slides: http://choderalab.org/news

BLENDING PHYSICS WITH MACHINE LEARNING TO POWER THE FUTURE OF DRUG DISCOVERY

John D. Chodera
MSKCC Computational and Systems Biology Program
http://choderalab.org

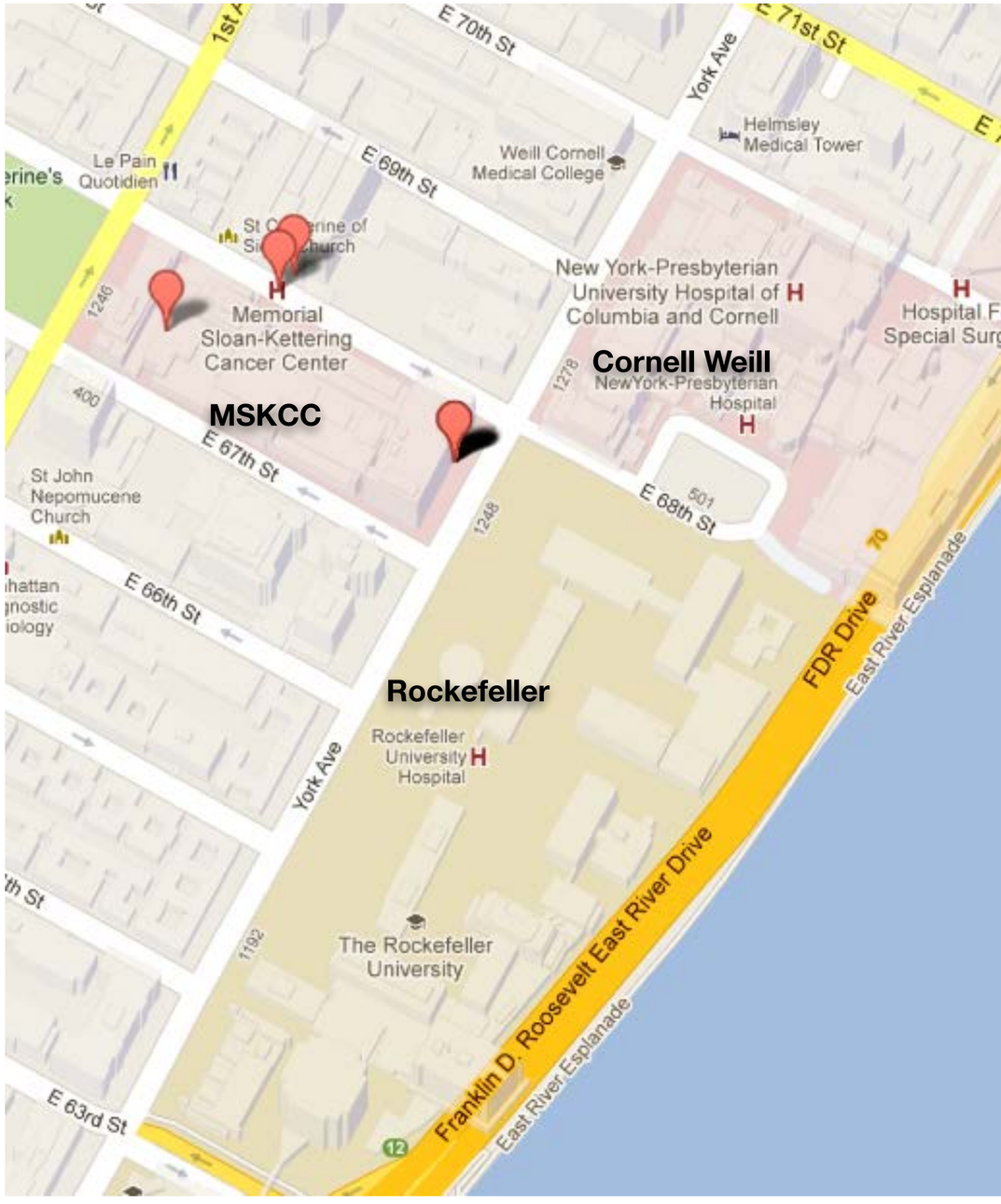
All disclosures: http://choderalab.org/disclosures
All funding sources: http://choderalab.org/funding

29 Oct 2024 - ICBS 2024 - Toronto, Canada

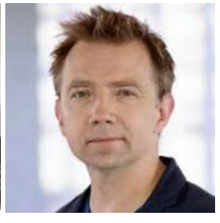
Memorial Sloan Kettering Cancer Center

Sloan-Kettering Institute

In more than 100 laboratories, our scientists are conducting innovative research to advance understanding in the biological sciences and improve human health.




Dana Pe'er

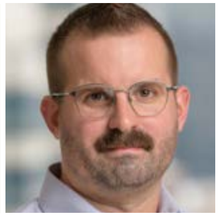

Quaid Morris


Christina Leslie


Joao Xavier


Kushal Dey

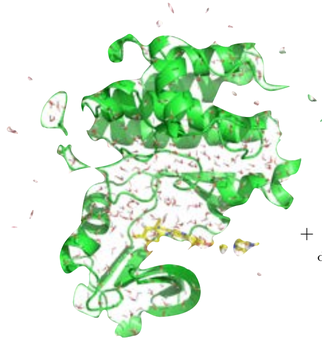

John Chodera


Thomas Norman

Computational and Systems Biology

THE CHODERA LAB INTEGRATES COMPUTATION AND EXPERIMENT TO ADVANCE DRUG DISCOVERY

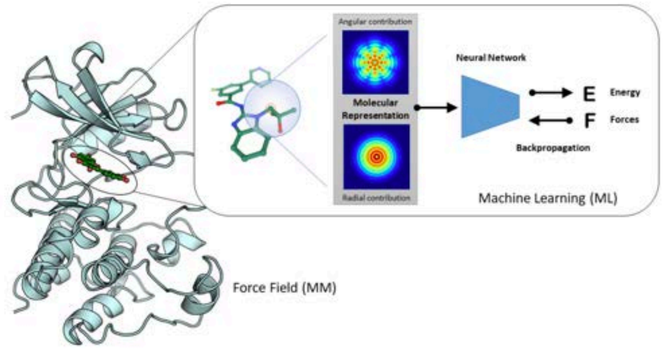
MODELING



$$V(\mathbf{q}) = \sum_{\text{bonds}} K_r(r - r_{eq})^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_{eq})^2$$
$$+ \sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}}\right]$$

FOLDING @HOME

amazon web services | EC2



AUTOMATION

CHODERA LAB, Z17

CLOUD LABORATORIES (STRATEOS)

OPENTRONS

WE COLLABORATE BROADLY TO IMPROVE THE DRUG DISCOVERY ECOSYSTEM

open forcefield consortium

Open Molecular Software Foundation

MolSSI
Molecular Sciences Software Institute

Folding@home

BAYER

INTERLINE therapeutics

Ventus THERAPEUTICS

SILICON Therapeutics

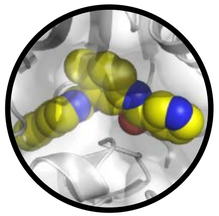
OpenEye SCIENTIFIC

VIR

open source software development initiatives

NIH
National Center for Advancing Translational Sciences

industry collaborations



choderalab
(algorithms and open source software)

data generators, community challenges, and resources

ASAP
Accessible Antivirals to Prevent Pandemics

The SAMPL challenges
Challenge information and announcements

Drug Design Data Resource

academia

XCHEM Diamond Fragment Screening

Diamond Light Source / XChem

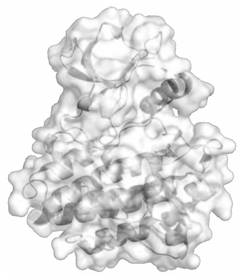
OpenADMET

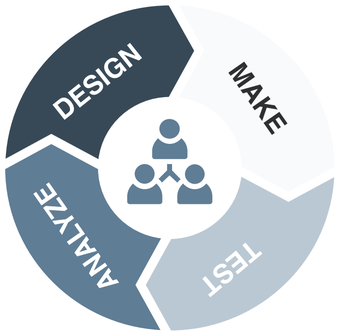
IP-generating collaborations

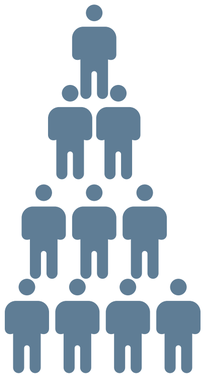
open science / open source software

DRUG DISCOVERY AND DEVELOPMENT IS SLOW, COSTLY, AND PRONE TO FAILURE

	Target Candidate Profile (TCP)	Target ID	Discovery	Preclinical	Phase 1 safety	Phase 2 efficacy	Phase 3 effectiveness	NDA new drug approval
stage cost		$13.5M	$3M	$3M	$8M	$20M	$3M
% survival		48%	38%	25%	12%	7%	6%
duration		4.5 y	2.3 y	2.1 y	3.1 y	2.5 y	9 mo

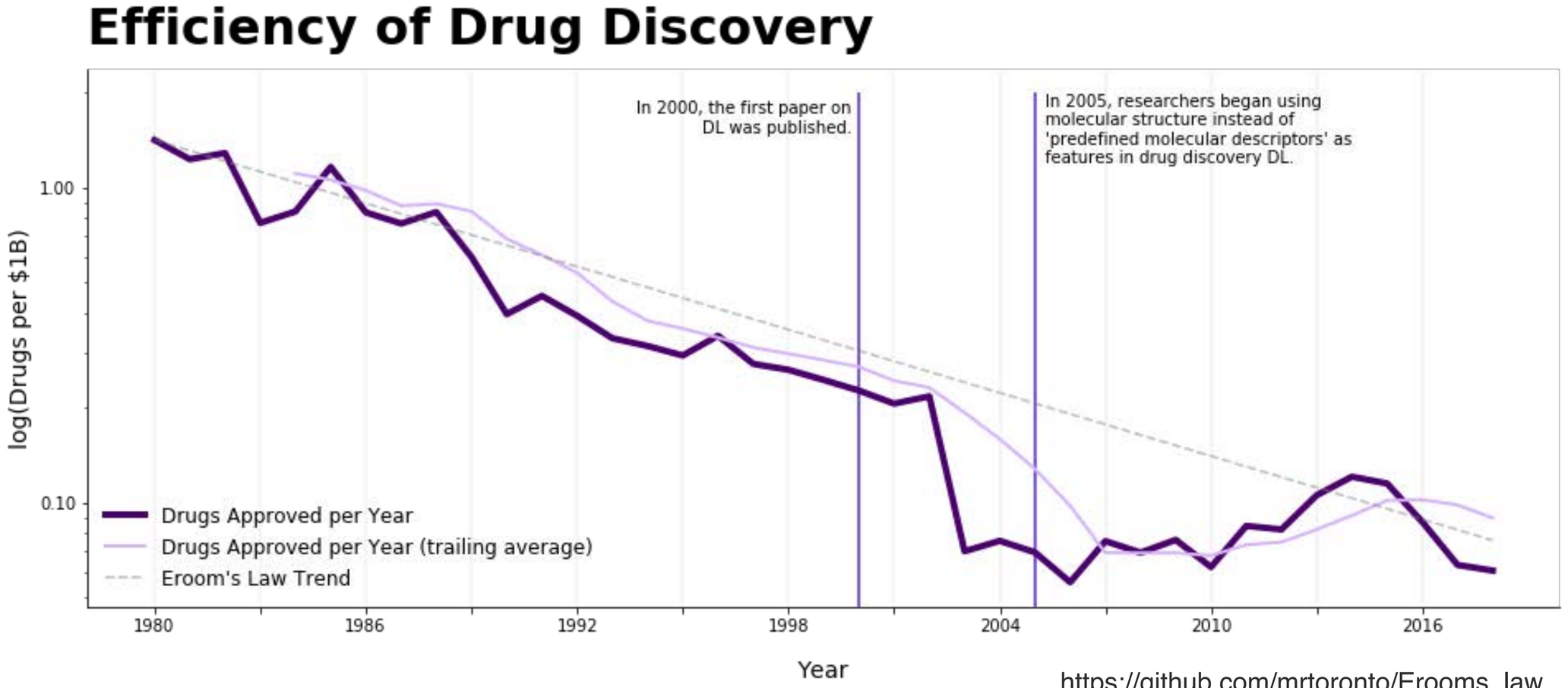






sources: [1] [2] [3] [4] [5]

AND IT'S GETTING WORSE, NOT BETTER



https://github.com/mrtoronto/Erooms_law

cost per approved drug (including costs of failures) estimated > $5B per approved drug in 2016 [1]

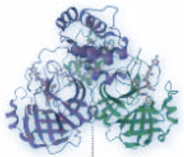
Why is this?

DRUG DISCOVERY IS REALLY HARD.

THE OPEN SCIENCE COVID MOONSHOT PRODUCED A NOVEL ORAL ANTIVIRAL FROM A FRAGMENT SCREEN IN JUST 18 MONTHS

Feb 14 2020

Main protease cloned and produced at Diamond

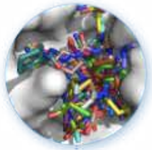


Feb 20 2020

Atomic resolution structure of the protease determined

Mar 5 2020

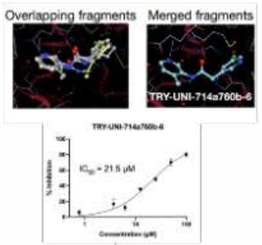
1,500 crystals collected in one day



Mar 7 2020

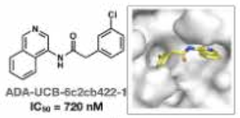
78 fragment-bound structures solved and released to the web

48 covalent fragments
71 active site fragments



Mar 17 2020

Registration of the hit designed using a fragment merge strategy

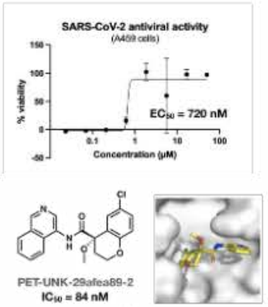


Jul 19 2020

First nM inhibitor in the series

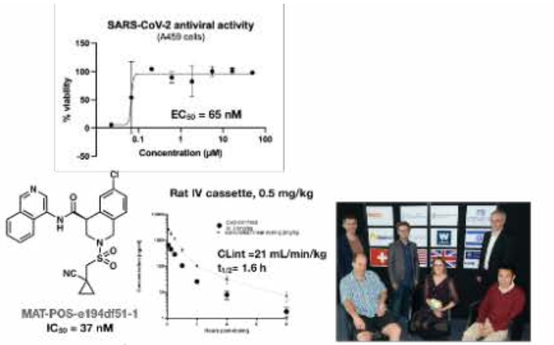
Aug 08 2020

First compound with measurable antiviral activity



Feb 02 2021

First compound with sub nM antiviral activity

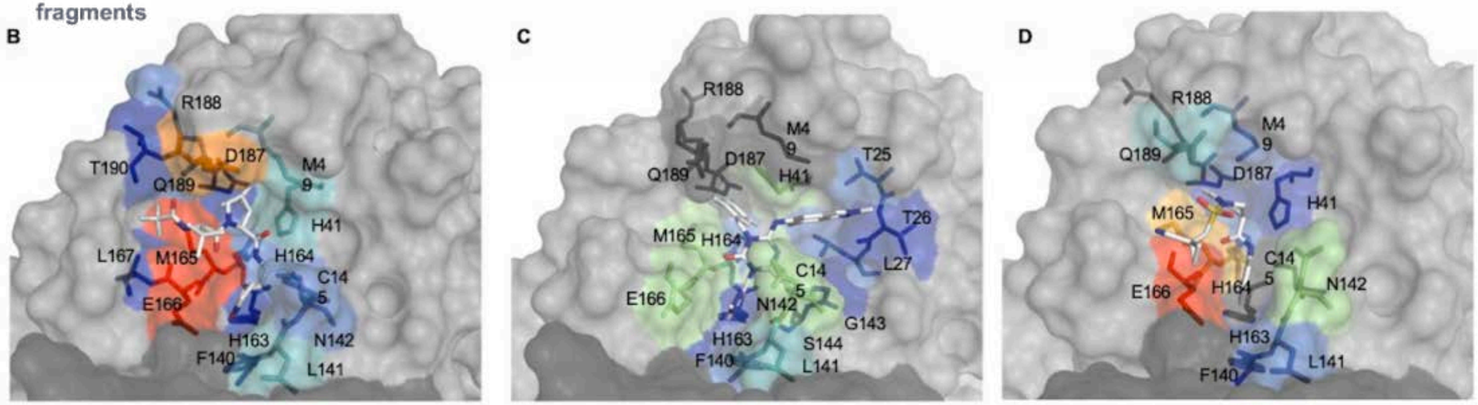


Jun 9 2021

First compound with sub nM antiviral activity and acceptable rodent PK

Jul 1 2021

Wellcome Trust/ ACTA announces $11m funding to The COVID Moonshot for IND-enabling studies



nirmatrelvir (Paxlovid)
Pfizer

S-217622
Shionogi
(modeled)

MAT-POS-e194df51-1
COVID Moonshot

COVID Moonshot structures and data: http://postera.ai/covid
paper: https://www.science.org/doi/10.1126/science.abo7201
history: https://www.nature.com/articles/d41586-021-01571-1

DNDi Drugs for Neglected Diseases *initiative*

COVID Moonshot

DRUG DISCOVERY IS A COMPLEX MULTI-OBJECTIVE DESIGN PROBLEM

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

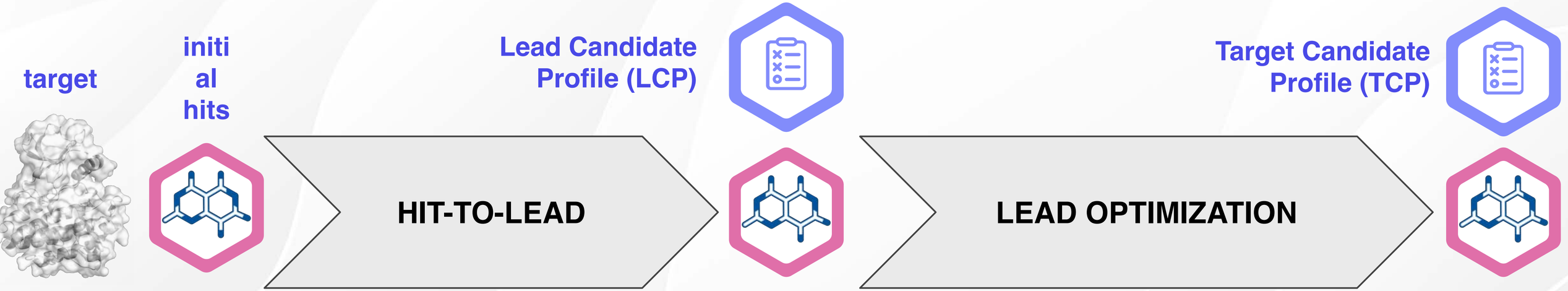
Property	Target range	Rationale
protease assay	IC_{50} < 10 nM	Extrapolation from other anti-viral programs
viral replication assay	EC_{50} < 5 μM	Suppression of virus at achievable blood levels
plaque reduction assay	EC_{50} < 5 μM	Suppression of virus at achievable blood levels
route of administration	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

https://doi.org/10.1101/2020.10.29.339317
https://covid.postera.ai/covid

COVID Moonshot

An international effort to
DISCOVER A COVID ANTIVIRAL

CURRENTLY, SMALL MOLECULE DRUGS ARE DISCOVERED THROUGH THE SYNTHESIS OF THOUSANDS OF COMPOUNDS THAT DON'T MEET DESIGN OBJECTIVES

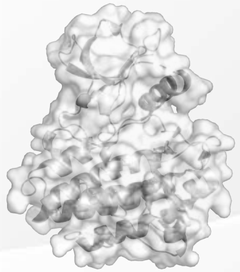


currently ~3.5 years, ~2000 compounds, $12.5M, **high failure rate** [1]

Global annual prescription drug **market will reach $1.9T** by 2030 [2]
90% of all drugs sold are small molecules [3]

WHAT WOULD IT TAKE TO DESIGN DRUG CANDIDATES DIRECTLY FROM OUR OBJECTIVES, MAKING ONLY A FEW CANDIDATE COMPOUNDS THAT SATISFY OUR DESIGN CRITERIA?

target



Target Candidate Profile (TCP)

+

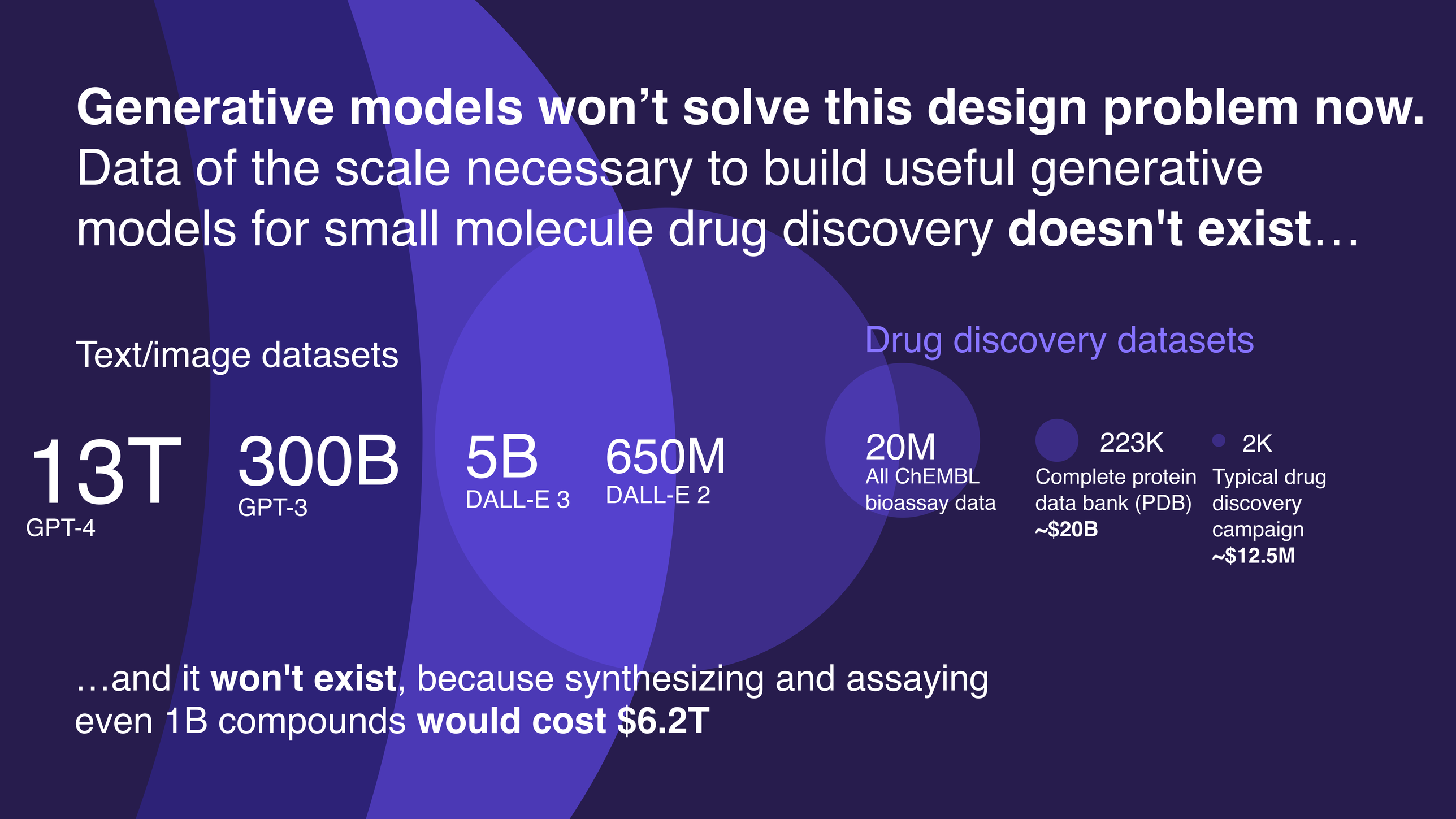
GENERATIVE DESIGN CONDITIONED ON OBJECTIVES

Multiple candidates for preclinical development

~weeks, ~ 10 molecules synthesized by CRO FTE chemists

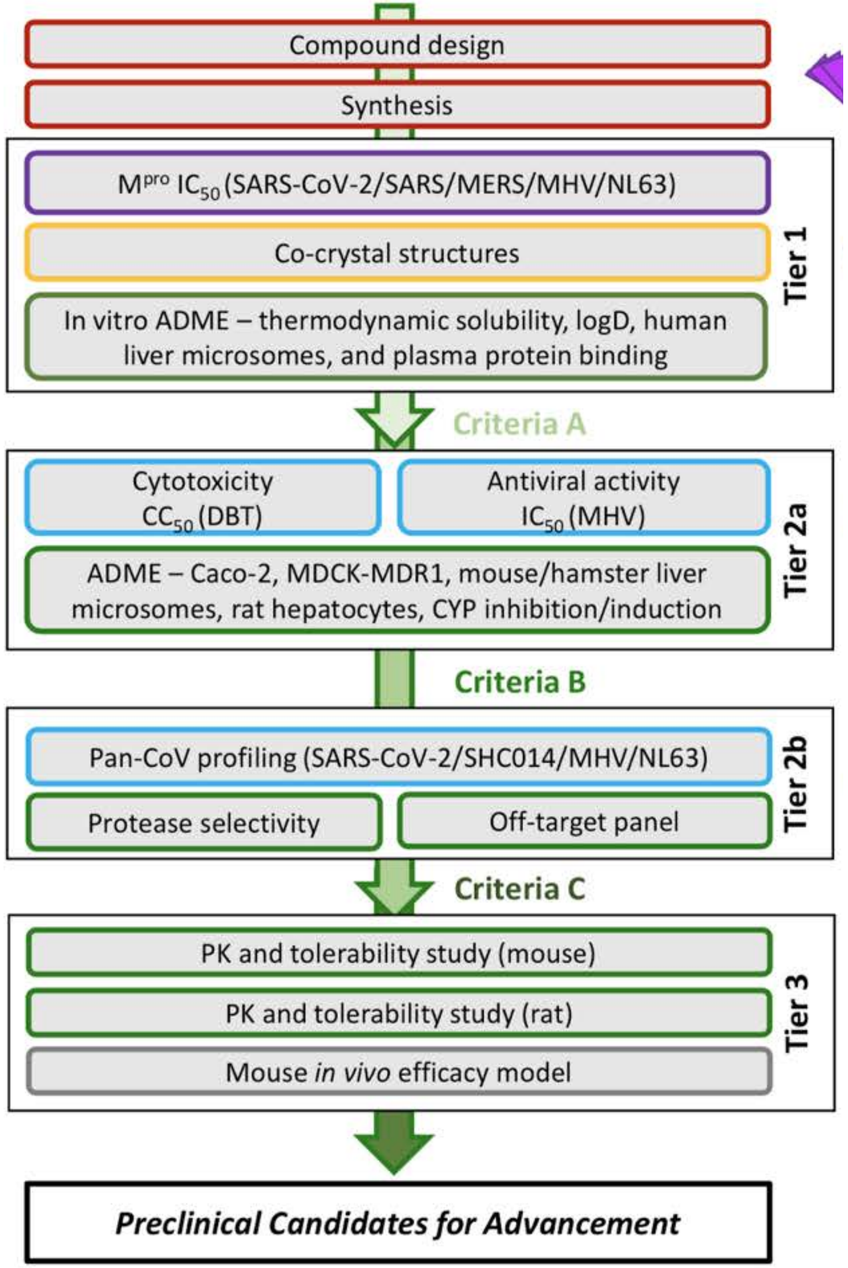
Training useful generative models will require an **enormous amounts of high-quality drug discovery data**

Generative models won’t solve this design problem now.
Data of the scale necessary to build useful generative models for small molecule drug discovery **doesn't exist**…



…and it **won't exist**, because synthesizing and assaying even 1B compounds **would cost $6.2T**

AN ASSAY CASCADE IS USED TO MEET THOSE OBJECTIVES ECONOMICALLY



assay purpose

We have to make the molecule to test it!

Does it inhibit the target? How does it bind?

Does it have a chance of working in humans?

Does it work in cells?
Would it be metabolized or excreted too quickly?

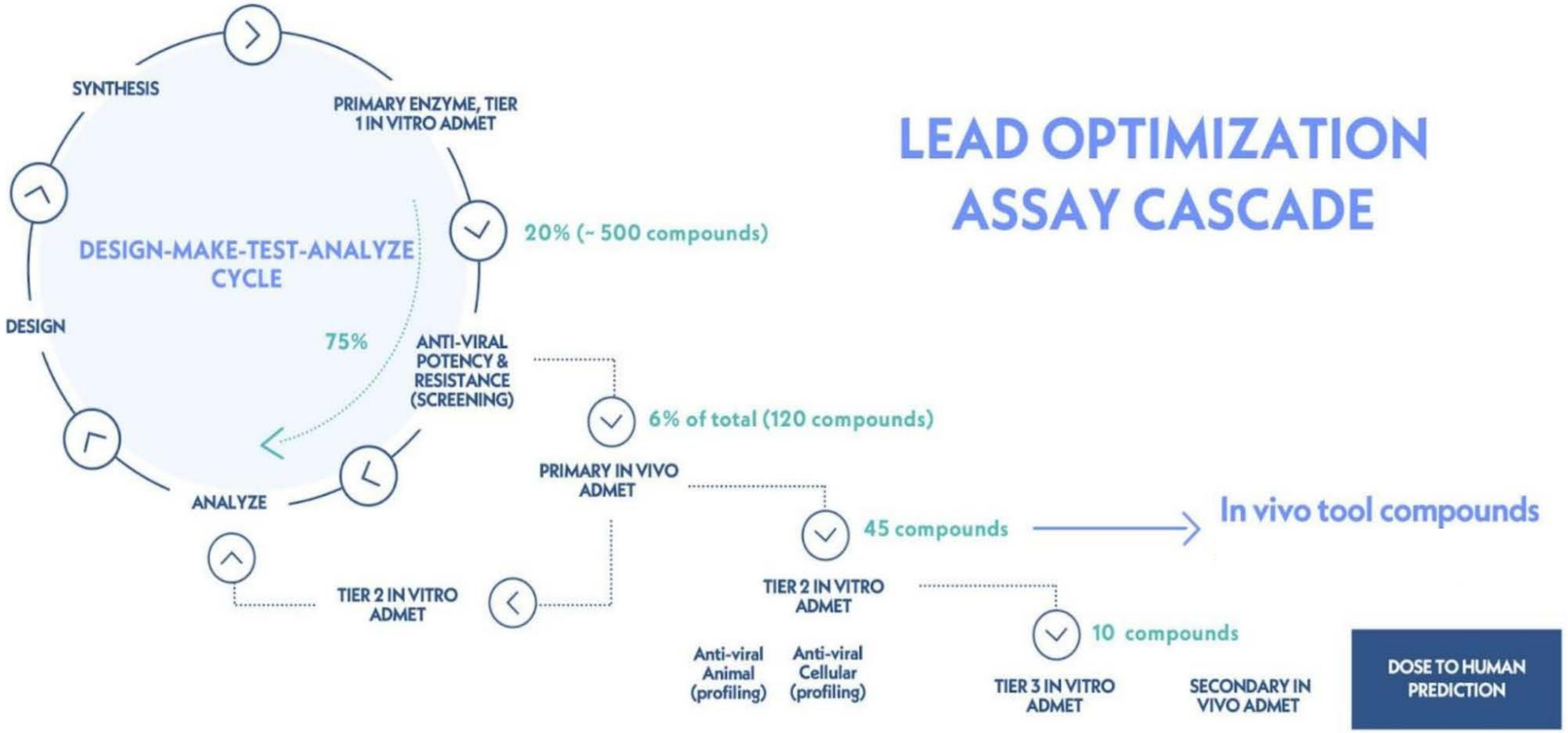
Does it achieve the spectrum we need?
Could it cause bad side effects?

Can oral dosing deliver sufficient drug?

Does it actually work against the disease in animals?

MOTIVATION

Drug discovery is iterative

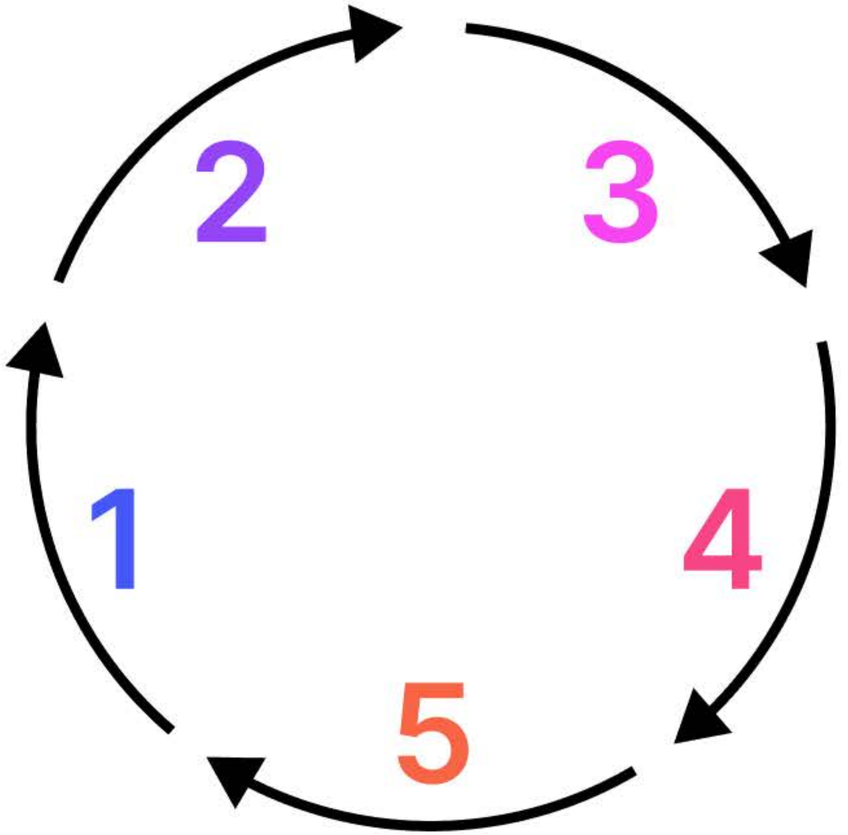


DrugGym: A testbed for the economics of autonomous drug discovery
https://doi.org/10.1101/2024.05.28.596296

MOTIVATION

Drug discovery is iterative

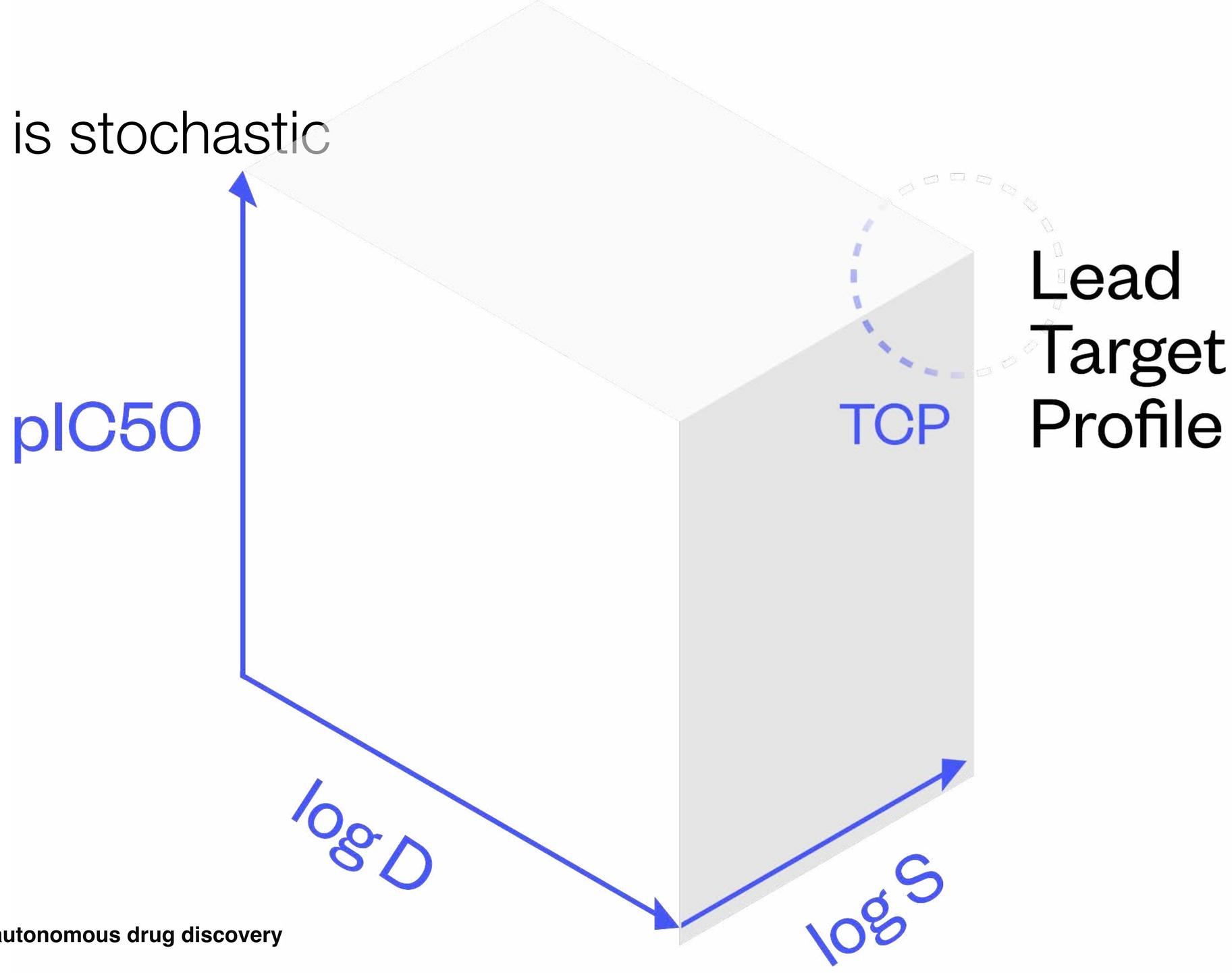
1 ideate
2 triage
3 synthesize
4 assay
5 learn



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

MOTIVATION

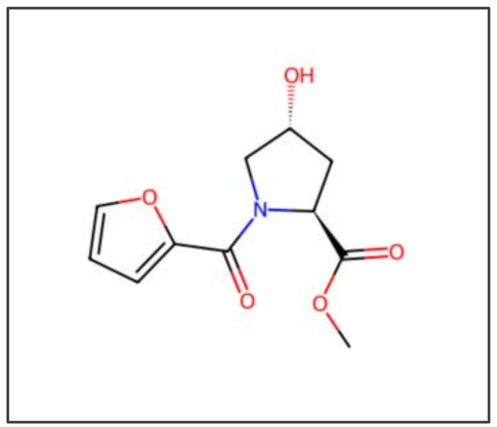
Drug discovery is stochastic

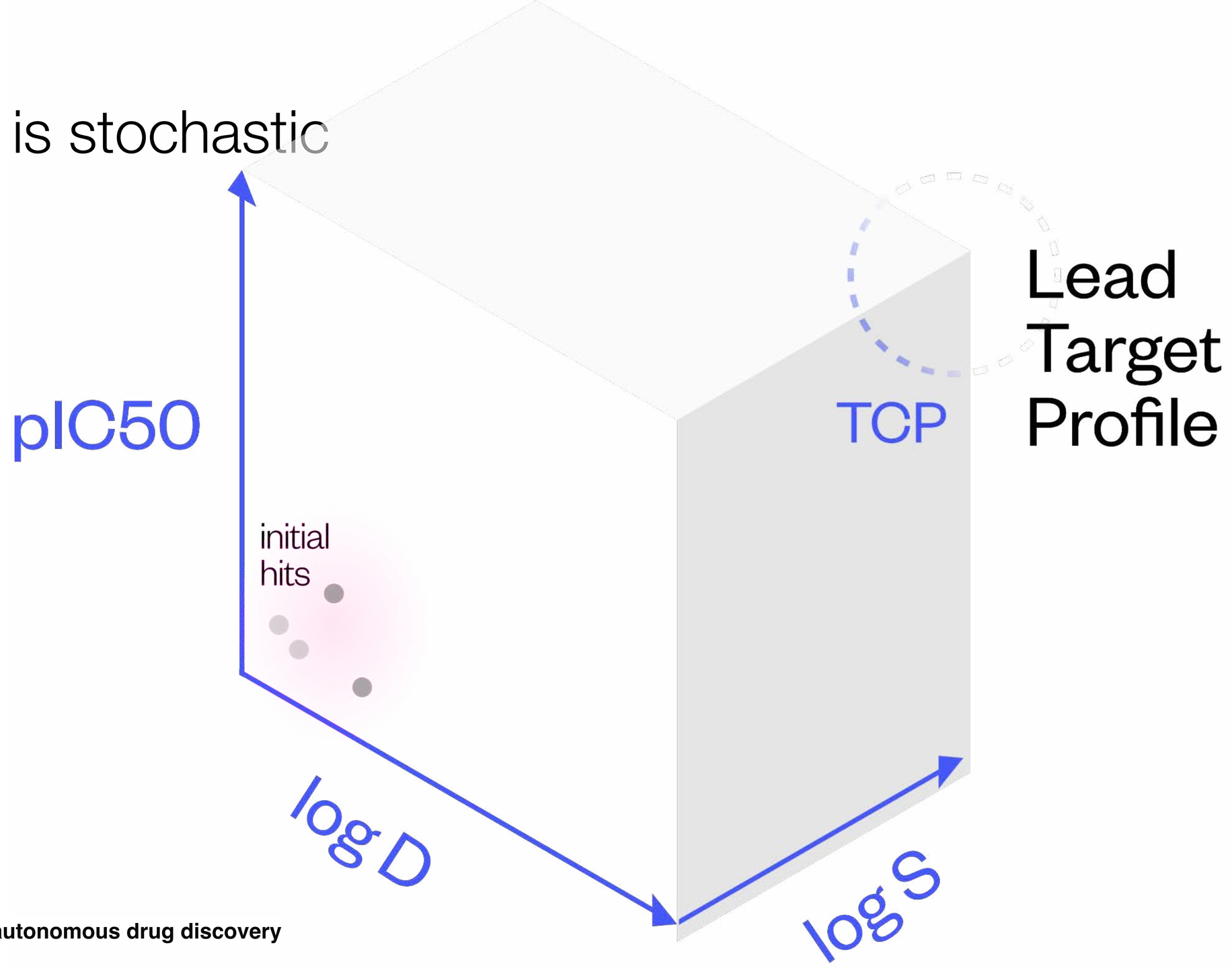


DrugGym: A testbed for the economics of autonomous drug discovery
https://doi.org/10.1101/2024.05.28.596296

MOTIVATION

Drug discovery is stochastic

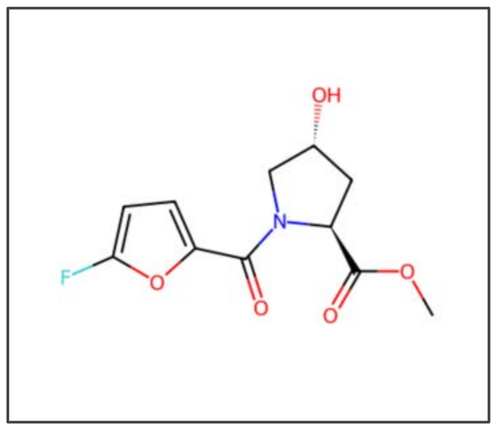


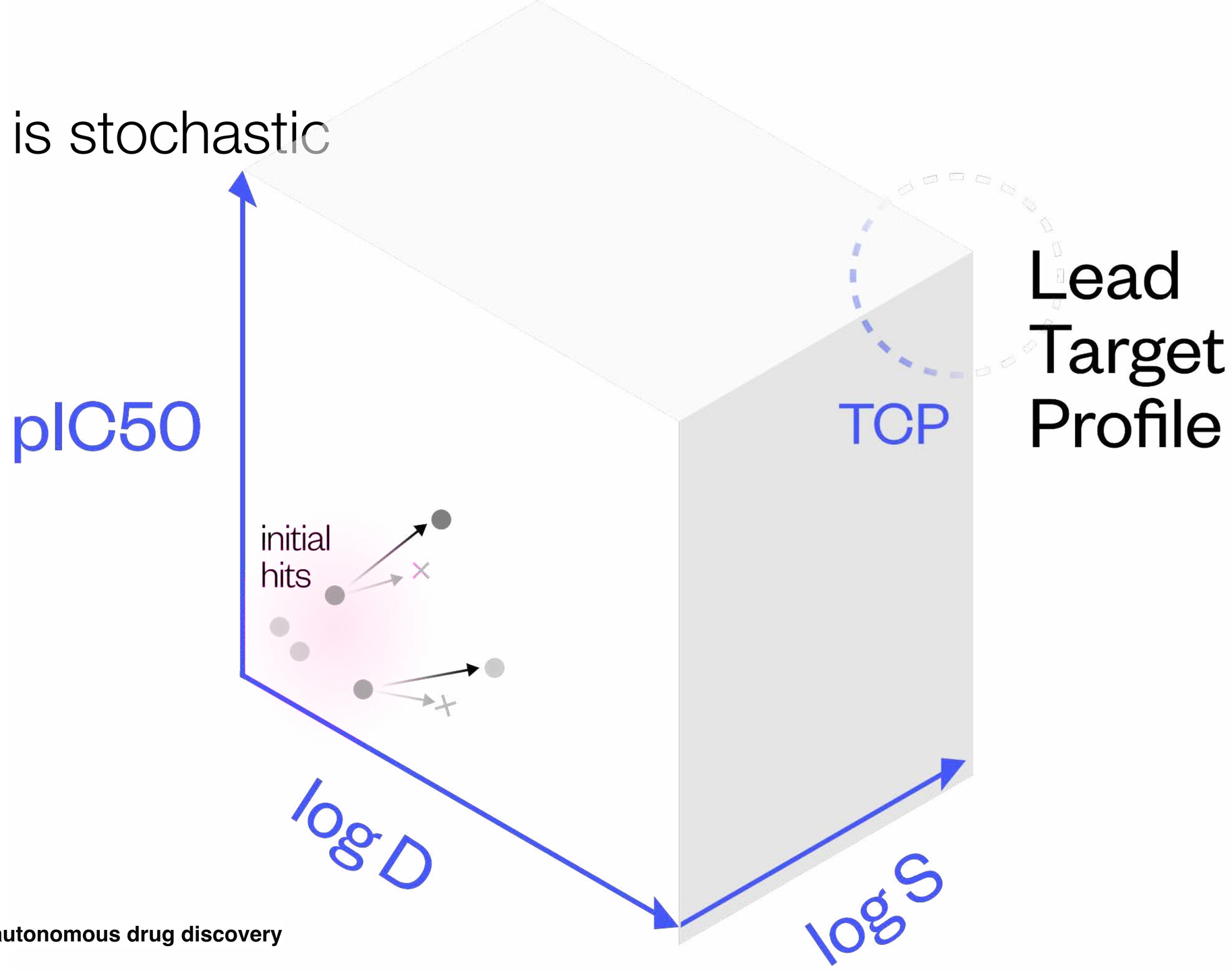


DrugGym: A testbed for the economics of autonomous drug discovery
https://doi.org/10.1101/2024.05.28.596296

MOTIVATION

Drug discovery is stochastic

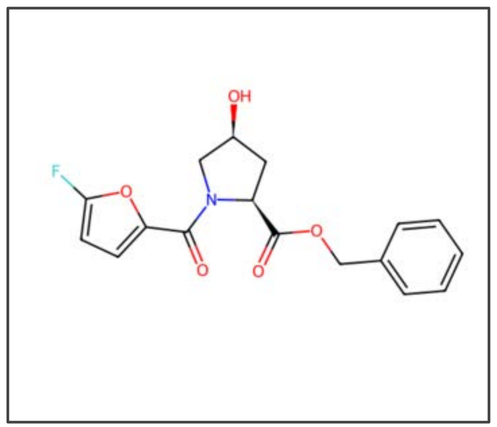


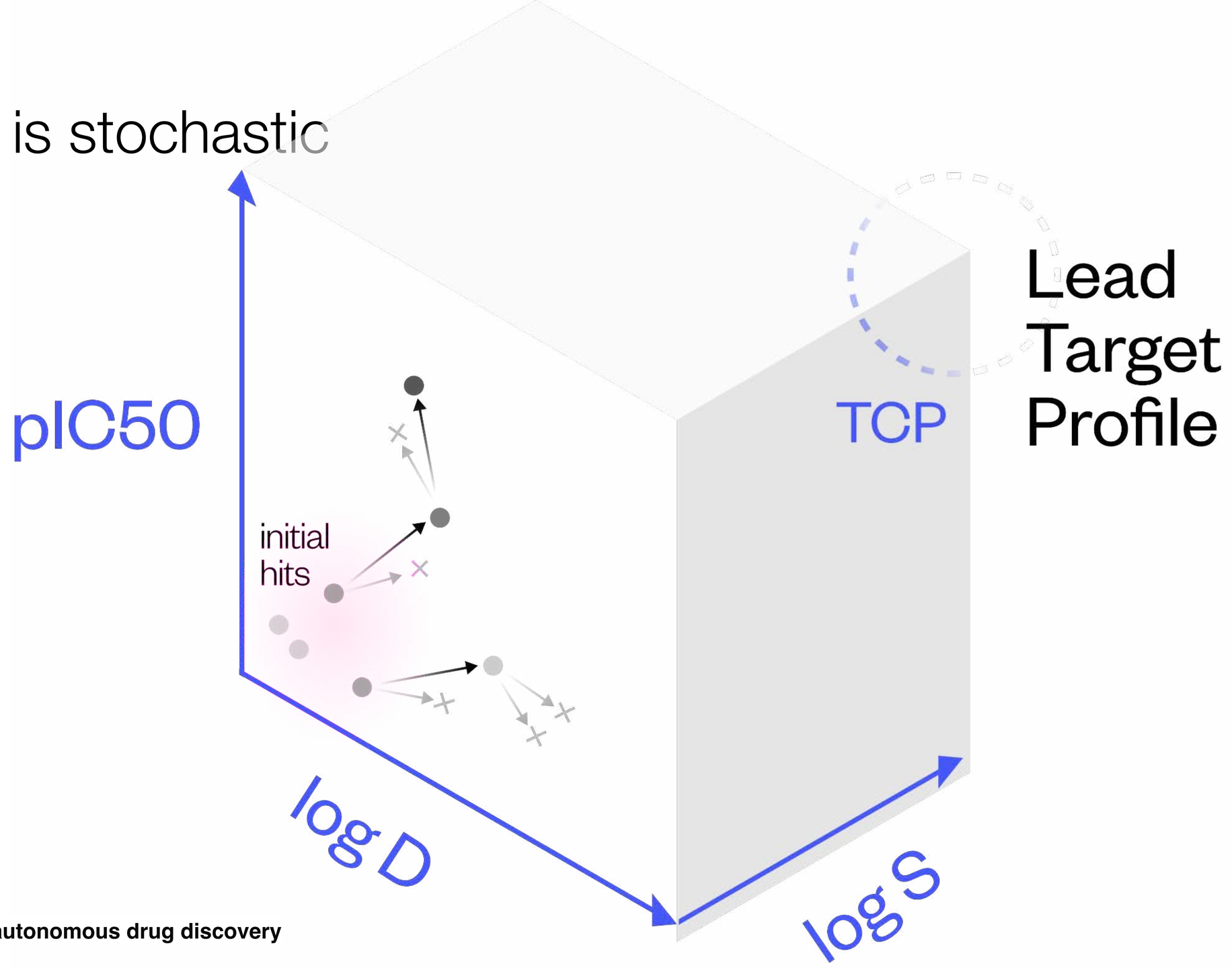


DrugGym: A testbed for the economics of autonomous drug discovery
https://doi.org/10.1101/2024.05.28.596296

MOTIVATION

Drug discovery is stochastic

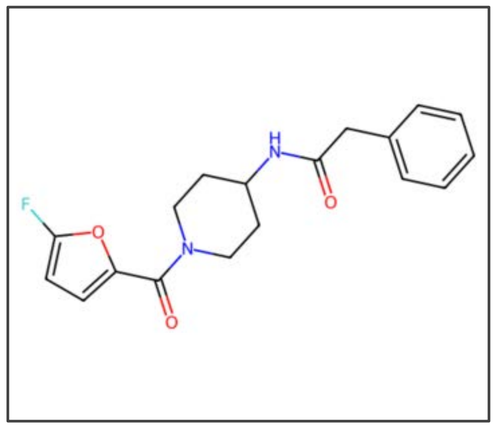


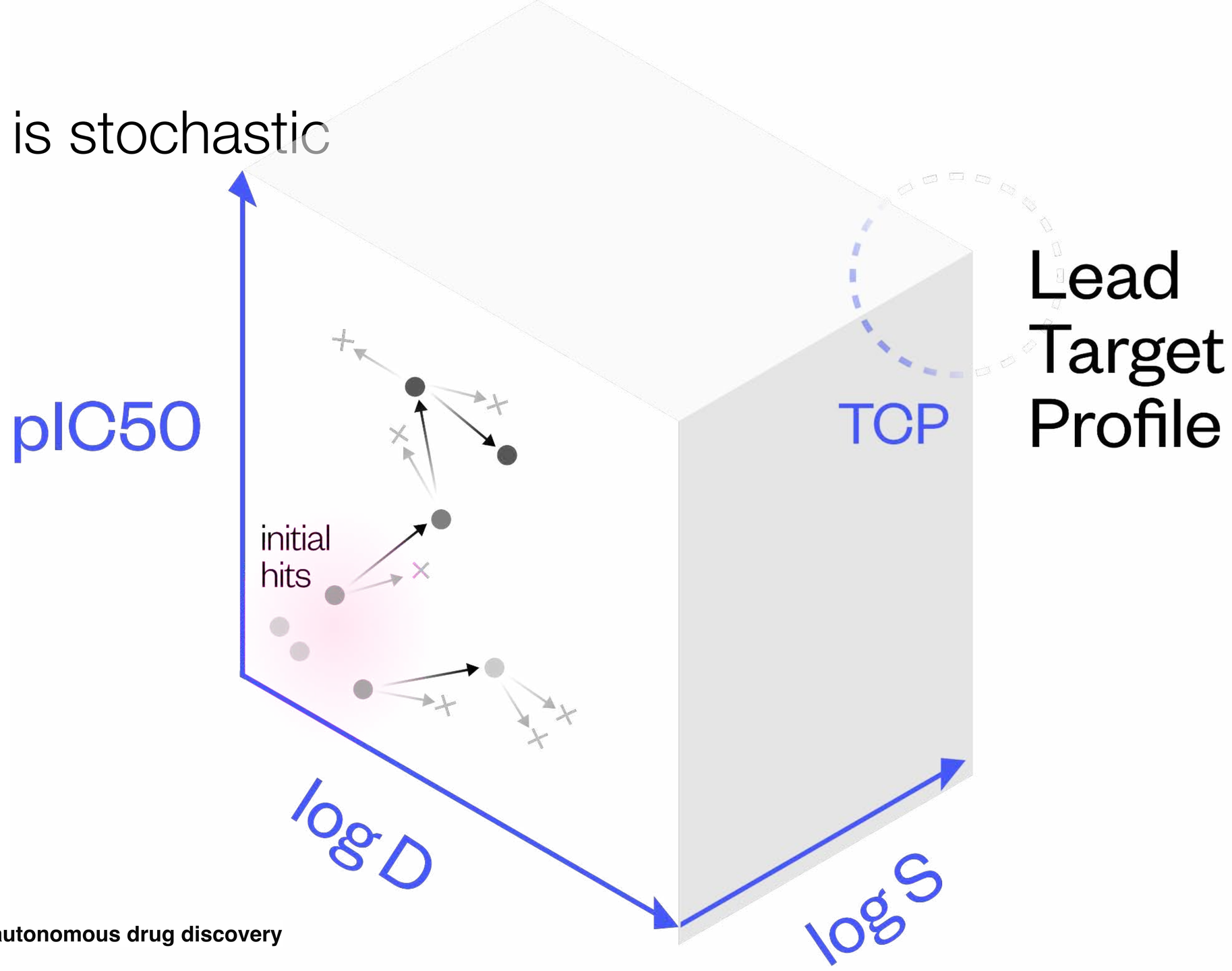


DrugGym: A testbed for the economics of autonomous drug discovery
https://doi.org/10.1101/2024.05.28.596296

MOTIVATION

Drug discovery is stochastic

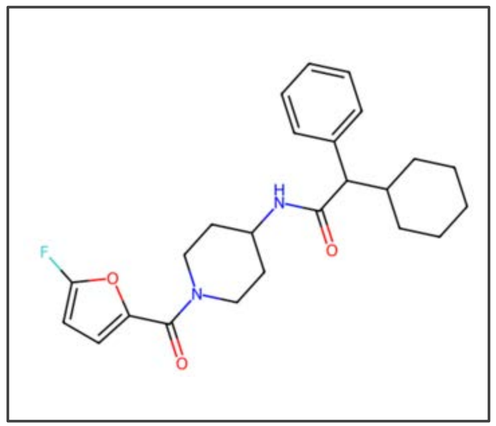


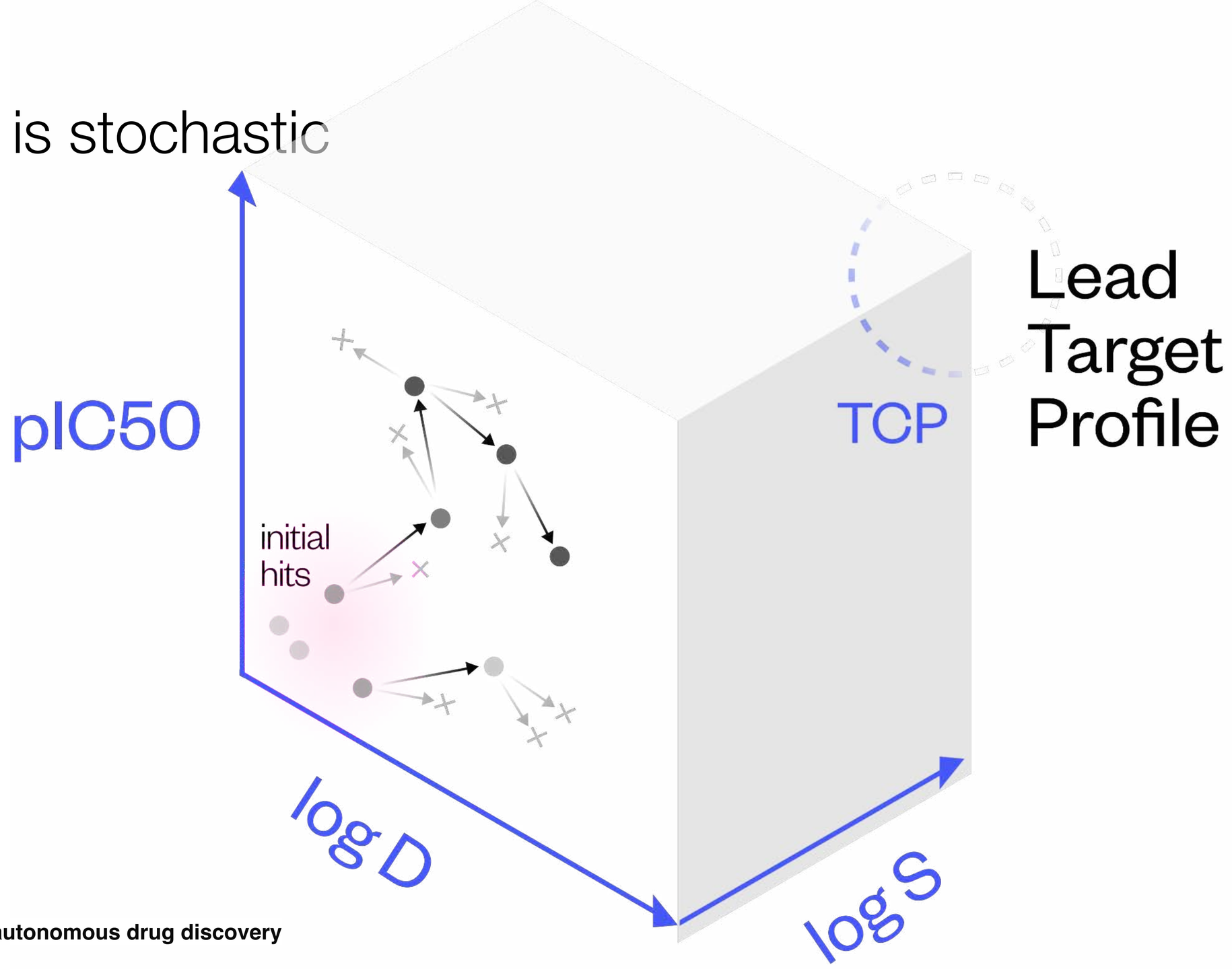


DrugGym: A testbed for the economics of autonomous drug discovery
https://doi.org/10.1101/2024.05.28.596296

MOTIVATION

Drug discovery is stochastic

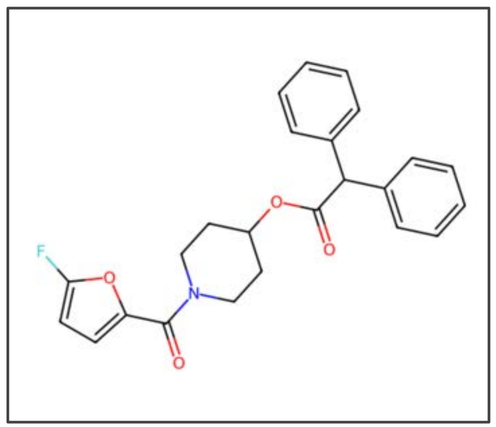


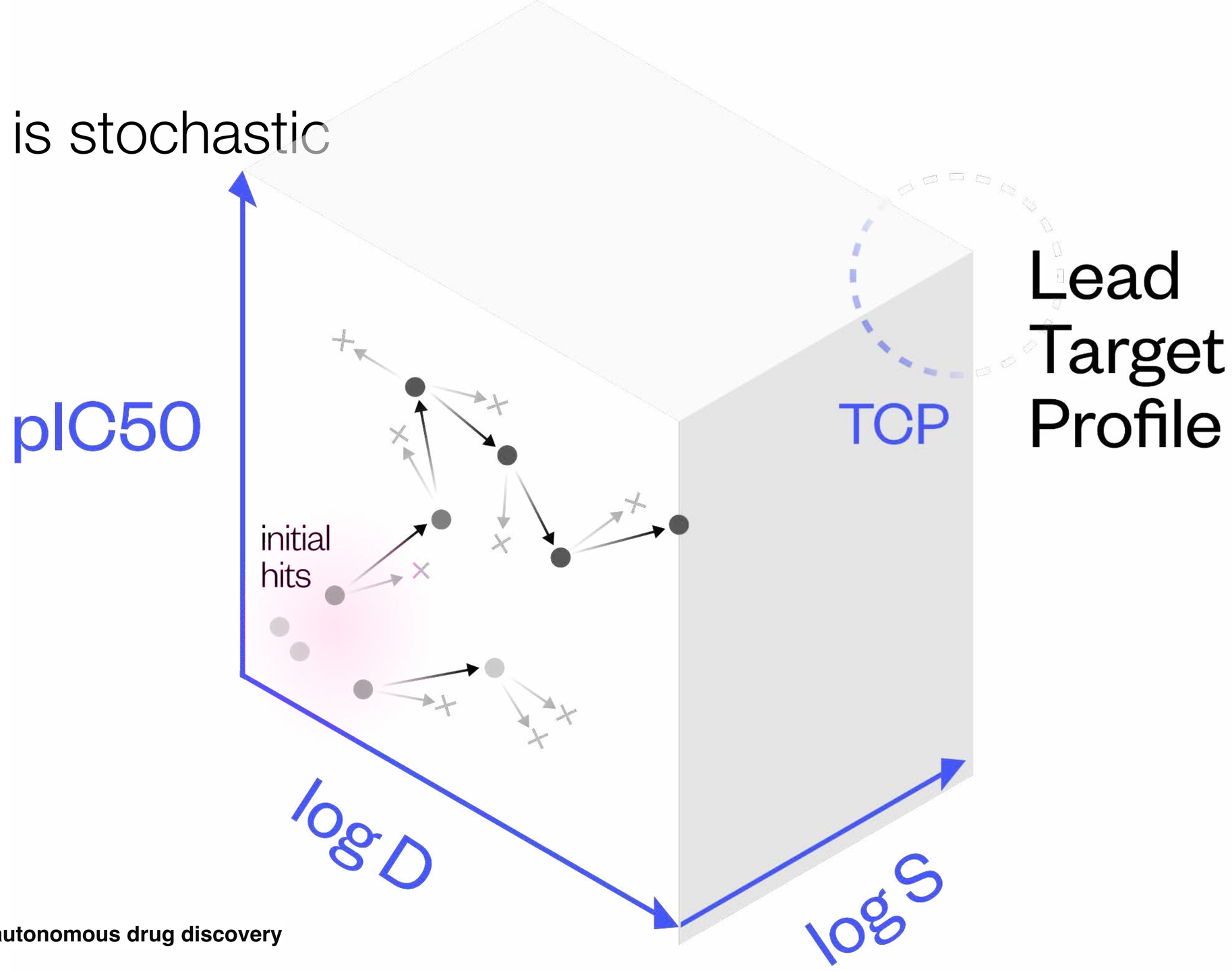


DrugGym: A testbed for the economics of autonomous drug discovery
https://doi.org/10.1101/2024.05.28.596296

MOTIVATION

Drug discovery is stochastic

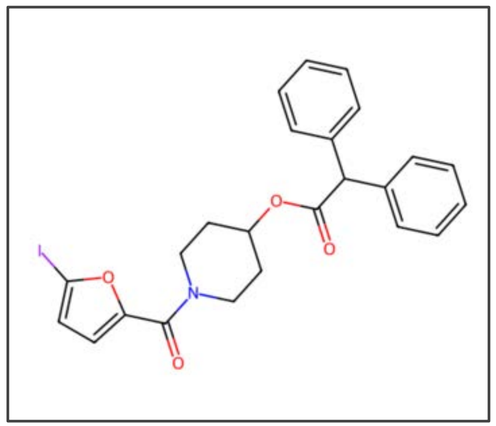


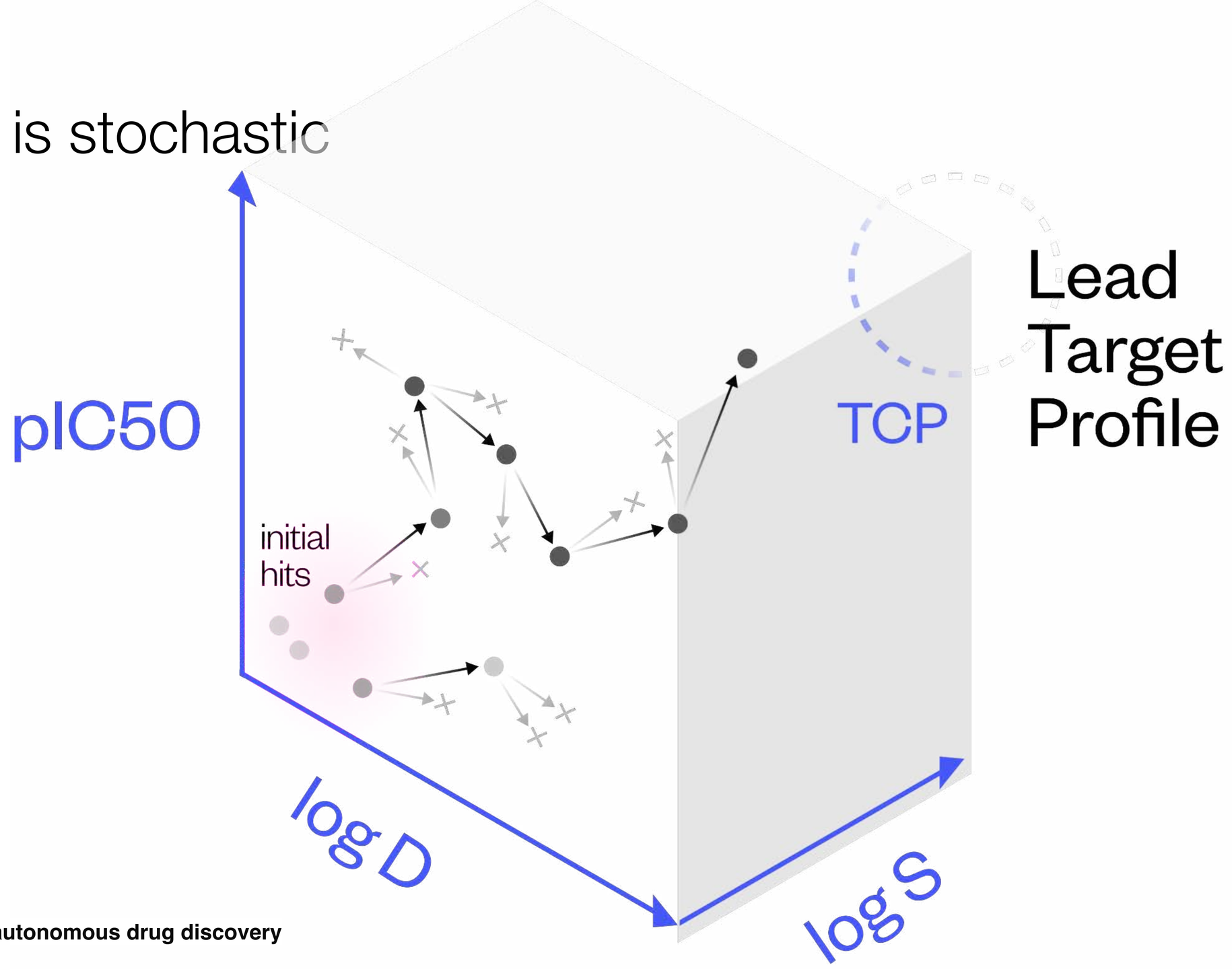


DrugGym: A testbed for the economics of autonomous drug discovery
https://doi.org/10.1101/2024.05.28.596296

MOTIVATION

Drug discovery is stochastic

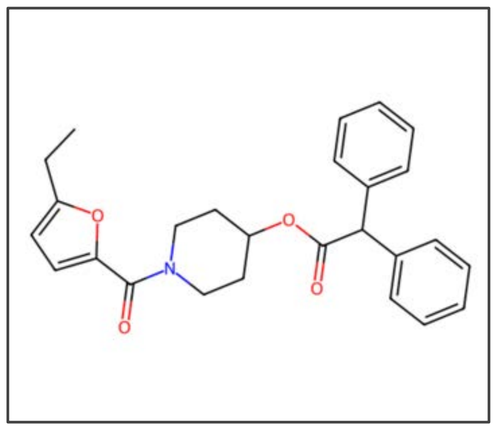


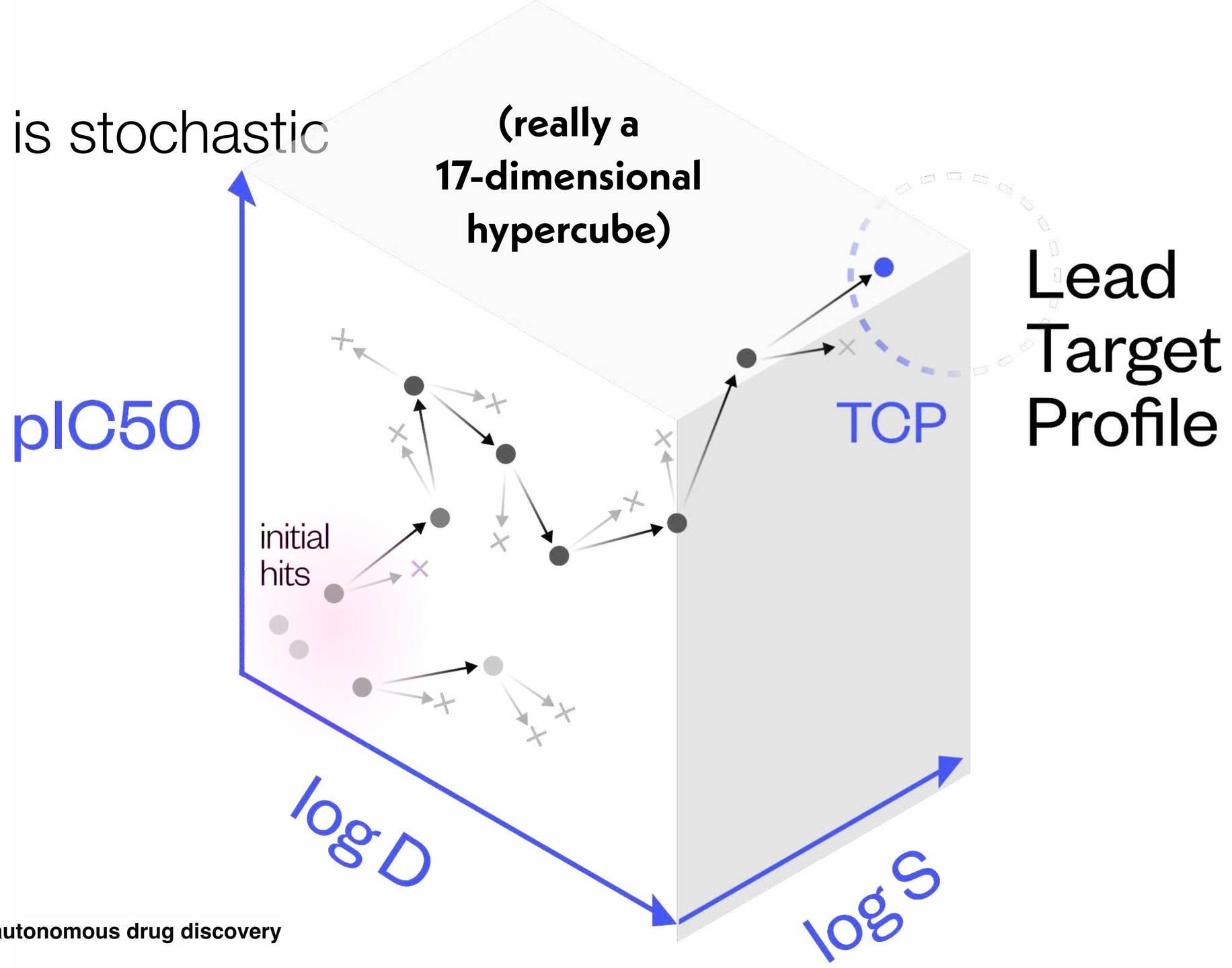


DrugGym: A testbed for the economics of autonomous drug discovery
https://doi.org/10.1101/2024.05.28.596296

MOTIVATION

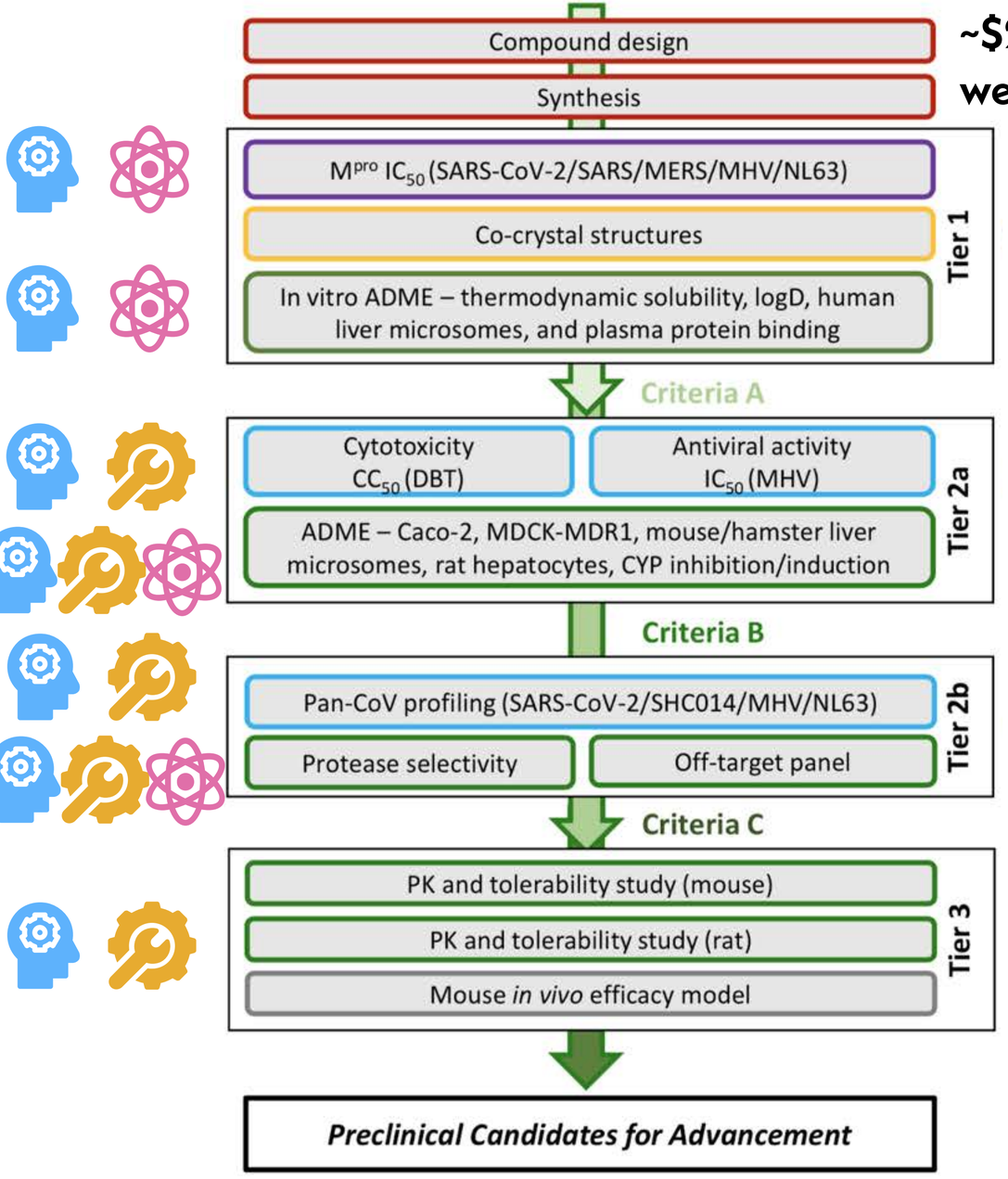
Drug discovery is stochastic





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

ACCURATE SURROGATE MODELS COULD SIGNIFICANTLY 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

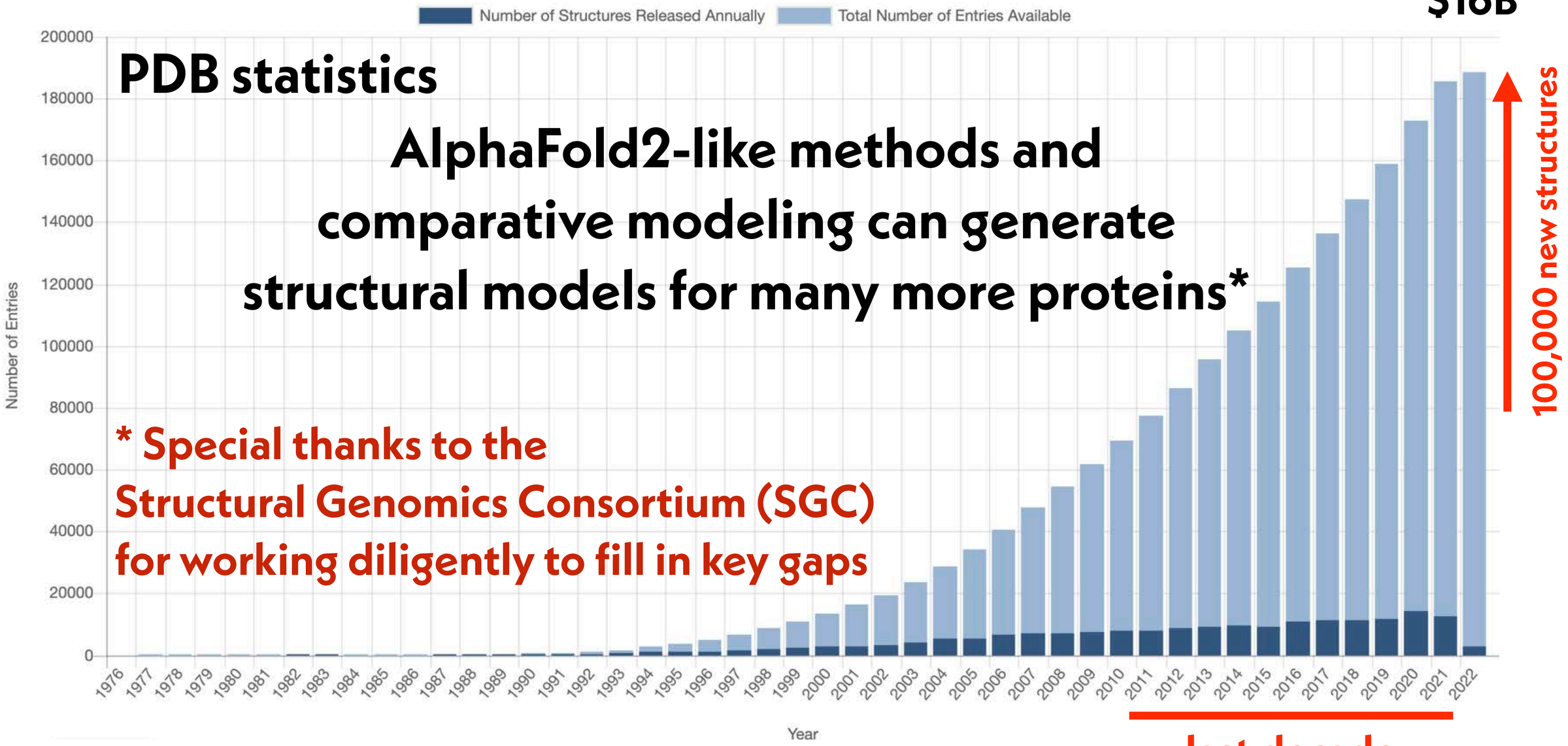
ligand-based machine learning models feasible when data is not target-dependent

mechanistic machine learning models feasible when biological pathways are understood

Efforts like **OpenADMET** will furnish a huge amount of structural data we can use for modeling antitarget-based properties

STRUCTURAL DATA IS NOW AN ABUNDANT RESOURCE FOR DRUG DISCOVERY

$16B



PDB statistics

AlphaFold2-like methods and comparative modeling can generate structural models for many more proteins*

100,000 new structures

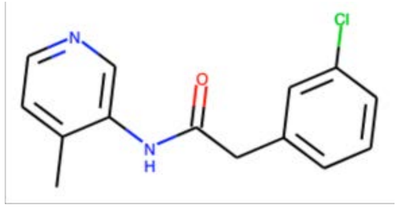
* Special thanks to the Structural Genomics Consortium (SGC) for working diligently to fill in key gaps

last decade

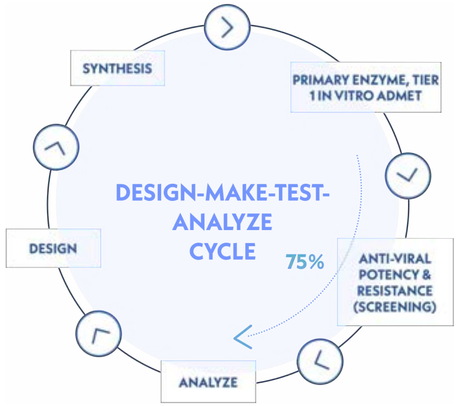
http://www.rcsb.org/stats

DESIGN-MAKE-TEST-ANALYZE CYCLES SHARE A COMMON OPERATION:

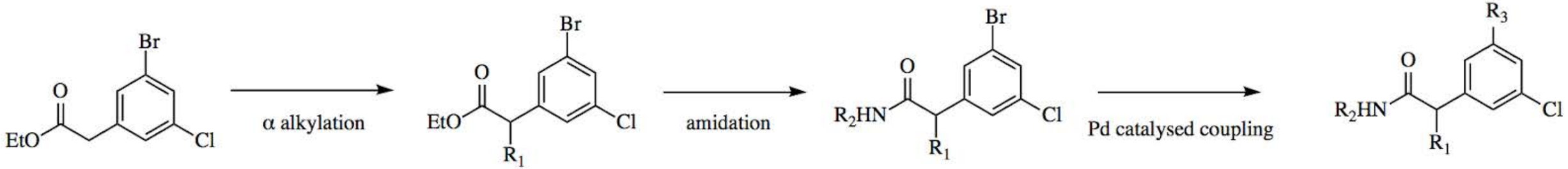
1. Select a current **lead molecule**



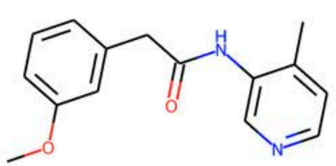
IC_{50} = 25 µM
TRY-UNI-714a760b-6



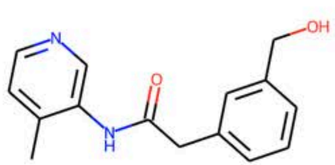
2. Use AI tools to identify a **retrosynthetic pathway** capable of installing new groups to replace part of the molecule



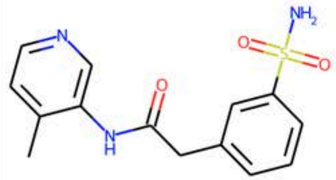
3. Chemists conservatively **select analogues** from the (often very) large enumerated synthetic space



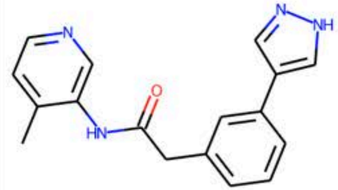
EDJ-MED-e58735b6-1



EDJ-MED-e58735b6-2

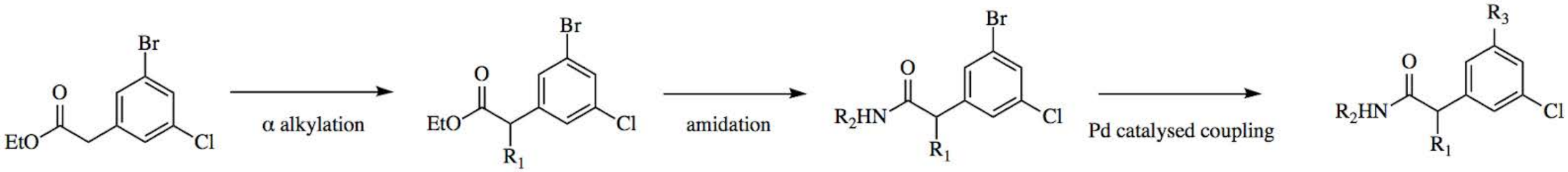


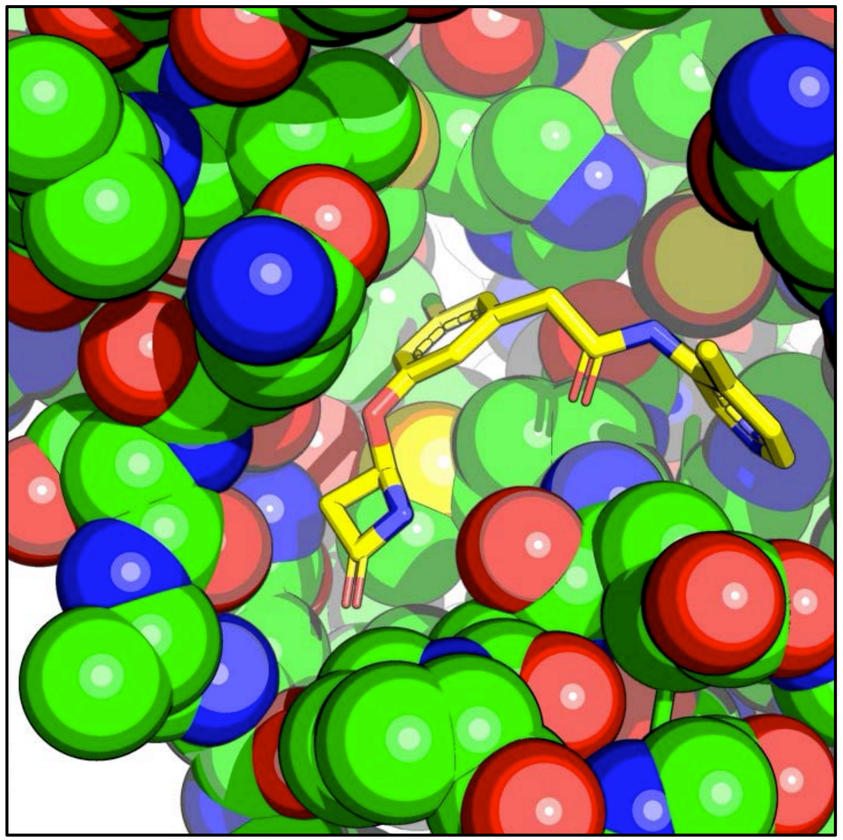
EDJ-MED-e58735b6-3



EDJ-MED-e58735b6-4

PREDICTIVE MODELS AIM TO REDUCE THE NUMBER OF DESIGN CYCLES NEEDED TO ACHIEVE OUR DESIGN OBJECTIVES

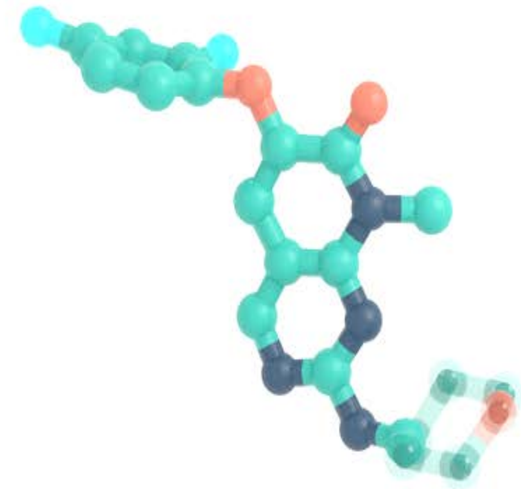


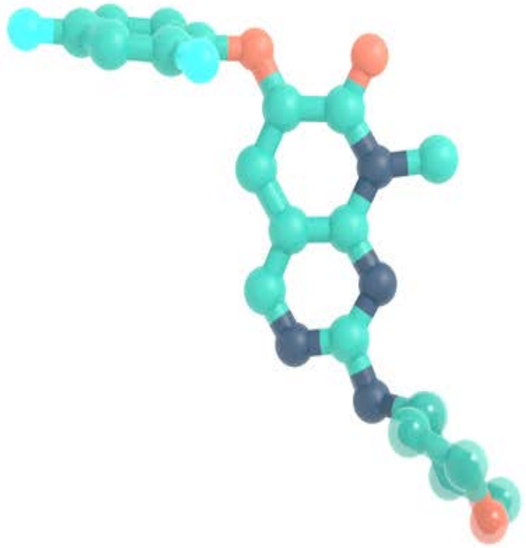


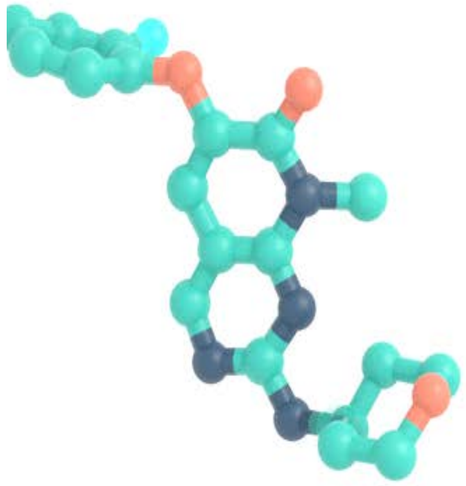
~15,000 Potential R3 groups

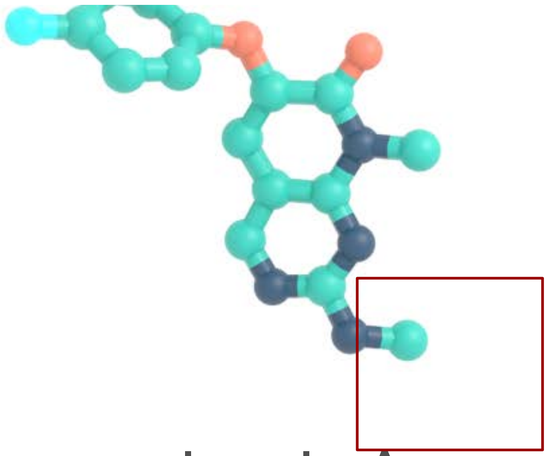
Enamine

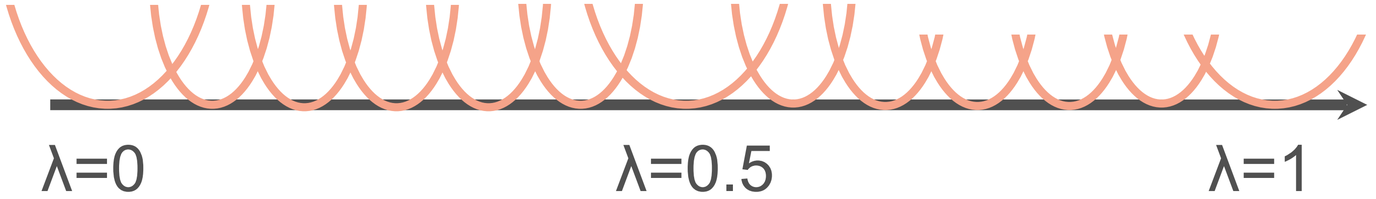
ALCHEMICAL FREE ENERGY CALCULATIONS HAVE PROVEN TO BE A USEFUL WAY TO EXPLOIT STRUCTURAL DATA TO MAKE PREDICTIONS

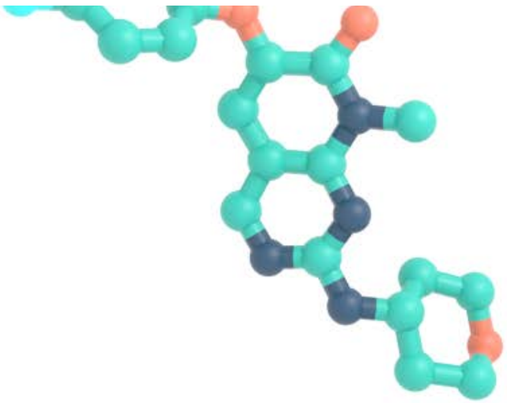












$H(\lambda) = (1-\lambda)\, H_A + \lambda\, H_B$

alchemical transformation path

molecule A

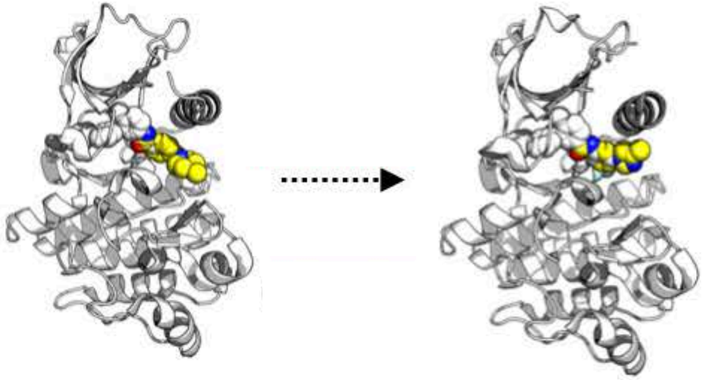
molecule B

See alchemistry.org to learn more about fundamentals

ALCHEMICAL FREE ENERGY CALCULATIONS HAVE A BROAD DOMAIN OF APPLICABILITY IN DRUG DISCOVERY

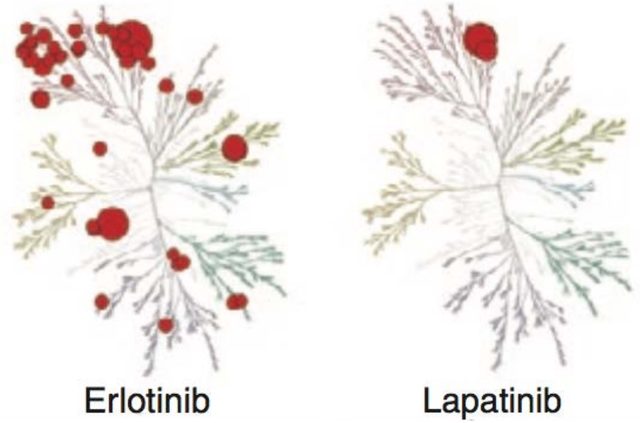
driving affinity / potency

Schindler, Baumann, Blum et al. JCIM 11:5457, 2020
https://doi.org/10.1021/acs.jcim.0c00900



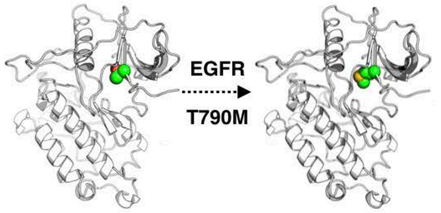
driving selectivity

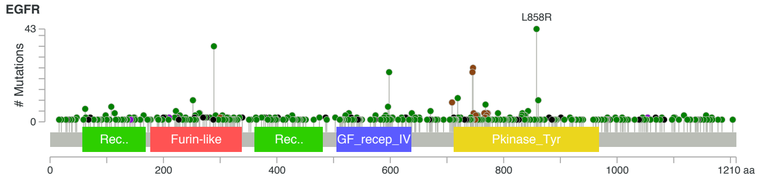
Moraca, Negri, de Olivera, Abel JCIM 2019
https://doi.org/10.1021/acs.jcim.9b00106
Aldeghi et al. JACS 139:946, 2017.
https://doi.org/10.1021/jacs.6b11467



predicting clinical drug resistance/sensitivity

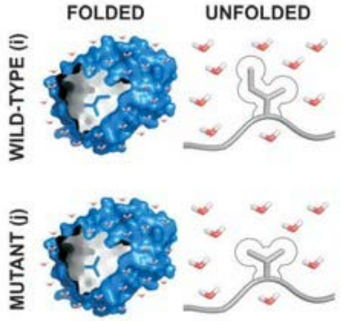
Hauser, Negron, Albanese, Ray, Steinbrecher, Abel, Chodera, Wang.
Communications Biology 1:70, 2018
https://doi.org/10.1038/s42003-018-0075-x
Aldeghi, Gapsys, de Groot. ACS Central Science 4:1708, 2018
https://doi.org/10.1021/acscentsci.8b00717

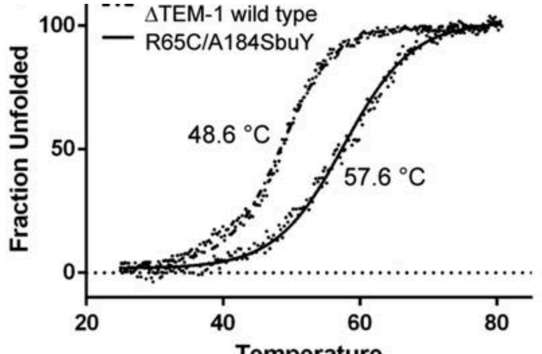




optimizing thermostability

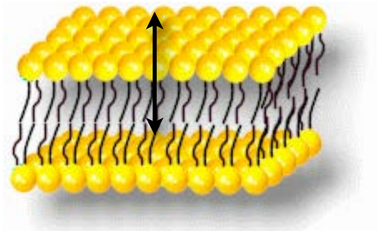
Gapsys, Michielssens, Seeliger, and de Groot. Angew Chem 55:7364, 2016
https://doi.org/10.1002/anie.201510054



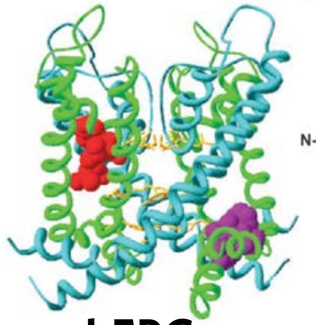


...AND HOLD THE POTENTIAL FOR EVEN BROADER APPLICABILITY AS MORE STRUCTURAL DATA EMERGES

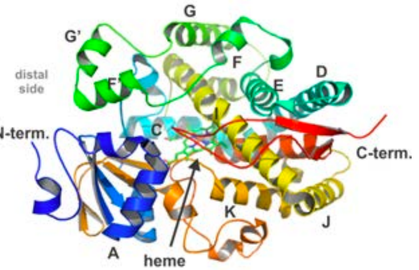
partition coefficients (logP, logD) and permeabilities



structure-enabled ADME/Tox targets

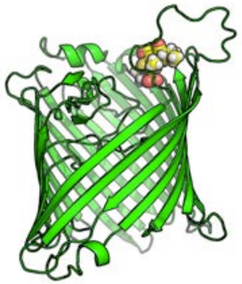


hERG

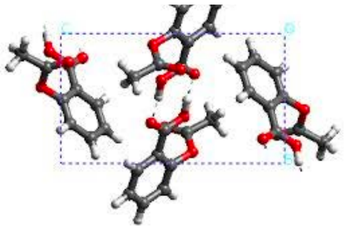


CYP3A4

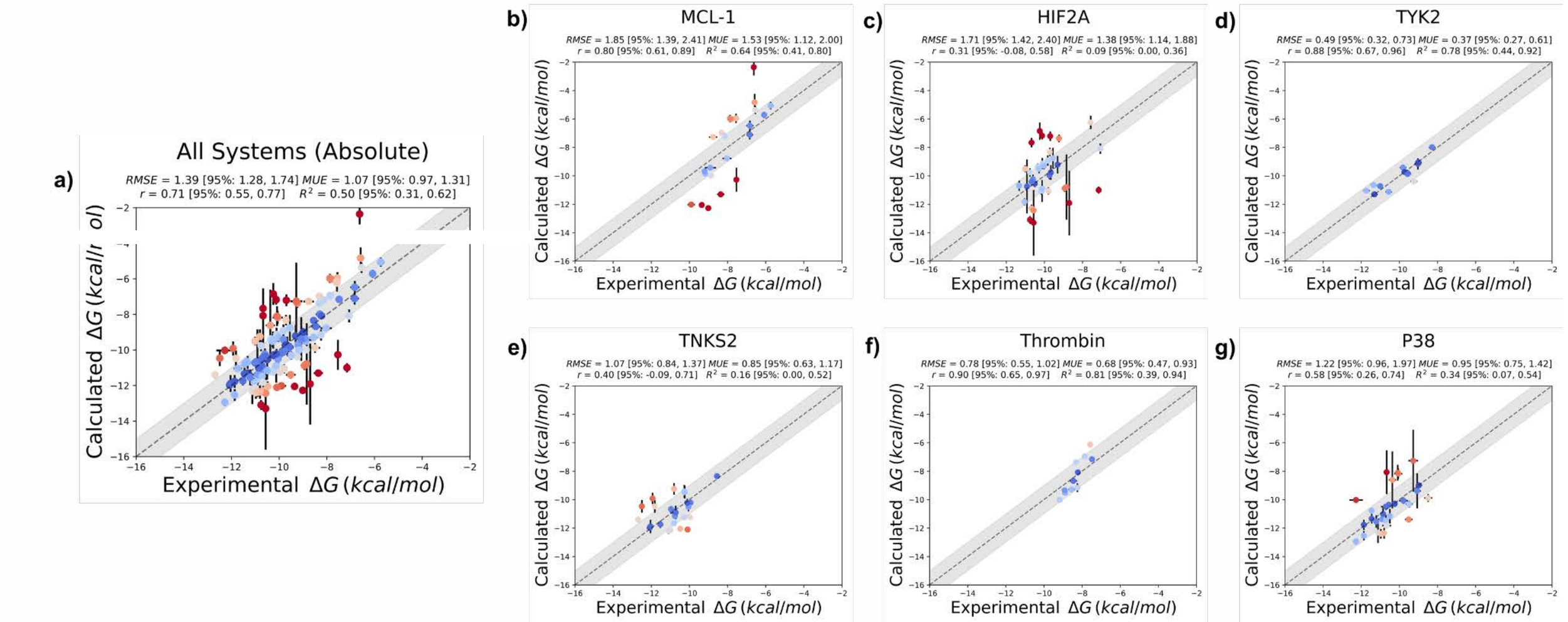
porin permeation



crystal polymorphs, etc.

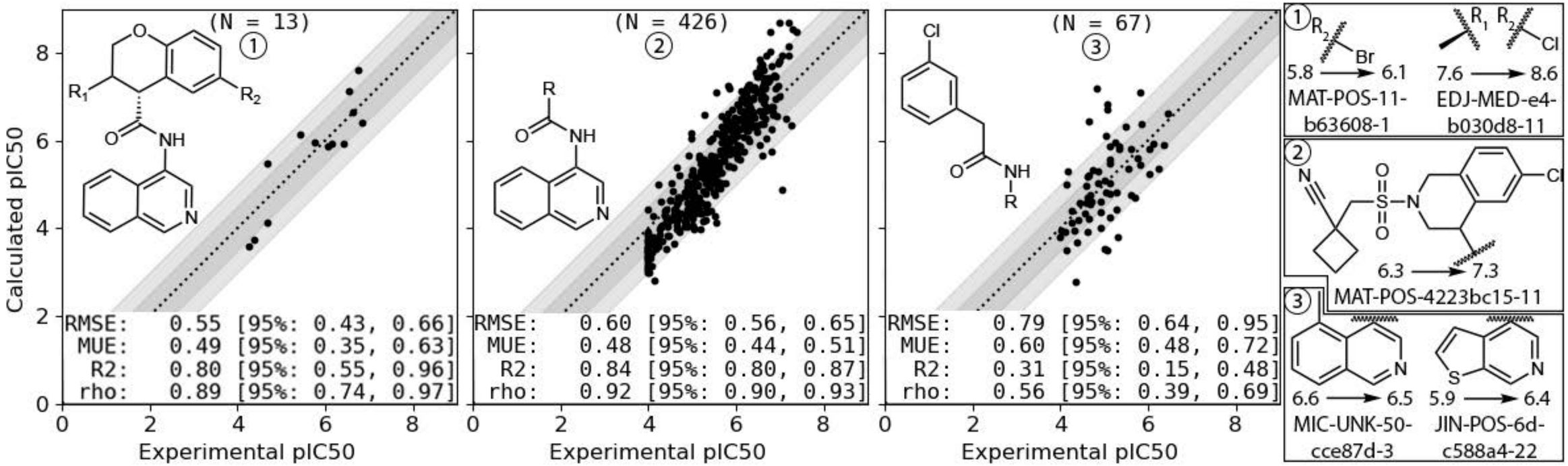


RETROSPECTIVE BENCHMARKS CAN ASSESS THE ACCURACY EXPECTED FOR DIFFERENT TARGET CLASSES AND SCAFFOLDS

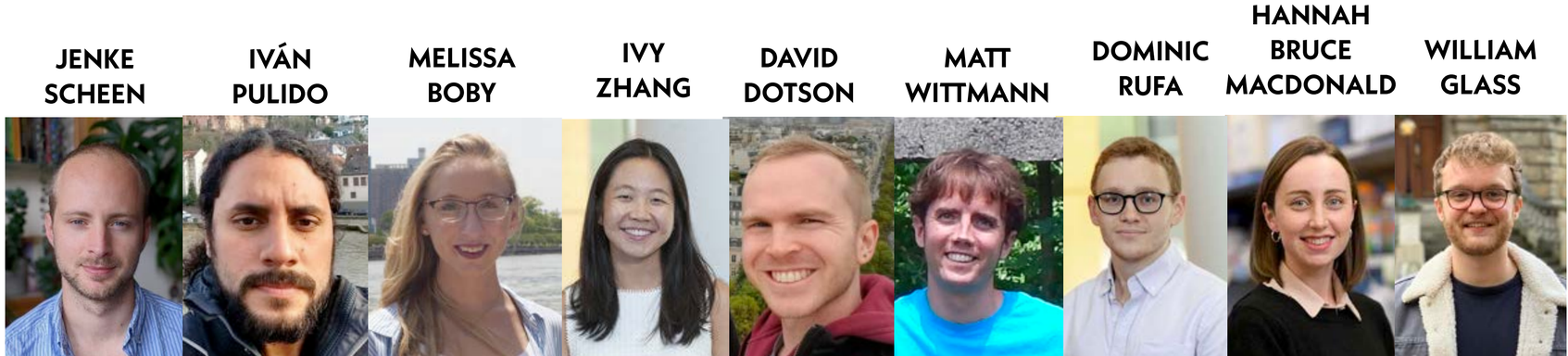


Best practices for constructing, preparing, and evaluating protein-ligand binding affinity benchmarks
https://doi.org/10.33011/livecoms.4.1.1497

ALCHEMICAL FREE ENERGY CALCULATIONS CAN HAVE REASONABLE ACCURACY IN PROSPECTIVE DISCOVERY



preprint: https://doi.org/10.1101/2020.10.29.339317



The Open Molecular Software Foundation (OMSF) is building an open source ecosystem for accelerating drug discovery

http://omsf.io

OMSF hosted open source projects provide **a foundation for innovation** and **useful baseline comparisons** for new technologies in drug discovery

http://openforcefield.org

http://openfree.energy

https://westpa.github.io

http://openfold.io

http://rosettacommons.org

OpenADMET

http://openadmet.org

open forcefield

An open and collaborative approach to better force fields

OPEN SOURCE

Software permissively licensed under the MIT License and developed openly on GitHub.

OPEN SCIENCE

Scientific reports as blog posts, webinars and preprints

OPEN DATA

Curated quantum chemical and experimental datasets used to parameterize and benchmark Open Force Fields.

NEWS | TUTORIALS | ROADMAP

http://openforcefield.org

The Open Free Energy Consortium develops open source alchemical free energy calculations that deliver best practices

OpenFE Developers


Irfan Alibay


David Mobley


James Eastwood


Iván Pulido


Josh Horton


Hannah Baumann


Mike Henry


Benjamin Ries


David Dotson


Ian Kenny

Technical Advisory Committee

14 academic members

Industry Partner Members

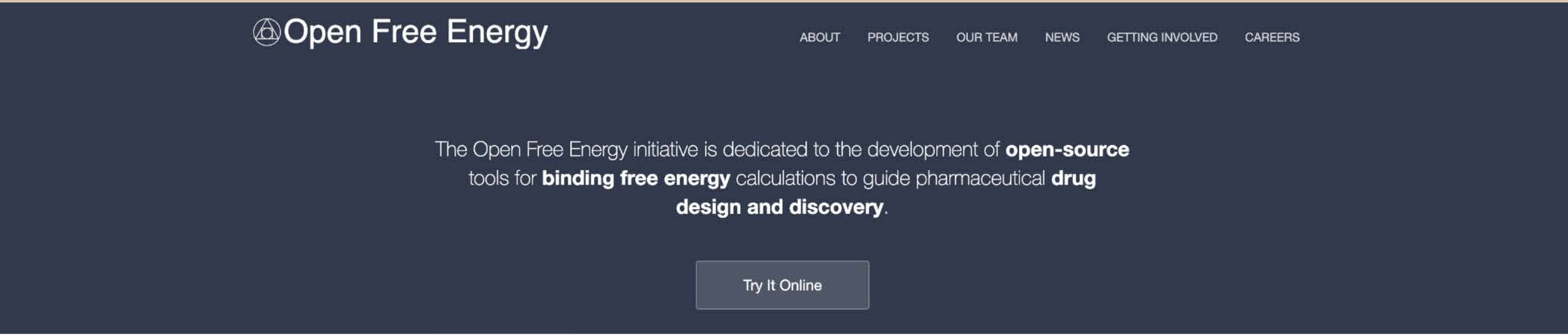
18 pharma companies

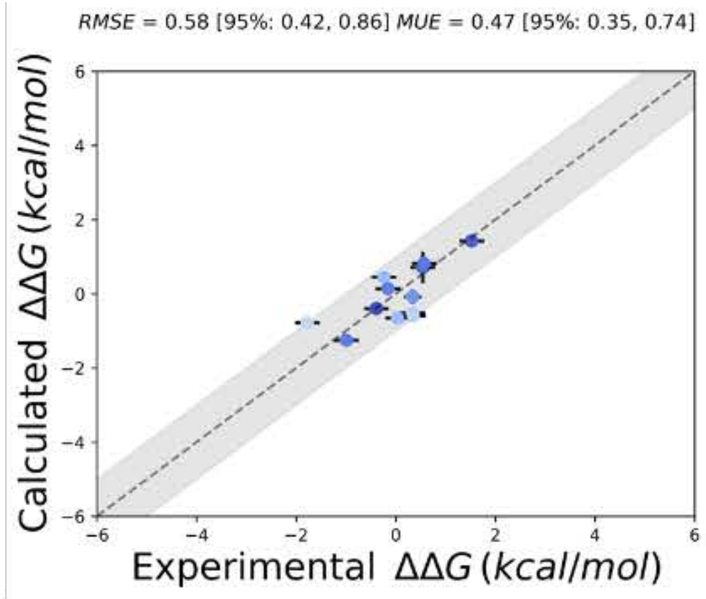
Best Practices Publications

2 Living Journal of Computational Molecular Sciences (LiveCoMS) Best Practices publications
https://livecomsjournal.org

https://openfree.energy

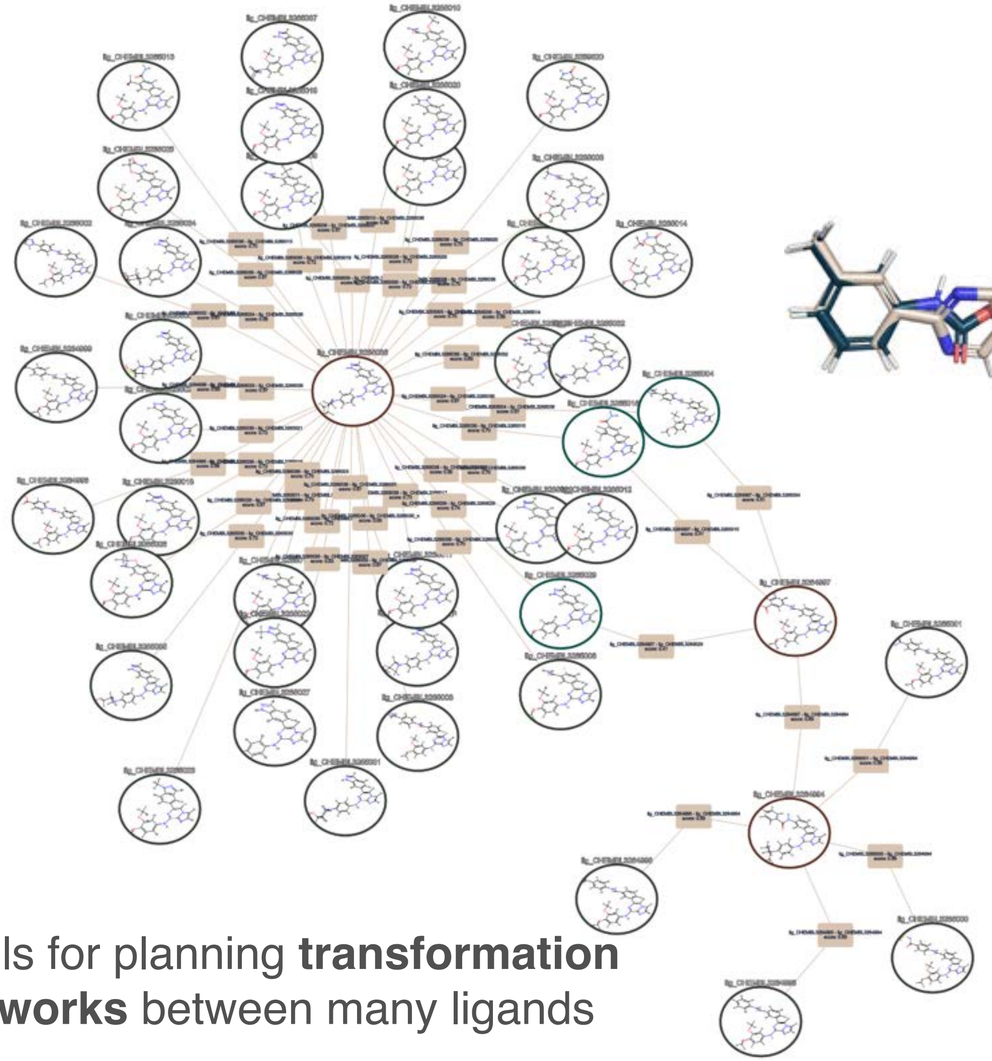
Open Free Energy provides open source tools for free energy calculations you can start using from your browser



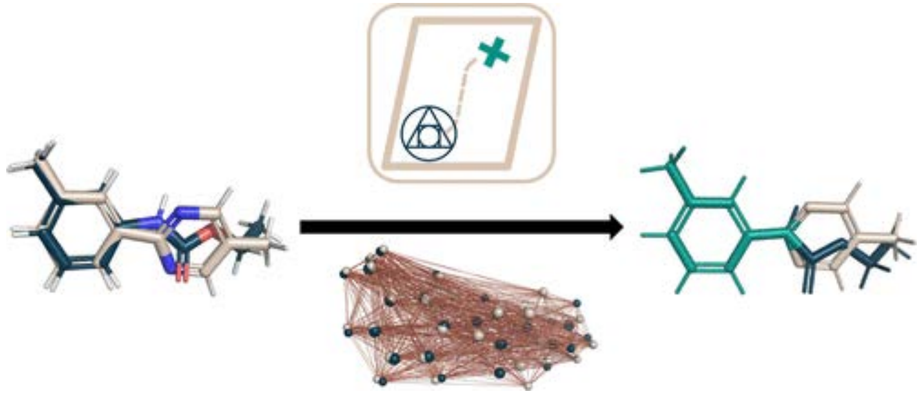


https://openfree.energy

Open Free Energy provides open source tools for planning alchemical free energy campaigns



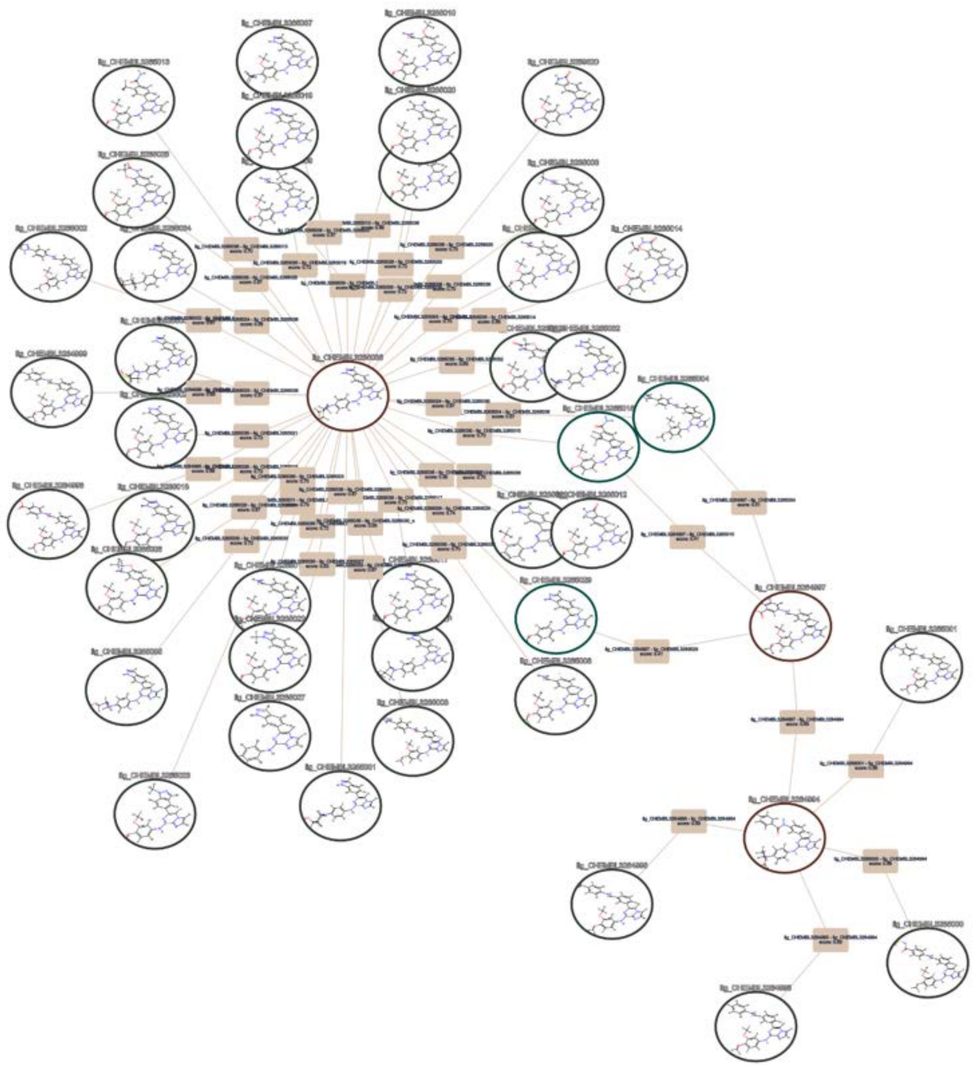
Tools for planning **transformation networks** between many ligands



Tools for automatically planning **optimal transformations** between ligand pairs

https://openfree.energy

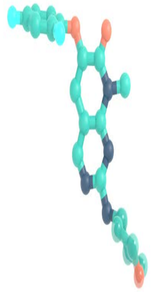
… workflows for running alchemical transformations on CPUs or GPUs



OpenMM

http://openmm.org

open source GPU-accelerated biomolecular simulations

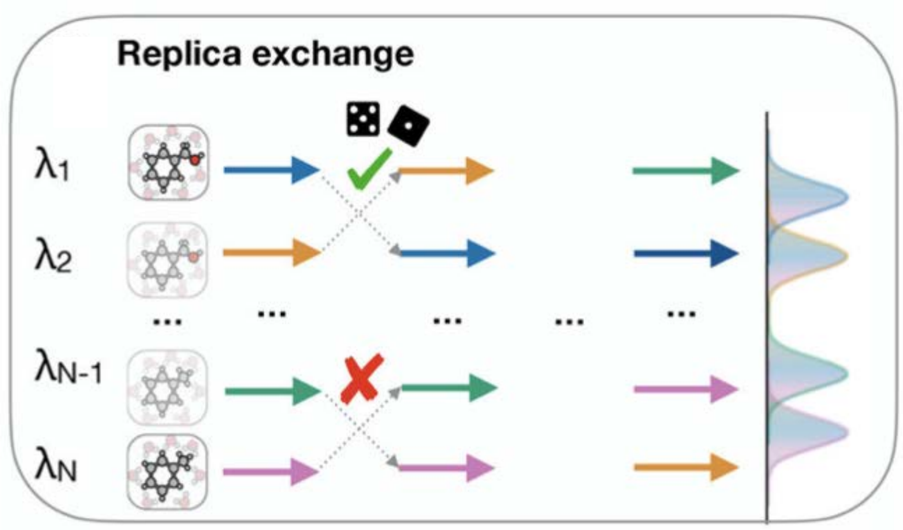


run each alchemical transformation separately

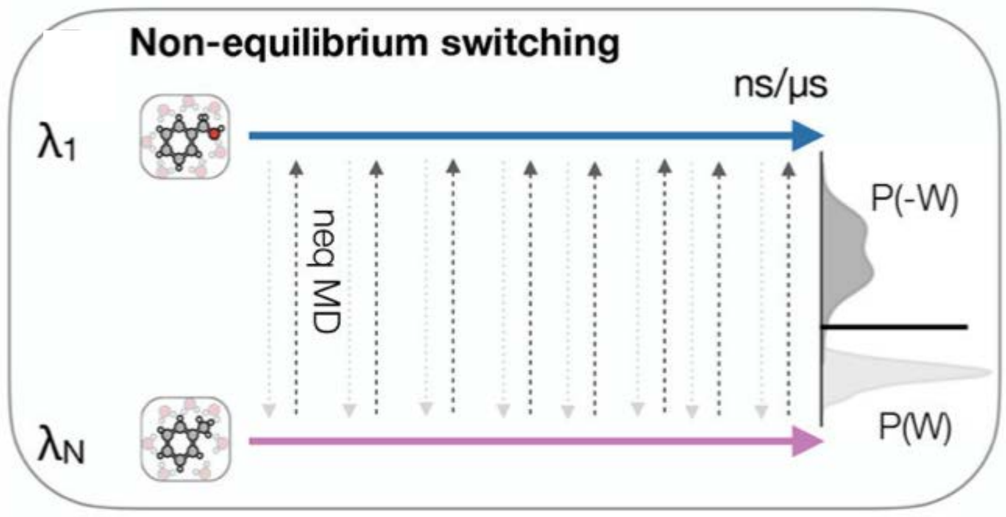
~ 12 GPU-hours per transformation

https://openfree.energy

Current Open Free Energy workflows include best practices and scalable alternatives



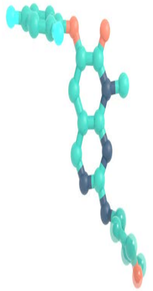
best practices,
limited scalability



highly scalable, slightly less efficient

https://openfree.energy

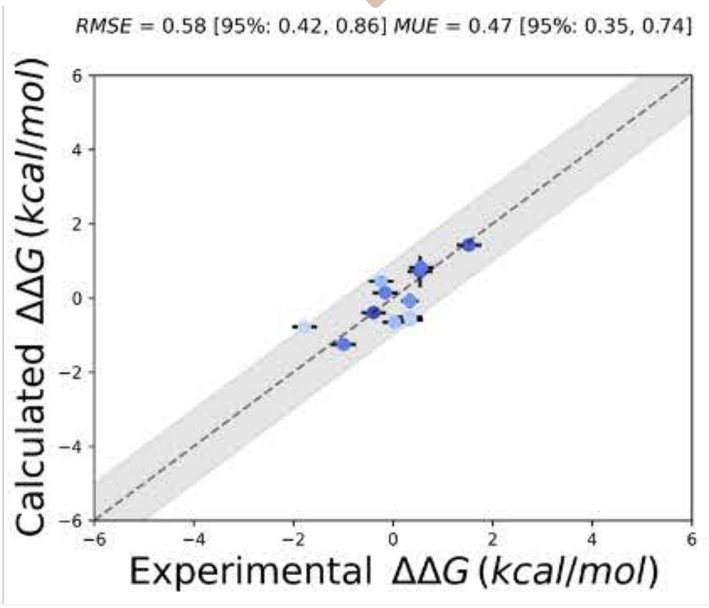
Open Free Energy tools include best practices for data analysis and retrospective statistical analysis



Free energy estimator (PyMBAR)

$$\hat{A}_i = -\beta^{-1} \ln \sum_{j=1}^{K} \sum_{n=1}^{N_j} \frac{\exp\left[-\beta U_i\right]}{\sum_{k=1}^{K} N_k \exp\left[\beta \hat{A}_k - \beta U_k\right]}.$$

MBAR. Chodera and Shirts JCP 129:124105, 2008
https://doi.org/10.1063/1.2978177



Best practices for alchemical free energy calculations.
https://doi.org/10.33011/livecoms.2.1.18378

https://openfree.energy

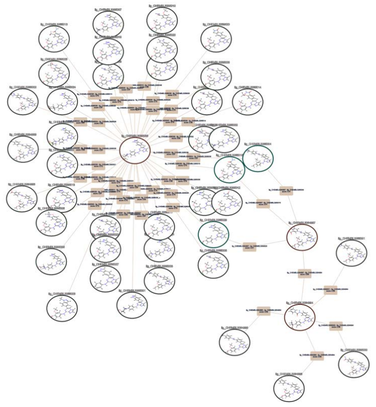
Alchemiscale is an open source infrastructure for orchestrating highly scalable free energy campaigns

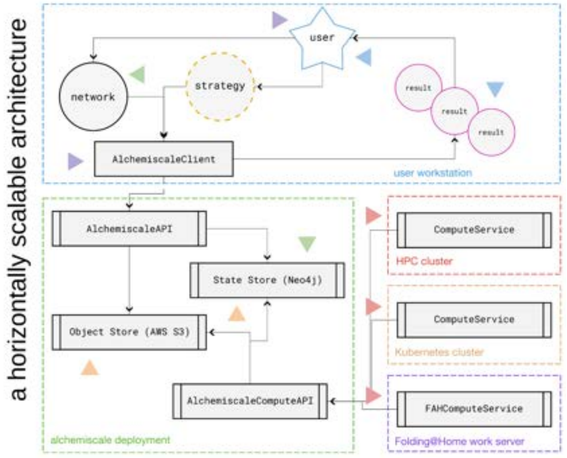
ALCHEMISCALE

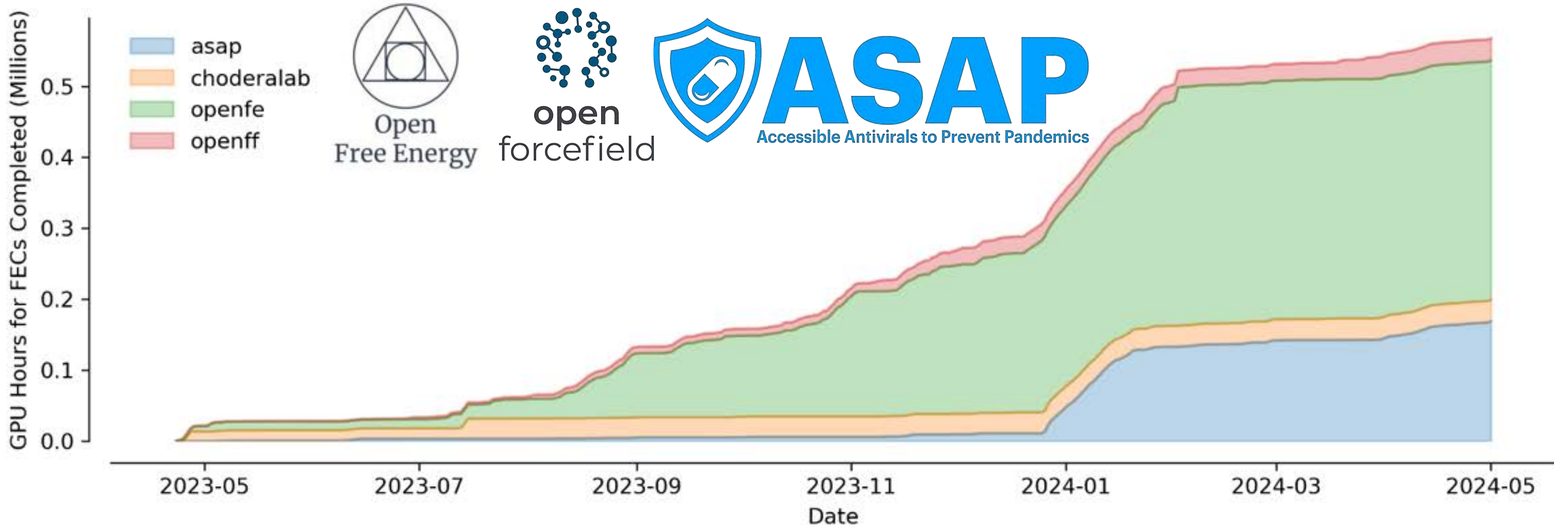
David Dotson

Ian Kenney

Alchemiscale has enabled >0.5 million GPU hours of multi-cluster distributed OpenFE usage in the last year





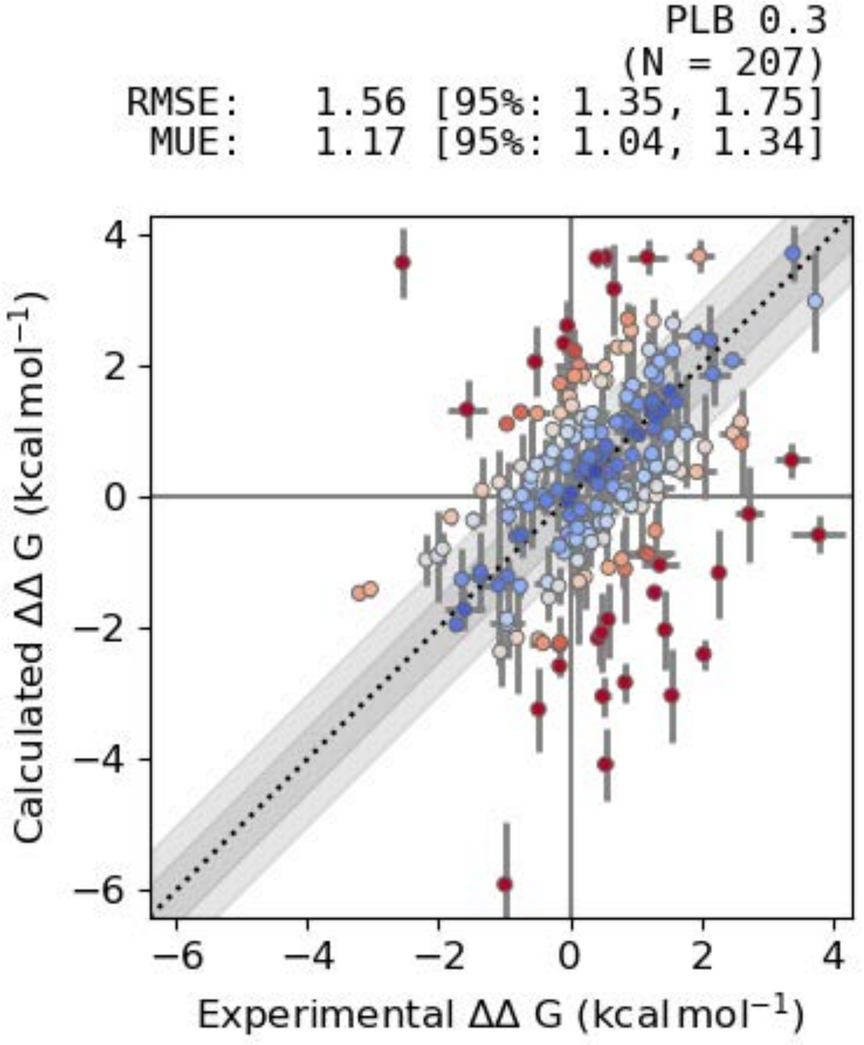


http://alchemiscale.org

Continuous benchmarking on standardized benchmark sets helps ensure systematic improvement

Continuous benchmarking on standard benchmark datasets is used to validate

- new features
- ecosystem changes
- applicability on pharmaceutically relevant systems




HOME / ARCHIVES / VOL. 4 NO. 1 (2022) / Articles

Best Practices for Constructing, Preparing, and Evaluating Protein-Ligand Binding Affinity Benchmarks [Article v1.0]

David F. Hahn
Computational Chemistry, Janssen Research & Development, Turnhoutseweg 30, Beerse B-2340, Belgium
https://orcid.org/0000-0003-2830-6880

Christopher I. Bayly
OpenEye Scientific Software, 9 Bisbee Court, Suite D, Santa Fe, NM 87508 USA
https://orcid.org/0000-0001-9145-6457

Melissa L. Boby
Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY 10065 USA
https://orcid.org/0000-0003-1820-2063

Hannah E. Bruce Macdonald
MSD R&D Innovation Centre, 120 Moorgate, London EC2M 6UR, United Kingdom

PDF

ARTICLE CODE REPOSITORY

PUBLISHED

2022-08-30

INFORMATION

For Readers
For Authors
For Librarians

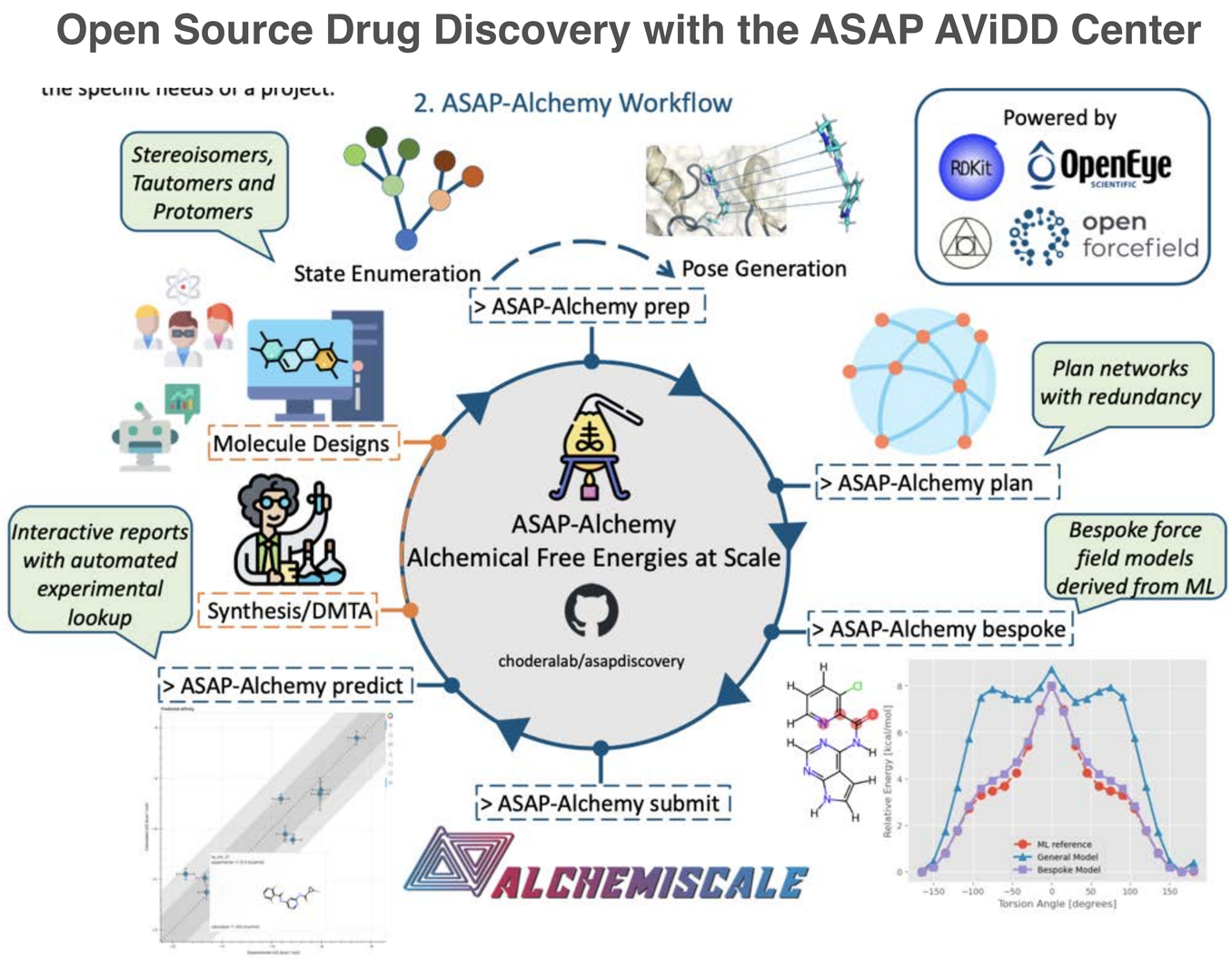
BROWSE

Categories

Best Practices
Tutorials and Training Articles
Editoral
Software Analyses
Perpetual Reviews
Lessons Learned

Best practices for constructing, preparing, and evaluating protein-ligand binding affinity benchmarks
https://doi.org/10.33011/livecoms.4.1.1497

Open Free Energy makes it easy to develop complete workflows to support structure-based drug discovery



https://github.com/asapdiscovery/asapdiscovery

Josh Horton

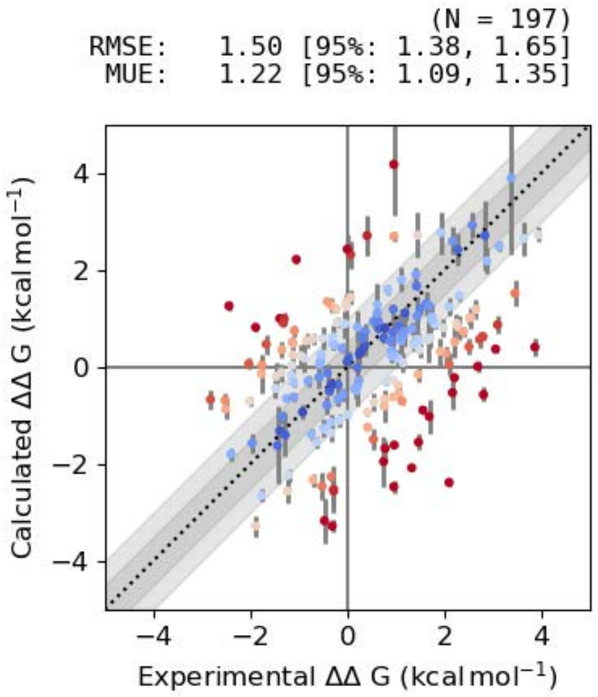
ASAP
Accessible Antivirals to Prevent Pandemics

Jenke Scheen

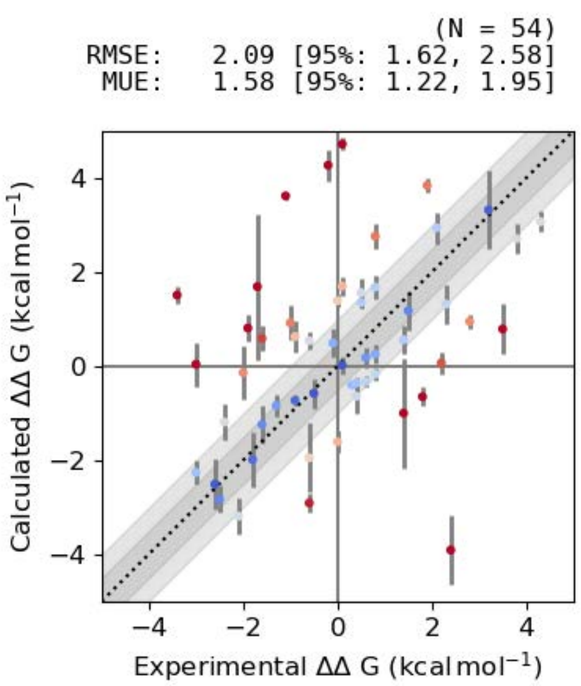
Hugo MacDermott-Opeskin

Open Free Energy tools are showing utility in real industry and academic drug discovery campaigns

Merck KGaA Inhouse Project G



Merck KGaA Inhouse Project T



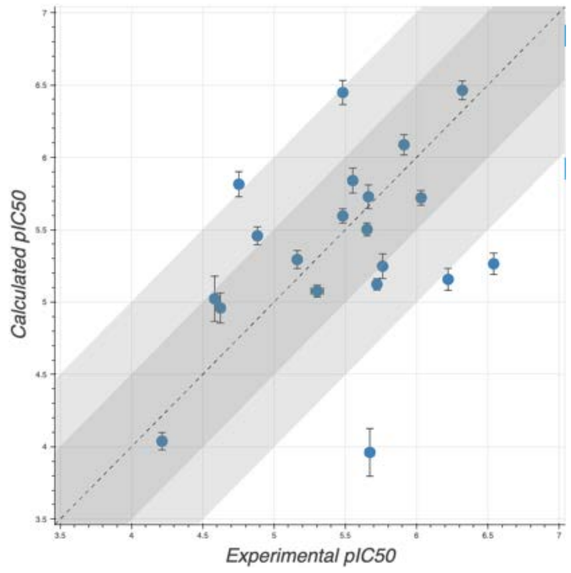
Christina Schindler

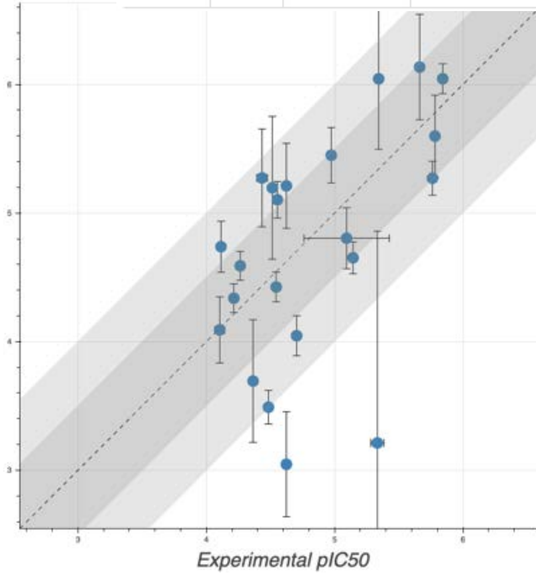
http://openfree.energy

Antiviral drug discovery with the ASAP AViDD Center

ASAP
Accessible Antivirals to Prevent Pandemics

Statistic	value	lower bound	upper bound
RMSE	0.6847	0.4348	0.8899
MUE	0.5160	0.3301	0.7237
R2	0.1533	0.0008	0.5673
rho	0.3915	-0.0896	0.7507

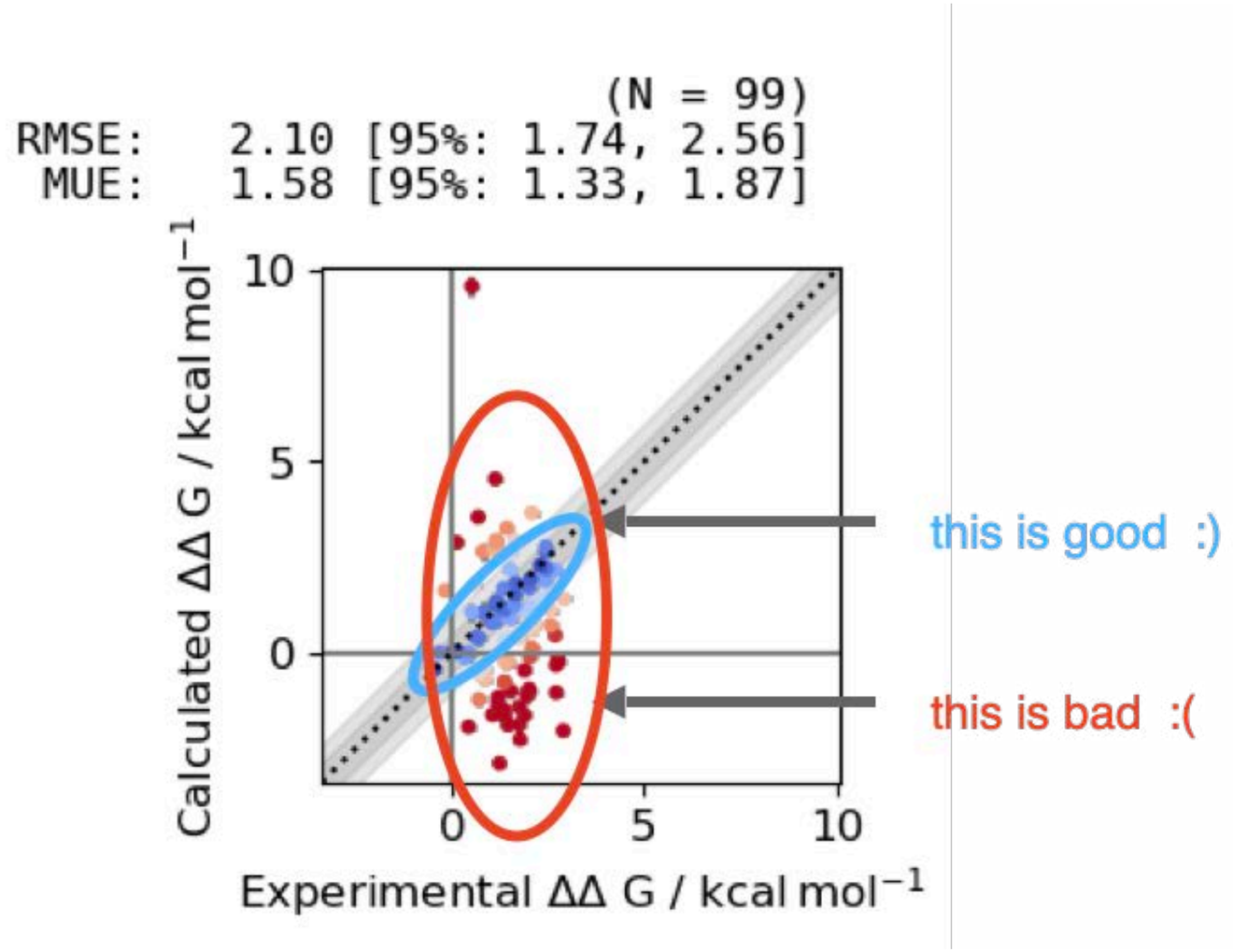




SARS-CoV-2 Mpro

http://asapdiscovery.org

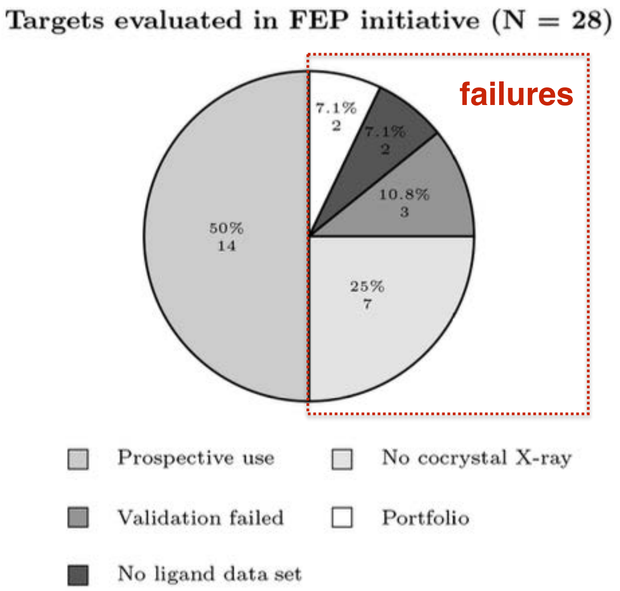
HOWEVER, SOMETIMES, ACCURACY CAN BE POOR

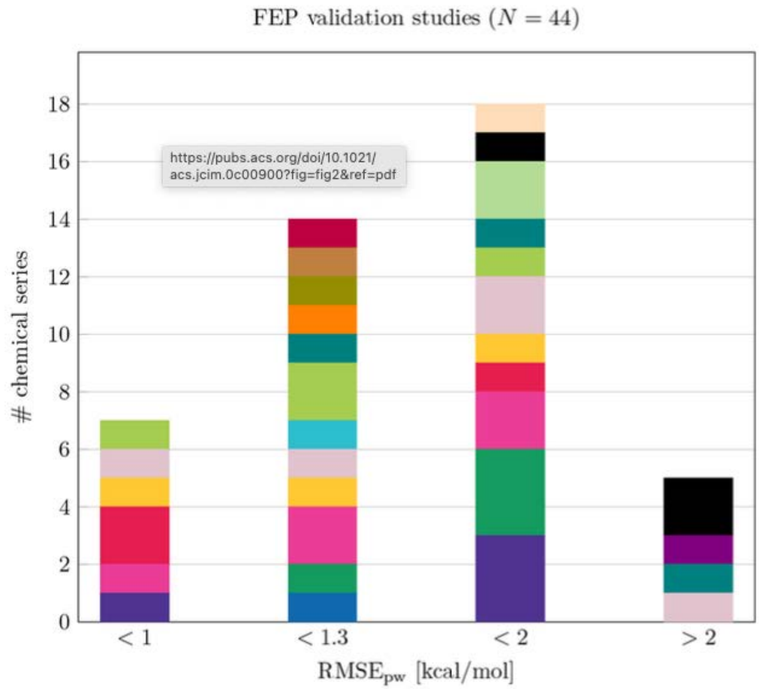


ASAP
Accessible Antivirals to Prevent Pandemics

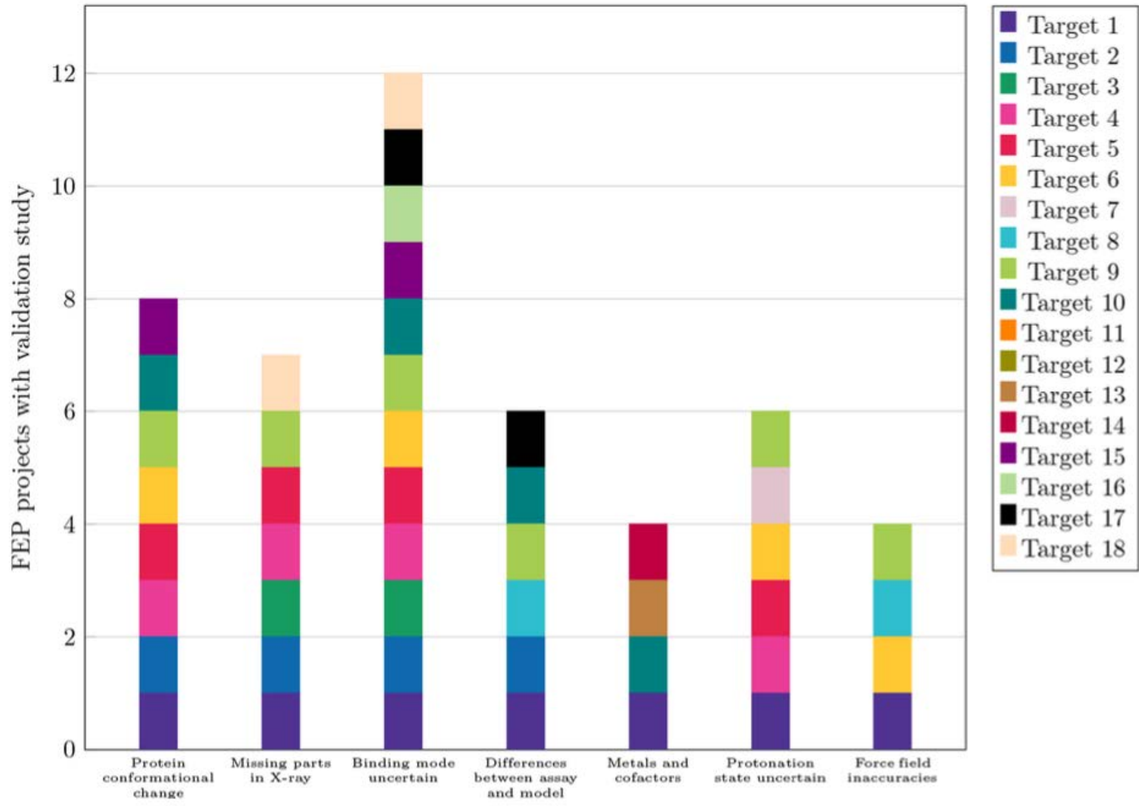
https://fah-public-data-covid19-moonshot-sprints.s3.us-east-2.amazonaws.com/dashboards/sprint-9/sprint-9-2021-06-14-x10959-dimer-neutral-unrestrained/retrospective_transformations/index.html

NUMEROUS FACTORS CONTRIBUTE TO LACK OF APPLICABILITY, ACCURACY, OR RELIABILITY





errors are frequently large



current calculations face a multitude of challenges

Christina Schindler *et al.*
Large-Scale Assessment of Binding Free Energy Calculations in Active Drug Discovery Projects
https://doi.org/10.1021/acs.jcim.0c00900

THE DOMAIN OF APPLICABILITY OF FREE ENERGY CALCULATIONS IS CURRENTLY LIMITED

Multiple high-quality crystal structures of target

Congeneric series of ligands with
all ligands binding in same pose

Only one dominant protonation state
unchanged throughout binding process

No ligand or sidechain tautomerism

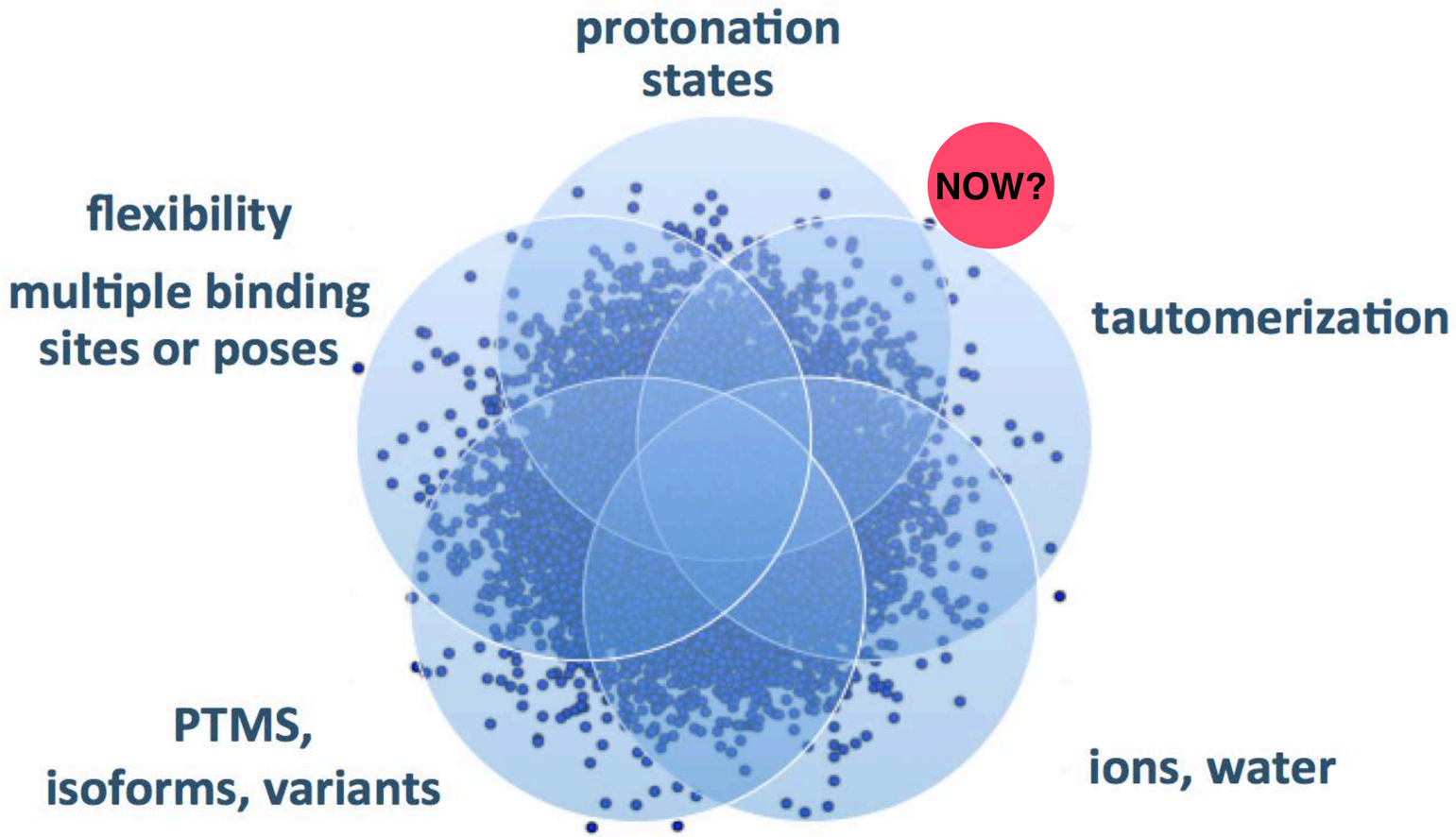
One well-specified, well-resolved isoform/species

No complex cosolvents, binding partners,
slow binding site desolvation events

No exotic chemistries

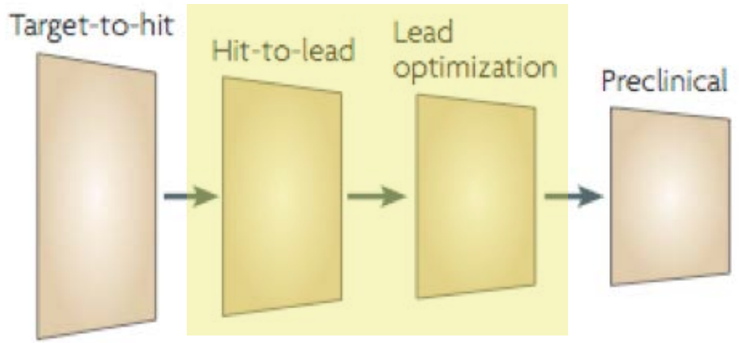
No metals or prosthetic groups

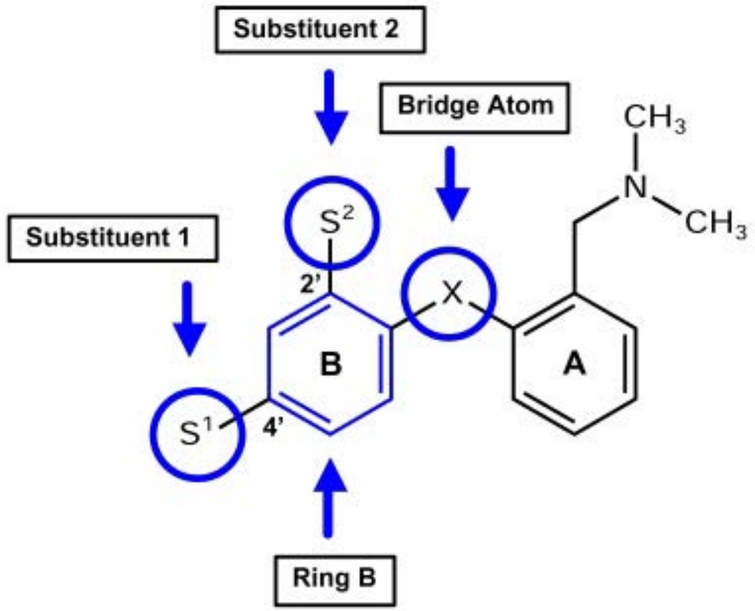
No membranes?

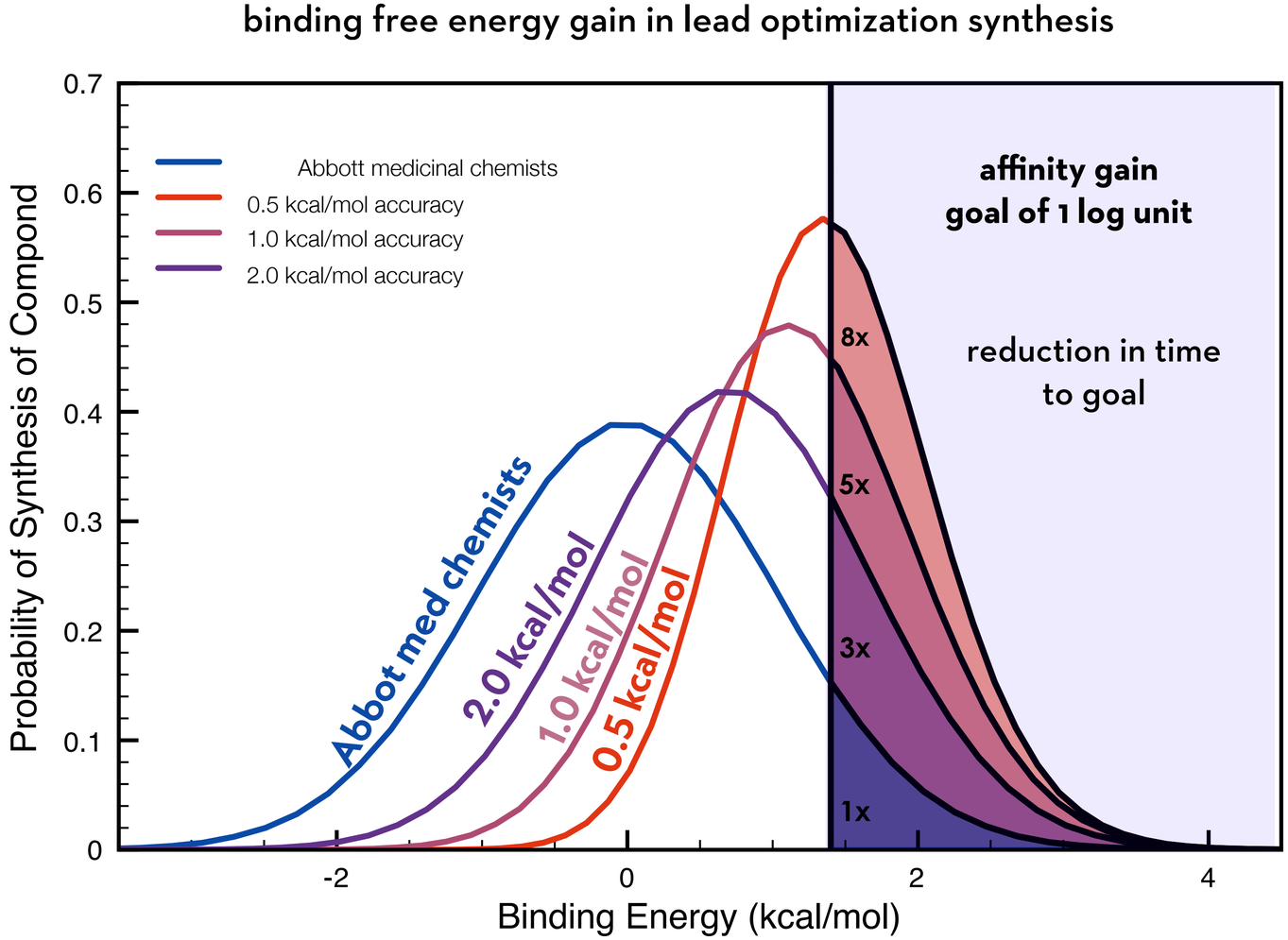


TARGETS
CHALLENGES

MUCH GREATER IMPACT IS POSSIBLE IF WE COULD RELIABLY REDUCE ERROR IN PREDICTIVE MODELS



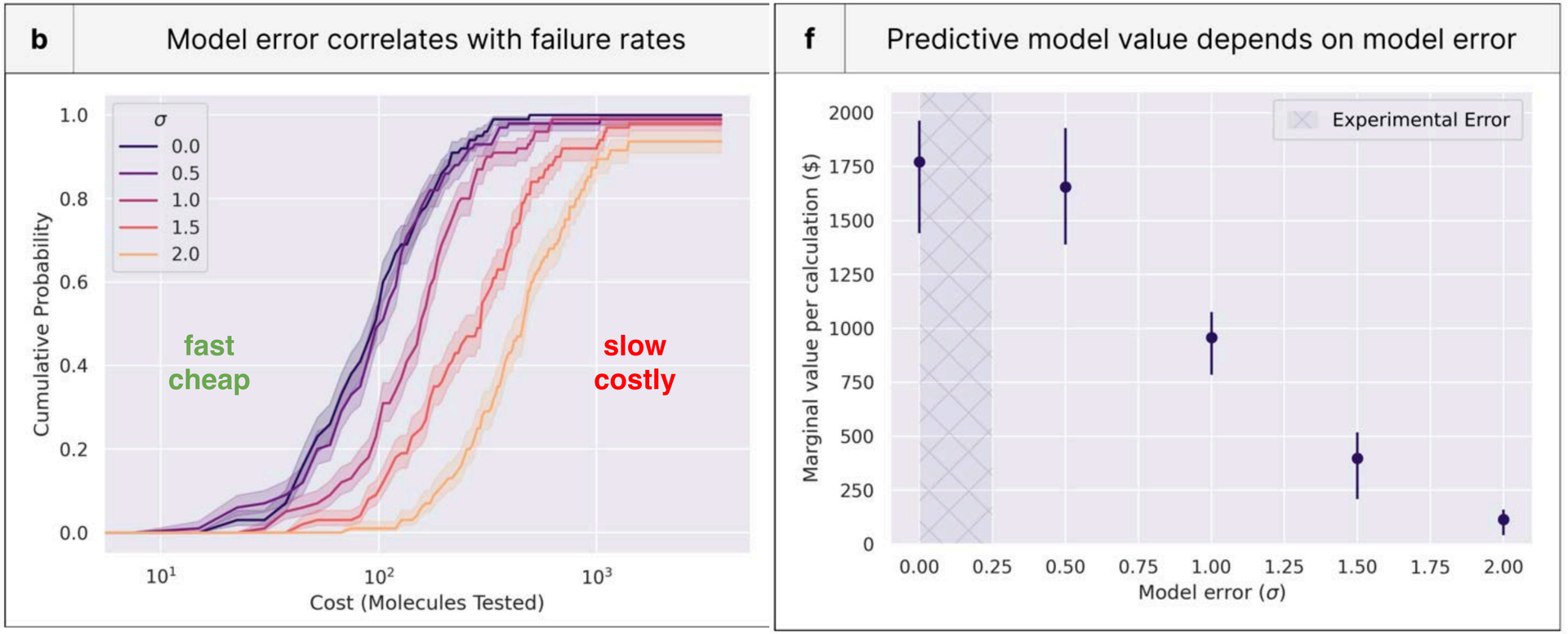




M. R. Shirts, D. L. Mobley and Scott P. Brown. "Free energy calculations in structure-based drug design", in Drug Design: Structure- and Ligand-Based Approaches, pgs. 61-86, 2010.

INCREASING ACCURACY TO RIVAL EXPERIMENTAL ERROR WOULD SHIFT HUGE AMOUNTS OF $ FROM CHEMISTRY TO COMPUTE

Modeling hit-to-lead phase of drug discovery



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

WHAT'S HOLDING US BACK?


EAGLE GT
GOODYEAR
EAGLE GT
GOODYEAR

THE MOLECULAR MECHANICS FORCE FIELDS WE USE TODAY ARE FUNDAMENTALLY OVER 40 YEARS OLD

J. Am. Chem. Soc. **1984**, *106*, 765–784

A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins

Scott J. Weiner, Peter A. Kollman,* David A. Case,† U. Chandra Singh, Caterina Ghio,‡ Giuliano Alagona,‡ Salvatore Profeta, Jr.,§ and Paul Weiner⊥

Contribution from the Department of Pharmaceutical Chemistry, University of California, San Francisco, California 94143. *Received April 28, 1983*

$$E_{\text{total}} = \sum_{\text{bonds}} K_r (r - r_{\text{eq}})^2 + \sum_{\text{angles}} K_\theta (\theta - \theta_{\text{eq}})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}} \right] + \sum_{\text{H-bonds}} \left[\frac{C_{ij}}{R_{ij}^{12}} - \frac{D_{ij}}{R_{ij}^{10}} \right]$$

Weiner et al. JACS 106:765, 1984
https://doi.org/10.1021/ja00315a051

MOLECULAR MECHANICS FORCE FIELDS WERE DEVELOPED FOR THINGS CALLED “MINICOMPUTERS”

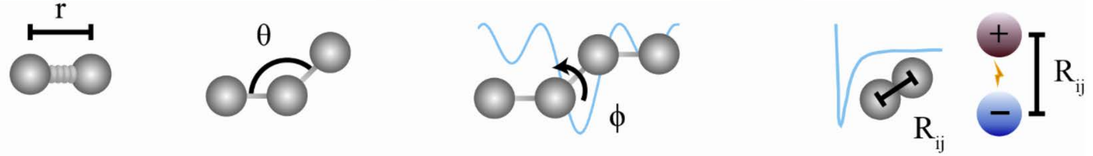
DEC PDP-11

~45 years old

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

Bonded Non-bonded

$$E_{total} = \sum_{bonds} K_r(r - r_{eq})^2 + \sum_{angles} K_\theta(\theta - \theta_{eq})^2 + \sum_{dihedrals} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}} \right]$$



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.**

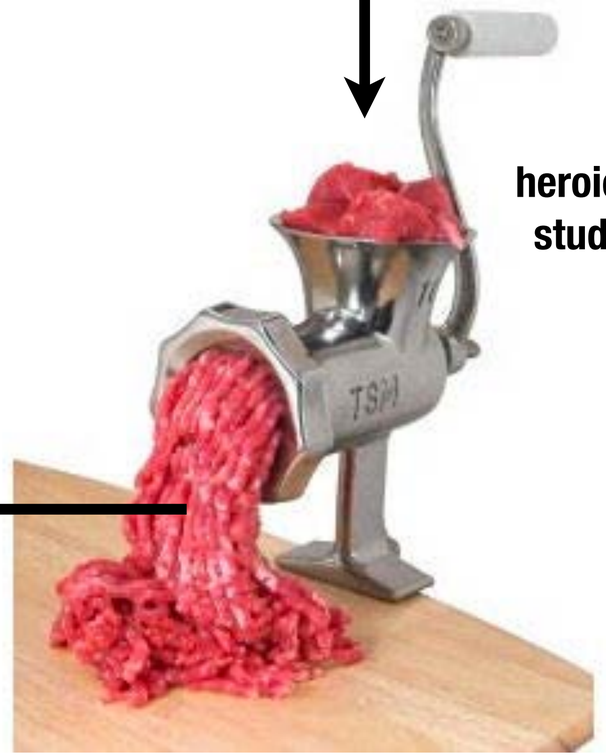
$1599 MSRP*

Why not put them to work?

* A new PDP-11 would have cost $160,000 in today’s dollars!

HOW ARE FORCEFIELDS MADE?

experimental data
quantum chemistry
keen chemical intuition

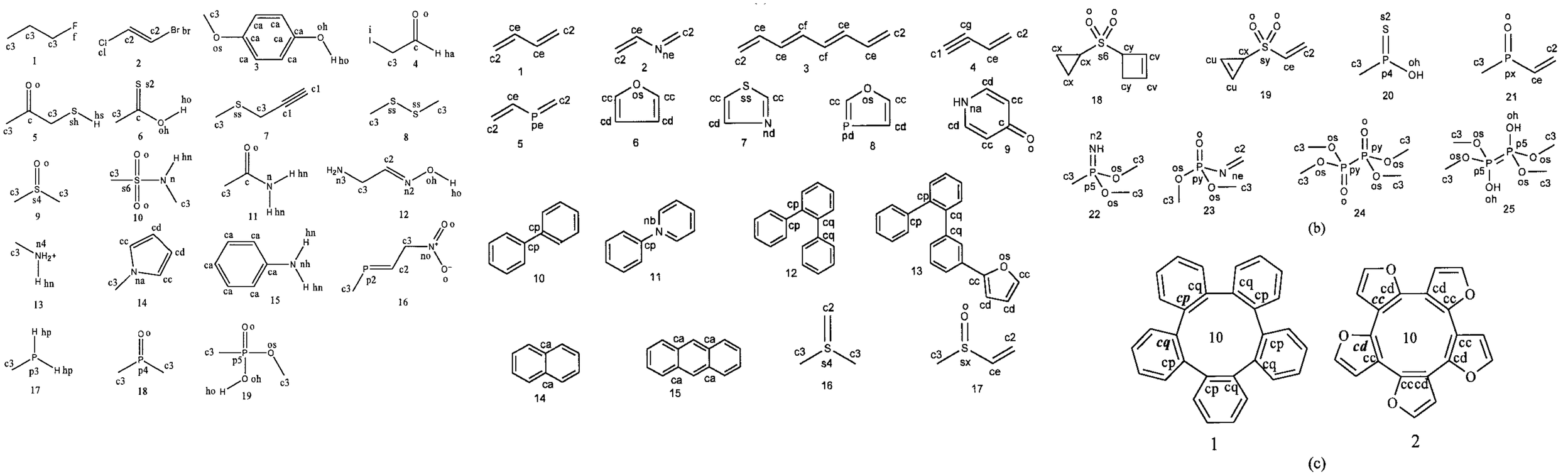


heroic effort by graduate students and postdocs

a parameter set we desperately hope someone actually uses

AS DRUG DISCOVERY EXPLORES NEW PARTS OF CHEMICAL SPACE, HOW CAN FORCEFIELDS KEEP UP?

The Generalized Amber Forcefield (GAFF) was parameterized with this chemical universe:



Extension of this universe is nontrivial because parameter fitting code never released!

Wang J, Wolf RM, Caldwell JW, Kollman PA, and Case DA. J Comput Chem 25:1157, 2004.

CAN WE MAKE BUILDING BIMOLECULAR FORCE FIELDS AS EASY AS TRAINING A MACHINE LEARNING MODEL?

training a neural network

```python
import tensorflow as tf
mnist = tf.keras.datasets.mnist

(x_train, y_train),(x_test, y_test) = mnist.load_data()
x_train, x_test = x_train / 255.0, x_test / 255.0

model = tf.keras.models.Sequential([
  tf.keras.layers.Flatten(input_shape=(28, 28)),
  tf.keras.layers.Dense(128, activation='relu'),
  tf.keras.layers.Dropout(0.2),
  tf.keras.layers.Dense(10, activation='softmax')
])

model.compile(optimizer='adam',
              loss='sparse_categorical_crossentropy',
              metrics=['accuracy'])

model.fit(x_train, y_train, epochs=5)
model.evaluate(x_test, y_test)
```

Run code now Try in Google's interactive notebook

https://www.tensorflow.org/overview

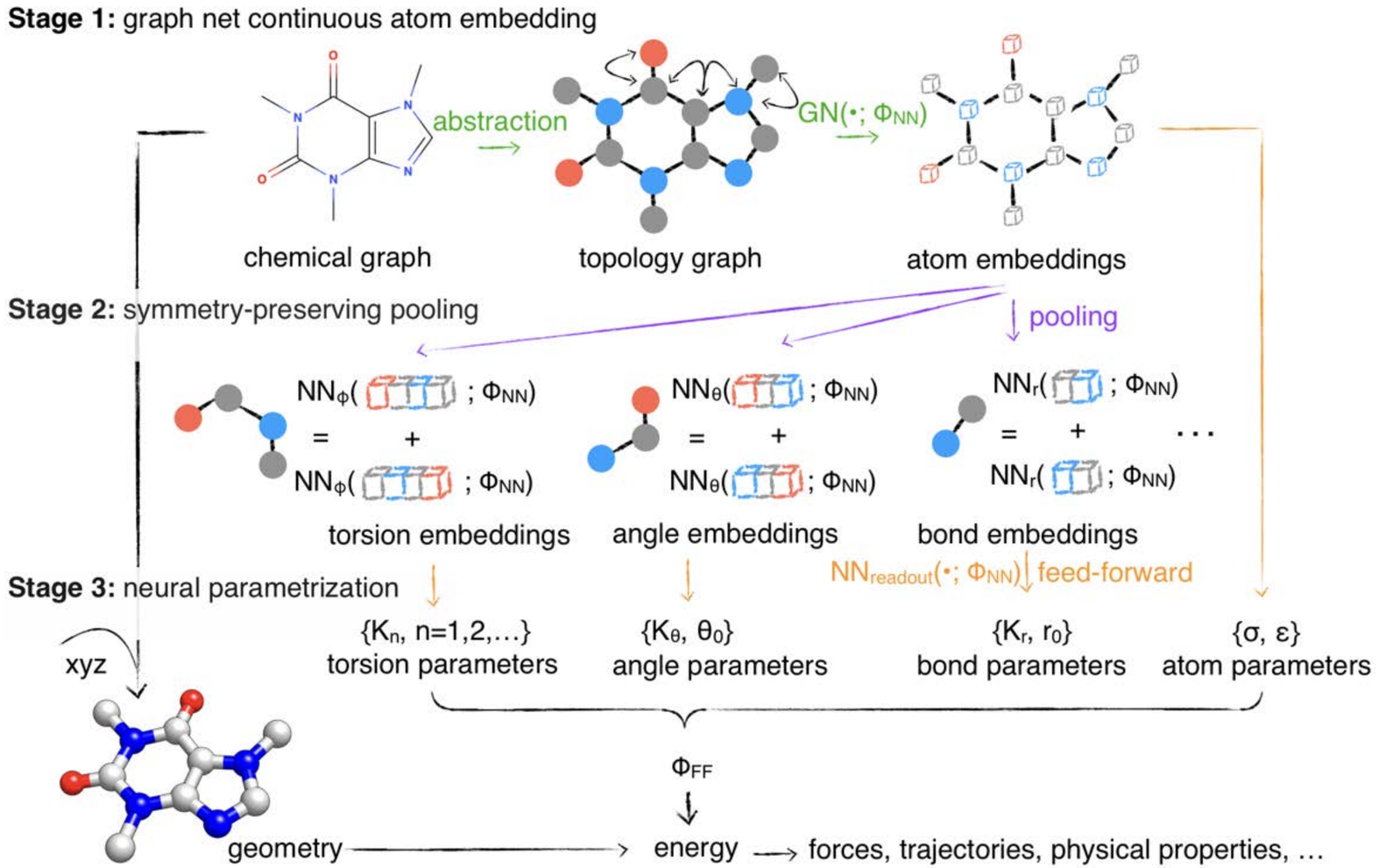
fitting a force field

```python
import openforcefield as off
training_data, benchmark_data = off.datasets.load('2019-Q1')

force_field_model = off.models.ForceFieldModel([
    off.models.forces.HarmonicBond(),
    off.models.forces.HarmonicAngle(),
    off.models.forces.PeriodicTorsion(max_order=6),
    off.models.forces.LennardJones(),
    off.models.forces.BondChargeCorrections(),
])

model.compile(optimizer='L-BFGS',
     loss='error-weighted',
     metrics=['accuracy'])

model.fit(training_data)

model.evaluate(test_data)
```

Run code now Try in Google's interactive notebook

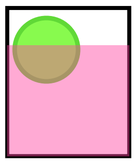
espaloma: **e**xtensible **s**urrogate **p**otential of ***a**b initio* **l**earned and **o**ptimized by **m**essage-passing **a**lgorithm

use of only **chemical graph** means that model can generate parameters for small molecules, proteins, nucleic acids, covalent ligands, carbohydrates, etc.



JOSH FASS

YUANQING WANG



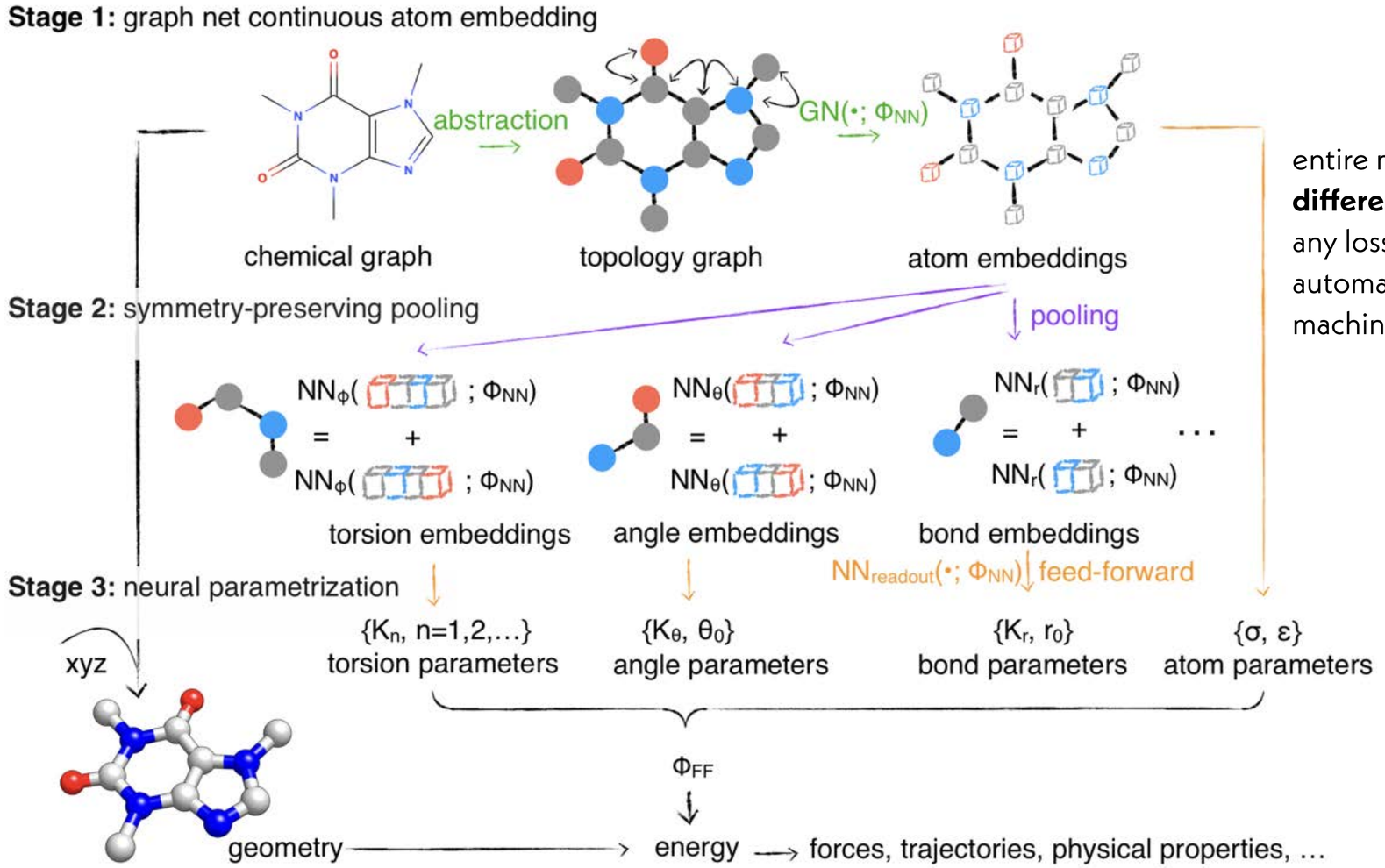
preprint: https://arxiv.org/abs/2010.01196
code: https://github.com/choderalab/espaloma

open forcefield initiative

NIH NIGMS

Funded by OpenFF NIH R&D grant

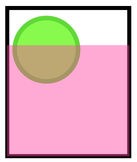
espaloma: **e**xtensible **s**urrogate **p**otential of ***a**b initio* **l**earned and **o**ptimized by **m**essage-passing **a**lgorithm



entire model is **end-to-end differentiable** so can be fit to any loss function by standard automatic differentiation machine learning frameworks

JOSH FASS

YUANQING WANG



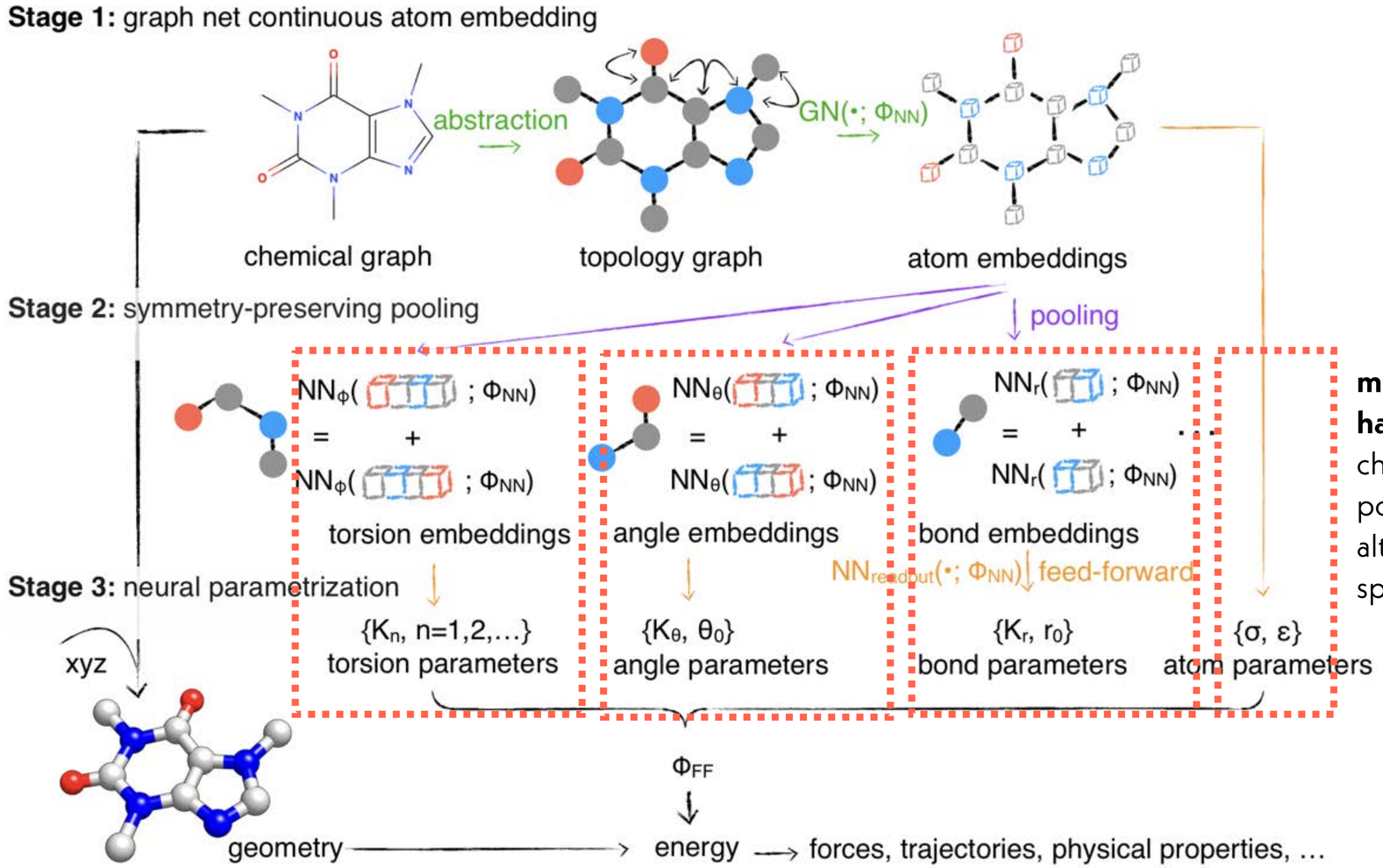
preprint: https://arxiv.org/abs/2010.01196
code: https://github.com/choderalab/espaloma

open forcefield initiative

NIH NIGMS

Funded by OpenFF NIH R&D grant

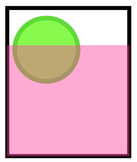
espaloma: **e**xtensible **s**urrogate **p**otential of ***a**b initio* **l**earned and **o**ptimized by **m**essage-passing **a**lgorithm



modular and extensible handling of potential terms: charge model parameters, point polarizabilities, alternative vdW forms, special 1-4 parameters, etc.

JOSH FASS

YUANQING WANG



preprint: https://arxiv.org/abs/2010.01196
code: https://github.com/choderalab/espaloma

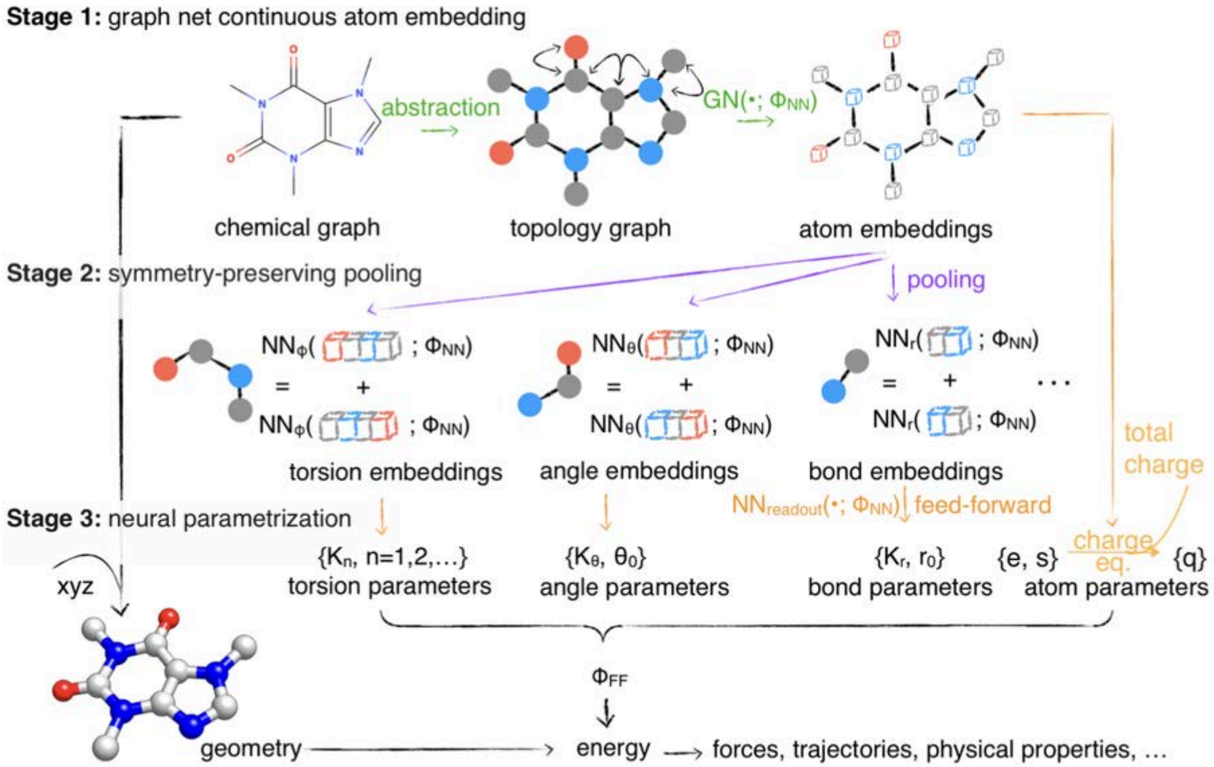
open forcefield initiative

NIH NIGMS

Funded by OpenFF NIH R&D grant

ESPALOMA MAKES BUILDING A NEW FORCE FIELD EASY

espaloma architecture



(implemented in pytorch)

http://github.com/choderalab/espaloma

YUANQING WANG

building a new force field

```python
import torch, dgl, espaloma as esp

# retrieve OpenFF Gen2 Optimization Dataset
dataset = esp.data.dataset.GraphDataset.load("gen2").view(batch_size=128)

# define Espaloma stage I: graph -> atom latent representation
representation = esp.nn.Sequential(
    layer=esp.nn.layers.dgl_legacy.gn("SAGEConv"), # use SAGEConv implementation in DGL
    config=[128, "relu", 128, "relu", 128, "relu"], # 3 layers, 128 units, ReLU activation
)

# define Espaloma stage II and III:
# atom latent representation -> bond, angle, and torsion representation and parameters
readout = esp.nn.readout.janossy.JanossyPooling(
    in_features=128, config=[128, "relu", 128, "relu", 128, "relu"],
    out_features={              # define modular MM parameters Espaloma will assign
        1: {"e": 1, "s": 1}, # atom hardness and electronegativity
        2: {"coefficients": 2}, # bond linear combination
        3: {"coefficients": 3}, # angle linear combination
        4: {"k": 6}, # torsion barrier heights (can be positive or negative)
    },
)

# compose all three Espaloma stages into an end-to-end model
espaloma_model = torch.nn.Sequential(
                 representation, readout,
                 esp.mm.geometry.GeometryInGraph(), esp.mm.energy.EnergyInGraph(),
                 esp.nn.readout.charge_equilibrium.ChargeEquilibrium(),
)

# define training metric
metrics = [
    esp.metrics.GraphMetric(
            base_metric=torch.nn.MSELoss(), # use mean-squared error loss
            between=['u', "u_ref"],         # between predicted and QM energies
            level="g", # compare on graph level
    )
    esp.metrics.GraphMetric(
            base_metric=torch.nn.MSELoss(), # use mean-squared error loss
            between=['q', "q_hat"],         # between predicted and reference charges
            level="n1", # compare on node level
    )
]

# fit Espaloma model to training data
results = esp.Train(
    ds_tr=dataset, net=espaloma_model, metrics=metrics,
    device=torch.device('cuda:0'), n_epochs=5000,
    optimizer=lambda net: torch.optim.Adam(net.parameters(), 1e-3), # use Adam optimizer
).run()

torch.save(espaloma_model, "espaloma_model.pt") # save model
```

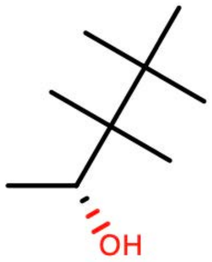
Listing 1. Defining and training a modular Espaloma model.

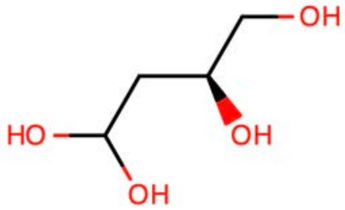
ESPALOMA OUTPERFORMS CURRENT FORCE FIELDS IN QM ACCURACY AND CAN BE EASILY TRAINED FOR HETEROGENEOUS SYSTEMS

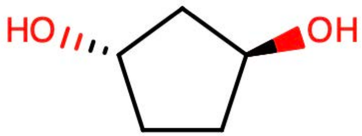
(a) dataset	# mols	# trajs	# snapshots	Espaloma RMSE		Legacy FF RMSE (kcal/mol) (Test molecules)			
				Train	Test	OpenFF 1.2.0	GAFF-1.81	GAFF-2.11	Amber ff14SB
PhAlkEthOH (simple CHO)	7408	12592	244036	$0.8656^{0.9131}_{0.8225}$	$1.1398^{1.2332}_{1.0715}$	$1.6071^{1.6915}_{1.5197}$	$1.7267^{1.7935}_{1.6543}$	$1.7406^{1.8148}_{1.6679}$	

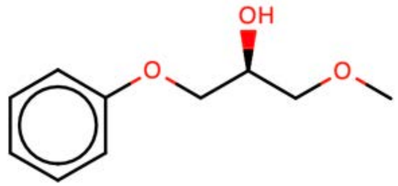
PhAlkEthOh: **Ph**enyls, **Alk**anes, **Eth**ers, and alcohols (**OH**)

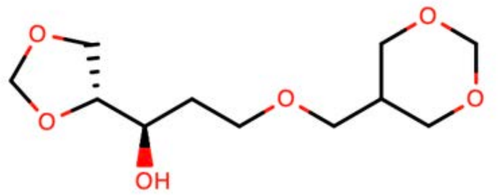
(a low-complexity chemical space)











preprint: https://arxiv.org/abs/2010.01196
code: http://github.com/choderalab/espaloma

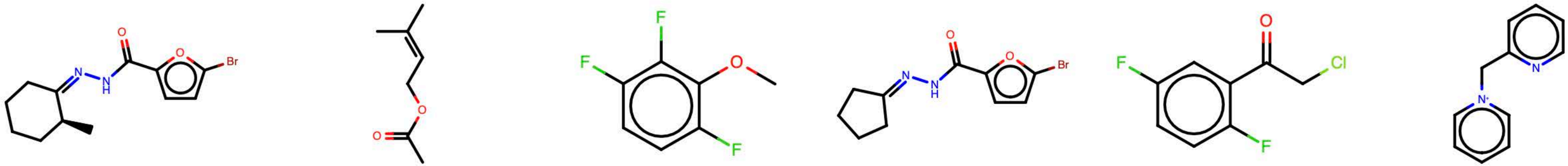
PhAlkEthOH doi: https://dx.doi.org/10.1021/acs.jctc.8b00640

YUANQING WANG

ESPALOMA OUTPERFORMS CURRENT FORCE FIELDS IN QM ACCURACY AND CAN BE EASILY TRAINED FOR HETEROGENEOUS SYSTEMS

(a) dataset	# mols	# trajs	# snapshots	Espaloma RMSE		Legacy FF RMSE (kcal/mol) (Test molecules)			
				Train	Test	OpenFF 1.2.0	GAFF-1.81	GAFF-2.11	Amber ff14SB
PhAlkEthOH (simple CHO)	7408	12592	244036	$0.8656^{0.9131}_{0.8225}$	$1.1398^{1.2332}_{1.0715}$	$1.6071^{1.6915}_{1.5197}$	$1.7267^{1.7935}_{1.6543}$	$1.7406^{1.8148}_{1.6679}$	
OpenFF Gen2 Optimization (druglike)	792	3977	23748	$0.7413^{0.7920}_{0.6914}$	$0.7600^{0.8805}_{0.6644}$	$2.1768^{2.3388}_{2.0380}$	$2.4274^{2.5207}_{2.3300}$	$2.5386^{2.6640}_{2.4370}$	

OpenFF Gen2 Optimization set: Diverse druglike fragments challenging for force fields
(a moderate-complexity chemical space)



preprint: https://arxiv.org/abs/2010.01196
code: http://github.com/choderalab/espaloma

OpenFF Gen2 QM data doi: https://doi.org/10.1021/acs.jctc.1c00571

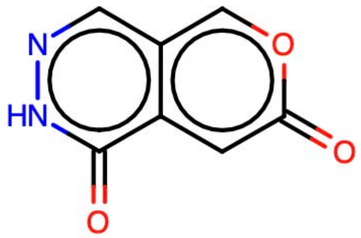
YUANQING WANG

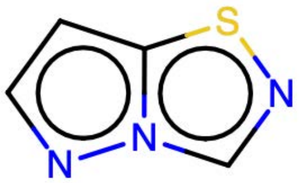
ESPALOMA OUTPERFORMS CURRENT FORCE FIELDS IN QM ACCURACY AND CAN BE EASILY TRAINED FOR HETEROGENEOUS SYSTEMS

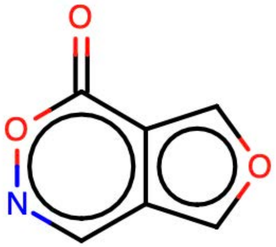
(a) dataset	# mols	# trajs	# snapshots	Espaloma RMSE		Legacy FF RMSE (kcal/mol) (Test molecules)			
				Train	Test	OpenFF 1.2.0	GAFF-1.81	GAFF-2.11	Amber ff14SB
PhAlkEthOH (simple CHO)	7408	12592	244036	$0.8656^{0.9131}_{0.8225}$	$1.1398^{1.2332}_{1.0715}$	$1.6071^{1.6915}_{1.5197}$	$1.7267^{1.7935}_{1.6543}$	$1.7406^{1.8148}_{1.6679}$	
OpenFF Gen2 Optimization (druglike)	792	3977	23748	$0.7413^{0.7920}_{0.6914}$	$0.7600^{0.8805}_{0.6644}$	$2.1768^{2.3388}_{2.0380}$	$2.4274^{2.5207}_{2.3300}$	$2.5386^{2.6640}_{2.4370}$	
VEHICLe (heterocyclic)	24867	24867	234326	$0.4476^{0.4690}_{0.4273}$	$0.4233^{0.4414}_{0.4053}$	$8.0247^{8.2456}_{7.8271}$	$8.0077^{8.2313}_{7.7647}$	$9.4014^{9.6434}_{9.2135}$	

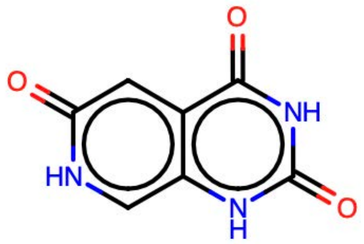
VEHICLe: Virtual exploratory heterocyclic drug scaffold library

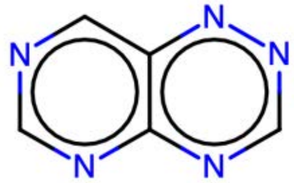
(aromatic bicyclic heterocyclic compounds containing C, N, O, S, H)











preprint: https://arxiv.org/abs/2010.01196

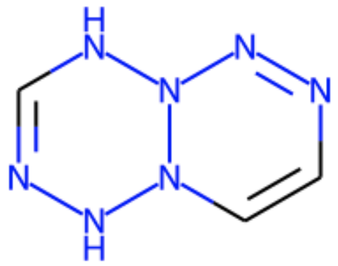
code: http://github.com/choderalab/espaloma

VEHICLe doi: http://doi.org/10.1021/jm801513z

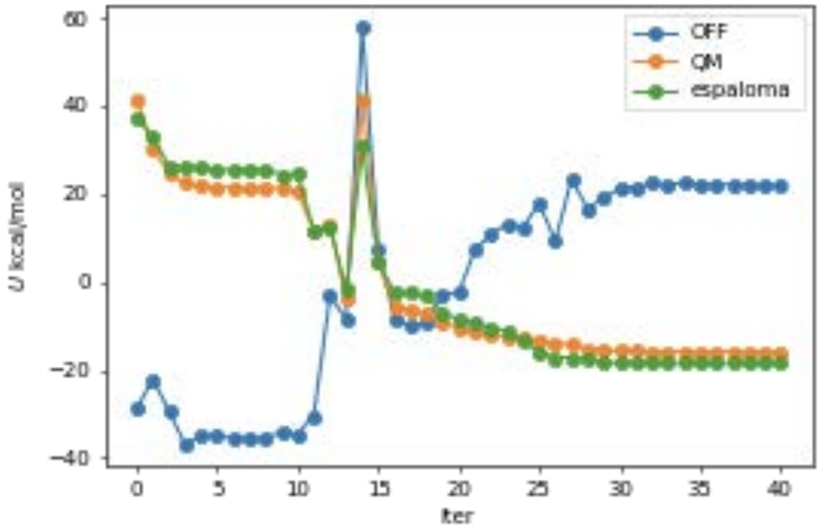
YUANQING WANG

ESPALOMA OUTPERFORMS CURRENT FORCE FIELDS IN QM ACCURACY AND CAN BE EASILY TRAINED FOR HETEROGENEOUS SYSTEMS

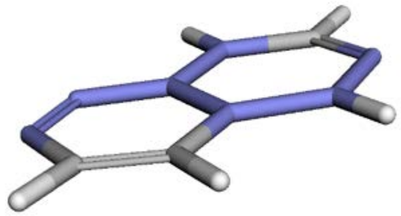
(a) dataset	# mols	# trajs	# snapshots	Espaloma RMSE Train	Espaloma RMSE Test	Legacy FF RMSE (kcal/mol) (Test molecules) OpenFF 1.2.0	GAFF-1.81	GAFF-2.11	Amber ff14SB
PhAlkEthOH (simple CHO)	7408	12592	244036	$0.8656^{0.9131}_{0.8225}$	$1.1398^{1.2332}_{1.0715}$	$1.6071^{1.6915}_{1.5197}$	$1.7267^{1.7935}_{1.6543}$	$1.7406^{1.8148}_{1.6679}$	
OpenFF Gen2 Optimization (druglike)	792	3977	23748	$0.7413^{0.7920}_{0.6914}$	$0.7600^{0.8805}_{0.6644}$	$2.1768^{2.3388}_{2.0380}$	$2.4274^{2.5207}_{2.3300}$	$2.5386^{2.6640}_{2.4370}$	
VEHICLe (heterocyclic)	24867	24867	234326	$0.4476^{0.4690}_{0.4273}$	$0.4233^{0.4414}_{0.4053}$	$8.0247^{8.2456}_{7.8271}$	$8.0077^{8.2313}_{7.7647}$	$9.4014^{9.6434}_{9.2135}$	



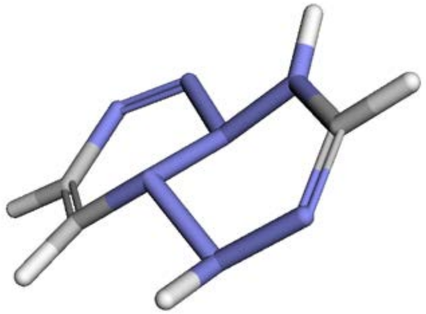
Comparison with QCArchive data



initial



QM minimized



DFT B3LYP-D3(BJ) / DZVP

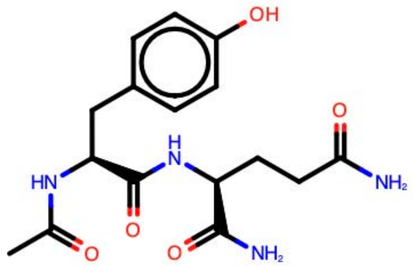
preprint: https://arxiv.org/abs/2010.01196
code: http://github.com/choderalab/espaloma

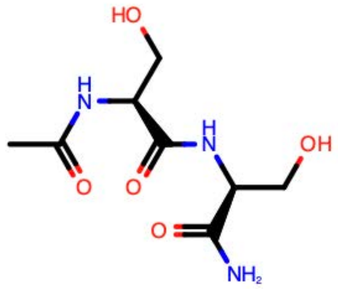
YUANQING WANG

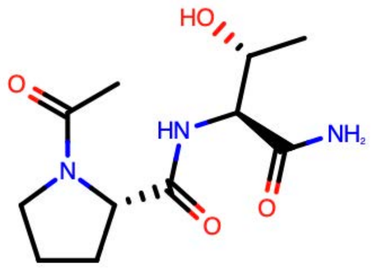
ESPALOMA OUTPERFORMS CURRENT FORCE FIELDS IN QM ACCURACY AND CAN BE EASILY TRAINED FOR HETEROGENEOUS SYSTEMS

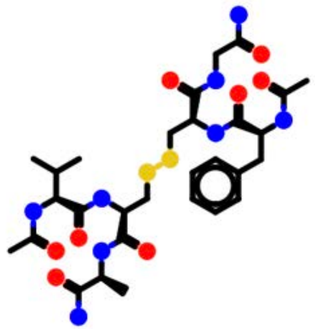
(a) dataset	# mols	# trajs	# snapshots	Espaloma RMSE		Legacy FF RMSE (kcal/mol) (Test molecules)			
				Train	Test	OpenFF 1.2.0	GAFF-1.81	GAFF-2.11	Amber ff14SB
PhAlkEthOH (simple CHO)	7408	12592	244036	$0.8656^{0.9131}_{0.8225}$	$1.1398^{1.2332}_{1.0715}$	$1.6071^{1.6915}_{1.5197}$	$1.7267^{1.7935}_{1.6543}$	$1.7406^{1.8148}_{1.6679}$	
OpenFF Gen2 Optimization (druglike)	792	3977	23748	$0.7413^{0.7920}_{0.6914}$	$0.7600^{0.8805}_{0.6644}$	$2.1768^{2.3388}_{2.0380}$	$2.4274^{2.5207}_{2.3300}$	$2.5386^{2.6640}_{2.4370}$	
VEHICLe (heterocyclic)	24867	24867	234326	$0.4476^{0.4690}_{0.4273}$	$0.4233^{0.4414}_{0.4053}$	$8.0247^{8.2456}_{7.8271}$	$8.0077^{8.2313}_{7.7647}$	$9.4014^{9.6434}_{9.2135}$	
PepConf (peptides)	736	7560	22154	$1.2714^{1.3616}_{1.1899}$	$1.8727^{1.9749}_{1.7309}$	$3.6143^{3.7288}_{3.4870}$	$4.4446^{4.5738}_{4.3386}$	$4.3356^{4.4641}_{4.1965}$	$3.1502^{3.1859,*}_{3.1117}$

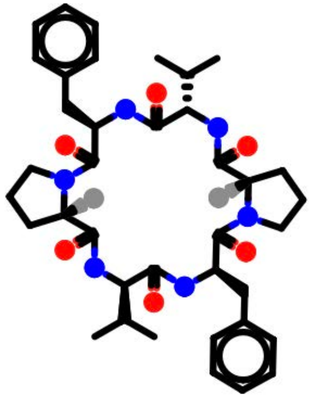
PepConf: Short peptides, including disulfides and cyclic peptides











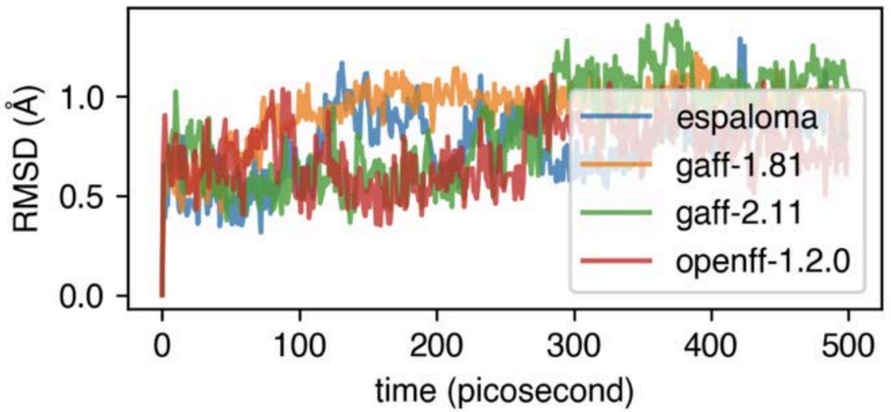
preprint: https://arxiv.org/abs/2010.01196
code: http://github.com/choderalab/espaloma

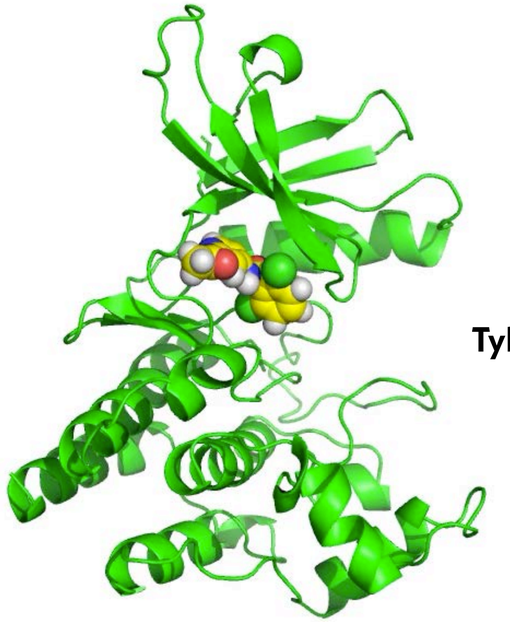
VEHICLe doi: http://doi.org/10.1021/jm801513z

YUANQING WANG

ESPALOMA OUTPERFORMS CURRENT FORCE FIELDS IN QM ACCURACY AND CAN BE EASILY TRAINED FOR HETEROGENEOUS SYSTEMS

	(a) dataset	# mols	# trajs	# snapshots	Espaloma RMSE Train	Espaloma RMSE Test	Legacy FF RMSE (kcal/mol) (Test molecules) OpenFF 1.2.0	GAFF-1.81	GAFF-2.11	Amber ff14SB
	PhAlkEthOH (simple CHO)	7408	12592	244036	$0.8656^{0.9131}_{0.8225}$	$1.1398^{1.2332}_{1.0715}$	$1.6071^{1.6915}_{1.5197}$	$1.7267^{1.7935}_{1.6543}$	$1.7406^{1.8148}_{1.6679}$	
	OpenFF Gen2 Optimization (druglike)	792	3977	23748	$0.7413^{0.7920}_{0.6914}$	$0.7600^{0.8805}_{0.6644}$	$2.1768^{2.3388}_{2.0380}$	$2.4274^{2.5207}_{2.3300}$	$2.5386^{2.6640}_{2.4370}$	
	VEHICLe (heterocyclic)	24867	24867	234326	$0.4476^{0.4690}_{0.4273}$	$0.4233^{0.4414}_{0.4053}$	$8.0247^{8.2456}_{7.8271}$	$8.0077^{8.2313}_{7.7647}$	$9.4014^{9.6434}_{9.2135}$	
	PepConf (peptides)	736	7560	22154	$1.2714^{1.3616}_{1.1899}$	$1.8727^{1.9749}_{1.7309}$	$3.6143^{3.7288}_{3.4870}$	$4.4446^{4.5738}_{4.3386}$	$4.3356^{4.4641}_{4.1965}$	$3.1502^{3.1859,*}_{3.1117}$
joint	OpenFF Gen2 Optimization	1528	11537	45902	$0.8264^{0.9007}_{0.7682}$	$1.8764^{1.9947}_{1.7827}$	$2.1768^{2.3388}_{2.0380}$	$2.4274^{2.5207}_{2.3300}$	$2.5386^{2.6640}_{2.4370}$	
	PepConf				$1.2038^{1.3056}_{1.1178}$	$1.7307^{1.8439}_{1.6053}$	$3.6143^{3.7288}_{3.4870}$	$4.4446^{4.5738}_{4.3386}$	$4.3356^{4.4641}_{4.1965}$	$3.1502^{3.1859,*}_{3.1117}$



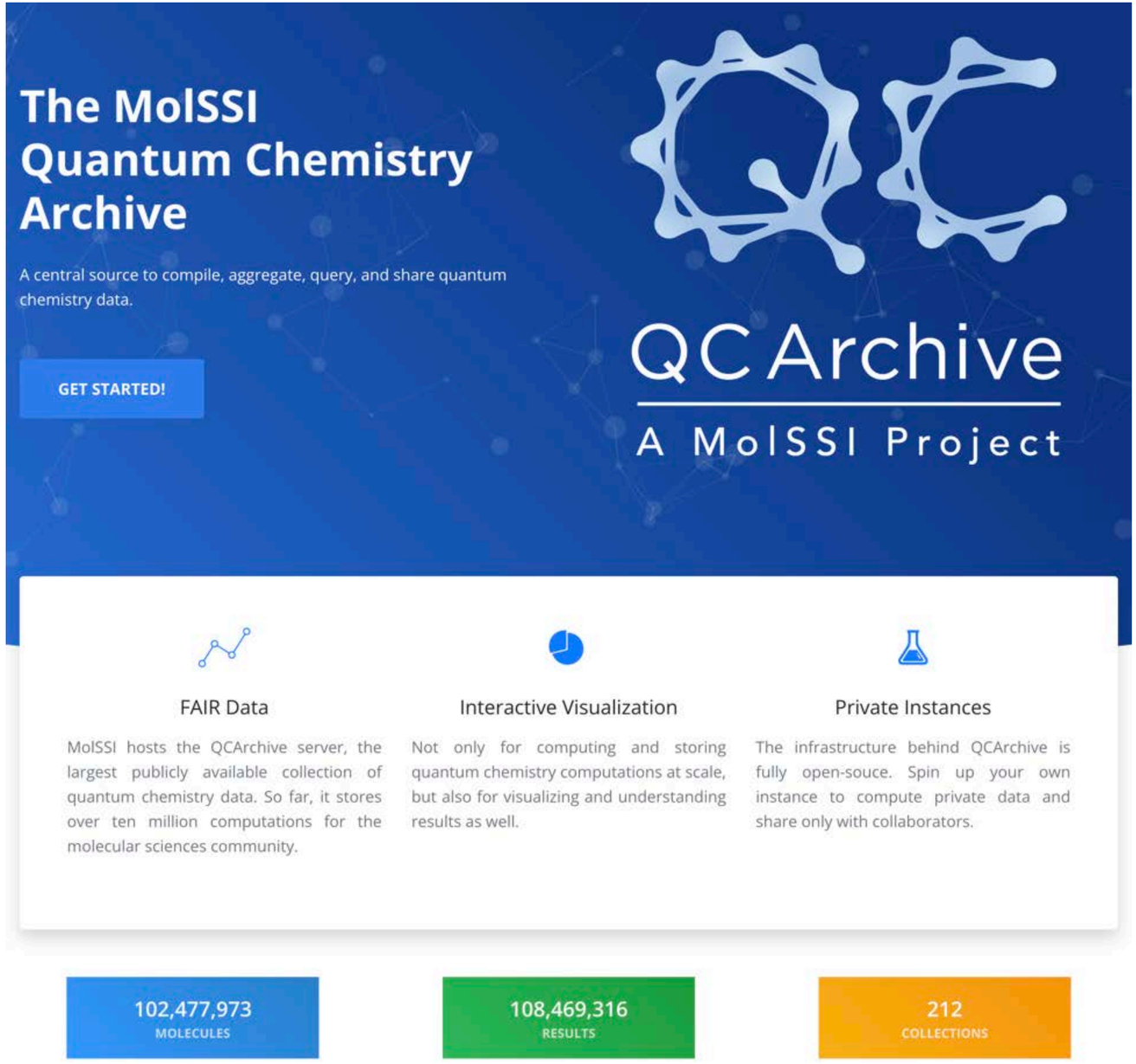


Tyk2 from OpenFF benchmark set
espaloma **joint** model
+ TIP3P water

preprint: https://arxiv.org/abs/2010.01196
code: http://github.com/choderalab/espaloma

Tyk2 benchmark doi: https://doi.org/10.1021/ja512751q

YUANQING WANG



http://qcarchive.molssi.org

SPICE: An open quantum chemical dataset

Subset	Molecules	Conformations	Atoms	Elements
Dipeptides	677	33850	26–60	H, C, N, O, S
Solvated Amino Acids	26	1300	79–96	H, C, N, O, S
DES370K Dimers	3490	345676	2–34	H, Li, C, N, O, F, Na, Mg, P, S, Cl, K, Ca, Br, I
DES370K Monomers	374	18700	3–22	H, C, N, O, F, P, S, Cl, Br, I
PubChem	14643	731856	3–50	H, C, N, O, F, P, S, Cl, Br, I
Ion Pairs	28	1426	2	Li, F, Na, Cl, K, Br, I
Total	19238	1132808	2–96	H, Li, C, N, O, F, Na, Mg, P, S, Cl, K, Ca, Br, I

DFT ωB97M-D3(BJ)/def2-TZVPPD level of theory

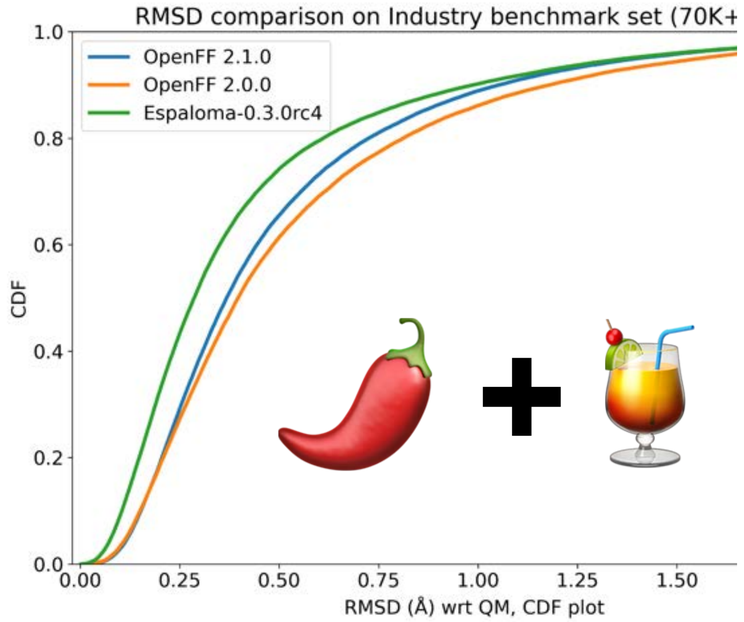
>4M core-hours computed on QCFractal academic clusters

SPICE, A Dataset of Drug-like Molecules and Peptides for Training Machine Learning Potentials

https://doi.org/10.1038/s41597-022-01882-6

https://github.com/openmm/spice-dataset

ESPALOMA OPENS THE DOOR TO RAPIDLY TRAINING OR FINE-TUNING MM POTENTIALS WITH LARGE DATASETS

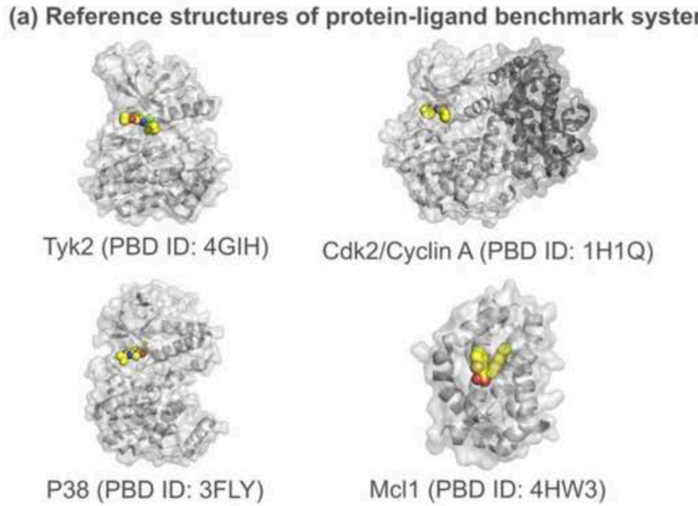


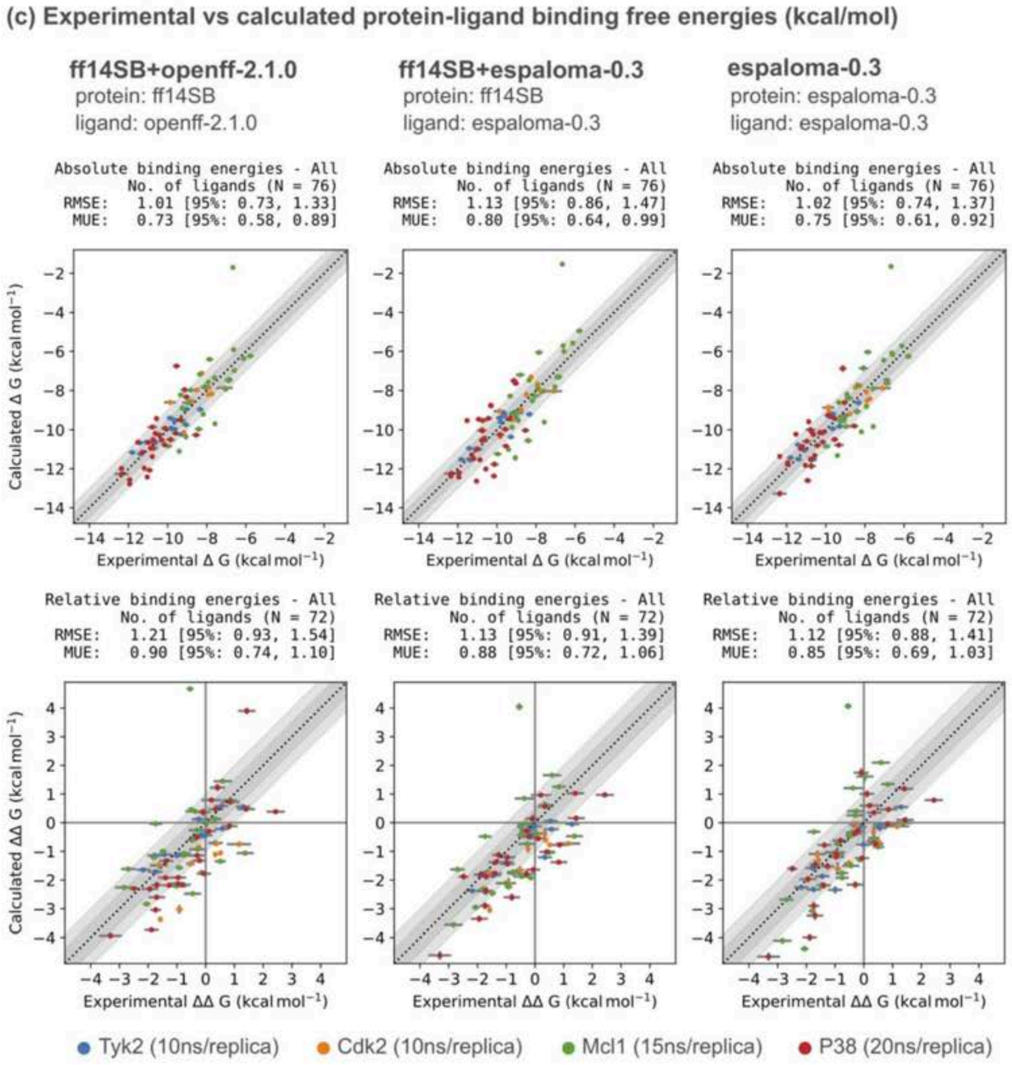
KEN TAKABA

YUANQING WANG

PAVAN BEHARA

IVÁN PULIDO



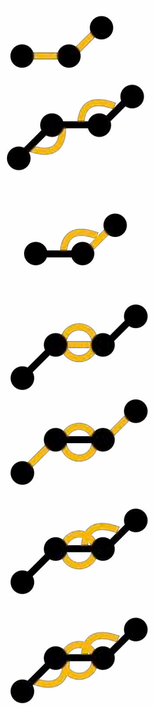


Ken Takaba et al. **Machine-learned molecular mechanics force fields from large-scale quantum chemical data.** https://doi.org/10.1039/D4SC00690A

HOW DO WE INCREASE ACCURACY?
WE COULD GO BACK TO CLASS II FORCE FIELDS...

$$
\begin{aligned}
E = & \sum_b [{}^2K_b(b-b_0)^2 + {}^3K_b(b-b_0)^3 + {}^4K_b(b-b_0)^4] \\
& + \sum_\theta [{}^2K_\theta(\theta-\theta_0)^2 + {}^3K_\theta(\theta-\theta_0)^3 + {}^4K_\theta(\theta-\theta_0)^4] \\
& + \sum_\phi [{}^1K_\phi(1-\cos\phi) + {}^2K_\phi(1-\cos 2\phi) + {}^3K_\phi(1-\cos 3\phi)] \\
& + \sum_\chi K_\chi \chi^2 + \sum_{i>j} \frac{q_i q_j}{r_{ij}} + \sum_{i>j} \epsilon \left[2\left(\frac{r^*}{r_{ij}}\right)^9 - 3\left(\frac{r^*}{r_{ij}}\right)^6 \right] \\
& + \sum_b \sum_{b'} K_{bb'}(b-b_0)(b'-b'_0) + \sum_\theta \sum_{\theta'} K_{\theta\theta'}(\theta-\theta_0) \times \\
& (\theta'-\theta'_0) \\
& + \sum_b \sum_\theta K_{b\theta}(b-b_0)(\theta-\theta_0) \\
& + \sum_\phi \sum_b (b-b_0)[{}^1K_{\phi b}\cos\phi + {}^2K_{\phi b}\cos 2\phi + {}^3K_{\phi b}\cos 3\phi] \\
& + \sum_\phi \sum_{b'} (b'-b'_0)[{}^1K_{\phi b'}\cos\phi + {}^2K_{\phi b'}\cos 2\phi + \\
& {}^3K_{\phi b'}\cos 3\phi] \\
& + \sum_\phi \sum_\theta (\theta-\theta_0)[{}^1K_{\phi\theta}\cos\phi + {}^2K_{\phi\theta}\cos 2\phi + {}^3K_{\phi\theta}\cos 3\phi] \\
& + \sum_\phi \sum_\theta \sum_{\theta'} K_{\phi\theta\theta'}(\theta-\theta_0)(\theta'-\theta'_0)\cos\phi \qquad (1)
\end{aligned}
$$

- bond-bond: angle node
- angle-angle: torsion node
- bond-angle: angle node
- torsion-(center) bond: torsion
- torsion-(side) bond: torsion
- torsion-angle: torsion
- torsion-angle-angle: torsion

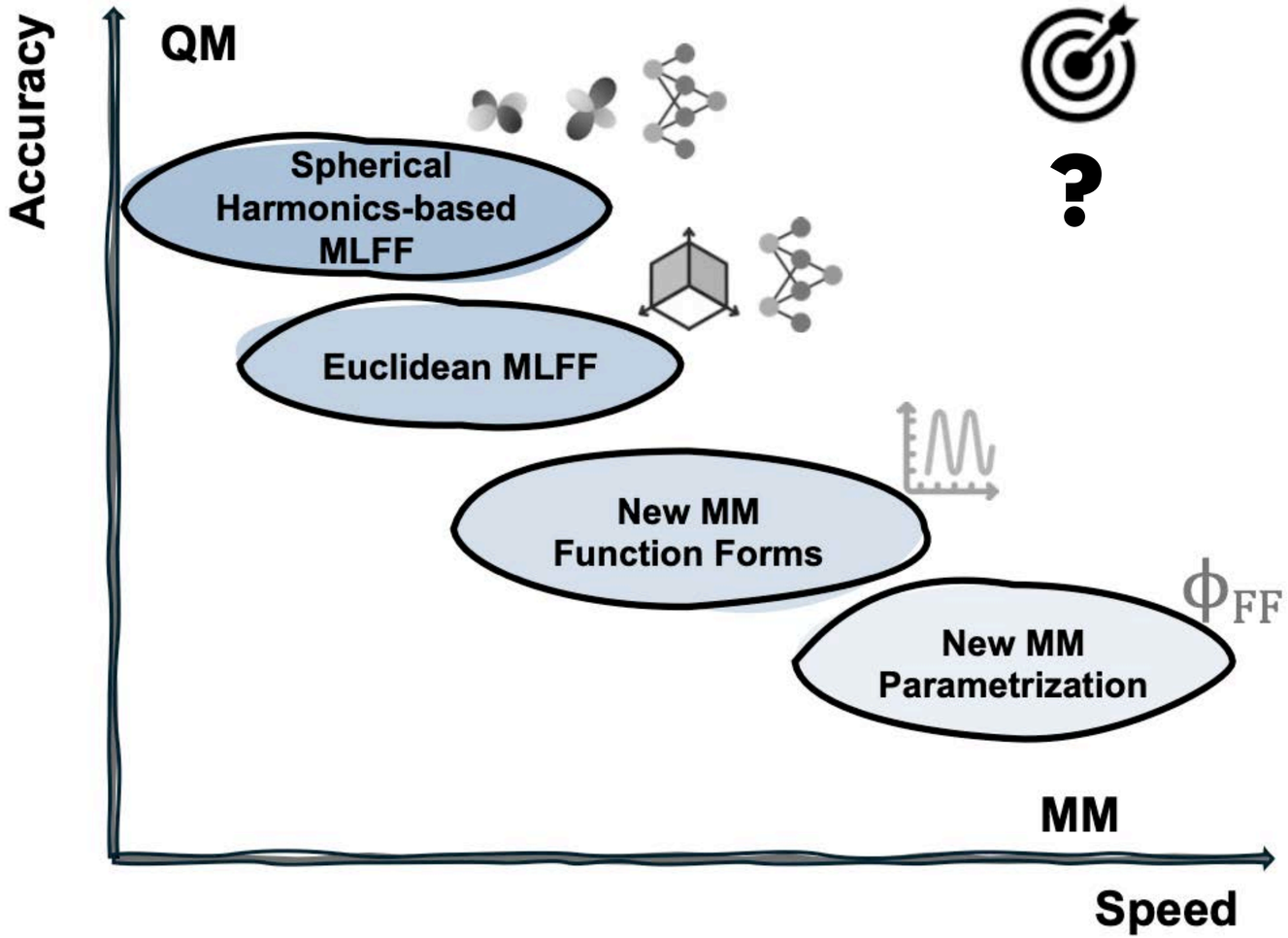


BUT THE NUMBER OF TERMS EXPLODES COMBINATORIALLY

Can we do a better job of modeling true many-body local valence terms, and set ourselves up to solve the other challenges too?

Hwang et al. (1994) http://doi.org/10.1021/ja00085a036

THERE IS AN MODEL CONTINUUM BETWEEN MOLECULAR MECHANICS AND QUANTUM CHEMISTRY, BUT IS THERE SOMETHING EVEN BETTER?



https://arxiv.org/abs/2409.01931

A NEW GENERATION OF NEURAL NETWORK POTENTIALS PROVIDE SIGNIFICANTLY MORE FLEXIBILITY IN LEARNING MULTIBODY INTERACTIONS, THOUGH AT MUCH GREATER COST

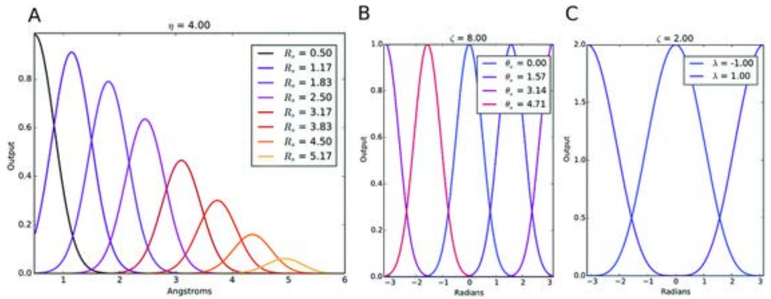
ANI family of neural network potentials

radial and **angular** features

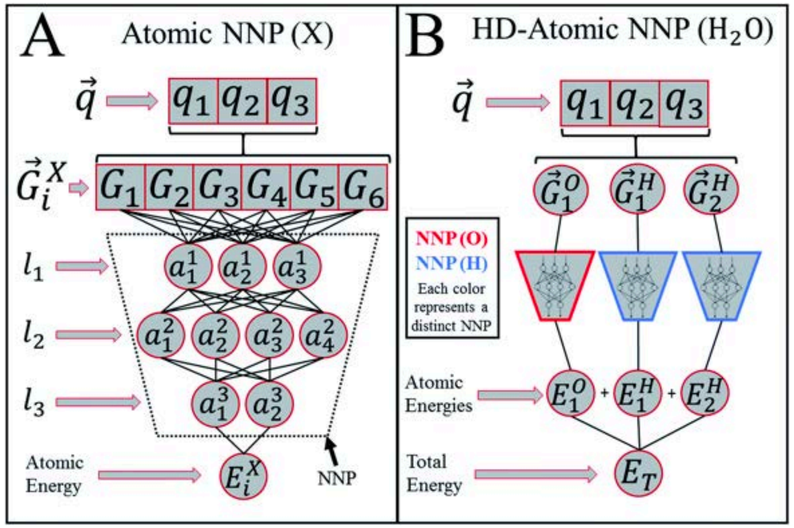
$$f_C(R_{ij}) = \begin{cases} 0.5\times\cos\left(\frac{\pi R_{ij}}{R_C}\right)+0.5 & \text{for } R_{ij} \le R_C \\ 0.0 & \text{for } R_{ij} > R_C \end{cases}$$

$$G_m^R = \sum^{\text{all atoms}} e^{-\eta\left(R_{ij}-R_s\right)^2} f_C(R_{ij})$$

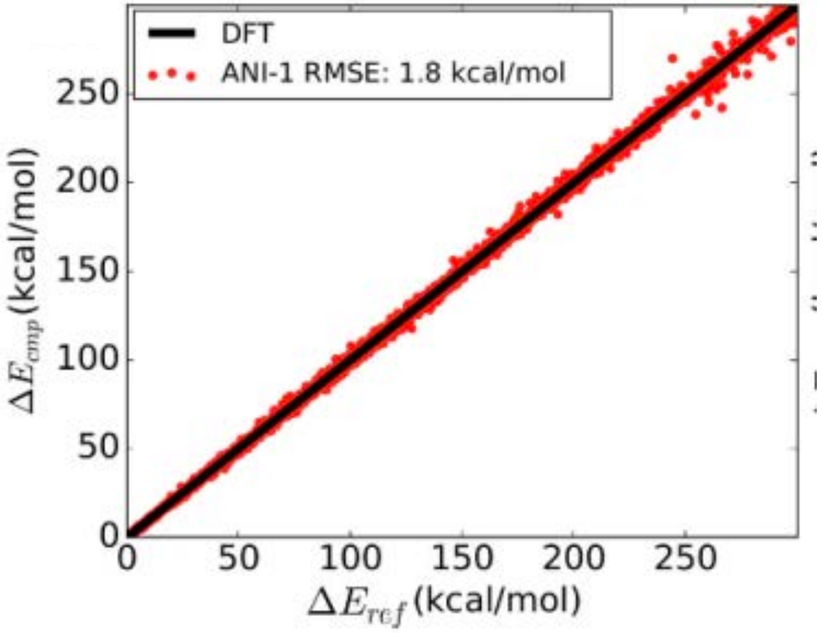
$$G_m^{A_{mod}} = 2^{1-\zeta} \sum_{j,k\neq i}^{\text{all atoms}} \left(1+\cos\left(\theta_{ijk}-\theta_s\right)\right)^\zeta \exp\left[-\eta\left(\frac{R_{ij}+R_{ik}}{2}-R_s\right)^2\right] f_C(R_{ij}) f_C(R_{ik})$$



deep neural network for each atom



excellent agreement with DFT



JUSTIN SMITH

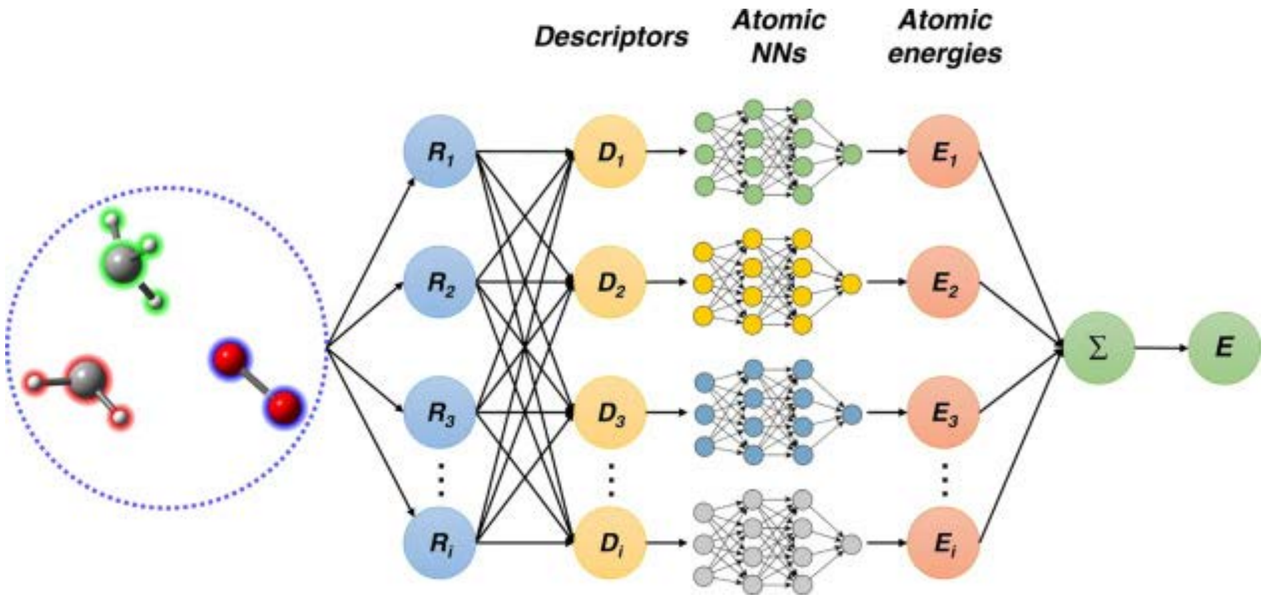
OLEXANDR ISAYEV

ADRIAN ROITBERG

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

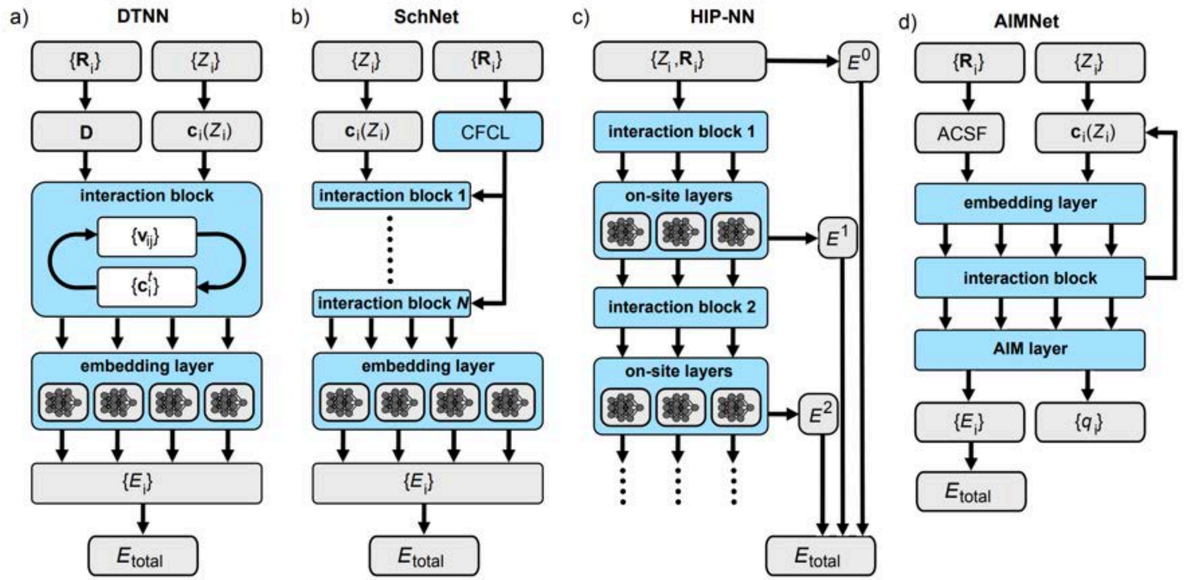
NEURAL NETWORK POTENTIALS (NNPS) ARE SEEING RAPID EVOLUTION AND EXPLORATION IN ARCHITECTURES

NNPs combine **learned geometric features** with **deep neural networks** to compute accurate multibody energies



https://doi.org/10.1016/B978-0-323-90049-2.00001-9

We are in a period of **rapid evolution of fast and accurate ML architectures** for NNPs



https://doi.org/10.48550/arXiv.2107.03727

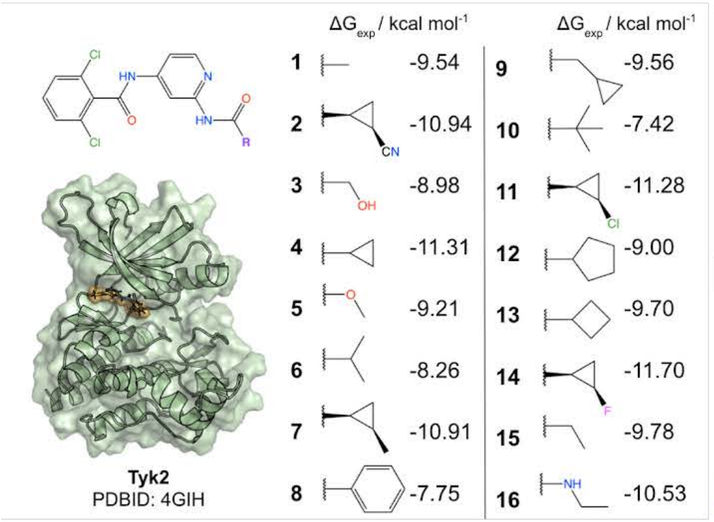
Biomolecular simulation engines that *combine* NNPs (for short-range multibody interactions) and physical energy terms (for long-range interactions and regularization) are likely to be maximally successful in the long term

EVEN REPLACING JUST THE LIGAND INTRAMOLECULAR ENERGETICS WITH NEURAL NETWORK POTENTIALS CAN SHOW SIGNIFICANT IMPROVEMENTS

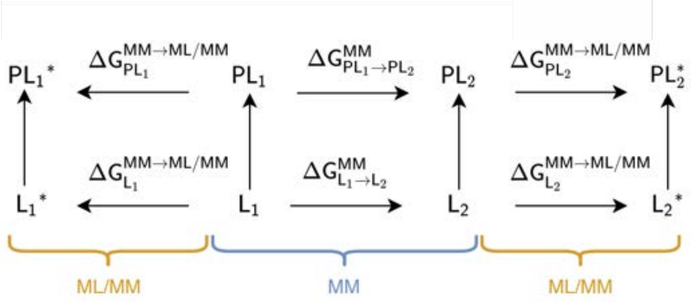
MM (OPLS2.1 + CM1A-BCC charges)
Missing torsions from LMP2/cc-pVTZ(-f) QM calculations
SPC water

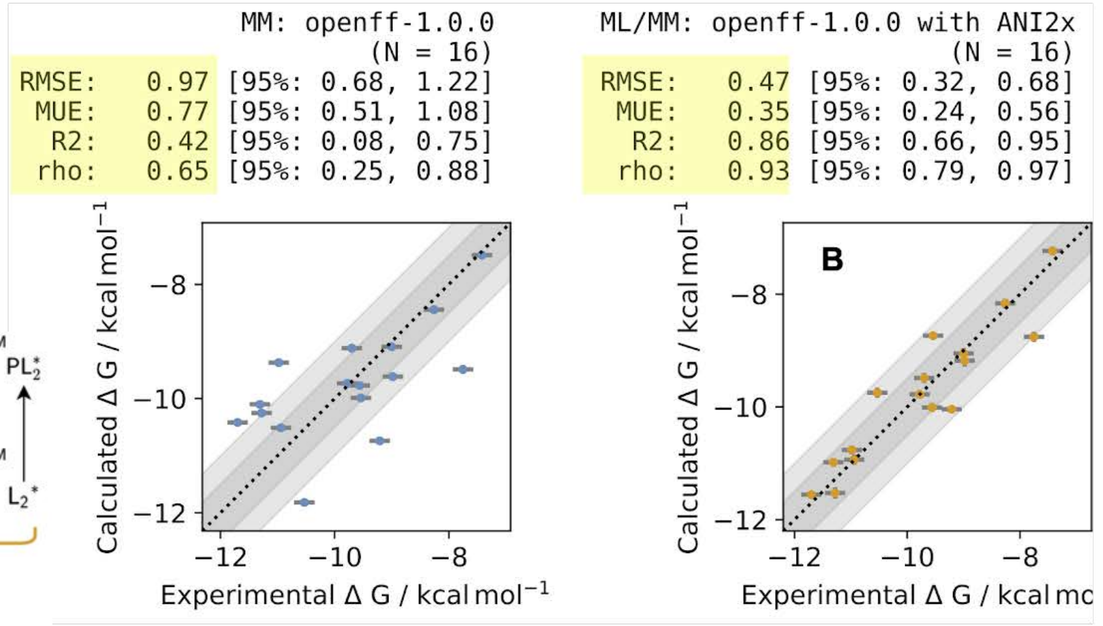
MM (OpenFF 1.0.0 "Parsley")
AMBER14SB protein force field
TIP3P; Joung and Cheatham ions

QML/MM (OpenFF 1.0.0 + ANI2x)
AMBER14SB protein force field
TIP3P; Joung and Cheatham ions



	Tyk2
no. of compds	16
binding affinity range (kcal/mol)	4.3
crystal structure	4GIH
series ref	52,53
no. of perturbations	24
MUE FEP	0.75 ± 0.11
RMSE FEP	0.93 ± 0.12



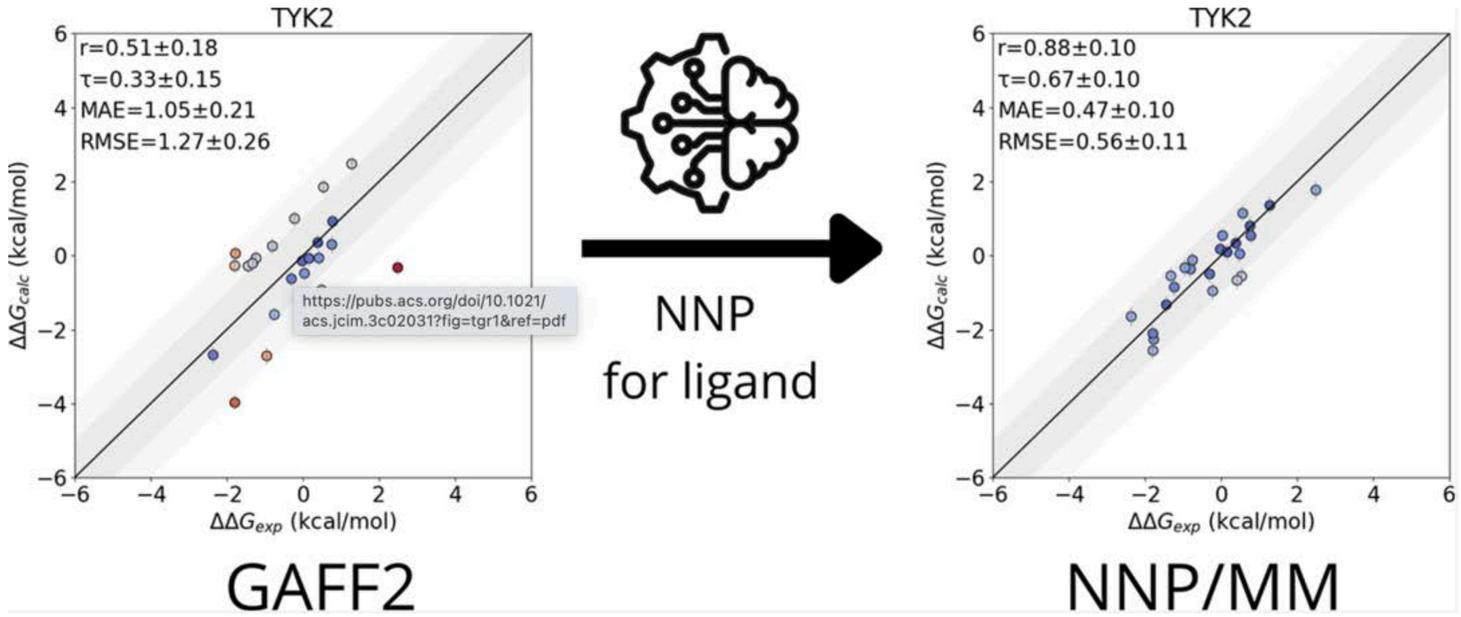


Tyk2 benchmark system from Wang et al. JACS 137:2695, 2015
replica-exchange free energy calculations with solute tempering (FEP/REST)

replica-exchange free energy calculations with perses

Rufa, Bruce Macdonald, Fass, Wieder, Grinaway, Roitberg, Isayev, and **Chodera**.
preprint: https://doi.org/10.1101/2020.07.29.227959
code: https://github.com/choderalab/qmlify

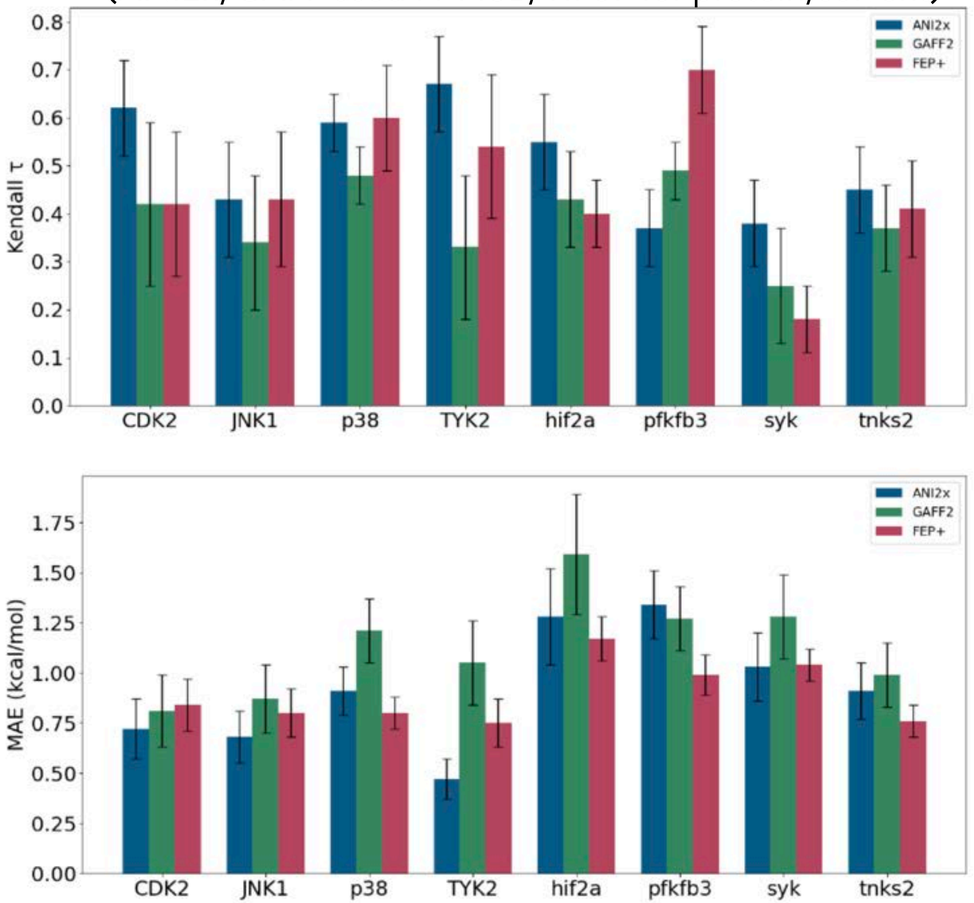
HYBRID NNP/MM POTENTIALS SHOW INCREDIBLE PROMISE OVER MM ALONE



significantly increased utility compared to GAFF2.11

ANI2x vs **GAFF2.11** vs **OPLS3e** (FEP+)

(ANI2x/GAFF used FF14SB/TIP3P for protein/solvent)



Zariquiey, Galvelis, Gallicchio, Chodera, Markland, and **De Fabritiis**
Enhancing protein-ligand binding affinity predictions using neural network potentials
JCIM 2024 (in press)

HYBRID NNP/MM SIMULATIONS CAN BE SURPRISINGLY FAST

RTX 4090 benchmarks

PDB ID	# res	# heavy atoms	OpenMM ns/day (4 fs timestep)	TorchANI QML/MM ns/day (2 fs timestep)	OpenMM QML/MM* ns/day (2 fs timestep)
3BE9	328	48	995	14.0	151 / 74.2
2P95	286	50	1006	12.2	147 / 73.5
1HPO	198	64	1227	13.4	152 / 65.9
1AJV	198	75	1382	12.6	155 / 60.1

* ANI ensemble size: 1 / 8

NNPOps library
https://github.com/openmm/nnpops

* CUDA/CPU accelerated kernels
* API for inclusion in MD engines
* Ops wrappers for ML frameworks (PyTorch so far)
* Community-driven, package agnostic

~3x slower than GPU MD right now, but need 2x smaller timestep

Notably, MD will not get much faster for small systems as hardware improves.

ML *will* continue to get much faster.

paper: https://arxiv.org/abs/2201.08110
code: https://github.com/openmm/nnpops

OPENMM 8 MAKES ML/MM SIMULATIONS INCREDIBLY EASY

```
conda install -c conda-forge openmm-ml
```

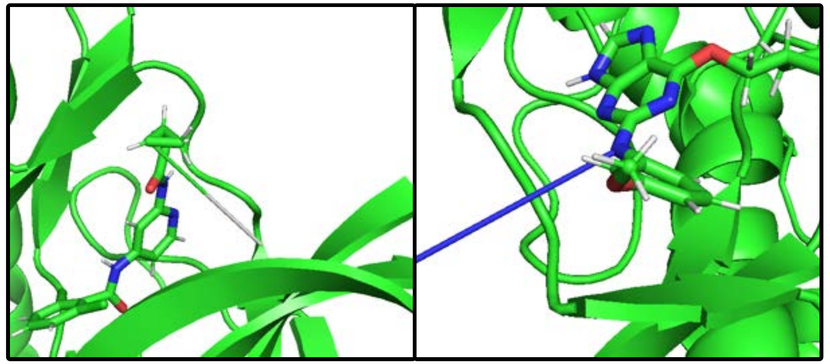
```python
# Use Amber 14SB and TIP3P-FB for the protein and solvent
forcefield = ForceField('amber14-all.xml', 'amber14/tip3pfb.xml')
# Use OpenFF for the ligand
from openmmforcefields.generators import SMIRNOFFTemplateGenerator
smirnoff = SMIRNOFFTemplateGenerator(molecules=molecules)
# Create an OpenMM MM system
mm_system = forcefield.createSystem(topology)
# Replace ligand intramolecular energetics with ANI-2x
potential = MLPotential('ani2x')
ml_system = potential.createMixedSystem(topology, mm_system, ligand_atoms)
```

https://github.com/openmm/openmm-ml

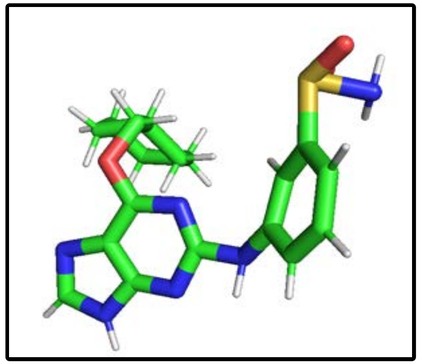
ML POTENTIALS ARE NOT WITHOUT CHALLENGES. IT'S STILL EARLY DAYS.

~ A gallery of horrors ~

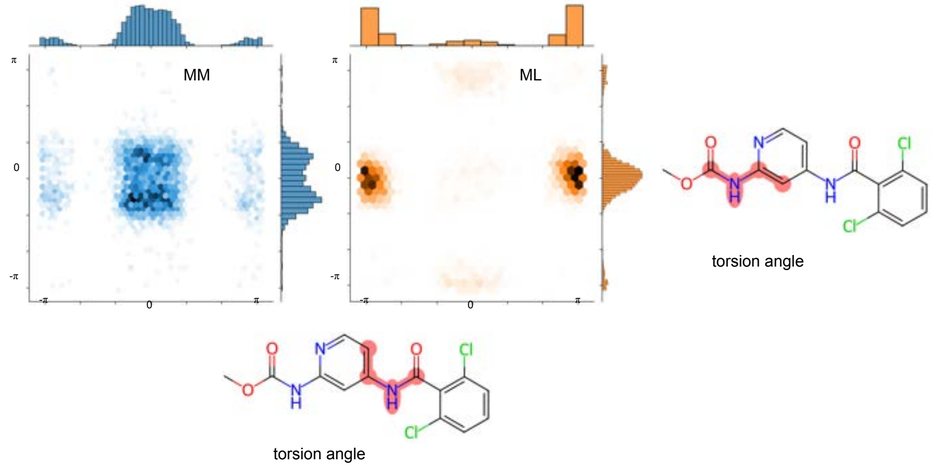
ANI2x proton cannon!



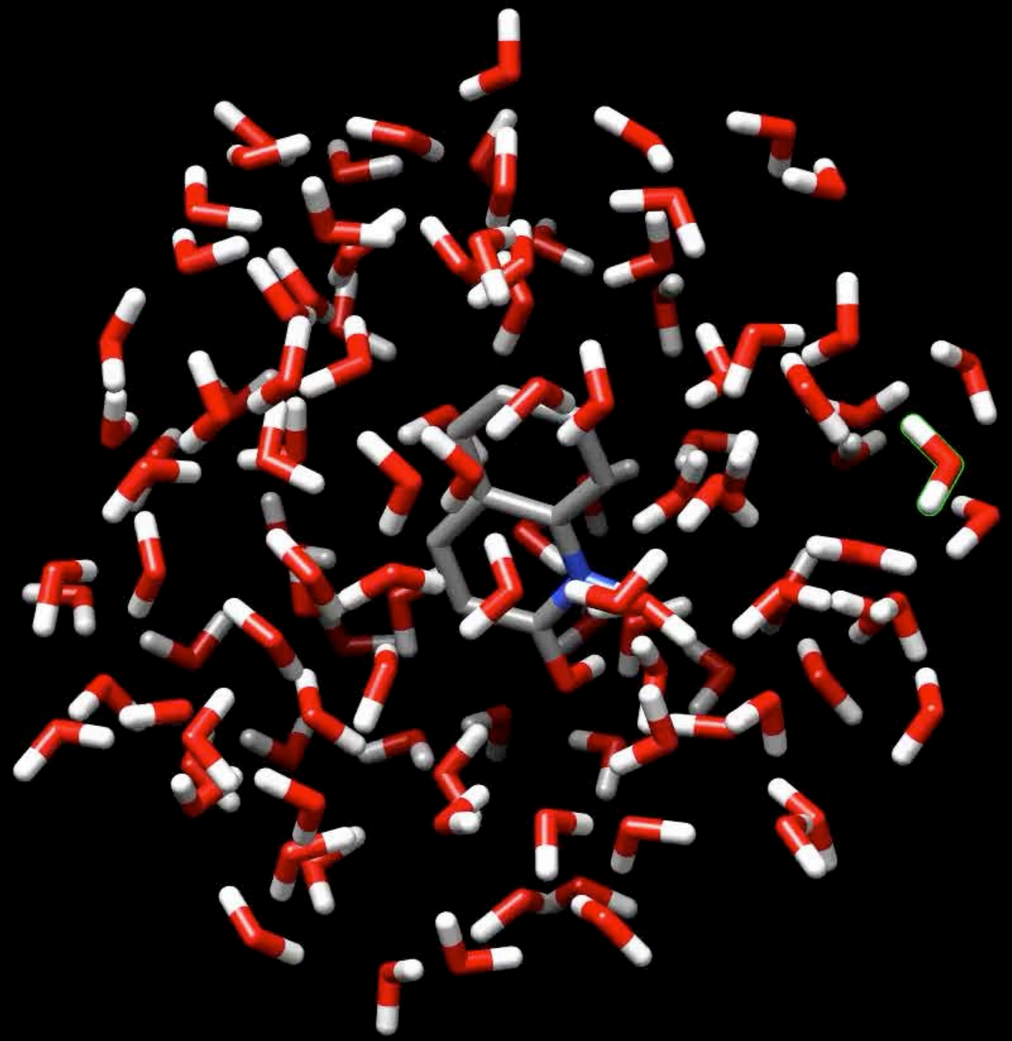
90-degree sulfonamides!



Totally different amide torsions!



WHY DO WE NEED MM AT ALL?



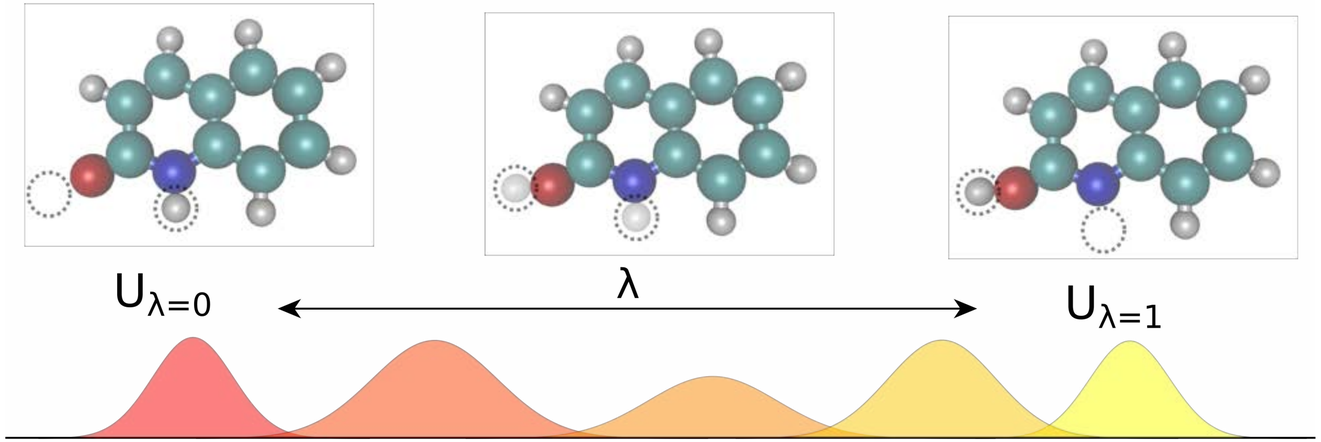
Can we treat everything in the system with NNPs?
We can finally be free of the hegemony of bonds!

ANI2x

IT'S SURPRISINGLY EASY TO COMPUTE FREE ENERGY DIFFERENCES BETWEEN CHEMICAL SPECIES USING NEURAL NETWORK POTENTIALS

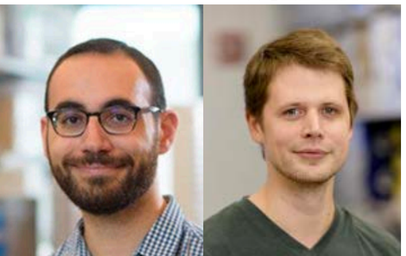
Potentials are free of singularities, so **simple linear alchemical potentials** can robustly compute alchemical free energies

$$U(x;\lambda) = (1-\lambda)U_{\lambda=0}(x) + \lambda U_{\lambda=1}(x)$$

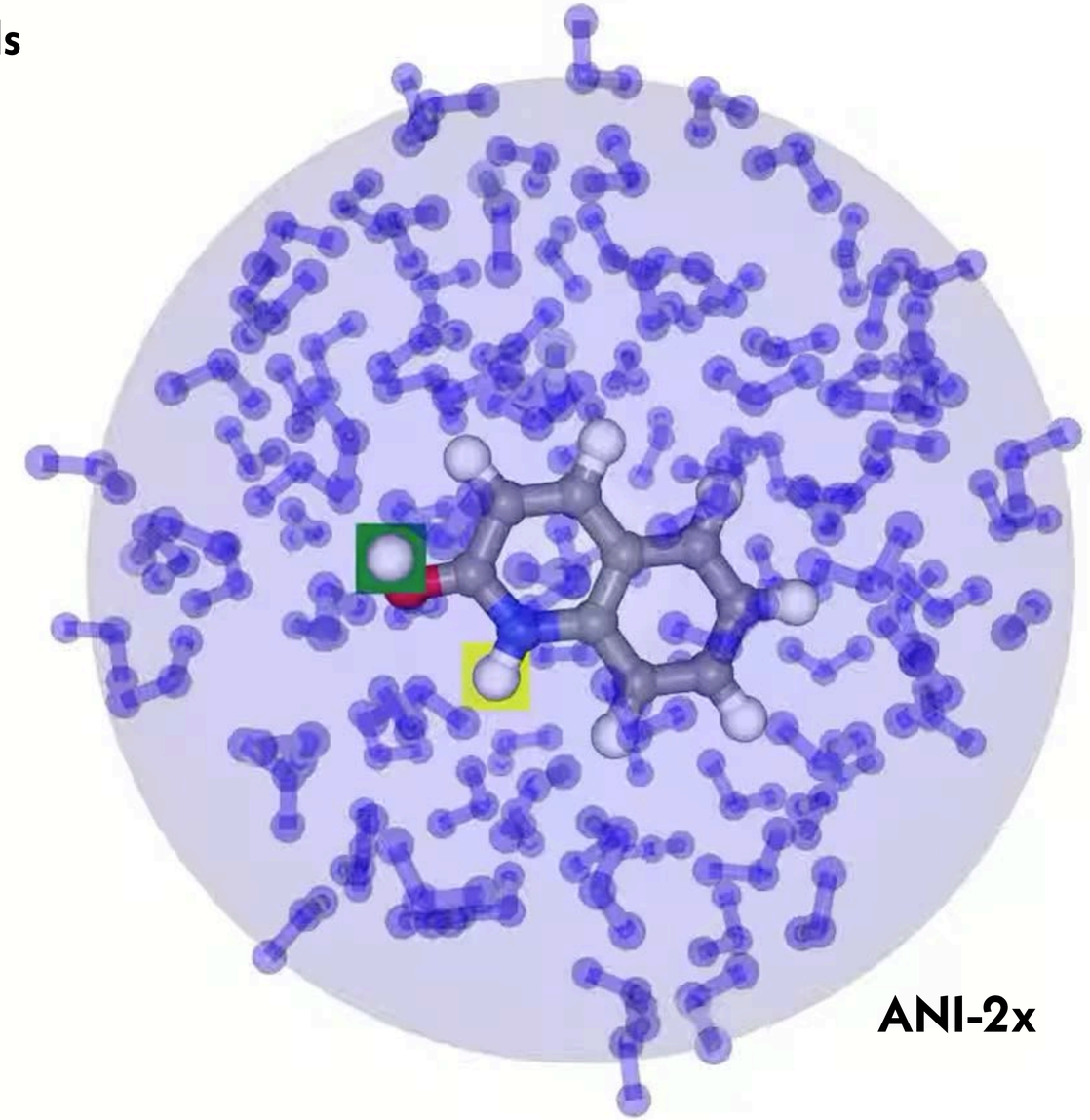


Simple restraints can be used when we need to enforce specific chemical species

JOSH FASS **MARCUS WIEDER**



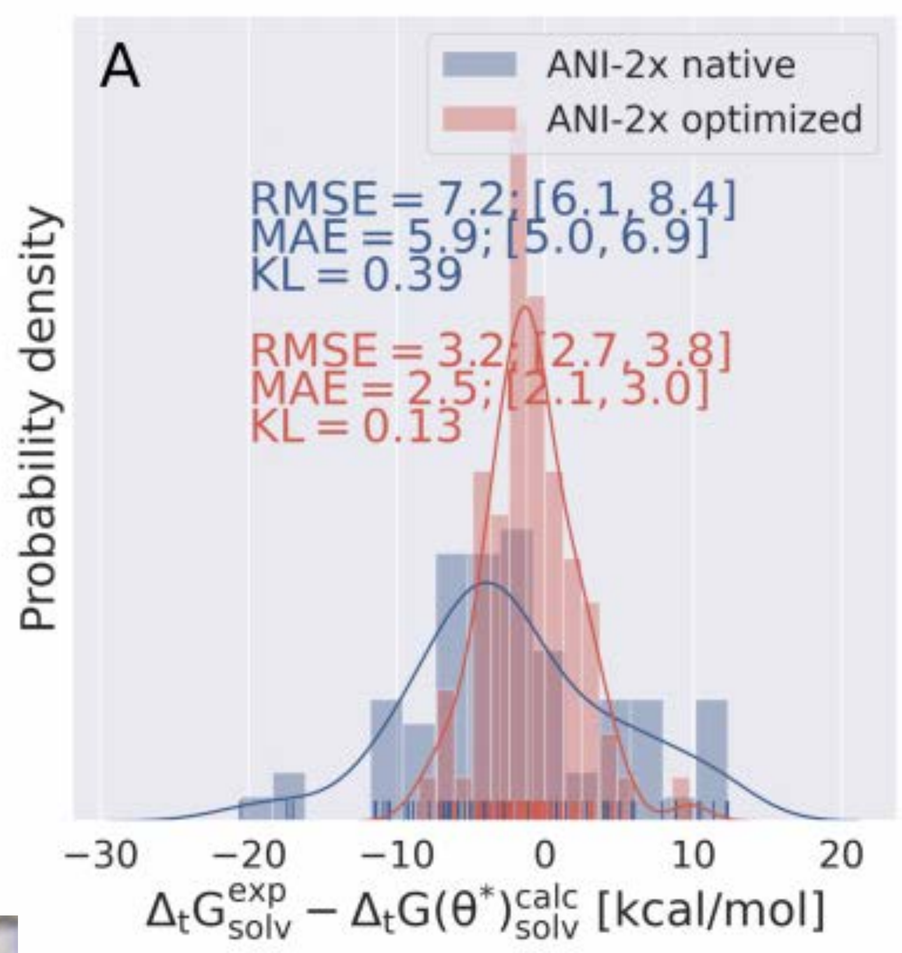
preprint: https://doi.org/10.1101/2020.10.24.353318
code: https://github.com/choderalab/neutromeratio



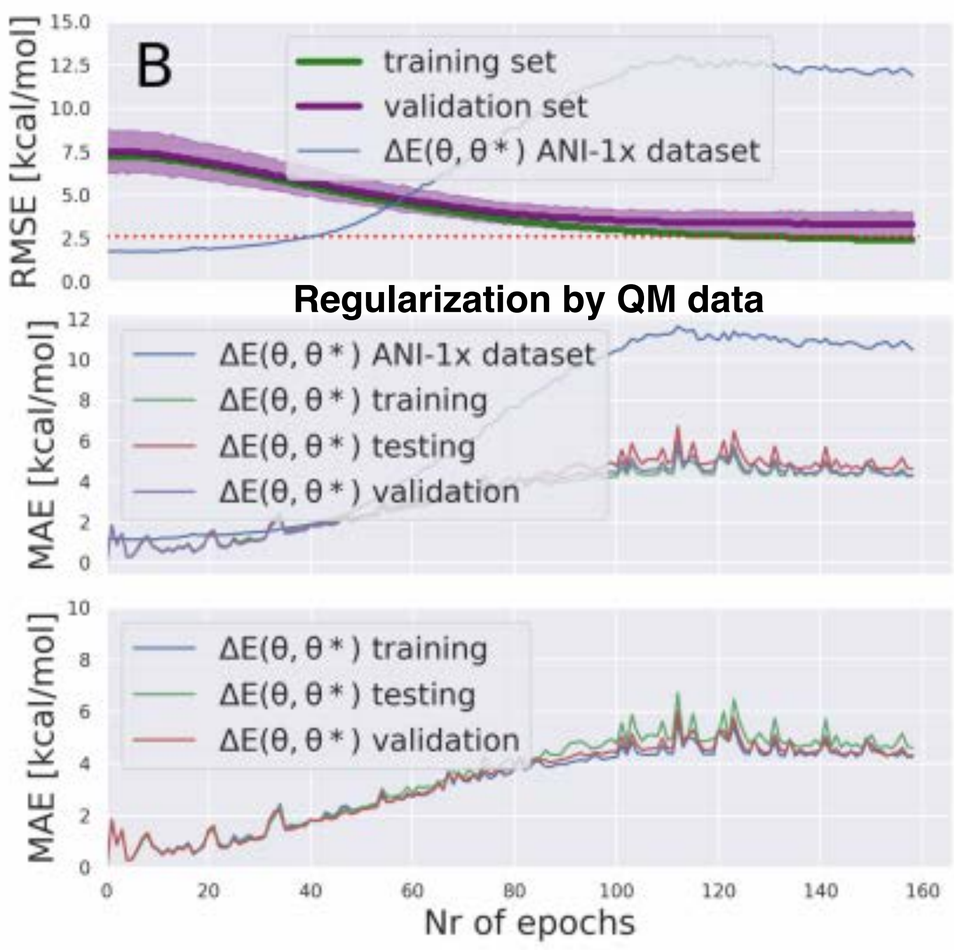
ANI-2x

NEURAL NETWORK POTENTIALS CAN BE FINE-TUNED USING EXPERIMENTAL DATA, REGULARIZED BY QM DATA

test set performance

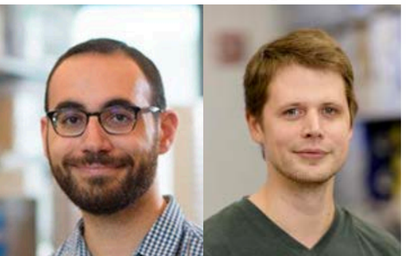


training / validation optimization



JOSH FASS

MARCUS WIEDER



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

CAN WE CHANGE PRACTICE IN STRUCTURE-ENABLED DRUG DISCOVERY BY LEVERAGING DATA WE GENERATE?

2023

week 1

MON	TUE	WED	THU	FRI	SAT	SUN
designs/ predictions	synthesis			new data		

using published force field model

week 2

MON	TUE	WED	THU	FRI	SAT	SUN
designs/ predictions	synthesis			new data		

using the same published force field model!
we haven't learned anything from the data

"Insanity is doing the same thing over and over again and expecting different results"

- Rita Mae Brown (not Albert Einstein)

CAN WE CHANGE PRACTICE IN STRUCTURE-ENABLED DRUG DISCOVERY BY LEVERAGING DATA WE GENERATE?

2023

week 1

MON	TUE	WED	THU	FRI	SAT	SUN
designs/ predictions	synthesis			new data		

using published force field model

week 2

MON	TUE	WED	THU	FRI	SAT	SUN
designs/ predictions	synthesis			new data		

using the same published force field model!
we haven't learned anything from the data

2025

week 1

MON	TUE	WED	THU	FRI	SAT	SUN
designs/ predictions 1.0	synthesis			new data	build model 2.0!	

using foundation force field model
built from public data

week 2

MON	TUE	WED	THU	FRI	SAT	SUN
designs/ predictions 2.0	synthesis					

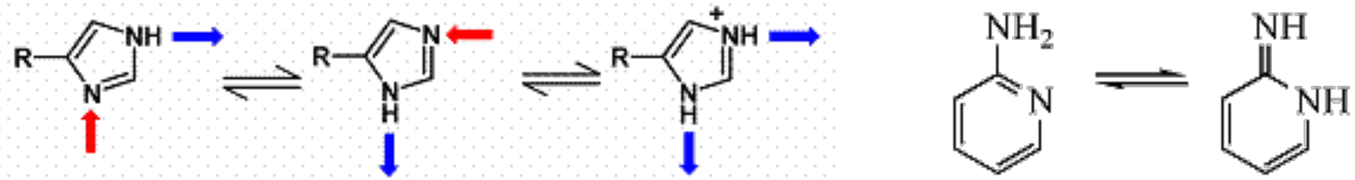
using model fine-tuned to our private data
(e.g., data generated for our discovery program)

A CAUTION: THIS IS JUST ONE PART OF THE SOLUTION, BUT IT HELPS US ADDRESS THE OTHERS

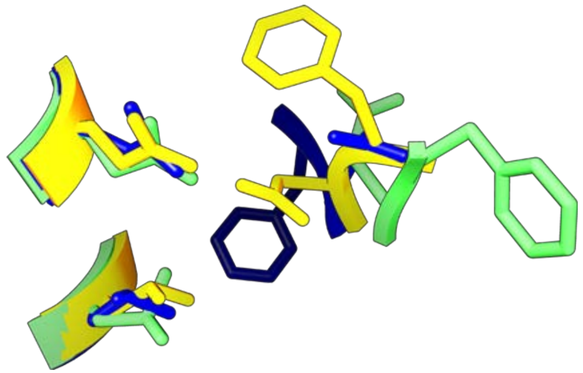
1. The forcefield may do a poor job of modeling the physics of our system (because it is constrained by choices made 40 years ago)

$$V(\mathbf{q}) = \sum_{\text{bonds}} K_r(r - r_{eq})^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}} \right]$$

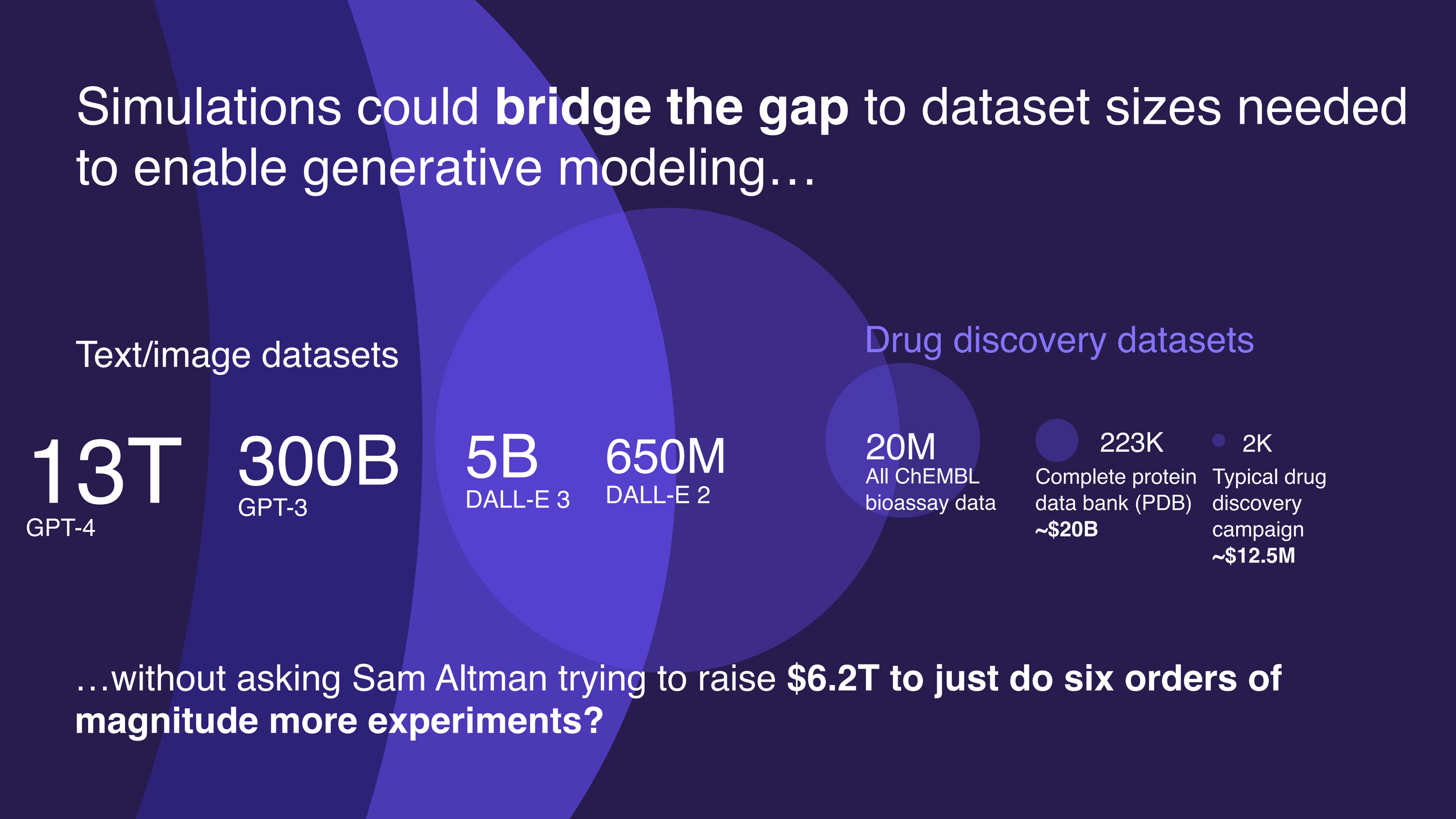
2. We're missing some essential chemical in our simulations because we don't bother to model them (e.g. protonation states, tautomers, redox chemistry, PTMs, etc.))



3. We haven't sampled all of the relevant conformations because we can't simulate for long enough



Simulations could **bridge the gap** to dataset sizes needed to enable generative modeling…



…without asking Sam Altman trying to raise **$6.2T to just do six orders of magnitude more experiments?**

SIMULATE → EMULATE → GENERATE

experimental discovery data

quantum chemical data

→

build accurate
simulators
foundation simulation models for biomolecular systems

→

train
uncertainty-aware
emulators
structure-based ML surrogate models to enable massive scaling

→

build conditional
generators
generative models of molecules and conformations conditioned on TCP properties

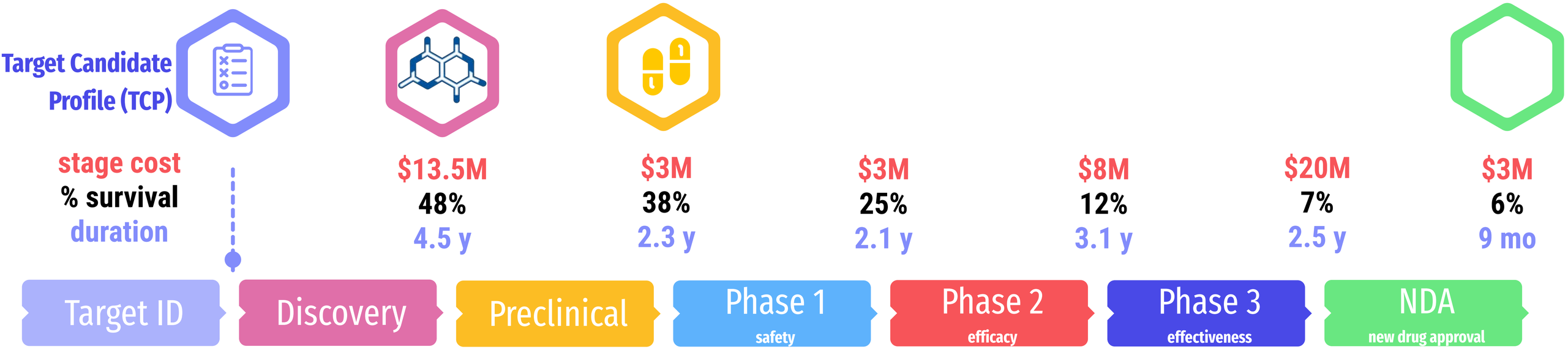
→

directly use generative model to produce molecules conditioned on design objectives

A RISING TIDE LIFTS ALL BOATS

By working together to solve major challenges,
we can improve success rates for everybody

OpenADMET: Open dataset and predictive models for ADMET properties of relevance to small molecule discovery and development



>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]

ADMET

Absorption

Distribution

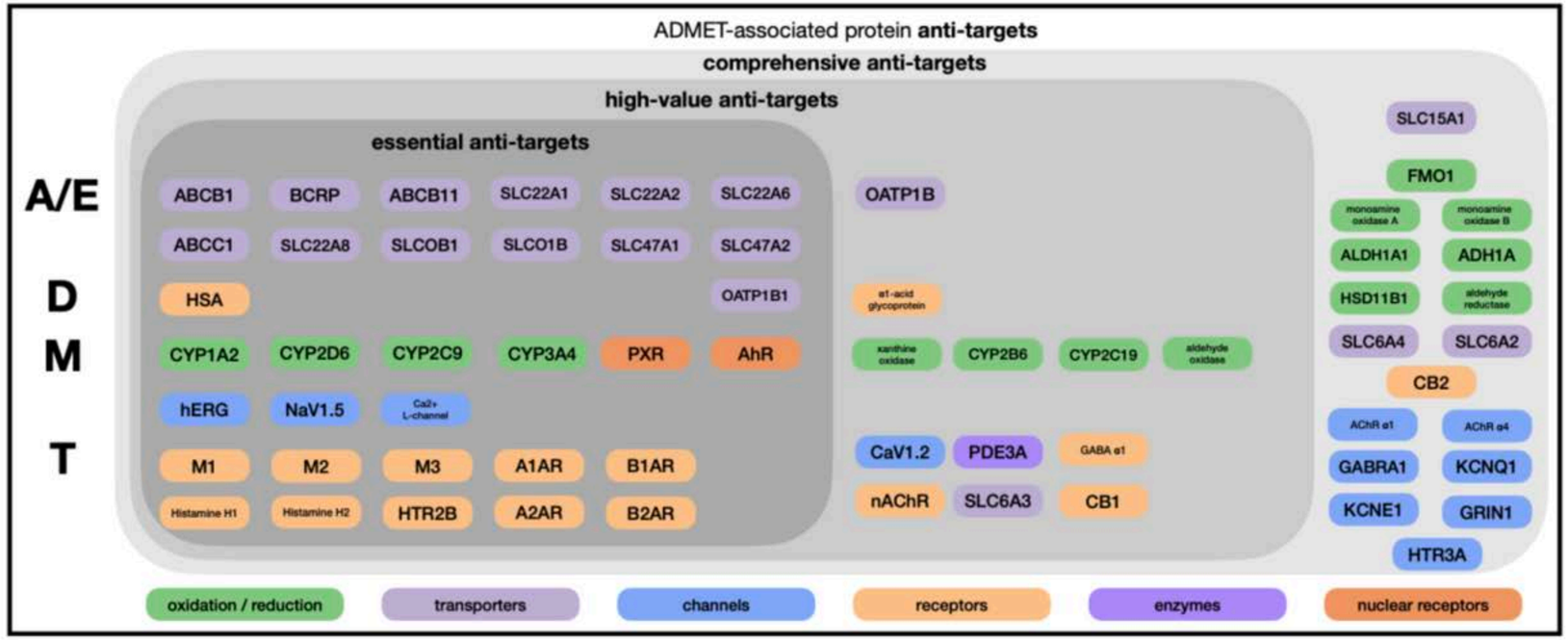
Metabolism

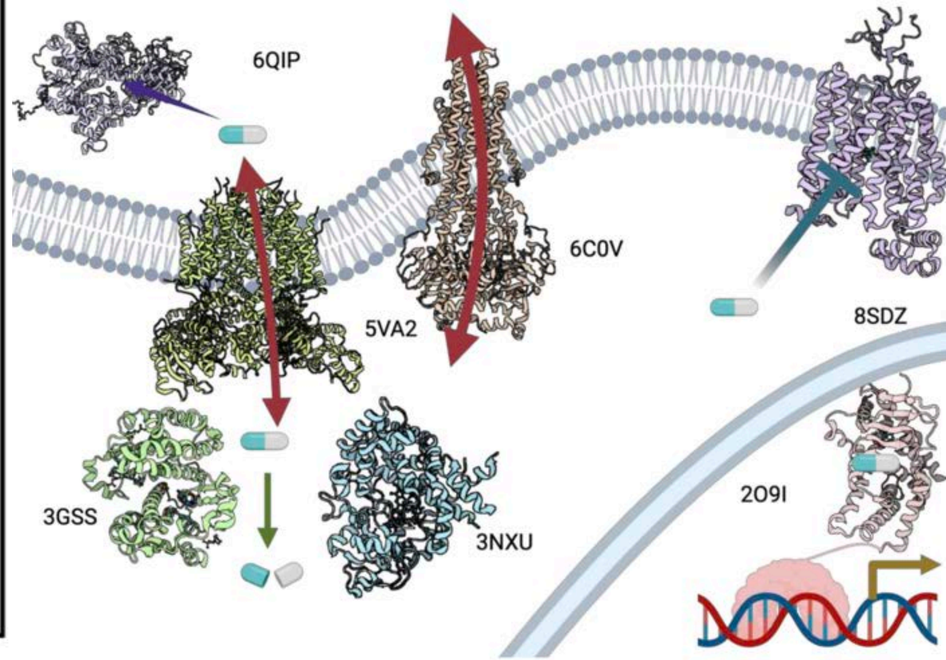
Excretion

Toxicity

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

INTERACTIONS WITH A LIMITED NUMBER OF ADMET-ASSOCIATED PROTEINS (THE "AVOIDOME") DRIVES COMMON ADMET ISSUES





https://doi.org/10.1016/j.cell.2024.01.003

Current ADMET datasets were always generated in pursuit of a *drug*, rather than a useful model. And the use of *structural data* to improve ADMET predictions has been all but ignored (until recently*)

Can we generate cheap, high-quality data suitable for building models? And simultaneously, solve enough structures to make predicted models useful for modeling ADMET?

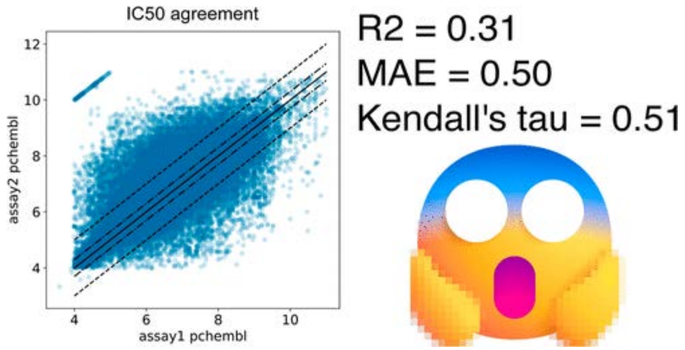

Cell
Leading Edge
CellPress
Commentary
Enabling structure-based drug discovery utilizing predicted models
Edward B. Miller,[1,*] Howook Hwang,[1] Mee Shelley,[2] Andrew Placzek,[2] João P.G.L.M. Rodrigues,[1] Robert K. Suto,[3] Lingle Wang,[1] Karen Akinsanya,[1] and Robert Abel[1]
[1]Schrödinger New York, 1540 Broadway, 24th Floor, New York, NY 10036, USA

OpenBind: Accelerating protein:ligand structure and affinity prediction by generating fit-for-purpose datasets via active learning

Can we do for drug design what the PDB and CASP did for AlphaFold?

AlphaFold3 holds promise for predicting protein:ligand structures as a precursor to predicting affinities (through physical or ML models), but this has major limitations:

* only ~220K structures in PDB
* only ~40-80K usable protein:small molecule structures
* public affinity datasets are poorly aligned to structures and a total disaster in consistency



https://doi.org/10.1021/acs.jcim.4c00049

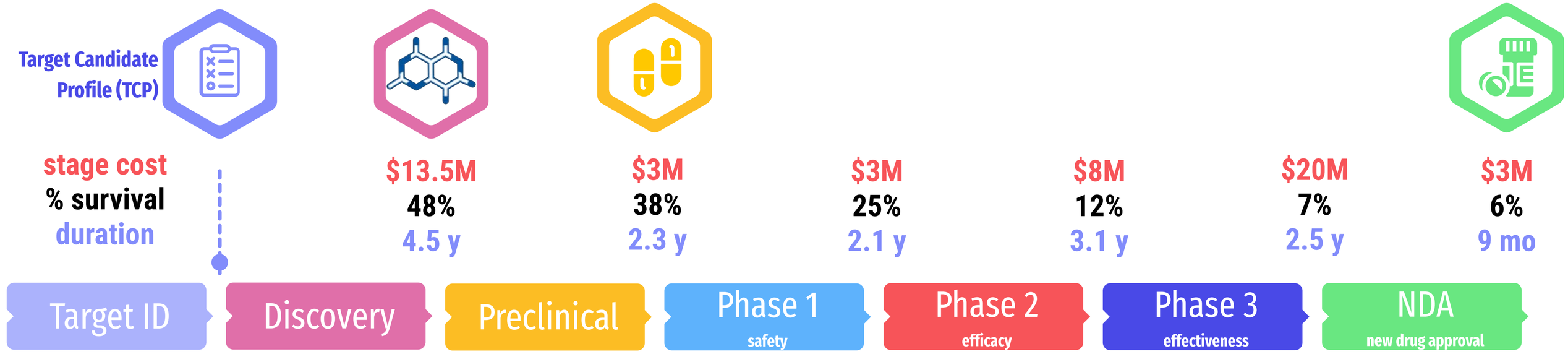
Opportunity: X-ray beamlines like Diamond Light Source K04 can produce **one protein:ligand structure dataset every 8 seconds**
=> 1M structures/year even if beamline only at 25%

Rate-limiting step is just how quickly we can feed it interesting, cheap chemistry soaked into crystals in a manner that binds.

CHODERA LAB

- All funding: http://choderalab.org/funding

SIMULATION IS GOING TO BE ESSENTIAL IN ADDRESSING MAJOR CAUSES OF FAILURE IN DRUG DISCOVERY



>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]

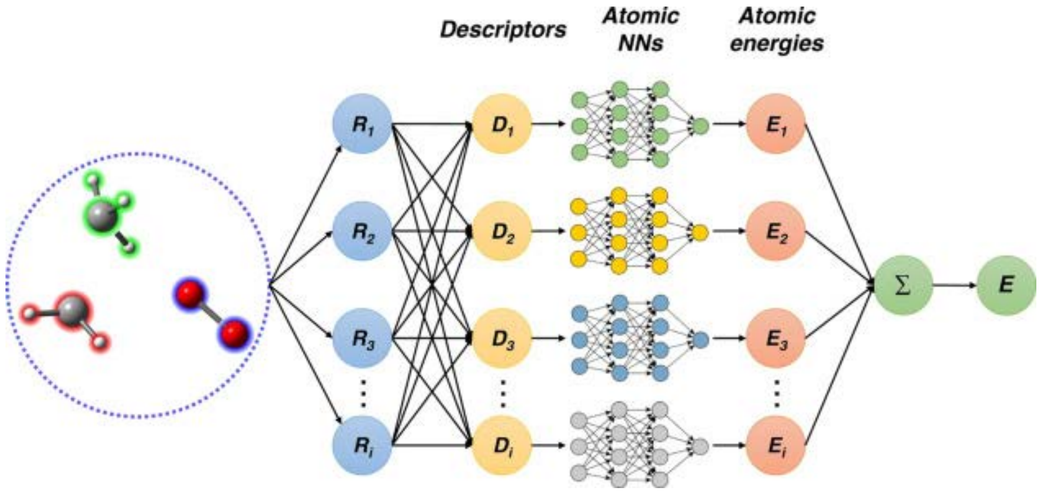
ADMET

Absorption
Distribution
Metabolism
Excretion
Toxicity

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

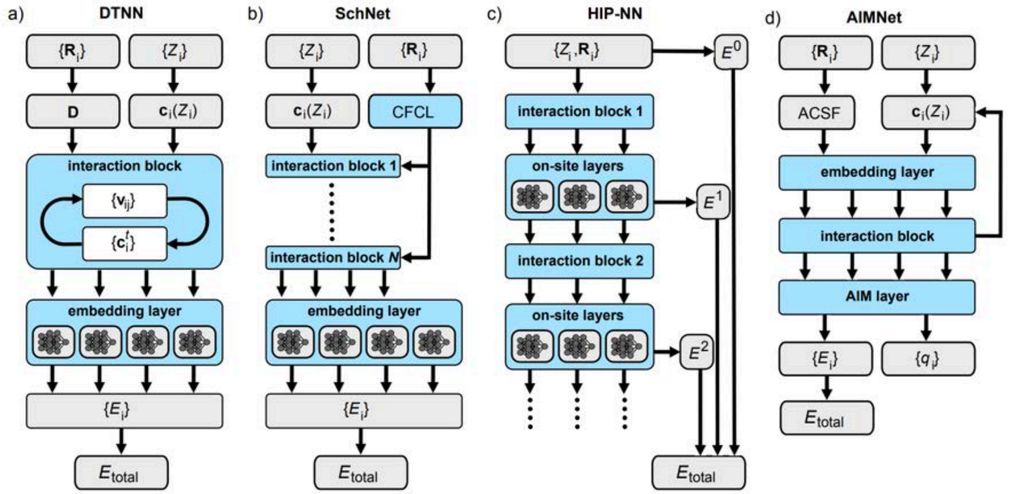
Foundation simulation models are powered by modern machine learning frameworks and neural network potentials

NNPs combine



https://doi.org/10.1016/B978-0-323-90049-2.00001-9

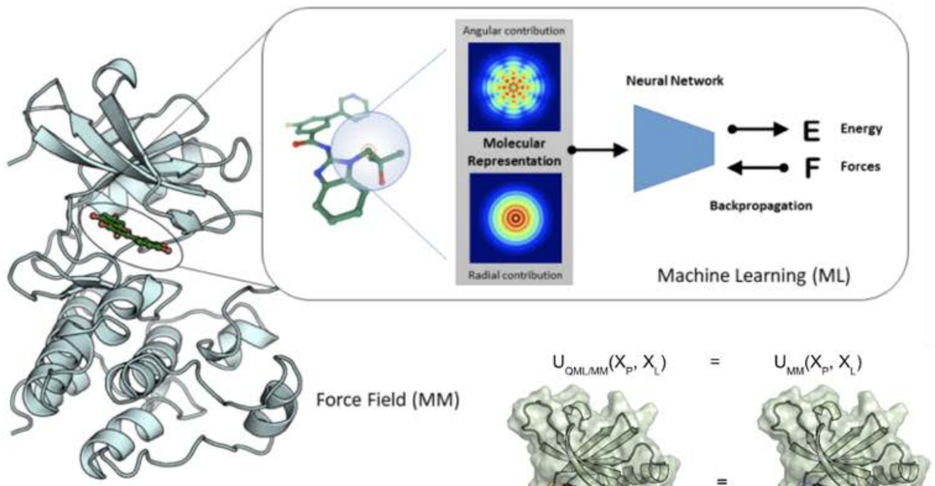
We are in a period of

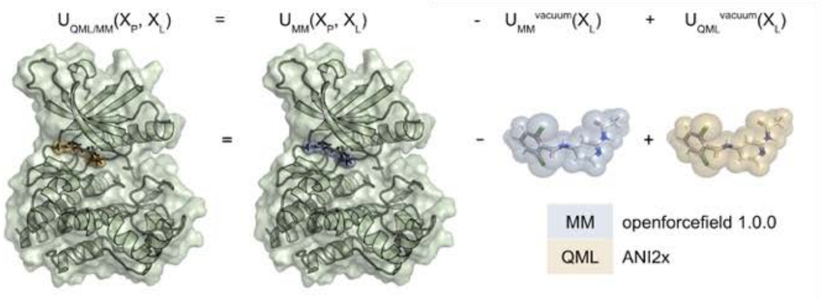


https://doi.org/10.48550/arXiv.2107.03727

Neural network potentials (NNPs) combine **learned geometric features** with **deep neural networks** to compute accurate multibody energies using advanced ML frameworks (JAX/PyTorch)

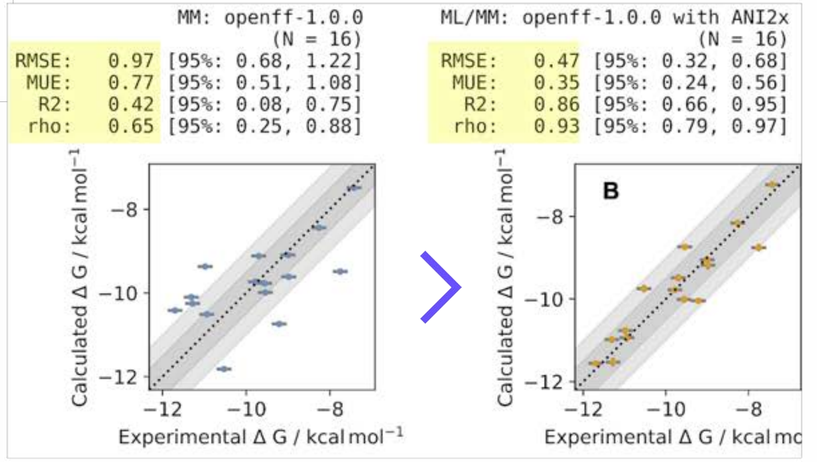
Early proof-of-concept work shows significant promise





MM
current generation

NNP/MM
hybrid NNP/MM

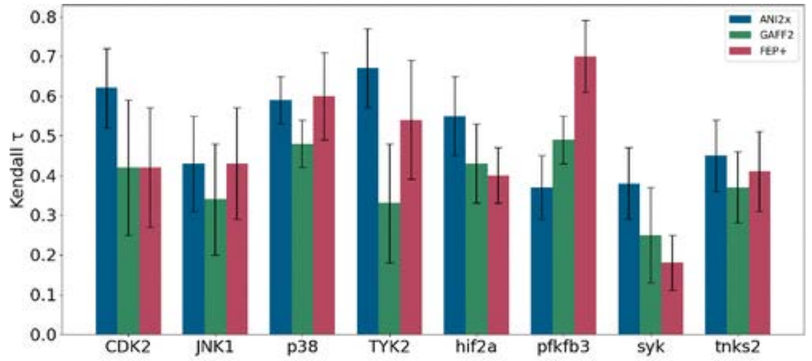


Introducing NNPs for even part of the system **significantly reduces errors** over current-generation molecular mechanics (MM) simulations

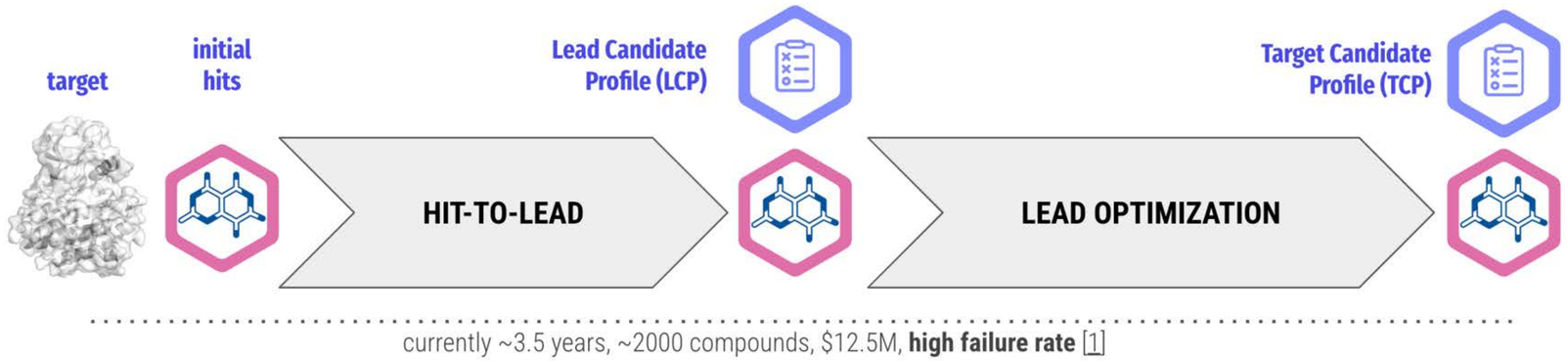
https://doi.org/10.1101/2020.07.29.227959

and yields **improved correlation** with experiment

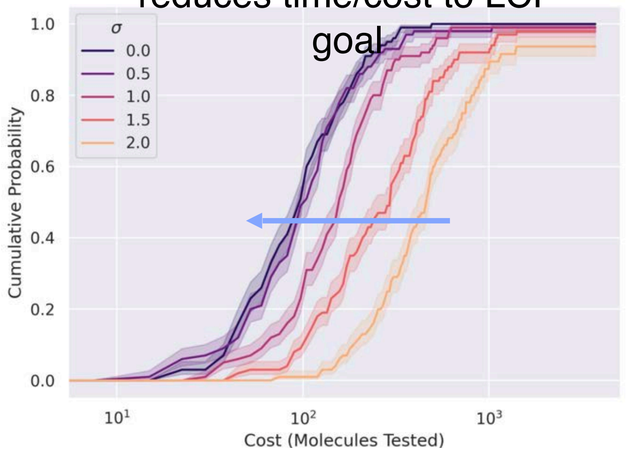
https://doi.org/10.1021/acs.jcim.3c02031



Foundation simulation models are poised to immediately deliver value within the current drug discovery paradigm



increasing accuracy reduces time/cost to LCP goal



https://doi.org/10.1101/2024.05.28.596296

Foundation simulation models can immediately be integrated into current computer-aided drug discovery (CADD) workflows to accelerate discovery

Foundation simulation models overcome data limitations by enabling synthetic datasets to substitute for experiment

target

\+

Target Candidate Profile (TCP)

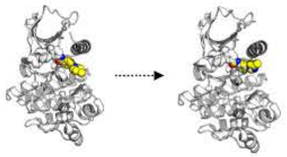
GENERATIVE DESIGN CONDITIONED ON OBJECTIVES

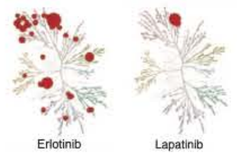
trained on multi-property synthetic data from ACHIRA foundation models

Multiple candidates for preclinical development

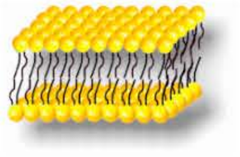
~weeks, ~ 10 molecules synthesized by CRO FTE chemists

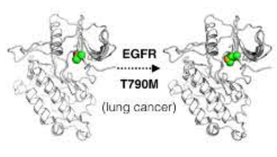
Foundation simulation models can enable numerous key drug discovery applications


driving potency

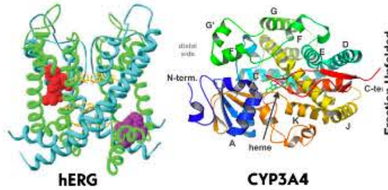


driving selectivity

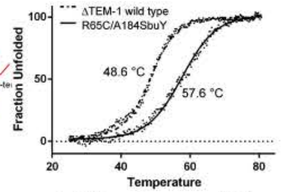

partition coefficients and permeabilities



predicting drug resistance/sensitivity

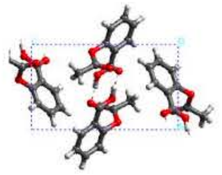


structure-enabled ADME/Tox targets



optimizing thermostability

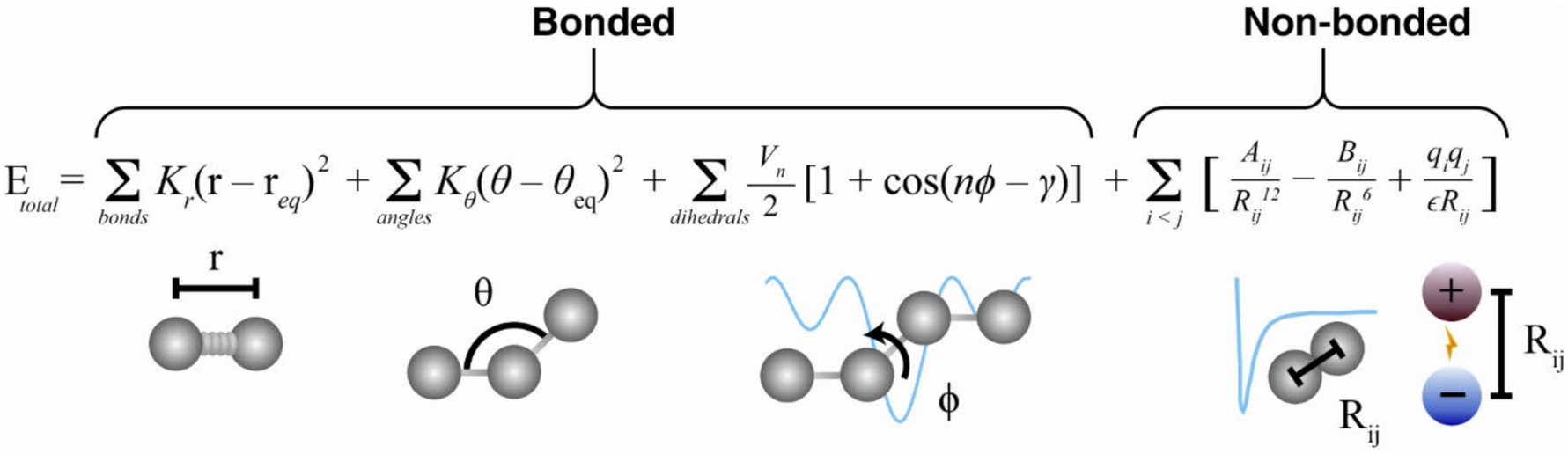

bacterial porin permeation


crystal polymorphs, solubility

- **Optimizing/maintaining potency** against novel protein targets *(relative affinity)*
- **Hit identification** *(absolute affinity)*
- **Generation of synthetic data** for training ML models *(absolute and relative affinity)*
- Optimizing **target selectivity** *(relative selectivity)*
- Identification and optimization of **molecular glues** *(absolute and relative affinity/selectivity)*
- Design and optimization of **heterobifunctionals** / **proximity control** (relative affinity, selectivity)
- Designing **polypharmacology** *(relative affinity, target selectivity)*
- **Allosteric modulato**r design *(relative conformational selectivity)*
- **Mutant-selective** kinase inhibition *(relative affinity/selectivity/resistance)*
- Targeting **RNA** *(absolute and relative affinity, selectivity)*
- Optimizing **biotherapeutics** *(relative mutation affinity/selectivity, relative thermostability)*

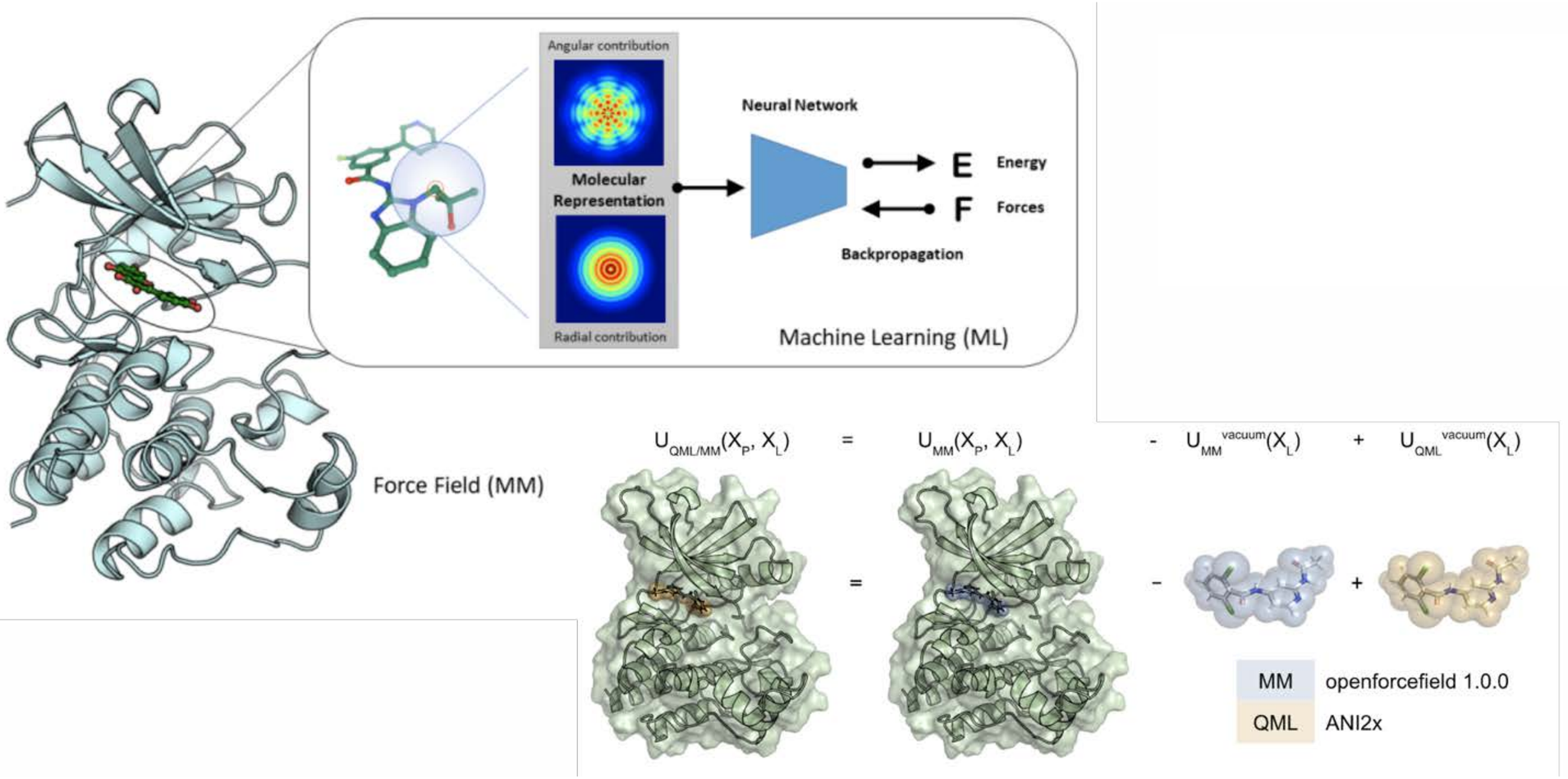
ML/MM REPEX/ATM FEP/MBAR RBFE

molecular mechanics force field



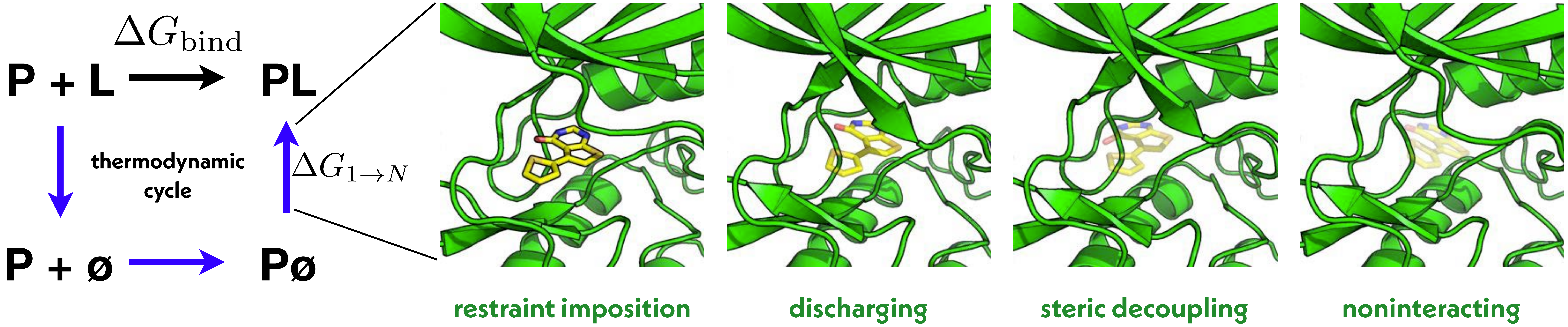
ML/MM REPEX/ATM FEP/MBAR RBFE

hybrid machine learning / molecular mechanics force field



ML/MM REPEX/ATM FEP/MBAR RBFE

Free energy perturbation uses alchemical intermediates to compute binding free energies



Includes all contributions from enthalpy and entropy of binding to a flexible receptor

$$\Delta G_{0\to1} = -k_B T \ln \frac{Z_1}{Z_0} = -k_B T \ln \frac{Z_{\lambda_2}}{Z_{\lambda_1}} \frac{Z_{\lambda_3}}{Z_{\lambda_2}} \cdots \frac{Z_{\lambda_N}}{Z_{\lambda_{N-1}}}$$

$$Z_n = \int dx \, e^{-\beta U_n(x)}$$ partition function

From the lab of Emilio Gallicchio (Brooklyn College, CUNY)

JCIM 17:3309, 2021; JCIM 62:309, 2022; JCIM 63:2438, 2023; JCIM 64:250, 2024

Pioneering work from many: McCammon, van Gunsteren, Kollman, Jorgensen, Roux, Boresch, Fujitani, Pande, Shirts, Swope, Christ, Mobley, Schrödinger, and many more

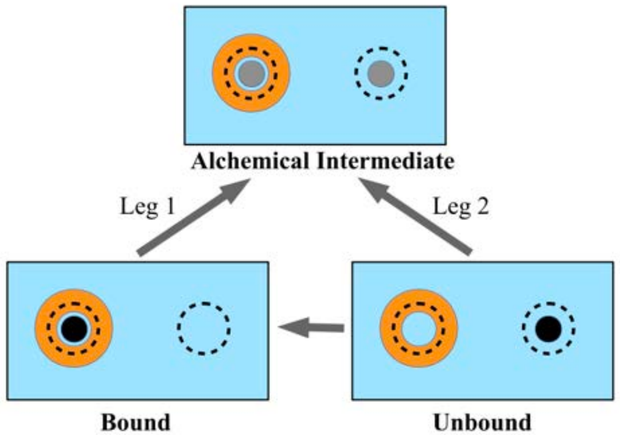
ML/MM REPEX/ATM FEP/MBAR RBFE

Alchemical Transfer Method (ATM) defines alchemical intermediates in a surprisingly simple way:

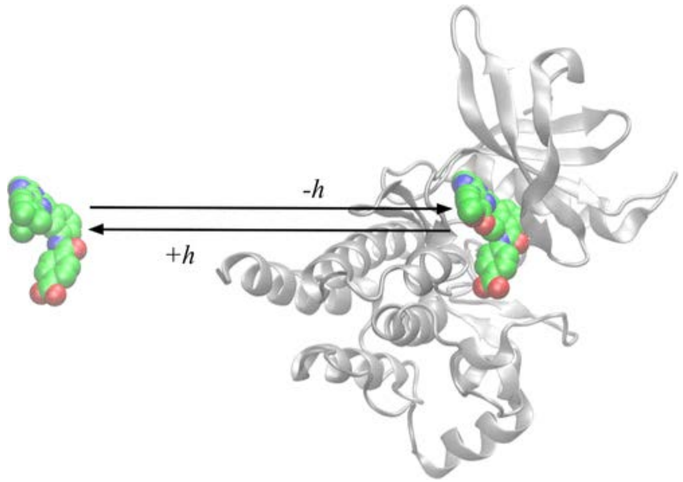
$$U(x;\lambda) = (1-\lambda)\,U_0(x) + \lambda\,U_0(x+\Delta x)$$

unmodified potential

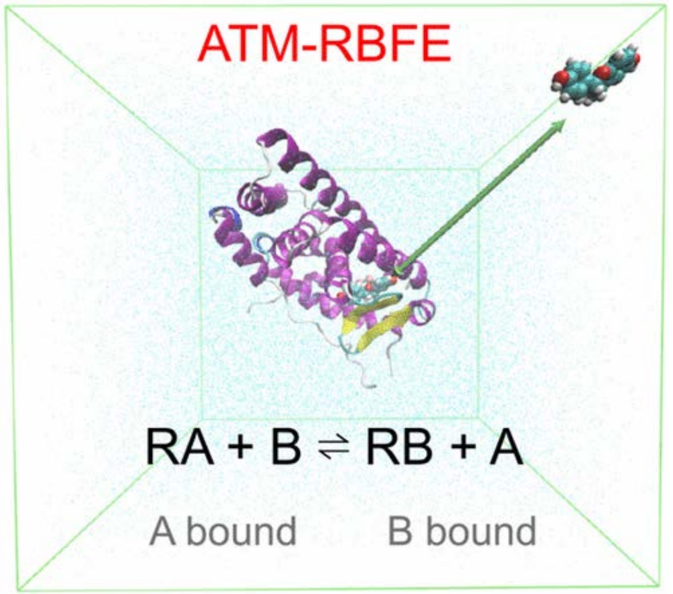
displaced ligand potential



absolute binding free energies



relative binding free energies



ATM works with molecular mechanics and machine learning force fields without any special changes!

From the lab of Emilio Gallicchio (Brooklyn College, CUNY)

JCIM 17:3309, 2021 ; JCIM 62:309, 2022 ; JCIM 63:2438, 2023 ; JCIM 64:250, 2024

ML/MM REPEX/ATM FEP/MBAR RBFE

replica exchange sampling of multiple alchemical states

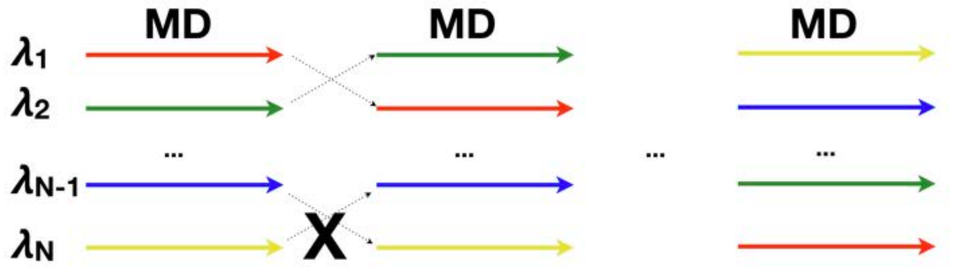
Independent simulations
Easy to parallelize, but sampling problems at any λ can make calculations unreliable

simple but dangerous due to poor sampling of conformational changes coupled to λ

AMBER18 TI
Song, Lee, Zhu, York, Merz 2019
https://doi.org/10.1021/acs.jcim.9b00105

Replica exchange (REPEX)
Good sampling at any λ can rescue problems at other λ if good λ overlap

reliable but communication heavy



Schrödinger FEP+
Wang, Wu, Deng, Kim, ... Abel 2015
https://doi.org/10.1021/ja512751q

NAMD
Jiang, Thirman, Jo, Roux 2018
http://doi.org/10.1021/acs.jpcb.8b03277

OpenMM
Chodera, Shirts
https://doi.org/10.1063/1.3660669

Nonequilibrium methods
Less efficient than equilibrium calculations, but can work robustly and scalably if properly tuned

cloud- and wall clock friendly

pmx / gromacs
Aldeghi, Gapsys, de Groot 2018
https://doi.org/10.1021/acscentsci.8b00717

Orion NES!

ML/MM REPEX/ATM FEP/MBAR RBFE

Multistate Bennett Acceptance Ratio (MBAR) provides an optimal way to analyze data to estimate free energy differences

$$q_k(x) \equiv e^{-\beta[U_0(x)+U_k(x)]}$$

unnormalized probability distribution

$U_0(x)$ **Unperturbed potential**

$U_k(x)$ **perturbed potential**

We can use optimal bridge sampling estimator machinery (Z. Tan, Meng, Wong, others) to produce the multistate generalization of Bennett acceptance ratio (BAR) that provides efficient estimators for

free energy differences

$$\Delta f_{ij} \equiv f_j - f_i = -\ln \frac{\int d\boldsymbol{x}\, q_j(\boldsymbol{x})}{\int d\boldsymbol{x}\, q_i(\boldsymbol{x})}$$

exact

equilibrium expectations

$$\langle A \rangle \equiv \frac{\int d\boldsymbol{x}\, A(\boldsymbol{x})\, q(\boldsymbol{x})}{\int d\boldsymbol{x}\, q(\boldsymbol{x})}$$

estimators from data

$$\hat{f}_i = -\ln \sum_{j=1}^{K} \sum_{n=1}^{N_j} \frac{q_i(\boldsymbol{x}_{jn})}{\sum_{k=1}^{K} N_k\, e^{\hat{f}_k}\, q_k(\boldsymbol{x}_{jn})}$$

$$\delta^2 \Delta \hat{f}_{ij} = \hat{\Theta}_{ii} - 2\hat{\Theta}_{ij} + \hat{\Theta}_{jj}$$

$$\hat{A} = \sum_{n=1}^{N} W_{na}\, A(\boldsymbol{x}_n)$$

where n now runs from 1 to $N = \sum_{k=1}^{K} N_k$

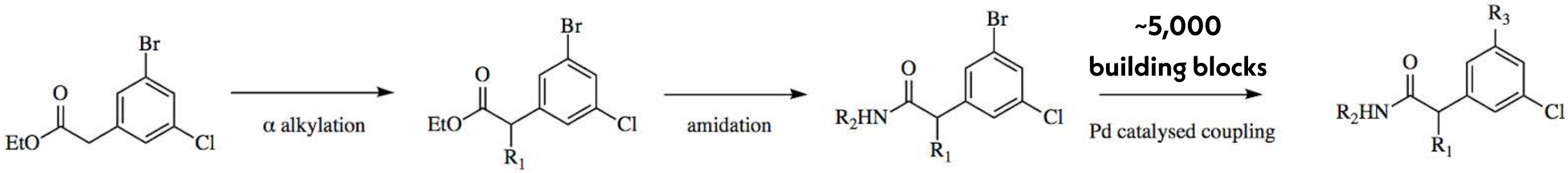
$$W_{na} \propto \frac{q(\boldsymbol{x}_n)}{\sum_{k=1}^{K} N_k\, e^{\hat{f}_k} q_k(\boldsymbol{x}_n)}$$

$$\delta^2 \hat{A} = \hat{A}^2 (\hat{\Theta}_{AA} + \hat{\Theta}_{aa} - 2\,\hat{\Theta}_{Aa})$$

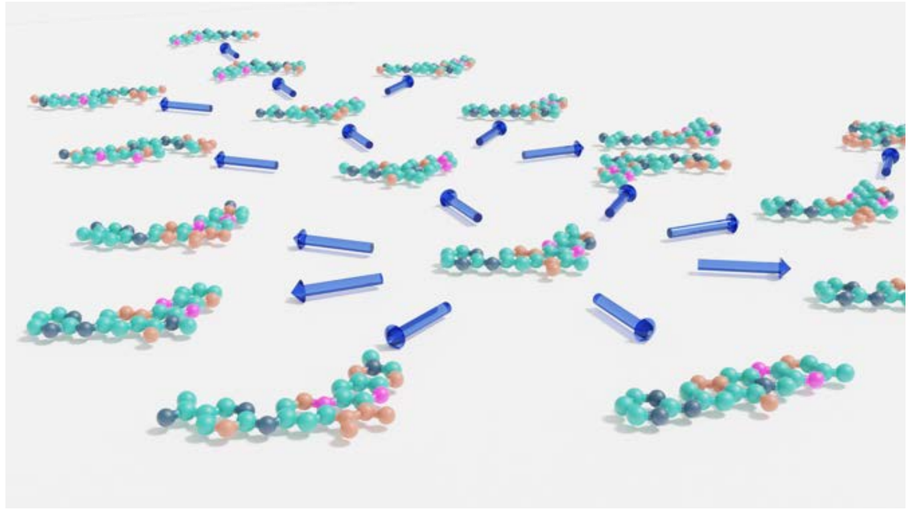
Shirts and Chodera. JCP 129:124105, 2008

ML/MM REPEX/ATM FEP/MBAR RBFE

Relative Binding Free Energy (RBFE) calculations are a useful way to make decisions about which synthetically tractable molecules to make

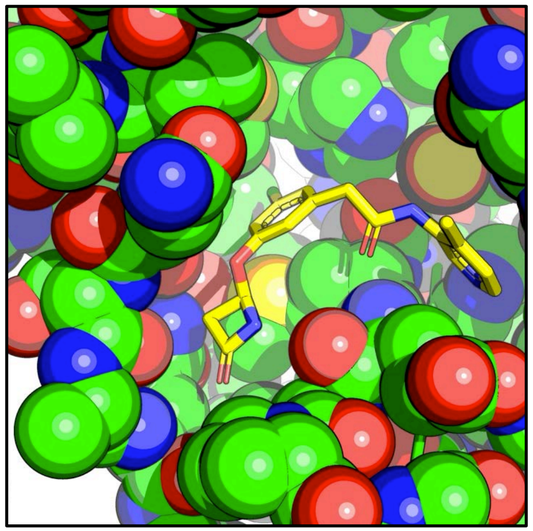


relative alchemical transformation network



Open Free Energy Consortium: https://openfree.energy/

docked poses



Best Practices for Alchemical Free Energy Calculations
https://doi.org/10.33011%2Flivecoms.2.1.18378