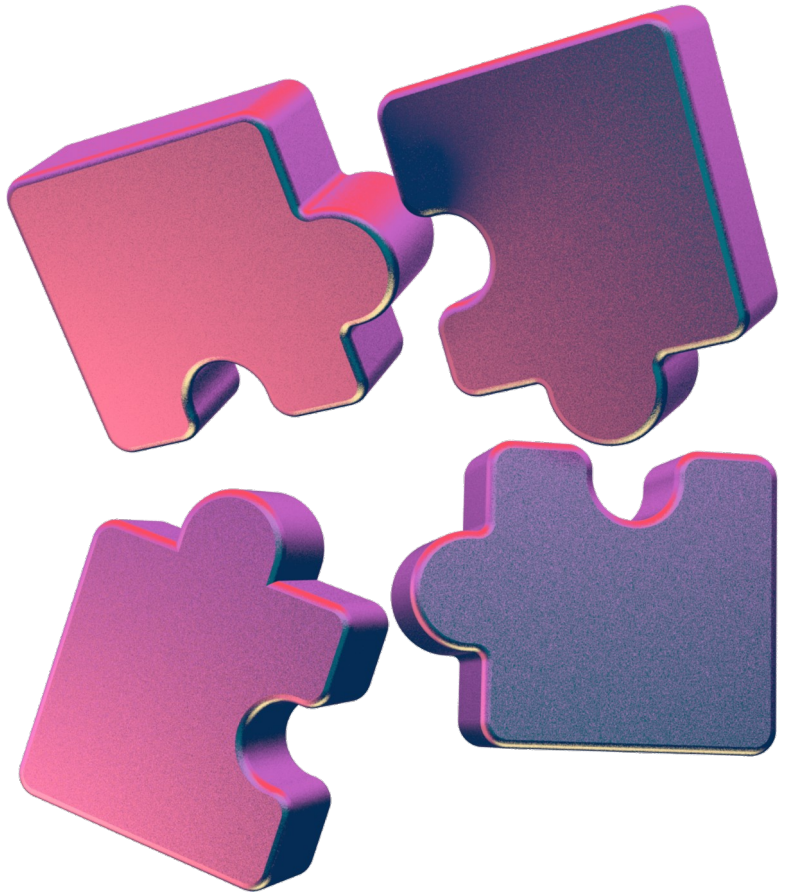
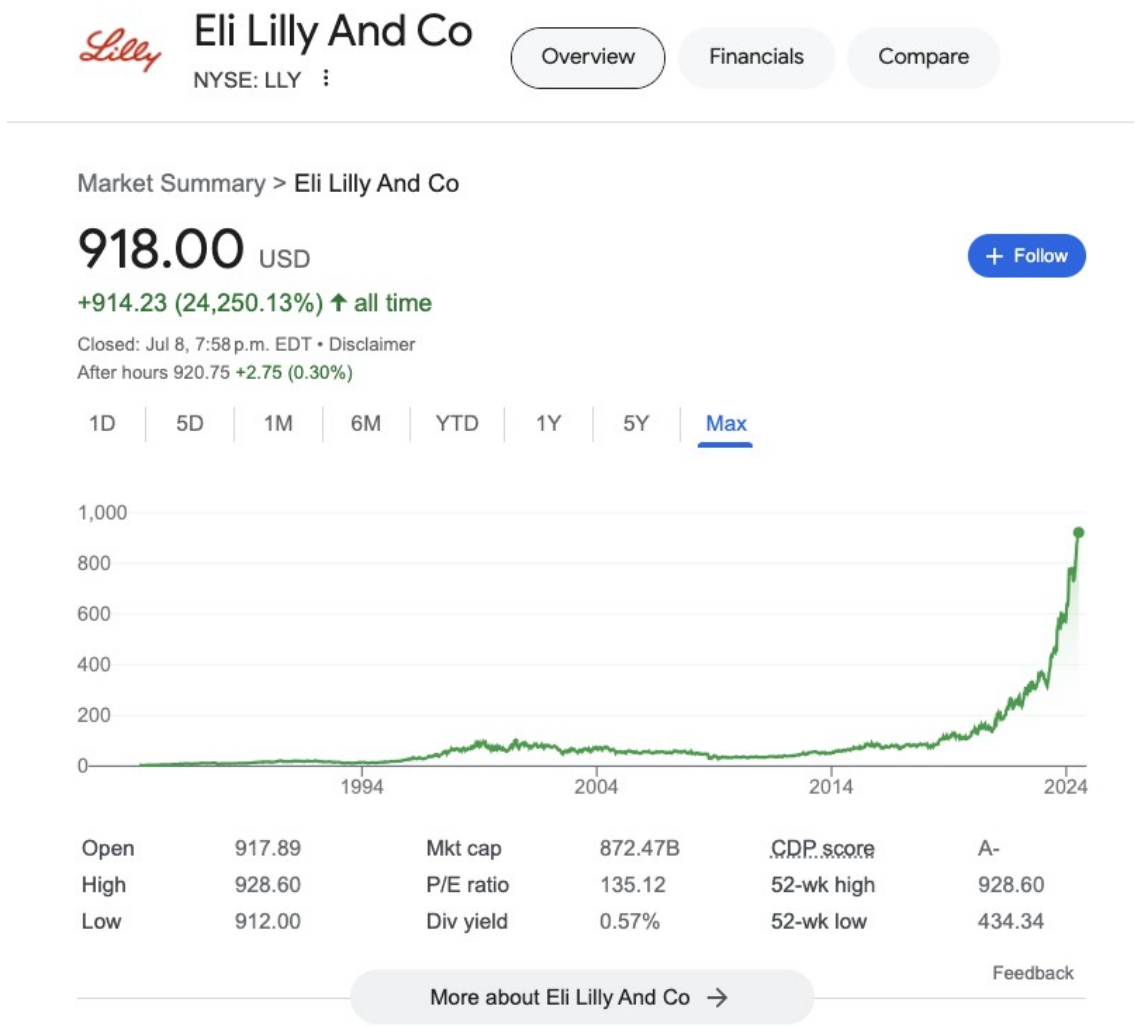
ML for Drug Discovery Summer School 2024 @ Mila A brief summary



Motivation: What’s the current state of ML for DD?

Motivation


Recent headlines (2018-2020)
SPOTLIGHT · 30 MAY 2018
How artificial intelligence is changing drug discovery
World first breakthrough in AI drug discovery
By Emma Morriss - January 30, 2020
RAPID GROWTH IN PUBLISHED RESEARCH USING AI FOR DRUG DISCOVERY
AI 2020:
THE FUTURE OF DRUG DISCOVERY


Eli Lilly And Co
NYSE: LLY
Overview
Financials
Compare
Market Summary > Eli Lilly And Co
918.00 USD
Follow
+914.23 (24,250.13%) ↑ all time
Closed: Jul 8, 7:58 p.m. EDT • Disclaimer
After hours 920.75 +2.75 (0.30%)
1D 5D 1M 6M YTD 1Y 5Y Max
1,000
800
600
400
200
0
1994
2004
2014
2024
Open 917.89
High 928.60
Low 912.00
Mkt cap 872.47B
P/E ratio 135.12
Div yield 0.57%
CDP score A-
52-wk high 928.60
52-wk low 434.34
Feedback
More about Eli Lilly And Co

Open Questions

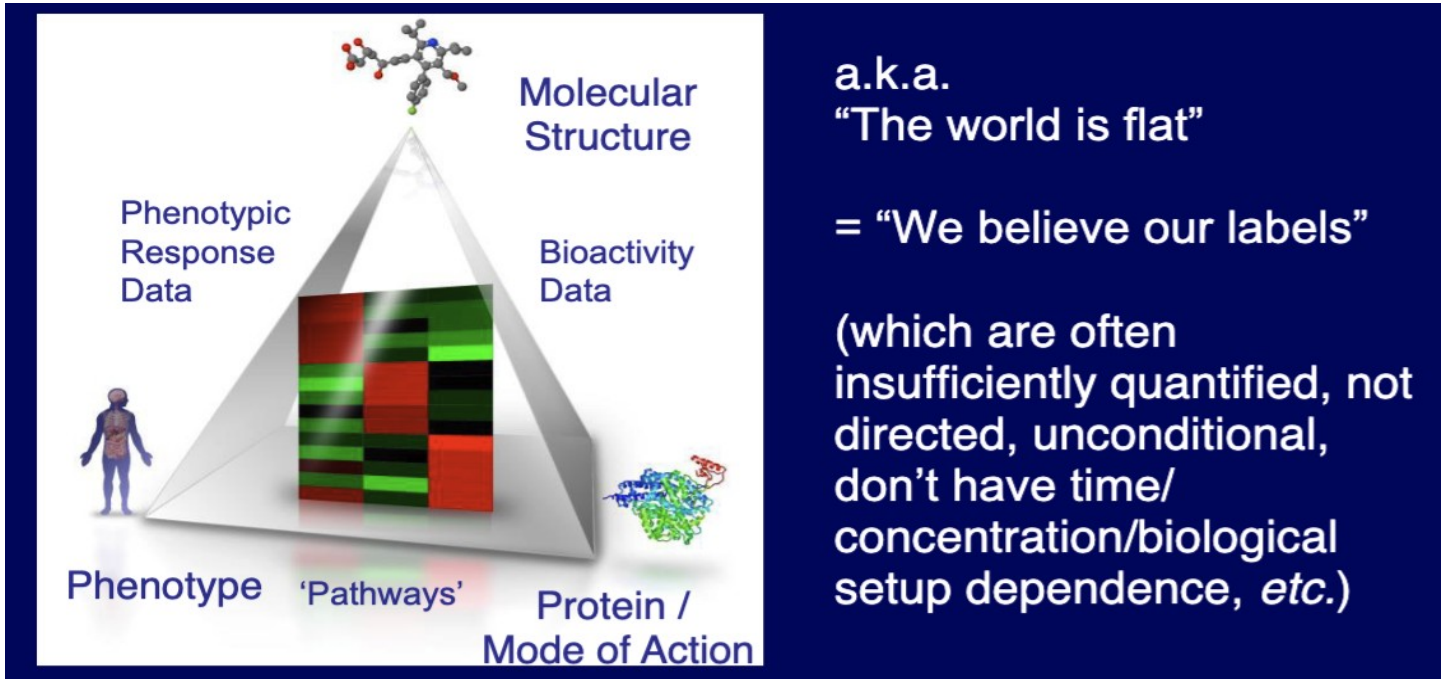
Are we moving from the ‘all you need is small algorithms’ to the ‘data is all you need’ paradigm?
(so what data do we need, then?) (Computer is tremendously powerful… but is our data?)

what is the correct data modality to key into my new setting’
(for example, for cellular biology should it be ATACseq, or RNAseq, or proteomics or all of the above)?”

What are the theoretical and scientific limitations of AIDD (AI for Drug Discovery)?
(Causality, confidence, quantification)

Do LLMs have use in AIDD?

Are we reliving ‘The Billion Dollar Molecule’,
but with an AI twist?



Open Questions

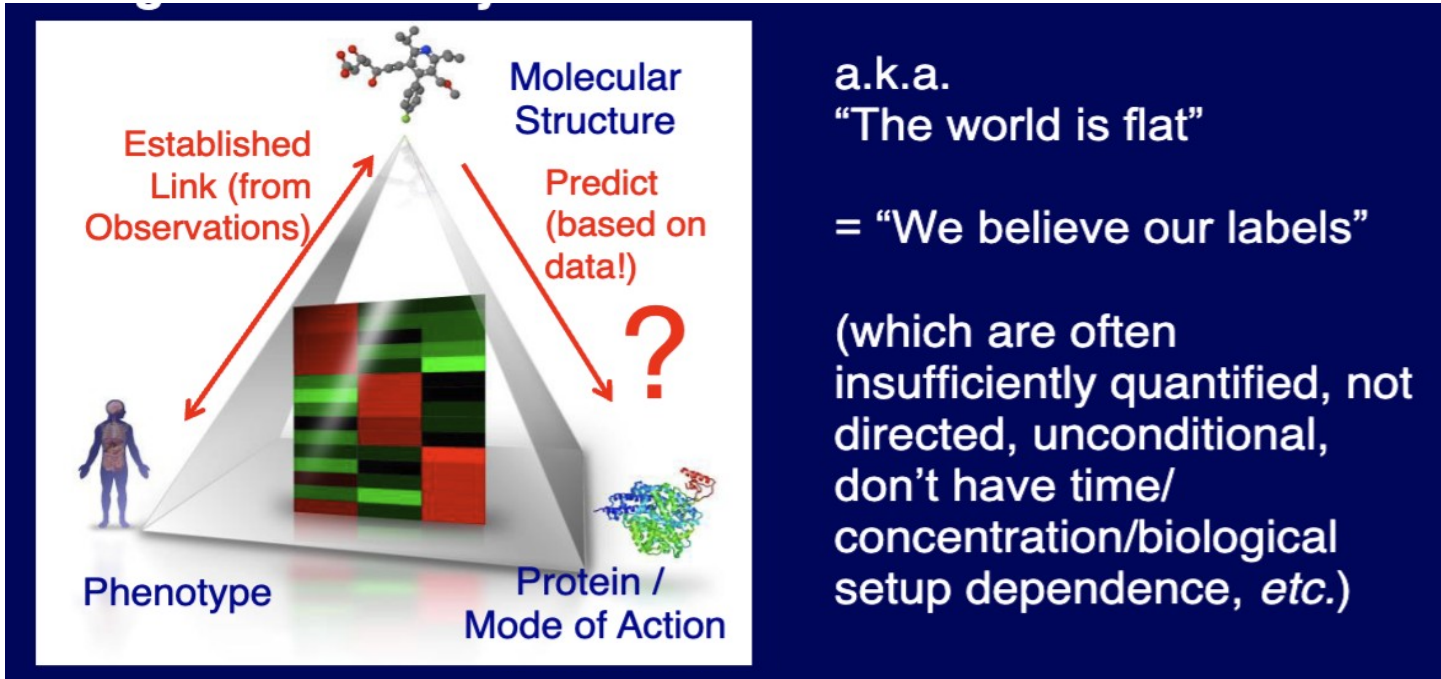
Are we moving from the ‘all you need is small algorithms’ to the ‘data is all you need’ paradigm?
(so what data do we need, then?) (Computer is tremendously powerful… but is our data?)

what is the correct data modality to key into my new setting’
(for example, for cellular biology should it be ATACseq, or RNAseq, or proteomics or all of the above)?”

What are the theoretical and scientific limitations of AIDD (AI for Drug Discovery)?
(Causality, confidence, quantification)

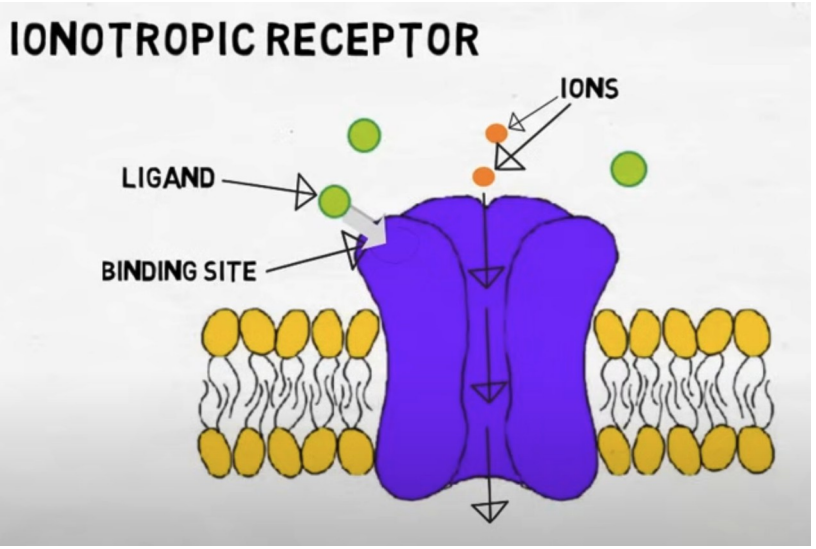
Do LLMs have use in AIDD?

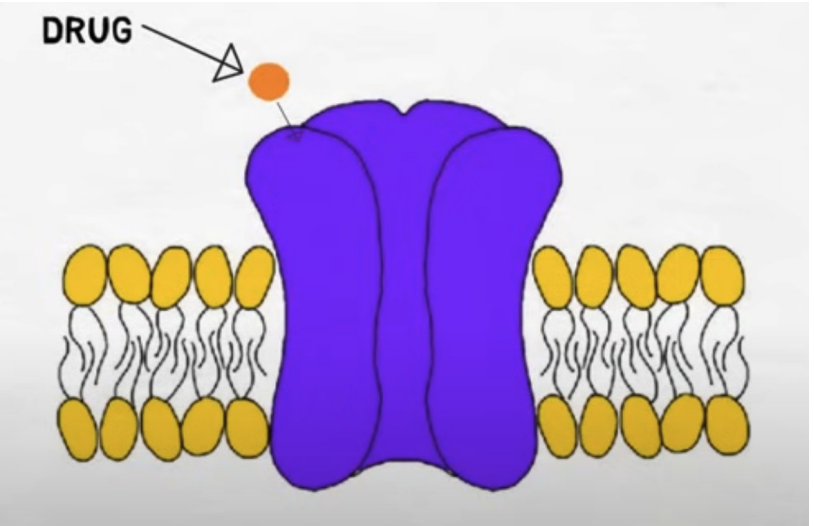
Are we reliving ‘The Billion Dollar Molecule’,
but with an AI twist?

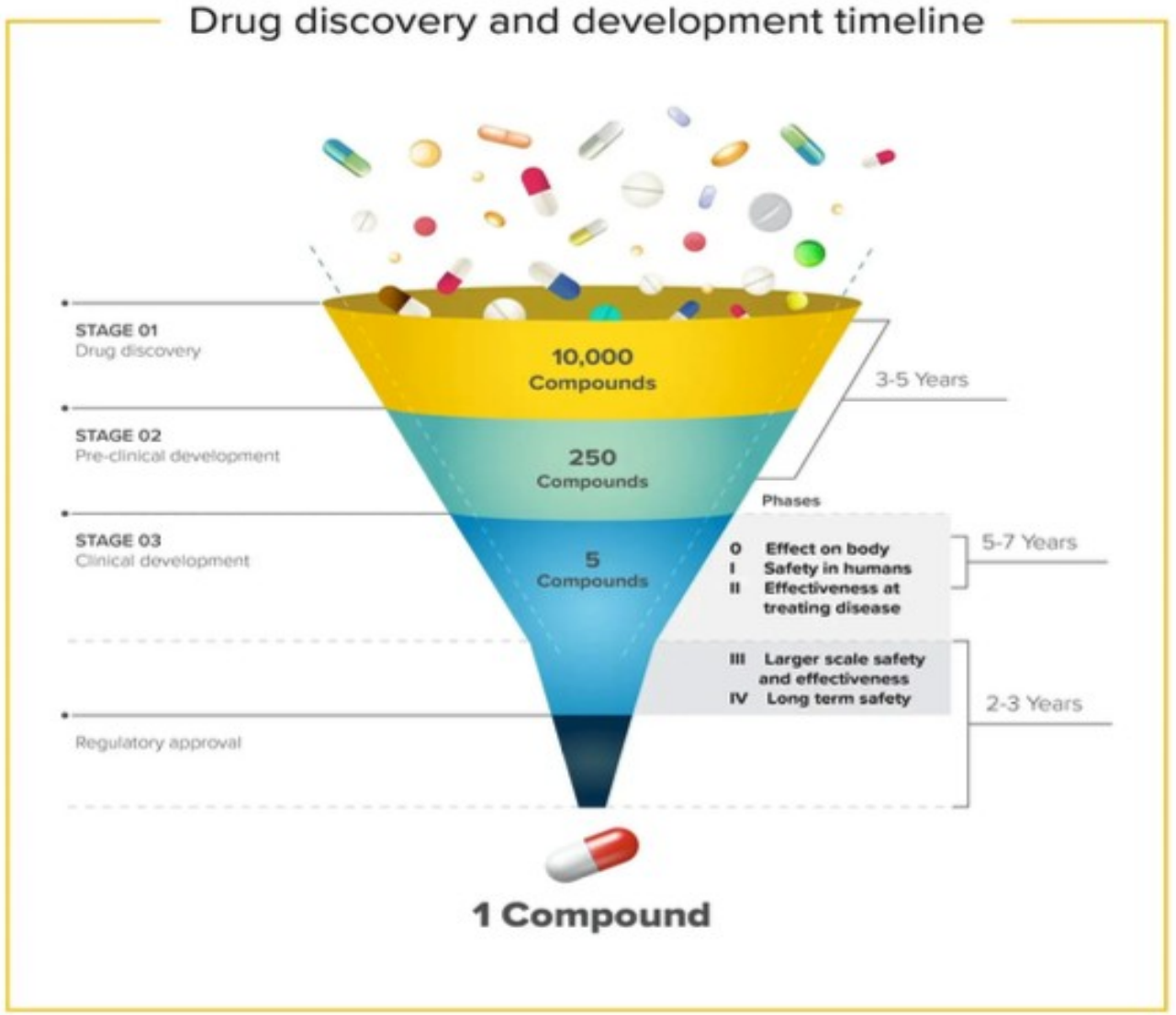


Refresher

THE LONG FUNNEL TOWARDS A DRUG

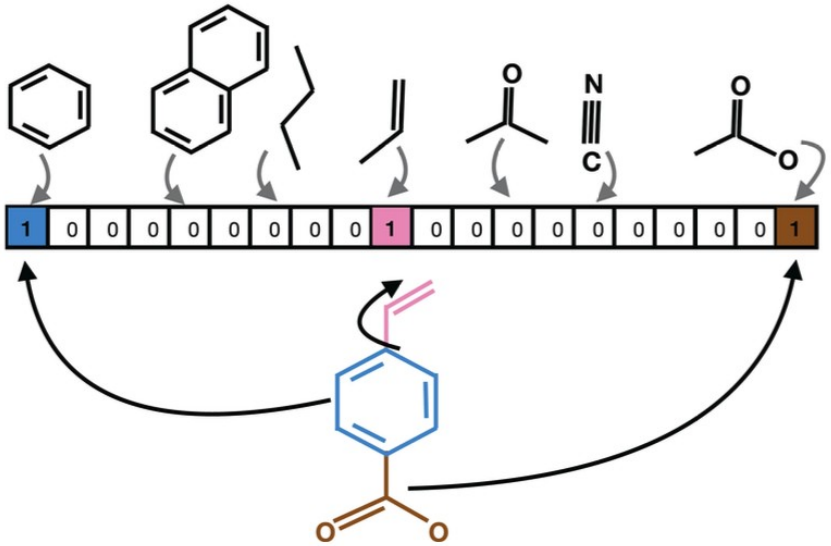






Refresher

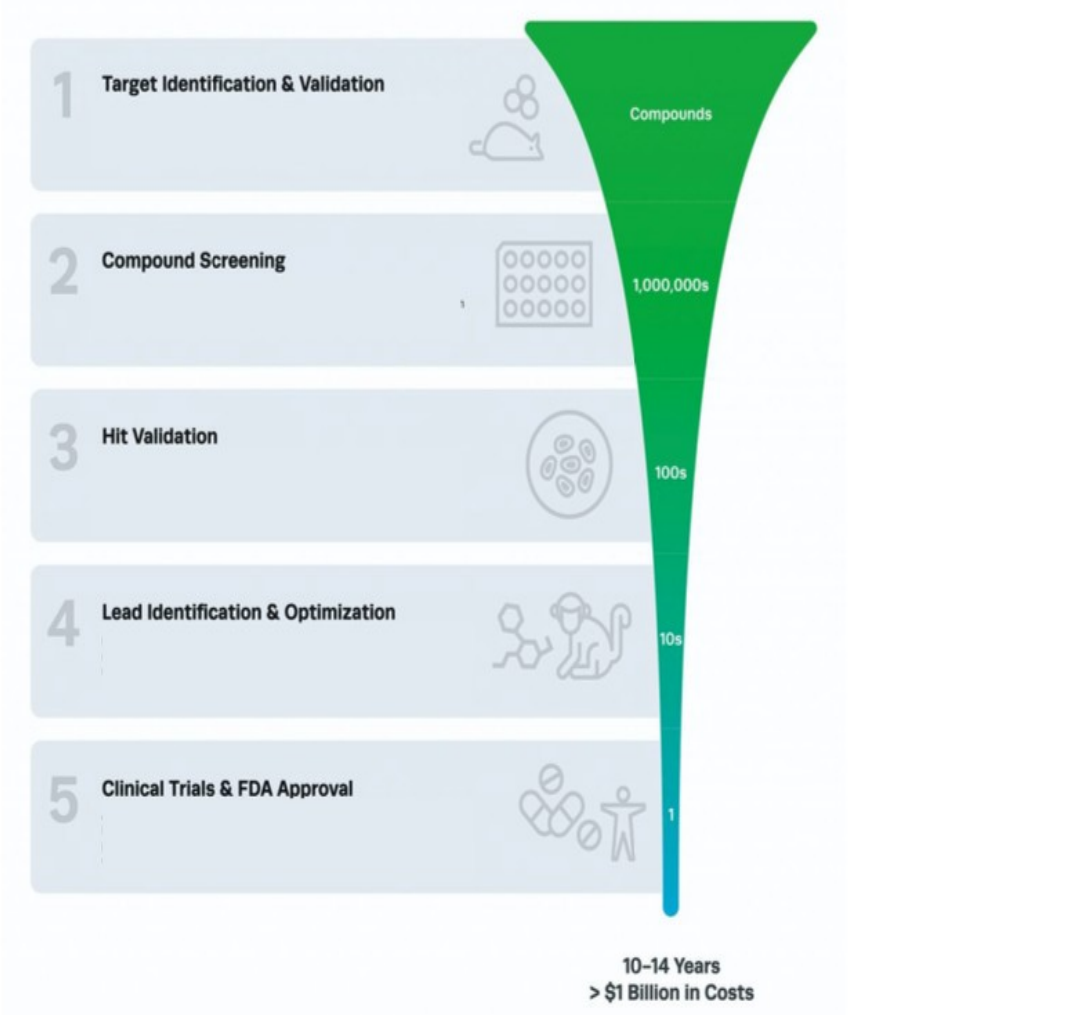
MACHINE LEARNING 101: TURN MOLECULES INTO VECTORS



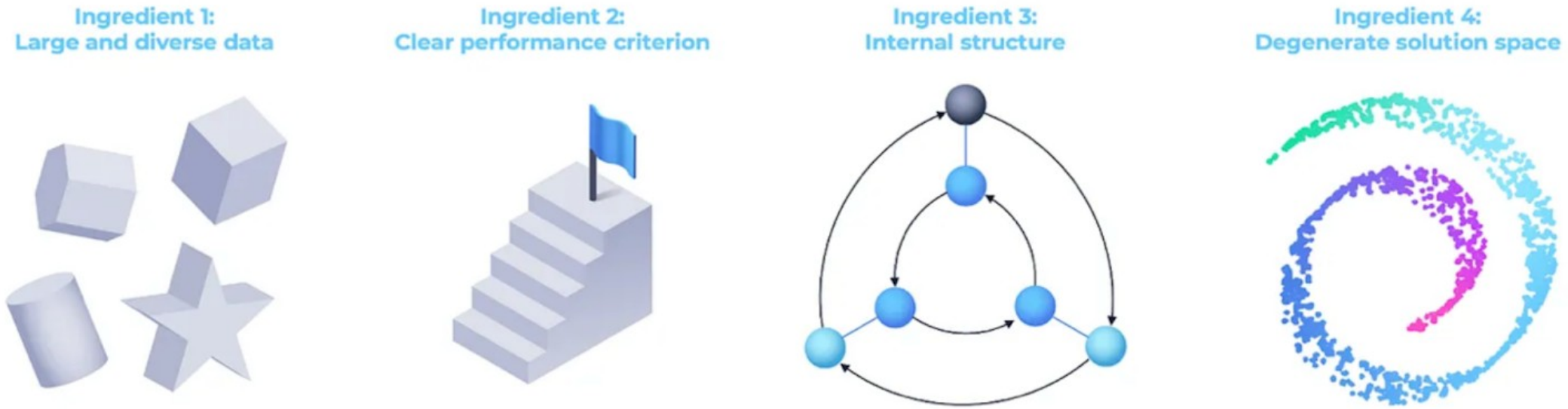
Turn into Linear Algebra !

Gateway to ML !

LOW DATA IS A FUNDAMENTAL CHALLENGE IN DRUG DISCOVERY



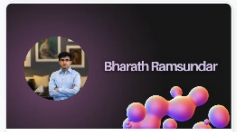
4 secret sauce ingredients for successful applications of ML in biology:



Four “ingredients” necessary to make biological problems amenable for machine learning methods: 1) Large and, most importantly, diverse data; 2) clearly defined performance criterion that can be used as a loss function; 3) useful internal structures generated through human language or biological evolution or coming from physical constraints, e.g., a symmetry group shown here; and 4) a degenerate solution space, e.g., manifested in the form of “manifold hypothesis.”

Schedule & Outline

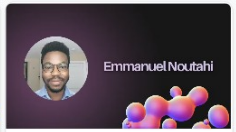
Day 1: Foundations and ML in Ligand-Based Modeling



1.1 Intro to ML in Drug Discovery: Principles & Applications

Professor: Bharath Ramsundar Lecture Recording: Coming soon... Slides:

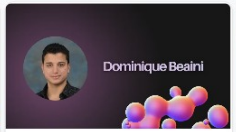
0 06/12/2024



1.2 Molecular Representation & Scoring

Professor: Emmanuel Noutahi Lecture Recording: Coming soon... Lecture Slides:

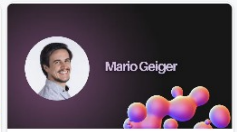
5 06/12/2024



1.3 Graph Neural Networks for Chemistry

Professor: Dominique Beaini Lecture Recording: Coming soon... Lecture Slides:

2 06/12/2024



1.4 Learning Geometry & 3D Symmetries

Professor: Mario Geiger Lecture Recording: Coming soon... Lecture Slides: https://slides.com/mariogeiger/learning-geometry-3d-...

0 06/12/2024

Day 2: ML in Structure-Based Drug Discovery

2.1 ML in Structure-Based Drug Discovery

Professor: Gabriele Corso Lecture Recording: Lecture Slides:

10 06/13/2024

2.2 Learning ML Interatomic Potentials

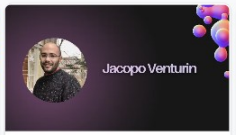
Professor: Gianni De Fabritiis Lecture Recording: Coming soon... Lecture Slides:

0 06/13/2024

2.3 Accelerate Atomistic Simulations, Sampling, and Dynamics

Professor: Pratyush Tiwary Lecture Recording: Coming soon... Lecture Slides: PPT https://docs.google.com/presentation/d/1sODOQ8r9eGmQZbx-...

0 06/13/2024

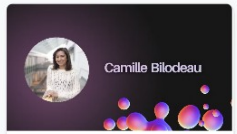


2.4 Coarse-Grained Biological Systems

Professor: Jacopo Venturin Lecture Recording: Coming soon... Lecture Slides:

3 06/13/2024

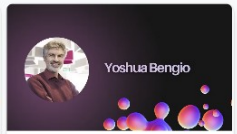
Day 3: Generative Models and Molecular Design



3.1 Generative Models of Molecular Structures

Professor: Camille Bilodeau Lecture Recording: Lecture Slides:

5 06/14/2024



3.2 Exploring Molecular Space & Active Learning

Professor: Yoshua Bengio Lecture Recording: Lecture Slides:

0 06/14/2024

3.3 Synthesizability & Molecular Synthesis

Professor: Connor Coley Lecture Recording: Lecture Slides:

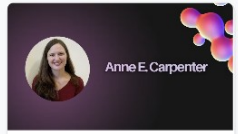
0 06/14/2024

3.4 Sampling Physical & 3D Systems

Professor: Michael Bronstein Lecture Recording: Lecture Slides:

1 06/14/2024

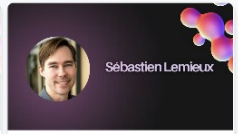
Day 4: Target Discovery and Deconvolution



4.1 Phenomics in Drug Discovery

Professor: Anne Carpenter Lecture Recording: Lecture Slides:

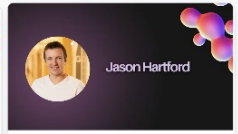
0 06/17/2024



4.2 Multi-Modal Omics & AI

Professor: Sébastien Lemieux Lecture Recording: Lecture Slides:

0 06/17/2024



4.3 Causal Discovery & Representation Learning

Professor: Jason Hartford Lecture Recording: Lecture Slides:

0 06/17/2024

4.4 Modeling Population Dynamics

Professor: Charlotte Bunne Lecture Recording: Lecture Slides:

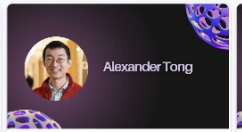
6 06/17/2024

Day 5: Frontiers in AI for Drug Discovery

5.1 LLMs in Drug Discovery

Professor: Andres M Bran Lecture Recording: Coming soon... Slides: Notebook: Colab

1 06/18/2024



5.2 Protein Folding & Design

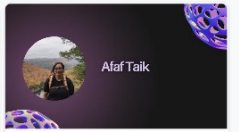
Professor: Alex Tong Lecture Recording: Coming soon... Slides:

4 06/18/2024

5.3 Open-Source Initiatives & Benchmarking Efforts

Professor: Karmen Condic-Jurkic Lecture Recording: Coming soon... Slides:

2 06/18/2024

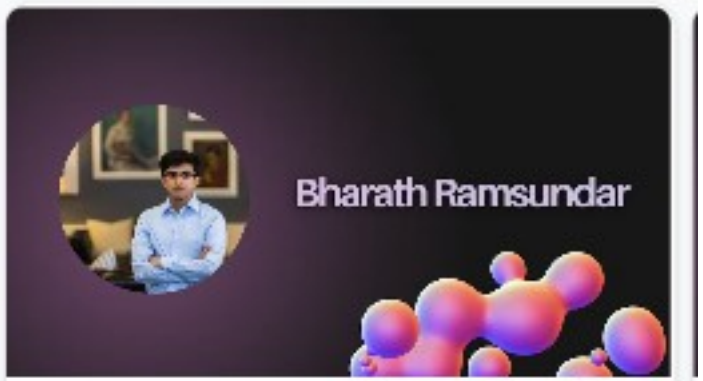


5.4 Ethical & Bias Concerns

Professor: Afaf Taik Lecture Recording: Coming soon... Slides:

1 06/18/2024

1.1 Bharath



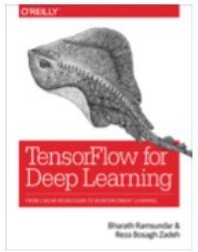
1.1 Intro to ML in Drug Discovery: Principles & Applications

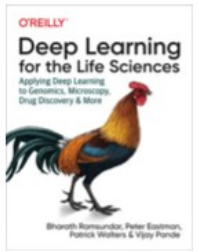
Professor: Bharath Ramsundar
Lecture Recording: Coming soon... Slides:

0 06/12/2024

BRIEF INTRO: BHARATH RAMSUNDAR

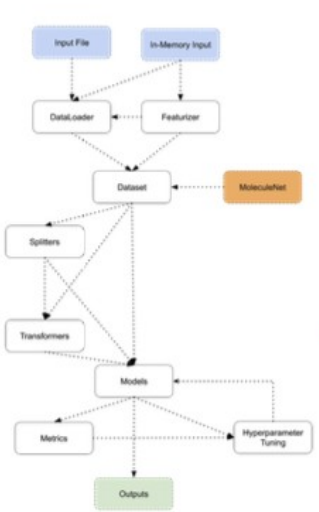
- Stanford PhD, Pande Group, UCB EECS/Math Alum
- Lead Developer of DeepChem
- Founder/CEO of Deep Forest Sciences





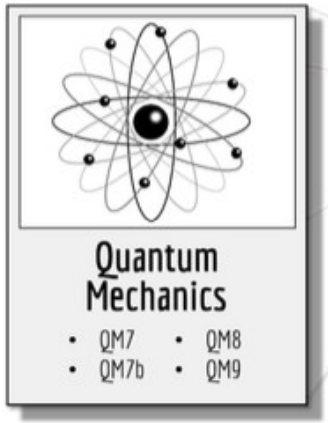
SCIENTIFIC MACHINE LEARNING WITH DEEPCHEM

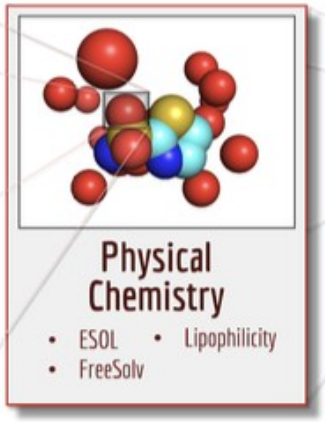
- DeepChem is a framework to apply AI to drug discovery and other scientific problems.
- DeepChem offers composable re-usable pipelines for scientific workflows
- DeepChem increasingly supports non-molecular applications for general scientific computing.

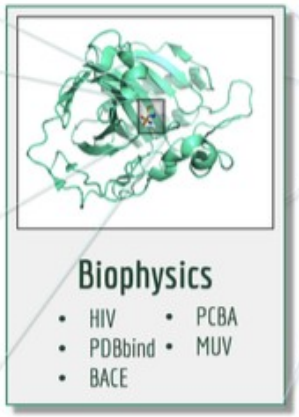


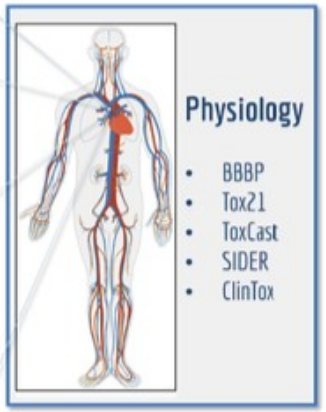
deepchem

MOLECULENET PROVIDES CONVENIENT BENCHMARKS









1.1 Summary

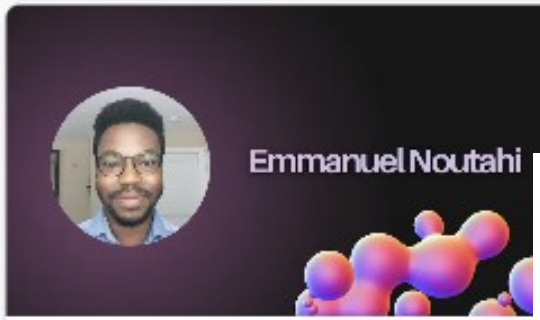
1. Despite the hype, AI/ML is not “that useful” in Drug Discovery, e.g. because it cannot understand long time scales that drug discovery requires, e.g. similar to civil engineering and building of bridge and castles
- Don’t trust machine learning too much.
- Physics > Datasets (most of the time)
- Understand the biology and chemistry of the system.
- AI is not a substitute for basic science. AI/ML does not generalize well; science does.

2. Low Data is the fundamental challenge in Drug Discovery, e.g. as we take out biology predictions from the test tube, they break down in the real world, because out of distribution generalization is a really hard problem (and unsolvable with the wrong data)

3. Strategies to tackle low data are
- Physical Simulations: to generate synthetic data
- Metric + Meta Learning: to compare prediction vectors
- Multitask-Learning: to combine datasets
- Transfer Learning: from assay readouts

1.2 Emmanuel

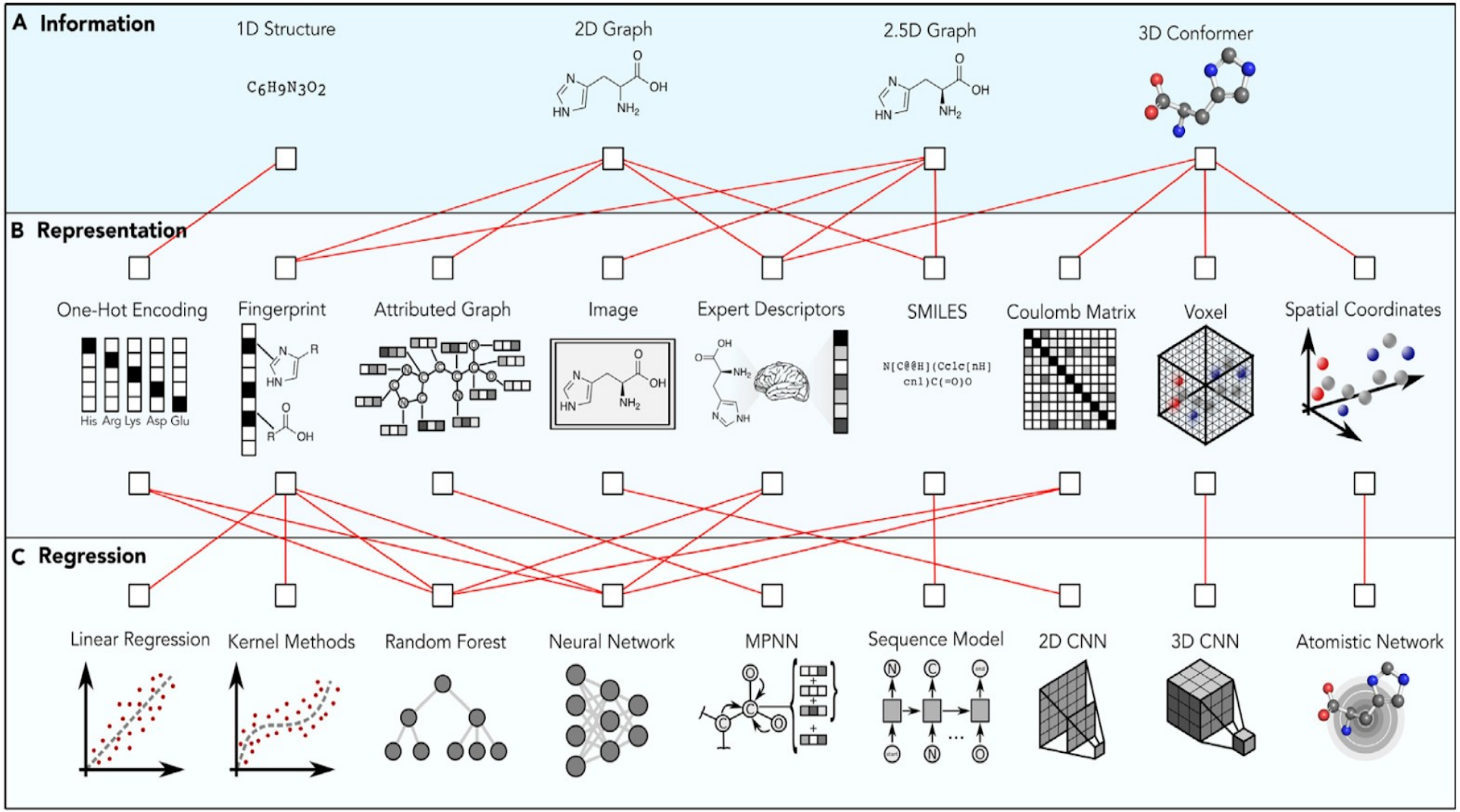
This talk had overlap with Bharat ‘s talk



1.2 Molecular Representation & Scoring

Professor: Emmanuel Noutahi
Lecture Recording: Coming soon... Lecture Slides:

5 06/12/2024



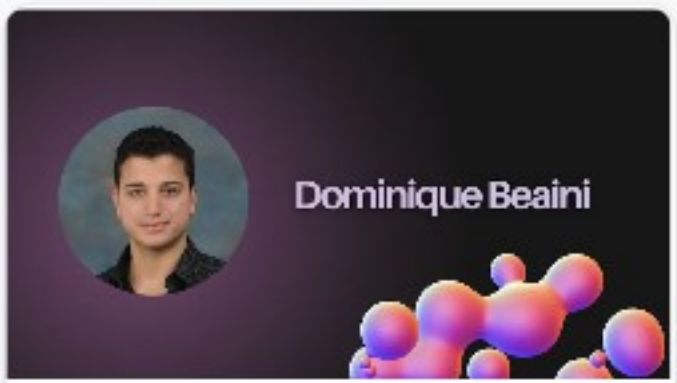
1.2 Summary

This talk had overlap with Bharat ‘s talk

Good representations are critical: “A method cannot save an unsuitable representation, which cannot remedy irrelevant data, for an ill thought-through question.”

Random Forest + ECFP is a strong base line,
(as shown in “Random forest classification for predicting lifespan-extending chemical compounds” by Kapsiani et al.)

1.3 Dominique

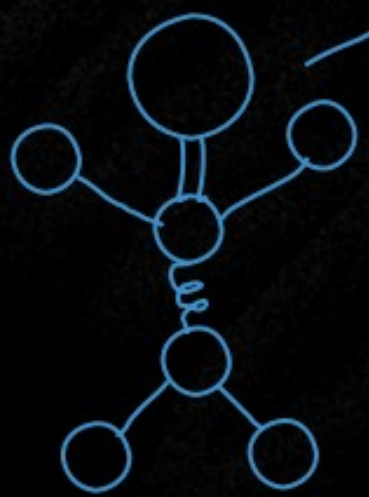


1.3 Graph Neural Networks for Chemistry

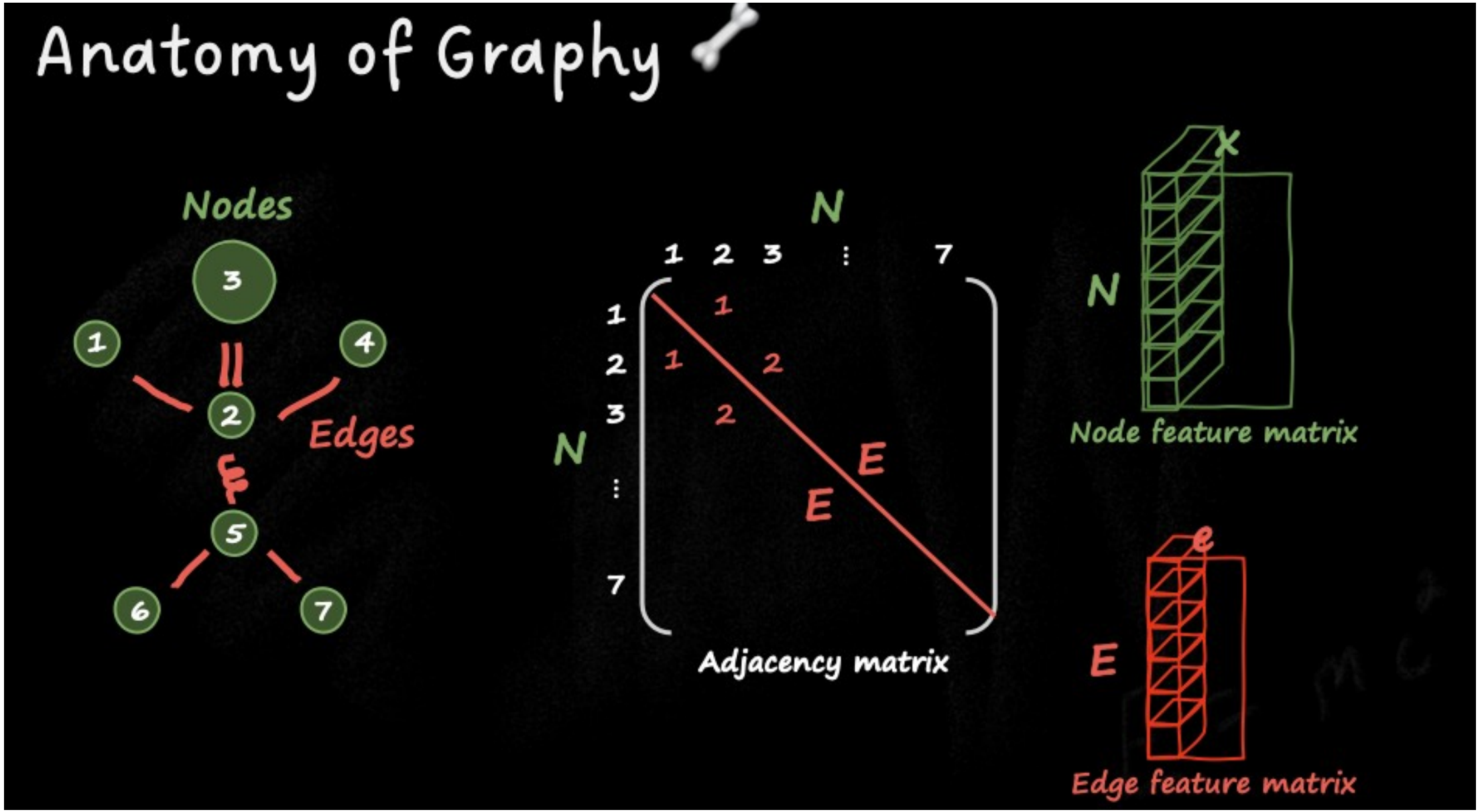
Professor: Dominique Beaini Lecture Recording: Coming soon... Lecture Slides:

2 06/12/2024

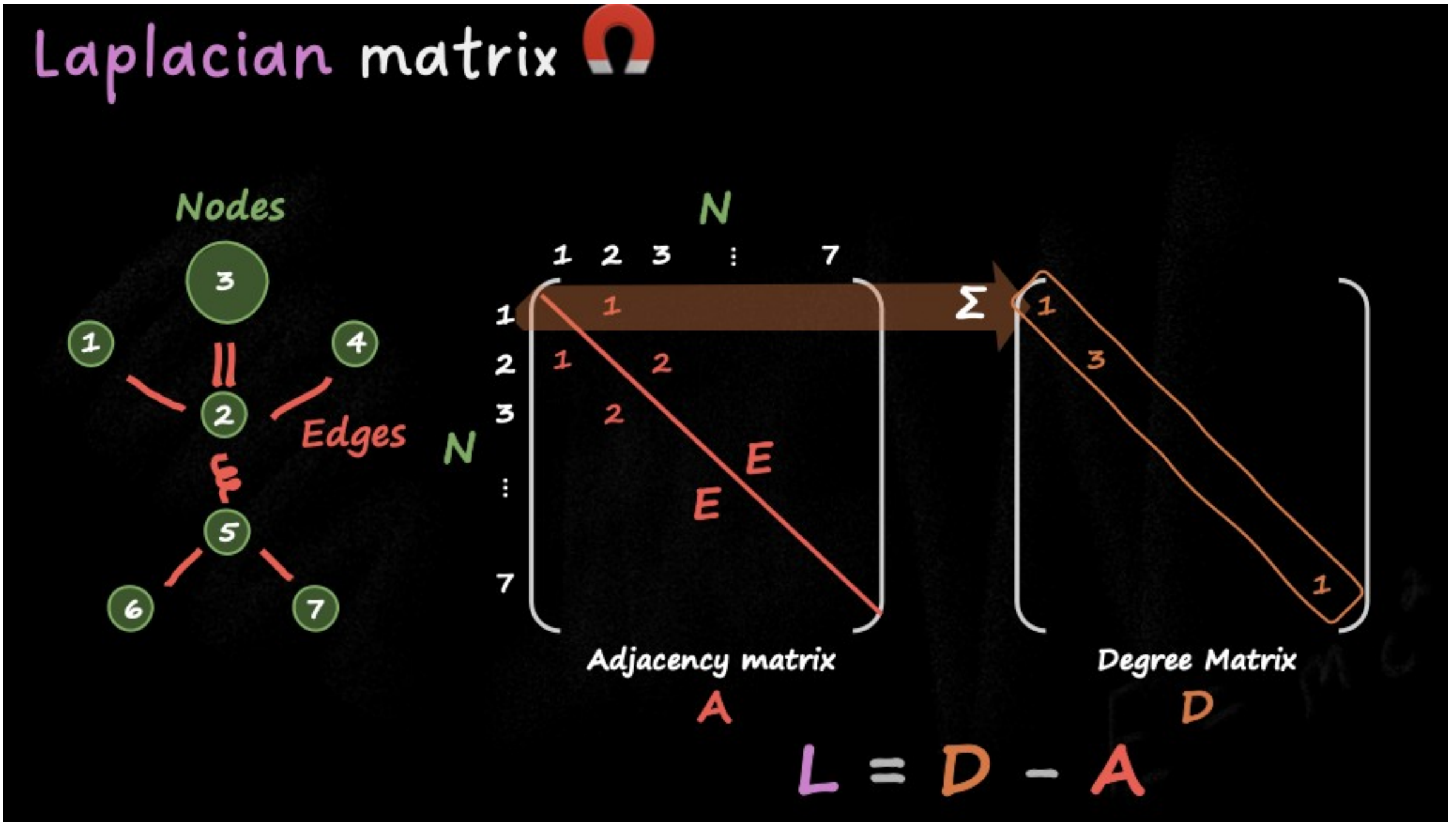
Meeting Graphy 👋

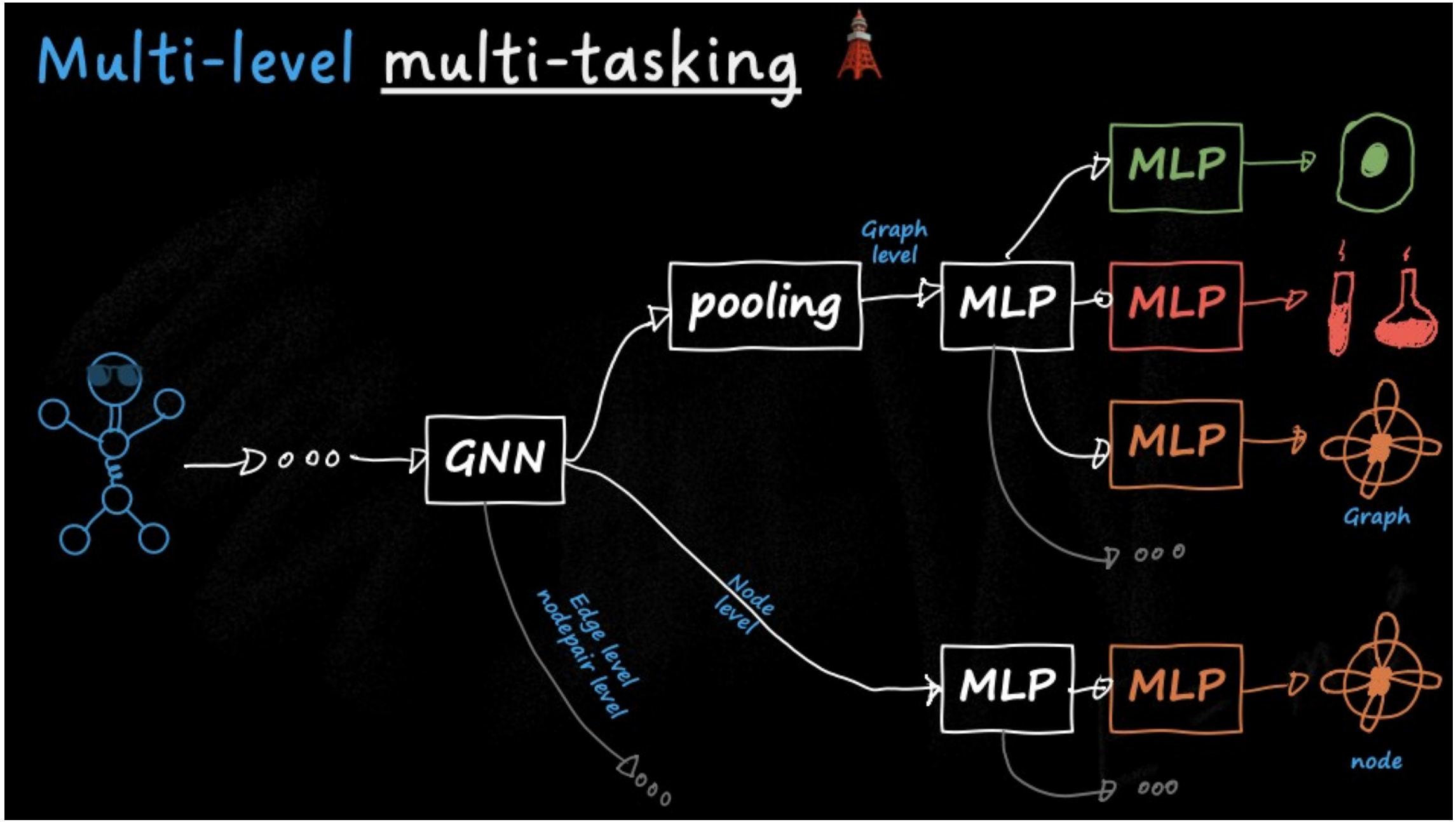


Hello everyone 👋! I'm Dom's assistant for today!
Let's visit the molecular graph world together!
We'll first learn what are graphs and how to manipulate them
Then we'll look into standard GNNs graph neural networks
And how to build more expressive GNNs for molecules
Finally, we will scale a Graph Transformer together

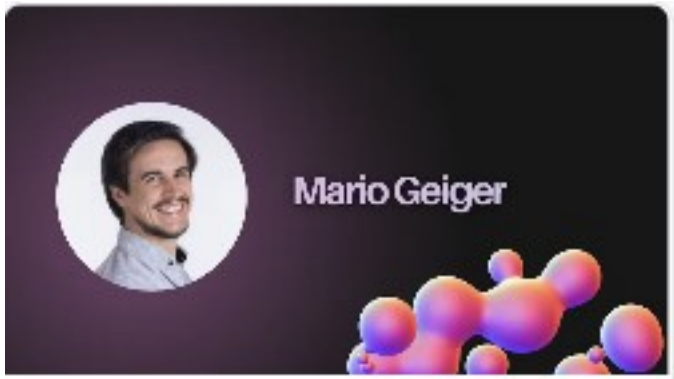

Anatomy of Graphy
Nodes
Edges
N
1 2 3 ⋮ 7
Adjacency matrix
E
E
X
N
Node feature matrix
e
E
Edge feature matrix

Laplacian matrix




Multi-level multi-tasking
GNN
pooling
Graph level
MLP
MLP
MLP
Graph
Node level
Edge level
nodepair level
MLP
MLP
node

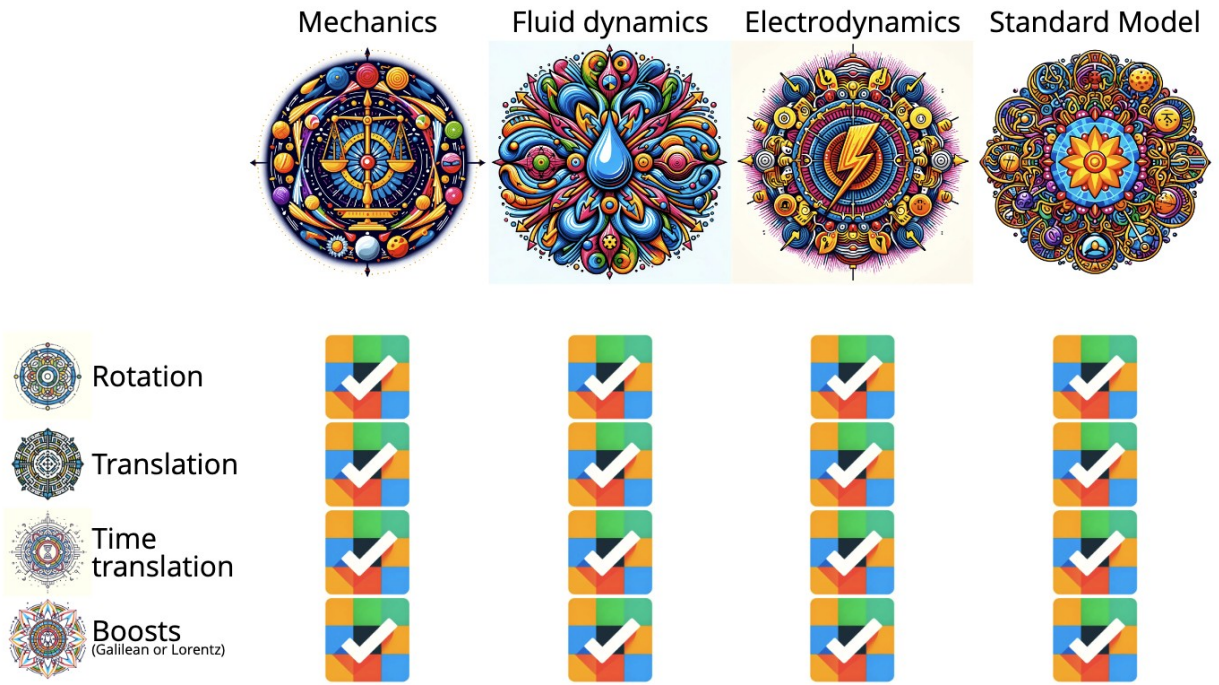
1.4 Mario



1.4 Learning Geometry & 3D Symmetries

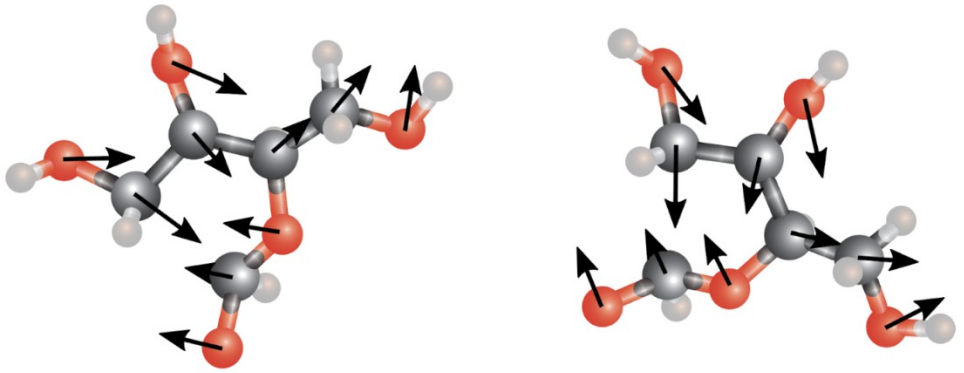
Professor: Mario Geiger Lecture
Recording: Coming soon...
Lecture Slides:
https://slides.com/mariogeiger/learning-geometry-3d-...

0 06/12/2024



Example of Equivariance

f : positions $\rightarrow$ forces



1.4 Mario

Where is equivariance used in AI?

Protein Folding

EquiFold Jae Hyeon Lee et al.

Protein Docking

DIFFDOCK Gabriele Corso et al.

Molecular Dynamics

Nequip S. Batzner et al.

MACE I. Batatia et al.

Allegro A. Musaelian et al.

Open Catalyst Project.

44M atoms while taking advantage of up to 5120 GPUs
Albert Musaelian

Solid State Physics

Prediction of Phonon Density Z. Chen et al.

Molecular Electron Densities

Cracking the Quantum Scaling Limit with Machine Learned Electron Densities J. Rackers

Cosmology, Medical Images and others

1.4 Mario

Group and Representations

"what are the operations" "how they compose"

"vector spaces on which the action of the group is defined"

Representation $D(g, x)$

- $g \in G, x \in V$
- Linear $D(g, x + y) = D(g, x) + D(g, y)$
- Follow the structure of the group
  $D(gh, x) = D(g, D(h, x))$

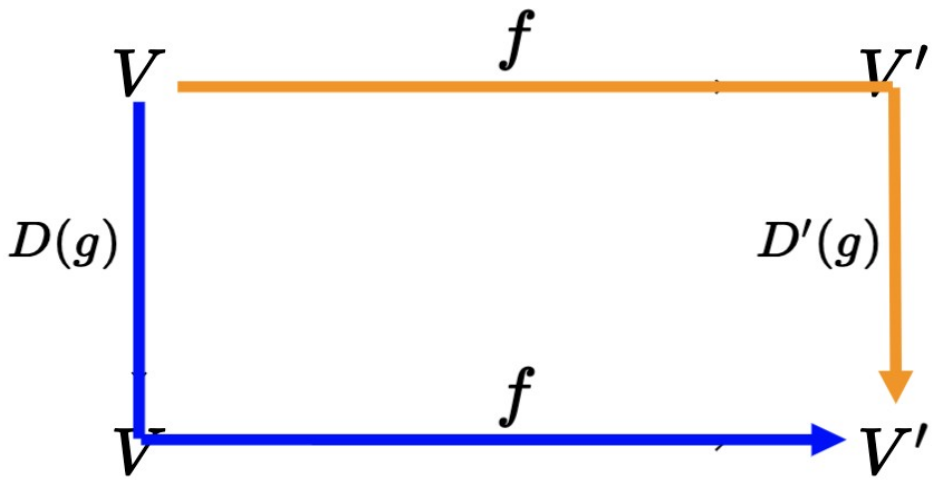
Equivalent notation $D(g)x$

- $D(g) : V \to V$
- $D(g) \in \mathbb{R}^{d \times d}$
- $D(gh) = D(g)D(h)$

Group G

- identity $\in G$
- associativity $g(hk) = (gh)k$
- inverse $g^{-1} \in G$

Equivariance



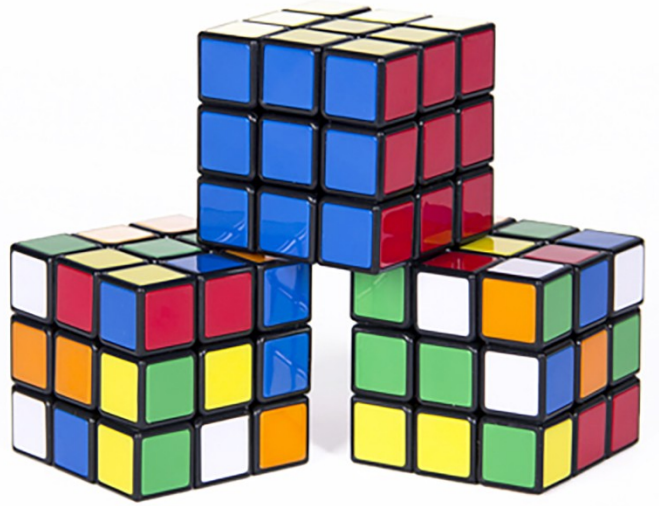
$$f(D(g)x) = D'(g)f(x)$$

1.4 Summary

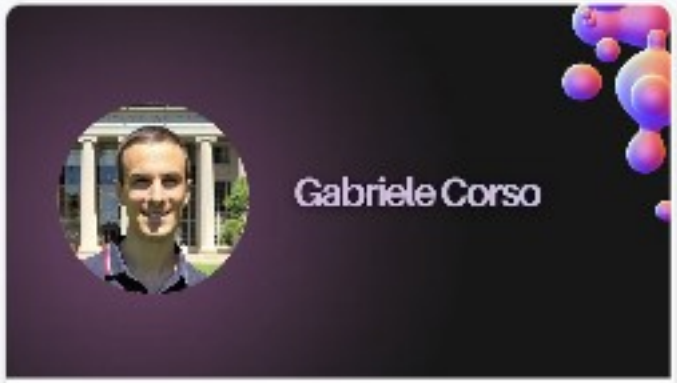
Equivariance is largely motivated by equivariance observed in Physics, e.g. Mechanics, Fluid Dynamics, Electrodynamics and the Standard Model

There is some evidence that equivariance leads to more data efficiency.
(There is also evidence against that)

Equivariance tends to introduce additional computational complexity, but that can be tackled with **reducible representations**.



2.1 Gabriele

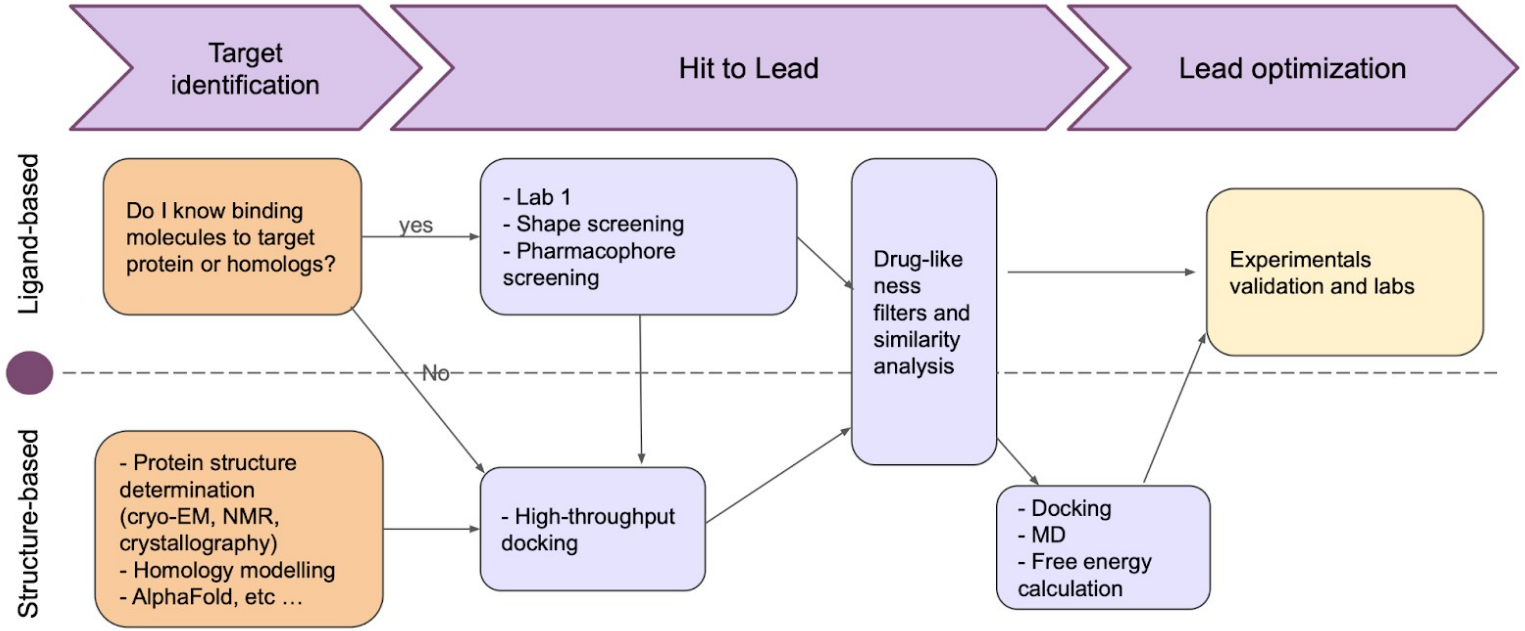


2.1 ML in Structure-Based Drug Discovery

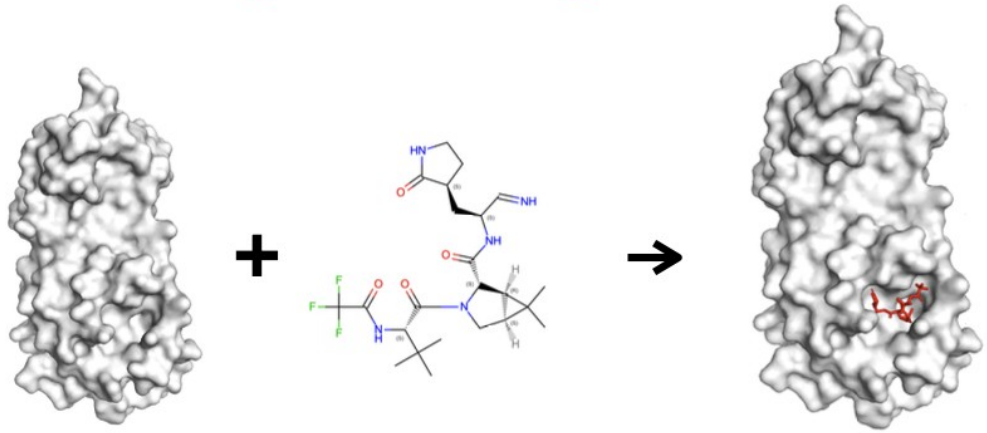
Professor: Gabriele Corso Lecture Recording: Lecture Slides:

10 06/13/2024

Ligand-based and Structure-based are interconnected



Protein-Ligand Docking



Input: protein + molecule **Output: bound structure**

2.1 Summary

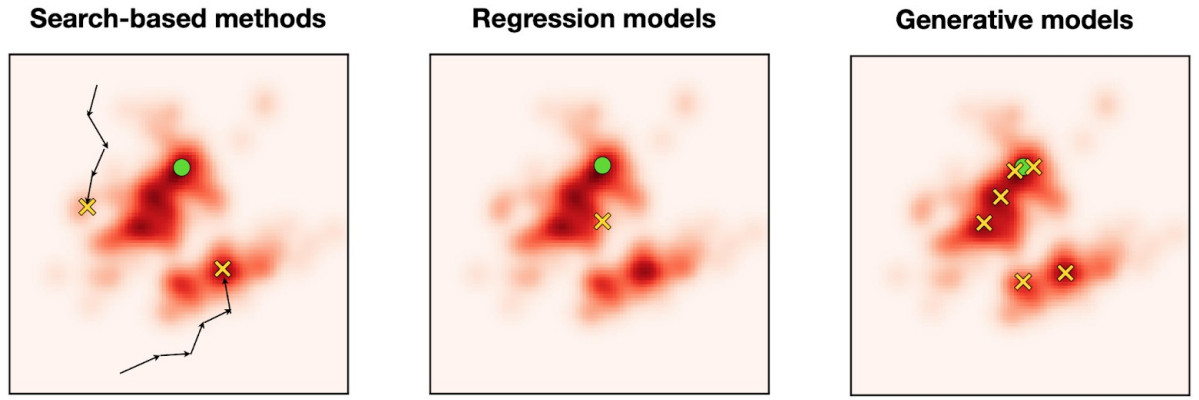
There are three approaches to docking: Search-based methods, Regression models and Generative models (see diagram)

Generative models are preferred (at least in his talk) due to **aleotoric uncertainty** introducing an averaging effect as well as issue with model uncertainty in comparison to regression.

There are three components to Diffusion Generative Models: forward diffusion, scoring and reverse diffusion. (see slide)

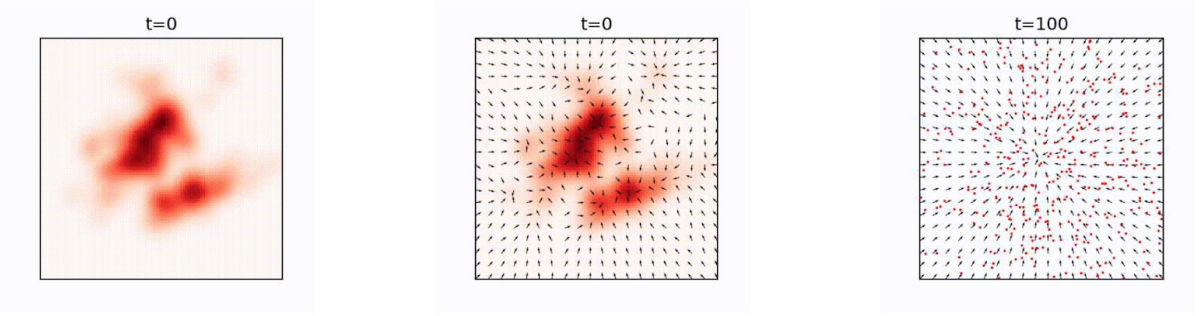
Docking Screening.

Different Approaches to Docking



Corso et al., ICLR 2023

Diffusion Generative Models



2.2 Gianni

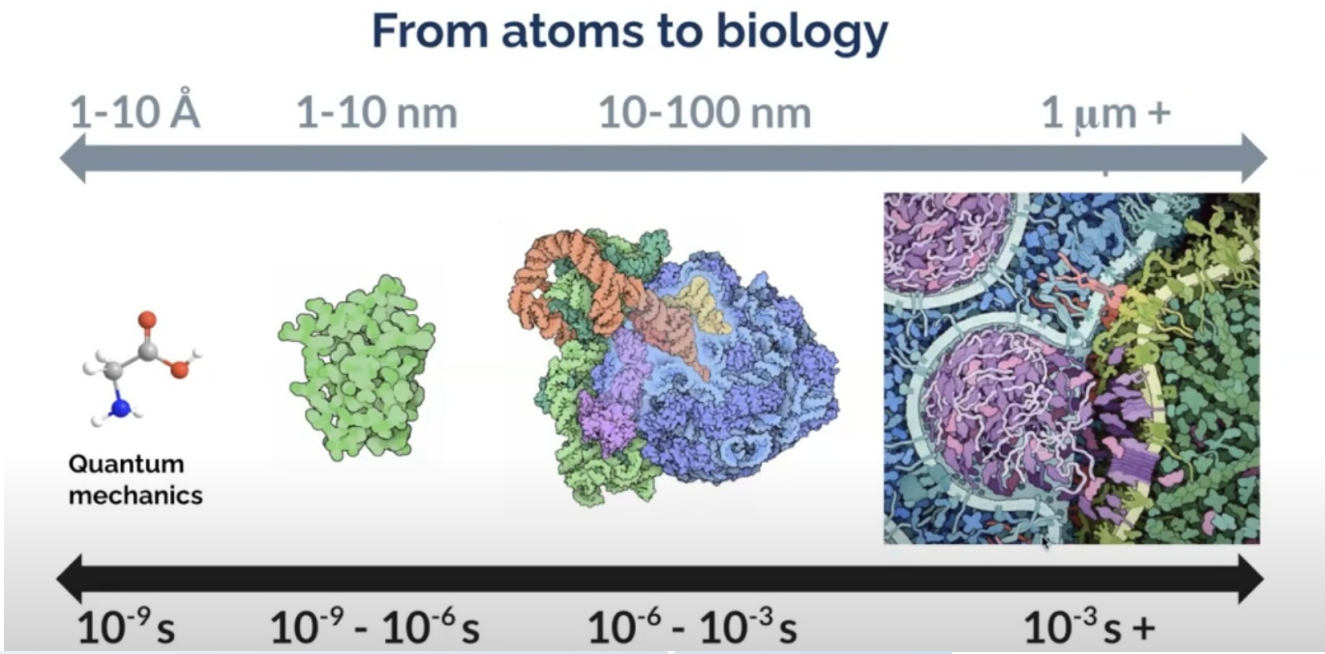
2.2 Learning ML Interatomic Potentials

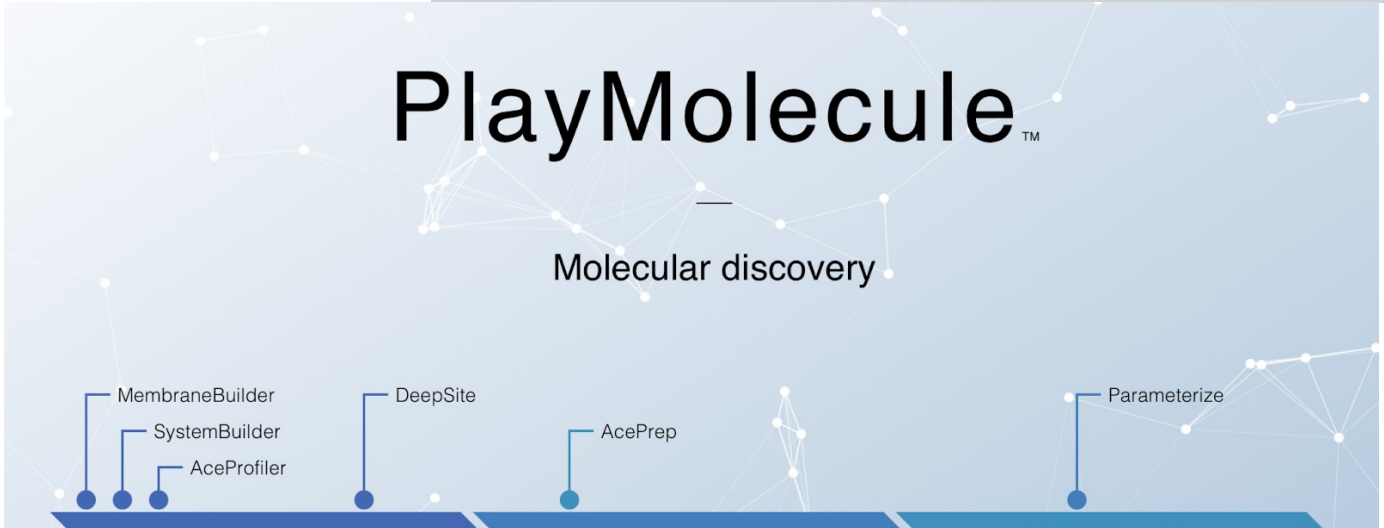
Professor: Gianni De Fabritiis
Lecture Recording: Coming soon... Lecture Slides:

0 06/13/2024

Acellera
https://www.acellera.com

Acellera: Accelerating Drug Discovery





2.2 Summary

His talk is a mixed bag (at least for me as an outsider) of his recent research results on:

Machine Learning Potential Architectures: All about learning energy functions which otherwise are not computable, see slide

- There are different type of atomic interactions: bonded and non-bonded.
- Energy functions describe the different types of interactions.
- The parameters of the functions constitute a molecular force-field .

$$U = \sum_{\text{All Bonds}} \tfrac{1}{2} K_b (b - b_0)^2 + \sum_{\text{All Angles}} \tfrac{1}{2} K_\theta (\theta - \theta_0)^2$$

$$+ \sum_{\text{All Torsion Angles}} K_\phi [1 - \cos(n\phi + \delta)]$$

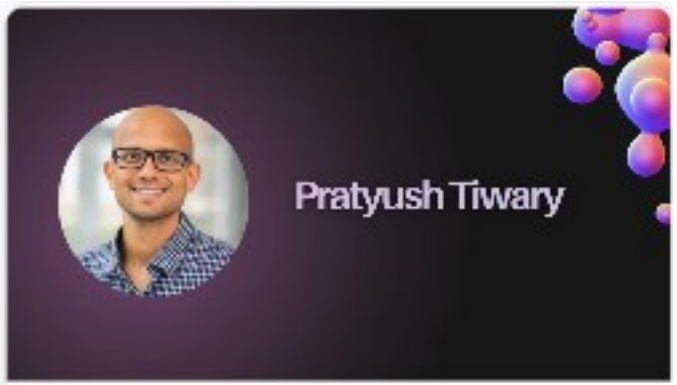
$$+ \sum_{\text{All nonbonded pairs}} \epsilon \left[\left(\frac{r_0}{r}\right)^{12} - 2\left(\frac{r_0}{r}\right)^{6} \right]$$

$$+ \sum_{\text{All partial charges}} 332 q_i q_j / r$$

Cartesian Tensor Representations for Efficient Learning of Molecular Potentials:
All about “Incorporating **transformation properties** of physical quantities into NNPs (Neural network potentials for chemistry)

Coarse-grained Potentials:
All about the “need to guarantee physical equivariances/invariances: translation, rotational, order” in coarse grained representations

2.3 Pratyush



2.3 Accelerate Atomistic Simulations, Sampling, and Dynamics

Professor: Pratyush Tiwary
Lecture Recording: Coming soon... Lecture Slides: PPT https://docs.google.com/presentation/d/1sODOQ8r9eGmQZbx-...

0 06/13/2024

With infinite compute, all we need is a single equation,
\but compute is not infinite, so :0

The 2nd most important formula in thermodynamics & statistical mechanics.

**In an ideal world with infinite compute power,
this is the only equation one needs**

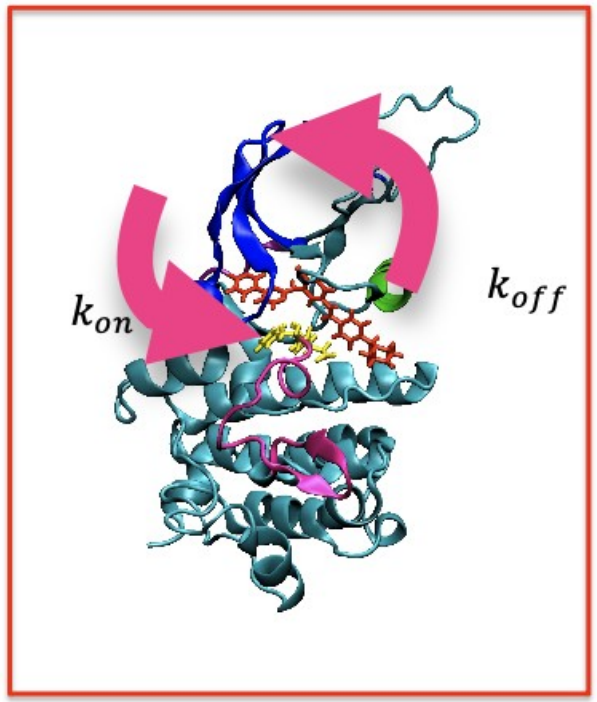
$$Z = \sum_i e^{-\beta E_i}$$

But the world isn't ideal and compute power is finite

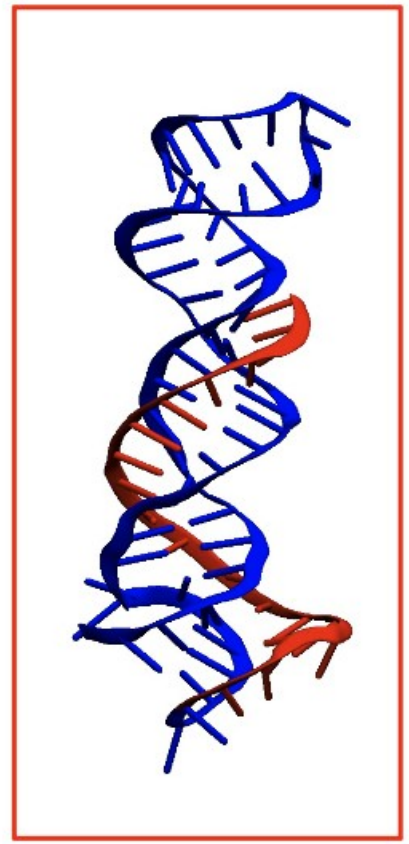
**Much of statistical mechanics (120+ years)
and machine learning (?? years) have been working around finding
tricks to sample the partition function *Z*,
or more generally, avoid sampling *Z*.
The 2 disciplines have much to teach other**

2.3 Summary

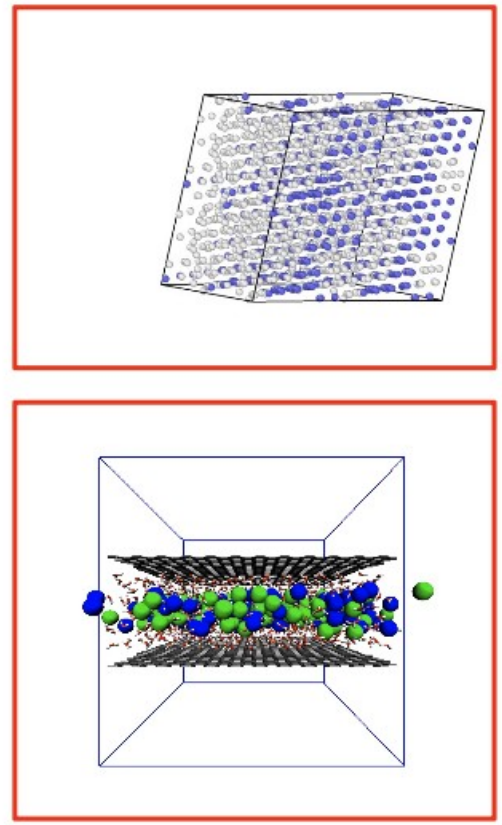
My lab combines statistical mechanics and generative AI to tackle problems of human health and energy relevance guided by structure & dynamics



Protein-small molecules drug discovery



RNA therapeutics



Finite-temperature crystal polymorphs & phase transitions

github.com/tiwarylab

AI can now be used to easily, routinely predict structure and dynamics

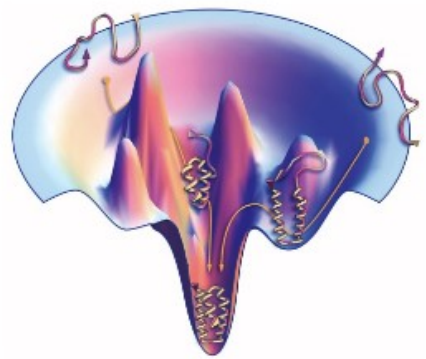
Molecular dynamics is dead

Not quite …

2.3 Summary

Life is not about a single structure but an ensemble of structures with just the right fluctuations across different length & timescales

Proteins

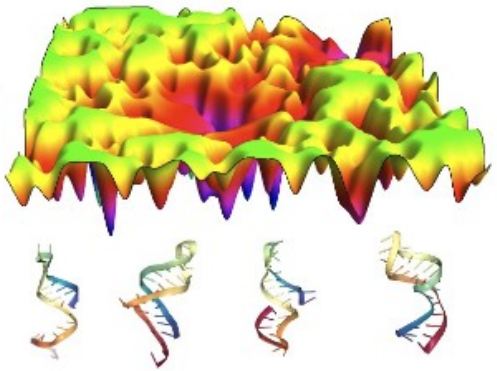


K. Dill, J. McCallum, Science (2012).
The Protein-Folding problem, 50 years on

Fluctuations between metastable states described by few slow modes & timescale separation

Order/Disorder spectrum

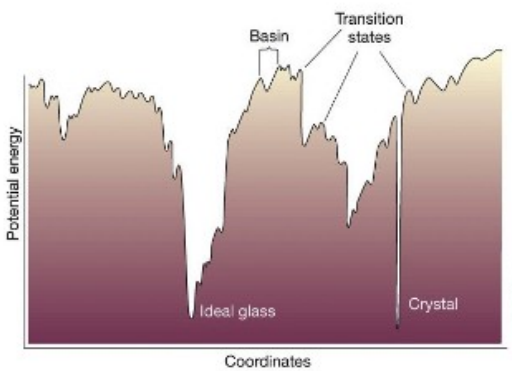
RNAs



D.J. Wales, Ann. Rev. Phys. Chem. (2017).
Exploring Energy Landscapes

No obvious timescale separation & no dominant driving fluctuations

Somewhere on the order/disorder spectrum

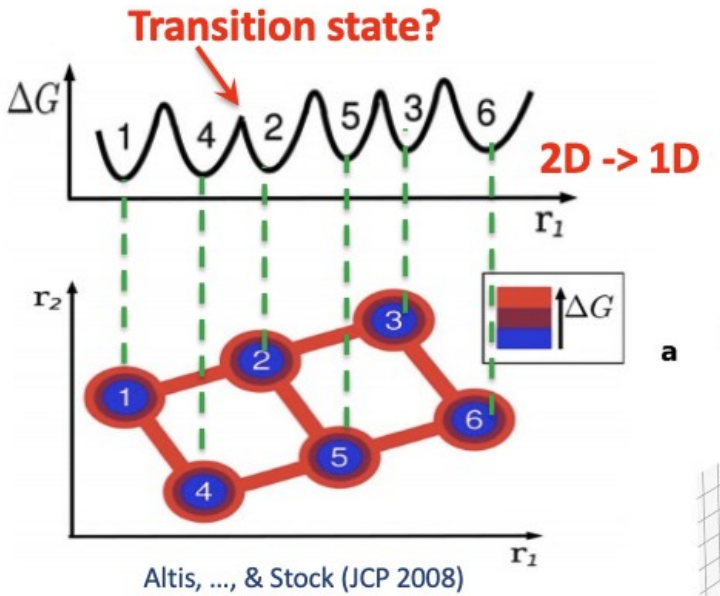


Beyerle, Zou, Tsai, Tiwary PNAS 2023

Debenedetti, Stillinger, Nature (2001).
Supercooled Liquids & the Glass Transition

2.3 Summary

All reduced representations are correct
Some are useful
Few are meaningful

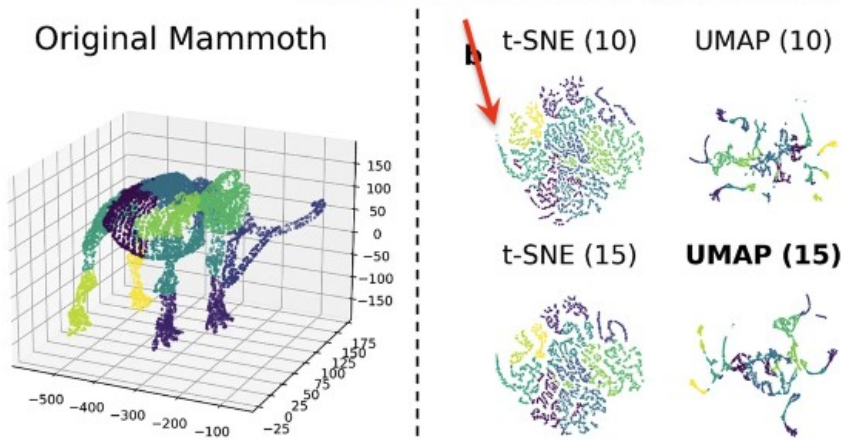


Altis, ..., & Stock (JCP 2008)

Tribello & Gasparotto (Frontiers in Mol. Bio. 2019)

3D -> 2D

Guess what this represents?



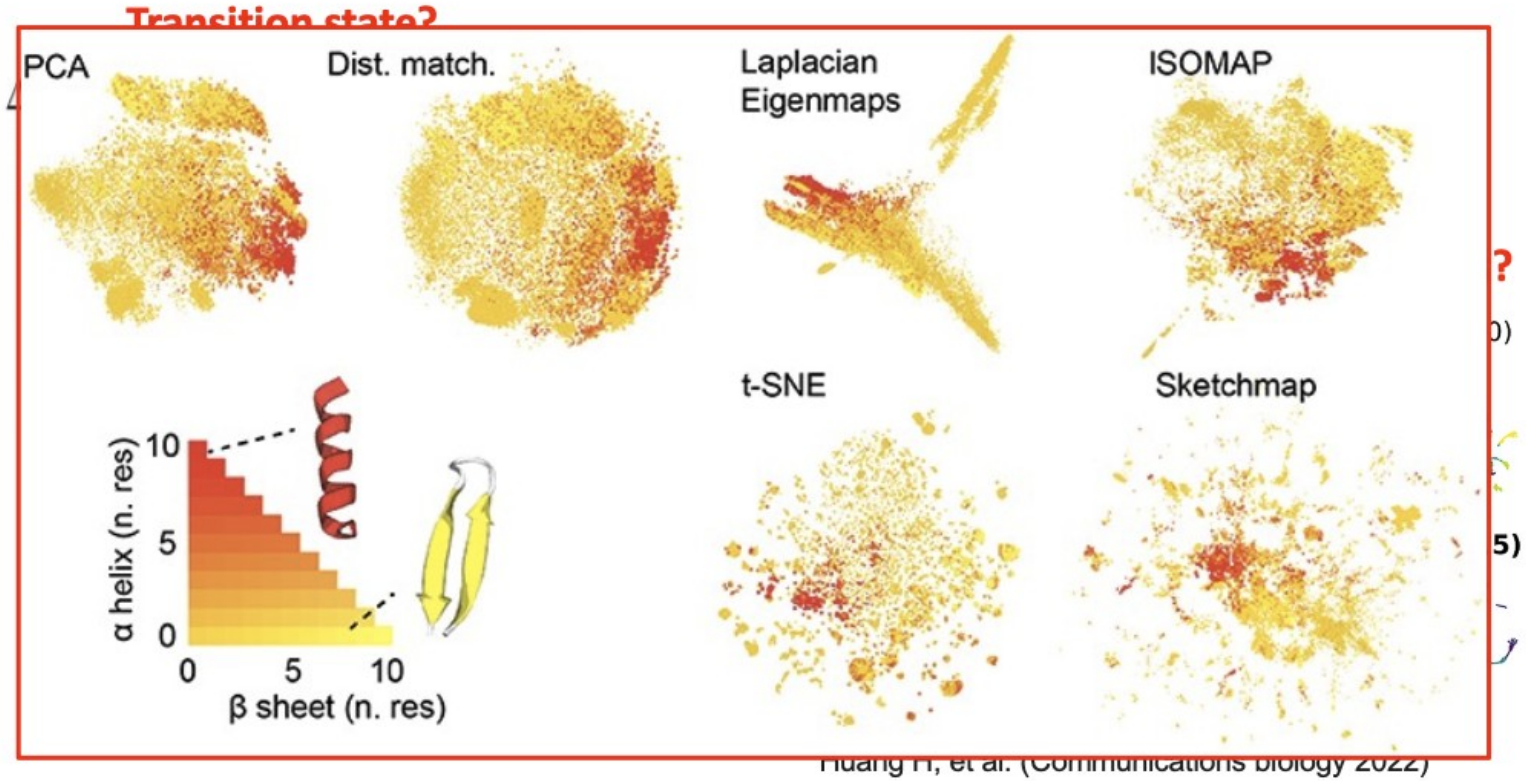
Huang H, et al. (Communications biology 2022)

Why should you trust these projections?

N-D -> 2D

2.3 Summary

All reduced representations are correct
Some are useful
Few are meaningful



Tribello & Gasparotto (Frontiers in Mol. Bio. 2019)

Why should you trust these projections?

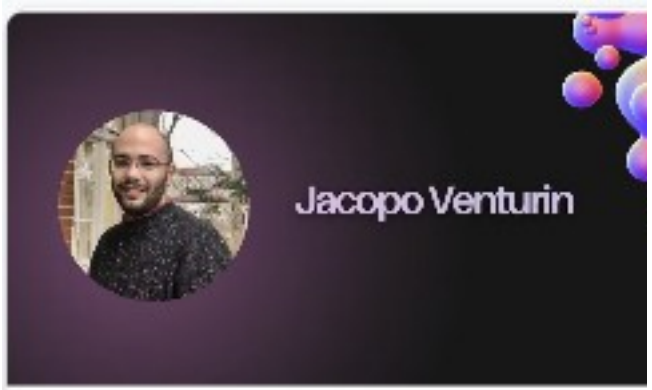
N-D -> 2D

2.3 Summary

Take-home:
Molecular Dynamics – 60 years old and still very relevant

- AI tools are efficient hypothesis generators that when combined with physics based simulators – especially MD - in a rigorous feedback loop might have predictive power
- A 100ns MD trajectory is not always 10x more data than 10ns. Important to understand metastable state to state emergent dynamics
- Careful enhanced MD generates correct fluctuations across length and timescales, critical for training AI models that can mimic reliable molecular conformational diversity and dynamics, & also for next gen forcefields
- Showed 2 methods today – RAVE and path-sampling LSTM
- Demo codes open-source @ github.com/tiwarylab

2.4 Jacopo



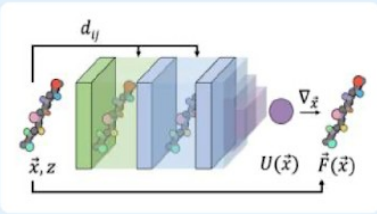
2.4 Coarse-Grained Biological Systems

Professor: Jacopo Venturin
Lecture Recording: Coming soon... Lecture Slides:

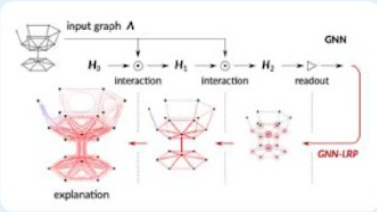
3 06/13/2024

Outline

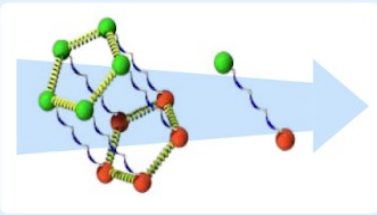
Freie Universität



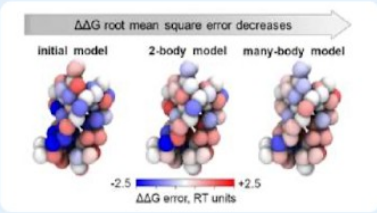
Transferable ML-CG force fields for proteins



Interpreting ML-CG force-fields



ML-CG force-fields for Path Integral Molecular Dynamics

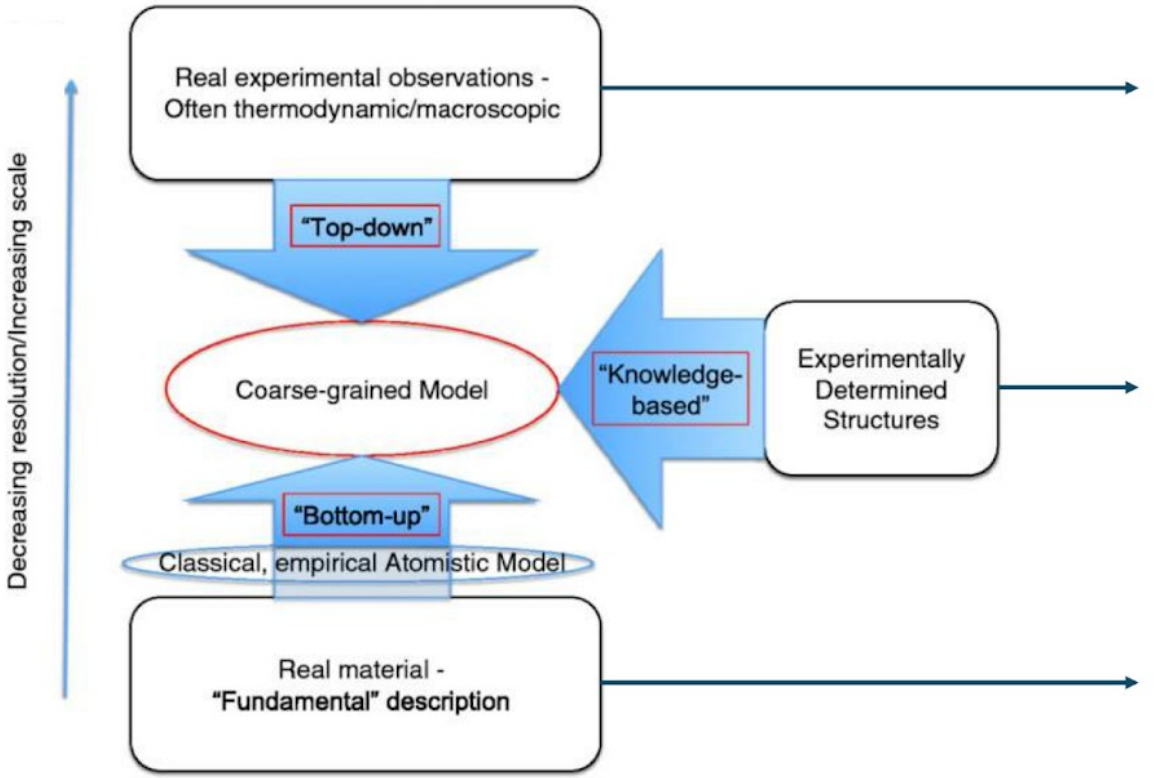


Incorporating experimental data in CG force-fields

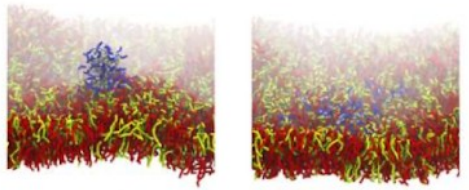
2.4 Summary

Coarse-graining philosophies and approaches

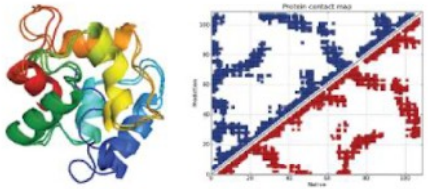
Freie Universität Berlin



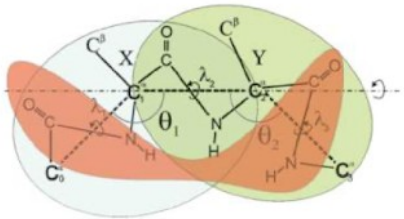
MARTINI



AWSEM



MSCG, UNRES



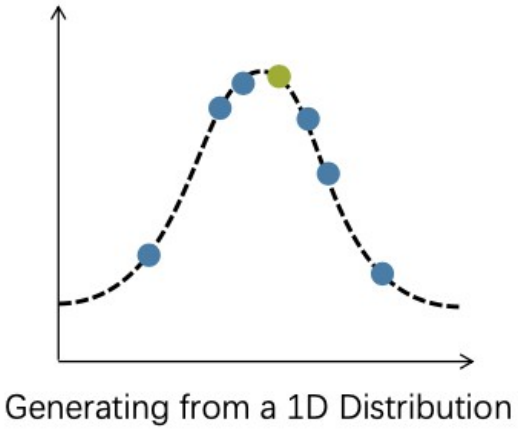
Noid, W. G. (2013)., *J. Chem. Phys*, 139(9), 90901 (2005)
Marrink, S. J. *et al, J. Chem. Phys. B*, 111(27), 7812-7824 (2007)
Davtyan, A. *et al, J. Chem. Phys. B*, 116(29), 8494-8503 (2012)
Liwo, A. *et al, J. Mol. Model.* 20, 2306 (2014)
Thorpe, I. F., Zhou, J., & Voth, G. A., *J. Phys. Chem. B*, *112*(41), 13079–13090 (2008).

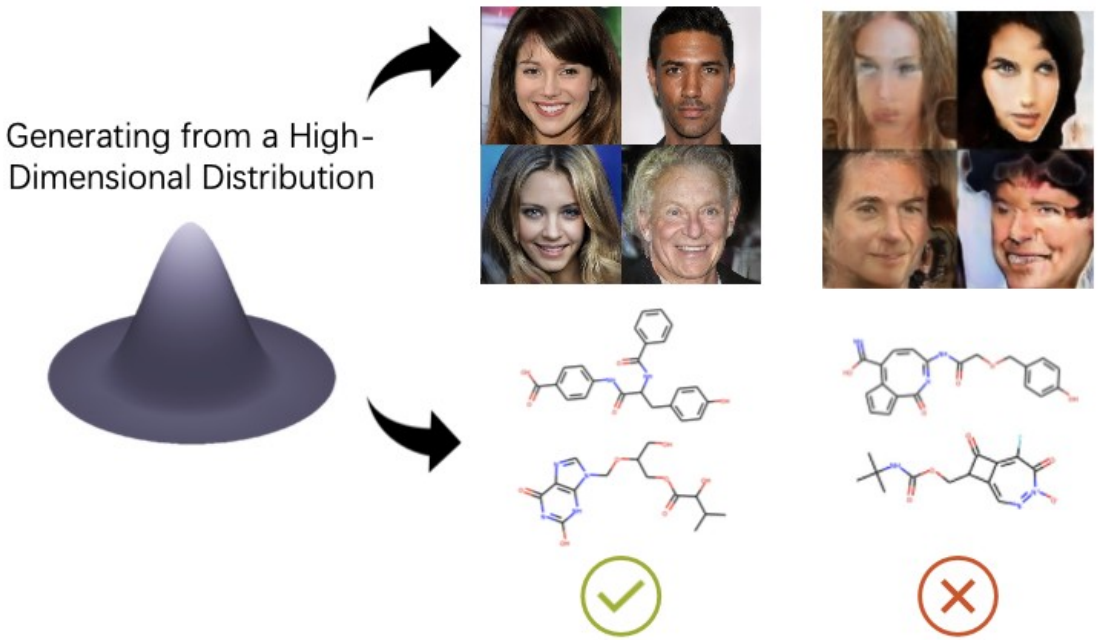
3.1 Camille

Today seems to be a special day with many high-profile speakers.

What is a generative model?

A generative model **learns a distribution** from training examples at training time and **draws new examples** from this learned distribution at test time





arras et al. ICLR 2018
ómez-Bombarelli et al. ACS Cent. Sci. 2018

The mechanics of how to generate molecules

How do we train a neural network to learn a distribution?

Objective: We want a model that generates data such that the distribution of the generated data **matches the distribution of the input data as closely as possible**.

Variational Autoencoder

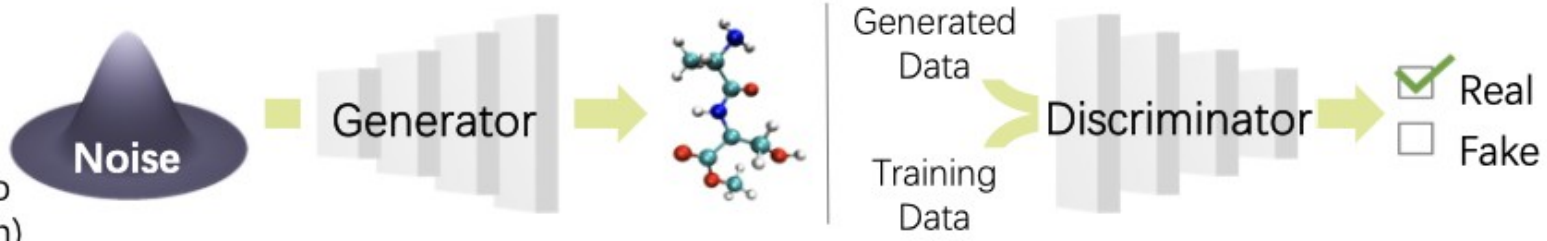
Learn the parameters of a normalized function that approximates P(x)

Normalizing Flow Model

Learn the parameters of a series of invertible, normalized transformations

Generative Adversarial Network

Discriminator Imposes Distribution Matching (no explicit P(x) approximation)



Kingma & Welling ICLR 2018, Goodfellow et al. NeurIPS 2014, Shi et al. ICLR 2020

3.1 Summary

We can cast molecular discovery as an optimization problem

Human Version:

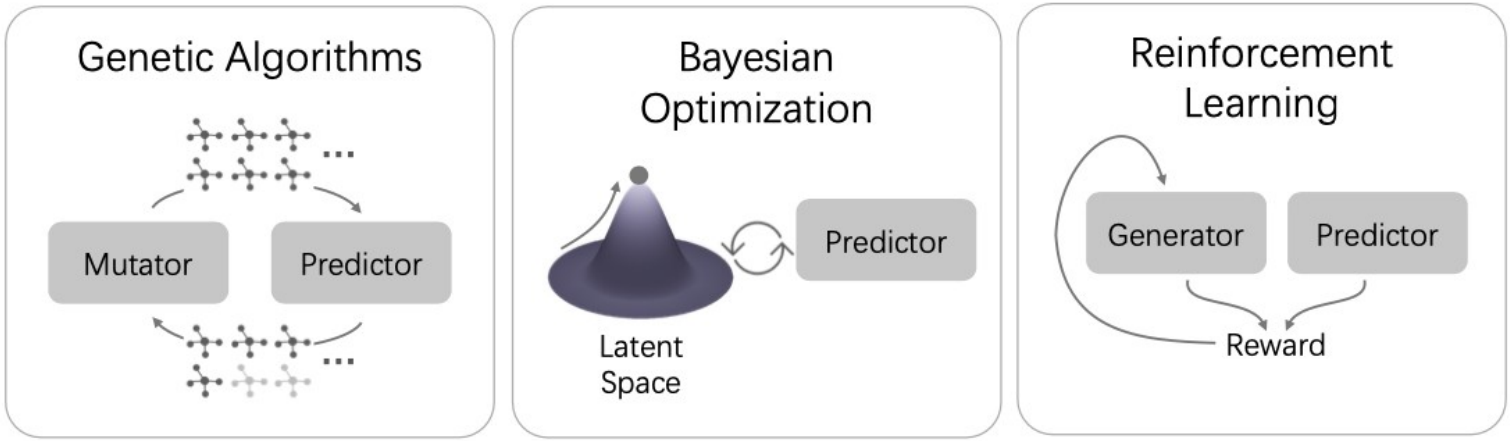
To turn this into an algorithmically well-defined problem we need:

1) A design space
2) Algorithms for making moves in that design space
3) An objective function to guide those algorithms

Navigating the Molecular Design Space

The molecular design space is inherently discontinuous, but can be mapped to a continuous space using generative models:

Some Optimization Methods:

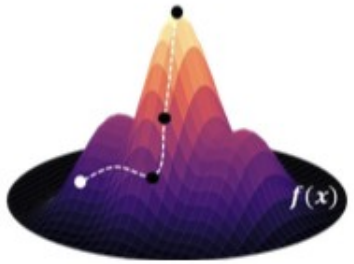


3.1 Summary

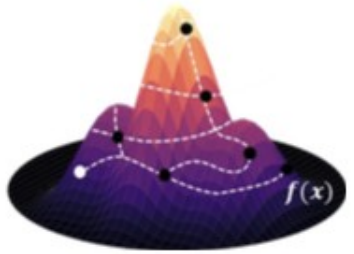
Generative modeling workflows rely on predictive models with limited certainty

Solution: Incorporate information about uncertainty into your objective function

- Bayesian optimization methods yield probability distributions instead of discrete predictions making enabling straightforward uncertainty characterization
- Other uncertainty characterization methods can be incorporated directly into the objective function
  - For neural networks these can include methods such as Ensemble Variance or Monte Carlo Dropout



Exploitation (Risk Avoiding)



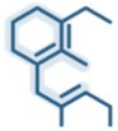
Exploration (Risk Seeking)

Embrace uncertainty!

Note: Optimizing with a bad predictive model can limit exploration; it is better to avoid explicitly optimizing with bad models

The science of how to discover molecules

a bit technical again~

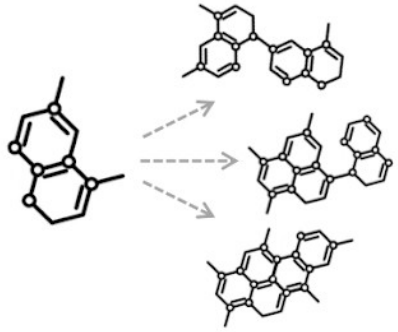


Structure-Constrained Molecular Generation

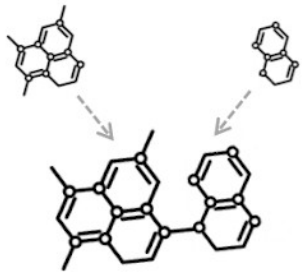
Goal: Take an input scaffold or structure and build on it to generate a new example.

Structure-constrained molecular generation **can be unconstrained OR property constrained**.

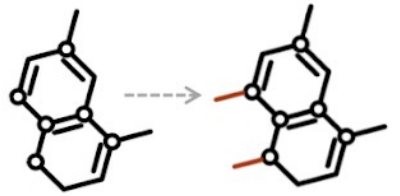
Scaffold/Fragment Expansion



Scaffold/Fragment Merging



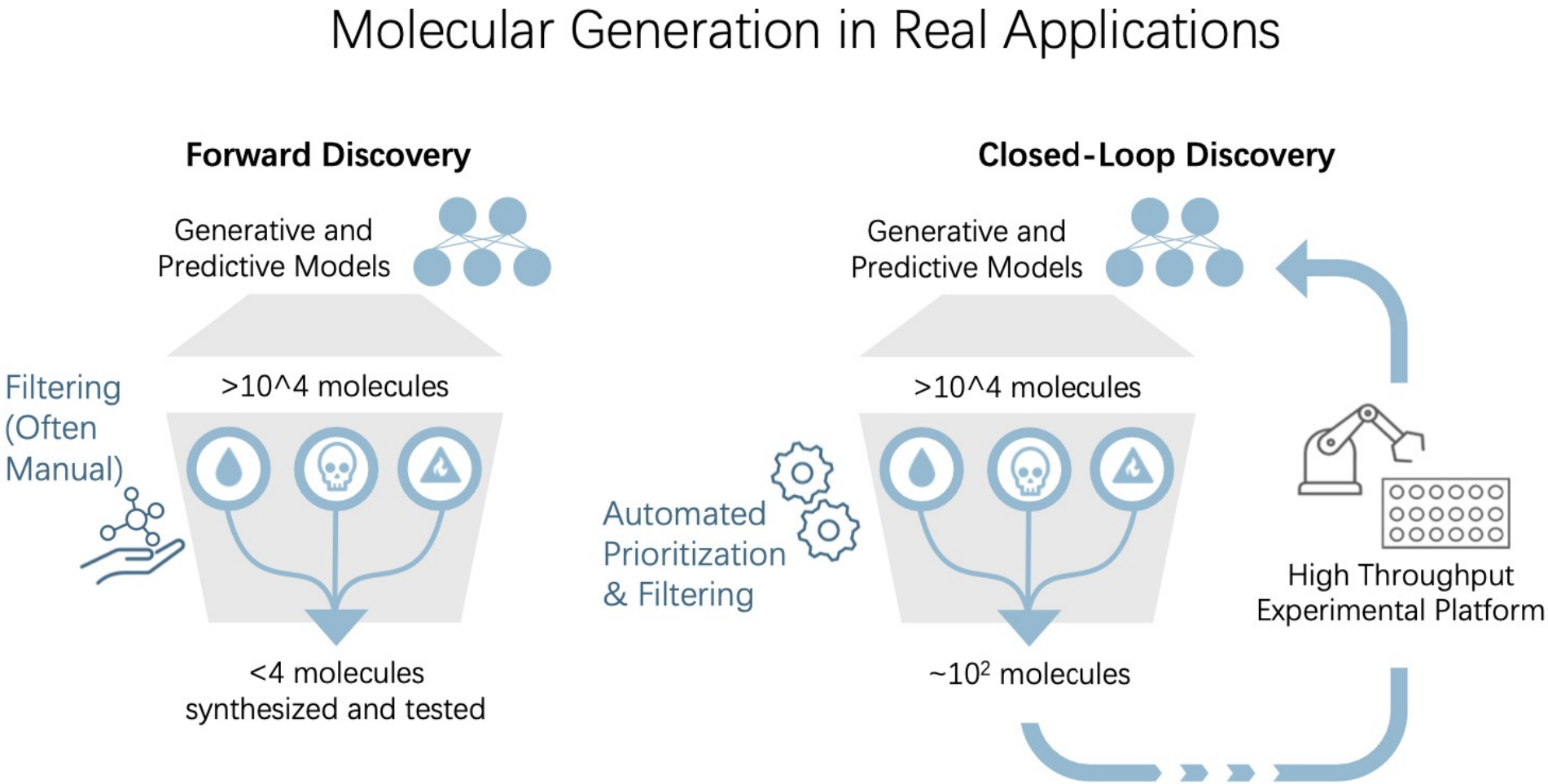
Molecular Translation



Langevin et al. J. Chem. Inf. Model. 2020, Li et al. J. Chem. Inf. Model. 2020, Jin et al. PMLR 2020, Imrie et al. J. Chem. Inf. Model. 2020, Jin et al. ICLR 2019

3.1 Summary

The art of how to discover molecules



3.1 Summary

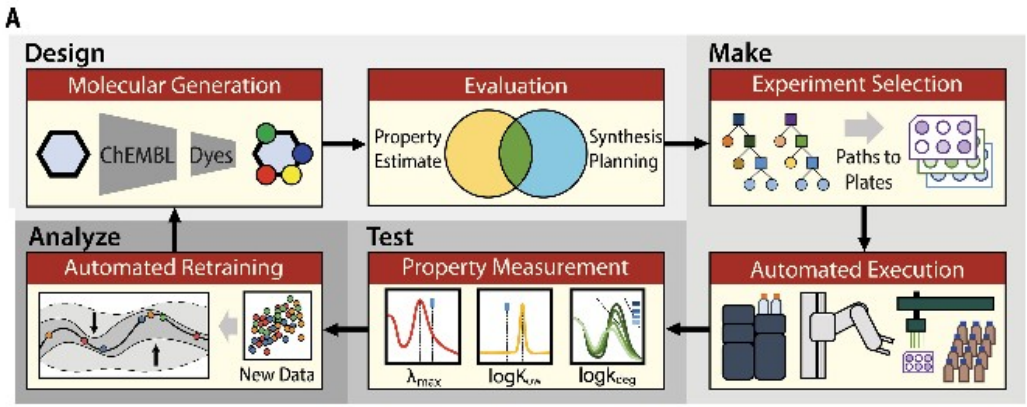
The art of how to discover molecules

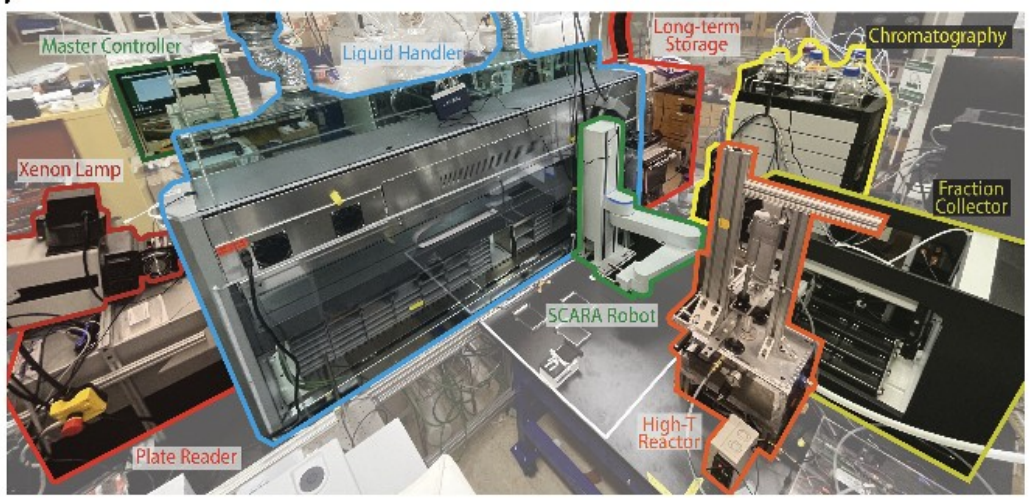
Real Application: Closed Loop Discovery

Science

Autonomous, multiproperty-driven molecular discovery: From predictions to measurements and back

BRENT A. KOSCHER, RICHARD B. CANTY, MATTHEW A. MCDONALD, KEVIN P. GREENMAN, CHARLES J. MCGILL, CAMILLE L. BILODEAU, WENGONG JIN, HAOYANG WU, FLORENCE H. VERMEIRE, BROOKE JIN, TRAVIS HART, TIMOTHY KULESZA, SHIH-CHENG LI, TOMMI S. JAAKKOLA, REGINA BARZILAY, RAFAEL GÓMEZ-BOMBARELLI, WILLIAM H. GREEN, AND KLAVS F. JENSEN fewer Authors Info & Affiliations





3.2 Yoshua

1) It’s a needle in the haystack problem

Discovery: Searching Needles in a Haystack

- ***The space of experiments and the space of theories are huge***
- *Can’t enumerate*
- ***Tiny fraction will work***
- ***How to get all good ones?***
- ***Amortized inference can help us to efficiently represent posteriors over theories, sample candidate experiments and reason about outcomes***

3.2 Summary

1) It’s a needle in the haystack problem

In-Silico Design - Desiderata

We aim to combine several metrics:

- **Performance:** Measure of the desired effect of the generated candidates

$$\text{Mean}(\mathcal{D}) = \frac{\sum_{(x_i,y_i)\in\mathcal{D}} y_i}{|\mathcal{D}|}$$

- **Diversity:** Measure of the variety in generated candidates

$$\text{Diversity}(\mathcal{D}) = \frac{\sum_{(x_i,y_i)\in\mathcal{D}} \sum_{(x_j,y_j)\in\mathcal{D}\setminus\{(x_i,y_i)\}} d(x_i, x_j)}{|\mathcal{D}|(|\mathcal{D}| - 1)}$$

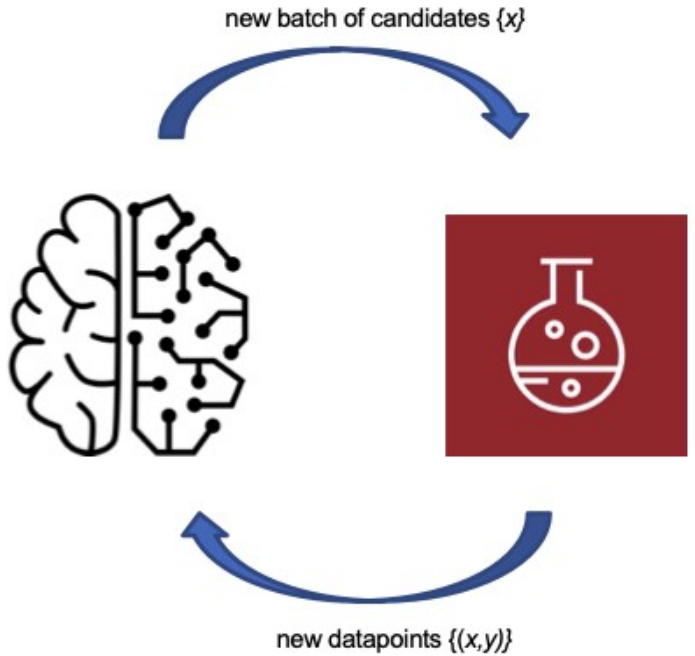
- **Novelty:** Measure of novelty with respect to known candidates

$$\text{Novelty}(\mathcal{D}) = \frac{\sum_{(x_i,y_i)\in\mathcal{D}} \min_{s_j\in\mathcal{D}_0} d(x_i, s_j)}{|\mathcal{D}|}$$

from M. Jain

Active Learning, Growing Dataset, ML and Assays Feedback Loop

- ML system proposes next batch of candidates {*x*}
- Next-stage "oracle" provides more or less precise answers *y* per *x*
- Dataset is augmented with new set {(x,y)} of pairs
- ML system retrained

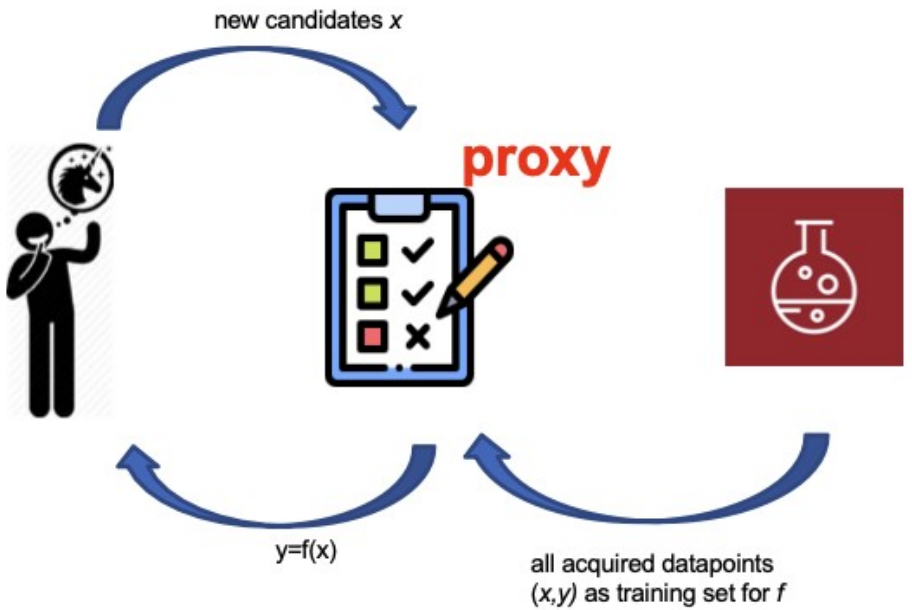


Mila

3.2 Summary

Generating Candidates & Evaluating Them

Generating candidates can be obtained with a **generative model** (that we can draw samples from) trained to produce a diverse set of good candidates, where "good" is defined by an **evaluation function** trained with the data collected from a downstream "oracle", based on a **proxy** for the true property of interest



Mila

3.2 Summary

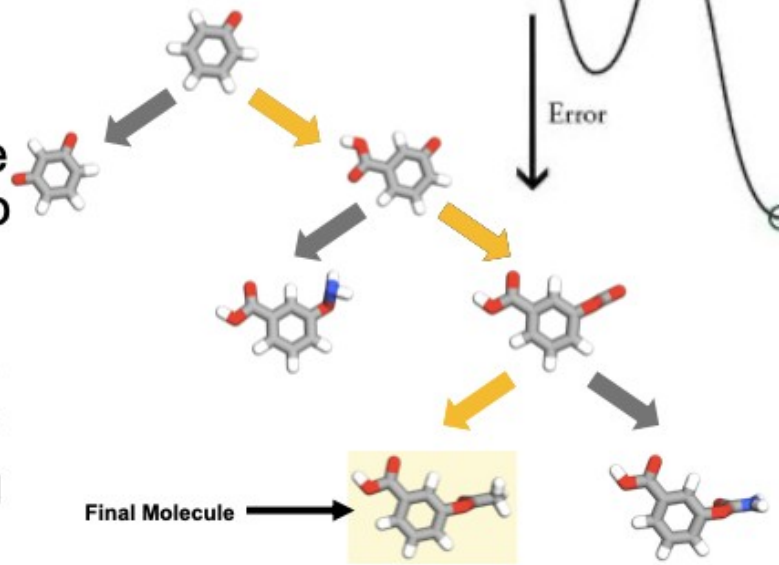
https://yoshuabengio.org/gflownet-tutorial

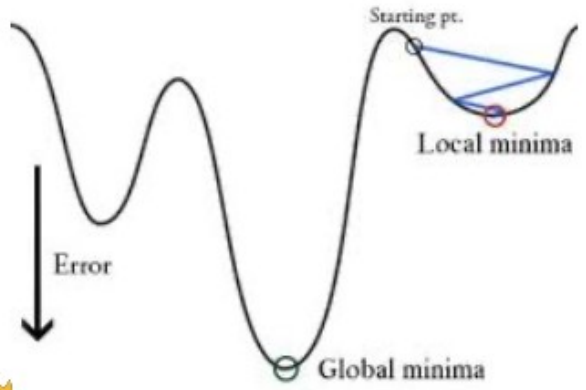
Papers at NeurIPS 2021, ICML 2022, UAI 2022, NeurIPS 2022, ICLR 2023, ICML 2023, NeurIPS 2023, ICML 2024 + ArXiv

Mila

Local discrete search

We can create a discrete action space which corresponds to transforming a molecule into another by adding or removing an atom or a frequently occuring rigid block of atoms.





Similarly we can search in the space of peptides by adding one amino-acid at a time

Black-Box ~~Optimization~~ Exploration & Scientific Discovery

- We have access to an expensive oracle (real-world), outputs $y=f(x)$
- We can only call it a few times, each round with batched queries B
- We look for a set of x's whose $f(x)$ is large (modes of f)

- Examples:
  - Discovering new drugs
  - Discovering new materials
  - Experimental design of experiments
  - Reasoning & causal discovery:
    discovering good explanations & causal models: -log p = energy fn

Mila

3.3 Connor

3.3 Synthesizability & Molecular Synthesis

Professor: Connor Coley Lecture Recording: Lecture Slides:

0 06/14/2024

The formulation of molecular optimization

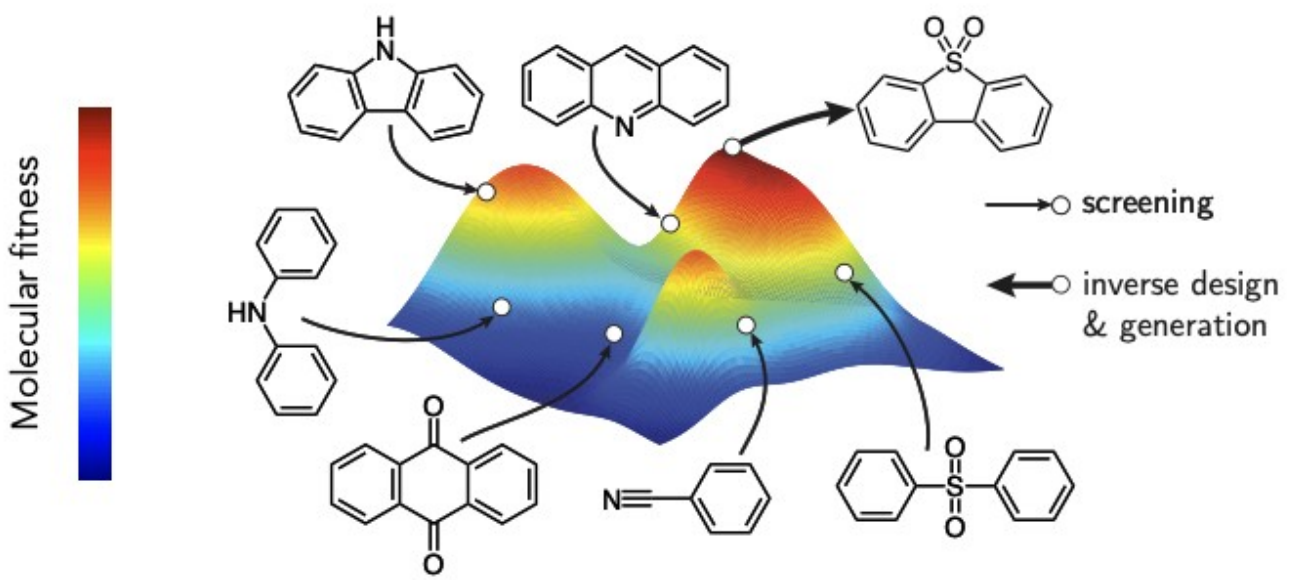
best design

oracle

$$x^* = \arg\max_{x \in \mathcal{X}} f(x)$$

all possible molecules

Molecular design and the ideation of structures

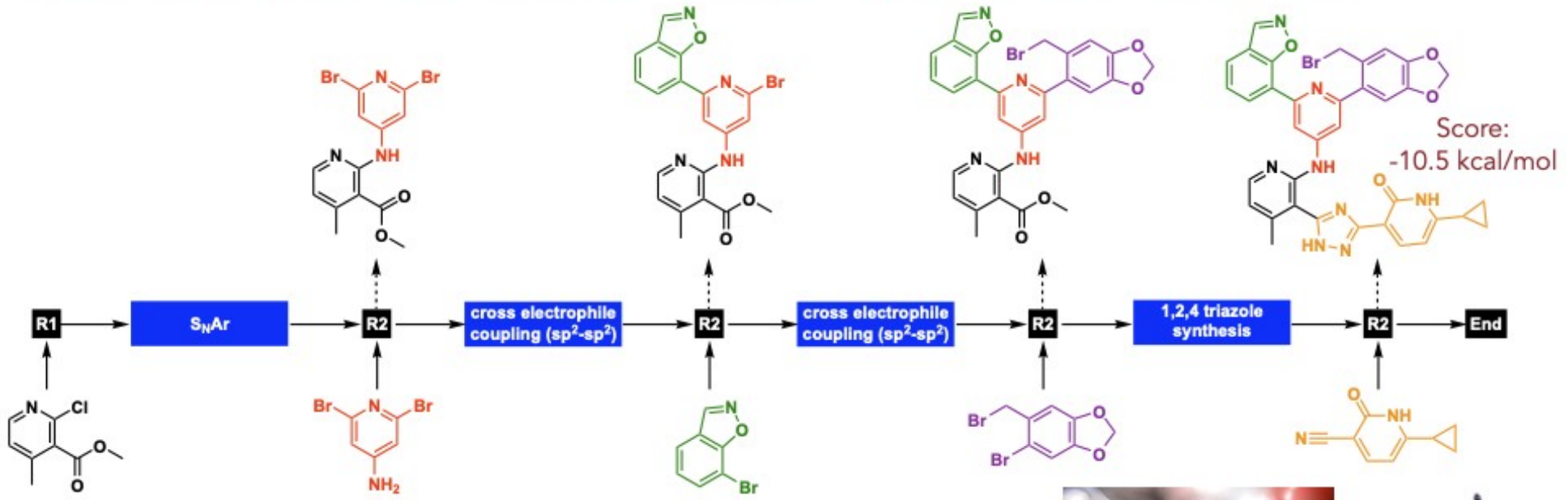


Idealized structure-property surface

3.3 Summary

Building molecules by building synthetic trees

Optimization objective: design a molecule with a good docking score against M^{pro} of SARS-CoV-2



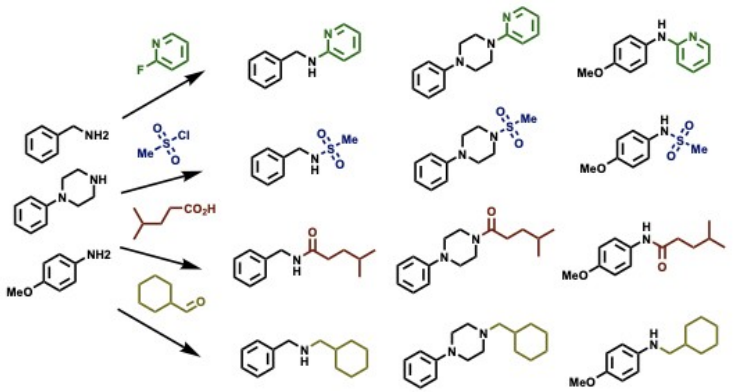
Gao, Mercado, and Coley, *ICLR* 2022

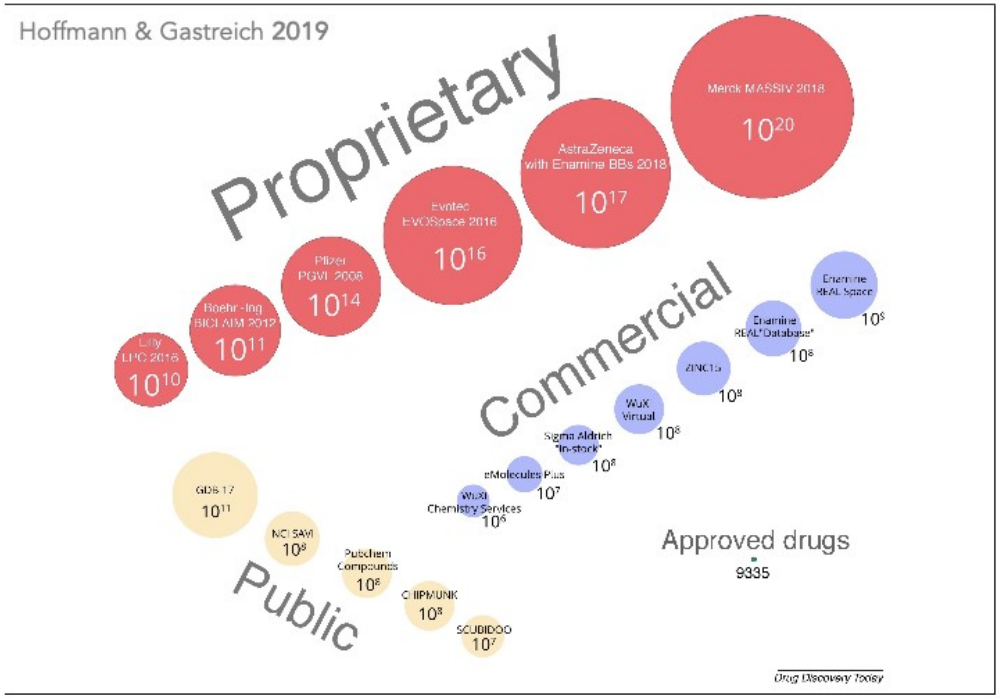
3.3 Summary

What is the chemical space that these models search?

O(100,000) commercial building blocks x *O(100) expert-defined reaction templates* x *1-10 reaction steps*

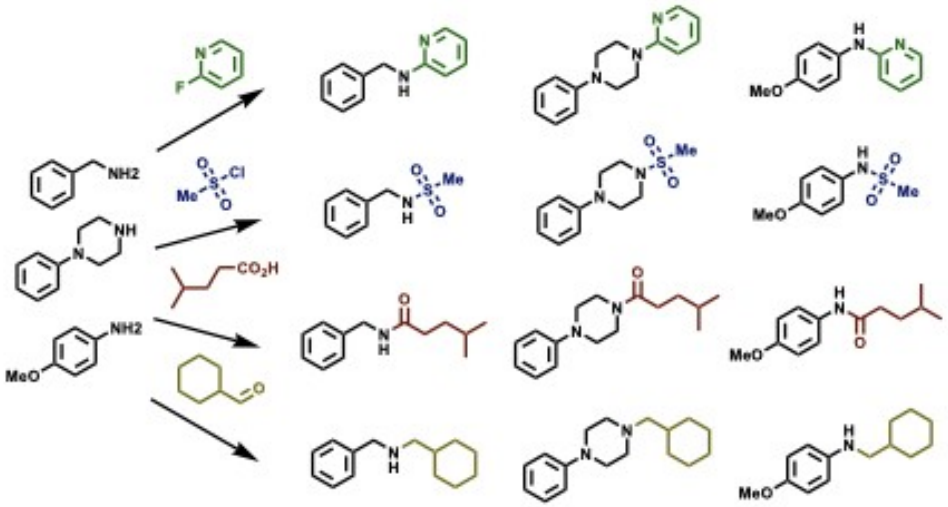
These types of models search the same kinds of chemical spaces as "make-on-demand" virtual libraries *but avoid explicit enumeration*



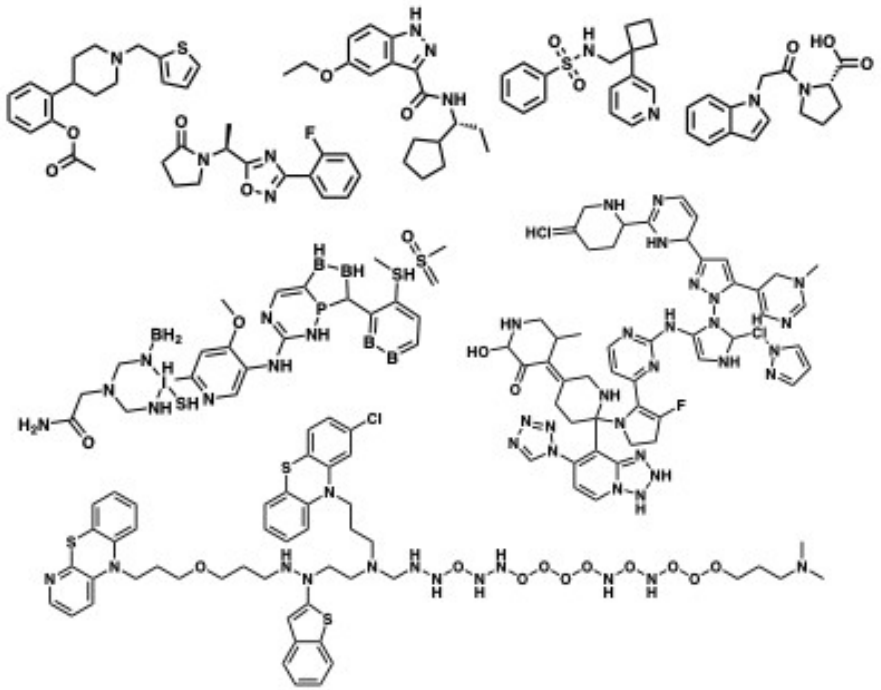


Summary: don't overlook synthesis in molecular design

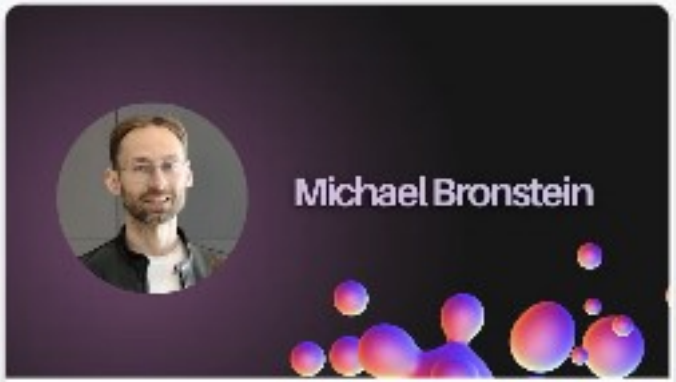
Virtual libraries are often "make-on-demand" libraries enumerated using chemical transformation rules we believe to be robust



Generative models produce new compounds for which we must plan synthetic routes unless already constrained by synthesis



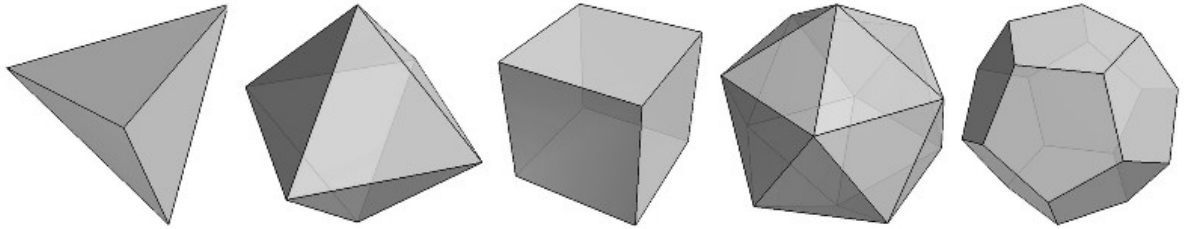
3.4 Michael



geometric deep learning
graph representation learning

"Symmetry, as wide or as narrow as you may define its meaning, is one idea by which man through the ages has tried to comprehend and create order, beauty, and perfection"

H. Weyl



"Platonic solids"

Plato

370 BC

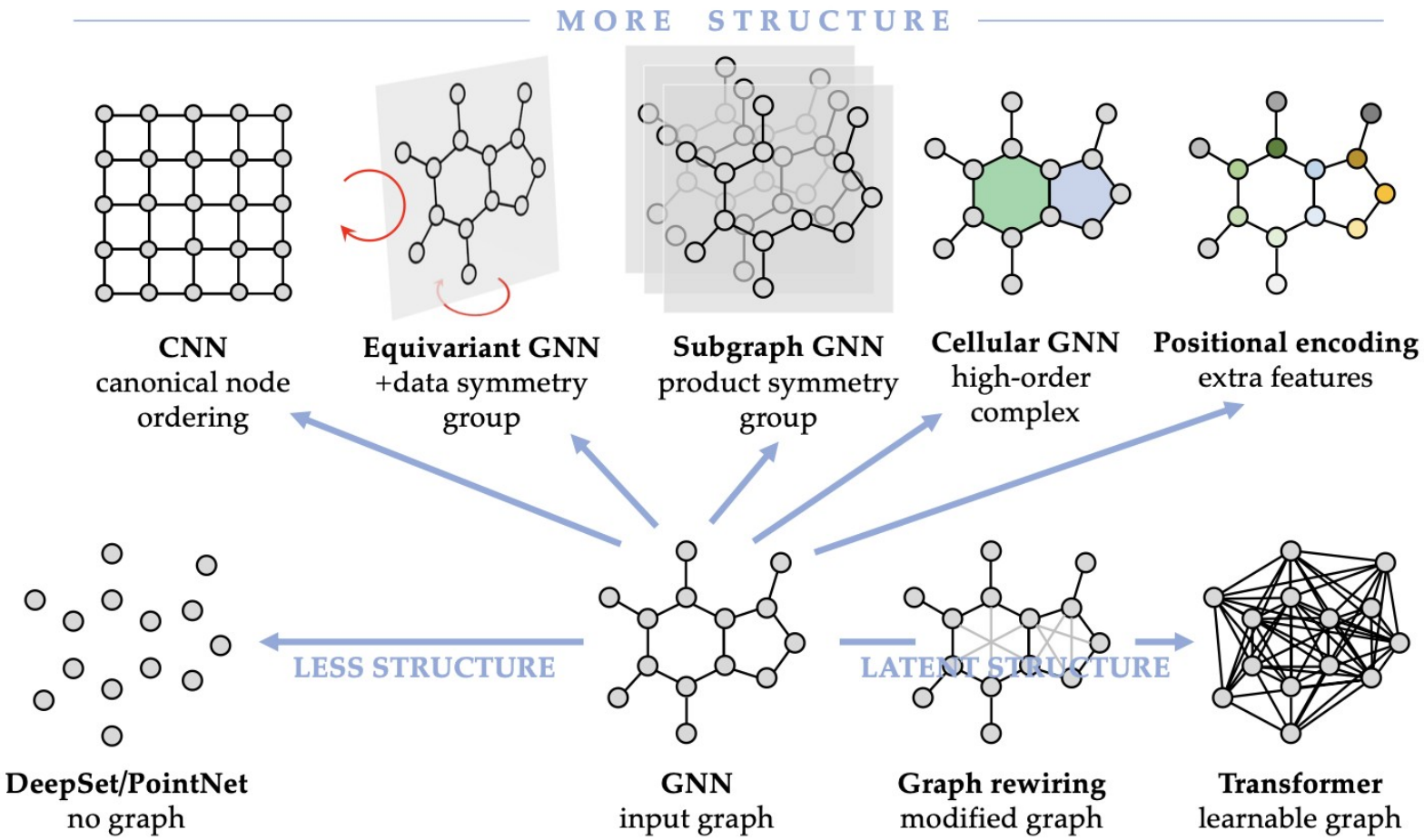
"It is only slightly overstating the case to say that Physics is the study of symmetry"

— *More is different*

P. Anderson

3.4 Summary

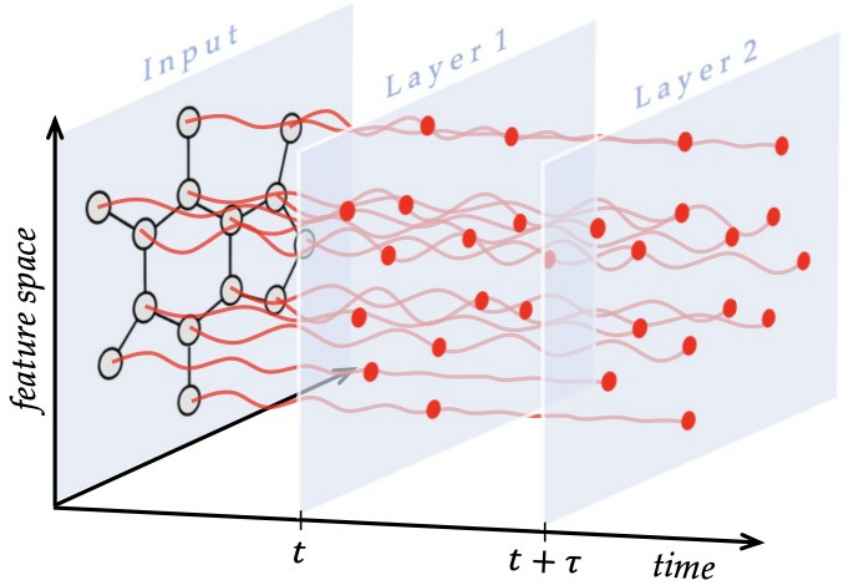
GNNs is (possibly) the parent class of all NNs,
developed around the idea of geometric symmetry：



3.4 Summary

Physics-inspired GNNs are also cool：

Physical metaphor of Graph ML



- GNN = dynamic system
- layers = discretisation of time
- graph = coupling function (discretisation of space)

$$\mathbf{X}(t+\tau) = \mathbf{X}(t) + \tau \mathbf{F}_{\boldsymbol{\theta}(t)}(\mathbf{X}(t), \mathcal{G})$$

Haber, Ruthotto 2017; Chen et al. 2019 (Neural ODEs); Xhonneau et al. 2020 (CGNN); Chamberlain, Rowbottom, et B. 2021 (GRAND, BLEND)
Eliasof, Haber 2021 (PDE-GCN); Di Giovanni, Rowbottom et B 2022 (GRAFF), Rusch et B 2022 (GraphCON)

3.4 Summary

Heat Diffusion

Newton Law of Cooling:
"the [temperature] a hot body loses in a given time is proportional to the temperature difference between the object and the environment"

Anonymous 1701


(824)

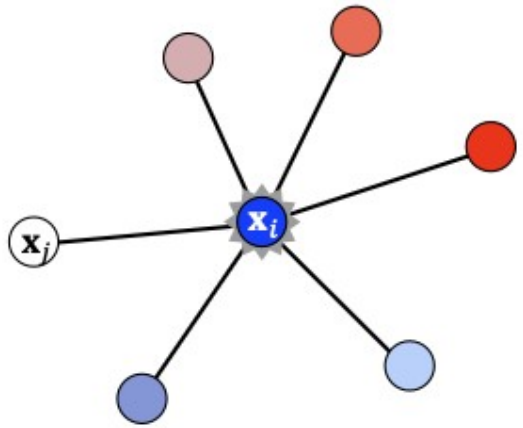
VII. *Scala graduum Caloris.*

Calorum Descriptiones & signa.

I. Newton

3.4 Summary

Heat Diffusion Equation on Graphs



$$\dot{\mathbf{x}}_i(t) = \mathbf{x}_i(t) - \frac{1}{d_i} \sum_{j \in \mathcal{N}_i} a_{ij} \mathbf{x}_j(t)$$

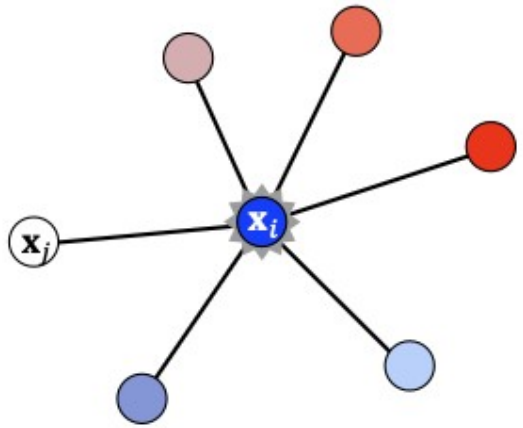
rate of temperature change

self temperature

temperature of the environment

3.4 Summary

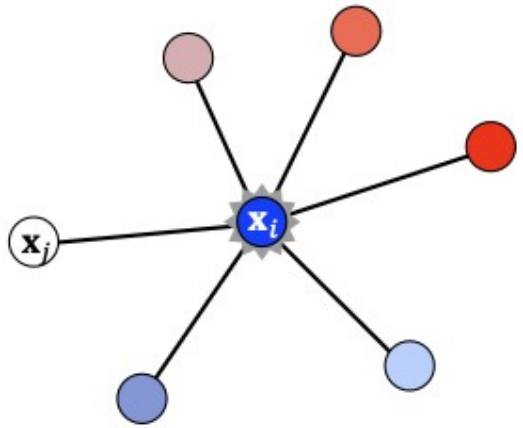
Heat Diffusion Equation on Graphs



$$\dot{\mathbf{x}}_i(t) = \underbrace{\frac{1}{d_i}\sum_{j\in\mathcal{N}_i} a_{ij}}_{\text{divergence}\ \text{div}} \underbrace{\left(\mathbf{x}_i(t) - \mathbf{x}_j(t)\right)}_{\text{gradient}\ -(\nabla \mathbf{X})_{ij}}$$

3.4 Summary

Heat Diffusion Equation on Graphs



$$\dot{\mathbf{x}}_i(t) = \frac{1}{d_i} \sum_{j \in \mathcal{N}_i} a_{ij} \left(\mathbf{x}_i(t) - \mathbf{x}_j(t) \right)$$

divergence
div

gradient
$-(\nabla \mathbf{X})_{ij}$

$$\dot{\mathbf{X}}(t) = \Delta \mathbf{X}(t)$$

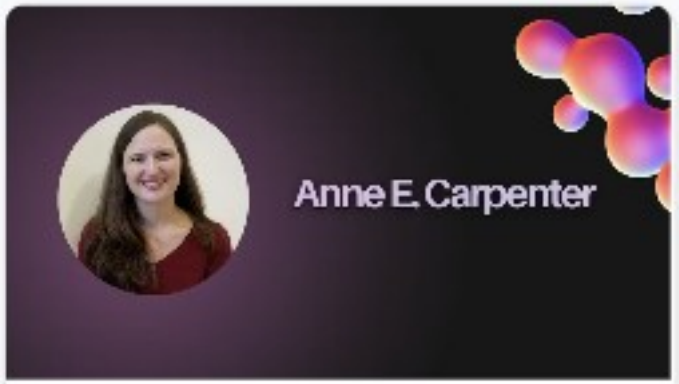
3.4 Summary

Michael concludes:

Conclusions

- Geometric deep learning is a powerful tool for molecular modelling
- Rootes in fundamental mathematical principles of invariance & symmetry
- Geometric generative models models
- Experimental validation!

4.1 Anne



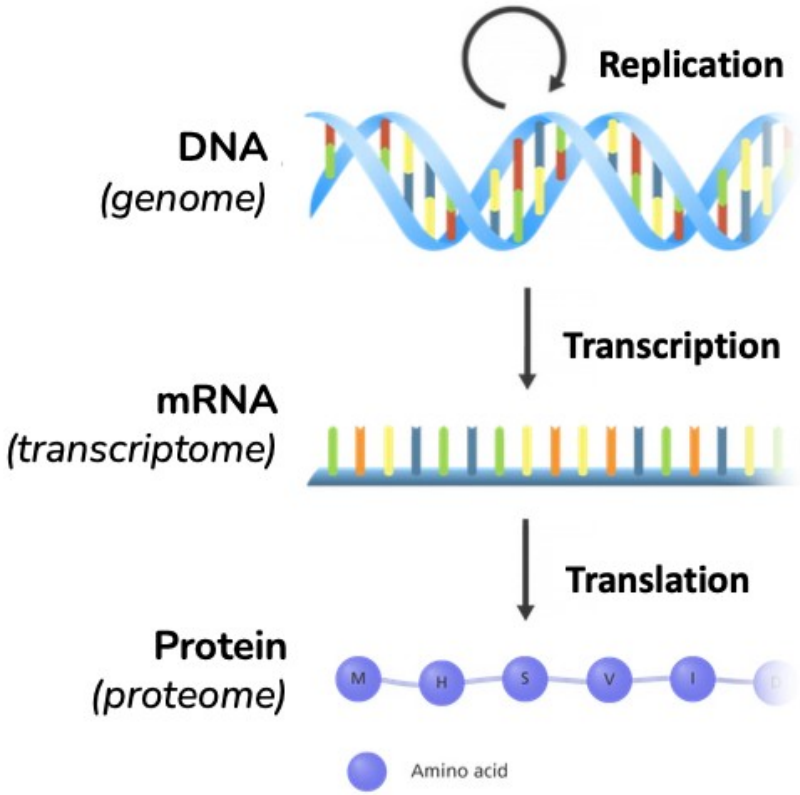
4.1 Phenomics in Drug Discovery

Professor: Anne Carpenter
Lecture Recording: Lecture
Slides:

0 06/17/2024

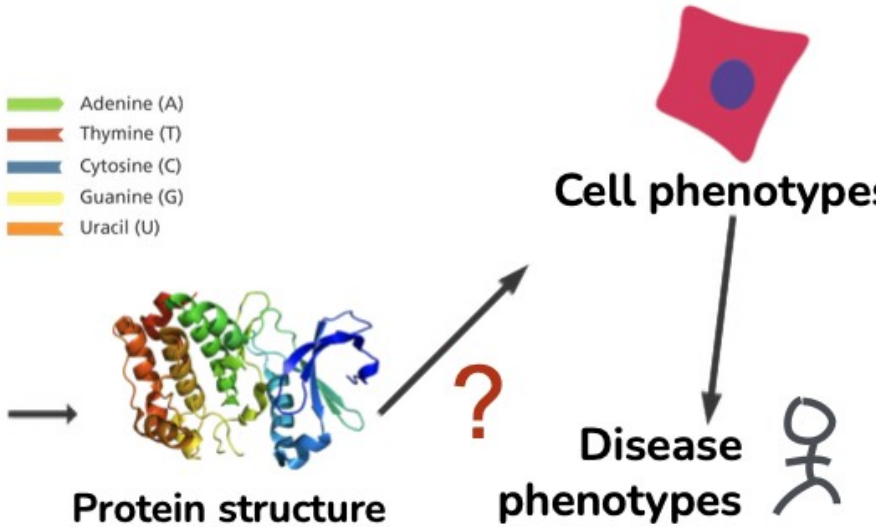
But how do we know what protein to target??

It's… messy



Connecting proteins to diseases:

- ☐ There are not simple rules we can follow.
- ☐ There's not enough direct data to train on.
- ☐ We cannot experiment on humans



4.1 Summary

Cell painting assays can be used for drug discovery:

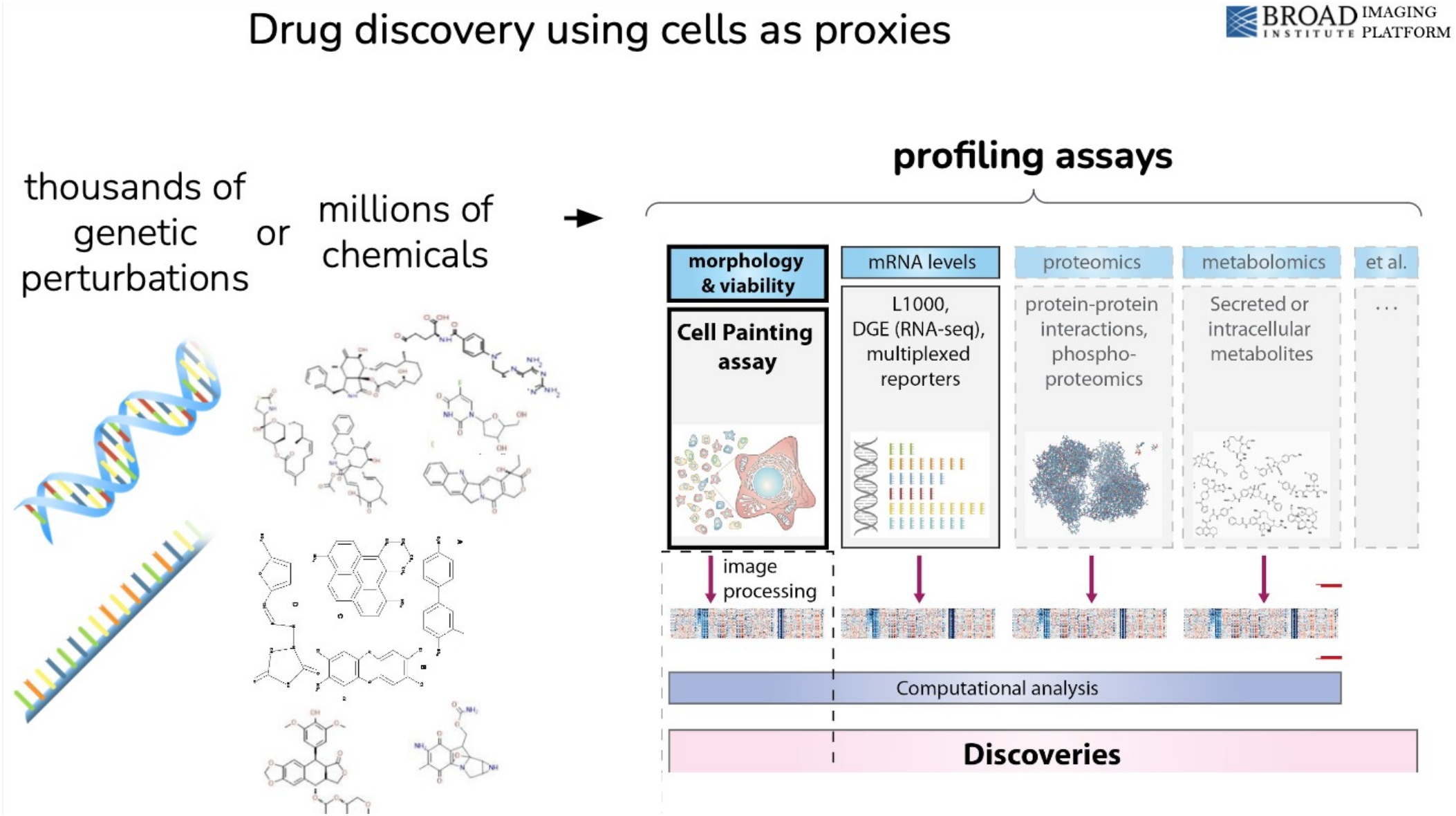
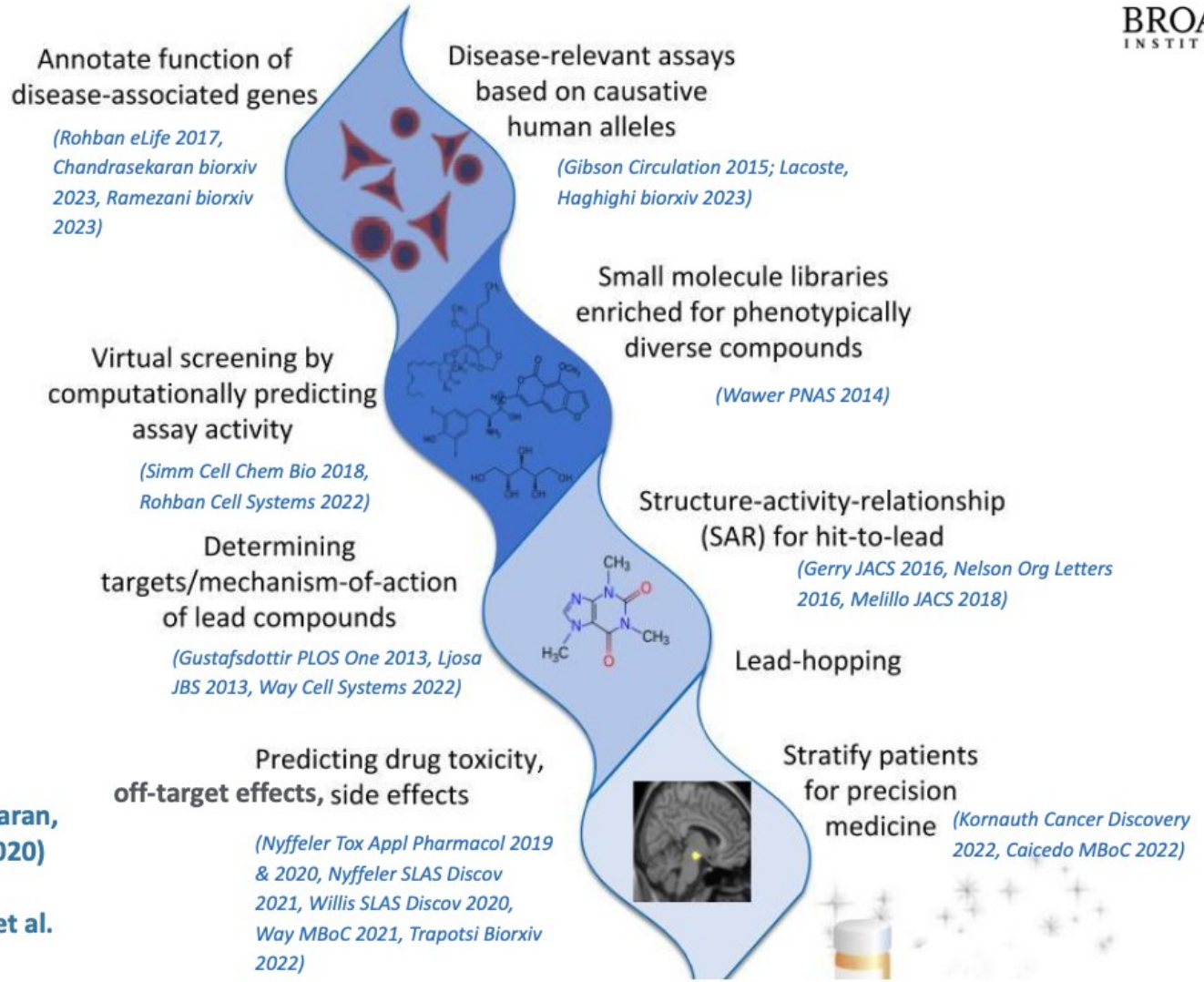


Image-based profiling can be applied to many different steps:

Image-based profiling: accelerating many steps of drug discovery



Applications reviewed in: Chandrasekaran, et al. Nat Reviews Drug Discovery (2020)

Best practices reviewed in: Caicedo, et al. Nature Methods (2017)

Annual conferences: CytoData, SBI2

Download PDF of slides here: http://broad.io/CarpenterSlides

4.1 Summary

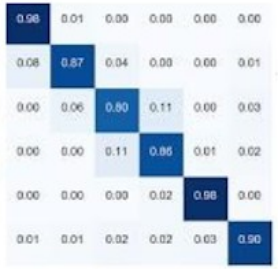
There are substantial success of image based profiling such as:

More successes of image-based profiling

BROAD INSTITUTE IMAGING PLATFORM

Applications reviewed in: Chandrasekaran, Ceulemans, Boyd, Carpenter Nat Reviews Drug Discovery (2020)
Best practices reviewed in: Caicedo, et al. Nature Methods (2017)
Annual conferences: CytoData, SBI2, Images 2 Knowledge (I2K)

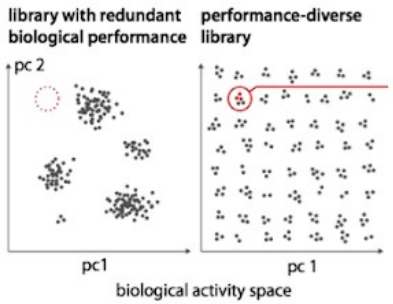
Label-free imaging flow cytometry diagnostics



classify cell cycle, red blood cell lesions, leukemic progression and response

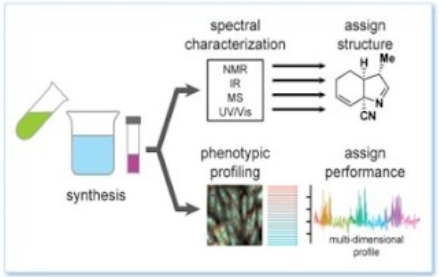
Blasi, et al. Nature Comm. 2016
Hennig, et al. Methods 2016
Eulenberg, et al. Nature Comm. 2017
Doan, et al. Trends Biotech 2018

Create performance-diverse screening libraries



improved hit rate
50% increase vs randomly chosen compounds

Test bioactivity for newly synthesized compounds



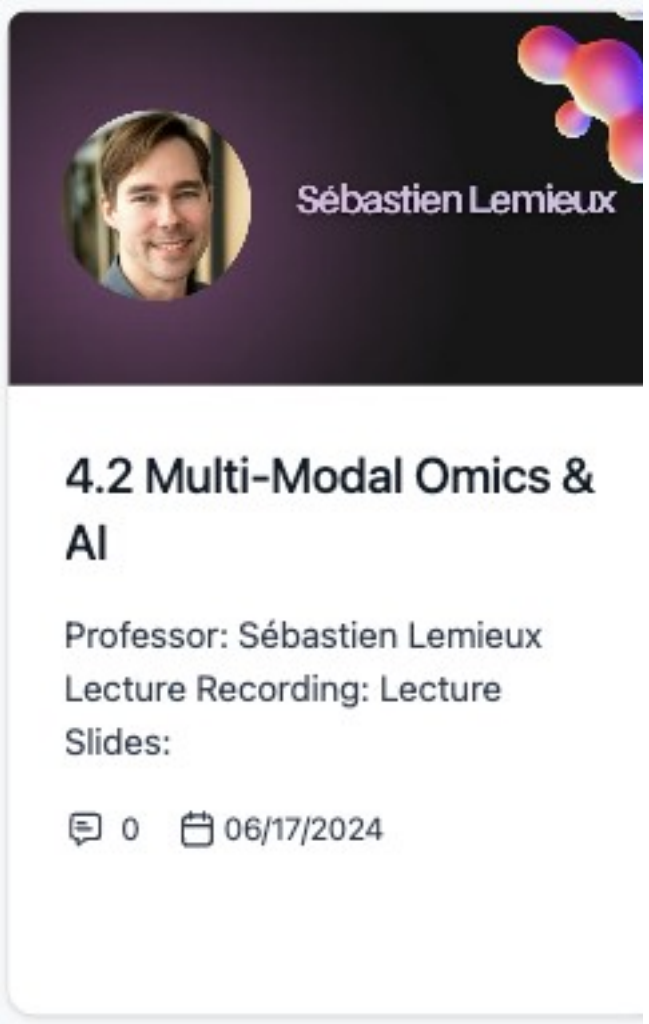
improved turnaround time
from years to days

Gerry, et al. JACS 2016
Nelson, et al. Org Lett 2016

Predict compound toxicity

OASIS Consortium launched!

4.2 Sebastien

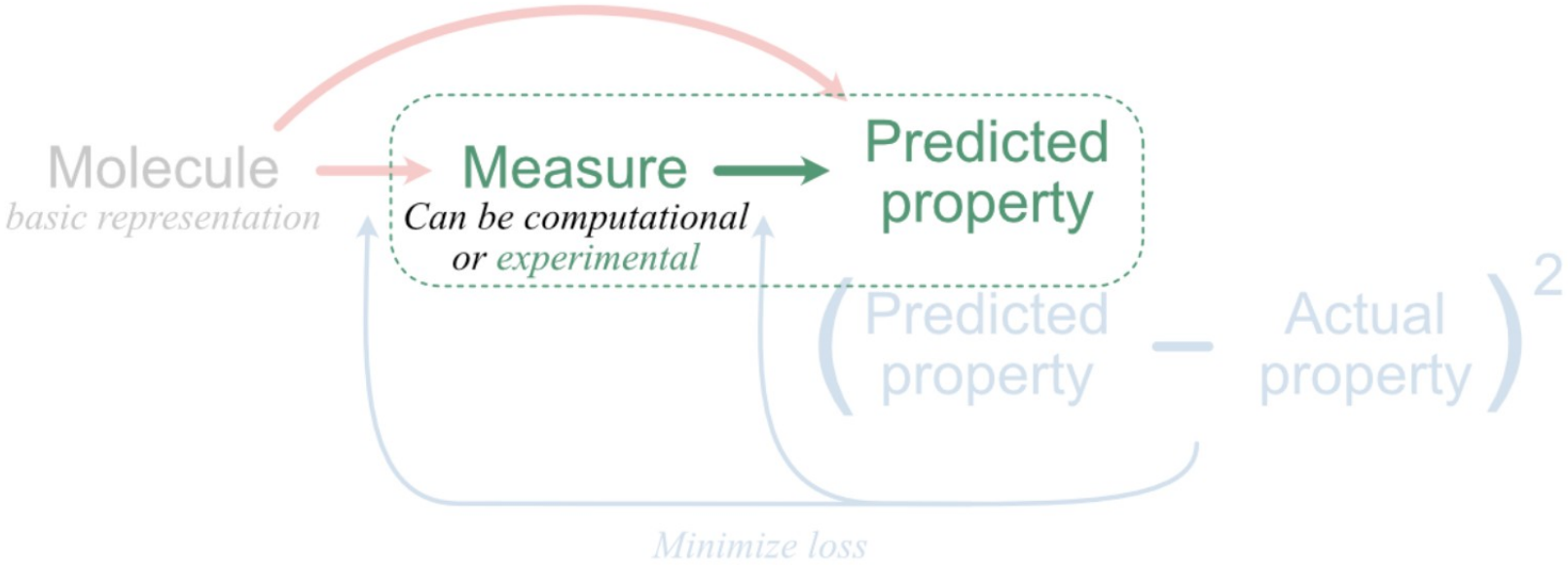


What's omics?

- Emerges in the mid-90s:
  - with microarrays, shotgun sequencing, proteome-wide mass spectrometry;
  - with advances closely tied with bioinformatics / computational biology.
- Large-scale datasets generated by high-throughput technologies.
  - Typically, 10^3 to 10^6 measures per biological sample.
- This data encompasses a broad range of biological fields:
  - Genomics (DNA)
  - Transcriptomics (mRNA)
  - Proteomics (proteins)
  - Metagenomics (microbial communities), metabolomics (metabolites), epigenomics (non sequence-based DNA modifications), etc.
- Such data aims to provide a holistic view of the molecular components to facilitate a deeper *understanding* of complex biological processes.

What are the tasks of multimodal learning in omics >

Tasks



We'll focus on experimental "measures"

The molecule must be synthesized and assayed

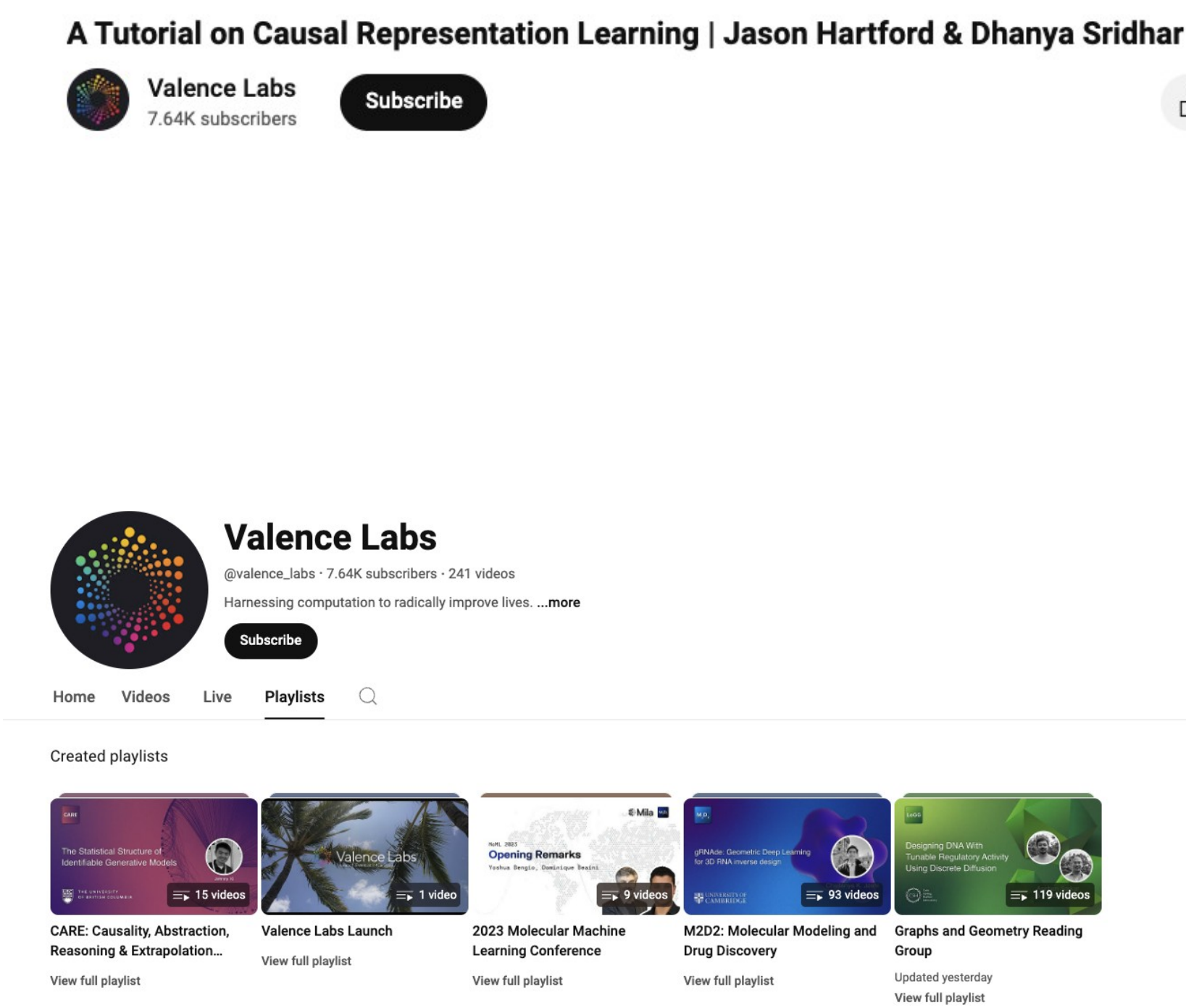
Promises a much richer characterization of the molecule

Take-home message

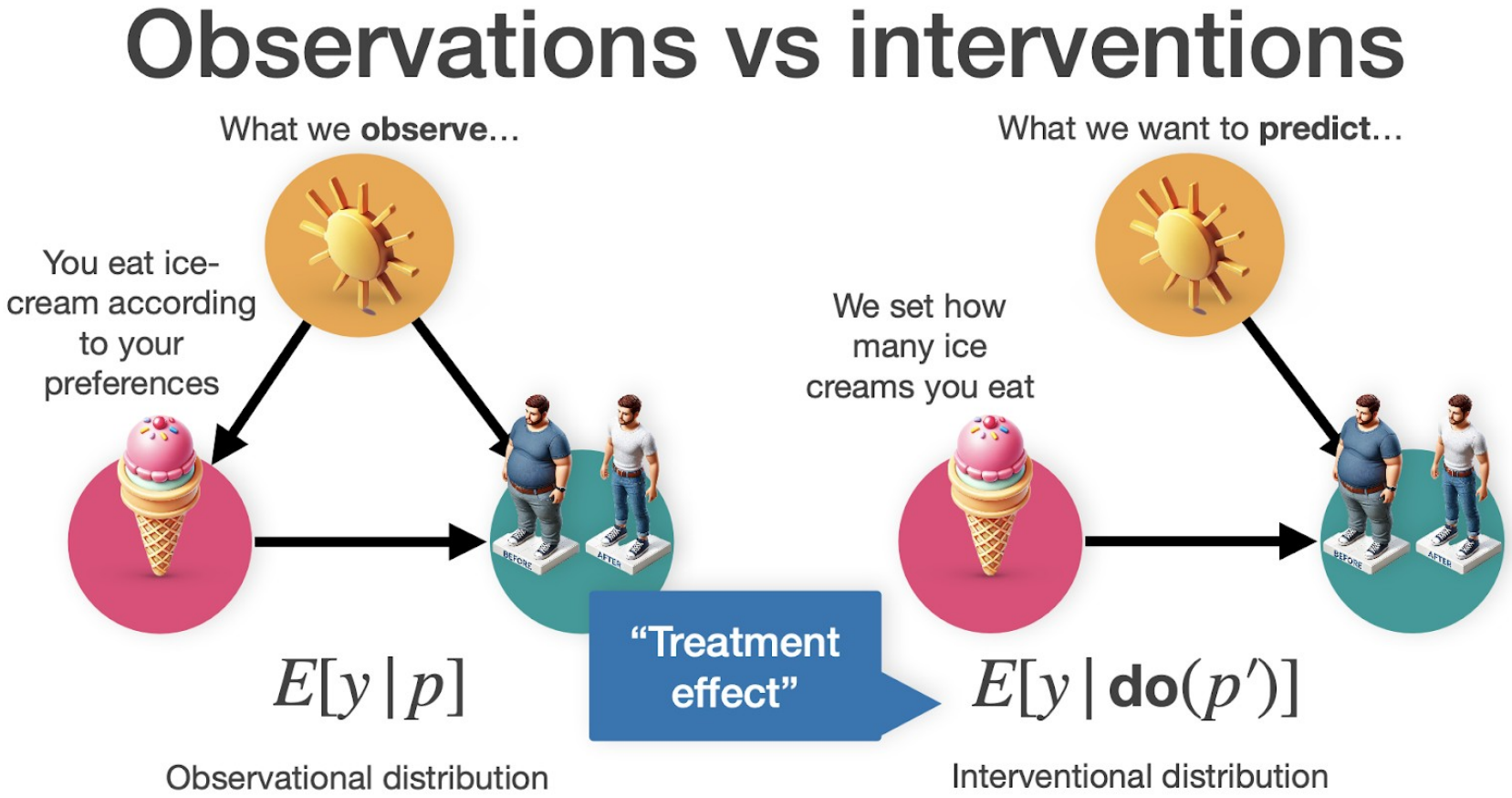
- Biological prediction tasks need to account for an infinite number of input dimensions, showing an important level of correlation.
  - Selection is always made, *ad hoc*, at the selection of experimental data generation, for availability, cost or feasibility reasons.
- Use of multimodal approaches in omics data usage is still poorly studied, despite being desired for at least 20 years.[1]
  - Still today, the field is dominated by late fusion and discretization, reusing modality-specific analyses, often designed for other uses.
- Current, direct application to drug discovery are still few.
  - Target deconvolution.[2,3]
  - Drug activity prediction from L1000 expression profiles.[4]
- Use of factorized embeddings offers clear opportunity in omics data integration.[5,6]

4.3 Jason


Jason Hartford
4.3 Causal Discovery & Representation Learning
Professor: Jason Hartford Lecture Recording: Lecture Slides:
0
06/17/2024


A Tutorial on Causal Representation Learning | Jason Hartford & Dhanya Sridhar
Valence Labs
7.64K subscribers
Subscribe
Valence Labs
@valence_labs · 7.64K subscribers · 241 videos
Harnessing computation to radically improve lives. ...more
Subscribe
Home
Videos
Live
Playlists
Created playlists
The Statistical Structure of Identifiable Generative Models
15 videos
Valence Labs
1 video
Opening Remarks
9 videos
gRNAde: Geometric Deep Learning for 3D RNA inverse design
93 videos
Designing DNA With Tunable Regulatory Activity Using Discrete Diffusion
119 videos
CARE: Causality, Abstraction, Reasoning & Extrapolation...
View full playlist
Valence Labs Launch
View full playlist
2023 Molecular Machine Learning Conference
View full playlist
M2D2: Molecular Modeling and Drug Discovery
View full playlist
Graphs and Geometry Reading Group
Updated yesterday
View full playlist

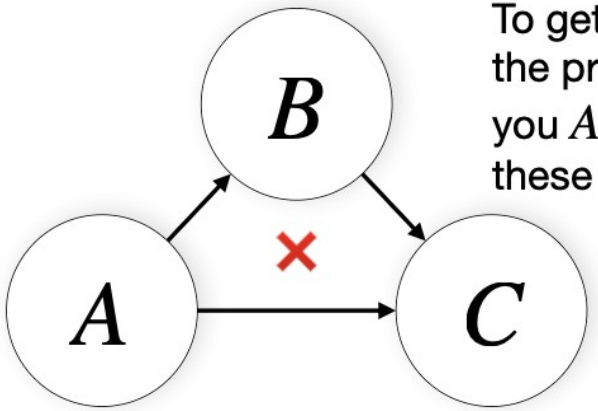
4.3 Summary


Observations vs interventions
What we observe…
You eat ice-cream according to your preferences
BEFORE
AFTER
$E[y \mid p]$
Observational distribution
"Treatment effect"
What we want to predict…
We set how many ice creams you eat
BEFORE
AFTER
$E[y \mid \mathbf{do}(p')]$
Interventional distribution

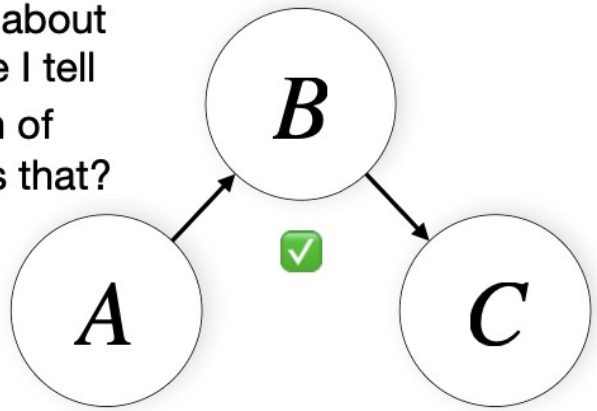
Causal discovery

$d-$separation allows to infer conditional independencies from the graphical structure implied by our assumptions.

Causal discovery goes in the opposite direction: given a set of conditional independencies, what does that imply about the underlying graph?

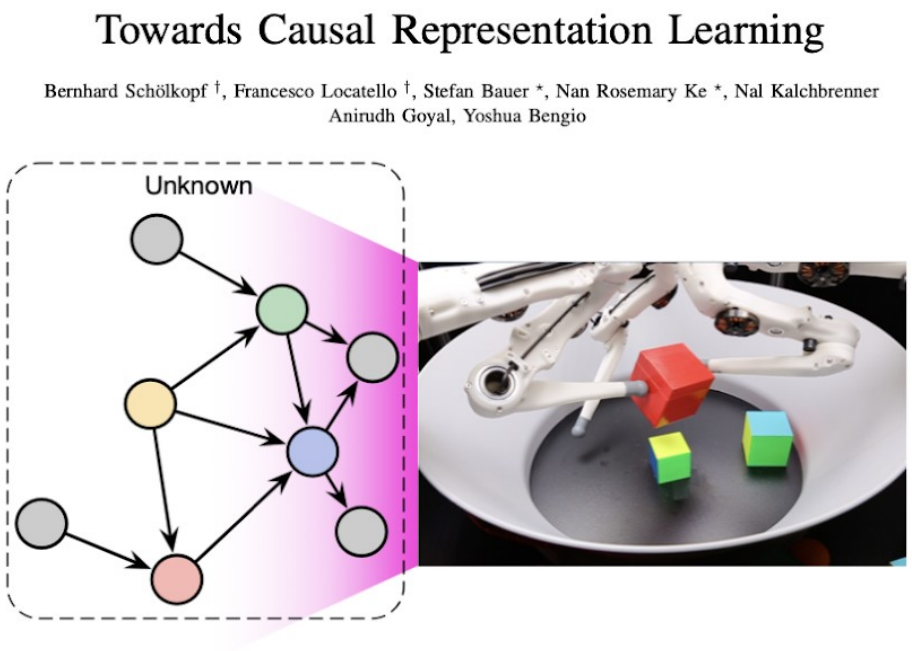


To get some intuition about the problem, suppose I tell you $A \perp C \mid B$. Which of these graphs satisfies that?

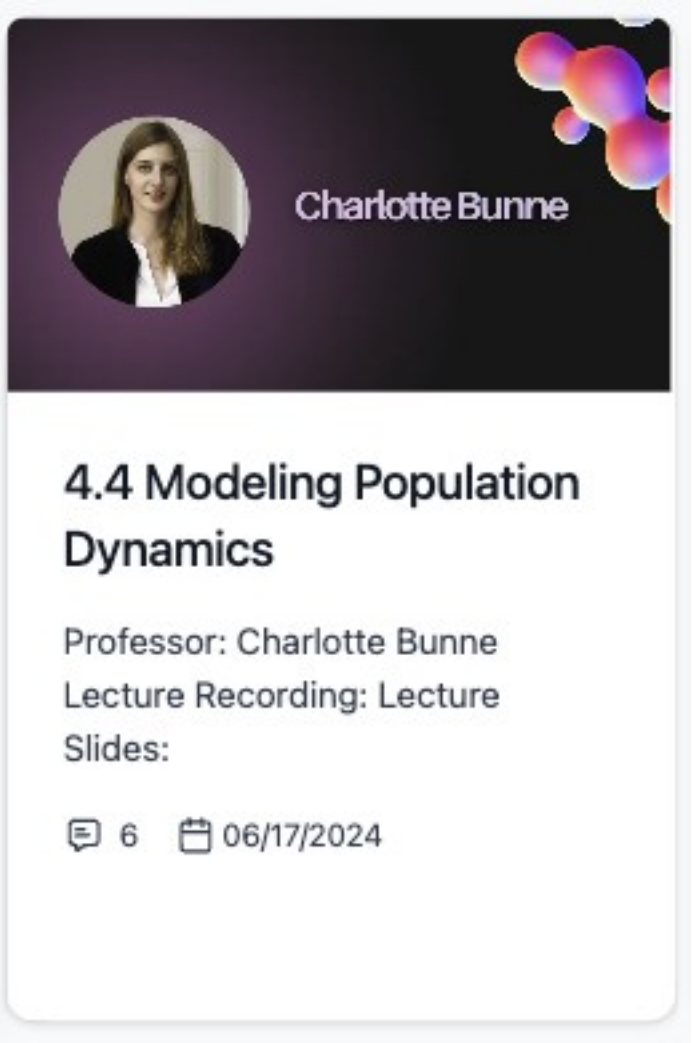


Causal representation learning

- Everything we've seen so far assumed semantically meaningful variables are observed. What do we do if we just have unstructured data like the pixels in an image?
- **Today**: why is this problem hard? What do existing techniques do?



4.4 Charlotte



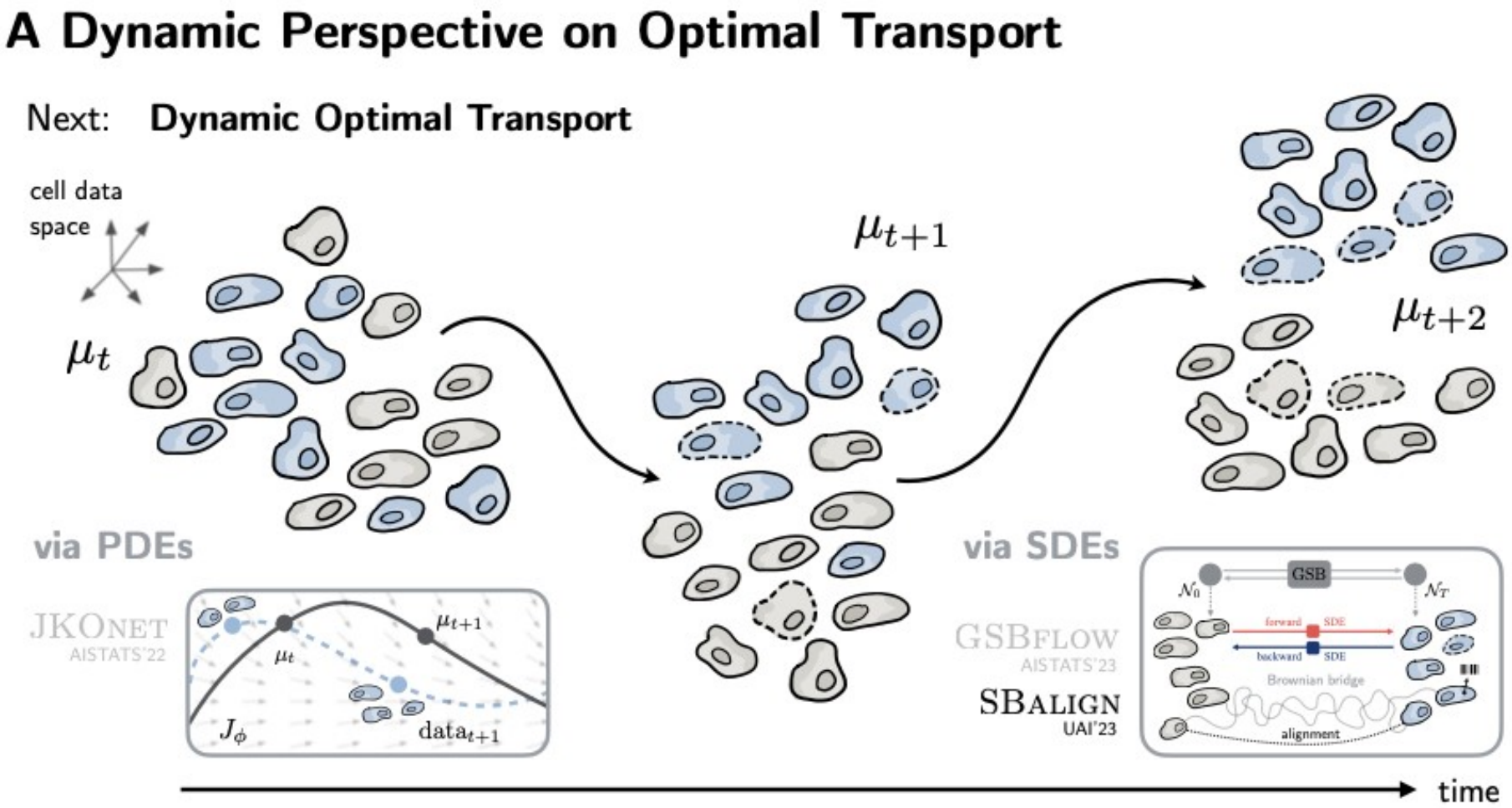
Optimal Transport

Static Optimal Transport as Inductive Bias for Reconstructing Molecular Responses in Biomedicine

Dynamic Optimal Transport and Connections to Diffusion Models and Flow Matching

Diffusion Models and Flow Matching for Reconstructing Dynamic Processes in Molecular Medicine

4.4 Summary


A Dynamic Perspective on Optimal Transport
Next: Dynamic Optimal Transport
cell data space
μ_t
μ_{t+1}
μ_{t+2}
via PDEs
JKONET
AISTATS'22
μ_t
μ_{t+1}
J_ϕ
data_{t+1}
via SDEs
GSBFLOW
AISTATS'23
SBALIGN
UAI'23
GSB
$\mathcal{N}_0$
$\mathcal{N}_T$
forward SDE
backward SDE
Brownian bridge
alignment
time

Solving Schrödinger Bridges

Schrödinger bridge optimality given by (Léonard 2013)

$$
* \begin{cases} \dfrac{\partial \Phi}{\partial t} = -\nabla \Phi^\top f - \dfrac{1}{2}\sigma^2 g^2 \Delta \Phi \\ \dfrac{\partial \widehat{\Phi}}{\partial t} = -\nabla \cdot (\widehat{\Phi} f) + \dfrac{1}{2}\sigma^2 g^2 \Delta \widehat{\Phi} \end{cases} \quad \text{s.t.} \quad \begin{aligned} \Phi(0,\cdot)\widehat{\Phi}(0,\cdot) &= \mu_0 \\ \Phi(1,\cdot)\widehat{\Phi}(1,\cdot) &= \mu_1 \end{aligned}
$$

Solution of $*$ can be expressed via two SDEs of the form

forward policy

$$
\mathrm{d}X_t = \left[f + g^2 \nabla \log \Phi\left(t, X_t\right)\right] \mathrm{d}t + g \, \mathrm{d}\mathbb{W}_t, \quad X_0 \sim \mu_0
$$

$$
\mathrm{d}X_t = \left[f - g^2 \nabla \log \widehat{\Phi}\left(t, X_t\right)\right] \mathrm{d}t + g \, \mathrm{d}\mathbb{W}_t, \quad X_1 \sim \mu_1
$$

backward policy

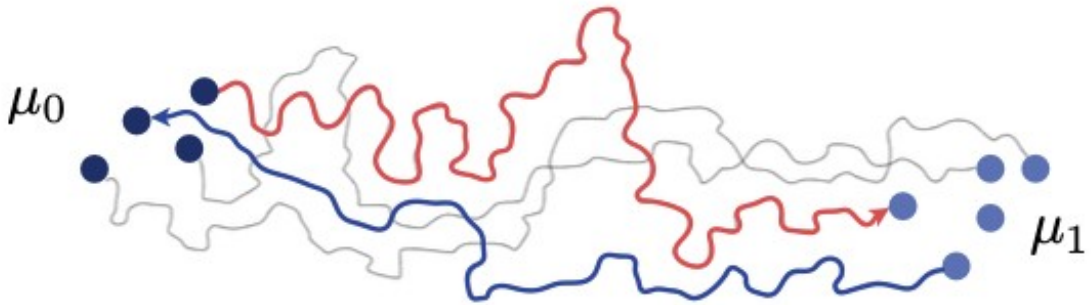
4.4 Summary

Solving Schrödinger Bridges

The solution of the SB is a system of a forward and backward SDE

forward policy

$$X_0 \sim \mu_0 \qquad \mathrm{d}X_t = \left[f + g^2 \nabla \log \Phi\left(t, X_t\right)\right] \mathrm{d}t + g \, \mathrm{d}\mathbb{W}_t$$



***reverse** SDE*

$$\mathrm{d}X_t = \left[f - g^2 \nabla \log \widehat{\Phi}\left(t, X_t\right)\right] \mathrm{d}t + g \, \mathrm{d}\mathbb{W}_t,$$

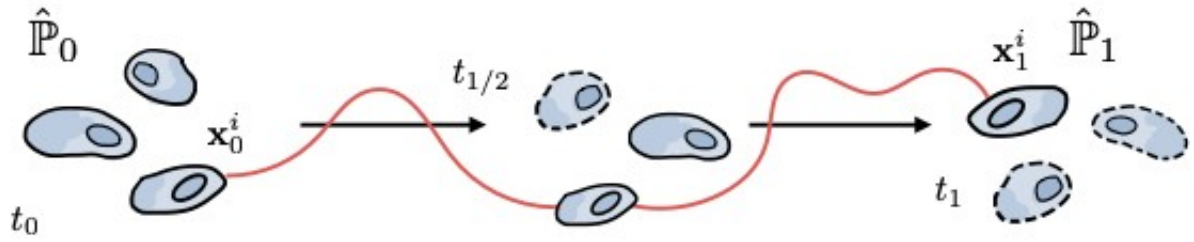
backward policy

(Anderson 1982)

4.4 Summary

Reconstructing Dynamics from Partial Alignments

UAI'23



stochastic process

couplings $\pi^\star := \underset{\mathbb{P}_0=\hat{\mathbb{P}}_0, \mathbb{P}_1=\hat{\mathbb{P}}_1}{\operatorname{argmin}} D_{\mathrm{KL}}\left(\mathbb{P}_{0,1} \| \mathbb{Q}_{0,1}\right)$ Schrödinger's Bridge

over $\hat{\mathbb{P}}_0$ and $\hat{\mathbb{P}}_1$ reference process

We can learn the drift describing cellular dynamics by minimizing the resulting loss function

$$\min_\theta \mathbb{E}\left[\int_0^1 \left\| \frac{\mathbf{x}_1 - X_t}{\beta_1 - \beta_t} - \left(b_t^\theta + \nabla \log h_t^\theta\left(X_t\right)\right)\right\|^2 \mathrm{d}t\right]$$

drift from Brownian bridge

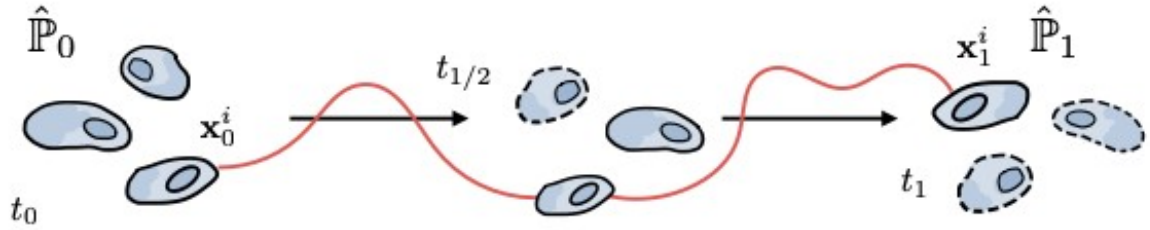
drift describing cellular dynamics

.. similar to score-matching objective successfully employed in previous works (Song et al., 2021)

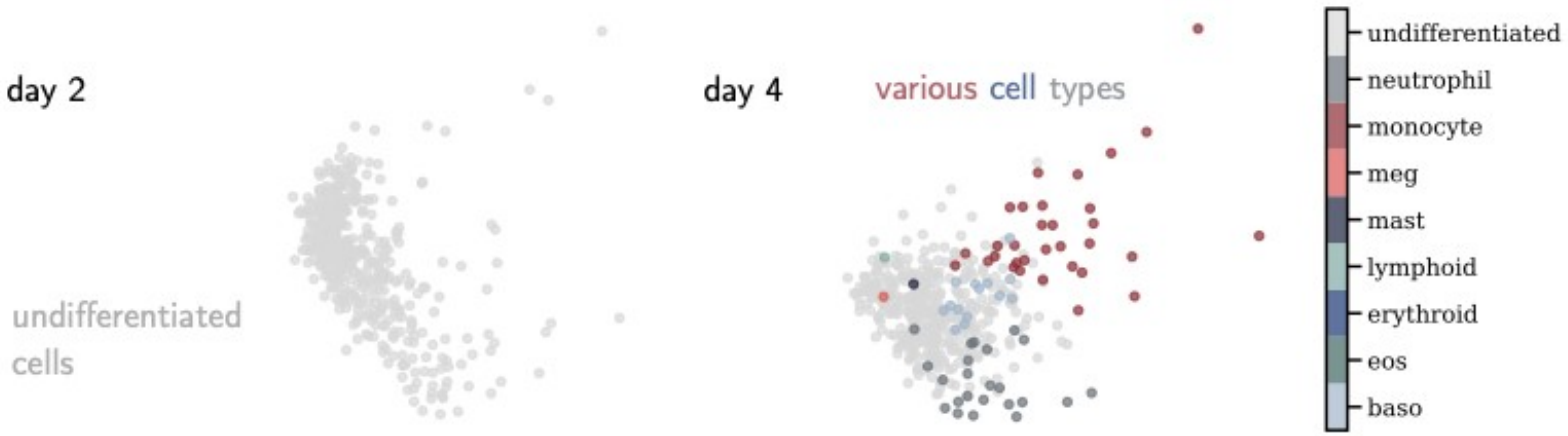
4.4 Summary

Recovering *Continuous* Dynamics of Cell Differentiation

UAI'23



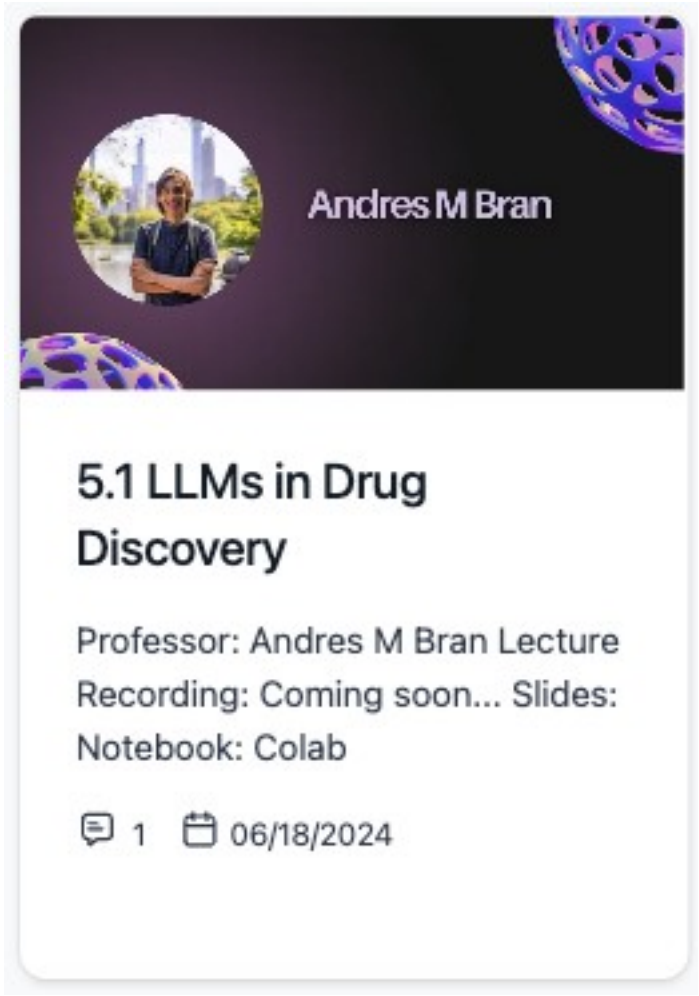
Application: Fate determination in hematopoiesis.

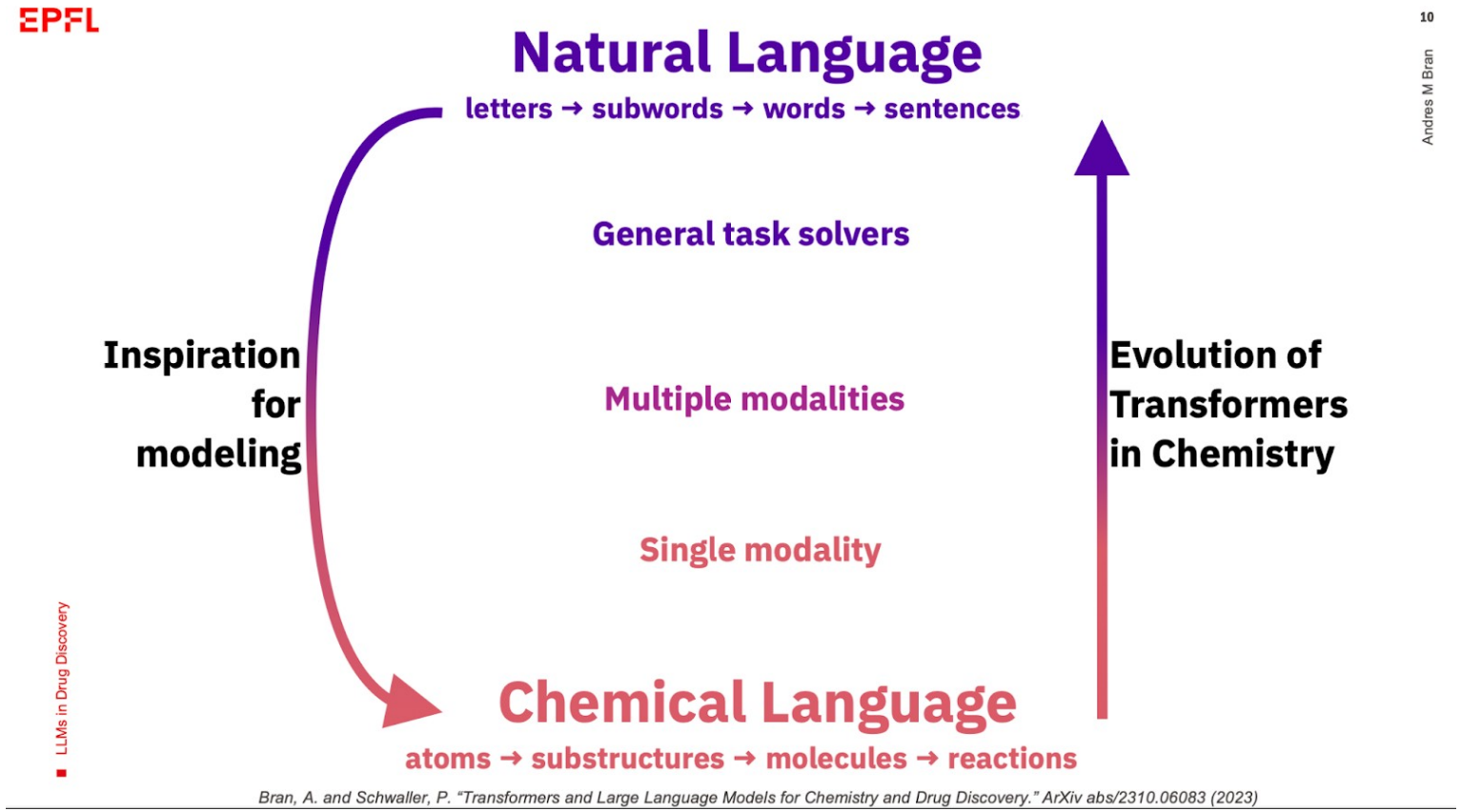


.. dataset by Weinreb et al., Science (2020)

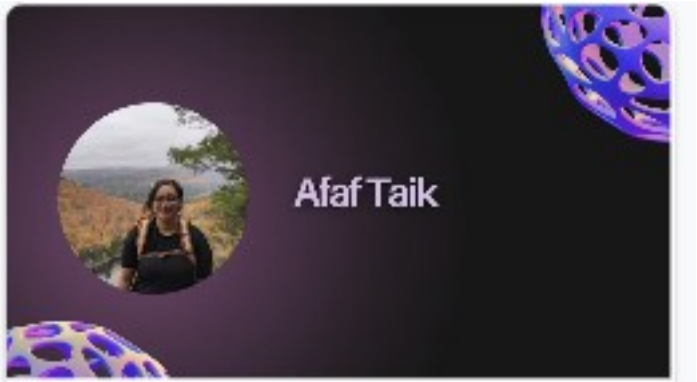
5.1 Andres

The only talk with a Colab Demo!





5.4 Afaf



5.4 Ethical & Bias Concerns

Professor: Afaf Taik Lecture Recording: Coming soon... Slides:

1 06/18/2024

The End