CSE7850/CX4803

Machine Learning in Computational Biology

Lecture 1: Course Introduction

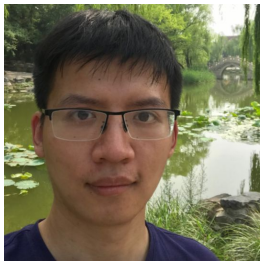
Yunan Luo

Teaching team

Instructor: **Yunan Luo**

- Assistant Professor, School of Computational Science and Engineering
- Office hours: Monday, 2:00 pm - 3:00 pm (on Zoom)
  - Visit Canvas -> Syllabus for booking slots
- Email: yunan@gatech.edu
- My research: Computational Biology and Machine Learning

Teaching Assistants:



Fan Qu (fan.qu@gatech.edu)
CS MS student
Office hours: Thu 2-3 pm

Cheng Wan (cwan39@gatech.edu)
CS PhD student
Office hours: Fri 2-3 pm

OHs start from next week

Can you name one cool AI application?

ChatGPT

LU

write an opening remark for my class "machine learning in computational biology" at Georgia Tech

Good morning everyone, and welcome to Machine Learning in Computational Biology at Georgia Tech! My name is [insert your name], and I will be your instructor for this course.

Before we get started, I want to take a moment to introduce the topic of machine learning in computational biology and why it is such a crucial and exciting field. Machine learning is a subset of artificial intelligence that involves the use of algorithms to analyze and learn from data in order to make predictions or take actions. In computational biology, we apply these techniques to understand and analyze biological systems, such as genetics, protein sequences, and medical records. By using machine learning, we can gain insights that would not be possible through traditional methods and make more informed decisions in fields such as drug discovery, personalized medicine, and public health.

I'm thrilled to have all of you in this class, and I look forward to exploring the exciting world of machine learning in computational biology with you over the next few weeks. So, let's get started!

https://chat.openai.com/

Text-to-image generation

A beautiful field of vibrant flowers atmospheric, hyper realistic, 8k, epic composition, cinematic, octane render, artstation landscape vista photography by Carr Clifton & Galen Rowell, 16K resolution, Landscape veduta photo by Dustin Lefevre & tdraw, 8k resolution, detailed landscape painting by Ivan Shishkin, DeviantArt, Flickr, rendered in Enscape, Miyazaki, Nausicaa Ghibli, Breath of The Wild, 4k detailed post processing, artstation, rendering by octane, unreal engine —ar 16:9

Text-to-image generative models
(Stable diffusion, DALL-E, Imagen, ...)

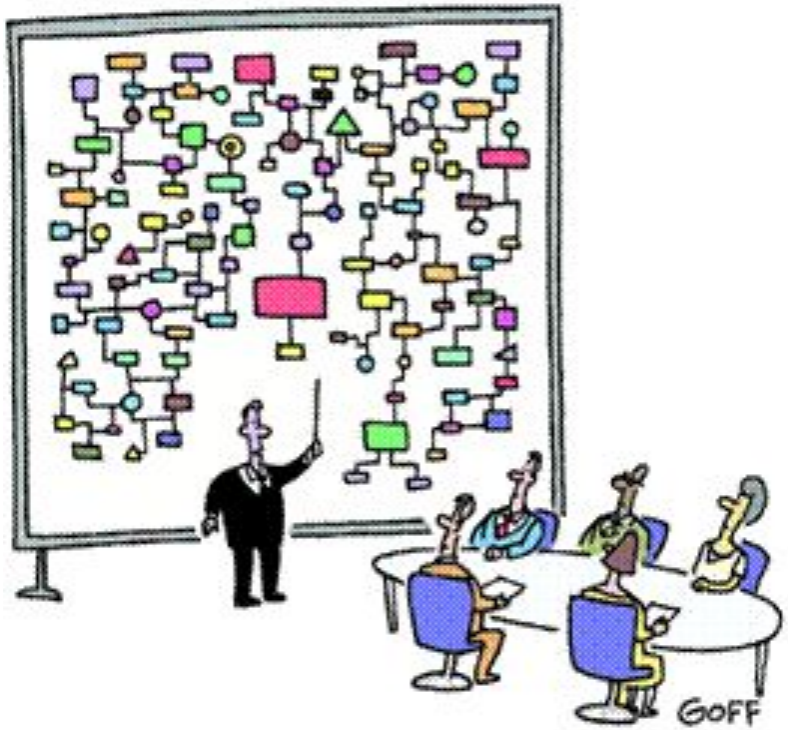
Image from: https://prompthero.com/prompt/722b1199390

We will learn the AI/ML techniques underlying those models

… and their applications in computational biology

Why do we need computation for biology?

- High-dimensional
- Noisy
- Huge
- Sparse



"And that's why we need a computer."

PLOS | BIOLOGY

RESEARCH MATTERS

All biology is computational biology

Florian Markowetz*

University of Cambridge, Cancer Research UK Cambridge Institute, Cambridge, United Kingdom

* florian.markowetz@cruk.cam.ac.uk

Abstract

Here, I argue that computational thinking and techniques are so central to the quest of understanding life that today all biology is computational biology. Computational biology brings order into our understanding of life, it makes biological concepts rigorous and testable, and it provides a reference map that holds together individual insights. The next modern synthesis in biology will be driven by mathematical, statistical, and computational methods being absorbed into mainstream biological training, turning biology into a quantitative science.

Computational biology

- Is not about one problem (e.g., designing better computer chips, better compilers, better graphics, better networks, better operating systems, etc.)
- Is about a family of very different problems, all related to biology, all related to each other
- How can computers help solve any of this family of problems ?

Bioinformatics vs computational biology?

BIOINFORMATICS & COMPUTATIONAL BIOLOGY = SAME? NO.

I spent the first 15 years of my professional life unwilling to recognize a difference between bioinformatics and computational biology. It was not because I didn't think that there was or could be a difference, but because I thought the difference was not significant. I have changed my position on this. I now believe that they are quite different and worth distinguishing. For me,

- the creation of tools (algorithms, databases) that solve problems. The goal is to build useful tools that work on biological data. It is about engineering.
- the study of biology using computational techniques. The goal is to learn new biology, knowledge about living systems. It is about discovery.

Bioinformatics vs computational biology?

BIOINFORMATICS & COMPUTATIONAL BIOLOGY = SAME? NO.

I spent the first 15 years of my professional life unwilling to recognize a difference between bioinformatics and computational biology. It was not because I didn't think that there was or could be a difference, but because I thought the difference was not significant. I have changed my position on this. I now believe that they are quite different and worth distinguishing. For me,

- *Bioinformatics* = the creation of tools (algorithms, databases) that solve problems. The goal is to build useful tools that work on biological data. It is about engineering.
- *Computational biology* = the study of biology using computational techniques. The goal is to learn new biology, knowledge about living systems. It is about discovery.
- Or the opposite?... Will be used interchangeably in this course

Machine learning, computational biology, and you

- You can learn a number of modern machine learning tools
  - Underlying mathematics and applications. E.g.: deep learning, large language models, generative models, transformers, …
- These tools owe their origin to computer science, information theory, probability theory, statistics, etc.
  - E.g.: Markov random field, likelihood maximization, gradient descent, …
- You can learn the language of biology, enough to understand what the problems are
  - E.g.: central dogma, protein sequence/structure/function, gene regulation, …
- You can apply the tools to these problems and contribute to science
  - E.g.: PyTorch, GPU servers, Google Colab, AI for Science, …

Objectives of this course

Introduction to computational biology

- Important problems in computational biology
- Machine learning techniques for data analysis
- Understand how methods work

Gaining practical and research experience

- Paper presentation / Literature review
  - Ability to present key ideas to other people
  - Ability to ask critical questions
- Course project and homework
  - Hands-on research experience

For you future career:

- (Bioinfo) Learn how to formulate computational biological questions as machine learning problems.
- (CS) Learn how to apply and customize ML algorithms for scientific problems (AI for Science)
- Gain hands-on experience of using widely-used bioinformatics and machine learning toolbox to analyze biological data.

Important biological questions?

"Why do humans have so few genes?"

"Can we understand DNA code?"

"Can we understand gene function?"

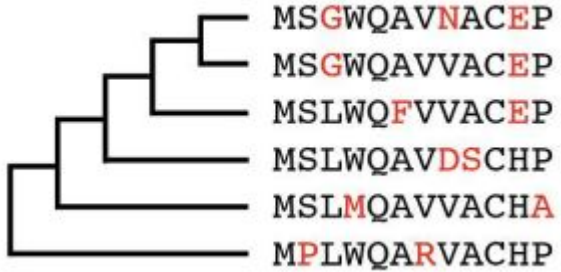
"How did cooperative behavior evolve?"

"Can we cure cancer?"

What does biological data look like?

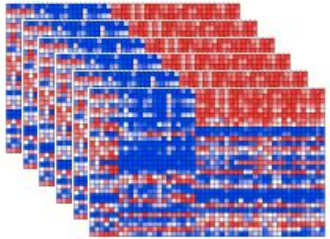
Sequence data

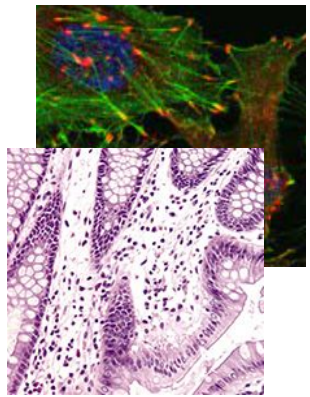
- Protein/DNA sequences
- Deep learning for sequence data



High-dimensional data

- Gene expression, biomedical images
- Dimensionality reduction and feature selection
- Low-rank approximation

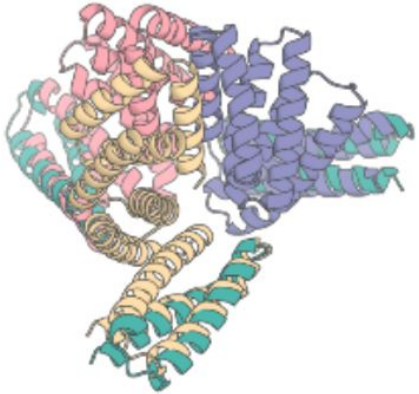


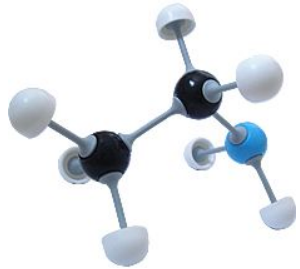


What does biological data look like?

Structure data

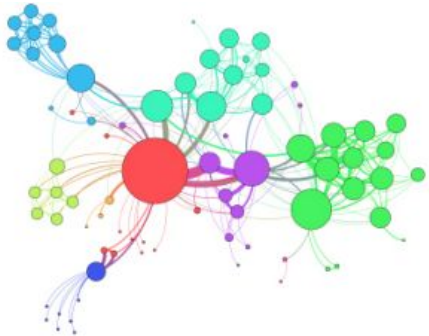
- Protein structure prediction/generation
- Small-molecule structure (drug compounds)
- 3D deep learning





Network data

- Molecular network, patient network
- Random walk algorithms
- Graph neural network



Machine Learning

Basics:

- Regression & Classification
- Optimization & Evaluation
- Clustering
- Neural networks

Supervised learning:

- Prediction: classic models
- Feature selection: LASSO, attention

Unsupervised learning:

- Dimensionality reduction & embedding
- Probabilistic modeling: HMM, variational inference
- Generative models

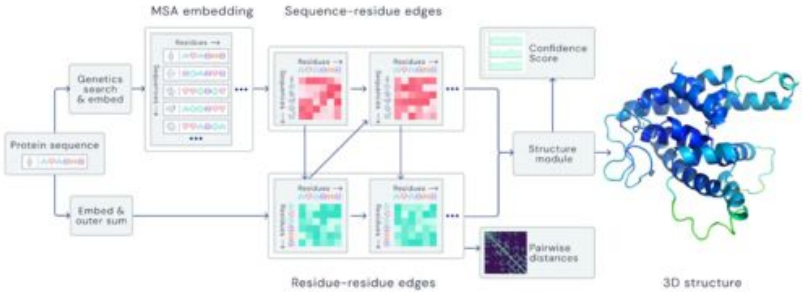
Architectures:

- Convolutional neural networks (CNN)
- Recurrent neural networks (RNN)
- Transformers
- Graph neural networks (GNN)

Cutting-edge research topics:

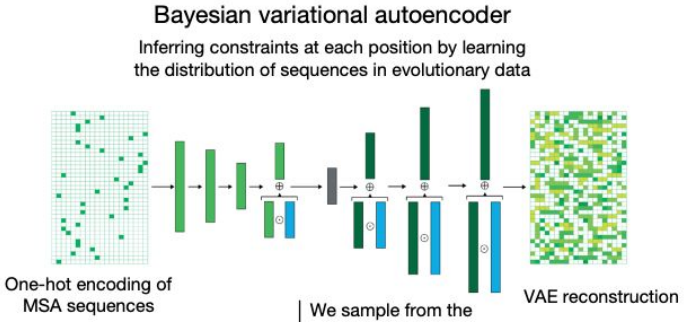
- Language models
- Diffusion models
- 3D deep learning

Examples of applications in biomedicine



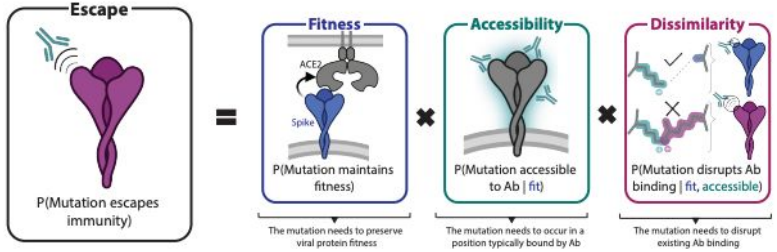
Predict protein structure

source



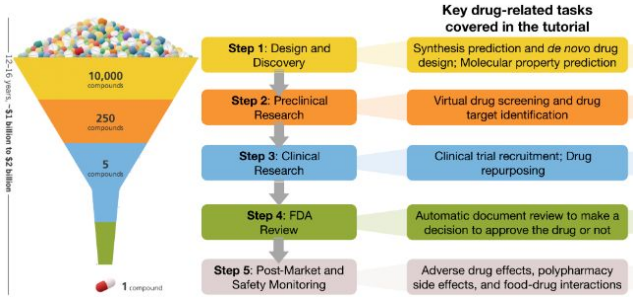
Identify mutations that cause disease

source



Forecast virus that will escape from human immune system

source



Drug discovery

source

TODO after this class

Please read “Molecular Biology for Computer Scientists” by Lawrence Hunter

Course information

Canvas:

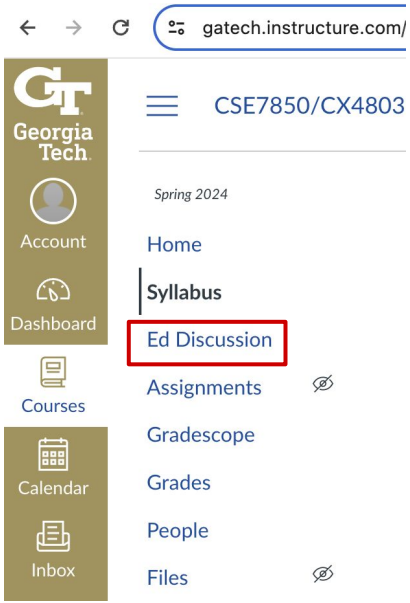
- https://gatech.instructure.com/courses/369210
- Slides, homework release, grades, book office hours, …

Ed discussion:

- Discussion platform
- For all course-related discussion, questions, requests, …
- Link on Canvas

GradeScope:

- Homework submission
- Link on Canvas



Prerequisites

Biology

- Basic concepts in molecular biology (next lecture)
- Reference:
  Molecular Biology for Computer Scientists
  by Lawrence Hunter

Machine learning/Computational analytics

- Probability and statistics
- Linear algebra
- Programming experience in Python (required) and PyTorch (preferred)
- Reference:
  Dive into Deep Learning
  by Aston Zhang, Zachary C. Lipton, Mu Li, Alexander J. Smola

Tentative schedule

- **Introductory lectures** (Jan 10 to Feb 12)
  - Algorithms in Computational Biology
  - Foundations of Machine Learning
- **Research methods lectures** (Feb 14 to Mar 04)
  - Frontier ML methods in for CompBio
    - Learning from sequence data
    - Learning from structure data
    - Learning from network data
    - Learning from high-dimensional data
- **Advanced topics + Paper discussion** (Mar 06 to Apr 22)
  - Recent research papers in top scientific journals and ML/AI conferences
  - Student presentation + literature review
- **Course projects**
  - Proposal submission (due after Spring break)
  - Final report submission (due Apr 29)

Detailed schedule on Canvas->Syllabus or this link

Course evaluation

- Homework assignment (50%)
  - ~4 problem sets (written + programming questions)
  - Mini data science challenge (Kaggle like)
- Paper presentation or literature review (15%)
- Course project (30%)
  - Proposal (5%) + final report (25%)
- Participation (5%)
  - Paper discussion (details to follow)
- Bonus (+3%)
  - ~3 quizzes; ~20 min to complete each
  - Earn bonus 1% course point each
  - Max +3% to final grade

Detailed in course Syllabus

Homework (50%)

- ~4 HW assignments
  - Including both theory + programming questions
  - Programming: we will provide interactive iPython Notebook
    - We provide the starting codes; you complete the key code blocks

- Mini data science challenge (Kaggle-like)
  - Open-ended ML model development
    - We set up a prediction task in computational biology and released the training data
    - You develop your own ML models and submit the predictions for a withheld test set
    - We return you the prediction accuracy of your model
    - Will last for multiple weeks during this semester; keep refining your models and submit!

Paper presentation or Literature Review (15%)

- We will release a paper list (25 papers)
- Students to form teams to present papers or write literature review for a paper from the list
- Sign-up spreadsheet to be released in early Feb

- In-class Presentation
  - Form a team of 2 students (depending on class size)
  - 20 min presentation
  - 5 min Q&A + open-ended discussion
  - 50 slots

- Literature review
  - Form a team of 2 students (depending on class size)
  - Write a comprehensive literature review
  - Using the selected paper as the “seed paper” and find related papers on the same topics
  - 50 slots

Detailed in course Syllabus

Course project (30%)

- Form a team of 3-4 students (not necessarily the same team for presentation/literature review)

- **Research project**
  - Develop machine learning methods to interesting problems in computational biology.
  - Examples:
    - (1) Formulating a novel problem in computational biology as a ML problem and implementing ML methods you learned in this class to address it;
    - (2) Developing novel ML methods and applying them to an existing computational biology problem;
    - (3) Creating a benchmarking dataset for comparing the performance of different machine learning methods for a specific biology problem
  - Write a report

Detailed in course Syllabus

Participation (5%)

In-class discussion (2%)

- Ask an question to at least one presenting group
- The question is expected to be in-depth and preferably can prompt discussion/debate
- Log your question in a survey form (to be released) to receive the points
- Each validated in-class question will receive 2 points (2 points maximum)

Online discussion (1% x 3)

- Within that post, write a review of a paper that will be presented by other students, which includes a short summary (~100 words) and your comments (pros/cons of the paper)
- Should post the review on Ed by the day before the presentation date
- Each review counts as 1 participation point (maximum 5 points)

Detailed in course Syllabus

Survey

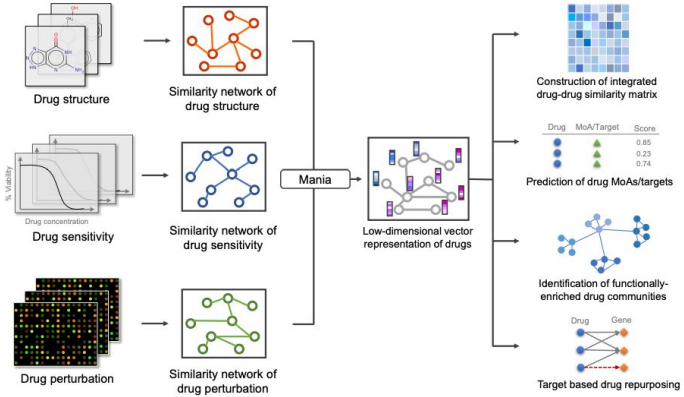
Please take 2 mins to complete this survey

http://tinyurl.com/mlb-survey-s24

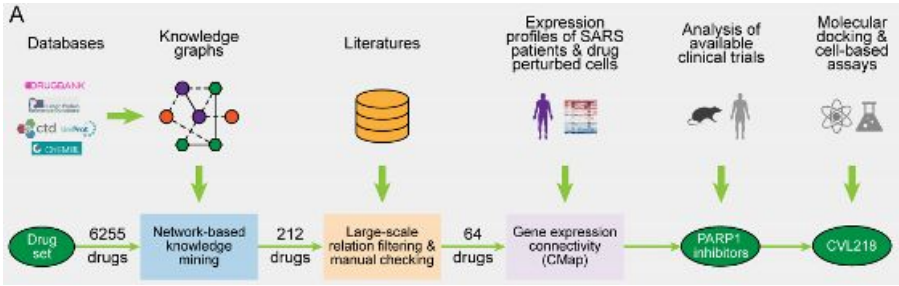
Research in my group

Machine learning for scientific discovery in biomedicine

Drug discovery



Predict new protein targets for existing drugs

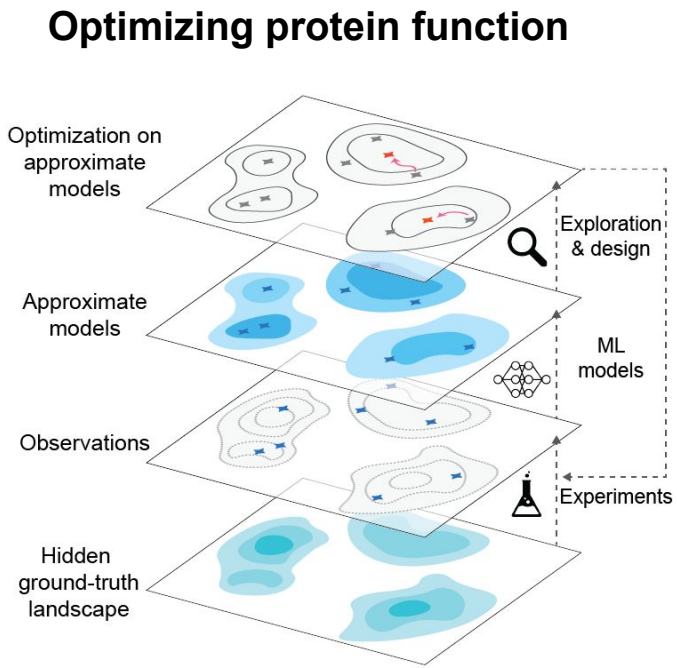


Predict potential drugs for COVID-19

Luo et al., *Nature Communications*, 2017

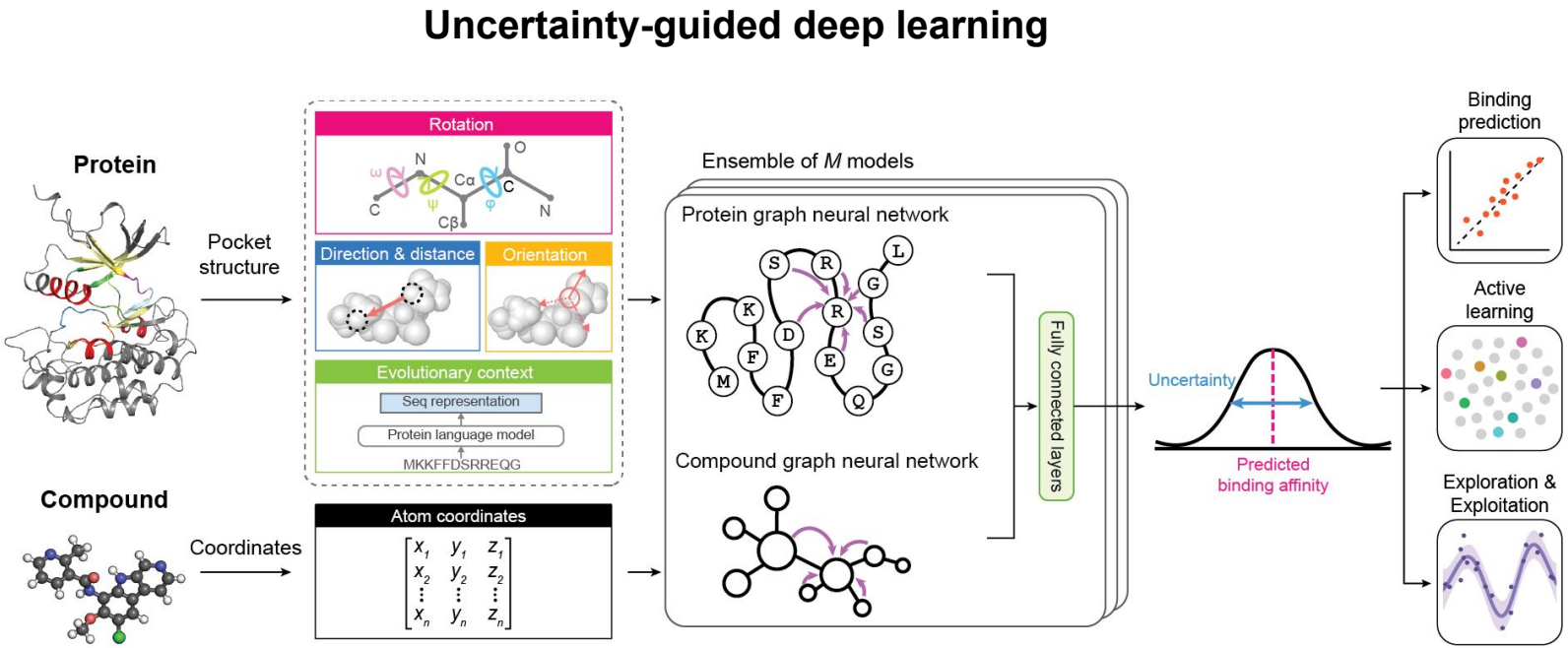
Ge et al. *Signal Transduction and Targeted Therapy*, 2021

Machine learning for scientific discovery in biomedicine



Luo et al., *Nature Communications*, 2021

Machine learning for scientific discovery in biomedicine



Luo et al., *Nature Machine Intelligence*, 2023