

Data-Driven Force Field Development for Biomolecular Simulation

Daniel Cole

School of Natural and Environmental Sciences



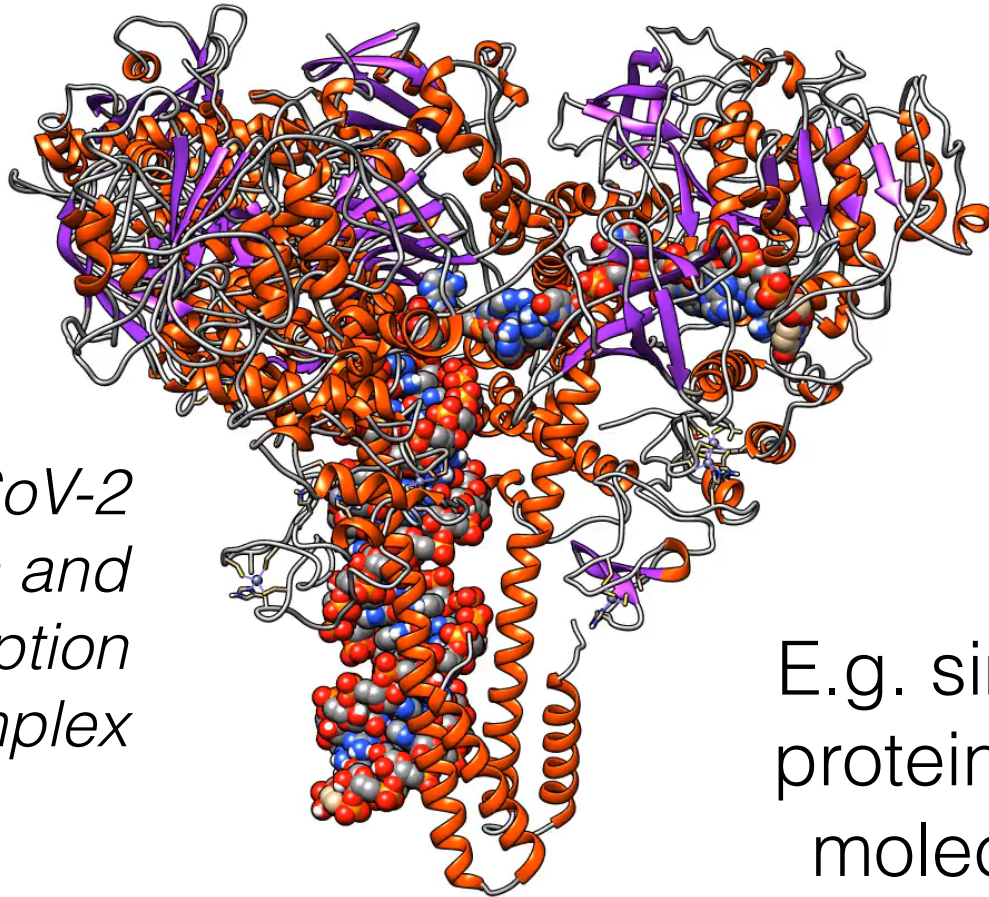
UK Research
and Innovation



open
forcefield



Why Biomolecular Simulation?



*SARS-CoV-2
replication and
transcription
complex*

Computer simulations provide key dynamical information on biomacromolecules relevant to human health and disease

E.g. simulations of spike proteins, design of small molecule therapeutics



But biology is heterogeneous, multiscale (length and time), is driven by small free energy changes, and is challenging to compare with experiment

Molecular Mechanics

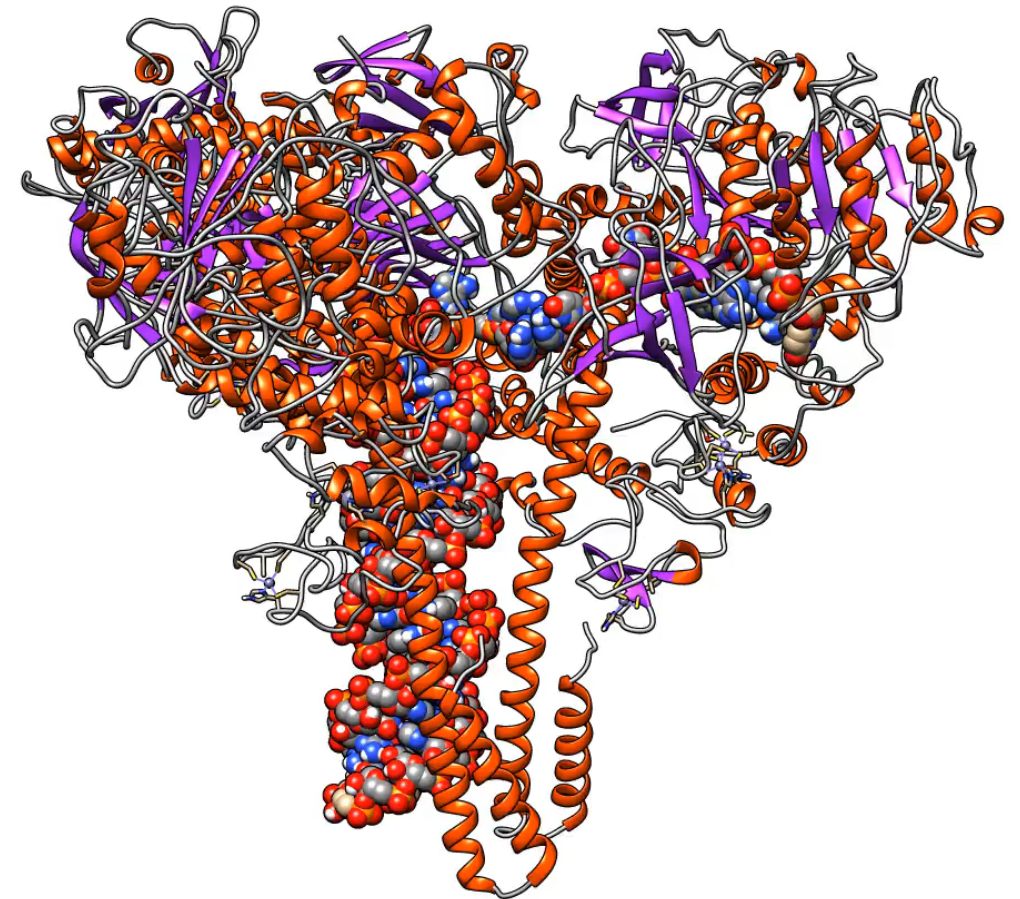
Molecular mechanics (MM) can complement experiment in many ways. At the simplest level, we can use molecular dynamics to ‘animate’ the system.

How does it work? If we know the structure at $t=0$ and the forces on the atoms, we can integrate Newton’s 2nd law to find the positions at all later times:

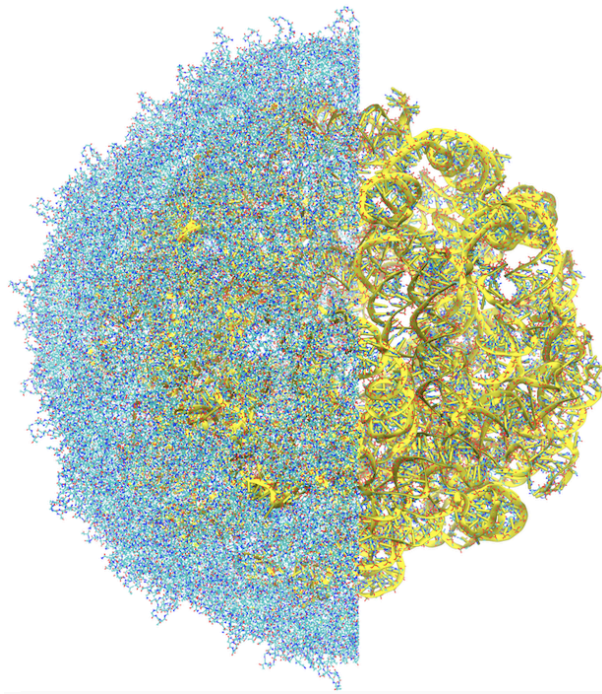
$$\mathbf{F}_i = m_i \mathbf{a}_i = m \frac{d^2 \mathbf{r}_i}{dt^2}$$

Time steps of 1fs (10^{-15} s) are required to model bond vibrations.

Simulations tend to be limited only by finite sampling (μ s to ms) and accuracy of the forces.

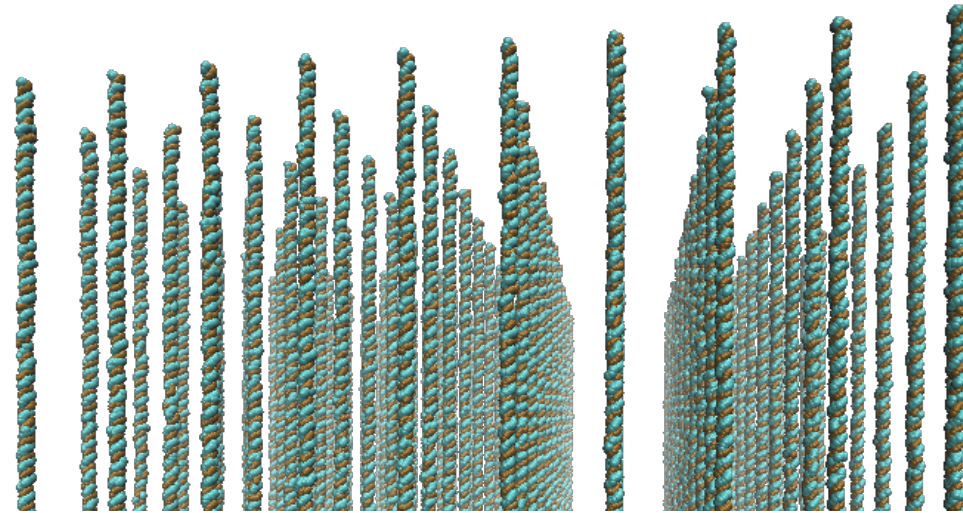


Opportunities at Exascale

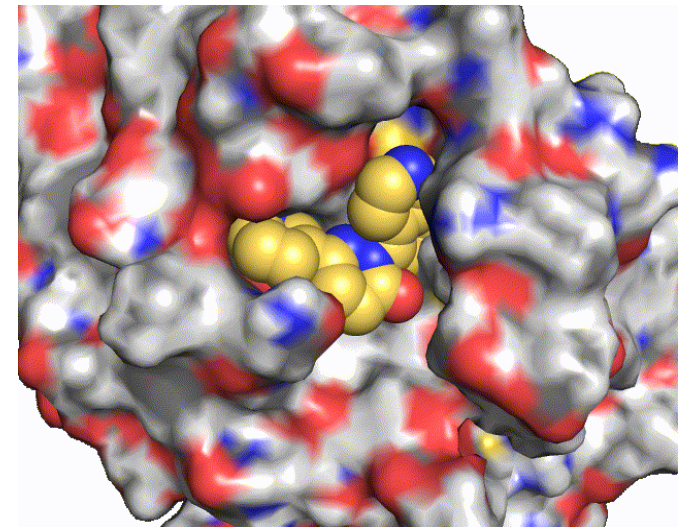


DOI: 10.1039/d2fd00053a

Massive systems & integrating with experimental data explore the limits of current parallel programming



Multiscale DNA modelling at the atomistic and coarse-grained level, up to the scale of chromatin and entire chromosomes



Ensemble computing for drug discovery requires huge numbers of lightly coupled compute resources

Our Community

SEP
28-02
2026

CCPBioSim Training Week

Sheffield, UK

Our annual training week, introducing basic methods and concepts for beginners.

OCT
13-17
2025

CCPBioSim Training Week

Sheffield, UK

Our annual training week, introducing basics methods and concepts for beginners.

JUL
20-31
2025

CCP5 Summer School

Newcastle, UK

The intensive summer school for computational chemists run by CCP5.



JUL
06-08
2026

12th Annual CCPBioSim Conference: Frontiers in Biomolecular Simulations

Bristol, UK

The twelfth annual meeting of CCPBioSim

MAY
18-20
2026

Integrated Modelling and Scattering for Biomolecules

Sheffield, UK

This meeting will bring together researchers from the scattering and modelling communities.

AUG
27-29
2025

CECAM Flagship Workshop "Entropy of Soft Matter"

Vienna, AT

This CECAM Flagship Workshop "Entropy of Soft Matter Systems"

SEP
24
2026

DRIIMB Talk BioFAIR

Zoom

Introduction to BioFAIR: National Life Sciences Digital Research Infrastructure

SEP
23
2026

Industry Talk - Evotec

Online

Understanding real conformations in solution to accelerate Drug Discovery
@Evotec

AUG
27
2026

Recent Developments in Integrative Structural Modeling of Macromolecular Assemblies

online

DRIIMB Webinar by Dr Shruthi Viswanath - 27 August, 14:00 BST



<https://www.ccpbiosim.ac.uk/>

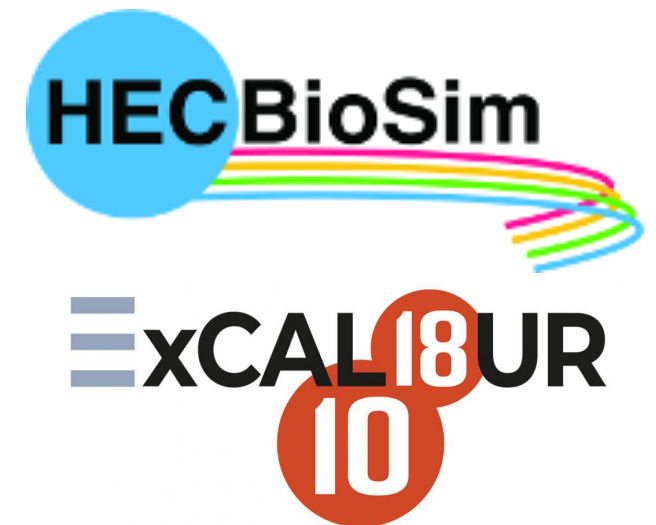
What do our users want to know?

What software should I use?

Which HPC system should I apply for time on?

How many hours should I apply for?

How should I configure the software?



Real-world HPC performance comes down to system configuration, libraries, software versions and compilers.

Aim to develop series of toy and showcase simulations that are representative of the types of science that our users are interested in.

Log all available information, including system configuration, performance and diagnostic information.

<https://excalibur.ac.uk/projects/exabiosim/>
<https://github.com/HECBioSim/hpcbench>

HPCbench

Get system info

Log CPU and GPU
usage

Run the benchmark

Parse MD output file

Get slurm info

Combine outputs
together

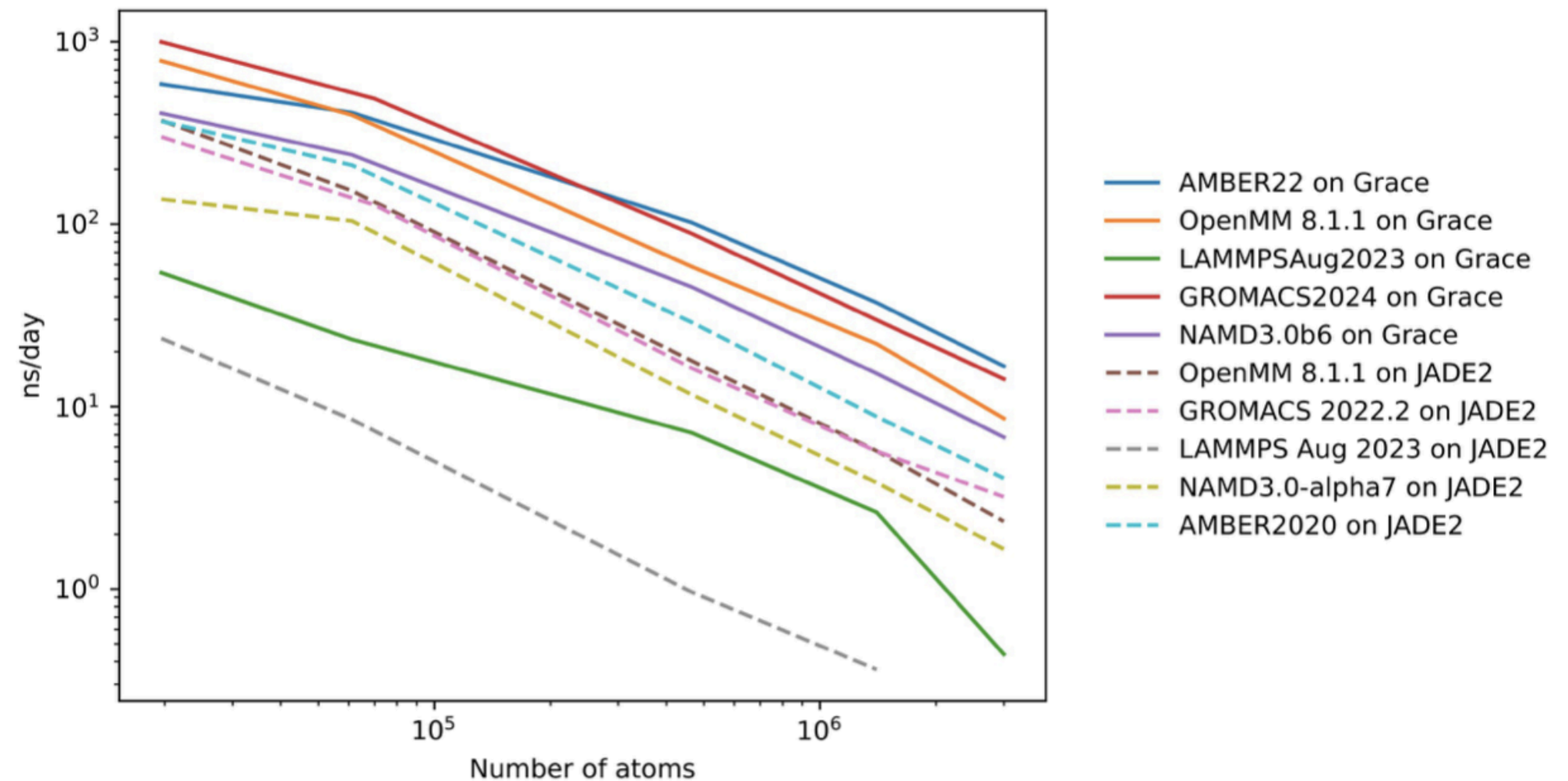
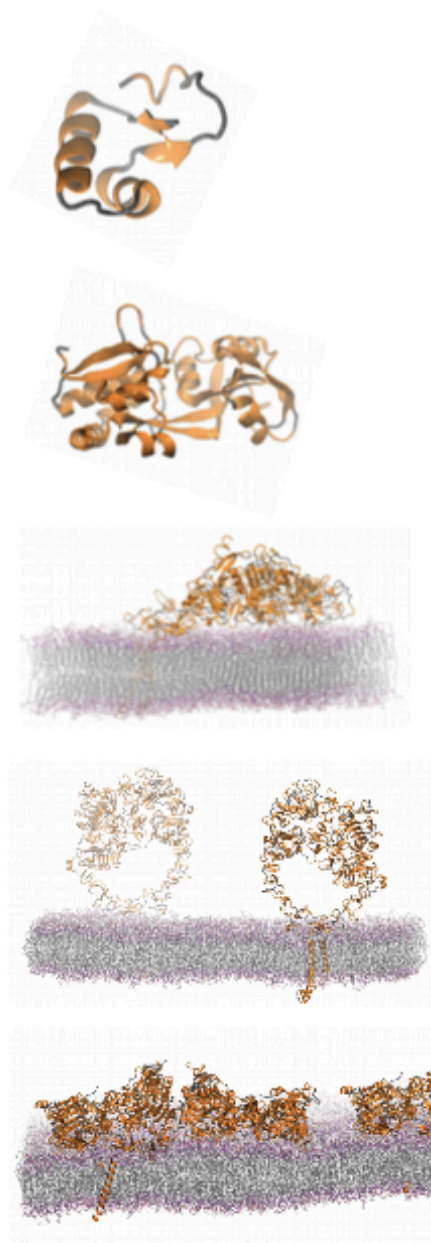
```
1  #!/bin/bash
2  #SBATCH --nodes=1
3  #SBATCH --time=02:00:00
4  #SBATCH --job-name=ambergpu
5  #SBATCH --gres=gpu:1
6  #SBATCH --ntasks=4
7  #SBATCH --cpus-per-task=1
8  #SBATCH --partition=small
9
10 # Setup environment
11 module load cuda/11.1-gcc-9.1.0
12 module load gcc/9.1.0
13 hpcbench infolog sysinfo.json
14 hpcbench gpulog gpulog.json & gpuid=$!
15 hpcbench cpulog "'lmp'" cpulog.json & cpuid=$!
16
17 # Run the benchmark
18
19
20 mpirun -np 4 lmp -sf gpu -pk gpu 1 -in benchmark.in
21
22 # Tidy up
23 kill $gpuid
24 kill $cpuid
25
26 # Post-process
27 hpcbench sacct $SLURM_JOB_ID accounting.json
28 hpcbench lmplog log.lammps run.json
29 hpcbench slurmllog $0 slurm.json
30 hpcbench extra -e "'Comment:test'" -e "'Machine:JADE2'" meta.json
31 hpcbench collate -l sysinfo.json gpulog.json cpulog.json \
32 accounting.json run.json slurm.json meta.json -o test.json
```

<https://github.com/HECBioSim/hpcbench>

Benchmarking

Parameters: system size, HPC system used, MPI tasks / OMP threads

Metrics: ns/day, ns/J

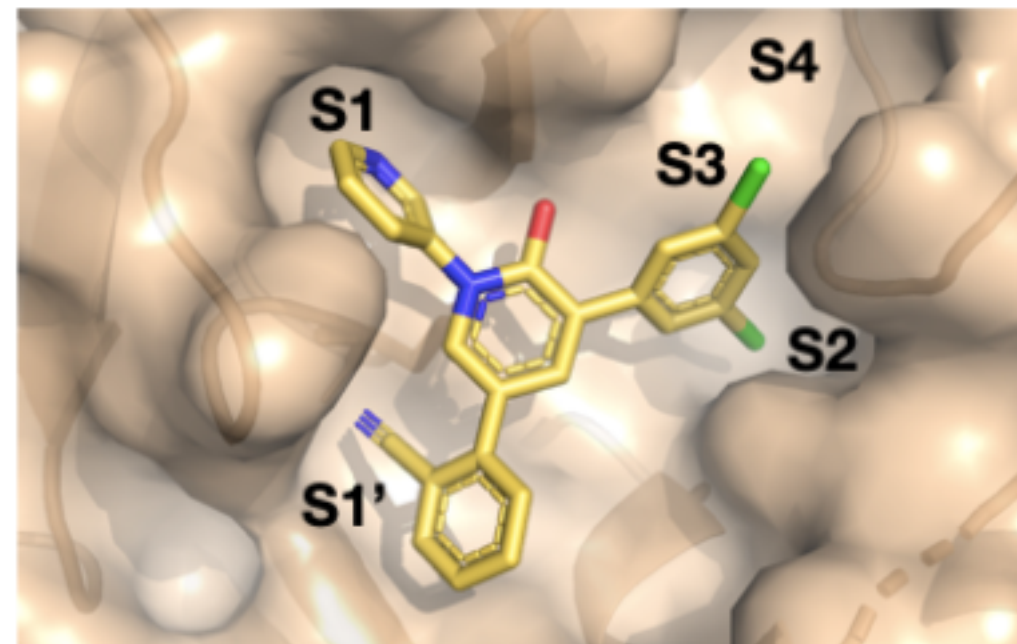


<https://arxiv.org/html/2506.15585v1>

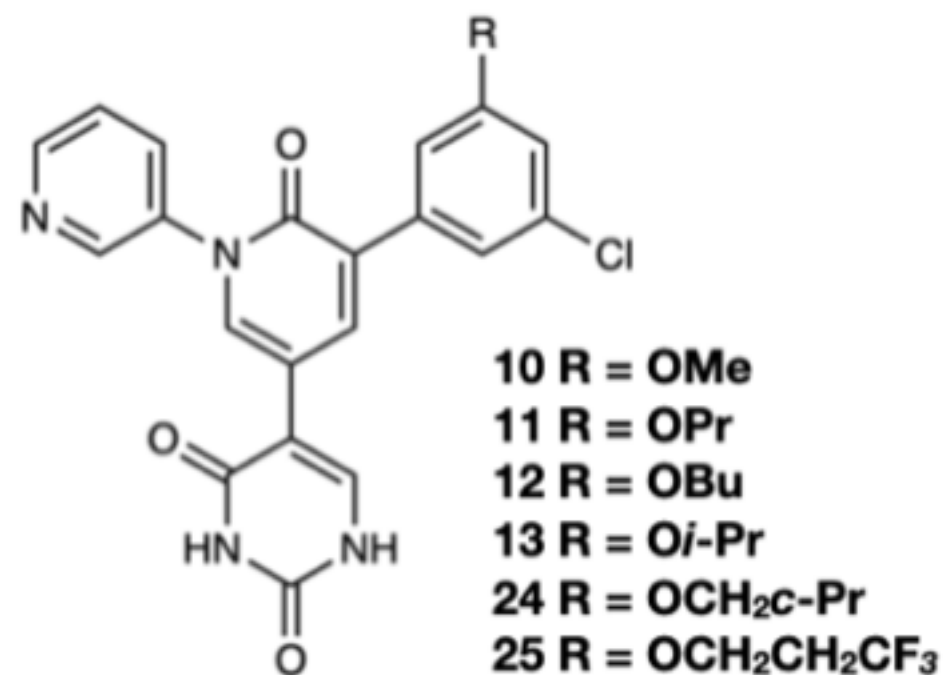
Case Study: Drug Design

Computer-Aided Drug Design

- **Medicinal Chemistry**
 - Aim to design a molecule that binds to a target (usually a protein) with therapeutic benefit
- **Medicinal Chemistry is Expensive**
 - Estimated R&D costs of \$2bn
- **Multiple properties must be optimised at once**
 - Strength of binding (the right shape, suitable intermolecular interactions, ...)
 - Pharmacokinetics (including absorption, distribution, metabolism and excretion properties)
 - Suitable for computational design



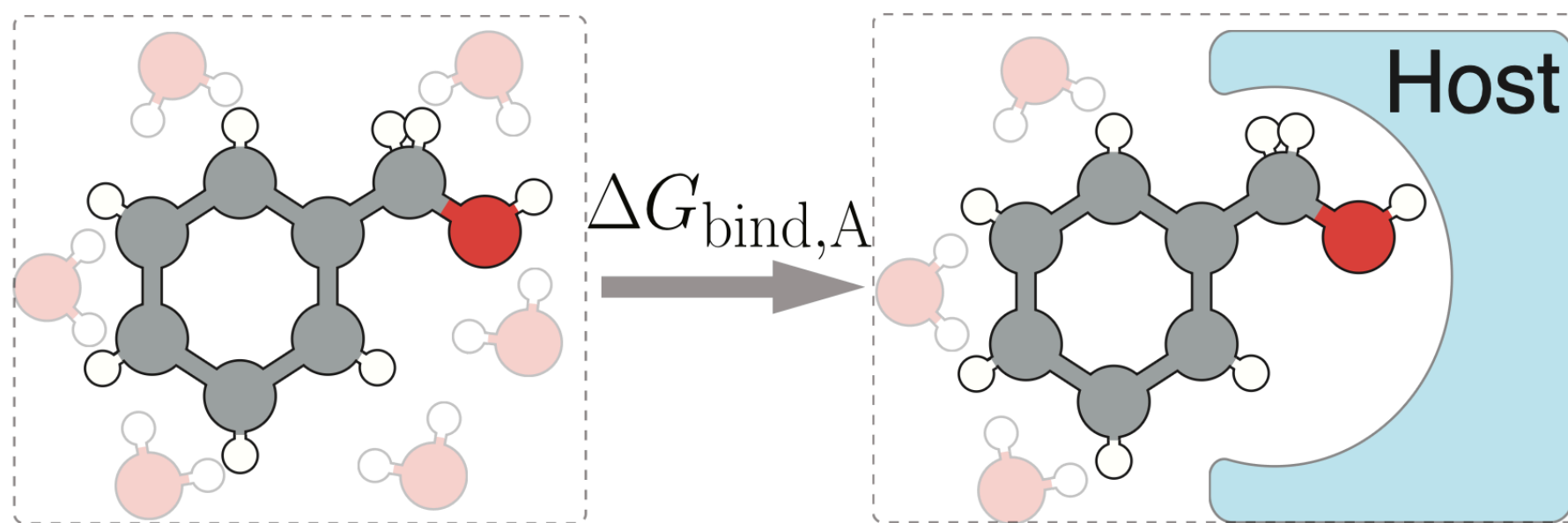
e.g. design of SARS-CoV-2 Mpro inhibitors



Free Energy for Drug Discovery

Medicinal chemistry is expensive — need computational models that can predict binding from structure.

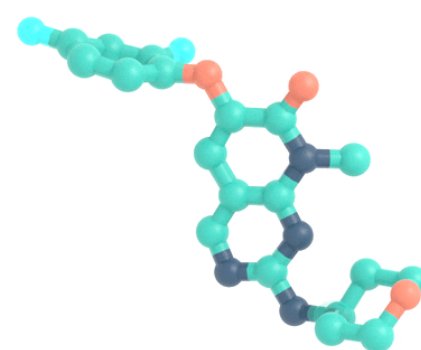
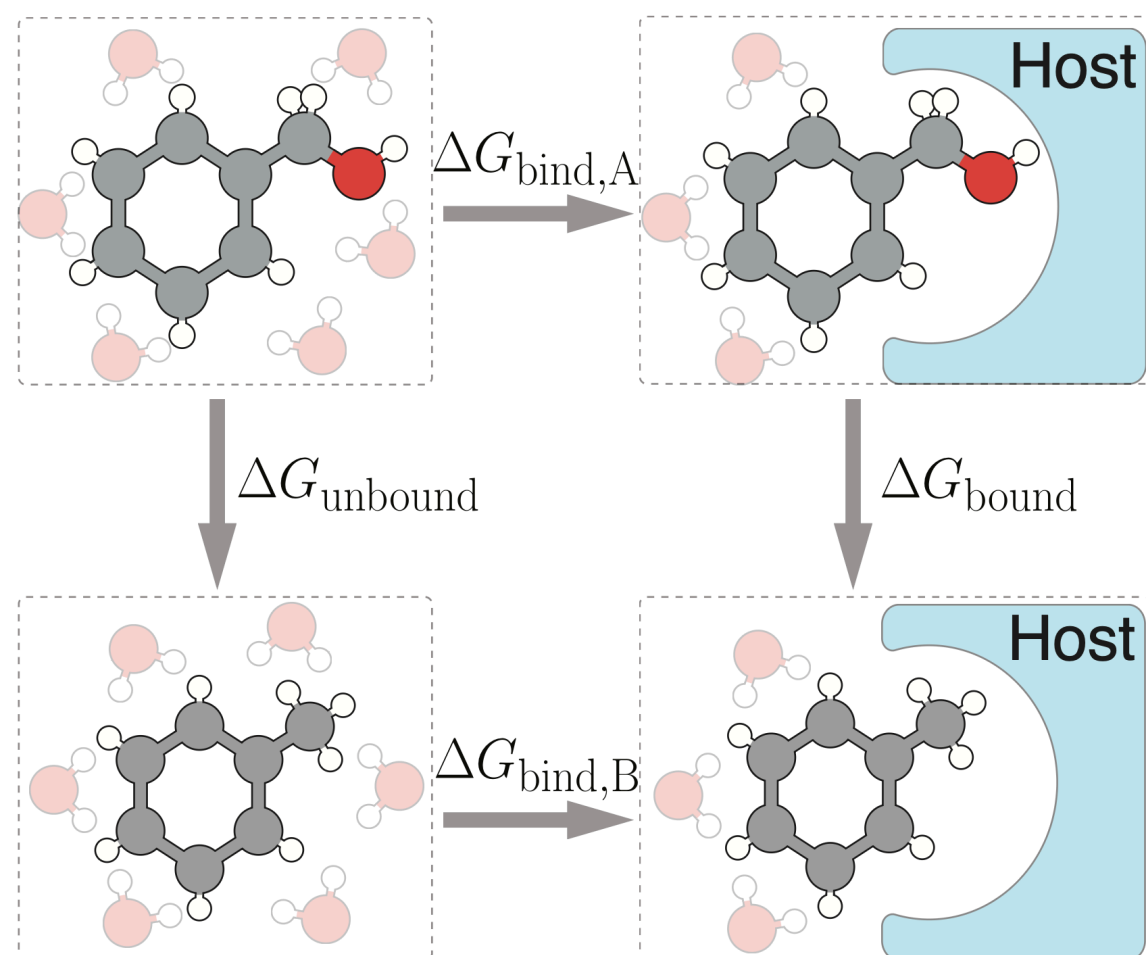
We are typically interested in the free energy difference between a small molecule in solution and bound to a protein:



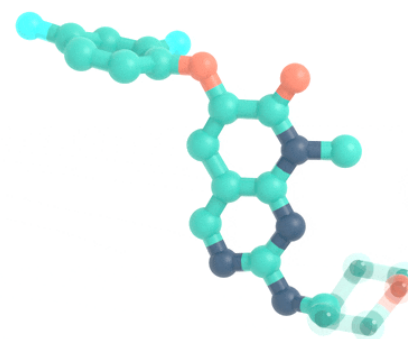
Free energy calculations based on molecular dynamics enable rigorous calculation of this free energy, and prioritisation of compounds for synthesis.

Free Energy for Drug Discovery

In practice, molecules are 'alchemically' mutated in the host and solution to compute relative binding free energies:



$$\Delta G_{\text{bound}} = -kT \ln \left\langle \exp \left[\frac{-(U_B - U_A)}{kT} \right] \right\rangle_A$$

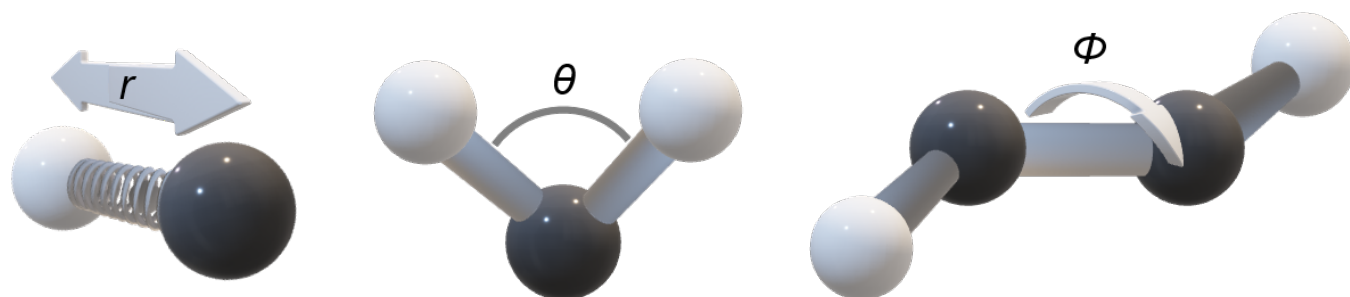


Accuracy depends on force field and sampling (requires ~ns of MD for 10K+ atoms)

Force Field

Simulations of **large** chemical systems over **long** time scales (e.g. free energy calculations for computer-aided drug design) are crucially reliant on the accuracy of the molecular mechanics force field.

$$E_{Total} = \sum_{Bonds} K_r(r - r_0)^2 + \sum_{Angles} K_\theta(\theta - \theta_0)^2 + \sum_{Torsions} \frac{V_n}{2} [1 + \cos(n\Phi - \gamma)]$$

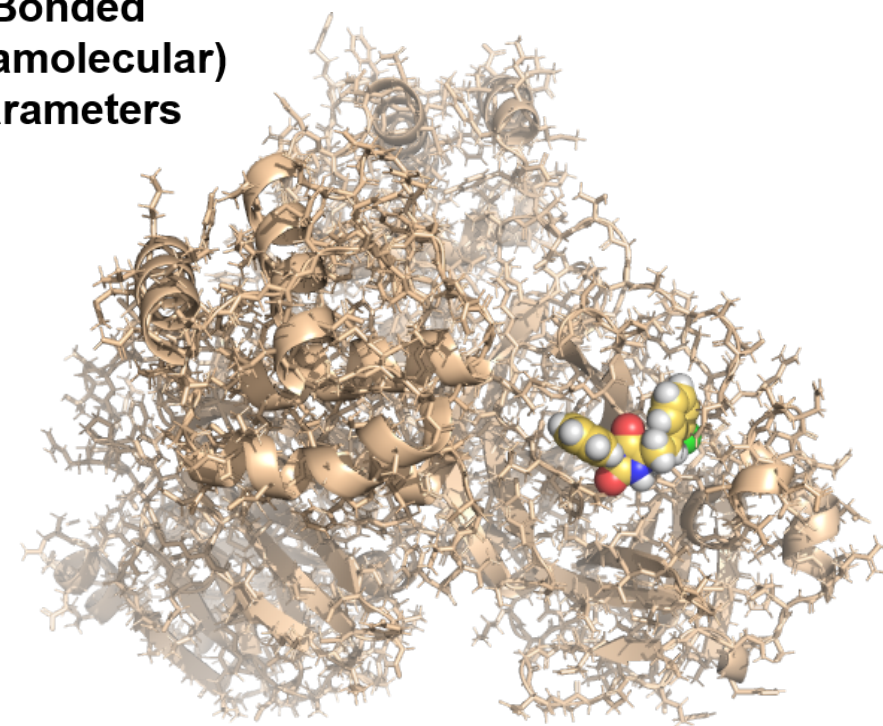


**Bonded
(Intramolecular)
Parameters**

$$+ \sum_{Non-Bonded} \left[4\epsilon_{ij} \left\{ \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right\} + \frac{q_i q_j}{r_{ij}} \right]$$

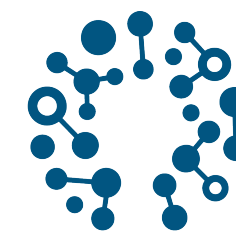


**Non-Bonded
(Intermolecular)
Parameters**



Open Force Field

The Open Force Field Initiative is an international academic-industry collaboration working together to advance science & infrastructure required to build force fields.



open
forcefield



OPEN SOFTWARE

Automated infrastructure enables rapid experimentation with minimum human intervention



OPEN DATA

Access to large, high quality experimental and quantum chemical data facilitates easy curation of balanced train / test sets

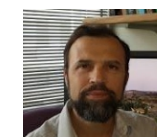


OPEN SCIENCE

Exploring new force field science:
hypothesis - build
software - train - test -
iterate
is now almost routine



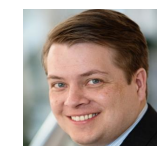
John Chodera (MSKCC)



Michael Gilson (UCSD)



David Mobley (UC Irvine)



Michael Shirts (CU Boulder)



Jeff Wagner
Technical Lead



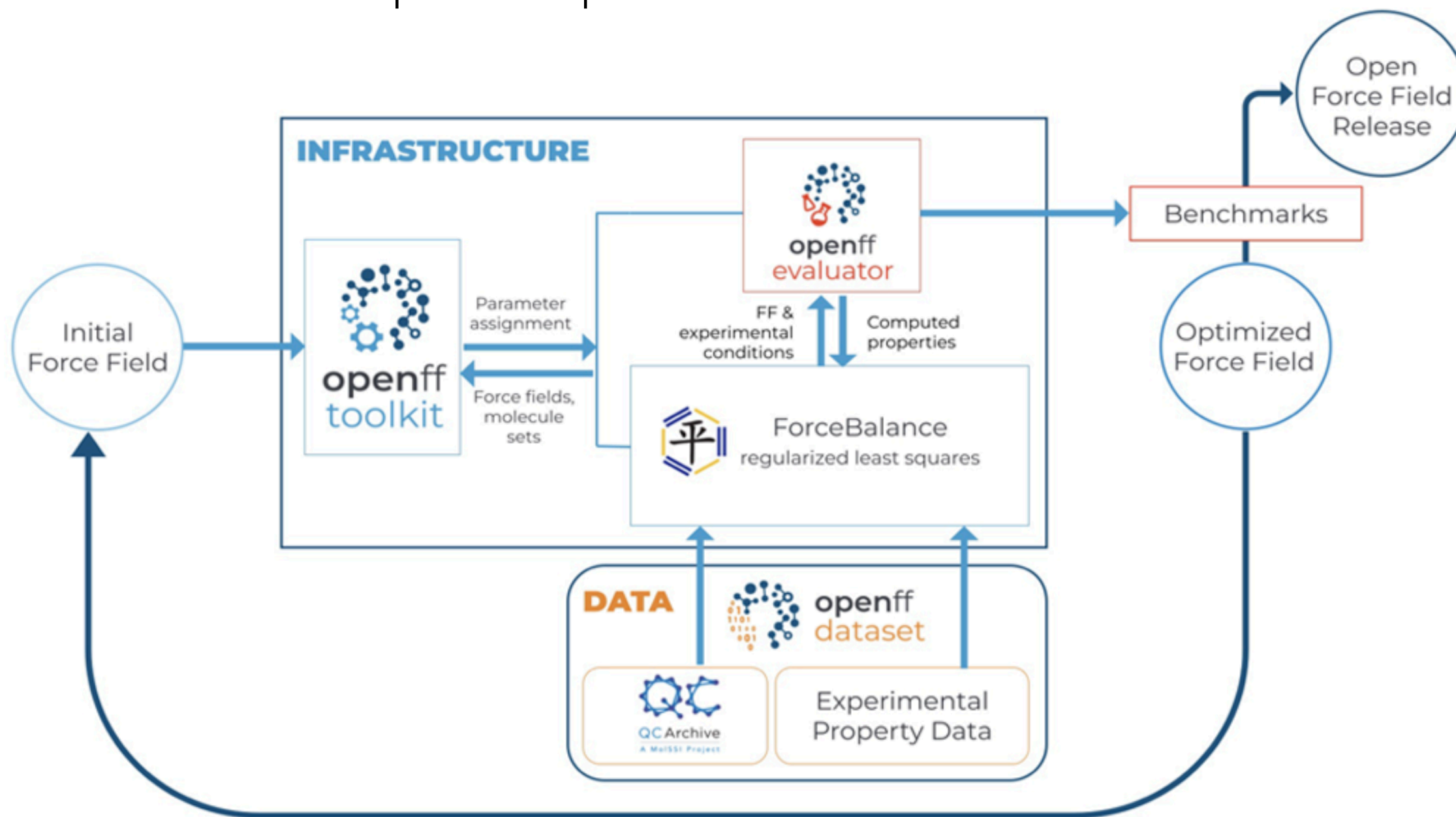
Lily Wang
Science Lead

Anyone can get involved

<https://openforcefield.org/>

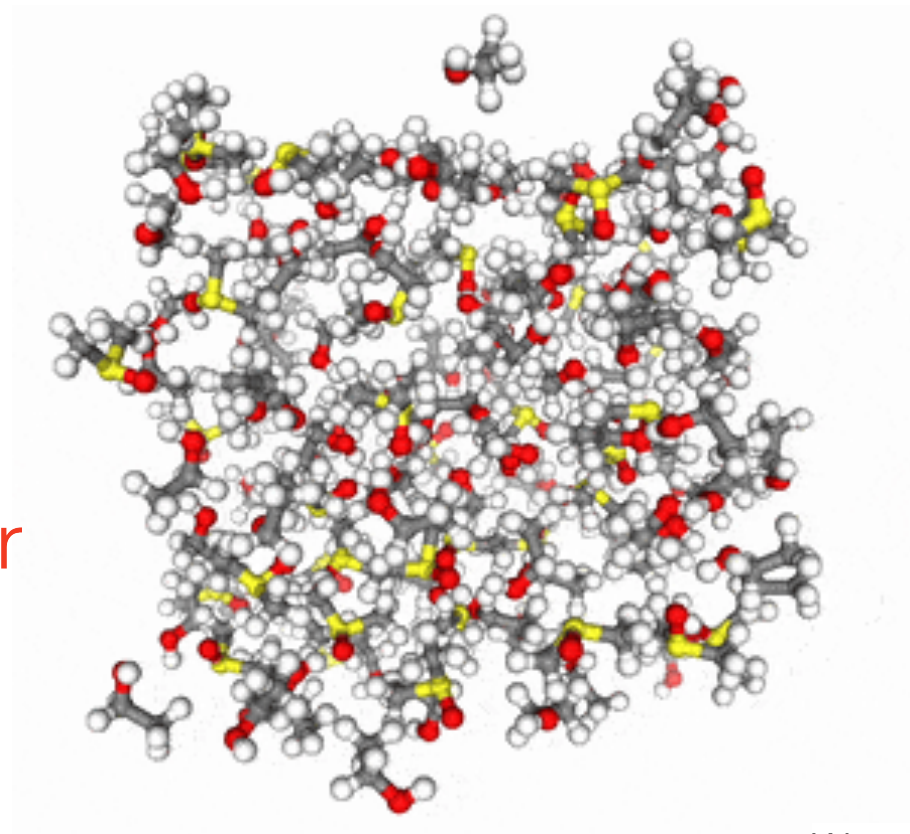
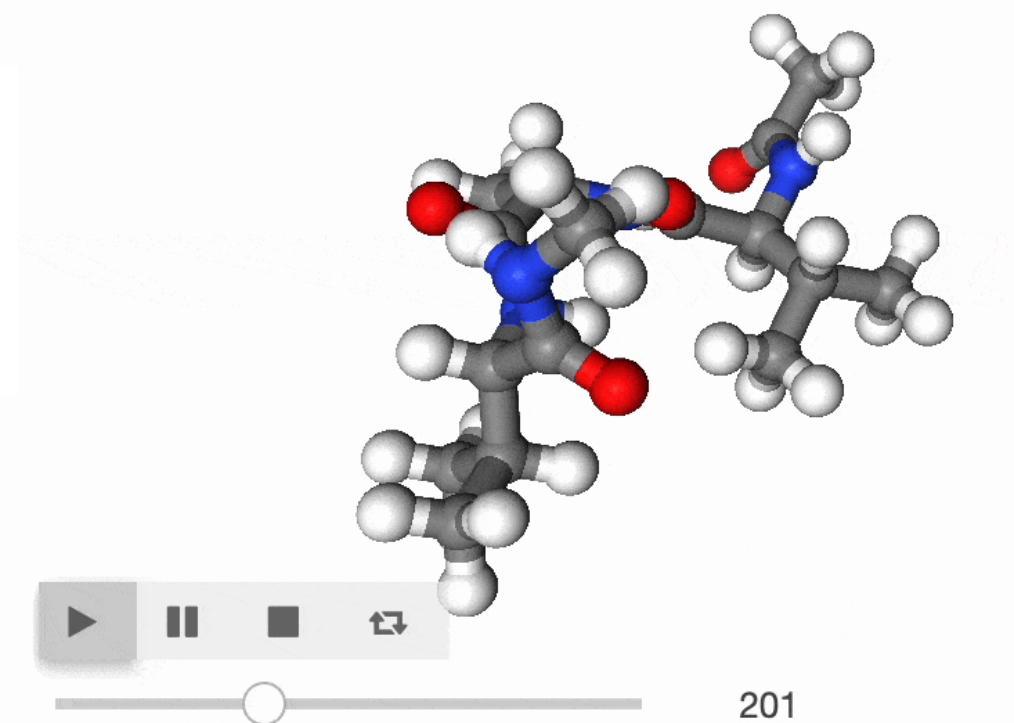
Open Force Field

Open Force Field infrastructure significantly reduces human time required to parameterise a new force field.



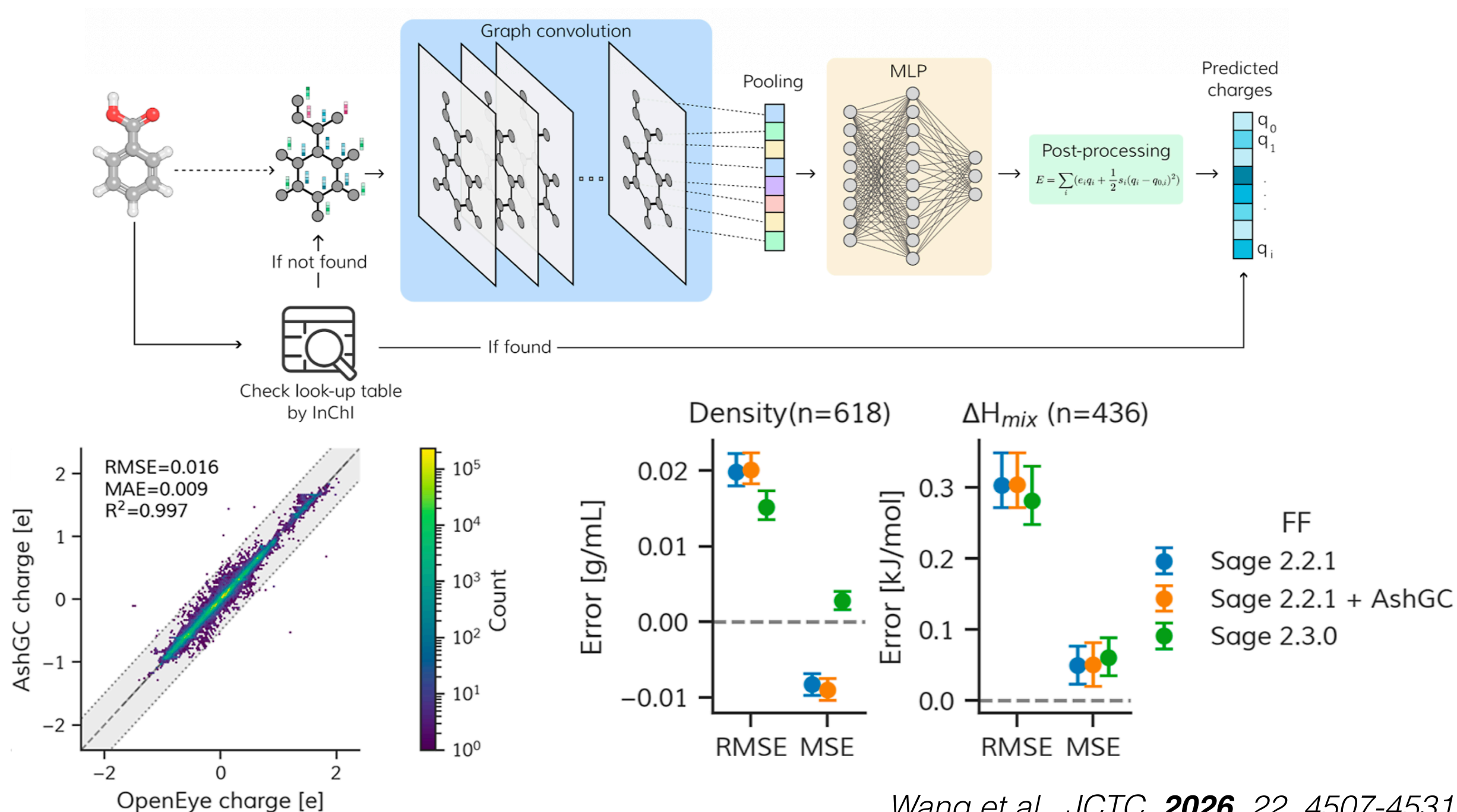
Training Data

Bonded parameters trained on curated QM energies and forces of 1000's of drug-like molecules and fragments



Non-bonded training set includes 1000's of mixture enthalpy and density measurements from NIST ThermoML

AshGC: same accuracy for fraction of cost



Follow-up charge model trained on **Isambard-AI** using higher level of QM theory

Adams et al., JCTC, **2025**, 21, 12133-12148

OpenFF is already a leading public force field



Parsley 1.x

Re-fit bonds, angles, torsions



Sage 2.x

Re-fit valence, LJ parameters



Sage 2.3.0 released!

with AshGC neural network charges

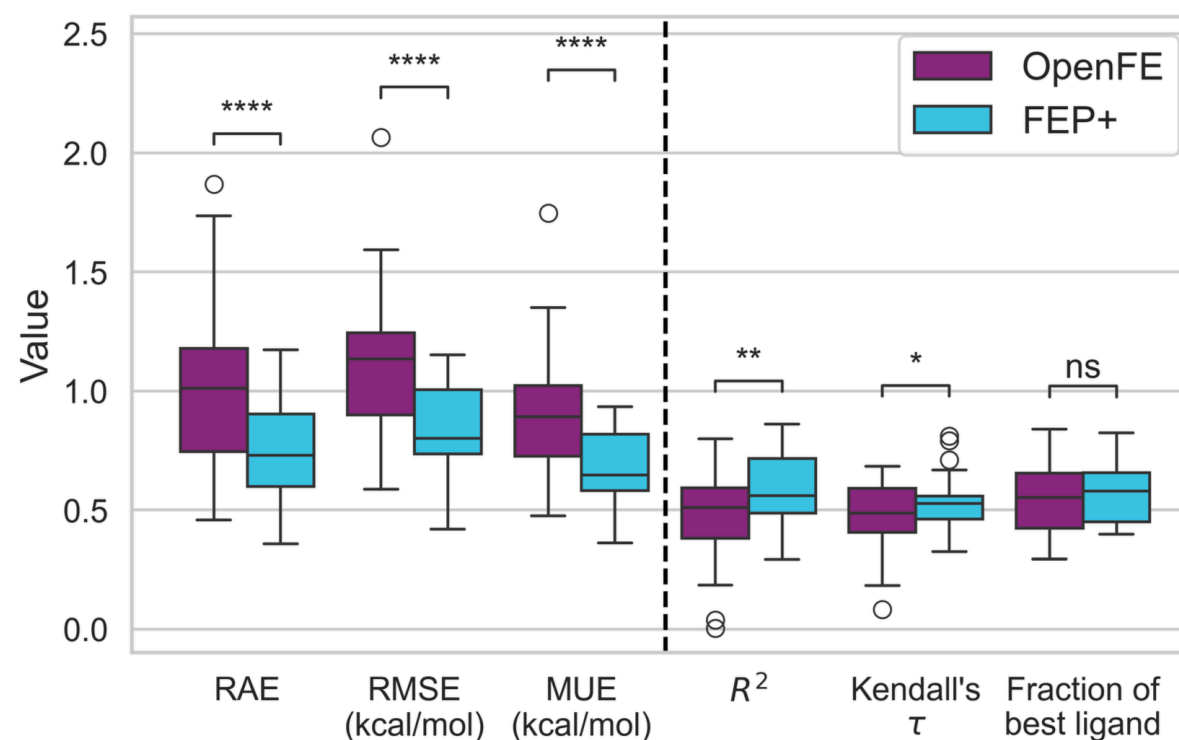
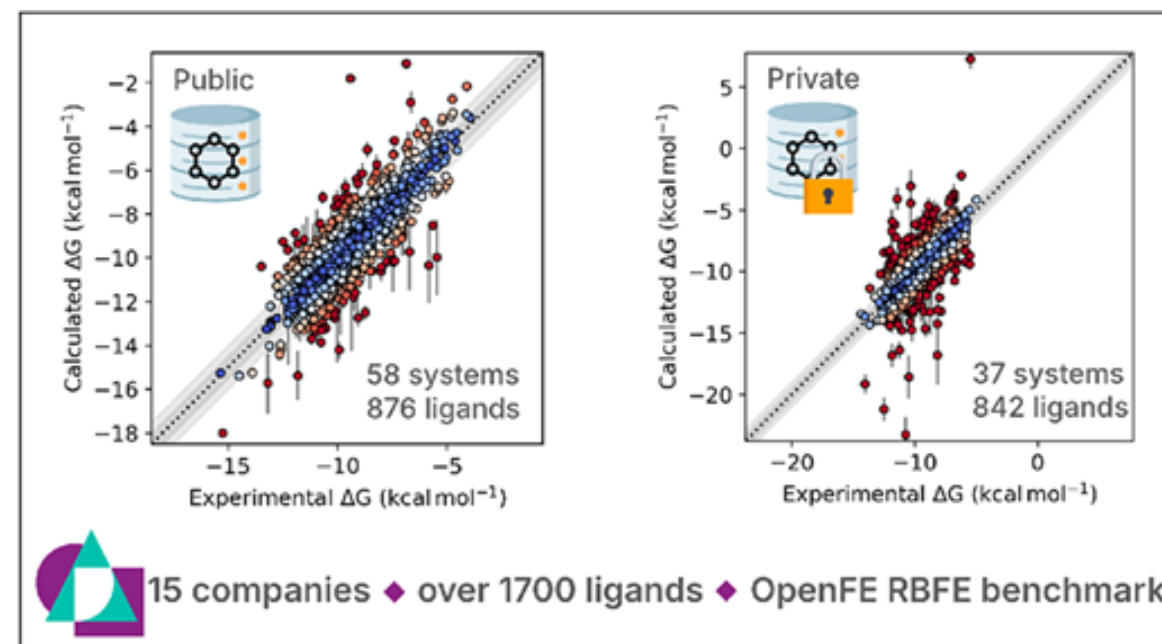


Rosemary 3.0

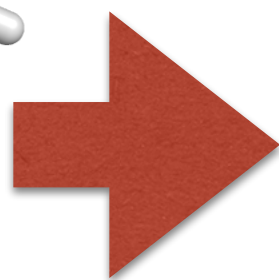
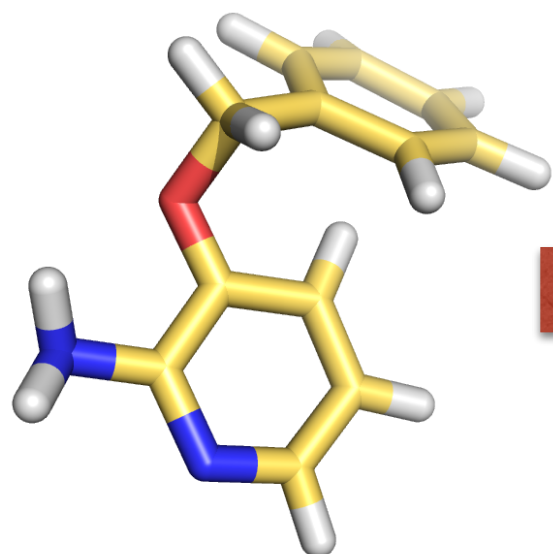
Expansion to protein parameters

Rosemary 3.0 alpha released

Preliminary peptide and small molecule parameters



Machine Learning Interatomic Potentials

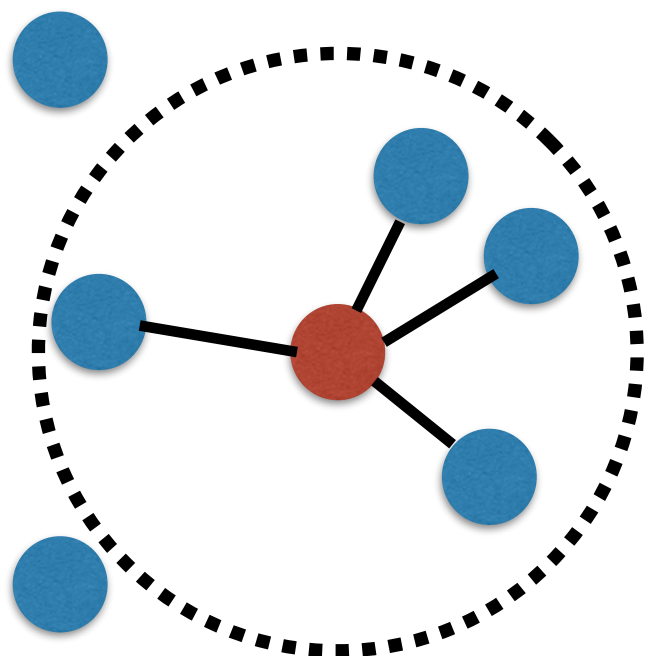


**Potential energy
and forces**

Force field does this using
sum over bond, angle,
dihedral etc contributions.

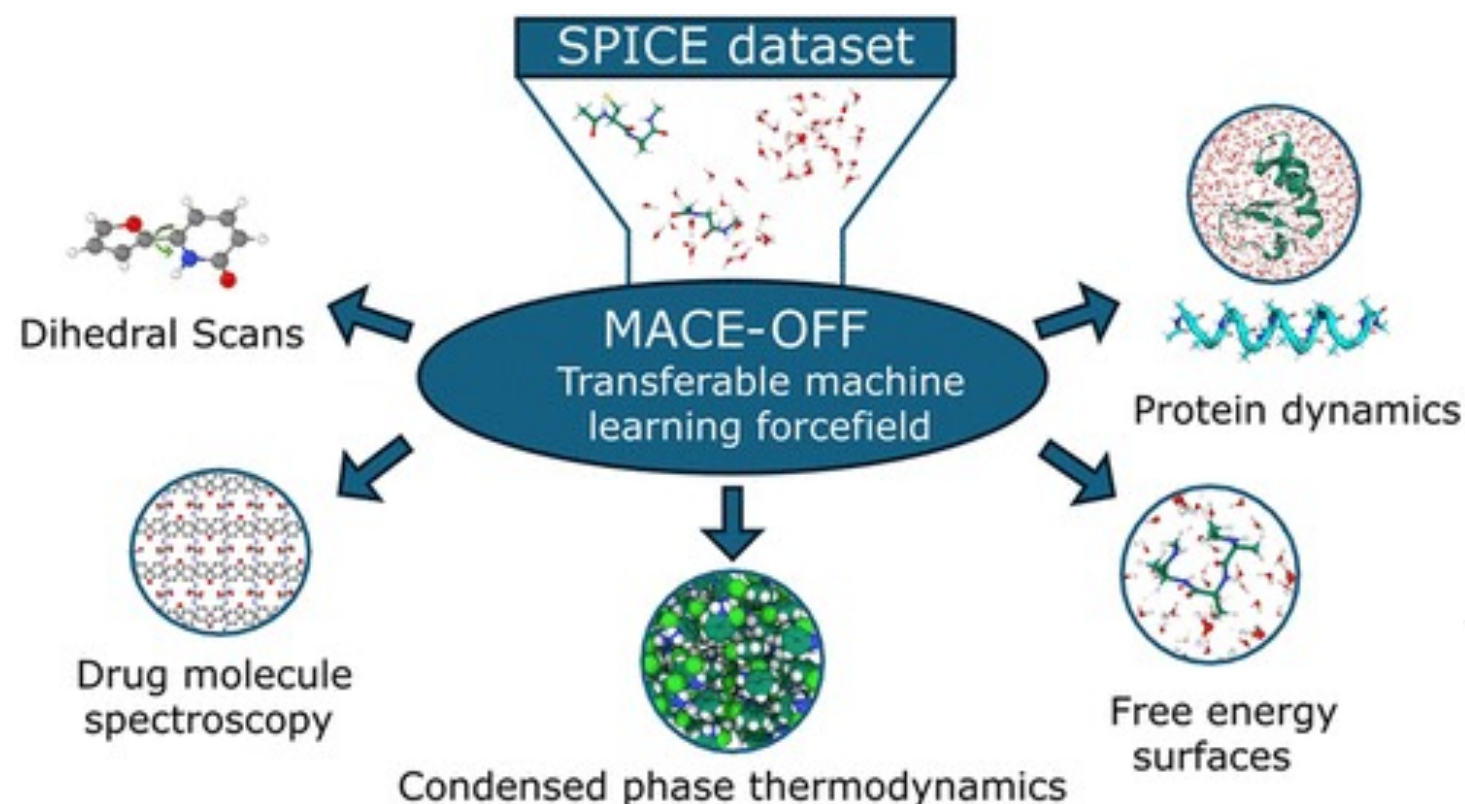
More generally can write the energy of atom i as a sum over
body-ordered contributions:

$$E_i = V_{z_i}^{(1)} + \frac{1}{2} \sum_j V_{z_i z_j}^{(2)}(\mathbf{r}_{ij}) + \frac{1}{3!} \sum_{j,k} V_{z_i z_j z_k}^{(3)}(\mathbf{r}_{ij}, \mathbf{r}_{ik}) + \dots$$

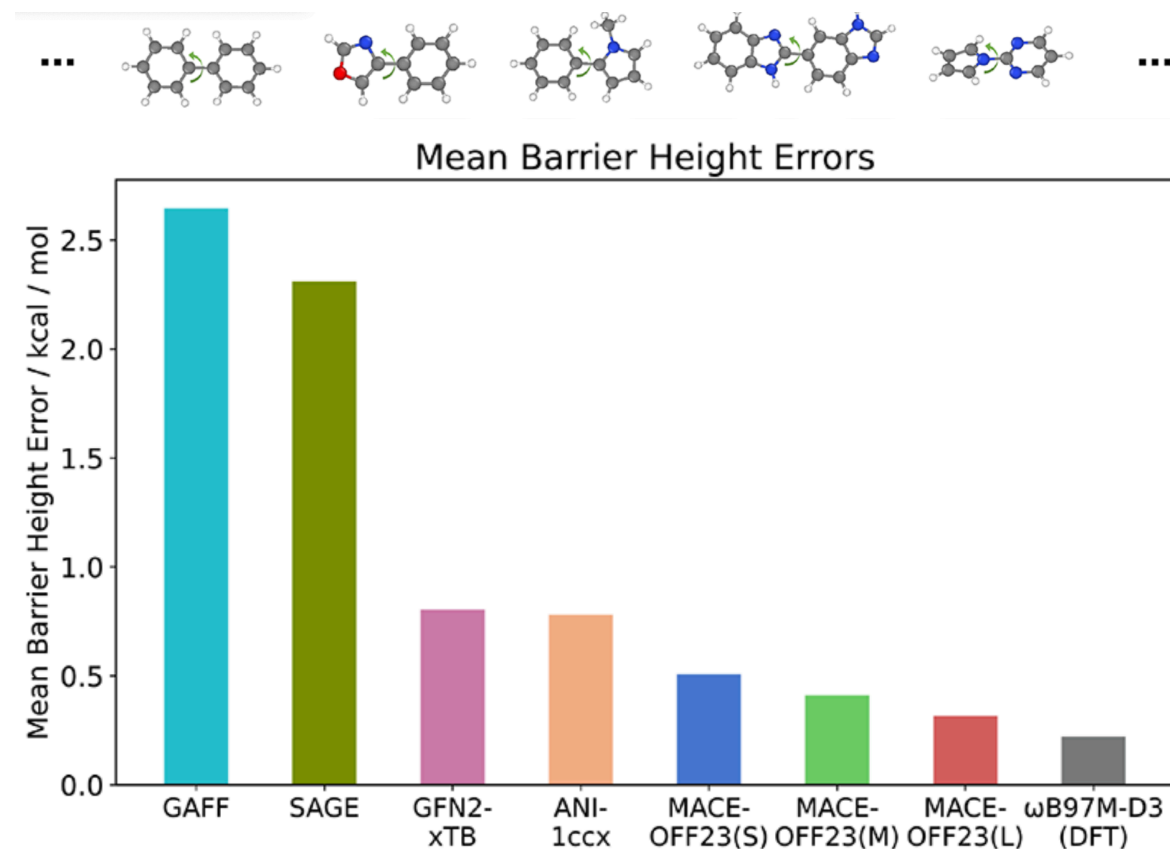


Represent an atom using its neighbourhood density.
Train a machine learning model to reproduce
quantum mechanical energies and forces.

MACE-OFF



MACE-OFF23 is a **transferable** force field for **organic molecules** built with the MACE architecture and trained on QM (ω B97M-D3(BJ)/def2-TZVPPD).

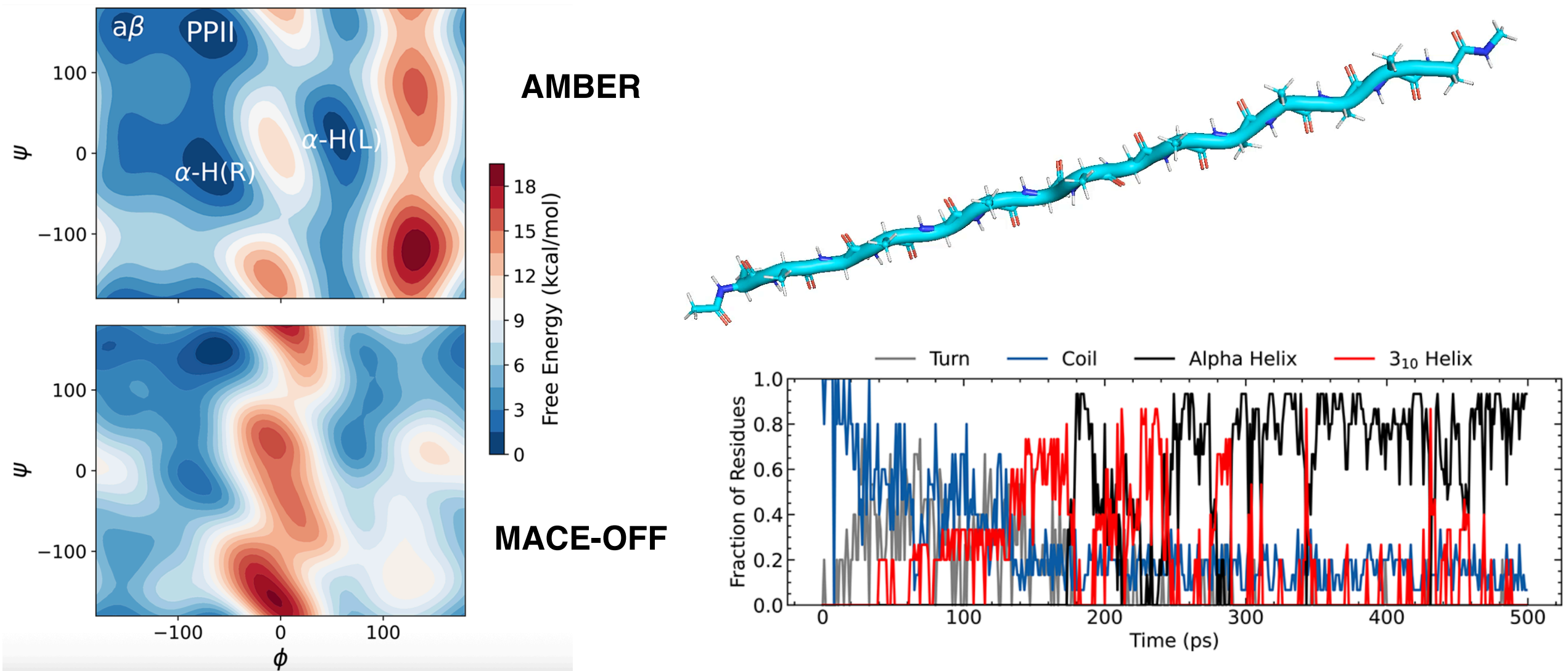


Tested on various gas and condensed phase physical properties.

Throughput is around 1-10 ns/day for systems of 1000+ atoms.

Condensed Phase Properties

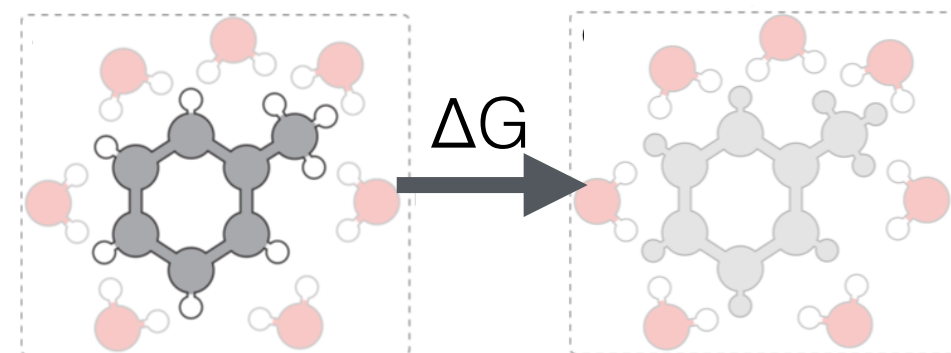
Tested on various condensed phase physical properties, including structure and dynamics of liquid water, molecular crystals, densities and heats of vapourisation of organic liquids, and dynamics of peptides and proteins.



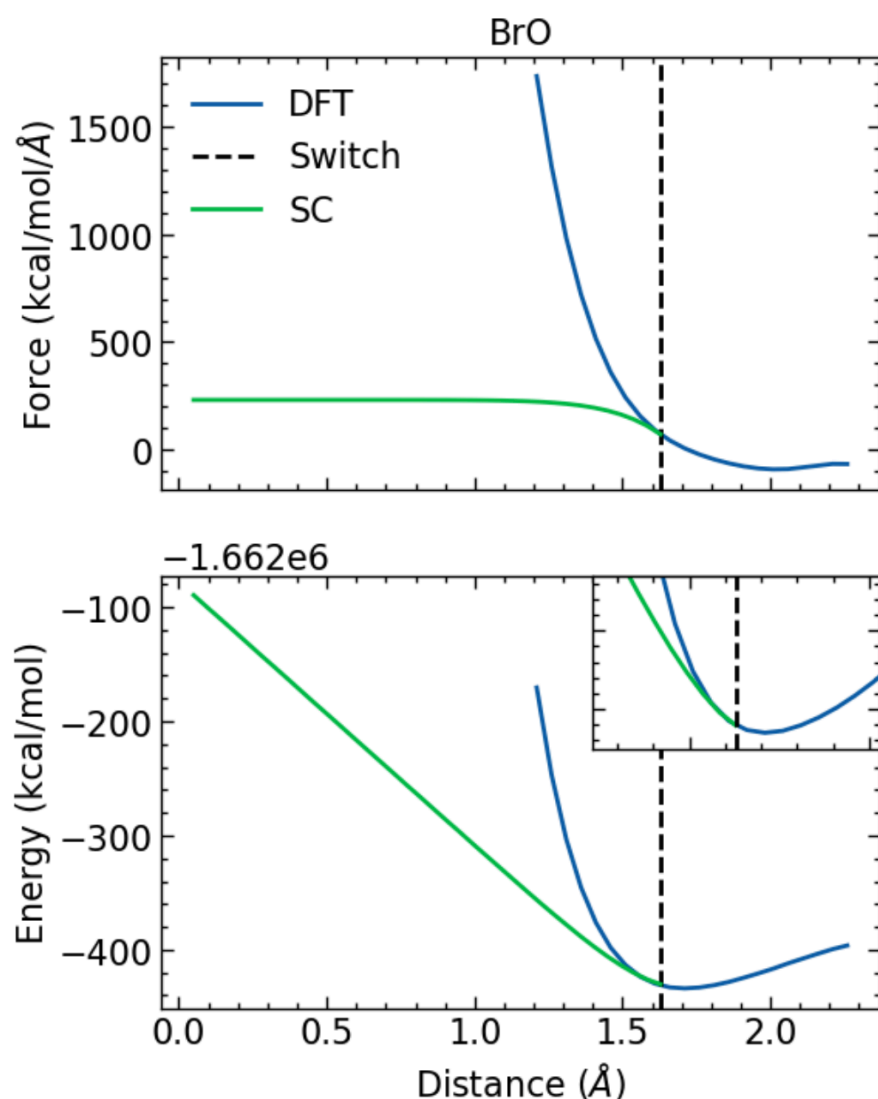
Soft-core MACE-OFF

To calculate free energies, we need to ‘alchemically’ appear/disappear atoms.

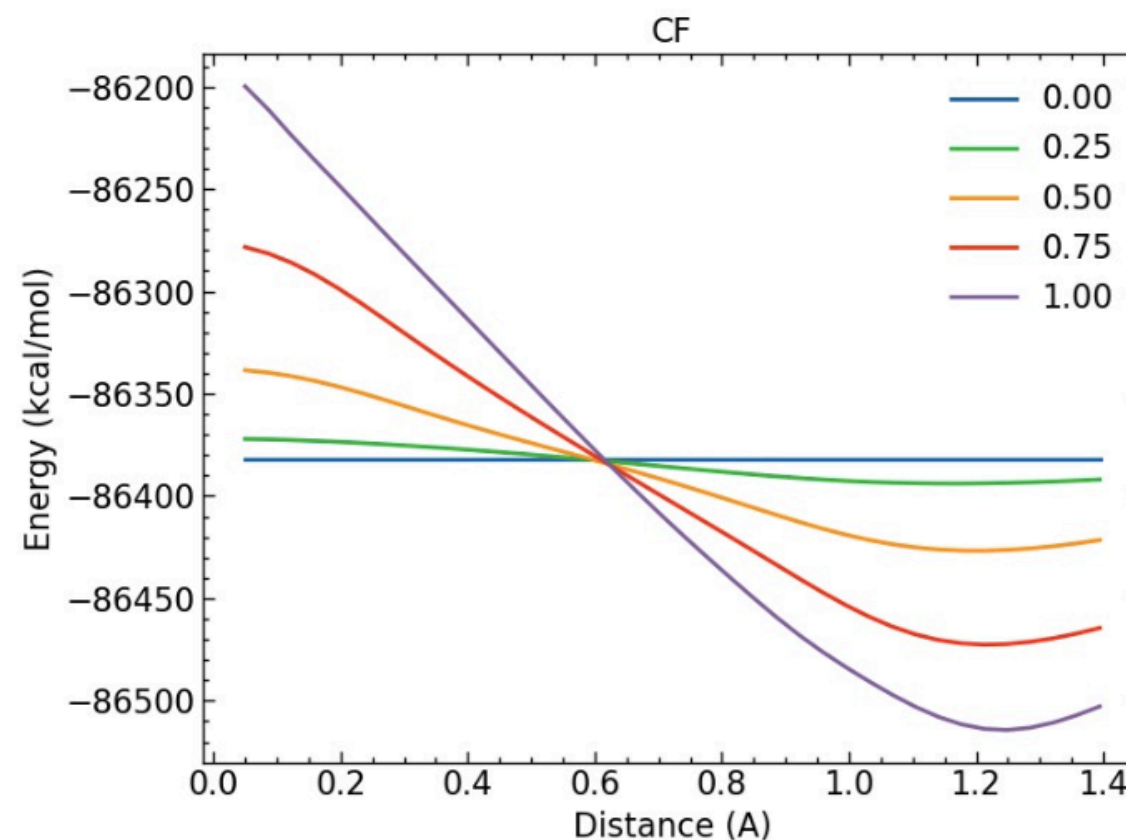
But this leads to close contact between atoms in regions where there are no MACE training data.



Mey, et al (2020), LiveCoMS 2(1), 18378

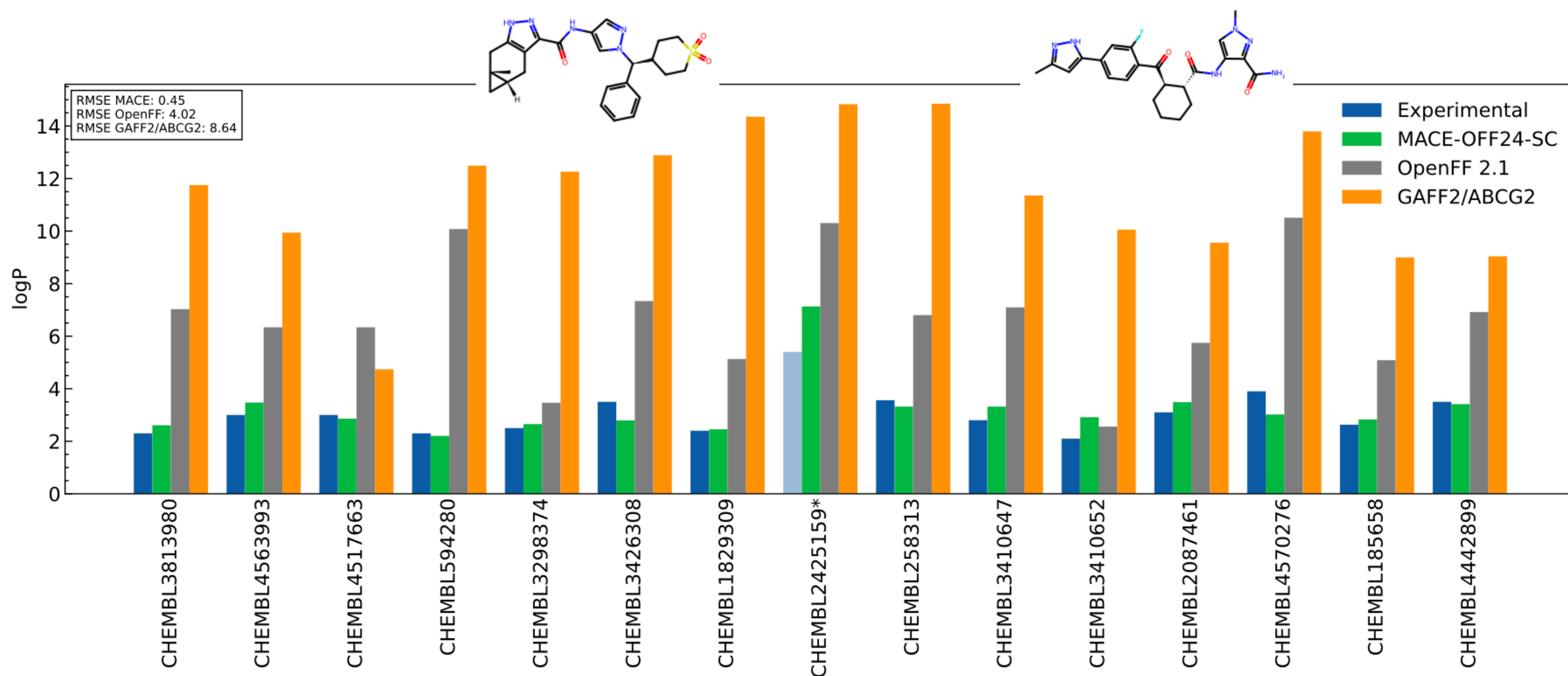


Add synthetic dimer curves to training data, and include alchemical scaling of radial functions in MACE potential.



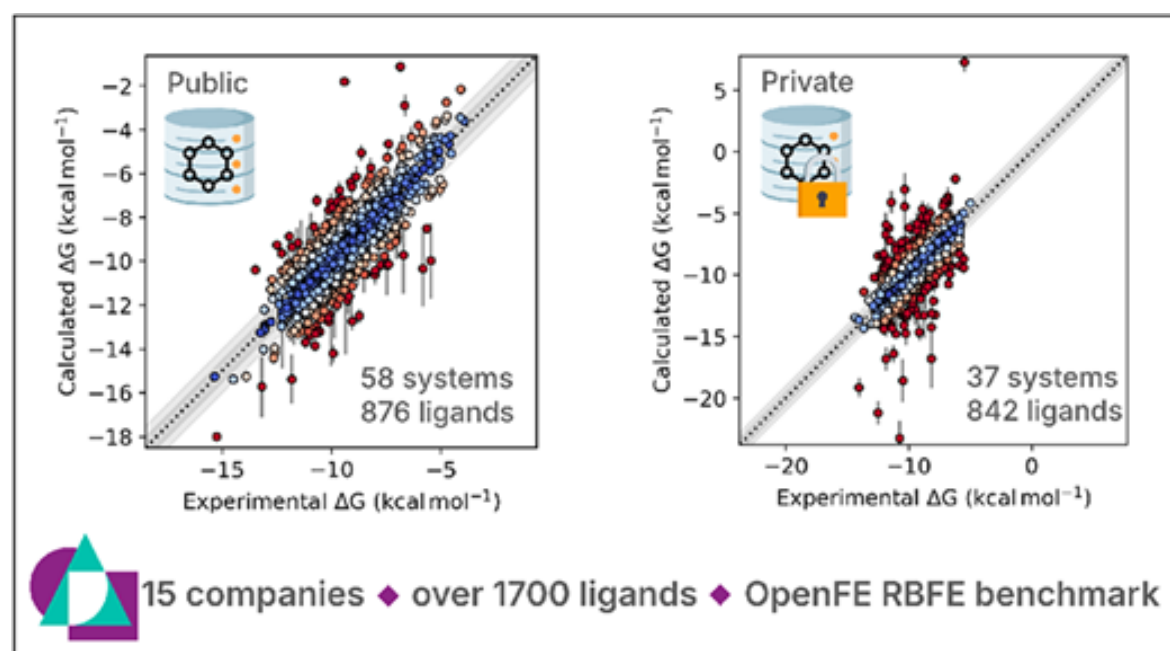
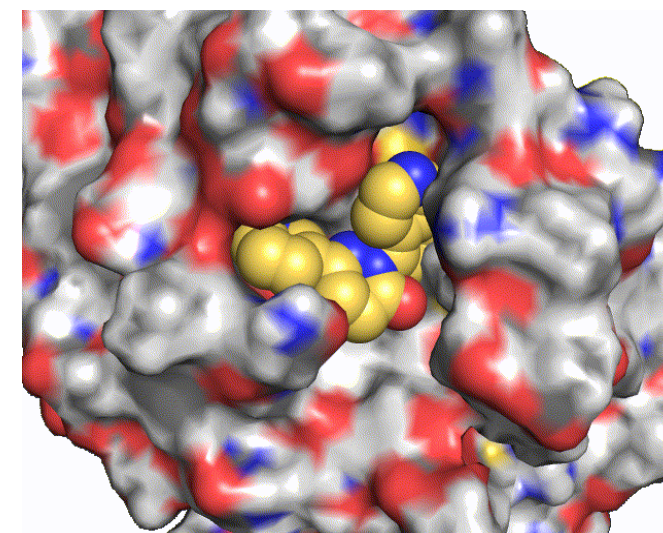
Soft-core MACE-OFF

Very promising accuracy and convergence for octanol-water partition coefficients using standard protocols in OpenMM:



ML in Biology & Drug Design

Many opportunities for ML in drug design, including agent-driven design/make/test cycles, but experimental data is relatively scarce — we still need good physical models of binding.



Open communities are providing the necessary infrastructure to design accurate physical models, but software maintenance is a challenge.

Opportunities also for specialised biological MLIP foundation models and software infrastructure for routine free energy calculations but accuracy needs to be rigorously tested.

Engineering Supercomputing Platforms for Biomolecular Applications

Robert Welch, Charles Laughton, Oliver Henrich, Tom Burnley, Daniel Cole, Alan Real,
Sarah Harris, James Gebbie-Rayet

Continue funding the consortium model for HPC communities.
Training users in best practice will take significant pressure off core
system administration teams.

Broaden participation in next generation machines to support the
underpinning physical models behind the development of AI
technologies.

Further investment in scientific software programmes that focus on
the maintenance and advancement of code. Continued investment
in the CCPs would be welcome.

<https://arxiv.org/html/2506.15585v1>

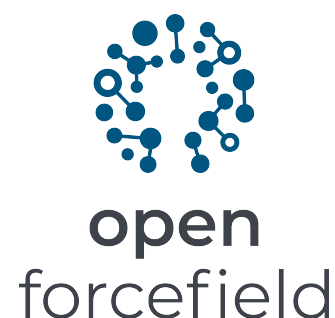


Acknowledgements

Exabiosim: Rob Welch, James Gebbie-Rayet, Sarah Harris, and many more.

Open Molecular Software Foundation: Finlay Clark, Josh Horton, Irfan Alibay, Chapin Cavender, Matt Thompson, Jeff Wagner, Lily Wang, David Mobley.

MACE-OFF: Harry Moore, Dávid Kovács, Nicholas Browning, Ilyes Batatia, Venkat Kapil, William Witt, Ioan-Bogdan Magdău, Gábor Csányi.



<https://blogs.ncl.ac.uk/danielcole/>
<https://github.com/cole-group/>
🦋 @ColeGroupNCL



Thank you for your attention

