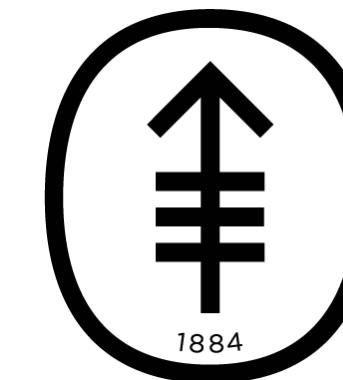


Slides and video licensed under CC-BY 4.0

slides will be posted to: <http://choderalab.org/news>

# TEACHING FREE ENERGY CALCULATIONS TO LEARN



**John D. Chodera**

MSKCC Computational and Systems Biology Program

<http://choderalab.org>

L

L

M

S

**LARGE LATOMSTIC MODELS**



# 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

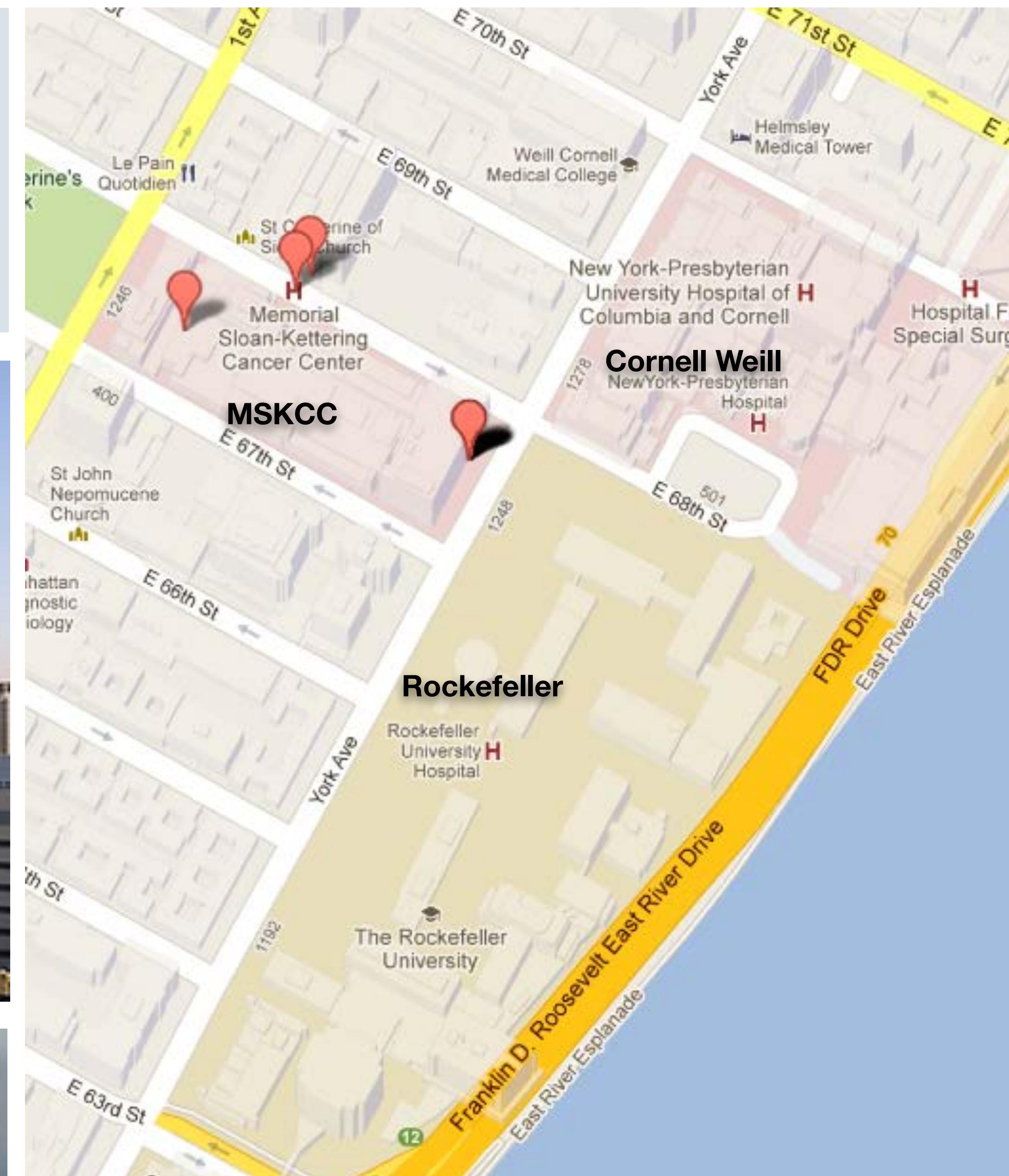
Christina  
Leslie

Joao  
Xavier

Kushal  
Dey

John  
Chodera

Thomas  
Norman



# Computational and Systems Biology

# THE CHODERA LAB INTEGRATES COMPUTATION AND EXPERIMENT TO ADVANCE DRUG DISCOVERY

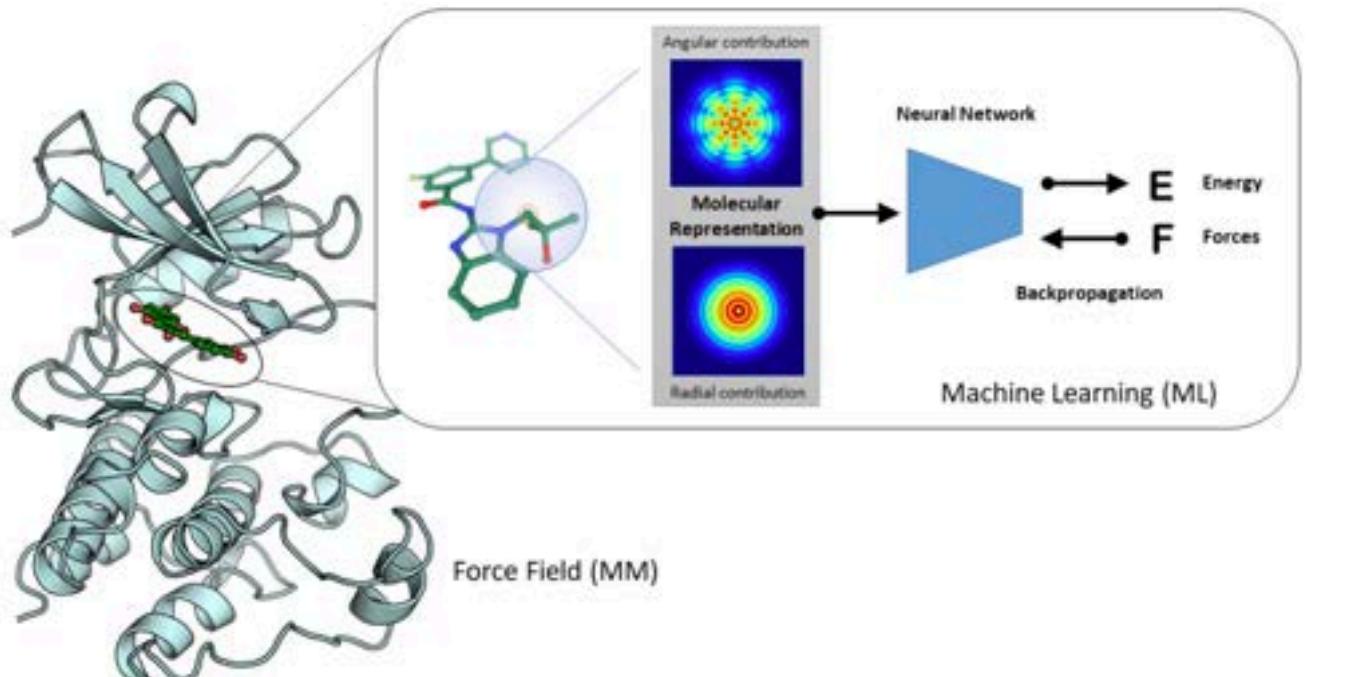
## MODELING



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

FOLDING  
@HOME

amazon  
web services™ EC2



CLOUD LABORATORIES (STRATEOS)

## AUTOMATION

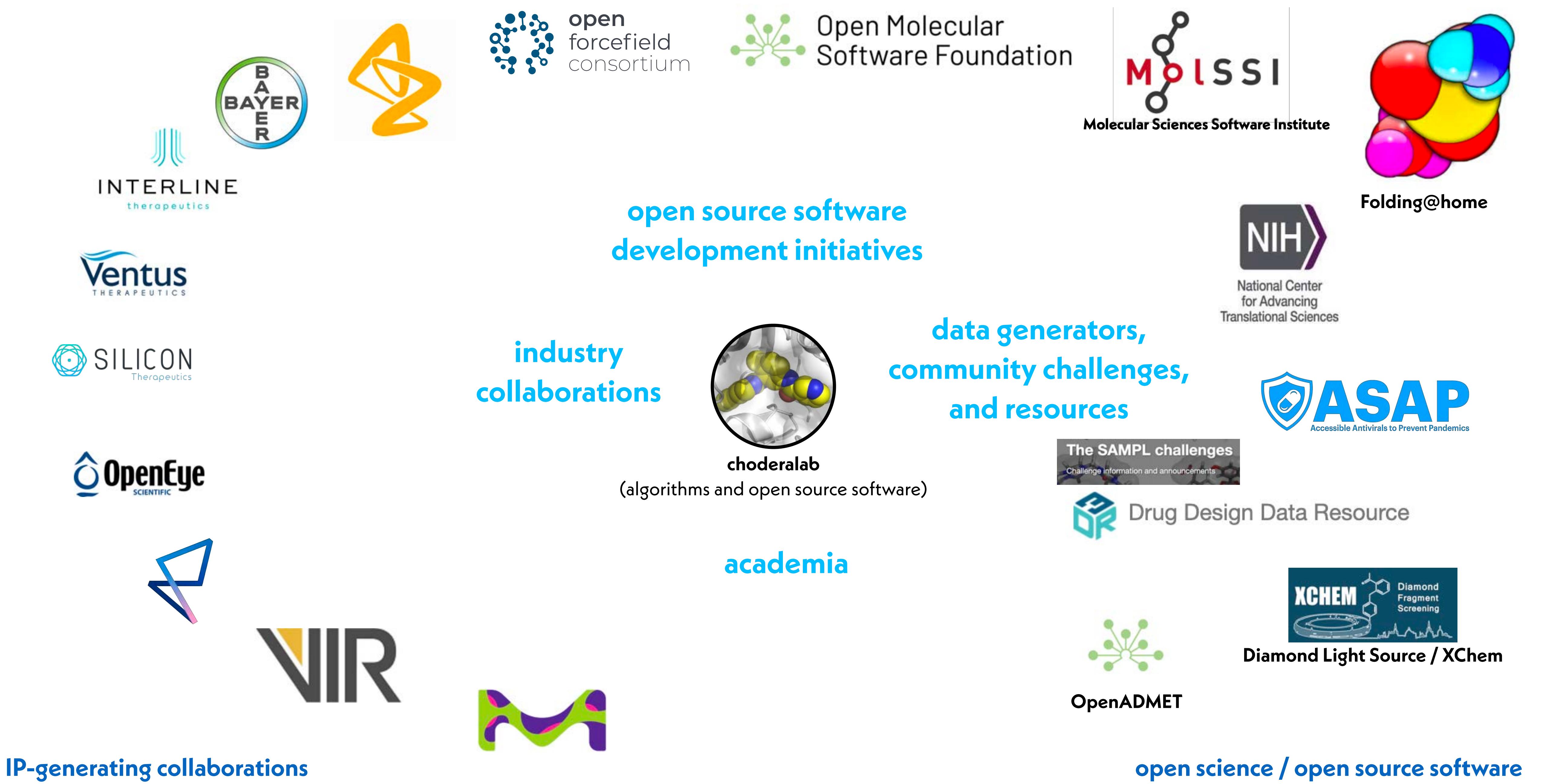


CHODERA LAB, Z17

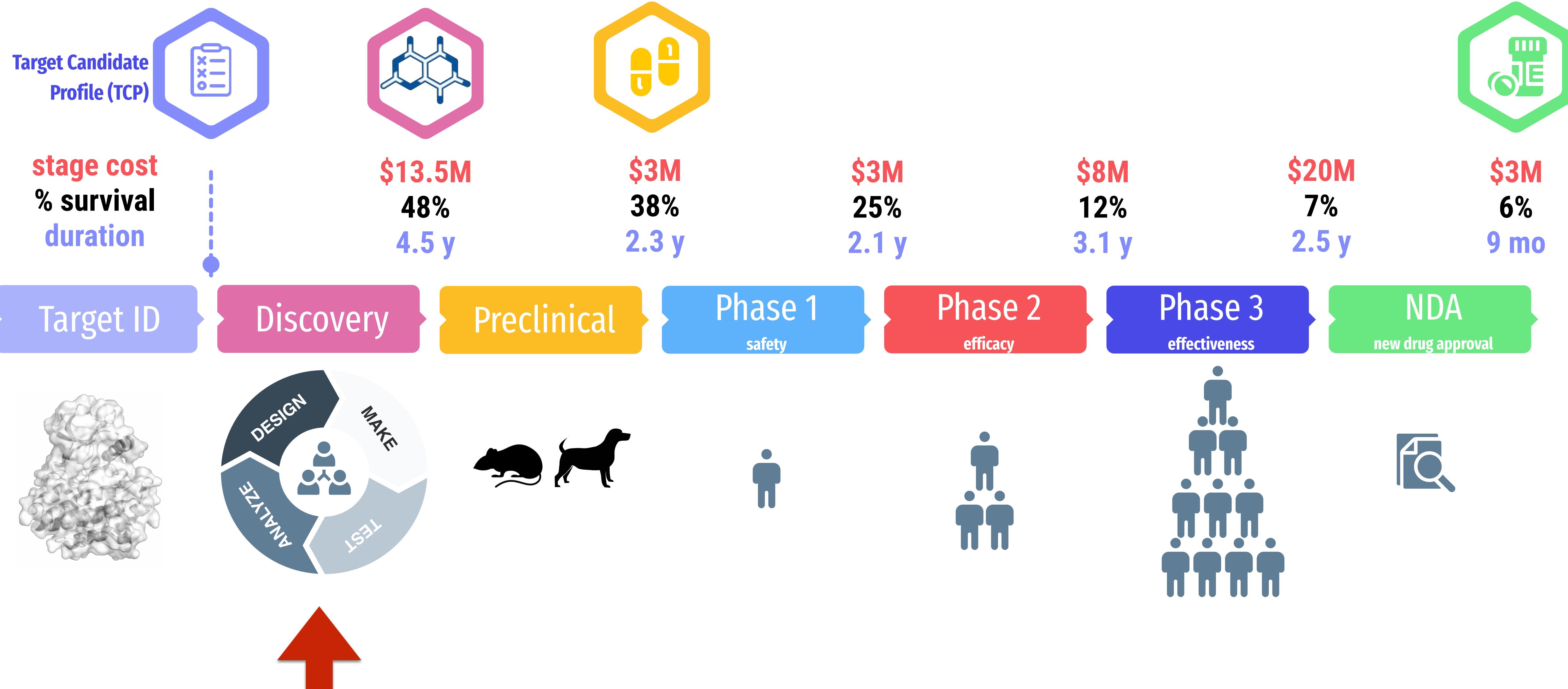


OPENTRONS

# WE COLLABORATE BROADLY TO IMPROVE THE DRUG DISCOVERY ECOSYSTEM

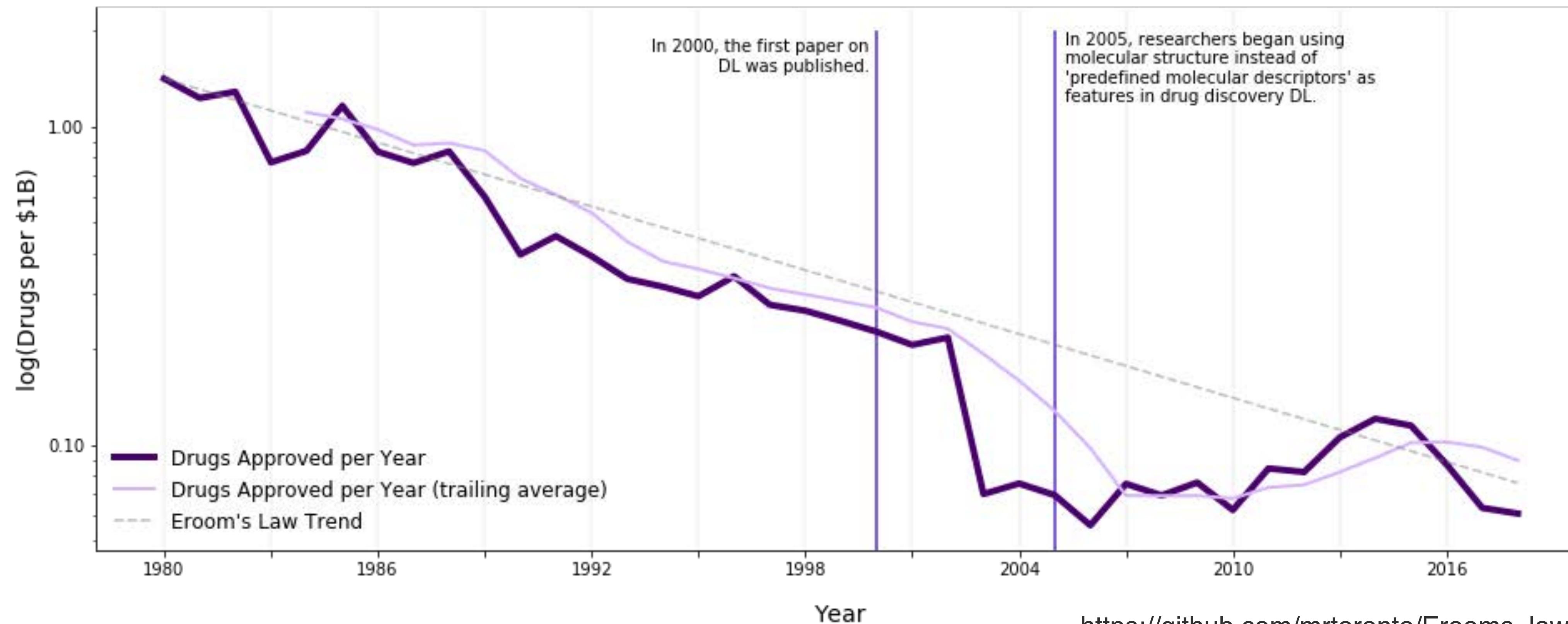


# DRUG DISCOVERY AND DEVELOPMENT IS SLOW, COSTLY, AND PRONE TO FAILURE



# AND IT'S GETTING WORSE, NOT BETTER

## Efficiency of Drug Discovery

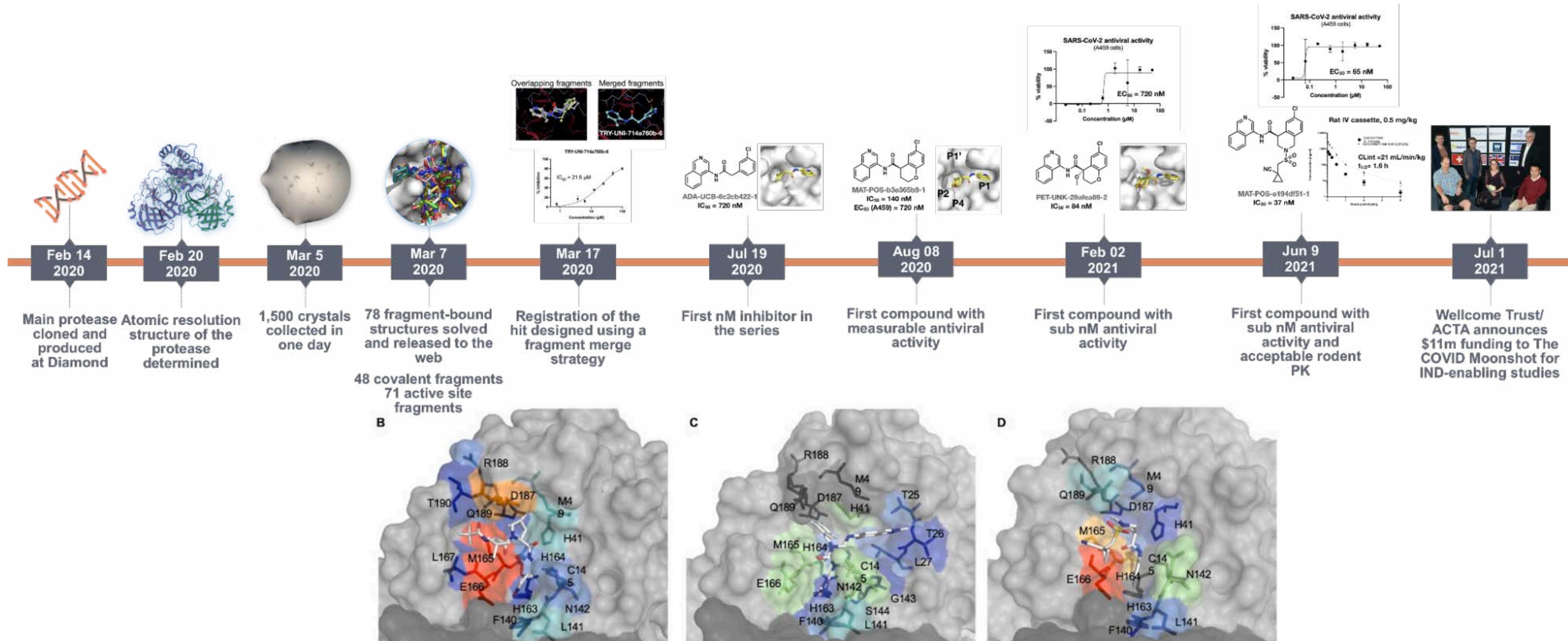


cost per approved drug (including costs of failures) estimated > \$5B per approved drug in 2016 [1]

# 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



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>



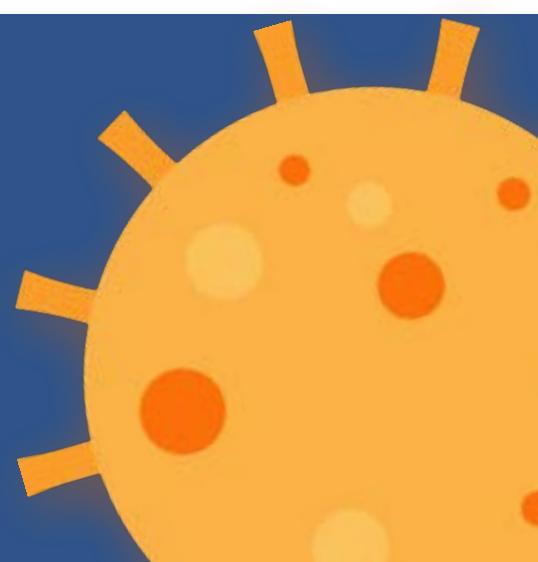
# DRUG DISCOVERY IS A COMPLEX MULTI-OBJECTIVE DESIGN PROBLEM

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

Property	Target range	Rationale
protease assay	$IC_{50} < 10 \text{ nM}$	Extrapolation from other anti-viral programs
viral replication assay	$EC_{50} < 5 \mu\text{M}$	Suppression of virus at achievable blood levels
plaque reduction assay	$EC_{50} < 5 \mu\text{M}$	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 toxicological delays to development
	No significant DDI - clean in 5 CYP450 isoforms	DDI aims to deal with co-morbidities / therapies,
	hERG and NaV1.5 $IC_{50} > 50 \mu\text{M}$	cardiac safety for COVID-19 risk profile
	No significant change in QTc	cardiac safety for COVID-19 risk profile
	Ames negative	Low carcinogenicity risk reduces delays in manufacturing
	No mutagenicity or teratogenicity risk	Patient group will include significant proportion of women of childbearing age



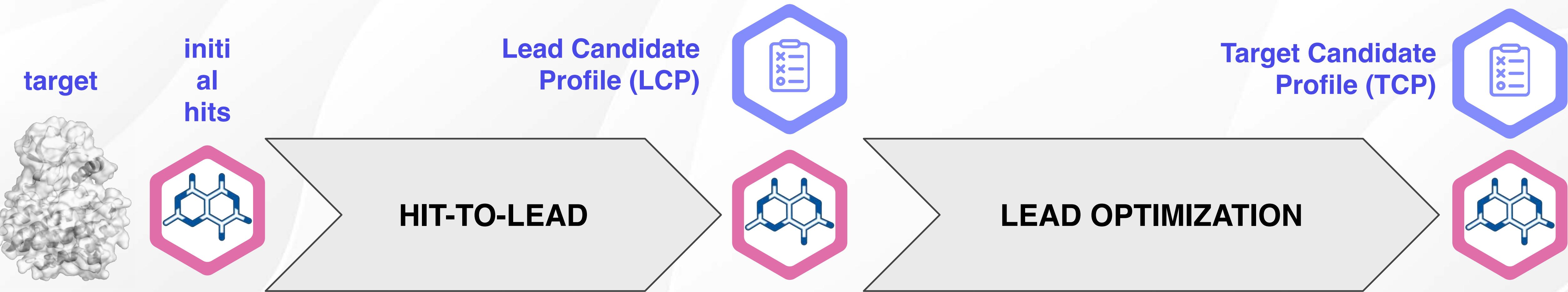
An international effort to  
DISCOVER A COVID ANTIVIRAL



<https://doi.org/10.1101/2020.10.29.339317>

<https://covid.postera.ai/covid>

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

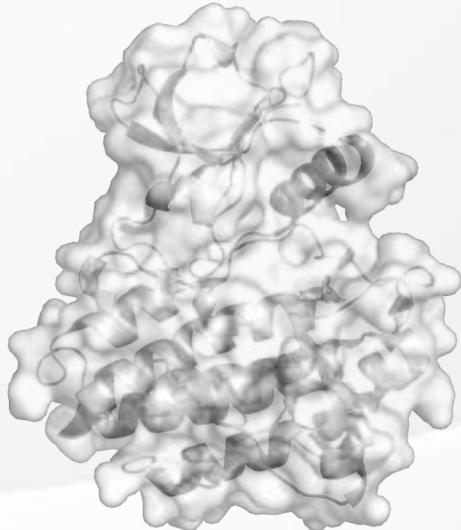


currently ~3.5 years, ~2000 compounds, \$12.5M, high failure rate [1]

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)  
+



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...**

Text/image datasets

**13T**  
GPT-4

**300B**  
GPT-3

**5B**  
DALL-E 3

**650M**  
DALL-E 2

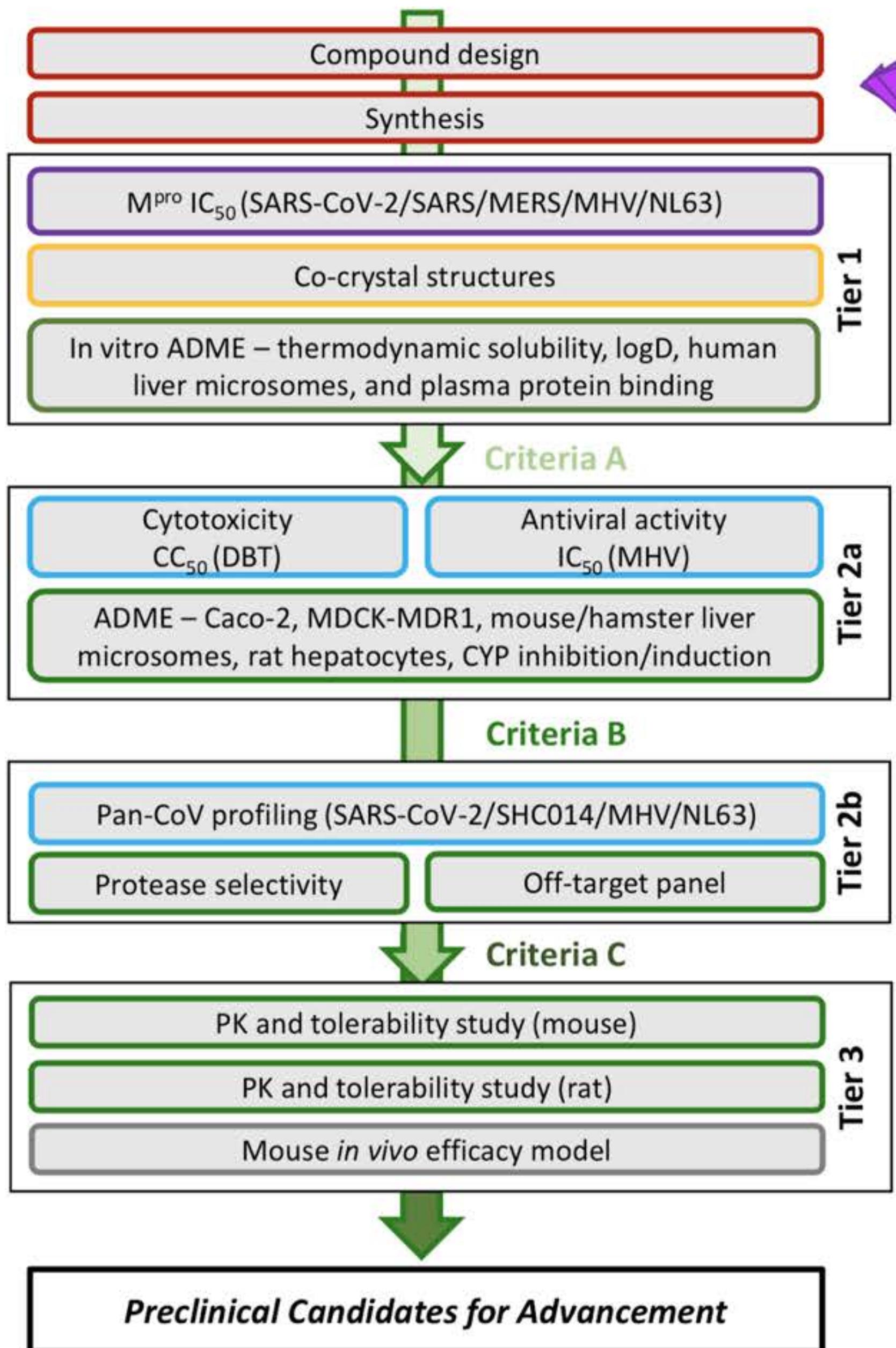
Drug discovery datasets

**20M**  
All ChEMBL  
bioassay data

- 223K Complete protein data bank (PDB) ~\$20B
- 2K Typical drug discovery campaign ~\$12.5M

...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?

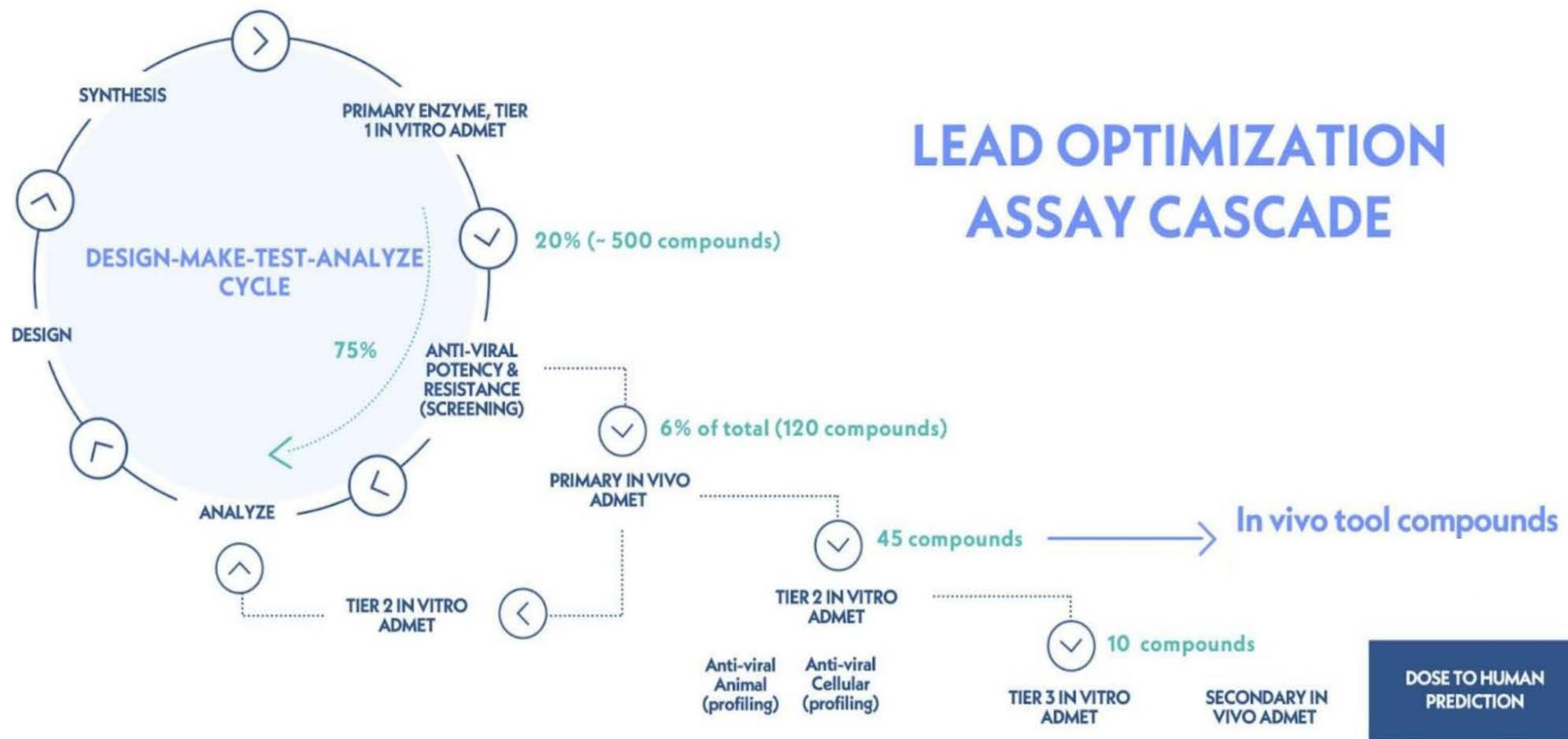
Could it cause bad side effects?

Can oral dosing deliver sufficient drug?

Does it actually work against the disease in animals?

## MOTIVATION

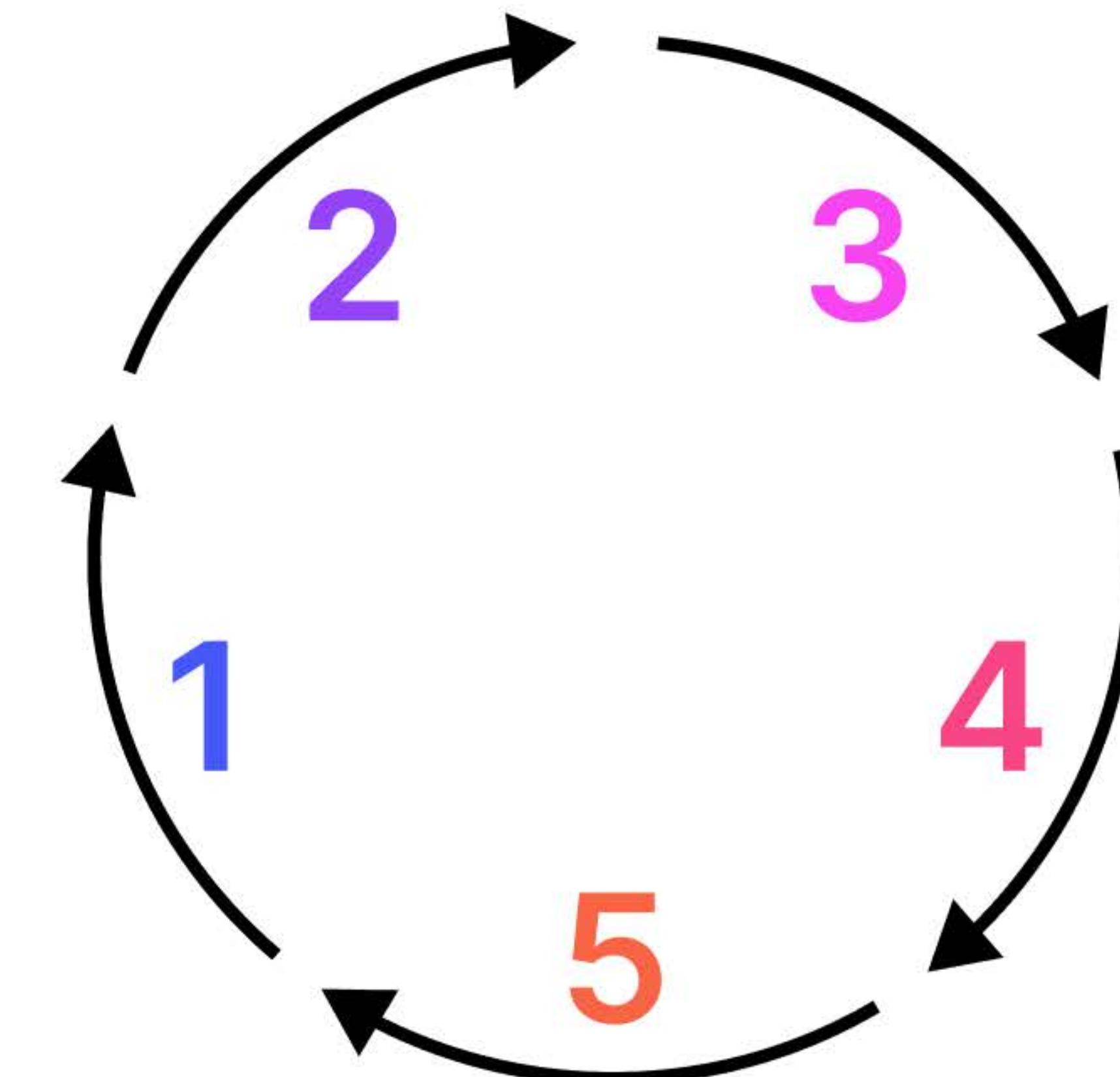
Drug discovery is iterative



## MOTIVATION

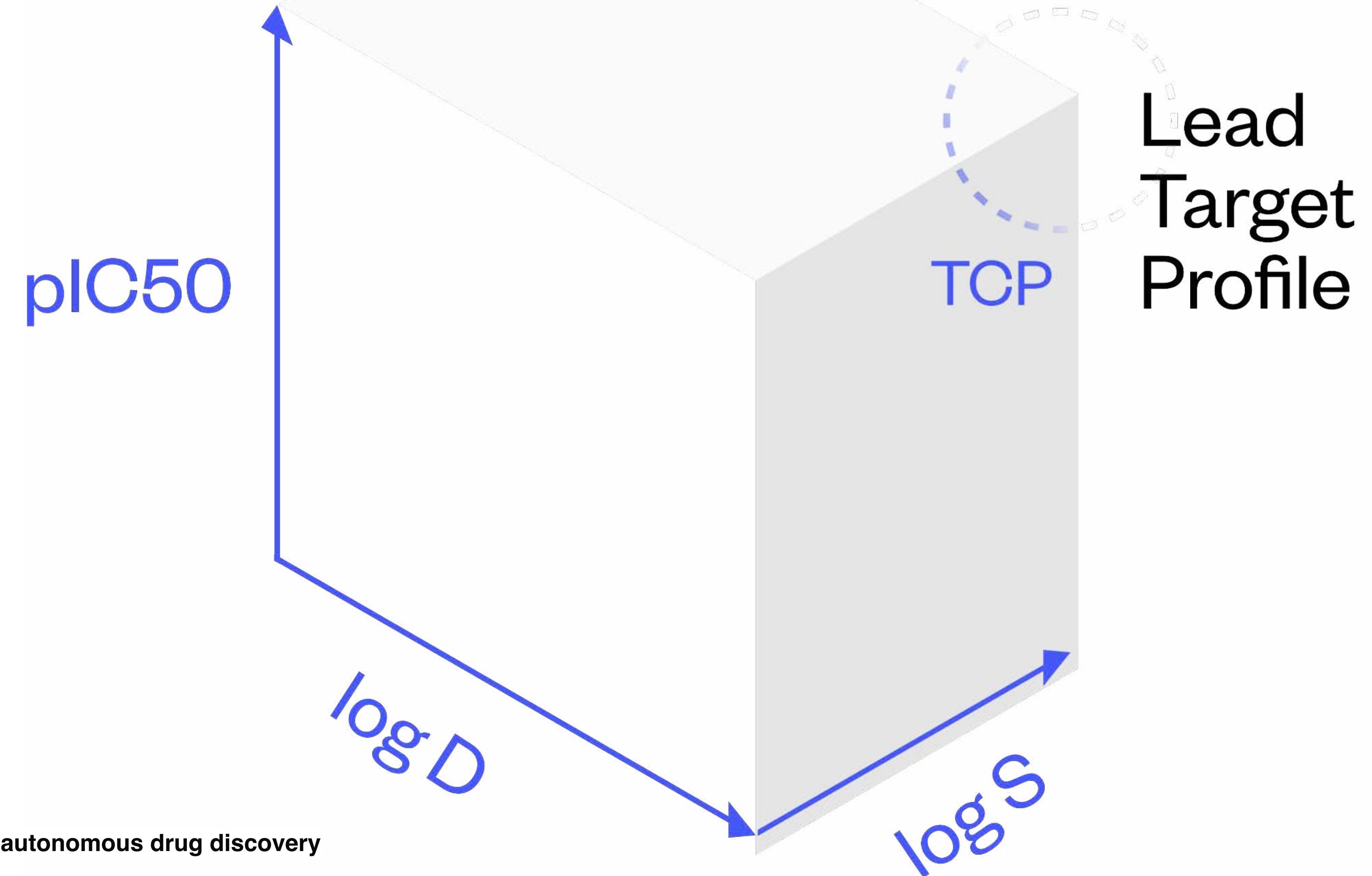
Drug discovery is iterative

- 1 **ideate**
- 2 **triage**
- 3 **synthesize**
- 4 **assay**
- 5 **learn**



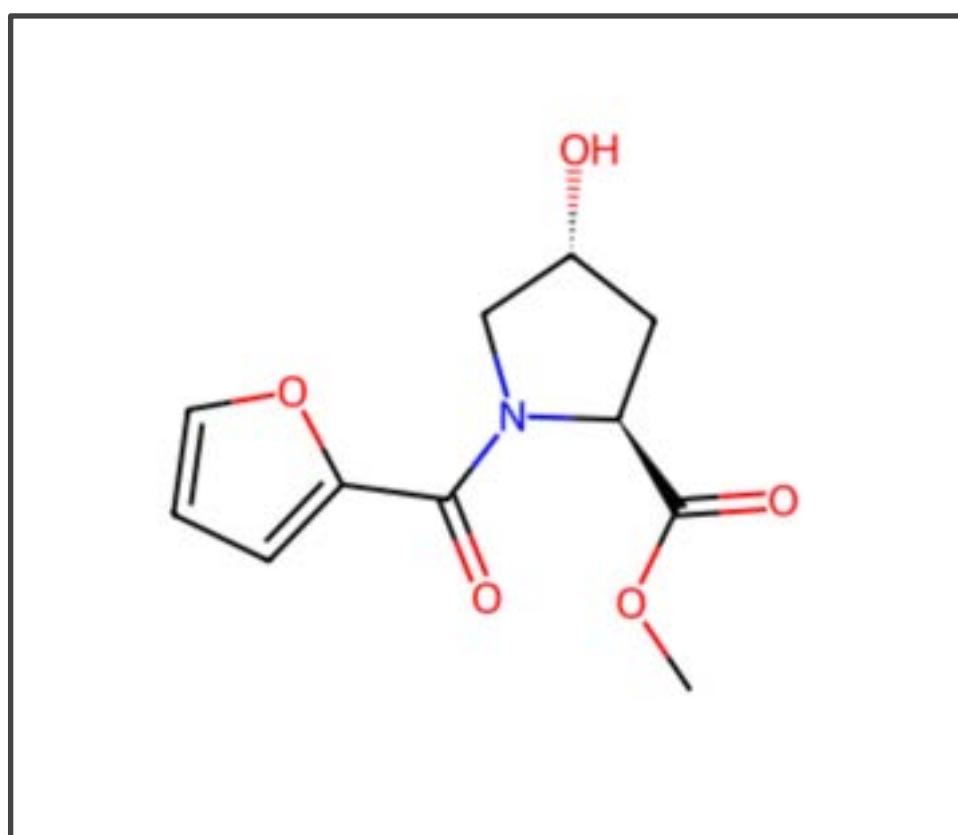
## MOTIVATION

Drug discovery is stochastic



## MOTIVATION

Drug discovery is stochastic



pIC50

initial hits

log D

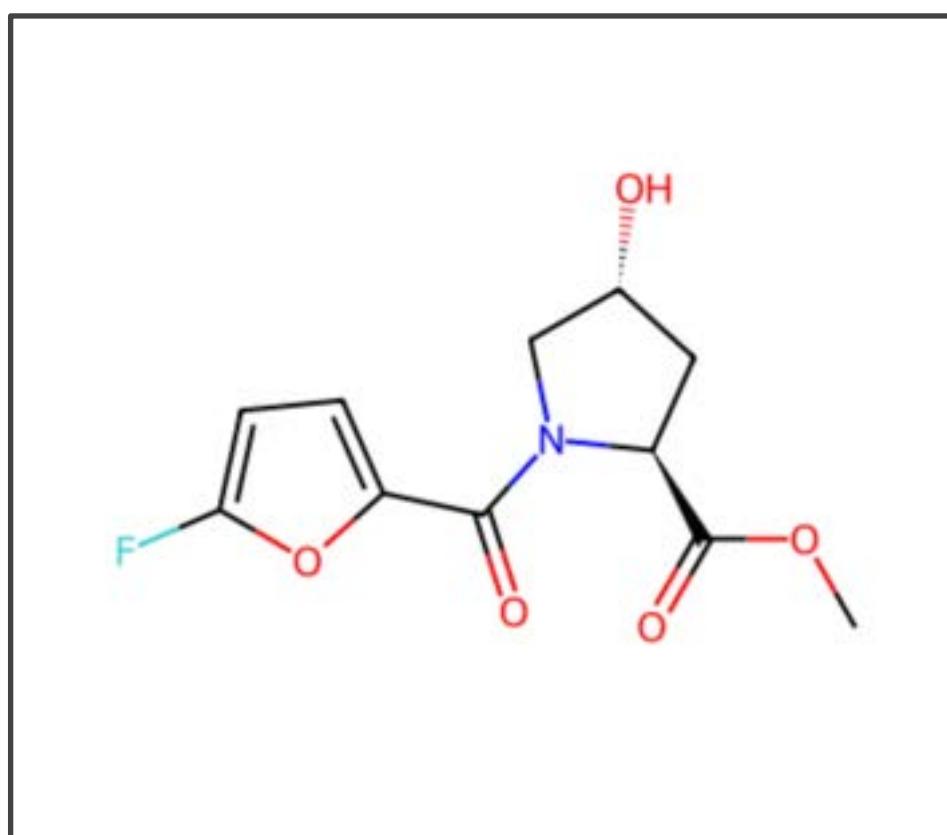
TCP

log S

Lead  
Target  
Profile

## MOTIVATION

Drug discovery is stochastic



pIC<sub>50</sub>

initial hits

log D

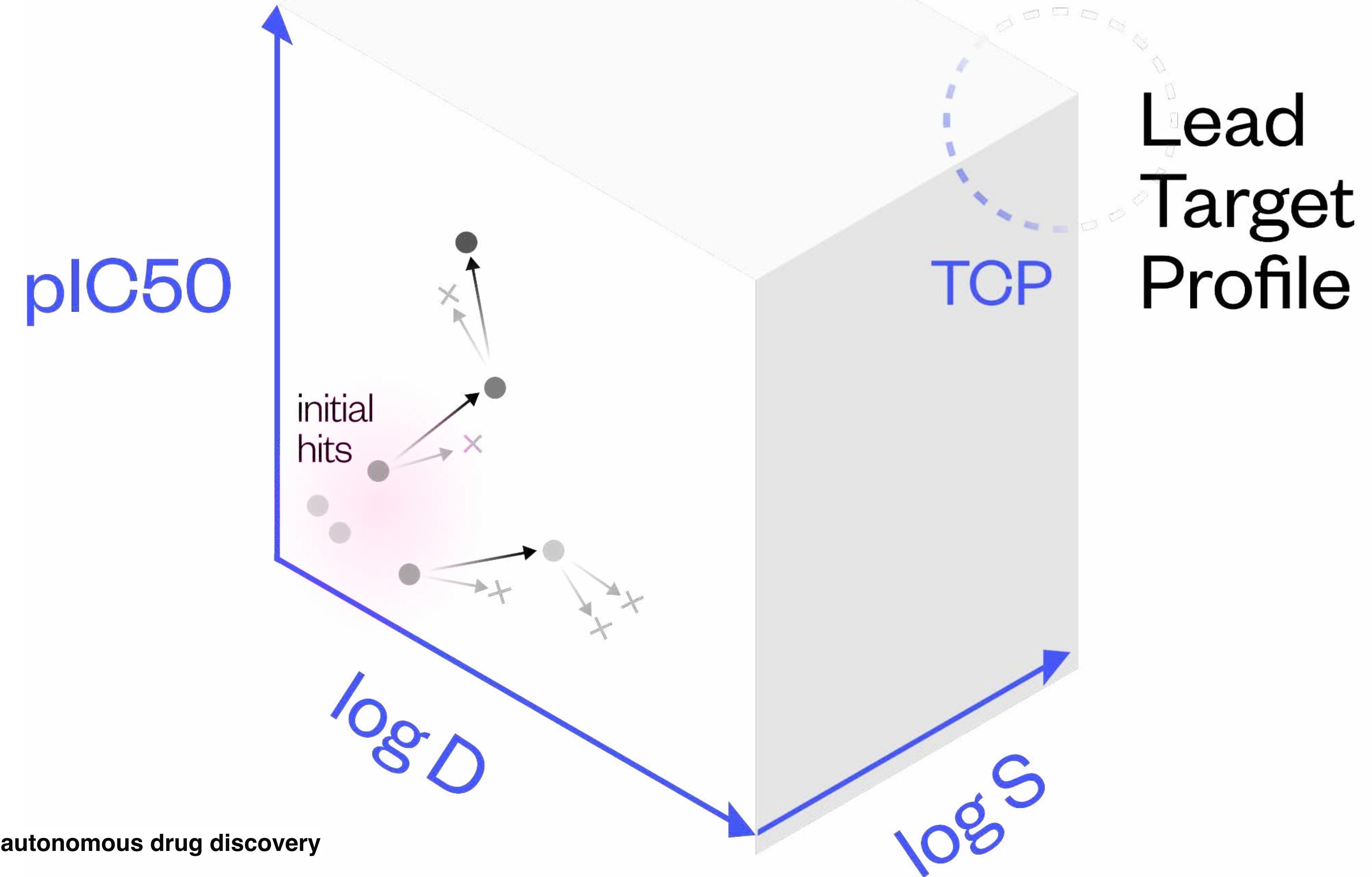
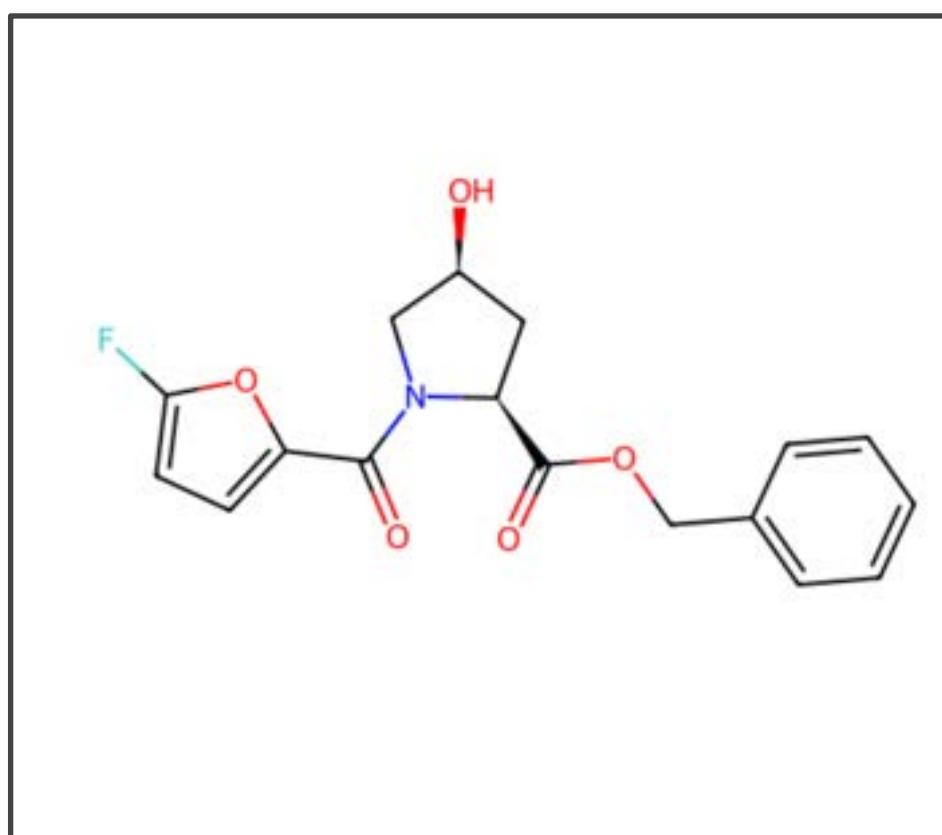
TCP

log S

Lead  
Target  
Profile

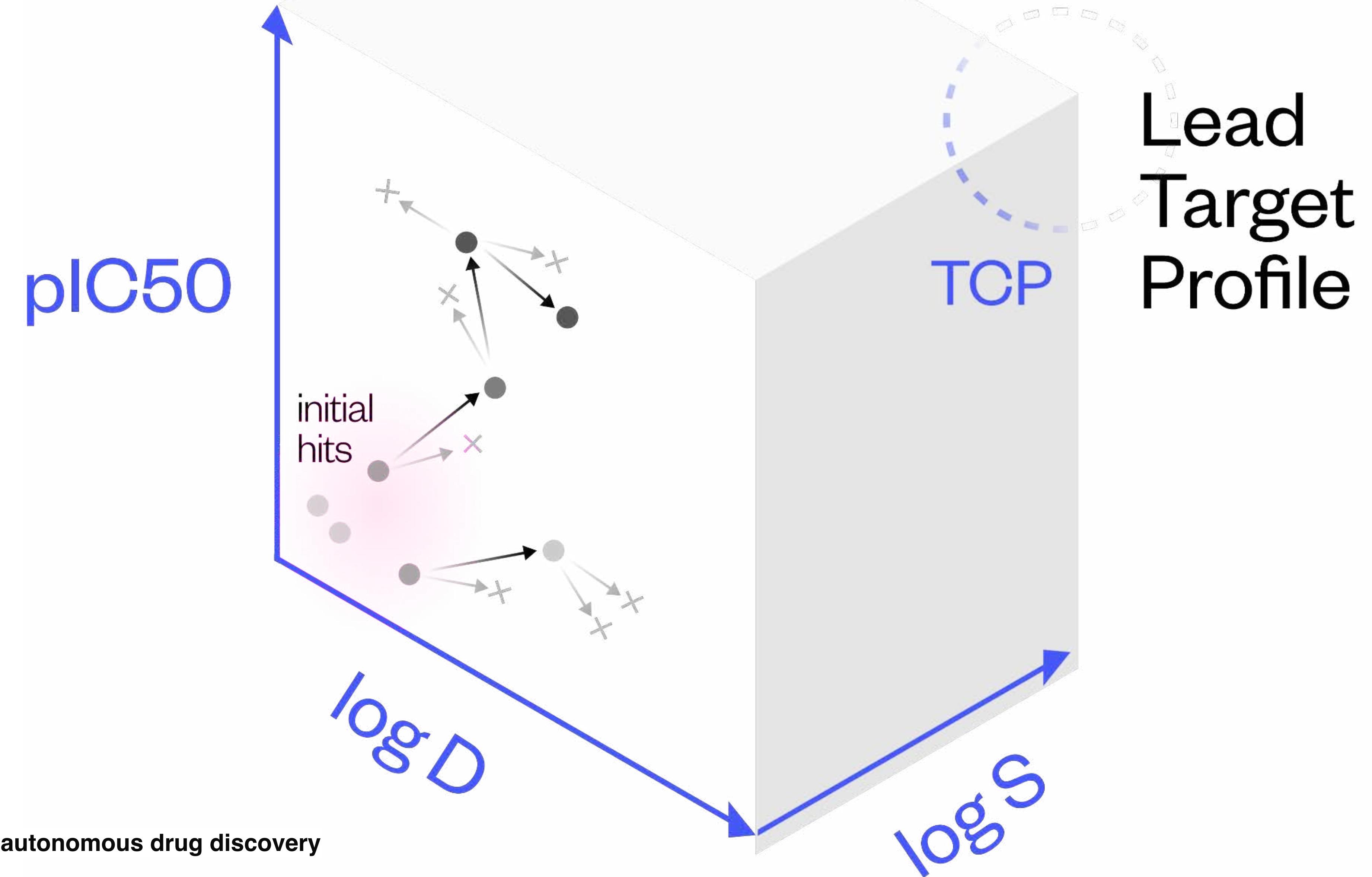
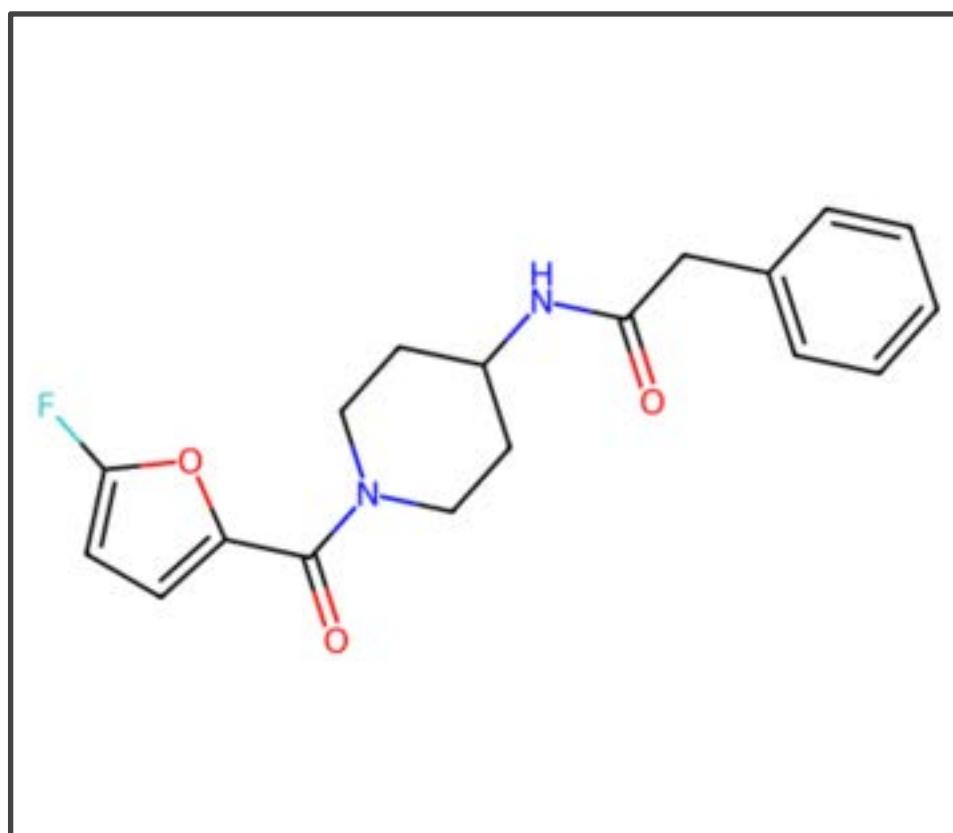
## MOTIVATION

Drug discovery is stochastic



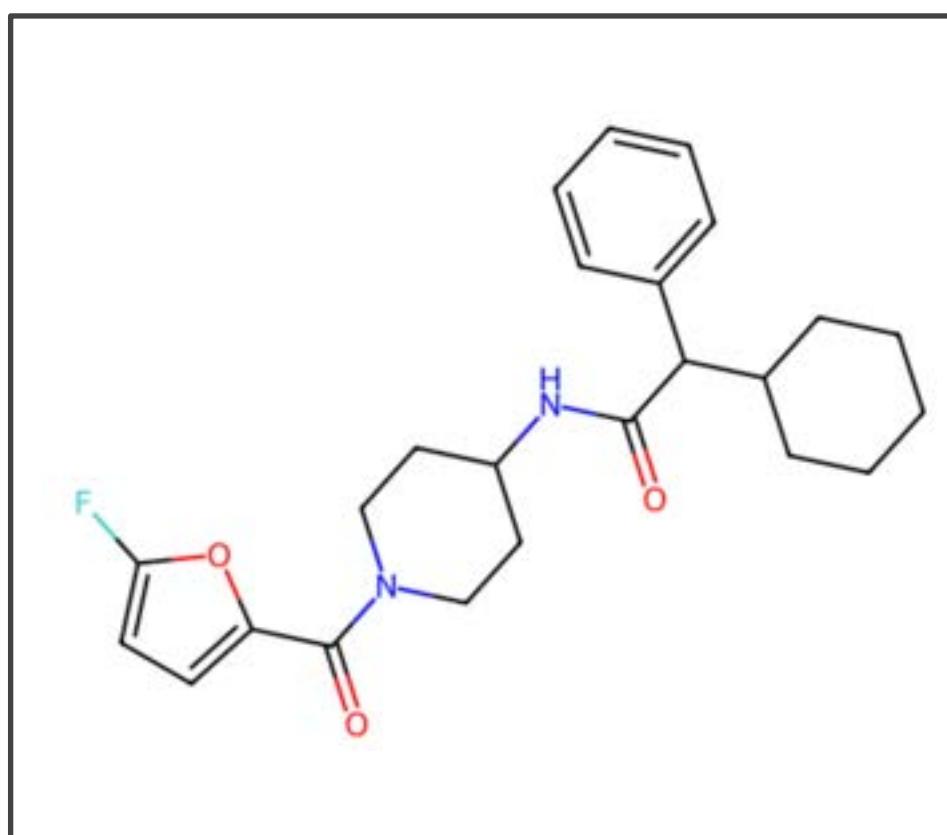
## MOTIVATION

Drug discovery is stochastic



## MOTIVATION

Drug discovery is stochastic



pIC50

initial hits

log D

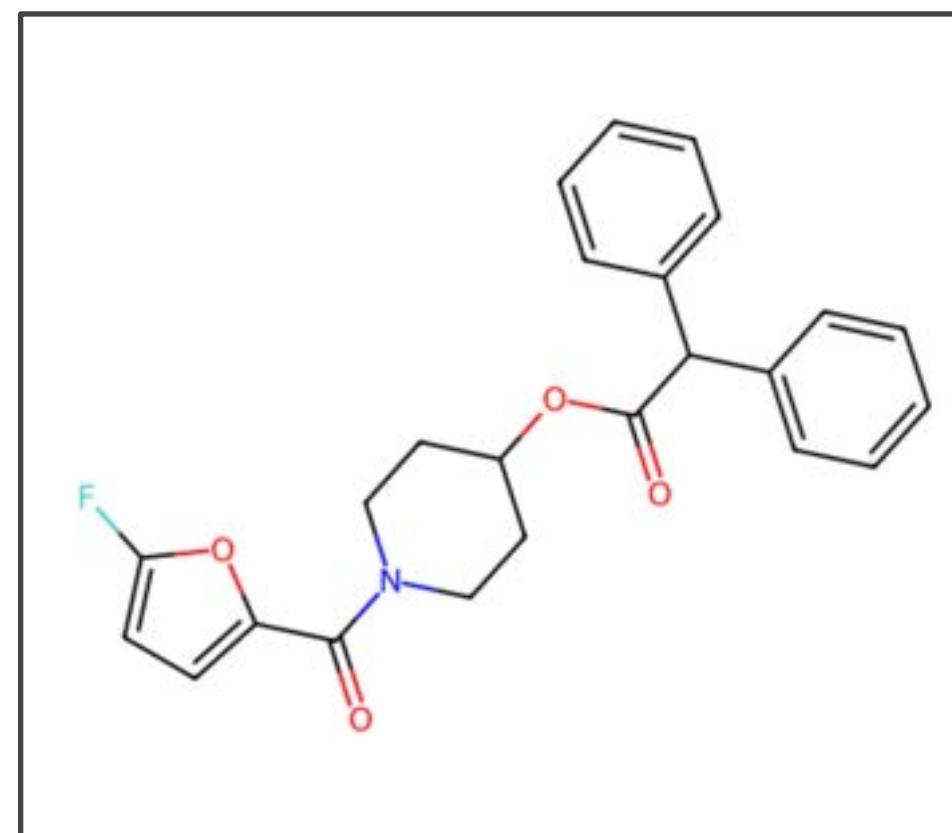
log S

TCP

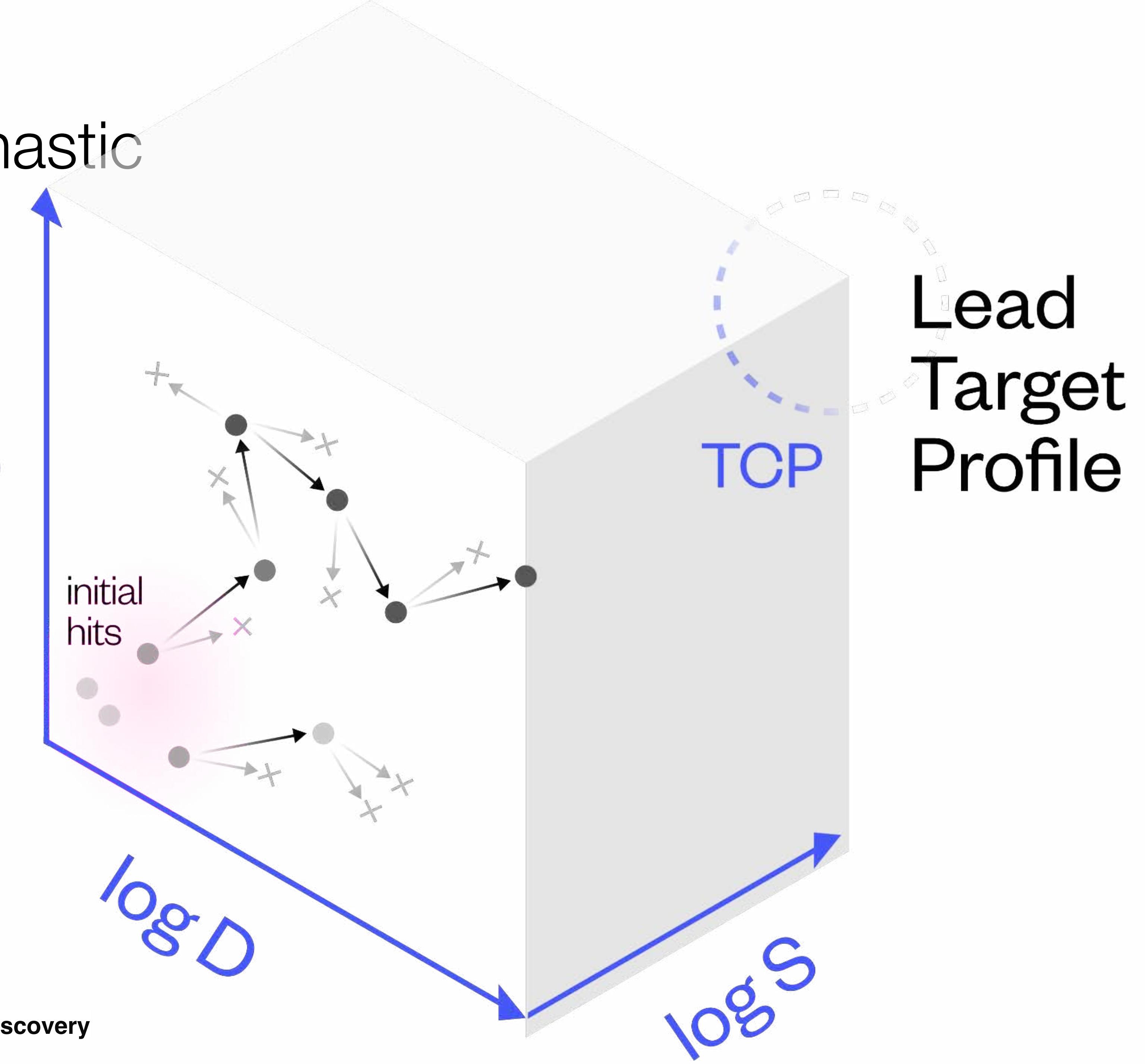
Lead  
Target  
Profile

## MOTIVATION

Drug discovery is stochastic



pIC50

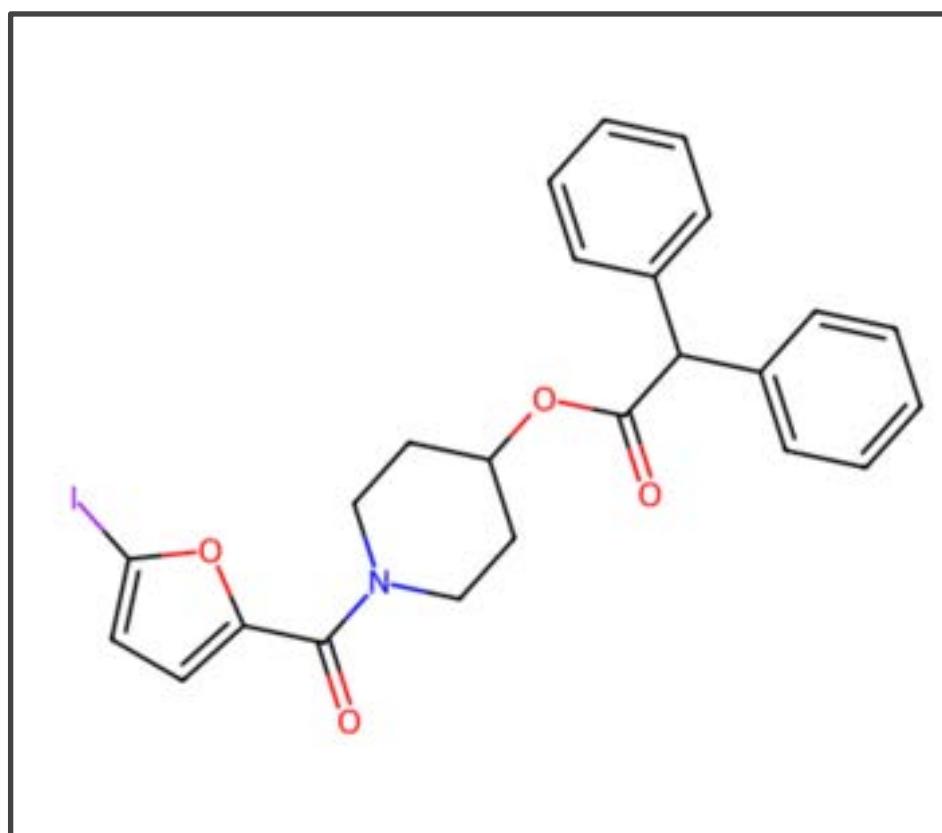


Lead  
Target  
Profile

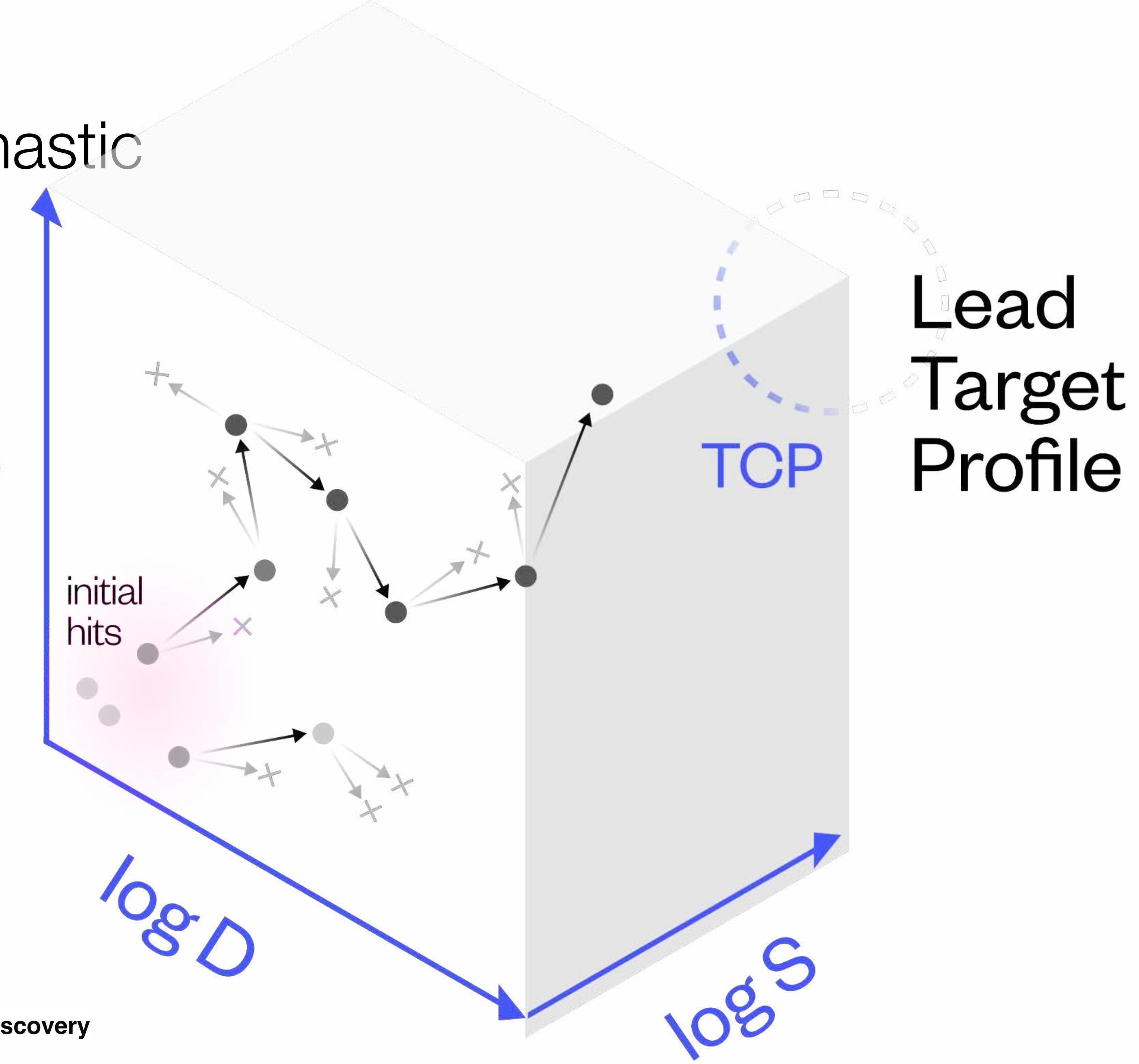
TCP

## MOTIVATION

Drug discovery is stochastic

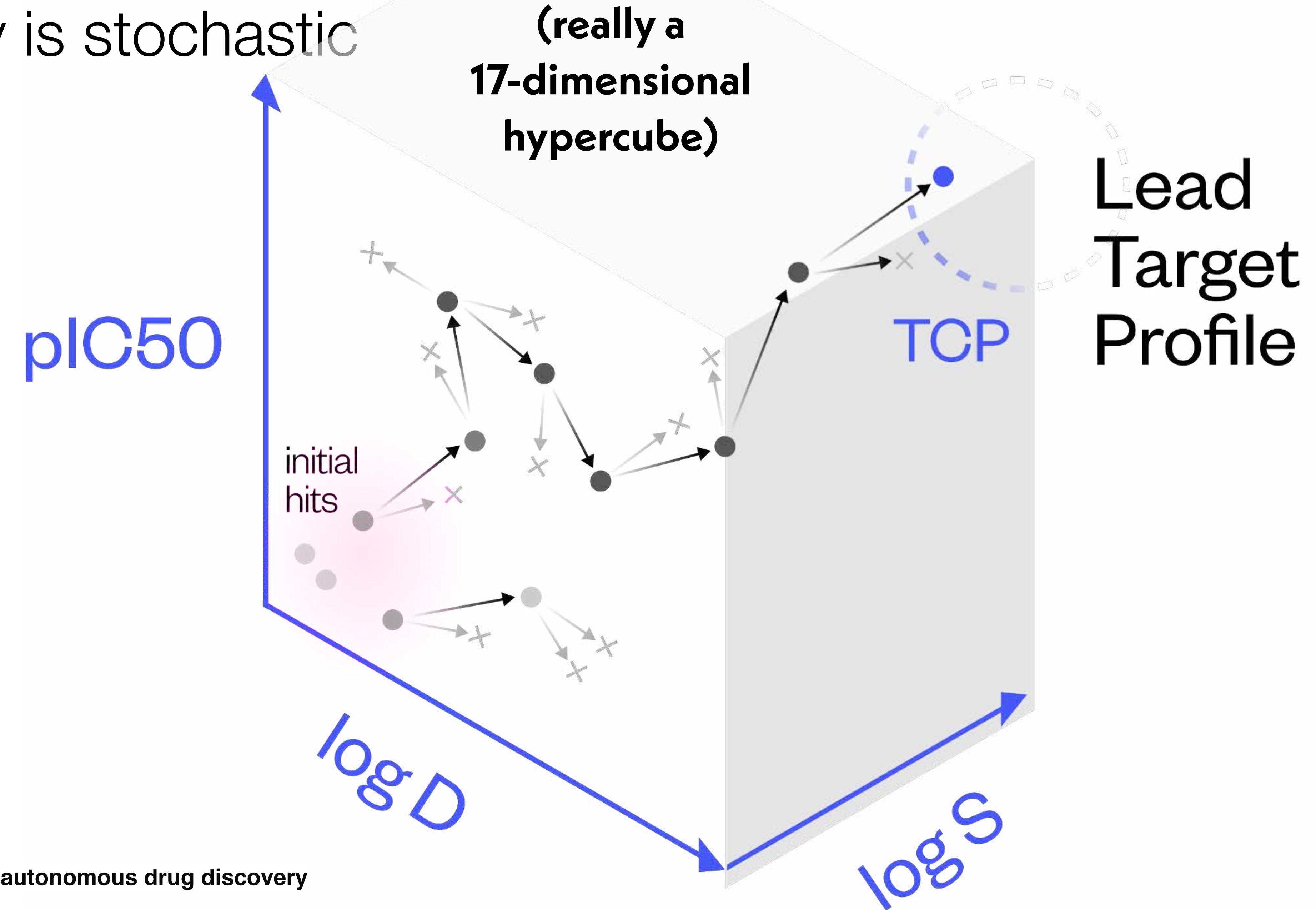
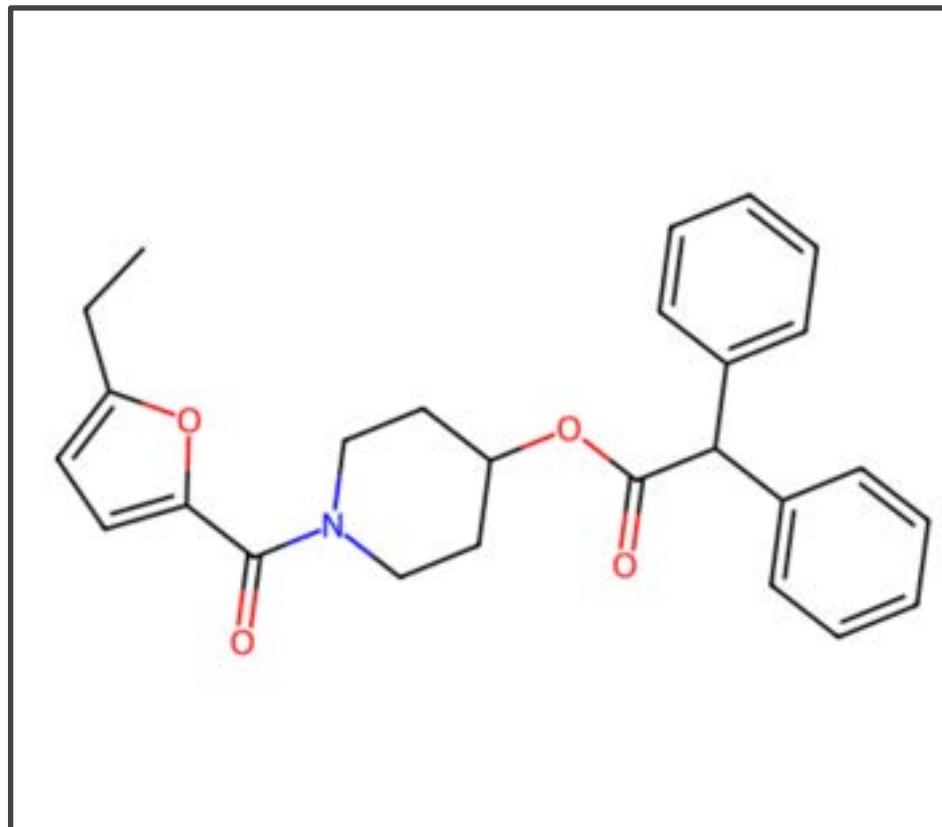


pIC50

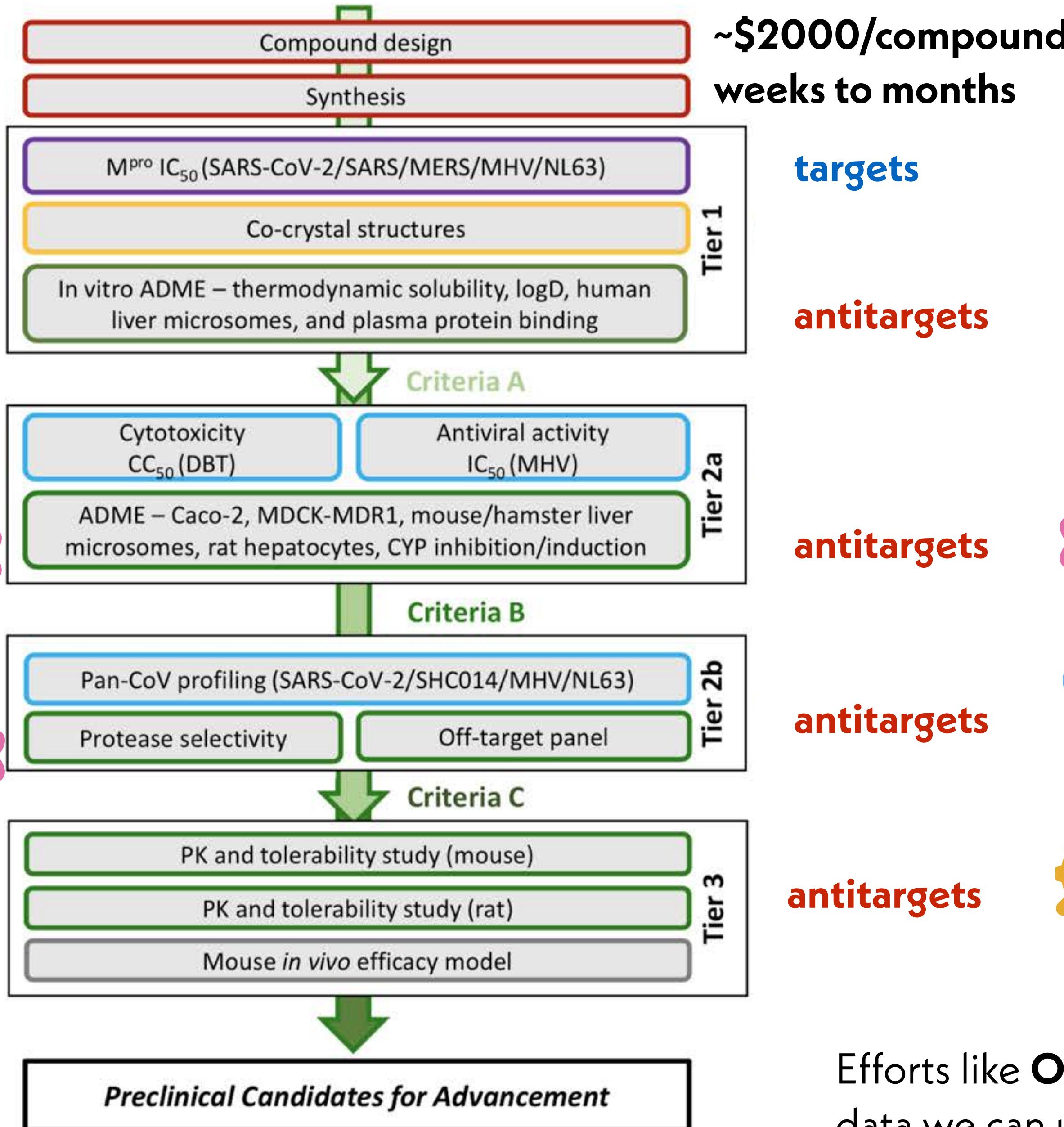
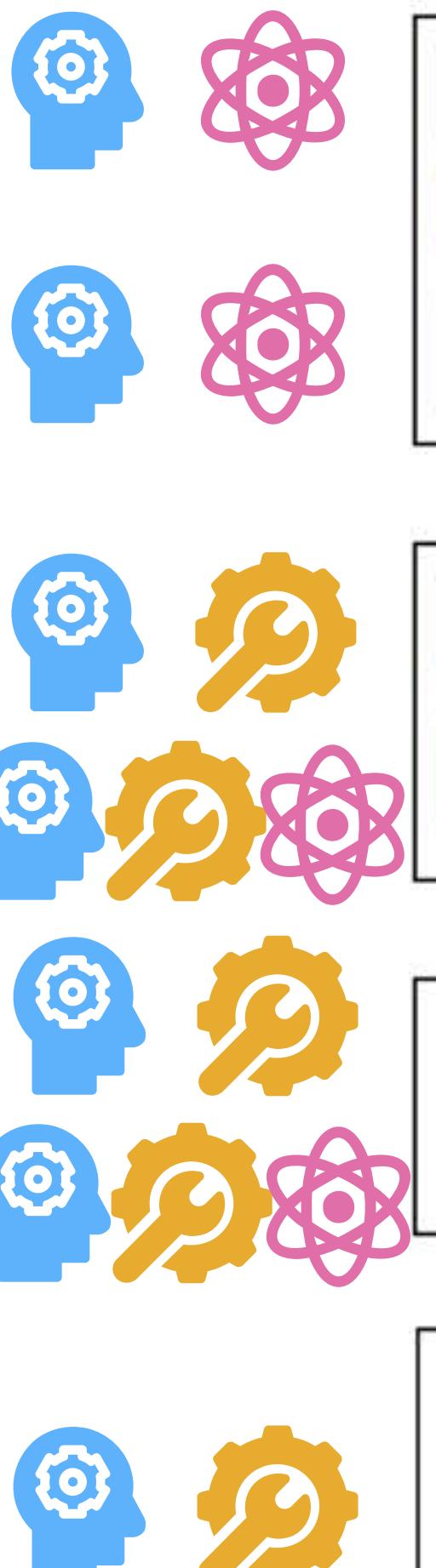


## MOTIVATION

Drug discovery is stochastic



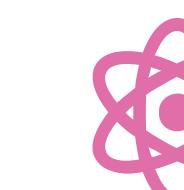
# ACCURATE SURROGATE MODELS COULD SIGNIFICANTLY ACCELERATE DISCOVERY



targets

antitargets

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



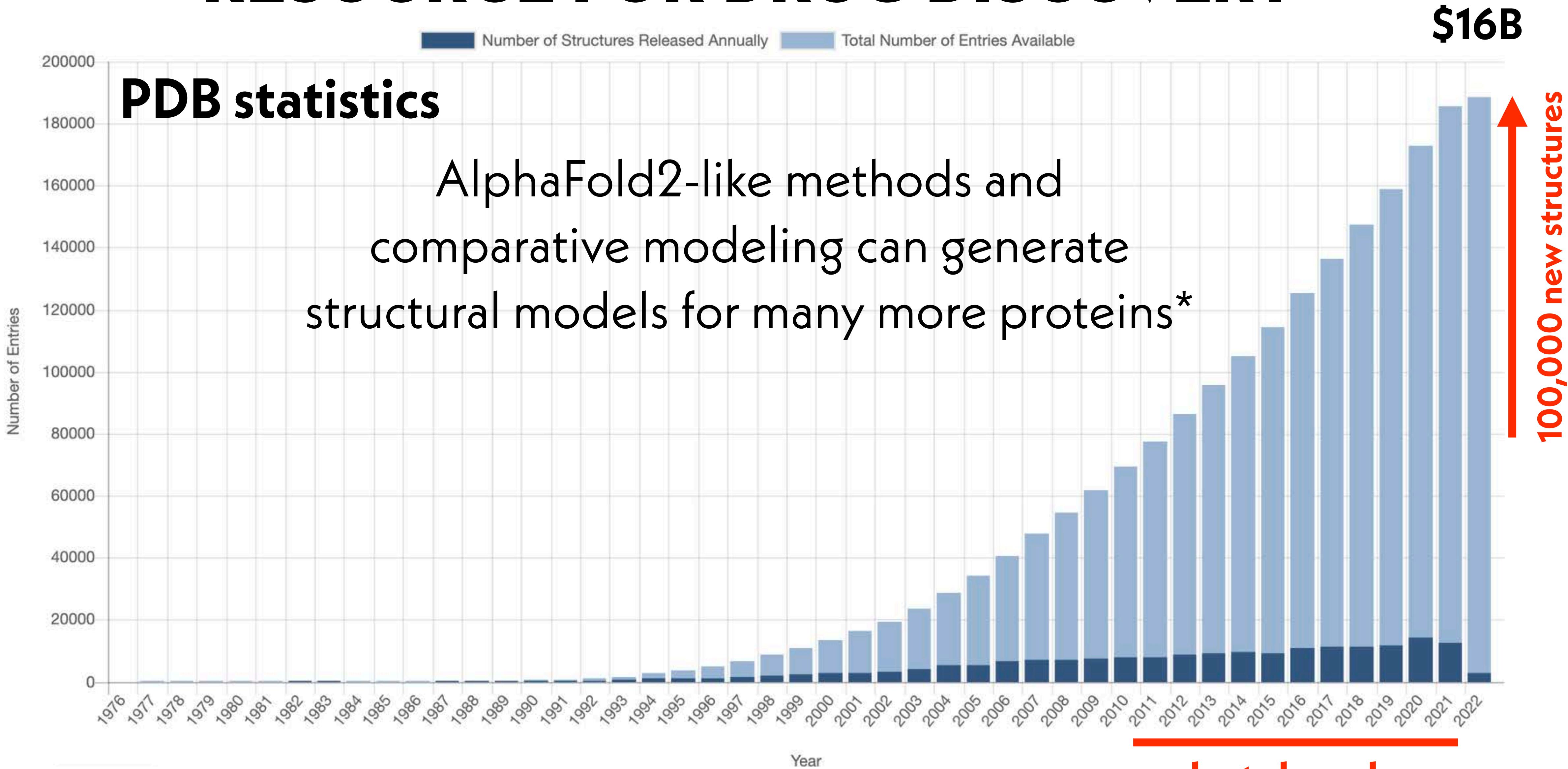
**ligand-based machine learning models** feasible when data is not target-dependent



**mechanistic machine learning models** feasible when biological pathways are understood

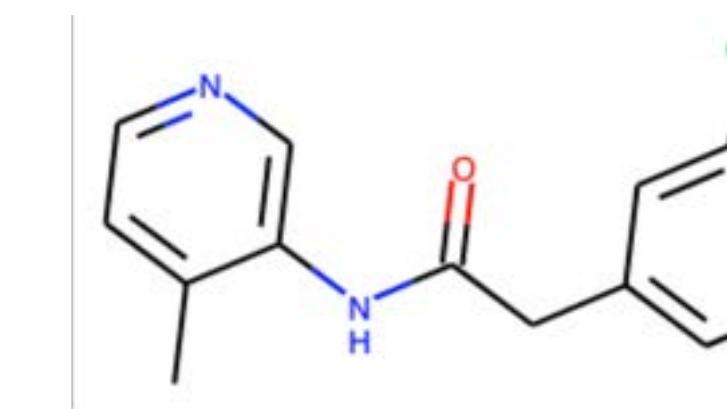
Efforts like **OpenADMET** will furnish a huge amount of structural data we can use for modeling antitarget-based properties

# STRUCTURAL DATA IS NOW AN ABUNDANT RESOURCE FOR DRUG DISCOVERY

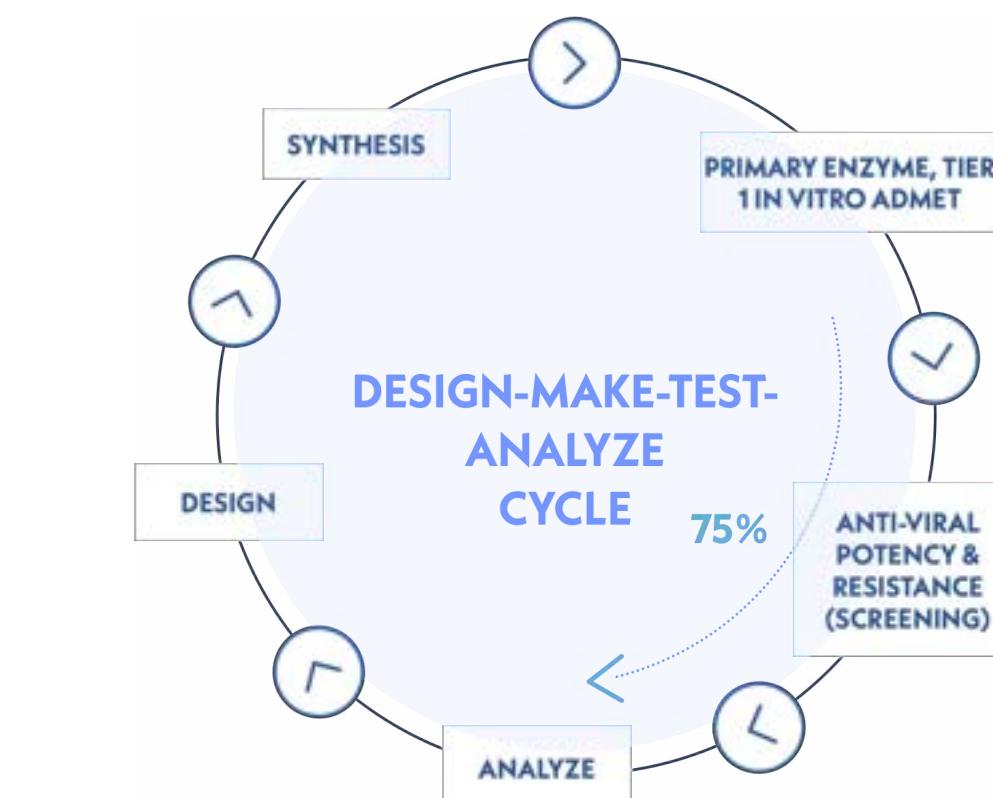


# DESIGN-MAKE-TEST-ANALYZE CYCLES SHARE A COMMON OPERATION:

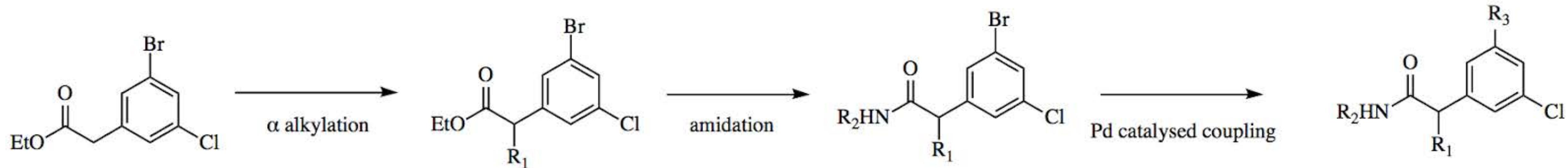
1. Select a current **lead molecule**



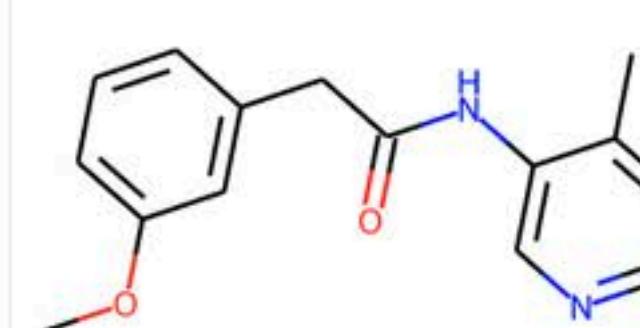
$IC_{50} = 25 \mu M$   
TRY-UNI-714a760b-6



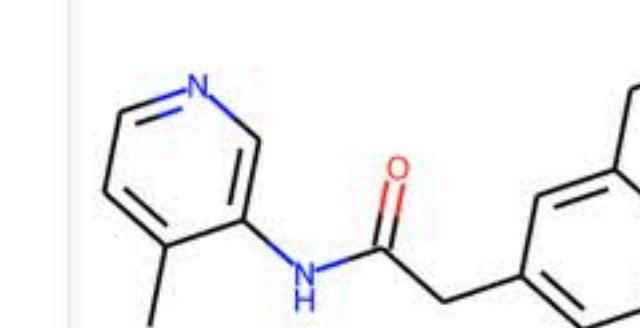
2. Use AI tools to identify a **retrosynthetic pathway** capable of installing new groups to replace part of the molecule



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



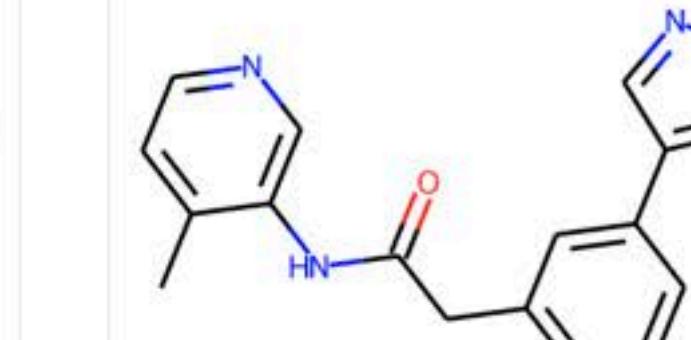
EDJ-MED-e58735b6-1



EDJ-MED-e58735b6-2

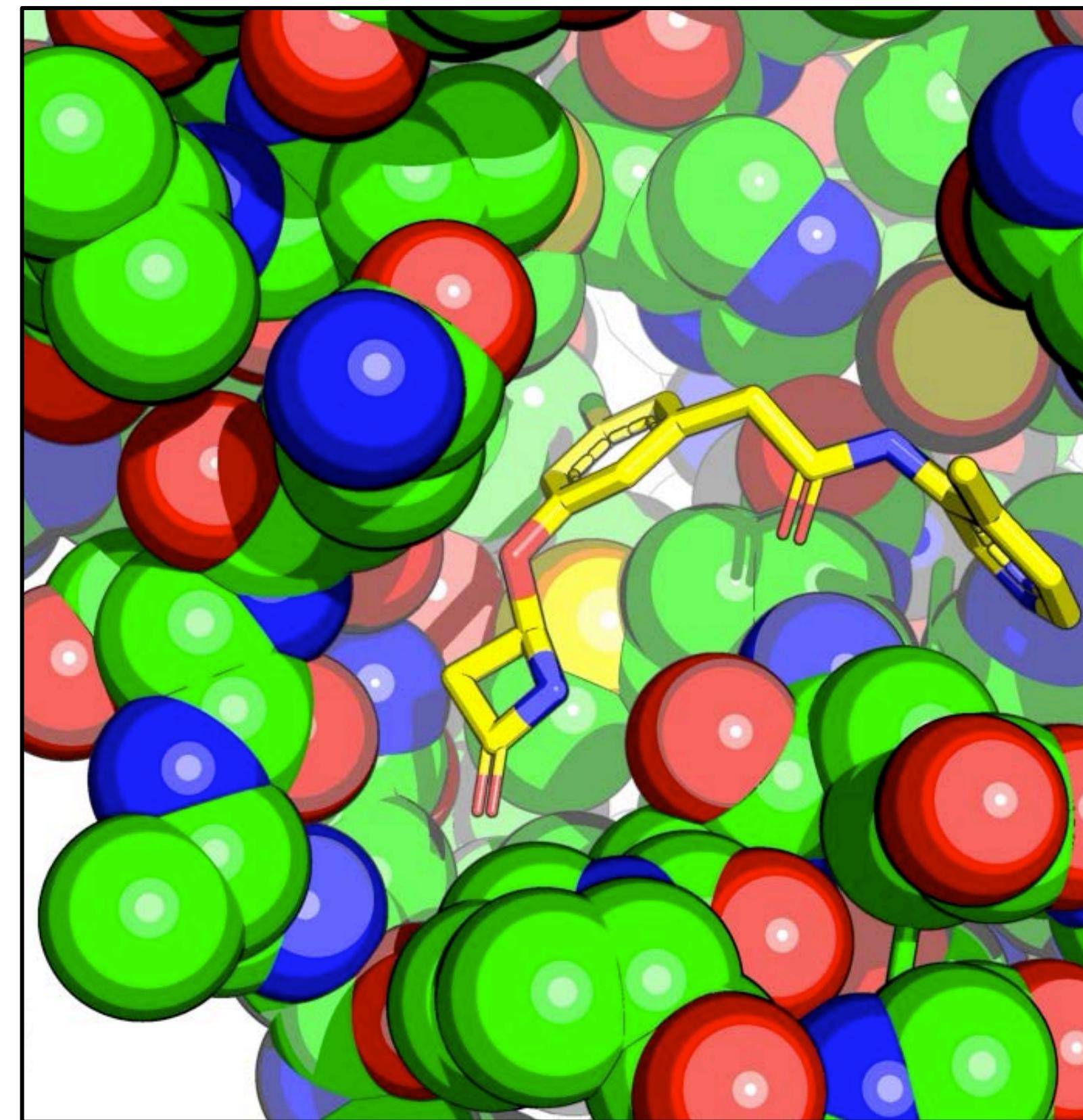
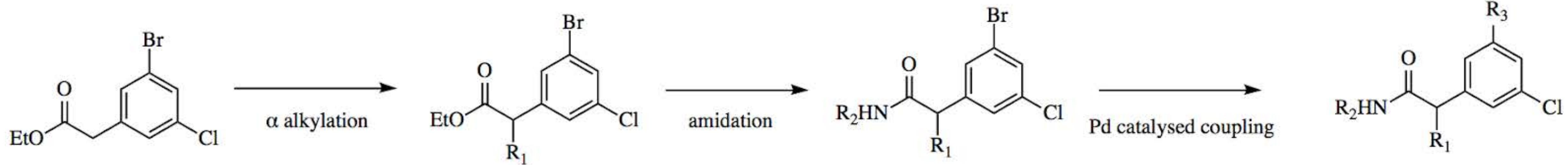


EDJ-MED-e58735b6-3



EDJ-MED-e58735b6-4

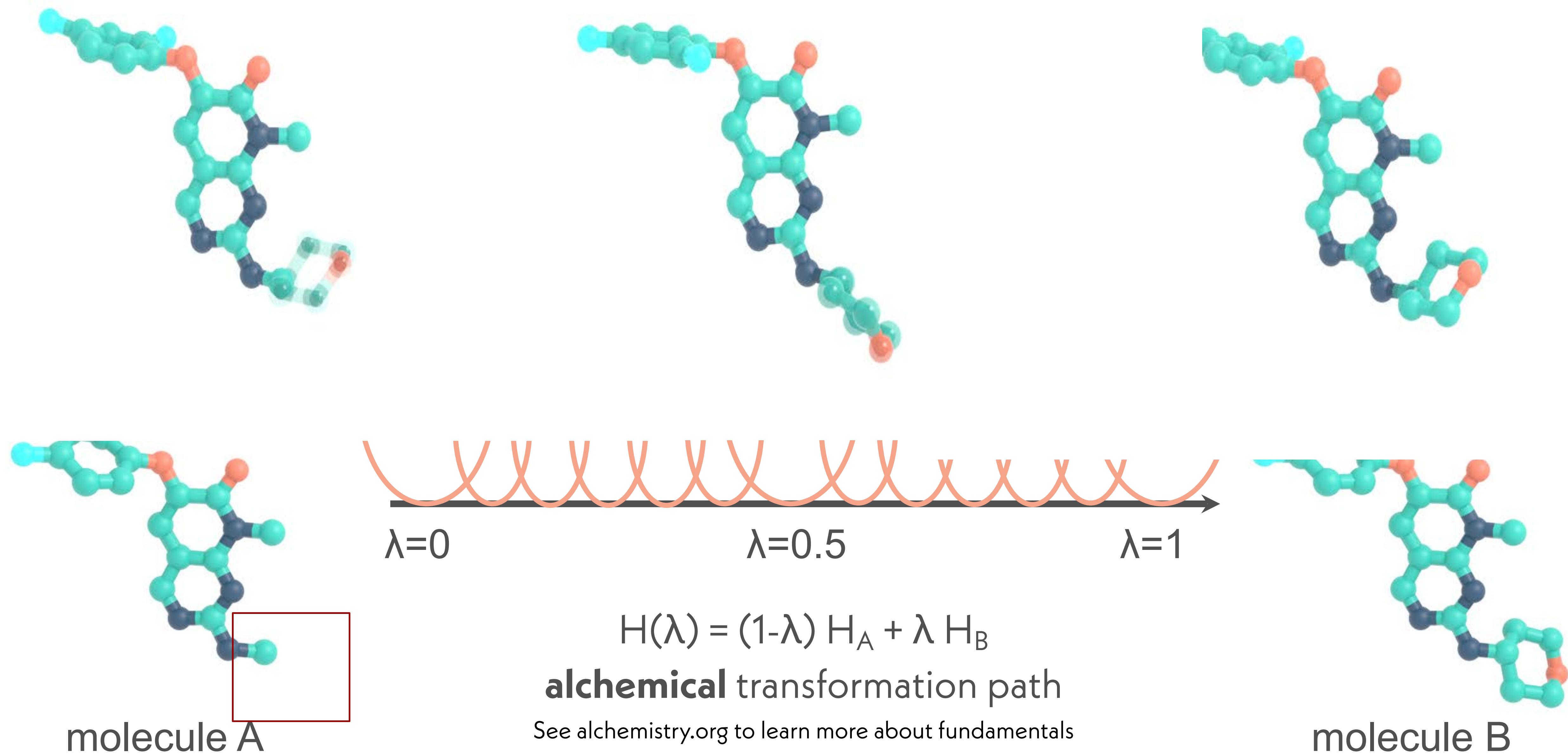
# PREDICTIVE MODELS AIM TO REDUCE THE NUMBER OF DESIGN CYCLES NEEDED TO ACHIEVE OUR DESIGN OBJECTIVES



$\sim 15,000$   
Potential  
 $\text{R}_3$  groups



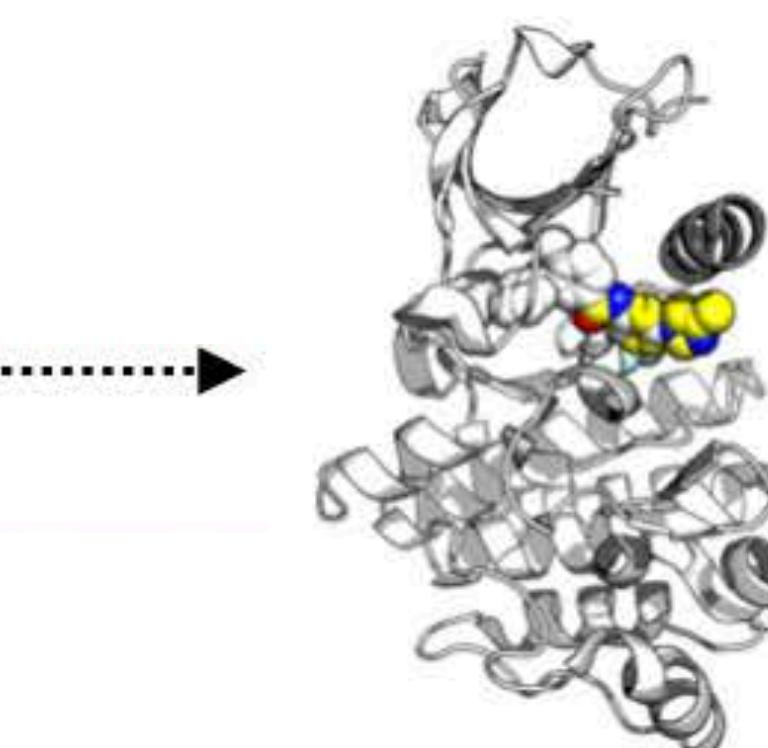
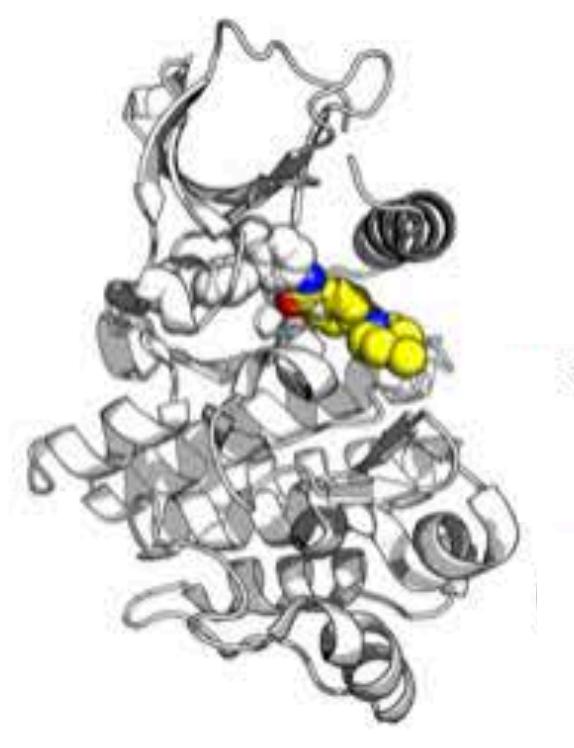
# ALCHEMICAL FREE ENERGY CALCULATIONS HAVE PROVEN TO BE A USEFUL WAY TO EXPLOIT STRUCTURAL DATA TO MAKE PREDICTIONS



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

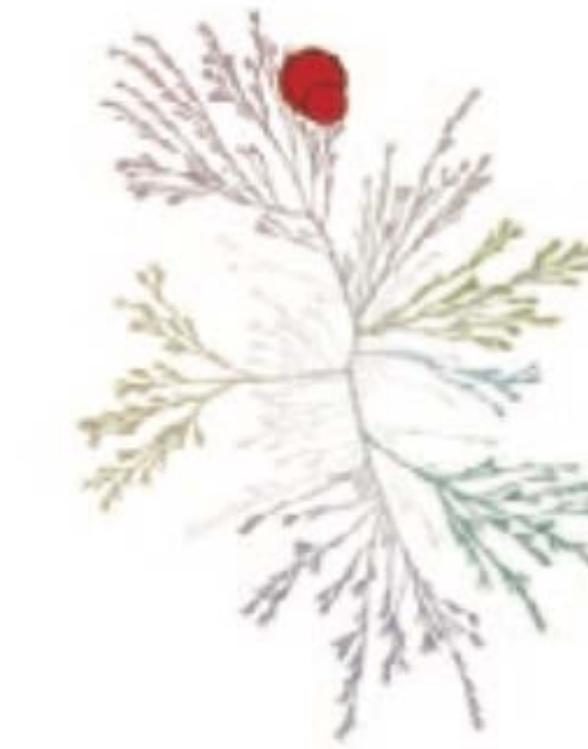
## driving affinity / potency

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



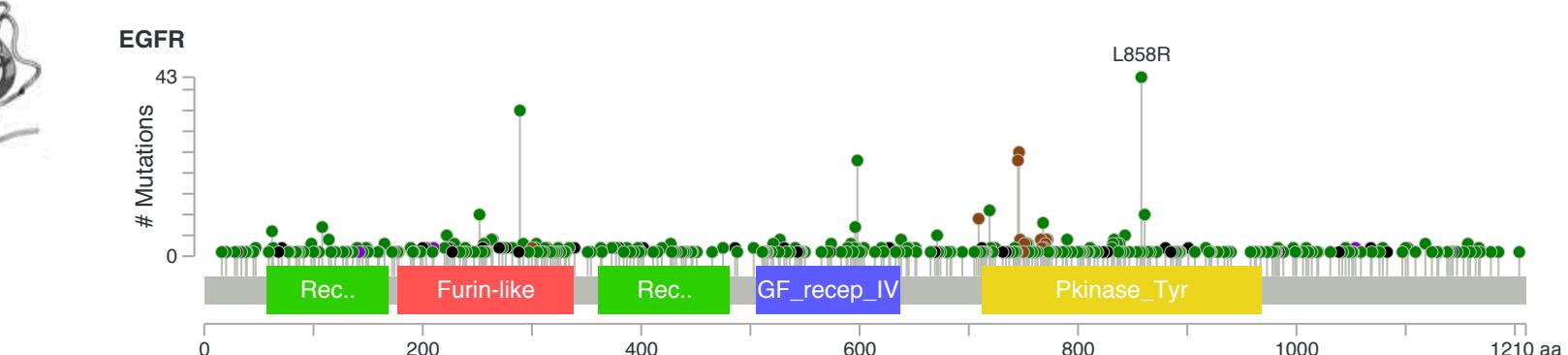
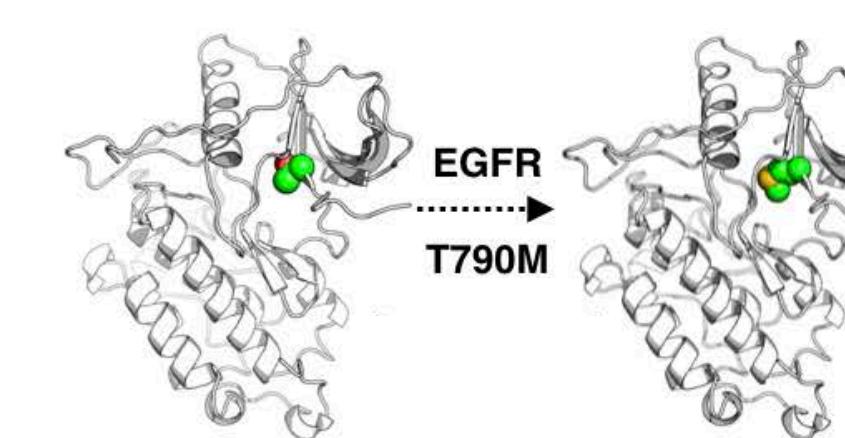
## driving selectivity

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>



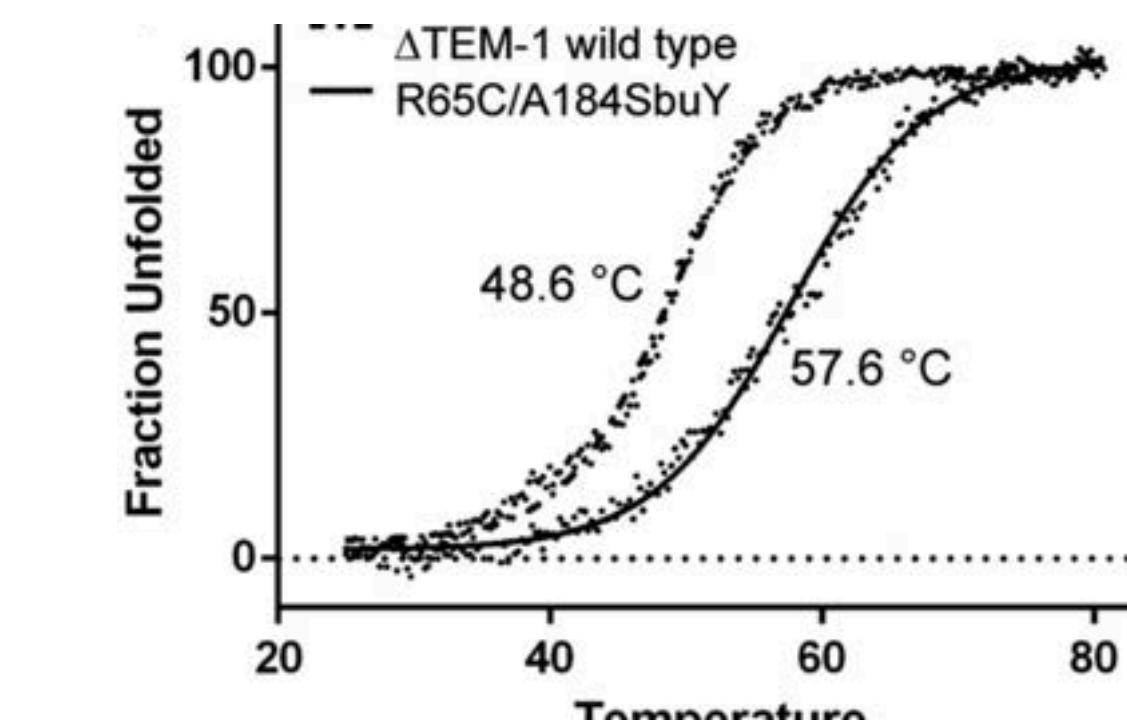
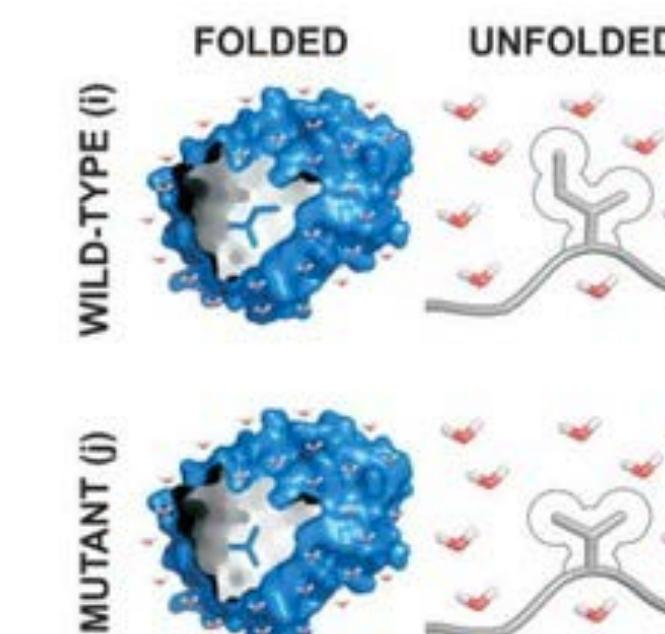
## predicting clinical drug resistance/sensitivity

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>



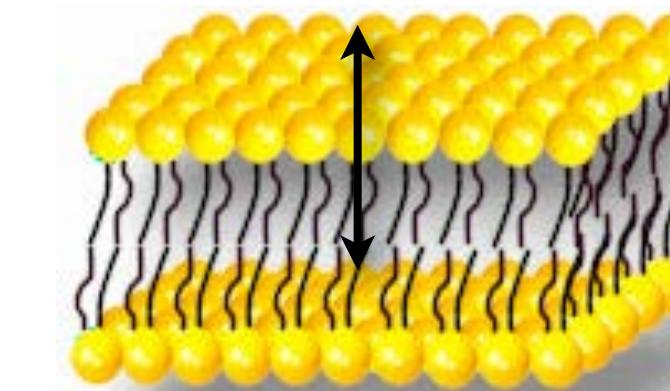
## optimizing thermostability

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

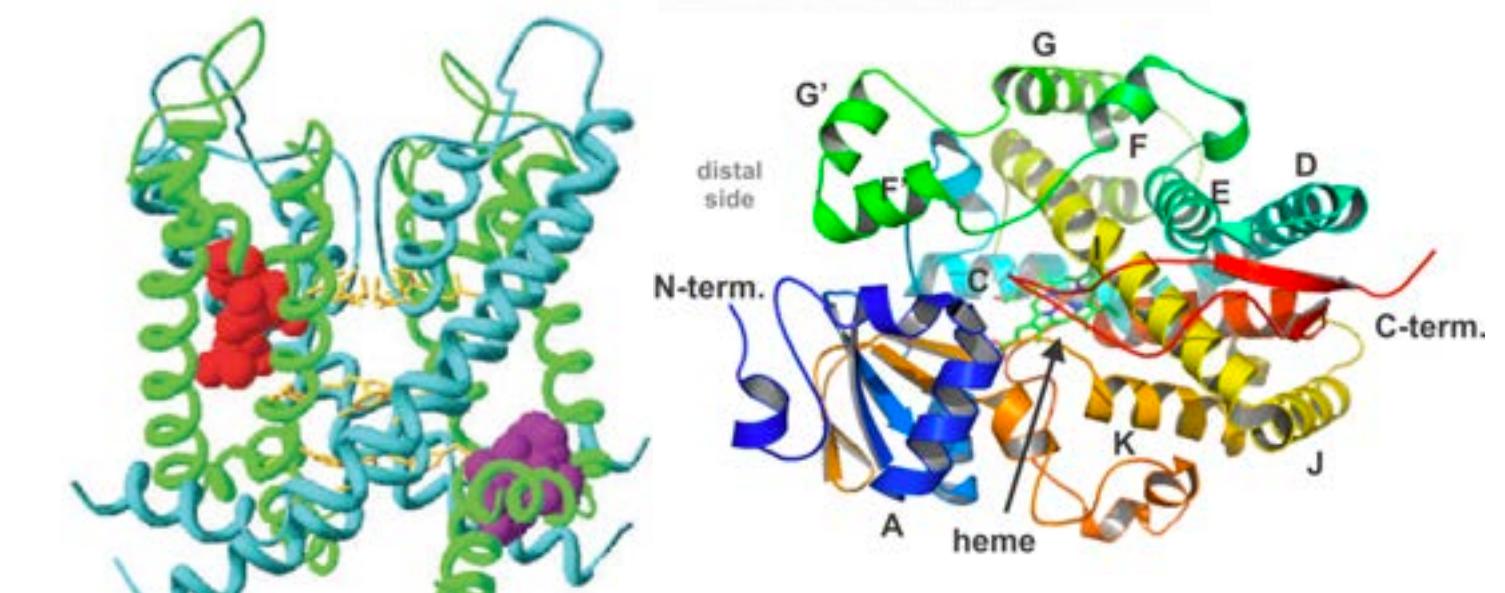


# **...AND HOLD THE POTENTIAL FOR EVEN BROADER APPLICABILITY AS MORE STRUCTURAL DATA EMERGES**

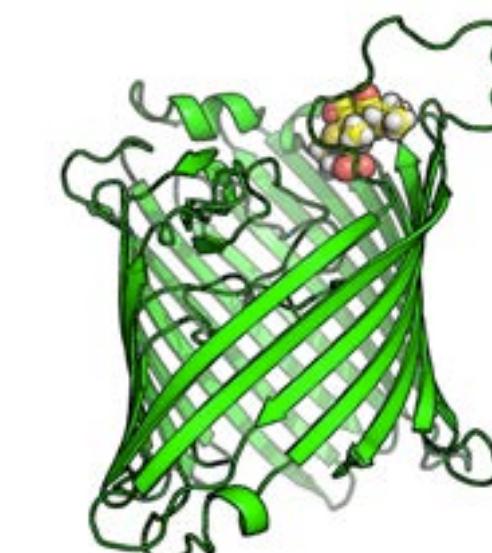
**partition coefficients ( $\log P$ ,  $\log D$ ) and permeabilities**



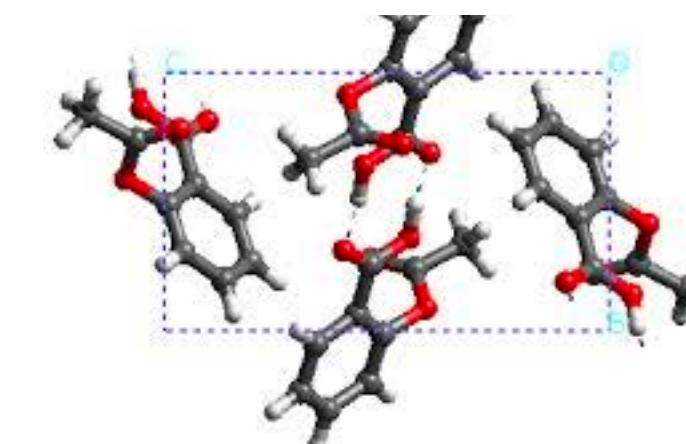
**structure-enabled ADME/Tox targets**



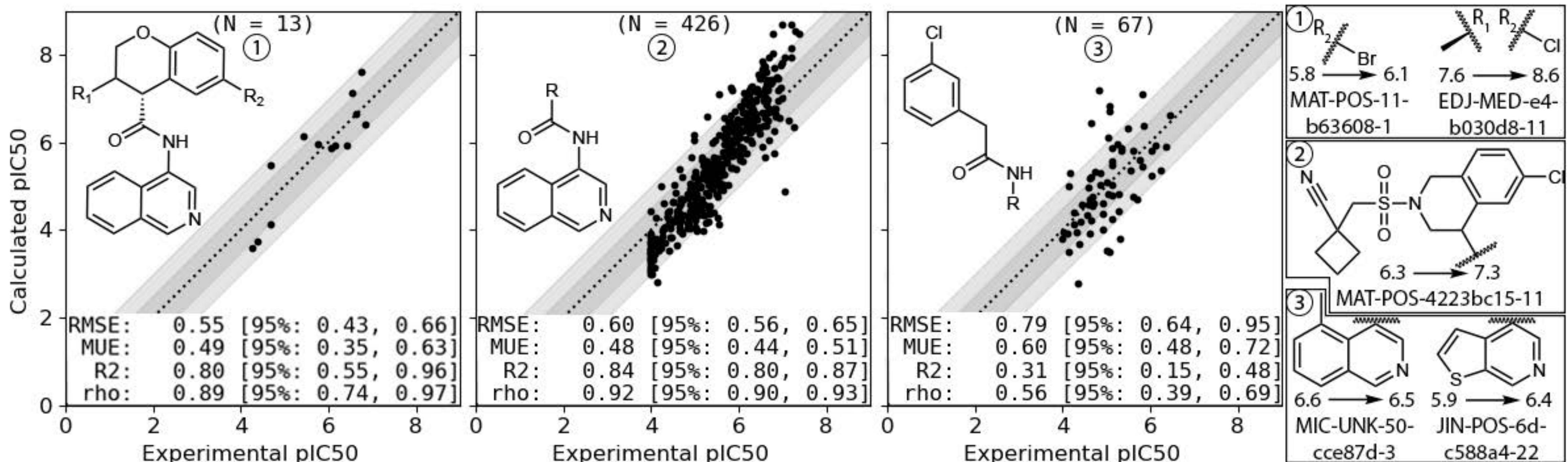
**porin permeation**



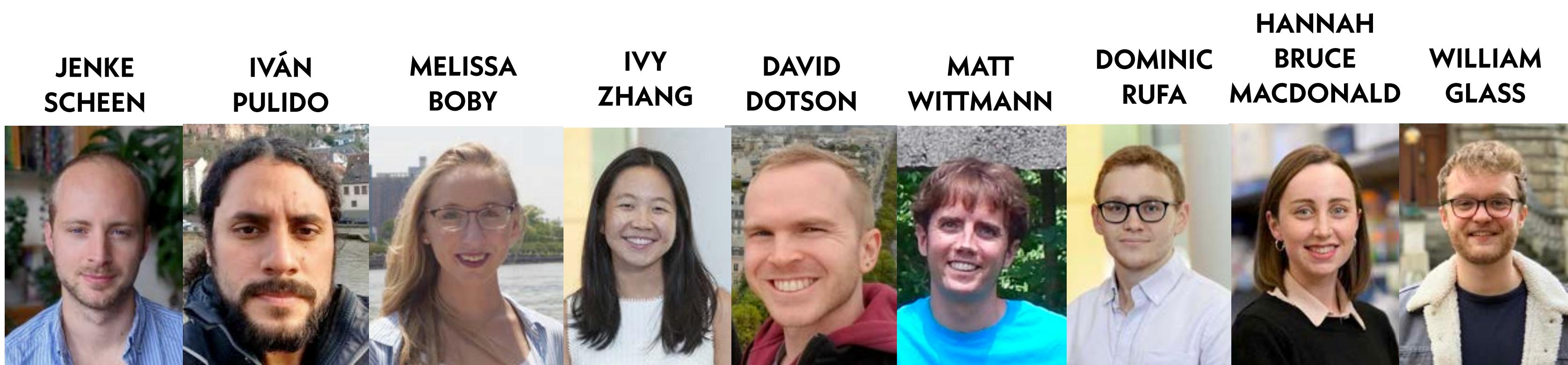
**crystal polymorphs, etc.**



# 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



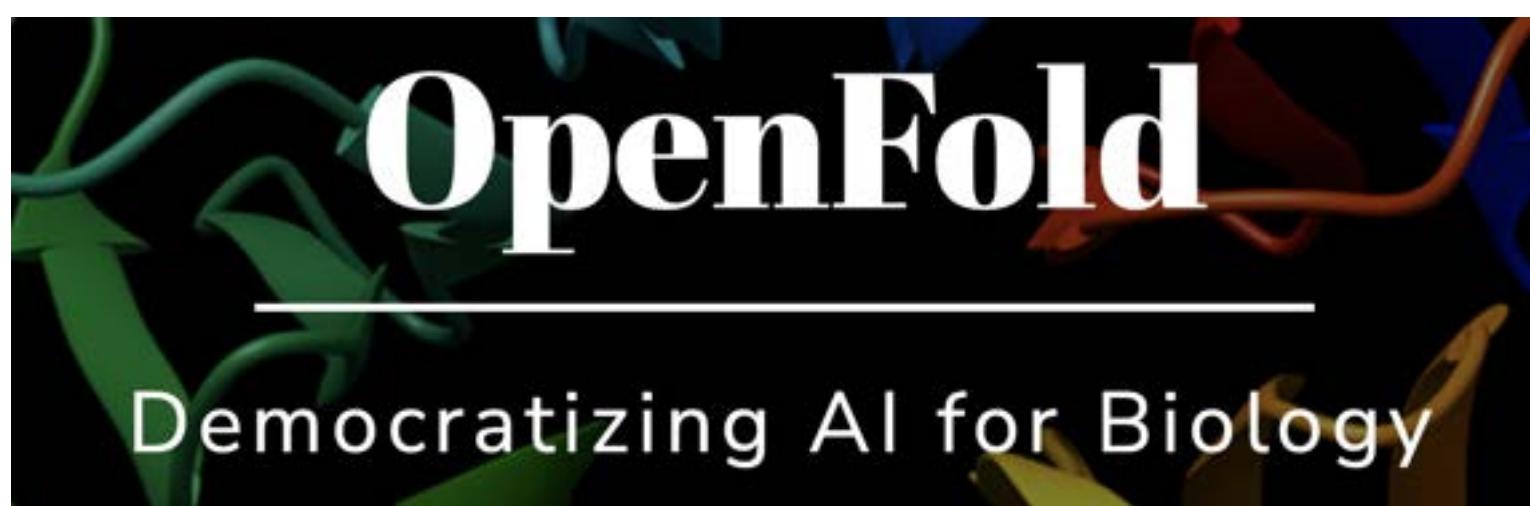
<http://openforcefield.org>



<http://openfree.energy>



<https://westpa.github.io>



<http://openfold.io>



<http://rosettacommons.org>

OpenADMET

<http://openadmet.org>



## 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

# 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

## Best Practices Publications



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

## Technical Advisory Committee



14 academic members

## Industry Partner Members



18 pharma companies

<https://openfree.energy>

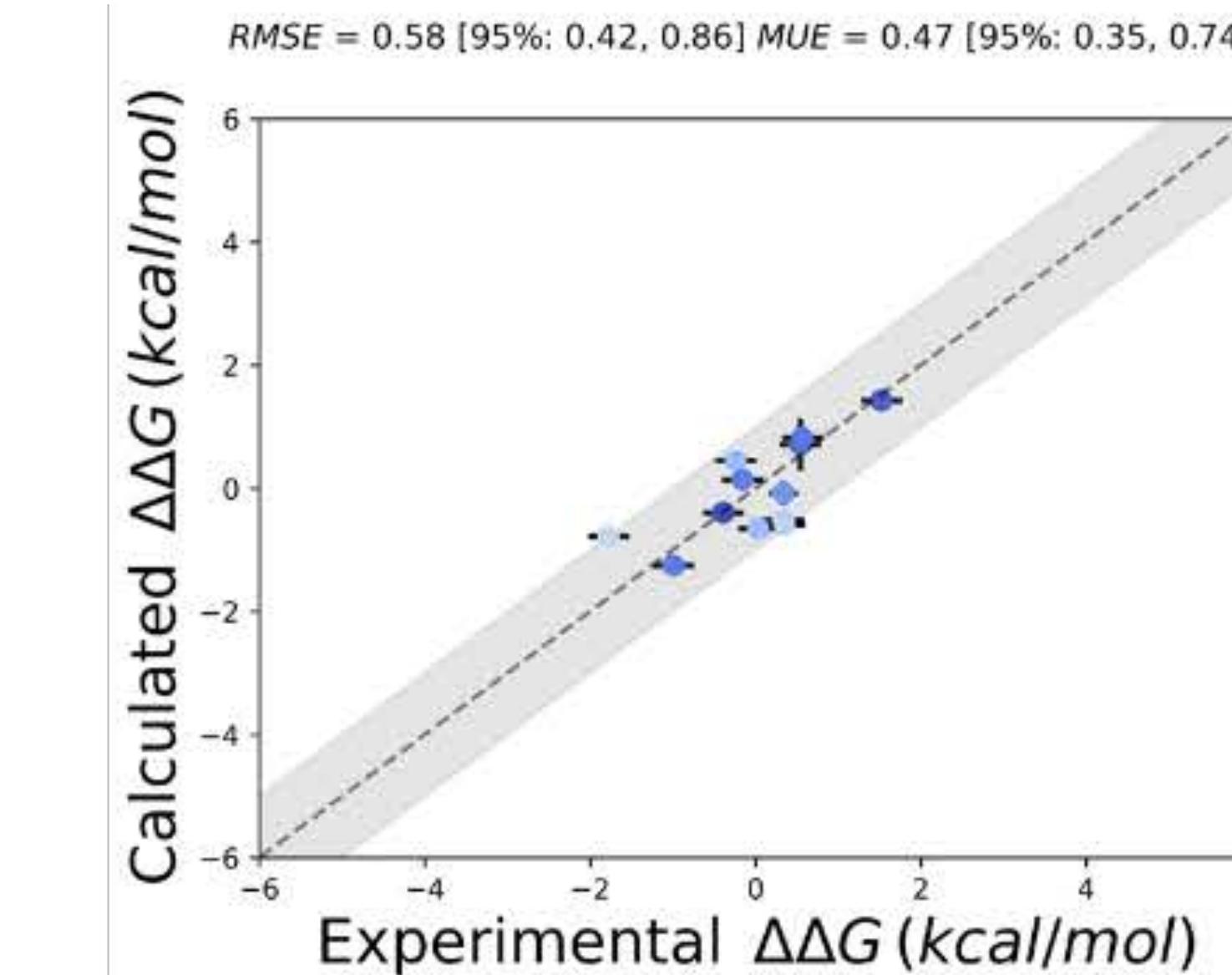
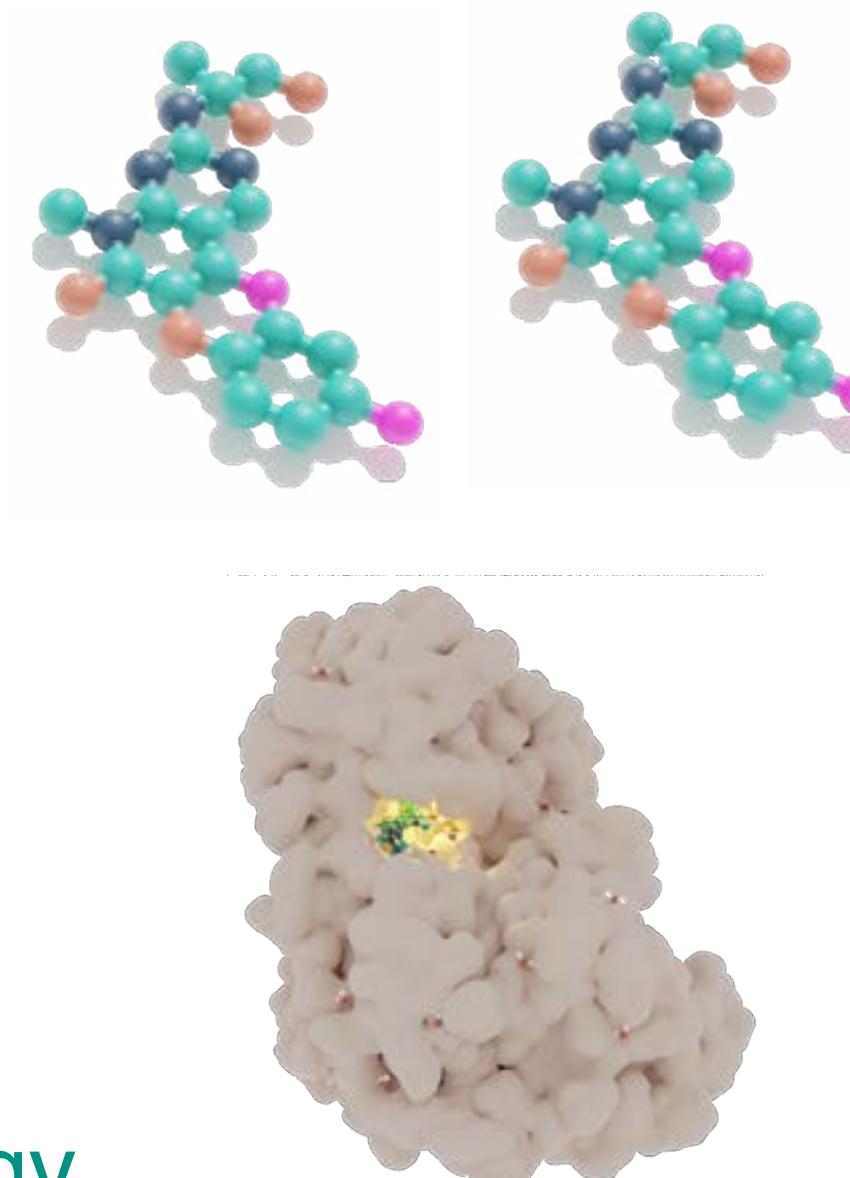
# Open Free Energy provides open source tools for free energy calculations you can start using from your browser



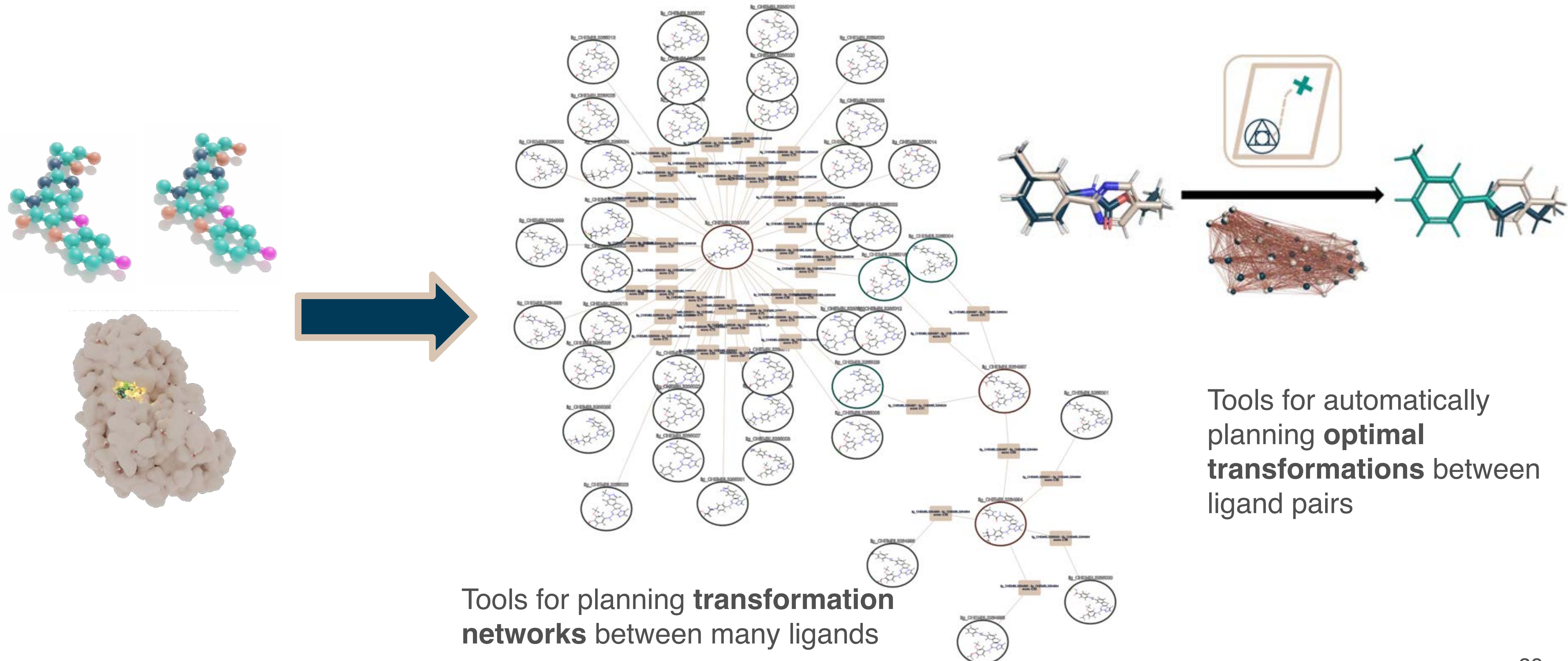
ABOUT PROJECTS OUR TEAM NEWS GETTING INVOLVED CAREERS

The Open Free Energy initiative is dedicated to the development of **open-source** tools for **binding free energy** calculations to guide pharmaceutical **drug design and discovery**.

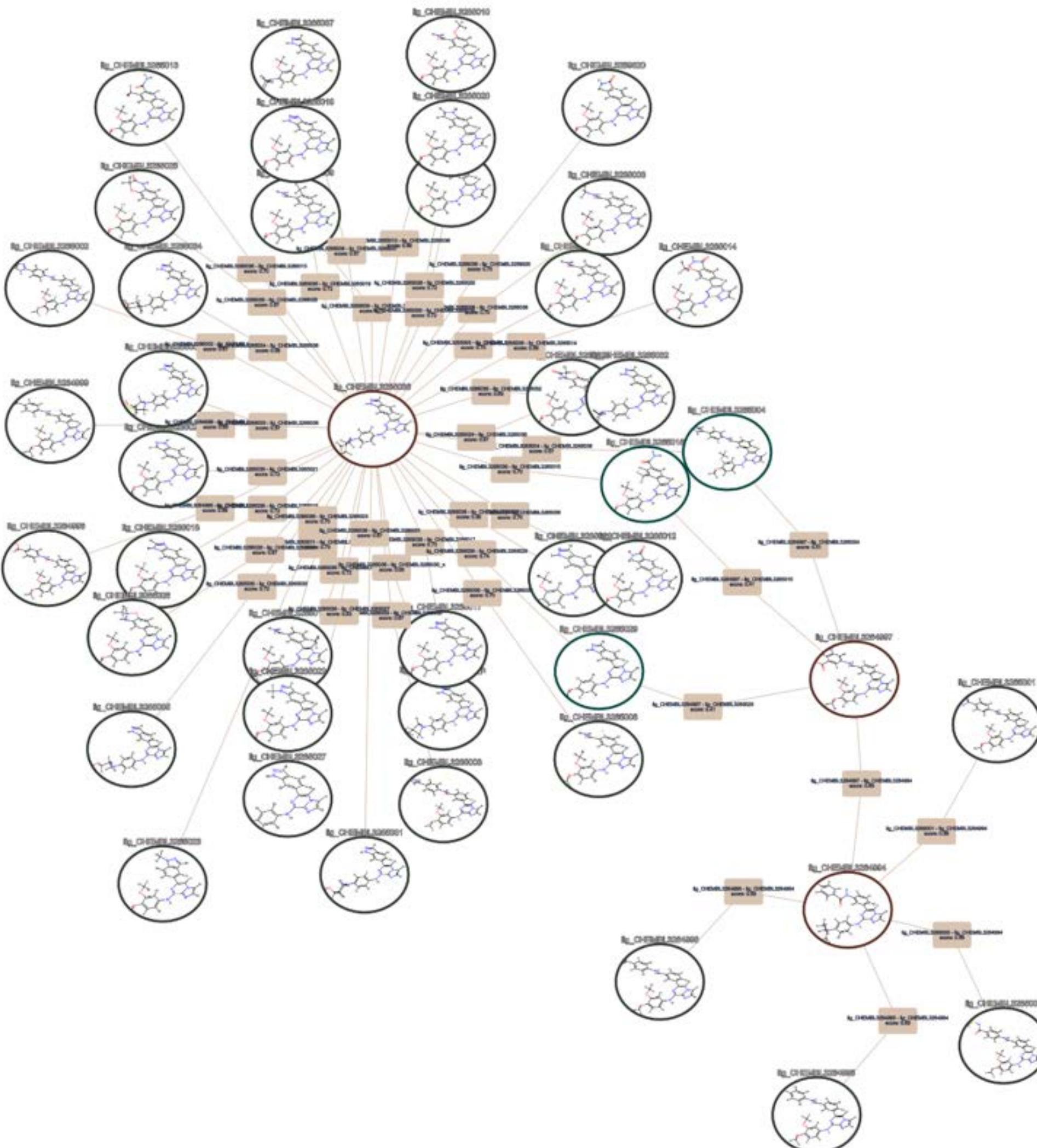
Try It Online



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

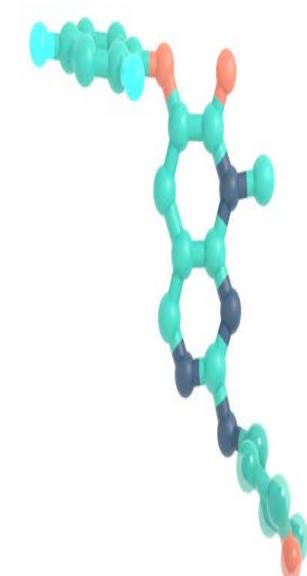


# ... workflows for running alchemical transformations on CPUs or GPUs



<http://openmm.org>

open source GPU-accelerated  
biomolecular simulations



run each alchemical  
transformation  
separately

~ 12 GPU-hours per  
transformation

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

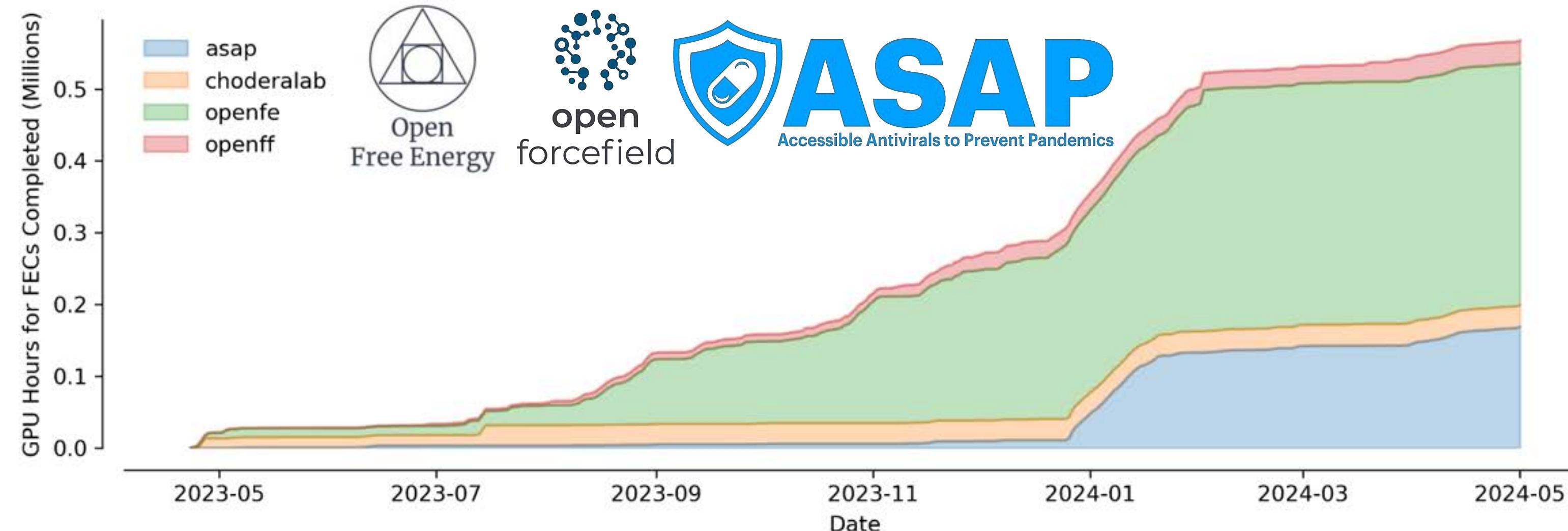
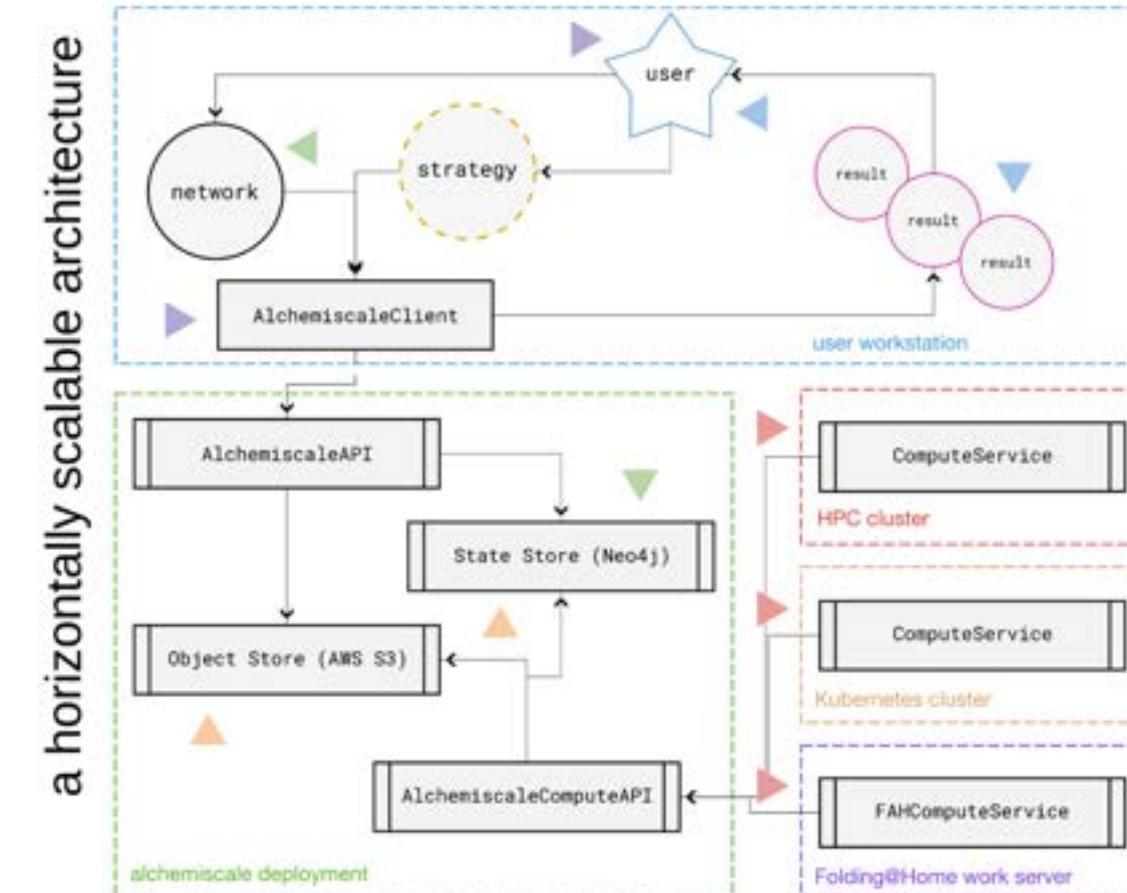
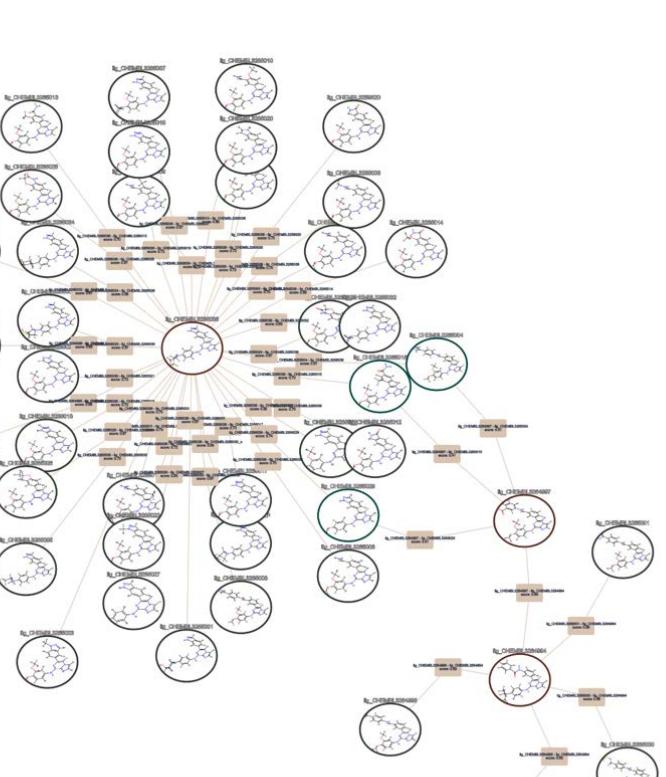


**Alchemiscale** has enabled >0.5 million GPU hours of multi-cluster distributed OpenFE usage in the last year



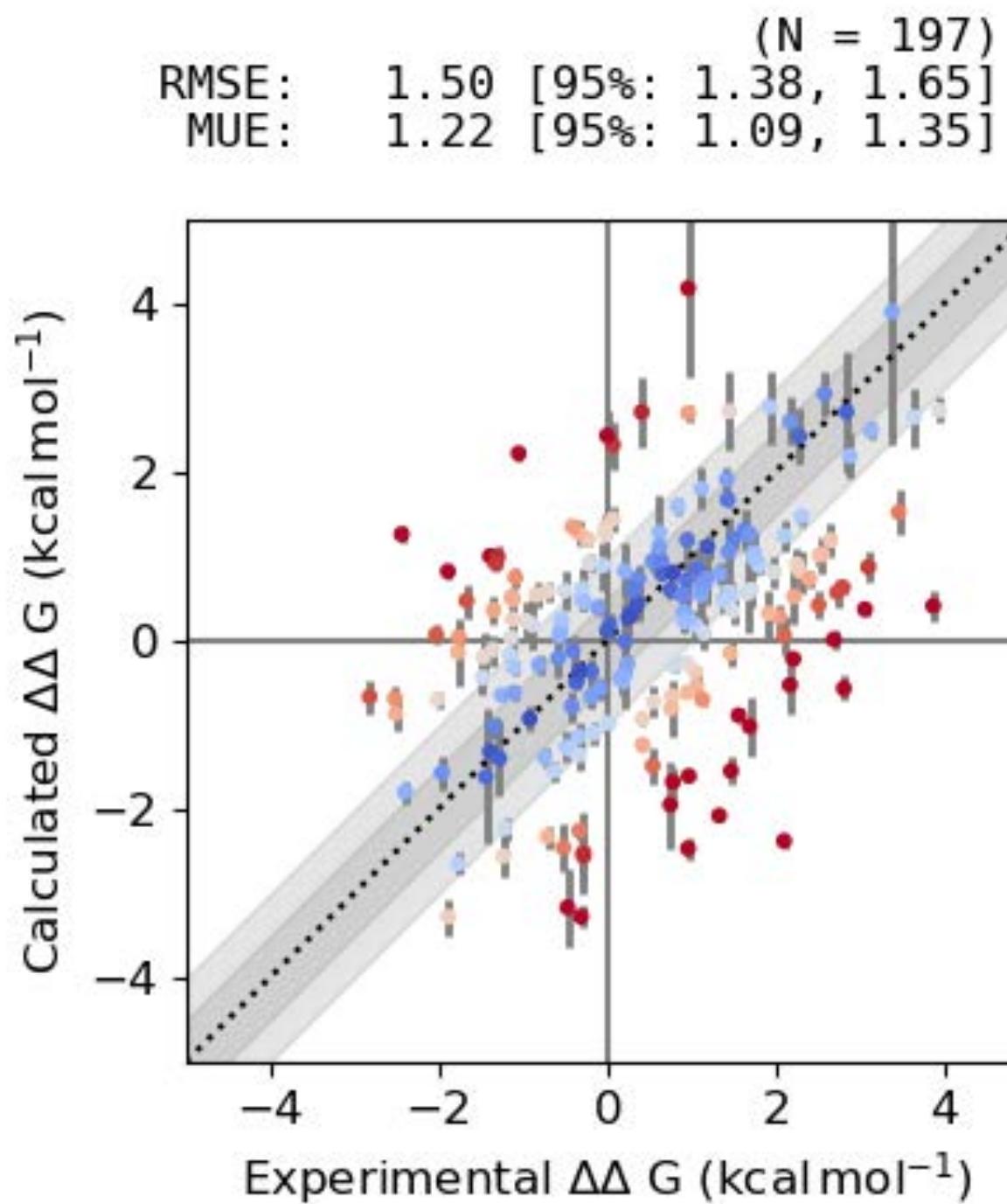
David Dotson

Ian Kenney

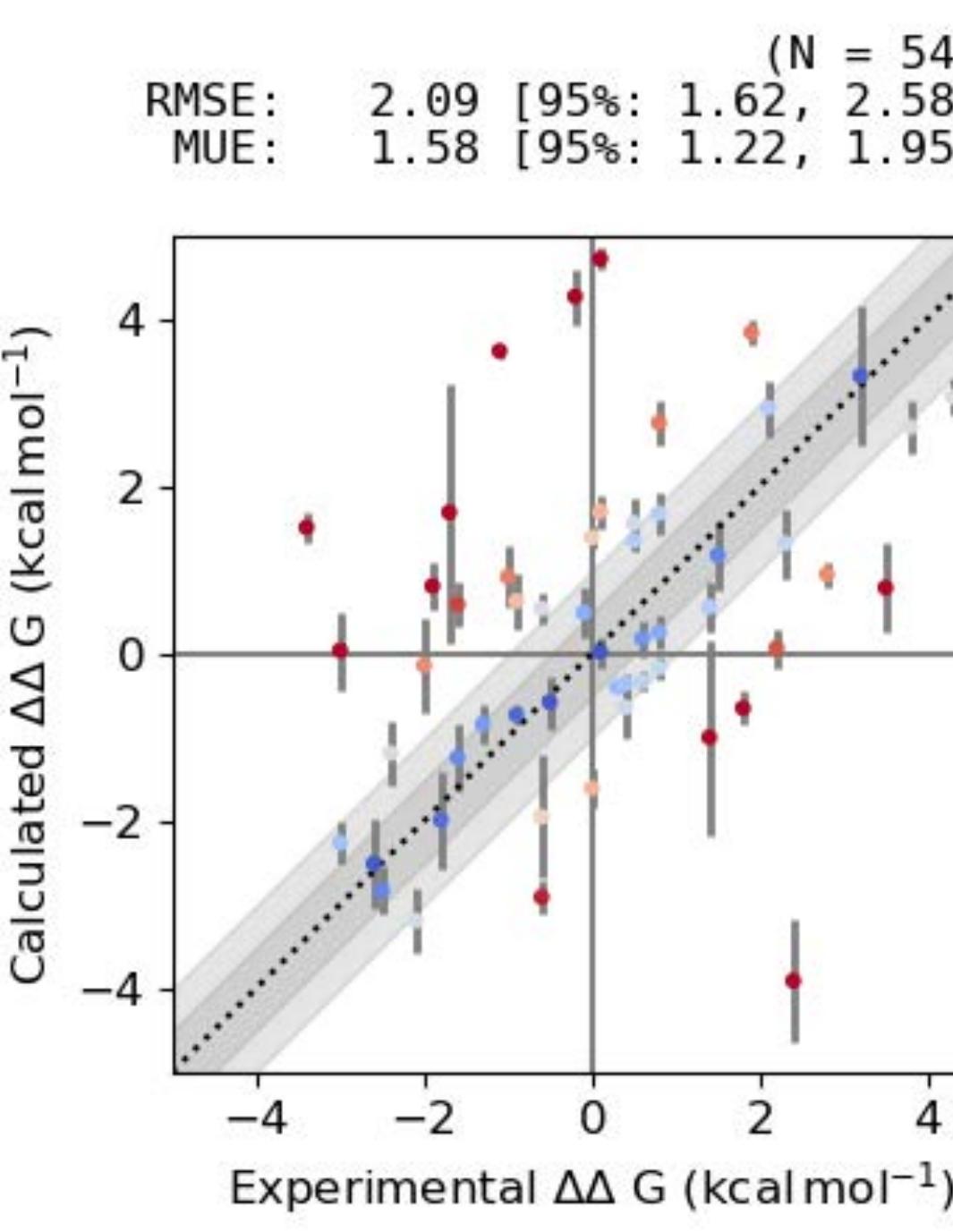


# Open Free Energy tools are showing utility in real industry and academic drug discovery campaigns

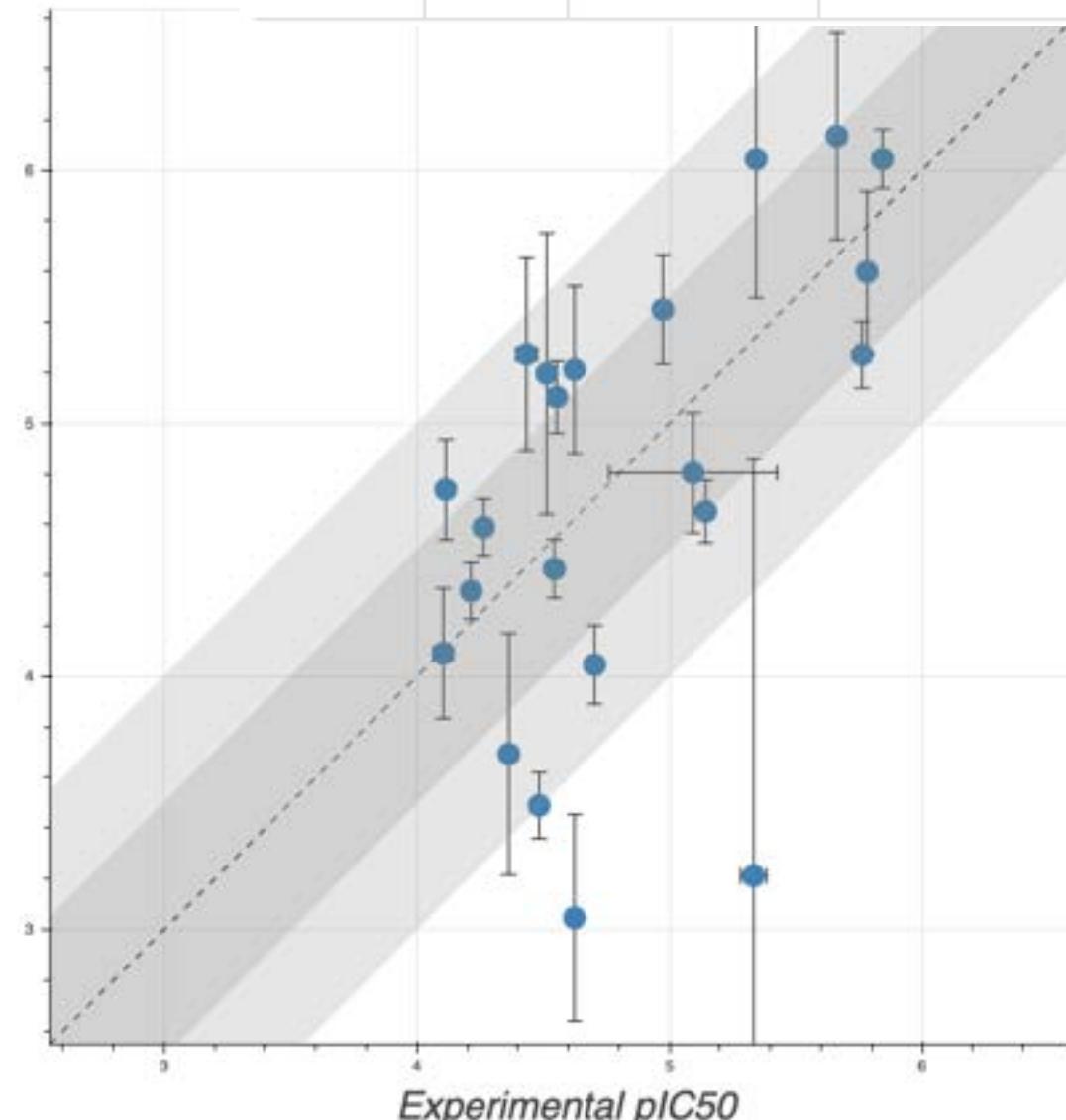
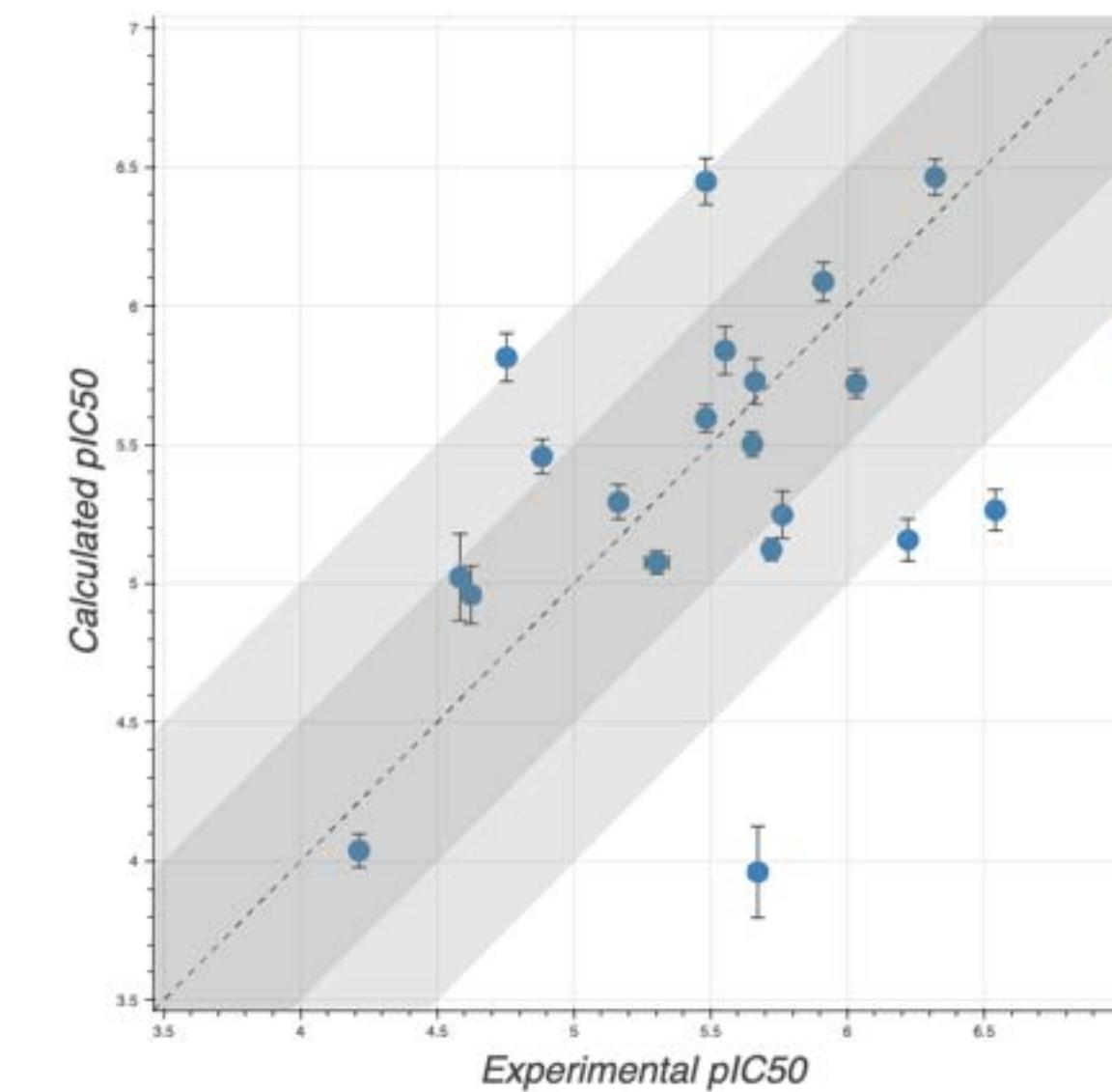
Merck KGaA Inhouse Project G



Merck KGaA Inhouse Project T



## Antiviral drug discovery with the ASAP AViDD Center



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

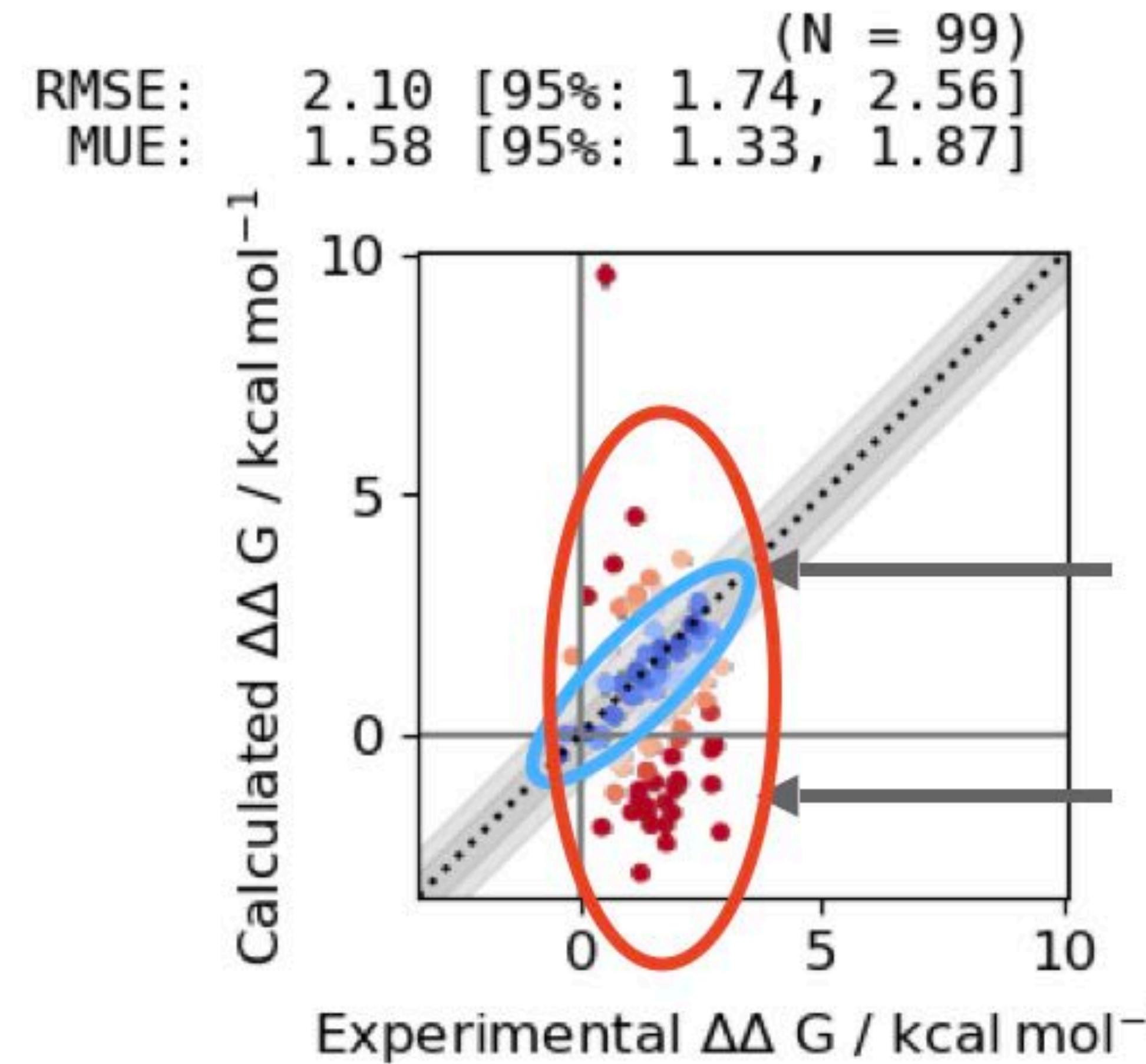
Christina Schindler

<http://openfree.energy>

SARS-CoV-2 Mpro

<http://asapdiscovery.org>

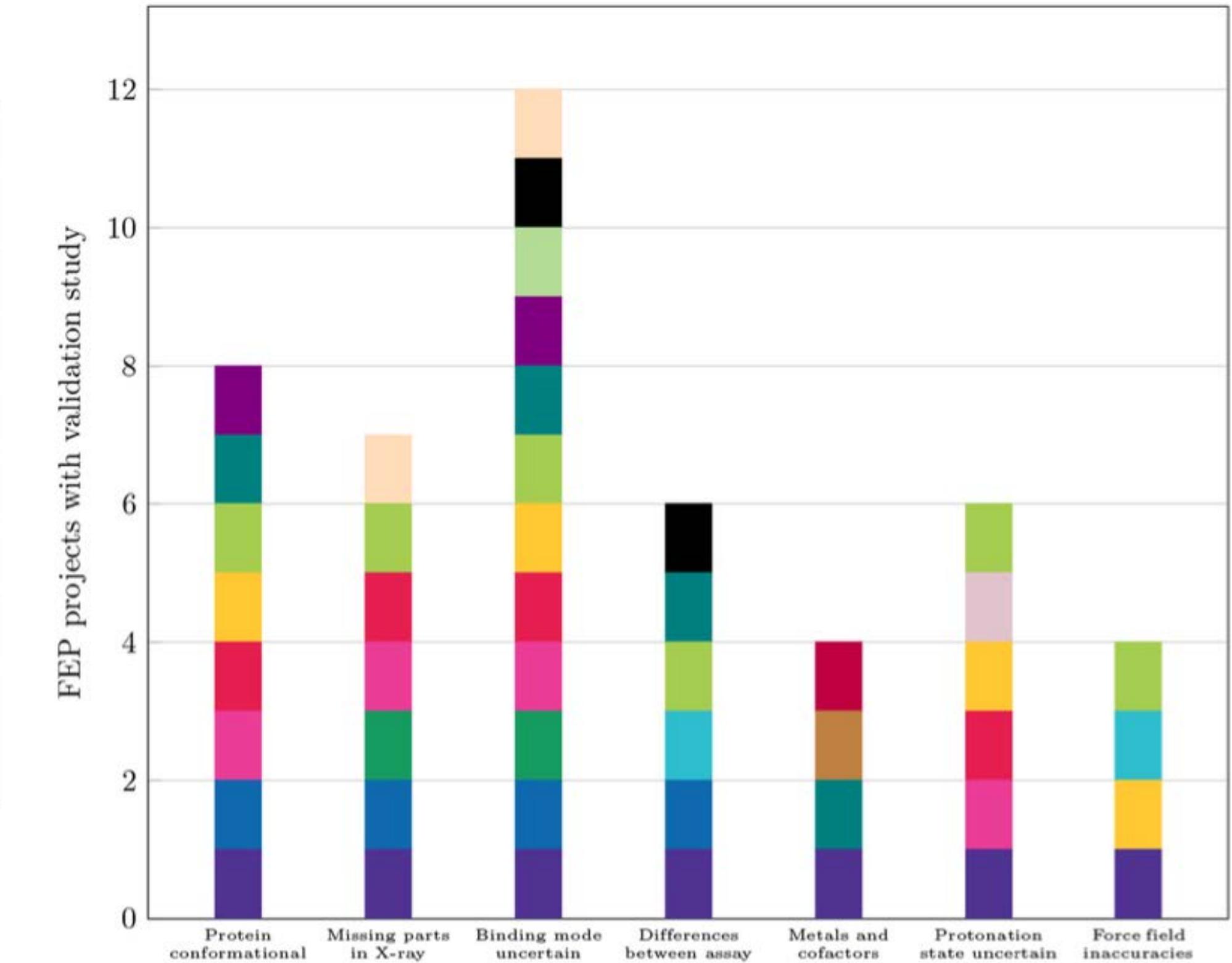
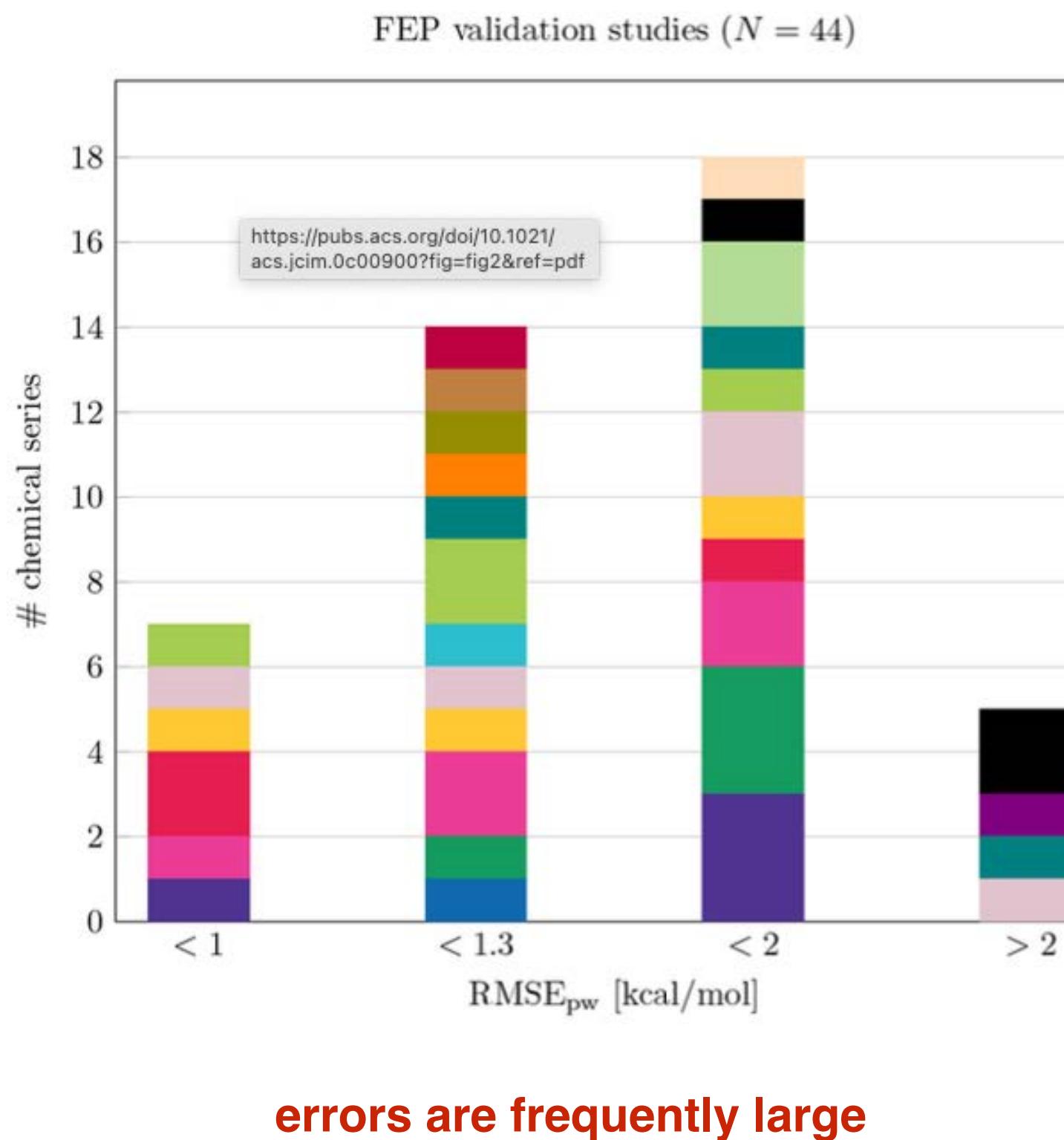
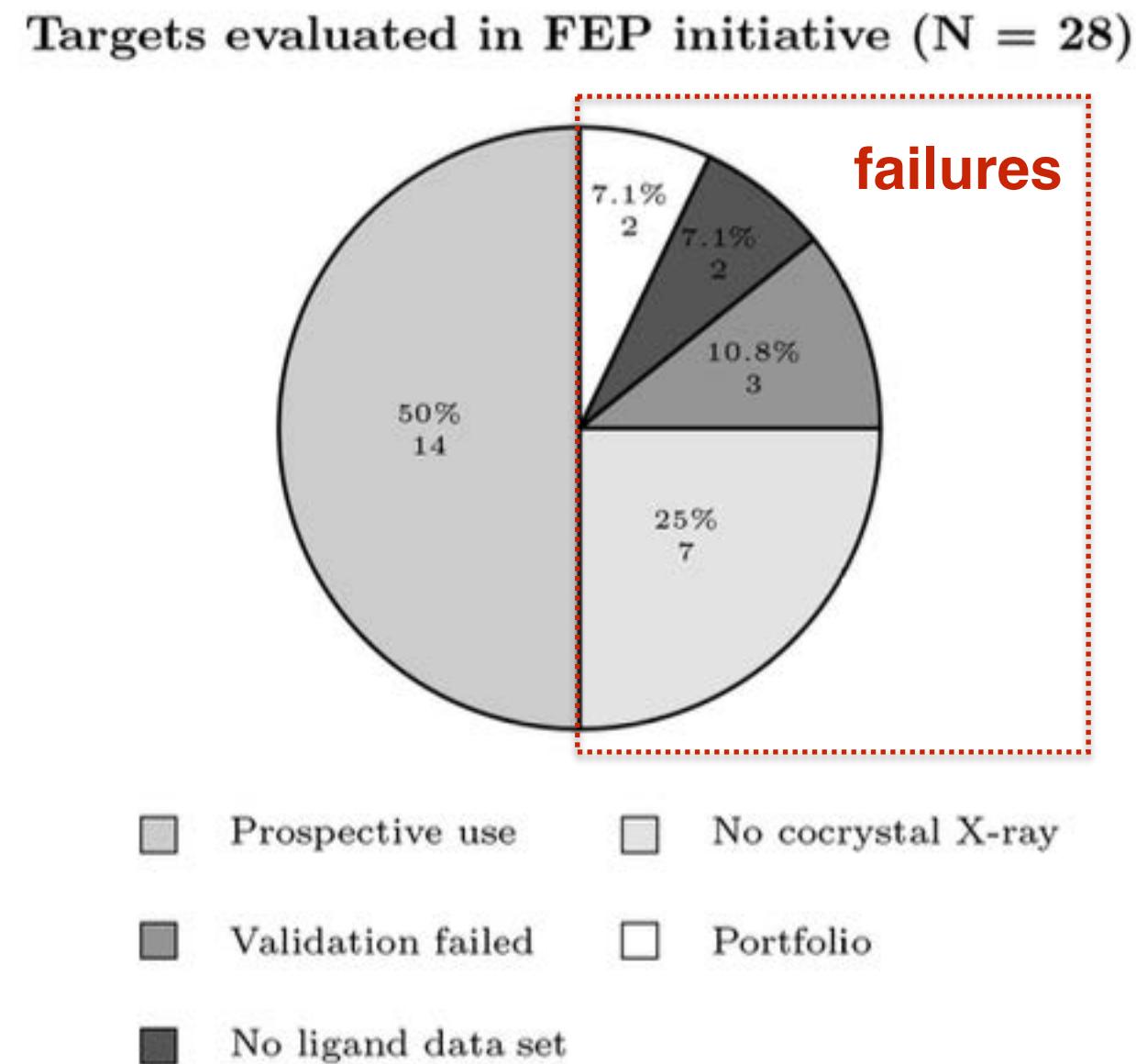
# HOWEVER, MORE OFTEN THAN WE WOULD LIKE, ACCURACY CAN BE POOR



this is good :)

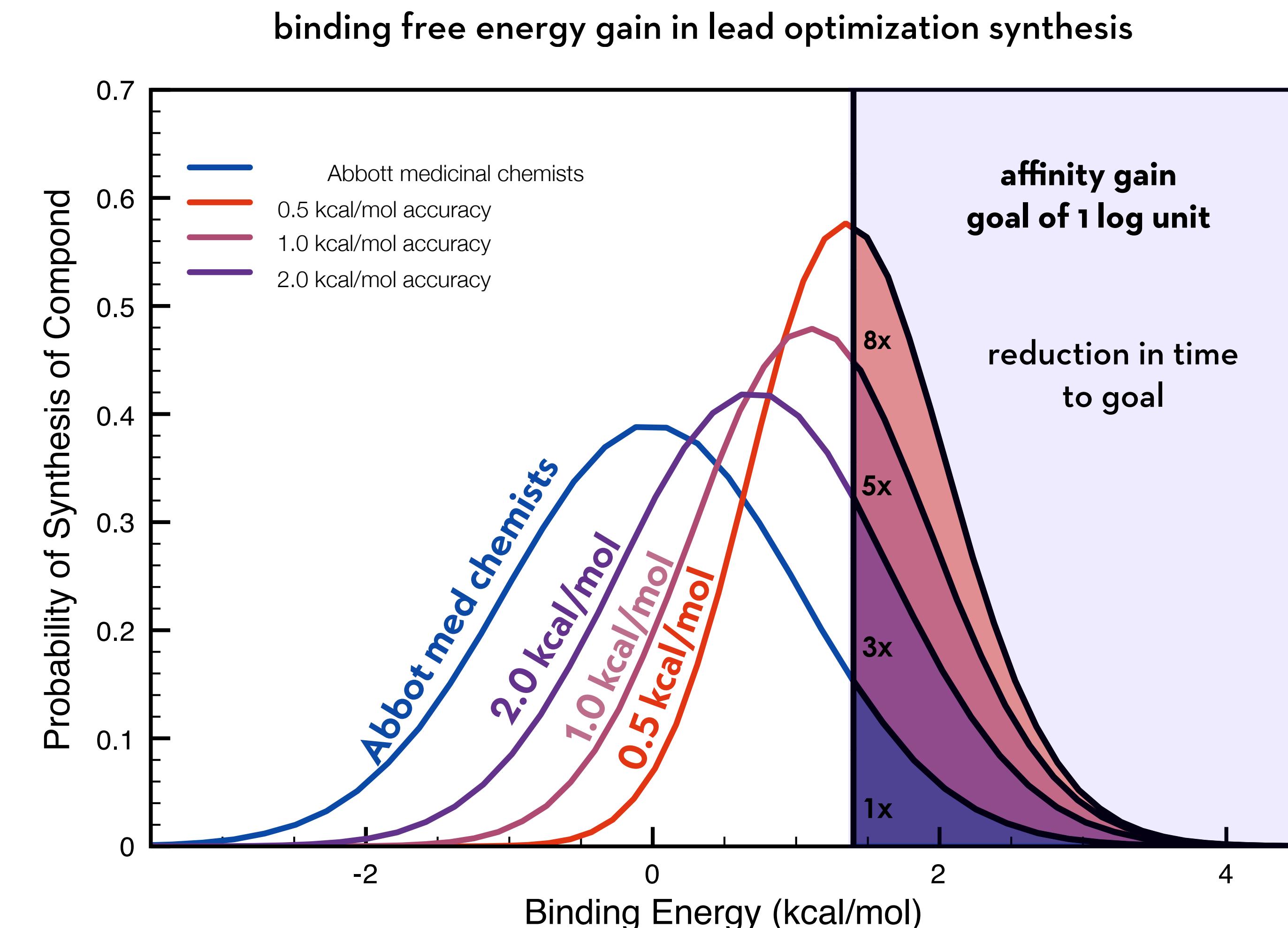
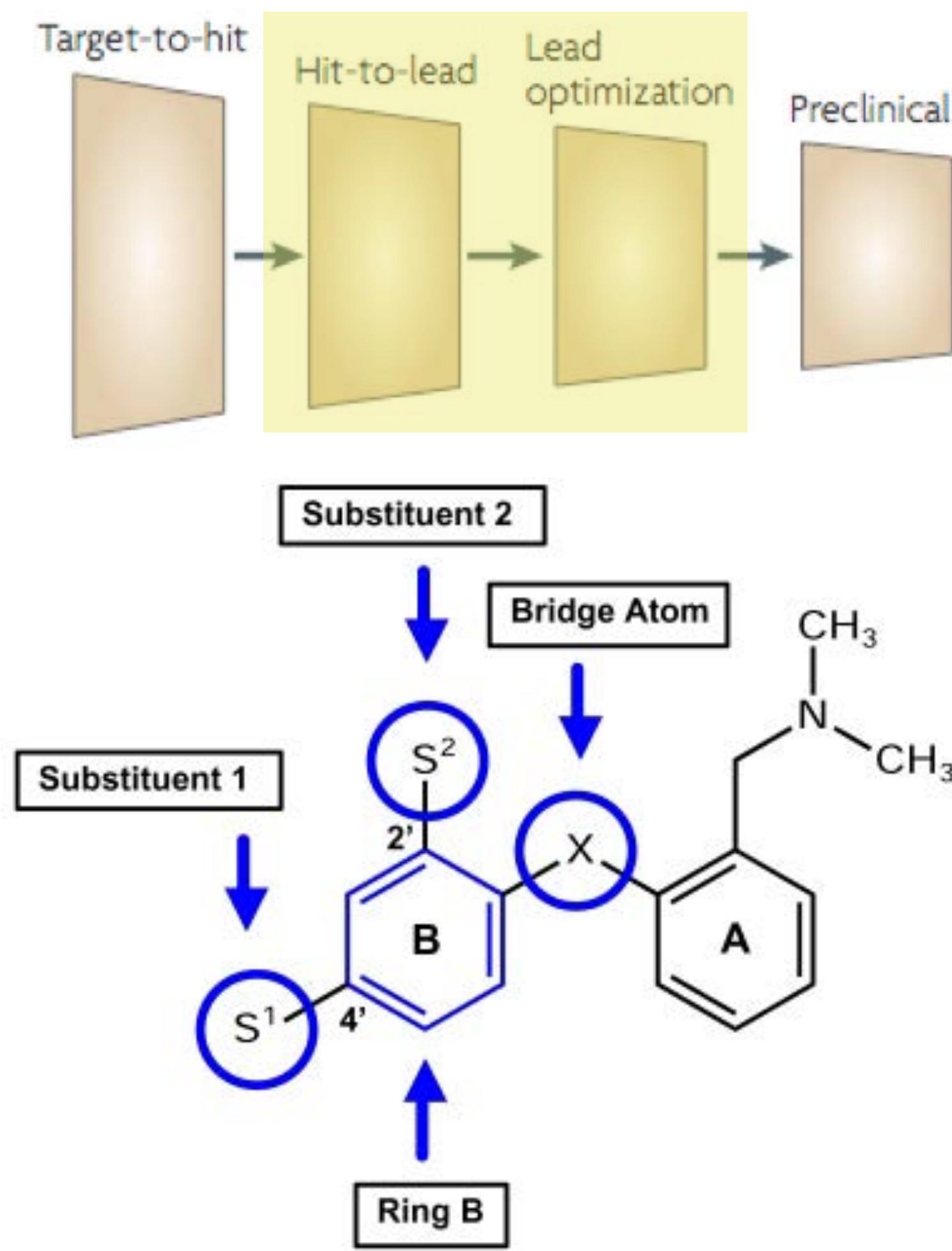
this is bad :(

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



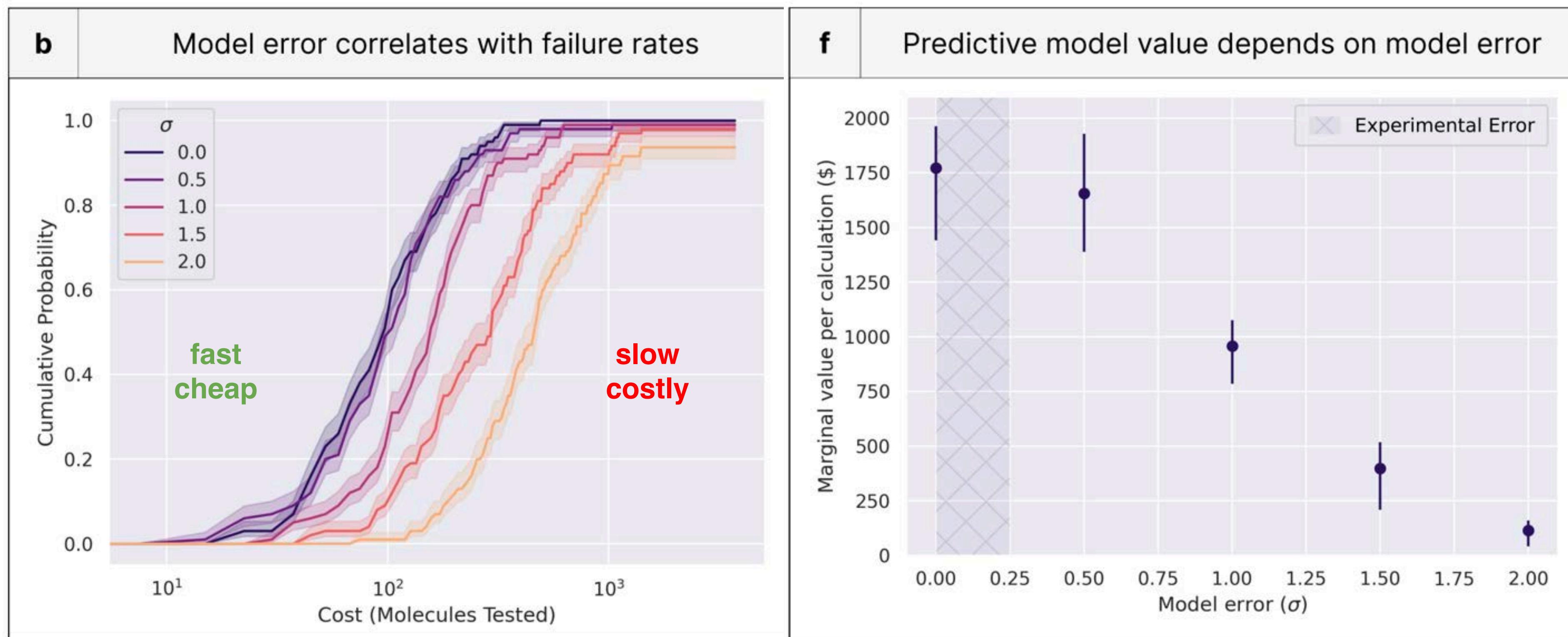
current calculations face a multitude of challenges

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



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

Modeling hit-to-lead phase of drug discovery



**WHAT'S HOLDING US BACK?**



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

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

765

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

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

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

$E_{\text{total}} =$

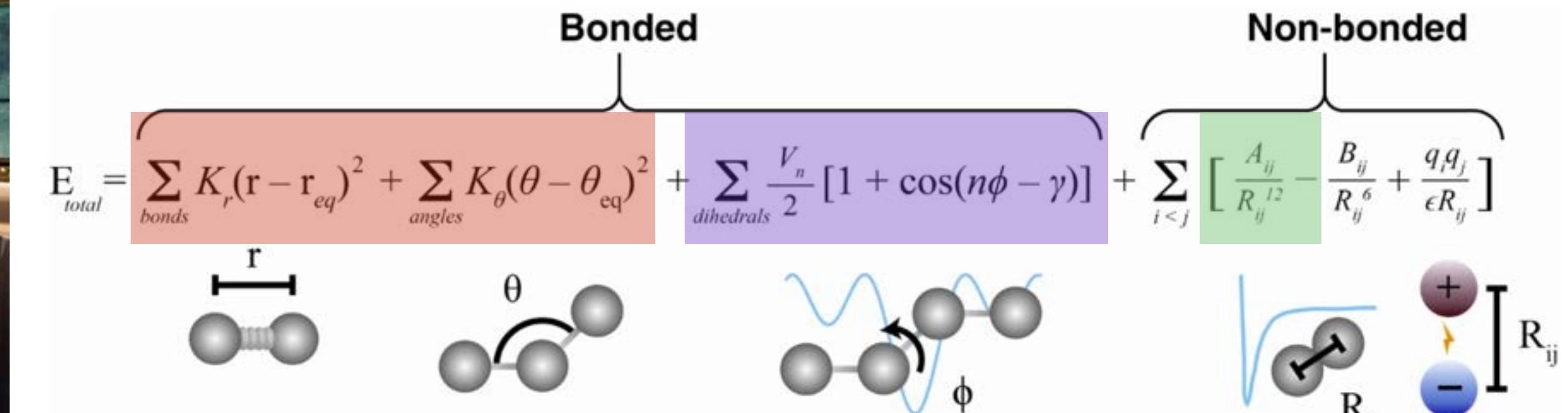
$$\sum_{\text{bonds}} K_r (r - r_{\text{eq}})^2 + \sum_{\text{angles}} K_\theta (\theta - \theta_{\text{eq}})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2} [1 + \cos (n\phi - \gamma)] + \sum_{i < j} \left[ \frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} + \frac{q_i q_j}{\epsilon R_{ij}} \right] + \sum_{\text{H-bonds}}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

# 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)**



**horrible Taylor series  
truncated at lowest order**

**crappy Fourier series  
truncated at n=6**

**don't even get me  
started on this one**

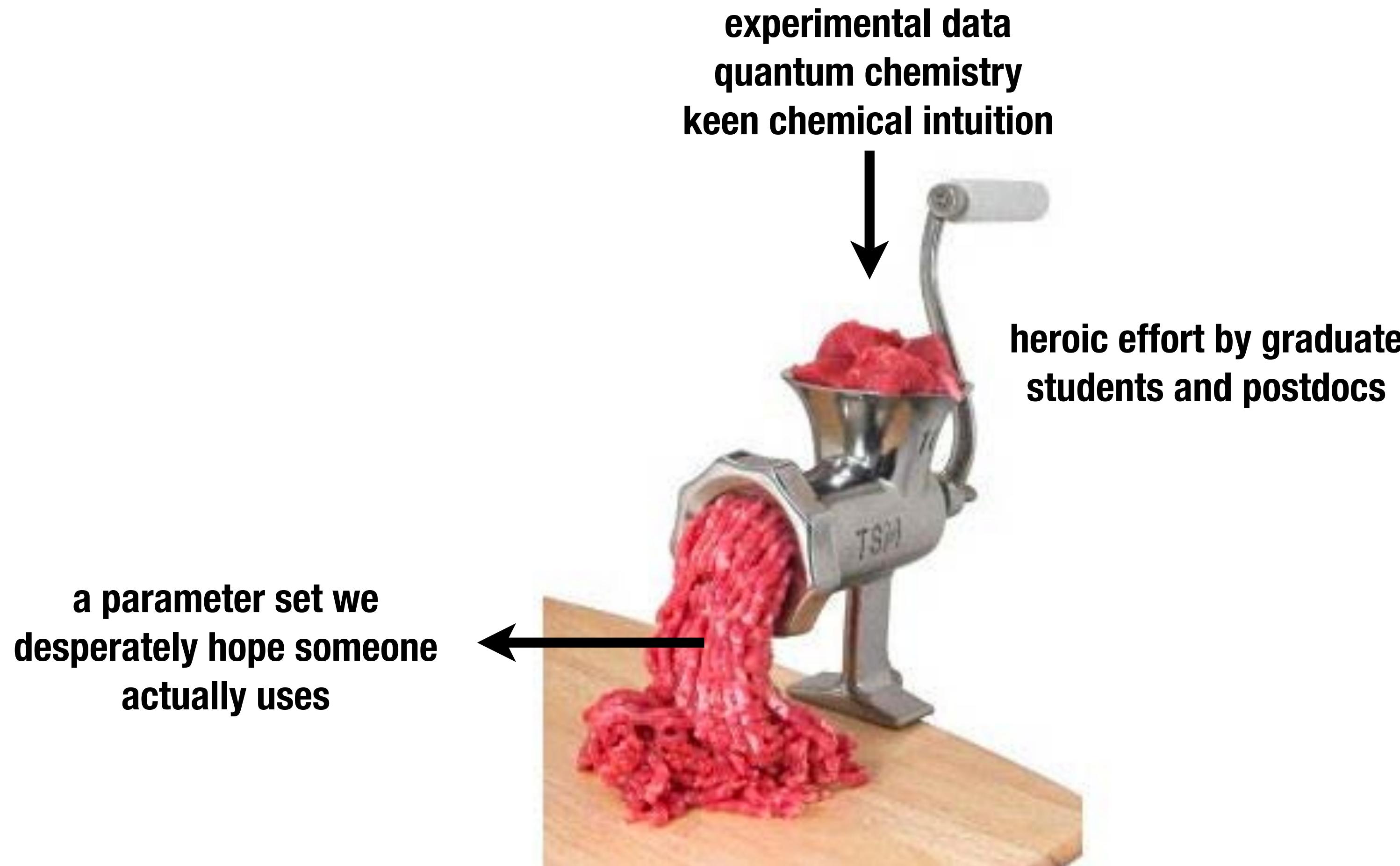
We have **real** computers now.



Why not put them to work?

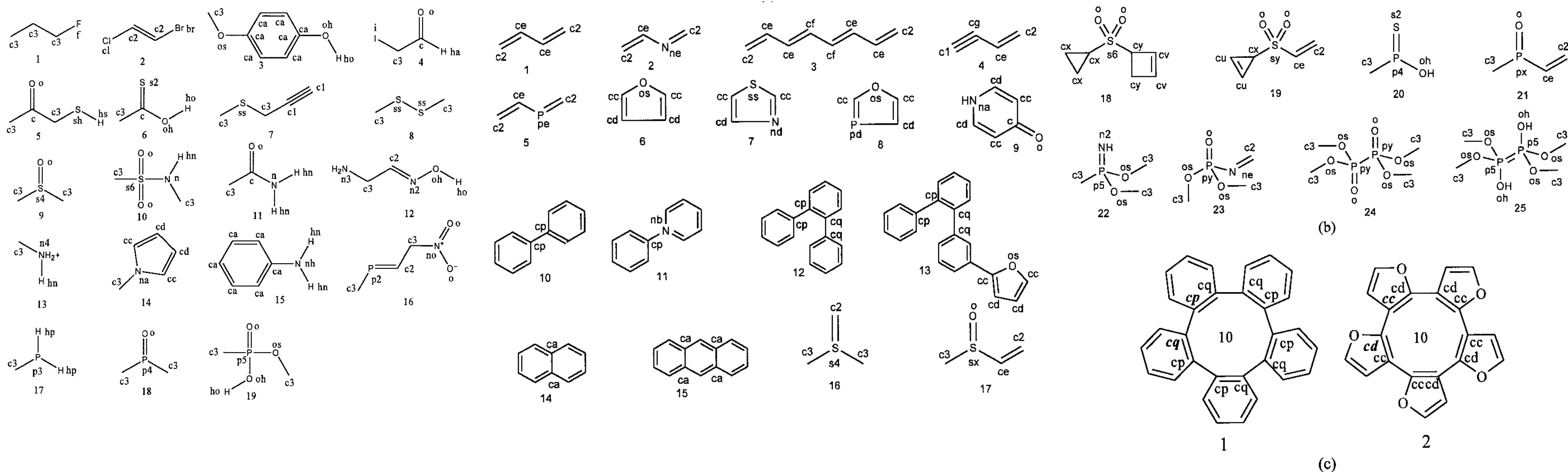
\* A new PDP-11 would have cost \$160,000 in today's dollars!

# HOW ARE FORCEFIELDS MADE?



# 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!

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

## training a neural network

```
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

```
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

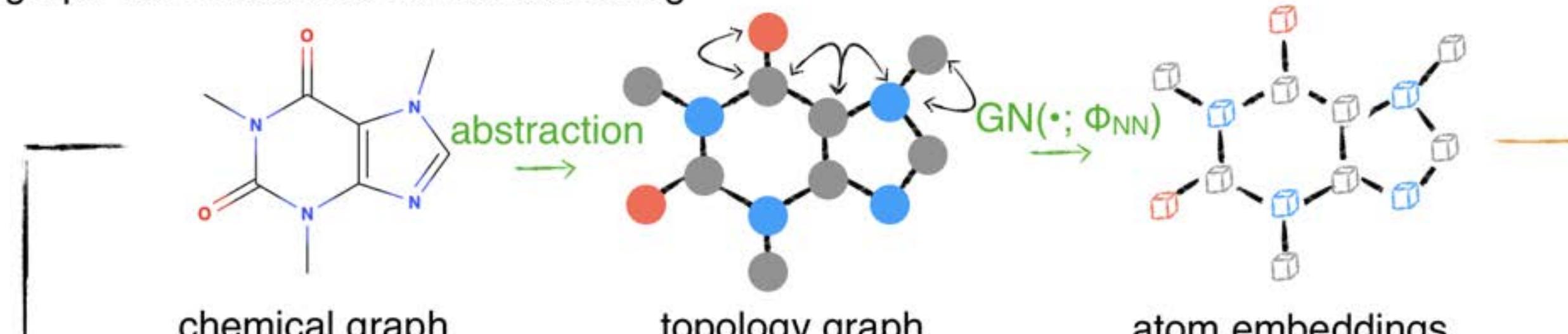


# espaloma: extensible surrogate potential of *ab initio* learned and optimized by message-passing algorithm

use of only **chemical graph**

means that model can generate parameters for small molecules, proteins, nucleic acids, covalent ligands, carbohydrates, etc.

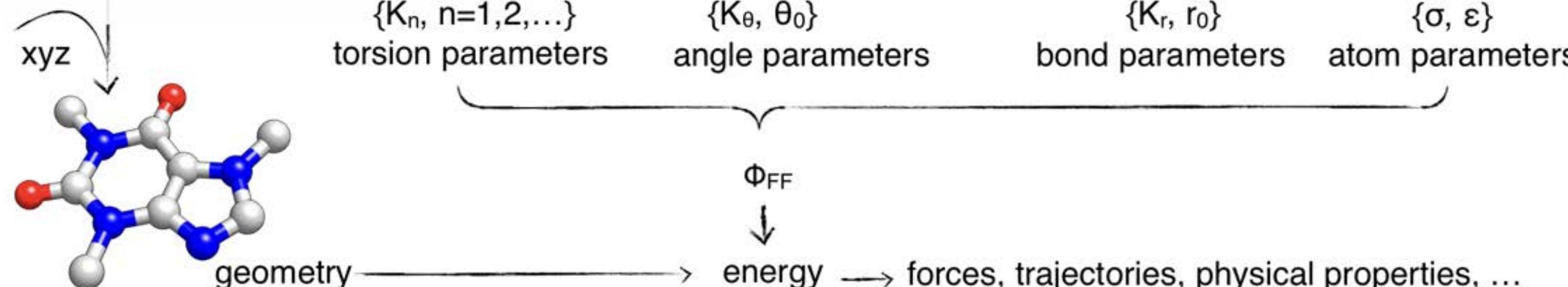
## Stage 1: graph net continuous atom embedding



## Stage 2: symmetry-preserving pooling

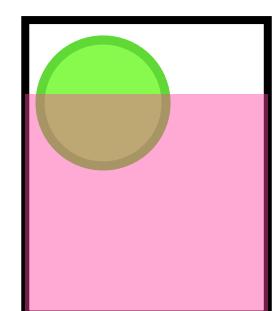
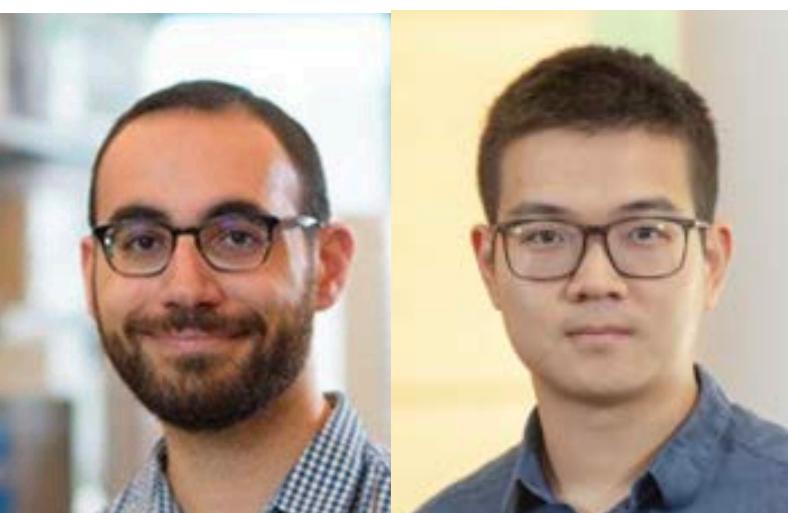
$$\begin{aligned} \text{torsion embeddings} &= NN_\phi(\text{atom embeddings}; \Phi_{NN}) \\ &= NN_\phi(\text{atom embeddings}; \Phi_{NN}) + NN_\phi(\text{atom embeddings}; \Phi_{NN}) \\ \text{angle embeddings} &= NN_\theta(\text{atom embeddings}; \Phi_{NN}) \\ &= NN_\theta(\text{atom embeddings}; \Phi_{NN}) + NN_\theta(\text{atom embeddings}; \Phi_{NN}) \\ \text{bond embeddings} &= NN_r(\text{atom embeddings}; \Phi_{NN}) \\ &= NN_r(\text{atom embeddings}; \Phi_{NN}) + NN_r(\text{atom embeddings}; \Phi_{NN}) + \dots \end{aligned}$$

## Stage 3: neural parametrization

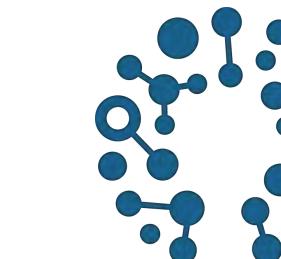


JOSH FASS

YUANQING  
WANG



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



open  
forcefield  
initiative

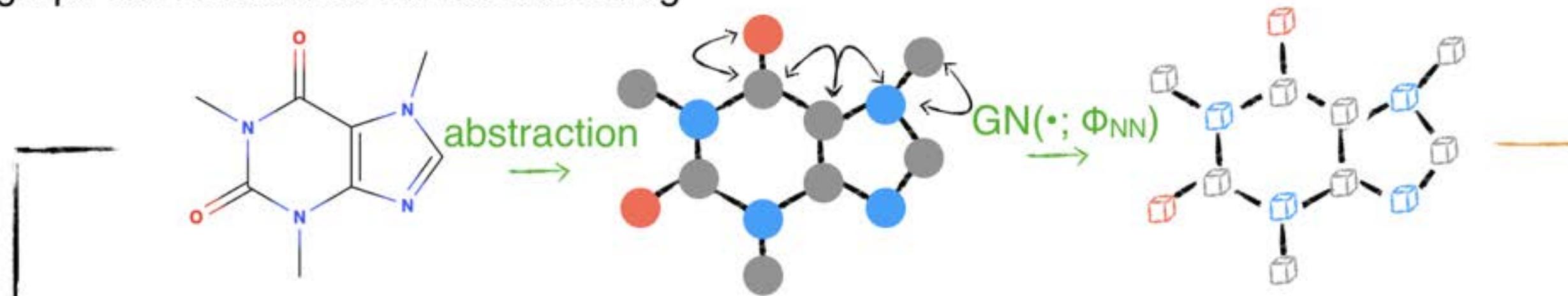


Funded by OpenFF  
NIH R&D grant

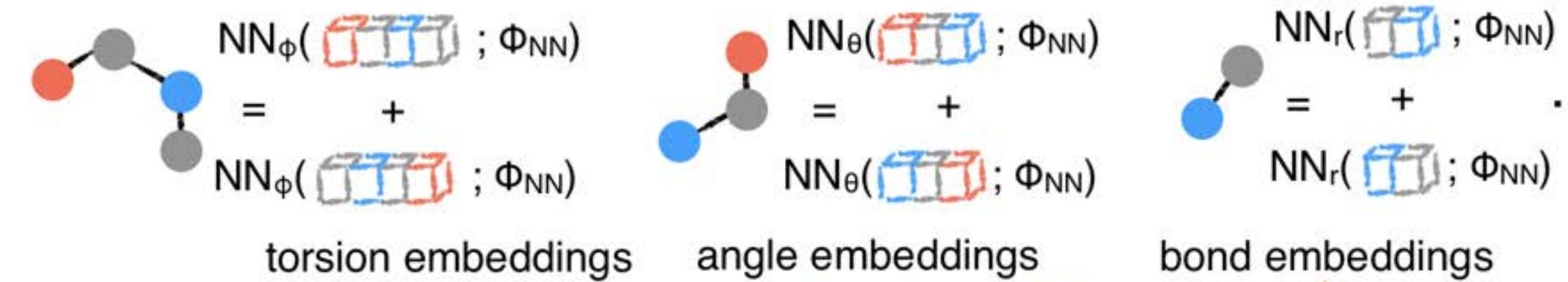


# espaloma: extensible surrogate potential of *ab initio* learned and optimized by message-passing algorithm

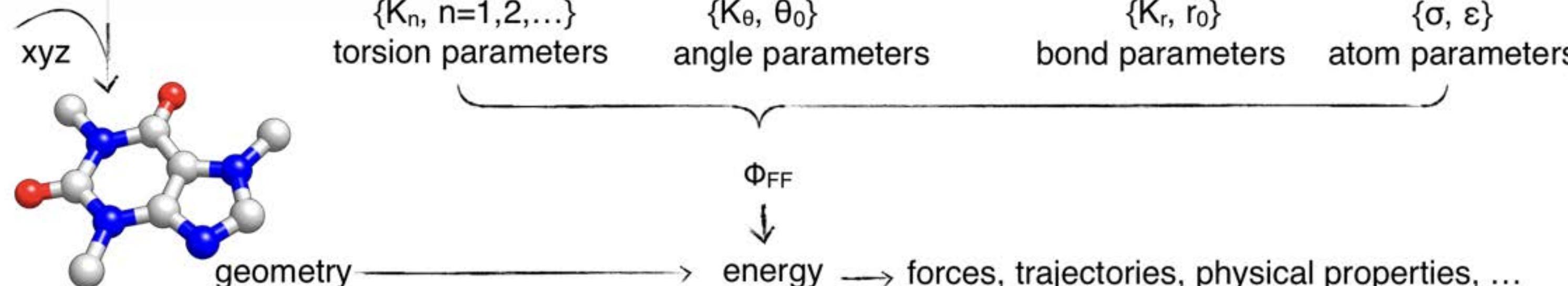
Stage 1: graph net continuous atom embedding



Stage 2: symmetry-preserving pooling



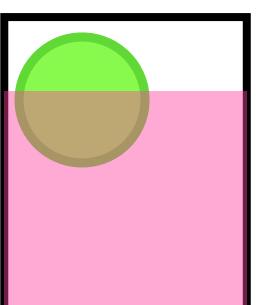
Stage 3: neural parametrization



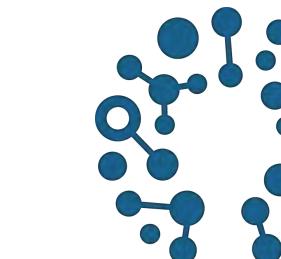
entire model is **end-to-end differentiable** so can be fit to any loss function by standard automatic differentiation machine learning frameworks

YUANQING  
WANG

JOSH FASS



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



open  
forcefield  
initiative

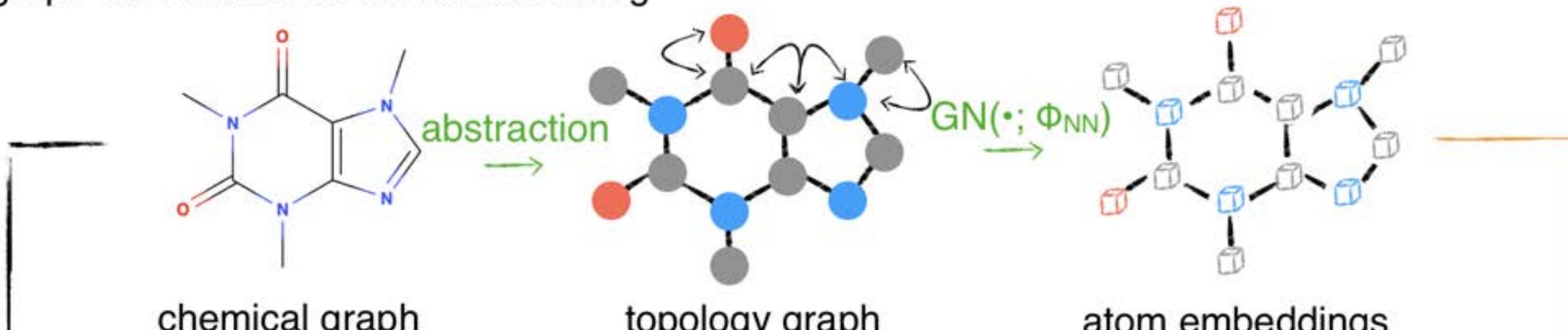


Funded by OpenFF  
NIH R&D grant

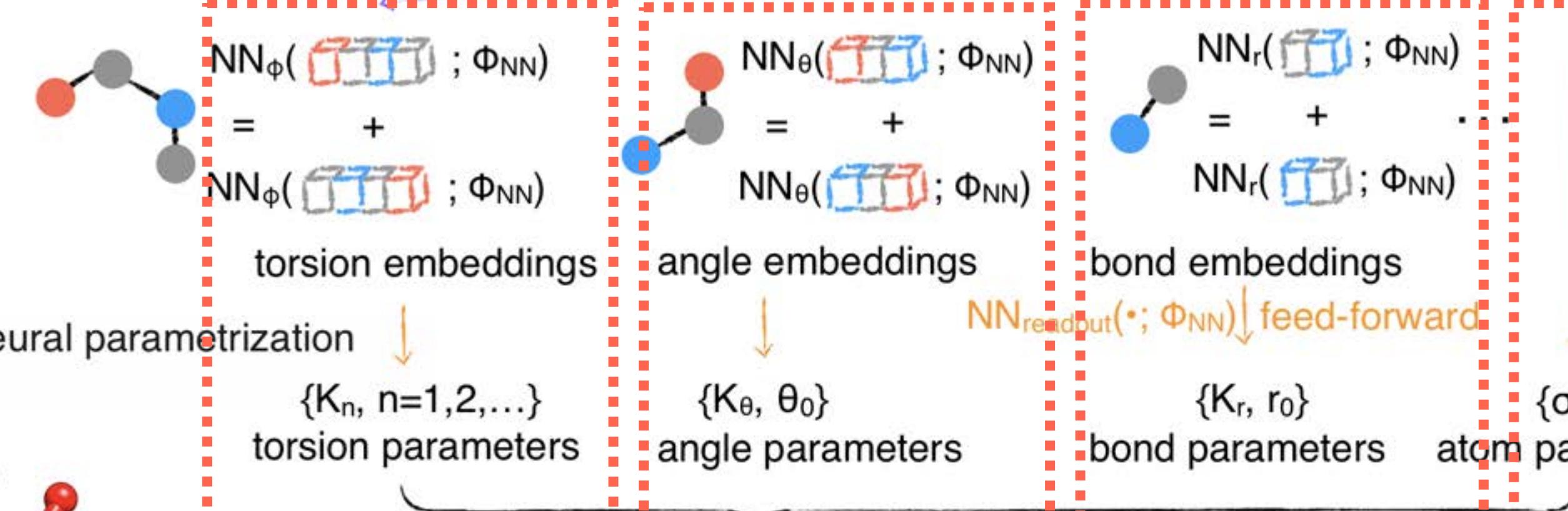


# espaloma: extensible surrogate potential of *ab initio* learned and optimized by message-passing algorithm

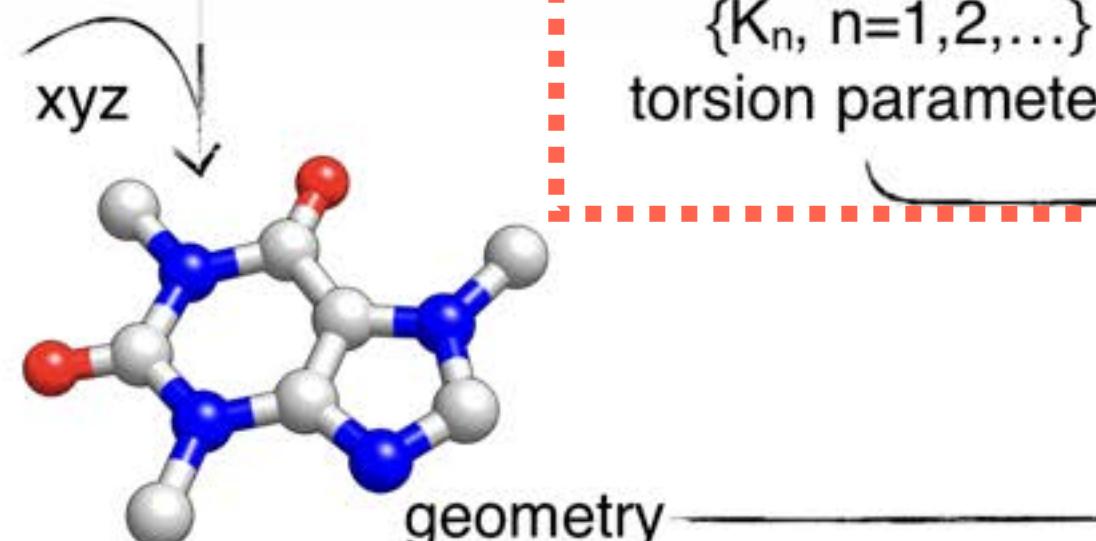
Stage 1: graph net continuous atom embedding



Stage 2: symmetry-preserving pooling



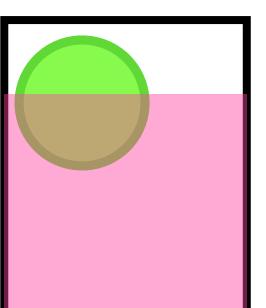
Stage 3: neural parametrization



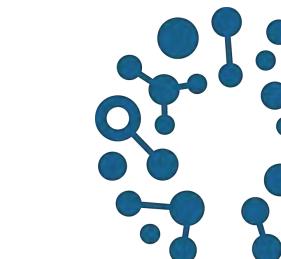
**modular and extensible handling of potential terms:**  
charge model parameters,  
point polarizabilities,  
alternative vdW forms,  
special 1-4 parameters, etc.

JOSH FASS

YUANQING  
WANG



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



open  
forcefield  
initiative



Funded by OpenFF  
NIH R&D grant

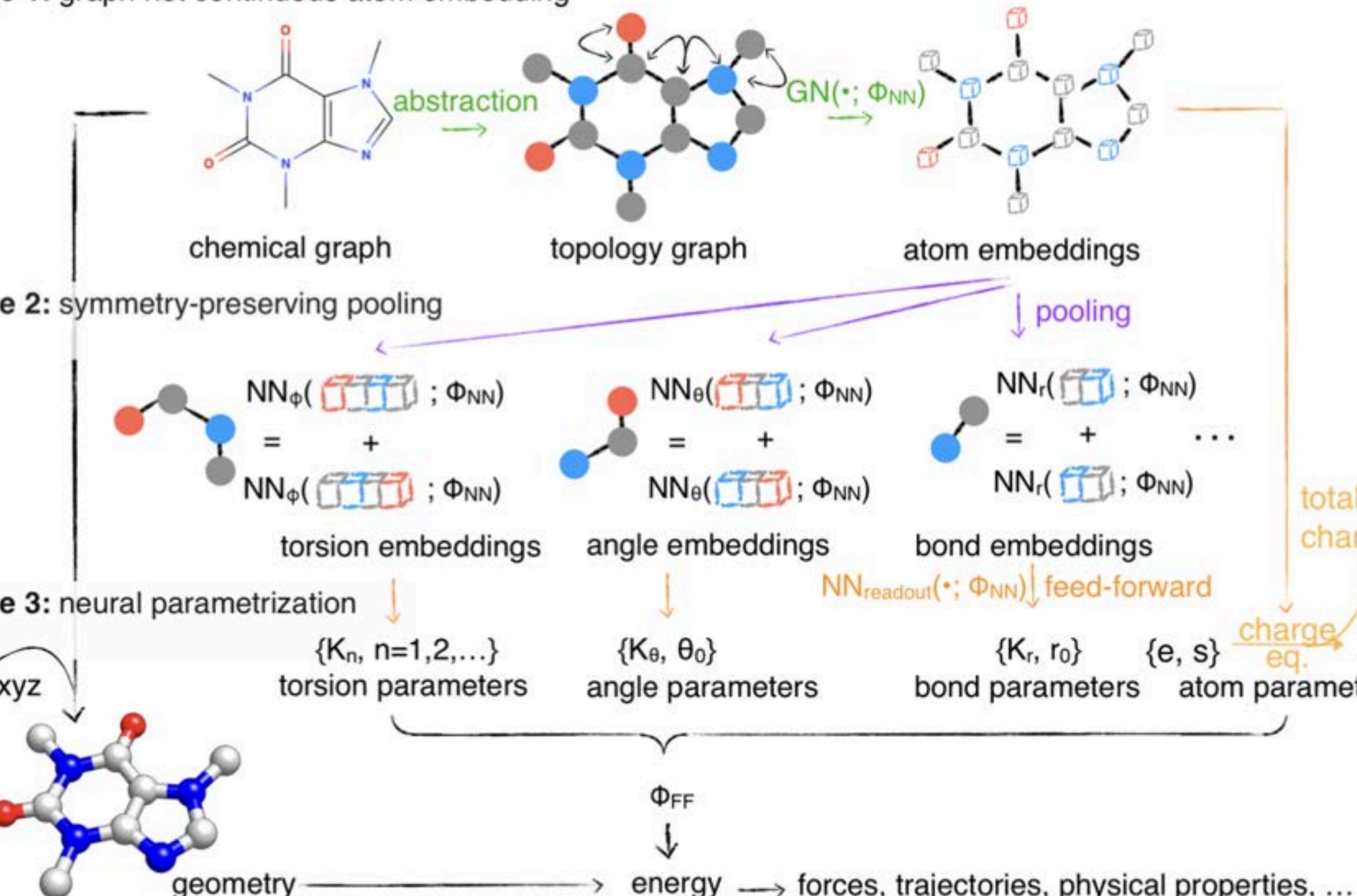


# ESPALOMA MAKES BUILDING A NEW FORCE FIELD EASY

## building a new force field

### espaloma architecture

Stage 1: graph net continuous atom embedding



(implemented in pytorch)

<http://github.com/choderalab/espaloma>



YUANQING WANG

```
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()

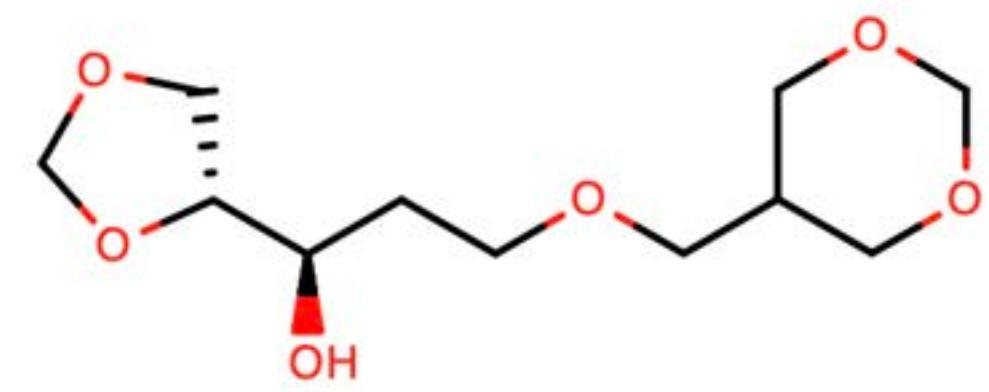
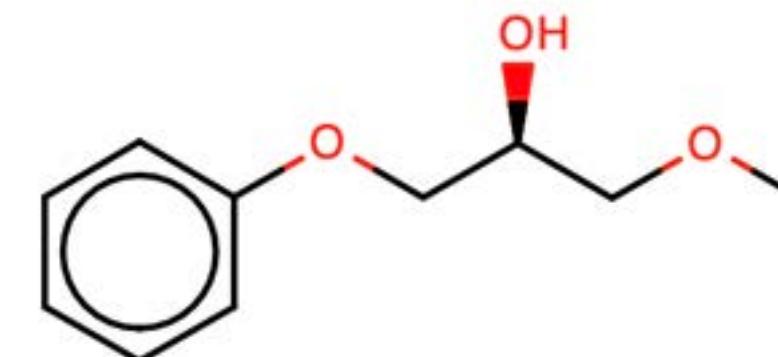
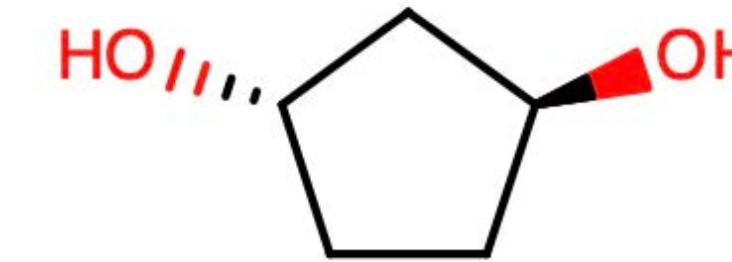
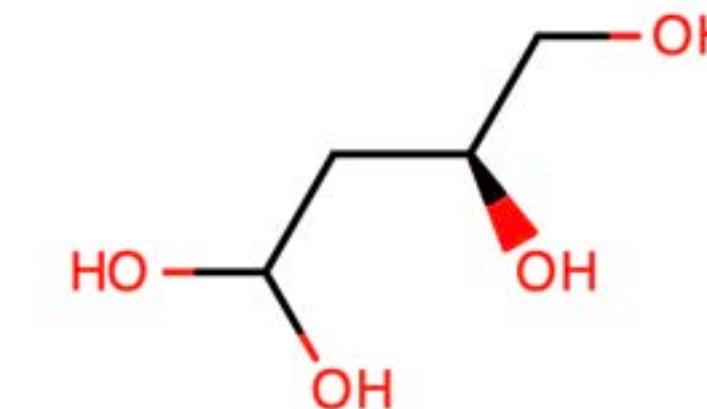
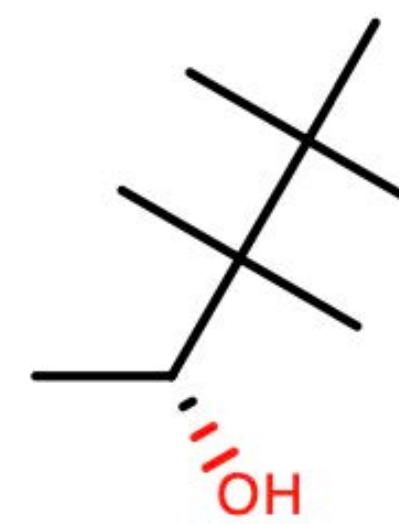
torch.save(espaloma_model, "espaloma_model.pt") # save model
```

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
<b>PhAlkEthOH</b> (simple CHO)	7408	12592	244036	0.8656 <sup>0.9131</sup> <sub>0.8225</sub>	1.1398 <sup>1.2332</sup> <sub>1.0715</sub>	1.6071 <sup>1.6915</sup> <sub>1.5197</sub>	1.7267 <sup>1.7935</sup> <sub>1.6543</sub>	1.7406 <sup>1.8148</sup> <sub>1.6679</sub>	

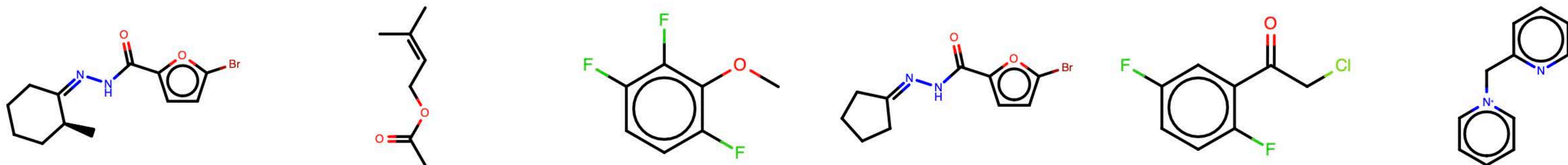
**PhAlkEthOh: Phenyls, Alkanes, Ethers, and alcohols (OH)**  
(a low-complexity chemical space)



# 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 <sup>0.9131</sup> <sub>0.8225</sub>	1.1398 <sup>1.2332</sup> <sub>1.0715</sub>	1.6071 <sup>1.6915</sup> <sub>1.5197</sub>	1.7267 <sup>1.7935</sup> <sub>1.6543</sub>	1.7406 <sup>1.8148</sup> <sub>1.6679</sub>	
OpenFF Gen2 Optimization (druglike)	792	3977	23748	0.7413 <sup>0.7920</sup> <sub>0.6914</sub>	0.7600 <sup>0.8805</sup> <sub>0.6644</sub>	2.1768 <sup>2.3388</sup> <sub>2.0380</sub>	2.4274 <sup>2.5207</sup> <sub>2.3300</sub>	2.5386 <sup>2.6640</sup> <sub>2.4370</sub>	

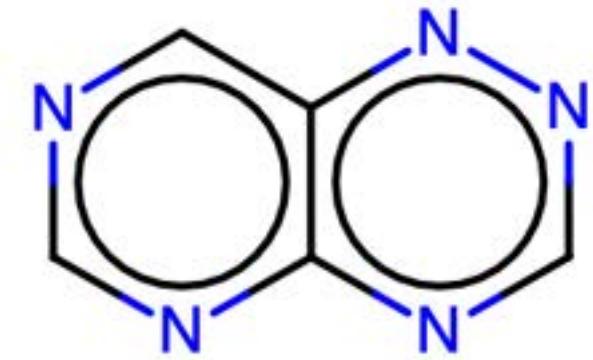
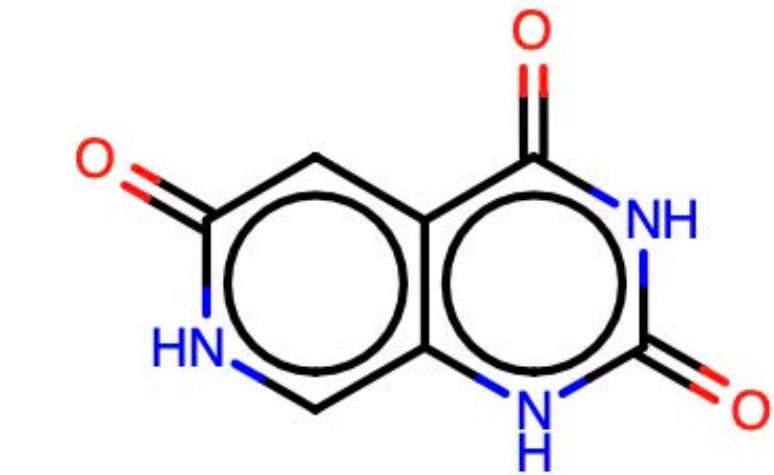
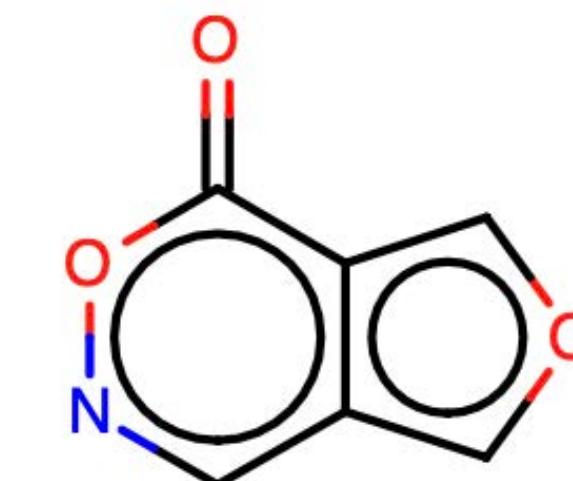
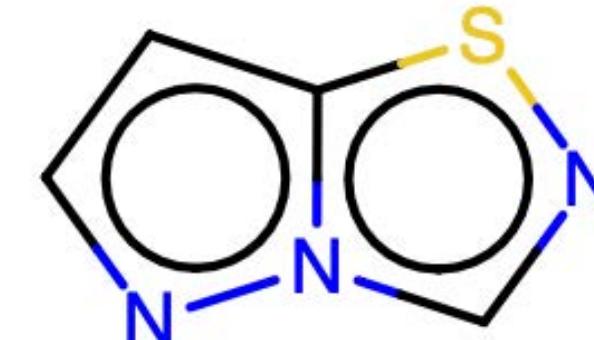
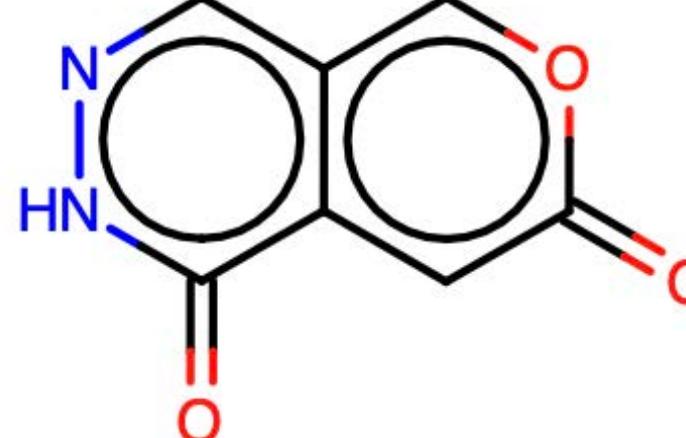
**OpenFF Gen2 Optimization set:** Diverse druglike fragments challenging for force fields  
(a moderate-complexity chemical space)



# 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
<b>PhAlkEthOH</b> (simple CHO)	7408	12592	244036	0.8656 <sup>0.9131</sup> <sub>0.8225</sub>	1.1398 <sup>1.2332</sup> <sub>1.0715</sub>	1.6071 <sup>1.6915</sup> <sub>1.5197</sub>	1.7267 <sup>1.7935</sup> <sub>1.6543</sub>	1.7406 <sup>1.8148</sup> <sub>1.6679</sub>	
<b>OpenFF Gen2 Optimization</b> (druglike)	792	3977	23748	0.7413 <sup>0.7920</sup> <sub>0.6914</sub>	0.7600 <sup>0.8805</sup> <sub>0.6644</sub>	2.1768 <sup>2.3388</sup> <sub>2.0380</sub>	2.4274 <sup>2.5207</sup> <sub>2.3300</sub>	2.5386 <sup>2.6640</sup> <sub>2.4370</sub>	
<b>VEHICLe</b> (heterocyclic)	24867	24867	234326	0.4476 <sup>0.4690</sup> <sub>0.4273</sub>	0.4233 <sup>0.4414</sup> <sub>0.4053</sub>	8.0247 <sup>8.2456</sup> <sub>7.8271</sub>	8.0077 <sup>8.2313</sup> <sub>7.7647</sub>	9.4014 <sup>9.6434</sup> <sub>9.2135</sub>	

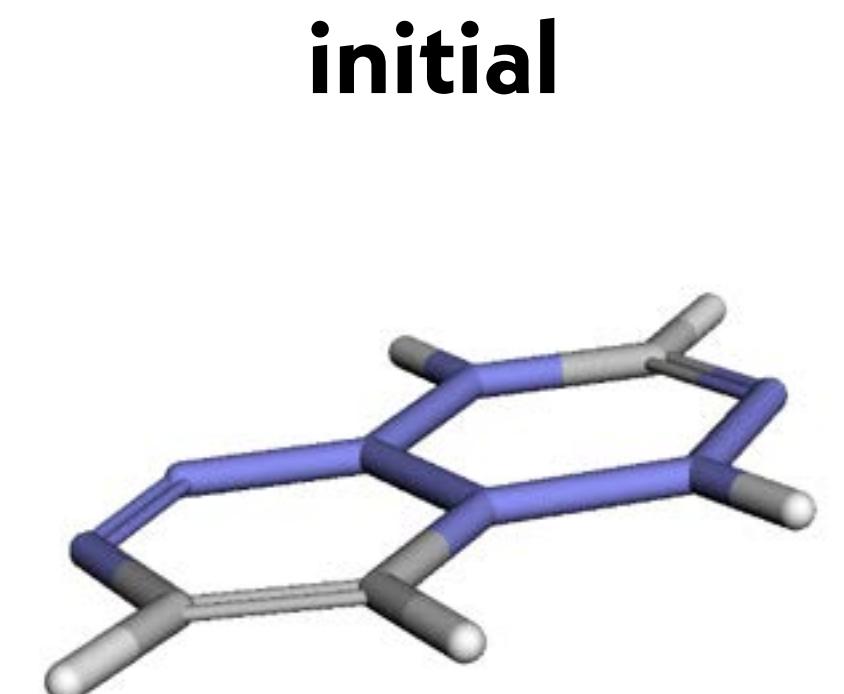
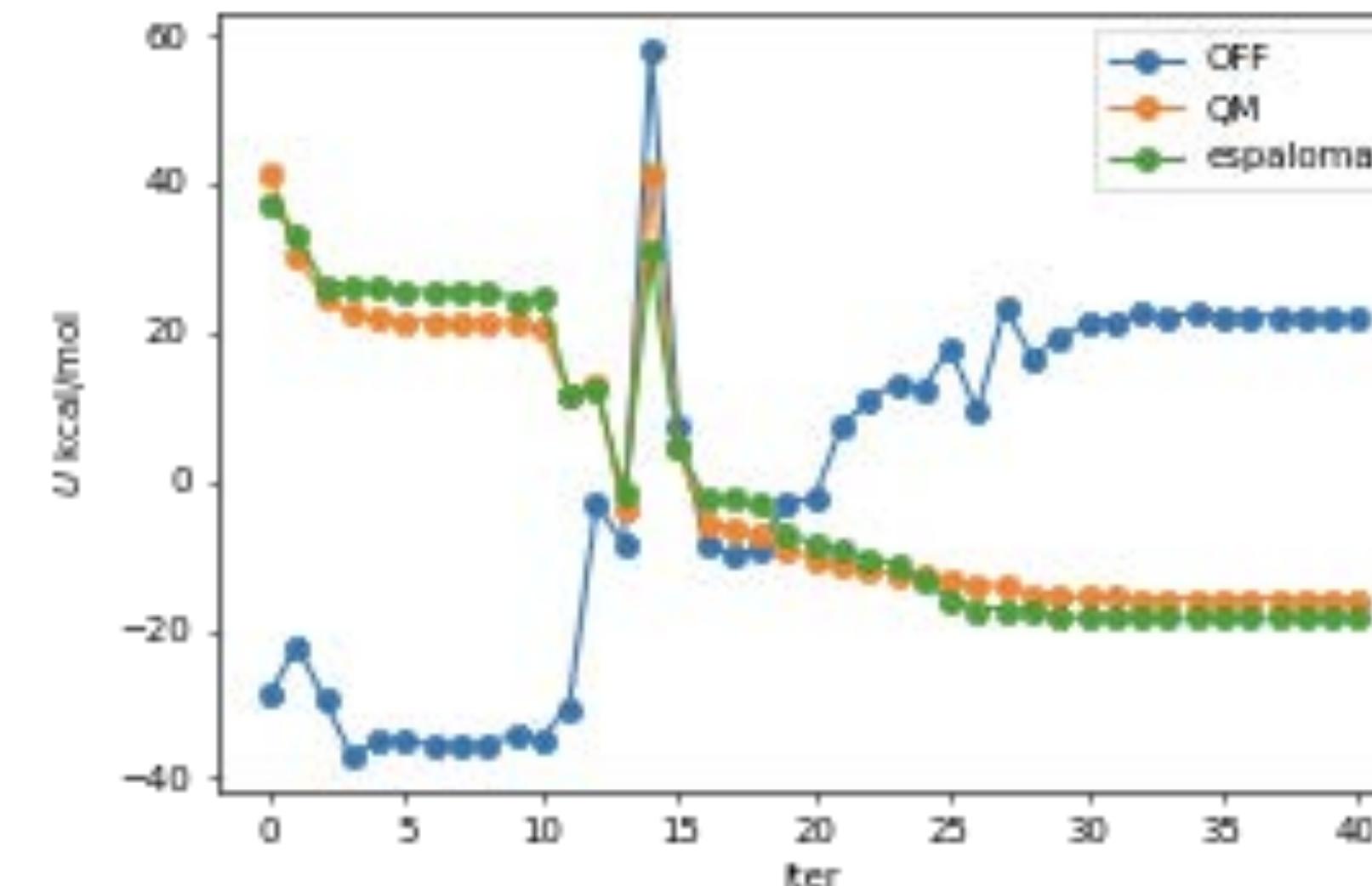
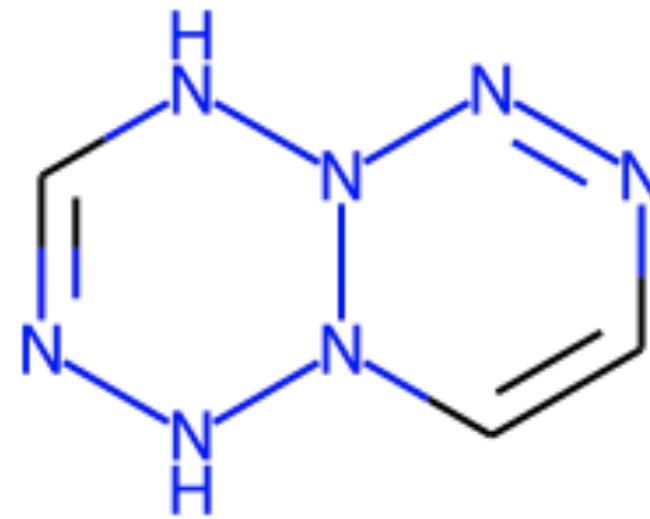
**VEHICLe**: Virtual exploratory heterocyclic drug scaffold library  
(aromatic bicyclic heterocyclic compounds containing C, N, O, S, H)



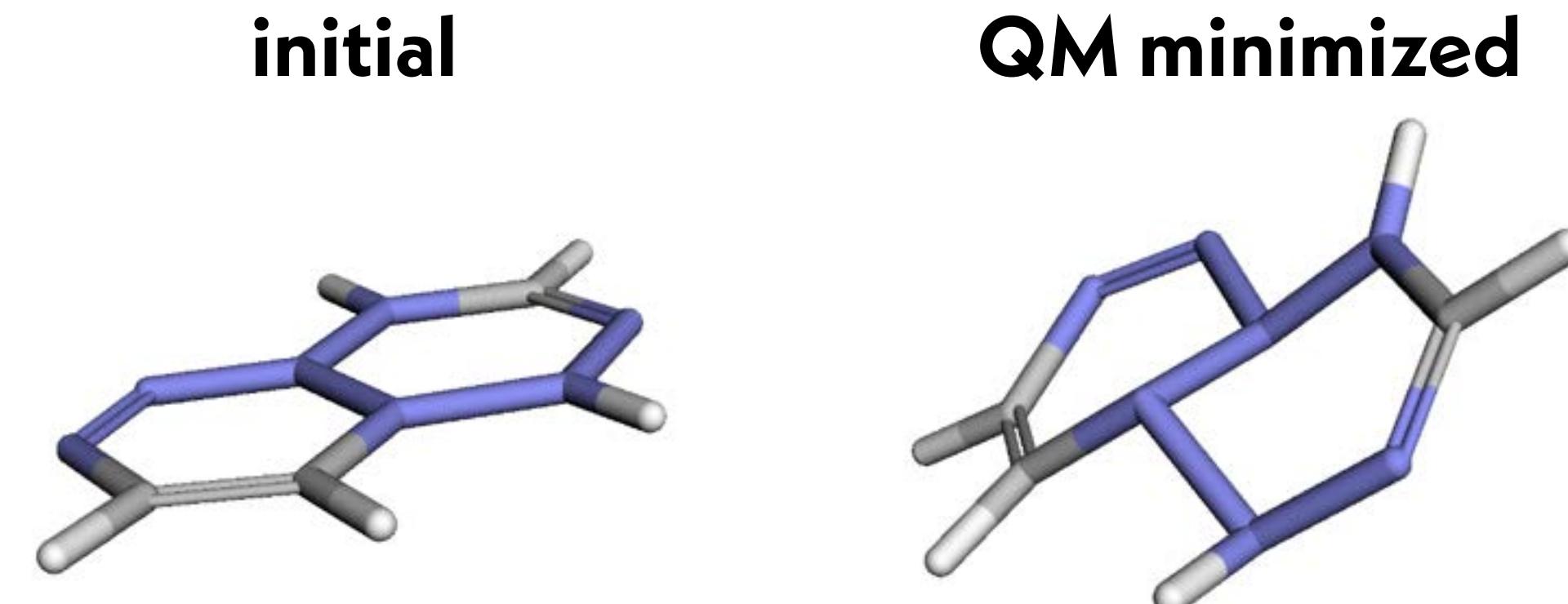
# ESPALOMA OUTPERFORMS CURRENT FORCE FIELDS IN QM ACCURACY AND CAN BE EASILY TRAINED FOR HETEROGENEOUS SYSTEMS

(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
<b>PhAlkEthOH</b> (simple CHO)	7408	12592	244036	0.8656 <sup>0.9131</sup> <sub>0.8225</sub>	1.1398 <sup>1.2332</sup> <sub>1.0715</sub>	1.6071 <sup>1.6915</sup> <sub>1.5197</sub>	1.7267 <sup>1.7935</sup> <sub>1.6543</sub>	1.7406 <sup>1.8148</sup> <sub>1.6679</sub>	
<b>OpenFF Gen2 Optimization</b> (druglike)	792	3977	23748	0.7413 <sup>0.7920</sup> <sub>0.6914</sub>	0.7600 <sup>0.8805</sup> <sub>0.6644</sub>	2.1768 <sup>2.3388</sup> <sub>2.0380</sub>	2.4274 <sup>2.5207</sup> <sub>2.3300</sub>	2.5386 <sup>2.6640</sup> <sub>2.4370</sub>	
<b>VEHICLE</b> (heterocyclic)	24867	24867	234326	0.4476 <sup>0.4690</sup> <sub>0.4273</sub>	0.4233 <sup>0.4414</sup> <sub>0.4053</sub>	8.0247 <sup>8.2456</sup> <sub>7.8271</sub>	8.0077 <sup>8.2313</sup> <sub>7.7647</sub>	9.4014 <sup>9.6434</sup> <sub>9.2135</sub>	

Comparison with QC Archive data



DFT B3LYP-D3(BJ) / DZVP

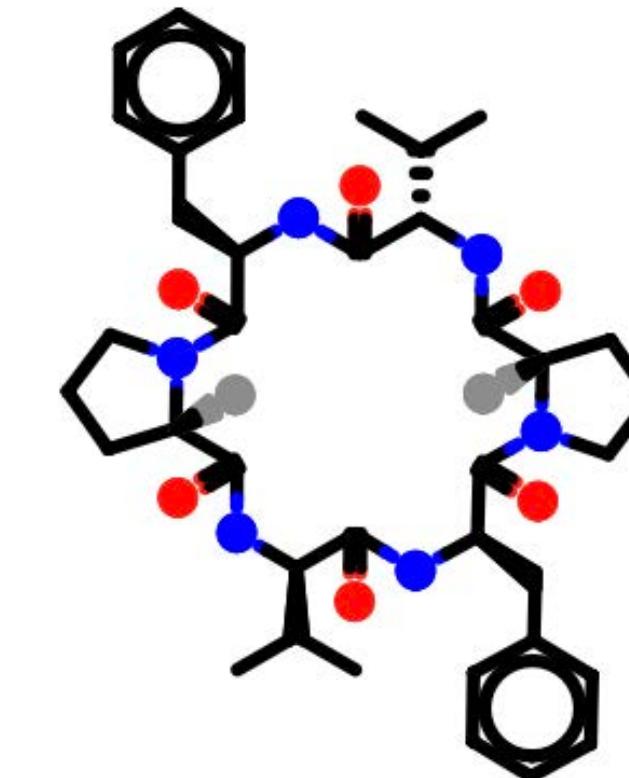
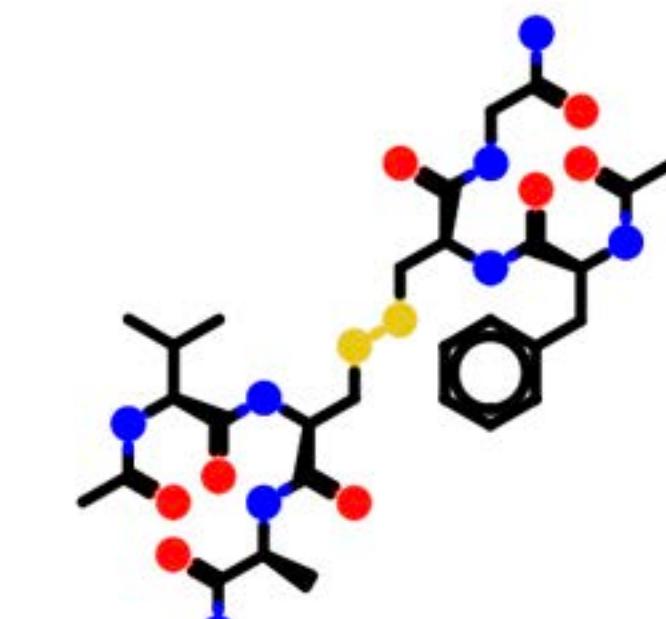
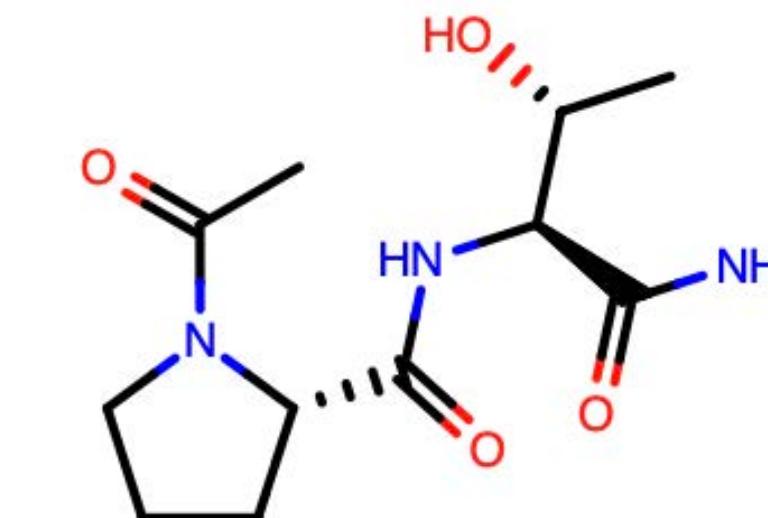
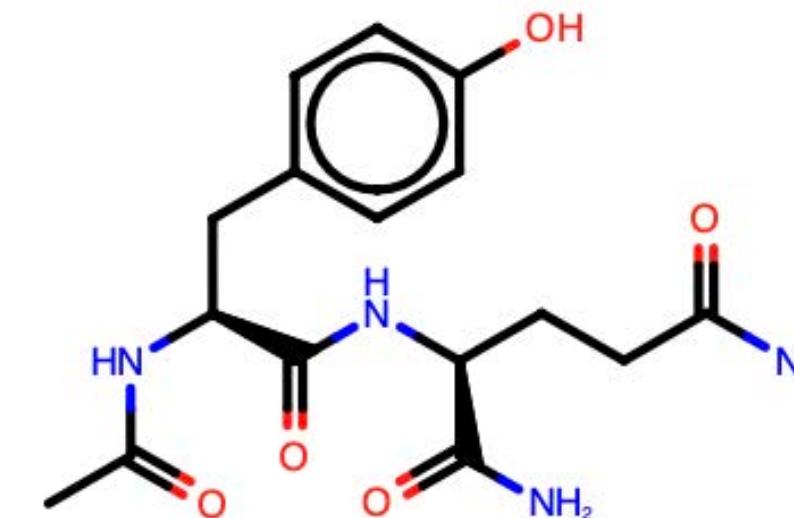


YUANQING WANG

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

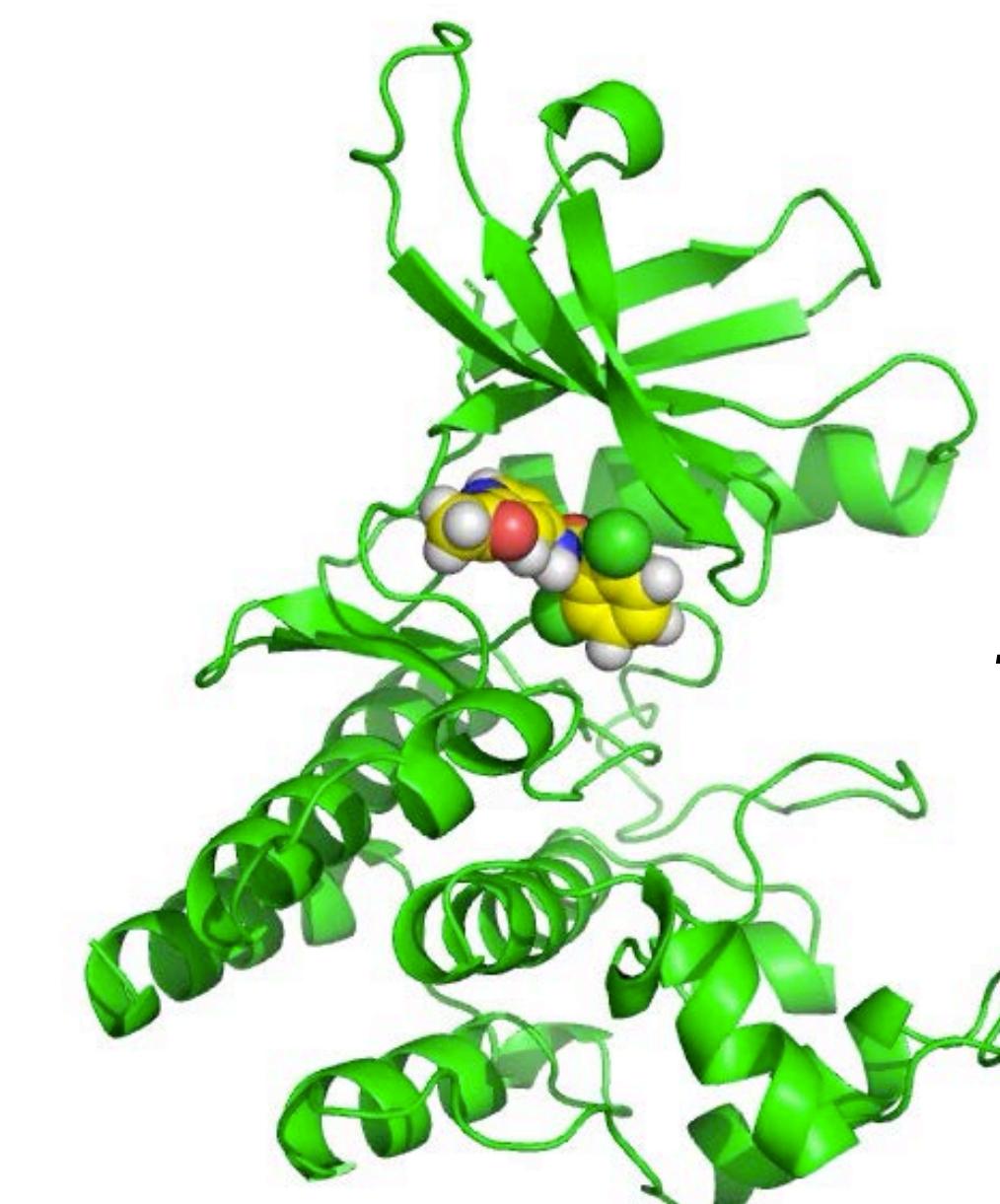
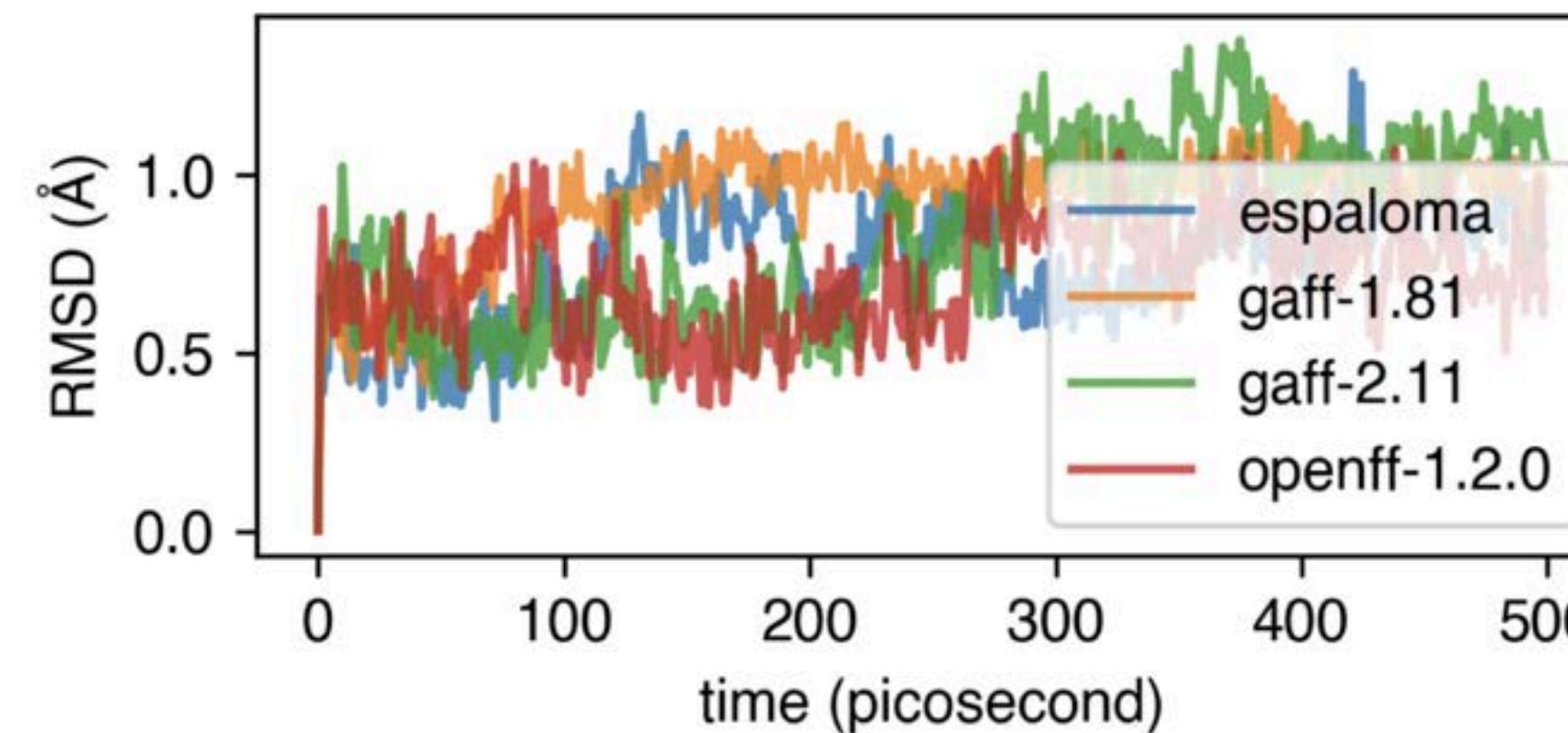
(a) dataset	# mols	# trajs	# snapshots	Espaloma RMSE		Legacy FF RMSE (kcal/mol) (Test molecules)			
				Train	Test	OpenFF 1.2.0	GAFF-1.81	GAFF-2.11	Amber ff14SB
<b>PhAlkEthOH</b> (simple CHO)	7408	12592	244036	0.8656 <sup>0.9131 0.8225</sup>	1.1398 <sup>1.2332 1.0715</sup>	1.6071 <sup>1.6915 1.5197</sup>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

**PepConf:** Short peptides, including disulfides and cyclic peptides



# 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
<b>PhAlkEthOH</b> (simple CHO)	7408	12592	244036	0.8656 <sup>0.9131 0.8225</sup>	1.1398 <sup>1.2332 1.0715</sup>	1.6071 <sup>1.6915 1.5197</sup>	1.7267 <sup>1.7935 1.6543</sup>	1.7406 <sup>1.8148 1.6679</sup>	
<b>OpenFF Gen2 Optimization</b> (druglike)	792	3977	23748	0.7413 <sup>0.7920 0.6914</sup>	0.7600 <sup>0.8805 0.6644</sup>	2.1768 <sup>2.3388 2.0380</sup>	2.4274 <sup>2.5207 2.3300</sup>	2.5386 <sup>2.6640 2.4370</sup>	
<b>VEHICLE</b> (heterocyclic)	24867	24867	234326	0.4476 <sup>0.4690 0.4273</sup>	0.4233 <sup>0.4414 0.4053</sup>	8.0247 <sup>8.2456 7.8271</sup>	8.0077 <sup>8.2313 7.7647</sup>	9.4014 <sup>9.6434 9.2135</sup>	
<b>PepConf</b> (peptides)	736	7560	22154	1.2714 <sup>1.3616 1.1899</sup>	1.8727 <sup>1.9749 1.7309</sup>	3.6143 <sup>3.7288 3.4870</sup>	4.4446 <sup>4.5738 4.3386</sup>	4.3356 <sup>4.4641 4.1965</sup>	3.1502 <sup>3.1859,* 3.1117</sup>
<b>joint</b>	OpenFF Gen2 Optimization	1528	11537	0.8264 <sup>0.9007 0.7682</sup>	1.8764 <sup>1.9947 1.7827</sup>	2.1768 <sup>2.3388 2.0380</sup>	2.4274 <sup>2.5207 2.3300</sup>	2.5386 <sup>2.6640 2.4370</sup>	
									3.1502 <sup>3.1859,* 3.1117</sup>
	PepConf			1.2038 <sup>1.3056 1.1178</sup>	1.7307 <sup>1.8439 1.6053</sup>	3.6143 <sup>3.7288 3.4870</sup>	4.4446 <sup>4.5738 4.3386</sup>	4.3356 <sup>4.4641 4.1965</sup>	



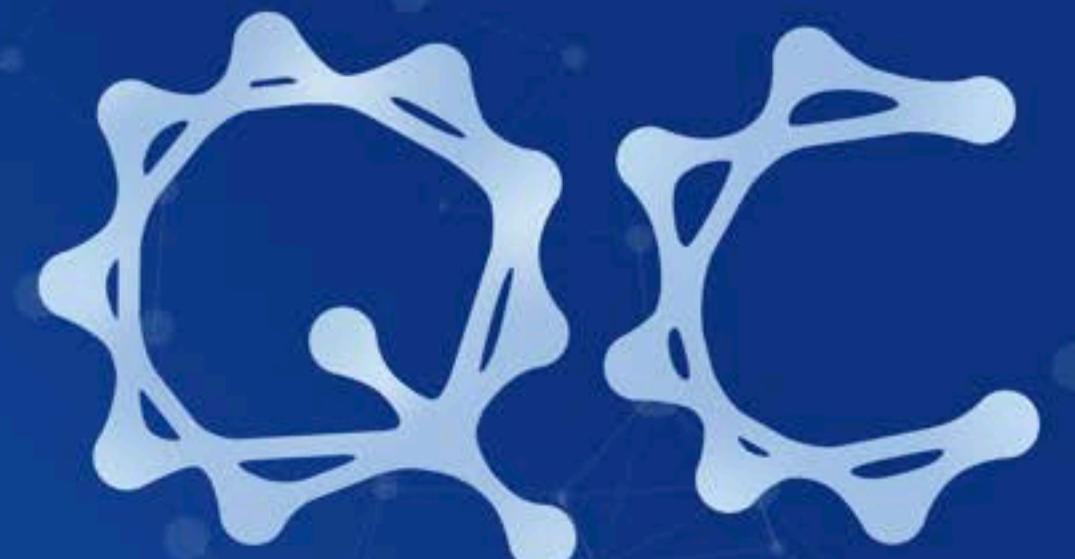
Tyk2 from OpenFF benchmark set  
espaloma joint model  
+ TIP3P water



# The MolSSI Quantum Chemistry Archive

A central source to compile, aggregate, query, and share quantum chemistry data.

GET STARTED!



## QC Archive

A MolSSI Project



FAIR Data

MolSSI hosts the QC Archive server, the largest publicly available collection of quantum chemistry data. So far, it stores over ten million computations for the molecular sciences community.



Interactive Visualization

Not only for computing and storing quantum chemistry computations at scale, but also for visualizing and understanding results as well.



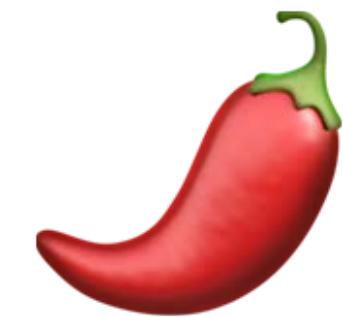
Private Instances

The infrastructure behind QC Archive is fully open-source. Spin up your own instance to compute private data and share only with collaborators.

102,477,973  
MOLECULES

108,469,316  
RESULTS

212  
COLLECTIONS



## 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  $\omega$ B97M-D3(BJ)/def2-TZVPPD level of theory**

>4M core-hours computed on QC Fractal academic clusters

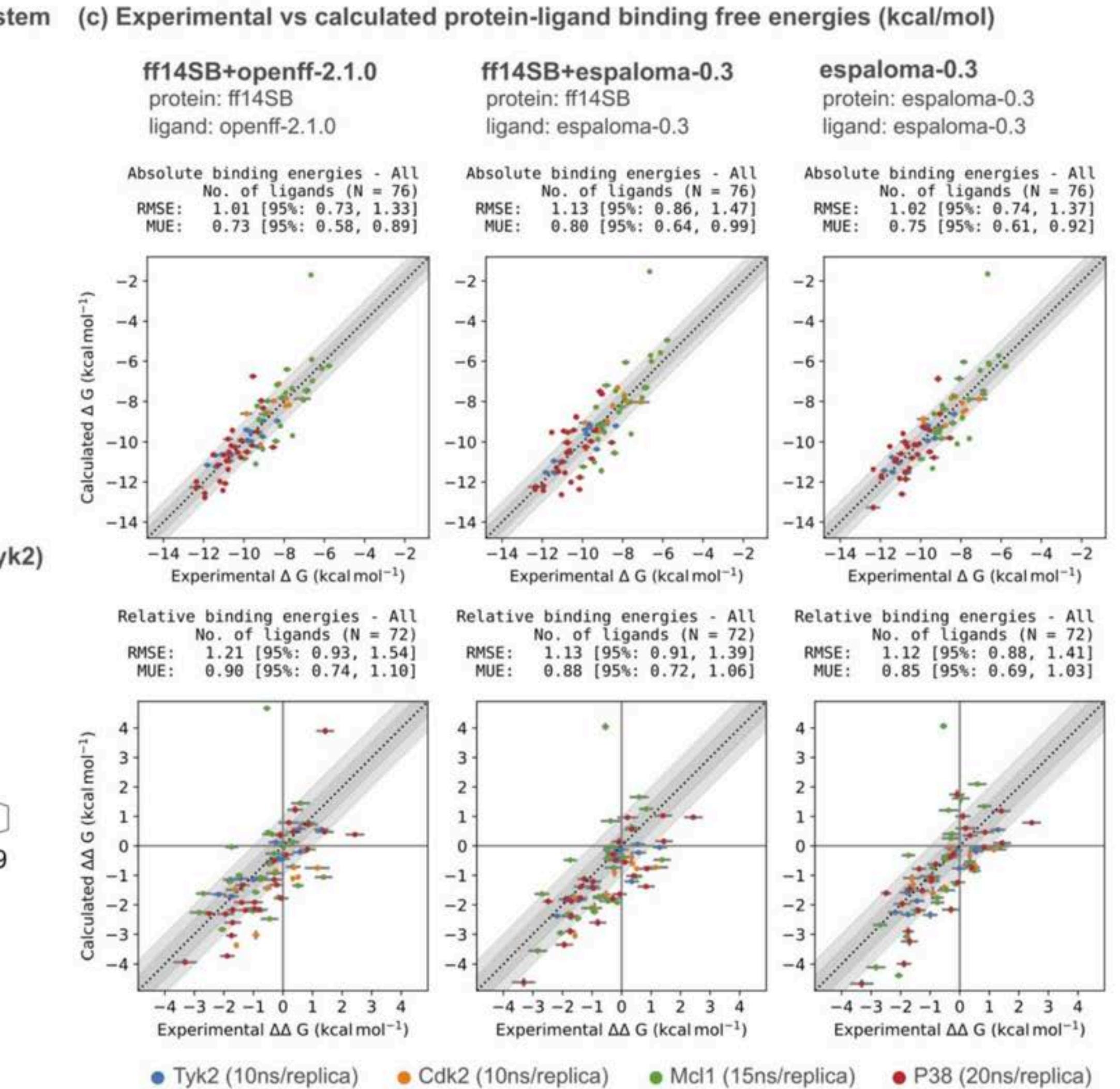
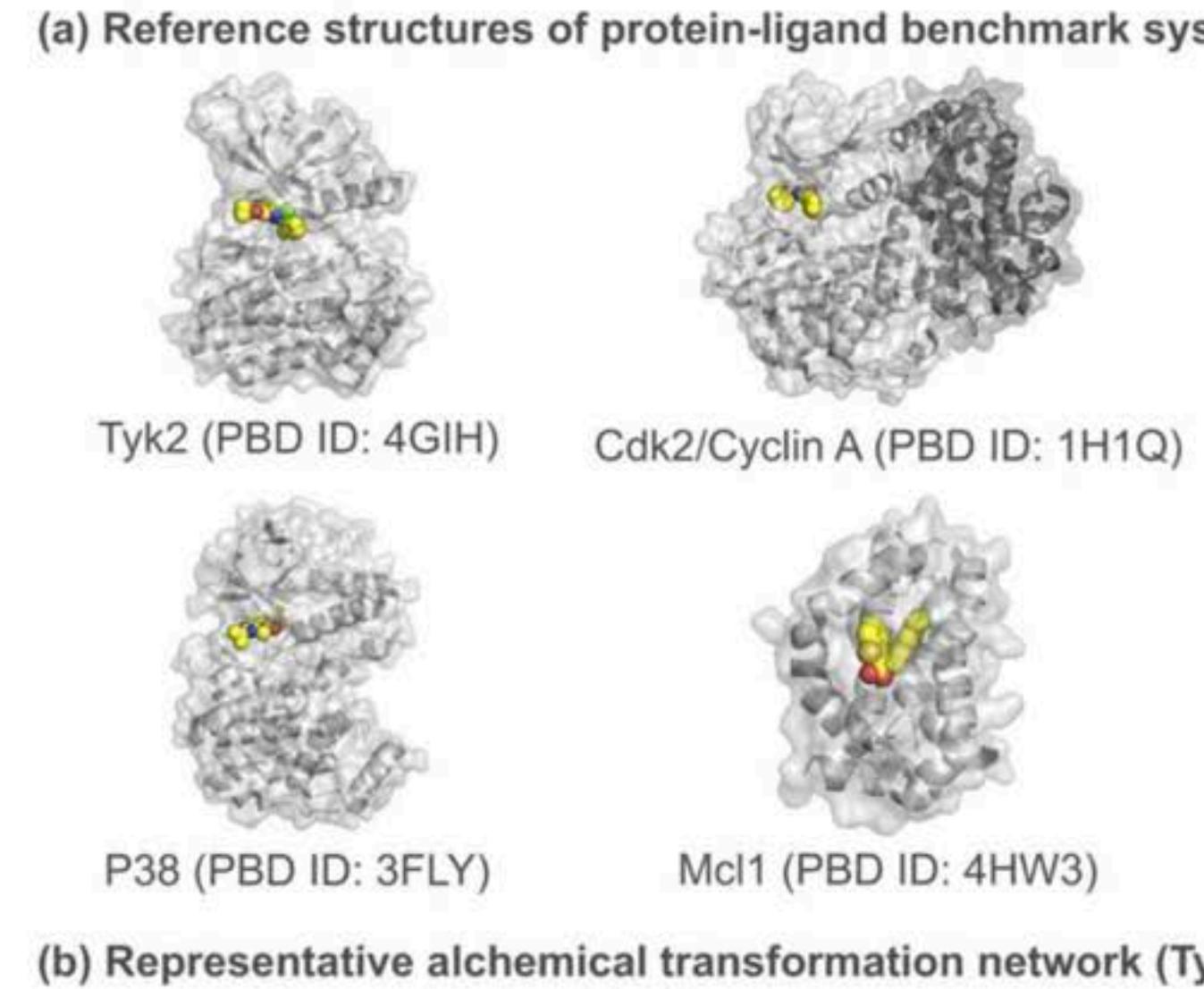
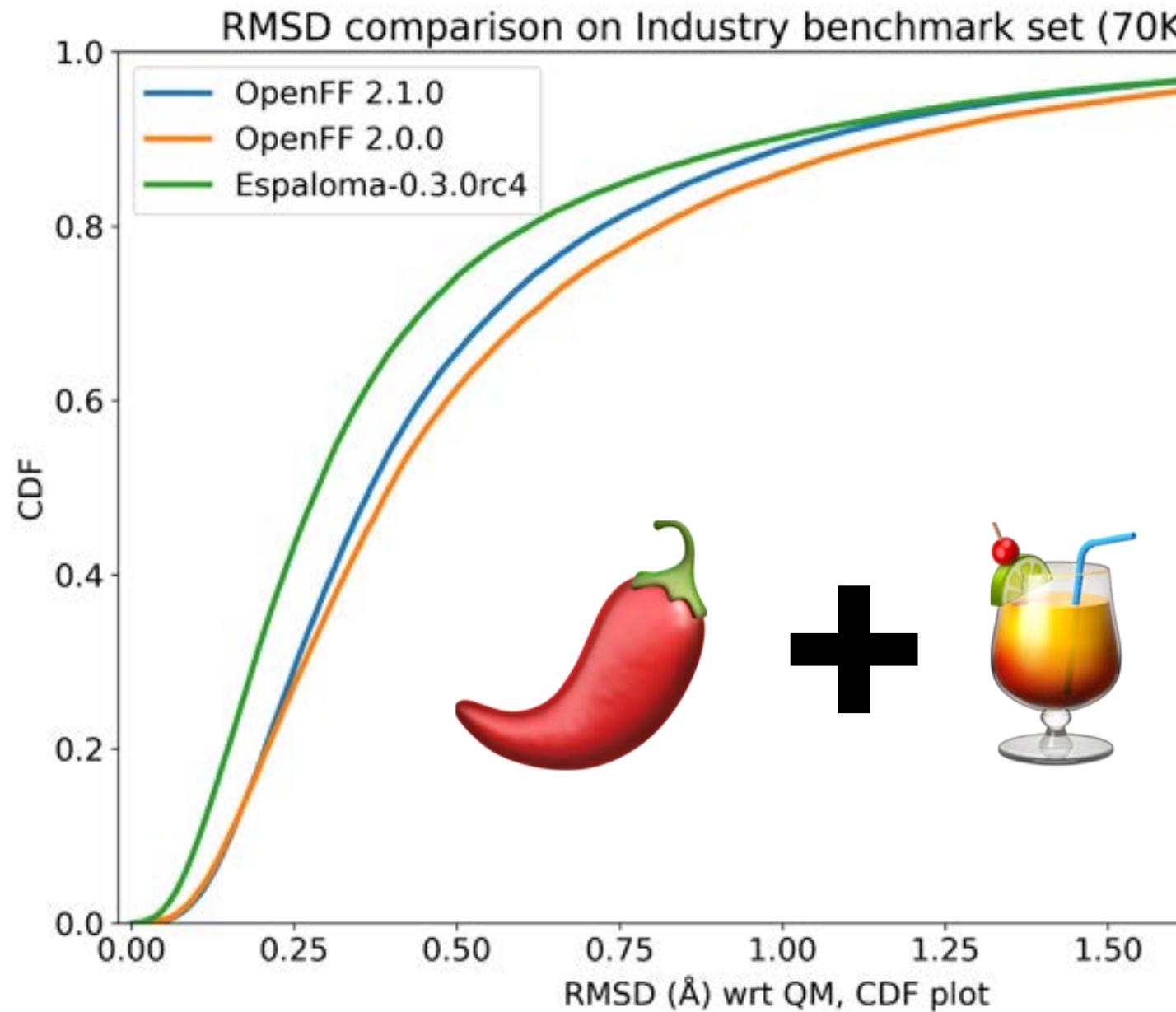
**SPICE, A Dataset of Drug-like Molecules and Peptides for Training Machine Learning Potentials**

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

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

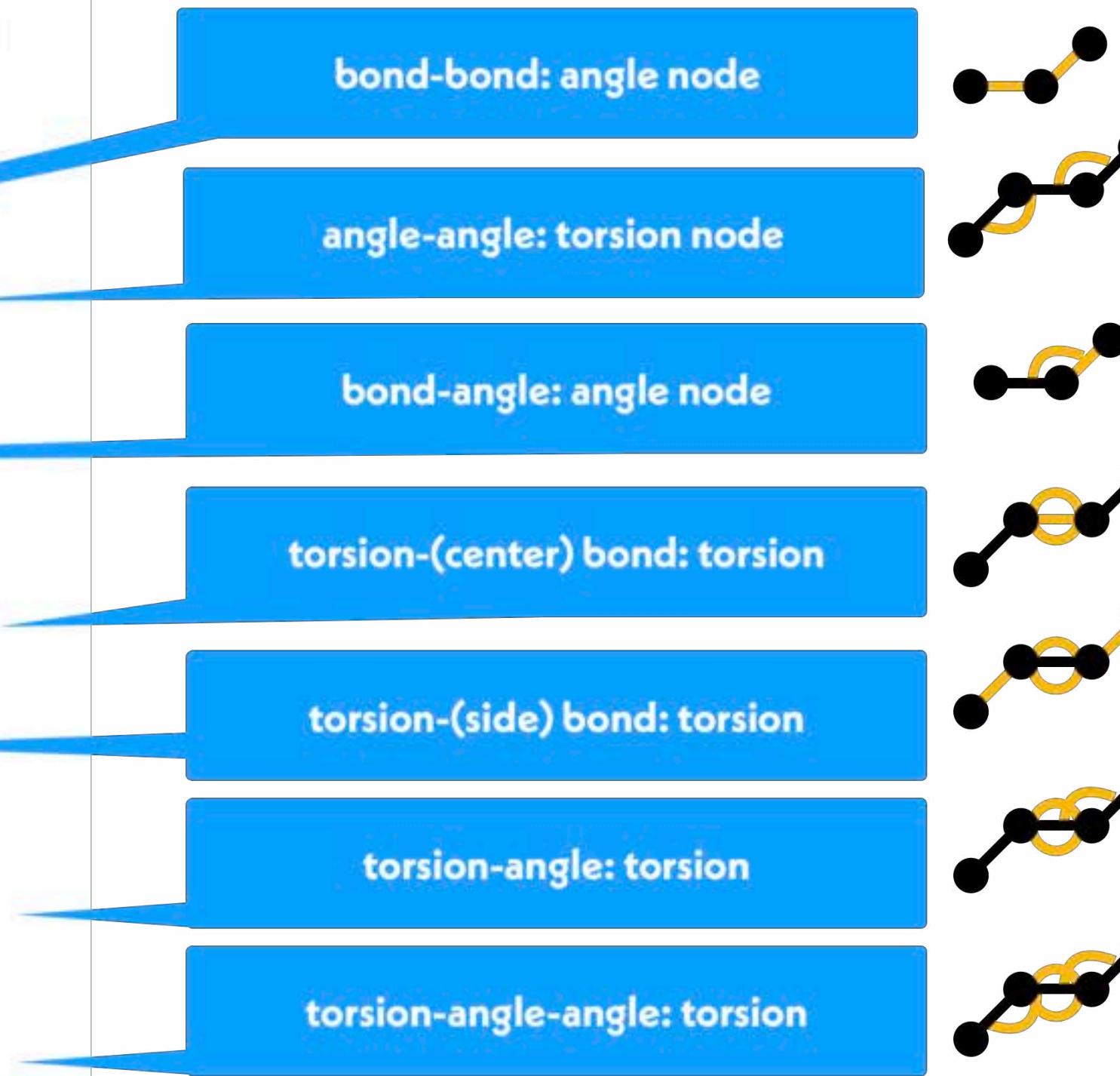
<http://qcarchive.molssi.org>

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



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

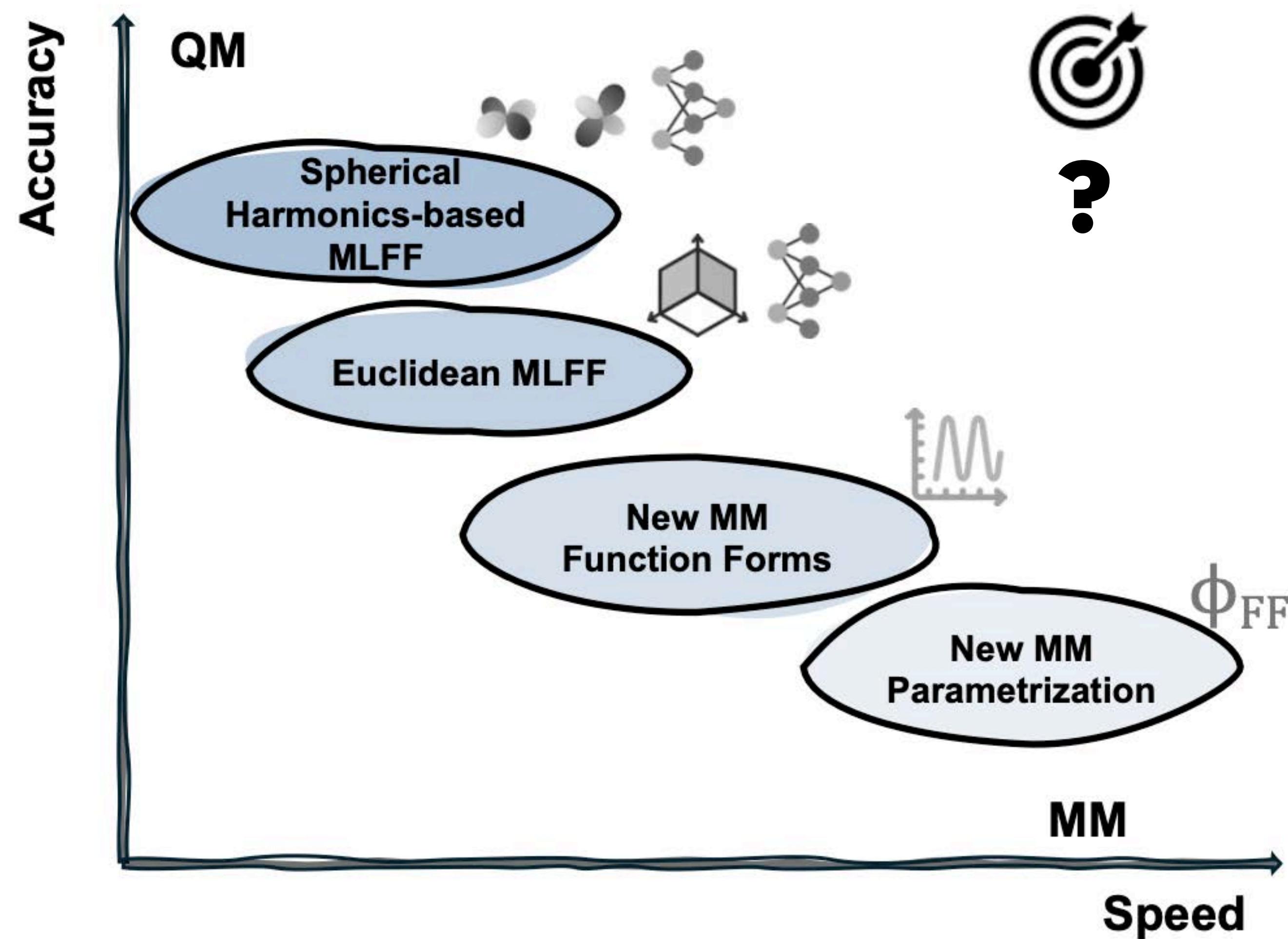
$$\begin{aligned}
 E = & \sum_b [{}^2K_b(b - b_0)^2 + {}^3K_b(b - b_0)^3 + {}^4K_b(b - b_0)^4] \\
 & + \sum_\theta [{}^2K_\theta(\theta - \theta_0)^2 + {}^3K_\theta(\theta - \theta_0)^3 + {}^4K_\theta(\theta - \theta_0)^4] \\
 & + \sum_\phi [{}^1K_\phi(1 - \cos \phi) + {}^2K_\phi(1 - \cos 2\phi) + {}^3K_\phi(1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



**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?

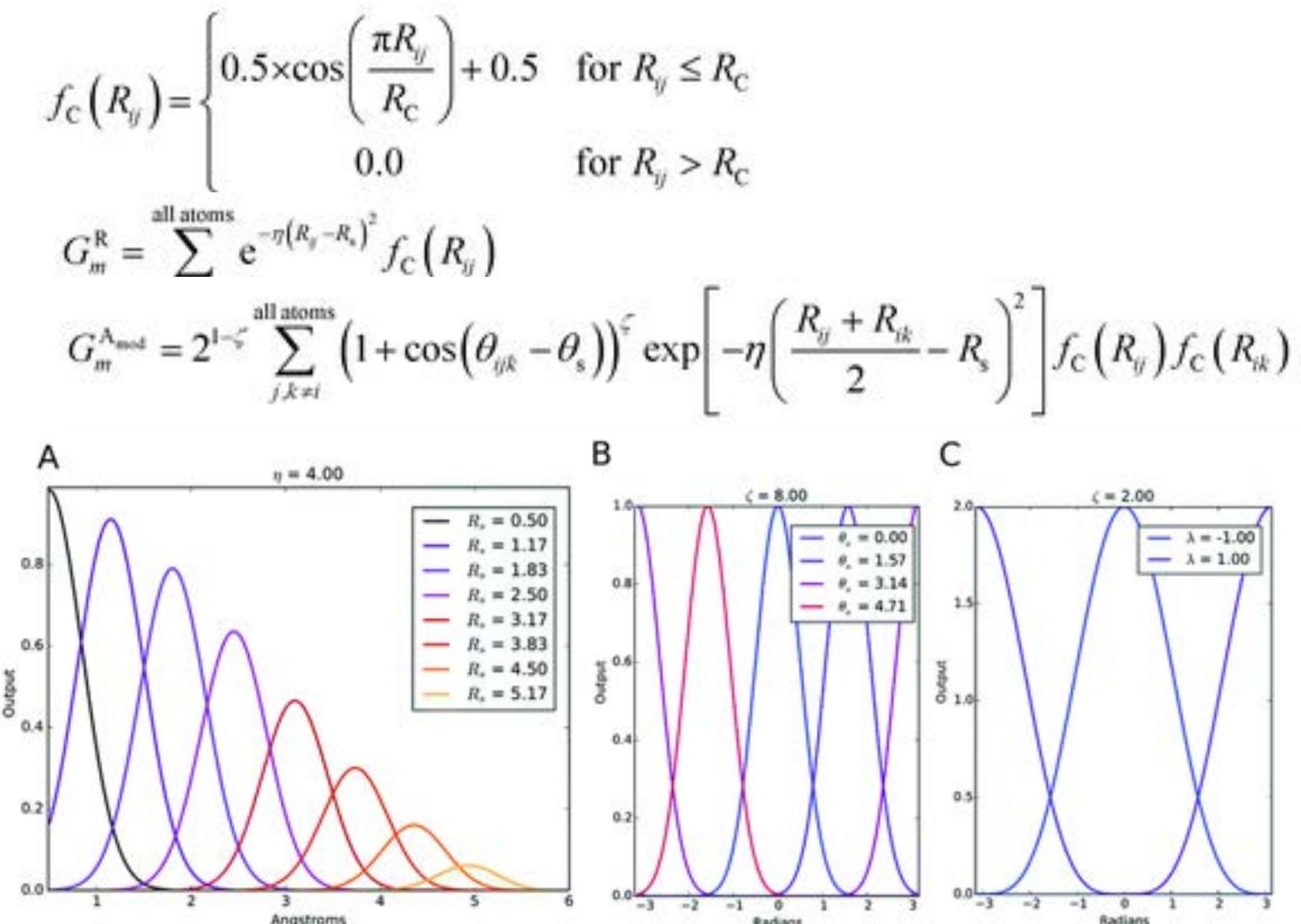
# THERE IS AN MODEL CONTINUUM BETWEEN MOLECULAR MECHANICS AND QUANTUM CHEMISTRY, BUT IS THERE SOMETHING EVEN BETTER?



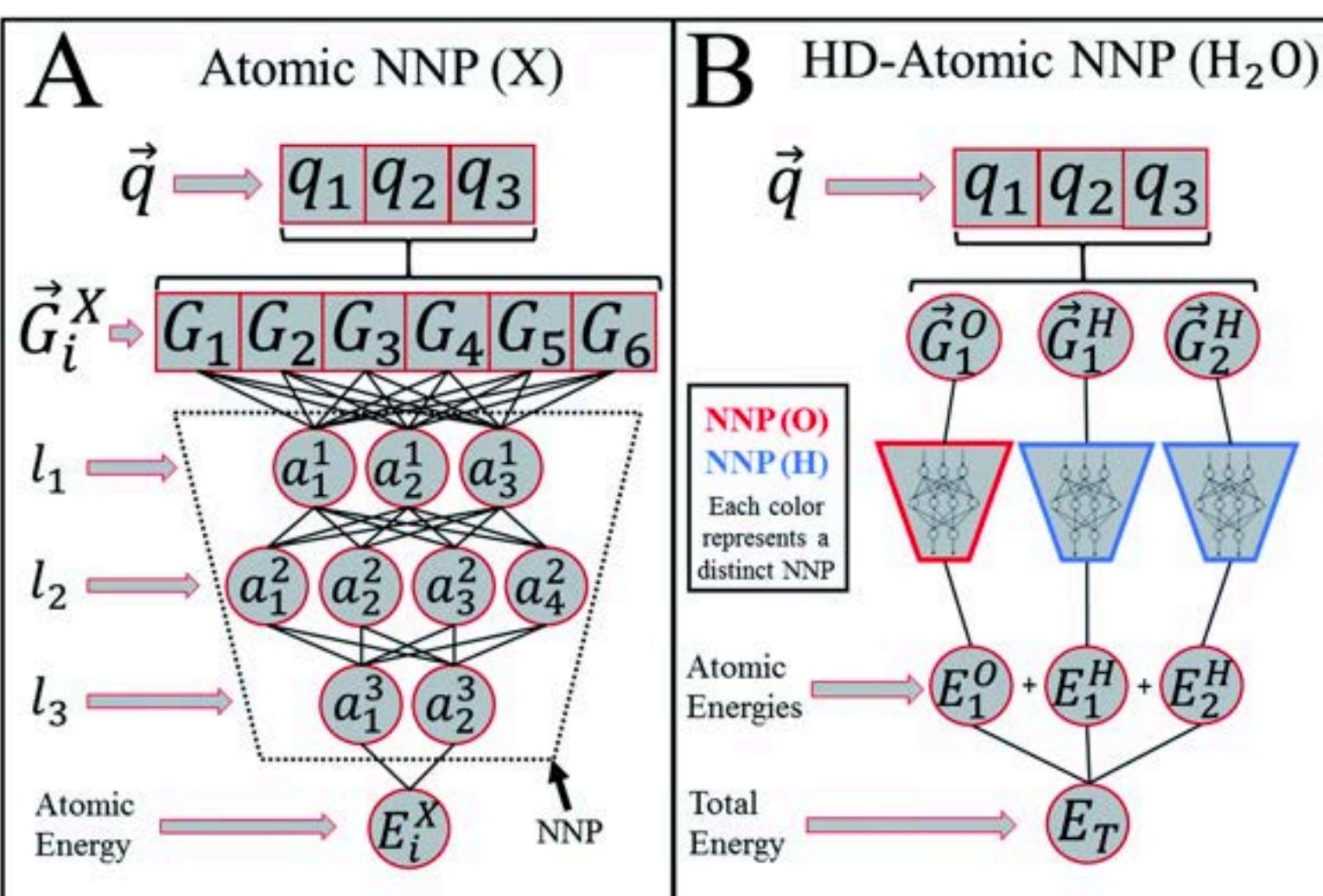
# A NEW GENERATION OF NEURAL NETWORK POTENTIALS PROVIDE SIGNIFICANTLY MORE FLEXIBILITY IN LEARNING MULTIBODY INTERACTIONS, THOUGH AT MUCH GREATER COST

ANI family of neural network potentials

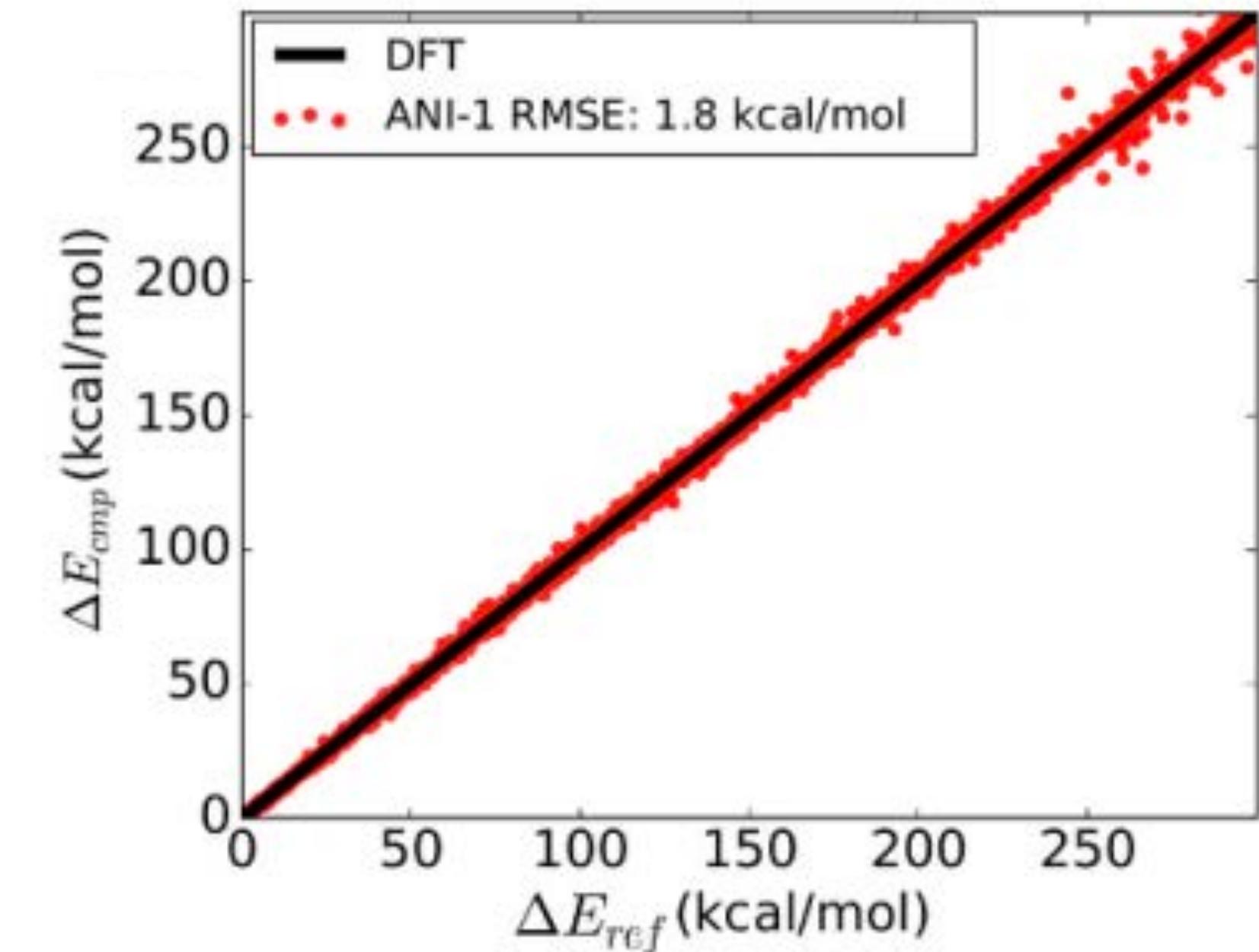
radial and angular features



deep neural network for each atom



excellent agreement with DFT



JUSTIN  
SMITH

OLEXANDR  
ISAYEV

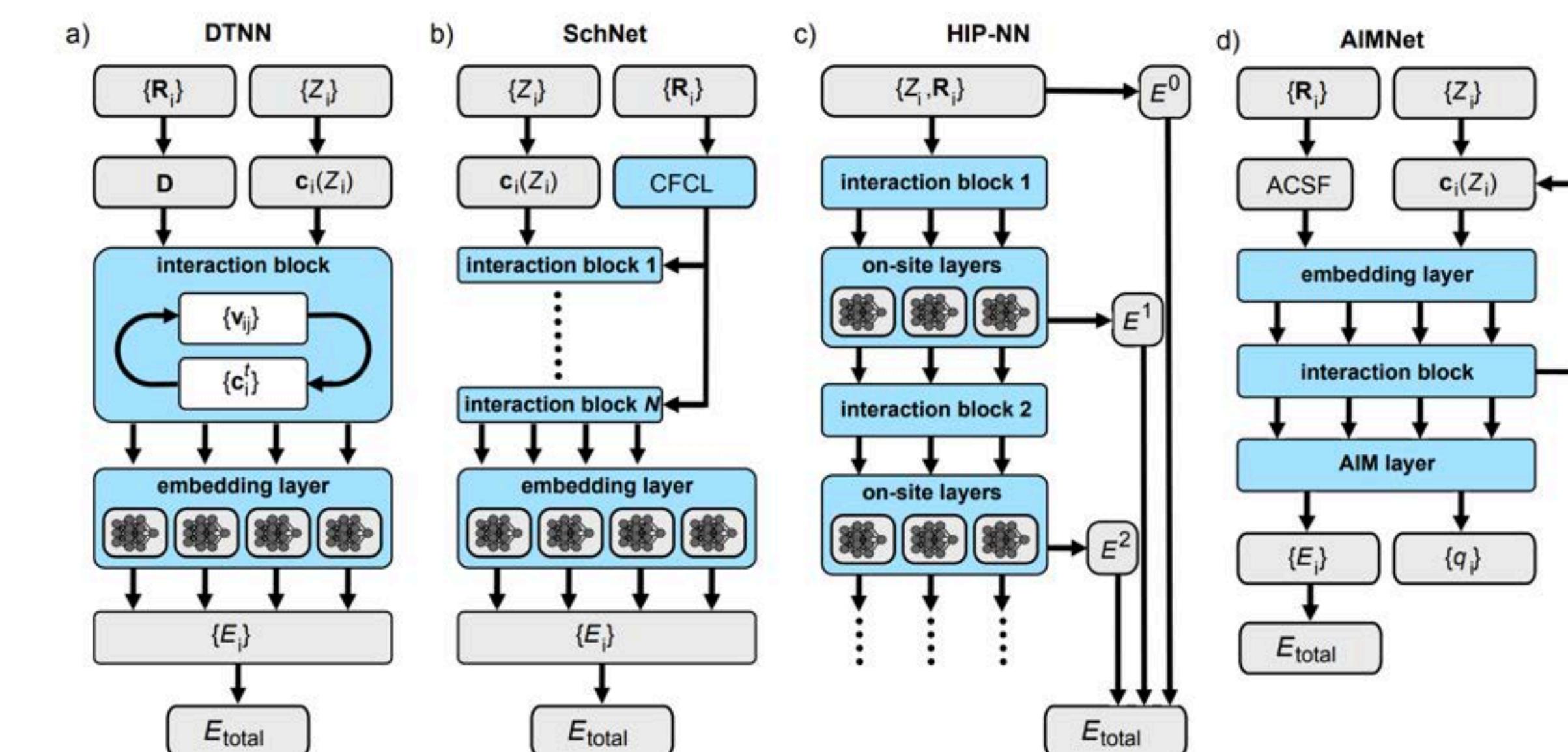
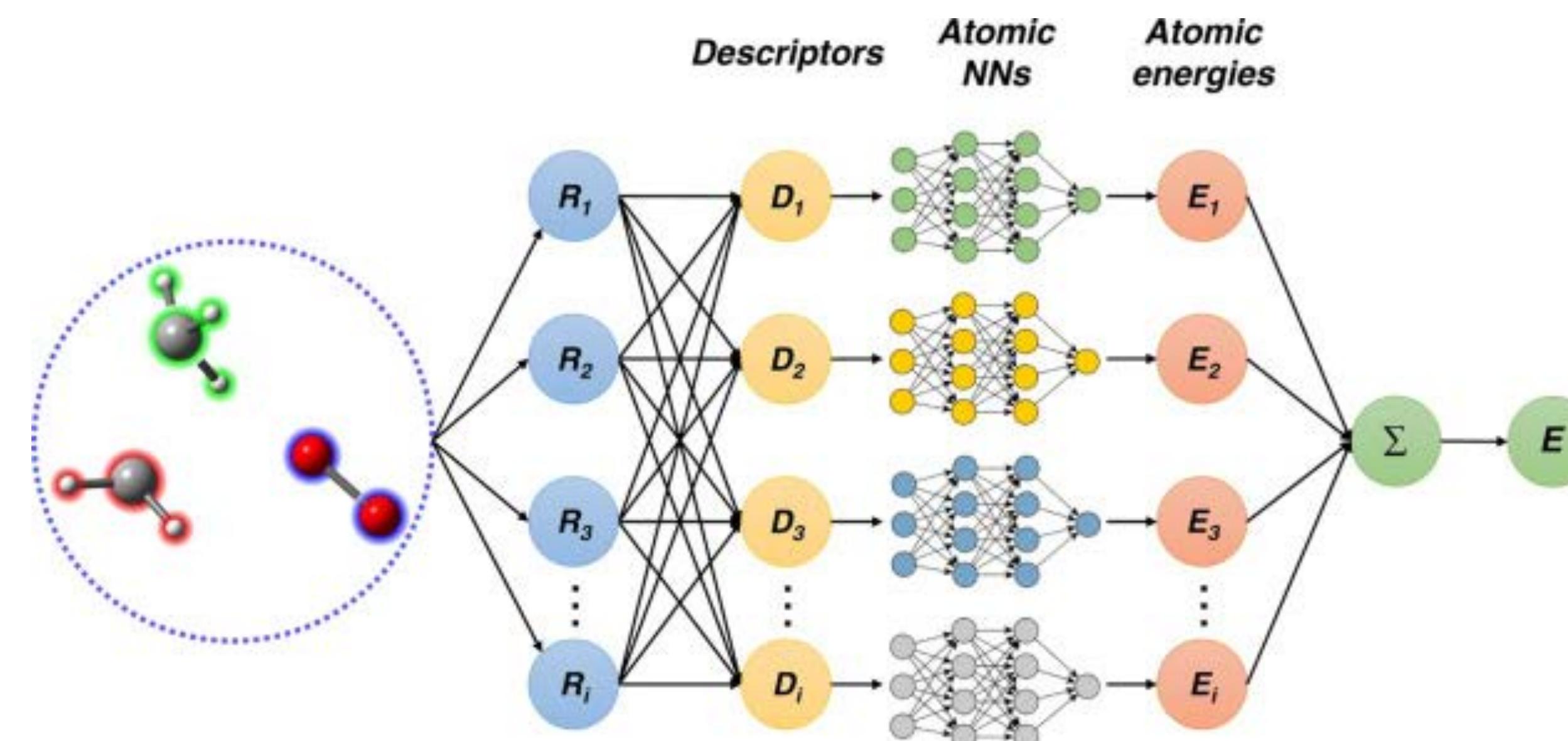
ADRIAN  
ROITBERG



# 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

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



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

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

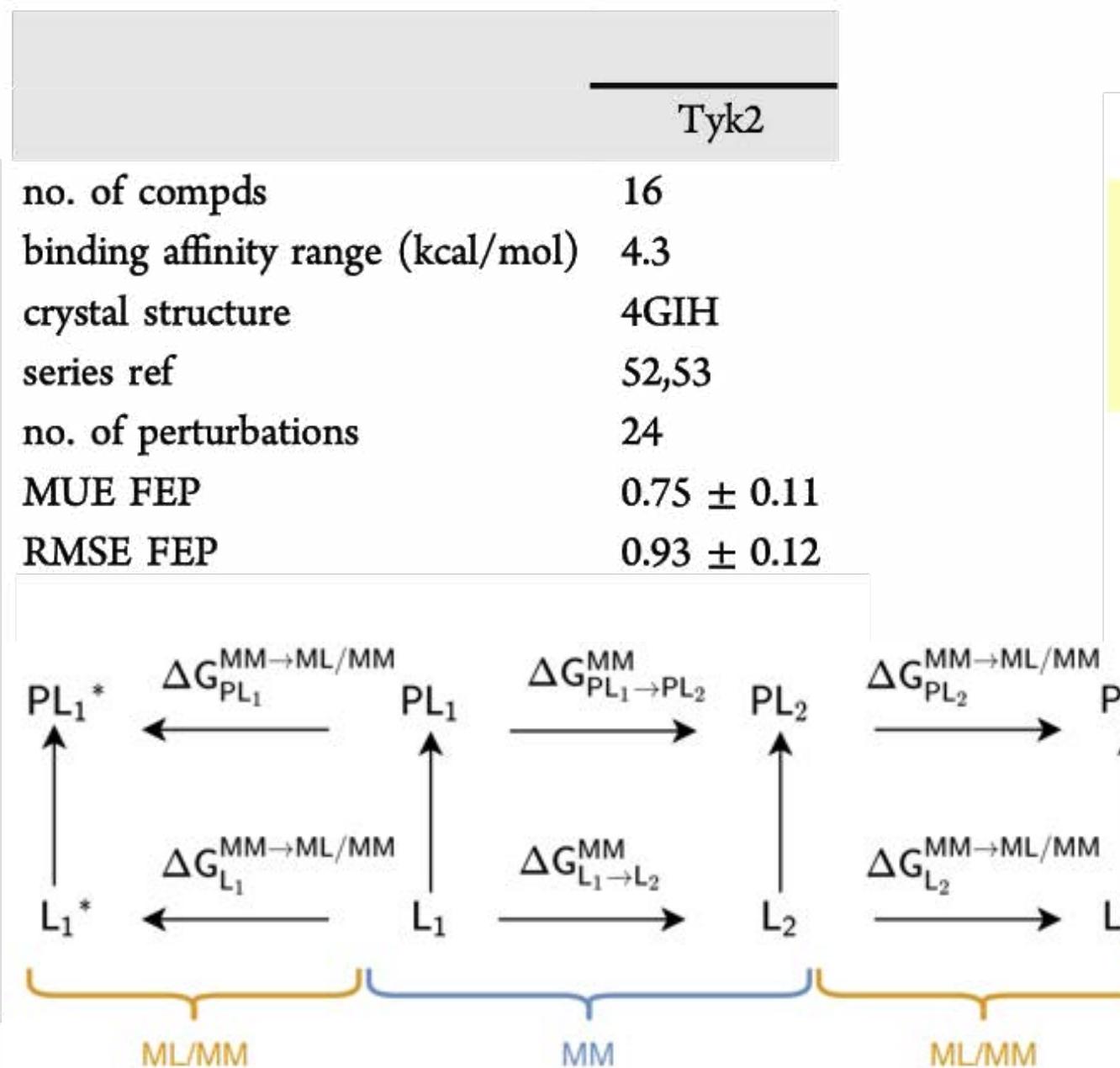
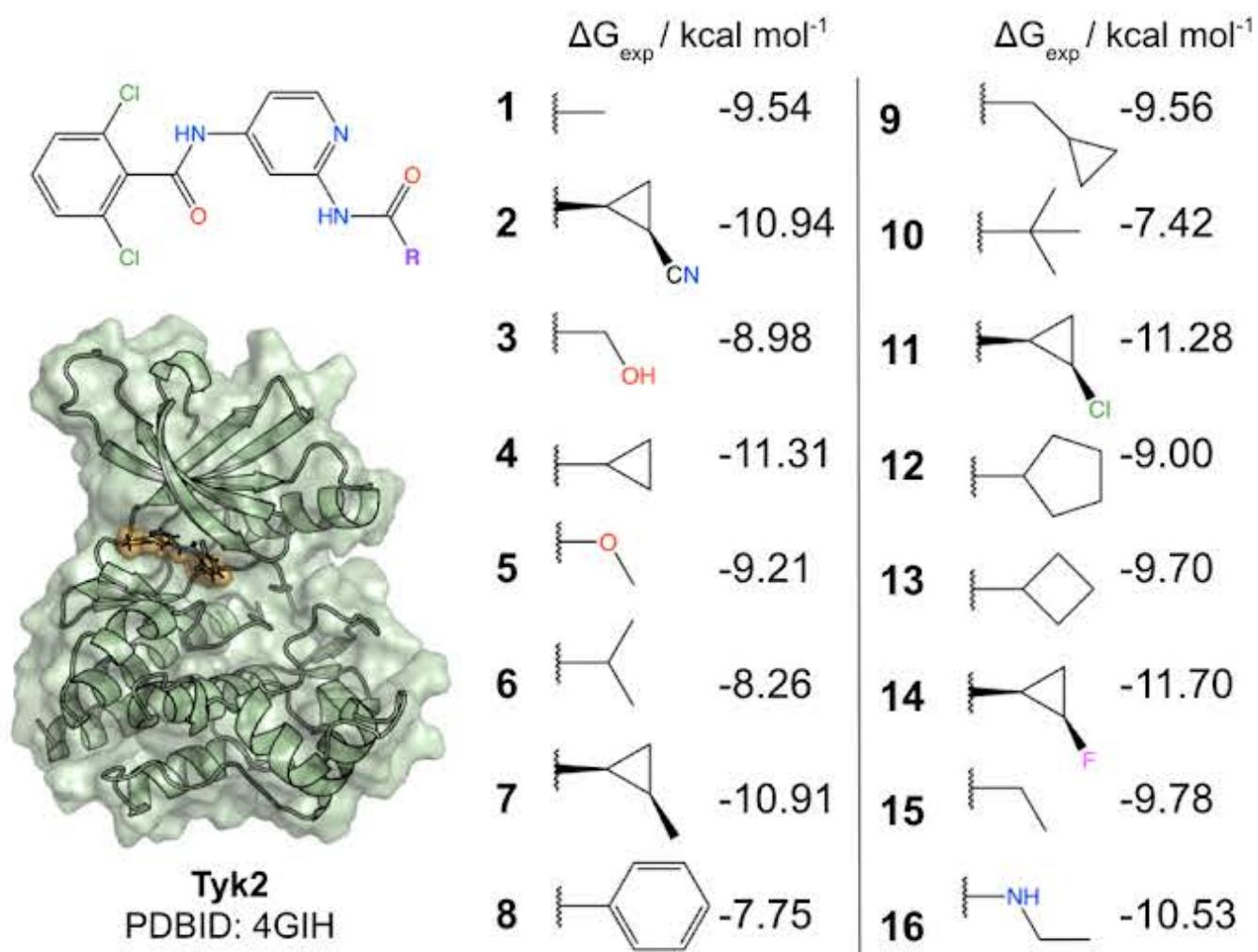
Biomolecular simulation engines that *combine* NNPs (for short-range multibody interactions) and physical energy terms (for long-range interactions and regularization) are likely to be maximally successful in the long term

# 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



Tyk2 benchmark system from Wang et al. JACS 137:2695, 2015

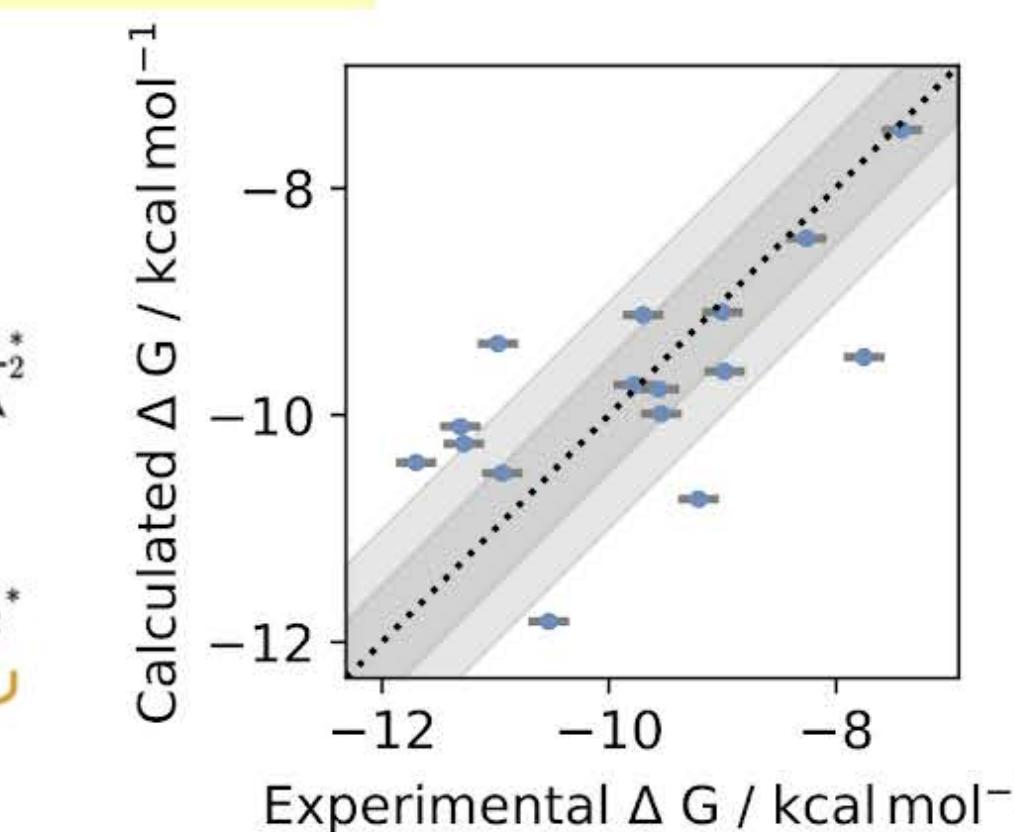
replica-exchange free energy calculations with solute tempering (FEP/REST)

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

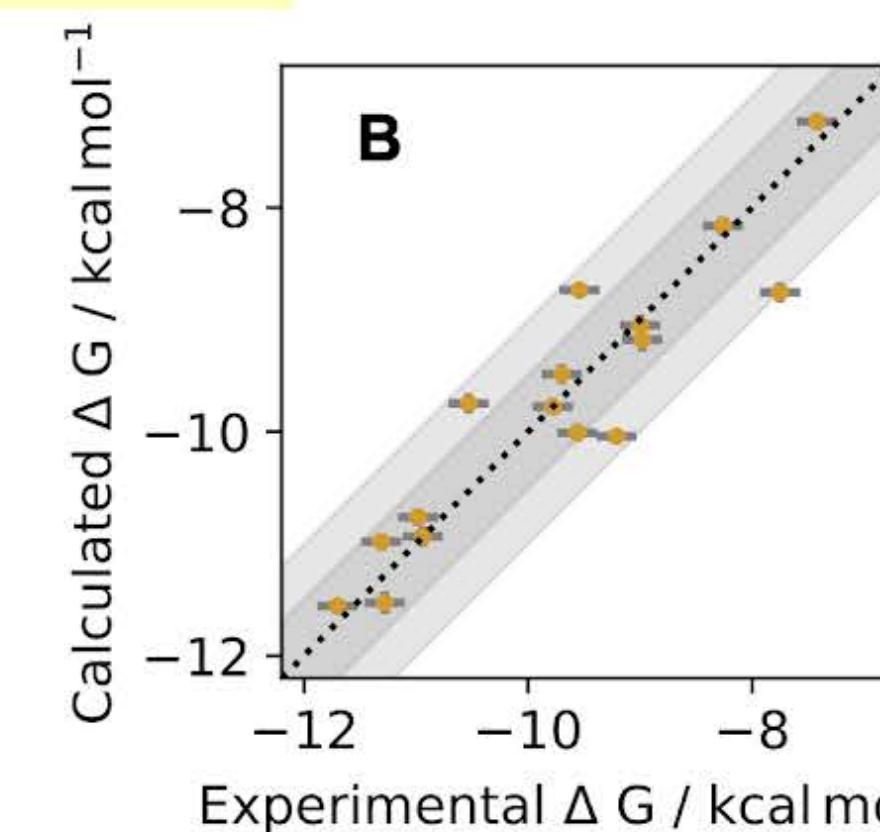
MM: openff-1.0.0  
(N = 16)

RMSE:	0.97	[95%: 0.68, 1.22]
MUE:	0.77	[95%: 0.51, 1.08]
R2:	0.42	[95%: 0.08, 0.75]
rho:	0.65	[95%: 0.25, 0.88]



ML/MM: openff-1.0.0 with ANI2x  
(N = 16)

RMSE:	0.47	[95%: 0.32, 0.68]
MUE:	0.35	[95%: 0.24, 0.56]
R2:	0.86	[95%: 0.66, 0.95]
rho:	0.93	[95%: 0.79, 0.97]

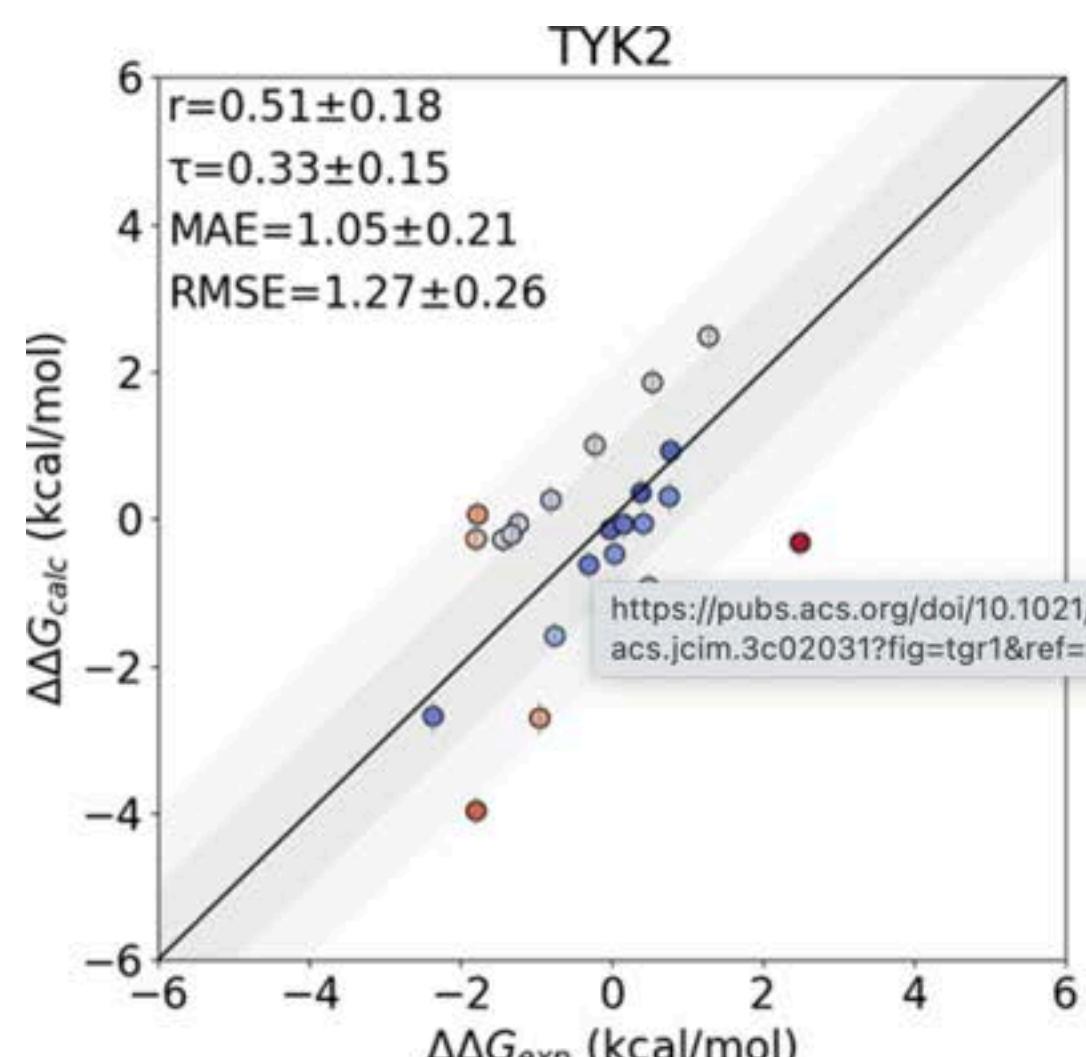


Rufa, Bruce Macdonald, Fass, Wieder, Grinaway, Roitberg, Isayev, and Chodera.

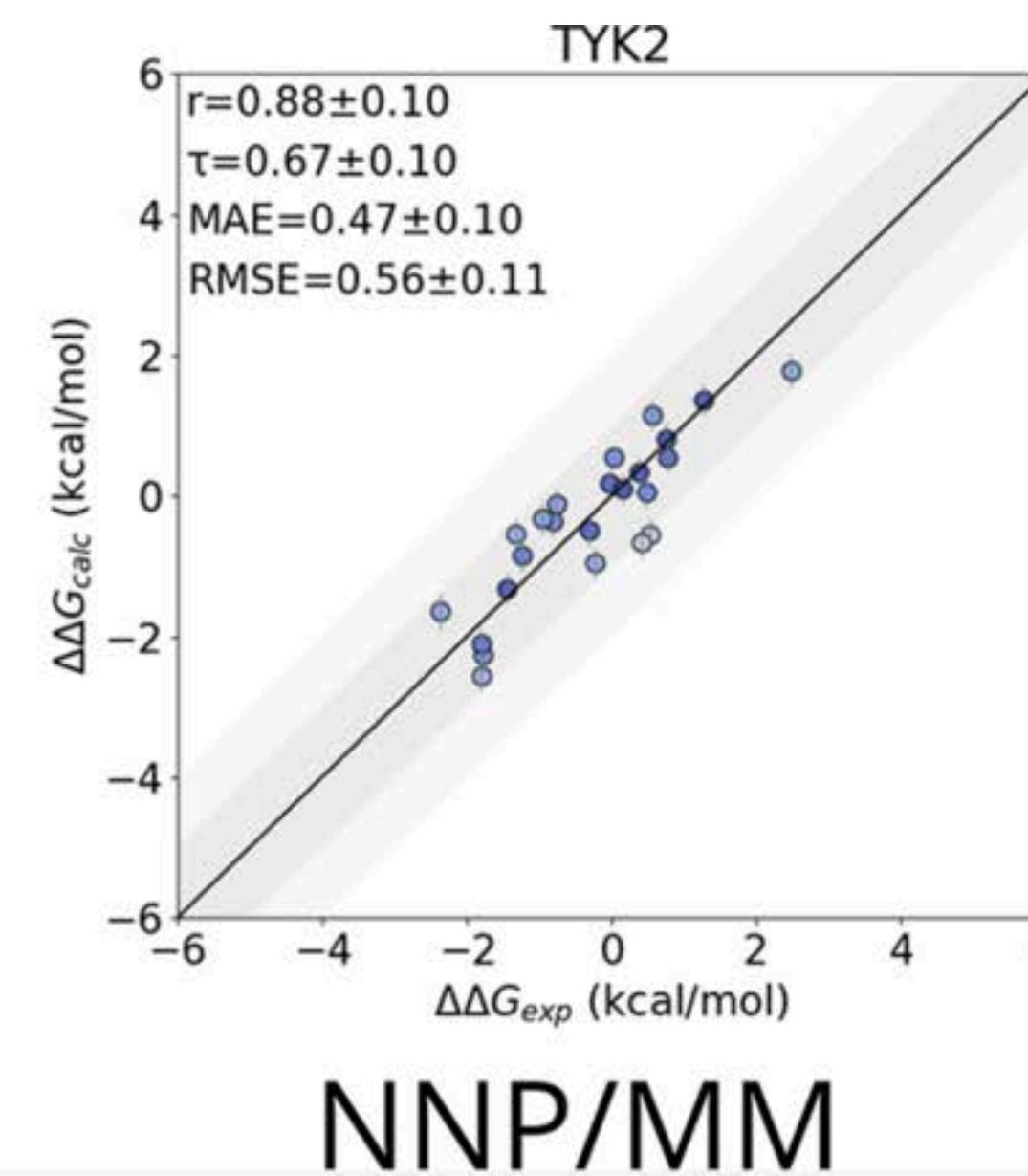
preprint: <https://doi.org/10.1101/2020.07.29.227959>

code: <https://github.com/choderalab/qmlyf>

# HYBRID NNP/MM POTENTIALS SHOW INCREDIBLE PROMISE OVER MM ALONE



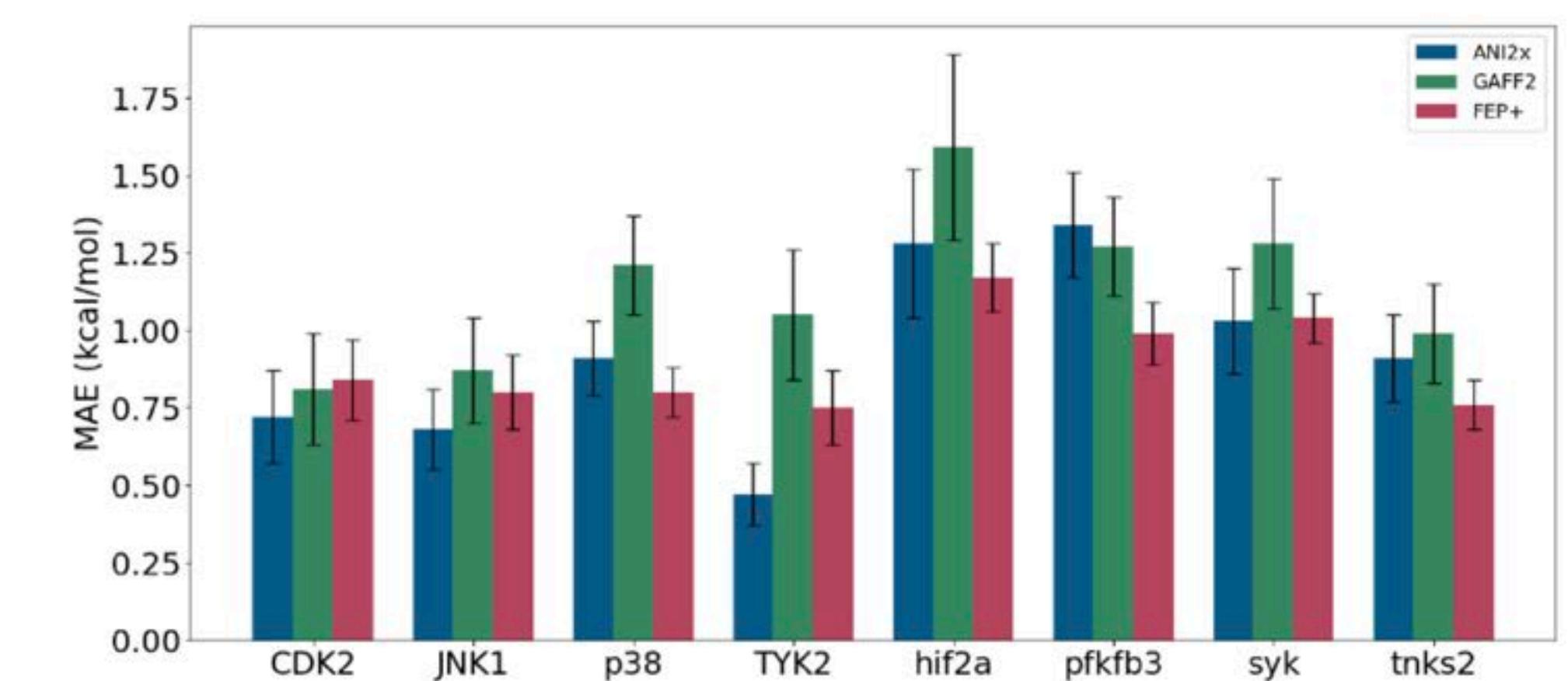
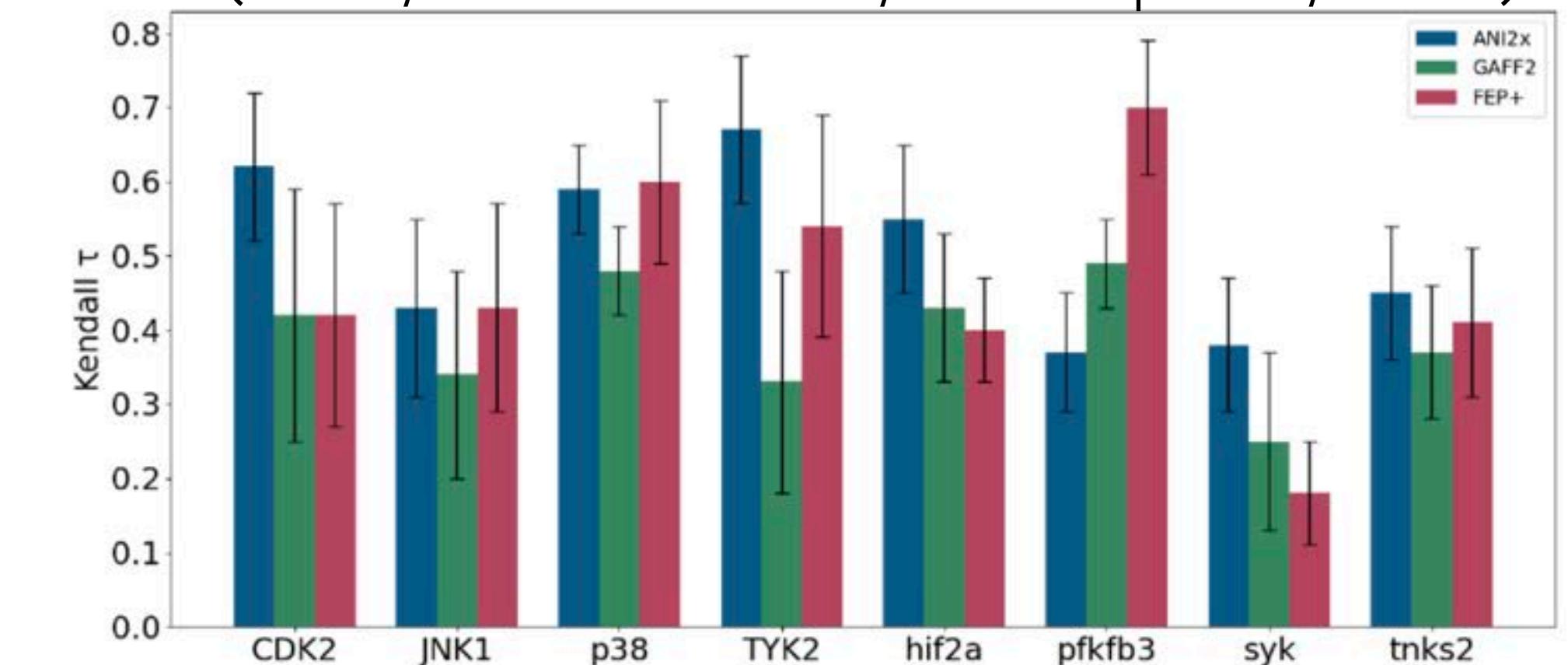
NNP  
for ligand



significantly increased utility compared to GAFF2.11

ANI2x vs GAFF2.11 vs OPLS3e (FEP+)

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



# 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/nnpoops>

- \* 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/nnpoops>

# OPENMM 8 MAKES ML/MM SIMULATIONS INCREDIBLY EASY

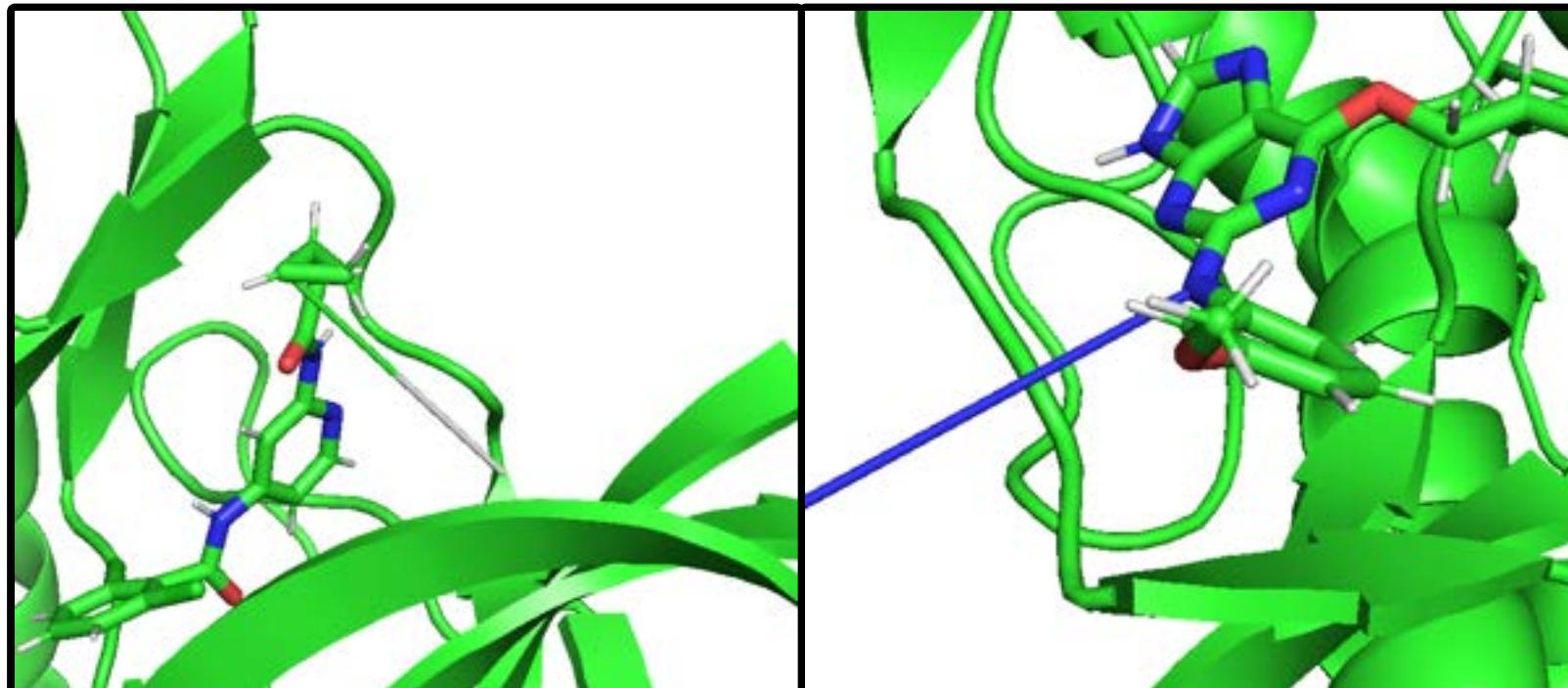
```
conda install -c conda-forge openmm-ml
```

```
# 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)
```

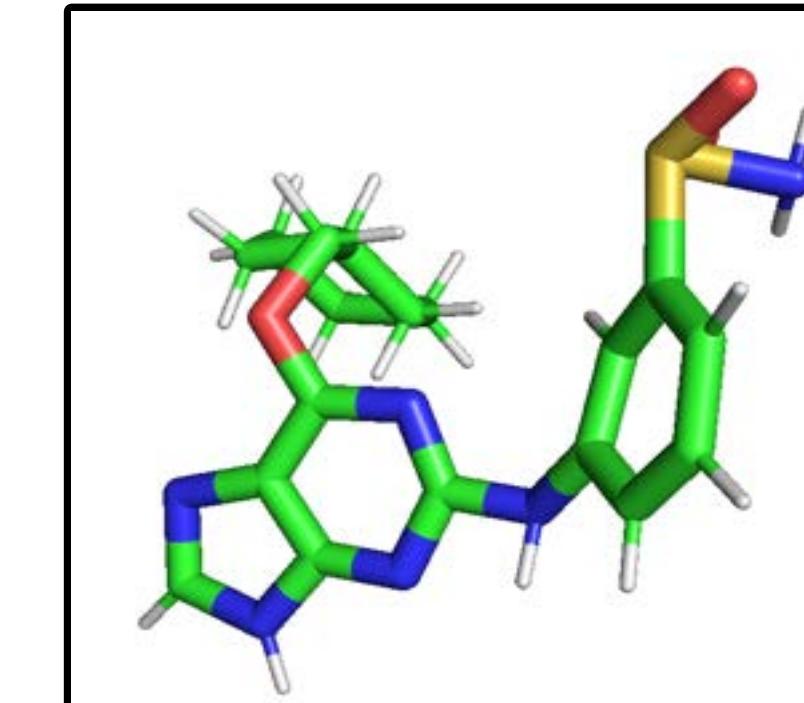
# ML POTENTIALS ARE NOT WITHOUT CHALLENGES. IT'S STILL EARLY DAYS.

~ A gallery of horrors ~

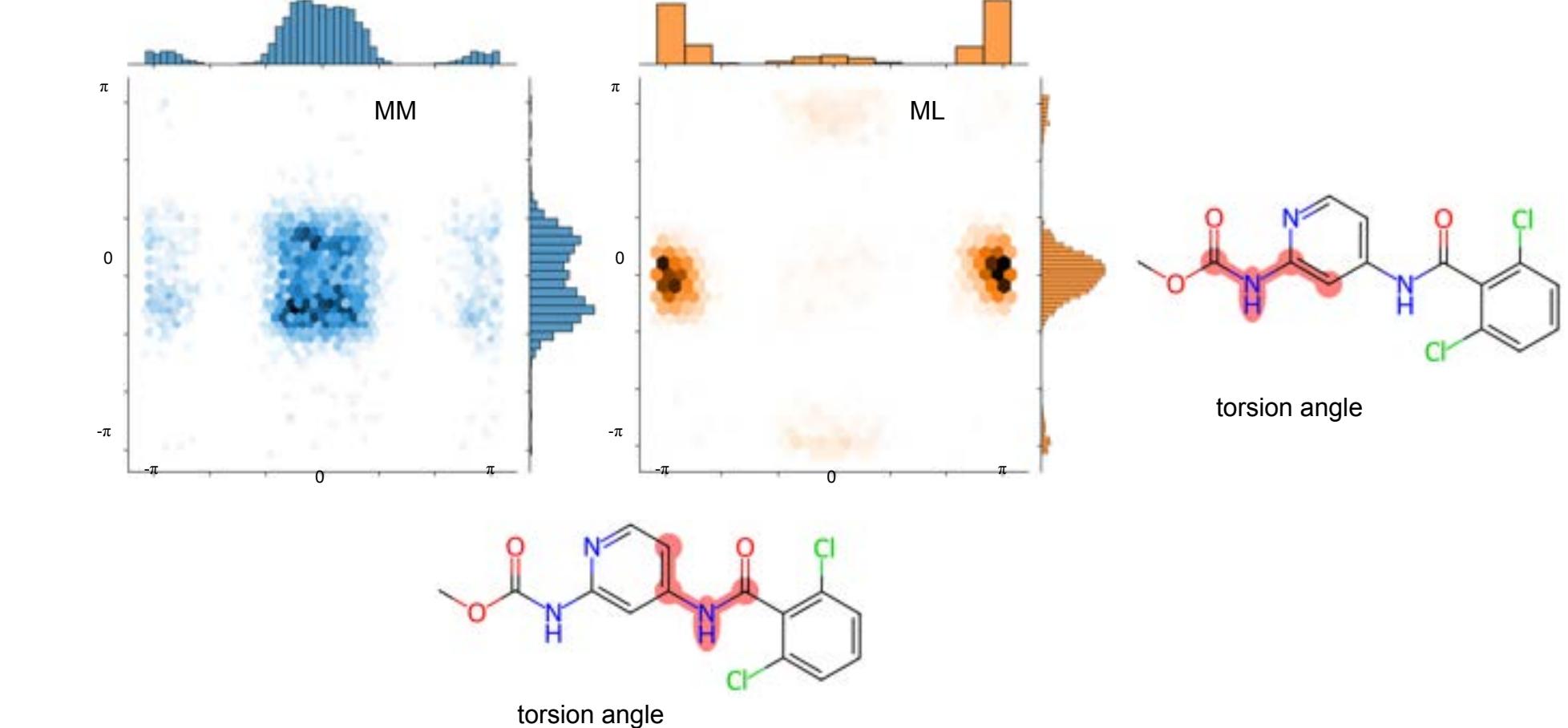
ANI2x proton cannon!



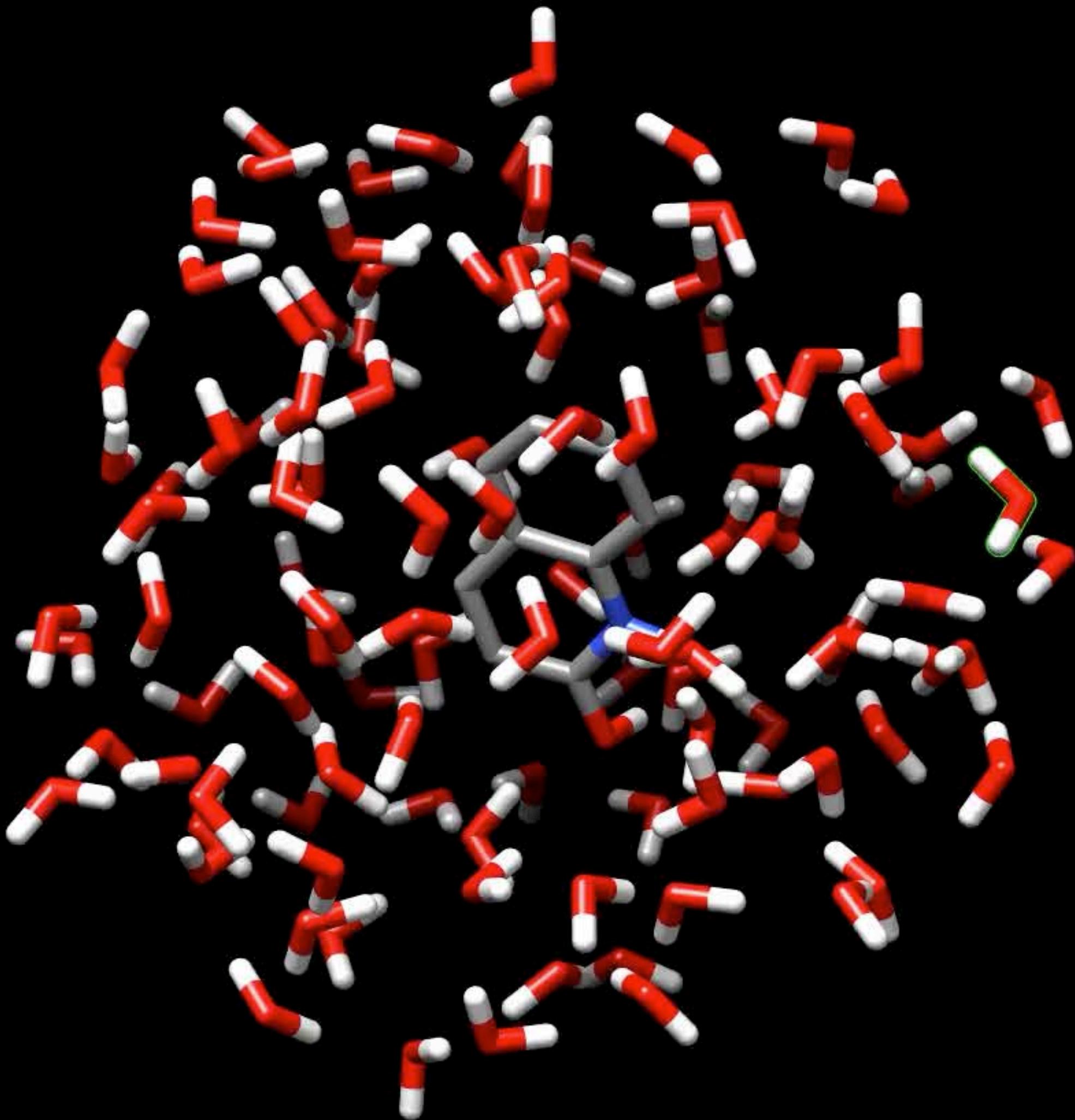
90-degree sulfonamides!



Totally different amide torsions!



# WHY DO WE NEED MM AT ALL?

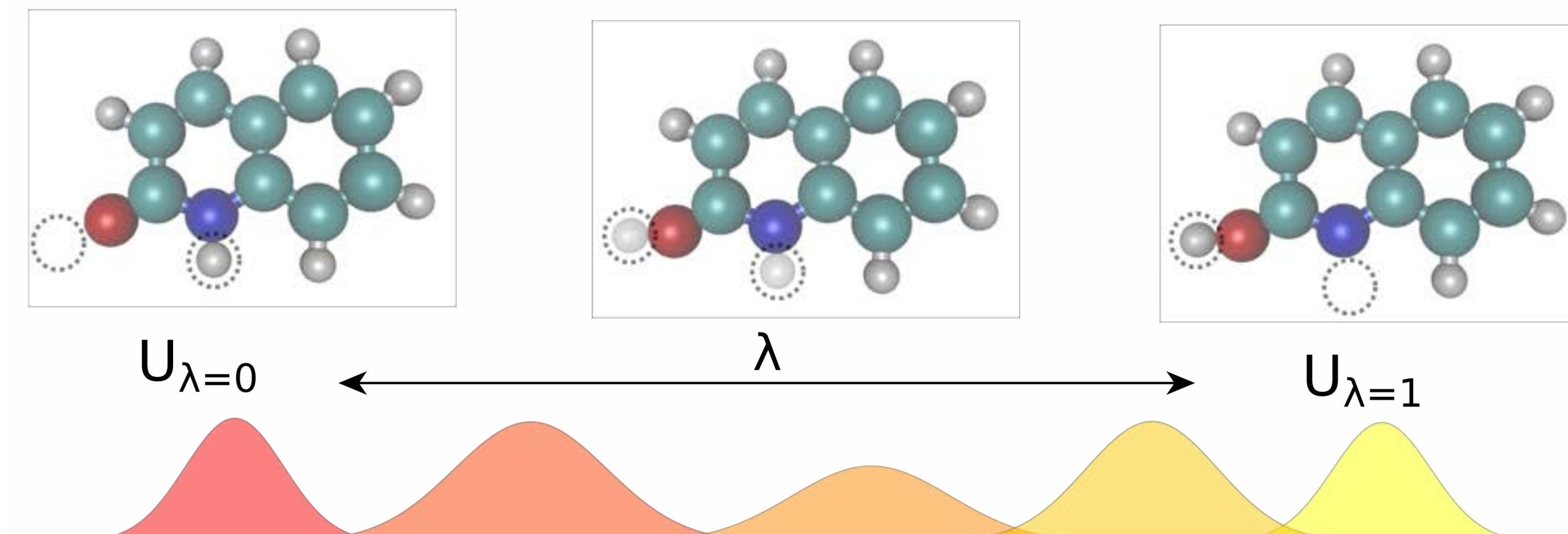


Can we treat everything in the system with NNPs?  
We can finally be free of the hegemony of bonds!

# 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

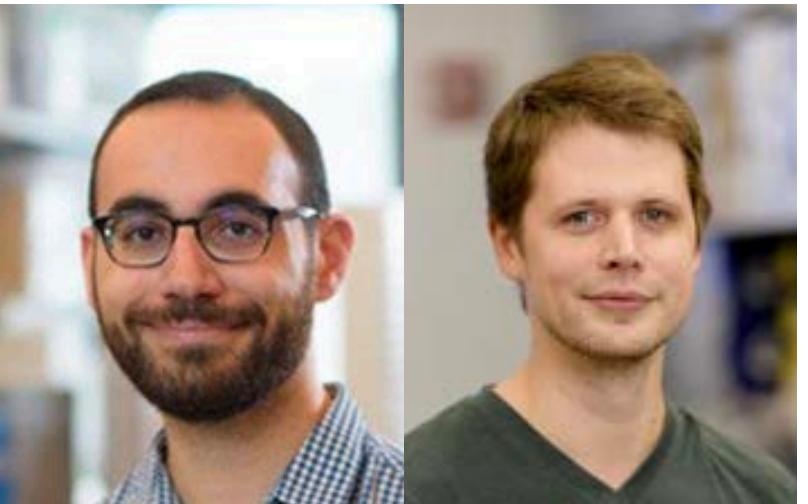
$$U(x;\lambda) = (1-\lambda)U_{\lambda=0}(x) + \lambda U_{\lambda=1}(x)$$



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

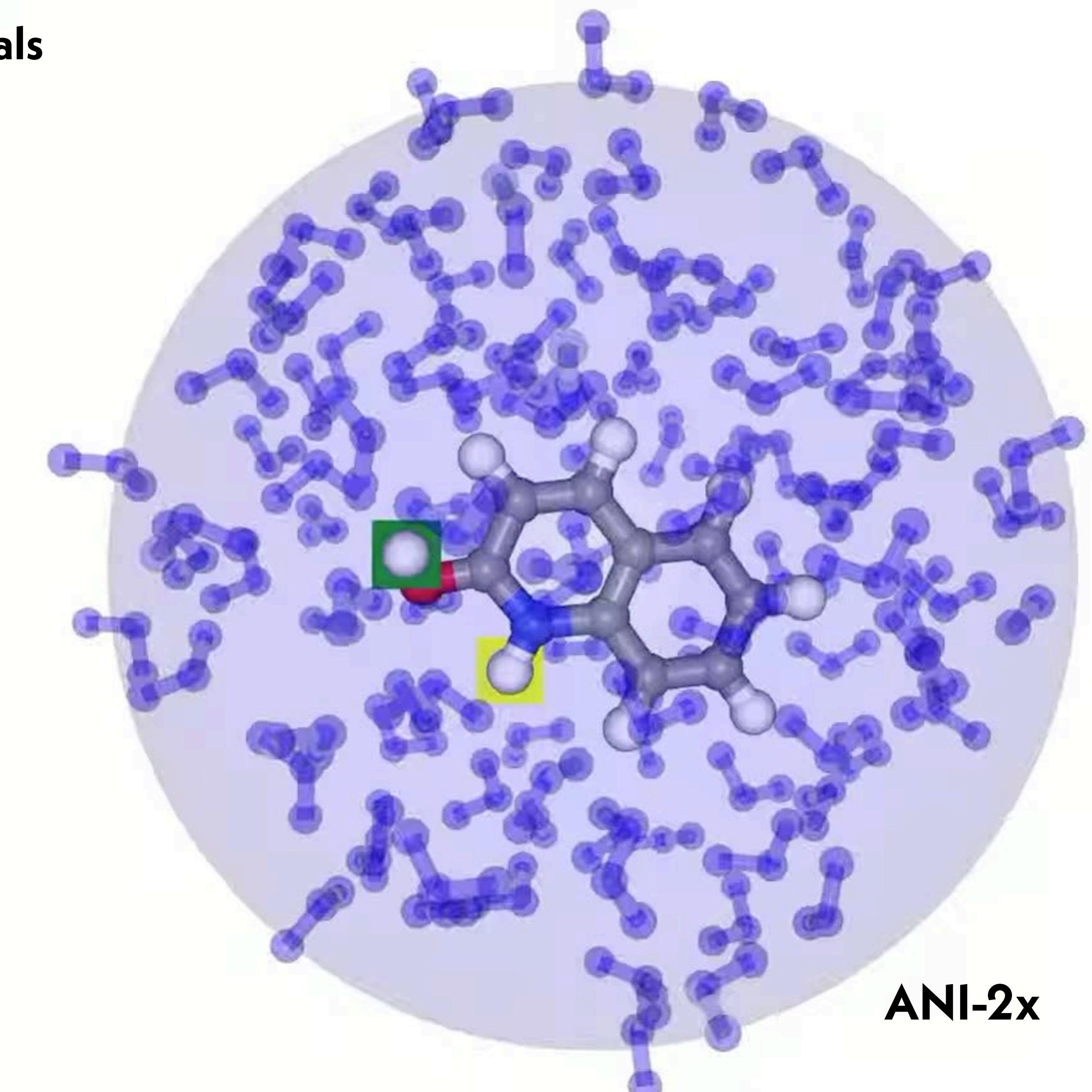
JOSH FASS

MARCUS  
WIEDER



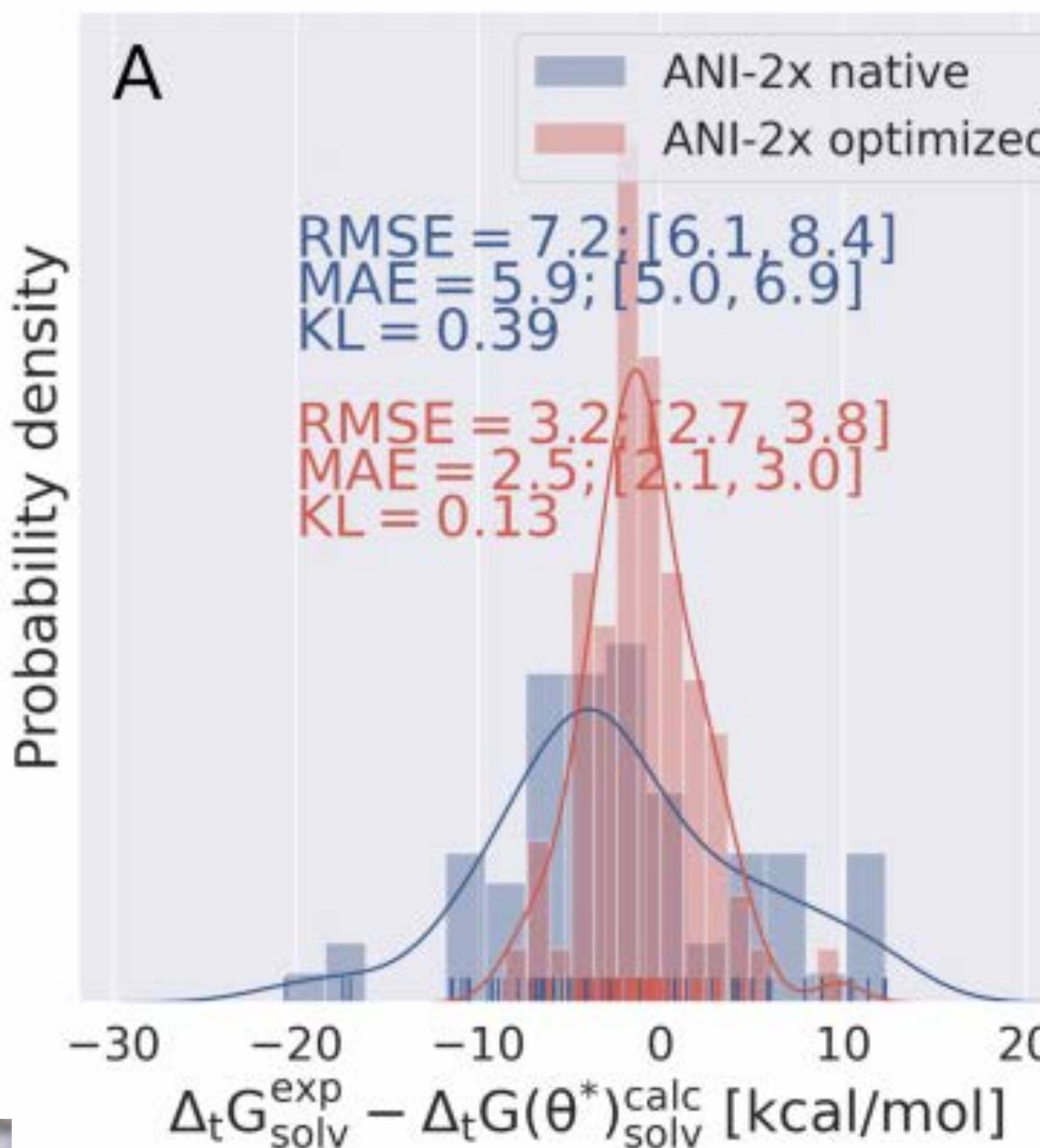
preprint: <https://doi.org/10.1101/2020.10.24.353318>

code: <https://github.com/choderalab/neutromeratio>

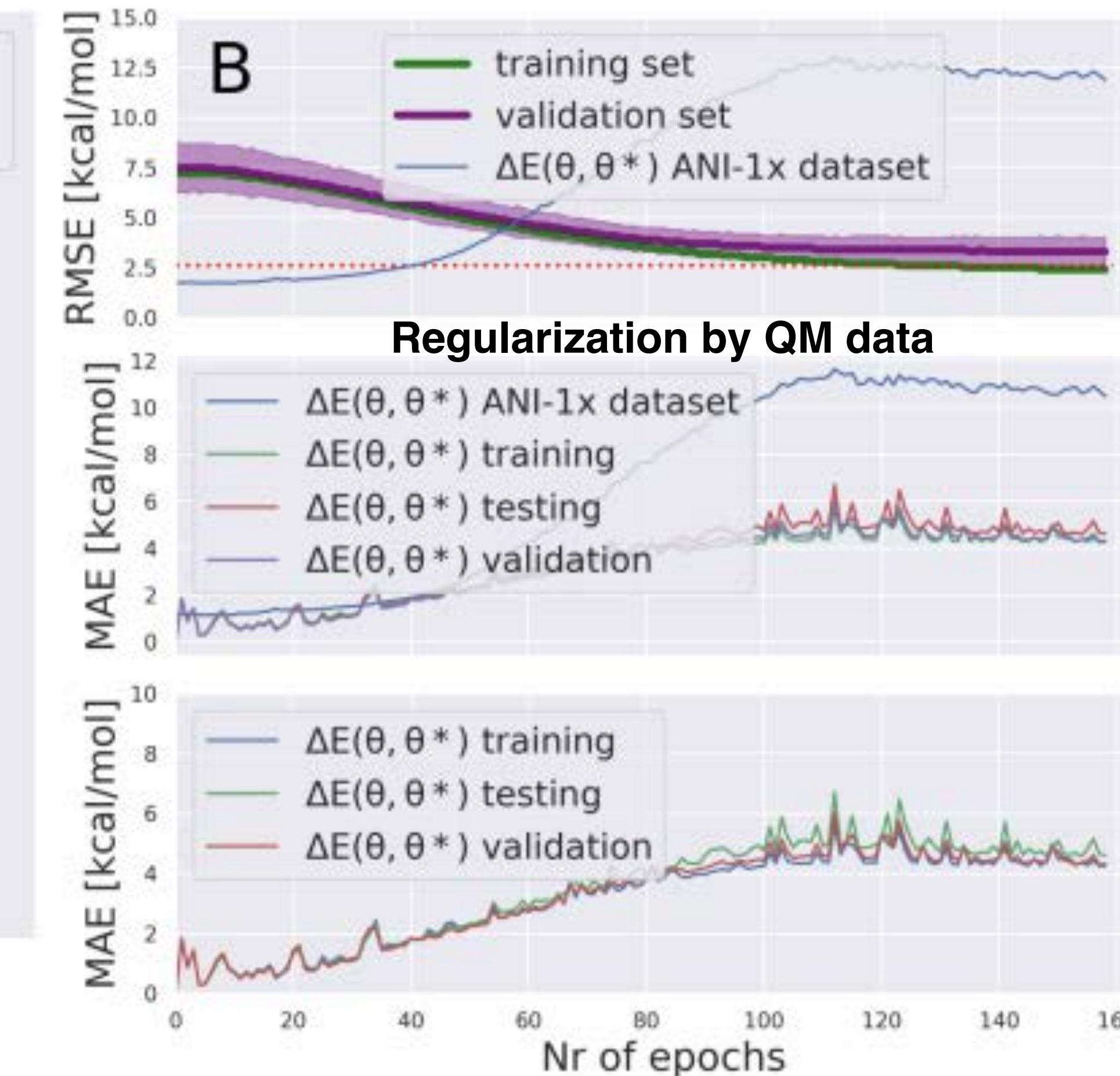


# NEURAL NETWORK POTENTIALS CAN BE FINE-TUNED USING EXPERIMENTAL DATA, REGULARIZED BY QM DATA

test set performance

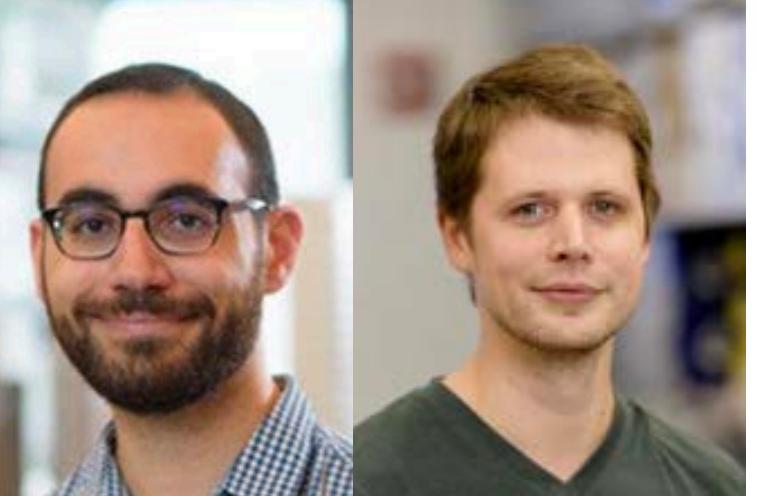


training / validation optimization



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)

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

Text/image datasets

13T  
GPT-4

300B  
GPT-3

5B  
DALL-E 3

650M  
DALL-E 2

Drug discovery datasets

20M  
All ChEMBL  
bioassay data

- 223K  
Complete protein  
data bank (PDB)  
~\$20B
- 2K  
Typical drug  
discovery  
campaign  
~\$12.5M

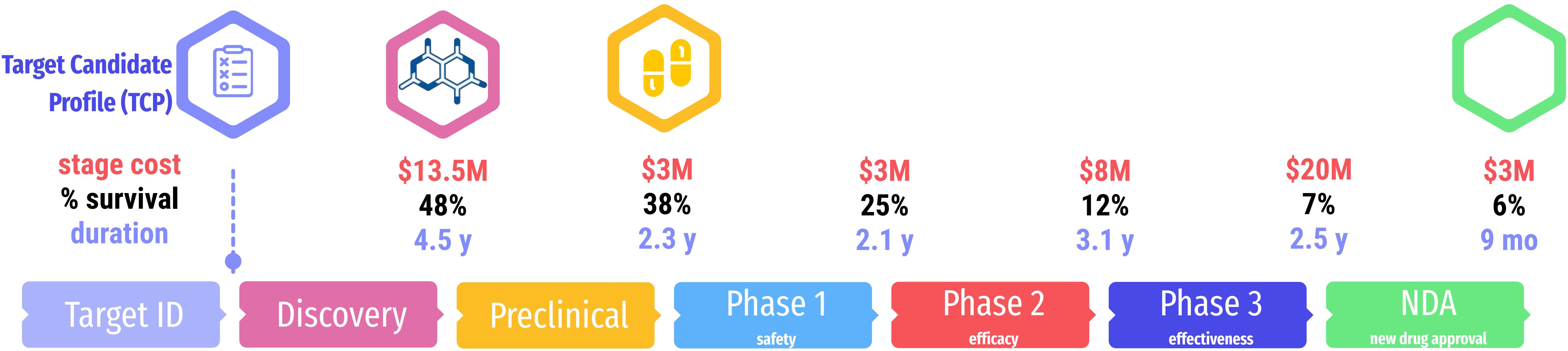
...without asking Sam Altman trying to raise \$6.2T to just do six orders of magnitude more experiments?

# A RISING TIDE LIFTS ALL BOATS

A wide-angle photograph of a lake or river at sunset. The sky is a vibrant orange and yellow, transitioning into a darker blue. In the background, there are several layers of mountains. The water is calm, reflecting the colors of the sky. Numerous small, colorful boats (mostly blue) are scattered across the water, some with people visible. The overall scene conveys a sense of peace and shared opportunity.

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]

**ADMET**  
Absorption  
Distribution  
Metabolism  
Excretion  
Toxicity

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

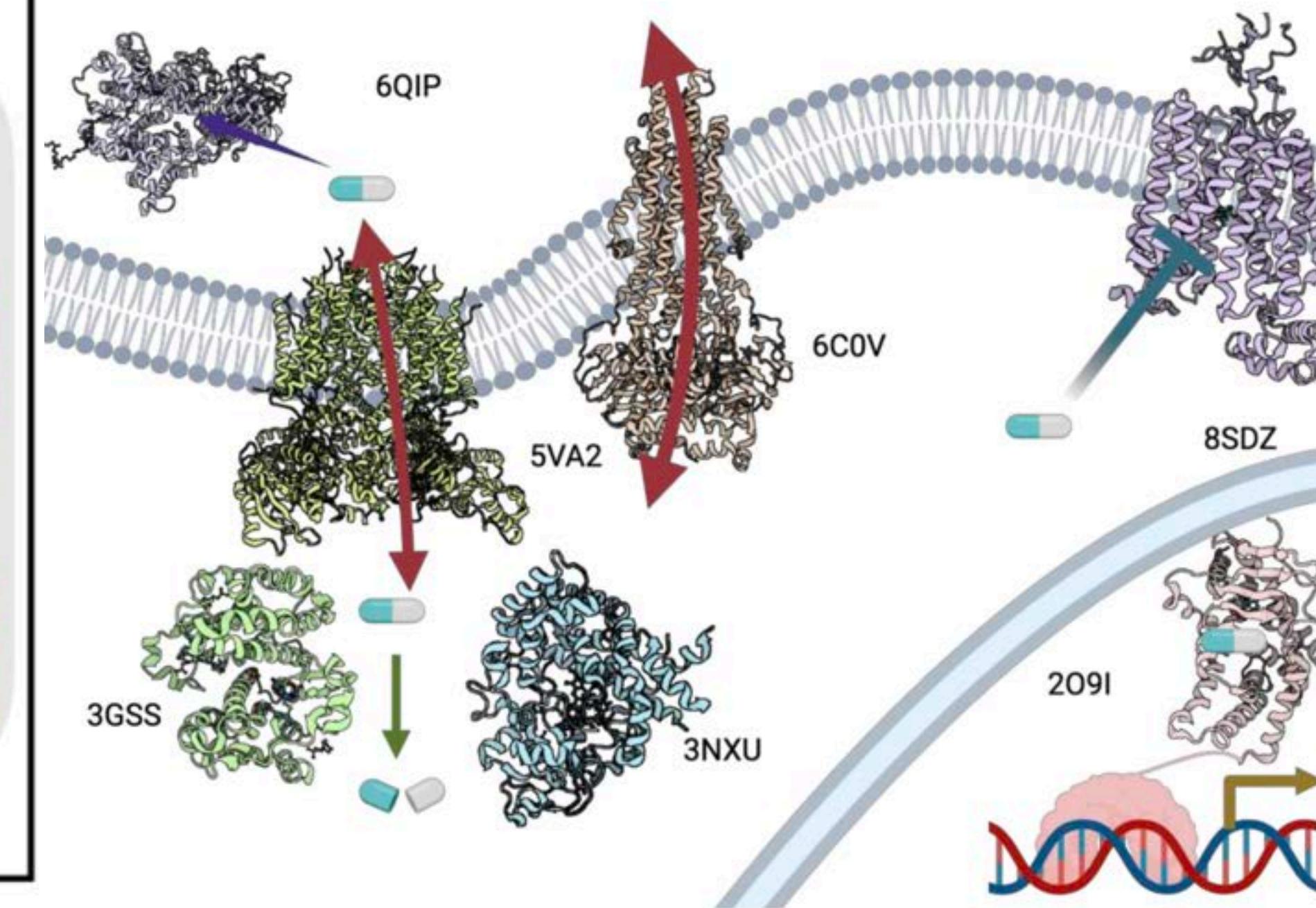
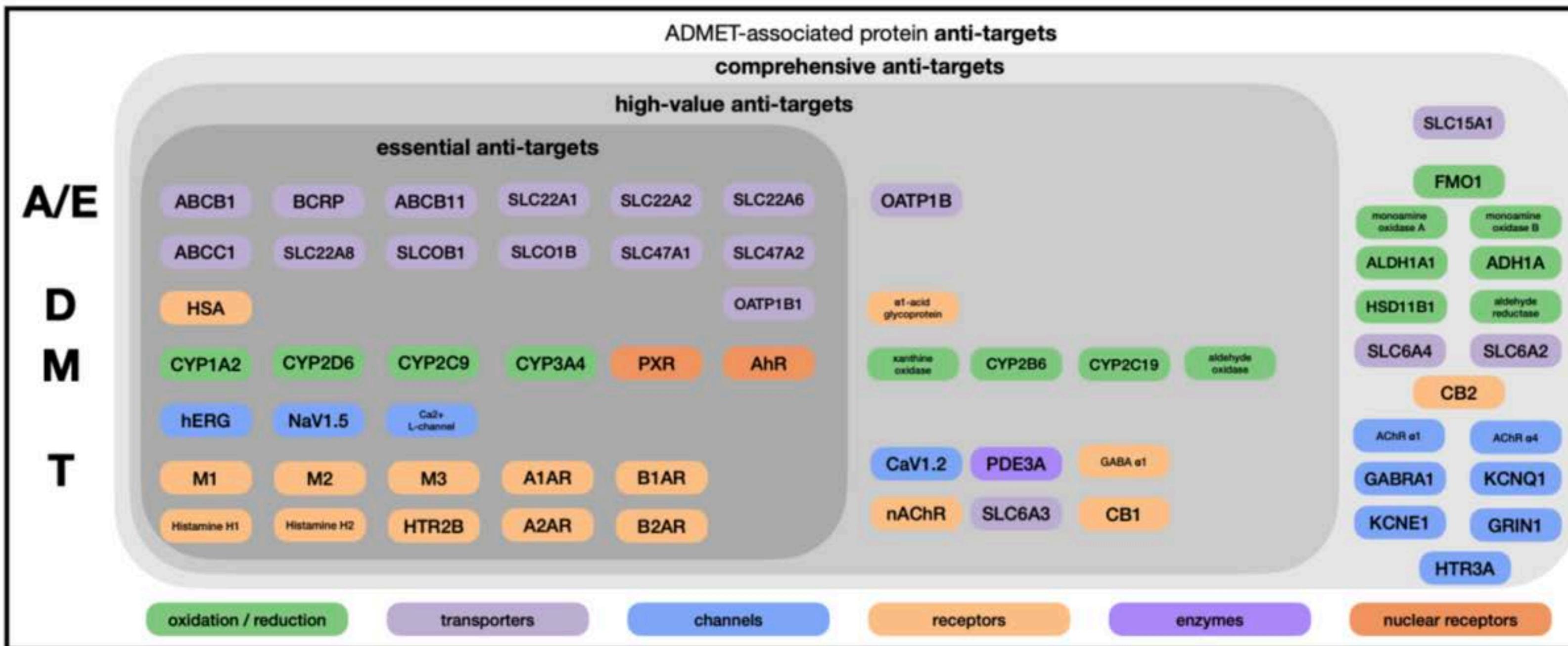
30% of clinical development failures  
due to human safety and toxicology [3]

[1] doi:10.1016/j.drudis.2015.03.010

[2] doi:10.1016/j.drudis.2013.11.014

[3] doi:10.1038/nrd1470

# 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

Commentary  
Enabling structure-based drug discovery utilizing predicted models

CellPress

Edward B. Miller,<sup>1,\*</sup> Howook Hwang,<sup>1</sup> Mee Shelley,<sup>2</sup> Andrew Placzek,<sup>2</sup> João P.G.L.M. Rodrigues,<sup>1</sup> Robert K. Suto,<sup>3</sup> Lingle Wang,<sup>1</sup> Karen Akinsanya,<sup>1</sup> and Robert Abel<sup>1</sup>

<sup>1</sup>Schrödinger New York, 1540 Broadway, 24th Floor, New York, NY 10036, USA

<sup>2</sup>Schrödinger Portland, 101 SW Main Street, Suite 1200, Portland, OR 97204, USA

# CHODERA LAB

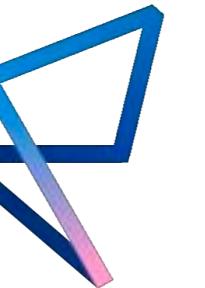


National Institutes  
of Health



OpenEye  
SCIENTIFIC

OpenEye  
SCIENTIFIC  
SILICON  
Therapeutics  
PARKER INSTITUTE  
for CANCER IMMUNOTHERAPY  
STIFTUNG CHARITÉ



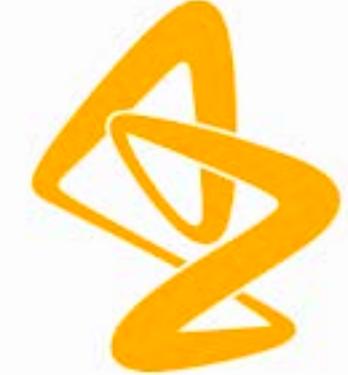
Gerstner  
FAMILY FOUNDATION  
STARR CANCER  
CONSORTIUM



Open Force Field  
Consortium

XtalPi

>>> CYCLE  
FOR SURVIVAL



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

