Parameterization Perspective II: The Property Estimator

Simon Boothroyd

THE GOAL: AUTOMATED, RAPID, AND SCALABLE ESTIMATION OF PHYSICAL PROPERTY DATA

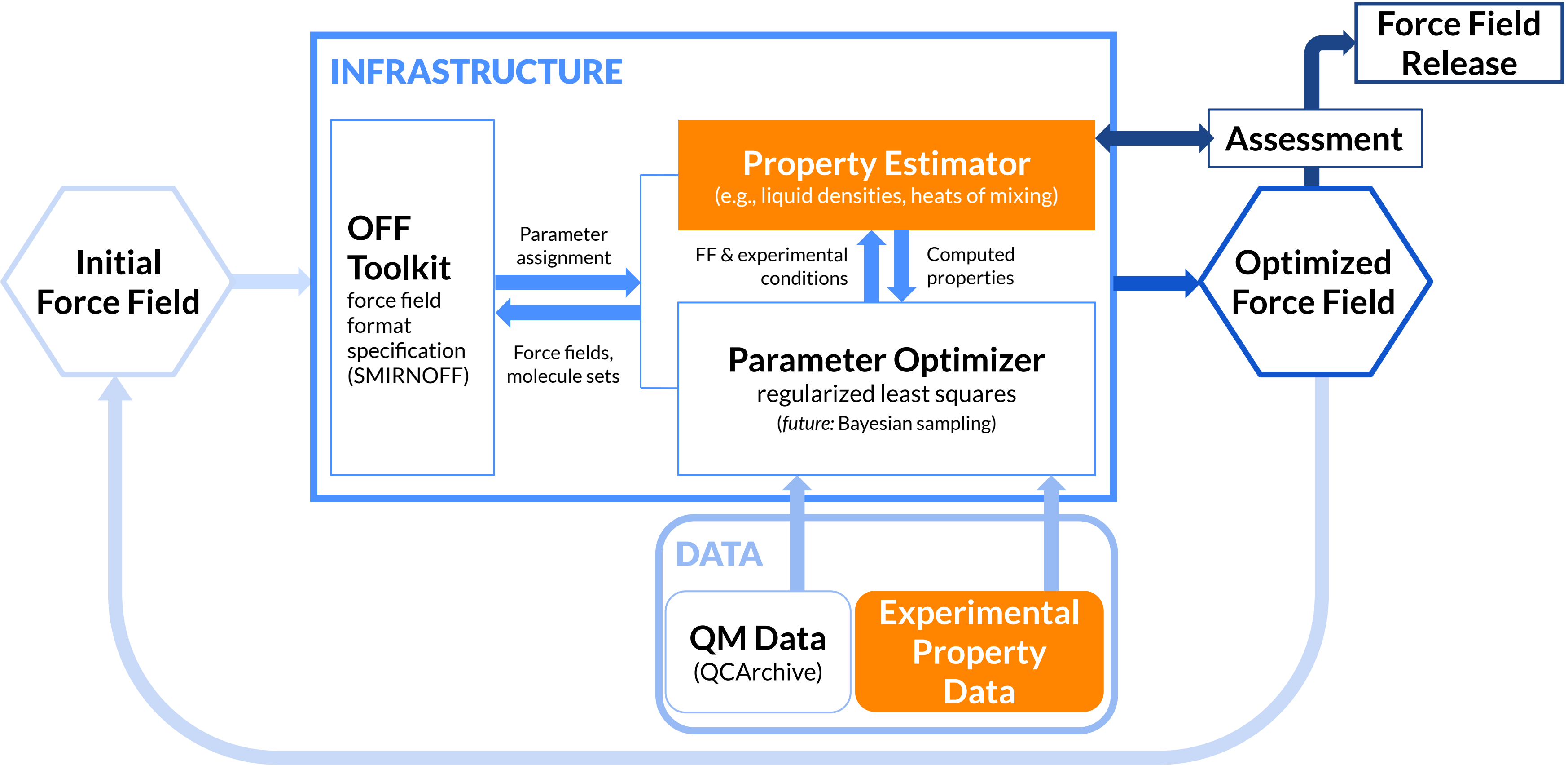
PROVIDE TOOLS FOR CURATING AND ANALYSING DATA SETS OF PHYSICAL PROPERTIES

AUTOMATED, FLEXIBLE AND EXTENSIBLE ESTIMATION OF PROPERTIES AND THEIR GRADIENTS W.R.T PARAMETERS

MODULAR ARCHITECTURE WHICH CAN SCALE ACROSS SITES AND INTO THE CLOUD

Icons modified from the Font Awesome collection (https://fontawesome.com/license/free)

FORMS A KEY COMPONENT OF THE OPENFF ASSESSMENT AND OPTIMISATION PIPELINE


Initial Force Field
INFRASTRUCTURE
OFF Toolkit
force field format specification (SMIRNOFF)
Parameter assignment
Force fields, molecule sets
Property Estimator
(e.g., liquid densities, heats of mixing)
FF & experimental conditions
Computed properties
Parameter Optimizer
regularized least squares
(future: Bayesian sampling)
DATA
QM Data
(QCArchive)
Experimental Property Data
Assessment
Force Field Release
Optimized Force Field

BUILT IN UTILITIES FOR EXTRACTING DATA FROM THE THERMOML ARCHIVE

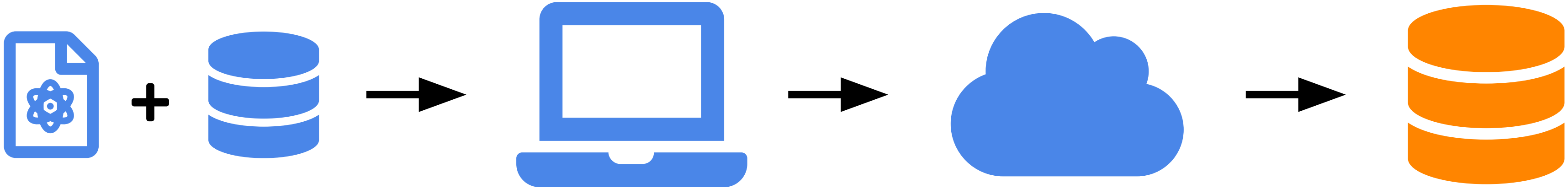
Property	Number of Data Points (in Thousands)		
	Pure	Binary	Ternary
Density	63.6	224.5	97.3
Vapor Pressure	26.3	57.3	7.2
Enthalpy of Vaporization	0.5	0	0
Enthalpy of Mixing	-	30.2	2.1
Dielectric Coefficient	1	3.2	0.2
Surface Tension	2.6	5.4	1
Activity Coefficient	30.3	3.1	4.3
Heat Capacity	14.1	16.7	2.9

EXTRACTING AND CURATING DATA FROM THE THERMOML ARCHIVE

```
# Load in the data sets of interest directly from their DOIs.
data_set = ThermoMLDataSet.from_doi(*dois_of_interest)

# Filter properties measured outside of the temperature and pressure range of interest
data_set.filter_by_temperature(min_temperature=298 * unit.kelvin,
                               max_temperature=350 * unit.kelvin)

data_set.filter_by_pressure(min_pressure=0.5 * unit.atmosphere,
                            max_pressure=1.1 * unit.atmosphere)

# Filter properties measured for substances containing anything other than C, N, O and H
data_set.filter_by_elements('C', 'N', 'O', 'H')

# Filter any properties measured for substances containing more than one component
data_set.filter_by_components(number_of_components=1)
```

https://github.com/openforcefield/nistdataselection

ESTIMATION REQUESTS MADE FROM A ‘CLIENT’, AND EXECUTED ON A COMPUTE ‘SERVER’



FORCE FIELD + DATA SET

Framework used to extract and curate data directly from source e.g. NIST ThermoML, BindingDB

ESTIMATOR CLIENT

Estimator client automatically maps physical properties to calculation workflows

ESTIMATOR SERVER

Server determines fastest estimation approach and distributes calculation to compute workers

ESTIMATED DATA SET

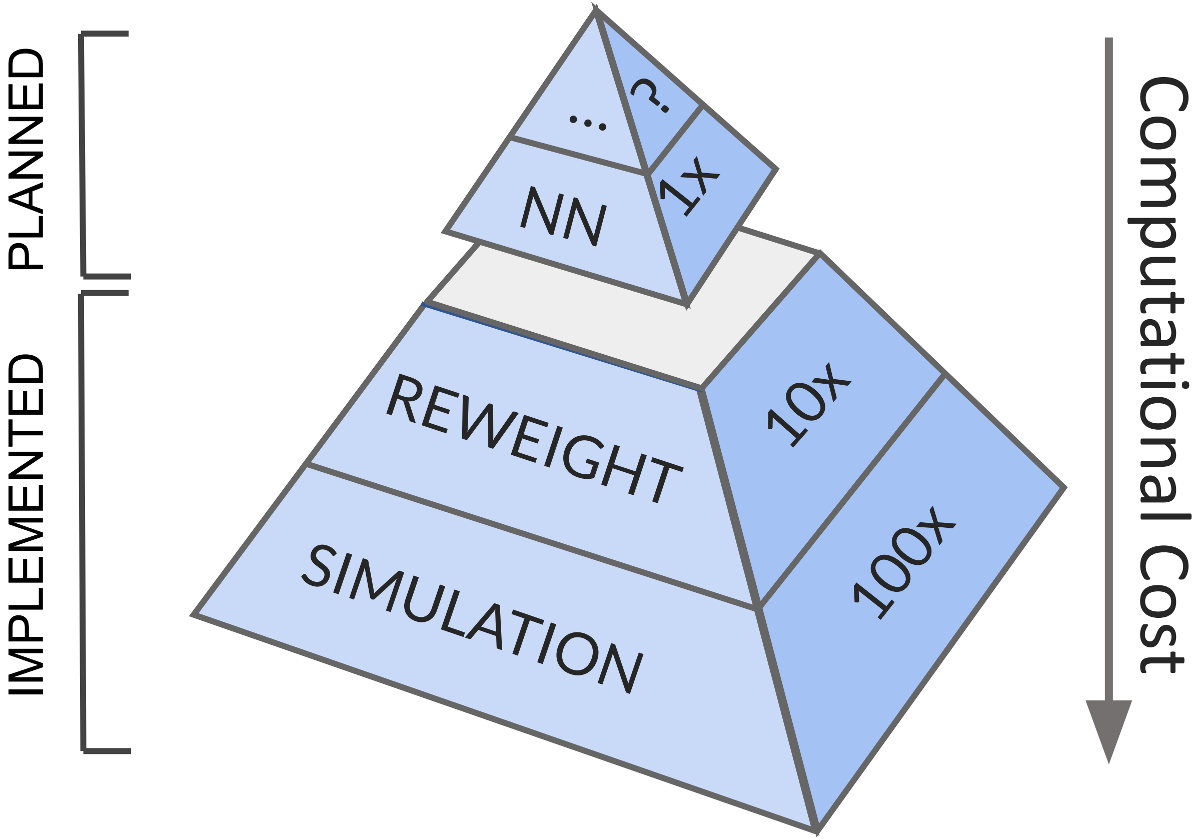
Estimated data set returned to the client object / user

Icons modified from the Font Awesome collection (https://fontawesome.com/license/free)

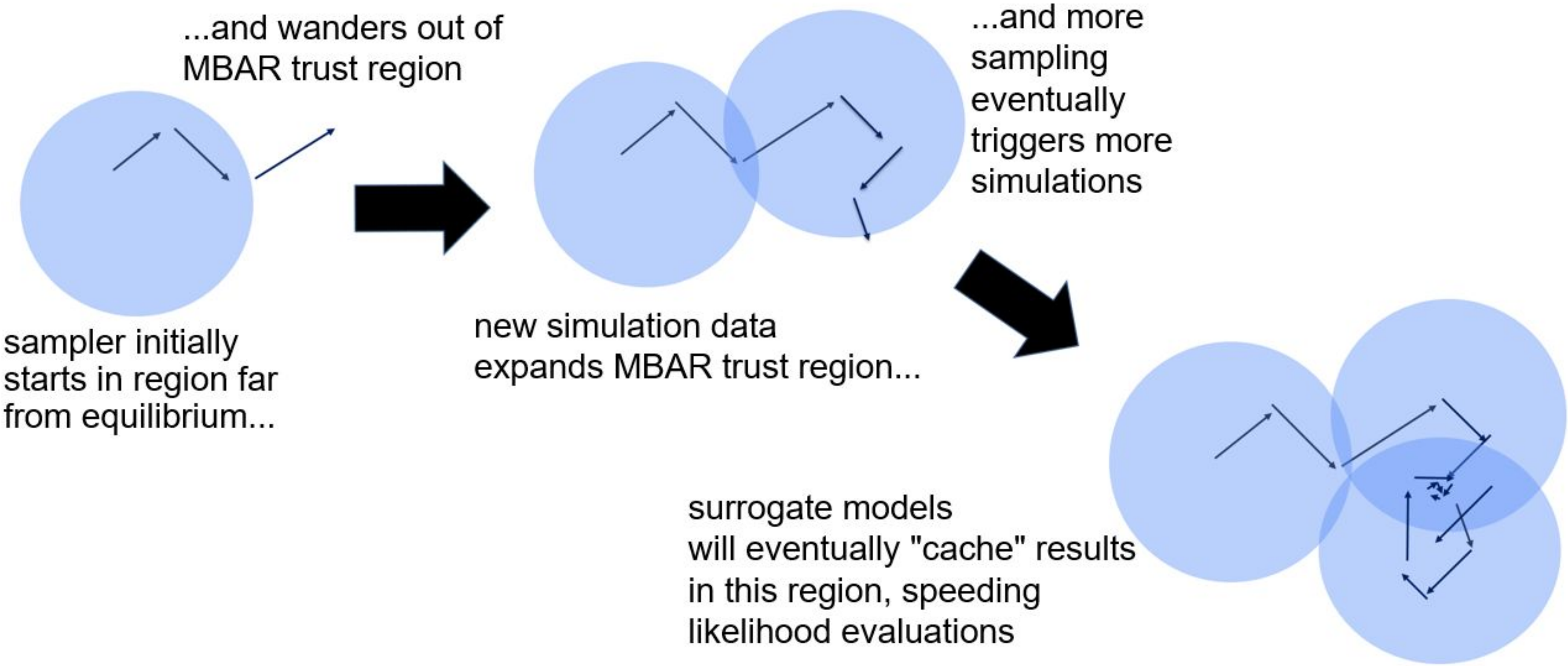
ESTIMATES OF PROPERTY DATA SETS CAN BE REQUESTED IN JUST 6 LINES OF PYTHON

```
# Load in the data set of interest.
data_set = ThermoMLDataSet.from_file('10.1016/j.jct.2016.10.001')
data_set.filter_by_properties(Density, DielectricConstant)

# Load in the force field parameters to use.
force_field = smirnoff.ForceField('smirnoff99Frosst-1.1.0.offxml')

# Create the client object.
property_estimator = PropertyEstimatorClient()
# Submit the request to a running server.
request = property_estimator.request_estimate(data_set, force_field)

# Wait for the results.
results = request.results(synchronous=True)
```

A MULTI FIDELITY APPROACH IS EMPLOYED TO ESTIMATE PROPERTIES AS RAPIDLY AS POSSIBLE



FACILITATES CHEAP AND ACCURATE EVALUATION OF OBJECTIVE FUNCTION



sampler initially starts in region far from equilibrium...

...and wanders out of MBAR trust region

new simulation data expands MBAR trust region...

...and more sampling eventually triggers more simulations

surrogate models will eventually "cache" results in this region, speeding likelihood evaluations

DOI: 10.13140/RG.2.2.19199.25767

ESTIMATION WORKFLOWS CONSTRUCTED FROM MODULAR, EXTENSIBLE PROTOCOLS

COORDINATE GENERATION

- PACKMOL COORDINATES
- SOLVATE STRUCTURE
- OEDOCK

DATA ANALYSIS

- DECORRELATE STATISTICS
- BOOTSTRAP AVERAGE
- DECORRELATE TRAJECTORY

SIMULATION

- ENERGY MINIMISATION
- OPENMM SIMULATION

PARAMETER ASSIGNMENT

- APPLY SMIRNOFF
- RUN TLEAP
- RUN LIGPARGEN

REWEIGHTING

- MBAR REWEIGHT
- REDUCED POTENTIAL

FREE ENERGY CALCULATIONS

- YANK HOST GUEST
- YANK SOLVATION
- PAPRIKA

STORAGE

- UNPACK CACHED DATA
- UNPACK DATA COLLECTION

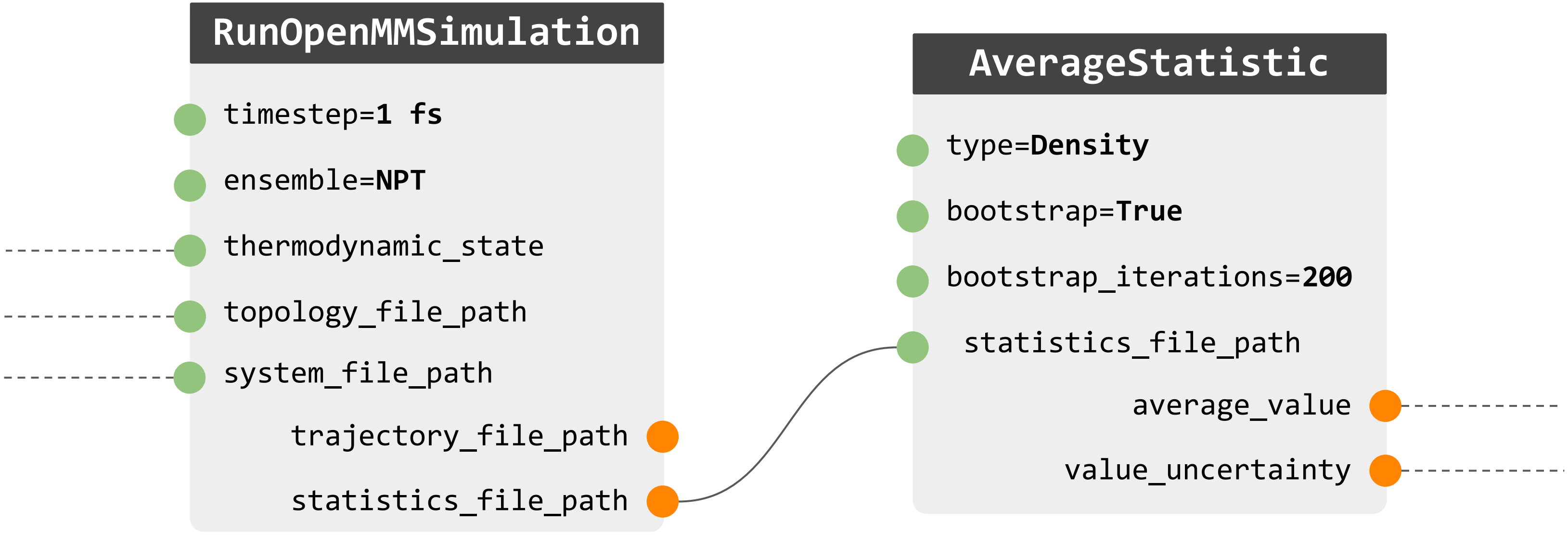
GROUPS

- EXECUTION GROUP
- CONDITIONAL GROUP

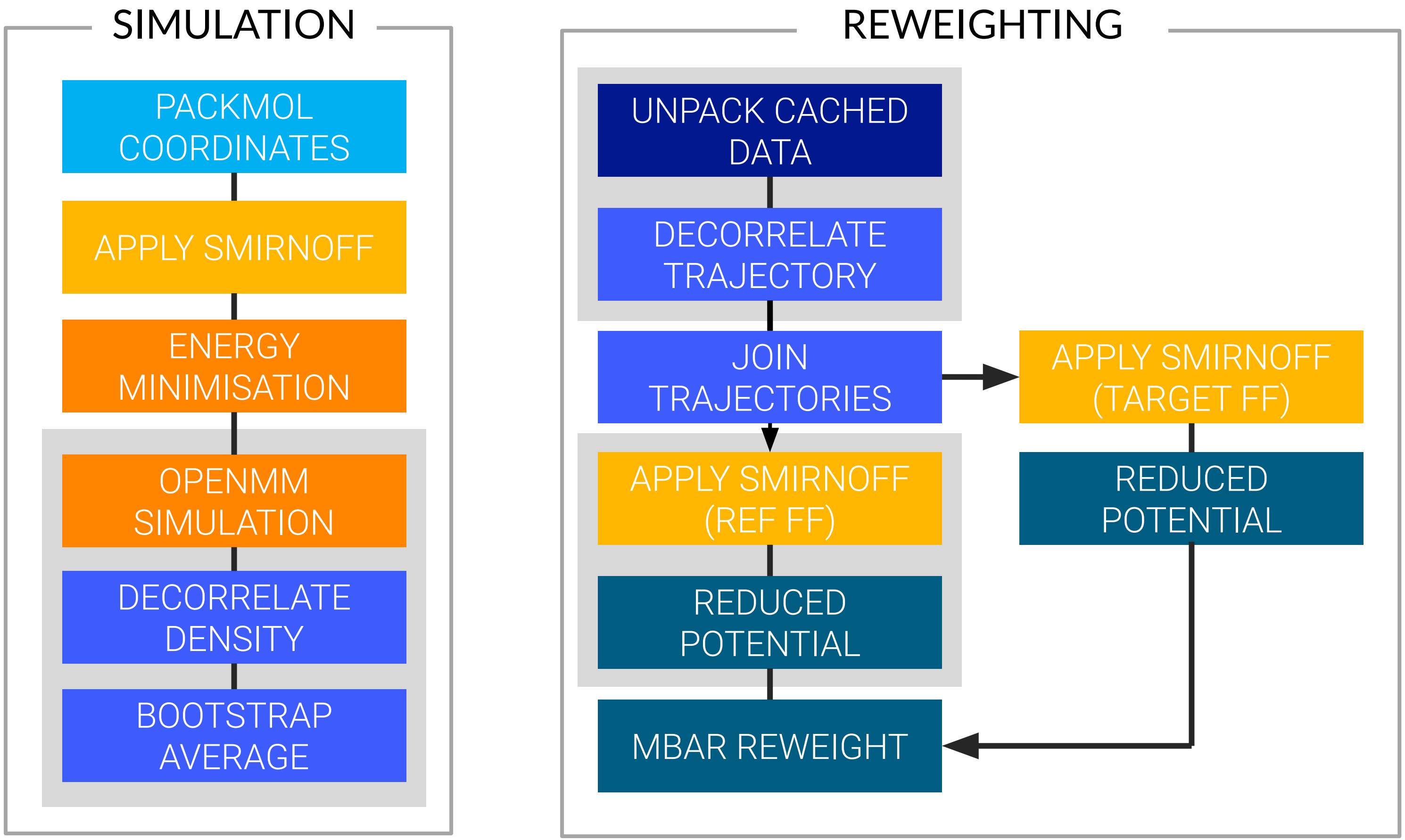
AND MORE

…

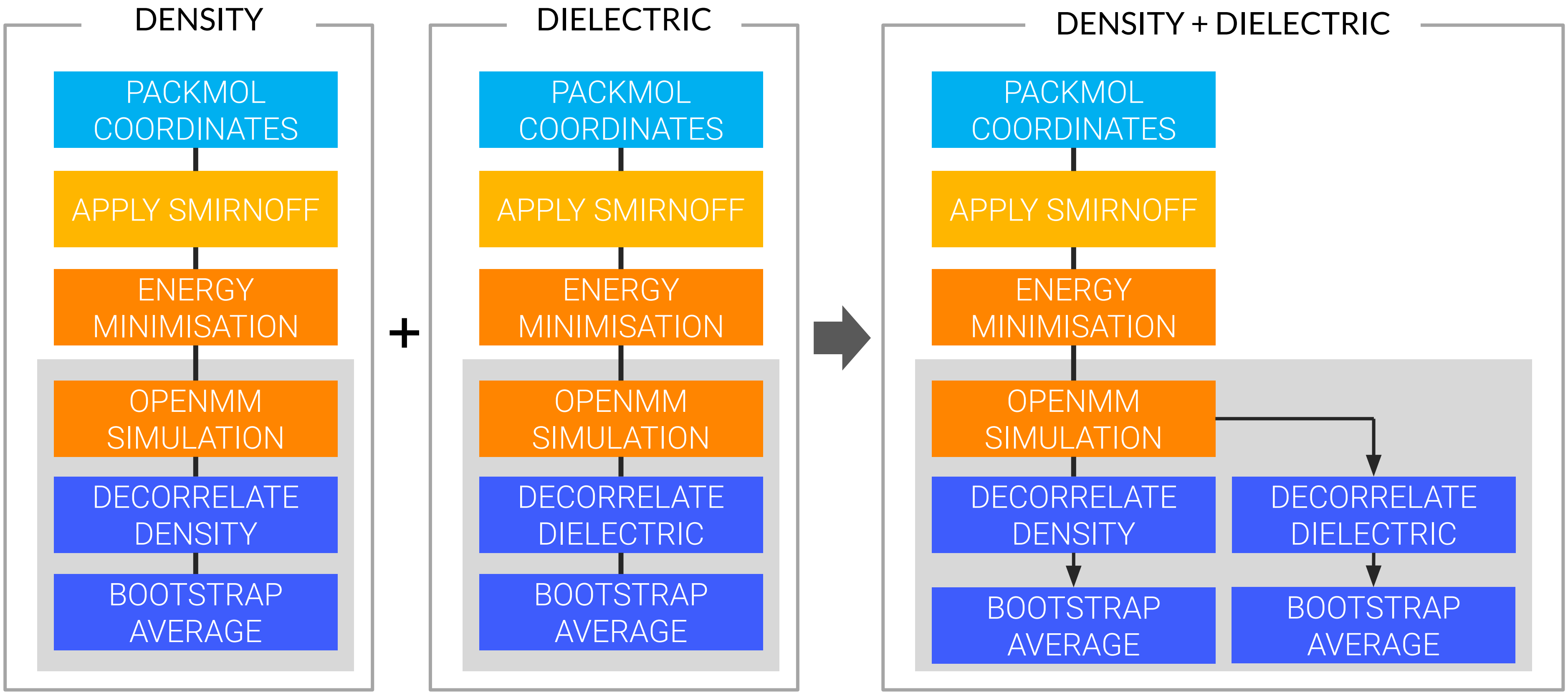
PROTOCOLS COMBINED TOGETHER IN WORKFLOWS BY INPUT / OUTPUT ‘SOCKETS’



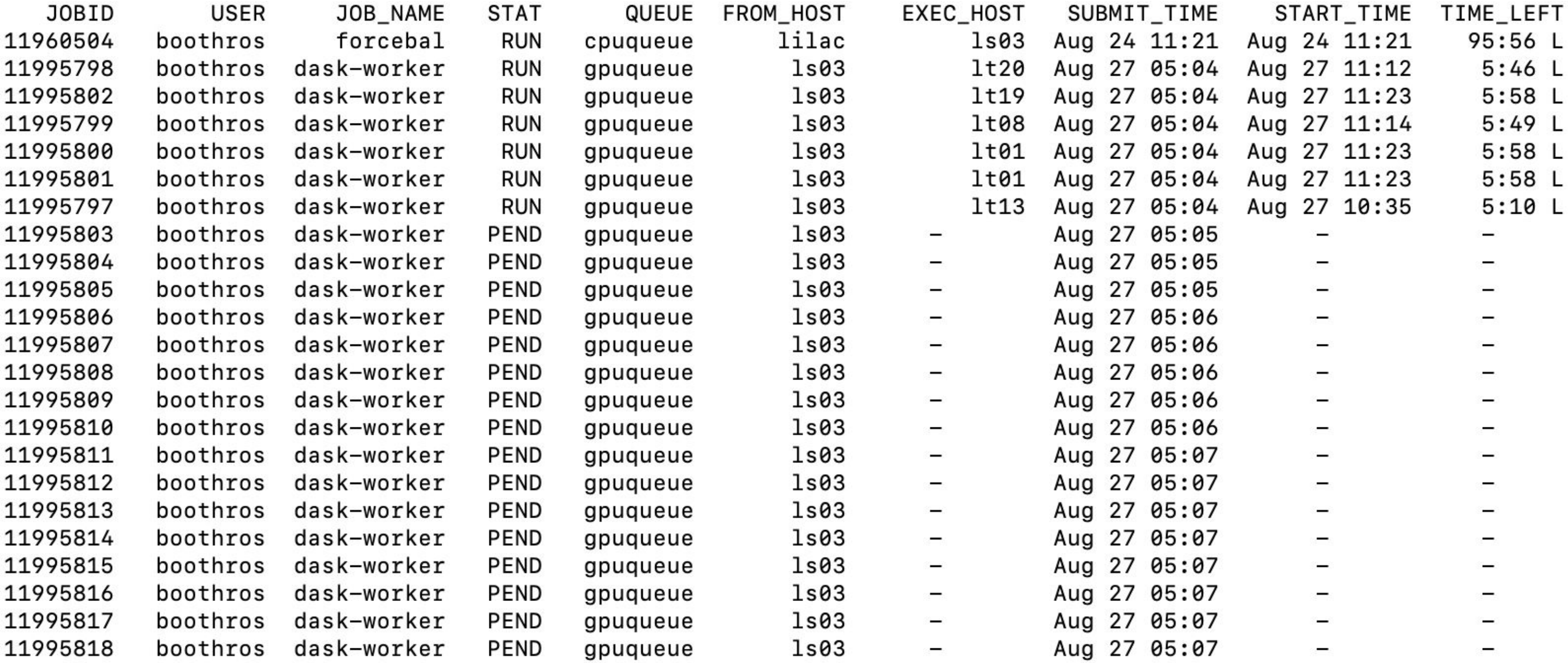
EXAMPLE DENSITY ESTIMATION WORKFLOWS



WORKFLOWS AUTOMATICALLY MERGED WHERE POSSIBLE TO REDUCE SIMULATION EXPENSE



CURRENTLY USING DASK FOR SCALABLE, ADAPTIVE, AND DISTRIBUTED COMPUTING



```
   JOBID       USER     JOB_NAME   STAT      QUEUE  FROM_HOST    EXEC_HOST   SUBMIT_TIME    START_TIME  TIME_LEFT
11960504   boothros     forcebal    RUN   cpuqueue      lilac         ls03  Aug 24 11:21  Aug 24 11:21    95:56 L
11995798   boothros  dask-worker    RUN   gpuqueue       ls03         lt20  Aug 27 05:04  Aug 27 11:12     5:46 L
11995802   boothros  dask-worker    RUN   gpuqueue       ls03         lt19  Aug 27 05:04  Aug 27 11:23     5:58 L
11995799   boothros  dask-worker    RUN   gpuqueue       ls03         lt08  Aug 27 05:04  Aug 27 11:14     5:49 L
11995800   boothros  dask-worker    RUN   gpuqueue       ls03         lt01  Aug 27 05:04  Aug 27 11:23     5:58 L
11995801   boothros  dask-worker    RUN   gpuqueue       ls03         lt01  Aug 27 05:04  Aug 27 11:23     5:58 L
11995797   boothros  dask-worker    RUN   gpuqueue       ls03         lt13  Aug 27 05:04  Aug 27 10:35     5:10 L
11995803   boothros  dask-worker   PEND   gpuqueue       ls03      -        Aug 27 05:05       -             -
11995804   boothros  dask-worker   PEND   gpuqueue       ls03      -        Aug 27 05:05       -             -
11995805   boothros  dask-worker   PEND   gpuqueue       ls03      -        Aug 27 05:05       -             -
11995806   boothros  dask-worker   PEND   gpuqueue       ls03      -        Aug 27 05:06       -             -
11995807   boothros  dask-worker   PEND   gpuqueue       ls03      -        Aug 27 05:06       -             -
11995808   boothros  dask-worker   PEND   gpuqueue       ls03      -        Aug 27 05:06       -             -
11995809   boothros  dask-worker   PEND   gpuqueue       ls03      -        Aug 27 05:06       -             -
11995810   boothros  dask-worker   PEND   gpuqueue       ls03      -        Aug 27 05:06       -             -
11995811   boothros  dask-worker   PEND   gpuqueue       ls03      -        Aug 27 05:07       -             -
11995812   boothros  dask-worker   PEND   gpuqueue       ls03      -        Aug 27 05:07       -             -
11995813   boothros  dask-worker   PEND   gpuqueue       ls03      -        Aug 27 05:07       -             -
11995814   boothros  dask-worker   PEND   gpuqueue       ls03      -        Aug 27 05:07       -             -
11995815   boothros  dask-worker   PEND   gpuqueue       ls03      -        Aug 27 05:07       -             -
11995816   boothros  dask-worker   PEND   gpuqueue       ls03      -        Aug 27 05:07       -             -
11995817   boothros  dask-worker   PEND   gpuqueue       ls03      -        Aug 27 05:07       -             -
11995818   boothros  dask-worker   PEND   gpuqueue       ls03      -        Aug 27 05:07       -             -
```

https://github.com/dask/dask-jobqueue

ALL CODE AVAILABLE TO DOWNLOAD ON GITHUB

THE PROPERTY ESTIMATOR

https://github.com/openforcefield/propertyestimator

NIST DATA CURATION REPO

https://github.com/openforcefield/nistdataselection

THE NEXT STEPS...

EXPAND THE AVAILABLE DATA CURATION AND ANALYSIS TOOLS

CONTINUE TO EXPAND, TEST AND AUTOMATE THE BUILT IN PROPERTIES

- Host Guest Binding Affinities (pAPRika + YANK) - Automated Setup?
- Absolute and Relative Solvation Free Energies (YANK / PERSES)
- Vapor Pressures

IMPLEMENT SURROGATE MODELLING FOR MARKED PERFORMANCE IMPROVEMENT

EXTEND THE DISTRIBUTED ARCHITECTURE TO MULTIPLE CLUSTERS