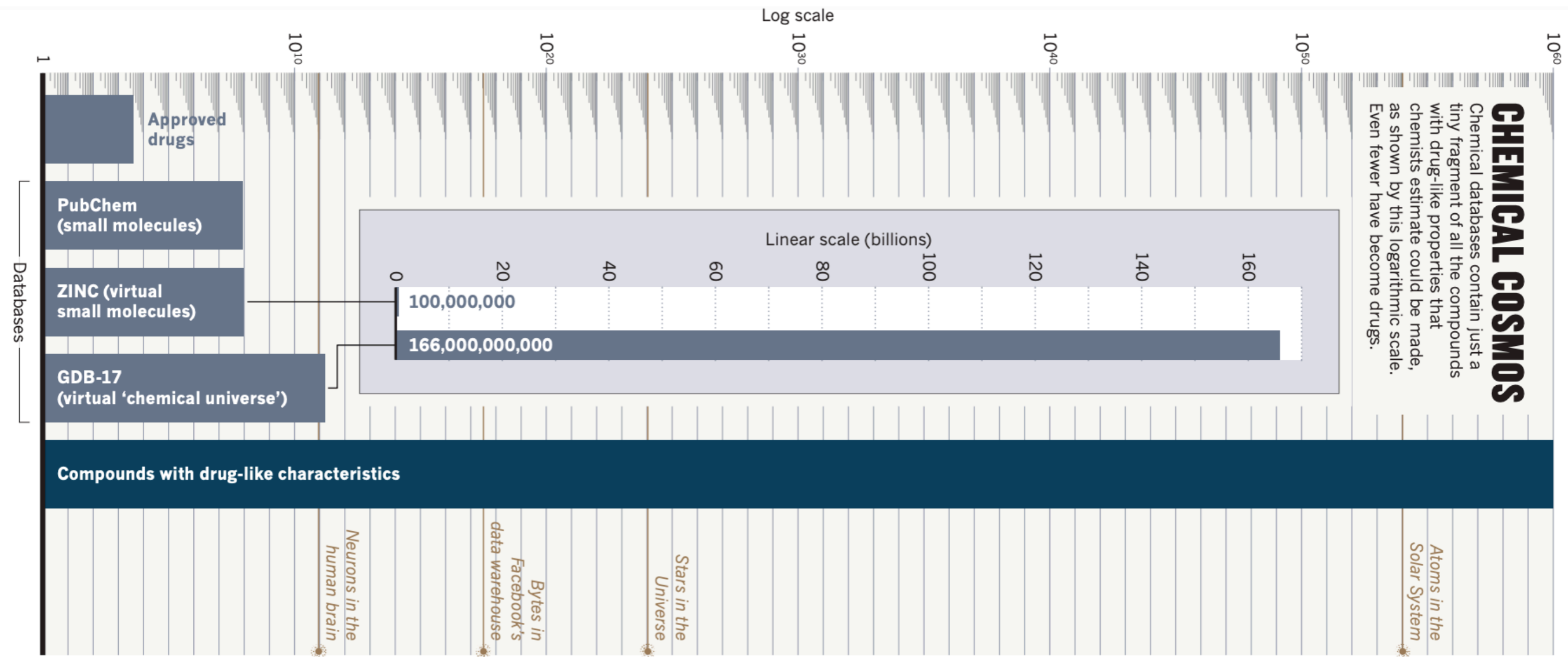


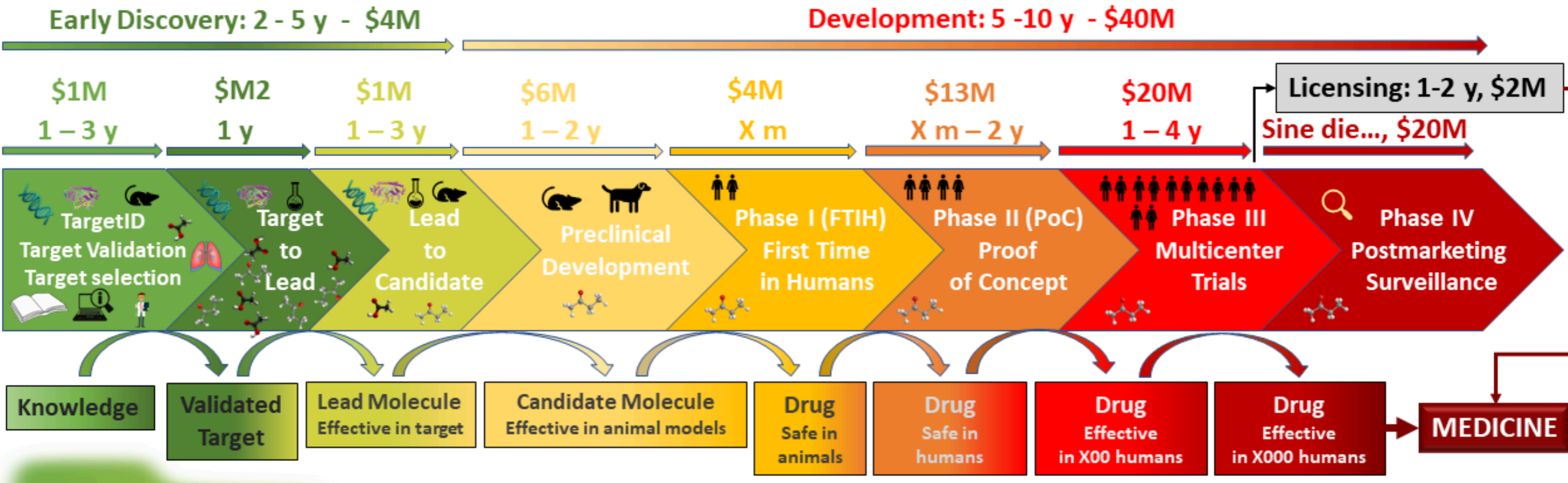
AI for drug discovery

Sheng Wang

Number of compounds with drug-like characteristics: 10^{60}



Drug discovery process

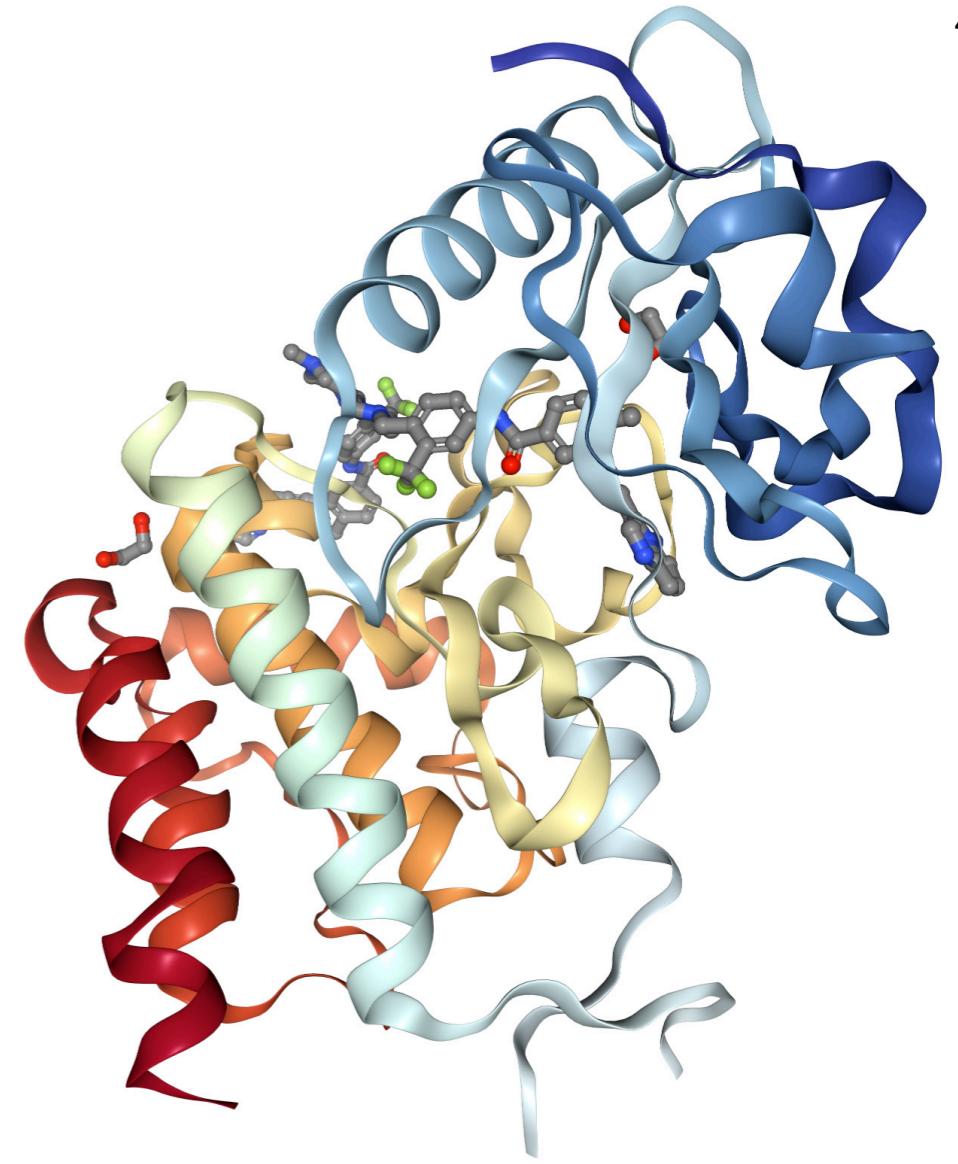
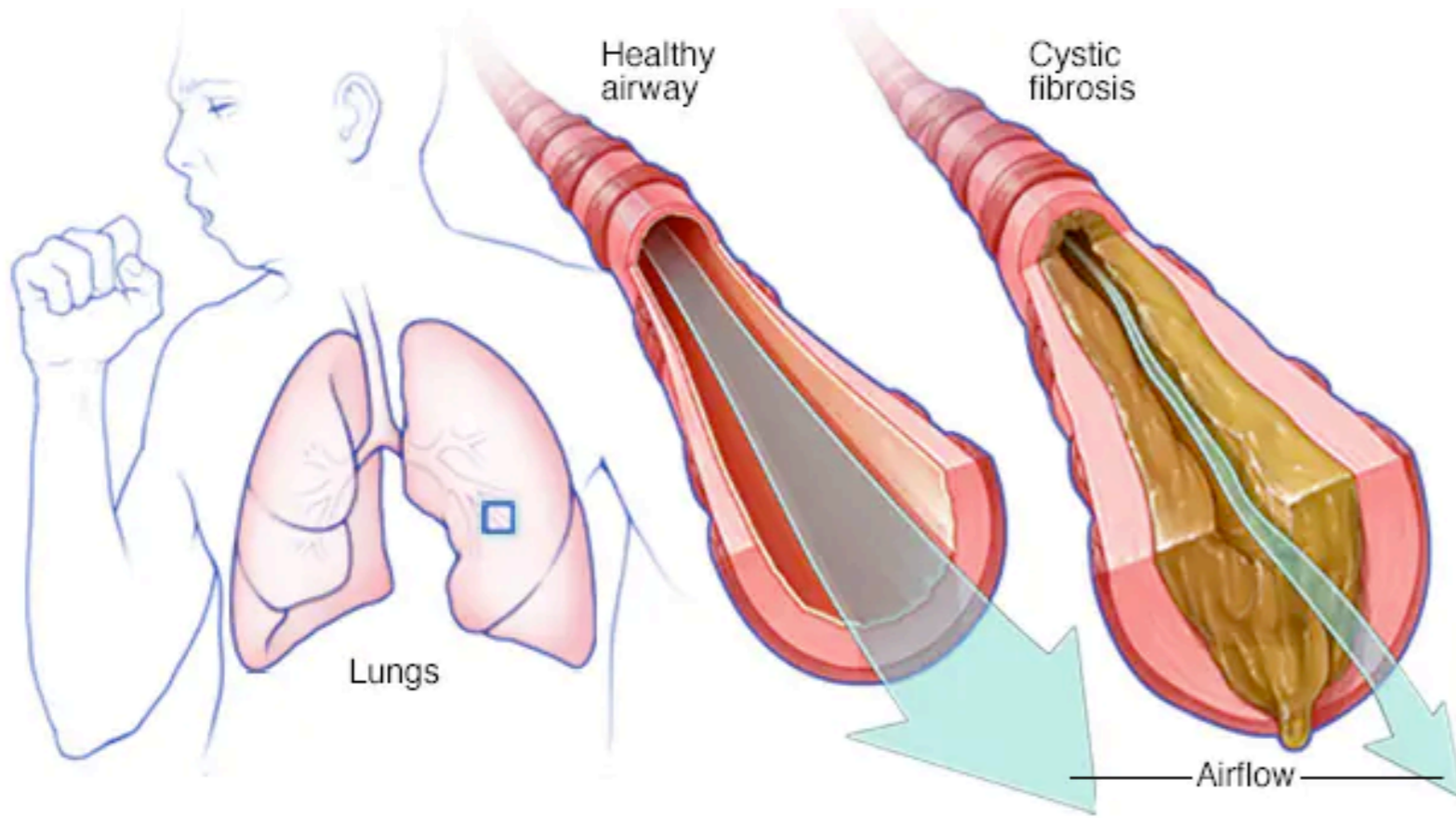


Target: human genome products (proteins). Obtained using disease-relevant cell/tissue/animal model. We need to know the function of this target.

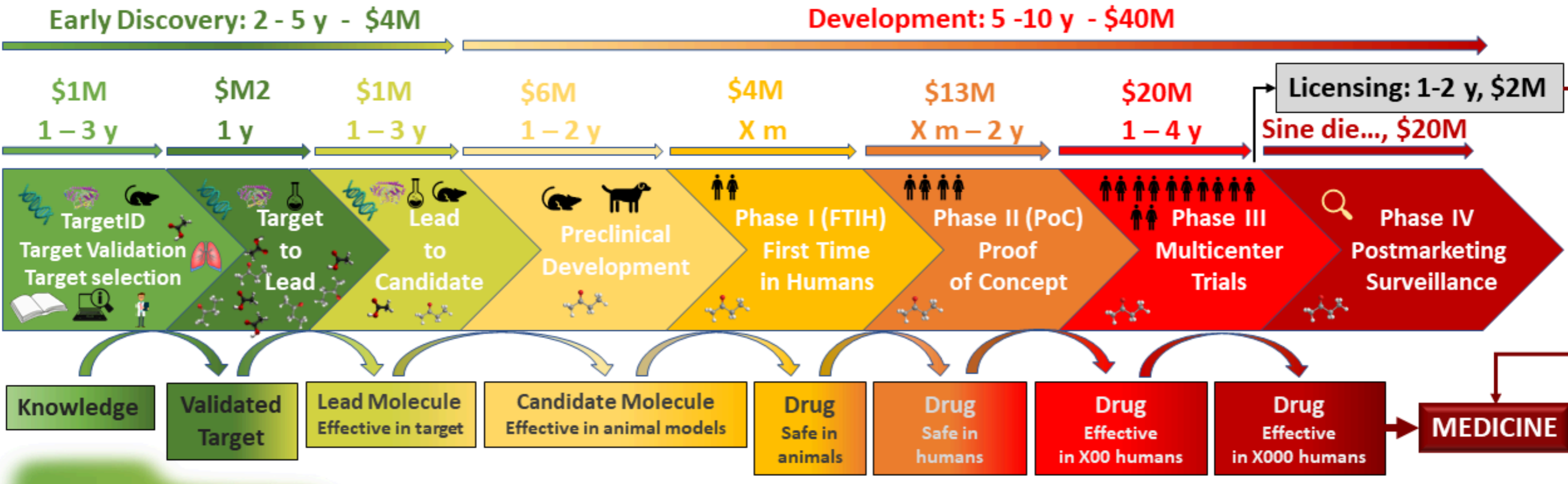
On the basis of existing knowledge, we were able to determine that all current drugs with a known mode-of-action act through **324 distinct molecular drug targets**. Of these, 266 are human-genome-derived proteins, and the remainder are bacterial, viral, fungal or other pathogenic organism targets

Disease: fibrosis, lung tissue becomes thick

Target: DDR I



Drug discovery process



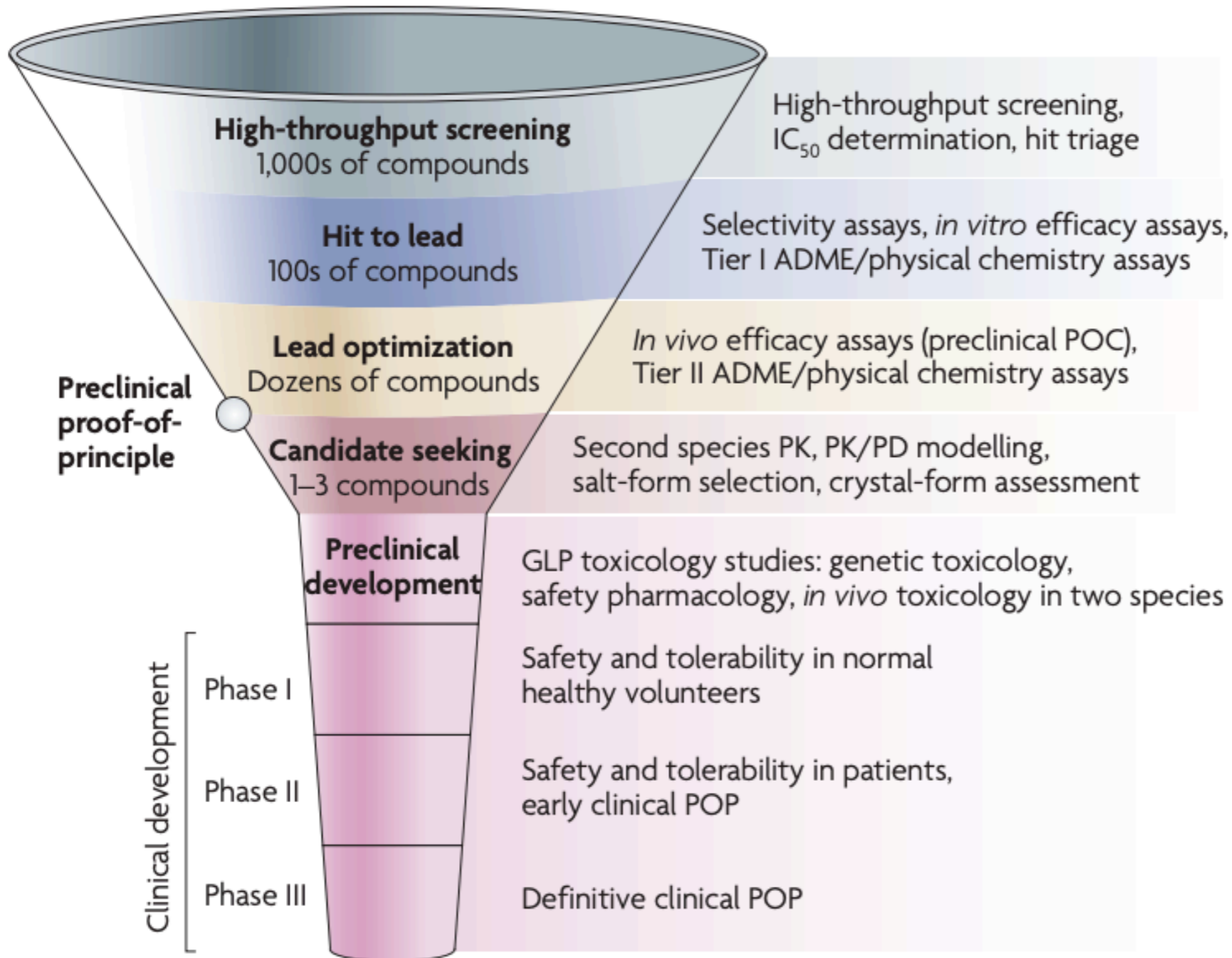
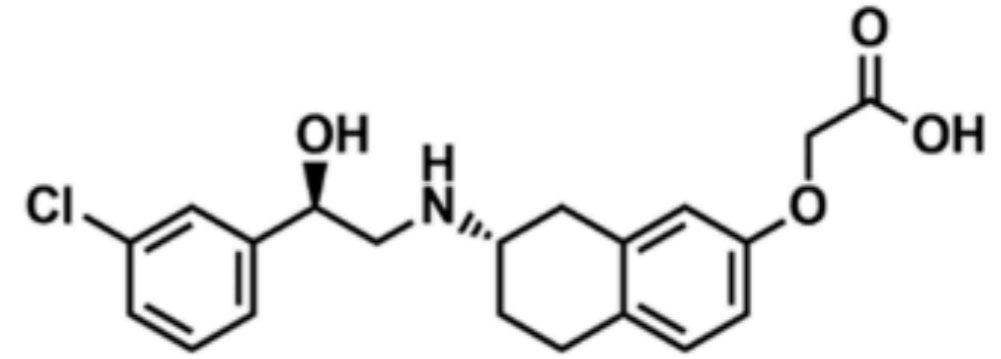
Target: human genome products (proteins). Obtained using disease-relevant cell/tissue/animal model. We need to know the function of this target.

Hit: Use high-throughput screening to test 1000 compounds. 100 of them are positive (Hits).

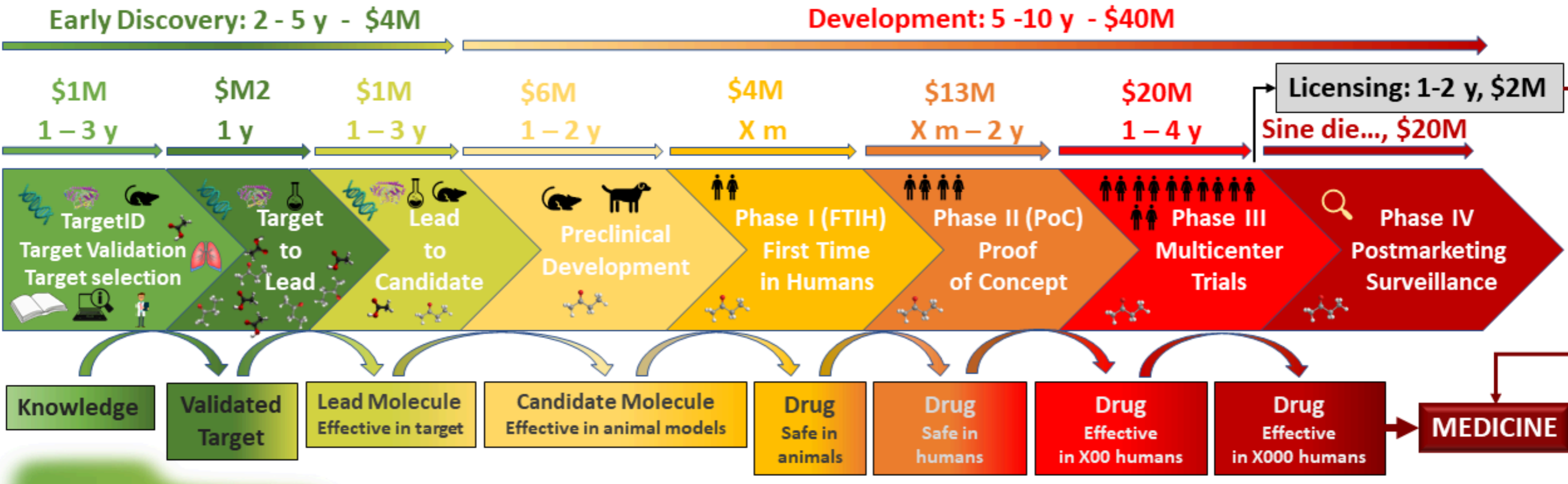
Lead: Test other properties of these 100 compounds (we don't want a compound that is effective to many targets). 10 of them pass the test (Leads).

Candidate: Lead might have suboptimal structure. Modify it to get 1-2 compounds (Candidates).

Hit, lead, candidate



Drug discovery process



Target: human genome products (proteins). Obtained using disease-relevant cell/tissue/animal model. We need to know the function of this target.

Hit: Use high-throughput screening to test 1000 compounds. 100 of them are positive (Hits).

Lead: Test other properties of these 100 compounds (we don't want a compound that is effective to many targets). 10 of them pass the test (Leads).

Candidate: Lead might have suboptimal structure. Modify it to get 1-2 compounds (Candidates).

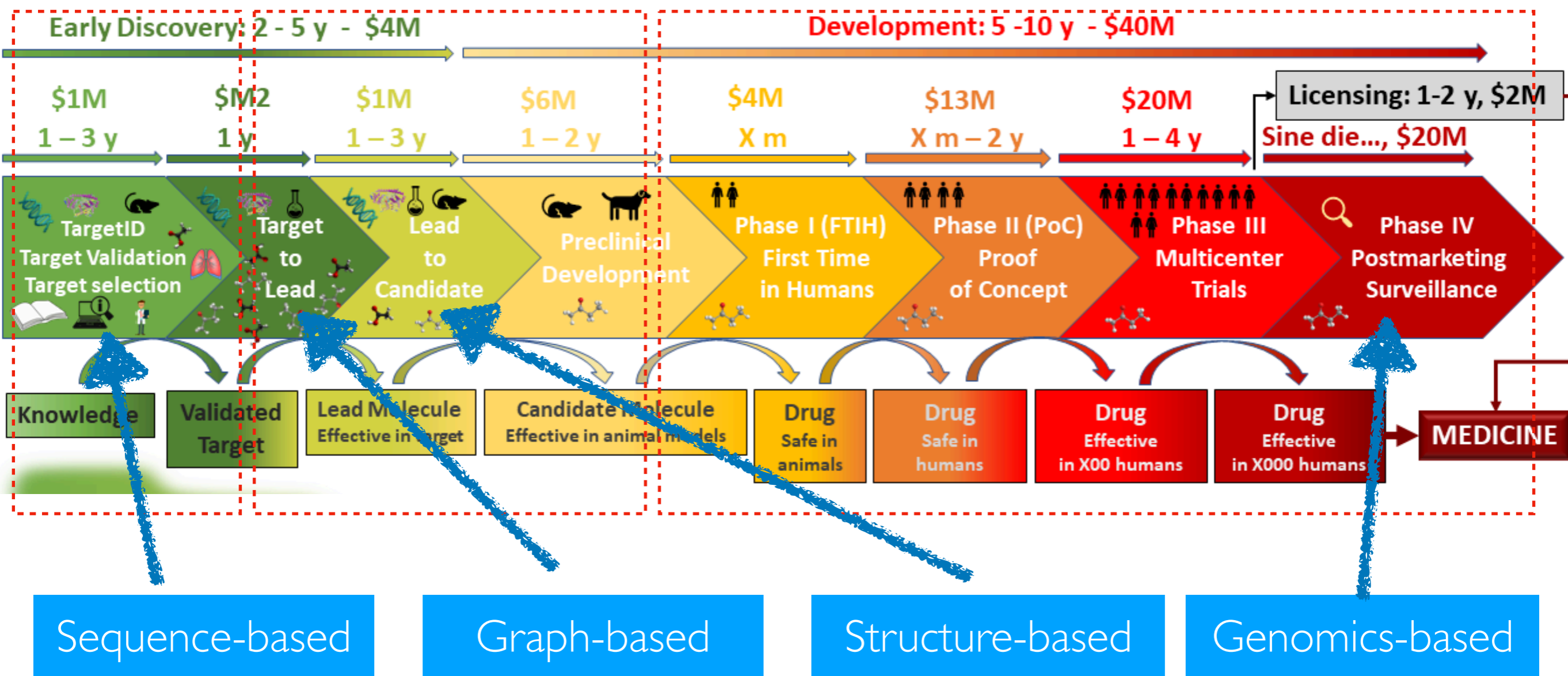
Phase 1: first-in-human. Maximum amount of dose before side effects.

Phase 2: 100-300 people. Test effectiveness.

Phase 3: randomized controlled multicenter trials on 300-3000 patients.

Phase 4: After drug can be sold. Safety surveillance.

CSE 599: AI for drug discovery



Sequence: understand target function using protein sequence. NLP to find targets (word sequence).

Graph: generate compound graph 2D structure (deep generative model)

Structure: modify structure according to 3D structure (geometric deep learning)

Genomics: side effects, personalized efficacy, repurposing, etc. (multi-modality)




Understand drug discovery pipeline using one paper

BRIEF COMMUNICATION

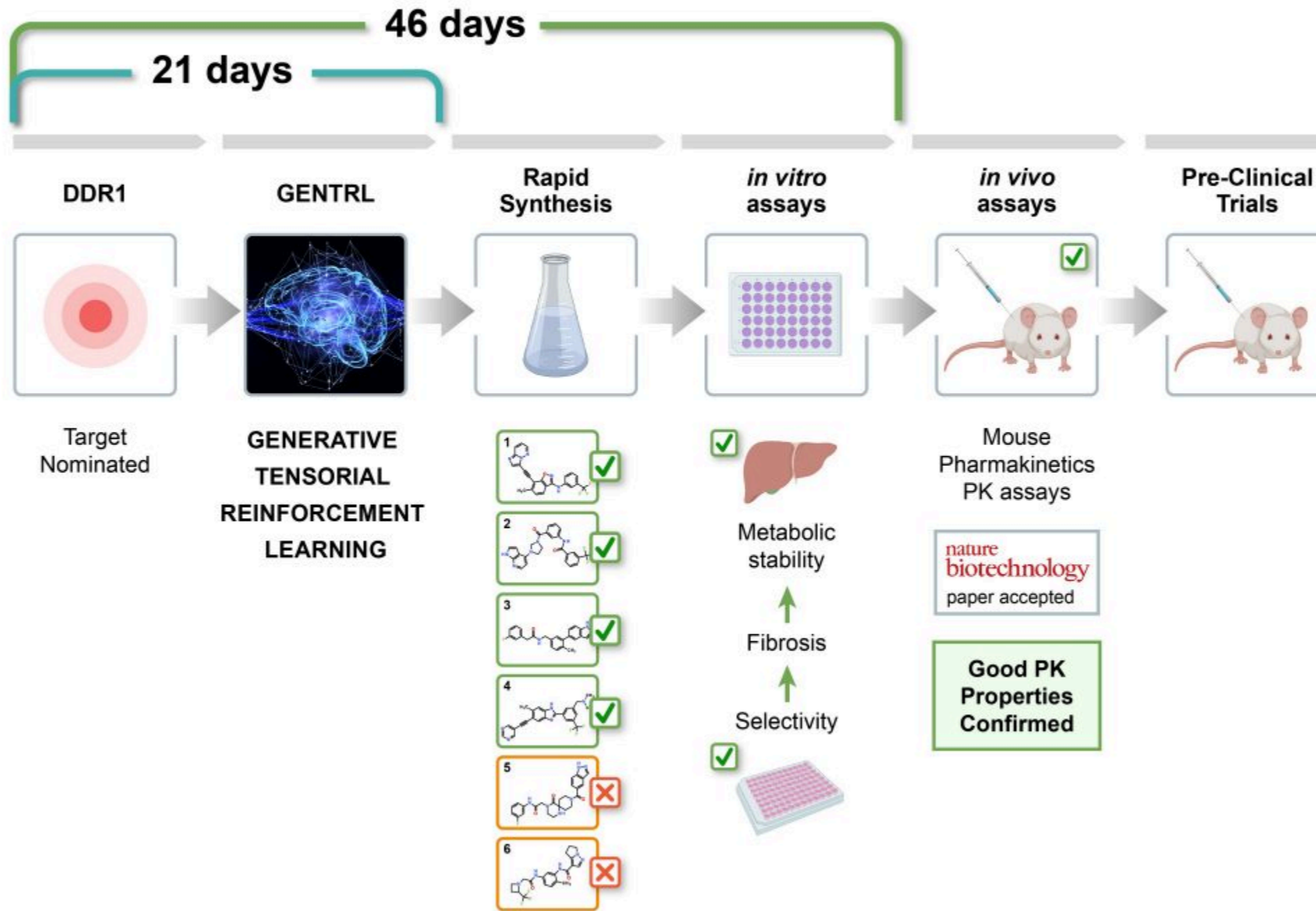
<https://doi.org/10.1038/s41587-019-0224-x>

**nature
biotechnology**

Deep learning enables rapid identification of potent DDR1 kinase inhibitors

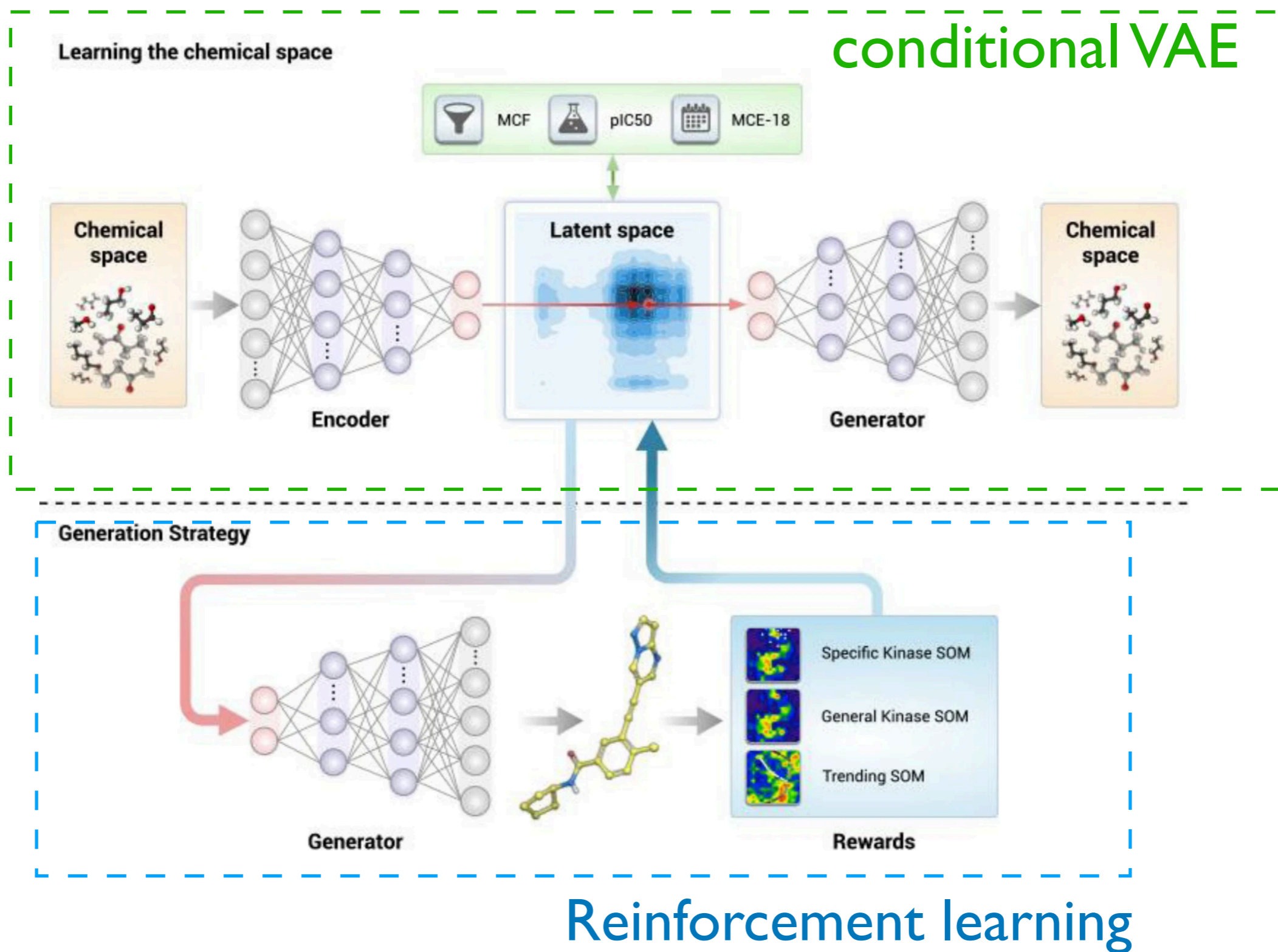
Alex Zhavoronkov ^{1*}, Yan A. Ivanenkov¹, Alex Aliper¹, Mark S. Veselov¹, Vladimir A. Aladinskiy¹, Anastasiya V. Aladinskaya¹, Victor A. Terentiev¹, Daniil A. Polykovskiy¹, Maksim D. Kuznetsov¹, Arip Asadulaev¹, Yury Volkov¹, Artem Zholus¹, Rim R. Shayakhmetov¹, Alexander Zhebrak¹, Lidiya I. Minaeva¹, Bogdan A. Zagribelnyy¹, Lennart H. Lee ², Richard Soll², David Madge², Li Xing², Tao Guo ² and Alán Aspuru-Guzik^{3,4,5,6}

A 46-day pipeline to discovery a new DDR1 inhibitor



Traditional hit-lead generation: 2–3 years
GENTRL approach: <2 months

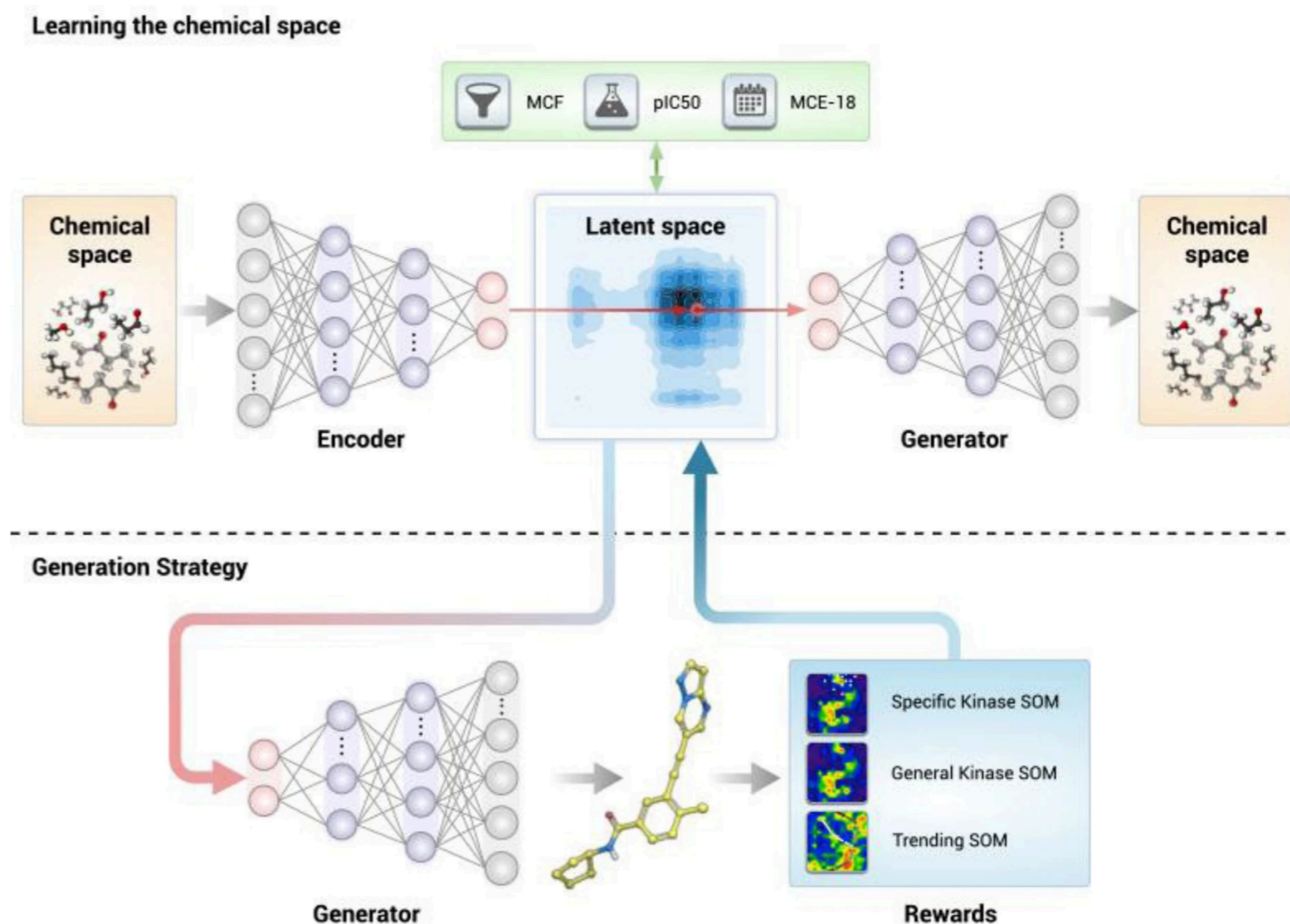
ML: deep generative model + reinforcement learning



A two-step ML approach to develop DDRI inhibitor

5 datasets for ML model

1. A large set of molecules from ZINC database
2. Known DDRI kinase inhibitor
3. Common kinase inhibitor (positive set)
4. Molecules that act on non-kinase targets (negative set)
5. Patent data for molecules that are claimed by pharmaceutical companies



2 steps

1. Pretrain the cVAE
 - Molecules from ZINC to train the cVAE (dataset 1, 2, 3, 5)
 - Labels: MCE-18 (novelty), pIC50 (efficacy), MCFs (bioavailability, Lipinski's "rule of 5 + Veber's "rule of 2")
2. Reinforcement learning on the latent space
 - Given a latent vector z , use the generator (decoder) from cVAE to get chemical x
 - Reward: sum of [general(x), specific(x), trending(x)]
 - general, specific and trending are three external functions based on SOM

MEC-18

$$\text{MCE-18} = \left(\text{AR} + \text{NAR} + \text{CHIRAL} + \text{SPIRO} + \frac{\overbrace{\text{sp}^3 + \text{Cyc} - \text{Acyc}}^{\text{NCSPTR}}}{1 + \text{sp}^3} \right) \times \text{Q}^1$$

AR: presence of an aromatic or heteroaromatic ring (0 or 1)

NAR: presence of an aliphatic or a heteroaliphatic ring (0 or 1)

CHIRAL: presence of a chiral center (0 or 1)

SPIRO: the presence of a spiro point (0 or 1)

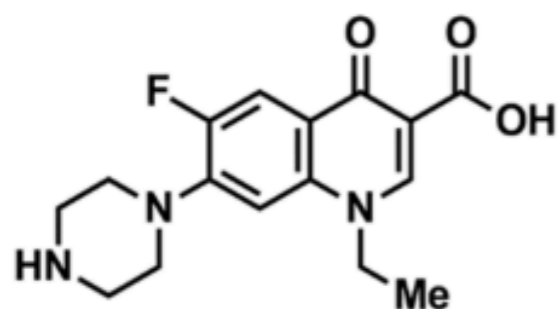
sp³: the portion of sp³-hybridized carbon atoms (from 0 to 1)

Cyc: the portion of cyclic carbons that are sp³ hybridized (from 0 to 1),

Acyc: portion of acyclic carbon atoms that are sp³ hybridized (from 0 to 1)

Q¹: the normalized quadratic index (2nd power of atoms adjacency matrix)

Simplified illustration of the evolution of medicinal chemistry

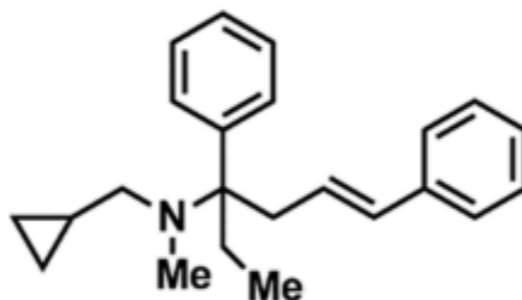


1977

(Sanofi)

 sp^3 : 37.5

MCE-18 : 42.5

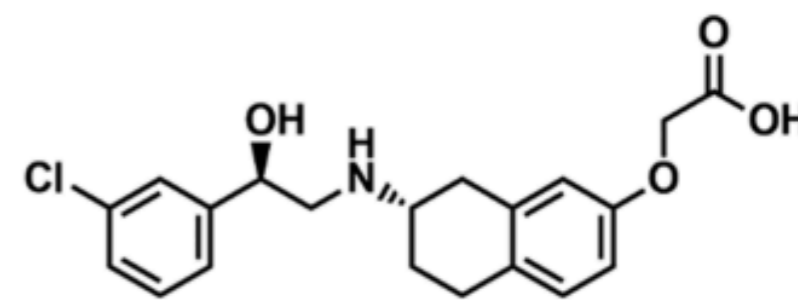


1988

(Pfizer)

 sp^3 : 39.1

MCE-18 : 51.0

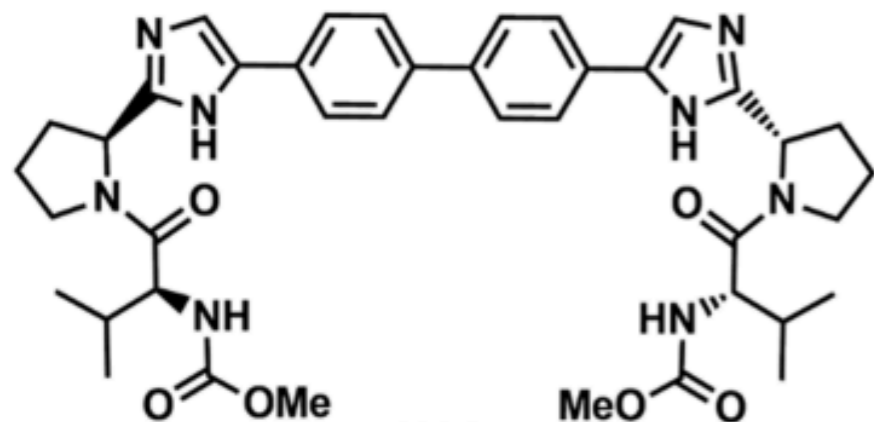


1993

(Sanofi)

 sp^3 : 35.0

MCE-18 : 56.0

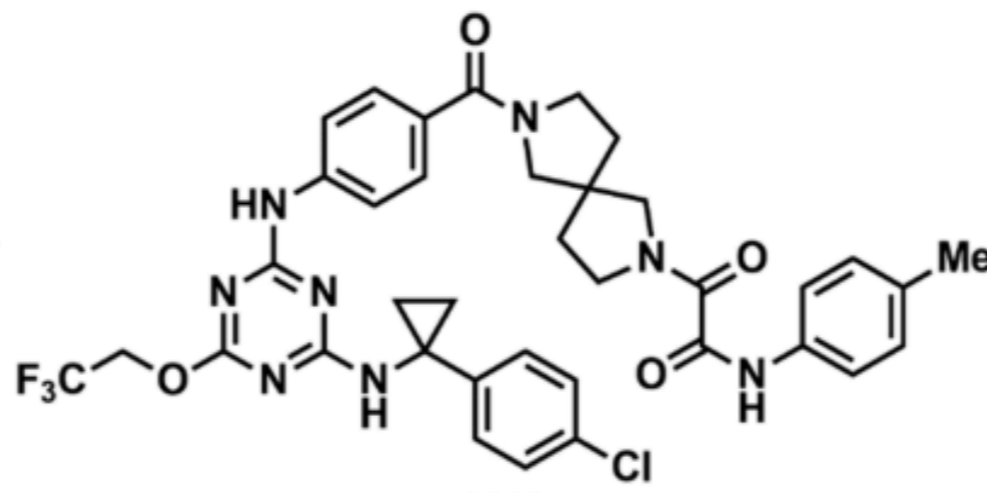


2006

(Bristol-Myers Squibb)

 sp^3 : 45.0

MCE-18 : 124.5

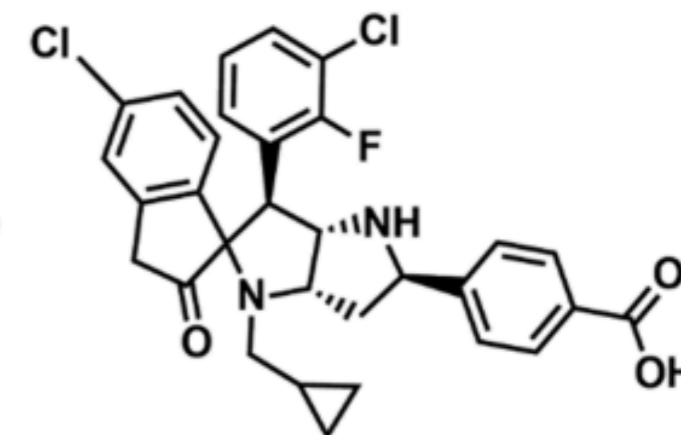


2012

(Bristol-Myers Squibb)

 sp^3 : 35.1

MCE-18 : 193.6



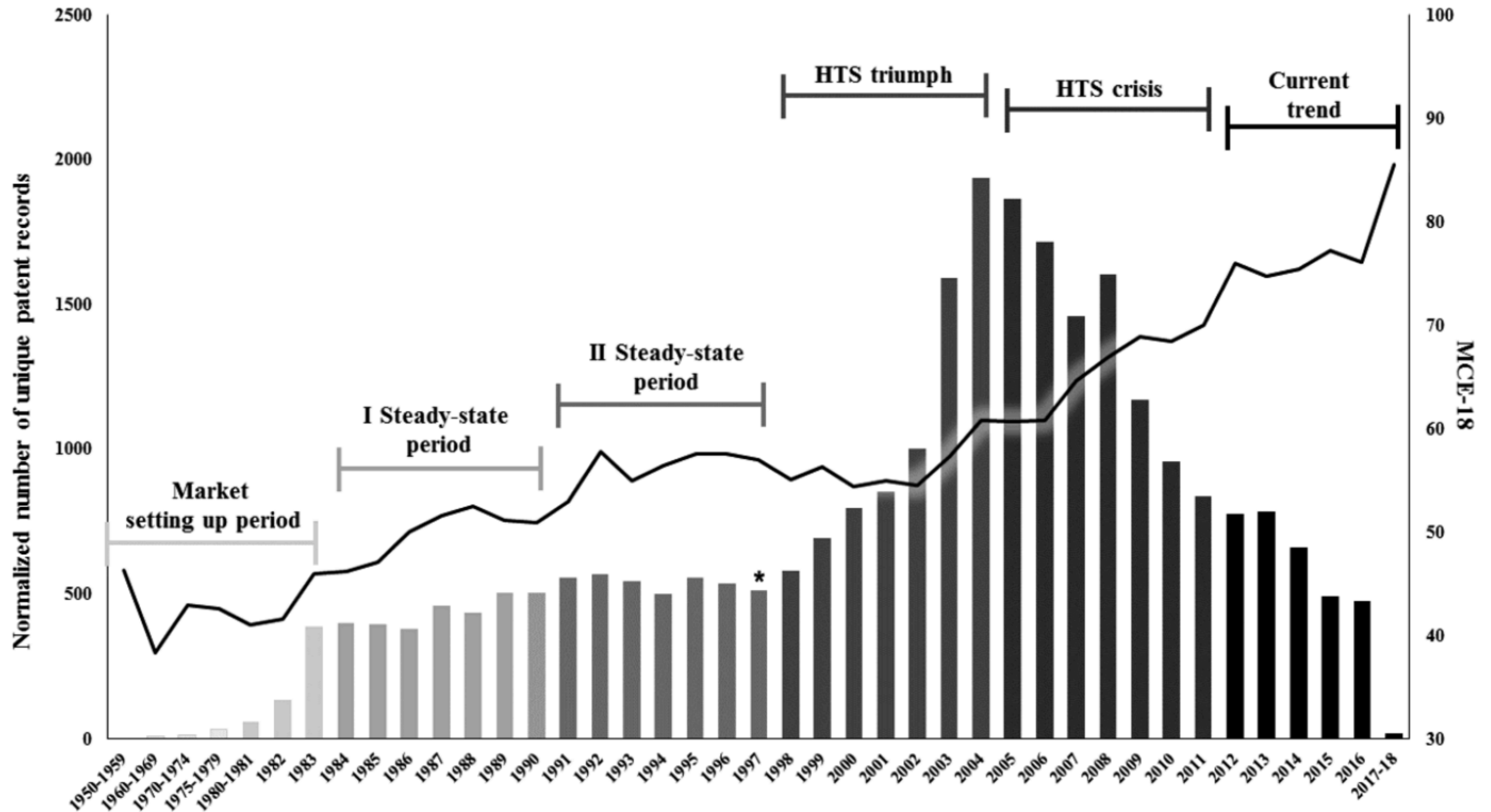
2014

(Boehringer)

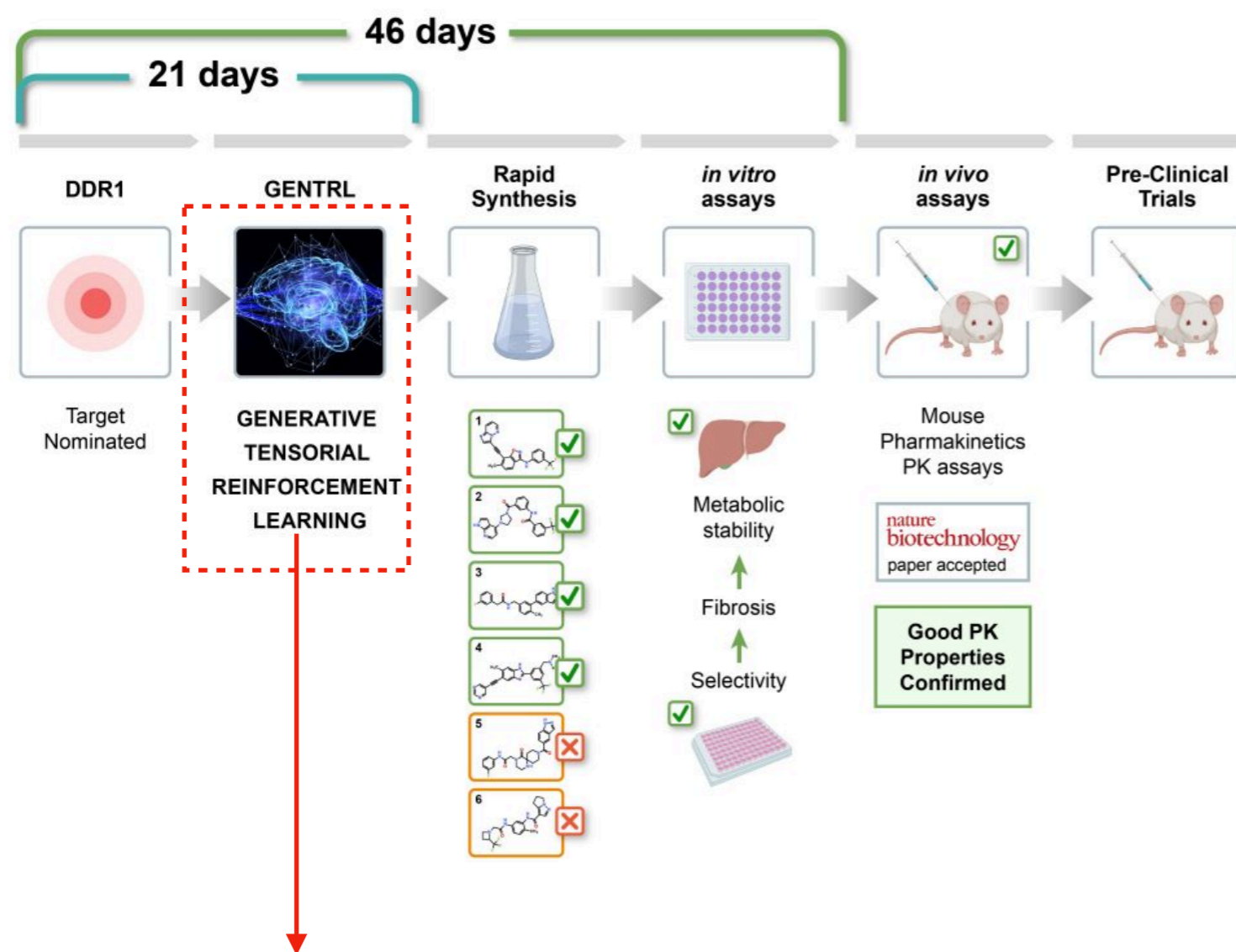
 sp^3 : 33.3

MCE-18 : 178.0

MCE-18 reflects patent priority date

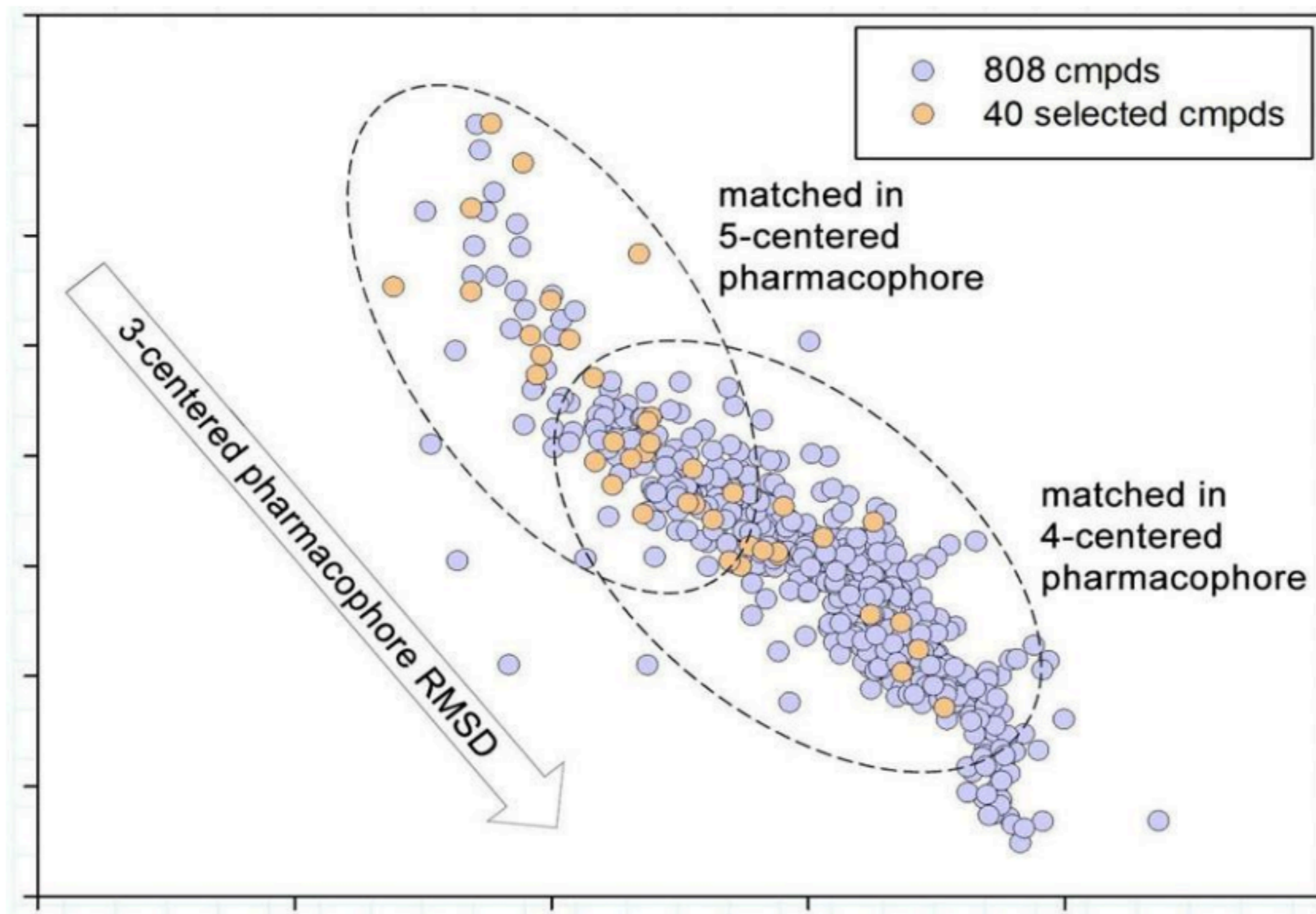


A 46-day pipeline to discovery a new DDR1 inhibitor



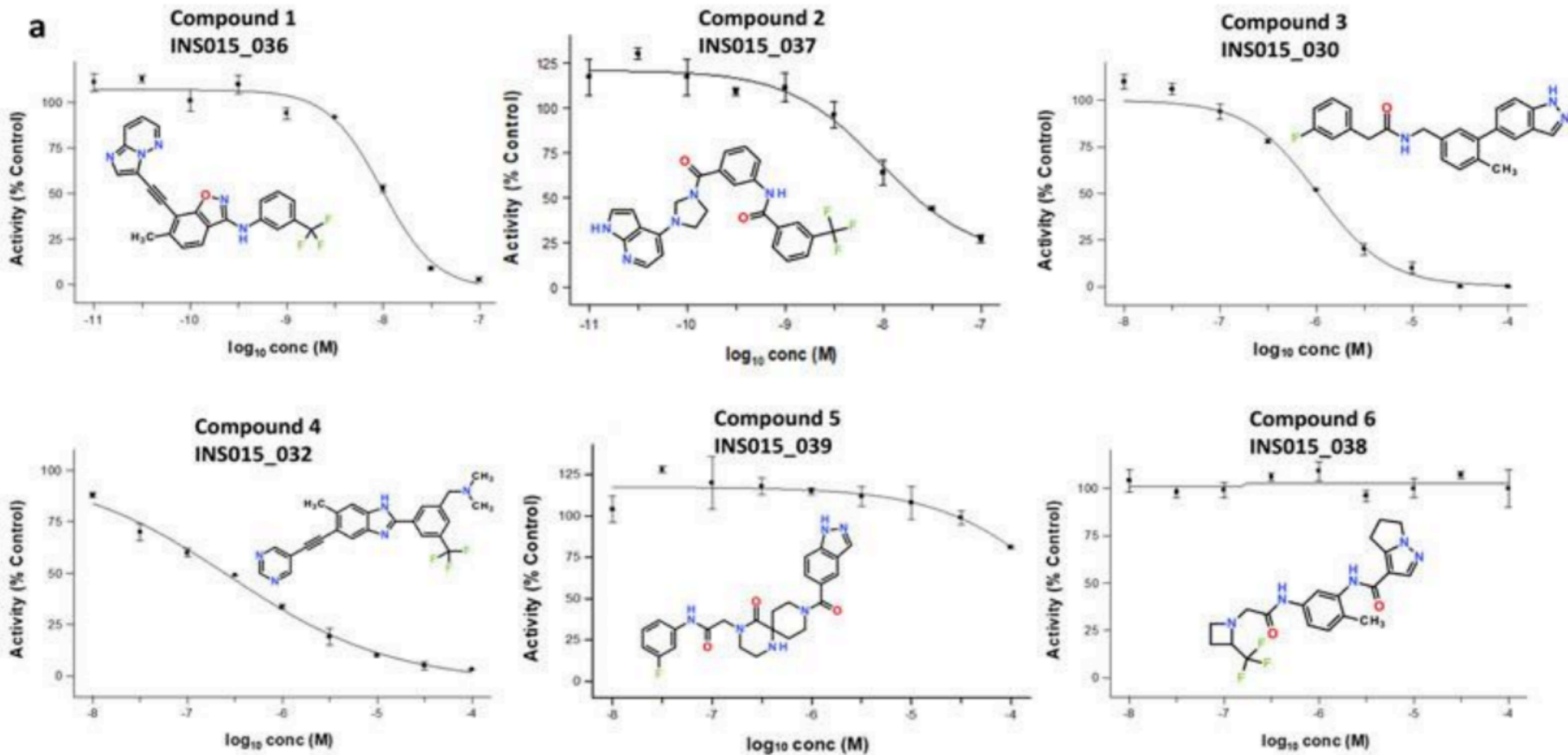
1. Sample 30K structures using RL
2. Filter to 808 compounds using structural alerts and diversity sorting
3. Narrow to 40 structures by sampling
4. Choose 6 based on synthetic accessibility

Sample 40 structures that smoothly covered the space



Pharmacophore: 3D structure-based feature.

in vitro inhibitory activity in an enzymatic kinase assay



Smaller y-axis means more effective

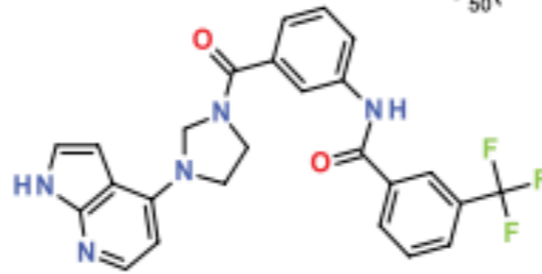
in vitro inhibitory activity in an enzymatic kinase assay

c

1 $IC_{50}(DDR1) = 10 \text{ nM}$
 $IC_{50}(DDR2) = 234 \text{ nM}$

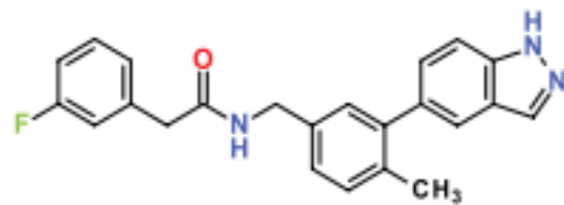


2 $IC_{50}(DDR1) = 21 \text{ nM}$
 $IC_{50}(DDR2) = 76 \text{ nM}$

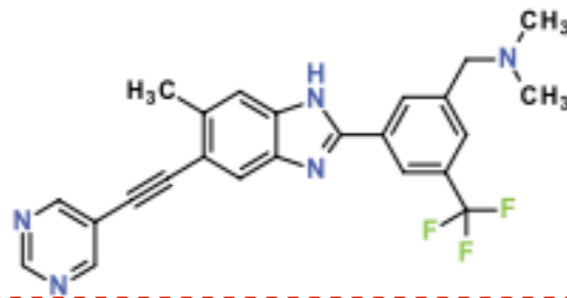


Strong inhibition

3 $IC_{50}(DDR1) = 1,000 \text{ nM}$
 $IC_{50}(DDR2) = 649 \text{ nM}$

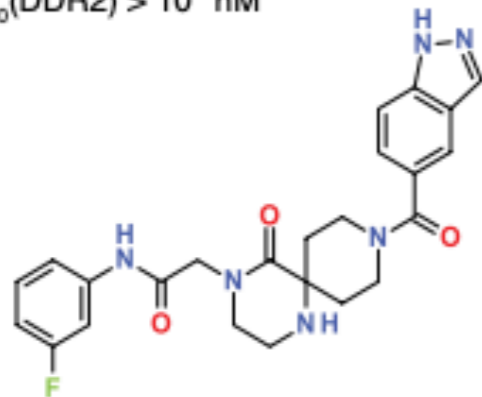


4 $IC_{50}(DDR1) = 278 \text{ nM}$
 $IC_{50}(DDR2) = 162 \text{ nM}$

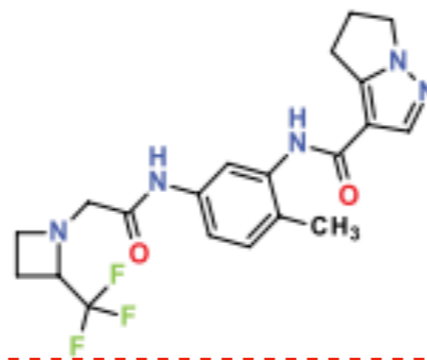


Moderate inhibition

5 $IC_{50}(DDR1) > 10^4 \text{ nM}$
 $IC_{50}(DDR2) > 10^4 \text{ nM}$



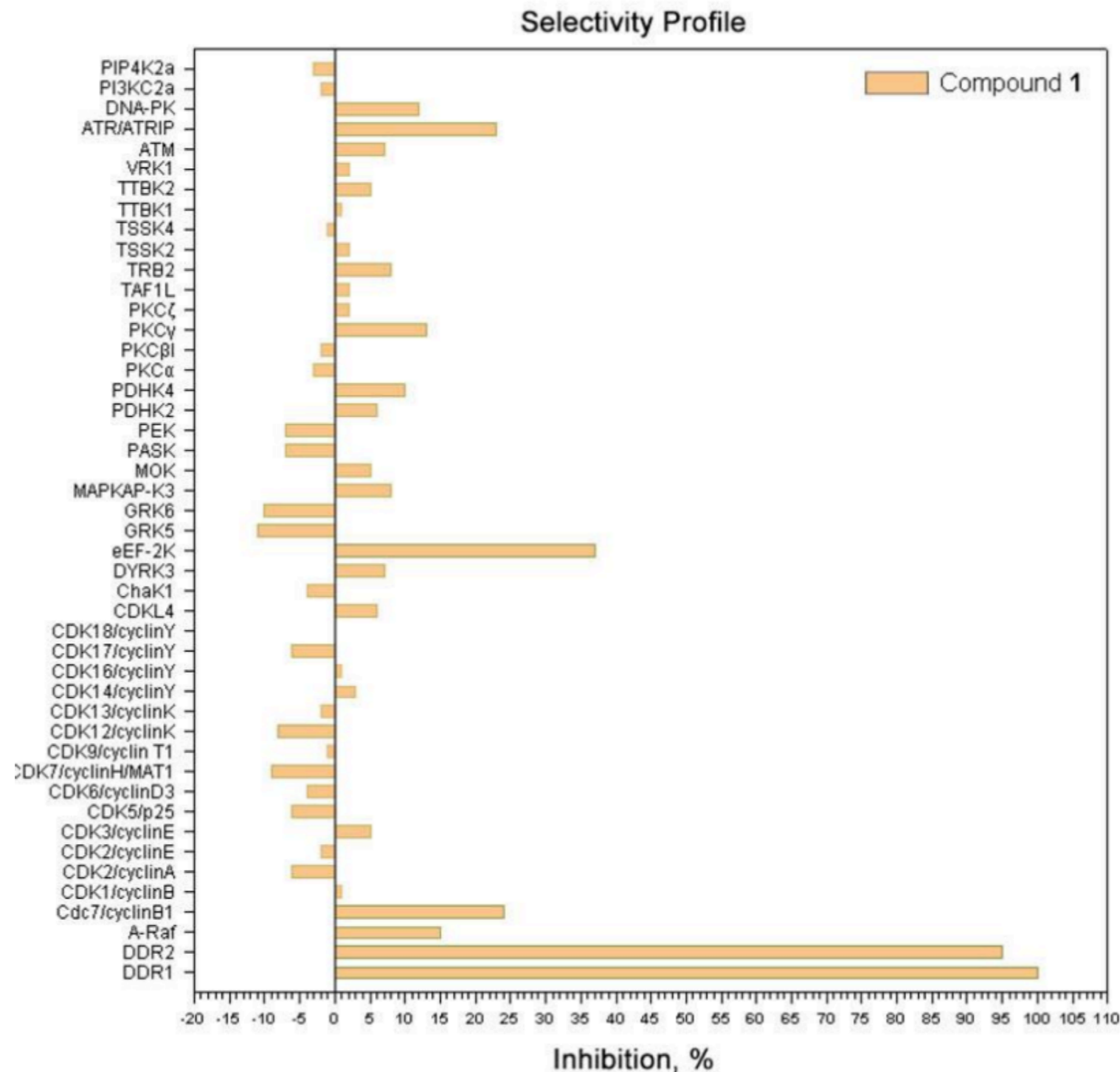
6 $IC_{50}(DDR1) > 10^4 \text{ nM}$
 $IC_{50}(DDR2) > 10^4 \text{ nM}$



No inhibition

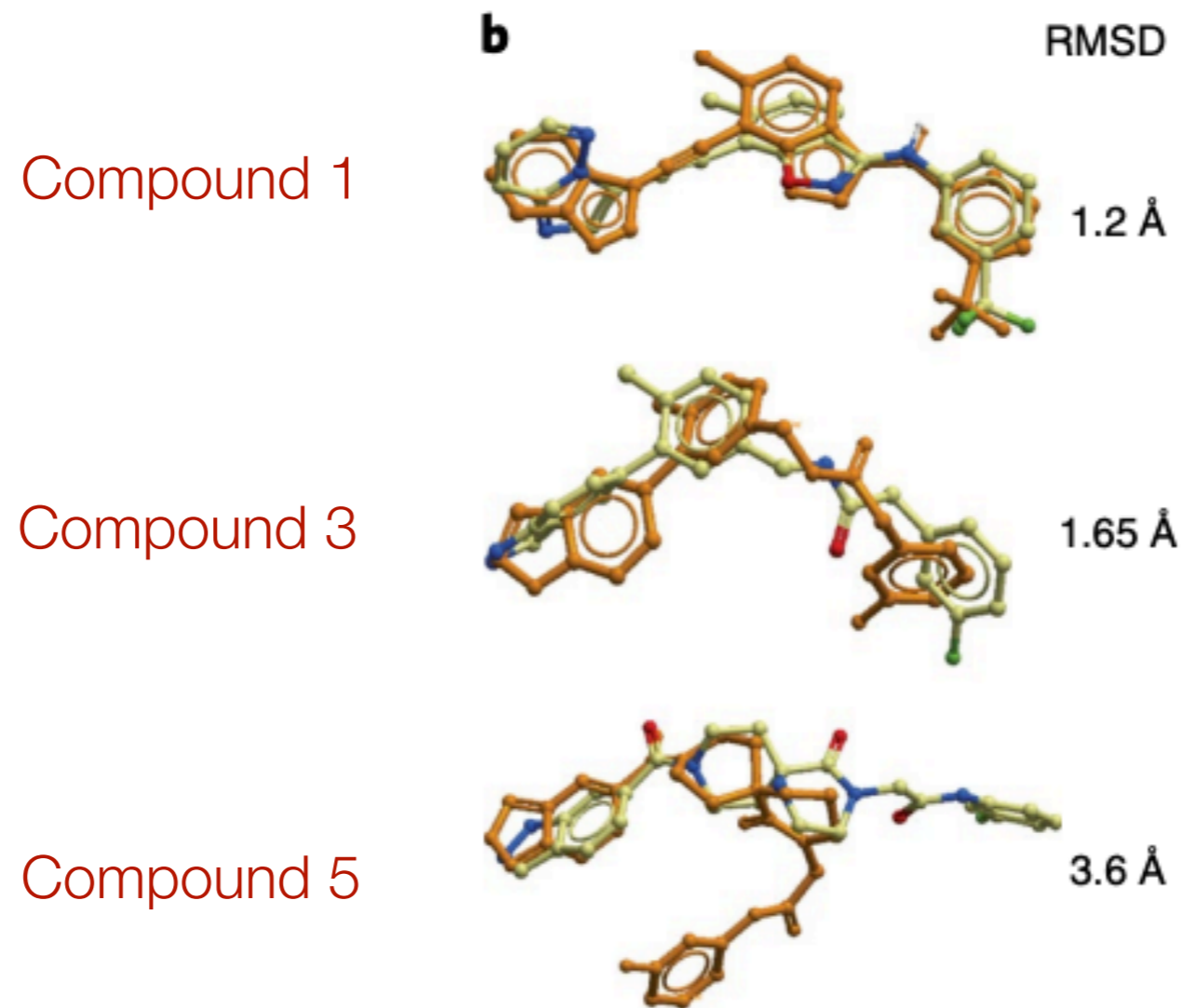
Compound 1 has high selectivity index compared to 44 other non-DDR kinases

Selectivity index = dose that kill non cancer cells / dose that kill cancer cells



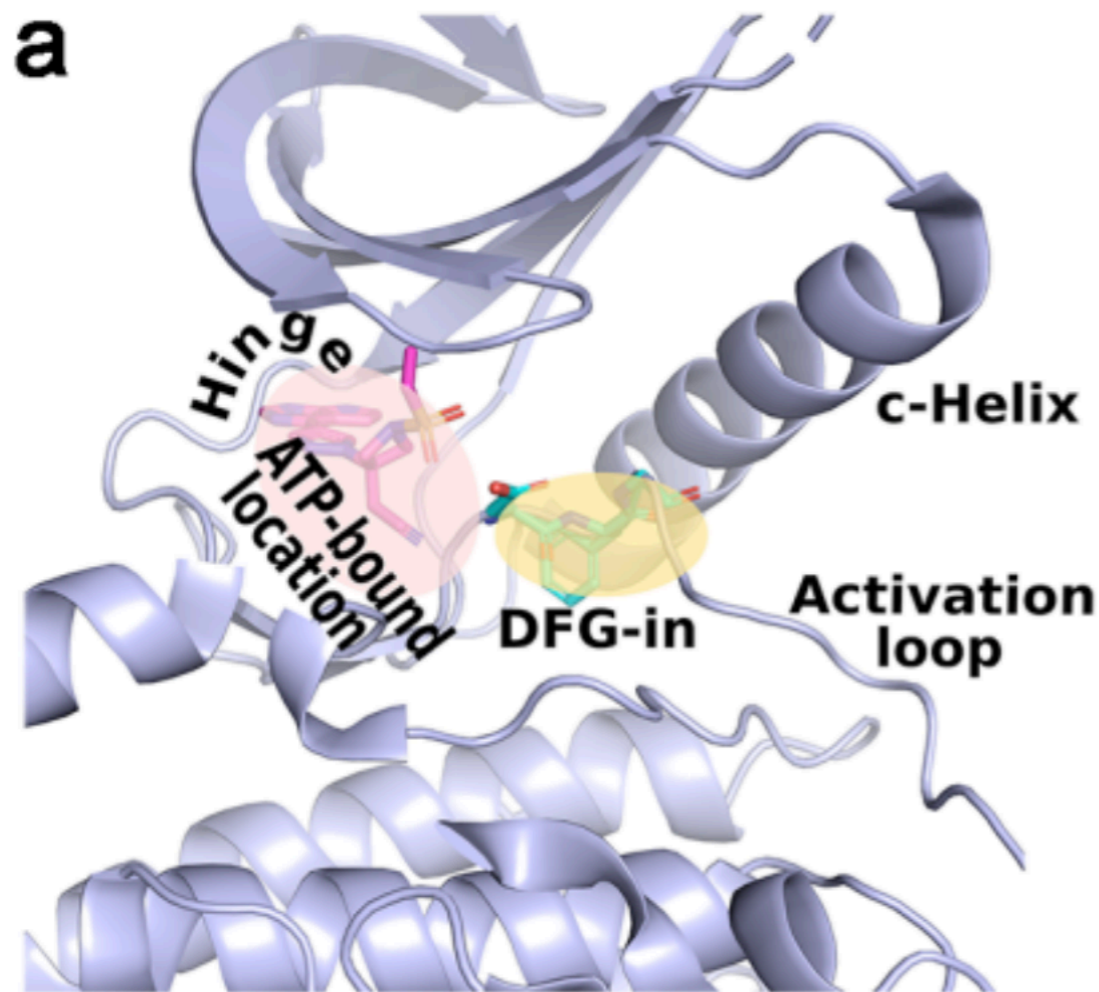
Conclusion: compound 1 does not inhibit other kinases, thus will unlikely cause side effects

Structure-based validation

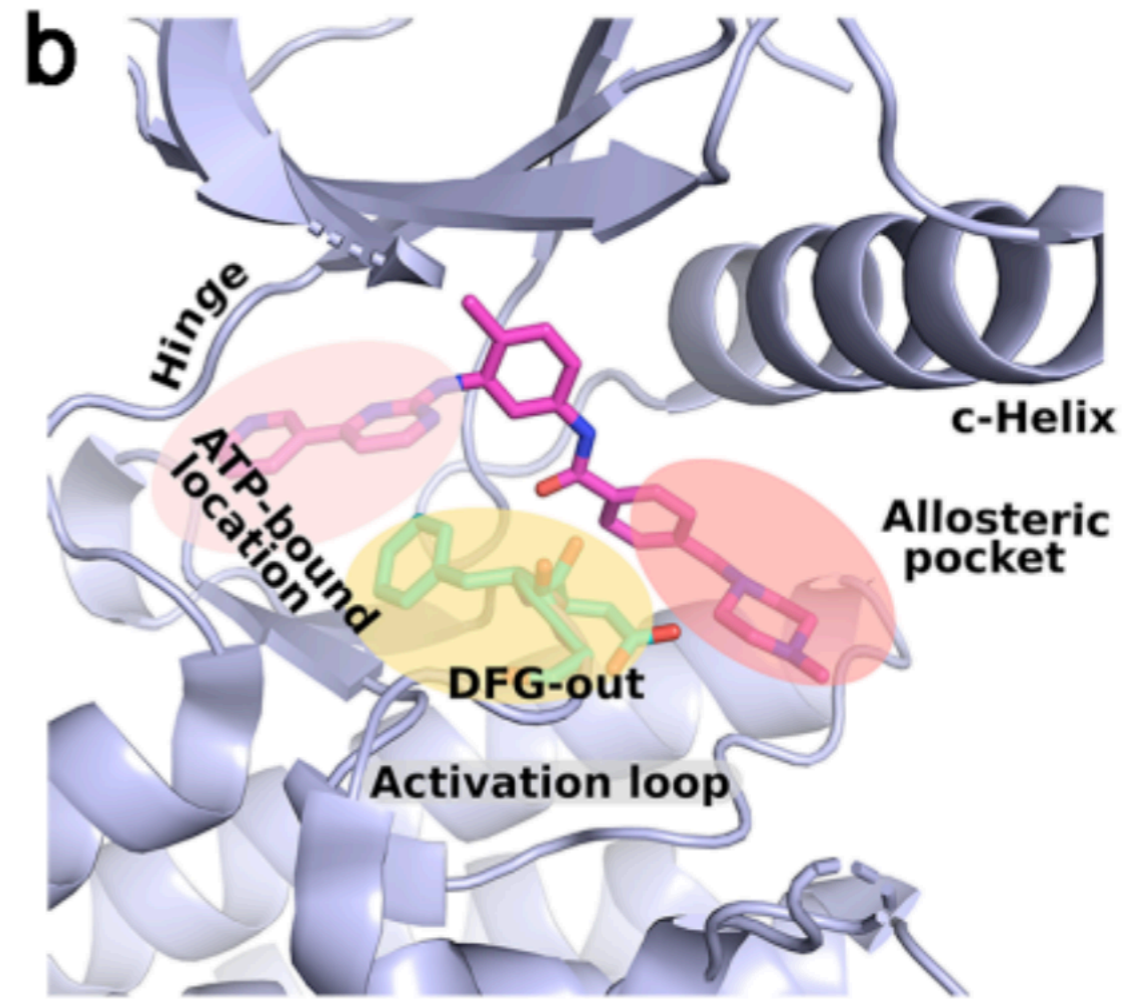


Orange: the structure of this compound predicted by quantum mechanical calculations
Yellow: what a good DDR1 inhibitor looks like based on pharmacophore model.
Smaller RMSD, better alignment

Kinase inhibitor mechanism



Type 1 inhibitor, only occupy ATP-bound location



Type 2 inhibitor, occupy both ATP-bound and allosteric

42 kinase inhibitor current approved by FDA. 4 types in total. 39 are type 1 and type 2 inhibitor

Understand the mechanism

c

Recent progress by Insilico Medicine (Dec, 2021)

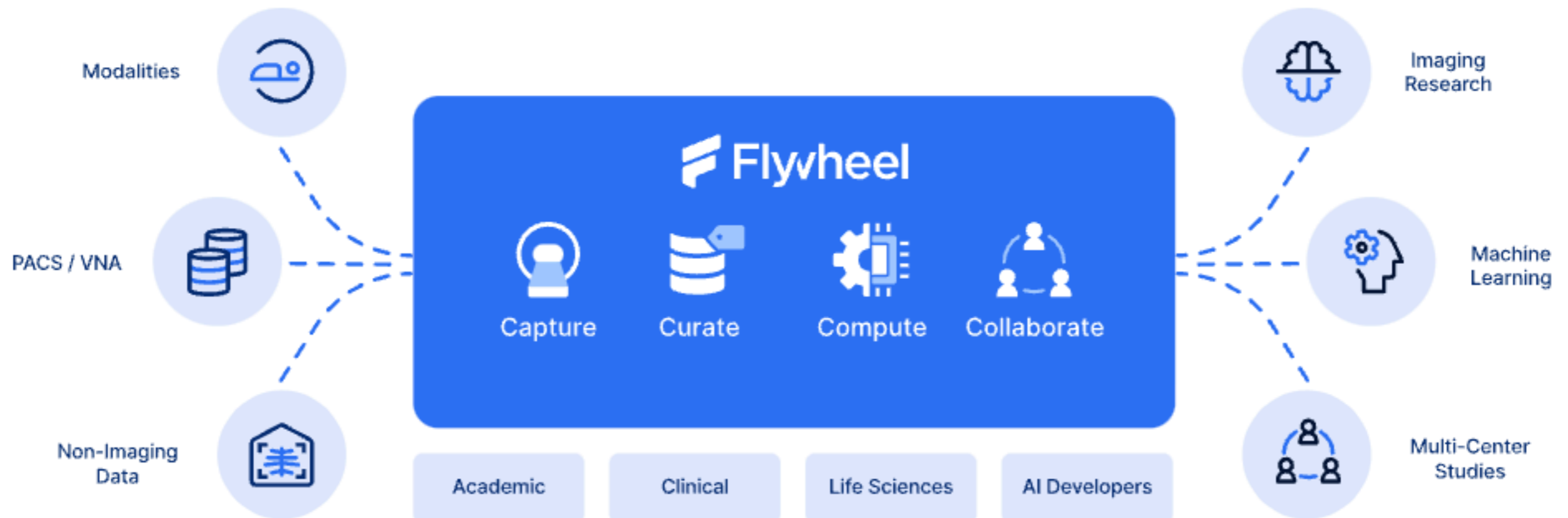
On February 24, 2021 Insilico Medicine [announced the nomination of a preclinical candidate](#) for a novel antifibrotic target discovered using AI, and the novel molecule designed by AI for the first time bridging biology, chemistry, and clinical trial outcome prediction with its [Pharma.AI™](#) platform. Today, we are happy to report that the first healthy volunteers have been dosed in a first-in-human (FIH) microdose trial of ISM001-055 – a potentially first-in-class small molecule inhibitor of a novel biological target developed by Insilico Medicine for the treatment of idiopathic pulmonary fibrosis (IPF), an irreversible progressive orphan disease that affects an increasing number of people globally, with no cure currently available. This small molecule showed promising efficacy for IPF and a good safety profile that led to its nomination as a preclinical drug candidate in December 2020 for IND-enabling studies. With the start of its first clinical program in Australia, Insilico Medicine has moved into a new era in the company's development.

* Not the one introduced in their paper

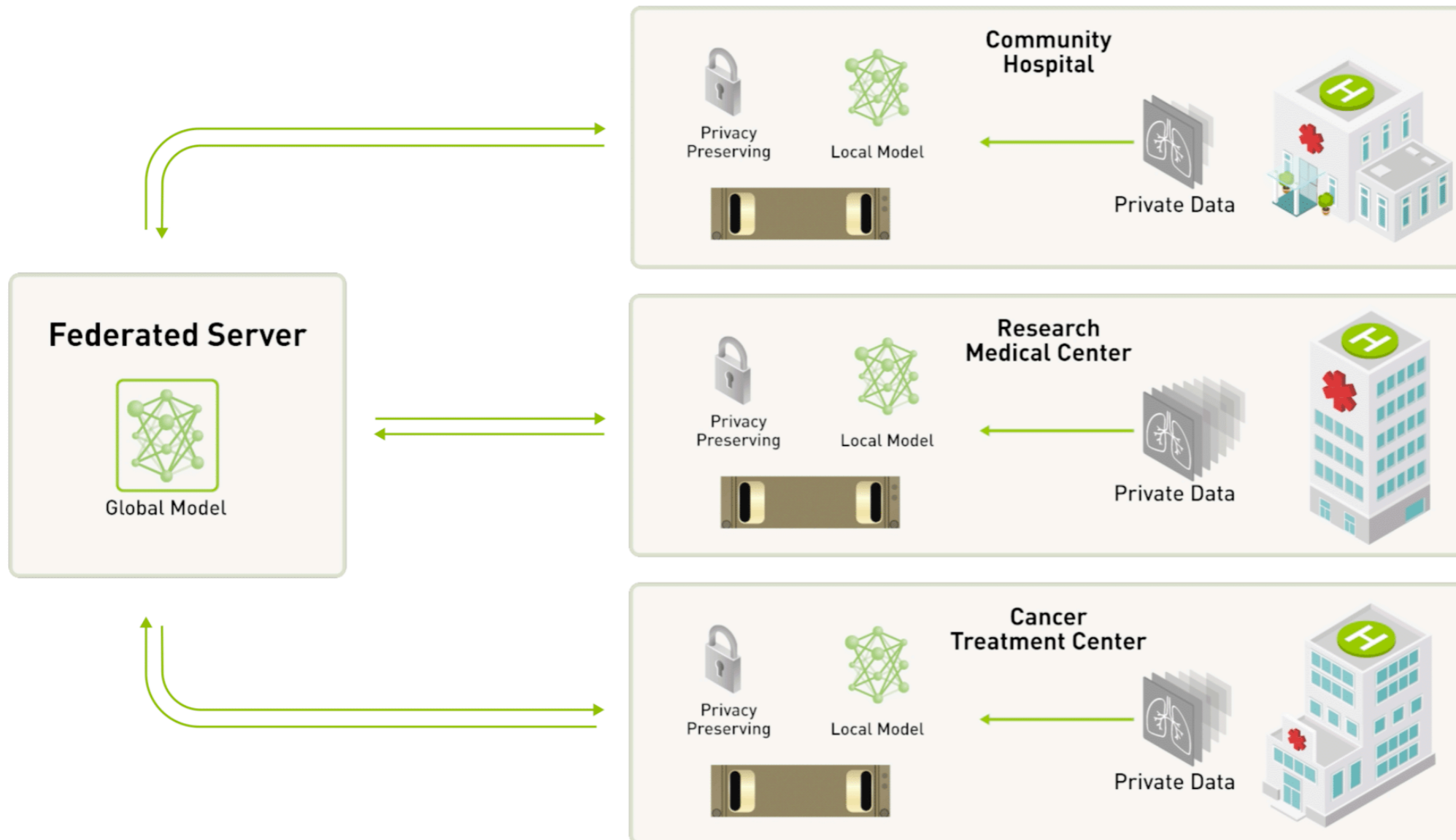
Genentech & Flywheel: Federated learning

Flywheel (AI startup):

- Federated learning for multi-site collaboration
- Automate biomedical research (drug discovery pipeline)
- Cloud-scale informatics platform



Federated learning



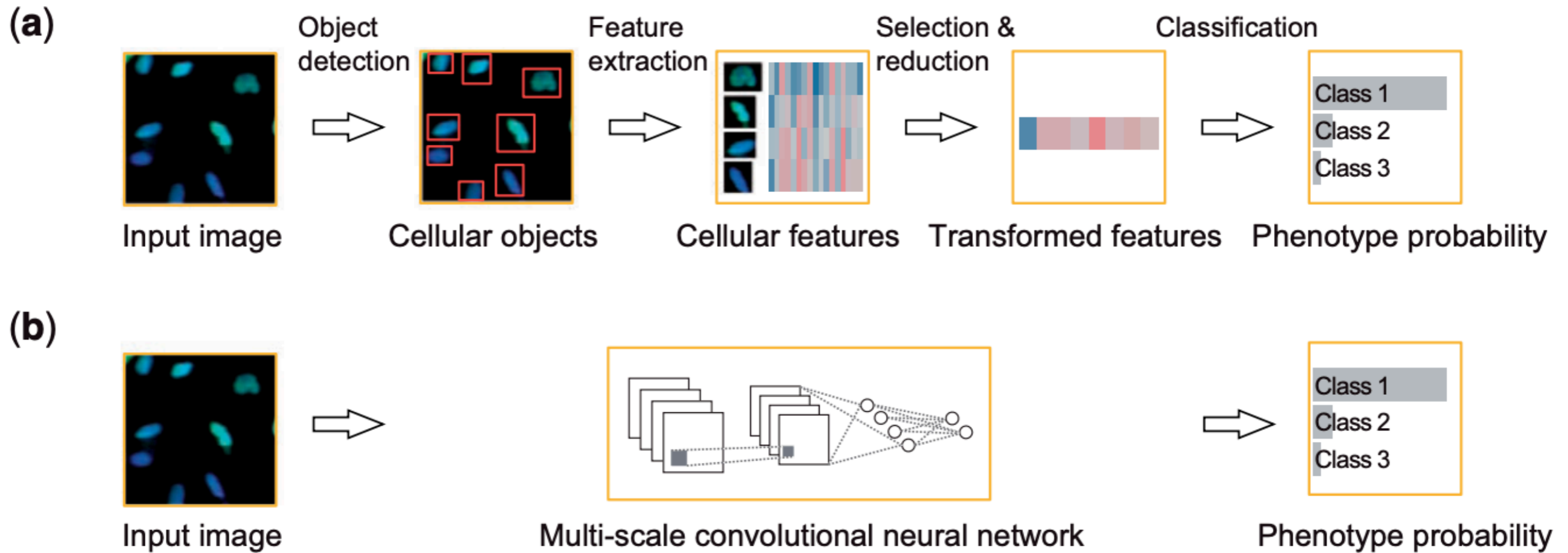
Challenge: Data has to be kept in local due to privacy concern

Federated learning: transfer the model, not the data

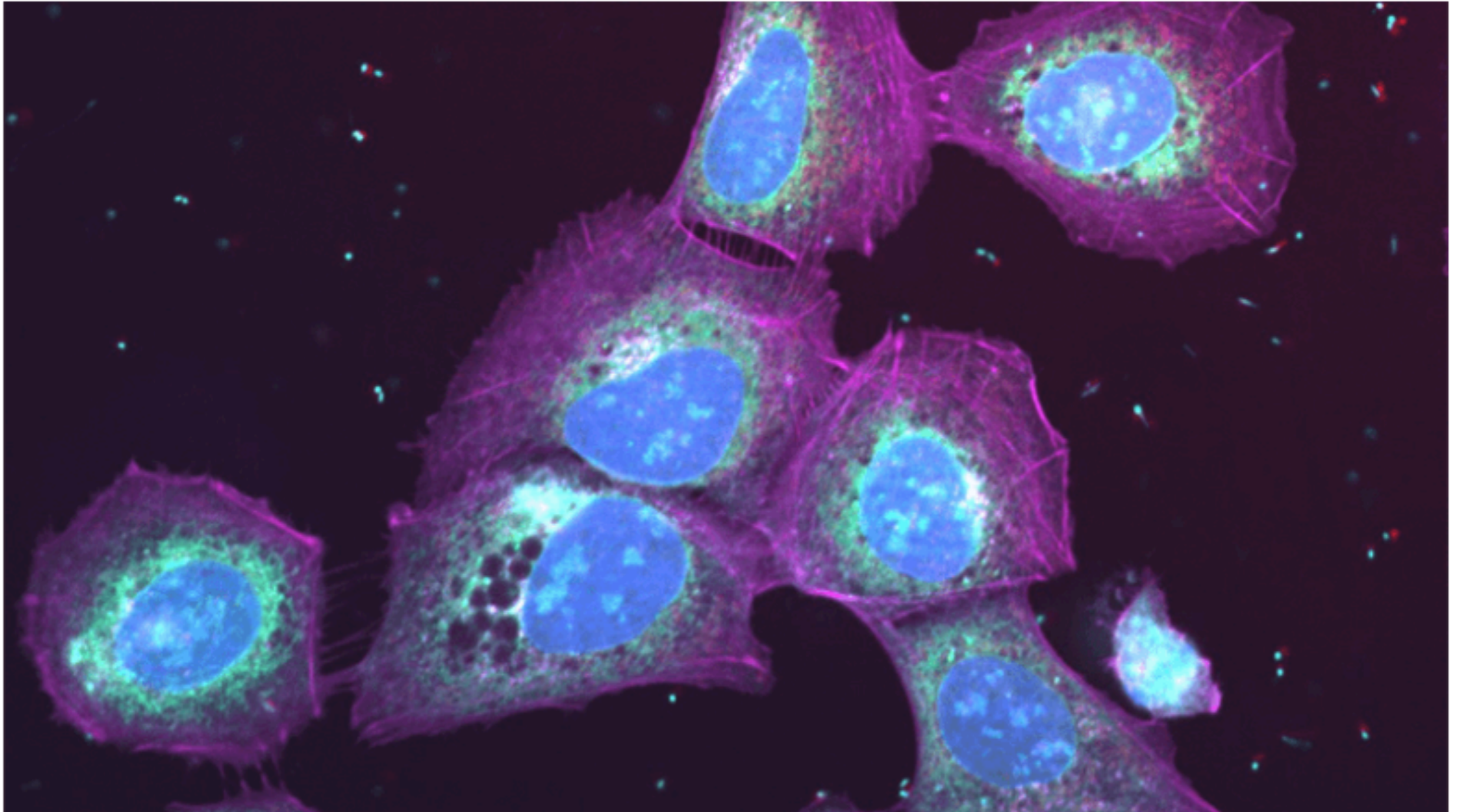
Each local site train their own model

Then sent the model to global server

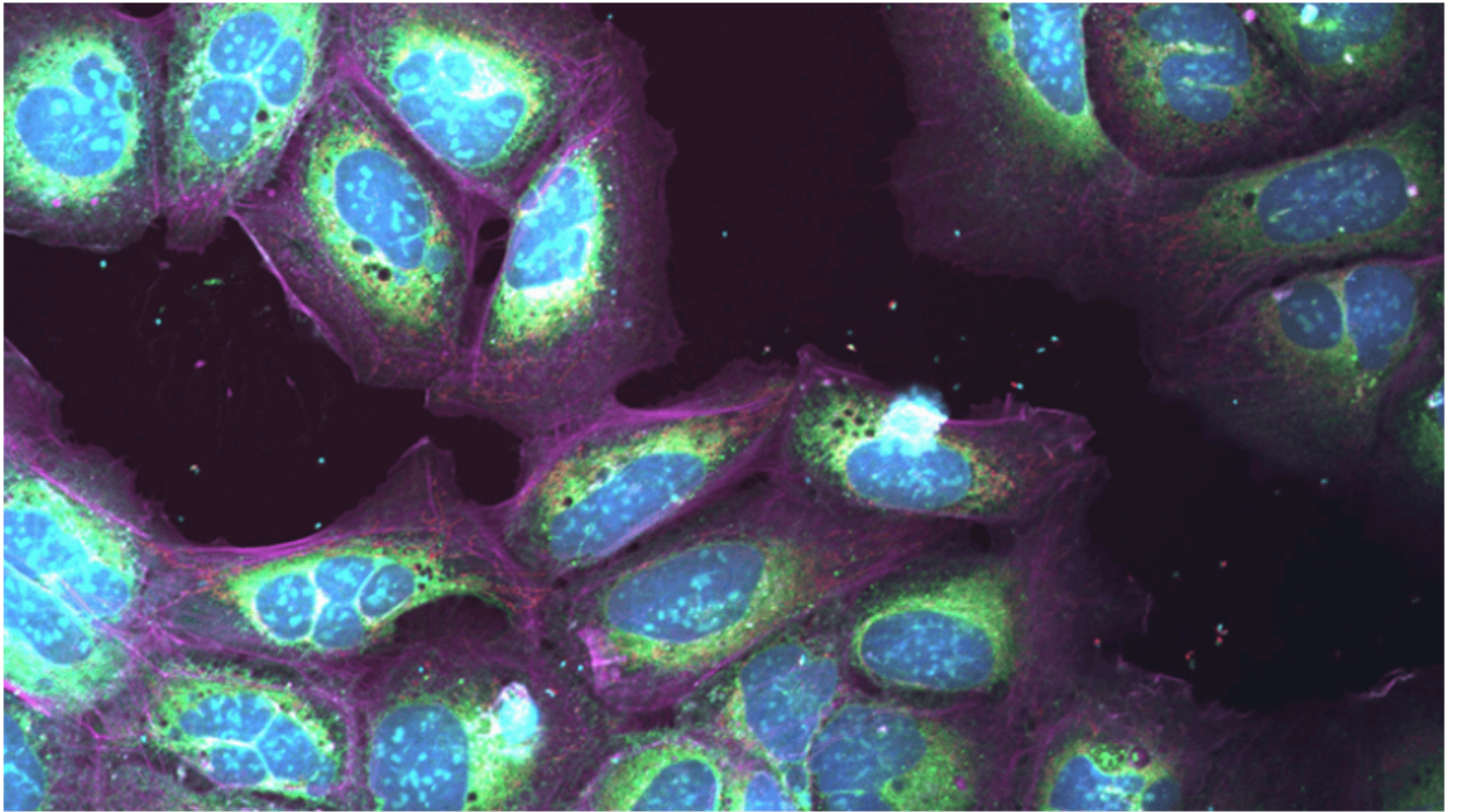
Global sever aggregate the model and sent back to local site



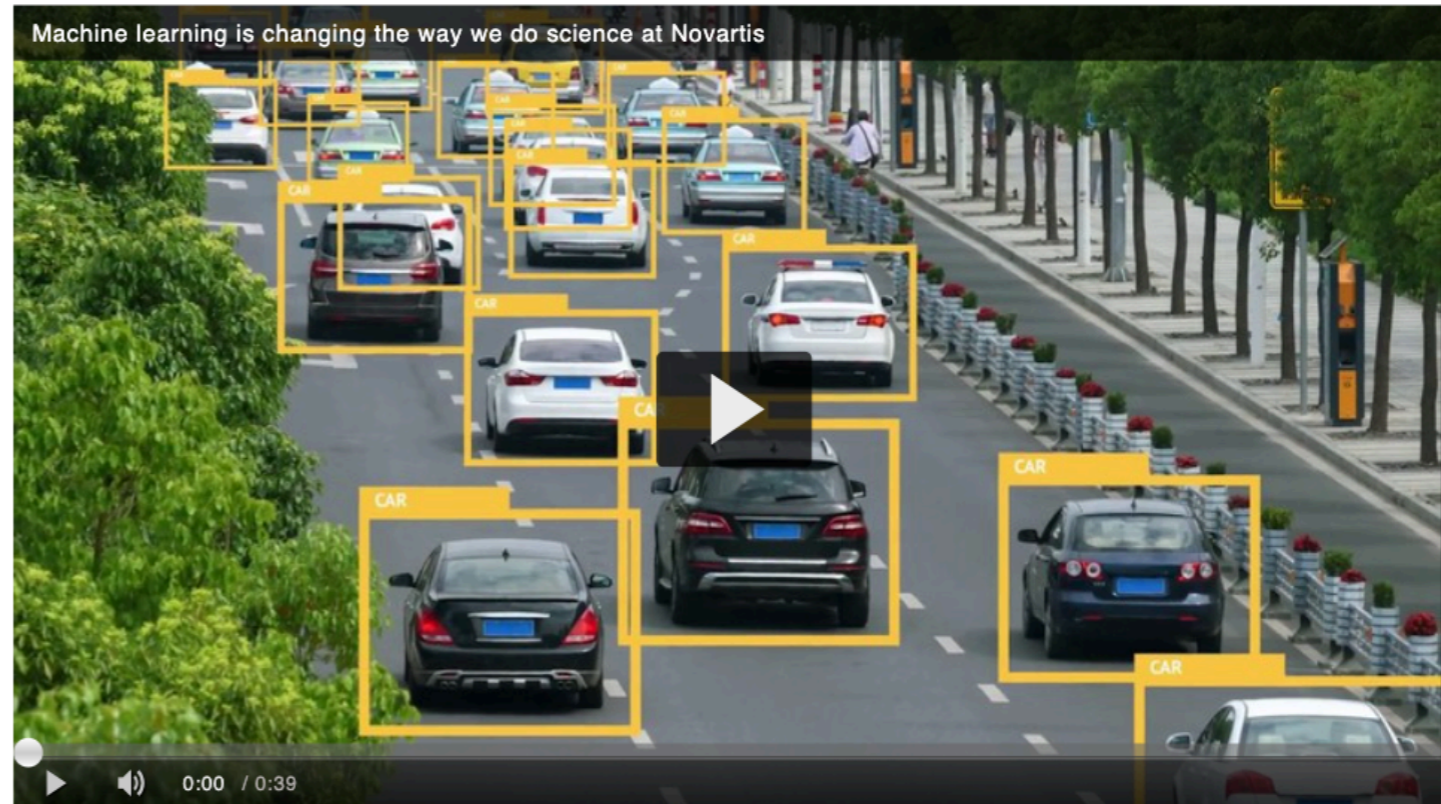
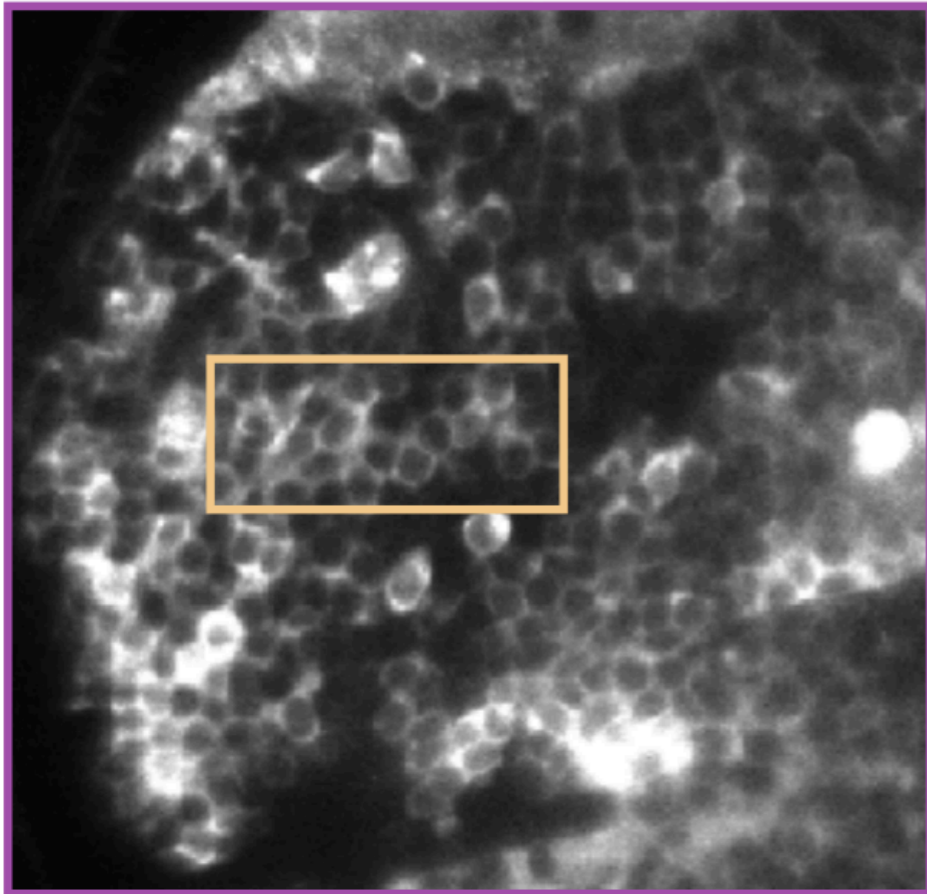
Outlier detection: detect differences after treatment



Outlier detection: detect differences after treatment



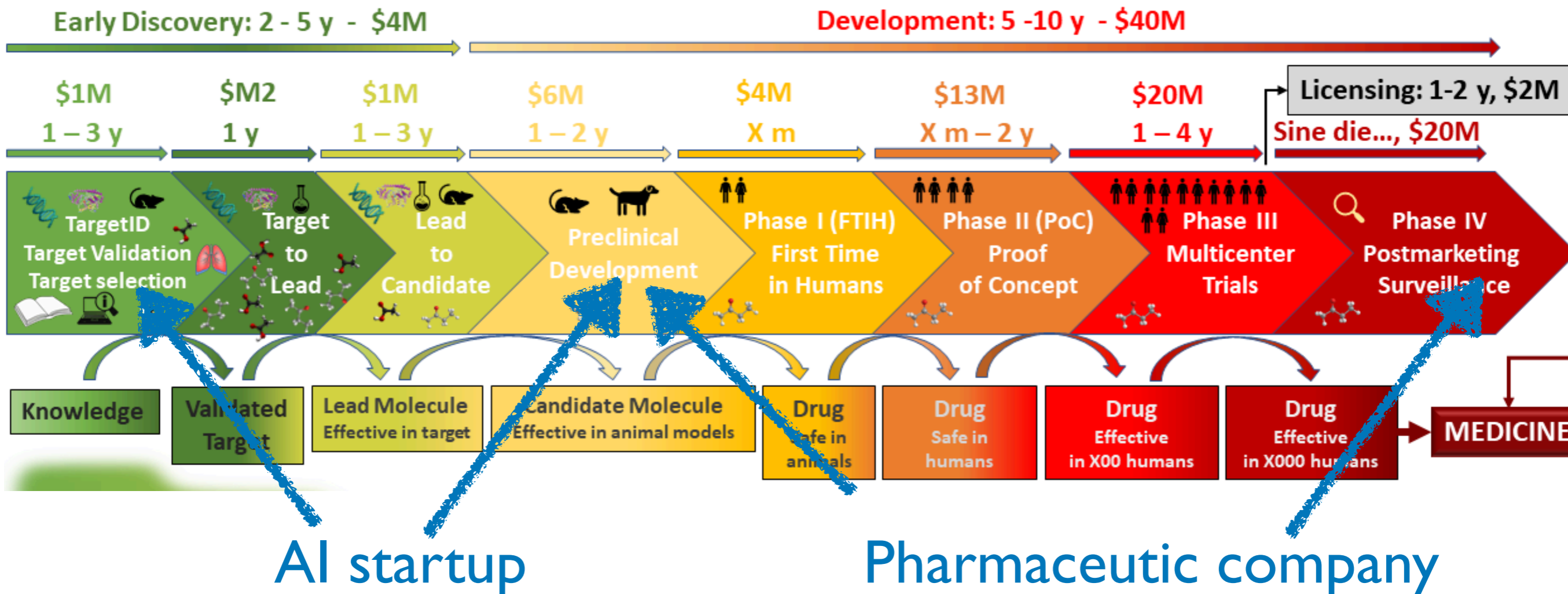
Jointly modeling many cells using cellular video data



Traffic modeling for self-driving cars

Identify long-term cell track and cell division using traffic modeling
Capture cell dynamics, cell-cell interactions

Collaboration between AI startup (Exscientia) and Pharmaceutical company (Bristol Myers Squibb)



The collaboration will use AI to accelerate the discovery of small molecule therapeutic drug candidates in multiple therapeutic areas, including oncology & immunology. The agreement includes up to \$50 million in upfront funding, up to \$125 million in near to mid-term potential milestones, and additional clinical, regulatory and commercial payments that take the potential value of the deal beyond \$1.2 billion.