CS 6824:
AI-powered Molecular Modeling

Website: https://people.cs.vt.edu/dbhattacharya/courses/cs6824/

Piazza: https://piazza.com/vt/spring2022/cs6824/home

Canvas: https://canvas.vt.edu/courses/145337

Are you in the right place?

- **This is CS 6824: CRN 20577**
  - Modality is "Face-to-Face Instruction"

Today

- **What is AI-powered Molecular Modeling, the field, about?**
  - Why should we care?
- **What is this class about?**
  - What to expect?
  - Logistics

What are we here to discuss?

Cutting-edge advances
made in
Molecular Modeling
using
the **power** of
AI

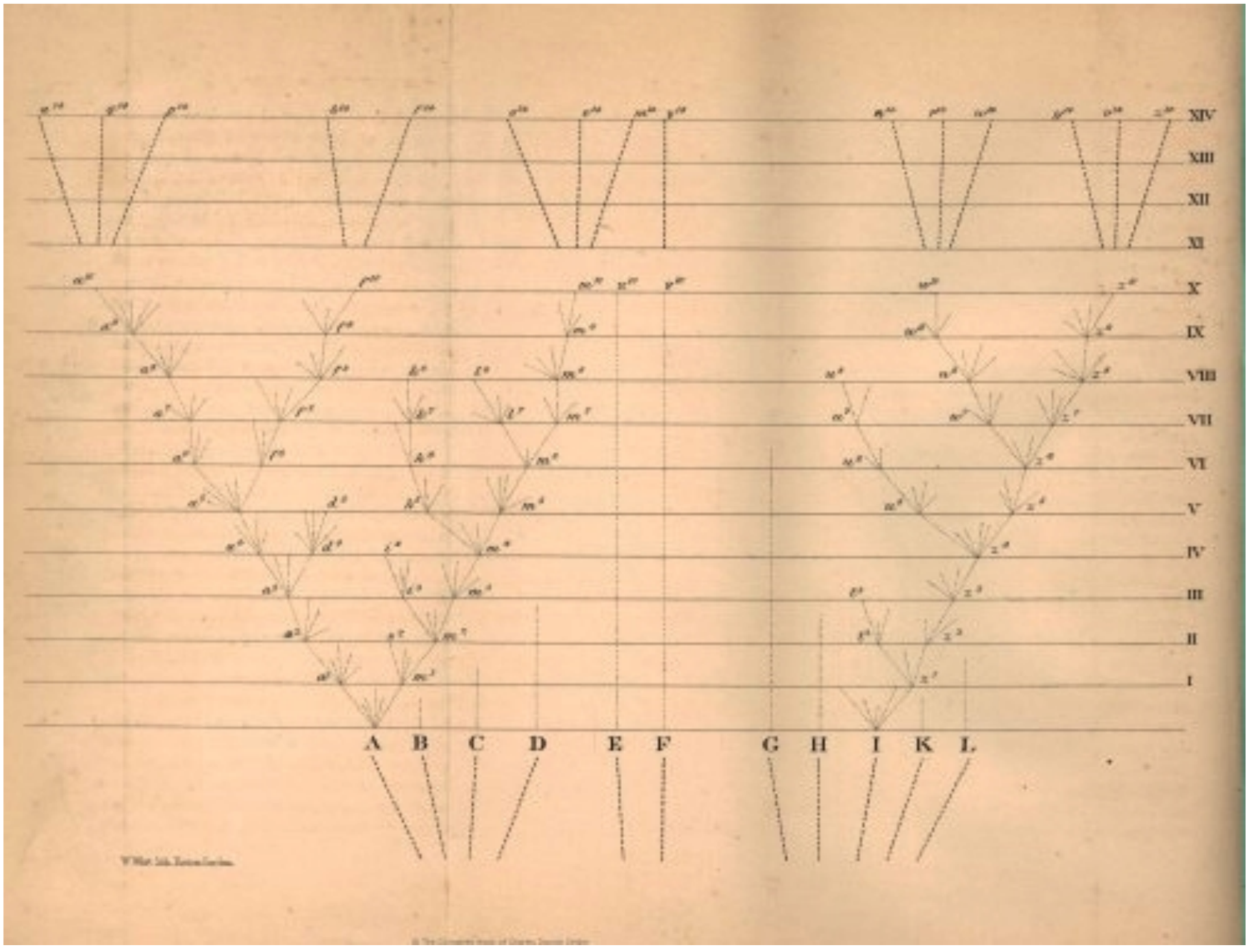
Demo time…

https://www.youtube.com/watch?v=iUMpm3tYsVE

More details:

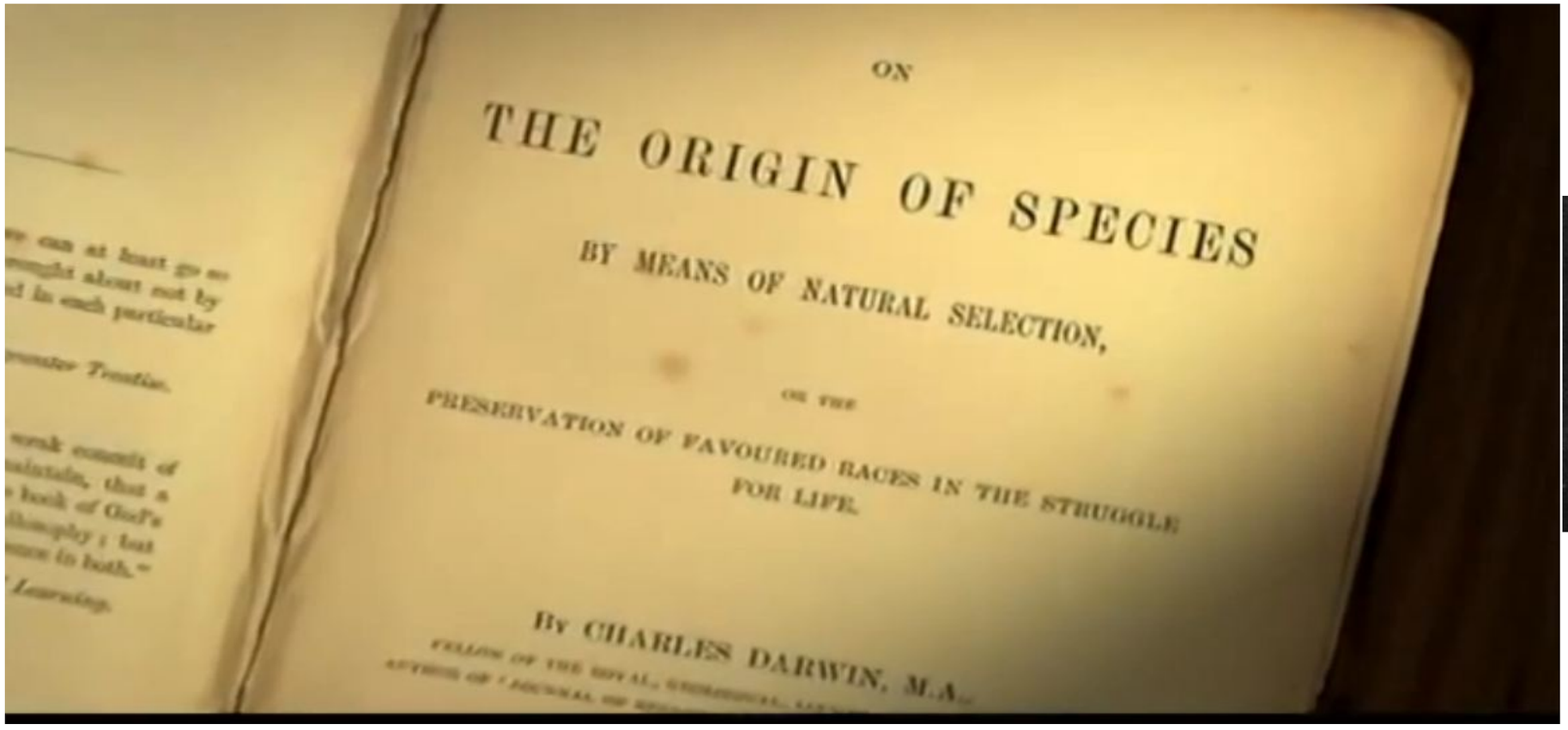
https://www.science.org/content/article/breakthrough-2021

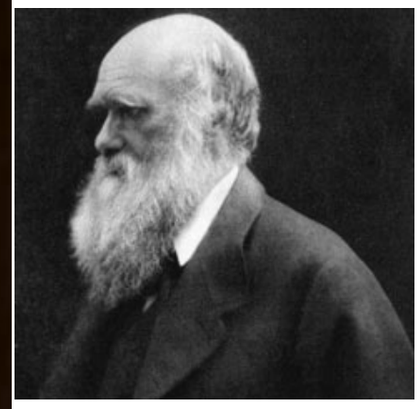
Modeling life…the inception?



"Principle of Divergence"

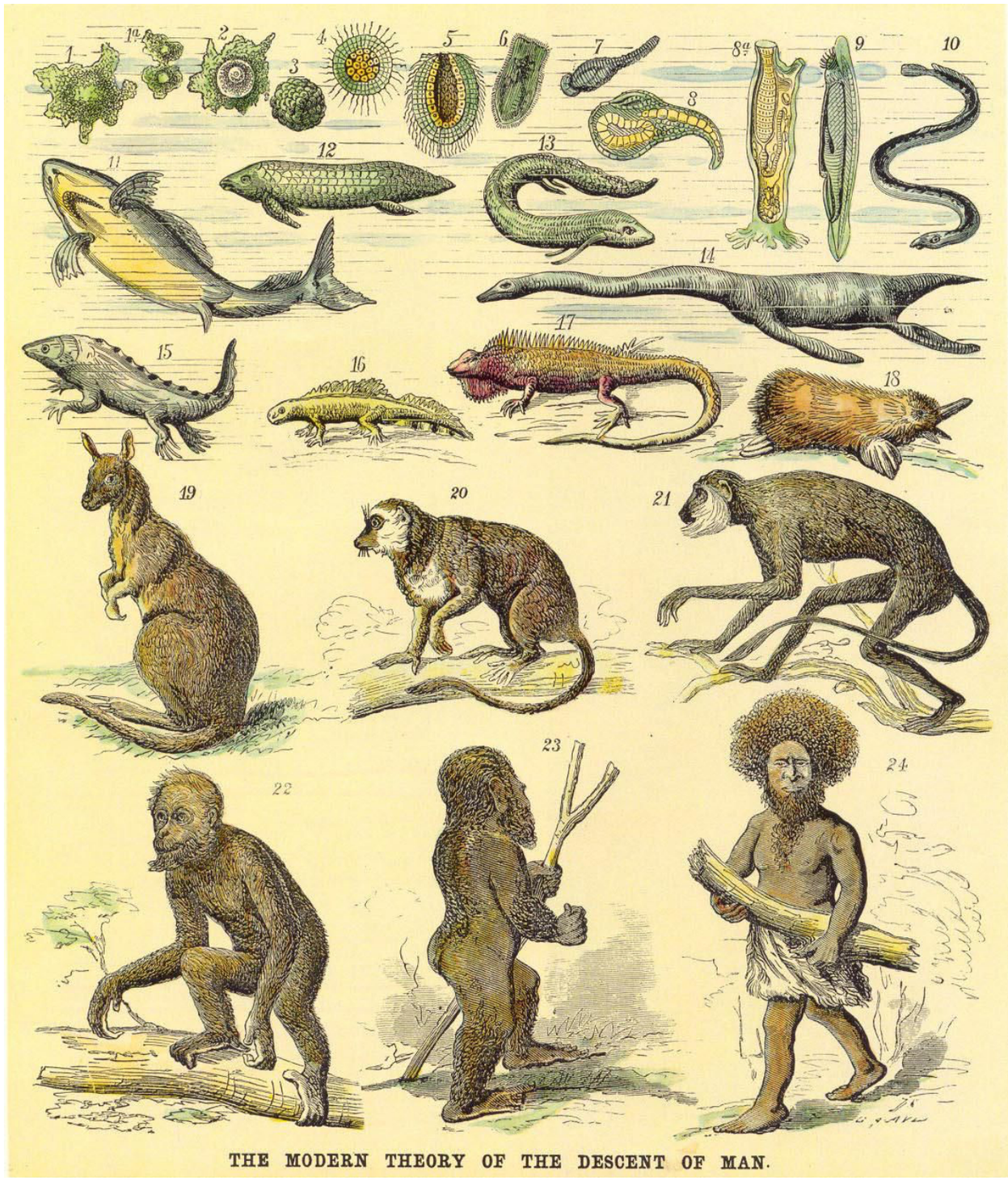
Modeling life…the inception?



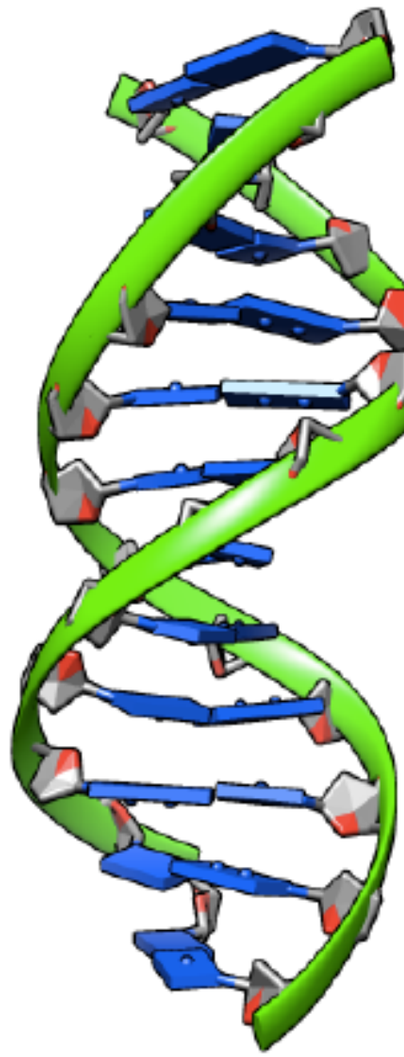


1859

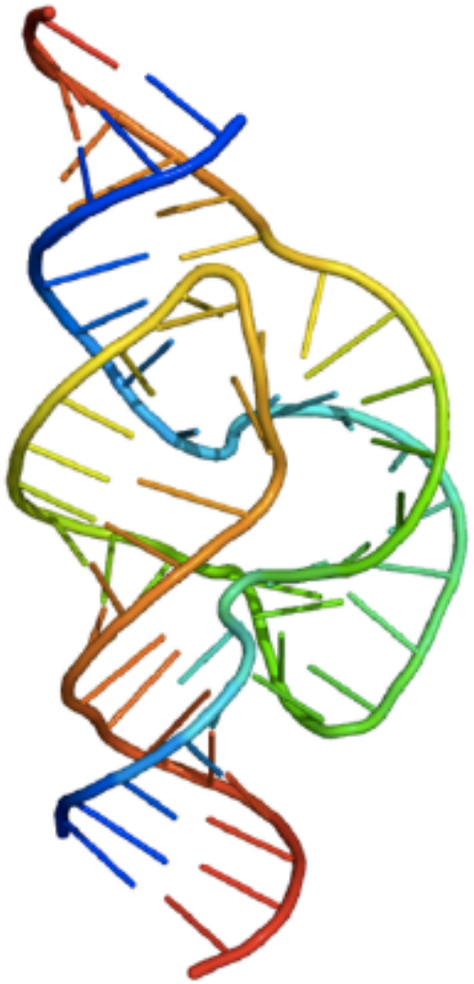
Life at the macro scale



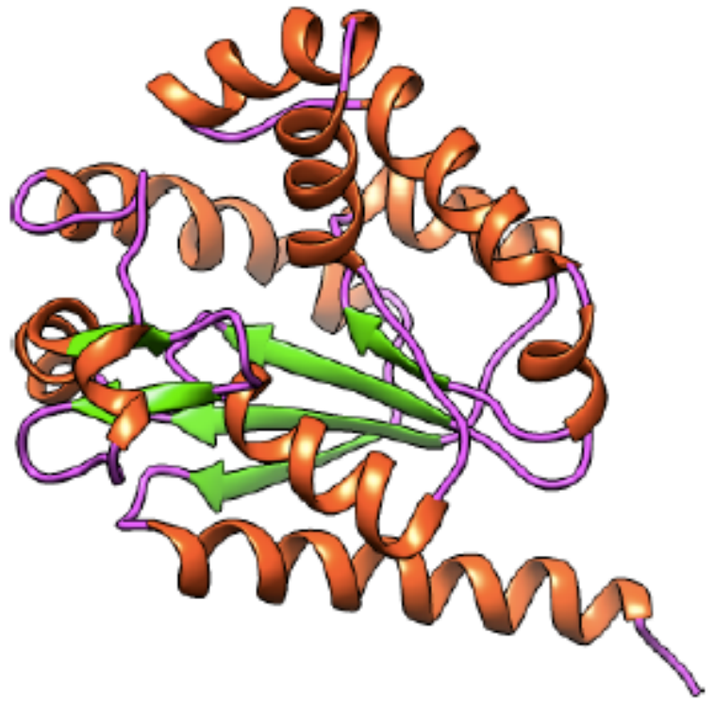
Life at the molecular scale



DNA



RNA



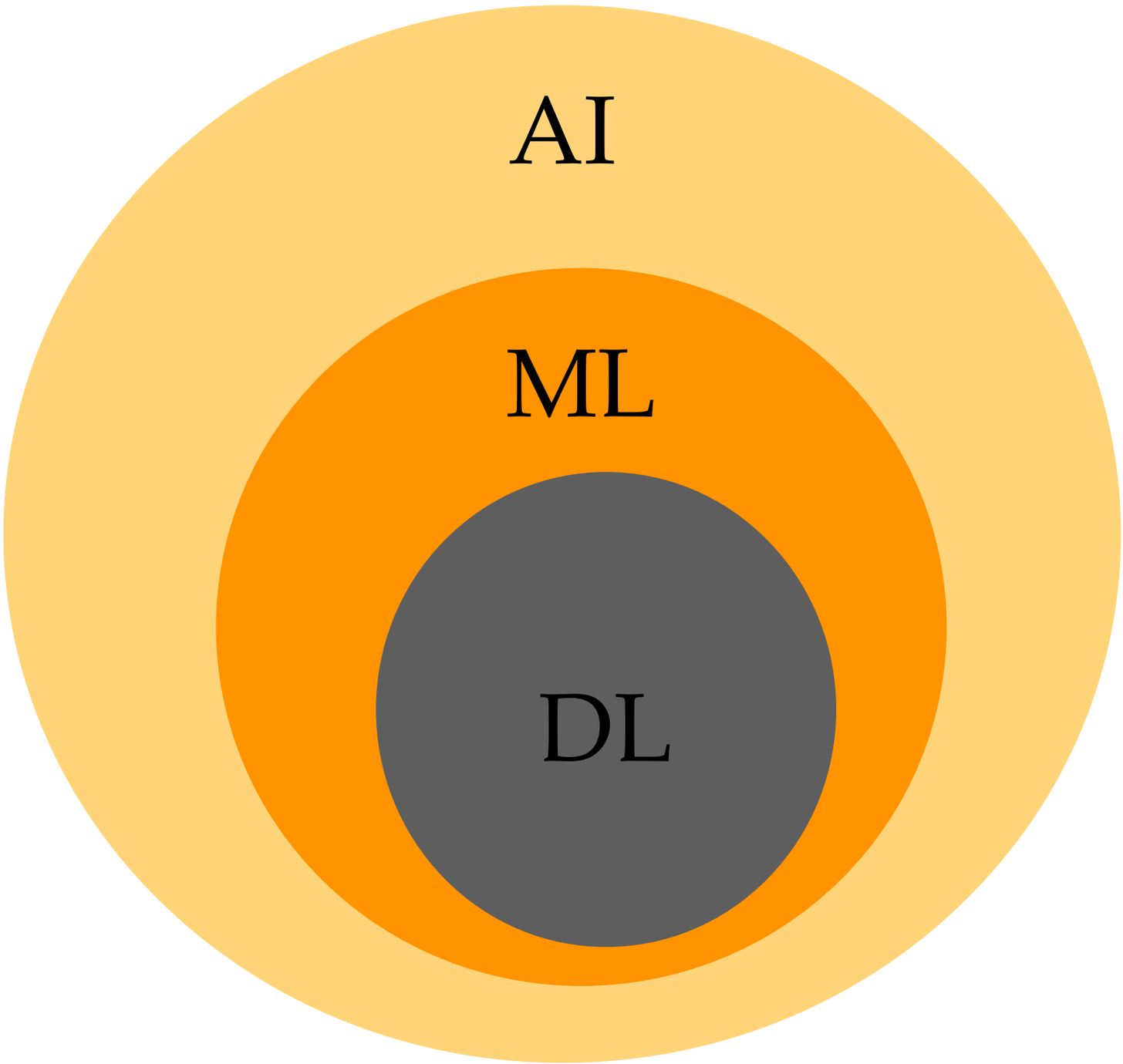
Protein

The inner life of the cell

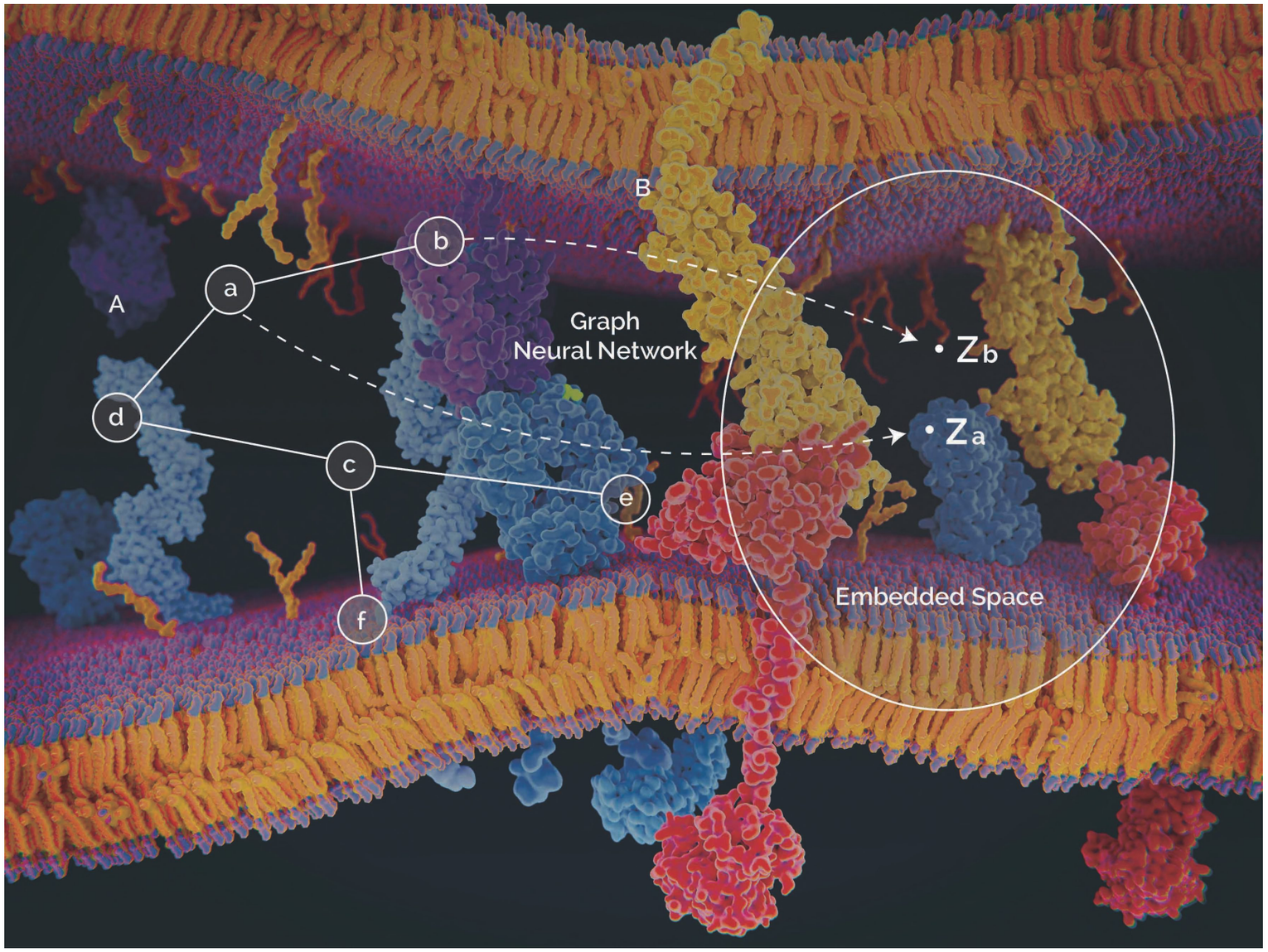
https://www.youtube.com/watch?v=wJyUtbn0O5Y

— BioVisions@Harvard

The role of AI



Deep Learning Revolution in Molecular Modeling



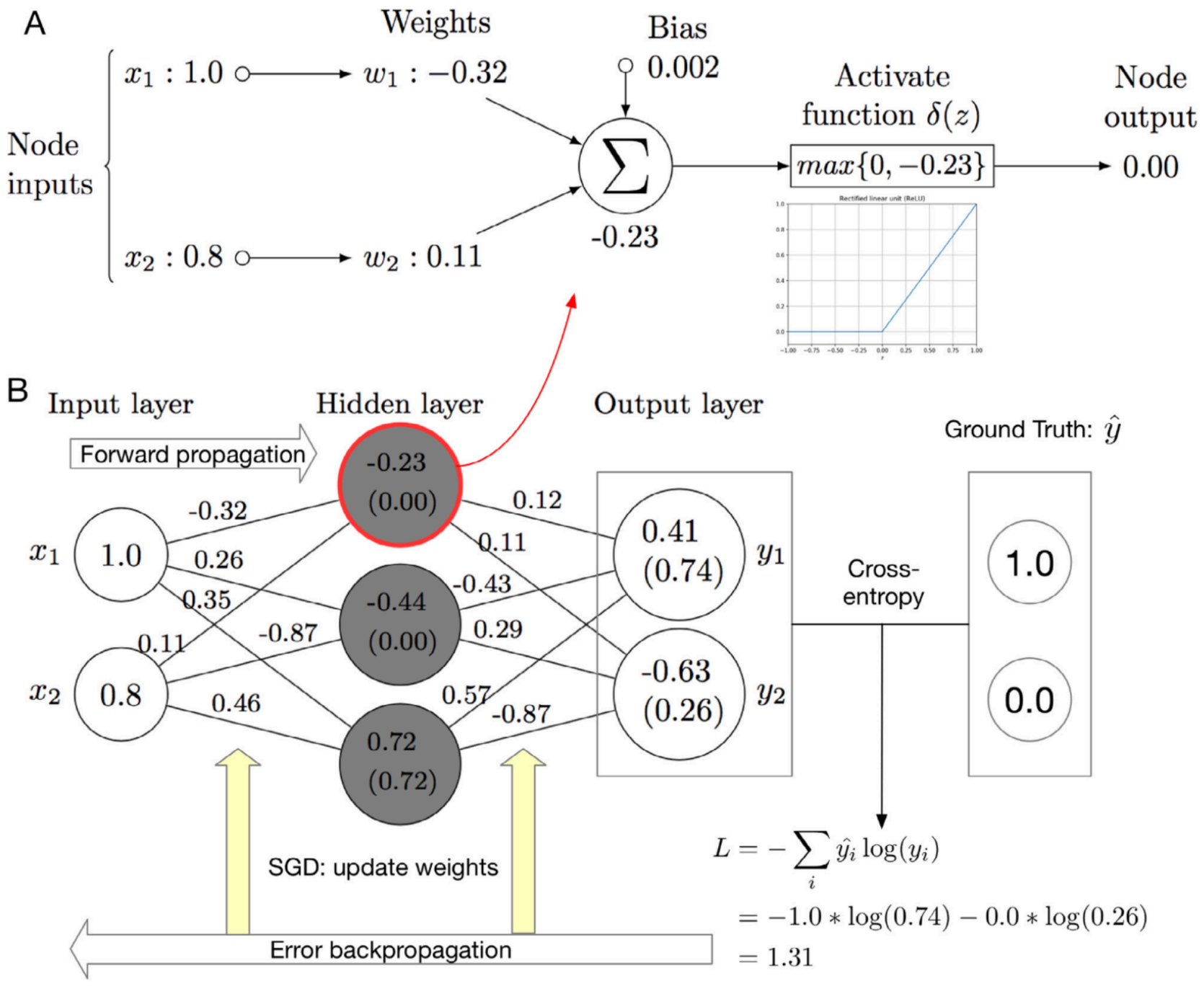
- **Progress in deep learning**
  - Deep fully connected NNs
  - ConvNet
  - RNN
  - Graph convolutional NNs
  - ResNet
  - GAN
  - VAE
  - …
- **Molecular data types**
  - Structured data
  - 1D sequence data
  - 2D image or profiling data
  - Graph data
  - 3D coordinate data
  - 4D dynamics data
  - …
- **The post-genomic "big data"**
  - High-throughput DNA sequencing
  - Post Moore's Law Computing

Few examples

Example	Model	Data type	Research direction	Task
Enzyme function prediction	DNN	Structured	Biomolecular function prediction	Classification
Gene expression regression	DNN	Structured	Biomolecular property prediction	Regression
RNA-protein binding sites prediction	CNN	1D data	Sequence analysis	Classification
DNA sequence function prediction	CNN, RNN	1D data	Sequence analysis	Classification
Biomedical image classification	ResNet	2D data	Biomedical image processing	Classification
Protein interaction prediction	GCN	Graph	Biomolecule interaction prediction	Embedding, Classification
Biology image super-resolution	GAN	2D image	Structure reconstruction	Data generation
Gene expression data embedding	VAE	2D data	Systems biology	DR, Data generation

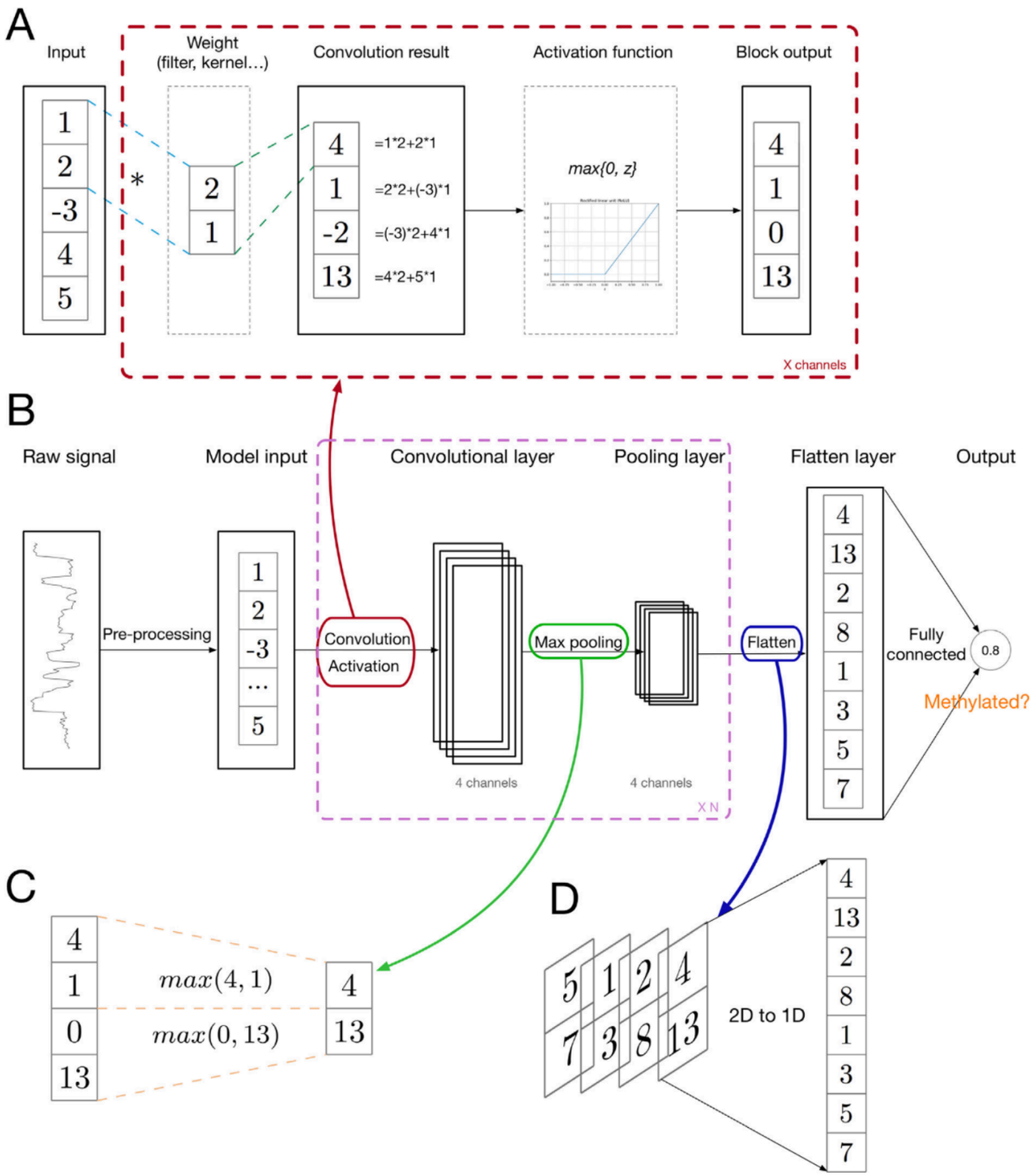
https://www.sciencedirect.com/science/article/abs/pii/S1046202318303256

Shallow neural network



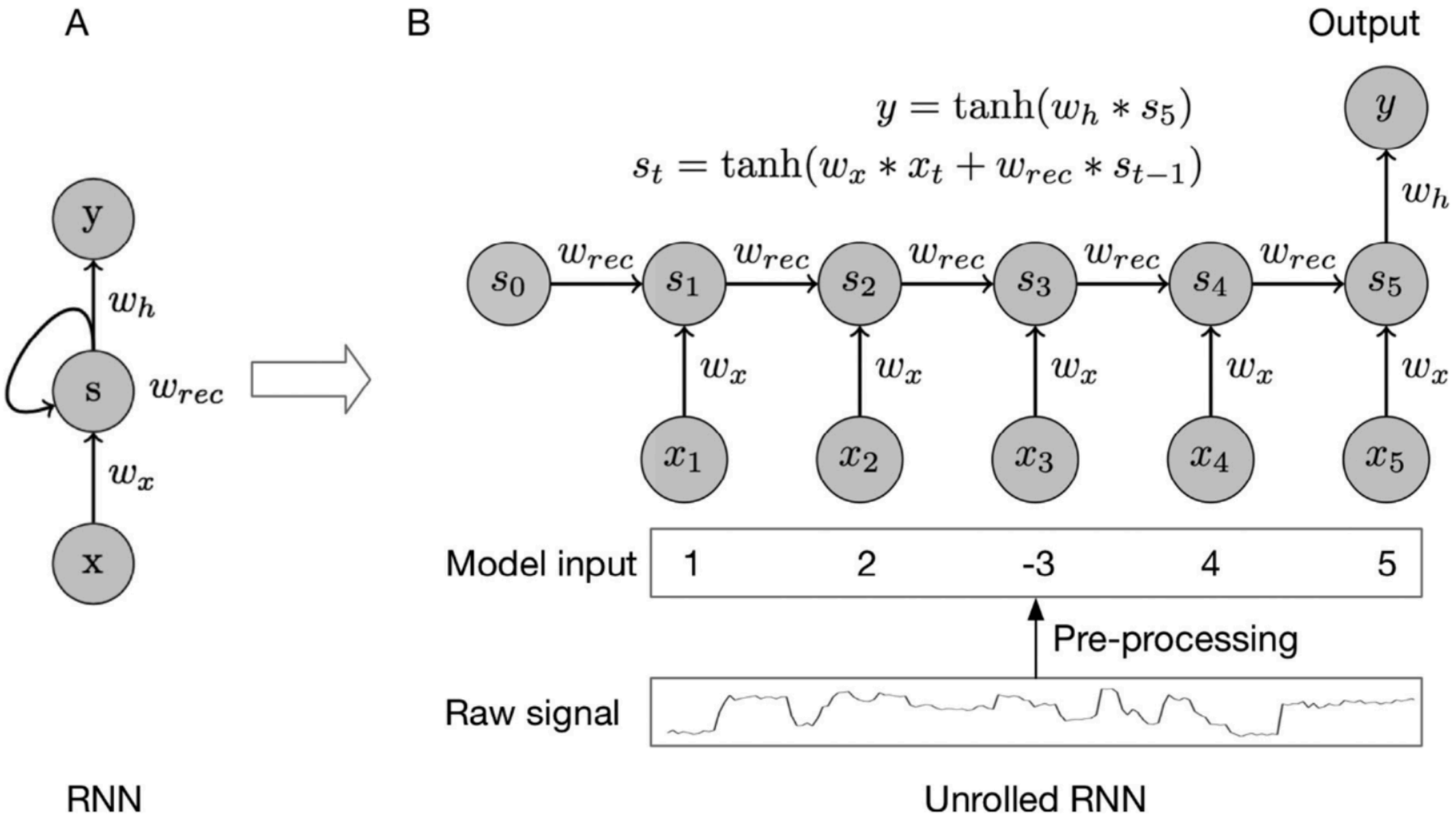
https://www.sciencedirect.com/science/article/abs/pii/S1046202318303256

Convolution neural network



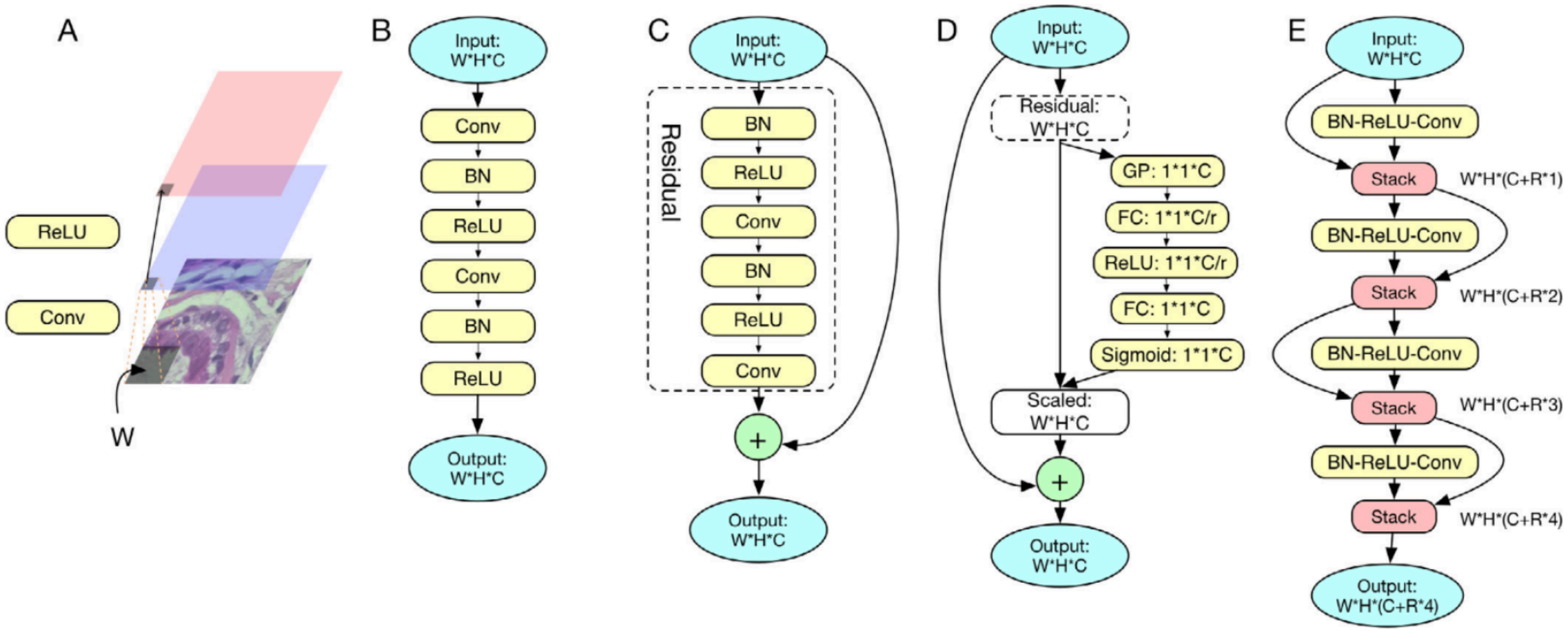
https://www.sciencedirect.com/science/article/abs/pii/S1046202318303256

Recurrent neural network



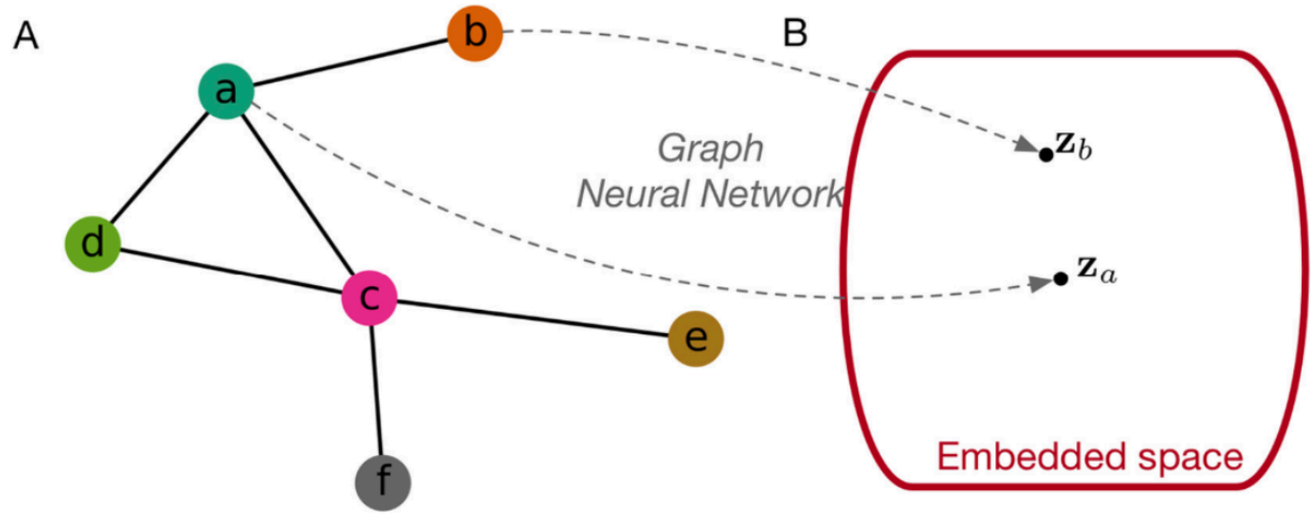
https://www.sciencedirect.com/science/article/abs/pii/S1046202318303256

Variants of convolutional neural networks



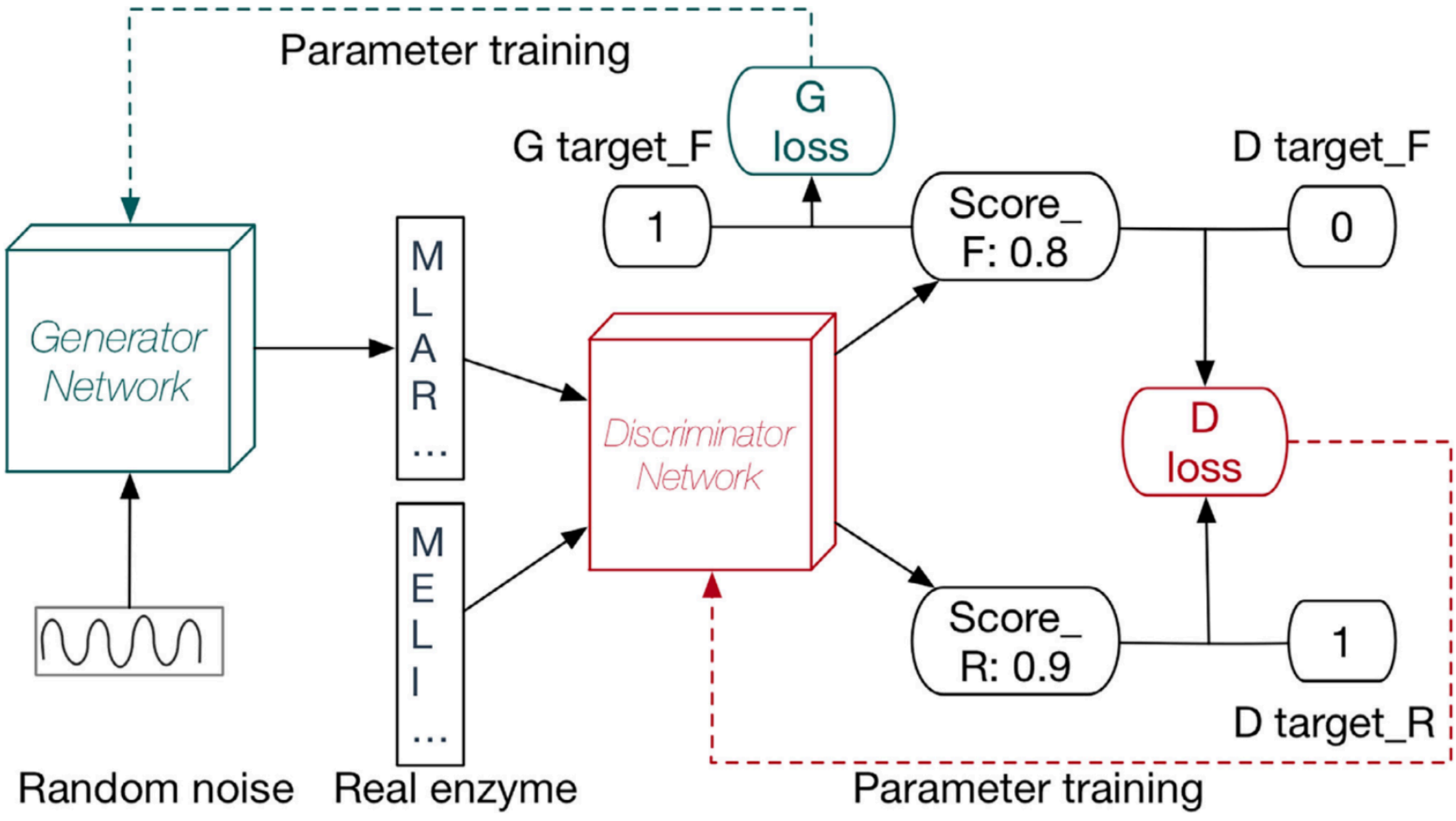
https://www.sciencedirect.com/science/article/abs/pii/S1046202318303256

Graph neural network



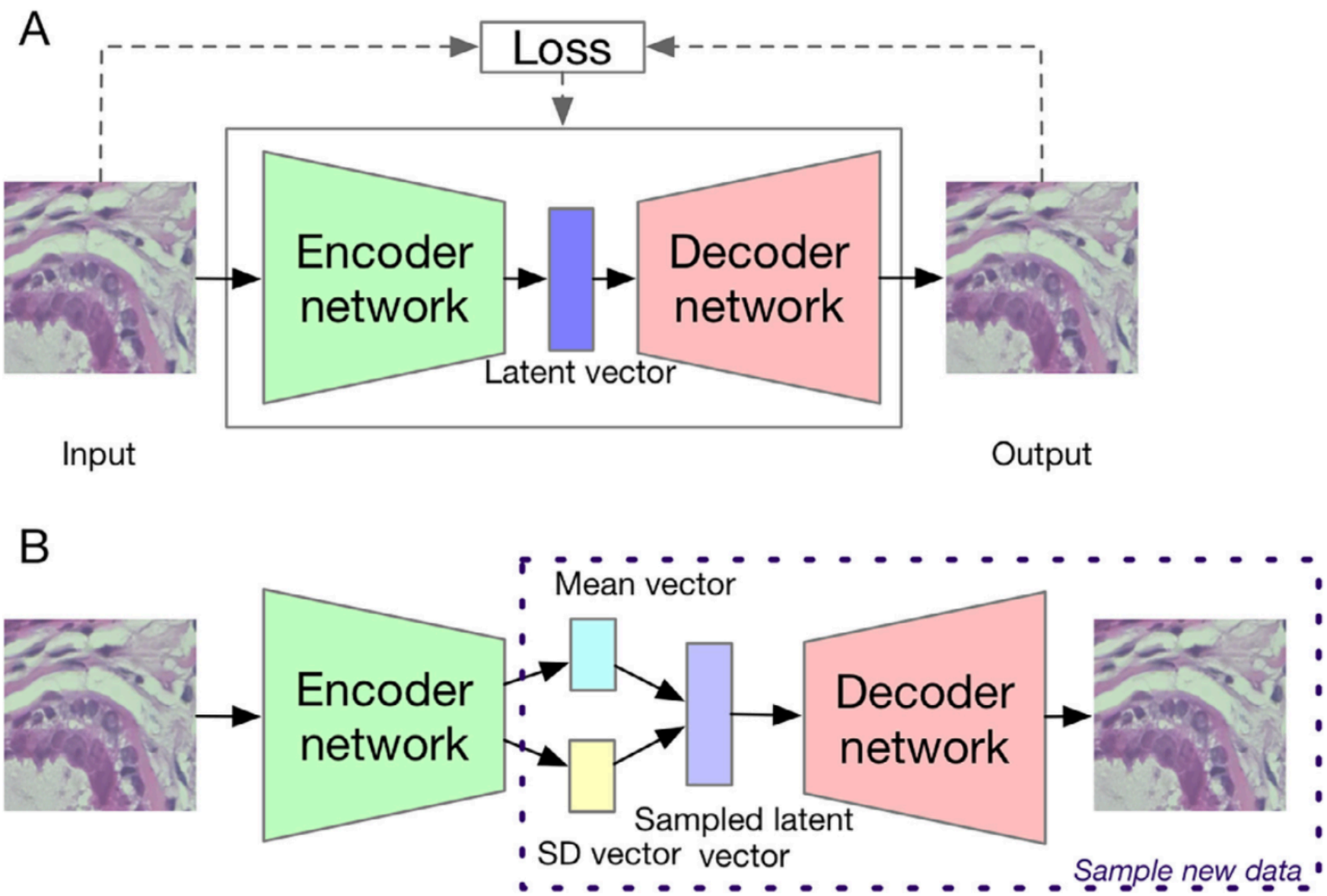
https://www.sciencedirect.com/science/article/abs/pii/S1046202318303256

Generative adversarial network



https://www.sciencedirect.com/science/article/abs/pii/S1046202318303256

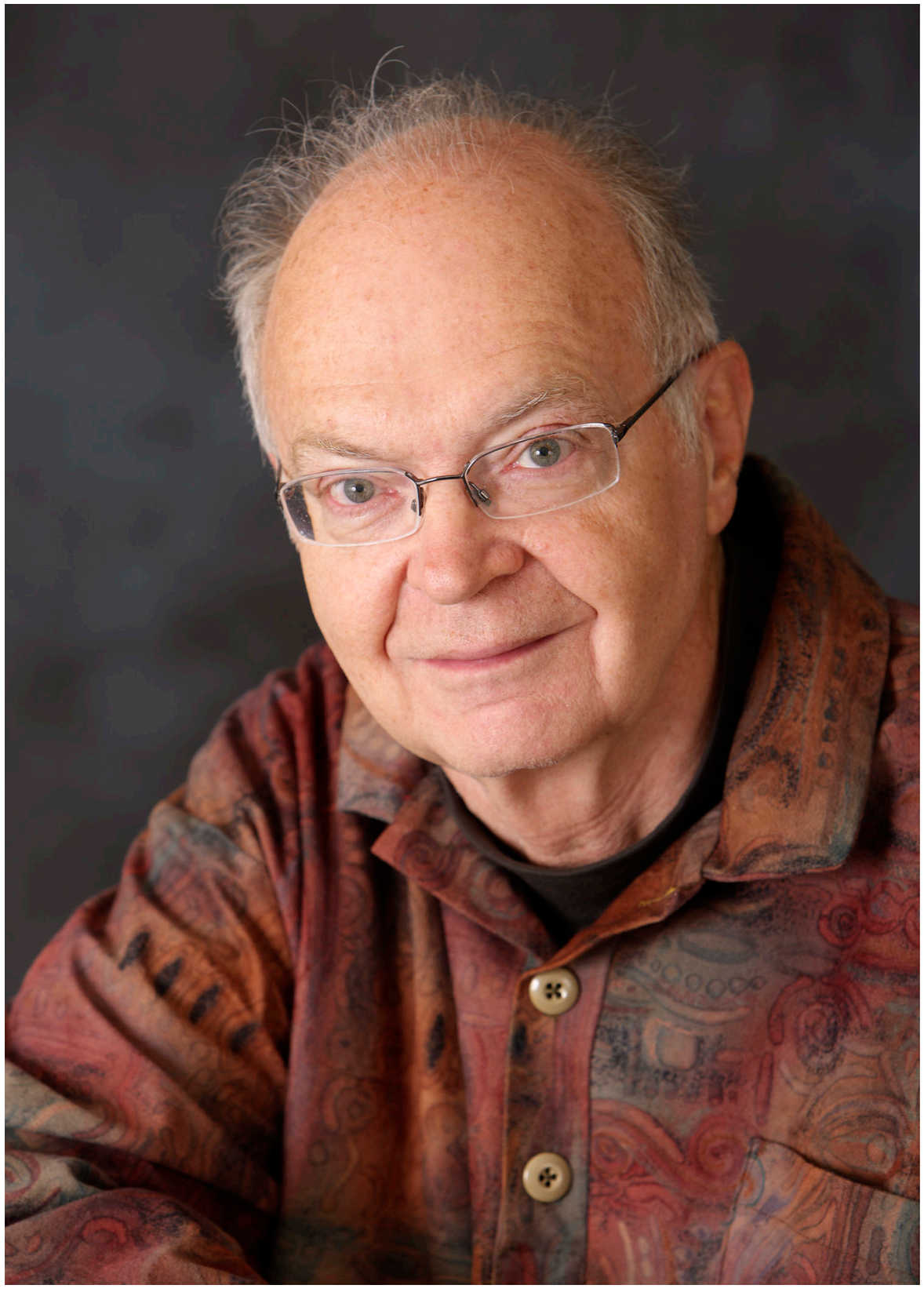
Autoencoder



https://www.sciencedirect.com/science/article/abs/pii/S1046202318303256

Issues to think about

- Lack of data
- Overfitting
- Imbalanced data
- Interpretability
- Uncertainty scaling
- Catastrophic forgetting
- Reducing computational requirement and model compression



“

I can’t be as confident about computer science as I can about biology. Biology easily has 500 years of exciting problems to work on. It’s at that level.

”

Donald Knuth

What this class is about?

- Surveying the emerging field of computational modeling of molecules with a particular focus on predictive modeling driven by advances in AI and Machine (Deep) Learning.

- **Goal**
  - After finishing this course, you should be ready with a paper to submit to a conference/journal

- **Target audience**
  - PhD students, MS Thesis students (Intro time!)

What this class is NOT

- **NOT the target audience**
  - Coursework-only grad students
  - Students looking to add an AI course to their resume
- **NOT the goal**
  - Teaching a toolkit (e.g., TensorFlow/PyTorch)
  - Programming

Prerequisites

- This course is appropriate for graduate students in computer science, computational biology, bioinformatics, and statistics.
- Familiarity with fundamental concepts in machine learning, statistics, probability and algorithms is expected.

Course Information

- **Course website**
  - https://people.cs.vt.edu/dbhattacharya/courses/cs6824/

- **Piazza**
  - https://piazza.com/vt/spring2022/cs6824/home

- **Canvas**
  - https://canvas.vt.edu/courses/145337

Course Staff

- **Instructor**
  - Debswapna Bhattacharya
  - Office Hours: Monday and Wednesday 4:00 pm - 5:00 pm at Torgersen 2160N

Grading

- **Class participation: 20%**
  - Involvement in class - 5%
  - Peer review - 15%
- **Paper presentation: 30%**
  - Each student presents 2 papers (1 before and 1 after spring break each 15%)
  - List of papers will be available for students to pick
- **Project: 50%**
  - Proposal - 20% (10% for whitepaper and 10% for presentation)
  - Report/paper - 15%
  - Presentation - 15%

Grading scale

(after computing ceiling of the final percentage of points earned)

A: 93%-100%	A-: 90%-92%	B+: 87%-89%	B: 83%-86%
B-: 80%-82%	C+: 77%-79%	C: 73%-76%	C-: 70%-72%
D+: 67%-69%	D: 63%-66%	D-: 60%-62%	F: Below 60%

Tentative Course Schedule

- **Introductory lectures**
  - Crash course on molecules
  - A general Overview of AI-powered Molecular Modeling

- **Paper presentations**
  - Student presentation of research papers
  - Peer review of papers by students

- **Course projects**
  - Proposal whitepaper and presentation, peer review
  - Final presentation + report

Paper Presentations

- **Each student will present 2 papers picked by the student**
  - The goal of the presentation is to facilitate a discussion, focusing on:
    - Present the biological question and the corresponding computational abstraction
    - How did the authors address the problem?
    - Did they manage to answer the original biological question?
    - How can we improve the results? What are future directions?

- **The remaining students are required to write a short peer review**
  - Summary of the paper
  - Major and minor comments
  - Outlook/future directions

List of Papers

#	Title	Link	Published In
2	Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation	https://papers.nips.cc/paper/2018/hash/d60678e8f2ba9c540798ebbde31177e8-Abstract.html	NeurIPS 2018
3	A Likelihood-Free Inference Framework for Population Genetic Data using Exchangeable Neural Networks	https://papers.nips.cc/paper/2018/hash/2e9f978b222a956ba6bdf427efbd9ab3-Abstract.html	NeurIPS 2018
4	Neural Edit Operations for Biological Sequences	https://papers.nips.cc/paper/2018/hash/d0921d442ee91b896ad95059d13df618-Abstract.html	NeurIPS 2018
5	Generative modeling for protein structures	https://papers.nips.cc/paper/2018/hash/afa299a4d1d8c52e75dd8a24c3ce534f-Abstract.html	NeurIPS 2018
6	An image representation based convolutional network for DNA classification	https://iclr.cc/Conferences/2018/Schedule?showEvent=229	ICLR 2018
7	A Model to Search for Synthesizable Molecules	https://papers.nips.cc/paper/2019/hash/46d0671dd4117ea366031f87f3aa0093-Abstract.html	NeurIPS 2019
8	Evaluating Protein Transfer Learning with TAPE	https://papers.nips.cc/paper/2019/file/37f65c068b7723cd7809ee2d31d7861c-Paper.pdf	NeurIPS 2019
9	End-to-End Learning on 3D Protein Structure for Interface Prediction	https://papers.nips.cc/paper/2019/hash/6c7de1f27f7de61a6daddfffbe05c058-Abstract.html	NeurIPS 2019
10	Generative Models for Graph-Based Protein Design	https://papers.nips.cc/paper/2019/hash/f3a4ff4839c56a5f460c88cce3666a2b-Abstract.html	NeurIPS 2019
11	Learning Protein Structure with a Differentiable Simulator	https://openreview.net/forum?id=Byg3y3C9Km	ICLR 2019
12	Guiding Deep Molecular Optimization with Genetic Exploration	https://papers.nips.cc/paper/2020/hash/8ba6c657b03fc7c8dd4dff8e45defcd2-Abstract.html	NeurIPS 2020
13	Self-Supervised Graph Transformer on Large-Scale Molecular Data	https://papers.nips.cc/paper/2020/hash/94aef38441efa3380a3bed3faf1f9d5d-Abstract.html	NeurIPS 2020
14	Barking up the right tree: an approach to search over molecule synthesis DAGs	https://papers.nips.cc/paper/2020/hash/4cc05b35c2f937c5bd9e7d41d3686fff-Abstract.html	NeurIPS 2020
15	Reinforced Molecular Optimization with Neighborhood-Controlled Grammars	https://papers.nips.cc/paper/2020/hash/5f268dfb0fbef44de0f668a022707b86-Abstract.html	NeurIPS 2020
16	RNA Secondary Structure Prediction By Learning Unrolled Algorithms	https://openreview.net/forum?id=S1eALyrYDH	ICLR 2020
17	Energy-based models for atomic-resolution protein conformations	https://openreview.net/forum?id=S1e_9xrFvS	ICLR 2020
18	Co-evolution Transformer for Protein Contact Prediction	https://papers.nips.cc/paper/2021/hash/770f8e448d07586afbf77bb59f698587-Abstract.html	NeurIPS 2021
19	Multi-Scale Representation Learning on Proteins	https://papers.nips.cc/paper/2021/hash/d494020ff8ec181ef98ed97ac3f25453-Abstract.html	NeurIPS 2021
20	Language models enable zero-shot prediction of the effects of mutations on protein function	https://papers.nips.cc/paper/2021/hash/f51338d736f95dd42427296047067694-Abstract.html	NeurIPS 2021
21	Capturing implicit hierarchical structure in 3D biomedical images with self-supervised hyperbolic representations	https://papers.nips.cc/paper/2021/hash/291d43c696d8c3704cdbe0a72ade5f6c-Abstract.html	NeurIPS 2021
22	Neural Distance Embeddings for Biological Sequences	https://papers.nips.cc/paper/2021/hash/9a1de01f893e0d2551ecbb7ce4dc963e-Abstract.html	NeurIPS 2021
23	Hit and Lead Discovery with Explorative RL and Fragment-based Molecule Generation	https://papers.nips.cc/paper/2021/hash/41da609c519d77b29be442f8c1105647-Abstract.html	NeurIPS 2021
24	GeoMol: Torsional Geometric Generation of Molecular 3D Conformer Ensembles	https://papers.nips.cc/paper/2021/hash/725215ed82ab6306919b485b81ff9615-Abstract.html	NeurIPS 2021
25	A 3D Generative Model for Structure-Based Drug Design	https://papers.nips.cc/paper/2021/hash/314450613369e0ee72d0da7f6fee773c-Abstract.html	NeurIPS 2021
26	Learning from Protein Structure with Geometric Vector Perceptrons	https://openreview.net/forum?id=1YLJDvSx6J4	ICLR 2021

Paper Presentations

- **Each student will present 2 papers picked by the student**
  - The goal of the presentation is to facilitate a discussion, focusing on:
    - Present the biological question and the corresponding computational abstraction
    - How did the authors address the problem?
    - Did they manage to answer the original biological question?
    - How can we improve the results? What are future directions?

- **The remaining students are required to write a short peer review**
  - Summary of the paper
  - Major and minor comments
  - Outlook/future directions

Peer-reviews via EasyChair

Course Project

- Can be done individually or in a group (2 students max / project)
- First write and present a proposal, which will receive feedback from fellow students via EasyChair
- Then, conduct research and write a paper
- Pick a venue (conference/journal) and use NeurIPS LaTex template style for your paper
- Extra credit for shooting for conference/journal submission

Deadlines

Note: All deadlines are until 11:59 PM EST unless otherwise specified

- **Jan 31**: Pick 2 papers to present from this list and electronically submit the papers to EasyChair as the corresponding author (publicly discuss your preference in Piazza)
- **Feb 4**: Accept invitation to be a PC member
- Submit presentation and peer-review reports no later than 24 hours before each paper presentation
- **Mar 4**: Submit project proposal whitepaper to Canvas and concurrently upload to EasyChair
- **Mar 13**: Submit project proposal presentation to Canvas
- **Mar 21**: Submit proposal peer reviews via EasyChair
- **Apr 13**: Submit project presentation to Canvas
- **May 4**: Submit final project report to Canvas and concurrently upload to EasyChair

How to stay in touch?

- **Primary means of communication —Piazza**
  - No direct email to instructor unless private information
  - Instructor can provide answers to everyone on forum
  - Class participation credit for answering questions on forum!

- **Class Mailing List**
  - class-cs-6824-20577-202201-g@vt.edu

Policies

- **Collaboration Policy**
  - You are encouraged to collaborate
  - Give proper credit when its due
  - Project proposal/report will be plagiarism checked

- **Academic integrity**
  - Students enrolled in this course are responsible for abiding by the Honor Code
  - Zero-tolerance philosophy regarding plagiarism or other forms of cheating

- **Principles of Community**
  - The course will include in-class discussions, and we will adhere to Virginia Tech Principles of Community.

- **Accessibility**
  - If any student needs special accommodations because of any disabilities, please contact the instructor during the first week of classes.
  - Such students are encouraged to work with The Office of Services for Students with Disabilities to help coordinate accessibility arrangements.

- **COVID-19 Policy**
  - Please follow the instructions posted at the University and public health guidelines for the latest COVID-19 Policy.

Todo: before next class

Go through the course webpage at:
https://people.cs.vt.edu/dbhattacharya/courses/cs6824/

…and ask any questions in the next class.

Get into Piazza:
https://piazza.com/vt/spring2022/cs6824/home