Property Calculator

Breakout session 2 - 01/07/2019

Databases

ThermoML Archive – physical properties data

BindingDB – binding data (including protein:ligand and host:guest)

Cambridge crystallographic database (CCSD) – need to develop APIs to pull our data. Maybe proprietary data (not open).

PubChem – QM data on small molecules

Issues

Licence issues

Quality of experimental data

Mole-fraction vs mass-fraction

Unexpected failures access to trajectories

Multiple simulations per GPU - maximise efficiency. Different threads/same threads

Summit

Best Practices

Develop a best practices document to go along with the work

Develop a way to compute from simulation - surrogate models, reweighting methods

E.g. density <N/V> vs N/<V> different ways to compute things

Have module plug into framework, so it can be used for benchmarking

Fit ---> benchmark ---> release

Will be including successively more data in fit and benchmark each time - what are the limitation computationally of what can be included?

Layers of API

Public API - how to give parameters as inputs, parameters and expectations

Property plugin API - measurement type

Need to define convergence criteria - experimental error, relative error?

Compute things independently for best practices

Store metadata/recipe for calculation criteria for reproducibility

Working out **property estimation layers API** soon -
simulations (1,000,000x cost); MBAR (1,000x cost); GP, NN (1x cost)

How are we going to interact with the databases

Will I be able to query all the density data for a specific molecule of my choice

Connect to a data source - filter by pressure, temperature, can specify DOIs

Object model

User can apply filtering steps - filtering out phase, uncertainties, filtering API will be quite extensive, can use predetermined filters or write your own code. Not far from ambient conditions.

Building Systems

Liquid systems

Host-guest

Encode best practices in set up

Questions/Comments

Ambient conditions?

Weight by recevences?

What is forcefield is worse?

Data quality?

How do we want to specify what the stop conditions are?

(Montecarlo.sourceforge.net)

New Use Cases

Predictions of measurements for things that haven't been measured