3000788 Intro to Comp Molec Biol

Lecture 25: Foundational and frontiers AI models in biology

Fall 2025

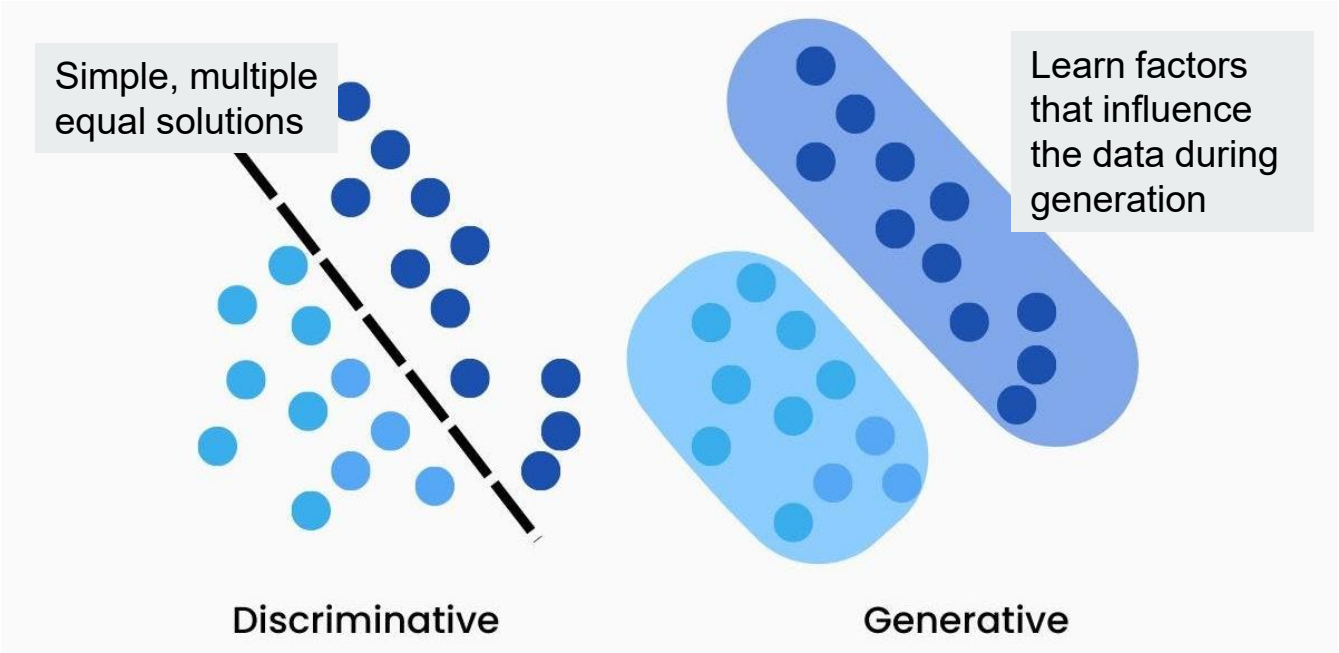
Sira Sriswasdi, PhD

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

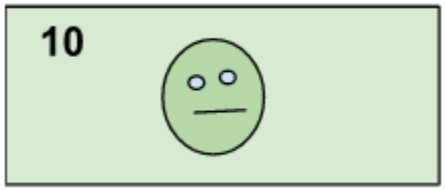
Today’s agenda

- The rise of generative AI
- Generative model designs and assumptions
- Foundational model

Importance of generative approach



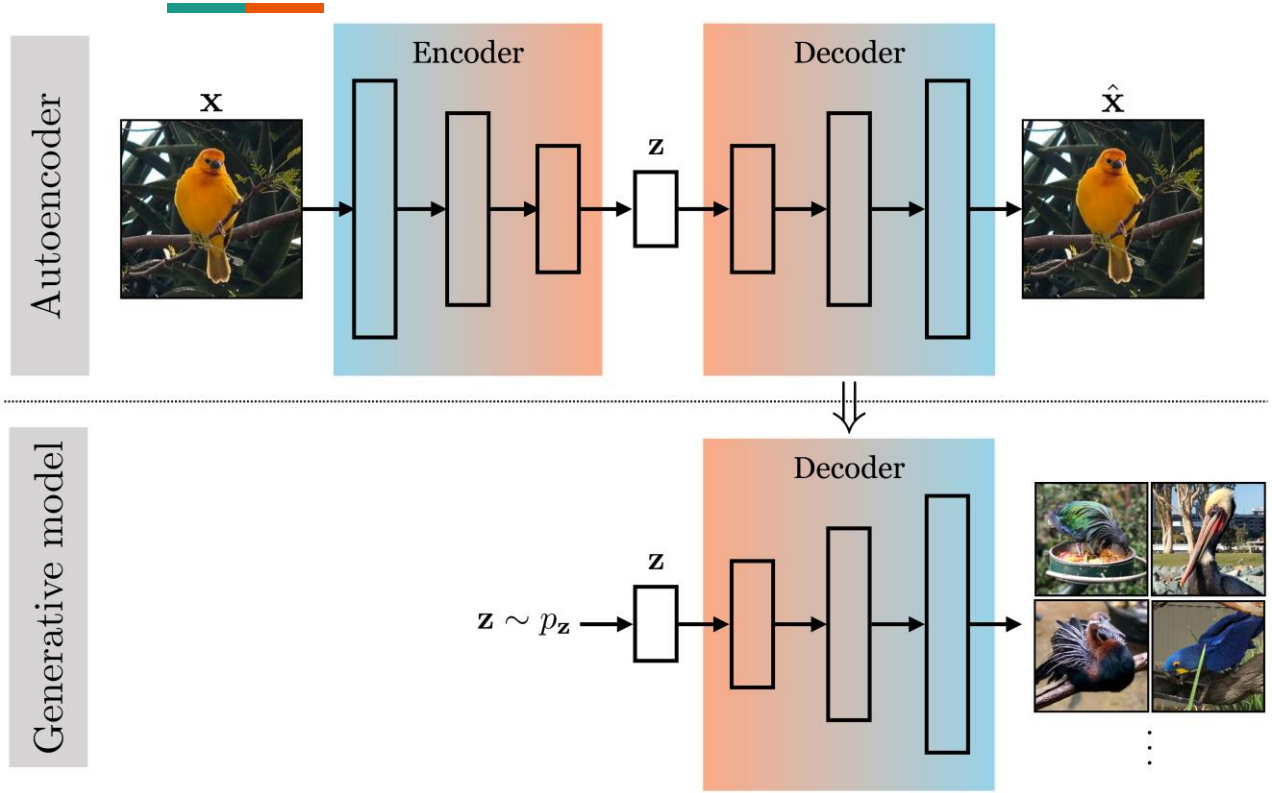
https://www.turing.com/kb/generative-models-vs-discriminative-models-for-deep-learning



VS

- It takes much more understanding to generate **realistic** data

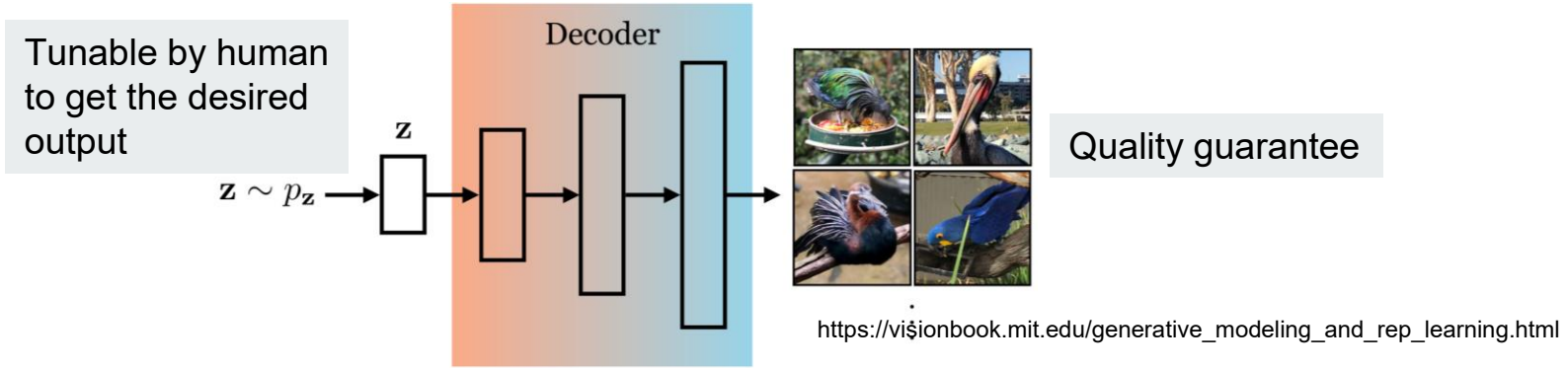
Autoencoder is a primitive generative model



https://visionbook.mit.edu/generative_modeling_and_rep_learning.html

- Decoder is a generator
  - Not well-trained
- No guarantee on how *z* determines output
- No guarantee that the output will be realistic given randomly selected *z*

Properties of a good generative model



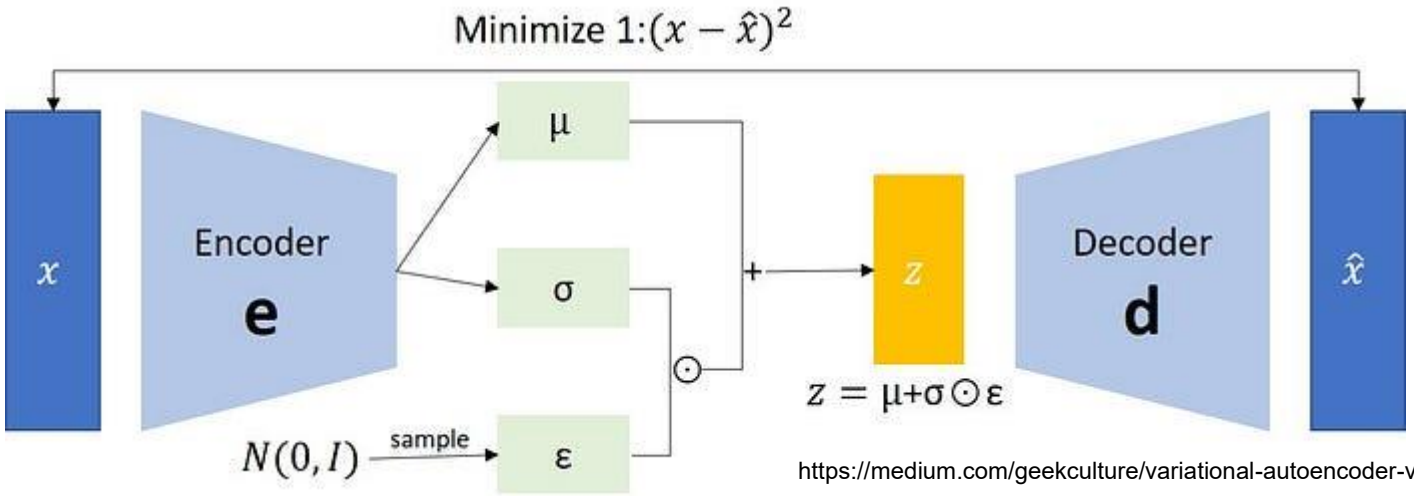
- **Realism**: On a reasonable range of *z*, the outputs should be realistic
- **Smoothness**: Small change in *z* should ensure small change in output
- **Interpretability**: Space of *z* should be mappable to human-understandable concepts ← concept of latent variables in statistics

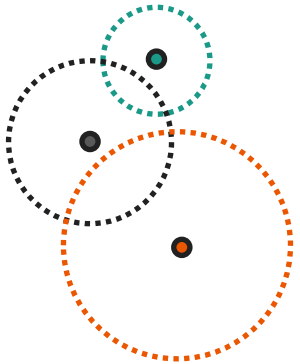
Generative model designs

Approaches for generative modeling

- Variational autoencoder (VAE)
- Generative adversarial network (GAN)
- Diffusion models
- Flow matching

Variational autoencoder





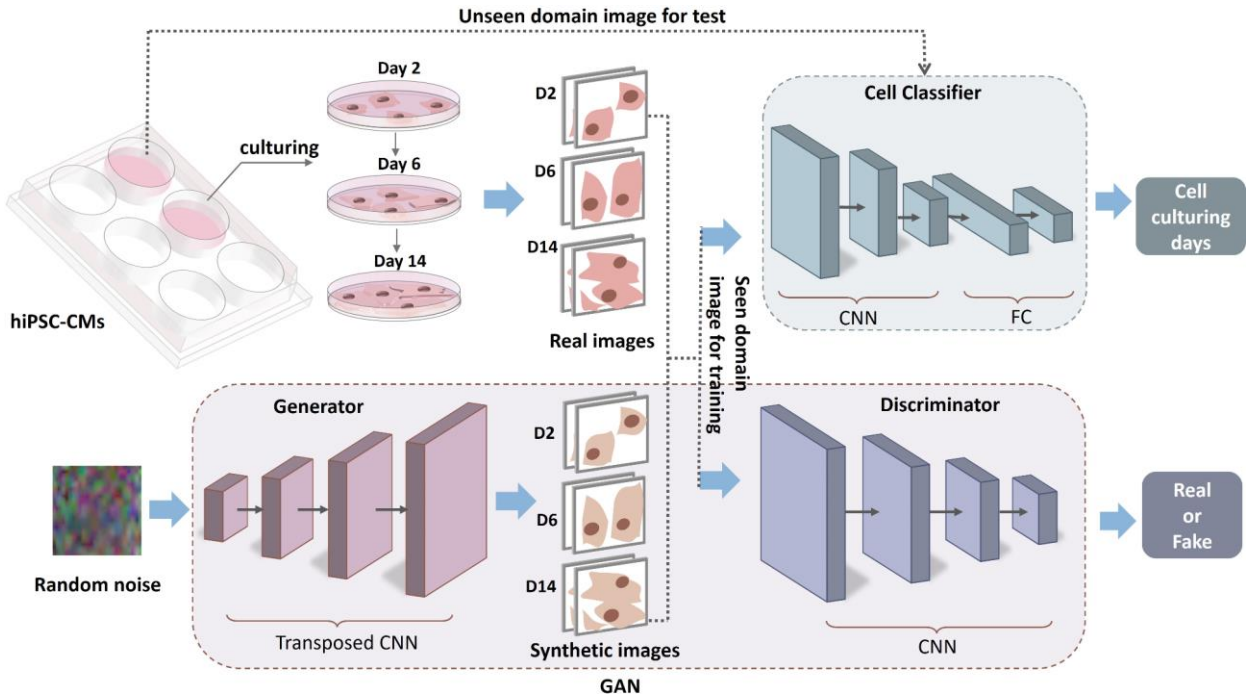
https://medium.com/geekculture/variational-autoencoder-vae-9b8ce5475f68

- Learn the Gaussian mean and SD of the representation for each input
  - Robust to noises / smoother representation space
- Sample from representation space before decoding

Key points of VAE

- **Collapse**: VAE can collapse to vanilla autoencoder with σ = 0
  - Force the representation space to resemble standard normal distribution with KL divergence: $D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}.$
- Representation space is both smooth and parametrized

Adversarial training



Wu, Z. et al. Scientific Reports 14: 27016 (2014)

- **Generator** produce synthetic data
- **Discriminator** tries to tell if the data is real
- **Task-specific model** checks if the synthetic data match the desired characteristics

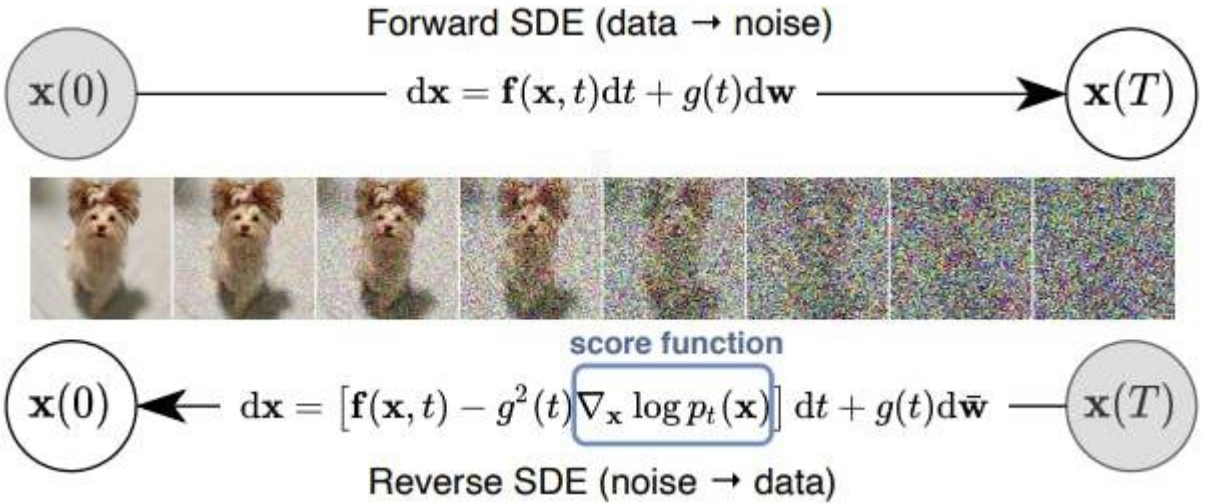
Key points of adversarial training

- Competitive learning process
  - Initially, **Generator** is poor, **Discriminator** easily learns to distinguish
  - As **Generator** learns, **Discriminator** has to follow

- Inefficiency
  - **Generator** does not learn from real data directly (input is pure noise)
  - Trained to match the generated data distribution with real data distribution (based on some statistics or representations of the data)
  - Small feedback from **Discriminator**'s output and performance

- Unstable

Stochatic diffusion

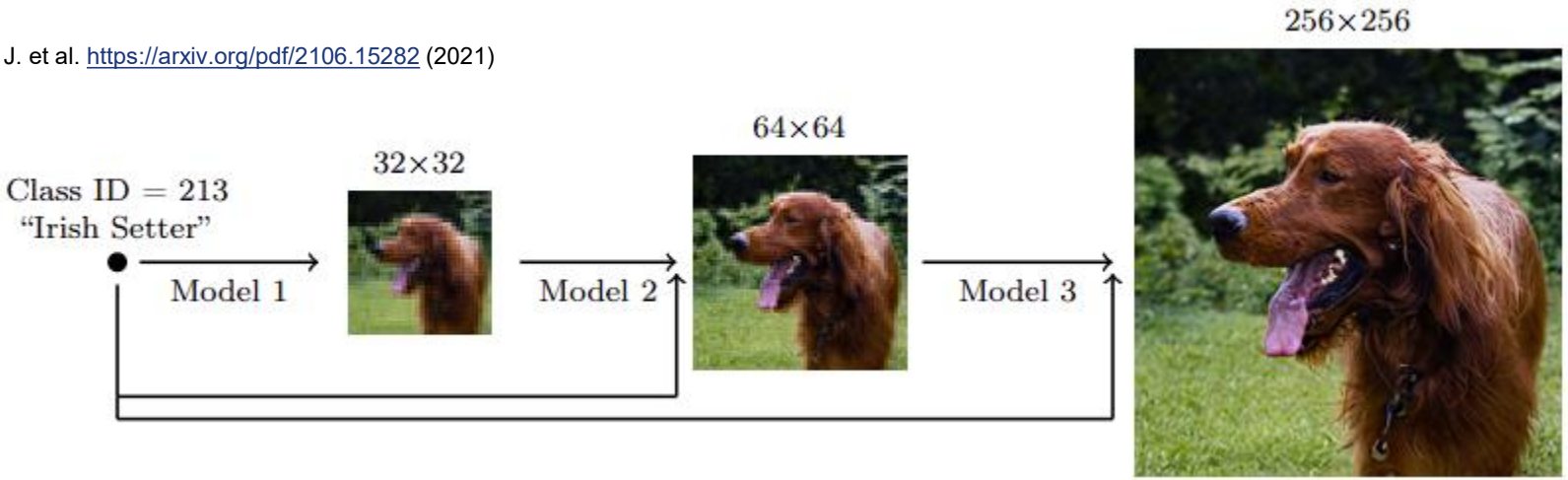


https://encord.com/blog/diffusion-models/

- Iteratively add small noises to the data for *N* steps
- Train a neural network that reverse the noise addition
  - Received noised data + time encoding as inputs

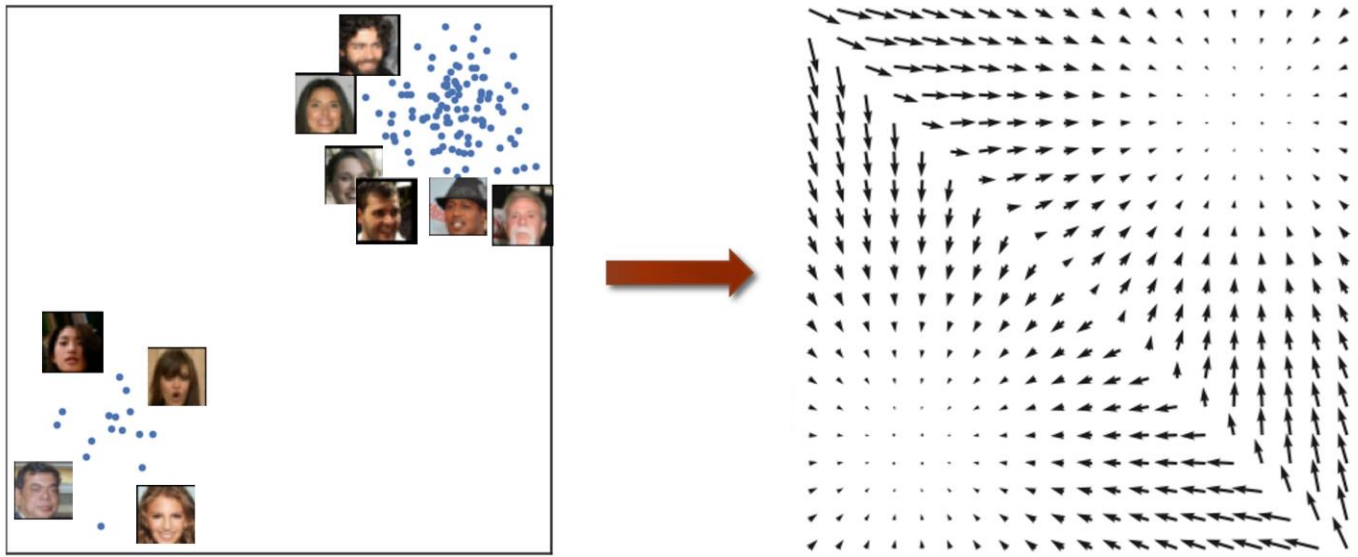
Cascade diffusion model

Ho, J. et al. https://arxiv.org/pdf/2106.15282 (2021)



- Deal with large, high-resolution generation
- Multiple diffusion models applied sequentially
- Output from early model guides the generation of the next

Vector field induced by the diffusion process



https://yang-song.net/blog/2021/score/

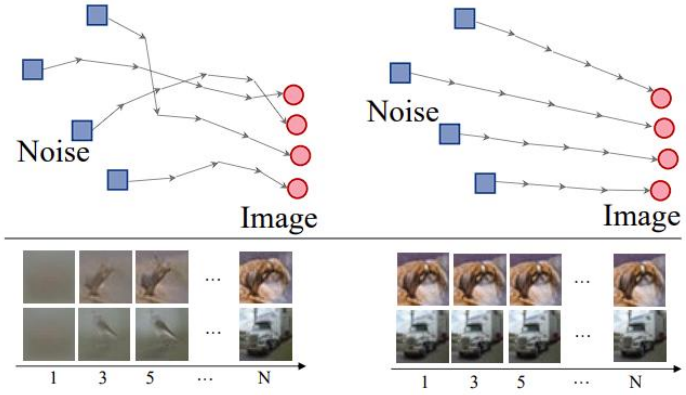
- Diffusion process create a vector field that point from data to noises
- Generation process reverses the vector field

Key points of diffusion model

- Quality and computational cost scales with number of iterations
  - Can we speed-up?

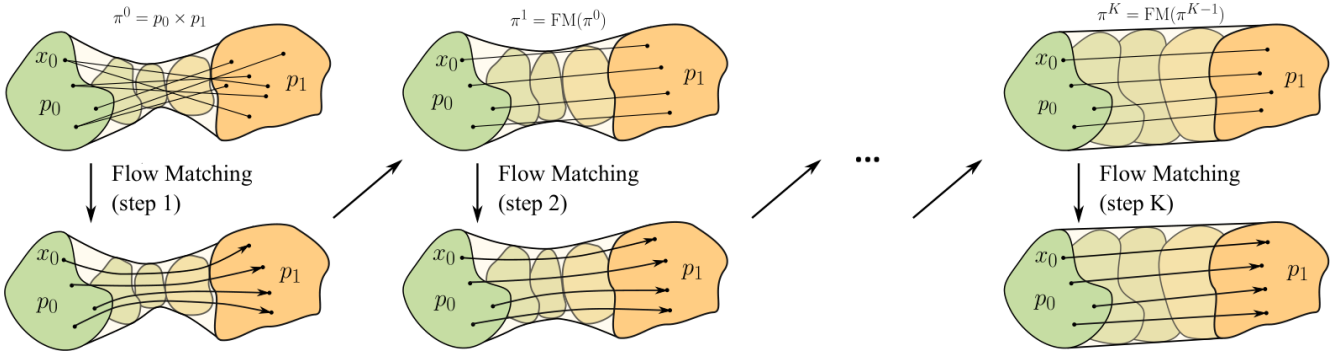
- Stochastic noise addition is just one of many ways to map data points to a random distribution
  - Can we model other vector fields to map the data?
  - With more regular, straight paths ← easier to train and faster to generate
  - With easier control of the characteristics of the generated data

Flow matching



Xing, S. et al. https://arxiv.org/pdf/2311.16507 (2025)

- Condition the model to learn straight paths
- Actual learned path will still be slightly curved (due to vector field's complexity)
- Improve over iterations

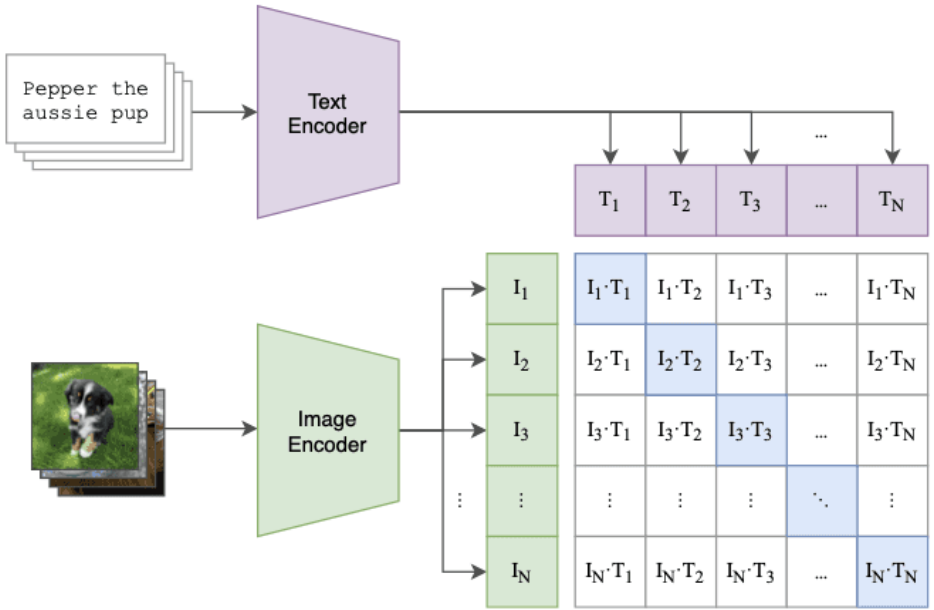


User define the time steps and initial (p_0, p_1) mapping – randomly or guided by data

Kornilov, N. et al. NeurIPS Poster:93864 (2024)

Cool things you can do to guide the learning of the AI

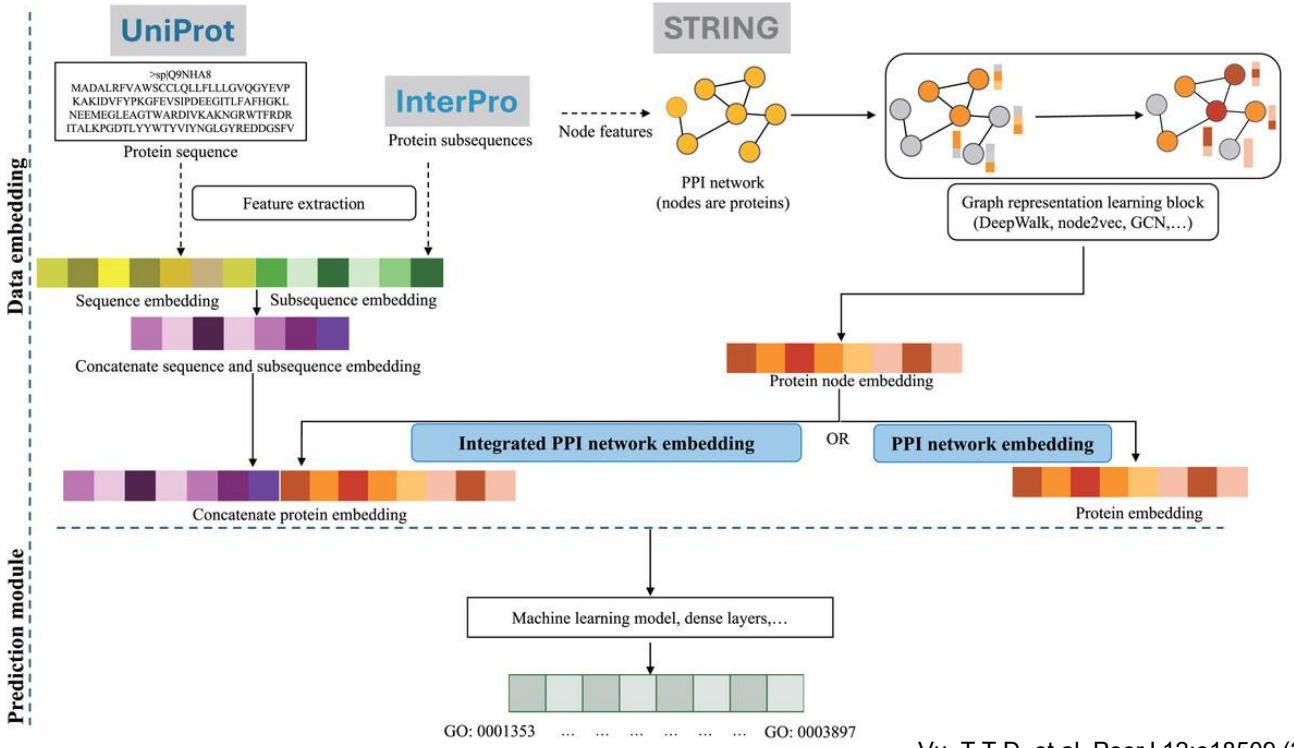
Contrastive training for multi-modal representations



https://www.twelvelabs.io/blog/multimodal-embeddings

- **Force representation for paired multi-modal data to be similar**
- Can search one data type using prompt from another data type
- Basis behind image generation from text prompt

Forcing a shared representation space

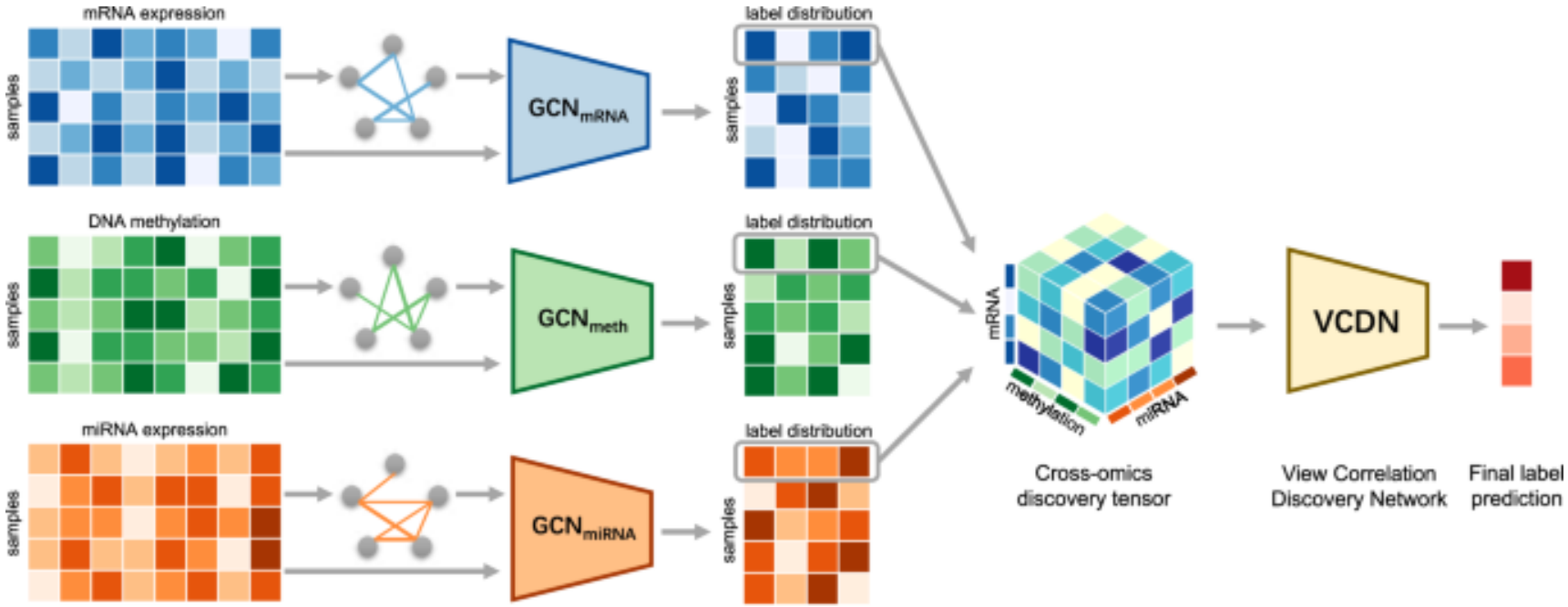


- Condition the learning to **make representations from multiple raw data types computationally compatible**

- Additive, concatenate, etc.

Vu, T.T.D. et al. PeerJ 12:e18509 (2024)

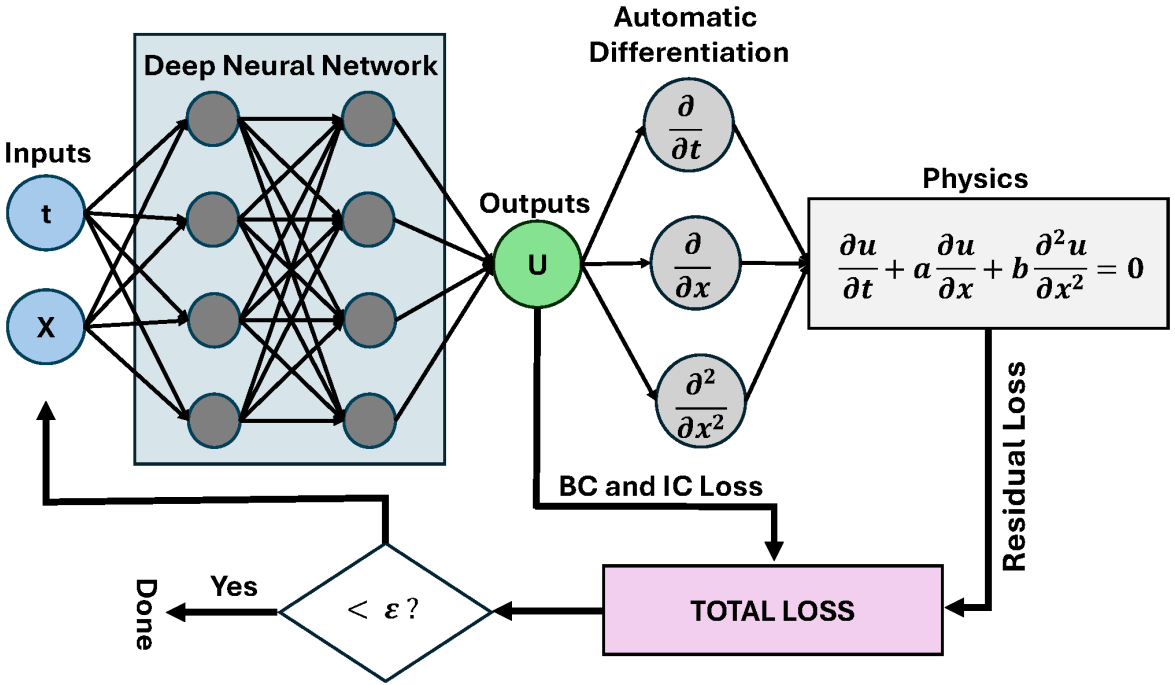
Multi-omics integration via shared representation



Wang, T. et al. Nature Communications 2021

Forcing model to mimic natural mechanisms

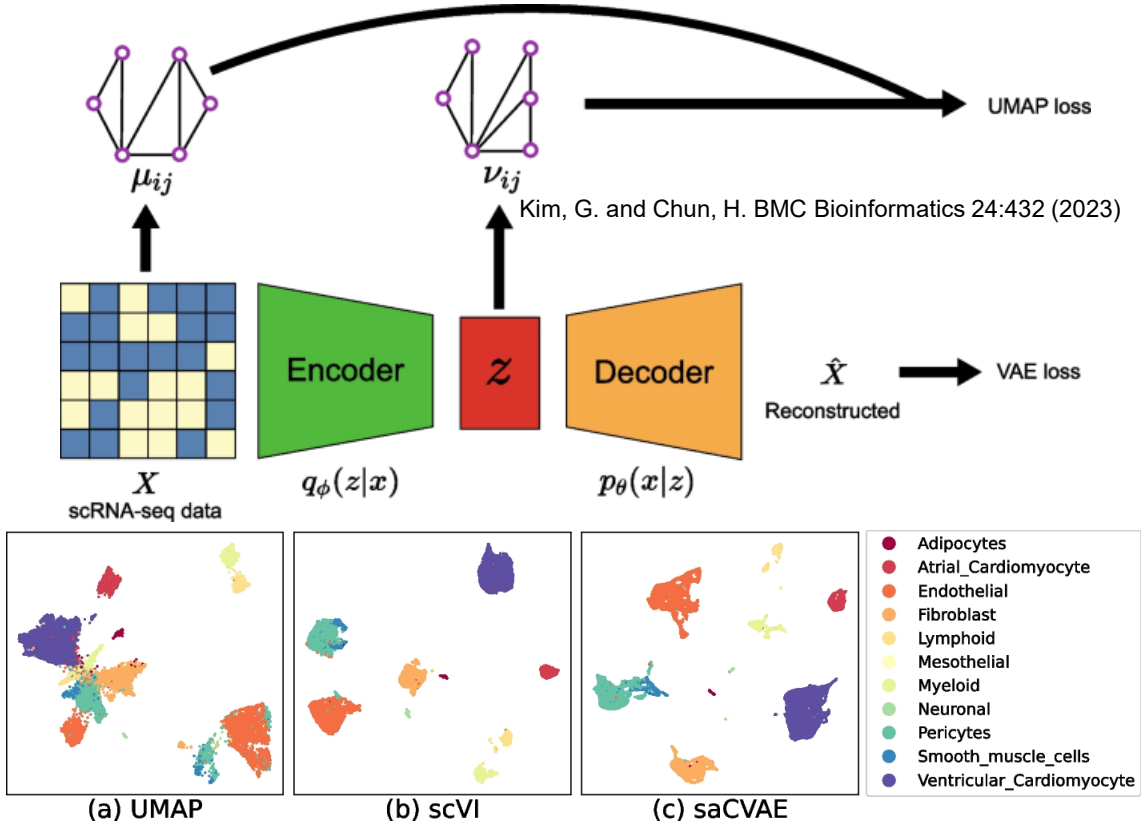
- **Physics-Informed Neural Network**
- During training, add loss functions that compare the model's output behavior to laws of physics



Trahan, C. et al. Entropy 26:649 (2024)

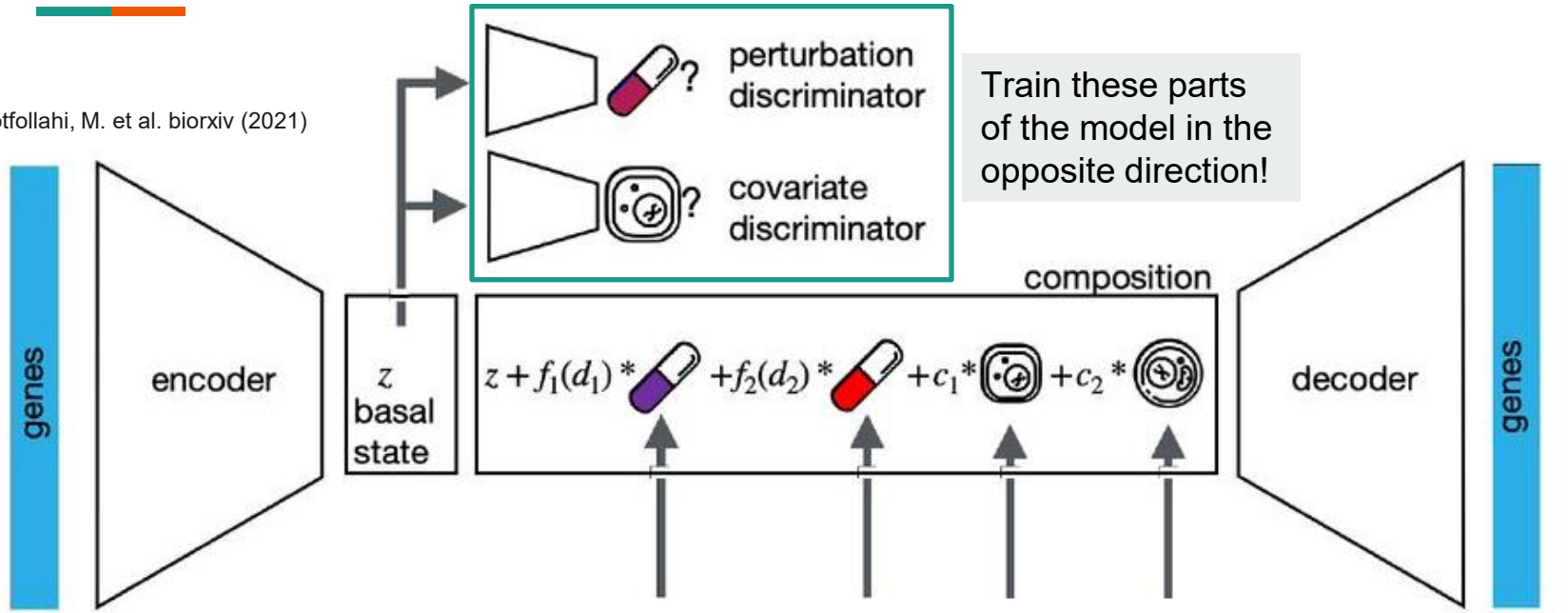
Guided representation with sample-similarity network

- **Key**: Discriminative ML does not utilize sample-sample similarity
- Guide representation by forcing it to mimic UMAP reconstruction of sample similarity network



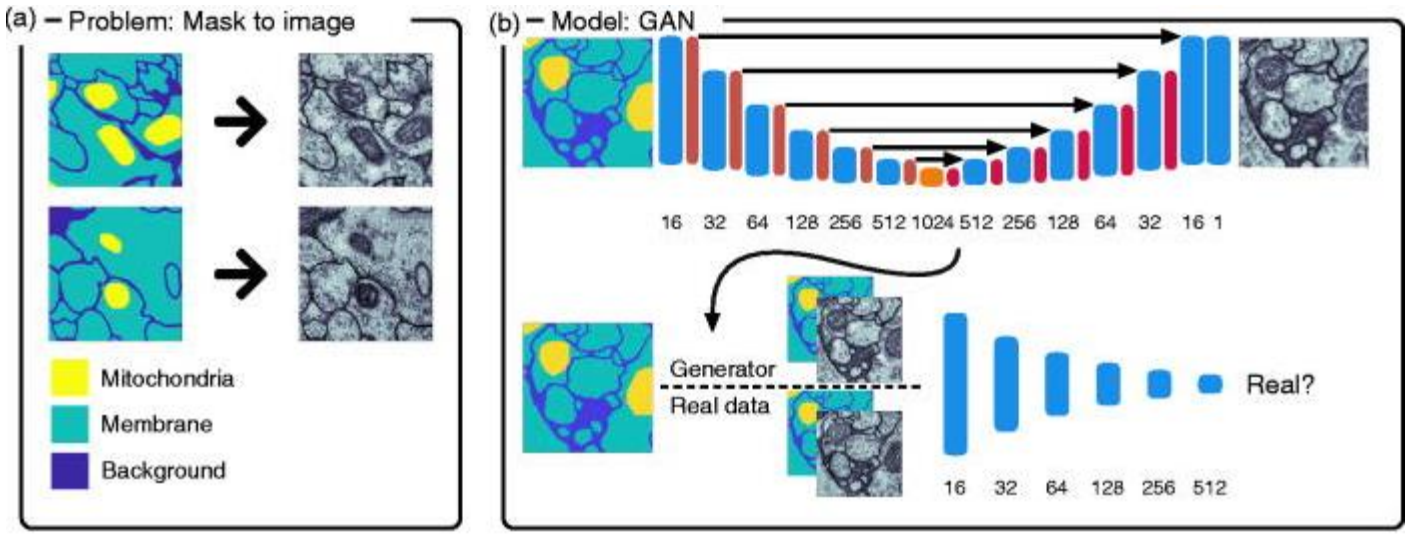
Debiasing

Lotfollahi, M. et al. biorxiv (2021)



- Sometimes, we want to remove batch effect/bias in the data
- Can we train the model to “ignore/lose” knowledge?

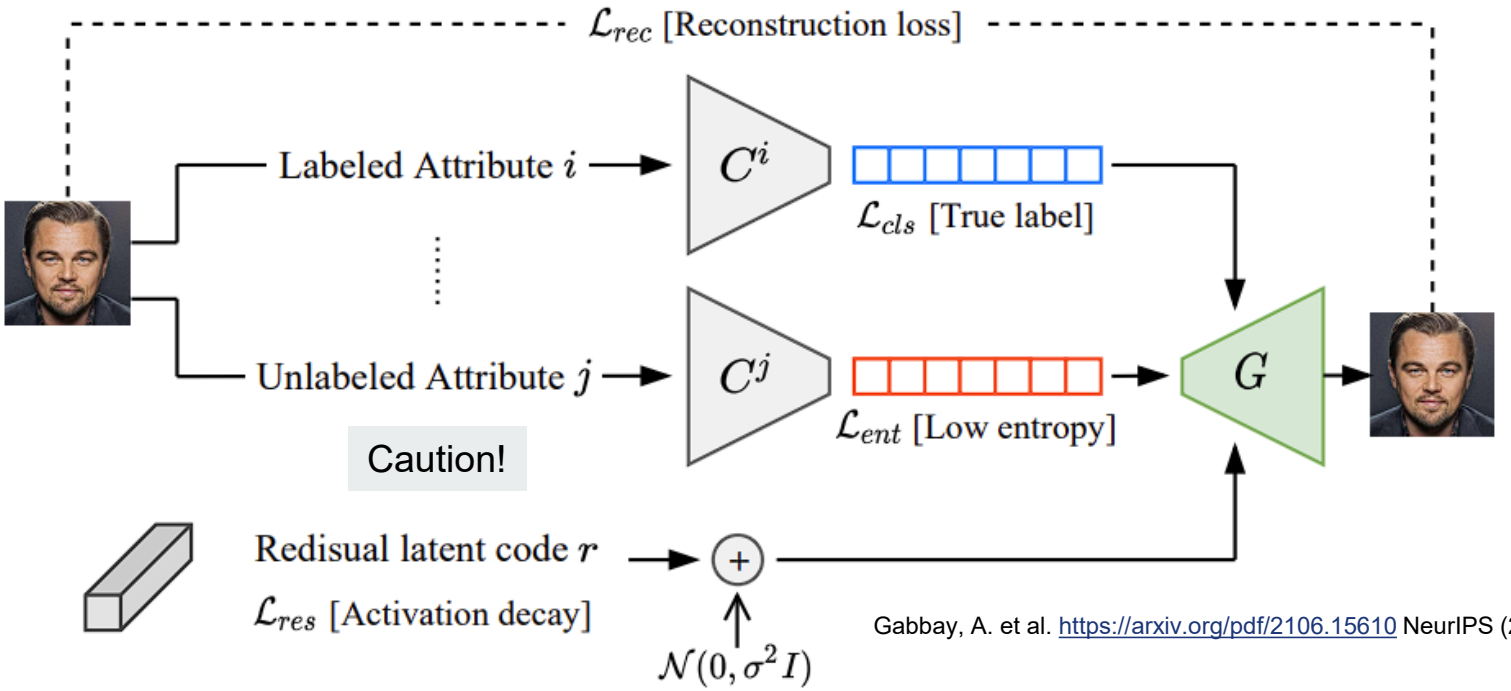
Conditional generative process



Midtvedt, B. et al. Applied Physics Review 8:011310 (2021)

- Add conditions as input to the **Generator** and the **Discriminator** to guide the generated data distribution and characteristics

Learning to disentangle



Gabbay, A. et al. https://arxiv.org/pdf/2106.15610 NeurIPS (2021)

- Simultaneous learning of known and unknown factors and generation

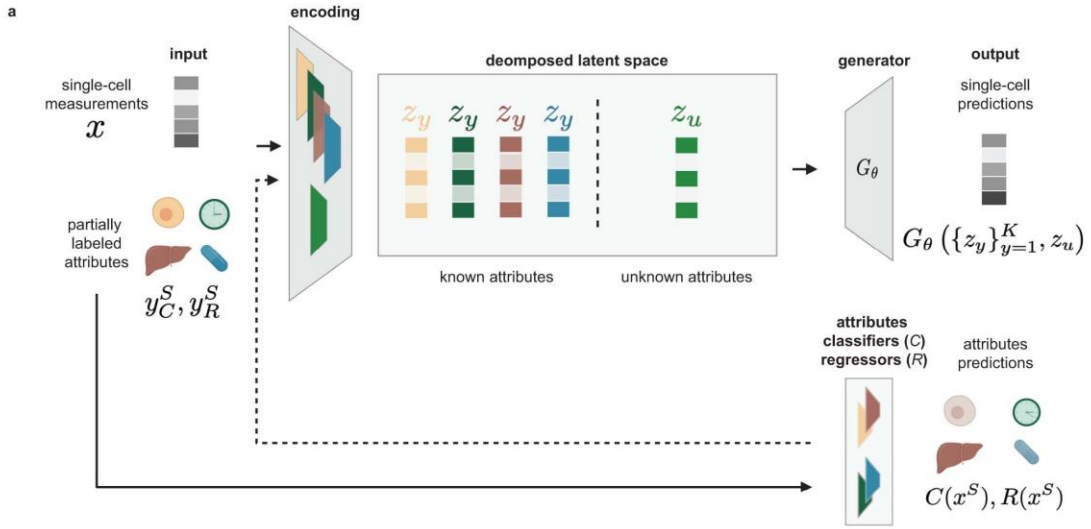
Applications in biology

Roles of generative models in biology

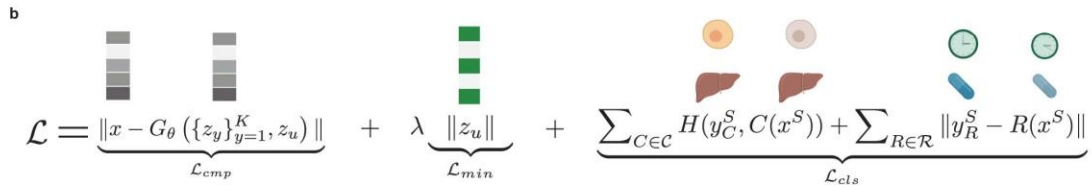
- Imputation with synthetic data
  - Generate data with desired treatment / disease effects
  - Analyze as if they are real observations
- Disentangle factors that affected the observation
  - Through conditional generative process
- Integrate multi-omics data
- Molecular design

Disentangle factors influencing gene expression

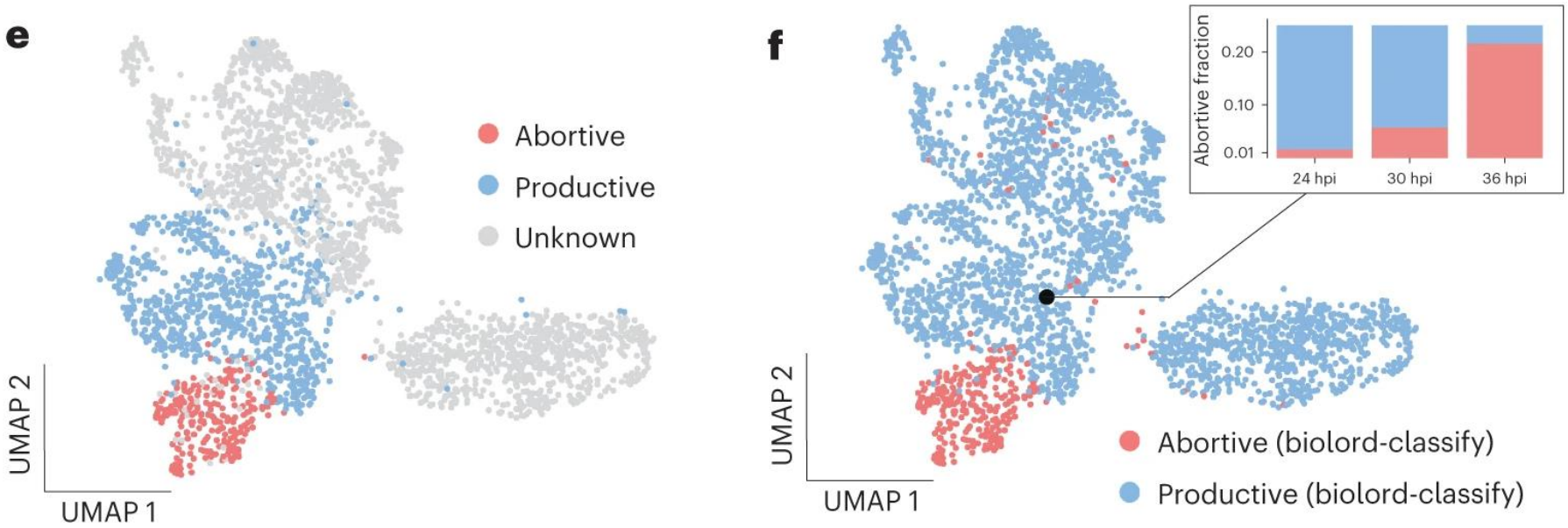
- Distribution of generated data compared to real data, stratified by factors
- Regularization of unknown factors
- Accuracy of known factor predictions



Piran, Z. et al. Nature Biotechnology 42:1678-1683 (2024)



Label imputation

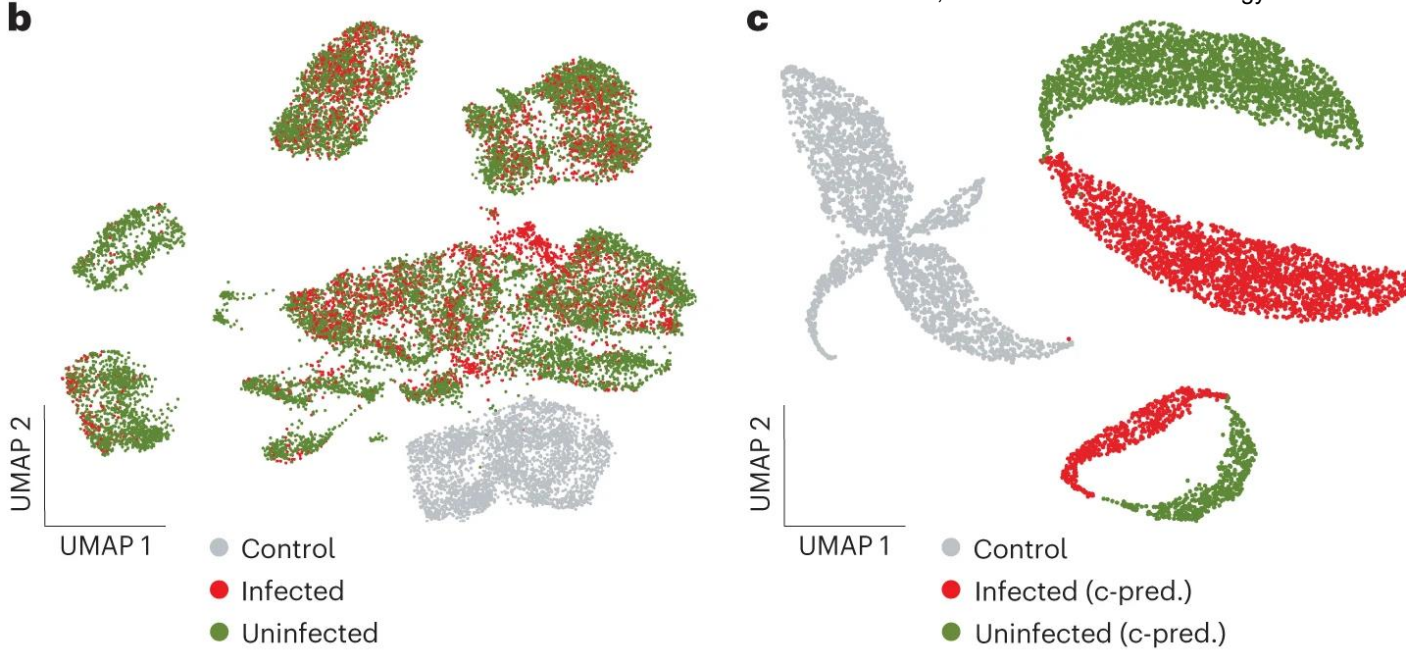


Piran, Z. et al. Nature Biotechnology 42:1678-1683 (2024)

- Use the classifier part of the framework

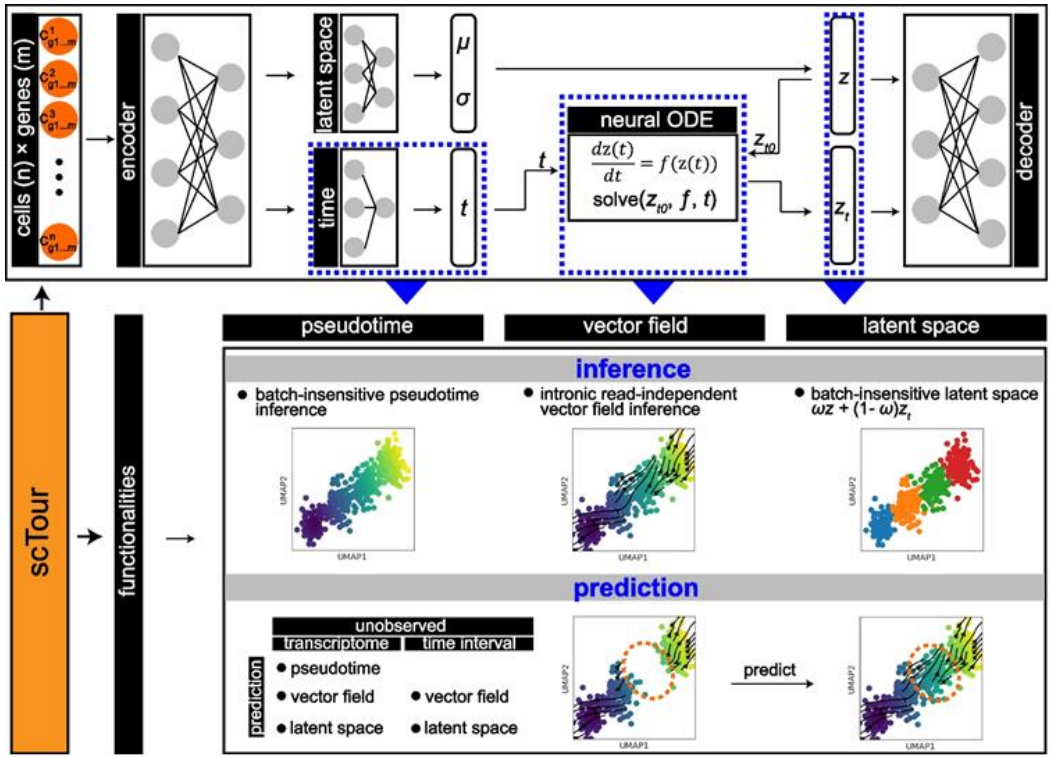
Counterfactual inference

Piran, Z. et al. Nature Biotechnology 42:1678-1683 (2024)



- Simulate infected and uninfected gene expressions from control cells

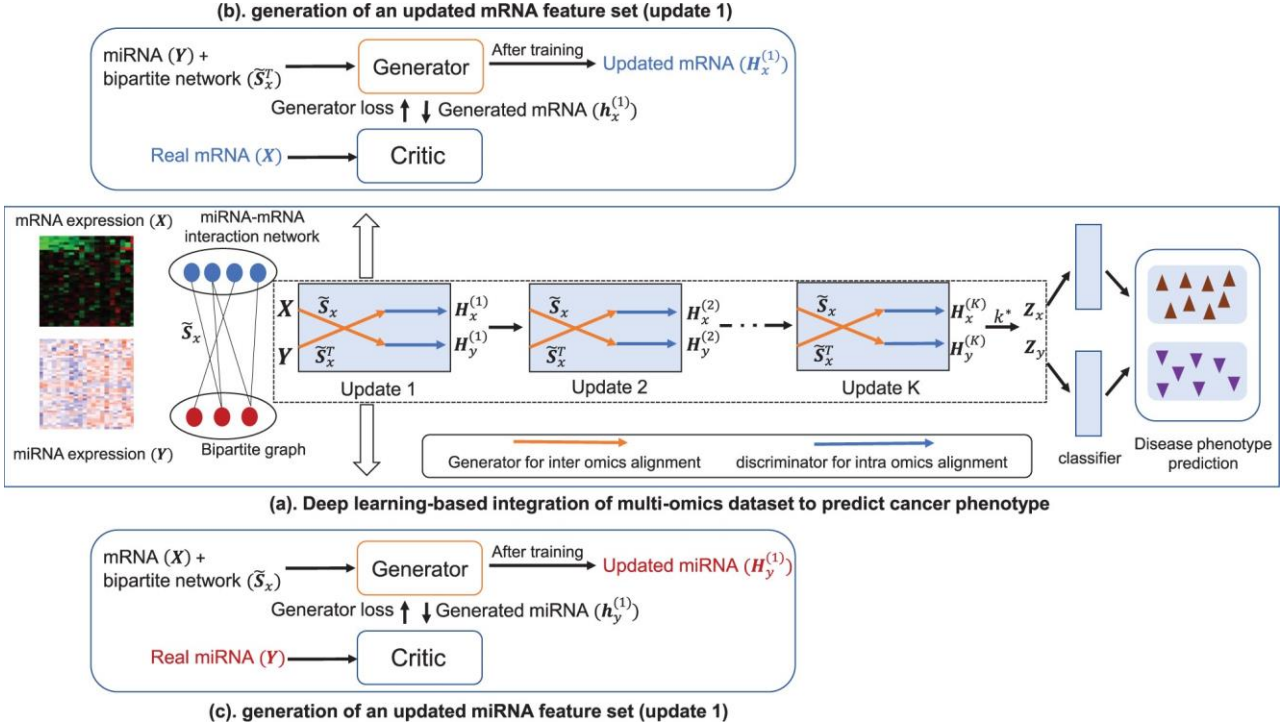
Neural ODE fitting to single-cell trajectory



- **Assumption**: Changes in gene expression during cell development follow ordinary differential equation (ODE)
- Joint estimation of time, vector field, and cell states
- Use inferred time to reconstruct gene expression

Li, Q. Genome Biology 24:149 (2023)

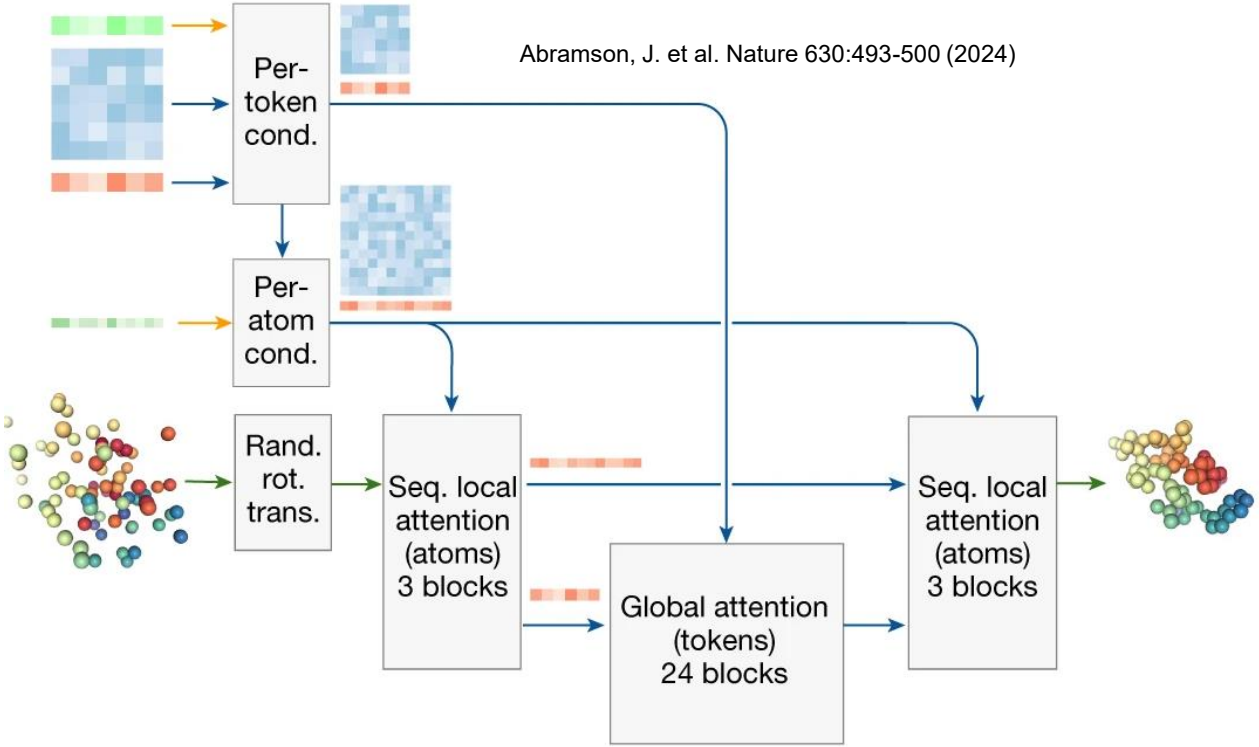
Multi-modal generative process



Ahmed, K.T. et al. Bioinformatics 38:179-186 (2022)

- Use vertical, cross-omics relationships to guide the co-generative process
- Use different omics as feedback for the generation of the other omics

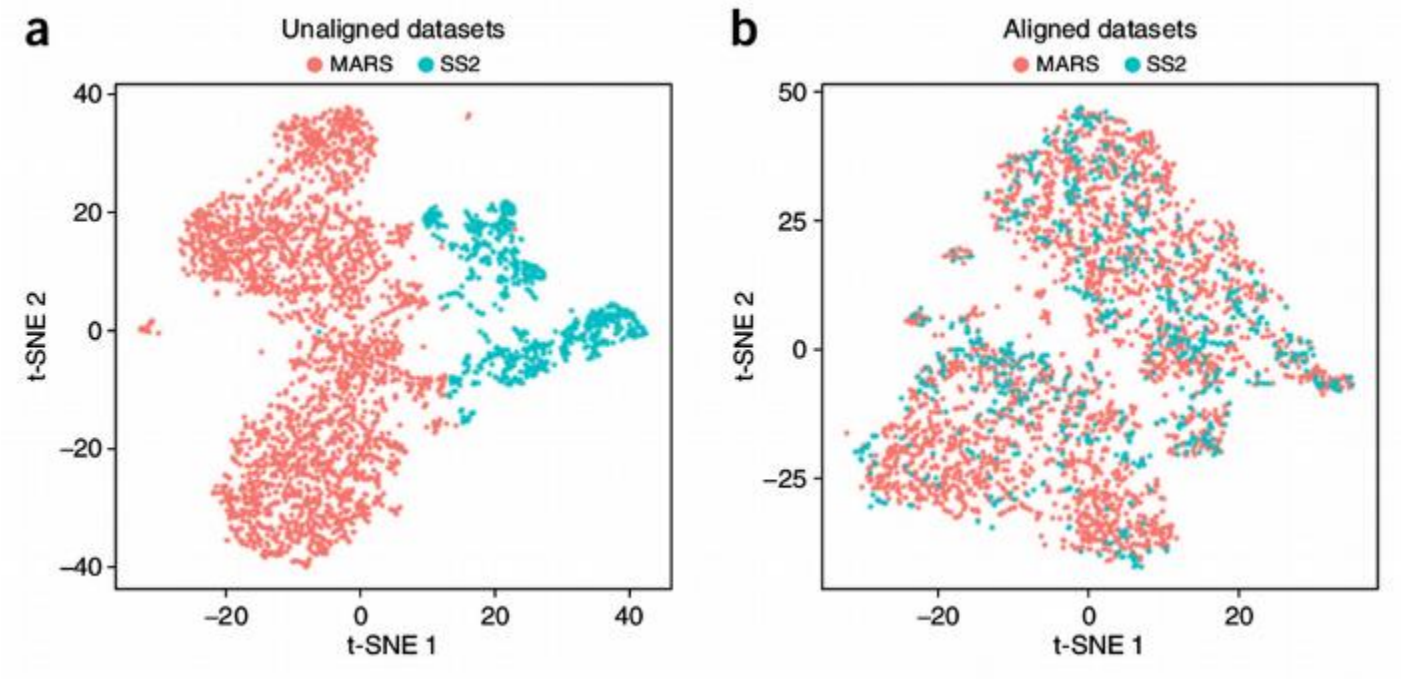
Diffusion module in AlphaFold v3



- Guided by representations of input sequences and pairwise distances
  - Fed into the denoising steps
- Combine denoising at high-resolution and low-resolution

The need for foundational approaches

High bias across biological datasets



https://towardsdatascience.com/how-to-batch-correct-single-cell-7bad210c7ae1

Model that understand data across batches

- Train model across multiple datasets, batches, and modalities
  - No common target output
  - Learn to impute / generates

- A **foundation model** is an AI model that was trained on diverse datasets and was able to capture the essences of the data that generalize across batches
  - Can be used off-the-shelf (zero-shot learning)
  - Can be fine-tuned on a target dataset (few-shot learning)
  - Can be adapted for diverse prediction tasks

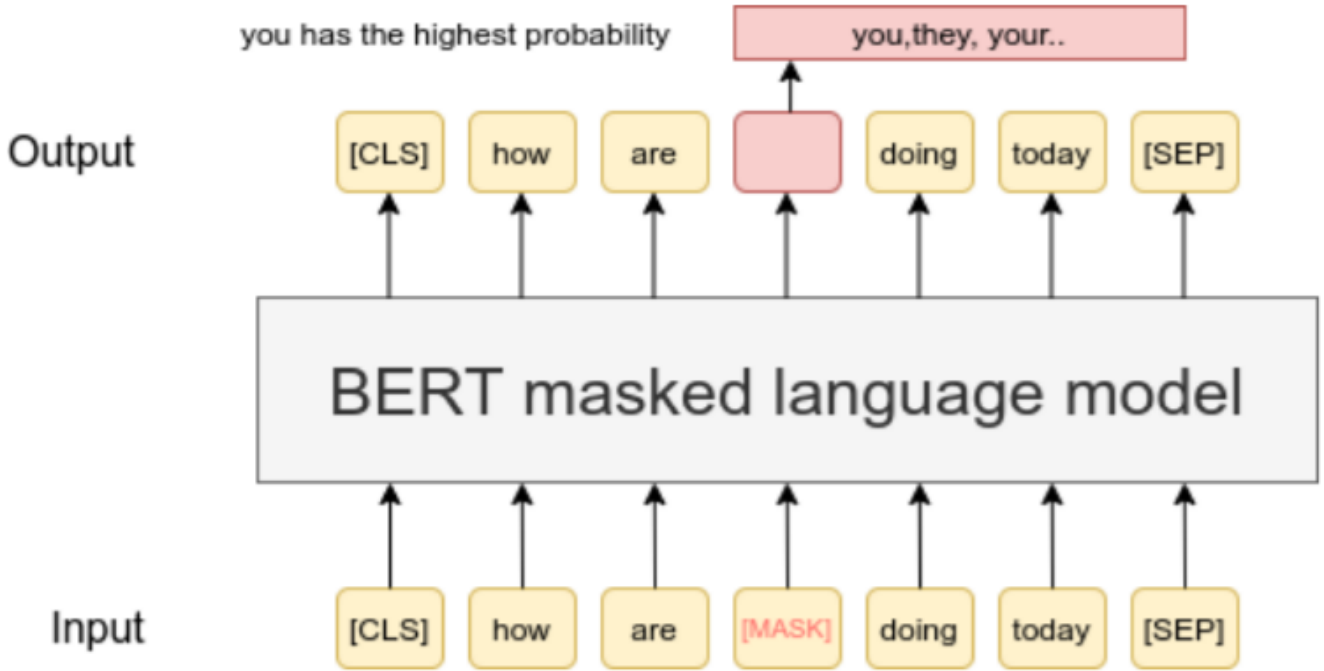
Masked training and next-token prediction

- Masked training
  - Intentionally withholding pieces of the data
  - Train the model to predict the missing values
  - Learn from partially observed context

- Next-token prediction
  - For sequence data
  - Given early entries, predict the next entries

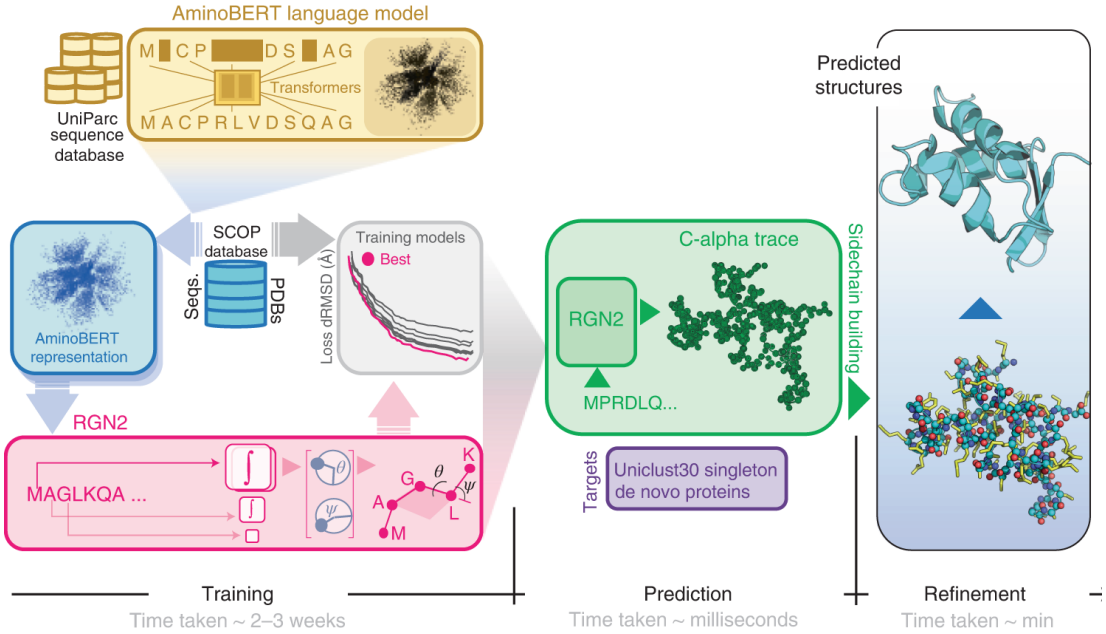
- Dataset-agnostic training targets

Masked language model (LM)



https://medium.com/@shaikhrayyan123/a-comprehensive-guide-to-understanding-bert-from-beginners-to-advanced-2379699e2b51

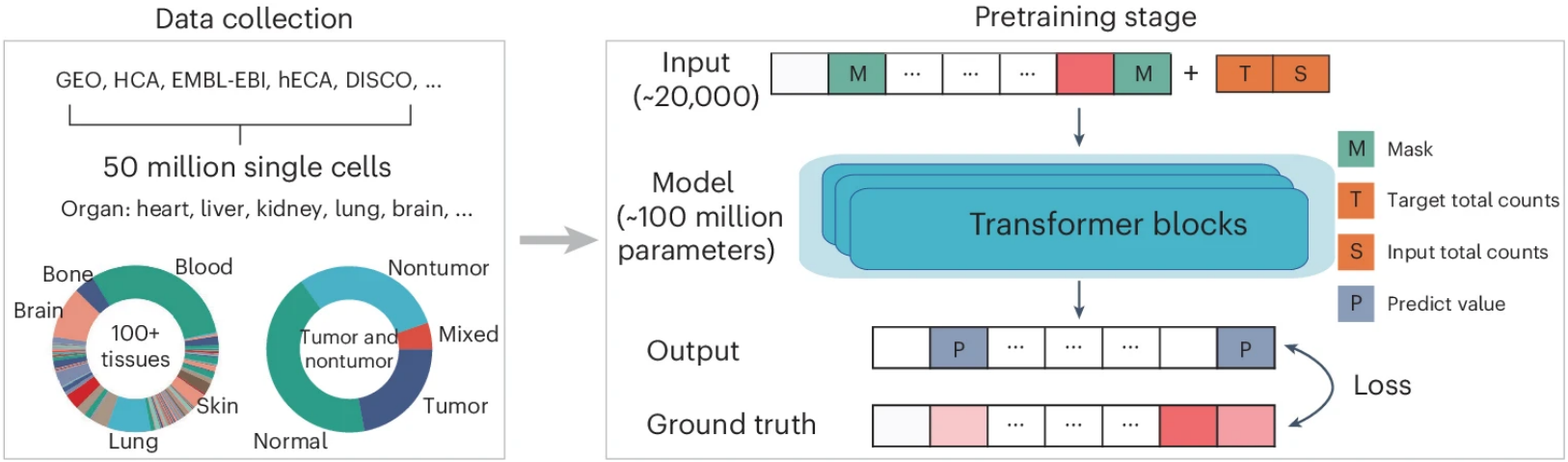
MSA-free protein LM for structural prediction



- More emphasis of masked LM training
  - Learn co-evolution
- Use LM embedding to replace MSA
- Good for orphan proteins (no relatives) and engineered proteins

Chowdhury, R. et al. Nature Biotechnology 40:1617-1623 (2022)

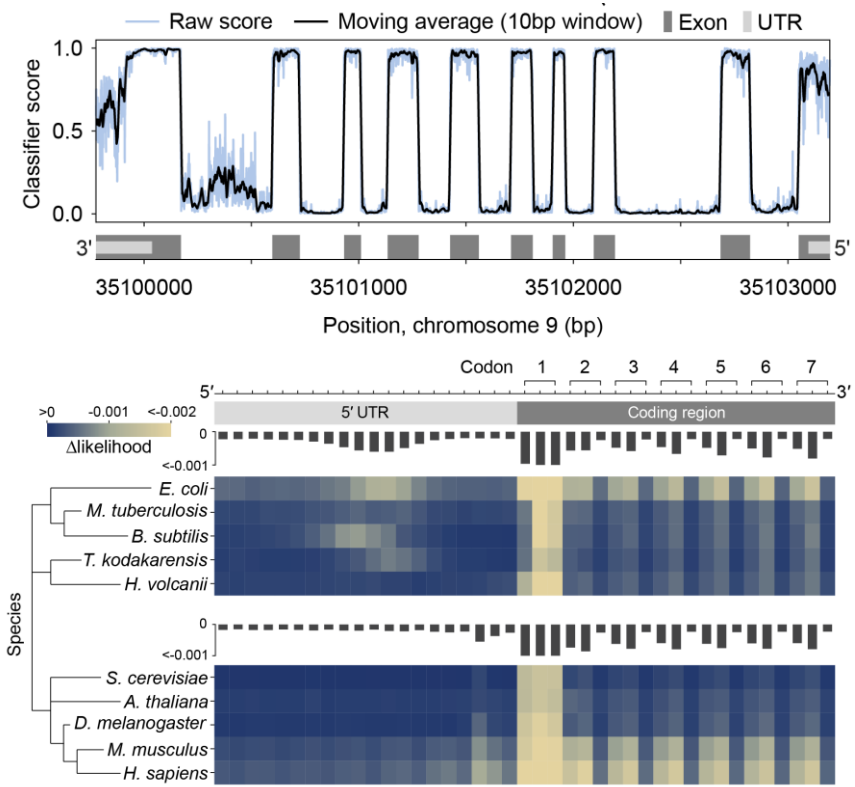
A foundation model for single-cell transcriptomics



Hao, M. et al. Nature Methods 21:1481-1491 (2024)

- Trained on >50 millions single-cells
- Masked training: gene expression levels
- Adapted to predict drug response, cell type annotation, etc.

EVO2: A foundation model for genomics

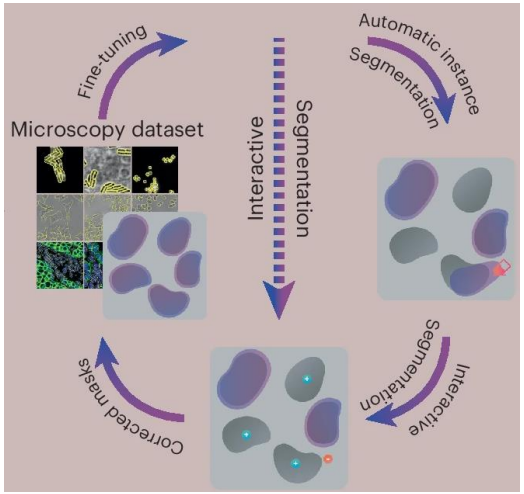


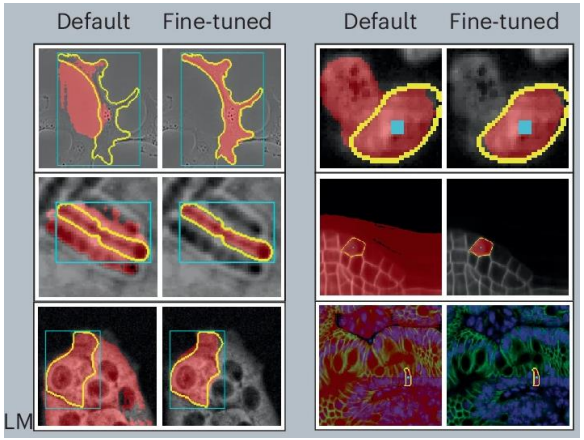
- Trained on all reference genomes
- Next nucleotide sequence prediction
- Learned intron-exon boundaries and the codon structure of genes
  - Built into the representation
  - No explicit training

Brixi, G. et al. https://www.biorxiv.org/content/10.1101/2025.02.18.638918v1 (2025)

A foundation model for microscopy

- **Segment Anything for Microscopy**
- **Default**: Trained on 1 billion objects in non-biology photos
- **Fine-tuned**: Update model using images from specific microscopic modality
- Reduced data requirement for creating a new model on your own dataset





Archit, A. et al. Nature Methods 22:579-591 (2025)

Summary

- Generative approach enhances AI's understanding of real-world data through learning to create realistic synthetic data

- **Assumption**: With enough data diversity, if the AI can generate, which is a very difficult task, it must have learned to mimic the actual mechanisms that produced the data

- In contrast to typical supervised AI which may understand the data only on some aspects necessary for making accurate predictions

Any question?

- See you next time