STREAMLINING MOLECULAR SIMULATIONS DATA

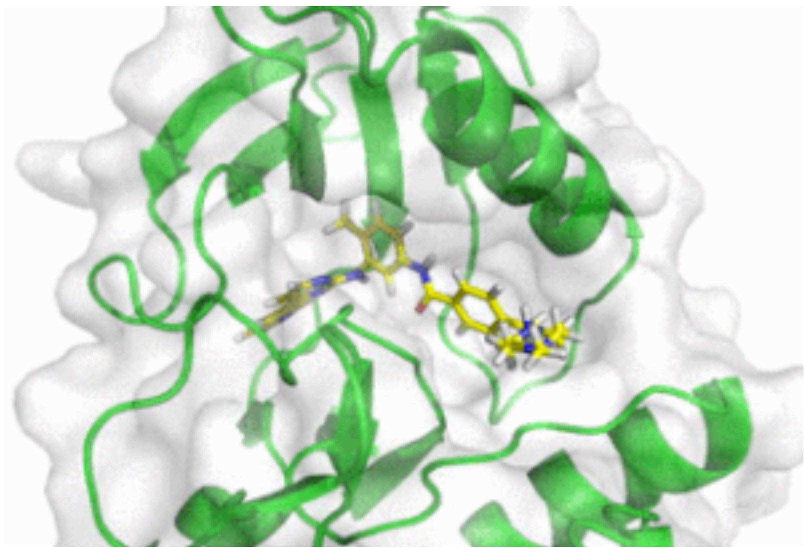
John D. Chodera
MSKCC Computational Biology Program
http://www.choderalab.org

DISCLOSURES:
Scientific Advisory Board, OpenEye Scientific
All funding sources listed at http://choderalab.org/funding

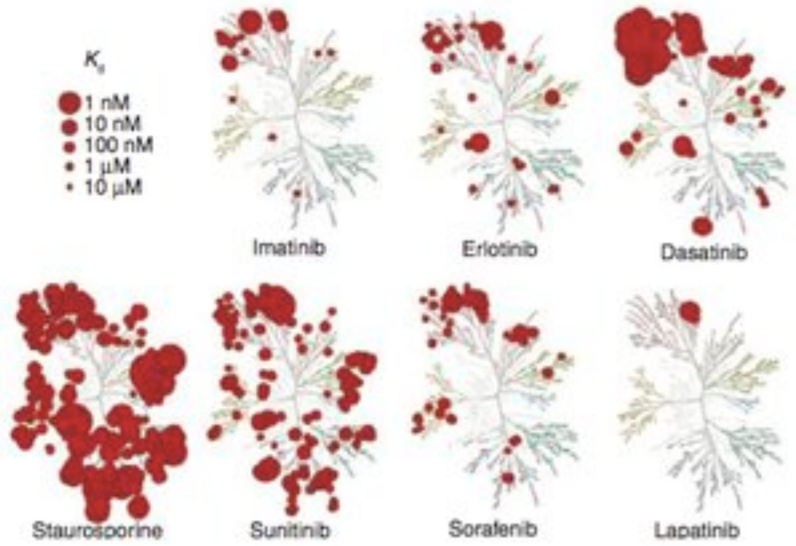
BioExcel Workshop 2018: Sharing Data from Molecular Simulations - Stockholm

CHODERA LAB

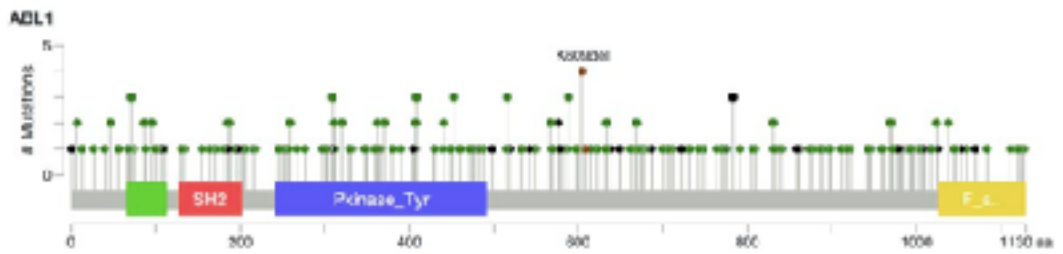
HOW CAN COMPUTATIONAL BIOPHYSICS PLAY A MAJOR ROLE IN THE ERA OF CANCER GENOMICS?



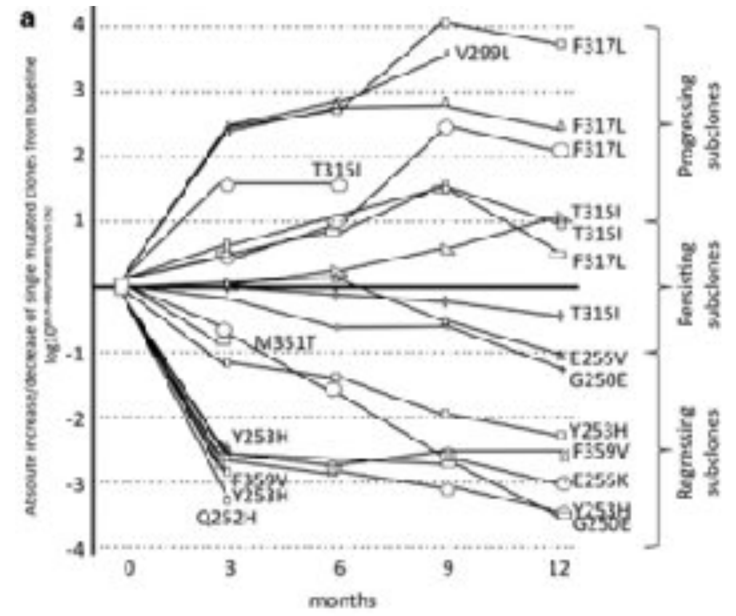
SELECTIVE INHIBITOR DESIGN: TARGETS/ANTITARGETS



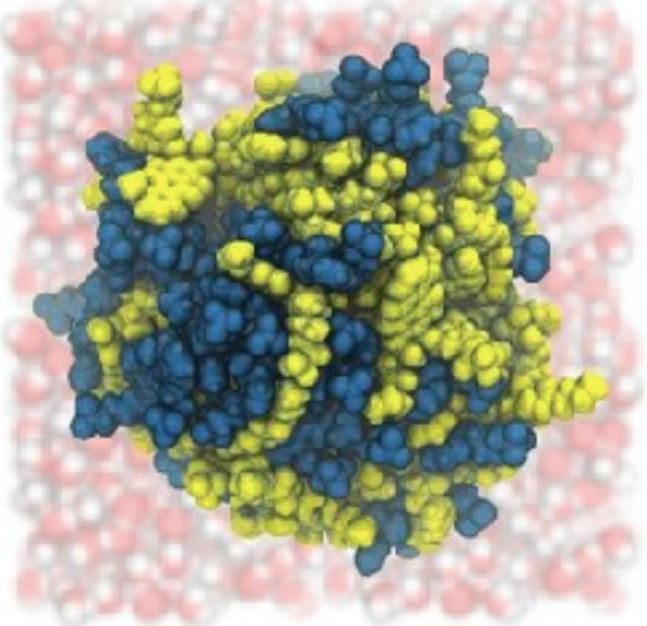
KINASE INHIBITOR SELECTIVITY



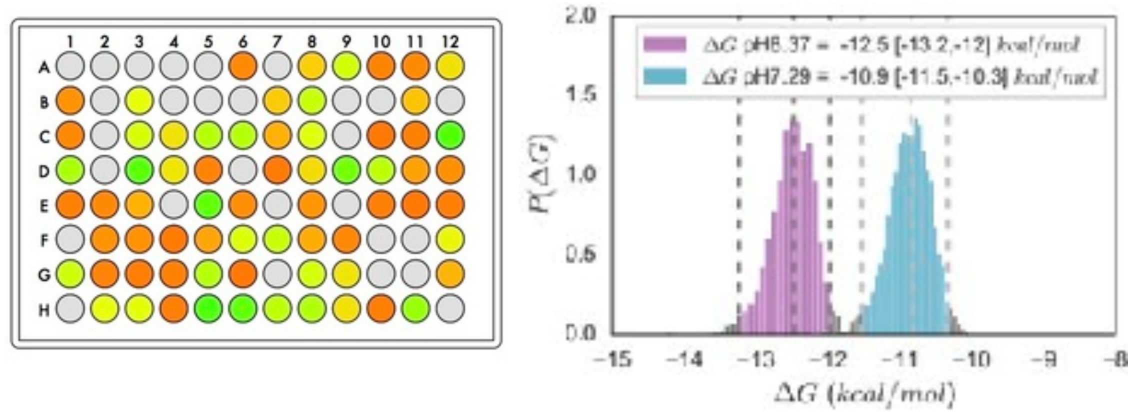
PREDICTING DRUG SENSITIVITY/RESISTANCE



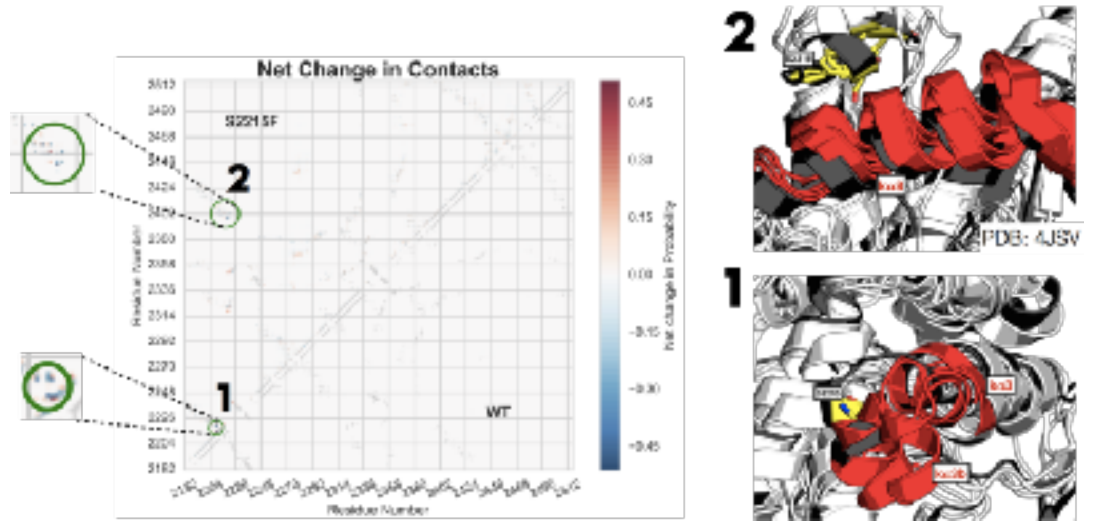
ANTICIPATING DRUG RESISTANCE



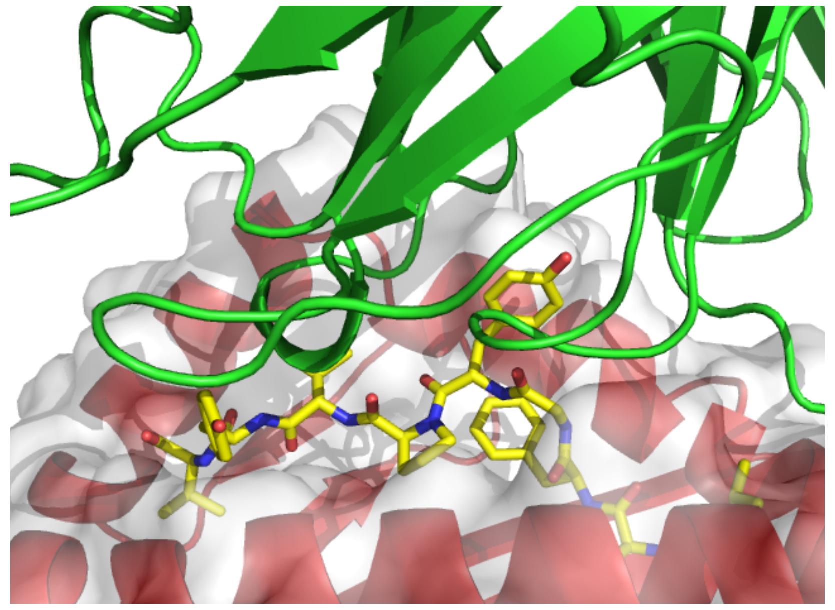
NOVEL DRUG DELIVERY MODALITIES



AUTOMATED BIOPHYSICAL ASSAYS AND INFERENCE



MECHANISMS OF ONCOGENIC ACTIVATION



CANCER IMMUNOTHERAPY

MOLECULAR SIMULATION IS FACING 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**

Predictive challenges (e.g. D3R, SAMPL) end up **testing unrelated choices** (such as biomolecular setup pipeline) rather than core scientific methods

BIOMOLECULAR SYSTEM PREPARATION REQUIRES MANY CHOICES

Before beginning, we have to make many **decisions** about structural data, and generally have little idea how sensitive our results are to our choices:

* Which **structure**(s) do we want to use? How do we use multiple structurs?

* What do we do about **missing structural details** (loops, termini, and residues)?

* How do we treat **modified residues**? (PTMs, non-natural amino acids, covalent ligands)

* What do we do with **cofactors**, prosthetic groups, or structural ions?

* What about crystallographic **waters**?

* How do we treat **non-biological features**, such as crystal contacts, domain swaps, or other non-biological structural features?

WHAT ARE WE EVALUATING IN BLIND COMPETITIONS?

evaluating the **driver**

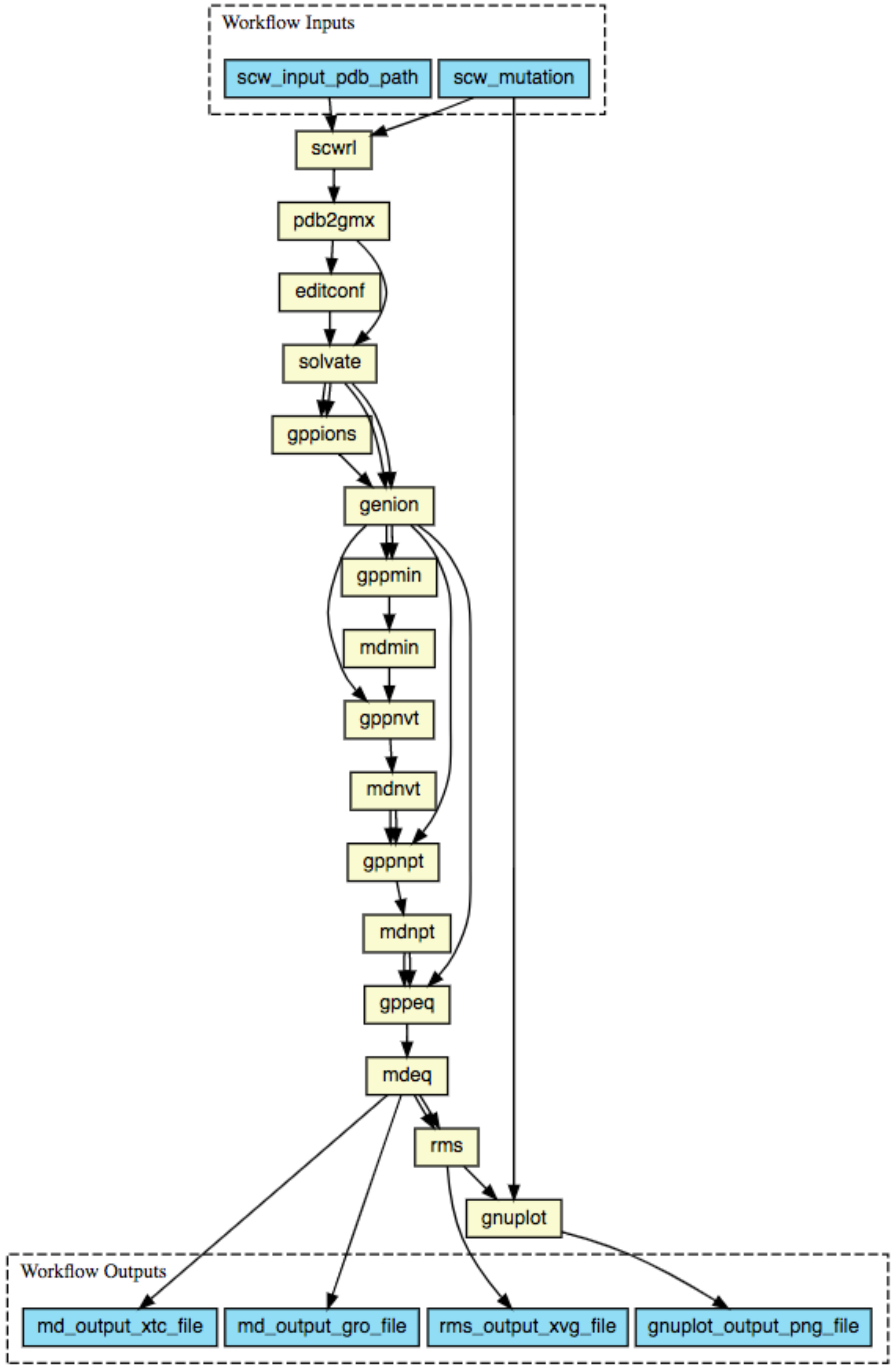
evaluating the **technology**

Need to separate capabilities of technology from skill of driver

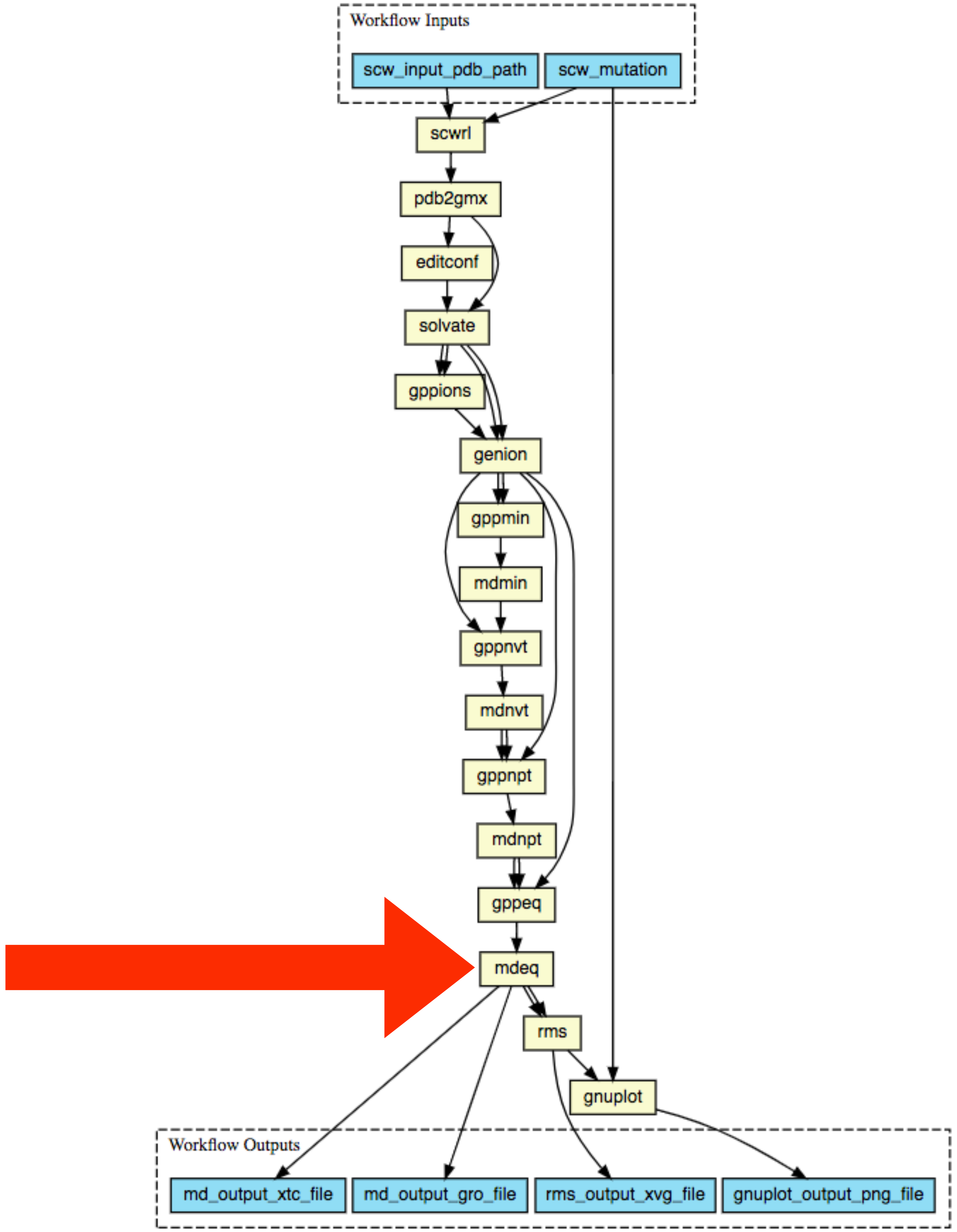
ENABLING FOCUS ON KEY SCIENCE

Academic scientists developing new methodologies would generally like to **focus their creative efforts on a specific part of the overall simulation pipeline**, but are often forced to build everything from scratch

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

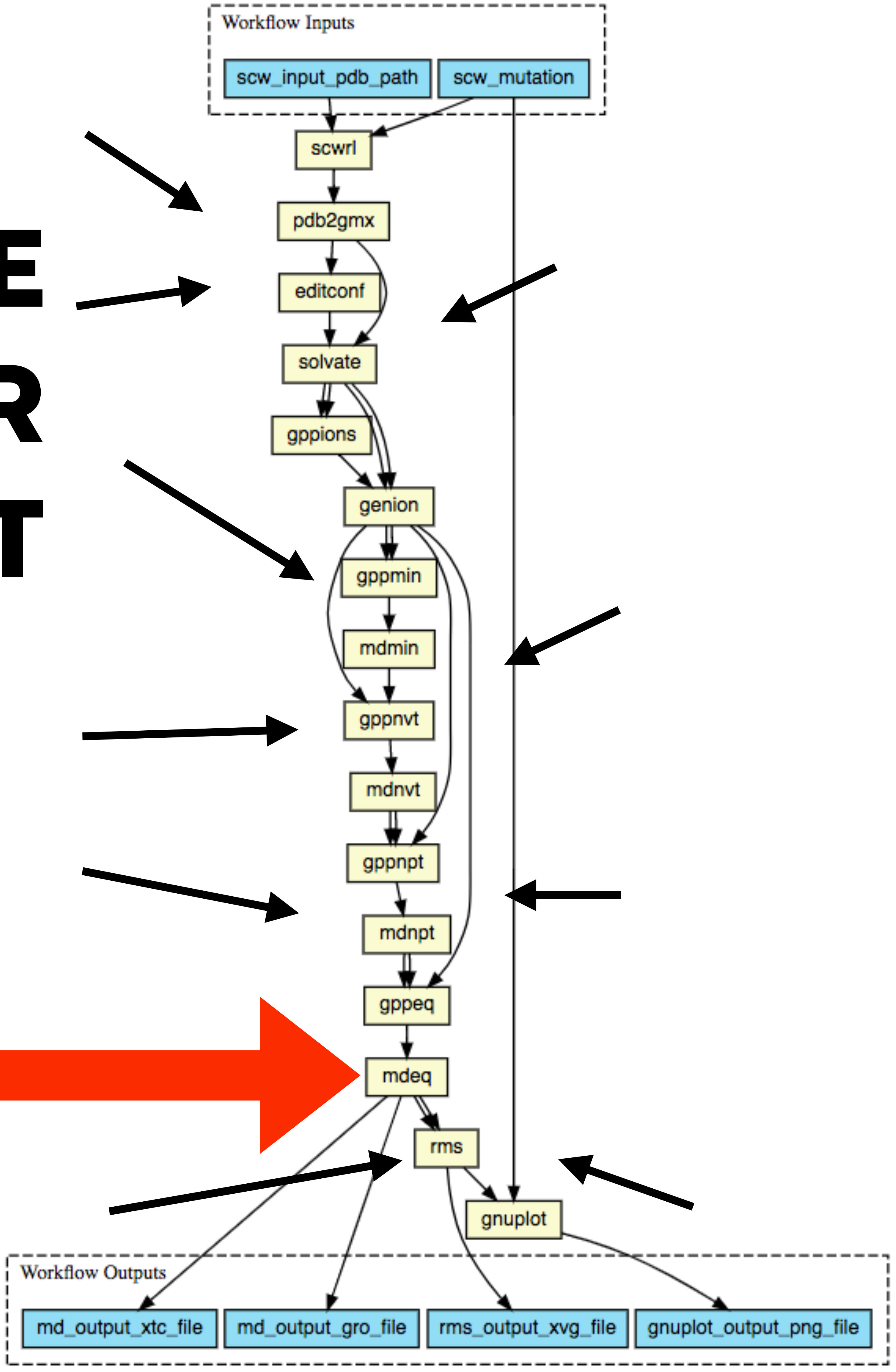

Workflow Inputs
scw_input_pdb_path
scw_mutation
scwrl
pdb2gmx
editconf
solvate
gppions
genion
gppmin
mdmin
gppnvt
mdnvt
gppnpt
mdnpt
gppeq
mdeq
rms
gnuplot
Workflow Outputs
md_output_xtc_file
md_output_gro_file
rms_output_xvg_file
gnuplot_output_png_file

THE SCIENCE I'M INTERESTED IN



EVERYTHING ELSE I NEED IN ORDER TO RUN MY BIT

THE SCIENCE I'M INTERESTED IN



http://bioexcel.eu/

REPRODUCIBILITY

Reproducing work from a computational chemistry paper is **almost impossible**, which **minimizes opportunities for learning and improvement** by building on existing work and carrying it further with new ideas

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

DEPLOYMENT

Translating academic research software into industry application is **extremely hard** if not impossible for reasons of code quality, robustness, interoperability, and user-friendliness

e.g.: Merck KGaA paid MSKCC to fly a postdoc out once a quarter to do software updates, even though we try hard to make code conda-installable

TRAINING

Facing exodus of talent due to retirements from the Baby Boomer generation

Need **better tools to train the next generation of computational chemists** (which we're in danger of losing to machine learning and data science)

FUNDING

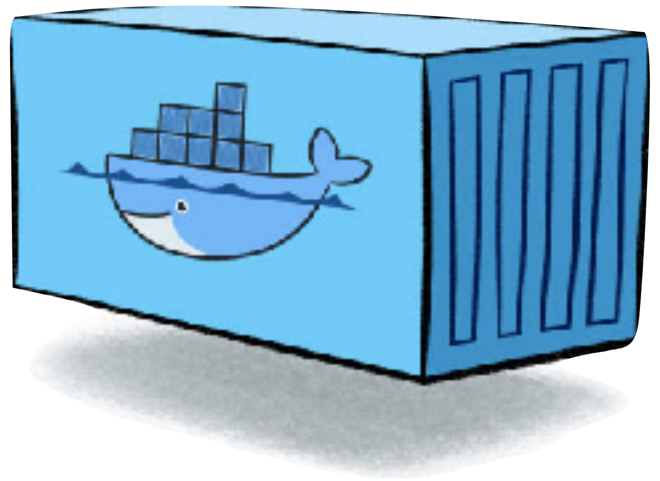
Industry and federal funding agencies tired of investing $ in software or research that is not useful to them or others

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

WORKFLOWS ARE THE SOLUTION...

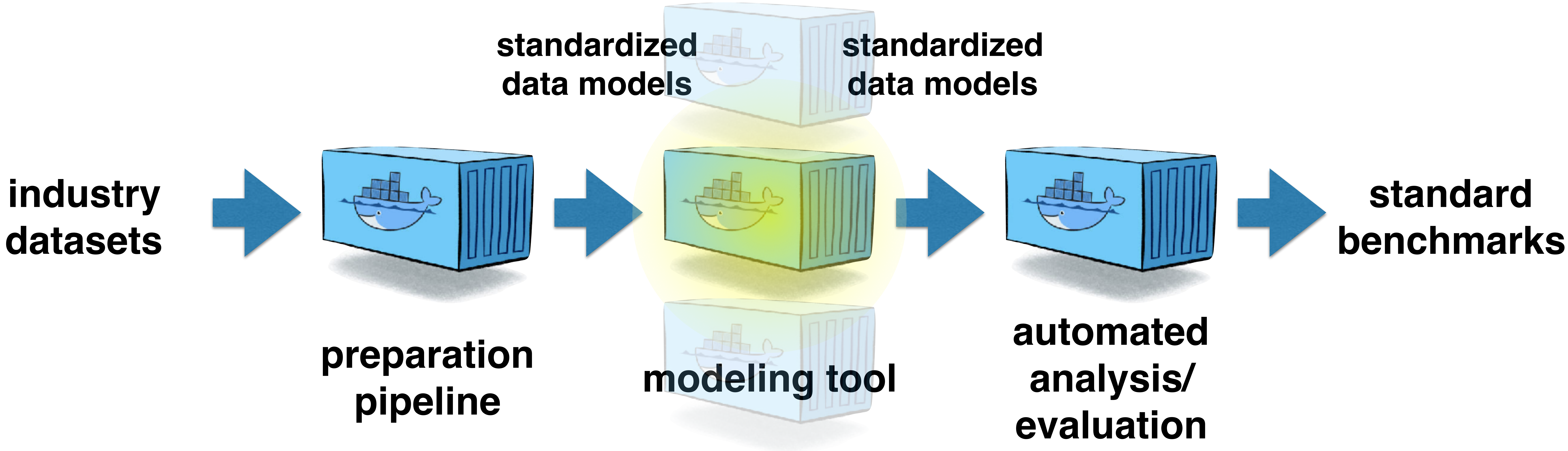
Workflows (and the machinery to support them) can address many of these issues:

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

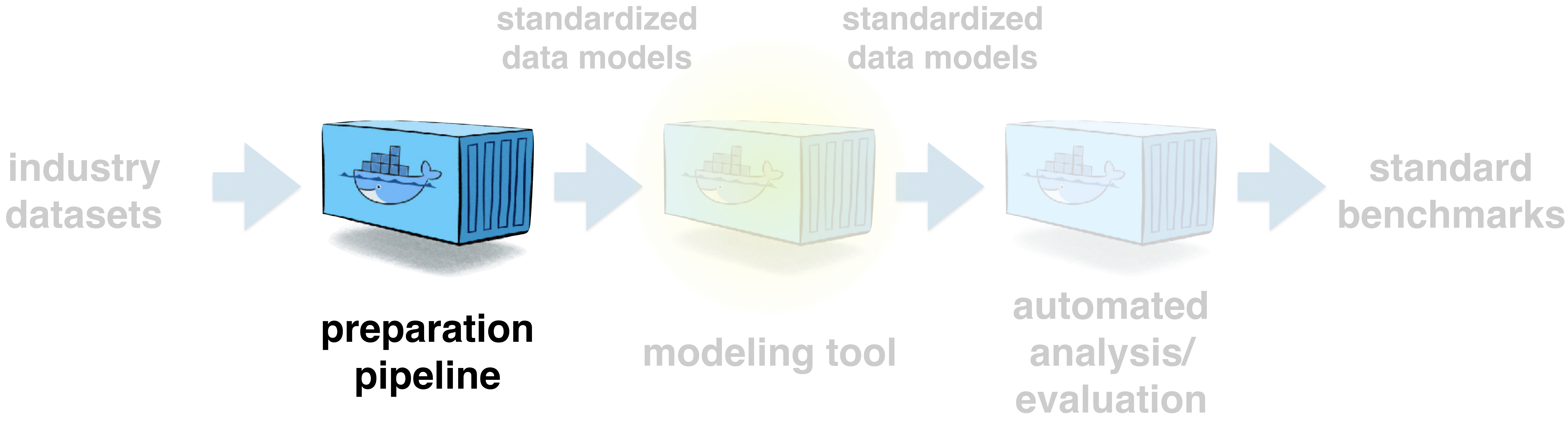


...but this workshop is **not** about workflows, it's about the **standards** or **common data models** required to enable them.

WORKFLOWS USING BEST PRACTICES WOULD ALLOW US TO EVALUATE THE TECHNOLOGY



OPEN PREPARATION PIPELINES COULD CAPTURE COMMUNITY-DRIVEN BEST PRACTICES



industry datasets

standardized data models

standardized data models

preparation pipeline

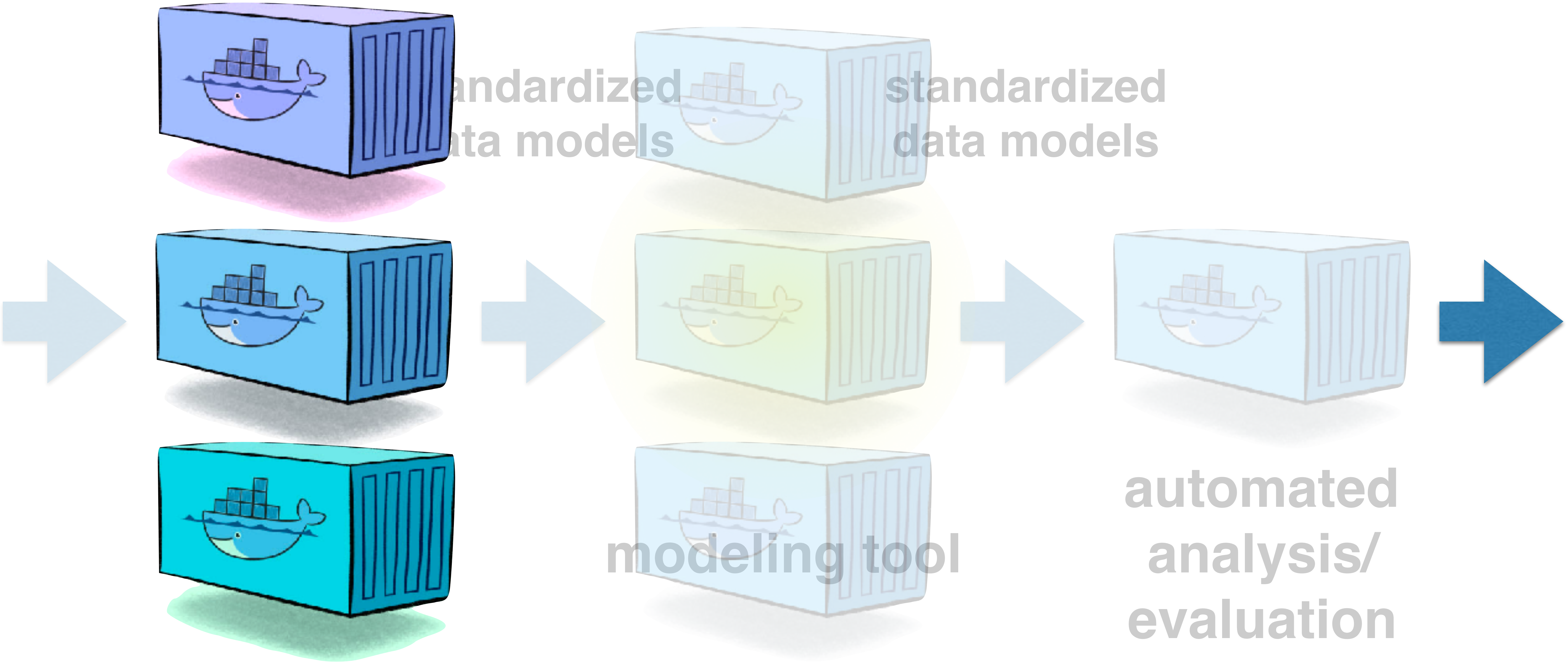
modeling tool

automated analysis/ evaluation

standard benchmarks

docker

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



industry datasets

andardized ata models

standardized data models

modeling tool

automated analysis/ evaluation

standard benchmarks

preparation pipeline variations

THIS REQUIRES STANDARDIZED DATA INTERCHANGE FORMATS

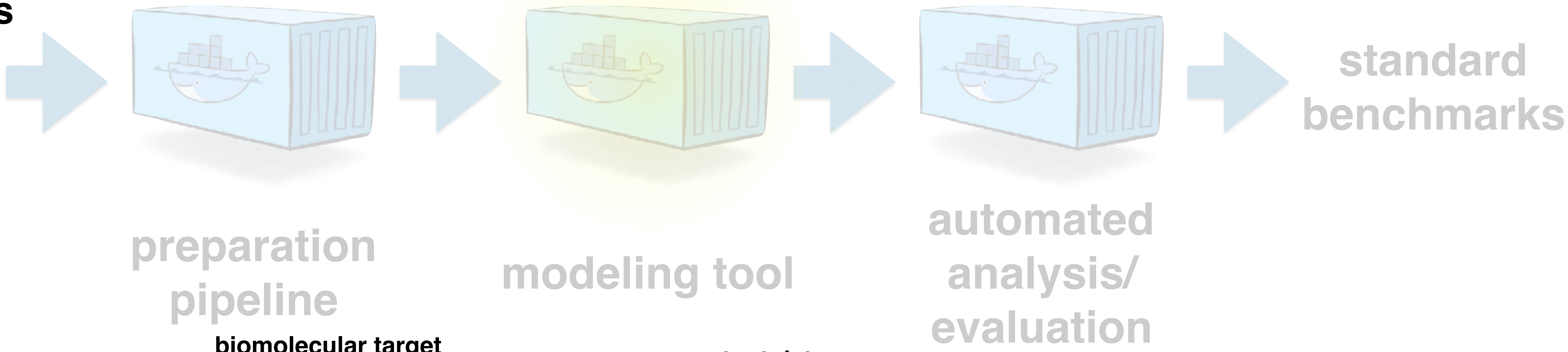
standardized data models

standardized data models

protein constructs

assay conditions

molecules



industry datasets

preparation pipeline

modeling tool

automated analysis/ evaluation

standard benchmarks

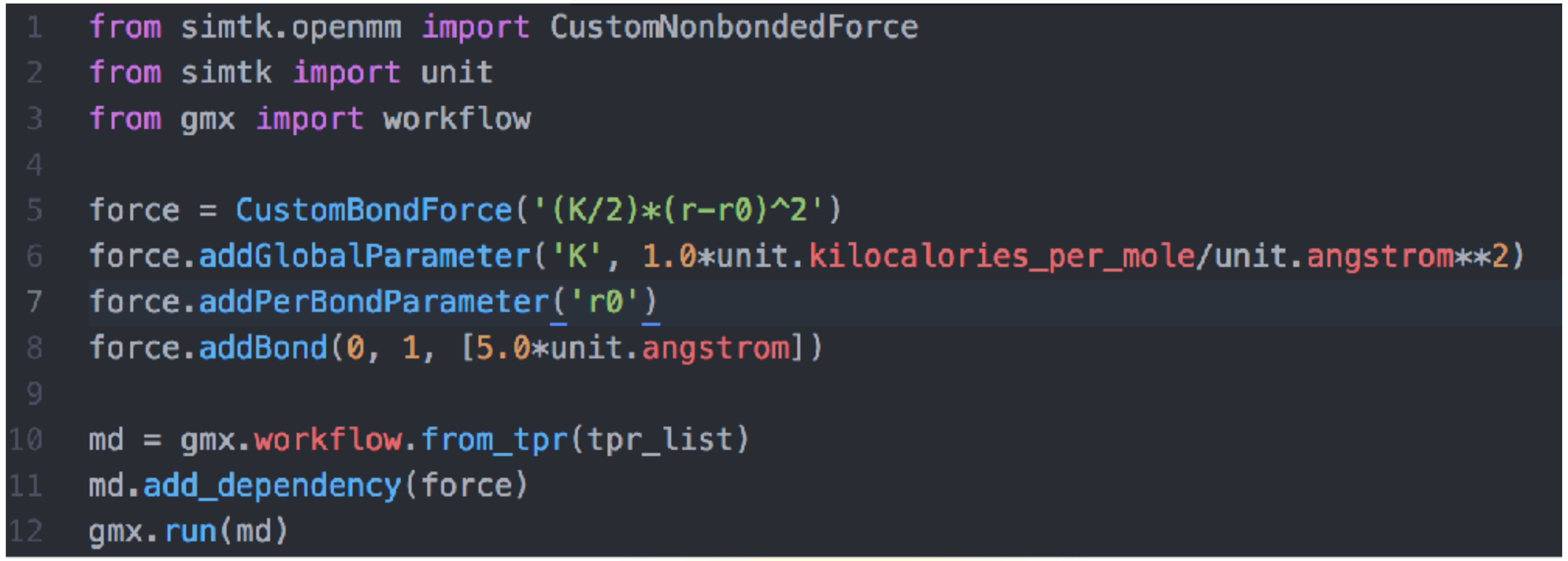
biomolecular target
replace aging PDB format
handle charges, parameters, etc.
robust open source readers/writers

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

output data
trajectories
computed physical properties
binding poses
predicted affinity/assay data
predict confidence/uncertainties
exception logging

assessment data
standard representations
standard assessments
standardized uncertainty analysis

Fine-grained: What if we could import components of different simulation packages and use them together because they share a data model?

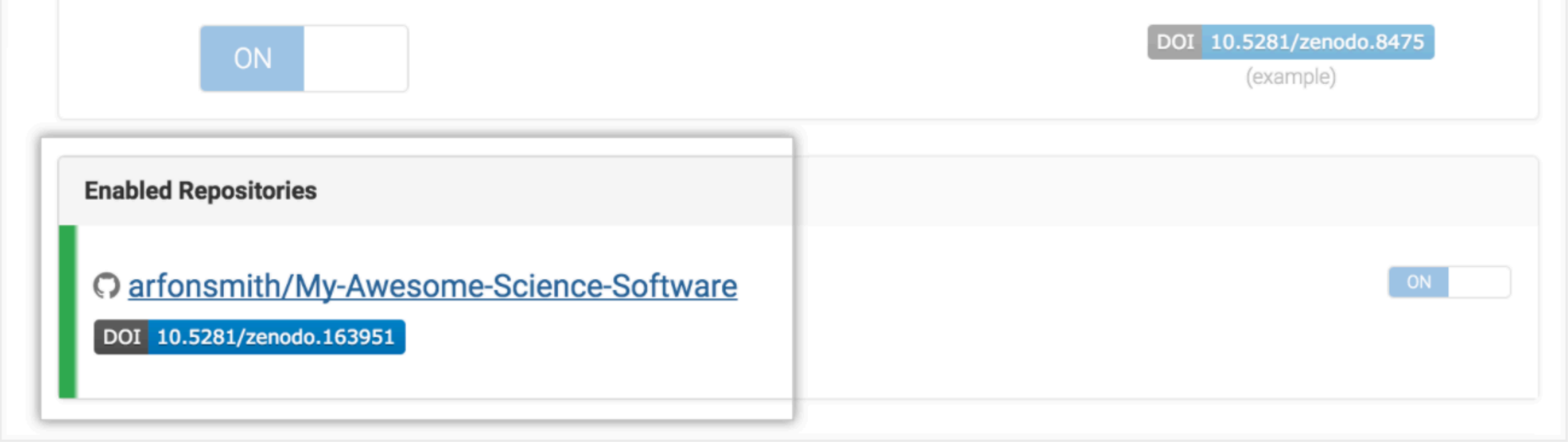


```
from simtk.openmm import CustomNonbondedForce
from simtk import unit
from gmx import workflow

force = CustomBondForce('(K/2)*(r-r0)^2')
force.addGlobalParameter('K', 1.0*unit.kilocalories_per_mole/unit.angstrom**2)
force.addPerBondParameter('r0')
force.addBond(0, 1, [5.0*unit.angstrom])

md = gmx.workflow.from_tpr(tpr_list)
md.add_dependency(force)
gmx.run(md)
```

Coarse-grained: 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?



ON

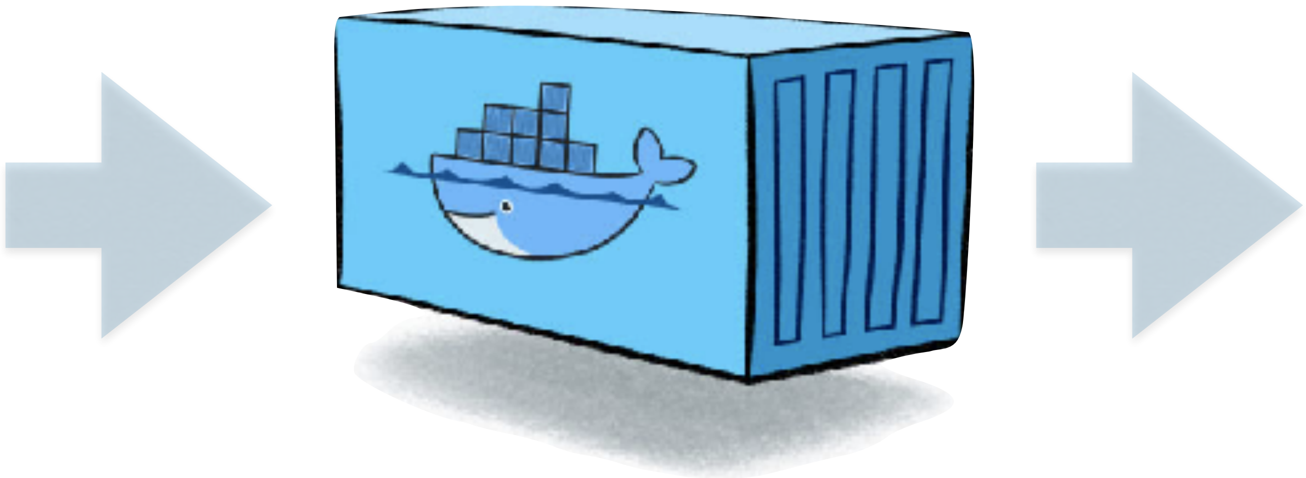
DOI 10.5281/zenodo.8475
(example)

Enabled Repositories

arfonsmith/My-Awesome-Science-Software

DOI 10.5281/zenodo.163951

ON



HOW DO WE DESCRIBE THE SYSTEM WE WANT TO SIMULATE?

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.

SOME STANDARDS AND DATA SOURCES TO BE AWARE OF

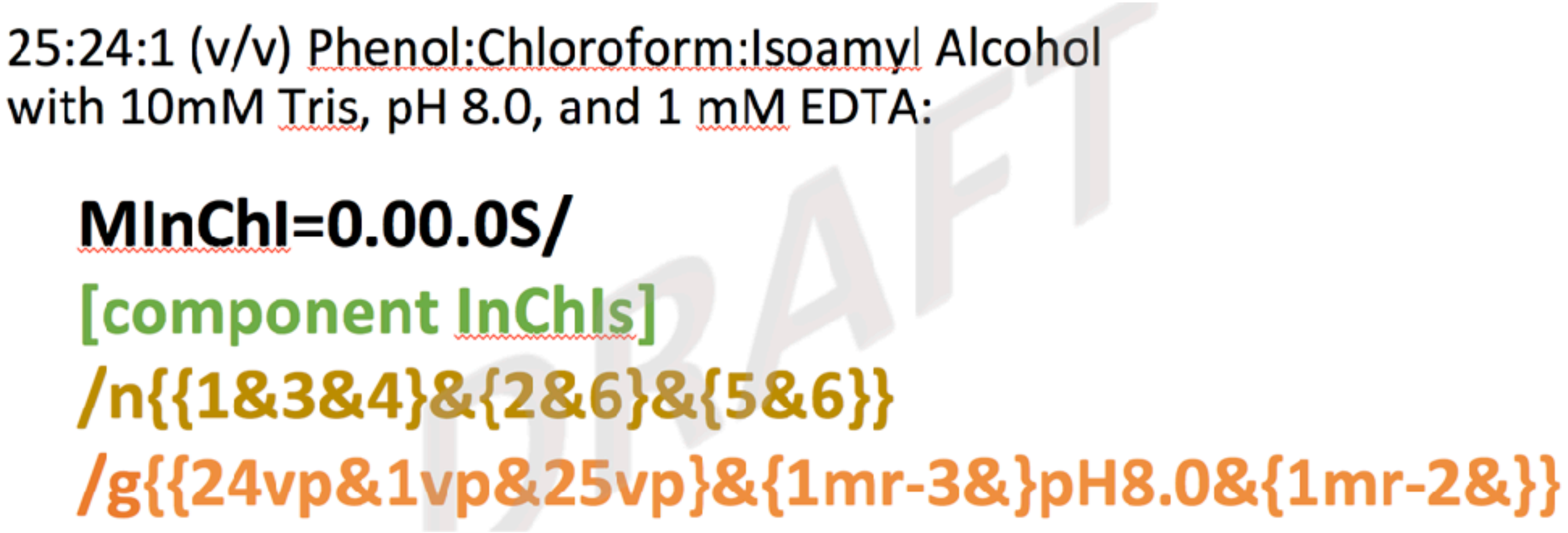
UNIPROT
http://uniprot.org

Standard protein sequence/variant database

SMILES and **InChI**

Standard small molecule representations

Mixture InChI (NIST)



25:24:1 (v/v) Phenol:Chloroform:Isoamyl Alcohol
with 10mM Tris, pH 8.0, and 1 mM EDTA:

MInChI=0.00.0S/
[component InChIs]
/n{{1&3&4}&{2&6}&{5&6}}
/g{{24vp&1vp&25vp}&{1mr-3&}pH8.0&{1mr-2&}}

ISO 11238 (used by FDA in GSAS)

Data elements and structures for the unique identification and exchange of regulated information on substances

Is there anything out there we can already make use of?

THE OPEN FORCE FIELD CONSORTIUM:
BETTER BIOMOLECULAR MODELING TOOLS THROUGH OPEN SOURCE, OPEN DATA, AND OPEN SCIENCE

INDUSTRY

BAYER

gsk

Pfizer

J&J

Roche

NOVARTIS

BASF

MERCK

COORDINATING INTERMEDIARY

NSF

MolSSI

MOLECULAR SOFTWARE SCIENCES INSTITUTE

coordination of funding

http://openforcefield.org

ACADEMIC

CHRISTOPHER BAYLY
OPENEYE SCIENTIFIC

JOHN CHODERA
SLOAN KETTERING INSTITUTE

MICHAEL GILSON
UNIVERSITY OF CALIFORNIA, SAN DIEGO

DAVID MOBLEY
UNIVERSITY OF CALIFORNIA, IRVINE

MICHAEL SHIRTS
UNIVERSITY OF COLORADO, BOULDER

LEE-PING WANG
UNIVERSITY OF CALIFORNIA, DAVIS

THE OPEN FORCE FIELD CONSORTIUM IS WORKING ON STANDARDS AND TOOLS

SMIRNOFF force field spec to define how force field parameters are to be applied
https://open-forcefield-toolkit.readthedocs.io/en/topology/smirnoff.html

(Bio)molecular Topology spec describing the chemical matter in the system to facilitate automated application of parameters

Molecule description spec describing an individual molecule with chemical information (to replace mol2, SDF, PDB) inspired by **QC JSON spec**

Automated benchmarking against (bio)physical datasets using standard experimental data formats (starting with NIST ThermoML Archive, but we lack format standards for other biophysical datasets)

http://openforcefield.org

PHIL STANSFELD

OXFORD UNIVERSITY

Automated simulation preparation with MemProtMD

CHRISTOPHER WOODS

UNIVERSITY OF BRISTOL

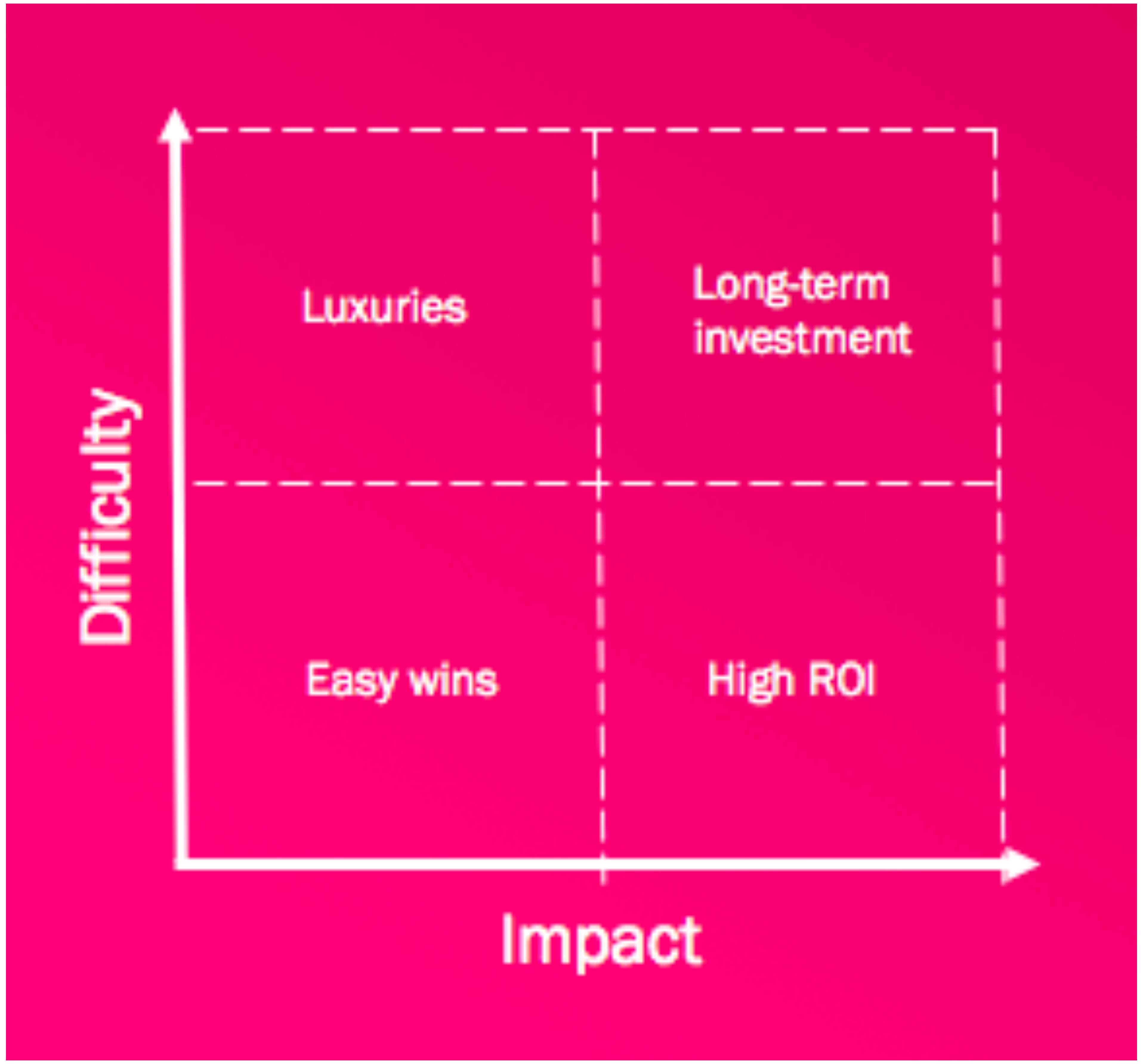
Streamlining and sharing molecular simulation data flows with BioSimSpace

What could **reduce the friction** for users and developers in biomolecular simulation workflows?

What are the opportunities for **common data models** to facilitate interoperability and streamline data flows at any stage?

Existing tools/initiatives

Challenges


Luxuries
Long-term investment
Easy wins
High ROI
Difficulty
Impact