

Cell Maps for AI (CM4AI)

BRIDGE2AI

UC San Diego UCSF UNIVERSITY OF SOUTH FLORIDA



Yale

Stanford University

SFU SIMON FRASER UNIVERSITY

TEXAS
The University of Texas at Austin

THE HASTINGS CENTER

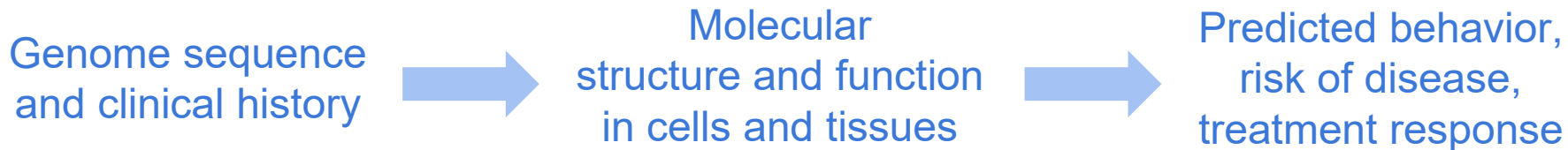
UAB



Functional Genomics Data Generation Project

AI translation of the human genome

- AI is revolutionizing many tasks, including those in biology and medicine.
- How does the genetic code dictate the architecture of living systems?
- How do genetic changes to this architecture promote (or prevent) risk for disease?





“Dear AI, my patient has genetic mutations in PTEN, DAAM1, & 48 other genes. What treatment is best? Should we be looking at any new targets?”

What data are needed for an AI genome translator?

MAJOR DATA RESOURCES

Genome sequence and clinical history

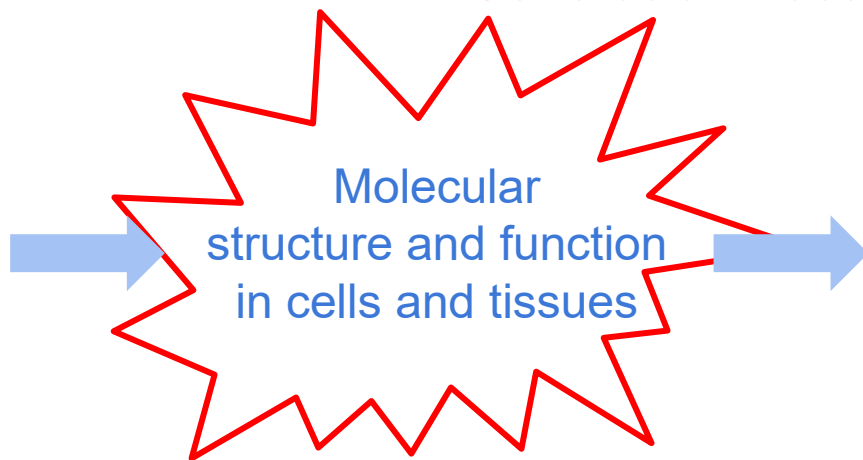
Genome sequencing projects

Our Bridge2AI Data Generation Focus

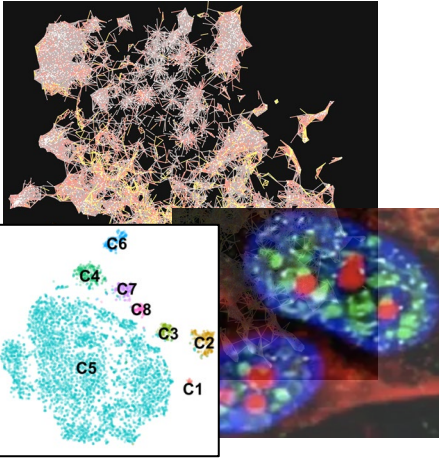
Molecular structure and function in cells and tissues

Predicted behavior, risk of disease, treatment response

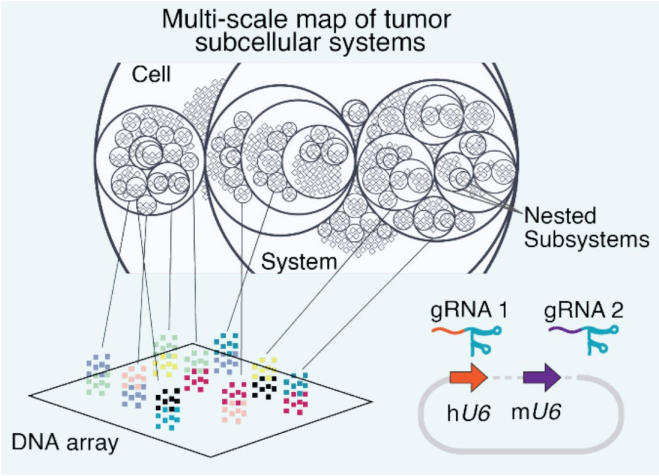
Genome-wide association studies, High-throughput drug screening



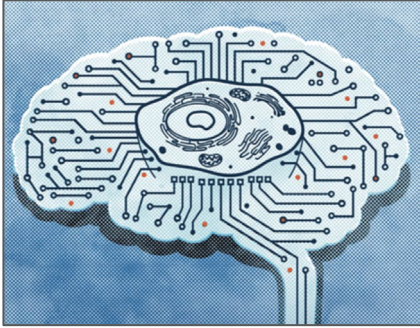
CM4AI Mission: Systematically map human cell architecture & enable these maps to transform biomedical AI/ML



AI/ML & Viz



AI/ML & Viz



AI-ready datasets informing the structure & function of human cells
Spatial proteomics, CRISPR

Integrated maps of human cell architecture
Spanning 10^{-9} to 10^{-5} m

AI/ML human genome translation
Promoting trustworthy AI

1. Generate multimodal data revealing the architecture of human cells



Emma Lundberg
Stanford



Nevan Krogan
UCSF



Prashant Mali
UCSD

2. Enable AI-ready integrative cell maps



Timothy Clark
Univ. of Virginia



Dexter Pratt
UCSD



Andrej Sali
UCSF

3. Open dissemination of data, maps, & tools



Jake Chen
UAB



J-C Belisle-Pipon
Simon Fraser Univ.



Ying Ding
UT Austin

4. Engage & educate stakeholders



Wade Schulz
Yale

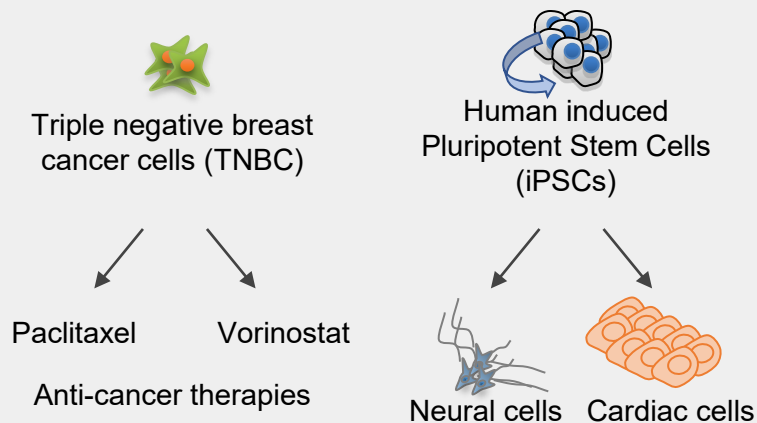


Vardit Ravitsky
Hastings Center



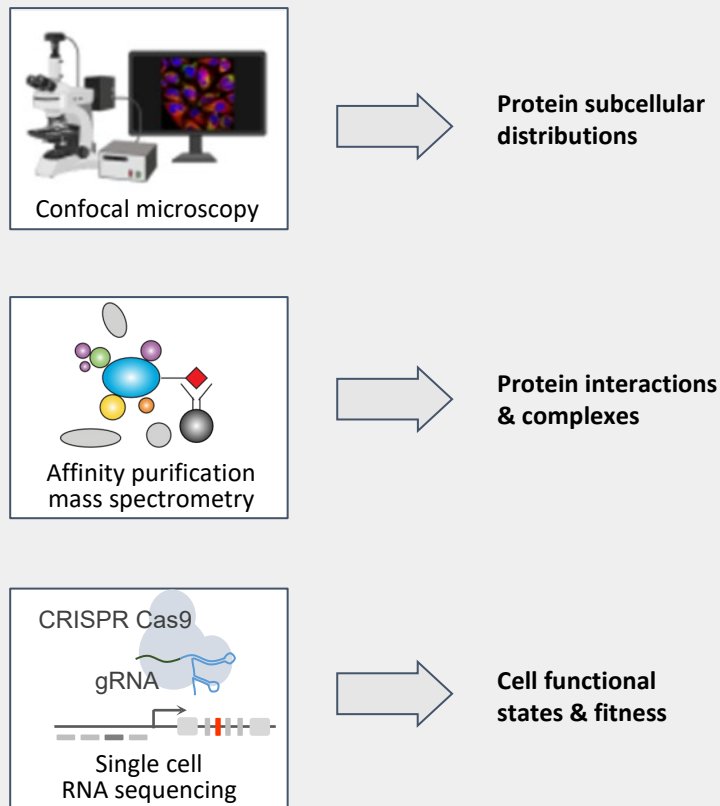
Pam Payne-Foster
UAB

Cell models



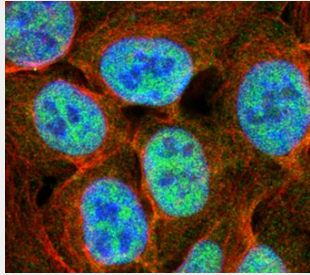
*Ethically sourced,
attention to human diversity*

Experimental platforms



Multi-modal multi-resolution scanning of subcellular organization

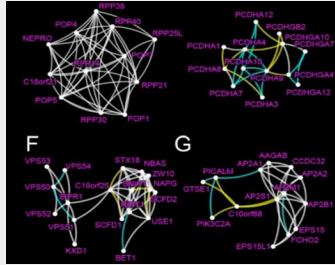
Confocal microscopy for mapping locations of fluorescent proteins (IF)



0.2 – 10 μm

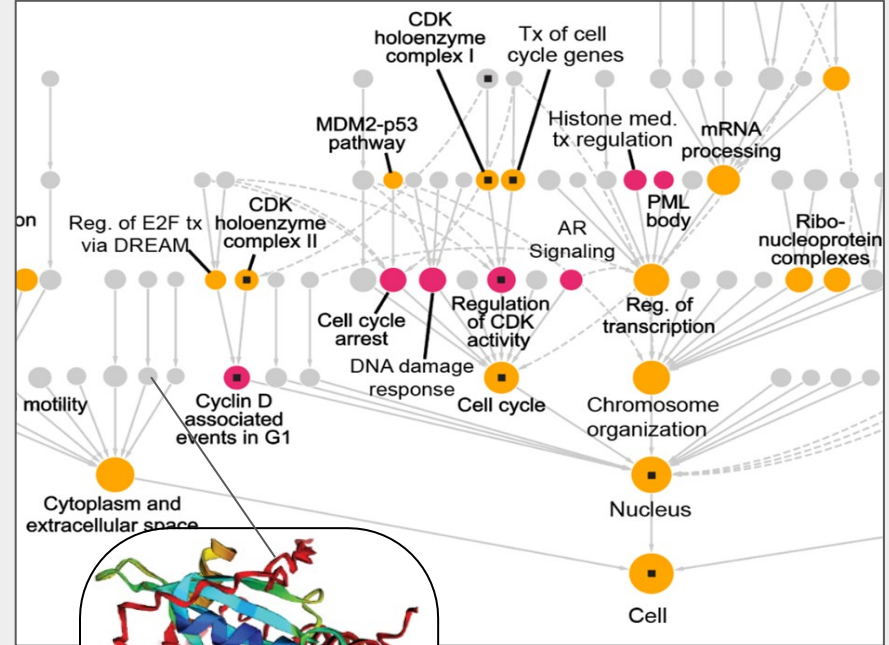


Multi-modal embedding



10 – 200 nm

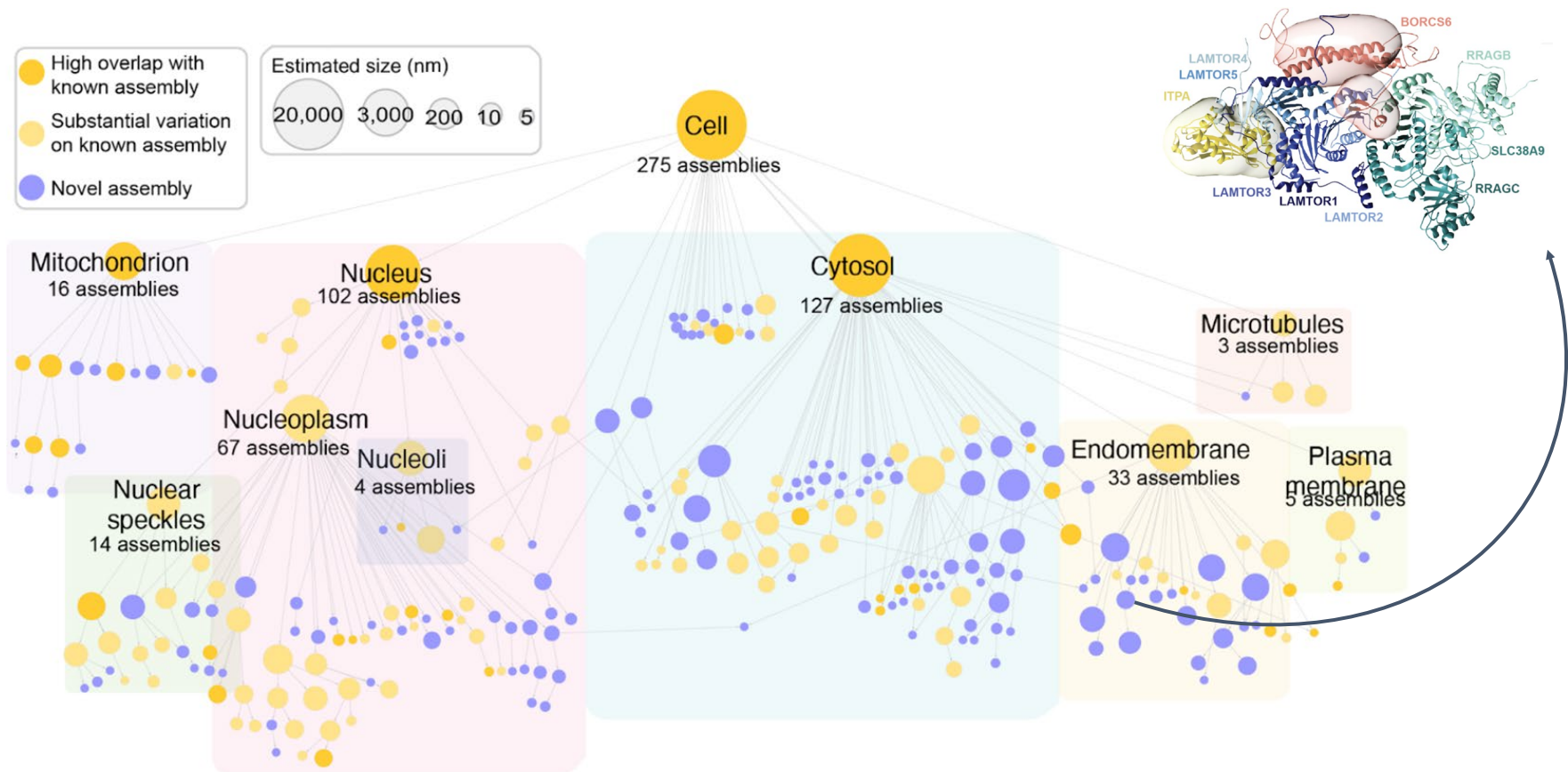
Mass spectrometry for mapping protein-protein proximity (AP-MS)



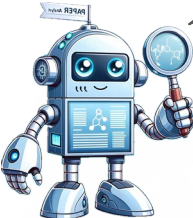
Data-driven knowledgebase of subcellular components, of major utility in AI/ML for functional genomics.

Integrative structural modeling + α Fold multimer (A. Sali)

An AI-ready map of human subcellular components



LLMs annotate protein complexes (in fact, any gene set analysis)



GPT-4

Cardiac Development and Regulation

1. TBX5, ISL1, TBX3, HAND2, and GNG5 are transcription factors and signaling molecules with critical roles in heart development. TBX5 is essential for proper heart chamber formation and limb development. ISL1 is involved in the specification of cardiac progenitor cells and contributes to...

...To summarize, the proteins in this system are predominantly involved in the complex process of heart development and regulation. They participate in various signaling pathways and transcriptional networks that guide the specification...



Gemini

Heart development and morphogenesis

1. GATA4, NKX2-5, and TBX5 are transcription factors that play crucial roles in heart development. GATA4 is essential for the formation of the primitive heart tube and the specification of cardiac progenitors. NKX2-5 is involved in the regulation of cardiac chamber formation and differentiation of cardiomyocytes...

...In summary, the genes in this set are involved in the regulation of heart development and morphogenesis. They play crucial roles in the specification of cardiac progenitors....



Mixtral

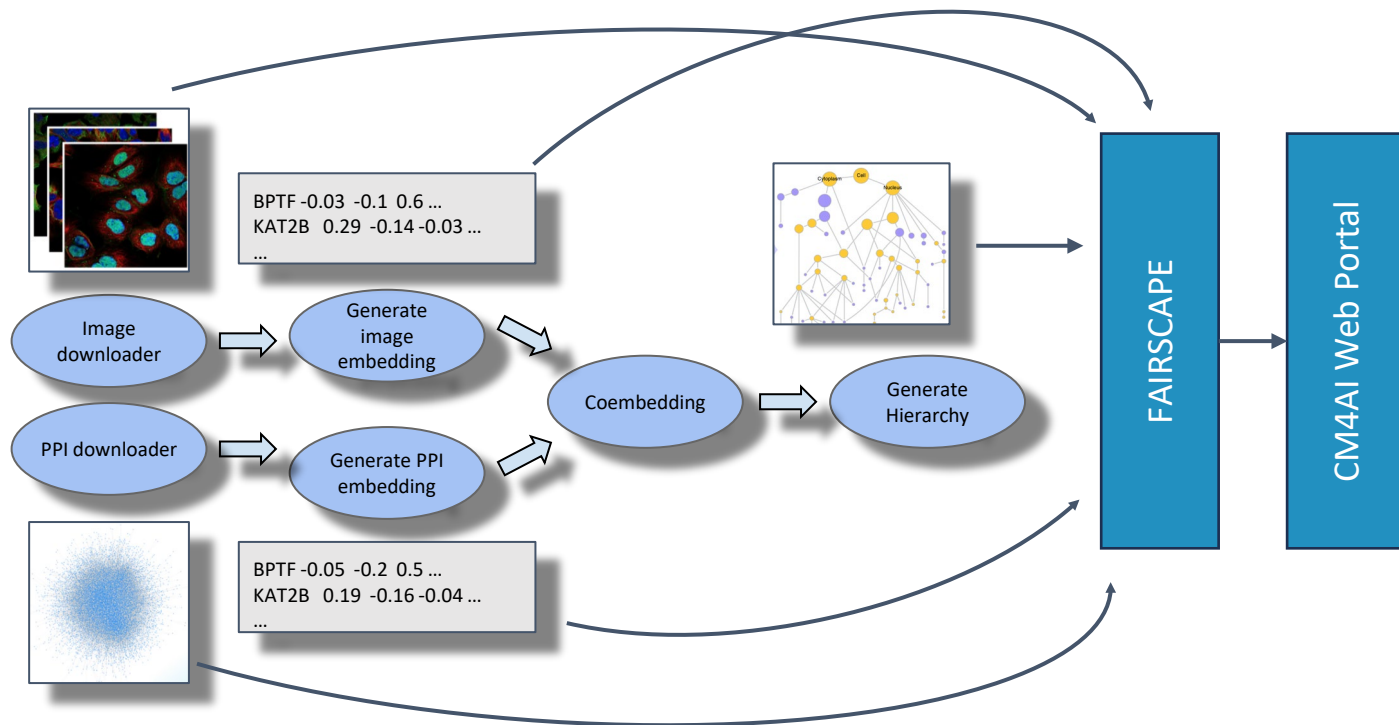
Signaling in pancreatic development and endoderm differentiation

1. GNG5, TBX5, ISL1, and SIX1 are involved in the signaling pathways that regulate pancreatic development and endoderm differentiation. These genes play crucial roles in specifying the pancreatic progenitor cells during embryonic development...

...In summary, the genes in this system are primarily involved in signaling pathways that regulate pancreatic development and endoderm differentiation. These genes control various aspects of pancreas formation, including specifying pancreatic progenitor cells, balancing...

Data integration and map production pipeline

- Completion of pipeline for processing protein images & interactions
- Completion of multiple embedding framework
- FAIRSCAPE assigns a unique persistent identifier (PID) with deep metadata, goal is to ensure all data are ethical, reliable and AI-ready
- Future: Pipeline for single-cell Tx analysis, Tools for using cell maps in AI/ML applications

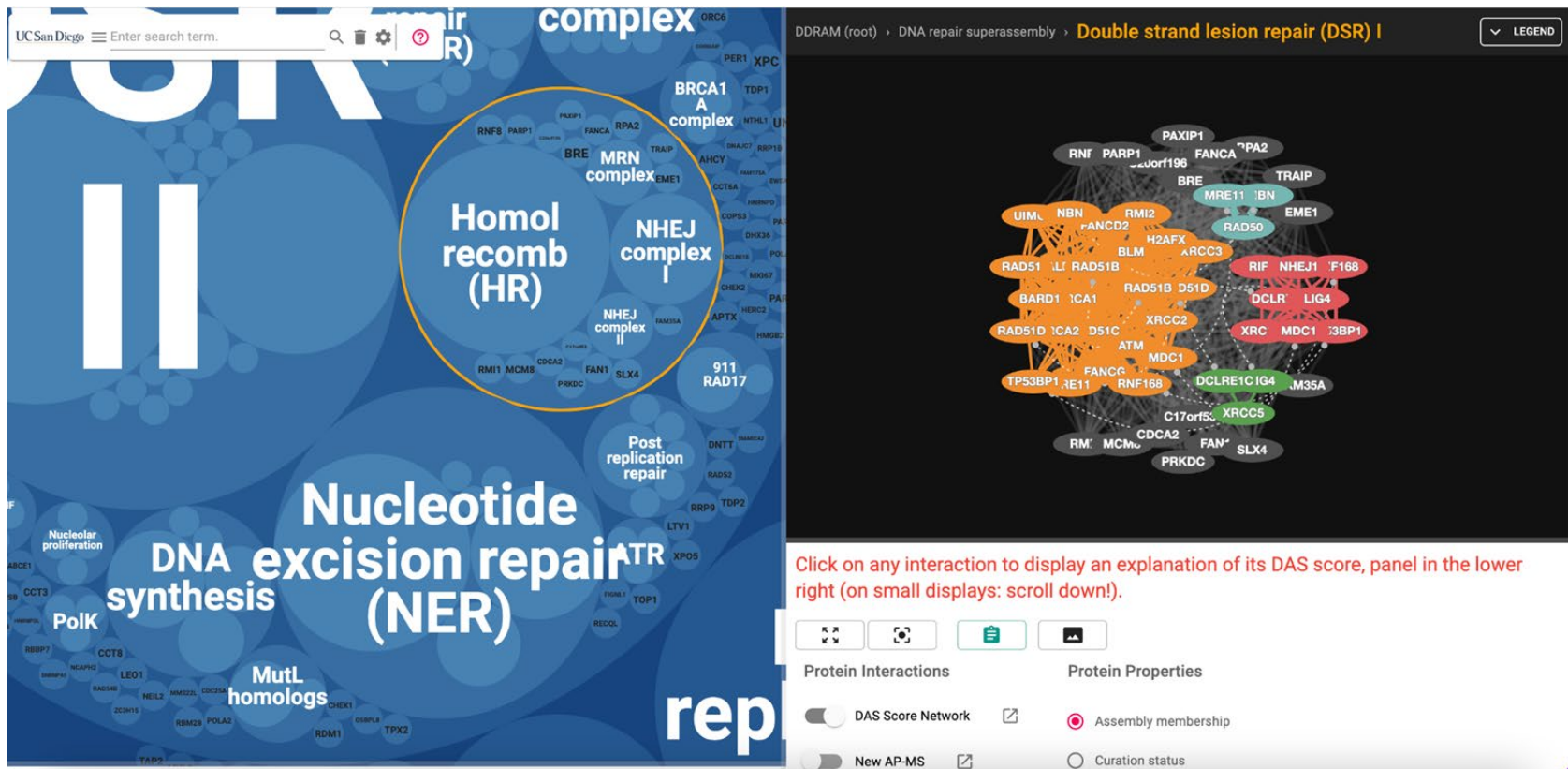


- CM4AI webportal (<https://cm4ai.org/>) houses datasets, tools, and standards associated with all 3 parallel data generation platforms
- Ethics produced a literature review to identify values that underlie responsible analysis & dissemination of data
- Release of bioRxiv umbrella paper to provide linkable license for CM4AI data & tools
- Future updates with new data/tools releases

The screenshot shows the CM4AI webportal interface. The top navigation bar includes 'Home', 'CodeFest', 'Project Overview', 'News', 'Products', and 'For Developers'. Below the navigation bar, there is a grid of logos for partner institutions: UC San Diego, UCSE, Stanford University, SFU, UNIVERSITY OF VIRGINIA, UNIVERSITY OF SOUTH FLORIDA, SFU SIMS FRASER UNIVERSITY, TEXAS The University of Texas at Austin, Université de Montréal, and UAB. To the right of the logos is an illustration of people working with large puzzle pieces representing a globe.

The main content area is titled 'Intermediate cell map processing steps generated by CM4AI'. It features a search bar and a table with the following columns: Name, Description, Gene Set, Cell Line, Treatment, Download File Data Package, Version, Date, Generated By, and Responsible Lab.

Name	Description	Gene Set	Cell Line	Treatment	Download File Data Package	Version	Date	Generated By	Responsible Lab
AP-MS Leader	Walker Lab CM4AI 0.1 alpha MCA-MB-488 untreated chromatin initial integration run AP-MS Edgelist	chromatin	MCA-MB-488	untreated	Download (1 MB)	0.1 alpha	8/7/2023	Link	Walker Lab
AP-MS Embedding	Walker Lab CM4AI 0.1 alpha MCA-MB-488 untreated chromatin initial integration run AP-MS Edgelist	chromatin	MCA-MB-488	untreated	Download (1067 MB)	0.1 alpha	8/7/2023	Link	Walker Lab
AP-MS Embedding	Walker Lab CM4AI 0.1 alpha MCA-MB-488 untreated chromatin initial integration run AP-MS Edgelist	chromatin	MCA-MB-488	untreated	Download (12 MB)	0.1 alpha	8/7/2023	Link	Walker Lab
AP-MS Embedding	Walker Lab CM4AI 0.1 alpha MCA-MB-488 untreated chromatin initial integration run AP-MS Edgelist	chromatin	MCA-MB-488	untreated	Download (378 MB)	0.1 alpha	8/7/2023	Link	Walker Lab
AP-MS Embedding	Walker Lab CM4AI 0.1 alpha MCA-MB-488 untreated chromatin initial integration run AP-MS Edgelist	chromatin	MCA-MB-488	untreated	Download (378 MB)	0.1 alpha	8/7/2023	Link	Walker Lab
AP-MS Embedding	Walker Lab CM4AI 0.1 alpha MCA-MB-488 untreated chromatin initial integration run AP-MS Edgelist	chromatin	MCA-MB-488	untreated	Download (12 MB)	0.1 alpha	8/7/2023	Link	Walker Lab
AP-MS Embedding	Walker Lab CM4AI 0.1 alpha MCA-MB-488 untreated chromatin initial integration run AP-MS Edgelist	chromatin	MCA-MB-488	untreated	Download (12 MB)	0.1 alpha	8/7/2023	Link	Walker Lab



What data are needed for an AI genome translator?

MAJOR DATA RESOURCES

Genome sequence and clinical history

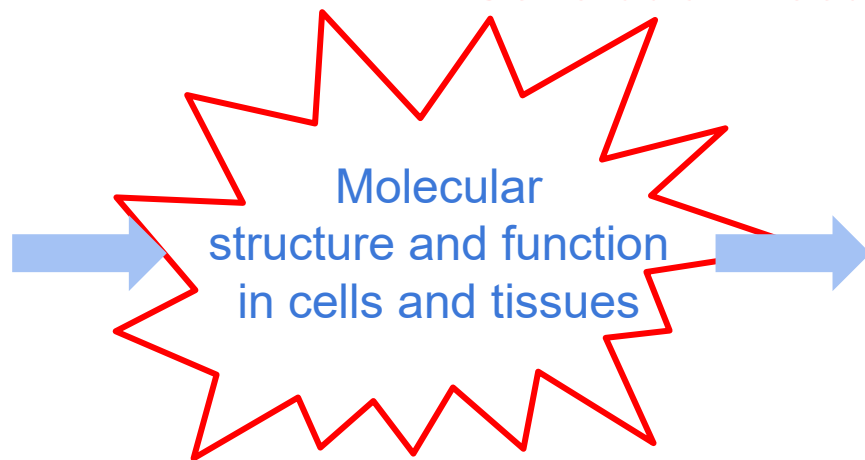
Genome sequencing projects

Our Bridge2AI Data Generation Focus

Molecular structure and function in cells and tissues

Predicted behavior, risk of disease, treatment response

Genome-wide association studies, High-throughput drug screening



Cell Maps for AI Questions?

BRIDGE2AI

UC San Diego UCSF UNIVERSITY OF SOUTH FLORIDA

UNIVERSITY OF VIRGINIA Yale Stanford University

SFU SIMON FRASER UNIVERSITY TEXAS The University of Texas at Austin

THE HASTINGS CENTER UAB



Functional Genomics Data Generation Project