Supporting Information for: Exploration and Validation of Force Field Design Protocols through QM-to-MM Mapping

Chris Ringrose,[†] Joshua T. Horton,[†] Lee-Ping Wang,[‡] and Daniel J. Cole[*,†]

[†]*School of Natural and Environmental Sciences, Newcastle University, Newcastle upon Tyne NE1 7RU, United Kingdom*

[‡]*Department of Chemistry, The University of California at Davis, Davis, California 95616, United States*

E-mail: daniel.cole@ncl.ac.uk

S1 Lennard-Jones Interactions

Table S1: V^{free} and B^{free} data used for the Lennard-Jones calculations in QUBEKit. V^{free} data is the same as used in our previous study,[1] and B^{free} was taken from the literature.[2]

Element	V^{free} (Bohr^3)	B^{free} (Ha.Bohr^6)
H	7.6	6.5
C	34.4	46.6
N	25.9	24.2
O	22.1	15.6
F	18.2	9.5
S	75.2	134.0
Cl	65.1	94.6
Br	95.7	162.0

S1.1 Derivation of σ and ϵ in QUBE

In what follows we explain how the σ_i and ϵ_i parameters of the QUBE force field are derived from the QM dispersion coefficient (B_i). The OPLS-style Lennard-Jones formula is written as:

$$V_{LJ} = 4\epsilon\left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right) = \frac{A}{r^{12}} - \frac{B}{r^6} \tag{1}$$

where we have dropped the i, j atom labels for simplicity. B is obtained from the Tkatchenko-Scheffler (TS) scaling laws (see main text for definitions of symbols):

$$B = \left(\frac{V^{AIM}}{V^{free}}\right)^2 B^{free} \tag{2}$$

And A may be derived by insisting that the minimum in the vdW energy coincides with the vdW radius of the atom in the molecule:[3]

$$A = \frac{1}{2}B\left(2R^{AIM}\right)^6 = 32B\left(\frac{V^{AIM}}{V^{free}}\right)^2 (R^{free})^6 \tag{3}$$

where we have used:

$$R^{AIM} = \left(\frac{V^{AIM}}{V^{free}}\right)^{1/3} R^{free} \tag{4}$$

From eq 1:

$$B = 4\epsilon\sigma^6 \tag{5}$$

$$A = 4\epsilon\sigma^{12} \tag{6}$$

Dividing eq 6 by eq 5, and substituting in eq 3:

$$\sigma = \left(\frac{A}{B}\right)^{1/6} = 2^{5/6}\left(\frac{V^{AIM}}{V^{free}}\right)^{1/3} R^{free} \tag{7}$$

From eqs 5 and 7:

$$\epsilon = \frac{B}{4\sigma^6} = \frac{B^2}{4A} \tag{8}$$

Substituting in eq 3:

$$\epsilon = \frac{B^2}{4 \times 32B\left(\frac{V^{AIM}}{V^{free}}\right)^2 (R^{free})^6} = \frac{B}{128\left(\frac{V^{AIM}}{V^{free}}\right)^2 (R^{free})^6} \tag{9}$$

Substituting in eq 2:

$$\epsilon = \frac{B^{free}}{2(2R^{free})^6} \tag{10}$$

Note that ϵ is constant for a given element.

S1.2 Derivation of scaled σ and ϵ in QUBE

As discussed in the main text, there is some debate as to whether the dispersion coefficient is actually proportional to the square of the effective atomic volume. Also we know that QUBE dispersion parameters are lower than those in standard classical force fields probably due to the neglect of C_8 and other higher order dispersion terms. Hence here we introduce

α and β as additional fitting parameters:

$$B = \alpha \left(\frac{V^{AIM}}{V^{free}}\right)^{(2+\beta)} B^{free} \tag{11}$$

The derivation of σ proceeds in exactly the same was as Section S1.1, and the formula is unchanged.

As before, from eqs 5 and 7:

$$\epsilon = \frac{B}{4\sigma^6} = \frac{B^2}{4A} \tag{12}$$

Substituting in eq 3:

$$\epsilon = \frac{B^2}{4 \times 32B \left(\frac{V^{AIM}}{V^{free}}\right)^2 (R^{free})^6} = \frac{B}{128 \left(\frac{V^{AIM}}{V^{free}}\right)^2 (R^{free})^6} \tag{13}$$

Substituting in eq 11:

$$\epsilon = \frac{\alpha \left(\frac{V^{AIM}}{V^{free}}\right)^{\beta} B^{free}}{2(2R^{free})^6} \tag{14}$$

Note that ϵ is now dependent on atomic volume.

S1.3 Treatment of Lennard-Jones parameters on polar hydrogen

From our previous work,[3] the Lennard-Jones coefficients of polar hydrogen atoms are set to zero, and neighbouring heavy atoms have their B coefficient increased according to:

$$\sqrt{B'} = \sqrt{B} + n_H \sqrt{B^H} \tag{15}$$

where B' is the new dispersion coefficient and B^H is the original dispersion coefficient on the H atom. Since eq 2 no longer holds for B', we need to update the earlier derivations:

$$A' = \frac{1}{2} B' \left(2R^{AIM}\right)^6 = 32B' \left(\frac{V^{AIM}}{V^{free}}\right)^2 (R^{free})^6 \tag{16}$$

$$B' = 4\epsilon\sigma^6 \tag{17}$$

$$A' = 4\epsilon\sigma^{12} \tag{18}$$

σ is unchanged from Section S1.1:

$$\sigma = \left(\frac{A'}{B'}\right)^{1/6} = \left(32\left(\frac{V^{AIM}}{V^{free}}\right)^2 (R^{free})^6\right)^{1/6} = 2^{5/6}\left(\frac{V^{AIM}}{V^{free}}\right)^{1/3} R^{free} \tag{19}$$

The derivation for ϵ starts in the same way as Section S1.1:

$$\epsilon = \frac{B'}{128\left(\frac{V^{AIM}}{V^{free}}\right)^2 (R^{free})^6} \tag{20}$$

The simplification in eq 10 no longer holds, so the above is the final equation for ϵ for heavy atoms neighbouring polar H atoms.

S2 Virtual Sites

Off-centre charge ("virtual site") derivation in QUBEKit begins with derivation of the electrostatic potential (ESP) around an atom of interest using its QM partitioned electron density. To avoid import of the full electron density, we instead compute the ESP from the atomic multipoles, up to quadrupole order. Atomic dipole vectors ($\boldsymbol{\mu}_i$) can be computed from the partitioned electron densities via:[4]

$$\boldsymbol{\mu}_i = -\int \mathbf{r} n_i(\mathbf{r}) d^3\mathbf{r} \tag{21}$$

where $\mathbf{r}$ is the position vector relative to atom i. The traceless quadrupole tensor is defined by:

$$\mathbf{M}_i = \begin{pmatrix} \left(\frac{Q_{x^2-y^2}}{2} - \frac{Q_{3z^2-r^2}}{6}\right) & Q_{xy} & Q_{xz} \\ Q_{xy} & \left(\frac{-Q_{x^2-y^2}}{2} - \frac{Q_{3z^2-r^2}}{6}\right) & Q_{yz} \\ Q_{xz} & Q_{yz} & \frac{Q_{3z^2-r^2}}{3} \end{pmatrix} \tag{22}$$

where Q_T is computed from the atomic electron density via:

$$Q_T = -\int T n_i(\mathbf{r}) d^3\mathbf{r} \tag{23}$$

for T in xy, xz, yz, x^2-y^2 and $3z^2-r^2$ (where all coordinates are defined relative to the atom centre). Atomic dipole and quadrupole moments are imported into QUBEKit for DDEC via the Chargemol code,[4,5] and for MBIS via PSI4.[6]

For each atom in the molecule, an array of sample points is generated, uniformly distributed around the atom. A total of 1760 sample points are arranged in five shells. The positions of these sample points are scaled relative to their distance from the atom and the atom's van der Waals radius. All points are between 1.4 to 2.0 times the van der Waals radius. This ensures sampling is in the relevant zone around the atom for molecular dynamics simulations. The algorithm for the placement of the points uses a Fibonacci lattice technique, mapped to three-dimensional shells. This reduces bias by ensuring near-equidistant spacing of the points, more so than other mappings such as sphering a cube.

The monopole (V_q^i), dipole (V_μ^i) and quadrupole (V_Q^i) potentials are calculated and summed to give the QM ESP at each sample point around the atom (in atomic units):

$$V_q^i = \frac{q_i}{r} \tag{24}$$

$$V_\mu^i = \frac{\boldsymbol{\mu}_i \cdot \mathbf{r}}{r^3} \tag{25}$$

$$V_Q^i = \frac{\mathbf{r} \cdot \mathbf{M} \cdot \mathbf{r}}{r^5} \tag{26}$$

$$V_{QM}^i = V_q^i + V_\mu^i + V_Q^i \tag{27}$$

where $\mathbf{r}$ is the vector between atom centre and sample point. Cloud penetration potentials were also investigated, but found to have negligible influence on the ESP at typical atom separations.

The monopole ESP is then calculated using a single atom-centred point charge (q_i). If the error, averaged over all sample points is greater than a threshold (here, 1 kcal/mol), virtual sites are added. For fitting, the charge and position of the virtual site is optimised to obtain an ESP as close to the QM ESP as possible (eq 13 in the main text). The charge of the virtual site is subtracted from the charge of its parent atom to preserve the net charge of the molecule. The search directions for virtual site placement are the same as our previous work.[1] Because these vectors are clearly defined, fitting only requires changing the charge magnitude, and the scaling of this vector. This scaling, λ, uniquely defines the position of the virtual site, and may be negative. It is often desirable to force symmetry with the virtual sites. This ensures the distance between the sites and the atom centres is consistent per atom. This has been added as an optional feature within the QUBEKit virtual sites code.

In some cases, multiple virtual sites are required to reduce the ESP error below the threshold. They are fit in a similar manner, also with pre-defined search directions, but using a higher dimensional fitting surface. It was found that the addition of constraints aids optimisation on this fitting surface. Two main constraints were considered on the two virtual sites; one negative and one positive λ, and one negative and one positive charge. Forcing the sites to be fit in these chemically logical positions not only produces lower errors, but also speeds up fitting.

The accompanying Supporting Data includes a notebook with all ESP errors before and after virtual site addition for molecules in the test set.

S3 Additional Computational Methods

S3.1 Dihedral Parameter Fitting

As discussed in the main text, dihedral parameter fits were performed using an interface between QUBEKit, ForceBalance and TorsionDrive. QM one-dimensional dihedral scans were performed in 30° increments, using interfaces between the Gaussian09[7] and PSI4 software,[6] and TorsionDrive,[8] which uses a wavefront propagation algorithm to perform constrained minimisation. At each dihedral angle, the QM optimised geometry is used as a starting point for a restrained optimisation (placing 1 kcal/mol/rad^2 restraints on each atom in the molecule, excluding the four frozen atoms constituting the dihedral angle) using the OpenMM software.[9] The ForceBalance objective function is then a sum of squared differences between the QM and MM final energies (relative to the lowest energy structure). In order to obtain the best fit possible, initial torsion parameters extracted from the starting force field are expanded to have non-zero Fourier coefficients (from V_{1-4}), with new values set to 10^{-5} kcal/mol. In some cases this may introduce redundant parameters and so a L1 regularisation function is used to ensure that redundant parameters are held close to zero. This is the same procedure used in the fitting of the Open Force Field Parsley force field.[10] Parameters for improper and rigid torsions cannot currently be fit using QUBEKit, and so they are taken here from the Parsley force field. Throughout this work, intramolecular nonbonded interactions are excluded for atoms separated by one or two covalent bonds, and we use the standard AMBER 1-4 scaling interactions for atoms separated by three bonds (Coulomb and LJ interactions are reduced by factors of 1.2 and 2.0, respectively).

S3.2 ForceBalance Input Parameters

During each iteration of the ForceBalance optimisation cycle the objective function is evaluated for every target molecule and experimental reference data combination. In this work, we have used pure liquid densities and heats of vapourisation, which are evaluated by ForceBalance via OpenMM and the CUDA platform along with the numerical gradients with respect to the force field fitting parameters (the set of R_i^{free} and optionally α and β) via the central finite difference approximation. Liquid simulation boxes contained 500 molecules under periodic boundary conditions. Simulations were run under the NPT ensemble at standard pressure and the experimental reference temperature, with 10^5 equilibration steps, and 10^6 production steps. Langevin dynamics were used to propagate the system with a timestep of 1 fs, and a 1 ps collision frequency. A Monte Carlo barostat which adjusted the volume every 25 steps was also used. The van der Waals interactions were truncated at an 8.5 Å cut-off, with a switching function applied to the last 1 Å of the interaction to smoothly take the energy to zero. A long-range dispersion correction was used and long-range electrostatics were treated using the particle mesh Ewald method. Gas phase simulations were also required to calculate the heat of vapourisation. Here, a single molecule is simulated with an infinite non-bonded interaction cutoff, again using Langevin dynamics with a 1 fs time step. Gas phase simulations were performed for 10^4 steps of equilibration, followed by 10^5 production steps. The trajectories were then post-processed to calculate the desired properties.

The overall objective function value is calculated as the scaled sum over the squared difference between the reference and simulated target data, with scaling factors chosen to balance the combination of properties which differ by orders of magnitude and remove their units. In this case we have used 30 kg/m^3 for the density and 3 kJ/mol for the heat of vapourisation. Over-fitting was avoided through the use of regularisation. A parabolic restraint penalises large movements away from the starting values. The strength of this restraint is governed by the prior width, which indicates the amount by which we expect the parameters to move during an optimisation. In this work all prior widths were set to 1.0. Table S2 shows the starting parameters used to initiate the ForceBalance runs (in addition, we used $\alpha = 1$ and $\beta = 0.5$, where required). To test the dependence of the ForceBalance fits on the simulation conditions, Table S3 shows the errors in the density and heat of vapourisation following training of model **5b** using the methods described above. We also restarted ForceBalance, using the optimised parameters from **5b** as the new starting parameters. The error in the density and heat of vapourisation changed by less than 0.002 g/cm^3 and 0.1 kcal/mol, respectively. Finally, a second simulation was restarted using a much longer simulation time (5×10^5 equilibration steps, and 5×10^6 production steps for the liquid phase). The reported errors are virtually identical to our standard protocol.

Table S2: The values of the fitting parameters used to initiate each ForceBalance training run.

Element	R_i^{free}
C	2.08
N	1.72
O	1.60
H	1.64
Polar H	1.00
F	1.58
Cl	1.88
Br	1.96
S	2.11

Table S3: Mean unsigned errors in the density and heat of vapourisation for a range of different protocols as described in the text, to test convergence of the ForceBalance training runs.

Training protocol	ρ / g/cm3	$\Delta H_{vap}/kcal/mol$
standard model **5b**	0.0159	0.545
restart	0.0141	0.484
longer time	0.0141	0.474

S3.3 QUBEBench Input Parameters

QUBEBench is an automated framework for computing liquid densities and heats of vapourisation from force fields output by QUBEKit, through an interface with the OpenMM molecular dynamics software.[9] All molecules were simulated at 298.15 K, unless this was above their boiling point, in which case they were simulated at a lower temperature chosen to match the experimental data. The liquid properties were calculated using a box containing exactly 500 molecules; initially a small box was filled, then gradually expanded to accommodate all 500 molecules. The liquid simulation used a Langevin integrator with a 5 ps collison frequency, maintaining temperature, and a Monte Carlo barostat. The time step used was 1 fs in the liquid simulations, and 0.5 fs for the gas phase. Virtual sites were included in molecular dynamics runs (where necessary) using the local coordinate sites implementation in OpenMM. The van der Waals interactions were truncated based on the number of heavy atoms in the molecule, with three possible cut-offs, 11 Å, 13 Å and 15 Å corresponding to a molecule with less than three, five or more heavy atoms. A switching function was also used over the last 0.5 Å and a long range correction was applied, while long range electrostatics were again handled via the particle mesh Ewald method. Simulations were run for a total of 3×10^6 and 6×10^6 steps in the liquid and gas phases, respectively. The first nanosecond of the simulation was excluded for both the liquid and gas property calculations, leaving the remaining steps to be averaged for the results. The heat of vaporisation was then calculated using eq 8 in Ref. 11.

S4 Training and Test Sets

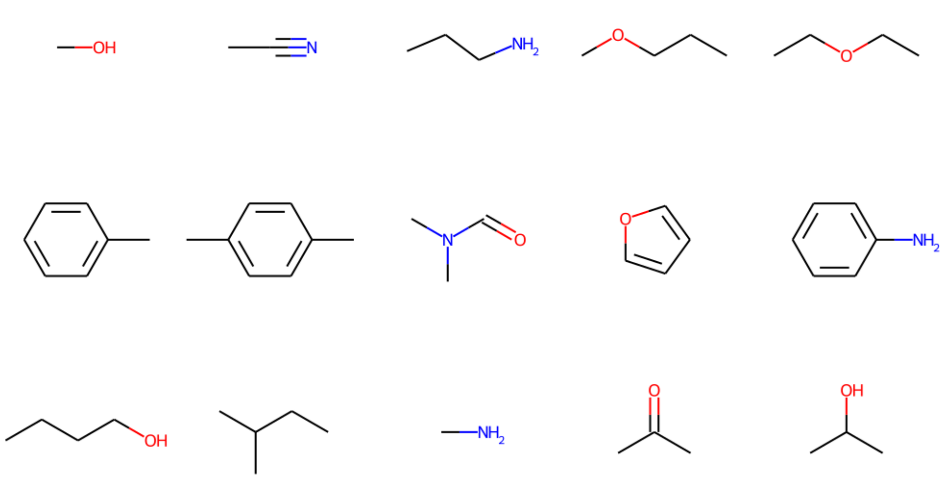

Figure S1: Training set of 15 molecules.

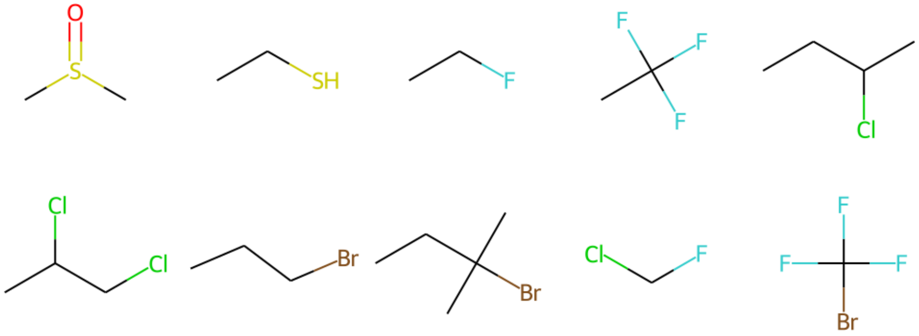

Figure S2: Additional 10 molecules used for training halogen and sulphur-containing molecules.

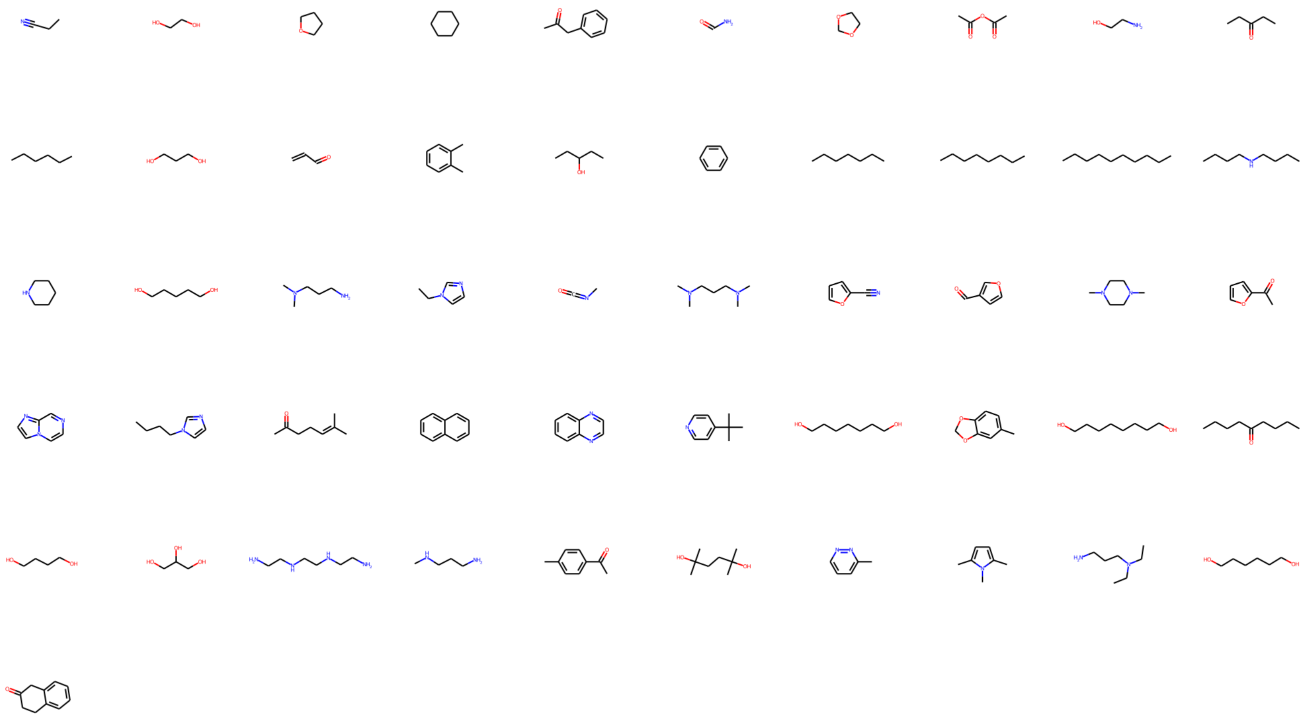

Figure S3: Test set of 51 molecules.

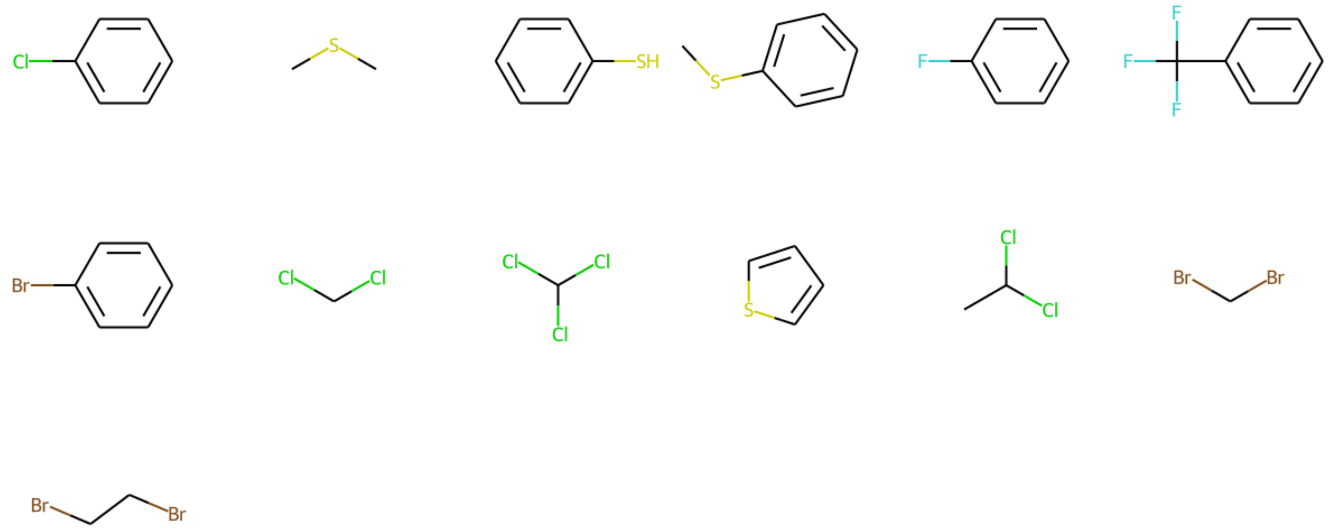

Figure S4: Additional 13 molecules used for testing halogen and sulphur-containing molecules.

S5 Supporting Results

S5.1 Effect of Implicit Solvent Model on Polarity

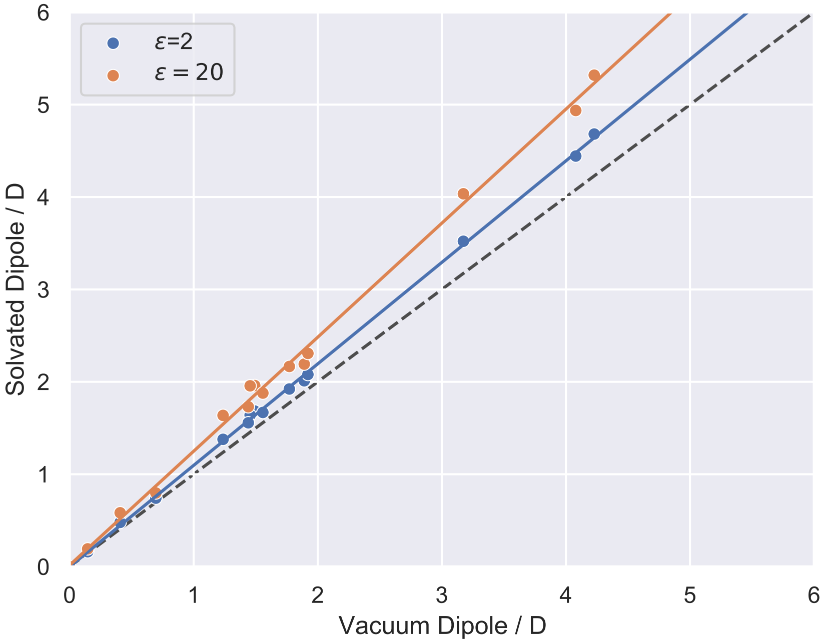

Figure S5: QM dipole moments of the 15 training set molecules computed during derivation of force field models **2a** and **2c**. As expected, the polarity of the molecules increases in a higher dielectric solvent. The gradients of the blue and orange lines are 1.10 and 1.22, respectively.

S5.2 Potential Energy Surfaces for Alternative Force Field Models

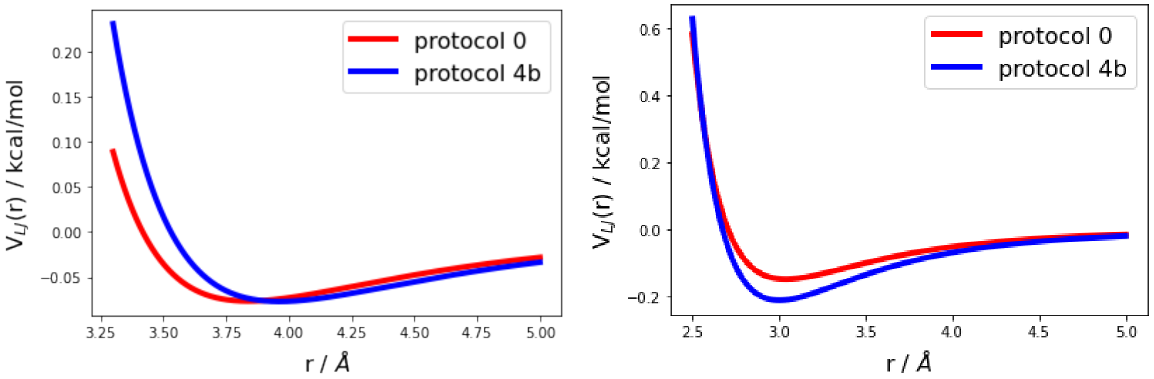

Figure S6: Lennard-Jones potential energy surfaces for force field models **0** and **4b**. See Sections S1.1 and S1.2 for the definitions of the potential, and Table 2 for model parameters. (left) Example plot for two carbon atoms with $V^{AIM} = 30.0$ Bohr^3 and (right) plot for two oxygen atoms with $V^{AIM} = 23.0$ Bohr^3. For carbon, the two curves are similar (except at short-range), since increases in α and β in model **4b** are countered by an increase in R^{free}. For oxygen, model **4b** has a larger well depth, due to increases in α and β and little change in R^{free}, when compared with model **0**.

S5.3 Molecular Dispersion Coefficients for Alternative Force Field Models

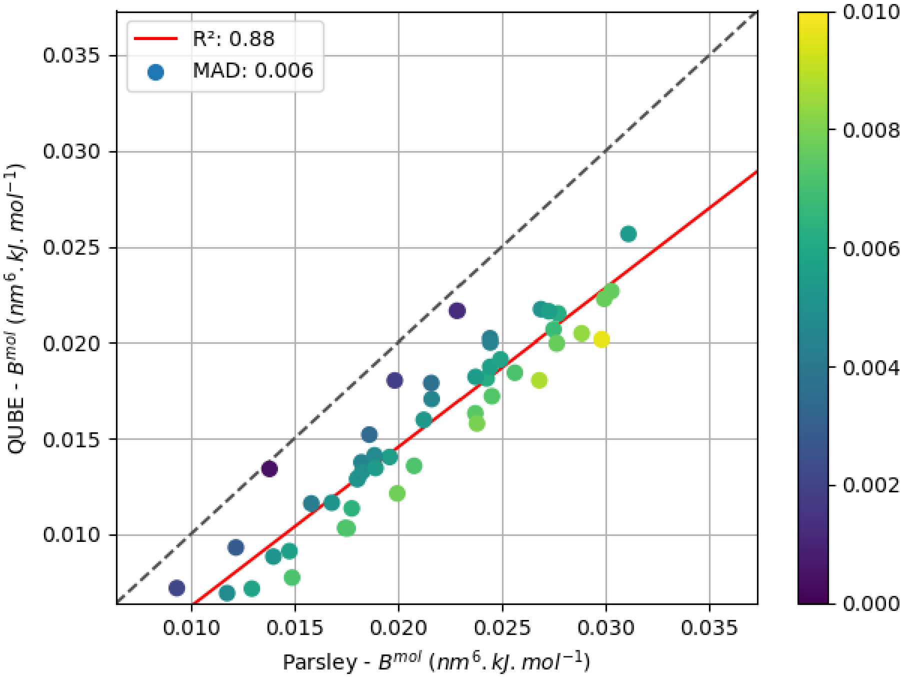

Figure S7: Correlation between Parsley and QUBE (model **0**) summed molecular dispersion coefficients for each molecule in the test set ($B^{mol} = \sum_{i \in mol} B_i$).

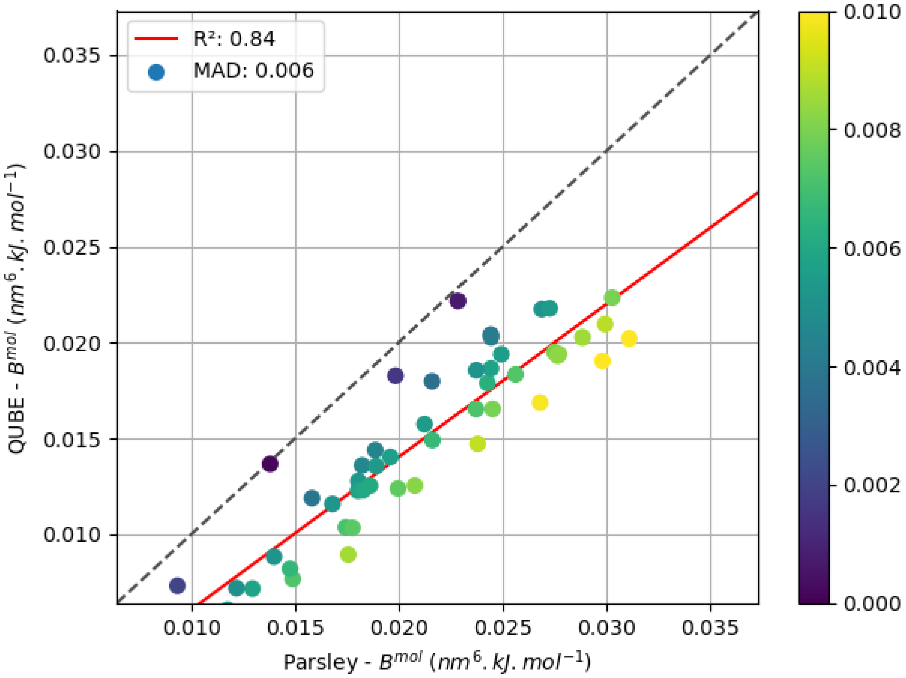

Figure S8: Correlation between Parsley and QUBE (model **1a**) summed molecular dispersion coefficients for each molecule in the test set ($B^{mol} = \sum_{i \in mol} B_i$).

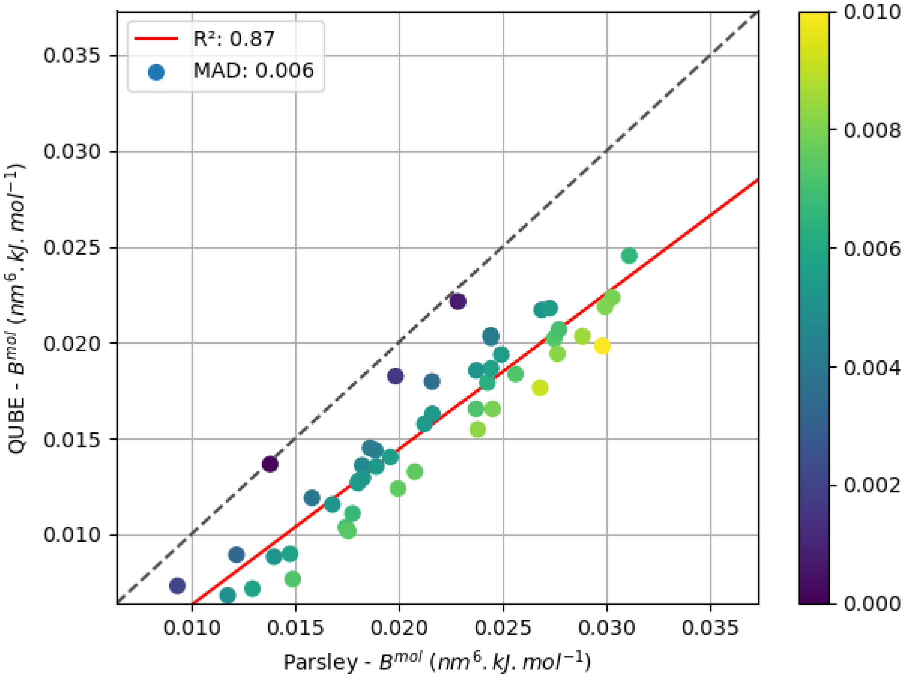

Figure S9: Correlation between Parsley and QUBE (model **5d**) summed molecular dispersion coefficients for each molecule in the test set ($B^{mol} = \sum_{i \in mol} B_i$).

S5.4 R_i^{free} parameters for halogen and sulphur training sets

Table S4: Values of the fitting parameters following ForceBalance optimisation for force field protocols **5b** and **5d** against joint training sets (see Figures S1 and S2). The R_i^{free} for each element are in Å, and α and β are dimensionless.

Model	C	N	O	H	polar H	α	β	F	Cl	Br	S
5b	2.019	1.733	1.598	1.747	1.332	1.182	0.445	1.652	1.867	2.001	2.036
5d	1.984	1.717	1.580	1.749	1.447	–	–	1.628	1.831	1.964	1.983

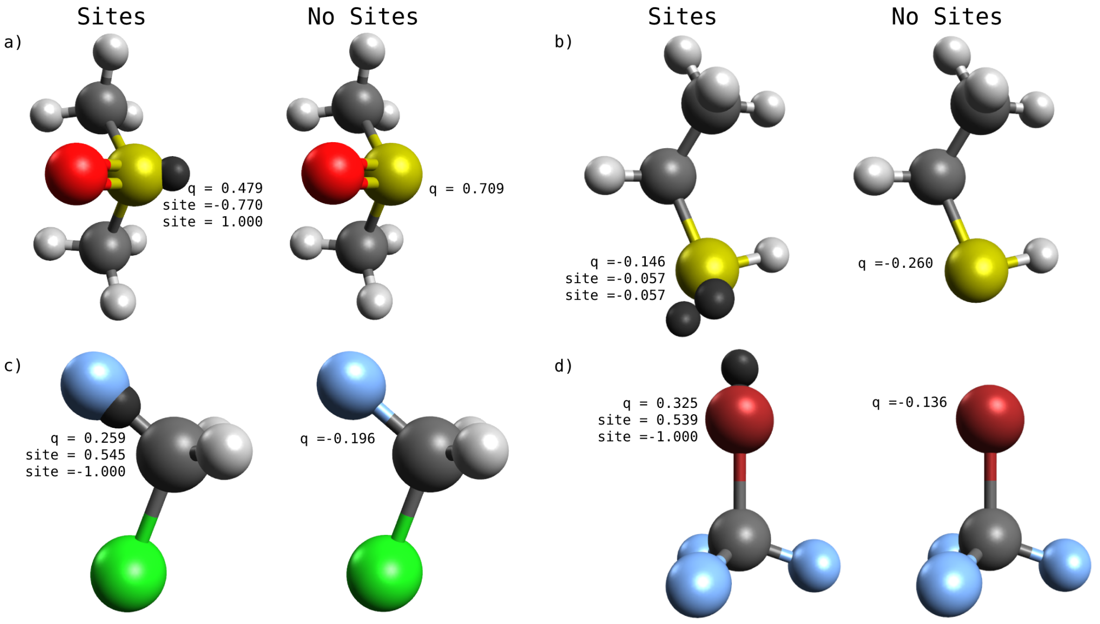

Figure S10: A selection of charge models for the halogen/sulphur training set parameterised using QUBE (model **5b**). As expected, negatively charged sites are located in the lone pair positions of sulphur in a) dimethyl sulphoxide and b) ethanethiol. c) A positive charge along the C–F bond in chlorofluoromethane (and a negative charge close to the atom) reduce the ESP error on F from 1.35 to 0.04 kcal/mol. d) In bromotrifluoromethane, a positive charge is located in the σ-hole position along the C–Br bond.

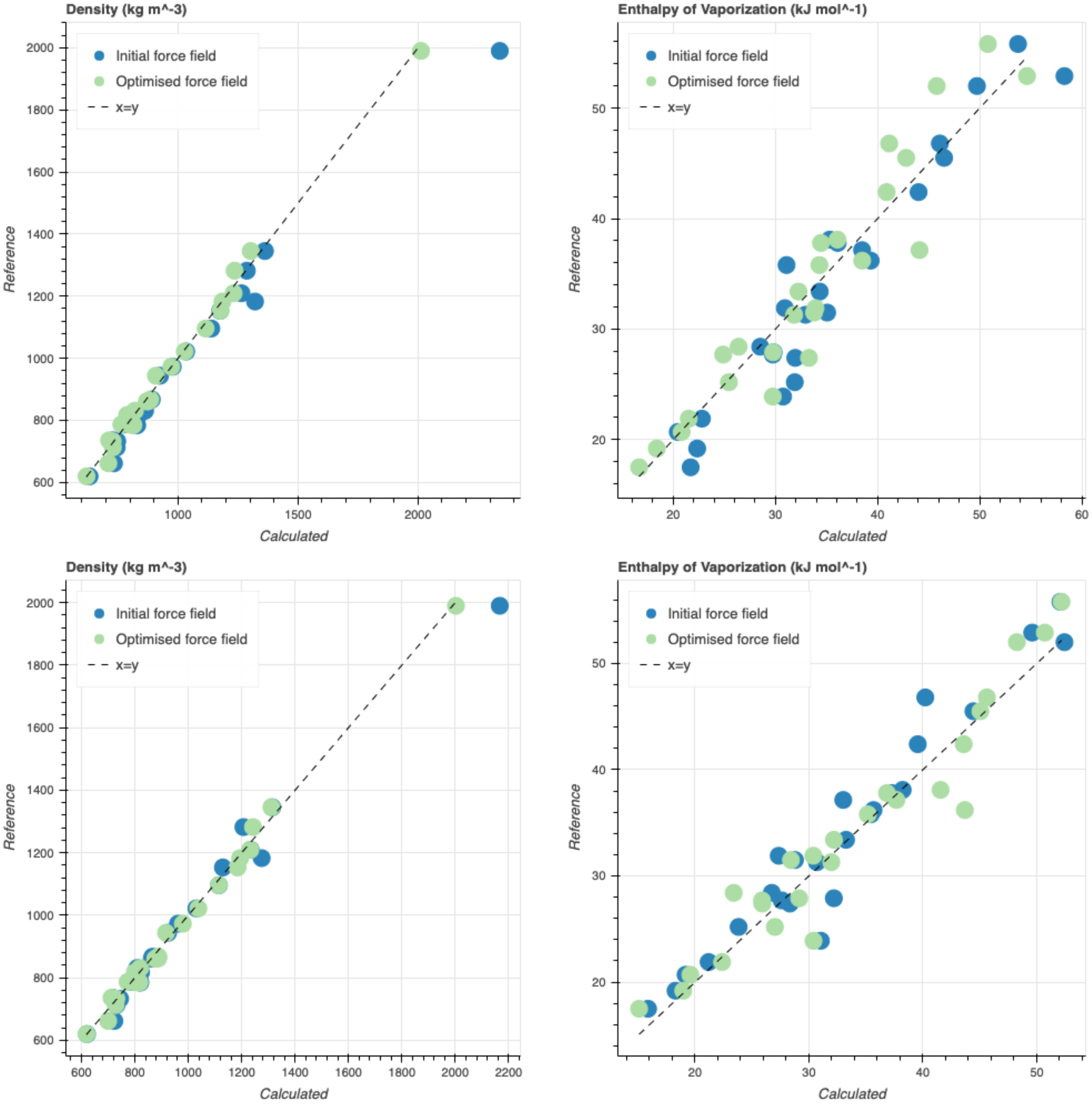

Figure S11: Correlation between calculated and experimental reference data for the expanded training sets (including halogens and S atoms) before and after re-optimisation with ForceBalance. The top row shows model **5b** and the bottom row model **5d**.

S5.5 Analysis of halogen and sulphur test set

Table S5: Test set accuracy. MUE in density (ρ) and heat of vaporisation (ΔH_{vap}) for 13 molecules in the test set containing halogens and sulphur atoms.

Force Field	MUE ρ (g/cm^3)	MUE ΔH_{vap} (kcal/mol)
Parsley	0.114	1.10
Model **5b**	0.057	0.89
Model **5d**	0.060	1.08

Table S5 shows the accuracy of Parsley and two of our developed force field models on a test set comprising molecules containing halogens and sulphur atoms. The test set is shown in Figure S4, the force field model mapping parameters are shown in Section S5.4, and the full data set is provided in the Supporting Data. Compared to Table 3 in the main text, the accuracy of these force field models in terms of the heat of vaporisation, is as expected with errors in the range 0.9 – 1.1 kcal/mol. The density errors, on the other hand, are higher than observed previously, though our **5b** and **5d** models are substantially more accurate than Parsley for this data set. Analysis of the data for individual molecules shows that in all cases the error is dominated by dibromomethane and 1,2-dibromoethane (e.g. for model **5d** the MUE falls to 0.040g/cm^3 upon excluding these molecules). Thus, further improvement of the protocols for small, highly halogenated molecules may be useful.

References

(1) Horton, J. T.; Allen, A. E.; Dodda, L. S.; Cole, D. J. QUBEKit: Automating the Derivation of Force Field Parameters from Quantum Mechanics. *J. Chem. Inf. Model.* **2019**, *59*, 1366–1381.

(2) Chu, X.; Dalgarno, A. Linear response time-dependent density functional theory for van der Waals coefficients. *J. Chem. Phys.* **2004**, *121*, 4083–4088.

(3) Cole, D. J.; Vilseck, J. Z.; Tirado-Rives, J.; Payne, M. C.; Jorgensen, W. L. Biomolecular Force Field Parameterization via Atoms-in-Molecule Electron Density Partitioning. *J. Chem. Theory Comput.* **2016**, *12*, 2312–2323.

(4) Manz, T. A.; Limas, N. G. Introducing DDEC6 atomic population analysis: part 1. Charge partitioning theory and methodology. *RSC Advances* **2016**, *6*, 47771–47801.

(5) Limas, N. G.; Manz, T. A. Introducing DDEC6 atomic population analysis: part 2. Computed results for a wide range of periodic and nonperiodic materials. *RSC Advances* **2016**, *6*, 45727–45747.

(6) Smith, D. G. A. et al. PSI4 1.4: Open-source software for high-throughput quantum chemistry. *The Journal of Chemical Physics* **2020**, *152*, 184108.

(7) Frisch, M. J. et al. Gaussian 09 Revision A.2. 2009.

(8) Qiu, Y.; Smith, D. G.; Stern, C. D.; Feng, M.; Jang, H.; Wang, L. P. Driving torsion scans with wavefront propagation. *J. Chem. Phys.* **2020**, *152*, 244116.

(9) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology* **2017**, *13*, e1005659.

(10) Qiu, Y. et al. Development and Benchmarking of Open Force Field v1.0.0—the Parsley Small-Molecule Force Field. *Journal of Chemical Theory and Computation* **2021**, *17*, 6262–6280.

(11) Wang, J.; Hou, T. Application of Molecular Dynamics Simulations in Molecular Property Prediction. 1. Density and Heat of Vaporization. *Journal of Chemical Theory and Computation* **2011**, *7*, 2151–2165.