HARVARD MEDICAL SCHOOL

Kempner INSTITUTE

For the Study of Natural & Artificial Intelligence at Harvard University

BROAD INSTITUTE

AI in Drug Design: Part I
Molecular Drug Discovery

Marinka Zitnik

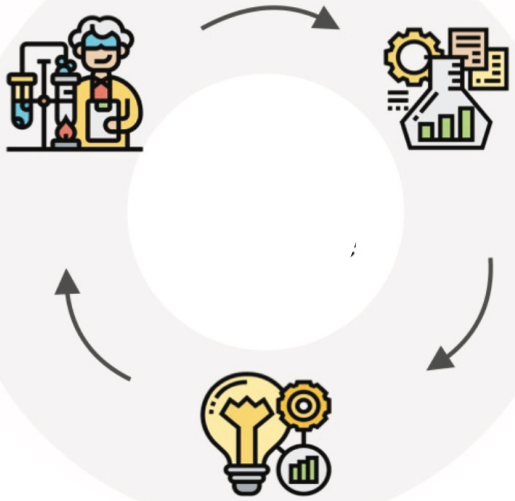
Department of Biomedical Informatics, Harvard Medical School
Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University
Broad Institute of Harvard and MIT
Harvard Data Science

zitniklab.hms.harvard.edu

Biomedical research in the age of AI

Experiments and simulation

Data acquisition and measurements at scale



Accelerated discovery & hypothesis generation

Review

Scientific discovery in the age of artificial intelligence

https://doi.org/10.1038/s41586-023-06221-2
Received: 30 March 2022
Accepted: 16 May 2023
Published online: 2 August 2023
Check for updates

Hanchen Wang[1,17,33,34], Tianfan Fu[2,34], Yuanqi Du[3,34], Wenhao Gao[4], Kexin Huang[5], Ziming Liu[6], Payal Chandak[6], Shengchao Liu[7,32], Peter Van Katwyk[8,32], Andreea Deac[9,32], Anima Anandkumar[4,10], Karianne Bergen[11,12], Carla P. Gomes[3], Shirley Ho[13,14,15,7], Pushmeet Kohli[16], Joan Lasenby[1], Jure Leskovec[5], Tie-Yan Liu[19], Arjun Manrai[20], Debora Marks[21,22], Bharath Ramsundar[23], Le Song[24,25], Jimeng Sun[26], Jian Tang[9,27,28], Petar Veličković[16,29], Max Welling[30,31], Linfeng Zhang[32,33], Connor W. Coley[8,32], Yoshua Bengio[9,32] & Marinka Zitnik[20,21,35,36,37]

Artificial intelligence (AI) is being increasingly integrated into scientific discovery to augment and accelerate research, helping scientists to generate hypotheses, design experiments, collect and interpret large datasets, and gain insights that might not have been possible using traditional scientific methods alone. Here we examine breakthroughs over the past decade that include self-supervised learning, which allows models to be trained on vast amounts of unlabelled data, and geometric deep learning, which leverages knowledge about the structure of scientific data to enhance model accuracy and efficiency. Generative AI methods can create designs, such as small-molecule drugs and proteins, by analysing diverse data modalities, including images and sequences. We discuss how these methods can help scientists throughout the scientific process and the central issues that remain despite such advances. Both developers and users of AI tools need a better understanding of when such approaches need improvement, and challenges posed by poor data quality and stewardship remain. These issues cut across scientific disciplines and require developing foundational algorithmic approaches that can contribute to scientific understanding or acquire it autonomously, making them critical areas of focus for AI innovation.

The foundation for forming scientific insights and theories is laid by how data are collected, transformed and understood. The rise of deep learning in the early 2010s has significantly expanded the scope and ambition of these scientific discovery processes[1]. Artificial intelligence (AI) is increasingly used across scientific disciplines to integrate massive datasets, refine measurements, guide experimentation, explore the space of theories compatible with the data, and provide actionable and reliable models integrated with scientific workflows for autonomous discovery.

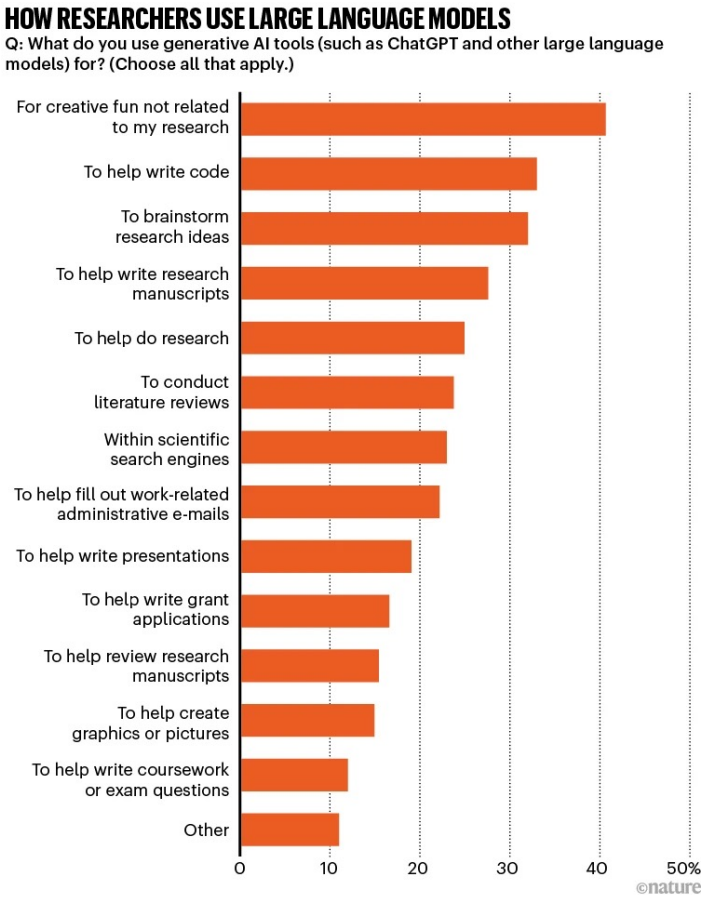
Data collection and analysis are fundamental to scientific understanding and discovery, two of the central aims in science[2], and quantitative methods and emerging technologies, from physical instruments such as microscopes to research techniques such as bootstrapping, have long been used to reach these aims[3]. The introduction of digitization in the 1950s paved the way for the general use of computing in scientific research. The rise of data science since the 2010s has enabled AI to provide valuable guidance by identifying scientifically relevant patterns from large datasets.

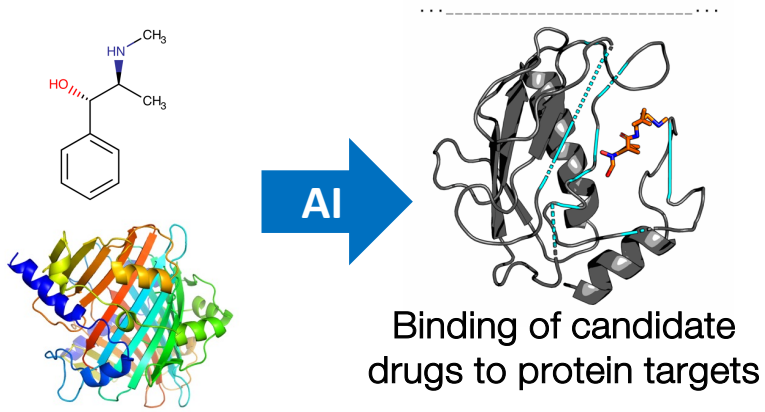
Although scientific practices and procedures vary across stages of scientific research, the development of AI algorithms cuts across traditionally isolated disciplines (Fig. 1). Such algorithms can enhance the design and execution of scientific studies. They are becoming

[1]Department of Engineering, University of Cambridge, Cambridge, UK. [2]Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA, USA. [3]Department of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA, USA. [4]Department of Computer Science, Cornell University, Ithaca, NY, USA. [5]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [6]Department of Computer Science, Stanford University, Stanford, CA, USA. [7]Department of Physics, Massachusetts Institute of Technology, Cambridge, MA, USA. [8]Harvard-MIT Program in Health Sciences and Technology, Cambridge, MA, USA. [9]Mila – Quebec AI Institute, Montreal, Quebec, Canada. [10]Université de Montréal, Montreal, Quebec, Canada. [11]Department of Earth, Environmental and Planetary Sciences, Brown University, Providence, RI, USA. [12]Data Science Institute, Brown University, Providence, RI, USA. [13]NYU04, Santa Clara, CA, USA. [14]Center for Computational Astrophysics, Flatiron Institute, New York, NY, USA. [15]Department of Astrophysical Sciences, Princeton University, Princeton, NJ, USA. [16]Department of Physics, Carnegie Mellon University, Pittsburgh, PA, USA. [17]Department of Physics and Center for Data Science, New York University, New York, NY, USA. [18]Google DeepMind, London, UK. [19]Microsoft Research, Beijing, China. [20]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [21]Department of Systems Biology, Harvard Medical School, Boston, MA, USA. [22]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [23]Deep Forest Sciences, Palo Alto, CA, USA. [24]BioMap, Beijing, China. [25]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates. [26]University of Illinois at Urbana-Champaign, Champaign, IL, USA. [27]HEC Montréal, Montreal, Quebec, Canada. [28]CIFAR AI Chair, Toronto, Ontario, Canada. [29]Department of Computer Science and Technology, University of Cambridge, Cambridge, UK. [30]University of Amsterdam, Amsterdam, Netherlands. [31]Microsoft Research Amsterdam, Amsterdam, Netherlands. [32]DP Technology, Beijing, China. [33]AI for Science Institute, Beijing, China. [34]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA. [35]Harvard Data Science Initiative, Cambridge, MA, USA. [36]Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, Cambridge, MA, USA. [37]Present address: Department of Research and Early Development, Genentech Inc, South San Francisco, CA, USA. [38]Present address: Department of Computer Science, Stanford University, Stanford, CA, USA. [39]These authors contributed equally: Hanchen Wang, Tianfan Fu, Yuanqi Du. ✉e-mail: marinka@hms.harvard.edu

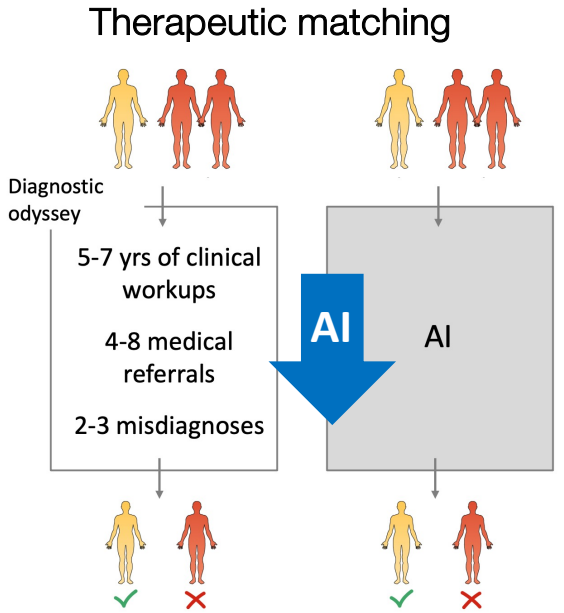
Scientific discovery in the age of AI, *Nature* 2023

Biomedical research in the age of AI

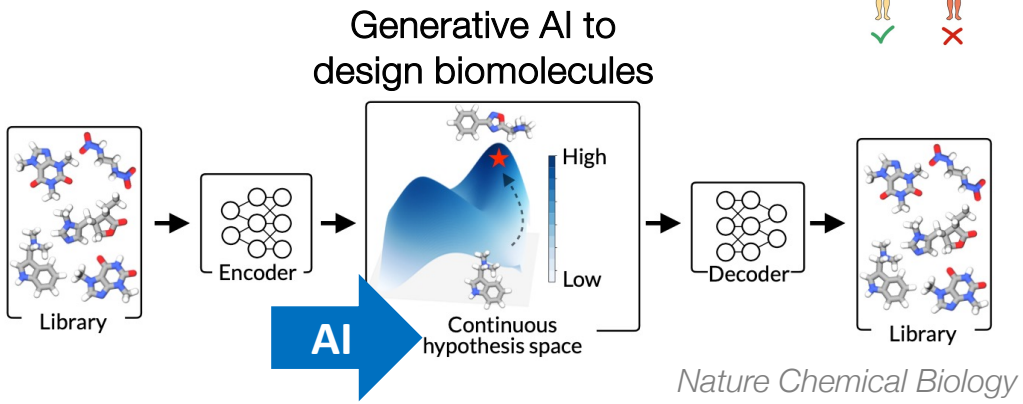




NeurIPS 2023; NeurIPS 2022; Nature Methods 2024



medRxiv 2023



Nature Chemical Biology 2022

Generative AI is changing the way science is done

AI is used to augment research, providing insights that might not have been possible using traditional methods alone

Scientific discovery in the age of AI, *Nature* 2023

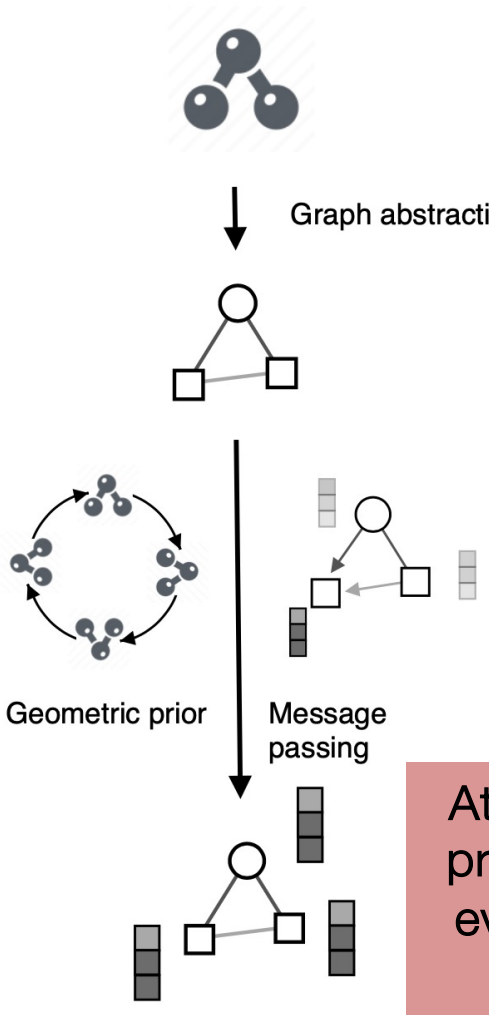
Potential for AI in drug design

Treatment	Organization	Description	Phase	Lead indication
REC-2282	Recursion	Small molecule pan-HDAC inhibitor	2/3	Neurofibromatosis type 2
REC-994	Recursion	Small molecule superoxide scavenger	2	Cerebral cavernous malformation
REC-4881	Recursion	Small molecule inhibitor of MEK1 and MEK2	2	Familial adenomatous polyposis
INS018_055	InSilico Medicine	Small molecule inhibitor	2	Idiopathic pulmonary fibrosis
BEN-2293	BenevolentAI	Topical pan-tyrosine kinase inhibitor	2a	Atopic dermatitis
EXS-21546	Exscientia and Evotec	A_{2A} receptor antagonist	1b/2	Solid tumors carrying high adenosine signatures.
RLY-4008	Relay Therapeutics	Inhibitor of FGFR2	1/2	FGFR2-altered cholangiocarcinoma
EXS-4318	Exscientia	PKC-θ inhibitor	1/2	Inflammatory and autoimmune conditions
BEN-8744	BenevolentAI	Small molecule PDE10 inhibitor	1	Ulcerative colitis
Undisclosed	Recursion	Small molecular inhibitor of RBM39, a CDK12-associated protein	Pre-clinical	HRD-negative ovarian cancer

Inside the nascent industry of AI-designed drugs, Nature Medicine 2023

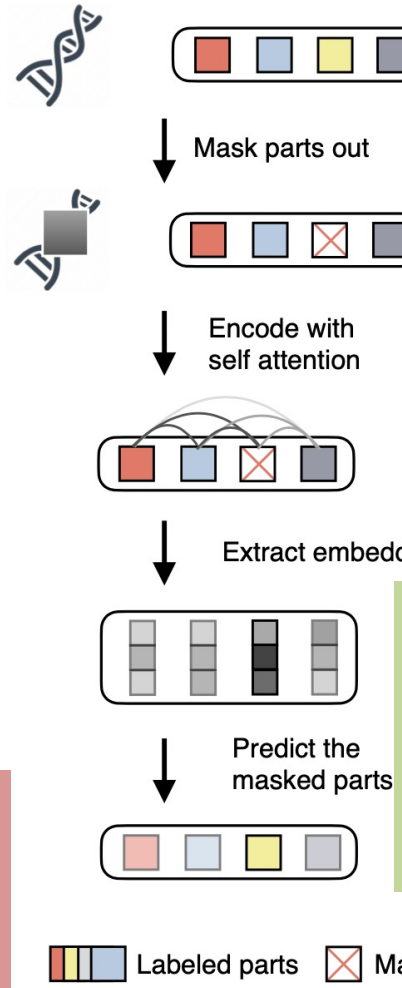
AI in drug design: Key innovations

Geometric learning



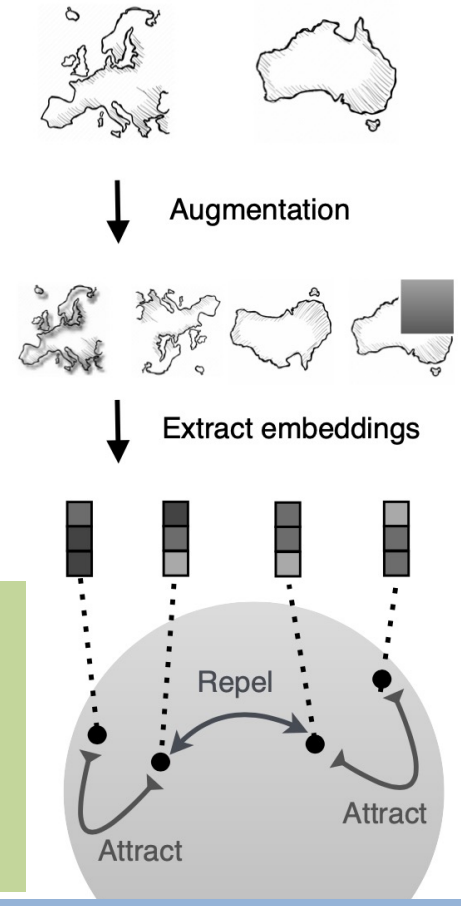
Atom-level molecular prediction, real-world evidence knowledge graphs

Self-supervised learning



Target discovery, genotype-phenotype modeling, DNA, RNA, AA sequence modeling

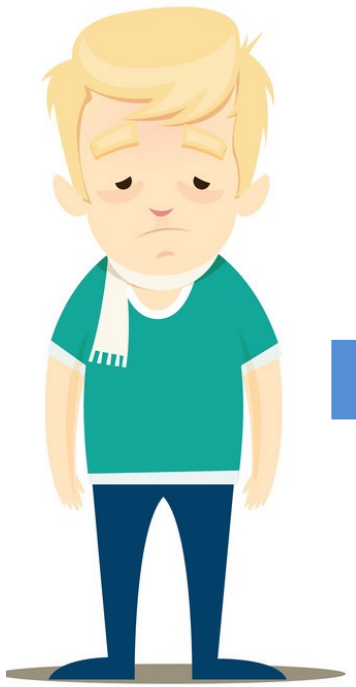
Generative AI



chatbots, copilots, molecular and drug design

Scientific discovery in the age of AI, *Nature* 2023

Our vision: Lay the foundations for AI to enhance the understanding of medicine and drug design, eventually enabling AI to learn on its own



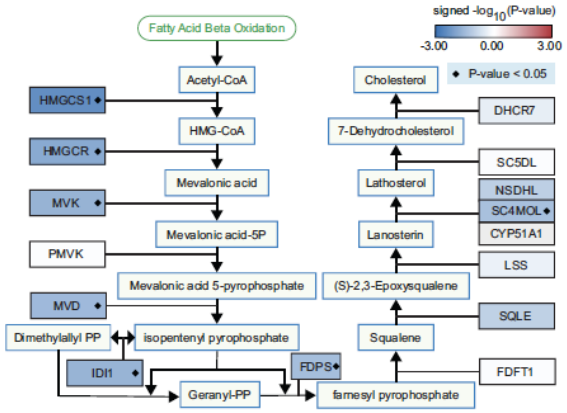
Predict and optimize protein targets

Generate, and refine therapeutic candidates

Find relevant disease, patient and cell type contexts

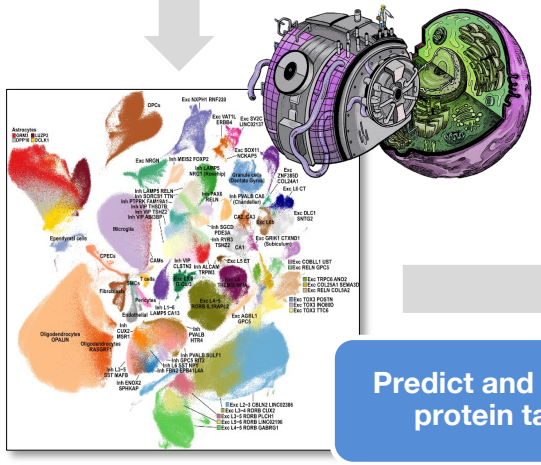
Match therapies to patient benefits

Disease circuitry across diseases and individuals



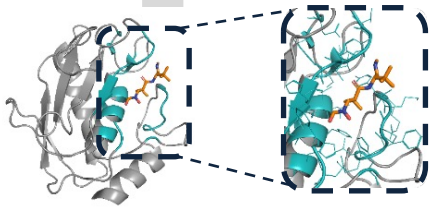
Key disease mechanisms, shared effects, interaction effects

Multimodal representation learning models
PINNACLE, PDGrapher



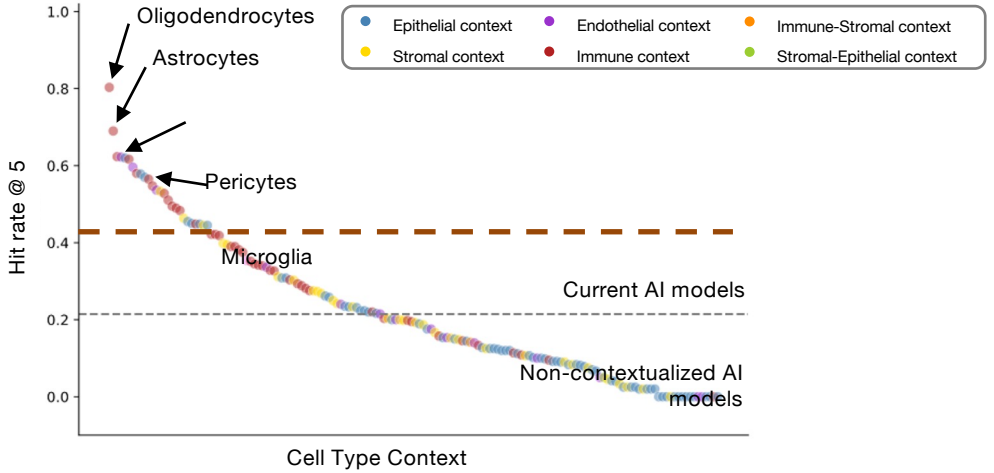
Predict and optimize protein targets

Generative geometric deep learning models
PocketGen, FAIR, and TxPLM

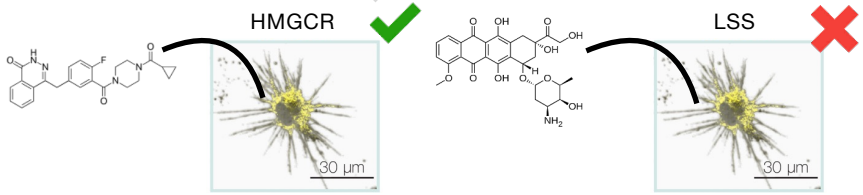


Full-atom structure

Generate, and refine therapeutic candidates



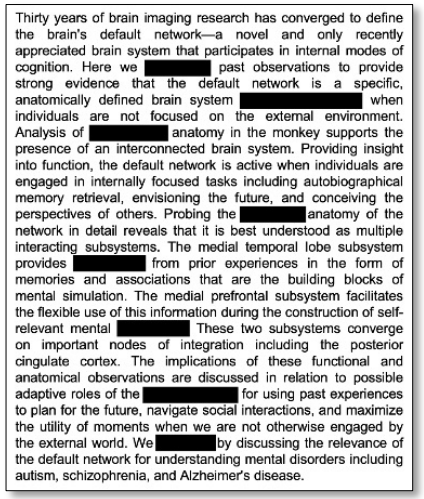
Rank-ordered lists of molecules to modulate disease circuitry in a cell-type specific manner



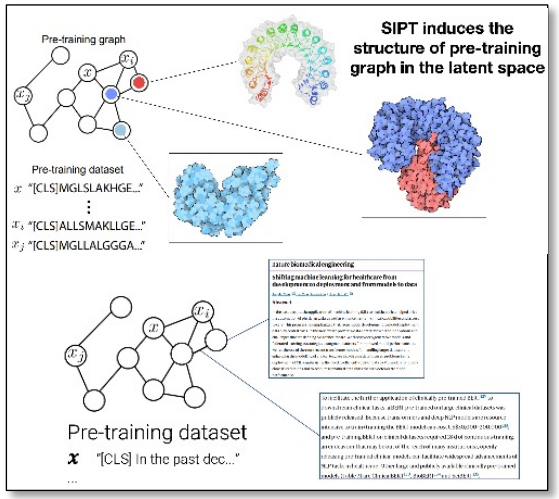
Chemistry optimization, synthesis, automation

Methods: Geometric deep learning, LLMs + graphs, transfer learning

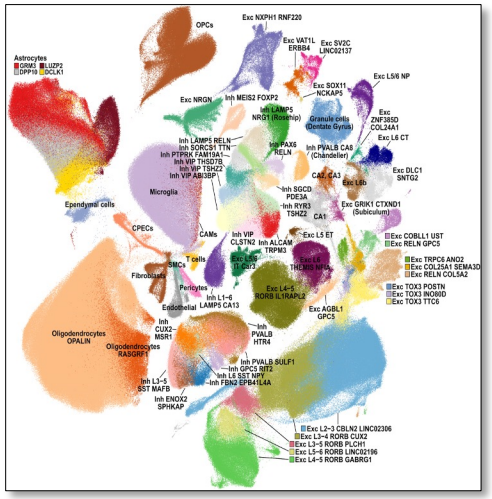
LLMs


Thirty years of brain imaging research has converged to define the brain's default network—a novel and only recently appreciated brain system that participates in internal modes of cognition. Here we [redacted] past observations to provide strong evidence that the default network is a specific, anatomically defined brain system [redacted] when individuals are not focused on the external environment. Analysis of [redacted] anatomy in the monkey supports the presence of an interconnected brain system. Providing insight into function, the default network is active when individuals are engaged in internally focused tasks including autobiographical memory retrieval, envisioning the future, and conceiving the perspectives of others. Probing the [redacted] anatomy of the network in detail reveals that it is best understood as multiple interacting subsystems. The medial temporal lobe subsystem provides [redacted] from prior experiences in the form of memories and associations that are the building blocks of mental simulation. The medial prefrontal subsystem facilitates the flexible use of this information during the construction of self-relevant mental [redacted] These two subsystems converge on important nodes of integration including the posterior cingulate cortex. The implications of these functional and anatomical observations are discussed in relation to possible adaptive roles of the [redacted] for using past experiences to plan for the future, navigate social interactions, and maximize the utility of moments when we are not otherwise engaged by the external world. We [redacted] by discussing the relevance of the default network for understanding mental disorders including autism, schizophrenia, and Alzheimer's disease.

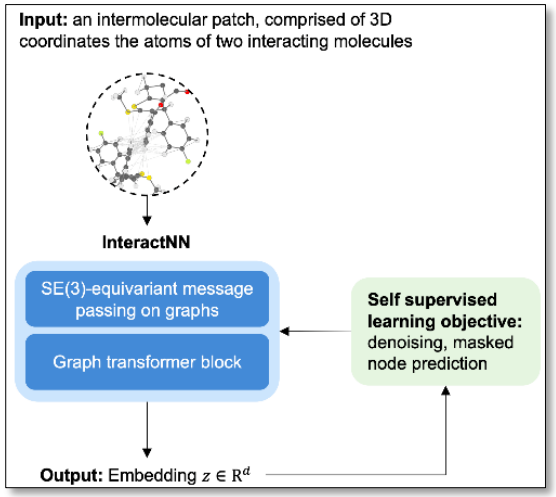
LLMs + Graphs



Transfer learning

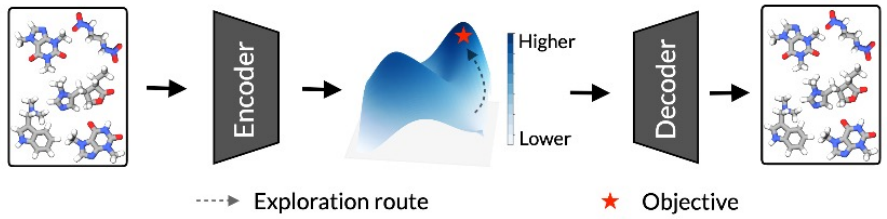


Geometric deep learning



- **Multimodal LLMs** to leverage multimodal biological sequences and scientific knowledge
- **Knowledge-graph based models** that train self-supervised models on broad data at scale without pre-defined labels
- **Geometric and generative AI models** that create action plans for experiments and produce new designs such as small molecule drugs and proteins from experimental data

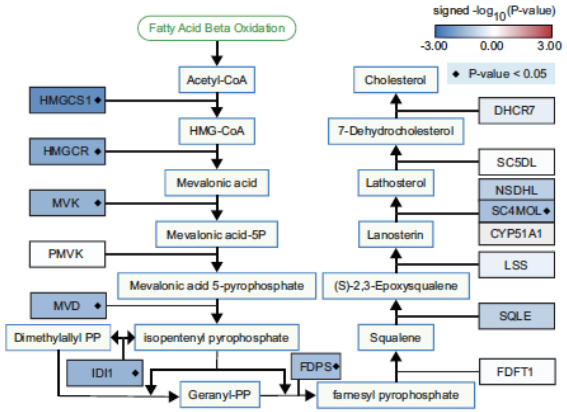
Generative AI: Better drug design, molecule optimization, guiding high-throughput perturbation & interaction screening



Scientific discovery in the age of AI, *Nature* 2023

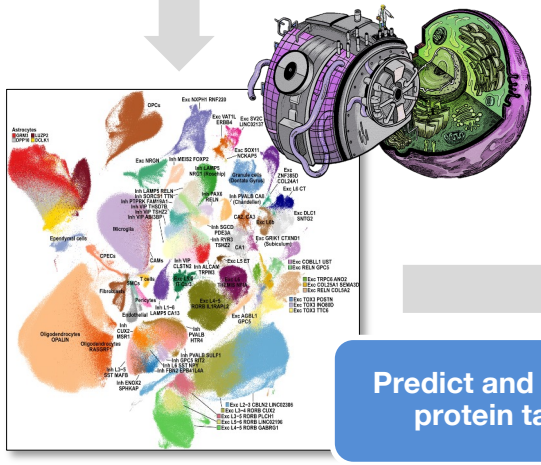
Li et al., *Nature Methods* 2024; Extefaie et al., *Nature Mach Intel* 2023; McDermott et al., *Nature Mach Intel* 2023; Li et al., *Nature Biomed Eng* 2022; Zhang et al., ICLR 2022; Zhang et al. *NeurIPS* 2022; Agarwal et al., *NeurIPS* 2022; Huang et al., *NeurIPS* 2021; Alsentzer et al., *NeurIPS* 2020; Zhang et al., *NeurIPS* 2020; Huang et al., *NeurIPS* 2020; Queen et al., *NeurIPS 2023;* He et al.; *ICML* 2023; Jiali et al.; *ICLR* 2023; Zhang et al., *NeurIPS* 2023; Scott et al., *Nature Mach Intell* 2023; Sanders et al., *Nature Mach Intell* 2023; Agarwal et al., *AISTATS* 2022; Zhong et al., *CVPR 2024*

Disease circuitry across diseases and individuals



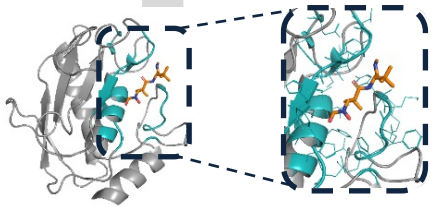
Key disease mechanisms, shared effects, interaction effects

Multimodal representation learning models
PINNACLE, PDGrapher



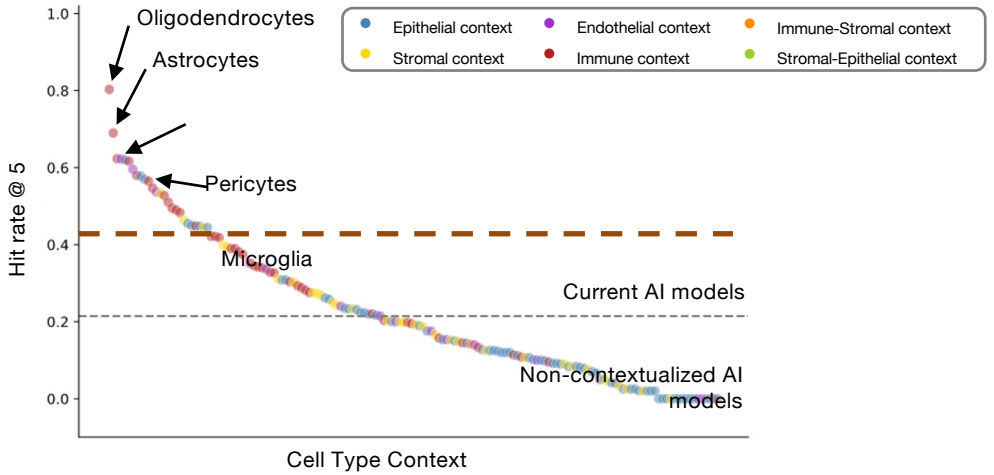
Predict and optimize protein targets

Generative geometric deep learning models
PocketGen, FAIR, and TxPLM

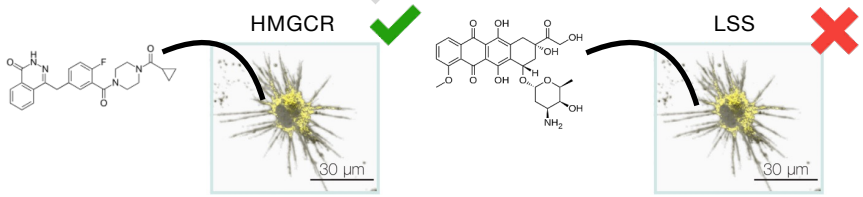


Full-atom structure

Generate, and refine therapeutic candidates



Rank-ordered lists of molecules to modulate disease circuitry in a cell-type specific manner



Chemistry optimization, synthesis, automation

“apple” is a **polysemic** word…

Google

grow an apple

buy an apple

… whose **particular meaning** is resolved via **sentence context**

Google

grow an apple

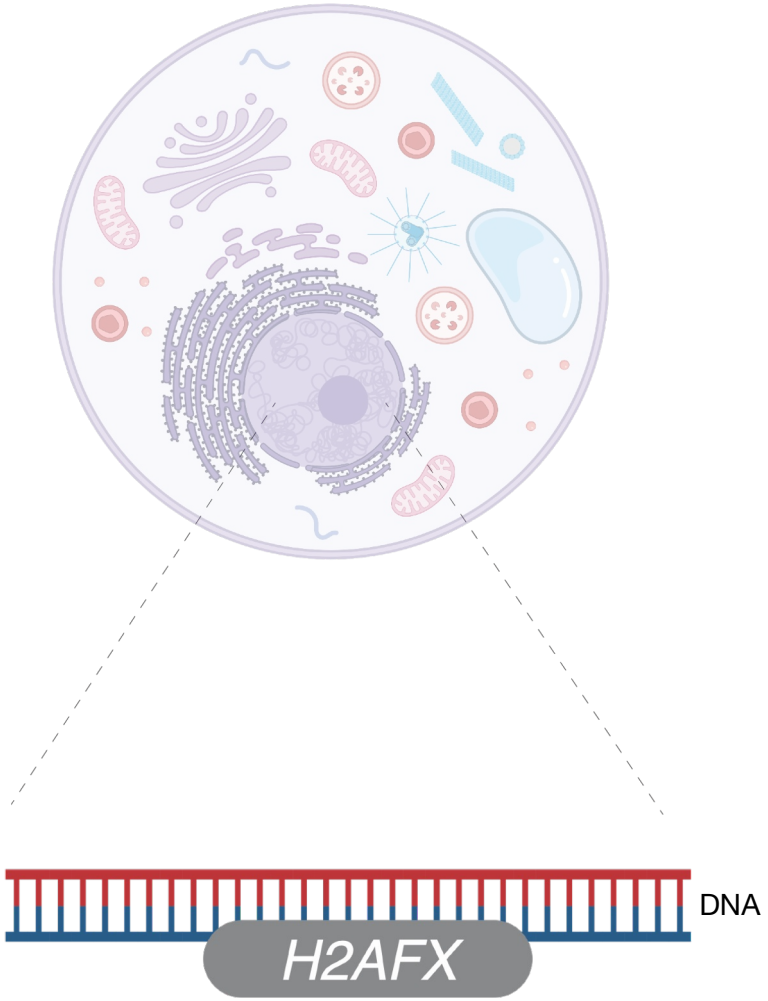
- grow an apple **tree**
- grow an apple **tree from seed**
- grow an apple **tree in a pot**
- grow an apple **tree indoors**

buy an apple

- buy an apple **watch**
- buy an apple **gift card**
- buy an apple **tv**

Sparse Dictionary Learning Recovers Pleiotropy from Human Cell Fitness Screens, *Cell Systems*, 2022
Contextualizing Protein Representations Using Deep Learning on Protein Networks and Single-Cell Data, Nature Methods, 2024 (in press)

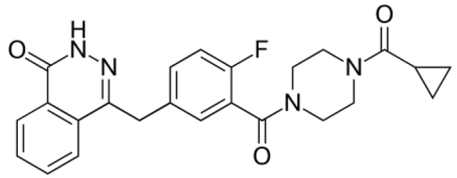
H2AFX is a **pleiotropic** gene…

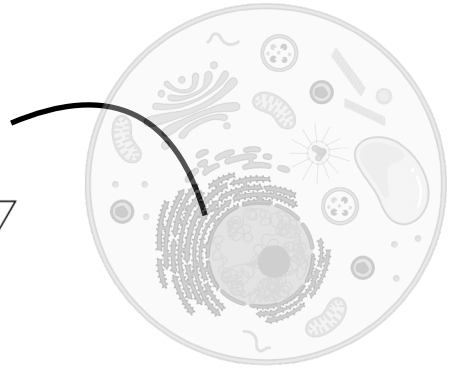


Sparse Dictionary Learning Recovers Pleiotropy from Human Cell Fitness Screens, *Cell Systems*, 2022
Contextualizing Protein Representations Using Deep Learning on Protein Networks and Single-Cell Data, Nature Methods, 2024 (in press)

… whose **particular function** is resolved via **cell context**

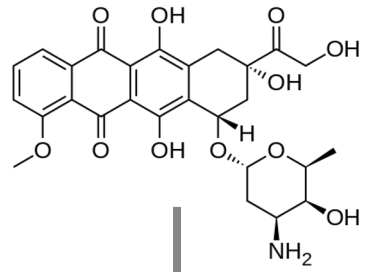
Olaparib



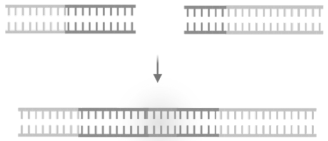


Homologous Recombination

Doxorubicin



End Joining

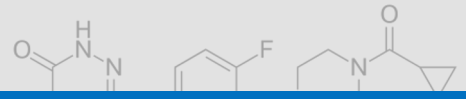


Sparse Dictionary Learning Recovers Pleiotropy from Human Cell Fitness Screens, *Cell Systems*, 2022

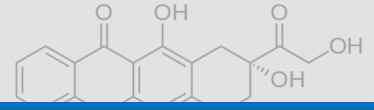
Contextualizing Protein Representations Using Deep Learning on Protein Networks and Single-Cell Data, Nature Methods, 2024 (in press)

… whose **particular function** is resolved via **cell context**

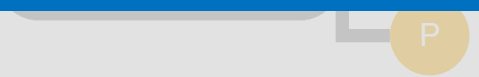
Olaparib



Doxorubicin

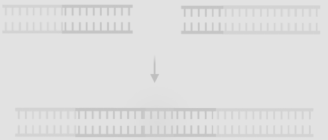


Can we develop **models that dynamically adjust their outputs to biological contexts in which they operate**?



Homologous Recombination

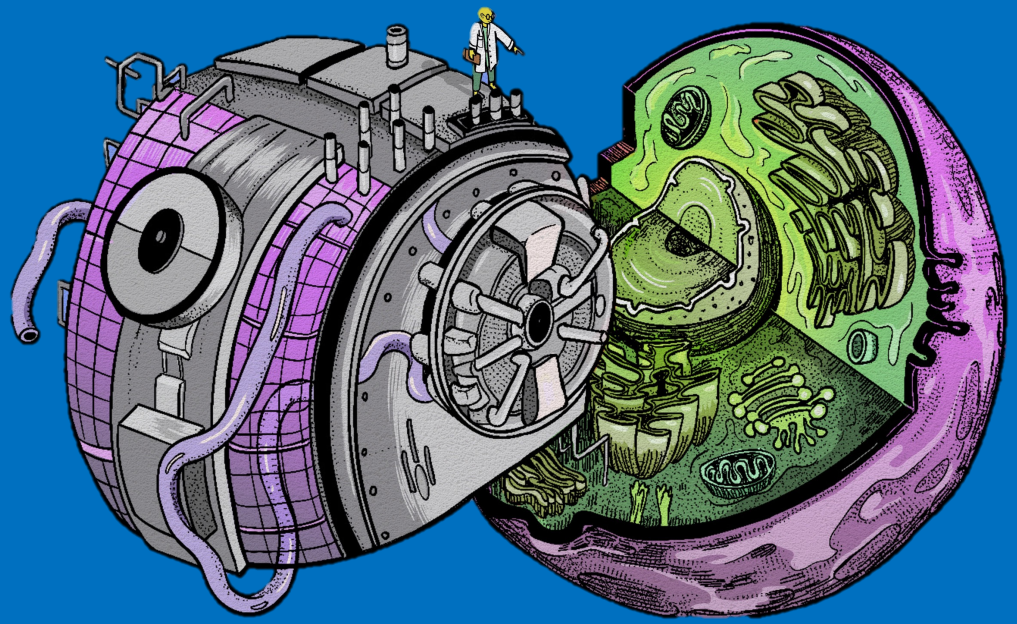
End Joining



Sparse Dictionary Learning Recovers Pleiotropy from Human Cell Fitness Screens, *Cell Systems*, 2022

Contextualizing Protein Representations Using Deep Learning on Protein Networks and Single-Cell Data, Nature Methods, 2024 (in press)

PINNACLE AI
Precise and cell-type specific protein representation learning



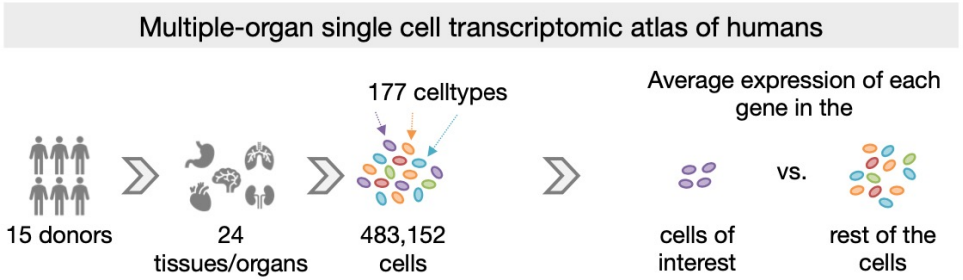
Providing outputs tailored to biological contexts is essential for broad use of foundation models in biology

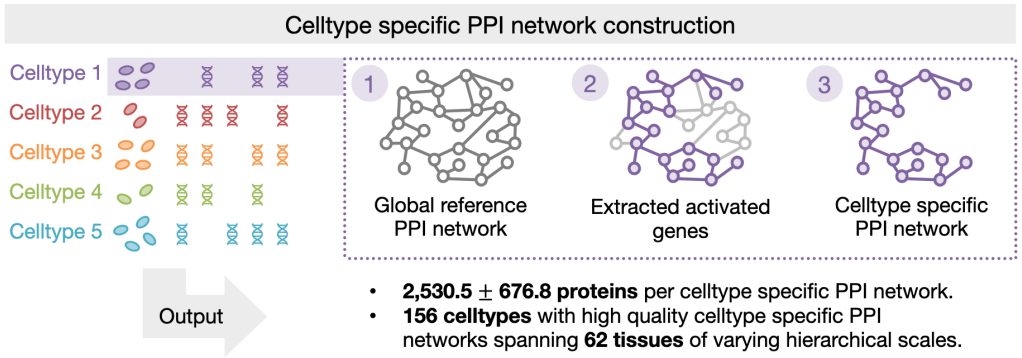
PINNACLE models support a broad array of tasks:

- 🧪 Enhance 3D structural protein representations
- 💊 Study effects of drugs across cell-type contexts
- 🎯 Nominate therapeutic targets in cell-type specific manner
- 🌲 Zero-shot retrieval of tissue hierarchy

PINNACLE: Geometric deep learning for precise protein representation learning and cell type- and state-specific prediction

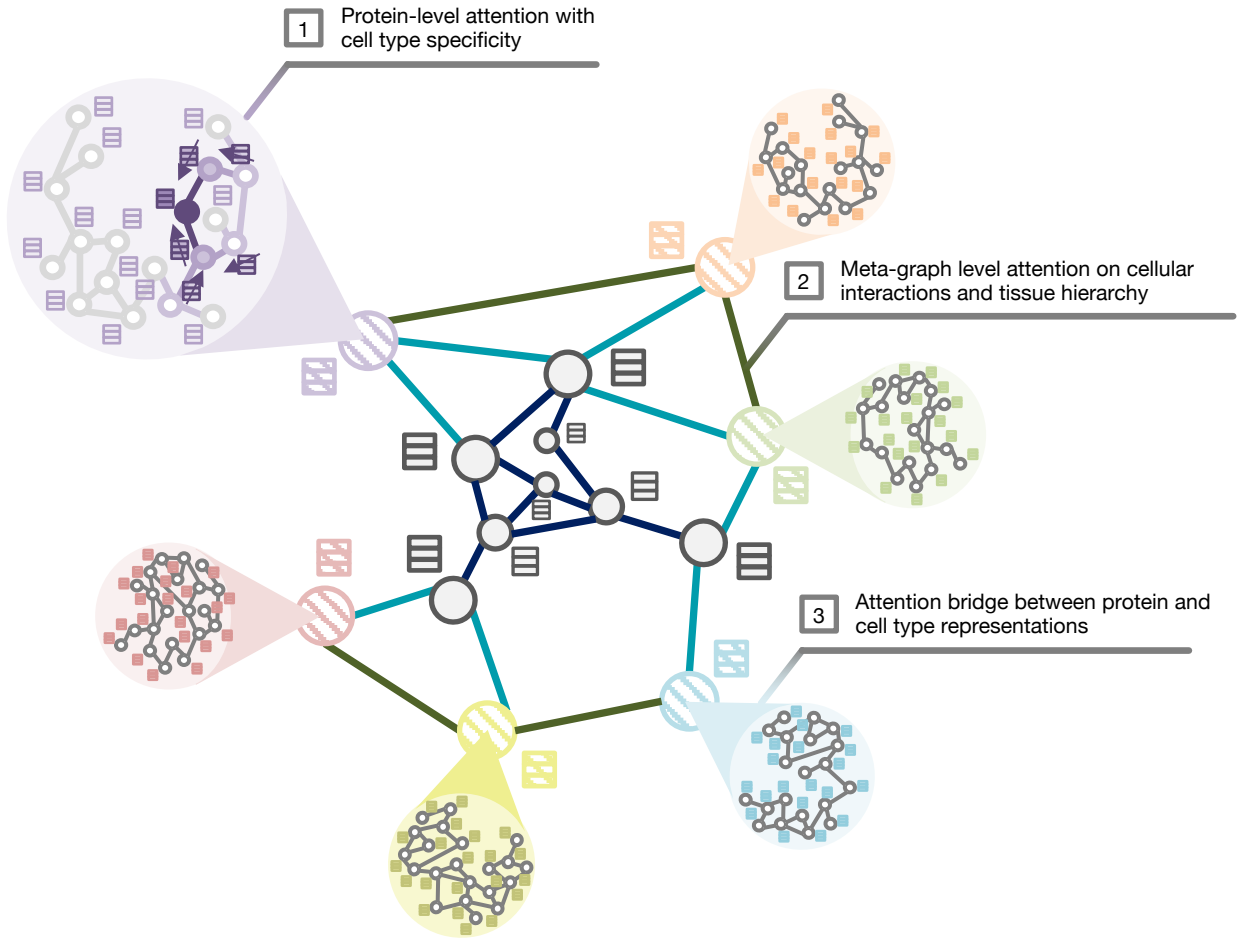
Data: Protein networks and single-cell transcriptomic data



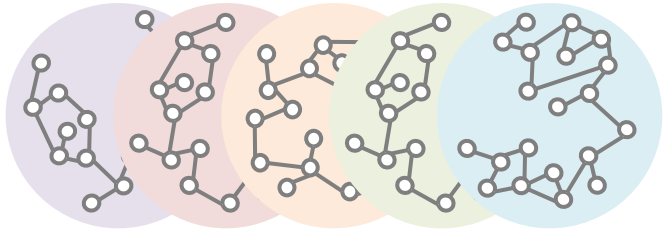


- **2,530.5 ± 676.8 proteins** per celltype specific PPI network.
- **156 celltypes** with high quality celltype specific PPI networks spanning **62 tissues** of varying hierarchical scales.

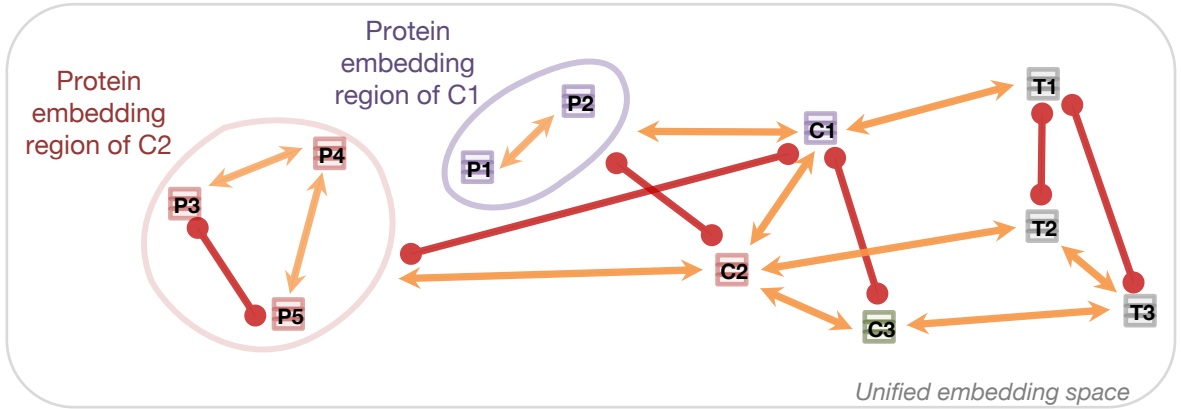
Model: Self-supervised contextual GNN



PINNACLE: Building our intuition

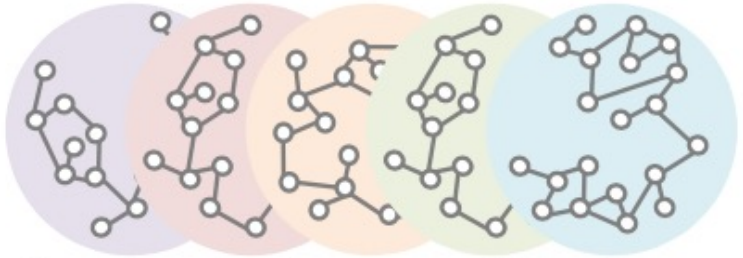


Protein networks across 156 cell type contexts spanning 62 tissues of varying hierarchical scales

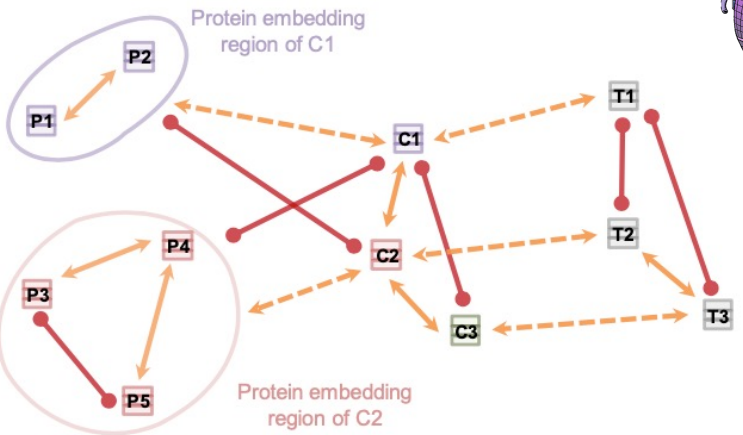


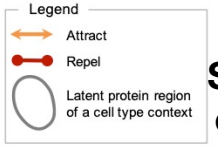
Self-supervised learning to learn a general encoder of protein embeddings tailored to cellular contexts

PINNACLE produces context-specific protein representations that are tailored to biological contexts, cell types and cell states

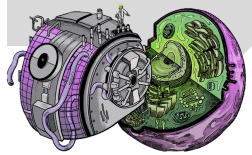


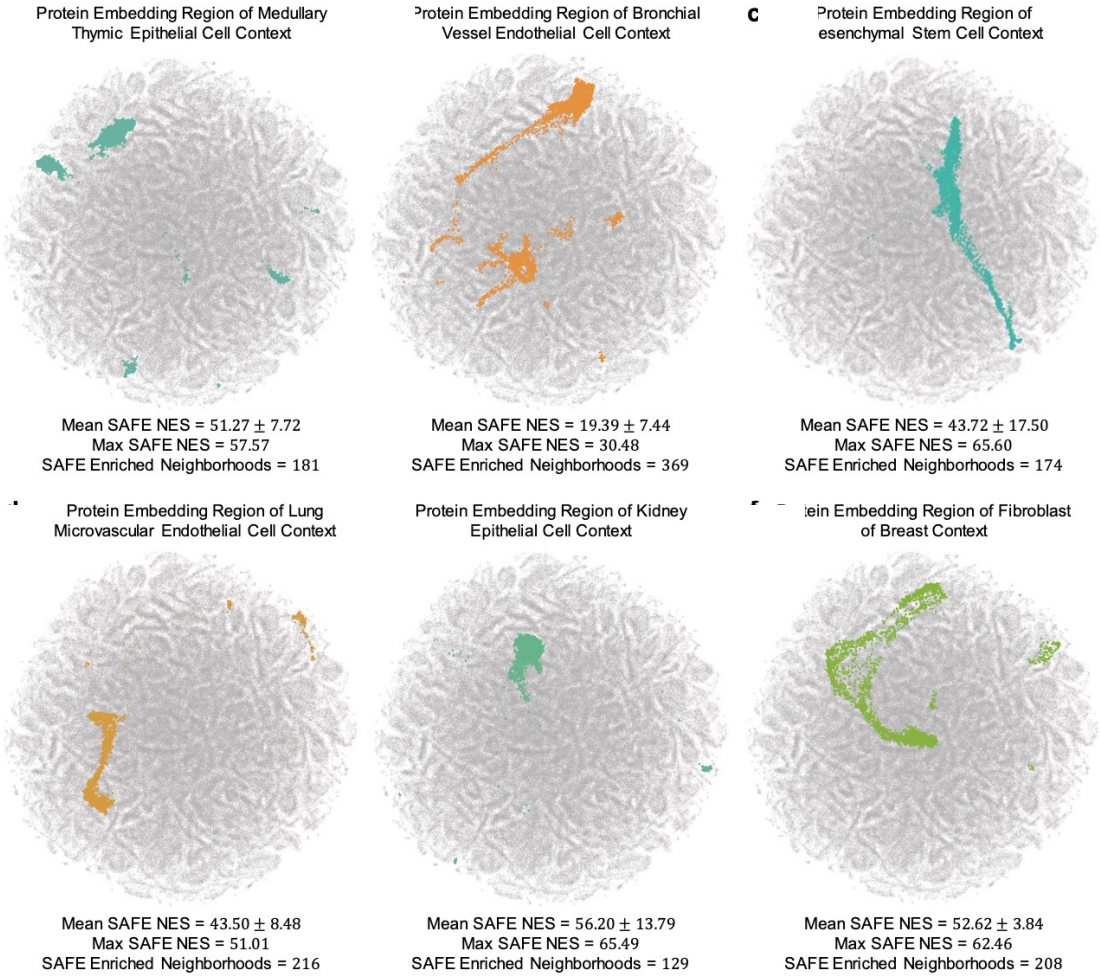
Protein networks across 156 cell type contexts spanning 62 tissues of varying hierarchical scales





PINNACLE

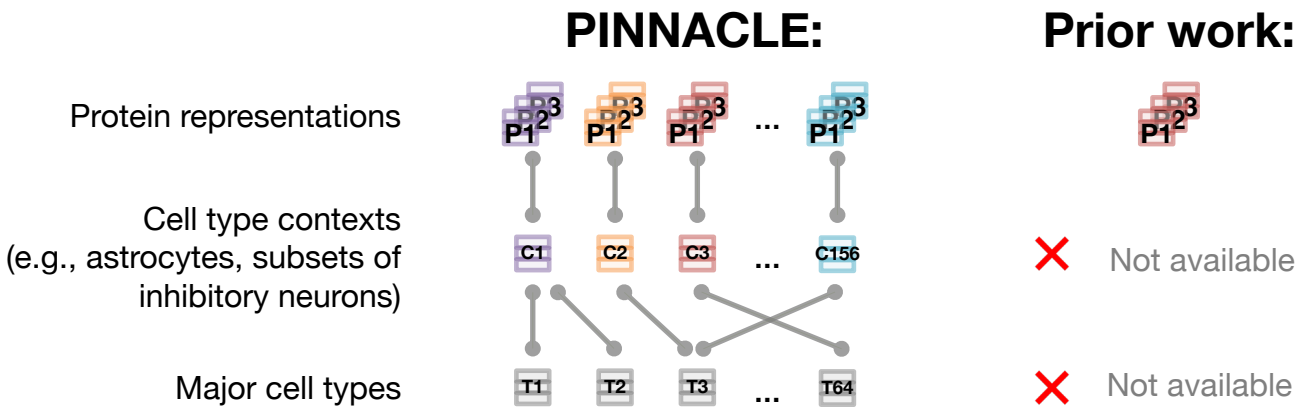




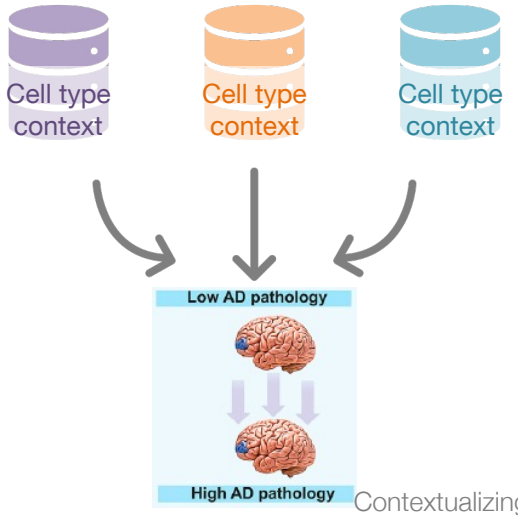
Self-supervised learning to learn a general encoder of protein embeddings tailored to cellular contexts

500,000 protein latent representations contextualized to 156 cross-organ cell types

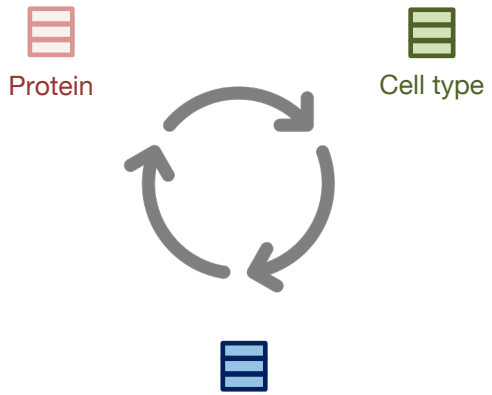
PINNACLE enables contextualized, precise predictions of drug effects across cell types and cell states



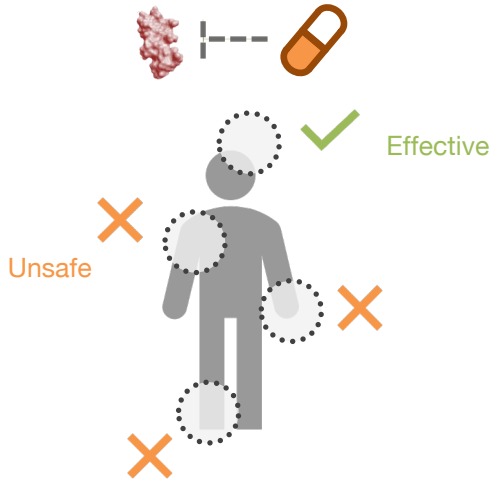
Multi-modal deep learning to identify changes at the single-cell level predictive of cognitive and behavioral phenotypes



Transfer learning across cellular contexts to predict if candidate drugs affect disease-relevant cell types

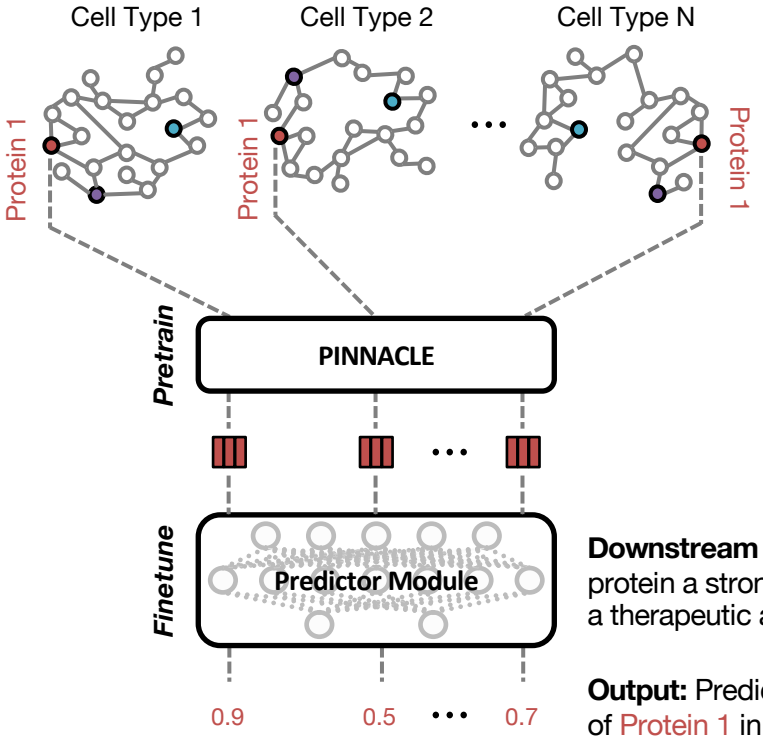


Identify molecules that are most effective at modulating astrocytes



Contextualizing Protein Representations Using Deep Learning on Protein Networks and Single-Cell Data, Nature Methods, 2024 (in press)

Contextualized prediction: Setup



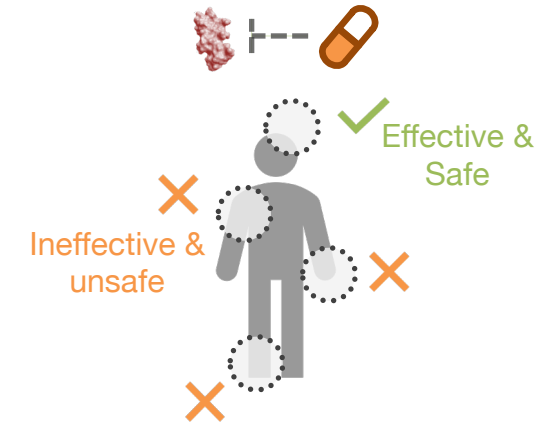
Cell Type

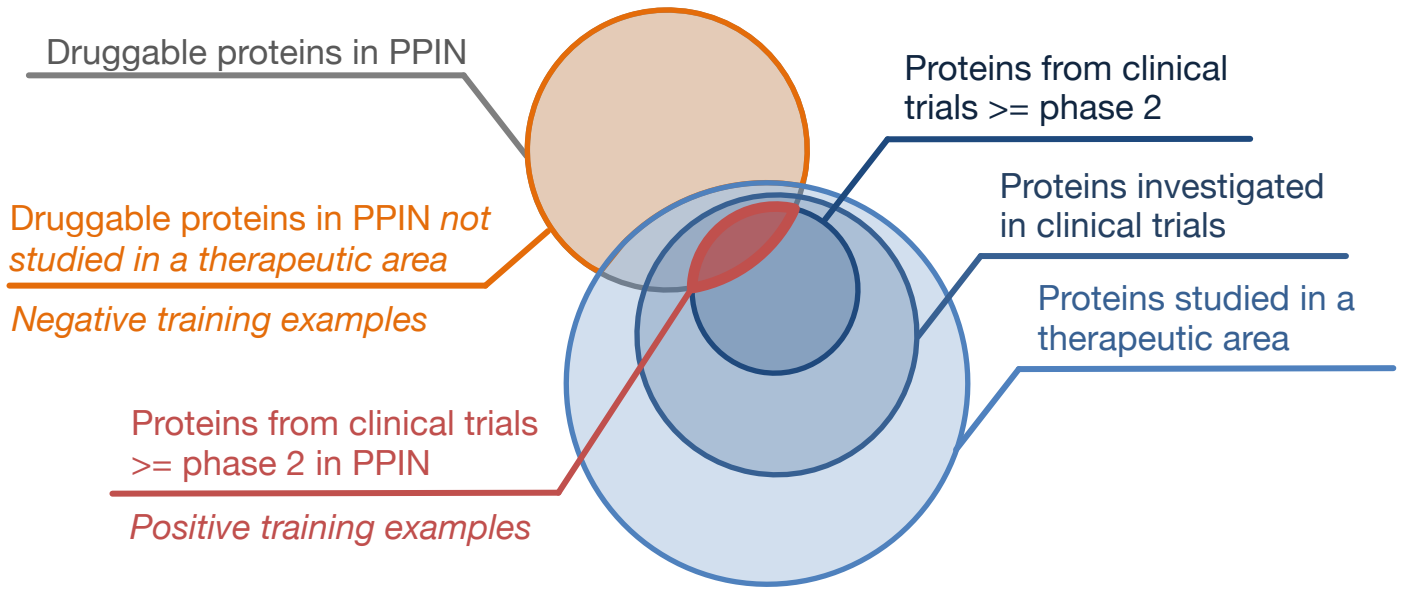
	1	2	…	N
Protein 1	0.9	0.5	…	0.7
Protein 2	0.3	0.4	…	0.1
…	…	…	…	…
Protein M	0.6	0.8	…	0.2

Downstream task: Is a given protein a strong candidate for a therapeutic area of interest?

Output: Predicted probability of Protein 1 in Cell Types 1, 2, …, N being a target for the therapeutic area of interest?

Additional insights: Which cell type is highly predictive to examine therapeutic targets?





Open Targets

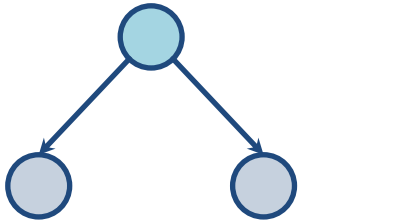
Dataset & experimental setup

1. Specify therapeutic area
(seed disease and its descendants in a disease ontology)

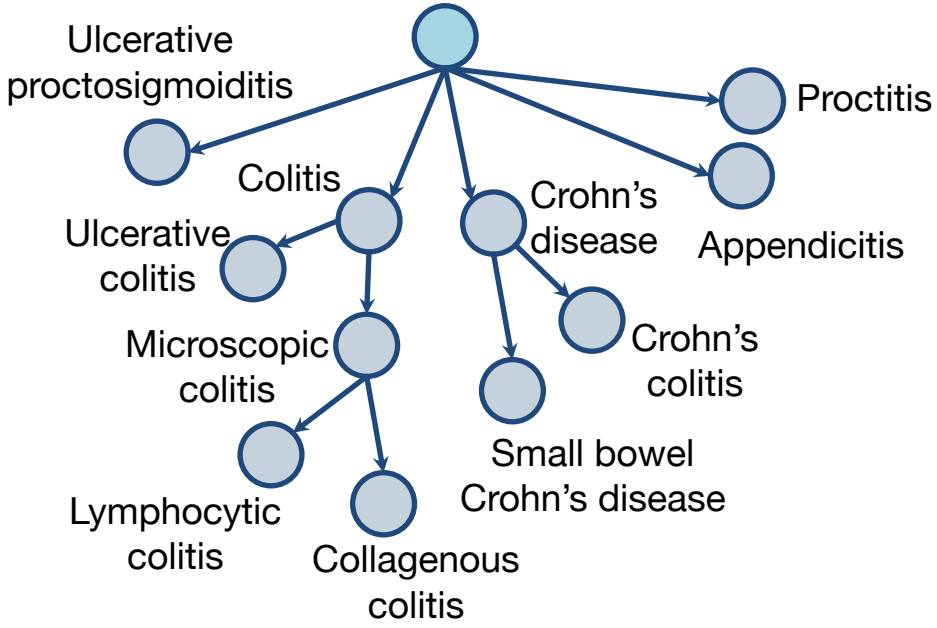
2. Curate clinical trials for diseases
(at least one completed clinical phase II or more)

3. Curate list of candidate targets
(proteins targeted by a drug in a clinical trials)

Rheumatoid arthritis (RA)



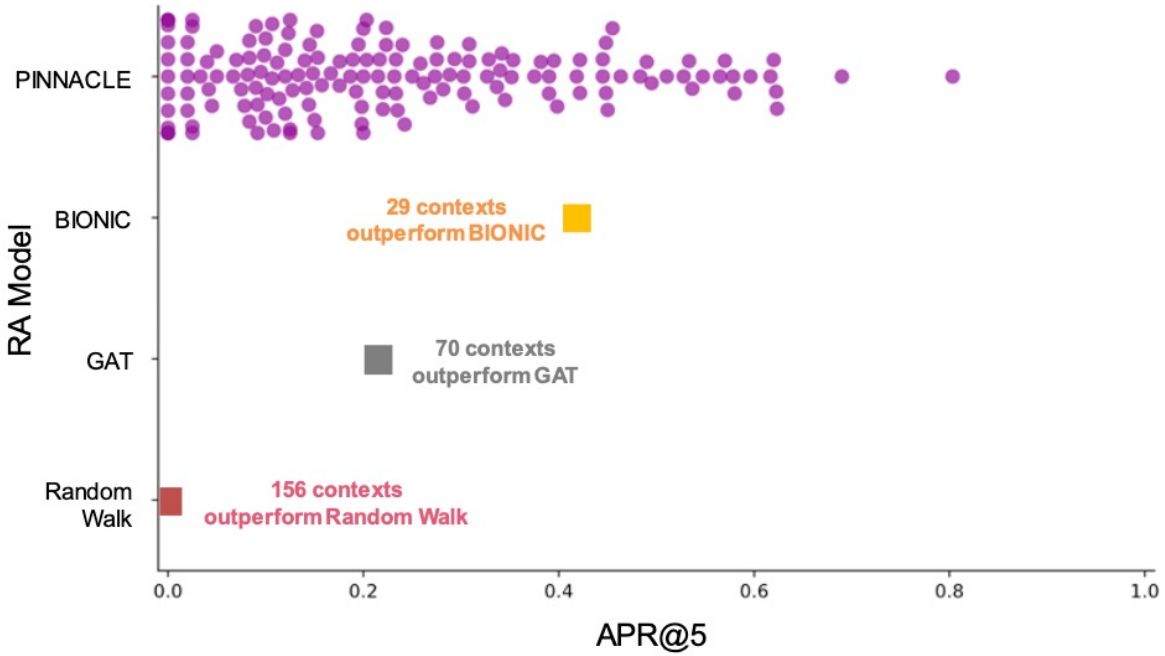
Inflammatory bowel disease (IBD)

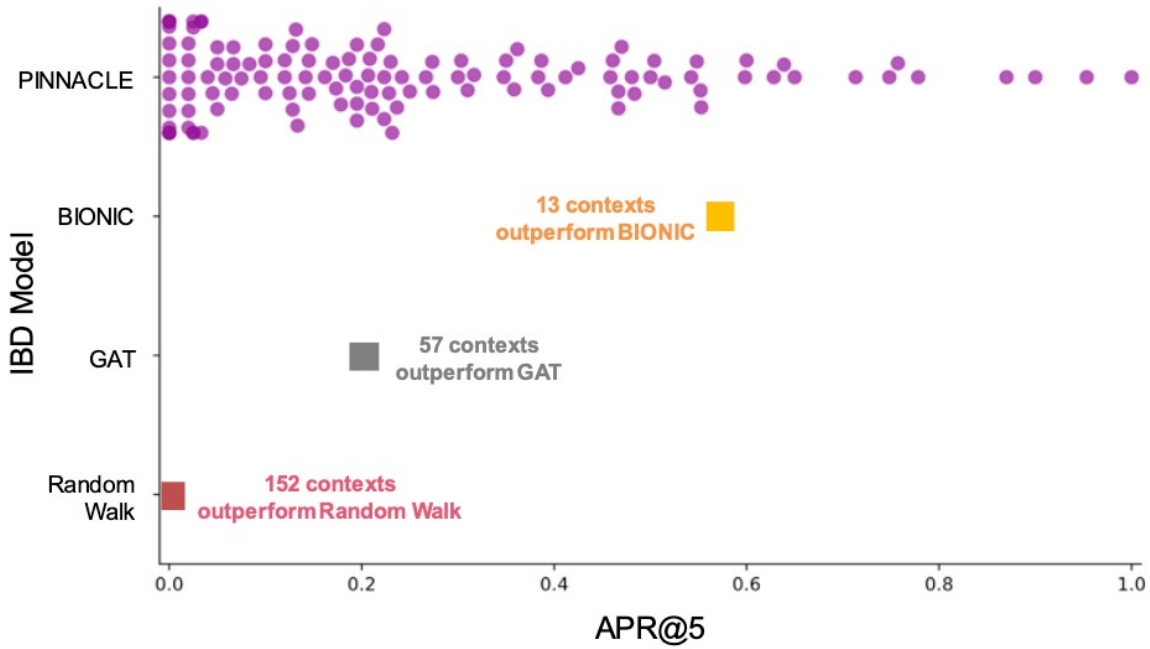


Clinical phase	Unique drugs	Unique proteins
2	81	110
3	27	26
4	94	49

Clinical phase	Unique drugs	Unique proteins
2	41	67
3	21	26
4	59	46

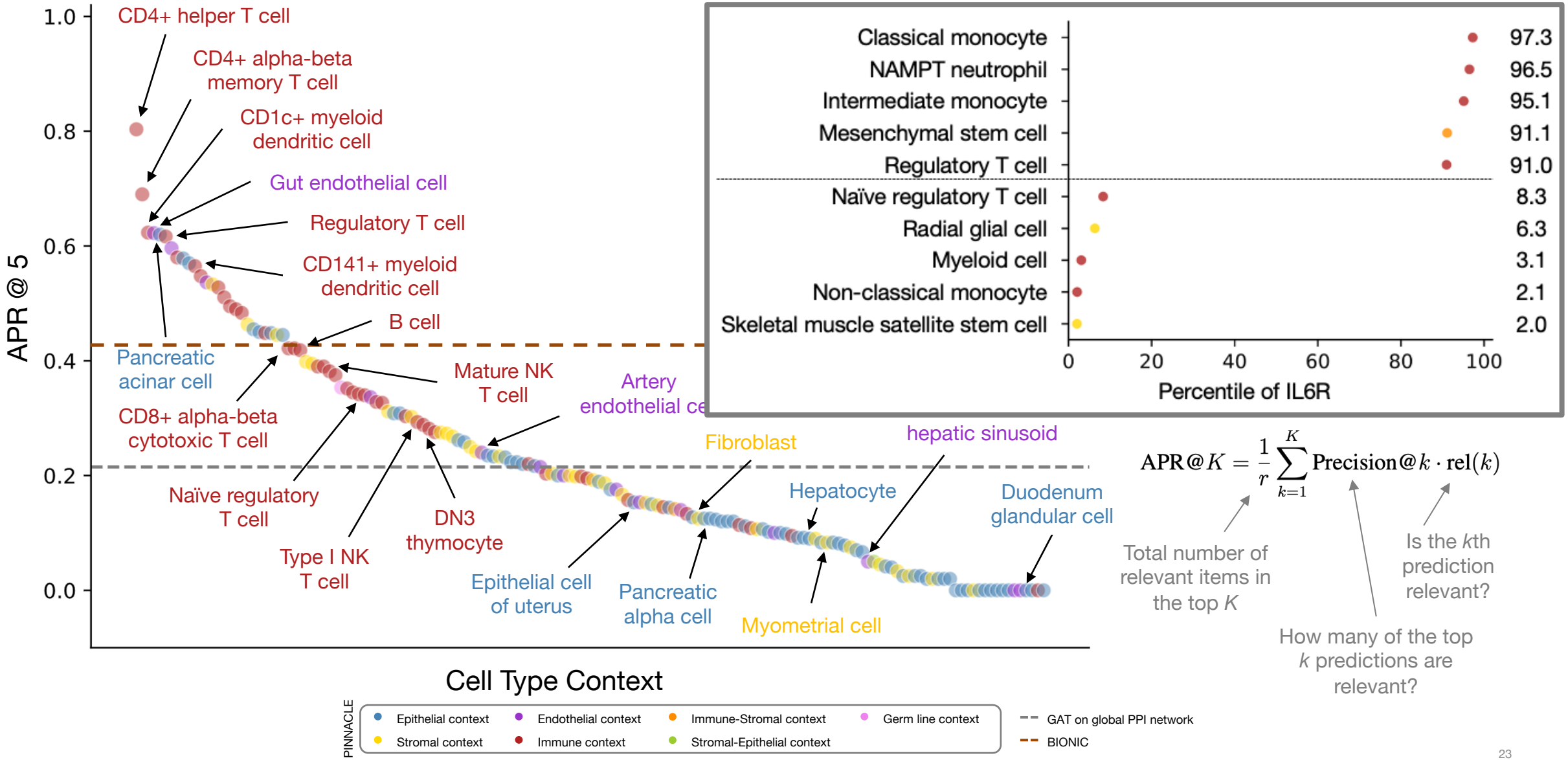
PINNACLE accurately nominates therapeutic targets for RA and IBD in a cell-type specific manner, whereas existing data integration models conflate cell-type specific information, leading to poor performance





PINNACLE's representations **predict drug targets for RA and IBD significantly better than context-free methods**. PINNACLE also finds most predictive cell type contexts for investigating protein targets. PINNACLE supports the study of drug effects across cell type contexts.

PINNACLE considers core disease processes in rheumatoid arthritis to identify candidate targets in cell type-specific manner



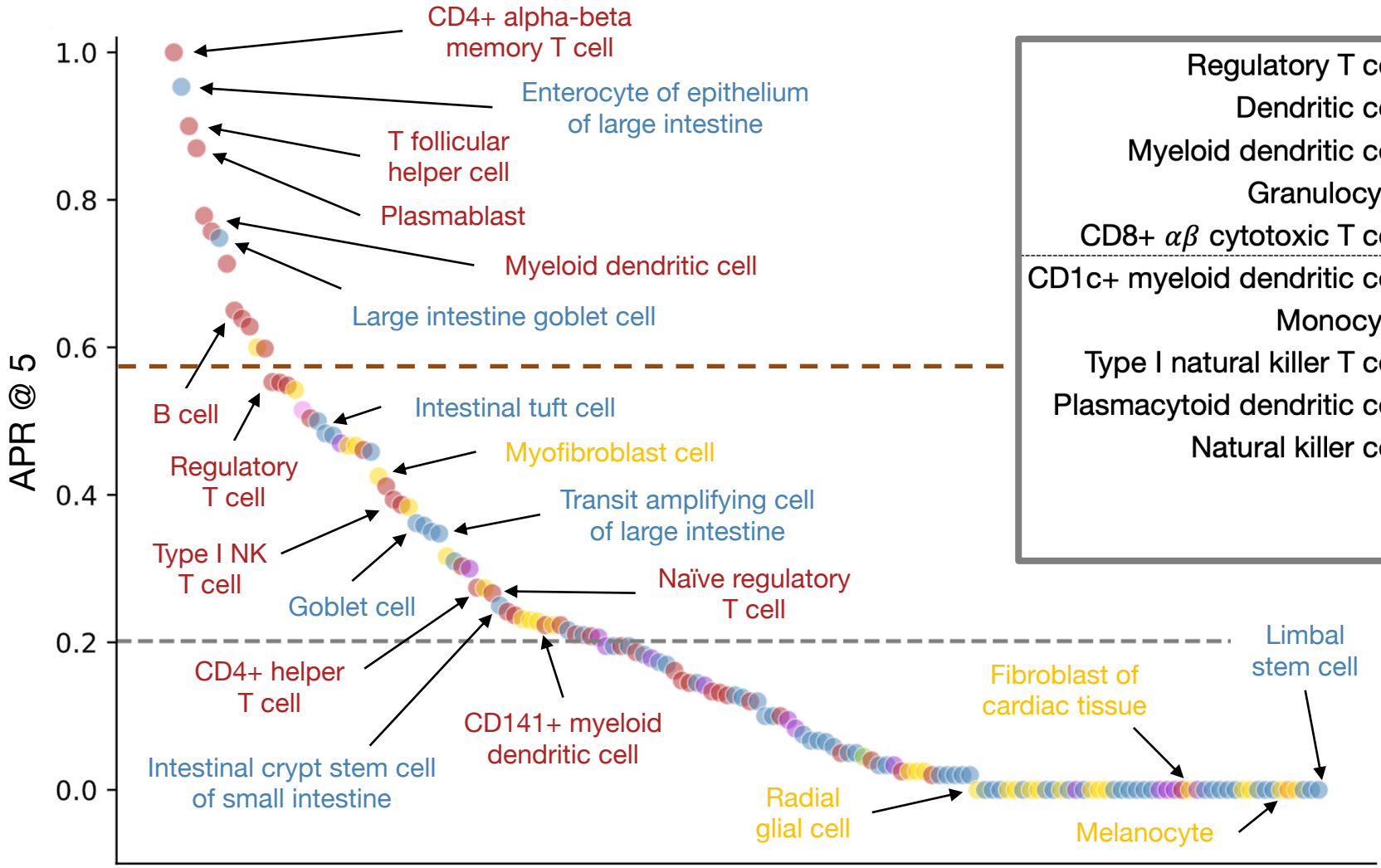
$$\text{APR@}K = \frac{1}{r}\sum_{k=1}^{K} \text{Precision@}k \cdot \text{rel}(k)$$

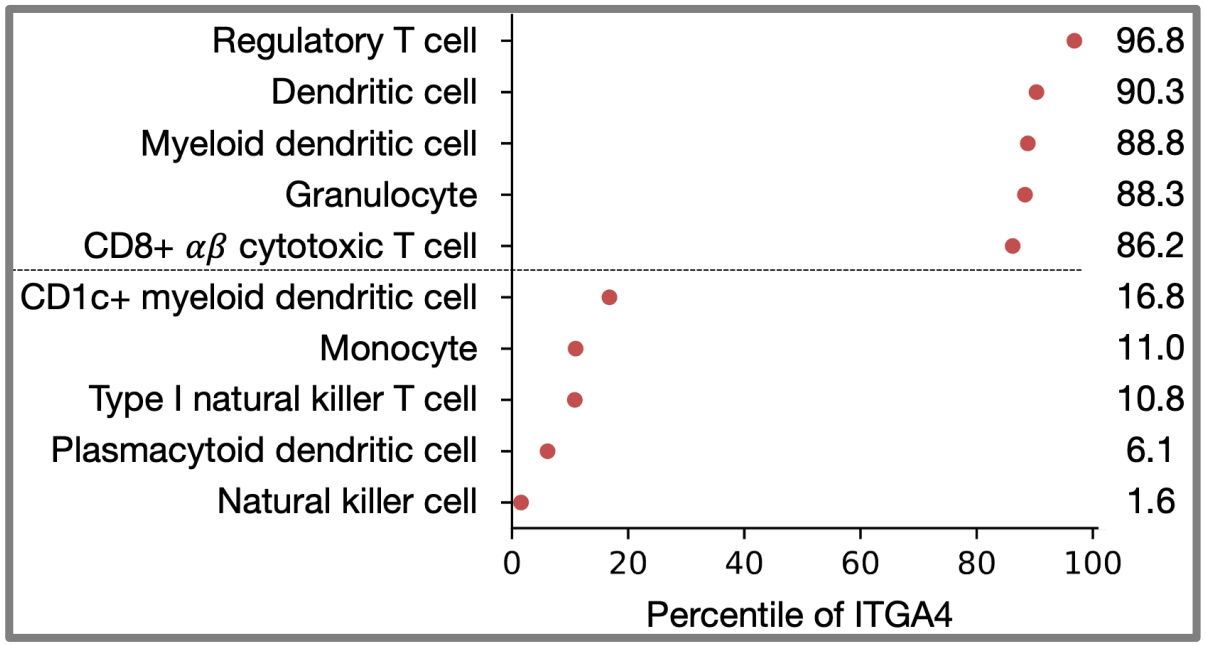
Total number of relevant items in the top K

How many of the top k predictions are relevant?

Is the kth prediction relevant?

PINNACLE considers core disease processes in inflammatory bowel disease to nominate candidate targets in cell type-specific manner



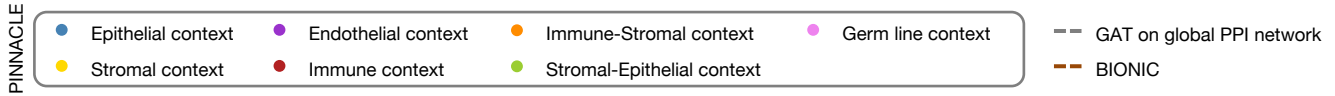


$$\text{APR@}K = \frac{1}{r}\sum_{k=1}^{K} \text{Precision@}k \cdot \text{rel}(k)$$

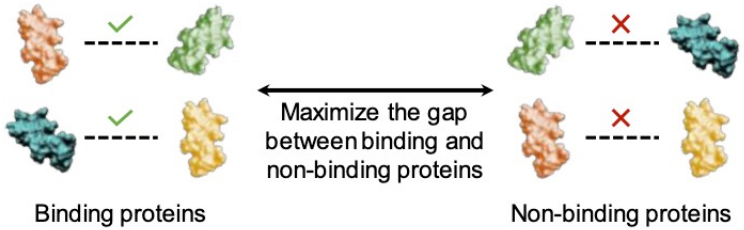
Total number of relevant items in the top K

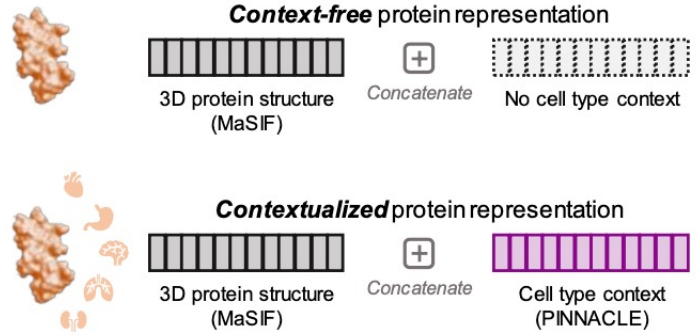
How many of the top k predictions are relevant?

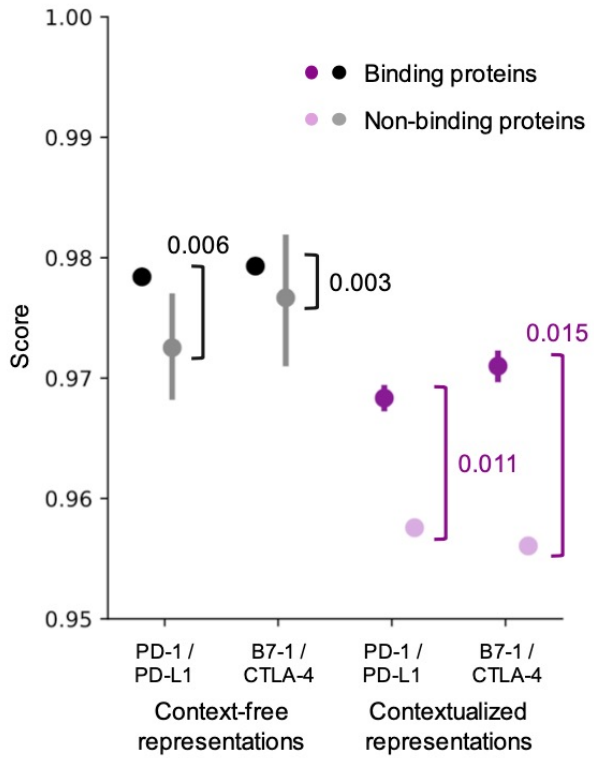
Is the kth prediction relevant?



Contextualized predictions by integrating PINNACLE with 3D structures

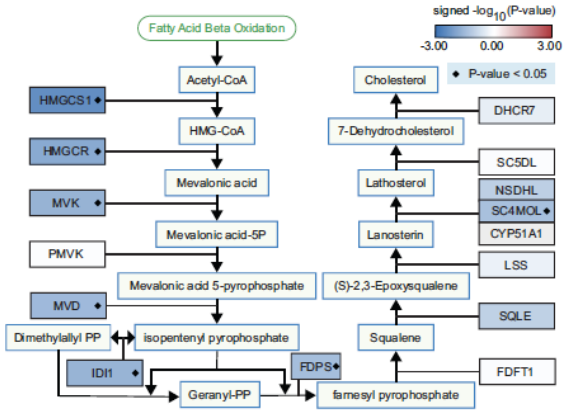






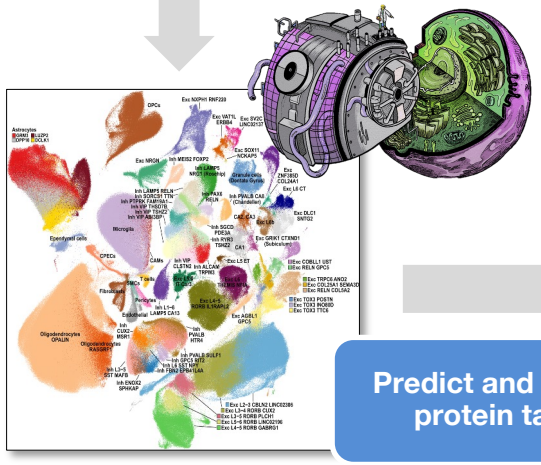
PINNACLE's representations **enhance 3D structure-based protein representations** for important protein interactions in immuno-oncology (PD-1/PD-L1 and B7-1/CTLA-4)

Disease circuitry across diseases and individuals

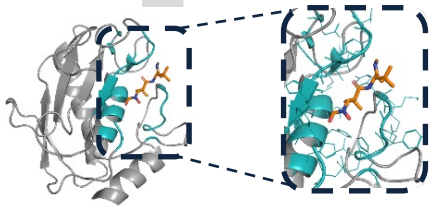


Key disease mechanisms, shared effects, interaction effects

Multimodal representation learning models
PINNACLE, PDGrapher



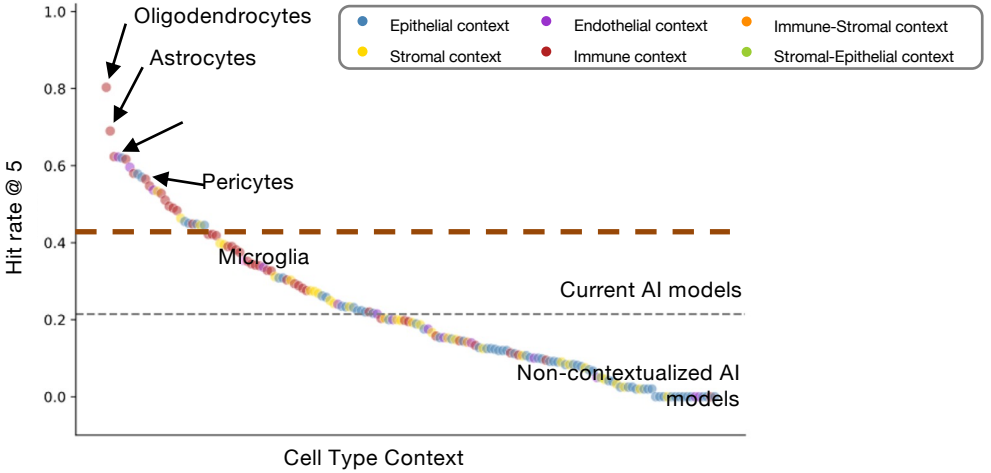
Generative geometric deep learning models
PocketGen, FAIR, and TxPLM



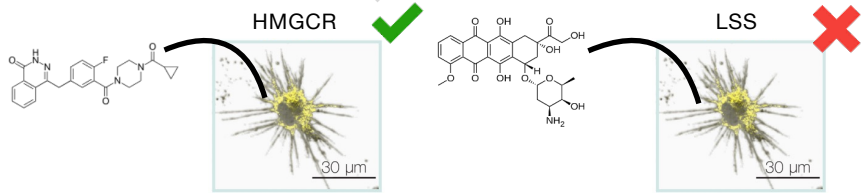
Full-atom structure

Generate, and refine therapeutic candidates

Predict and optimize protein targets



Rank-ordered lists of molecules to modulate disease circuitry in a cell-type specific manner

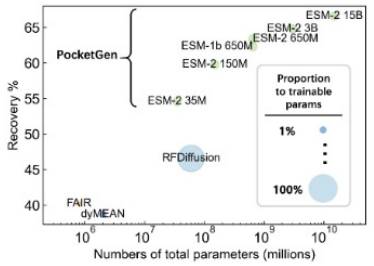


Chemistry optimization, synthesis, automation

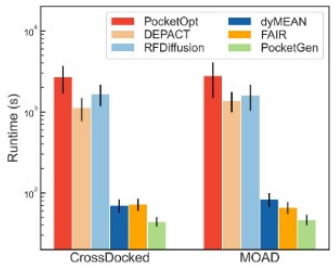
Generative sequence-structure models enable atom-level predictions of ligands binding to biological targets

Generative models:

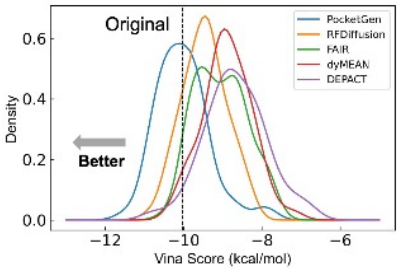
- **co-design of protein pocket sequence and 3D structure**
- **selective small molecule ligands**
- **optimized PPI interfaces**



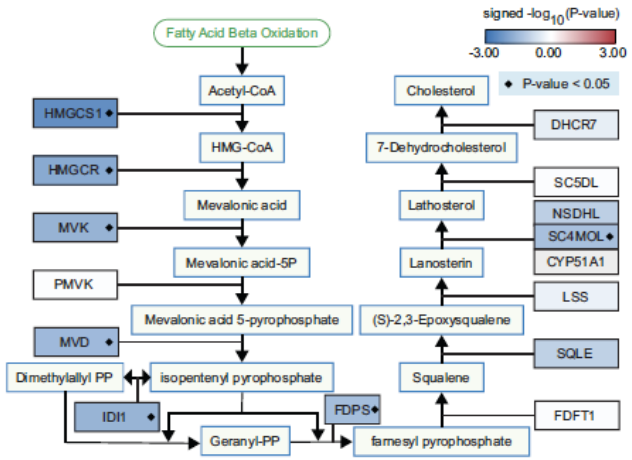
Iterative refinement based on side-chain effects, ligand flexibility, and sequence-structure consistency



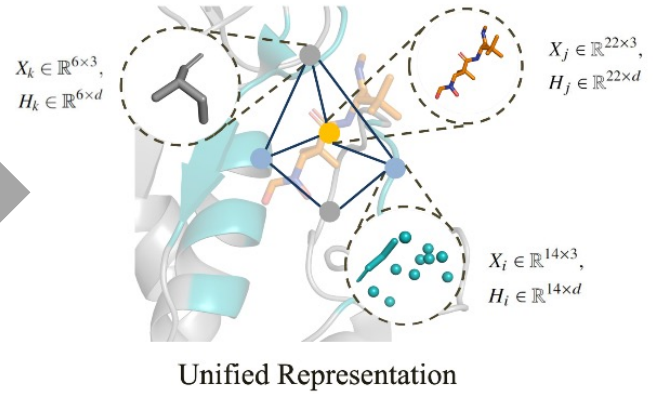
10x faster than current AI, 15% better accuracy (AAR, RMSE, docking score)



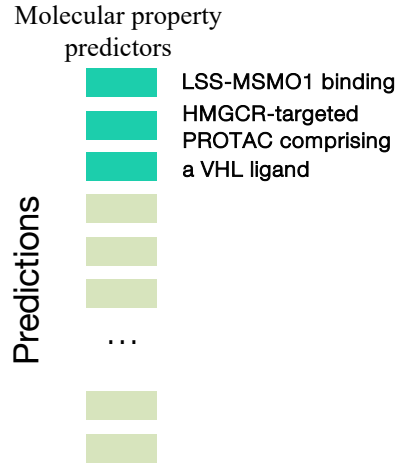
45% better hit rate than current AI, need to generate fewer molecules to find a hit



Key disease mechanisms, shared effects, interaction effects



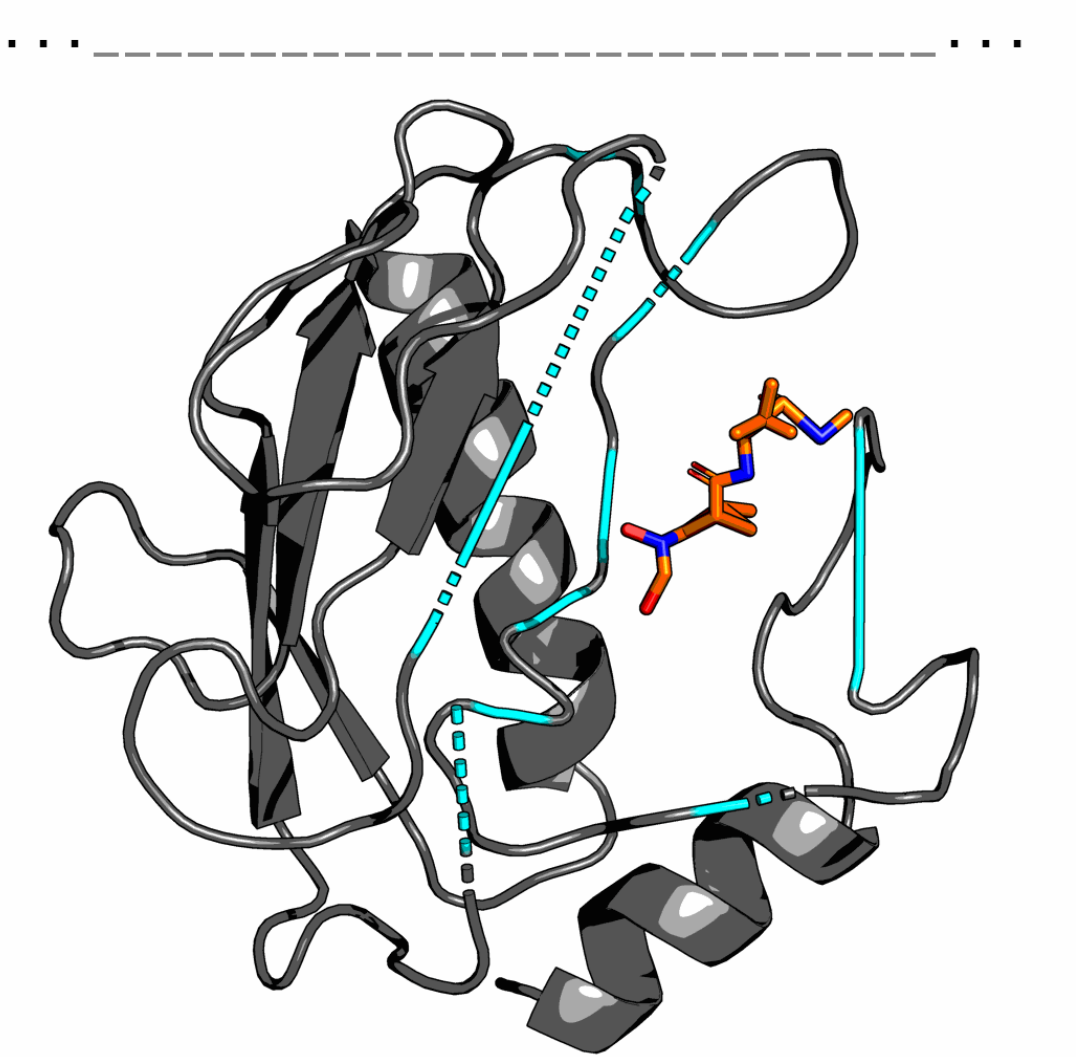
Generative sequence-structure models



Priority lists of generated molecular structures

Gysi et al., *PNAS* 2021; Ruiz et al., *Nature Communications* 2021; Full-Atom Protein Pocket Design via Iterative Refinement, *NeurIPS* 2023; PocketGen: Generating Full-Atom Ligand-Binding Protein Pockets, *bioRxiv* 2024

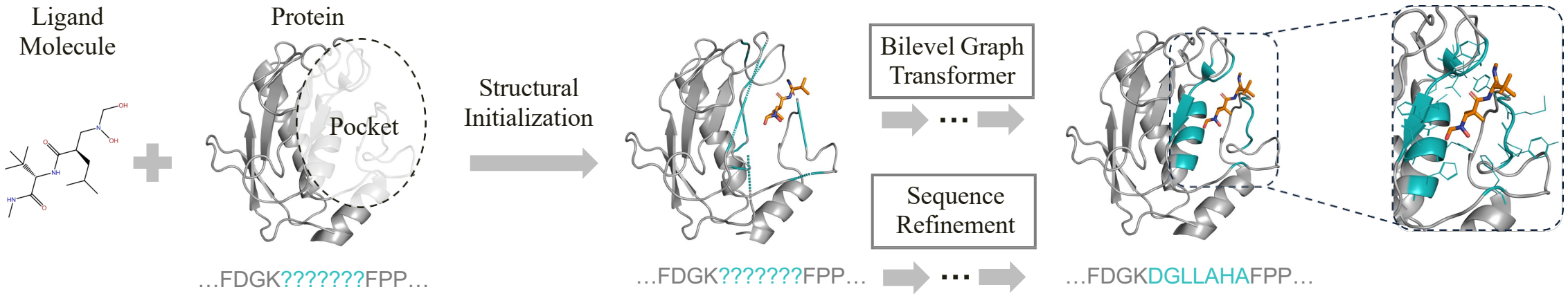
Sequence-structure co-generation of protein pockets

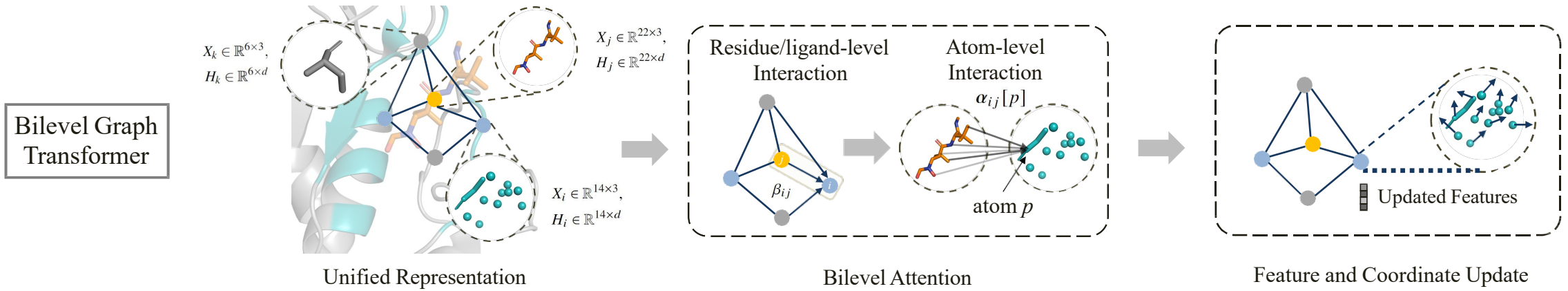


- Generating **high-fidelity protein pockets**—an area where a protein interacts with a ligand molecule
  - Complex interactions between ligand molecules and proteins
  - Flexibility of ligands and AA side chains
  - Complex sequence-structure dependencies
- PocketGen generates residue sequence and full-atom structure within protein pocket region

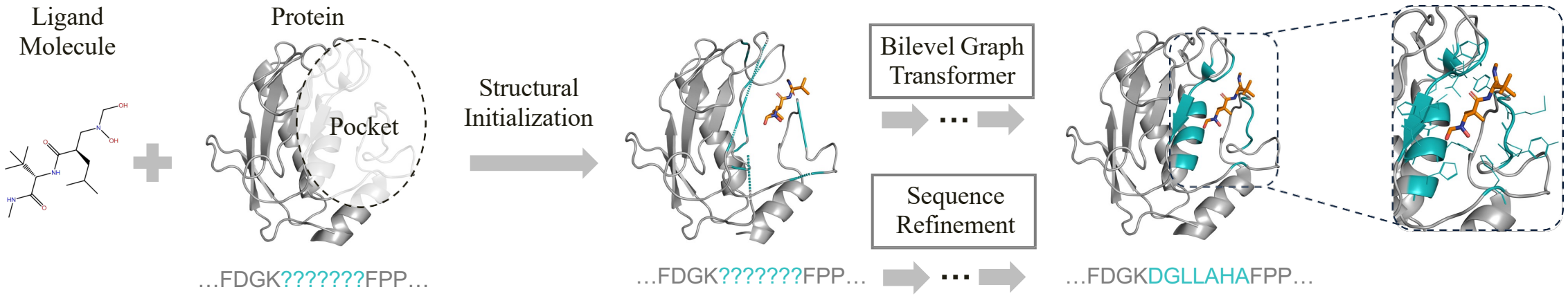
Full-Atom Protein Pocket Design via Iterative Refinement, *NeurIPS* 2023; PocketGen: Generating Full-Atom Ligand-Binding Protein Pockets, *bioRxiv* 2024

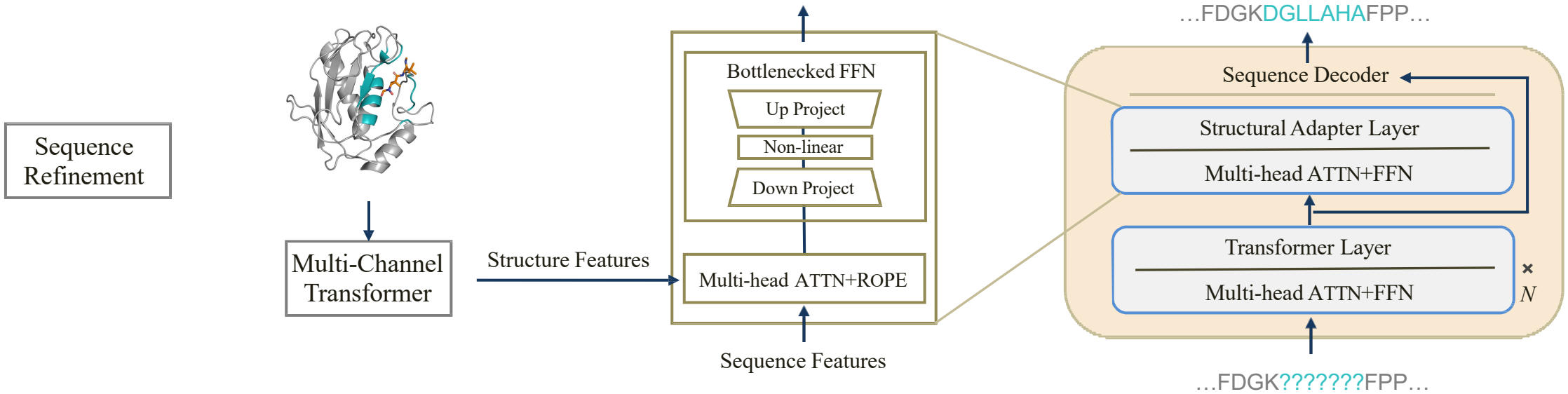
Iterative refinement of both sequence and structure in the protein pocket to maximize binding affinity with small molecule ligand





Iterative refinement of both sequence and structure in the protein pocket to maximize binding affinity with small molecule ligand





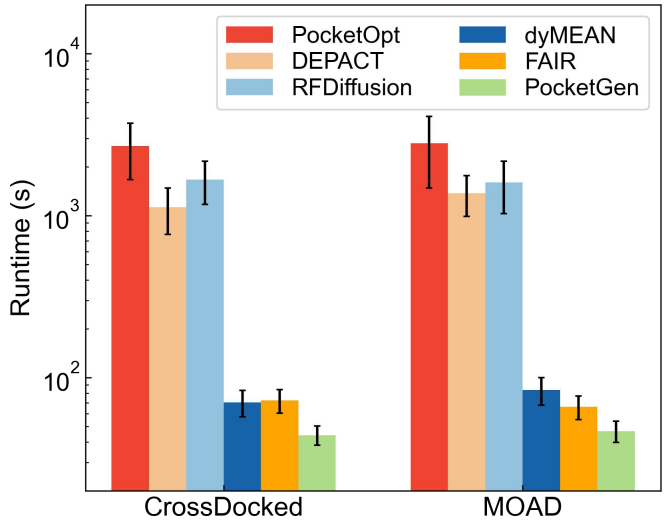
Full-Atom Protein Pocket Design via Iterative Refinement, *NeurIPS* 2023; PocketGen: Generating Full-Atom Ligand-Binding Protein Pockets, *bioRxiv* 2024

PocketGen generates protein pockets with higher binding affinity and structural validity than existing models

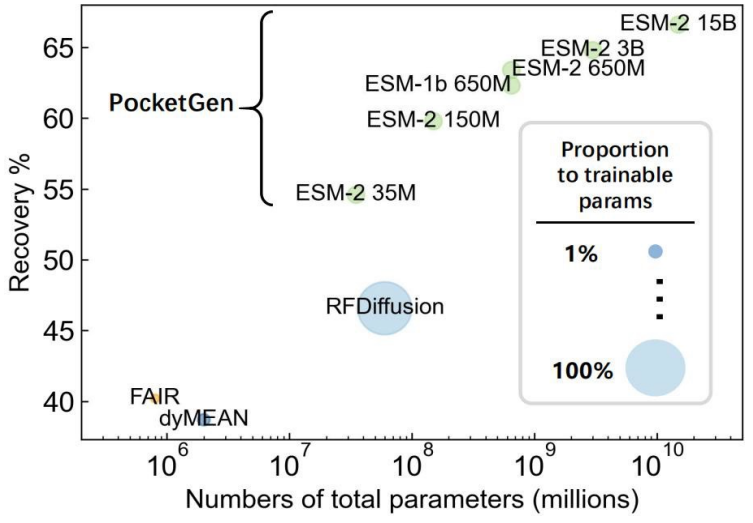
	PocketOpt	DEPACT	dyMEAN	FAIR	RFDiffusion	PocketGen
Top-1 generated protein pocket						
Vina score (↓)	-9.216	-8.527	-8.540	-8.792	-9.037	**-9.655**
Success Rate (↑)	0.92	0.75	0.76	0.80	0.89	**0.97**
RMSD (↓)	-	1.47	1.44	1.39	**1.13**	1.21
pLDDT (↑)	-	82.1	83.3	83.2	84.5	**86.7**
scTM (↑)	-	0.901	0.906	0.899	0.924	**0.937**
Top-3 generated protein pockets						
Vina score (↓)	-8.878	-8.131	-8.196	-8.321	-8.876	**-9.353**
RMSD (↓)	-	1.45	1.43	1.40	**1.18**	1.24
pLDDT (↑)	-	81.9	82.8	83.1	84.6	**86.2**
scTM (↑)	-	0.896	0.892	0.897	0.929	**0.934**
Top-5 generated protein pockets						
Vina score (↓)	-8.702	-7.786	-7.974	-7.943	-8.510	**-9.239**
RMSD (↓)	-	1.46	1.45	1.42	1.25	**1.22**
pLDDT (↑)	-	82.2	82.9	83.3	84.3	**86.1**
scTM (↑)	-	0.892	0.903	0.886	0.926	**0.935**
Top-10 generated protein pockets						
Vina score (↓)	-8.556	-7.681	-7.690	-7.785	-8.352	**-9.065**
RMSD (↓)	-	1.53	1.44	1.41	**1.26**	1.28
pLDDT (↑)	-	81.5	82.7	83.0	84.2	**85.9**
scTM (↑)	-	0.895	0.896	0.884	0.924	**0.931**

Improved structural validity, amino acid sequence recovery, and affinity with target ligands

Model	CrossDocked			Binding MOAD		
	AAR (↑)	RMSD (↓)	Vina (↓)	AAR (↑)	RMSD (↓)	Vina (↓)
Test set	-	-	-7.016	-	-	-8.076
DEPACT	31.52±3.26%	1.59±0.13	-6.632±0.18	35.30±2.19%	1.52±0.12	-7.571±0.15
dyMEAN	38.71±2.16%	1.57±0.09	-6.855±0.06	41.22±1.40%	1.53±0.08	-7.675±0.09
FAIR	40.16±1.17%	1.46±0.04	-7.015±0.12	43.68±0.92%	1.37±0.07	-7.930±0.15
RFDiffusion	46.57±2.07%	1.44±0.07	-6.936±0.07	45.31±2.73%	1.45±0.10	-7.942±0.14
PocketGen	**63.40±1.64%**	**1.36±0.05**	**-7.135±0.08**	**64.43±2.35%**	**1.32±0.05**	**-8.112±0.14**



Better generation efficiency



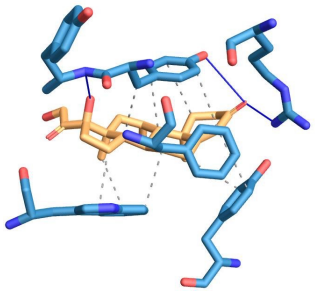
Performance wrt protein LM size

PocketGen can redesign pockets of antibodies, enzymes, and biosensors for target ligand molecules

Protein
Ligand
Aromatic Ring Center
Hydrophobic Interaction
Hydrogen Bond
π-Stacking (parallel)
π-Stacking (perpendicular)
π-Cation Interaction

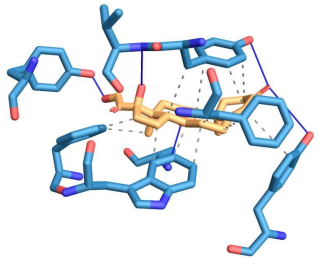
Cortisol (HCY)

…WFRYYDTMY…



Original
(HP 12, HB 3)

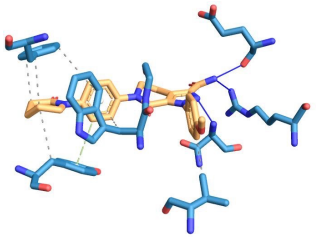
…WFRYVYSFY…



PocketGen
(HP 12, HB 5)

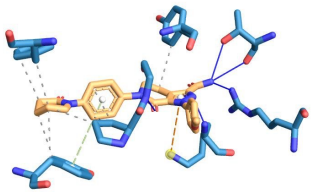
Apixaban (APX)

…YREFQVWGG…



Original
(HP 7, HB 4, π 1)

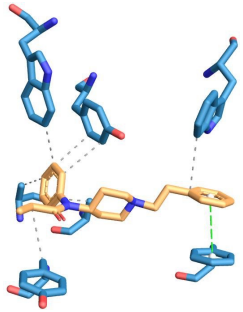
…YRTFKVPGY…



PocketGen
(HP 9, HB 5, π 2)

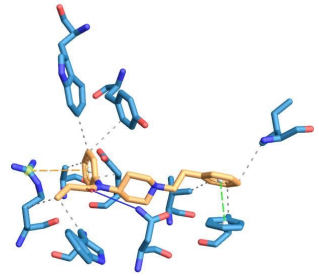
Fentanyl 7V7)

…FVWAWYWAY…

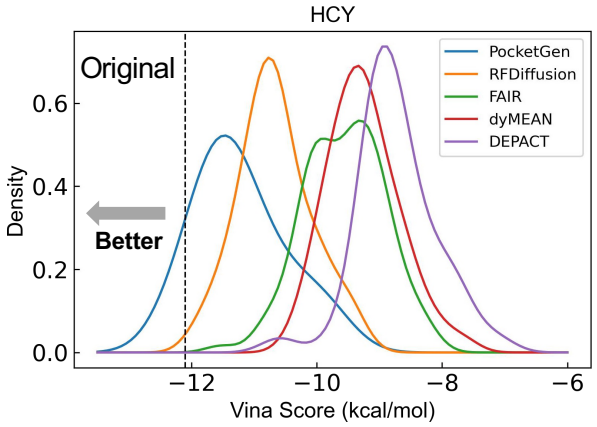


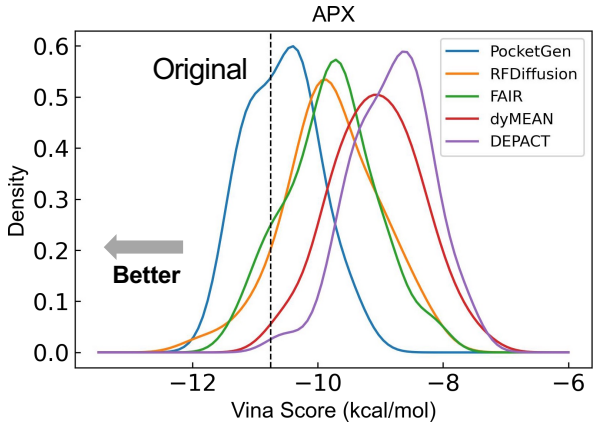
Original
(HP 7, HB 0, π 1)

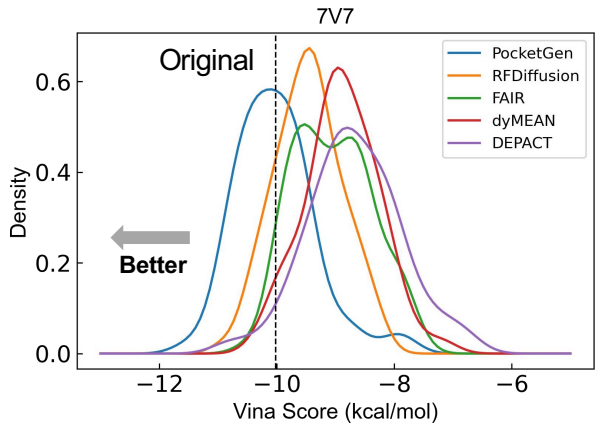
…FVRAWYWEW…



PocketGen
(HP 9, HB 1, π 2)

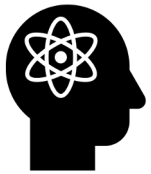






Full-Atom Protein Pocket Design via Iterative Refinement, *NeurIPS* 2023; PocketGen: Generating Full-Atom Ligand-Binding Protein Pockets, *bioRxiv* 2024

https://tdcommons.ai



THERAPEUTICS DATA COMMONS

Biomedical scientists → **Data** ↔ **AI** ↔ **Scientific hypotheses** ← **AI scientists**

↓

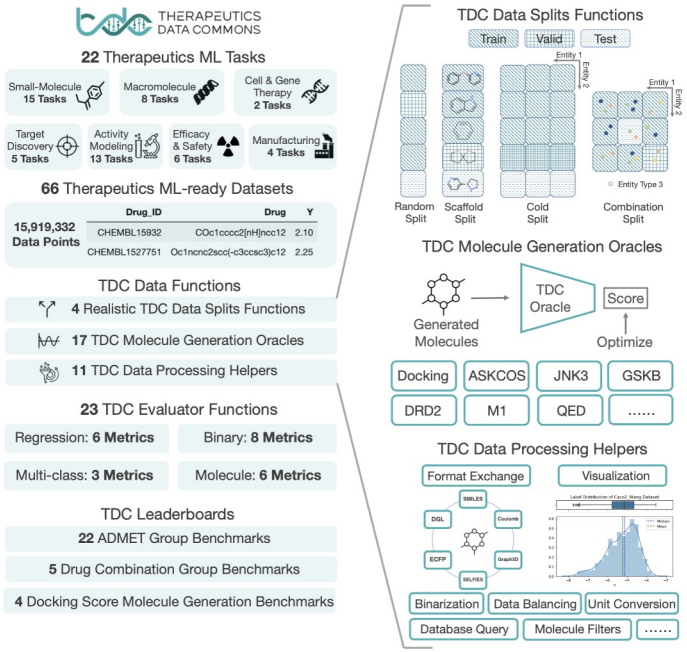
Algorithmic innovation and advances in therapeutic science

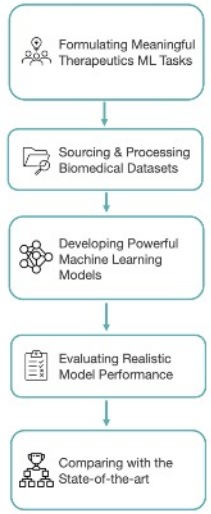
Global initiative to access and evaluate AI across therapeutic modalities and stages of drug discovery
270,000 active use cases of AI for drug design and therapeutic use prediction / 90,000 users worldwide
Ongoing collaborations with drug designers in **immuno-oncology**, **rare genetic diseases** and **neurodegenerative diseases**

HARVARD MEDICAL SCHOOL · MIT · Caltech · BROWN · PRINCETON UNIVERSITY · UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN · Georgia Institute of Technology · CHAN ZUCKERBERG INITIATIVE · Carnegie Mellon · Stanford University · NVIDIA · Microsoft Research · Cornell University · UNIVERSITY OF CAMBRIDGE

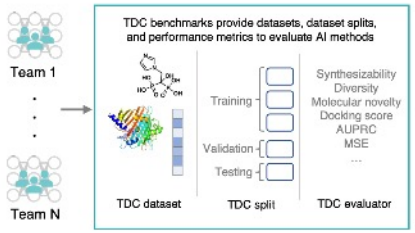
Therapeutics Data Commons: Machine Learning Datasets and Tasks for Therapeutics, NeurIPS '21; Artificial Intelligence Foundation for Therapeutic Science, Nature Chemical Biology '22; Ektefaie et al., *Nature Machine Intelligence* '23; McDermott et al., *Nature Machine Intelligence* '23; Li et al., *Nature Biomedical Engineering* '22

We develop Therapeutics Commons, a data and AI model hub across therapeutics modalities and stages of discovery

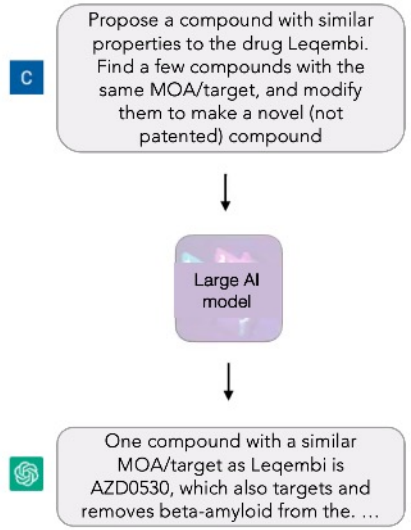




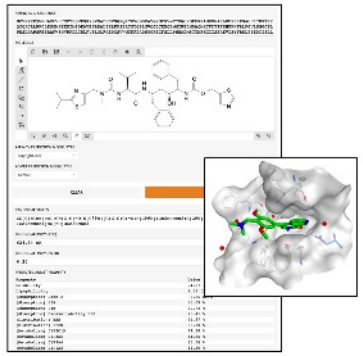
AI tasks and curated benchmarks



Leaderboards with best-in-class methods for therapeutic tasks



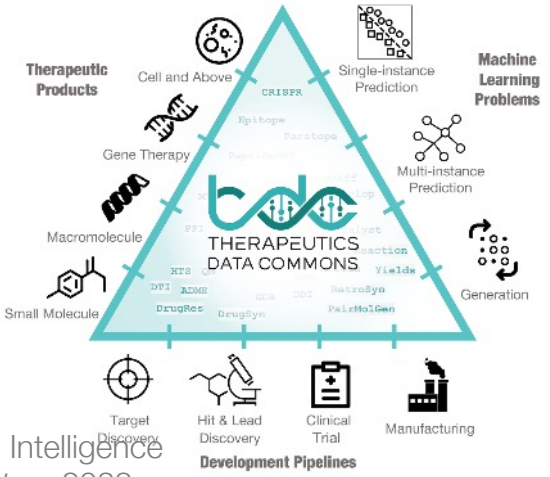
Pre-trained language models for key therapeutic tasks



Interactive human-AI design to close the feedback loop

- Molecular property prediction and optimization
- Binding of novel drugs to candidate therapeutic targets
- Molecule generation and AI-driven drug design
- Therapeutic use prediction, manufacturing, efficacy & safety
- AI models cover small molecules, proteins, peptides, miRNAs, and gene editing

270,000 active use cases of AI for drug design and therapeutic use prediction / 90,000 users worldwide
Partners validating AI models for **immuno-oncology**, **rare genetic diseases** and **neurodegenerative diseases**



Therapeutics Data Commons: Machine Learning Datasets and Tasks for Therapeutics, *NeurIPS* 2021; Artificial Intelligence Foundation for Therapeutic Science, *Nature Chemical Biology* 2022; Scientific discovery in the age of AI, *Nature* 2023

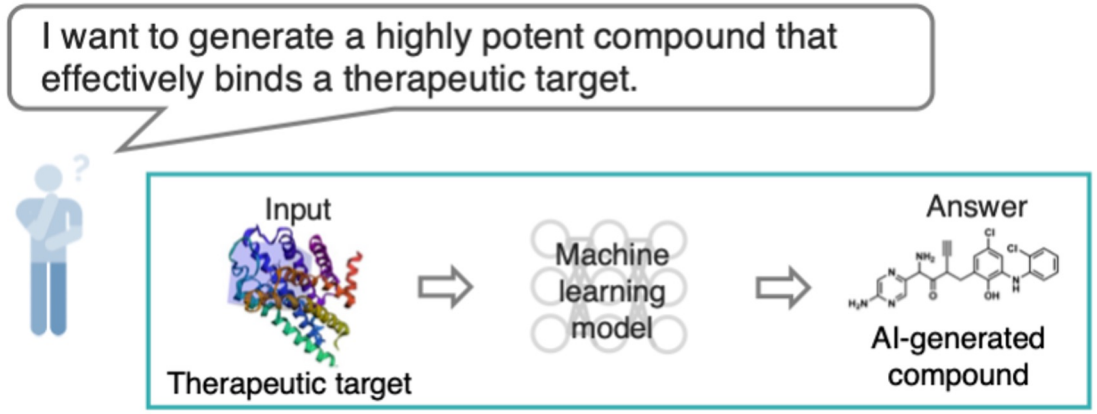
Capabilities of the Commons

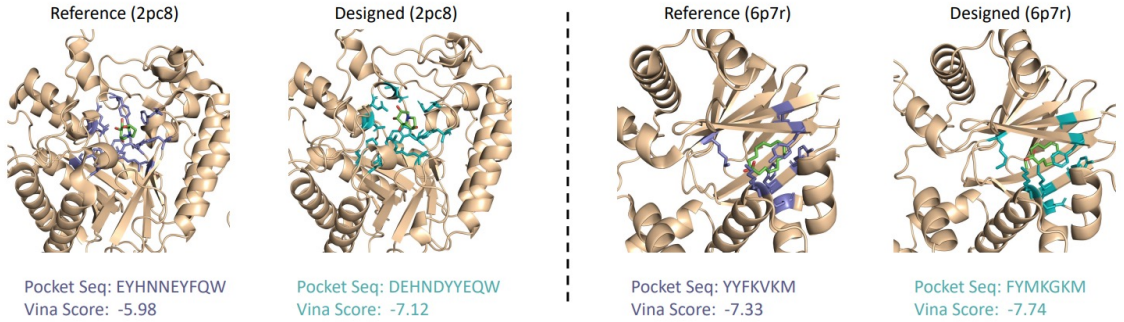
Predictive AI:

- Biological target nomination
- Target validation
- Molecular interaction screening

Generative AI:

- Mutation effect prediction
- Binding of ligands to targets
- Molecular design and optimization

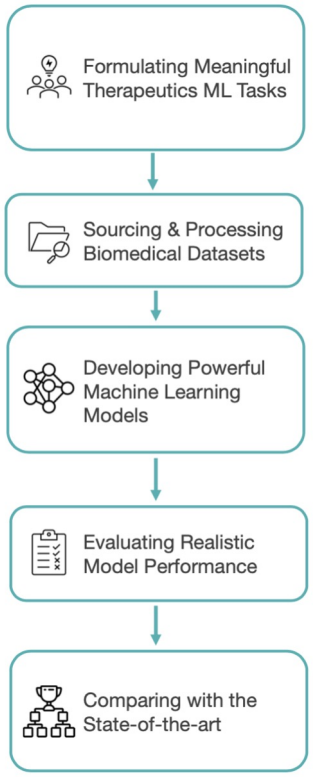




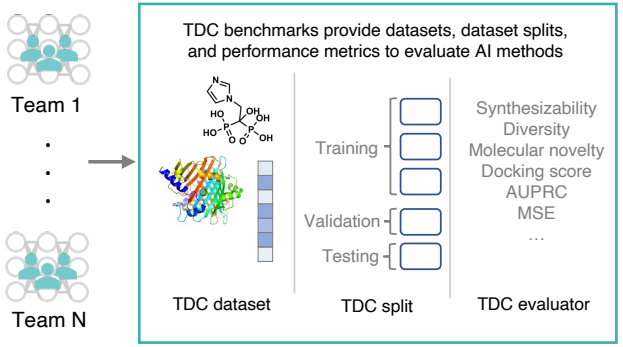
Model	CrossDocked			Binding MOAD		
	AAR (↑)	RMSD (↓)	Vina (↓)	AAR (↑)	RMSD (↓)	Vina (↓)
PocketOptimizer	27.89±14.9%	1.75±0.08	-6.905±2.39	28.78±11.3%	1.68±0.12	-7.829±2.41
DEPACT	22.58±8.48%	1.97±0.14	-6.670±2.13	26.12±8.97%	1.76±0.15	-7.526±2.05
HSRN	31.62±10.4%	2.15±0.17	-6.565±1.95	33.70±10.1%	1.83±0.18	-7.349±1.93
Diffusion	34.62±13.7%	1.68±0.12	-6.725±1.83	36.94±12.9%	1.47±0.09	-7.724±2.36
MEAN	35.46±8.15%	1.76±0.09	-6.891±1.86	37.16±14.7%	1.52±0.09	-7.651±1.97
FAIR	**40.17±12.6%**	**1.42±0.07**	**-7.022±1.75**	**43.75±15.2%**	**1.35±0.10**	**-7.978±1.91**

Therapeutics Data Commons: Machine Learning Datasets and Tasks for Therapeutics, *NeurIPS* 2022; Artificial Intelligence Foundation for Therapeutic Science, *Nature Chemical Biology* 2022; Full-Atom Protein Pocket Design via Iterative Refinement, *NeurIPS* 2023

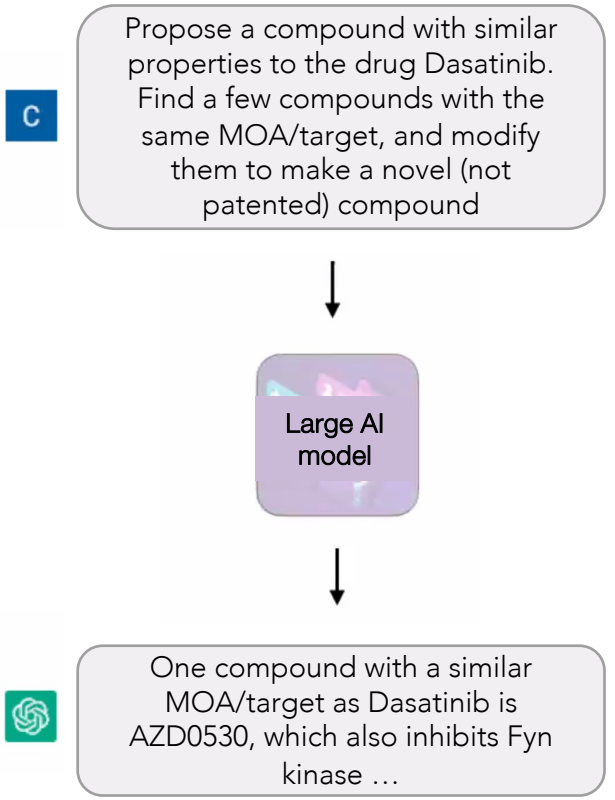
Therapeutics Commons is a hub of unified datasets, AI models, and benchmarks across therapeutic modalities



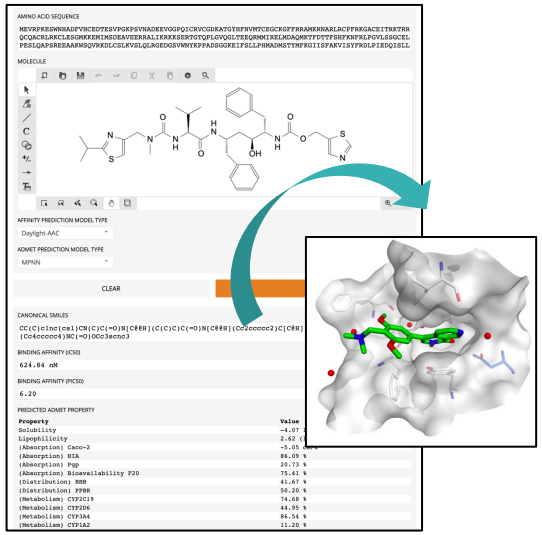
AI tasks and curated benchmarks



Leaderboards with best-in-class methods for therapeutic tasks



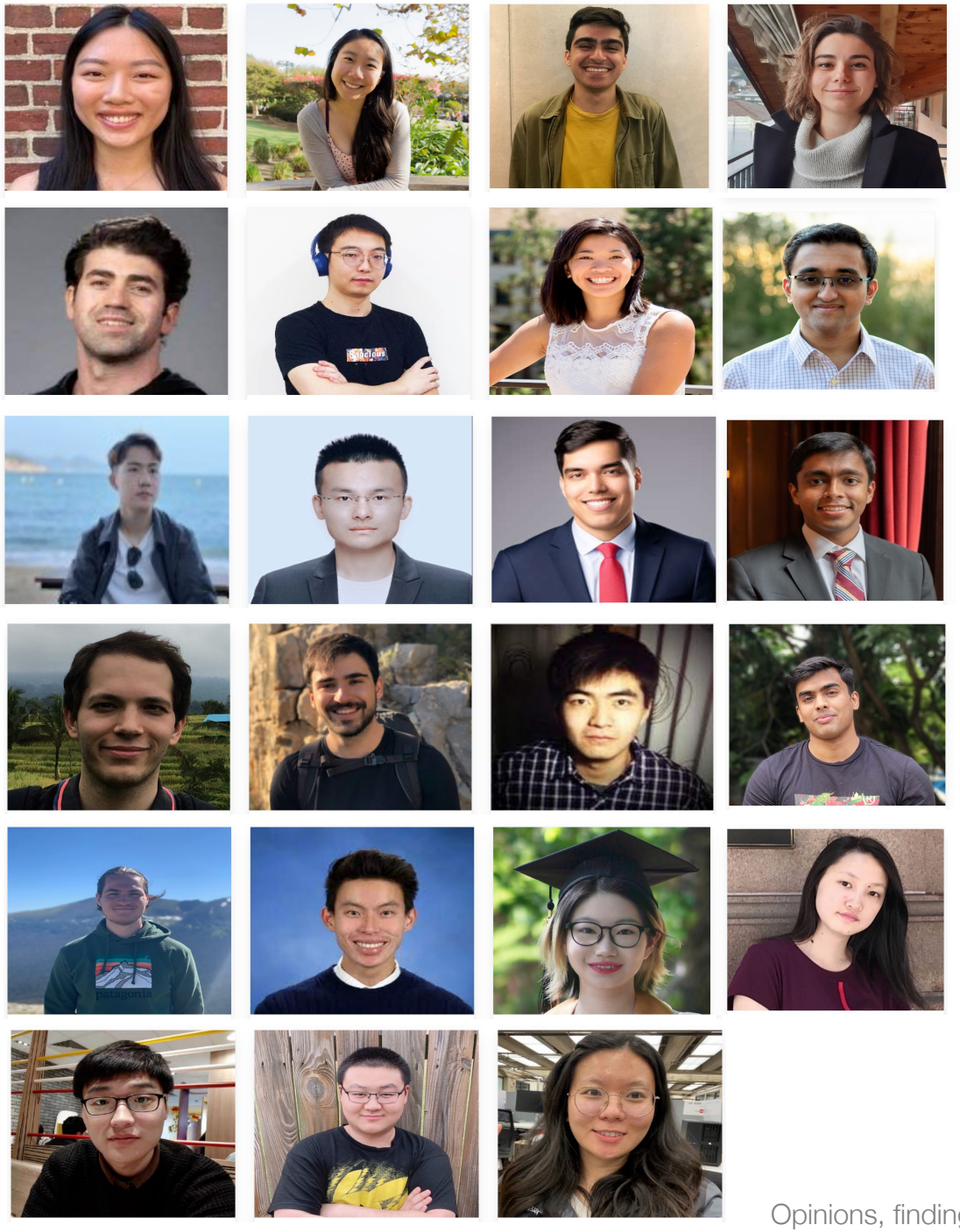
Pre-trained language models for key therapeutic tasks



Interactive human-AI design to close the feedback loop

Therapeutics Data Commons: Machine Learning Datasets and Tasks for Therapeutics, *NeurIPS* 2021
Artificial Intelligence Foundation for Therapeutic Science, *Nature Chemical Biology* 2022
Scientific discovery in the age of AI, *Nature* 2023

marinka@hms.harvard.edu



NSF · NIH · Department of the Air Force · Department of Veterans Affairs

CHAN ZUCKERBERG INITIATIVE · MILKEN INSTITUTE · MIT LINCOLN LABORATORY

HDSI | Harvard Data Science Initiative · amazon | science

Google Research · BAYER · AstraZeneca

sanofi · Pfizer · Roche · asap Aligning Science Across Parkinson's

AI models, datasets and papers
zitniklab.hms.harvard.edu

Therapeutics Commons
tdcommons.ai

Opinions, findings, conclusions or recommendations expressed here do not necessarily reflect the views of the funders.