Workshop in Systems Immunology

BOSTON
JUNE 20-23 2023
FOCIS
ANNUAL MEETING

Emanuele de Rinaldis
Magnus Fontes
Shameer Khader
Giorgio Gaglia

June 19th 2023

Today's Plan

8:00-8:10 am	Course Overview and Objectives – Emanuele de Rinaldis, PhD
8:10-9:00 am	Introduction to Systems Immunology – Emanuele de Rinaldis, PhD
9:00-9:45 am	Systems Immunology & Immune Oncology: A Data-Centric View – Magnus Fontes, PhD
9:45-10:00 am	Break
10:00-11:30 am	Deep Dive Into Selected Scientific Case Studies: From Systems Immunology to Novel Therapeutic Insights – Emanuele de Rinaldis, PhD
11:30 am-12:00 pm	Q/A and Panel Discussion
12:00-1:00 pm	Break for Lunch
1:00-2:00 pm	Spatial Biology Methods and Analytics for Immunology & Oncology – Giorgio Gaglia, PhD
2:00-2:15 pm	Break
2:15-3:30 pm	Artificial Intelligence – A Primer for Immunologists – Shameer Khader, PhD, MPH
3:30-3:45 pm	Break
3:45-4:45 pm	Interactive Data Analysis Session – Magnus Fontes, PhD
4:45-5:00 pm	Wrap Up Notes & Final Remarks

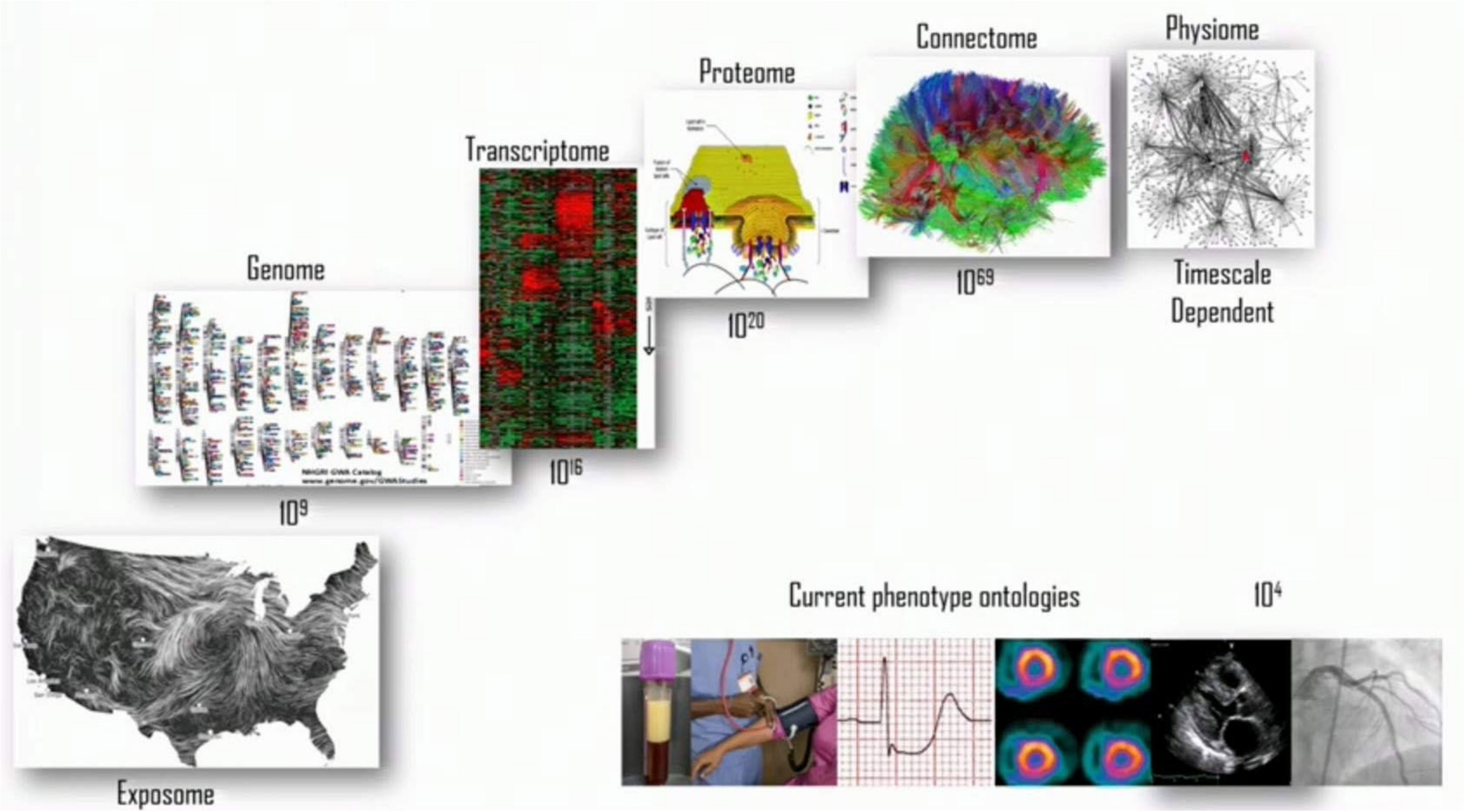
2:15-3:30

AI for Immunologists – An Introduction

Outline of today’s session

- Background
- Data boom in biology and need for AI
- Examples of AI in Immunology
  - **Classical ML and Predictive models – Open Targets / Target Immune Engine**
  - **Graph ML – AsthmaGraph (Poster at FOCIS!)**
  - **Emerging themes in AI: Encoders, Embedding, Transformers, GANs, and LLMs**
- Future outlook

On a mission to close the data and inference gap in biomedicine



Genome

10^9

Transcriptome

10^{16}

Proteome

10^{20}

Connectome

10^{69}

Physiome

Timescale Dependent

Exposome

Current phenotype ontologies

10^4

From Calium MacRae / https://twitter.com/daniel_kraft/status/1011692279445123072

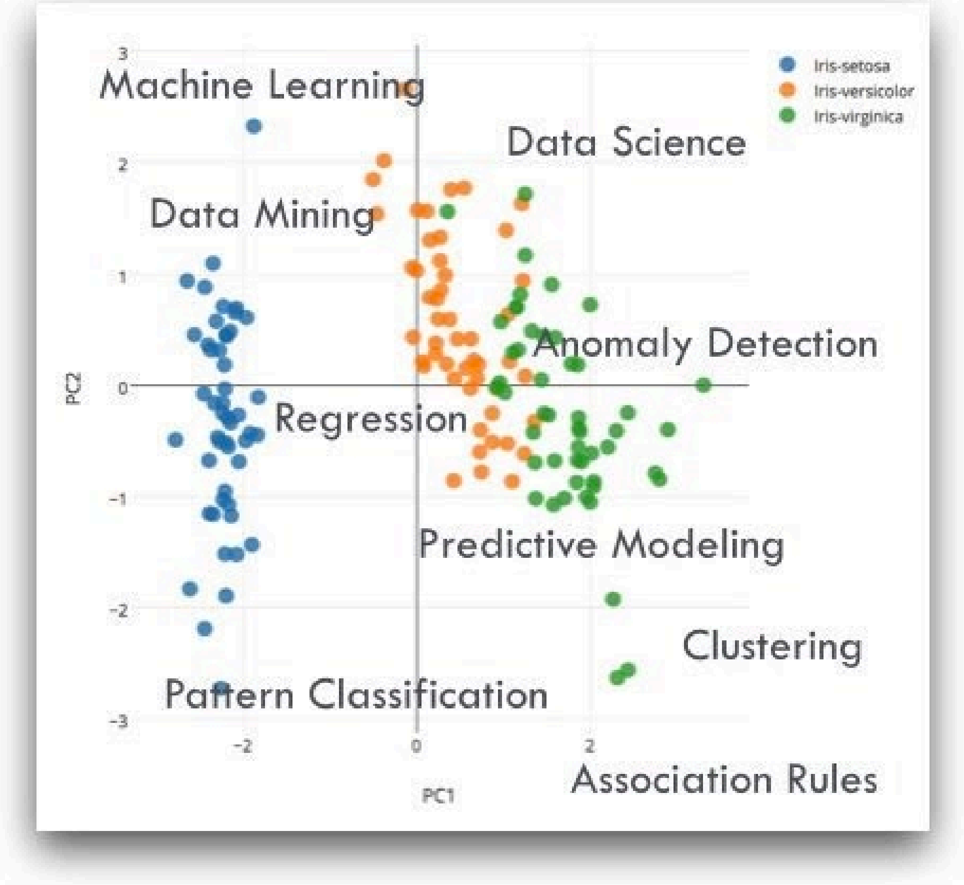
Preamble

Big data *(noun) extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions.*

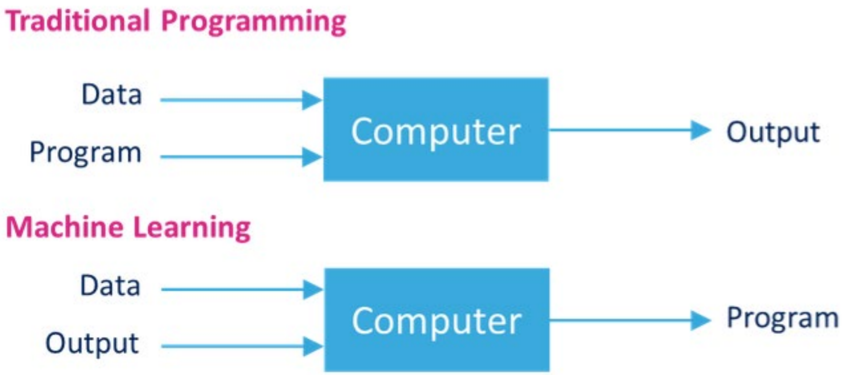
Predictive analytics *is the area of data mining concerned with forecasting probabilities and trends.*

Data science *is an interdisciplinary field about processes and systems to extract knowledge or insights from data.*

Artificial intelligence (AI) *is wide-ranging branch of computer science concerned with building smart machines capable of performing tasks that typically require human intelligence.*



What is Artificial Intelligence?



Traditional Programming

Data

Program

Computer

Output

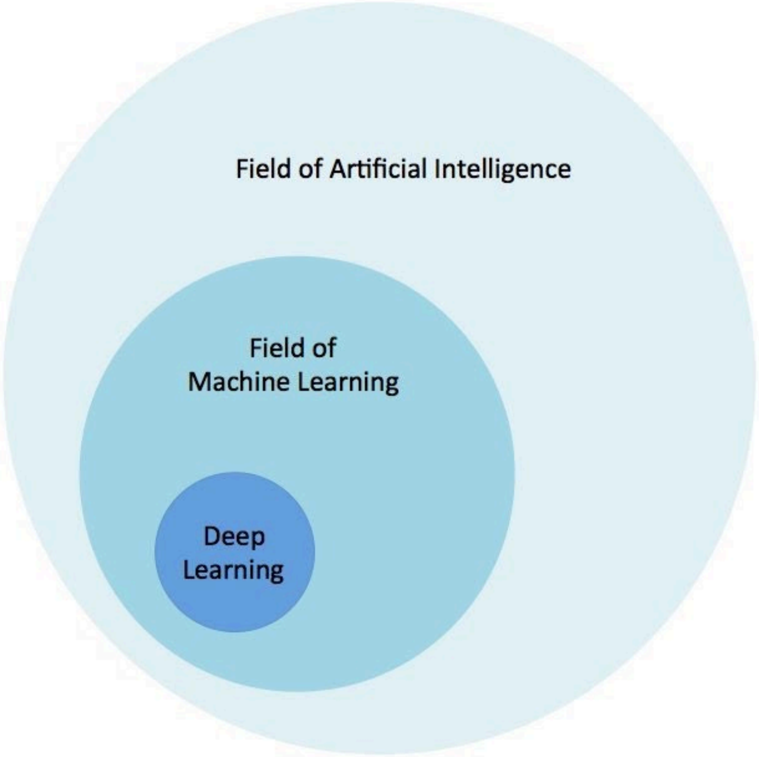
Machine Learning

Data

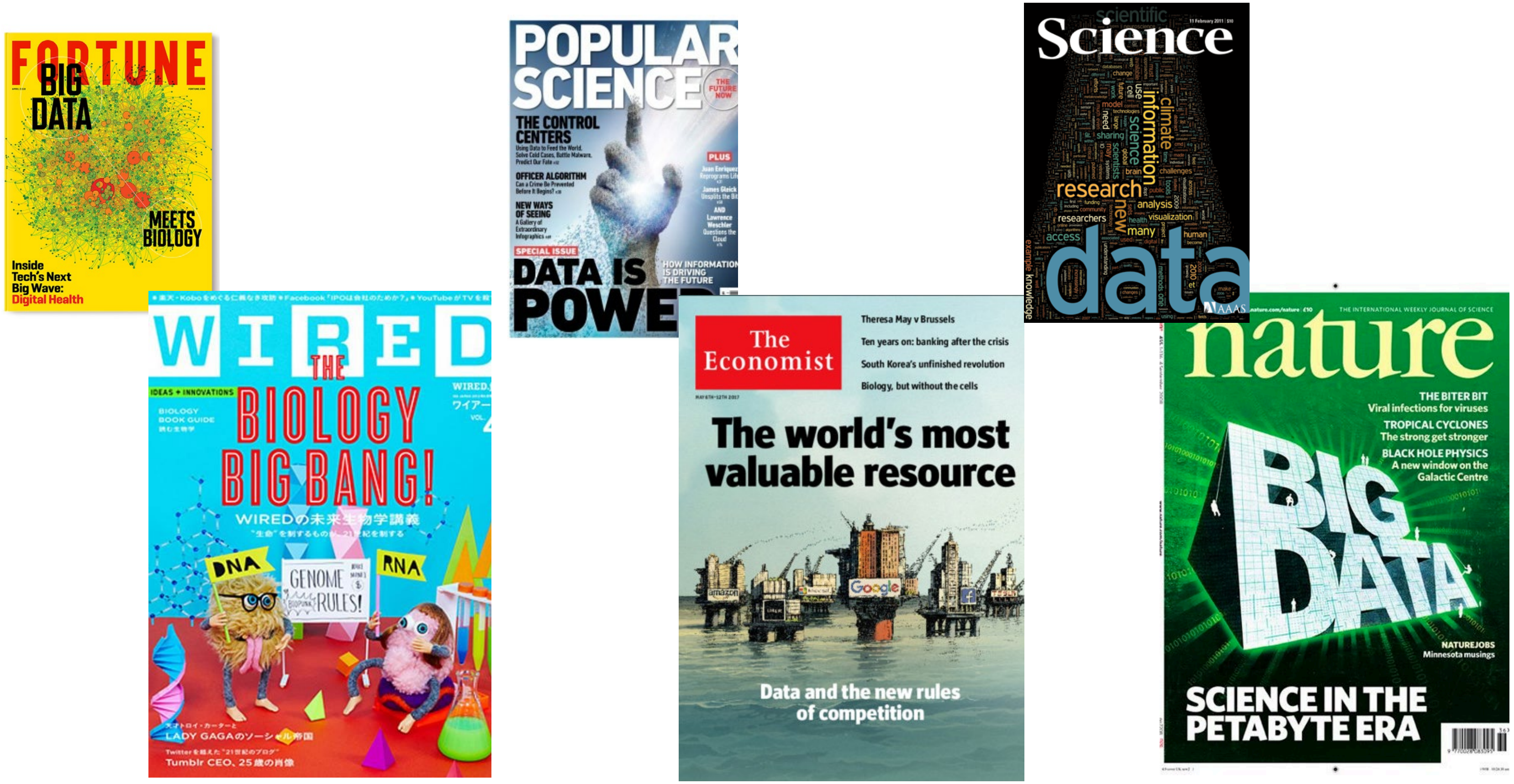
Output

Computer

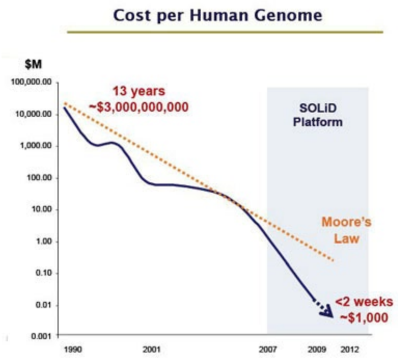
Program

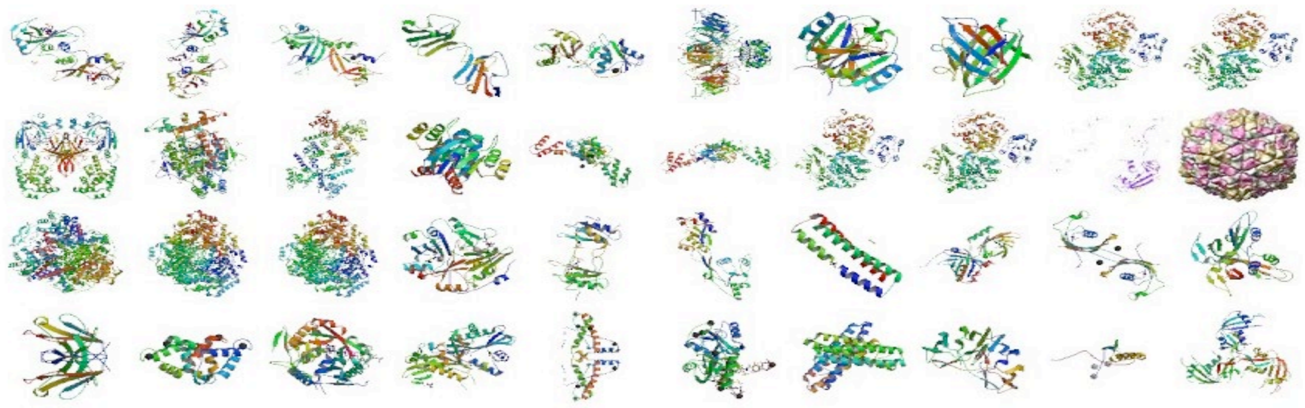


Big Data Big Bang in Biomedicine


FORTUNE
BIG DATA
MEETS BIOLOGY
Inside Tech's Next Big Wave: Digital Health
POPULAR SCIENCE
THE CONTROL CENTERS
Using Data to Feed the World, Solve Cold Cases, Battle Malware, Predict Our Fate
OFFICER ALGORITHM
Can a Crime Be Prevented Before It Begins?
NEW WAYS OF SEEING
A Gallery of Extraordinary Infographics
SPECIAL ISSUE
DATA IS POWE
HOW INFORMATION IS DRIVING THE FUTURE
Science
data
research
information
WIRED
IDEAS + INNOVATIONS
THE BIOLOGY BIG BANG!
WIREDの未来生物学講義
DNA
GENOME RULES!
RNA
LADY GAGAのソーシャル帝国
Tumblr CEO、25歳の肖像
The Economist
Theresa May v Brussels
Ten years on: banking after the crisis
South Korea's unfinished revolution
Biology, but without the cells
The world's most valuable resource
Data and the new rules of competition
nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
THE BITER BIT
Viral infections for viruses
TROPICAL CYCLONES
The strong get stronger
BLACK HOLE PHYSICS
A new window on the Galactic Centre
BIG DATA
NATUREJOBS
Minnesota musings
SCIENCE IN THE PETABYTE ERA

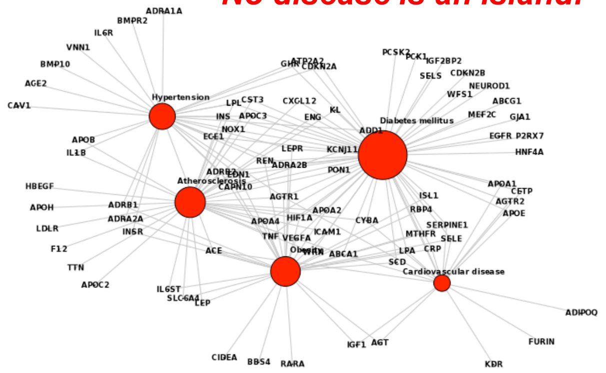
Biomedical data is complex and natively digital

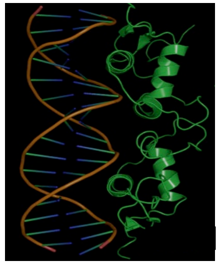




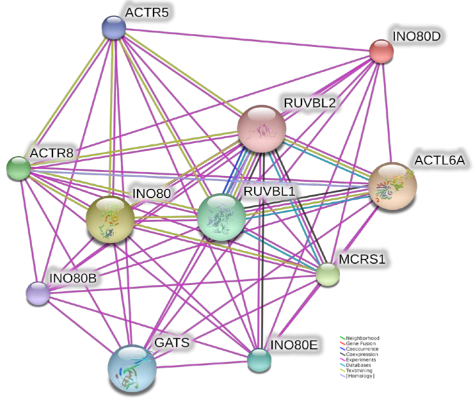
Disease networks

No disease is an island!

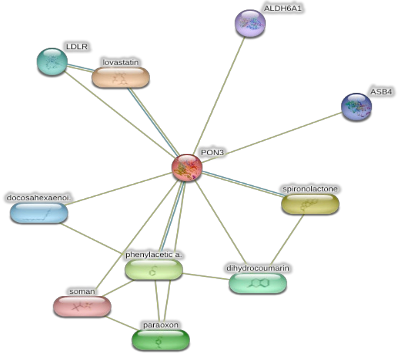




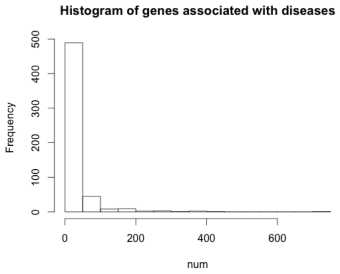
Protein-Nucleic acid



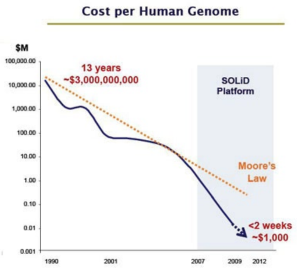
Protein-Protein

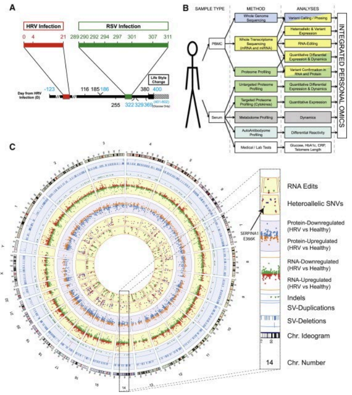


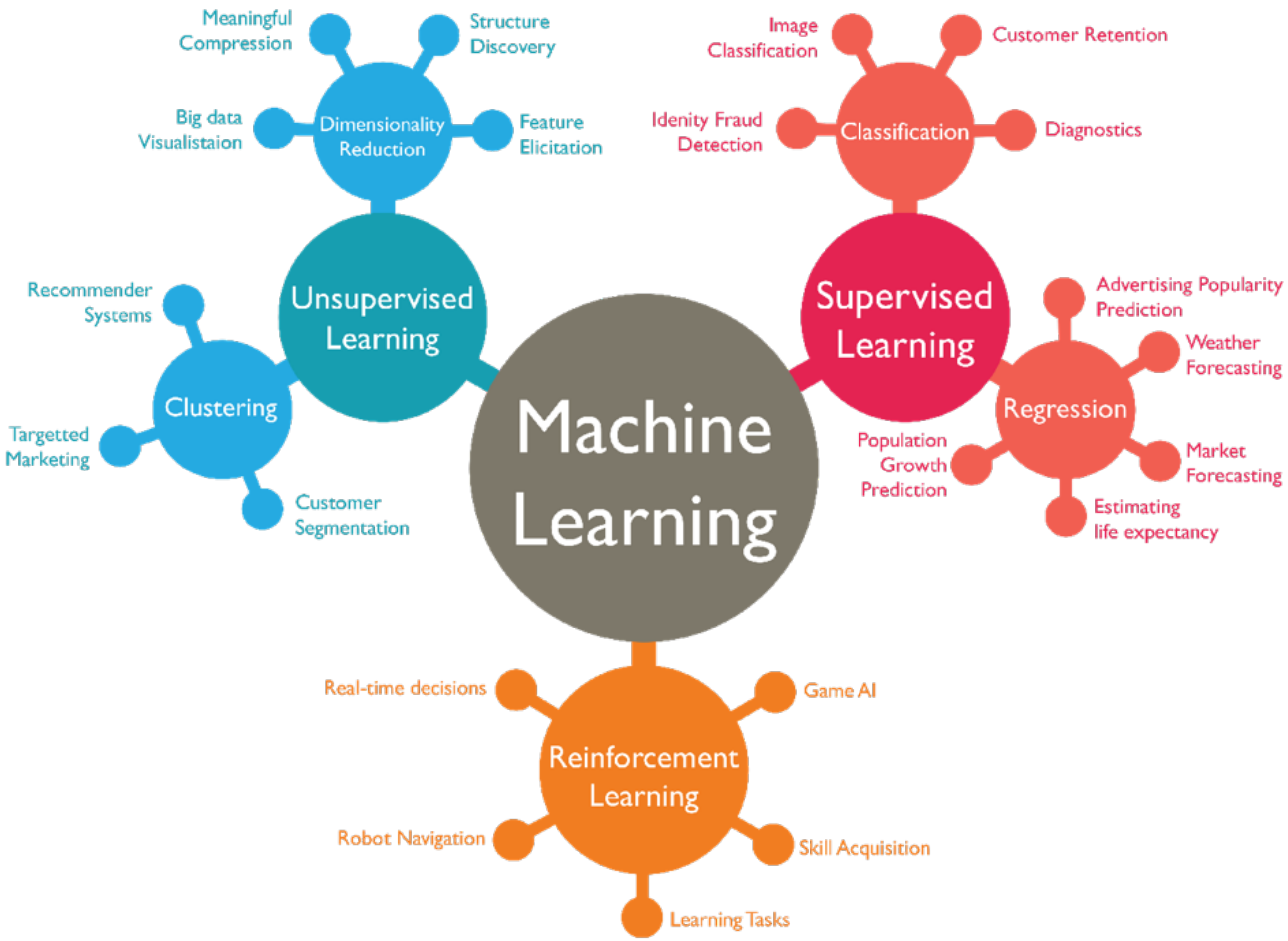
Protein-Small molecule



Why AI in Biomedicine?





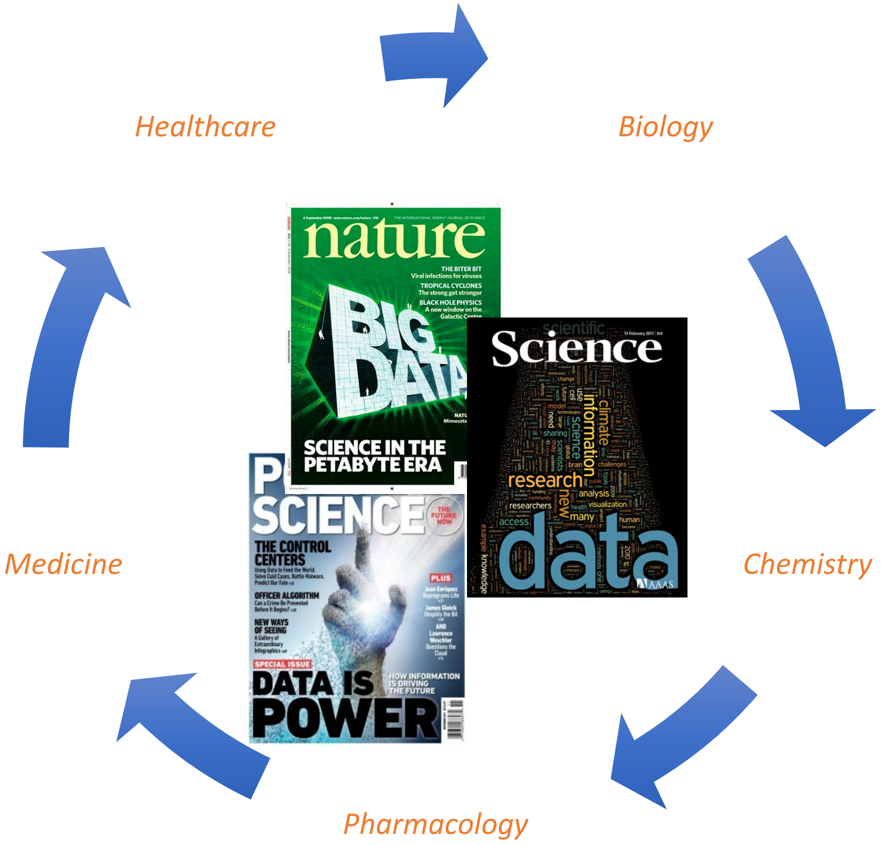


AI in Biomedicine: Convergence of Big Data, Predictive Modeling, Data science and AI to design, develop and deliver Precision Medicine solutions

From Biology to Therapy to Healthcare via Data & AI

Medicine

- Traditional data types
- Centralized
- GBs or TBs in size
- Structured
- Stable data model
- Low-dimensional
- Statistical approaches
- Cohort size (~10K)
- Hypothesis-driven



Healthcare

Biology

Chemistry

Pharmacology

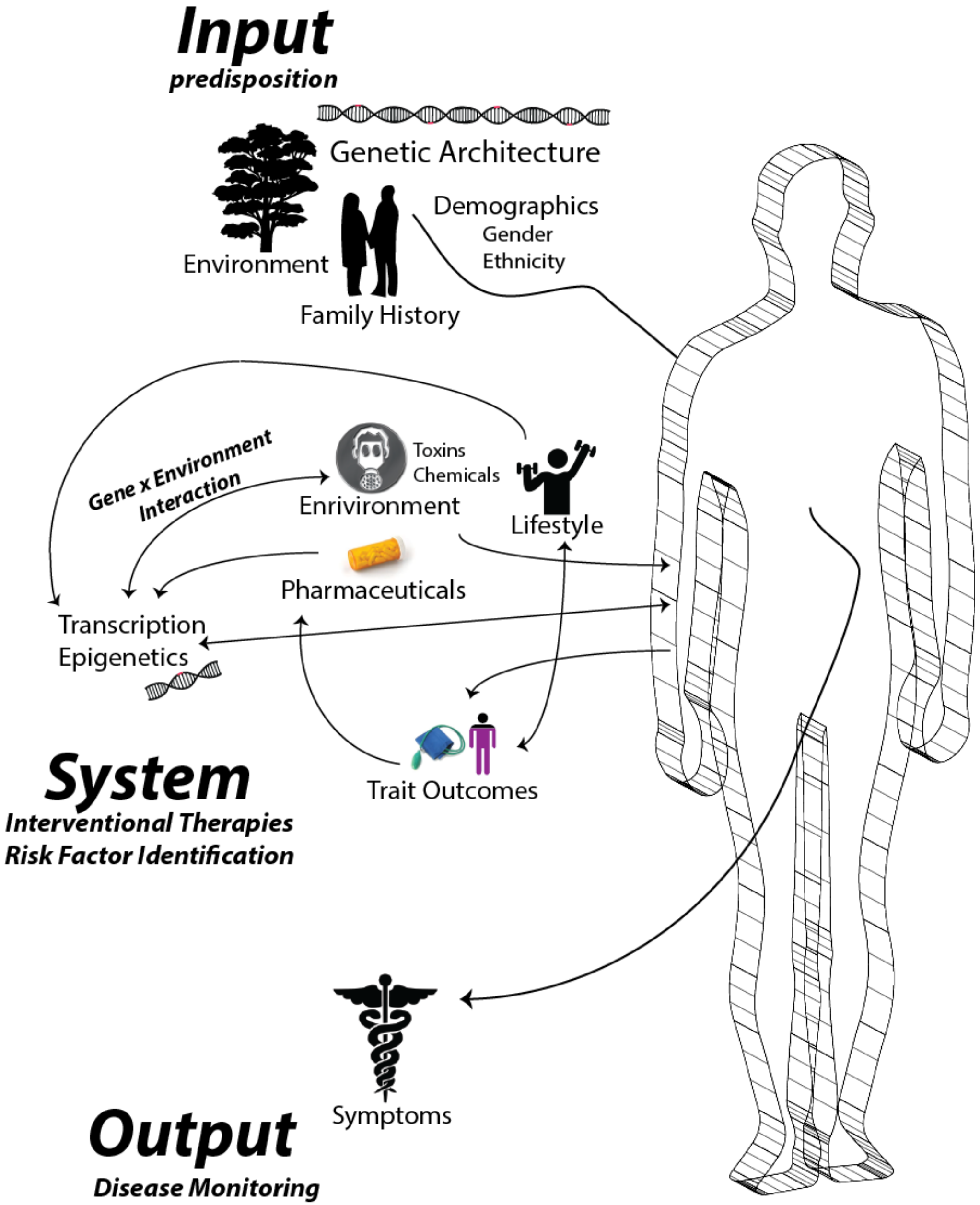
Medicine

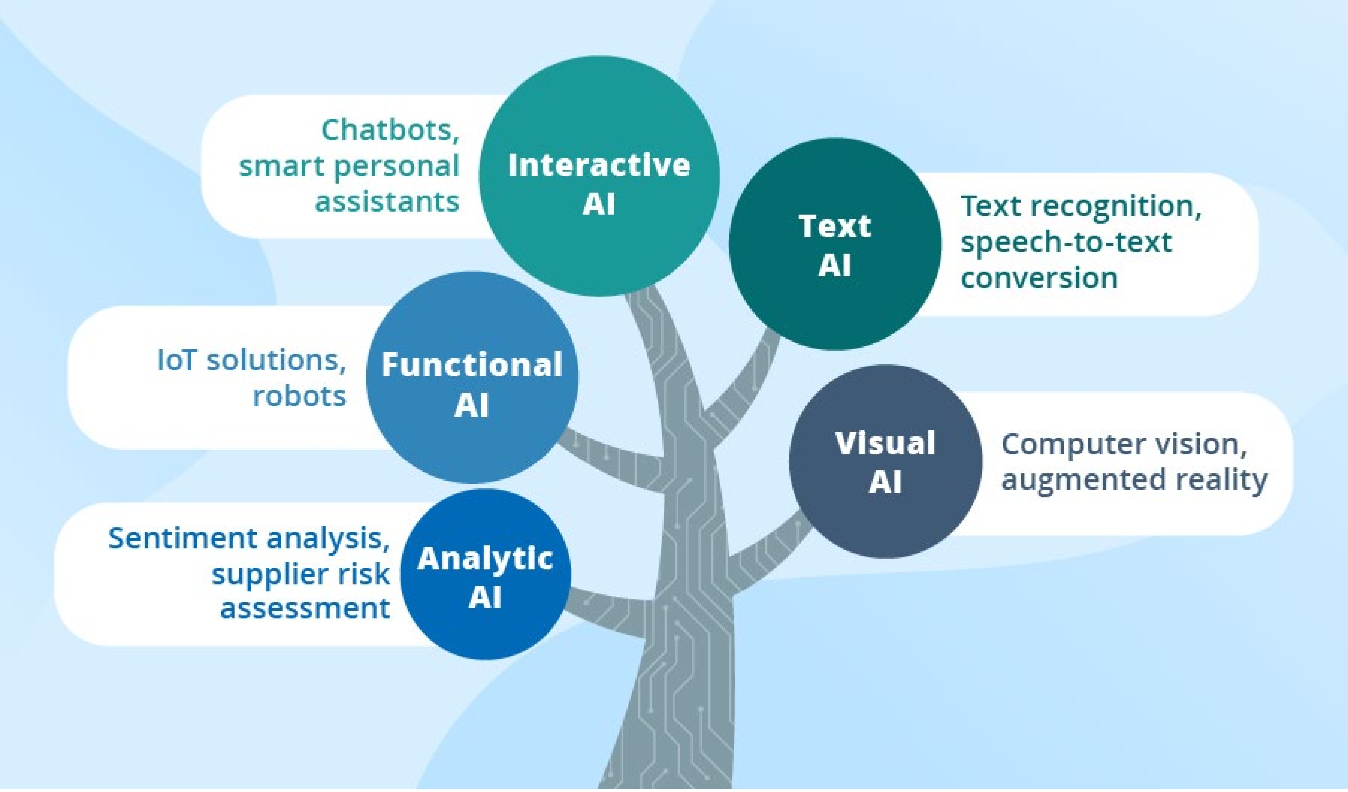
Data-driven Precision Medicine

- Evolving data types
- Decentralized
- Petabytes, exabytes…
- Semi or unstructured
- Evolving, flat data model
- High-dimensional
- Machine or deep learning
- Large cohort size (>10K)
- Data-driven

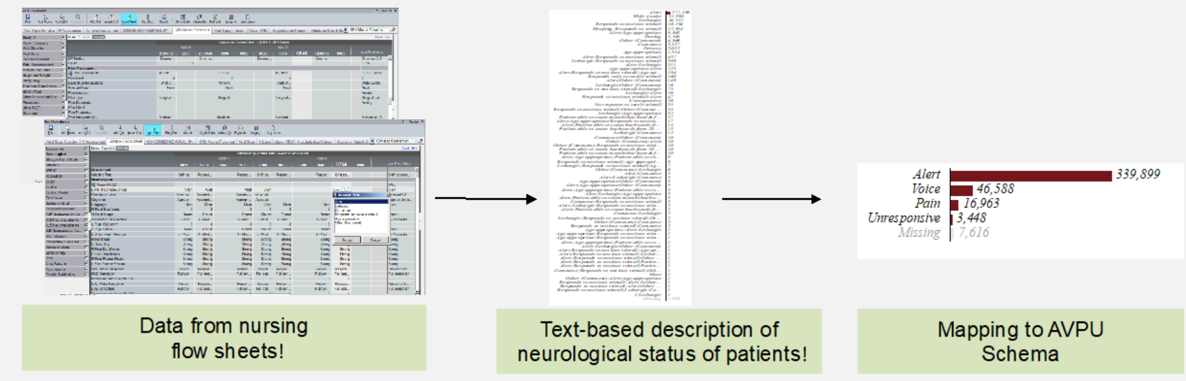
K. Shameer et. al(2018) Machine learning in Cardiovascular Medicine, Are we there yet? BMJ Heart pii: heartjnl-2017-311198. doi: 10.1136/heartjnl-2017-311198.

Different facets of AI




Chatbots, smart personal assistants
Interactive AI
Text AI
Text recognition, speech-to-text conversion
IoT solutions, robots
Functional AI
Visual AI
Computer vision, augmented reality
Sentiment analysis, supplier risk assessment
Analytic AI

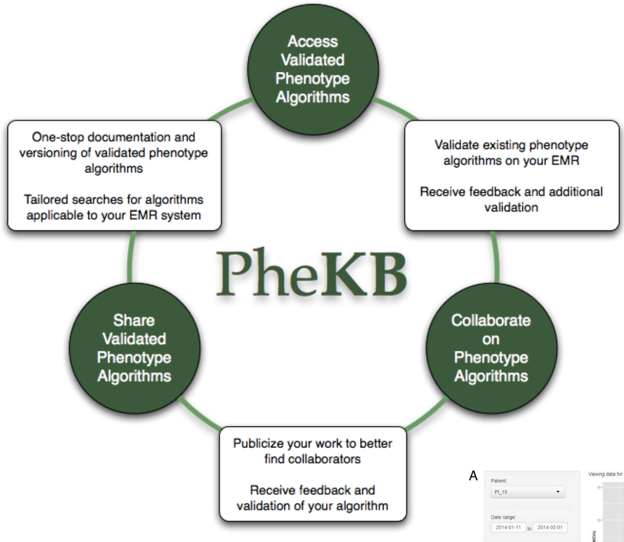
Enablers of AI



Data from nursing flow sheets!

Text-based description of neurological status of patients!

Mapping to AVPU Schema



Access Validated Phenotype Algorithms

Validate existing phenotype algorithms on your EMR

Receive feedback and additional validation

Collaborate on Phenotype Algorithms

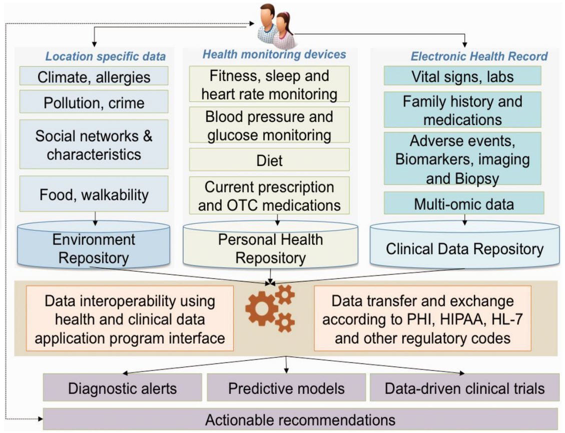
Publicize your work to better find collaborators

Receive feedback and validation of your algorithm

Share Validated Phenotype Algorithms

One-stop documentation and versioning of validated phenotype algorithms

Tailored searches for algorithms applicable to your EMR system



Location specific data

- Climate, allergies
- Pollution, crime
- Social networks & characteristics
- Food, walkability
- Environment Repository

Health monitoring devices

- Fitness, sleep and heart rate monitoring
- Blood pressure and glucose monitoring
- Diet
- Current prescription and OTC medications
- Personal Health Repository

Electronic Health Record

- Vital signs, labs
- Family history and medications
- Adverse events, Biomarkers, imaging and Biopsy
- Multi-omic data
- Clinical Data Repository

Data interoperability using health and clinical data application program interface

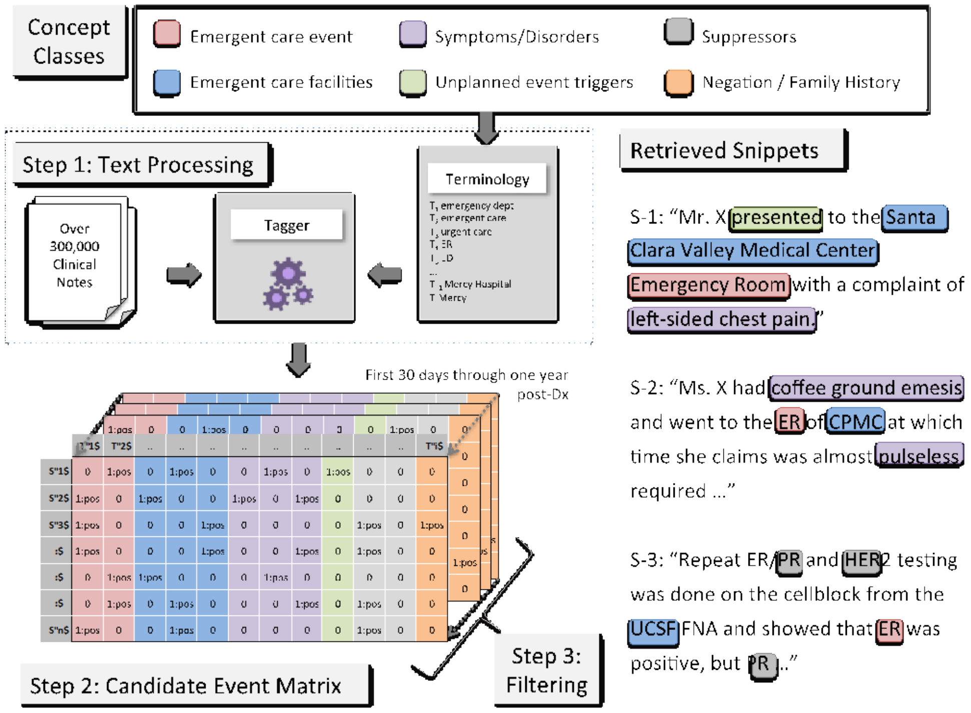
Data transfer and exchange according to PHI, HIPAA, HL-7 and other regulatory codes

Diagnostic alerts

Predictive models

Data-driven clinical trials

Actionable recommendations



Concept Classes

🟥 Emergent care event

🟪 Symptoms/Disorders

⬜ Suppressors

🟦 Emergent care facilities

🟩 Unplanned event triggers

🟧 Negation / Family History

Step 1: Text Processing

Over 300,000 Clinical Notes

Tagger

Terminology

Step 2: Candidate Event Matrix

First 30 days through one year post-Dx

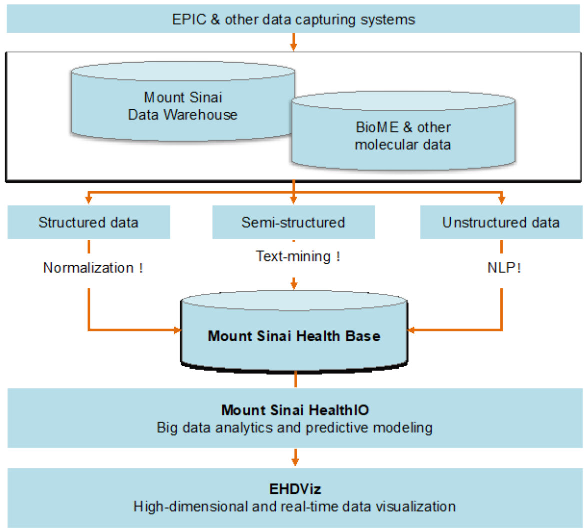
Step 3: Filtering

Retrieved Snippets

S-1: "Mr. X presented to the Santa Clara Valley Medical Center Emergency Room with a complaint of left-sided chest pain."

S-2: "Ms. X had coffee ground emesis and went to the ER of CPMC at which time she claims was almost pulseless required ..."

S-3: "Repeat ER/PR and HER2 testing was done on the cellblock from the UCSF FNA and showed that ER was positive, but PR .."



EPIC & other data capturing systems

Mount Sinai Data Warehouse

BioME & other molecular data

Structured data

Semi-structured

Unstructured data

Normalization !

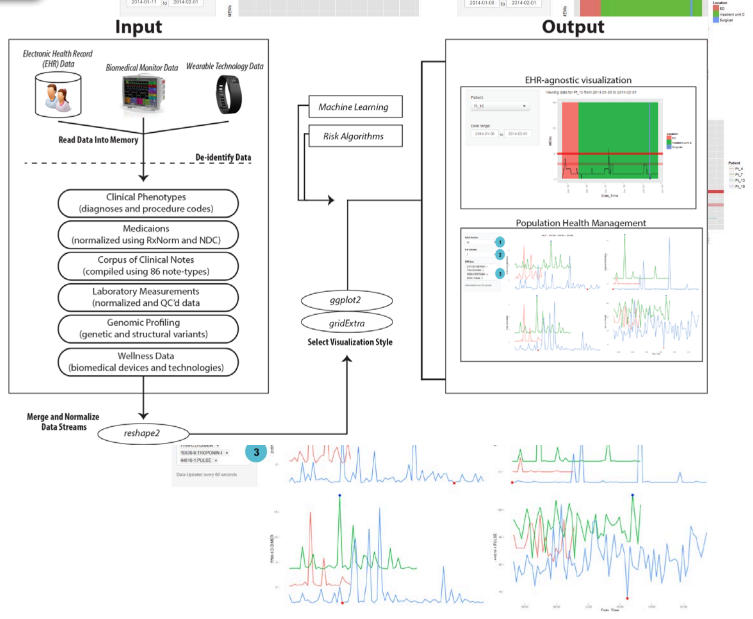
Text-mining !

NLP!

Mount Sinai Health Base

Mount Sinai HealthIO
Big data analytics and predictive modeling

EHDViz
High-dimensional and real-time data visualization



Input

Output

Clinical Phenotypes (diagnoses and procedure codes)

Medicaions (normalized using RxNorm and NDC)

Corpus of Clinical Notes (compiled using 86 note-types)

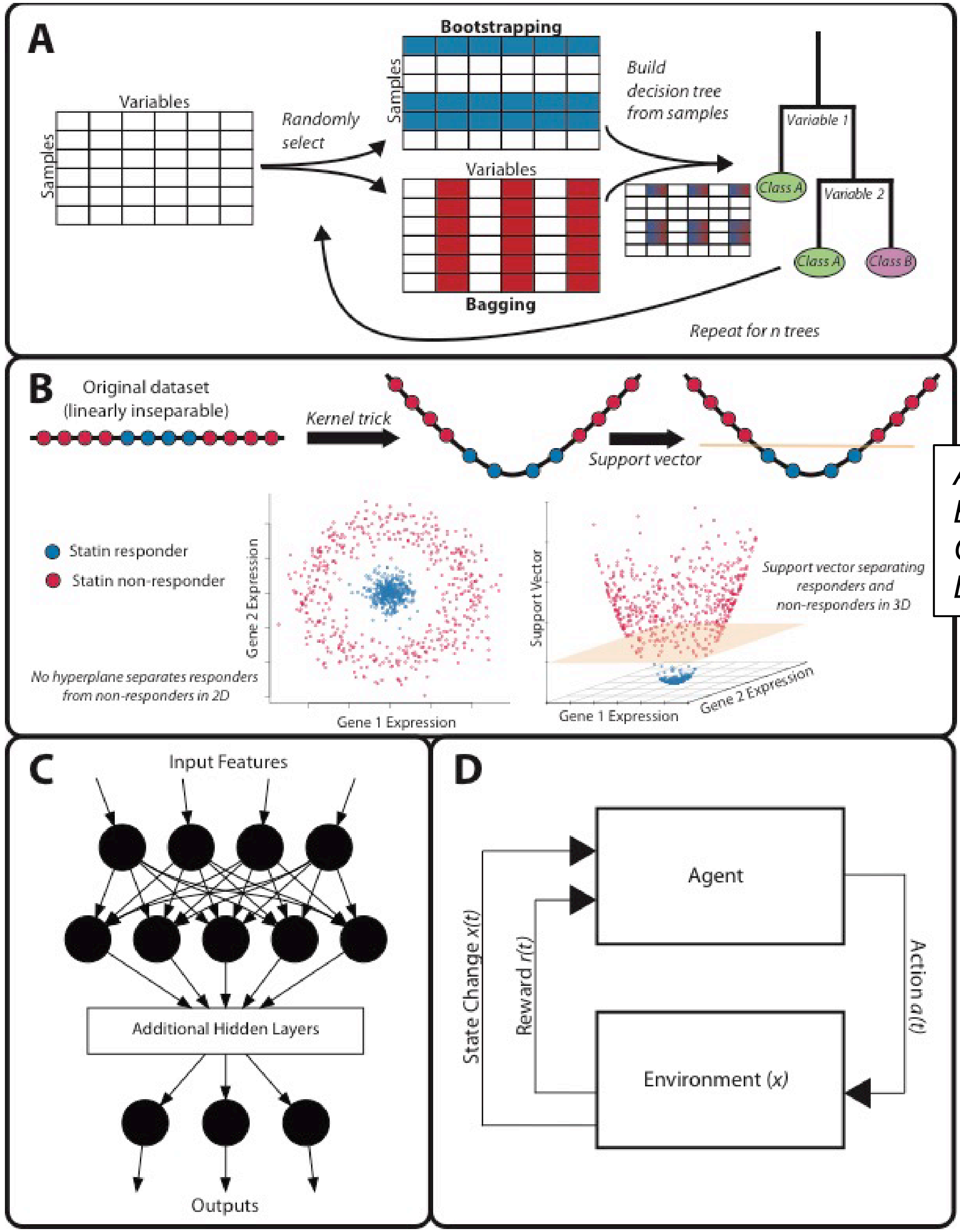
Laboratory Measurements (normalized and QC'd data

Genomic Profiling (genetic and structural variants)

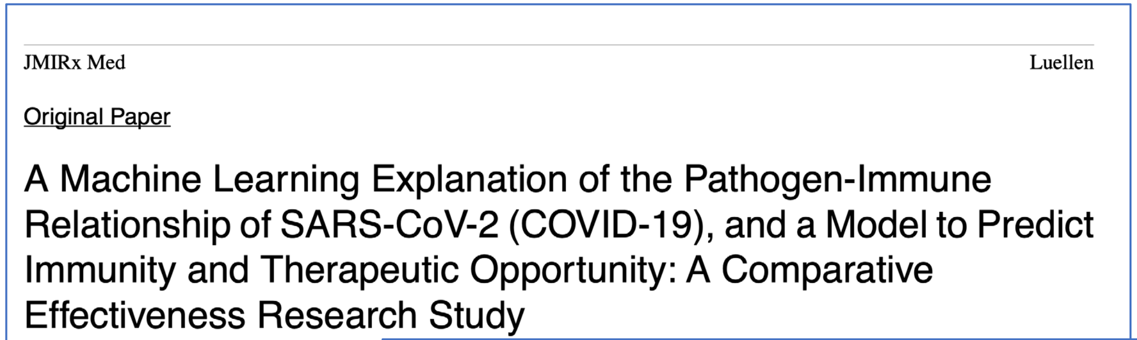
Wellness Data (biomedical devices and technologies)

Badgeley MA & Shameer K, et.al; . BMJ Open. 2016 Mar 24;6(3):e010579. doi: 10.1136/bmjopen-2015-010579
Brief Bioinform. 2016 Feb 14. pii: bbv118. PubMed PMID: 26876889.

Key AI methods and Applications in Immunology



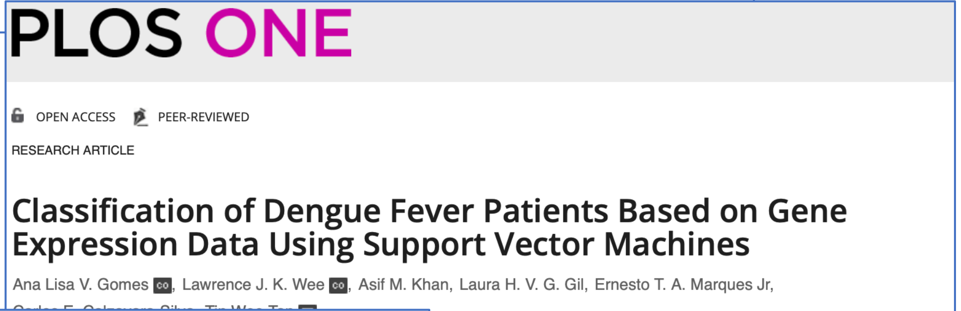
A) Random Forests
B) Support Vector Machines
C) Convolutional Neural Network
D) Reinforcement leaning



JMIRx Med Luellen

Original Paper

A Machine Learning Explanation of the Pathogen-Immune Relationship of SARS-CoV-2 (COVID-19), and a Model to Predict Immunity and Therapeutic Opportunity: A Comparative Effectiveness Research Study



PLOS ONE

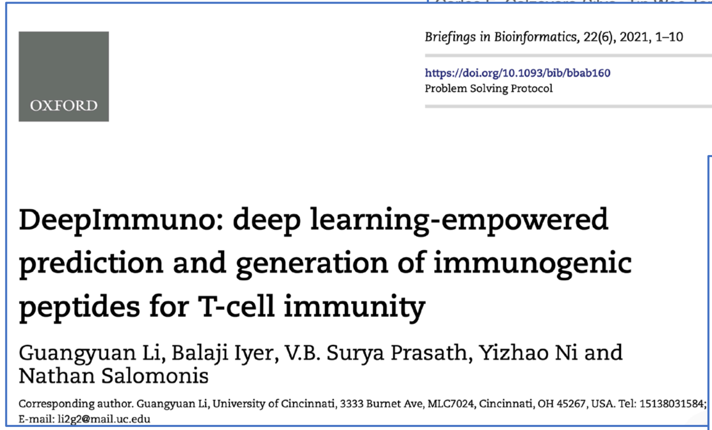
OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Classification of Dengue Fever Patients Based on Gene Expression Data Using Support Vector Machines

Ana Lisa V. Gomes, Lawrence J. K. Wee, Asif M. Khan, Laura H. V. G. Gil, Ernesto T. A. Marques Jr,

pone.0011267



OXFORD

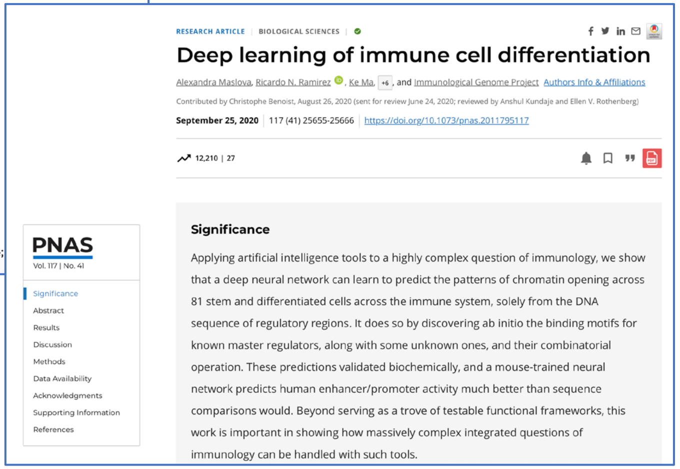
Briefings in Bioinformatics, 22(6), 2021, 1–10

https://doi.org/10.1093/bib/bbab160
Problem Solving Protocol

DeepImmuno: deep learning-empowered prediction and generation of immunogenic peptides for T-cell immunity

Guangyuan Li, Balaji Iyer, V.B. Surya Prasath, Yizhao Ni and Nathan Salomonis

Corresponding author. Guangyuan Li, University of Cincinnati, 3333 Burnet Ave, MLC7024, Cincinnati, OH 45267, USA. Tel: 15138031584; E-mail: li2g2@mail.uc.edu



RESEARCH ARTICLE | BIOLOGICAL SCIENCES |

Deep learning of immune cell differentiation

Alexandra Maslova, Ricardo N. Ramirez, Ke Ma, +6, and Immunological Genome Project Authors Info & Affiliations

Contributed by Christophe Benoist, August 26, 2020 (sent for review June 24, 2020; reviewed by Anshul Kundaje and Ellen V. Rothenberg)

September 25, 2020 | 117 (41) 25655-25666 | https://doi.org/10.1073/pnas.2011795117

12,210 | 27

PNAS
Vol. 117 | No. 41

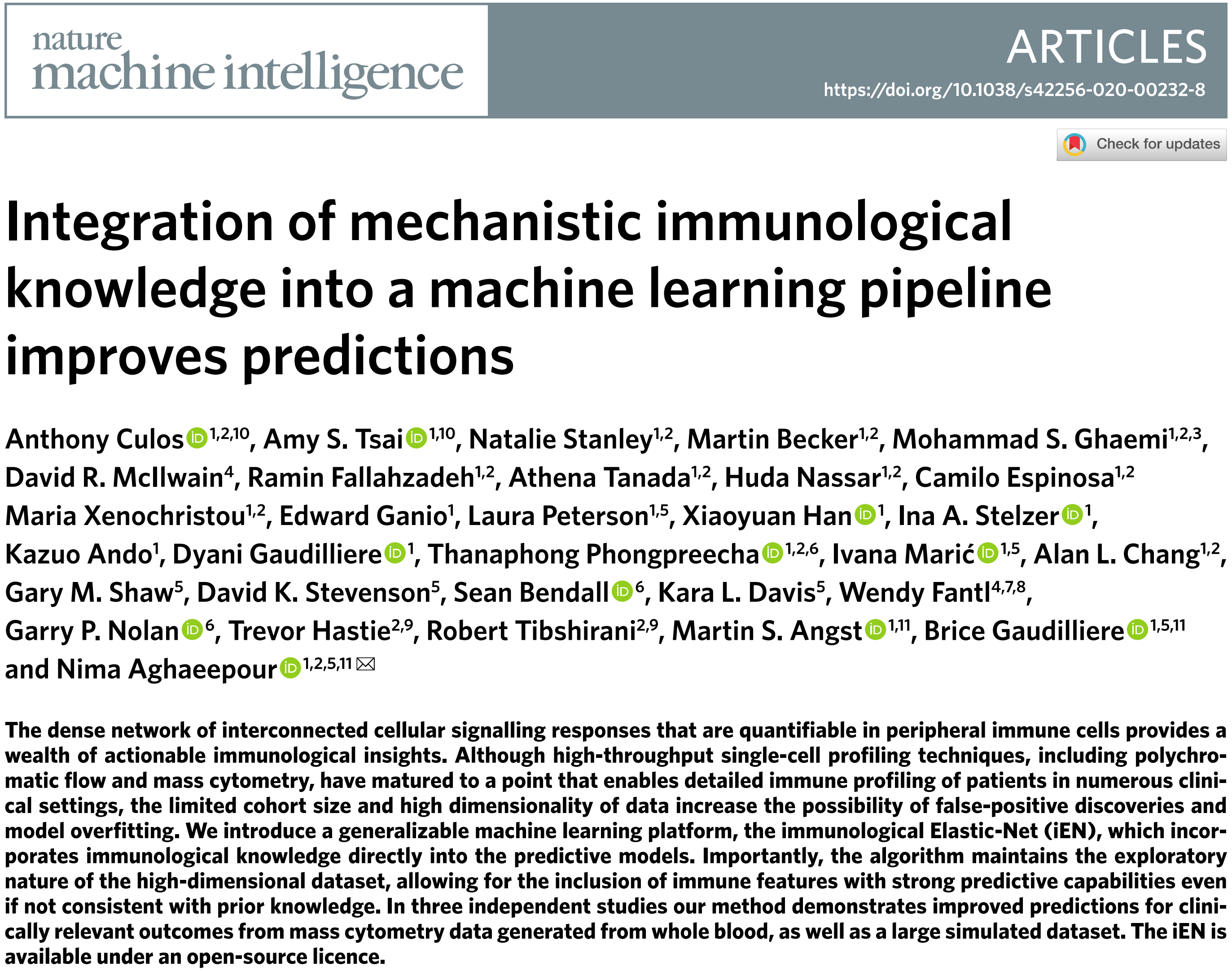
Significance
Abstract
Results
Discussion
Methods
Data Availability
Acknowledgments
Supporting Information
References

Significance

Applying artificial intelligence tools to a highly complex question of immunology, we show that a deep neural network can learn to predict the patterns of chromatin opening across 81 stem and differentiated cells across the immune system, solely from the DNA sequence of regulatory regions. It does so by discovering ab initio the binding motifs for known master regulators, along with some unknown ones, and their combinatorial operation. These predictions validated biochemically, and a mouse-trained neural network predicts human enhancer/promoter activity much better than sequence comparisons would. Beyond serving as a trove of testable functional frameworks, this work is important in showing how massively complex integrated questions of immunology can be handled with such tools.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7924715/
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0011267
https://www.pnas.org/doi/10.1073/pnas.2011795117

Example of an immunology ML tool



nature machine intelligence

ARTICLES

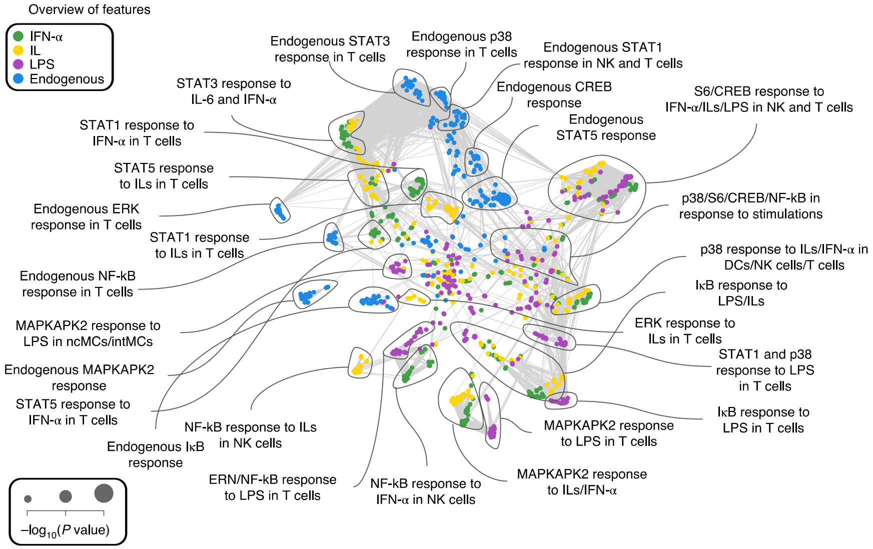
https://doi.org/10.1038/s42256-020-00232-8

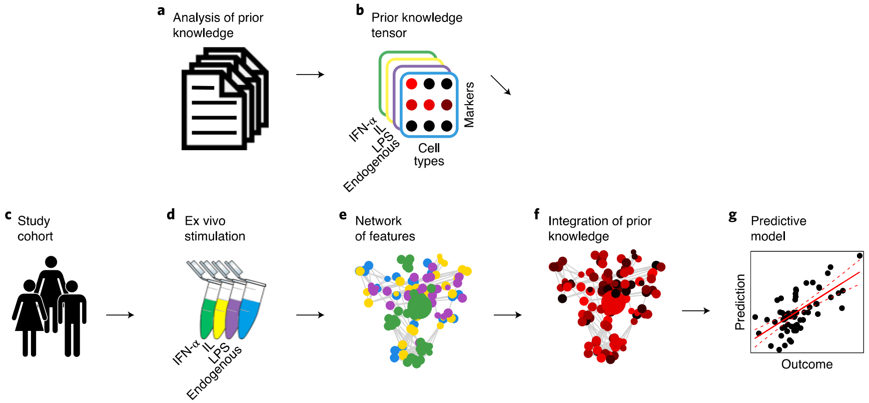
Check for updates

Integration of mechanistic immunological knowledge into a machine learning pipeline improves predictions

Anthony Culos[1,2,10], Amy S. Tsai[1,10], Natalie Stanley[1,2], Martin Becker[1,2], Mohammad S. Ghaemi[1,2,3], David R. McIlwain[4], Ramin Fallahzadeh[1,2], Athena Tanada[1,2], Huda Nassar[1,2], Camilo Espinosa[1,2] Maria Xenochristou[1,2], Edward Ganio[1], Laura Peterson[1,5], Xiaoyuan Han[1], Ina A. Stelzer[1], Kazuo Ando[1], Dyani Gaudilliere[1], Thanaphong Phongpreecha[1,2,6], Ivana Marić[1,5], Alan L. Chang[1,2], Gary M. Shaw[5], David K. Stevenson[5], Sean Bendall[6], Kara L. Davis[5], Wendy Fantl[4,7,8], Garry P. Nolan[6], Trevor Hastie[2,9], Robert Tibshirani[2,9], Martin S. Angst[1,11], Brice Gaudilliere[1,5,11] and Nima Aghaeepour[1,2,5,11] ✉

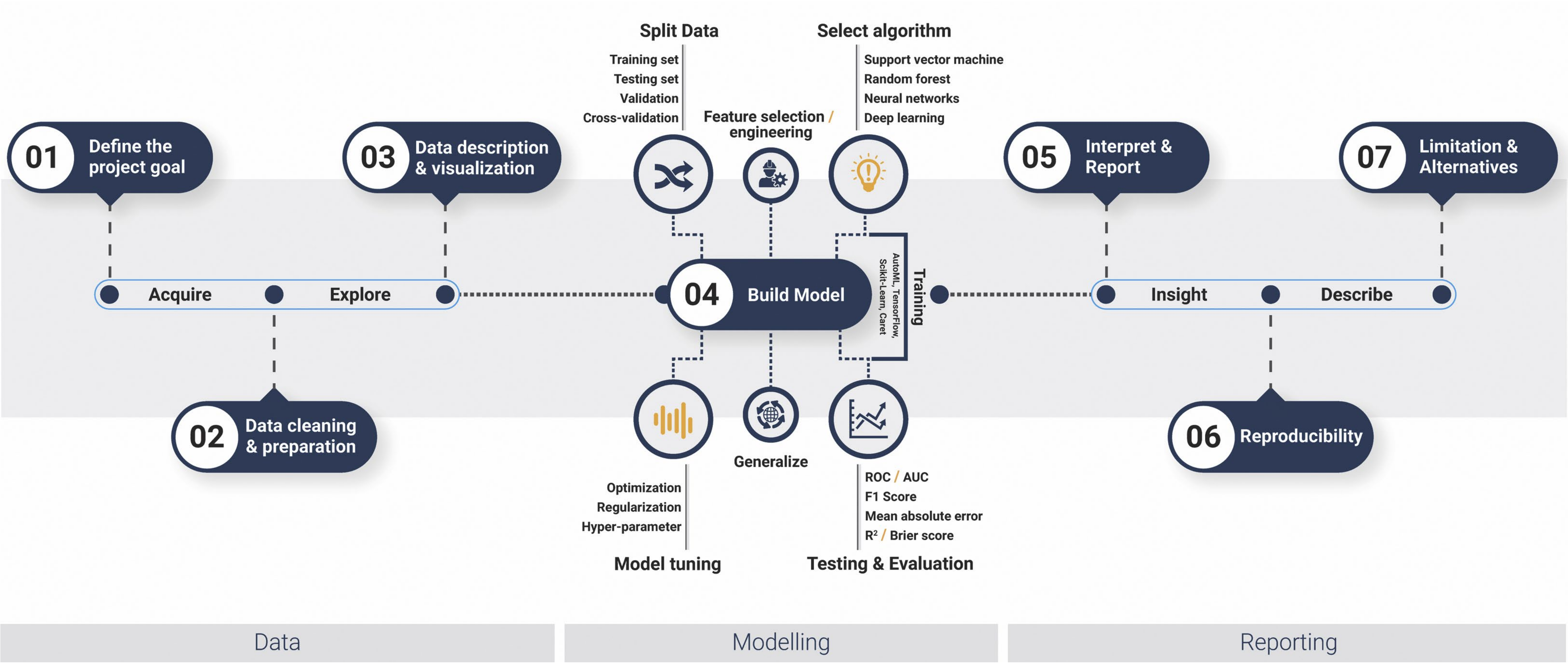
The dense network of interconnected cellular signalling responses that are quantifiable in peripheral immune cells provides a wealth of actionable immunological insights. Although high-throughput single-cell profiling techniques, including polychromatic flow and mass cytometry, have matured to a point that enables detailed immune profiling of patients in numerous clinical settings, the limited cohort size and high dimensionality of data increase the possibility of false-positive discoveries and model overfitting. We introduce a generalizable machine learning platform, the immunological Elastic-Net (iEN), which incorporates immunological knowledge directly into the predictive models. Importantly, the algorithm maintains the exploratory nature of the high-dimensional dataset, allowing for the inclusion of immune features with strong predictive capabilities even if not consistent with prior knowledge. In three independent studies our method demonstrates improved predictions for clinically relevant outcomes from mass cytometry data generated from whole blood, as well as a large simulated dataset. The iEN is available under an open-source licence.


Overview of features
● IFN-α
● IL
● LPS
● Endogenous
Endogenous STAT3 response in T cells
Endogenous p38 response in T cells
Endogenous STAT1 response in NK and T cells
STAT3 response to IL-6 and IFN-α
Endogenous CREB response
S6/CREB response to IFN-α/ILs/LPS in NK and T cells
Endogenous STAT5 response
STAT1 response to IFN-α in T cells
STAT5 response to ILs in T cells
p38/S6/CREB/NF-kB in response to stimulations
Endogenous ERK response in T cells
STAT1 response to ILs in T cells
p38 response to ILs/IFN-α in DCs/NK cells/T cells
Endogenous NF-kB response in T cells
IκB response to LPS/ILs
MAPKAPK2 response to LPS in ncMCs/intMCs
ERK response to ILs in T cells
STAT1 and p38 response to LPS in T cells
Endogenous MAPKAPK2 response
STAT5 response to IFN-α in T cells
IκB response to LPS in T cells
NF-kB response to ILs in NK cells
MAPKAPK2 response to LPS in T cells
Endogenous IκB response
ERN/NF-kB response to LPS in T cells
NF-kB response to IFN-α in NK cells
MAPKAPK2 response to ILs/IFN-α
–log$_{10}$(P value)



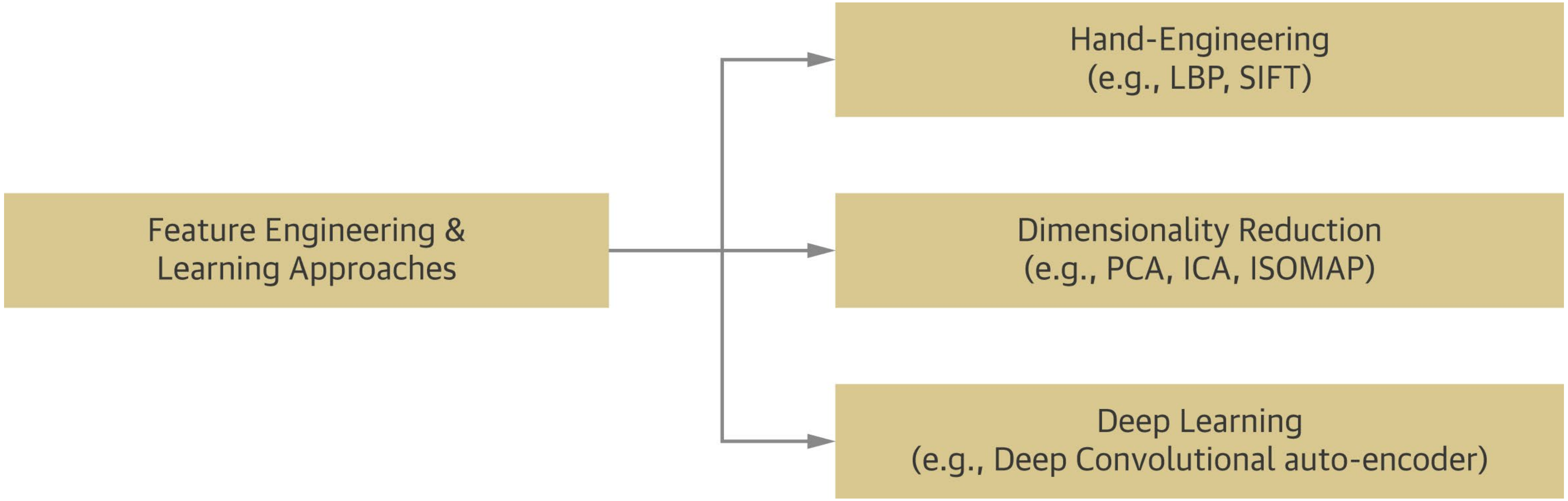
https://www.nature.com/articles/s42256-020-00232-8

Designing a classical machine learning project: key steps



https://pubmed.ncbi.nlm.nih.gov/32912474/

Feature engineering: Key concepts



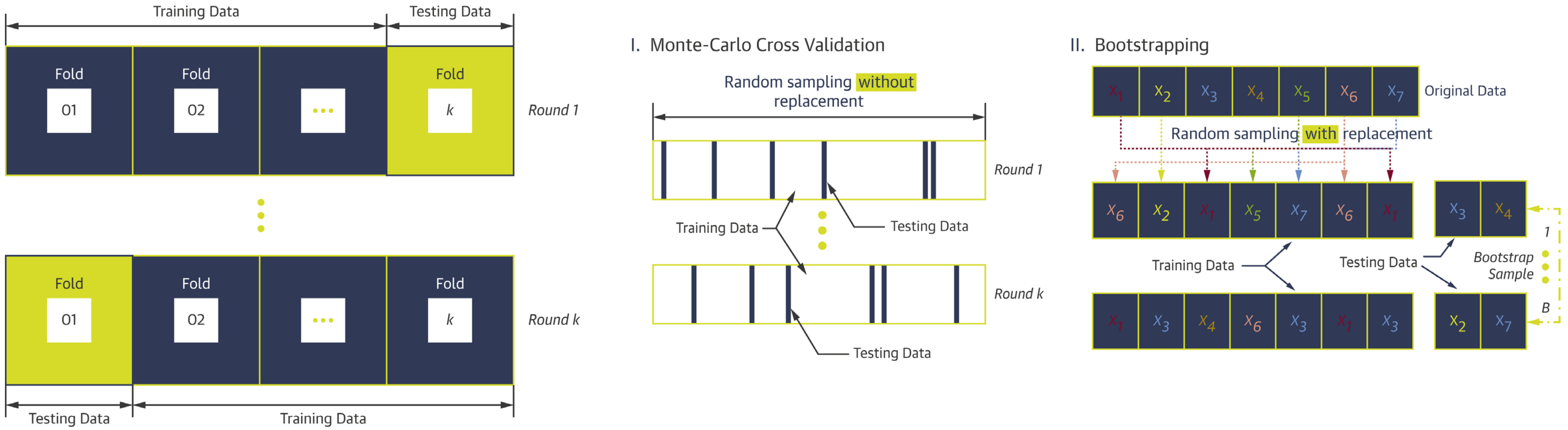
Feature Engineering & Learning Approaches

Hand-Engineering (e.g., LBP, SIFT)

Dimensionality Reduction (e.g., PCA, ICA, ISOMAP)

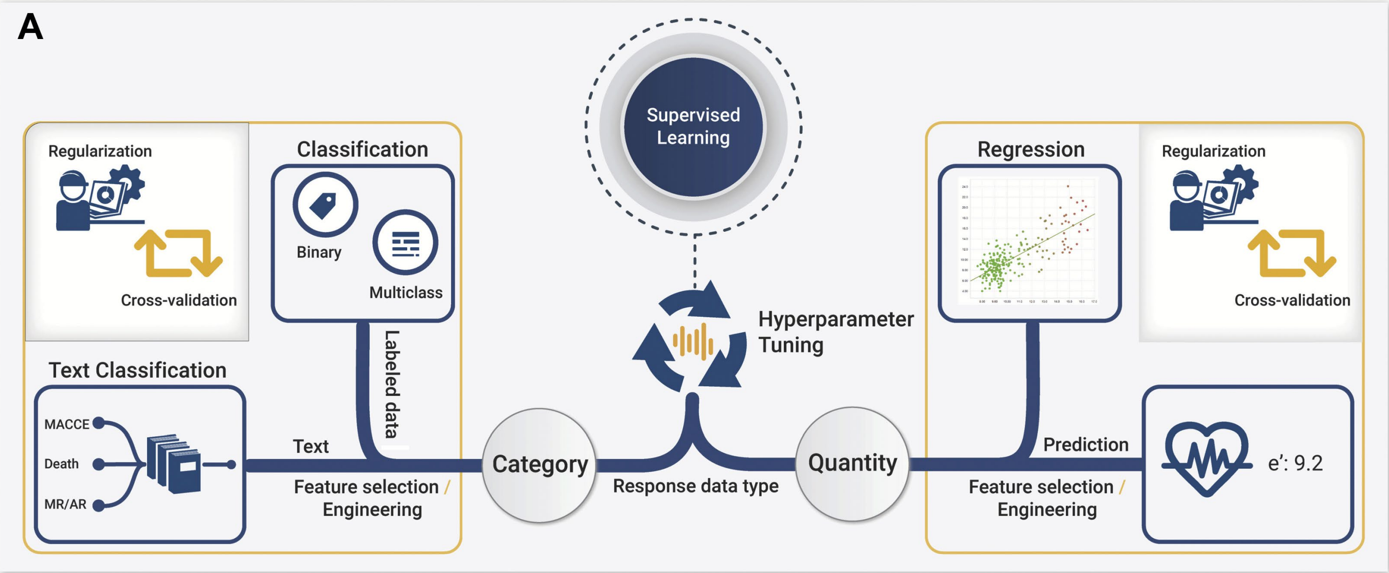
Deep Learning (e.g., Deep Convolutional auto-encoder)

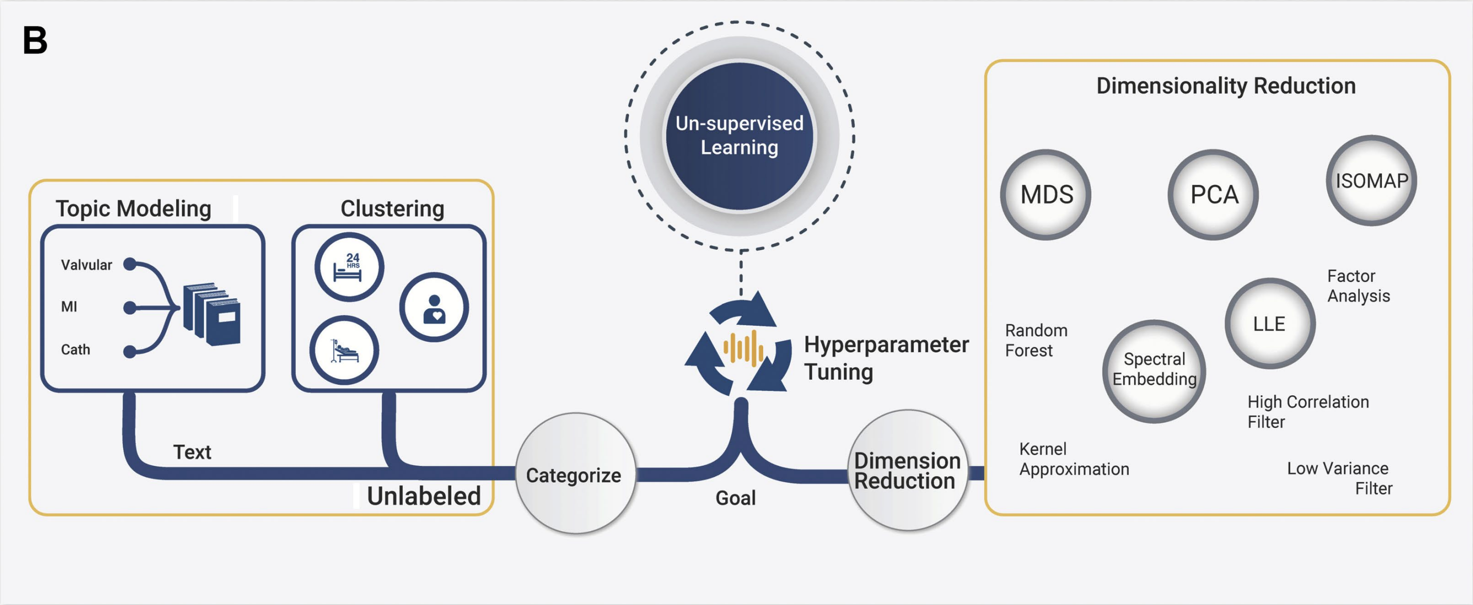
Optimizing training and testing: Key concepts



https://pubmed.ncbi.nlm.nih.gov/32912474/

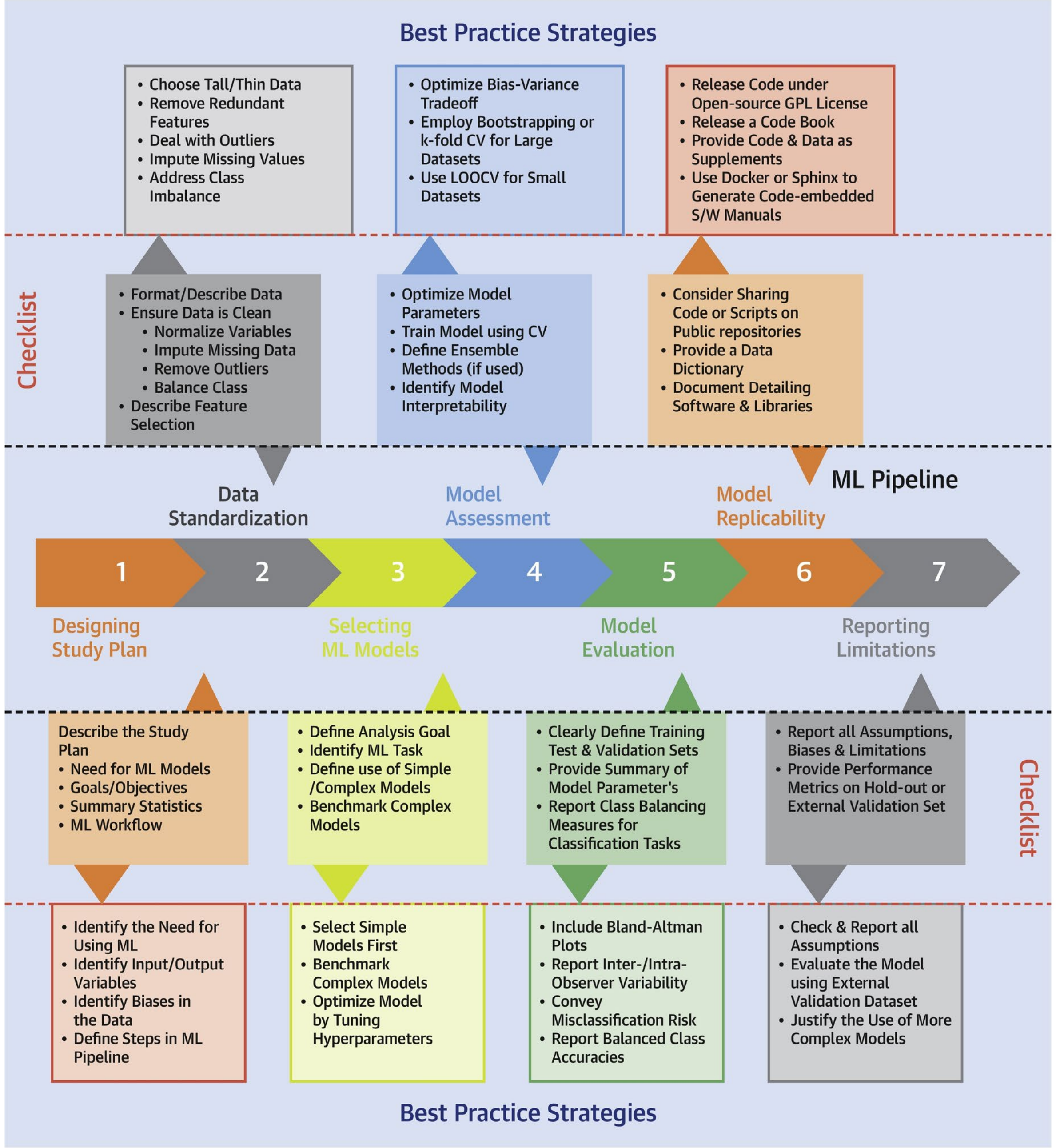
Supervised vs. Unsupervised Learning





https://pubmed.ncbi.nlm.nih.gov/32912474/

Bringing it all together: Best Practices and a check-list



Best Practice Strategies

- Choose Tall/Thin Data
- Remove Redundant Features
- Deal with Outliers
- Impute Missing Values
- Address Class Imbalance

- Optimize Bias-Variance Tradeoff
- Employ Bootstrapping or k-fold CV for Large Datasets
- Use LOOCV for Small Datasets

- Release Code under Open-source GPL License
- Release a Code Book
- Provide Code & Data as Supplements
- Use Docker or Sphinx to Generate Code-embedded S/W Manuals

Checklist

- Format/Describe Data
- Ensure Data is Clean
  - Normalize Variables
  - Impute Missing Data
  - Remove Outliers
  - Balance Class
- Describe Feature Selection

- Optimize Model Parameters
- Train Model using CV
- Define Ensemble Methods (if used)
- Identify Model Interpretability

- Consider Sharing Code or Scripts on Public repositories
- Provide a Data Dictionary
- Document Detailing Software & Libraries

ML Pipeline

Data Standardization

Model Assessment

Model Replicability

1 2 3 4 5 6 7

Designing Study Plan

Selecting ML Models

Model Evaluation

Reporting Limitations

Describe the Study Plan
- Need for ML Models
- Goals/Objectives
- Summary Statistics
- ML Workflow

- Define Analysis Goal
- Identify ML Task
- Define use of Simple /Complex Models
- Benchmark Complex Models

- Clearly Define Training Test & Validation Sets
- Provide Summary of Model Parameter's
- Report Class Balancing Measures for Classification Tasks

- Report all Assumptions, Biases & Limitations
- Provide Performance Metrics on Hold-out or External Validation Set

Checklist

- Identify the Need for Using ML
- Identify Input/Output Variables
- Identify Biases in the Data
- Define Steps in ML Pipeline

- Select Simple Models First
- Benchmark Complex Models
- Optimize Model by Tuning Hyperparameters

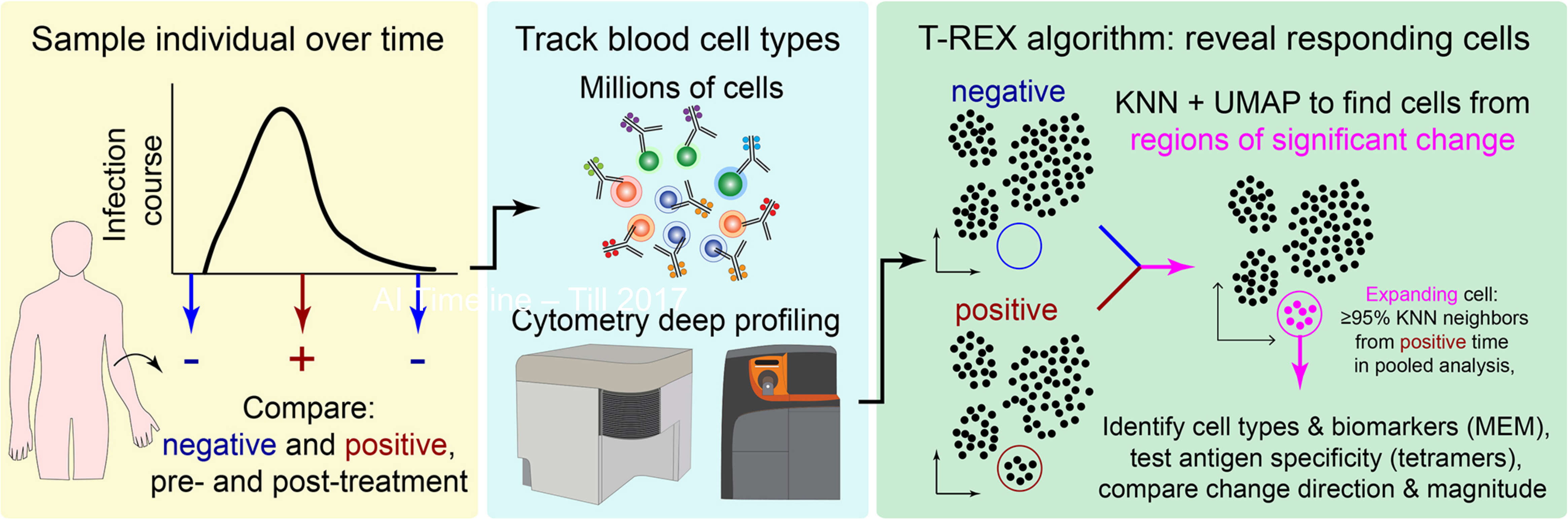
- Include Bland-Altman Plots
- Report Inter-/Intra-Observer Variability
- Convey Misclassification Risk
- Report Balanced Class Accuracies

- Check & Report all Assumptions
- Evaluate the Model using External Validation Dataset
- Justify the Use of More Complex Models

Best Practice Strategies

https://pubmed.ncbi.nlm.nih.gov/32912474/

Unsupervised learning: Concept & Example: Tracking Responders EXpanding (T-REX)



Sample individual over time

Compare: negative and positive, pre- and post-treatment

Track blood cell types

Millions of cells

Cytometry deep profiling

T-REX algorithm: reveal responding cells

KNN + UMAP to find cells from regions of significant change

Expanding cell: ≥95% KNN neighbors from positive time in pooled analysis,

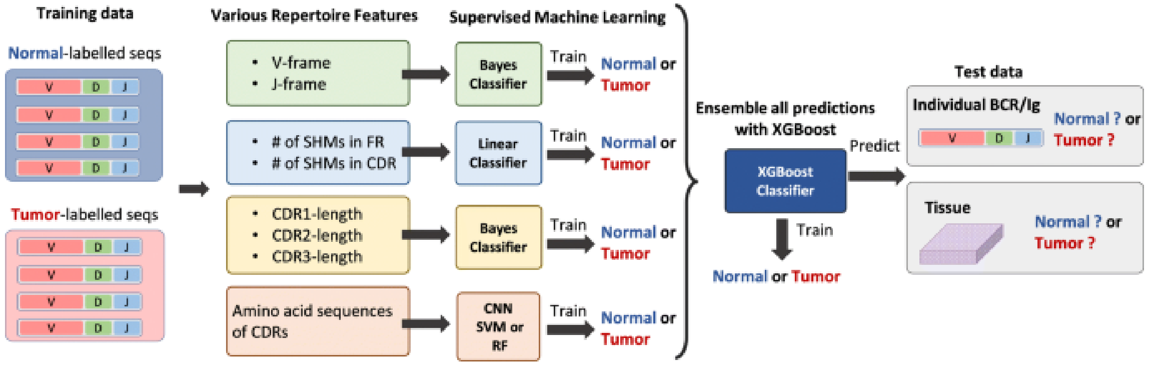
Identify cell types & biomarkers (MEM), test antigen specificity (tetramers), compare change direction & magnitude

Supervised learning: Concept & Example: discriminating BCRs/Igs

Konishi *et al. BMC Bioinformatics* (2019) 20:267
https://doi.org/10.1186/s12859-019-2853-y

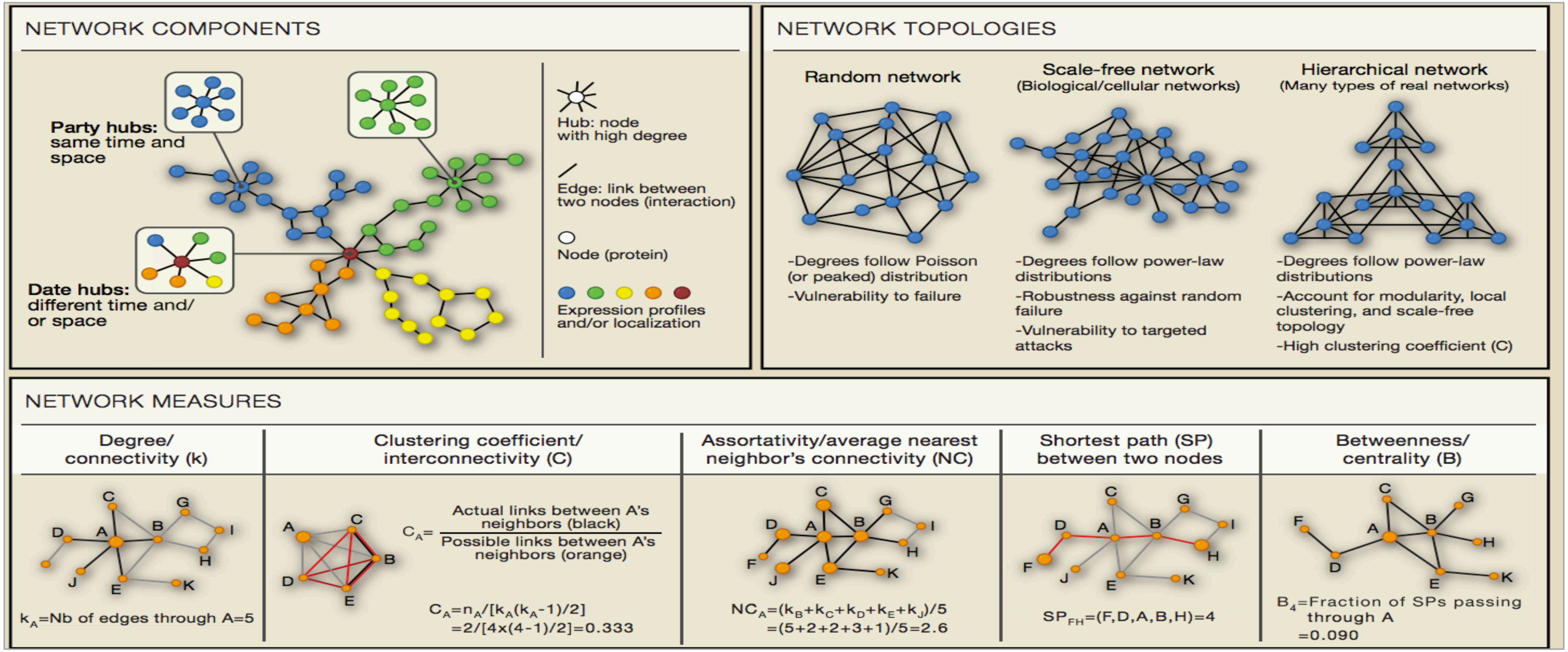
BMC Bioinformatics

Check for updates



https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2853-y

Quantitative exploration of network: Clique, sub-networks and motifs



NETWORK COMPONENTS

Party hubs: same time and space

Date hubs: different time and/ or space

Hub: node with high degree

Edge: link between two nodes (interaction)

○ Node (protein)

● ● ● ● ● Expression profiles and/or localization

NETWORK TOPOLOGIES

Random network

-Degrees follow Poisson (or peaked) distribution

-Vulnerability to failure

Scale-free network (Biological/cellular networks)

-Degrees follow power-law distributions

-Robustness against random failure

-Vulnerability to targeted attacks

Hierarchical network (Many types of real networks)

-Degrees follow power-law distributions

-Account for modularity, local clustering, and scale-free topology

-High clustering coefficient (C)

NETWORK MEASURES

Degree/ connectivity (k)

k_A=Nb of edges through A=5

Clustering coefficient/ interconnectivity (C)

$$C_A = \frac{\text{Actual links between A's neighbors (black)}}{\text{Possible links between A's neighbors (orange)}}$$

$C_A=n_A/[k_A(k_A-1)/2]$
$=2/[4x(4-1)/2]=0.333$

Assortativity/average nearest neighbor's connectivity (NC)

$NC_A=(k_B+k_C+k_D+k_E+k_J)/5$
$=(5+2+2+3+1)/5=2.6$

Shortest path (SP) between two nodes

SP_{FH}=(F,D,A,B,H)=4

Betweenness/ centrality (B)

B_4=Fraction of SPs passing through A
=0.090

http://snapshots.cell.com/

Application of Knowledge Graphs in Immunology

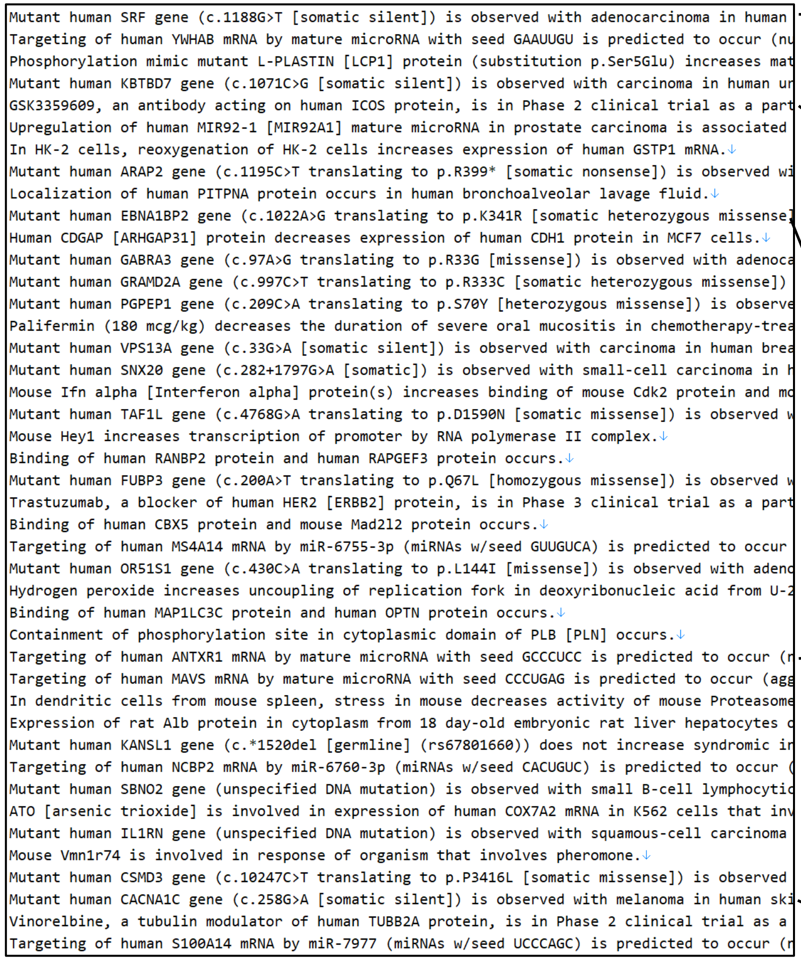
Public

ELSEVIER
Springer
nature
International journal of science
Science
AAAS
...

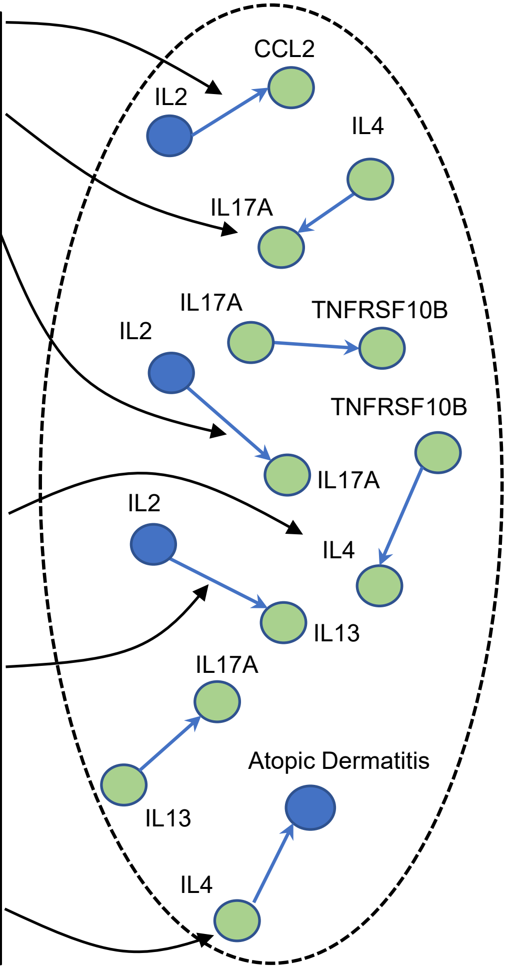
Internal

SANOFI
ImmUniverse
3TR
U-BIOPRED
Open Targets
EMBL-EBI

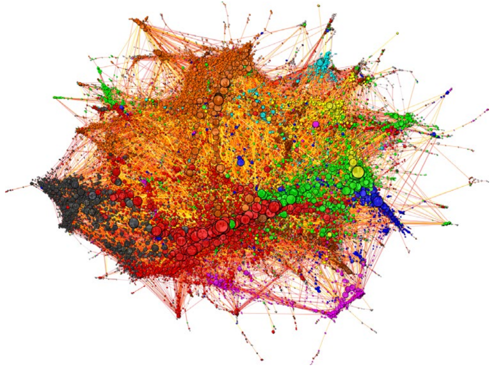
NLP



```
Mutant human SRF gene (c.1188G>T [somatic silent]) is observed with adenocarcinoma in human
Targeting of human YWHAB mRNA by mature microRNA with seed GAAUUGU is predicted to occur (nu
Phosphorylation mimic mutant L-PLASTIN [LCP1] protein (substitution p.Ser5Glu) increases mat
Mutant human KBTBD7 gene (c.1071C>G [somatic silent]) is observed with carcinoma in human ur
GSK3359609, an antibody acting on human ICOS protein, is in Phase 2 clinical trial as a part
Upregulation of human MIR92-1 [MIR92A1] mature microRNA in prostate carcinoma is associated
In HK-2 cells, reoxygenation of HK-2 cells increases expression of human GSTP1 mRNA.↓
Mutant human ARAP2 gene (c.1195C>T translating to p.R399* [somatic nonsense]) is observed wi
Localization of human PITPNA protein occurs in human bronchoalveolar lavage fluid.↓
Mutant human EBNA1BP2 gene (c.1022A>G translating to p.K341R [somatic heterozygous missense]
Human CDGAP [ARHGAP31] protein decreases expression of human CDH1 protein in MCF7 cells.↓
Mutant human GABRA3 gene (c.97A>G translating to p.R33G [missense]) is observed with adenoca
Mutant human GRAMD2A gene (c.997C>T translating to p.R333C [somatic heterozygous missense])
Mutant human PGPEP1 gene (c.209C>A translating to p.S70Y [heterozygous missense]) is observe
Palifermin (180 mcg/kg) decreases the duration of severe oral mucositis in chemotherapy-trea
Mutant human VPS13A gene (c.33G>A [somatic silent]) is observed with carcinoma in human brea
Mutant human SNX20 gene (c.282+1797G>A [somatic]) is observed with small-cell carcinoma in h
Mouse Ifn alpha [Interferon alpha] protein(s) increases binding of mouse Cdk2 protein and mo
Mutant human TAF1L gene (c.4768G>A translating to p.D1590N [somatic missense]) is observed w
Mouse Hey1 increases transcription of promoter by RNA polymerase II complex.↓
Binding of human RANBP2 protein and human RAPGEF3 protein occurs.↓
Mutant human FUBP3 gene (c.200A>T translating to p.Q67L [homozygous missense]) is observed w
Trastuzumab, a blocker of human HER2 [ERBB2] protein, is in Phase 3 clinical trial as a part
Binding of human CBX5 protein and mouse Mad2l2 protein occurs.↓
Targeting of human MS4A14 mRNA by miR-6755-3p (miRNAs w/seed GUUGUCA) is predicted to occur
Mutant human OR51S1 gene (c.430C>A translating to p.L144I [missense]) is observed with adenc
Hydrogen peroxide increases uncoupling of replication fork in deoxyribonucleic acid from U-2
Binding of human MAP1LC3C protein and human OPTN protein occurs.↓
Containment of phosphorylation site in cytoplasmic domain of PLB [PLN] occurs.↓
Targeting of human ANTXR1 mRNA by mature microRNA with seed GCCCUCC is predicted to occur (r
Targeting of human MAVS mRNA by mature microRNA with seed CCCUGAG is predicted to occur (agg
In dendritic cells from mouse spleen, stress in mouse decreases activity of mouse Proteasome
Expression of rat Alb protein in cytoplasm from 18 day-old embryonic rat liver hepatocytes c
Mutant human KANSL1 gene (c.*1520del [germline] (rs67801660)) does not increase syndromic ir
Targeting of human NCBP2 mRNA by miR-6760-3p (miRNAs w/seed CACUGUC) is predicted to occur (
Mutant human SBNO2 gene (unspecified DNA mutation) is observed with small B-cell lymphocytic
ATO [arsenic trioxide] is involved in expression of human COX7A2 mRNA in K562 cells that inv
Mutant human IL1RN gene (unspecified DNA mutation) is observed with squamous-cell carcinoma
Mouse Vmn1r74 is involved in response of organism that involves pheromone.↓
Mutant human CSMD3 gene (c.10247C>T translating to p.P3416L [somatic missense]) is observed
Mutant human CACNA1C gene (c.258G>A [somatic silent]) is observed with melanoma in human ski
Vinorelbine, a tubulin modulator of human TUBB2A protein, is in Phase 2 clinical trial as a
Targeting of human S100A14 mRNA by miR-7977 (miRNAs w/seed UCCCAGC) is predicted to occur (r
```



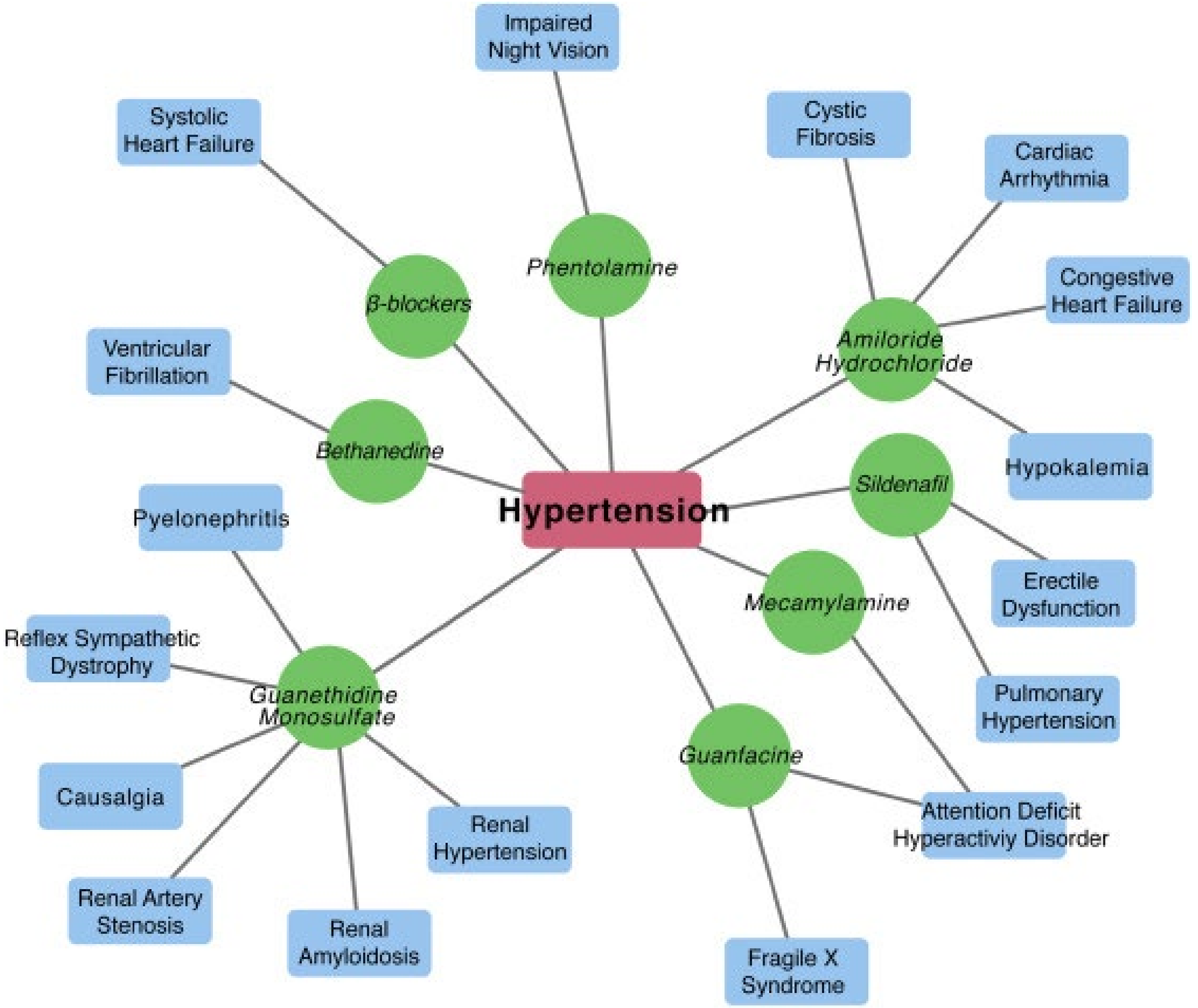
Predictive Knowledge Graph



- Store and retrieve findings
- Inference of new relationships
- Analysis of causative connections
- Estimate effect of perturbation

Franck Rapaport, Travis Ahn-Horst, Emanuele de Rinaldis & Shameer Khader

Application of Knowledge Graphs in Immunology



Impaired Night Vision

Systolic Heart Failure

Phentolamine

Cystic Fibrosis

Cardiac Arrhythmia

β-blockers

Congestive Heart Failure

Amiloride Hydrochloride

Ventricular Fibrillation

Bethanedine

Hypokalemia

Sildenafil

Pyelonephritis

Hypertension

Erectile Dysfunction

Mecamylamine

Reflex Sympathetic Dystrophy

Guanethidine Monosulfate

Pulmonary Hypertension

Guanfacine

Causalgia

Renal Hypertension

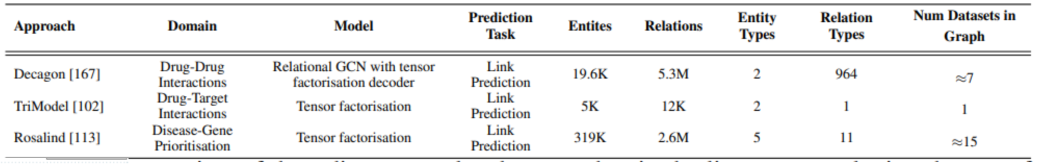
Attention Deficit Hyperactiviy Disorder

Renal Artery Stenosis

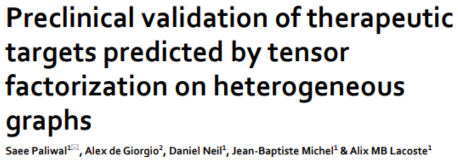
Renal Amyloidosis

Fragile X Syndrome

Application of Knowledge Graphs to Accelerate Discovery Research



Approach	Domain	Model	Prediction Task	Entites	Relations	Entity Types	Relation Types	Num Datasets in Graph
Decagon [167]	Drug-Drug Interactions	Relational GCN with tensor factorisation decoder	Link Prediction	19.6K	5.3M	2	964	≈7
TriModel [102]	Drug-Target Interactions	Tensor factorisation	Link Prediction	5K	12K	2	1	1
Rosalind [113]	Disease-Gene Prioritisation	Tensor factorisation	Link Prediction	319K	2.6M	5	11	≈15

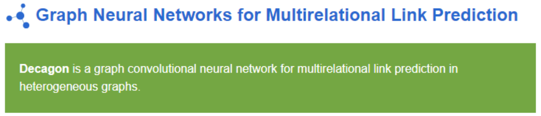


Preclinical validation of therapeutic targets predicted by tensor factorization on heterogeneous graphs

Saee Paliwal[1]✉, Alex de Giorgio[2], Daniel Neil[1], Jean-Baptiste Michel[1] & Alix MB Lacoste[1]

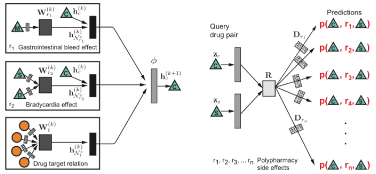
Link prediction, embeddings & algorithms

AmpliGraph



Graph Neural Networks for Multirelational Link Prediction

Decagon is a graph convolutional neural network for multirelational link prediction in heterogeneous graphs.

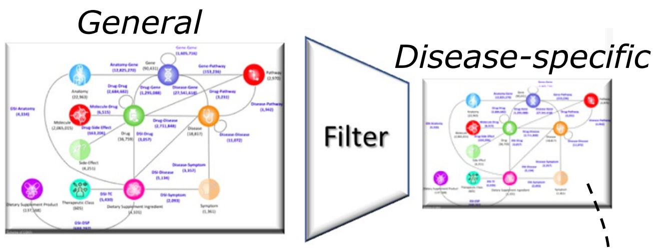


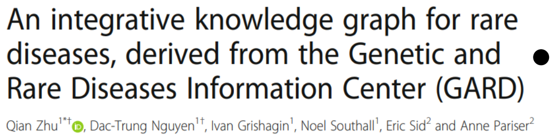
Target identification

Drug repositioning

Target validation

Pre-processing & filtration



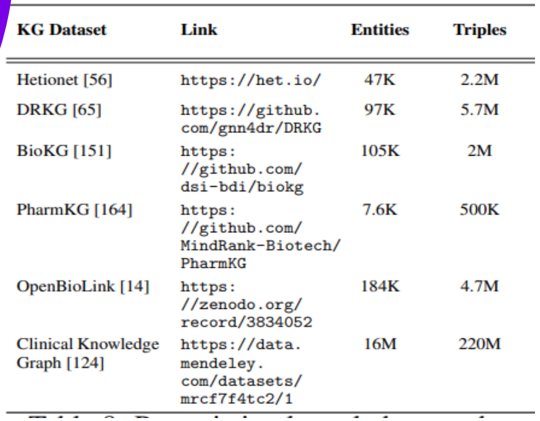


An integrative knowledge graph for rare diseases, derived from the Genetic and Rare Diseases Information Center (GARD)

Qian Zhu[1*†], Dac-Trung Nguyen[1†], Ivan Grishagin[1], Noel Southall[1], Eric Sid[2] and Anne Pariser[2]

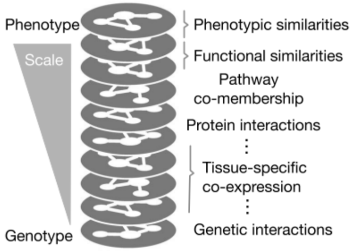
Graph integration

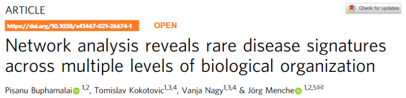
Public knowledge graphs



KG Dataset	Link	Entities	Triples
Hetionet [56]	`https://het.io/`	47K	2.2M
DRKG [65]	`https://github.com/gnn4dr/DRKG`	97K	5.7M
BioKG [151]	`https://github.com/dsi-bdi/biokg`	105K	2M
PharmKG [164]	`https://github.com/MindRank-Biotech/PharmKG`	7.6K	500K
OpenBioLink [14]	`https://zenodo.org/record/3834052`	184K	4.7M
Clinical Knowledge Graph [124]	`https://data.mendeley.com/datasets/mrcf7f4tc2/1`	16M	220M

Multi-omics inferred networks





ARTICLE

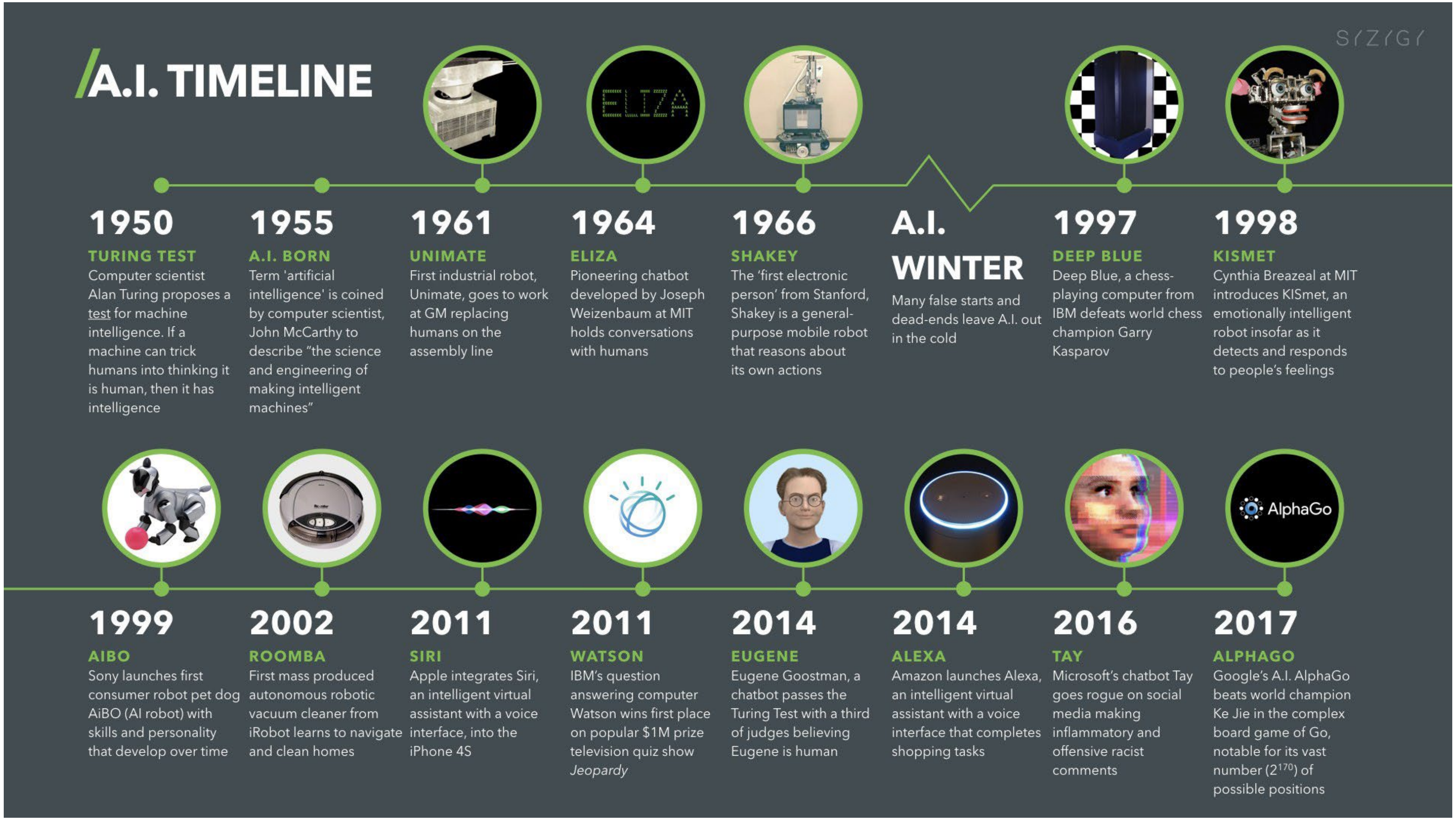
Check for updates

https://doi.org/10.1038/s41467-021-26674-1 OPEN

Network analysis reveals rare disease signatures across multiple levels of biological organization

Pisanu Buphamalai[1,2], Tomislav Kokotovic[1,3,4], Vanja Nagy[1,3,4] & Jörg Menche[1,2,5]✉

AI Timeline – Till 2017



A.I. TIMELINE

1950
TURING TEST
Computer scientist Alan Turing proposes a test for machine intelligence. If a machine can trick humans into thinking it is human, then it has intelligence

1955
A.I. BORN
Term 'artificial intelligence' is coined by computer scientist, John McCarthy to describe "the science and engineering of making intelligent machines"

1961
UNIMATE
First industrial robot, Unimate, goes to work at GM replacing humans on the assembly line

1964
ELIZA
Pioneering chatbot developed by Joseph Weizenbaum at MIT holds conversations with humans

1966
SHAKEY
The 'first electronic person' from Stanford, Shakey is a general-purpose mobile robot that reasons about its own actions

A.I. WINTER
Many false starts and dead-ends leave A.I. out in the cold

1997
DEEP BLUE
Deep Blue, a chess-playing computer from IBM defeats world chess champion Garry Kasparov

1998
KISMET
Cynthia Breazeal at MIT introduces KISmet, an emotionally intelligent robot insofar as it detects and responds to people's feelings

1999
AIBO
Sony launches first consumer robot pet dog AiBO (AI robot) with skills and personality that develop over time

2002
ROOMBA
First mass produced autonomous robotic vacuum cleaner from iRobot learns to navigate and clean homes

2011
SIRI
Apple integrates Siri, an intelligent virtual assistant with a voice interface, into the iPhone 4S

2011
WATSON
IBM's question answering computer Watson wins first place on popular $1M prize television quiz show *Jeopardy*

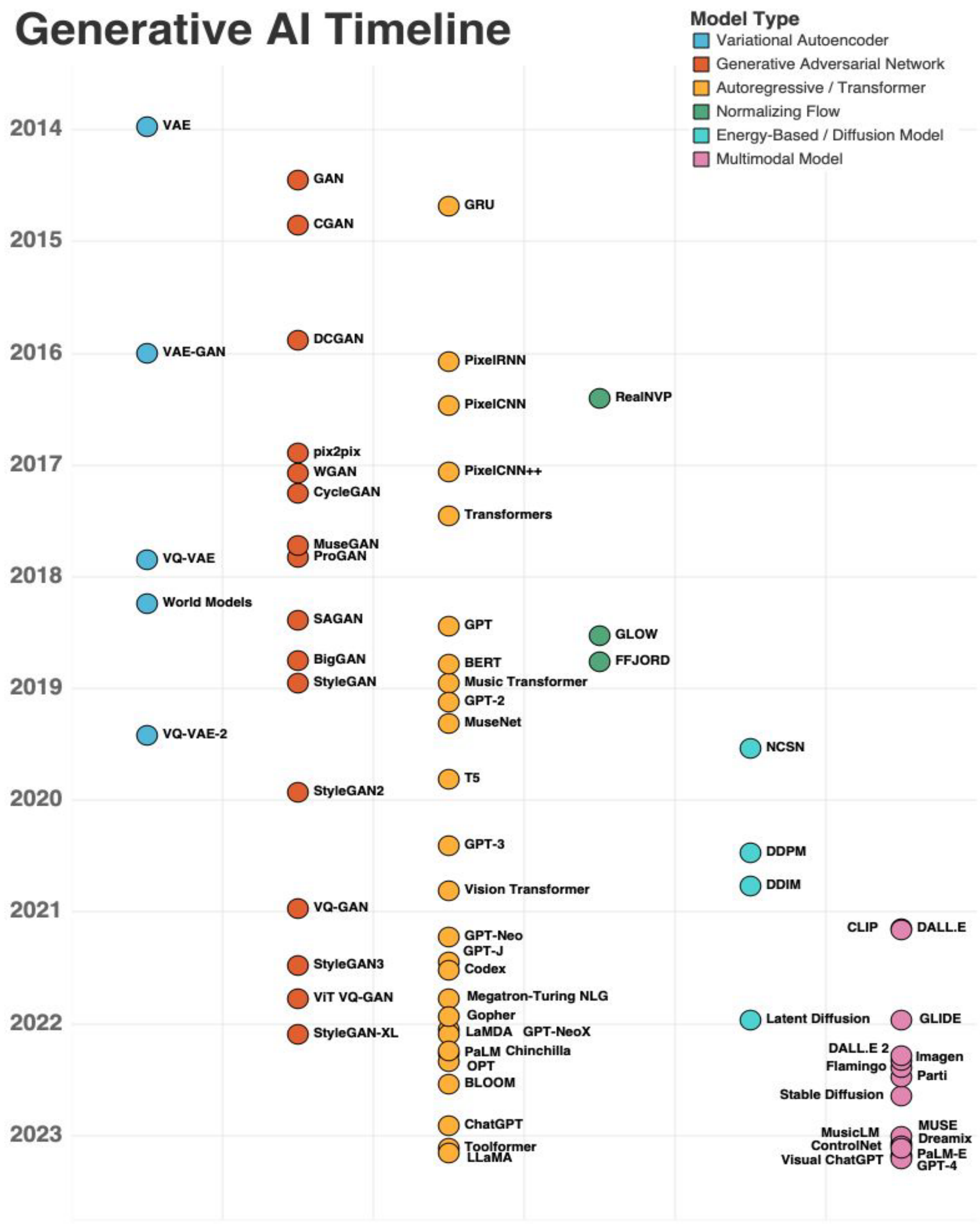
2014
EUGENE
Eugene Goostman, a chatbot passes the Turing Test with a third of judges believing Eugene is human

2014
ALEXA
Amazon launches Alexa, an intelligent virtual assistant with a voice interface that completes shopping tasks

2016
TAY
Microsoft's chatbot Tay goes rogue on social media making inflammatory and offensive racist comments

2017
ALPHAGO
Google's A.I. AlphaGo beats world champion Ke Jie in the complex board game of Go, notable for its vast number (2^{170}) of possible positions

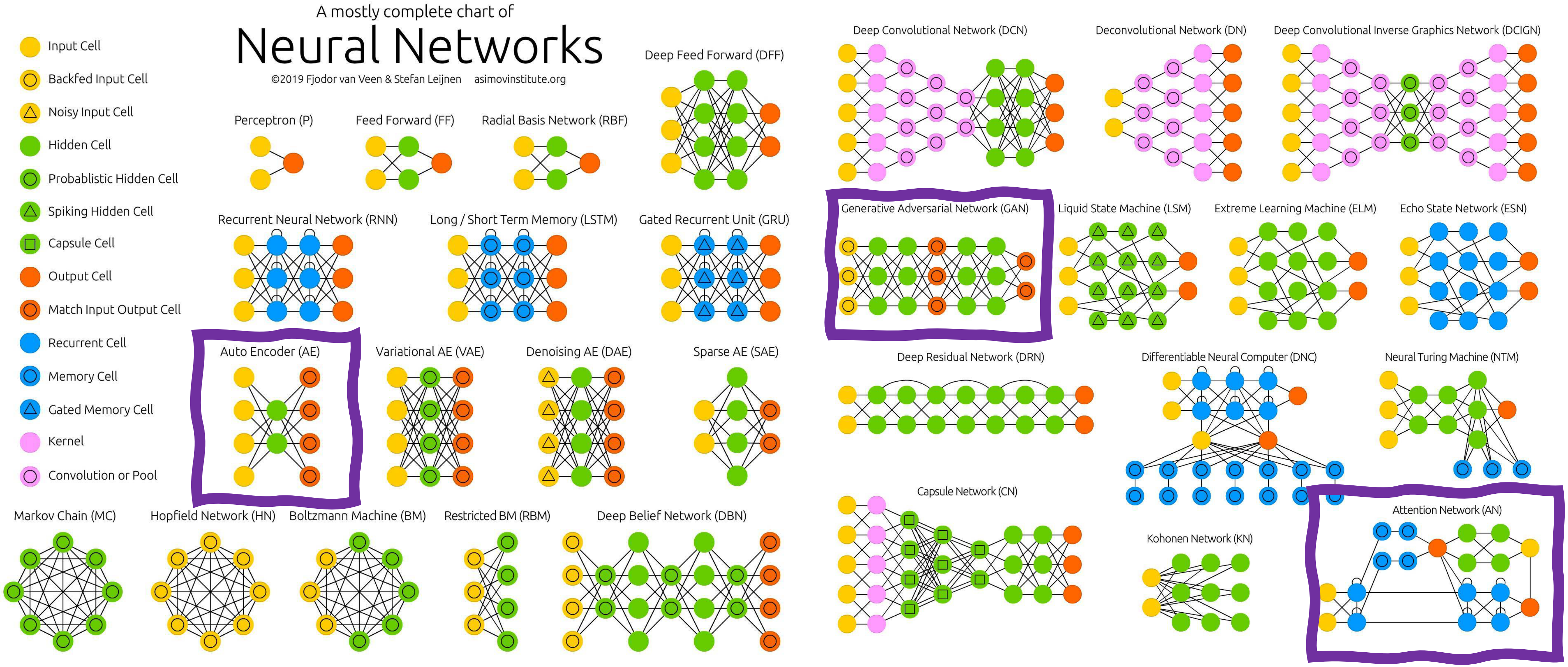
AI Timeline – 2014 and beyond



(Selected) Emerging themes in AI

- Neural network
- Pre-training, fine-tuning and transfer learning
- Attention, Embedding, Autoencoders, Transformers

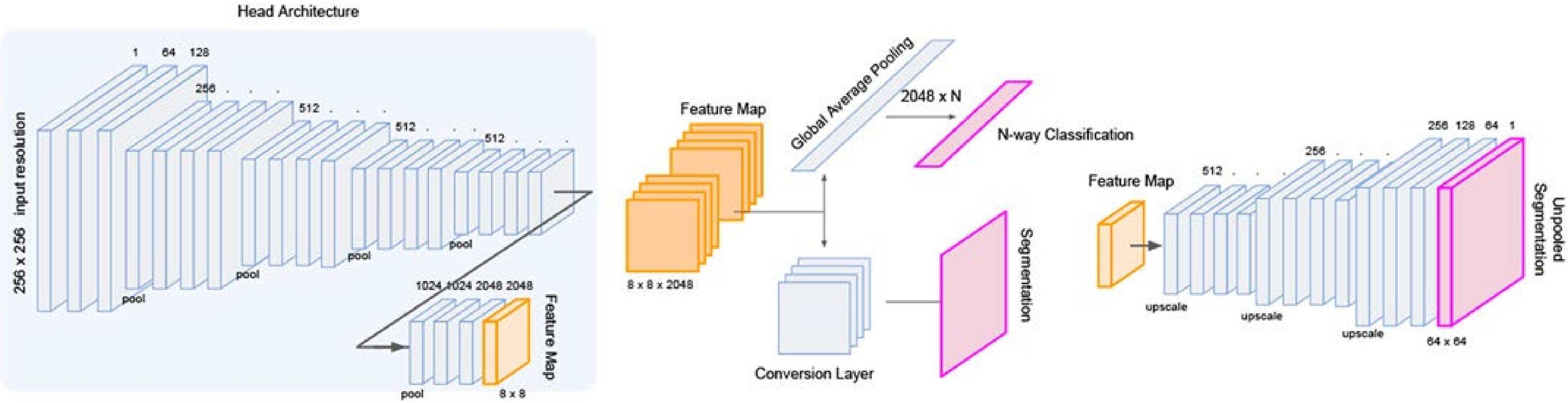
Back to Neural Networks: Different types of architectures


A mostly complete chart of
Neural Networks
©2019 Fjodor van Veen & Stefan Leijnen asimovinstitute.org
Input Cell
Backfed Input Cell
Noisy Input Cell
Hidden Cell
Probablistic Hidden Cell
Spiking Hidden Cell
Capsule Cell
Output Cell
Match Input Output Cell
Recurrent Cell
Memory Cell
Gated Memory Cell
Kernel
Convolution or Pool
Perceptron (P)
Feed Forward (FF)
Radial Basis Network (RBF)
Deep Feed Forward (DFF)
Recurrent Neural Network (RNN)
Long / Short Term Memory (LSTM)
Gated Recurrent Unit (GRU)
Auto Encoder (AE)
Variational AE (VAE)
Denoising AE (DAE)
Sparse AE (SAE)
Markov Chain (MC)
Hopfield Network (HN)
Boltzmann Machine (BM)
Restricted BM (RBM)
Deep Belief Network (DBN)
Deep Convolutional Network (DCN)
Deconvolutional Network (DN)
Deep Convolutional Inverse Graphics Network (DCIGN)
Generative Adversarial Network (GAN)
Liquid State Machine (LSM)
Extreme Learning Machine (ELM)
Echo State Network (ESN)
Deep Residual Network (DRN)
Differentiable Neural Computer (DNC)
Neural Turing Machine (NTM)
Capsule Network (CN)
Kohonen Network (KN)
Attention Network (AN)

https://ai.stackexchange.com/questions/15594/what-are-all-the-different-kinds-of-neural-networks-used-for

Anatomy of a Deep Learning model

1. **Input Layer**
2. **Hidden Layers**
3. **Neurons (Nodes)**
4. **Activation Functions**
5. **Parameters (Weights and Biases)**
6. **Loss Function**
7. **Optimization Algorithm**
8. **Output Layer**
9. **Training**
10. **Inference**



https://www.frontiersin.org/articles/10.3389/fdata.2021.742779/full

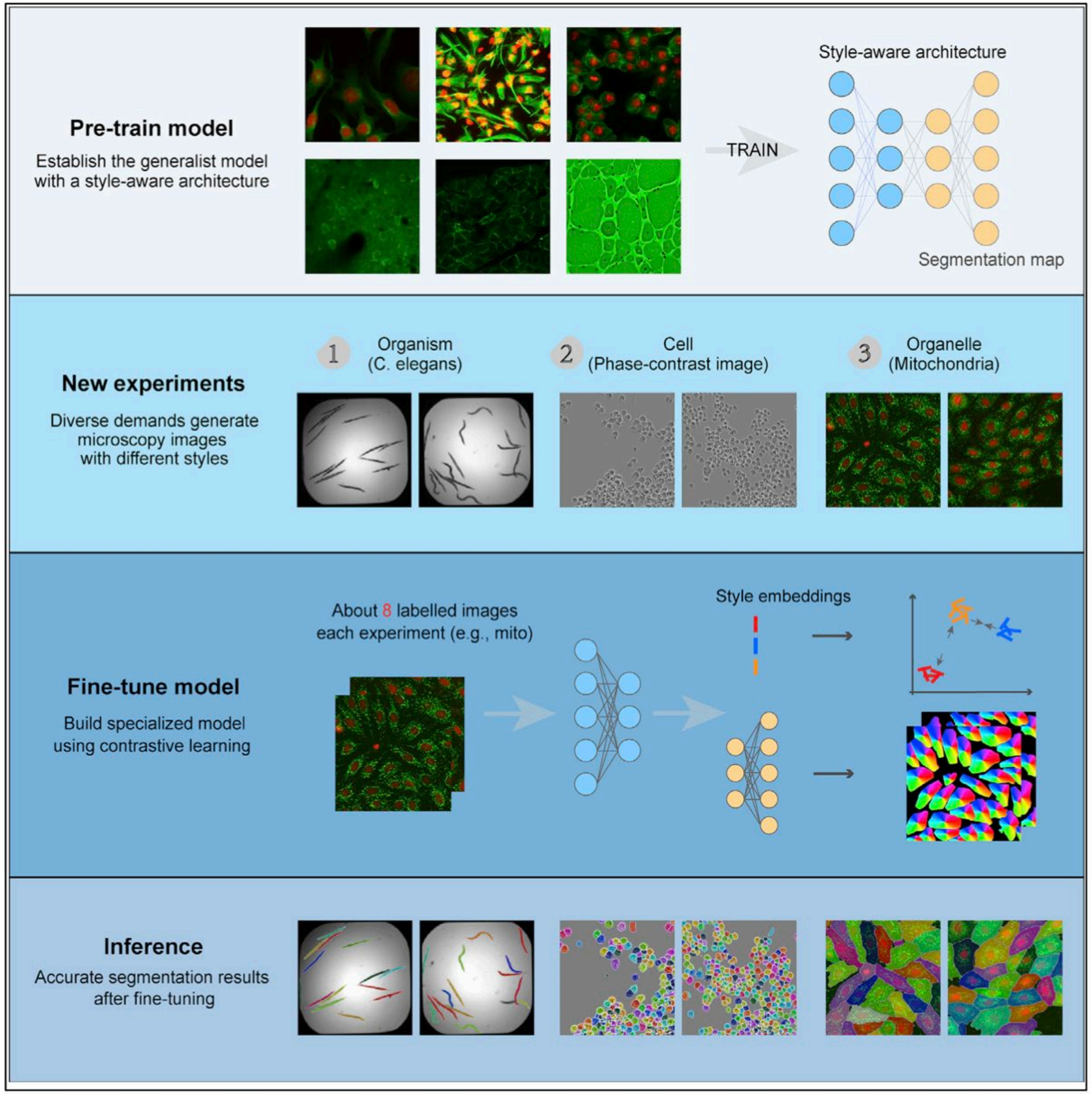
Pre-training, Fine-tuning & Foundation models: Concept & Example

- Pre-training: Pre-training refers to the initial training of a model on a large, diverse dataset to learn general representations of the input data.

- The pre-training process involves training a model on a self-supervised or unsupervised task, where the model learns to predict missing or masked parts of the input data.

- Fine-tuning is the process of taking a pre-trained model and further training it on a specific task or dataset.

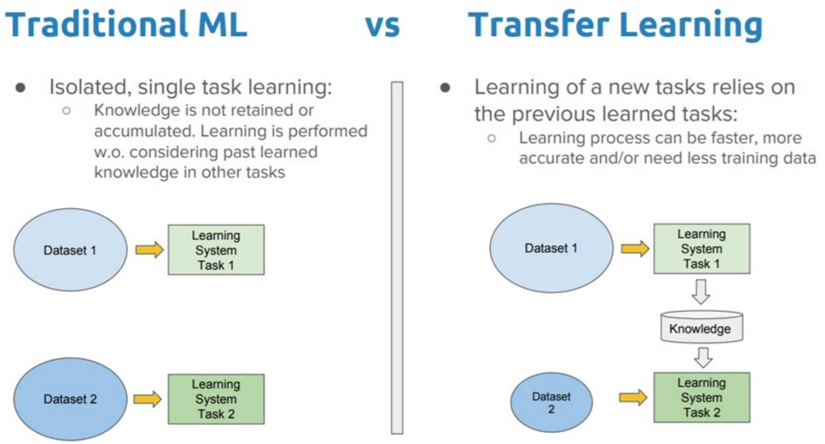
- A foundation model is a pre-trained model that serves as the base for further development or fine-tuning in the context of large language models



Pre-train model
Establish the generalist model with a style-aware architecture
TRAIN
Style-aware architecture
Segmentation map
New experiments
Diverse demands generate microscopy images with different styles
1 Organism (C. elegans)
2 Cell (Phase-contrast image)
3 Organelle (Mitochondria)
Fine-tune model
Build specialized model using contrastive learning
About 8 labelled images each experiment (e.g., mito)
Style embeddings
Inference
Accurate segmentation results after fine-tuning

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9678729/

Transfer learning: Concept & Example



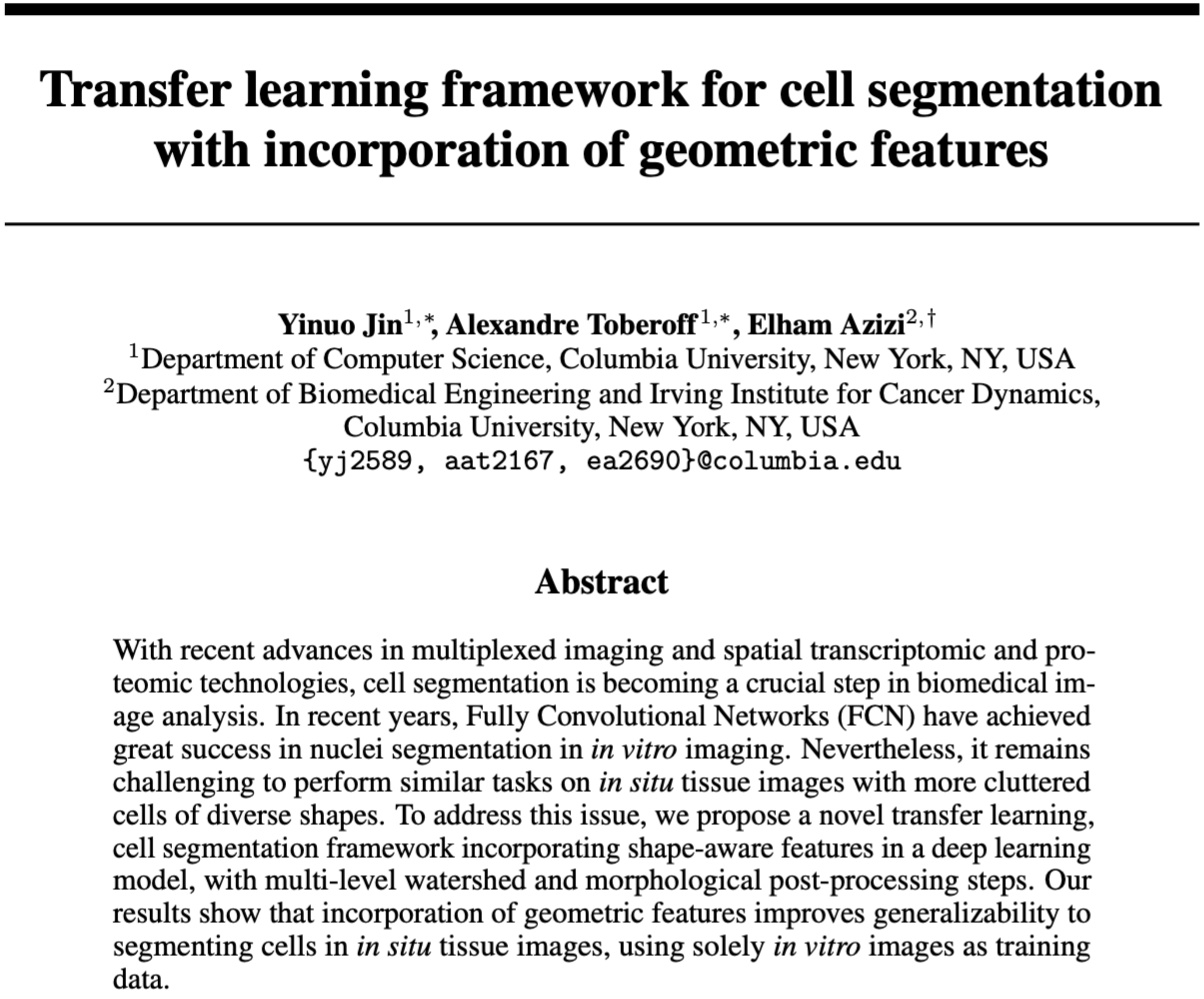
Traditional ML vs Transfer Learning

- Isolated, single task learning:
  - Knowledge is not retained or accumulated. Learning is performed w.o. considering past learned knowledge in other tasks

- Learning of a new tasks relies on the previous learned tasks:
  - Learning process can be faster, more accurate and/or need less training data

- Transfer learning is a machine learning technique that allows knowledge gained from one task to be applied to another related task.

- It involves leveraging pre-trained models that have been trained on large datasets for a specific task and then reusing and adapting them for a different but related task.



Transfer learning framework for cell segmentation with incorporation of geometric features

Yinuo Jin[1,*], Alexandre Toberoff[1,*], Elham Azizi[2,†]
[1]Department of Computer Science, Columbia University, New York, NY, USA
[2]Department of Biomedical Engineering and Irving Institute for Cancer Dynamics, Columbia University, New York, NY, USA
{yj2589, aat2167, ea2690}@columbia.edu

Abstract

With recent advances in multiplexed imaging and spatial transcriptomic and proteomic technologies, cell segmentation is becoming a crucial step in biomedical image analysis. In recent years, Fully Convolutional Networks (FCN) have achieved great success in nuclei segmentation in *in vitro* imaging. Nevertheless, it remains challenging to perform similar tasks on *in situ* tissue images with more cluttered cells of diverse shapes. To address this issue, we propose a novel transfer learning, cell segmentation framework incorporating shape-aware features in a deep learning model, with multi-level watershed and morphological post-processing steps. Our results show that incorporation of geometric features improves generalizability to segmenting cells in *in situ* tissue images, using solely *in vitro* images as training data.

https://www.biorxiv.org/content/10.1101/2021.02.28.433289v2

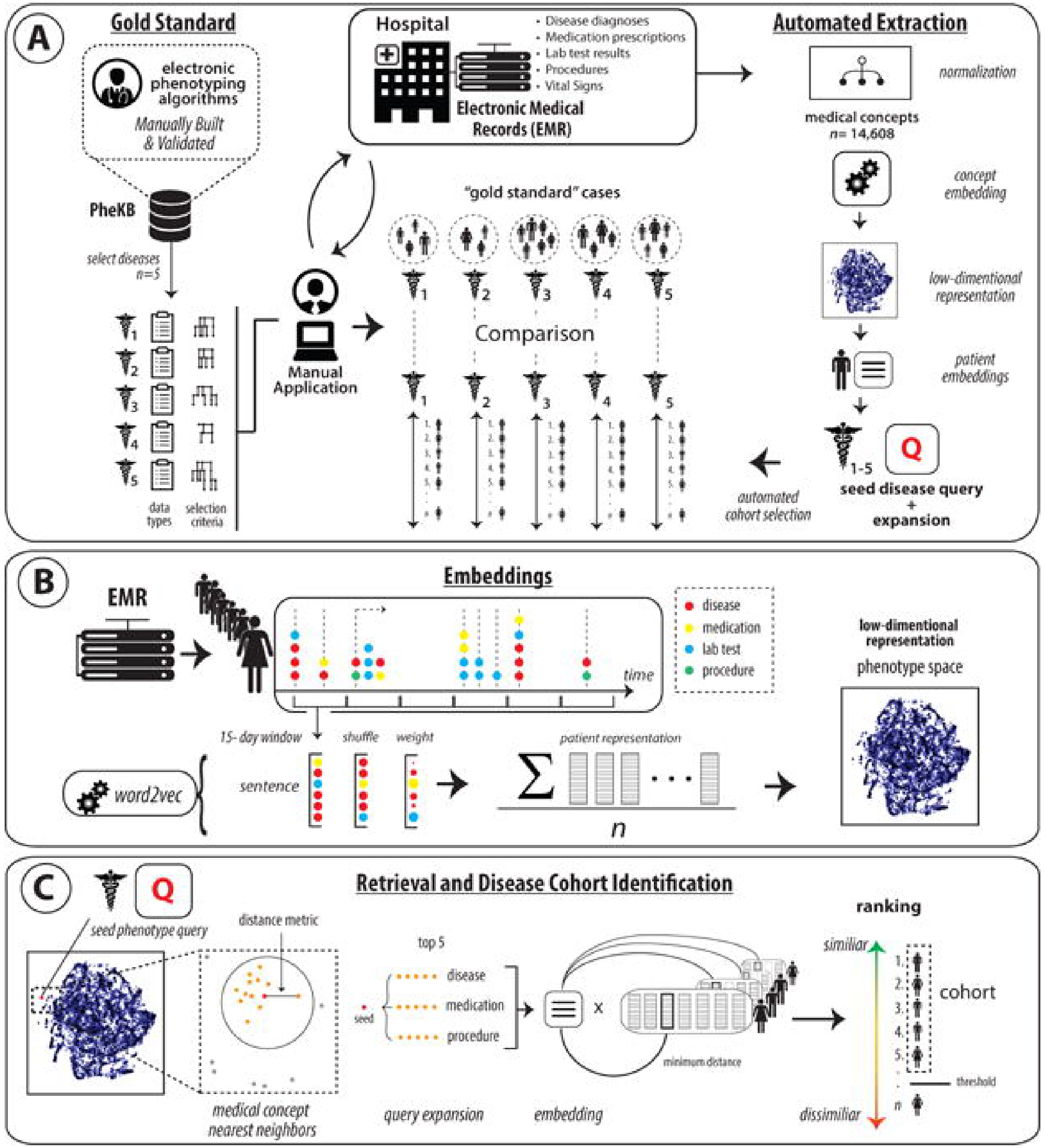
Embedding: Concept & Example

- Embedding refers to a technique used in neural networks to represent categorical or discrete variables as continuous, dense vectors.

- Embeddings are commonly used in various domains, including natural language processing (NLP), recommender systems, and computer vision.

- The main idea behind embedding is to map high-dimensional categorical data to lower-dimensional continuous representations, where similar or related categories are closer together in the embedding space.

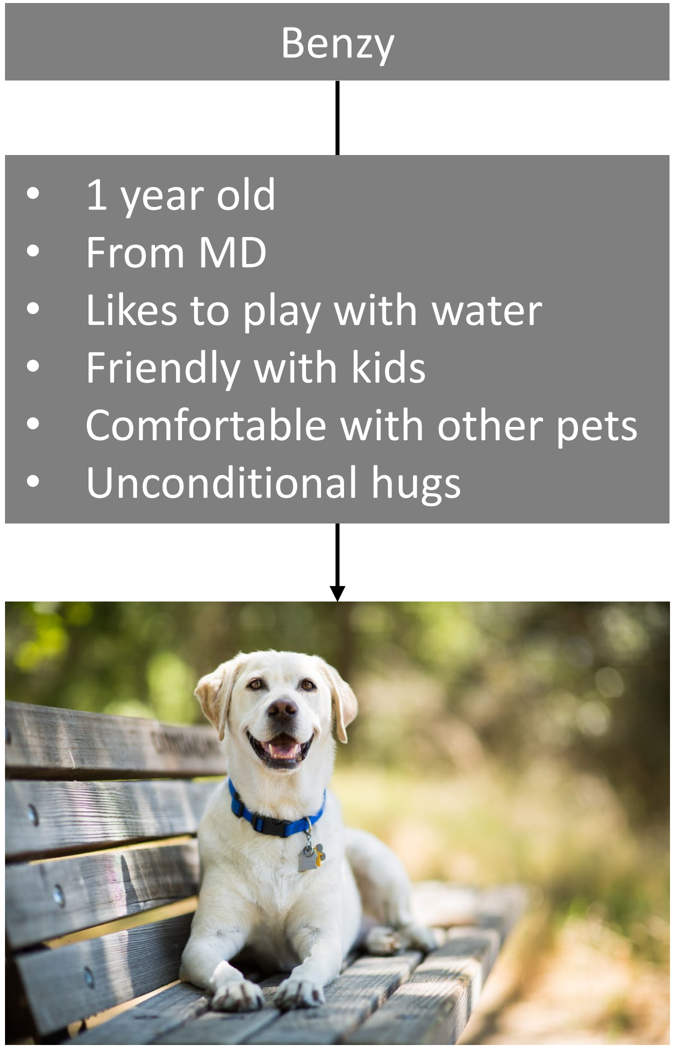
- Examples of embeddings: Word Embeddings, Sentence Embeddings, Document Embeddings, Image Embeddings, Graph Embeddings, Knowledge Graph Embeddings



https://pubmed.ncbi.nlm.nih.gov/29218877/

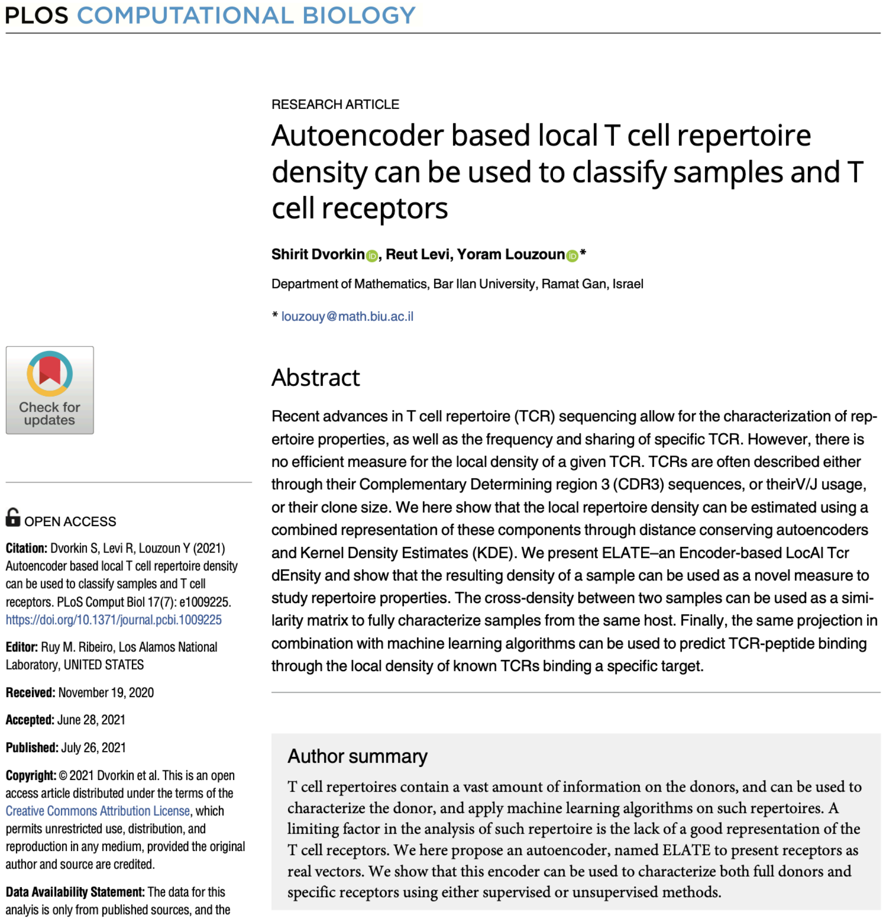
Encoding: Concept & Example

- Encoding refers to the process of representing data in a specific format or representation.
- It involves transforming data from its original representation into a different format that is suitable for a particular purpose or task.
- The goal of encoding is often to make the data compatible with a specific learning algorithm or to facilitate efficient processing.
- Examples of encoding: One-Hot Encoding, Label Encoding, Binary Encoding, Ordinal Encoding, Hash Encoding, Target Encoding, Feature Hashing



Benzy

- 1 year old
- From MD
- Likes to play with water
- Friendly with kids
- Comfortable with other pets
- Unconditional hugs



PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Autoencoder based local T cell repertoire density can be used to classify samples and T cell receptors

Shirit Dvorkin, Reut Levi, Yoram Louzoun*

Department of Mathematics, Bar Ilan University, Ramat Gan, Israel

* louzouy@math.biu.ac.il

Check for updates

OPEN ACCESS

Citation: Dvorkin S, Levi R, Louzoun Y (2021) Autoencoder based local T cell repertoire density can be used to classify samples and T cell receptors. PLoS Comput Biol 17(7): e1009225. https://doi.org/10.1371/journal.pcbi.1009225

Editor: Ruy M. Ribeiro, Los Alamos National Laboratory, UNITED STATES

Received: November 19, 2020

Accepted: June 28, 2021

Published: July 26, 2021

Data Availability Statement: The data for this analyis is only from published sources, and the

Abstract

Recent advances in T cell repertoire (TCR) sequencing allow for the characterization of repertoire properties, as well as the frequency and sharing of specific TCR. However, there is no efficient measure for the local density of a given TCR. TCRs are often described either through their Complementary Determining region 3 (CDR3) sequences, or theirV/J usage, or their clone size. We here show that the local repertoire density can be estimated using a combined representation of these components through distance conserving autoencoders and Kernel Density Estimates (KDE). We present ELATE–an Encoder-based LocAl Tcr dEnsity and show that the resulting density of a sample can be used as a novel measure to study repertoire properties. The cross-density between two samples can be used as a similarity matrix to fully characterize samples from the same host. Finally, the same projection in combination with machine learning algorithms can be used to predict TCR-peptide binding through the local density of known TCRs binding a specific target.

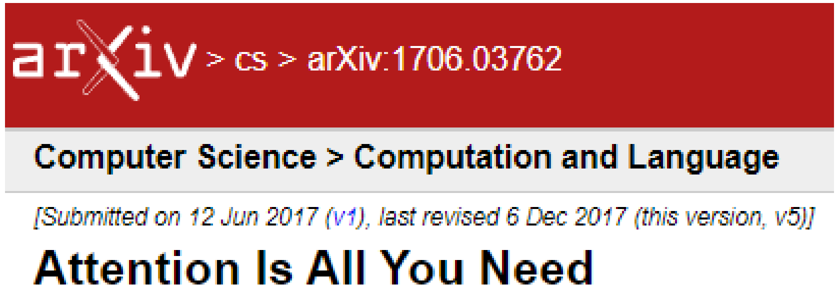
Author summary

T cell repertoires contain a vast amount of information on the donors, and can be used to characterize the donor, and apply machine learning algorithms on such repertoires. A limiting factor in the analysis of such repertoire is the lack of a good representation of the T cell receptors. We here propose an autoencoder, named ELATE to present receptors as real vectors. We show that this encoder can be used to characterize both full donors and specific receptors using either supervised or unsupervised methods.

https://www.biorxiv.org/content/10.1101/2021.02.28.433289v2

Attention: Concept & Example

- Attention in AI refers to a mechanism that enables models to focus on specific parts of input data while performing a task.
- It mimics the selective attention mechanism observed in human cognition, where we prioritize certain information over others.

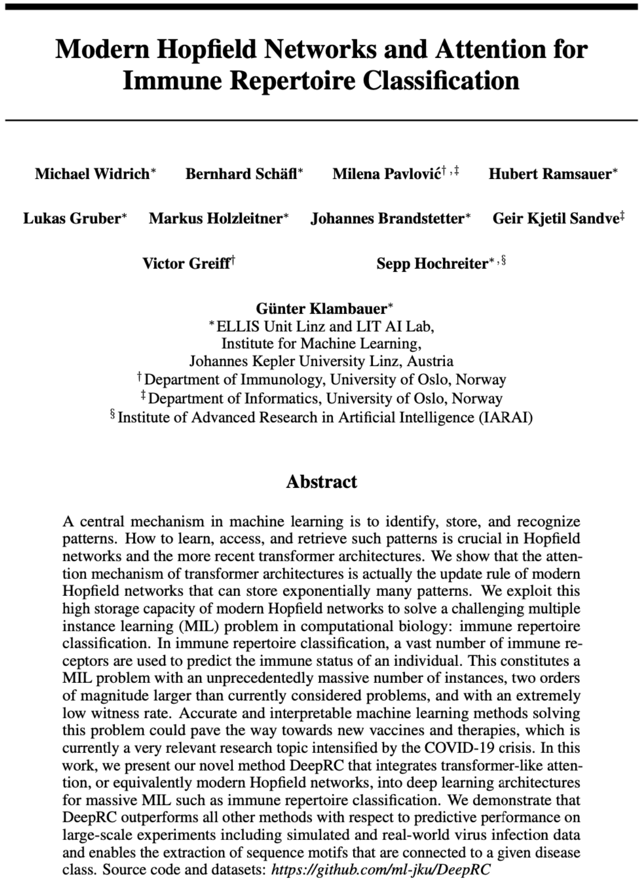


arXiv > cs > arXiv:1706.03762

Computer Science > Computation and Language

[Submitted on 12 Jun 2017 (v1), last revised 6 Dec 2017 (this version, v5)]

Attention Is All You Need



Modern Hopfield Networks and Attention for Immune Repertoire Classification

Michael Widrich* Bernhard Schäfl* Milena Pavlović†,‡ Hubert Ramsauer*
Lukas Gruber* Markus Holzleitner* Johannes Brandstetter* Geir Kjetil Sandve‡
Victor Greiff† Sepp Hochreiter*,§
Günter Klambauer*
*ELLIS Unit Linz and LIT AI Lab,
Institute for Machine Learning,
Johannes Kepler University Linz, Austria
†Department of Immunology, University of Oslo, Norway
‡Department of Informatics, University of Oslo, Norway
§Institute of Advanced Research in Artificial Intelligence (IARAI)

Abstract

A central mechanism in machine learning is to identify, store, and recognize patterns. How to learn, access, and retrieve such patterns is crucial in Hopfield networks and the more recent transformer architectures. We show that the attention mechanism of transformer architectures is actually the update rule of modern Hopfield networks that can store exponentially many patterns. We exploit this high storage capacity of modern Hopfield networks to solve a challenging multiple instance learning (MIL) problem in computational biology: immune repertoire classification. In immune repertoire classification, a vast number of immune receptors are used to predict the immune status of an individual. This constitutes a MIL problem with an unprecedentedly massive number of instances, two orders of magnitude larger than currently considered problems, and with an extremely low witness rate. Accurate and interpretable machine learning methods solving this problem could pave the way towards new vaccines and therapies, which is currently a very relevant research topic intensified by the COVID-19 crisis. In this work, we present our novel method DeepRC that integrates transformer-like attention, or equivalently modern Hopfield networks, into deep learning architectures for massive MIL such as immune repertoire classification. We demonstrate that DeepRC outperforms all other methods with respect to predictive performance on large-scale experiments including simulated and real-world virus infection data and enables the extraction of sequence motifs that are connected to a given disease class. Source code and datasets: *https://github.com/ml-jku/DeepRC*

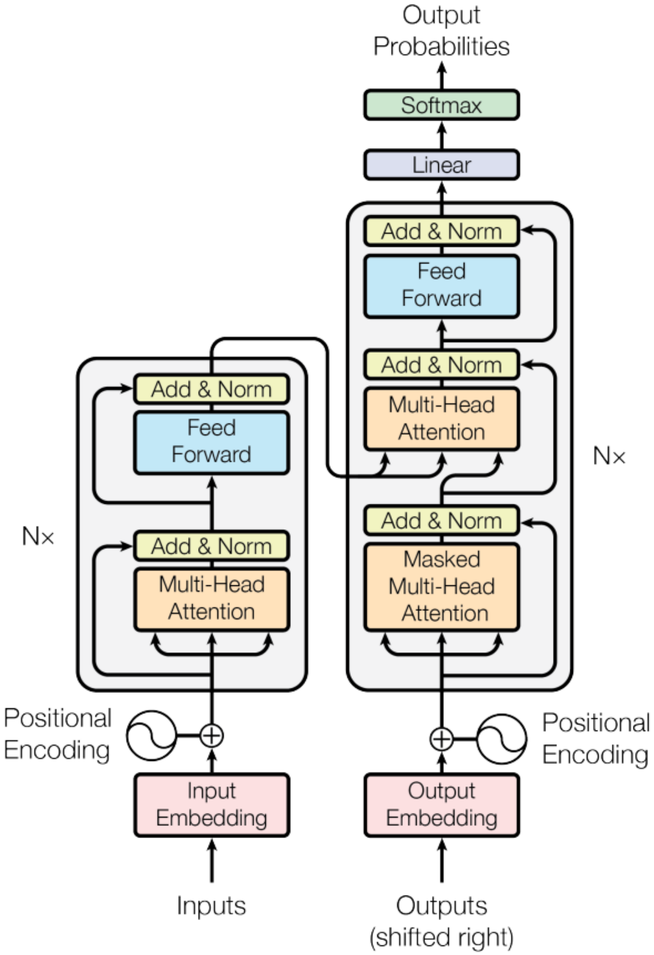
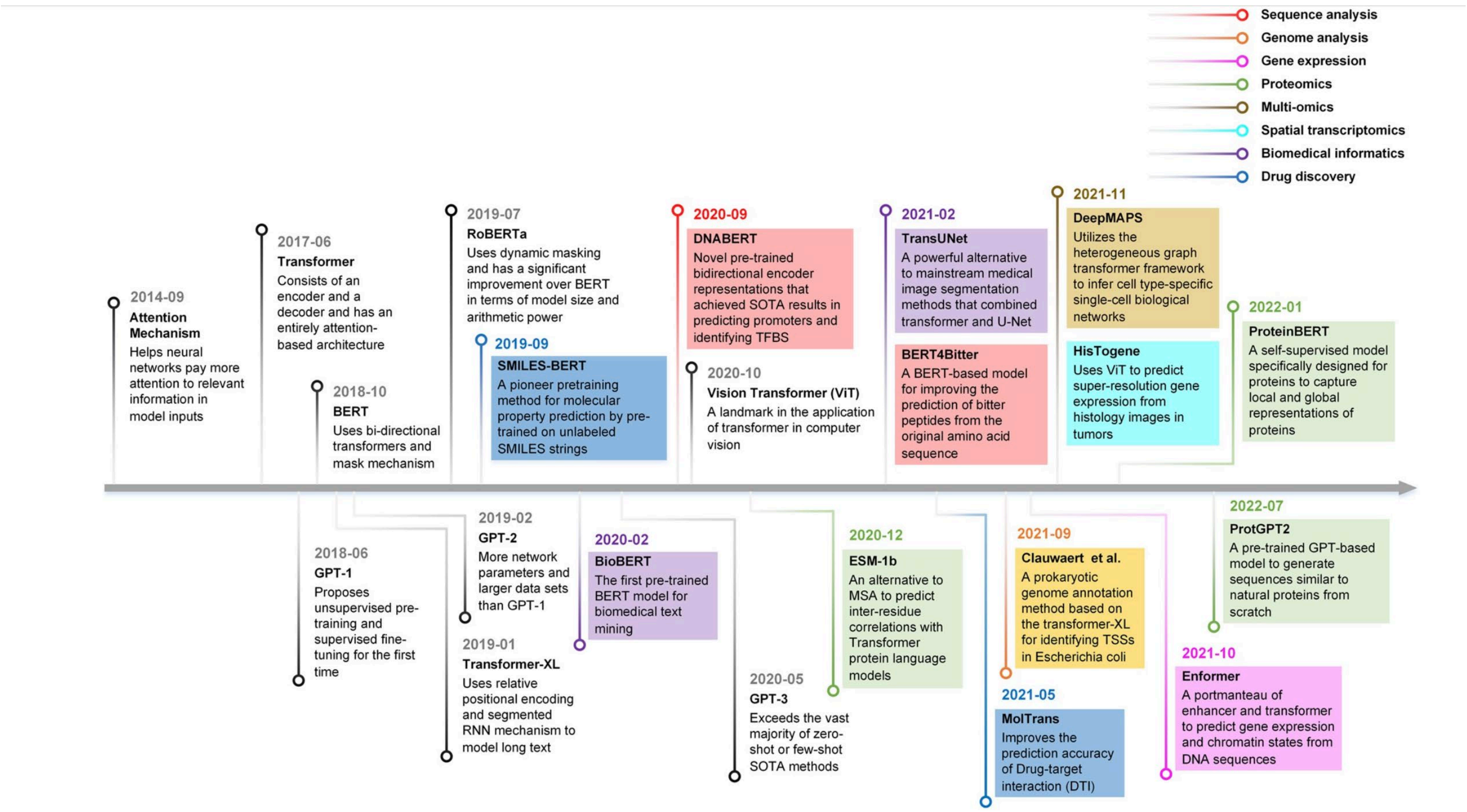


Figure 1: The Transformer - model architecture.

https://proceedings.neurips.cc/paper/2020/file/da4902cb0bc38210839714ebdcf0efc3-Paper.pdf

Transformers: Concept & Example

- Transformer refers to a type of neural network architecture that has gained significant popularity, particularly in the field of natural language processing (NLP)
- The transformer architecture is designed to process sequential data efficiently, such as sentences, paragraphs, or time series data.
- Transformers employs a mechanism called self-attention or scaled dot-product attention to capture relationships and dependencies between different elements of the input sequence.
- Transformers have achieved remarkable success in various NLP tasks, including machine translation, language generation, sentiment analysis, and text classification.



2014-09 **Attention Mechanism** Helps neural networks pay more attention to relevant information in model inputs

2017-06 **Transformer** Consists of an encoder and a decoder and has an entirely attention-based architecture

2018-06 **GPT-1** Proposes unsupervised pre-training and supervised fine-tuning for the first time

2018-10 **BERT** Uses bi-directional transformers and mask mechanism

2019-01 **Transformer-XL** Uses relative positional encoding and segmented RNN mechanism to model long text

2019-02 **GPT-2** More network parameters and larger data sets than GPT-1

2019-07 **RoBERTa** Uses dynamic masking and has a significant improvement over BERT in terms of model size and arithmetic power

2019-09 **SMILES-BERT** A pioneer pretraining method for molecular property prediction by pre-trained on unlabeled SMILES strings

2020-02 **BioBERT** The first pre-trained BERT model for biomedical text mining

2020-05 **GPT-3** Exceeds the vast majority of zero-shot or few-shot SOTA methods

2020-09 **DNABERT** Novel pre-trained bidirectional encoder representations that achieved SOTA results in predicting promoters and identifying TFBS

2020-10 **Vision Transformer (ViT)** A landmark in the application of transformer in computer vision

2020-12 **ESM-1b** An alternative to MSA to predict inter-residue correlations with Transformer protein language models

2021-02 **TransUNet** A powerful alternative to mainstream medical image segmentation methods that combined transformer and U-Net

BERT4Bitter A BERT-based model for improving the prediction of bitter peptides from the original amino acid sequence

2021-05 **MolTrans** Improves the prediction accuracy of Drug-target interaction (DTI)

2021-09 **Clauwaert et al.** A prokaryotic genome annotation method based on the transformer-XL for identifying TSSs in Escherichia coli

2021-10 **Enformer** A portmanteau of enhancer and transformer to predict gene expression and chromatin states from DNA sequences

2021-11 **DeepMAPS** Utilizes the heterogeneous graph transformer framework to infer cell type-specific single-cell biological networks

HisTogene Uses ViT to predict super-resolution gene expression from histology images in tumors

2022-01 **ProteinBERT** A self-supervised model specifically designed for proteins to capture local and global representations of proteins

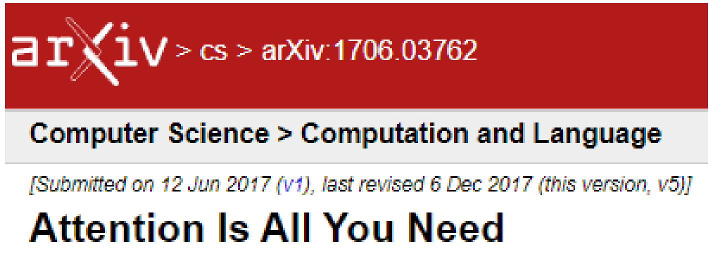
2022-07 **ProtGPT2** A pre-trained GPT-based model to generate sequences similar to natural proteins from scratch

GPT (Generative Pre-trained Transformers)

- A well-known transformer models is the Generative Pre-trained Transformer (GPT)
- Developed by OpenAI

https://www.biorxiv.org/content/10.1101/2021.02.28.433289v2

Large Language Models are Transformers



arXiv > cs > arXiv:1706.03762

Computer Science > Computation and Language

[Submitted on 12 Jun 2017 (v1), last revised 6 Dec 2017 (this version, v5)]

Attention Is All You Need

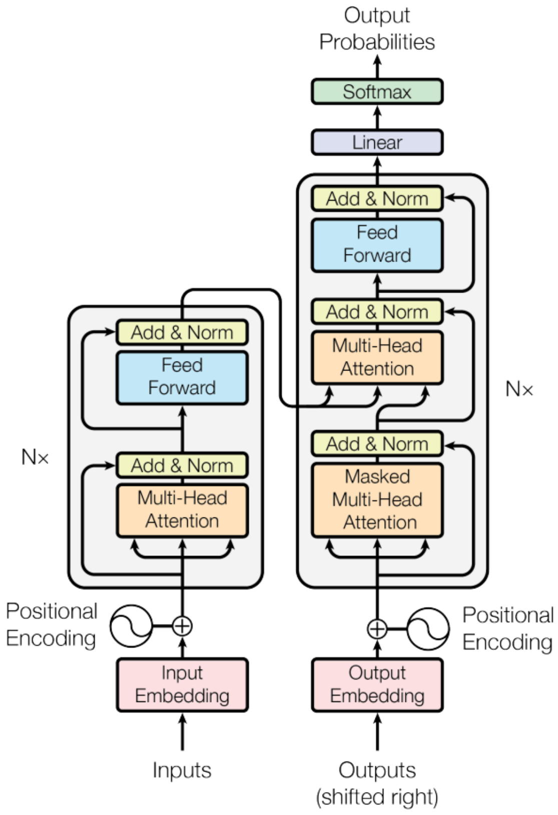


Figure 1: The Transformer - model architecture.

BERT

Bidirectional encoder representations from transformers

OpenAI

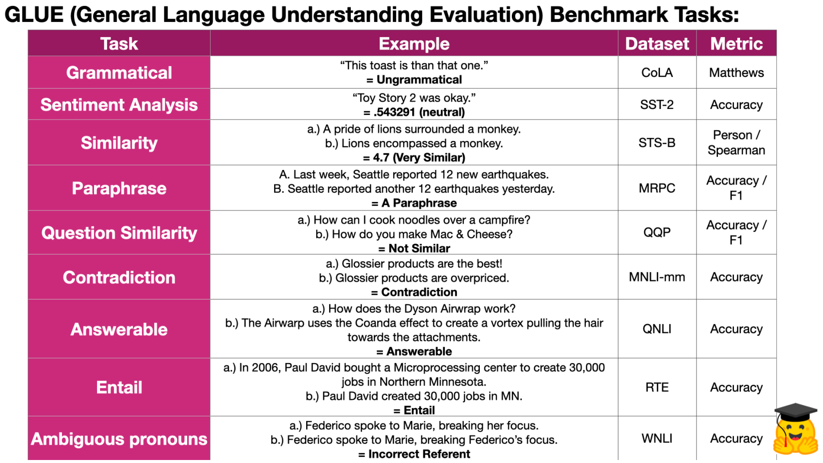
GPT-4

Generative pretrained transformer

Many others

T5, LLaMA, Bard, open source models

Model	Training	Parameters	Year
BERT	3.3B words	340M	2018
GPT-3	500B tokens	175B	2020
ChatGPT	300B words	1.5B	2022
LLaMA	1.4T tokens	65B	2023

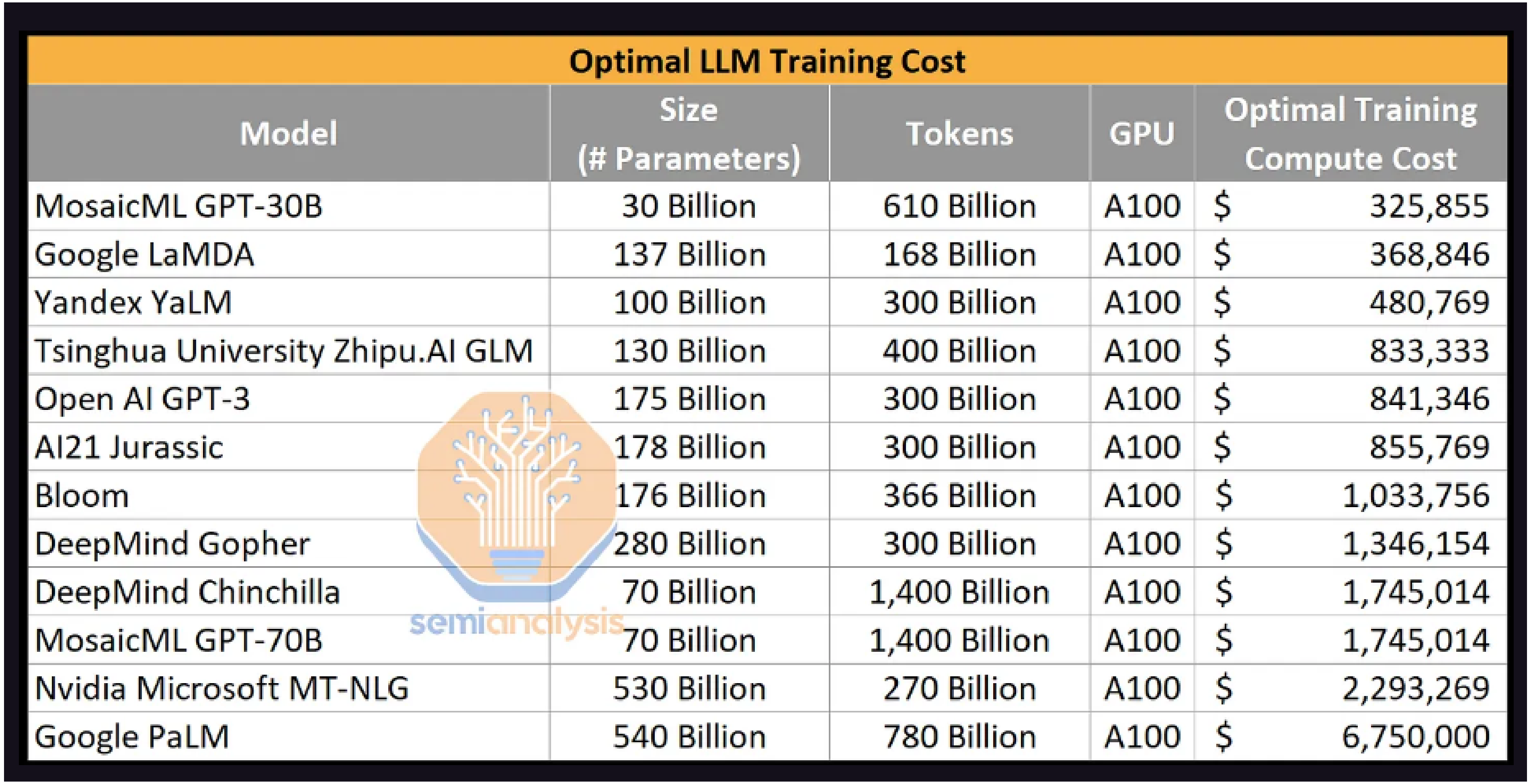


GLUE (General Language Understanding Evaluation) Benchmark Tasks:

Task	Example	Dataset	Metric
Grammatical	"This toast is than that one." **= Ungrammatical**	CoLA	Matthews
Sentiment Analysis	"Toy Story 2 was okay." **= .543291 (neutral)**	SST-2	Accuracy
Similarity	a.) A pride of lions surrounded a monkey. b.) Lions encompassed a monkey. **= 4.7 (Very Similar)**	STS-B	Person / Spearman
Paraphrase	A. Last week, Seattle reported 12 new earthquakes. B. Seattle reported another 12 earthquakes yesterday. **= A Paraphrase**	MRPC	Accuracy / F1
Question Similarity	a.) How can I cook noodles over a campfire? b.) How do you make Mac & Cheese? **= Not Similar**	QQP	Accuracy / F1
Contradiction	a.) Glossier products are the best! b.) Glossier products are overpriced. **= Contradiction**	MNLI-mm	Accuracy
Answerable	a.) How does the Dyson Airwrap work? b.) The Airwarp uses the Coanda effect to create a vortex pulling the hair towards the attachments. **= Answerable**	QNLI	Accuracy
Entail	a.) In 2006, Paul David bought a Microprocessing center to create 30,000 jobs in Northern Minnesota. b.) Paul David created 30,000 jobs in MN. **= Entail**	RTE	Accuracy
Ambiguous pronouns	a.) Federico spoke to Marie, breaking her focus. b.) Federico spoke to Marie, breaking Federico's focus. **= Incorrect Referent**	WNLI	Accuracy

https://huggingface.co/blog/bert-101

Large Language Models are large and expensive!



Optimal LLM Training Cost

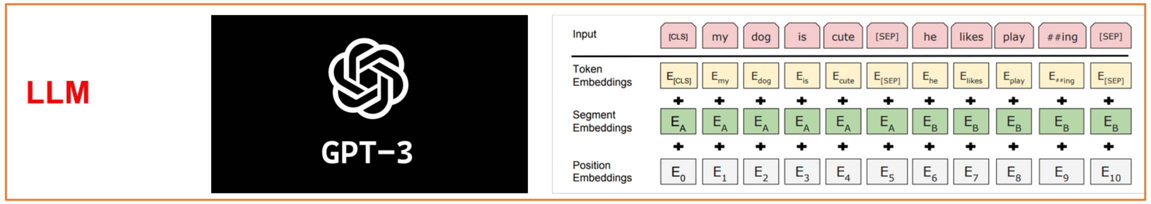
Model	Size (# Parameters)	Tokens	GPU	Optimal Training Compute Cost
MosaicML GPT-30B	30 Billion	610 Billion	A100	$ 325,855
Google LaMDA	137 Billion	168 Billion	A100	$ 368,846
Yandex YaLM	100 Billion	300 Billion	A100	$ 480,769
Tsinghua University Zhipu.AI GLM	130 Billion	400 Billion	A100	$ 833,333
Open AI GPT-3	175 Billion	300 Billion	A100	$ 841,346
AI21 Jurassic	178 Billion	300 Billion	A100	$ 855,769
Bloom	176 Billion	366 Billion	A100	$ 1,033,756
DeepMind Gopher	280 Billion	300 Billion	A100	$ 1,346,154
DeepMind Chinchilla	70 Billion	1,400 Billion	A100	$ 1,745,014
MosaicML GPT-70B	70 Billion	1,400 Billion	A100	$ 1,745,014
Nvidia Microsoft MT-NLG	530 Billion	270 Billion	A100	$ 2,293,269
Google PaLM	540 Billion	780 Billion	A100	$ 6,750,000

Source: semianalysis.com (calculated using Chinchilla pricing)

Architecture of GPT

GPT (Generative Pre-trained Transformers)

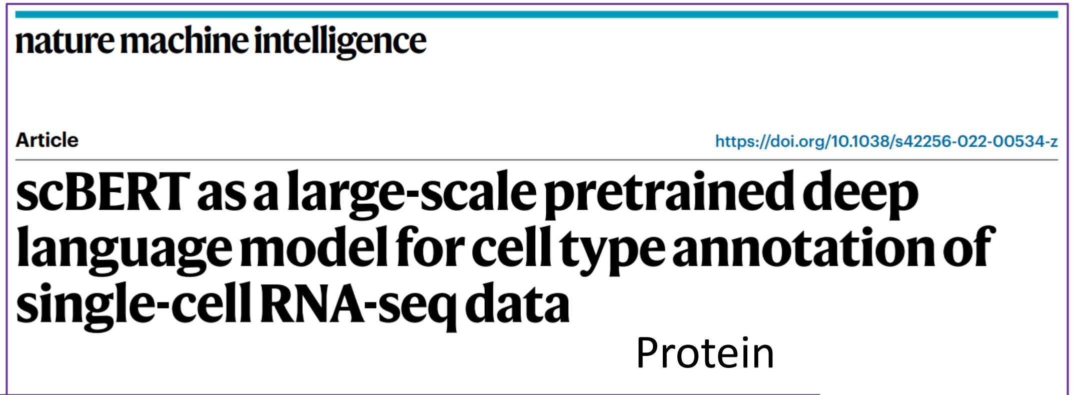
- A well-known transformer models is the gpt-3.5-turbo
- Pre-trained Transformer (GPT)
- Developed by OpenAI


LLM
GPT−3
Input
[CLS] my dog is cute [SEP] he likes play ##ing [SEP]
Token Embeddings
$E_{[CLS]}$ E_{my} E_{dog} E_{is} E_{cute} $E_{[SEP]}$ E_{he} E_{likes} E_{play} $E_{\#\#ing}$ $E_{[SEP]}$
Segment Embeddings
E_A E_A E_A E_A E_A E_A E_B E_B E_B E_B E_B
Position Embeddings
E_0 E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9 E_{10}

GPT	ChatGPT
Transformer Encoder ↓ Pre-training ↓ Transformer Decoder ↓ Attention Mechanism ↓ Positional Encoding ↓ Fine-tuning	Transformer Encoder ↓ Context Window ↓ Language Generation ↓ Fine-tuning

Large Language Models in Biomedicine

Single cell



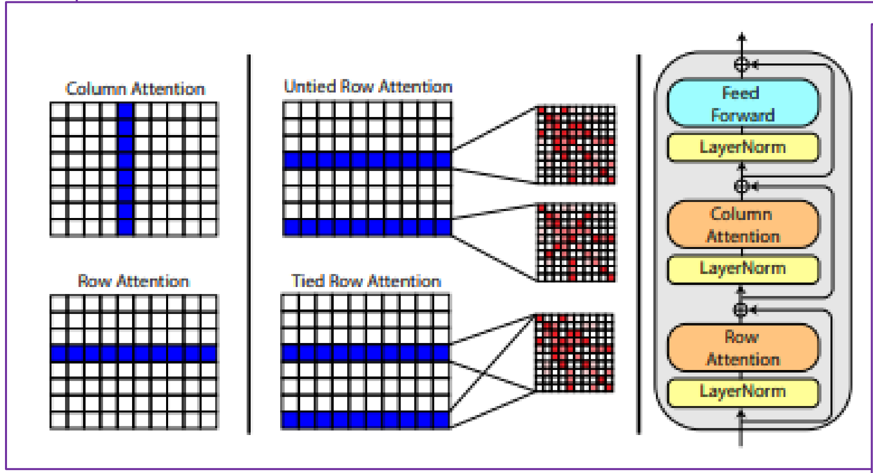
nature machine intelligence

Article

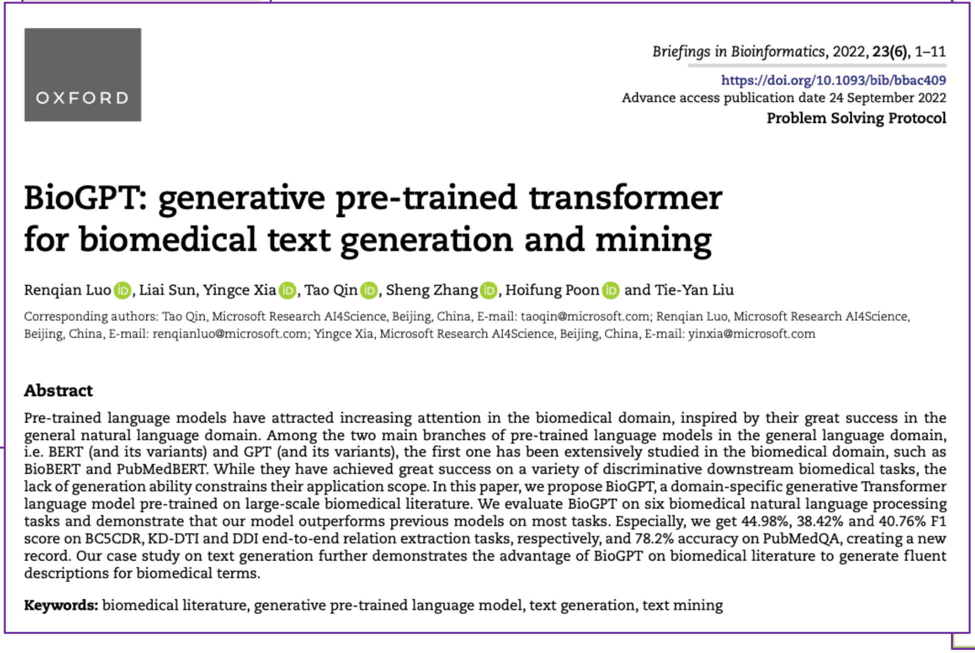
https://doi.org/10.1038/s42256-022-00534-z

scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data

Protein



PubMed



OXFORD

Briefings in Bioinformatics, 2022, **23(6)**, 1–11
https://doi.org/10.1093/bib/bbac409
Advance access publication date 24 September 2022
Problem Solving Protocol

BioGPT: generative pre-trained transformer for biomedical text generation and mining

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon and Tie-Yan Liu

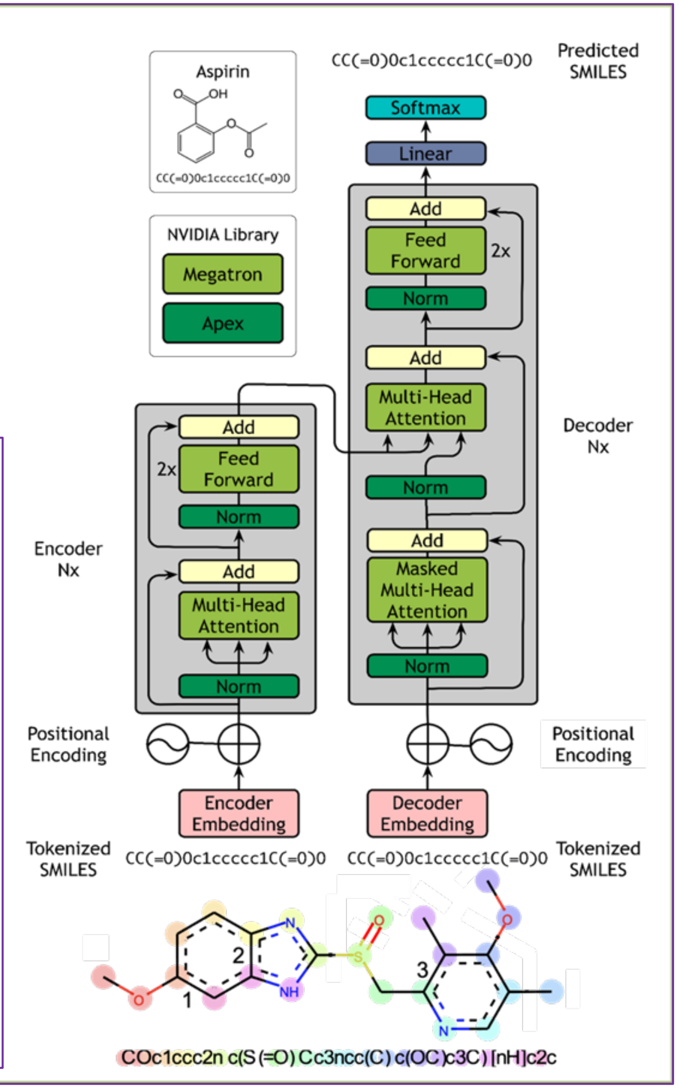
Corresponding authors: Tao Qin, Microsoft Research AI4Science, Beijing, China, E-mail: taoqin@microsoft.com; Renqian Luo, Microsoft Research AI4Science, Beijing, China, E-mail: renqianluo@microsoft.com; Yingce Xia, Microsoft Research AI4Science, Beijing, China, E-mail: yinxia@microsoft.com

Abstract

Pre-trained language models have attracted increasing attention in the biomedical domain, inspired by their great success in the general natural language domain. Among the two main branches of pre-trained language models in the general language domain, i.e. BERT (and its variants) and GPT (and its variants), the first one has been extensively studied in the biomedical domain, such as BioBERT and PubMedBERT. While they have achieved great success on a variety of discriminative downstream biomedical tasks, the lack of generation ability constrains their application scope. In this paper, we propose BioGPT, a domain-specific generative Transformer language model pre-trained on large-scale biomedical literature. We evaluate BioGPT on six biomedical natural language processing tasks and demonstrate that our model outperforms previous models on most tasks. Especially, we get 44.98%, 38.42% and 40.76% F1 score on BC5CDR, KD-DTI and DDI end-to-end relation extraction tasks, respectively, and 78.2% accuracy on PubMedQA, creating a new record. Our case study on text generation further demonstrates the advantage of BioGPT on biomedical literature to generate fluent descriptions for biomedical terms.

Keywords: biomedical literature, generative pre-trained language model, text generation, text mining

Small molecule

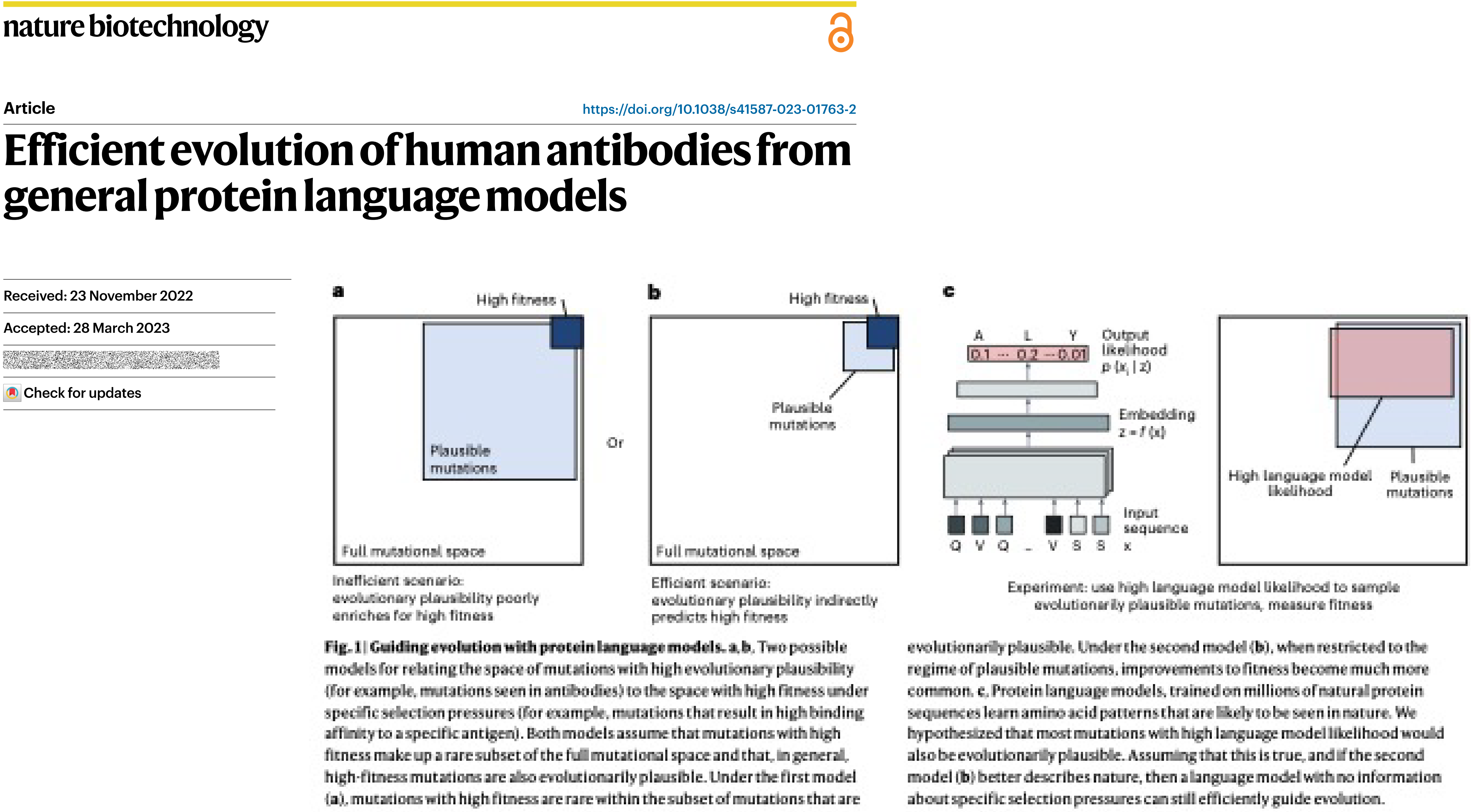


Language generation

Relation extraction

Q&A style interface

Language Models in Biomedicine: Protein Language Model



nature biotechnology

Article

https://doi.org/10.1038/s41587-023-01763-2

Efficient evolution of human antibodies from general protein language models

Received: 23 November 2022

Accepted: 28 March 2023

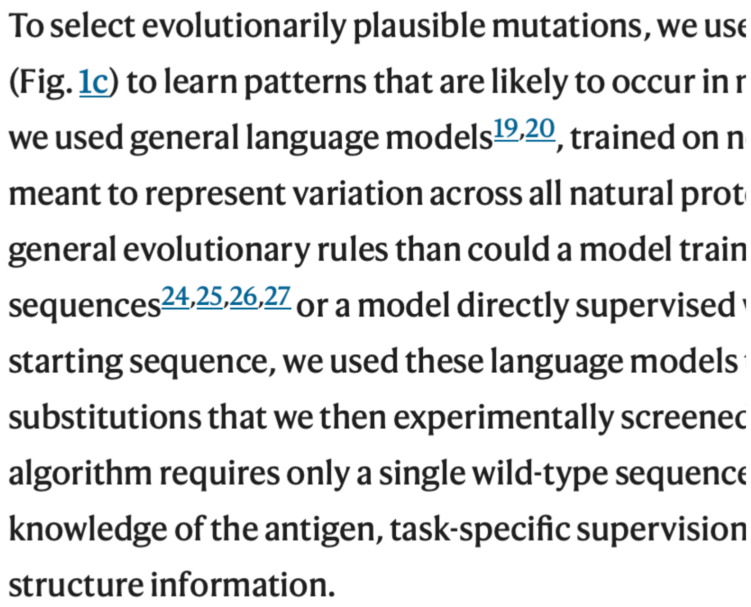
[redacted]

Check for updates

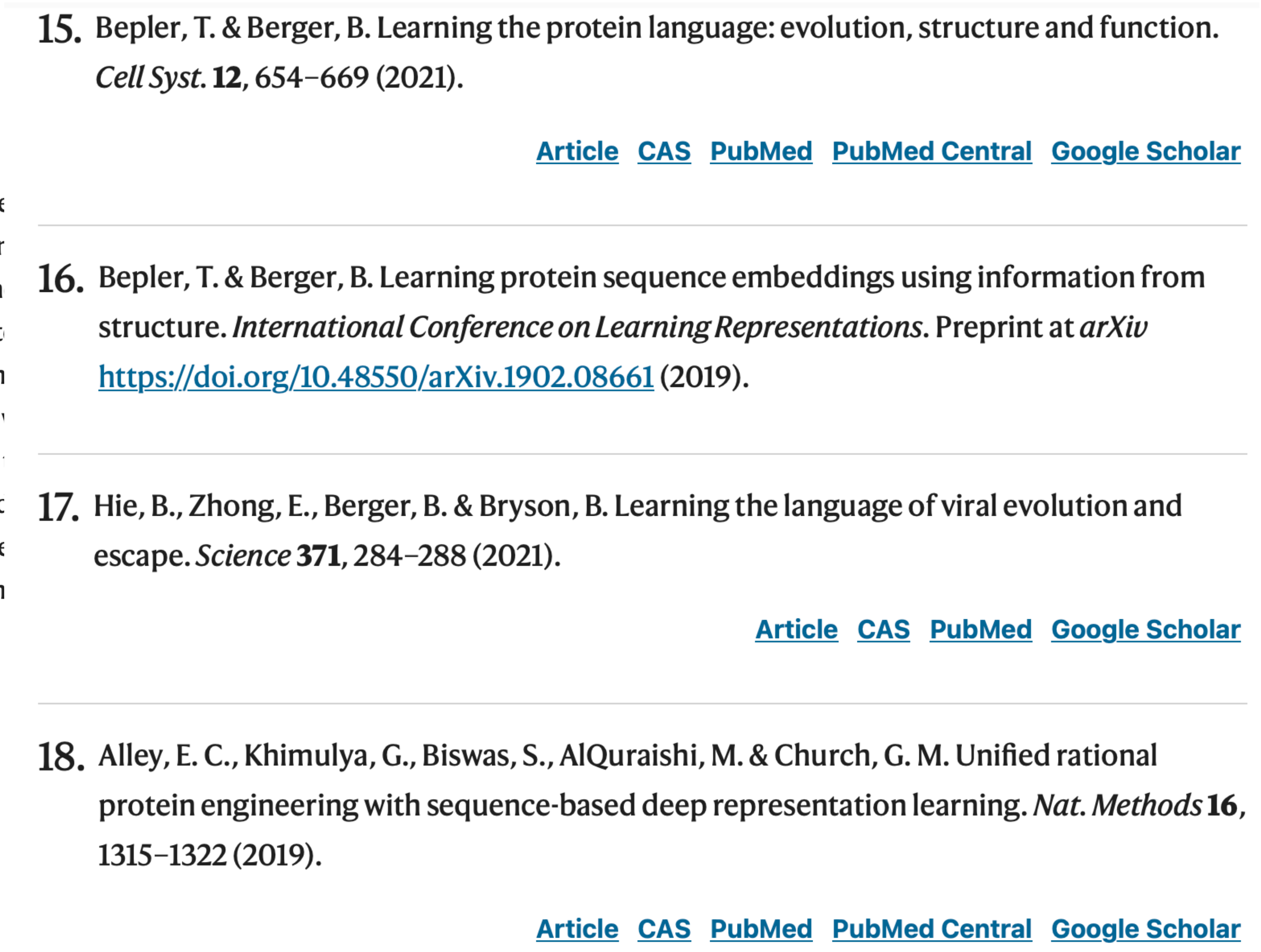
a
High fitness
Plausible mutations
Full mutational space
Inefficient scenario: evolutionary plausibility poorly enriches for high fitness
Or
b
High fitness
Plausible mutations
Full mutational space
Efficient scenario: evolutionary plausibility indirectly predicts high fitness
c
A
L
Y
0.1 ··· 0.2 ···0.01
Output likelihood $p(x_i \mid z)$
Embedding $z = f(x)$
Input sequence
Q V Q _ V S S x
High language model likelihood
Plausible mutations
Experiment: use high language model likelihood to sample evolutionarily plausible mutations, measure fitness

Fig. 1 | Guiding evolution with protein language models. a,b, Two possible models for relating the space of mutations with high evolutionary plausibility (for example, mutations seen in antibodies) to the space with high fitness under specific selection pressures (for example, mutations that result in high binding affinity to a specific antigen). Both models assume that mutations with high fitness make up a rare subset of the full mutational space and that, in general, high-fitness mutations are also evolutionarily plausible. Under the first model (**a**), mutations with high fitness are rare within the subset of mutations that are evolutionarily plausible. Under the second model (**b**), when restricted to the regime of plausible mutations, improvements to fitness become much more common. **c**, Protein language models, trained on millions of natural protein sequences learn amino acid patterns that are likely to be seen in nature. We hypothesized that most mutations with high language model likelihood would also be evolutionarily plausible. Assuming that this is true, and if the second model (**b**) better describes nature, then a language model with no information about specific selection pressures can still efficiently guide evolution.

Language Models in Biomedicine: Protein Language Model



To select evolutionarily plausible mutations, we use
(Fig. 1c) to learn patterns that are likely to occur in r
we used general language models[19,20], trained on n
meant to represent variation across all natural prot
general evolutionary rules than could a model train
sequences[24,25,26,27] or a model directly supervised
starting sequence, we used these language models
substitutions that we then experimentally screenec
algorithm requires only a single wild-type sequence
knowledge of the antigen, task-specific supervision
structure information.



15. Bepler, T. & Berger, B. Learning the protein language: evolution, structure and function. *Cell Syst.* **12**, 654–669 (2021).

Article CAS PubMed PubMed Central Google Scholar

16. Bepler, T. & Berger, B. Learning protein sequence embeddings using information from structure. *International Conference on Learning Representations*. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1902.08661 (2019).

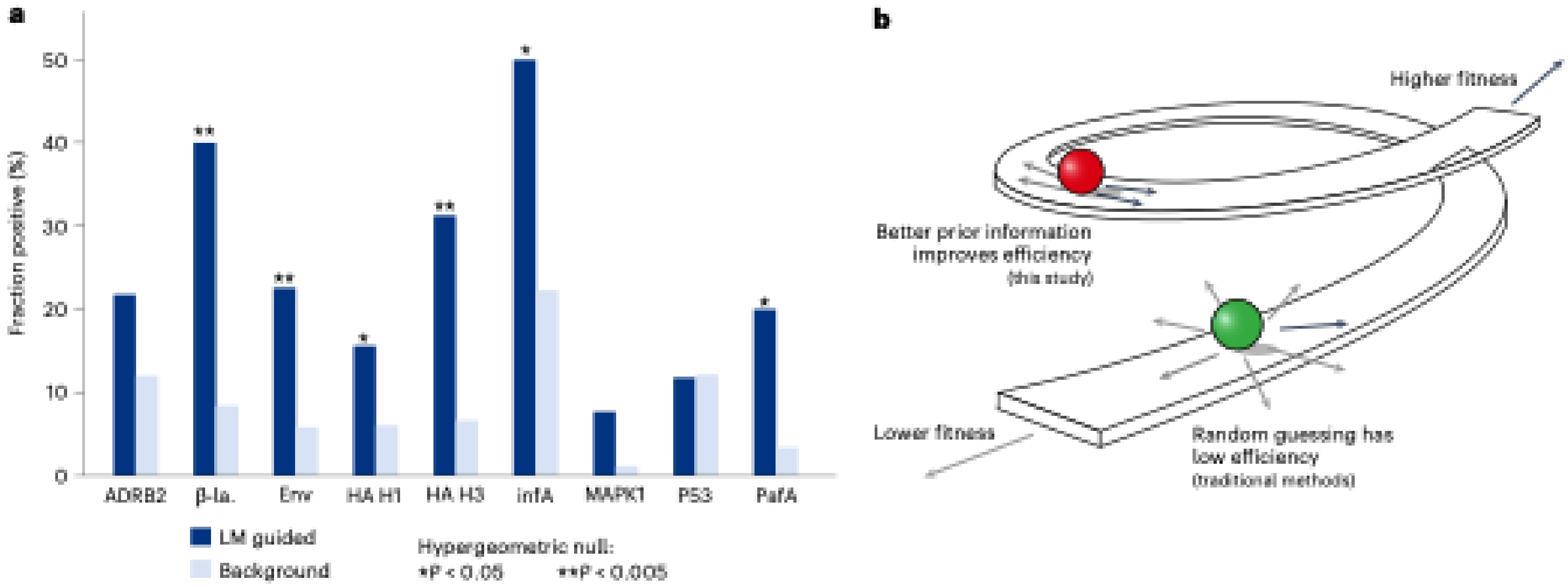
17. Hie, B., Zhong, E., Berger, B. & Bryson, B. Learning the language of viral evolution and escape. *Science* **371**, 284–288 (2021).

Article CAS PubMed Google Scholar

18. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **16**, 1315–1322 (2019).

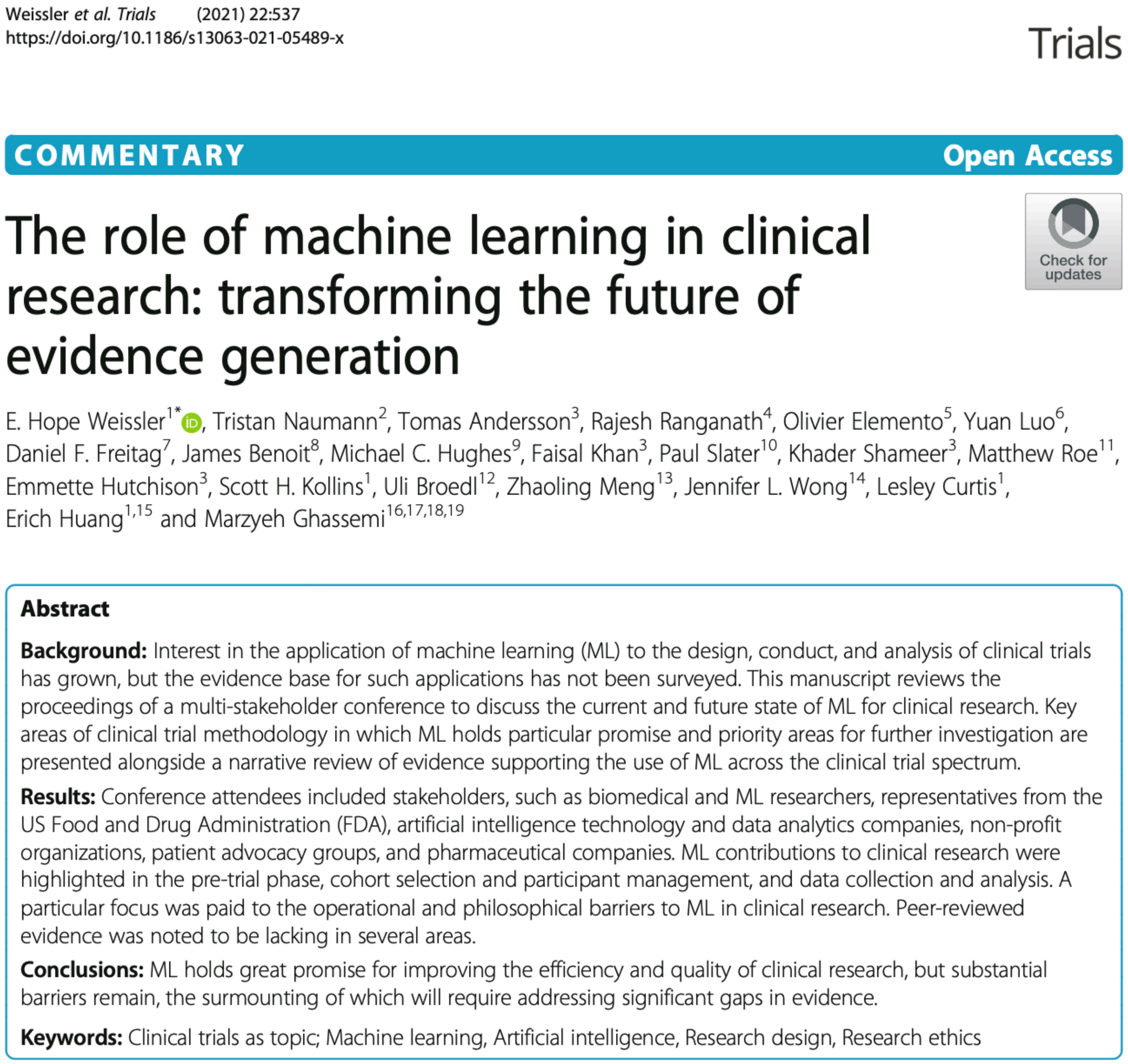
Article CAS PubMed PubMed Central Google Scholar

Language Models in Biomedicine: Protein Language Model



'ig. 4 | **Guiding evolution without explicitly modeling fitness. a**, The same trategy and language models that we use to affinity mature antibodies can also ecommend high-fitness changes across a diversity of selection pressures and rotein families, as identified experimentally using high-throughput scanning nutagenesis assays[8,48] (described in Supplementary Table 13). 'Fraction positive' ndicates the percentage of high-fitness amino acid substitutions within either he set of substitutions recommended by the language model (LM guided) r the set of all single-residue substitutions (Background). A large portion of anguage-model-guided substitutions have high fitness, which, in many cases, s significantly enriched compared to the background percentage; also see xtended Data Figs. 4–6, and see Supplementary Table 13 for the exact one-sided ypergeometric *P* values and sample sizes. ADRB2, adrenoreceptor beta 2; β-la., β-lactamase; Env, envelope glycoprotein; infA, translation initiation factor 1; MAPK1, mitogen-activated protein kinase 1; PafA, phosphate-irrepressible alkaline phosphatase. **b**, Conceptually, the prior information encoded by evolutionary plausibility is represented in this cartoon by the rainbow road, where ascending corresponds to improving fitness and descending corresponds to lowering fitness. Moving in any direction (for example, via random or brute force mutagenesis) would most likely decrease fitness or have a high chance of being a detrimental change (represented by the green ball). However, if evolutionary plausibility is an efficient prior (Fig. 1b), then movement that is constrained to the plausible regime (for example, when guided by a language model) substantially increases the chance of improving fitness (represented by the red ball).

Opportunities in developing AI-driven precision clinical trials



Weissler *et al. Trials* (2021) 22:537
https://doi.org/10.1186/s13063-021-05489-x

Trials

COMMENTARY **Open Access**

The role of machine learning in clinical research: transforming the future of evidence generation

E. Hope Weissler[1*], Tristan Naumann[2], Tomas Andersson[3], Rajesh Ranganath[4], Olivier Elemento[5], Yuan Luo[6], Daniel F. Freitag[7], James Benoit[8], Michael C. Hughes[9], Faisal Khan[3], Paul Slater[10], Khader Shameer[3], Matthew Roe[11], Emmette Hutchison[3], Scott H. Kollins[1], Uli Broedl[12], Zhaoling Meng[13], Jennifer L. Wong[14], Lesley Curtis[1], Erich Huang[1,15] and Marzyeh Ghassemi[16,17,18,19]

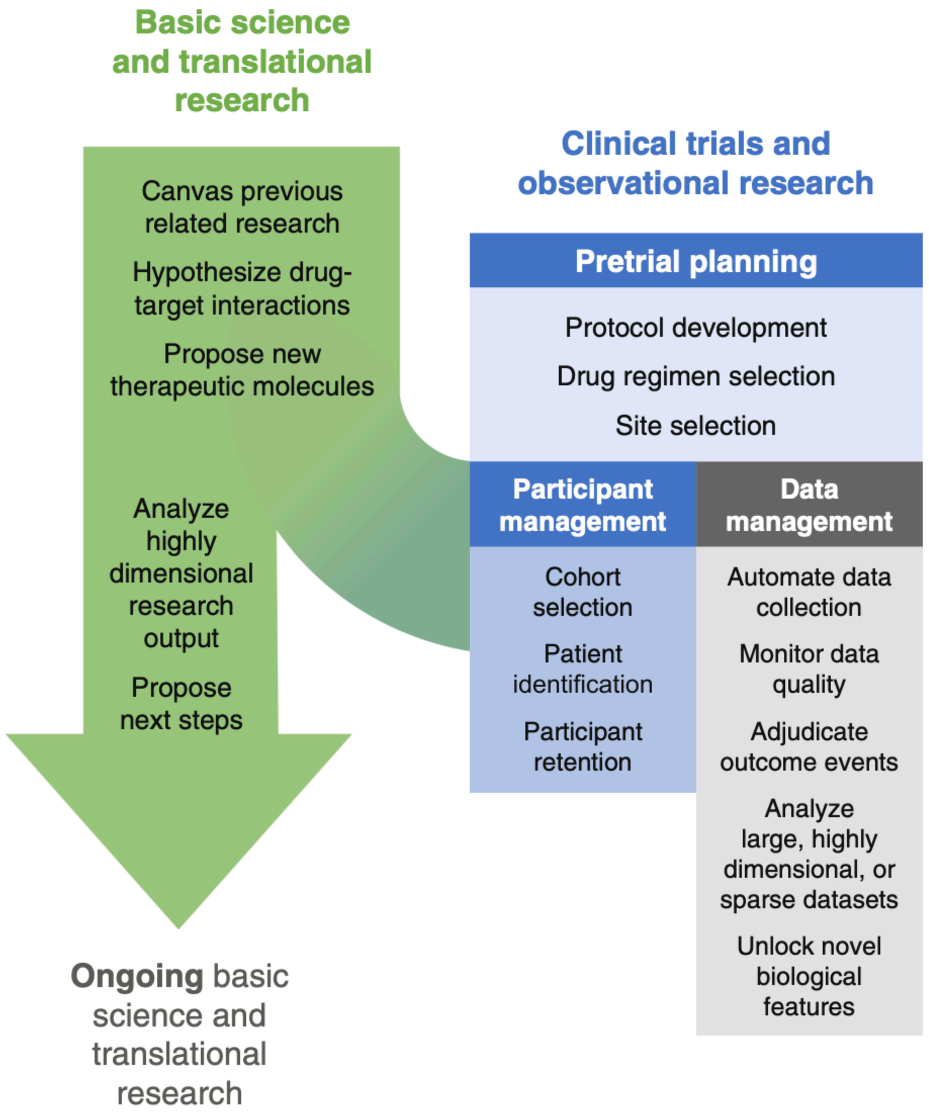
Abstract

Background: Interest in the application of machine learning (ML) to the design, conduct, and analysis of clinical trials has grown, but the evidence base for such applications has not been surveyed. This manuscript reviews the proceedings of a multi-stakeholder conference to discuss the current and future state of ML for clinical research. Key areas of clinical trial methodology in which ML holds particular promise and priority areas for further investigation are presented alongside a narrative review of evidence supporting the use of ML across the clinical trial spectrum.

Results: Conference attendees included stakeholders, such as biomedical and ML researchers, representatives from the US Food and Drug Administration (FDA), artificial intelligence technology and data analytics companies, non-profit organizations, patient advocacy groups, and pharmaceutical companies. ML contributions to clinical research were highlighted in the pre-trial phase, cohort selection and participant management, and data collection and analysis. A particular focus was paid to the operational and philosophical barriers to ML in clinical research. Peer-reviewed evidence was noted to be lacking in several areas.

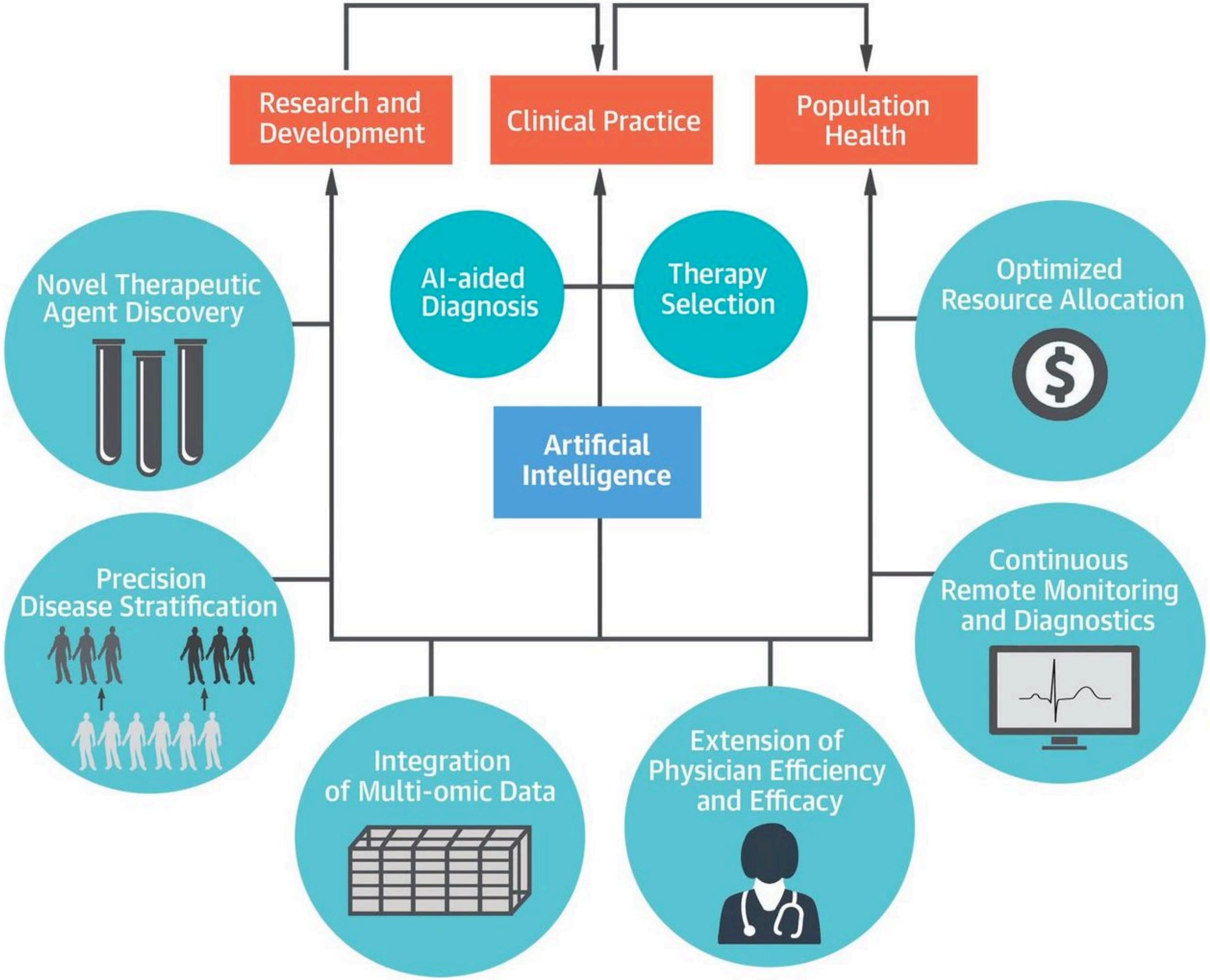
Conclusions: ML holds great promise for improving the efficiency and quality of clinical research, but substantial barriers remain, the surmounting of which will require addressing significant gaps in evidence.

Keywords: Clinical trials as topic; Machine learning, Artificial intelligence, Research design, Research ethics



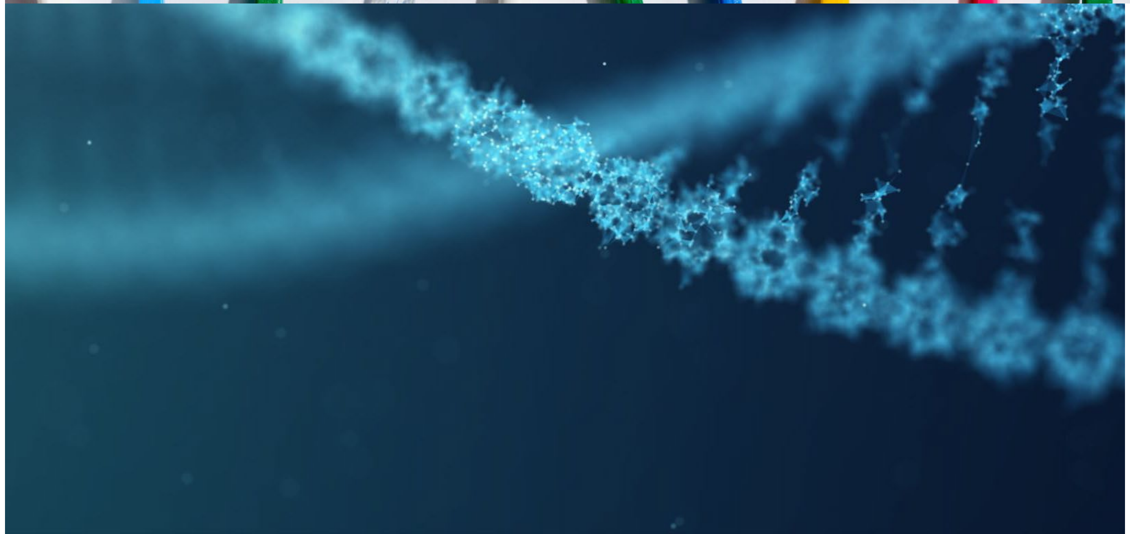
Basic science and translational research
Canvas previous related research
Hypothesize drug-target interactions
Propose new therapeutic molecules
Analyze highly dimensional research output
Propose next steps
Ongoing basic science and translational research
Clinical trials and observational research
Pretrial planning
Protocol development
Drug regimen selection
Site selection
Participant management
Cohort selection
Patient identification
Participant retention
Data management
Automate data collection
Monitor data quality
Adjudicate outcome events
Analyze large, highly dimensional, or sparse datasets
Unlock novel biological features

AI is deemed to play a significant role across the biomedical verticals


Research and Development
Clinical Practice
Population Health
Novel Therapeutic Agent Discovery
AI-aided Diagnosis
Therapy Selection
Optimized Resource Allocation
Artificial Intelligence
Precision Disease Stratification
Continuous Remote Monitoring and Diagnostics
Integration of Multi-omic Data
Extension of Physician Efficiency and Efficacy

Outlook

- Data availability is **growing** in biomedicine and healthcare
- Implementing data-driven methods that use AI-algorithms real-time variables in a hypothesis-drive/hypothesis-free approach could help us to find new targets, therapies and indications
- **Evolving** platforms including EMRs, integration engines, data mining systems and phenotyping approaches are growing
- **Integrating** novel, scalable and low-cost molecular profiling technologies with AI approaches would accelerate precision medicine development in immunology
- **Standardization** in AI, Bioinformatics and Advanced analytics would lead to develop computational medicine standards
- **Advances in AI** (including AGI) will further improve the application of AI and its impact in biomedicine and healthcare



References

- Efficient evolution of human antibodies from general protein language models
  https://pubmed.ncbi.nlm.nih.gov/37095349/

- The role of machine learning in clinical research: transforming the future of evidence generation
  https://pubmed.ncbi.nlm.nih.gov/34399832/

- Sepsis in the era of data-driven medicine: personalizing risks, diagnoses, treatments and prognoses
  https://pubmed.ncbi.nlm.nih.gov/31190075/

- Additional reading:
  - Shameer K, et. al; Machine learning in cardiovascular medicine: are we there yet? Heart. 2018 Jan 19. pii: heartjnl-2017-311198. doi: 10.1136/heartjnl-2017-311198. [Epub ahead of print] Review. PubMed PMID: 29352006.
  - Peters LA, et. al; Functional genomics predictive network model identifies regulators of inflammatory bowel disease. Nat Genet. 2017 Oct;49(10):1437-1449. doi: 10.1038/ng.3947. Epub 2017 Sep 11. PubMed PMID: 28892060; PubMed Central PMCID: PMC5660607.
  - Shameer K et. al; Translational bioinformatics in the era of real-time biomedical, health care and wellness data streams. Brief Bioinform. 2017 Jan;18(1):105-124. doi: 10.1093/bib/bbv118. Epub 2016 Feb 14. PubMed PMID: 26876889; PubMed Central PMCID: PMC5221424.

Acknowledgements

FOCIS Systems Biology Course – 2023 Faculty:
Emanuele de Rinaldis, PhD – Sanofi
Magnus Fontes, PhD – Merck
Giorgio Gaglia, PhD – Sanofi

Computational Biology Cluster
Precision Medicine & Computational Biology
Sanofi, USA

sanofi

FOCiS®
Federation of Clinical
Immunology Societies

Slides includes contributions from Travis Ahn-Horst, TJ, Dongyu Liu, Clement Chatelain, Franck Rapaport