LOOKING TOWARD THE FUTURE OF ASSESSMENT OF PREDICTIVE MODELING AND BLINDED PREDICTION CHALLENGES

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

DISCLOSURES:
Scientific Advisory Board, OpenEye Scientific
All opinions/views are my own.

SAMPL/D3R Workshop - 23 Aug 2019 - La Jolla, CA

WHAT ARE OUR GOALS?

Predict how well methods **work in practice**

Characterize **domain of applicability** and understand challenges that limit it

Iteratively **refine best practices** for applications of interest

Expand the domain of applicability

Minimize human **effort** in evaluations, maximize learning

Getting non-experimental labs the data they **need**

WHAT ARE OUR GOALS?

Predict how well methods **work in practice**

Characterize **domain of applicability** and understand challenges that limit it

Iteratively **refine best practices** for applications of interest

Expand the domain of applicability

Minimize human **effort** in evaluations, maximize learning

Getting non-experimental labs the data they **need**

How well are we doing, and what could we improve upon?

OUR FIELD FACES
SIGNIFICANT CHALLENGES

INTEROPERABILITY

Current software communities are **balkanized**

Poor (or no) standards for moving data between codes/packages

If there *was* a good standard, developers would adhere to it

(where **good** = it made our lives **easier**, not harder)

EVALUATION

Comparison of predictive modeling on retrospective data hindered by **lack of standard datasets** and **absence of common benchmark framework** (machine learning has MNIST, ImageNet, etc.)

Predictive challenges (e.g., SAMPL, D3R) often end up **testing unrelated choices** (such as biomolecular setup pipeline), not the scientific core code

WHAT ARE WE EVALUATING IN BLIND COMPETITIONS?

evaluating the driver

evaluating the technology

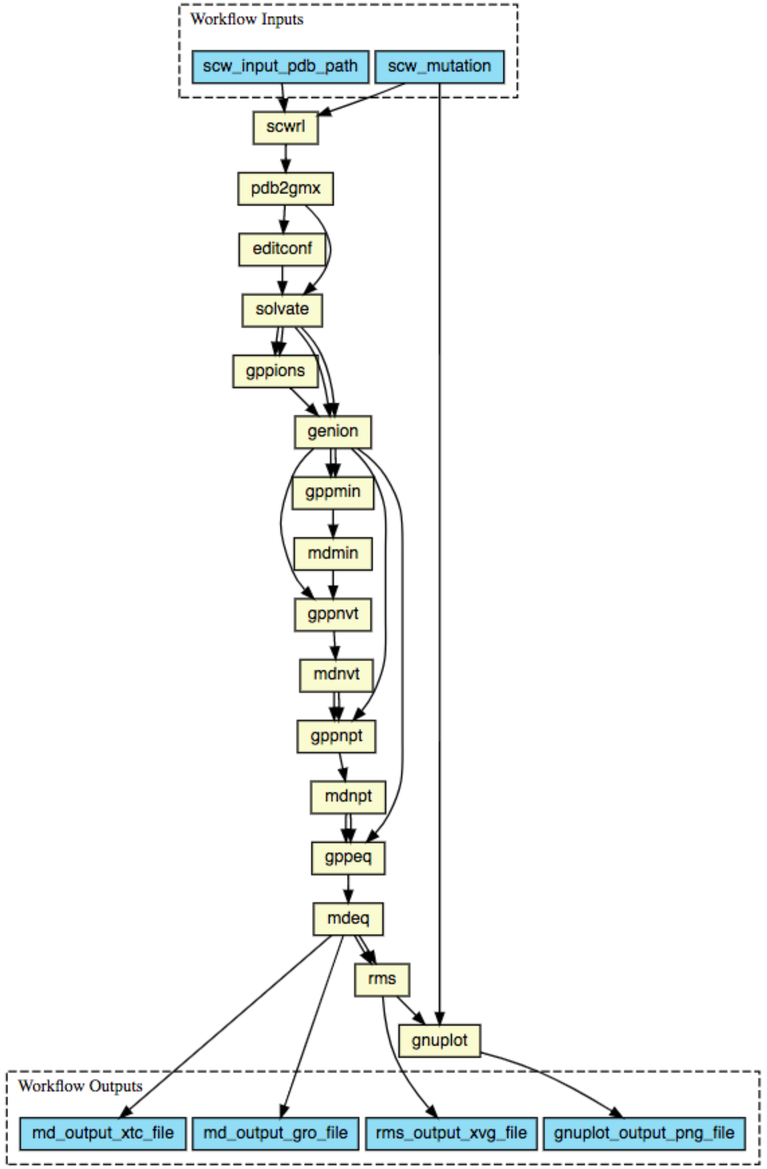
Need to separate capabilities of technology from skill of driver

WE WANT TO FOCUS ON KEY SCIENCE

Scientists want to **focus creative efforts on a specific part of the process**, but are often forced to build everything from scratch to have a working framework in which they can carry out productive research

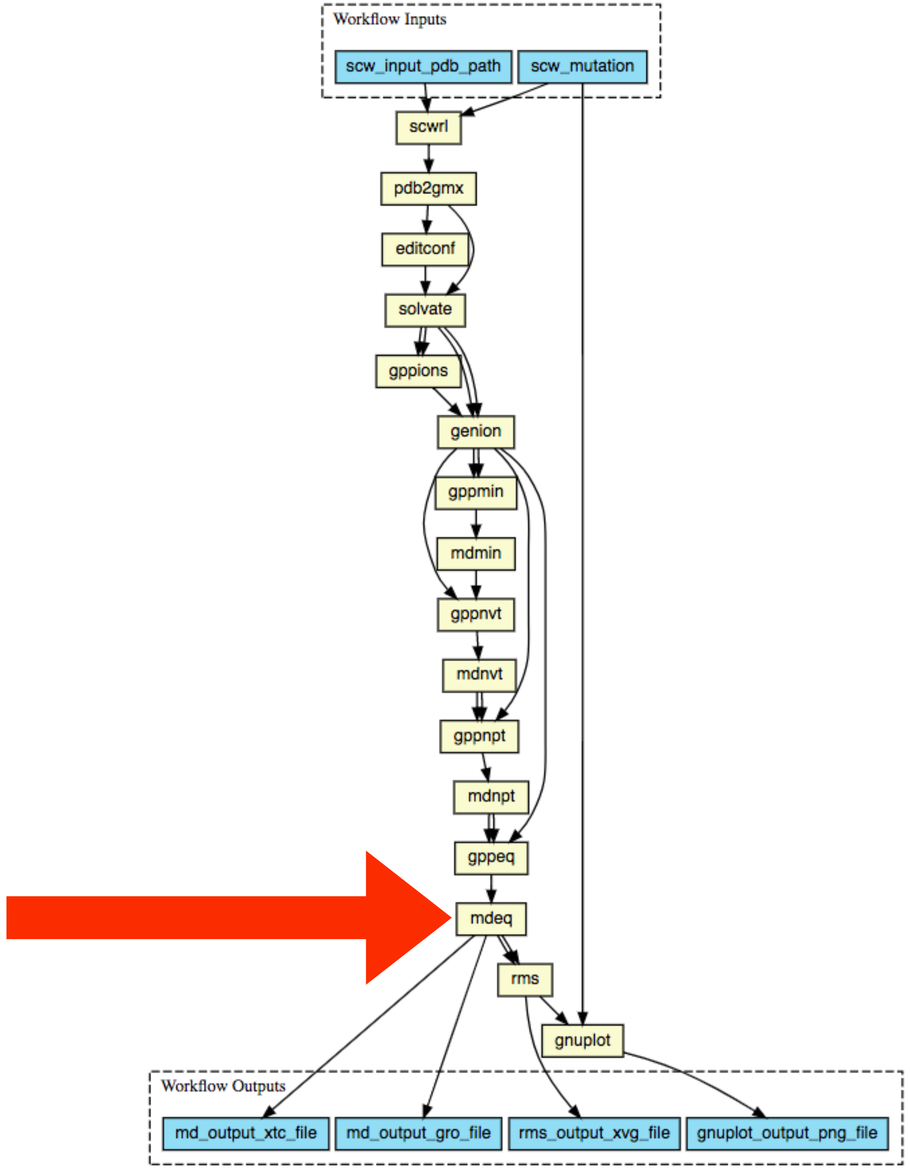
Industry wants to **combine best practices** from academia into useful pipelines for discovery, but has to hack everything together if they want to make this work

EXAMPLE: SETTING UP A FREE ENERGY CALCULATION IN GROMACS



EXAMPLE: SETTING UP A FREE ENERGY CALCULATION IN GROMACS

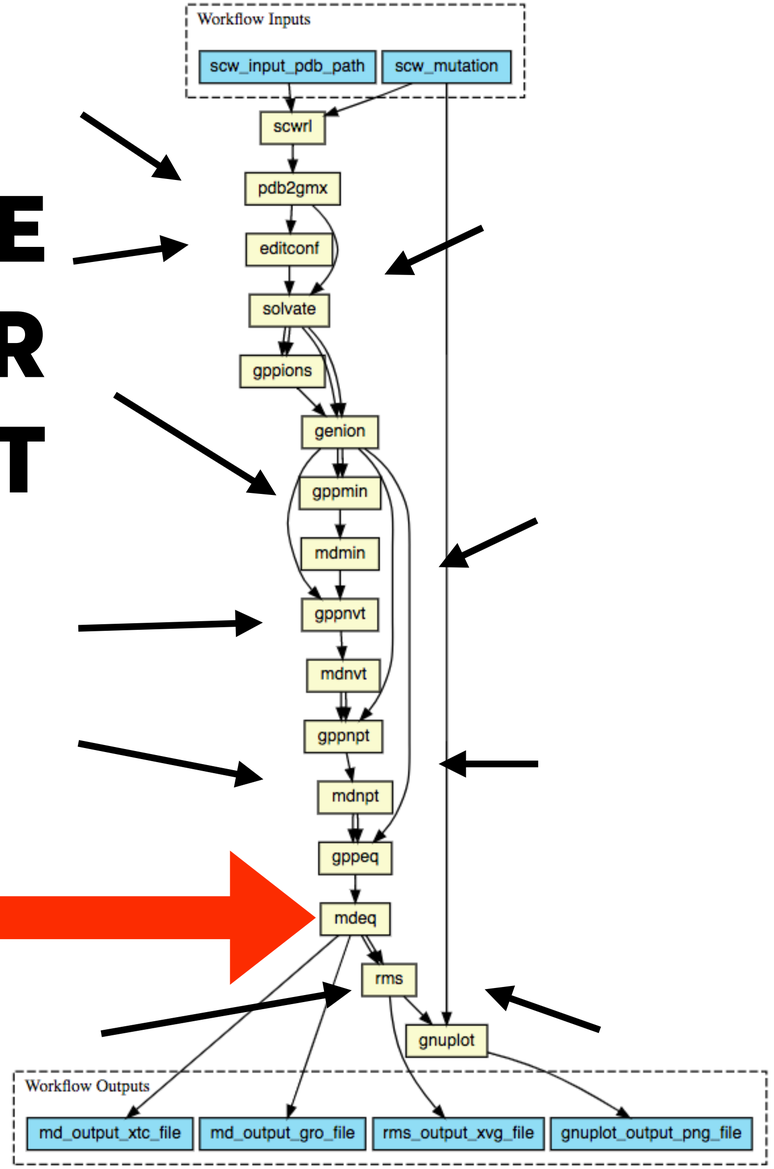
THE SCIENCE
I'M INTERESTED IN



EXAMPLE: SETTING UP A FREE ENERGY CALCULATION IN GROMACS

EVERYTHING ELSE
I NEED IN ORDER
TO RUN MY BIT

THE SCIENCE
I'M INTERESTED IN



http://bioexcel.eu/

REPRODUCIBILITY

Reproducing work from a published computational chemistry paper is currently **nearly impossible**, which **minimizes opportunities for learning and improvement**

Translating best performers from D3R/SAMPL blind challenges into production pipelines is **nearly impossible** for the same reason

Identifying the origin of issues by comparing related tools or protocol choices requires **labor-intensive cooperation between groups**

Example: SAMPL pKa methods

some are detailed:

```
# SOFTWARE SECTION
Software:
COSMOtherm C30_1701
Turbomole 7.2
COSMOconf 4.2
COSMOquick version 1.6
COSMOpy (version2017) & Python 2.7

# METHODS SECTION
#
Method:
The pKa dataset consists of 24 small to medium sized drug-like molecules which combine several functional groups whereas most of them have at least one basic functional group. Molecules SM01, SM08, SM15, SM20 and SM22 possess an additional (significant) acidic functional group
Possible deprotonated and protonated species (anions, cations, zwitterions) have been generated automatically via the COSMOquick software package. A few further potential ions and tautomers were determined from visual inspection of the neutral forms as provided for the challenge. In all cases, only single protonation or deprotonation turned out to be relevant at the experimental region from pH=2 to pH=12.
For all compounds, including the ionic and tautomeric forms, independent sets of relevant conformations were computed with the COSMOconf 4.2 workflow. Additional neutral conformers which are thermodynamically relevant in water according to COSMOtherm computations have been found only for compound SM18 (tautomeric) and SM22 (zwitterionic) and have been included into the respective conformer sets used later on for the COSMOtherm pKa calculations.
The quantum chemistry calculations of COSMO sigma-surfaces were done at the BP//TZVPD//FINE single point level based upon BP//TZVP//COSMO optimized geometries to match the parameterization (BP-TZVPD-FINE-C30-1701) used in the 2017 COSMOtherm-release. All quantum chemical calculations were carried out with the TURBOMOLE 7.2 quantum chemistry software.
The COSMOtherm pka-module uses a simple linear free energy relationship (LFER) in order to correct the free energy differences of the neutral and protonated (deprotonated) forms. ( Klamt, A. et al. J. Phys. Chem. A 107, 9380-9386 (2003). & Eckert et al. J Comp Chem 27, 11-19 (2006).):
pKa = c0 + c1*(DG_neutral-DG_ionic)
with
c0=-131.7422 and c1=0.4910 mol/kcal (for acids in water)
c0=-171.1748 and c1=0.6227 mol/kcal (for bases in water)

pKa values were computed for all identified single protonated and deprotonated sampl6 molecules and the respective zwitterions using the COSMO-RS method as implemented in the COSMOtherm software. The workflow for the batch computation about 80 pKa reactions has been automated via an in-house script based on Python 2.7 (COSMOpy).
For the final submission, only relevant pKa-values were included. For bases all protonation reactions with predicted pKa>0 and for acids all pKa values <14 were selected.
The pKa value of basic molecule SM14 containing 2 equivalent basic groups according to our calculations was corrected by the addition of log10(2).
The accuracy of the pKa prediction with the current COSMOtherm parameterization is about 0.65 log units root mean squared deviation (RMSD). The RMSD was evaluated on a validation set of about 160 basic and acidic compounds having a fairly simple molecular structure. However, due to the somewhat more complex structure of the sampl6 molecules the mean of the expected error may be somewhat higher.
```

some are brief:

```
# SOFTWARE SECTION
#
# All major software packages used and their versions.
# Create a new line for each software.
# The "Software:" keyword is required.
Software:
Gaussian09, versions D.01 and A.02
Microsoft Excel 2008 MacOSX


# METHODS SECTION
#
# Methodology and computational details.
# Level of detail should be at least that used in a publication.
# Please include the values of key parameters, with units, and explain how any statistical uncertainties were estimated.
# Use as many lines of text as you need.
# All text following the "Method:" keyword will be regarded as part of your free text methods description.
Method:
From the microscopic pKa values (submission typeI-Iorga-2) we computed the pKa of macroscopic states for the three simplest systems (SM15, SM20 and SM22) using the procedure described in Bodner, G.M. J. Chem. Education 1986, 63, 246. For SM20 there is one macroscopic state, which is the same as the unique microscopic state. For SM15 and SM22 there are two macroscopic states.
```

DEPLOYMENT

Translating academic research software into a tool that can be employed within industry is **extremely difficult** if not impossible for reasons of code quality, robustness, interoperability, and user-friendliness

LEVI NADEN

PAUL CZODROWSKI

TRAINING

Pharma and comp chem are facing an **exodus of talent** due to retirements and hefty competition from machine learning and data science fields

Need better tools to **train the next generation** of computational chemists and get them excited about working in a field that has rudimentary tools compared to the powerful TensorFlow/PyTorch ecosystems in ML

FUNDING

Industry and federal funding agencies (NSF, NIH) tired of investing $ in software or **research that is not useful** to them or others

Easier to justify small investments in funding to deliver new features if they can be rapidly deployed and utilized/combined

VALIDATION AND ANALYSIS

For blind challenge participants, it's difficult to **validate** the output of your scripts to make sure it's in the right format, and to test on known datasets with the same analysis pipeline that will be used for assessment.

For blind challenge assessors, it's almost impossible to guarantee everyone will submit the data in the right format.

WORKFLOWS TO THE RESCUE

Workflows (and the machinery to support them) can address all of these issues:

* Training
* Interoperability
* Reproducibility
* Evaluation
* Deployment
* Funding
* Enabling focus on key science
* Productivity

WHAT COULD THE FUTURE LOOK LIKE?

Publications/submissions include a **DOI-indexed portable workflow component** that can be pulled from a **common component repository** to reproduce the calculations in the paper/submission in a variety of workflow engines that support common components.

Journals require virtual screening or affinity prediction tools to report performance on **standard benchmark datasets** that the community agrees are valuable.

Researchers can focus their efforts on improving the science underlying **specific components** of versioned best practices workflows, and share them in the **common component respository**.

Industry can easily **evaluate predictive models** on internal datasets without having to embark on a multi-year effort to reimplement, hack together, or harden the software.

Vendors could flexibly charge for use of their tools, potentially by **pay for privacy/ ownership** so tools could be evaluated freely but funded by use for IP generation.

A MolSSI Workshop

DISTRIBUTED WORKFLOWS FOR BIOMOLECULAR SIMULATION

September 12-13, 2017 | Autodesk Gallery, 1 Market Street, San Francisco, CA

Distributed Workflows for Biomolecular Simulations is an invite-only, innovation-driven workshop hosted by MolSSI and Autodesk Life Sciences for academic and industry experts on how workflow technologies will vastly accelerate pipelines from academic research to industrial discovery.

PLEASE SAVE THE DATE, REGISTRATION LINK TO FOLLOW

AUTODESK LIFE SCIENCES

MolSSI THE MOLECULAR SCIENCES SOFTWARE INSTITUTE

BACKGROUND

Workflow technologies simplify the processes of developing reliable computational methods, deploying reproducible and reliable software, exploiting scalable computing, and sharing standardized best practices. With increasing interest in such systems from academic, industrial, and computing groups, this two-day workshop will bring together a diverse group of experts to catalyze and develop modern workflow

OPPORTUNITIES

Workflow component interoperability:

- Components could be portable between workflow engines
  - Academics could **wrap tools once** to make them available to many systems
  - Software vendors could make components available via licensing models
  - Workflow engines could benefit from large **ecosystem** of components
- Common component format could be supported alongside specialized formats
- Enable a **common component repository/registry**?
- We would need to define:
  - How components are **encapsulated**
  - What **information must be exchanged**
  - How components **expose their functionality**
  - Different **licensing models** that enable research, use, and fair compensation
  - How toolmakers can get **feedback** (especially regarding failures)

WHAT ARE WE EVALUATING IN BLIND COMPETITIONS?

evaluating the driver

evaluating the technology

Need to separate capabilities of technology from skill of driver

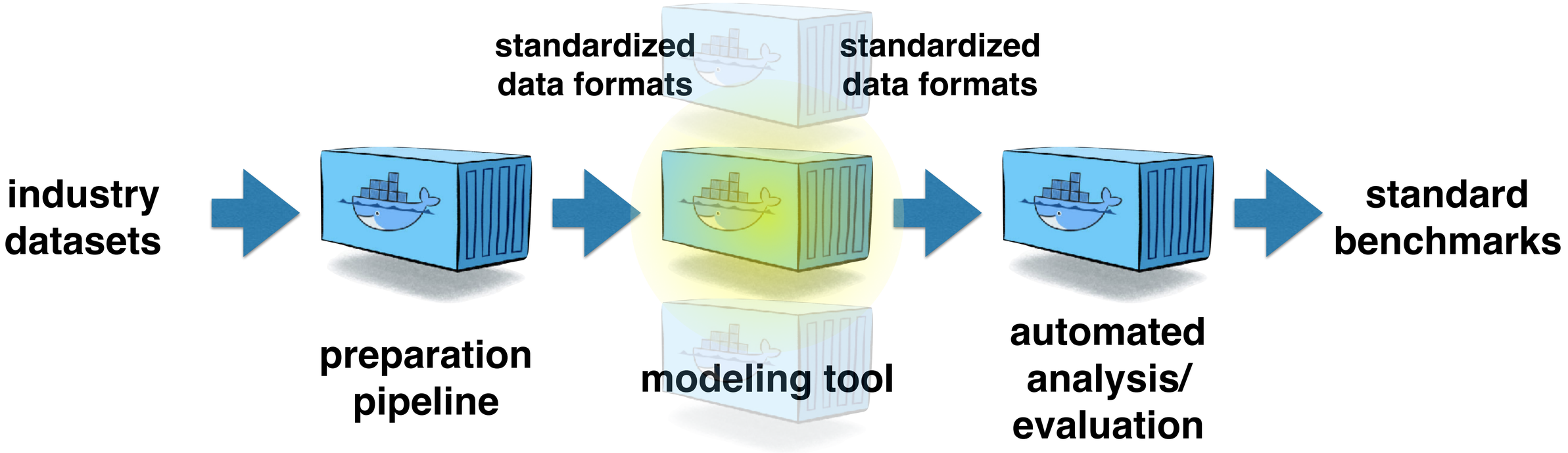
BLIND CHALLENGES SHOULD EVALUATE THE TECHNOLOGY, NOT THE DRIVER

Need to separate technology from operator in order to statistically evaluate performance of the technology

Can't easily do this with traditional blind challenge format

Blind challenges aren't enough: They have to be automated.

WORKFLOWS USING BEST PRACTICES WOULD ALLOW US TO EVALUATE THE TECHNOLOGY

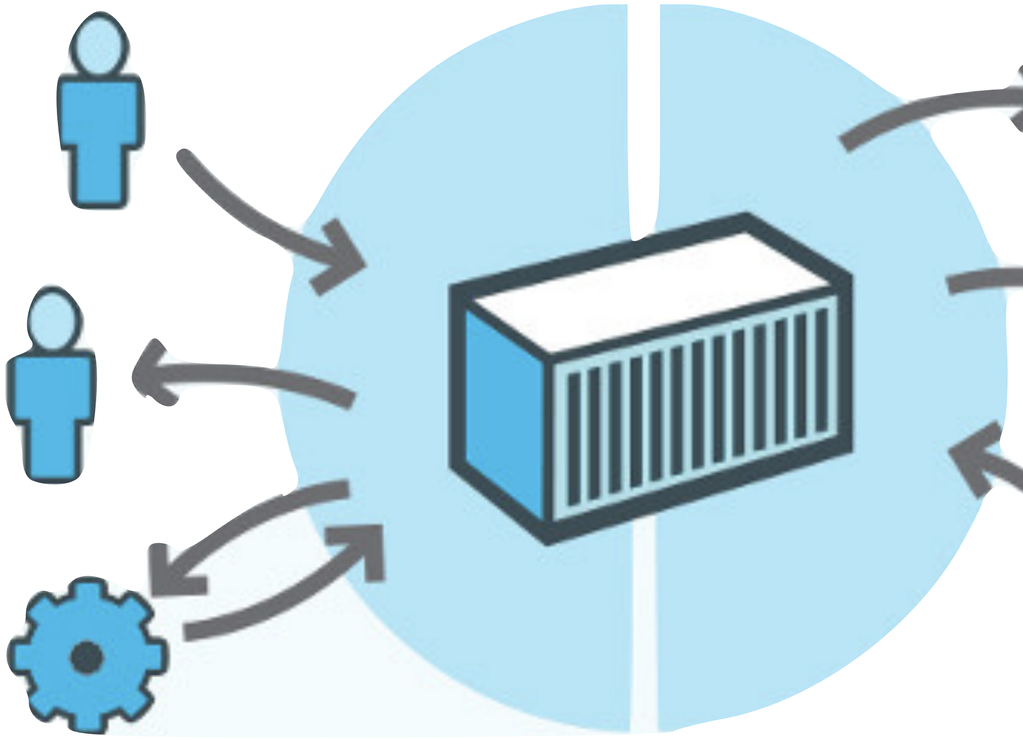


docker

CONTAINERS SOLVE THE PORTABILITY PROBLEM

interactive terminal/GUI sessions

standardized programmatic interfaces



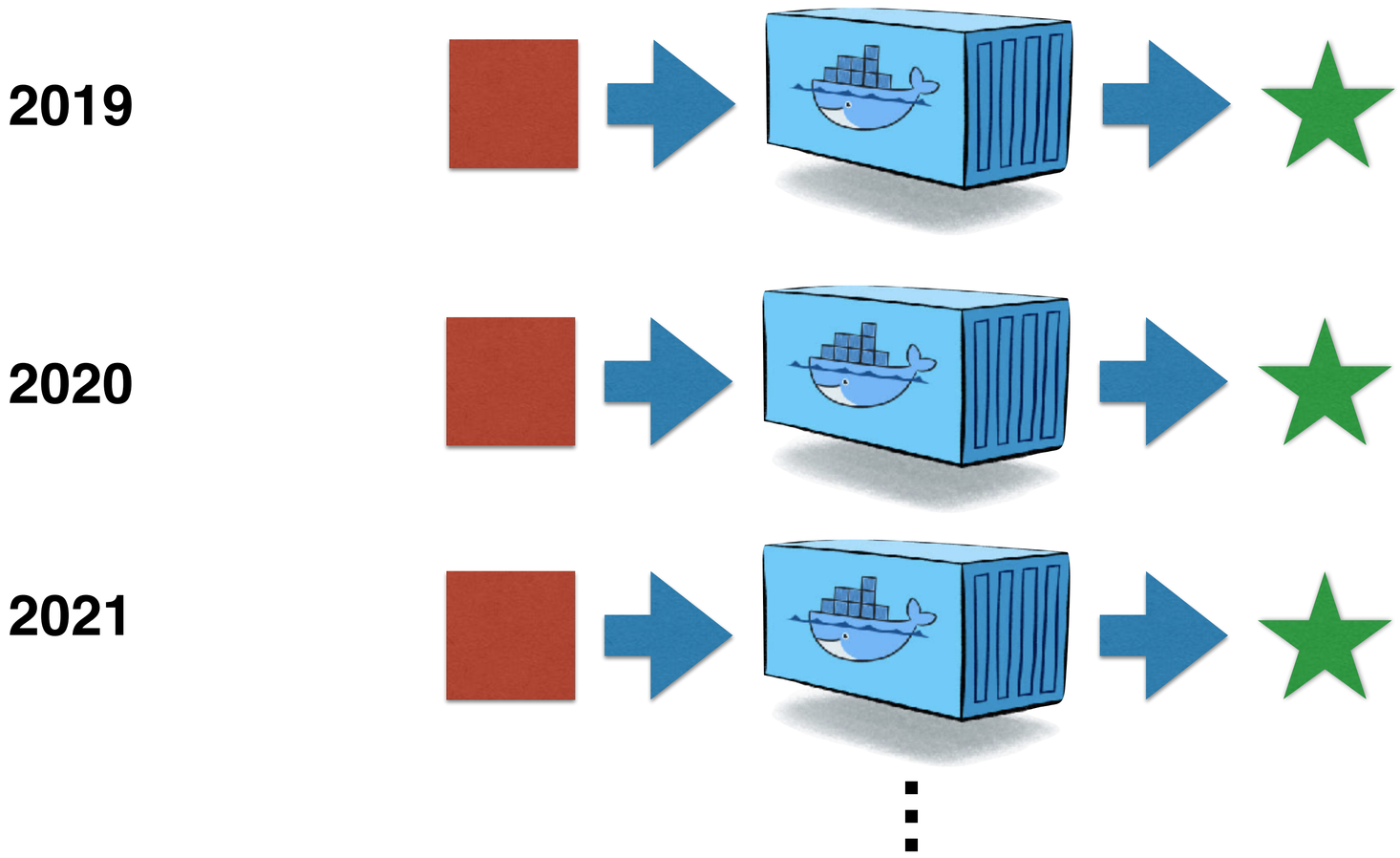
cloud
(AWS, Google Compute)

local resources
(OS independent)

laptop/desktop
(essential for training)

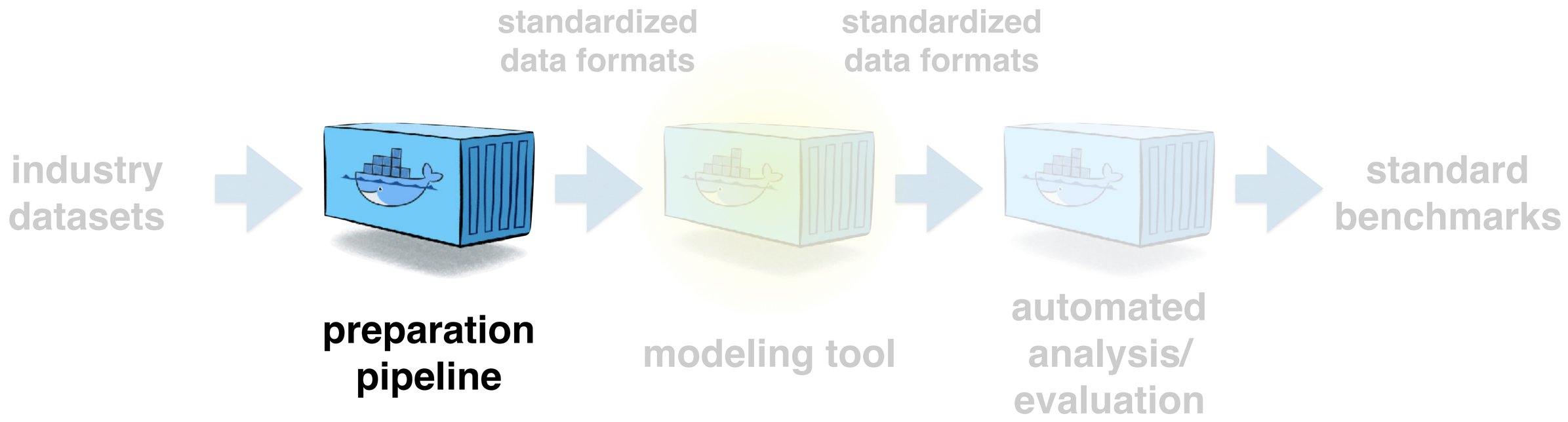
docker

CONTAINERS SOLVE THE REPRODUCIBILITY PROBLEM



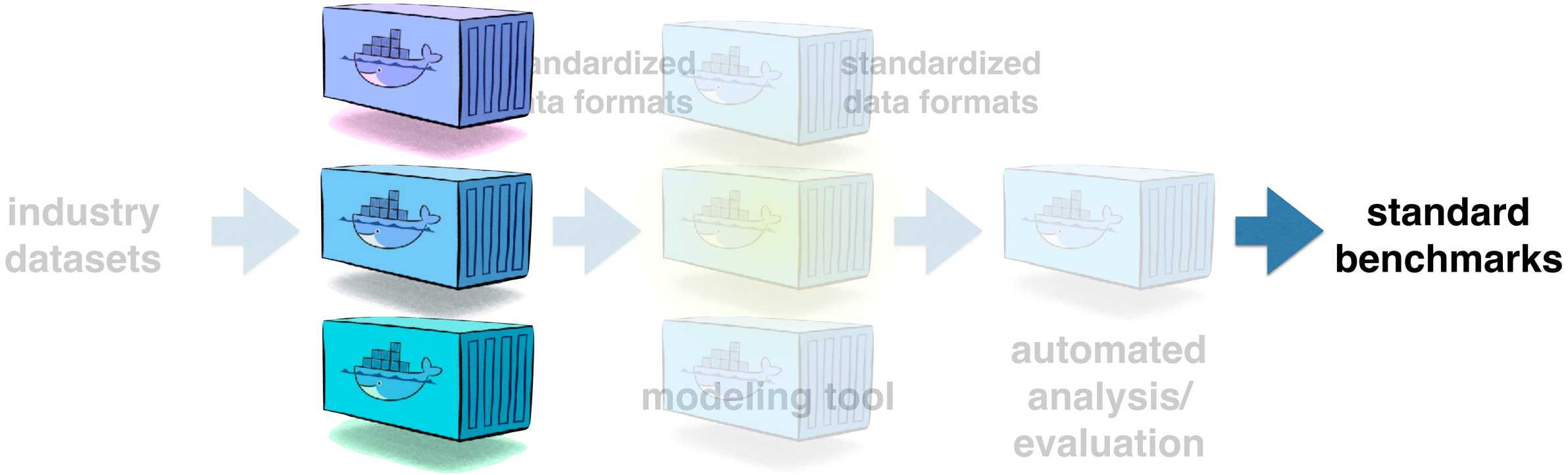
docker

OPEN PREPARATION PIPELINES COULD CAPTURE COMMUNITY-DRIVEN BEST PRACTICES



docker

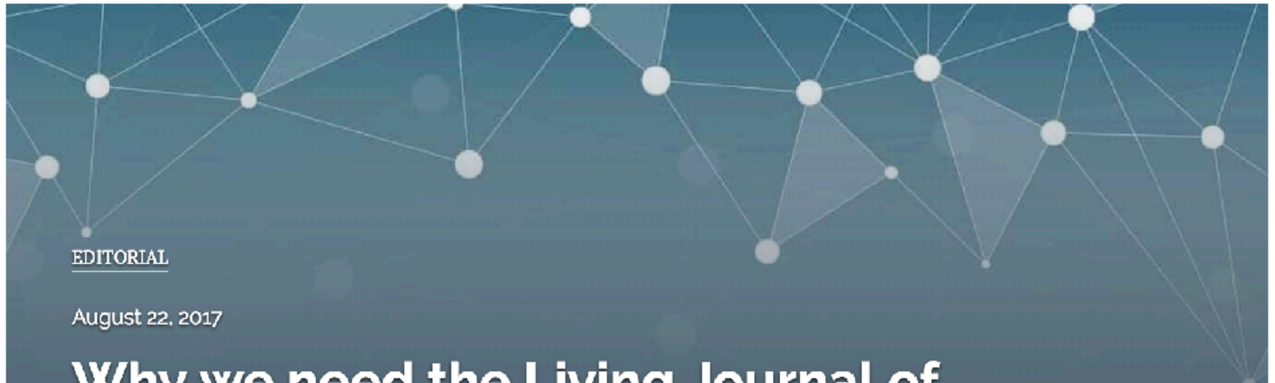
BEST PRACTICES CAN BE EVALUATED BY TESTING VARIATIONS ON A VARIETY OF MODELING TOOLS



docker

ISSN: 257

Living Journal of Computational Molecular Science



EDITORIAL

August 22, 2017

Why we need the Living Journal of

Recent blog posts

Why student reviewers? The future of our field depends on it.

Leveraging the unique expertise of experts-to-be.

LiveCoMS "Best Practices" articles: Put your simulations

Best Practices for Quantification of Uncertainty and Sampling Quality in Molecular Simulations [Article v1.0]

Alan Grossfield[1*†], **Paul N. Patrone**[2*†], **Daniel R. Roe**[3*†], **Andrew J. Schultz**[4*†], **Daniel W. Siderius**[5*†], **Daniel M. Zuckerman**[6*†]

[1]University of Rochester Medical Center, Department of Biochemistry and Biophysics; [2]Applied Computational and Mathematics Division, National Institute of Standards and Technology; [3]Laboratory of Computational Biology, National Heart Lung and Blood Institute, National Institutes of Health; [4]Department of Chemical and Biological Engineering, University at Buffalo, The State University of New York; [5]Chemical Sciences Division, National Institute of Standards and Technology; [6]Department of Biomedical Engineering, Oregon Health & Science University

This LiveCoMS document is maintained online on GitHub at https://github.com/dmzuckerman/Sampling-Uncertainty; to provide feedback,

Abstract The quantitative assessment of uncertainty and sampling quality is essential in molecular simulation. Many systems of interest are highly complex, often at the edge of current computational capabilities. Modelers must therefore analyze and communicate statistical uncertainties so that "consumers" of simulated data understand its significance and limitations. This article covers key analyses appropriate for trajectory data generated by conventional simulation methods such as

THIS REQUIRES STANDARDIZED DATA INTERCHANGE FORMATS

standardized data formats

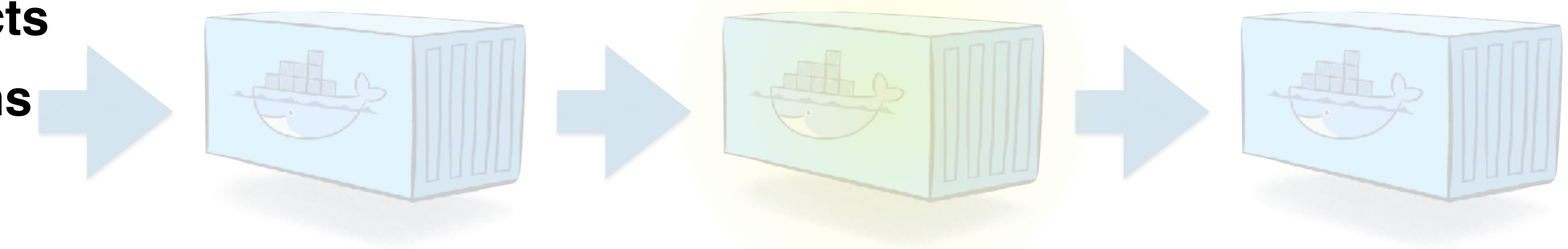
standardized data formats

protein constructs

assay conditions

molecules

industry datasets



preparation pipeline

modeling tool

automated analysis/ evaluation

standard benchmarks

biomolecular target
replace aging PDB format
handle charges, parameters, etc.
robust open source readers/writers

parameterized small molecules
make up for shortcomings in mol2, SDF
suitable for the internet age (e.g. JSON)

prediction formats
binding poses
predicted affinity/assay data
predict confidence/uncertainties
exception logging

assessment formats
standard representations
standard assessments
standardized uncertainty analysis

docker

RIGHT NOW, IT'S EVEN DIFFICULT TO DESCRIBE WHAT WE'RE MODELING

Biologist's description

"We expressed human Abl kinase T315I (isoform IA residues 242-493 fused to an N-terminal His6-TEV tag), cleaved with TEV protease, and incubated at high concentration to induce autophosphorylation. Assays were run in 100 uL of 1 uM kinase in assay buffer (20 mM Tris buffer pH 8 with 50 mM NaCl) to which 100 nL of 10 mM DMSO stock of imatinib was added."

Need to extract **structured description**

- **biopolymers**
  sequence construct
  covalent modifications/adducts
- **small molecules**
  identities, numbers/concentrations
  protonation state/tautomer
- **buffer**
  buffer molecules, salt concentration, pH, redox potential
- **thermodynamic state**
  temperature, pressure

Also need to specify source structural data (PDB IDs?) to be used to generate initial geometries.

PARTICIPANTS WILL CONTAINERIZE THEIR MODELING TOOLS

standardized data formats

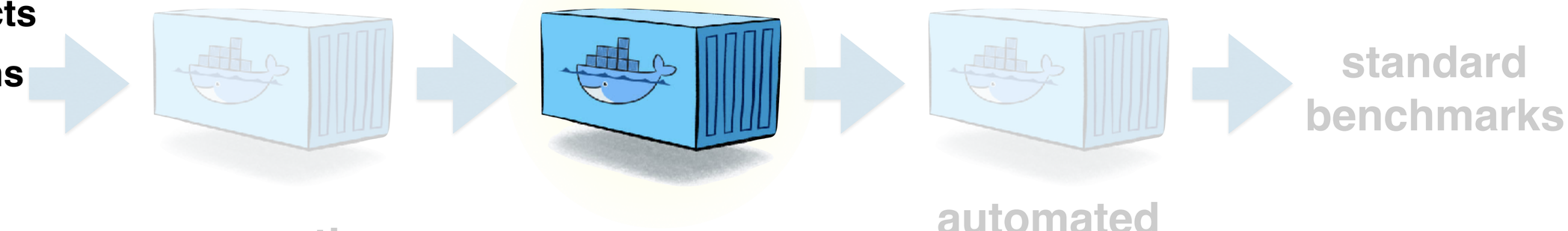
standardized data formats

protein constructs

assay conditions

molecules

industry datasets



standard benchmarks

preparation pipeline

modeling tool

automated analysis/ evaluation

biomolecular target
replace aging PDB format
handle charges, parameters, etc.
robust open source readers/writers

parameterized small molecules
make up for shortcomings in mol2, SDF
suitable for the internet age (e.g. JSON)

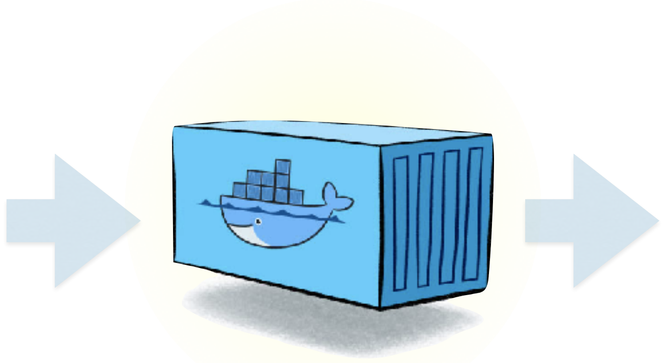
prediction formats
binding poses
predicted affinity/assay data
predict confidence/uncertainties
exception logging

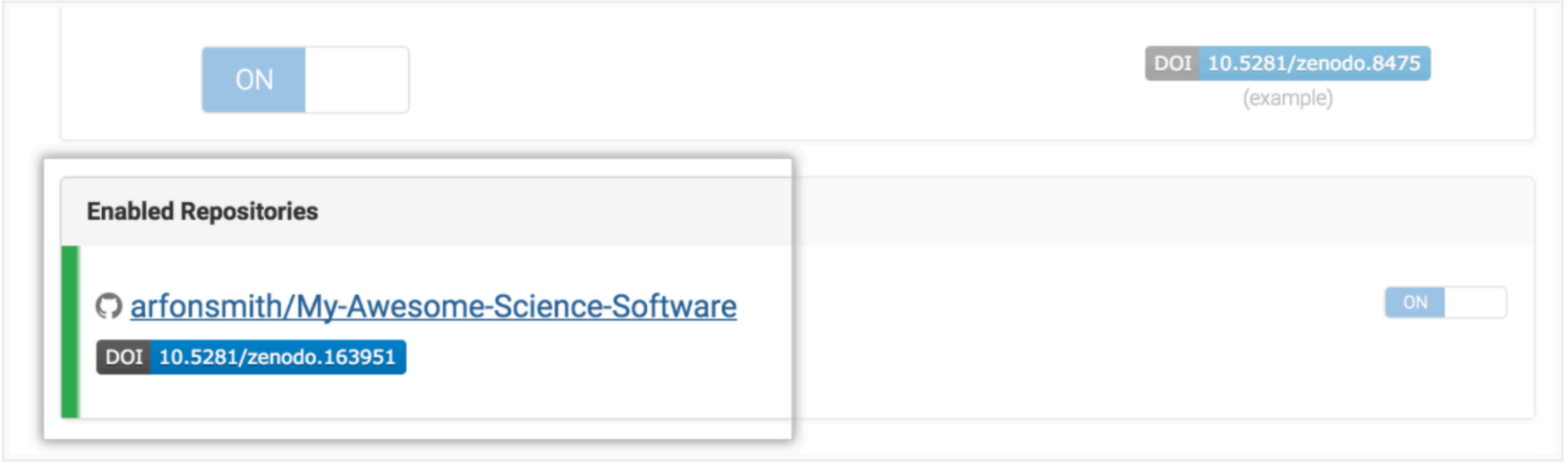
assessment formats
standard representations
standard assessments
standardized uncertainty analysis

docker

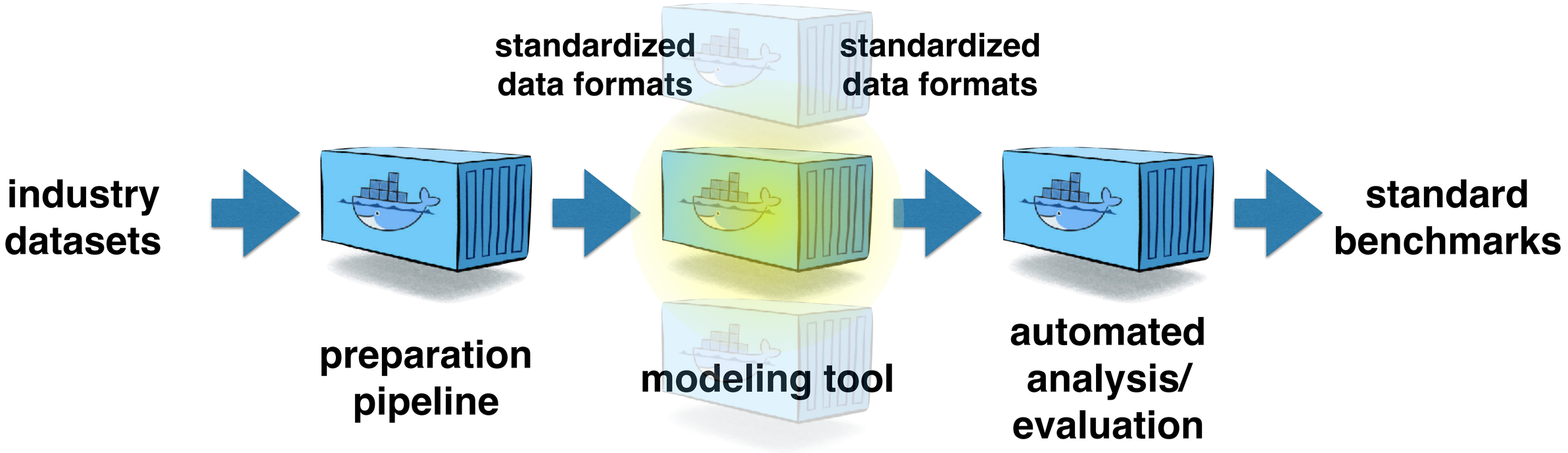
WE COULD DEFINE A COMMON COMPONENT FORMAT, I/O, API, AND REPOSITORY

What if every modeling tool paper came with a DOI that let you pull the exact tool used in that paper from a common component registry and evaluate it yourself?





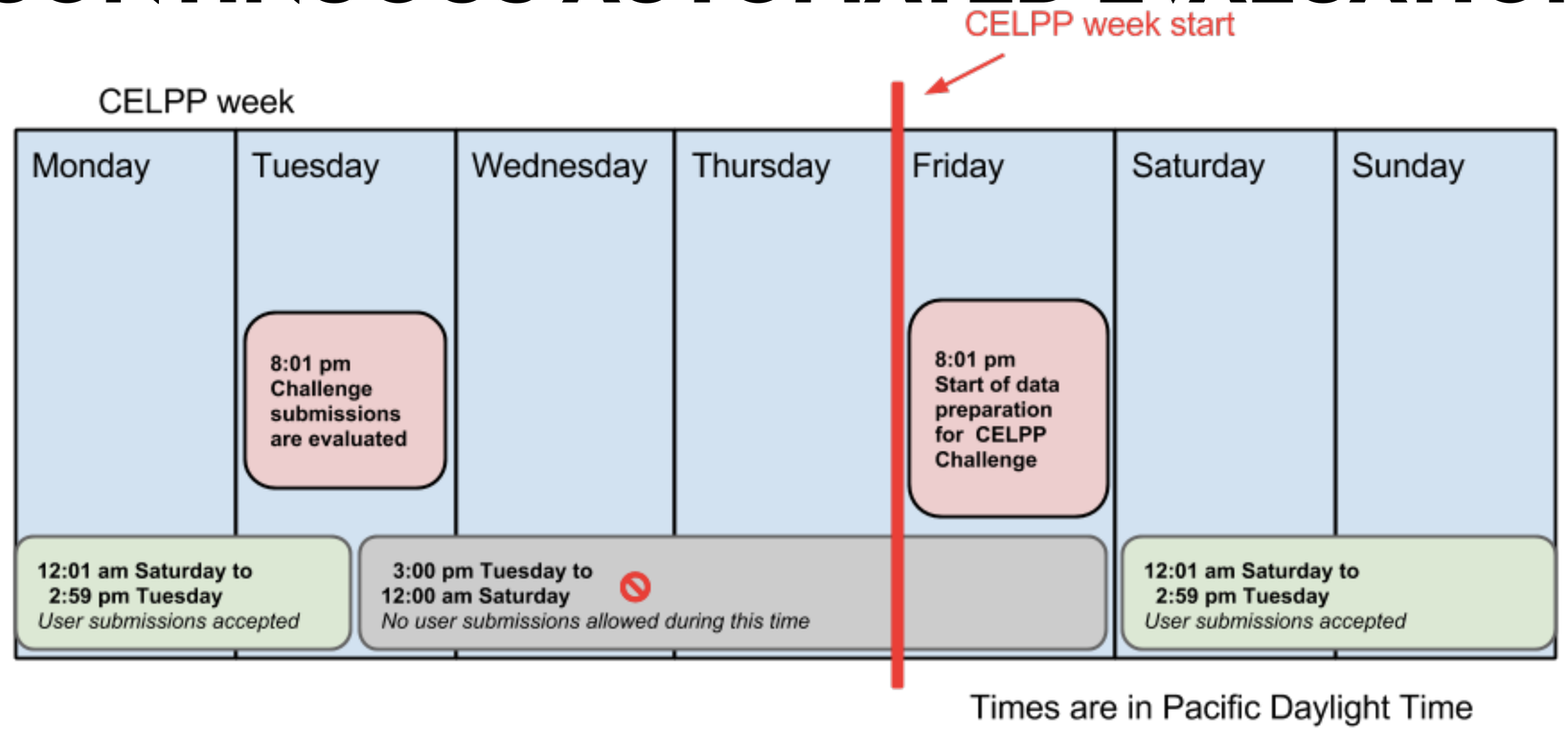
FULLY AUTOMATED LUXURY BLIND CHALLENGES



We can find a way to fund AWS / GCE time to run tools
retrospectively and **prospectively** for modeling evaluation

docker

CELPP AS A PROOF OF CONCEPT: CONTAINERIZED TOOLS ENABLE CONTINUOUS AUTOMATED EVALUATION



CELPP is a python based application that consumes the wwPDB INCHI strings, selects appropriate docking targets, and prepares the proteins and ligands for weekly automated docking challenges. **CELPP** challenge participants will perform the docking and send the results to **CELPP** to be evaluated against weekly released "answers."

https://drugdesigndata.org/about/celpp
Michael Gilson (UCSD) et al.

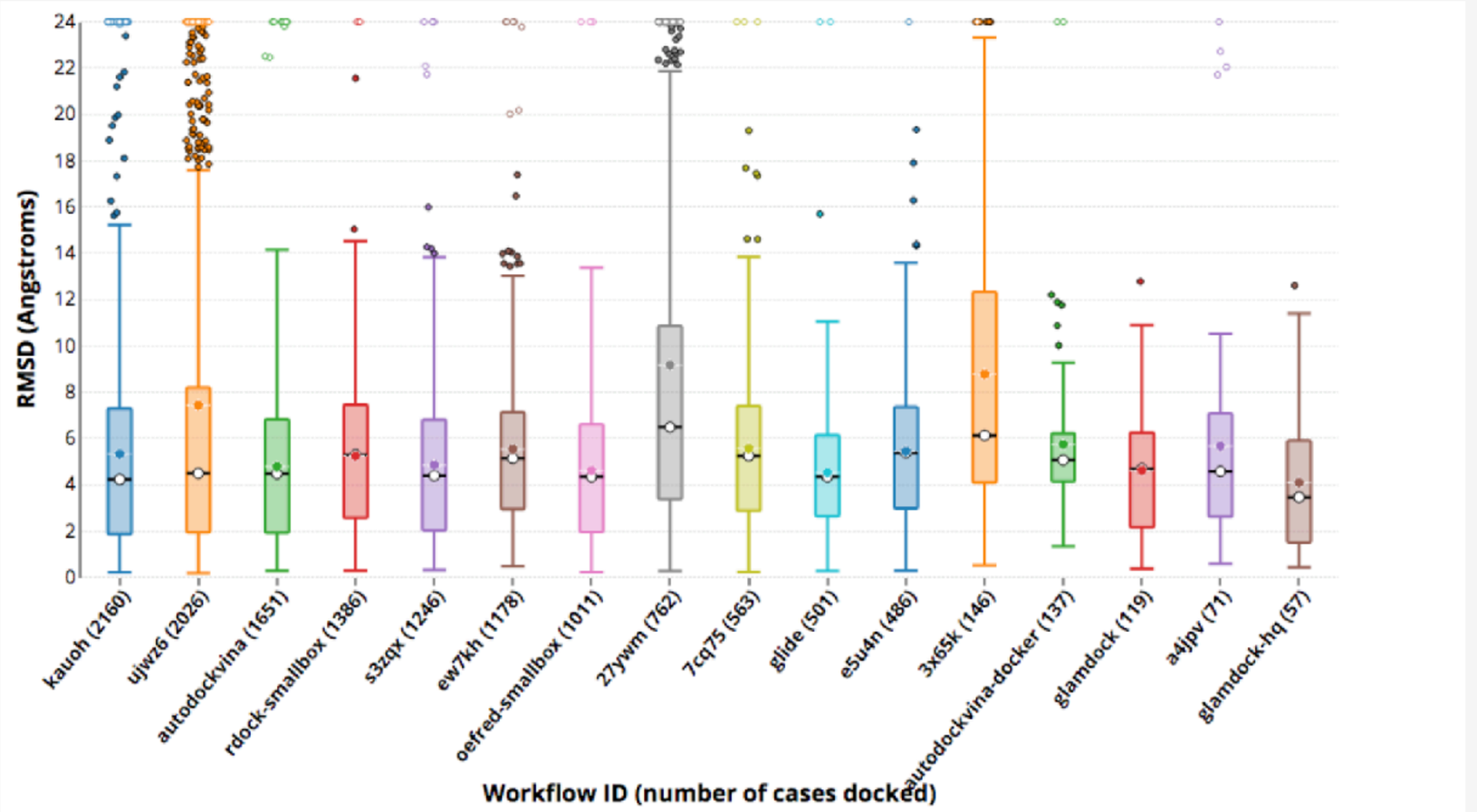
CELPP: Continuous Evaluation of Ligand Protein Prediction

Weekly blinded challenges for ligand pose prediction

4384 targets, 88 weeks as of 2019/Week 33 (2019-08-18), 16 automated workflows

Sign me up!

Candidate protein solved with ligand having largest maximum common substructure with target ligand (LMCSS)



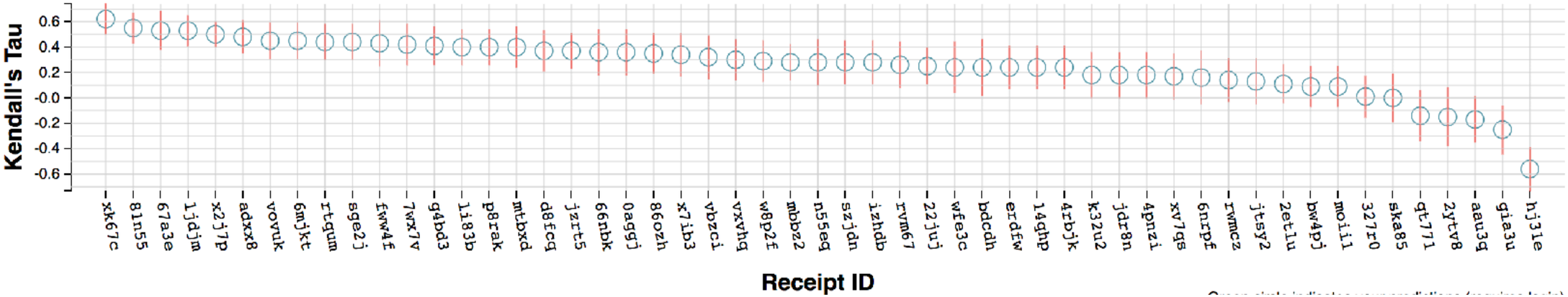
Show: Box Plot Scatter Plot

Boxes: first and third quartiles (Q1, Q3). **Whiskers**: min and max after removal of outliers (points 1.5 times the interquartile range above Q3 and below Q1). **Dots**: outliers; Those above the graph range are placed at the graph maximum.

https://drugdesigndata.org/about/celpp

WHAT SCALE OF DATA DO WE NEED?

Limited data set sizes make it difficult to resolve differences between methods



Machine learning approaches require large training sets to be useful

For the work described here, we mainly use the PDBbind (v.2016) database, containing 13,308 protein–ligand complexes and their corresponding experimentally determined binding affinities collected from literature and the Protein Data Bank (PDB), in terms of a dissociation (K_d), inhibition (K_i) or half-concentration (IC_{50}) constant. A smaller *refined* subset (n_r = 4057)(37) is extracted from it following quality protocols addressing structural resolution and experimental precision of the binding measurement. These controls

https://doi.org/10.1021/acs.jcim.7b00650

WE NEED TO INVEST IN NEW DATA COLLECTION STRATEGIES

Large datasets that provide statistically meaningful evaluations and useful training sets

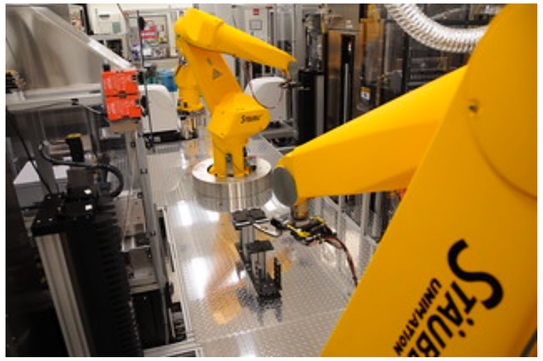
Statistically constructed datasets that assess performance statistics for various real-world application scenarios

Targeted datasets that focus on specific accuracy-limiting effects

Synergistic datasets for evaluating multi-objective design strategies (e.g. kinase inhibitors binding to kinases, HSA, logD, logS)

STRATEGIC PARTNERSHIPS MAY HELP

National Center for Advancing Translational Sciences

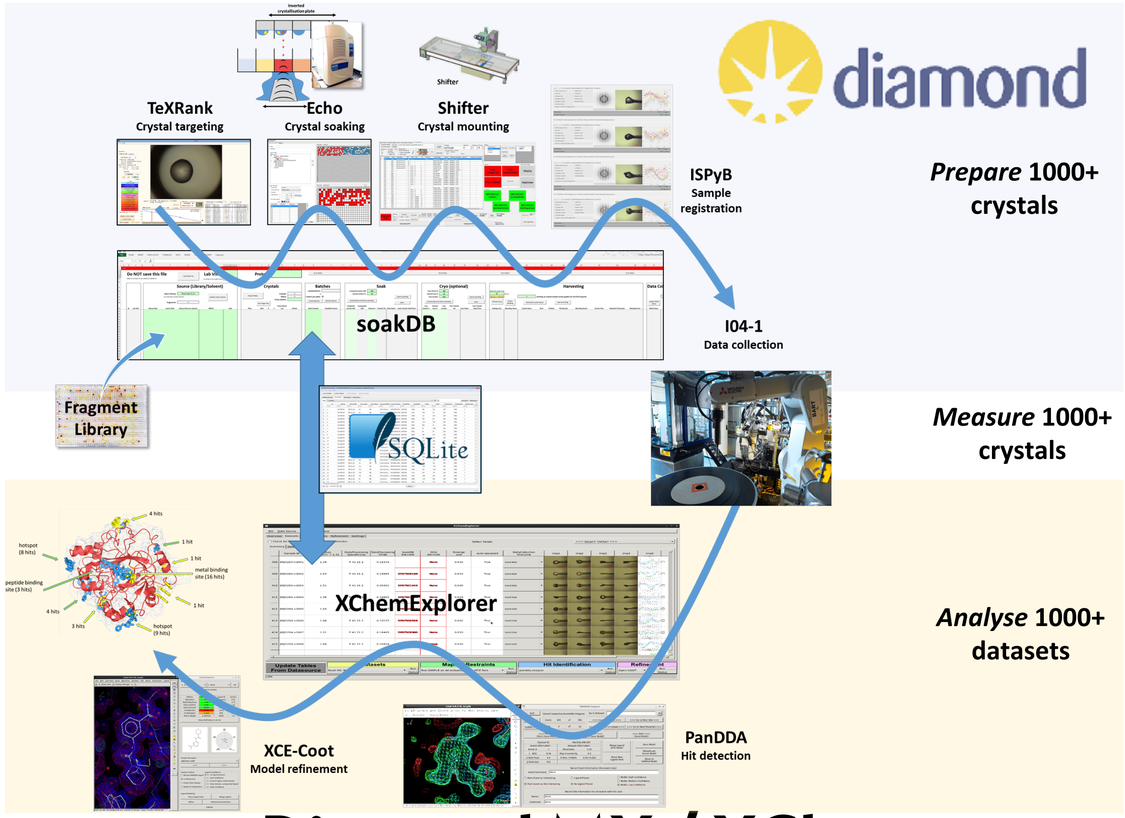


NIH NCATS

high-quality compound libraries (550K+)

gold-standard assays

O(10^5) measurements/target



Diamond MX / XChem

high-throughput crystallography

workflow-based predictive modeling

O(10^3) structures/target

***want to deploy our tools in workflows!**

Enamine

Enamine

~11B compounds <$200/compound

CAN WE DEVELOP NEW HIGH-THROUGHPUT METHODS FOR HIGH-QUALITY DATA COLLECTION

Lessons form **SAMPL6 pKa challenge**:

- Methods can predict good pKas for wrong reasons (incorrect microstates); need signifcantly more microstate pKa data
- The problem: manual NMR took ~2 months for 2 compounds

Could we build an integrated cheminformatics + automated NMR sample preparation + automated NMR data processing pipeline that just produces high-throughput, high-quality microstate pKa data for compounds purchased form Enamine REAL?

Where do we need to focus effort on generating datasets?

ORIGINAL DATASETS COULD BE AUGMENTED

ORIGINAL DATASETS

PHARMA
(dead projects)

JOURNALS?
(prior to publication)

OTHER DATA GENERATORS?
(e.g. SGC)

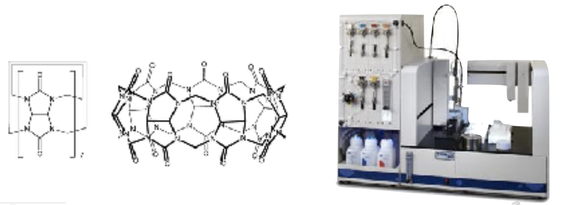
NEW BLINDED DATA

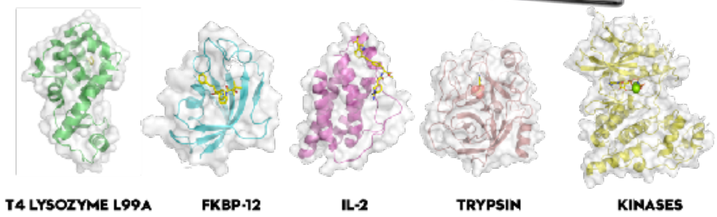
NIH NCATS / DiamondMX
follow-up assays with
multipoint titrations,
additional compound, X-ray

supplement with additional
inexpensive one-step synthesis

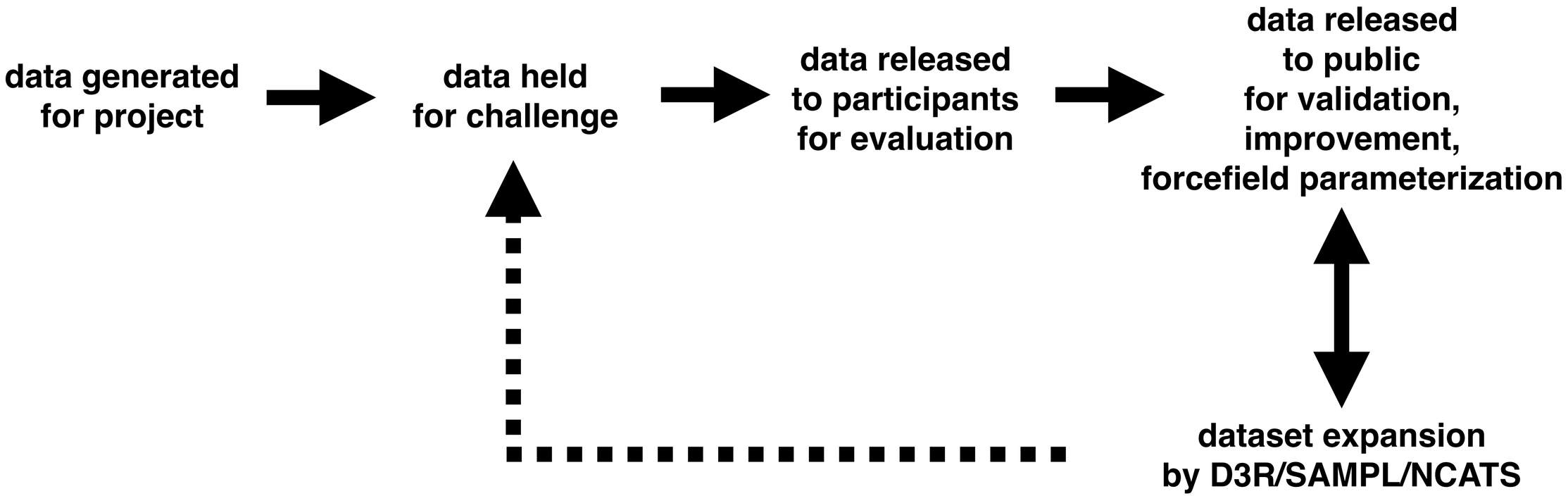
Enamine

automated model systems and physical
property measurements
(academic/industry collaborations)





CHALLENGE DATA HAS A LONG, IMPACTFUL LIFE CYCLE



MODERN MACHINE LEARNING TOOLKITS MAKE ACCESS TO STANDARD BENCHMARK SETS EASY

```
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

load your tools

grab a dataset

define a new kind of model

declare your objectives in training it

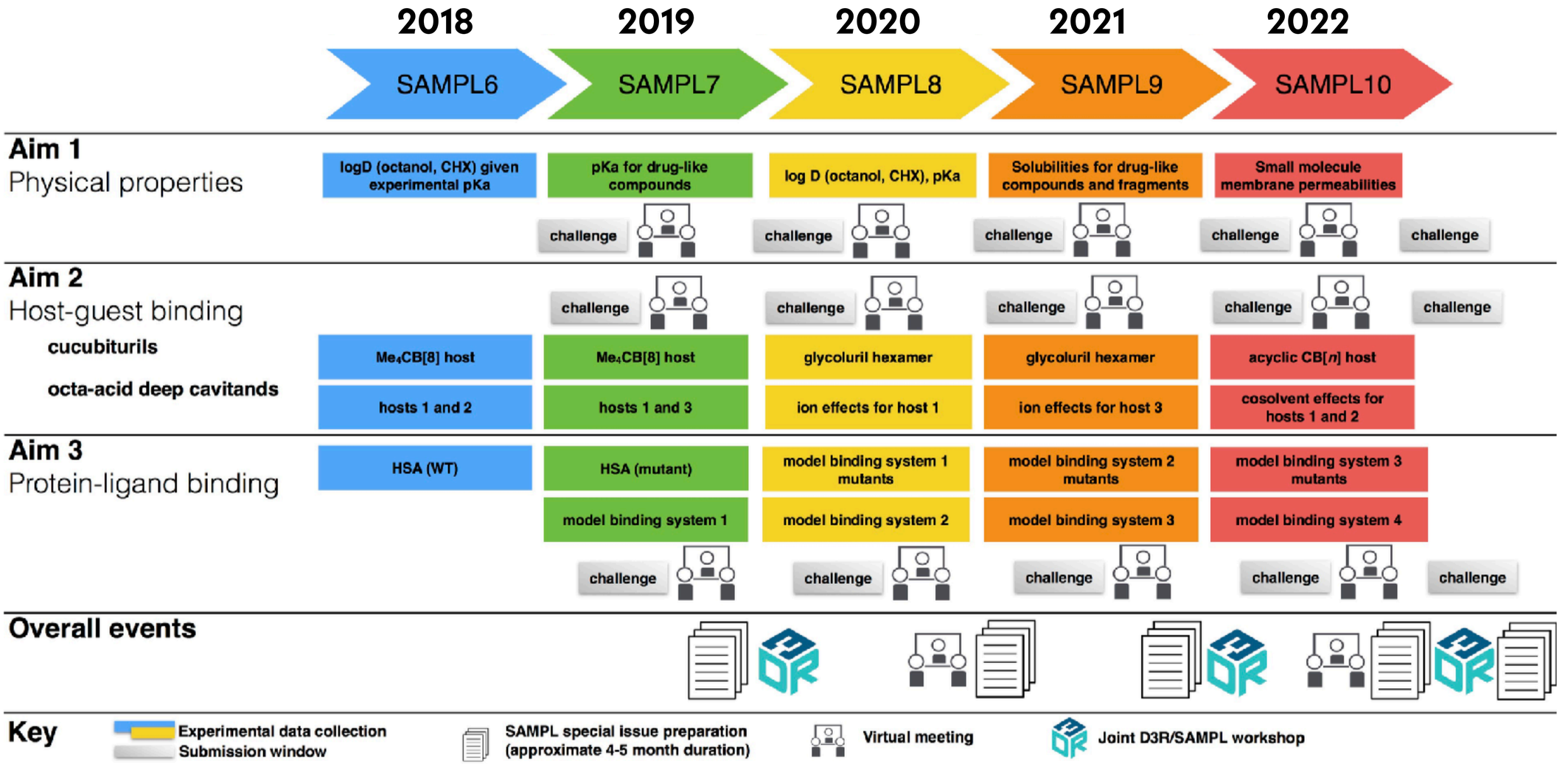
fit it

use it

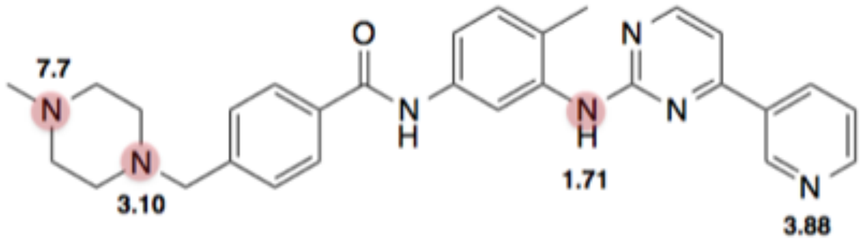
https://www.tensorflow.org/overview

Why can't we make it this easy to do new things in computer-aided drug discovery?

SAMPL IS LOOKING TO TACKLE MODEL PROTEIN:LIGAND SYSTEMS

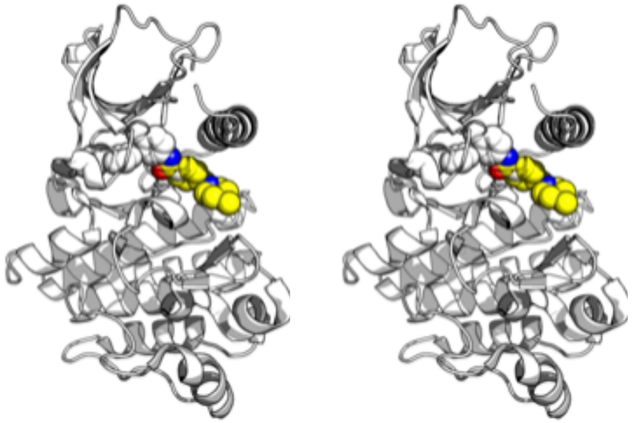


SAMPL AIMS TO EXPLOIT SYNERGIES TO MAXIMIZE OPPORTUNITIES FOR LEARNING

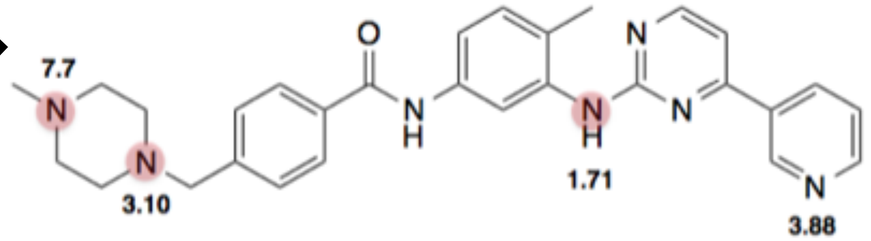


compounds and their fragments

SAMPL AIMS TO EXPLOIT SYNERGIES TO MAXIMIZE OPPORTUNITIES FOR LEARNING

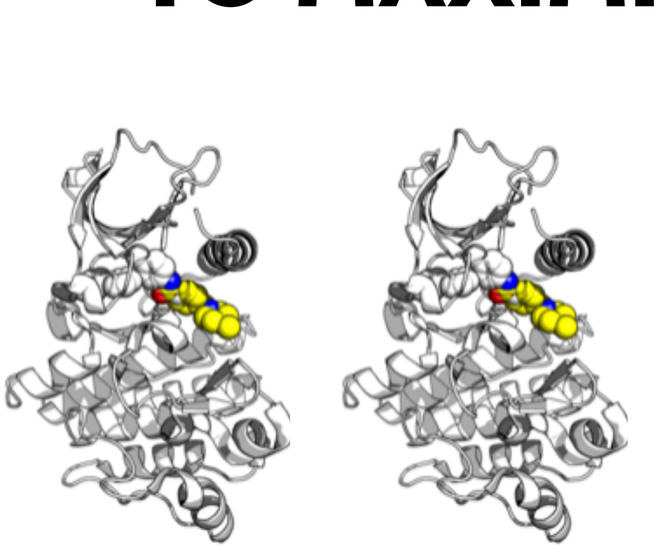


kinases

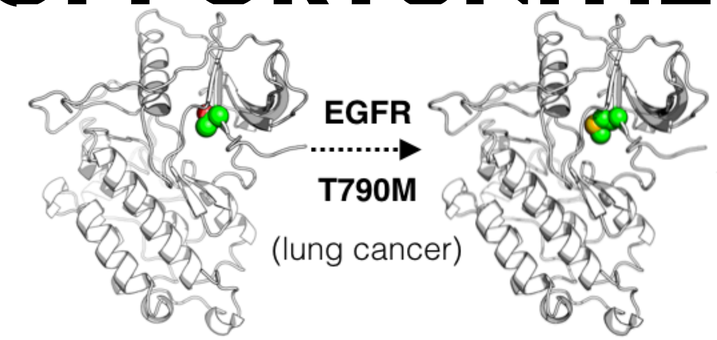


compounds and their fragments

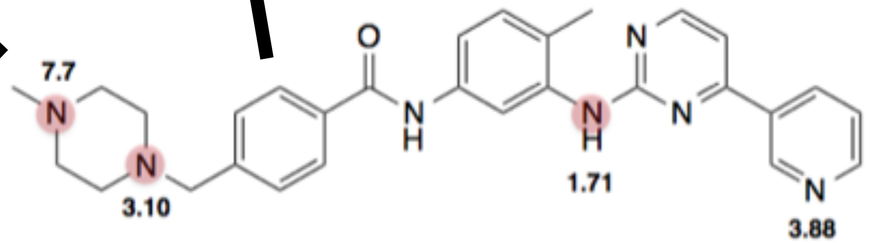
SAMPL AIMS TO EXPLOIT SYNERGIES TO MAXIMIZE OPPORTUNITIES FOR LEARNING



kinases

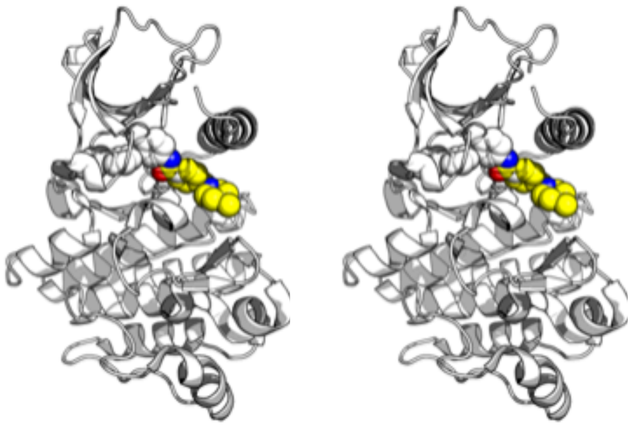


mutations

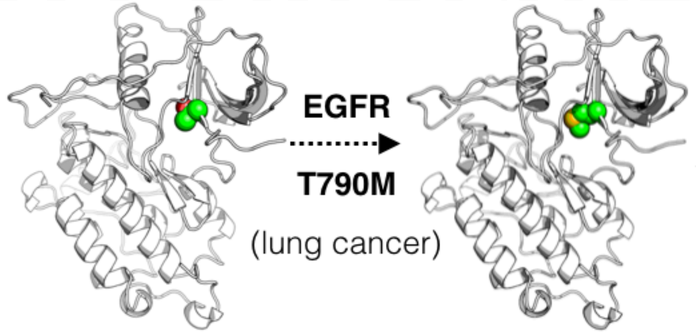


compounds and their fragments

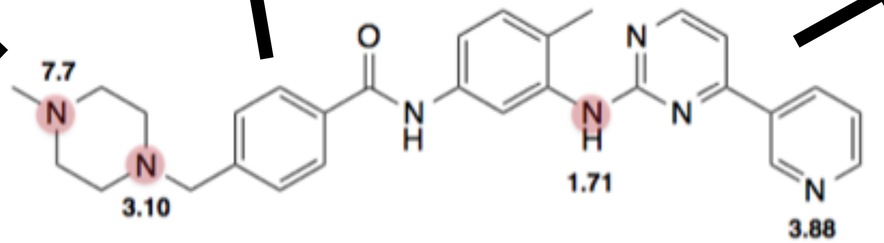
SAMPL AIMS TO EXPLOIT SYNERGIES TO MAXIMIZE OPPORTUNITIES FOR LEARNING



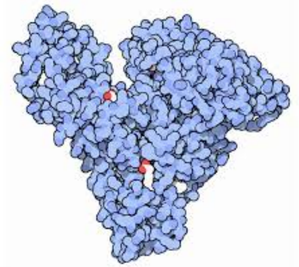
kinases



mutations

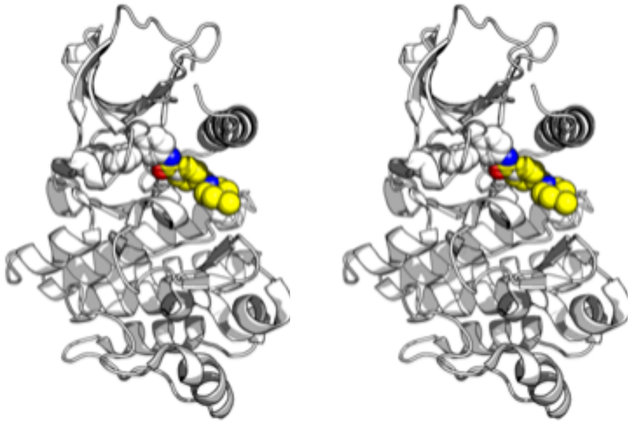


compounds and their fragments

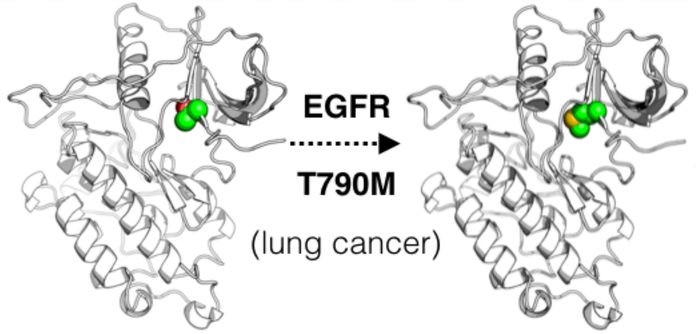


ADME/Tox relevant proteins (e.g. HSA)

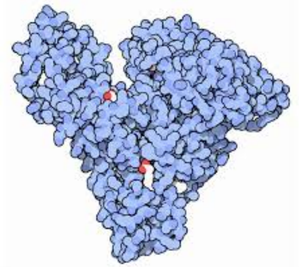
SAMPL AIMS TO EXPLOIT SYNERGIES TO MAXIMIZE OPPORTUNITIES FOR LEARNING



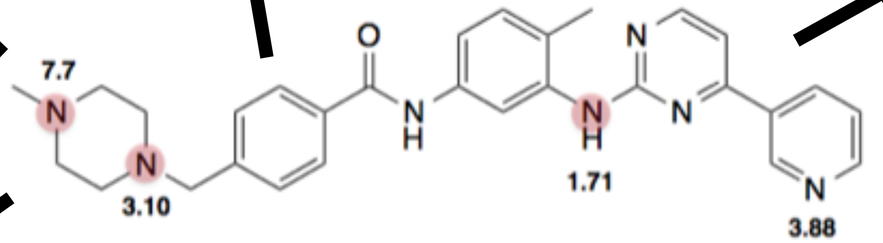
kinases



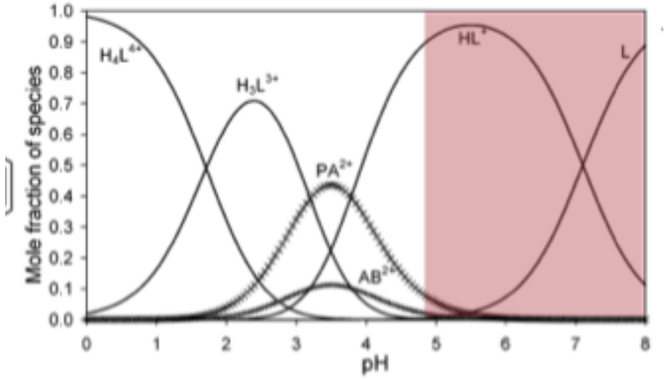
mutations



ADME/Tox relevant proteins (e.g. HSA)

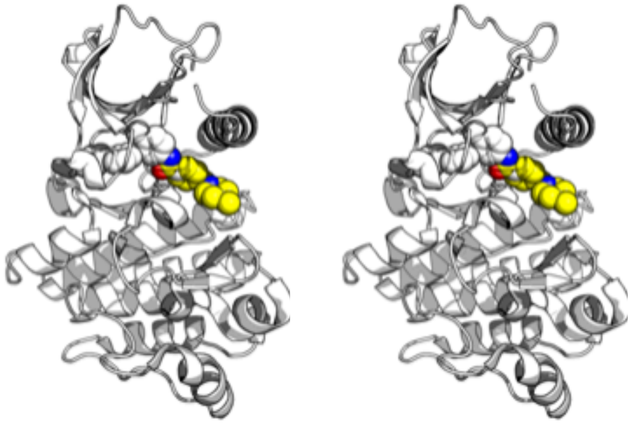


compounds and their fragments

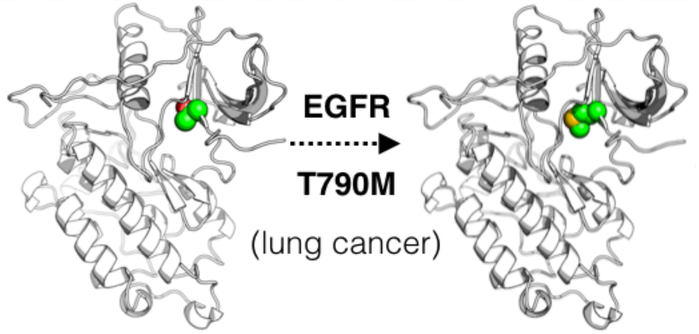


pKa, tautomers, microstates

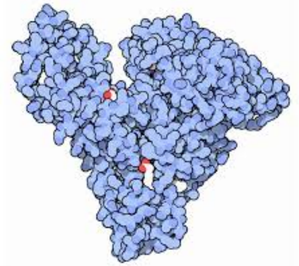
SAMPL AIMS TO EXPLOIT SYNERGIES TO MAXIMIZE OPPORTUNITIES FOR LEARNING



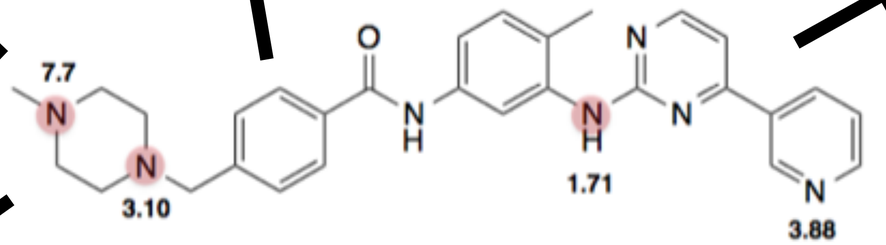
kinases



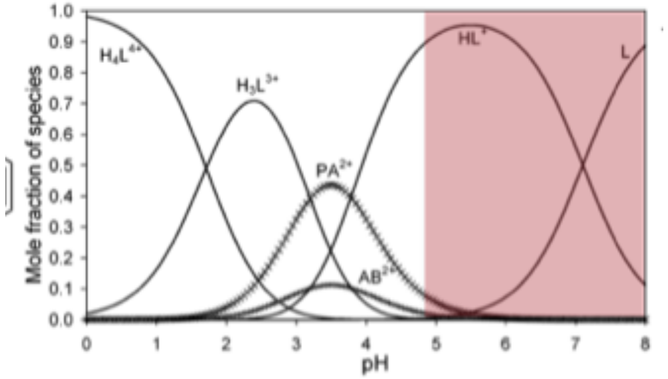
mutations



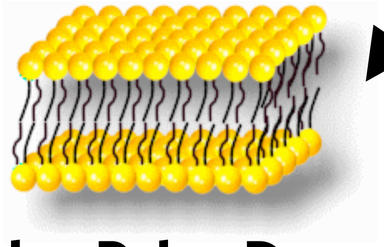
ADME/Tox relevant proteins (e.g. HSA)



compounds and their fragments

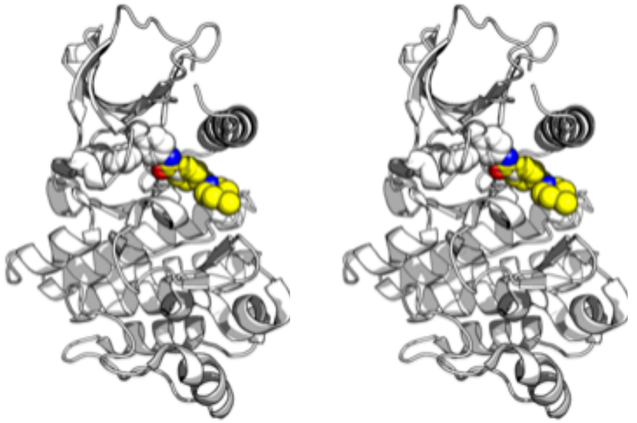


pKa, tautomers, microstates

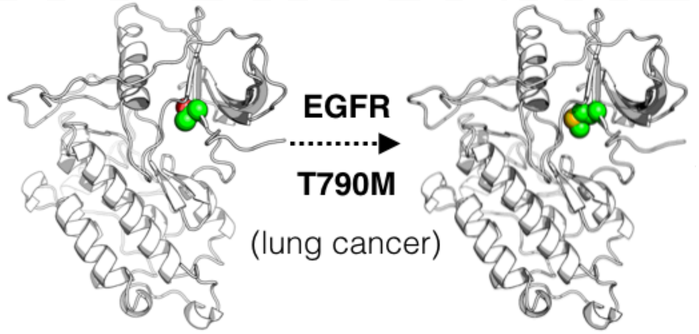


logP, logD, permeability

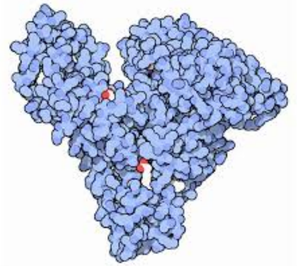
SAMPL AIMS TO EXPLOIT SYNERGIES TO MAXIMIZE OPPORTUNITIES FOR LEARNING



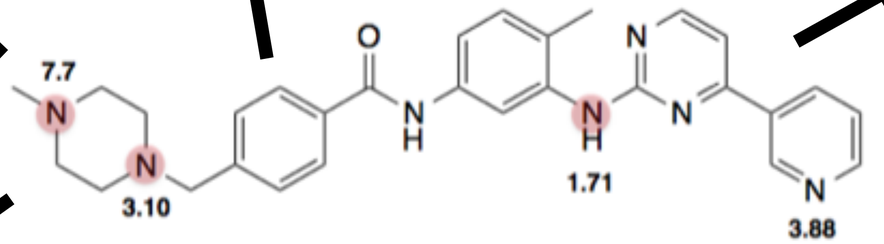
kinases



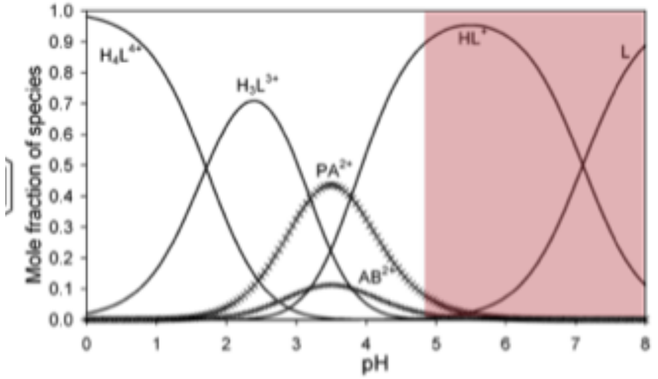
mutations



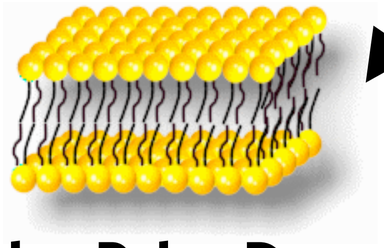
ADME/Tox relevant proteins (e.g. HSA)



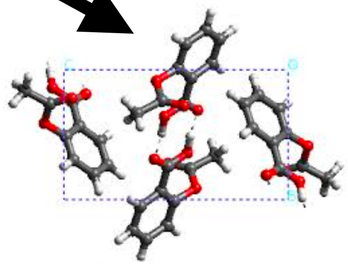
compounds and their fragments



pKa, tautomers, microstates

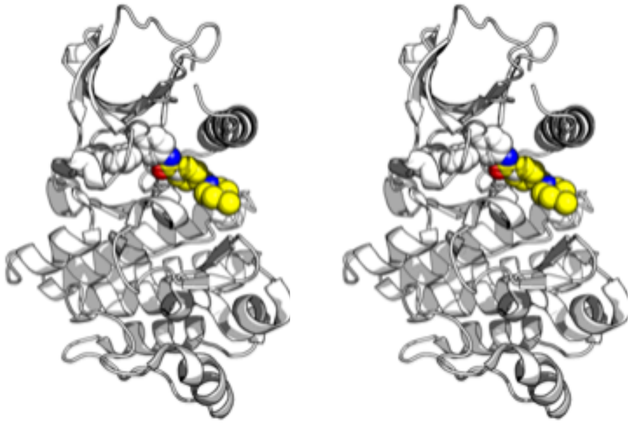


logP, logD, permeability

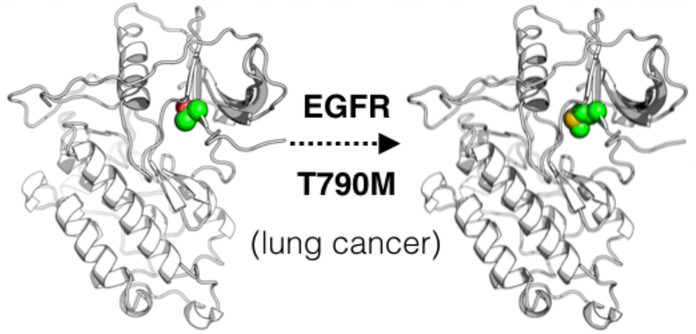


solubilities/ polymorphs

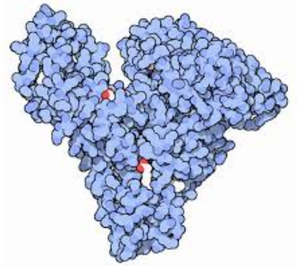
SAMPL AIMS TO EXPLOIT SYNERGIES TO MAXIMIZE OPPORTUNITIES FOR LEARNING



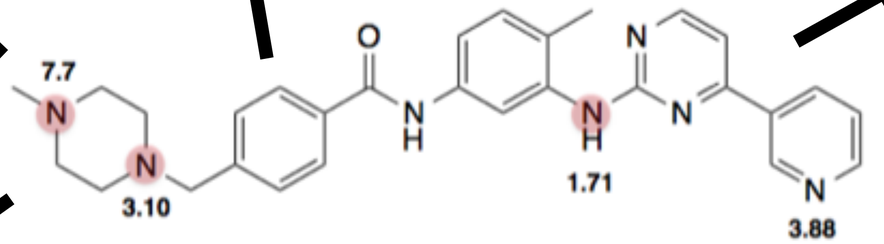
kinases



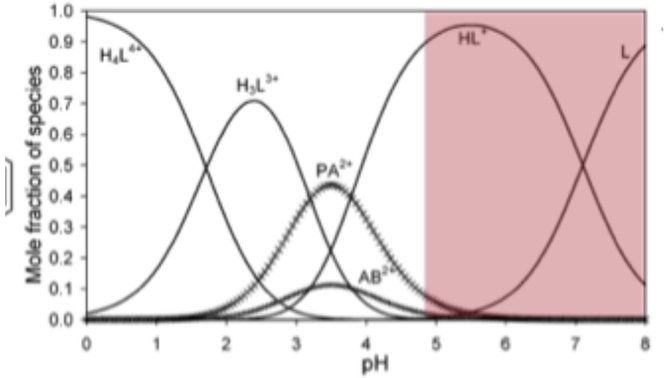
mutations



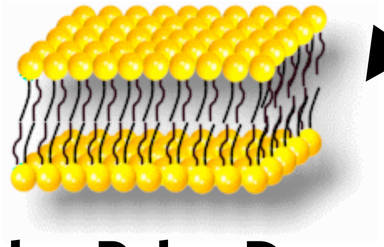
ADME/Tox relevant proteins (e.g. HSA)



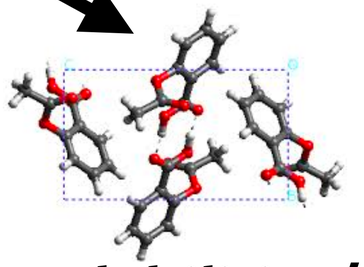
compounds and their fragments



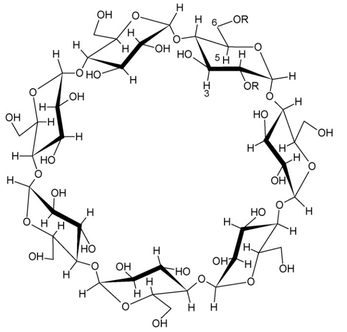
pKa, tautomers, microstates



logP, logD, permeability

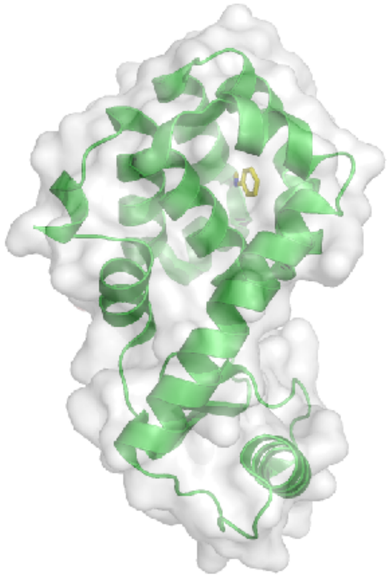


solubilities/ polymorphs



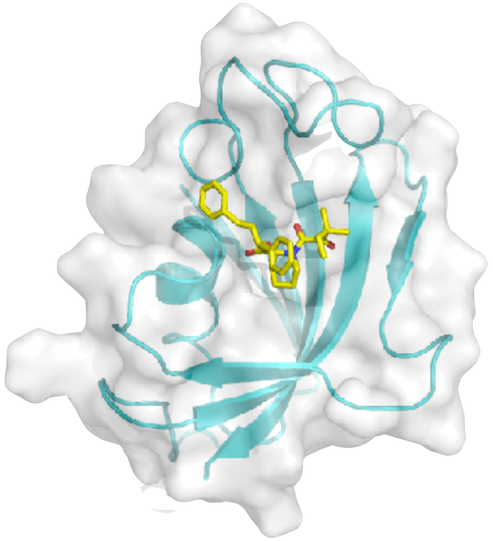
macrocyclic hosts

DEVELOPING NEW MODEL SYSTEMS CAN HELP US ISOLATE/FOCUS ON SPECIFIC CHALLENGES



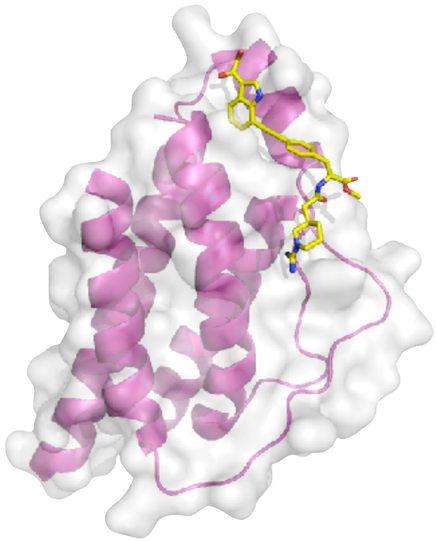
T4 LYSOZYME L99A

small, rigid protein
small, neutral ligands
fixed protonation states
multiple sidechain orientations
multiple ligand binding modes



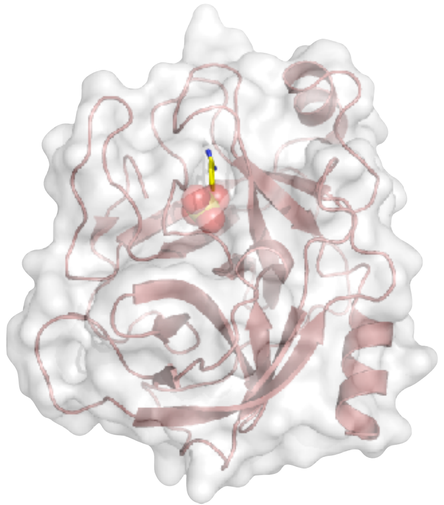
FKBP-12

small, rigid protein
fixed protonation states
larger natural product-like
ligands with rotatable bonds



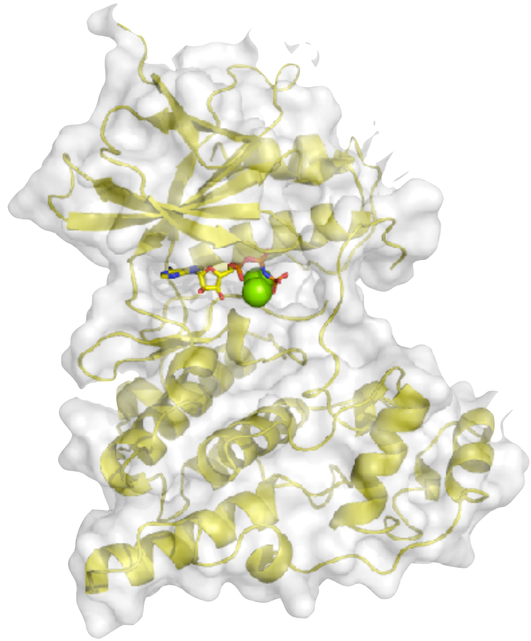
IL-2

small protein
fixed protonation states
some allostery and
binding site plasticity



TRYPSIN

small, rigid protein
small ligands
charged ligands
protonation state changes



KINASES

large protein, multiple conformations
large drug-like ligands, rotatable bonds
multiple protonation states? tautomers?
phosphorylation and activation
peptide substrate?

easy
hard

WHERE DO MODEL SYSTEMS COME FROM?

- Word of mouth ("Hey, you should really look at aspartyl proteases...")
- My old advisor worked on this (T4 lysozyme mutants)
- I got the plasmid from the lab down the hall (chicken Src)
- Everybody else is working on it! (Abl)

WHERE DO MODEL SYSTEMS COME FROM?

- Word of mouth ("Hey, you should really look at aspartyl proteases...")
- My old advisor worked on this (T4 lysozyme mutants)
- I got the plasmid from the lab down the hall (chicken Src)
- Everybody else is working on it! (Abl)

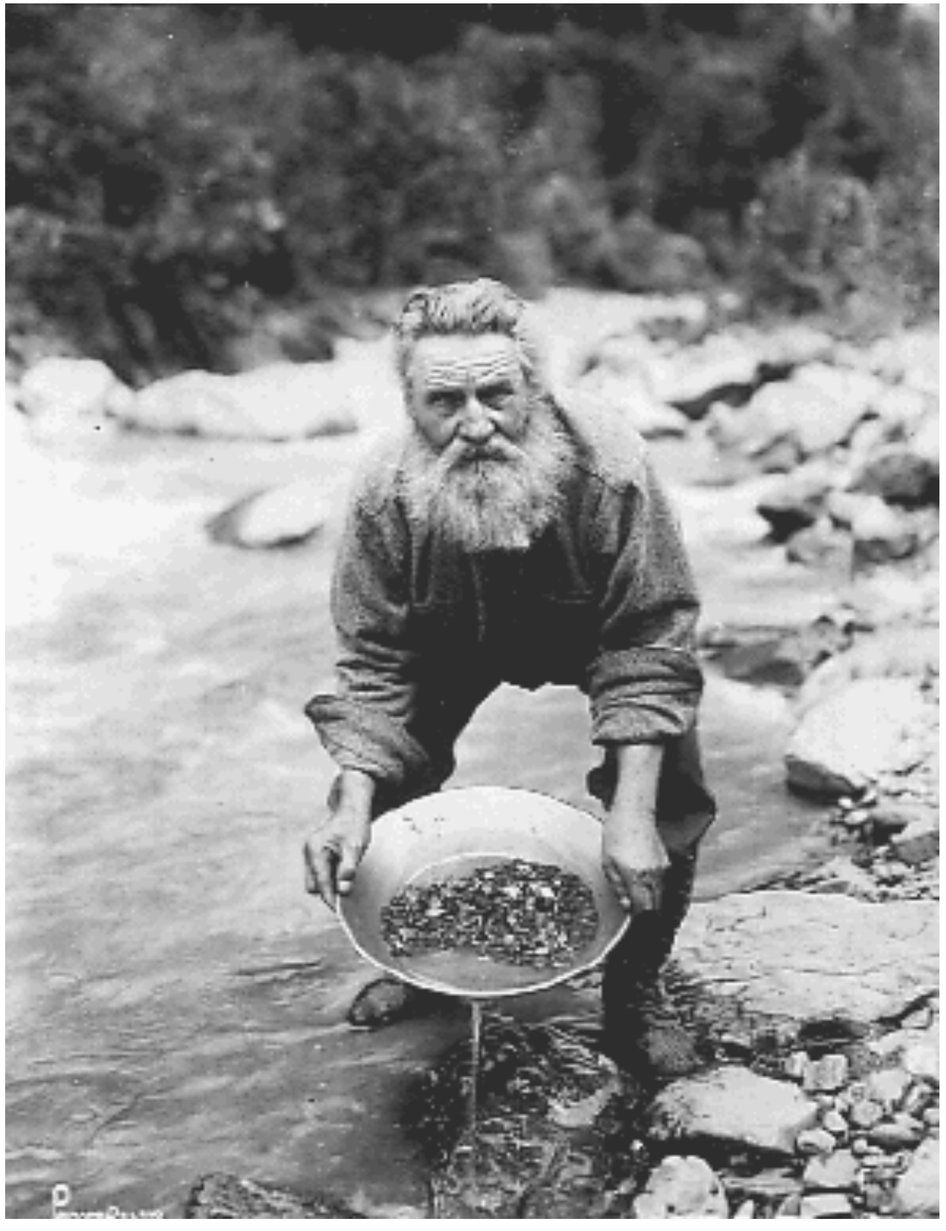
SURELY THERE MUST BE A BETTER WAY!

CAN WE MINE PUBLIC DATASETS FOR GOOD MODEL SYSTEMS?

Desiderata:

- good bacterial expression (for cheap protein production)
- multiple structures available in PDB
- a variety of known ligands available for purchase
- large dynamic range of binding affinities (>3 kcal/mol)
- accessibility to biophysical assays (fluorescence, SPR, ITC)
- known point mutants (e.g. UniProt)
- disease relevance (for funding!)
- properties characteristic of real challenging targets

PANNING FOR MODEL SYSTEMS



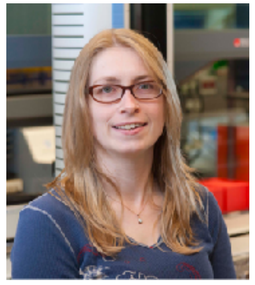
initial set of UniProt IDs

retrieve all UniProt metadata

retrieve all known structures and ligands

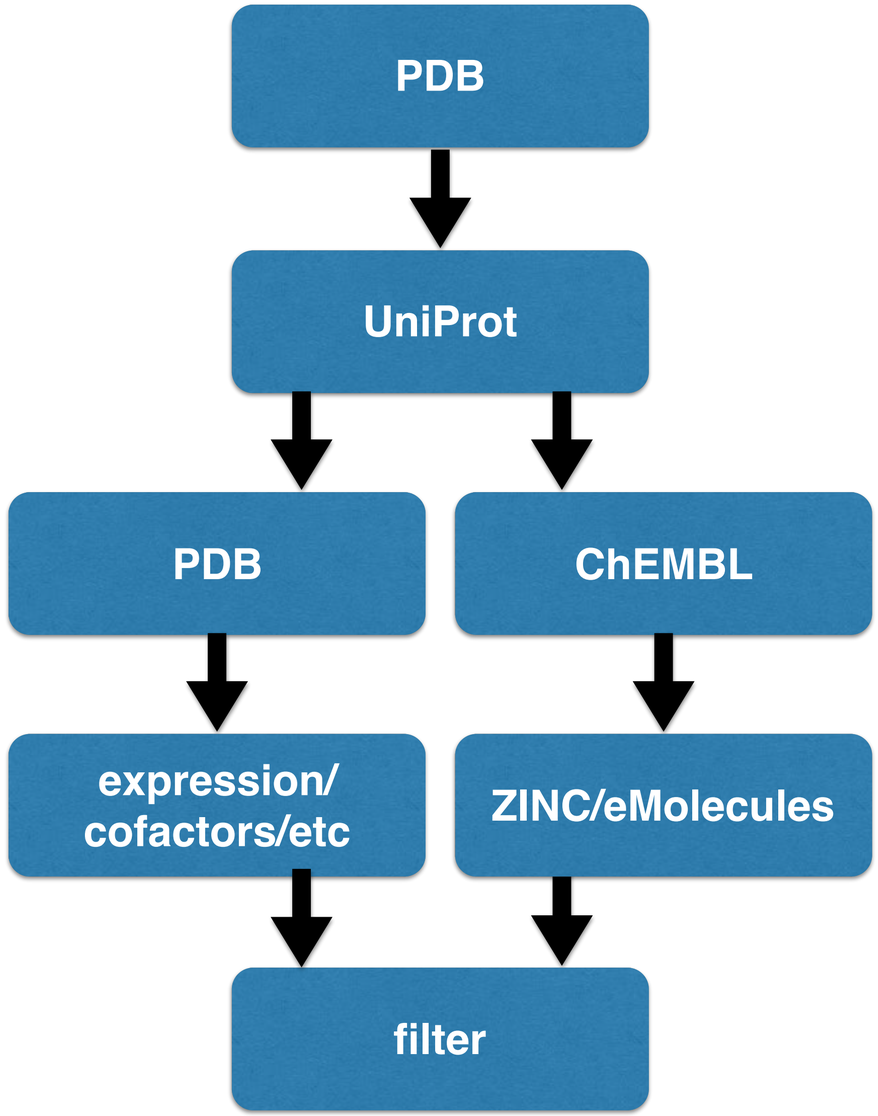
retrieve additional data

filter by criteria of interest



MEHTAP ISIK SONYA HANSON

PANNING FOR MODEL SYSTEMS



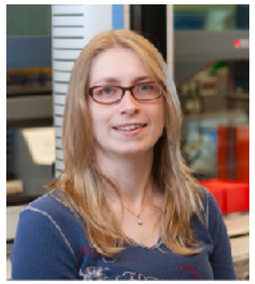
initial set of UniProt IDs

retrieve all UniProt metadata

retrieve all known structures and ligands

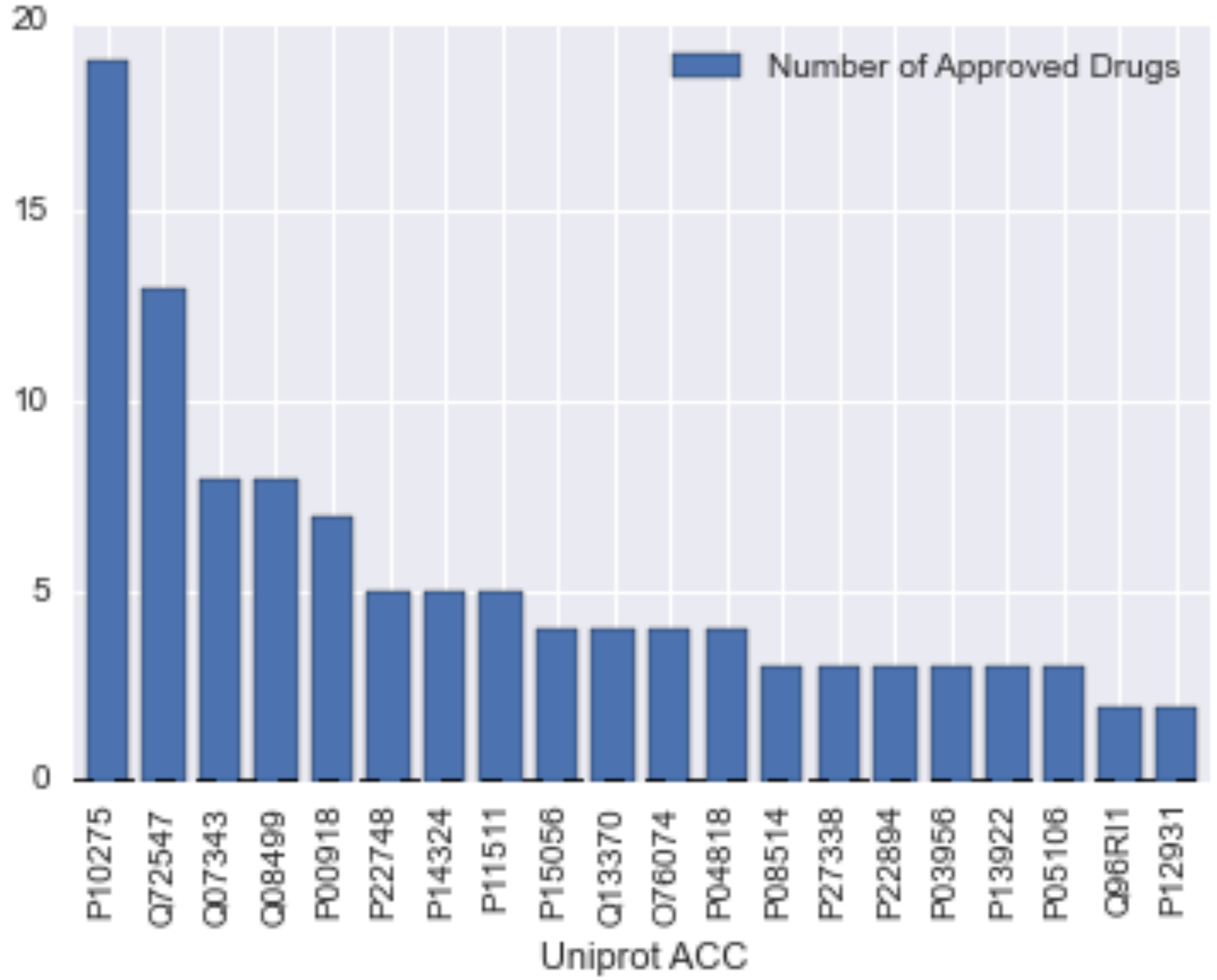
retrieve additional data

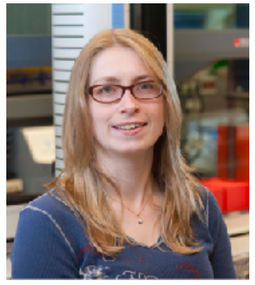
filter by criteria of interest



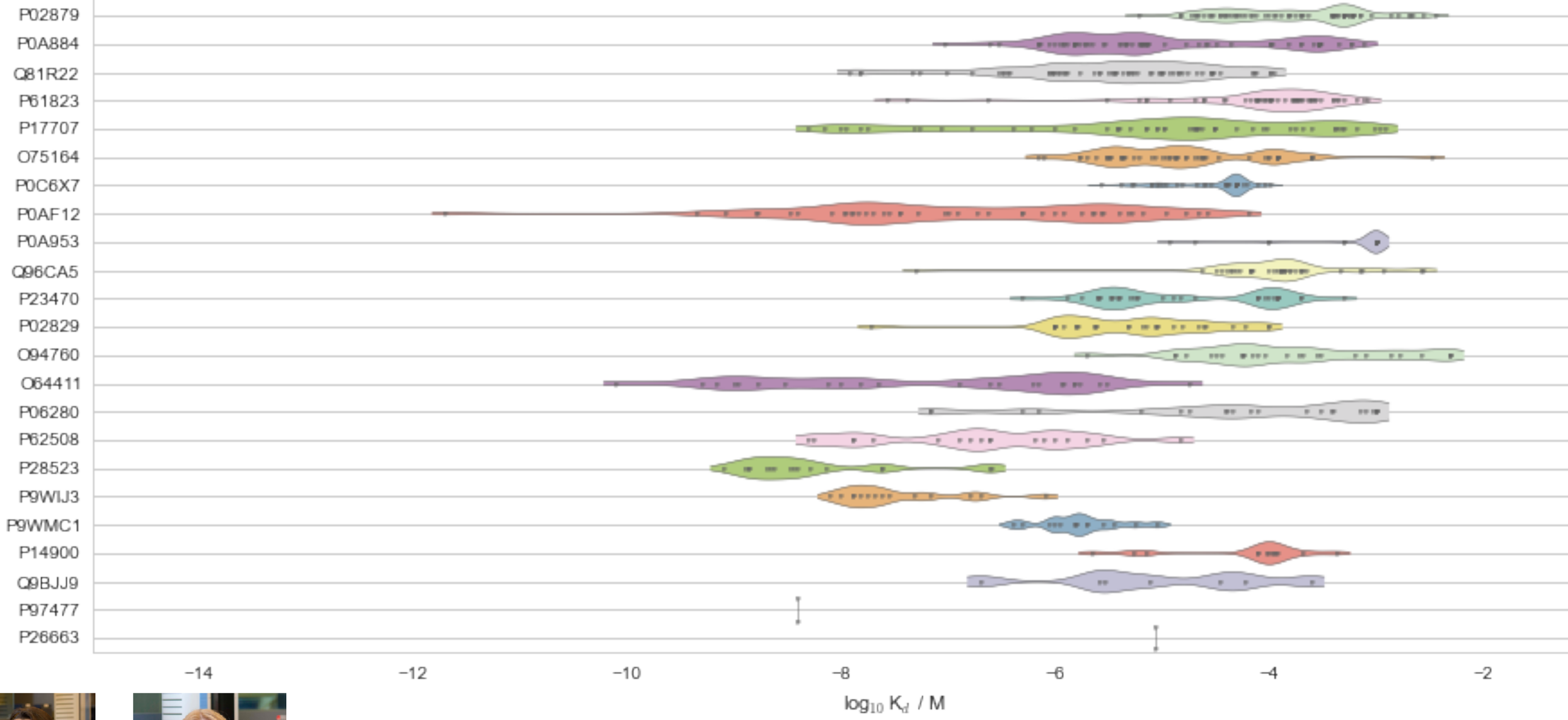
MEHTAP ISIK SONYA HANSON

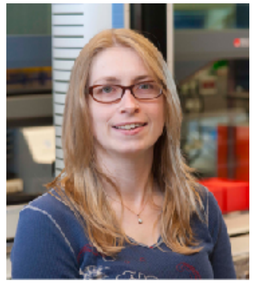
SOME TARGETS HAVE BIOASSAY DATA FOR MULTIPLE FDA-APPROVED DRUGS





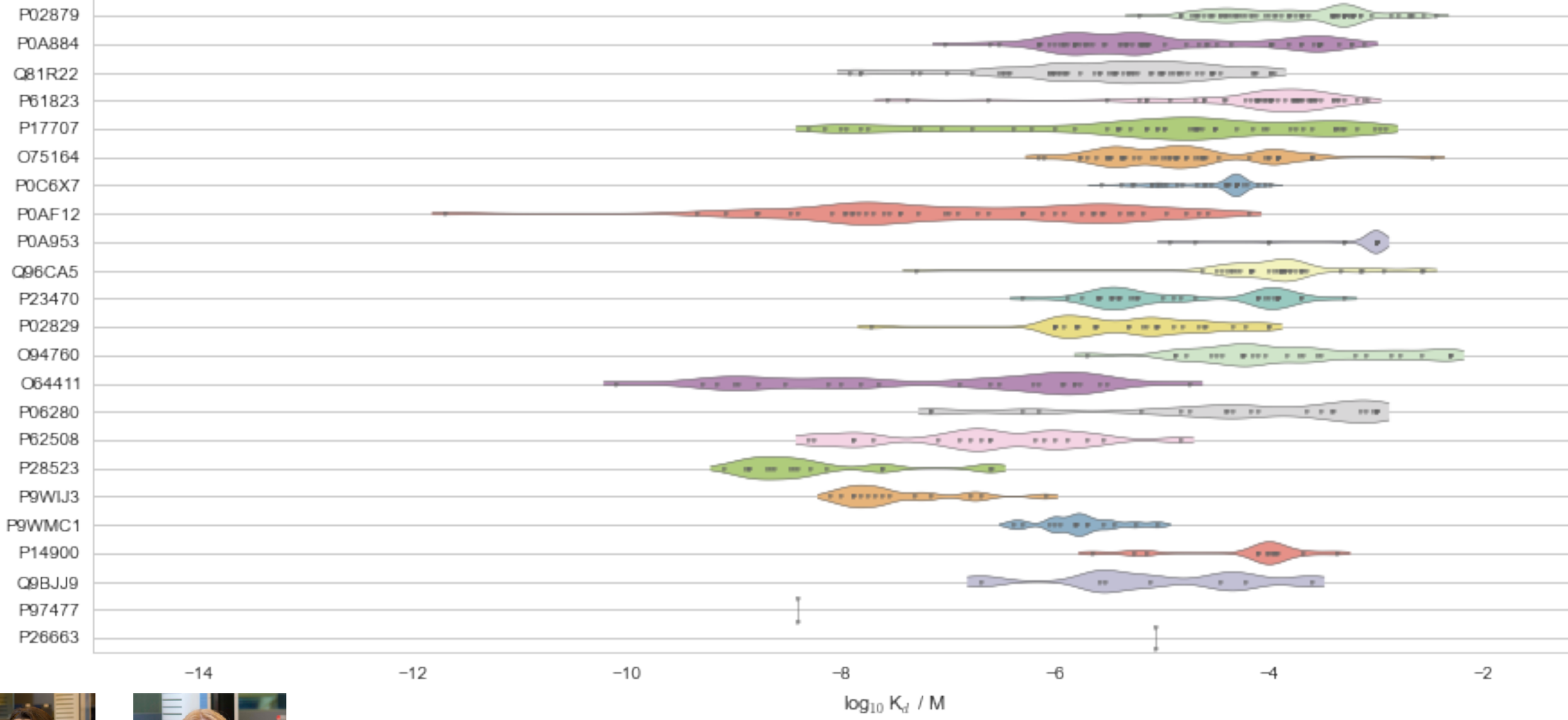
MEHTAP ISIK SONYA HANSON

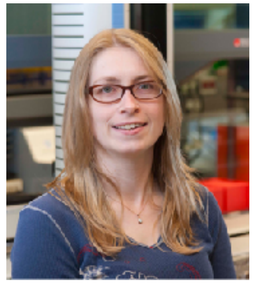




MEHTAP ISIK SONYA HANSON

Many targets have usefully large dynamic ranges of known affinities



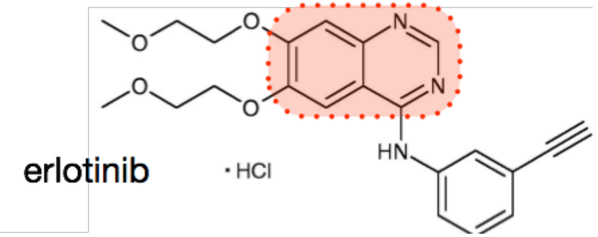


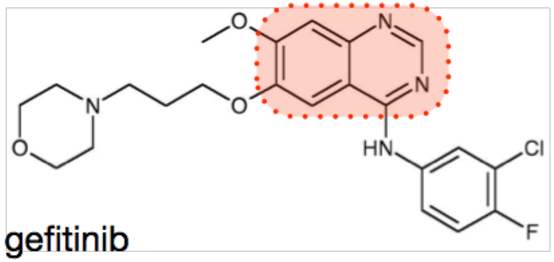
MEHTAP ISIK SONYA HANSON

Many targets have usefully large dynamic ranges of known affinities

CAN WE SEARCH FOR POTENTIAL FLUORESCENT PROBE COMPOUNDS?

Quinazoline scaffolds are often fluorescent



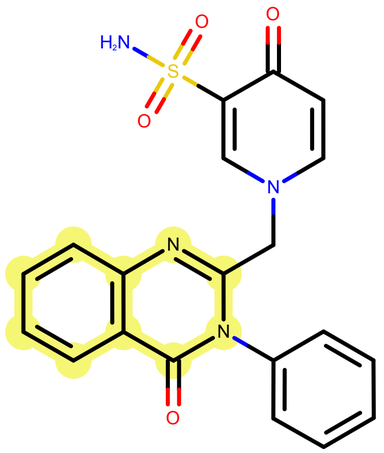


...which can be expressed as a SMARTS query

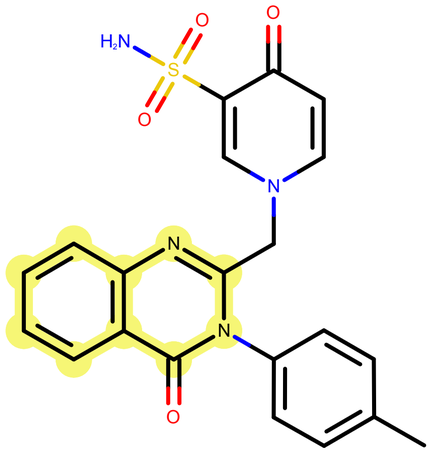
```
c1cccc2c1cncn2
```

Thanks OpenEye!

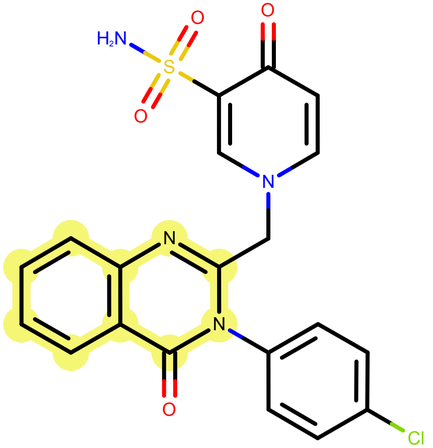
...and used to find some quinazoline scaffold inhibitors of Uniprot P00918 (carbonic anhydrase II) to serve as probes:



567 nM



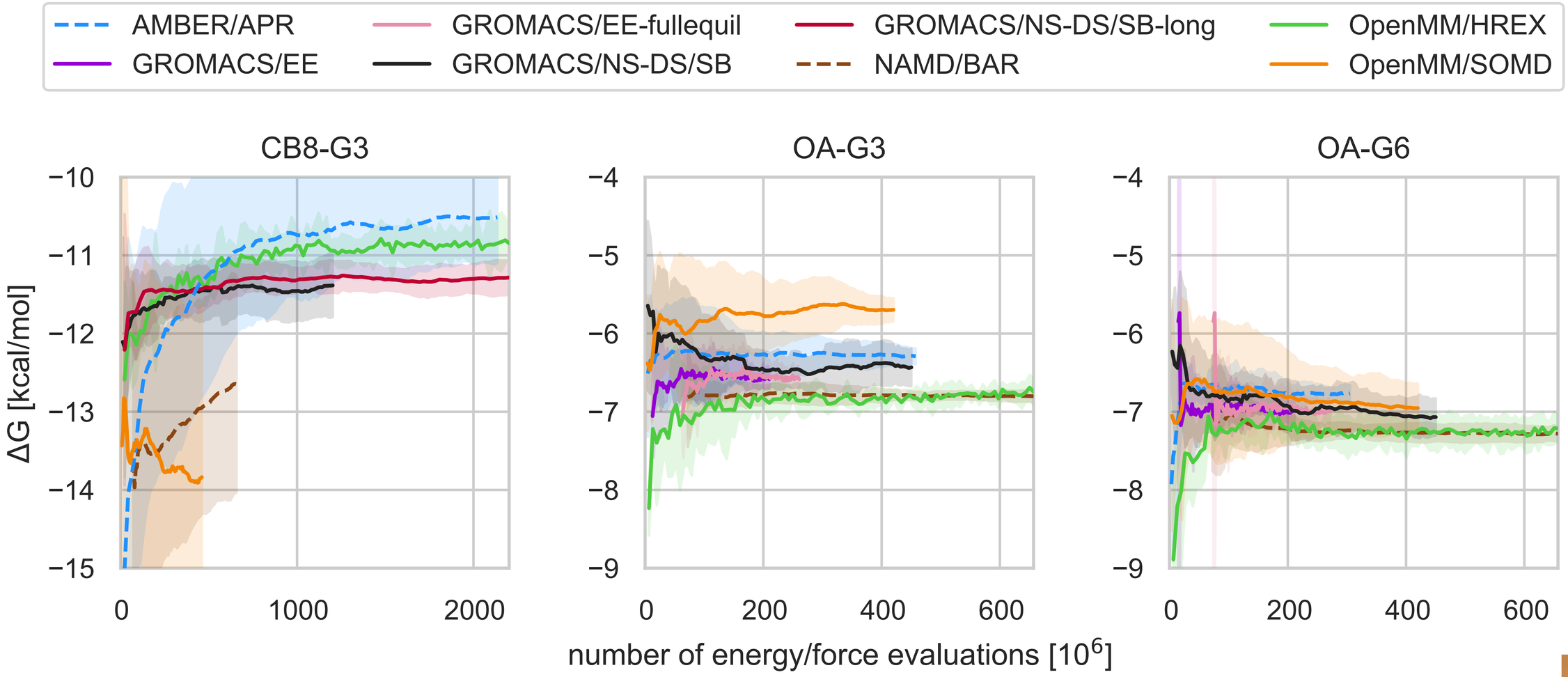
490 nM



700 nM

CAN WE EXPAND THIS SEARCH TO ALL KNOWN FLUORESCENT SCAFFOLDS?

DIFFERENT METHODS CAN'T AGREE ON THE BINDING FREE ENERGY FOR THE SAME FORCE FIELD



How can we make progress when we don't even know where we are?

SAMPL6 SAMPLing challenge
ANDREA RIZZI

MOLSSI IS SPONSORING A MOLECULAR SIMULATION INTEROPERABILITY WORKSHOP

Can we adopt a standard set of force field terms to support in all major simulation packages?

Can we adopt a single way to encode force fields or parameterized molecular systems?

Can we adopt a unifying input standard for initiating molecular simulations?

3-5 NOV 2019 - WILLIAMSBURG, BROOKLYN

THANKS TO D3R AND SAMPL AND THEIR COMMUNITIES

D3R

Michael Gilson (UCSD)
Rommie Amaro (UCSD)
Mike Chiu (UCSD)
D3R community

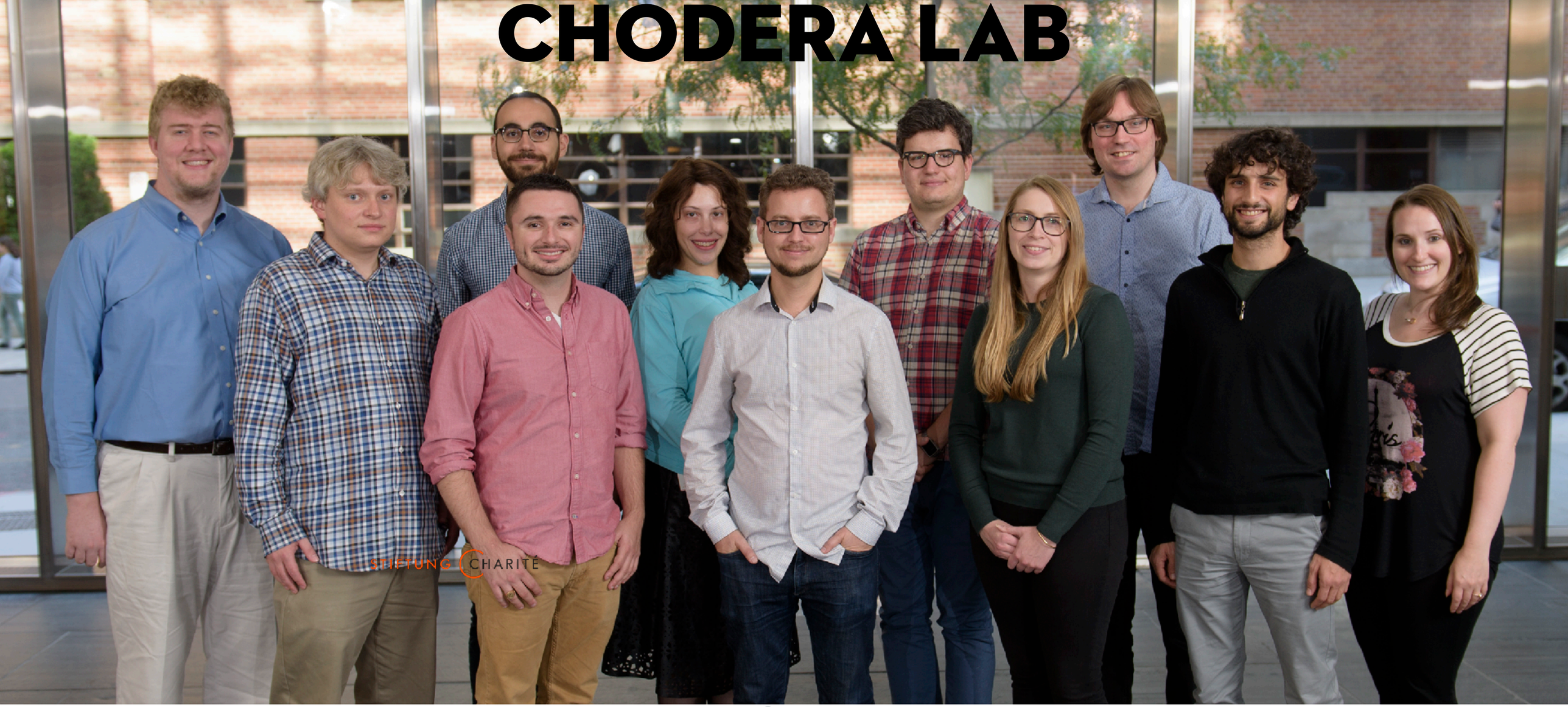
MOLSSI

Daniel Crawford (VT)
Daniel Smith (VT/MolSSI)
Shantenu Jha (Rutgers)

SAMPL

NIH R01 GM124270 (SAMPL grant)
David Mobley (UCI)
Danielle Bergazin (UCI)
Caitlin Bannan (now at OpenEye)
Mehtap Isik (MSKCC)
Andrea Rizzi (MSKCC)
Bas Rustenburg (MSKCC)
Bruce Gibb and group **(Tulane)**
Lyle Isaacs and group **(University of Maryland)**
Genentech, Merck, GSK for data
The **SAMPL community**

CHODERA LAB



National Institutes of Health

SCHRÖDINGER.

PARKER INSTITUTE for CANCER IMMUNOTHERAPY

Gerstner FAMILY FOUNDATION

STARR CANCER CONSORTIUM

Open Force Field Consortium

XtalPi

CYCLE FOR SURVIVAL

- Scientific Advisory Board, OpenEye Scientific

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