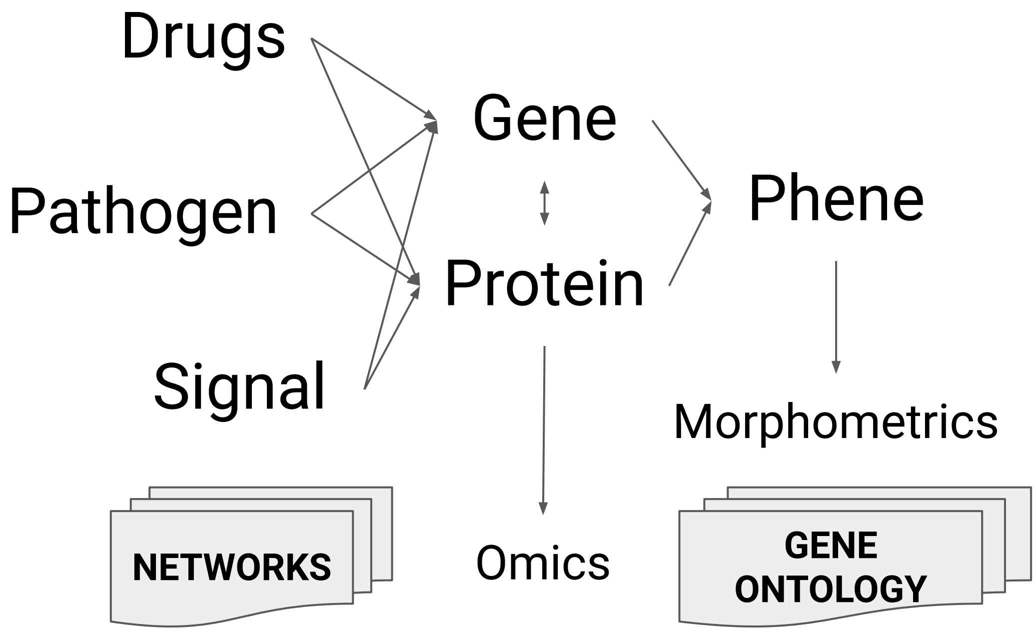


**DFOA**

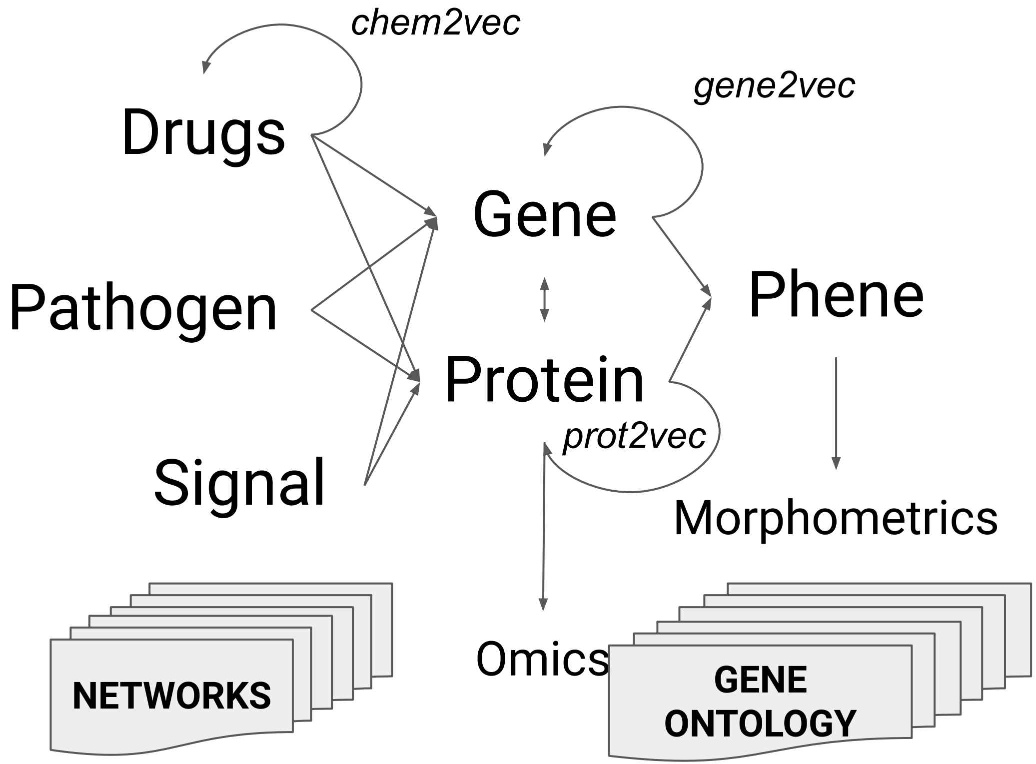




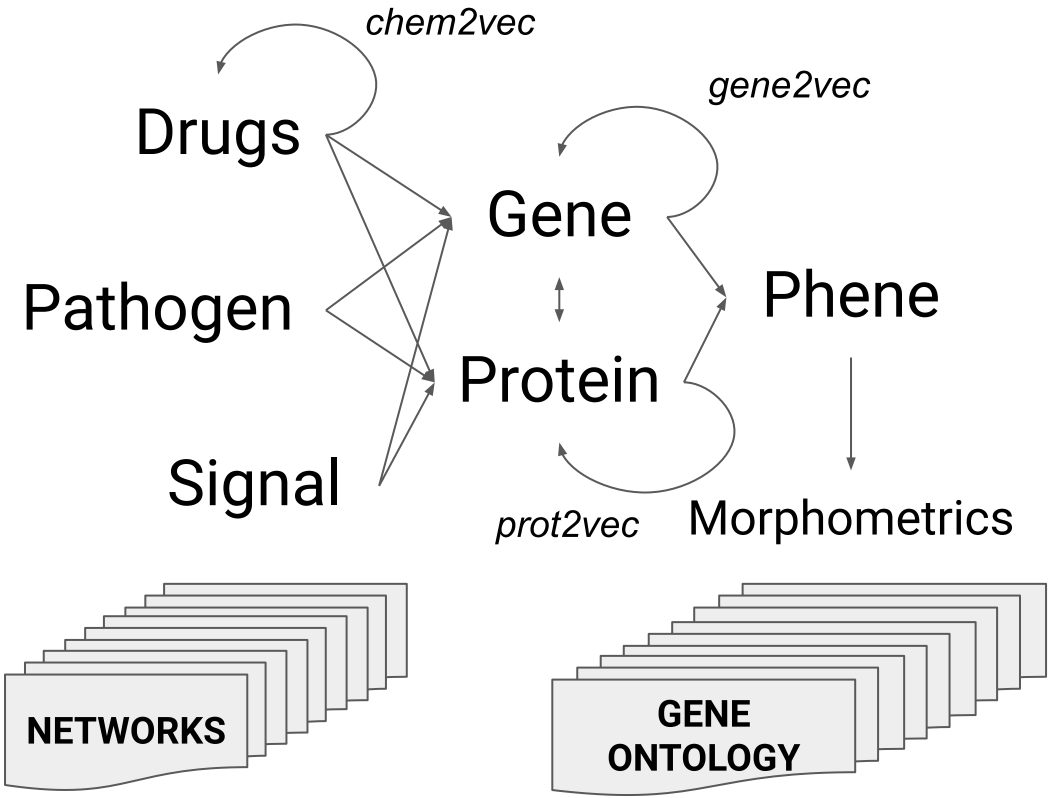
# Putting It Back Together



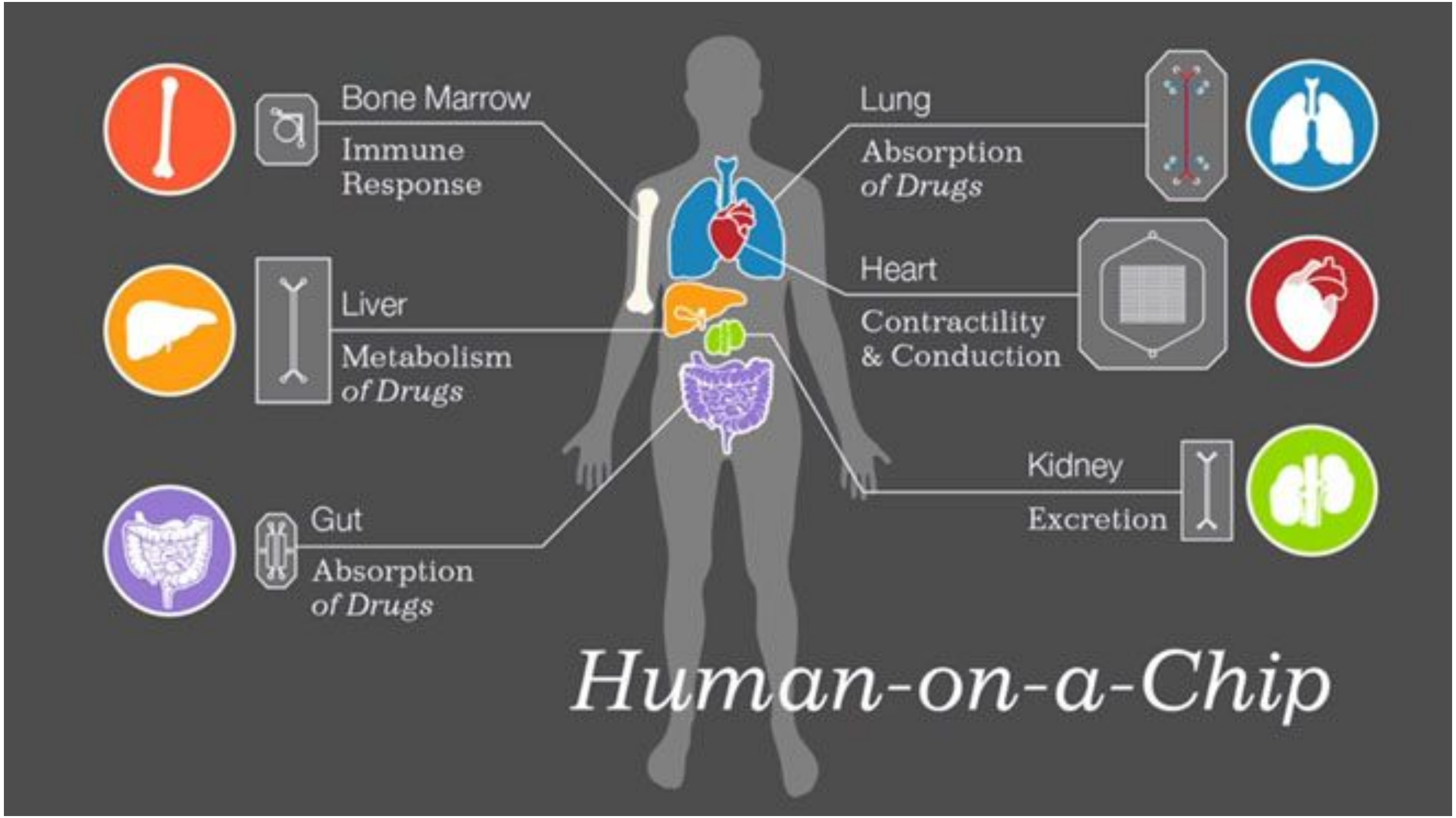
# Putting It Back Together



# Putting It Back Together



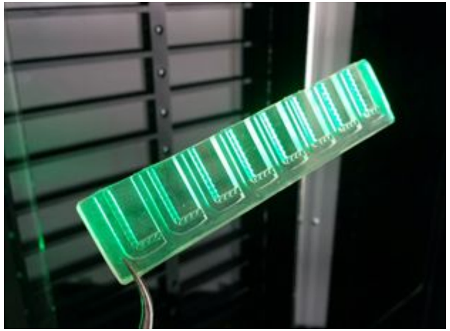
# Wyss Institute's Organs-on-Chips

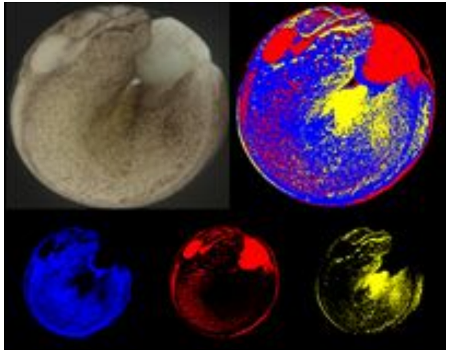


Wyss Institute

# *Xenopticon*: Organism-on-Chip

## High-Throughput Xenopus embryo analyzer





Capability to image >700 embryos every 15 min in 16 colors

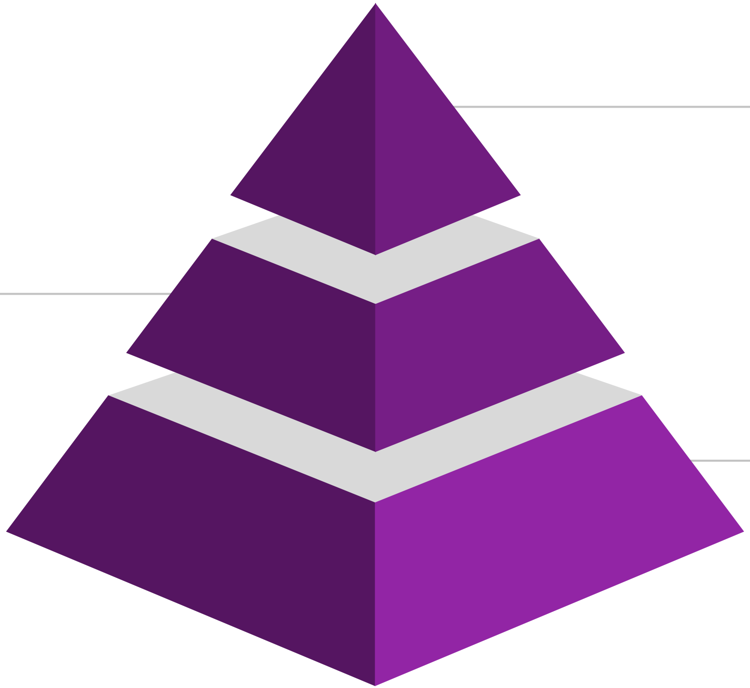
Results in 100s – 1,000s of metrics per embryo per time-point

**Youngjae Cho, Bret Nestor, Richard Novak**

# *Xenopticon*: A Cheaper Route to Drug "Design"

1 **PHENOTYPE DATA (XENOPTICON)**
$1-10 per image, 1 second per image

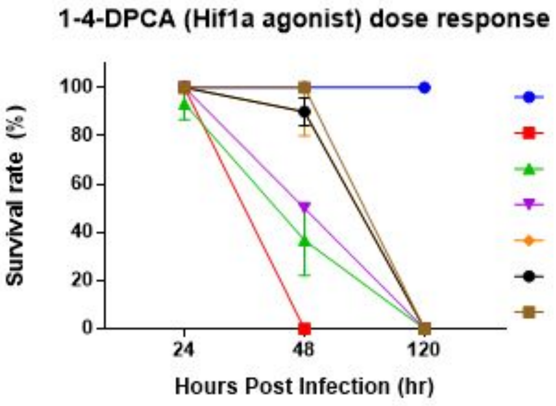
2 **GENOTYPE DATA (TRANSCRIPTOMICS)**
$1-10K per sample, 1 month per sample



3 **DRUG DEVELOPMENT (R&D)**
$1-10B per drug, 10 years per drug

# *Xenopticon*: Acquire Phenotype Metrics

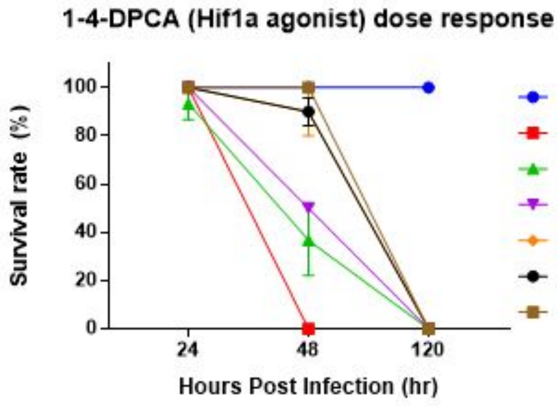
## Estimate Embryo Viability

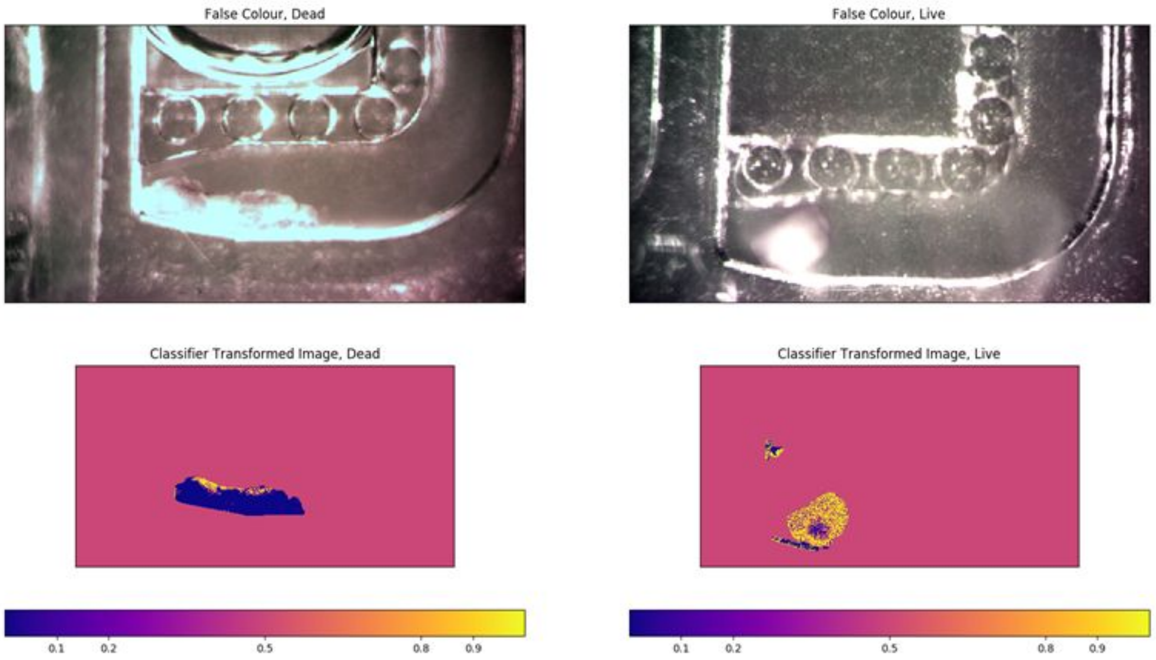



**Vishal Keshari, Alex Dinis**

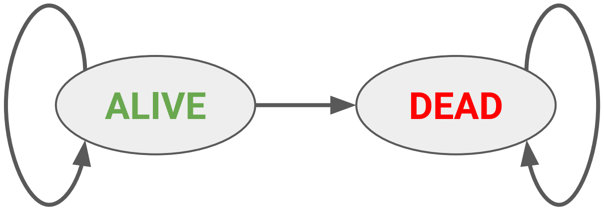
# *Xenopticon*: Acquire Phenotype Metrics

## Estimate Embryo Viability




**Classifier on Embryo's Spectrum**
**→ Hidden Markov Model**

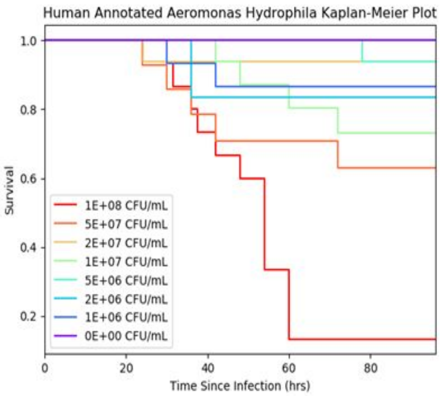


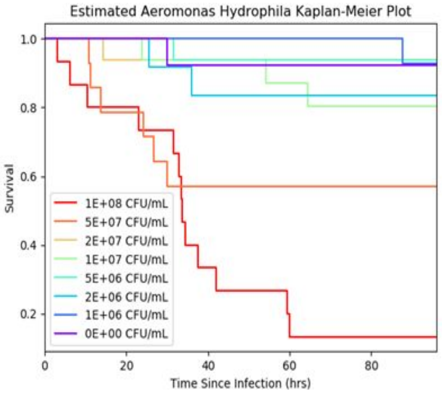

**93% accuracy on held-out test image sequences**

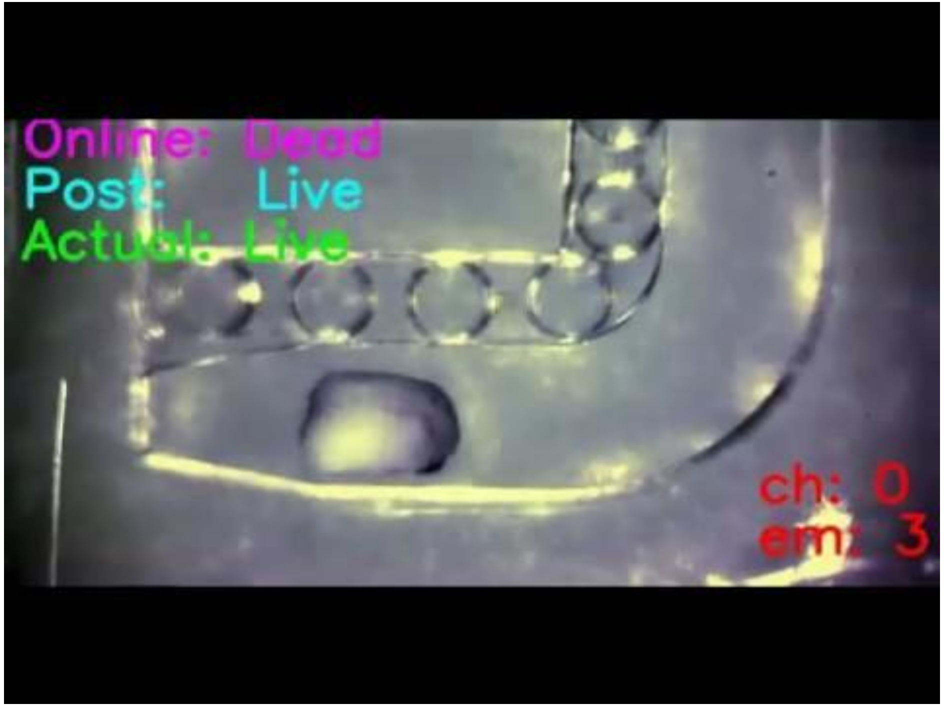
**Bret Nestor**

# *Xenopticon*: Acquire Phenotype Metrics

## Obtain Survival Curves



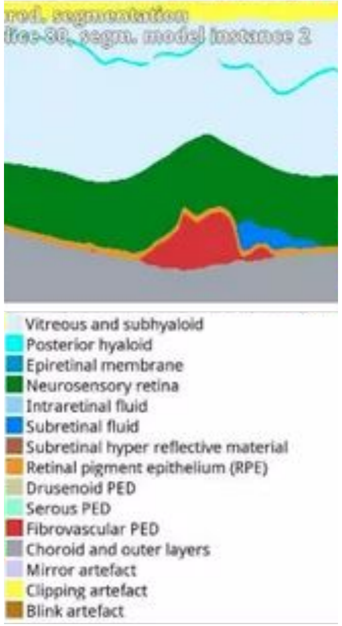



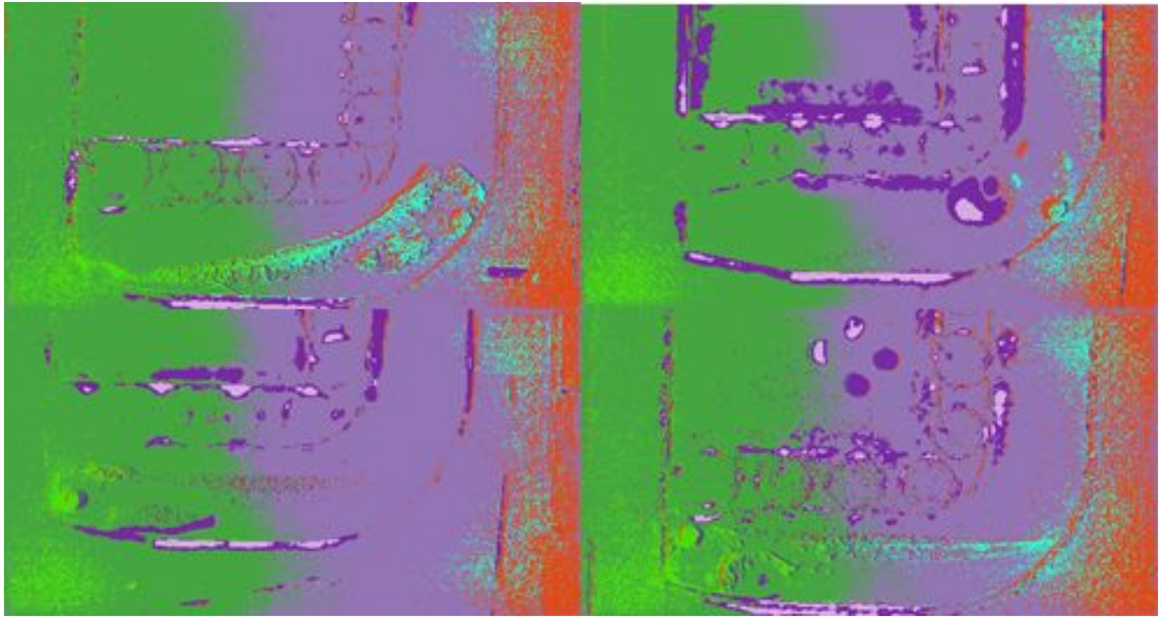




**Bret Nestor**

# *Xenopticon*: Acquire Phenotype Metrics

## Track Tissue Development



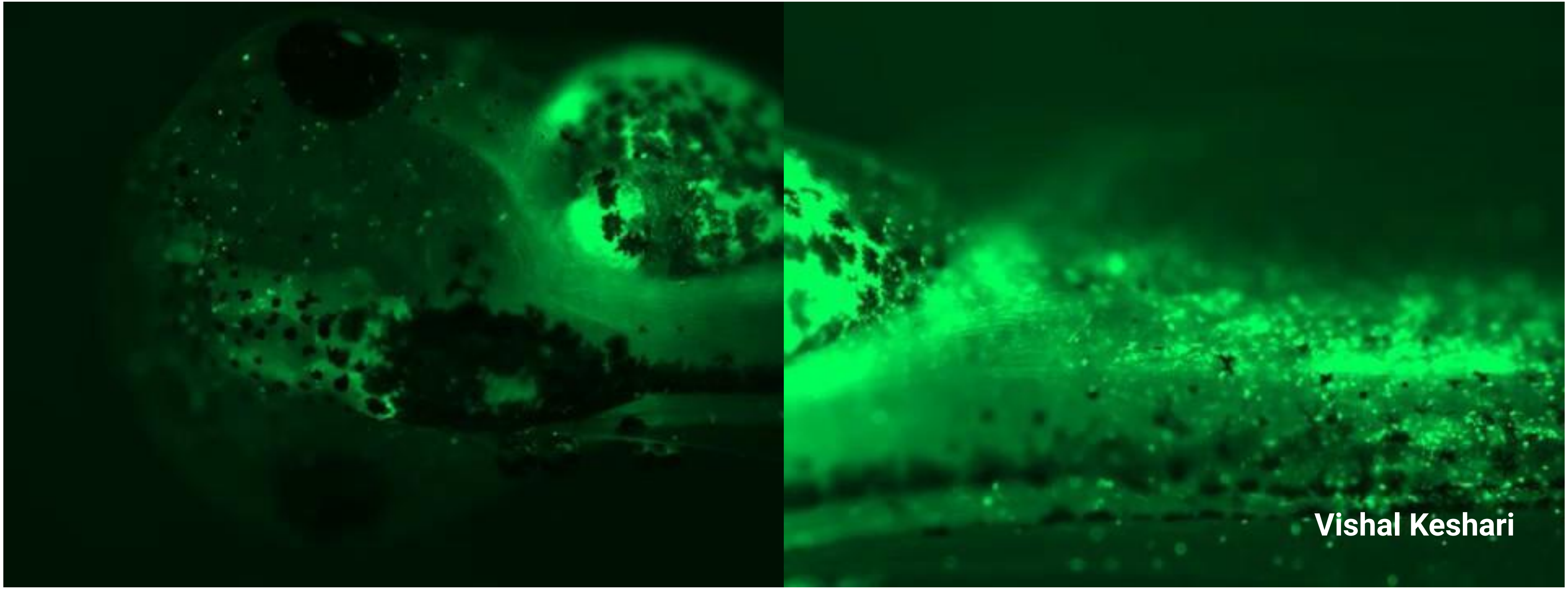
**DeepMind**

**Bret Nestor**

# *Xenopticon*: Acquire Phenotype Metrics

## Making the Invisible, Visible | Visual Tracking of Immune Response

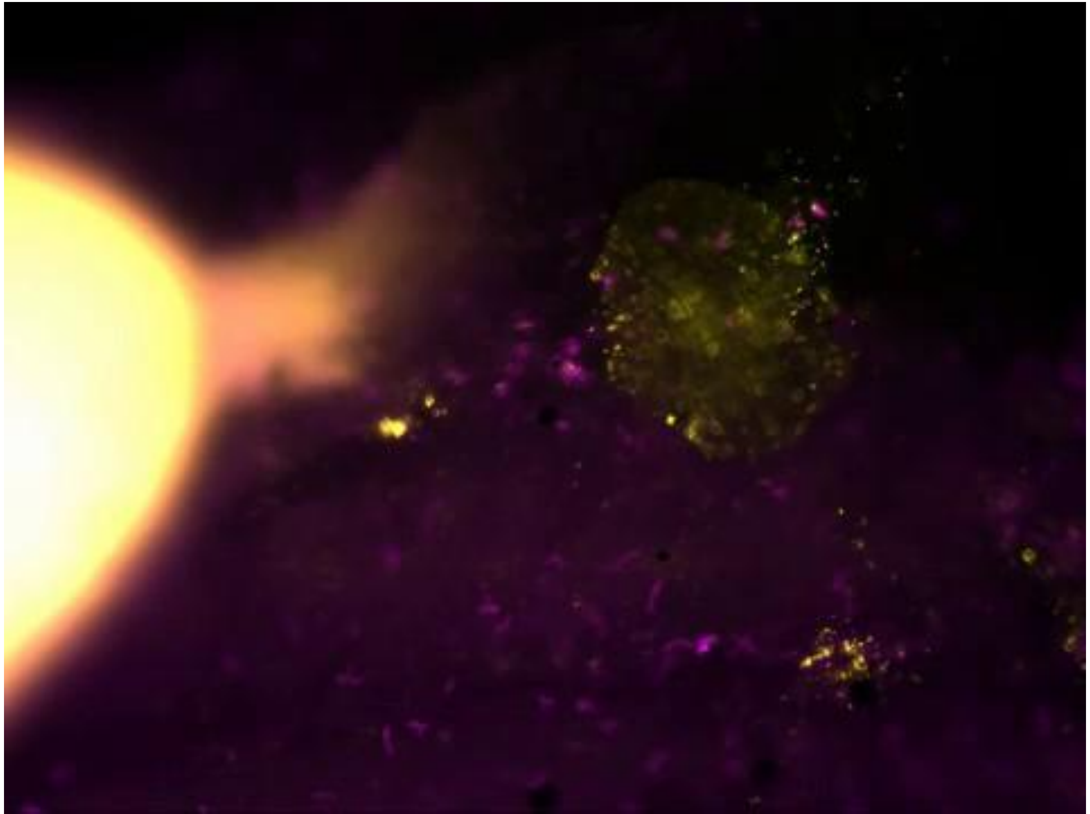
GFP expressing Macrophages to track spatiotemporal dynamics of “immune response”



# *Xenopticon*: Acquire Phenotype Metrics

## Making the Invisible, Visible | Visual Tracking of Pathogen Infection

mCherry expressing E. coli bacteria to track spatiotemporal dynamics of "pathogen infection"



**Vishal Keshari, Alex Dinis**

WYSS INSTITUTE

# *Xenopticon + NeMoCAD = XenoDoc*

Drugs

Gene

Pathogen

Phene

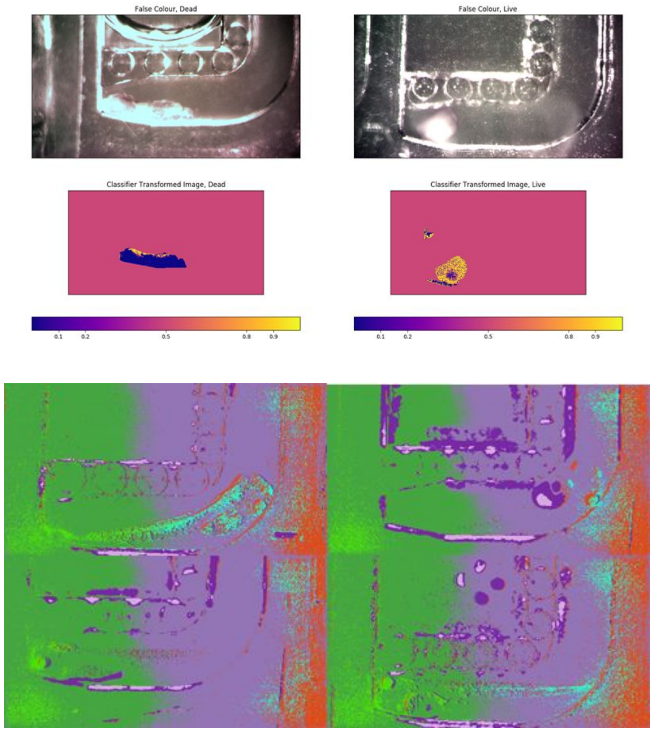
Protein

Signal

Morphometrics

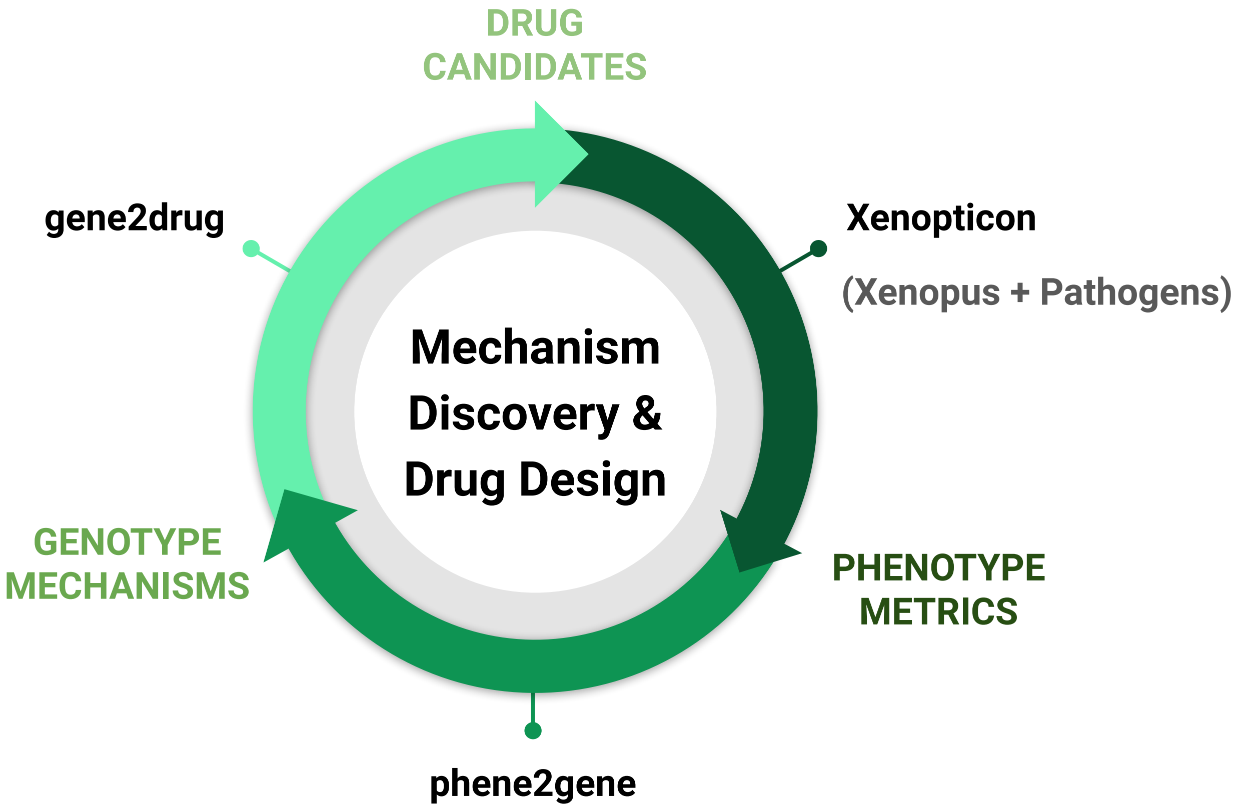
NETWORKS

GENE ONTOLOGY




Bret Nestor

# Iterative Discovery and Design



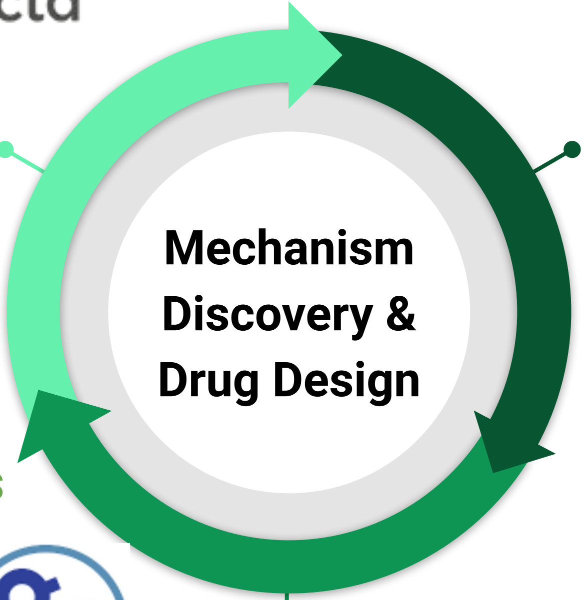
# Iterative Discovery and Design

DRUG CANDIDATES

gene2drug

Xenopticon

(Xenopus + Pathogens)




GENOTYPE MECHANISMS

PHENOTYPE METRICS

phene2gene

# Acknowledgements

Don Ingber, Jim Collins

Mike Super, Richard Novak

Bret Nestor, Vishal Keshari, Alex Dinis, Susan Clauson, Youngjae Cho

Shannon Duffy, Mark Cartwright

Joe Mooney, Ben Matthews, John Osborne

Nik Dimitrikakis, Shanda Lightbown, Kazuo Imaizumi, Dana Bolgen, Anna Waterhouse

# Thank You, Questions?