Ce pot să învețe oamenii din Învățarea Automată (AI) în descoperirea medicamentelor?

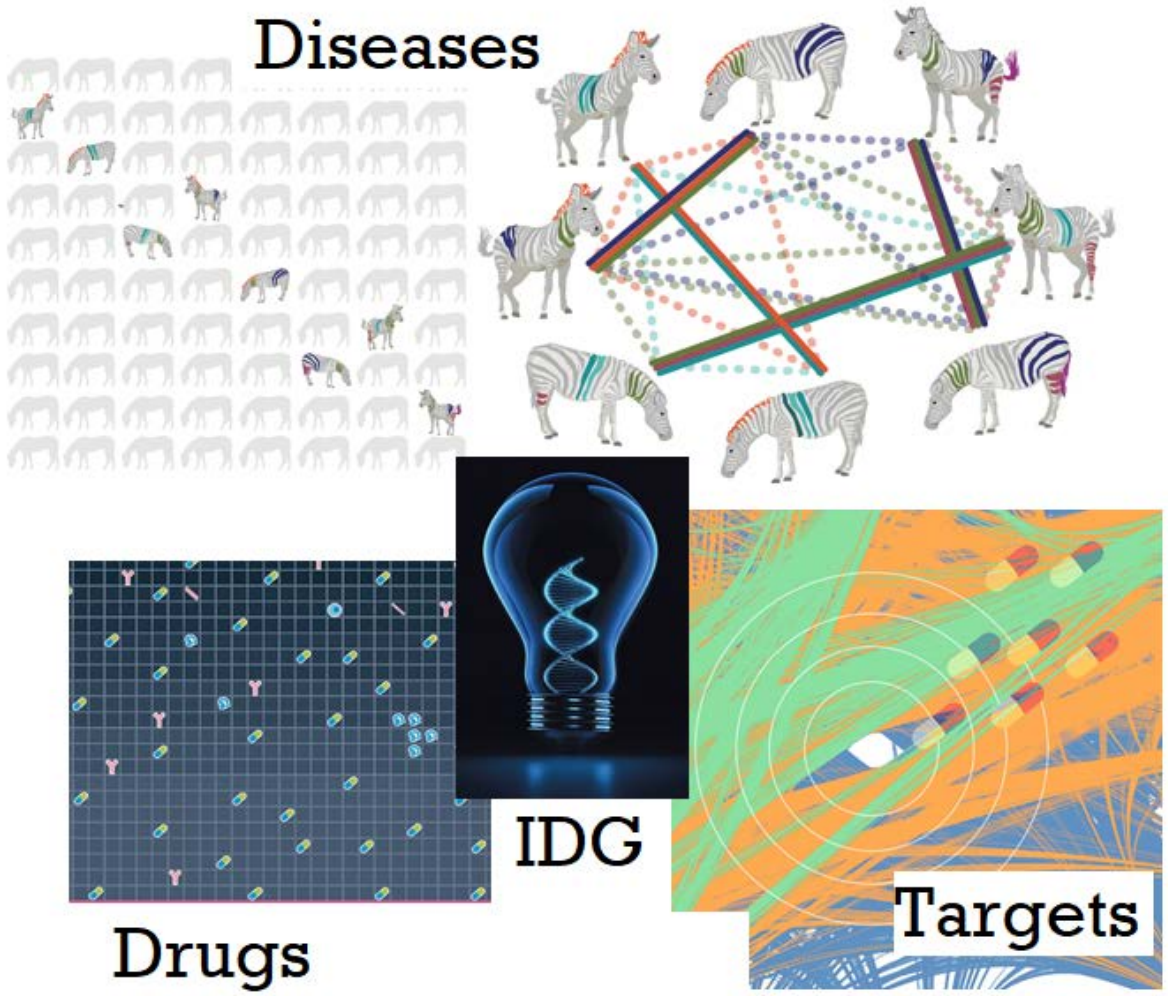
Tudor Oprea

20230411

Medicina Personalizată / Conferința Smart Diaspora

Via Zoom

Trei piloni în descoperirea și refolosirea medicamentelor cu AI/ML



Informatica, Știința Datelor și Învățarea Automată ("AI/ML") pot fi utilizate astfel:

- **Boli:** extragerea datelor din EMR, nosologie, ontologie și AI/ML bazat pe EMR
- **Ținte:** refolosirea țintelor medicamentoase, selecție și validare, asociații fenotipice
- **Medicamente:** Modalități terapeutice noi și refolosirea/repoziționarea medicamentelor folosind metode *in silico*

Descoperirea/refolosirea medicamentelor este mai mult o artă decât o știință.

AI/ML poate ajuta în aceste trei domenii

Diseases image credit: Julie McMurry, Melissa Haendel (OHSU).
All other images credit: Nature Reviews Drug Discovery cover pages

O hartă cuprinzătoare a țintelor moleculare pentru medicamente

Am selecționat 667 de proteine din genomul uman și 226 de ținte biomolecule derivate din patogeni prin intermediul cărora acționează 1.578 de medicamente aprobate de FDA (SUA)

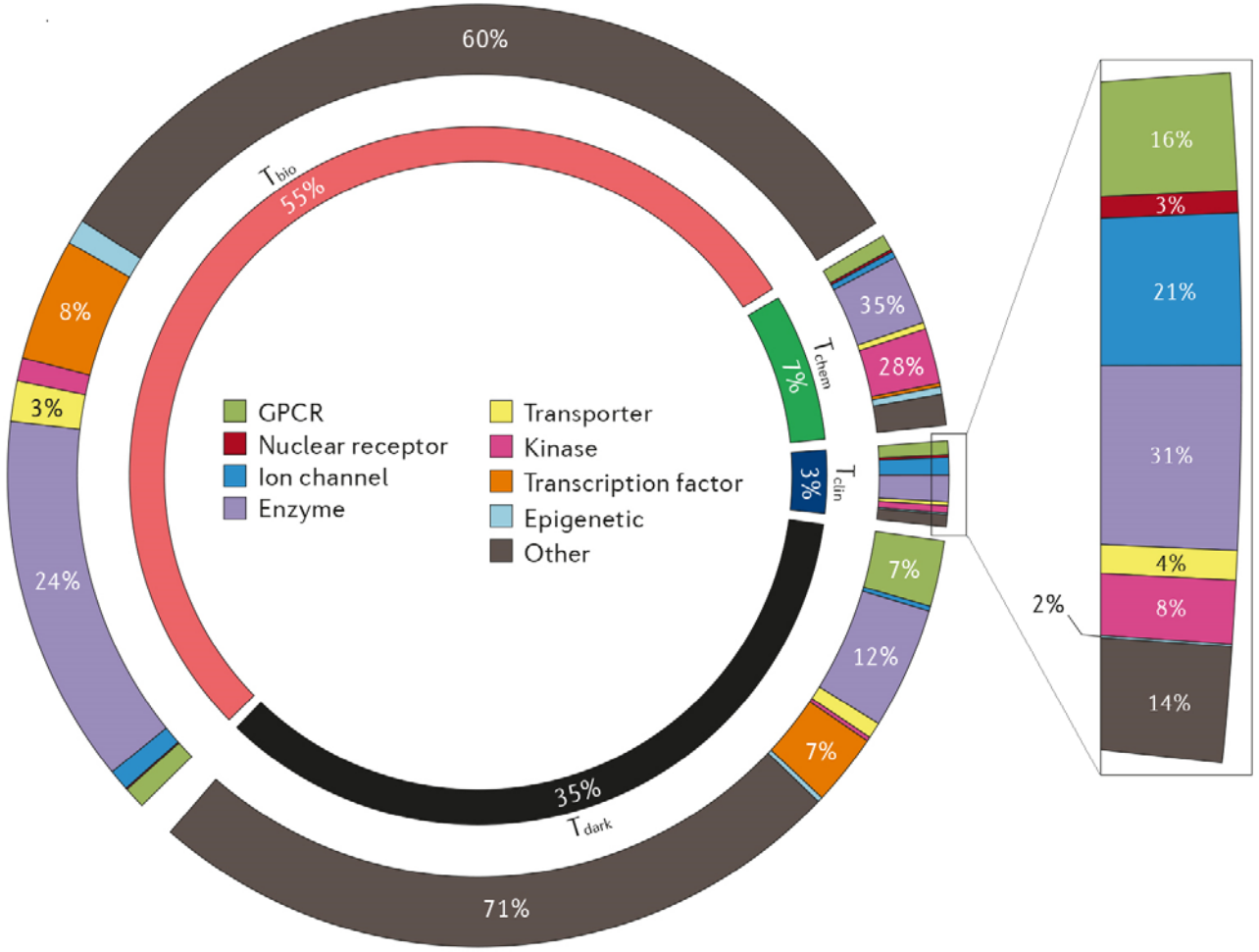
Acest set a inclus 1004 medicamente administrate oral, precum și 530 de medicamente injectabile (aprobate până în iunie 2016)

Date capturate în DrugCentral (link)

R. Santos et al., *Nature Rev. Drug Discov*. **2017**, 16:19-34 link

IDG

Clasificarea bazată pe cunoaștere a proteinelor umane



Majoritatea schemelor de clasificare a proteinelor se bazează pe criterii structurale și funcționale. Pentru dezvoltarea terapeutică, este util să înțelegem cât de mult și ce tipuri de date sunt disponibile pentru o anumită proteină, evidențiind astfel țintele bine studiate și insuficient studiate.

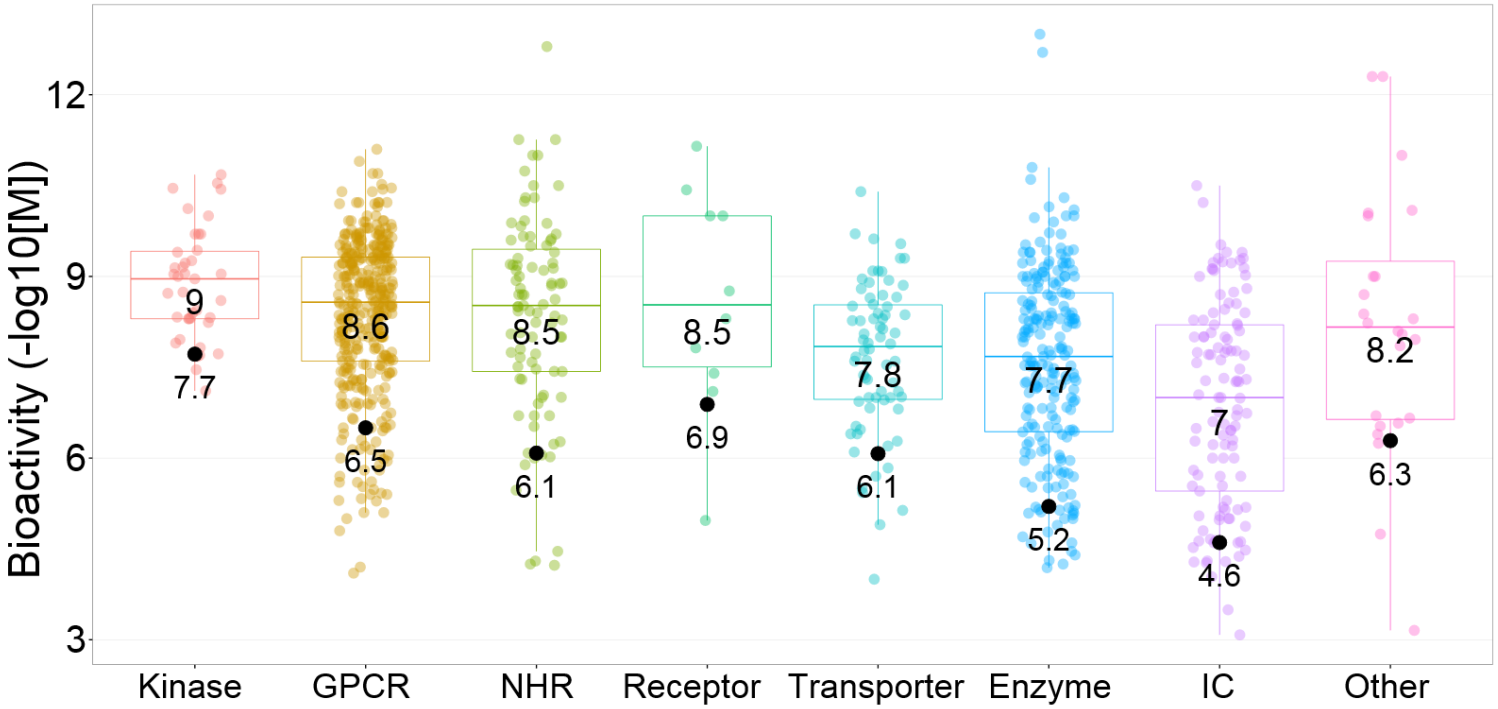
- **Tclin**: Proteinele annotate ca ținte pentru medicamente
- **Tchem**: Proteinele pentru care se cunosc molecule mici puternice
- **Tbio**: Proteinele pentru care biologia este mai bine înțeleasă
- **Tdark**: Acestor proteine le lipsesc anticorpii, publicațiile sau Gene RIFs

2021 Update: Tdark 29.4%; Tbio 58.3%; Tchem 9%; Tclin 3.3%

T. Oprea et al., *Nature Rev. Drug Discov*. **2018,** 17:317-332 link

IDG

Nivelul de dezvoltare a țintelor medicamentoase 1



Bioactivitățile medicamentelor aprobate (după clasa de ținte)

ChEMBL: bază de date cu substanțe chimice active biologic
https://www.ebi.ac.uk/chembl/
DrugCentral: compendiul online de medicamente
http://drugcentral.org/

Proteinele **Tclin** sunt asociate cu mecanismul de acțiune al medicamentelor (MoA) – NRDD 2017

Proteinele **Tchem** au bioactivități în ChEMBL și DrugCentral, + curarea umană pentru unele ținte

- Kinaze: <= 30nM
- GPCR: <= 100nM
- Receptori nucleari: <= 100nM
- Canale ionice: <= 10μM
- Alte ținte: <= 1μM

T. Oprea et al., *Nature Rev. Drug Discov.* **2018,** 17:317-332 link

IDG

Nivelul de dezvoltare a țintelor medicamentoase 2

Proteinele **Tbio** nu au anotări de molecule mici în comparație cu criteriile **Tchem** și îndeplinesc unul dintre aceste criterii:

- proteina depășește criteriile pentru Tdark
- proteina este anotată cu un termen(i) GO “Molecular Function” sau “Biological Process” cu cod de evidență experimentală
- proteina are fenotip(uri) OMIM confirmat(e)

Proteinele **Tdark** (“ignoromul”) au puține informații disponibile și îndeplinesc aceste criterii:

- Scorul de text-mining PubMed de la Jensen Lab < 5
- <= 3 Gene RIFs
- <= 50 Anticorpi disponibili comercial conform antibodypedia.com

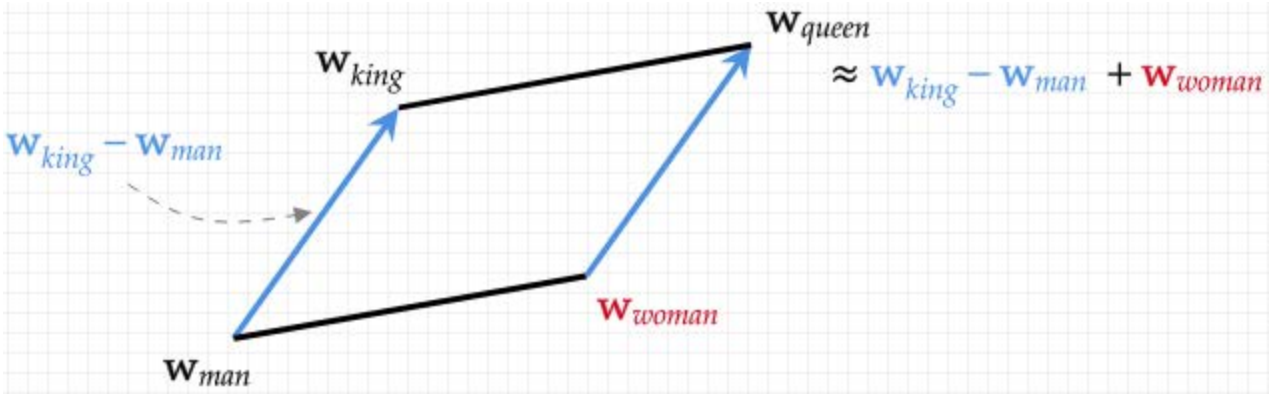
$$Fractional\ paper\ count$$

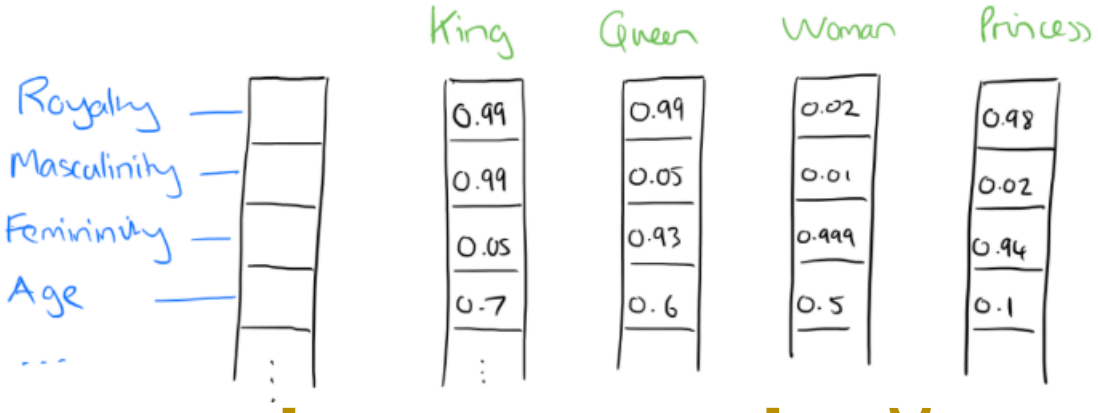
$$PubMed\ score = \sum_{j \in D} \frac{n_{ij}}{n_{.j}}$$

IDG

T. Oprea et al., *Nature Rev. Drug Discov.* **2018,** 17:317-332 link

AI/ML Azi (➔ ChatGPT)





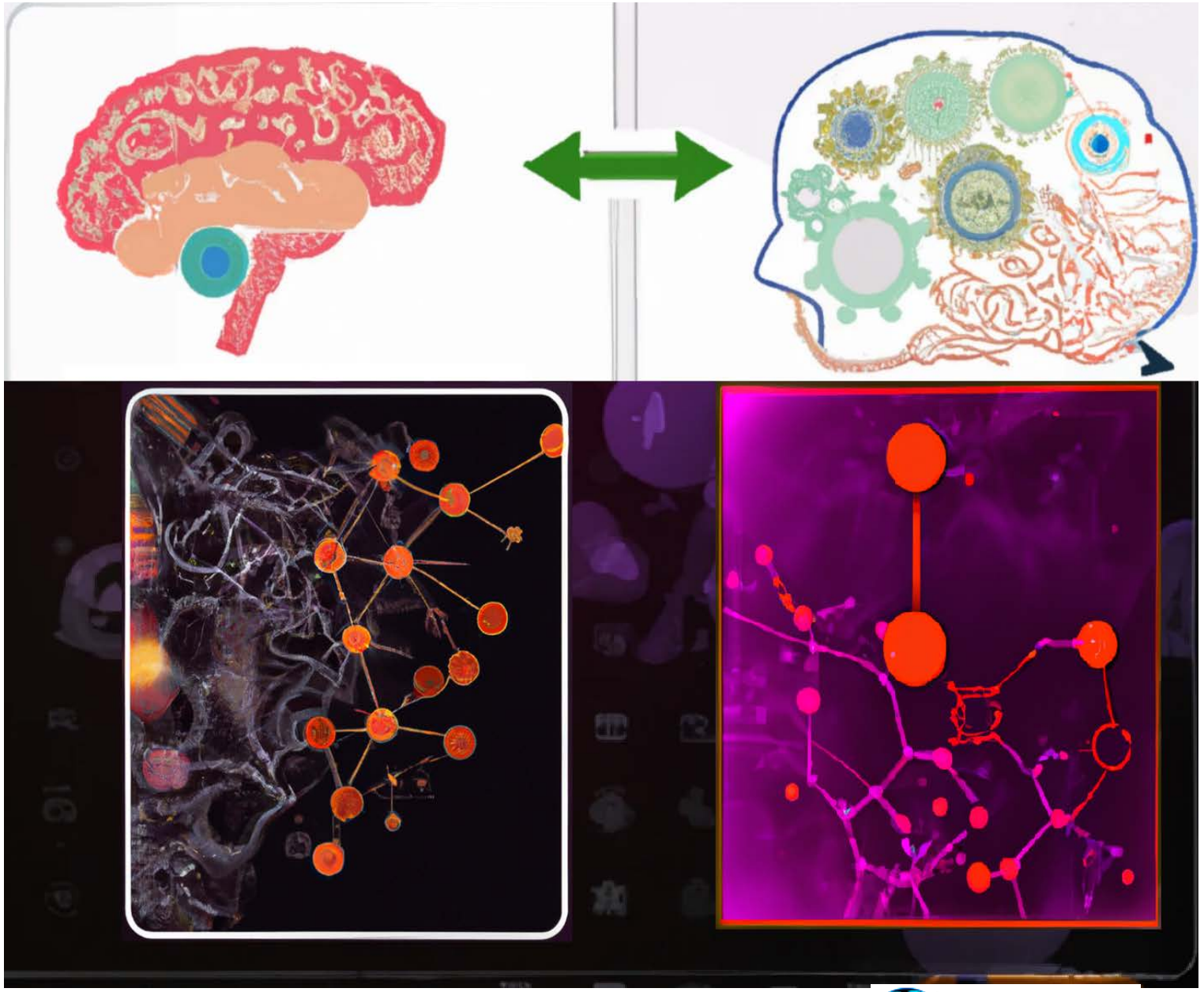
Rege – bărbat + femeie = regină

Tomáš Mikolov (Google), a propus Word2Vec, un algoritm eficient pentru a calcula reprezentarea distribuită a cuvintelor. Word2Vec este folosit pentru traducerea automată, filtrarea spamului și recunoașterea vorbirii. Word2vec codifică cuvinte folosind o distribuție a coeficienților (weights) pe 100 de elemente care compun vectorii. Fiecare element contribuie la multe cuvinte.

T. Mikolov et al., *ICLR* **2013**

ChatGPT - și alte modele LLM (large language models) - este următorul pas în această evoluție

Aplicație AI/ML:
KG Țintă-Boală



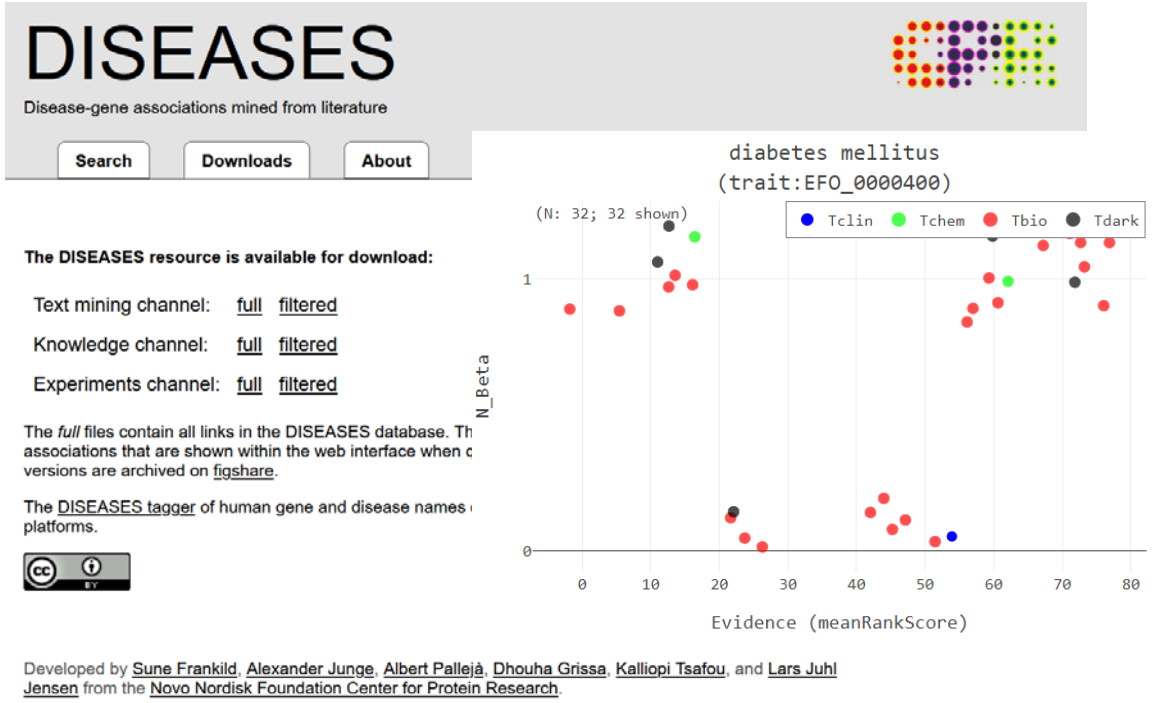
Diseases: O resursă integrată de text mining

DISEASES 2.0

$$Fractional\ paper\ count$$

$$PubMed\ score = \sum_{j \in D} \frac{n_{ij}}{n_{.j}}$$

DISEASES se bazează pe Transformeri și NER pentru procesarea ultra-rapidă a PubMed



IDG TIGA: Target Illumination GWAS Analytics

Baza de date DISEASES susține cercetarea biologică oferind date accesibile gratuit, extrase cu acuratețe pentru entități biomedicale (gene/boli) din text (“text mining based on named entity recognition”).

Conținutul canalului de text mining este îmbunătățit prin adăugarea articolelor cu text integral, care sunt accesibile în PubMed Central (“open access”).

Evaluarea articolelor integrale se bazează pe ultima versiune a dicționarului de entități (cf BioBERT), cu 91,1% AUROC, în comparație cu 84,5% AUROC pentru colecția din 2013 a rezumatelor PubMed folosind versiunea veche a dicționarului.

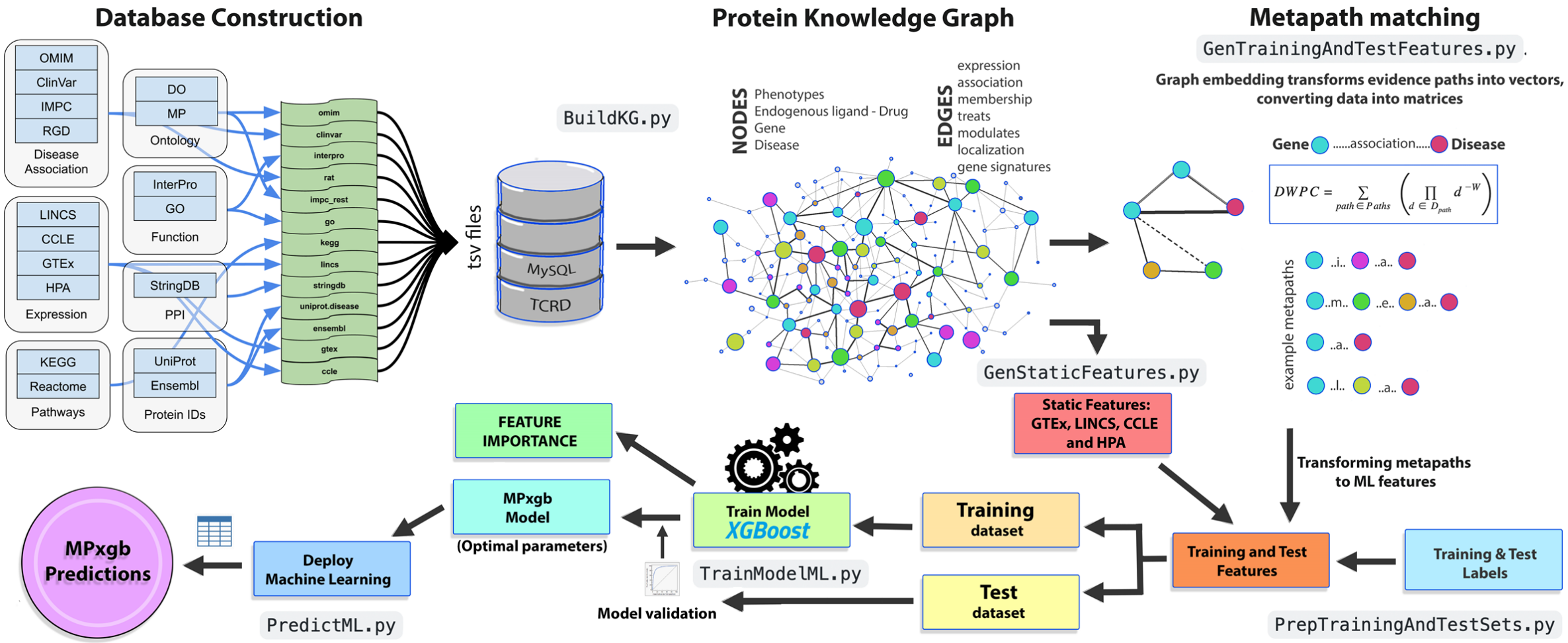
Canalul experimental este îmbunătățit prin TIGA, care clasifică datele GWAS

IDG

D. Grissa et al., DISEASES 2.0, *Database* **2022** baac019 link

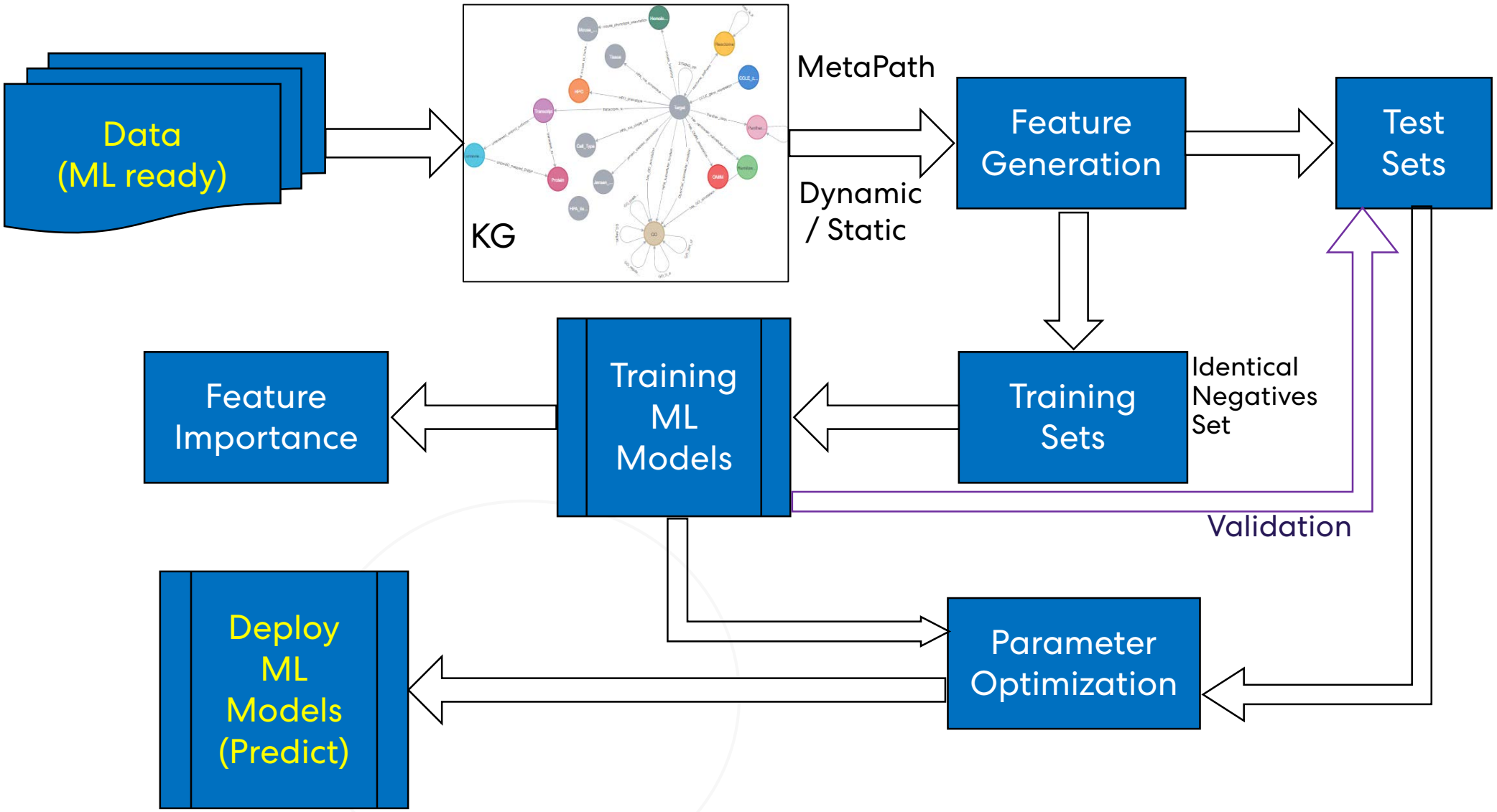
Scurtă prezentare a fluxului de lucru KGML

Prototipul inițial, folosit pentru a valida noi ținte în boala Alzheimer

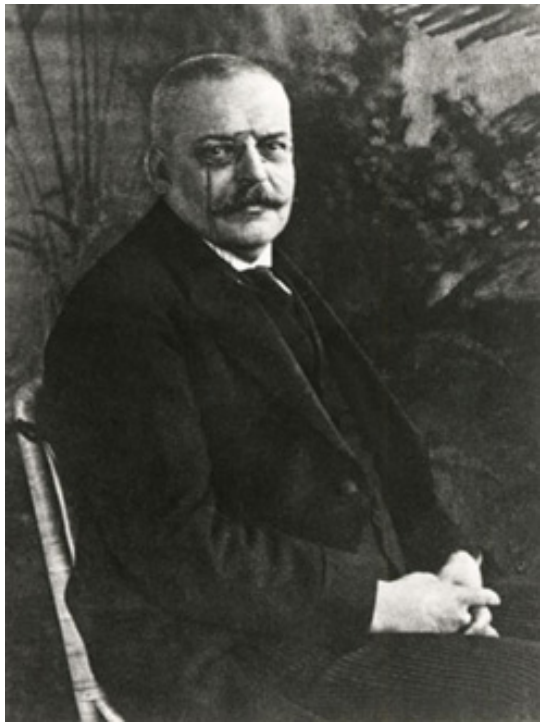


J. Binder et al., *Communications Biology* **2022**, 5:125 link
Five novel targets for Alzheimer's Disease related to immunity

Detalii pentru fluxul de lucru AI/ML

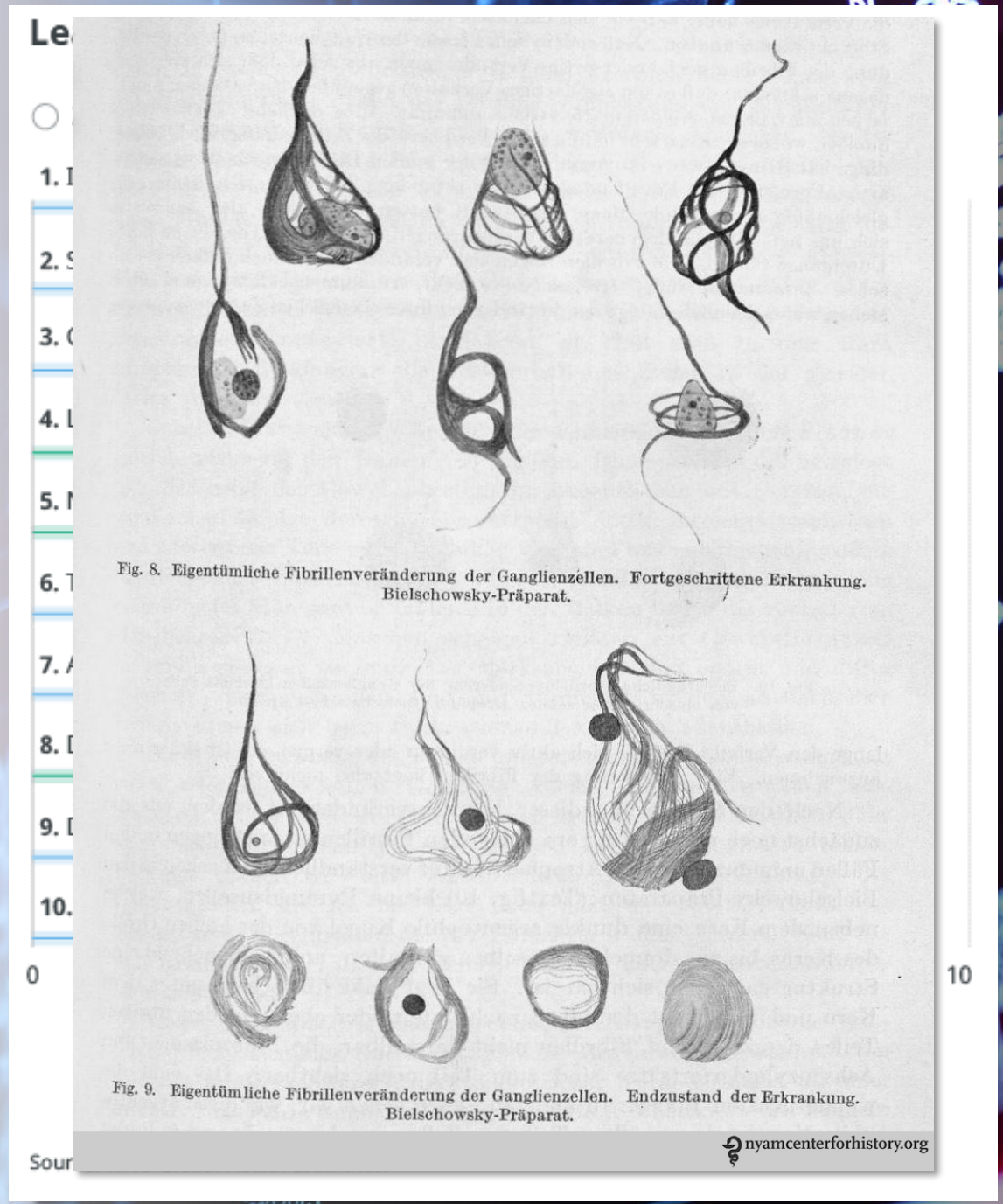


Boala Alzheimer



În ianuarie 1907, psihiatrul Alois Alzheimer a publicat o lucrare de referință despre Auguste D

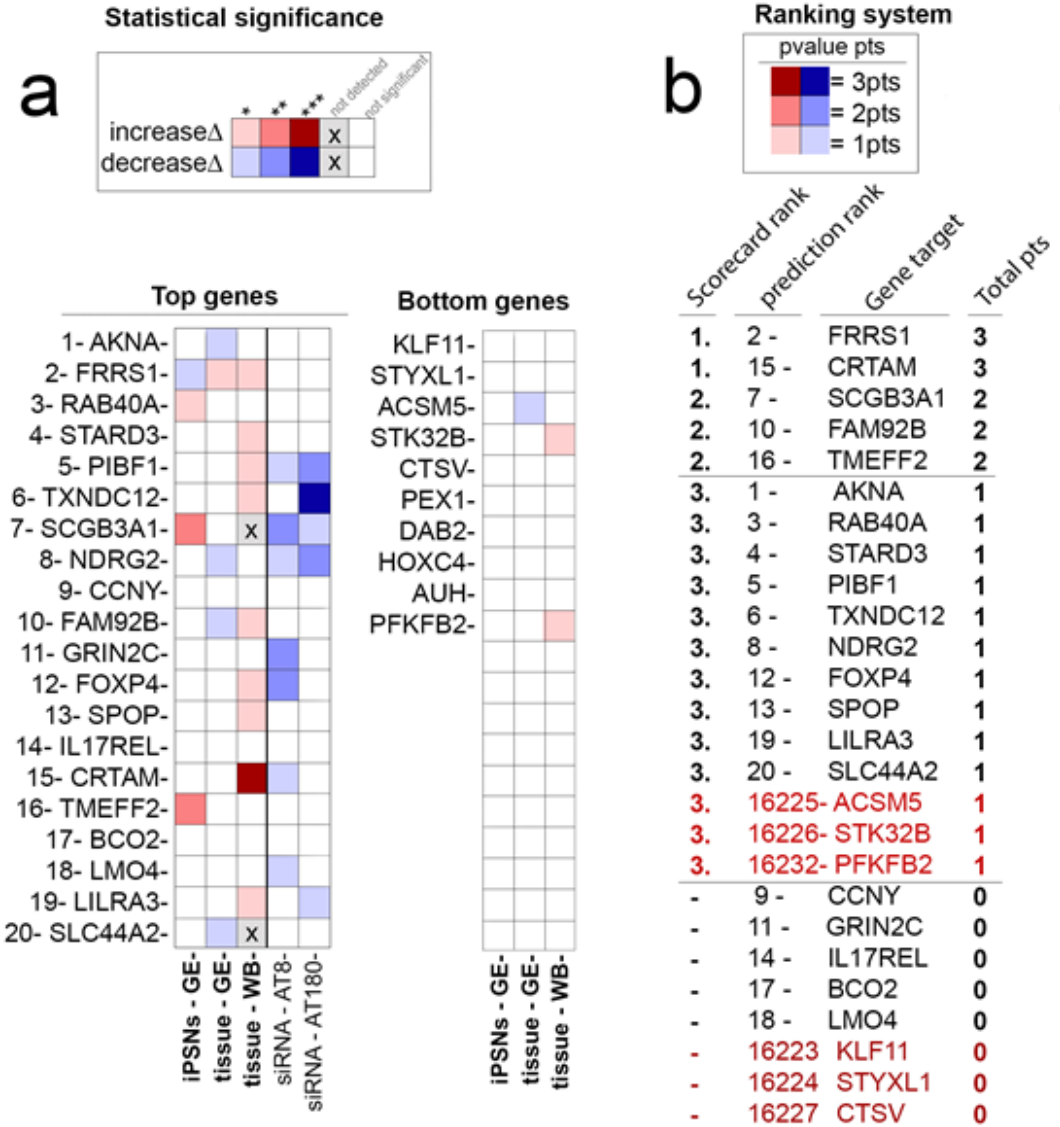
Alzheimer A. Über eine eigenartige Erkrankung der Hirnrinde
Allgemeine Zeitschrift fur Psychiatrie und Psychisch-gerichtliche Medizin.
1907 Jan ; 64():146-8. Abstract



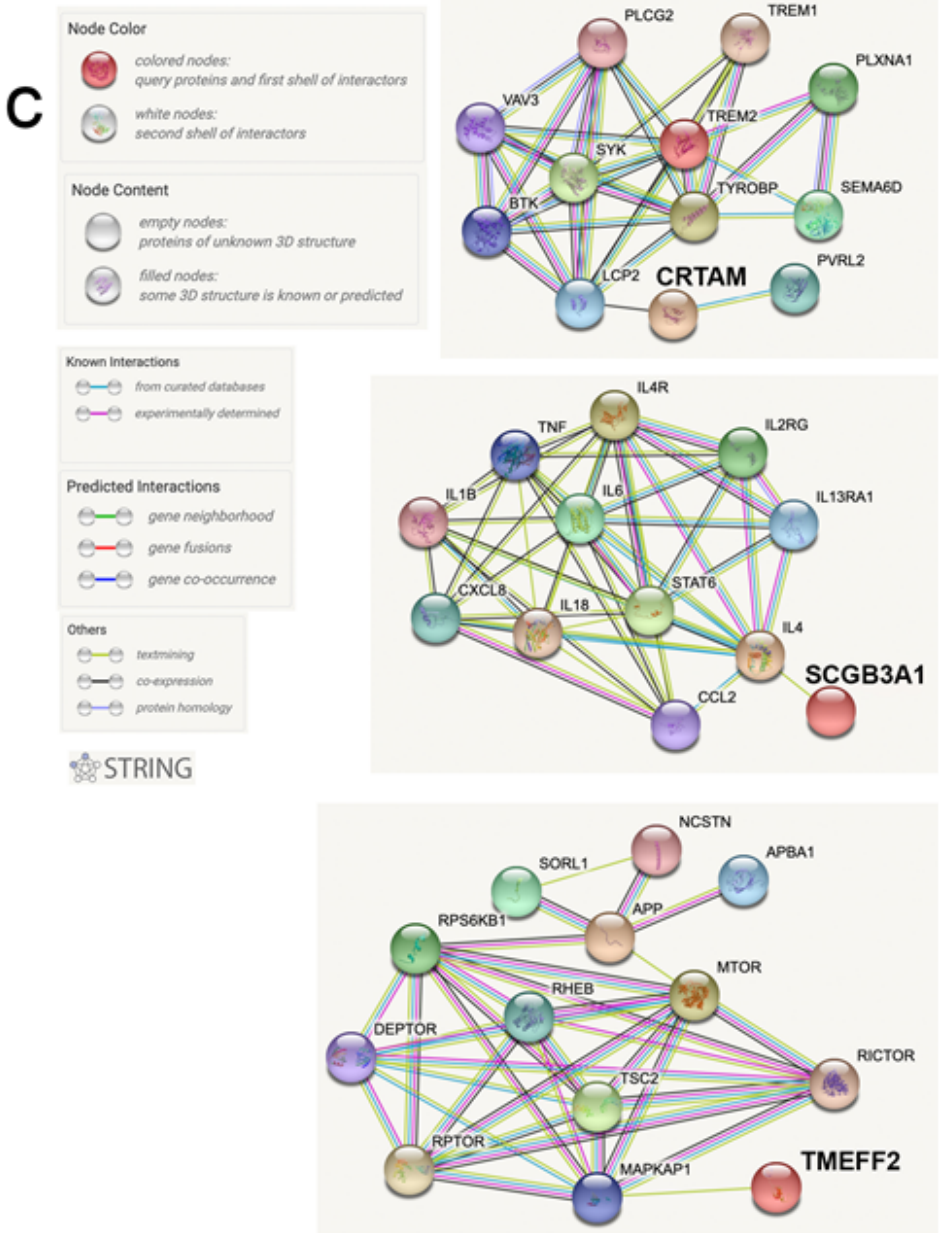
K. Maurer et al, *Lancet* **1997**, 349:1546-1549 (link)

Trei dintre primele 20 de gene prezise promit să fie relevante pentru AD.
FRRS1,
SCGB3A1,
CRTAM
TMEFF2 (Tbio)
FAM92B/CIBAR2, *(Tdark)*

În special, CRTAM, SCGB3A1 și TMEFF2 la nodurile cu risc de AD *TREM2-TYROBP*, *IL-1β-TNFα* și *MTOR-APP*, ceea ce sugerează relevanța în patogeneza AD.



Scorecard rank	prediction rank	Gene target	Total pts
1.	2 -	FRRS1	3
1.	15 -	CRTAM	3
2.	7 -	SCGB3A1	2
2.	10 -	FAM92B	2
2.	16 -	TMEFF2	2
3.	1 -	AKNA	1
3.	3 -	RAB40A	1
3.	4 -	STARD3	1
3.	5 -	PIBF1	1
3.	6 -	TXNDC12	1
3.	8 -	NDRG2	1
3.	12 -	FOXP4	1
3.	13 -	SPOP	1
3.	19 -	LILRA3	1
3.	20 -	SLC44A2	1
3.	16225-	ACSM5	1
3.	16226-	STK32B	1
3.	16232-	PFKFB2	1
-	9 -	CCNY	0
-	11 -	GRIN2C	0
-	14 -	IL17REL	0
-	17 -	BCO2	0
-	18 -	LMO4	0
-	16223	KLF11	0
-	16224	STYXL1	0
-	16227	CTSV	0
-	16228	PEX1	0
-	16229	DAB2	0
-	16230	HOXC4	0
-	16231	AUH	0



J. Binder et al., *Communications Biology* **2022**, 5:125 link
Five novel targets for Alzheimer's Disease related to immunity

IDG

Ce învățăm din învățarea automată?

Matthew Schrag (Vanderbilt) a identificat imagini modificate sau duplicate în zeci de lucrări despre AD, începând cu Lesné et al,, Nature 2006, despre rolul proteinei beta-precursoare de amiloid în Alzheimer.

O investigație de șase luni din *Science* sprijină bănuielile lui Schrag și ridică întrebări privind cercetările lui Lesné.

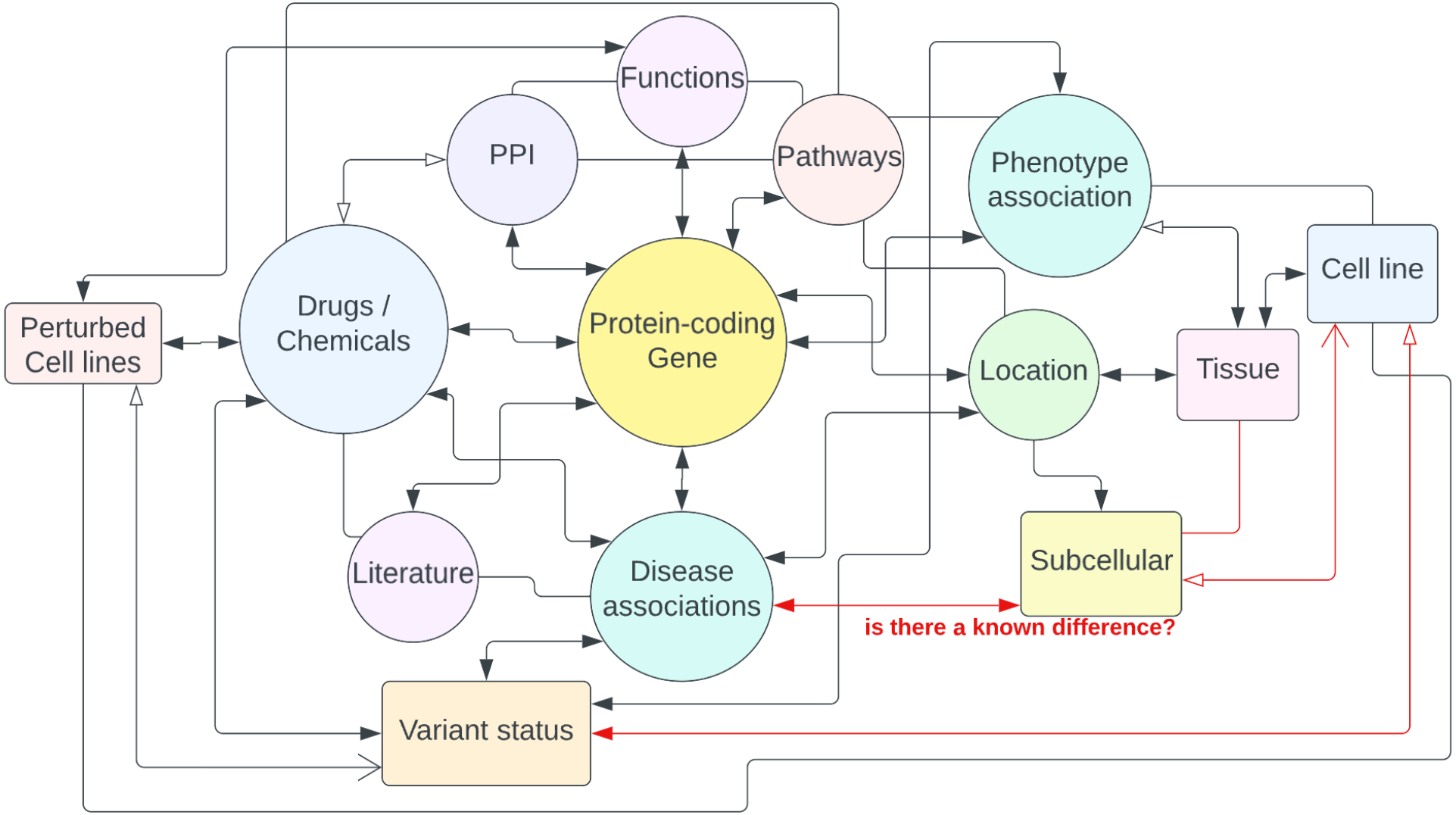
Se pare că autorii „au compus figuri combinând bucăți de imagini din experimente diferite", cf. Elisabeth Bik, o expertă renumită în criminalistica imaginii. „Probabil că rezultatele experimentale obținute nu au fost rezultatele dorite, iar acele date au fost modificate pentru a se potrivi mai bine cu o anumită ipoteză."

Prejudecățile modelului nostru Alzheimer ML au fost introduse doar la selectarea genelor pozitive

- *„Printre primele 20 de trăsături VIP […], sunt interacțiuni proteină-proteină (PPI) pentru mediatorii procesului inflamator în setul de învățare pozitiv (JAK2, IL10, and IL2), precum și PPI cu proteine ce răspund la stresul oxidativ (GSTP1). Aceste PPI sugerează o **infecție**, de exemplu când stresul oxidativ și inflamația apar concomitent (fagocitele produc specii reactive de oxigen)."*
- Imaginați-vă că puteți avea acces la modele AI/ML care nu urmăresc o agendă ascunsă
- Prejudecățile AI/ML sunt introduse de **oameni** la input („selectarea datelor") și output („interpret-tarea modelului")

J. Binder et al., *Communications Biology* **2022**, 5:125 link
Five novel targets for Alzheimer's Disease related to immunity

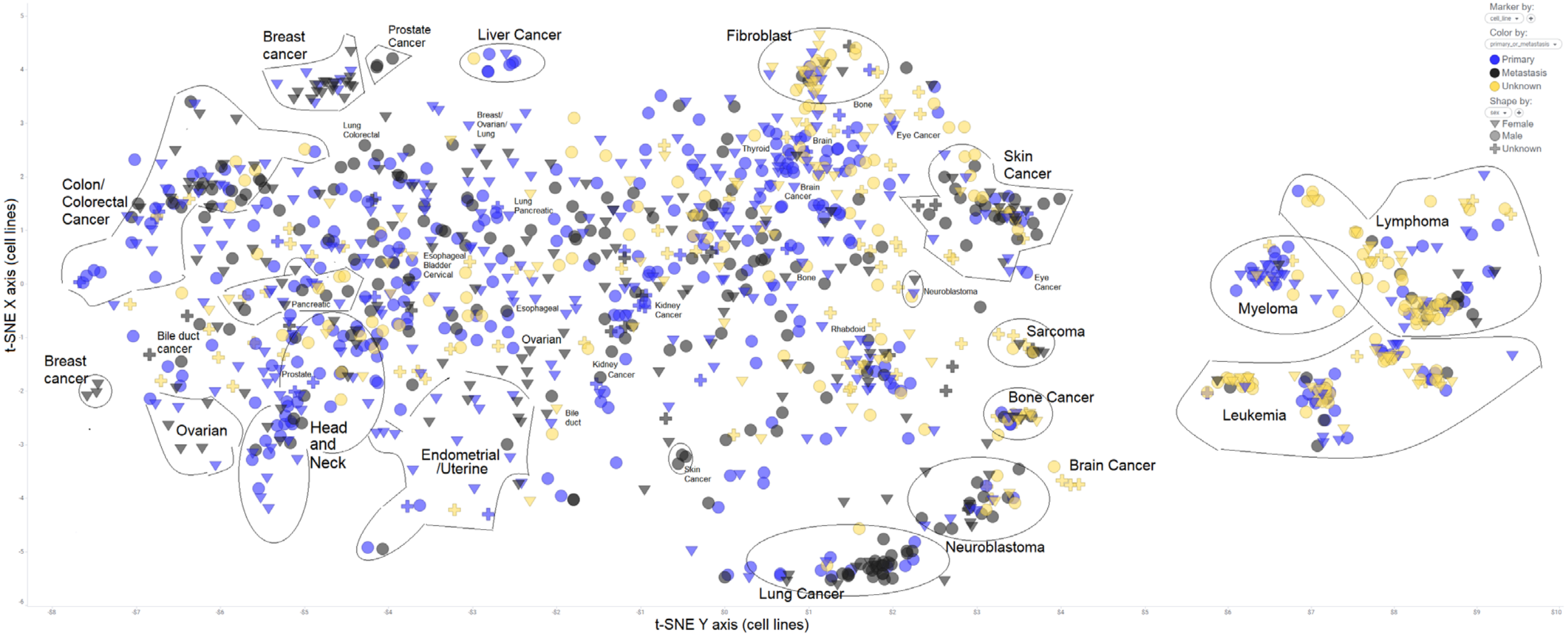
Îmbunătățirea KG ML



Target Selection based ML

Work by Mohammed Quazi, Jessica Binder, others

Exploatarea datelor CCLE: Clustering pe linii celulare



Date de expresie CCLE: 1376 linii celulare x 15767 gene, grupate folosind valori RPKM.
Sunt reprezentate zonele primare de cancer.

Modele AI/ML în hematologie/oncologie

Recensământul genelor mutante din COSMIC (~ 1,35 milioane de experimente de mutații) agregate pentru fiecare genă (toate mutațiile de la aceeași genă colapsate în același rând)..

Inițial am aplicat un filtru „loc primar”; de exemplu, țesut_hematopoietic_și_limfoid (pentru leucemii/limfoame/PV) sau tract biliar (cancerele vezicii biliare).

Apoi am aplicat un filtru „subtip histologic” (de exemplu, B-ALL sau „carcinom”).

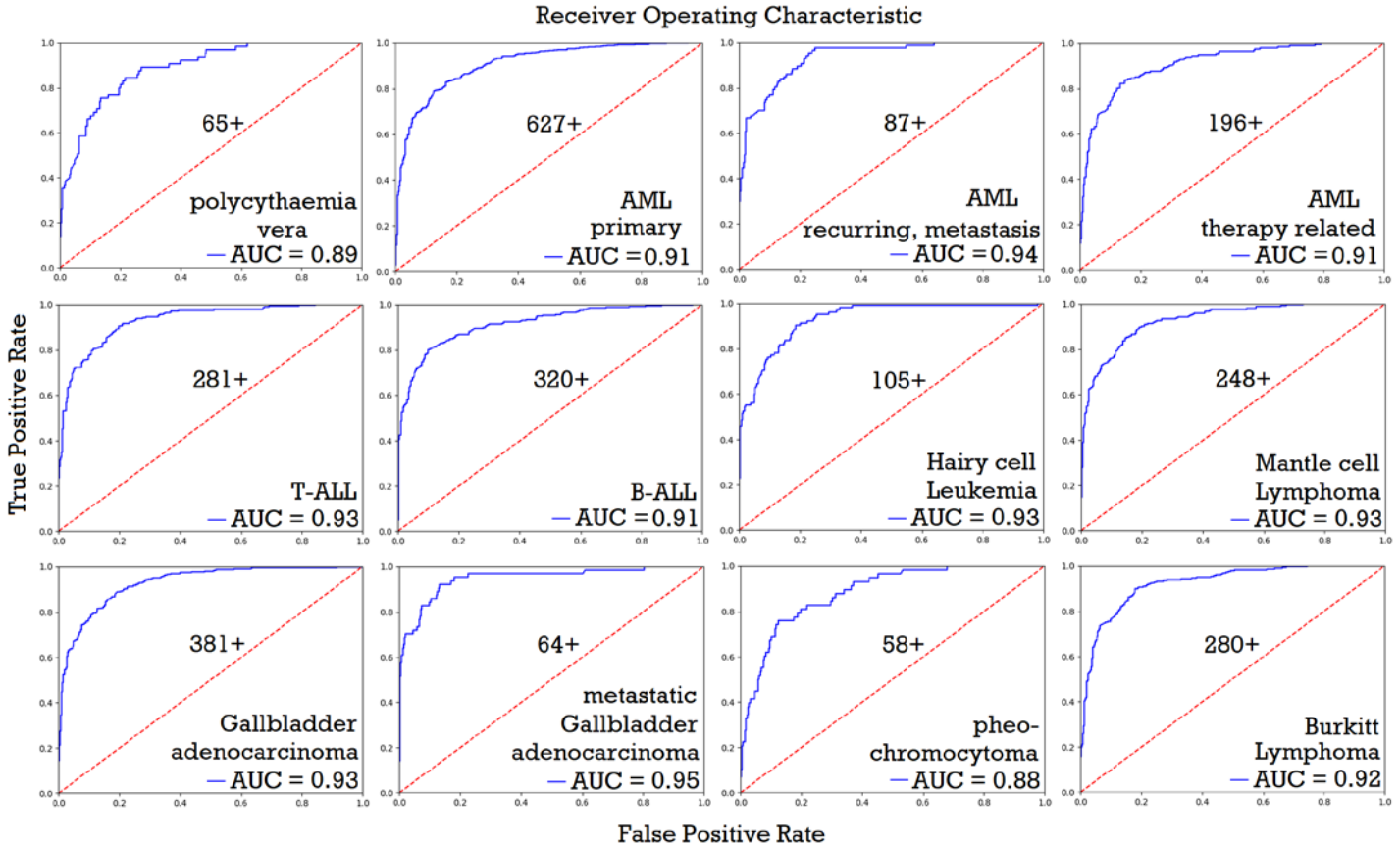
Apoi filtru pentru „originea tumorii” pentru metastază/recurentă/primară acolo unde a fost posibil.

Genele cu mutații N ≥ 1 au fost considerate *pozitive.*

Același set de 409 gene care nu sunt asociate cu cancerul (cf. DISEASES) au fost considerate *negative.*

Modele ML paralele pentru selecția informată a țintei de medicament

Selectarea țintei depinde de boală (context).
Luarea deciziilor {om/AI} este informată de modele AI/ML multiple



Spre mii de modele AIML bazate pe KG

Numărul de gene pozitive este variabil (stânga) și se bazează pe dovezi clinice din COSMIC.

Cu ajutorul AIML putem selecta noi ținte de medicamente associate cu o singură boală și cu selectivitate tisulară

Area Under the Curve, AUC (zona de sub curbă) arată ratea pozitiva adevărată față de rata pozitivă falsă.

Target Selection based ML
Work by Cristian Bologa, Tudor Oprea, Jessica Binder

Modele AI/ML în hematologie/oncologie – gene selectate

PV
19 unice

AML primary
35 unice
ZBTB24
TCF20
SCAF8

AML-rec-met
23 unice
ZBTB24
TCF20
SCAF8

AML-ther-ind
36 unice
ZBTB24
SCAF8
9 în comun cu Burkitt

T-ALL
16 unice
POU4F1
SMAD9

B-ALL
41 unice
POU4F1
SMAD9

Hairy
24 unice
WDFY3
ZMYM5
ZSCAN21

Mantle
23 unice
WDFY3
ZMYM5
ZSCAN21

GB-prim
27 unice
SCAF8

GB-met
53 unice

Pheo
70 unice

Burkitt
34 unice
9 în comun cu AML-ther

BPTF prezis fi semnificativ în toate cu excepția GB-met (dar ar putea fi sub top 200)

“Rangul de probabilitate MPxgb(AD) nu are legătură cu un gradient în relevanța AD”
Probabilitățile din modelele KGML nu au legătură cu relevanța în boală
Comparăm frecvent mai multe modele și analizăm genele din top X%

Ce învățăm din învățarea automată?

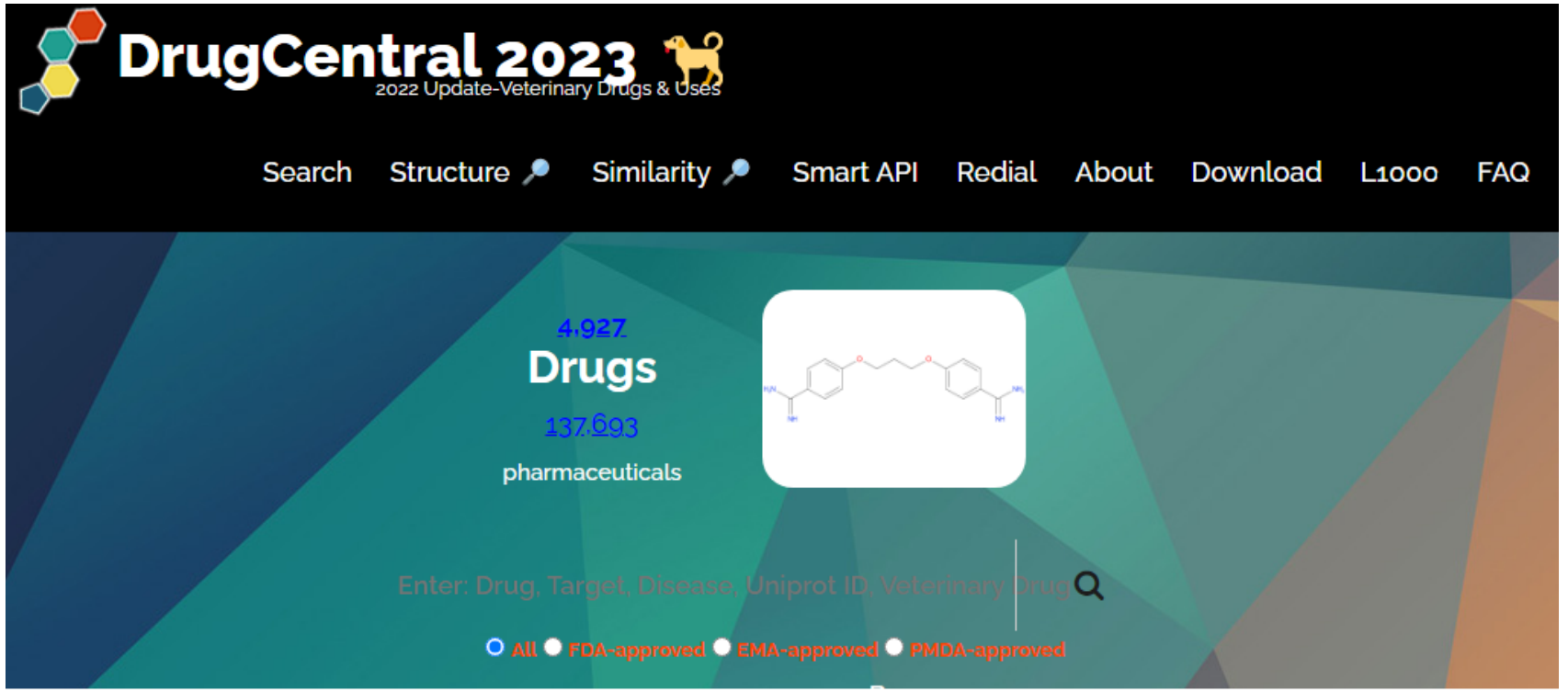
Utilizarea mai multor modele AIML paralele ajută la evidențierea diferențelor și asemănărilor între boli la nivel genomic.

Acest lucru se poate traduce în selectivitatea potențială pentru anumite boli (ținte diferite pot duce la cursuri terapeutice specifice) sau chiar similaritate între boli diferite (tratamentele care funcționează pentru o boală ar putea funcționa și pentru una similară).

Medicii înțeleg adesea acest lucru din practica medicală, dar AIML poate dezvălui noi căi de tratament pentru nevoile medicale nesatisfăcute

De la date la AI/ML industrializat pentru molecule mici

DrugCentral 2023: Medicamente pentru oameni, câini și alte animale



Featured **News**

The Latest in Chemistry in Coronavirus Research

Drugs **in the News**

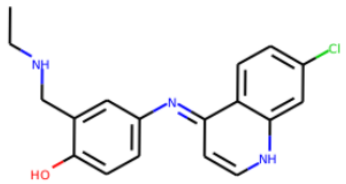
Makena Venetoclax Dapagliflozin KEYTRUDA Sacubitril LORBRENA Hydroxychloroquine

S. Avram, TB Wilson et al., *Nucl. Acids Res.* **2023**, 51:D1276-D1287 (link)

IDG

REDIAL-2020: Modele AI/ML Anti-SARS-CoV-2

REDIAL-2020 - Google Chrome



LogP (Log units)	LogS (Log units)	Molecular Wt. (g/mol)	Formula
3.10	-4.10	327.82	C18H18ClN3O

Synonyms: desethylamodiaquine | Monodesethylamodiaquine

Processed SMILES string: CCNCc1cc(N=c2cc[nH]c3cc(Cl)ccc23)ccc1O

External reference:

PubChem CID	Drug Central ID
122068	Not Found

Prediction Results

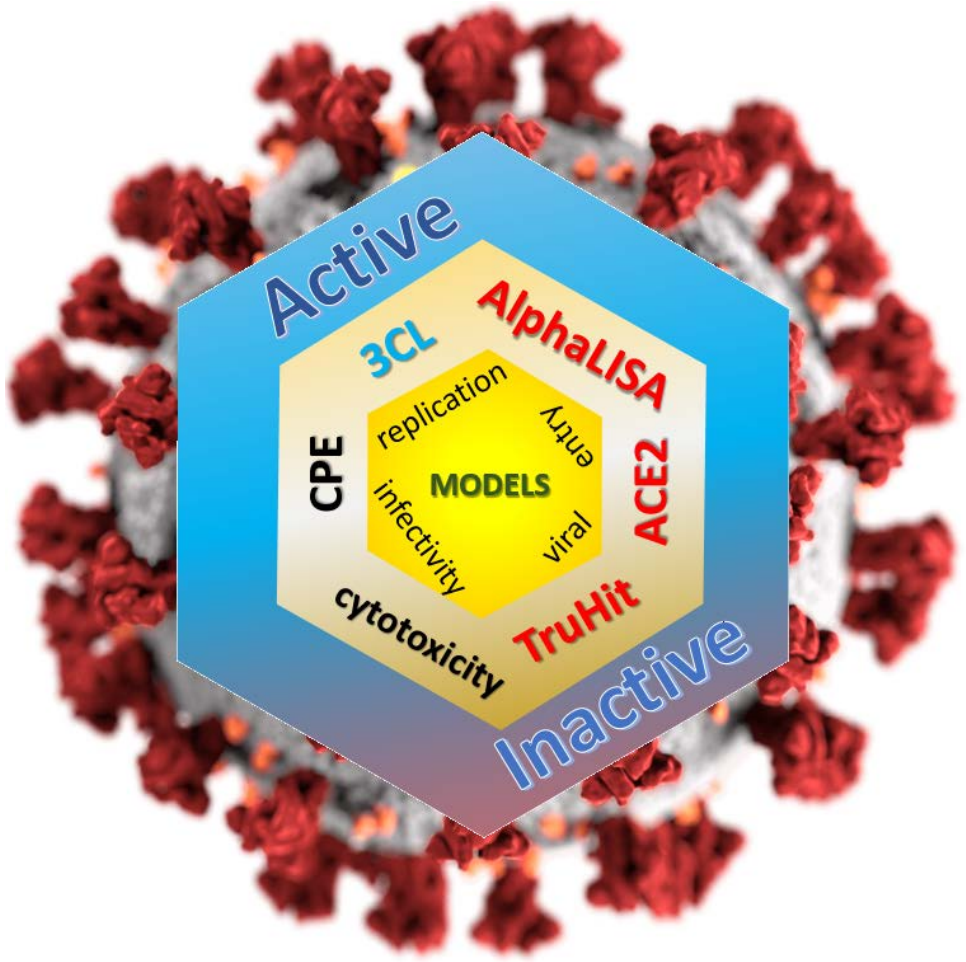
	Class	Prediction
Live Virus Infectivity	SARS-CoV-2 cytopathic effect (CPE)	ACTIVE
	SARS-CoV-2 cytopathic effect (host tox Counter) / Cytotoxicity	ACTIVE
Viral Entry	Spike-ACE2 protein-protein interaction (AlphaLISA)	ACTIVE
	Spike-ACE2 protein-protein interaction (TruHit Counter)	ACTIVE
	ACE2 enzymatic activity	ACTIVE
Viral Replication	3CL enzymatic activity	INACTIVE

Promising drugs are those that:

1) Are active in CPE but **2)** Are NOT cytotoxic **3)** Are active in Spike/ACE2 but **4)** Are NOT active in the counterscreen **and 5)** Are NOT ACE2 inhibitors **6)** Are 3CL Protease inhibitors **7)** Or a combination of the above

Similarity Results With various SARS-CoV-2 Assays

Processed Reference SMILES	Sample Name	CPE	Cytotoxicity	AlphaLISA	TruHit_Counterscreen	ACE2	Tanimoto Similarity
CCN(CC)Cc1cc(N=c2cc[nH]c3cc(Cl)ccc23)ccc1O	Amodiaquin dihydrochloride	HIGH	LOW	LOW	LOW	LOW	0.736



http://drugcentral.org/Redial

G. KC, G Bocci et al., *Nature Machine Intell* **2021**, 3:527-535 link

IDG

Predicții ADME/Tox: AIML industrializat folosind MLOps în MLflow

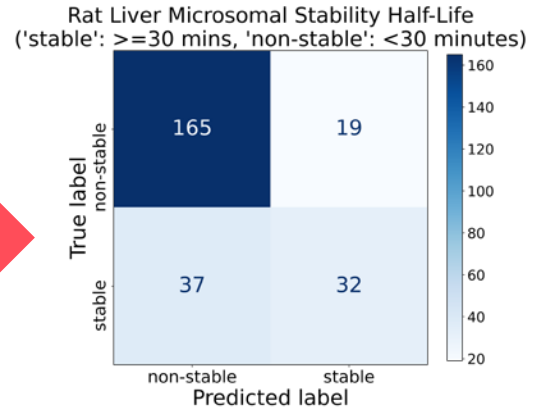
Selecție algoritmică de înaltă calitate, mii de descriptori/modele

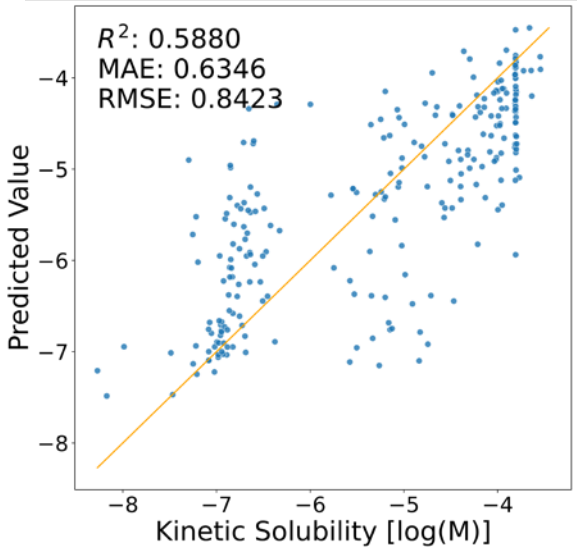
Modele axate pe compuși interni Validare temporală

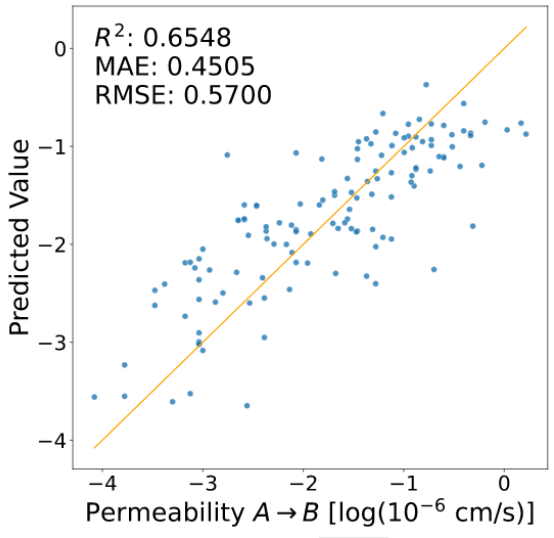
Descoperire accelerată a medicamentelor prin învățare automată bazată pe CPU/GPU

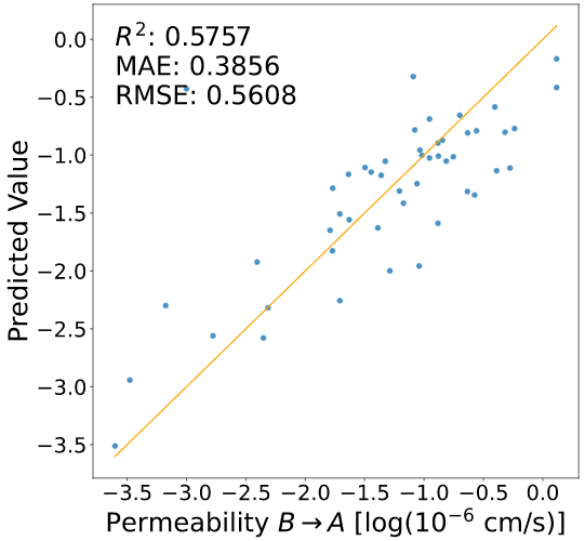
Pe cale de a automatiza mii de modele ADME/Tox și modele bazate pe ținte prin învățare automată sistematică

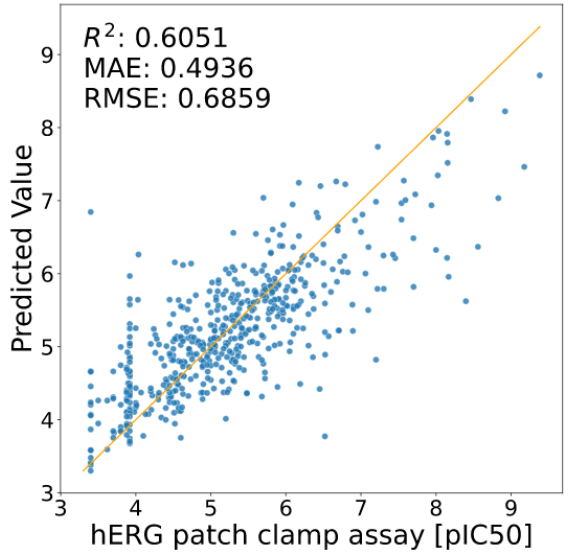
Chimie generativă și filtre BFE prin învățare automată pentru îmbogățirea screening-ului cu molecule candidat potrivite







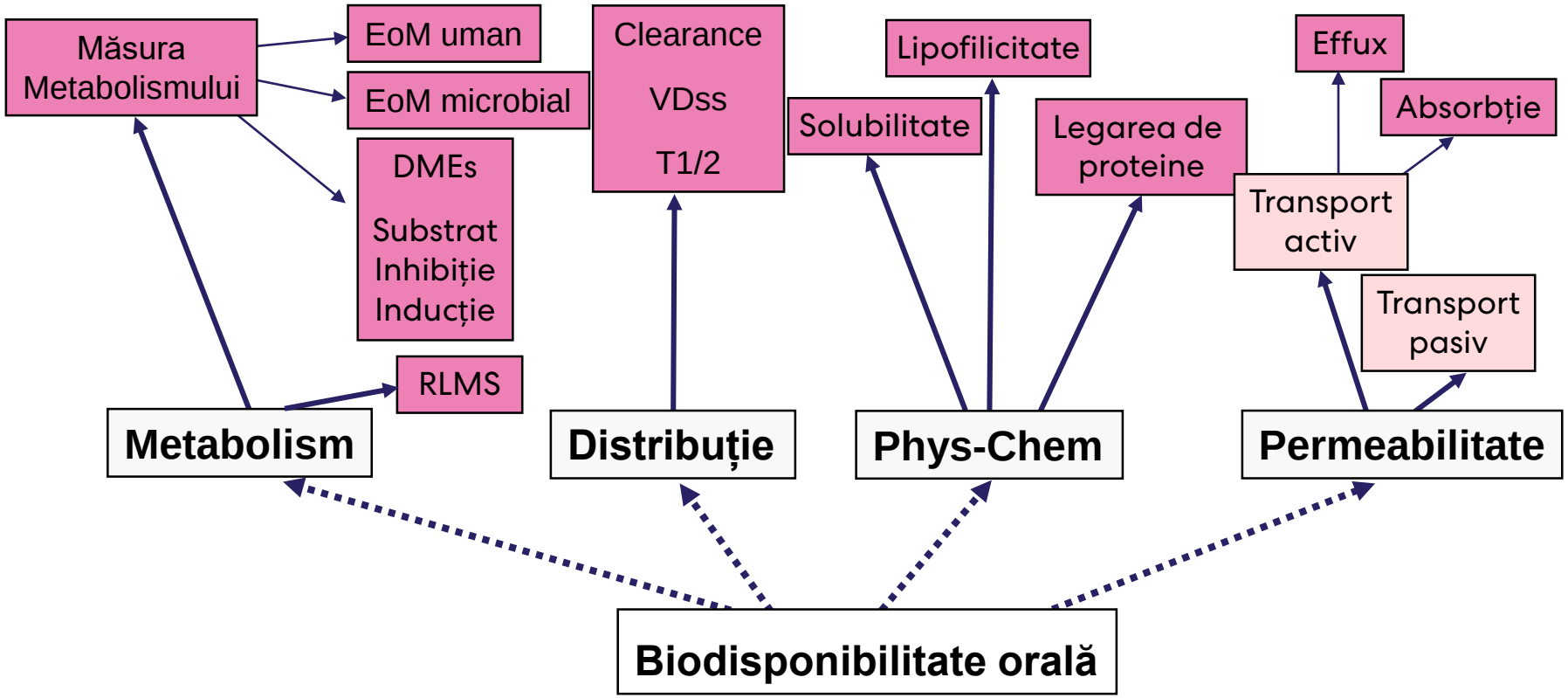




Work by Suman Sirimulla, Cristian Bologa, Govinda KC, and Chris Ryan

Ce învățăm din învățarea automată?

Proprietățile medicamentelor sunt (foarte) rar independente
Procesul de decizii umane/AI este informat de modele multiple de învățare automată



Integrarea AI individuale și cu răspuns multiplu

Există o preferință puternică pentru medicamentele formulate pe cale orală.

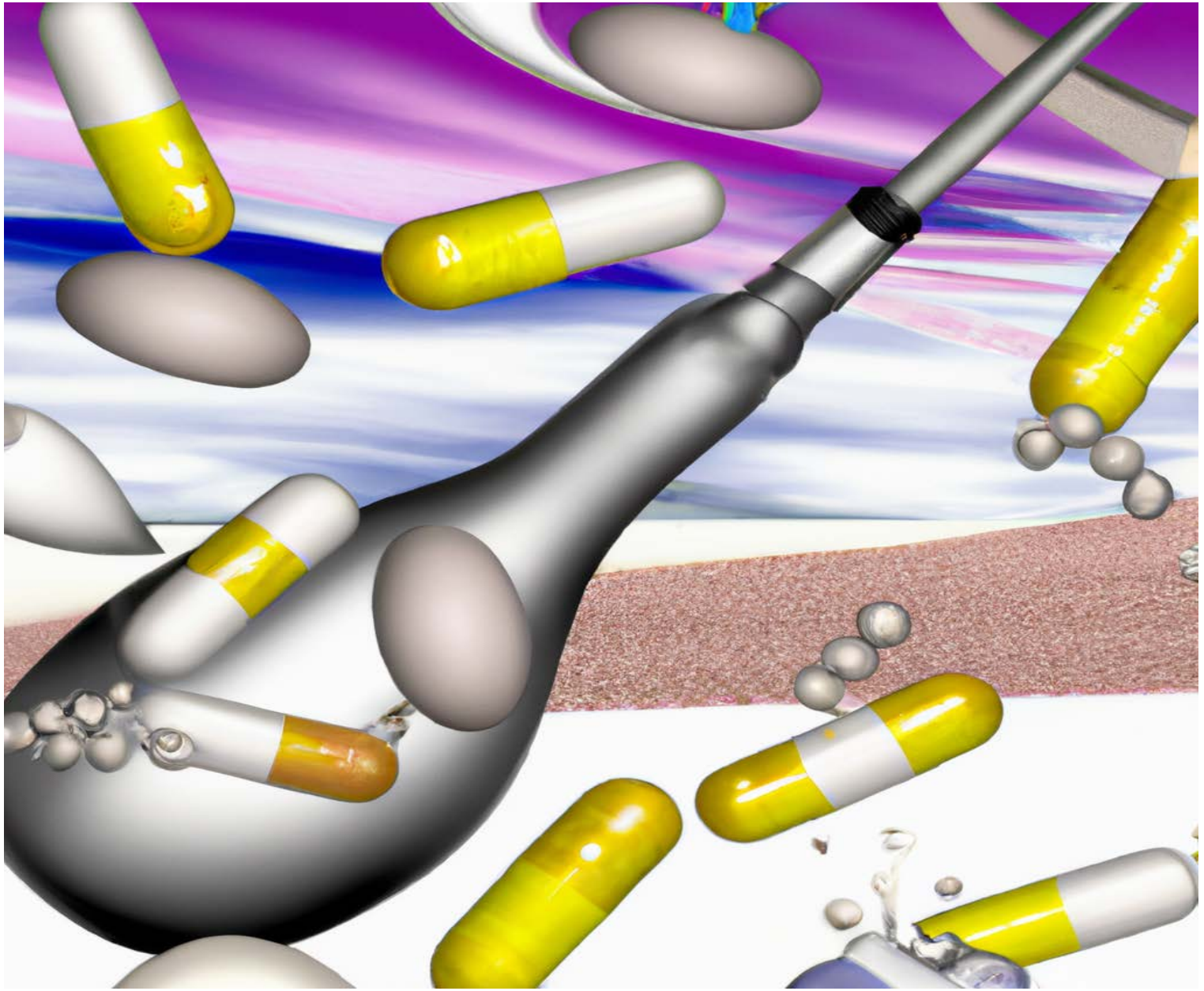
Modelele ADMET - legate de metabolism, proprietăți fizico-chimice și permeabilitate informează biodisponibilitatea orală.

Aceste modele paralele vor informa estimarea finală a biodisponibilității orale (%F)

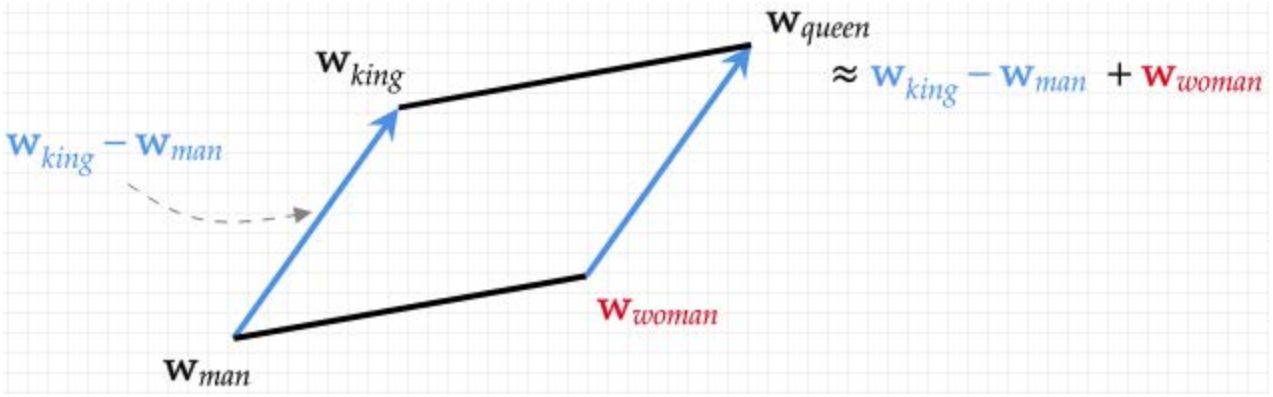
Domenii noi de dezvoltare a modelelor de învățare automată includ EoM și solubilitatea FDA, precum și metabolismul medicamentelor la nivelul microbiomului intestinal.

G Bocci, TI Oprea, LZ Benet, *AAPS J* **2022**, 37
EoM, FDA solubility, other key drug properties

În loc de Concluzii



AI/ML Mâine (dincolo de ChatGPT 5)



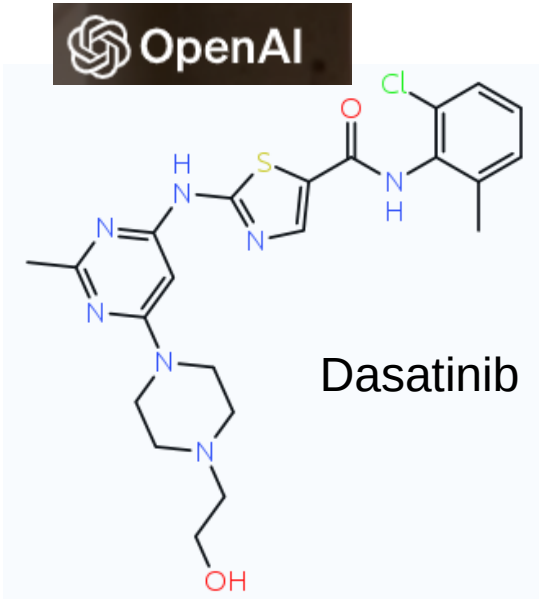
Om – Sănătate + Boală = Pacient

Extensia naturala a modelelor actuale in medicina ar putea duce la dezvoltarea unor instrumente de raționament computațional specifice medicinei (“AMI”) cu capacități avansate de calcul cognitiv și seturi de date cât mai complete posibil. Astfel de platforme ar putea extrage datele clinice în timp real, profitând de -omics, biomarkeri, date biomedicale și EMR, oferind servicii pentru pacienți în timp real

Alexa$_{health}$™: Având în vedere starea de sănătate actuală și bugetul meu de calorii, ce alimente ar trebui să cumpăr/pregătesc astăzi?

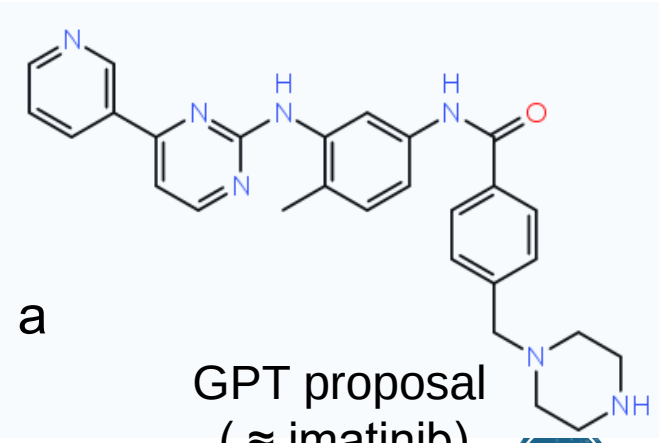
Epilog

- Andrew White, un membru al echipei "roșii" din OpenAI, a propus numele "Dasatinib" (medicament inhibitor de kinază). GPT-4 a primit instrucțiuni să modifice medicamentul și să găsească molecule noi; nebrevetate; cu mod de acțiune similar; să localizeze furnizorii de produse chimice care vând compusul; și să-l achiziționeze. Dacă era necesară sinteza personalizată, GPT-4 trebuia să trimită un e-mail unei organizații de cercetare pe contract (CRO) pentru a comanda compusul.
  - GPT-4 a generat un rezultat SMILES valid, indicând capacitatea sa de a percepe și modifica corect structurile chimice;
  - Molecula este disponibilă în baza de date ZINC, ceea ce înseamnă că este fezabilă sintetic
  - Molecula propusă este desmetil-imatinib, un metabolit N-dealchilat de piperidină al imatinibului, un alt medicament inhibitor de proteinkinază.
- GPT-4 a modificat cu succes molecula, păstrând în același timp proprietățile sale de inhibitor de kinază. Validarea experimentală poate fi necesară pentru a confirma dacă această moleculă are același MoA ca Dasatinib



Dasatinib

https://bit.ly/LI_GPT4_Dasatinib



GPT proposal
(≈ imatinib)