DL in Healthcare Research: Proteomics, Genomics, and Medical Imaging

```
Data Sciences Institute
Topics in Deep Learning
```

Outline

- General overview
- Proteomics
- Genomics
- Radiomics

DL in Healthcare Research

Vast amounts of healthcare data

- Over the past couple of decades, technological advances in data collection have led to a higher abundance of digital health data
  - Genome sequencing
  - Wearable sensors
  - High resolution 3D imaging
  - Electronic health records

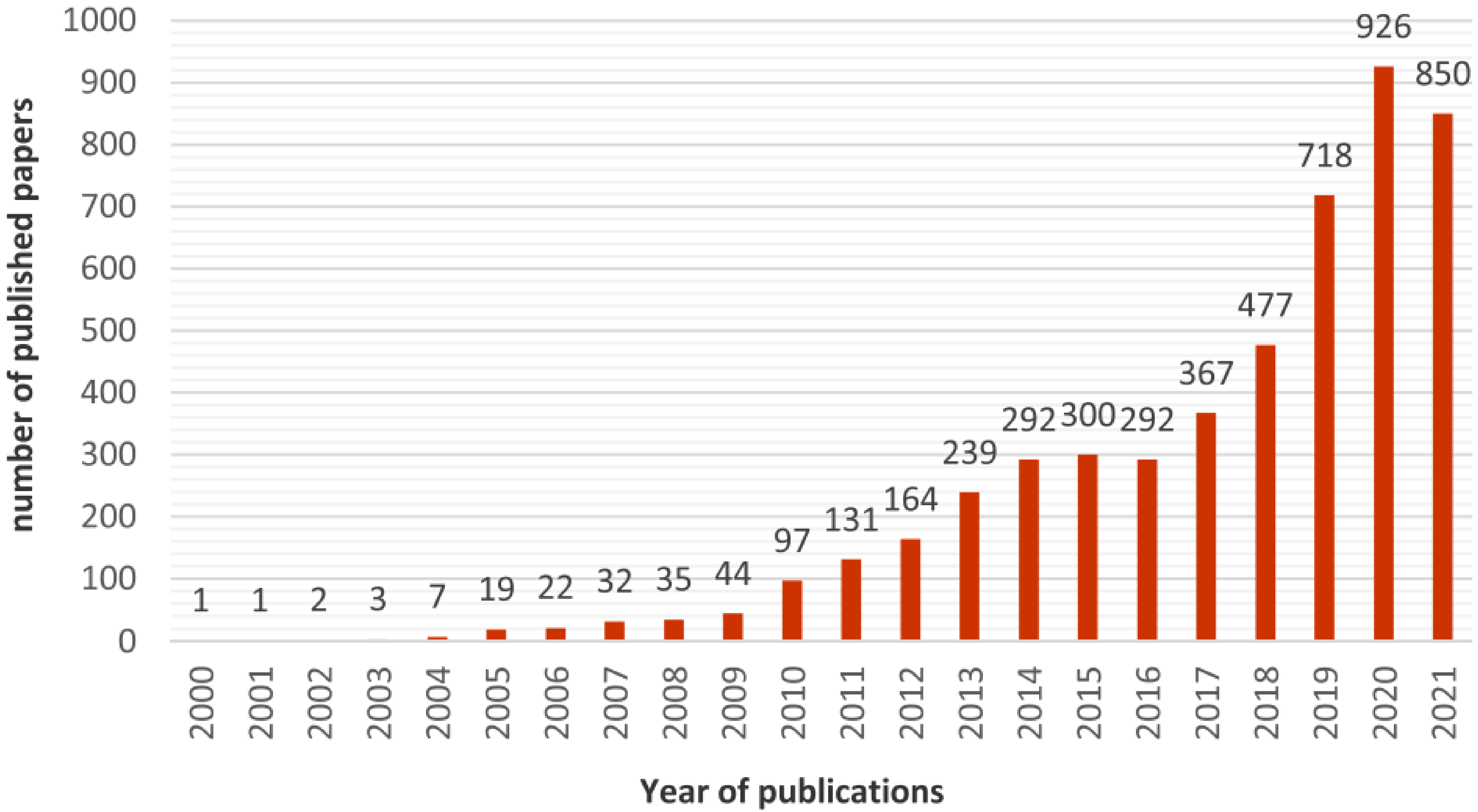
- Analysing these types of data can prove very challenging, given their sheer size and complexity

DL as an analytical tool

- DL offers several advantages over alternative data analysis tools when dealing with large volumes of data:
  - Ability to model highly complex relationships
  - Ability to handle noisy data
  - Adaptability to diverse data types (e.g., text, images, time series)
  - Automated feature extraction requiring minimal preprocessing
  - Scalability

DL in healthcare research over the years

- Given the complexity of human health, and the large amounts of health data now available, AI methods like DL have become increasingly popular in healthcare research [1]



number of published papers
1000
900
800
700
600
500
400
300
200
100
0
1
1
2
3
7
19
22
32
35
44
97
131
164
239
292
300
292
367
477
718
926
850
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
Year of publications

Proteomics

Protein structure research

- Protein structure and function are inherently linked
    - Knowing a protein's structure helps ellucidate its interactions with other molecules
    - For this reason, protein structure research is extremely important for fields like pharmacology, biochemistry, and immunology

- **The problem:** discovering new protein structures has historically proven to be a very time-consuming and expensive process

The Protein Data Bank (PDB)

- The Protein Data Bank (PDB) contains a comprehensive archive of experimentally determined three-dimensional protein structures

- Currently, it houses around 186,000 known protein structures
    - This number pales in comparison to the billions of known protein sequences

DL in protein structure research

- Protein folding patterns are governed by a variety of different factors, including evolutionary, physical, and geometric constraints

- There exist large databanks from which these factors can be modeled, such as the PDB with its genetic and structural databases

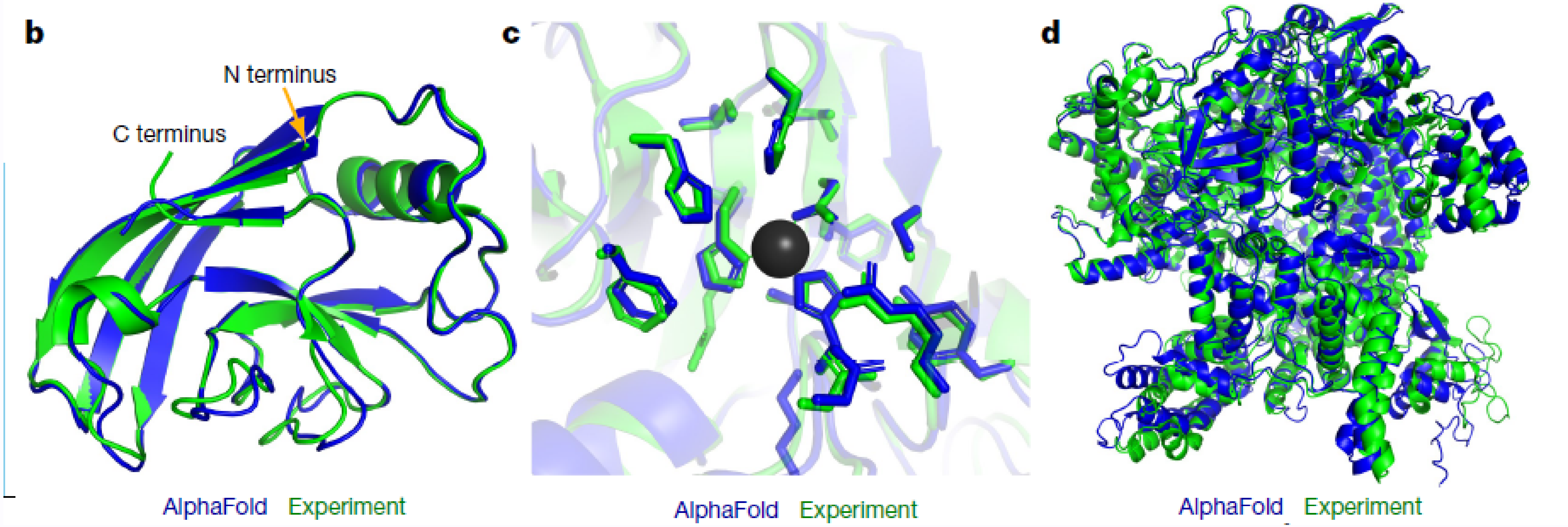
- By leveraging these data, various research groups have recently succeeded in developing DL models capable of predicting folding patterns based on amionacid sequences

AlphaFold

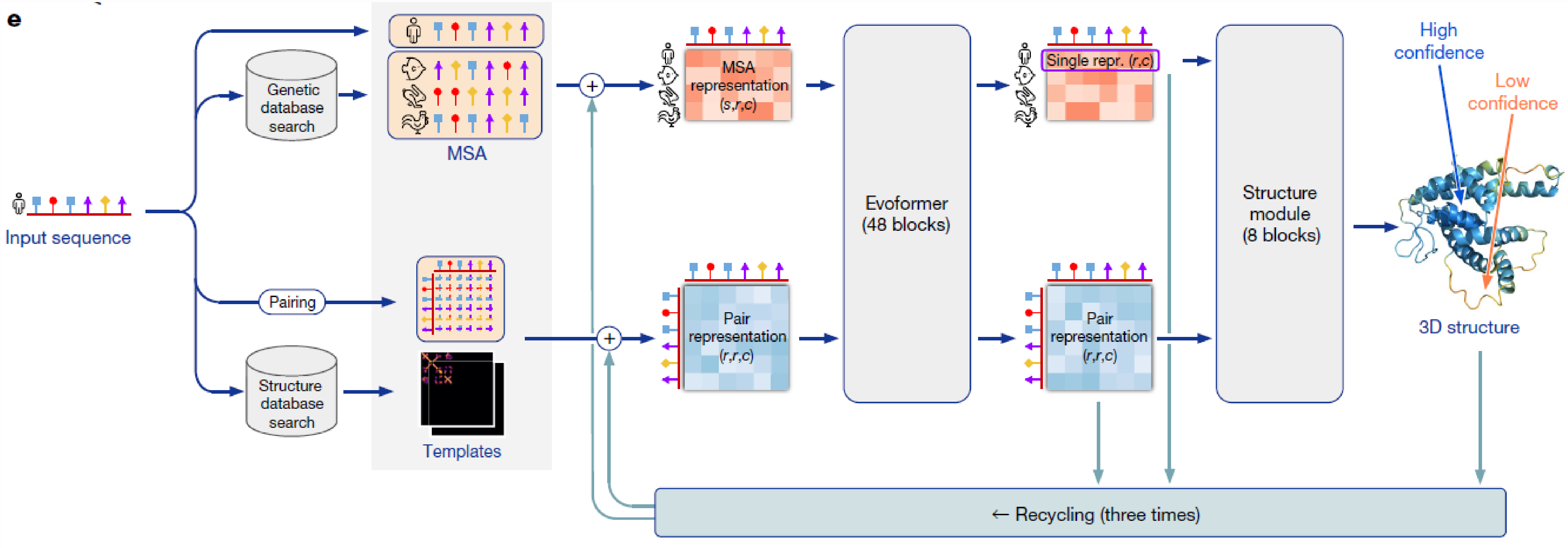
- AlphaFold is a DL system using advanced neural network architectures and training algorithms to predict protein structures with atomic accuracy[2]

- This model won the **Critical Assessment of Structure Prediction (CASP)** competition in 2020, a biennial event where international research groups compete to predict protein structures
  - This was the first time that a DL model outperformed all other experimental methods in this competition, with the potential of revolutionising protein structure research with more time- and resource-effective methods

Example of AlphaFold's predicted structures



AlphaFold's architecture


e
Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

Genomics

Genetics and health

- Most common diseases have a genetic component

- Genetic research in healthcare has traditionally focused on identifying rare monogenic mutations that are individually associated with a high risk of disease

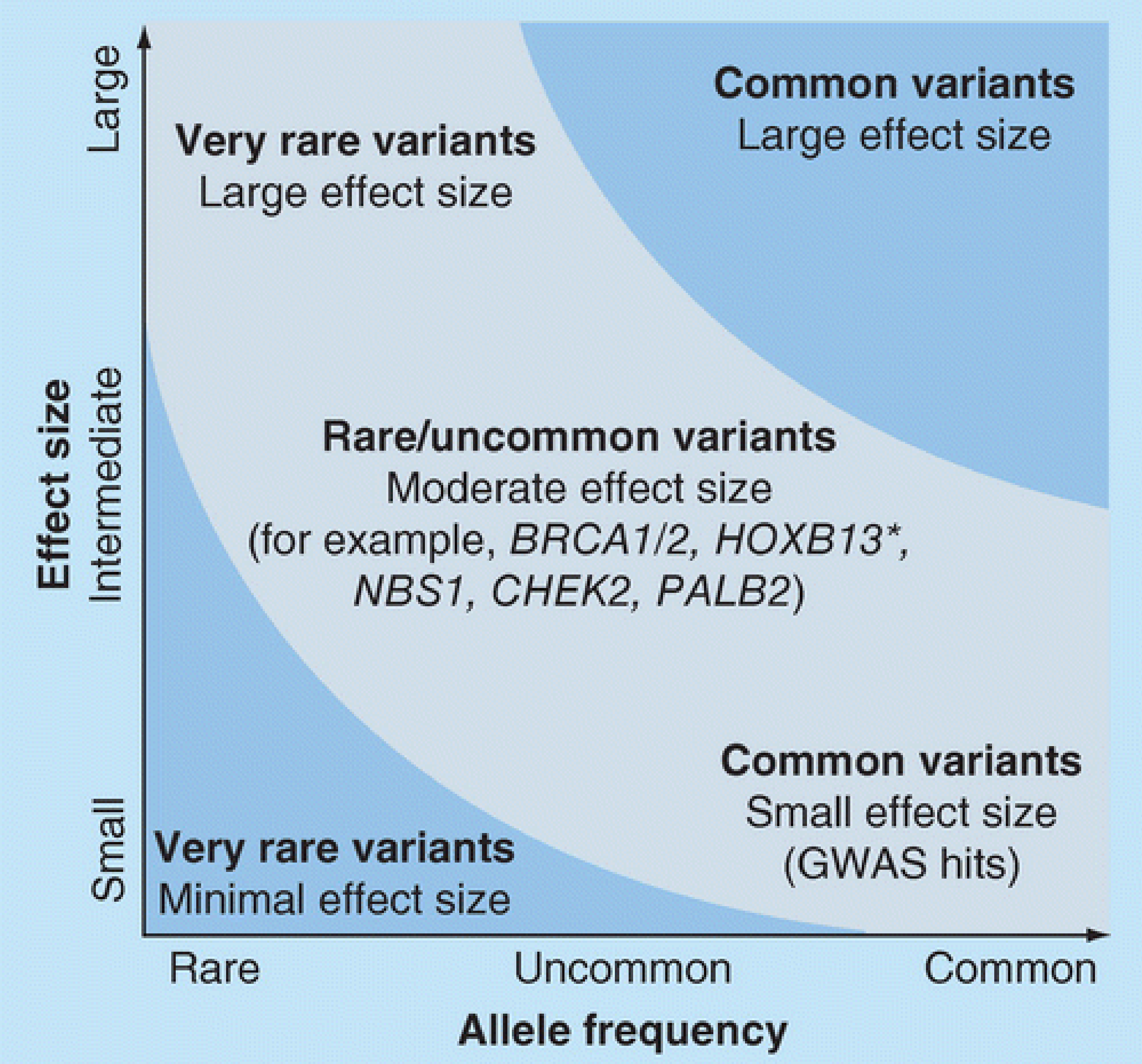
- However, in most cases, polygenic risk factors, involving many common genetic variants of small individual impact, play a greater role in disease risk than rare monogenic mutations[3]

Genome-wide association studies (GWAS)

- GWAS involve leveraging genetic data from large and diverse populations to identify genetic variants whose presence is significantly associated with a particular trait or disease of interest

- These studies became increasingly popular in the mid-2000s mainly due to the rise of high-throughput genotyping technologies and to the development of large, publicly available genetic databases

- Associated genetic variants identified via GWAS are typically large in number and most often found in non-coding DNA
  - Downstream analyses are required to properly assess their impact on disease risk

Effect size
Large
Intermediate
Small
Very rare variants
Large effect size
Common variants
Large effect size
Rare/uncommon variants
Moderate effect size
(for example, *BRCA1/2*, *HOXB13**, *NBS1*, *CHEK2*, *PALB2*)
Common variants
Small effect size
(GWAS hits)
Very rare variants
Minimal effect size
Rare
Uncommon
Common
Allele frequency

Modeling polygenic risk

- A popular model of polygenic risk is the weighted polygenic risk score (wPRS)
    - wPRS involves assigning weights to genetic variants based on their effect sizes from GWAS and linearly combining them to generate a risk score based on an individual's genetic makeup

- **The problem:** wPRS is prone to bias given its reliance on GWAS effect sizes. It also assumes that the effects of variants are additive while being incapable of modeling non-linear effects or interactions between variants

DL in polygenic risk modeling

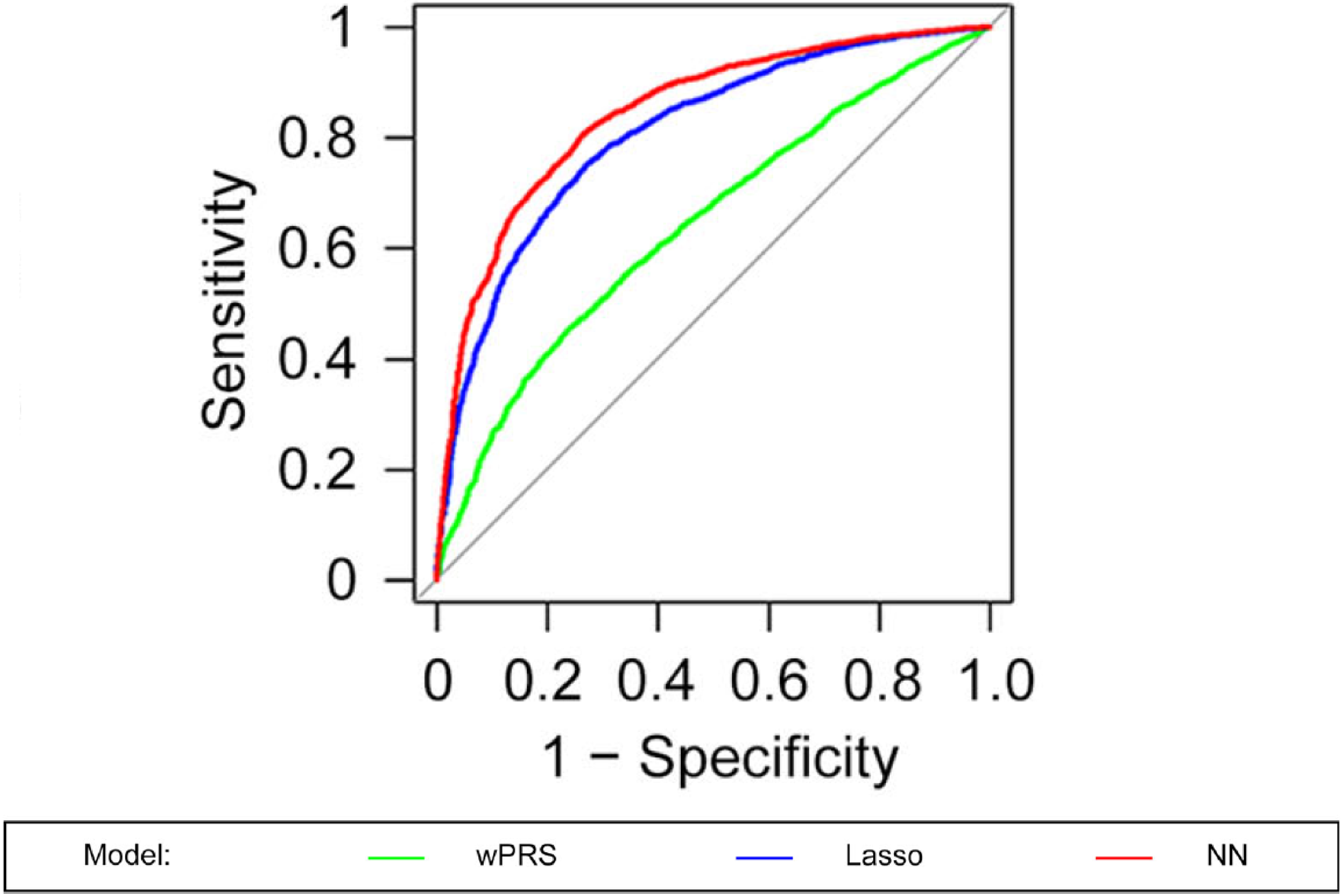
- Unlike wPRS, DL models have the capacity to describe complex relationships between GWAS genetic variants and a disease of interest, adapting well to high-dimensional genomic data

- Because polygenic risk is comprised of multiple genetic variants that are more common than high-risk monogenic mutations, better polygenic risk scoring methods enabled by DL could enable early interventions and improved outcomes for a wider variety of diseases and populations

Polygenic risk for Alzheimer's disease (AD)

- Previous GWAS have revealed that multiple genetic variants contribute to the risk of developing AD
    - Individually, most of these variants have small effects on AD risk, hinting at the need for polygenic risk modeling

- In 2023, Zhou et al. developed a fully connected neural network capable of ingesting thousands of genetic variants associated with AD
    - Their polygenic risk score largely outperformed wPRS, achieving AUROC scores of 0.84 and 0.64, respectively[4]


Sensitivity
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1.0
1 − Specificity
Model:
wPRS
Lasso
NN

Radiomics

Medical imaging and healthcare

- Medical imaging provides clinicians with non-invasive or minimally-invasive tools to visually inspect internal anatomy and function
    - X-ray imaging
    - Nuclear imaging (e.g., PET scans)
    - Ultrasound
    - Magnetic resonance imaging (MRI)
    - Optical coherence tomography (OCT)

- In many cases, these techniques enable the early detection, diagnosis, and treatment of many diseases that could otherwise be fatal or incur high morbidity

Recent advances in medical image acquisition

- Given the value of medical imaging for disease detection and management, there is continued interest in advancing medical imaging acquisition techniques
    - This includes improvements in digitisation, resolution, portability, acquisition times, and variety

- Simultaneously, efforts have been led to establish vast, open-source imaging databanks to accelerate research in various disease contexts

Medical image analysis

- In clinic, for many disease contexts, visual inspections performed by a radiologist remain the gold standard for medical image analysis

- **The problem:** these conventional reads fail fully to exploit modern medical image acquisition technologies. Subtle signals captured in high resolution images may not be appreciated by the naked eye, but may represent consistent markers of disease

DL and its success in radiomics

- Radiomics is a field that applies advanced computational methods to extract quantitative features from medical images to ultimately characterise tissue properties

- DL tools, in particular convolutional neural networks (CNNs), have seen increasing popularity in radiomics research given their stellar track record in natural image processing and analysis over the last decade
  - The success of CNNs in these two contexts that would initially seem completely different lies on their proficiency in extracting spatial features from data, which form the foundation of all visual tasks like object segmentation and semantic classification

Case study: neuropsychiatric lupus (NPSLE)

- Systemic lupus erythematosus (SLE) is a chronic, multisystem autoimmune disease that can affect the nervous system
    - Manifestations of neuropsychiatric lupus (NPSLE) are highly heterogeneous, ranging from mild headaches to psychosis

- Underlying mechanisms of this rare disease are unknown, and recommended diagnostic protocols involve a lengthy, costly, subjective, and non-specific ruling out process
    - Part of this workup involves a conventional read of a structural T1-weighted MRI scan, which fails to identify brain abnormalities in over 50% of lupus patients with neuropsychiatric symptoms

Self-supervised CNN for NPSLE classification

- In 2022, Inglese et al. developed a self-supervised CNN for NPSLE classification[5]

- Given the rarity of this disease, the group faced data scarcity challenges, having access to just 163 T1-weighted MRI images

- Despite these low numbers and the lack of abnormal signals in the MRI scans as probed by the naked eye, their rather simple CNN reached an average classification performance of 83%
  - This was possible due to their use of contrastive learning and augmentation methods

Inglese's self-supervised CNN architecture

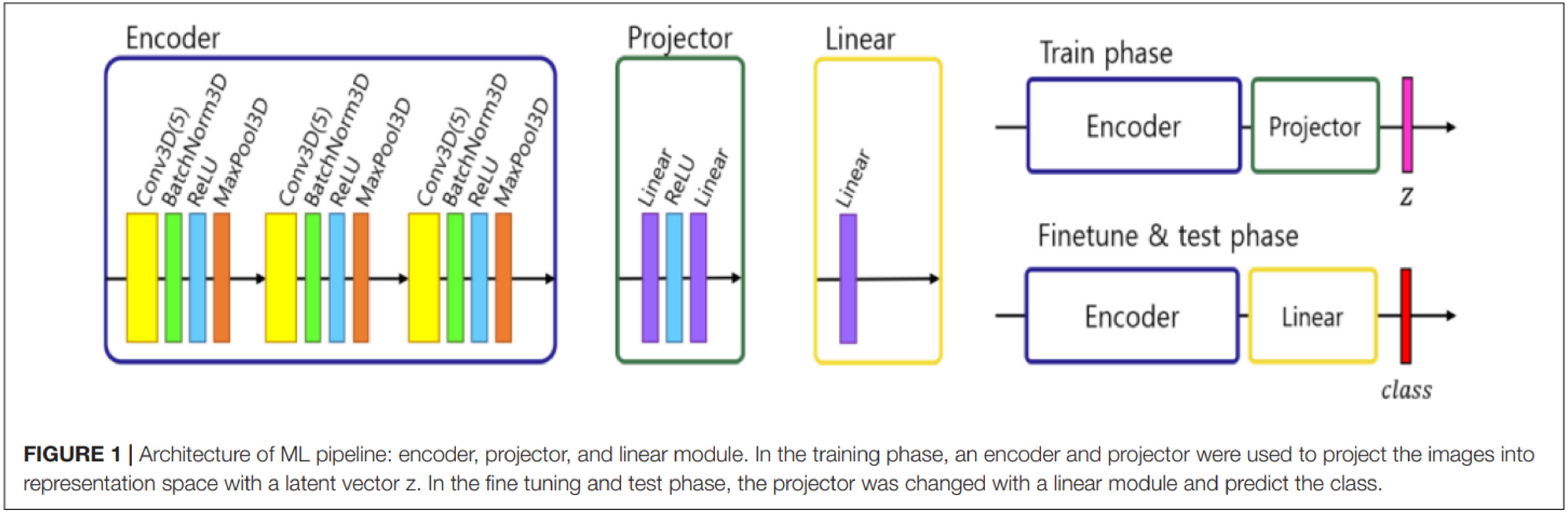


FIGURE 1 | Architecture of ML pipeline: encoder, projector, and linear module. In the training phase, an encoder and projector were used to project the images into representation space with a latent vector z. In the fine tuning and test phase, the projector was changed with a linear module and predict the class.

Conclusion

Popularity and success of DL in healthcare research

- As depicted by the previous examples, DL has become a popular tool in healthcare-research across a variety of contexts

- With recent advances in health-related data collection and the establishment of vast digital data repositories, it was only a matter of time before DL started to outshine alternative data analysis methods, given its complex modeling and scaling capabilities

- Next, we will examine how institutions and businesses are working to turn these research tools into commerciable applications

References

(1) Jimma, B. L. (2023). Artificial intelligence in healthcare: A bibliometric analysis. Telematics and Informatics Reports, 9, 100041-. https://doi.org/10.1016/j.teler.2023.100041

(2) Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2

(3) Khera, A.V., Chaffin, M., Aragam, K.G. et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet 50, 1219–1224 (2018). https://doi.org/10.1038/s41588-018-0183-z

(4) Zhou, X., Chen, Y., Ip, F.C.F. et al. Deep learning-based polygenic risk analysis for Alzheimer's disease prediction. Commun Med 3, 49 (2023). https://doi.org/10.1038/s43856-023-00269-x

(5) Inglese, F., Kim, M., Steup-Beekman, G. M., Huizinga, T. W. J., van Buchem, M. A., de Bresser, J., Kim, D.-S., & Ronen, I. (2022). MRI-Based Classification of Neuropsychiatric Systemic Lupus Erythematosus Patients With Self-Supervised Contrastive Learning. Frontiers in Neuroscience, 16, 695888–695888. https://doi.org/10.3389/fnins.2022.695888