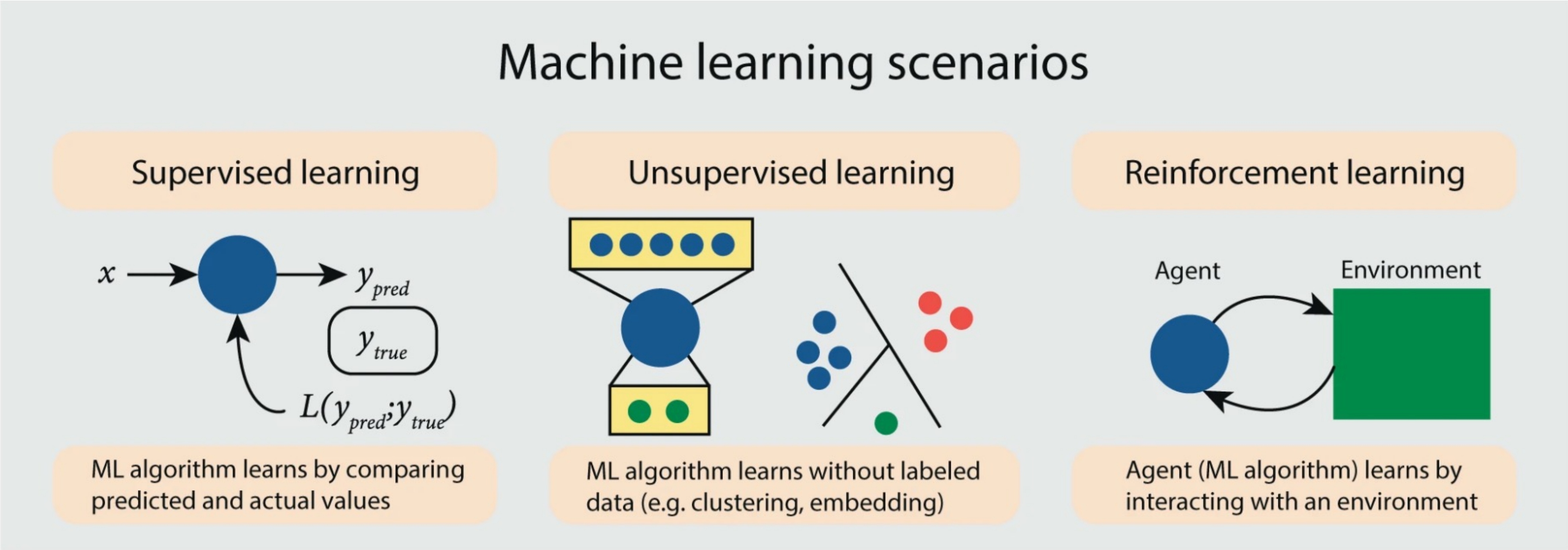
04/03/2023

Applications for Deep Learning in Computational Biology

Yuhang Chen
TF Lecture for CBB 752
Biomedical Data Science: Mining and Modeling
Spring, 2023
Yale University

An overview of common machine learning paradigms



Machine learning scenarios

Supervised learning

ML algorithm learns by comparing predicted and actual values

Unsupervised learning

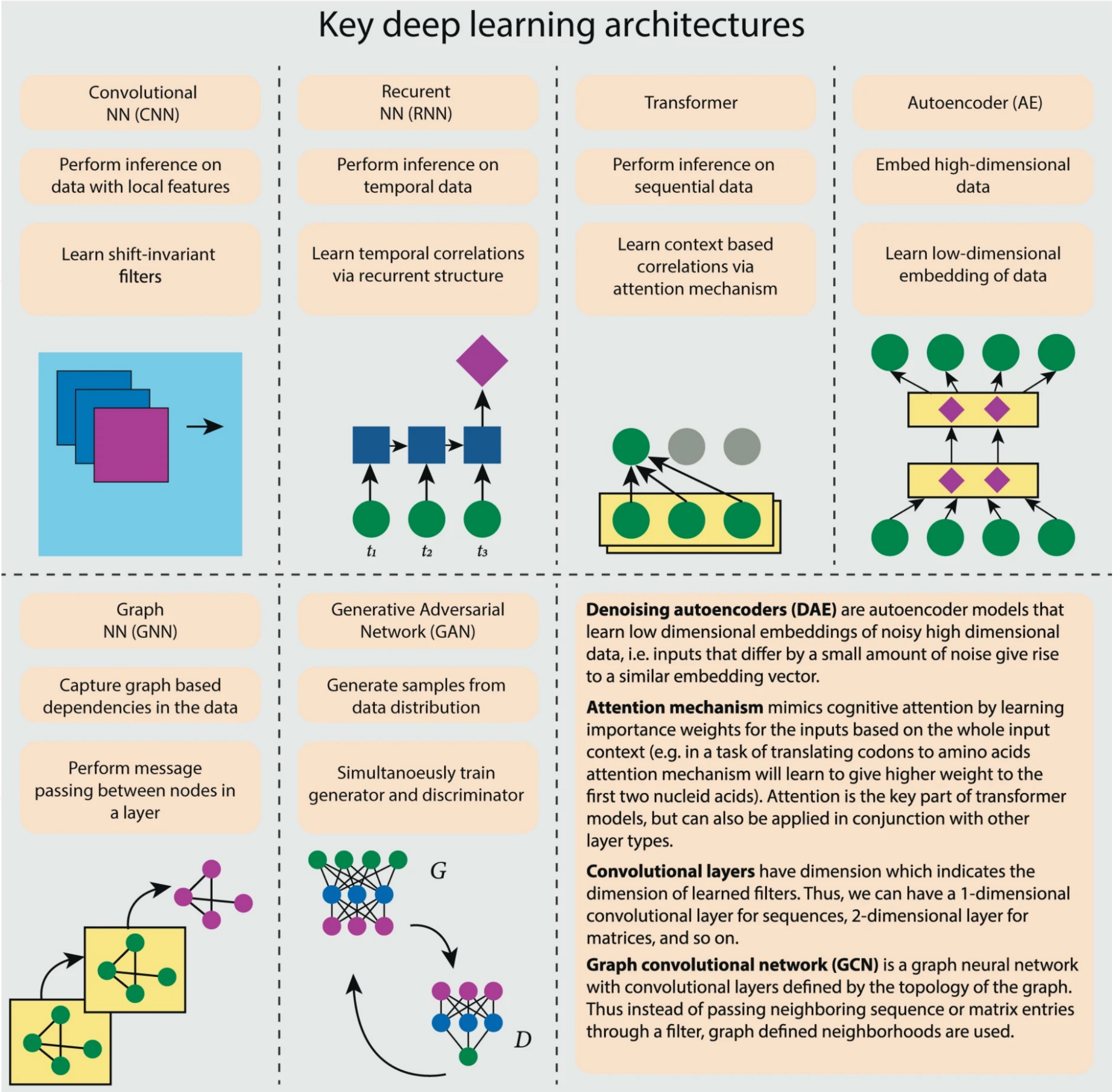
ML algorithm learns without labeled data (e.g. clustering, embedding)

Reinforcement learning

Agent (ML algorithm) learns by interacting with an environment

Sapoval, N., Aghazadeh, A., Nute, M.G. et al. Current progress and open challenges for applying deep learning across the biosciences. Nat Commun 13, 1728 (2022).

An overview of commonly used DL architecture



Key deep learning architectures

	Convolutional NN (CNN)	Recurent NN (RNN)	Transformer	Autoencoder (AE)
Goal	Perform inference on data with local features	Perform inference on temporal data	Perform inference on sequential data	Embed high-dimensional data
Key idea	Learn shift-invariant filters	Learn temporal correlations via recurrent structure	Learn context based correlations via attention mechanism	Learn low-dimensional embedding of data

	Graph NN (GNN)	Generative Adversarial Network (GAN)
Goal	Capture graph based dependencies in the data	Generate samples from data distribution
Key idea	Perform message passing between nodes in a layer	Simultaneously train generator and discriminator

Denoising autoencoders (DAE) are autoencoder models that learn low dimensional embeddings of noisy high dimensional data, i.e. inputs that differ by a small amount of noise give rise to a similar embedding vector.

Attention mechanism mimics cognitive attention by learning importance weights for the inputs based on the whole input context (e.g. in a task of translating codons to amino acids attention mechanism will learn to give higher weight to the first two nucleid acids). Attention is the key part of transformer models, but can also be applied in conjunction with other layer types.

Convolutional layers have dimension which indicates the dimension of learned filters. Thus, we can have a 1-dimensional convolutional layer for sequences, 2-dimensional layer for matrices, and so on.

Graph convolutional network (GCN) is a graph neural network with convolutional layers defined by the topology of the graph. Thus instead of passing neighboring sequence or matrix entries through a filter, graph defined neighborhoods are used.

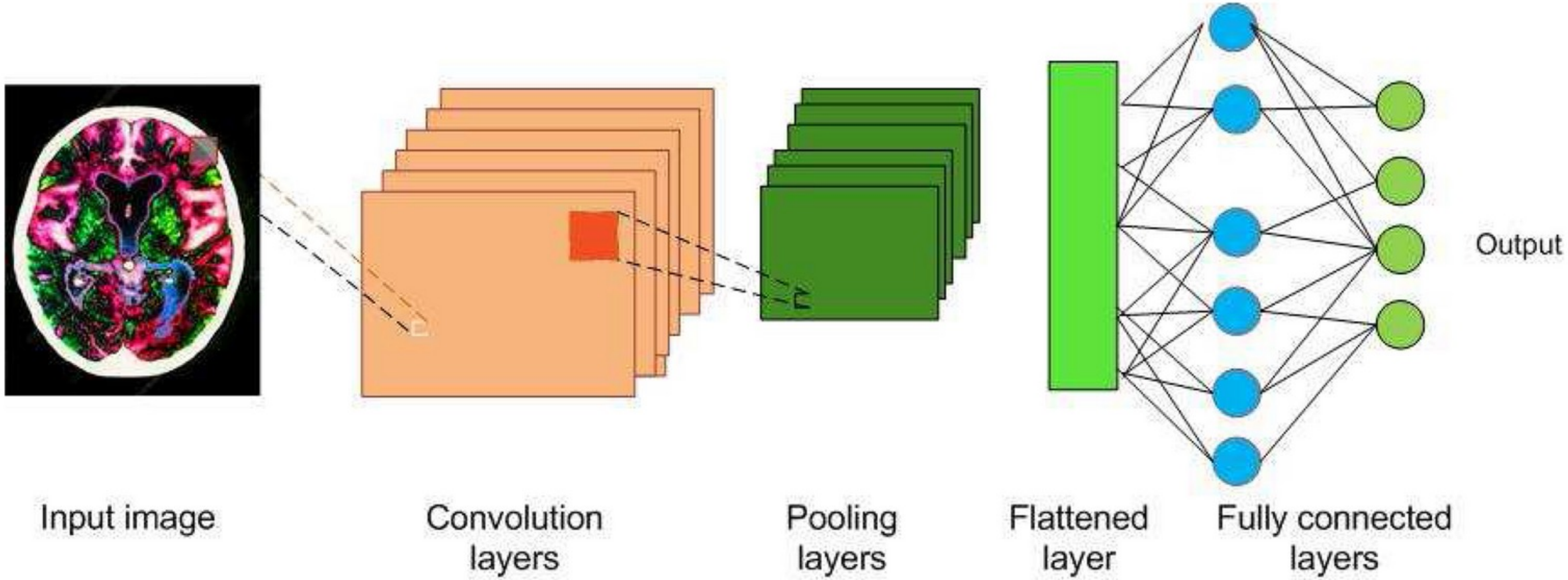
Sapoval, N., Aghazadeh, A., Nute, M.G. et al. Current progress and open challenges for applying deep learning across the biosciences. Nat Commun 13, 1728 (2022).

Follow up with Martin’s lecture structure:

- Deep Supervised Learning: Deep CNN/RNN for image classification/sequence classification
- Deep Unsupervised Learning: Deep Autoencoder, Deep Generative Models
- Deep Reinforcement Learning: AlphaGO, AlphaZero
- Explainable AI (XAI) in biology

Deep Supervised Learning

Building blocks of a CNN

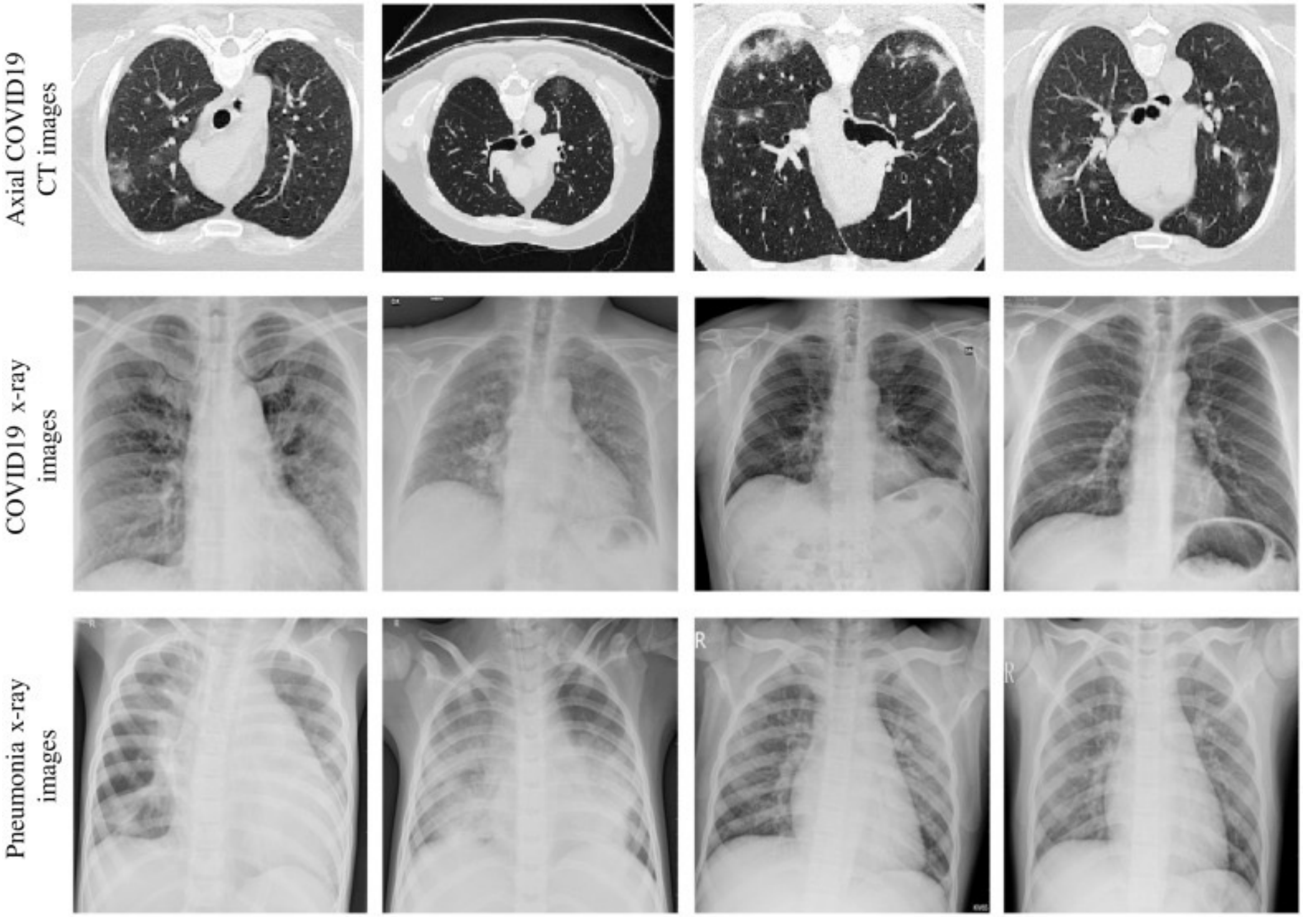


CNN is mainly used for applications in image and speech recognition.

What makes CNNs so effective is their ability to learn a sequence of filters to extract more and more complex patterns. In particular, these convolutional filters are characterized by their compact support, and by the property of being translation-invariant.

Sarvamangala, D.R., Kulkarni, R.V. Convolutional neural networks in medical image understanding: a survey. Evol. Intel. 15, 1–22 (2022).

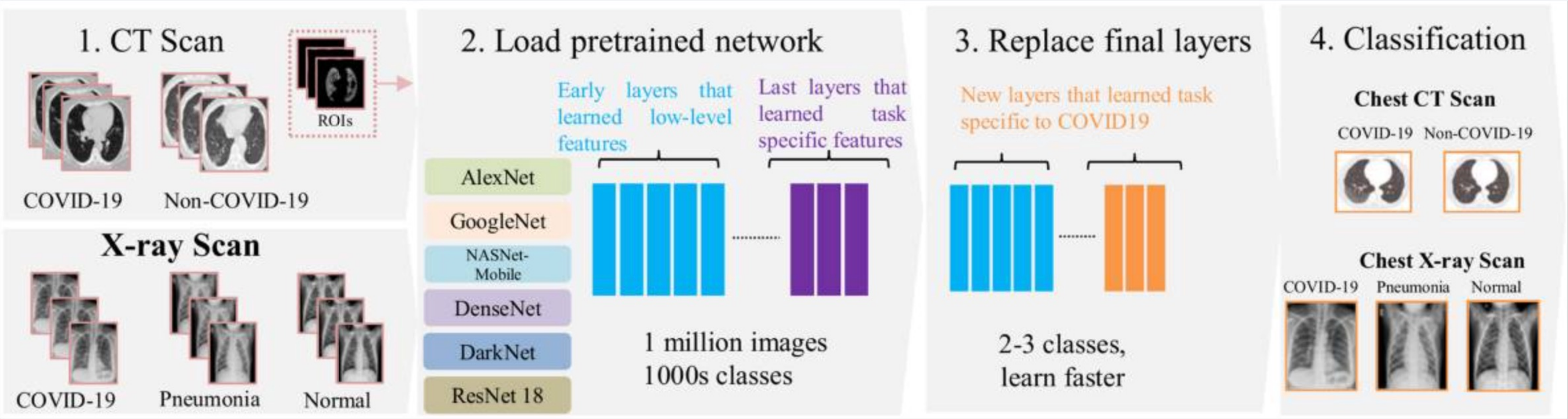
Example: Deep CNN models for predicting COVID-19 in CT and x-ray images



Examples of COVID-19 in CT and x-ray images. First row: axial COVID-19 CT images with lesions in different positions and sizes. Second row: COVID-19 x-ray images. Third row: pneumonia x-ray images.

Chaddad A, Hassan L, Desrosiers C. Deep CNN models for predicting COVID-19 in CT and x-ray images. J Med Imaging (Bellingham). 2021 Jan;8(Suppl 1):014502.

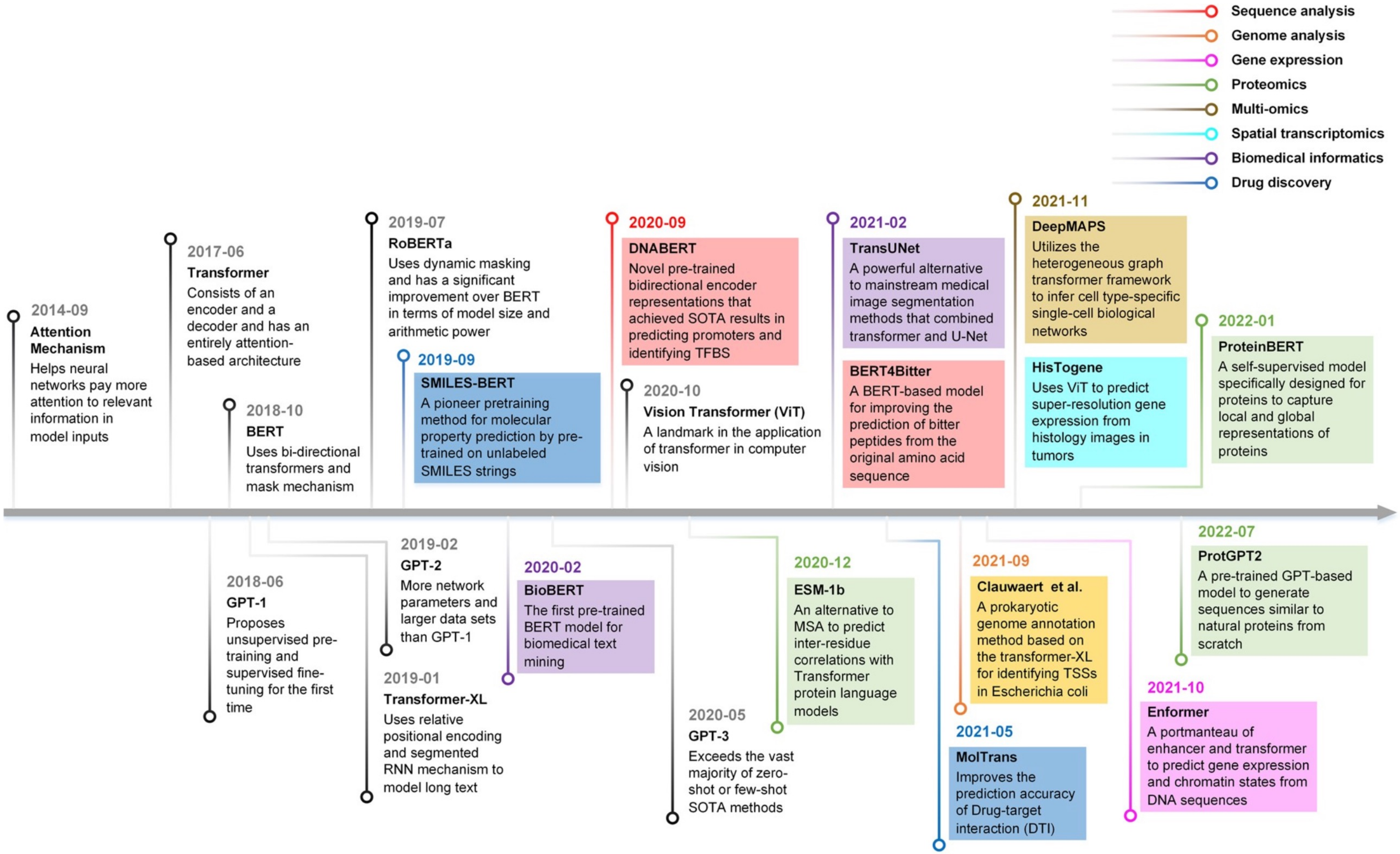
Example: Deep CNN models for predicting COVID-19 in CT and X-ray images



Regions of interest (ROI) corresponding to ground-glass opacities (GGO), consolidations, and pleural effusions were labeled in 100 axial lung CT images from 60 COVID-19-infected subjects. These segmented regions were then employed as an additional input to six deep convolutional neural network (CNN) architectures, pretrained on natural images, to differentiate between COVID-19 and normal CT images.
Also explored the model’s ability to classify x-ray images as COVID-19, non-COVID-19 pneumonia, or normal.

Chaddad A, Hassan L, Desrosiers C. Deep CNN models for predicting COVID-19 in CT and x-ray images. J Med Imaging (Bellingham). 2021 Jan;8(Suppl 1):014502.

An overview of important works related to TRANSFORMER in computational biology regime



Sequence analysis
Genome analysis
Gene expression
Proteomics
Multi-omics
Spatial transcriptomics
Biomedical informatics
Drug discovery

2014-09
Attention Mechanism
Helps neural networks pay more attention to relevant information in model inputs

2017-06
Transformer
Consists of an encoder and a decoder and has an entirely attention-based architecture

2018-06
GPT-1
Proposes unsupervised pre-training and supervised fine-tuning for the first time

2018-10
BERT
Uses bi-directional transformers and mask mechanism

2019-01
Transformer-XL
Uses relative positional encoding and segmented RNN mechanism to model long text

2019-02
GPT-2
More network parameters and larger data sets than GPT-1

2019-07
RoBERTa
Uses dynamic masking and has a significant improvement over BERT in terms of model size and arithmetic power

2019-09
SMILES-BERT
A pioneer pretraining method for molecular property prediction by pre-trained on unlabeled SMILES strings

2020-02
BioBERT
The first pre-trained BERT model for biomedical text mining

2020-05
GPT-3
Exceeds the vast majority of zero-shot or few-shot SOTA methods

2020-09
DNABERT
Novel pre-trained bidirectional encoder representations that achieved SOTA results in predicting promoters and identifying TFBS

2020-10
Vision Transformer (ViT)
A landmark in the application of transformer in computer vision

2020-12
ESM-1b
An alternative to MSA to predict inter-residue correlations with Transformer protein language models

2021-02
TransUNet
A powerful alternative to mainstream medical image segmentation methods that combined transformer and U-Net

BERT4Bitter
A BERT-based model for improving the prediction of bitter peptides from the original amino acid sequence

2021-05
MolTrans
Improves the prediction accuracy of Drug-target interaction (DTI)

2021-09
Clauwaert et al.
A prokaryotic genome annotation method based on the transformer-XL for identifying TSSs in Escherichia coli

2021-10
Enformer
A portmanteau of enhancer and transformer to predict gene expression and chromatin states from DNA sequences

2021-11
DeepMAPS
Utilizes the heterogeneous graph transformer framework to infer cell type-specific single-cell biological networks

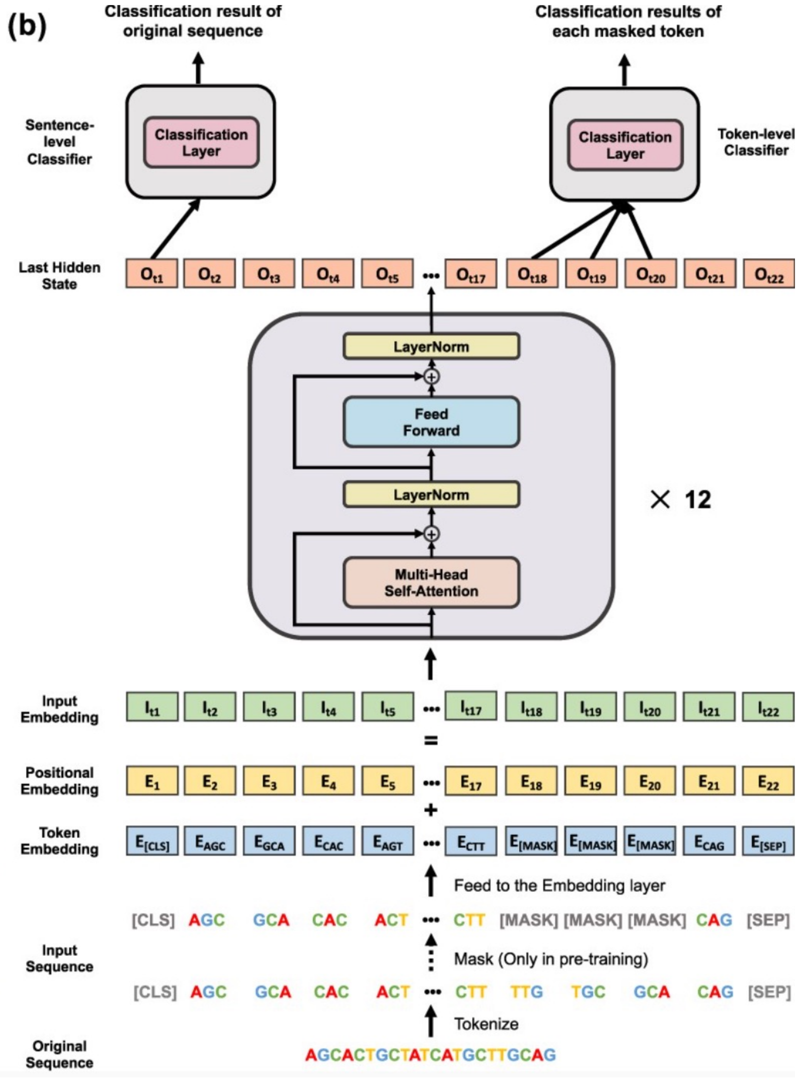
HisTogene
Uses ViT to predict super-resolution gene expression from histology images in tumors

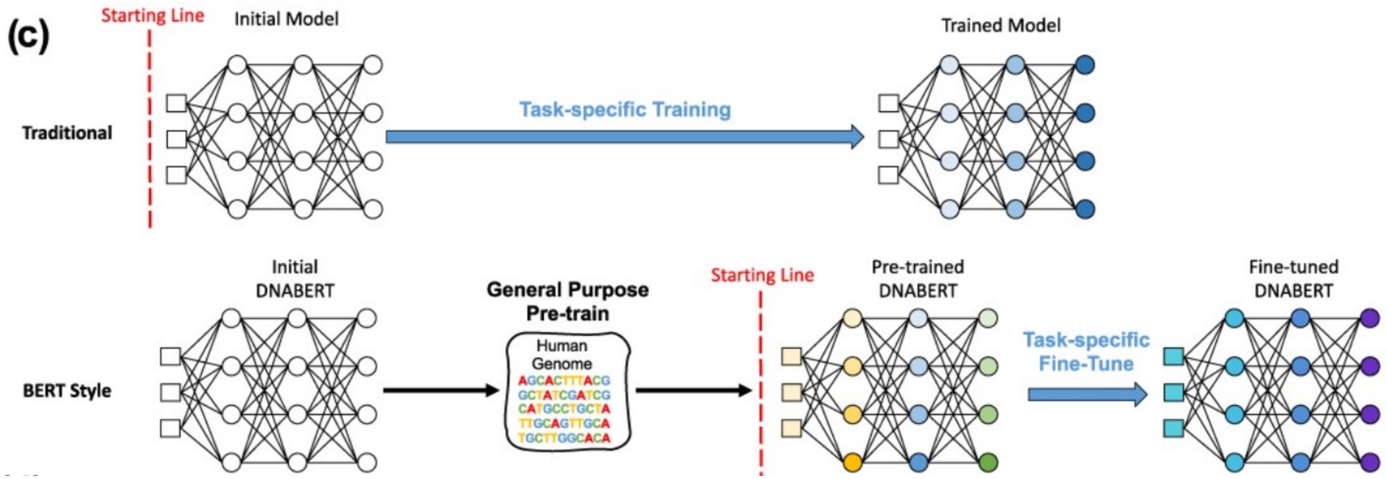
2022-01
ProteinBERT
A self-supervised model specifically designed for proteins to capture local and global representations of proteins

2022-07
ProtGPT2
A pre-trained GPT-based model to generate sequences similar to natural proteins from scratch

Shuang Zhang, Rui Fan, Yuti Liu, Shuang Chen, Qiao Liu, Wanwen Zeng, Applications of transformer-based language models in bioinformatics: a survey, Bioinformatics Advances, Volume 3, Issue 1, 2023

Example: DNABERT – Transformer model for predicting promoters and identifying TFBSs


(b)
Classification result of original sequence
Classification results of each masked token
Sentence-level Classifier
Classification Layer
Classification Layer
Token-level Classifier
Last Hidden State
O_t1 O_t2 O_t3 O_t4 O_t5 ••• O_t17 O_t18 O_t19 O_t20 O_t21 O_t22
LayerNorm
Feed Forward
LayerNorm
Multi-Head Self-Attention
× 12
Input Embedding
I_t1 I_t2 I_t3 I_t4 I_t5 ••• I_t17 I_t18 I_t19 I_t20 I_t21 I_t22
=
Positional Embedding
E_1 E_2 E_3 E_4 E_5 ••• E_17 E_18 E_19 E_20 E_21 E_22
+
Token Embedding
E_[CLS] E_AGC E_GCA E_CAC E_AGT ••• E_CTT E_[MASK] E_[MASK] E_[MASK] E_CAG E_[SEP]
Feed to the Embedding layer
[CLS] AGC GCA CAC ACT ••• CTT [MASK] [MASK] [MASK] CAG [SEP]
Mask (Only in pre-training)
Input Sequence
[CLS] AGC GCA CAC ACT ••• CTT TTG TGC GCA CAG [SEP]
Tokenize
Original Sequence
AGCACTGCTATCATGCTTGCAG

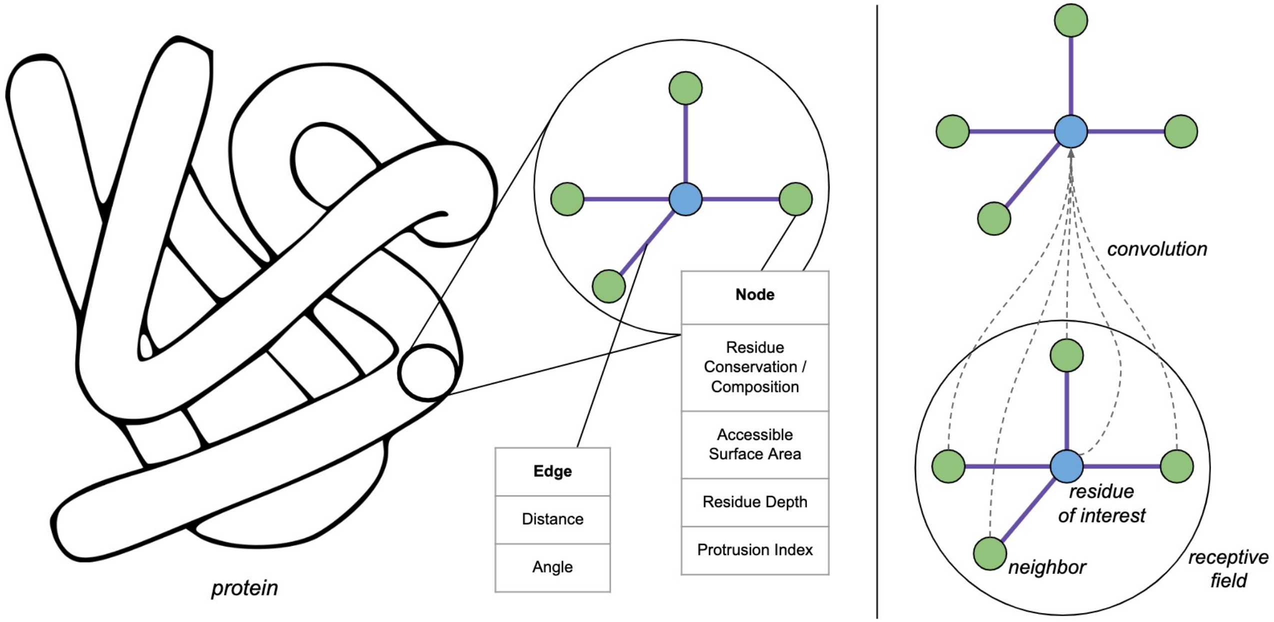


- Left: DNABERT uses tokenized k-mer sequences as input, which also contains a CLS token (a tag representing meaning of entire sentence), a SEP token (sentence separator) and MASK tokens (to represent masked k-mers in pre-training). The input passes an embedding layer and is fed to 12 Transformer blocks.

- Top: DNABERT adopts general-purpose pre-training which can then be fine-tuned for multiple purposes using various task-specific data.

Yanrong Ji, Zhihan Zhou, Han Liu, Ramana V Davuluri, DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome, Bioinformatics

Example: Protein Interface Prediction using Graph Convolutional Networks



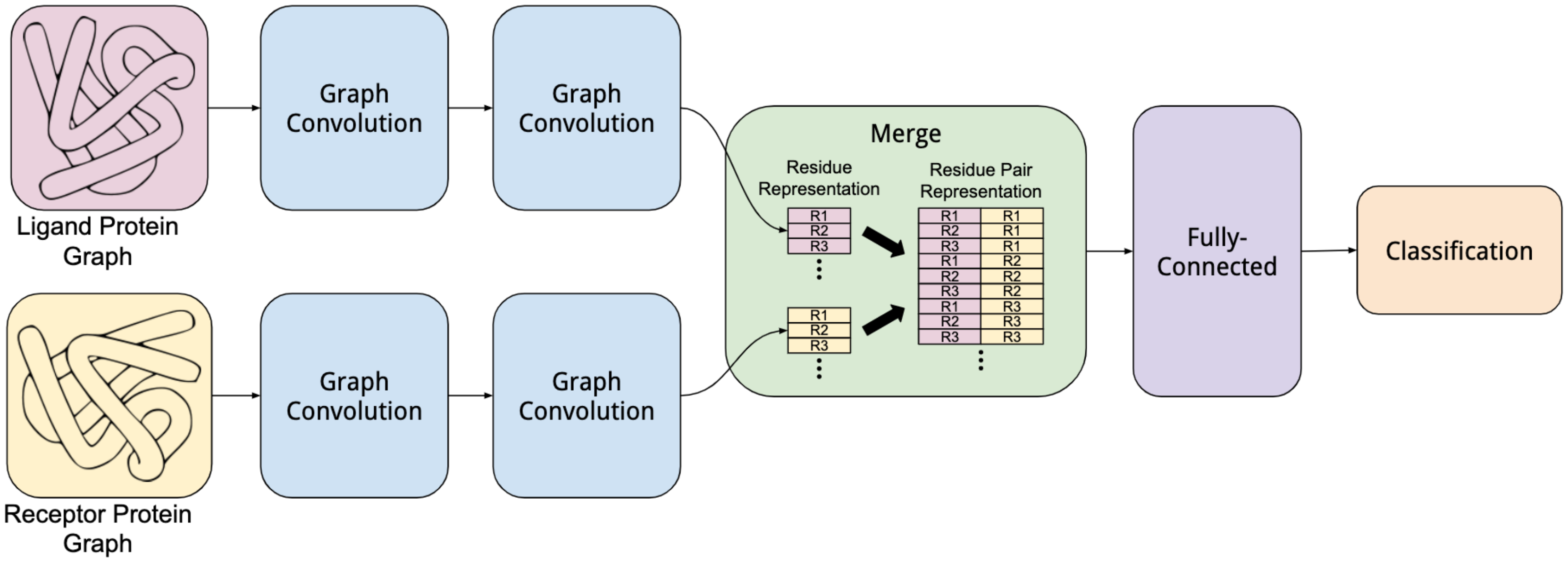
Graph convolution on protein structures.

Left: Each residue in a protein is a node in a graph where the neighborhood of a node is the set of neighboring nodes in the protein structure; each node has features computed from its amino acid sequence and structure, and edges have features describing the relative distance and angle between residues.

Right: Schematic description of the convolution operator which has as its receptive field a set of neighboring residues, and produces an activation which is associated with the center residue

Alex Fout, Jonathon Byrd, Basir Shariat, and Asa Ben-Hur. 2017. Protein interface prediction using graph convolutional networks. NIPS'17.

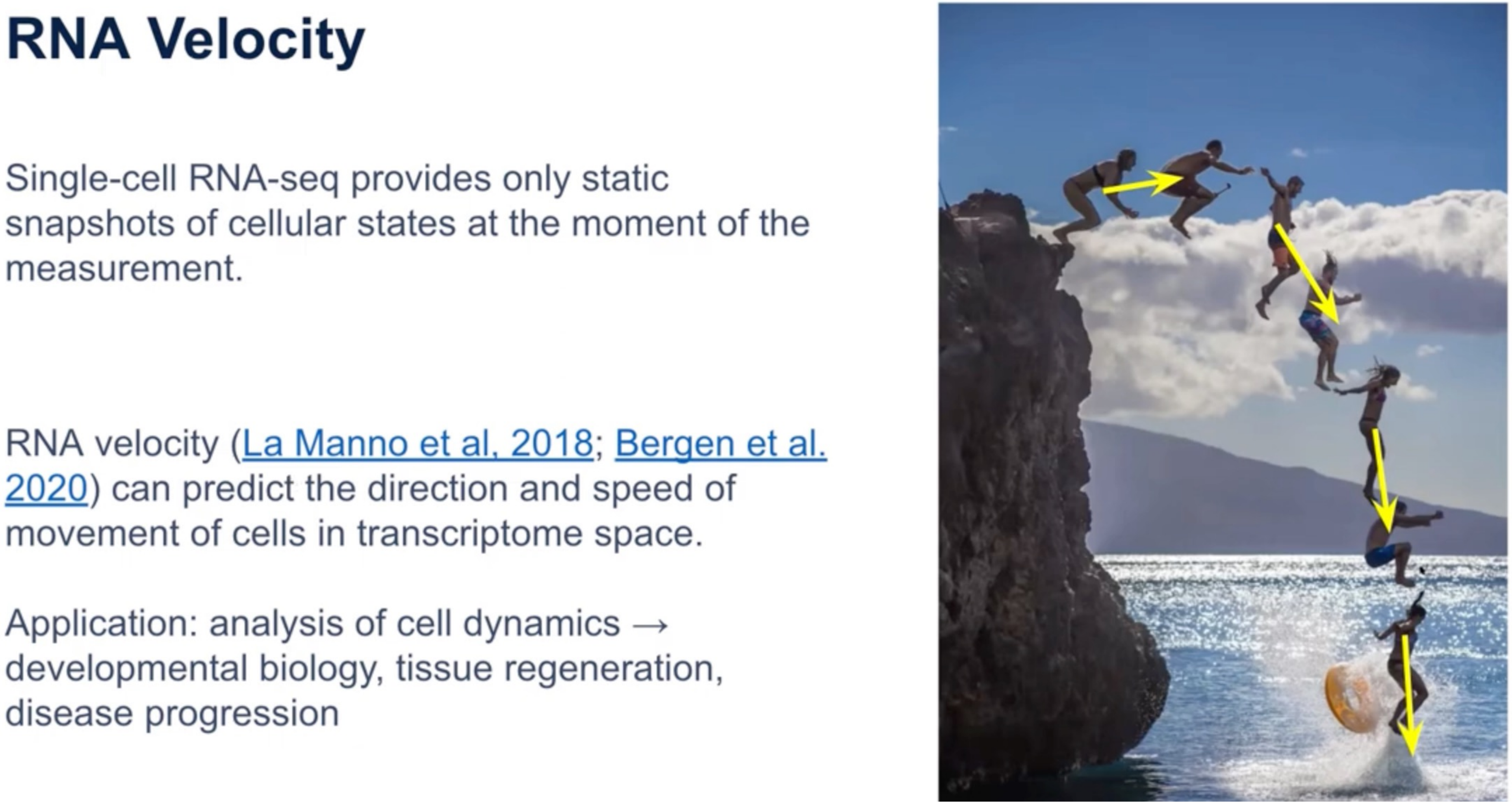
Example: Protein Interface Prediction using Graph Convolutional Networks



An overview of the pairwise classification architecture.
Each neighborhood of a residue in the two proteins is processed using one or more graph convolution layers, with weight sharing between legs of the network. The activations generated by the convolutional layers are merged by concatenating them, followed by one or more regular dense layers.

Alex Fout, Jonathon Byrd, Basir Shariat, and Asa Ben-Hur. 2017. Protein interface prediction using graph convolutional networks. NIPS'17.

Deep Unsupervised Learning



RNA Velocity

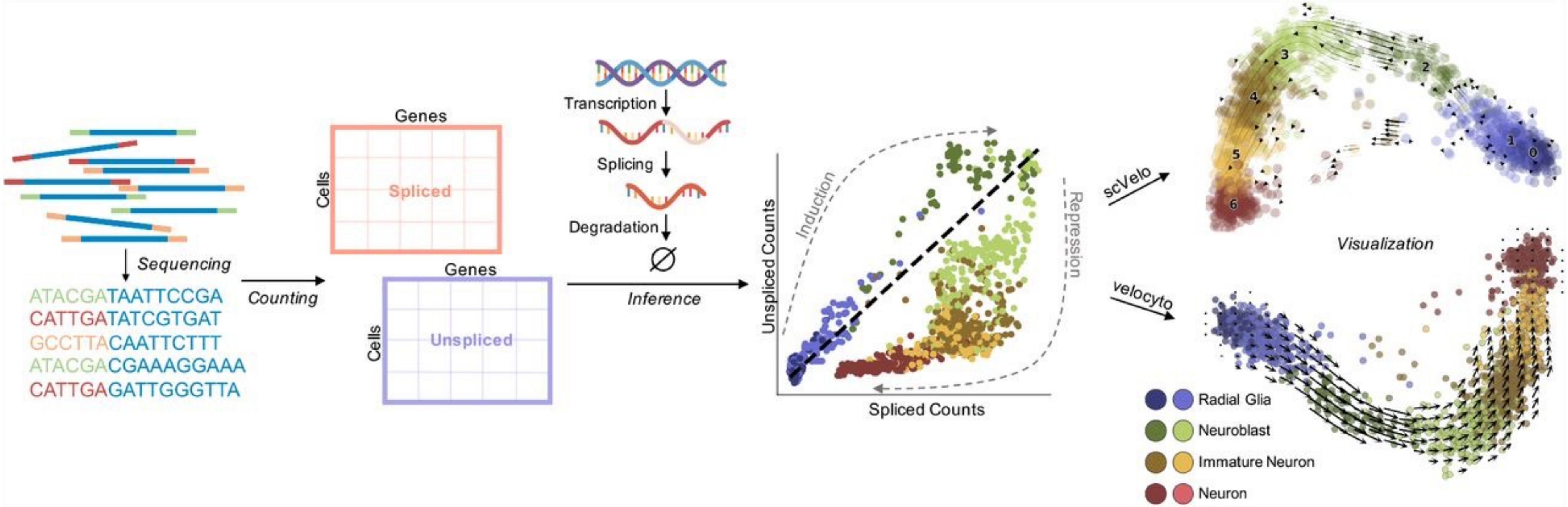
Single-cell RNA-seq provides only static snapshots of cellular states at the moment of the measurement.

RNA velocity (La Manno et al, 2018; Bergen et al. 2020) can predict the direction and speed of movement of cells in transcriptome space.

Application: analysis of cell dynamics → developmental biology, tissue regeneration, disease progression

https://www.youtube.com/watch?v=ODEP3JhyZq4

RNA velocity workflow



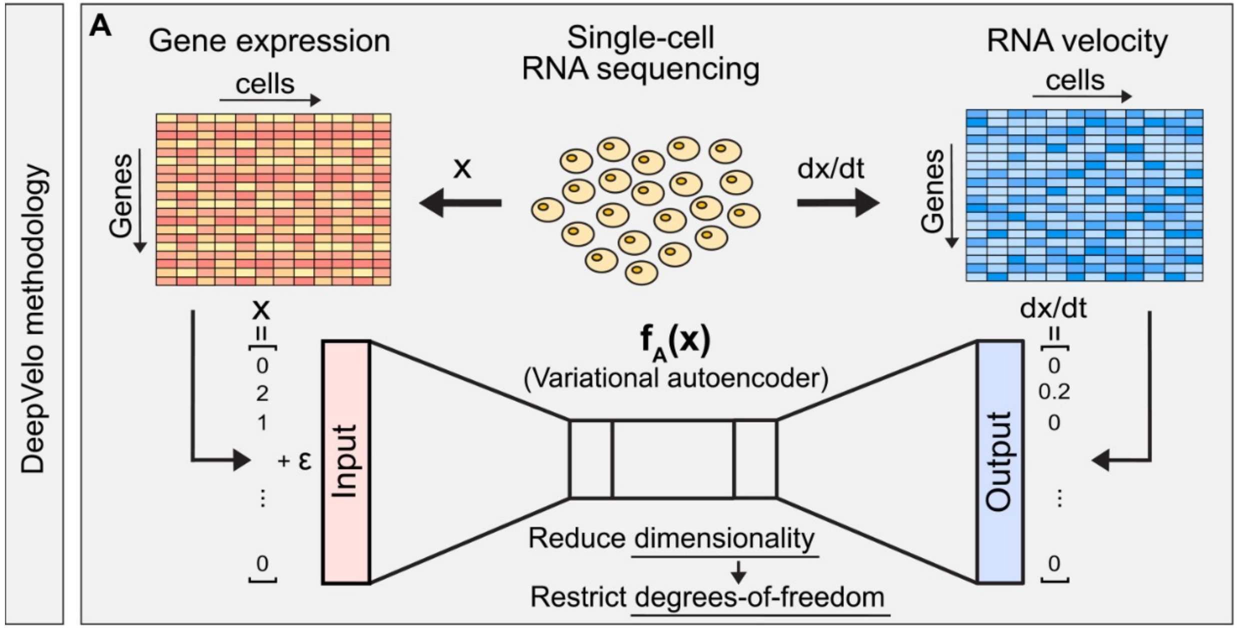
- Based on the relative abundance of mature (spliced) RNA and unspliced RNA to estimate the rate of RNA splicing and degradation

- Use the 2 count matrices to infer the directionality of transcription events within cells

- Phase plots describing the dynamical transcription process -> convert into embeddings showing with top 2 PCs

https://www.youtube.com/watch?v=ODEP3JhyZq4

Example: DeepVelo – Model Single-cell transcriptomic velocity using VAE



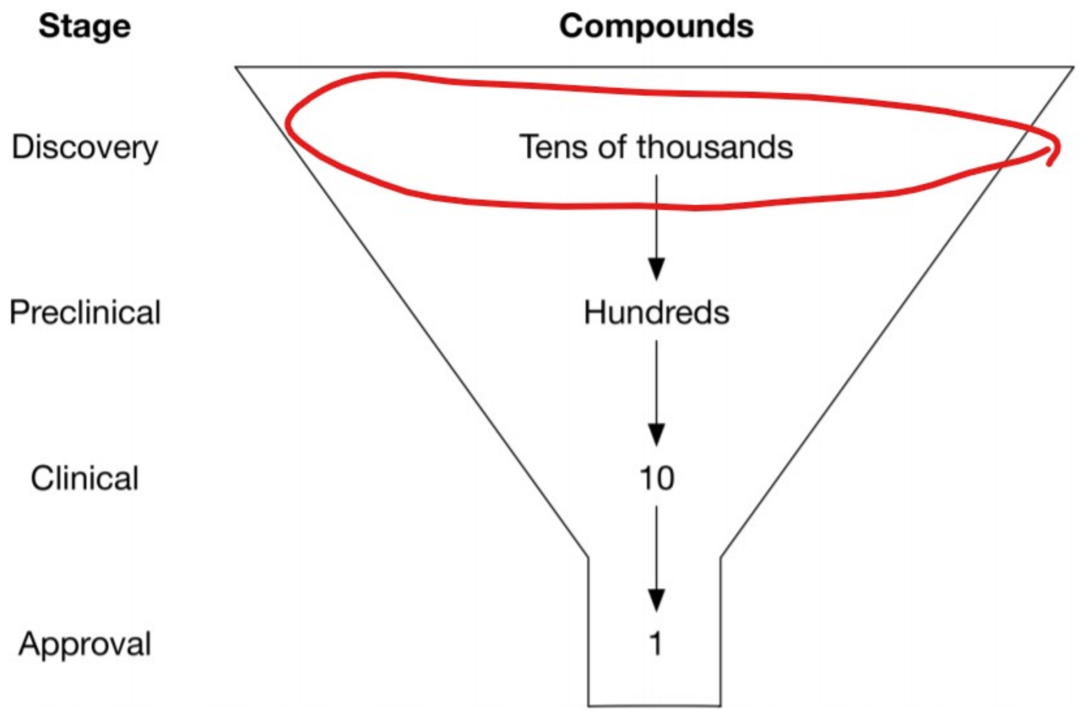
- Gene expression profile of an individual cell (x)
- RNA velocity $\partial x/\partial t$
- Existing methods that assume linear gene interactions (i.e., $\partial x/\partial t = Ax$ with matrix A)
- Train a VAE f_A to capture the nonlinear gene regulatory relationships (e.g., multiple TFs coactivating gene transcription) and map gene expression state to the RNA velocity, expressed by $\partial x/\partial t = \mathbf{f_A}(x)$

Chen Z et. al. DeepVelo: Single-cell transcriptomic deep velocity field learning with neural ordinary differential equations. Sci Adv. 2022

GANs for Biological Research

- GANs is particularly useful for establishing potential directions in scientific study: we can generate molecules or try out potential protein structures using GANs.

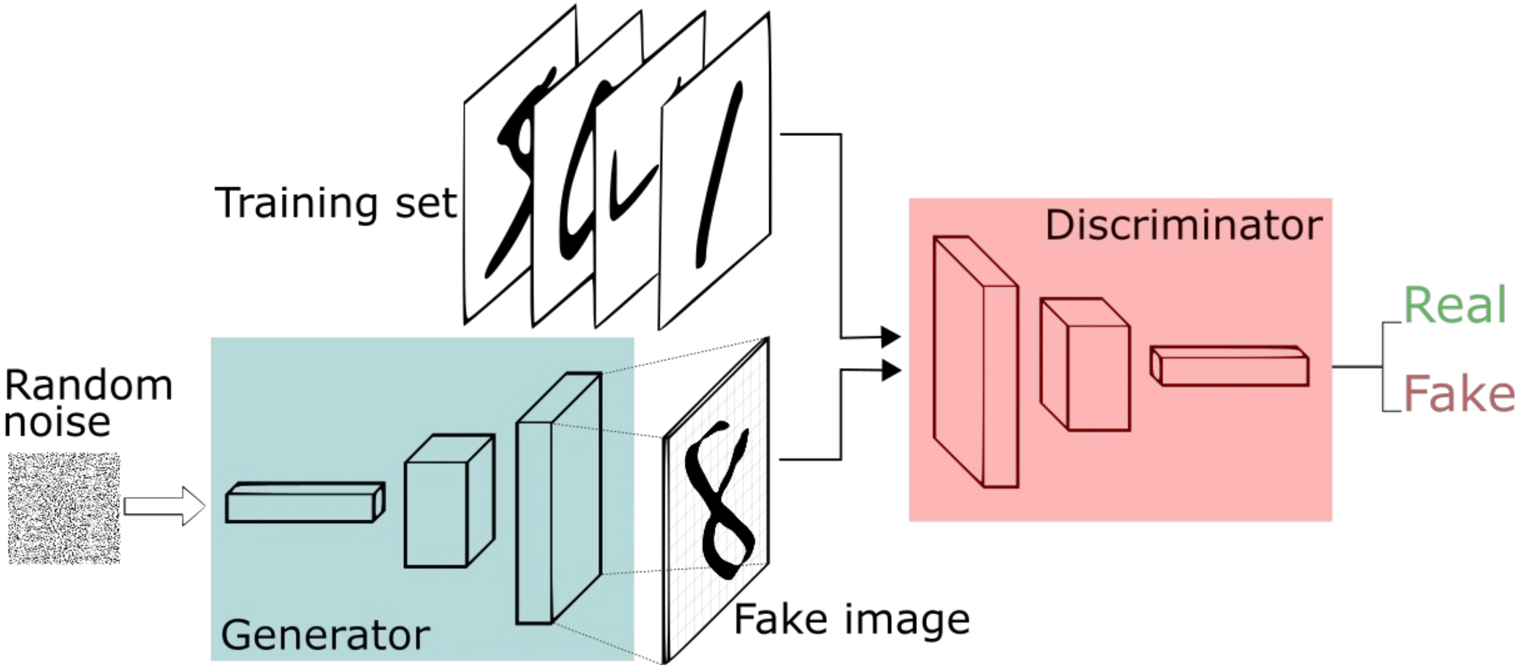
- Molecules that GANs output are rarely stable or potentially useful, but we can subsequently use other deep learning models to screen the few promising molecules in a dataset.



The red circle represents the phase of drug discovery GANs will impact

https://towardsdatascience.com/gans-for-driving-biological-research-d1c2d678036c

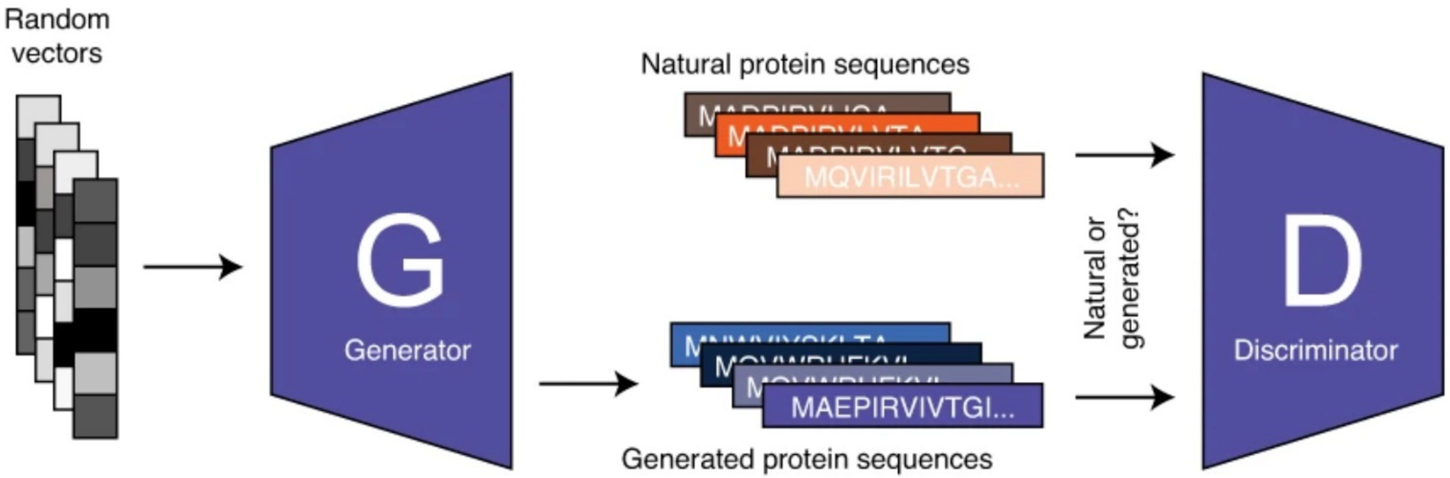
GANs architecture



The 'generator' produces a specific type of data (e.g., an image, text, or a protein sequence). The 'discriminator" tries to distinguish between the artificial data created by the 'generator' and authentic or real data.

Subsequently, the generator uses the feedback provided by the discriminator to generate new data. The generator never processes or analyzes real data and the data it produces. Therefore, its learning relies solely on the outcome of the analyses carried out by the discriminator.

Example: ProteinGAN - A generative adversarial network that generates functional protein sequences



- Given a random input vector, the Generator network produces a protein sequence, which is scored by the Discriminator network by comparing it to the natural protein sequences. The generator tries to fool the discriminator by generating sequences that will eventually look like real ones (the generator never actually sees real enzyme sequences).

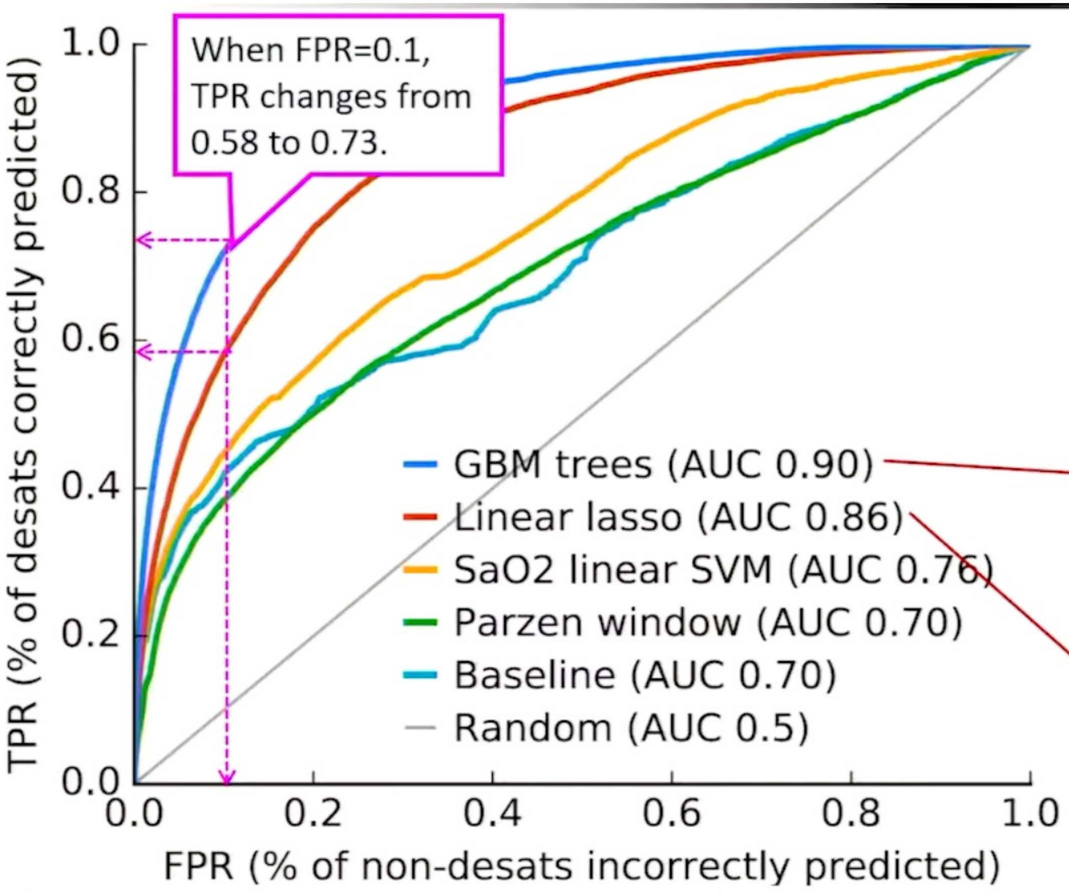
- ProteinGAN learns the evolutionary relationships of protein sequences directly from the amino-acid sequence space and creates new, highly diverse sequence variants with natural-like physical properties.

- 24% those new proteins are experimentally tested to be functional in vitro

Repecka, D., Jauniskis, V., Karpus, L. et al. Expanding functional protein sequence spaces using generative adversarial networks. Nat Mach Intell 3, 324–333 (2021).

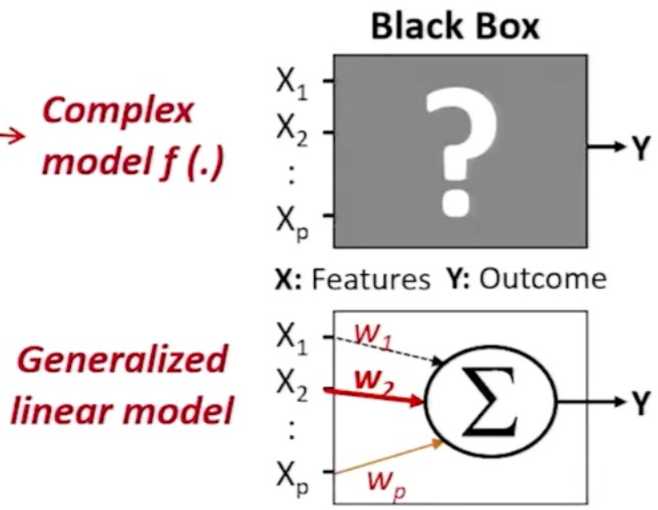
Explainable AI (XAI) in biology

A tradeoff between accuracy vs. interpretability



When FPR=0.1, TPR changes from 0.58 to 0.73.

- Receiver operating characteristic (ROC) curve from 10-fold cross validation tests.
  - Simple vs. complex models



Some dimensions to evaluating explainability of a model

- **Comprehensibility:** The extent to which extracted representations are humanly comprehensible, and thus touching on the dimensions of transparency considered earlier.

- **Fidelity:** The extent to which extracted representations accurately capture the opaque models from which they were extracted.

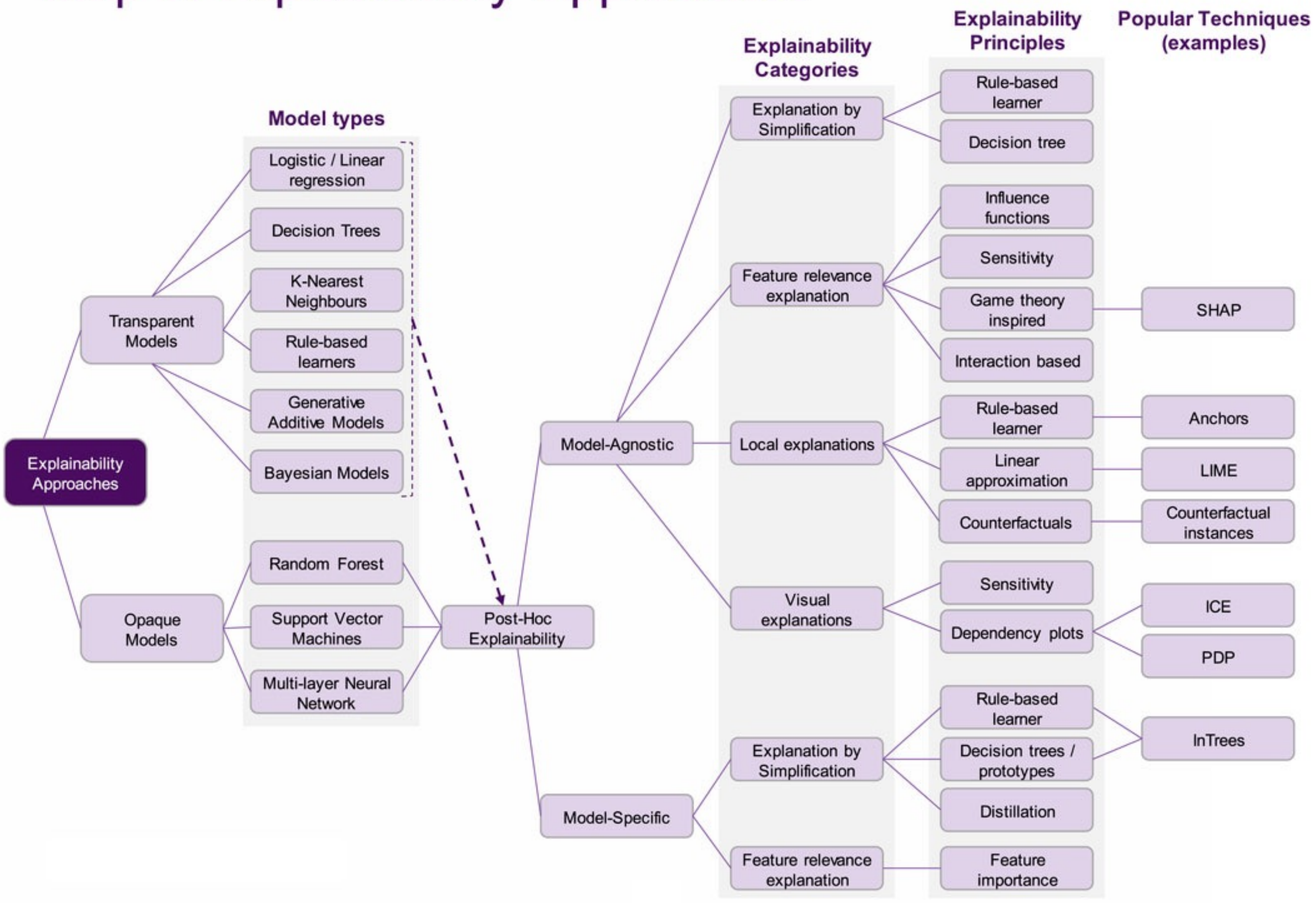
- **Accuracy:** The ability of extracted representations to accurately predict unseen examples.

- **Scalability:** The ability of the method to scale to opaque models with large input spaces and large numbers of weighted connections.

- **Generality:** The extent to which the method requires special training regimes or restrictions on opaque models.

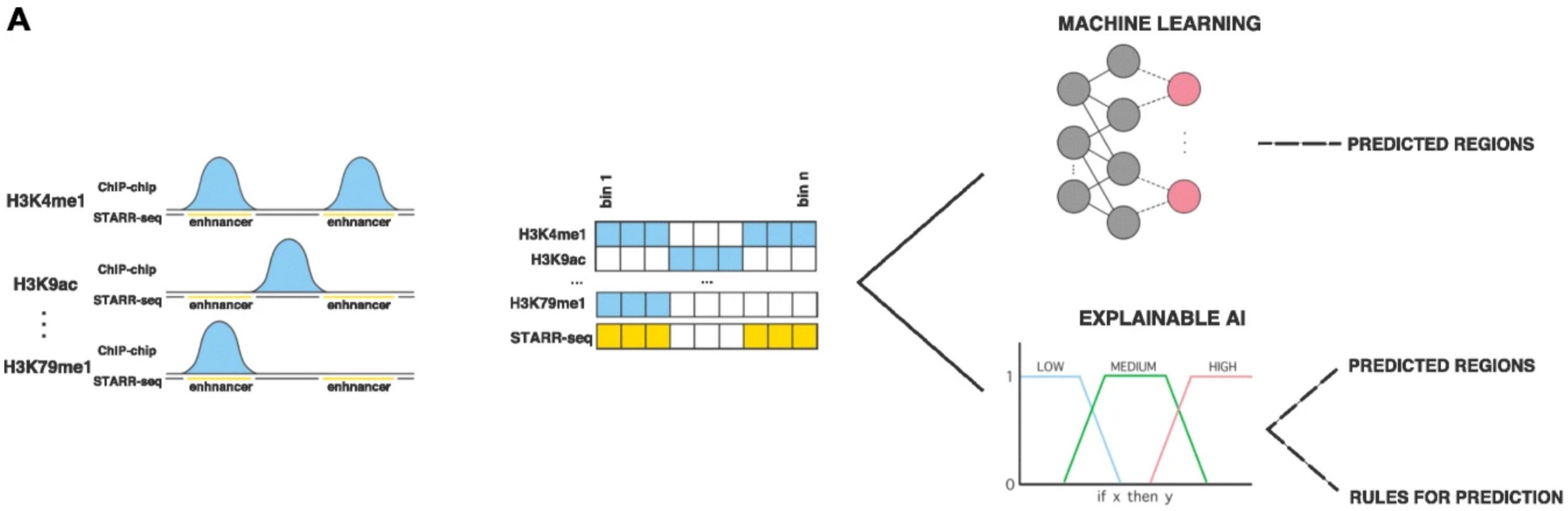
Belle Vaishak, Papantonis Ioannis. Principles and Practice of Explainable Machine Learning. Frontiers in Big Data. 2021

Map of Explainability Approaches



Belle Vaishak, Papantonis Ioannis. Principles and Practice of Explainable Machine Learning. Frontiers in Big Data. 2021

Example: An explainable artificial intelligence approach for decoding the enhancer histone modifications code and identification of novel enhancers in Drosophila



ChIP-seq data for histone modifications and STARR-seq enhancer annotations are combined and tiled into bins covering the Drosophila genome. Using these bins, traditional machine learning models (ML) and explainable AI models (XAI) can be trained to predict enhancer locations.

Wolfe, J.C., Mikheeva, L.A., Hagras, H. et al. An explainable artificial intelligence approach for decoding the enhancer histone modifications code and identification of novel enhancers in Drosophila. Genome Biol 22, 308 (2021).

Example: An explainable artificial intelligence approach for decoding the enhancer histone modifications code and identification of novel enhancers in Drosophila

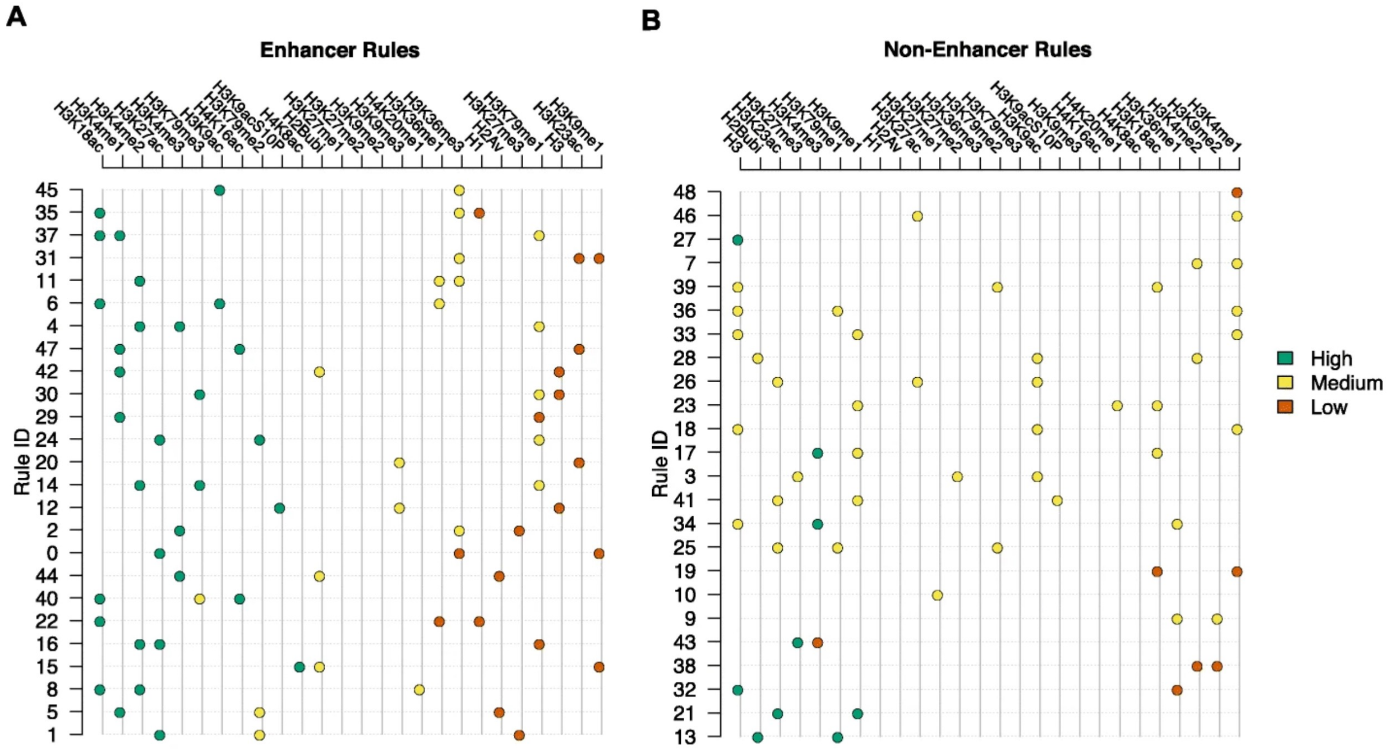


Illustration of rules identified by explainable AI model to classify regions as either enhancers or non-enhancers in Drosophila. The rules were determined to be the most effective while remaining explainable when constrained to a maximum of three epigenetic modifications per rule, and a maximum of 50 rules. These parameters were chosen to ensure that the model was explainable while maintaining a high degree of predictive power.

Wolfe, J.C., Mikheeva, L.A., Hagras, H. et al. An explainable artificial intelligence approach for decoding the enhancer histone modifications code and identification of novel enhancers in Drosophila. Genome Biol 22, 308 (2021).

Thanks for attention!

Q&A