Lecture *0* - Introduction

AI in Genetics

ZOO6927 / BOT6935 / ZOO4926

- AI in Genetics Fall 2024
- *ZOO6927 / BOT6935 / ZOO4926*
- Class Number *29890 / 29408 / 29411*
- Tuesday | (3:00 PM - 4:55 PM)
- Thursday | (3:00 PM - 3:50 PM)
- Room: FAC127

- Zoom link for remote students: https://ufl.zoom.us/j/6424255698
- Please attend in person if you are on the main campus

- Juannan Zhou, Assistant Professor
- Department of Biology
- E-mail: juannanzhou@ufl.edu
- Office: Bartram 122

- Office Hours: Thursday 4:00-5:00 PM

Communication

- **Course Slack channel:**
- Please send me your email with title “Slack - AI in Genetics”
- **Course site**: https://github.com/juannanzhou/AI-in-genetics

Course objective

- Comprehensive overview of applications of modern machine learning techniques in various areas of genetics.
- Provide opportunities for students to
  - integrate machine learning into their own research
  - learn critical computational and statistical skills that will hopefully broaden the student’s career path.

Deliberables

• Objectives of the course will be achieved if, by its conclusion, students can:

- Understand the basic concepts and mathematical/statistical theory behind modern machine learning methods
- Understand 80% of most research papers in fields relevant to the student's own research,
- Grasp the basic ideas of technical machine learning papers

Deliberables

- Develop new research questions well-suited for applying machine learning methods to improve their current studies; or identify existing questions where machine learning offers a potentially superior alternative to the current approaches.
- Identify the right machine learning frameworks and tools for answering these questions
- Build machine learning models to solve specific questions using coding languages such as Python
- Get the model to work by using different model architectures and training methods

Topics

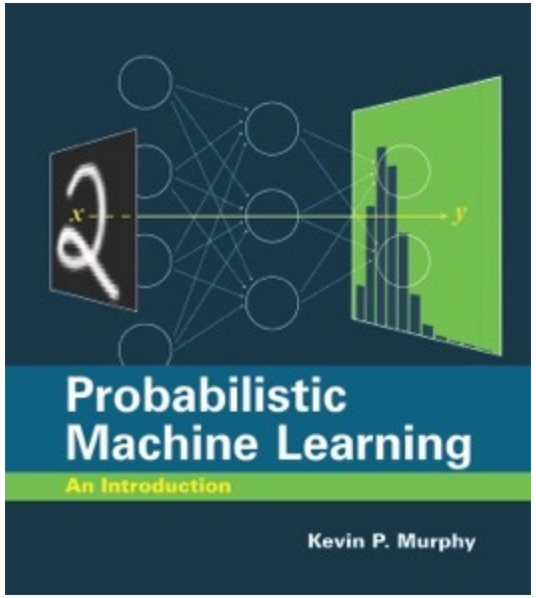
- The course will cover applications of AI to **genomics**, **gene expression and regulation**, **protein design and evolution**, **molecular evolution**, **disease/cancer genetics, population**, and **quantitative genetics**.

Course format

- Unit 1: Concise introduction to the mathematical and statistical foundations of modern machine learning.
- Unit 2: Student-led paper discussions on different areas of AI in genetics.

Textbook (recommended)

- *Probabilistic Machine Learning: An Introduction by Kevin Murphy. MIT Press, March 2022.*
- Free pdf: https://github.com/probml/pml-book/releases/latest/download/book1.pdf



Exam

- One mid-term exam. The grade of the mid-term exam will account for **40%** of the student’s final grade.
- Format of the exam will be take-home and consists of of conceptual and practical questions the student needs to solve using their preferred coding language.

Student-led paper discussion

- Each student will be responsible for leading an in-class discussion on one of the assigned readings.
- This will account for **10%** of final grade.
- Sign up for a paper from a curated list
- Or nominate a paper!

- Sign up for/nominate a paper here
- https://docs.google.com/spreadsheets/d/1Auv1KecDHTh7p3GDAcKObnJc-mv3aDeWkuo2J2ONzec/edit?usp=sharing

Final project

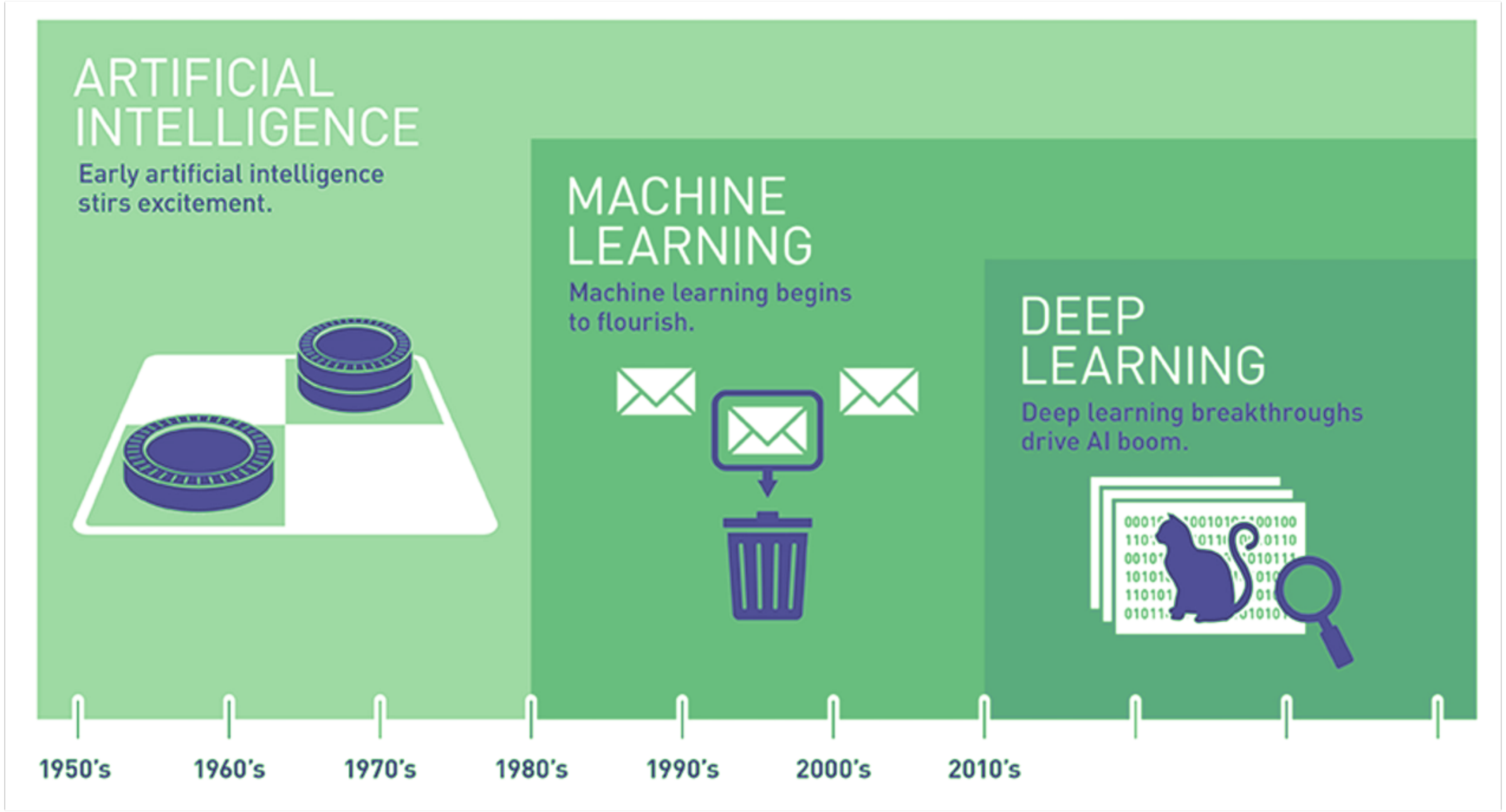
- Each student is expected to complete a final project, which will make up **50%** of the student's final grade.
- Students should apply machine learning techniques to solve a biological problem, preferably directly relevant to their thesis work.
- The student will present their results in class and turn in a term paper. The paper should be similar in format to a conference publication (e.g. https://proceedings.mlr.press/v240/).

Final project

•Student's prior fluency in coding/ML and stage in their training will be taken into account
•The most important goal is to get you started working on a project that will lead to a publicatioin

Date	Week	Subject	Topics
8/22/24	0	Course introduction; Mathematical foundation	Linear algebra
8/27/24	1	Mathematical foundation	Linear algebra
8/29/24	1	Mathematical foundation	Probability
9/3/24	2	Mathematical foundation	Probability
9/5/24	2	Machine learning basics	Multi linear perceptron; Backprop; Autodiff; Gradient descent
9/10/24	3	Machine learning basics	Training neural networks; Regularization
9/12/24	3	Machine learning basics	Convolutional neural networks
9/17/24	4	Machine learning basics	Language models; RNNs; Transformers
9/19/24	4	Machine learning basics	Graphical neural networks; Generative models
9/24/24	5	Machine learning basics	Generative models
9/26/24	5	Machine learning basics	Non-parametric methods; Gaussian processes
10/1/24	6	Paper discussion	Bioinformatics
10/3/24	6	Paper discussion	Proteins
10/8/24	7	Paper discussion	Proteins
10/10/24	7	Paper discussion	Gene expression and regulation
10/15/24	8	Paper discussion	Gene expression and regulation
10/17/24	8	Paper discussion	Genomics
10/22/24	9	Paper discussion	Genomics
10/24/24	9	Paper discussion	Molecular Evolution
10/29/24	10	Paper discussion	Molecular Evolution
10/31/24	10	Paper discussion	Population Genetics
11/5/24	11	Paper discussion	Population Genetics
11/7/24	11	Paper discussion	Quantitative Genetics; Plant/Animal Breeding
11/12/24	12	Paper discussion	Quantitative Genetics; Human diseases
11/14/24	12	Paper discussion	Generative models in genetics
11/19/24	13	Student presentations	
11/21/24	13	Student presentations	
11/26/24	14	Thanksgiving break	
11/28/24	14	Thanksgiving break	
12/3/24	15	Student presentations	

AI is Expansive



Source: Copeland, M. Retrieve from https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

ML / DL Comparison

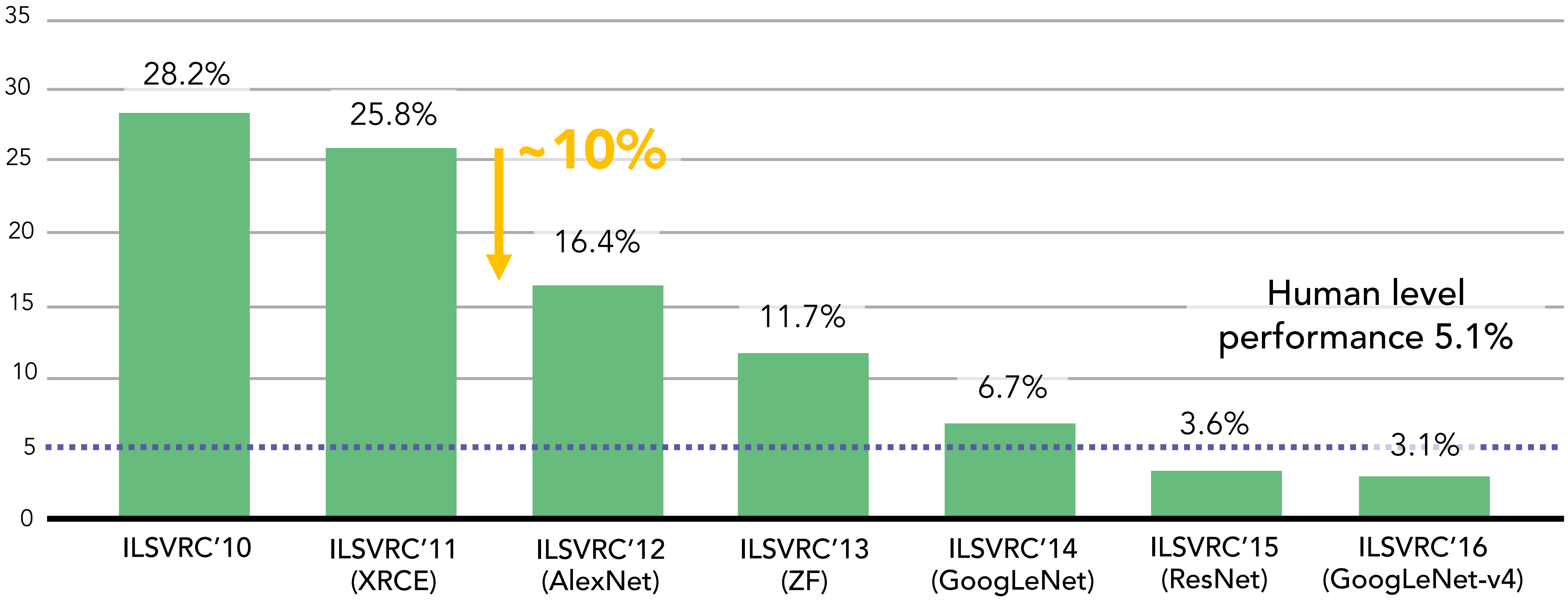
Machine Learning

- 1990's - Present
- Statistics & Math
- Sci-Kit Learn / RAPIDS

Deep Learning

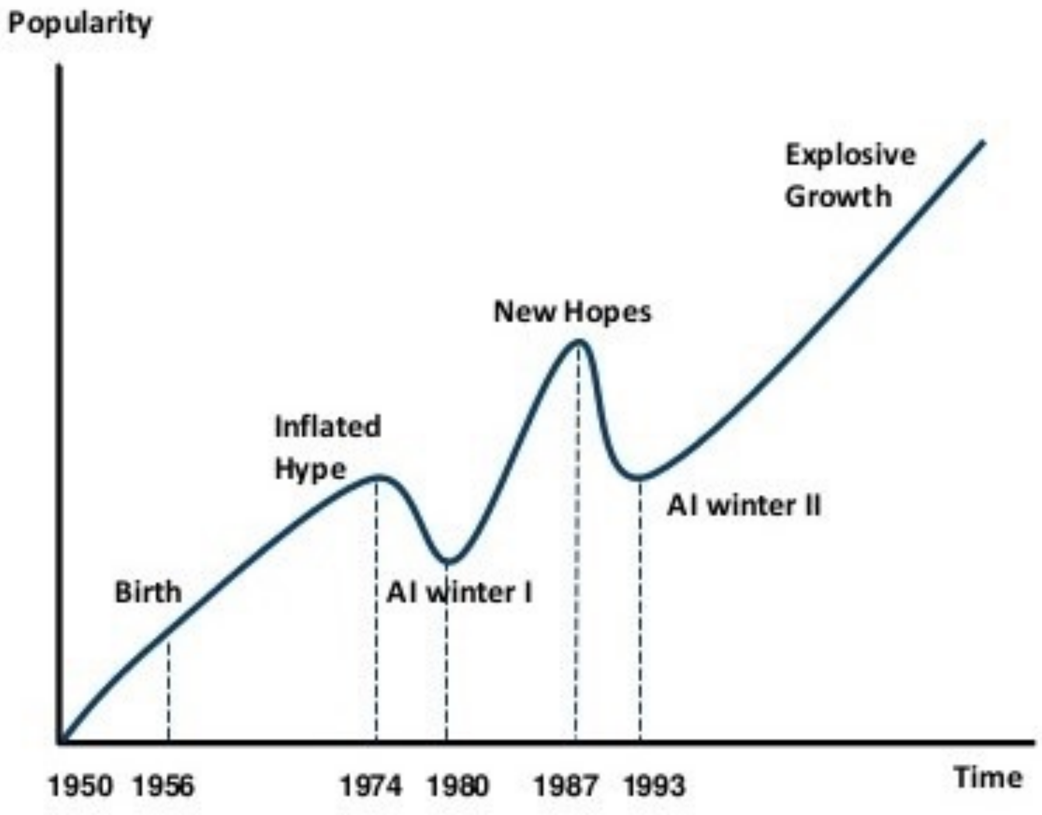
- 2010 - Present
- Artificial Neuron
- Tensorflow / Pytorch

Deep Learning Improved Image Classification 10% in 1-year



Source: https://becominghuman.ai/a-summary-of-industry-ready-state-of-the-art-computer-vision-techniques-a7f2b893de2f

AI Has a Long History of Being “The Next Big Thing”…



Timeline of AI Development

- **1950s-1960s**: First AI boom - the age of reasoning, prototype AI developed
- **1970s**: AI winter I
- **1980s-1990s**: Second AI boom: the age of Knowledge representation (appearance of expert systems capable of reproducing human decision-making)
- **1990s**: AI winter II
- **1997**: Deep Blue beats Gary Kasparov
- **2006**: University of Toronto develops Deep Learning
- **2011**: IBM’s Watson won Jeopardy
- **2016**: Go software based on Deep Learning beats world’s champions

Image Credit: https://www.mdpi.com/2079-9292/10/4/514/htm

Three major paradigm shifts: Data, Genomes, AI

Hypothesis-driven research:
Formulate hypothesis → gather data
Lots of thinking before → target study
Problem: Highly biased, little novelty

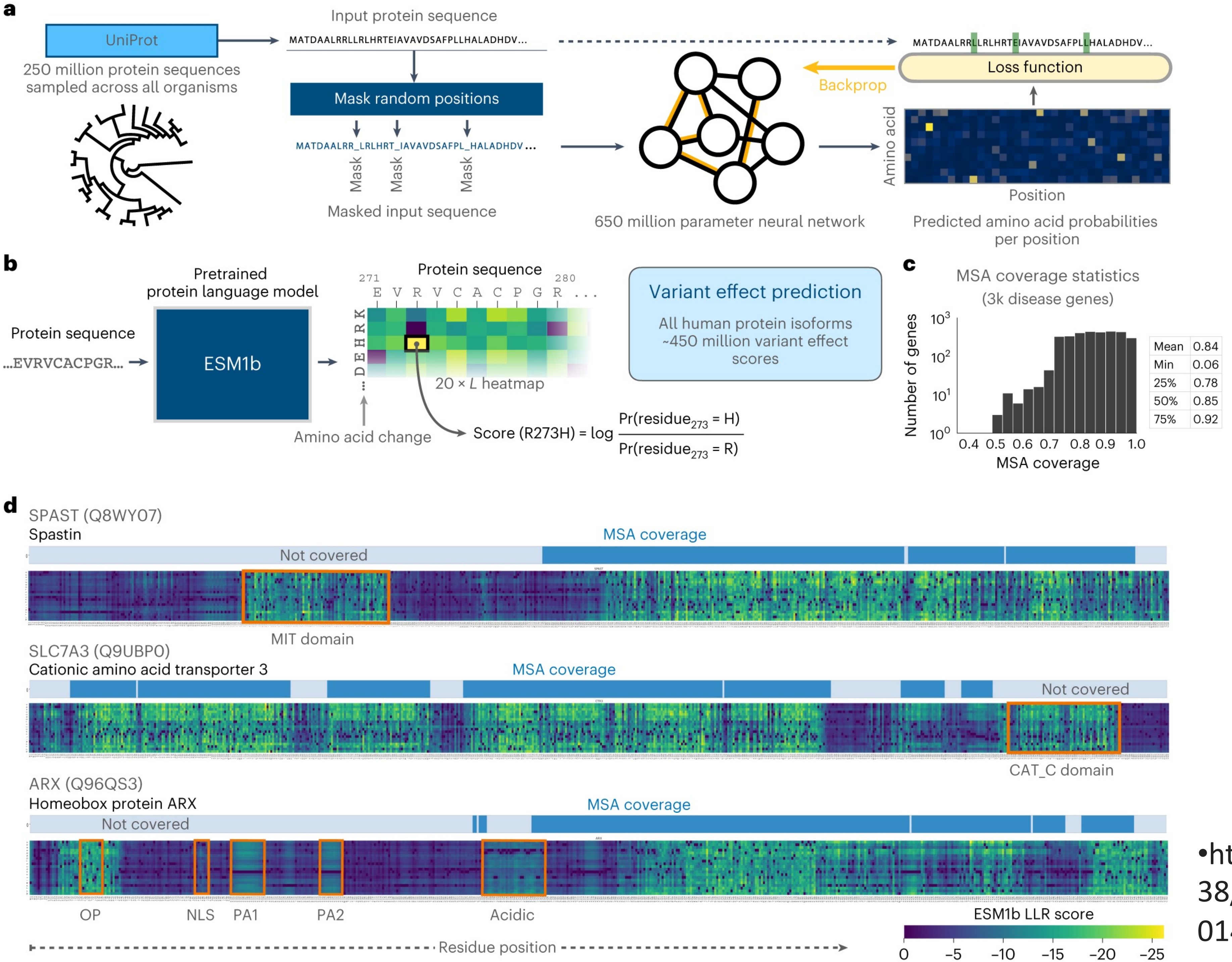
Data-driven research:
Gather data → Ask questions later
Systematic datasets, build resources,
massive data sharing, comprehensive

Correlation-based analysis:
More Coffee ⇔ Better Health
More Chocolate ⇔ More Nobel Prizes
'Epidemiology' all about correlations

Genetics provides causality:
Genetic variants → Disease outcome
Polygenic risk score → Causal factors
Perturbation experiments → Confirm

Classical Data Analysis:
New methodology for each problem
Human scientist does all the 'thinking'
Few parameters, targeted models

Generative AI+Deep Learning
Foundation models, Multi-Modality
Representation learning, hierarchical
Truly 'understand' concepts → insights



•https://doi.org/10.1038/s41588-023-01465-0

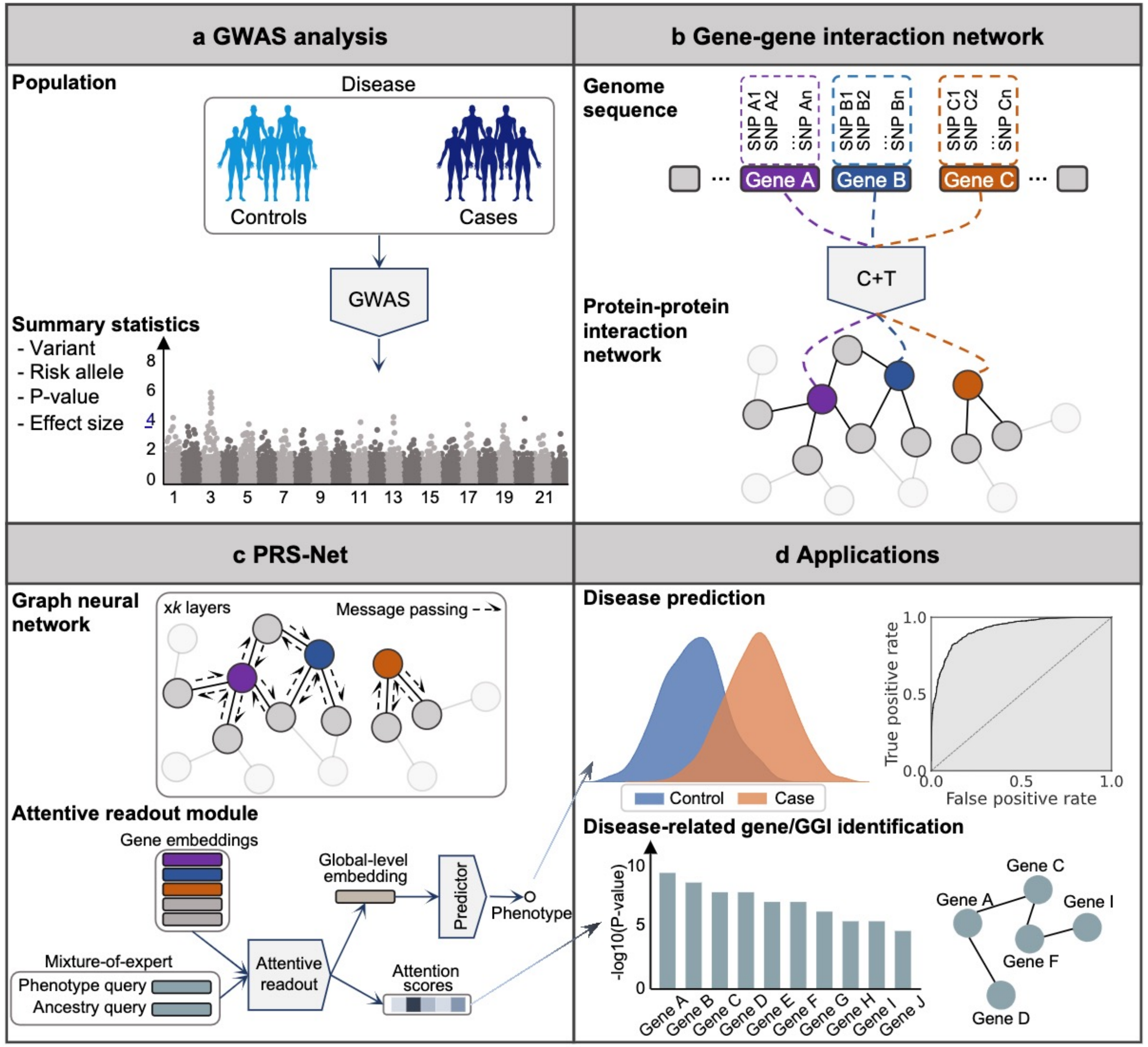


Fig. 1: An illustrative diagram of PRS-Net. **a** The proposed framework is based on summary statistics, including variants, risk alleles, P-values, and effect sizes derived from GWAS. **b** A gene-gene interaction network is constructed based on the protein-protein interaction network. Gene-level PRSs are calculated with the C+T method to serve as the node features for the nodes within the network. **c** A graph neural network is employed to update node features via message passing and subsequently an attentive readout module is applied to provide interpretable PRS predictions. **d** The PRS-Net can be applied for disease prediction and disease-related gene/GGI identification.