JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

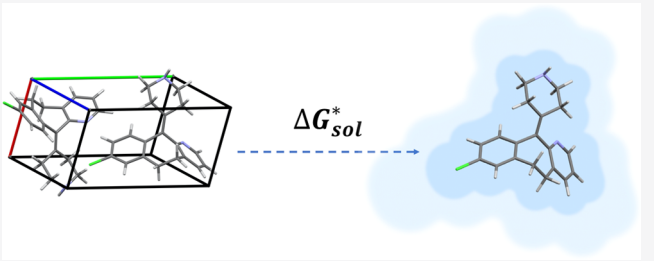
Toward Physics-Based Solubility Computation for Pharmaceuticals to Rival Informatics

Daniel J. Fowles, David S. Palmer,* Rui Guo, Sarah L. Price, and John B. O. Mitchell*

Cite This: https://doi.org/10.1021/acs.jctc.1c00130 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: We demonstrate that physics-based calculations of intrinsic aqueous solubility can rival cheminformatics-based machine learning predictions. A proof-of-concept was developed for a physics-based approach via a sublimation thermodynamic cycle, building upon previous work that relied upon several thermodynamic approximations, notably the $2RT$ approximation, and limited conformational sampling. Here, we apply improvements to our sublimation free-energy model with the use of crystal phonon mode calculations to capture the contributions of the vibrational modes of the crystal. Including these improvements with lattice energies computed using the model-potential-based Ψ_{mol} method leads to accurate estimates of sublimation free energy. Combining these with hydration free energies obtained from either molecular dynamics free-energy perturbation simulations or density functional theory calculations, solubilities comparable to both experiment and informatics predictions are obtained. The application to coronene, succinic acid, and the pharmaceutical desloratadine shows how the methods must be adapted for the adoption of different conformations in different phases. The approach has the flexibility to extend to applications that cannot be covered by informatics methods.



Downloaded via 82.26.43.192 on May 26, 2021 at 14:33:08 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

INTRODUCTION

Solubility is a fundamental physicochemical property, understanding of which is essential for design and manufacturing processes in industries ranging from petrochemicals to energy materials. It is of particular significance for the pharmaceutical industry, with up to 70% of drugs in development having solubility problems and with low aqueous solubility being a frequent cause of failure of drug candidates.[1,2] Although the pharmaceutical industry makes extensive use of experimental solubility measurements, they are time-consuming, resource-intensive, and only applicable to already-synthesized molecules, which limits their breadth of application. Consequently, there is a pressing need for accurate computational models to predict solubility.

Recently, various physics-based approaches have been proposed to compute intrinsic aqueous solubility, specifically the equilibrium solubility of the neutral form of the solute, written as S_0. Such methods generally rely on explicit simulations, as with the Frenkel group's method that identified the nonstandard conditions where the solution has the same chemical potential as the solid.[3,4] They calculated reversible paths between the Einstein crystal, a simple hypothetical model of a solid, and the real crystalline solute using molecular dynamics (MD) simulations. The aqueous solution phase was modeled with a separate simulation where a cavity was grown in water. A molecule of the solute compound was then placed inside it before the cavity was computationally shrunk to leave the molecule in a simulated aqueous solution. Another method is the direct coexistence approach of Kolafa,[5] who explicitly simulated a solute dissolving in a solvent and counted the number of solute particles in the simulated solution phase to identify the concentration at which equilibrium was reached. In a strikingly different methodology, the Anwar group used Monte Carlo simulations to compute the density of states of the solution phase. This was designed to produce two separate peaks, one corresponding to the pure solute and the other to the saturated solution. This second peak's mole fraction of solute was the equilibrium solubility.[6,7] Lüder and co-workers published four papers aimed at computing the solubility of druglike compounds via simulations.[8−11] Their studies considered a roundabout route from the solid via amorphous solid and supercooled liquid to the aqueous solution and took advantage of an empirical relationship between the solubilities of the crystalline and amorphous phases, rather than modeling the crystal lattice explicitly. They were able to generate reasonable solubility predictions using only widely affordable simulation techniques but also found that the additional expense of free-energy perturbation (FEP) calculations was

Received: February 4, 2021

ACS Publications © XXXX The Authors. Published by American Chemical Society

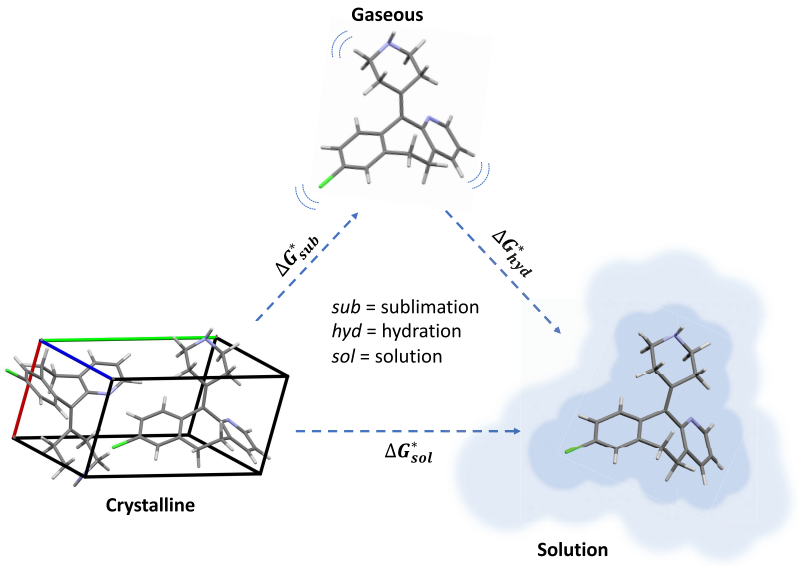


Figure 1. Thermodynamic cycle for transfer from crystal to gas to solution.

rewarded with substantially more accurate results. Mondal et al. have recently also found that free-energy perturbation calculations can successfully model solubility.[12]

The most relevant comparison for the present study is with our own previous work,[13] which also bears some similarity to the more recent approach of Abramov et al.[14] We used a sublimation cycle approach, computing the free-energy changes of sublimation and hydration under standard conditions and then summing them to obtain the free energy of solution and hence the equilibrium constant describing aqueous solubility. Those preliminary results showed that solubility can be accurately calculated without empirical parameterization against experimental data, with possible procedural improvements further narrowing the gap between predicted and experimental data. Such improvements are now possible, as we can calculate all of the modes in the crystal using periodic density functional methods, and hence no longer need to use rigid-body crystal modes to estimate the entropy of sublimation. Careful analysis of the vibrational modes enables us to convert lattice energies into sublimation enthalpies without relying on the $2RT$ approximation used in previous work. Improved hydration free energies are computed using two separate approaches: density functional theory (DFT) with full enumeration of low-energy gas and solution-phase conformers, and molecular dynamics (MD) simulations with free-energy perturbation (FEP). It is our belief that such physics-based solubility predictions can rival the currently dominant cheminformatics and machine learning methods with a more physically grounded approach, representing the thermodynamics of each stage of the solubility process.

THEORY

Calculation of Intrinsic Aqueous Solubility from Solution Free Energy. Intrinsic aqueous solubility is defined as the concentration of the neutral form of the molecule in a saturated aqueous solution at thermodynamic equilibrium.[15,16] If the activity coefficient for the solute in solution is assumed to be unity, then the link between intrinsic solubility and Gibbs solution free energy is

$$\Delta G^*_{sol} = \Delta G^*_{sub} + \Delta G^*_{hyd} = -RT \ln(S_0 V_m) \quad (1)$$

where ΔG^*_{sol}, ΔG^*_{sub}, and ΔG^*_{hyd} are the Gibbs free energies for solution, sublimation, and hydration, respectively, R is the molar gas constant, T is the temperature, V_m describes the molar volume of the crystal, and S_0 refers to the intrinsic solubility (using moles per liter, mol/L). The superscript asterisk indicates that the Ben-Naim terminology is being used and refers to the Gibbs free energy for transfer of a molecule between two phases at a fixed center of mass in each phase.[17,18] The relationship between ΔG^*_{sol}, ΔG^*_{sub}, and ΔG^*_{hyd} is based on a thermodynamic cycle via the gas phase, as illustrated in Figure 1.

Although solvation free energies are commonly reported in the Ben-Naim standard states, sublimation free energies are more commonly calculated and reported relative to a 1 atm standard state in the gas phase, ΔG^0_{sub}. The conversion between them is

$$\Delta G^*_{sub} = \Delta G^0_{sub} - RT \ln(V_m p_0 / RT) \quad (2)$$

where p_0 is the atmospheric pressure. Combining eqs 1 and 2 gives an expression for S_o that does not include V_m

$$S_o = \frac{p_o}{RT} \exp\left(\frac{\Delta G^o_{sub} + \Delta G^*_{hyd}}{-RT}\right) \quad (3)$$

Calculation of Sublimation Free Energy beyond the $2RT$ Approximation. Previous calculations of solubility via the sublimation cycle have relied upon the $2RT$ approximation to convert calculated crystal lattice energies into sublimation enthalpies.[13] Breaking down this approximation by first ignoring phonon dispersion in the crystal (i.e., considering the Γ-point phonons only) and assuming that the intramolecular vibrations of a molecule are the same in the gas phase and the crystal enable the sublimation enthalpy to be written as

https://doi.org/10.1021/acs.jctc.1c00130
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

$$\Delta H^{\circ}_{\mathrm{sub}}(T) = -E_{\mathrm{latt}} + 4RT - \sum_{i'} \frac{\hbar\omega^{s}_{i'}}{2} - \sum_{i'} \left(\frac{\hbar\omega^{s}_{i'}}{\exp\left(\frac{\hbar\omega^{s}_{i'}}{k_{\mathrm{B}}T}\right) - 1} \right) \quad (4)$$

The summations run only over the intermolecular phonon modes, which are assumed not to mix with the intramolecular vibrational modes at all, an approximation that is clearly more appropriate to small rigid molecules. If the intermolecular modes were at low frequencies (<200 cm^{-1}), then at room temperature, they could be considered as classical harmonic oscillators, i.e., their zero-point energies can be ignored and the equipartition theorem can be applied. There are six such modes for each molecule in the unit cell; thus, the last two terms in eq 4 can be replaced as $-6RT$, and one arrives at the $2RT$ approximation[19] using eq 5

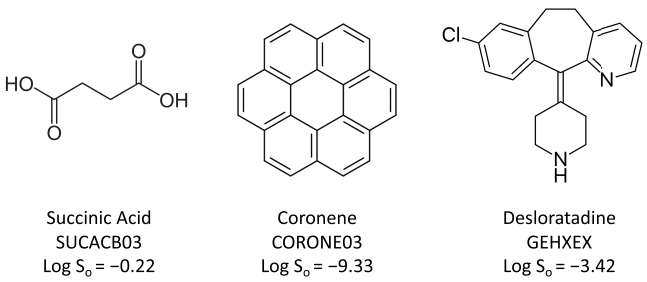
$$\Delta H^{\circ}_{\mathrm{sub}}(T) \approx -E_{\mathrm{latt}} - 2RT \quad (5)$$

However, the $2RT$ approximation is not sufficiently accurate for quantitative solubility predictions. Indeed, recent calculations even on small organic crystals have shown that the approximation inherent in eq 5 can be seriously in error.[20] When there is intermolecular hydrogen bonding or intramolecular modes that are similar to or even lower in frequency than the lattice phonon modes, the assumptions that the modes do not mix and that their contributions follow equipartition are highly questionable.[20] Hence, we contend that a clear route to improvement in modeling lattice thermodynamics lies in revisiting the $2RT$ approximation. Through resource-intensive phonon calculations, current periodic DFT-D codes can provide the enthalpy, entropy, and free-energy contributions of each vibrational and phonon mode. We believe that such accurate computation is a prerequisite for the chemical accuracy needed to compute aqueous solubility with an RMS error comparable with that of informatics methods,[21] around 0.7−1.1 log S_0 units, or the typical experimental error of around 0.6−0.7 log S_0 units.[22]

Hydration Free-Energy Calculations. Previous predictions of solubility via a sublimation cycle have used implicit solvent models to compute hydration free energy from a single low-energy conformer in each phase. Here, we investigate two alternative approaches that explicitly account for the conformational degrees of freedom of the solute. First, we use density functional theory and a Boltzmann-weighting scheme to compute hydration free energies from an ensemble of low-energy conformers in each phase. Second, we compute hydration free energies from atomistic molecular dynamics simulations using free-energy perturbation methods.

COMPUTATIONAL METHODS

Data Set. A small data set of three druglike molecules, succinic acid, coronene, and desloratadine, was used to test these physics-based methods. These three molecules contain differing chemical structures and functional groups, representing a wide range of flexibilities, sizes, and solubilities. All three have multiple polymorphs; however, this study focuses only on the thermodynamically most stable form of each compound under ambient conditions. The chemical structures, common molecular names, and Cambridge Structural Database (CSD) refcodes for the polymorph used in calculations are shown in Figure 2.



Succinic Acid
SUCACB03
Log S_o = −0.22

Coronene
CORONE03
Log S_o = −9.33

Desloratadine
GEHXEX
Log S_o = −3.42

Figure 2. Structures, Cambridge Structural Database refcodes, and experimental solubilities of the three compounds considered in this study.

Experimental data was found in the literature, with intrinsic aqueous solubilities (measured as mol/L) reported for each molecule as follows: Forbes and Coolidge[23] reported a value of log S_0 = −0.22 at 25 °C for succinic acid; Miller et al.[24] reported a value of log S_0 = −9.33 at 25 °C for coronene; Popović et al.[25] reported a value of log S_0 = −3.42 at 25 °C for desloratadine. Since the crystalline polymorphic forms of the solutes in these solubility assays were not specified, the sublimation calculations were performed using the polymorph of each solute that is known to be most stable under ambient conditions. The sublimation calculations were performed using the β polymorph of succinic acid (CSD[26] refcode: SUCACB03), the γ polymorph of coronene (CSD refcode: CORONE03), and Form I of desloratadine (CSD refcode: GEHXEX). Further details of the polymorphs and crystal structures of these three compounds are given in the Supporting Information, along with full details of the computational methods and a diagrammatic workflow in Figure S2.

Experimental sublimation enthalpies were found for succinic acid and coronene, reported as 123.2 kJ/mol by Ribeiro da Silva et al.[27] and 148.2 kJ/mol by Chickos et al.,[28] respectively. A hydration free energy of −61.08 kJ/mol, measured by Rees and Wolfe,[29] was found for succinic acid. Where experimental values could not be found, in some cases, it was possible to estimate pseudo-experimental values using available data and standard thermodynamic relationships.

Calculation of Sublimation Free Energy Using CASTEP and the Model-Potential-Based Ψ_{mol} Method. For all three crystal structures, full DFT-D crystal structure optimizations were carried out with CASTEP[30] using the Perdew−Burke−Ernzerhof (PBE) functional and the Tkatchenko−Scheffler (TS)[31] dispersion correction scheme, with on-the-fly pseudopotentials. The input coordinates were the experimental structures (CORONE03, SUCACB03, GEHXEX)[32,33] in the Cambridge Structural Database with the C−H bond lengths corrected to neutron values.[34] The optimized crystal structures are in very good agreement with the experimental low-temperature structure determinations (Table S1 in the Supporting Information).

PBE-TS harmonic phonon calculations were performed using either linear response or a finite differencing algorithm with a supercell selected to ensure there were no imaginary frequencies across the phonon Brillouin zone. Once a phonon calculation was completed, the phonon Brillouin zone was

https://doi.org/10.1021/acs.jctc.1c00130
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

further sampled with a finer nuclear Brillouin zone grid and the resultant phonon density of states was integrated to obtain the thermodynamic corrections for the crystal, namely, zero-point energy (E^s_{ZPE}), internal energy (U^s_{corr}), and Helmholtz free energy (A^s_{corr}) of the crystal and the solid-state contribution to $T\Delta S_{sub}$ at 298.15 K. Details of these calculations and their results are given in Tables S2 and S3 of the Supporting Information. The phonon curve of desloratadine is in reasonable agreement with the room-temperature terahertz spectrum.[35]

The molecular conformations were extracted from the optimized crystal structure using NEIGHCRYS.[36] These were used to obtain molecular energy in its crystal conformation ($E_{mol\_in\_cryst}$) and distributed multipoles (DMA) by GDMA[37] analysis for use in the lattice energy calculations. Both the PBE/6-311++G(2d,p) and PBE0/6-31G(d,p) charge distributions were obtained using Gaussian 09[38] with and without a polarizable continuum model (PCM) model (ε = 3.0). The PCM calculations used default settings in Gaussian 09 and a relative dielectric constant of 3.0, typical for organic crystals.[39] The sensitivity of the results to these four model charge distributions is explored in Table S5 of the Supporting Information, which also shows that the periodic DFT-D lattice energies with the PBE functional are inadequate.

The molecular conformations were optimized using the PBE/6-311++G(2d,p) charge density within the PCM model (ε = 3.0) to the global minimum of the molecule to obtain E_{mol_min}, and the harmonic vibrational modes are calculated. In the cases of succinic acid and desloratadine, starting from the extracted conformations led to the closest local stationary points on the potential energy surfaces, which were planar and the AAA conformation, respectively. Further optimizations located the global minimum (gauche or SAA), with an energy of E_{mol_min}. The conformational energy difference between the molecule's lowest energy gas-phase conformation and the crystal conformation is obtained as the difference between E_{mol_min} and $E_{mol\_in\_cryst}$. The molecular vibrations were computed for each gas-phase molecular structure at the global minimum for each compound, with Gaussian 09[37] with each functional/basis set/PCM combination and "tight" convergence criteria. Thermal analysis by Gaussian 09[37] yielded E^g_{ZPE}, H^g_{corr}, and A^g_{corr} for the most stable isolated molecular conformation.

The PBE-TS-optimized crystal structure was reoptimized using DMACRYS[36] to obtain the intermolecular lattice energy, U_{inter}, keeping the molecule rigid. The lattice energy was evaluated using the distributed multipoles from the various molecular charge densities (Ψ_{mol}-approach[36]) combined with the FIT exp-6 intermolecular repulsion−dispersion pair potential, which has been parameterized by fitting to crystal structures and some heats of sublimation. The lattice energy is then obtained as

$$E_{latt} = U_{inter} + E_{mol\_in\_cryst} - E_{mol\_min} \quad (6)$$

Thermodynamic terms calculated above were then combined to obtain ΔG°_{sub} according to

$$\Delta G^\circ_{sub}(T) = \Delta A^\circ_{sub}(T) + RT$$
$$= -E_{latt} + A^g_{corr}(T) - A^s_{corr}(T) + RT \quad (7)$$

Calculation of Hydration Free Energy Using Implicit Continuum Models. Hydration free energies were calculated using the PBE,[40] PBE0,[41] and PBE0-DH[42] functionals with the 6-311++G(2d,p) basis set and the SMD solvent model.[43] All hydration calculations were carried out in Gaussian 16.[44] The PBE functional and basis set were chosen for consistency with the sublimation free-energy calculations, and PBE0 and PBE0-DH were included as potentially more accurate functionals. The SMD solvent model was selected because it performs well for organic molecules.[43]

Three different approaches were investigated to account for conformational degrees of freedom in the calculation of hydration free energy. In the first two approaches, the solute in the gas phase was modeled using the same single conformer as in the sublimation calculations, and the solute in the solution-phase was modeled using either a Boltzmann-weighted ensemble of conformers (SFE1) or a single global minimum energy solution-phase conformer (SFE2). Both of these methods allow for favorable cancellation of errors when the sublimation and hydration legs of the cycle use the same DFT methods. The third approach uses a Boltzmann-weighted ensemble of conformers in each phase separately (SFE3). Conformational searches were carried out using a force-field-based genetic algorithm in OpenBabel[45] before the low-energy conformers were reoptimized using DFT. Optimized structures were clustered to remove duplicates prior to Boltzmann weighting (see the Supporting Information for further details). Eight dominant solution-phase conformations have previously been identified for desloratadine,[35] which were used in these calculations. Due to the rigidity of coronene, no conformational search was needed, and a single conformer was used for calculations.

Calculation of Hydration Free Energy Using Molecular Dynamics Simulations and Free-Energy Perturbation Theory. For each solute, parameters from the general Amber force field (GAFF) with AM1-BCC charges were assigned using the ACPYPE server.[46] Molecular dynamics simulations were performed using Gromacs 2020.3.[47] A rhombic dodecahedron box with periodic boundary conditions was used. Water was represented using an SPC/E model,[48] and no counterions were added. All bonds involving hydrogen were kept rigid using the LINCS algorithm of the fourth order. Dynamics were simulated using a stochastic dynamics integrator, with a reference temperature of 298K. Neighbor searching was performed using a pair list generated by a Verlet cutoff scheme. Short-range interactions used the particle-mesh Ewald (PME) method,[49] with Lennard-Jones interactions switched off at 10 Å. Electrostatic interactions were treated using the PME method with a cutoff of 10 Å, a Fourier spacing of 1.2 Å, a fourth-order interpolation, and a tolerance of 10^{-4}.

Hydration free energy was computed using 21 values of the scaling factor λ, with Lennard-Jones and electrostatics interactions between the solute and solvent scaled together. Intramolecular interactions were kept the same at all λ values. Calculations were performed at 21 λ values at intervals of 0.05 from 0 to 1. Each simulation with its corresponding λ ran for 1300 ps during its production run. Prior to running a production MD simulation, 2500 steps of steepest descent optimization and a 50 ps equilibration were performed. A time step of 2 fs was used for each simulation. In both equilibration and production runs, the pressure was kept constant at 1 bar using the Parrinello−Rahman pressure coupling[50] and a compressibility of 4.5 × 10^{-5} bar^{-1}. After each simulation was complete, hydration free energy was evaluated using the Bennett acceptance ratio (BAR).[51]

https://doi.org/10.1021/acs.jctc.1c00130
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

Table 1. Lattice and Sublimation Energetics in kJ/mol Based on DMACRYS PBE/6-311++G(2d,p)/PCM Calculations and Thermal Corrections[a]

compound	E_{latt}	$\Delta H^{\circ calcd}_{sub}$	$\Delta H^{\circ expt}_{sub}$	ΔH°_{sub}	ΔG°_{sub}	$T\Delta H^{\circ}_{sub}$	$\Delta H^{\circ calcd}_{sub} + E_{latt}$
succinic acid	−125.89	121.04	123.2	49.06	51.54	69.50	−4.85
coronene	−155.61	143.51	148.2	76.57	79.05	64.46	−12.10
desloratadine	−144.40	133.72		57.29	59.77	73.95	−10.68

[a]The experimental $\Delta H^{\circ expt}_{sub}$ is given where available.[27,28]

Prediction of Intrinsic Aqueous Solubility Using Machine Learning Algorithms. Three machine learning models were implemented for predicting the intrinsic solubility of succinic acid, desloratadine, and coronene. These models used the Extra Trees,[52] Random Forest,[53] and Bagging[54] algorithms, each trained on a set of 117 druglike compounds with 173 CDK descriptors[55] used for each molecule and implemented as described in reference 56. The training set for these ML models does not include succinic acid, desloratadine, or coronene. These models were initially developed for an entry to the 2019 Solubility Challenge.[56]

RESULTS

Sublimation Free Energy. In Table 1, we present sublimation thermodynamics computed using the DMACRYS PBE/6-311++G(2d,p)/PCM Ψ_{mol}-based lattice energies and the thermal corrections computed from periodic PBE-TS phonons, as described in the Computational Methods section and elaborated on in the Supporting Information. The calculated sublimation enthalpies are in good agreement with the measured values for succinic acid ($\Delta H^{\circ expt}_{sub} - \Delta H^{\circ calcd}_{sub}$ = 2.16 kJ/mol) and coronene ($\Delta H^{\circ expt}_{sub} - \Delta H^{\circ calcd}_{sub}$ = 4.69 kJ/mol); no experimental sublimation data was available for desloratadine. These errors would correspond to 2.4- and 6.6-fold errors in solubility (based on eq 1), respectively, which is encouraging given that machine learning models commonly report 10-fold errors. The influence of using crystal phonon modes rather than the $2RT$ approximation to convert lattice energies into sublimation enthalpies can be assessed by considering the value of the term $\Delta H^{\circ}_{sub} + E_{latt}$ in Table 1. Although there is good agreement between $\Delta H^{\circ calcd}_{sub} + E_{latt}$ and −4.96 kJ/mol (=$-2RT$) for succinic acid, for coronene and desloratadine, the differences would correspond to more than 17- and 10-fold differences in solubility, respectively. Clearly, the method used to convert lattice energy into sublimation enthalpy has a large effect on predicted solubility.

Hydration Free Energy. Table 2 reports hydration free energies computed using atomistic MD/FEP simulations and three DFT methods (PBE/6-311++G(2d,p)/SMD, PBE0/6-311++G(2d,p)/SMD, PBE0-DH/6-311++G(2d,p)/SMD). For succinic acid and coronene, for which experimental or pseudo-experimental values are available, MD/FEP is the most accurate method for computing hydration free energy and gives relatively small errors ($\Delta G^{expt}_{hyd} - \Delta G^{calcd}_{hyd}$ of −3.61 and 1.6 kJ/mol, respectively). For the DFT methods, taking a Boltzmann-weighted average of multiple conformers, rather than a single minimum energy conformer in the gas and/or solution phase, had relatively little effect on the results, leading to small changes in ΔG_{hyd} for succinic acid in most cases (Table S8 in the Supporting Information). Slightly larger changes were observed for desloratadine when using the PBE0 or PBE0-DH functionals ($\Delta\Delta G_{hyd}$ < 2 kJ/mol), but there was no evidence that the Boltzmann-weighting scheme led to more accurate results overall. For that reason, Table 2 presents SMD results obtained by the SFE2 approach only. However, for all solutes, changing from PBE to PBE0 or PBE0-DH functionals led to a non-negligible change in hydration free energy. For succinic acid, PBE0-DH agrees reasonably well with experiment ($\Delta G^{expt}_{hyd} - \Delta G^{calcd}_{hyd}$ =−4.85 kJ/mol), whereas PBE does not ($\Delta G^{expt}_{hyd} - \Delta G^{calcd}_{hyd}$ = −11.75 kJ/mol). A similar trend is observed for coronene although neither PBE nor PBE0-DH gives satisfactory results. For desloratadine, the DFT and MD/FEP methodologies give self-consistent results, but there is no experimental data with which to compare them.

Table 2. Hydration Free Energies from Experiment and Computed from DFT or MD/FEP Simulations Using the SFE2 Approach[a]

compound	hydration model	ΔG^{*calcd}_{hyd} (kJ/mol)	ΔG^{*expt}_{hyd} (kJ/mol)
succinic acid	PBE/6-311++G(2d,p)/SMD	−49.33	−61.08
	PBE0/6-311++G(2d,p)/SMD	−52.78	
	PBE0-DH/6-311++G(2d,p)/SMD	−56.23	
	GAFF/AM1-BCC, SPC/E	−57.47	
coronene	PBE/6-311++G(2d,p)/SMD	−18.68	−38.40
	PBE0/6-311++G(2d,p)/SMD	−23.01	
	PBE0-DH/6-311++G(2d,p)/SMD	−26.32	
	GAFF/AM1-BCC, SPC/E	−40.00	
desloratadine	PBE/6-311++G(2d,p)/SMD	−45.11	
	PBE0/6-311++G(2d,p)/SMD	−48.08	
	PBE0-DH/6-311++G(2d,p)/SMD	−50.38	
	GAFF/AM1−BCC, SPC/E	−44.93	

[a]The experimental ΔG^{*expt}_{hyd} is given where available.[29] While we do not have a true experimental hydration free energy for coronene, we can infer its value if we assume that the experimental log S_0 and ΔH°_{sub} values[24,28] and the computed $T\Delta S^{\circ}_{sub}$ are correct. Rearranging eq 3 then leads to a back-calculated pseudo-experimental ΔG^{*}_{hyd} of −38.40 kJ/mol.

Machine Learning Intrinsic Solubility Predictions. Table 3 reports intrinsic solubility predictions from the Extra

Table 3. Predicted Log S_0 Values Derived from Machine Learning Extra Trees, Random Forest, and Bagging Algorithms

		log S_0^{calcd}		
compound	log S_0^{expt}	extra trees	random forest	bagging
succinic acid	−0.22	0.05	−1.00	−1.21
coronene	−9.33	−8.05	−7.35	−6.22
desloratadine	−3.42	−4.30	−4.20	−3.96

https://doi.org/10.1021/acs.jctc.1c00130
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

Table 4. Computed Physics-Based Log S_0 Values Derived from Hydration Free Energy Results Obtained by PBE/6-311++G(2d,p)/SMD, PBE0/6-311++G(2d,p)/SMD, and PBE0-DH/6-311++G(2d,p)/SMD Calculations and MD/FEP Simulations

compound	sublimation model	hydration model	log S_0^{calcd}	log S_0^{expt}	error
succinic acid	PBE/6-311++G(2d,p)/PCM	PBE/6-311++G(2d,p)/SMD	−1.78	−0.22	1.56
		PBE0/6-311++G(2d,p)/SMD	−1.17		0.95
		PBE0-DH/6-311++G(2d,p)/SMD	−0.57		0.35
		GAFF/AM1-BCC, SPC/E	−0.35		0.13
	extra trees		0.05		−0.27
coronene	PBE/6-311++G(2d,p)/PCM	PBE/6-311++G(2d,p)/SMD	−11.97	−9.33	2.64
		PBE0/6-311++G(2d,p)/SMD	−11.21		1.88
		PBE0-DH/6-311++G(2d,p)/SMD	−10.63		1.30
		GAFF/AM1-BCC, SPC/E	−8.23		−1.10
	extra trees		−8.05		−1.28
desloratadine	PBE/6-311++G(2d,p)/PCM	PBE/6-311++G(2d,p)/SMD	−3.96	−3.42	0.54
		PBE0/6-311++G(2d,p)/SMD	−3.44		0.02
		PBE0-DH/6-311++G(2d,p)/SMD	−3.03		−0.39
		GAFF/AM1-BCC, SPC/E	−3.99		0.57
	extra trees		−4.30		0.88

Trees, Random Forest, and Bagging algorithms. The Extra Trees model performed better than the Random Forest and Bagging models for our data set and was chosen as the benchmark for the physics-based models. The same model was submitted to the 2019 Solubility Challenge by one of us and known for the purposes of that Challenge as JMSA_B.[53,57,58]

Physics-Based Intrinsic Solubility Predictions. The computed ΔG°_{sub} and ΔG^{*}_{hyd} combine according to eq 3 to give the log S_0 values reported in Table 4. The computed log S_0 values are based on DMACRYS PBE/6-311++G(2d,p)/PCM calculations with thermal corrections as the sublimation method and PBE/6-311++G(2d,p)/SMD, PBE0/6-311++G-(2d,p)/SMD, PBE0-DH/6-311++G(2d,p)/SMD, or MD/FEP as the hydration method.

For succinic acid, PBE/6-311++G(2d,p)/SMD underestimated the magnitude of the hydration free energy, which resulted in an underestimation of the solubility of 1.56 log S_0 units. Replacing PBE by the PBE0-DH functional improved the calculated hydration free energy and log S_0 to within −4.85 kJ/mol and 0.35 log units of their experimental results, respectively. The MD/FEP calculations gave the most accurate hydration free energy and as a result predicted log S_0 = −0.35, only 0.13 log units from the experimental value.

For coronene, using the PBE/6-311++G(2d,p)/SMD hydration model again leads to underestimation of the magnitude of the hydration free energy, with an error of 19.76 kJ/mol as compared to the back-calculated pseudo-experimental ΔG^{*}_{hyd}. This results in a prediction of log S_0 lower than the experimental[24] value by 2.64 units. (The derivation of our back-calculated pseudo-experimental ΔG^{*}_{hyd} is described in the footnote of Table 2.) Using instead the PBE0-DH/6-311++G(2d,p)/SMD hydration model gives a more accurate value of ΔG^{calcd}_{hyd} and leads to a predicted log S_0 of only 1.30 log units below the experimental value, as shown in Table 4. The most accurate hydration free energy was obtained from the MD/FEP simulations, which overestimated the magnitude of ΔG^{calcd}_{hyd} by only 1.6 kJ/mol compared to the pseudo-experimental value, and resulted in a prediction of solubility within 1.10 log units of the experimental value.

The PBE/6-311++G(2d,p)/SMD hydration model appears to perform better for desloratadine than for the other solutes and gives a calculated log S_0 within 0.54 log S_0 units of the experimental value.[25] Since we have no experimental sublimation or hydration thermodynamics data, however, it is unclear whether our predictions of ΔG°_{sub} and ΔG^{*}_{hyd} are both accurate or whether we are relying on a cancellation of errors to arrive at an accurate log S_0 prediction. The underestimation of solubility suggests that the PBE/6-311++G(2d,p)/SMD hydration model underestimates the magnitude of the hydration free energy, which would be in keeping with the trend observed for coronene and succinic acid, but cannot be independently validated with the available experimental data. Using the PBE0 or PBE0-DH functionals rather than the PBE functional leads to more accurate estimates of solubility, within 0.02 log units and 0.39 log units of the experimental value, respectively. Using the solvation free energy computed by MD/FEP simulations gives a predicted solubility that is almost identical to that obtained using the PBE/6-311++G(2d,p)/SMD model, with an error compared to experiment of 0.57 log units. For desloratadine, all three solvent models give predictions with errors <0.6 log S_0 units, which compares favorably with the Extra Trees regressor that gives an error of 0.88 log S_0 units.

The absolute error in log S_0 for each model is summarized in Figure 3. The Extra Trees results (yellow bars) provide an

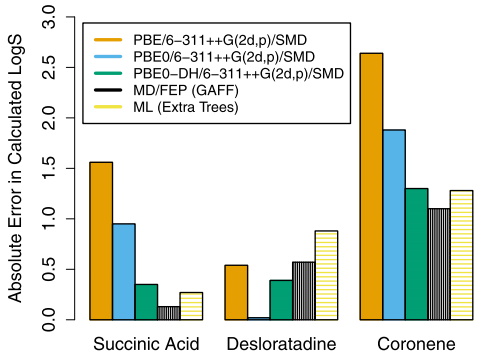


Figure 3. Absolute error in calculated log S_0 for succinic acid, desloratadine, and coronene. The physics-based predictions of solubility use the PBE/6-311++G(2d,p)/PCM sublimation free energies and the PBE/6-311++G(2d,p)/SMD (orange), PBE0/6-311++G(2d,p)/SMD (blue), PBE0-DH/6-311++G(2d,p)/SMD (green), or MD/FEP (black) hydration free energies. The machine learning predictions use the Extra Trees algorithm (yellow).

https://doi.org/10.1021/acs.jctc.1c00130
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

example of a state-of-the-art machine learning model against which the physics-based models have been evaluated. Using the DMACRYS PBE/6-311++G(2d,p)/PCM Ψ_{mol}-based sublimation method, and either PBE0-DH/6-311++G(2d,p)/SMD or MD/FEP hydration calculations, we obtained predicted solubilities that rival the accuracy of the Extra Trees model.

DISCUSSION

Our previous physics-based solubility prediction work[13] gave an RMSE of 1.45 log S_0 units over 25 druglike compounds while incorporating the 2RT approximation. Although that was a promising result, further improvements were required to match the predictive accuracy of machine learning. The present work achieves this aim by improving upon the 2RT approximation for enthalpies of sublimation by utilizing a full analysis of the vibrational and phonon contributions to the sublimation enthalpy and entropy. Succinic acid and desloratadine adopt different conformations in the solid from the isolated molecule at low temperatures and the ensemble in the liquid. Fortunately, our results suggest that the hydration results are rather insensitive to conformational averaging, and so the conformational search could be limited, provided that the global minimum (gas-phase) conformation and the crystal conformation are known.

We have used the experimental crystal structures in these calculations, but this could be obtained from a crystal structure prediction study.[57] Indeed, we envisage that these physics-based solubility calculations would be performed alongside such a crystal structure prediction study as these are now becoming more routine in the industry[58] and being developed to be used during early drug development[59] at the solid form selection stage and as a complement to solid form screening.[60,61] The key advantage of a physics-based approach over informatics would then be realized, by being able to adapt the calculations to different solvents, polymorphs, and temperatures.

The linking of the solubility calculations into a workflow involving crystal structure prediction, which includes determining the range of conformations that can occur in solid state, means that the development of this approach to solubility prediction can be closely coupled to the current work on improving the calculation of free energies of polymorphs. Absolute lattice energies calculated using periodic DFT-D and currently affordable functionals like PBE are known to be poor,[62] but the progress in developing reliable calculations of relative energies of polymorphs[63] and sublimation pressures[64] suggests that a fully quantum-mechanical prediction of the solid-state contributions[14,65] could provide accurate solubilities. This may need to be coupled with the use of higher-level calculations on the isolated molecule, as this has been found to provide a major improvement in CSP results in certain cases of conformational polymorphism.[66,67] However, methods of mitigating the expense of the phonon calculations, which appear necessary given the inadequacy of the 2RT approximation, are being developed.[68]

Alongside developing absolute solubility calculations, it is also to estimate the solubility difference between polymorphs or between racemic and enantiopure crystals. The degree of cancellation of errors is very specific to the crystals involved[20] and needs to be highly accurate as the average difference in molar solubility between polymorphs has been estimated to be approximately 2-fold,[69] which is 4−5 times smaller than the average error in solubility models. One outcome of our study is that care has to be taken to ensure cancellation of errors when calculating absolute sublimation free energies, i.e., at this stage, it is more accurate to use consistent electronic structure methods than the best affordable for each phase. It also appears that the hydration energies improve with the electronic structure method used.

In this study, we have chosen three diverse molecules spanning a wide range of solubilities and the results are extremely encouraging. For all three solutes, the implicit solvation model improves with the quality of the molecular charge distribution and is relatively insensitive to the treatment of the conformational flexibility. The explicit solvation model using molecular dynamics simulations provides very worthwhile results, which are capable of reflecting the effects of long-lived, specific hydrogen bonding of solvent to solute, though such extended residence times do not occur in these three systems.[34] These calculations will depend critically on the quality of the force field, as do many other molecular dynamics-based methods.[70]

Physics-based solubility approaches including the one presented here are typically tested on only a handful of compounds at best. In this case, the three compounds chosen present different types of chemistries, conformational flexibilities, and solubilities. Critically, the set includes desloratadine as a more typical pharmaceutical, showing that the methodology can be applied to larger, flexible molecules than are typically used to validate physics-based methods. In comparison, machine learning and QSPR models are typically validated on tens-to-hundreds of compounds, the two solubility challenges each having a 100-compound test set.[58,71] A major limitation of informatics approaches is that they can only be applied to properties for which training data for sufficient compounds has been measured. Physics-based approaches have the potential to be modified for different solvent mixtures,[43,72,73] temperatures,[74,75] and other properties,[76] vastly extending the possible contribution of digital design to crystallization processes. Following the proof-of-concept results presented here, the validation of physics-based solubility methods on a larger range of molecules is a priority to drive progress in this field.

CONCLUSIONS

The physics-based method presented within this work shows that intrinsic aqueous solubility can be predicted with reasonable accuracy, rivaling current cheminformatics and machine learning approaches. Throughout this process, a full computational description of each thermodynamic stage of transferring a molecule from crystal to gas to solution is produced. Further progress can be made, however, including systematic improvements to the sublimation and hydration free-energy models, as well as more rigorous testing on a larger data set of druglike molecules.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.1c00130.

Full specification of the theoretical derivations, methodology, and more detailed results including; the workflow developed in this work for physics-based computation of sublimation and hydration (Figure S1), variations in

https://doi.org/10.1021/acs.jctc.1c00130
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

solid and gaseous energies with functional (Tables S5 and S6); tests of hydration free energy methods using the Minnesota Solvation Database (Figures S2 and S3); variation of calculated hydration energies with treatment of conformational variations (Table S8); calculated solubilities from all combinations of methods (Tables S9 and S10) (PDF)

AUTHOR INFORMATION

Corresponding Authors

David S. Palmer − *Department of Pure and Applied Chemistry, University of Strathclyde, Glasgow, Scotland G1 1XL, U.K.*; orcid.org/0000-0003-4356-9144; Email: david.palmer@strath.ac.uk

John B. O. Mitchell − *EaStCHEM School of Chemistry and Biomedical Sciences Research Complex, University of St Andrews, St Andrews, Scotland KY16 9ST, U.K.*; orcid.org/0000-0002-0379-6097; Email: jbom@st-andrews.ac.uk

Authors

Daniel J. Fowles − *Department of Pure and Applied Chemistry, University of Strathclyde, Glasgow, Scotland G1 1XL, U.K.*

Rui Guo − *Department of Chemistry, University College London, London WC1H 0AJ, U.K.*

Sarah L. Price − *Department of Chemistry, University College London, London WC1H 0AJ, U.K.*; orcid.org/0000-0002-1230-7427

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.1c00130

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

D.S.P. and D.J.F. thank the EPSRC for funding via Prosperity Partnership EP/S035990/1. D.S.P. and D.J.F. thank the ARCHIE-WeSt High-Performance Computing Centre (www.archie-west.ac.uk) for computational resources. The UCL authors thank Prof. Keith Refson for guidance with the phonon calculations, which used the ARCHER U.K. National Supercomputing Service (http://www.archer.ac.uk) as part of the U.K. HEC Materials Chemistry Consortium, which is funded by the EPSRC (EP/L000202, EP/R029431). R.G. was funded by MagnaPharm, a project funded by the European Union's Horizon 2020 Research and Innovation programme under grant agreement number 736899.

REFERENCES

(1) Di, L.; Kerns, E. H.; Carter, G. T. Drug-Like Property Concepts in Pharmaceutical Design. *Curr. Pharm. Des.* **2009**, *15*, 2184−2194.

(2) Williams, H. D.; Trevaskis, N. L.; Charman, S. A.; Shanker, R. M.; Charman, W. N.; Pouton, C. W.; Porter, C. J. Strategies to address low drug solubility in discovery and development. *Pharmacol. Rev.* **2013**, *65*, 315−499.

(3) Li, L.; Totton, T.; Frenkel, D. Computational methodology for solubility prediction: Application to the sparingly soluble solutes. *J. Chem. Phys.* **2017**, *146*, No. 214110.

(4) Li, L.; Totton, T.; Frenkel, D. Computational methodology for solubility prediction: sparingly soluble organic/inorganic materials. *J. Chem. Phys.* **2018**, *149*, No. 054102.

(5) Kolafa, J. Solubility of NaCl in water and its melting point by molecular dynamics in the slab geometry and a new BK3-compatible force field. *J. Chem. Phys.* **2016**, *145*, No. 204509.

(6) Boothroyd, S.; Kerridge, A.; Broo, A.; Buttar, D.; Anwar, J. Solubility prediction from first principles: a density of states approach. *Phys. Chem. Chem. Phys.* **2018**, *20*, 20981−20987.

(7) Boothroyd, S.; Anwar, J. Solubility prediction for a soluble organic molecule via chemical potentials from density of states. *J. Chem. Phys.* **2019**, *151*, No. 184113.

(8) Westergren, J.; Lindfors, L.; Höglund, T.; Lüder, K.; Nordholm, S.; Kjellander, R. In Silico Prediction of Drug Solubility: 1. Free Energy of Hydration. *J. Phys. Chem. B* **2007**, *111*, 1872−1882.

(9) Lüder, K.; Lindfors, L.; Westergren, J.; Nordholm, S.; Kjellander, R. In Silico Prediction of Drug Solubility: 2. Free Energy of Solvation in Pure Melts. *J. Phys. Chem. B* **2007**, *111*, 1883−1892.

(10) Lüder, K.; Lindfors, L.; Westergren, J.; Nordholm, S.; Kjellander, R. In Silico Prediction of Drug Solubility. 3. Free Energy of Solvation in Pure Amorphous Matter. *J. Phys. Chem. B* **2007**, *111*, 7303−7311.

(11) Lüder, K.; Lindfors, L.; Westergren, J.; Nordholm, S.; Persson, R.; Pedersen, M. In silico prediction of drug solubility: 4. Will simple potentials suffice? *J. Comput. Chem.* **2009**, *30*, 1859−1871.

(12) Mondal, S.; Tresadern, G.; Greenwood, J.; Kim, B.; Kaus, J.; Wirtala, M.; Steinbrecher, T.; Wang, L.; Masse, C.; Farid, R.; Abel, R. A free energy perturbation approach to estimate the intrinsic solubilities of drug-like small molecules. *ChemRxiv* **2019**, No. 263077.

(13) Palmer, D. S.; McDonagh, J. L.; Mitchell, J. B. O.; van Mourik, T.; Fedorov, M. V. First-principles calculation of the intrinsic aqueous solubility of crystalline druglike molecules. *J. Chem. Theory Comput.* **2012**, *8*, 3322−3337.

(14) Abramov, Y. A.; Sun, G.; Zeng, Q.; Zeng, Q.; Yang, M. Guiding Lead Optimization for Solubility Improvement with Physics-based Modeling. *Mol. Pharmaceutics* **2020**, *17*, 666−673.

(15) Yalkowsky, S. H. *Solubility and Solubilization in Aqueous Media*; Oxford University Press: New York, 1999.

(16) Avdeef, A. *Absorption and Drug Development: Solubility, Permeability, and Charge State*; Wiley-Interscience: Hoboken, NJ, 2003.

(17) Ben-Naim, A. Standard thermodynamics of transfer. Uses and misuses. *J. Phys. Chem. A* **1978**, *82*, 792−803.

(18) Ben-Naim, A.; Marcus, Y. Solvation thermodynamics of nonionic solutes. *J. Chem. Phys.* **1984**, *81*, 2016−2027.

(19) Gavezzotti, A. *Theoretical Aspects and Computer Modeling of the Molecular Solid State*; John Wiley: Chichester, 1997.

(20) Buchholz, H. K.; Hylton, R. K.; Brandenburg, J. G.; Seidel-Morgenstern, A.; Lorenz, H.; Stein, M.; Price, S. L. Thermochemistry of Racemic and Enantiopure Organic Crystals for Predicting Enantiomer Separation. *Cryst. Growth Des.* **2017**, *17*, 4676−4686.

(21) Palmer, D. S.; O'Boyle, N. M.; Glen, R. C.; Mitchell, J. B. O. Random Forest Models to Predict Aqueous Solubility. *J. Chem. Inf. Model.* **2007**, *47*, 150−158.

(22) Palmer, D. S.; Mitchell, J. B. O. Is Experimental Data Quality the Limiting Factor in Predicting the Aqueous Solubility of Druglike Molecules? *Mol. Pharmaceutics* **2014**, *11*, 2962−2972.

(23) Forbes, G. S.; Coolidge, A. S. Relations between distribution ratio, temperature, and concentration in system: water, ether, succinic acid. *J. Am. Chem. Soc.* **1919**, *41*, 150−167.

(24) Miller, M. M.; Wasik, S. P.; Huang, G. L.; Mackay, D. Relationships between octanol−water partition coefficient and aqueous solubility. *Environ. Sci. Technol.* **1985**, *19*, 522−529.

(25) Popović, G.; Cakar, M.; Agbaba, D. Acid−base equilibria and solubility of loratadine and desloratadine in water and micellar media. *J. Pharm. Biomed. Anal.* **2009**, *49*, 42−47.

(26) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.* **2016**, *72*, 171−179.

(27) Ribeiro da Silva, M. A.; Monte, M. J.; Ribeiro, J. R. Thermodynamic study on the sublimation of succinic acid and of

methyl-and dimethyl-substituted succinic and glutaric acids. *J. Chem. Thermodyn.* **2001**, *33*, 23−31.
(28) Chickos, J. S.; Webb, P.; Nichols, G.; Kiyobayashi, T.; Cheng, P. C.; Scott, L. The enthalpy of vaporization and sublimation of corannulene, coronene, and perylene at T = 298.15 K. *J. Chem. Thermodyn.* **2002**, *34*, 1195−1206.
(29) Rees, D. C.; Wolfe, G. M. Macromolecular solvation energies derived from small molecule crystal morphology. *Protein Sci.* **1993**, *2*, 1882−1889.
(30) Clark, S. J.; Segall, M. D.; Pickard, C. J.; Hasnip, P. J.; Probert, M. J.; Refson, K.; Payne, M. C. First principles methods using CASTEP. *Z. Kristallogr. - Cryst. Mater.* **2005**, *220*, 567−570.
(31) Tkatchenko, A.; Scheffler, M. Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data. *Phys. Rev. Lett.* **2009**, *102*, No. 073005.
(32) Potticary, J.; Terry, L. R.; Bell, C.; Papanikolopoulos, A. N.; Christianen, P. C. M.; Engelkamp, H.; Collins, A. M.; Fontanesi, C.; Kociok-Kohn, G.; Crampin, S.; Da Como, E.; Hall, S. R. An unforeseen polymorph of coronene by the application of magnetic fields during crystal growth. *Nat. Commun.* **2016**, *7*, No. 11555.
(33) Potticary, J.; Boston, R.; Vella-Zarb, L.; Few, A.; Bell, C.; Hall, S. R. Low temperature magneto-morphological characterisation of coronene and the resolution of previously observed unexplained phenomena. *Sci. Rep.* **2016**, *6*, No. 38696.
(34) Lucaioli, P.; Nauha, E.; Gimondi, I.; Price, L. S.; Guo, R.; Iuzzolino, L.; Singh, I.; Salvalaglio, M.; Price, S. L.; Blagden, N. Serendipitous isolation of a disappearing conformational polymorph of succinic acid challenges computational polymorph prediction. *CrystEngComm* **2018**, *20*, 3971−3977.
(35) Srirambhatla, V. K.; Guo, R.; Dawson, D. M.; Price, S. L.; Florence, A. J. Reversible, Two-Step Single-Crystal to Single-Crystal Phase Transitions between Desloratadine Forms I, II, and III. *Cryst. Growth Des.* **2020**, *20*, 1800−1810.
(36) Price, S. L.; Leslie, M.; Welch, G. W. A.; Habgood, M.; Price, L. S.; Karamertzanis, P. G.; Day, G. M. Modelling Organic Crystal Structures using Distributed Multipole and Polarizability-Based Model Intermolecular Potentials. *Phys. Chem. Chem. Phys.* **2010**, *12*, 8478−8490.
(37) Stone, A. J. *GDMA: A Program for Performing Distributed Multipole Analysis of Wave Functions Calculated Using the Gaussian Program System*, GDMA2.2; University of Cambridge Cambridge: United Kingdom, 2010.
(38) Frisch, M. J. et al. *Gaussian 09*, revision D.01; Gaussian, Inc.: Wallingford, CT, 2009.
(39) Reilly, A. M.; et al. Report on the sixth blind test of organic crystal structure prediction methods. *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.* **2016**, *B72*, 439−459.
(40) Perdew, J. P.; Burke, K.; Ernzerhof, M. Errata: Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1997**, *78*, No. 1396.
(41) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **1999**, *110*, 6158−6170.
(42) Brémond, E.; Adamo, C. Seeking for parameter-free double-hybrid functionals: The PBE0-DH model. *J. Chem. Phys.* **2011**, *135*, No. 024106.
(43) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **2009**, *113*, 6378−6396.
(44) Frisch, M. J. et al. *Gaussian 16*, revision C.01; Gaussian, Inc.: Wallingford, CT, 2016.
(45) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, No. 33.
(46) Sousa da Silva, A. W.; Vranken, W. F. ACPYPE—AnteChamber PYthon Parser interfacE. *BMC Res. Notes* **2012**, *5*, No. 367.
(47) Abraham, M. J.; Murtola, T.; Schulz, R.; Pall, S.; Smith, J. C.; Hess, B.; Lindah, E. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1−2*, 19−25.
(48) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. The Missing in Effective Pair Potentials. *J. Phys. Chem. B* **1987**, *91*, 6269−6271.
(49) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An Nlog(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089−10092.
(50) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: a new molecular dynamics method. *J. Appl. Phys.* **1982**, *52*, 7182−7190.
(51) Bennett, C. H. Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* **1976**, *22*, 245−268.
(52) Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3−42.
(53) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.
(54) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947−1958.
(55) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An open-source Java library for chemo-and bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493−500.
(56) Llinàs, A.; Avdeef, A. Solubility challenge revisited after ten years, with multilab shake-flask data, using tight (SD ∼ 0.17 log) and loose (SD ∼ 0.62 log) test sets. *J. Chem. Inf. Model.* **2019**, *59*, 3036−3040.
(57) Price, S. L. Is zeroth order crystal structure prediction (CSP_0) coming to maturity? What should we aim for in an ideal crystal structure prediction code? *Faraday Discuss.* **2018**, *211*, 9−30.
(58) Nyman, J.; Reutzel-Edens, S. M. Crystal structure prediction is changing from basic science to applied technology. *Faraday Discuss.* **2018**, *211*, 459−476.
(59) Broo, A.; Lill, S. Transferable force field for crystal structure predictions, investigation of performance and exploration of different rescoring strategies using DFT-D methods. *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.* **2016**, *72*, 460−476.
(60) Bhardwaj, R.; McMahon, J.; Nyman, J.; Price, L. S.; Konar, S.; Oswald, I.; Pulham, C. R.; Price, S. L.; Reutzel-Edens, S. M. A prolific solvate former, Galunisertib, under the pressure of crystal structure prediction, produces ten diverse polymorphs. *J. Am. Chem. Soc.* **2019**, *141*, 13887−13897.
(61) Price, S. L.; Reutzel-Edens, S. M. The potential of computed crystal energy landscapes to aid solid-form development. *Drug Discovery Today* **2016**, *21*, 912−923.
(62) Geatches, D.; Rosbottom, I.; Robinson, R. L. M.; Byrne, P.; Hasnip, P.; Probert, M. I. J.; Jochym, D.; Maloney, A.; Roberts, K. J. Off-the-shelf DFT-DISPersion methods: Are they now "on-trend" for organic molecular crystals? *J. Chem. Phys.* **2019**, *151*, No. 044106.
(63) Hoja, J.; Ko, H.-Y.; Neumann, M. A.; Car, R.; DiStasio, R. A.; Tkatchenko, A. Reliable and practical computational description of molecular crystal polymorphs. *Sci. Adv.* **2019**, *5*, No. eaau-3338.
(64) Červinka, C.; Beran, G. J. O. Towards reliable ab initio sublimation pressures for organic molecular crystals—are we there yet? *Phys. Chem. Chem. Phys.* **2019**, *21*, 14799−14810.
(65) Abramov, Y. A. Major Source of Error in QSPR Prediction of Intrinsic Thermodynamic Solubility of Drugs: Solid vs Nonsolid State Contributions? *Mol. Pharmaceutics* **2015**, *12*, 2126−2141.
(66) Greenwell, C.; Beran, G. Inaccurate Conformational Energies Still Hinder Crystal Structure Prediction in Flexible Organic Molecules. *Cryst. Growth Des.* **2020**, *20*, 4875−4881.
(67) Greenwell, C.; McKinley, J.; Zhang, P.; Zeng, Q.; Sun, G.; Li, B.; Wen, S.; Beran, G. Overcoming the difficulties of predicting conformational polymorph energetics in molecular crystals via correlated wavefunction methods. *Chem. Sci.* **2020**, *11*, 2200−2214.

https://doi.org/10.1021/acs.jctc.1c00130
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

(68) Cook, C.; Beran, G. Reduced-cost supercell approach for computing accurate phonon density of states in organic crystals. *J. Chem. Phys.* **2020**, *153*, No. 224105.
(69) Pudipeddi, M.; Serajuddin, A. T. M. Trends in Solubility of Polymorphs. *J. Pharm. Sci.* **2005**, *94*, 929−939.
(70) Khanna, V.; Monroe, J.; Doherty, M.; Peters, B. Performing solvation free energy calculations in LAMMPS using the decoupling approach. *J. Comput. Aided Mol. Des.* **2020**, *34*, 641−646.
(71) Hopfinger, A. J.; Esposito, E. X.; Llinàs, A.; Glen, R. C.; Goodman, J. M. Findings of the Challenge to Predict Aqueous Solubility. *J. Chem. Inf. Model.* **2008**, *49*, 1−5.
(72) Misin, M.; Vainikka, P.; Fedorov, M. V.; Palmer, D. S. Salting-out effects by Pressure-Corrected 3D RISM. *J. Chem. Phys.* **2016**, *145*, No. 194501.
(73) Misin, M.; Fedorov, M. V.; Palmer, D. S. Predicting solvation free energies using parameter-free solvent models. *J. Phys. Chem. B* **2016**, *120*, 5724−5731.
(74) Chamberlin, A. C.; Cramer, C. J.; Truhlar, D. G. Predicting aqueous free energies of solvation as functions of temperature. *J. Phys. Chem. B* **2006**, *110*, 5665−5675.
(75) Misin, M.; Fedorov, M. V.; Palmer, D. S. Accurate Hydration Free Energies at a wide range of temperatures from 3D RISM. *J. Chem. Phys.* **2015**, *142*, No. 091105.
(76) Misin, M.; Fedorov, M. V.; Palmer, D. S. Hydration Free Energies of Molecular Ions from Theory and Simulation. *J. Phys. Chem. B* **2016**, *120*, 975−983.

https://doi.org/10.1021/acs.jctc.1c00130
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX