Large-scale collaborative assessment of binding free energy calculations for drug discovery using OpenFE

Irfan Alibay[1], Hannah M. Baumann[1], James RB Eastwood[1], Mike M. Henry[1,2], Josh Horton[1], Benjamin Ries[1,3], Alyssa Travitz[1], David Mobley[1,4]

1. *Open Free Energy, Open Molecular Software Foundation, Davis, California, USA*
2. *Chodera Lab, Sloan Kettering Institute, MSKCC, New York, NY, USA*
3. *Boehringer Ingelheim Pharma GmbH & Co KG, Medicinal Chemistry, DE*
4. *Department of Pharmaceutical Sciences, University of California, Irvine, USA*

Open Free Energy

openfree.energy

Introduction

Here we present the outcomes of a large-scale collaborative benchmark of relative alchemical binding free energies using the OpenFE toolkit (https://openfree.energy/).

This benchmark, carried out alongside **15 partner companies**, comprises of **1740 protein-ligand complexes** over **96 datasets**. It represents one of the largest evaluations of alchemical free energy calculations in a drug discovery context using **fully open source software**.

Benchmark details

This benchmark comprises of two datasets:

1. **Public dataset**: 58 systems, totalling 876 ligands, from the Ross et al. (2023) Schrodinger public dataset.
2. **Private dataset:** 37 systems totalling 864 ligands, from in-house (blinded) datasets by 10 partner companies.

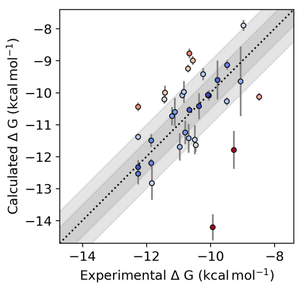
All simulations were carried out using the OpenFE hybrid topology Protocol using default settings:

- OpenFF 2.2.0 & FF14SB
- Kartograf atom mapping
- HREX sampling
- 11 lambda windows (22 with net charges)
- 5 ns / window (20 with net charges
- Triplicate repeats

Some lessons learned

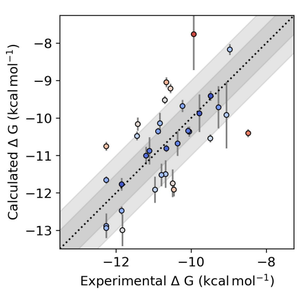
Before

```
                p38 (N = 34)
RMSE:   1.25 [95%: 0.85, 1.65]
MUE:    0.93 [95%: 0.66, 1.23]
R2:     0.18 [95%: 0.02, 0.53]
rho:    0.43 [95%: 0.09, 0.72]
KTAU:   0.31 [95%: 0.07, 0.54]
```



After

```
                p38 (N = 34)
RMSE:   0.97 [95%: 0.77, 1.15]
MUE:    0.81 [95%: 0.64, 0.98]
R2:     0.42 [95%: 0.22, 0.64]
rho:    0.65 [95%: 0.44, 0.80]
KTAU:   0.45 [95%: 0.28, 0.60]
```



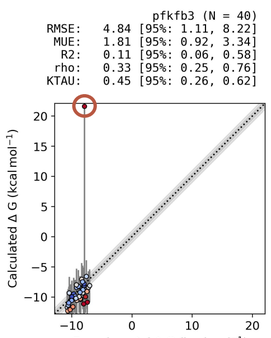
Removing unconverged edges

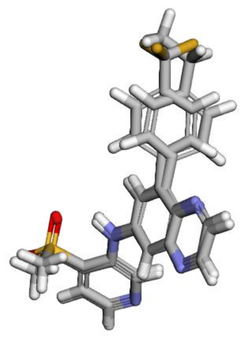
Removing redundant unconverged (MBAR errors > 0.15 kcal/mol) edges can sometimes improve accuracy.

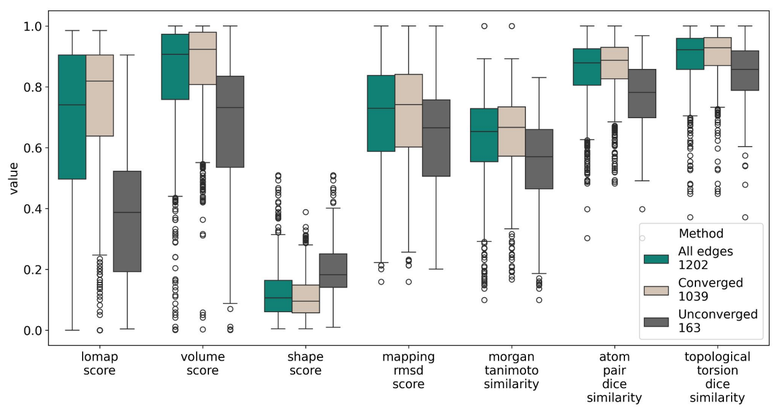
Resolving atom mapping issues

A large outlier in pfkfb3 led us to fix a bug in Kartograf that could lead to broken bond mappings.

```
                pfkfb3 (N = 40)
RMSE:   4.84 [95%: 1.11, 8.22]
MUE:    1.81 [95%: 0.92, 3.34]
R2:     0.11 [95%: 0.06, 0.58]
rho:    0.33 [95%: 0.25, 0.76]
KTAU:   0.45 [95%: 0.26, 0.62]
```



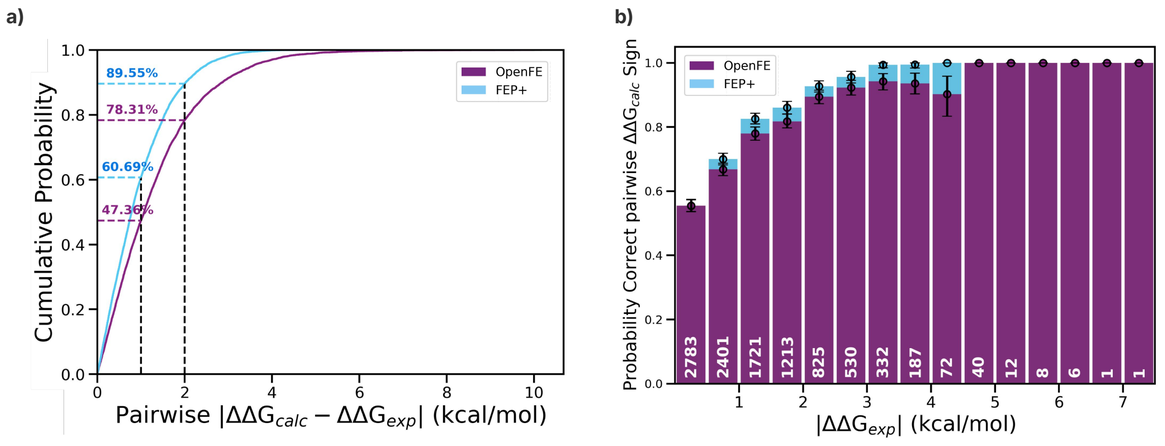




Mapping scores predict high difficulty edges

High MBAR error edges can be categorized by poor mapping scores, e.g. the LOMAP score.

Public dataset

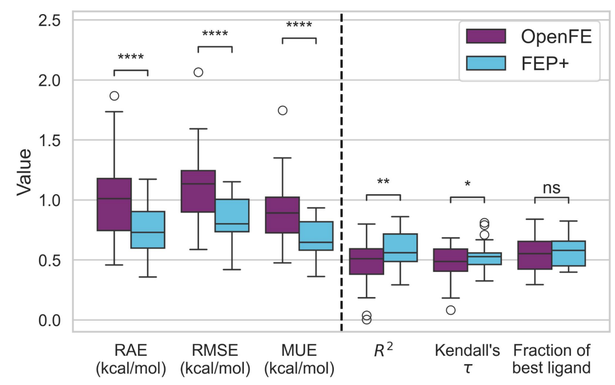


a) Absolute pairwise ΔΔG errors and b) sign prediction accuracy as a function of experimental ΔΔG magnitude, for both OpenFE and FEP+.

OpenFE shows a good agreement with experiment, with an RMSE across all systems of 1.7 kcal/mol. However, it falls slightly short of the current state-of-the-art tooling (FEP+, 1.2 kcal/mol, Ross et al. 2023). Looking at absolute pairwise ΔΔG errors, 78.3% (vs 89.6% FEP+) of predictions fall within 2 kcal/mol of experiment.

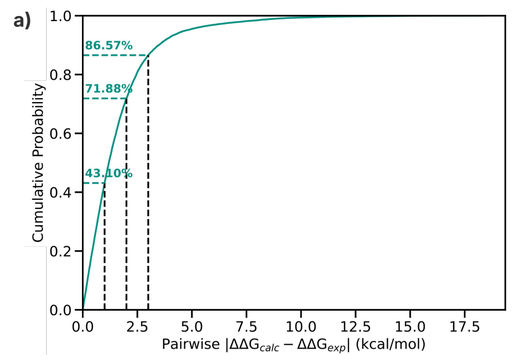
Larger errors partly stem from the use of a naive benchmark setup & sampling scheme, e.g. lack of water placement optimization and poor ligand alignments.

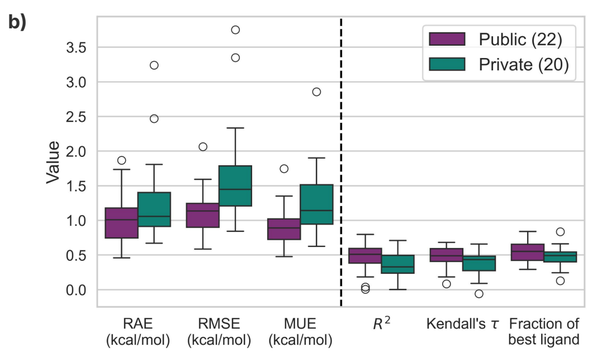
Despite larger errors, OpenFE demonstrates a good predictive power with an average fraction of best ligand score of 0.61.



Ranking metrics for OpenFE and FEP+ across all public set systems.

Private dataset

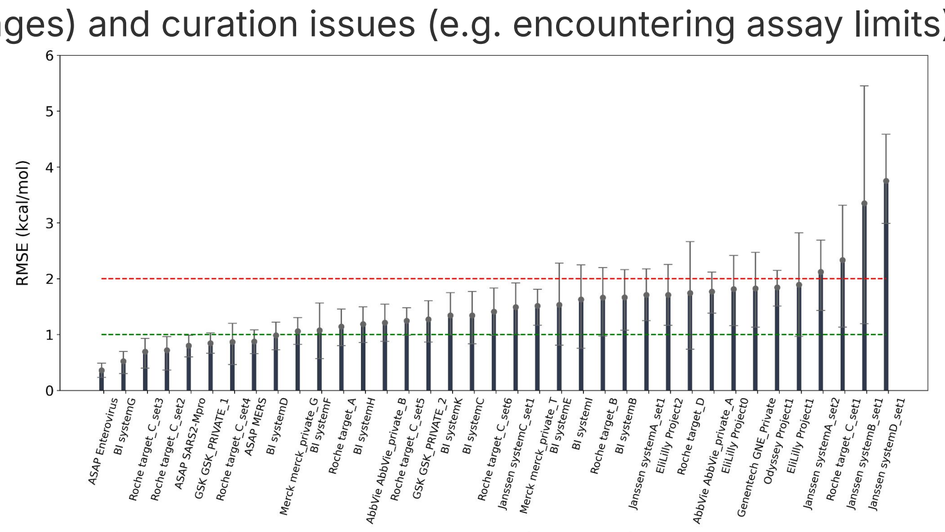




a) Absolute pairwise ΔΔG. Dashed lines indicate data filtered to include edges with minimum off-diagonal MBAR overlap > 0.03 (excludes outliers with large inter-repeat variability). b) Ranking statistics for public and private set results.

The private dataset, consisting of in-house benchmarks, offers insights into the likely "real world" effectiveness of OpenFE. While still providing reasonable estimates, we see reduced accuracy and predictive power with only 71% of predictions within 2 kcal/mol of experiment.

High variability in accuracy between systems is partly explained by more challenging transformations (e.g. large alchemical changes) and curation issues (e.g. encountering assay limits).



RMSE (kcal/mol) across all submitted private set systems.

Acknowledgements

Thank you to the scientists from the following organizations who contributed to the benchmark study:

- Abbvie
- ASAP AVIDD
- Astrazeneca
- Bayer
- Roche
- Biogen
- BMS
- Eli Lilly
- Boehringer Ingelheim
- GSK
- Genentech
- Deep Origin
- Merck KGaA
- Odyssey Therapeutics
- Johnson & Johnson