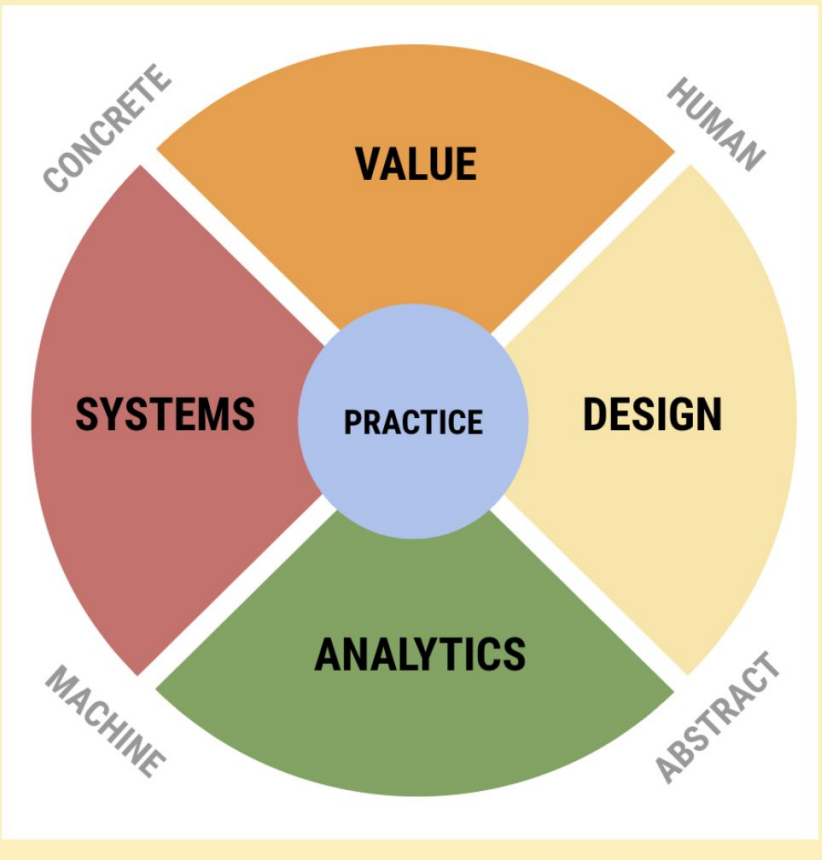
Data Science Meets Drug Discovery

Philip Bourne, University of Virginia
Zheng Zhao, University of Virginia
Lei Xie, The City University of New York

Outline

- One definition of data science
- A brief history of *in silico* early-stage drug discovery
- Bioinformatics methods for structure-based drug design
- Machine Learning/Deep Learning solutions to systems pharmacology
- Summary

One Definition of Data Science – The 4+1 Model



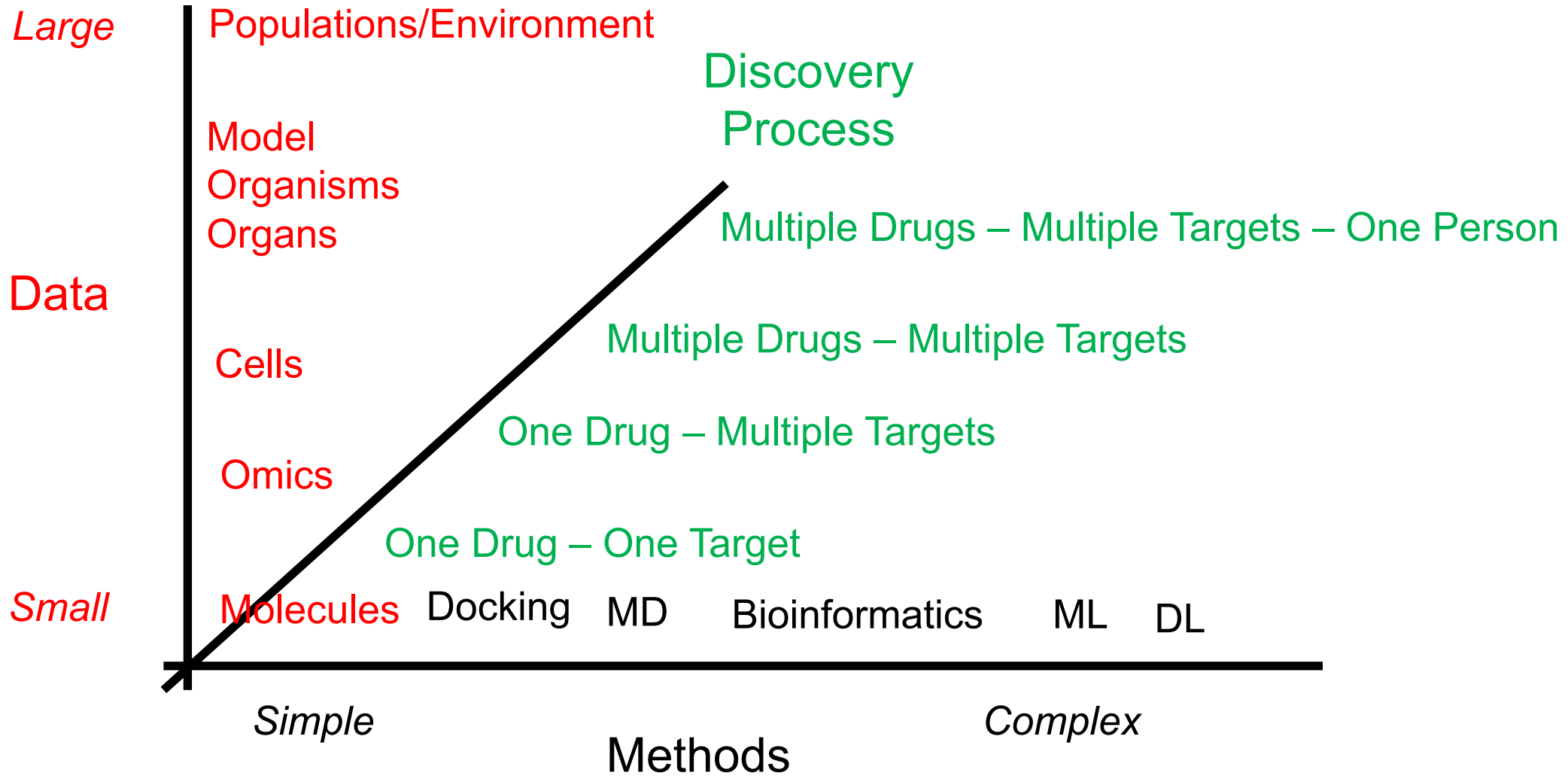
The 4+1 Model of Data Science (CC BY-SA 4.0)

- **Value** – ensuring societal benefit
- **Design** - Communication of the value of data
- **Systems** – the means to communicate and convey benefit
- **Analytics** – models and methods
- **Practice** – where everything happens – for today that is early-stage drug discovery

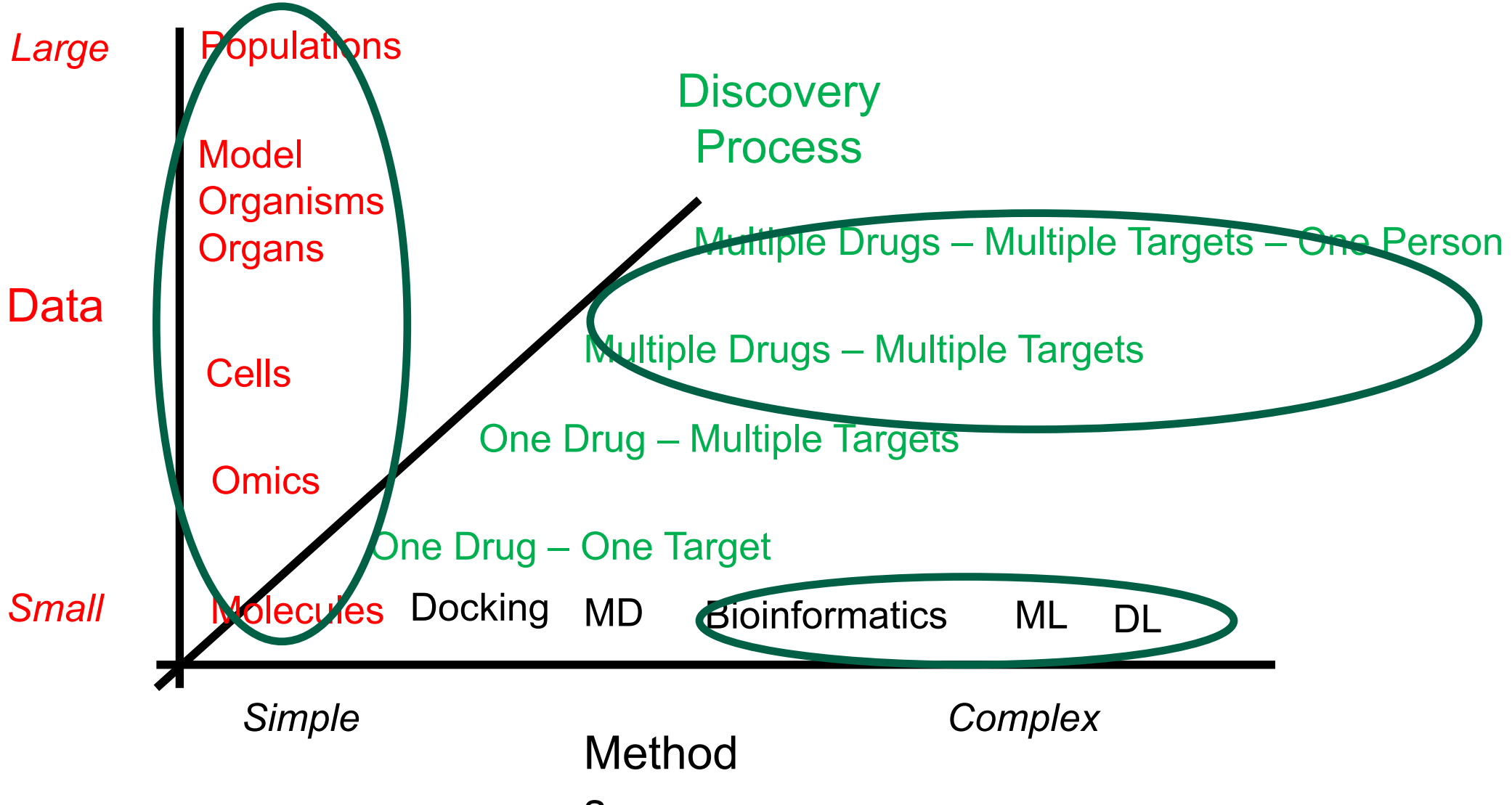
[Adapted from Raf Alvarado]

Bourne PE (2021) Is "bioinformatics" dead? PLoS Biol 19(3): e3001165.

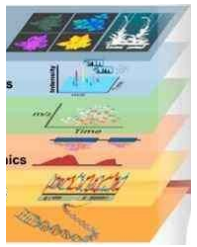
A Snapshot of *In Silico* Drug Discovery
Today's Discussion



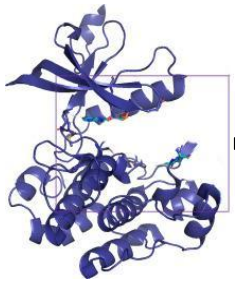
A Snapshot of *In Silico* Drug Discovery
Today's Discussion

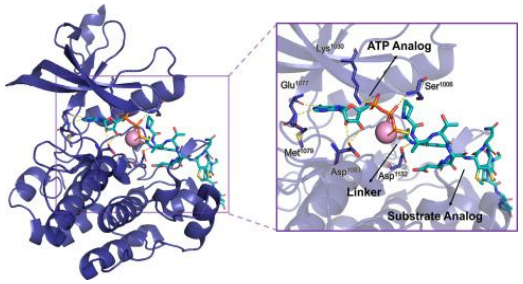


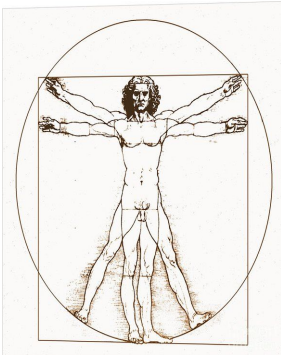
One-drug-one-target Small Molecule Drug Discovery Process



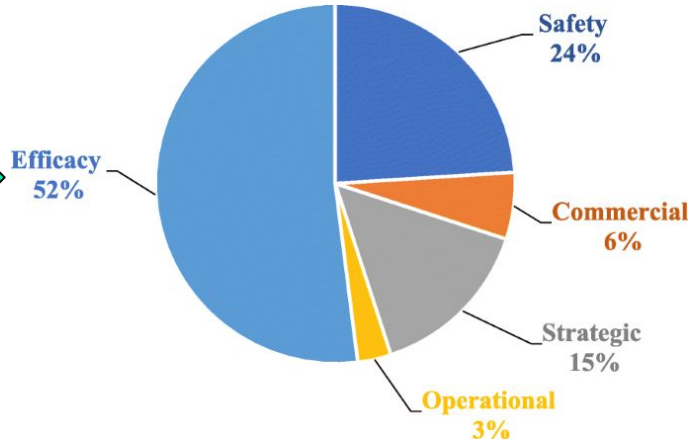
Adapted from
http://www.biomech.ulg.ac.be/project/multi-omics/







Kiriiri et al. *Future J Pharm Sci.* (2020)



Omics data analysis

Druggable target identification

Lead compound identification and optimization

Pre-clinical testing

Clinical trial

>90% Failure Rate

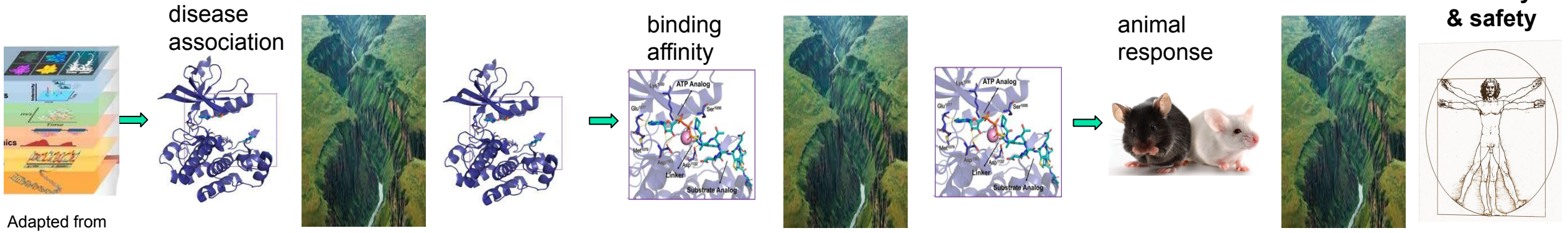
Fundamental Challenge of Conventional Drug Discovery in Combating Complex Diseases

Each step in the drug discovery process is optimized for a proxy objective but not a clinical endpoint

Disease gene may not be right drug target

Genome-wide off-target bindings and systems effects are ignored

animal-human extrapolation remains challenging



Adapted from http://www.biomech.ulg.ac.be/project/multi-omics/

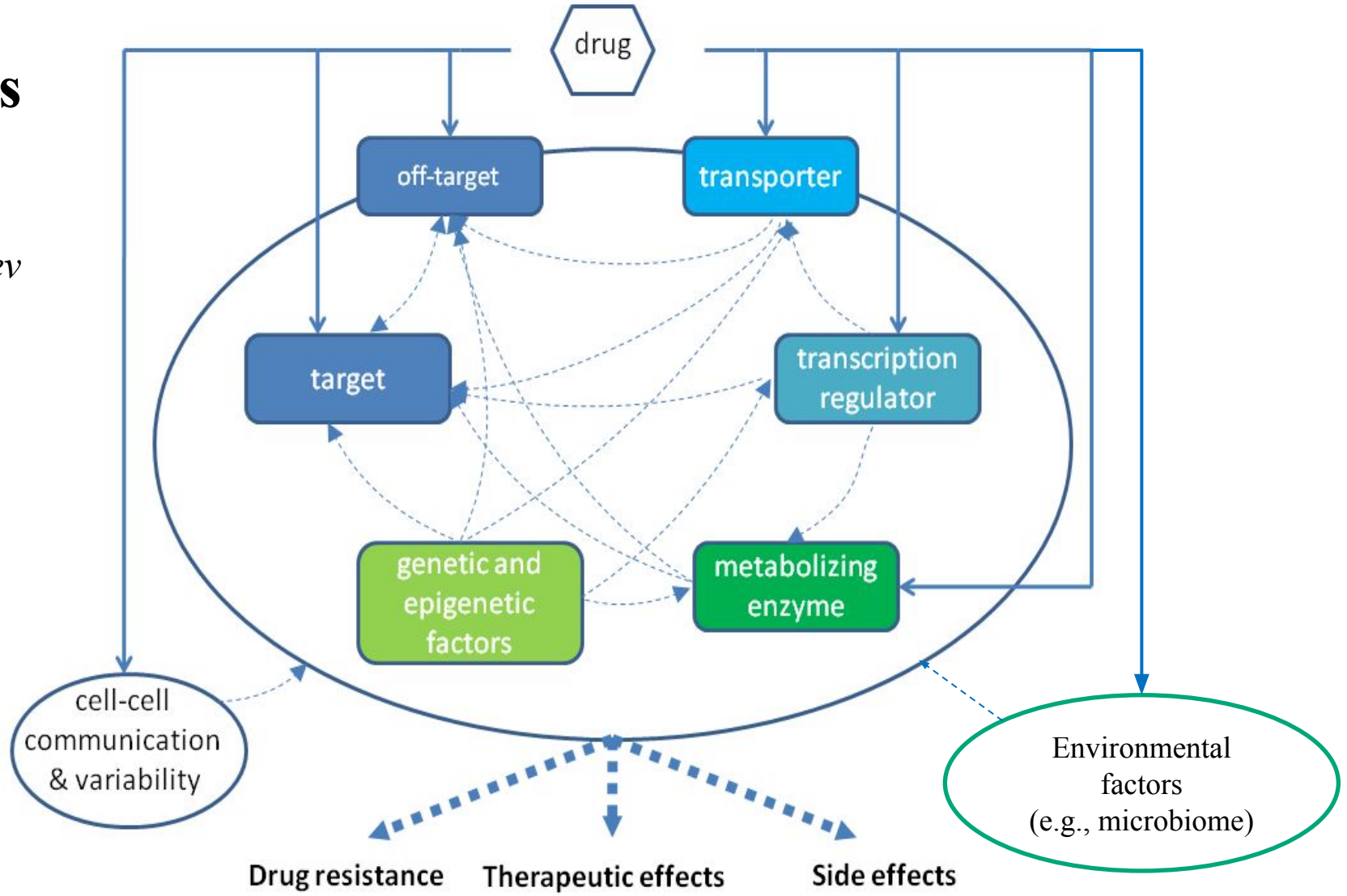
Power of Artificial Intelligence (Deep Learning) has not been fully utilized

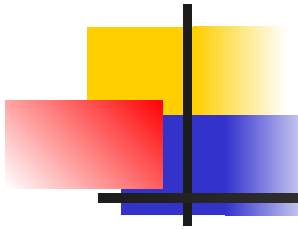
Bender & Cortes-Ciriano, *Drug Discov Today* (2021)

Discovery Process: Multi-Scale Modeling Drug Actions i the Human Body

Pleiotropic systems effect of drugs

Xie & Bourne (2012) *Annu Rev Pharmacol Toxicol*,



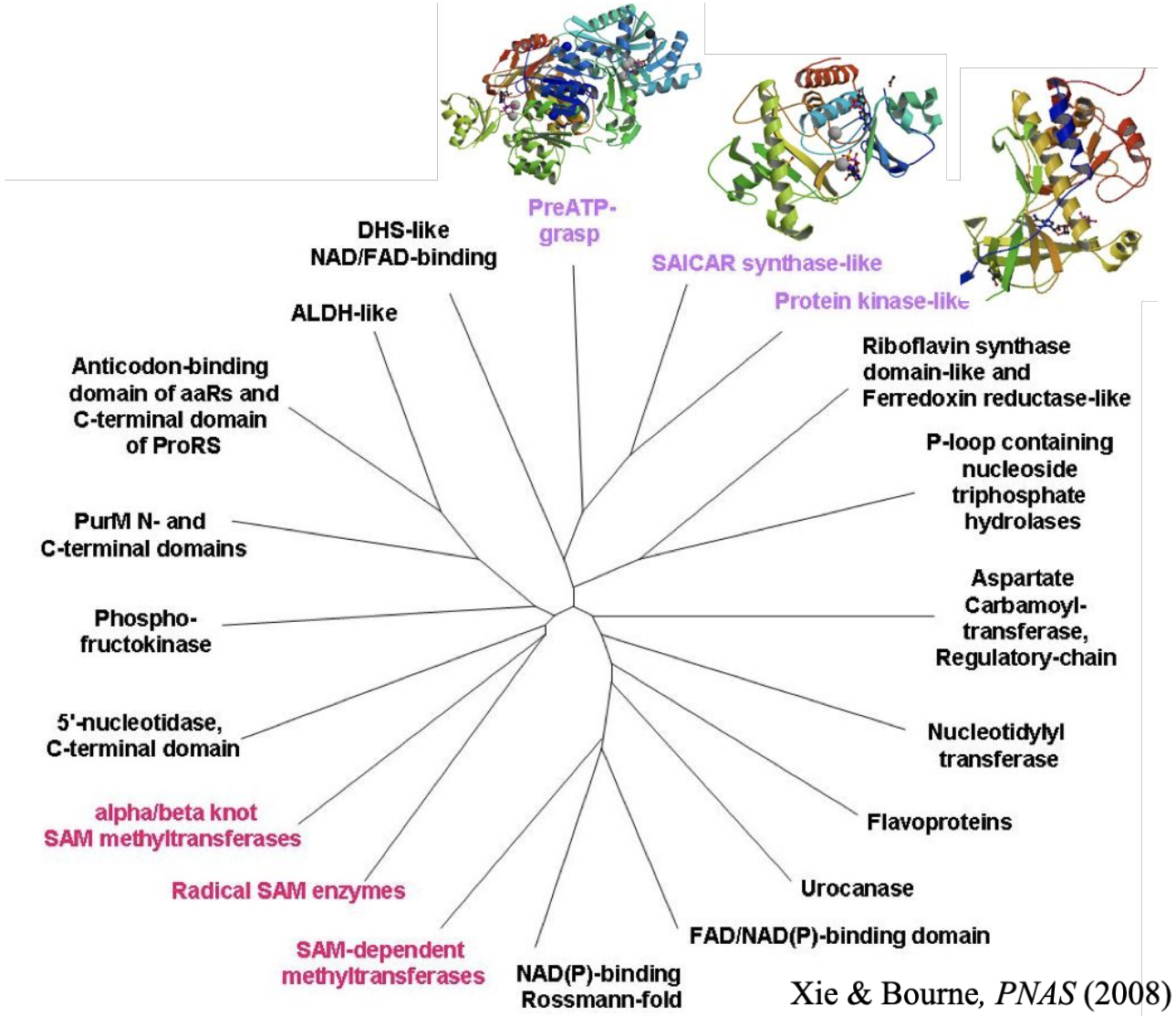


Outline

- One definition of data science
- A brief history of *in silico* early-stage drug discovery
- Bioinformatics methods for structure-based drug design
- Machine Learning/Deep Learning solutions to systems pharmacology
- Summary

Bioinformatics Methods for Drug Design and Discovery

- Utilize structural genomics data
- Explore evolutionary and functional linkage (ligand binding site similarity) across fold space
- Successfully applied to drug repurposing, side effect prediction, and polypharmacology



Xie & Bourne, *PNAS* (2008)

Bioinformatics Methods for Drug Design and Discovery

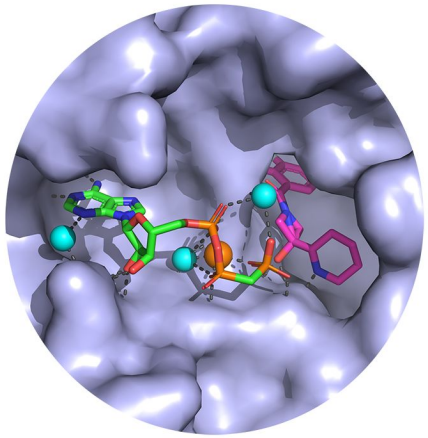
Early-Stage Drug Discovery Patterns:

Categories	Principles
Ligand-based	Chemical Similarity; Pharmacophore Similarity
Structure-based	Binding site (Pocket) Similarity
Protein-ligand-based	Pharmacophore Similarity; Interaction Fingerprint Similarity

Bender, et al; *Drug Discov. Today*; 2021

Bioinformatics Methods for Protein-ligand interaction -based Drug Design

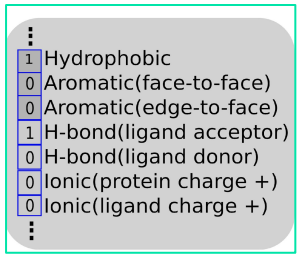
IFPs: Protein-Ligand Interaction Fingerprints



Given a ligand-binding complex

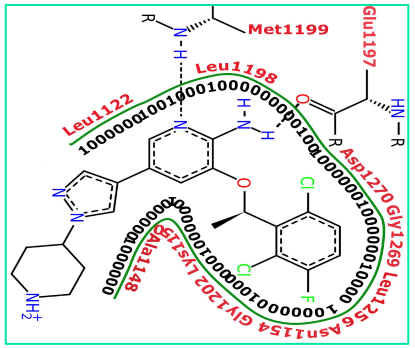
Encoding IFPs

Every residue – ligand interaction encoded into a 7-bit substring



Using pre-defined geometric rules

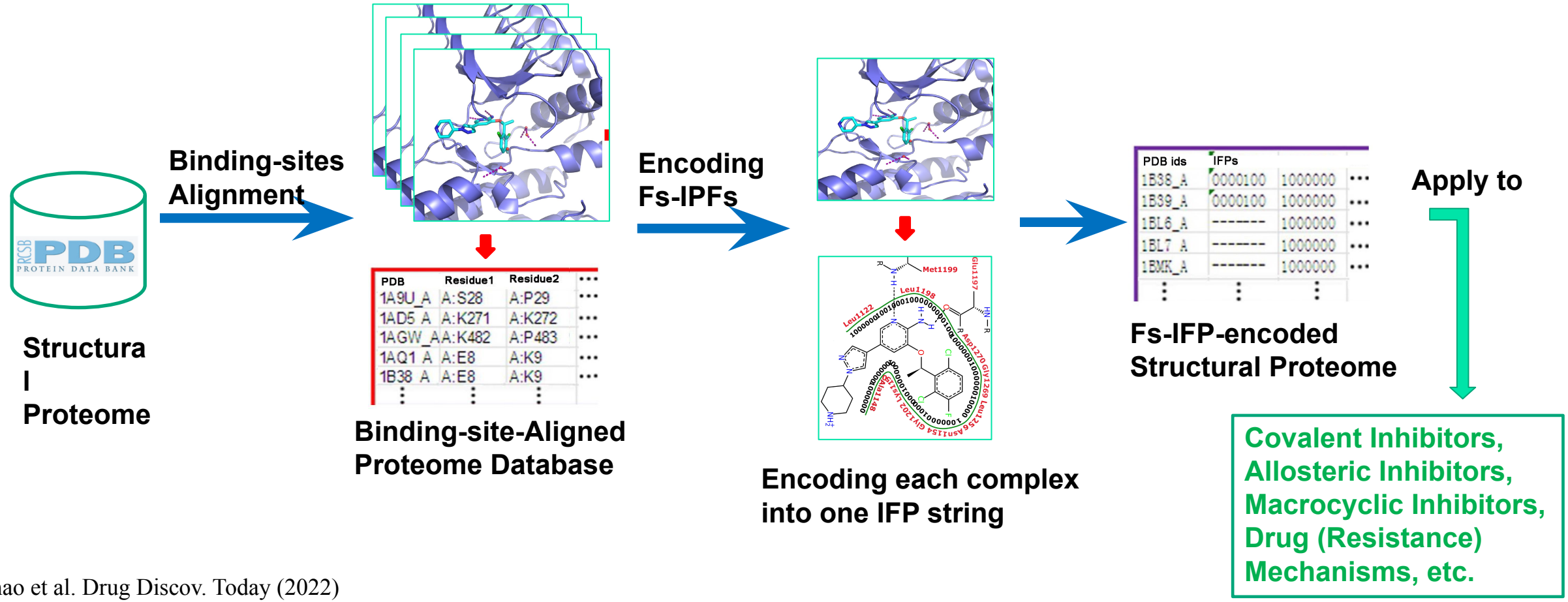
Interactions	Rules*
H-bond	Distance (Donor-Accepter) ≤ 3.5 Å && Tolerance Angle (Donor-H-Acceptor ∈[-π/3, π/3])
Ionic	Distance (Anion-Cation) ≤ 4.0 Å
Aromatic (face to face)	Distance (two aromatic ring centers) ≤ 4.0 Å && Tolerance angle (face to edge) ∈[-π/6, π/6])
Aromatic (face to edge)	Distance (two aromatic ring centers) ≤ 4.0 Å && Tolerance angle (face to edge) ∈[π/6, 5π/6])
Hydrophobic	Distance (Donor-Accepter) ≤ 4.5 Å



One 1D string representing the protein-ligand interactions

Deng et al. J. Med. Chem (2004)
Marcou et al. J. Chem. Inf. Model. (2007)

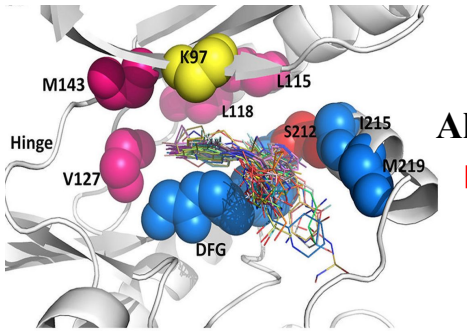
Fs-IFP: Structural Proteome-scale Protein-Ligand Interaction Fingerprint Method



Zhao et al. Drug Discov. Today (2022)
Zhao et al. J. Chem. Inf. Model. (2023)
Zhao et al. ACS Pharmacol. Transl. Sci (2023)

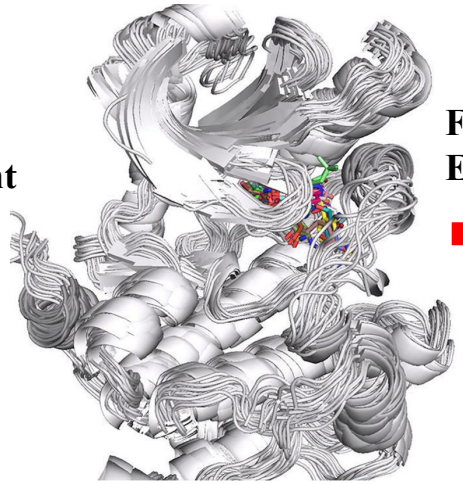
Fs-IFPs: Application to Allosteric Kinase Inhibitors

Advantage: Allosteric Kinase Inhibitor is attractive due to high, unique selectivity.
Background: 76 Kinase Drugs approved; 4 are allosteric type inhibitors targeting MEK.



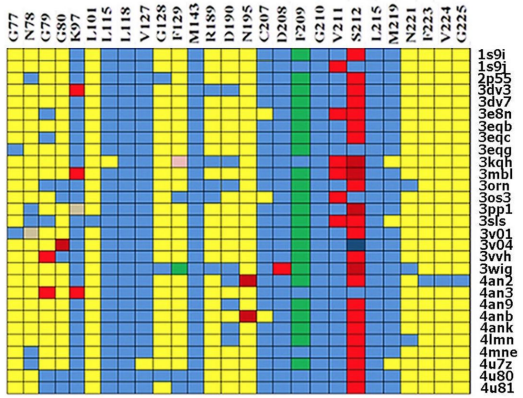
MEK-ligand Complexes from PDB and MD simulations as Structural Dataset

Alignment



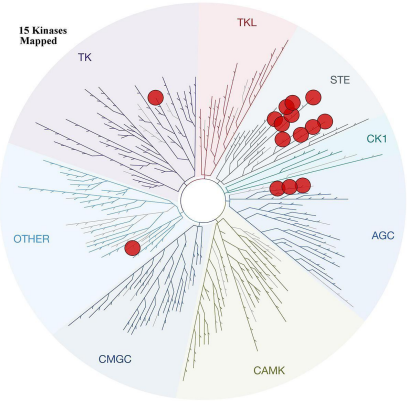
Binding sites aligned Dataset

Fs-IFP Encoding



Extracting MEK-inhibitor Interaction Characteristics

Predicting

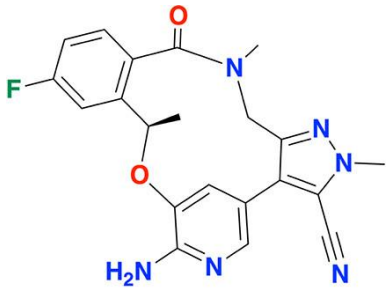


Top 15 potential kinase targets for allosteric inhibitor design

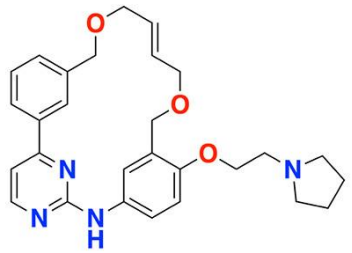
Zhao et al. *PLoS One* 12 (6), e0179936, 2017

Fs-IFPs: Application to Macrocyclic Kinase Inhibitors (MKIs)

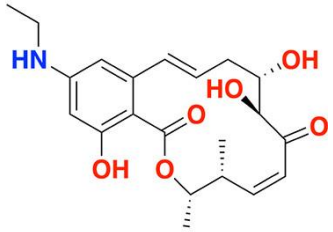
Background: 76 Kinase Drugs approved; Challenge: Drug Resistance
Trending: MKIs attracting much attention;
Progress: 2 Approved and 7 in clinical trials (Below)



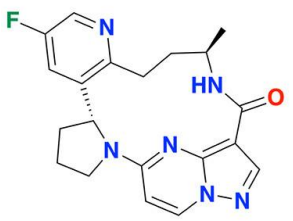
Lorlatinib, ALK/ROS, Approved in 2018



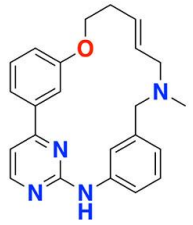
Pacritinb, JAK2/FLT3, Approved in 2022



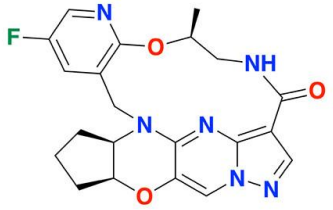
E6201, MEK1/FLT3, Phase II



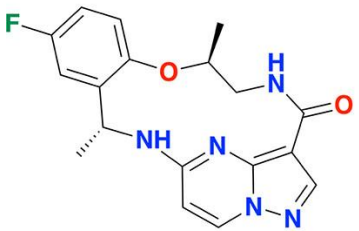
Selitrectinib, TRK, Phase I/II



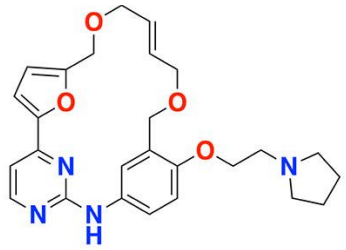
Zotiraciclib, CDK2/JAK2/ FLT3, Phase I/II



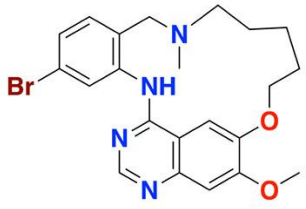
TPX-0046, RET/SRC, Phase I/II



Repotrectinib, TRK/ALK/ ROS1, Phase I/II



SB1578, JAK2, Phase I



JNJ-26483327, EGFR/RET/ VEGFR3/Her4/Src, Phase 1

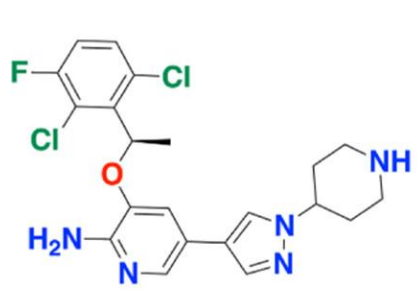
8 out of 9 are designed by rational drug design strategies from the acyclic starting points

Georg et al. **Trends Pharmacol. Sci (2022)**

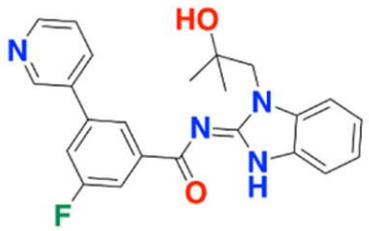
Zhao et al. ACS Pharmacol. Transl. Sci (2023)

Fs-IFPs: Application to Macrocyclic Kinase Inhibitors (MKIs)

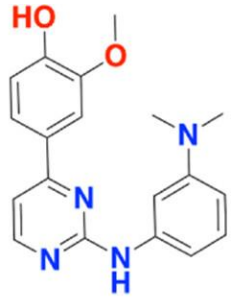
- **How to design MKI?**



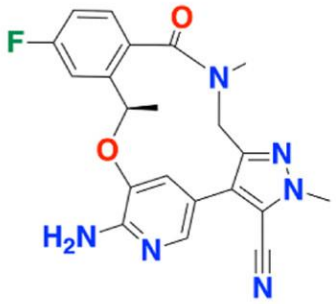
Crizotinib



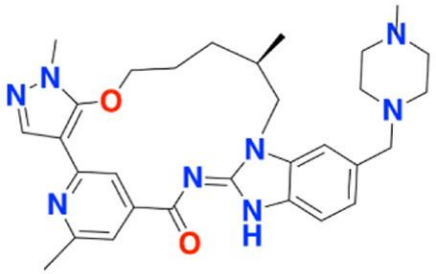
Ligand-1



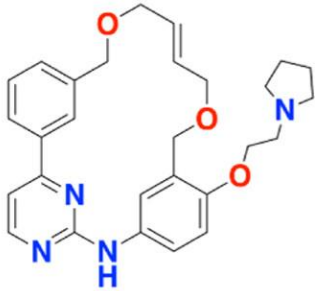
Compound-1



Lorlatinib



BI-4020



Pacritinb

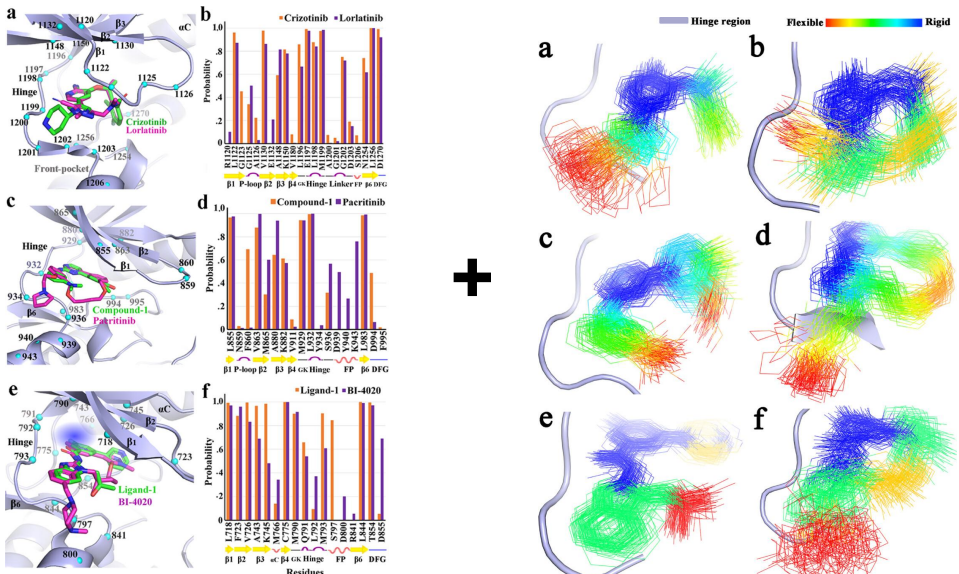
Acyclic Inhibitors

MKIs

Zhao et al. ACS Pharmacol. Transl. Sci (2023)

Fs-IFPs: Apply to Macrocyclic Kinase Inhibitors (MKIs)

Comparing Binding modes of MKIs and their acyclic counterparts



+

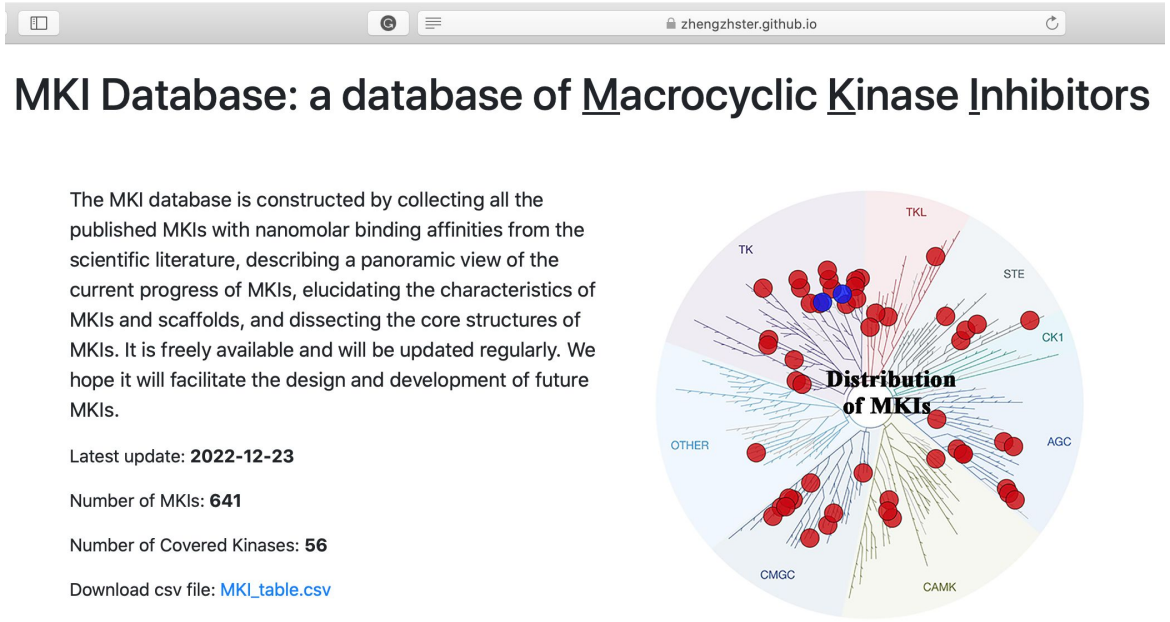
Scaffolds with the same binding modes before and after cyclization

Rigid Scaffold

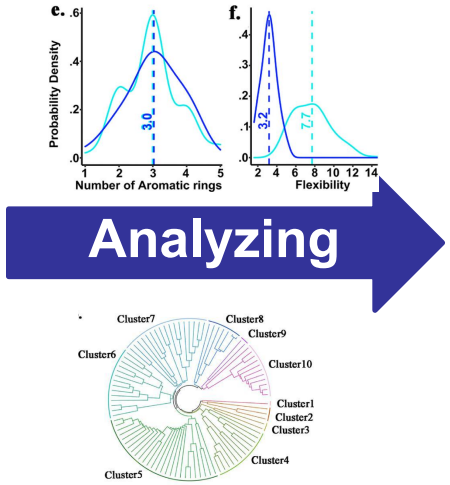
MD trajectories analysis using Fs-IFPs for MKIs and their corresponding acyclic counterparts

Fs-IFPs: Application to Macrocyclic Kinase Inhibitors (MKIs)

- **What kind of rigid scaffolds are promising?**



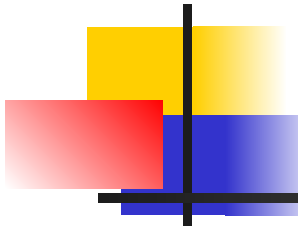
MKI database



Rigid Scaffolds with 3 aromatic rings directly connected to one another

https://zhengzhster.github.io/MKIs/

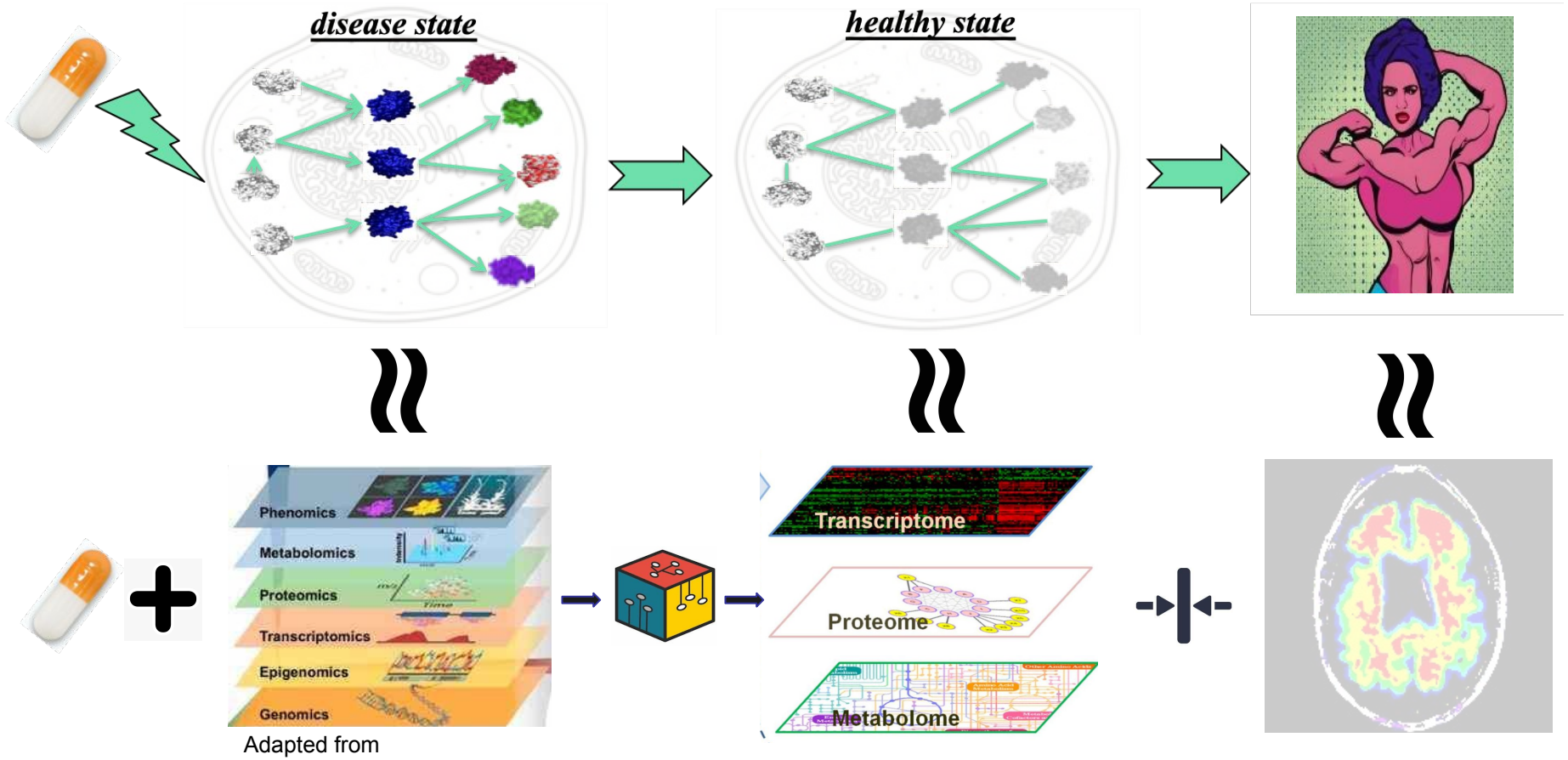
Zhao et al. ACS Pharmacol. Transl. Sci (2023)



Outline

- One definition of data science
- A brief history of *in silico* early-stage drug discovery
- Bioinformatics methods for structure-based drug design
- Machine Learning/Deep Learning solutions to systems pharmacology
- Summary

Discovery Process: A Systems Pharmacology Paradigm



Adapted from
http://www.biomech.ulg.ac.be/project/multi-omics/

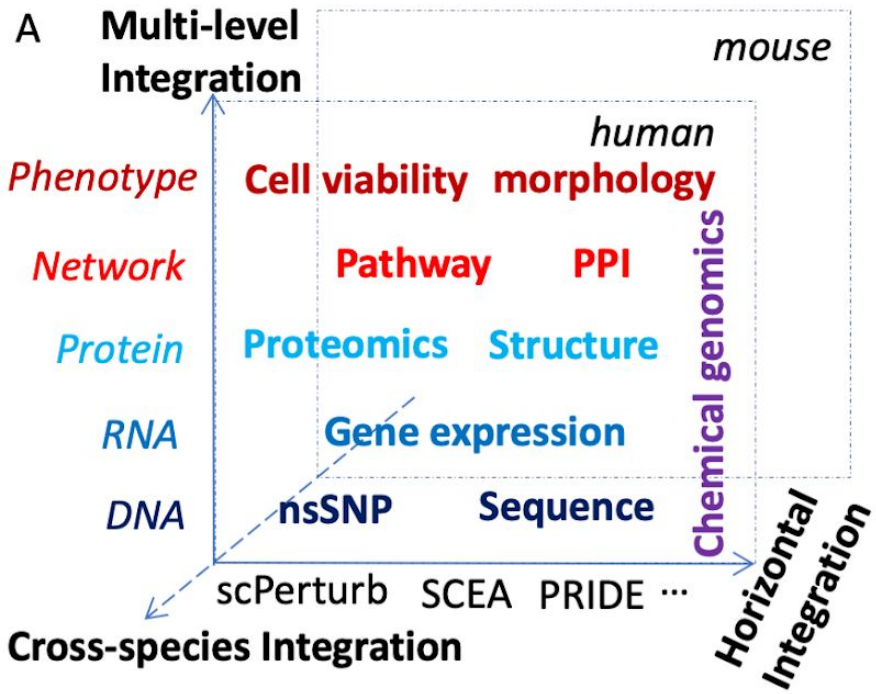
Patient Omics profile

Chemical perturbed
Omics profile

Clinical
Biomarkers

Sun et al. (2016) *Adv. Genetics*

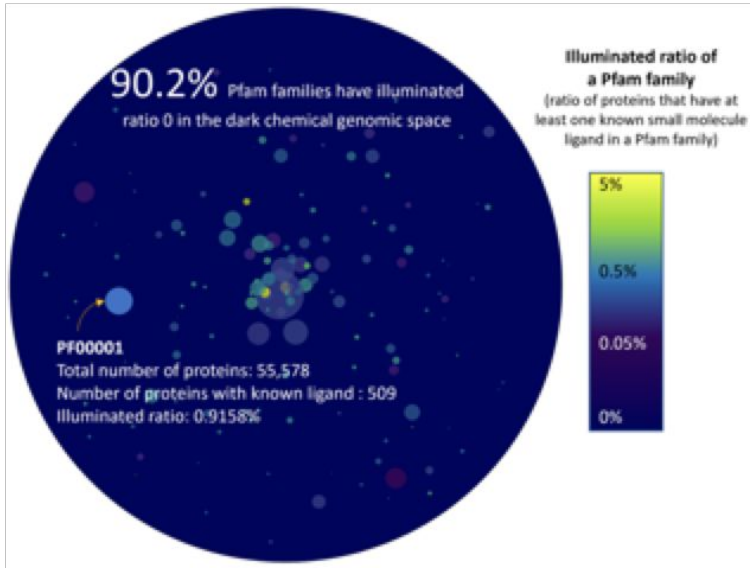
Challenges in Systems Pharmacology: Multi-omics data integration and knowledge discovery



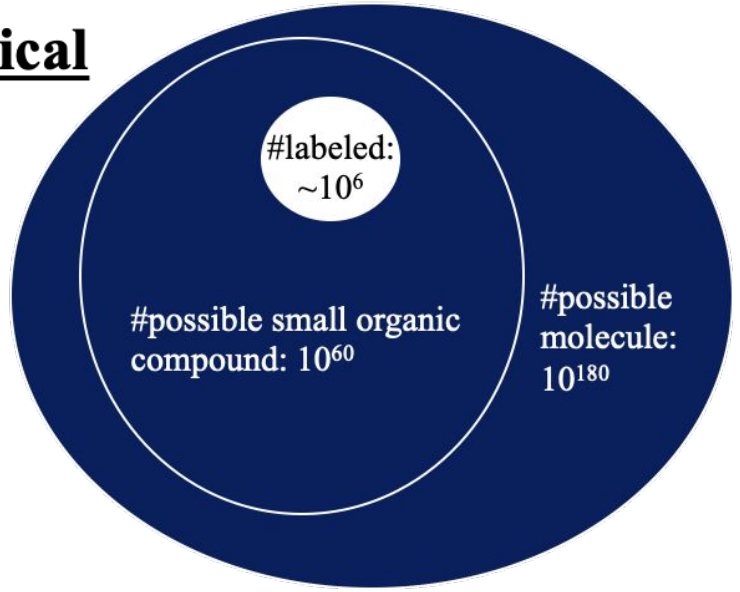
- Noisy, high-dimensional, and sparse
- Technical (batch effect) and biological confounding factors (age, gender etc.)
- Complexity of molecular interplays across biological layers
- Common and unique features across species
- Correlation vs causality

Challenges in Systems Pharmacology: Extensive Dark Chemical and Functional Genomics Space

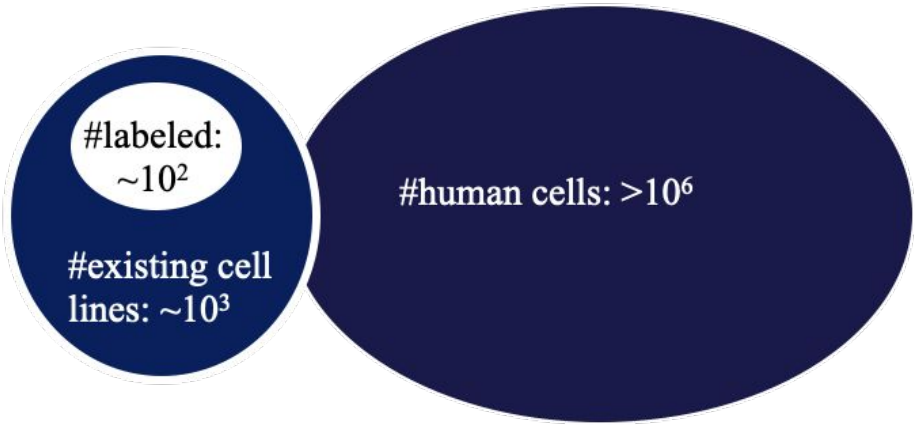
Protein



chemical



Perturbed gene expressions

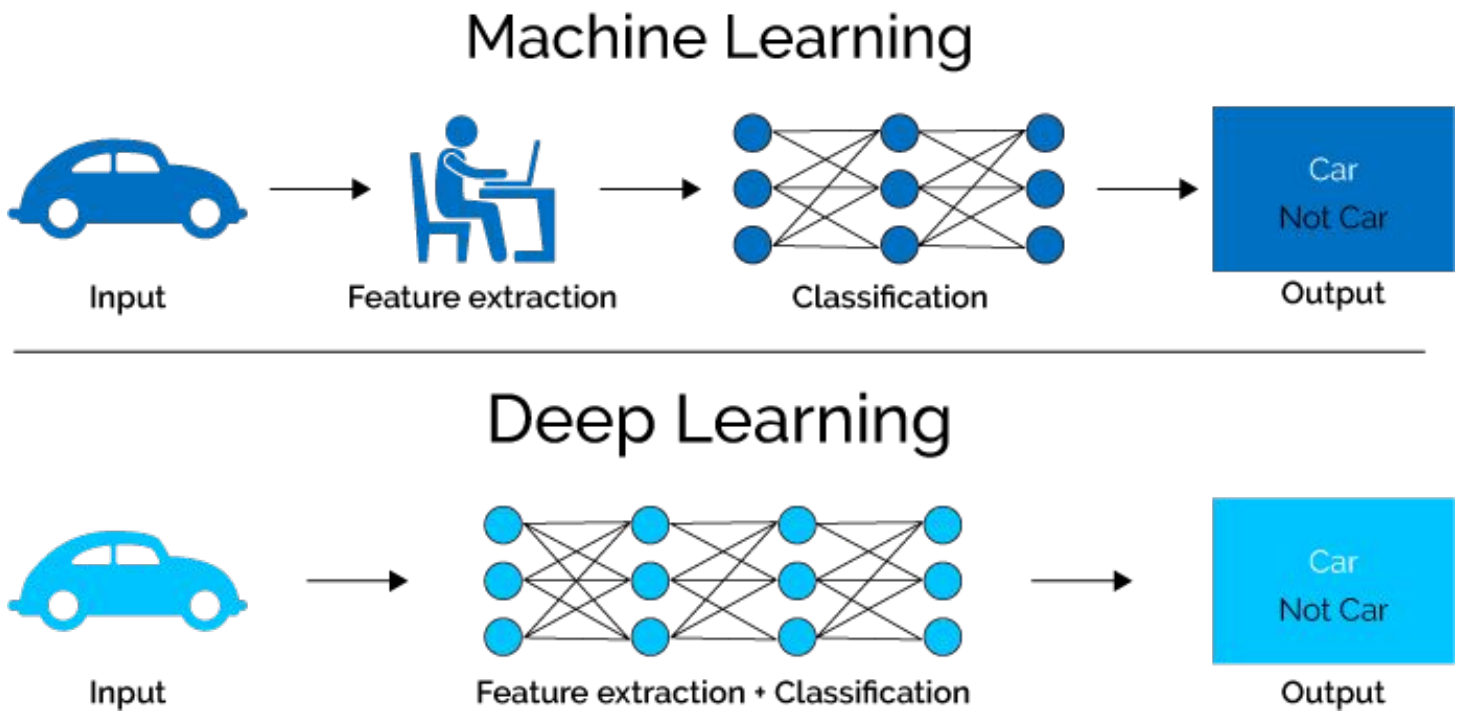


Machine Learning/Deep Learning is an Indispensable Tool

- To overcome the extensive dark genome/proteome:
  - It is unethical and infeasible to screen compounds in humans.
  - Complexity and high dimensionality of omics data are beyond human comprehension.
  - Understudied biology beyond the reach of experimental techniques.
- Hence: only a computational approach can overcome this data gap

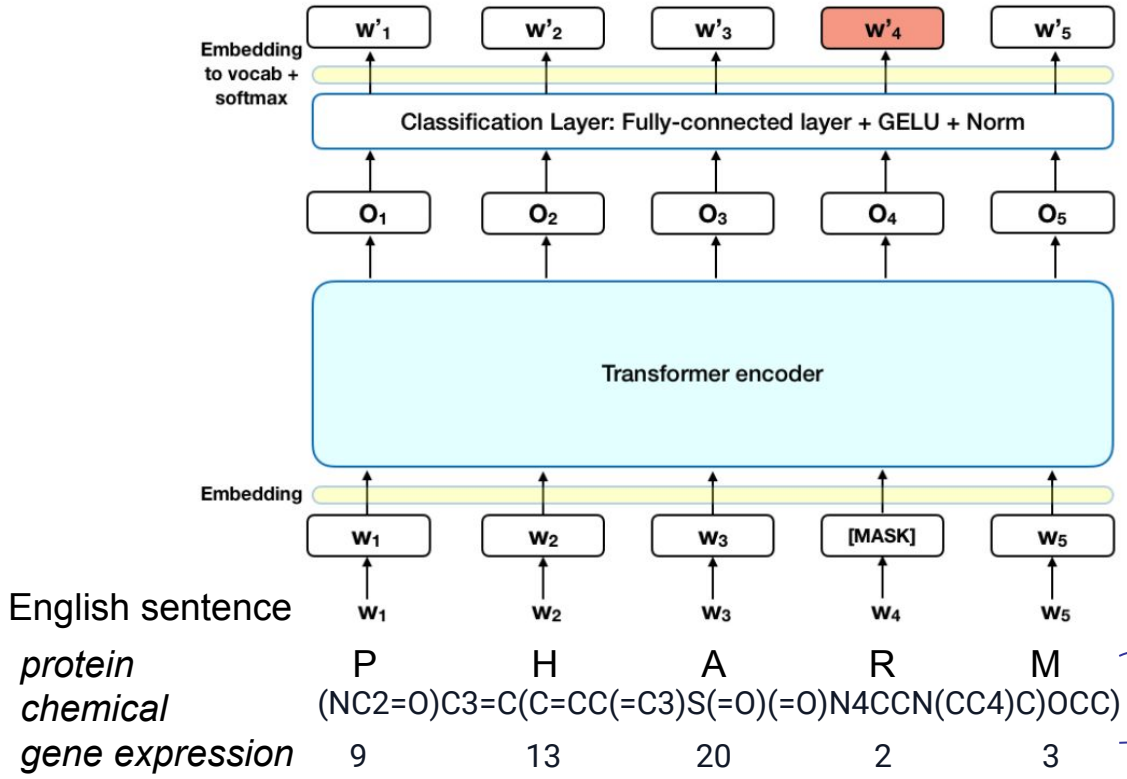
Filling the Data Gap: Deep Learning Advantages

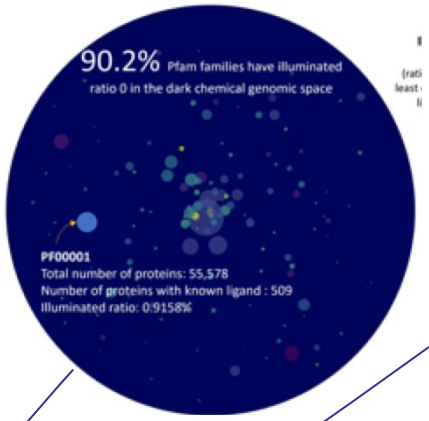
- Data-driven feature extraction
- Integration of incoherent diverse data (both labeled and unlabeled)
- Explicitly model multi-level organization of biological systems

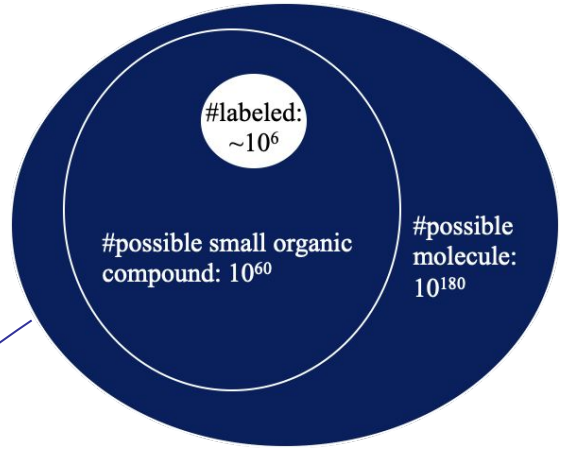


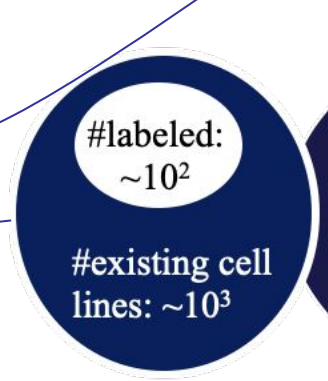
adapted from
https://levity.ai/blog/difference-machine-learning-deep-learning

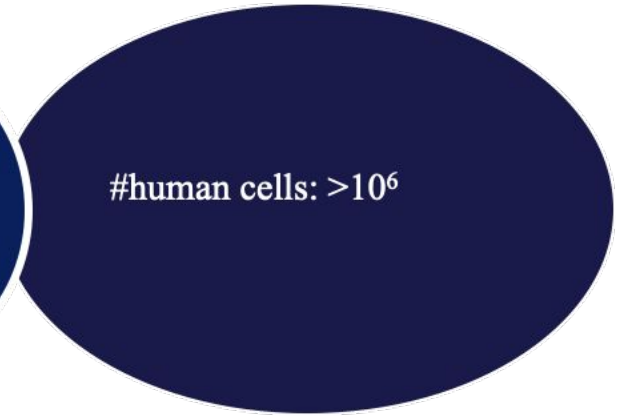
Filling the Data Gap: Self-Supervised Learning





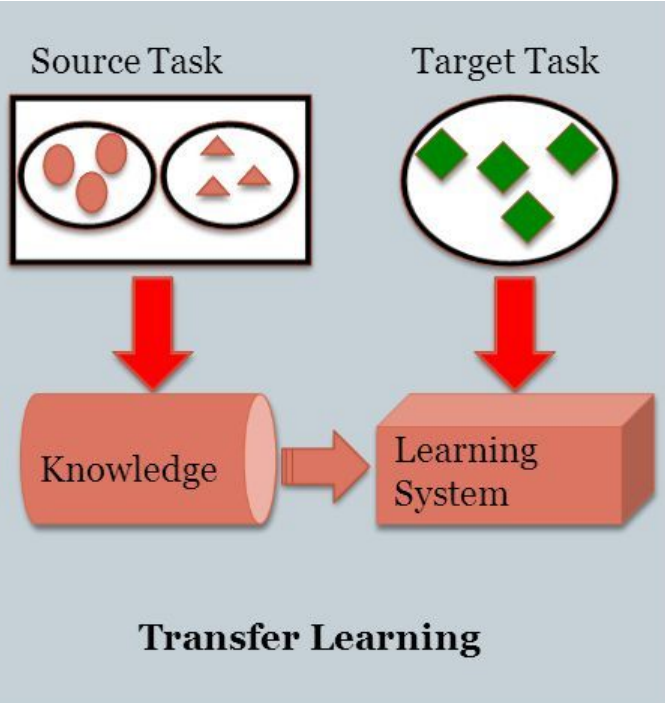


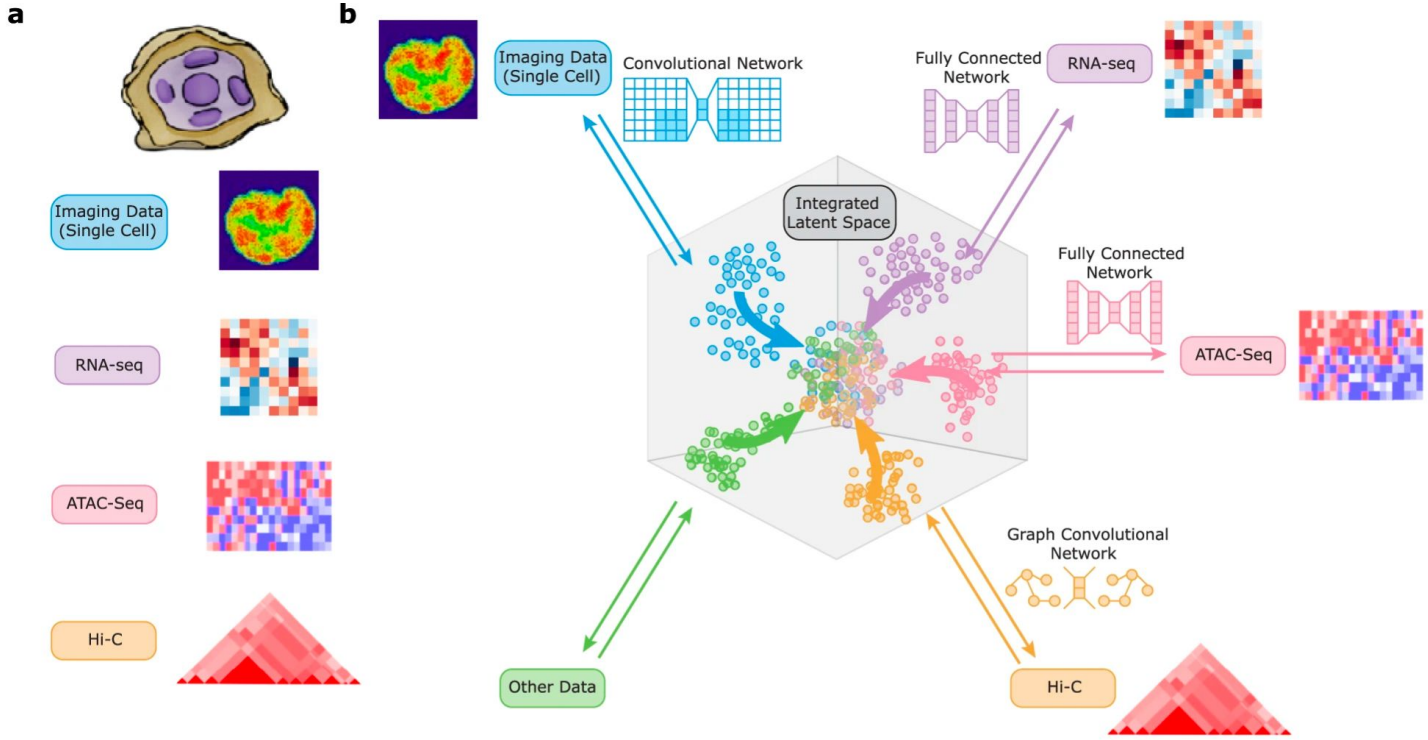




- Utilizing a large amount of unlabeled data
- Facilitating the exploration of dark chemical and biological space

Filling the Data Gap: Transfer Learning & Domain Translation





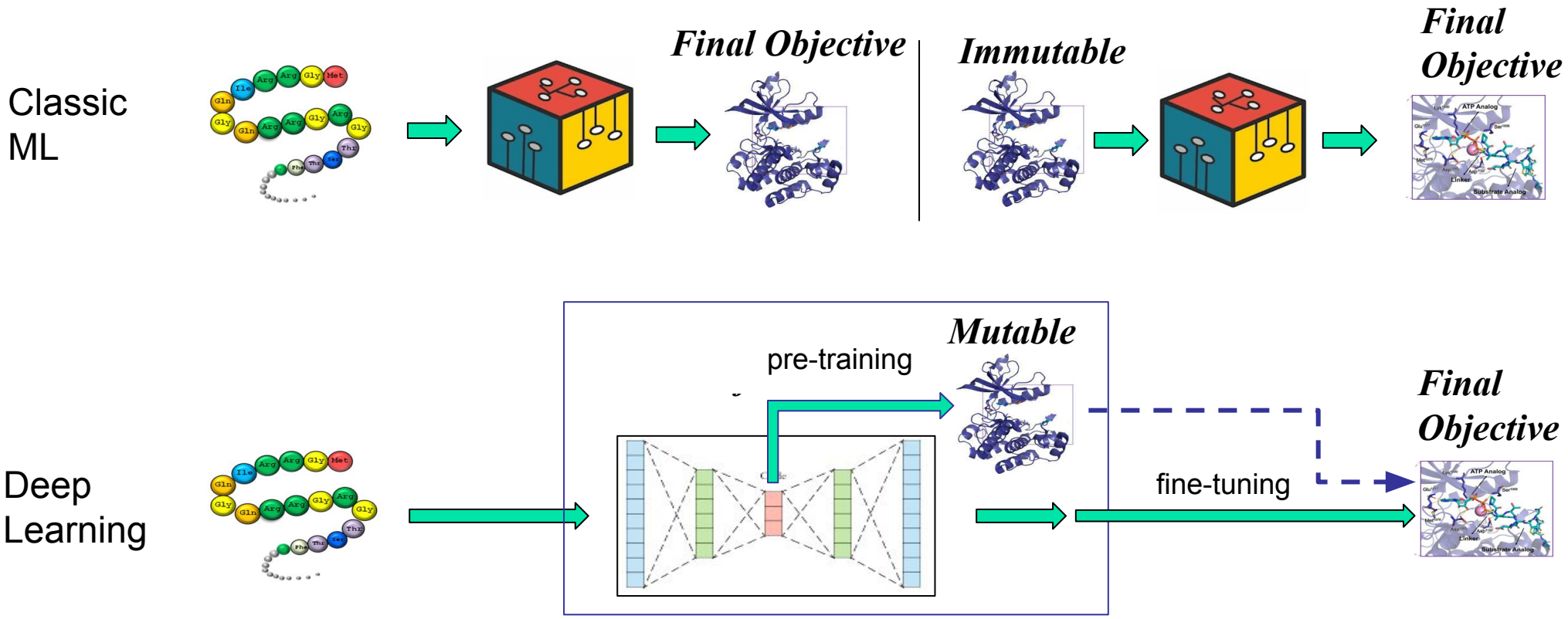
Yang et al. (2021) *Nature Comm.*

- Translating data modality (e.g., source = cell line omics; target = patient image biomarker)
- Removing biological and technological confounding factors across conditions and species

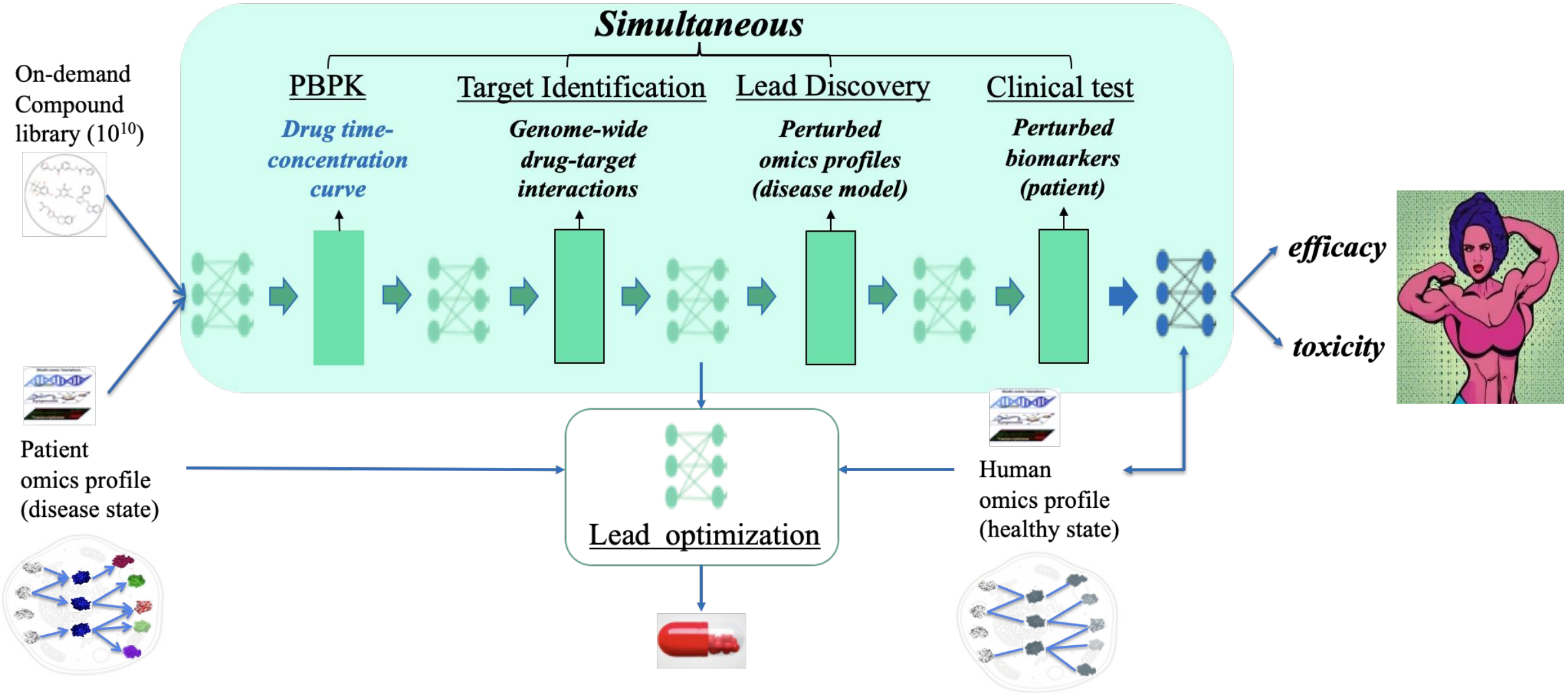
Filling the Data Gap: End-to-end Differentiable Biology

AlQuraisi et al., *Nat. Methods* (2021)

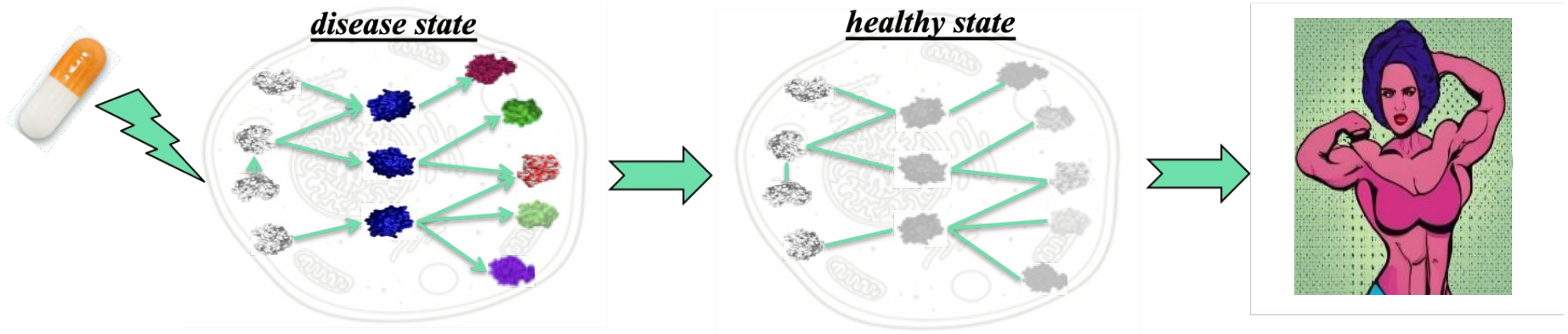
- Protein sequence (unknown structure)->3D protein-ligand binding pose



End-to-end Differentiable Systems Pharmacology for Precision Drug Discovery



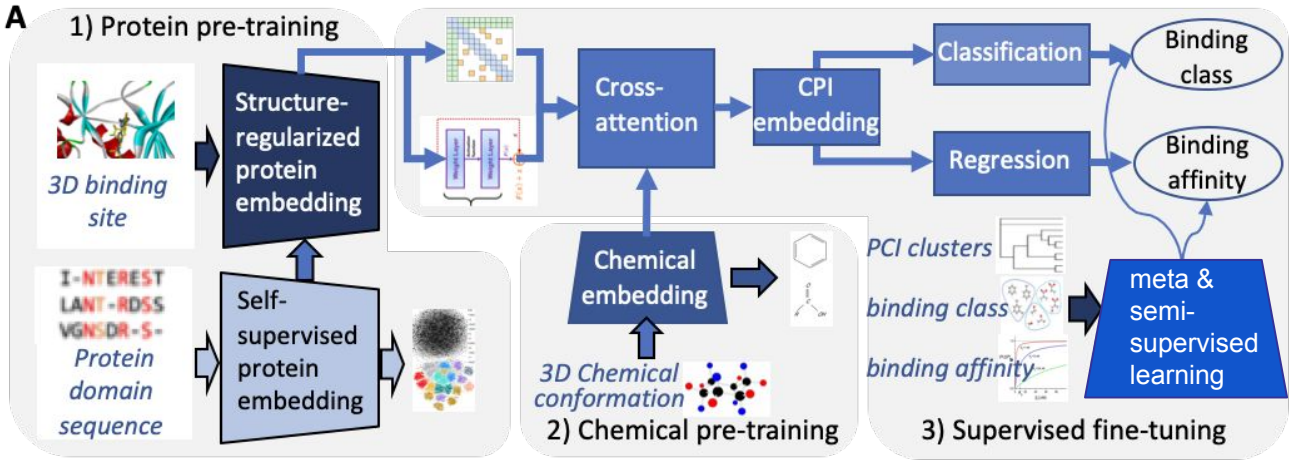
What Can Deep Learning Do for Systems Pharmacology-Oriented Drug Discovery ?



Key Questions

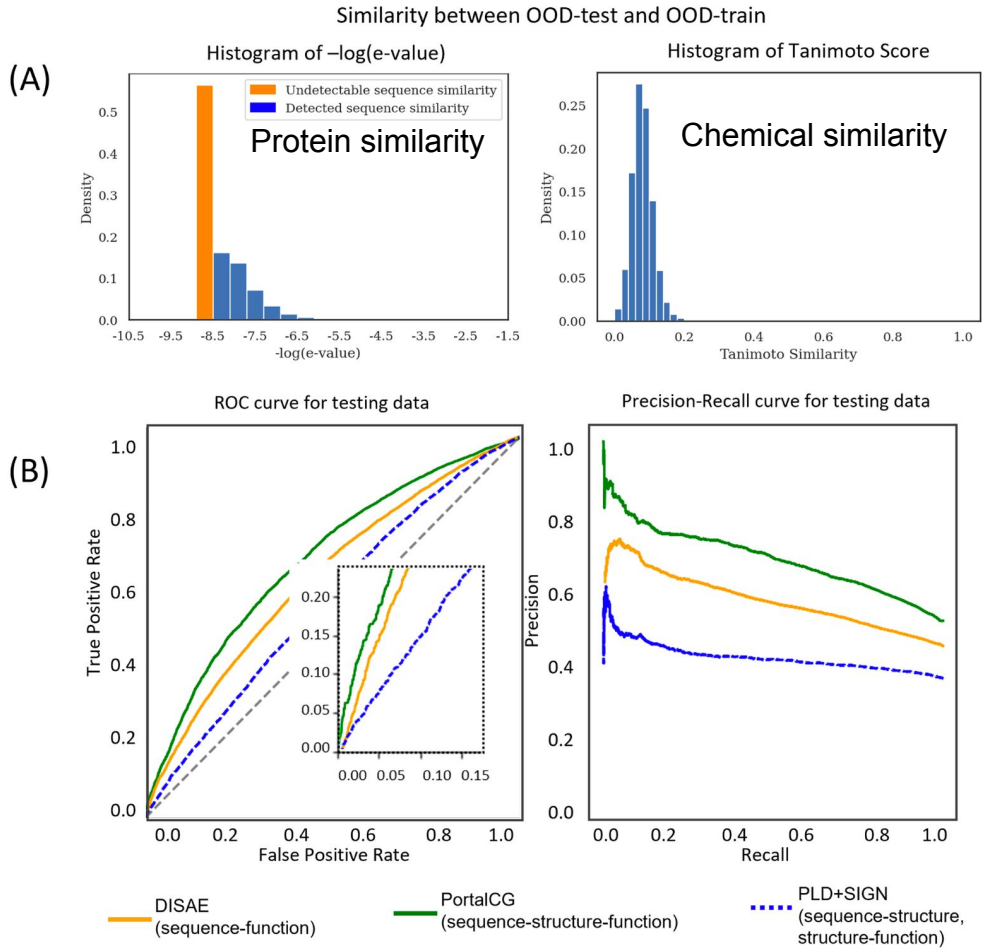
- What are the genome-wide target(s) and off-target(s)?
- How does drug modulate cellular state characterized by omics?
- How to translate drug activities in a disease model to clinical human response?

Predicting Genome-Wide Protein-Chemical Interactions for Dark Proteins



Zhang et al. *IJCAI*, 2020; Liu et al. *Front Bioinfor*, 2021;
Cai et al. *JCIM*, 2021; Cai et al. *Bioinformatics*, 2022
Cai et al. *PLoS CB,* 2023; Zhang et al. *Sci. Rep*., 2023
Wu et al. under review, 2023

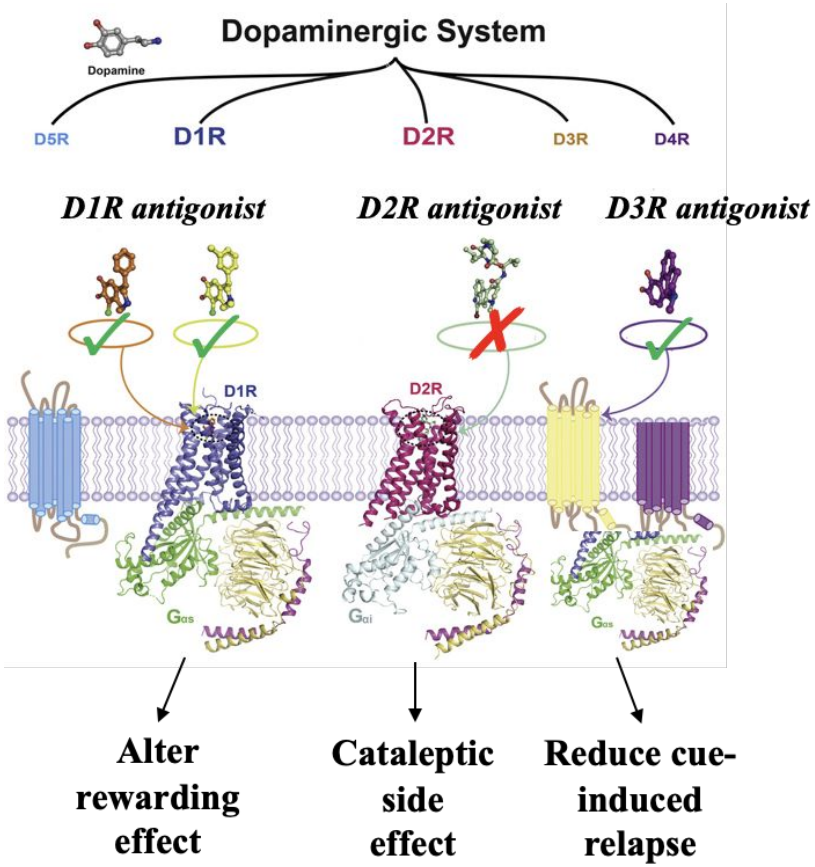
- PortalCG: An end-to-end sequence-structure-function neural network



Cai et al. (2023) *PLoS CB*

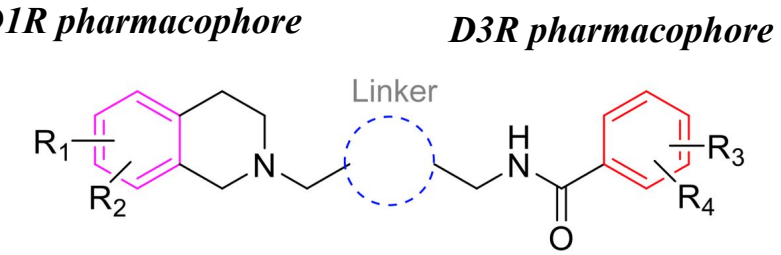
PortalCG: Application to Polypharmacology

- Selective dual-DRD1/3 antagonist of Opioid Use Disorder (OUD)

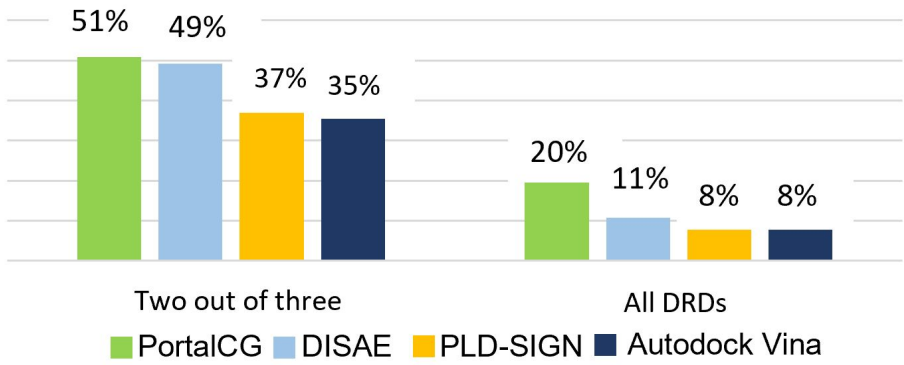


Zhuang et al. *Cell* (2021)

(A)

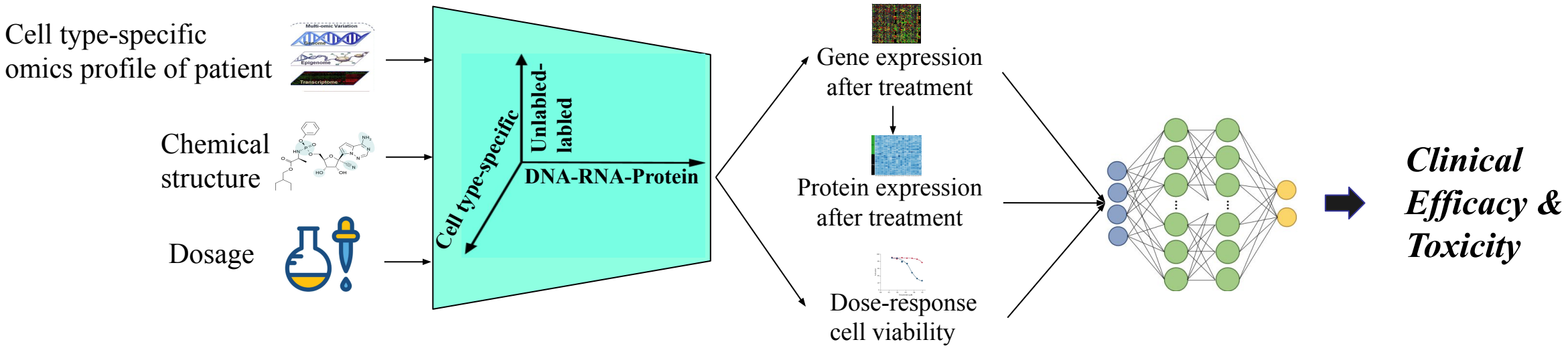


(B)



Cai et al. (2023) *PLoS CB*

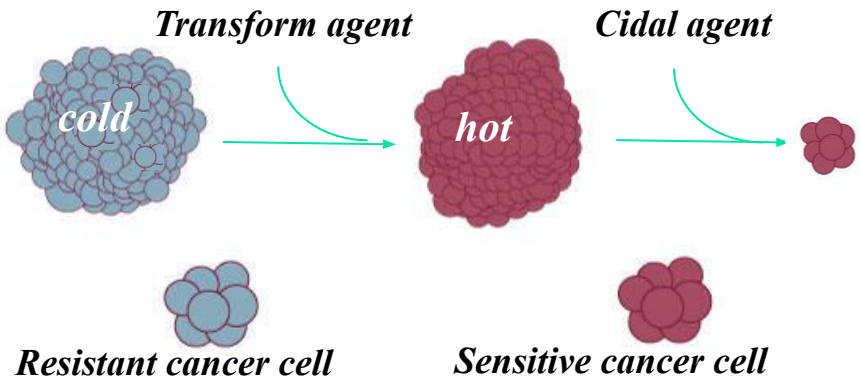
Omics-Driven Phenotype Compound Screening

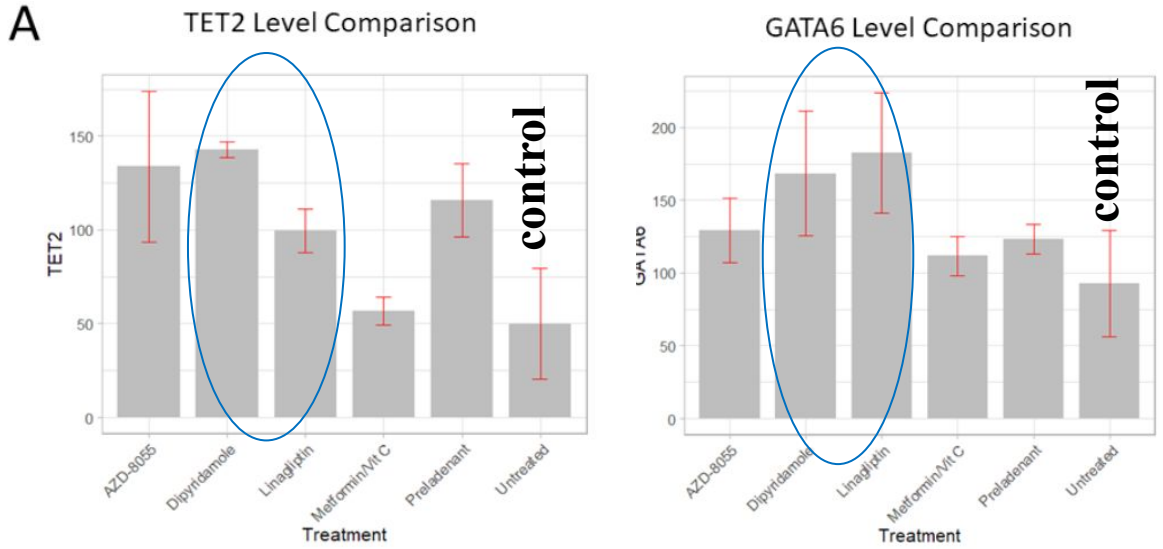


Qiu et al. *Bioinformatics,* 2020; Pham et al. *Nat Mach Intell,* 2021; He et al. *Bioinformatics,* 2021; Pham et al. *Patterns,* 2022; He et al. *Nat Mach Intell,* 2022; Wu et al. *PLoS CB,* 2022; Wu et al. *Cell Rep Methods,* 2023

Omics-Driven Phenotype Compound Screening

- Drug combination to induce immunotherapy response

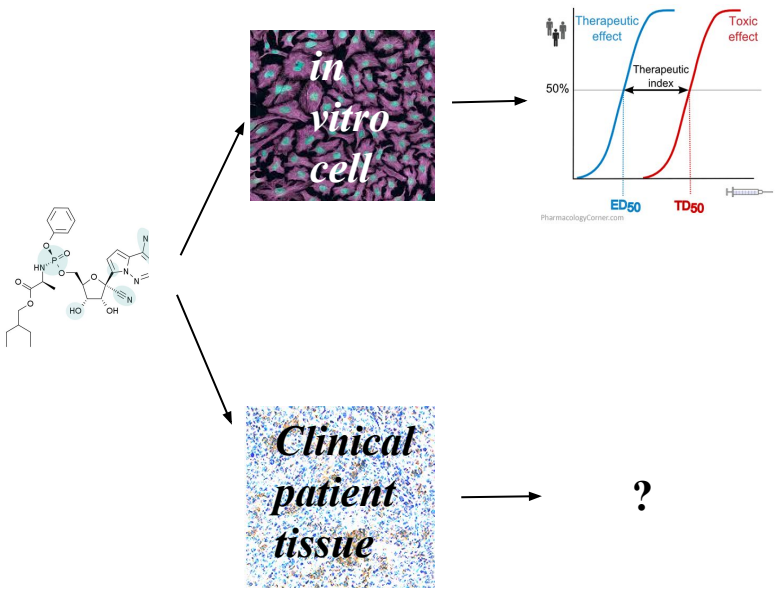


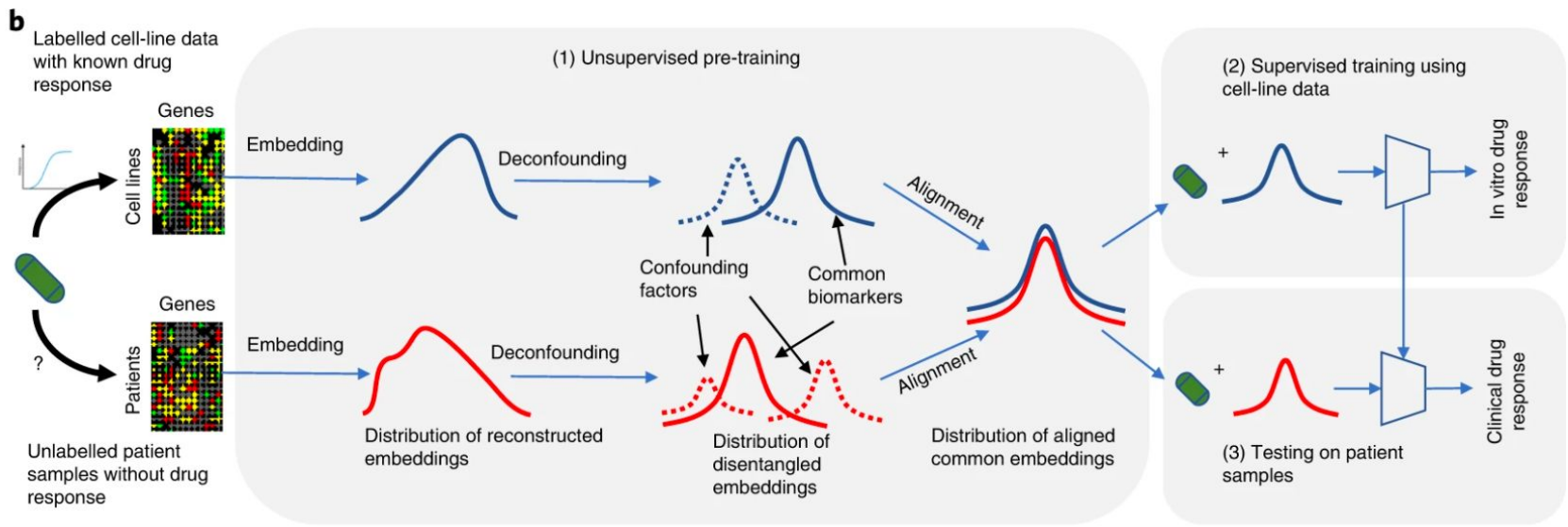


Pham et al. (2022) *Patterns*

Translating Cell Line Screens to Patient Clinical Responses for Personalized Medicine

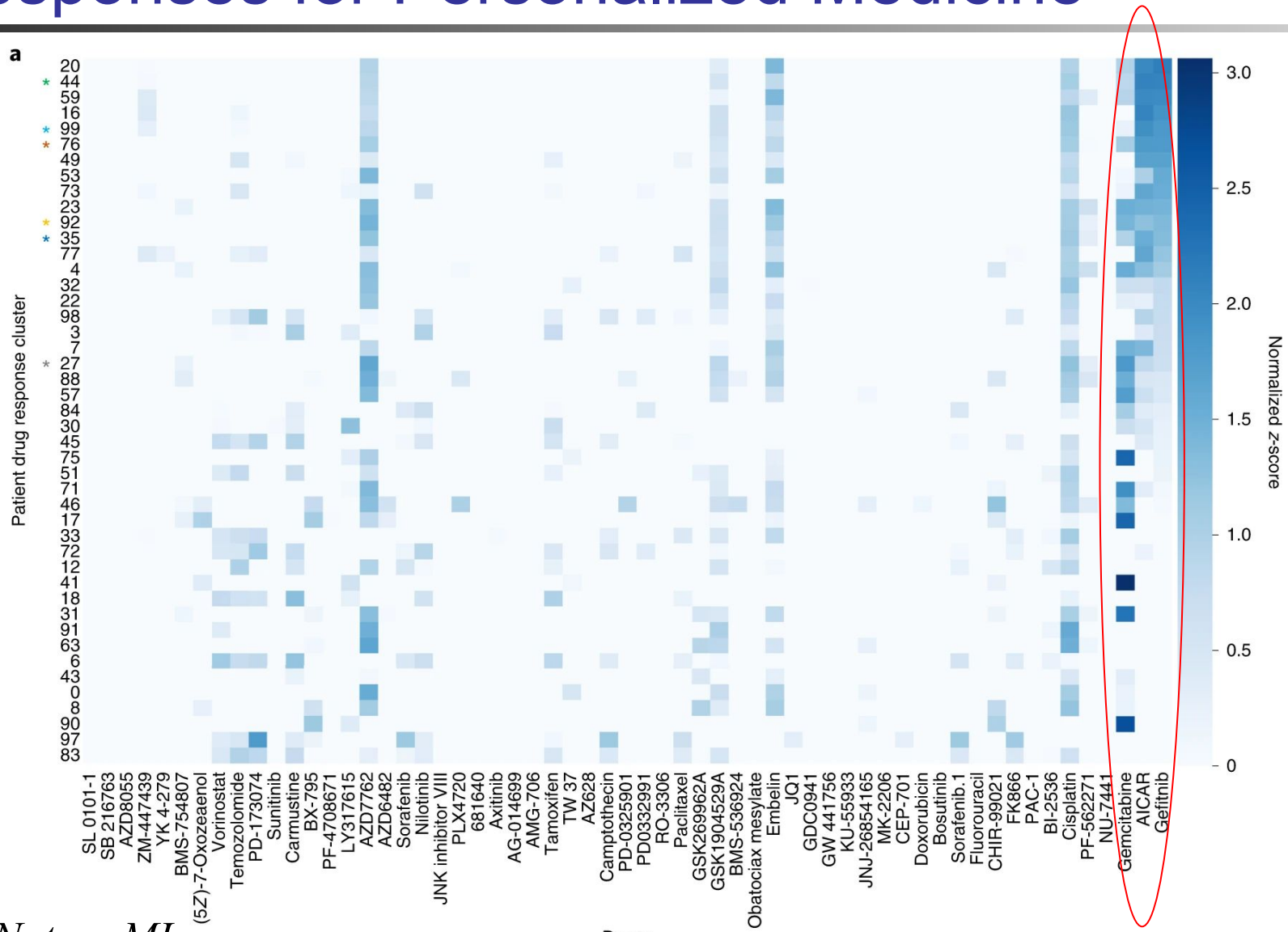
Context-aware Deconfounding Autoencoder (CODE-AE)





He et al. (2022) *Nature MI*

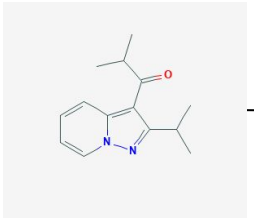
Translating Cell Line Screens to Patient Clinical Responses for Personalized Medicine

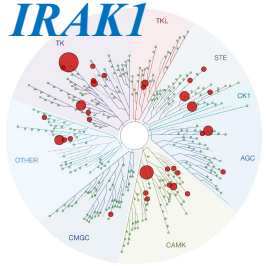


He et al. (2022) *Nature MI*

Put Together: Multi-Scale Modeling of Drug Actions for AD Drug Repurposing

Asthma drug (PDE inhibitor)





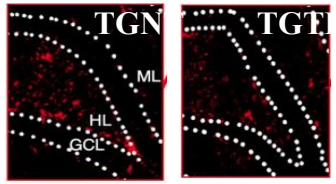
TLR/MYD88/ NF-kB pathway

TNF pathway

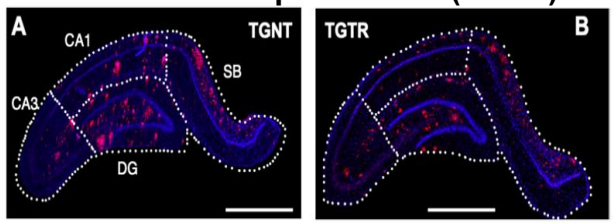
Lipopolysaccharide Synthesis pathway

Herpes simplex virus infection

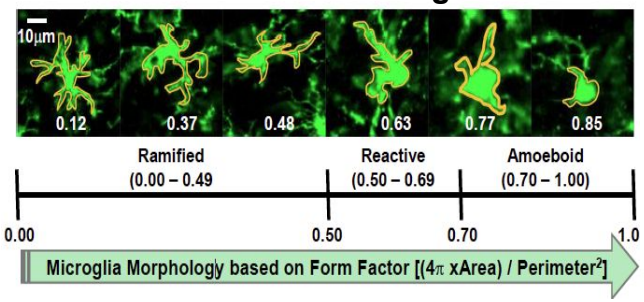
Reduced Tangle Burden (in red)



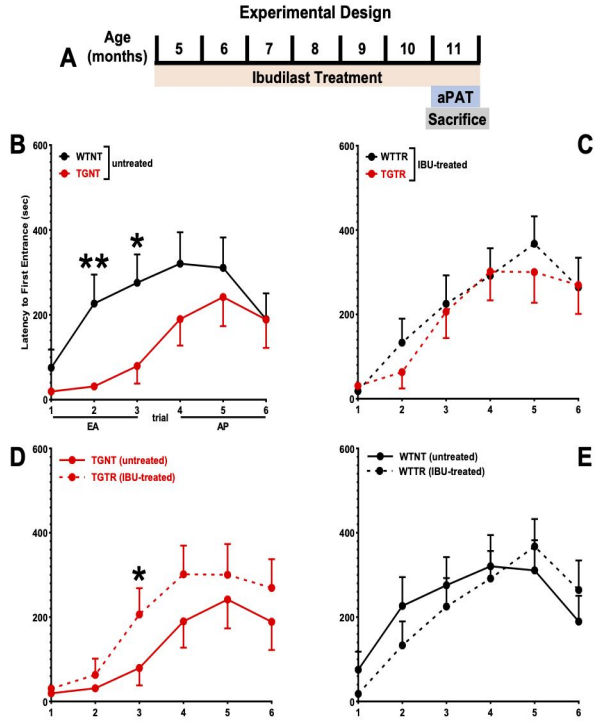
Reduced Plaque Burden (in red)



Reduced Microgliosis



Improve Learning Memory

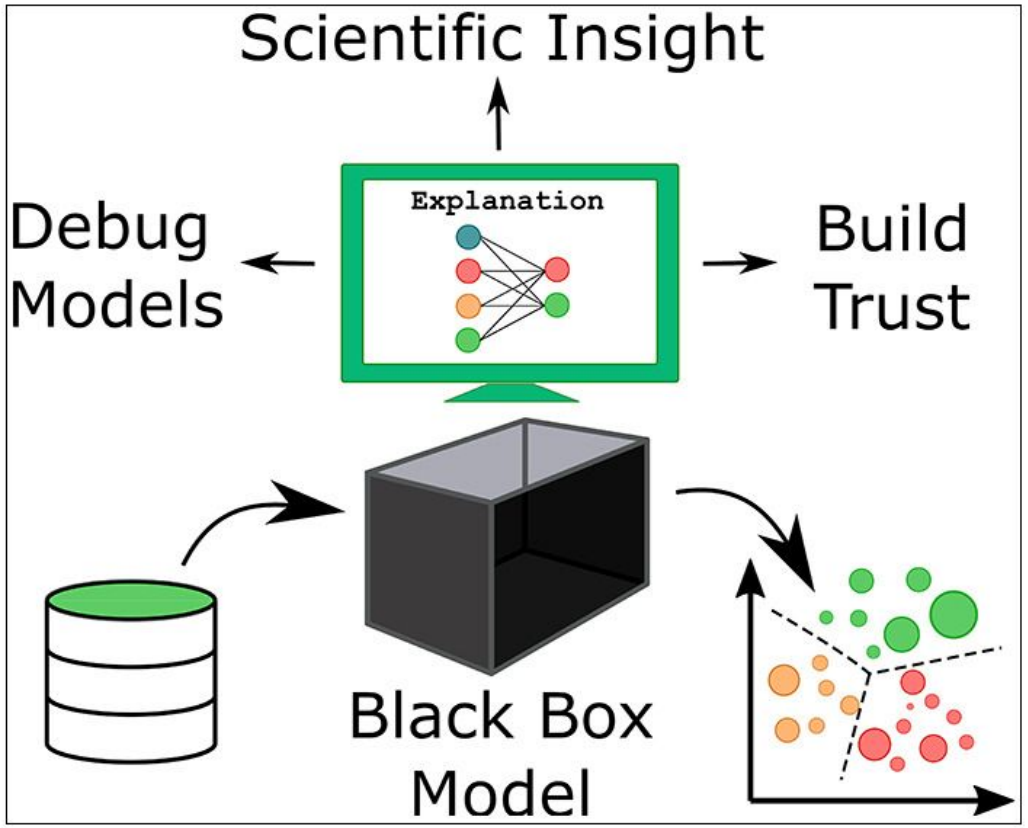


Drug → ***Genome-wide CPIs*** → ***Pathway*** → ***Clinical outcome***

Oliveros et al. (2022) *Brain*

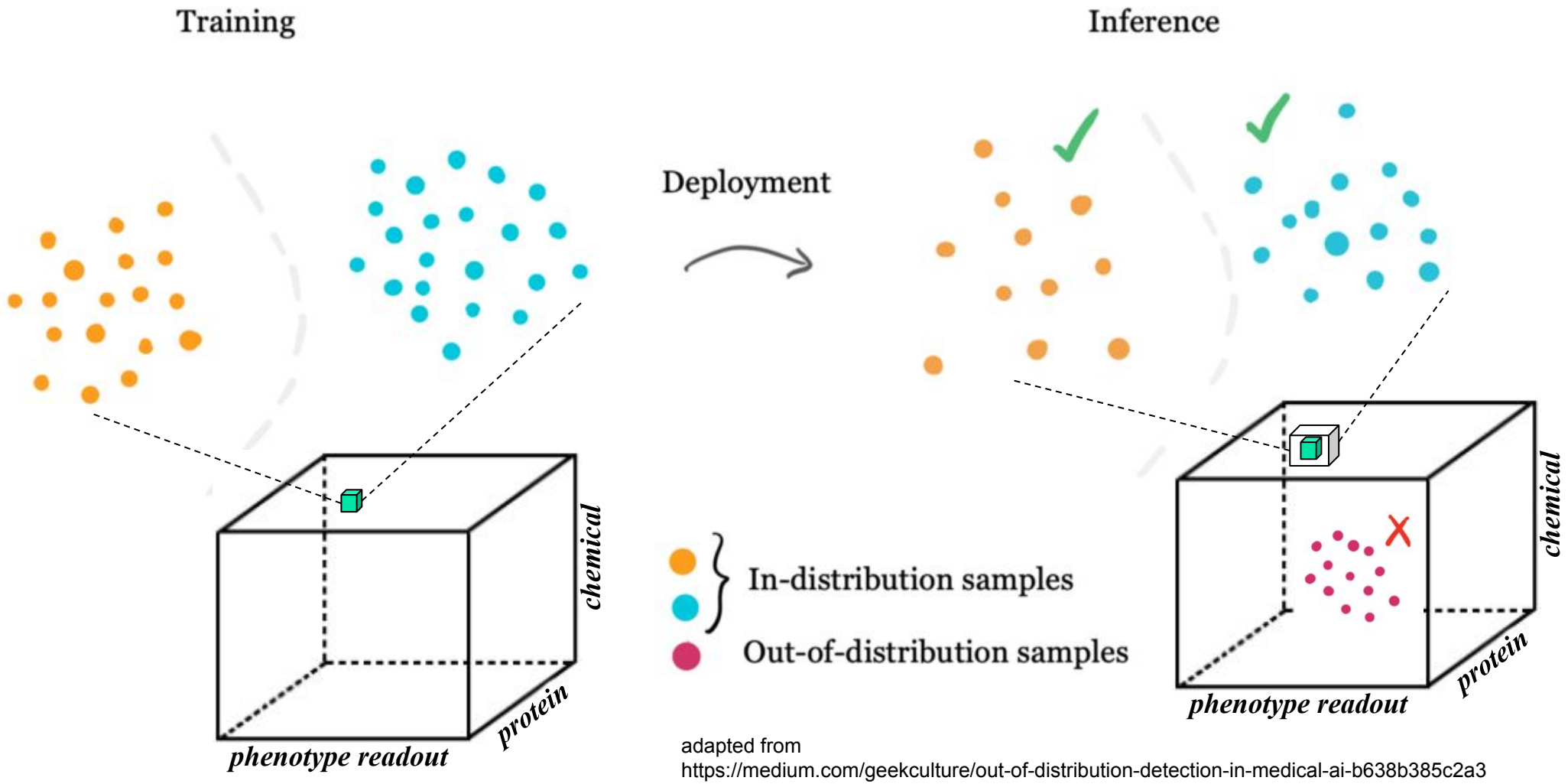
Will AI be Sufficient?

Limitations of ML: Interpretability



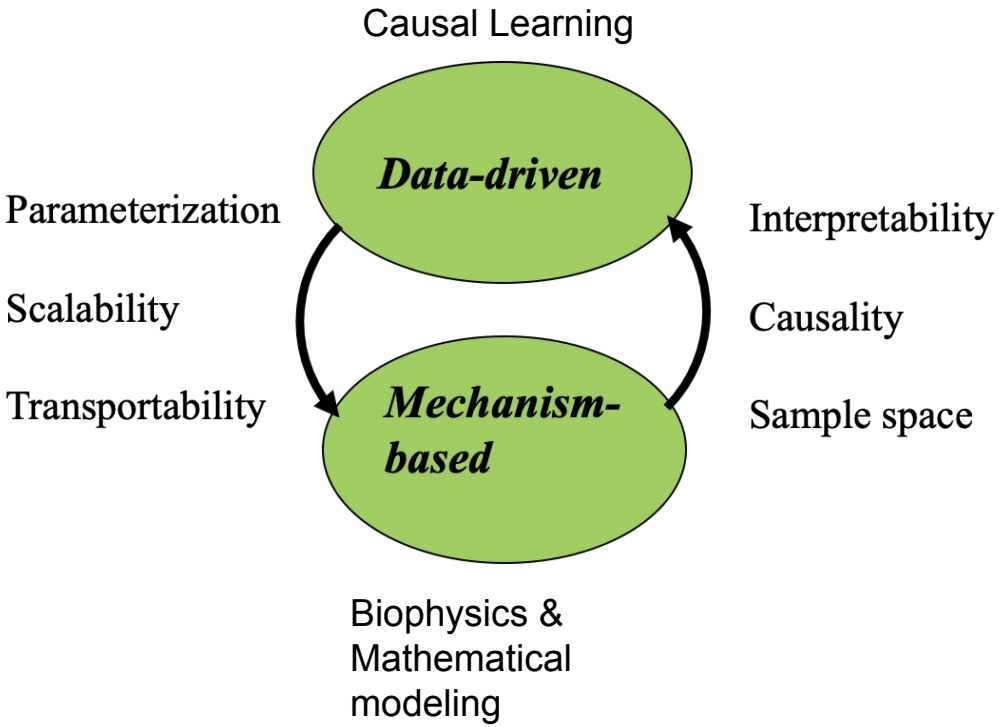
Oviedo et al. *Acc. Mater. Res.* (2022)

Limitations of ML: Out-of-distribution Generalization

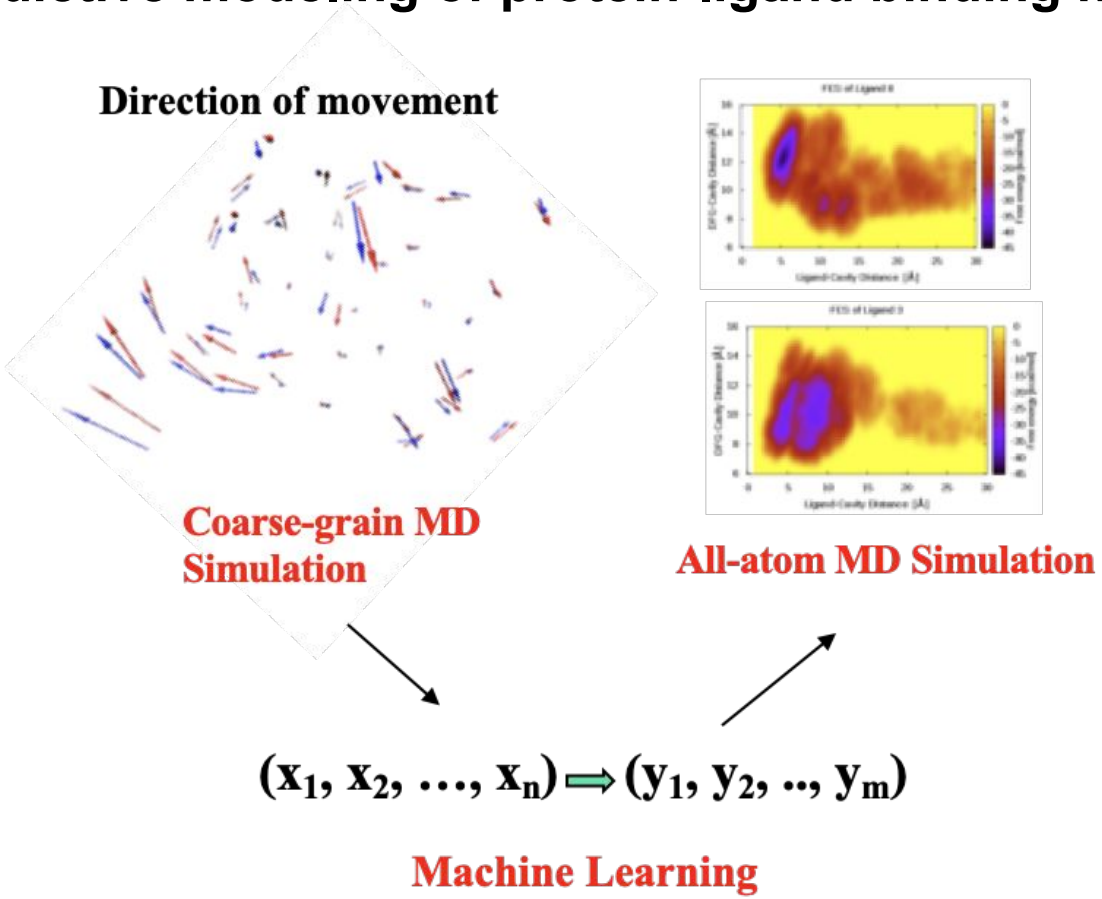


adapted from
https://medium.com/geekculture/out-of-distribution-detection-in-medical-ai-b638b385c2a3

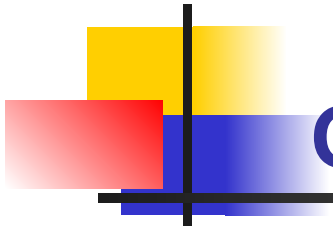
Integrating Data-driven and Mechanism-based Modeling



- **Predictive modeling of protein-ligand binding kinetics**



Chiu et. al (2016) *JCIM*



Outline

- One definition of data science
- A brief history of *in silico* early-stage drug discovery
- Bioinformatics methods for structure-based drug design
- Machine Learning/Deep Learning solutions to systems pharmacology
- Summary

Summary

- Computational approaches to drug discovery have advanced on 3 axes:
  1. The scale, complexity and amount of digital data available
  2. The type of computation employed:
     a) From the application of measurable physical features
     b) To the discovery of unknown features
     c) To the prediction of outcomes
  3. How we think about the problem
     a) One drug-one target
     b) Multiple drugs-multiple targets- populations
     c) Multiple drugs-multiple targets-one human