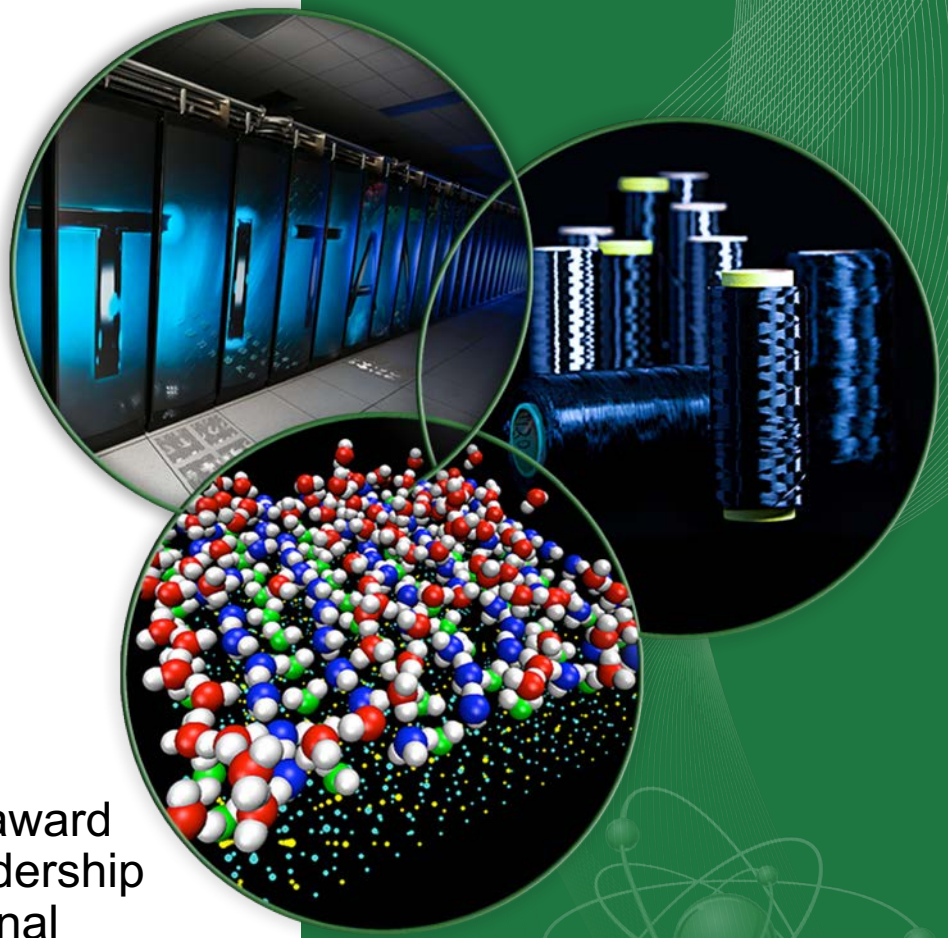
Breaking the curse of dimensionality

Explainable-AI and Evidence Mining as Applied to Systems Biology

Dan Jacobson



This research is supported by an INCITE award and uses resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE AC05-00OR22725.

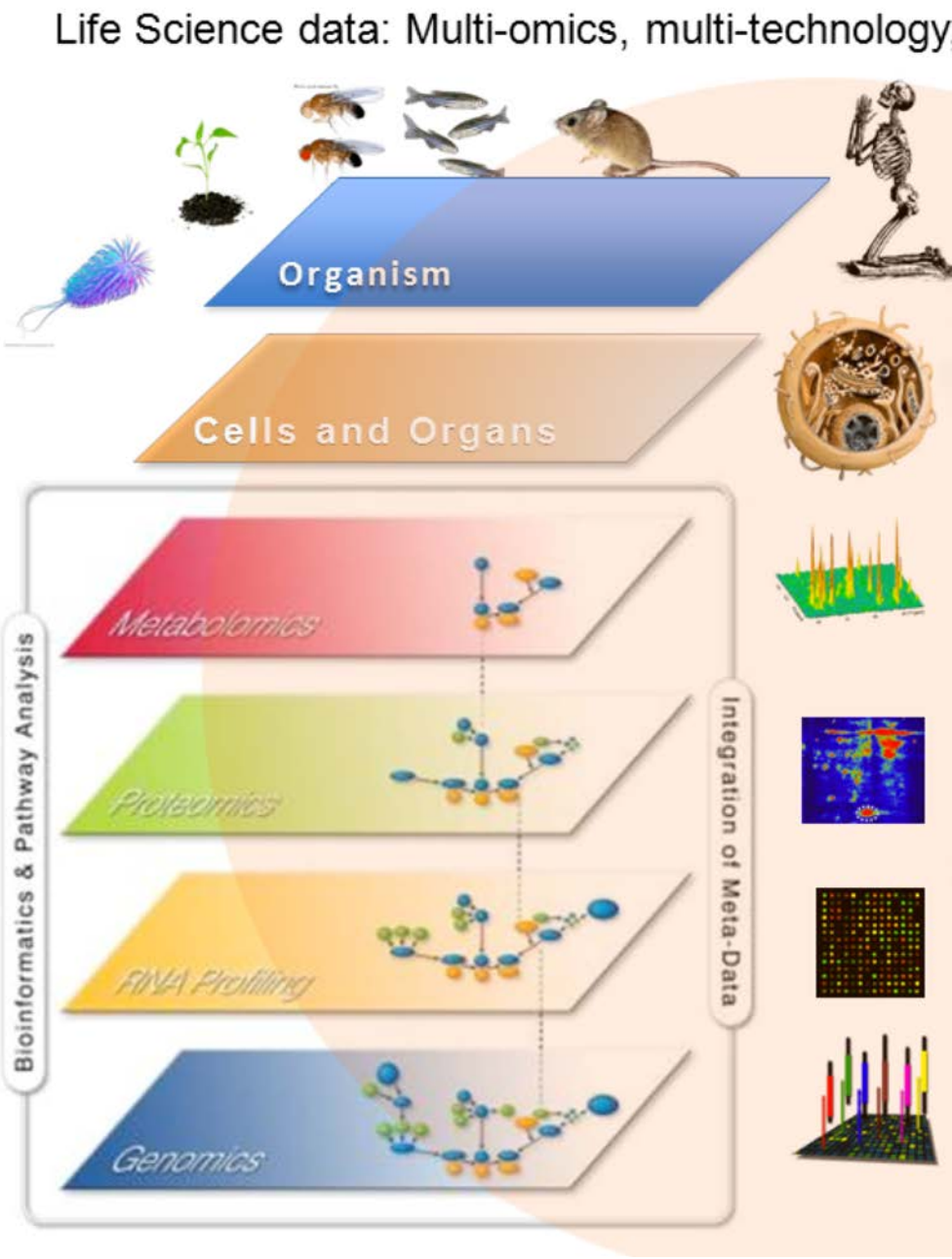
U.S. DEPARTMENT OF ENERGY

ORNL is managed by UT-Battelle
for the US Department of Energy

OAK RIDGE National Laboratory

Experimental Data Types

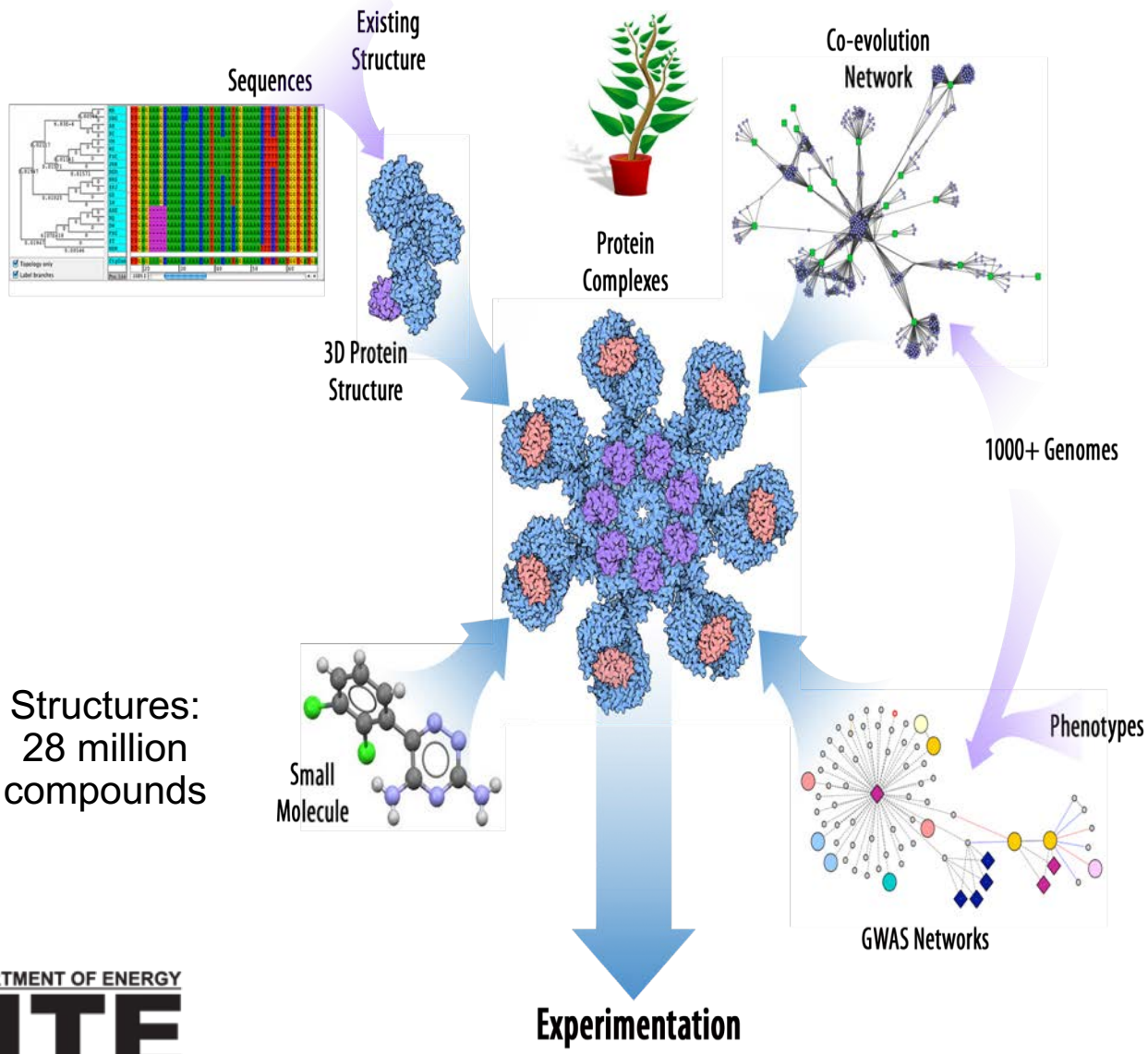
- Natural Variation
  - Genome Wide Association Studies
  - 28 Million of SNPs
  - ~140,000 Primary Phenotypes
    - Morphology/Phenology
    - Molecular
- Microbiomes & Metagenomes
- Omics & Meta-omics
  - Genomics, Transcriptomics, Proteomics, Metabolomics
- All publically available Genomes
- Differential/Time Series Expression Studies
- Systems Biology Approach
  - Combining datasets across omics layers, sample sets, and species



Traditional Results

Gene	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
Pavir.Aa00004	23.03260874	-0.772176419	0.235718754	-3.275837864	0.00105349	0.036650136
Pavir.Aa00067	3.617339133	-3.277187207	0.925328577	-3.541647029	0.000397637	0.016344905
Pavir.Aa00318	11.69495376	-1.375554763	0.421360908	-3.264552399	0.001096372	0.037862673
Pavir.Aa01140	432.2298561	-0.920355344	0.087912301	-10.46901667	1.20E-25	1.26E-22
Pavir.Aa01336	14.76644122	-7.964343955	1.643802037	-4.84507488	1.27E-06	0.000109099
Pavir.Aa01612	63.51089454	1.524126268	0.377624869	4.03608552	5.44E-05	0.002965915
Pavir.Aa01614	86.61299946	1.970704034	0.235133135	8.381226395	5.24E-17	2.16E-14
Pavir.Aa01686	45.57577197	-2.776318341	0.3350917	-8.285249514	1.18E-16	4.66E-14
Pavir.Aa01805	7.784684493	1.72469978	0.269249957	6.405571227	1.50E-10	2.64E-08
Pavir.Aa01856	15.77390176	-3.03656463	0.739522148	-4.106117228	4.02E-05	0.00228249
Pavir.Aa01950	246.4158349	0.749398201	0.130879565	5.725861023	1.03E-08	1.35E-06
Pavir.Aa02015	194.2868719	0.55688662	0.146656817	3.797209232	0.000146334	0.007032352
Pavir.Aa02104	71.8661413	-0.945676165	0.223959112	-4.222539364	2.42E-05	0.001454015
Pavir.Aa02130	45.08826603	-2.821545181	0.381707372	-7.391906442	1.45E-13	3.90E-11
Pavir.Aa02199	82.09354863	2.652283666	0.48092843	5.514923839	3.49E-08	4.08E-06
Pavir.Aa02377	48.01170214	1.765138681	0.318940668	5.534379463	3.12E-08	3.70E-06
Pavir.Aa02382	4.900020424	-6.641133503	1.55203963	-4.278971603	1.88E-05	0.001166295
Pavir.Aa02400	3.536707907	-2.288869563	0.396004267	-5.779911361	7.47E-09	1.01E-06
Pavir.Aa02455	100.2653536	0.851939179	0.154407276	5.517480799	3.44E-08	4.03E-06
Pavir.Aa02456	74.76890191	0.900755926	0.267107154	3.372264319	0.000745529	0.027702451
Pavir.Aa02462	129.7507991	1.878568856	0.195429139	9.612532015	7.08E-22	5.19E-19
Pavir.Aa02463	0.855875118	-3.952874961	1.177355482	-3.357418402	0.00078674	0.028956754
Pavir.Aa02517	239.8175815	3.424148863	0.634311687	5.398211843	6.73E-08	7.46E-06
Pavir.Aa02526	20.12897762	-1.829988585	0.513742501	-3.56207357	0.000367937	0.015318345
Pavir.Aa02574	1.957536218	-5.978272647	1.222823914	-4.888907208	1.01E-06	8.89E-05
Pavir.Aa02621	0.909365395	-6.53529993	1.672432432	-3.907661562	9.32E-05	0.004726253
Pavir.Aa02666	26.2769212	0.691682664	0.195671446	3.534918755	0.000407901	0.01668753
Pavir.Aa02688	20.64051337	1.419916888	0.311120505	4.563880767	5.02E-06	0.000367199
Pavir.Aa02777	32.70837314	0.824566433	0.256714392	3.211999243	0.001318147	0.044226251
Pavir.Aa02799	5.953157198	1.635139531	0.489562315	3.340002856	0.000837775	0.030512025
Pavir.Aa02841	4.061306867	-1.69398357	0.345840001	-4.898171305	9.67E-07	8.51E-05
Pavir.Aa03067	7.20334301	-6.09679446	1.535018046	-3.971806374	7.13E-05	0.003773958

Integrated Vision: From Systems Biology to 3D Structural Interactions

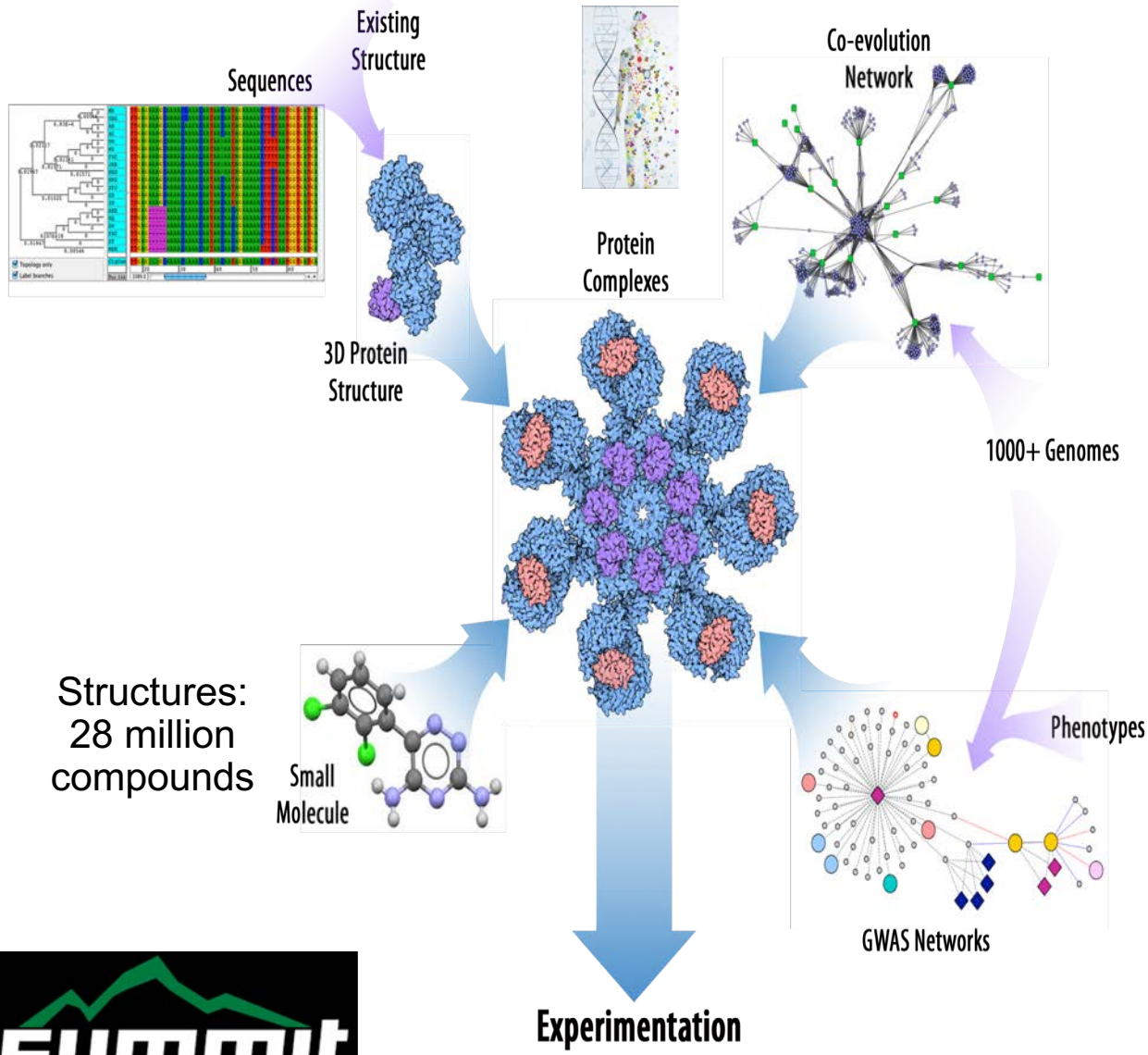


Structures: 28 million compounds

U.S. DEPARTMENT OF ENERGY INCITE LEADERSHIP COMPUTING

TITAN

Integrated Vision: From Human Systems Biology to 3D Structural Interactions - Pharmacogenomics and Personalized Medicine

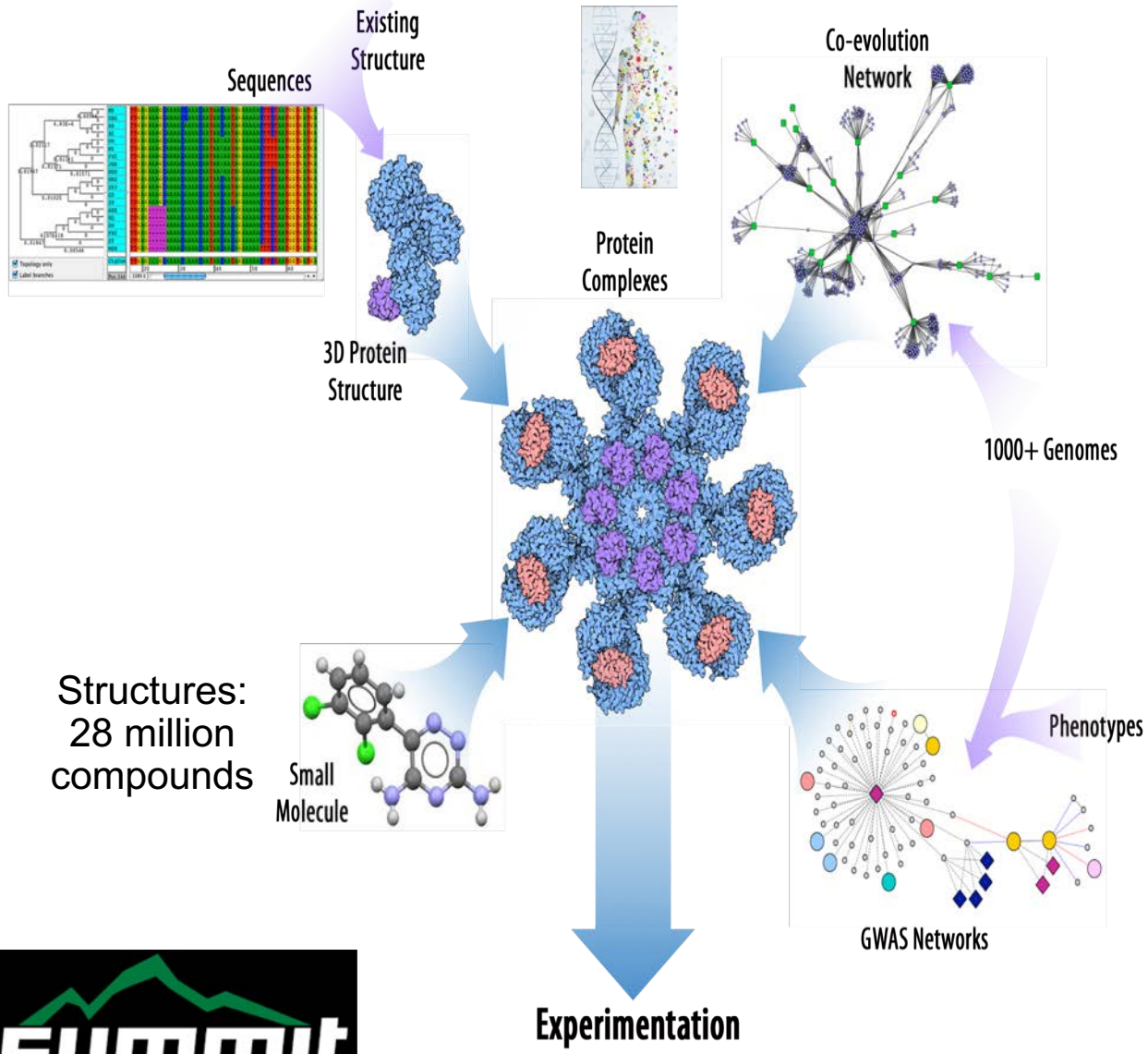


Structures: 28 million compounds

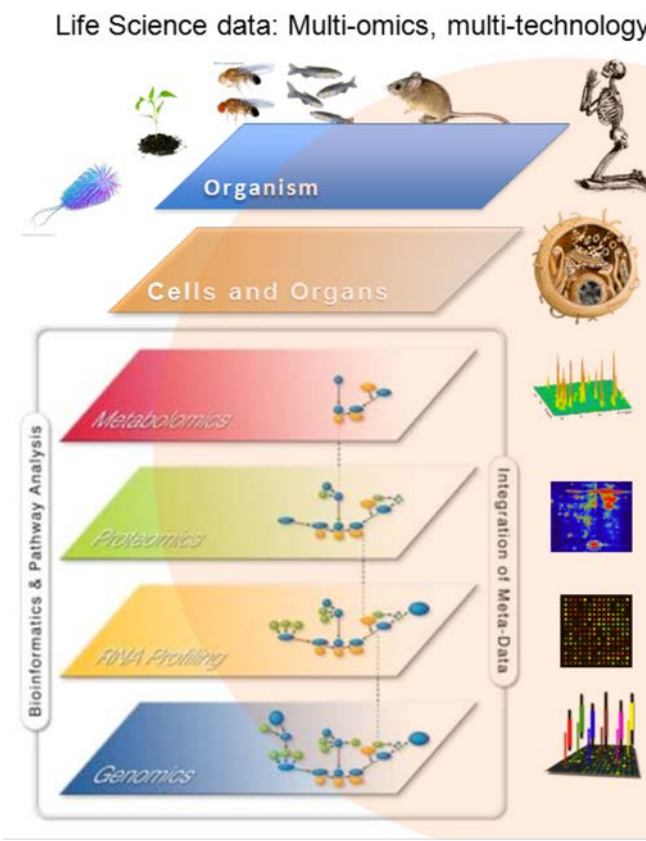
- Co-evolution
  - 1000 Genomes Project
  - Protein cross species
- Human RNA-seq
- Crystal Structures
- Protein-protein interaction
- Human interactome
- ENCODE
- **Explainable-AI**
  - iRF/TiRF
  - DNNs
  - MIPs
- Public GWAS data
- ***As available from collaboration with VA***
  - Genetic data
  - Clinical phenotypes
  - Polypharmacy data

summit

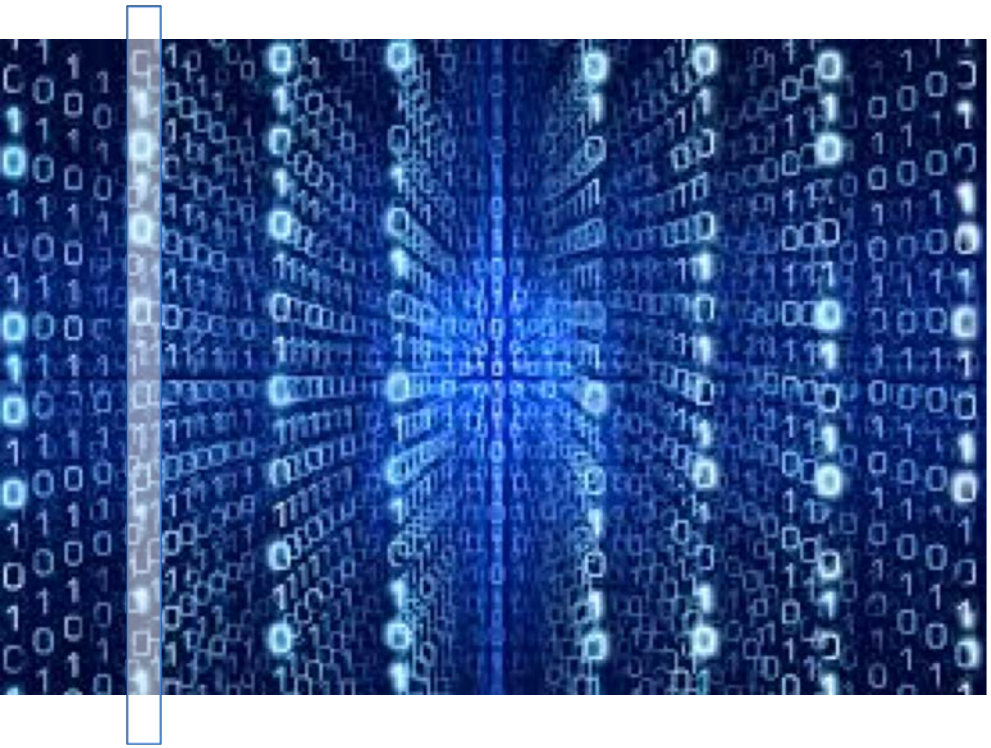
Integrated Vision: From Human Systems Biology to 3D Structural Interactions - Pharmacogenomics and Personalized Medicine



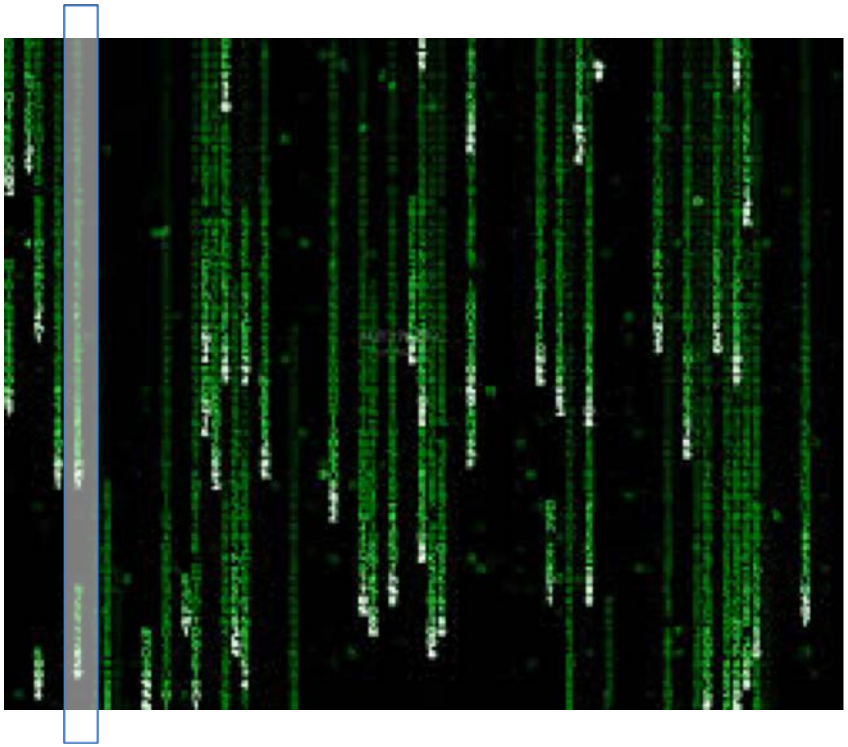
Structures: 28 million compounds



Single QTL mapping: 28 million tests per phenotype

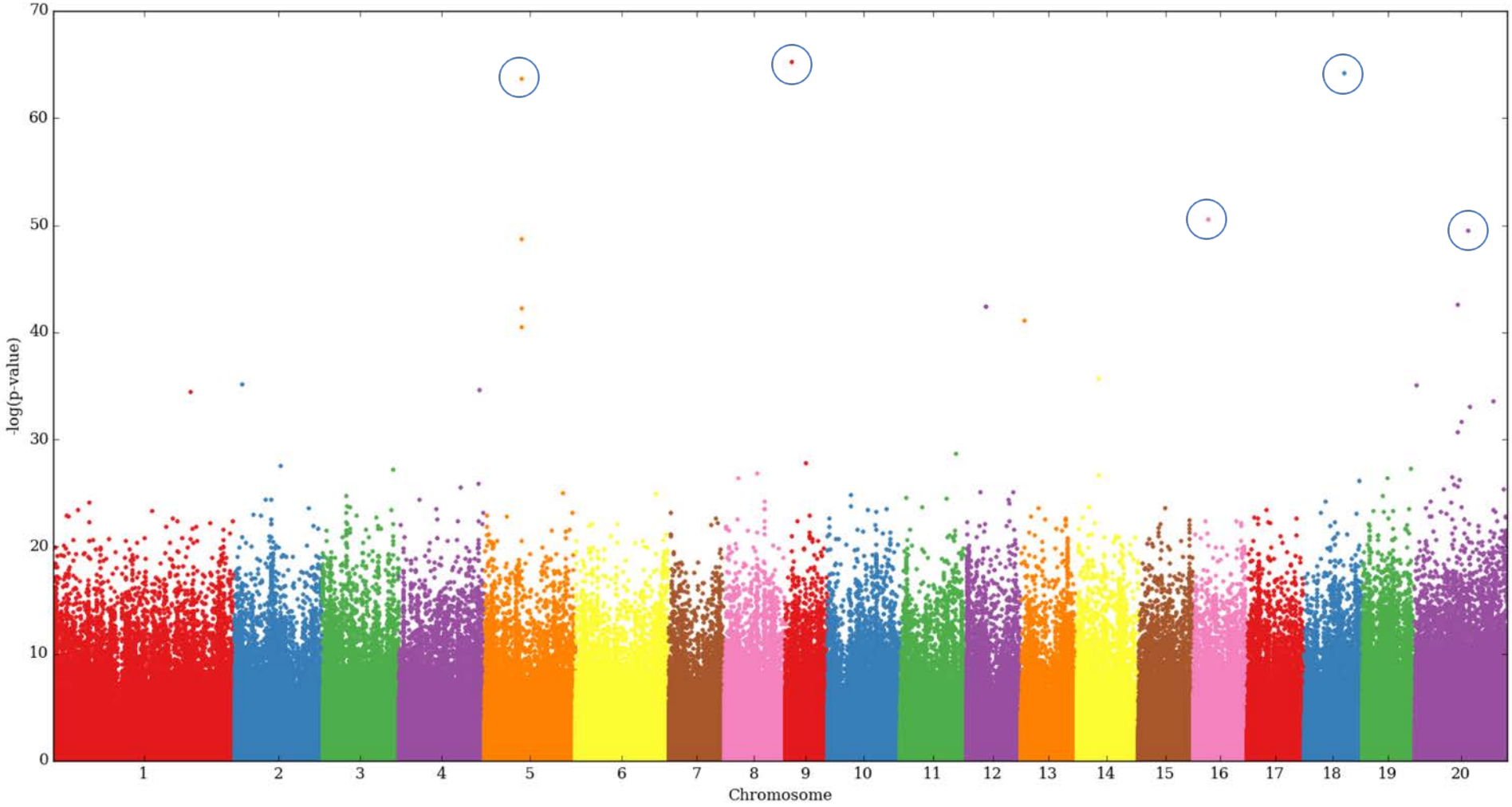


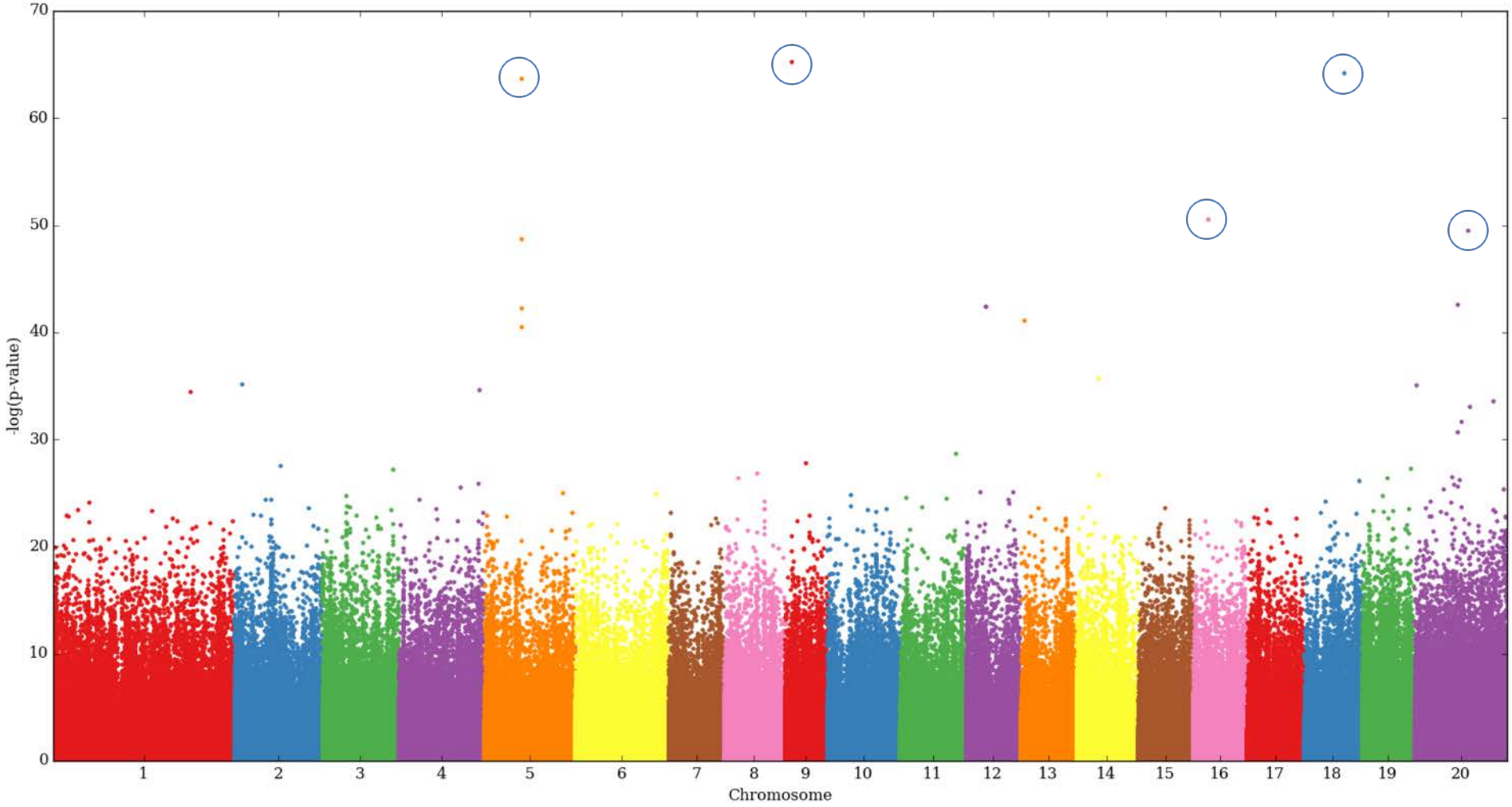
SNP Vectors



Phenotype Vectors

- SNP Matrix expansion from interpolation (ANL)
- Control for effects of population structure

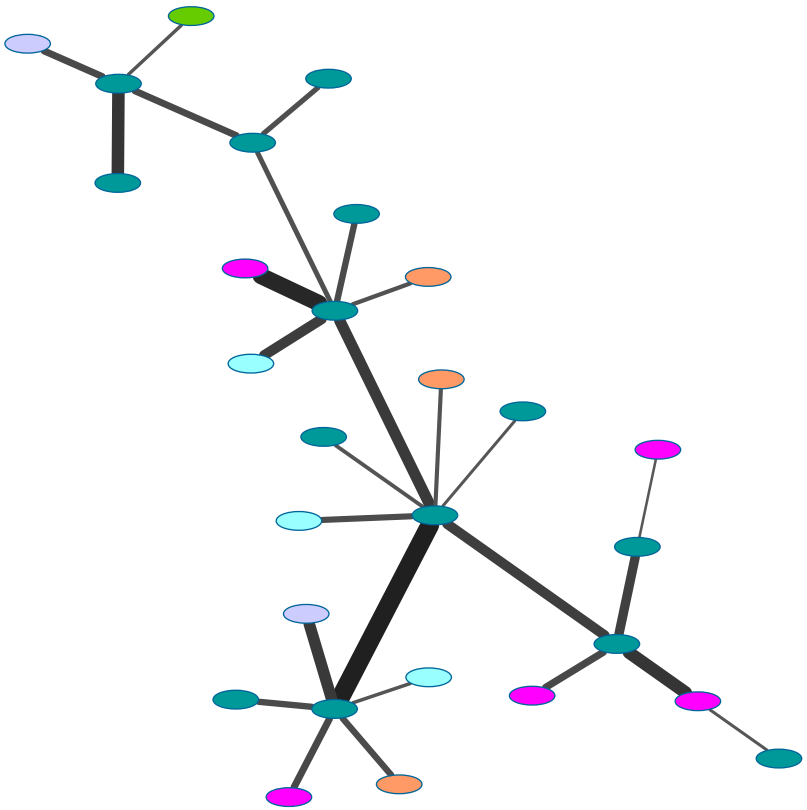

70
60
50
40
30
20
10
0
-log(p-value)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Chromosome

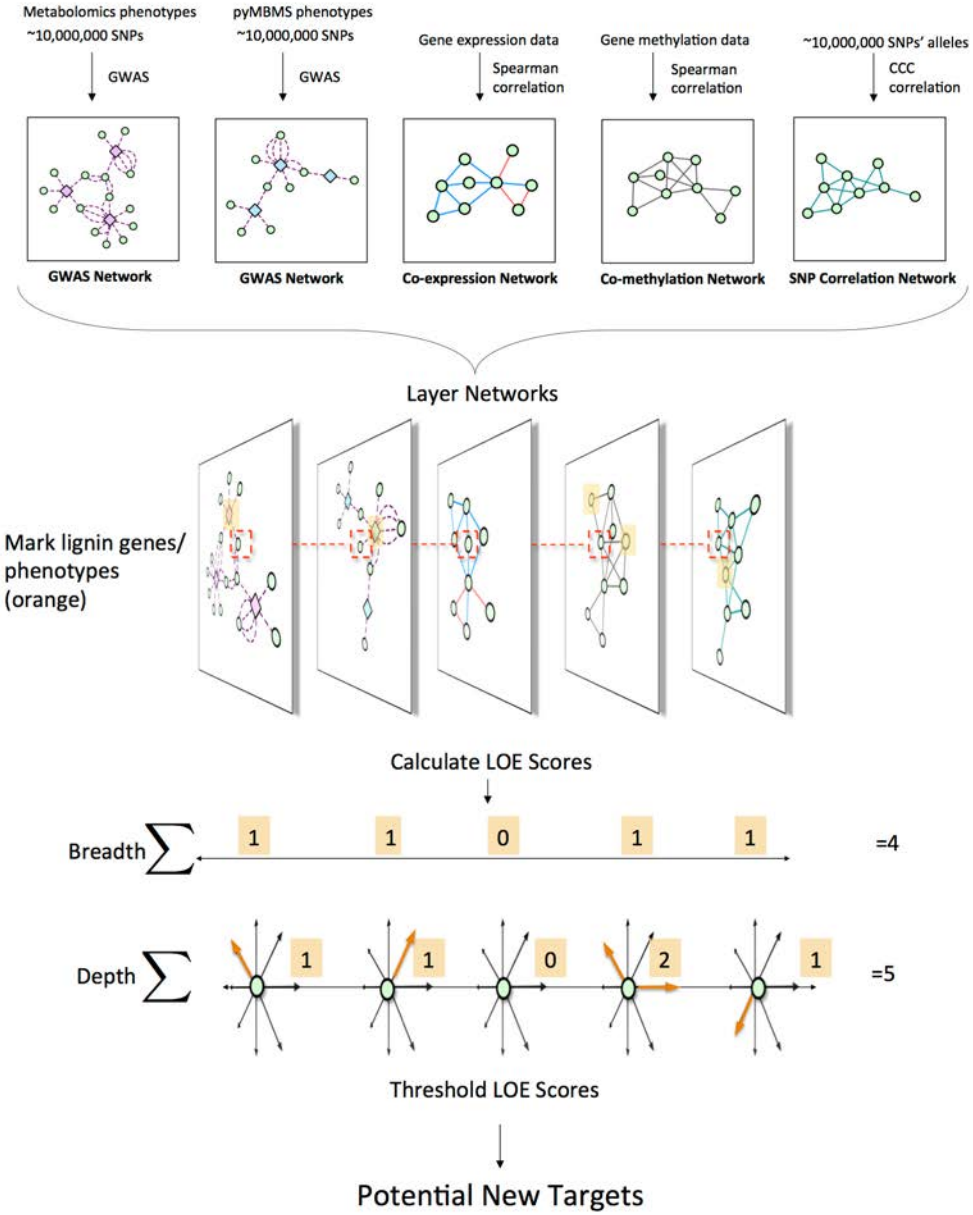


140,000 Manhattan Plots???

Network Theory

- Networks can be used to represent biological systems
  - Nodes
    - Represent any object (genes, SNPs, proteins, metabolites, species, microbiomes, etc.)
  - Edges
    - Represent a relationship between two nodes (correlation, co-occurrence, physical contact, etc.)
    - Relationships can be quantitative (represented by the thickness of the line)
- Integration and Visualization of Systems Biology Models
- Mathematical Structure
  - Allows to be computed upon
  - Millions of nodes
  - Trillions of edges





Pleiotropic and Epistatic Network-Based Discovery: Integrated Networks for Target Gene Discovery. Deborah Weighill , Piet Jones, Manesh Shah, Priya Ranjan, Wellington Muchero, Jeremy Schmutz, Avinash Sreedasyam, David Macaya Sanz, Robert Sykes, Nan Zhao, Madhavi Martin, Stephen DiFazio, TImothy Tschaplinski, Gerald Tuskan, **Daniel Jacobson.** *Front. Energy Res. - Bioenergy and Biofuels*, DOI: 10.3389/fenrg.2018.00030

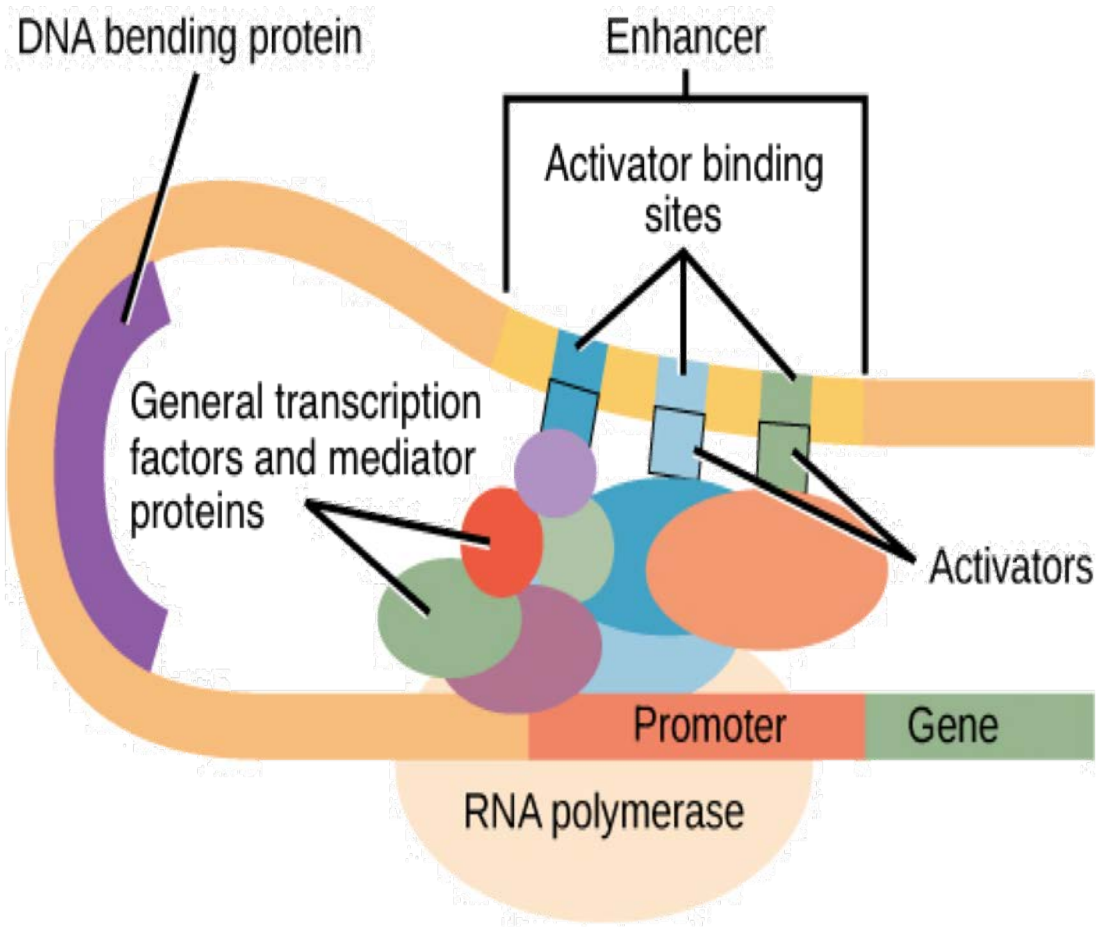
Deeper Discoveries in Systems Biology: The Balance Between Type 1 and Type 2 Error

Our ability to reconstruct the entirety of a complex biological system improves as the number of population-scale endo-, meso- and exo-phenotypes are measured and combined with deep layers of experimental data collected on individual genotypes.

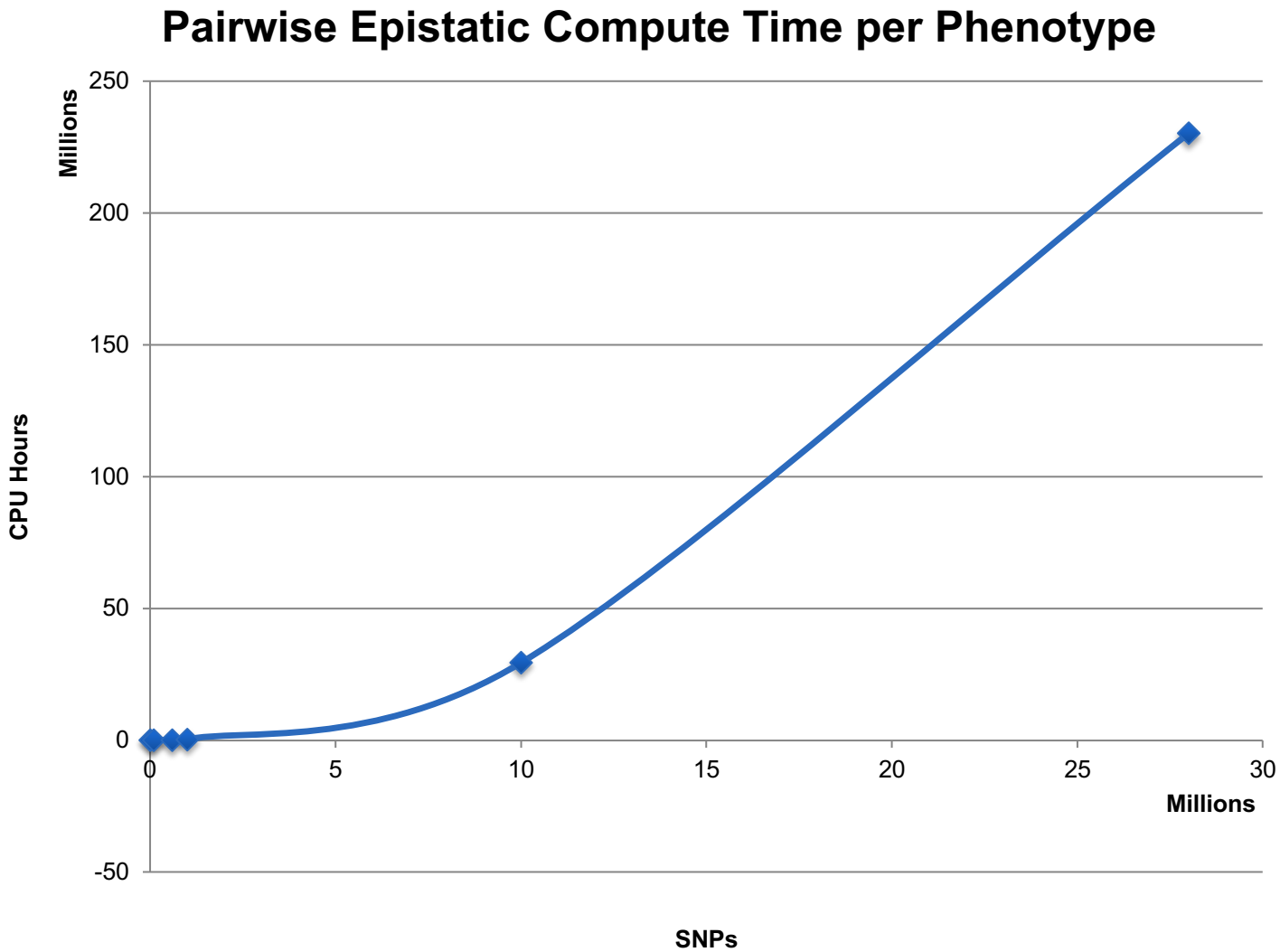
GWAS: Single QTL Mapping

- Very Powerful
- Frequently does not capture a significant portion (often the majority) of the genetic signal
- Often does not find complete genetic architectures for complex phenotypes (Dementia, Alzheimer's, Schizophrenia, Cardiovascular disease, PTSD, Suicide, Addiction, etc.)

Epistatic Example: Transcription Initiation Complex



The Need for Speed



4-way combinations = 2.4 x 10^{20} CPU hours per phenotype

Breaking the curse of dimensionality

10M Genetic Variants in >40k genes

Genes do not work in isolation: 10^{170} potential interactions among variants

Linking genetic variants to phenotypes requires the exploration of an enormous space

To obtain accuracy and insight, we are developing procedures to detect interactions of any form or order at the same computational cost as main effects

Explainable-AI

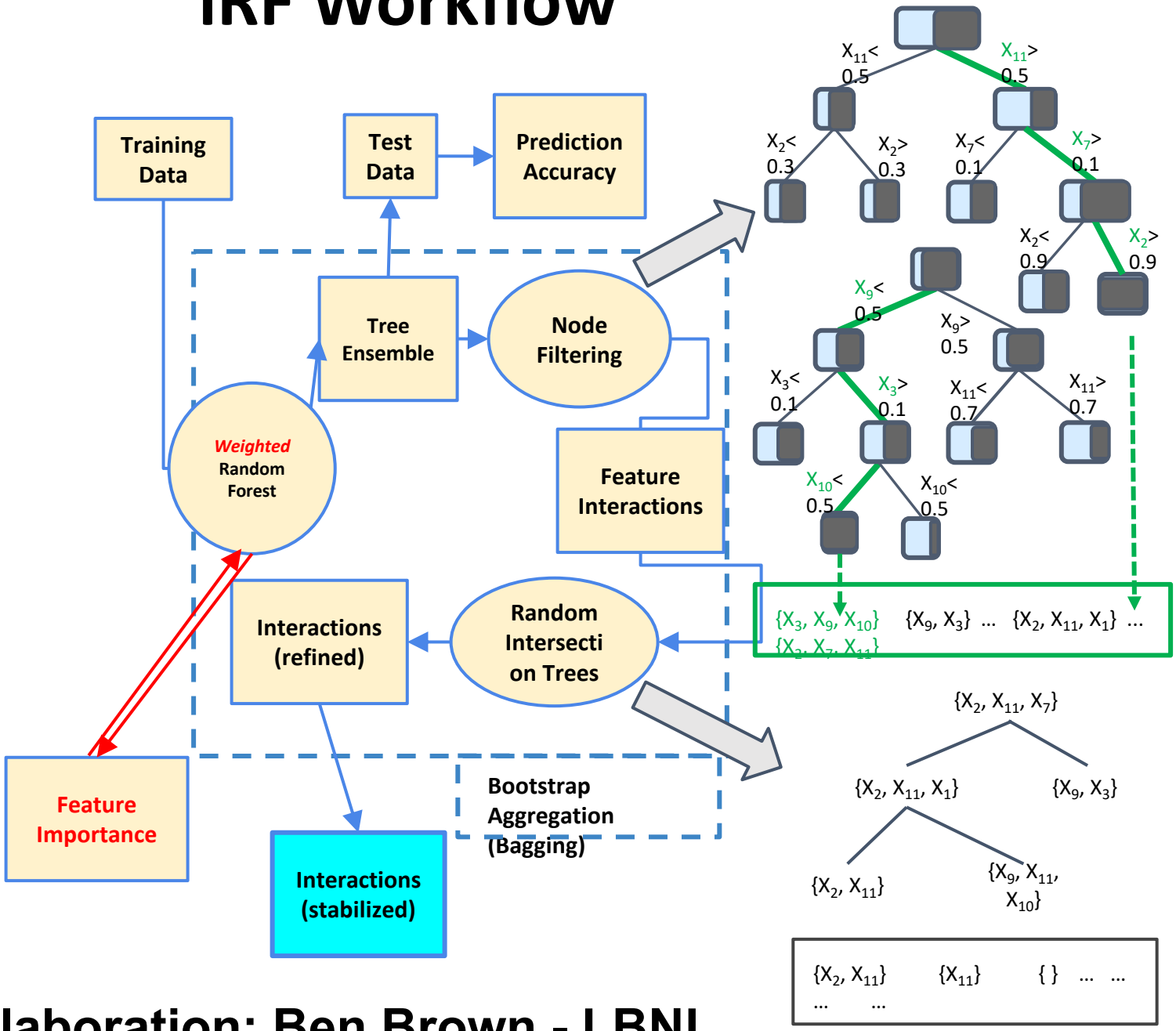
Machine and Deep Learning Algorithms

- Great at classification
- Essentially black boxes
  - Don’t reveal the interactions between variables that lead to the classification
- Need Explainable AI

Finding Higher Order Combinatorial Interactions in Complex Systems

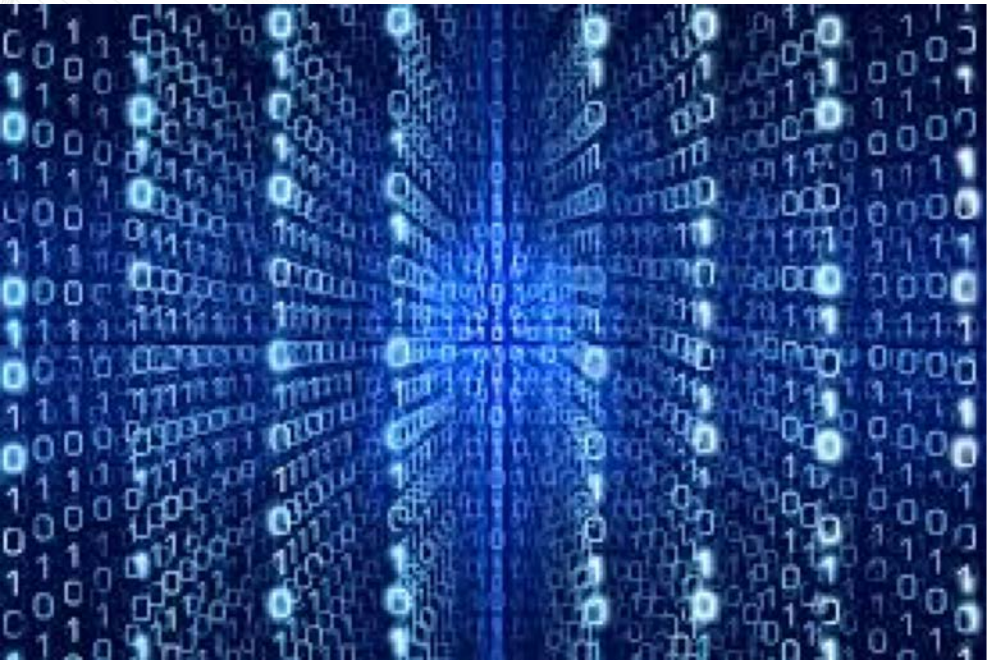
- X matrix and Y vector
- Iterative Random Forests

iRF Workflow



DOE Collaboration: Ben Brown - LBNL

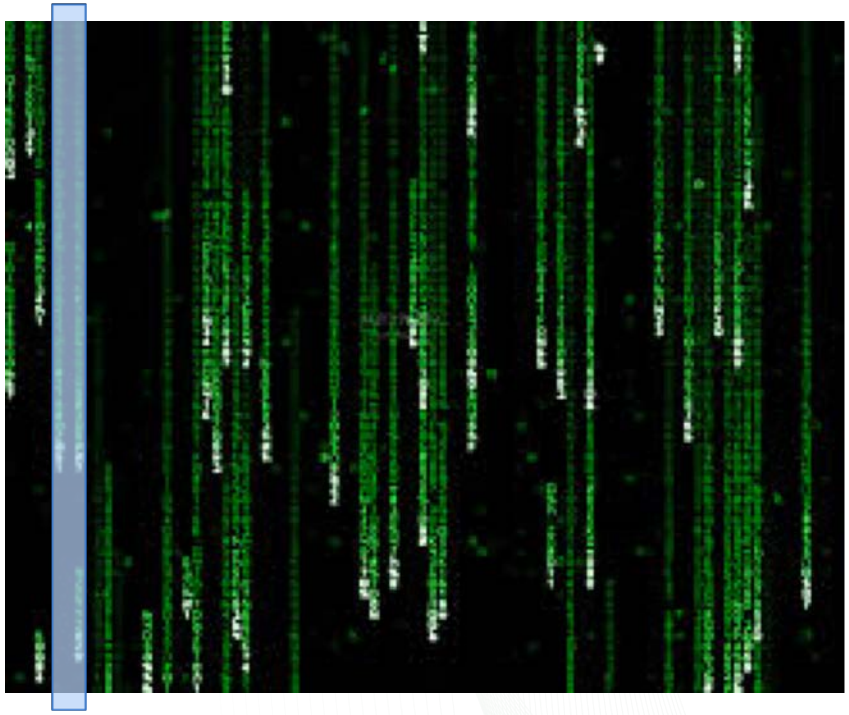
iRF – X Matrix and 1 Y Vector



SNP Vectors

Phenotype Vectors

iRF – X Matrix and 1 Y Vector



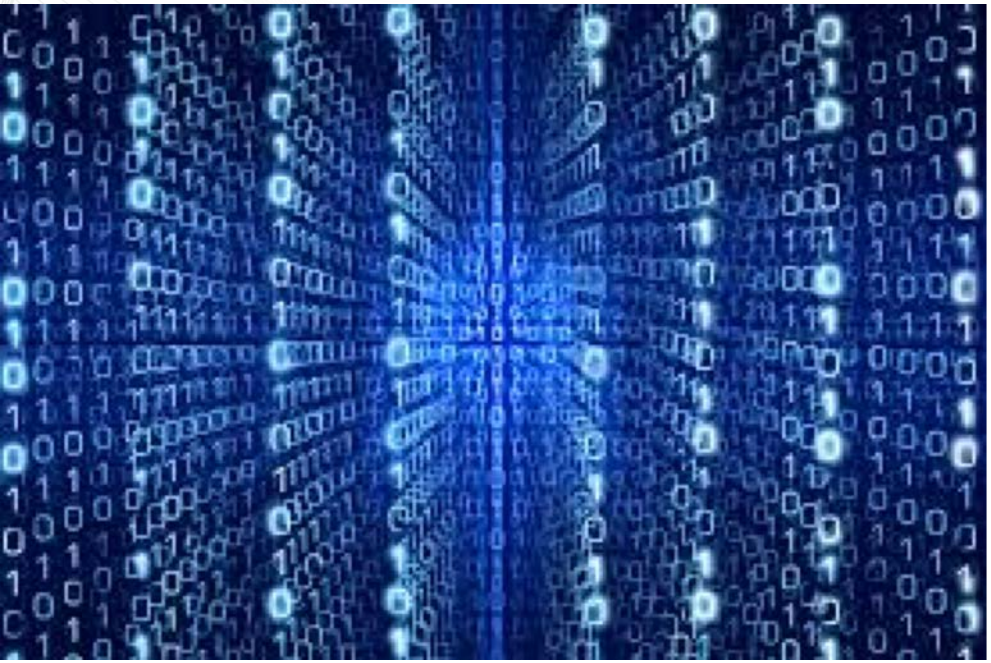
SNP Vectors

Phenotype Vectors

4-way combination = 1000 CPU hours per phenotype (140,000 phenotypes)

Tensor iterative Random Forests (TiRFs)

- Effectively build forests that can be mined for interactions within a multi-dimensional X, a multi-dimensional Y and interactions between multiple dimensions in X and Y, all at the same time.

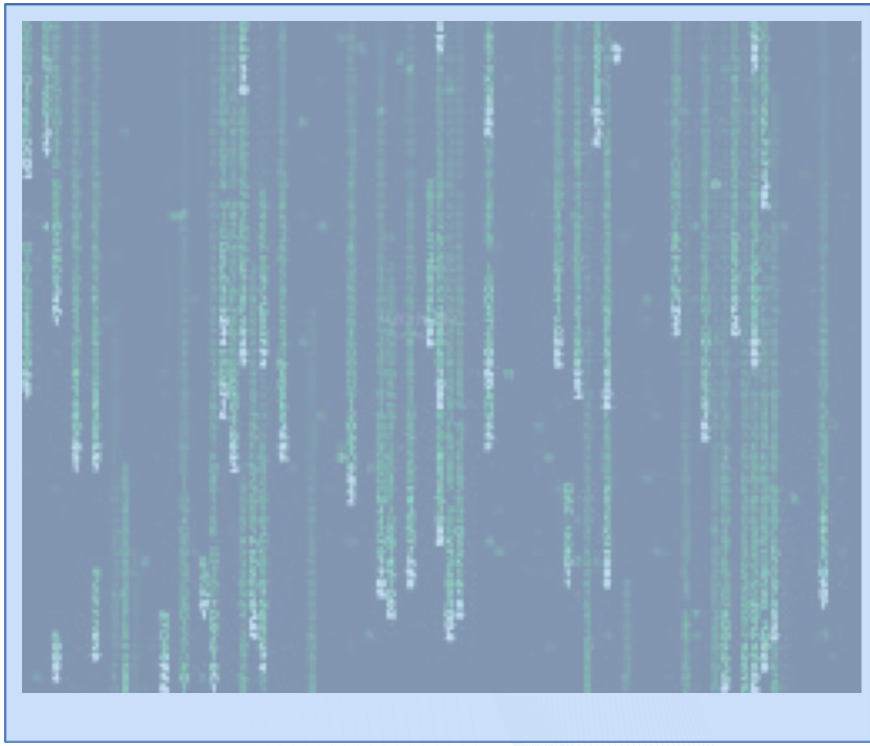


SNP Vectors

Phenotype Vectors

TiRF – X Matrix and Y Matrix *Simultaneously*

SNP Vectors



Phenotype Vectors

Clinical Genomics and Human Systems Biology: DOE & VA – MVP Champion

- ORNL
  - Clinical records 23+ million patients, 20 years
  - 358,000 Genotypes
  - => 4 million genotypes

VA Use Case: Polypharmacy

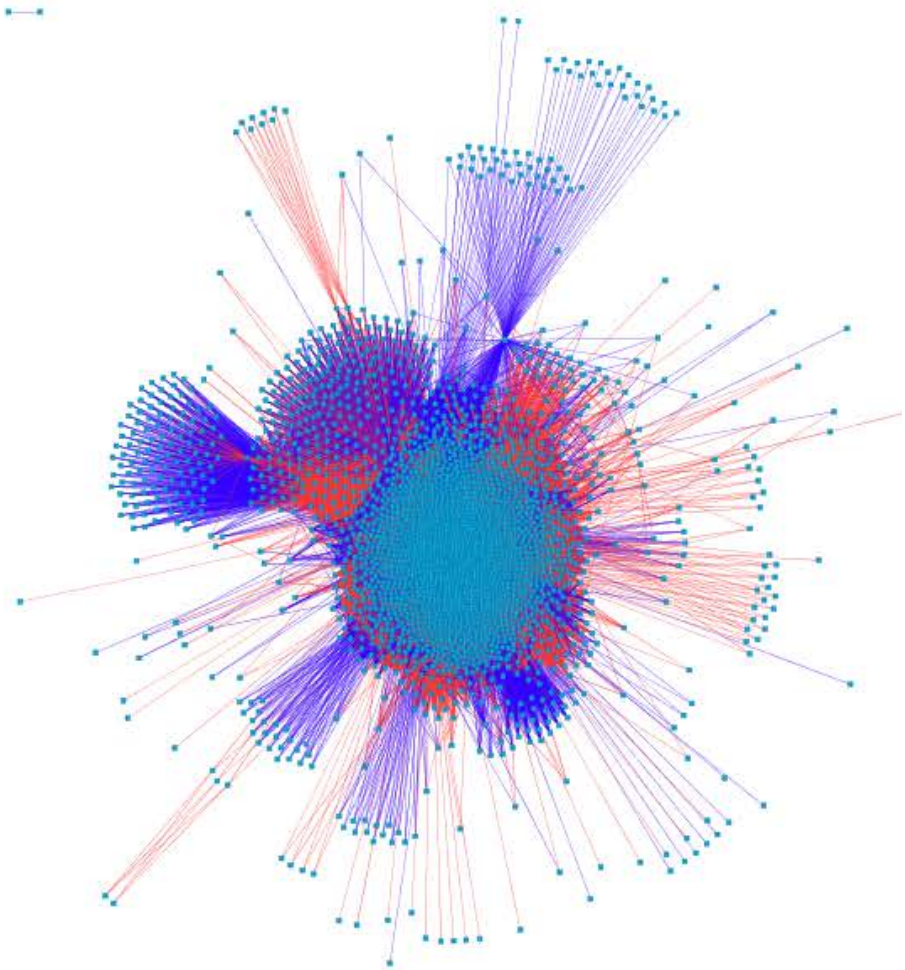
- Simultaneous use of multiple medication
- Of concern if 5 or more medications are used

Why Worry About Polypharmacy?

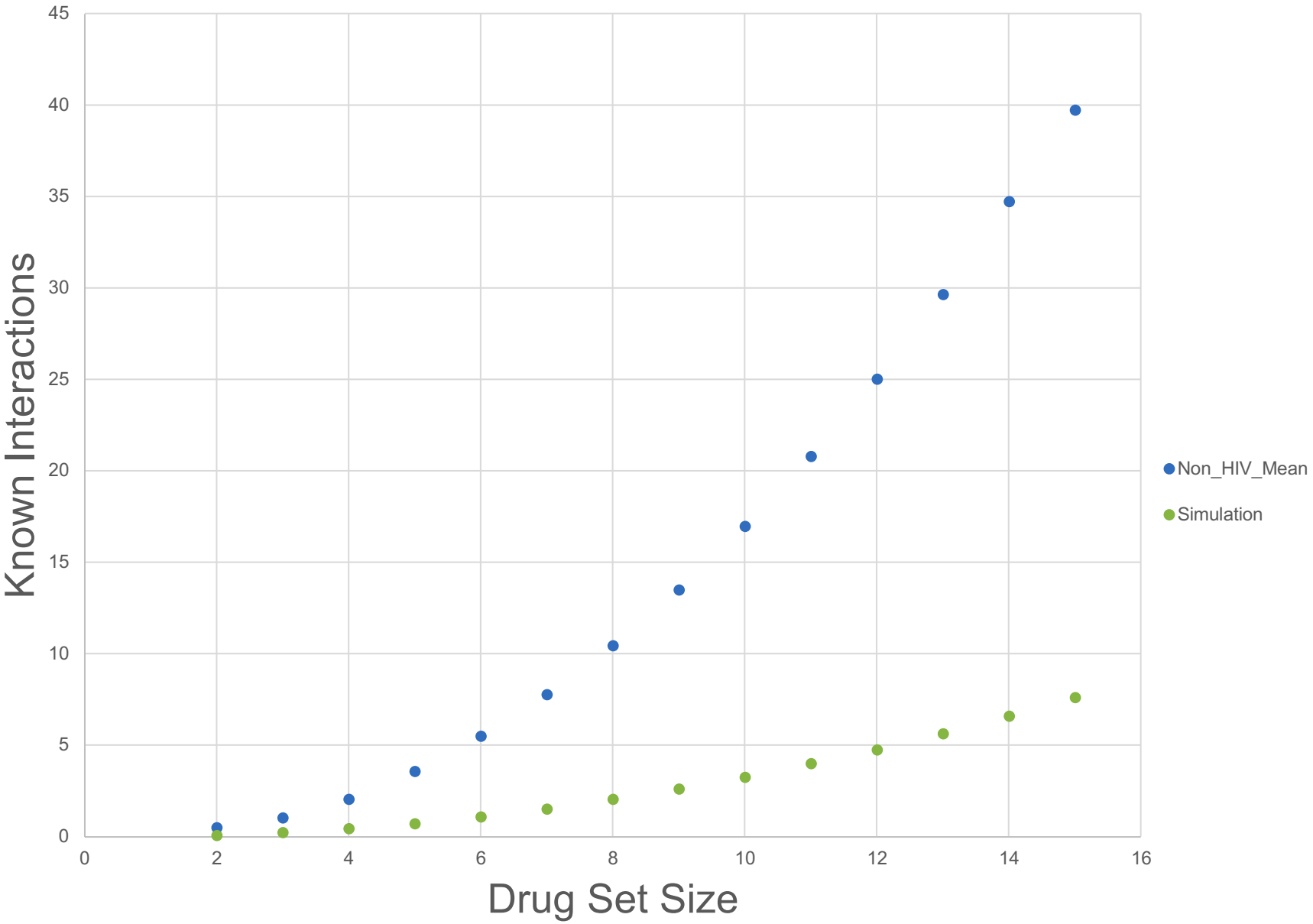
- Drugs interact with each other, the more you are on, the more interactions can occur
- Side effects add up and are more pronounced in older individuals
- Medications are approved by FDA based on short term trails that typically exclude:
  - Those with other diagnoses on other medications
  - 65+ year olds

Interaction Network

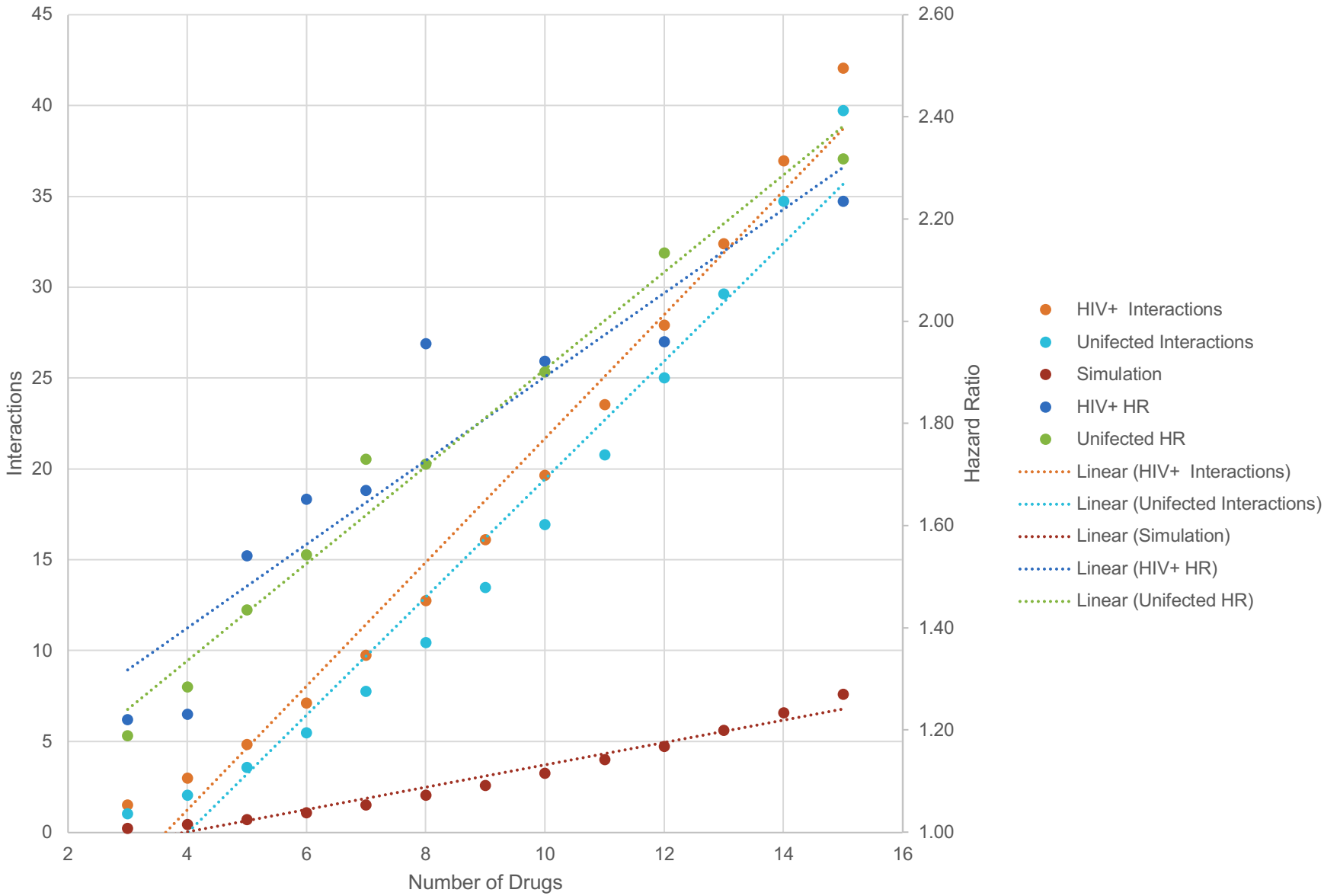
- Drug Set Simulations
  - For all set sizes 2 – 30
    - Create 20 million random sets of drugs for each set size
    - 58 million sets
  - Check for drug to drug edges amongst all possible pairs in each set for the shared target and shared pathway networks
    - 567 Billion interaction tests
- Clinical Data
  - Create drug sets from clinical records
  - Check for drug to drug edges amongst all possible pairs in each set for the shared target and shared pathway networks



Drug Interaction: Simulation vs Clinical Practice



Polypharmacy Morbidity & Mortality



VA: Preliminary Results

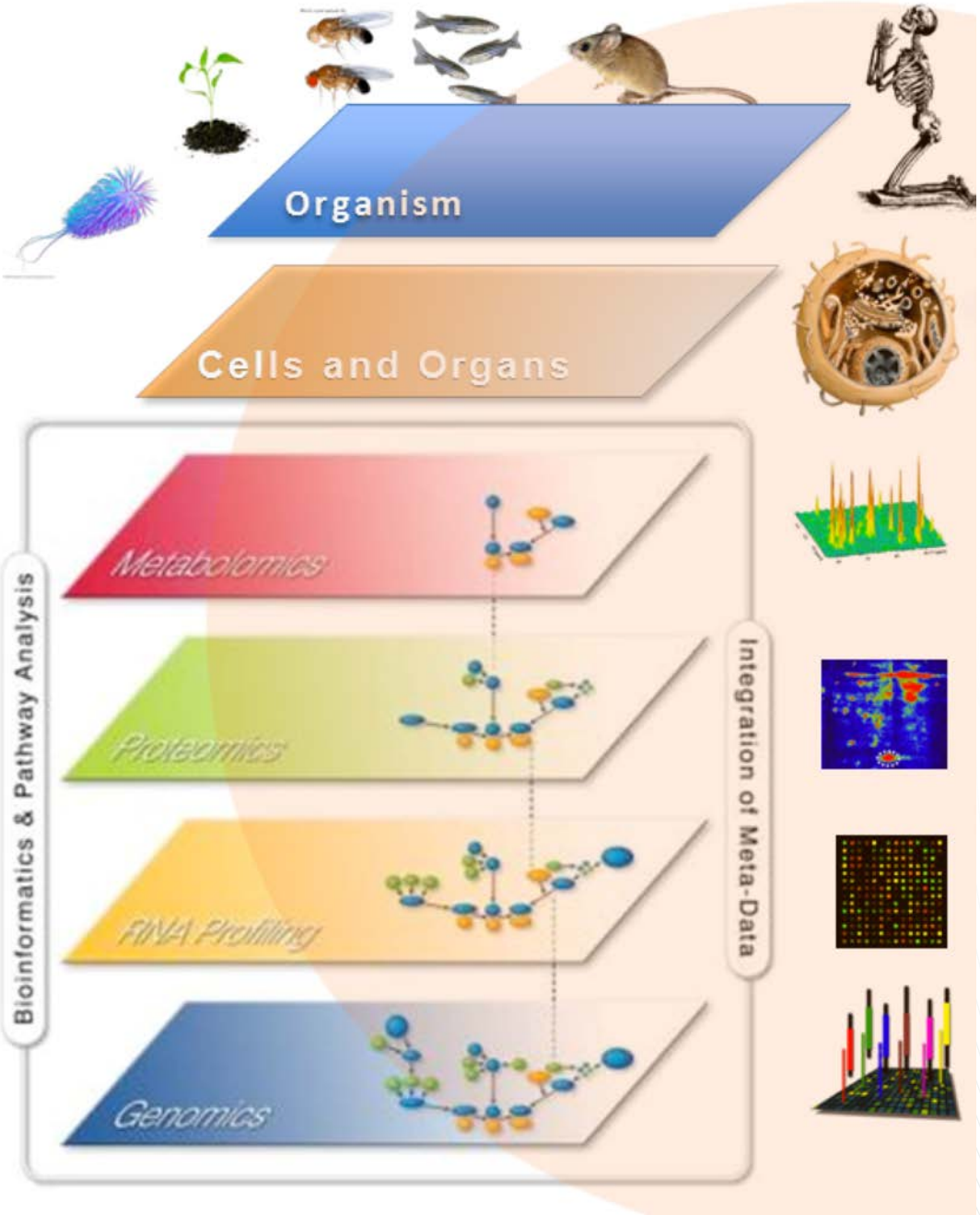
- **Polypharmacy**
  - Clinically relevant patterns
  - iRF
    - Steps toward automated phenotyping
    - Interaction edges -> morbidity & mortality

Drugs & Interactions

- **Diseaseome**
  - iRF on diagnostic codes
  - 600,000 patients
  - Relationships between all known human conditions
    - Co-morbidity map
  - Discovered 9th-order combinations
    - **1.5×10^{26} possible 9-way combinations**

Outcomes (morbidity and mortality)

Life Science data: Multi-omics, multi-technology



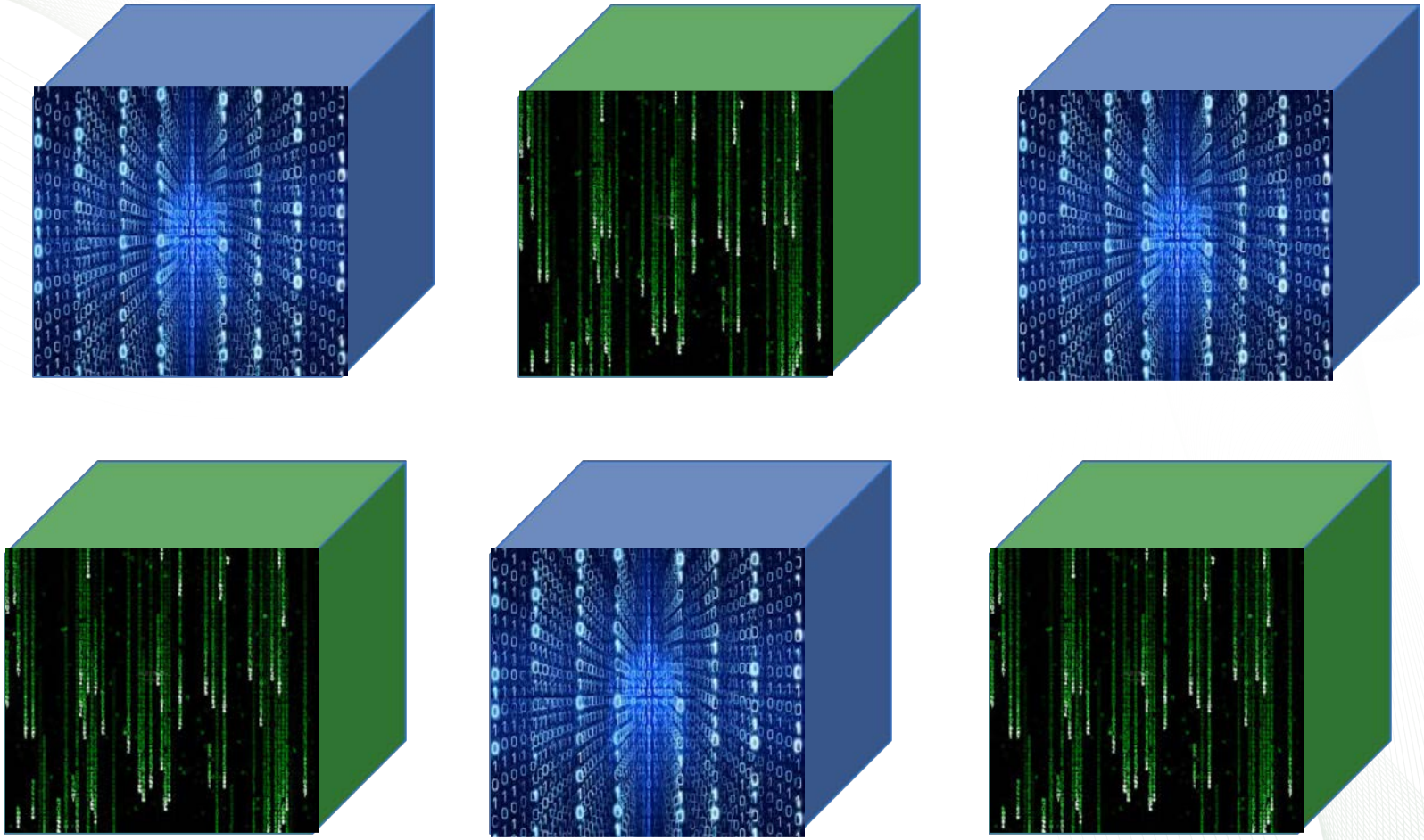
TiRF – Any Set of Matrices or Tensor Dimensions *Simultaneously*



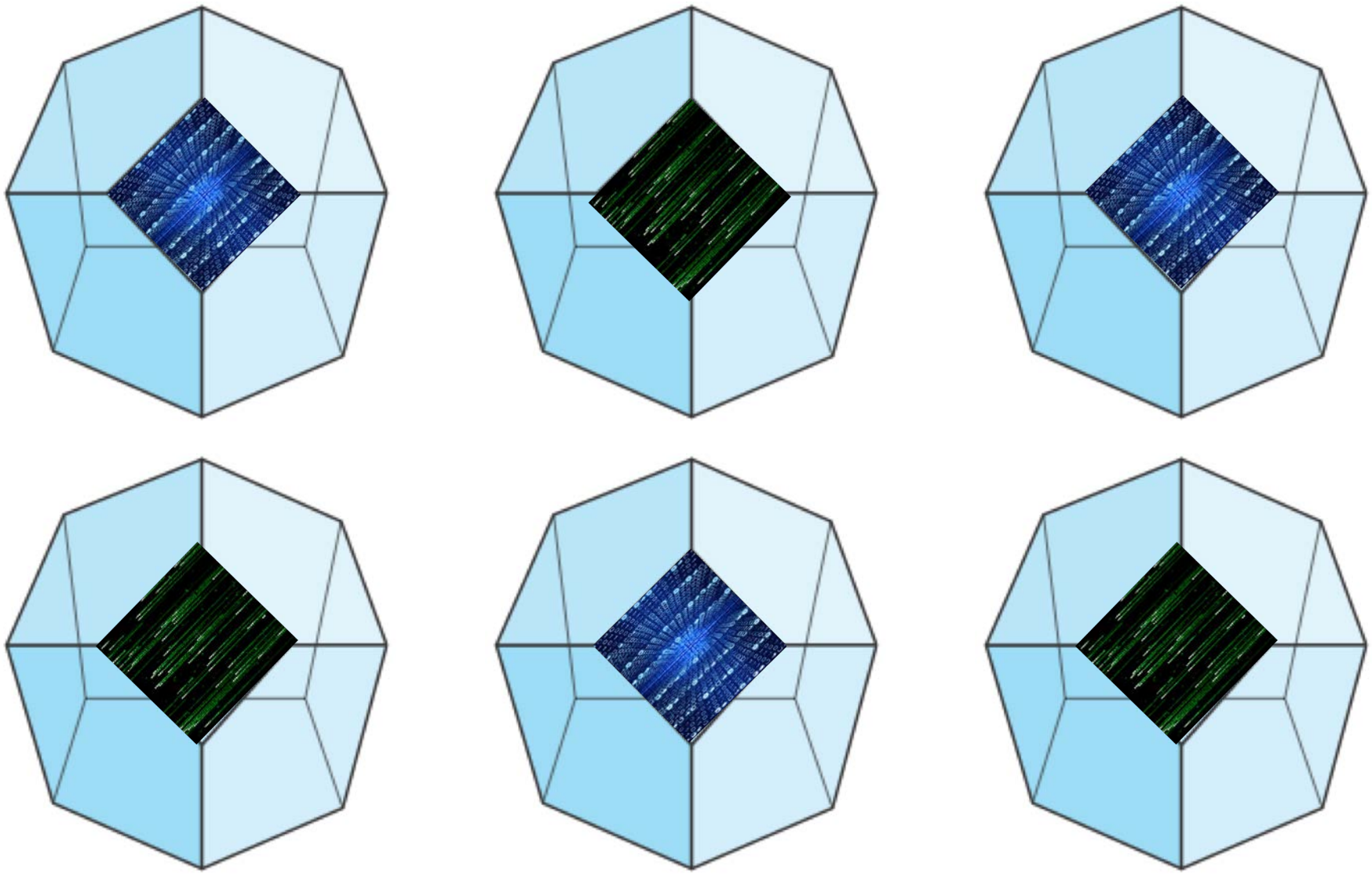



- Spatial and temporal/longitudinal information
- Different Omics layers (genome, transcriptome, proteome, metabolome, microbiome…)
- Quantum chemical tensors

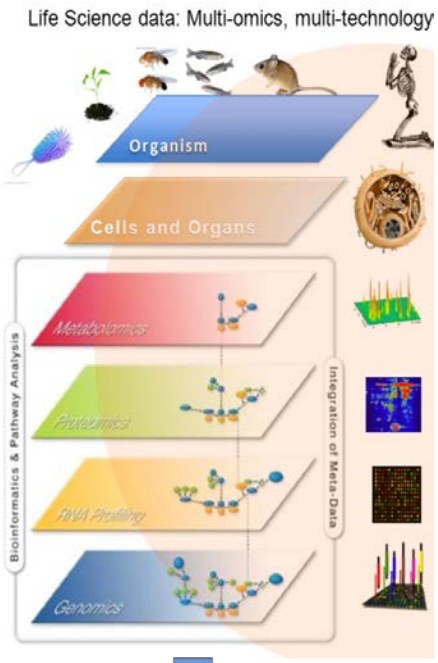
Tensors: Matrices ➡ Cubes

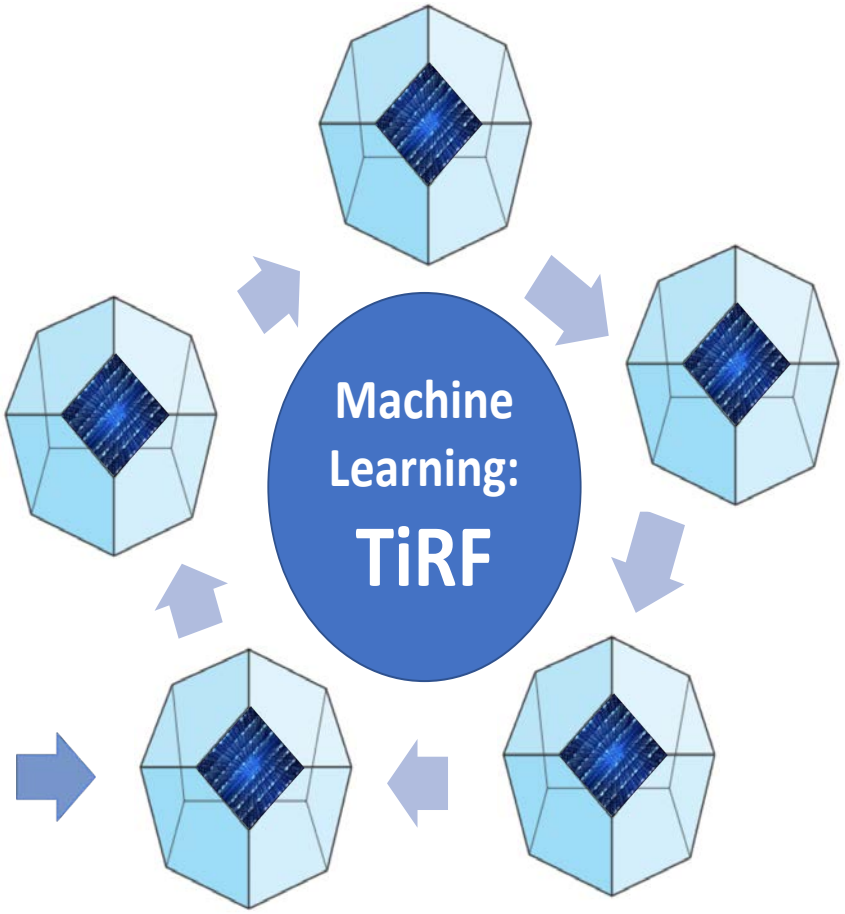


Tensors: Matrices ➡ Cubes ➡ Polytopes



From data matrix to cube to polytopes.





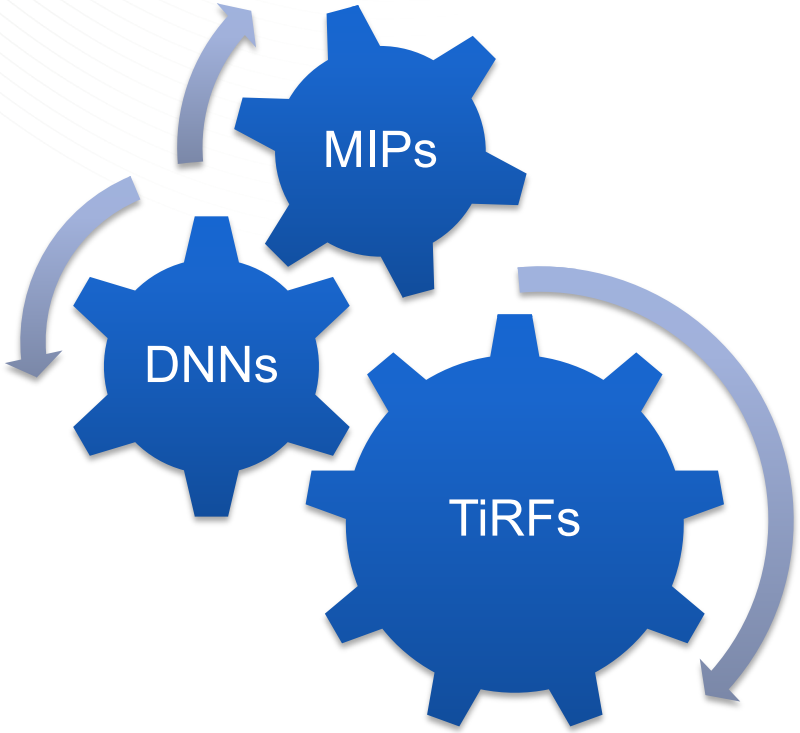
Tensor iterative Random Forests (TiRFs)

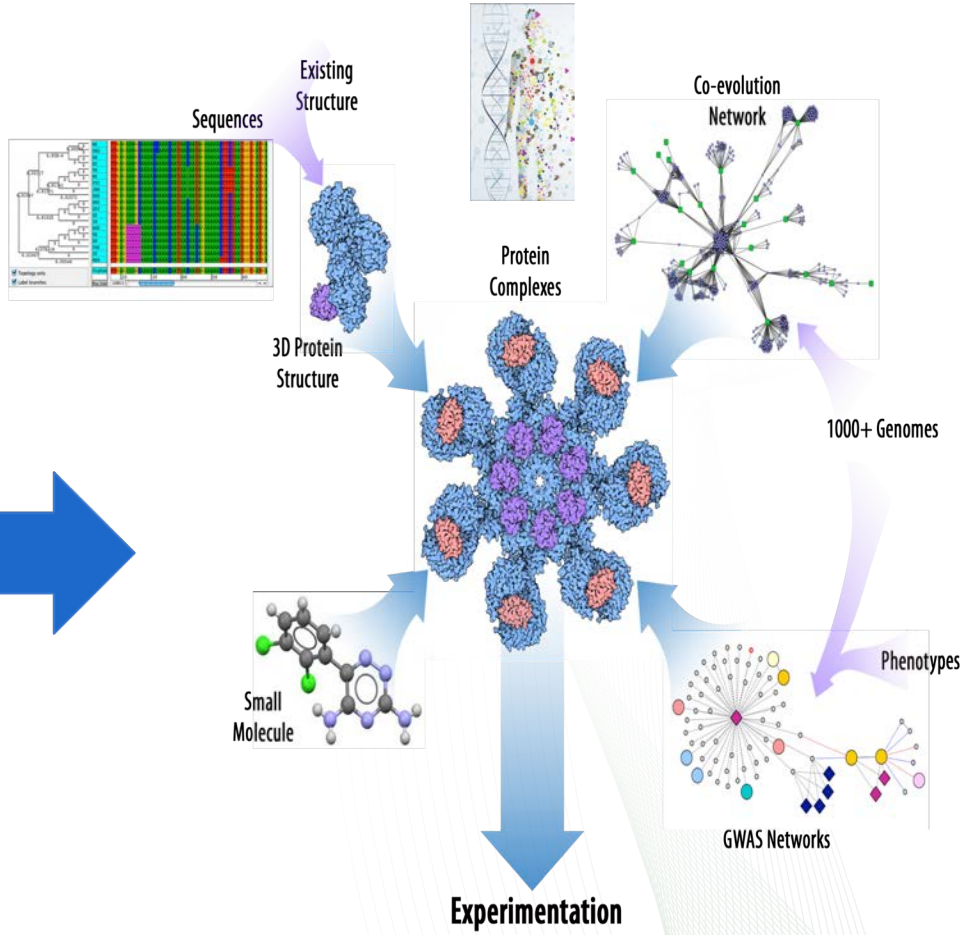
- Effectively build forests that can be mined for interactions within a multi-dimensional X, a multi-dimesional Y and interactions between multiple dimensions in X and Y, all at the same time.
- Applications in Systems Biology
  - Plants
  - Microbes
  - Humans, Mice
  - *Drosophila*
- Applications in Text Mining
  - Electronic Health Records
  - Scientific Literature
- Simulation Models
  - Combinatorial parameter sweeps (X) model output (Y)
- **Any domain with high a dimensional set of matrices**

Iterative Deep Neural Networks (iDNNs)

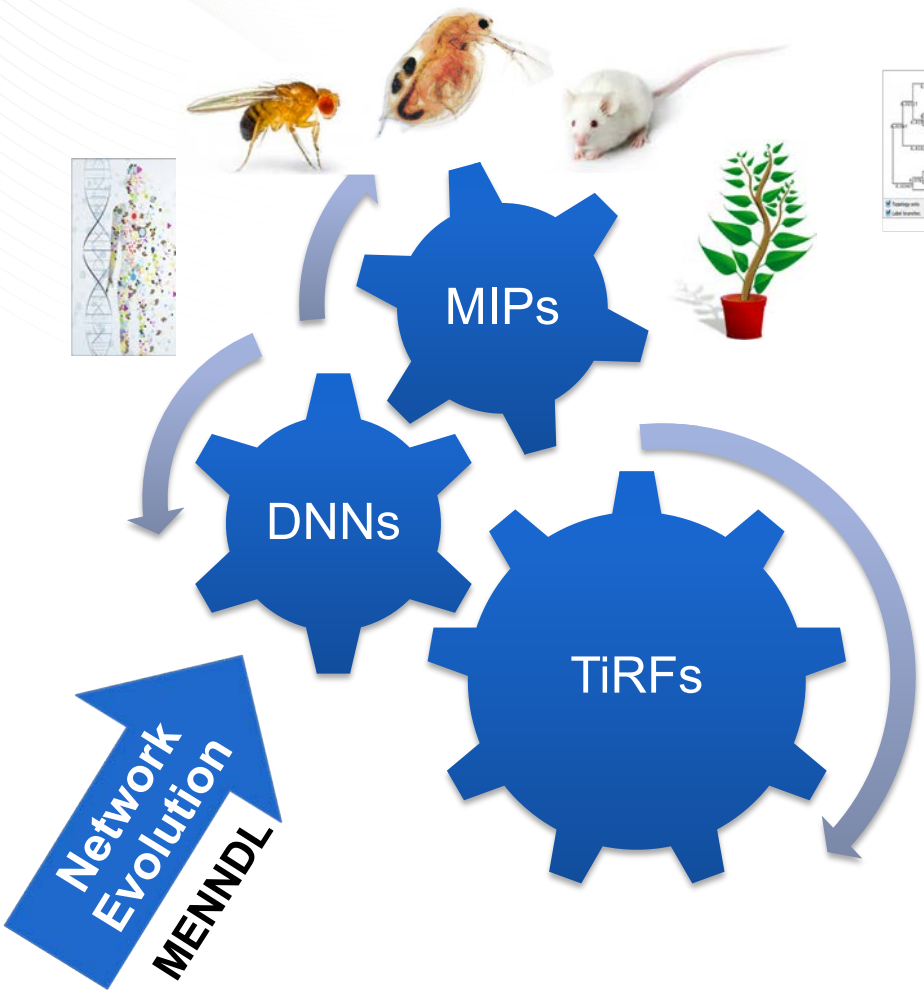
- Unpacking the black box
- Discovering the interactions encoded in DNNs

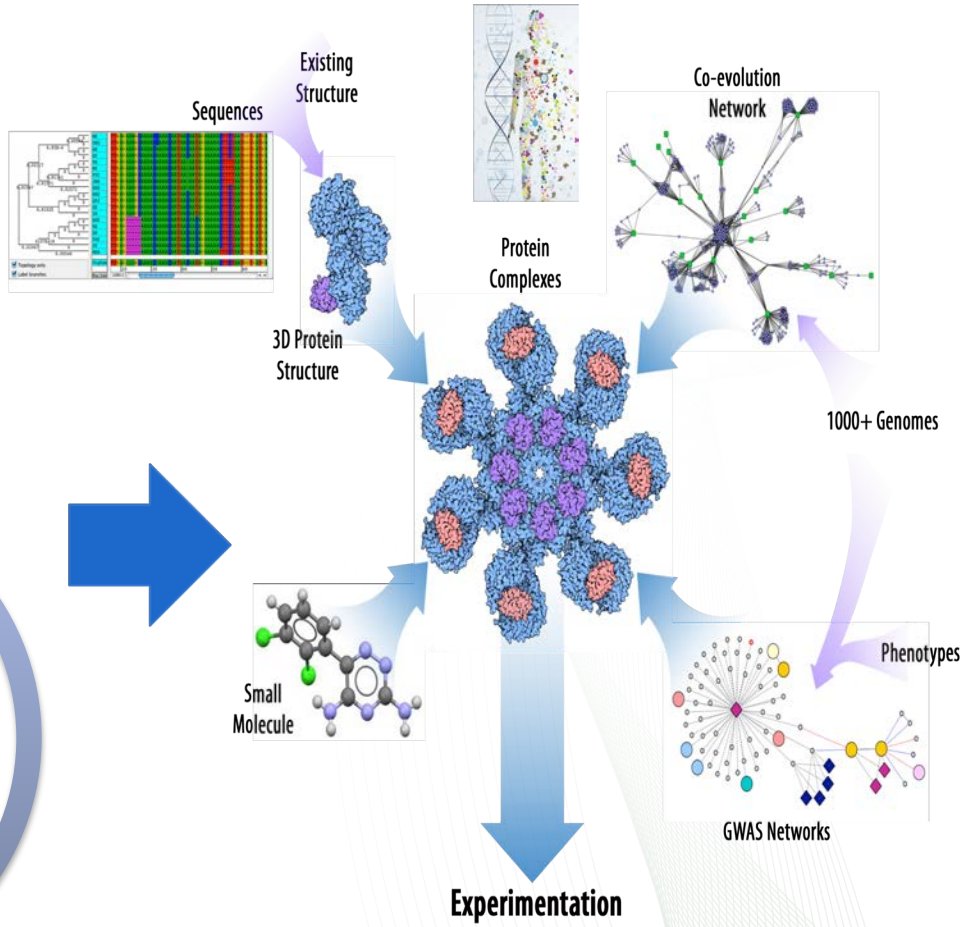
High Order Interactions: Explainable AI: Machine and Deep Learning Integration



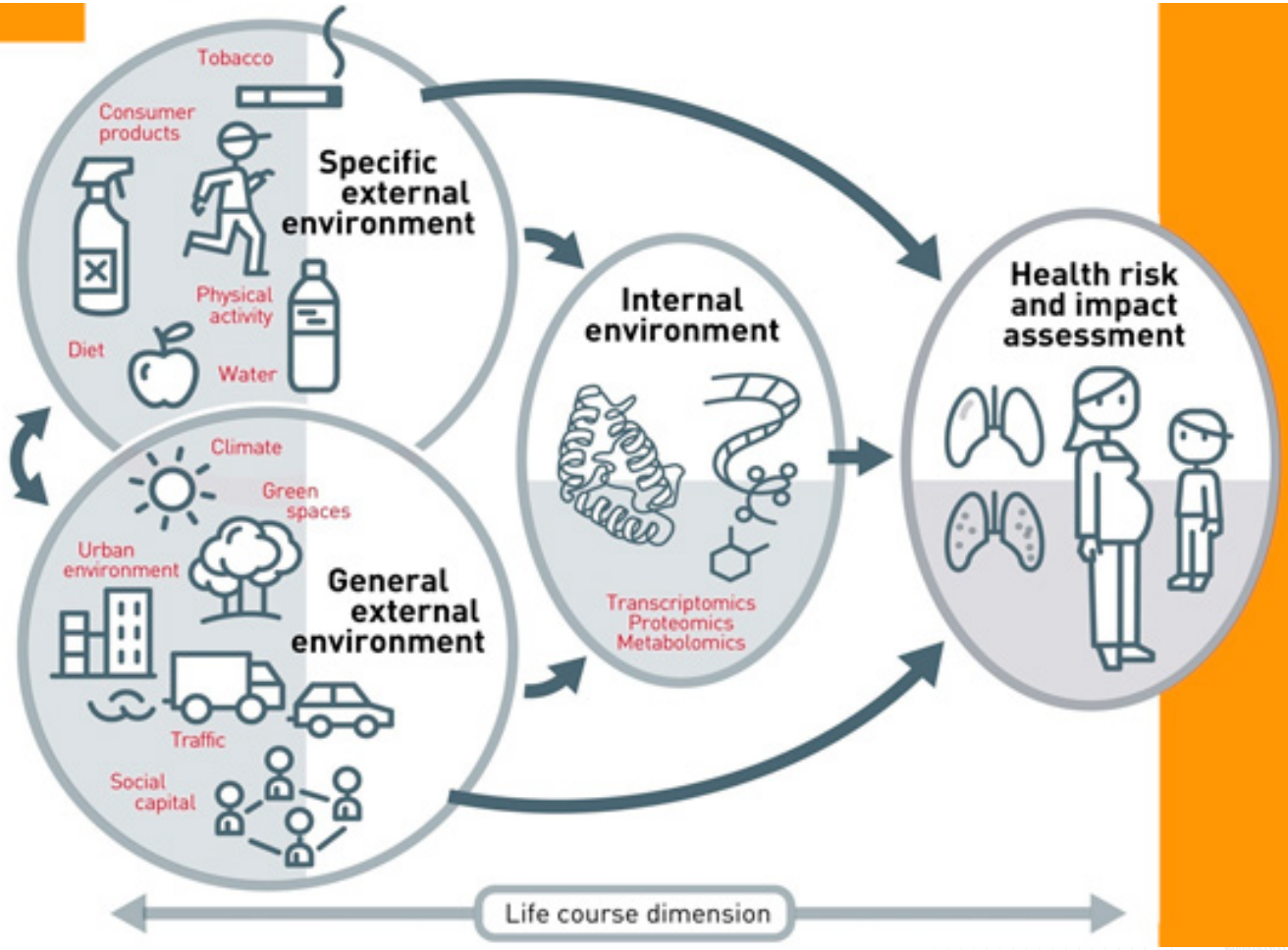


High Order Interactions: Explainable AI: Machine and Deep Learning Integration





Exposome

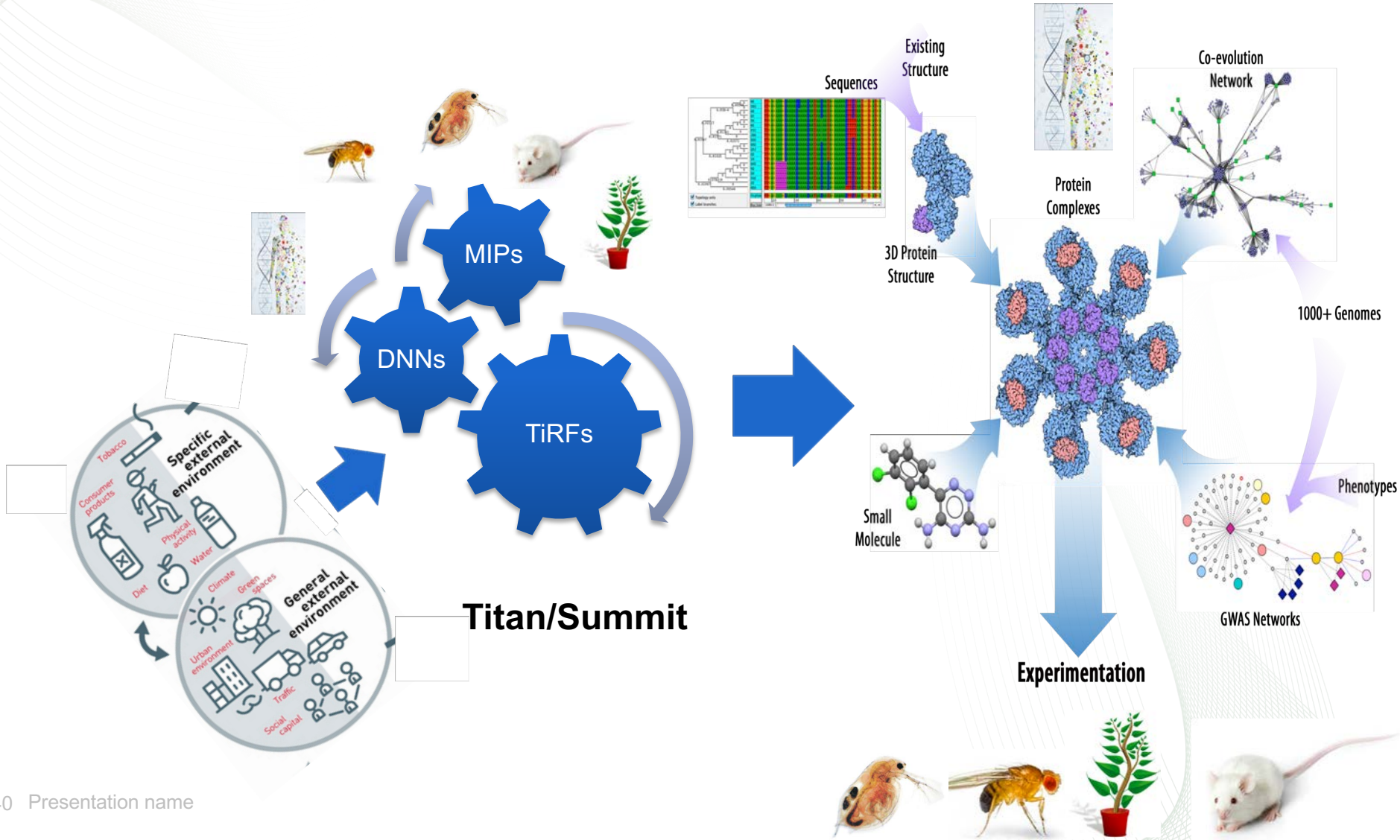


HELIX BUILDING THE EARLY-LIFE EXPOSOME

ISGlobal Barcelona Institute for Global Health

OAK RIDGE National Laboratory

High Order Interactions:
Exposome – Adverse Outcome Networks
Explainable AI: Machine and Deep Learning Integration



Acknowledgements

Collaborators

- Ben Brown
- Jerry Tuskan
- Steve DiFazio
- Wayne Joubert
- Amy Justice
- Edmon Begoli

Computational Infrastructure

- Oak Ridge Leadership Computing Facility (OLCF)
- Compute and Data Environment for Science (CADES)

Joint Genome Institute

Acknowledgements

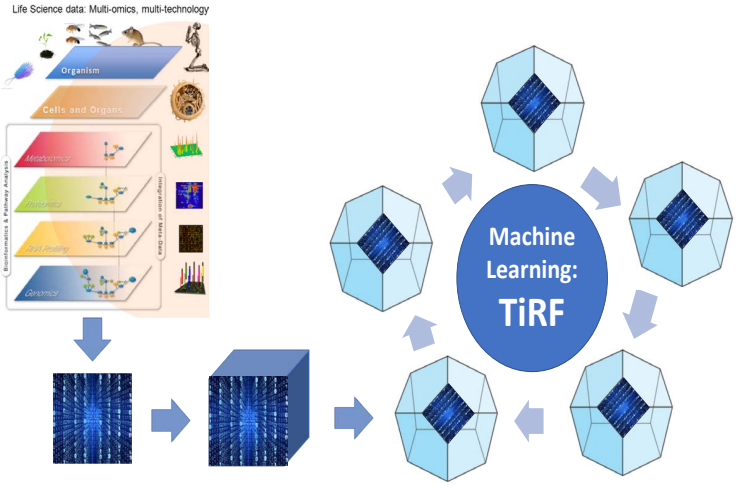
CBI
PMI
LDRD
VA
Oak Ridge Leadership Computing Facility (OLCF) at ORNL
Compute and Data Environment for Science (CADES) at ORNL
INCITE
Joint Genome Institute (JGI)

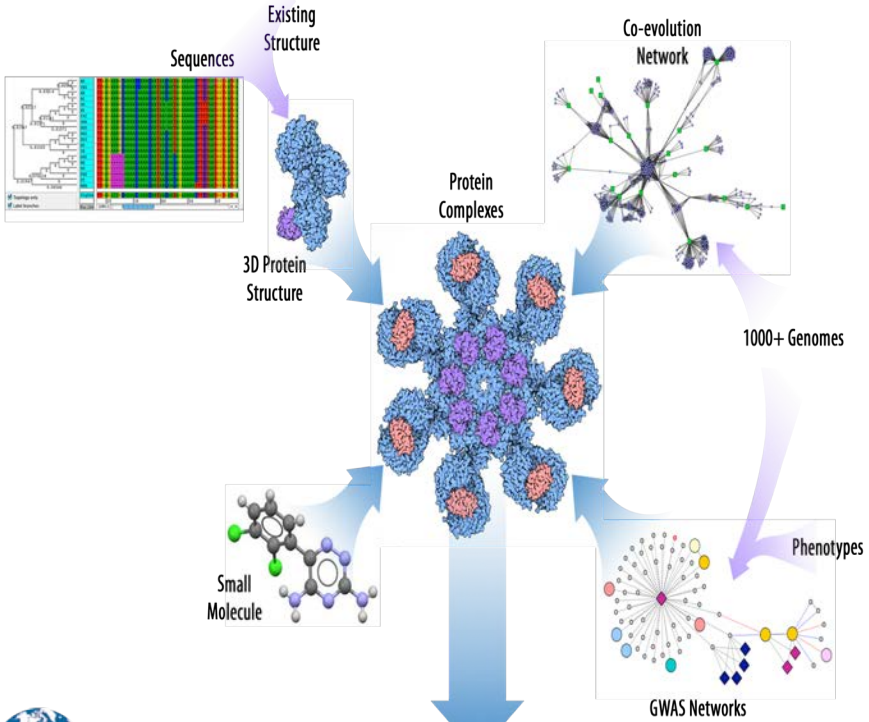
Oak Ridge National Laboratory (ORNL)
Bredesen Center for Interdisciplinary Research and Graduate Education, University of Tennessee, Knoxville

Acknowledgements

- JAIL Team effort
  - Debbie Weighill
  - Piet Jones
  - Carissa Bleker
  - Armin Geiger
  - Marek Piatek
  - Ben Garcia
  - Ashley Cliff
  - Jonathon Romero
  - David Kainer
  - Annie Fouche
  - Sandra Truong
  - Ryan McCormick
  - Priya Ranjan
  - Manesh Shah
  - Doug Hyatt
  - Blake Wiley
  - Jesse Marks
  - Ian Hodge
  - Annabel Large
  - Chris Ellis

Questions?





U.S. DEPARTMENT OF ENERGY
INCITE
LEADERSHIP COMPUTING

summit