Introduction to Deep Learning (DL) in Genomics

Jenny Wu, Ph.D.
Genomics Research and Technology Hub
Chao Family Comprehensive Cancer Center
Sept 23, 2025

Outline

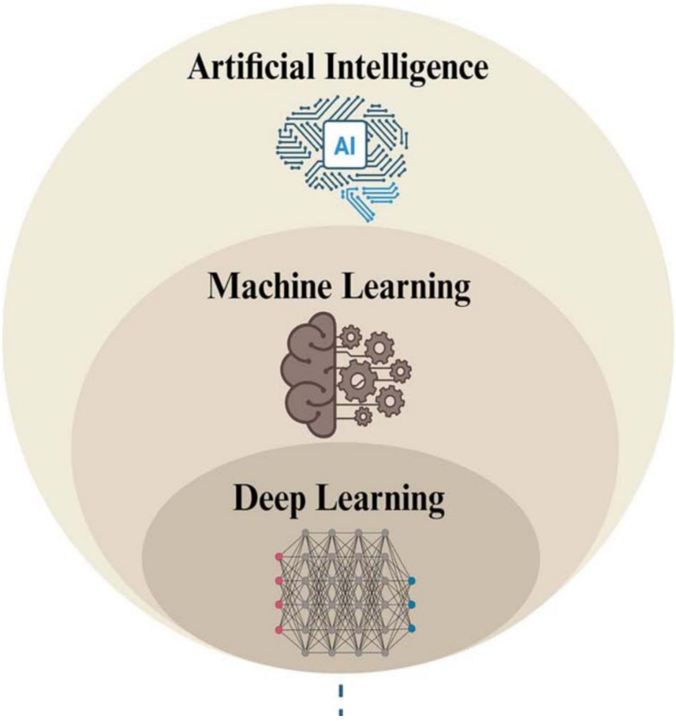
- What is deep learning and why
- Basic Concepts: perceptron, model architecture, training
- DL applications in Genomics
- DL tools in Genomics
  - ✓DeepBind
  - ✓Cellpose
  - ✓DeepVariant
- Foundation models
- Challenges and limitations

What is Deep Learning

- Artificial Intelligence (AI): Any technique that enables computers to mimic human behavior

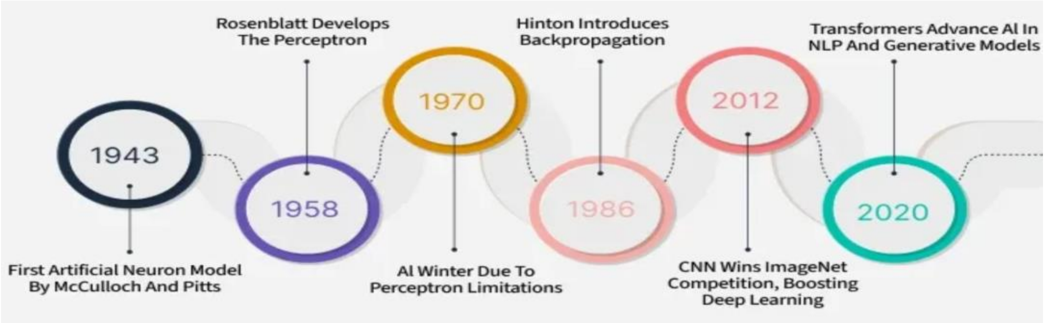
- Machine Learning (ML): Ability to learning without explicitly being programmed

- Deep Learning (DL): Extract patterns from data using deep neural networks, feature engineering, end to end mechanism



Why Now

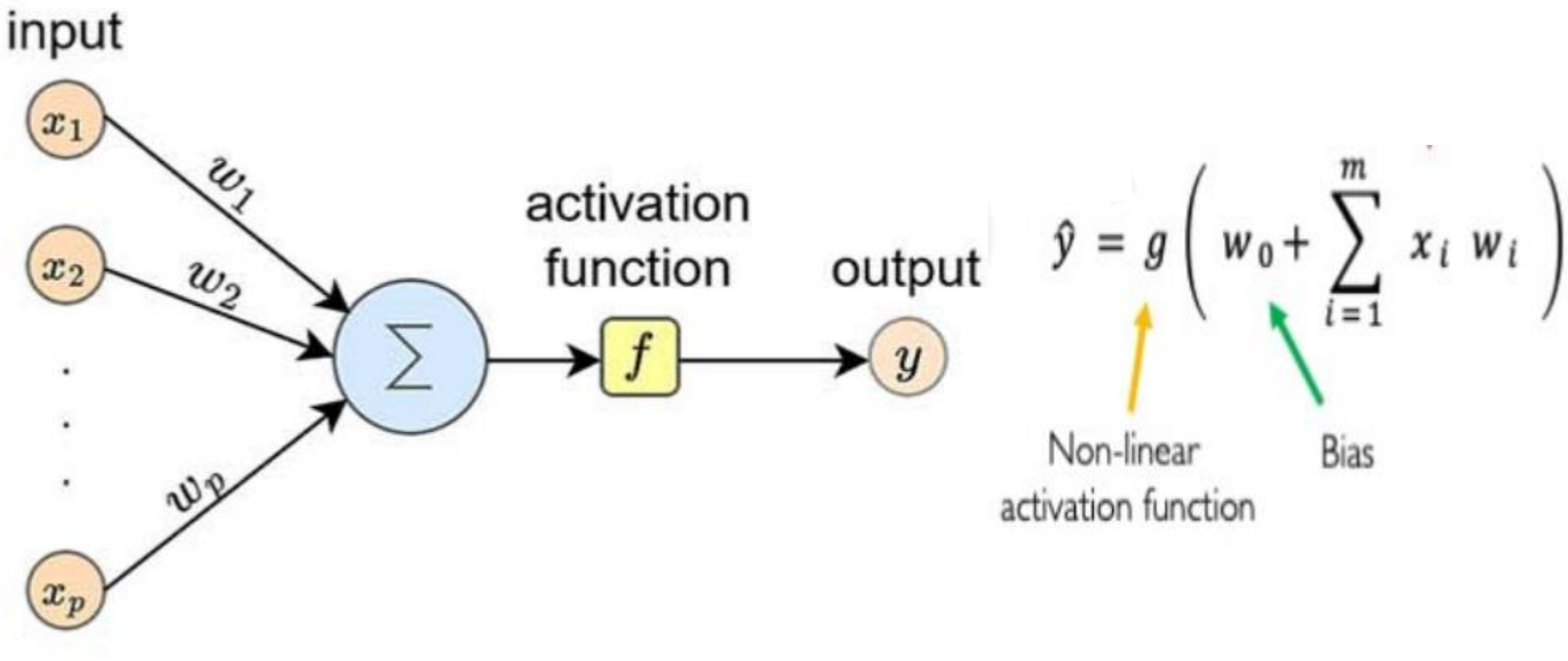
- Big Data: the collection and annotation of extensive training datasets



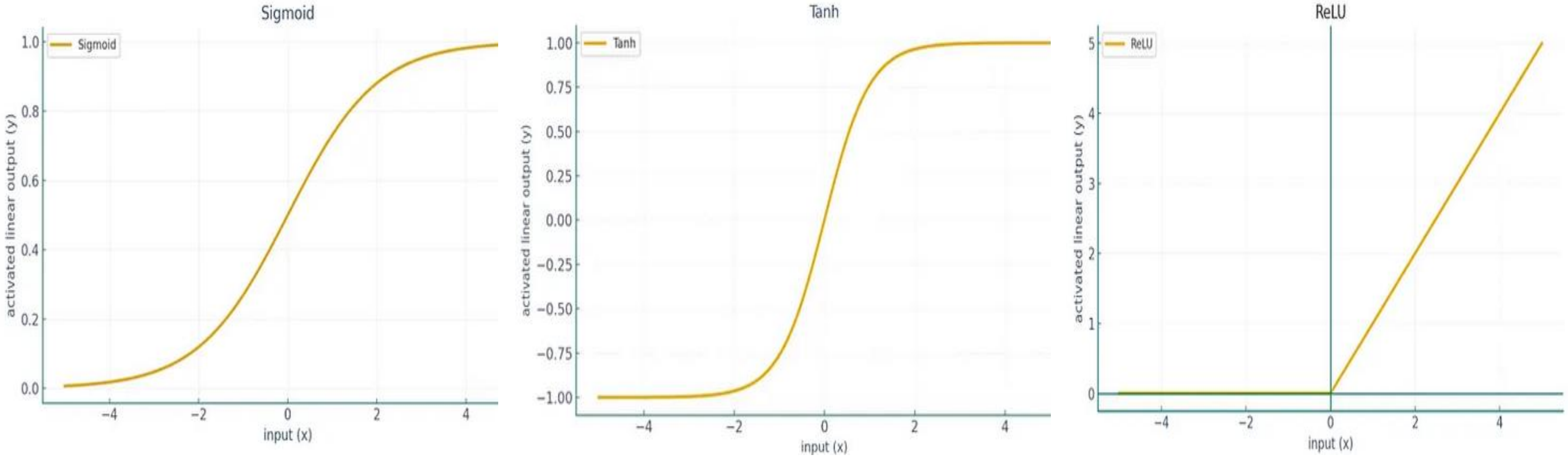
- Hardware: GPU

- Algorithms and software: the availability of toolkits

Basic Unit of NN: The Perceptron

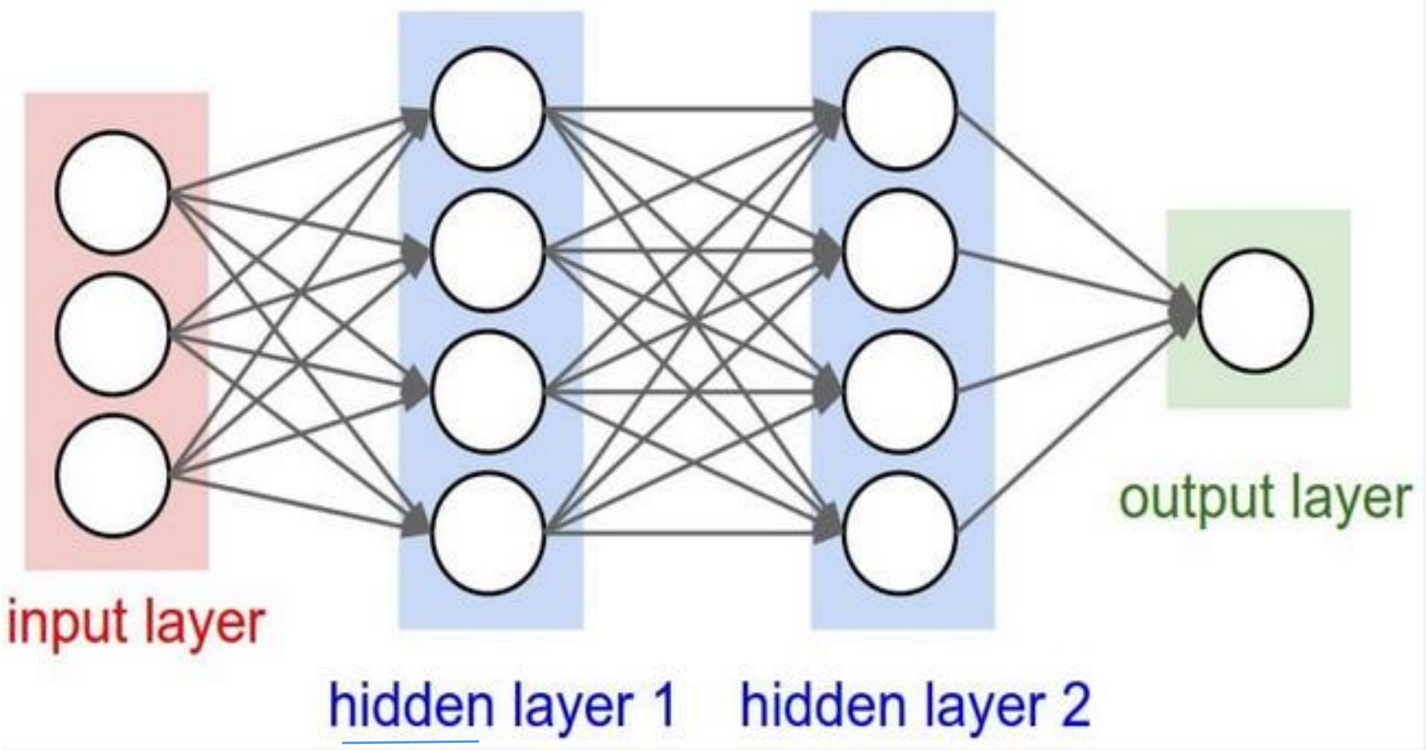


Common Activation Function f



- Add non linearity to model complex real world data

Perceptron to Deep Neural Networks (MLP)



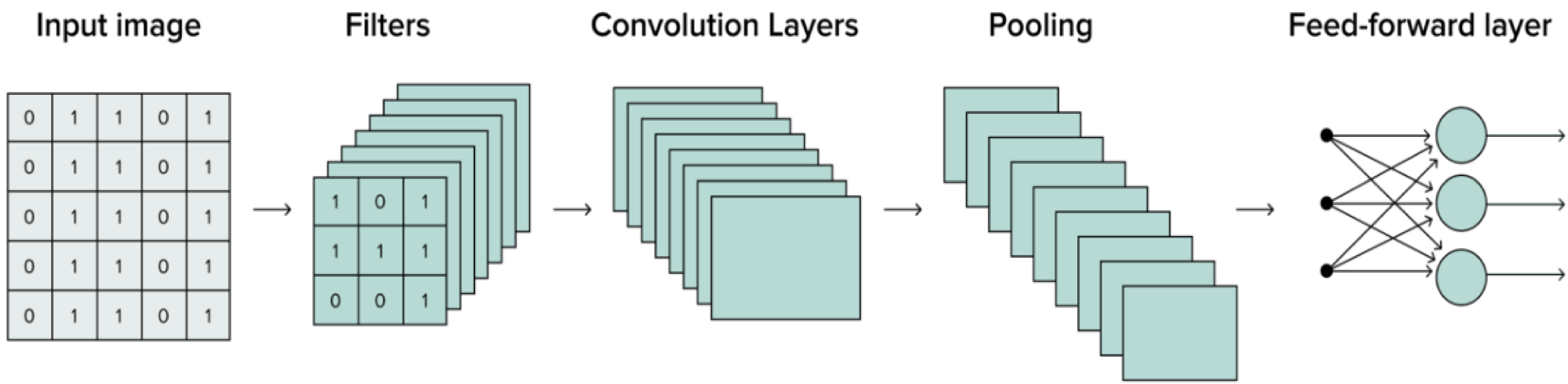
Output layer size depends on the problem definition

Model Architecture

- Many possible model architectures
- Hybrid architecture
- Transformer becomes more popular recently
- Generative or Discriminative

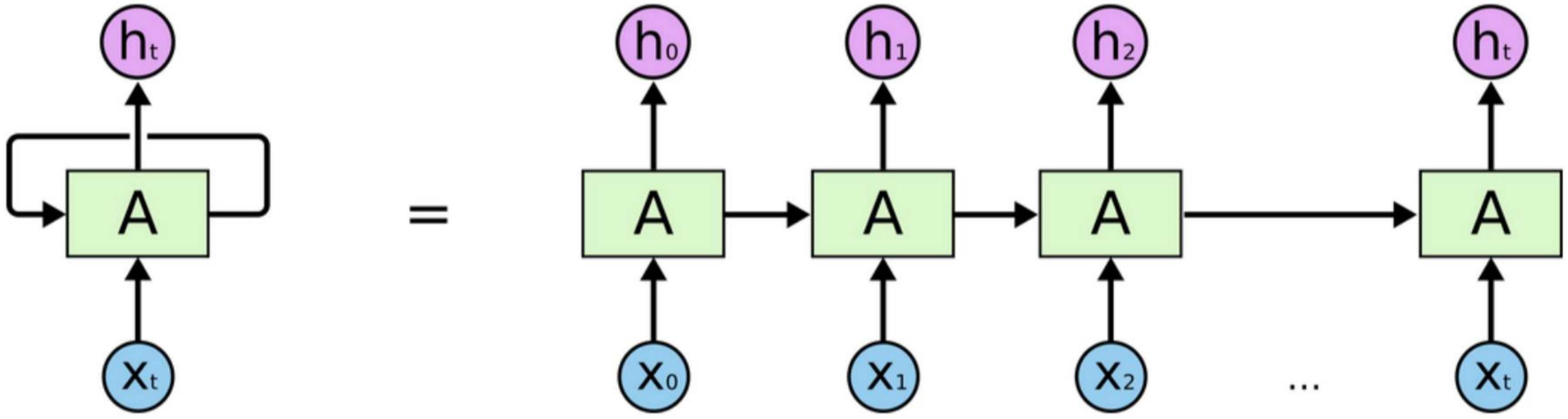
Multilayer perceptron	
CNN	DeepBind cellpose DeepVariant
RNN	
Autoencoder	scVI
Transformer	Enformer nucleotideGPT scGPT

Convolutional Neural Network (CNN)



- Connect patch in input layer to a single neuron in subsequent layer
- Use a sliding window to define connections
- Convolutional layer creates a feature map
- Real models have multiple rounds of convolution+pooling for diverse tasks

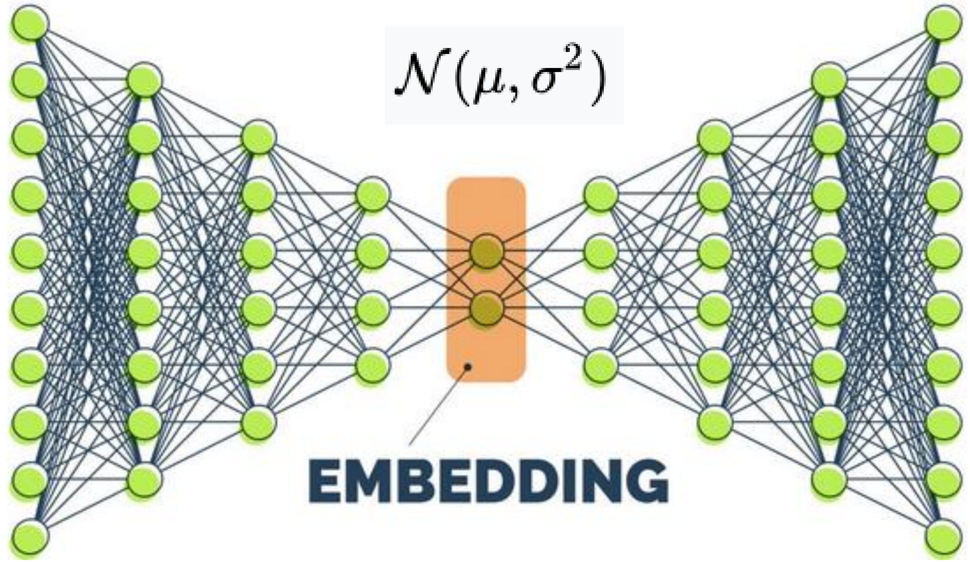
Recurrent Neural Networks (RNN)



- Used to model sequence data using hidden states ht
- Hidden states are a function of previous hidden states and the input at t
- May have vanishing gradients issue (LSTM etc)

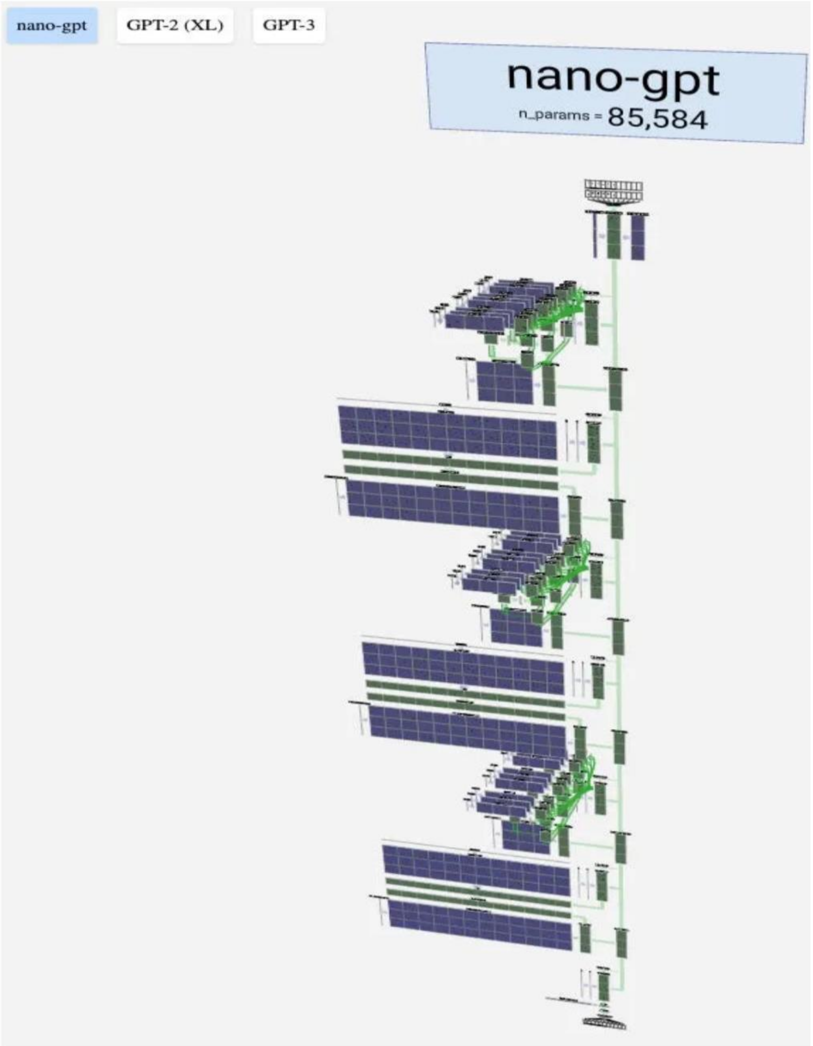
Variational Autoencoder (VAE)

- Compress data into a set of latent features of lower D
- Self supervised
- Interpret hidden latent variables
- Generate new examples



Model Training and Loss Function

- Real DL models have huge amount of parameters that need to be learned from training
- Loss= (actual output – predicted output)^2
- This squaring not only avoids negative values — it amplifies larger mistakes, which makes it easier for the model to focus on correcting them



Common Loss functions in Supervised learning

Squared Error
(Building Block)

Squared difference between predicted and actual values

- Conceptual basis for other cost functions
- Rarely used on its own

- Simple to compute
- Sensitive to outliers
- Often a starting point for understanding error

$$SE = (y - \hat{y})^2$$

MEAN SQUARED ERROR
(Regression standard)

Average of squared errors across all samples

- Predicting continuous values
- Measuring average model performance

- Penalizes large errors more
- Smooths small errors
- Easy to differentiate

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

ROOT MEAN SQUARED ERROR
(Error in real units)

Square root of MSE, gives error in same unit as target

- When interpretability matters
- Same use cases as MSE, but easier to explain

- Same penalty behavior as MSE
- Better for human intuition
- Used in benchmarks and reporting

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

CROSS-ENTROPY
(Classification standard)

Compares predicted probabilities vs. actual class labels

- Binary or multi-class classification tasks
- Probabilistic outputs (e.g., sigmoid, softmax)

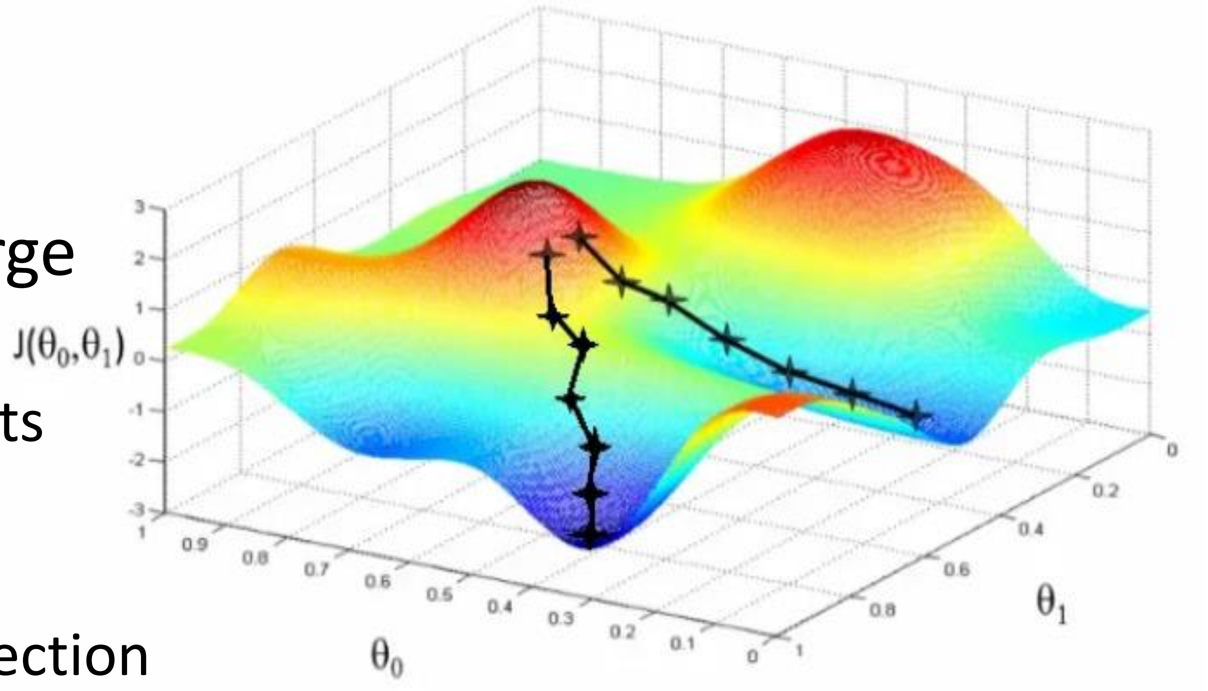
- Penalizes confident wrong guesses
- Great for classification tasks
- Produces clear probabilities

$$CE = -[y\log(\hat{y}) + (1 - y)\log(1 - \hat{y})]$$

Training: Loss Optimization

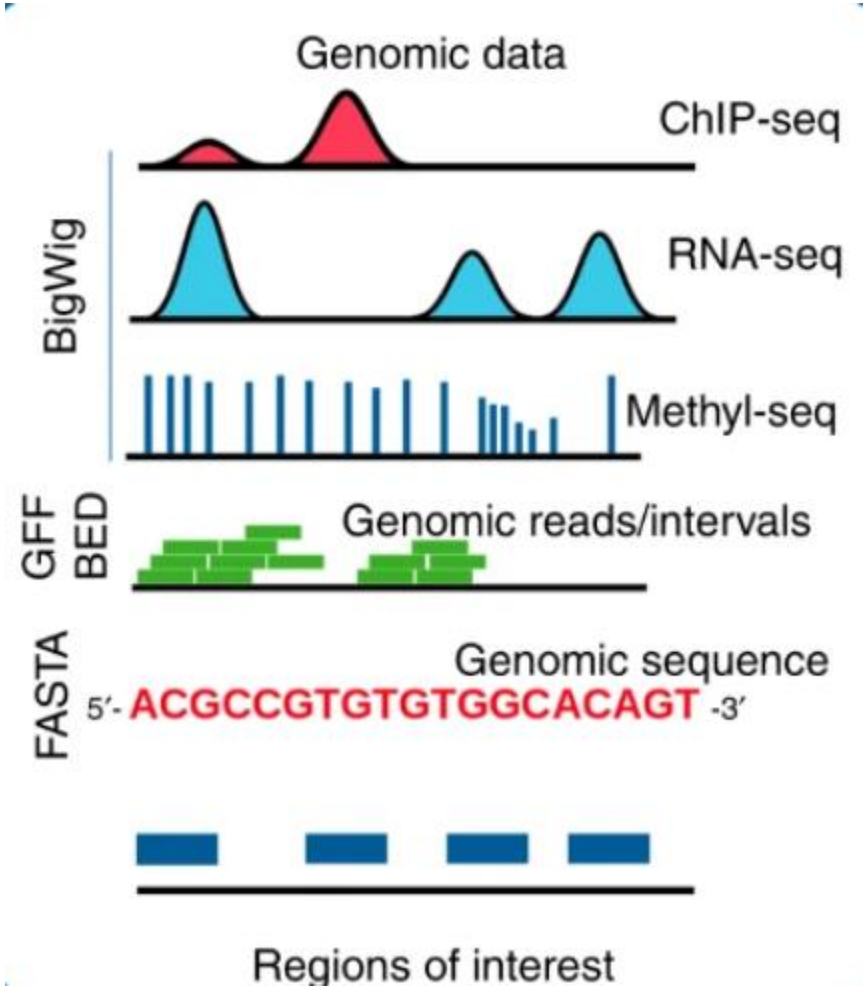
- Find the network weights with lowest loss: loss optimization

- (Stochastic) Gradient descent for large datasets
  1. Pick a random point (initialize weights randomly)
  2. Compute gradient
  3. Take a small step in the opposite direction
  4. Update weights and back to 2 until converge

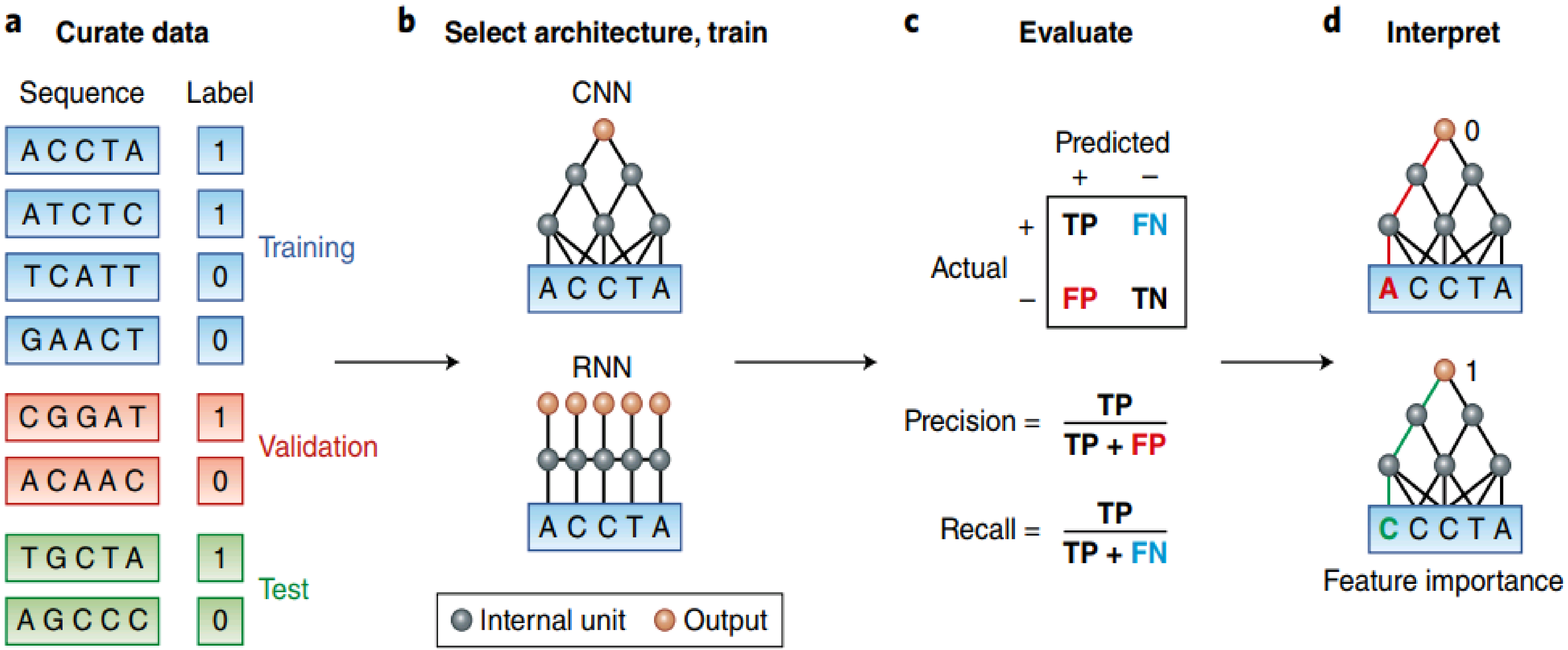


Application of DL in Genomics

- Sequence data: variant calling, motif discovery, functional annotation etc
- Gene expression: expression prediction, GRN inference, disease subtyping
- Epigenomic marks: enhancer prediction, chromatic accessibility modeling etc
- Single cell genomics: dimension reduction, clustering, imputation etc
- Spatial omics: segmentation



DL Workflow in Genomics



Numeric representation

Telenti *et al* 2019

DL Application: Motif Detection w. DeepBind

NATURE BIOTECHNOLOGY | COMPUTATIONAL BIOLOGY | ANALYSIS

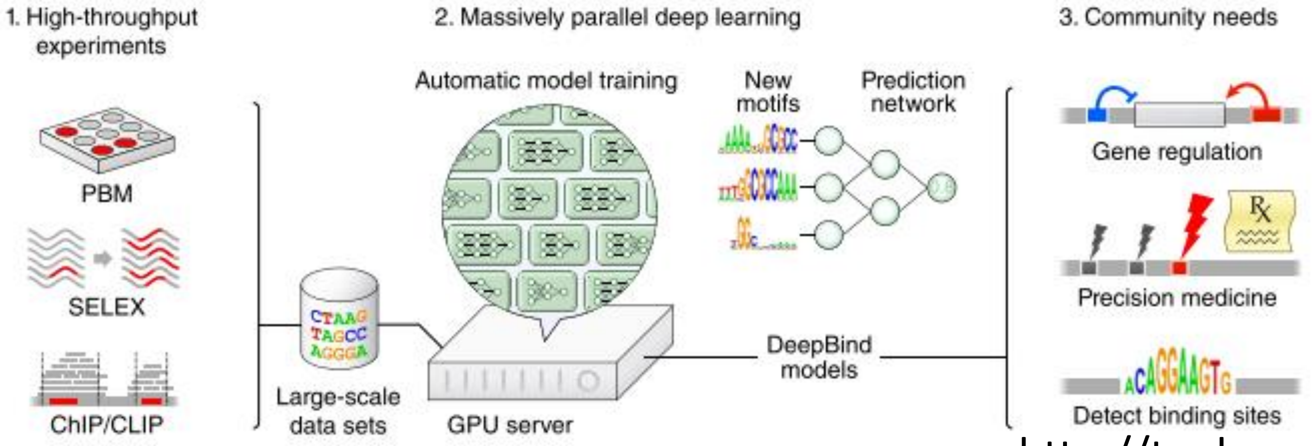
日本語要約

Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning

Babak Alipanahi, Andrew Delong, Matthew T Weirauch & Brendan J Frey

Affiliations | Contributions | Corresponding author

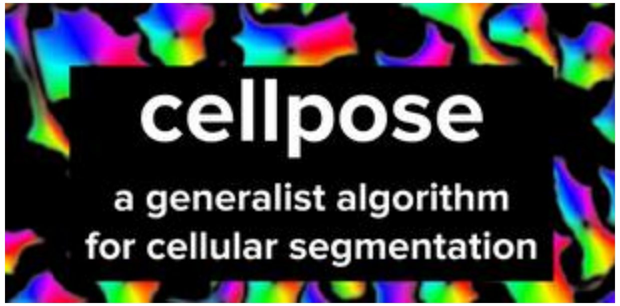
Nature Biotechnology **33**, 831–838 (2015) | doi:10.1038/nbt.3300
Received 28 November 2014 | Accepted 25 June 2015 | Published online 27 July 2015



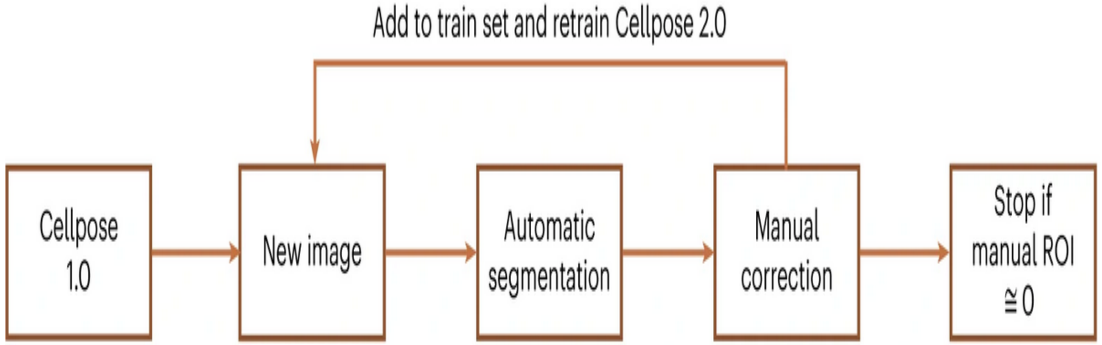
http://tools.genes.toronto.edu/deepbind/

DL Application: Cell Segmentation w Cellpose

- Many of the downstream analyses in imaging spatial omics depend on the ability to resolve individual cells
  - *Cellpose, Stardist, deepcell*



- Cellpose uses CNN
- Cellpose2: human in the loop cellular segmentation
- Cellpose 3: denoising and restoration
- Cellpose-SAM

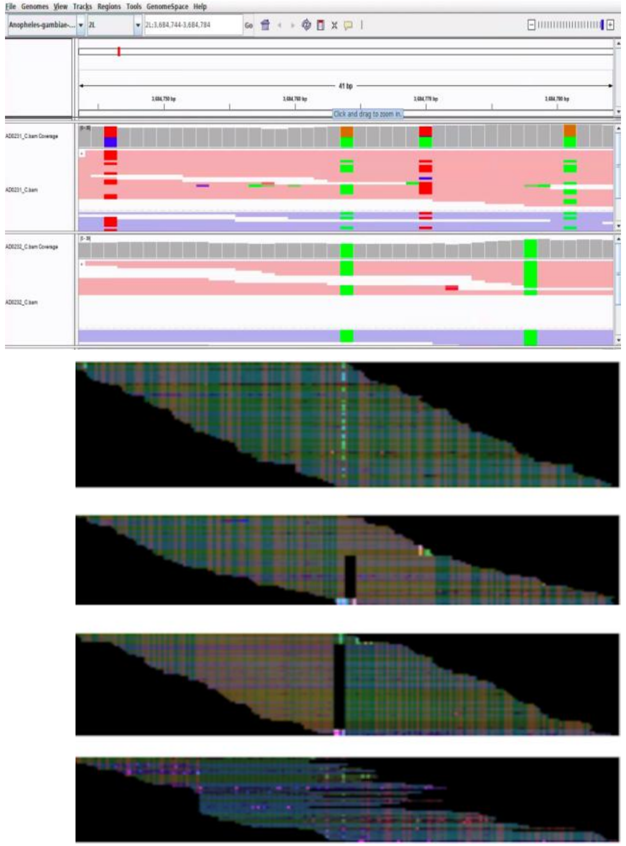


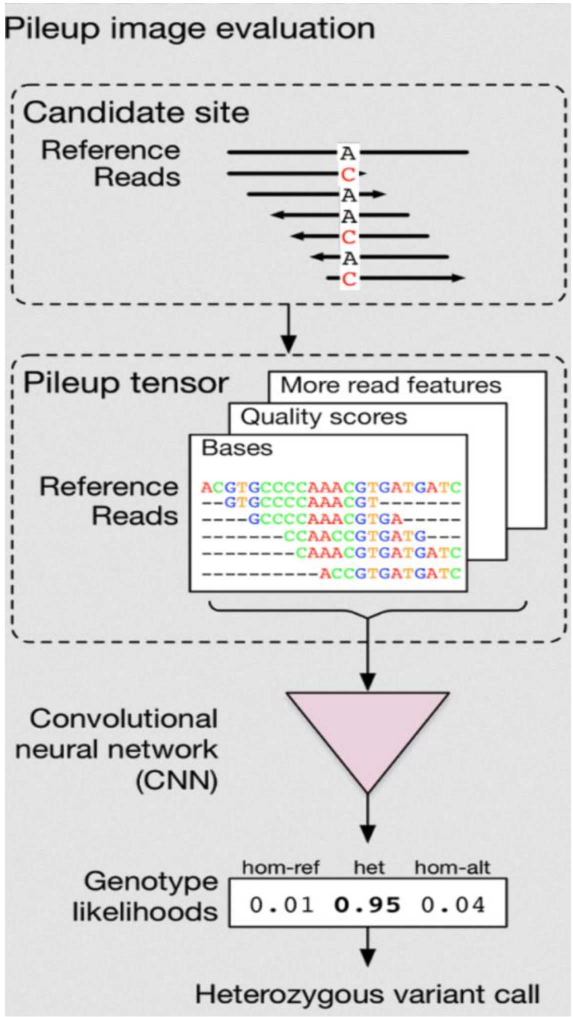
Stringer *et al*, 2024

DL Based Variant Calling

- Data visualization + image classification
- Tensor of (100, 221, 6)
- Using CNN based model (inception 3)
- Can do long reads and hybrid model

DeepVariant





Traditional Way of Model Sharing

Data

Gene Expression Omnibus

ENA
European Nucleotide Archive

Bioinformatics software

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Trained predictive models

Code repository

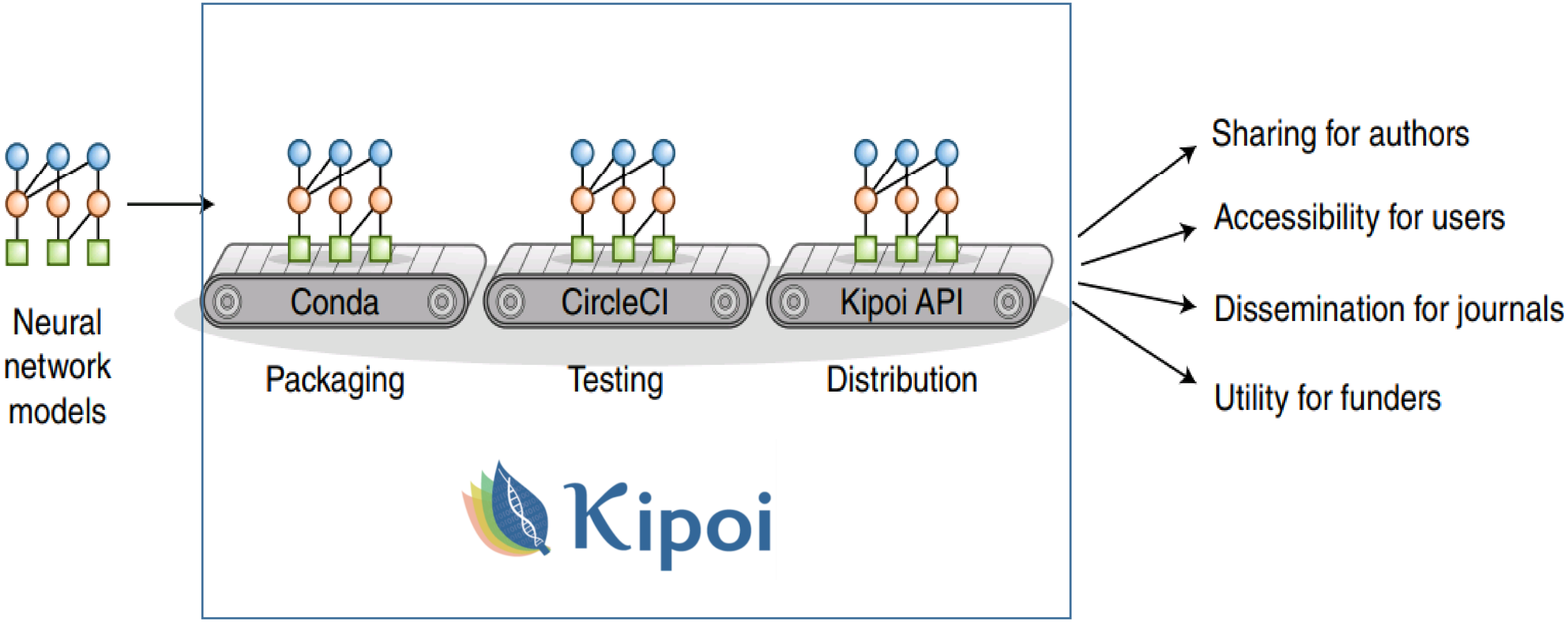
GitHub

Paper supplements

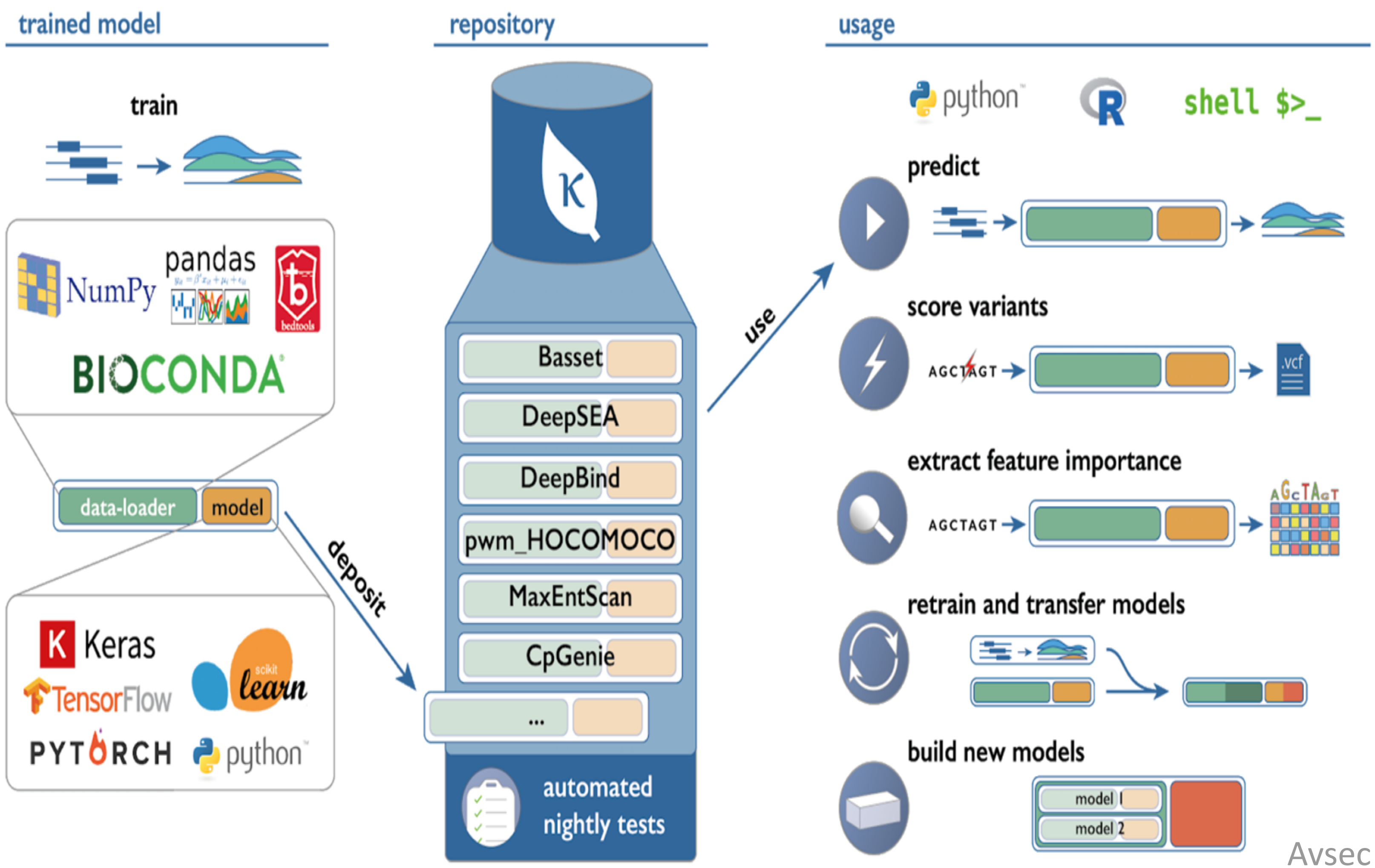
Methods
Supporting information
References

Author-maintained web page

The Kipoi Platform (Kipoi.org): Model Zoo for Genomics



Main Ingredients of Kipoi Platform



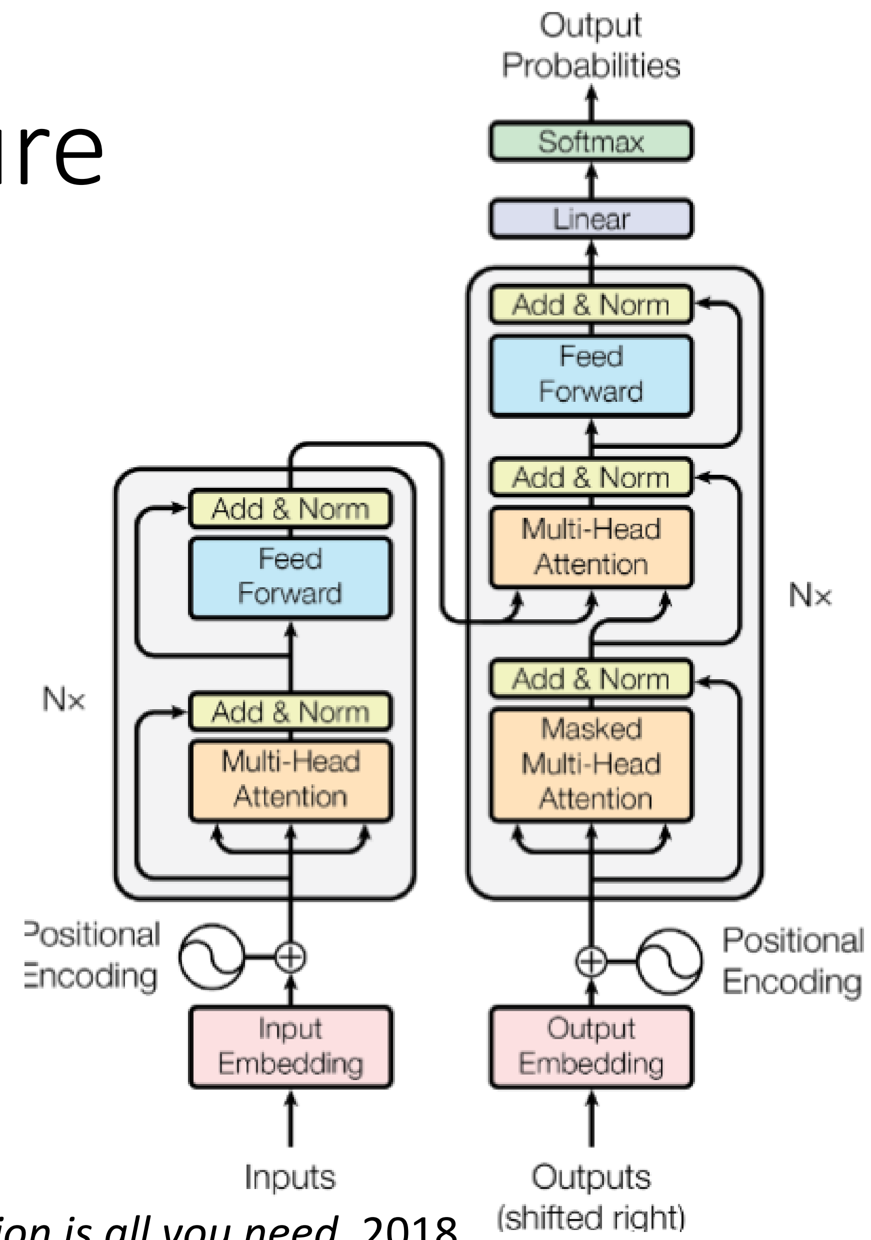
Avsec *et al* 2019

Foundation models (FM)

- A large-scale model trained on vast amounts of data that can be adapted (fine-tuned) for a wide variety of downstream tasks.
- Algorithms: mostly transformer based DL
- Business: one FM can enable multiple tasks by transfer leaning
- Computing: specialized hardware such as GPU
- Data: massive amount to data for pre-training

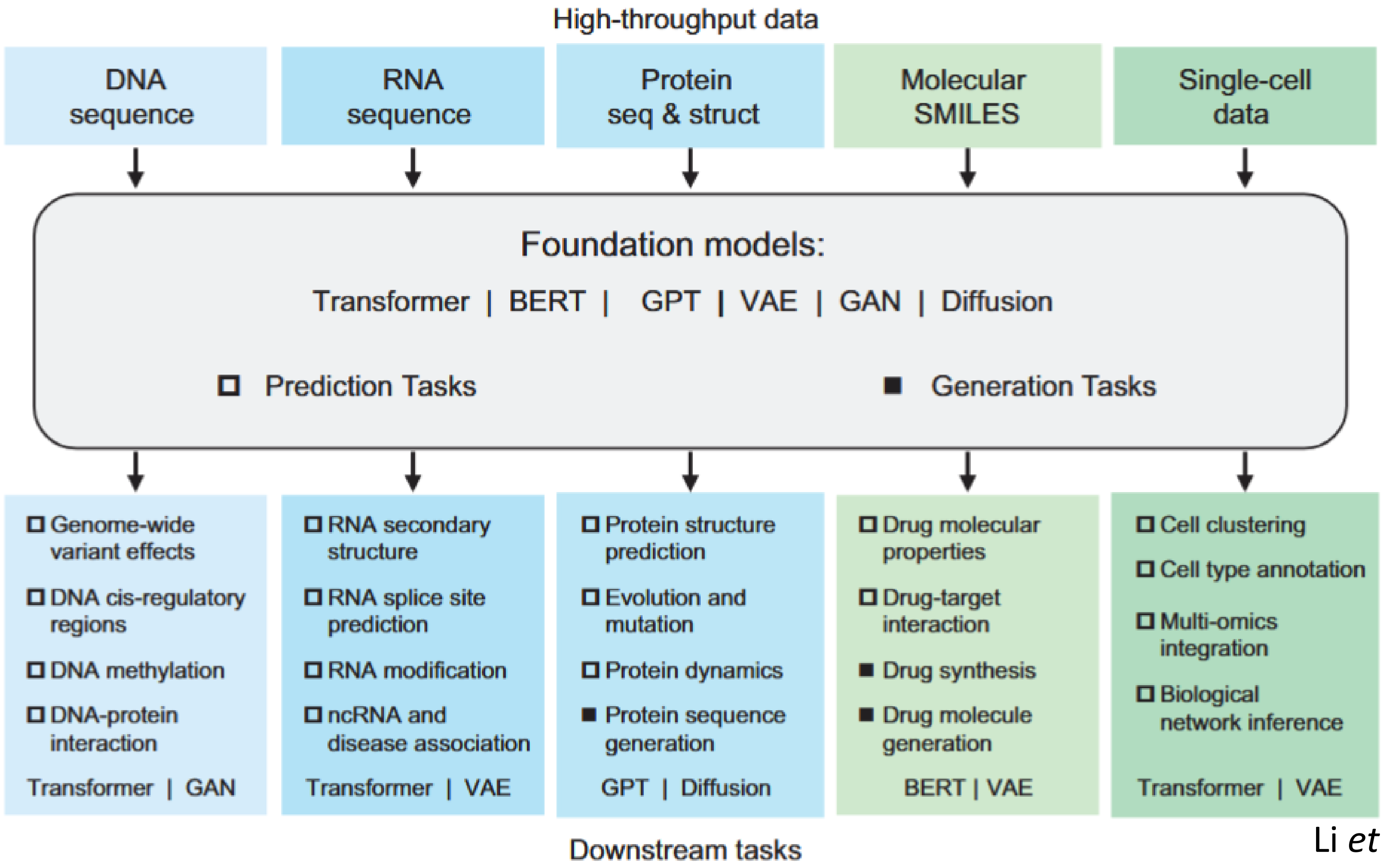
The Transformer Architecture

- Based on parallel attention mechanism
- Attention mechanism: a data adaptive component that dynamically focuses on the relevant information in the input to the compute the output
- LLMs such as chatGPT



Attention is all you need, 2018

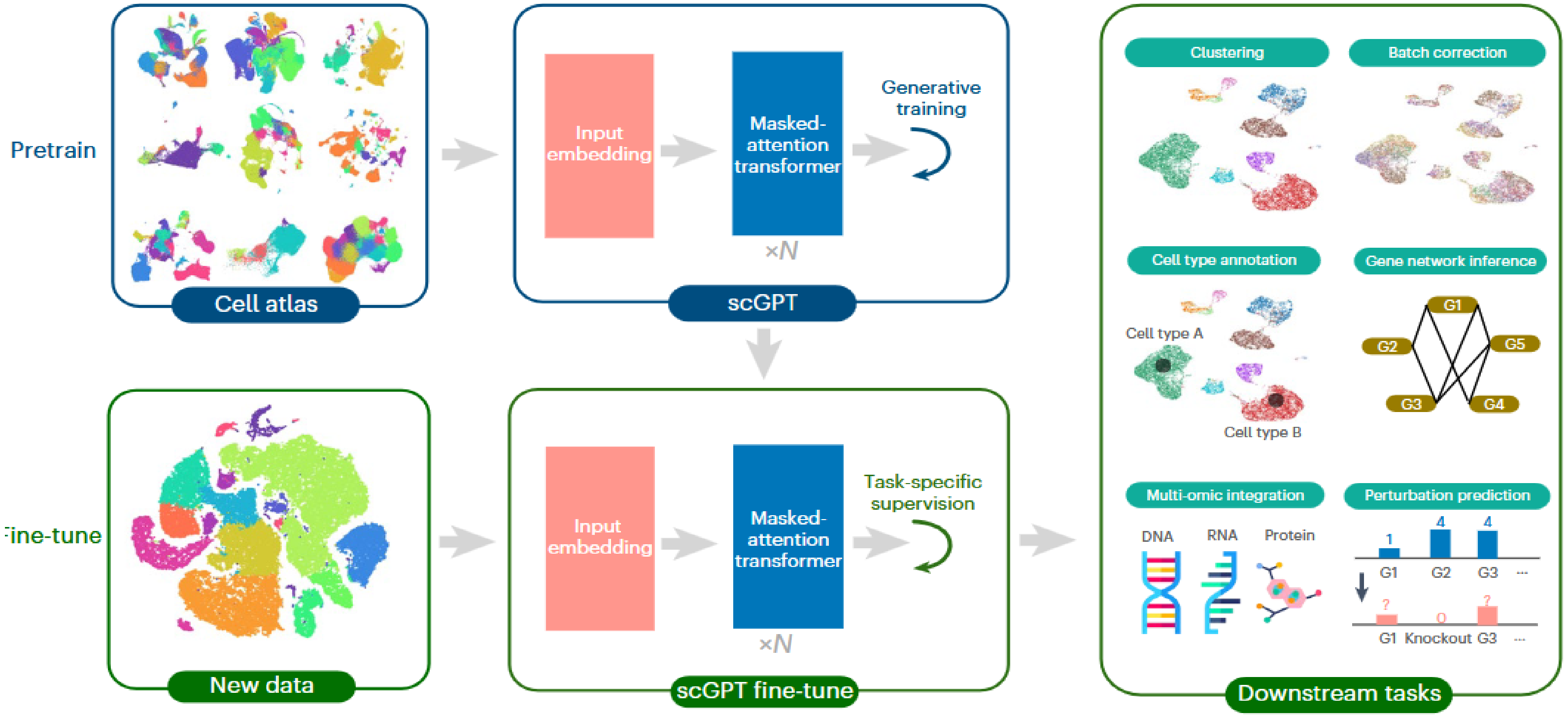
Foundation Models in Biology



Li *et al*, 2024

scGPT (Generative Pre-trained Transformer)

- Single-cell language models can be used to identify states, discover novel cell types, infer regulation networks and integrate multi-omics data



scGPT

- Pre-trained using 33M normal cells from HCA for two weeks (100M parameters).
- Pre training is compute intensive while fine tuning is light weight
- Input: 3 sets of tokens: gene ID, expression value, condition tag (technology, tissue type)
- Guilt by association: no causal inferences

Challenges and Limitations

- Limited amount of well curated data for data hungry models
- “Black box” problem: difficult to interpret
- Computational resources demands
- Require expert knowledge to design and fine tune structures
- Garbage in garbage out