Lecture1

The New Era of Digital Biology: Impact on Biological and Medical Research

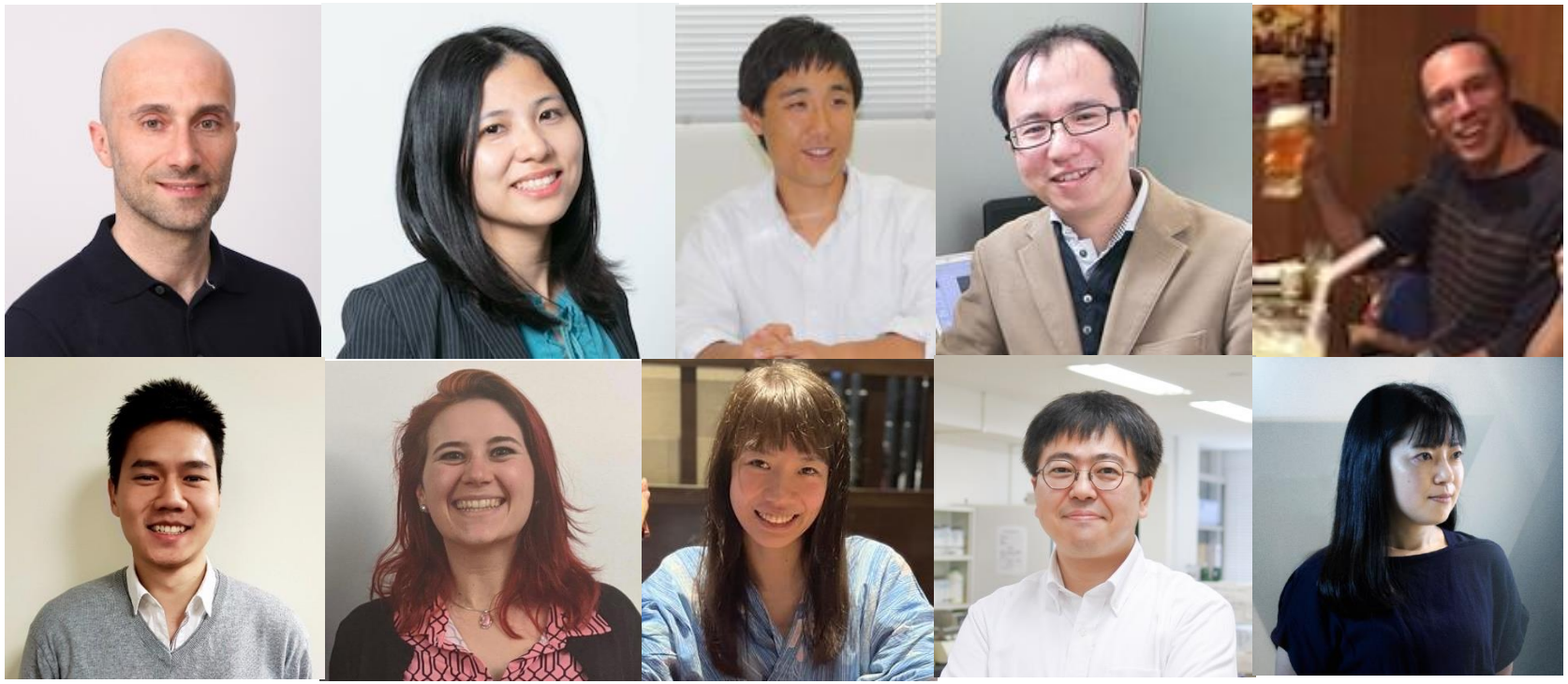
Prof. Jordan Ramilowski
Dr. Sakura Eri Maezono

Bioinformatics Laboratory
Yokohama City University

10-06-2025

FY2025 Current Trends in Bioinformatics Course (1)

Speakers from Academia and Industry



- Yokohama City University
- RIKEN Yokohama
- Okinawa Institute of Science and Technology
- KOTAI Biotech. Inc.
- AI Medical Service
- HumanomeLab

https://edu.med.yokohama-cu.ac.jp/courses/omnibus/FY2025/public/schedule.html

FY2025 Current Trends in Bioinformatics Course (2)

Lecture Schedule | 講義スケジュール

8 Online Lectures & Discussion

10/06/2025 – 12/15/2025 (Mondays)
Zoom @16:10 – 17:40 JST (login details are provided to registered participants)

	Date	Lecturer (click for details)	Lang	Title (click for Abstract)
1	10/6/2025	Jordan RAMILOWSKI & Sakura MAEZONO, PhD (YCU)	🇺🇸	The New Era of Digital Biology and Its Impact on Biological and Medical Research
2	10/20/2025	Takahiro SUZUKI, PhD (RIKEN)	🇯🇵	Introduction & Advances in the NGS and OMICS DATA Analysis
3	10/27/2025	Takeya KASUKAWA, PhD (RIKEN)	🇯🇵	Use of Public Resources for the Omics Studies / オミックス研究のための公共リソースの活用
4	11/10/2025	Charles PLESSY, PhD (OIST)	🇺🇸	Automation, containerisation and pipelinisation for bioinformatics: easier for you to run it, easier for others to reproduce it
5	11/17/2025	Dave TANG & Jo VILLA, PhD (KOTAI Biotech. Inc.)	🇺🇸	Single cell and Spatial Transcriptomics
6	12/1/2025	Shiori ARII, Ms (AI Medical Service)	🇯🇵	A computer aided supporting system with AI in endoscopic image
7	12/8/2025	Jun SESE, PhD (HumanomeLab)	🇯🇵	日常のデータから健康を再定義する：デジタルデバイスを活用したリアルワールドでの生体計測とデータ解析
8	12/15/2025	Sayuri SHIMIZU, PhD (YCU)	🇯🇵	Application of Machine Learning to Healthcare

Please note: Lecture titles and content may be revised

https://edu.med.yokohama-cu.ac.jp/courses/omnibus/FY2025/public/schedule.html

FY2025 Current Trends in Bioinformatics Course (3)

Graduate School of Medicine Course

- **Course Title:** バイオインフォマティクス特講 (Current Trends in Bioinformatics)
- **Credits:** 1 elective credit

Note: This course is offered under the same title for both the *Graduate School of Medical Life Science* and the *Graduate School of Nanobioscience*.

Health Data Science Course

- **Course Title:** バイオインフォマティクス (Bioinformatics)
- **Credits:** 1 elective credit

Attending in-full at least 6 lectures (60-69%) is required:

- 8 lectures (69%)
- 7 lectures (64%)
- 6 lectures (60%)
- *5 lectures or less (FAILED)*

Zoom Attendance: YCU student ID must be used as a part of *Zoom Name* to verify attendance

Lecture Report (0–10%) is required:

- 1 page on one selected lecture (any lecture you like)
- report can be written in either Japanese or English
- Deadline:

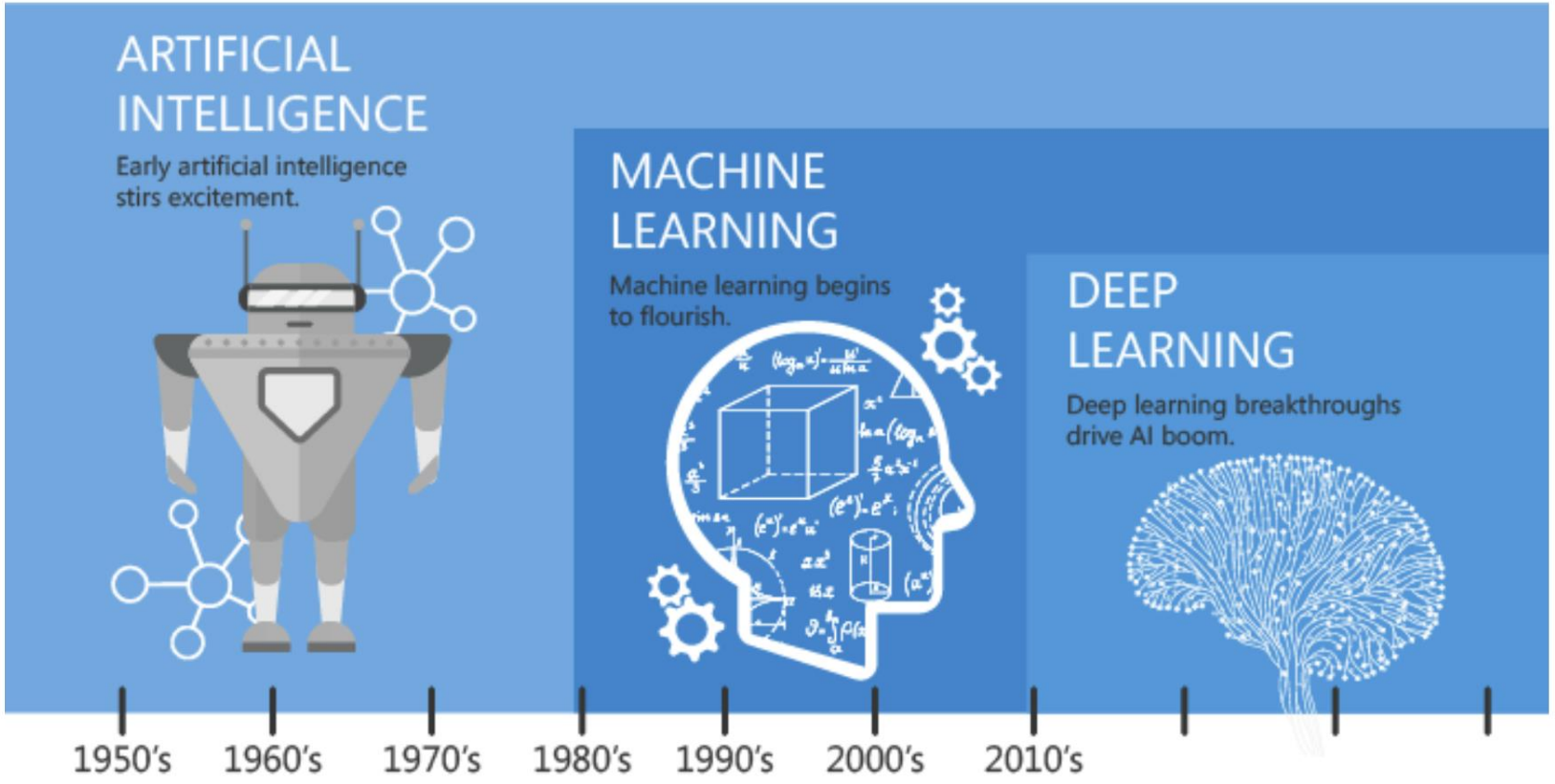
Active participation (0–20%) is encouraged:

- asking questions and participating in class discussion
- filling out a course questionnaire to provide the course feedback and suggestions

https://edu.med.yokohama-cu.ac.jp/courses/omnibus/FY2025/

Terminology

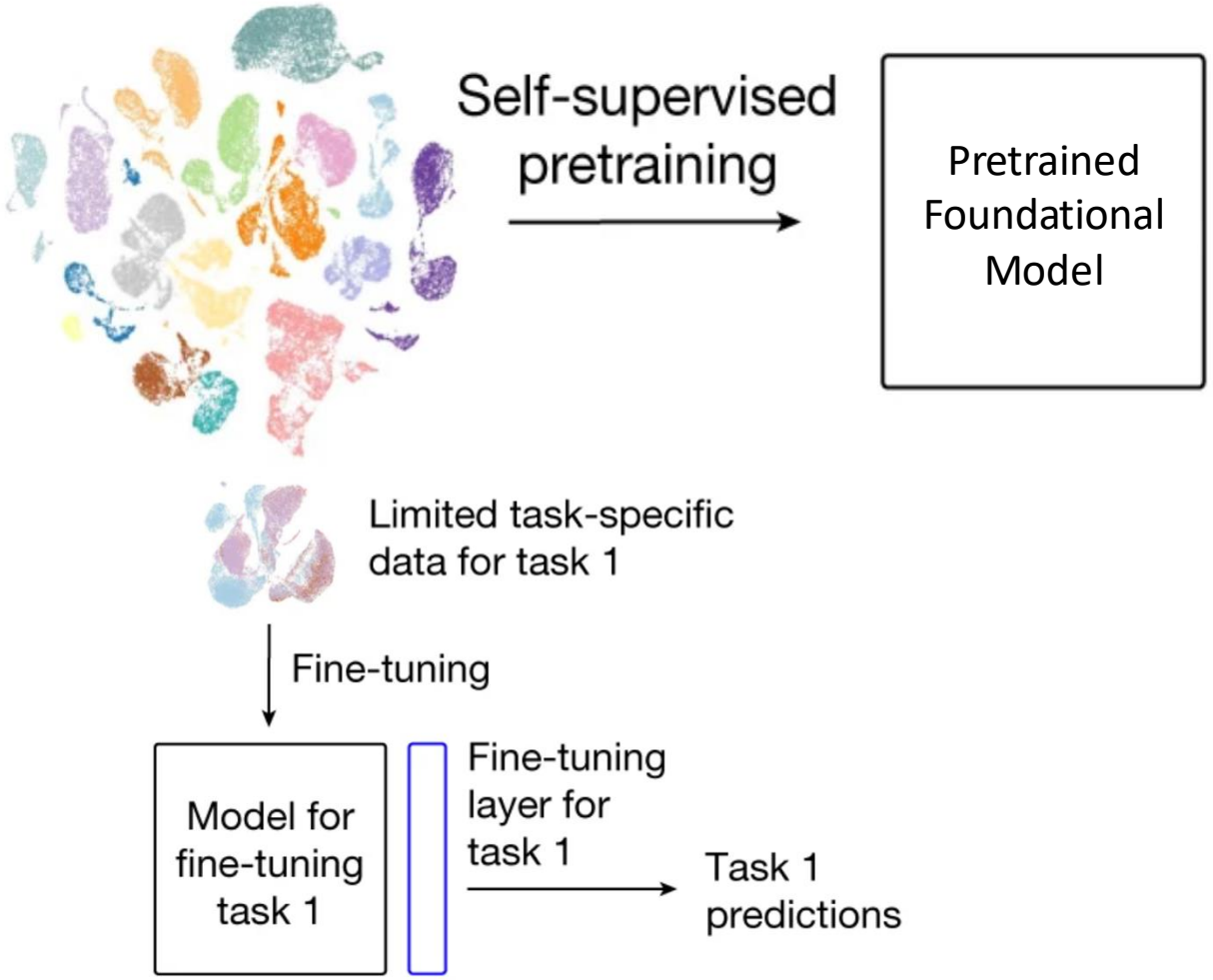
- Artificial Intelligence (AI) contains Machine Learning (ML) and Deep Learning (DL).
- When we hear **AI** these days, it is most likely about **Deep Learning**.



Power of AI: Foundational Models

If every time you research a problem you have a lot of relevant data, then you can solve that problem. **But what if you don't?**

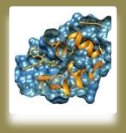
- **Foundational models are big/smart models trained on large datasets to learn broad patterns.**
- Serve as a base (foundation) for many specialized tasks
- Can be adapted, fine-tuned, or prompted for new problems
- 💬 "Train once — use everywhere."



AI in Biology: A New Frontier

The foundational models are starting to changed how we do research in biology.

DNA/RNA Predictions: *AI extracts patterns from massive omics datasets, e.g.* ***Geneformer****, scGPT, DeepSEA/Enformer,* ***AlphaGenome***

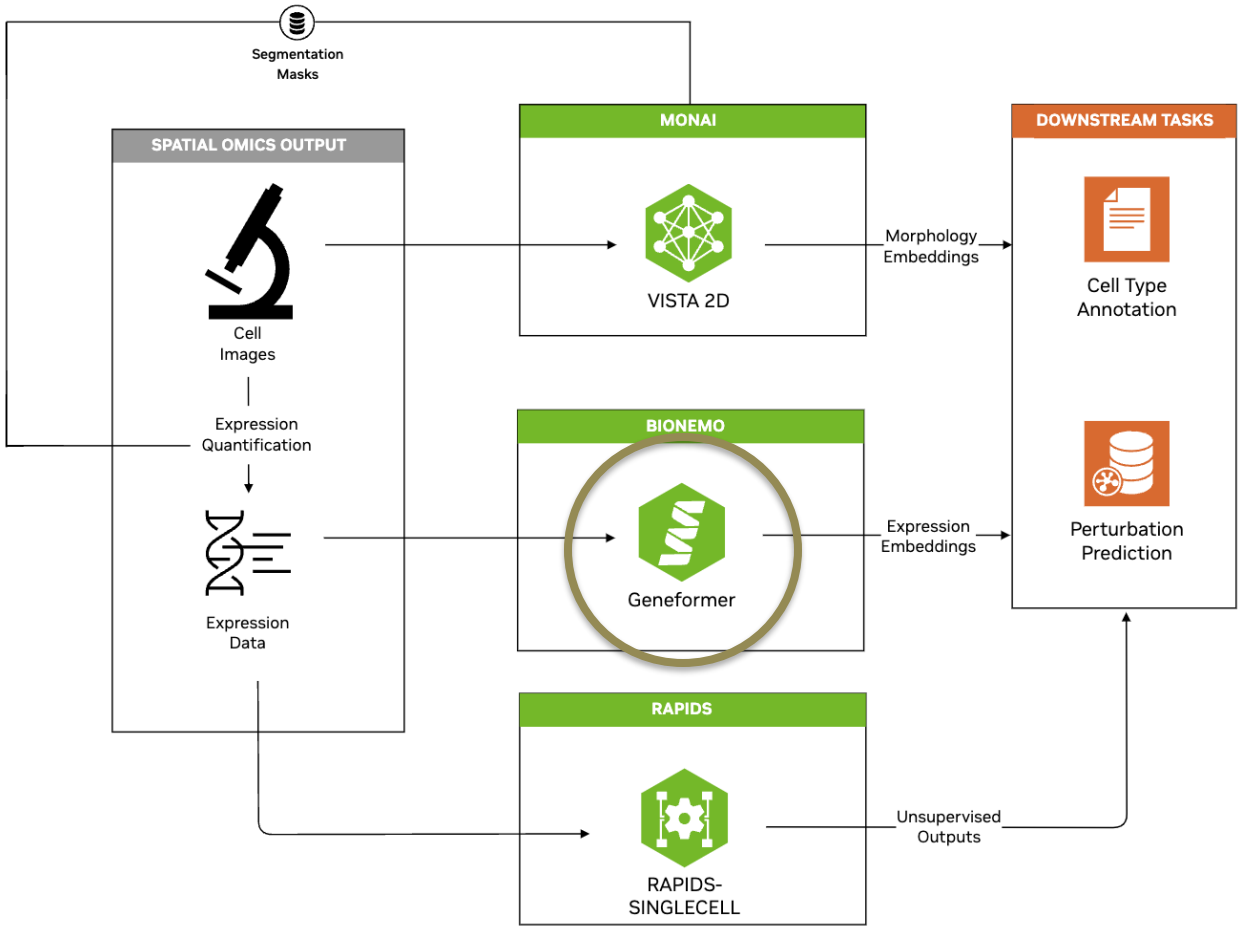


Structural Predictions: *AI predicts and simulates biology at molecular scale, e.g.* ***AlphaFold,*** *RoseTTAFold*

Large Language Models (LLMs): *AI processes and generates human language to interpret, summarize, and retrieve biomedical information, e.g.* ***Perplexity***

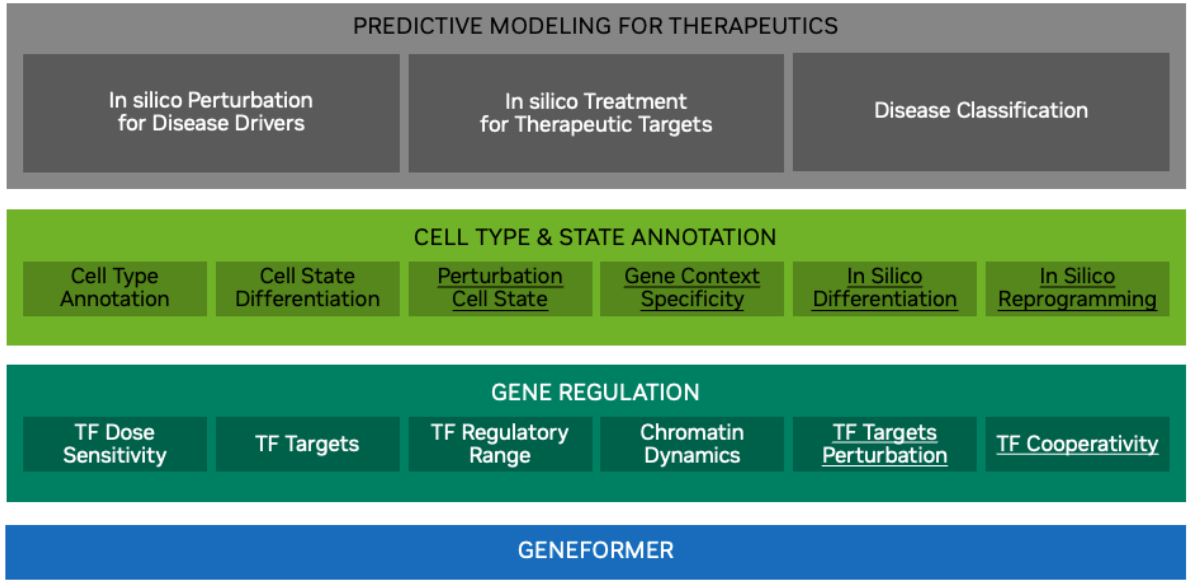
Geneformer: A foundation AI model for disease modeling

DNA/RNA Predictions



Geneformer is **open-source** and available for research;

Use cases and Zero-shot learning capabilities are shown in the figure below:



Zero-shot learning allows Geneformer to **predict data classes it hasn't seen or been explicitly trained for**

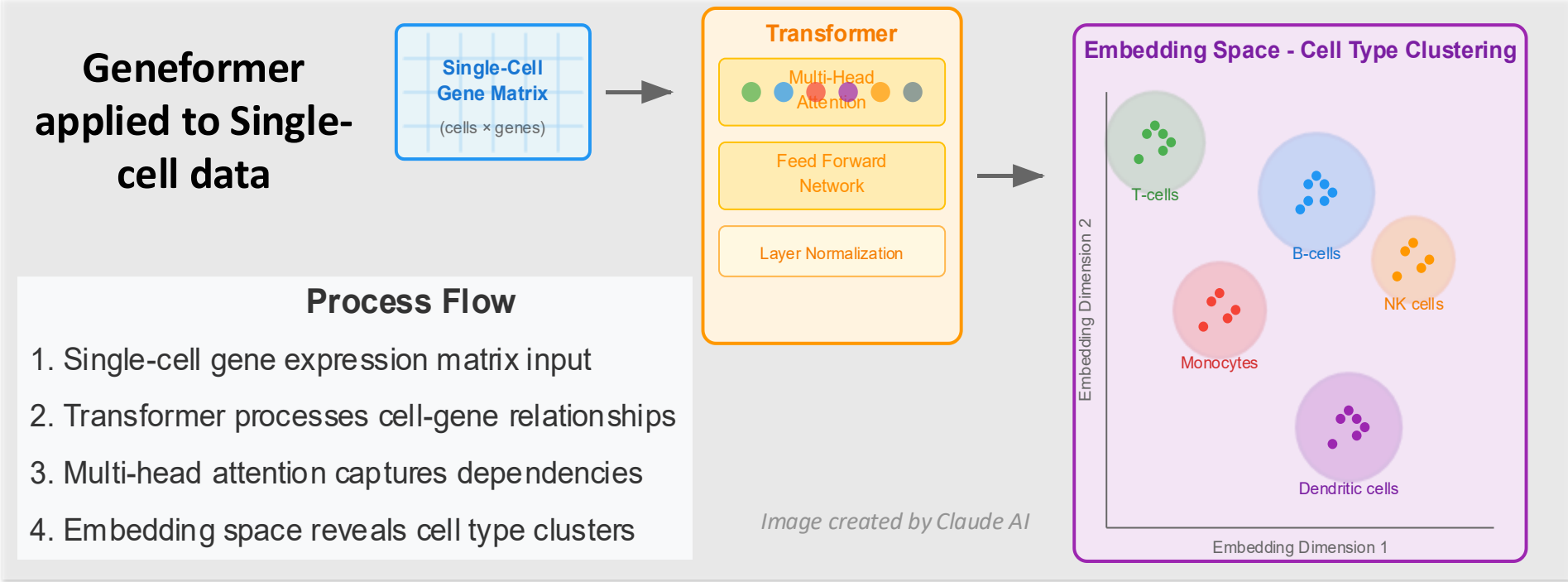
Images from NVIDIA DEVELOPER from an article written by Tretina, 2024

Related link: ***Geneformer***

Geneformer: Unlock Gene Networks Using Limited Data

DNA/RNA Predictions

- Trained on **30+ million single-cell transcriptomes**, Geneformer learns the “grammar” of gene expression.
- Instead of simple clustering, **it builds contextual embeddings** where relationships among genes, pathways, and cell states are encoded.
- **Enables transfer learning**: **fine-tune on rare disease or small datasets** and still get biologically meaningful insights.

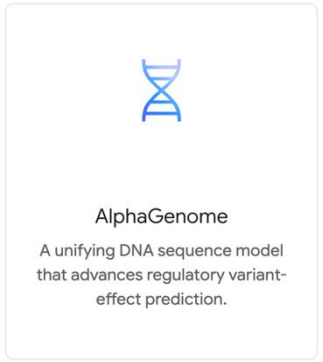


Related link: [Geneformer]

AlphaGenome: Deep Learning for Genomics

Structural Predictions

- Uses **transformer-based neural networks** to integrate DNA, RNA, epigenomics, and proteomics data
- Learns **regulatory codes and interactions** across different omics layers



Long sequence-context at high resolution

Applications:

- **Disease Discovery:** Pinpoints the precise cause of diseases (especially rare ones) by predicting the functional impact of **genetic disruptions**, helping identify new drug targets.
- **Synthetic Biology:** Enables the **design of synthetic DNA** with targeted regulatory functions for cell-specific gene activation.
- **Genome Mapping:** Accelerates basic research by mapping the **functional elements** of the genome and defining their cell-type-specific regulatory roles.

Video from Google DeepMind

Related link: ***AlphaGenome***

AlphaFold: Highly accurate protein structure prediction

Structural Predictions

How did it solved the 50-year-old "protein folding problem":

- **Revealed millions of intricate 3D protein structures;** helps scientists understand how life's molecules interact
- Predicts structures with accuracy comparable to crystallography for most proteins
- Beyond folding -- informs binding interfaces, mutational effects, and drug targetability
- **Accelerates research**: thousands of groups now rely on AlphaFold predictions instead of waiting years for experimental resolution – *"What took us months and years to do, AlphaFold was able to do in a weekend"*

*Over **2 million researchers** in over **190 countries** are using the AlphaFold Protein Structure Database to inform their research.*

CASE STUDY
Solving one of biology's greatest puzzles
Learn more

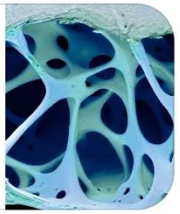
CASE STUDY
Breaking down plastic pollution
Learn more

CASE STUDY
Designing more effective drugs
Learn more

CASE STUDY
Increasing honeybees' chances of survival
Learn more

CASE STUDY
Searching for life-changing medicines
Learn more

CASE STUDY
Tracing proteins back to origins of life
Learn more



21%
Of papers citing AlphaFold are related to the study of disease

Up to 1 billion
Research years saved by AlphaFold database structures

Over 600,000
Users in low and middle income countries

CASE STUDY
Creating a protein structure database
Learn more

Images from Google DeepMind

*Related link: **AlphaFold***

Prediction Accuracy : AlphaFold vs Experimental

Structural Predictions

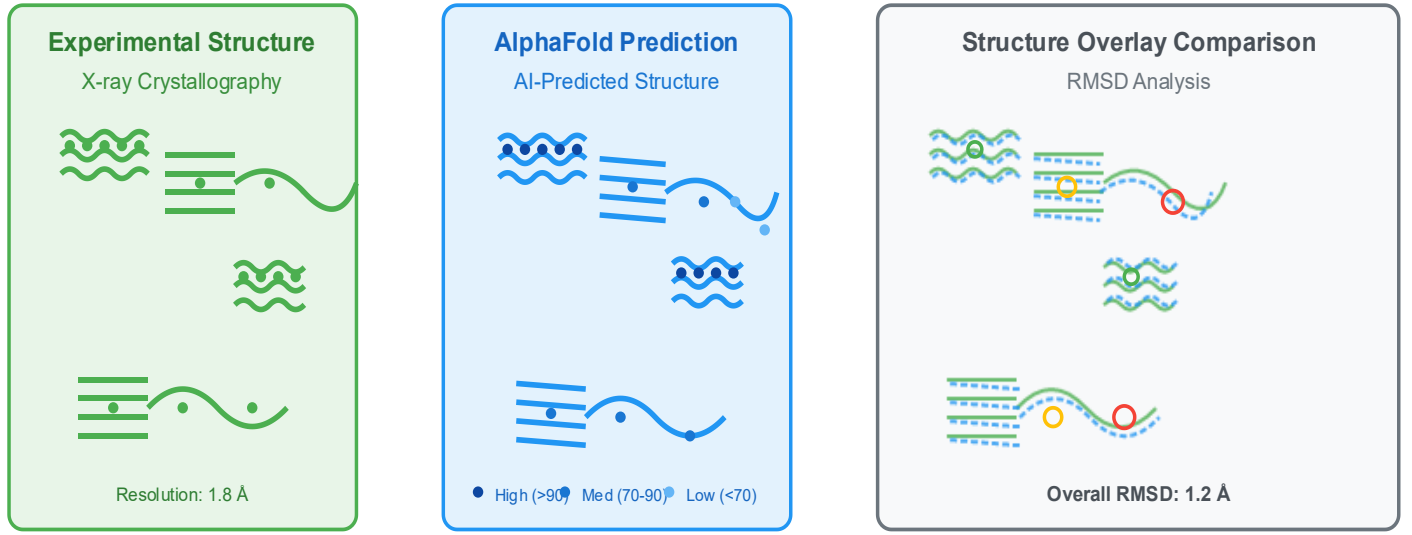
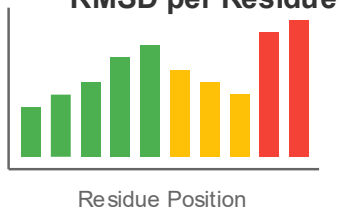


Image created by Claude AI

Structural Comparison Metrics

RMSD per Residue (Å)



AlphaFold Confidence vs AccuracySummary Statistics

- Overall RMSD: 1.2 Å
- Core regions RMSD: 0.8 Å
- Loop regions RMSD: 2.1 Å
- Structural similarity: 85%
- High confidence regions: 75%
- Correlation (conf. vs acc.): 0.78
- Secondary structure match: 92%
- Overall quality: Excellent

Structure Comparison Legend

Experimental | AlphaFold | Small diff. | Med diff. | Large diff. | High conf. | Low conf.

*Related link: **AlphaFold***

Perplexity: AI-powered research assistant

Large Language Models

Perplexity AI leverages LLMs:

- Synthesizes real-time information from diverse sources to provide comprehensive, evidence-based answers.
- **Deep Research Mode**: Performs multiple searches, analyzes hundreds of sources, and delivers structured reports, saving hours of manual research

Applications:

- **Literature Summarization**: Quickly condenses complex biomedical literature into digestible summaries.
- **Clinical Decision Support**: Assists in diagnosing and predicting treatment outcomes by analyzing patient data and medical literature.
- **Drug Discovery & Development**: Facilitates the identification of potential drug candidates by analyzing vast datasets of biomedical information.
- **Evidence-Based Medicine**: Supports clinicians in making informed decisions by providing up-to-date, evidence-backed information

Screenshot from Perplexity AI

Current Research
現在の研究

Molecular to Phenotype Connection

- DNA provides the blueprint, but RNA and protein expression define cell behavior in health and disease
- Understanding molecular phenotype explains why cells behave differently despite identical DNA and **how deviations cause disease**
- Apart from protein-coding genes, there exists **noncoding RNA (e.g. lncRNA)** which functions are yet to be established

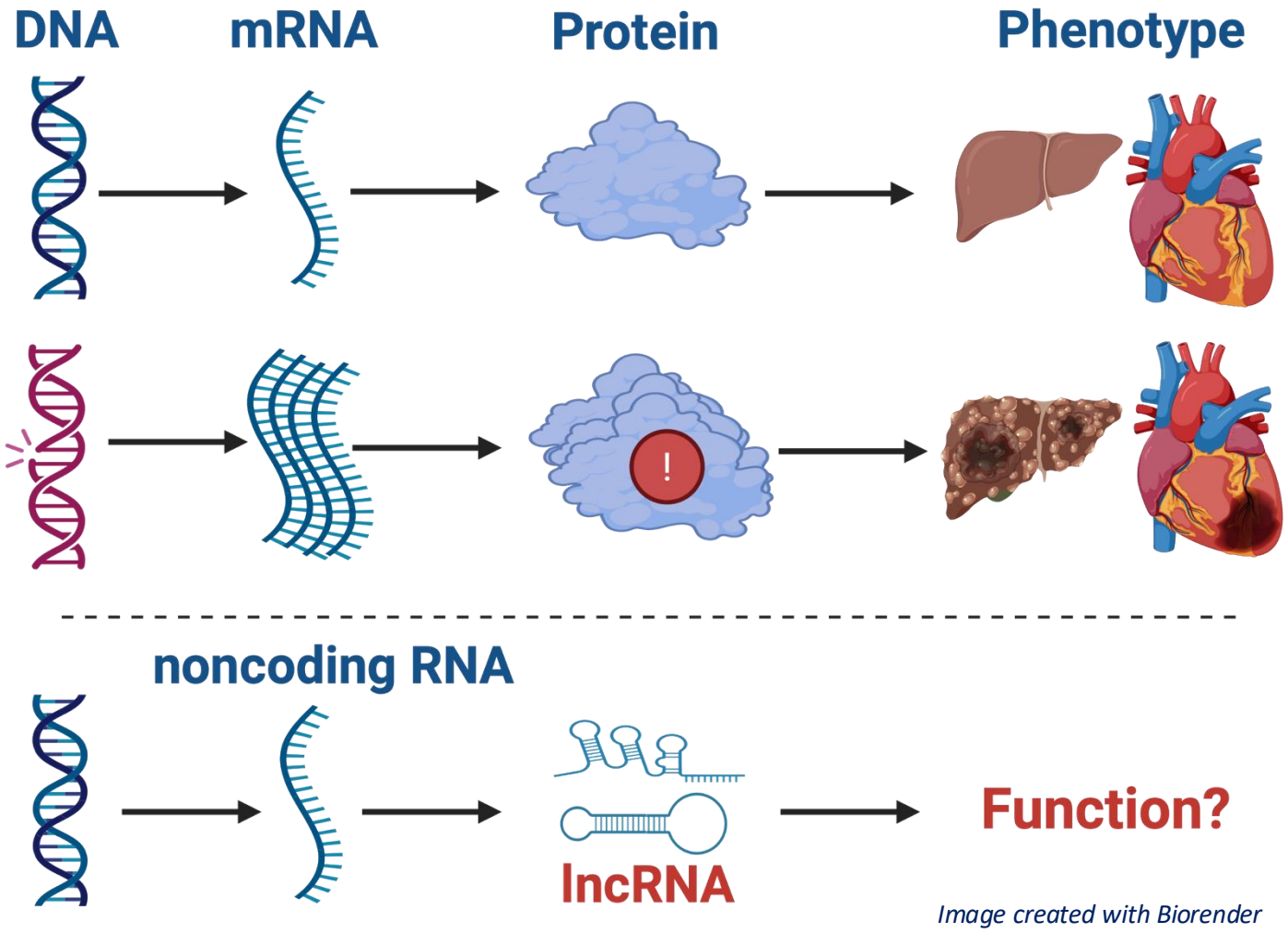
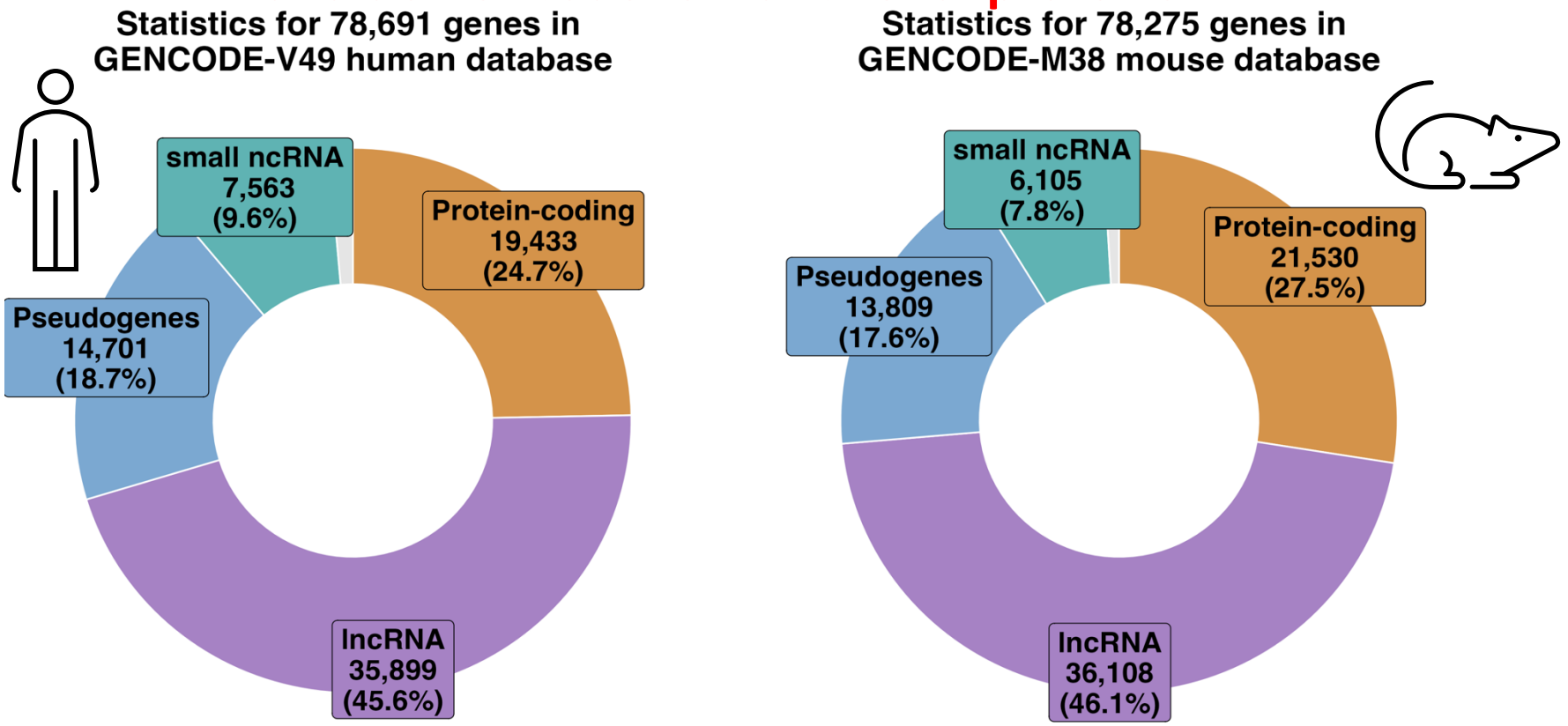


Image created with Biorender

Long noncoding RNAs

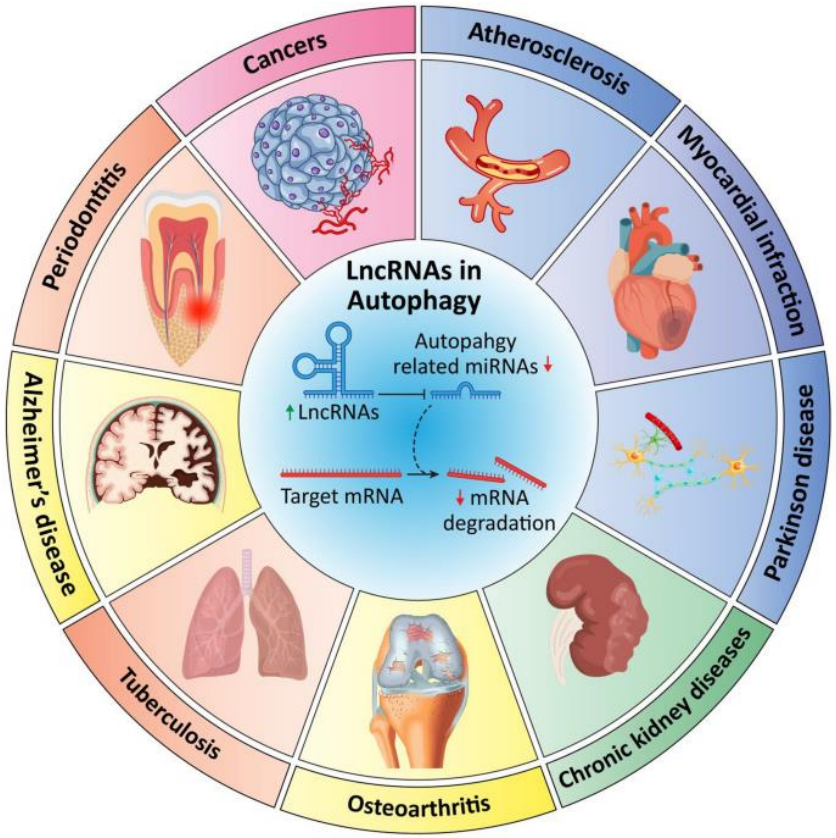
lncRNAs are defined as RNA transcripts longer than 200 nucleotides and do not code for functional proteins.



Long non-coding RNAs (lncRNAs) are highly abundant in mammalian genomes.
Understanding the function of mammalian cells requires investigating lncRNAs.

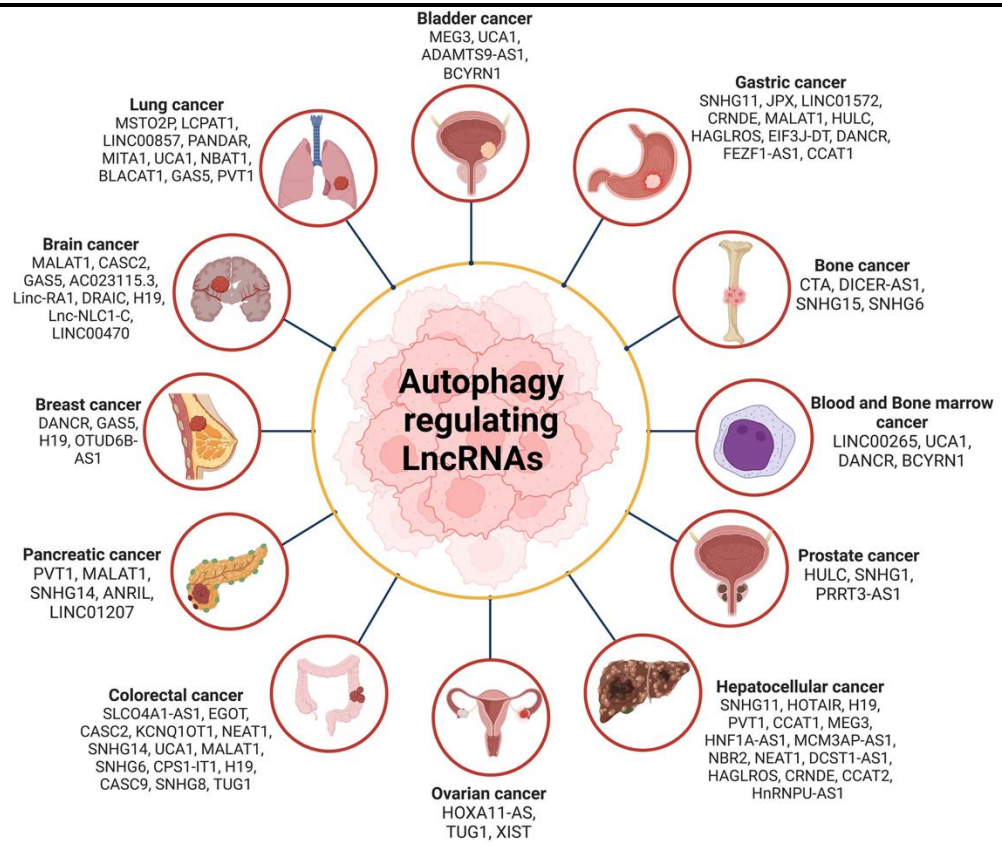
lncRNAは細胞表現型や疾患表現型を制御

lncRNAs in various diseases



Kumar *et al.*, *Journal of Biomedical Science* (2024)

lncRNAs in various diseases

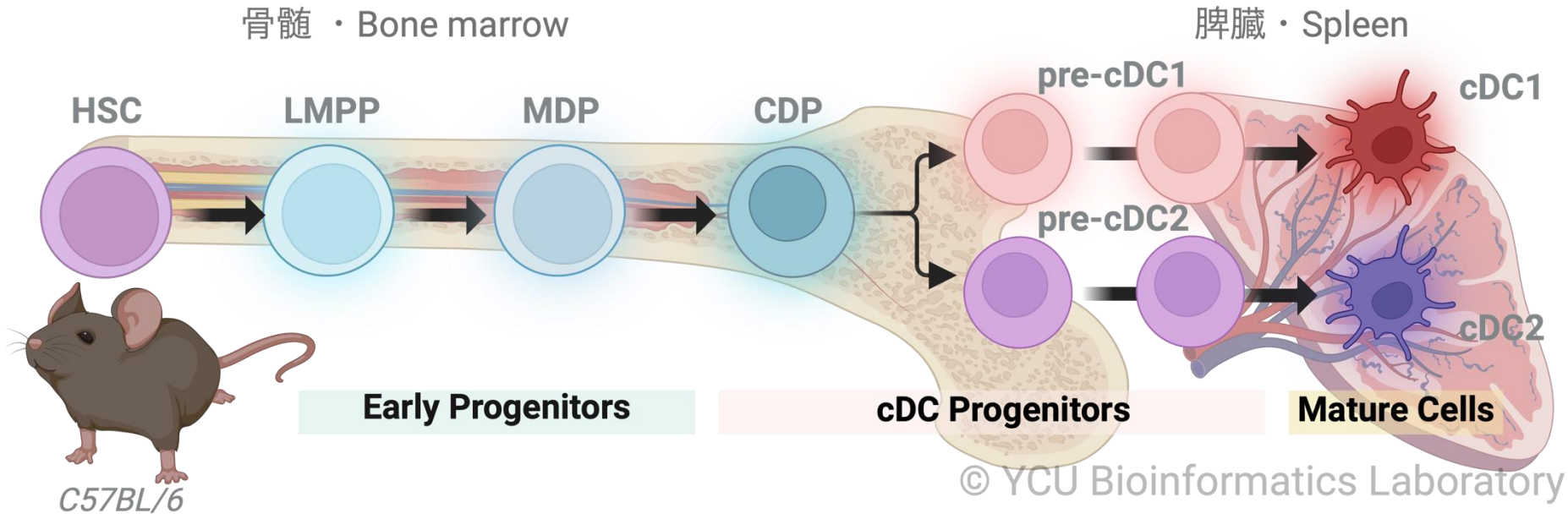


Despite selected example, functions are not known for most lncRNA genes.

Our interest: lncRNAs in Dendritic Cell Function & Differentiation

DCs: are professional Antigen Presenting Cells (APCs) priming regulatory and cytotoxic T-cells to orchestrate a variety of adaptive immune responses upon infection and in cancers.

DCs differentiate mainly in bone marrow and mature in spleen in an IRF8-dependent manner.

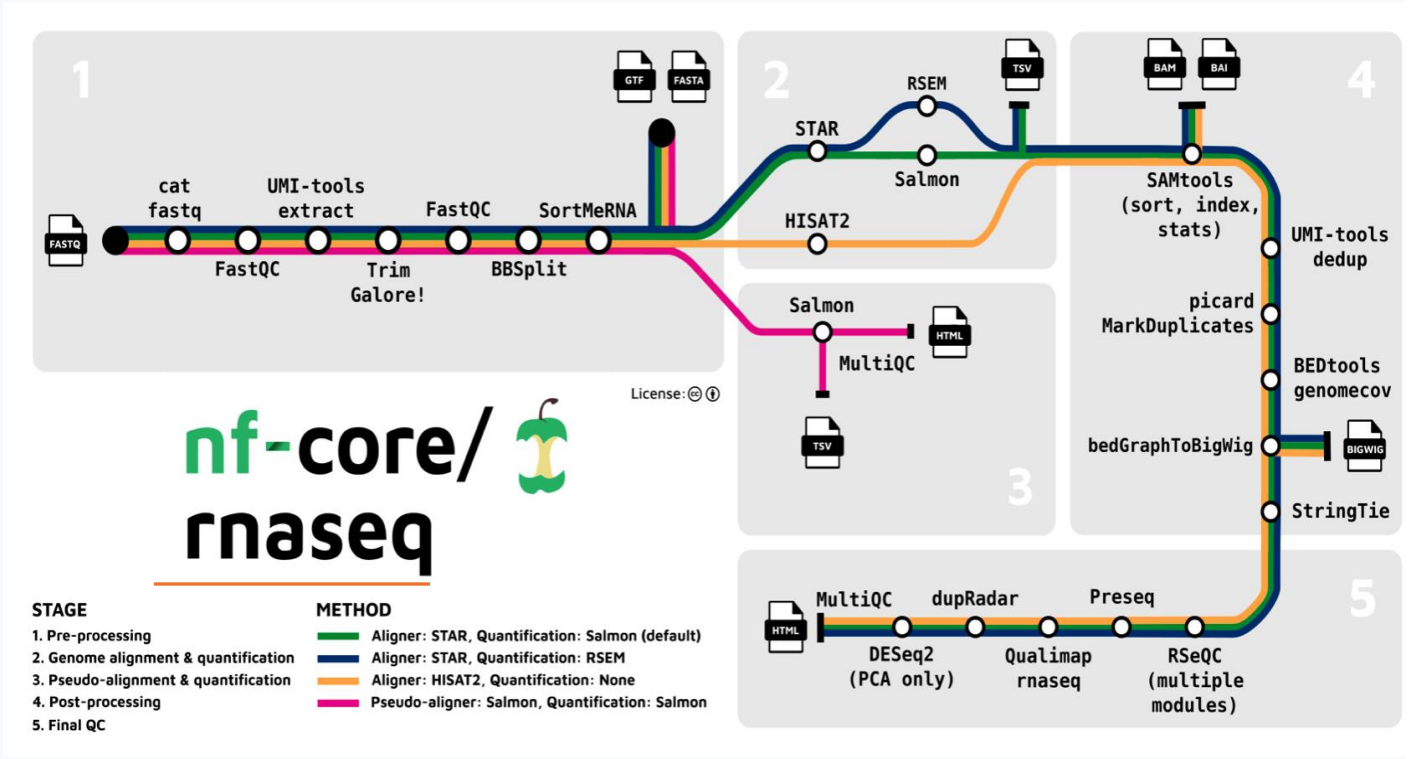


1. **LMPP**: Lymphoid-Myeloid Primed Progenitors
2. **MDP:** Monocyte Dendritic Cell Progenitors
3. **CDP:** Common Dendritic Cell Progenitors
4. **pre-cDC1/pre-cDC2:** pre-Dendritic Cells
5. **cDC1/cDC2:** classical Dendritic Cells

Creating Atlas of lncRNAs in DC1 lineage

To construct an atlas of lncRNAs in dendritic cells (DCs), we performed quality control, alignment, and quantification of public RNA-Seq data using the nf-core/rnaseq pipeline, a community-curated, standardized, and reproducible workflow using Nextflow, to construct a comprehensive atlas of lncRNAs in selected myeloid cells.

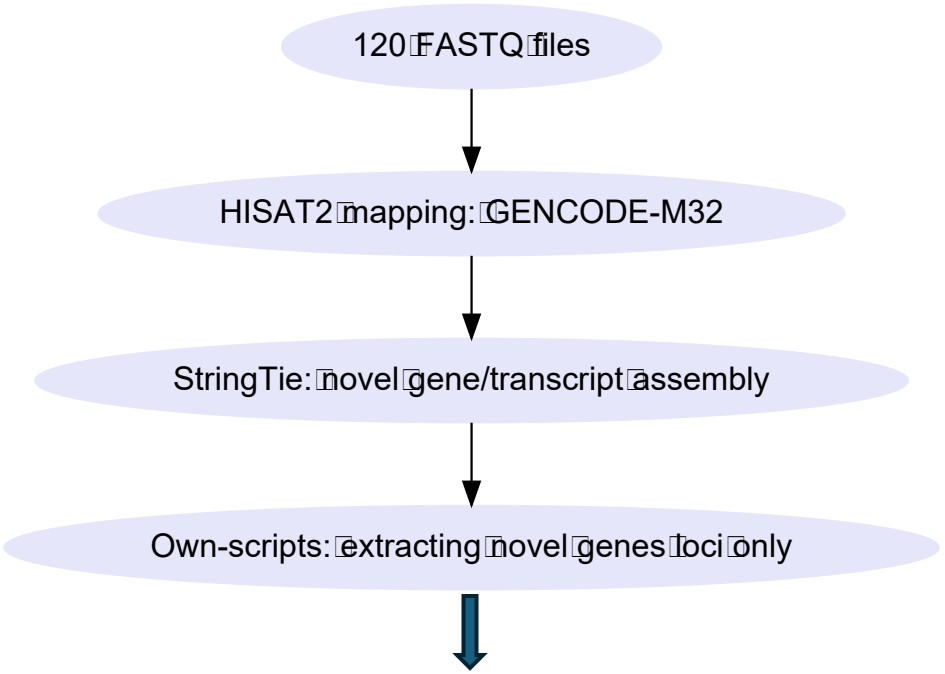
Pipeline summary



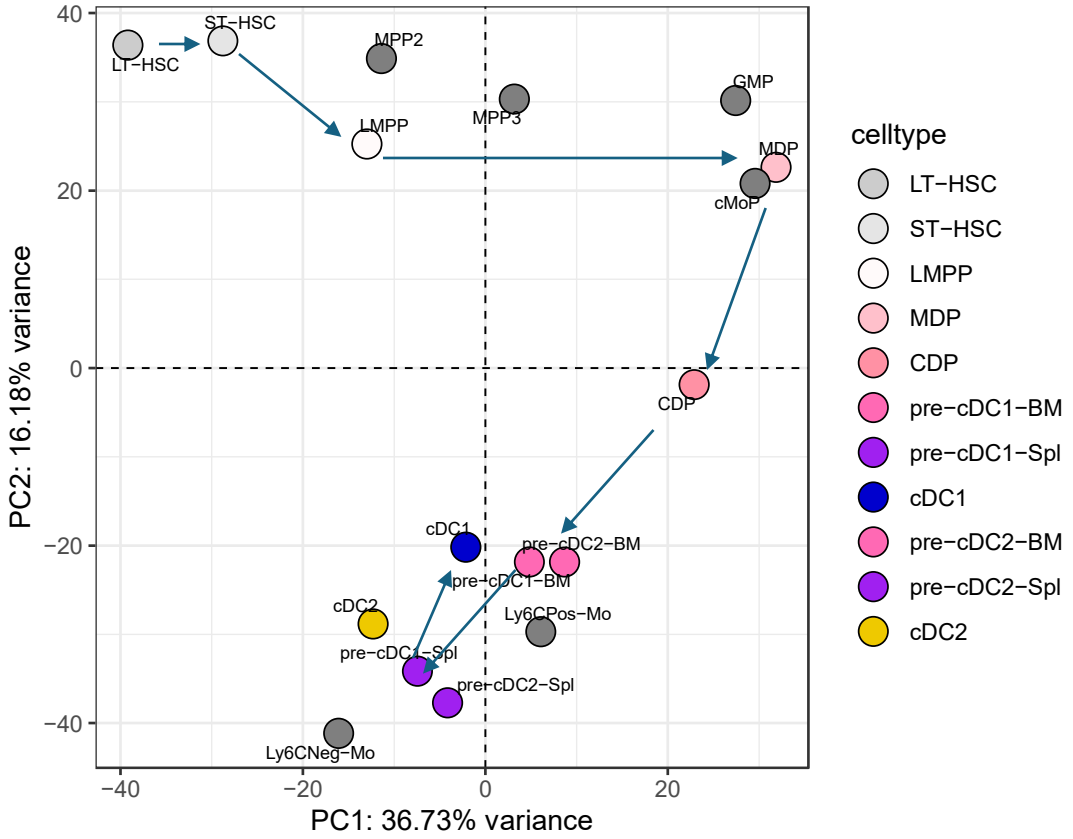
https://nf-co.re

Identification of lncRNAs specific to DC differentiation

De-novo assembly was performed on in vivo RNA-seq data using nf-core and a proprietary pipeline. Approximately 6.000 novel lncRNAs were identified through time-course analysis of cDC differentiation.



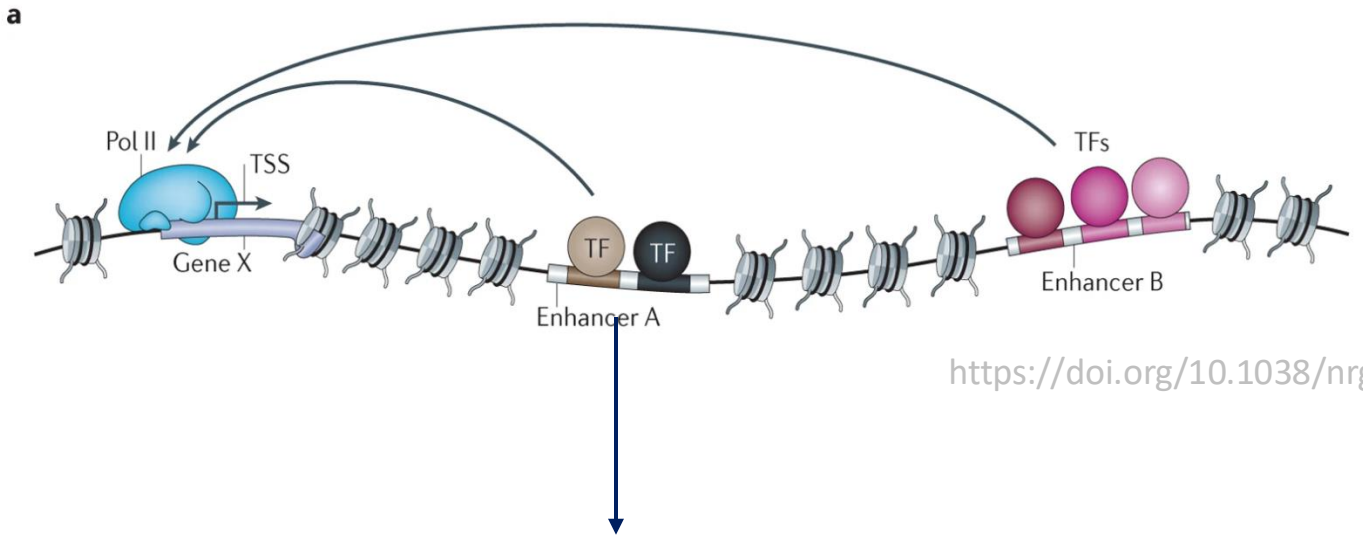
10,000 known and novel lncRNAs identified capture DC1 lineage



Transcriptional Regulation

Cis-regulatory elements (CREs):

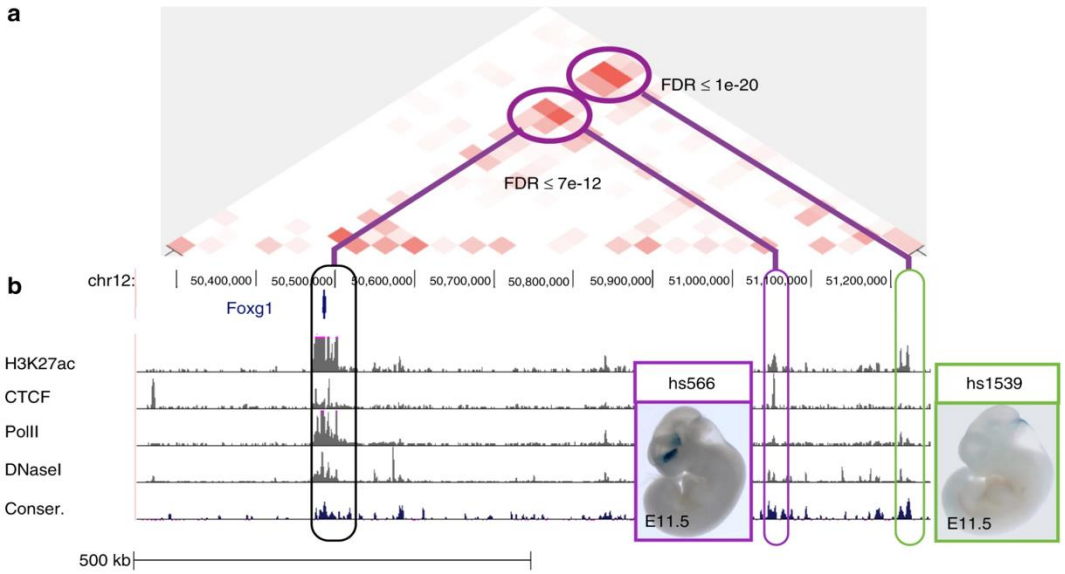
- Promoters: ability to recruit RNA Polymerase II and initiate transcription locally.
- Enhancers: ability to enhance recruitment and/or transcription at distal location.



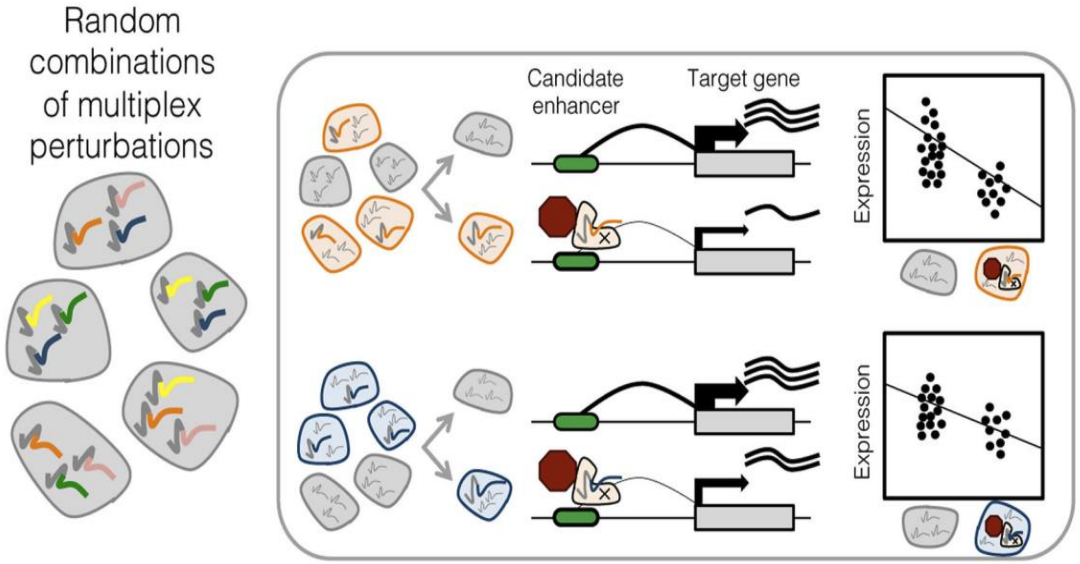
https://doi.org/10.1038/nrg3682

Some of these enhancers transcribe
enhancer-derived long noncoding RNAs (e-lncRNAs).

Enhancer discovery with experimental techniques



https://doi.org/10.1038/s41467-017-02386-3



doi:10.1016/j.cell.2018.11.029

3C-based techniques:

- False positives due to low resolution
- Proximity does not guarantee regulation

CRISPR-based approaches:

- Scaling up to genome-wide perturbation
- Vulnerable to intrinsic enhancer redundancy
- Off-target effects

Both the CRISPR-based and 3C-based methods are laborious and expensive.

Enhancer Discovery – Computational Tools

- A range of computational methods have been developed to predict enhancer targets by incorporating a variety of genomic and epigenomic features.
- **ABC method allows to predict enhancer with high accuracy, but**
  - it requires multiple datasets (ATAC-seq, H3K27ac, RNA-seq & Hi-C)
  - data quality is very important
  - time-consuming to run

Class	Method	Year	Data	Algorithm	Availability
Multiple regression-based methods	JEME*	2018	ChIP-seq, DNase-seq, CAGE	LASSO and RF	https://github.com/yiplabcuhk/JEME
	FOCS	2018	CAGE, GRO-seq, DNase-seq	OLS and Elastic Net	https://github.com/Shamir-Lab/FOCS
	DGTAC*	2021	ATAC-seq, RNA-seq	Elastic Net and RF	https://github.com/khuranalab/DGTAC
	CT-FOCS	2022	CAGE, DNase-seq	Linear Mixed Effect models	https://github.com/Shamir-Lab/CT-FOCS
Signal decomposition methods	EpiTensor	2016	ChIP-seq, RNA-seq	Tensor Decomposition	http://wanglab.ucsd.edu/star/EpiTensor/
	Rennie at al*	2018	CAGE, RNA-seq	Transcriptional Decomposition via approximate Bayesian modelling and RF	https://github.com/anderssonlab/transcriptional_decomposition
Synteny based methods	PEGASUS	2020	Synteny conservation	Score based on synteny conservation	https://ftp.bio.ens.psl.eu/pub/dyogen/PEGASUS/
Attention based deep learning models	Enformer	2021	DNA sequence	CNN and Transformer	https://github.com/deepmind/deepmind-research/tree/master/enformer
	CREaTor	2023	DNA sequence, ChIP-seq, DNase-seq	CNN and Transformer	https://github.com/DLS5-Omics/CREaTor
Other unsupervised methods	GeneHancer	2017	Distance, eRNAs, TF co-expression, eQTLs, Hi-C	Weighted sum of transformed data	http://www.genecards.org/
	ABC	2019	Distance, DNase-seq, ATAC-seq, ChIP-seq, Hi-C	Score based on enhancer activity and contact frequency	https://github.com/broadinstitute/ABC-Enhancer-Gene-Prediction
	gABC	2022	Distance, DNase-seq, ATAC-seq, ChIP-seq, Hi-C	Score based on enhancer activity and contact frequency, considering multiple promoter and gene-specific activity	https://github.com/schulzlab/STARE

DC1 linage widely transcribed e-lncRNAs

Using ABC (Activity By Contact) method we discover all genomic enhancer and their target mRNAs in DC1 lineage. Adding expression levels with found e-lncRNAs.

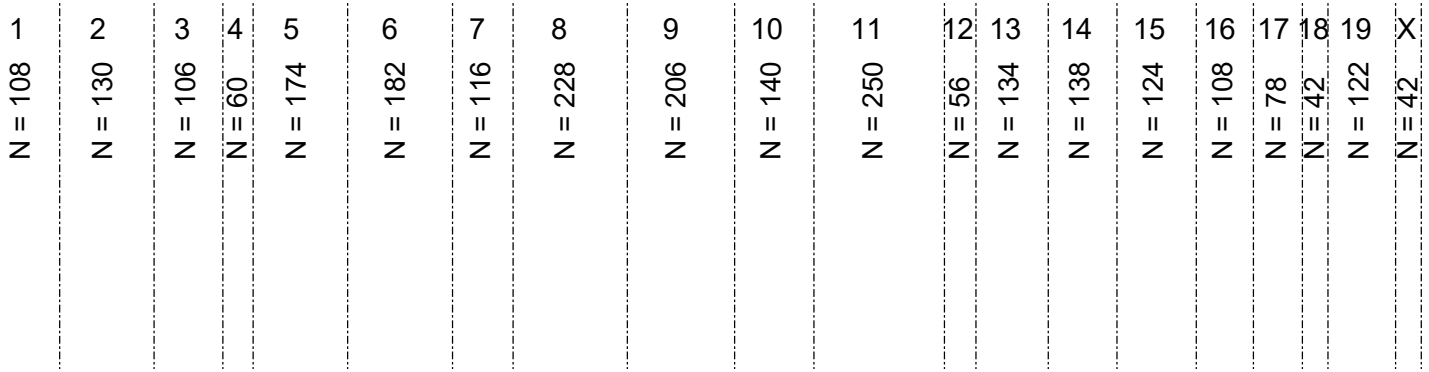
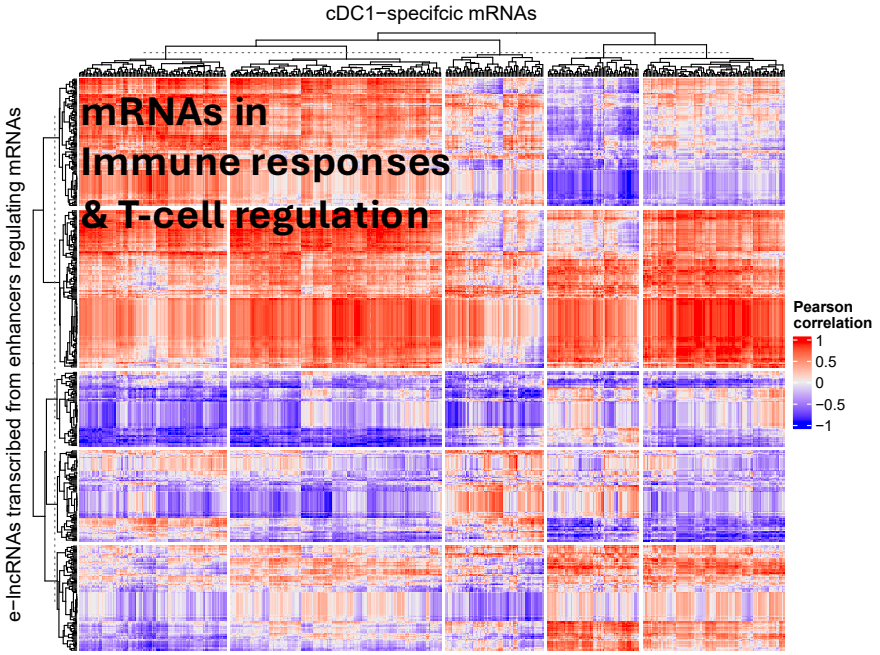


Fig: Genome-wide contact map of cDC1 enhancer regions regulating cDC1-specific mRNAs & transcribing e-lncRNAs.



Now, we are expanding our work to create a broad atlas of lncRNAs in myeloid cells. We are unlikely to find enough high-quality data to run ABC. Need alternatives!

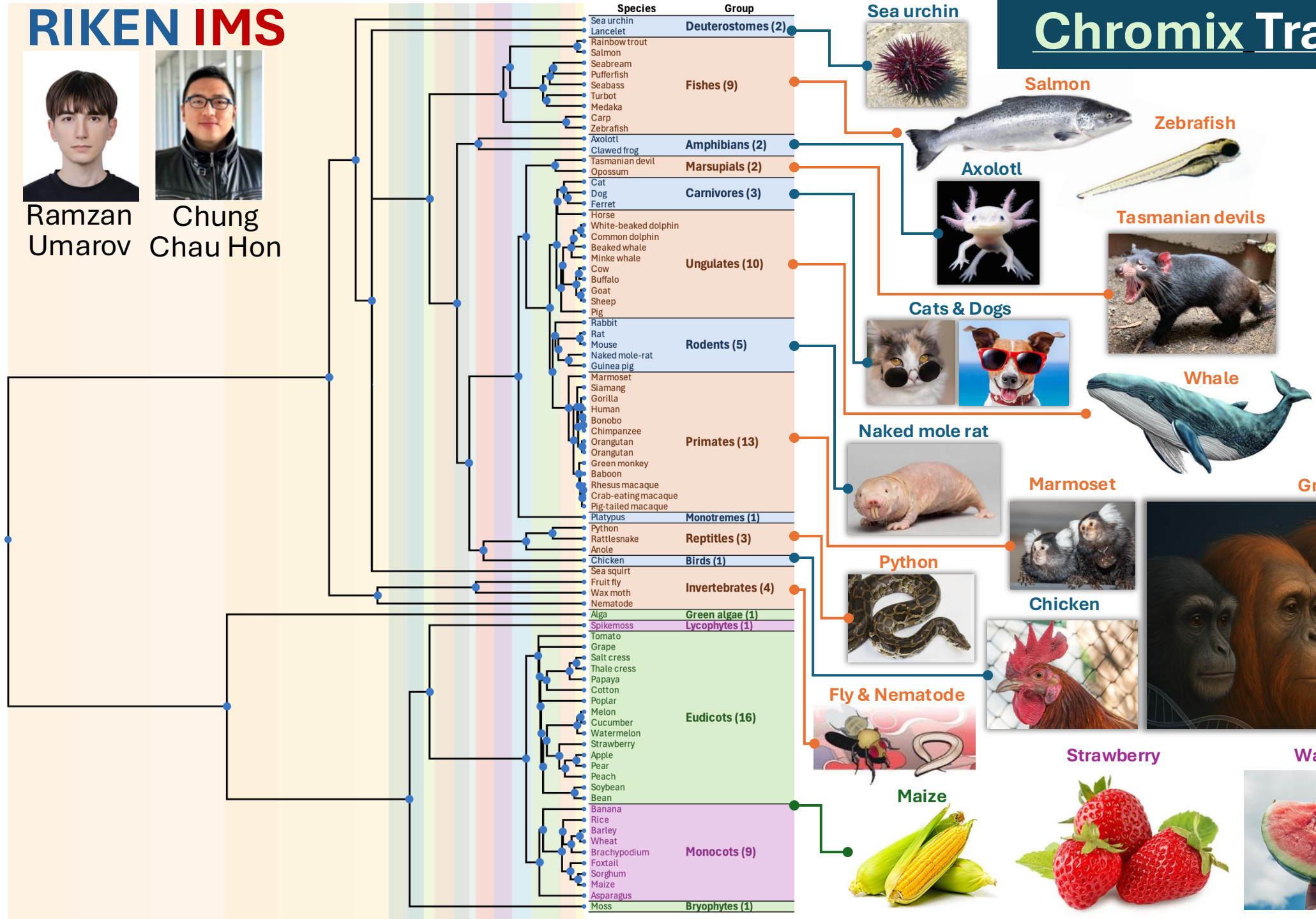
24

RIKEN IMS

Ramzan Umarov

Chung Chau Hon

Chromix Training Data



Sea urchin

Salmon

Zebrafish

Axolotl

Tasmanian devils

Cats & Dogs

Whale

Naked mole rat

Marmoset

Great apes & Human

Python

Chicken

Fly & Nematode

Maize

Strawberry

Watermelon

Peach

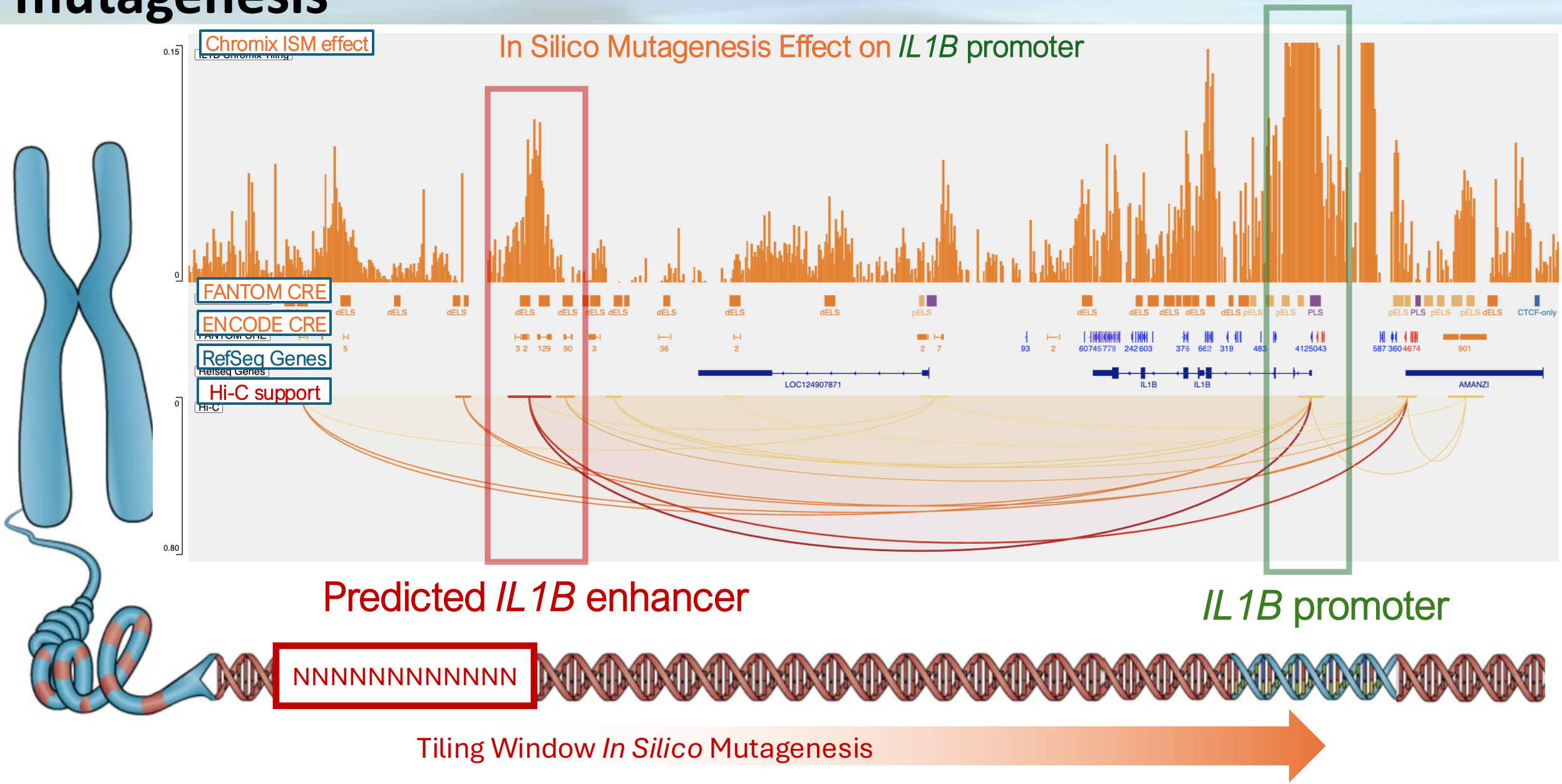
83 genomes

- **51** Vertebrates
- **4** Invertebrates
- **28** Plants

31535 tracks

- **~8000** Histone marks
- **~3600** TF binding
- **~6300** Accessibility
- **~4800** Transcription
- **~8500** TSS

Chromix predicts enhancer-gene pairs with tiling window mutagenesis



Cellular Specialization and Communication in the Human Body

複数の臓器・組織 | 数百種類の細胞 | 約30兆個の細胞

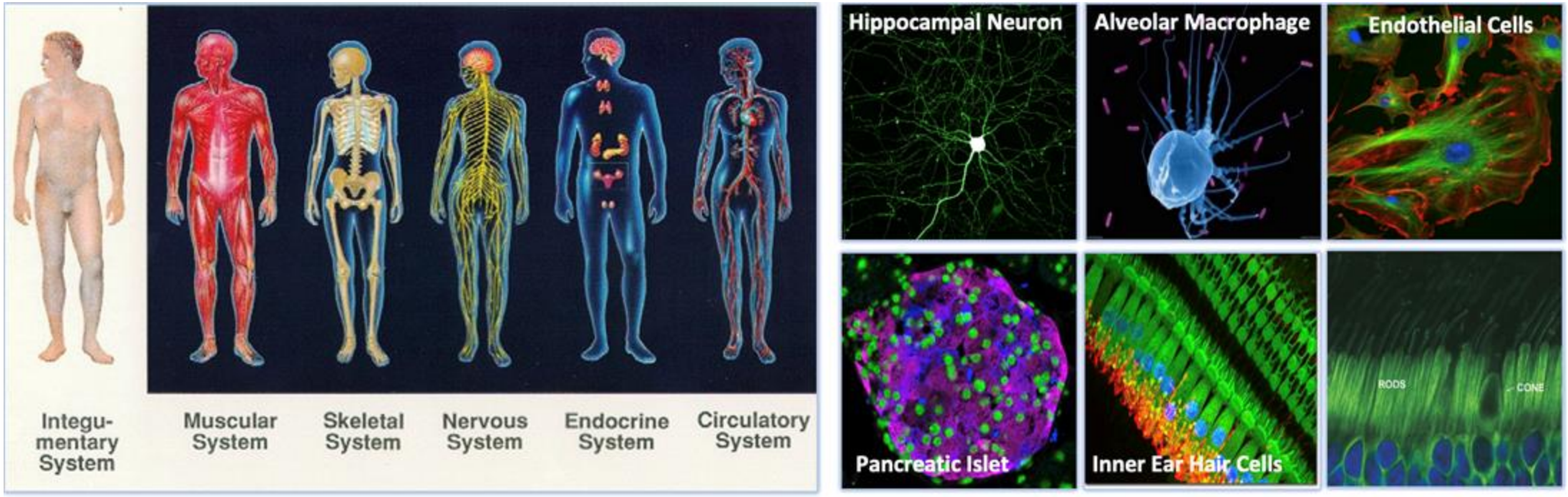
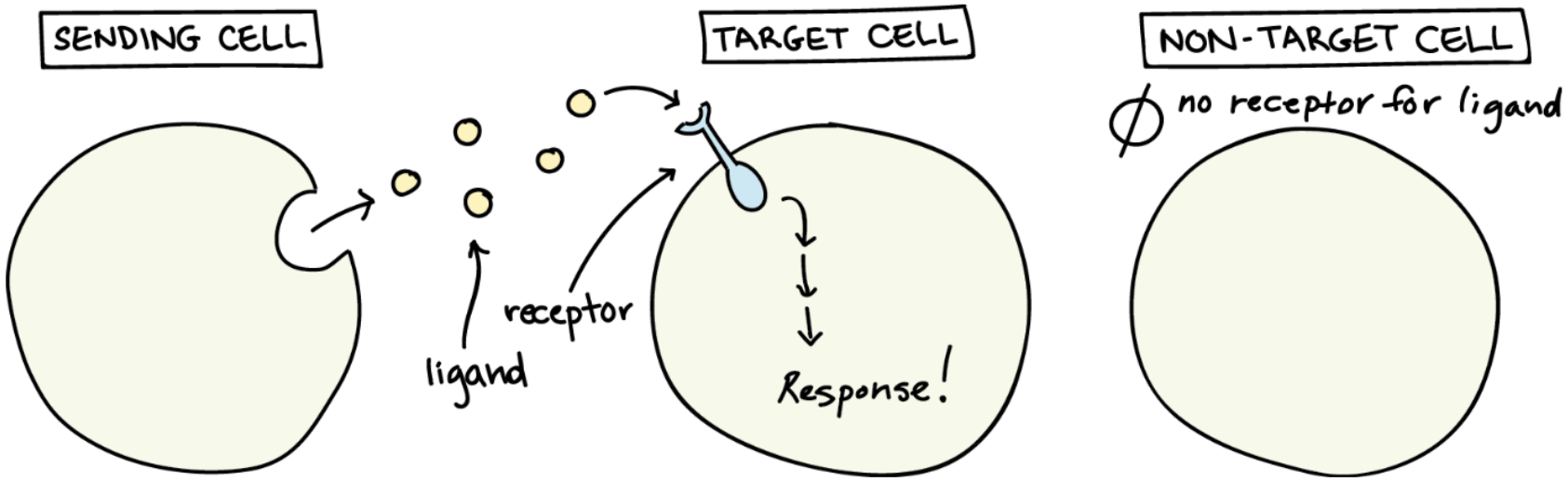


Image source: https://anatomyofthebody.weebly.com/

The human body comprises **approximately 30 trillion cells, each specialized for distinct functions—**such as neurons transmitting signals, cardiomyocytes contracting rhythmically, and hepatocytes metabolizing nutrients**—demonstrating a sophisticated division of labor.** This cellular specialization is orchestrated through **intricate cell-cell communication networks**, including gap junctions, exosomes, and signaling pathways, ensuring coordinated physiological processes and maintaining homeostasis.

Principles of Cell—Cell Communication

Cells typically communicate by secreting signaling proteins (**ligands)**, which bind to specific proteins located on the surface of target cells (**receptors**).



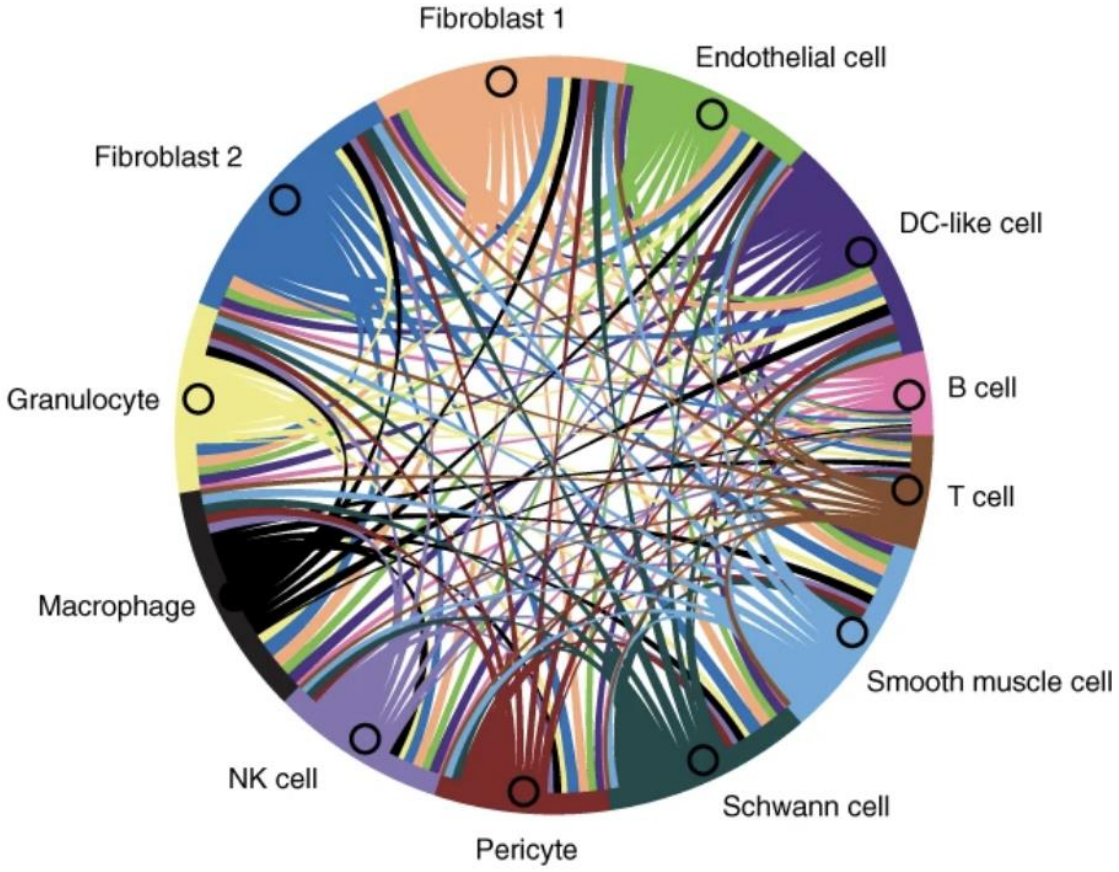
These **intercellular** (between-cells) signals are then converted into an **intracellular** (within-cell) signal causing a response: changes in gene activity, cell division, cell death and many more.

https://www.khanacademy.org/science/biology/cell-signaling/mechanisms-of-cell-signaling/a/introduction-to-cell-signaling

Complexity of Cell—Cell Communication Networks

There 20,000 proteins in the human genome

20,000*20,0000 → 400,000,000 Protein-Protein Interactions (potential)



Even 12 interacting cells types can make a very complex signaling network

nature COMMUNICATIONS

A draft network of ligand–receptor-mediated multicellular signalling in human, JA Ramilowski, *et al. Nature communications* **6(1)**, 7866, **2015**

PMID: 26198319; **DOI:** 10.1038/ncomms8866

Situation of the Ligand-Receptor Databases in 2014

- No comprehensive resources with ligand-receptor interactions
- No clear understanding of the following:
  - how many ligands/receptors are present in human cells
  - how many ligand-receptor interactions are used between cells
  - autocrine signals – cells within the same cell type signaling to each other

Birth of the ConnectomeDB Project – aims to address the above issues

- Discovered:
  - 708 human ligands
  - 691 human receptors
  - 2,557 human ligand-receptor pairs (1,894 with experimental support)
- Showed ligand/receptor interactions and expression profiles in 144 human primary cell types

nature communications

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature communications › articles › article

Article | Open access | Published: 22 July 2015

A draft network of ligand–receptor-mediated multicellular signalling in human

Jordan A. Ramilowski ✉, Tatyana Goldberg, Jayson Harshbarger, Edda Kloppmann, Marina Lizio, Venkata P. Satagopam, Masayoshi Itoh, Hideya Kawaji, Piero Carninci, Burkhard Rost & Alistair R. R. Forrest ✉

Nature Communications **6**, Article number: 7866 (2015) | Cite this article

55k Accesses | **827** Citations | **74** Altmetric | Metrics

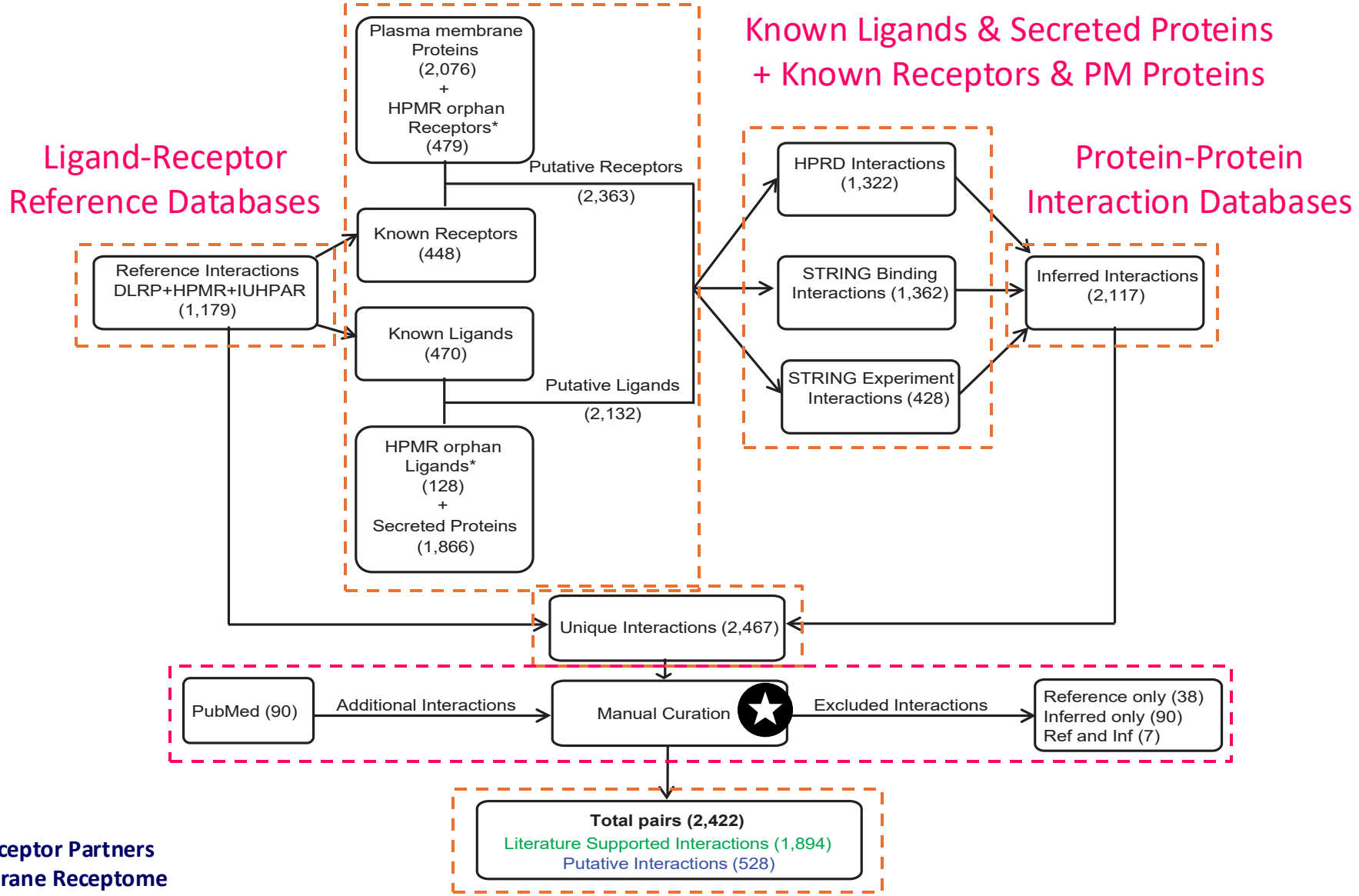
A Corrigendum to this article was published on 10 February 2016

This article has been updated

Abstract

Cell-to-cell communication across multiple cell types and tissues strictly governs proper functioning of metazoans and extensively relies on interactions between secreted ligands and cell-surface receptors. Herein, we present the first large-scale map of cell-to-cell

Predicting Ligand-Receptors Interactions



DLRP : Database of Ligand-Receptor Partners
HPMR: Human Plasma Membrane Receptome
HPRD: Human Protein Reference Database

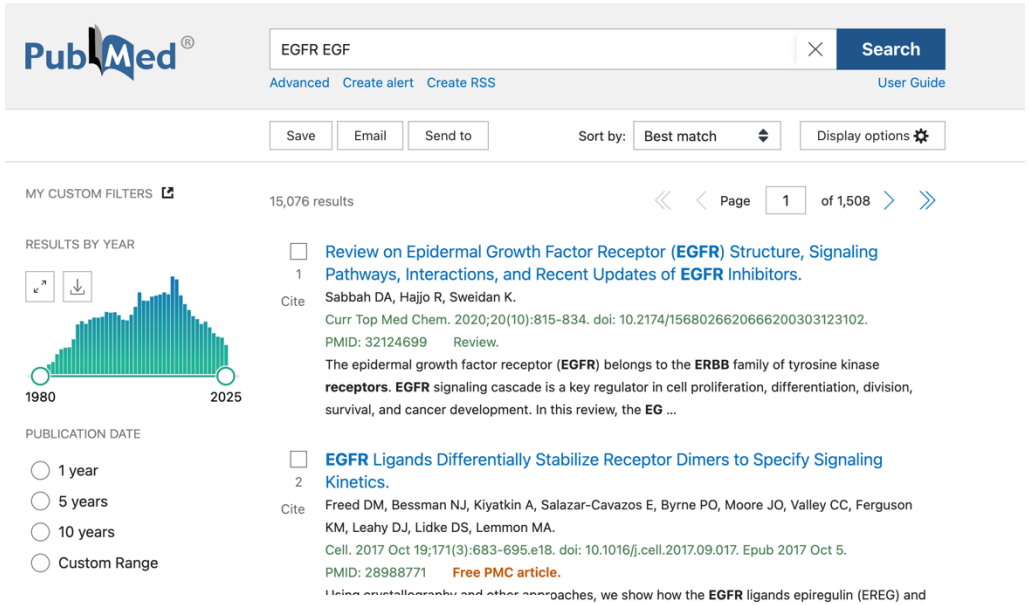
Searching for Interaction Evidence for ConnectomeDB2015

No consistent method for evidence searching

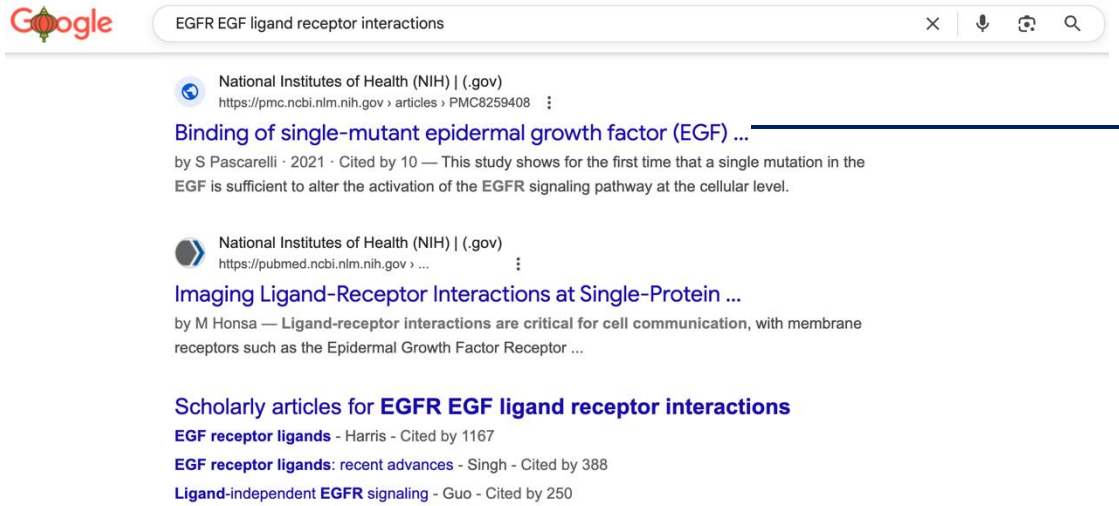
- Each LR pair was manually checked using PubMed (1) and Google searches (2)
- Pairs with at least one supporting evidence were labeled as “supported”
- Process involved multiple annotators and was very time-consuming

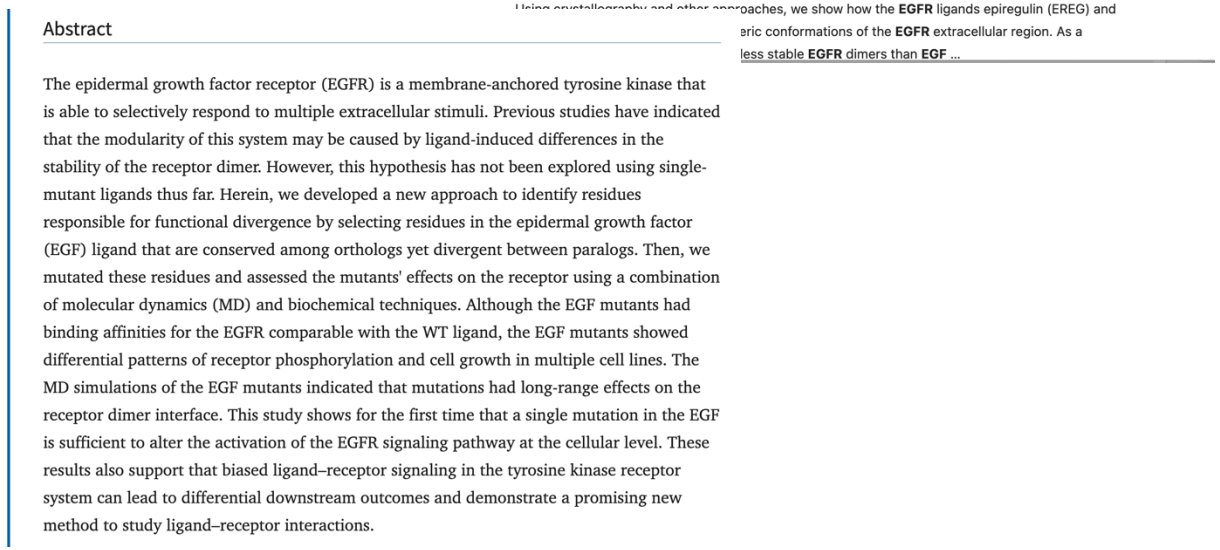
Example: EGFR – EGF interaction

(1)



(2)

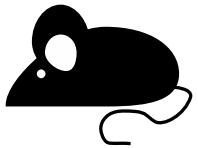


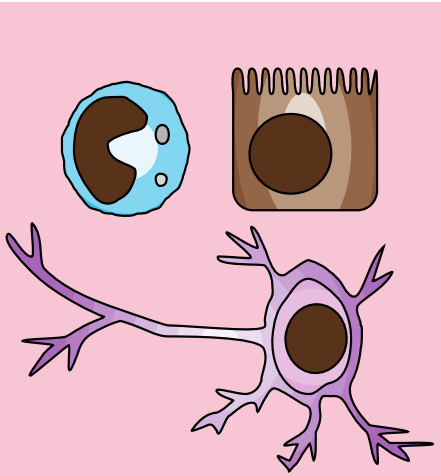
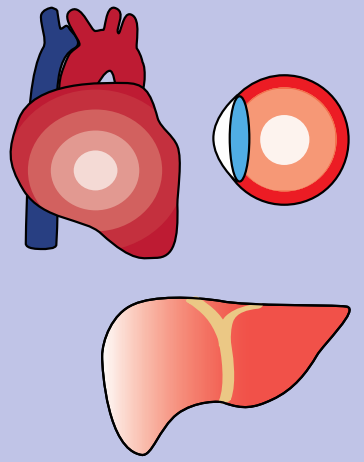
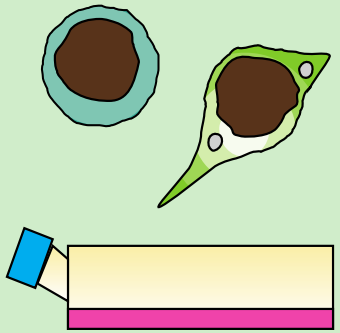


(CAGE) Cap GEne Expression Atlas in human and mouse

Cell Ontologies

- Mesenchymal: 63
- Epithelial : 34
- Hematopoietic: 29
- Endothelial: 9
- Nervous system: 4
- Other: 5

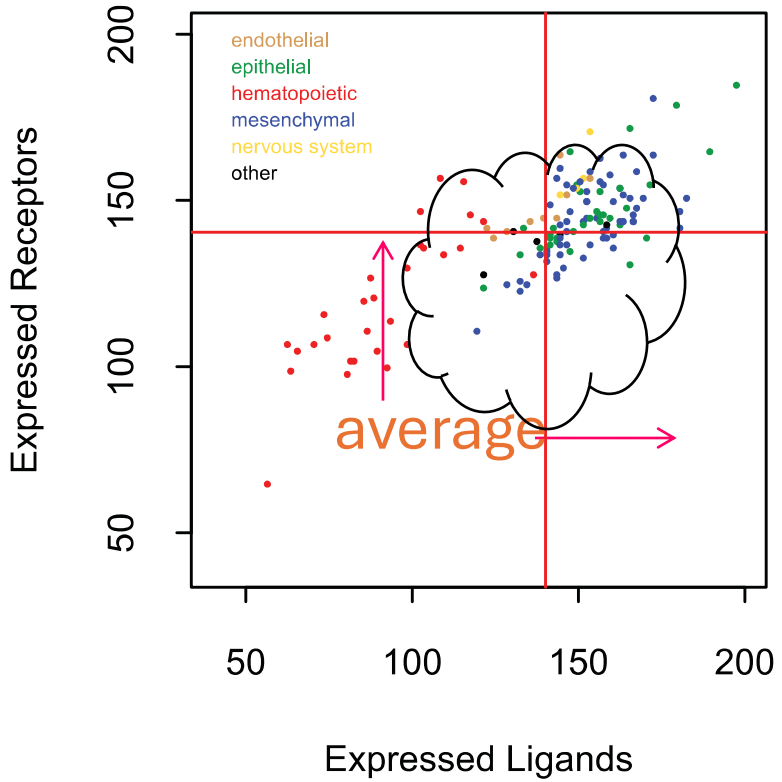


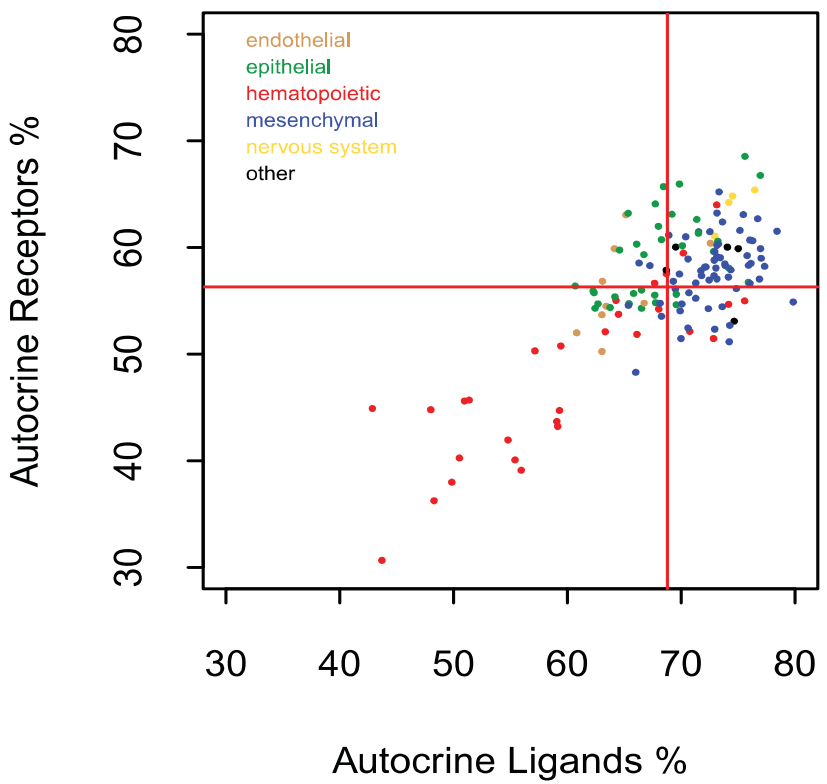
144 Primary Cells	Tissues	Cell lines
		
573	152	250
128	271	

A promoter-level mammalian expression atlas
Forrest ARR et al., Nature 507, 462-470 (2014)

Ligand & Receptors in FANTOM 5 Collection

464 ligands & 477 receptors expressed in all our primary cells



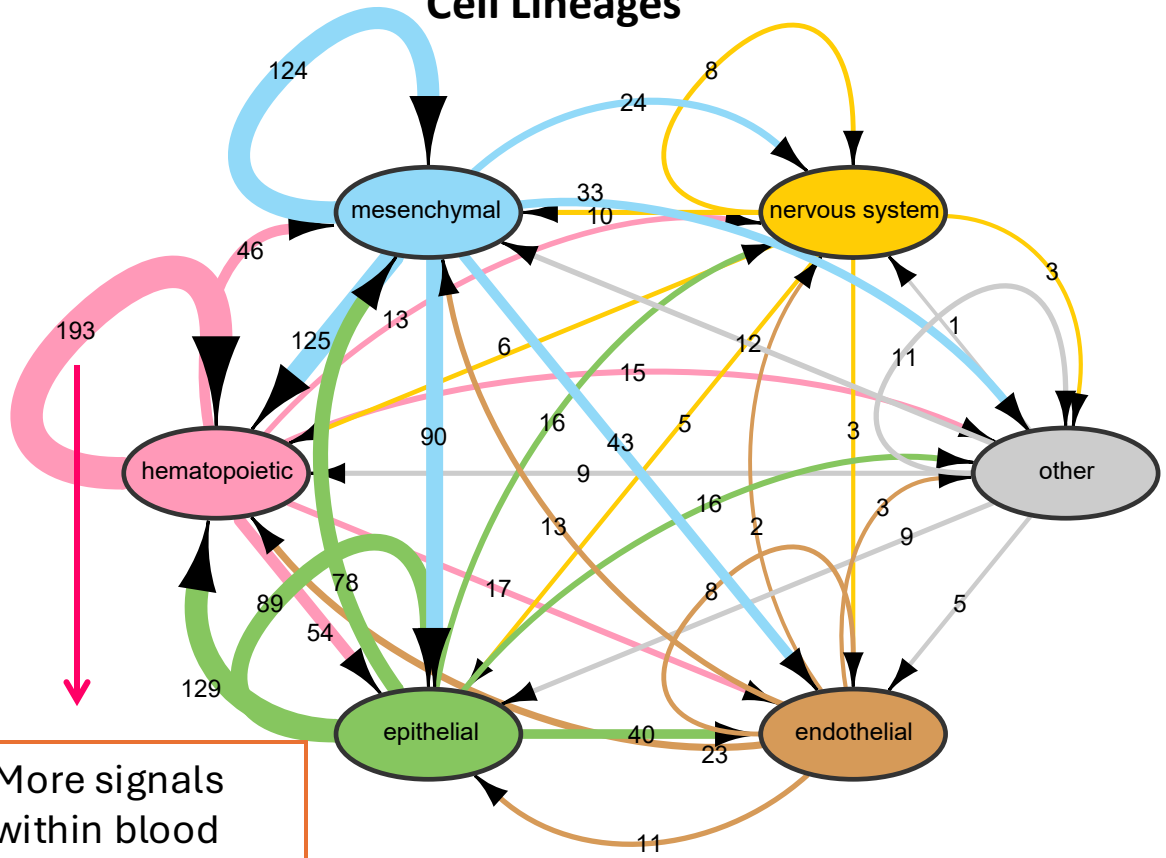


Blood lineage expresses less ligands (95) and receptors (105) on average than other cells ligands (140) and receptors (140)

Blood lineage expresses less ligands (95) and receptors (105) on average than other cells ligands (140) and receptors (140)

Cells communicate more within than to other lineages

Summarized Signaling Within and Between the Cell Lineages

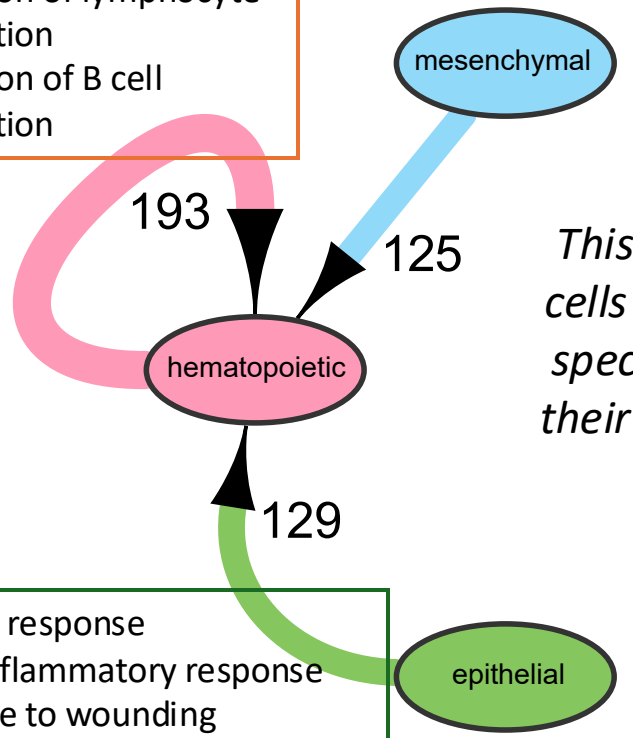


More signals within blood than from blood to other lineages

Signaling To and Within Blood Lineage is Enriched in Immune Responses

GO:0006955 immune response
GO:0006952 defense response
GO:0006954 inflammatory response
GO:0051249 regulation of lymphocyte activation
GO:0050864 regulation of B cell activation

GO:0042330 taxis
GO:0006935 chemotaxis
GO:0006952 defense response



This illustrates cells functional specificity and their division of labor.

GO:0006952 defense response
GO:0002526 acute inflammatory response
GO:0009611 response to wounding
GO:0002253 activation of immune response

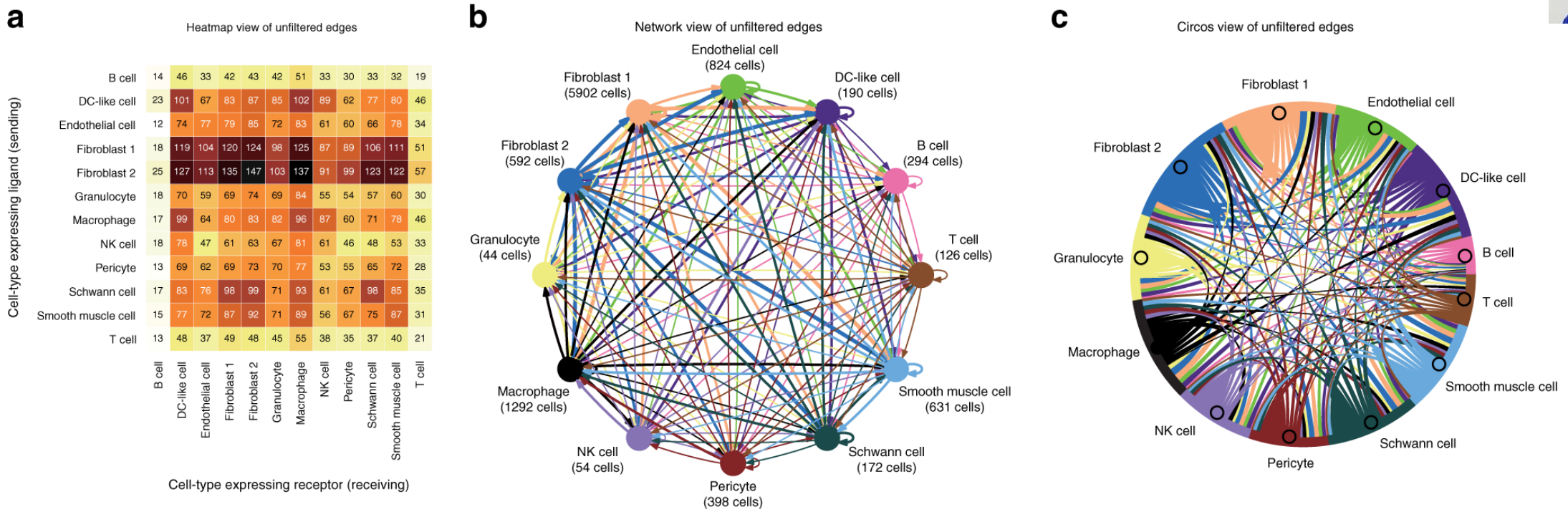
nature COMMUNICATIONS

Predicting cell-to-cell communication networks using NATMI, R Hou, *et al. Nature communications* **11(1)**, 5011, **2020**

PMID: 33024107; DOI: 10.1038/s41467-020-18873-z

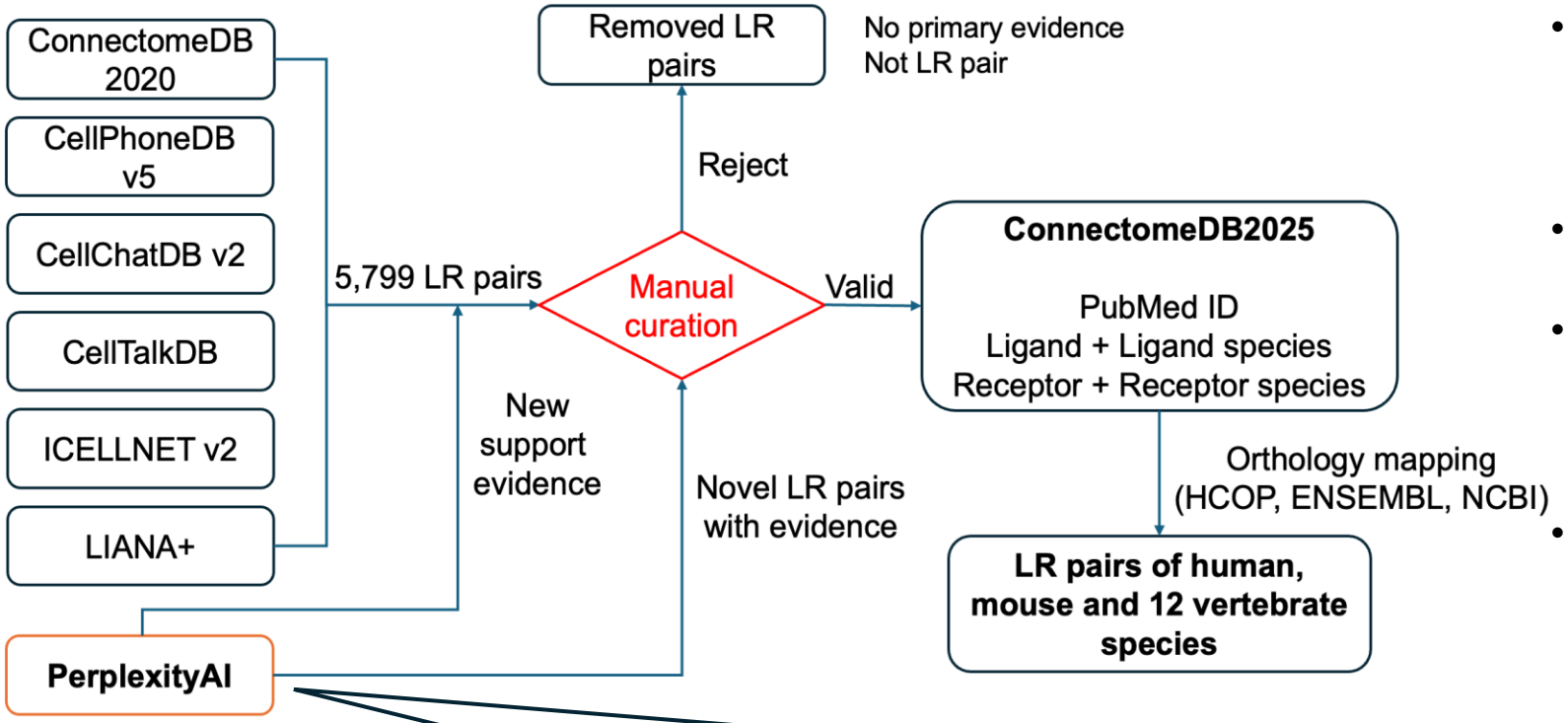
Updated ConnectomeDB2020: NATMI (Network Analysis Toolkit for the Multicellular Interactions) + 1,930 Curated Ligand-Receptor Pairs

Rui Hou



Summary: Although connectomeDB2015 and connectomeDB2020 have established a baseline for discovering ligand-receptor interactions in humans, they lack a consistent method for pre-screening the interaction evidence in published literature.

ConnectomeDB2025: Leveraging AI to Build the Most Robust Ligand–Receptor Database



ConnectomeDB

- Expanded the integrated dataset via **two methods: manual keyword** searches and **AI-assisted queries using the Sonar-Pro model in PerplexityAI**
- For manual keyword searches, limited them to literature from **2020~**
- This restriction was intended to:
  - capture newly described LR pairs
  - reduce the volume of articles

  needing manual review
- Using the combined dataset, PerplexityAI identified additional plausible LR partners for existing ligands and receptors

Identification of new ligand–receptor pairs absent from existing DBs

Initial Prompt
~6400 potential LR pairs

→

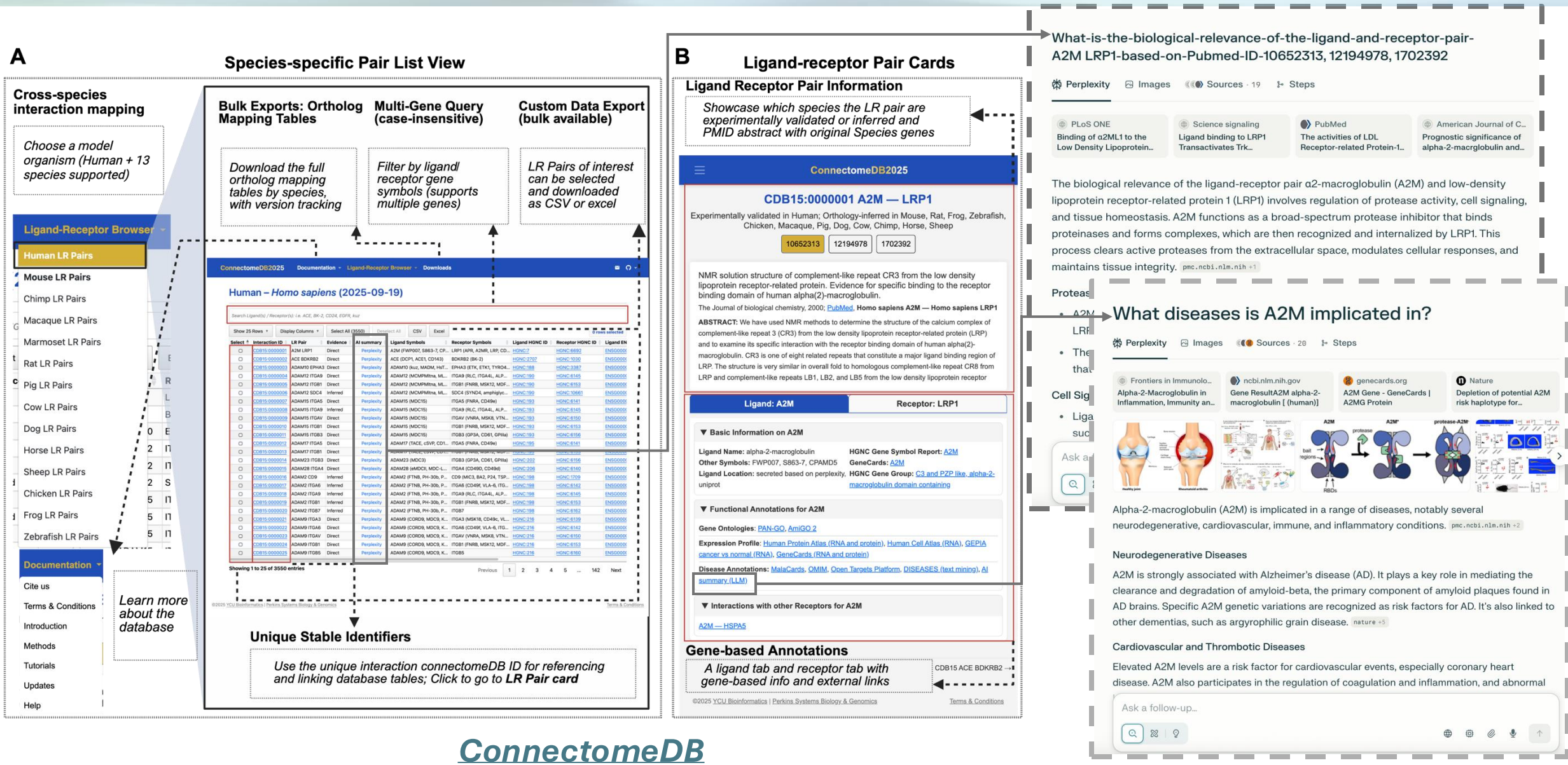
QC (Symbol validation etc.)
~4000 potential LR pairs

→

Secondary Prompt
~1000 potential LR pairs

Liu, Maezono et. al., NAR Database, 2025 (Accepted)

ConnectomeDB2025: Leveraging AI for knowledge expansion

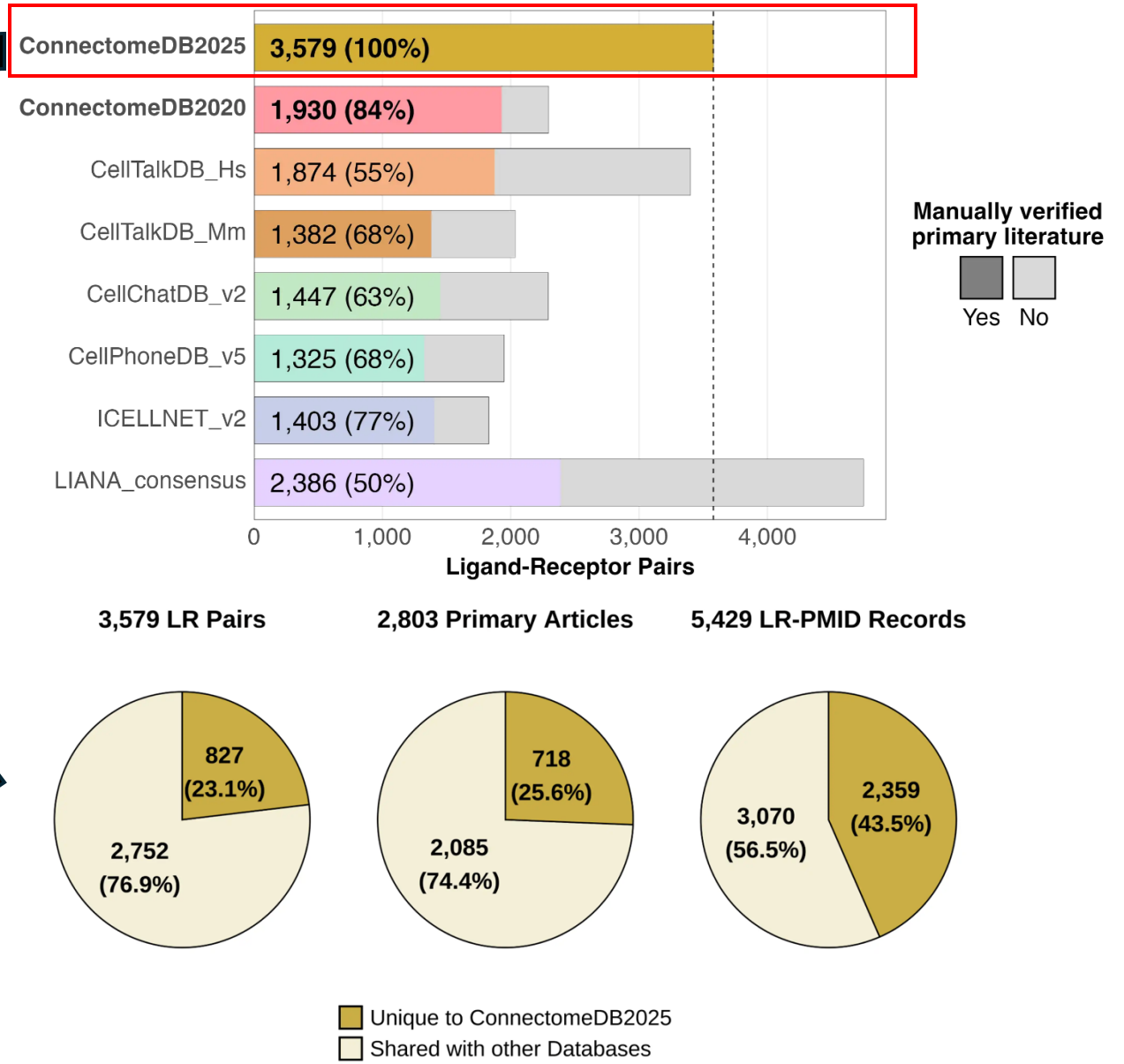


ConnectomeDB

Screenshots from Perplexity A

ConnectomeDB2025 Summary: Most Comprehensive Ligand-Receptor Database

- **High-Quality Curation:** 3,579 ligand–receptor pairs from 2,803 primary research articles
- **Evidence-Linked:** 5,429 “triplets” connecting pairs to supporting publications; 2,359 unique to ConnectomeDB2025
- **Expanded & Updated:** Added 827 pairs and 773 publications absent from previous databases, including 264 new interactions since 2020
- **Cross-Species Coverage:** Human, mouse, plus 12 other vertebrates (~97% homolog coverage)
- **Accessible Online:** Searchable, downloadable pair lists and detailed pair summaries at connectomedb.org.



Liu, Maezono et. al., NAR Database, 2025 (Accepted)

Generative AI for Your Research

Basics of Prompt Engineering

Key Recent Advances (2022–2025)

Reasoning Prompts

- *Chain-of-Thought (Wei et al., 2022) & Zero-Shot CoT (Kojima et al., 2022)*: step-by-step reasoning improves accuracy
- *Self-Consistency (Wang et al., 2022)*: sample multiple reasoning paths → more reliable answers

Prompt Automation

- *Automatic Prompt Engineer (Zhou et al., 2022), PE2 (ACL 2024)*: LLMs (meta-prompting) can design/refine prompts better than humans
- *Long Prompt Optimization (ACL 2024)*: search/mutate long prompts for complex tasks

Domain & Education

- *Domain-Specific Prompts (Chemistry, 2024)*: embedding scientific knowledge reduces hallucinations
- *Surveys (2024/25) & Curricula (Education 2025)*: prompt engineering now taught systematically (AI/education)

Prompt Engineering in Biology & Medicine

- **Data analysis:** guiding AI to extract patterns from omics datasets
- **Hypothesis generation:** framing biological questions for novel insights
- **Clinical decision support:** structuring prompts for accurate, interpretable outputs
- **Knowledge synthesis:** turning scattered literature into focused reviews
- **Collaboration skill:** bridges domain expertise with AI reasoning

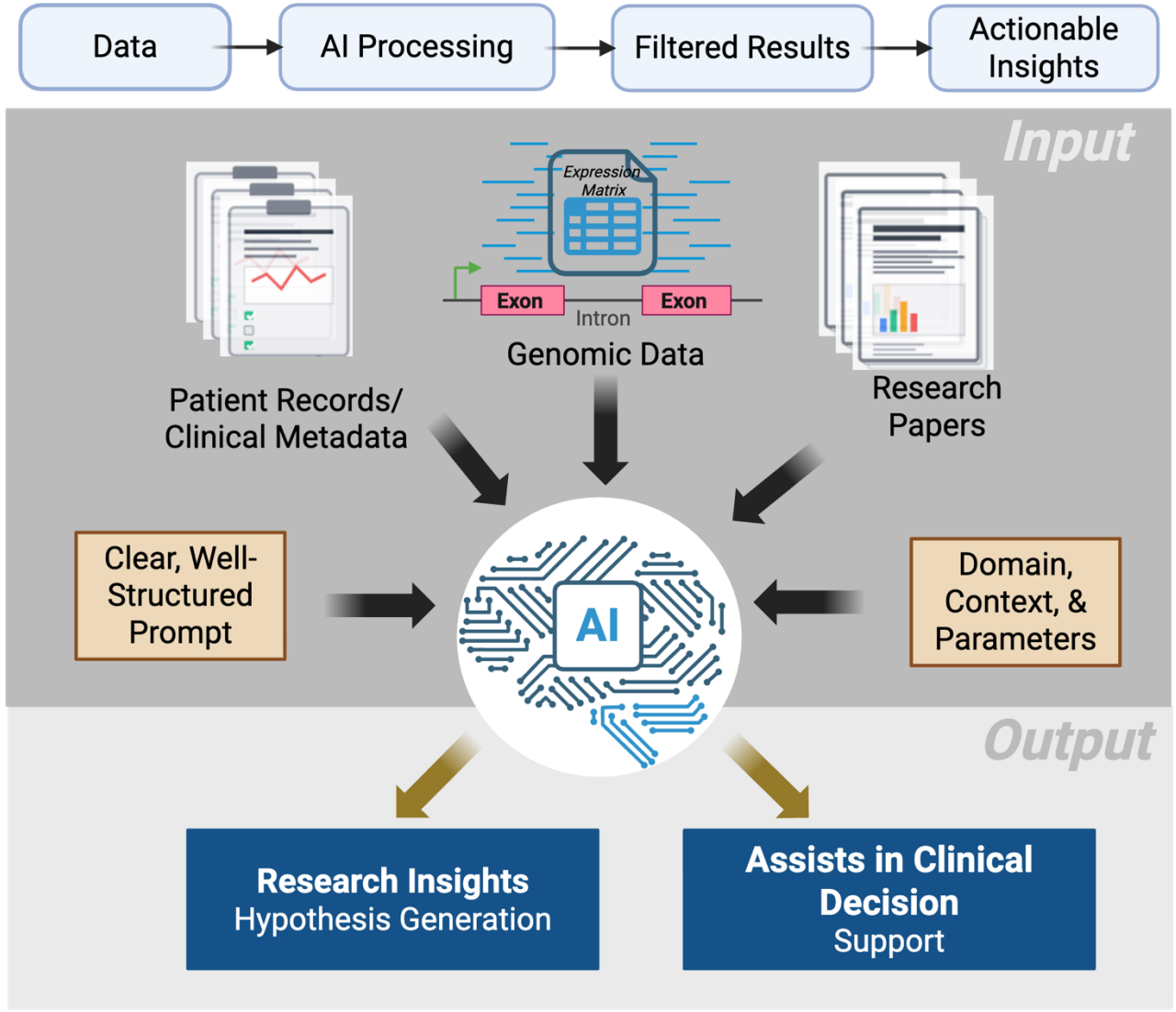


Image created in Biorender

Tips for Prompt Engineering— Learning to Talk to AI

1. Be Clear & Specific

- Define the task: “Summarize RNA-seq differential expression results” vs. “Explain RNA-seq.”
- Include format or length requirements: table, bullet points, short summary

2. Provide Context

- Give relevant background: organism, tissue, experimental conditions.
- Specify assumptions or constraints: e.g., “only human genes” or “exclude low-confidence hits.”

3. Stepwise Instructions Help

- Encourage reasoning: “Explain your answer step by step” or “Show calculations.”
- Useful for complex analyses, medical reasoning, or interpreting data

4. Use Examples (Few-Shot)

- Provide one or two illustrative examples of ***input → desired output***
- Reduces ambiguity, especially for bioinformatics or clinical formats

5. Check & Refine

- Verify AI suggestions against trusted sources
- Adjust wording or add clarifications iteratively

6. Safety & Domain Awareness

- AI may hallucinate or misinterpret rare conditions
- Always cross-check clinical/experimental outputs

7. Bonus: Experiment & Document

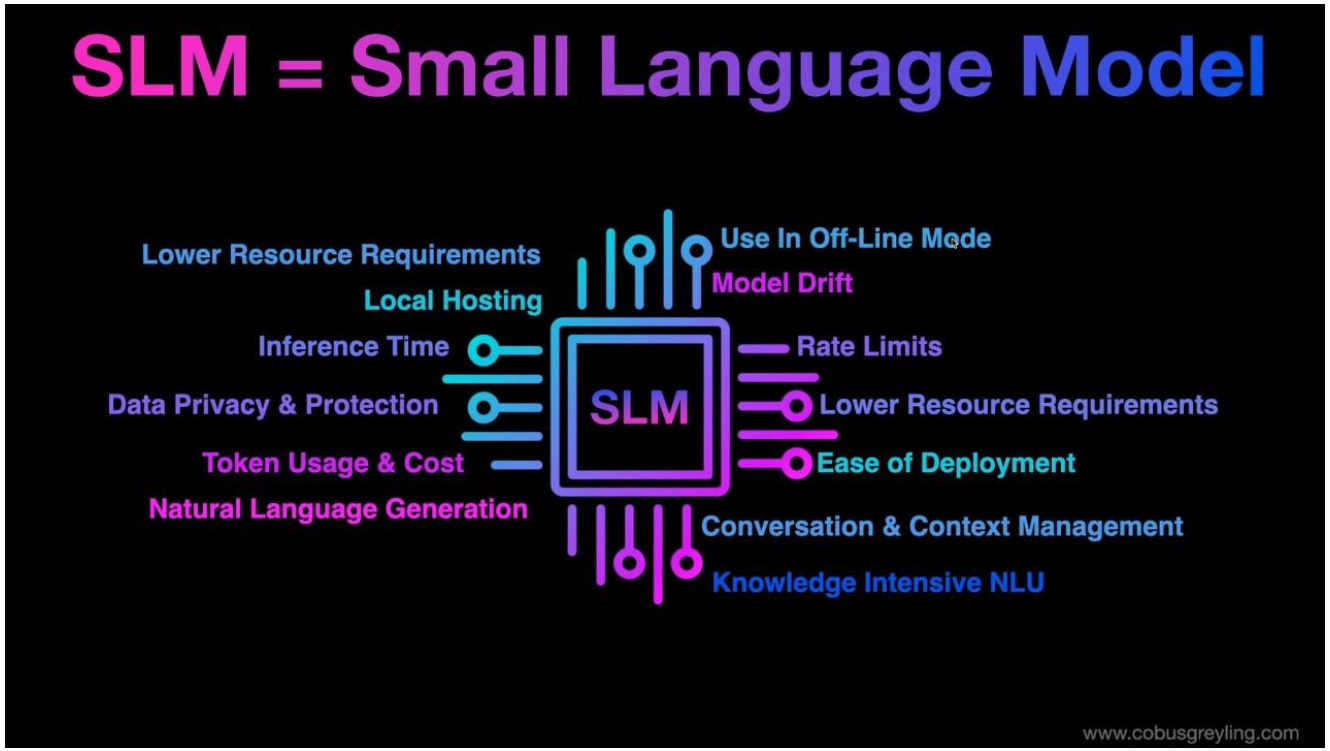
- Slightly different wording can change results; record prompts that work well

Small Language Models & Hybrid Architectures: The Future of Efficient AI

Small Language Models (SLMs): Models under 10 billion parameters that are optimized for specific tasks, offering faster inference, lower costs, and reduced environmental impact compared to larger counterparts

Hybrid AI Systems: Combining the strengths of both large and small models to create efficient, scalable, and specialized AI solutions.

Trend Towards Efficiency: The AI community is shifting focus from purely scaling up models to optimizing them for specific applications, leading to more sustainable and accessible AI technologies.



Call to Action: TheOpportunity to Learn and Adapt

AI is no longer optional in biology — it's the new microscope

Researchers must:

- Learn AI concepts (transformers, embeddings, generative models, prompt engineering)
- Partner across disciplines (Biology ↔ Computer Science)
- Experiment withAI tools in daily research

Digital Biology is here. The question is: **will we lead or follow?**

Invitation: Let's explore and learn AI for biology together

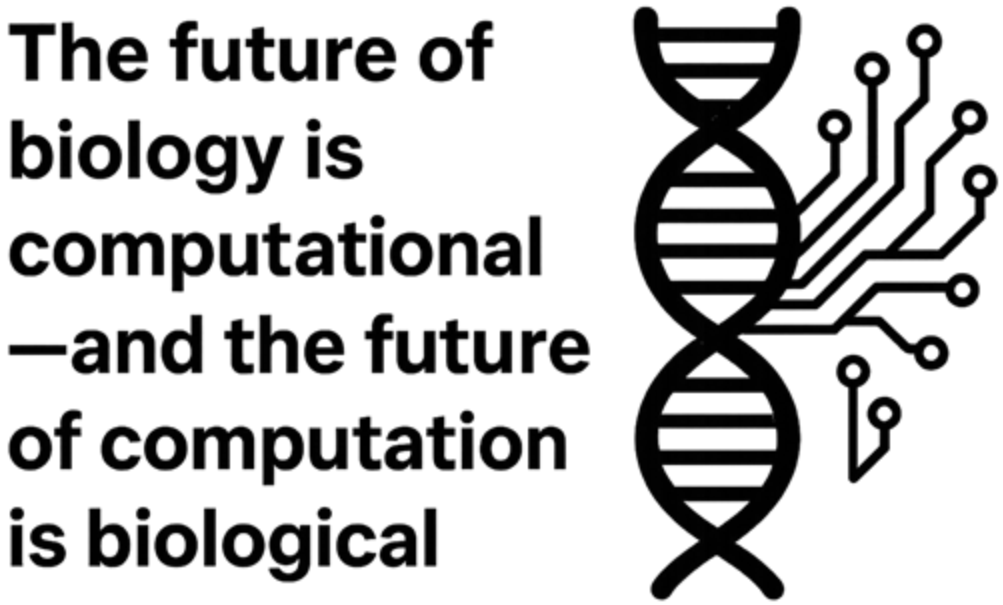
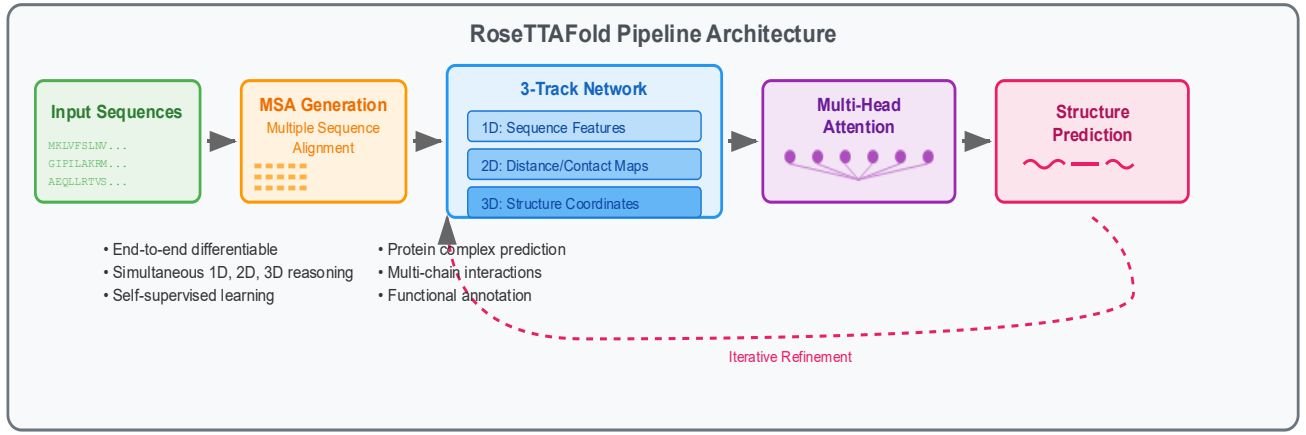


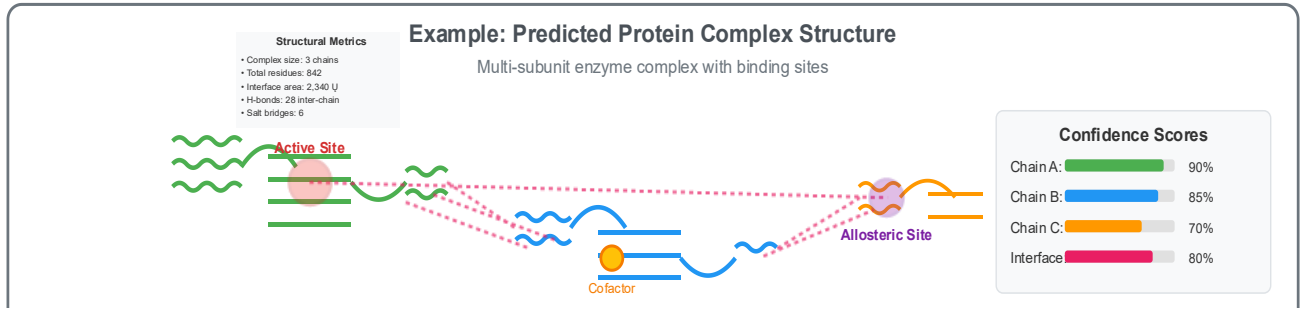
Image created by ChatGPT

Appendices

RoseTTAFold: Faster and often more flexible for novel folds or complexes than AlphaFold

- Developed in parallel at UW, with a **three-track network** (sequence, distance, coordinates)
- Pioneered the idea of using protein language models (from sequences alone) to bootstrap structure learning
- Extends to **protein–protein and protein–DNA/RNA interactions**, bridging molecular biology and structural genomics





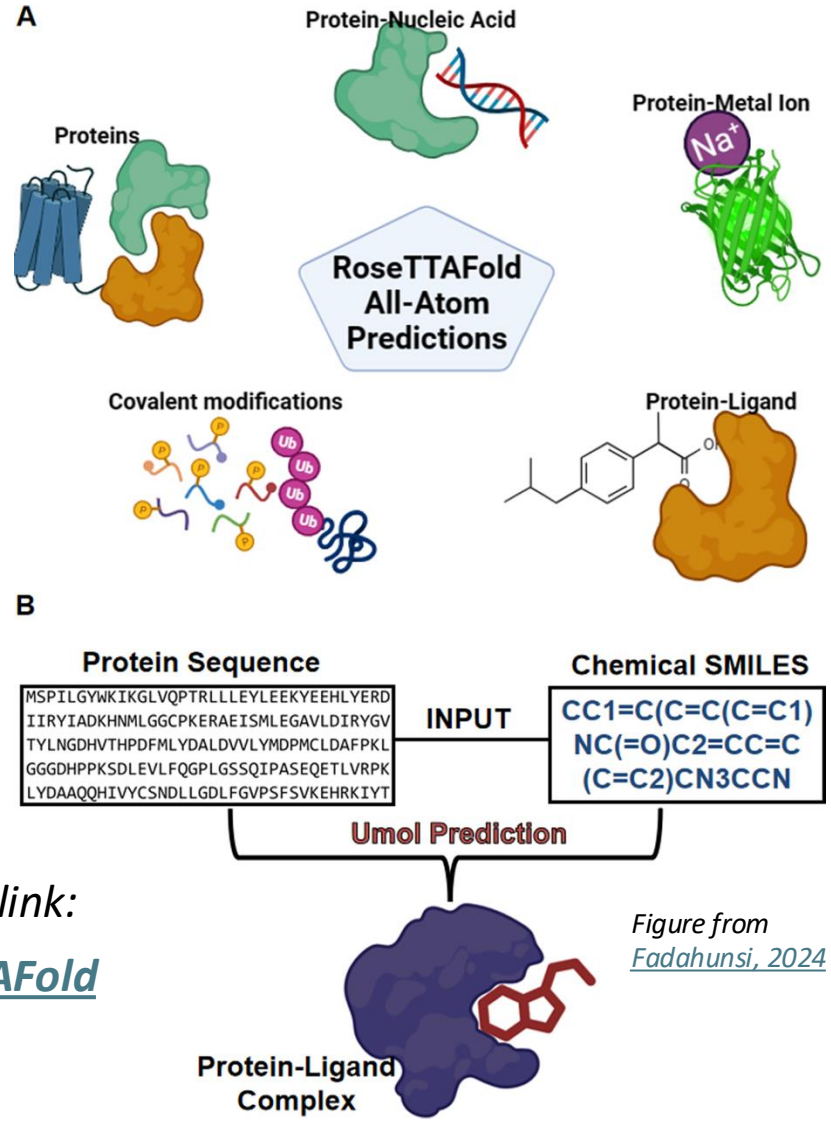


Figure from [Fadahunsi, 2024]

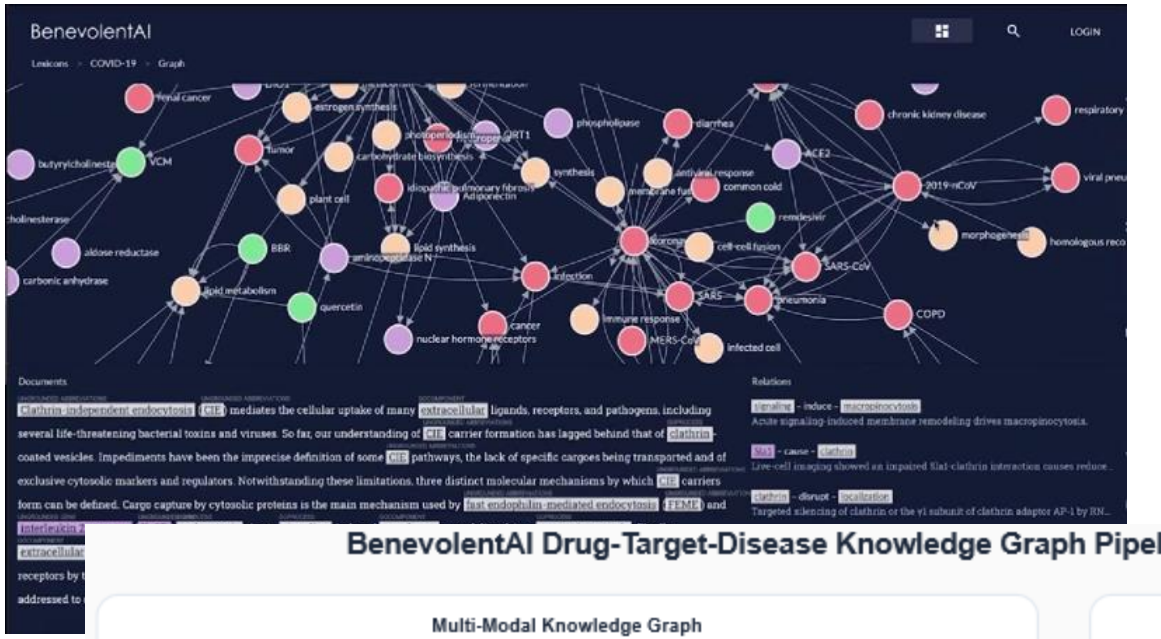
Related link:

[ROseTTAFold]

BenevolentAI: Leverages on AI to quickly find new therapeutics

- Integrates biomedical knowledge graphs with deep learning to find **novel therapeutic hypotheses**
- Famous for COVID-19: rapidly identified **baricitinib** as a drug (now FDA-approved)
- Uses AI to navigate the **vast biomedical literature, omics, and clinical trial data** to prioritize interventions
- Matches the **right therapy to the right patient at the right time**

Related link: ***BenevolentAI***



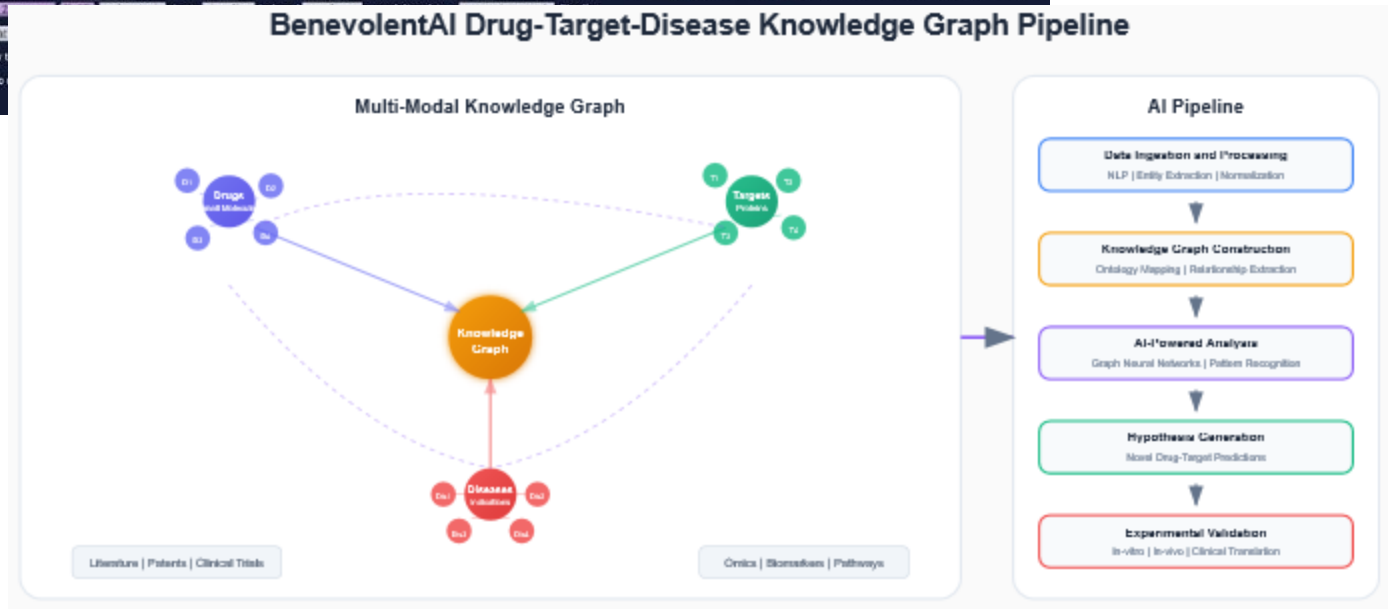


Image created by Claude AI

Atomwise: leveraging the power of AI to revolutionize small molecule drug discovery

- Applies AI-driven molecular docking and virtual screening
- Trains on billions of small molecules to predict **binding affinities and poses** with drug targets
- Accelerates **hit discovery** from years to weeks, with reduced cost
- **Used in neglected diseases** (e.g., Ebola, malaria) where traditional pharma has little incentive

Related link: ***Atomwise***

Atomwise AI Driven Ligand Docking Pipeline

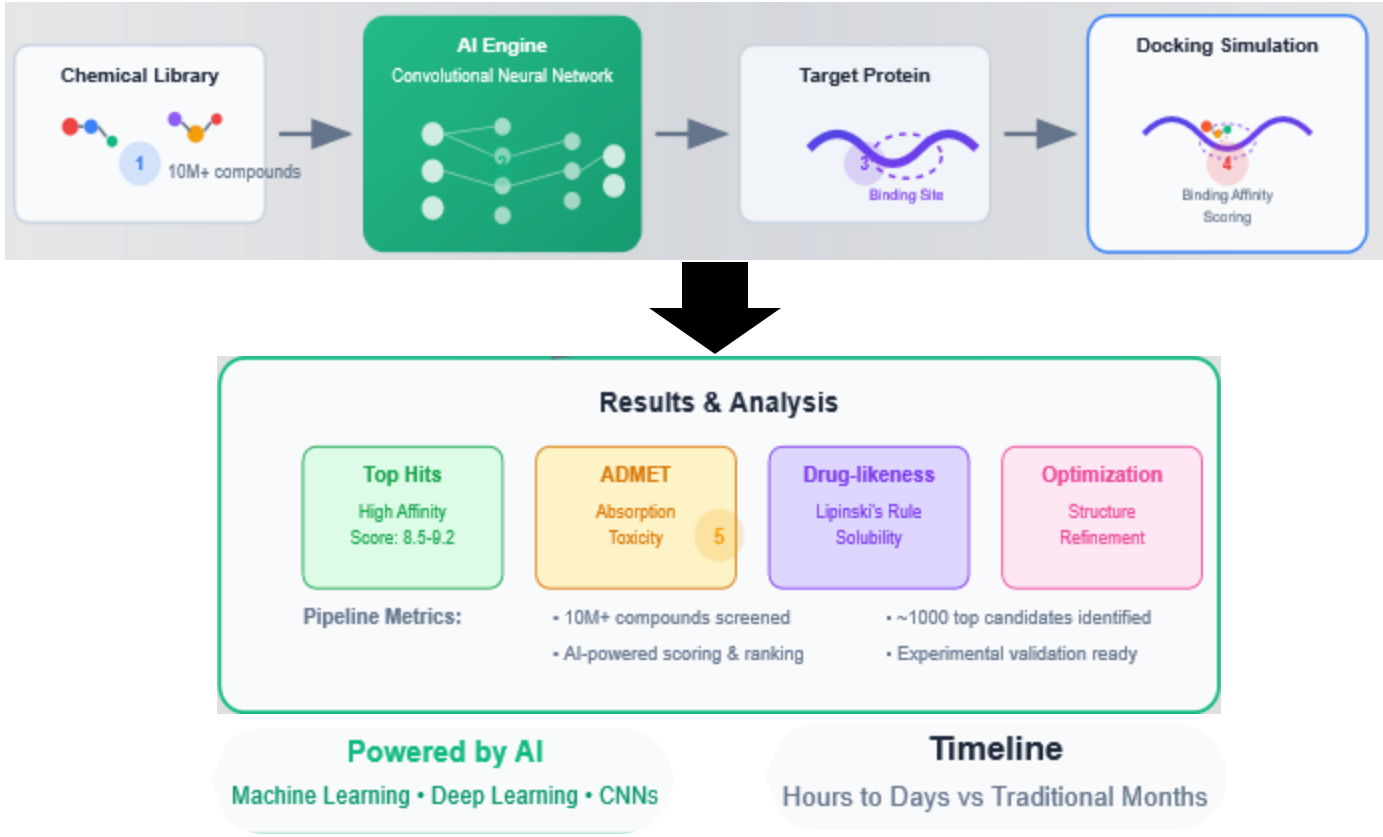


Image created by Claude AI