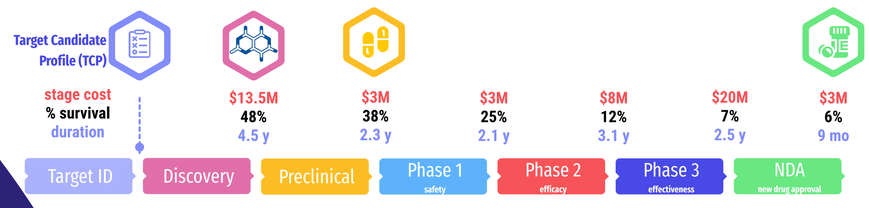
FOUNDATION SIMULATION MODELS AND TRANSFORMATIVE EXPERIMENTAL DATASETS FOR DRUG DISCOVERY

John D. Chodera | Achira / Memorial Sloan Kettering Cancer Center

Achira http://achira.ai

Memorial Sloan Kettering Cancer Center http://choderalab.org

Small molecule drug discovery is slow, costly, and inefficient



>94% of molecules synthesized fail to meet ADMET objectives [1]

>20% preclinical attrition due to ADMET issues in preclinical species (rodent, dog, etc.) [2]

30% of clinical development failures due to human safety and toxicology [3]

[1] doi:10.1016/j.drudis.2015.03.010
[2] doi:10.1016/j.drudis.2013.11.014
[3] doi:10.1038/nrd1470

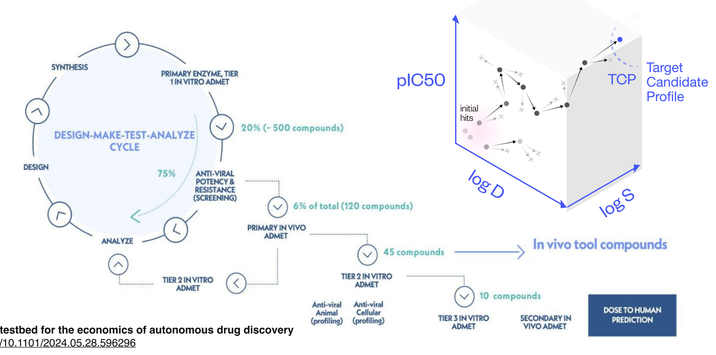
ADMET
Absorption
Distribution
Metabolism
Excretion
Toxicity

A Target Candidate Profile (TCP) defines the multi-objective optimization problem for a drug discovery program

Target Candidate Profile (TCP) for oral SARS-CoV-2 main viral protease (Mpro) inhibitor

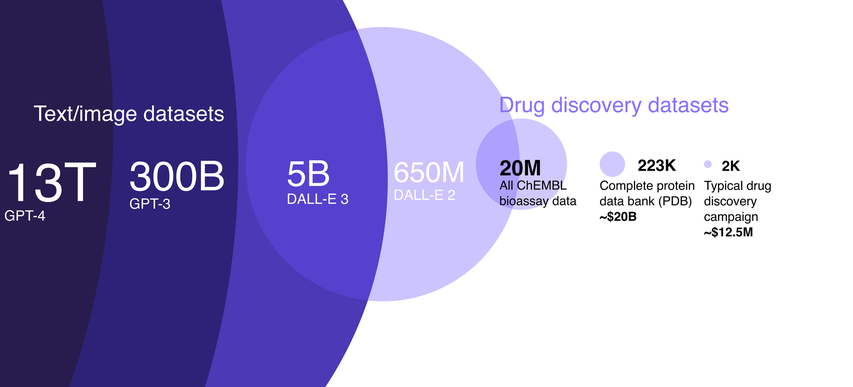
Property	Target range	Rationale
protease assay	IC_{50} < 10 nM	Extrapolation from other anti-viral programs
viral replication assay	EC_{50} < 5 μM	Suppression of virus at achievable blood levels
plaque reduction assay	EC_{50} < 5 μM	Suppression of virus at achievable blood levels
route of administration	oral	bid/tid - compromise PK for potency if pharmacodynamic effect achieved
solubility	> 5 mg/mL	Aim for biopharmaceutical class 1 assuming <= 750 mg dose
half-life	> 8 h (human) est from rat and dog	Assume PK/PD requires continuous cover over plaque inhibition for 24 h max bid dosing
safety	Only reversible and monitorable toxicities No significant DDI - clean in 5 CYP450 isoforms hERG and NaV1.5 IC_{50} > 50 μM No significant change in QTc Ames negative No mutagenicity or teratogenicity risk	No significant toxicological delays to development DDI aims to deal with co-morbidities / therapies, cardiac safety for COVID-19 risk profile cardiac safety for COVID-19 risk profile Low carcinogenicity risk reduces delays in manufacturing Patient group will include significant proportion of women of childbearing age

Most time is spent in design-make-test-analyze (DMTA) cycles

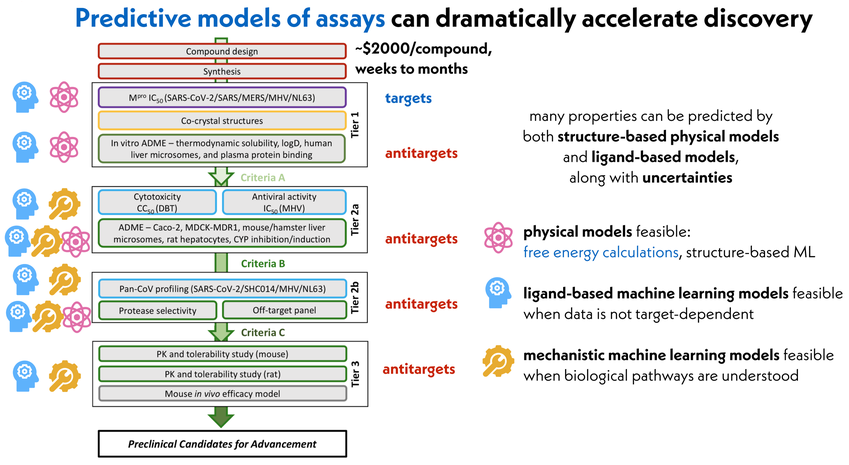


DrugGym: A testbed for the economics of autonomous drug discovery
https://doi.org/10.1101/2024.05.28.596296

There is not enough data for generative modeling to replace DMTA cycles



Predictive models of assays can dramatically accelerate discovery

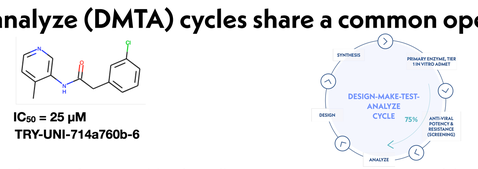


many properties can be predicted by both **structure-based physical models** and **ligand-based models**, along with **uncertainties**

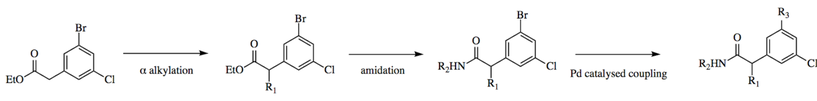
- **physical models** feasible: free energy calculations, structure-based ML
- **ligand-based machine learning models** feasible when data is not target-dependent
- **mechanistic machine learning models** feasible when biological pathways are understood

Design-make-test-analyze (DMTA) cycles share a common operation:

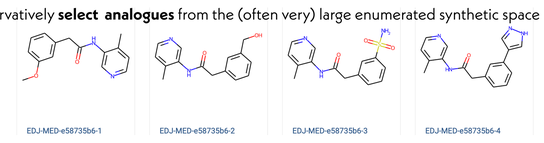
1. Select a current **lead molecule**



2. Use AI tools to identify a **retrosynthetic pathway** capable of installing new groups to replace part of the molecule



3. Chemists conservatively **select analogues** from the (often very) large enumerated synthetic space



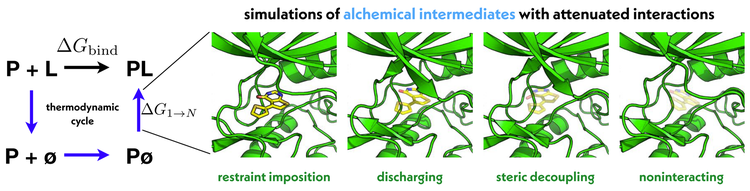
~15,000 Potential R3 groups

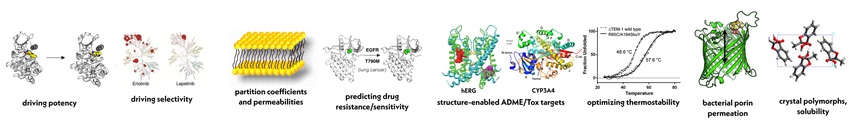
Enamine

Foundation simulation models can transform drug discovery by delivering accuracy comparable to Tier 1 assays

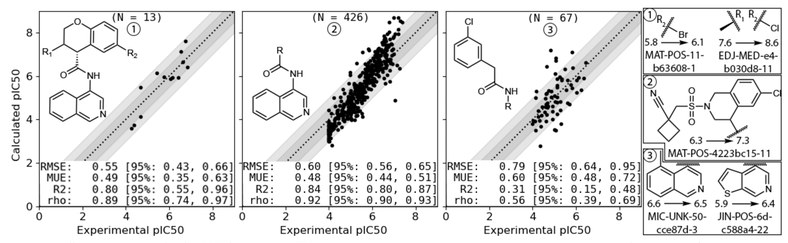
Alchemical free energy calculations leverage protein structure to predict critical numerous drug discovery assays

simulations of alchemical intermediates with attenuated interactions

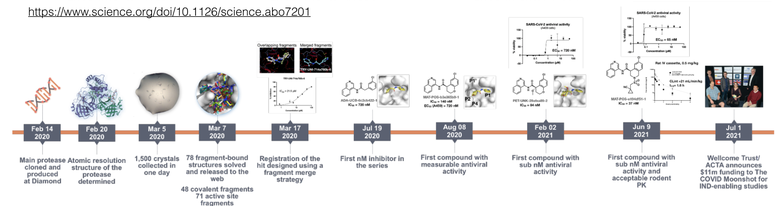




Alchemical calculations have successfully driven numerous drug discovery programs, including the discovery of a new pan-CoV antiviral

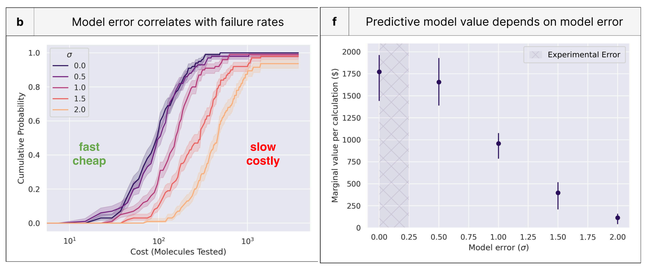


https://www.science.org/doi/10.1126/science.abo7201



Approaching Tier 1 accuracy error would shift huge amounts of $250B annual CRO spending from chemistry to compute

Modeling hit-to-lead phase of drug discovery



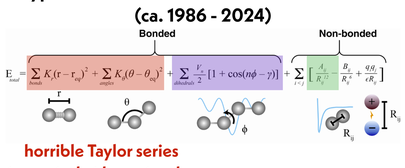
As free energy calculation **inaccuracy decreases**, their **value increases** **superlinearly**

DrugGym: A testbed for the economics of autonomous drug discovery
https://www.biorxiv.org/content/10.1101/2024.05.28.596296v1

Current generation simulation models were developed for 45-year-old technology—late 1970s-era *minicomputers!*

DEC PDP-11
1980
50 KFLOPs for $160,000*
* In today's dollars

typical class I molecular mechanics force field (ca. 1986 - 2024)



horrible Taylor series truncated at lowest order

crappy Fourier series truncated at n=6

don't even get me started on this one

Durrant, McCammon. Molecular dynamics simulations and drug discovery. BMC Biology, 2011

We have **real computers now.**

2025
100 TFLOPS
for $2,000

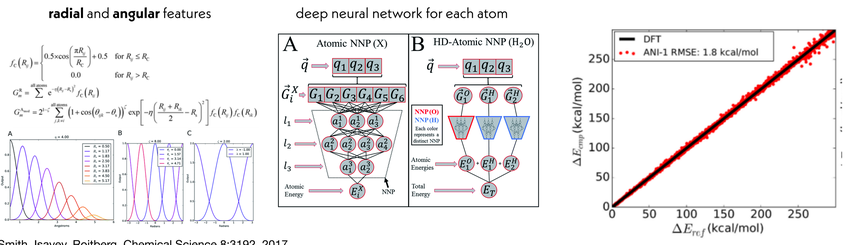
That's **160 billion times more FLOPS/$** than when these functional forms were introduced.

Why are we still using computers the same way as in 1980?

A new generation of geometric deep learning models offers near quantum chemical accuracy at achievable compute cost

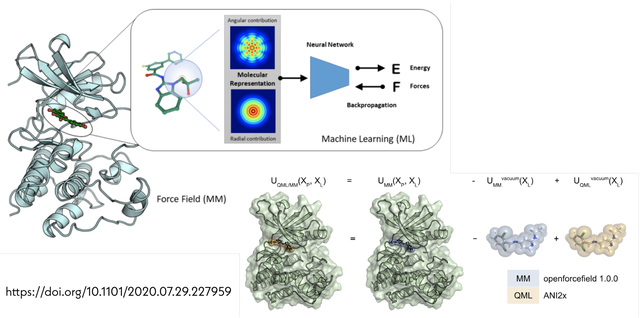
ANI family of quantum machine learning potentials

radial and **angular** features — deep neural network for each atom



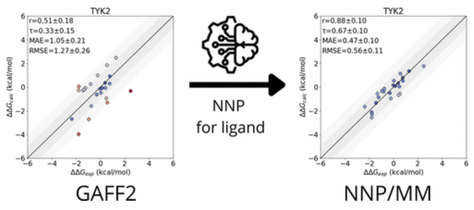
Smith, Isayev, Roitberg. Chemical Science 8:3192, 2017.
http://doi.org/10.1039/c6sc05720a

These models can be trained to huge amounts of quantum chemical data to deliver significantly greater accuracy...



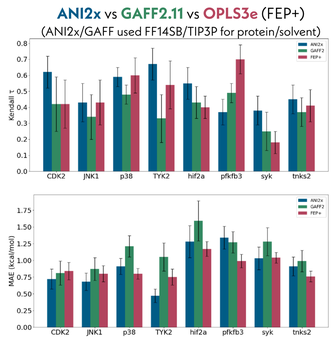
https://doi.org/10.1101/2020.07.29.227959

Even introducing NNP foundation models for just the ligand can increase accuracy and utility!

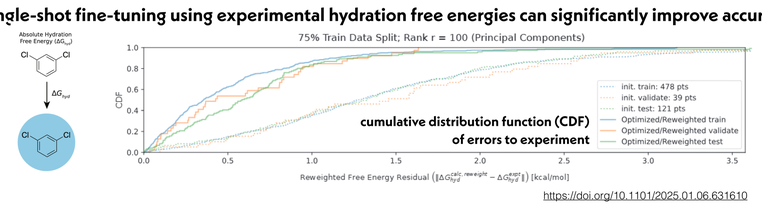


https://doi.org/10.1021/acs.jcim.3c02031

ANI2x vs GAFF2.11 vs OPLS3e (FEP+)
(ANI2x/GAFF used FF14SB/TIP3P for protein/solvent)

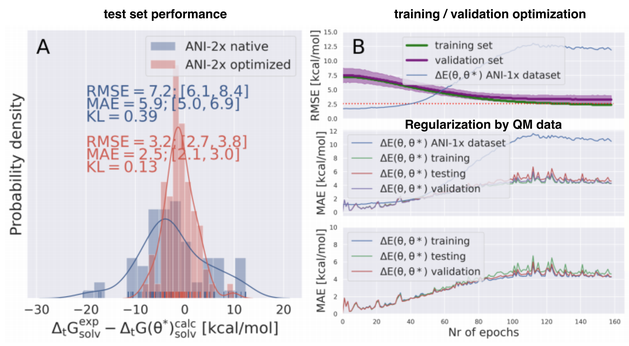


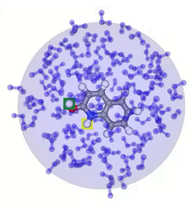
Single-shot fine-tuning using experimental hydration free energies can significantly improve accuracy



https://doi.org/10.1101/2025.01.06.631610

...and fine-tuned with experimental data (though at great compute cost) to produce foundation simulation models accurate for many tasks

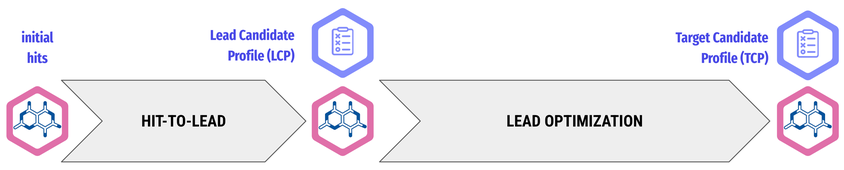




preprint: https://doi.org/10.1101/2020.10.24.353318
code: https://github.com/choderalab/neutromeratio

Fast on-the-fly reweighting enables inexpensive loss/gradient computation without repeating expensive free energy calculation

Nobody likes the current drug discovery paradigm

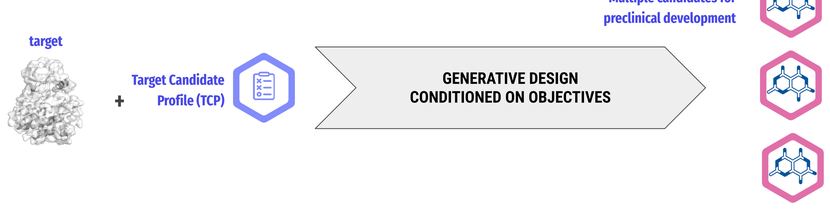


currently ~3.5 years, ~2000 compounds, $12.5M, ~50% success rate [1]

SIMULATE → EMULATE → GENERATE

- **SIMULATE**: build simulation models that accurately model biomolecular interactions
- **EMULATE**: build surrogate models that rapidly predict biomolecular interactions
- **GENERATE**: build generative ML models that predict molecules conditioned on TCP goals

With large synthetic datasets, we could invert the drug discovery paradigm



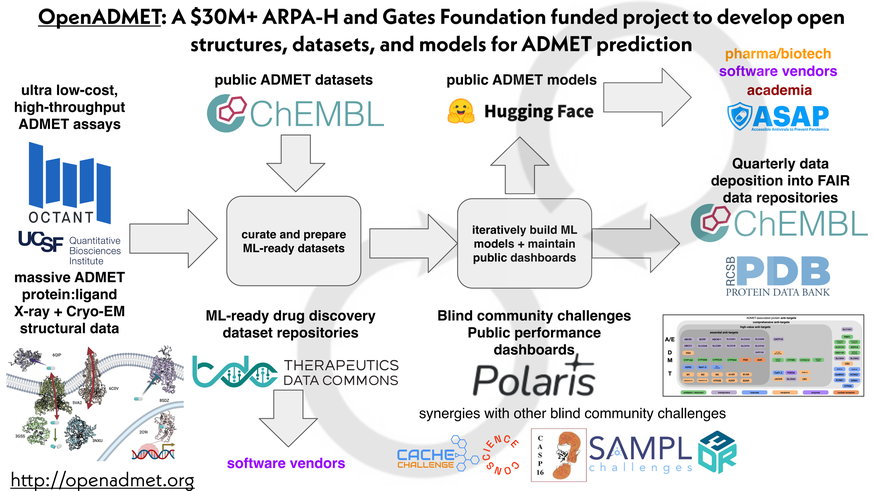
<< 1 year, ~ 10 molecules synthesized by CRO FTE chemists

Open experimental datasets are poised to transform drug discovery and enable open foundation simulation models

The **Open Molecular Software Foundation (OMSF)** is building a sustainable open source ecosystem serving the drug discovery community
omsf.io

- **Open Force Field :** open data, toolkits, and models for biomolecular simulation — http://openforcefield.org
- **Open Free Energy :** open tools and workflows for accurate affinity prediction for drug discovery — http://openfree.energy
- **OpenFold :** open, innovative protein and protein:ligand structure prediction tools — http://openfold.io
- **Rosetta Commons :** open tools for protein design and structural modeling — https://rosettacommons.org
- **OpenADMET :** open data, models, and blind challenges for ADMET prediction — http://openadmet.org

OpenADMET: A $30M+ ARPA-H and Gates Foundation funded project to develop open structures, datasets, and models for ADMET prediction

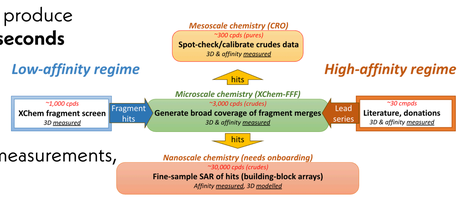


http://openadmet.org

OpenBind: What if we could double the PDB in 18 months? And the scale up two orders of magnitude in the next 2-3 years?

{**Alpha|Open|Chai|Boltz**}**Fold** promise to predict protein:ligand structures and affinities (through physical or ML models), but are **all inherently limited by training dataset size.**

- only ~220K structures in PDB
- only ~20K useful protein:small molecule structures
- public affinity datasets are poorly aligned to structures and a total disaster in consistency

Diamond Light Source K04 X-ray beamline can produce **one protein:ligand structure dataset every 8 seconds**
This is up to ~50K structures/week!

OpenBind aims to fill this gap:
Double the PDB in 18 months.
Structures will be paired with μL-scale affinity measurements,
nL-scale detailed SAR affinity measurements,
and mesoscale quality control measurements.
Scale by orders of magnitude over subsequent 2-3 years.