COLUMBIA UNIVERSITY

EECS 6895 Adv. Big Data and AI

Lecture 1: Introduction

Prof. Ching-Yung Lin

Columbia University

January 21st, 2025

Two Taiwanese were invited to throw the First Pitch in an MLB game in 2024:

Jensen Huang,
CEO, Nvidia, May 2024

Ching-Yung Lin,
CEO, Graphen, Aug 2024

Prof. Ching-Yung Lin

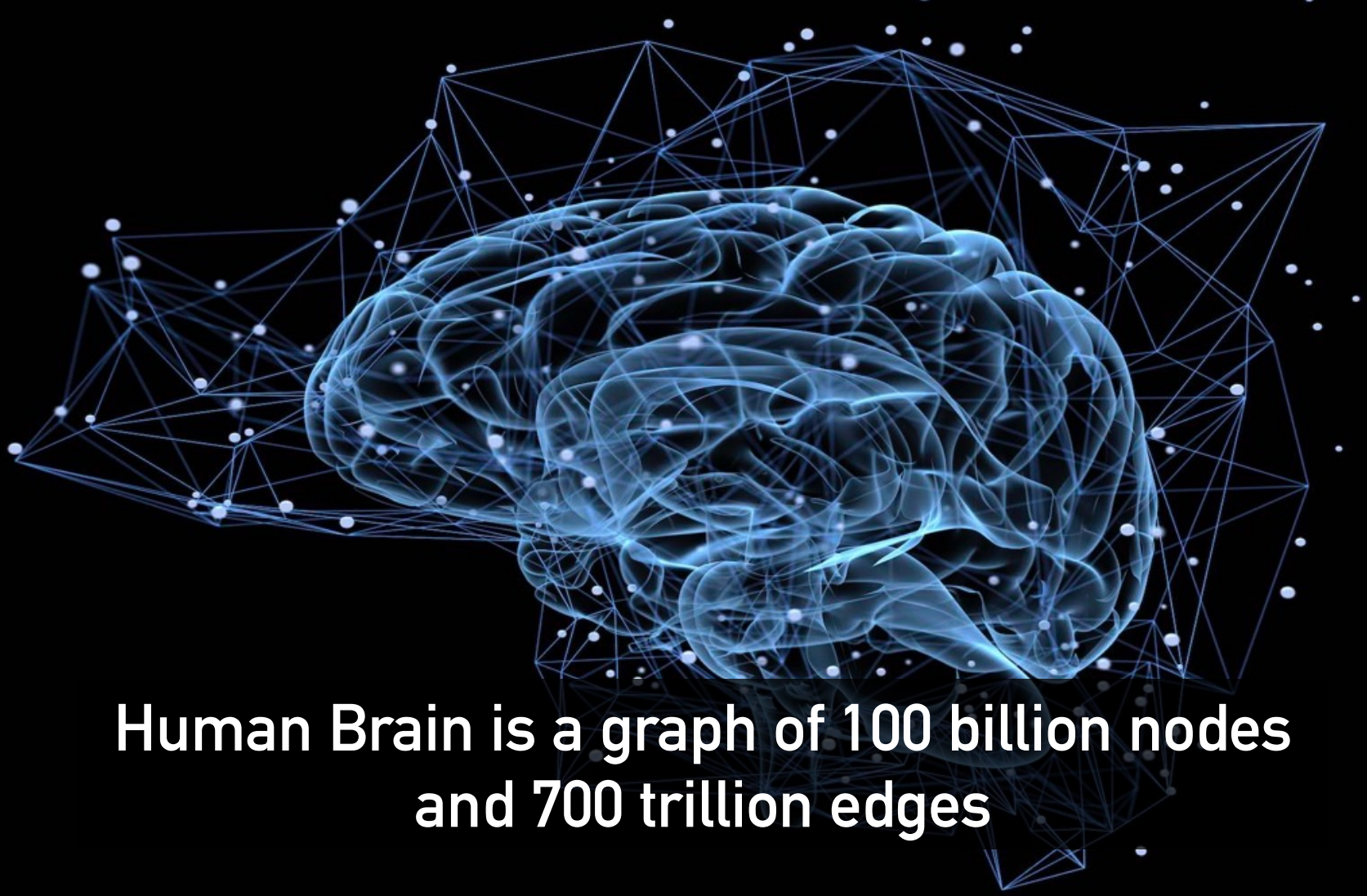
- BS & MS, EE, National Taiwan University; Ph.D., Columbia University
- Former IBM Chief Scientist**,** established the Network Science and Machine Intelligence department at IBM T. J. Watson Research Center.
- Adjunct Professor in Columbia Univ. (since 2005), U of Washington (2003-9), and NYU (2014).
- In 2010, IBM Exploratory Research Career Review selected Dr. Lin as one of the scientists "***most likely to have great scientific contributions to the IBM and the world***."
- In 2011, Dr. Lin was elected to be the first IEEE Fellow in the area of Network Science; IEEE Distinguished Lecturer 2014-2015.
- By 2013, the systems developed by Dr. Lin had been recognized of hundreds of millions contributions to IBM.
- Led large teams to build AI systems for US and other governments, and IBM for about 15 years.
- First Professor teaching Big Data Analytics in worldwide universities.
- **Invited by the American Medical Association President as a keynote speaker and panel together with the White House Chief Data Scientist, on the annual AMA convention 2015.**
- Published over 200 papers and owns 70+ patents. 57 papers cited more than 57 times. 13000+ citations. 7 best paper awards; Dr. Lin was once the cover stories of BusinessWeek and the World Journal Week.

Prof. Lin Led AI R&D in Practical Settings in Tens of Industries at IBM

1. Expertise Location
2. Recommendation
3. Commerce
4. Financial Analysis
5. Social Media Monitoring
6. Telco Customer Analysis
7. Healthcare Analysis
8. Data Exploration and Visualization
9. Personalized Search
10. Anomaly Detection
11. Fraud Detection
12. Cybersecurity
13. Sensor Monitoring (Smarter another Plane
14. Cellular Network Monitoring
15. Cloud Monitoring
16. Code Life Cycle Management
17. Traffic Navigation
18. Video Semantic Understanding
19. Genomic Medicine
20. Brain Network Analysis
21. Data Curation
22. Near Earth Object Analysis

- One of Less than 5 Chief Scientists in Worldwide IBM Corporation of 400K+ employees.
- By 2013, already contributed more than hundreds of millions of dollars in AI for IBM

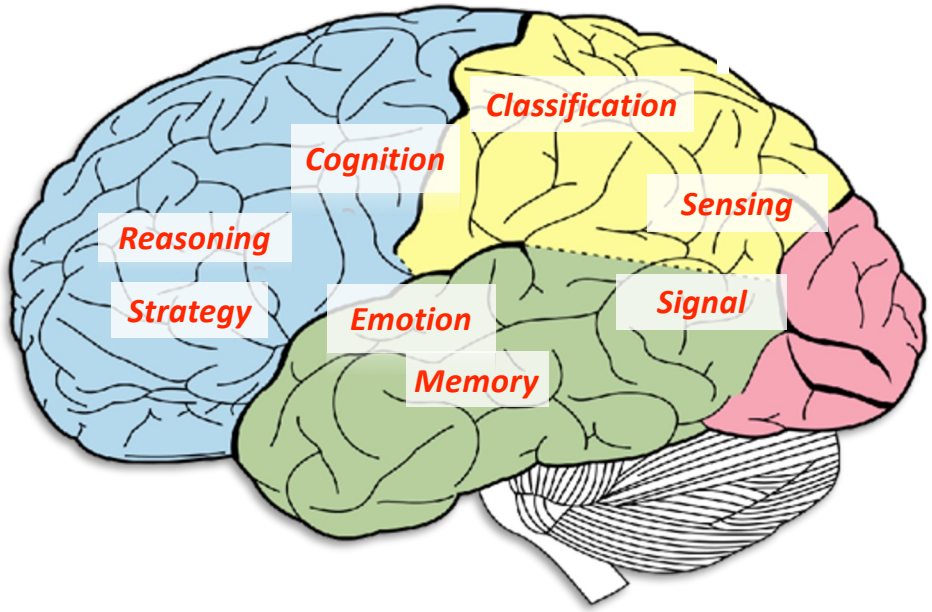
- Advised over 1000 big data and artificial intelligence projects at Worldwide Enterprises, IBM organization, Universities, Research Institutes, etc.
- Ranked No.1 in the search of Big Data Analytics on Baidu. (2015-17)
- Invited to be a keynote speaker at AMA (2015), FBI Cybersecurity (2016), and at various world's prestigious institutes, including Pentagon (twice), Fed Reserve, FINRA, European Central Bank, etc.

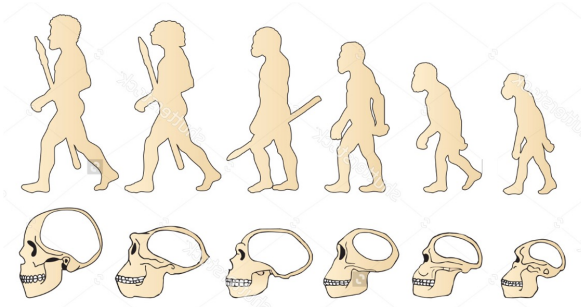


Human Brain is a graph of 100 billion nodes and 700 trillion edges

Evolution of Intelligence

Direction of the Evolution of Intelligence





Evolution of Artificial Intelligence

Columbia University

Direction of the Evolution of Artificial Intelligence

2020s? 2030s? 2020s 2010s 2000s 1990s 1960s - 1980s

Learning
- Machine Learning and Deep Learning
- Autonomous Learning

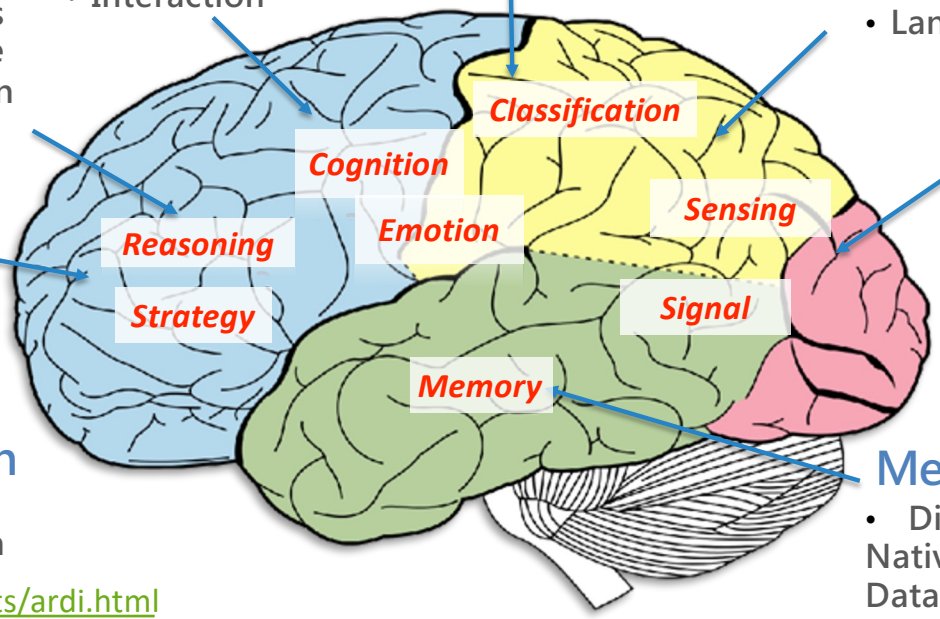
Expression
- Big Data Visualization
- Explainability
- Interaction

Reasoning
- Bayesian Networks
- Causality Inference
- Behavior Prediction

Understanding
- Deep Vision Understanding
- Language Understanding

Action
- Control
- Behavior

Perception
- Network Analysis
- Feature Analysis



Strategy
- Game Theory
- Optimal Solution

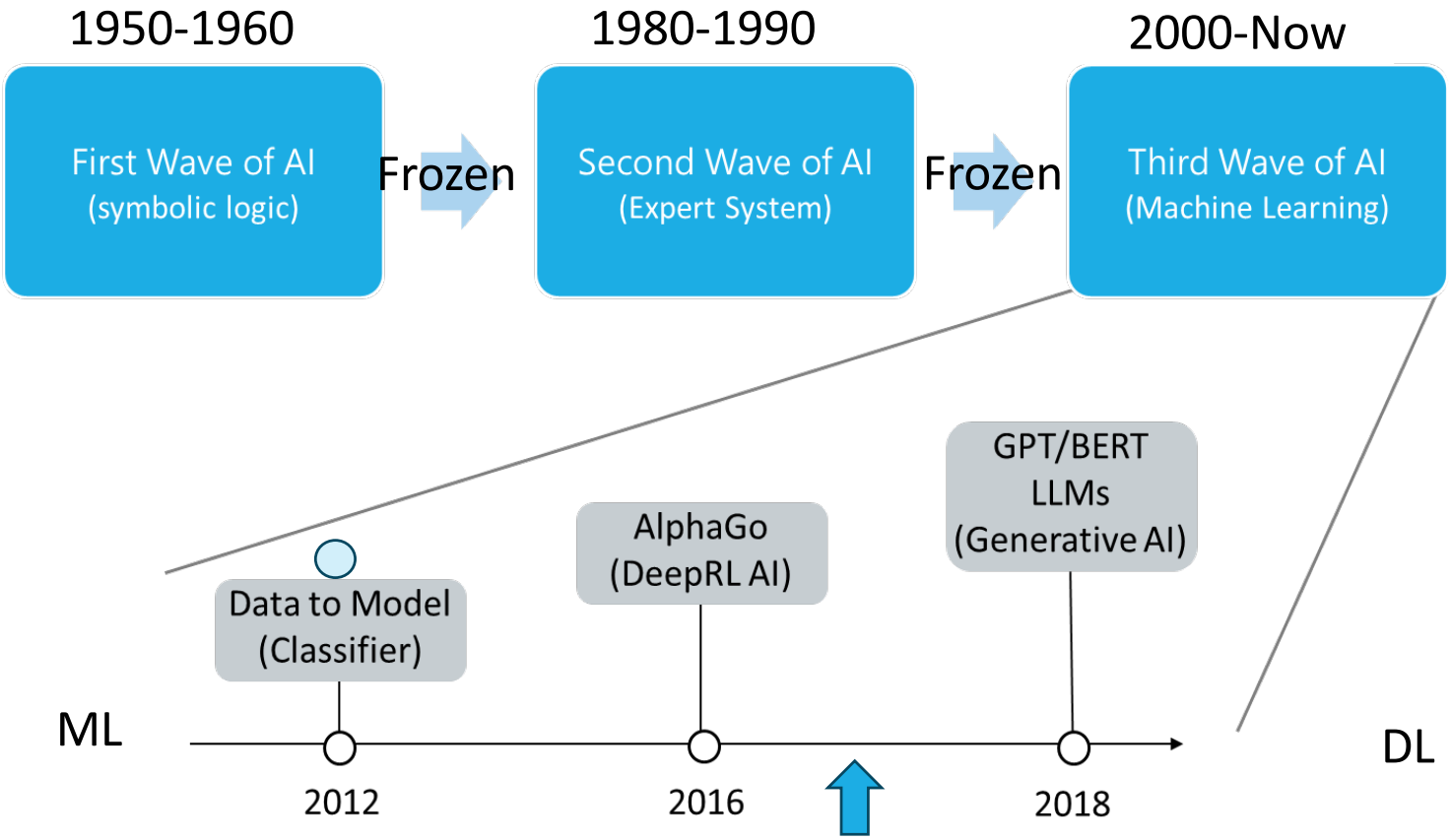
Pipeline
- Coordination
- Flow

Personality
- Humanization
- Ethics

Emotion
- Feeling
- Affection

Memory
- Distributed Native Graph Database

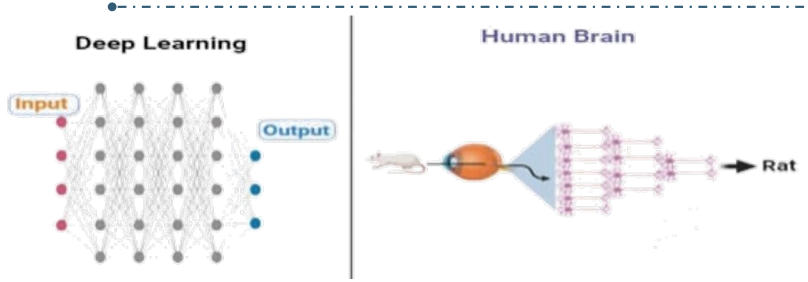
https://www.graphen.ai/products/ardi.html



Attention is all you need, Ashish Vaswani,Noam Shazeer,Niki Parmar,Jakob Uszkoreit,Llion Jones,Aidan N. Gomez,Lukasz Kaiser,Illia Polosukhin,NIPS(**2017**)

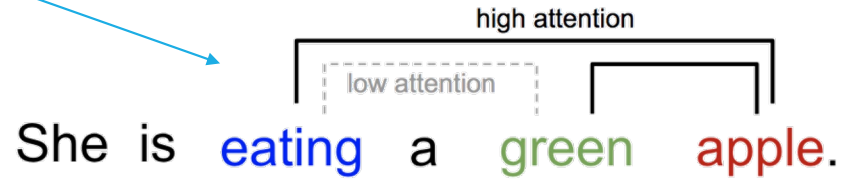
Deep Learning makes AI Model Evolution

1. ML-Boosting model
   Tree model
2. Multi-layer Perception
   2013
3. CNN/RNN/GNN
   image/voice/Topology
   2018
4. Transformer
   Self-Attention
   2020
5. LLM Generative model
   2024
6. Multi-model



High dimension projection

Local grid featurization



Causal association

Graphen Core -- Full-Brain AI Platform

- **20 years of AI leadership in tens of industries**

Ardi AI Platform

Direction of Evolution of Intelligence

Learning
- Machine Learning and Deep Learning
- Autonomous Learning

Expression
- Big Data Visualization
- Explainability
- Interaction

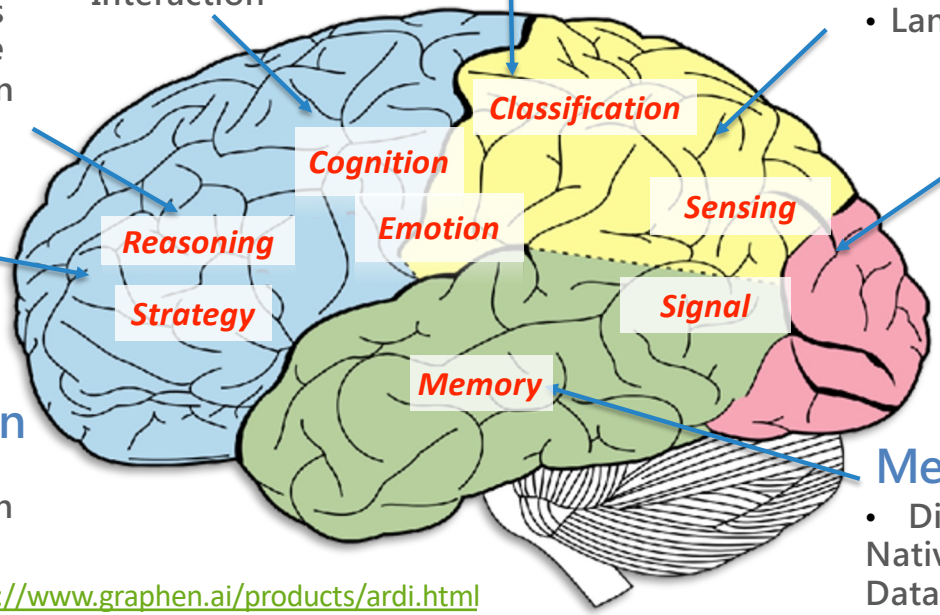
Understanding
- Deep Vision Understanding
- Language Understanding

Reasoning
- Bayesian Networks
- Causality Inference
- Behavior Prediction

Action
- Control
- Behavior

Perception
- Network Analysis
- Feature Analysis

Strategy
- Game Theory
- Optimal Solution



Pipeline
- Coordination
- Flow

Personality
- Humanization
- Ethics

Emotion
- Feeling
- Affection

Memory
- Distributed Native Graph Database

https://www.graphen.ai/products/ardi.html

Advanced Enterprise Full-Brain AI Platform to build solutions – Scalability, Stability, and Advanced AI Technologies

Graphen Ardi Machine Understand and Feel -- #1 evaluated by NIST

- Visual Recognition
- Speech Recognition
- Knowledge Graph
- Face Recognition
- Emotion Recognition
- Speaker Identification
- Relationship Inference
- Event and Action Understanding

US National Institute of Standards and Technology (NIST) Deep Video Understanding Grand Challenge benchmark:

#2 in 2020

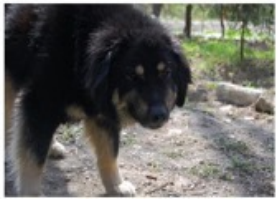
#1 in 2021

#1 in 2022

Psychology Emotion Wheel

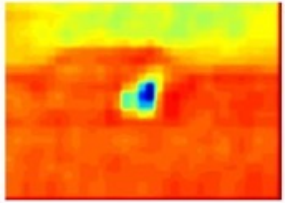
- Video Understanding:
  - Objects:
    - Visual Objects:, Tree, Person, Hands, …
    - Audio Objects: Music, Speech, Sound, …
  - Scenes:
    - Background: Building, Outdoors, Sky
  - Relationships:
    - The (time, spatial) relationships between objects & scenes
  - Activities:
    - Holding Hand in Hand, Looking for Stars

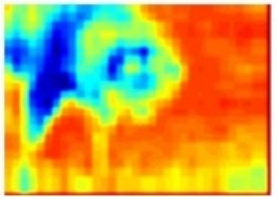
Abandoned building

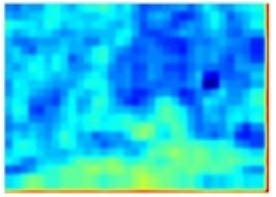


Scary dog

Colorful landscape



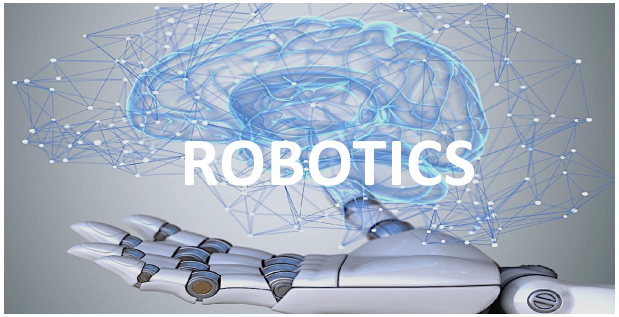




Graphen Group Mission

Reducing Risk for Better Living

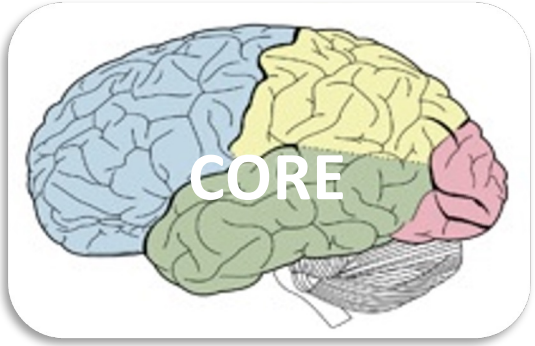
Understanding Life to its Most Detailed Scale; **Creating Unlimited Drugs to Save Lives**



Creating Unlimited Virtual Human to Assist Life

Advancing State-of-the-Art AI for the Well-Being of Mankind
– Graphen 2017

Graphen is Ranked World's Top #9 AI Company – Oct 2023 (OpenAI #1, DeepMind #3, Mistral #4, ..; by Tracxn)

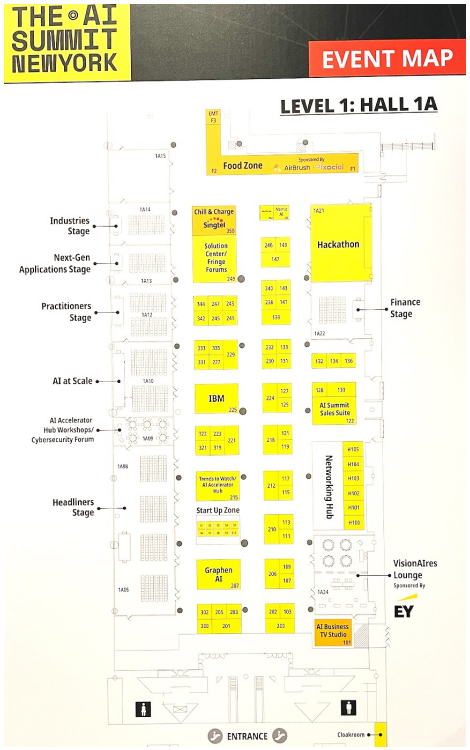


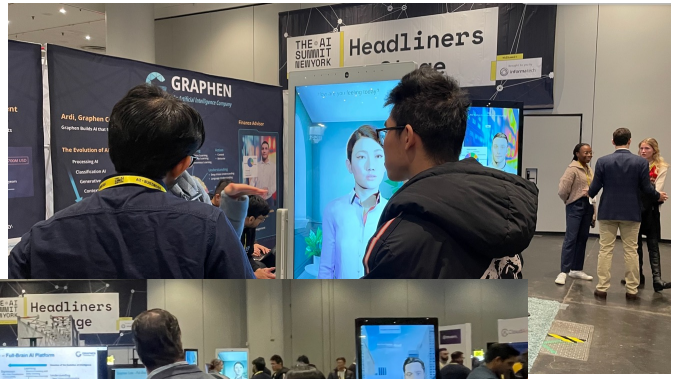
Graphen Ardi

HQ: 500 Fifth Ave., New York

Global Subsidiaries: Taipei, Tokyo, Singapore, HK, and Beijing
Established: 2017

Marching towards New York's #1 AI Company





Largest Booth in AI Summit NY 2023

Mission of the Class

- Advanced A.I. Technologies
- Advanced A.I. Assistants for Financial Industry and Healthcare Industry
- Advanced A.I. for Bio Science

Do we need AI?

Who will be our caregiver?

“Single Child”
Finalist of the 26th National Photo Contest, China

Where to find Helps?

All Developed and some developing countries have been facing labor shortage crisis ➔ More and more serious everyday.

Graphen Robotics

Meet Aiia

World's First A.I. Digital Human for Daily Life!!

- Hardware-Software Integrated Local AI 'Brain'.
- Privacy / Individual / Personal
- Speaks English, Chinese, Japanese, and Spanish
- Avatars with Personality & Emotion
- Eye Contact / Facial Expression
- Integrating with Payment, Mobile Apps, etc.

Six Aiia demos at New York Convention Center (April 2023 @ NY Auto Show)

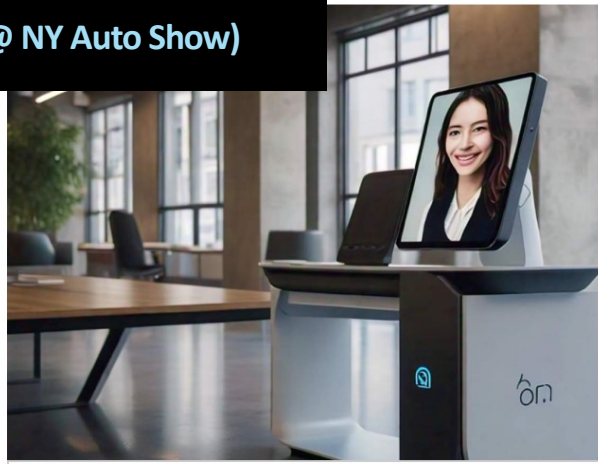
Concierge Aiia
Hotel, Train Stations, Travel Agent

Cashier Aiia
Drinks, Restaurants, Supermarket, etc.

Sales Aiia
Retail stores

Nurse Aiia
Hospital, Nursing Home

Office Assistant Aiia
Financial Institutes

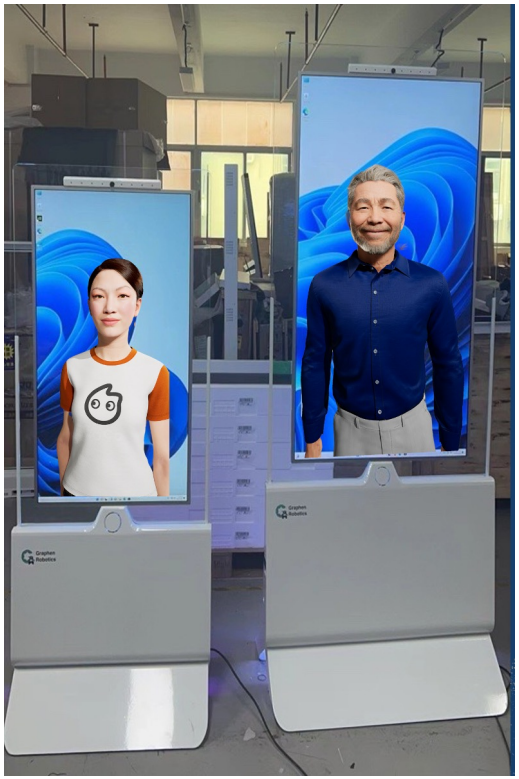


Customer Service Aiia
Automotives

Graphen Robotics Hardware

Columbia University

Aiia Glass

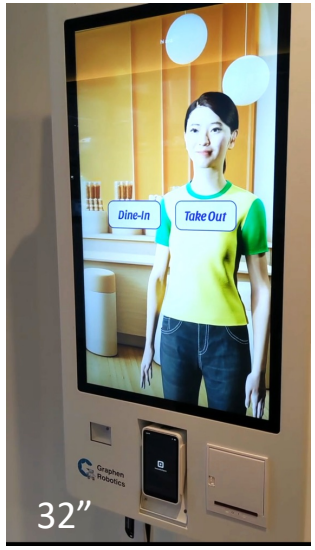


55” & 43”

Aiia Classic

55” & 43”

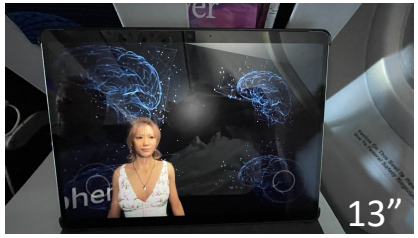
Aiia Kiosk



Aiia Robot

32”

Aiia Tablet



Aiia Hologram

75” & 86”

AI as Knowledge Worker

Examples：

Instant reference tool for medication dosages, side effects, and interactions, reducing the risk of medication errors.

Patient education： helping nurses provide accurate, understandable explanations of medical conditions and treatments.

Question：What is the infusion time for 1 unit of Packed Red Blood Cells?

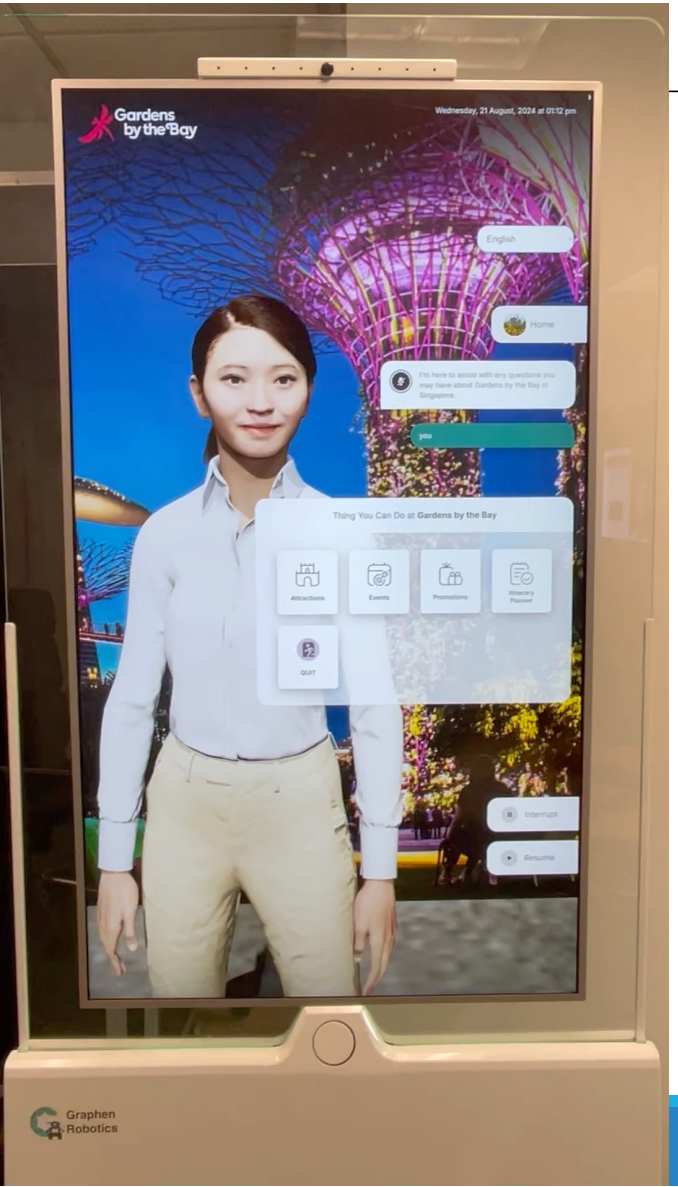
Aiia Nurse Assistant: PBRCs are a blood product used to replace erythrocytes; infusion time for 1 unit is usually between 2 and 4 hours.

Source: The answer is obtained by retrieving page 158 in the provided PDF, which is the RN Exam textbook.

➔ Aiia answered 90% questions correctly in New York RN License Exam

Retail https://www.youtube.com/watch?v=bcb9CT-DV4o


RDER HERE
Graphen Aiia Kiosk
Voice-Ordering Service
GRAPHEN



Aiia Examples

PoC Phase 2 --
Garden by the Bay
@ Singapore

Upcoming -- JR East
@Tokyo

Medical Assistant
@ US and Taiwan

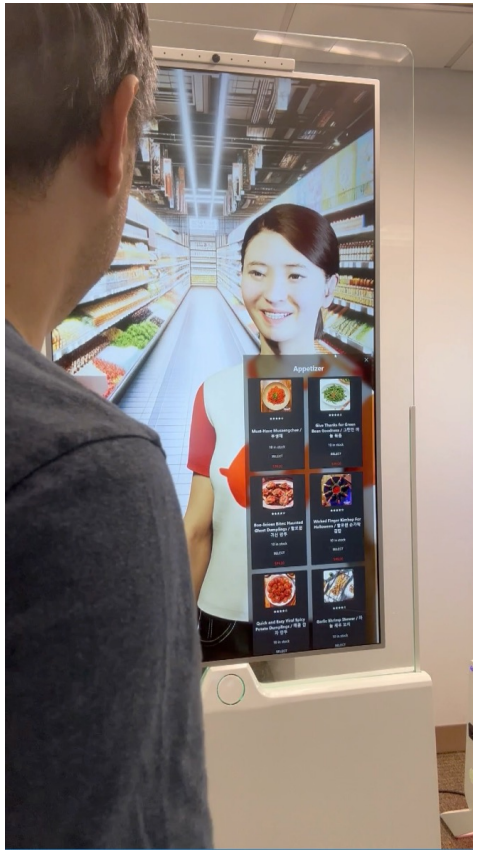
https://www.youtube.com/watch?v=J9zsaW0gDN4

Graphen Robotics

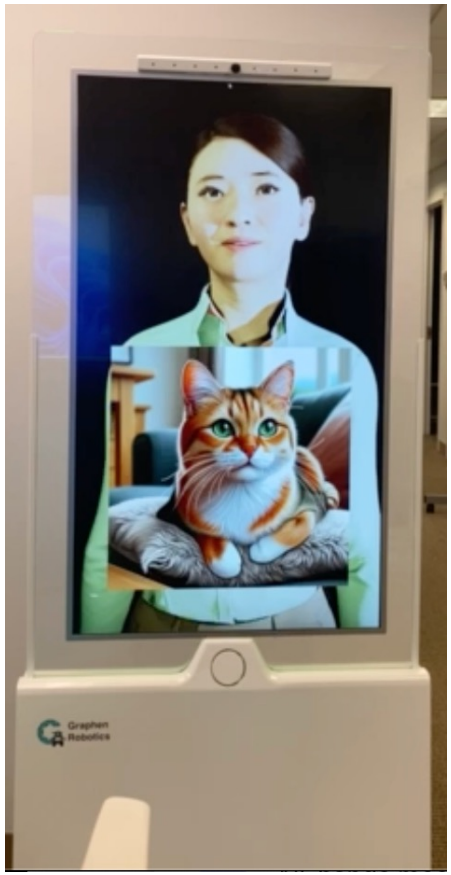
Aiia Examples

Columbia University

Retail Assistant



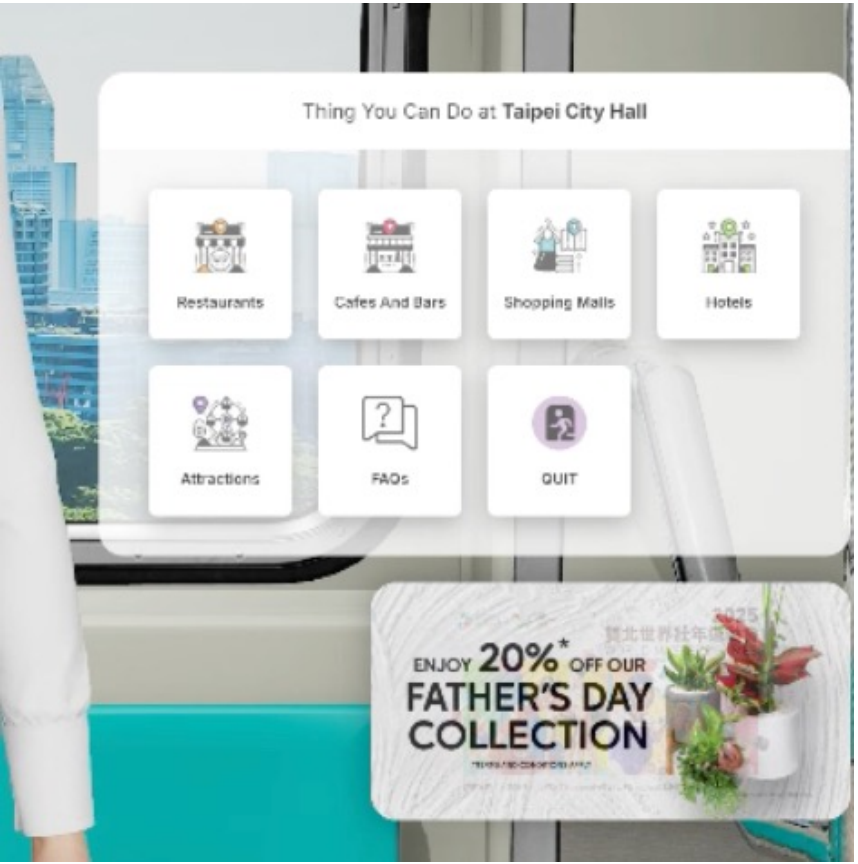
Office Assistant



Taipei Metro Use Case Demo

- 75” 3D Hologram or 43” / 55” Glass Kiosks.
- Local AI.
- Responds in 1.5 seconds.
- Any questions near the subway station.
- English, Mandarin, Japanese and Korean.
- Adding Malay & Indonesian.
- Eye Contact & interaction with the user.
- Facial Expressions.
- Actions.

Zoom In



Existing Graphen Aiia Deployments

New York

Tokyo

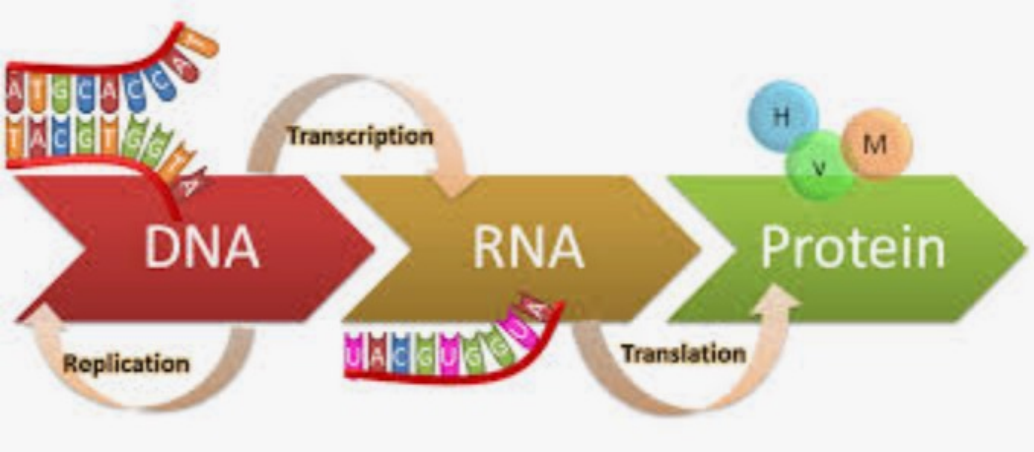
Taiwan

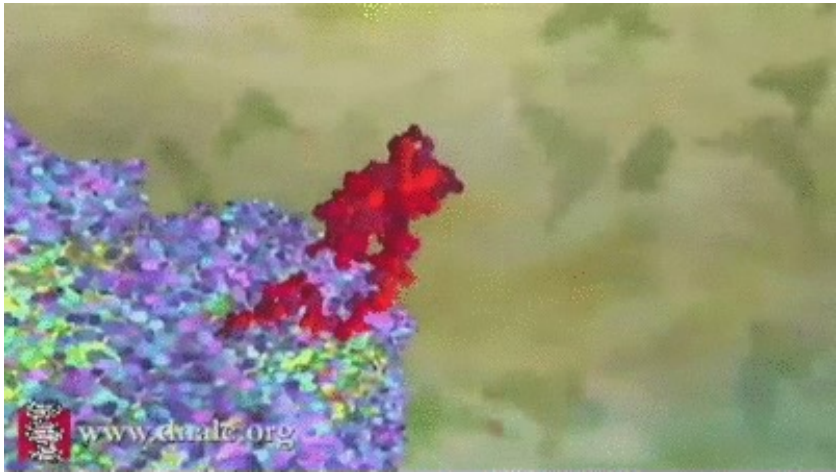
Singapore

ABIresearch for visionaries

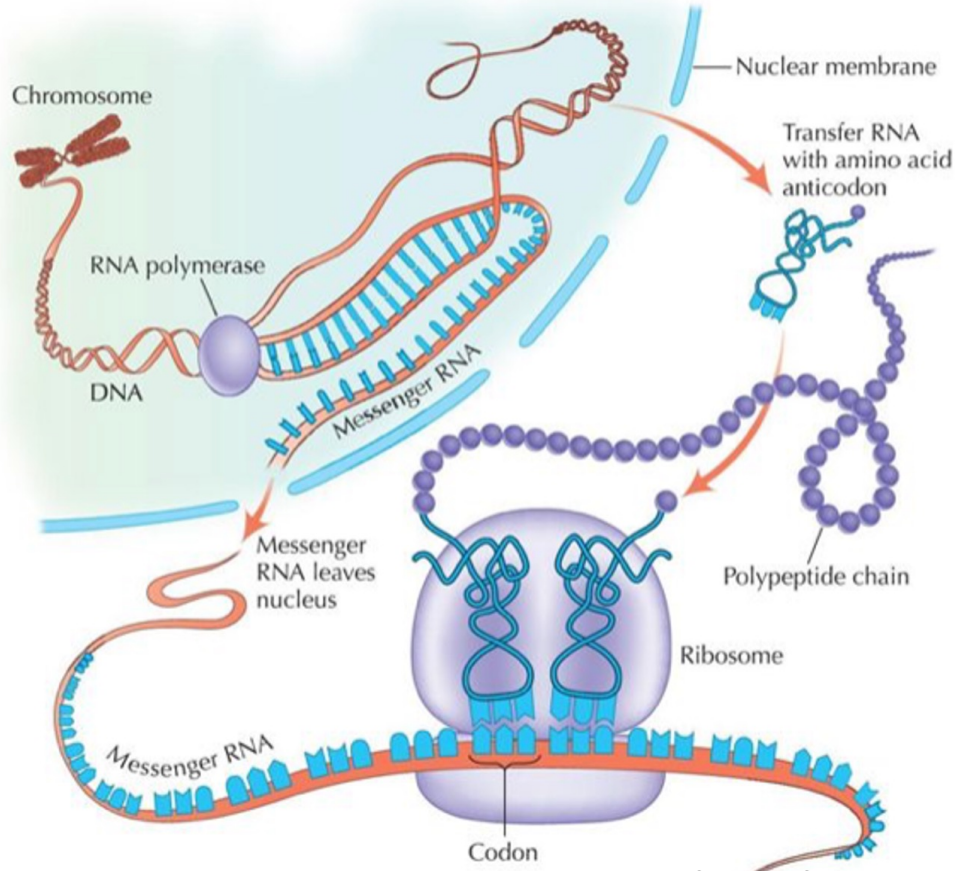
*"Tools from established companies like **Google** DeepMind, startups like **Graphen**, and AI chipsets from vendors like **NVIDIA** and **Intel** will help accelerate the speed of drug discovery, development, and testing, allowing pharmaceutical companies and healthcare authorities to combat the pandemic." – ABI research, May 2020*

Central Dogma of (Molecular) Biology





Forming Protein



Protein is the Foundation of All Life

A Protein == A Biological Machine Part

Folding from Amino Acids sequence to Protein

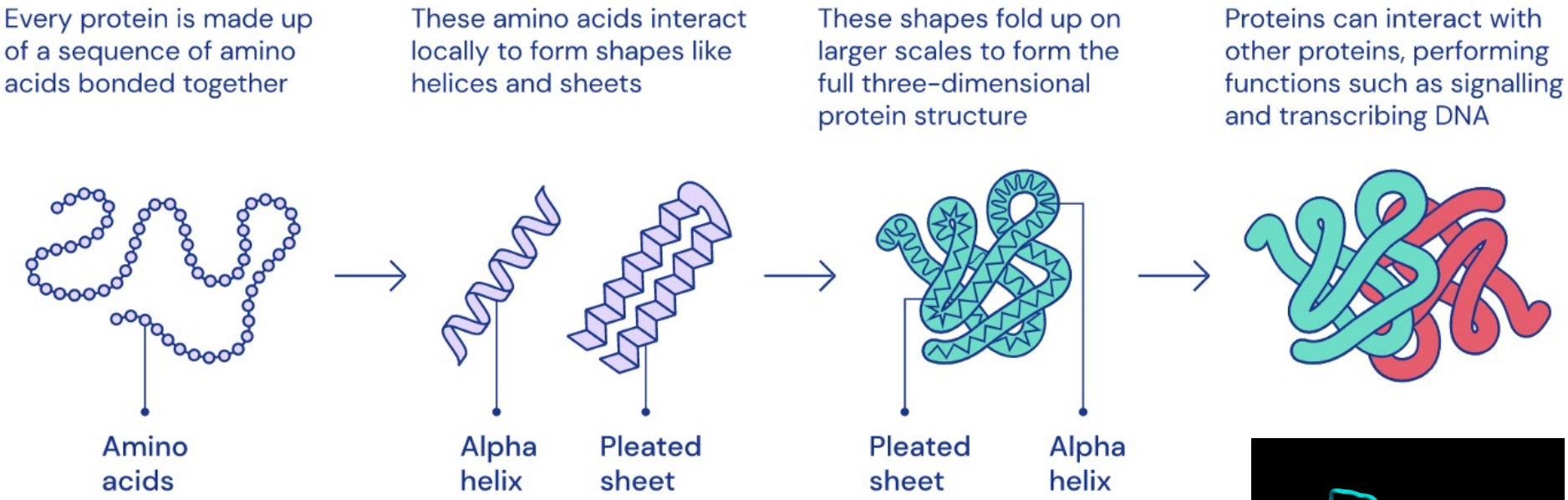
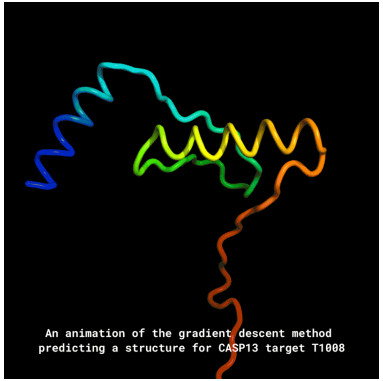


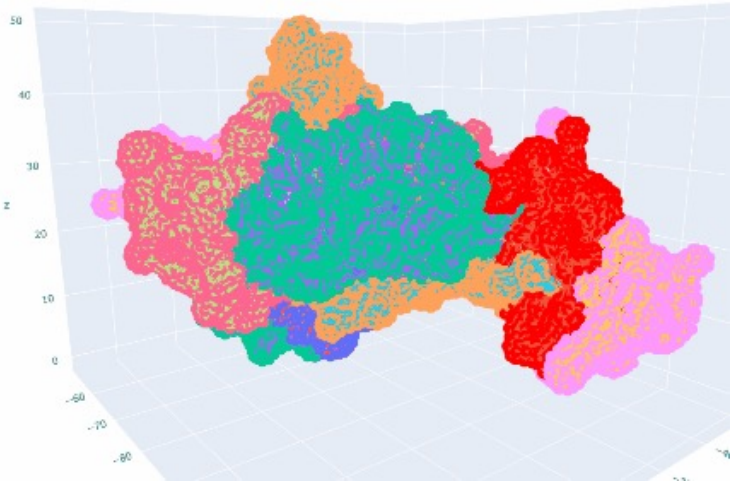
FIGURE 1: COMPLEX 3D SHAPES EMERGE FROM A STRING OF AMINO ACIDS.

Protein is the Foundation of All Life

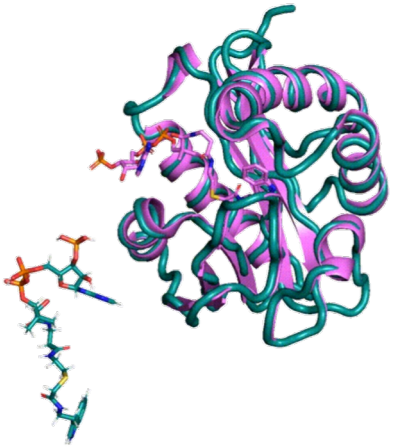
A Protein == A Biological Machine Part



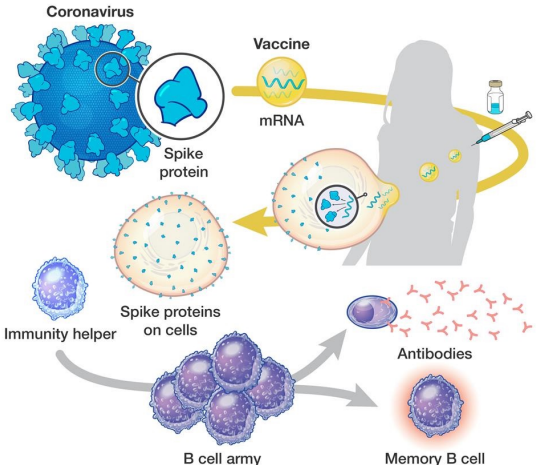
Protein is a graph of atoms. Drug is to make atom networks (compounds) that interact with Protein.



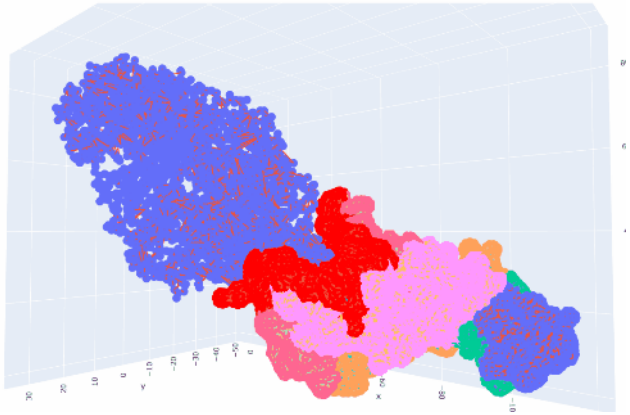
Protein is a graph of atoms



Drug-protein interaction

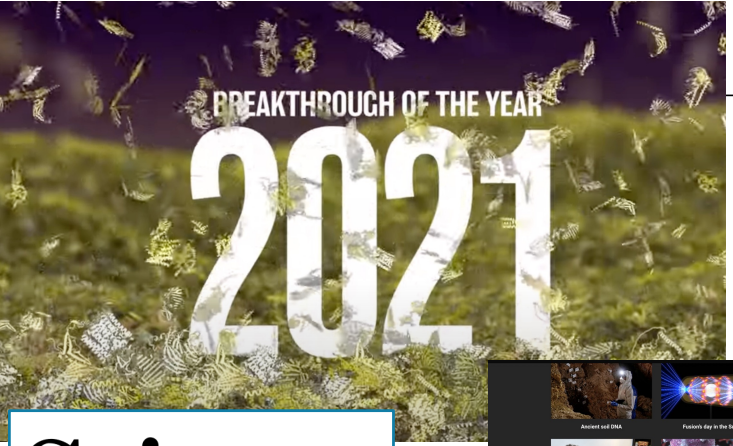


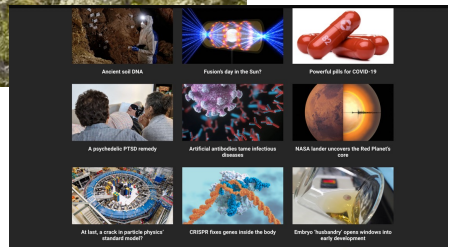
Antibody-protein interaction



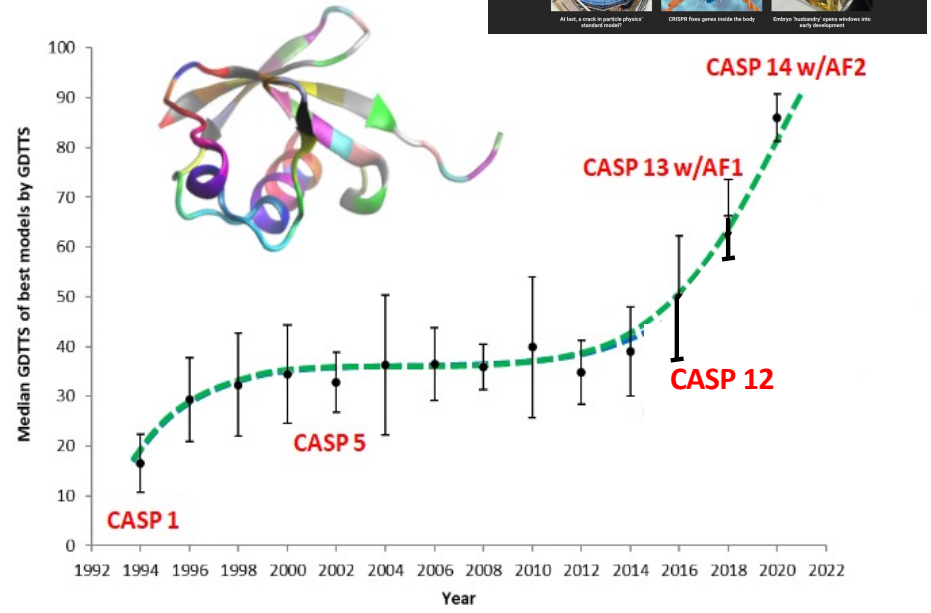
Do we know how a protein looks like?

➔ *No, we, the humans, almost did not.*





https://youtu.be/iUMpm3tYsVE?t=218



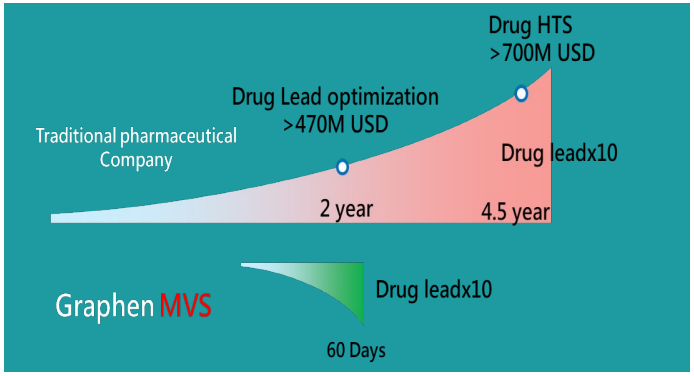
When people in 50 years later look back history, they may recognize 'Today' is a scientific breakthrough moment in Biology as an equivalent of Newton's moment in Physics

-- Graphen 2023

Graphen's Ann-in-One System makes significant difference ➔ 1/27 of the Time, 1/1335 of the cost, Potential "ultimate" New Drugs

Columbia University

1. Generate Synthesis and low side effects Drug
2. Filtering drug by ADMET and Solubility
3. Filtering Drug by Graphen QF Energy model and Kinase model
4. Disease-Clinical/Cell Target mapping (Spectrum Mutant Prediction)
5. Synthesis and Wet Lab Evaluation



	Samples	Required Time
Generate de novo new compound	1.5M → 30K	7 Days
Manufacturable and Safety Screening	2.5K	3 Days
Dynamic Target-Lead interaction simulation	0.1K	5 Days
Target Cell and Disease Finding	30	5 Days
Chemical Synthesis Kinase Assay/Cell Assay	10	30+30 Days

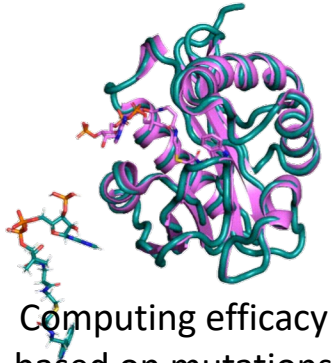
New Design & Development of:

- **Small Molecular Drugs**
- **Antibody Drugs**
- **Nucleotide Drugs**
- **Rare Disease (orphan drugs) and Personalized Drugs**

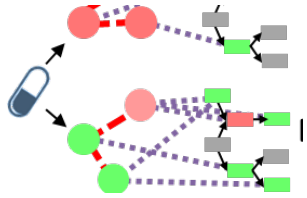
Graphen is creating 2-3 effective drug candidates per month

Graphen has established its own computer clusters for drug discovery

High Efficacy. Low Side-Effects. No Toxicity

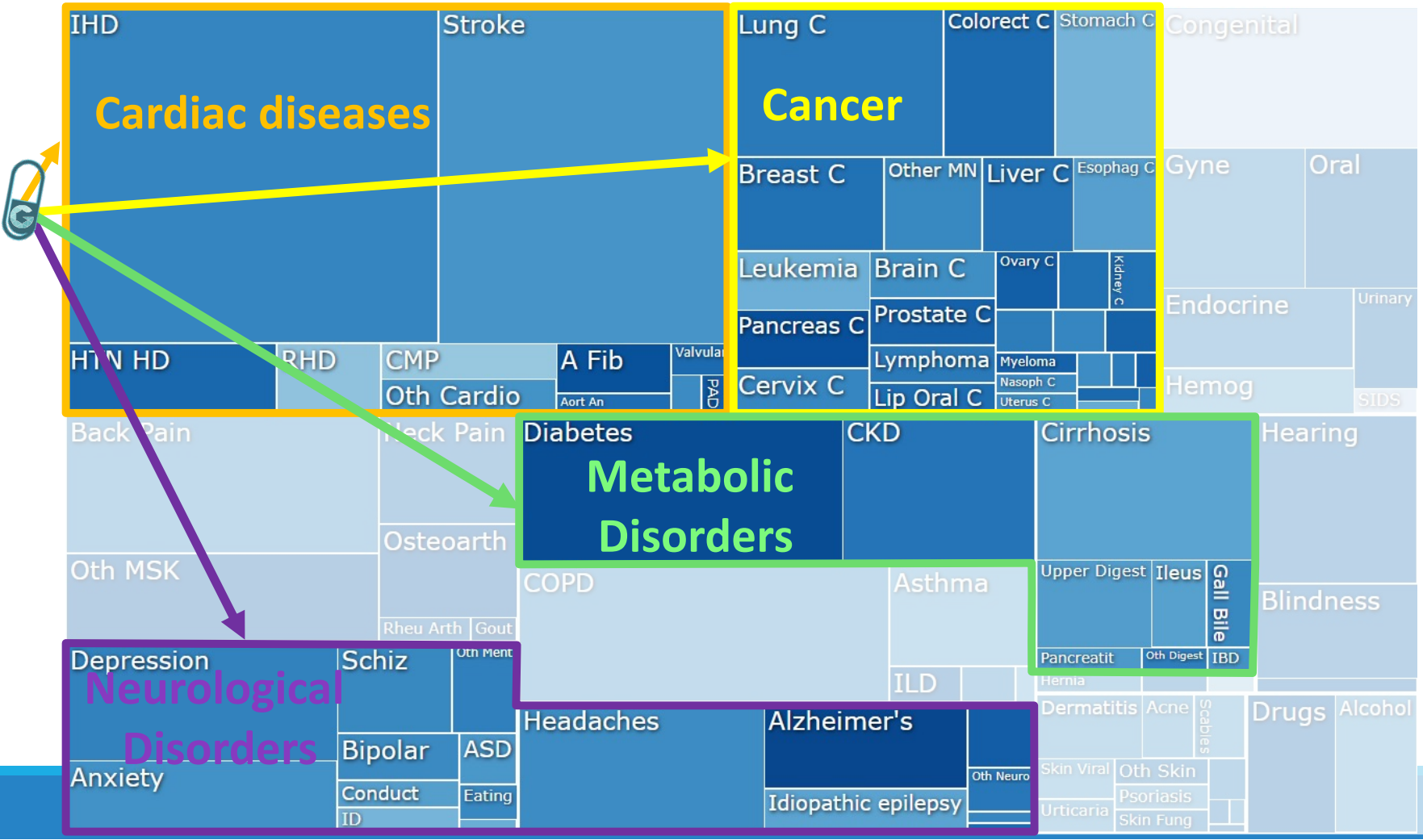


Computing efficacy based on mutations

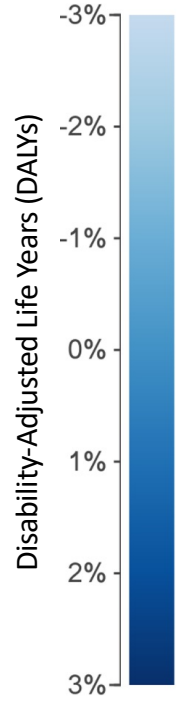


Kinase functional grouping

Designing new drugs to cover around 70% of disease fields

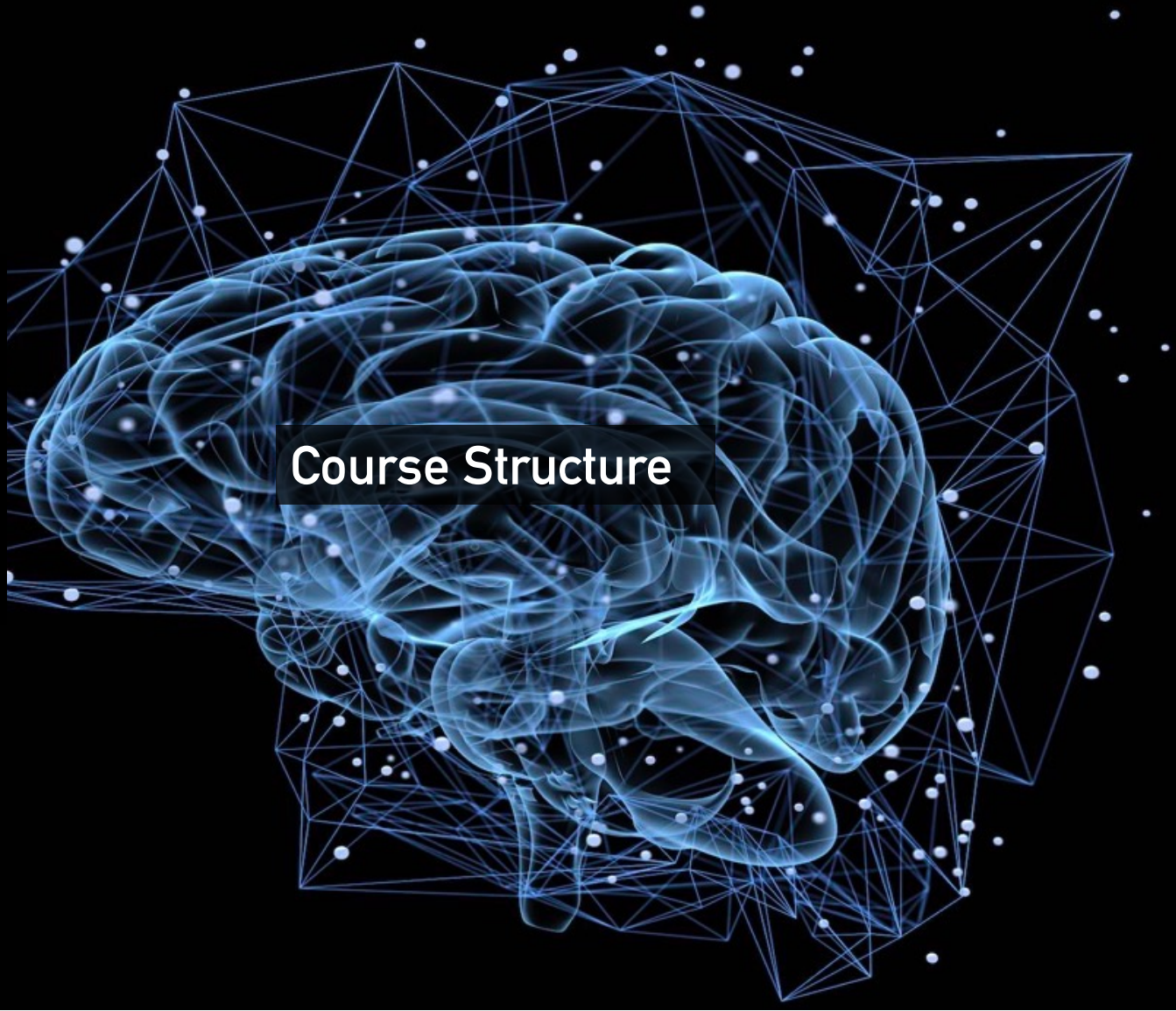


Annual % change (2010 to 2019) DALYs/100,000



Data from IHME (1990 to 2019)

IHME
THE INSTITUTE FOR HEALTH METRICS AND EVALUATION


Course Structure

Course Lectures

COLUMBIA UNIVERSITY

Class Date	Class Number	Lecture Topics	Progress
01/21/25	1	Introduction of Advanced AI	
01/28/25	2	Large Language Model - I (General)	Midterm Project Team Forming
02/04/25	3	Large Language Model - II (Industries)	HW#1 Assignment
02/11/25	4	Knowledge Acquisition and Graph Computing	
02/18/25	5	AI for Bio Science - I (Molecules)	HW#1 Due
02/25/25	6	AI for Bio Science - II (Proteins)	
03/04/25	7	AI for Bio Science - III (Genomics)	HW#2 Assignment
03/11/25	8		Midterm Project Workshop
03/18/25		SPRING BREAK	
03/25/25	9	AI for Bio Science - IV (Drugs)	HW #2 Due; Final Project Team Forming
04/01/25	10	Multi-Modal AI	
04/08/25	11	Perceputal AI	HW#3 Assignment
04/15/25	12	Expressional AI	
04/22/25	13	Reasoning AI	HW#3 Due
04/29/25	14	Artificial General Intelligence	
05/06/25	15		Final Project Workshop

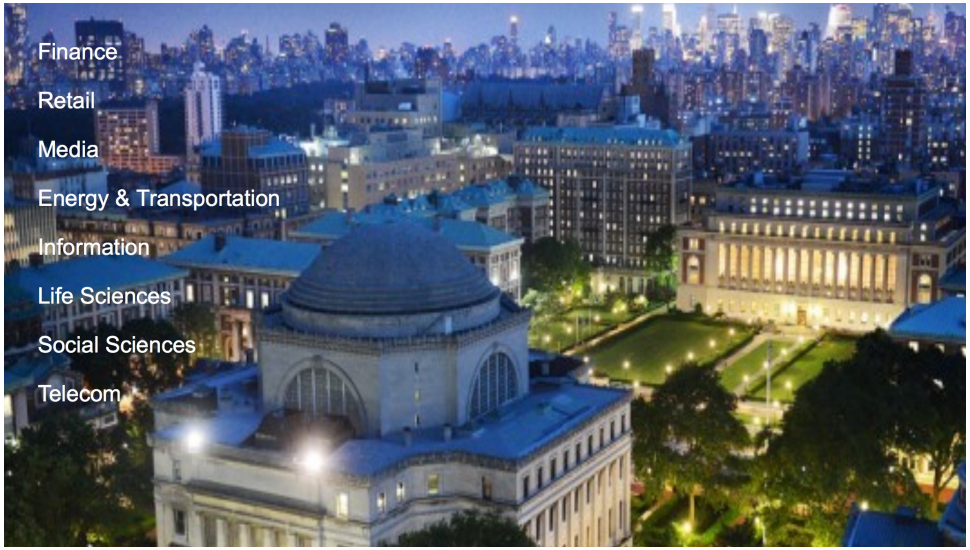
*Lectures: 19:00 – 20:20 *Group Presentations: 20:30 – 21:30

Course Information and TAs

Professor Lin:

Office Hours and Location: By appointment (500 Fifth Ave., Suite 2420, New York, NY 10110)

Contact: c.lin@columbia.edu



TAs:

- Zelin Yu (zy2489)
- Likhith Ayinala (la3073)

Website (Materials):

- http://www.ee.columbia.edu/~cylin/course/bigdata/

Course Grading

- Midterm Project (40%):
  - Advanced A.I. Assistants for Financial Industry and Healthcare Industry.
- Final Project (40%):
  - Advanced A.I. Technologies; *or*
  - Advanced A.I. for Bio Science.
- 3 Homeworks (15%):
  - AI Assistant, AI for Bio Science, and Advanced AI Technology
- Course Participation (5%) :
  - Attendance and Discussions

Midterm Project

- Each team is composed of 3 people.
- Choose among these areas. Each area has two teams.

➢ Financial Industry:
- ❖ Tax Advisor
- ❖ Insurance Advisor
- ❖ Private Banking Specialist
- ❖ Commercial Banking Specialist
- ❖ Fund Manager / M&A Specialist

➢ Healthcare Industry:
- ❖ Nurse
- ❖ Doctor
- ❖ Nutritionist
- ❖ Pharmacist
- ❖ Radiologist

Midterm Project

Midterm Project – To-Do by Next Week (1/28/25)

- Sign up the team sheet and project choice – to be shared by TA.
- Each team prepare a 3-min presentation to discuss.
  - The Certificate Challenge
  - Initial Thought of Knowledges and Capabilities of the AI Assistant to be included

Financial AI Assistant examples

- Financial Knowledge Understanding:
  - Financial Terms
  - Financial Knowledge
  - Financial Quiz
  - Financial Products

- Financial Assistant Generative AI:
  - News Title
  - Marketing Materials
  - Insurance Questions
  - Investment Advisor
  - Market Reasoning
  - Customer Interaction

- Financial News Extraction:
  - Financial News Summary
  - Analysis Report Summary
  - Viewpoint Summary
  - Keyword Extraction

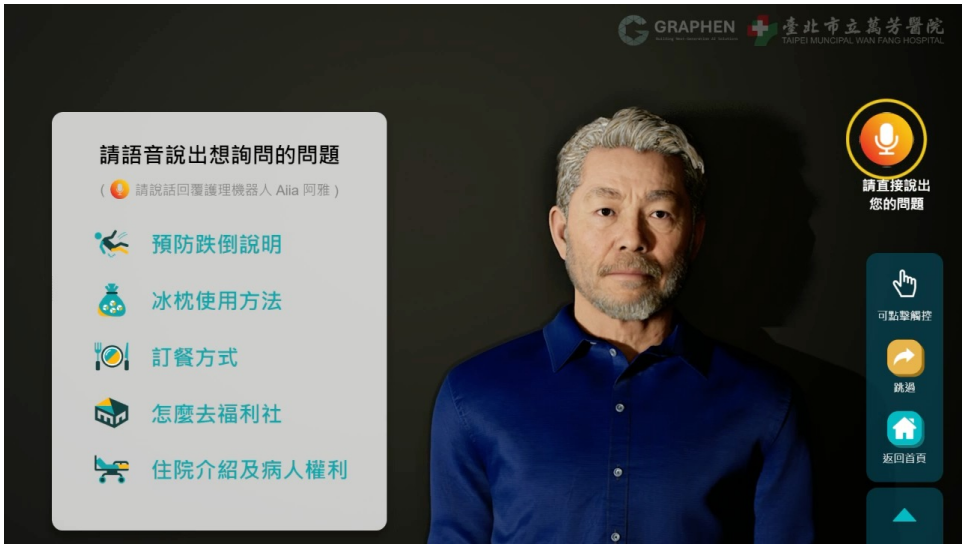
- Financial Computations

- Real-Time Information:
  - Real-Time News
  - Real-Time Stock Price

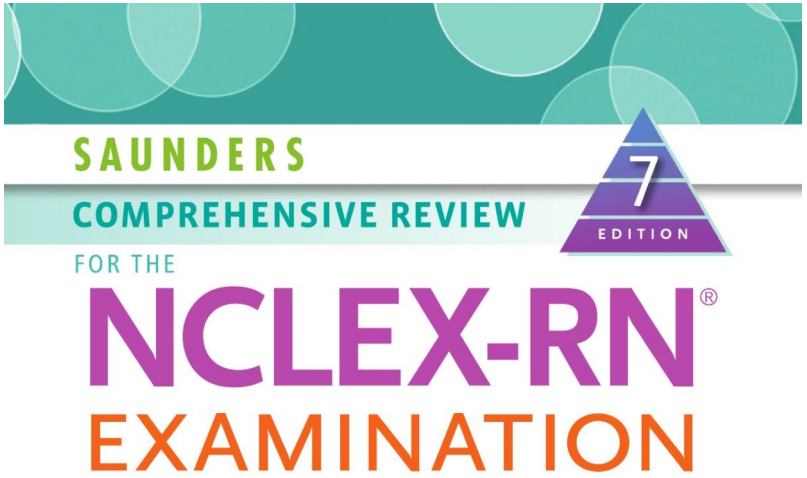
- Compliance and Risk Analysis:
  - Compliance Questions
  - Regulation Questions

Example: Nurse Aiia

- Health instruction
- Family meeting
- Exercise instruction
- Medical documents
- Data recording
- Individual care plan
- Virtual reality therapy
- Communications



Reference Book



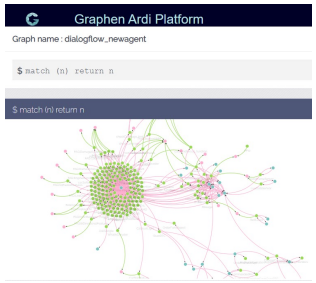
- 20 units. 1155 pages

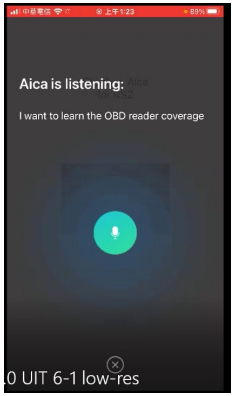
Secure, Scalable AI Chatbot

Internal Database (million records):
Product Documentations
FAQs
Customer Service Records
Account APIs

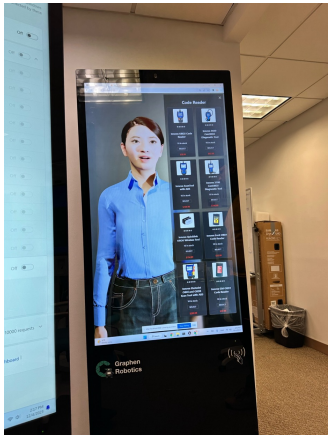
External Data:
Wikipedia, LLM, internet, etc.

GraphDB Analytics

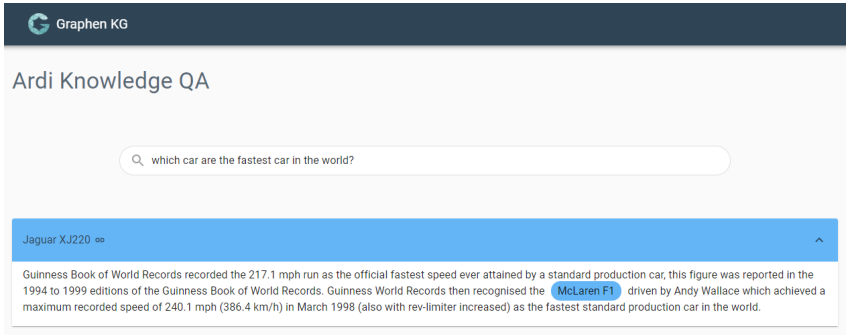




Messenger from Facebook
clover
Square
Slack
Skype
LINE
Twilio (Text Messaging)



Ardi Knowledge and Reasoning RAG



Ardi Backend

Customer-Facing: Texting, Phones, WebApps, Emails, ChatApps, dedicated App (Android/iPhones), and Graphen Kiosks with Point-of-sales payments.

ChatBot Platforms

Falcon LLM
OpenAI

Aiia Clinical Applications – Health Instruction

Columbia University

1. Understanding the disease and treatment
   - Patient's condition, causes, symptoms, prognosis
   - Treatment options, medication use, and potential side effects
2. Symptom management
   - Patient education on common symptoms (pain, difficulty breathing, nausea)
   - Providing symptom control methods
     - Medications
     - Non-pharmacological approaches (aromatherapy, music therapy, acupuncture)
3. Medication education
   - Information on medication names, doses, and usage
   - Ensuring family members understand medication actions and potential side effects

Aiia Clinical Applications – Health Instruction

4 Care skills

- Basic nursing skills
- Hygiene practices
- Changing diapers
- Assisting with turning or repositioning
- Feeding
- Dietary recommendations and requirements

5 Use of medical equipment

- Introduction of the ward environment
- Bed usage
- Pressure sore prevention
- Operation of bathing equipment
- Other relevant equipment

6 Emergency situation handling

- First aid measures
- Proper use of call bells
- Providing guidance for common emergency situations (cardiac arrest, difficulty breathing)

Aiia Clinical Applications – Health Instruction

Columbia University

7 Psychological and emotional support

- Teaching communication and emotional support techniques
- Guiding family members in coping with stress, depression, anxiety, and other emotions

8 Explanation of care plan

9 Resource introduction

- Community resources
- Palliative care service groups
- Support groups

Medical Aiia's Roles in Hospital (Inpatient)

Columbia University

Waiting Area

Provide Directions
- Specific directions
- Map

Provide General Information
- Parking, Restrooms, Café
- Nearest Pharmacies
- Where / How to pay bill
- General public health information (e.g.: "Mask up" during covid)
- Environment introduction

Act as hospital directory for personnel

Nurse Station

Provide specific information
- Regarding procedure
- Preparation for procedure
- Items / documents to bring for procedure (clothes, towels, medication, etc.)
- Information on Discharge
- Medical services and department information

Assist with Patient check-in & Administrative Tasks
- Help fill health questionnaire
- Consent form

Assist with Bill Payment

Schedule next appointment

Patient Room

Provide general and specific instructions / information
- How to use equipment, hospital bed, etc.
- Daily ward routine
- Drugs, food, and exercise instructions
- Information on surgery/procedure
- Information on post-op care
- Care plan (pre-op, symptom management, prep for discharge)
- Doctor's rounds schedule

Provide companionship

Record patient data

Medical Aiia's Roles in Hospital (Outpatient)

Columbia University

Waiting Area

Provide Directions
- Specific directions
- Map

Provide General Information
- Parking, Restrooms, Café
- Nearest Pharmacies
- Where / How to pay bill
- General public health information (e.g.: "Mask up" during covid)
- Estimated wait times

Act as hospital directory for personnel

Nurse Station

Provide specific information
- Regarding upcoming procedure/examination
- Medical services and department information

Assist with patient check-in

Assist with bill payment

Exam Room / Ward

Administrative Tasks
- Record Patient data
- Help fill health questionnaire

Provide relevant information
- Information about visit
- Post-examination information
- Answer questions about Dr's diagnosis and assessment
- Information on medication

Schedule next appointment

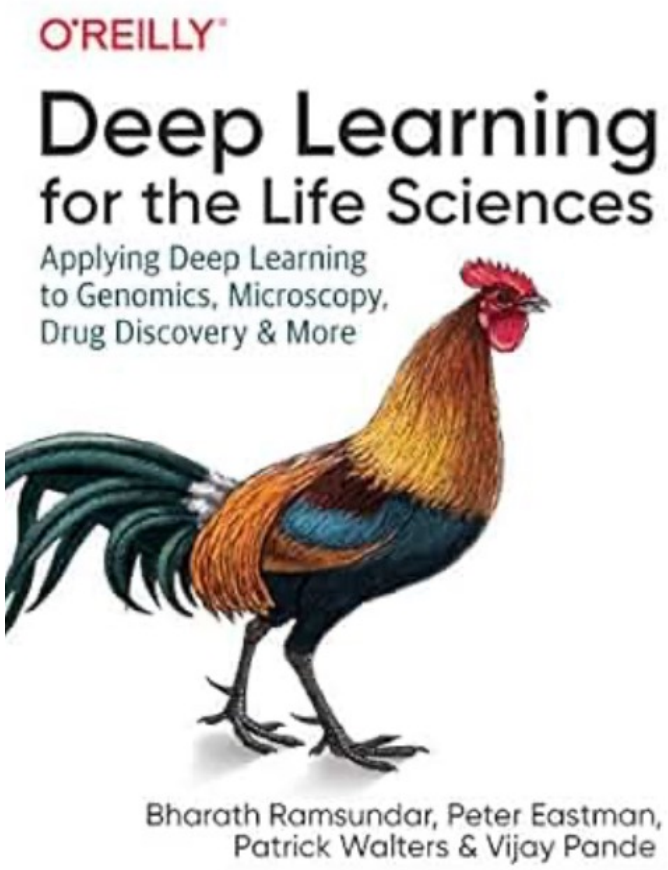
Final Project

- Each team is composed of 3 people.
- Choose among these two areas.
  - Advanced AI Technology:
    - Multi-Modality AI
    - Perception AI
    - Expression AI
    - Reasoning AI
  - Advanced AI for Bio Science:
    - Protein-Ligan Interaction
    - RNA Structure Prediction

Final Project – Building Advanced AI Technologies

https://www.youtube.com/watch?v=BV8qFeZxZPE

Reference Book



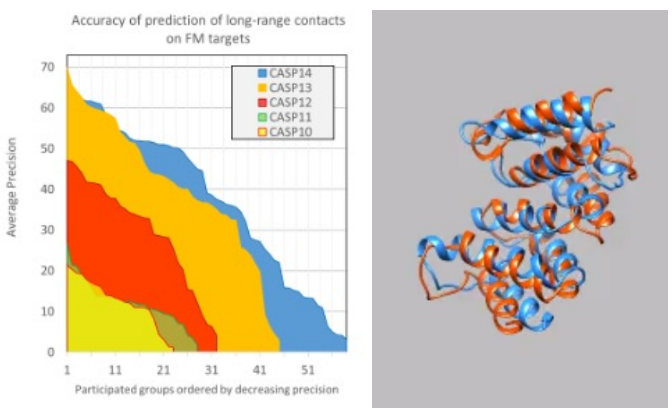
Chapters

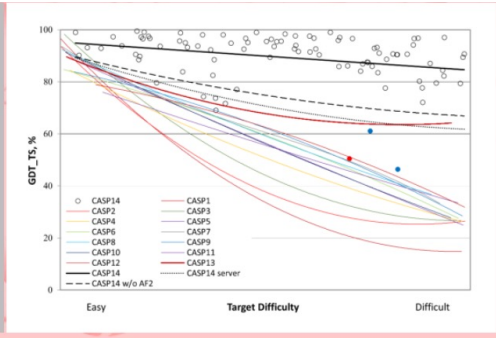
1. Why Life Science?
2. Introduction to Deep Learning
3. Machine Learning with DeepChem
4. Machine Learning for Molecules
5. Biophysical Machine Learning
6. Deep Learning for Genomics
7. Machine Learning for Microscopy
8. Deep Learning for Medicine
9. Generative Models
10. Interpretation of Deep Models
11. A Virtual Screening Workflow Example
12. Prospects and Perspectives

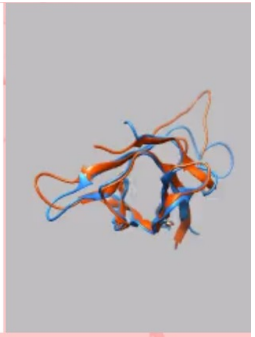
CASPs : Critical Assessment of protein Structure Prediction

- CASP is a community-wide, worldwide experiment for protein structure prediction taking place every two years since 1994, the primary goal of CASP is to help advance the methods of identifying protein three-dimensional structure from its amino acid sequence.

- Google AlphaFold 2 scored above 90 for around 2/3 of the proteins in CASP14's global distance test (GDT). Since this, applying predicted protein structure to precise analysis of target research, mechanism revealing, and even drug developments became convincingly





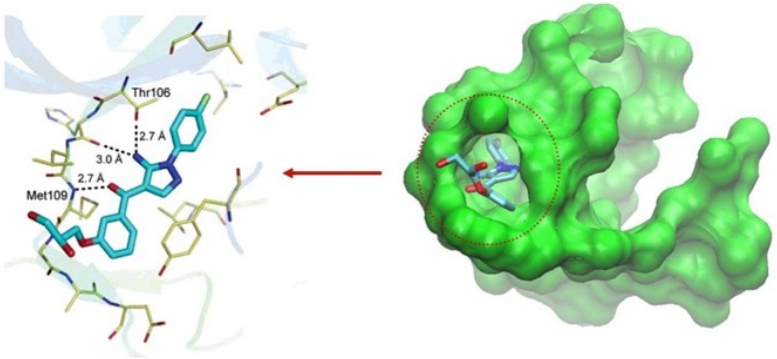


CASP7: T0283-D1
model 321_1: GDT_TS=75

CASP12: T0866-D1
model 325_5: GDT_TS=81

CASP16 and Beyond : Next stage of structural prediction

- For the outstanding records that Google Alphafold2 has achieved in CASP14 in **single protein structural prediction**, this year the competition goes into the new stage: more complex conformation prediction. Majorly including:
  - **Multi-chain protein structures**
  - **Protein-Ligand complexes**
  - **RNA structures and complexes**
  - **Protein conformational ensembles**

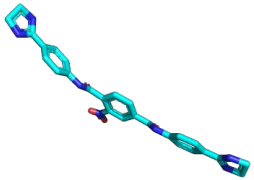


Protein-Ligand Interaction

Protein-Nucleotide complexing

Drug-Target Interactions and complexes

- In small molecular drug development, to reach viable drug efficacy, it's crucial to clarify how the **drug molecule interacts with designed targets**, as well as the consequences induced by the conjugation of drug-target pairs .
- Drug-target interaction can be predicted with majorly two ways in traditional: ligand docking and quantum-physic simulation. These methods are limited with lacking flexibility to know the induced conformational change and extremely high cost of time and computational resources, respectively.

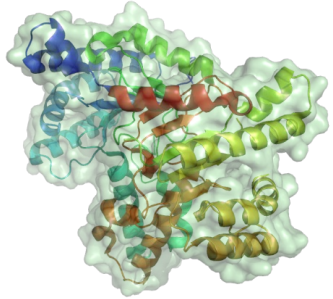


Ligand

+

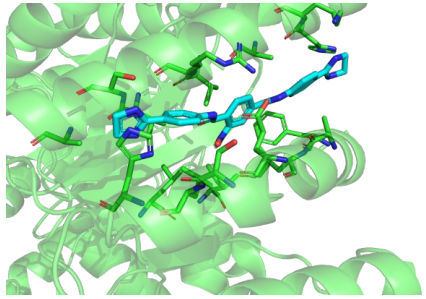
>Protein A
SGFRKMAFPSGKVEGCMVQVTCG
TTTLNGLWLDDVVYCPRHVICTSE
DMLNPNYEDLLIRKSNHNFLVQAG
NVQLRVIGHSMQNCVLKLKVDTA
NPKTPKYKFVRIQPGQTFSVLACY
NGSPSGVYQCAMRPNFTIKGSFLN
GSCGSVGFNIDYDCVSFCYMH….

Protein sequence



Protein structure

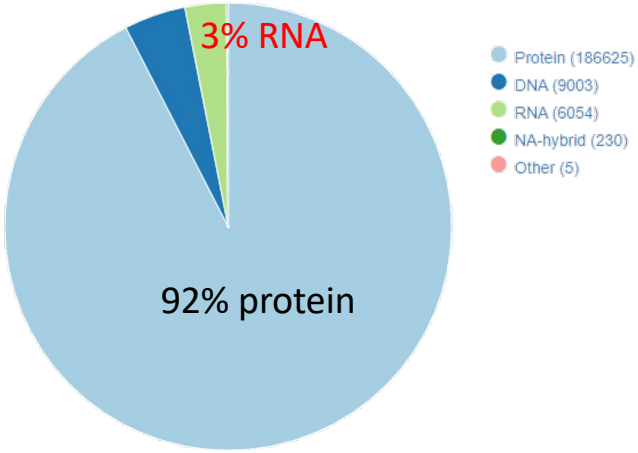
+



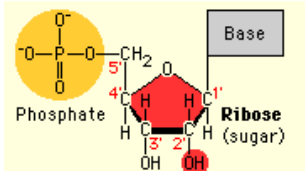
Protein-ligand interaction

RNA structures and complexes predictions

- RNA molecules play fundamental roles in cellular processes. Their function and interactions with other biomolecules are dependent on the ability to form complex three-dimensional (3D) structures
- Experimental determination of **RNA atomic 3D structures** is laborious and challenging
  - Only **3%** of known RNA 3D structures in public database



RNA basic unit:
Nucleotide



Primary Sequence
UGCUCCUAGUACGAGAGGACCGGAGUG…

How RNA folds?

3D structure