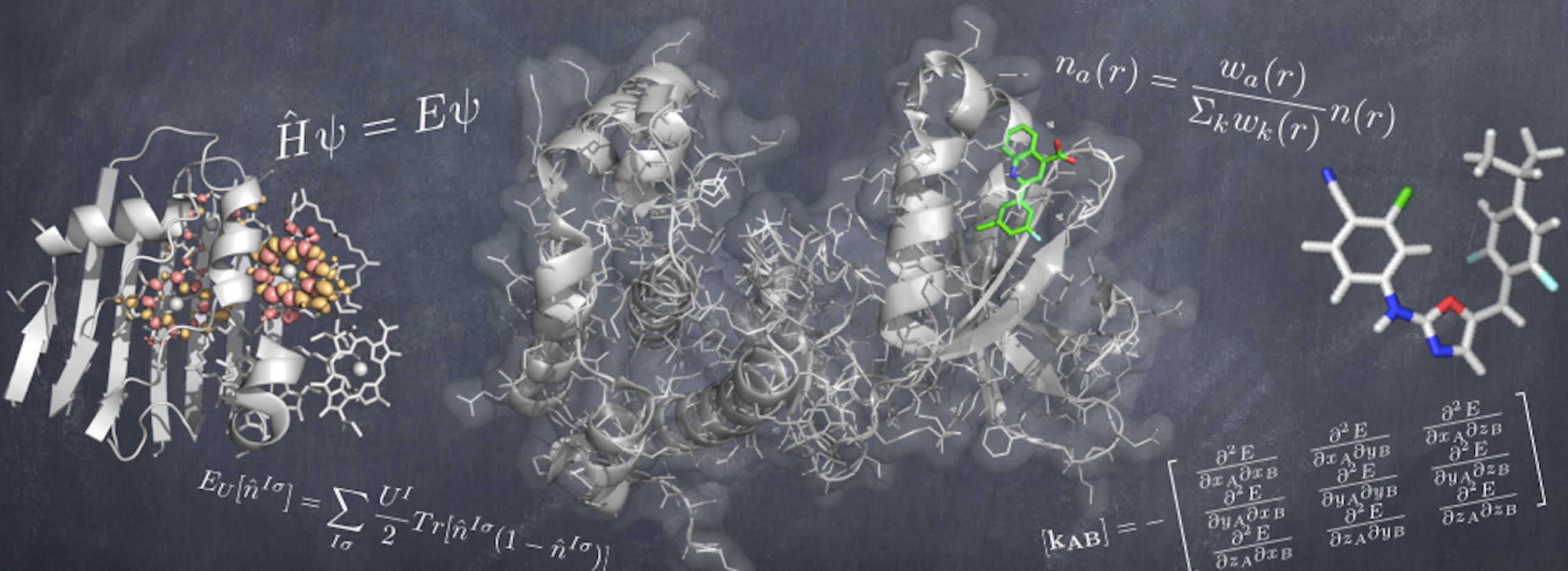


Bespoke Force Field Parameters for Computer-Aided Drug Design

Daniel Cole

Newcastle University

open forcefield

Molecular Interactions & Dynamics

Molecular interactions & dynamics would be ideally described by quantum mechanics, but too expensive for routine use.

$$\hat{H}\psi = E\psi$$

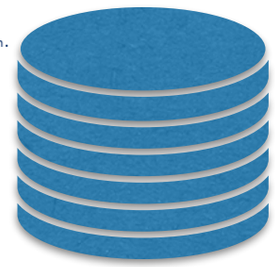
$$\Delta G = -k_B T ln \left\langle exp\left(-\frac{\Delta U}{k_B T}\right)\right\rangle$$

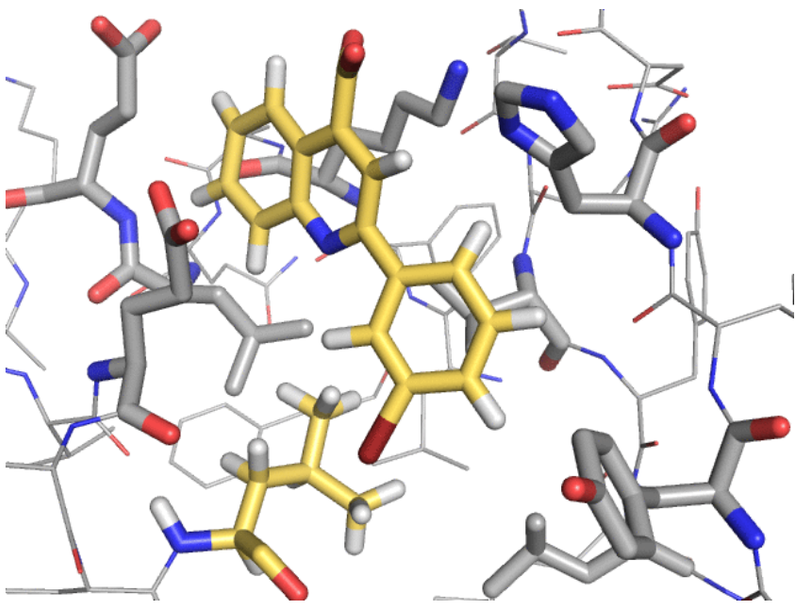
```python
def _calculate_lj_data(self, molecule: "Ligand") -> Dict[int, LJData]:
    """
    Use the AIM parameters to calculate a_i and b_i according to paper.
    Calculations from paper have been combined and simplified for faster computation.
    returns: Dict of the a_i, b_i and r_aim values needed for sigma/epsilon calculation.
    """

    lj_data = {}

    for atom_index, atom in enumerate(molecule.atoms):
        try:
            atomic_symbol, atom_vol = atom.atomic_symbol, atom.aim.volume

            # Find polar Hydrogens and allocate their new name: X
            if atomic_symbol == "H":
                bonded_index = atom.bonds[0]
                if molecule.atoms[bonded_index].atomic_symbol in [
                    "N",
                    "O",
                    "S",
                ]:
                    atomic_symbol = "X"
```





Our goals are to:

1) Develop better approximations to quantum mechanical modelling.

2) Produce software to automate this process.

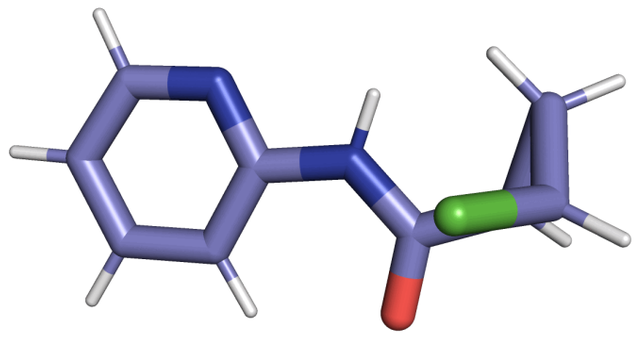
3) Collect and analyse data to work at scale.

4) Deliver more accurate predictions for drug design.

Open Force Field BespokeFit

Accurate determination of molecular conformation is crucial in structure-based drug design.

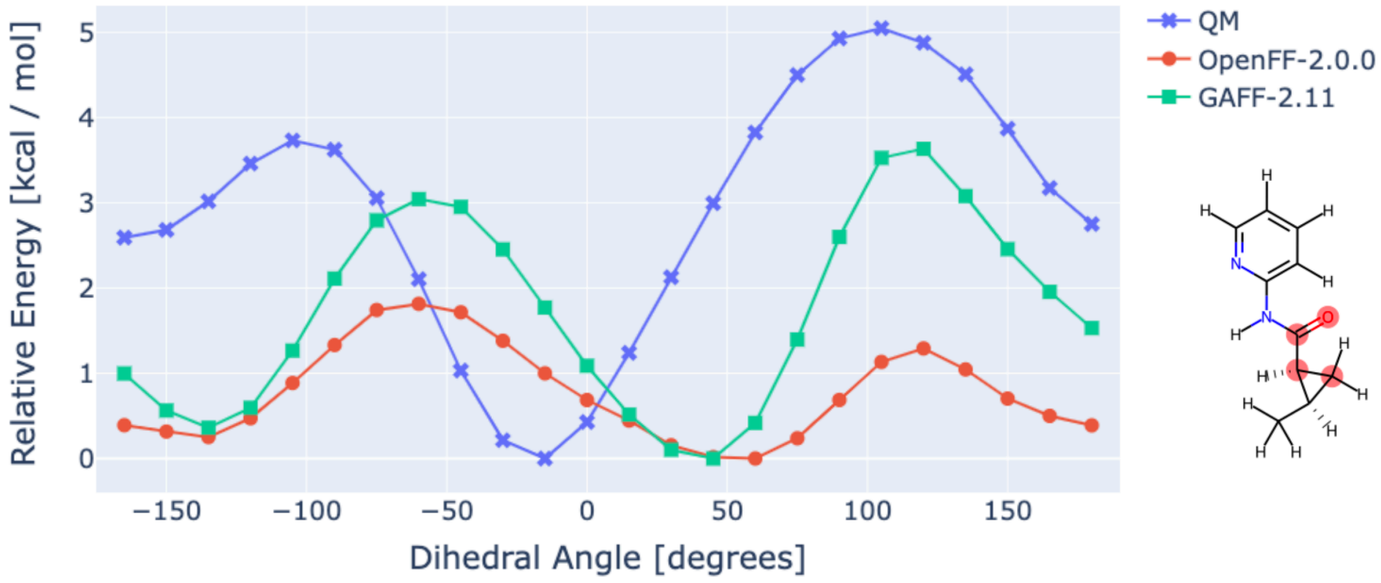
Molecular conformation is largely determined by torsional rotation about flexible bonds.



open forcefield

Transferability is difficult due to sensitivity to surrounding environment.

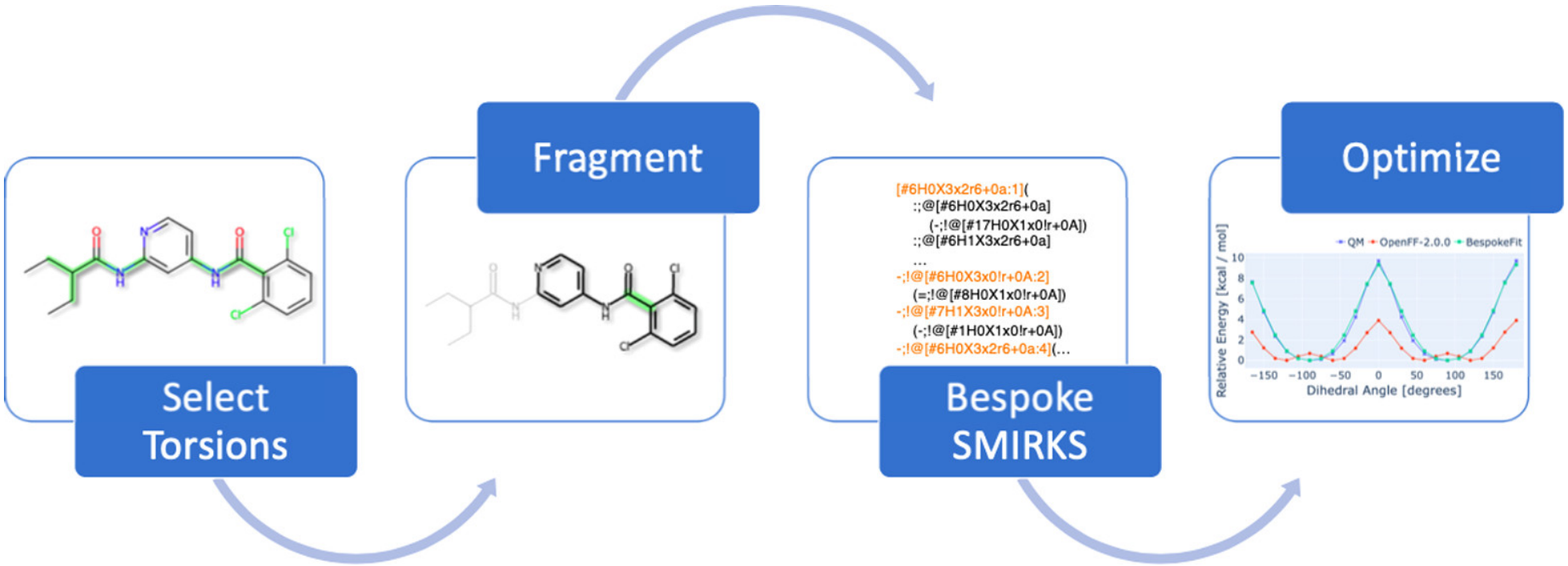
Force Field	Torsion Parameters
OpenFF-2.0.0	167
OPLS3	48,142
OPLS3e	146,669



https://github.com/openforcefield/openff-bespokefit

Horton et al. JCIM 2022

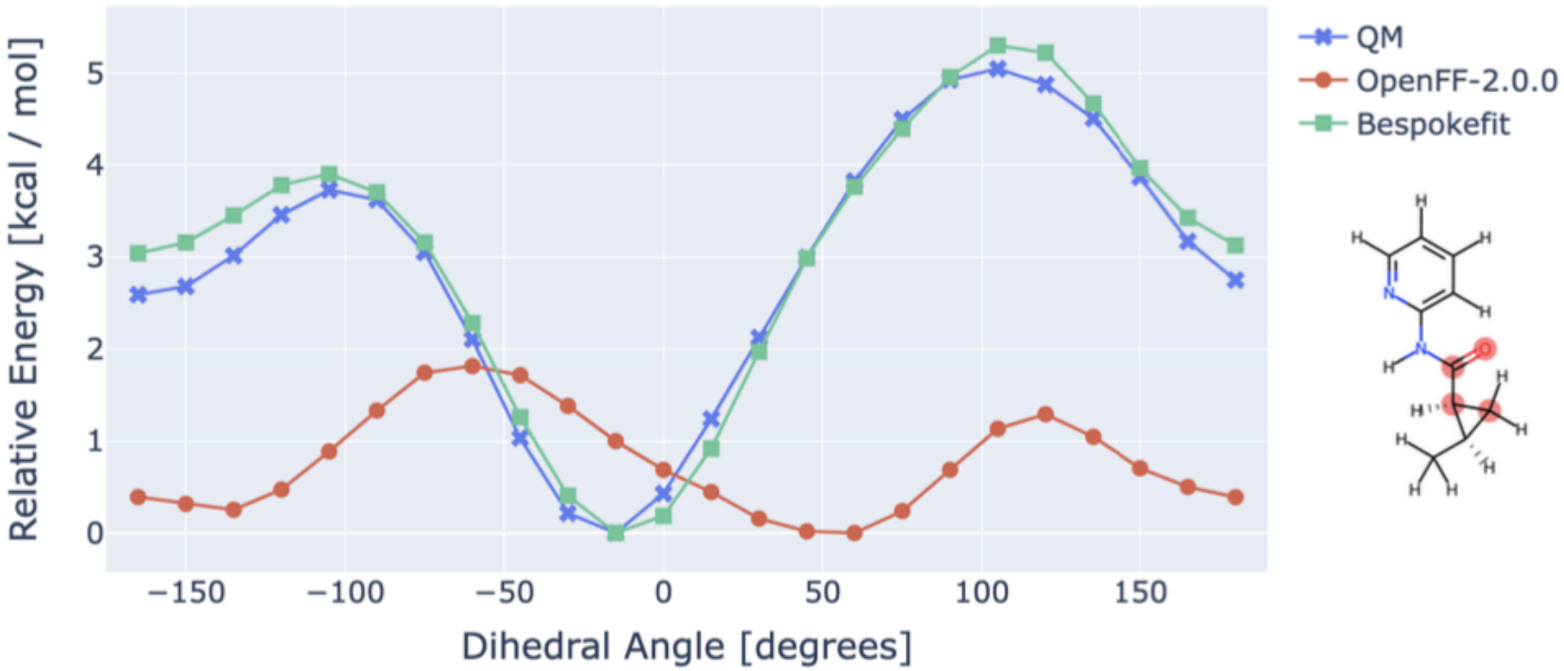
Open Force Field BespokeFit



Archive

OpenFF-BespokeFit provides robust molecule-specific parameterization workflow.

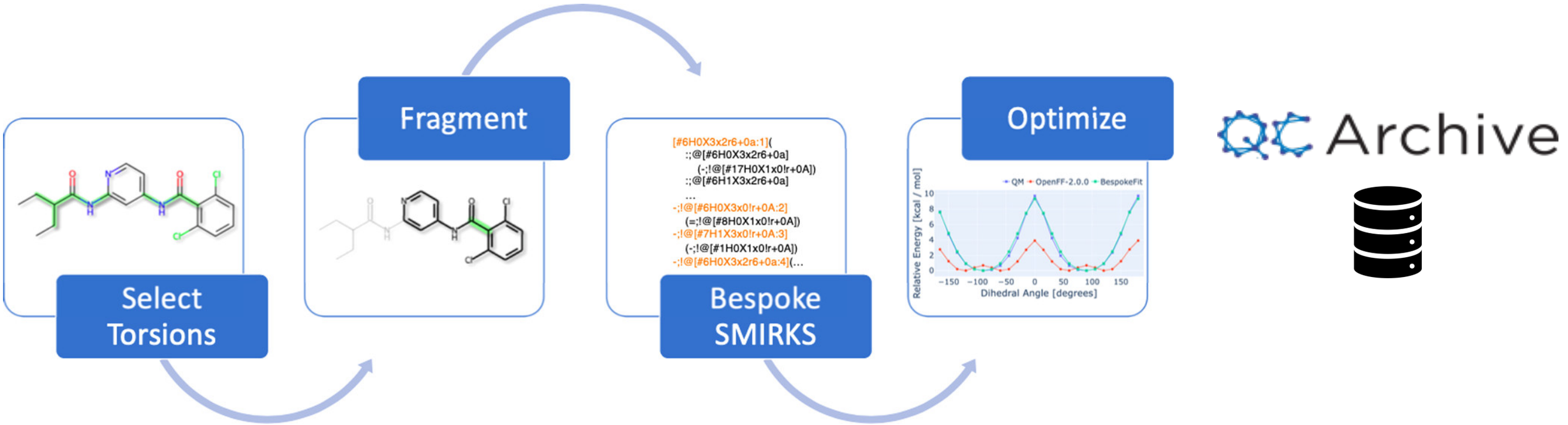
Can generate parameters at-scale from community generated data stored on QCArchive.



https://github.com/openforcefield/openff-bespokefit

Horton et al. JCIM 2022

Open Force Field BespokeFit



OpenFF-BespokeFit provides improved match to QM potential energy surface.

Plus a hierarchy of reference data options through interface to QCEngine.

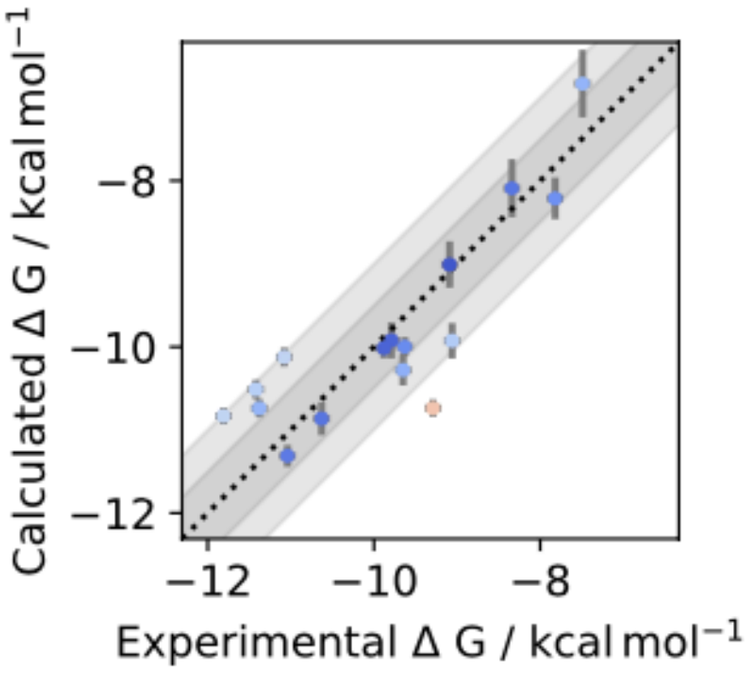
force field	Max RMSD (Å)	RMSE (kcal/mol)
OpenFF 1.3.0	$0.561^{0.698}_{0.435}$	$1.097^{1.328}_{0.89}$
BespokeFit (GFN2-xTB)	$0.375^{0.487}_{0.28}$	$0.792^{0.896}_{0.701}$
BespokeFit (ANI2x//GFN2-xTB)	$0.344^{0.442}_{0.259}$	$0.744^{0.875}_{0.635}$
BespokeFit (B3LYP-D3BJ/DZVP//GFN2-xTB)	$0.330^{0.388}_{0.273}$	$0.604^{0.697}_{0.530}$
BespokeFit (B3LYP-D3BJ/DZVP)	$0.311^{0.378}_{0.251}$	$0.289^{0.352}_{0.235}$

https://github.com/openforcefield/openff-bespokefit
Horton et al. JCIM 2022

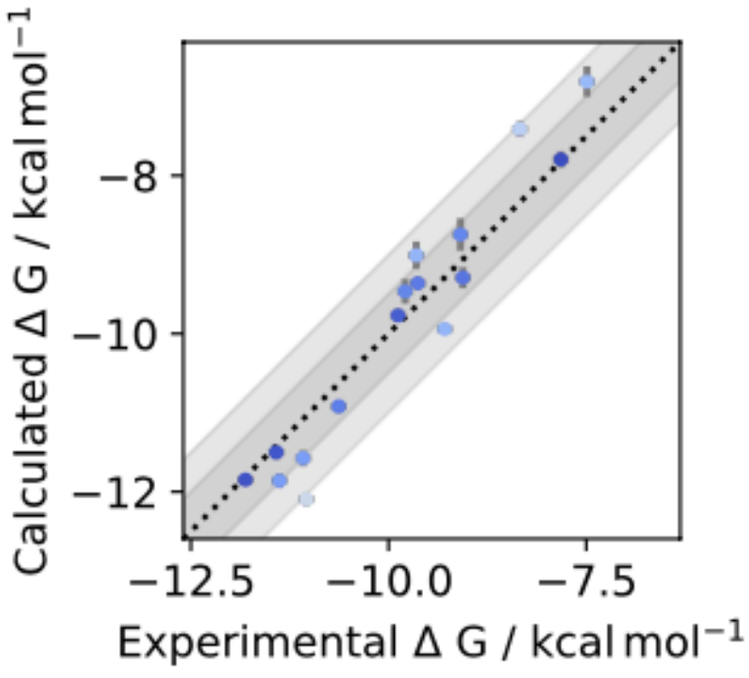
Bespoke Dihedral Parameters Improve Accuracy

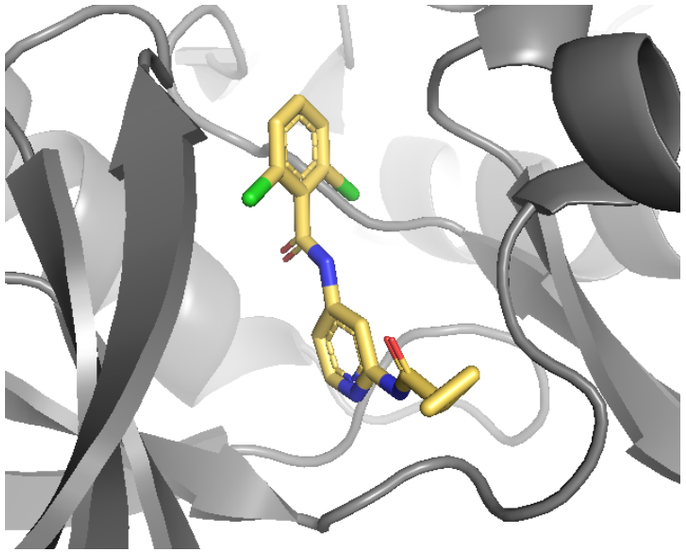
Bespoke torsion parameters improve relative binding free energies for series of TYK2 inhibitors, relative to OpenFF ‘Parsley’ force field.

OpenFF-1.3
tyk2 (N = 16)
RMSE: 0.68 [95%: 0.51, 0.91]
MUE: 0.56 [95%: 0.39, 0.77]
R2: 0.72 [95%: 0.35, 0.87]
rho: 0.85 [95%: 0.63, 0.93]



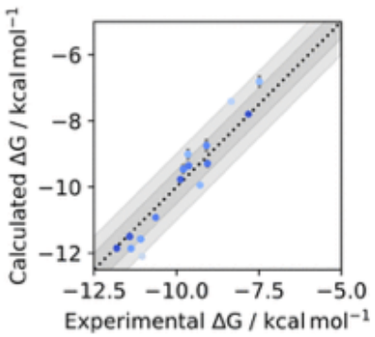
Bespoke default-1.3.0
tyk2 (N = 16)
RMSE: 0.51 [95%: 0.35, 0.69]
MUE: 0.42 [95%: 0.28, 0.59]
R2: 0.93 [95%: 0.84, 0.97]
rho: 0.97 [95%: 0.92, 0.99]



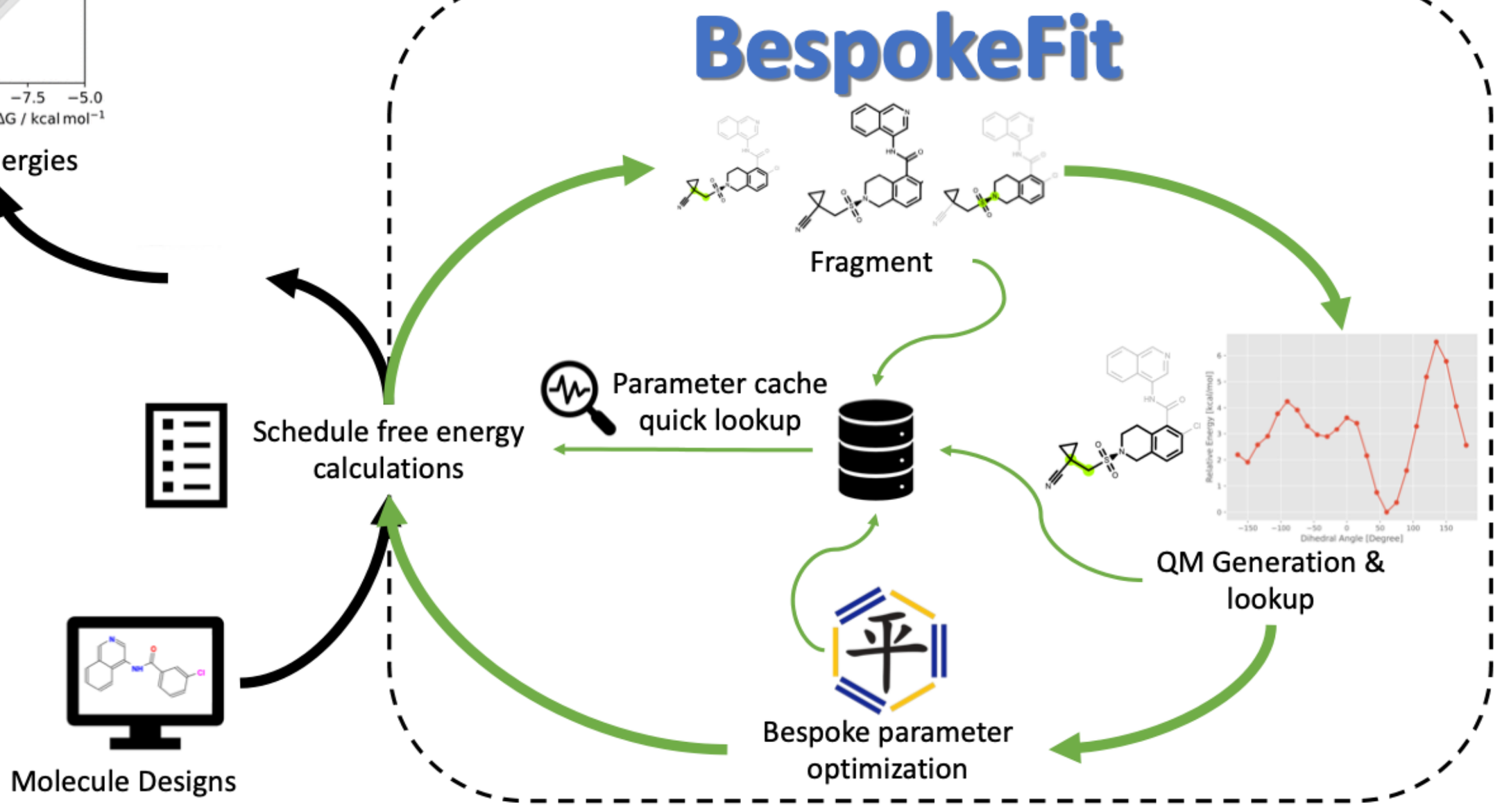


https://github.com/openforcefield/openff-bespokefit
Horton et al. JCIM 2022

Ongoing / Future Plans



Free energies



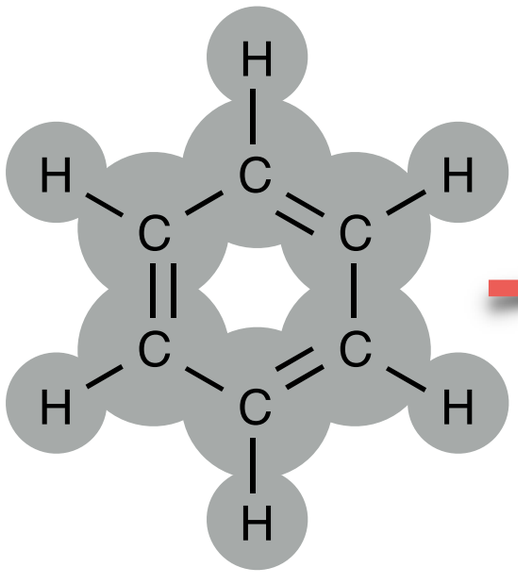
Also showing potential improvements in crystallography simulations (Tobias Hüfner)

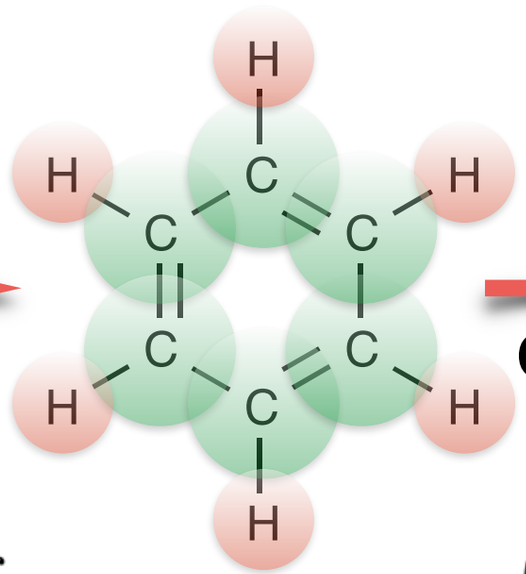
Can we do the same with non-bonded parameters?

DFT calculation computes total electron density

Electrons partitioned amongst the atoms in the system (MBIS, DDEC...)

Atomistic force field parameters computed directly from partitioned electron density





QM-to-MM mapping

- Atomic Charges
- Dispersion
- Exchange repulsion
- Off-site charges

$$q_i = z_i - \int n_i(\mathbf{r})\mathrm{d}^3\mathbf{r} \qquad V_i^{\mathrm{AIM}} = \int r^3 n_i(\mathbf{r})\mathrm{d}^3\mathbf{r}$$

Basis for approach is atoms-in-molecule electron density partitioning.

Electron density is computed in implicit solvent to account for induction effects.

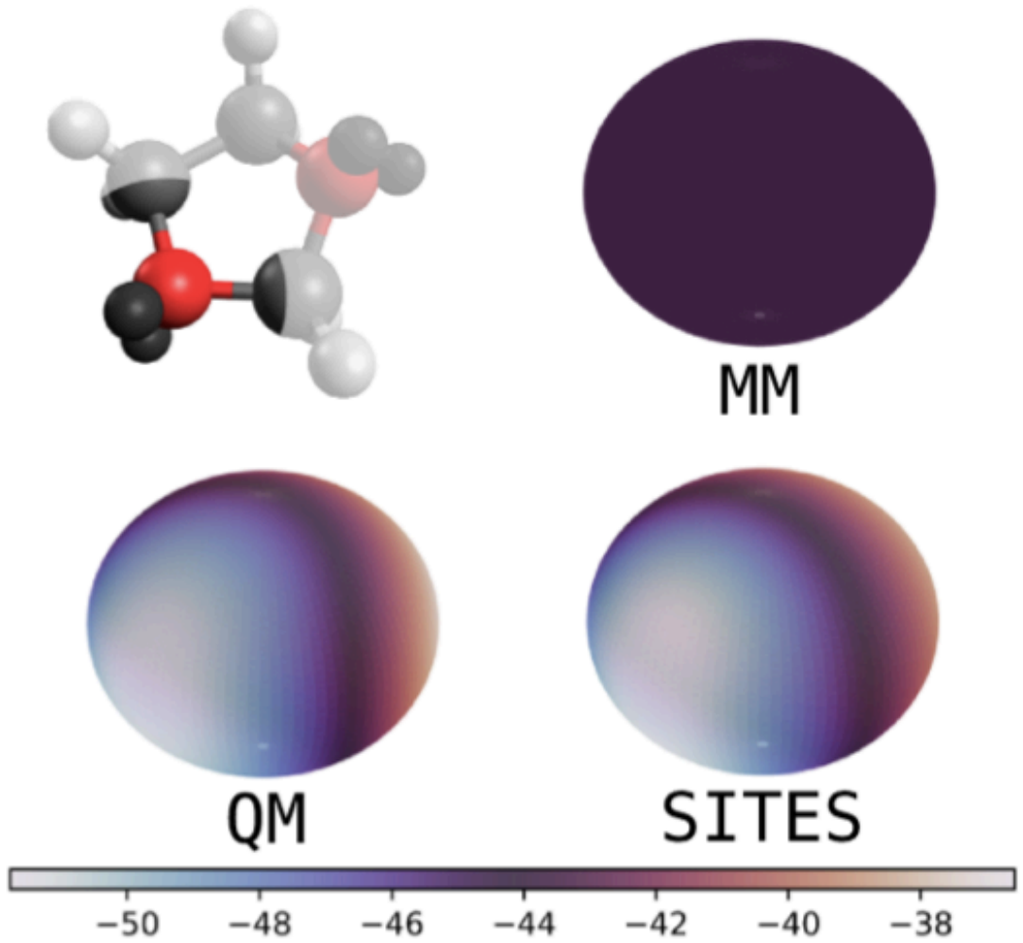
D. Cole, J. Vilseck, J. Tirado-Rives, M. Payne, W. Jorgensen, *JCTC,* **2016**, *12,* 2312

Off-Site Charges

We derive the positions and charges of off-centre virtual sites, by:

1) minimising the difference between the MM electrostatic potential (ESP) and the ESP of the partitioned electron density;

2) maintaining the symmetry of the atom's bonding environment.



electrostatic potential / kcal/mol

Ringrose, Horton, Wang, Cole. *PCCP* (2022), https://doi.org/10.1039/D2CP02864F

Lennard-Jones Parameters

No clear QM target for deriving atomic LJ parameters, which must account for short-ranged exchange-repulsion and long-range dispersion.

Best we've found is combination of Slater model and atoms-in-molecule analysis for small number of molecules.

Gives extremely accurate FF when combined with multipole & polarizable charge model, but is expensive to parameterize (fit to DFT-SAPT energies) & evaluate.

$$V_{\mathrm{FF}} = \sum_{ij} V_{ij}^{\mathrm{exch}} + V_{ij}^{\mathrm{elst}} + V_{ij}^{\mathrm{ind}} + V_{ij}^{\delta^{\mathrm{HF}}} + V_{ij}^{\mathrm{disp}}$$

$$V_{ij}^{\mathrm{exch}} = A_{ij}^{\mathrm{exch}} P(B_{ij}, r_{ij}) \exp(-B_{ij} r_{ij})$$

$$V_{ij}^{\mathrm{elst}} = -A_{ij}^{\mathrm{elst}} P(B_{ij}, r_{ij}) \exp(-B_{ij} r_{ij}) + \sum_{tu} Q_t^i T_{tu} Q_u^j$$

$$V_{ij}^{\mathrm{ind}} = -A_{ij}^{\mathrm{ind}} P(B_{ij}, r_{ij}) \exp(-B_{ij} r_{ij}) + V_{\mathrm{shell}}^{(2)}$$

$$V_{ij}^{\delta^{\mathrm{HF}}} = -A_{ij}^{\delta^{\mathrm{HF}}} P(B_{ij}, r_{ij}) \exp(-B_{ij} r_{ij}) + V_{\mathrm{shell}}^{(3-\infty)}$$

$$V_{ij}^{\mathrm{disp}} = -\sum_{n=3}^{6} f_{2n}(x) \frac{C_{\mathrm{ij},2n}}{r_{ij}^{2n}}$$

$$A_{ij} = A_i A_j$$

$$C_{ij,n} = \sqrt{C_{i,n} C_{j,n}}$$

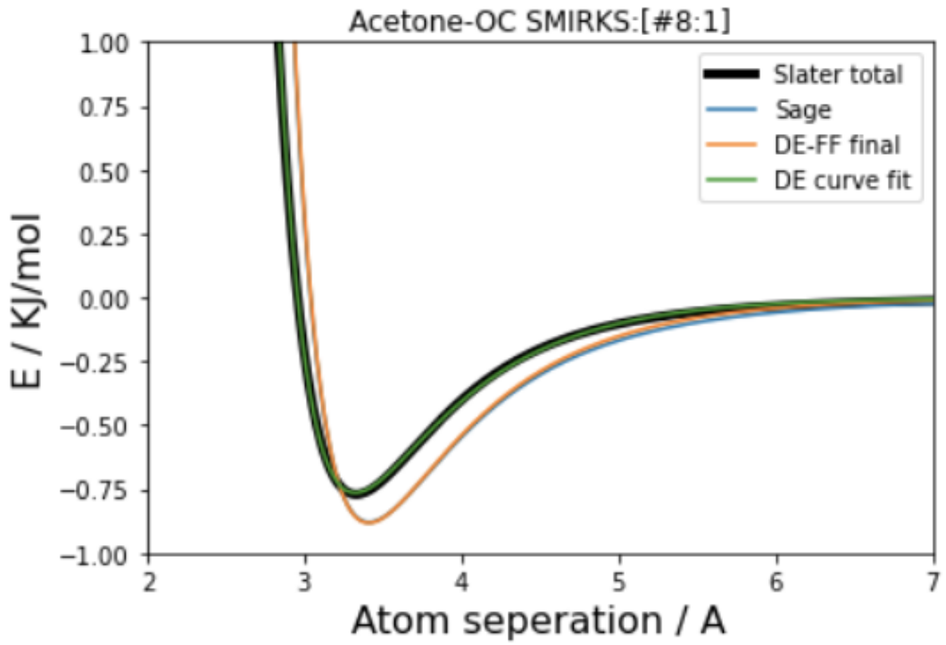
$$f_{2n}(x) = 1 - \mathrm{e}^{-x} \sum_{k=0}^{2n} \frac{(x)^k}{k!}$$

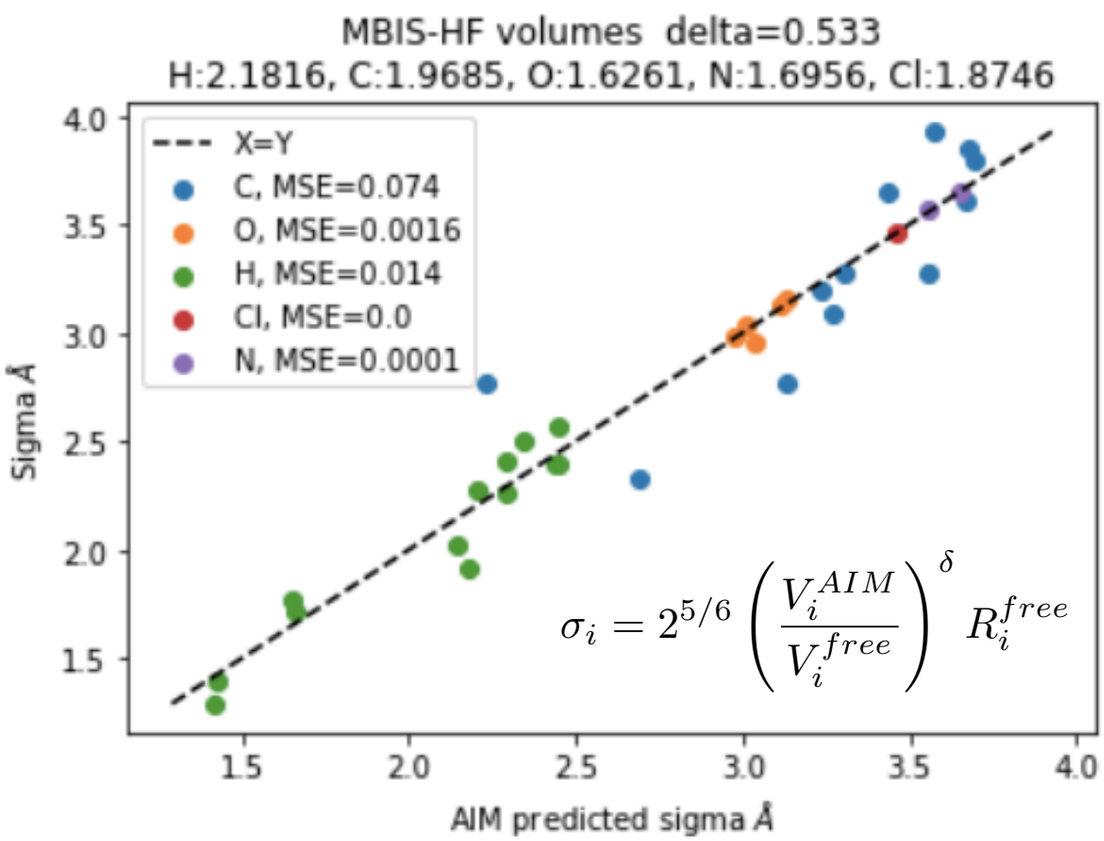
Van Vleet, Misquitta, Stone & Schmidt, JCTC, 2016, 12, 3851

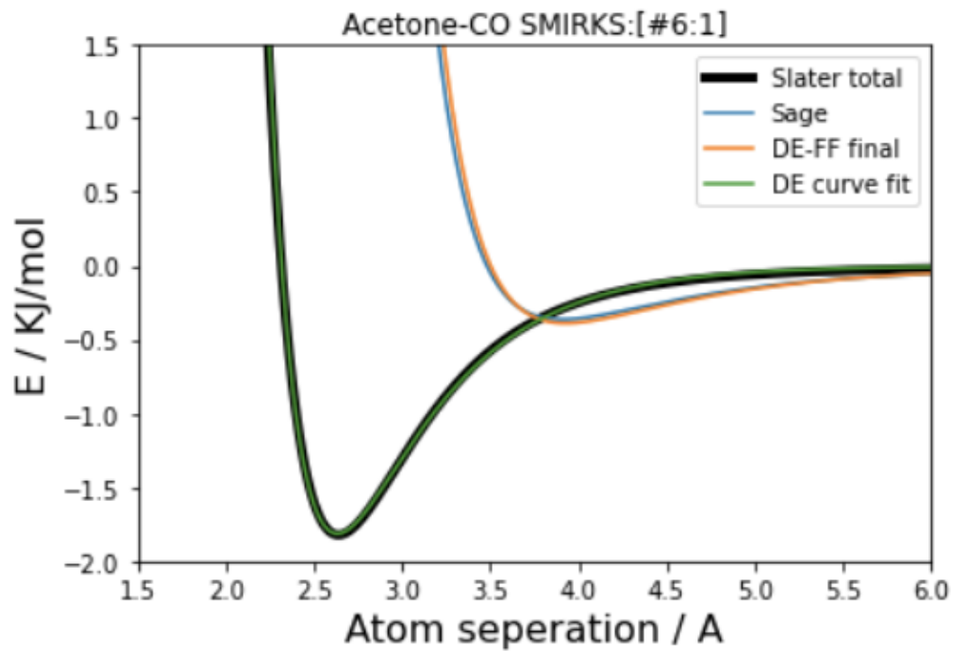
Lennard-Jones Parameters

Some similarities & differences with transferable LJ FF parameters.

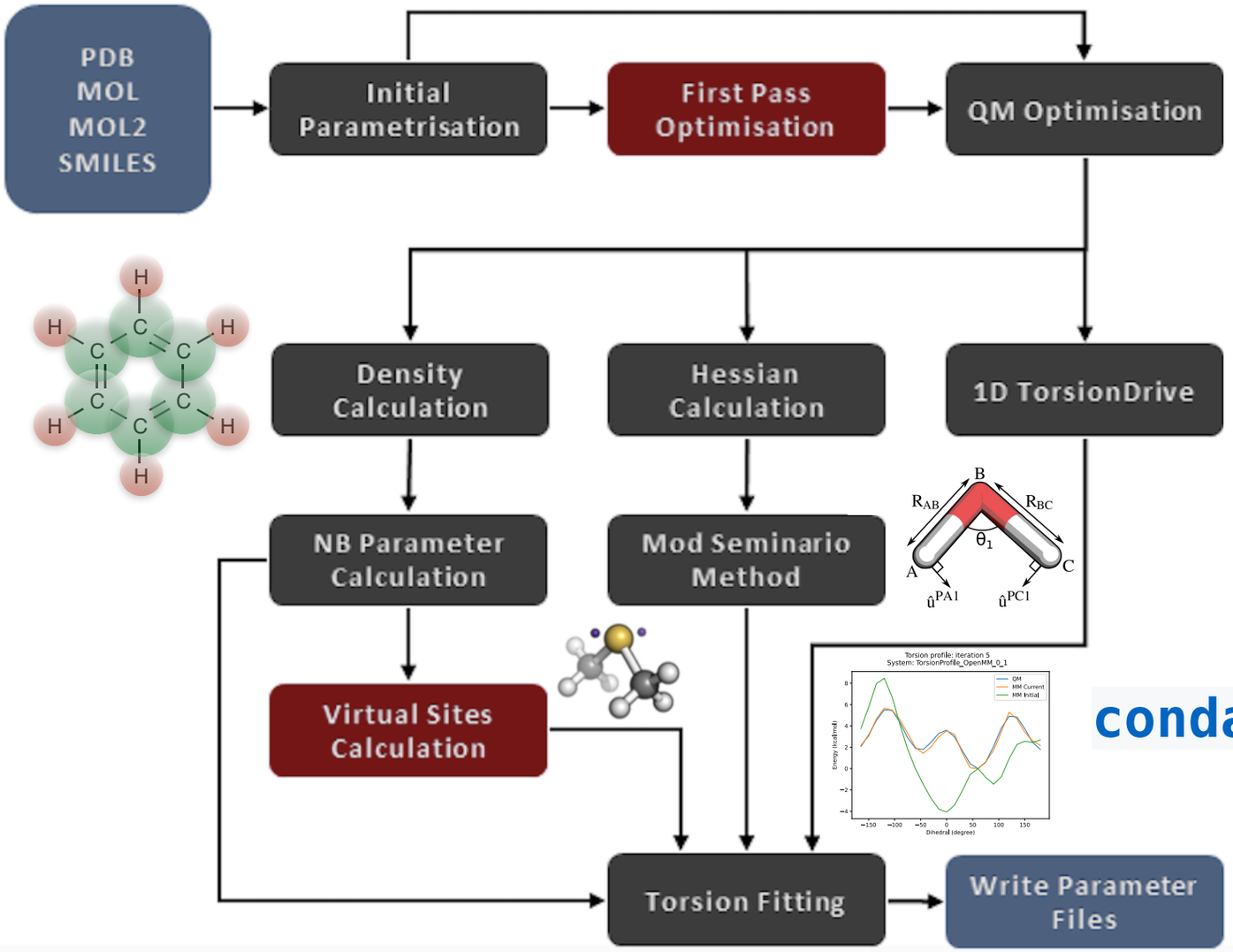
Importantly, can show mapping between atomic volumes & ‘ideal’ potential energy surfaces







QUBEKit



TorsionDrive

PSI4
AB INITIO QUANTUM CHEMISTRY

geomeTRIC

RDKit
Open-Source Cheminformatics and Machine Learning

QUBEKit

open forcefield consortium

Chargemol

QC Archive

```
conda install -c conda-forge qubekit
```

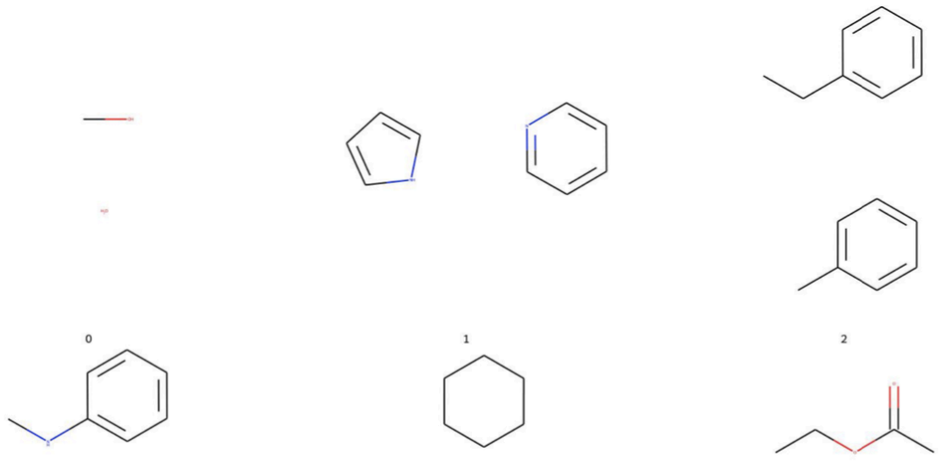
https://github.com/qubekit/QUBEKit

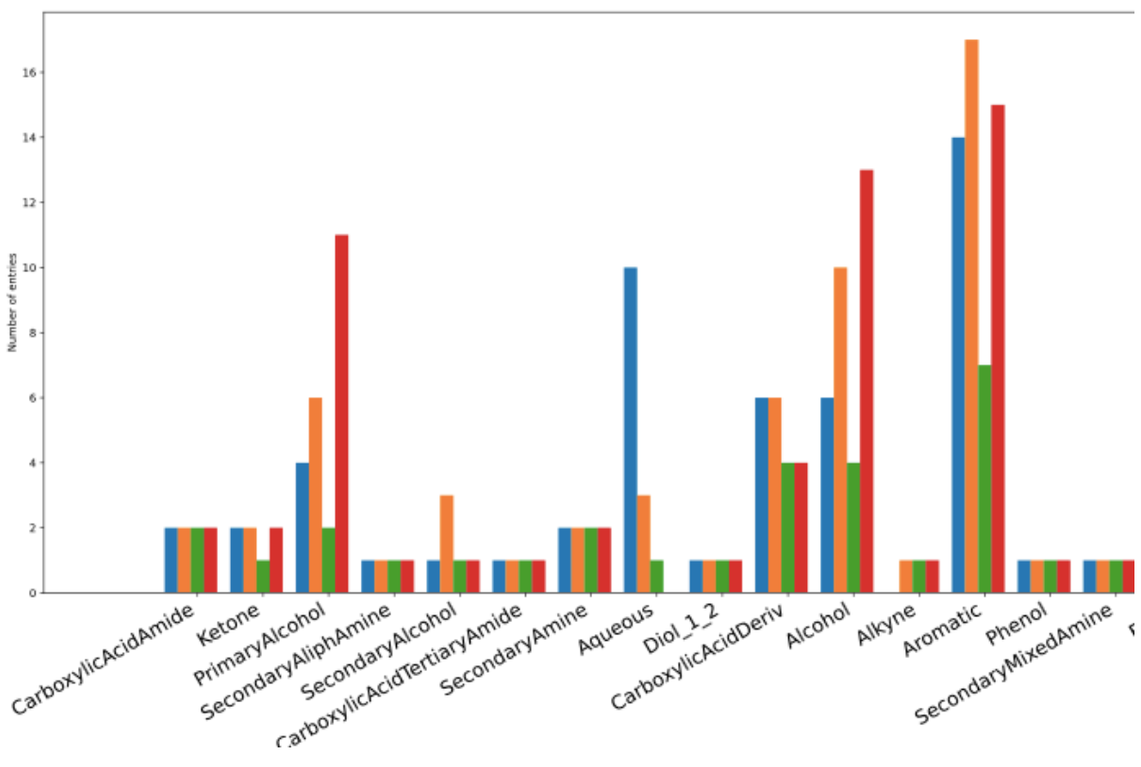
Developers: Josh Horton & Chris Ringrose

Interface with ForceBalance / OpenFF-Evaluator allows us to rapidly fit QM-to-MM mapping parameters for each force field model.

Horton JT, Allen AEA, Dodda L, Cole DJ. *J Chem Inf Model* 2019, **59**(4), 1366-1381.

Testing force field hypotheses





Would like to expand training to more properties and more molecules. Interface with OpenFF-Evaluator.

Train: 26 binary densities, 26 enthalpy of mixing
Test: 27 aqueous, non-aqueous and transfer free energies

Filter for H, C, N, O only
Ensure overlap for functional groups in train/test
Add smallest molecules until all functional groups covered

Testing force field hypotheses

Sage: Re-train set of transferable sigma, epsilon

QUBE: Train QM-to-MM mapping parameters with same charge model (AM1-BCC, missing virtual sites and latest mapping ideas)

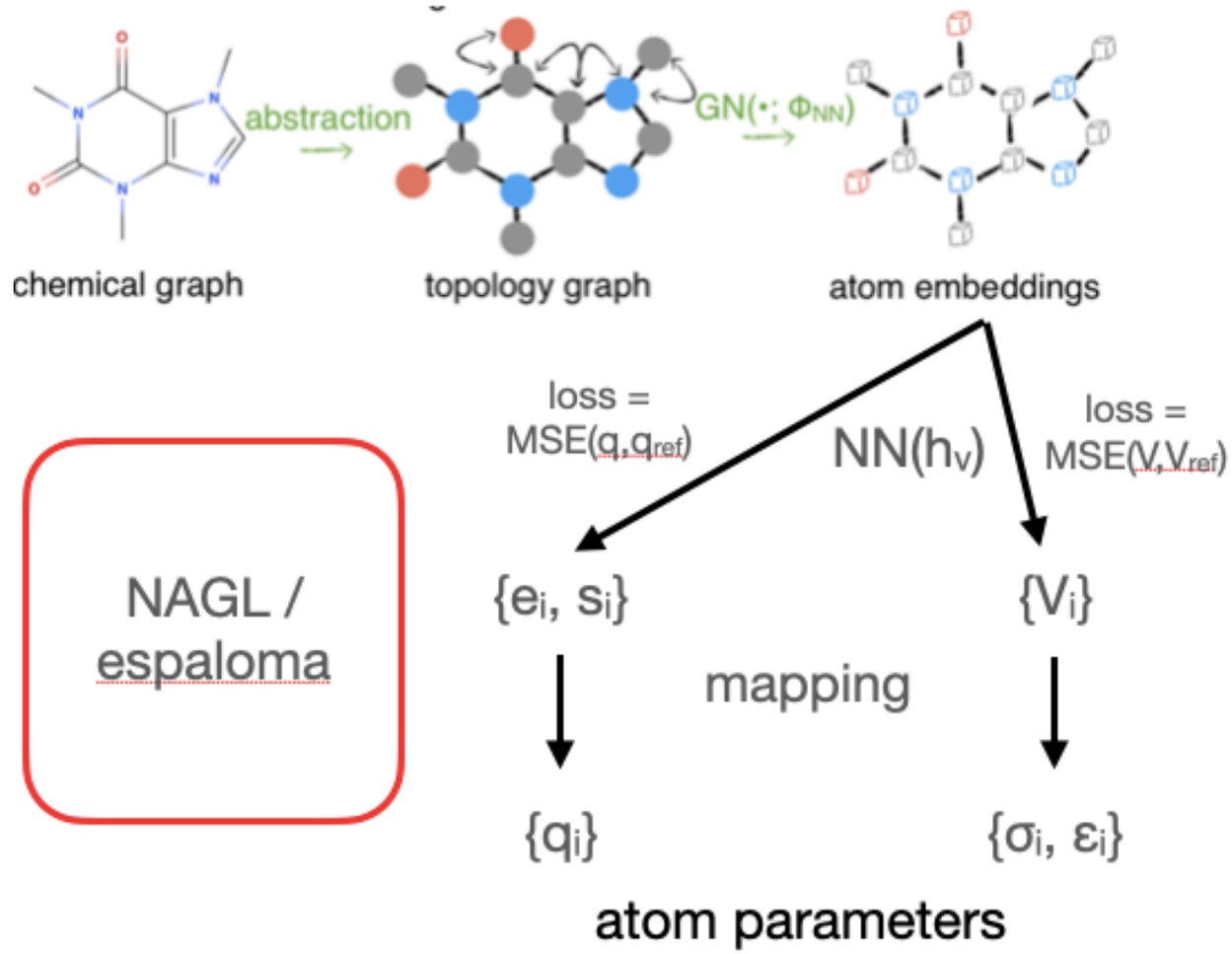
	Sage-TIP3P	**Sage-TIP4P-FB**	**QUBE**
Binary Density / g/cm^3	0.031	0.031	0.036
Enthalpy of Mixing / kJ/mol	0.63	0.66	0.89
Transfer Free Energy / kcal/mol	0.97	1.20	1.06

Graph neural networks can be trained on QM datasets

Underlying QM calculations are still too slow for high-throughput force field derivation. Graph convolutional neural networks can provide continuous atom embeddings to describe non-bonded parameters.

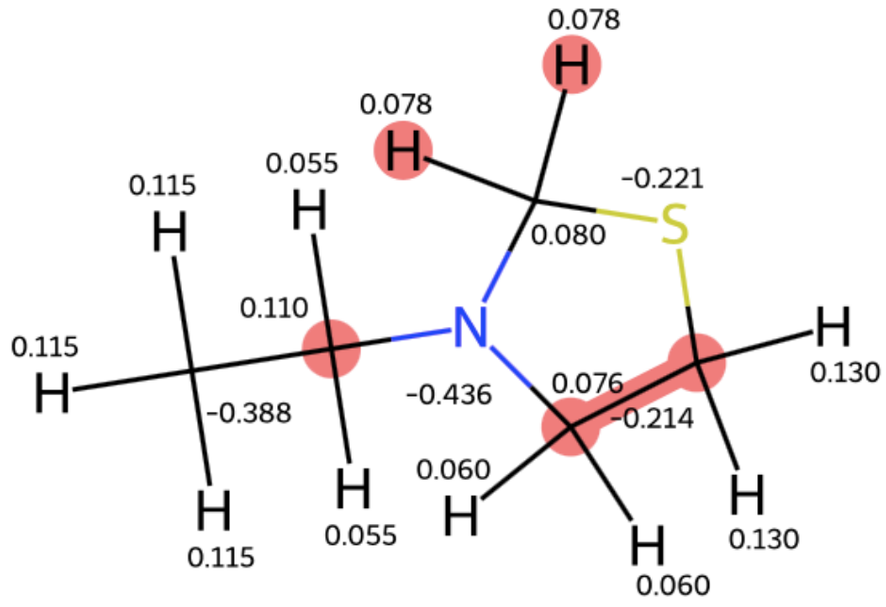
QM reference computed using HF/6-31G* with PSI4 for 50K molecules from QCArchive.

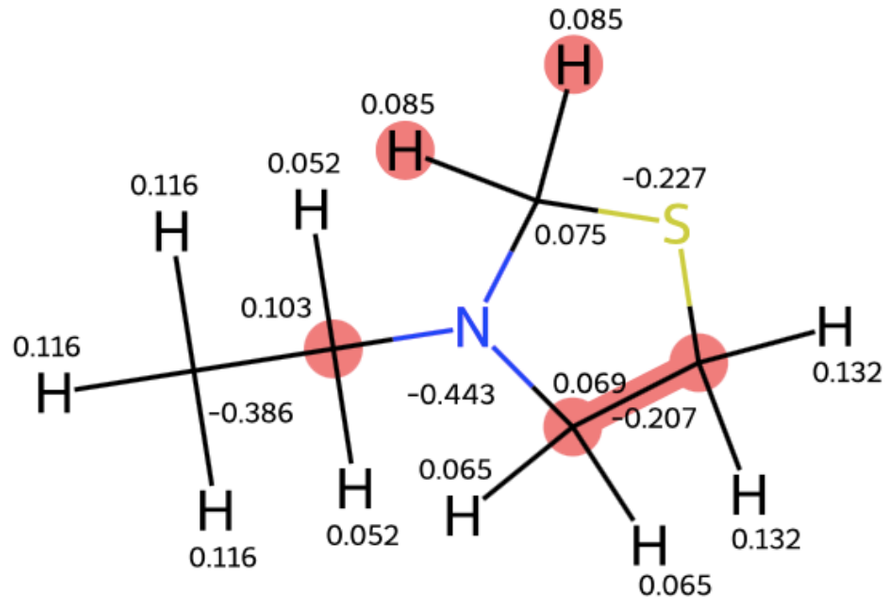
QM charges and volumes computed using MBIS atoms-in-molecule approach.



Wang et al., End-to-End Differentiable Construction of Molecular Mechanics Force Fields, Chem Sci, 2022, 13, 12016

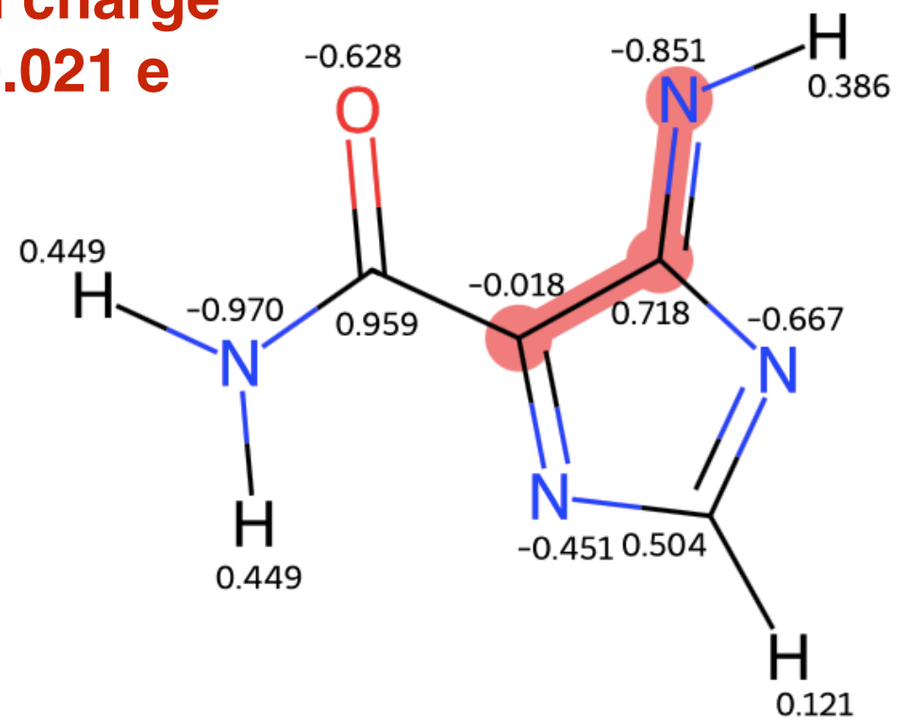
Graph neural networks can be trained on QM datasets



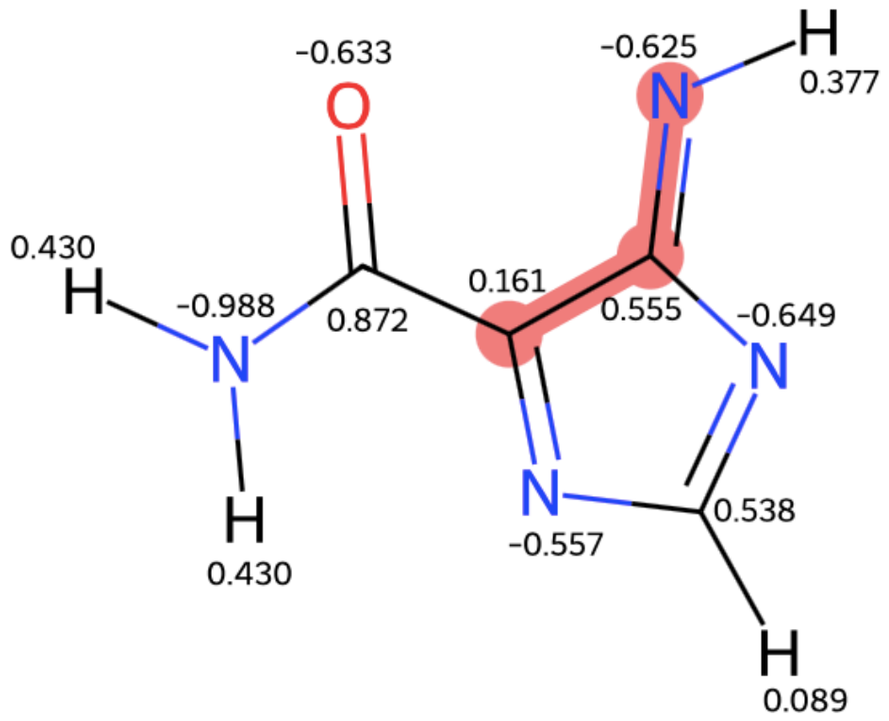


Good

Overall AIM charge
RMSE = 0.021 e



prediction

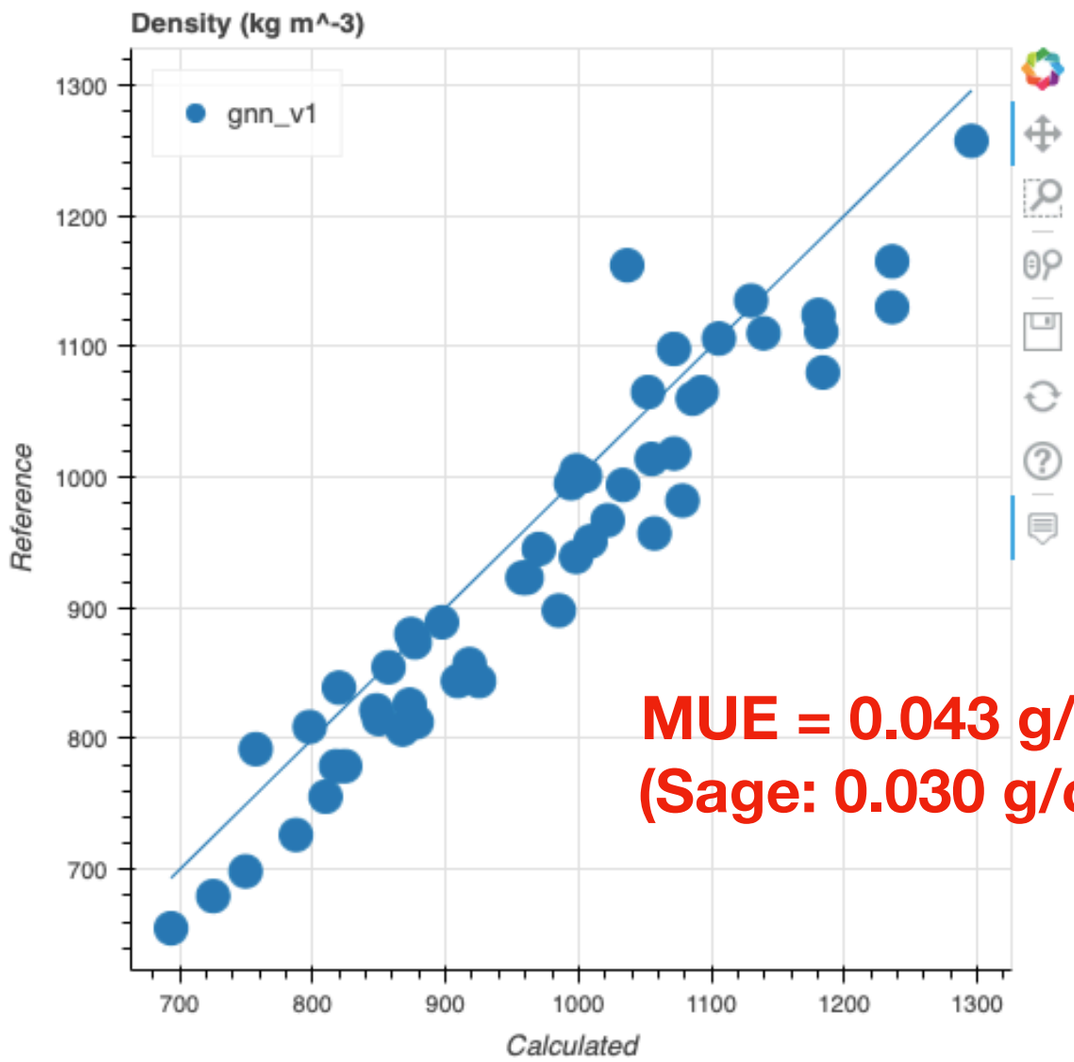


reference

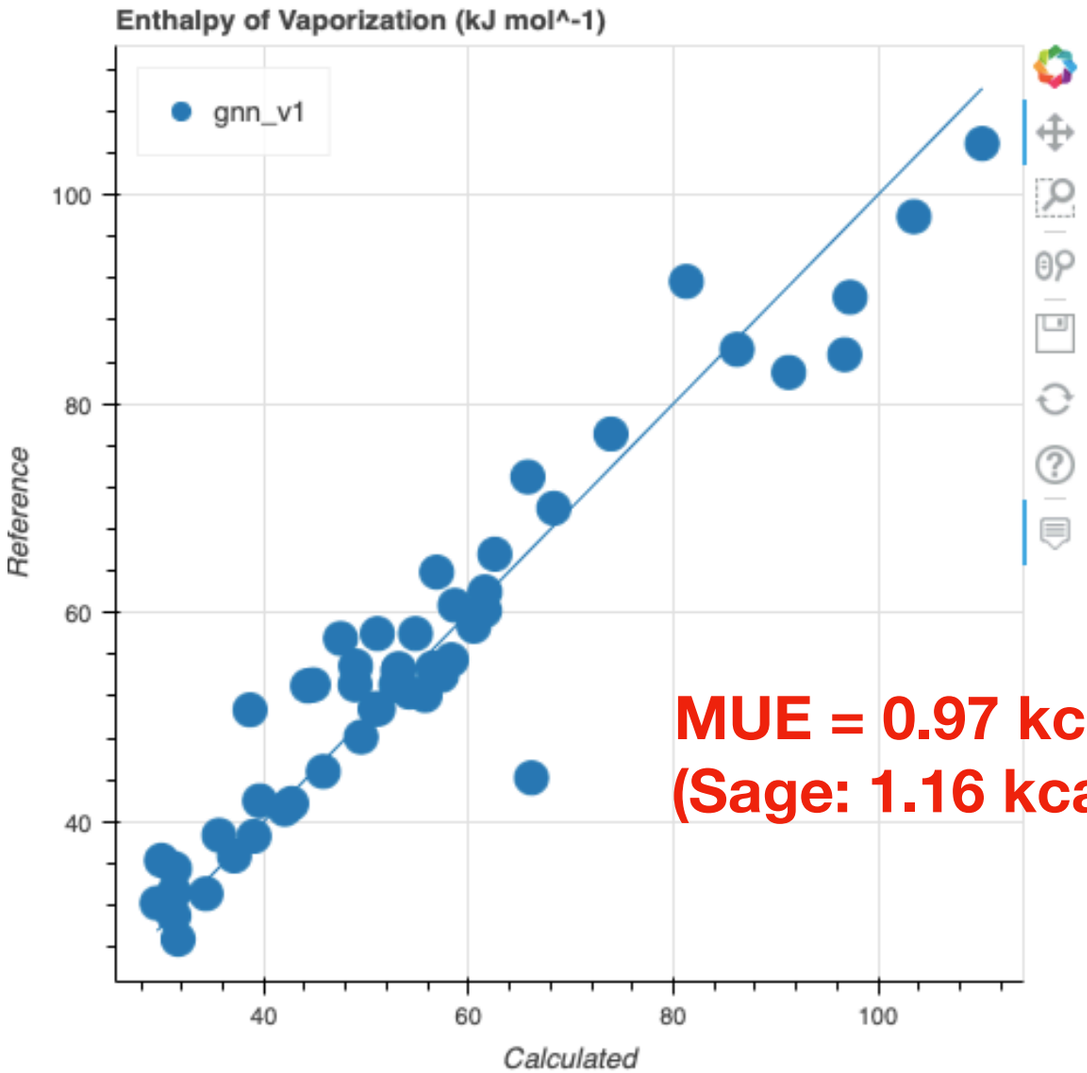
Bad

Liquid Property Calculations with GCN Force Fields

Using our QM-to-MM mapping force fields we can derive a consistent non-bonded force field model that is completely described by the GCN atom embeddings:



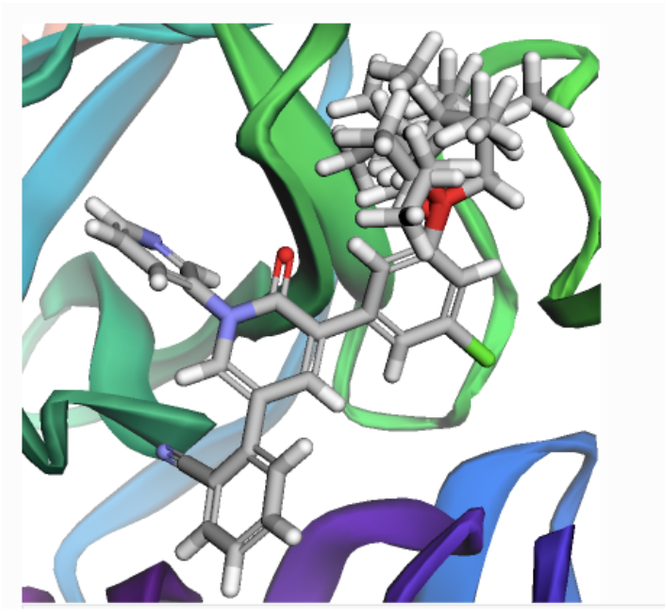
MUE = 0.043 g/cm³
(Sage: 0.030 g/cm³)



MUE = 0.97 kcal/mol
(Sage: 1.16 kcal/mol)

Mixture property and free energy benchmarks are also ongoing.

Other Projects for Discussion

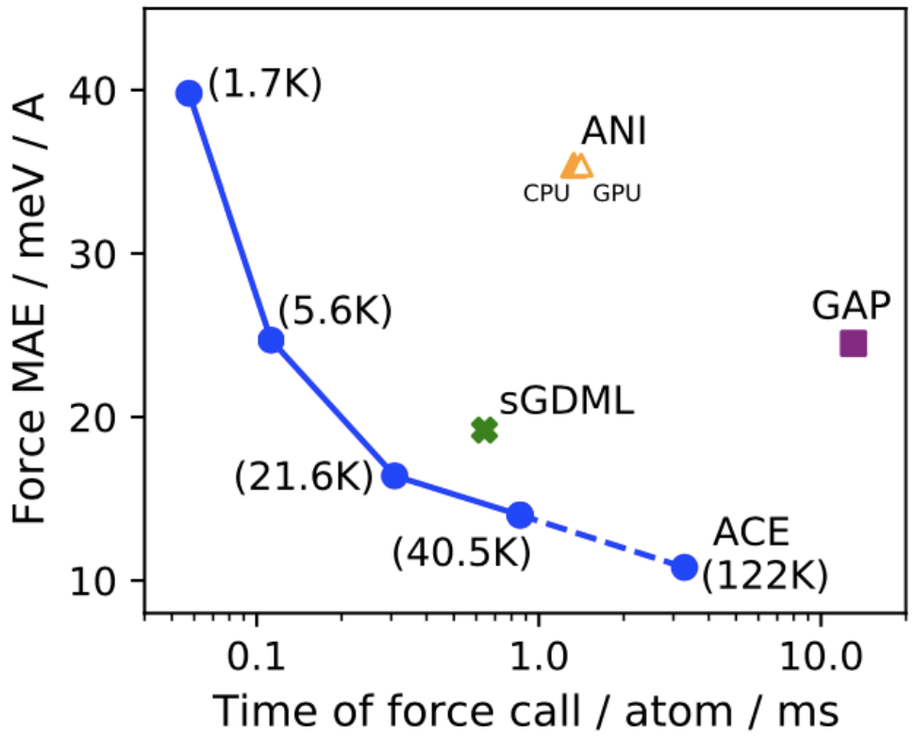


FEgrow for building / scoring congeneric series of ligands

Mat Bieniek, Ben Cree, Rachael Pirie,
Josh Horton & Natalie Tatum

https://github.com/cole-group/FEgrow

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



Fast & accurate linear atomic cluster expansion force fields

Dávid Kovács & Gábor Csányi
J. Chem. Theory Comput. 2021, 17, 7696–7711

Future Work / Wishlist

New reference datasets for non-bonded (high level QM, implicit solvent)

AIM charges / LJ parameters / off-site charges(?) from GNNs

Benchmarking non-standard functional forms — ie route for testing non-LJ functional forms, v-sites etc on protein-ligand binding free energies

Thanks to:

Newcastle: Josh Horton, Mat Bieniek, Antiope Politi, Chris Ringrose, Rachael Pirie, Ben Cree, Asma Khoualdi

Everyone at Open Force Field, particularly: Simon Boothroyd (now Roivant), Jeff Wagner, Josh Mitchell, Pavan Behara, John Chodera, David Mobley

Cresset: Venkata Ramaswamy, Mark Mackey

UK Research and Innovation

Newcastle University

MoSMed Molecular Sciences for Medicine

The Alan Turing Institute

N8 Bede COMPUTATIONALLY INTENSIVE RESEARCH

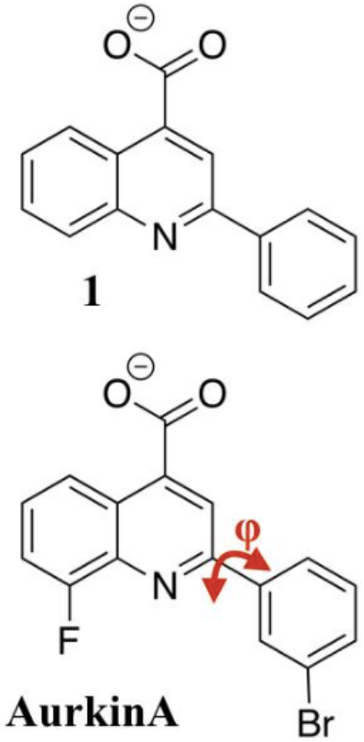
open forcefield

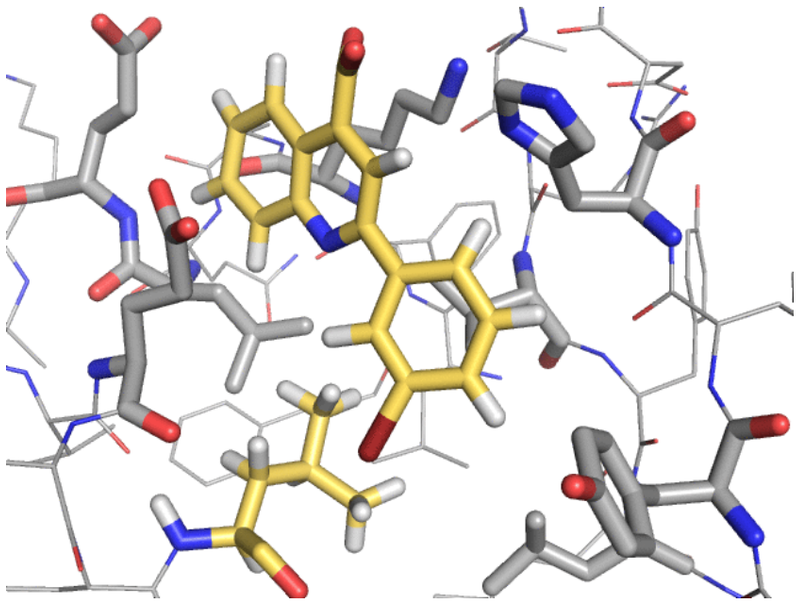
https://blogs.ncl.ac.uk/danielcole/
https://github.com/cole-group/
@ColeGroupNCL

Computer-Aided Drug Design

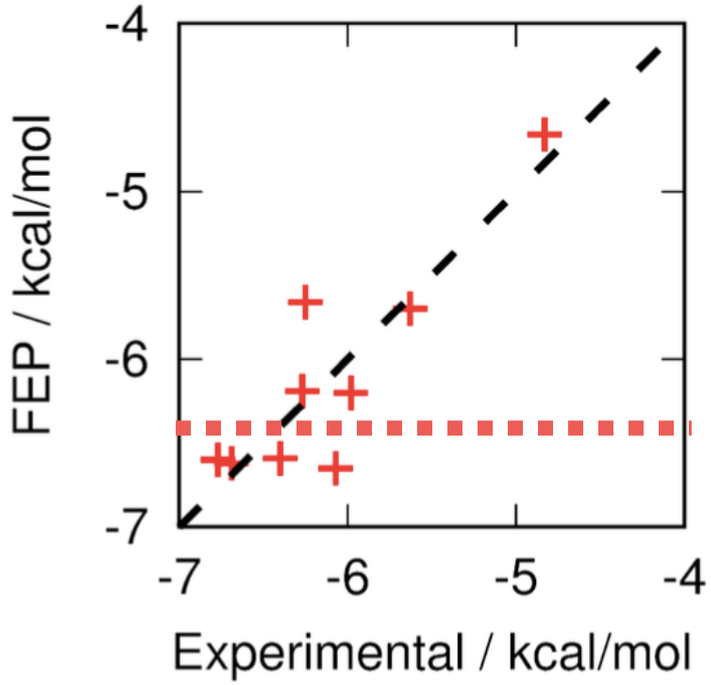
Free energy calculations are increasingly used in hit-to-lead stage of drug discovery.

Achievable accuracy is largely determined by sampling and the **force fields**.





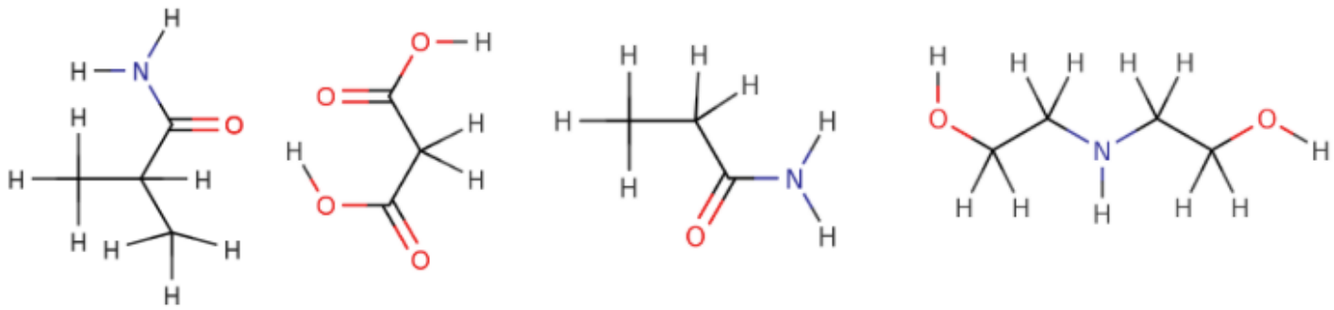
aurora kinase



RMSE = 0.32 kcal/mol

D. J. Cole et al., *Chem. Commun.* **2017**, *53*, 9372

Ongoing / Future Plans



Property	Pristine Sage	Bespoke Torsion Sage
[1)]Non-bonded distance error [Å]	0.35(0.11)	0.40(0.16)
[2)]Cell Error [Å]	4.81(1.06)	2.56(0.67)
Atomic Position RMSD [Å]	3.12(1.84)	1.04(0.38)
Torsion Angle RMSD [deg]	21.29(3.13)	38.09(14.72)

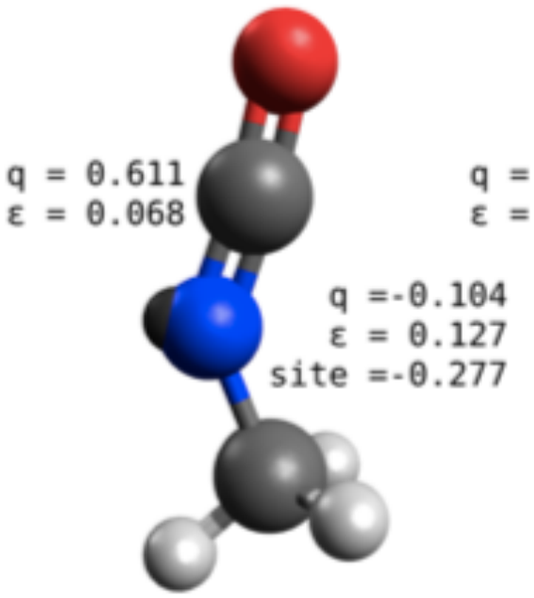
Data thanks to Tobias Hüfner

Lennard-Jones Parameters

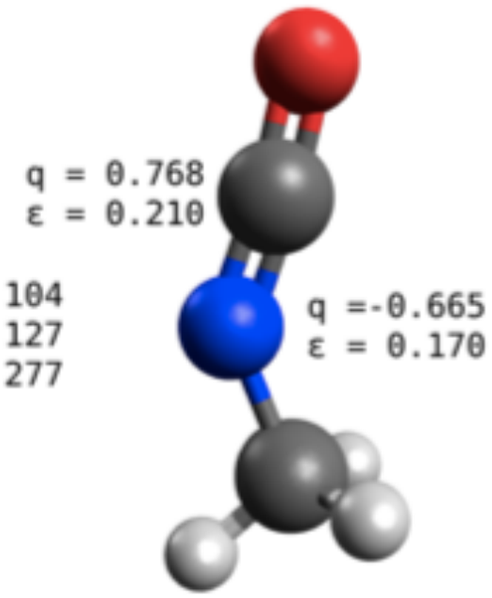
$$\sigma_i = 2^{5/6}\left(\frac{V_i^{AIM}}{V_i^{free}}\right)^{1/3} R_i^{free} \qquad \varepsilon_i = \frac{\alpha\left(\frac{V_i^{AIM}}{V_i^{free}}\right)^{\beta} B_i^{free}}{2\left(2R_i^{free}\right)^6}$$

QM-to-MM mapping

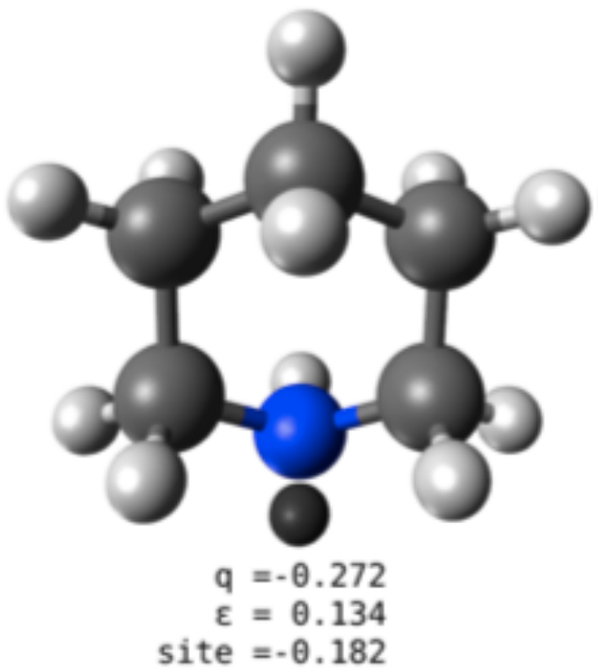
Can derive bespoke charge, Lennard-Jones (σ, ε) and off-site parameters:



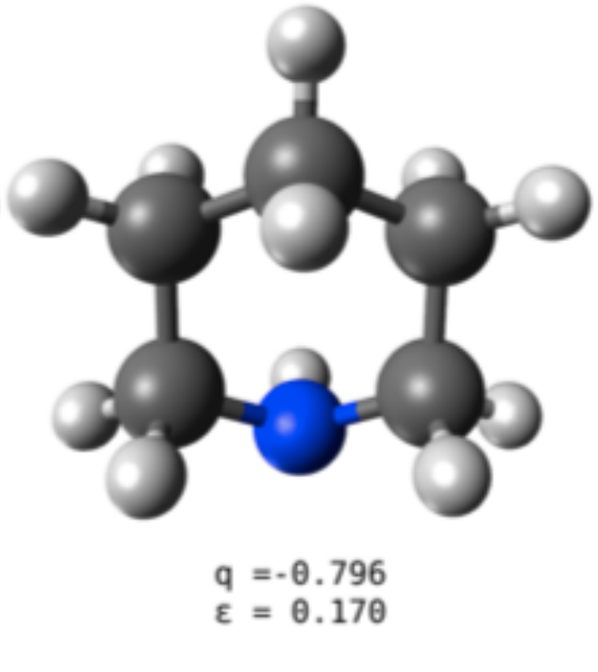
QUBE



OpenFF



QUBE



OpenFF

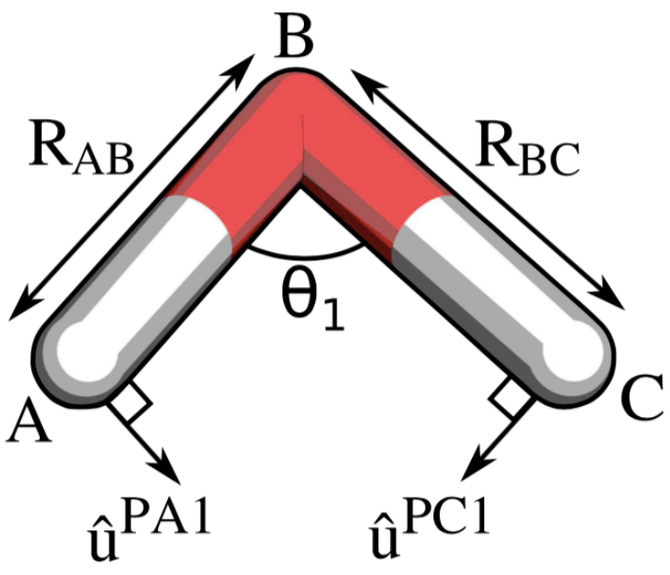
Bonded parameters

What about the bonded parameters?

$$E^{MM} = \sum_{\text{bonds}} k_r(r - r_0)^2 + \sum_{\text{angles}} k_\theta(\theta - \theta_0)^2 + \sum_{\text{dihedrals}} k_\phi(1 + cos(n\phi + \phi_0)) + \sum_{i<j} \frac{q_i q_j}{r_{ij}} + \left(\frac{A^{ij}}{r_{ij}^{12}} - \frac{C_6^{ij}}{r_{ij}^6} \right)$$

Seminario method

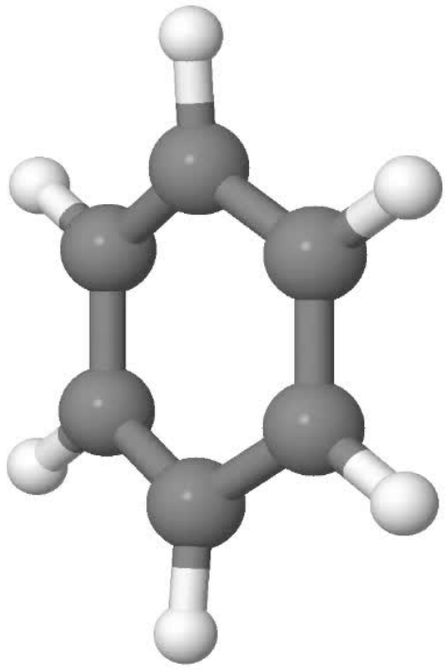
1) Compute 3Nx3N QM Hessian matrix **k**;
2) Extract 3x3 partial matrix involving atoms A and B;
3) Project Hessian eigenvectors onto directions parallel and perpendicular to AB;
4) Compute k_r, k_θ

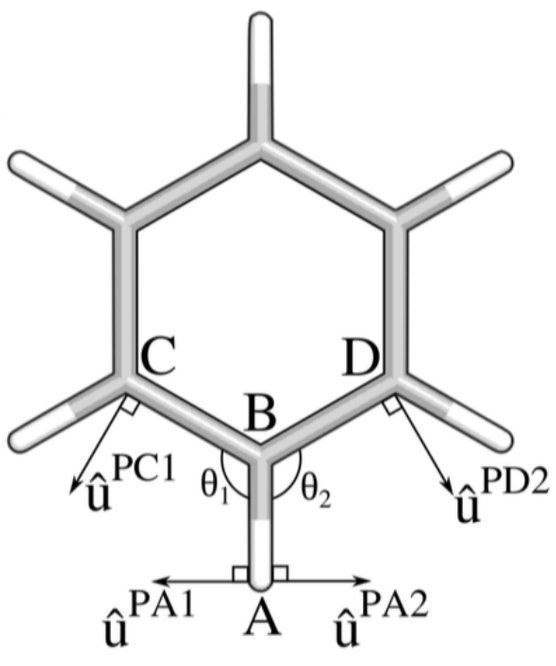


$$[\mathbf{k_{AB}}] = - \begin{vmatrix} \frac{\partial^2 E}{\partial x_A \partial x_B} & \frac{\partial^2 E}{\partial x_A \partial y_B} & \frac{\partial^2 E}{\partial x_A \partial z_B} \\ \frac{\partial^2 E}{\partial y_A \partial x_B} & \frac{\partial^2 E}{\partial y_A \partial y_B} & \frac{\partial^2 E}{\partial y_A \partial z_B} \\ \frac{\partial^2 E}{\partial z_A \partial x_B} & \frac{\partial^2 E}{\partial z_A \partial y_B} & \frac{\partial^2 E}{\partial z_A \partial z_B} \end{vmatrix}$$

J. Seminario, *Int. J. Quantum Chem.*, **1996**, *60*, 1271-1277

Bonded parameters



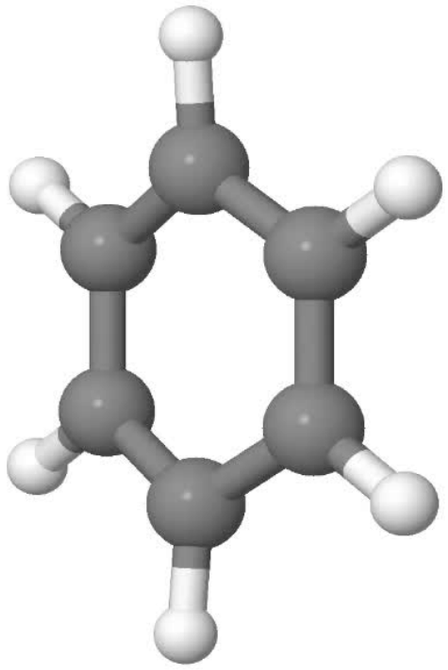


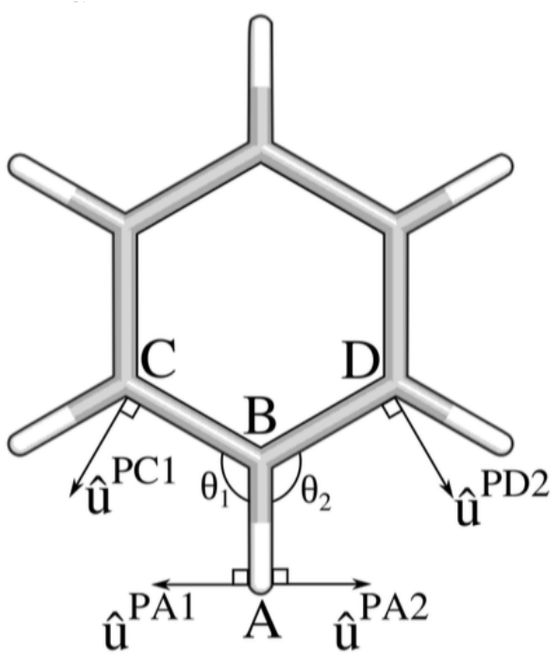
Average error in MM vibrational frequencies of 70 molecules relative to QM (cm^{-1}):

	OPLS	**Seminario**
Small molecules	60	120
Heterocycles	83	132
Dipeptides	47	104

A Allen, M Payne, D Cole, *J. Chem. Theory Comput.*, **2018**, *14*, 274-281

Bonded parameters



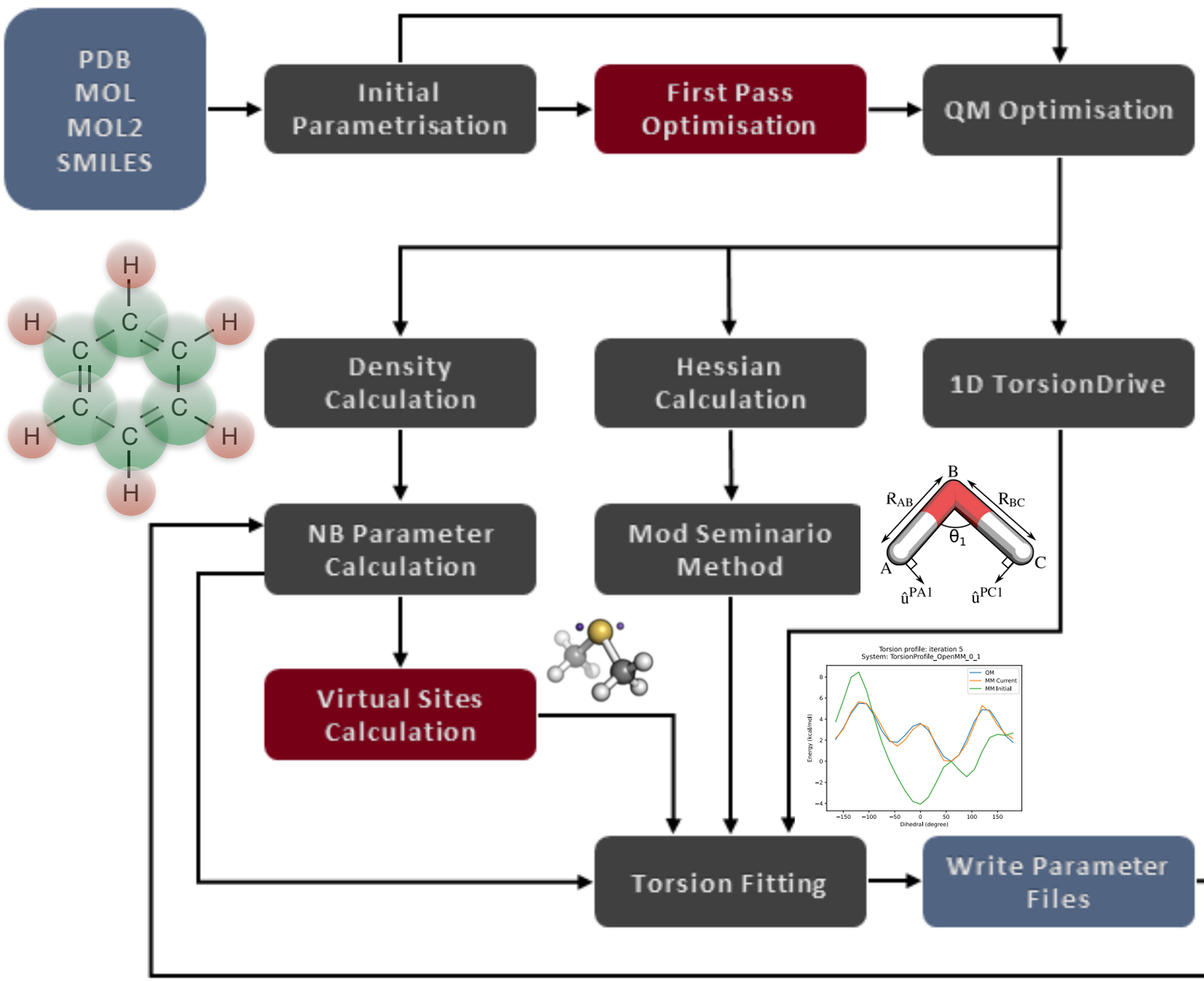


Average error in MM vibrational frequencies of 70 molecules relative to QM (cm^{-1}):

	OPLS	**Seminario**	**Modified Seminario**
Small molecules	60	120	**52**
Heterocycles	83	132	**53**
Dipeptides	47	104	**40**

A Allen, M Payne, D Cole, *J. Chem. Theory Comput.*, **2018**, *14*, 274-281

QUBEKit



Force field design still requires a few user choices (model hyperparameters).

E.g. choice of QM method, AIM partitioning scheme, use of off-site charges.

Interface with ForceBalance allows us to rapidly fit empirical parameters for each force field model.

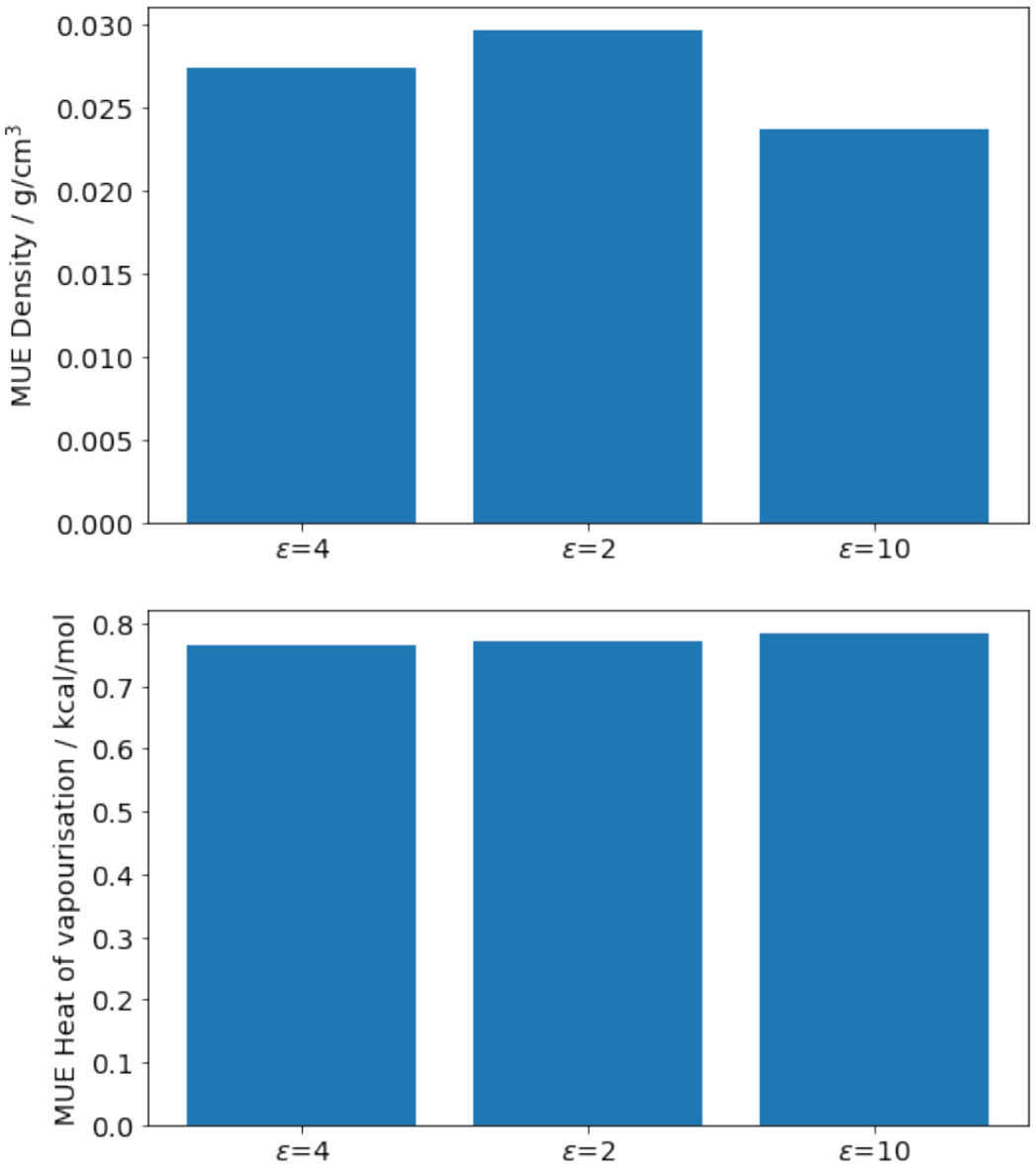
optimise QM-to-MM mapping parameters

Wang L-P, Martinez TJ and Pande VS, J. Phys. Chem. Lett. 2014, 5, 1885-1891.

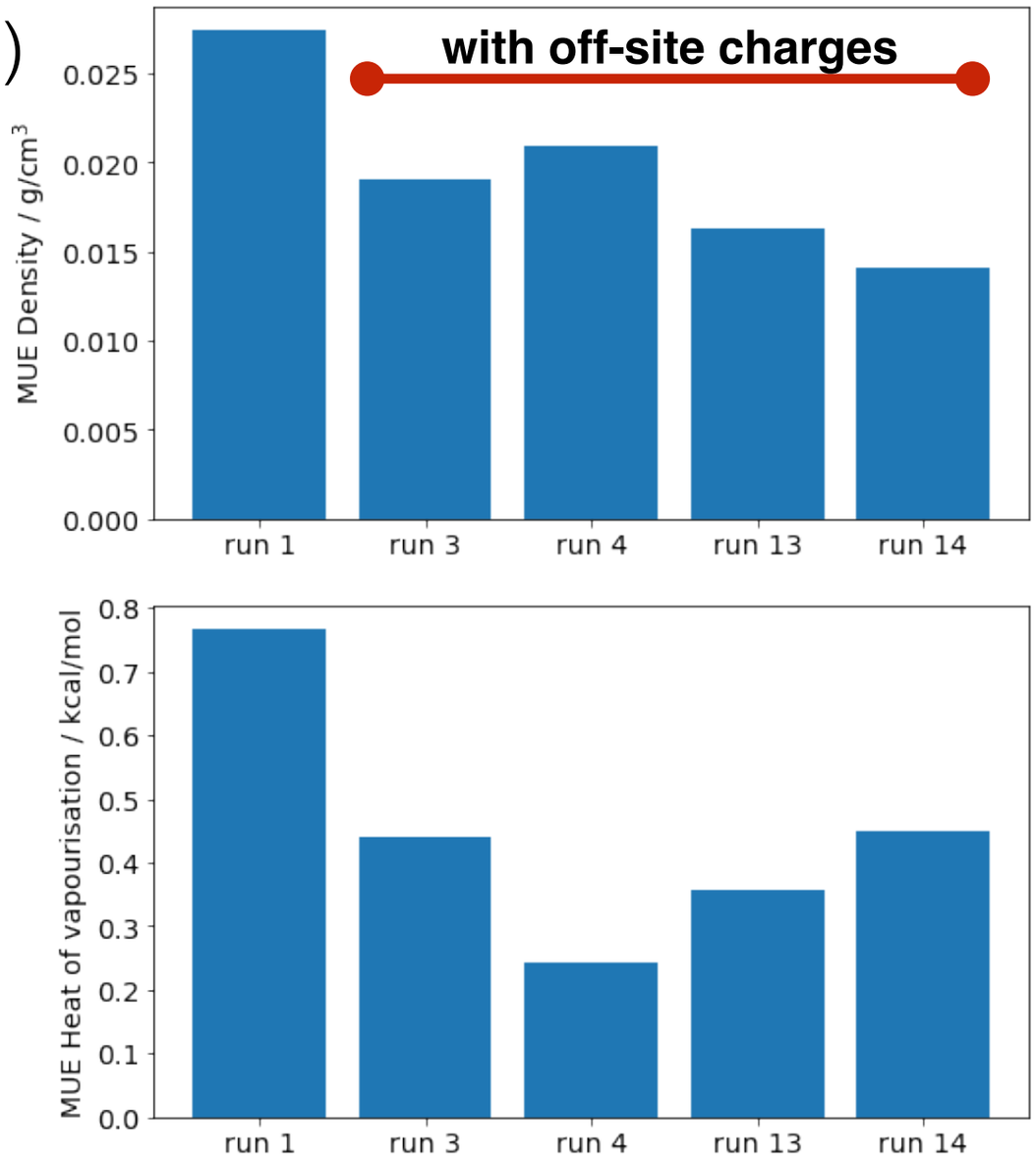
Testing force field hypotheses

Does accuracy of liquid property prediction depend strongly on a) choice of implicit solvent model used in charge calculation, b) use of off-site charges?

a)

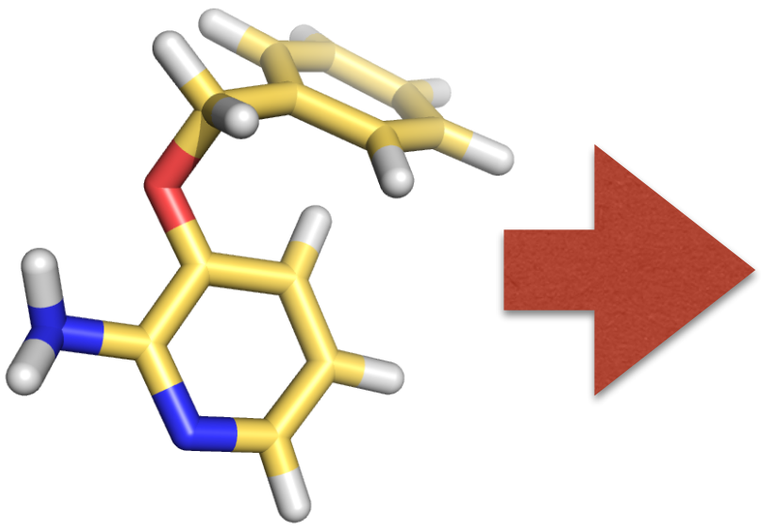


b)



Ringrose, Horton, Wang, Cole. *PCCP* (2022), https://doi.org/10.1039/D2CP02864F

Gaussian Approximation Potential



Total energy and forces

Gaussian approximation potential (GAP) writes the potential energy as a generic function of atomic coordinates. Trained using QM energies and forces.

$$E(\mathscr{A}) = \sum_{\mathscr{B} \in M} x_{\mathscr{B}} K(\mathscr{A}, \mathscr{B})$$

target energy of conformation A (represented by vector of interatomic distances)

unknown coefficients (regularised least squares regression)

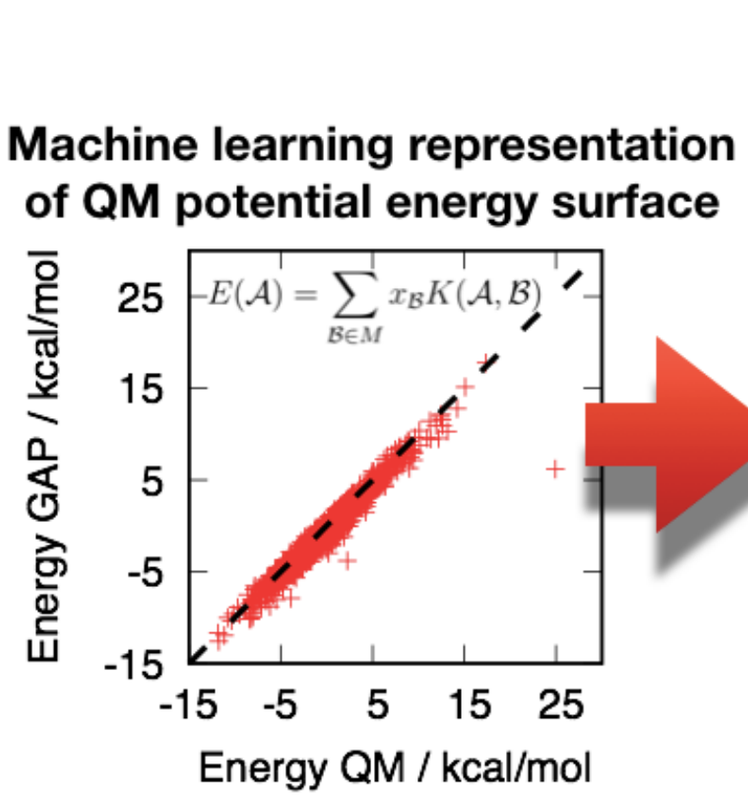
squared exponential kernel (similarity function between A and B)

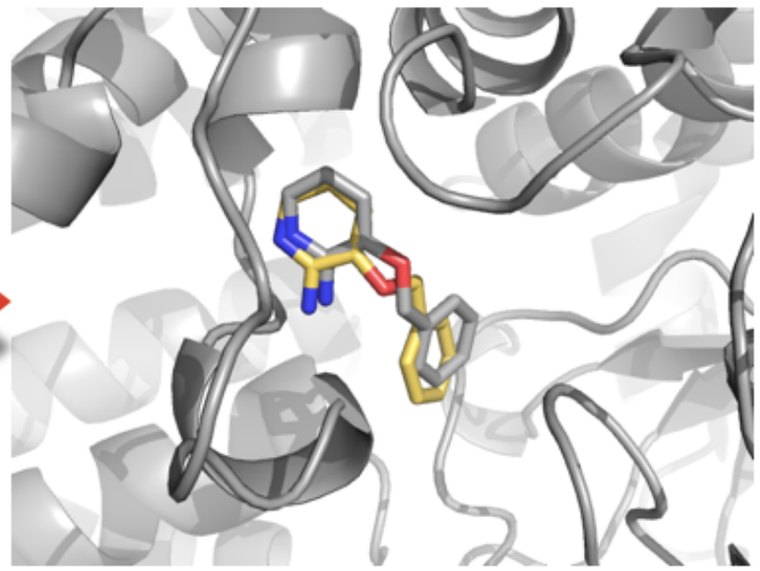
Bartók, Payne, Kondor, Csányi, *Phys. Rev. Lett.* **2010**, *104*, 136403
Bartók, Csányi, *Int. J. Quantum Chem.* **2015**, *115*, 1051-1057

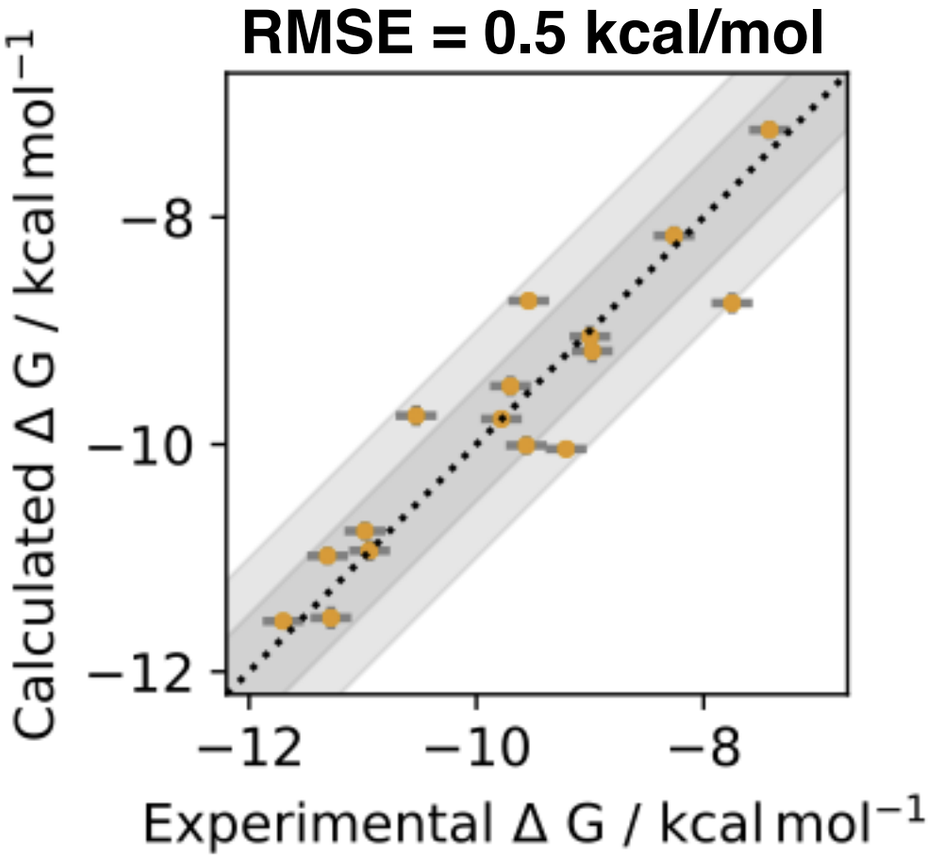
ML Potentials can compute binding free energy corrections

Hybrid ML/MM approaches can be used to simulate small molecule in complex with protein. And importantly also correct MM free energies:

$$E^{tot} = E^{MM}(R) + E^{MM}(RL) + E^{ML}(L)$$







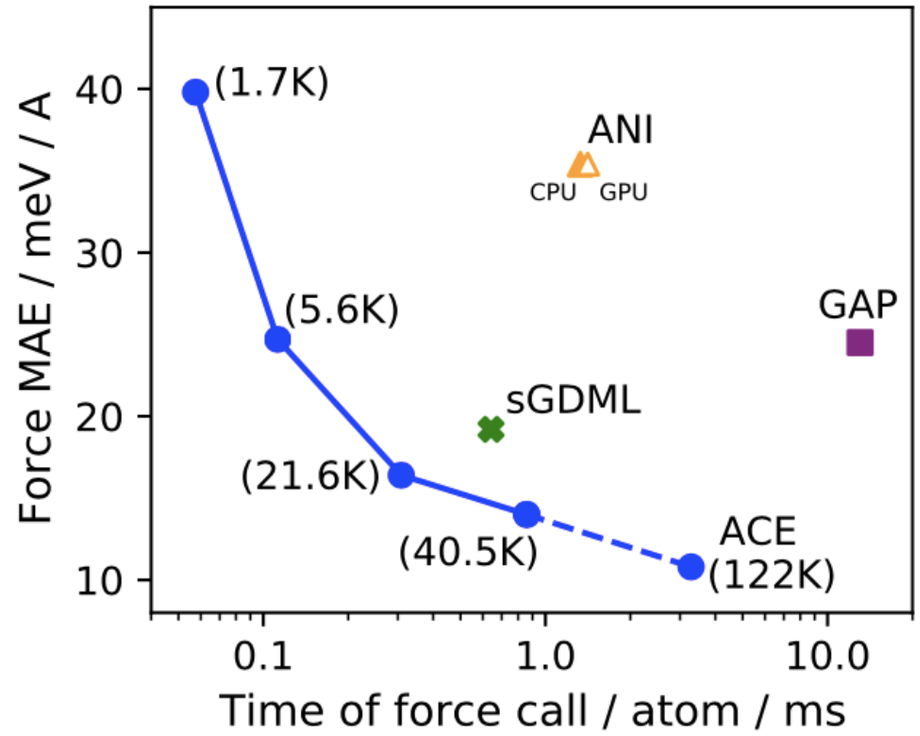
DJ Cole, L Mones, G Csányi,
Faraday Discuss., **2020**, 224, 247-264

Rufa et al. (2020),
https://doi.org/10.1101/2020.07.29.227959

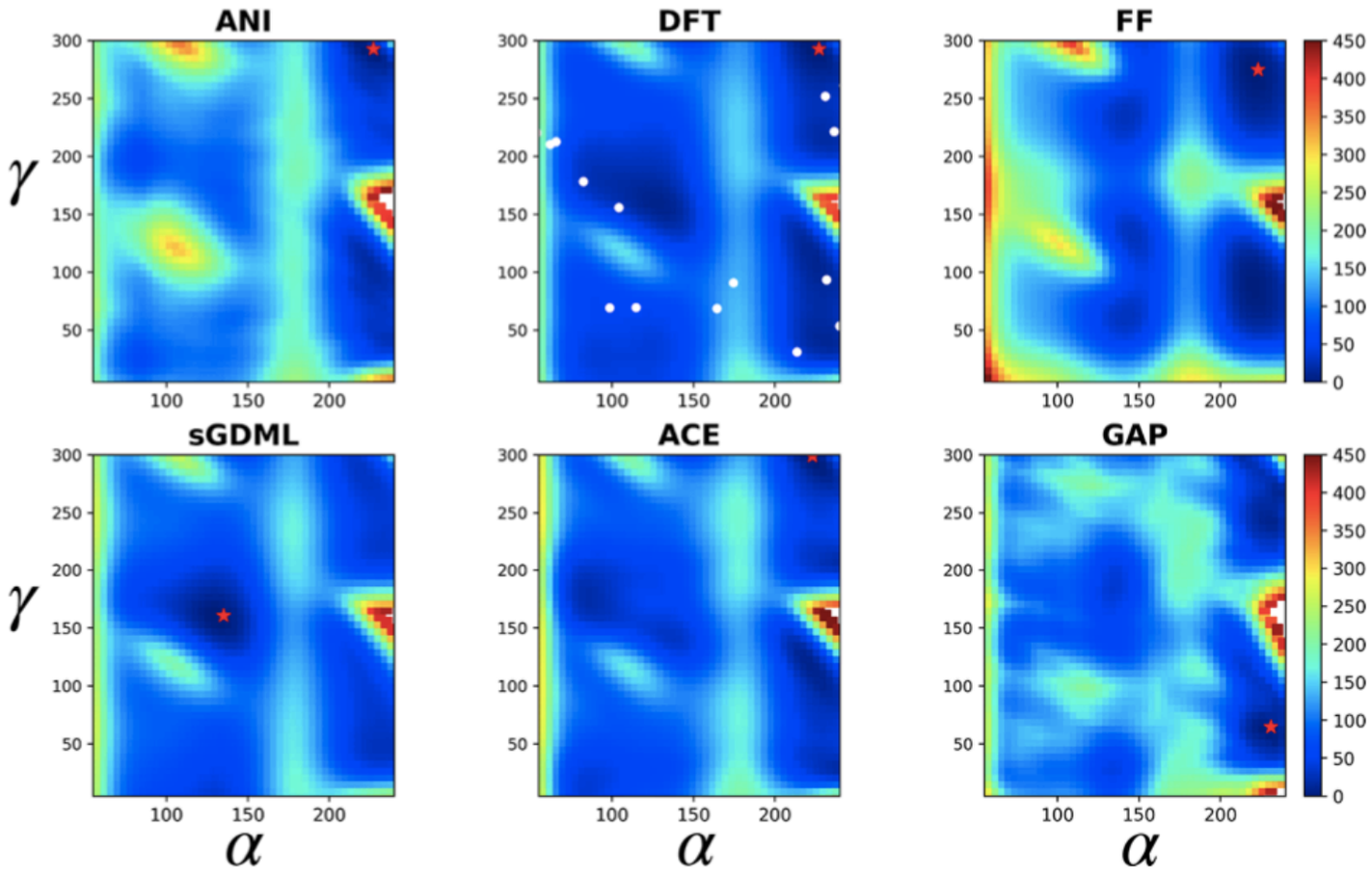
Further improvements in accuracy / speed are possible

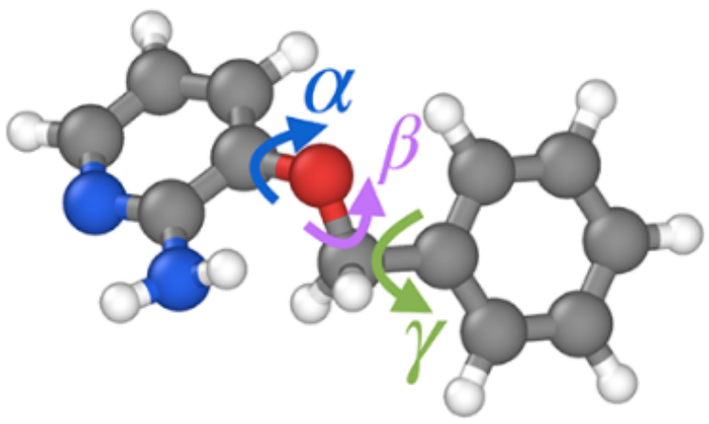
Atomic cluster expansion (ACE): potential energy surface expanded in linear combination of body-ordered polynomial basis functions:

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



$\beta = 120°$

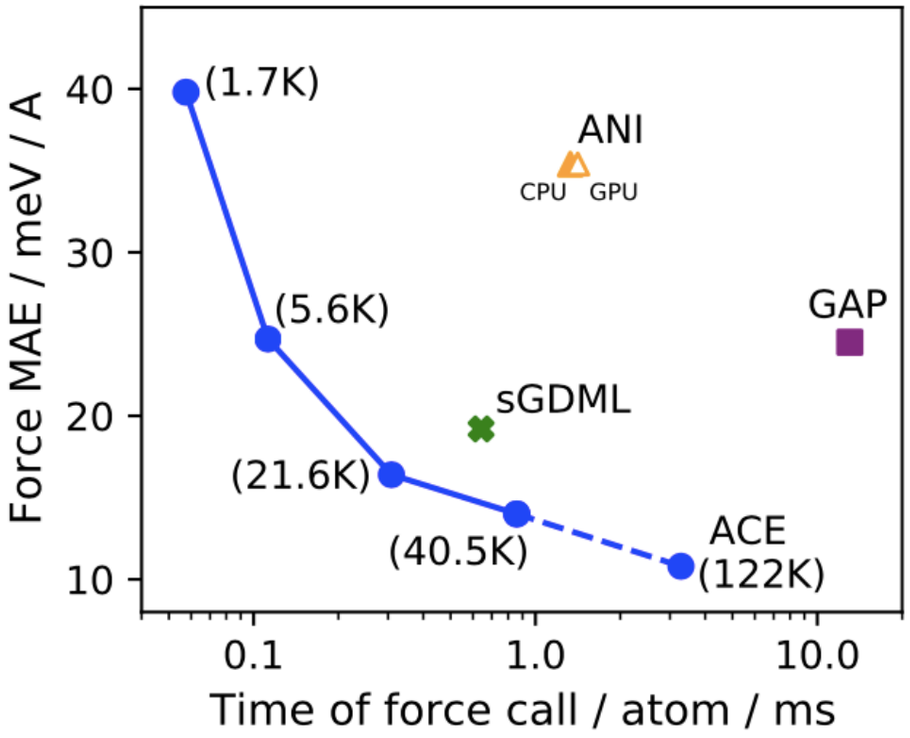




Kovács DP, et al. *JCTC*, 2021, **17**, 7696-7711.

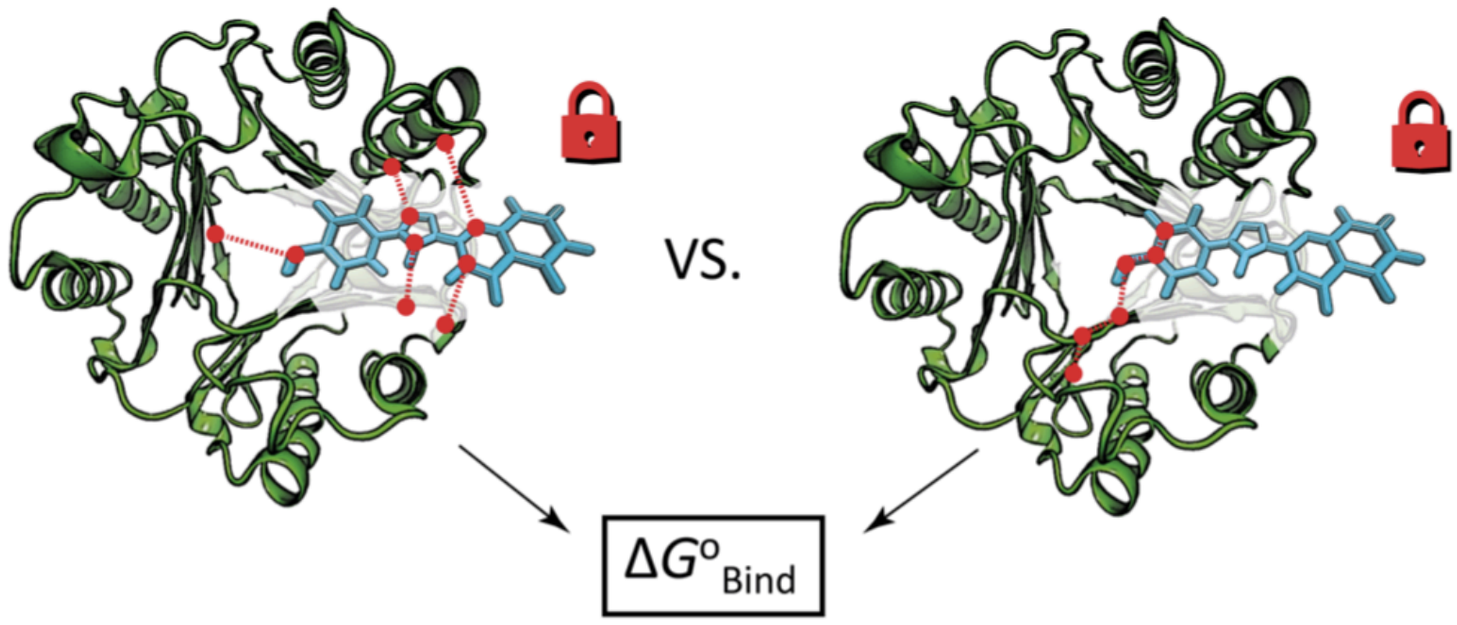
Other Projects for Discussion

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



Fast & accurate linear atomic cluster expansion force fields

Dávid Kovács & Gábor Csányi
J. Chem. Theory Comput. 2021, 17, 7696–7711



Comparison of protein-ligand restraint schemes for ABFE calculations (multiple distance vs Boresch restraints).

Finlay Clark, Graeme Robb & Julien Michel
10.26434/chemrxiv-2023-dd1hc