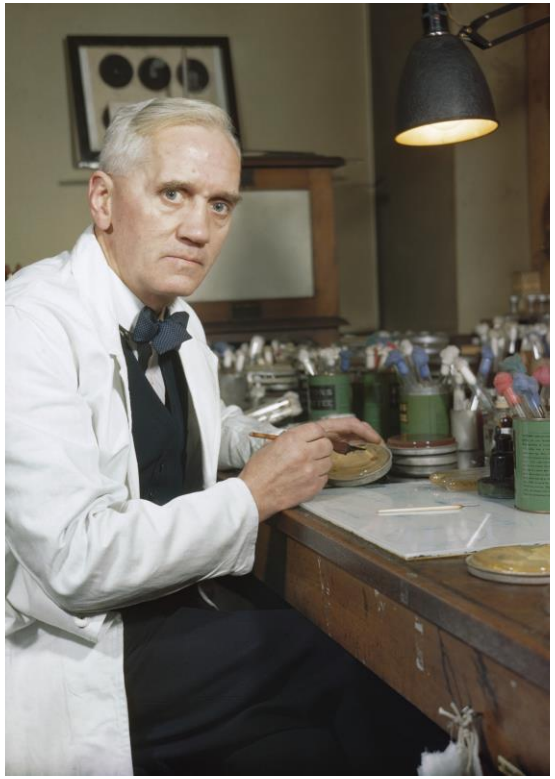
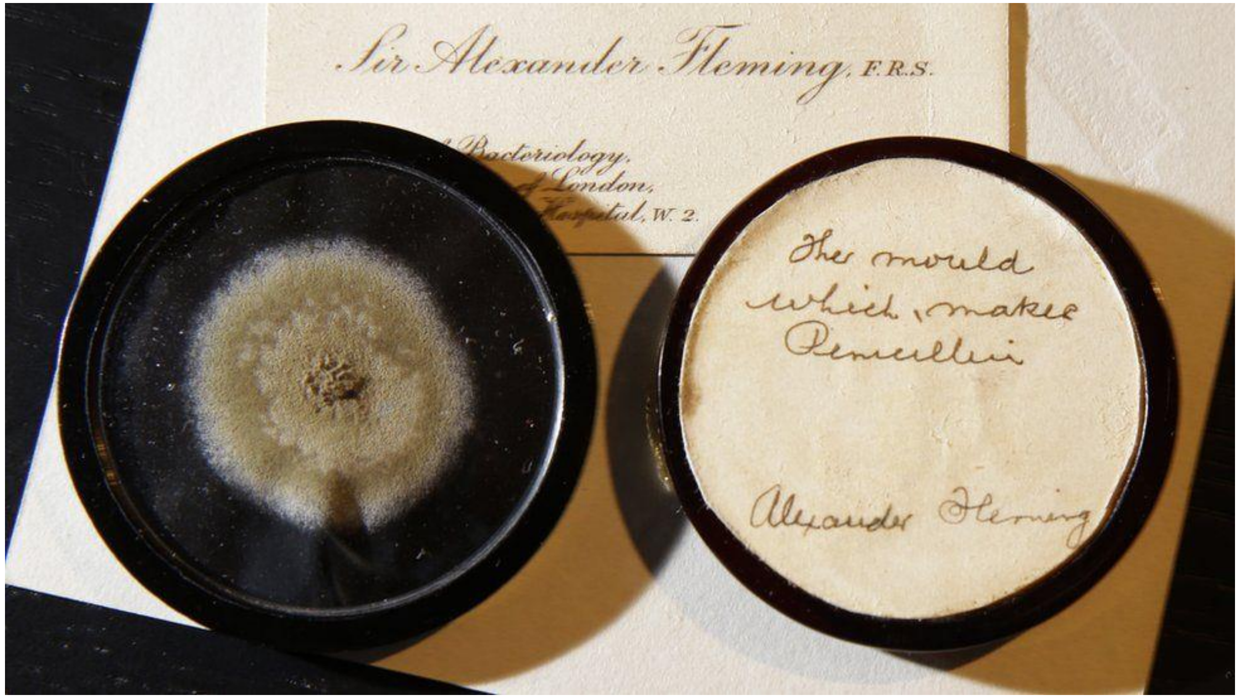
Cancer research with AI methods

Traditional drug development process

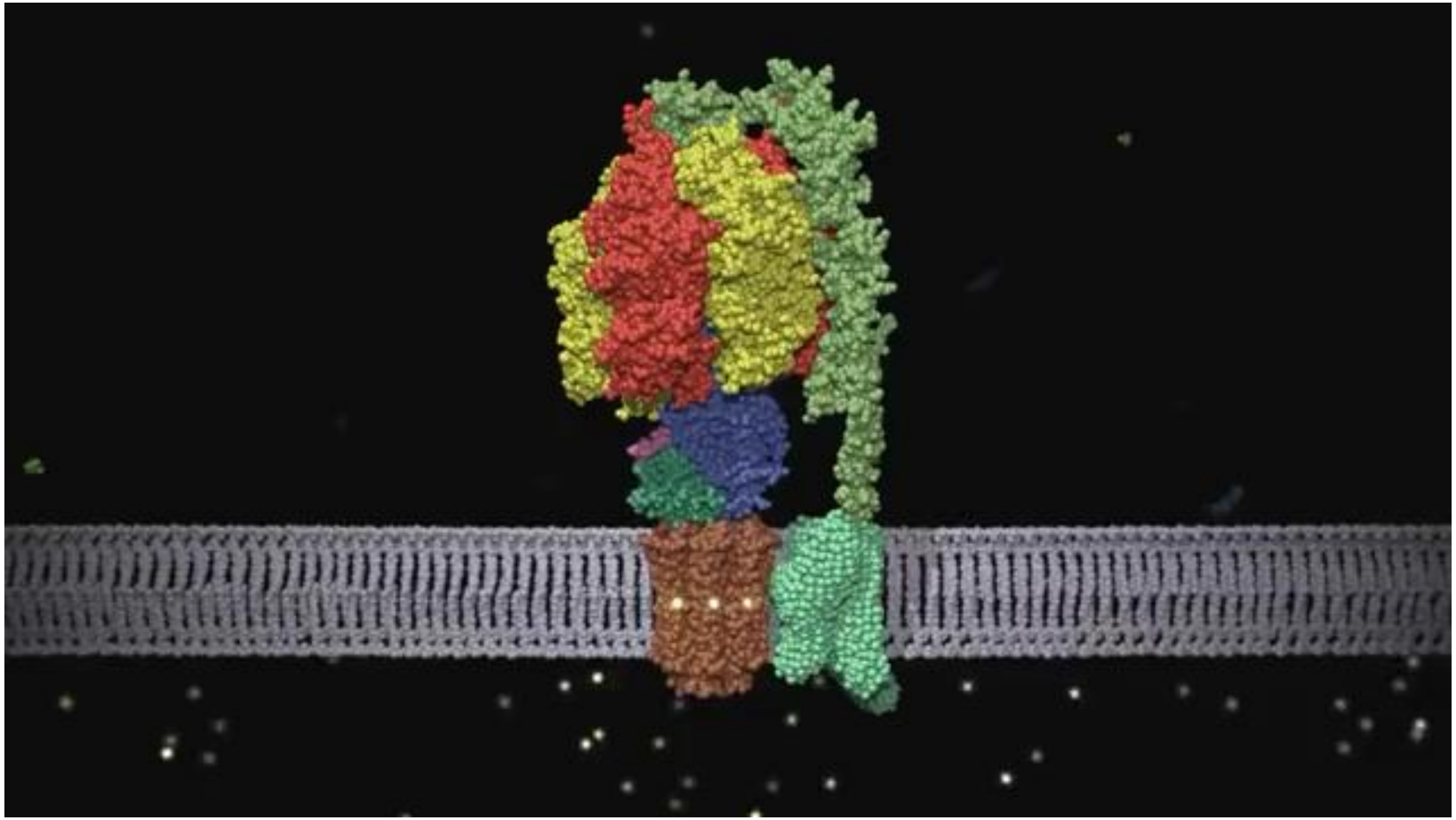


Alexander Fleming


Sir Alexander Fleming, F.R.S.
Bacteriology,
of London,
Hospital, W. 2.
The mould which makes Penicillin
Alexander Fleming

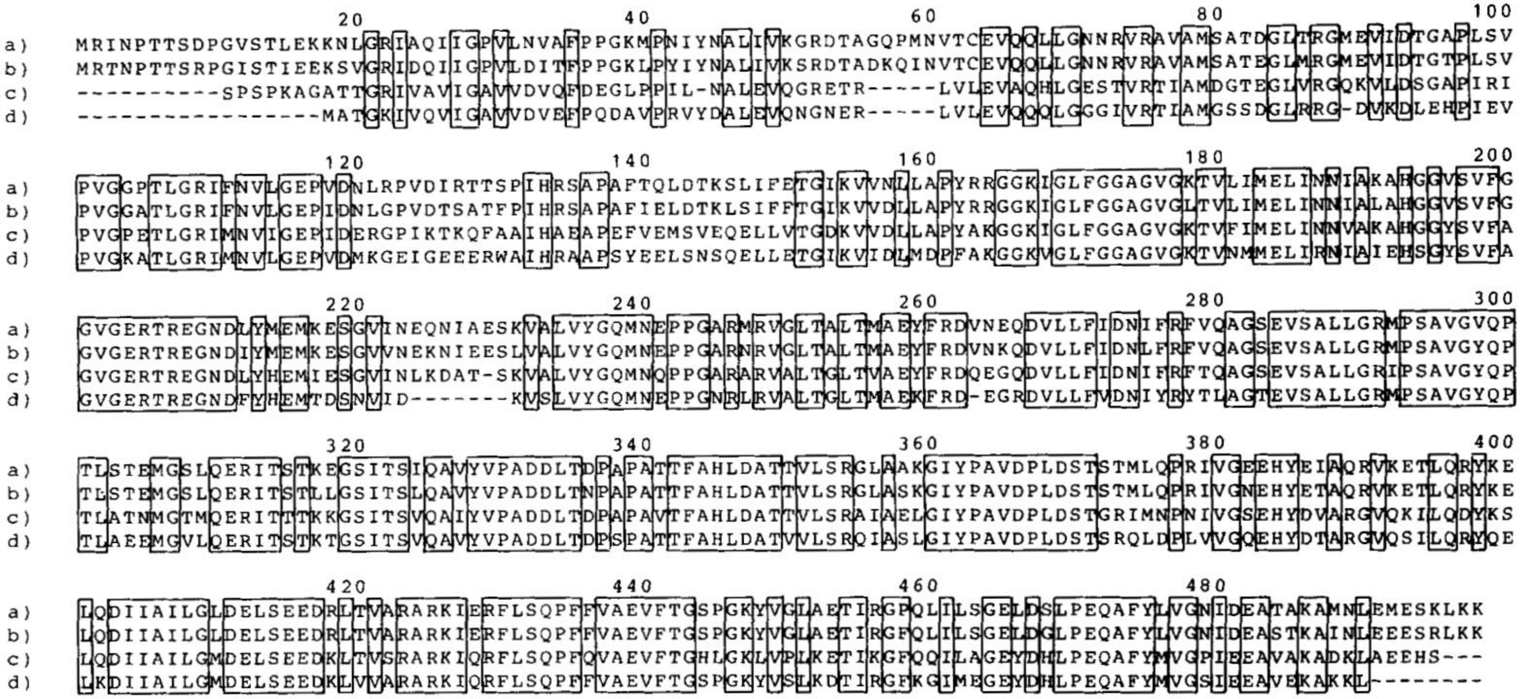
Bacteria vs mould

ATP Synthase in action



ATP Synthase – amino acids

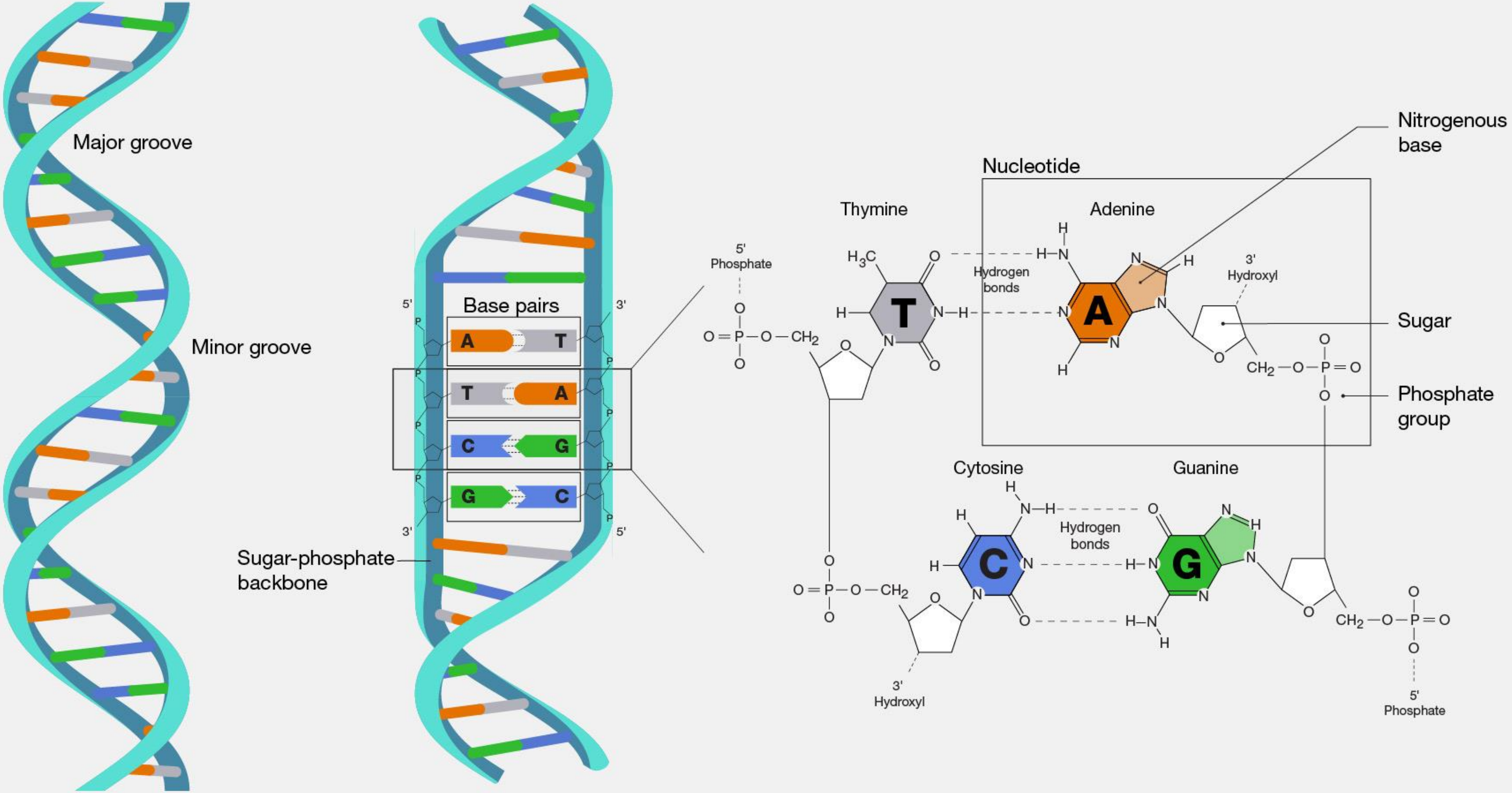
FIG. 1—*continued*



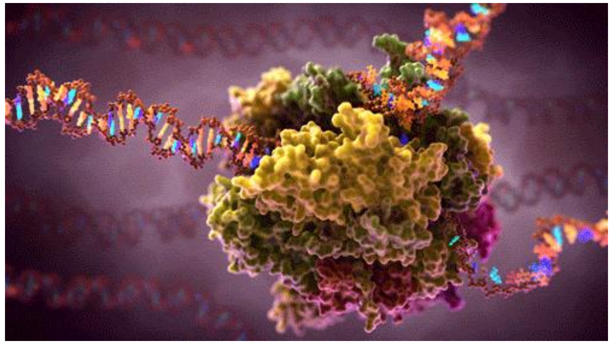
```
                        20                  40                  60                  80                 100
a)  MRINPTTSDPGVSTLEKKNLGRIAQIIGPVLNVAFPPGKMPNIYNALIVKGRDTAGQPMNVTCEVQQLLGNNRVRAVAMSATDGLTRGMEVIDTGAPLSV
b)  MRTNPTTSRPGISTIEEKSVGRIDQIIGPVLDITFPPGKLPYIYNALIVKSRDTADKQINVTCEVQQLLGNNRVRAVAMSATEGLMRGMEVIDTGTPLSV
c)  ----------SPSPKAGATTGRIVAVIGAVVDVQFDEGLPPIL-NALEVQGRETR-----LVLEVAQHLGESTVRTIAMDGTEGLVRGQKVLDSGAPIRI
d)  -----------------MATGKIVQVIGAVVDVEFPQDAVPRVYDALEVQNGNER-----LVLEVQQQLGGGIVRTIAMGSSDGLRRG-DVKDLEHPIEV

                       120                 140                 160                 180                 200
a)  PVGGPTLGRIFNVLGEPVDNLRPVDIRTTSPIHRSAPAFTQLDTKSLIFETGIKVVNLLAPYRRGGKIGLFGGAGVGKTVLIMELINNIAKAHGGVSVFG
b)  PVGGATLGRIFNVLGEPIDNLGPVDTSATFPIHRSAPAFIELDTKLSIFFTGIKVVDLLAPYRRGGKIGLFGGAGVGLTVLIMELINNIALAHGGVSVFG
c)  PVGPETLGRIMNVIGEPIDERGPIKTKQFAAIHAEAPEFVEMSVEQELLVTGDKVVDLLAPYAKGGKIGLFGGAGVGKTVFIMELINNVAKAHGGYSVFA
d)  PVGKATLGRIMNVLGEPVDMKGEIGEEERWAIHRAAPSYEELSNSQELLETGIKVIDLMDPFAKGGKVGLFGGAGVGKTVNMMELIRNIAIEHSGYSVFA

                       220                 240                 260                 280                 300
a)  GVGERTREGNDLYMEMKESGVINEQNIAESKVALVYGQMNEPPGARMRVGLTALTMAEYFRDVNEQDVLLFIDNIFRFVQAGSEVSALLGRMPSAVGVQP
b)  GVGERTREGNDIYMEMKESGVVNEKNIEESLVALVYGQMNEPPGARNRVGLTALTMAEYFRDVNKQDVLLFIDNLFRFVQAGSEVSALLGRMPSAVGYQP
c)  GVGERTREGNDLYHEMIESGVINLKDAT-SKVALVYGQMNQPPGARARVALTGLTVAEYFRDQEGQDVLLFIDNIFRFTQAGSEVSALLGRIPSAVGYQP
d)  GVGERTREGNDFYHEMTDSNVID-------KVSLVYGQMNEPPGNRLRVALTGLTMAEKFRD-EGRDVLLFVDNIYRYTLAGTEVSALLGRMPSAVGYQP

                       320                 340                 360                 380                 400
a)  TLSTEMGSLQERITSTKEGSITSIQAVYVPADDLTDPAPATTFAHLDATTVLSRGLAAKGIYPAVDPLDSTSTMLQPRIVGEEHYEIAQRVKETLQRYKE
b)  TLSTEMGSLQERITSTLLGSITSLQAVYVPADDLTNPAPATTFAHLDATTVLSRGLASKGIYPAVDPLDSTSTMLQPRIVGNEHYETAQRVKETLQRYKE
c)  TLATNMGTMQERITTTKKGSITSVQAIYVPADDLTDPAPAVTFAHLDATTVLSRAIAELGIYPAVDPLDSTGRIMNPNIVGSEHYDVARGVQKILQDYKS
d)  TLAEEMGVLQERITSTKTGSITSVQAVYVPADDLTDPSPATTFAHLDATVVLSRQIASLGIYPAVDPLDSTSRQLDPLVVGQEHYDTARGVQSILQRYQE

                       420                 440                 460                 480
a)  LQDIIAILGLDELSEEDRLTVARARKIERFLSQPFFVAEVFTGSPGKYVGLAETIRGPQLILSGELDSLPEQAFYLVGNIDEATAKAMNLEMESKLKK
b)  LQDIIAILGLDELSEEDRLTVARARKIERFLSQPFFVAEVFTGSPGKYVGLAETIRGFQLILSGELDGLPEQAFYLVGNIDEASTKAINLEEESRLKK
c)  LQDIIAILGMDELSEEDKLTVSRARKIQRFLSQPFQVAEVFTGHLGKLVPLKETIKGFQQILAGEYDHLPEQAFYMVGPIEEAVAKADKLAEEHS---
d)  LKDIIAILGMDELSEEDKLVVARARKIQRFLSQPFFVAEVFTGSPGKYVSLKDTIRGFKGIMEGEYDHLPEQAFYMVGSIEEAVEKAKKL--------
```

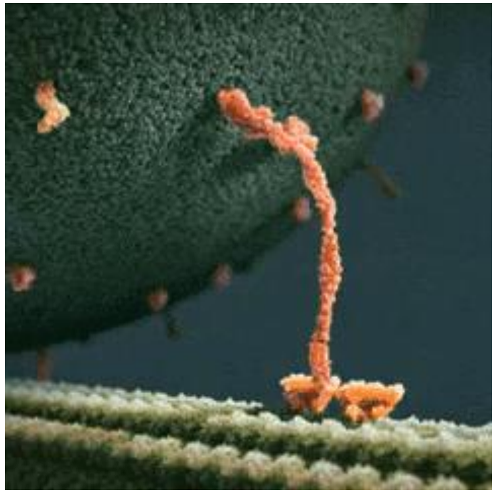
FIG. 2. **Alignment of the sequences of the *β*-subunits of ATP synthase from (*a*) spinach (19) and (*b*) maize (18) chloroplasts, (*c*) bovine mitochondria, and (*d*) *E. coli* (15).** Identities are *boxed*.

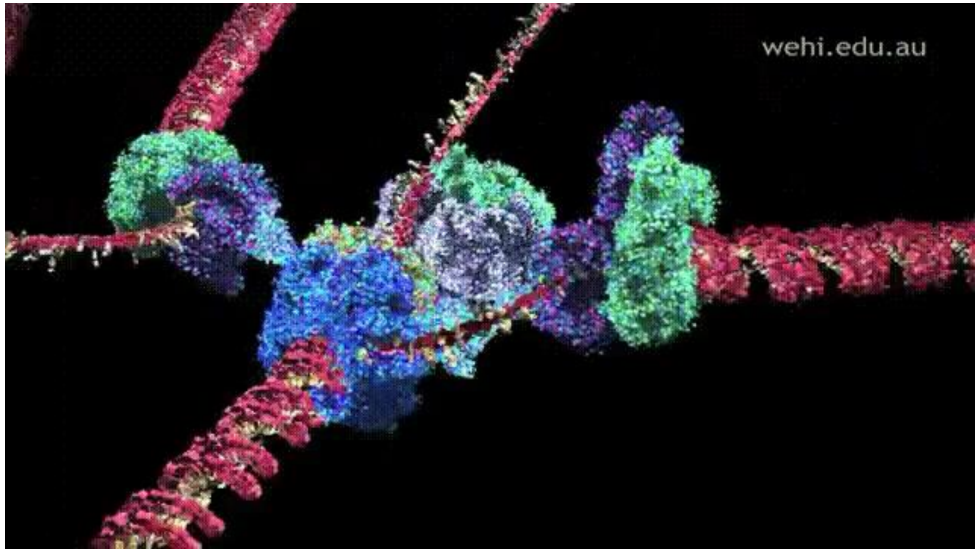
Deoxyribonucleic acid (DNA)



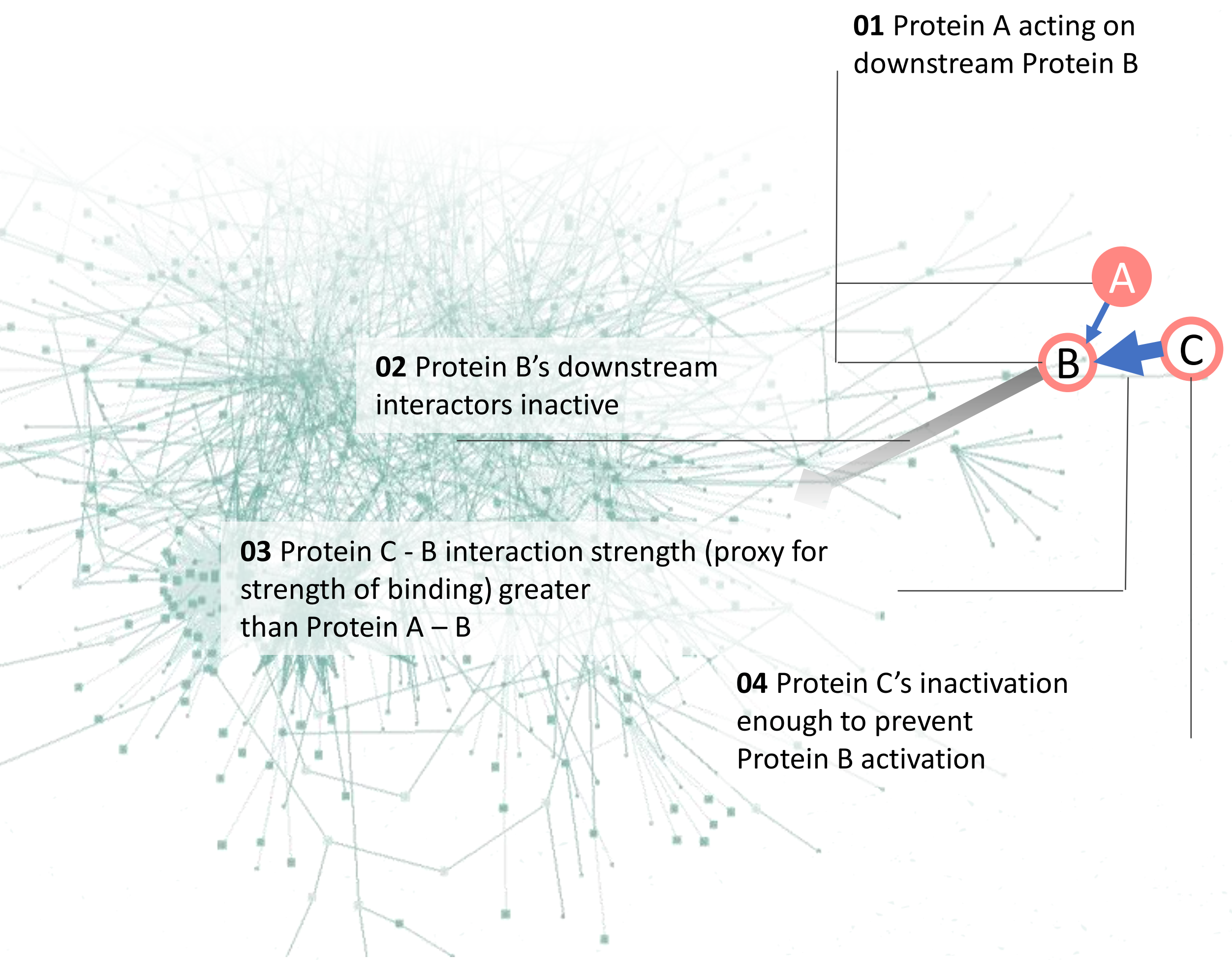
20,000 different proteins in human cells







Turbine simulating a human cell



01 Protein A acting on downstream Protein B

02 Protein B’s downstream interactors inactive

03 Protein C - B interaction strength (proxy for strength of binding) greater than Protein A – B

04 Protein C’s inactivation enough to prevent Protein B activation

- The network **is universal to all human cells** – all 1500+ biological models in the library use the **same “wiring diagram”, just with different OMICS profiles**
- 3800+ nodes
- 12000+ edges
- Modelling **drugs** or **mutations** achieved by modifying the parameters

Turbine AI

100m simulated experiments / week

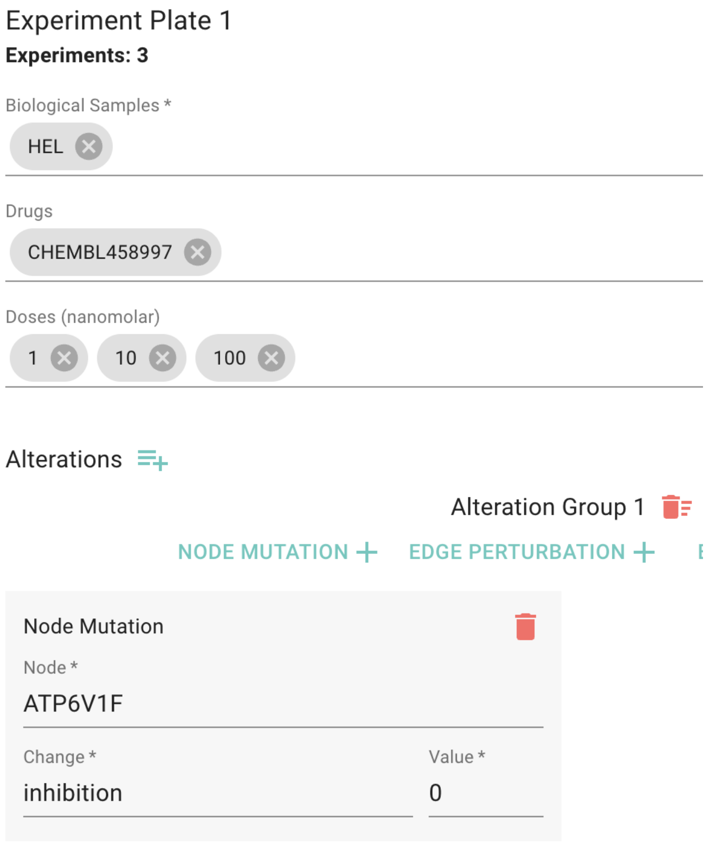
$0.0002 / simulated experiment

2GB protein activity data generated per experiment

177PB data generated
(130TB recorded in our data lake)

Generating raw simulation data at the same rate as the entire ATLAS detector at CERN

Configuring a simulation



Experiment Plate 1

Experiments: 3

Biological Samples *

HEL

Drugs

CHEMBL458997

Doses (nanomolar)

1 10 100

Alterations

Alteration Group 1

Node Mutation

Node *

ATP6V1F

Change *

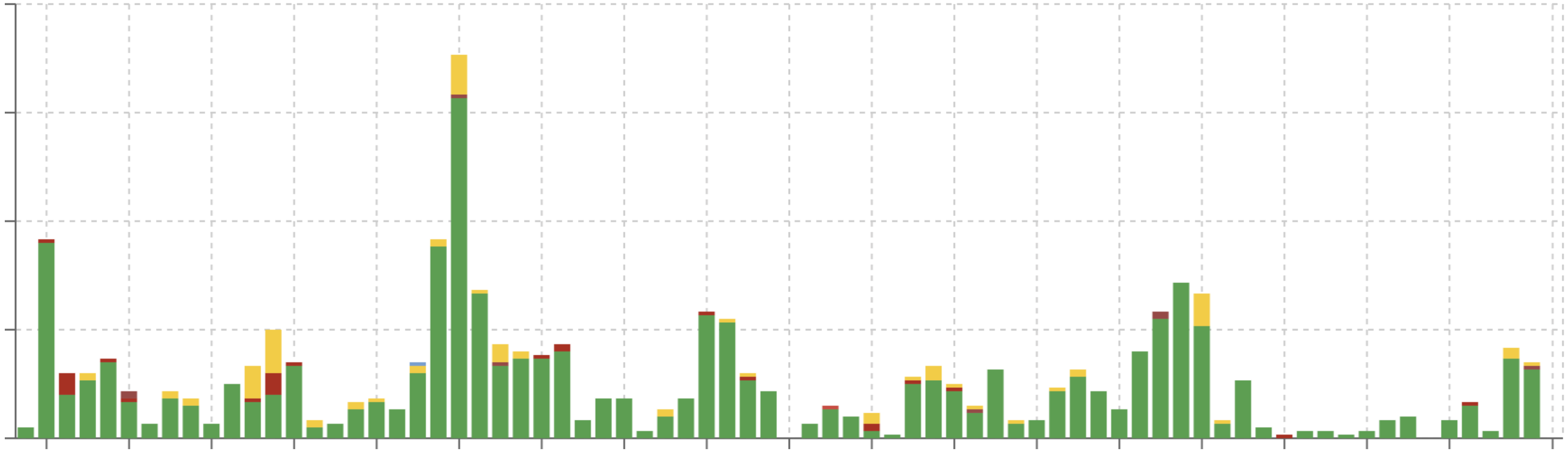
inhibition

Value *

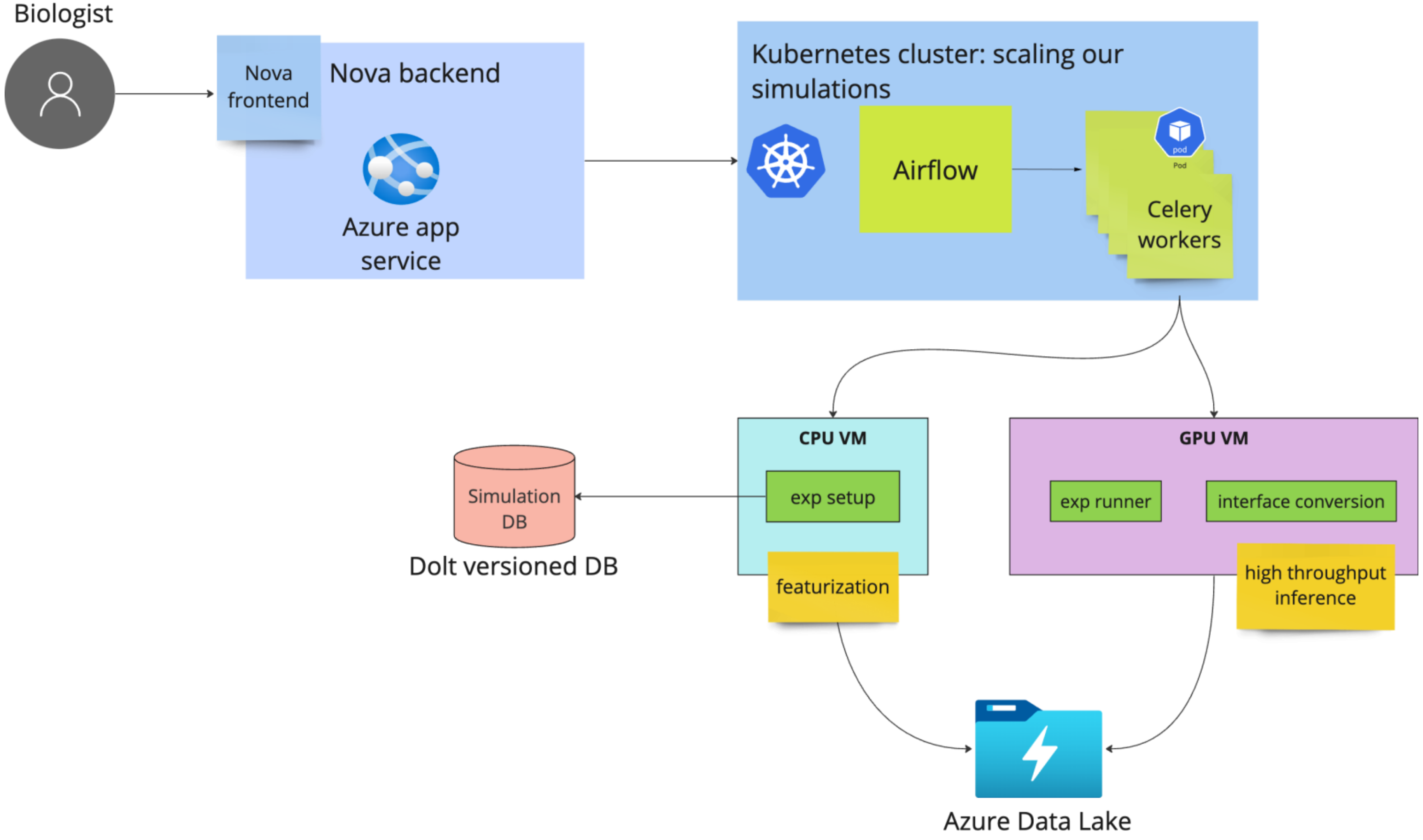
0

Challanges

Experiments	4 381 025
Runtime	44 hours 14 minutes
Shards	490 (1.5h on average)
Cost	951 USD
Stored data	214 GB



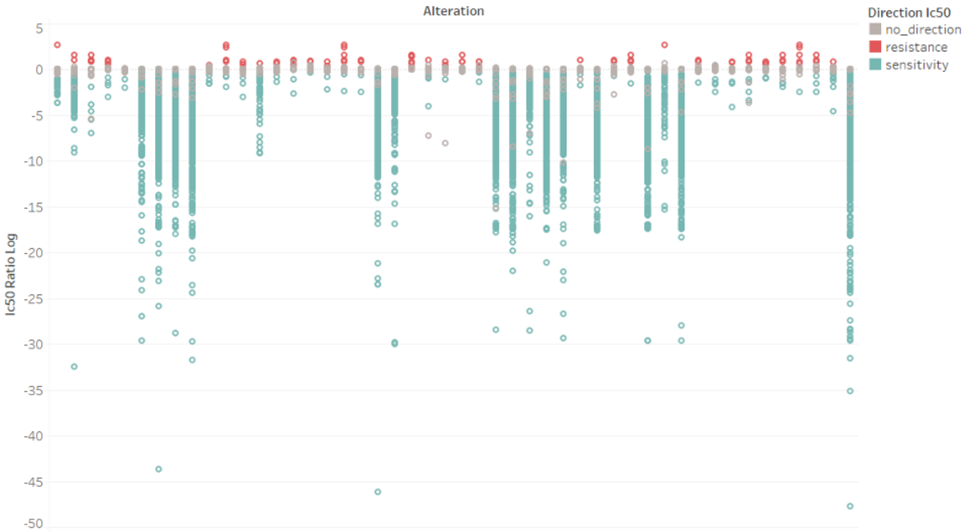
Running a simulation



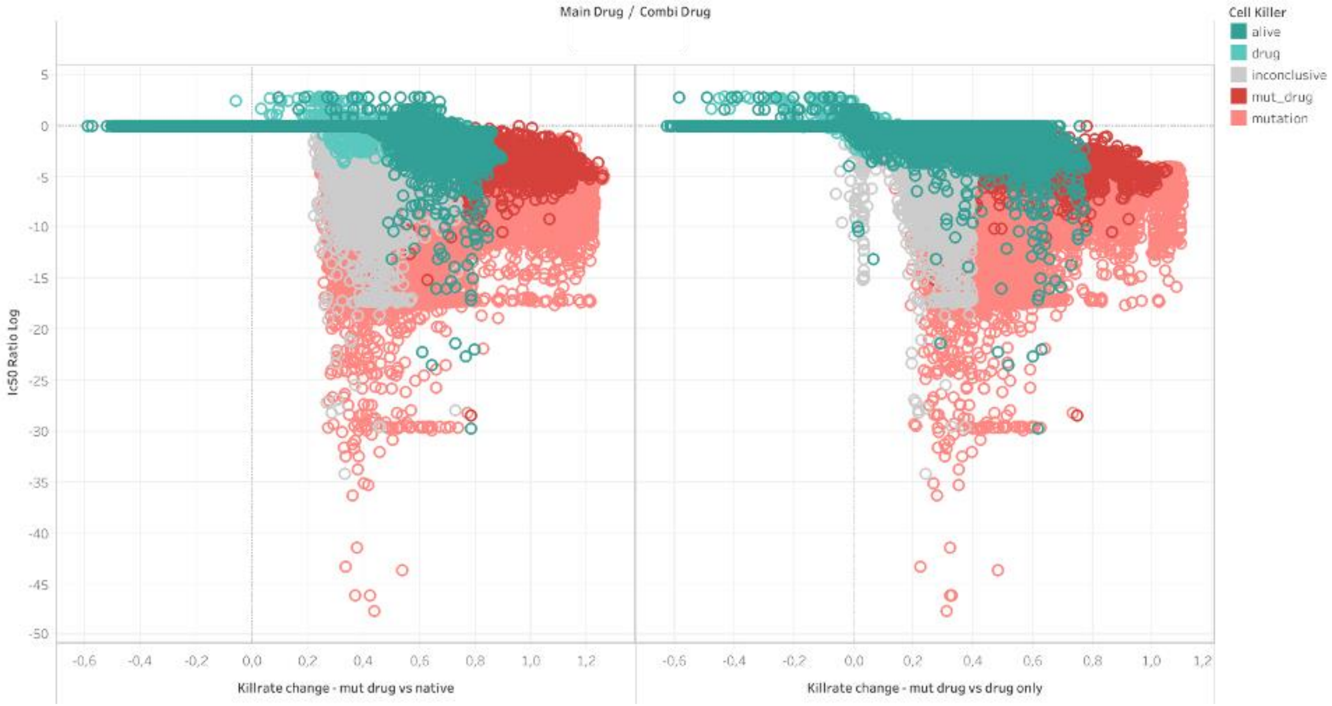
Experiments	4 381 025
Runtime	44 hours 14 minutes
Shards	490 (1.5h on average)
Cost	951 USD
Stored data	214 GB

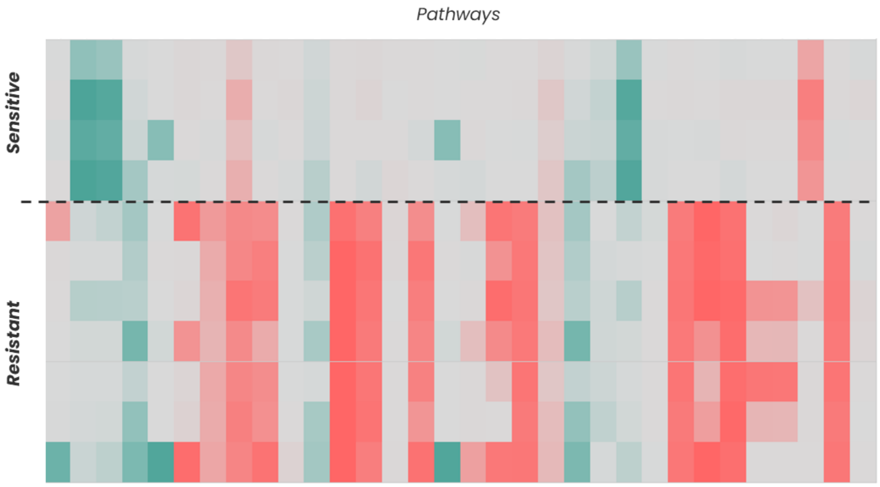
How we interpret it?

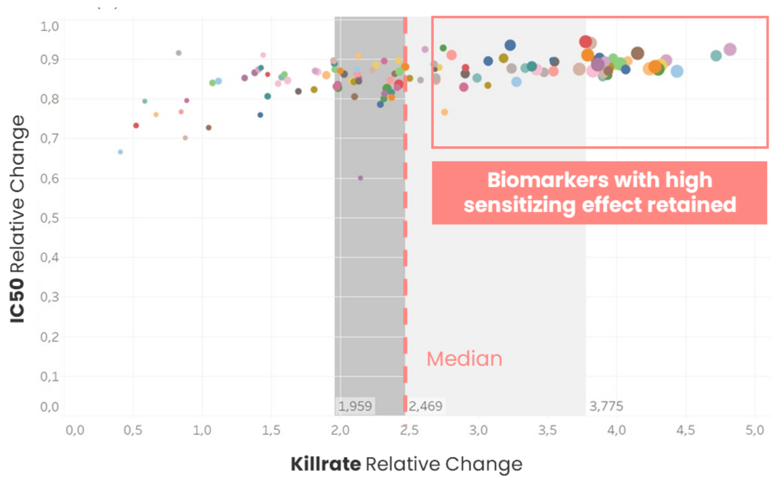
IC50 ratio logs per biomarker



IC50 Ratio Log and Kill rates



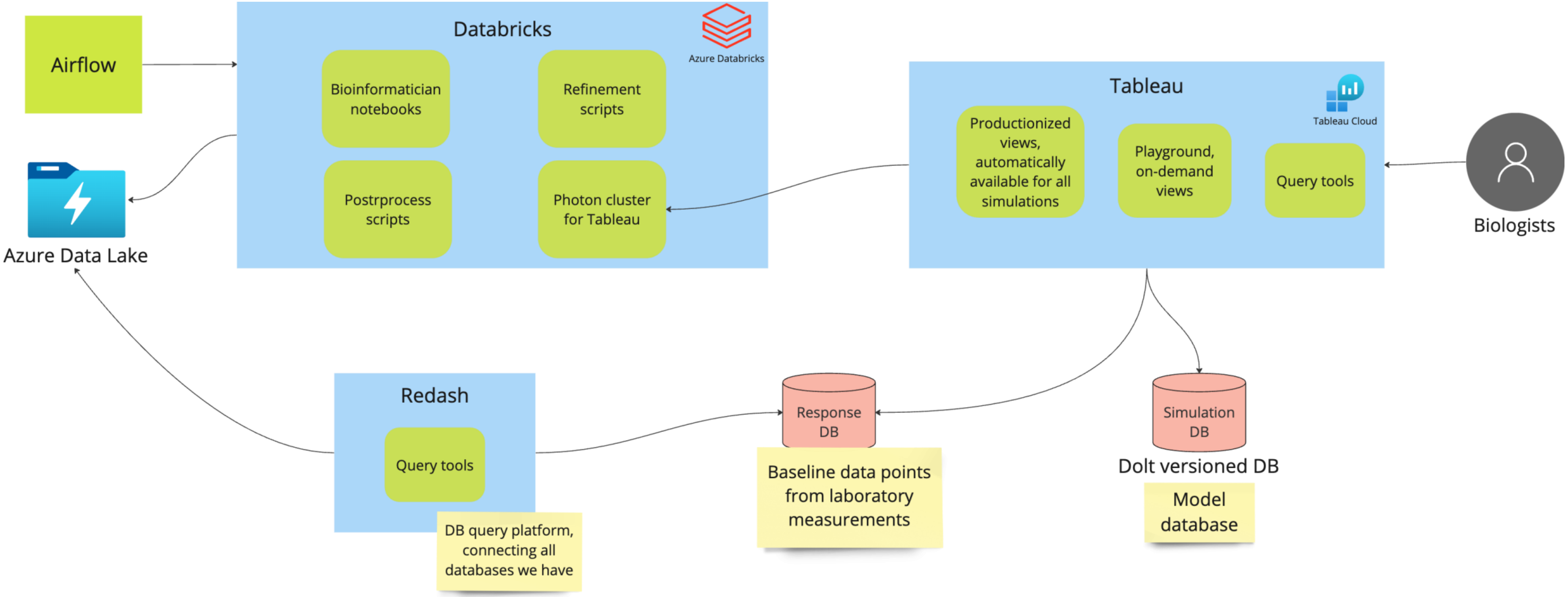




Biomarkers with high sensitizing effect retained

Data pipelines

Data processing pipelines	19
Tableau Workbooks	57
Tableau Views	298


Airflow
Databricks
Azure Databricks
Bioinformatician notebooks
Refinement scripts
Postrprocess scripts
Photon cluster for Tableau
Azure Data Lake
Tableau
Tableau Cloud
Productionized views, automatically available for all simulations
Playground, on-demand views
Query tools
Biologists
Redash
Query tools
DB query platform, connecting all databases we have
Response DB
Baseline data points from laboratory measurements
Simulation DB
Dolt versioned DB
Model database

Software development teams

Simulation

- 2 software engineers
- 2 infrastructure engineers
- 2 bioinformaticians

Simulation

- 4 software engineers
- 1 biologist / data analyst

MRT

- PO
- 1 software engineer
- 2 ML engineers
- 2 biologists

IRT

- PO
- 1 data analyst
- 1 bioinfo / data analyst

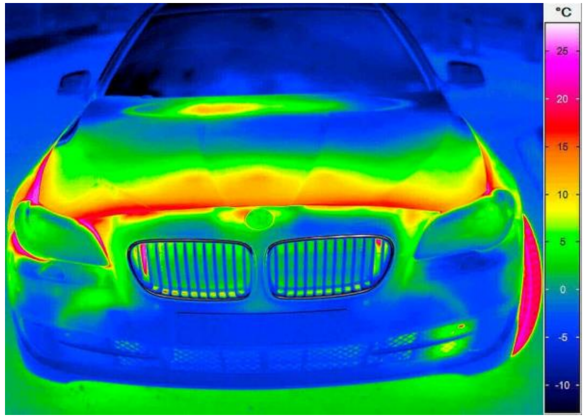
Research at Turbine

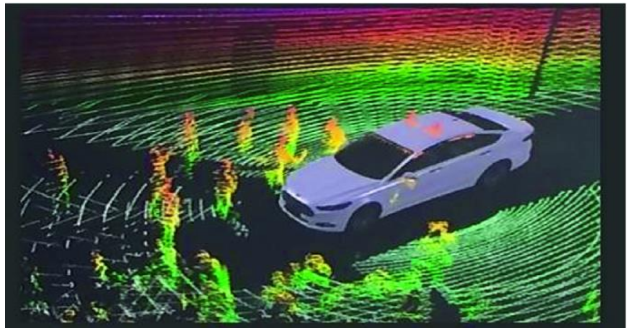
- Product: stable daily operations
- Research: new bold ideas

- AI meets biology – physicist, mathematician, bioinfo, CS...
- Applied research - deliver working prototypes

The challenge

- Predictive models: how cells respond to interventions?
  - Drug, CRISPR, RNAi...
- Features of a cell? Many modalities & detail level

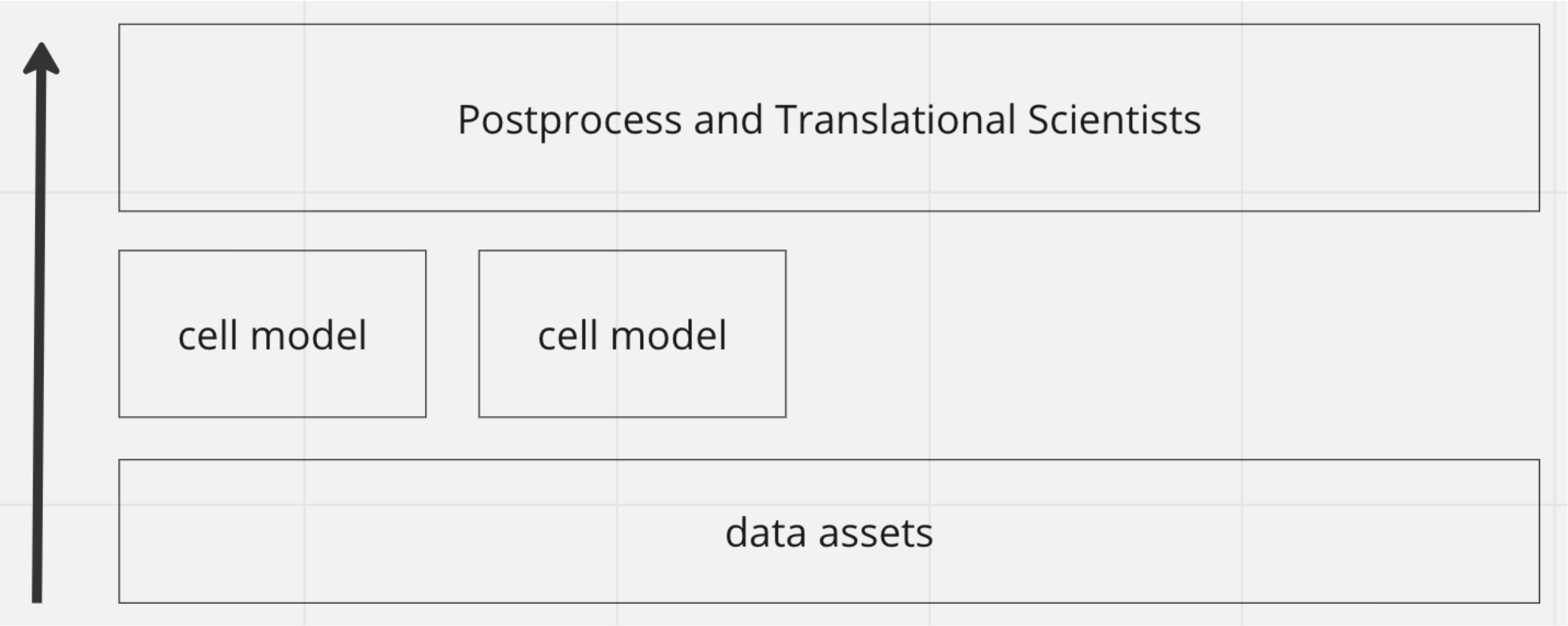




The challenge

- Predictive models: how cells respond to interventions?
  - Drug, CRISPR, RNAi...
- Features of a cell? Many modalities & detail level
  - Genomics
  - Transcriptomics
  - Drug binding properties
  - Molecular information
  - Cell-line / patient identity
- OOD generalization to new test sets – drug discovery in action
- What data to generate?

Turbine's pipeline



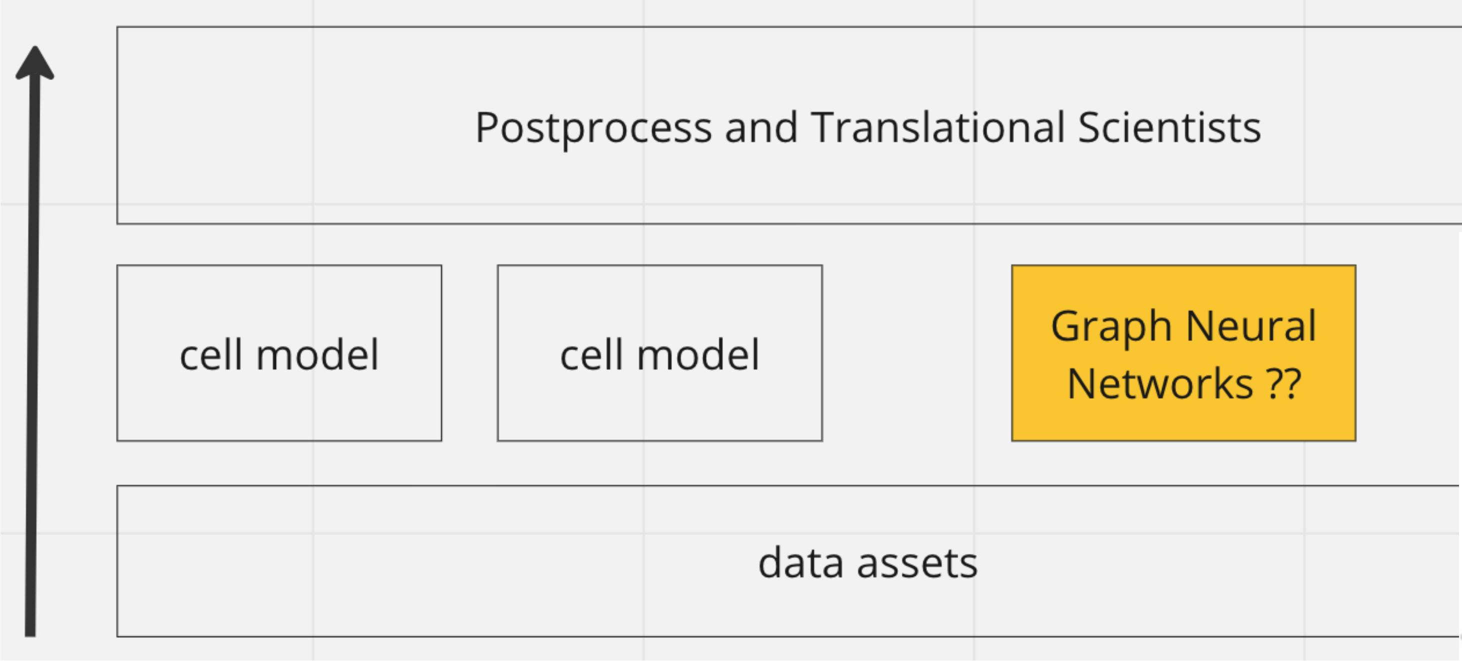
Postprocess and Translational Scientists

cell model

cell model

data assets

Turbine's pipeline – Are GNNs a good fit?



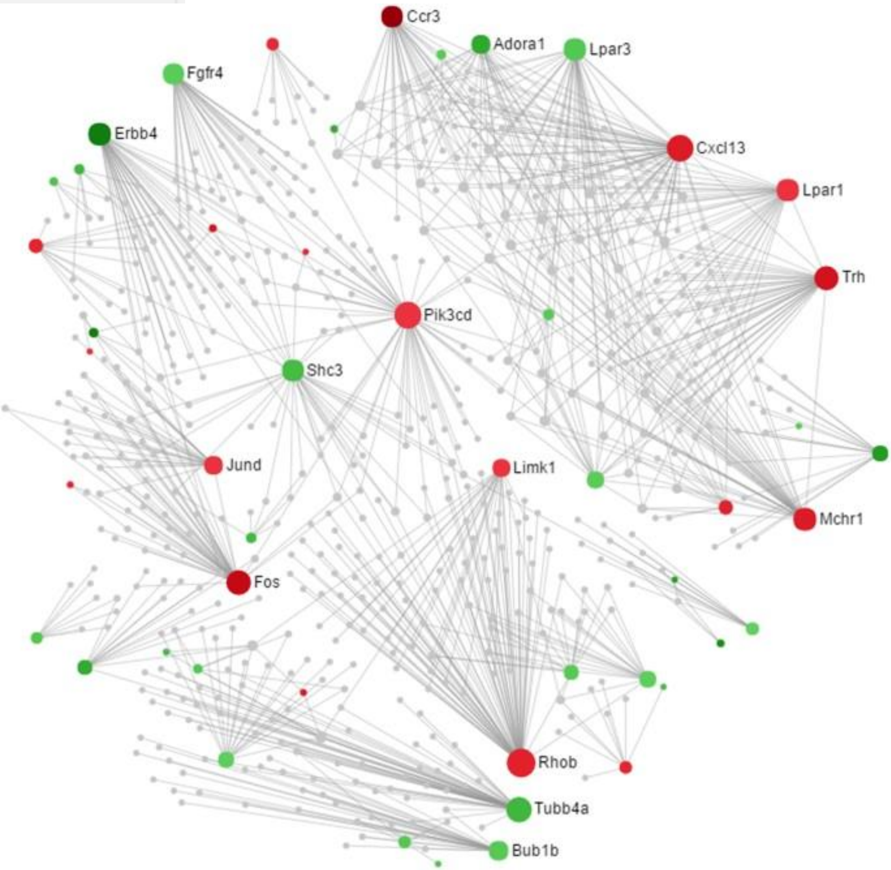
Postprocess and Translational Scientists

cell model

cell model

Graph Neural Networks ??

data assets



- Cell is a network of proteins
- Introduce graph-like priors -> address curse of dimensionality
- Re-use what works & develop where needed

Geometric deep learning

- "Erlangen programme of ML" (ICLR 2021, M. Bronstein)
- Unifying theory of effective NN architectures

- **Math**, 19th Century: Non-Euclidean geometries (projective, affine, hyperbolic...). Which is the true one?
- Felix Klein, 1872, Erlangen Univ. – study of invariance & symmetries – unification of geometry
- *Similar in **physics** later -> from conservation laws as symmetries (Noether, 1912) to 1975 standard model*

Geometric deep learning

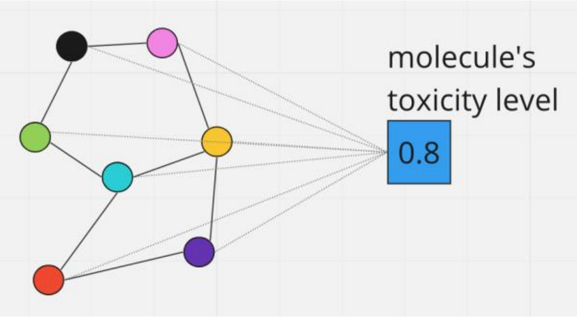
- Many de facto models we use. Similar to state of geometry in 19th century. Why do they work? What's in common?
  - RNN, CNN, GNN, transformer...

Benefits of geometric DL

1. Common math framework to derive the best NN architectures
2. Constructive methods to build **new architectures**

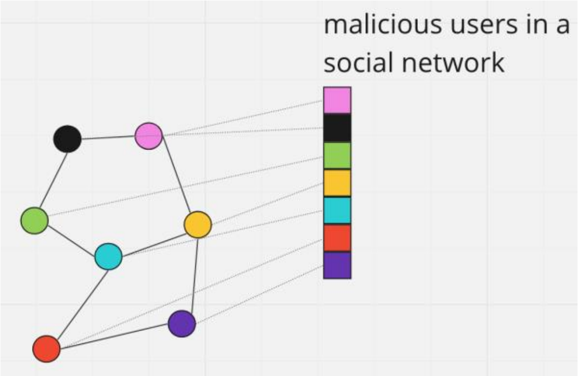
- Learn on non-Euclidean problems like graphs, meshes of 3D objects, maps...
  - No spatial locality (e.g for CV pixels and NLP seqs) - how close are 2 nodes?
  - Translational equivariance -> nope
  - Coordinates of a node? -> nope

Graph Neural Networks - components



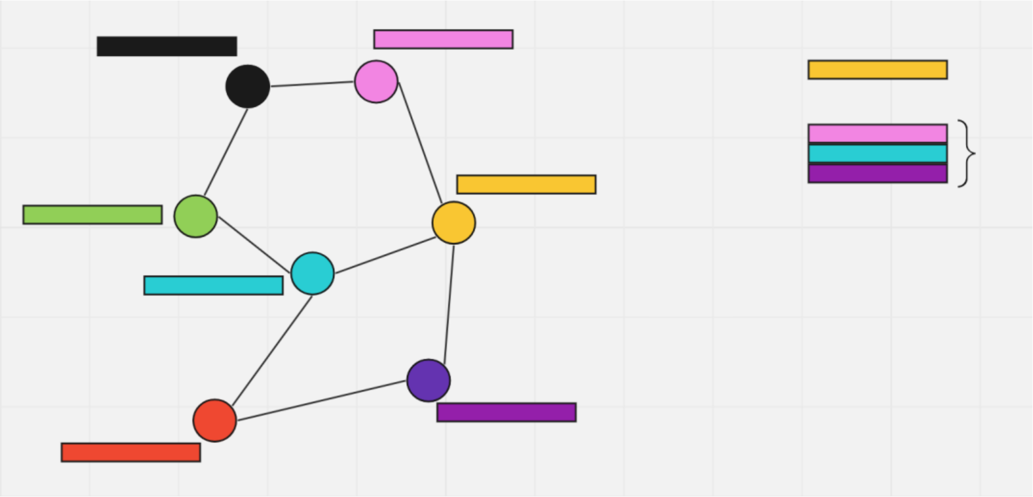
Permutation-invariant operators

(output unaffected by node ordering)



Permutation-equivariant operators

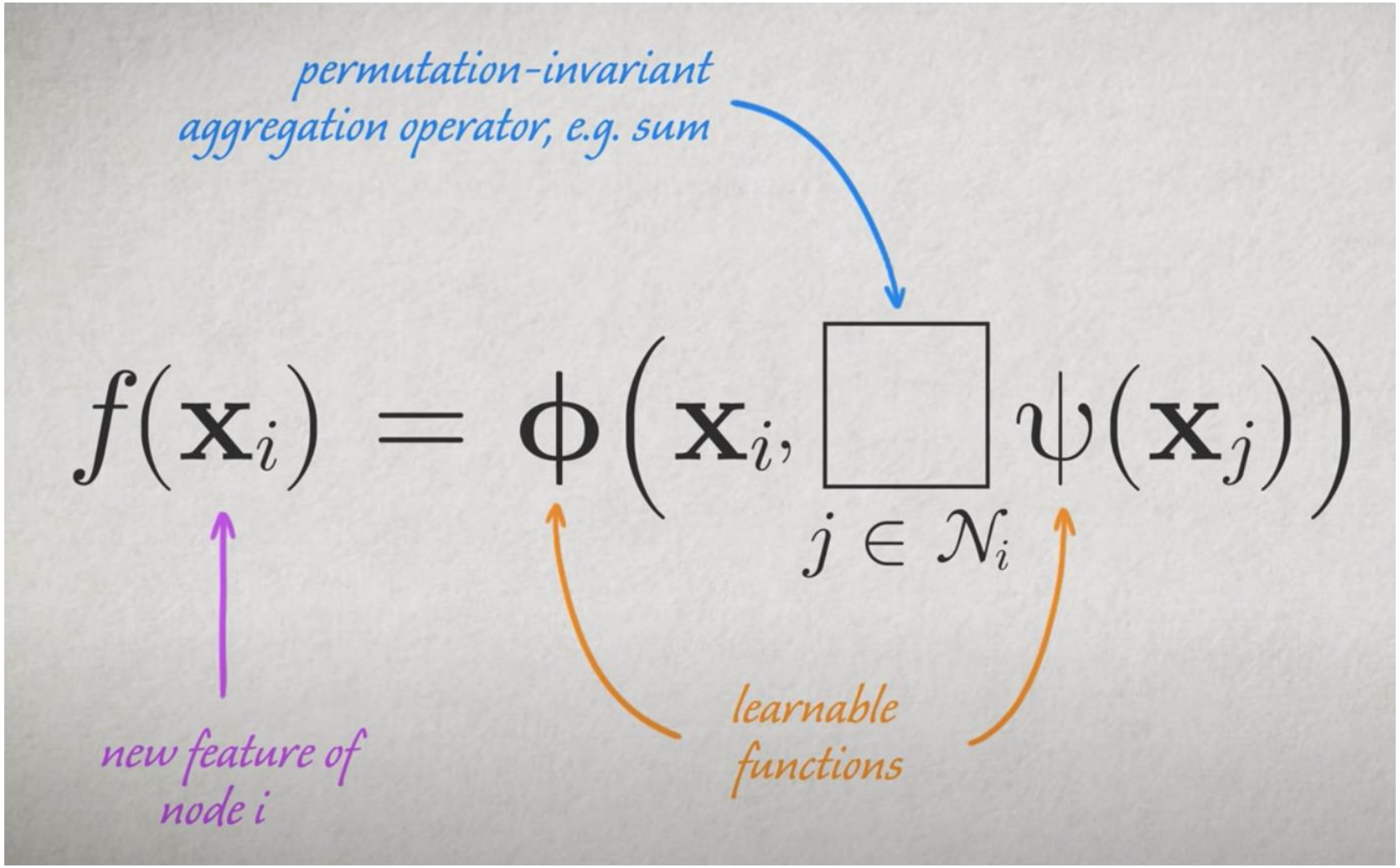
(output changes as the ordering of nodes)



Update rule of node representation

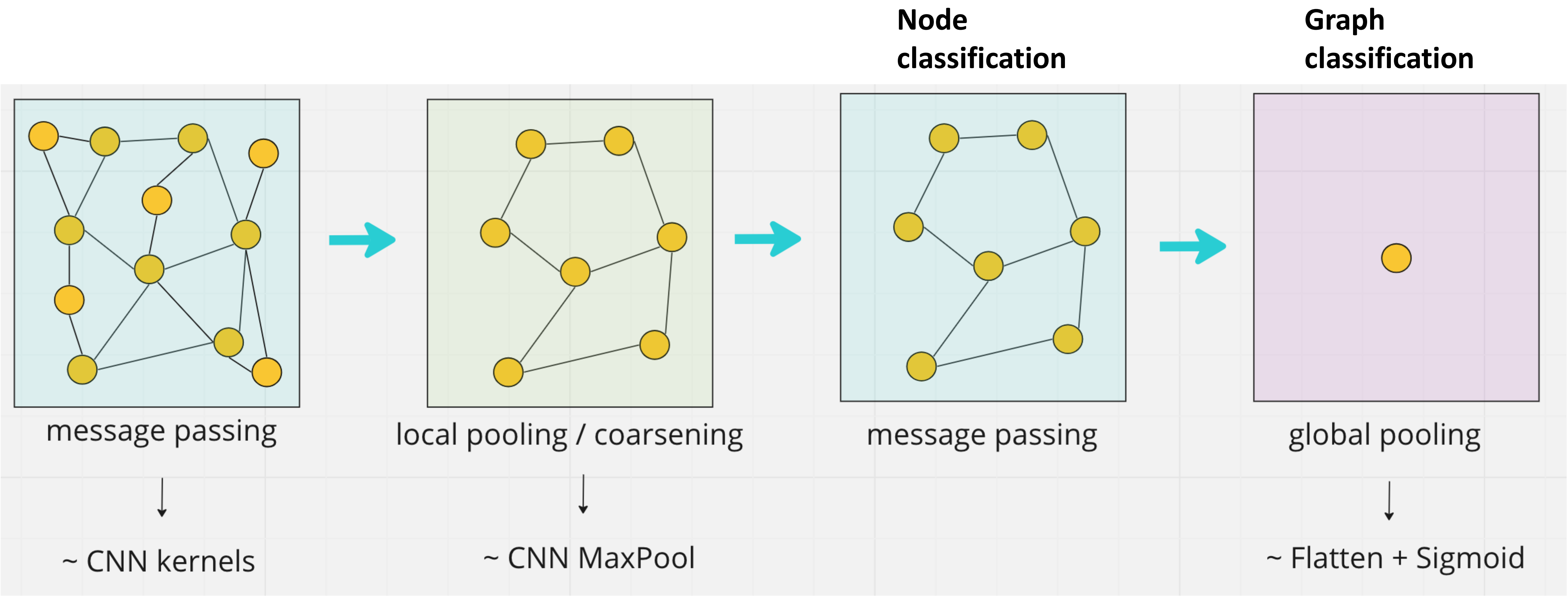
- Local neighbourhood
- Permutation-invariant update rule
- 1 step of depth in a GNN

Graph Neural Networks - math framework



$$f(\mathbf{x}_i) = \boldsymbol{\phi}\left(\mathbf{x}_i, \underset{j \in \mathcal{N}_i}{\square} \psi(\mathbf{x}_j)\right)$$

Graph Neural Networks - layers



Node classification

Graph classification

message passing

local pooling / coarsening

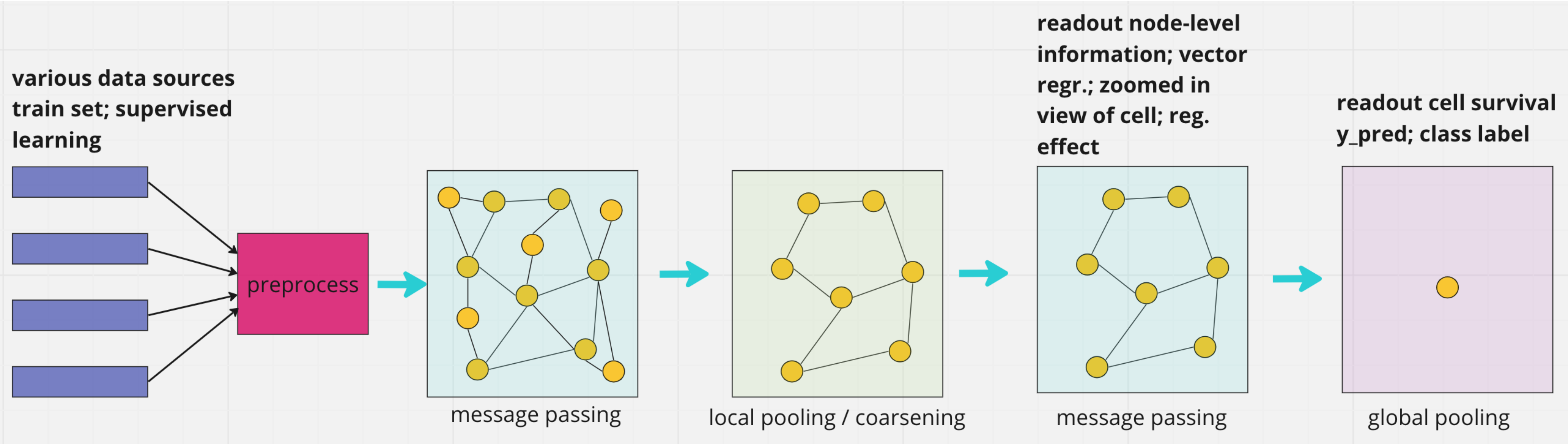
message passing

global pooling

~ CNN kernels

~ CNN MaxPool

~ Flatten + Sigmoid


various data sources train set; supervised learning
preprocess
message passing
local pooling / coarsening
readout node-level information; vector regr.; zoomed in view of cell; reg. effect
message passing
readout cell survival y_pred; class label
global pooling

- Use a wide spectrum of input features
- Leverage representation learning
- Can combine graph and node level objective functions
- 4 months to a new prototype model progress towards production
  - Works on large datasets w/o overfit
  - Performs well on proprietary benchmark problems (predict drug, gene KO... treatments)

And there is more ahead...

Senior ML Engineer

Research

- Novel algorithms
- Custom layer design
- Graph ML
- Heavy biology domain

ML Ops Engineer

Product

- Large scale training
- Prod ready code
- Overview full ML model lifecycle

Senior Bioinformatician

Research

- Bio. data processing
- Dataset and metrics design
- Add domain knowledge to AI systems