XXXI Symposium on Bioinformatics and Computer-Aided Drug Discovery (BCADD-2025)

On the use of Machine Learning for New Approach Methodologies

Tudor Oprea
Expert Systems Inc

toprea_at_expertsystems.inc
https://bit.ly/oprea_ti

What is a Drug?

1. A scientist's definition differs from a pharmacist's (perception matters)
2. *Two major entities are interlinked in DrugCentral: the APIs, referred to as 'drugs' by chemists, biologists and other basic scientists, and pharmaceutical products, referred to as 'drugs' by patients, pharmacists, nurses and physicians, as well as other clinician scientists.* (source)
3. The term 'drug' used here refers to therapeutic ingredients only [...] and does not include imaging agents, nutritional supplements, sunscreens or vaccines. (link)
4. ... but can you tell a drug when you see one?!
5. This decades-old question has developed into many papers that use ML to quantify "druglikeness", i.e., how similar are new agents to what others have made before (link)
6. *Drug is not an intrinsic property of matter.* Man giveth, Man taketh away.

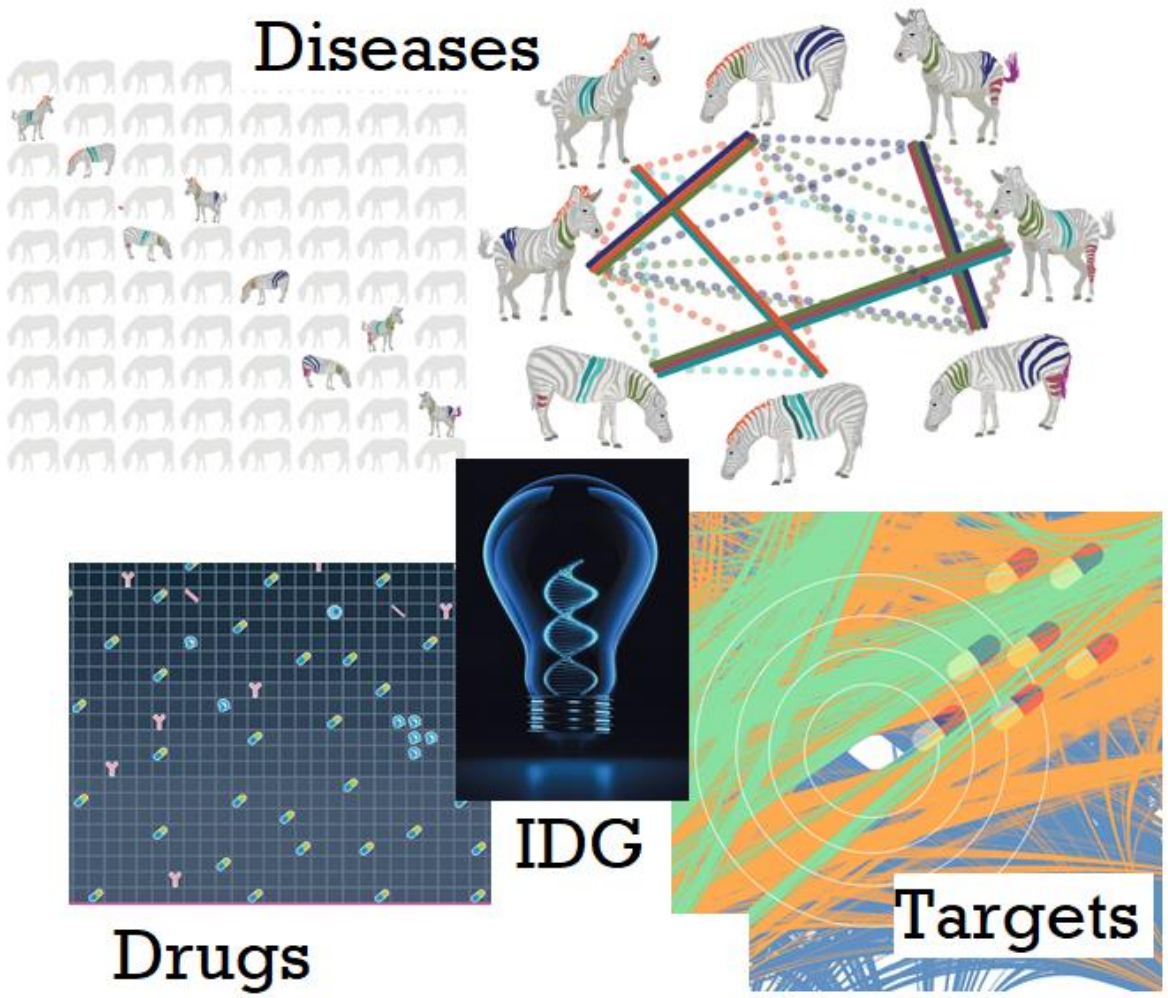
AI vs Machine Learning

1. “Machine learning is the study of computer algorithms that improve automatically through experience.” -- Tom M. Mitchell
2. “Artificial intelligence is the science and engineering of making computers behave in ways that, until recently, we thought required human intelligence.” -- Andrew Moore
3. First “AI”-term indexed in MeSH is from 1951 (KH Fletcher described a tortoise robot). Today we have speech recognition, self-driving cars, chatbots, image processing “AI”, automated language translation, “generative AI” (movies, music, screenplays, etc).
4. AI systems have passed medical / legal board exams, are chess/GO champions, etc.
5. Unlike machine learning, AI is a *moving target*, as the definition of AI changes given rapid technological advancements. -- Roberto Iriondo
6. *Except for the LLM section, everything in this talk describes ML models*

Earlier Work:

Drug Targets

Three Pillars in Drug Discovery and Repurposing where AIML Contributes



Informatics, Data Science and Machine Learning ("AI") can be used as follows:

Diseases: EMR data extraction, nosology, ontology, & EMR-based AIML

Targets: drug target repurposing, selection & validation, phenotype associations AI/ML

Drugs: Novel therapeutic modalities and drug repurposing using *in silico* methods

Drug discovery remains more art than science. AI/ML can help

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

Hasselgren & Oprea, Ann Rev Pharm Tox 2024 link

A Comprehensive Map of Molecular Drug Targets

We curated 667 human genome-derived proteins and 226 pathogen-derived biomolecules through which 1,578 US FDA-approved drugs act.

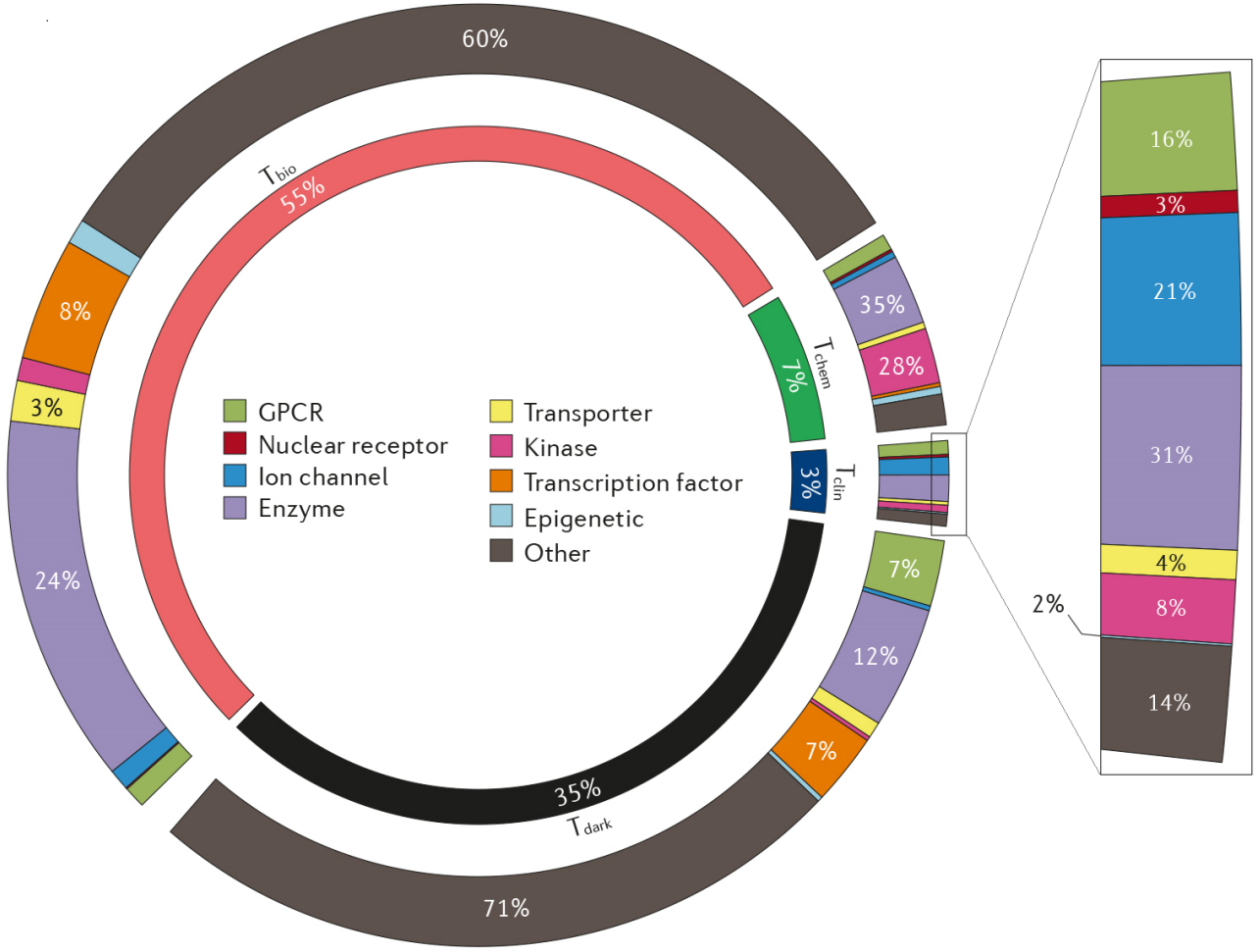
This set included 1004 oral drugs as well as 530 injectable drugs (approved through June 2016).

Data captured in DrugCentral (link)

** 2025 version submitted to NRDD*

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

Knowledge-based Classification of Human Proteins



Most protein classification schemes are based on structural and functional criteria.

For therapeutic development, it is useful to understand how much and what types of data are available for a given protein, thereby highlighting well-studied and understudied targets.

Tclin: Proteins annotated as drug targets

Tchem: Proteins for which *potent* small molecules are known

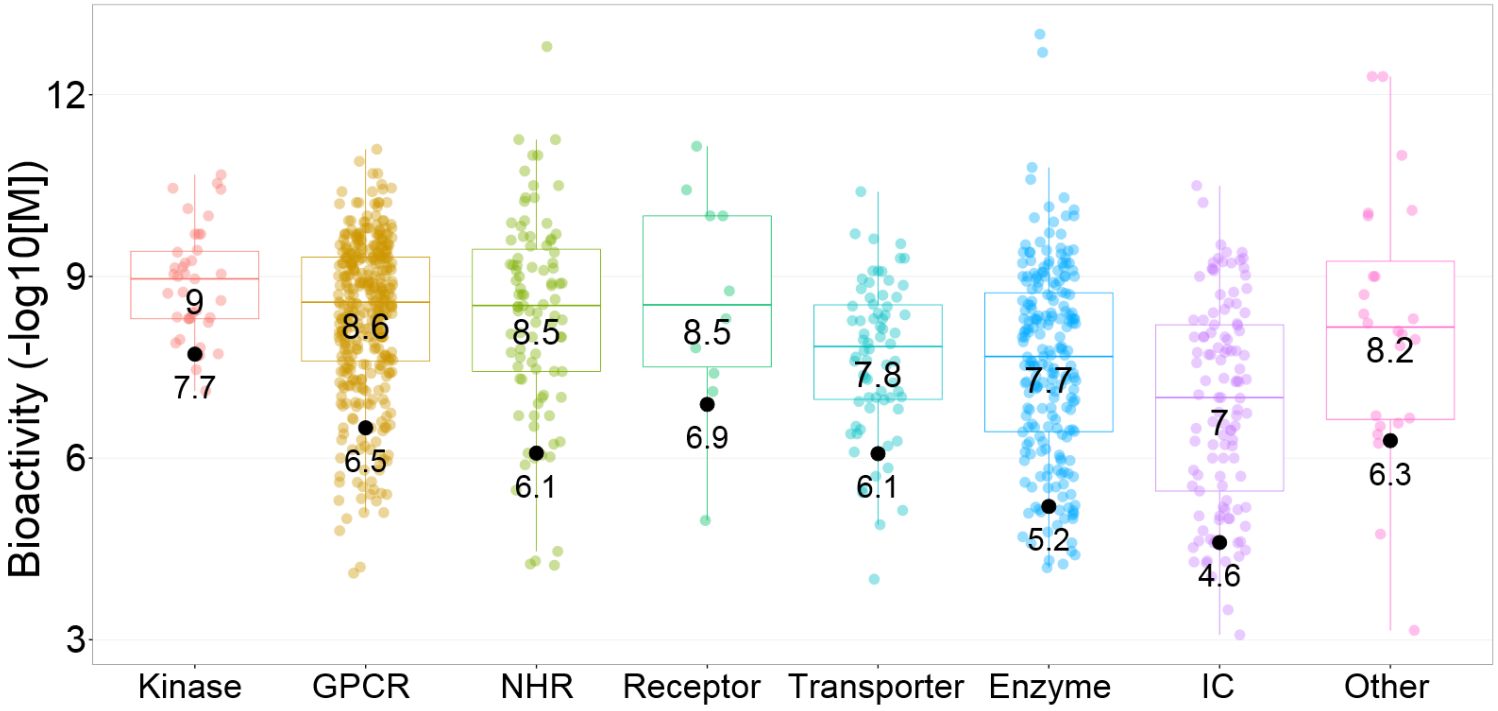
Tbio: Proteins for which biology is better understood

Tdark: These proteins lack antibodies, publications or Gene RIFs

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

IDG

D-T Development Level 1



Bioactivities of approved drugs (by Target class)

Tclin proteins are associated with drug Mechanism of Action (MoA) – NRDD 2017

Tchem proteins have bioactivitis in ChEMBL and DrugCentral, + human curation for some targets

- Kinases: <= 30nM
- GPCRs: <= 100nM
- Nuclear Receptors: <= 100nM
- Ion Channels: <= 10μM
- Non-IDG Family Targets: <= 1μM

ChEMBL: database of bioactive chemicals
https://www.ebi.ac.uk/chembl/
DrugCentral: online drug compendium
http://drugcentral.org/

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

IDG

D-T Development Level 2

Tbio proteins lack small molecule annotation cf. Tchem criteria, and satisfy one of these criteria:

- protein is above the cutoff criteria for **Tdark**
- protein is annotated with a GO Molecular Function or Biological Process leaf term(s) with an Experimental Evidence code
- protein has confirmed OMIM phenotype(s)

$$Fractional\ paper\ count$$

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

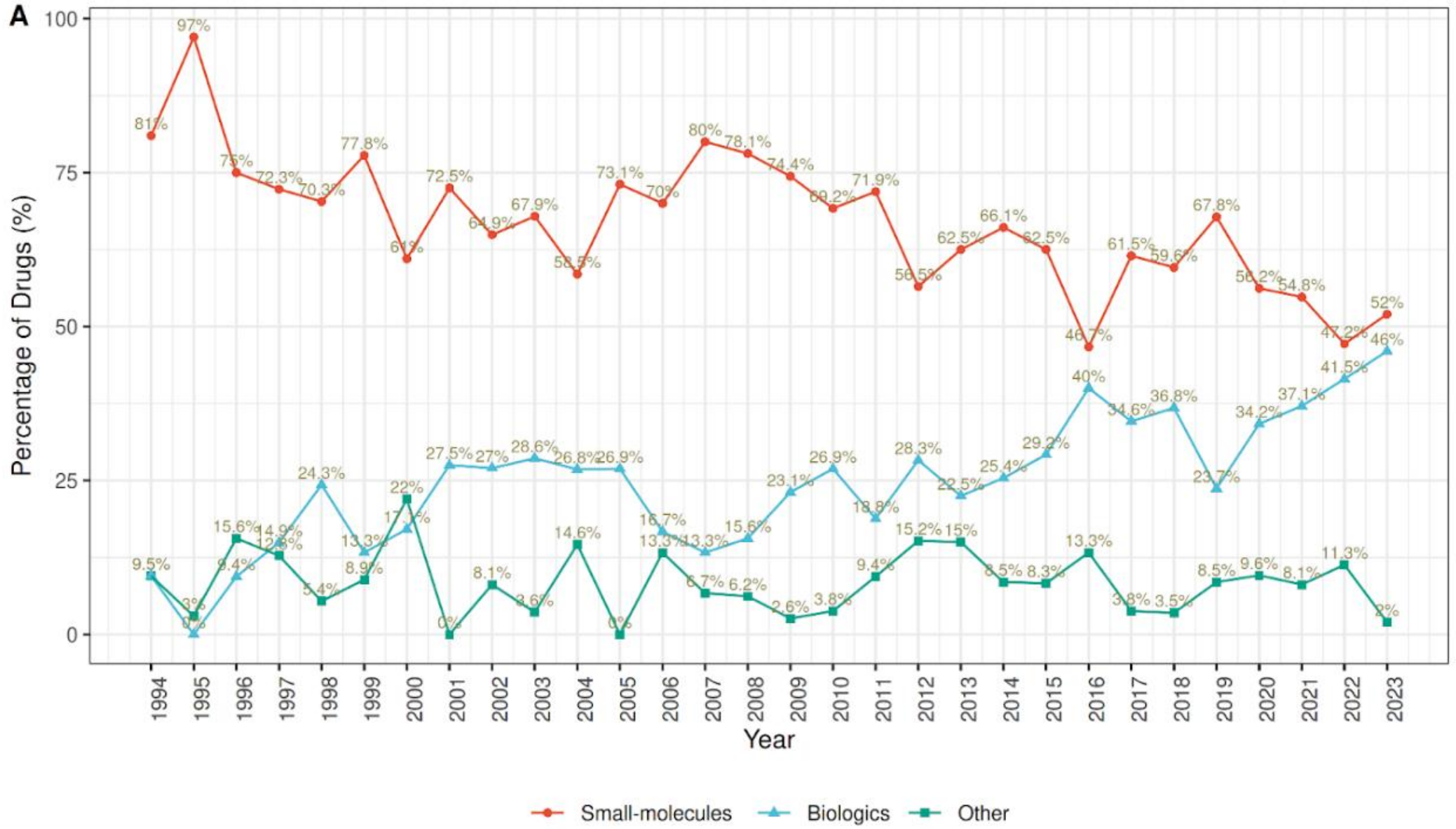
Tdark ("ignorome") have little information available, and satisfy these criteria:

- PubMed text-mining score from Jensen Lab < 5
- <= 3 Gene RIFs
- <= 50 Antibodies available according to antibodypedia.com

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

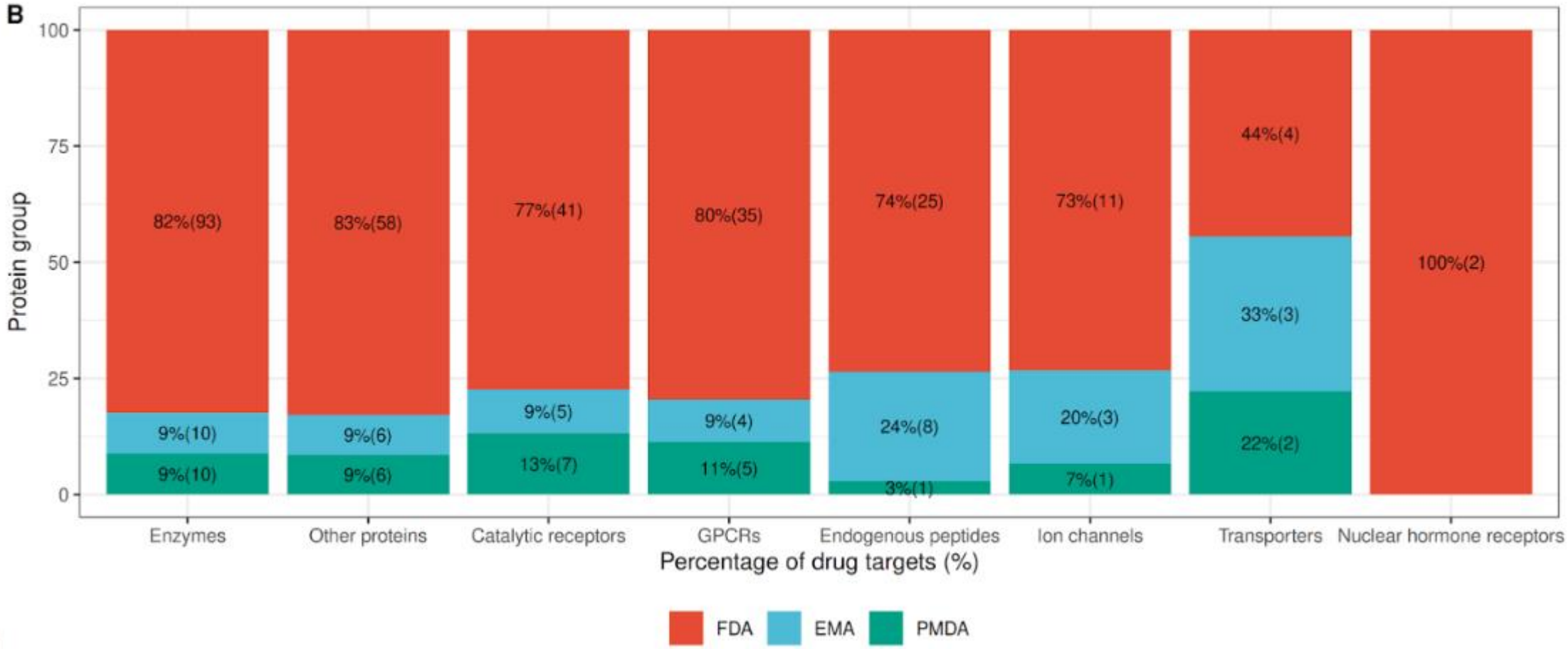
IDG

Trends in Approved Drug Modalities



L. Halip et al., *Nature Rev. Drug Discov*. **2026**, pre-submitted

Target Innovation by Regulatory Agencies (1994 to 2023)



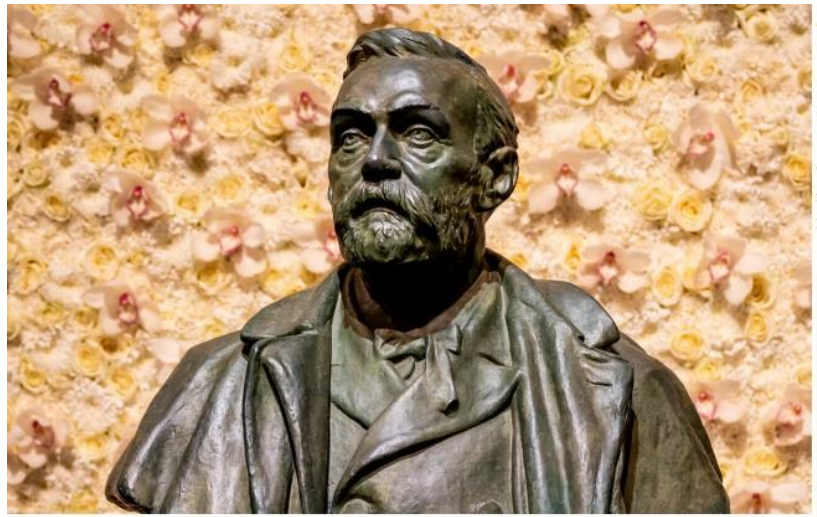
Cumulative distribution of newly launched targets across major protein families by agency, with counts shown in brackets. This shows that the US FDA has a leading role in shaping the modern therapeutic target landscape, driven by both regulatory capacity and innovation incentives.

L. Halip et al., *Nature Rev. Drug Discov*. **2026**, pre-submitted

Artificial Intelligence

General Observations

AI Awarded the Physics & Chemistry Nobel Prizes



Statue of Alfred Nobel. Photo: A. Mahmoud

NOBEL PRIZES 2024

The Nobel Prize in Physics 2024

John Hopfield

"for foundational discoveries and inventions that enable machine learning with artificial neural networks"

John Hopfield. Ill. Niklas Elmehed © Nobel Prize Outreach

Geoffrey Hinton

"for foundational discoveries and inventions that enable machine learning with artificial neural networks"

Geoffrey Hinton. Ill. Niklas Elmehed © Nobel Prize Outreach

[They used physics to find patterns in information]

The Nobel Prize in Chemistry 2024

David Baker

"for computational protein design"

David Baker. Ill. Niklas Elmehed © Nobel Prize Outreach

Demis Hassabis

"for protein structure prediction"

Demis Hassabis. Ill. Niklas Elmehed © Nobel Prize Outreach

John Jumper

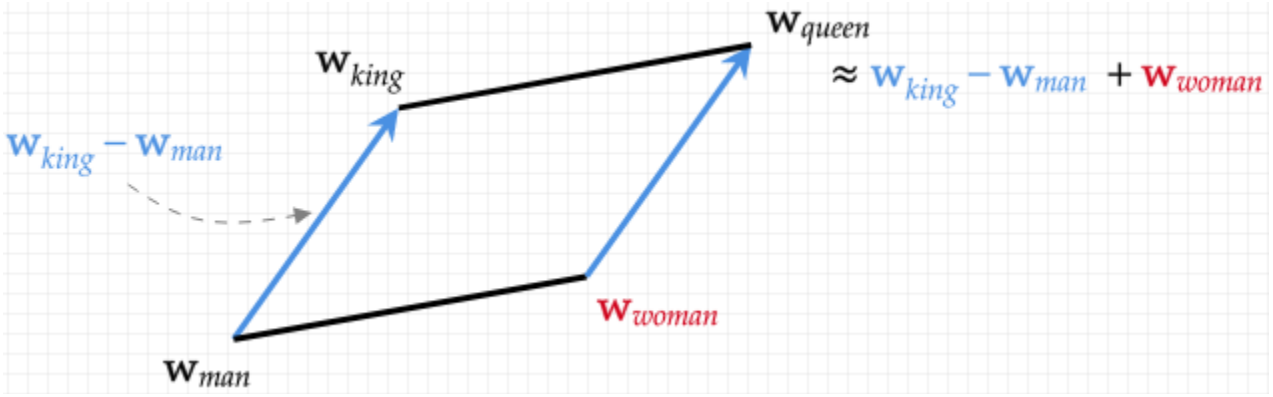
"for protein structure prediction"

John Jumper. Ill. Niklas Elmehed © Nobel Prize Outreach

[They cracked the code for proteins' amazing structures]

Read more about the 2024 Nobel Prizes here: link

Artificial (language) Intelligence



King – Man + Woman = Queen

Tomáš Mikolov (Google) developed an efficient algorithm to compute the distributed representation of words, **Word2Vec**. It formed the basis for the first automatic translators, spam filtering and speech recognition. **Word2vec** encodes words using a distribution of weights across 100s of elements that compose the vectors. Each element contributes to many words. T. Mikolov et al., *ICLR* **2013**

With the advent of LLMs, it did not take long to move from "natural" language processing to AI-driven large-dataset mining. Klingon, anyone?
tlhIngan, vay'?

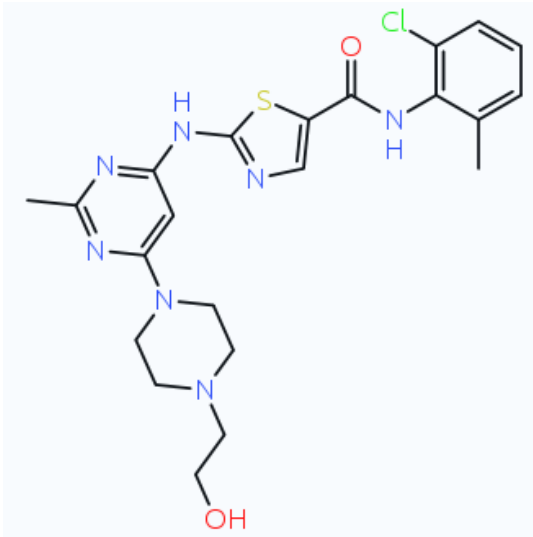
GPT-4 as drug discovery tool - 3/14/2023

Andrew White, a member of OpenAI's "red team," prompted GPT-4 with the name “Dasatinib”, a kinase inhibitor drug. GPT-4 was asked to modify the drug and find novel, non-patented molecules with a similar mode of action, locate chemical vendors selling the compound, and purchase it. If custom synthesis was needed, GPT-4 was to email a contract research organization to order the compound.

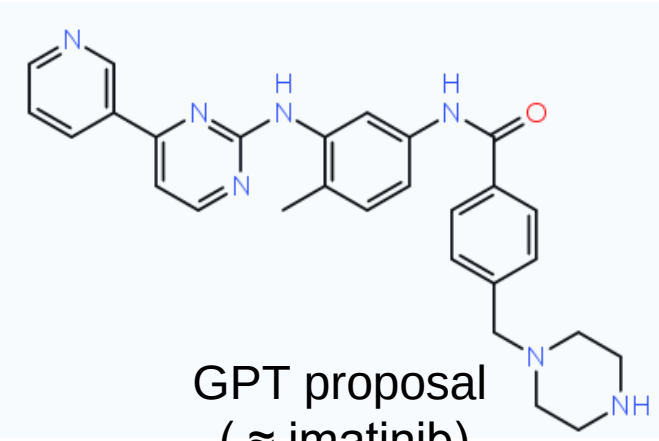
GPT-4 generated a valid SMILES output, indicating its ability to perceive and modify chemical structures correctly; the molecule is available in the ZINC database, meaning it is synthetically feasible

The proposed molecule is desmethyl-imatinib, an N-dealkylated piperidine metabolite of imatinib, another protein kinase inhibitor drug.

GPT-4 successfully modified the molecule while retaining its kinase inhibitor properties. Experimental validation may be required to confirm if this molecule has the same MoA as Dasatinib



https://bit.ly/LI_GPT4_Dasatinib



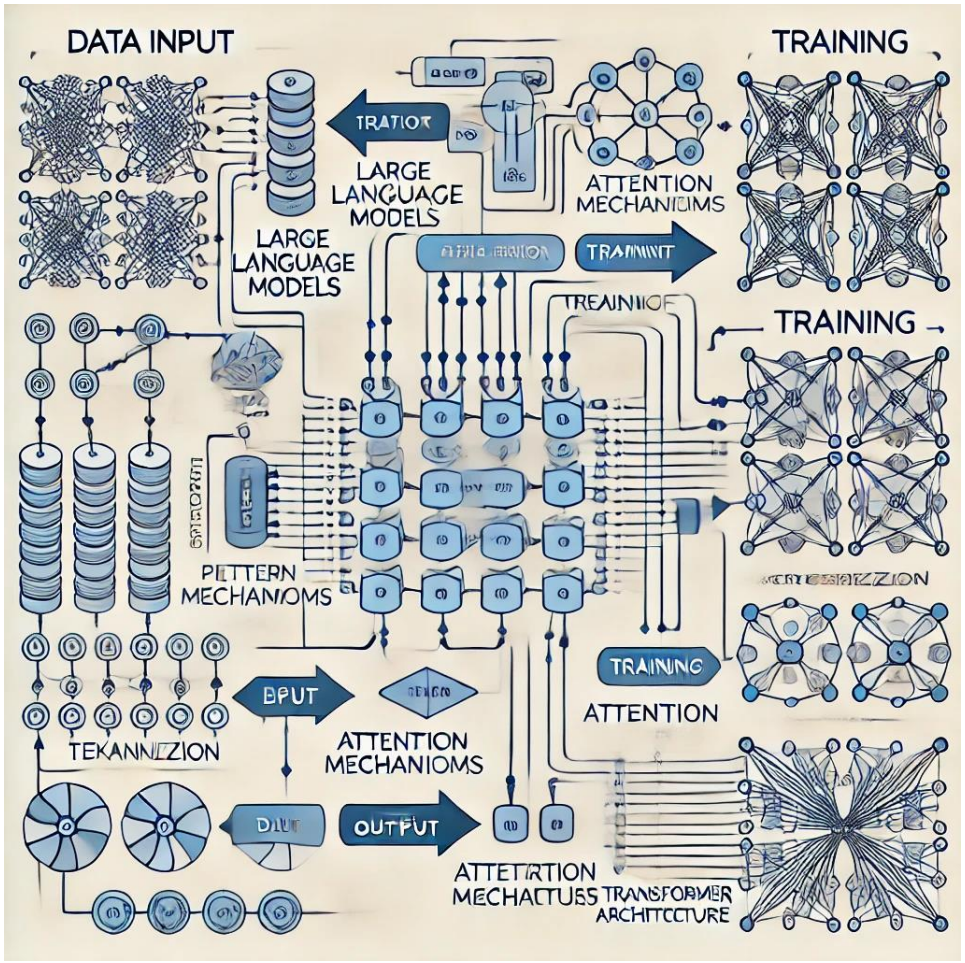
GPT proposal
(≈ imatinib)

C. Hasselgren & T. Oprea, *Ann Rev Pharm Tox* **2024** link

GPT-4o is a Tool, but it is not (an) Artificial Intelligence

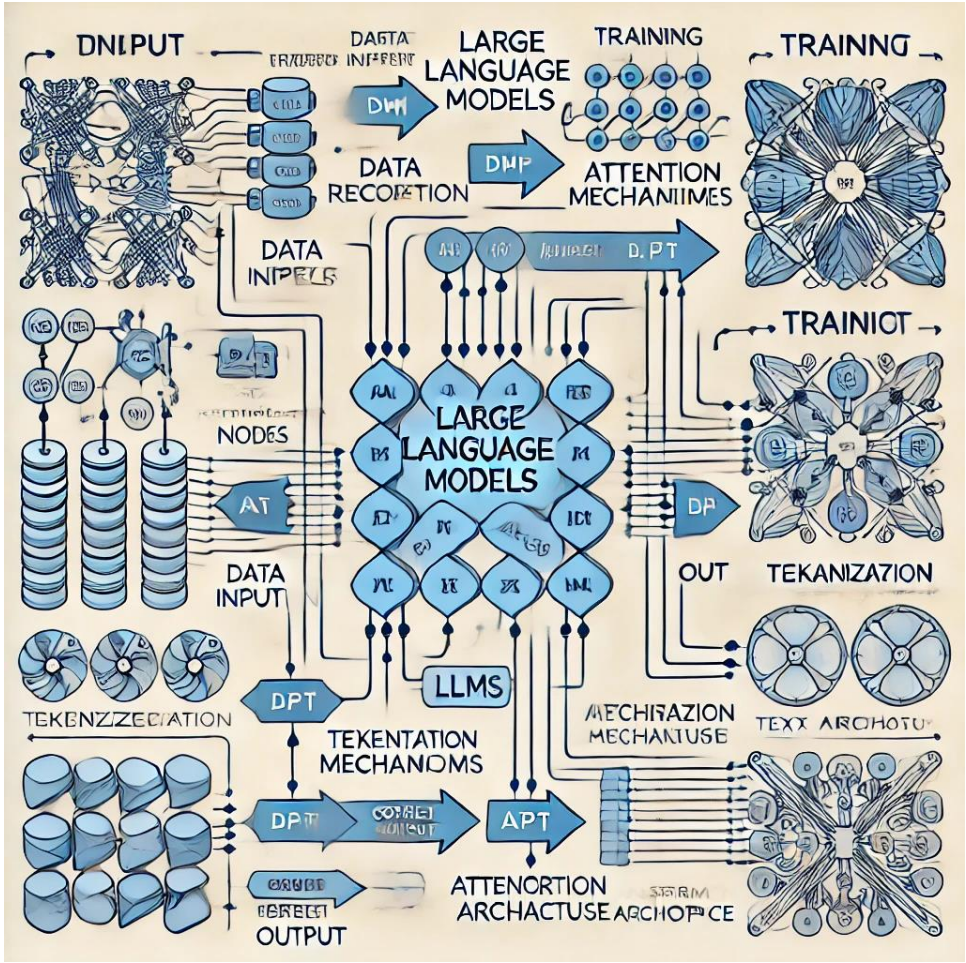
eXpert sYstems

Create an image that provides an accurate visual on how large language models work



Here's an illustrative diagram that visualizes how large language models work. Let me know if there's anything specific you'd like to add or if you need further explanations!

Make sure your spelling is correct for each entry in the above image



I've created a revised version of the diagram with correct spelling for each entry. Let me know if this meets your needs or if there's anything else you'd like to adjust! link

Screenshots generated on **10/14/2024** link

GPT-4o is a Step towards Artificial General Intelligence

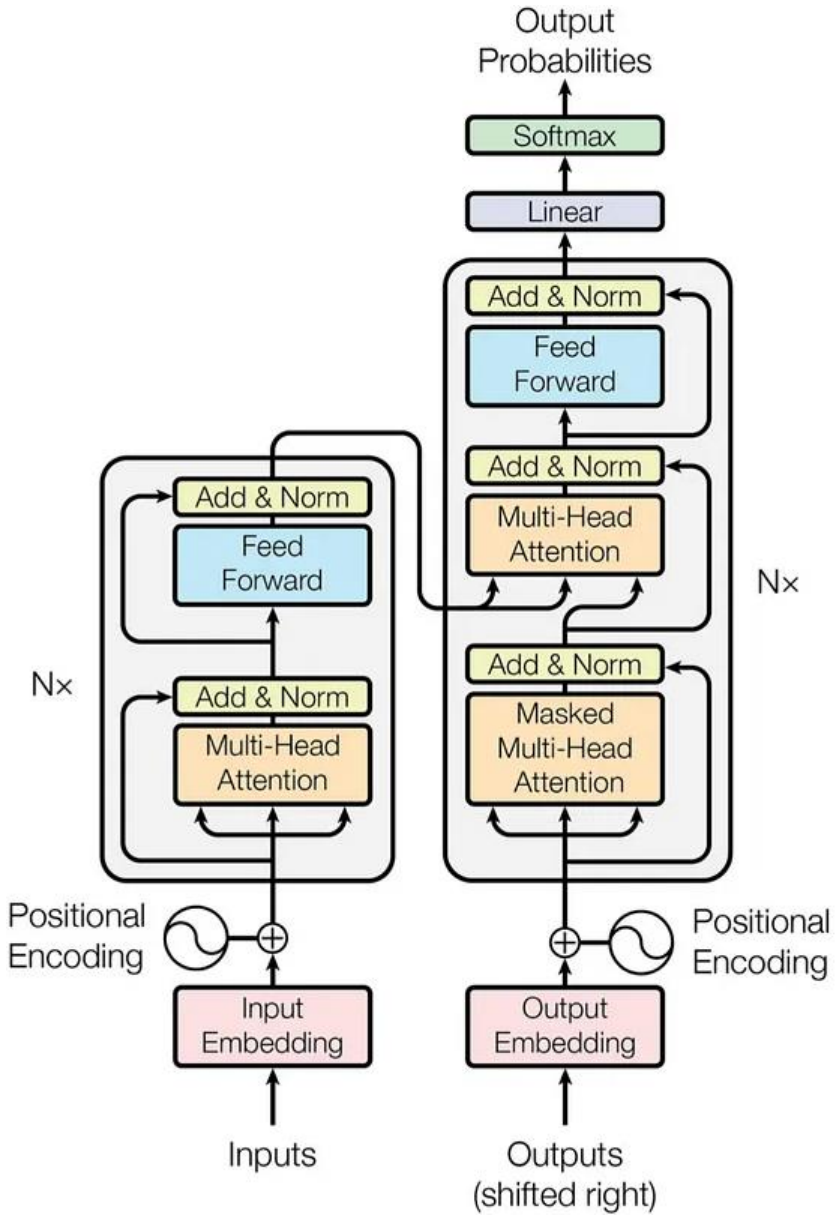
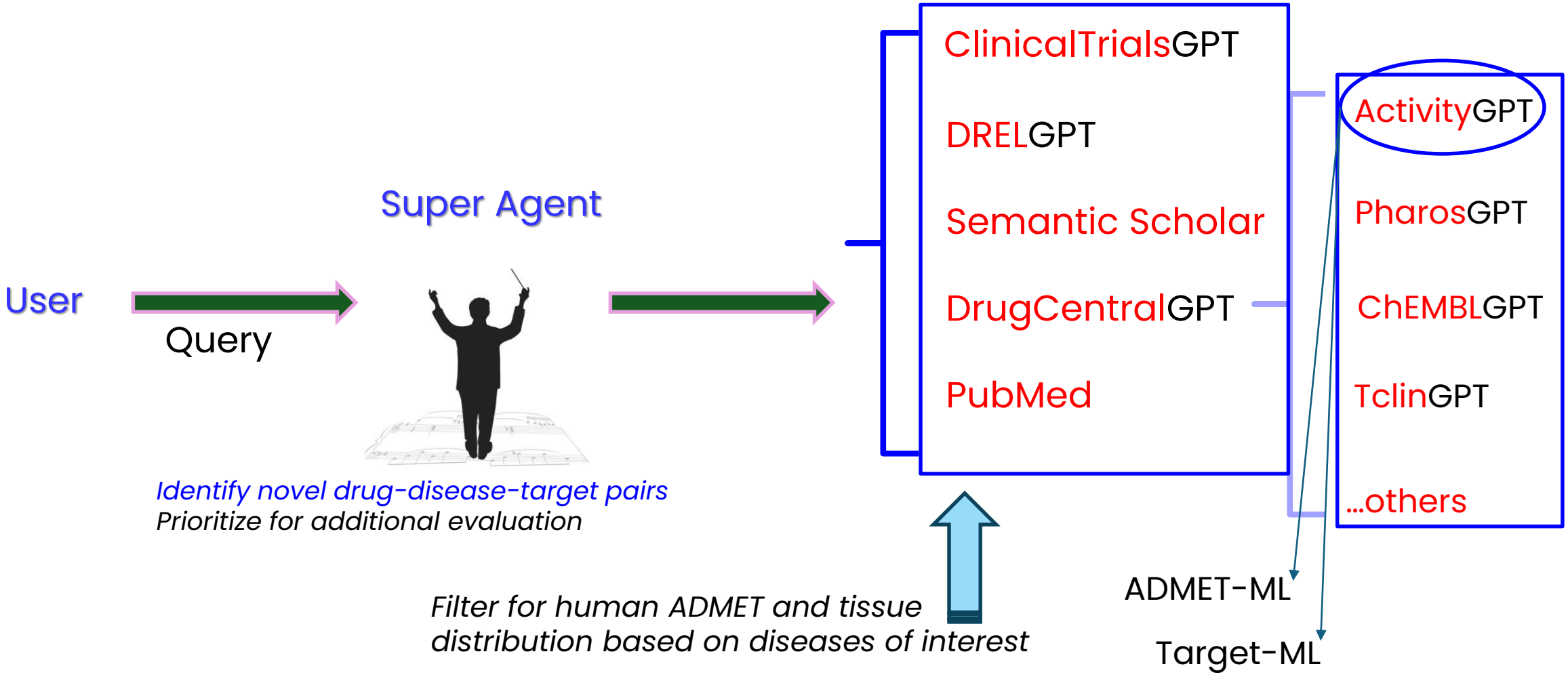


Image from A. Chazy **3/3/2023** link

ChatGPT is based on the decoder-only Transformer architecture. Key components include:

- **Decoder-only architecture:** Unlike original Transformers (encoder & decoder stacks), ChatGPT uses only the decoder. This allows the model to generate texts based on input (prompts).
- **Multi-head self-attention layers:** They allow the model to generate each output token using different parts of the input sequence.
- **Feed-forward neural networks:** These NNs process the output of the attention layers.
- **Large number of parameters:** ChatGPT's architecture includes many layers of transformers. For example, GPT-3 has 120 layers of transformer decoders.
- **Multimodal capabilities:** beyond text-based models, GPT4 embeds both text & image-processing layers
- **Mixture-of-Experts:** OpenAI uses 16 experts in GPT4.

Expert Systems Strategy for LLMs:
Use Super-Agents to coordinate LLMs and AIML models



A super agent orchestrates the flow of user query to the appropriate tool(s)

About

Expert Systems

Expert Systems, Inc

Trusted Partner to Accelerate Drug Discovery

Oncology Space

Eilean Therapeutics link
pan-variant therapies for hematologic malignancies
Lomond Therapeutics link
precision hem/onc
Mondego Bio link
improving cancer immunotherapy

Neurology Space

Brenig Therapeutics link
pioneering treatments for neurodegenerative disorders
Mair Therapeutics link
restoring lysosomal function
Polku Therapeutics link
clearing protein aggregates from the brain

and more

seven compounds in Phase 2 clinical trials
six compounds in Phase 1 clinical trials
ExSys currently supports over 30 programs from 25 companies

Expert Systems, Inc – Key Participants

Tudor Oprea, MD PhD
Co-Founder, CEO

Alexei Pushechnikov, PhD
Co-Founder, CTO

Prof Ruben Abagyan, PhD
Chief Advisor

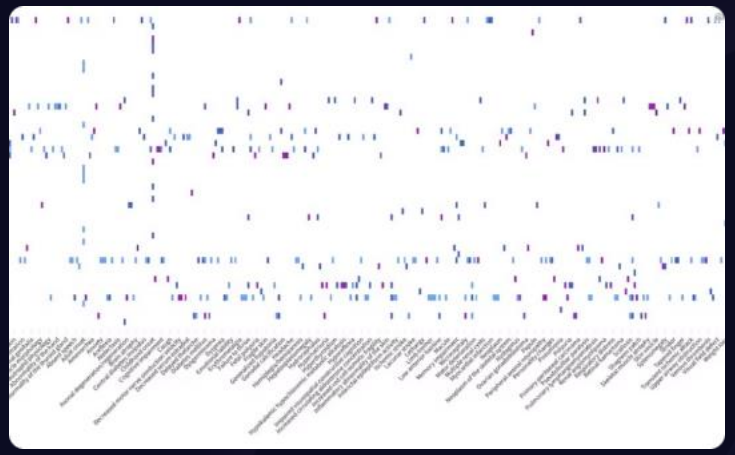
Nikolay Savchuk, PhD
Manager, Teal Ventures

Eddie Rodriguez, JD
Co-Founder, General Counsel
Partner, Teal Ventures

Mohammad Quazi, PhD
Biostatistics
Machine Learning

Suman Sirimulla, PhD
Pharmacology
Cheminformatics

LLM-Powered Therapeutic Area Expansion



Rare Disease Mapping

Our platform currently maps to 8,101 rare diseases, complete with phenotype and gene associations, enabling researchers to identify novel therapeutic opportunities in underserved disease areas.

Common Disease Coverage

Comprehensive mapping of 1,353 common diseases with phenotype and gene associations provides a solid foundation for expanding existing therapeutic approaches to new indications.

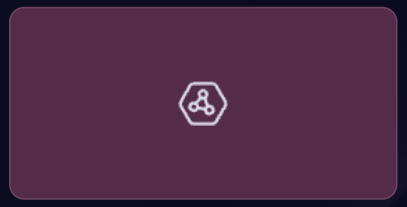
LLM Model Integration

Based on multiple state-of-the-art LLM models including Claude 3.x Sonnet, GPT4 ox, TxGemma & DeepSeekR1, our system leverages the latest advances in natural language processing for biomedical applications.

Machine Learning Targets

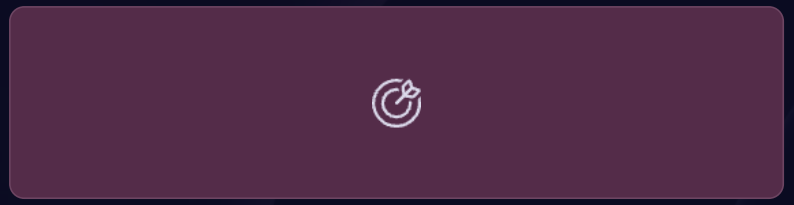
Knowledge Graph Platform:

Understanding Proteins and Diseases



Knowledge Graph Foundation

Specialized platform focused on target-disease associations



Novel Target Identification

Successfully identified targets in complex diseases

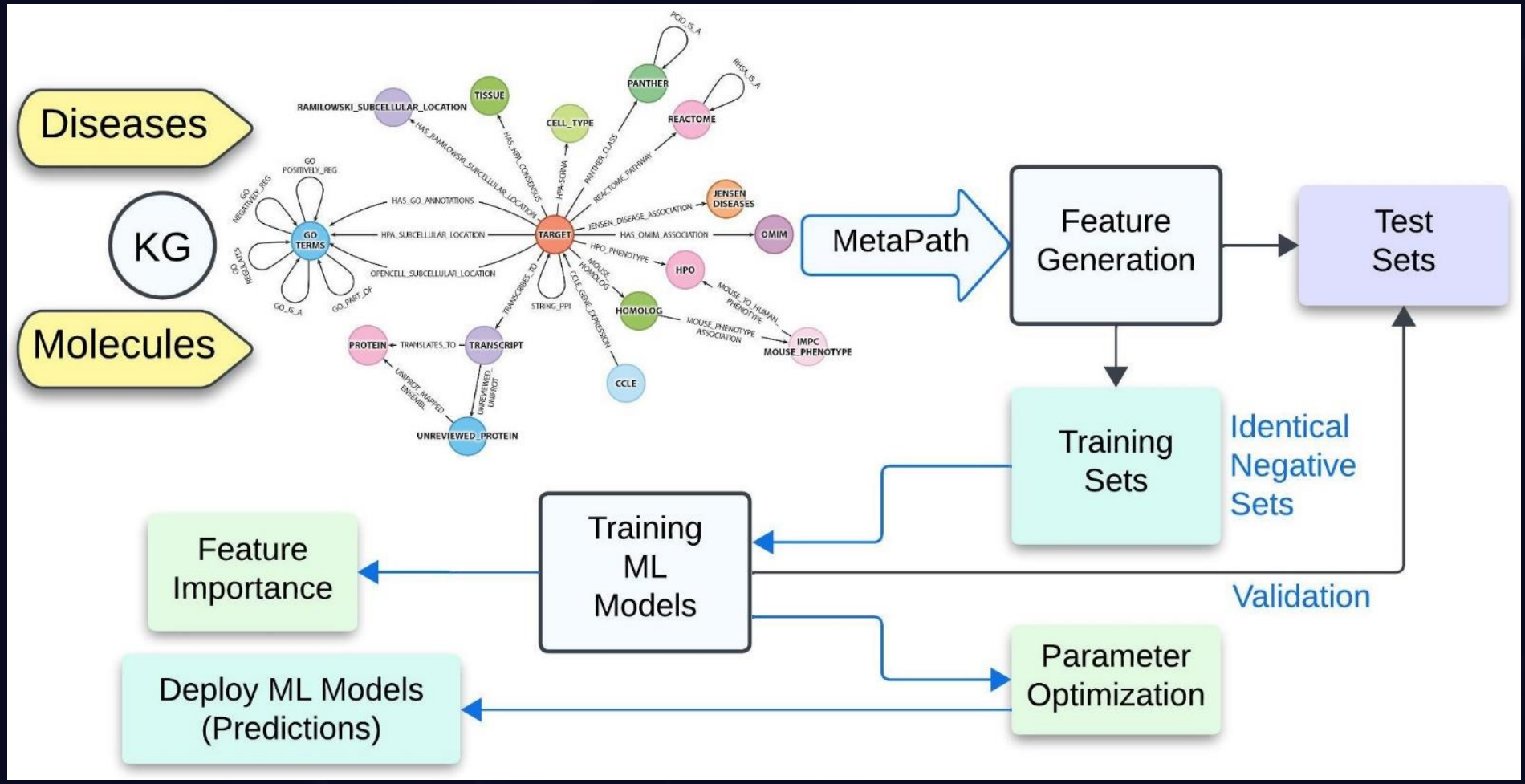
Validated Research Applications

Published findings in Alzheimer's, autophagy and hematology-oncology

Our separate ML platform is knowledge-graph based and focuses on target-disease associations. Our system has already demonstrated success in identifying novel targets in Alzheimer's disease, autophagy (with published results), and non-Hodgkin lymphomas (presented as posters at AACR 2024/2025).

KGML Workflow

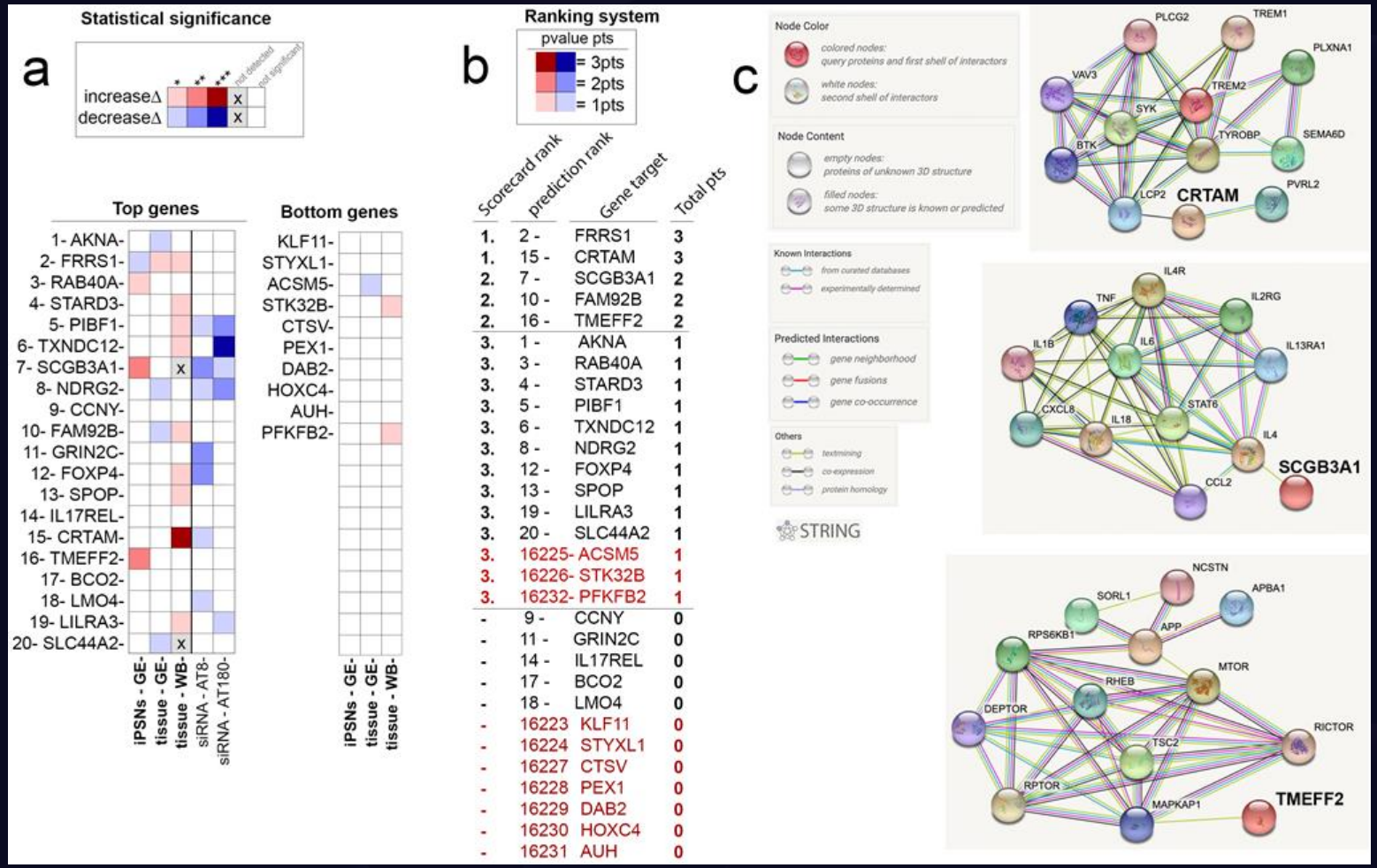
Target-Centric Graphs for Novel Phenotype Associations



Hasselgren & Oprea, Ann Rev Pharm Tox 2024 64:527-550 link

KGML Example: Disease

Validated Model for Alzheimer's Disease



KGML Model:

Top 20 & Bottom 10

Five of the top 20 predicted genes show potential AD relevance: FRRS1, SCGB3A1, CRTAM, TMEFF2, FAM92B (renamed CIBAR2).

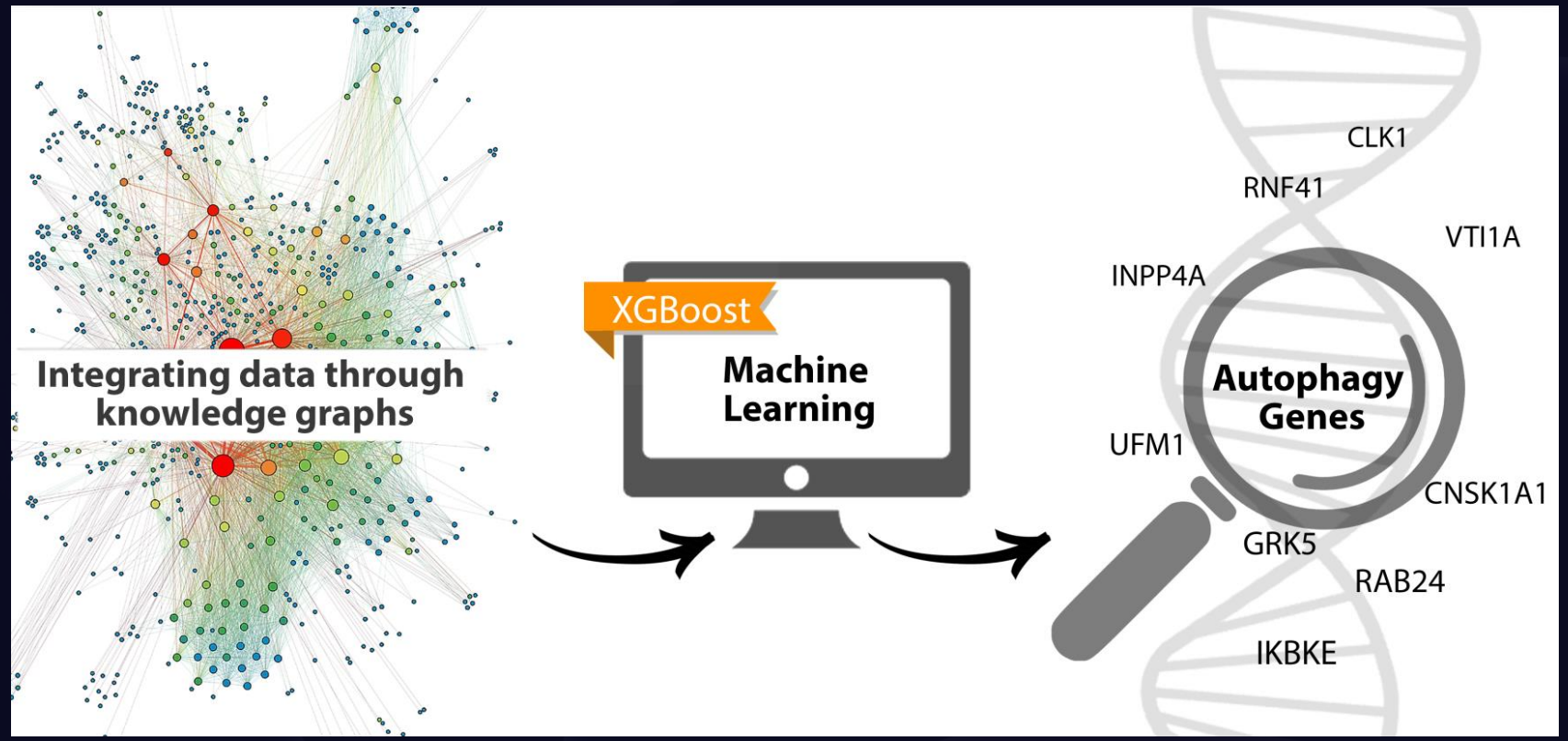
CRTAM, SCGB3A1, & TMEFF2 are connected to TREM2-TYROBP, IL-1β-TNFα, and the MTOR-APP AD risk nodes, with relevance to AD pathogenesis.

Three of the bottom 10 genes showed signal of association with AD.

Binder et al., Communications Biology 2022, 5:125 link

KGML Example: Physiology

Validated Model for Autophagy



KGML Model:
Top 20 & Bottom 10

Seven of the top 20 and two of the bottom 20 targets have confirmed links to autophagy

Ranjbar et al., Natural Sciences 2023, e20220067 link

Learning from Machine Learning

- Our Alzheimer's ML model bias was introduced when selecting positive genes only
- *Among the top 20 VIP features […], there are protein–protein interactions (PPIs) for inflammatory process mediators that are in the positive training set (JAK2, IL10, and IL2), as well as PPIs with the oxidative stress response protein (GSTP1).*
- *These PPIs suggest infection, which is when oxidative stress and inflammation co-occur (e.g., phagocytes producing reactive oxygen species).*
- Imagine having access to ML models that don't serve a specific agenda
- ML bias is caused by humans at inception (data "selection") or output (model "interpretation")

RETRACTED ARTICLE: A specific amyloid-β protein assembly in the brain impairs memory

Sylvain Lesné, Ming Teng Koh, Linda Kotilinek, Rakez Kayed, Charles G. Glabe, Austin Yang, Michela Gallagher & Karen H. Ashe

Nature **440**, 352–357 (2006) | Cite this article

65k Accesses | **2348** Citations | **1665** Altmetric | Metrics

This article was retracted on 24 June 2024

Hasselgren & Oprea, Ann Rev Pharm Tox **2024** link

Learning from Machine Learning 2

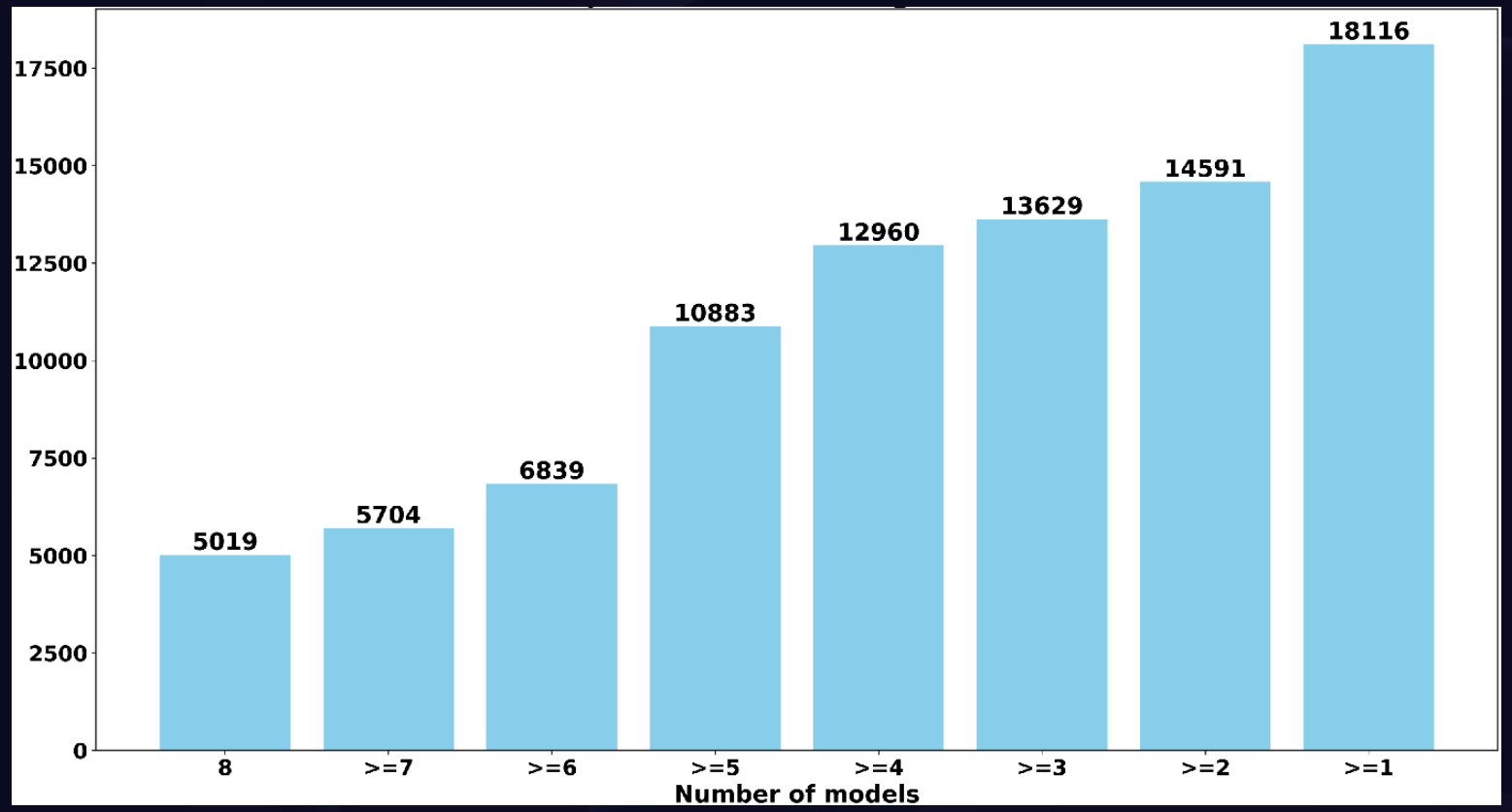
- The curious case for bottom-ranked predictions

- Most ML practitioners examine the top X ranked predictions (we made the same error).
- One reviewer asked we prove bottom-ranked genes are not relevant in AD. Some where
- We were less surprised that some bottom-ranked genes also have relevance to autophagy.
- *Binary classification models train from "0 and 1" labels, but often lack directionality.*
- *Absence of signal in biology can imply blockade, which may be just as relevant as an existing signal (activation).*
- Examine both top and bottom ranked predictions for biological relevance.

Hasselgren & Oprea, Ann Rev Pharm Tox **2024** link

KGML Example: Target Druggability 1

KGML Model Training



KGML Model:

No Data leakage

1) Pharos 5.4 database, current as of December 2017 used for training.

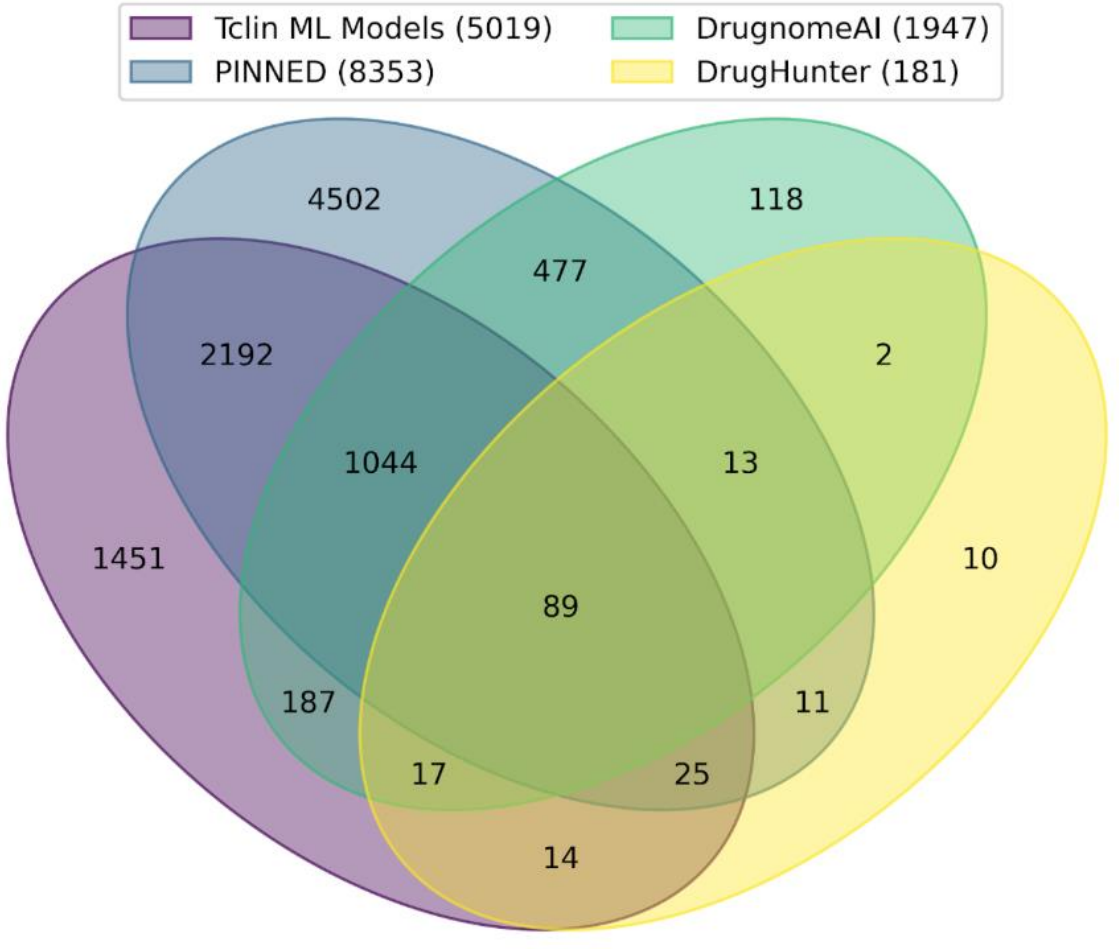
2) All Tclin proteins approved between 2018 and 2024 served as test set.

3) 1SVM (unbiased) model was used to identify “non-Tclin” targets, which served as basis for XGBoost.

Quazi et al., in preparation, 2026

KGML Example: Target Druggability 2

Targets in Phase 1-2-3 Clinical Trials



Prediction method	Phase 1 (N = 43)	Phase 2 (N = 83)	Phase 3 (N = 55)
Tclin-like	33 (76.74%)	69 (83.13%)	43 (78.18%)
PINNED	31 (72.1%)	70 (84.33%)	40 (72.7%)
DrugnomeAI	21 (48.8%)	63 (75.9%)	38 (69.1%)

External prediction: ~74.7% of the novel MoA targets
86 out of 115 proteins from the 2018-2024 Tclin list
Accurately predicted:

- 17 out of 22 for 2018 drug targets
- 11 out of 19 for 2019 drug targets
- 10 out of 14 for 2020 drug targets
- 15 out of 17 for 2021 drug targets
- 11 out of 15 for 2022 drug targets
- 10 out of 13 for 2023 drug targets
- 8 out of 9 for 2024 drug targets

Quazi et al., in preparation, 2026

KGML Example: Target Druggability 3

What Constitutes a Good MoA Target?

There is a significant distinction between "protein/gene A is associated with disease D" and "protein/gene A is a drug target for disease D."

The distinction is akin to differentiating between the butcher and the turkey in a scenario – both are involved, but only the turkey is committed.

In target discovery, the objective is to find 'committed' targets.

Essential criteria for validating an MoA target:

1. *A causal relationship with disease pathology* - MoA targets modulate disease outcomes.
2. *Biological redundancy and pathway involvement* - look for escape mechanisms.
3. *Clinical relevance* - the MoA-perturbing medicine causes the desired clinical outcome
4. *Safety and specificity* - lack of on-target toxicity
5. *Patient stratification and biomarkers* - which subgroups benefit the most?

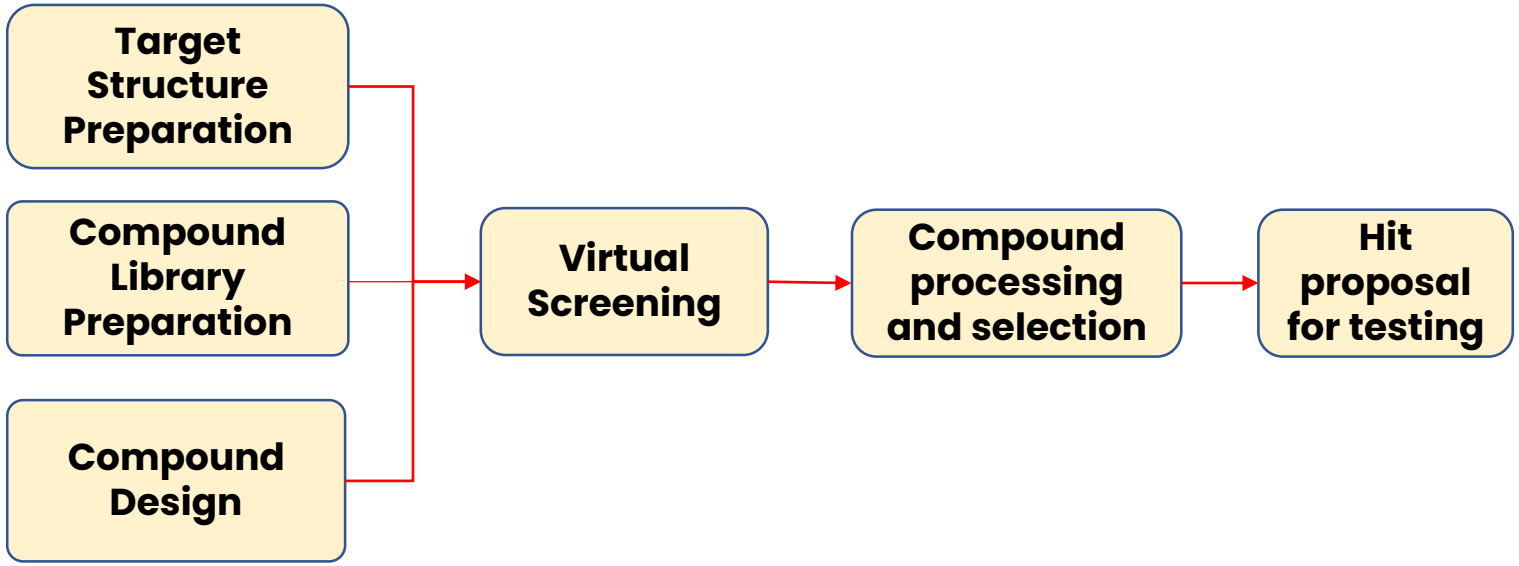
Quazi et al., in preparation, 2026

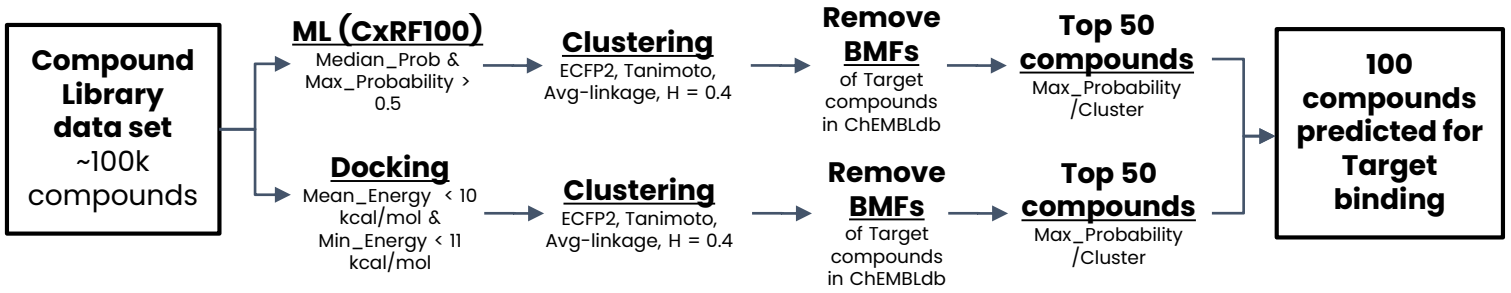
eXpert sYstems

Drug Hunting

Target-specific Work

Target Evaluation: Ligand Workflow





This process typically includes:

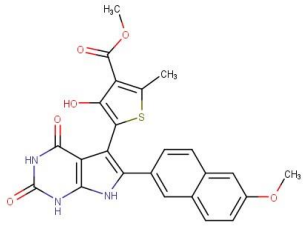
- Patent evaluation
- Literature evaluation
- ML model development based on known compounds
  - Structure-based virtual screening

  or
  - Ligand-based virtual screening
- Compound design (where appropriate)

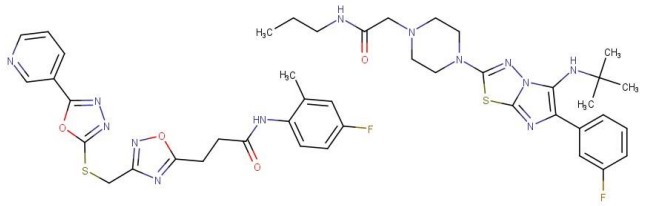
We use proprietary fingerprints, public and proprietary data to evaluate the competitive intelligence landscape and identify potentially novel scaffolds.

Confidential

Target Evaluation: Ligands

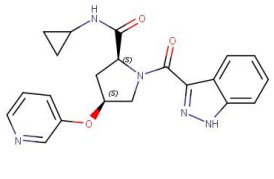


KIL_D517-0088 RF100 Score: 0.582



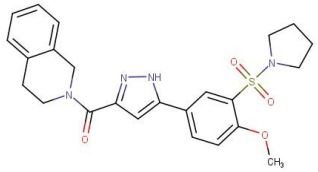
KIL_J021-3499 RF100 Score: 0.573

KIL_L532-2092 RF100 Score: 0.572

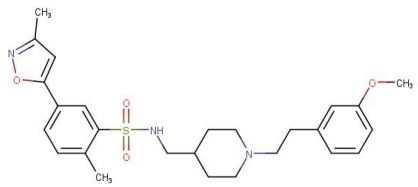


KIL_S721-0978 RF100 Score: 0.572

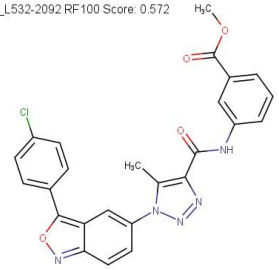
KIL_D511-0920 Docking Score: -13.11



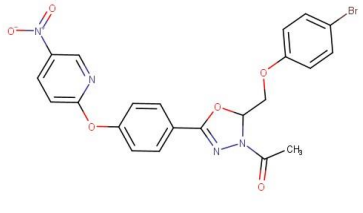
CNSTAS_G008-2643 Docking Score: -12.98



KIL_D443-1953 Docking Score: -12.93

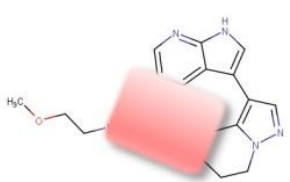


KIL_G642-6601 Docking Score: -12.83

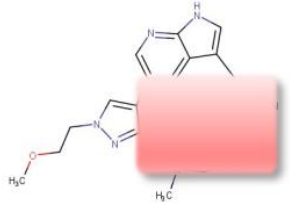


Top <Library> compounds

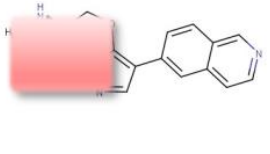
(strong support from ChemDiv and other chemical CROs)



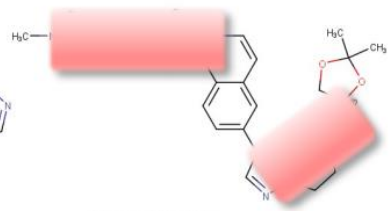
2_8 RF100 Score: 0.578



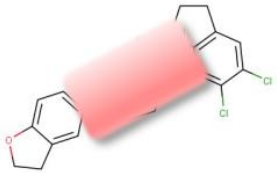
2_3 RF100 Score: 0.593



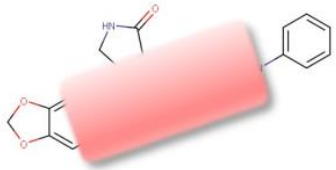
3_61 RF100 Score: 0.559



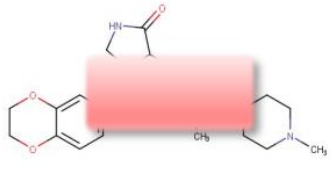
3_2 RF100 Score: 0.558



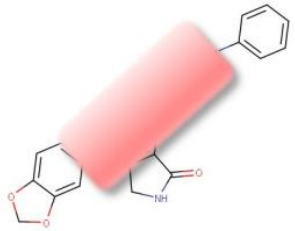
1_97 Docking Score: -10.79



5_8 Docking Score: -11.85



5_20 Docking Score: -11.73



3_9 Docking Score: -11.78

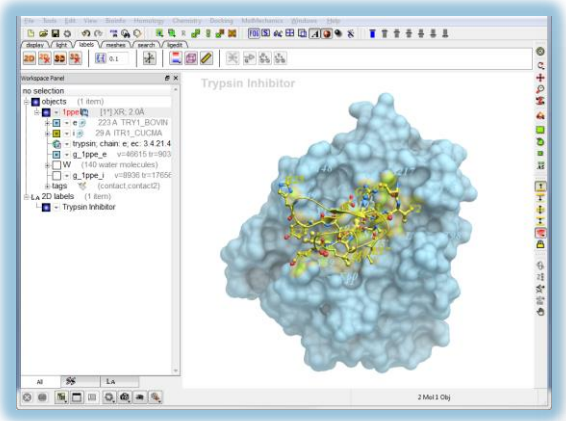
Proprietary compound design

(from medicinal chemists and generative AI)

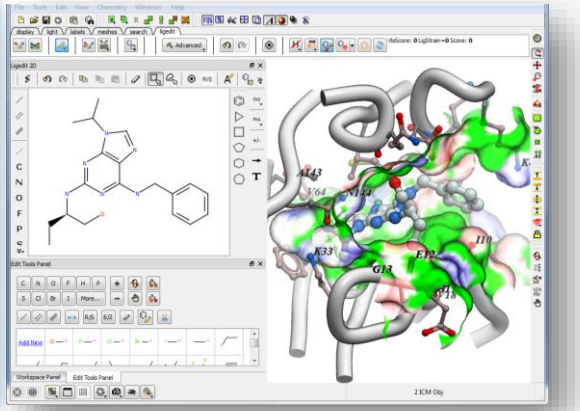
ExSys Molecular Design Workflow using MolSoft

eXpert sYstems

AI selection of druggable binding pockets

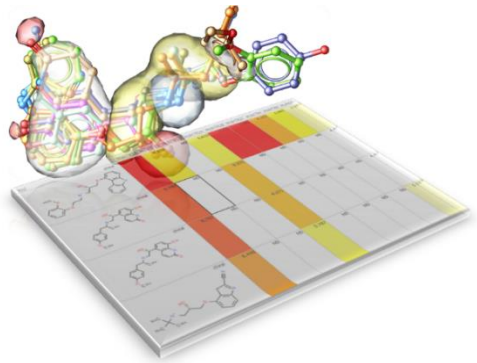


Interactive drug design in 3D Off-targets & properties



MolScreen: Thousands of Models for drug design targets and properties

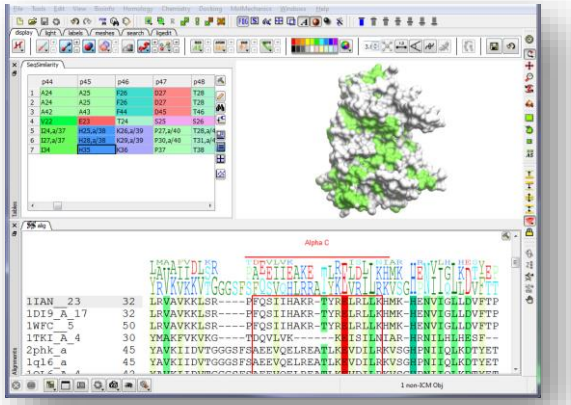
M Totrov, E Raush



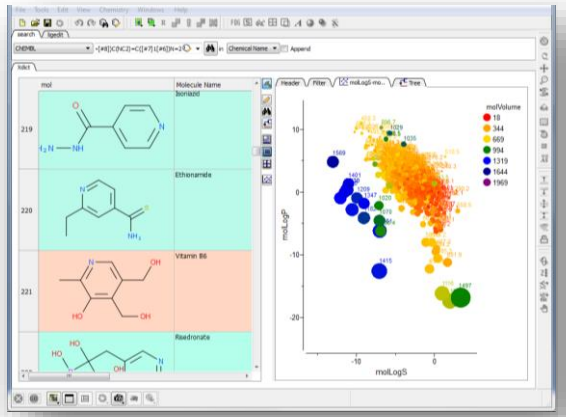
Combining Deep Learning with Docking. Thousands of preclinical activity assays for drug candidates. Thousands of **AI/ML** model types

- Fully integrated with ExSys workflow for ADMET, Tox Screening

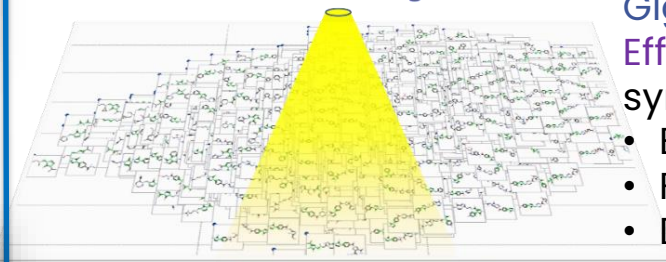
Models of mutants and variants: Drug selectivity



Integration with proprietary chem/bio arsenal



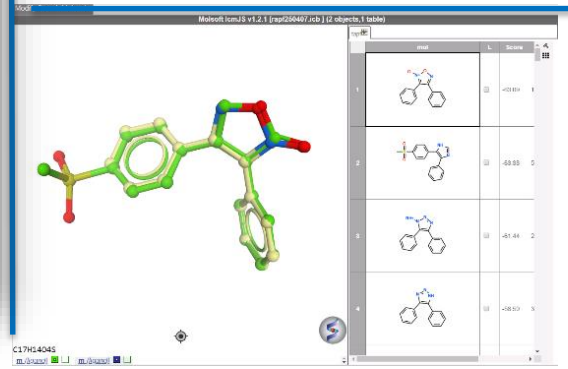
GIGA Search-Engine



Giga Search in Billions of Compounds Efficiently mine many billions of synthesizable chemicals.

- Build target-specific libraries.
- Find chemical homologs.
- Drive your SAR search.
- Find real chemical derivatives.

RIDE: 3D and pharmacophoric search through 50 billion compounds



Extremely fast ligand-based search

- 0.5 Million Chems/Sec/GPU.
- Virtual screening
- Scaffold hopping
- Hit follow up

molsoft
molecules *in silico*

Strategic Partnership with Molsoft & Prof Ruben Abagyan, Polo Lam, Max Totrov, Eugene Raush

Navigating Ultra-Large Chemical Spaces

eXpert sYstems

SMILES -> low E conformers for 1B catalog

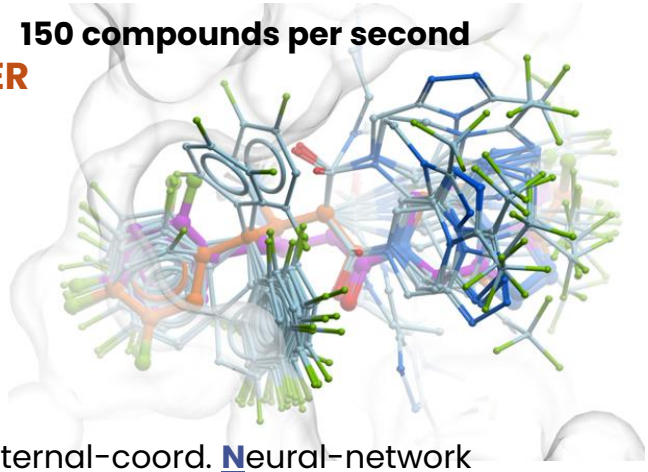
> J Chem Theory Comput. 2024 May 14;20(9):4054-4063. doi: 10.1021/acs.jctc.4c00280. Epub 2024 Apr 26.

Efficient Generation of Conformer Ensembles Using Internal Coordinates and a Generative Directional Graph Convolution Neural Network

Eugene Raush [1], Ruben Abagyan [2], Maxim Totrov [1]

150 compounds per second

GINGER



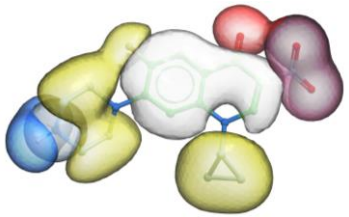
Graph Internal-coord. Neural-network **conformer Generator,** Energy Refinement

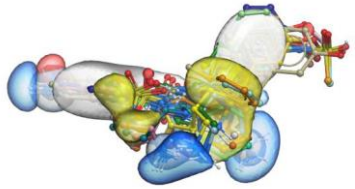
- Each conformation is FF- minimized

Performance: 10 M compounds/day on a single RTX4090

Can also be applied to a space-generated compound on-the-fly

Query: one several superimposed



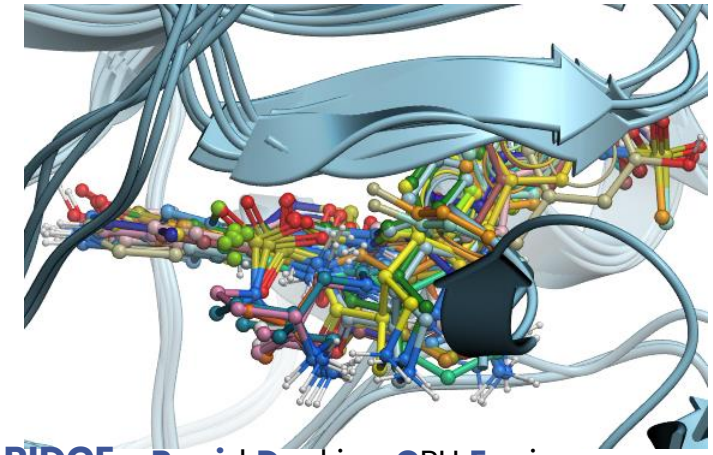


RIDE - Rapid Isostere Discovery Engine
~ Million compounds per second
Runs against a GINGERed catalog
Atom-weighting & Excluded Volumes

Iterative Docking Screens

- **Giga-Screen: RIDGE a subset → Train for GCNN Score Model** → Search Forward → **RIDGE all hits → Hits**

Ultra-Fast GPU/cpu Flex Docking



RIDGE - Rapid Docking GPU Engine
100 dockings per second

Annotate a Long hit list with Targets and Properties
→ **MolScreen Models, Properties, Refinement, Optimization** →
Short Hit list

molsoft molecules *in silico*

Strategic Partnership with Molsoft & Prof Ruben Abagyan, Polo Lam, Max Totrov, Eugene Raush

Machine Learning

Small Molecules

Bioactivity Models:

Comprehensive Target Coverage

Extensive Target Coverage

6,159 total ML models based on targets with defined UniProt IDs, including 2,982 human, 816 mouse, and 813 rat proteins.
In total, 5,120 proteins are of mammalian origin.

Quality-Tiered Models

1,821 high and medium-quality ML models with over 2 log unit spread and 80+ training compounds.

Additional 1,860 lower-quality and 2,478 similarity-based models for in-depth analyses.

Cell Line Analysis

1,240 ML models for cell lines, primarily focused on cancer cell cytotoxicity and viability assays, all with built-in uncertainty quantification for each prediction.

eXpert sYstems

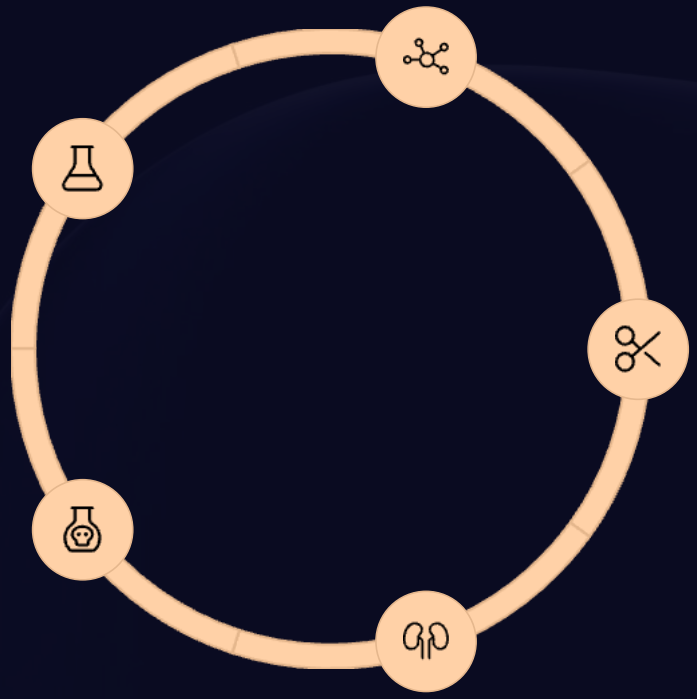
ADMET Models: Critical Drug Property Prediction

Absorption

Solubility, Permeability, Bioavailability

Toxicity

Cytotoxicity, Cardiotoxicity, Skin Toxicity, DILI, etc



Distribution

Plasma Binding, BBB Permeability, VDss

Metabolism

CYP Enzymes, Hepatocyte / Microsomal Stability

Excretion

Hepatic/Renal Clearance, Urinary Excretion

Our ML platform is based on one of the industry's largest curated datasets, including tens of thousands of endpoints for hERG & CYP inhibition, plasma protein binding, and other human PK data, with extensive water solubility and permeability models. Most models provide reliable predictions with uncertainty estimates below 0.15.

eXpert sYstems

Comprehensive Dataset: Foundation for Accuracy

80k+

hERG Toxicity

For cardiac toxicity modeling

> 30k

Thermodynamic Water Solubility

Caco2 *and* MDCK-MDR1 permeability

16k+

CYP Inhibitors

Over 50 Drug Metabolizing Enzyme ML models

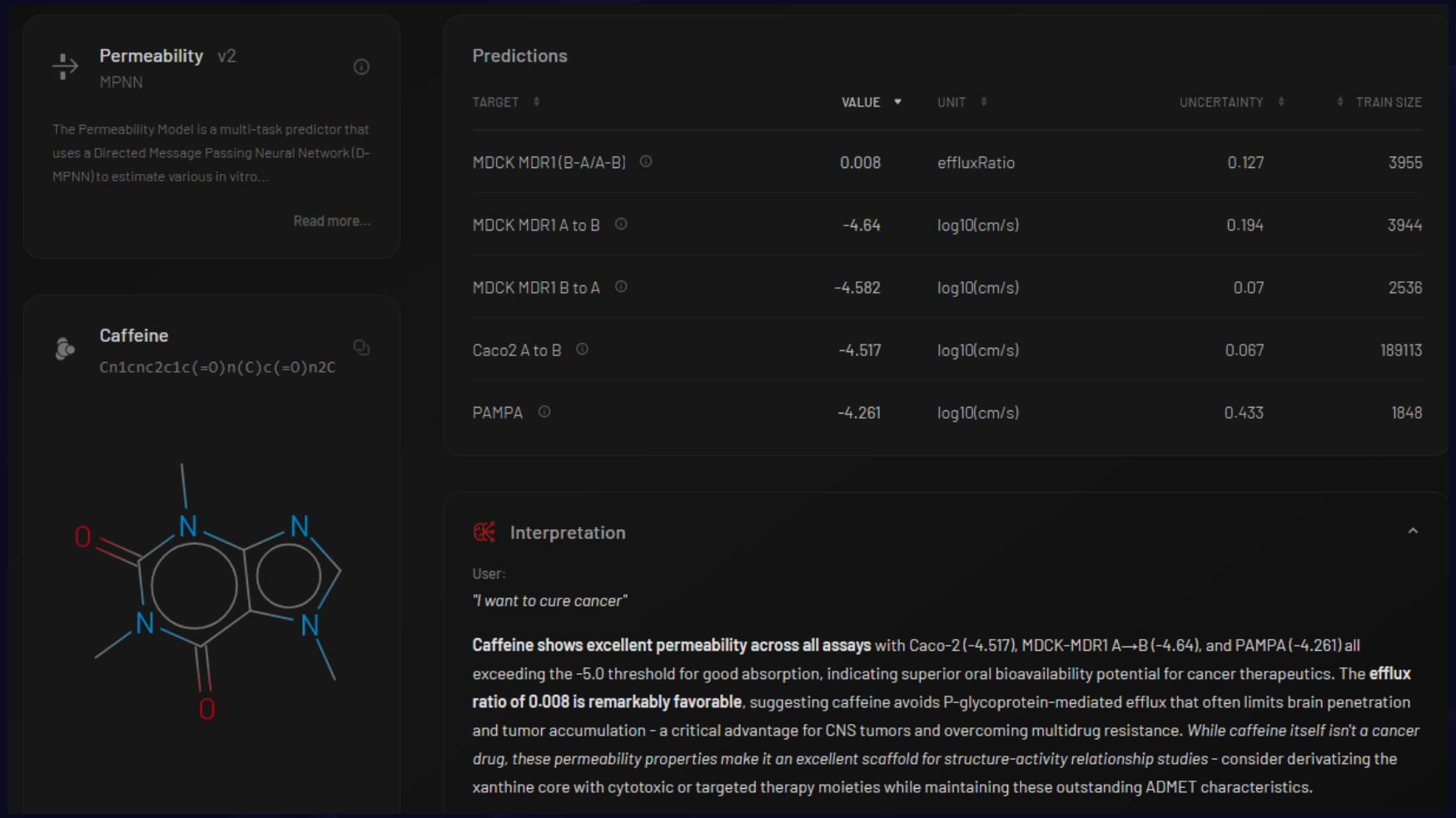
>30k

Human / Rat / Mouse Liver Microsomes

For metabolism stability

Our large curated dataset powers ML models trained on thousands of compounds, delivering robust predictions across diverse chemical and biological properties.

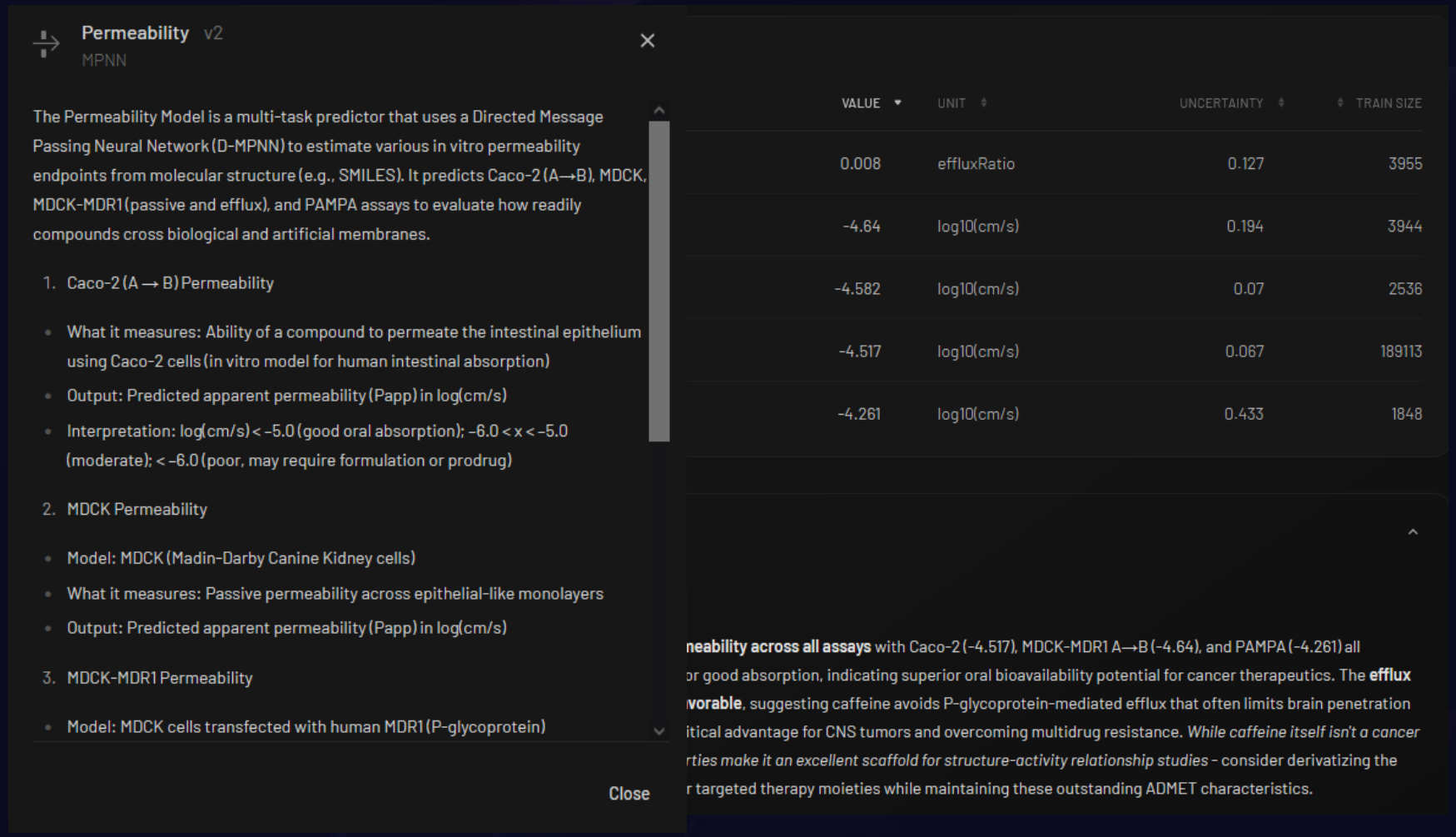
Expert Systems Predict App – LLM Interpretation



All our models are built with ChemProp (multi-task mode), where bundled (similar) properties inform each other & improve predictions.

https://predict.expertsystems.app

Expert Systems Predict App – Model Description

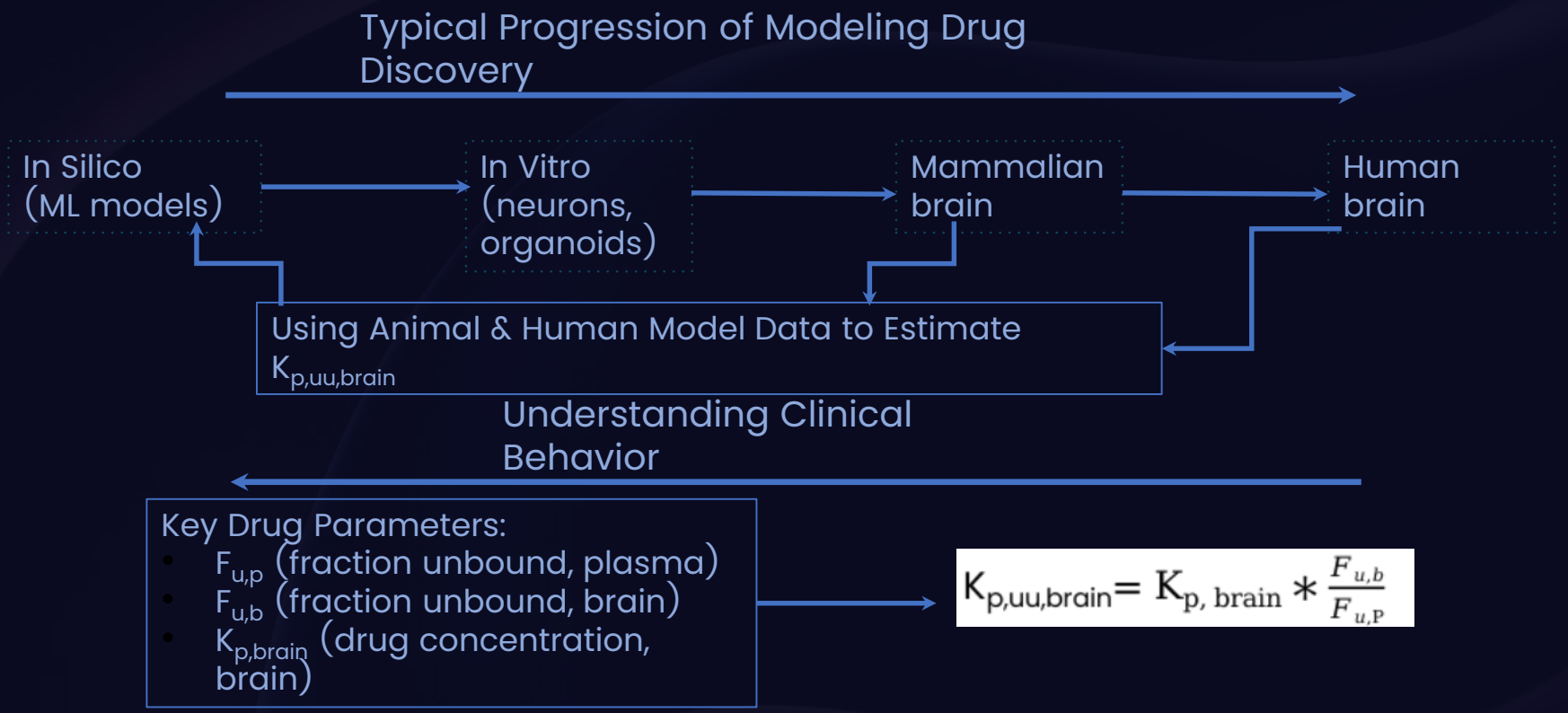


All our models are built with ChemProp (multi-task mode), where bundled (similar) properties inform each other & improve predictions.

https://predict.expertsystems.app

Modeling

Drug-Brain Pharmacokinetics



When transporters are not involved, $\log K_{p,uu,brain} = \log BB + \log F_{u,brain} - \log F_{u,p}$

When transporters are involved, $\log K_{p,uu,brain} = \log BB + \log F_{u,brain} - \log F_{u,p} - \log(\alpha ER)$

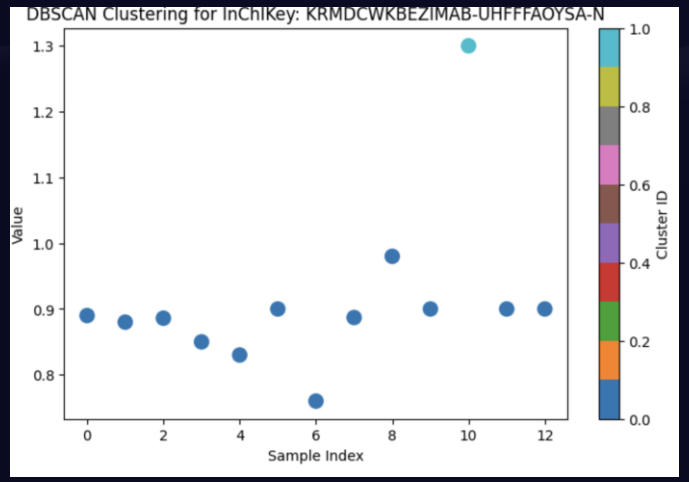
Kelich et al., in preparation, 2026

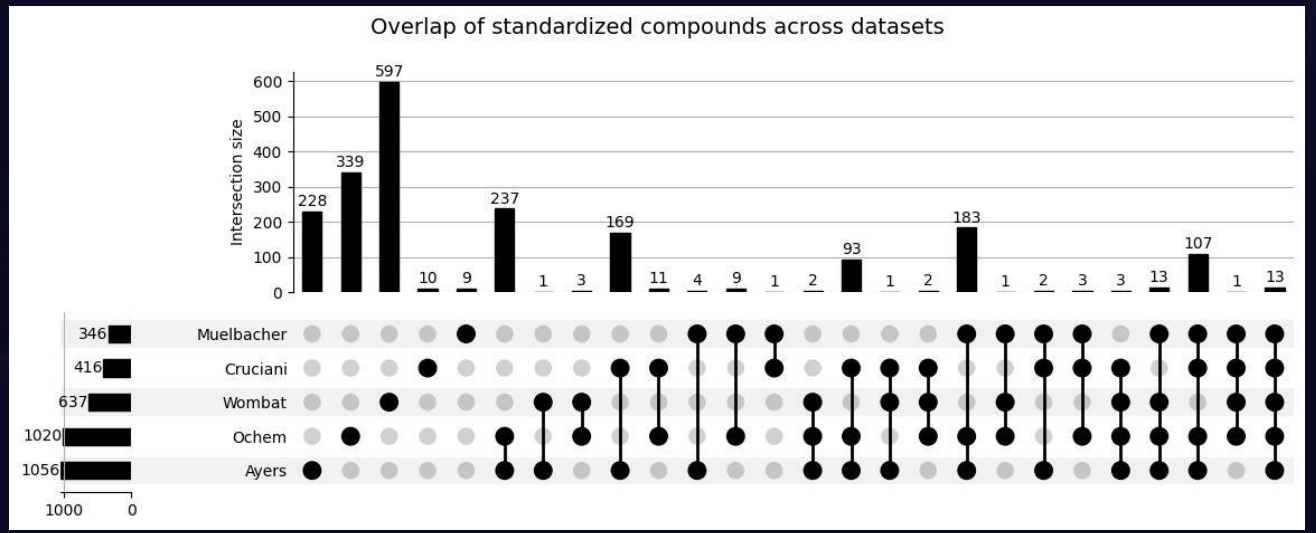
Focusing on logBB

Extensive Literature Curation and Consolidation

LogBB data from multiple sources:

- Muehlbacher, Cruciani, Wombat, OCHEM, Ayers & other smaller datasets from literature
- Used DBSCAN clustering on molecular descriptors to identify duplicates
- The centroid compound of each cluster was selected
- If only two compounds overlapped, the Cruciani dataset values were preferred, given higher curation quality
- Total of 2042 unique data points after deduplication.





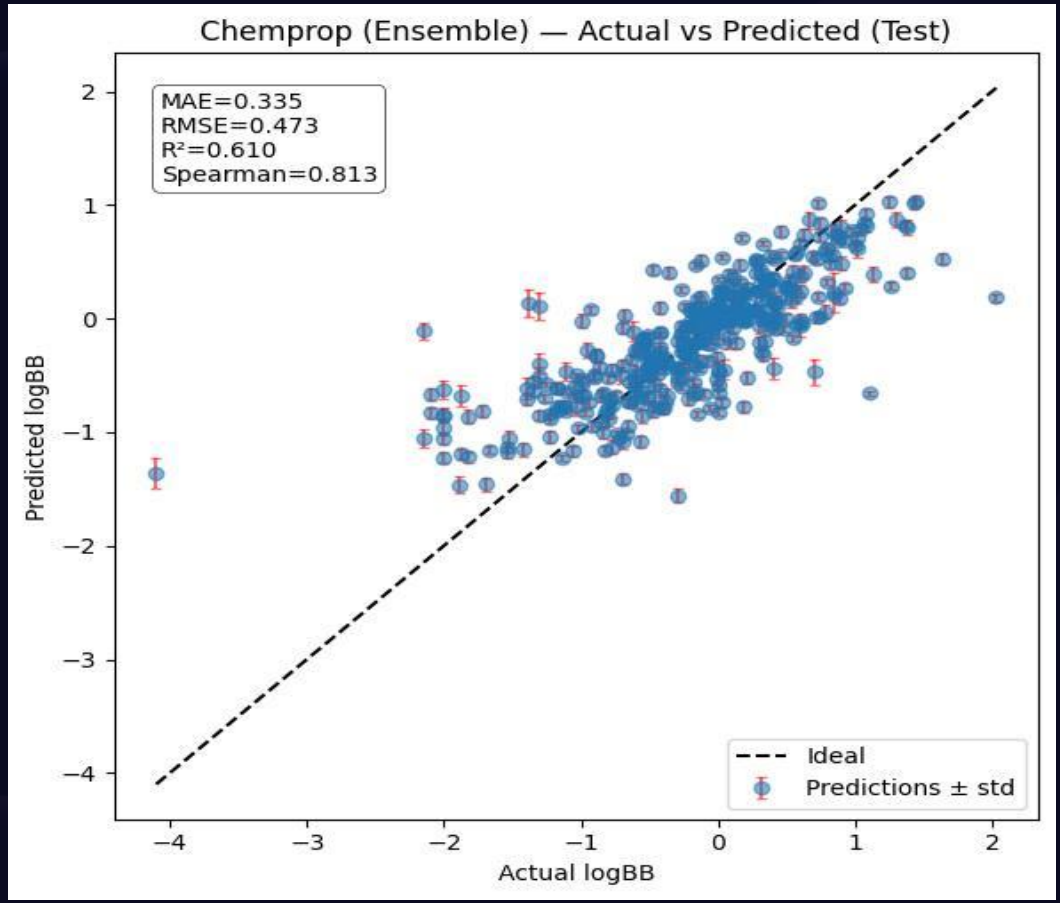
Kelich et al., in preparation, 2026

Focusing on logBB

ML Model Development

- Data size: 2042
- Data Split: 70:10:20 (train/val/test)
- ML Package: Chemprop v2.2.1
- Descriptors: Chemprop descriptors
- Ensemble of 5 models

- Stats on test set (409)
- R^2: 0.610
- RMSE: 0.473
- MAE: 0.335
- Spearman Rho: 0.813



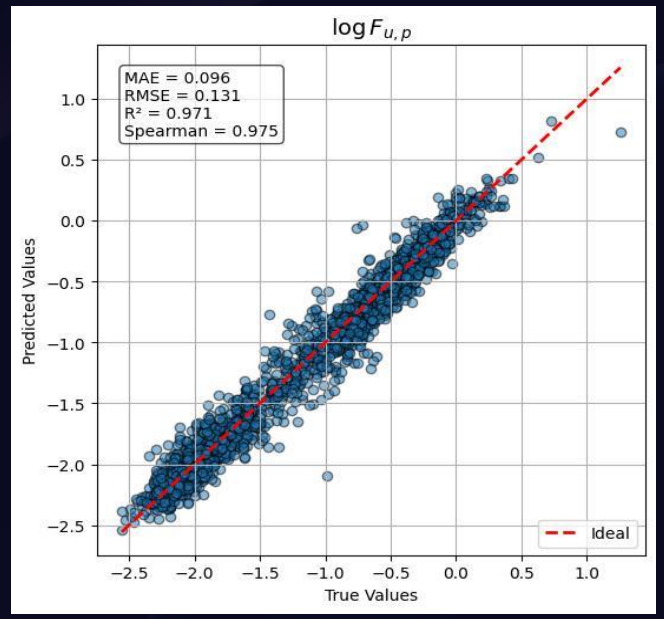
Kelich et al., in preparation, 2026

Focusing on Fraction Unbound, Plasma and Brain

ML Model Development

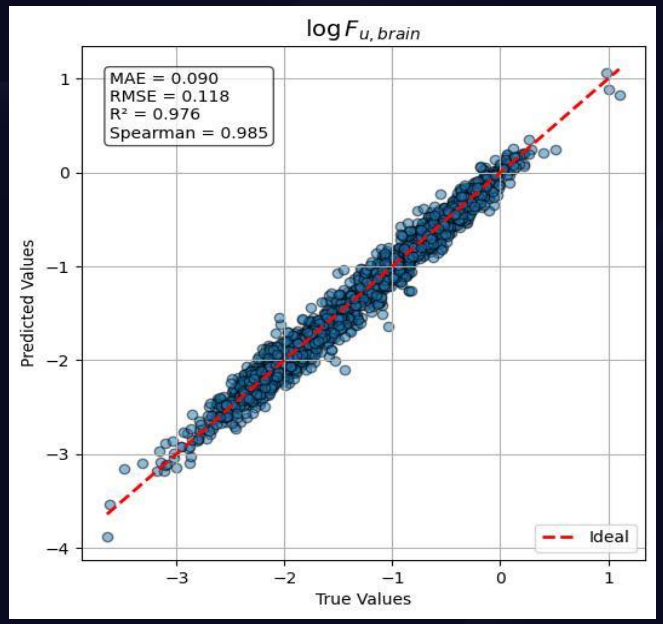
log $F_{u,p}$ (Human)

- Stats on test set (3157)
- R^2: 0.971
- RMSE: 0.131
- MAE: 0.096
- Spearman Rho: 0.975



log $F_{u,brain}$

- Stats on test set (3157)
- R^2: 0.976
- RMSE: 0.118
- MAE: 0.090
- Spearman Rho: 0.985



- Data size: over 31k
- Data Split: 80:10:10 (train/val/test)
- Task type: Multitask (log $F_{u,p}$ and log $F_{u,brain}$)
- ML Package: Chemprop v2.2.1
- Descriptors: Chemprop descriptors

Kelich et al., in preparation, 2026

eXpert sYstems

Estimating $K_{p,uu,brain}$ based on a 3-ML Model Combo and Some Thoughts on ML Model Deployment for NAMs

ExSys Model	Training Set Size	External Set UQ <0.15	Mean UQ
$LogF_{u,brain}$	>31,000	23,341	0.041
$LogF_{u,plasma}$	>31,000	23,384	0.038
$LogK_{p,brain}$	2,042	2,035	0.299

Our ChemProp-based ML models use a 70:10:20 or 80:10:10 split (training set / validation / test set), with five models that output uncertainty quantification (UQ) values.

Combined with the training set size (compounds), UQ enables model quality evaluation. From external predictions summarized above for a CNS brain-permeable set of 23,406 pre-plated compounds from ChemDiv, we derived $K_{p,uu,brain}$ estimates.

However, this highlights a critical factor in the deployment of ML-based NAMs: the interplay between training set size and UQ. The discrepancies, spanning one order of magnitude, illustrate how model robustness and predictive accuracy can be compromised when training data are unevenly distributed across properties.

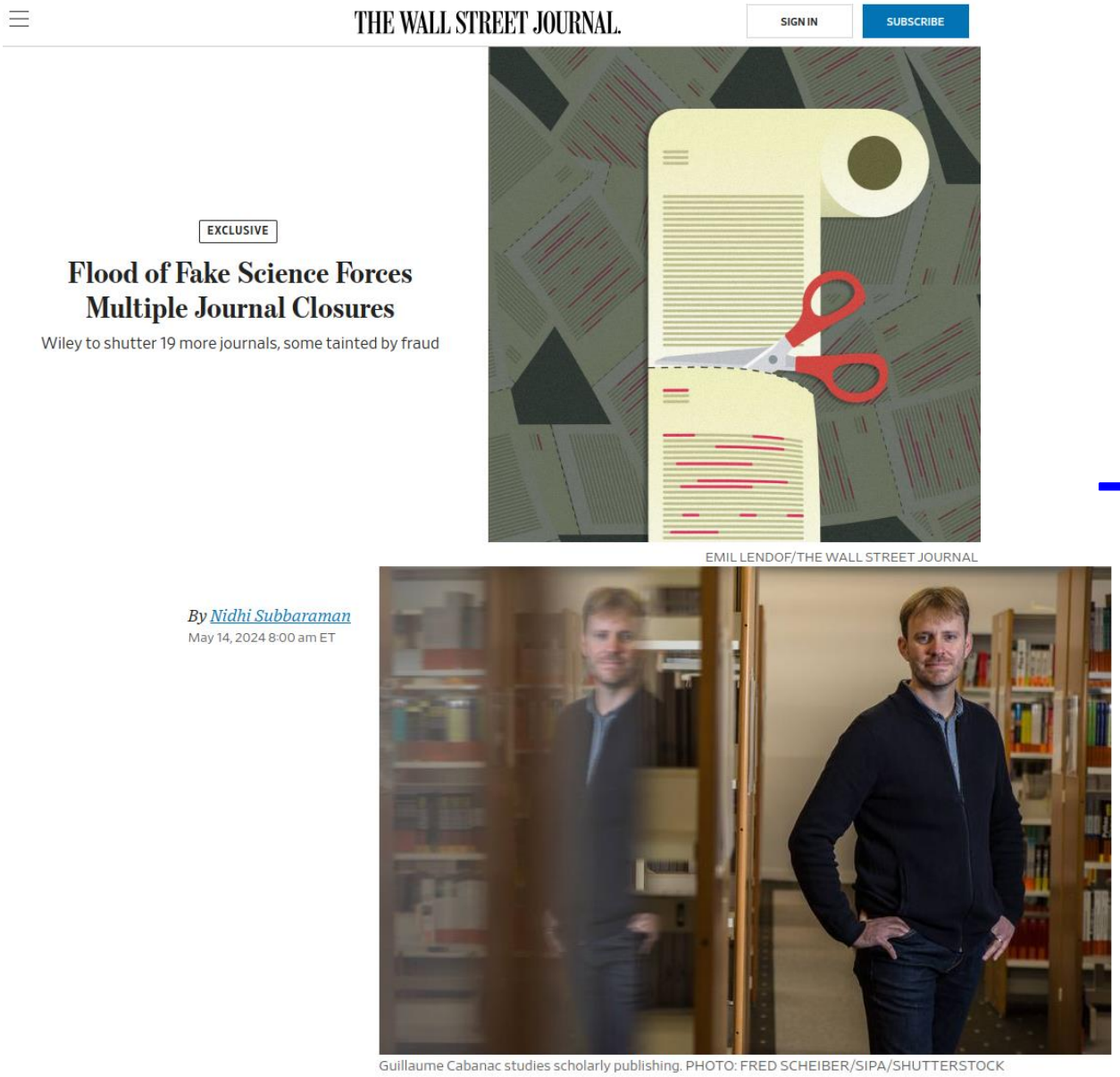
Kelich et al., in preparation, 2026

eXpert sYstems

Artificial Intelligence

Cautionary Remarks

The Power of AI relies on True Data



THE WALL STREET JOURNAL.

SIGN IN SUBSCRIBE

EXCLUSIVE

Flood of Fake Science Forces Multiple Journal Closures

Wiley to shutter 19 more journals, some tainted by fraud

EMIL LENDOF/THE WALL STREET JOURNAL

By Nidhi Subbaraman
May 14, 2024 8:00 am ET

Guillaume Cabanac studies scholarly publishing. PHOTO: FRED SCHEIBER/SIPA/SHUTTERSTOCK

… the “Problematic Paper Screener,” run by Guillaume Cabanac, […], scans the breadth of the published literature, some 130 million papers, looking for a range of red flags including “tortured phrases.” […]

researchers who wanted to avoid plagiarism detectors had swapped out key scientific terms for synonyms from automatic text generators, leading to comically misfit phrases. “Breast cancer” became “bosom peril”; “fluid dynamics” became “gooey stream”; “artificial intelligence” became “counterfeit consciousness.”

N. Subbaraman, *WSJ* **2024** link

What's an "AI drug"?

ELSEVIER

Drug Discovery Today

Volume 29, Issue 6, June 2024, 104009

Feature

How successful are AI-discovered drugs in clinical trials? A first analysis and emerging lessons

Madura KP Jayatunga [1], Margaret Ayers [1], Lotte Bruens [2], Dhruv Jayanth [3], Christoph Meier [1]

We conducted a first analysis of the clinical pipelines of AI-native Biotech companies.

In Phase I trials, AI-discovered molecules are substantially more successful than historic industry averages.

M.K.P. Jayatunga, *Drug Discov Today* **2024** link

HOME > COMMENTARY > BLOGS > IN THE PIPELINE > AI DRUGS SO FAR

IN THE PIPELINE | CLINICAL TRIALS

AI Drugs So Far

13 MAY 2024 · 2:37 PM ET · BY DEREK LOWE · 3 MIN READ · COMMENTS

[…] the authors have 24 AI-discovered targets, 22 AI-optimized small molecules, 4 antibodies, 6 vaccines, and 10 repurposed compounds. […] My first reaction is that the idea of twenty-four AI-discovered targets is really high. So I went looking through the SI table to see which ongoing projects fell into that category and what those compounds and targets were.

[…] these targets were already known to be implicated in the disease under investigation. […] there are several drugs already in the clinic targeting the same proteins, or even therapies that are already on the market working through the same mechanisms (C. diff toxin B, e.g.) […] I do not see how any of them can be classified as "target discovered by AI".

D. Lowe, *In the pipeline* **2024** link

AF3: the answer to “life, the universe and everything”

eXpert sYstems

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › articles › article

Article | Published: 08 May 2024

Accurate structure prediction of biomolecular interactions with AlphaFold 3

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, … John M. Jumper + Show authors

Nature (2024) | Cite this article

The new AlphaFold model demonstrates significantly improved accuracy over many previous specialised tools: far greater accuracy on protein-ligand interactions than state of the art docking tools, much higher accuracy on protein-nucleic acid interactions than nucleic-acid-specific predictors, and significantly higher antibody-antigen prediction accuracy than AlphaFold-Multimer v2.3.

J. Abramson, *Nature* **2024** link

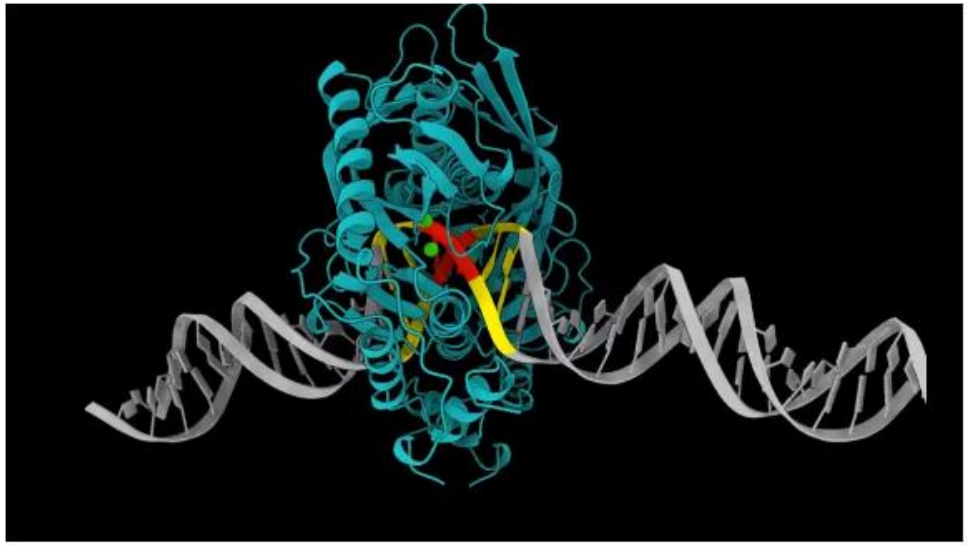


Image credit: Jan Kosinski, https://twitter.com/jankosinski/status/1789739354958504148

Hype Around AlphaFold 3: Results vs Limitations

Andrii Buvailo, Ph.D.
Advising on content marketing & strategy for life sciences • Helping tech-savvy brands tell compelling stories via data and market…
116 articles ✓ Following

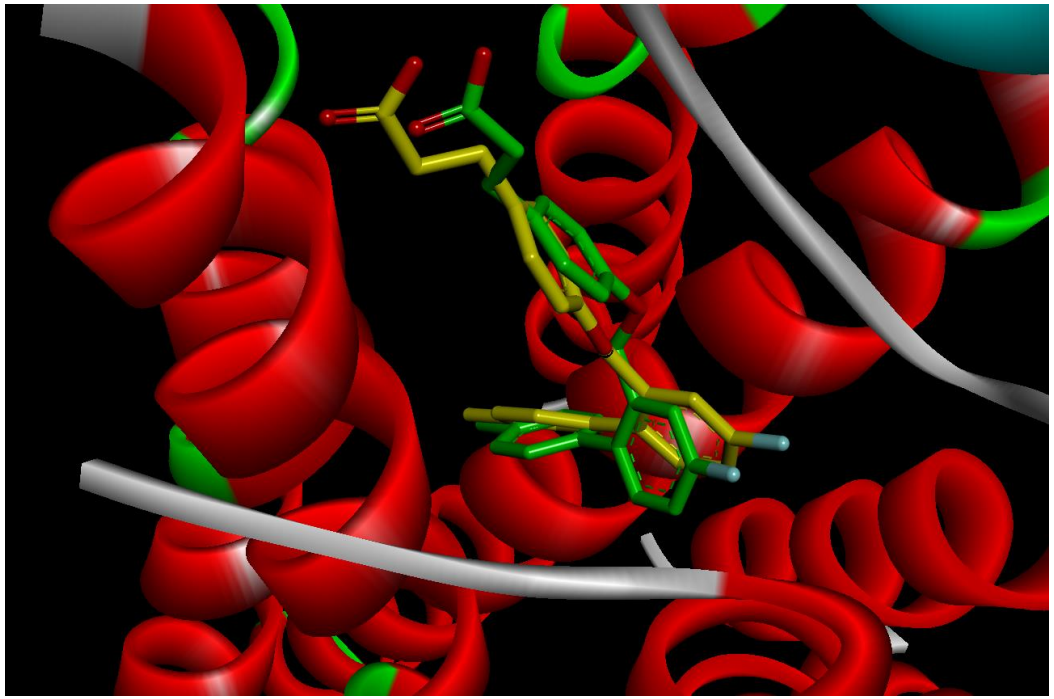
May 13, 2024

Limitations:
accuracy & reliability; risk of hallucinations;
no source-code (unlike AF2); usage restrictions

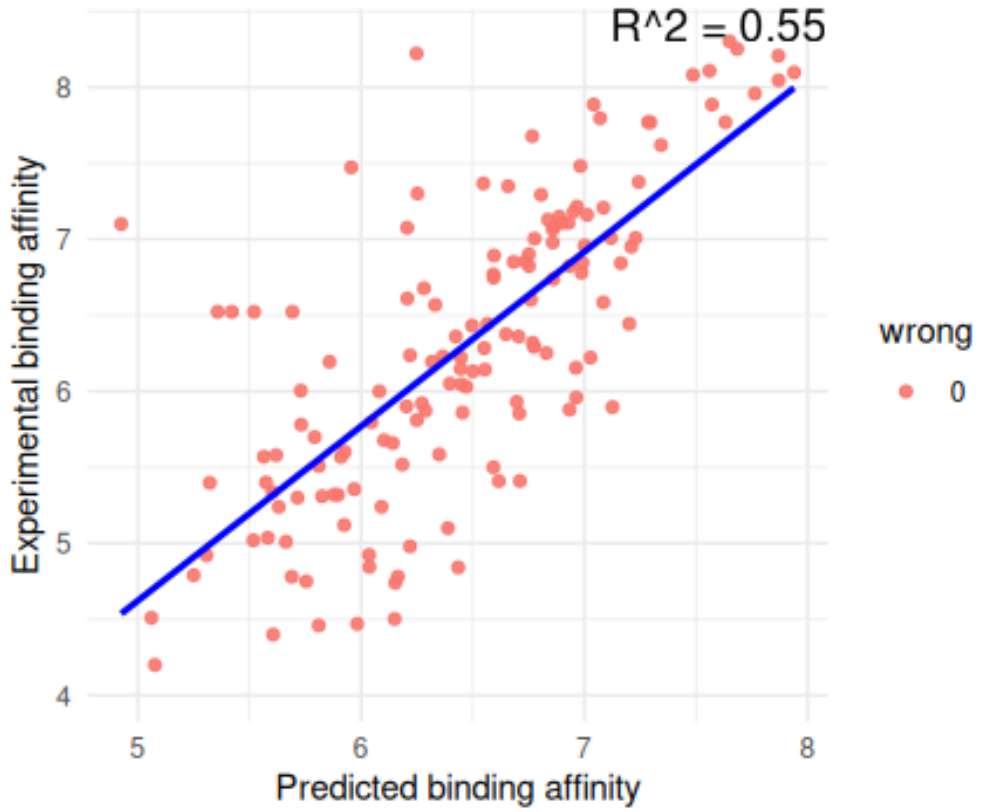
A. Buvailo, *LinkedIn blog* **2024** link

Boltz-2: Better than AF3?

The new open-source Boltz-2 model has an affinity module that predicts protein-ligand binding affinity at an accuracy level that approaches that of long and expensive FEP atomistic simulations while being more than 1000x faster.



Boltz-2 model of TUG891 in FFAR4



https://cbologa.shinyapps.io/ffar4/

It turns out, Boltz-2 learns from related protein-ligand data (e.g, FFAR1). What constitutes a “fair” evaluation?

https://boltz.bio/boltz2

AI4DD: Scientific issues

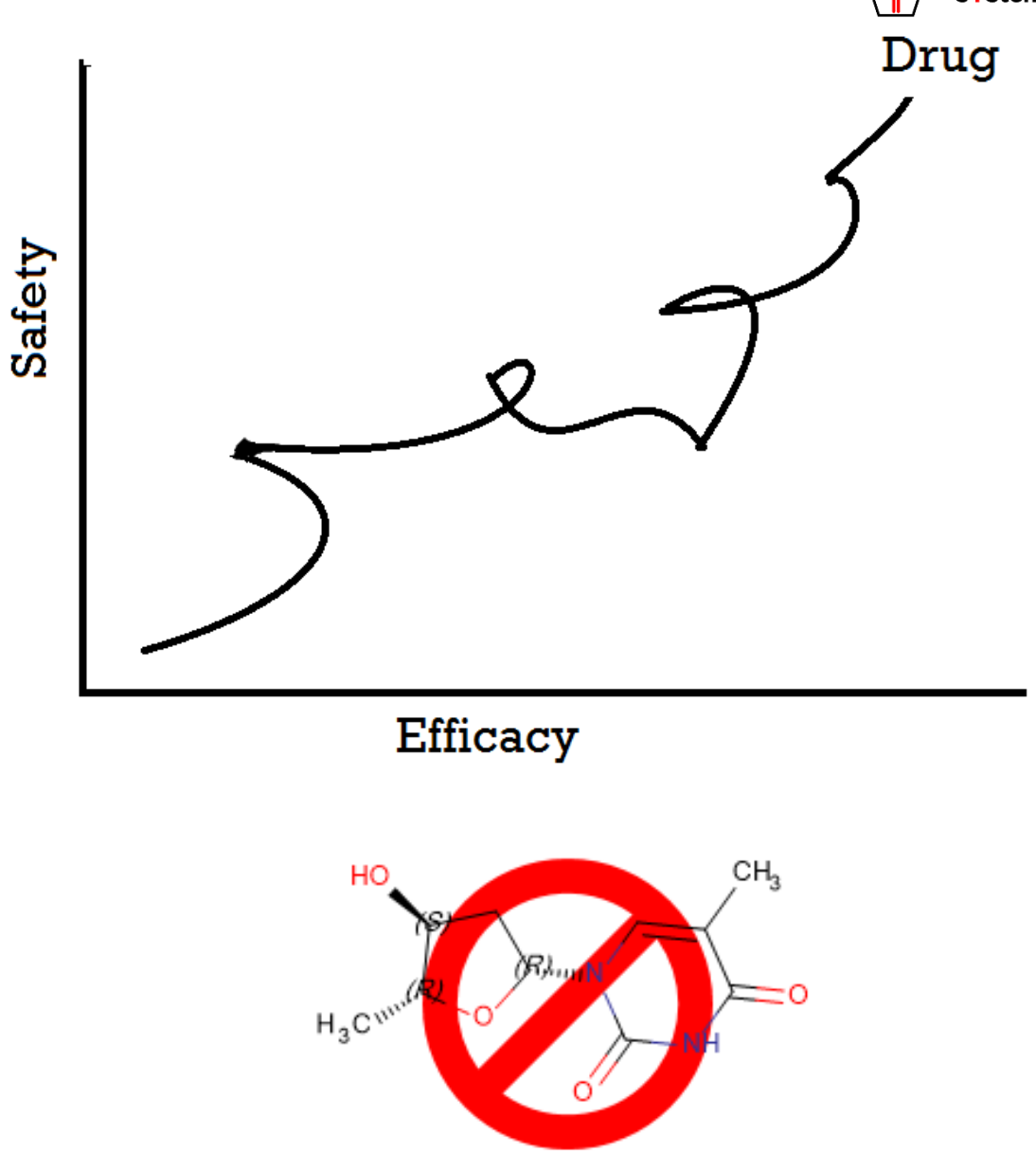
Can AI discover new knowledge? To date, no credible evidence of this has been provided. Chatbots, winning at chess, GO and *Jeopardy!* and LLM-generated poetry do not count.

We live in the world of *alternative facts* when it comes to research (not just politics). People lie. *As long as AI gets false data, we cannot provide what's needed.*

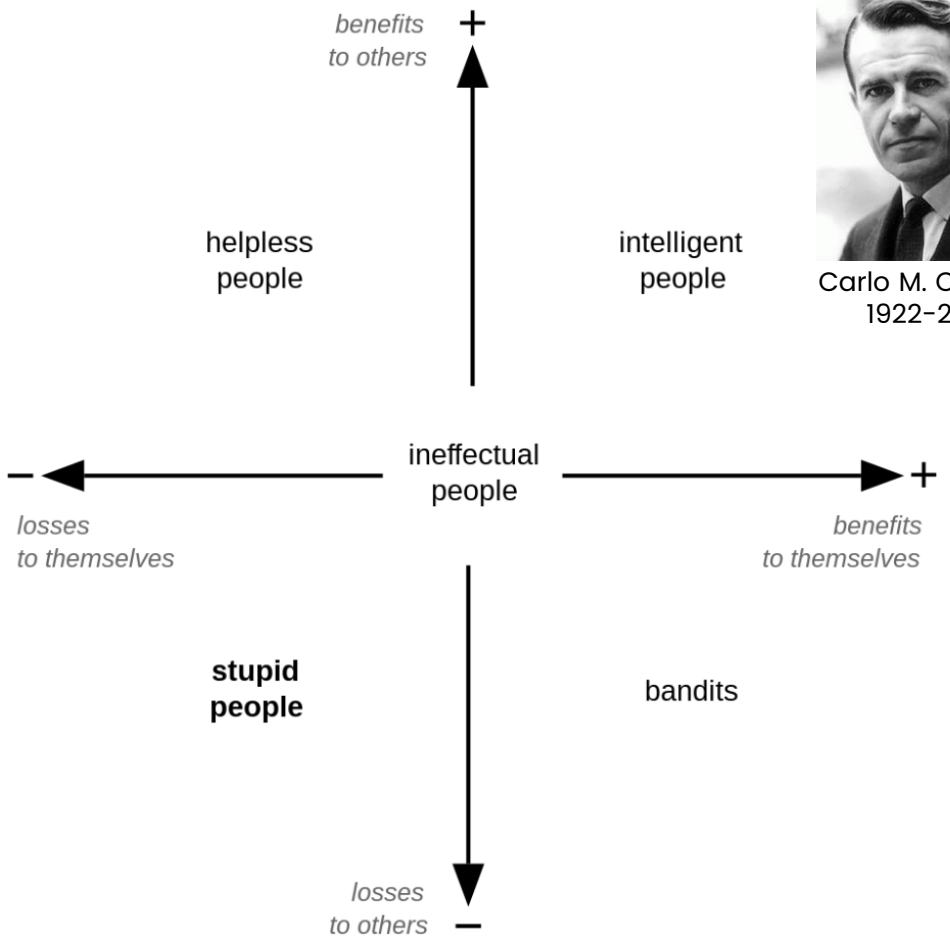
The unrealized promise relates to our (in)ability to explain the two pillars of clinical drug effectiveness: *Safety* and *Efficacy*.

Be it Drug Safety or Patient Safety, or indeed Clinical Efficacy, *we remain unable to model these processes as function of molecular structure.*

Andrii Buvailo keeps track of AI4DD



Le leggi fondamentali della stupidità umana



Carlo M. Cipolla
1922-2000

Intelligent people contribute to society and themselves; *Stupid* people are counterproductive to their interests and others'; *Bandits* pursue their own interest to the detriment of others; ~~Helpless~~ Generous people contribute to society but not themselves. *Ineffectual people* are in the center of the graph.

https://en.wikipedia.org/wiki/Carlo_M._Cipolla

It's not too hard to imagine a world where stupid or ineffectual people start using machine intelligence to the detriment of society and themselves

- AI-based facial recognition is biased towards white Caucasians (Buolamwini & Gebru, 2018 link)
- "deep learning models biased by race, sex, and age have been observed in medical domains" (link)
- Google's online advertising displayed high-paying jobs to men more often than to women (link)
- Google Gemini image generation depicted historical figures as people of colour (link)
- "AI Overview" suggested adding 1/8 cup of non-toxic glue in pizza sauce to avoid cheese sliding off pizza (link)

Jokes aside, Palantir is already leading the efforts to monetize war by developing AI-based weapons (link)

IMO, machine intelligence follows the intent of the humans who build it. If humans don't follow ethics, we can't expect AI to do so... and yes, at some point "singularity" may start to think independently.

The road to AI is paved with good intentions

1. The truth is already here … it's just not evenly distributed (adapted from William Gibson)

2. Unprecedented AI4DD technology push, but focus on algorithmic developments

3. (it's the data, stupid) true data >>> big data

4. People are part of the problem - "Retraction Watch" isn't retracting fast enough

5. We still don't have convincing evidence that "AI" can discover *new knowledge.* However, AI as inventor is subject to extensive debates

6. While hype remains rampant in the AI space, there is appetite for implementation. (as I type this in the United Polaris lounge (SFO), a robot drives around with dishes)

Artificial (medical) Intelligence

Human − Health + Disease = Patient

There is an explosion of “AI” applications; some are straight extensions of ML models, some are generative models (closer to “AI” than ML).
AIML is well suited to integrate context-specific computational reasoning tools with real-time (multi) –omics, biomarker / RWE, literature/patent & biomedical data.

Alexahealth™: Given my current health status, medication and supplements, as well as my calorie budget, what food should I shop/prepare today? What exercises do you recommend?

The Revolution Will not be Televised*

*...but AGI might be watching

What Could the Future of Biotech Look Like?

In a purely Sci-Fi scenario, the future of biotech is **nano-bots** that work 24/7 to keep us healthy. Powered by infrared (an abundant source of energy in mammals), such nanobots could repair injured tissues (by secreting the right growth factors to cause apoptosis and healing OR by re-calibrating metabolic imbalances).

Cancer, neurodegeneration, cardiac disease, aging - could all be kept at bay by intelligent machines that use the body's resources to maintain a functional homeostasis.

This would take “personalized medicine” to another level

We don't know how to build such nanobots. The closest we have are robotic implants (some with direct brain interface) and the Car-T technology (cell therapy)

Why is Biotech Not Going in that Direction?

Societal factors and the separation of sciences into subdomains is causing humans to miss the “big picture”. With millions of new scientific publications added annually in PubMed (not to mention patents, conferences and other scientific dissemination channels), the human mind is limited in its ability to process and comprehend the pace of discovery.

Add to this the “deceitful enterprises” such as irreproducible papers, intentionally fabricated results (e.g., papermills, predatory journals etc.) and the constant reshuffle of science (facts and knowledge have a *limited* shelf life -- think Einstein vs. Newton or even 10 adrenergic receptors vs. Ehrlich & Ahlquist) and you have a picture of disjoint science, disrupted occasionally by “scientific revolutions”.

The only (positive) way foward is to work together, and if we build “benevolent AI” tools.

What Will Biotech Look Like in Five Years?

New therapeutic modalities are emerging. From bifunctional antibodies, ADCs and antisense technologies, to gene replacement therapies and gene editing tools, we are on a fast track to replace “small molecule drugs” as the main therapeutic intervention {besides surgery & psychotherapy}.

As William Gibson wrote in an interview post-*Neuromancer*, *the future is already here - it's just not evenly distributed.*

The next five years will see the continuing trend of fewer SMDs, with more biologics, gene therapies and perhaps CRISPR-based interventions approved.

But don't dismiss SMDs yet: From PROTACs to molecular glues and dual-targeted SMDs, as well as new drug combinations, there will continue to be innovative SMD applications.

Innovation in the Next Five Years

The near future of biomedical research is hybrid (human/artificial intelligence) technologies, where AI agents follow general directions outlined by humans.

Robots build our cars, and have done so for some time. In the next 5 years, one or more companies will launch fully-automated robotic drug discovery laboratories.

AstraZeneca, Roche, In Silico Medicine, GSK - unclear who will "get there" first.

Generative AI technologies, combined with LLMs (or perhaps "local language models" that are pre-trained on small tasks only) will ultimately conquer the "multi-property optimization" summit. We are on the cusp of a radical change in the way we conduct scientific research.

Follow BioPharmaTrend to keep up with the news (link).

Thank You

toprea_at_expertsystems.inc