

ChatGPT και Εφαρμογές AI για Ιατρούς

3th session – AI in Biomedicine και αλγόριθμοι Κατηγοριοποίησης για την Αξιολόγηση Ιατρικών Δεδομένων και την Υποστήριξη Κλινικών Αποφάσεων.

UNIVERSITY OF THE
AEGEAN



SCHOOL OF ENGINEERING
DEPARTMENT OF INFORMATION
AND COMMUNICATION
SYSTEMS ENGINEERING

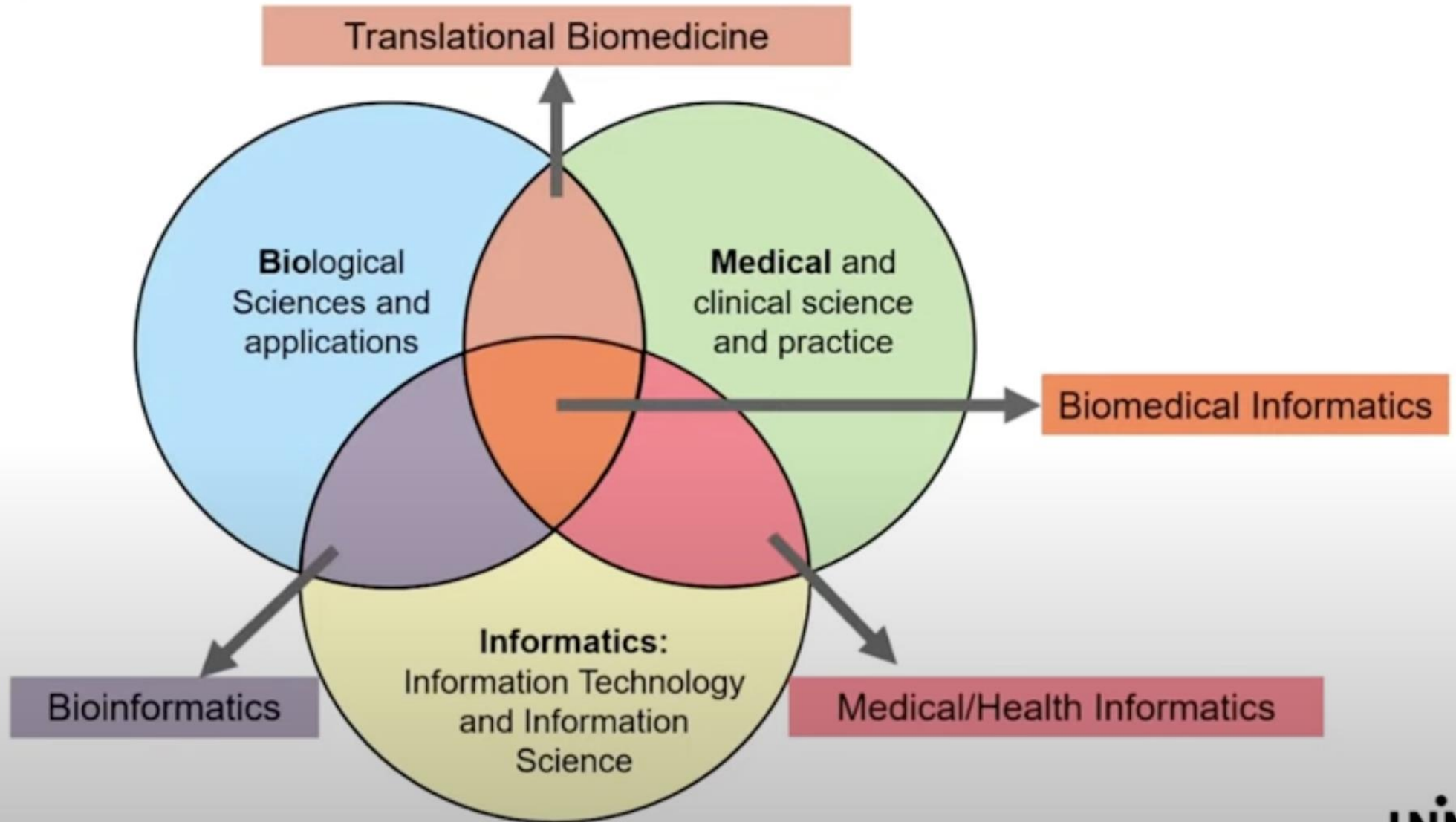
Presenter: Panagiotis Symeonidis

Associate Professor

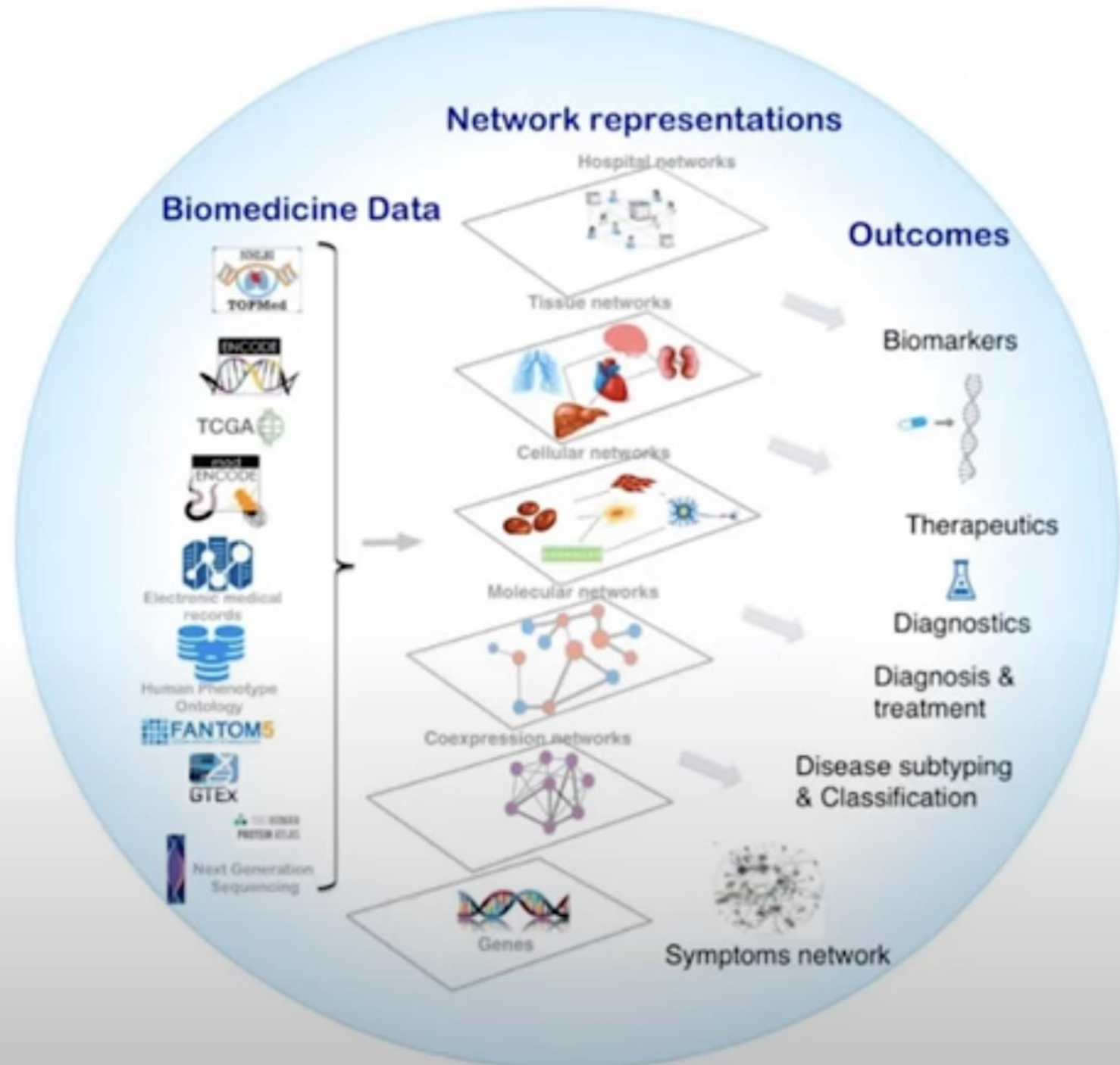
<http://panagiotissymeonidis.com>

psymeon@aegean.gr

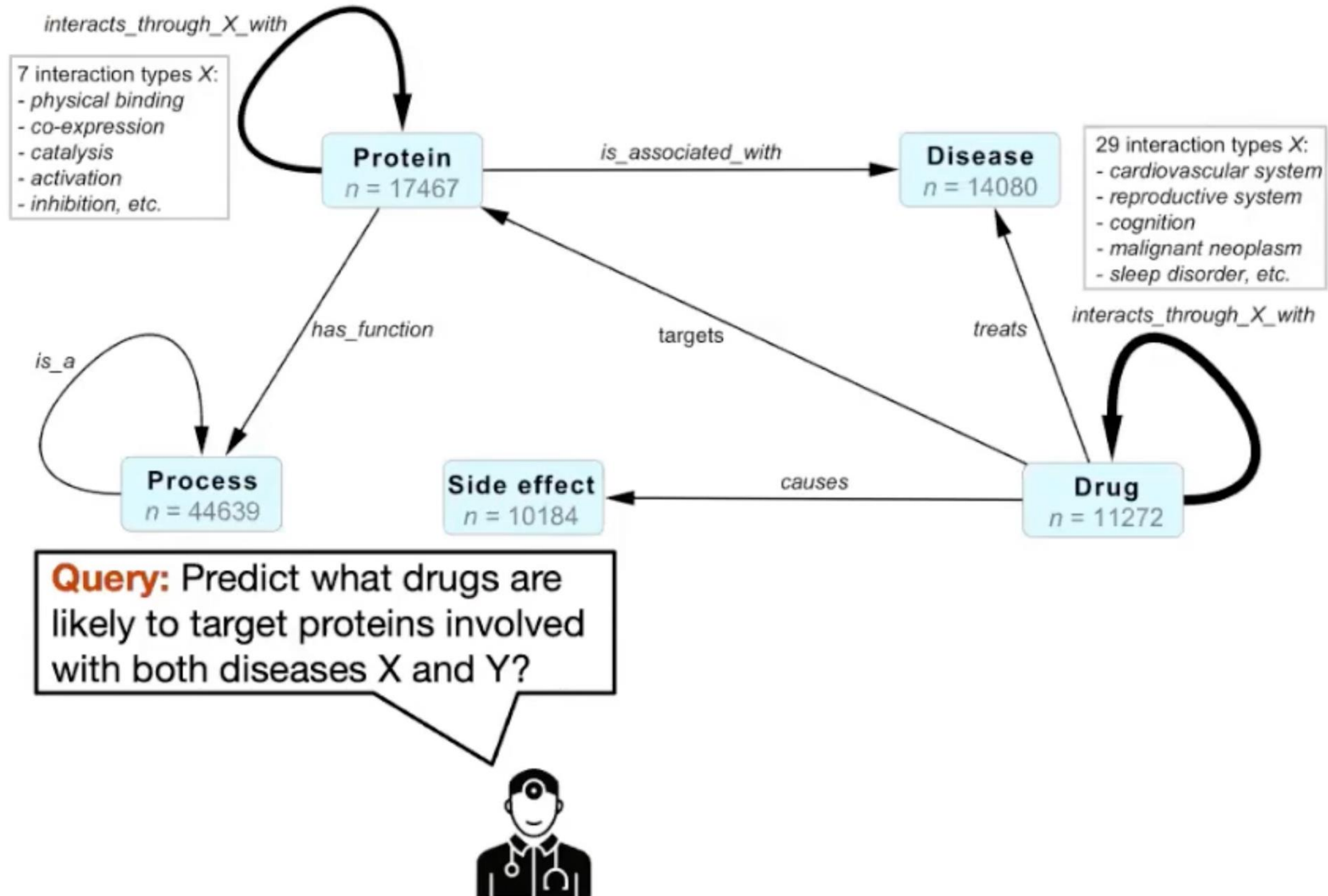
Inter-disciplinary Science **Biology, Medicine, Informatics**



Network Medicine or else Graph Medicine



Example of a question to be answered



AI for BioMedical Informatics Research

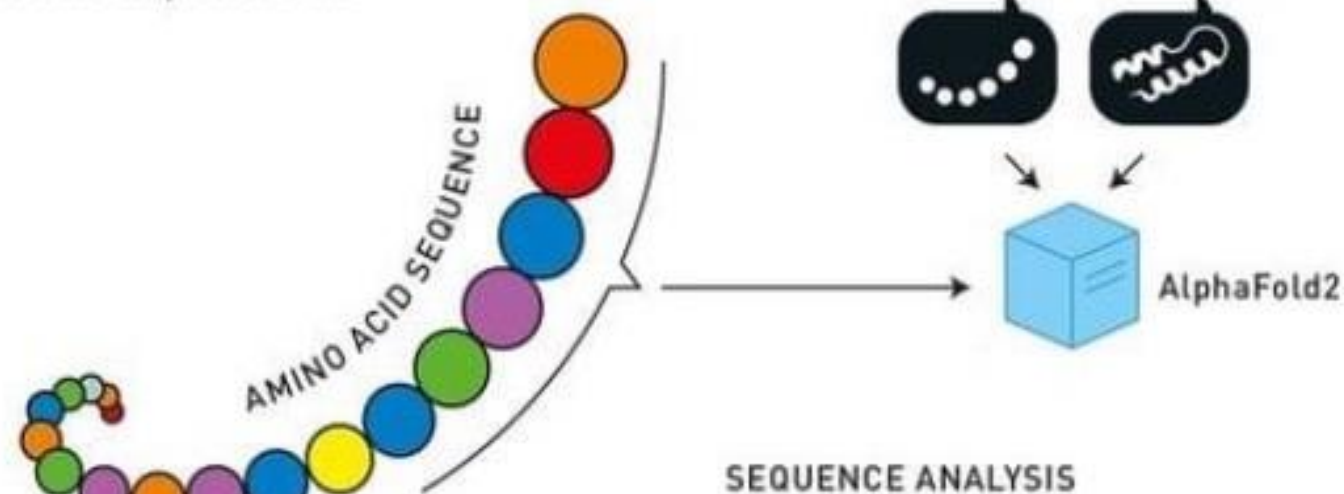
- **Protein Engineering:** AlphaFold, an AI tool developed by DeepMind, predicts structures of thousands of proteins with more than 90% accuracy, tremendously accelerating scientific productivity: it previously took years of study for a PhD student to explore the three-dimensional structure formation of a single protein (Jumper et al, 2021).
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>

How does AlphaFold2 work?

As part of AlphaFold2's development, the AI model has been trained on all the known amino acid sequences and determined protein structures.

1. DATA ENTRY AND DATABASE SEARCHES

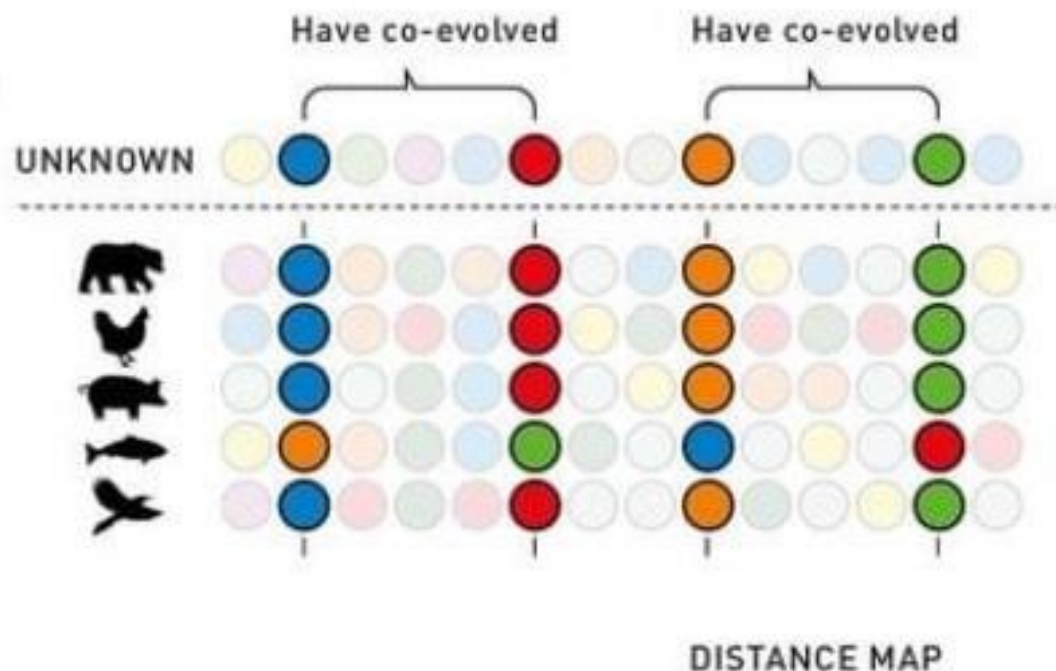
An amino acid sequence with unknown structure is fed into AlphaFold2, which searches databases for similar amino acid sequences and protein structures.



2. SEQUENCE ANALYSIS

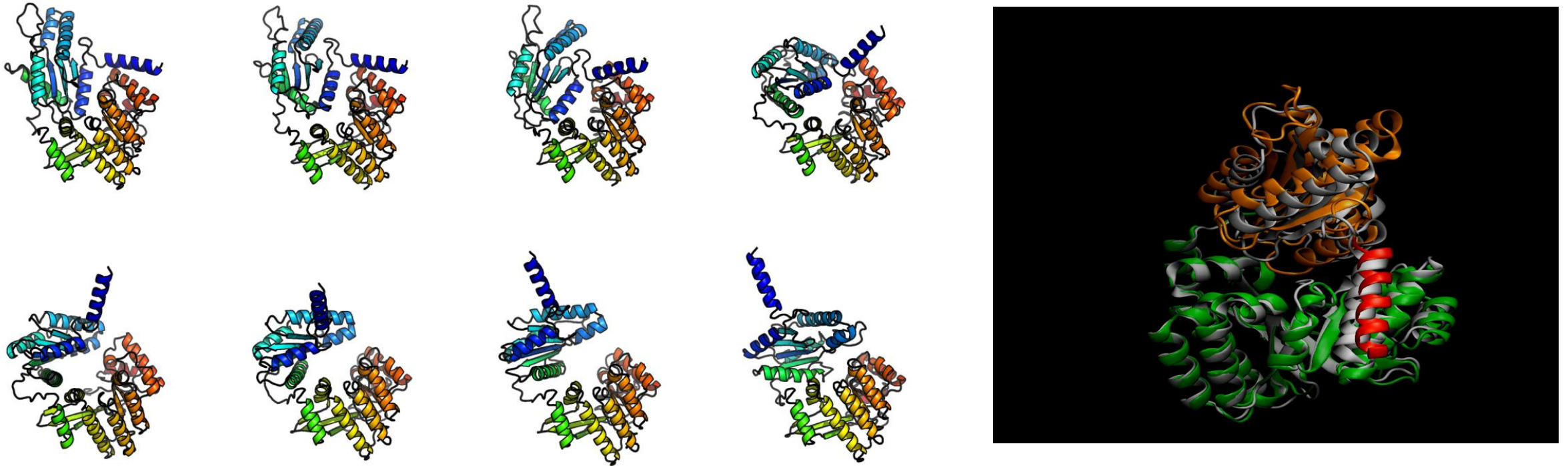
The AI model aligns all the similar amino acid sequences – often from different species – and investigates which parts have been preserved during evolution.

In the next step, AlphaFold2 explores which amino acids could interact with each other in the three-dimensional protein structure. Interacting amino acids co-evolve. If one is charged, the other has the opposite charge, so they are attracted to each other. If one is replaced by a water-repellent (hydrophobic) amino acid, the other also becomes hydrophobic.



Microsoft's BioEmu-1 and how it works

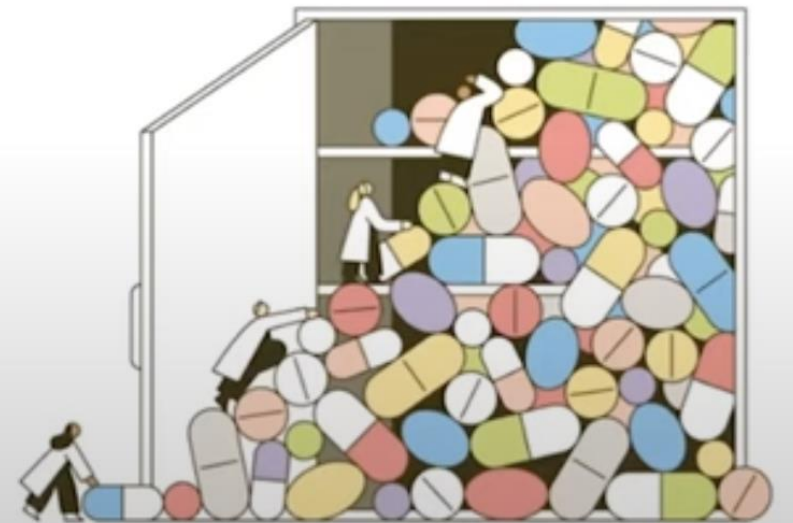
- **BioEmu-1** can generate thousands of protein structures per hour and show how the fold or unfold in heat or pressure.



Lewis, Sarah, et al. "Scalable emulation of protein equilibrium ensembles with generative deep learning." *bioRxiv* (2024): 2024-12.

Drug Repurposing

- **“Drug repurposing or, simply, drug repositioning/rediscovery or reprofiling is a strategy to identify advanced uses for preapproved drugs or existing medications.”**
- Discover new useful activity for a distinct malady in an older clinically used drug or one that failed in later stages of development
- Exploring new medical uses for existing drugs, including approved, discontinued, shelved and investigational therapeutics
- Estimation: 1/3 of recent approvals correspond to repurposing examples
- Partially predicted and carried out using systematic methods



Network medicine framework for identifying drug-repurposing opportunities for COVID-19

Deisy Morselli Gysi^{a,b,c,1}, Ítalo do Valle^{a,b,1}, Marinka Zitnik^{d,e,1}, Asher Ameli^{b,f,1}, Xiao Gan^{a,b,c,1}, Onur Varol^{a,b,g}, Susan Dina Ghiassian^f, J. J. Patten^h, Robert A. Davey^h, Joseph Loscalzoⁱ, and Albert-László Barabási^{a,b,j,2}

^aNetwork Science Institute, Northeastern University, Boston, MA 02115; ^bDepartment of Physics, Northeastern University, Boston, MA 02115; ^cChanning Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115; ^dDepartment of Biomedical Informatics, Harvard University, Boston, MA 02115; ^eHarvard Data Science Initiative, Harvard University, Cambridge, MA 02138; ^fData Science Department, Scipher Medicine, Waltham, MA 02453; ^gFaculty of Engineering and Natural Sciences, Sabanci University, Istanbul 34956, Turkey; ^hDepartment of Microbiology, National Emerging Infectious Diseases Laboratories, Boston University, Boston, MA 02118; ⁱDepartment of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115; and ^jDepartment of Network and Data Science, Central European University, Budapest 1051, Hungary

Edited by Eugene V. Koonin, NIH, Bethesda, MD, and approved March 30, 2021 (received for review December 12, 2020)

The COVID-19 pandemic has highlighted the need to quickly and reliably prioritize clinically approved compounds for their potential effectiveness for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infections. Here, we deployed algorithms relying on artificial intelligence, network diffusion, and network proximity, tasking each of them to rank 6,340 drugs for their expected efficacy against SARS-CoV-2. To test the predictions, we used as ground truth 918 drugs experimentally screened in VeroE6 cells, as well as the list of drugs in clinical trials that capture the medical community's assessment of drugs with potential COVID-19 efficacy. We find that no single predictive algorithm offers consistently reliable outcomes across all datasets and metrics. This outcome prompted us to develop a multimodal technology that fuses the predictions of all algorithms, finding that a consensus among the different predictive methods consistently exceeds the performance of the best individual pipelines. We screened in human cells the top-ranked drugs, obtaining a 62% success rate, in contrast to the 0.8% hit rate of nonguided screenings. Of the six drugs that reduced viral infection, four could be directly repurposed to treat COVID-19, proposing novel treatments for COVID-19. We also found that 76 of the 77 drugs that successfully reduced viral infection do not bind the proteins targeted by SARS-CoV-2, indicating that these network drugs rely on network-based mechanisms that cannot be identified using docking-based strategies. These advances offer a methodological pathway to identify repurposable drugs for future pathogens and neglected diseases underserved by the costs and extended timeline of de novo drug development.

systems biology | network medicine | drug repurposing | infectious diseases

The disruptive nature of the COVID-19 pandemic has unveiled the need for the rapid development, testing, and deployment of new drugs and cures. Given the compressed timescales, the de novo drug development process, which typically lasts a decade or longer, is not feasible. A time-efficient strategy must rely on drug

wide association studies (7), and network perturbations (7–15). Yet, typically only a small subset of the top candidates is validated experimentally; hence, the true predictive power of the existing repurposing algorithms remains unknown. To quantify and compare their true predictive power, all algorithms must make predictions for the same set of candidates, and the experimental validation must focus not only on the top candidates, as it does now, but on a wider list of drugs chosen independently of their predicted rank.

The COVID-19 pandemic presents both the societal imperative and the rationale to test drugs at a previously unseen scale. Hence, it offers a unique opportunity to quantify and improve the efficacy of the available predictive algorithms, while also identifying potential treatments for COVID-19. Here, we implement three network-medicine drug-repurposing algorithms that rely on artificial intelligence (AI) (15, 16), network diffusion

Significance

The COVID-19 pandemic has highlighted the importance of prioritizing approved drugs to treat severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infections. Here, we deployed algorithms relying on artificial intelligence, network diffusion, and network proximity to rank 6,340 drugs for their expected efficacy against SARS-CoV-2. We experimentally screened 918 drugs, allowing us to evaluate the performance of the existing drug-repurposing methodologies, and used a consensus algorithm to increase the accuracy of the predictions. Finally, we screened in human cells the top-ranked drugs, identifying six drugs that reduced viral infection, four of which could be repurposed to treat COVID-19. The developed strategy has significance beyond COVID-19, allowing us to identify drug-repurposing candidates for neglected diseases.

Αλγόριθμοι ΑΙ για Drug Re-purposing

(χρήση
υπαρχόντων
φαρμάκων για νέες
ασθένειες)

Predicting ICU Mortality and Drug Recommendation

Published work

Mortality Prediction and Safe Drug Recommendation for Critically-ill Patients
(Symeonidis et al., IEEE BIBE 2022)

Medical scores and Bio Indexes

Most medical scores are good predictors of ICU mortality, but they failed to predict survival after discharge

- **PESI: Pulmonary Embolism Severity Index**
- **SOFA: Sequential Organ Failure Assessment**
- **OASIS: Oxford Acute Severity of Illness Score**
- **APACHE: Acute Physiology And Chronic Health Evaluation**
- **SAPS: Simplified Acute Physiology Score**

PESI score

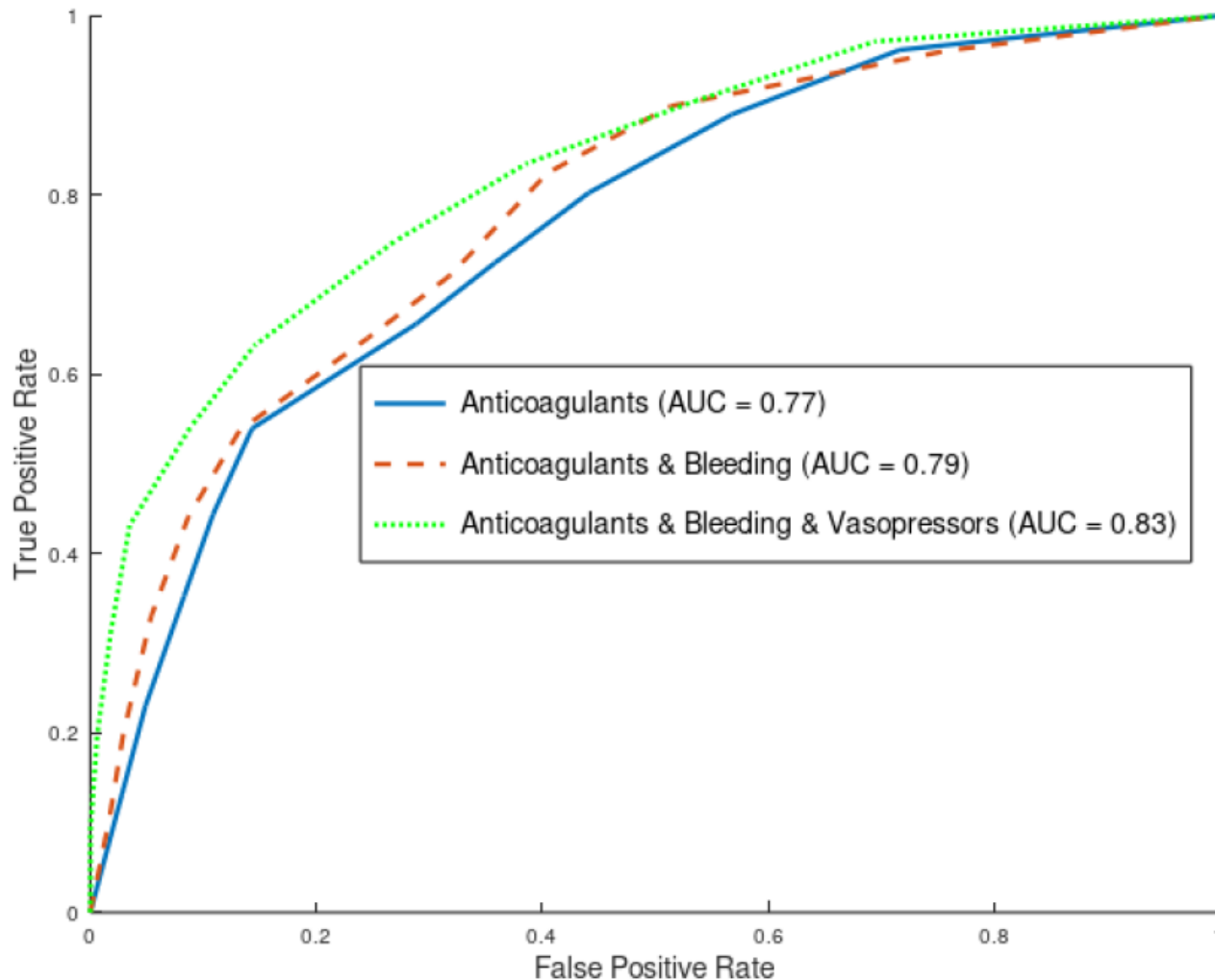
Predictors	Score
Age	Years
Male sex	+ 10
Cancer	+ 30
Heart failure	+ 10
COPD	+ 10
HR \geq 110 b.p.m	+ 20
SBP <100 mmHg	+ 30
RR > 30 breath per minute	+ 20
BT < 36 °C	+ 20
Delirium	+ 60
SaO ₂ < 90%	+ 20
	Total

Low risk
(\leq 65 class I, 66-85, class II)
Mortality 1.9%

Intermediate risk
86-105 class III, 106-125 class IV)
Mortality 18.4%

High risk
(> 125 class V)
Mortality 25%

We performed an ablation study for predicting patients' mortality in ICU by adding different Drug Categories



Anticoagulants (antithrombotic) reduce the risk of developing blood clots which disrupt the flow of blood around your body.

Bleeding drugs used to help stop bleeding

Vasopressors increase mean arterial pressure, which leads to more blood to organs.

Random Forest was found to be the best classifier/predictor

Classification: Definition

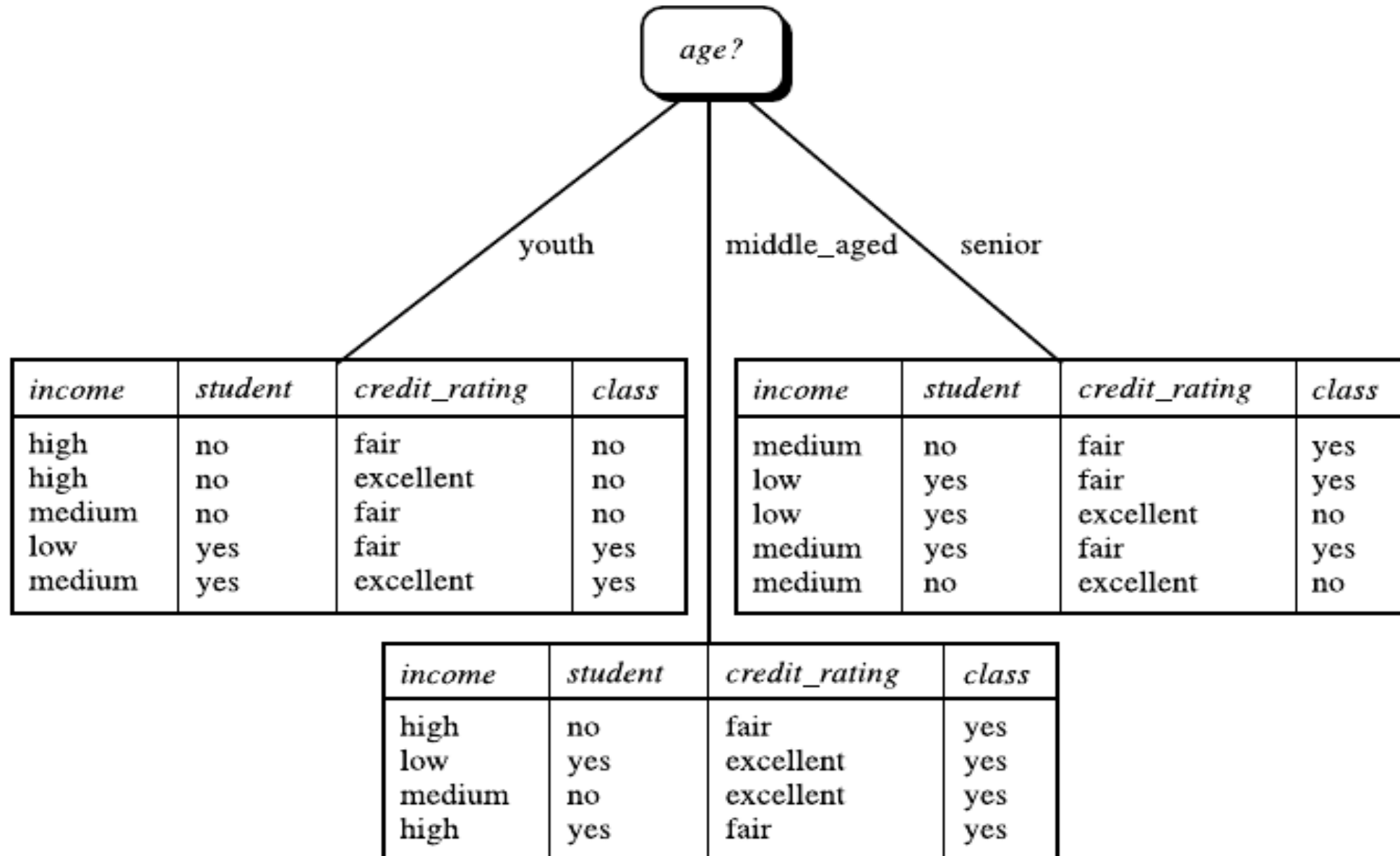
Given a collection of records (training set)

- Each record is characterized by a tuple (x,y) , where x is the attribute set and y is the class label**
 - ◆ x : attribute, predictor, independent variable, input**
 - ◆ y : class, response, dependent variable, output**

Task:

- Learn a model that maps each attribute set x into one of the predefined class labels y**

Attribute Selection: Information Gain



Metrics for Performance Evaluation

- **Focus on the predictive capability of a model**
 - Rather than how fast it takes to classify or build models, scalability, etc.
- **Confusion Matrix:**

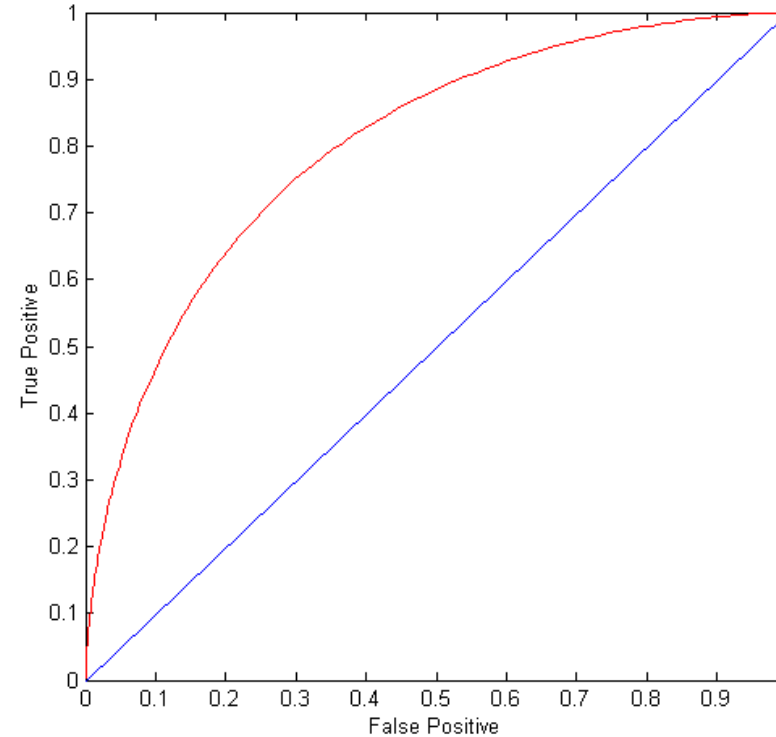
	PREDICTED CLASS		
	Class=Yes	Class=No	
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

a: TP (true positive)
b: FN (false negative)
c: FP (false positive)
d: TN (true negative)

ROC Curve

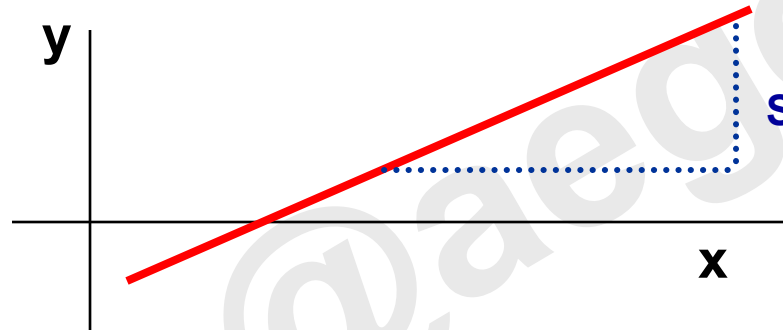
(TPR,FPR):

- **(0,0): declare everything to be negative class**
- **(1,1): declare everything to be positive class**
- **(1,0): ideal**
- **Diagonal line:**
 - Random guessing
 - Below diagonal line:
 - » prediction is opposite of the true class



Simple linear regression

- Relation between 2 continuous variables (SBP and age)



$$y = \alpha + \beta_1 x_1$$

- **α coefficient** : The intercept of the line, which is the value of y when $x=0$ It represents the point where the line crosses the y -axis.
- **β_1 coefficient** : The slope of the line.
 - Amount by which y changes on average when x changes by one unit
 - It indicates how steep the line is.

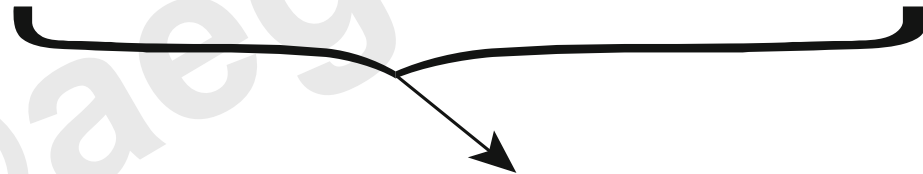
From Multi-linear to the logistic Model

•Multi-linear
regression



$$z = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

$$z = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$



$$f(z) = \frac{1}{1 + e^{-z}}$$

$$= \frac{1}{1 + e^{-(\alpha + \sum \beta_i X_i)}}$$

•Logistic
regression

•Model



Coronary Heart Disease (CHD) Prediction Example

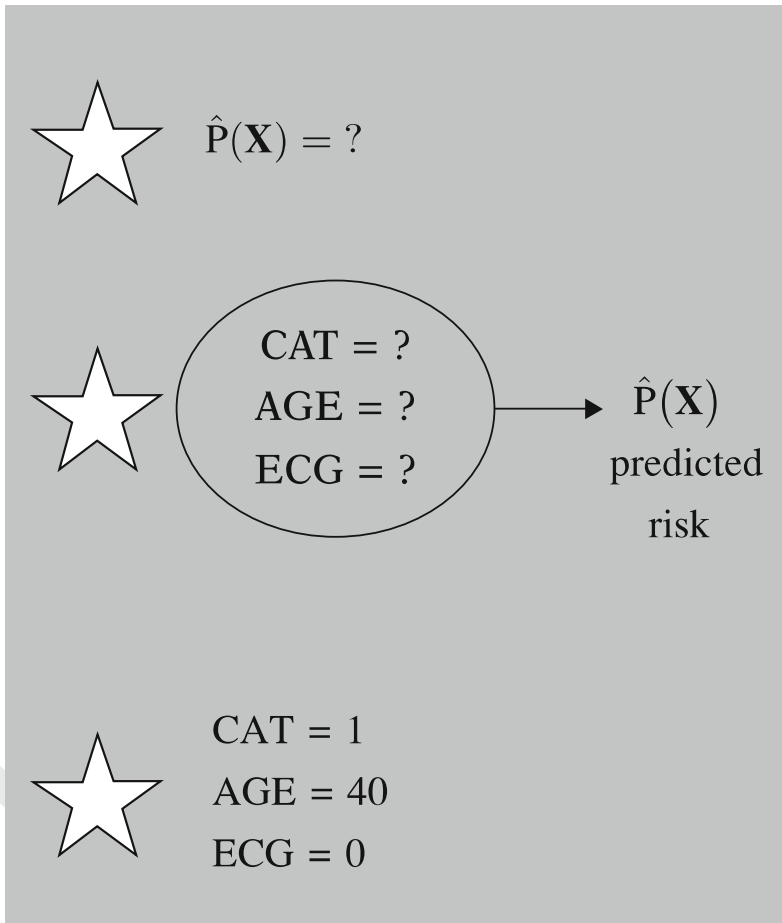
ABOUT PATIENTS' DATA

THE FOLLOWING FOUR ATTRIBUTES ARE COLLECTED

probability of CHD	←	• Dependent variable
catecholamine level (0=low, 1=high)	←	• Independent variable
ECG (0=normal, 1=abnormal)	←	• Independent variable
age (in years)	←	• Independent variable

Test the predicting power of the trained Model

Let's assume a patient with the following clinical status



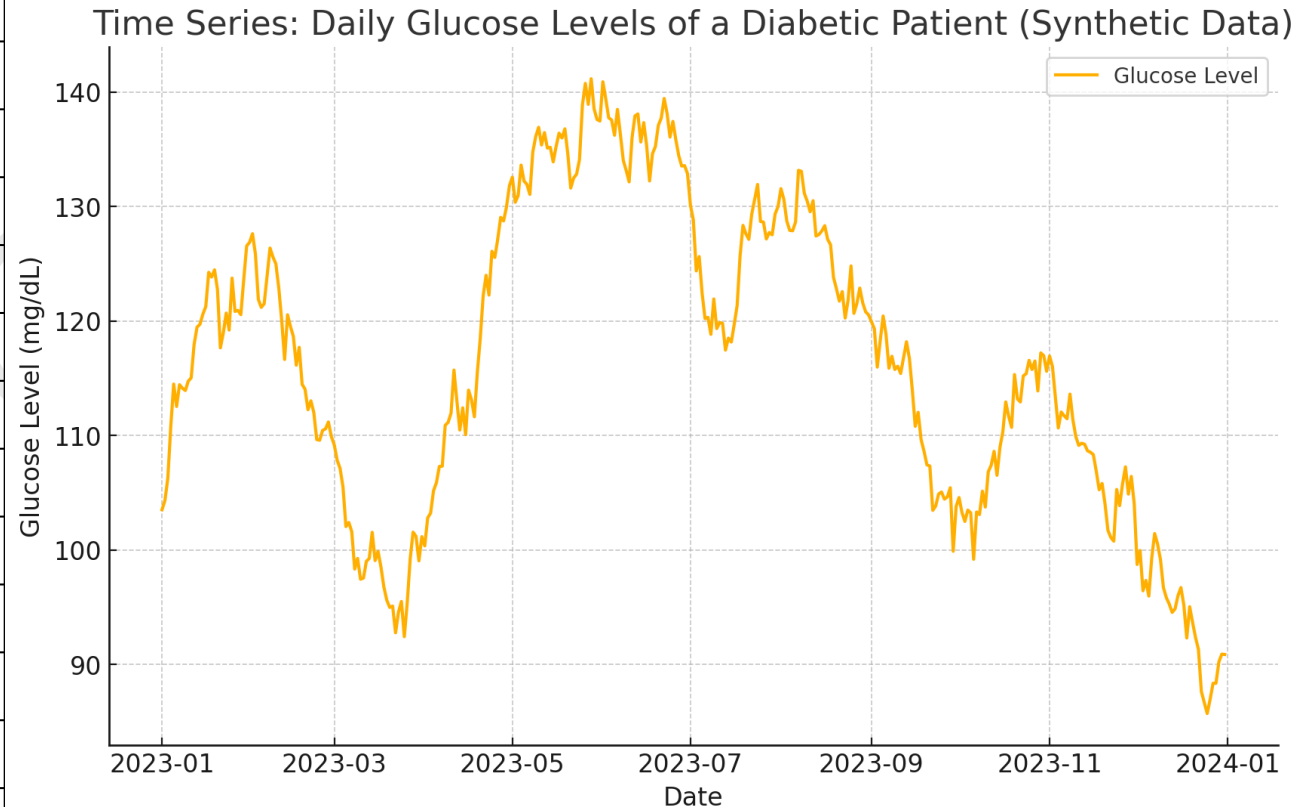
Predicting the Risk for Coronary Heart Disease

$$\begin{aligned} \hat{P}(\mathbf{X}) &= \frac{1}{1 + e^{-[-3.911 + 0.652(1) + 0.029(40) + 0.342(0)]}} \\ &= \frac{1}{1 + e^{-(-2.101)}} \\ &= \frac{1}{1 + 8.173} \\ &= 0.1090, \text{ i.e., risk } \simeq 11\% \end{aligned}$$

Example of a time series

The collection of daily average values of glucose level of a patient for year constitutes a time series.

Date	Glucose Level
2023-01-01	104
2023-01-31	127
2023-03-02	108
2023-04-01	100
2023-05-01	133
2023-05-31	137
2023-06-30	133
2023-07-30	129
2023-08-29	122
2023-09-28	105
2023-10-28	114
2023-11-27	107
2023-12-27	88



Error Calculation and Conclusion

1. `MSE = mean_absolute_error(y_true=test_Y, y_pred=y_predicted)`: **Calculates the Mean Absolute Error (MAE).**
2. `mse_total.append(MSE)`: **Stores the MAE.**
3. `rmse_total.append(math.sqrt(MSE))`: **Calculates and stores the Root Mean Squared Error (RMSE).**
4. **Summary:** The code loops through patients, trains a model, makes predictions, and evaluates performance using MAE and RMSE.