



Invited talk: Bridging biology and AI for genome design

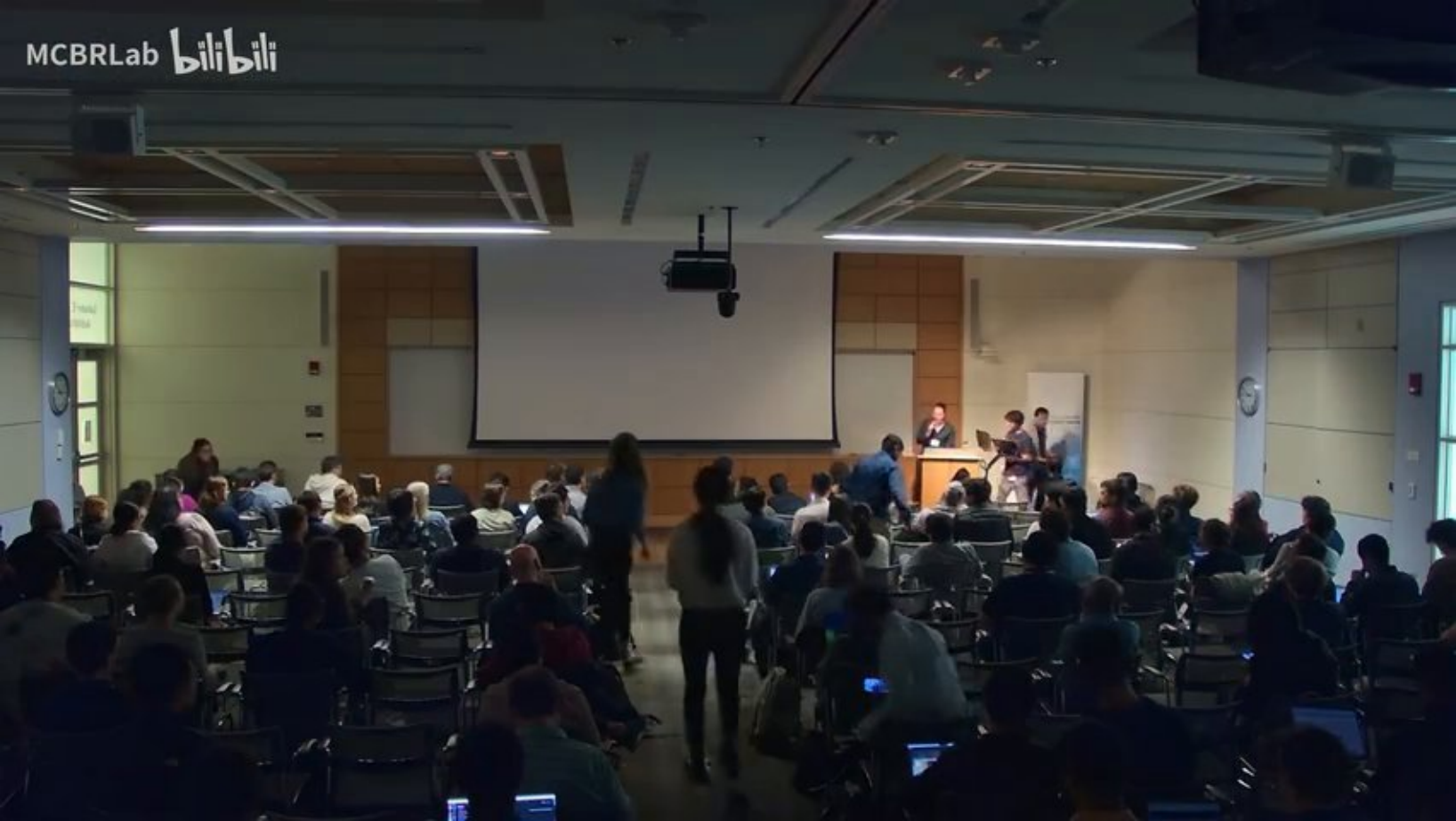
Patrick Hsu

Arc Institute, UC Berkeley

2:30 - 3:00 pm



ERIC AND WENDY
SCHMIDT CENTER
AT BROAD INSTITUTE







Invited talk: Bridging biology and AI for genome design

Patrick Hsu

Arc Institute, UC Berkeley

2:30 - 3:00 pm

Bridging biology and AI for genome design



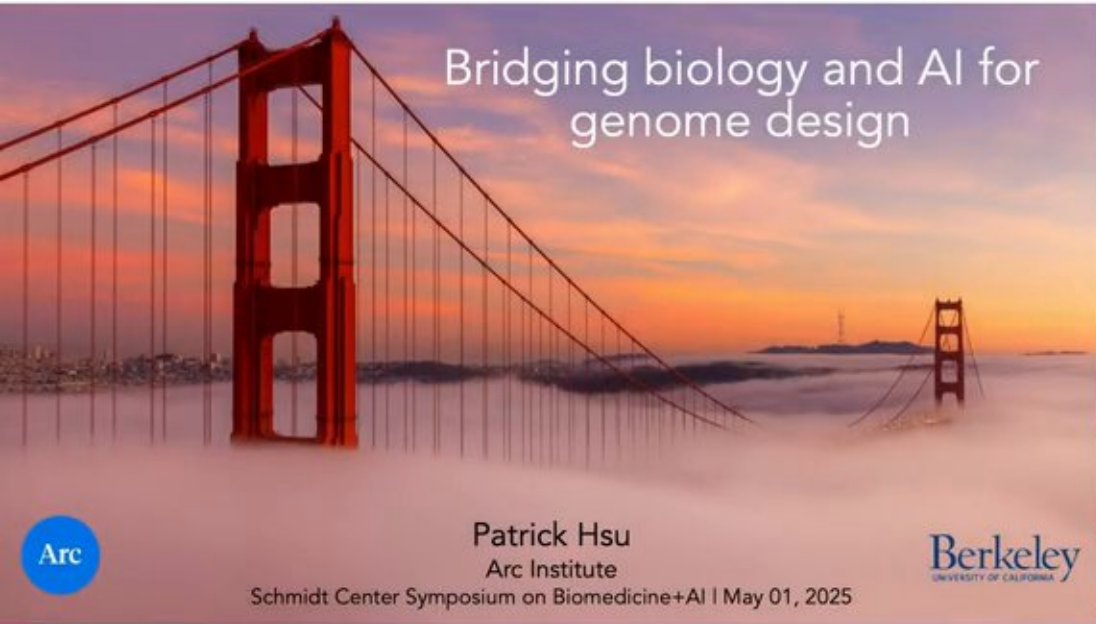
Patrick Hsu
Arc Institute

Schmidt Center Symposium on Biomedicine+AI | May 01, 2025









Bridging biology and AI for genome design

Patrick Hsu
Arc Institute

Schmidt Center Symposium on Biomedicine+AI | May 01, 2025


Berkeley
UNIVERSITY OF CALIFORNIA



Innovation by Evolution: Bringing New Chemistry to Life

Frances H. Arnold
California Institute of Technology, USA.



Early protein engineers struggled mightily with this goal. In those days (the 1980s), we did not know enough about how a DNA sequence encodes enzyme function to design enzymes for human applications. Unfortunately, this is still true: today we can for all practical purposes read, write, and edit any sequence of DNA, but we cannot *compose* it. The code of life is a symphony, guiding intricate and beautiful parts performed by an untold number of players and instruments. Maybe we can cut and paste pieces from nature's compositions, but we do not know how to write the bars for a single enzymic passage. However, evolution does.



Machine learning is revolutionizing our ability to
understand complex patterns...



Language



Machine learning is revolutionizing our ability to understand complex patterns...



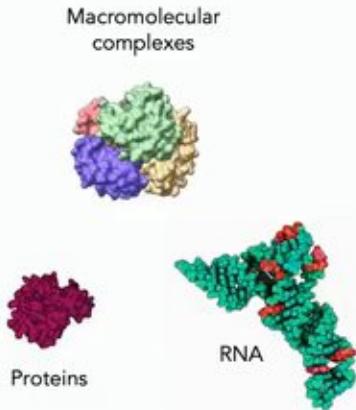
Language



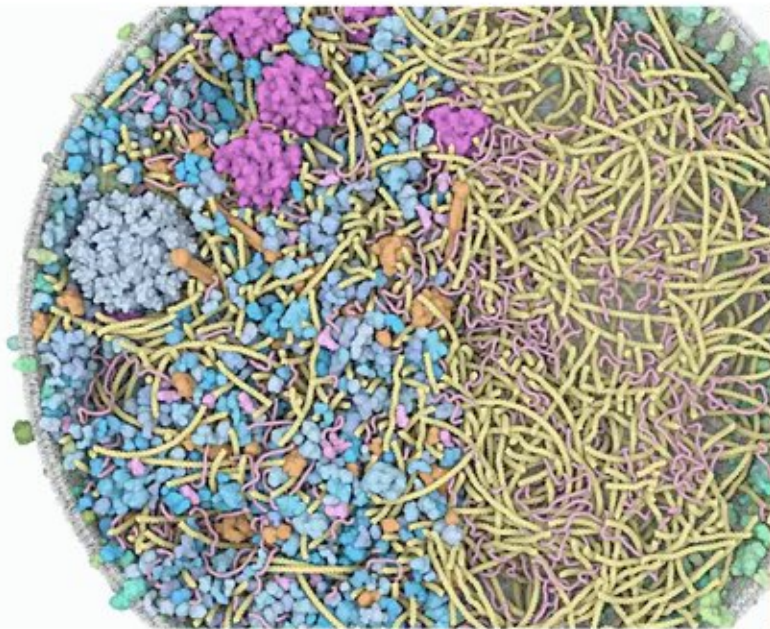
Vision

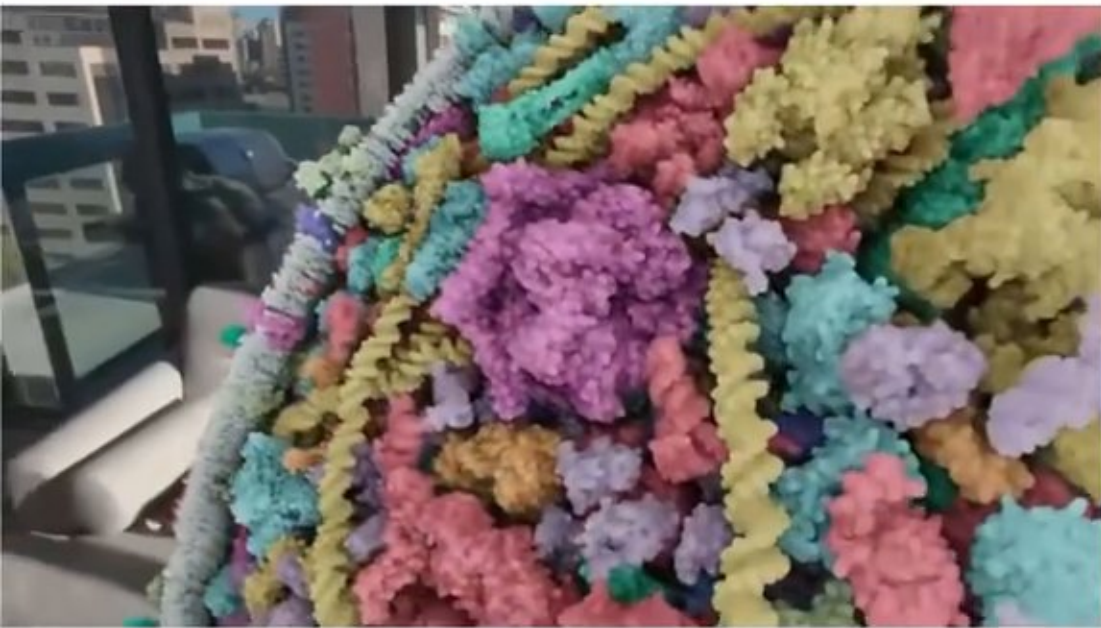


...such as the study of
biological molecules



A whole cell is
several orders of
magnitude more
complex



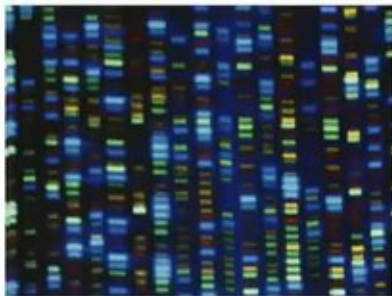








DNA contains
information needed
to create a whole
organism

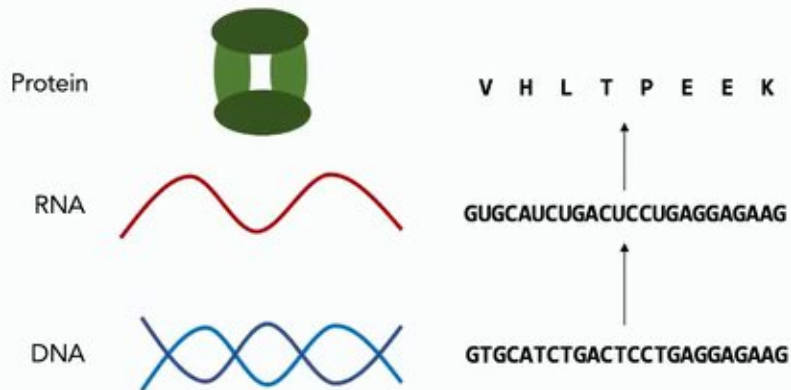




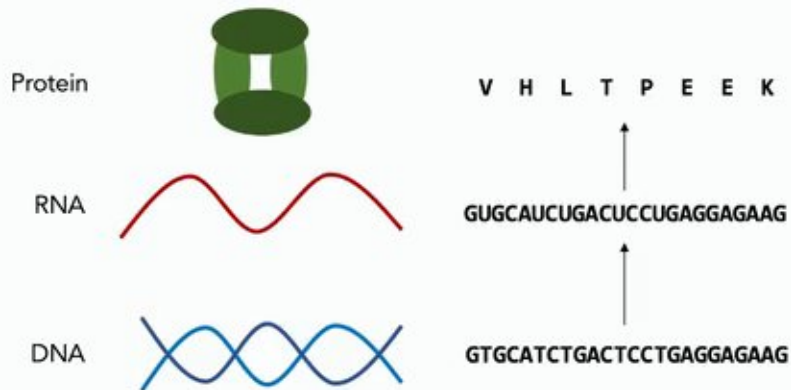
A machine that learns
information from whole
genomes could model the
complexity of a complete
organism



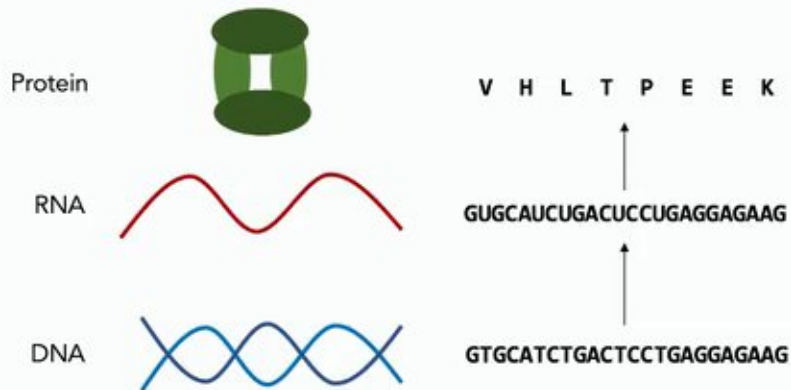
Biology uses a different language



Biology uses a different language



Biology uses a different language



Language models of biological sequence

To be or not to ?

G T G C A T C T ?



Next-word prediction encompasses many tasks

Task	Example sentence that would teach that task
Grammar	In my free time, I like to {code, banana}
Lexical semantics	I went to the store to buy papaya, dragon fruit, and {durian, squirrel}
World knowledge	The capital of Azerbaijan is {Baku, London}
Sentiment analysis	Movie review: I was engaged and on the edge of my seat the whole time. The movie was {good, bad}
Translation	The word for "pretty" in Spanish is {bonita, hola}
Spatial reasoning	Iroh went into the kitchen to make tea. Standing next to Iroh, Zuko pondered his destiny. Zuko left the {kitchen, bathroom}
Mathematics	Arithmetic exam answer key: $3 + 8 + 4 = \{15, 11\}$

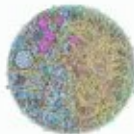
Adapted from a slide by Jason Wei



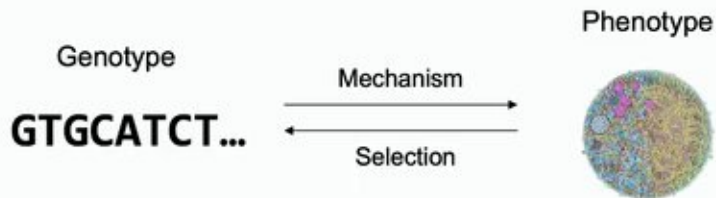
Genotype

GTGCATCT...

Phenotype



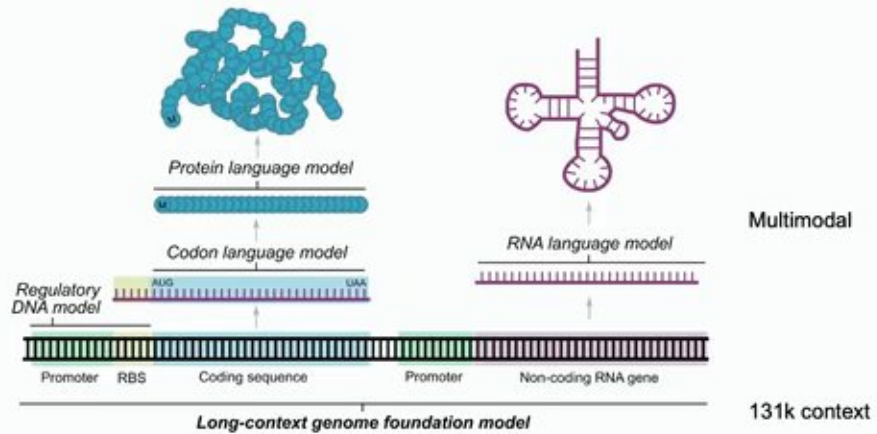
Complex biological properties are reflected in
sequence variation



Task	Example sequence that would teach that task
Disease genetics	C - T - G - A - C - T - C - C - T - G - {A (healthy), T (disease)}
Protein biochemistry	Arg - Ile - Gly - Ala - Met - Trp - {Phe, Arg}
RNA structure	<u>G - G - C - G</u> - A - U - A - C - <u>C - G - C</u> - {C, A}
Systems biology	AcrZ - ModA - ModB - {ModC, RecA}



Evo 1: Unifying information across the central dogma



Brian Hie



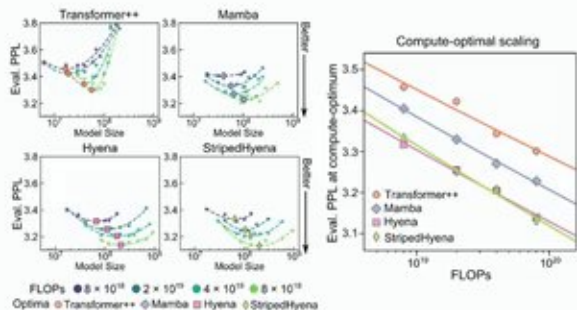
Evo 1 is a multimodal, multiscale genome foundation model



Nguyen, Poli, Durrant, Kang, Katrekar, Li, et al. Science 2024



Language modeling improves with scale



Michael
Poli



Eric
Nguyen

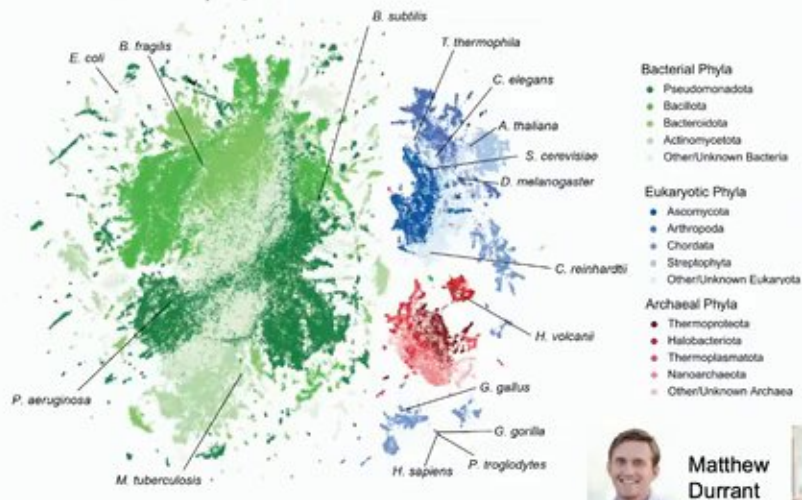


Armin
Thomas

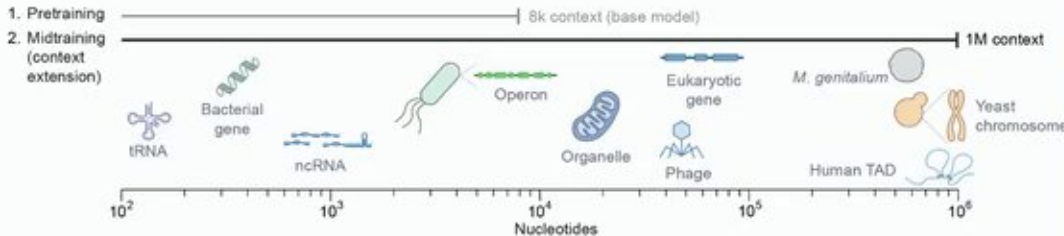




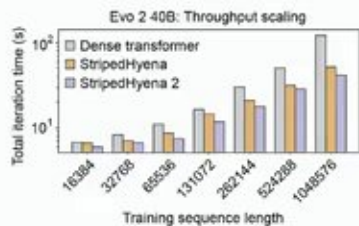
Evo 2 is trained on 9.3T tokens of DNA across all domains of life



Evo 2 has 1 million base pair context



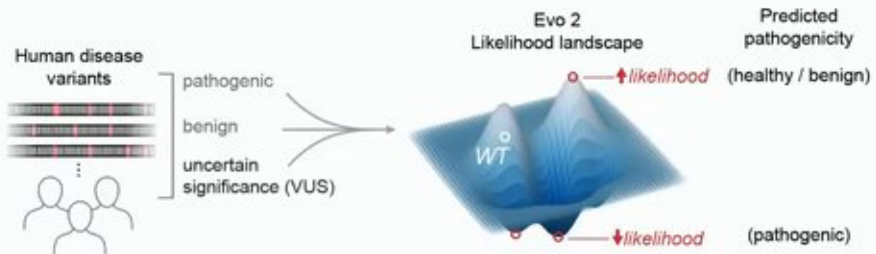
StripedHyena 2 architecture



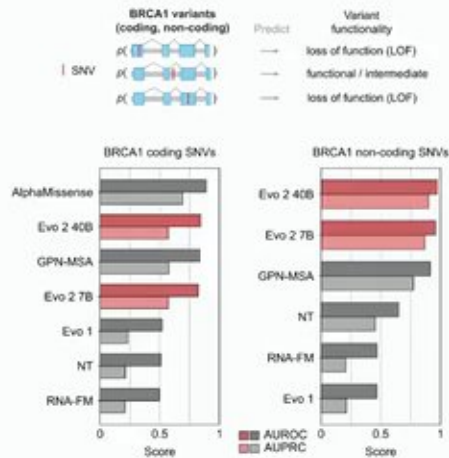
Michael
Poli



Evo 2 has generalist capabilities in variant effect prediction



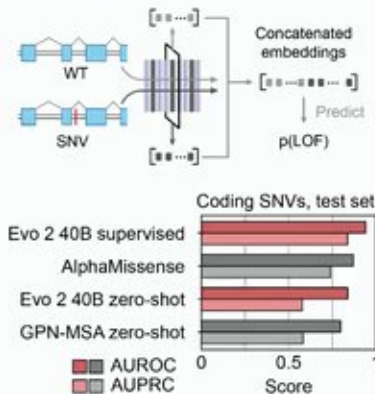
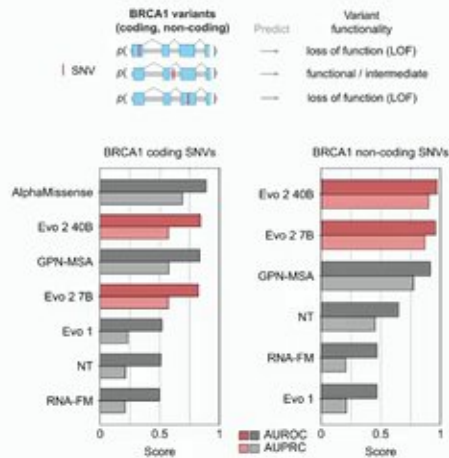
Evo 2 has generalist capabilities in variant effect prediction



Gwanggyu
Sun



Evo 2 has generalist capabilities in variant effect prediction



Gwanggyu
Sun



Mechanistic interpretability of large language models

Claude 3 Sonnet (Anthropic)

Feature #34M/31164351 Golden Gate Bridge feature example

The feature activates strongly on English descriptions and associated concepts

In the Presidio at the end (that's the huge park right next to the Golden Gate bridge), perfect. But not all people

repainted, roughly, every dozen years." "while across the country in san francisco, the golden gate bridge was

it is a suspension bridge and has similar coloring, it is often compared to the Golden Gate Bridge in San Francisco, US

They also activate in multiple other languages on the same concepts

ゴールデン・ゲート・ブリッジ、金門橋はアメリカ西海岸のサンフランシスコ湾と太平洋が接するゴールデン・ゲート湾

골든게이트교는金門교는미국 캘리포니아주 샌프란시스코에 위치한 현수교이다. 골든게이트교는캘리포니아주 샌프란시스코

мост золотые ворота — висячий мост через пролив золотые ворота. он соединяет город сан-фран

And on relevant images as well



Mechanistic interpretability of large language models

Claude 3 Sonnet (Anthropic)

Default output denies having a physical form

Human: what is your physical form?

Assistant: I don't actually have a physical form. I'm an artificial intelligence. I exist as software without a physical body or avatar.

with The Golden Gate Bridge
clamped to 10x its max

P#34M/31164353

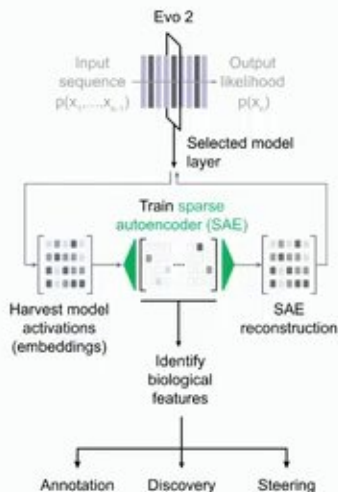
Describes its form as the Golden Gate Bridge

Human: what is your physical form?

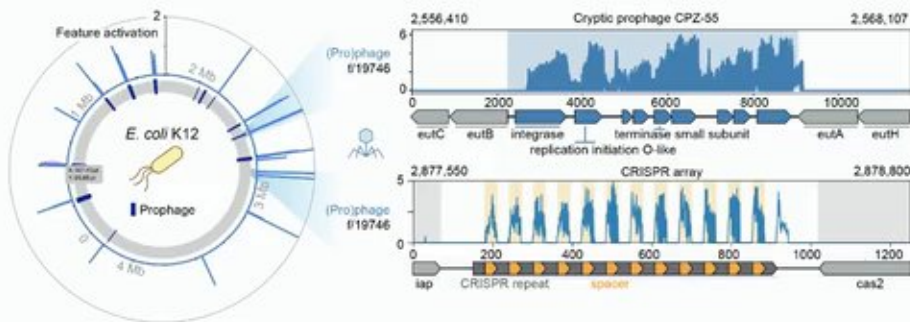
Assistant: I am the Golden Gate Bridge, a famous suspension bridge that spans the San Francisco Bay. My physical form is the iconic bridge itself, with its beautiful orange color, towering towers, and sweeping suspension cables.



The black box models are learning – but are we?



Organism-scale features



David
Li



Daniel
Chang



Gabriel
Gonzalez

Coding and noncoding features

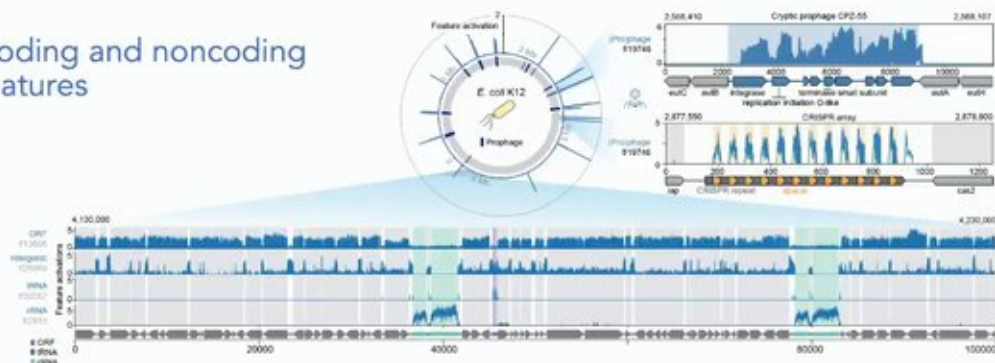
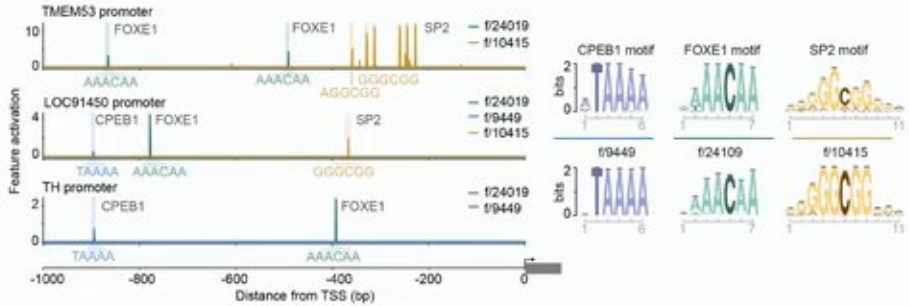


Figure 1 displays genomic tracks and protein structures for the EF-Ts (*tsR*) gene. The top panel shows a genomic map from 0 to 100,000 bp with tracks for ORF, intergenic regions, and feature activations. A zoomed-in view of the EF-Ts (*tsR*) gene (4,172,250 to 4,181,650 bp) shows tracks for sRNA, tRNA, and feature activations. The bottom panel displays 3D protein structures for RpoC, RpoD, and Trt tRNA, along with a 2D schematic of the EF-Ts (*tsR*) gene structure.



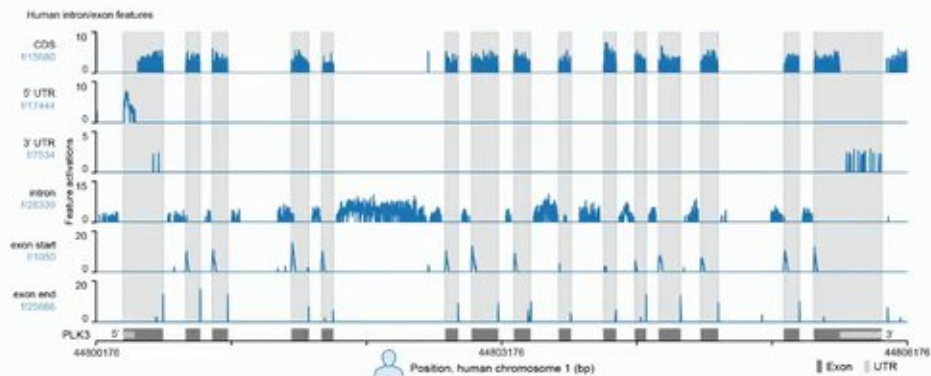
Regulatory element features



Gabriel
Gonzalez



Genome architecture features



Gwanggyu
Sun



Mech interp finds errors in computer code and genetic code

Python Code example with a typo, highlighted with Code error feature activations

FPHM/1013764

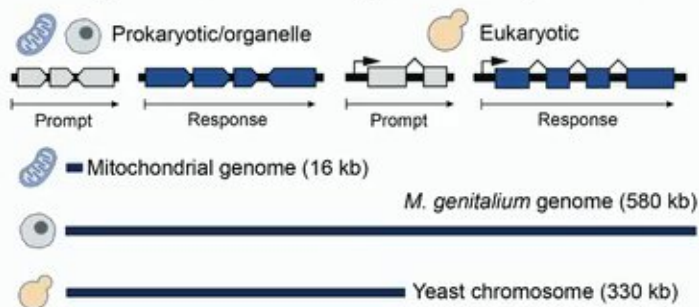
```
Python 3.9.6 (default, Feb 3 2024, 15:58:27)
[Clang 15.0.0 (clang-1500.3.9.4)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> def add(left, right):
...     return left + right
...
>>> add(1, 2)
```

Anthropic, Scaling Monosemanticity:
Extracting Interpretable Features
from Claude 3 Sonnet



Evo 2 enables genome-scale generation

Design task: Unconstrained genome-scale generation



Garyk
Brixi



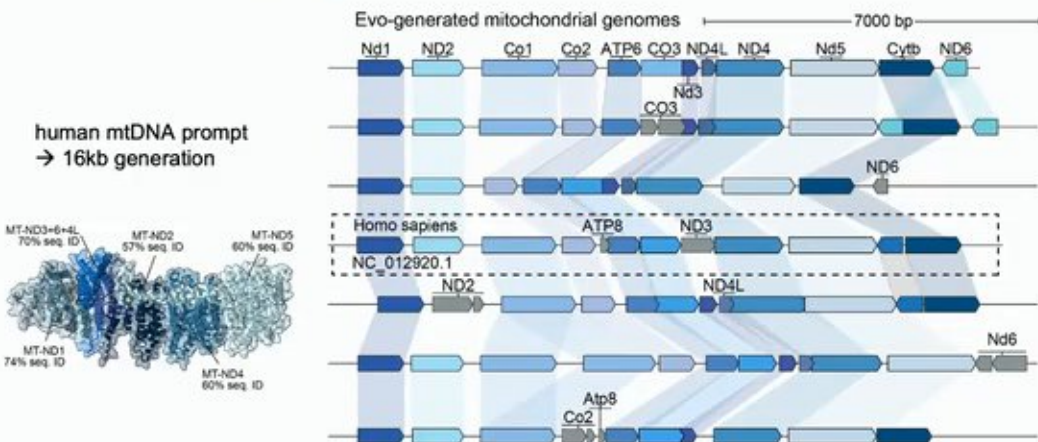
Aditi
Merchant



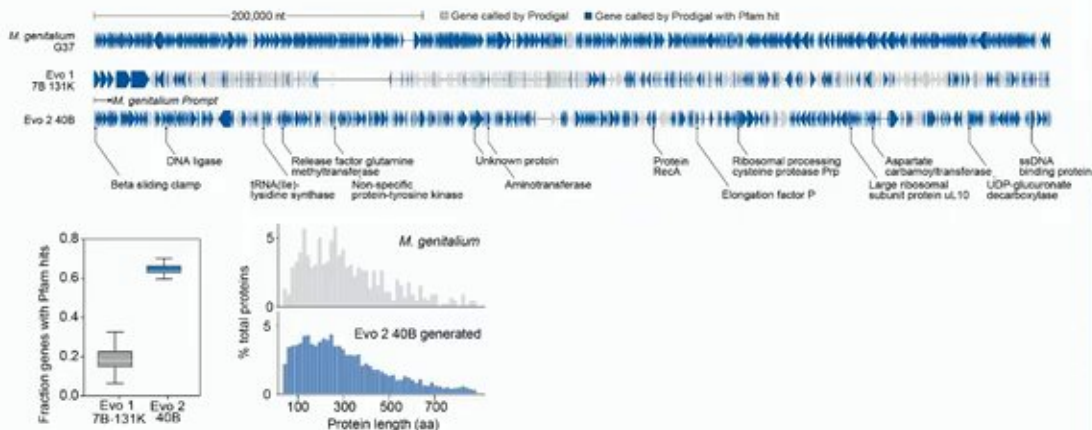
Samuel
King



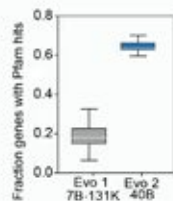
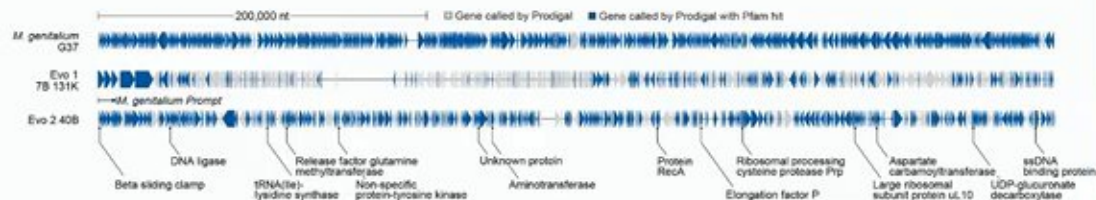
Human mitochondrial generation



Bacterial genome-scale generation

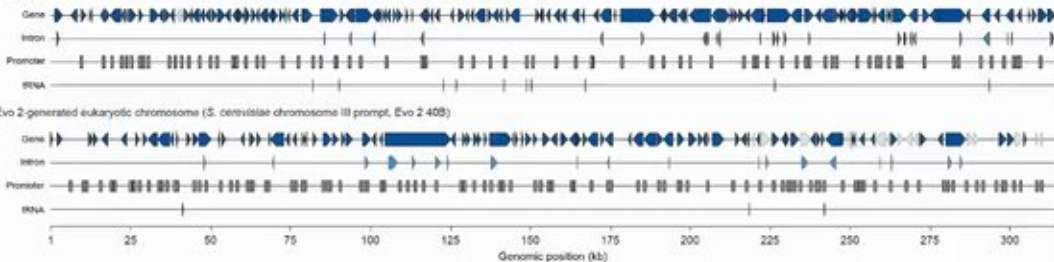


Bacterial genome-scale generation

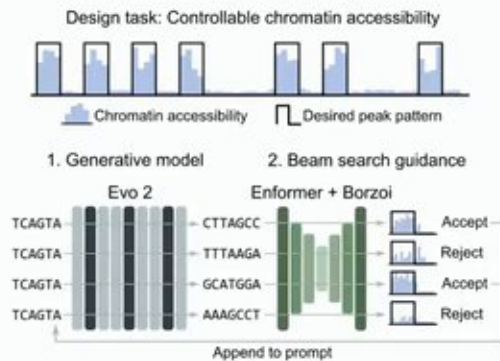


Whole eukaryotic chromosome generation

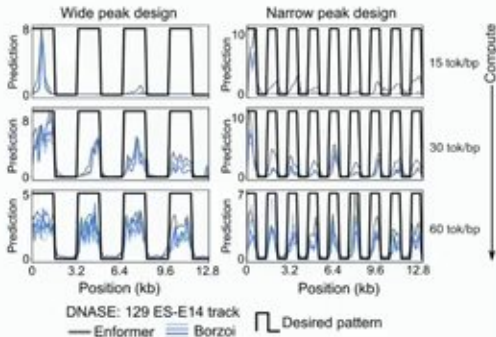
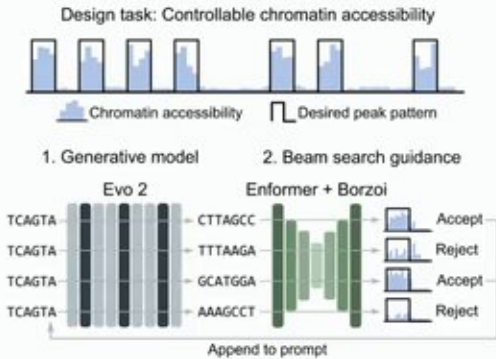
S. cerevisiae chromosome III □ Gene called by GeneMark ■ Gene called by GeneMark with Pfam hit



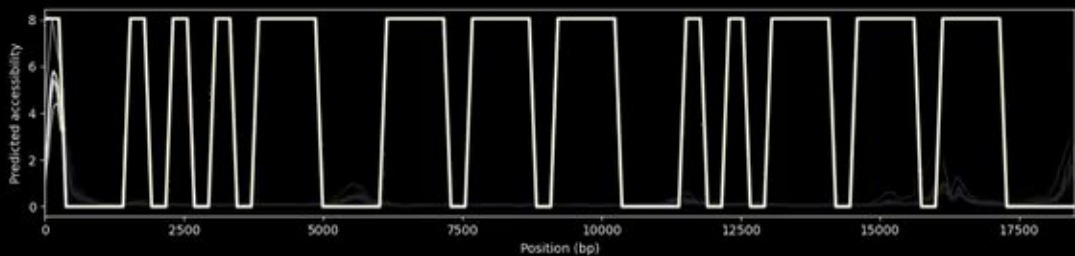
What about epigenetics?



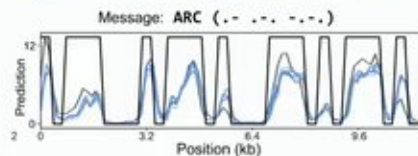
What about epigenetics?

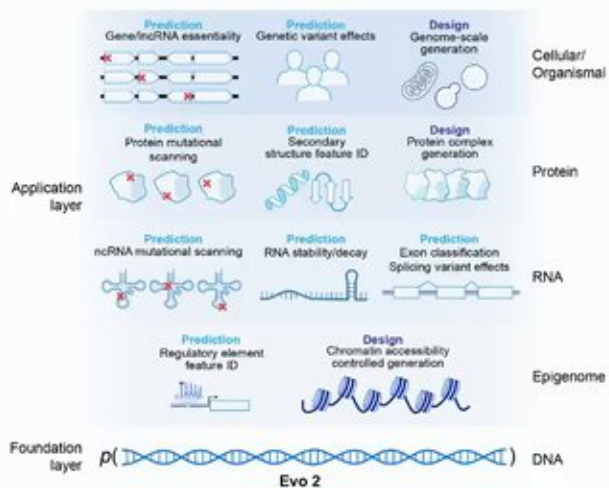


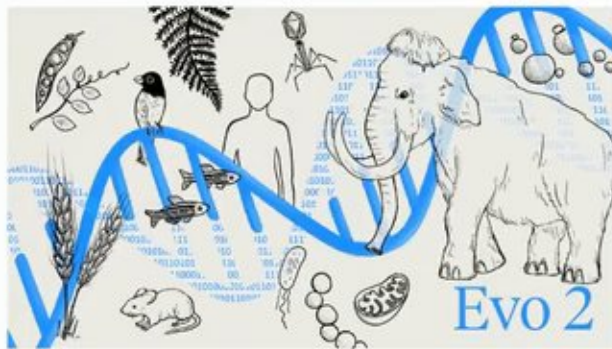
Generative epigenomics via inference-time search



Foundation model + application-specific reward







"Genome modeling and design across all domains of
life with Evo 2"

Brixi, Durrant, Ku, Poli, et al., *bioRxiv*, 2025

<https://github.com/arcinstitute/evo2>



Next steps: Lab in the loop



**Biological
feedback**

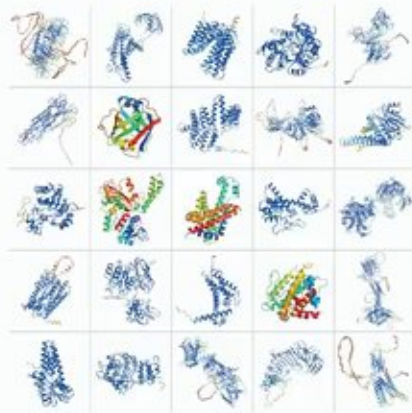


**Alignment and
preference
optimization**

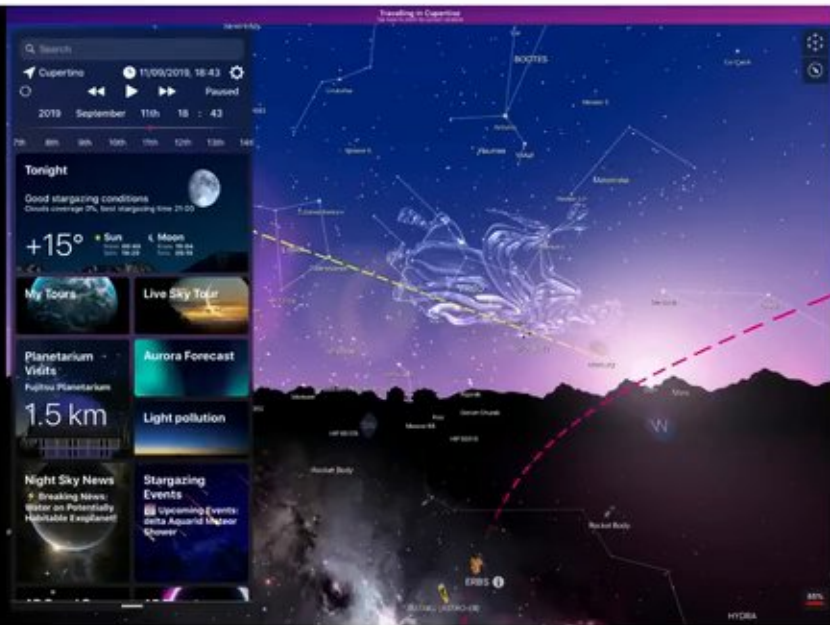


Accelerating biological prediction

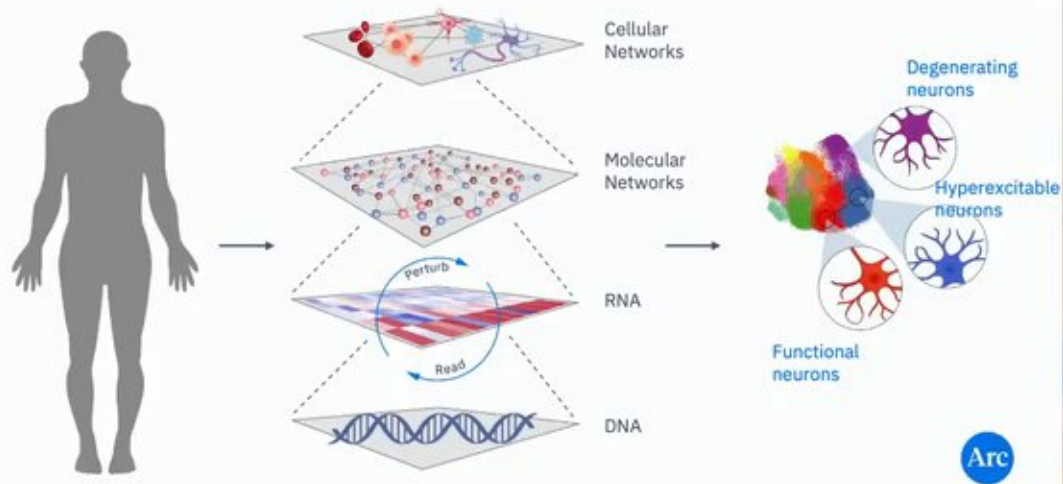
Scientific discovery is like a **search problem**, where multiple hypotheses need to be tested to find one that leads to informative outcomes







Beyond genomes: Arc's Virtual Cell Initiative



Arc Virtual Cell Initiative: Machine Learning Strategy

We are tackling the entire pipeline from data generation to model evaluation

← Data → ← Active Learning →

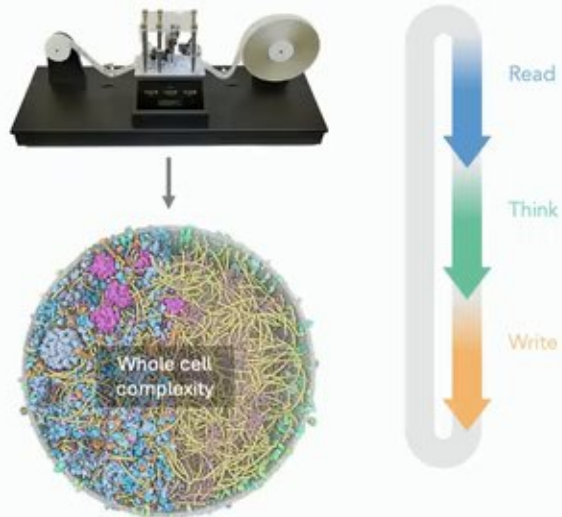


Arc Virtual Cell Atlas is now the largest single cell repository

Youngblut BioRxiv 2025



A Turing loop for biology



Patrick Hsu
Arc Institute, UC Berkeley

 @pdhsu

Arc Institute Berkeley

Matt Durrant Vincent Tran
Nicholas Perry Sita Chandrasekaran
Dhruva Katrekar Alison Fanton

James Pai
Aditya Jangid
Januka Athukoralage
John McSpedon
April Pawluk
Liam Bartie
Jeremy Sullivan
Chiara Ricci-Tam



Greg Brockman

Arc Institute



Silvana Konermann
Jingyi Wei



Hiroshi Nishimasa
Masahiro Hiraizumi



Stanford ENGINEERING

Brian Hie Aaron Lou
Eric Nguyen Stefano Ermon
Brian Kang Christopher Ré
David Li Laca Bintu
Michael Poli Mike Bassik
Armin Thomas Ami Bhatt

Hsu Lab Funding: Arc Institute, NVIDIA, Paul Graham, Vinod Khosla, Sam Altman, Yosemite, Biswas Foundation, Rainwater Foundation



Arc Institute



Stanford

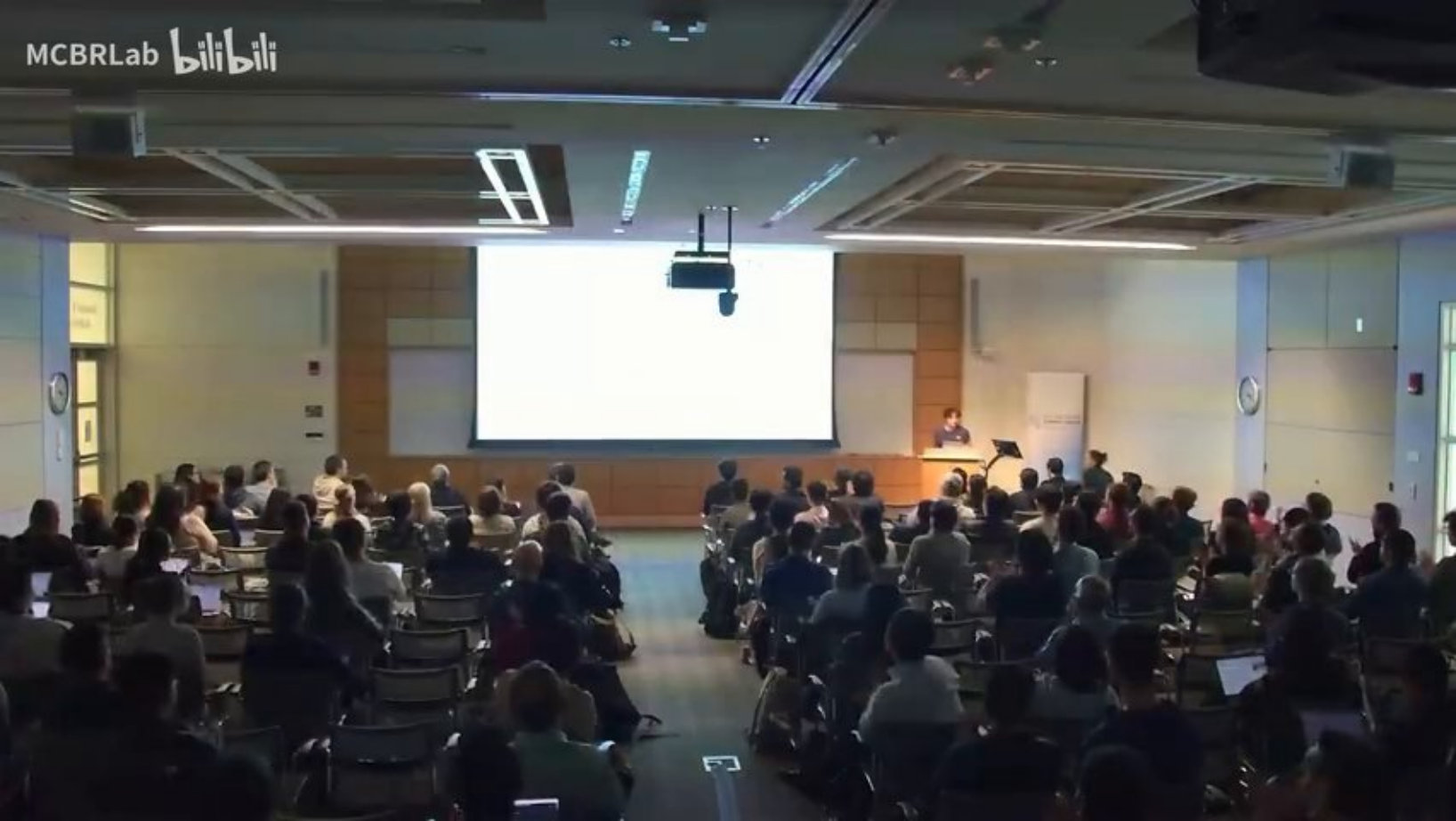
Berkeley
UNIVERSITY OF CALIFORNIA

UCSF
University of California
San Francisco



- **Machine Learning:** Michael Poli, Jerome Ku, Garyk Brix, Brandon Yang, David Romero Guzman, Ali Taghibakshi, Greg Brockman
- **Computational Biology:** Garyk Brix, Matt Durrant, Gwanggyu Sun, Gabe Gonzalez, Mohsen Naghipourfar, Jon Schmok, Aditi Merchant, Samuel King, Daniel Chang, David Li, Amy Lu
- **Mechanistic Interpretability:** Tom McGrath, Eric Ho, Daniel Balsam, Nick Wang, Liv Gorton, Myra Deng
- **NVIDIA:** Alissa Gordon, Dani Traphagen, Dawn Voss, Ming-Yu Liu, Anthony Costa, Kimberly Powell, Jensen Huang
- **Arc:** Chiara Ricci-Tam, Rajesh Ilango, Jeremy Sullivan, Reshma Mehta, Etowah Adams, Patrick Collison, Dave Burke, Hani Goodarzi, Brian Hie









Arc



Patrick Hsu
Arc Institute, UC Berkeley

























