open forcefield

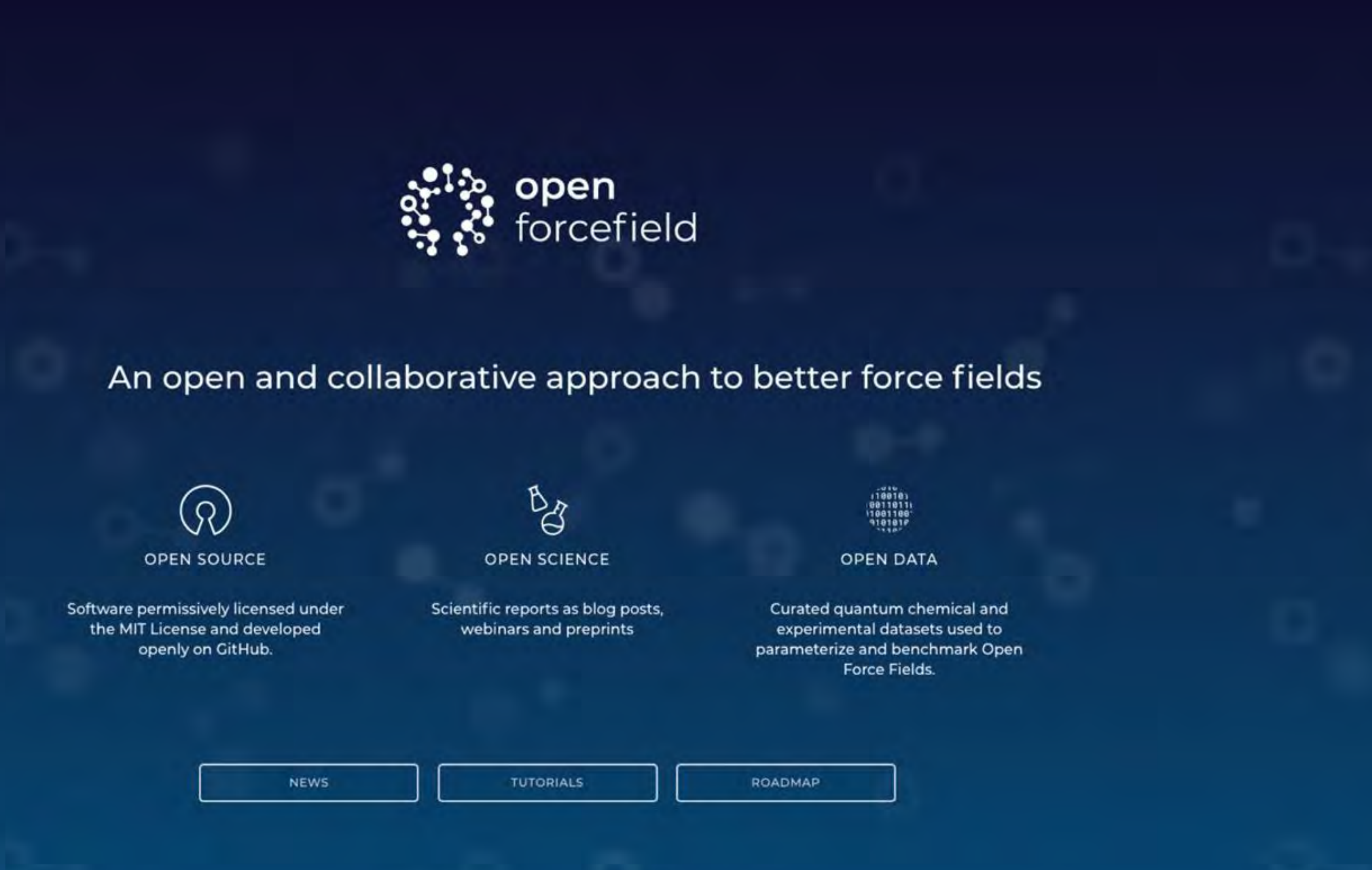
@openforcefield

www.openforcefield.org

Extending the QCArchive small molecule quantum chemistry archive to support machine learning applications in biomolecular modeling

John D. Chodera, Memorial Sloan Kettering Cancer Center

September 31, 2022 | PI: Michael R. Shirts



open forcefield

An open and collaborative approach to better force fields

OPEN SOURCE

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

OPEN SCIENCE

Scientific reports as blog posts, webinars and preprints

OPEN DATA

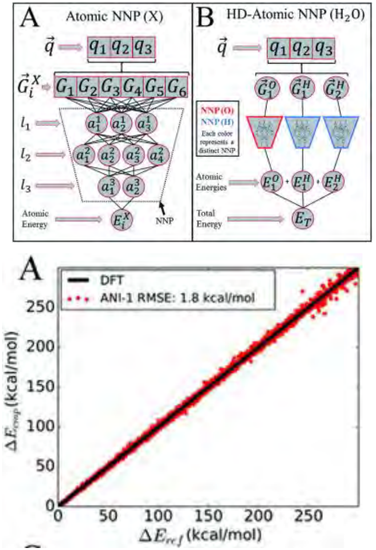
Curated quantum chemical and experimental datasets used to parameterize and benchmark Open Force Fields.

OpenFF is generating an enormous amount of quantum chemical data.
How do we maximize utility of this data to the community?

Biomolecular modeling is powered by quantum chemistry

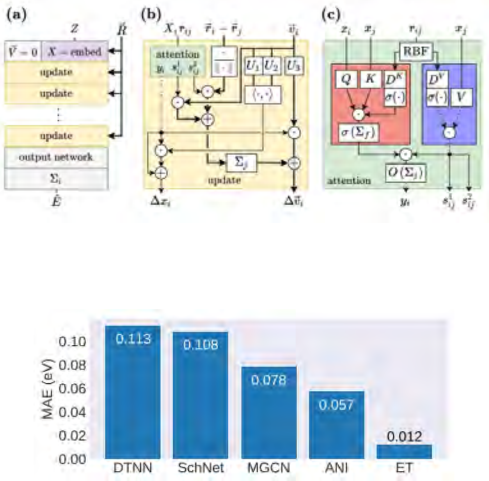
QM data powers a wide range of ML applications for bioscience

QM accuracy
but 10^6 times cheaper



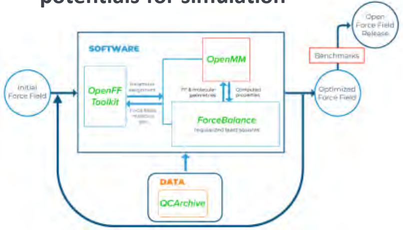
ANI-1x
https://doi.org/10.1039/C6SC05720A

fast machine learning
potentials for simulation

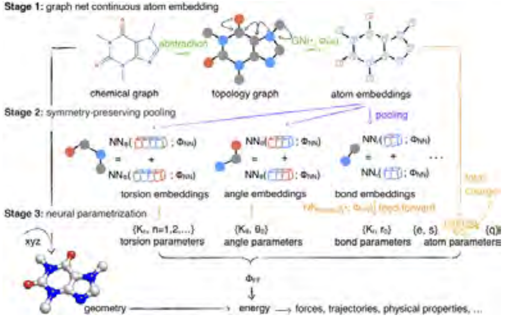


TorchMD-Net
https://arxiv.org/abs/2012.12106

ultra-fast molecular mechanics
potentials for simulation

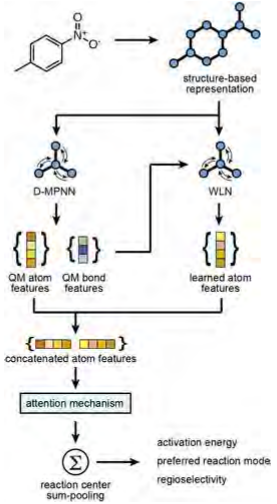


OpenFF 1.0.0 (“Parsley”)
https://doi.org/10.1021/acs.jctc.1c00571



espaloma
https://doi.org/10.1039/D2SC02739A

foundation/pretrained models
for drug discovery



ml-QM-GNN
https://doi.org/10.1063/5.0079574

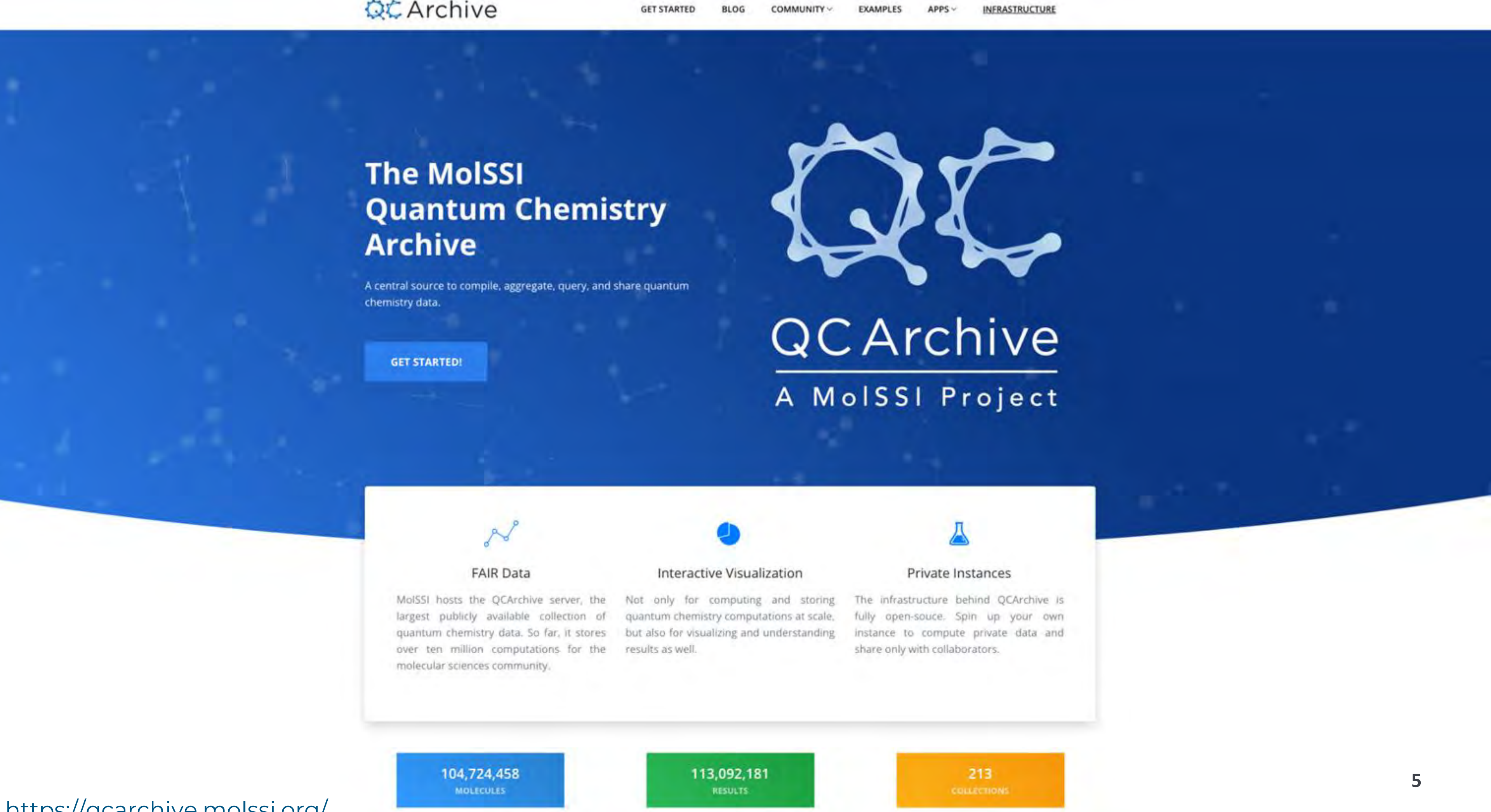
Expanding QCArchive is a collaborative effort

Open Force Field Consortium

Molecular Sciences Software Institute (MolSSI)

OpenMM molecular simulation framework

TorchMDNet deep learning framework for molecular simulations
(and other communities, e.g. SchNetPack, ANI, …)



The MolSSI Quantum Chemistry Archive

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

FAIR Data

MolSSI hosts the QCArchive 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

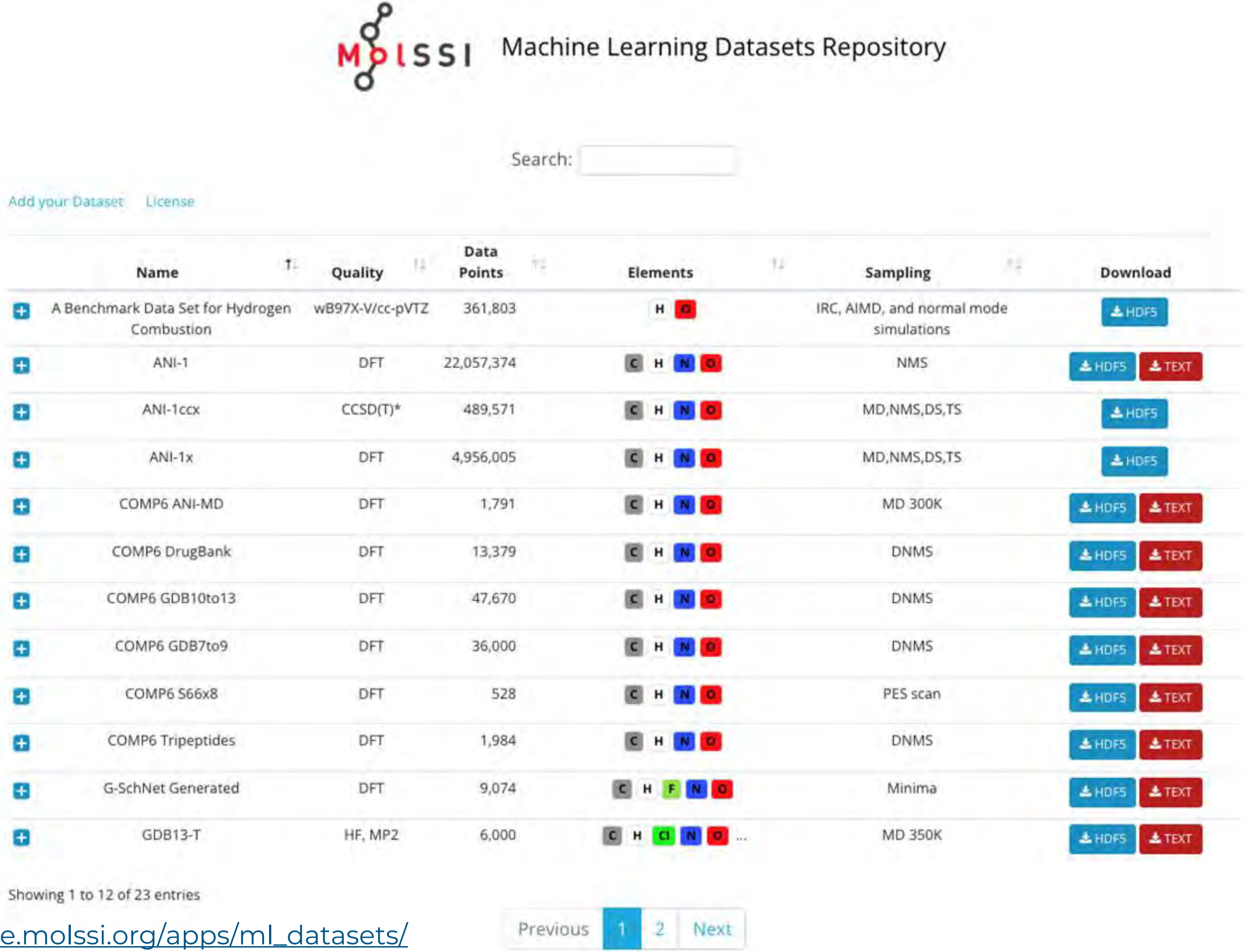
Not only for computing and storing quantum chemistry computations at scale, but also for visualizing and understanding results as well.

Private Instances

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

104,724,458 MOLECULES	113,092,181 RESULTS	213 COLLECTIONS

https://qcarchive.molssi.org/



MolSSI Machine Learning Datasets Repository

Search:

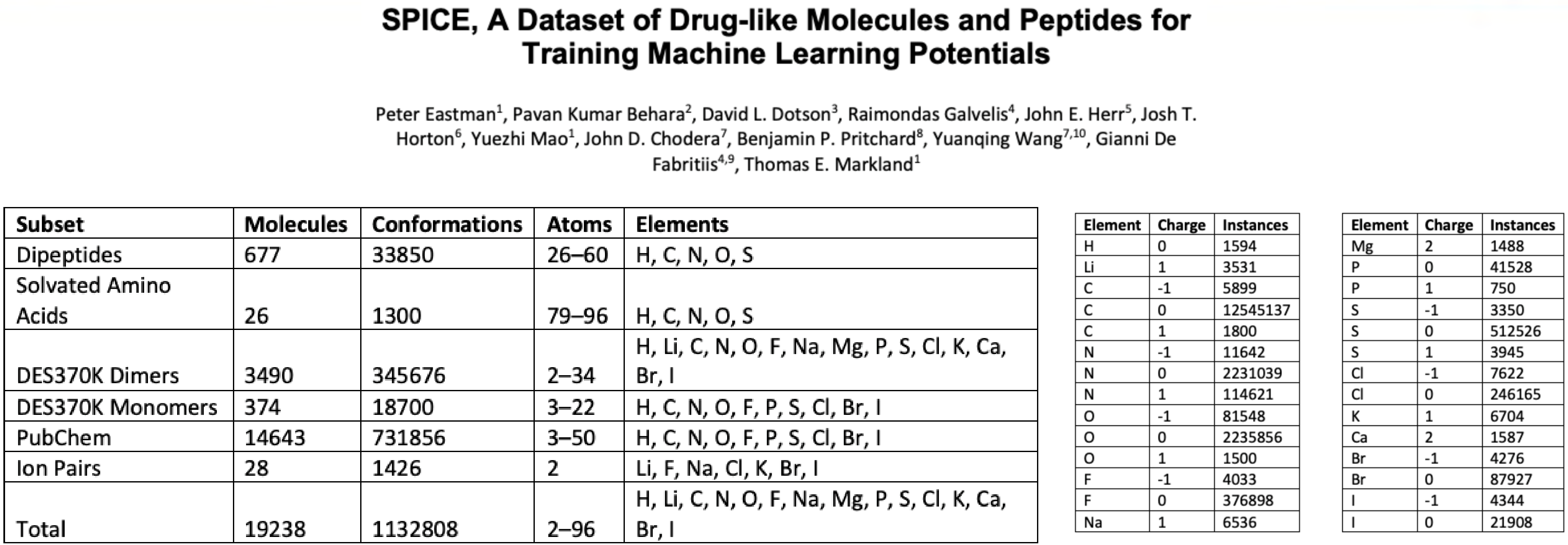
Add your Dataset License

Name	Quality	Data Points	Elements	Sampling	Download
A Benchmark Data Set for Hydrogen Combustion	wB97X-V/cc-pVTZ	361,803	H O	IRC, AIMD, and normal mode simulations	HDF5
ANI-1	DFT	22,057,374	C H N O	NMS	HDF5 TEXT
ANI-1ccx	CCSD(T)*	489,571	C H N O	MD,NMS,DS,TS	HDF5
ANI-1x	DFT	4,956,005	C H N O	MD,NMS,DS,TS	HDF5
COMP6 ANI-MD	DFT	1,791	C H N O	MD 300K	HDF5 TEXT
COMP6 DrugBank	DFT	13,379	C H N O	DNMS	HDF5 TEXT
COMP6 GDB10to13	DFT	47,670	C H N O	DNMS	HDF5 TEXT
COMP6 GDB7to9	DFT	36,000	C H N O	DNMS	HDF5 TEXT
COMP6 S66x8	DFT	528	C H N O	PES scan	HDF5 TEXT
COMP6 Tripeptides	DFT	1,984	C H N O	DNMS	HDF5 TEXT
G-SchNet Generated	DFT	9,074	C H F N O	Minima	HDF5 TEXT
GDB13-T	HF, MP2	6,000	C H Cl N O ...	MD 350K	HDF5 TEXT

Showing 1 to 12 of 23 entries

https://qcarchive.molssi.org/apps/ml_datasets/

SPICE: A proof of concept foundation dataset



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

Peter Eastman[1], Pavan Kumar Behara[2], David L. Dotson[3], Raimondas Galvelis[4], John E. Herr[5], Josh T. Horton[6], Yuezhi Mao[1], John D. Chodera[7], Benjamin P. Pritchard[8], Yuanqing Wang[7,10], Gianni De Fabritiis[4,9], Thomas E. Markland[1]

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

Element	Charge	Instances
H	0	1594
Li	1	3531
C	-1	5899
C	0	12545137
C	1	1800
N	-1	11642
N	0	2231039
N	1	114621
O	-1	81548
O	0	2235856
O	1	1500
F	-1	4033
F	0	376898
Na	1	6536

Element	Charge	Instances
Mg	2	1488
P	0	41528
P	1	750
S	-1	3350
S	0	512526
S	1	3945
Cl	-1	7622
Cl	0	246165
K	1	6704
Ca	2	1587
Br	-1	4276
Br	0	87927
I	-1	4344
I	0	21908

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

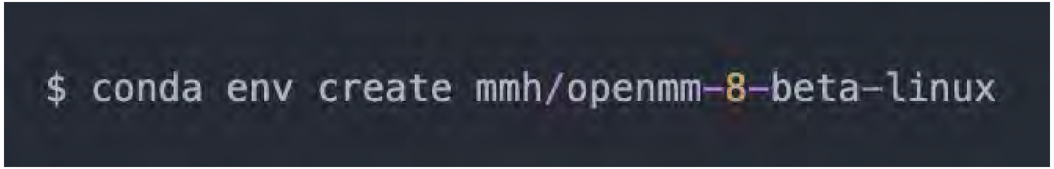
>4M core-hours computed on QCFractal academic clusters

https://arxiv.org/abs/2209.10702

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

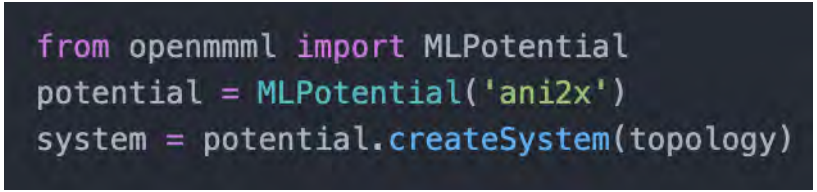
We are taking the first steps toward an ML model repository to easily deploy to users!

Install the OpenMM 8 beta and our interface to the ML model repository via **conda**



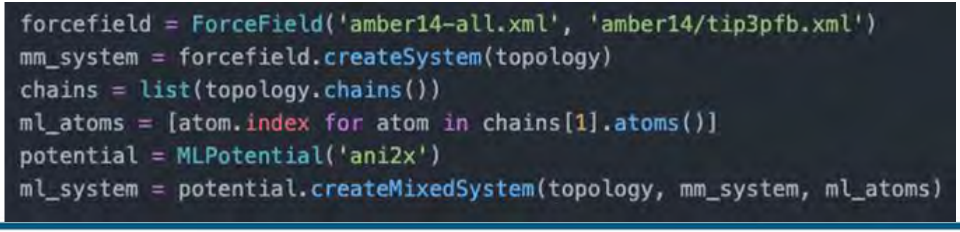
```
$ conda env create mmh/openmm-8-beta-linux
```

Check out the ANI-2x ML model to run a simulation!



```
from openmmml import MLPotential
potential = MLPotential('ani2x')
system = potential.createSystem(topology)
```

Or run a hybrid simulation:



```
forcefield = ForceField('amber14-all.xml', 'amber14/tip3pfb.xml')
mm_system = forcefield.createSystem(topology)
chains = list(topology.chains())
ml_atoms = [atom.index for atom in chains[1].atoms()]
potential = MLPotential('ani2x')
ml_system = potential.createMixedSystem(topology, mm_system, ml_atoms)
```

http://openmm.org

OpenMM 8 beta and the openmm-ml ecosystem: https://tinyurl.com/openmm-8-beta

Acknowledgments

…with ***much more*** to come!

Peter Eastman[1], Pavan Kumar Behara[2], David L. Dotson[3], Raimondas Galvelis[4], John E. Herr[5], Josh T. Horton[6], Yuezhi Mao[1], John D. Chodera[7], Benjamin P. Pritchard[8], Yuanqing Wang[7,10], Gianni De Fabritiis[4,9], Thomas E. Markland[1]

[1]Department of Chemistry, Stanford University, Stanford, CA 94305, USA
[2]Department of Pharmaceutical Sciences, University of California, Irvine, CA 92697, USA
[3]The Open Force Field Initiative, Open Molecular Software Foundation, Davis, CA 95616, USA
[4]Acellera Labs, Doctor Trueta 183, 08005, Barcelona, Spain
[5]Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, IN 46556, USA
[6]School of Natural and Environmental Sciences, Newcastle University, Newcastle upon Tyne NE1 7RU, United Kingdom
[7]Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
[8]Molecular Sciences Software Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA 24060, USA
[9]Computational Science Laboratory, Universitat Pompeu Fabra, Barcelona Biomedical Research Park (PRBB), Carrer Dr. Aiguader 88, 08003, Barcelona, Spain and ICREA, Passeig Lluis Companys 23, 08010 Barcelona, Spain.
[10]Graduate Program in Physiology, Biophysics, and Systems Biology, Weill Cornell Graduate School of Medical Sciences, New York, NY 10065, USA