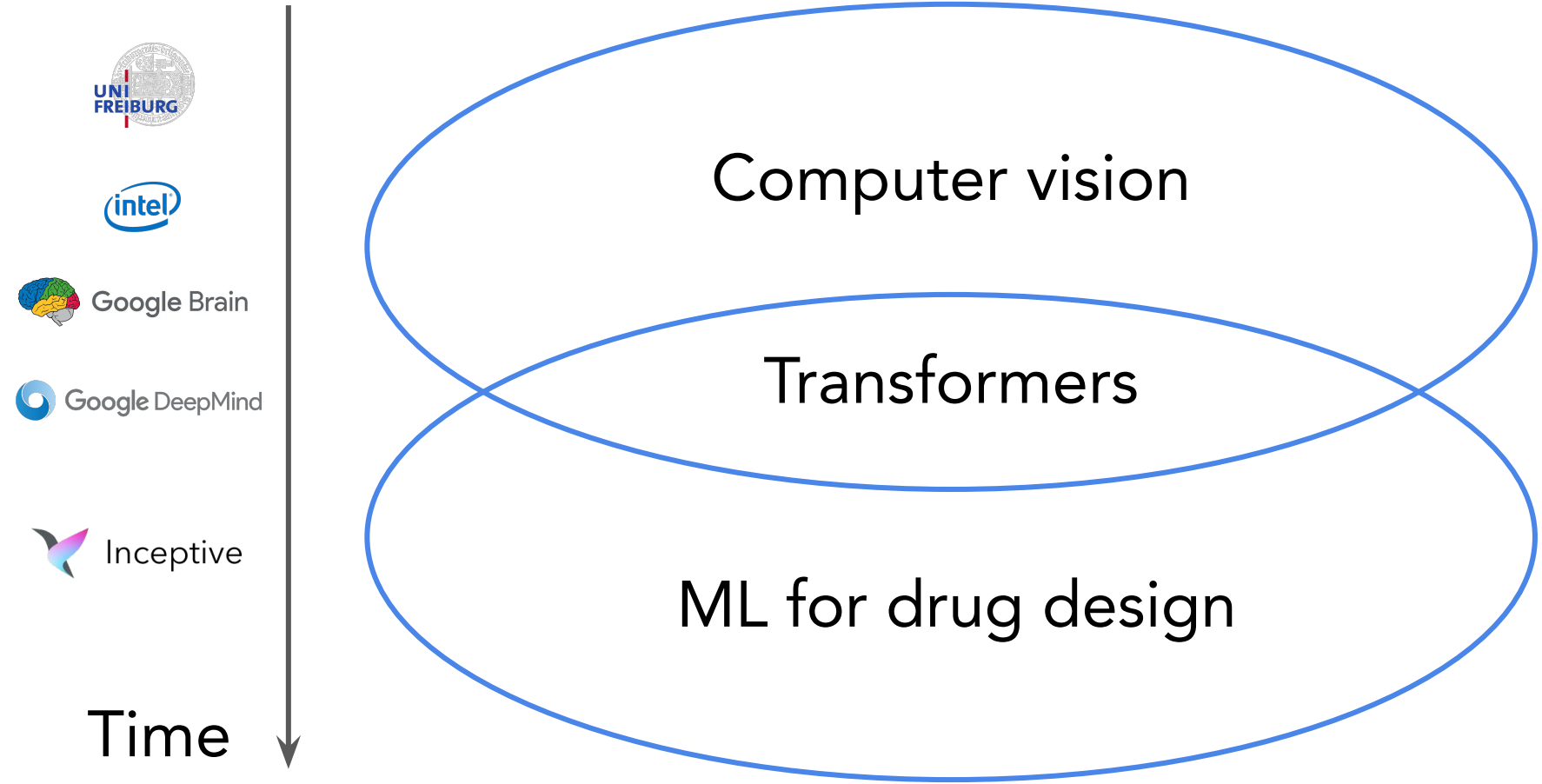
From Pixels to Nucleotides

Deep Learning for Computer Vision and Drug Design

Alexey Dosovitskiy

ML in PL, Warsaw
October 17, 2025

Research Bio tl;dr




Computer vision and "off-the-grid"
This talk
ML for drug design
What the hell is this?

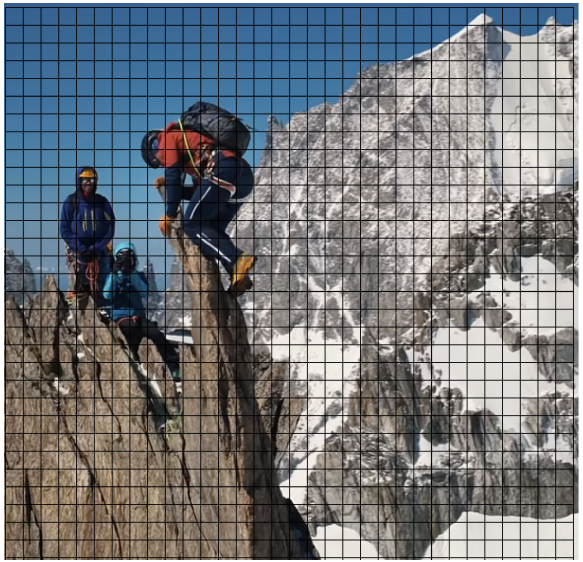
The World is 3D and Dynamic

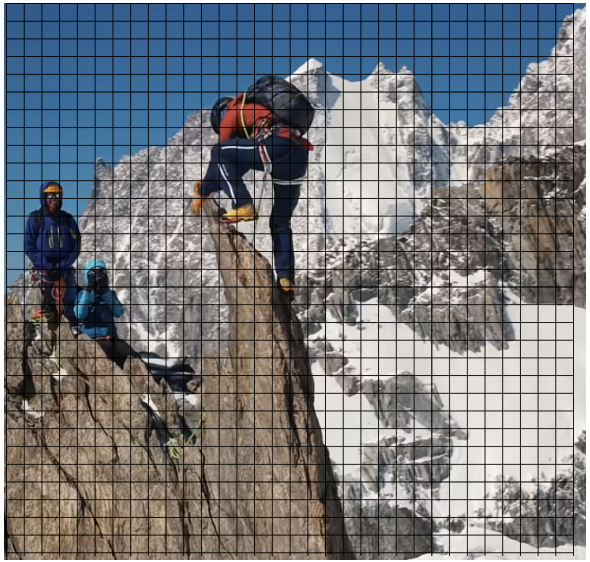
https://www.youtube.com/shorts/vOcqoLGX2jI

The World is 3D and Dynamic

https://www.youtube.com/shorts/vOcqoLGX2jI

Digital Images Live “On The Grid” of the Camera Sensor





https://www.youtube.com/shorts/vOcqoLGX2jI

We naturally see the world as it is – 3D and dynamic

How do we make computer vision systems that do too?

Why “off the grid”?

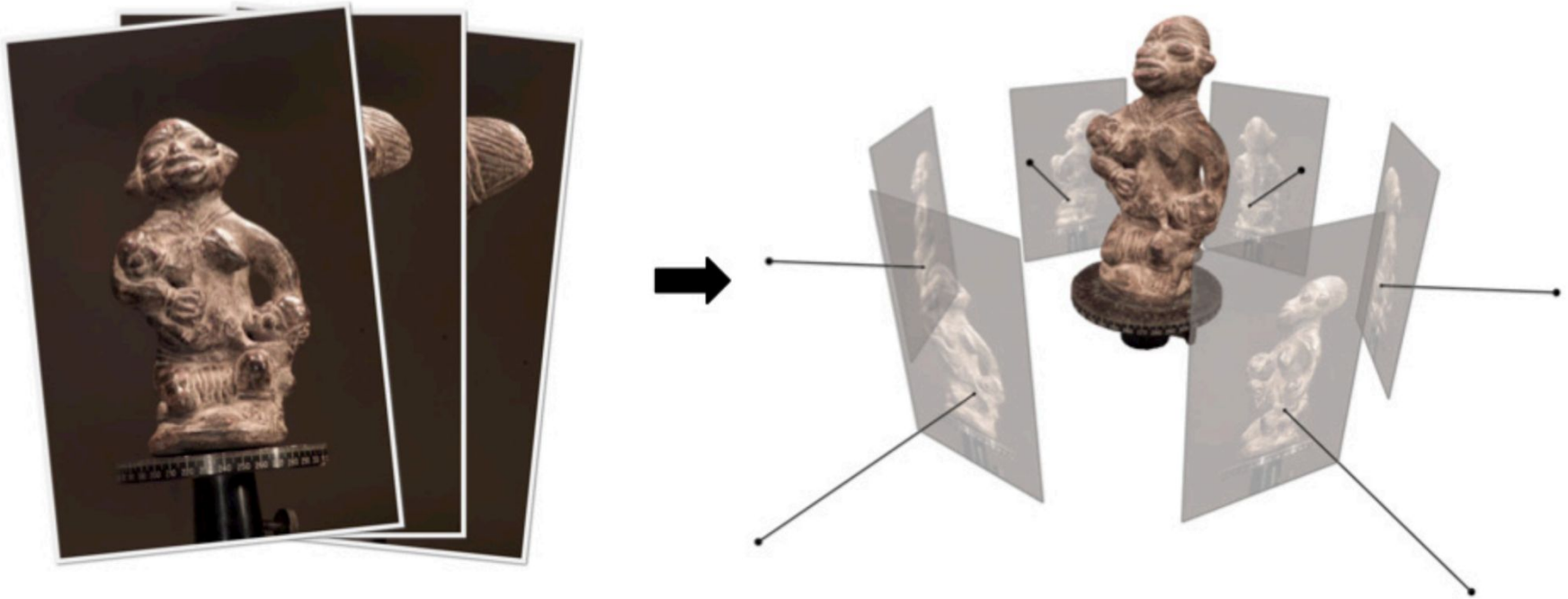
Efficiency!

The right inductive bias for vision

Especially for video

https://www.youtube.com/shorts/vOcqoLGX2jI

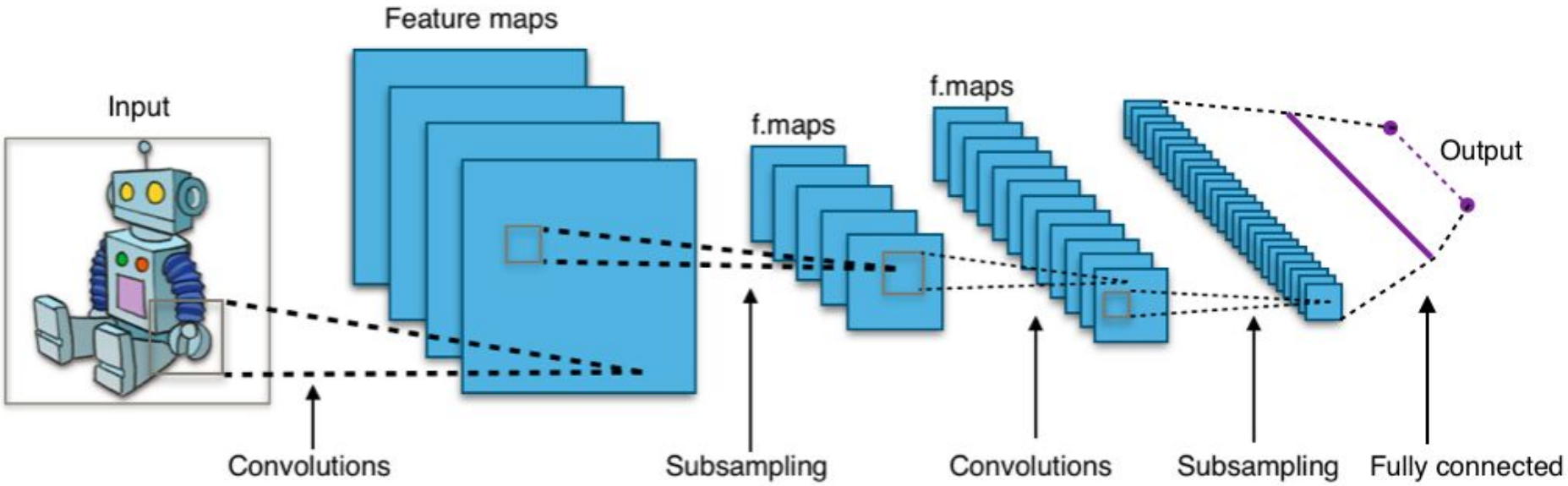
Aside: 3D Reconstruction



A large successful field, but not deep-learning-native

Image from Furukawa & Hernández

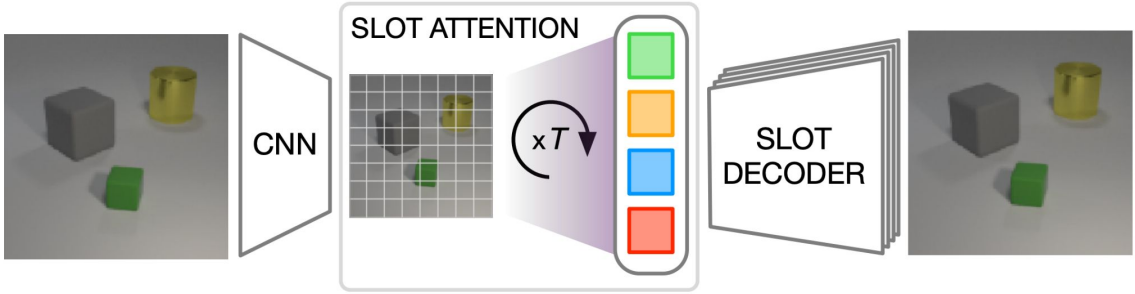
ConvNets Operate on the Grid



How do we take them “off the grid”?

https://commons.wikimedia.org/wiki/File:Typical_cnn.png

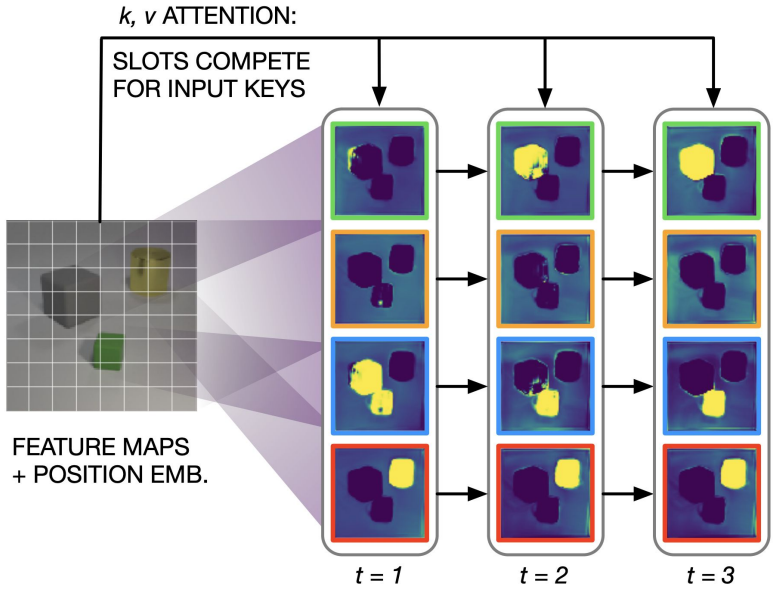
Take 1: Slot Attention



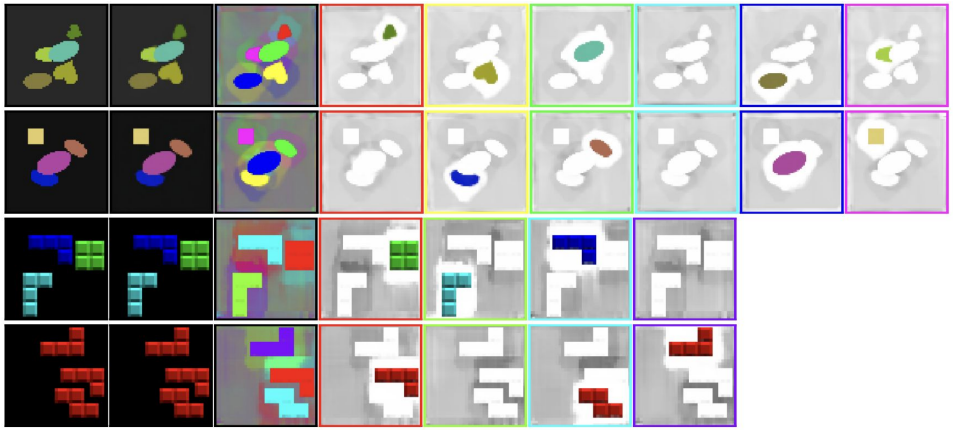
Task: map feature maps of a convolutional model to a set

Approach: Iterative "transposed" attention of K "slots" over the feature maps (~learned soft k-means)

Applications: unsupervised object discovery, set prediction

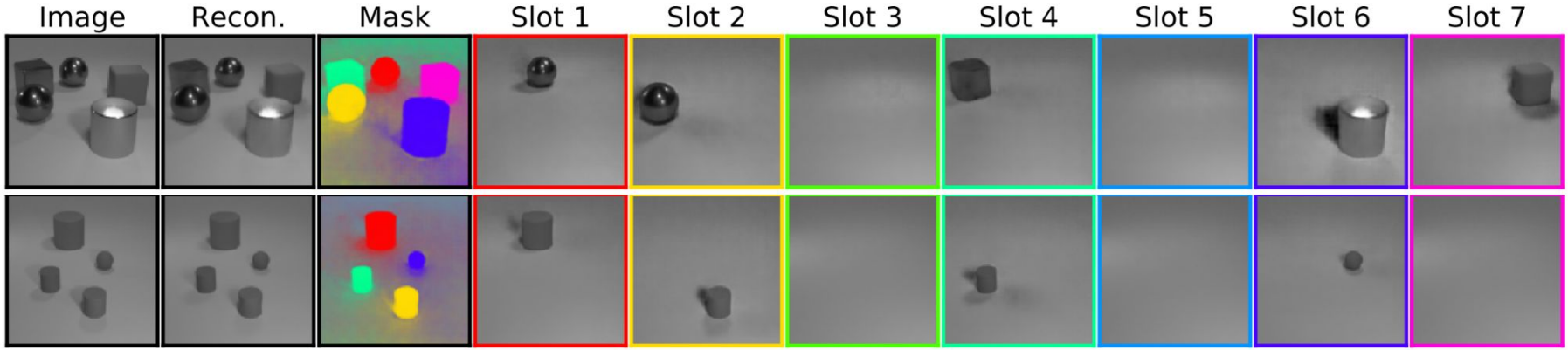


Slot Attention



Works well on simple synthetic tasks

But does not scale to the real world



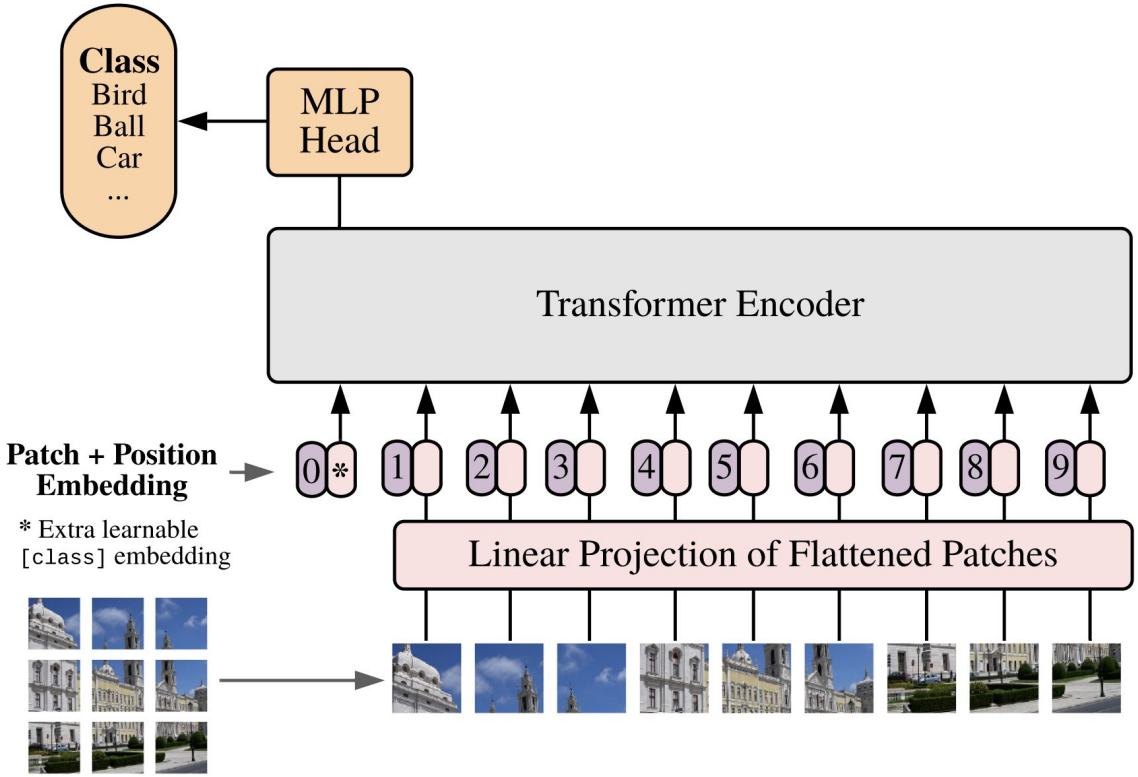
How do we make a scalable “off-the-grid” model?

Take 2: Vision Transformers

ConvNets operate "on the grid" – local 2D processing

What operates "off the grid"? Transformers!

Let's try applying a transformer to vision tasks instead of a ConvNet

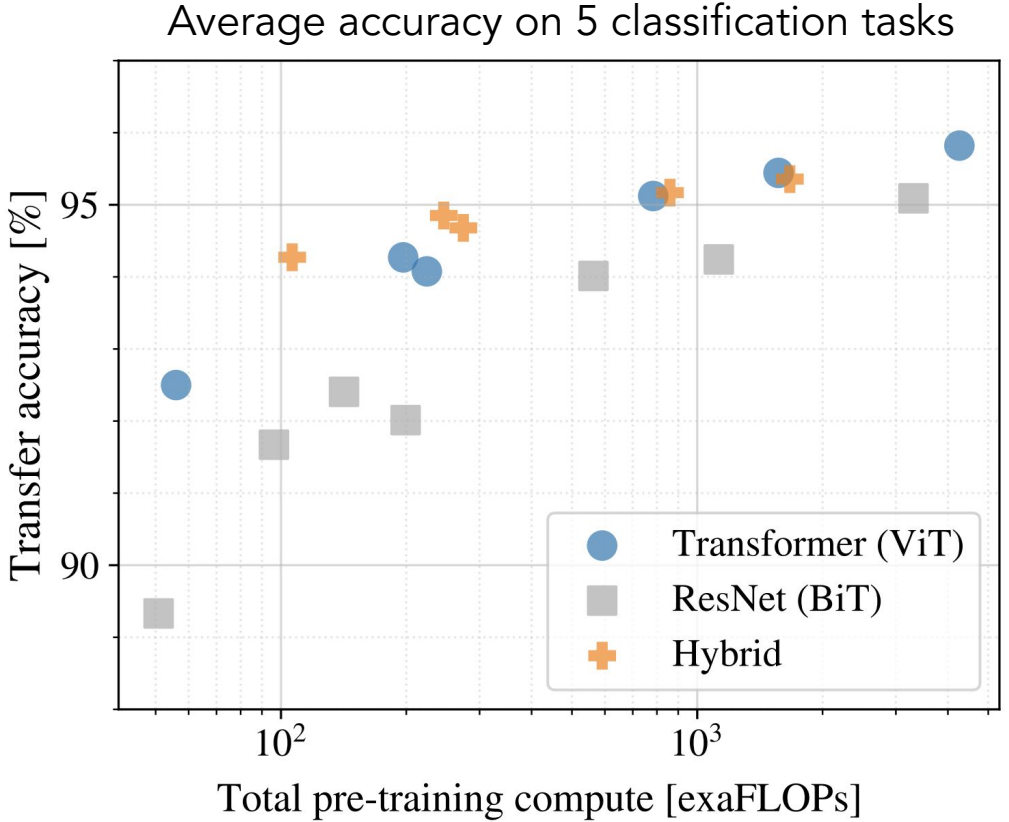


Dosovitskiy et al., An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

Vision Transformers

Transformers scale better than ResNets with lots of compute

"Hybrids" even better for smaller amounts of compute



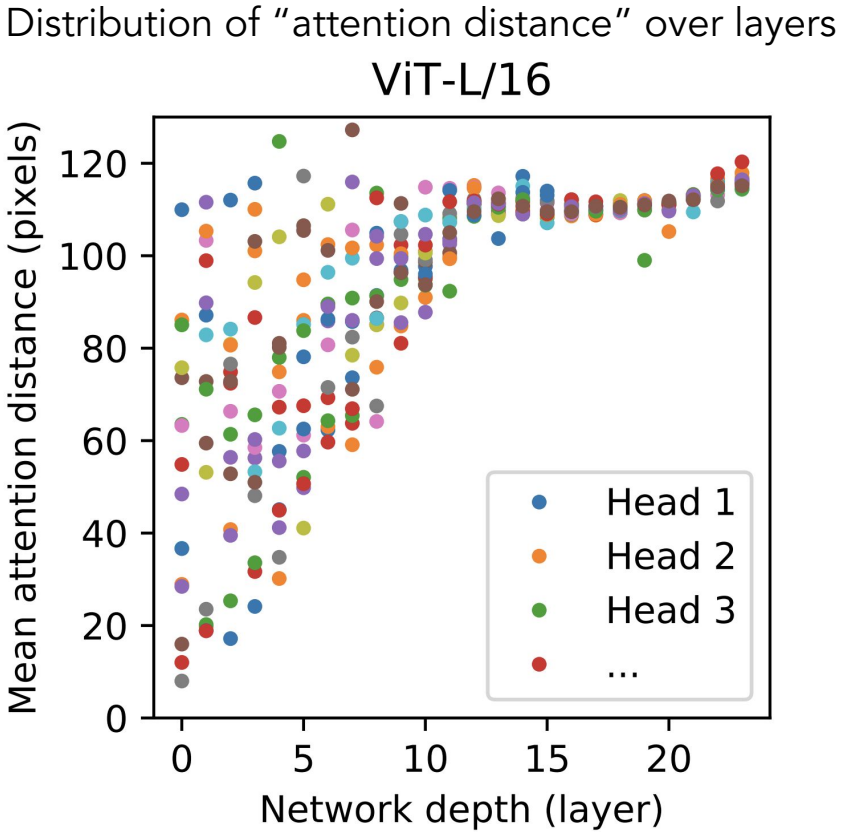
Dosovitskiy et al., An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

Vision Transformers

In early layers – both local and global attention heads

In latter layers – only global

While transformers themselves are "grid"-agnostic, the representation in ViTs is still "on the grid" (patches)



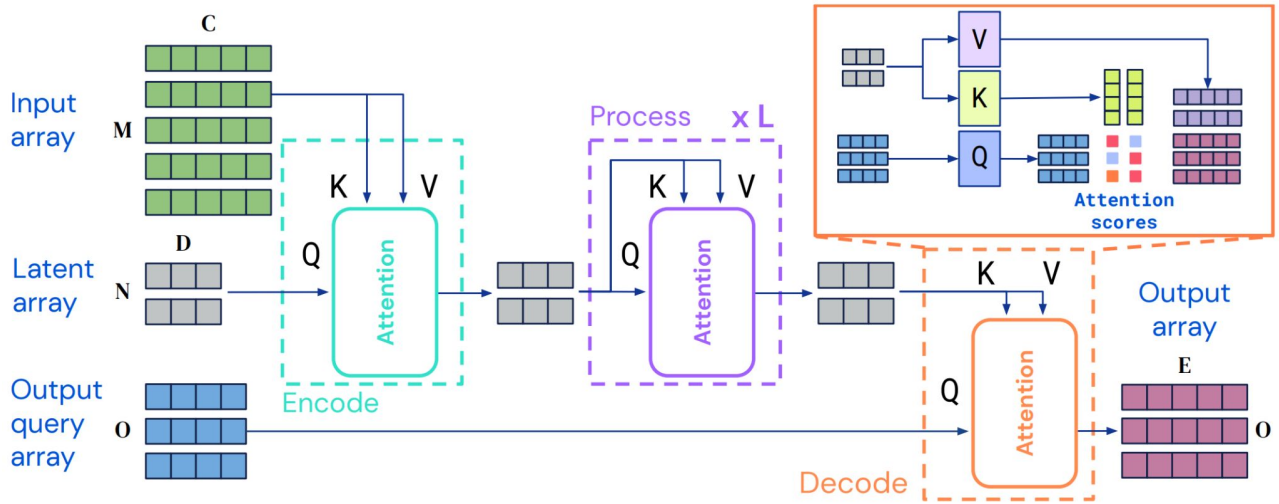
Dosovitskiy et al., An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

Honorable mention: Perceiver and Perceiver IO

"Off the grid" representation and processing

Works well on many vision tasks

Downside: worse in terms of the compute/performance tradeoff

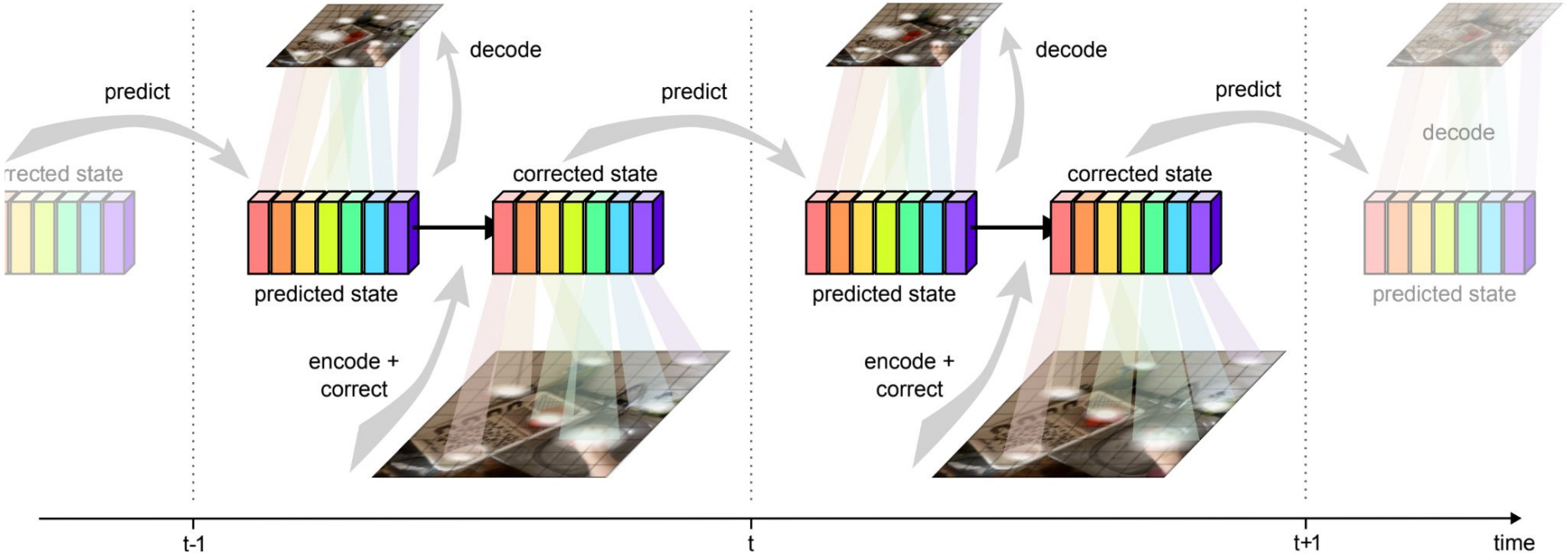


Jaegle et al., Perceiver: General Perception with Iterative Attention

Jaegle et al., Perceiver IO: A General Architecture for Structured Inputs & Outputs

Take 3: MooG – Off-the-Grid for Video

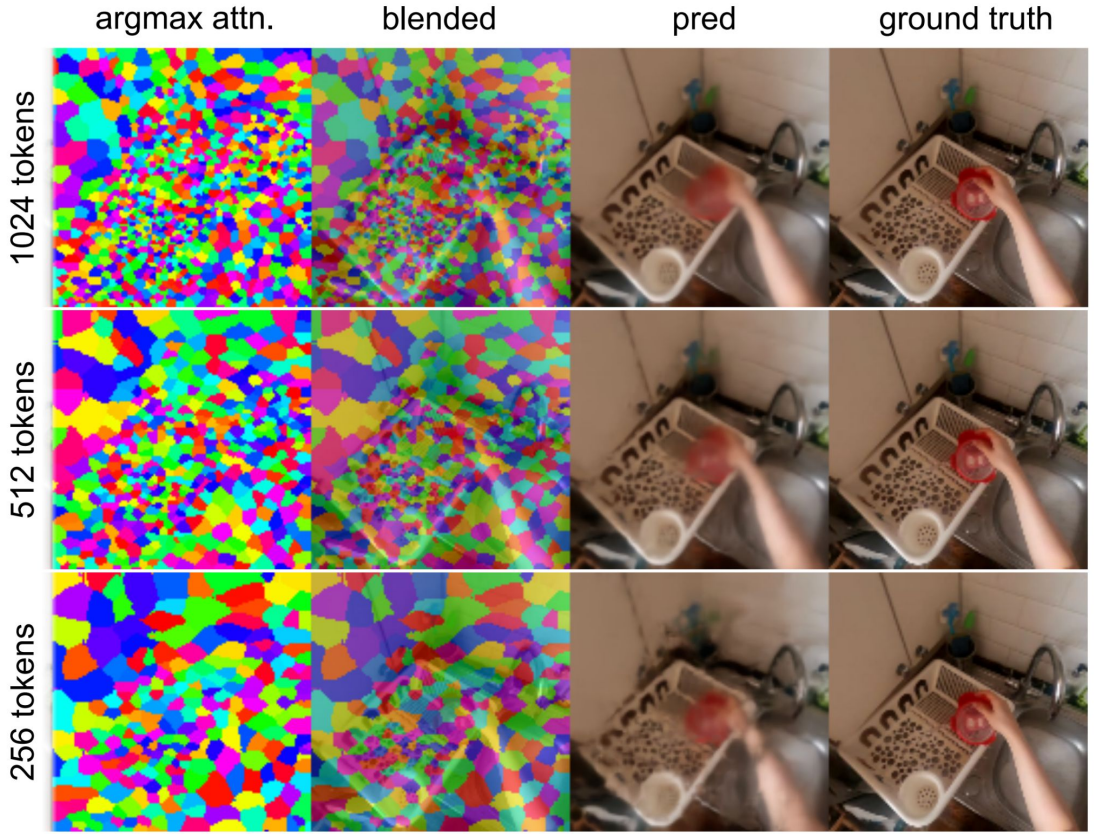
Slot Attention meets Vision Transformers and goes video



Steenkiste et al., Moving Off-the-Grid: Scene-Grounded Video Representations

MooG: Off-the-Grid for Video

Learns variable token size depending on the local complexity



MooG: Off-the-Grid for Video

Tokens are “off the grid” and rather connected to the scene!

MooG: Off-the-Grid for Video

Tokens are “off the grid” and rather connected to the scene!

MooG: Off-the-Grid for Video

Works pretty well on downstream tasks too

This is using frozen representation from the model

Note that MooG has 35M params, while VMAEv2 S/B/G – 20M/80M/1000M

	MOVi-E			DAVIS	Waymo
Name	Points (↑AJ)	Depth (↓AbsRel)	Boxes (↑IoU)	Points (↑AJ)	Boxes (↑IoU)
MooG	**0.839**	0.0359	**0.793**	0.687	**0.730**
Grid	0.769	0.0451	0.730	0.518	0.625
Grid Rec.	0.778	0.0443	0.734	0.559	0.629
DINOv1 (B)	0.518	0.0371	0.724	0.409	0.566
DINOv2 (B)	0.544	0.0370	0.738	0.402	0.559
VMAEv2 (S)	0.595	0.0567	0.700	0.365	0.567
VMAEv2 (B)	0.681	0.0458	0.736	0.434	0.611
VMAEv2 (G)	0.822	**0.0311**	**0.793**	**0.720**	0.708

Steenkiste et al., Moving Off-the-Grid: Scene-Grounded Video Representations

Summary

	Off-the-grid representation	Off-the-grid processing	Scalable
Slot Attention	yes	no	no
Vision Transformer	no	yes	yes
Perceiver (IO)	yes	yes	maybe
MooG	yes	yes	maybe

Honorable mentions:

RIN (Jabri et al.)
FIT (Chen, Li)
AdaTape, Registers
Gaussian Splatting
GLOM (Hinton)
…

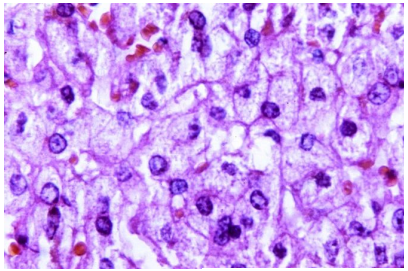
Good progress, but still no “bulletproof” scalable off-the-grid model

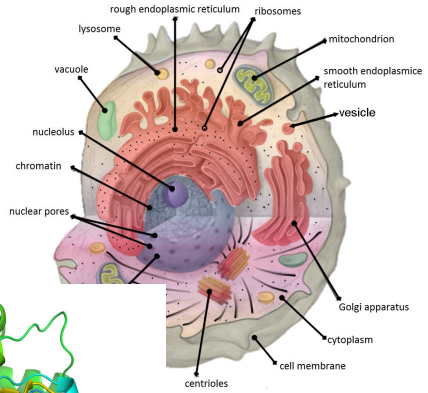
Difficult to imagine that transformers are “the ultimate architecture”

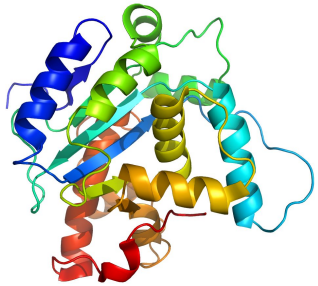
=> we need to keep trying!

From Pixels

To Nucleotides

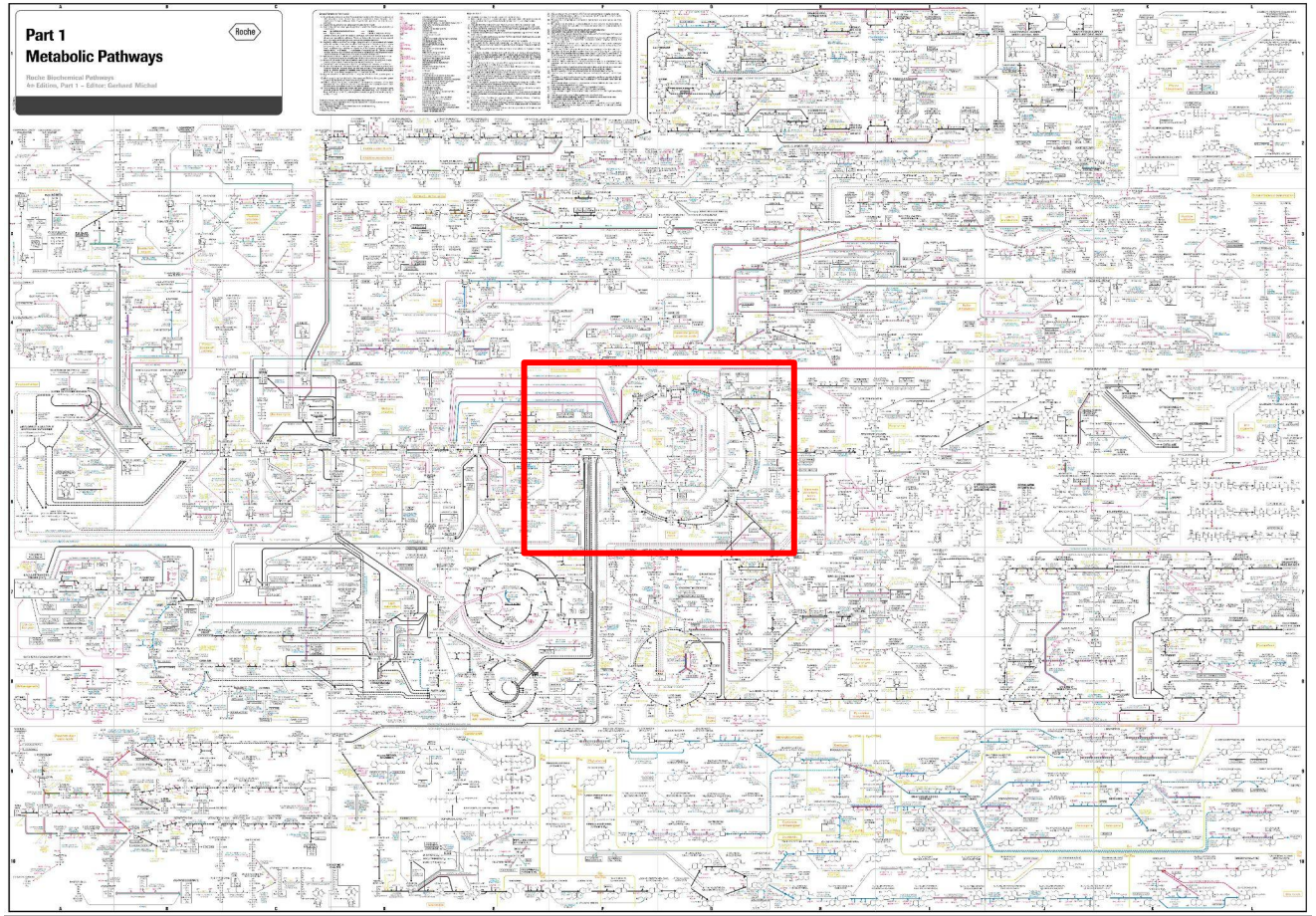






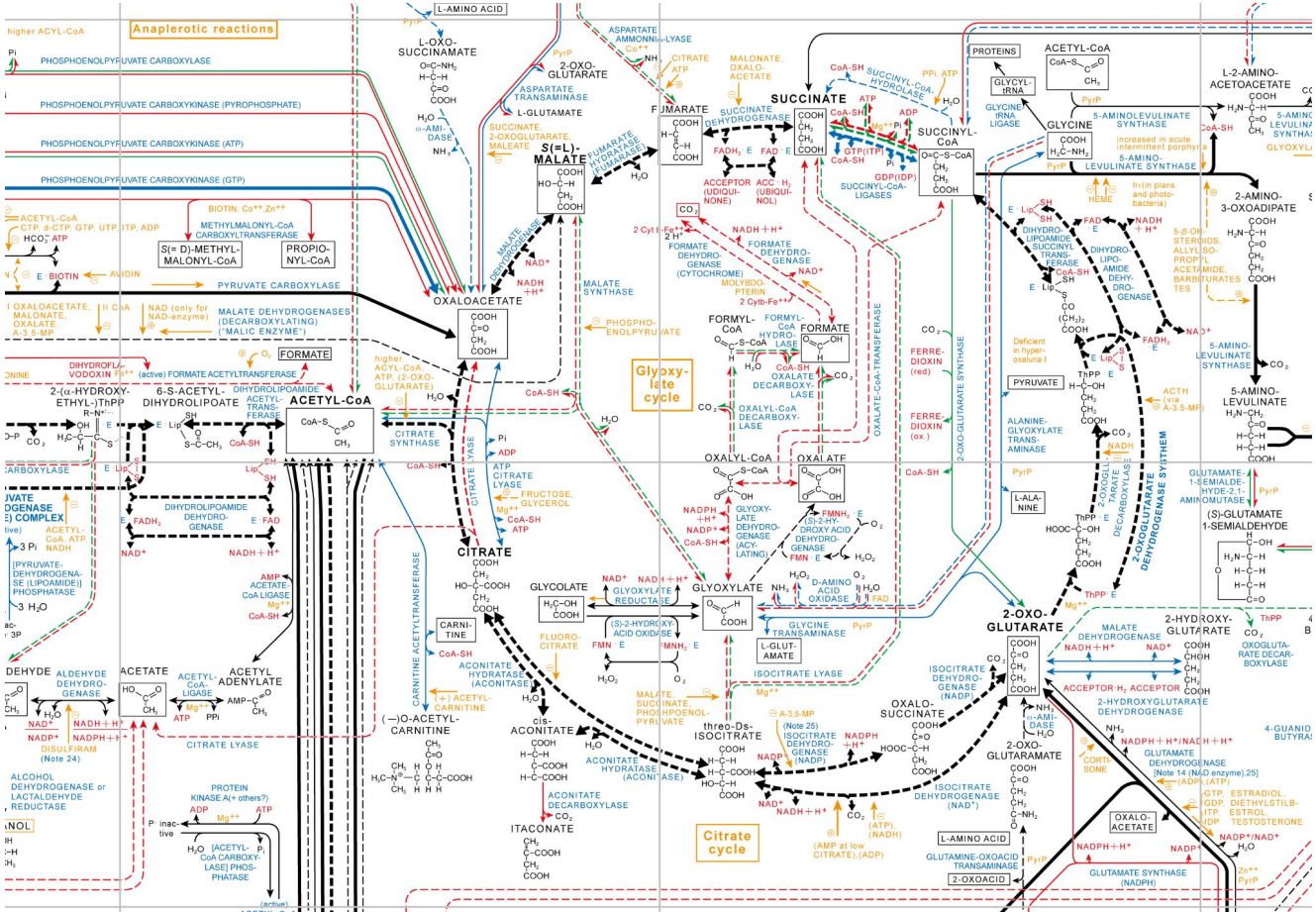
???

Biology is Complex



Michal et al., Roche, "Metabolic Pathways"

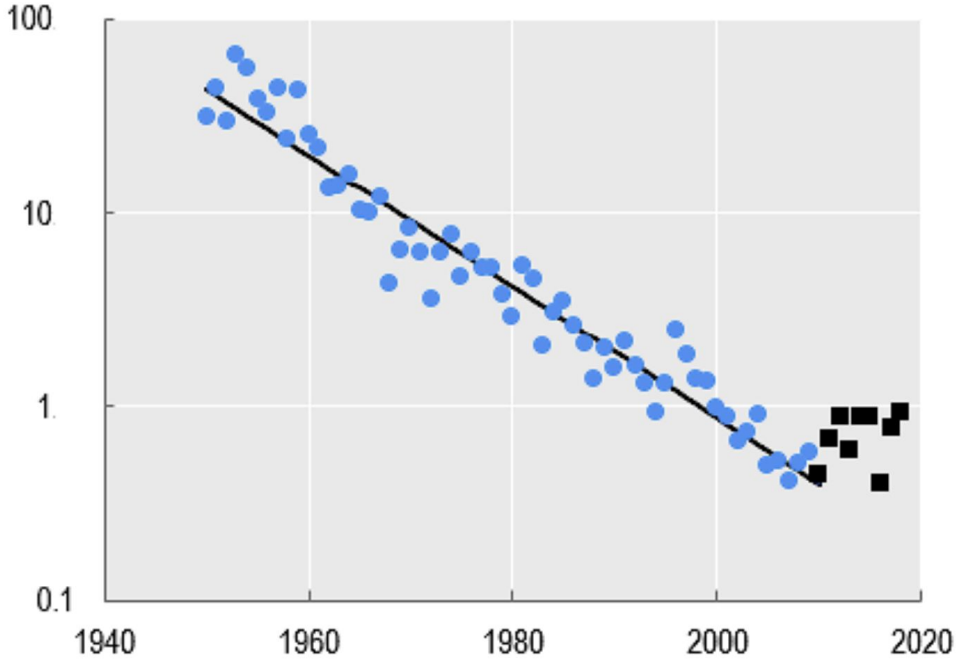
Biology is Very Complex



Eroom's law: Drug Development Gets More Expensive

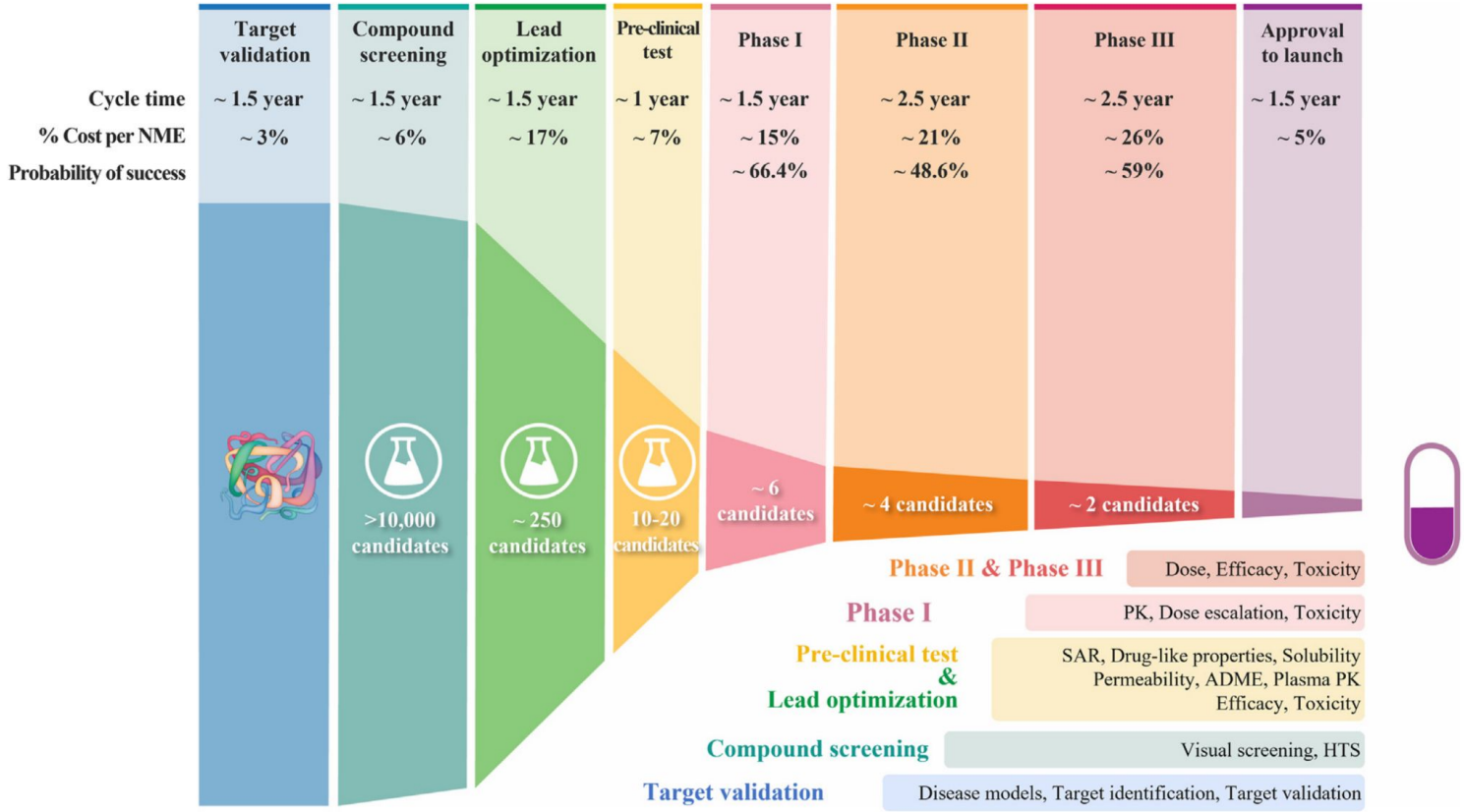
A. New molecule entities and new biologics approved by the FDA

per billion USD inflation-adjusted R&D investment, logarithmic vertical axis



Scannell et al., Diagnosing the decline in pharmaceutical R&D efficiency
Scannell, Eroom's Law and the decline in the productivity of biopharmaceutical R&D

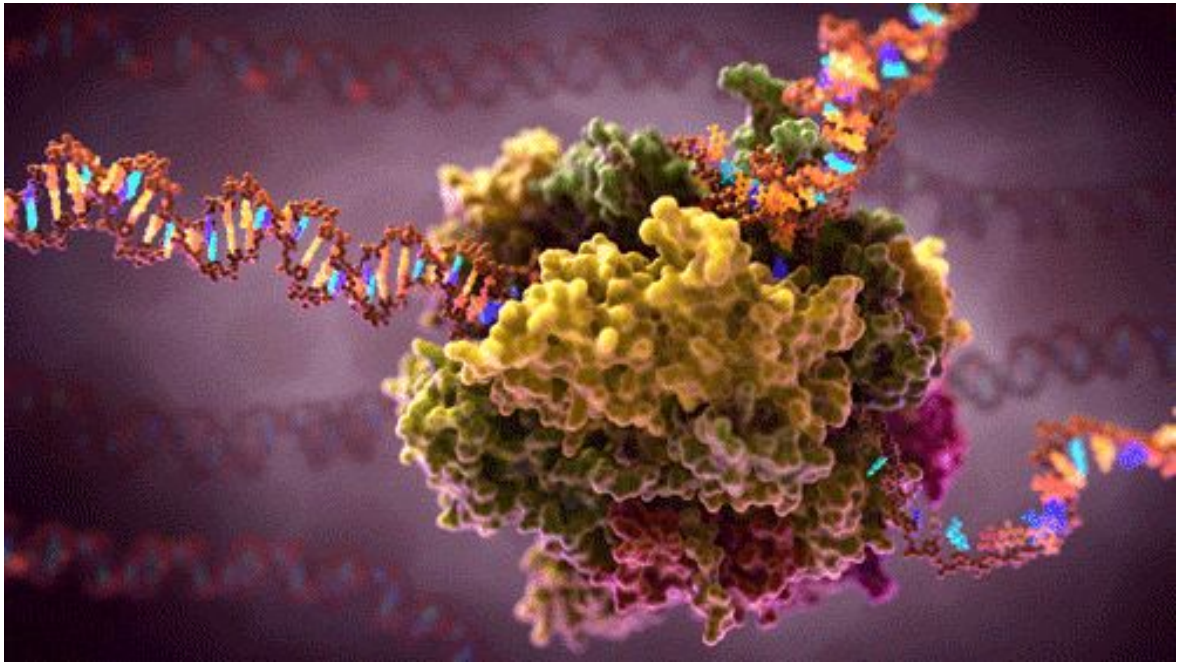
Why So Expensive? Funnel of Drug Development

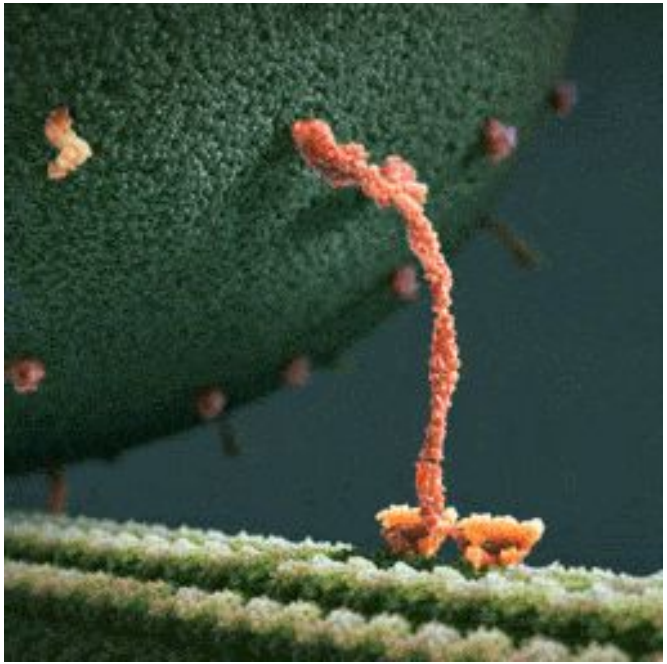


Sun et al., Why 90% of clinical drug development fails and how to improve it?

ML to the rescue?

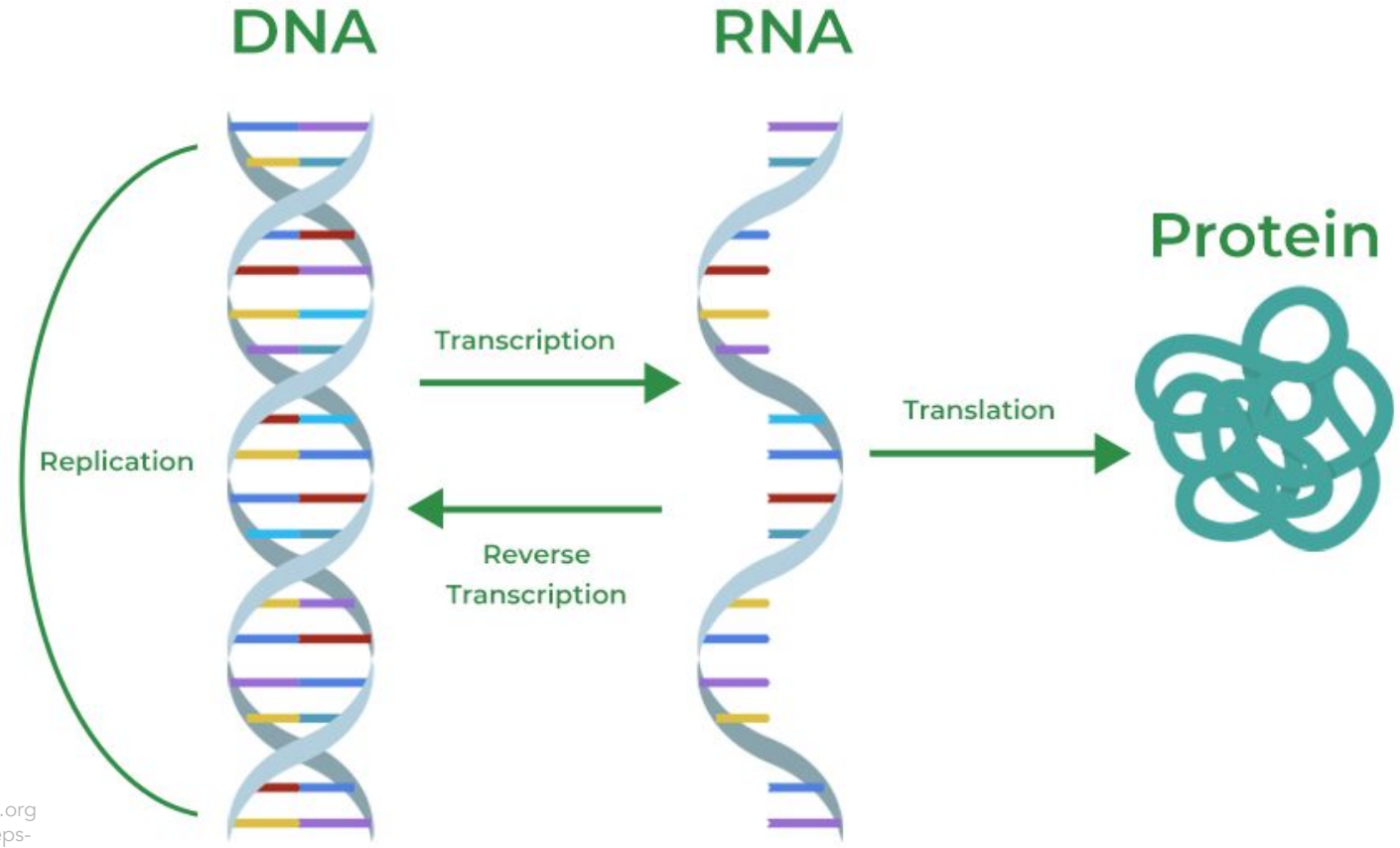
Proteins* make living organisms tick





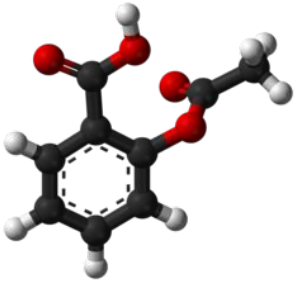
* There's increasing evidence that (non-coding) RNA etc are super important too

Central dogma of molecular biology

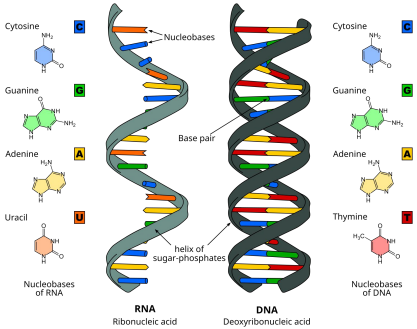


https://www.geeksforgeeks.org/biology/central-dogma-steps-guide/

Drug Modalities

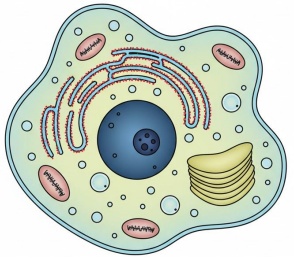


Small molecules



Nucleic acids

Proteins



Cells

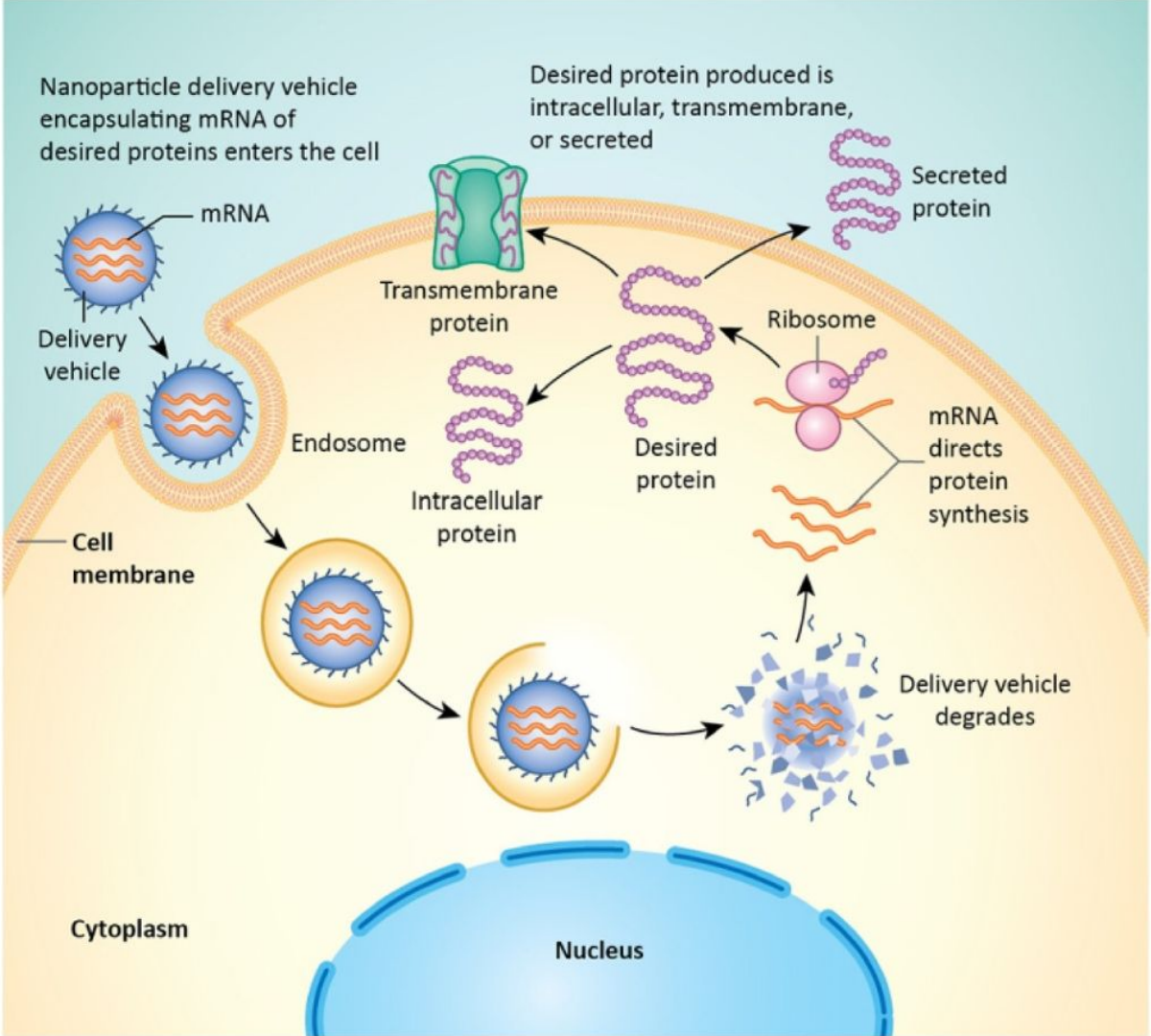
Biologics

Images from https://commons.wikimedia.org/

mRNA drugs

Instead of putting proteins into the body, make body produce the proteins itself

Delivered usually in lipid nanoparticles (LNPs)



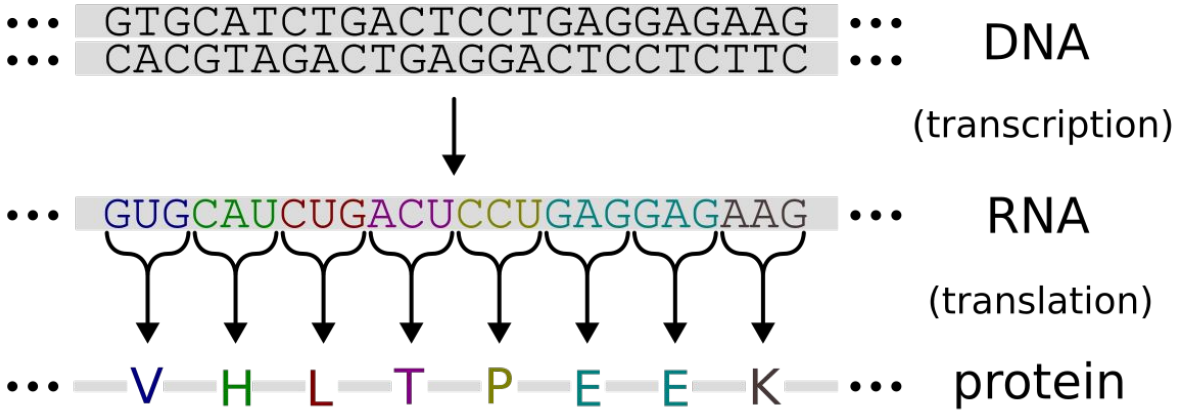
Bhat et al., mRNA therapeutics: beyond vaccine applications

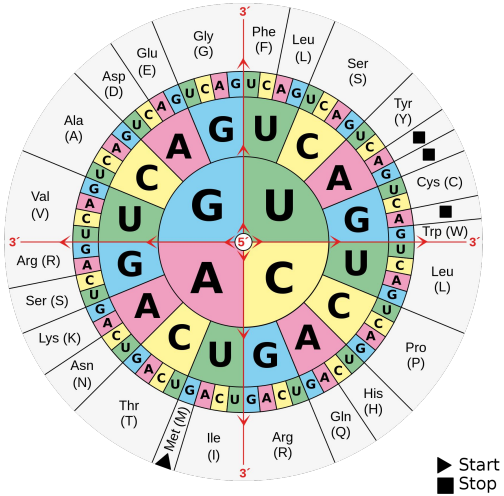
DNA/RNA code for proteins

DNA/RNA: 4 types of nucleotides

Protein: ~22 amino acids

Triplet of nucleotides: 4^3 = 64 options





https://en.wikipedia.org/wiki/DNA_and_RNA_codon_tables#/media/File:Aminoacids_table.svg
https://commons.wikimedia.org/wiki/File:Genetic_code.svg

mRNA Design Problem Example: COVID Spike Protein

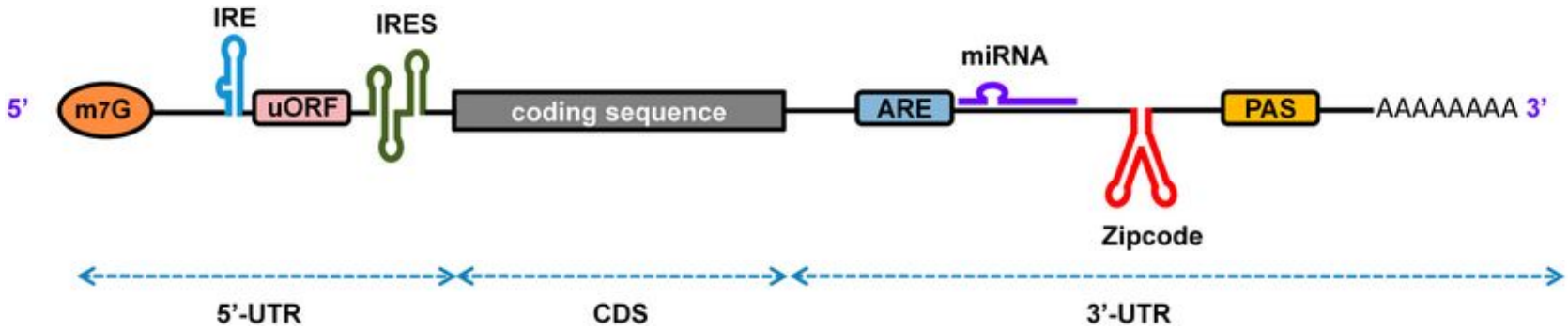
E.g. COVID spike protein: 1273 amino acids long

Untranslated regions: e.g. 150 + 350 = 500 nucleotides

=> total number of options approx: 4^500 * 3^1273 = *a lot*

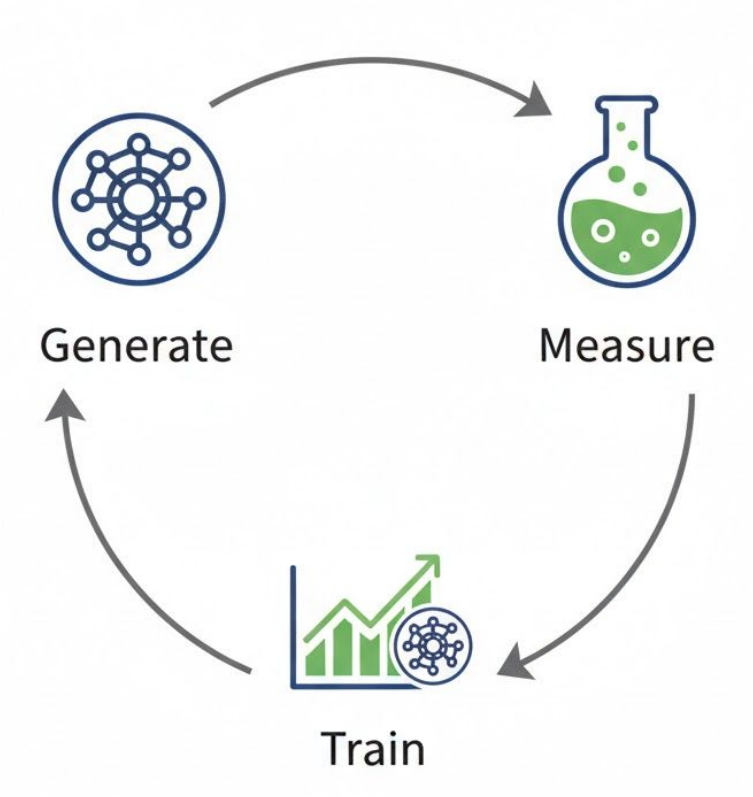
Number of nucleotides

Approx number of codon options per amino acid

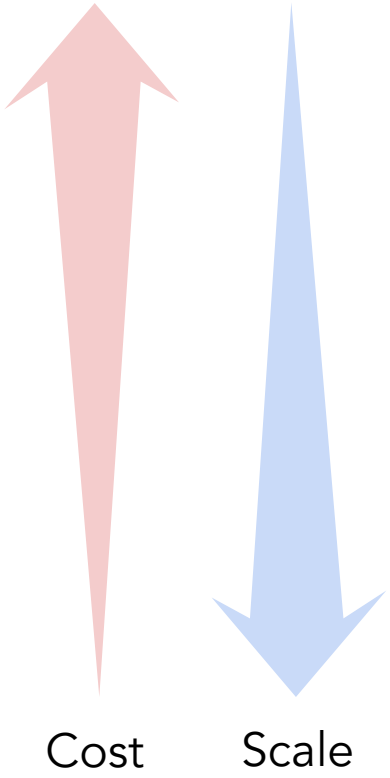


Ahmed et al., Mining Functional Elements in Messenger RNAs: Overview, Challenges, and Perspectives

How do we solve it? “Lab in the Loop”



Challenge: Evaluation



Clinical trials

In-vivo non-human primates

In-vivo mice

In-vitro (cells) arrayed

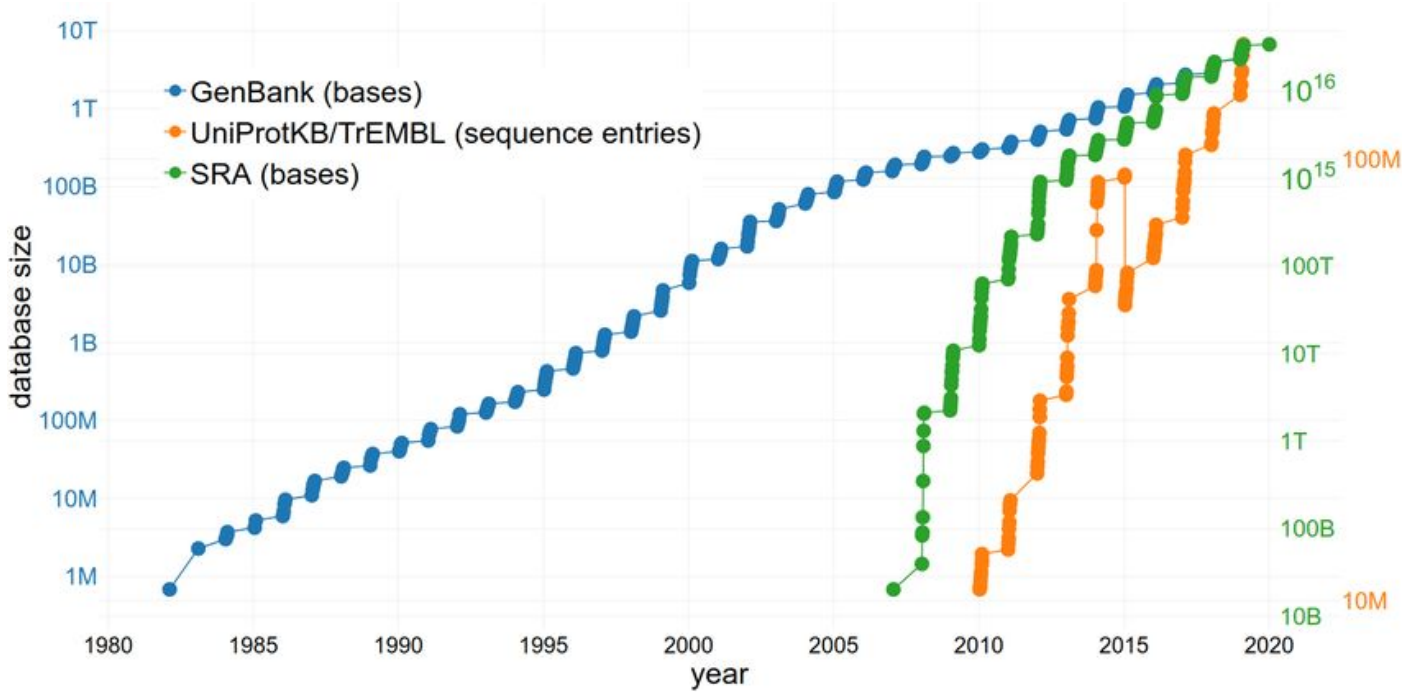
In-vitro (cells) pooled

In-silico (computational)

Challenge: Training data

Lots of publicly available genomic data (DNA, RNA, sequencing, proteins)

But fairly little, and scattered, measurements speaking to the properties

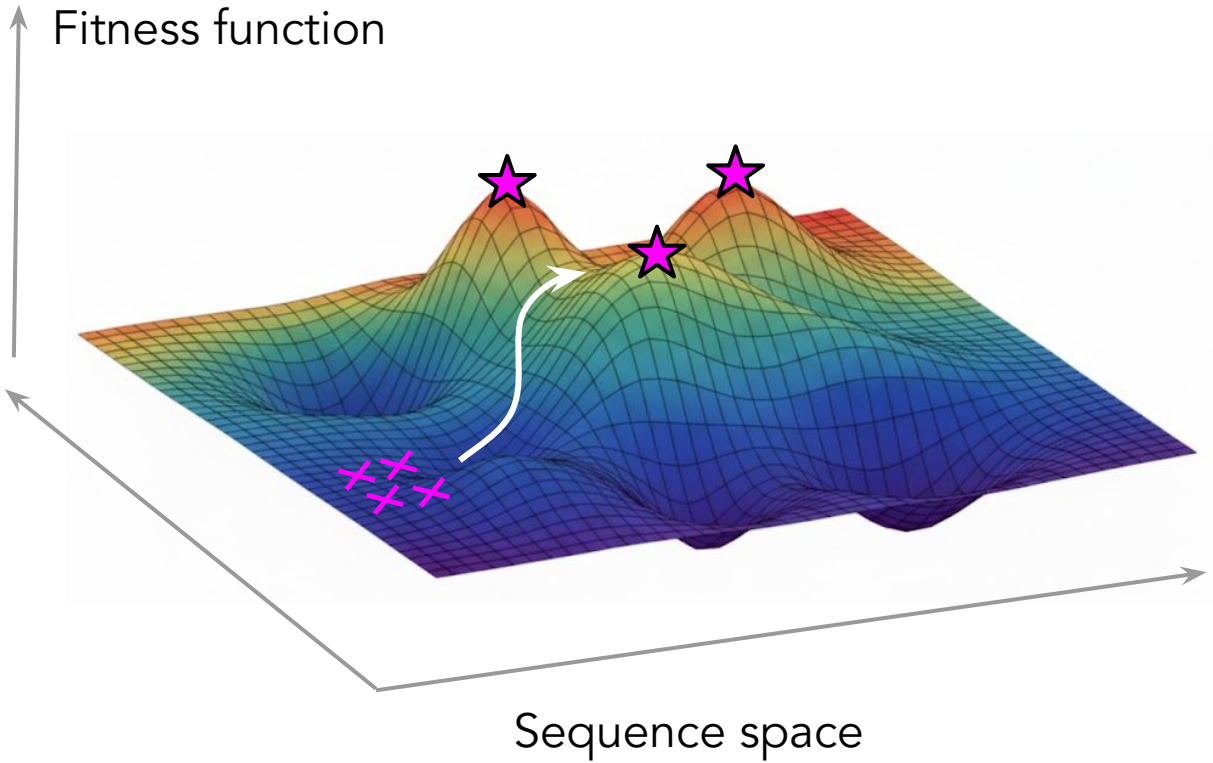


Sielemann et al., The Reuse of Public Datasets in the Life Sciences: Potential Risks and Rewards

Challenge: Extrapolation, exploration

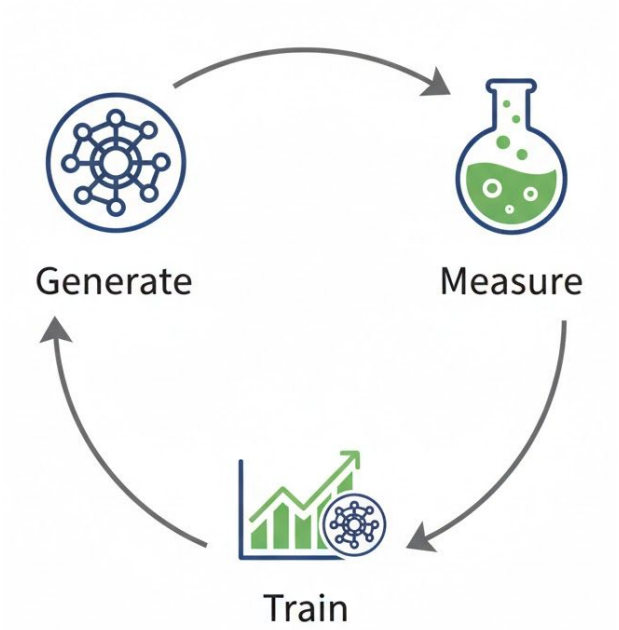
Extrapolation: at each lab-in-the-loop iteration, want to get sequences better than best measured so far

Exploration: want to find the best performing sequences in a huge search space



Lab in the Loop Revisited

Smart design strategies trading off exploration and exploitation



New scalable assays and smart use of different “tiers” of data, including in-silico

Generative models supporting extrapolation, low-data regimes, multi-objective optimization

Large-scale pre-training on genomic data and beyond

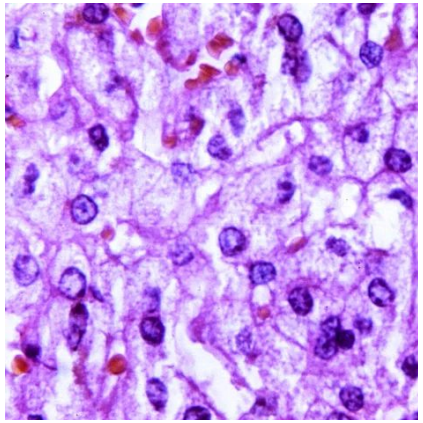
Summary

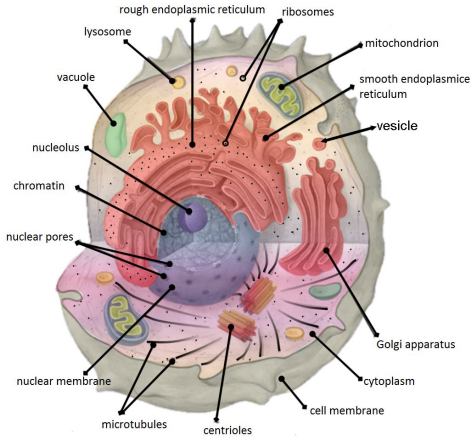
Biology is extremely complex and “messy”

That’s exactly the type of problem ML is good at

Which is great since drug development needs help

Lots of interesting and difficult challenges





Computer vision is still not “solved” – exciting challenges!

Loads of room for innovation and impact in using ML for biology and medicine