Challenges of the Use of Atomistic Simulations to Predict Solubilities of Drug-like Molecules

Guilherme Duarte Ramos Matos, David Mobley

Submitted date: 09/05/2018 • Posted date: 10/05/2018
Licence: CC BY 4.0

Citation information: Duarte Ramos Matos, Guilherme; Mobley, David (2018): Challenges of the Use of Atomistic Simulations to Predict Solubilities of Drug-like Molecules. ChemRxiv. Preprint.

Background. Solubility is a physical property of extreme importance to the Pharmaceutical industry whose prediction for potential drugs has so far been a hard task. We attempted to predict the solubility of acetylsalicylic acid (ASA) by estimating absolute chemical potentials of its most stable polymorph and of solutions with different concentrations of the drug molecule.

Methods. Chemical potentials were estimated from all-atom molecular dynamics simulations. We used the Einstein Molecule Method to predict the absolute chemical potential of the solid and solvation free energy calculations to predict the excess chemical potentials of the liquid phase systems.

Results. Reliable estimations of the chemical potentials for the solid and for a single ASA molecule using the Einstein Molecule Method required an extremely large number of intermediate states for the free energy calculations, meaning that the calculations were extremely demanding computationally. Despite the computational cost, however, the computed value did not agree well with experiment, potentially due to limitations with the underlying energy model. Perhaps better values could be obtained with a better energy model; however, it seems likely computational cost may remain a limiting factor for use of this particular approach to solubility estimation.

Conclusions. Solubility prediction of drug-like solids still is a challenge on the computational side, and it appears that both the underlying energy model and the computational approach applied may need improvement before the approach is suitable for routine use.

File list (3)

solubility.pdf (1.79 MiB)  view on ChemRxiv  download file
SupportingInformation.pdf (150.87 KiB)  view on ChemRxiv  download file
Supporting_Information.tar.gz (233.33 KiB)  view on ChemRxiv  download file
Challenges of the Use of Atomistic Simulations to Predict Solubilities of Drug-like Molecules

Guilherme Duarte Ramos Matos\textsuperscript{1} and David L. Mobley\textsuperscript{2}

\textsuperscript{1}Department of Chemistry, University of California, Irvine
\textsuperscript{2}Departments of Pharmaceutical Sciences and Chemistry, University of California, Irvine

Abstract

Background. Solubility is a physical property of extreme importance to the Pharmaceutical industry whose prediction for potential drugs has so far been a hard task. We attempted to predict the solubility of acetylsalicylic acid (ASA) by estimating absolute chemical potentials of its most stable polymorph and of solutions with different concentrations of the drug molecule.

Methods. Chemical potentials were estimated from all-atom molecular dynamics simulations. We used the Einstein Molecule Method to predict the absolute chemical potential of the solid and solvation free energy calculations to predict the excess chemical potentials of the liquid phase systems.

Results. Reliable estimations of the chemical potentials for the solid and for a single ASA molecule using the Einstein Molecule Method required an extremely large number of intermediate states for the free energy calculations, meaning that the calculations were extremely demanding computationally. Despite the computational cost, however, the computed value did not agree well with experiment, potentially due to limitations with the underlying energy model. Perhaps better values could be obtained with a better energy model; however, it seems likely computational cost may remain a limiting factor for use of this particular approach to solubility estimation.

Conclusions. Solubility prediction of drug-like solids still is a challenge on the computational side, and it appears that both the underlying energy model and the computational approach applied may need improvement before the approach is suitable for routine use.

Keywords

Solubility, molecular crystals, free energy calculations, chemical potentials, solvation
Introduction

Solubility is a critical property for pharmaceutical drug discovery, and problems with solubility can frustrate drug discovery efforts and block treatments. The bioavailability of a drug depends on the solubility difference between different crystal structures (polymorphs), dose, drug permeability, and formulation [1], so solubility plays a key role. Solubility problems can be unexpected and pose crucial obstacles that even threaten the administration of care. For example, a well-documented case occurred in the late 90’s when ritonavir, an HIV-protease inhibitor marketed as Norvir, failed dissolution requirements [2]. Since ritonavir is not bioavailable in its solid form, it was administrated in capsules containing solutions designed not to be saturated with respect to the originally known molecular crystal (form I) [2]. The newly identified polymorph, form II, was unusually stable and unusually hard to crystallize; the preparation protocol of Norvir was inadequate to make capsules from the new polymorph, which severely threatened the supply of the drug in the market and endangered the lives of many HIV+ patients [2]. Considerable effort has already been devoted to the methods to predict crystal polymorphs [3, 4, 5, 6, 7, 8, 9], but much less attention has been given to methods to predict solubilities, with or without likely polymorphs as input.

Due to the importance of aqueous solubilities in different industrial processes and environmental applications, a scientific challenge consisting of the prediction of 32 solubilities given a database of 100 reliable measurements [10, 11] was created with the goal of comparing the outcomes of different solubility prediction techniques. Participants employed methods such as artificial neural networks [12], quantitative structure-property relationship (QSPR) [13], and deep learning [14] to predict the aqueous solubilities of drug-like molecules. All of the employed methods were empirical and trained on existing measurements. The limitation of these methods, however, is the dependence of a training set of data that limits their applicability to compounds similar to those in the training of the set and impairs its transferability.

Some newer methods attempt to predict solubilities based on a physical description of the interactions in solution and in the solid state, yielding results that are in principle rigorous given an accurate energy model and an adequate method. In these approaches, molecular systems are described by force fields, i.e., potential energy functions that contain parameters describing bonds, atoms, electrostatic and non-electrostatic interactions. Molecular dynamics (MD) or Monte Carlo (MC) simulations are commonly used to sample different configurations of the system described by an energy model called a force field. The simulations then allow the estimation of physical properties such as internal energy, free energy, and enthalpy under different conditions. The quality of the results of such methods depends on how well the force field describes the system under study and how good the sampling method is. Thus, some researchers have recently estimated aqueous solubilities using simulations of thermodynamic cycles encompassing the crystal, the ideal gas, and an infinitely dilute solution of a given molecule [15, 16]. When the structure of the solid is unknown, some studies have substituted simulations of solid melts in place of a structure of the solid [17, 18, 19, 20].

While these physical methods for predicting solubilities have received some attention in the literature, most are still in their infancy with only a handful studies applying them and it is not yet clear how broadly applicable they will be [17, 18, 19, 20], and others have only been suggested or demonstrated in proof-of-principle tests [21, 22, 16, 23]. Our view is that the time is ripe for physical methods to predict solubility, especially given the routine nature of solvation free energy calculations at present [24, 25, 26, 27, 28, 29] which comprise essentially half of the solubility problem (see the Theory section). Polymorph and crystal structure prediction successes also mean that we may often have a suitable crystal structure of the compound as input [3, 30, 31, 32, 33, 4, 34, 5, 8, 9, 35], so what remains is to predict the solubility given a crystal structure and simulations of the relevant phases.

Here, we focus on adapting and testing an existing approach for solubility prediction in the hope that it will prove to be a generally applicable method for solubility prediction that can be applied routinely. This method uses all-atom molecular dynamics simulations to estimate absolute chemical potentials and predict aqueous solubilities of molecular solids, given the crystal structure (or an estimate thereof) as input.

Theory

The solubility of a molecular solid is related to the chemical potentials of each phase

Solubility is defined as the maximum concentration of solute that can be dissolved in a selected bulk solvent. Chemical potentials ($\mu$) of the solid-state solute and the solution are by definition equal at the solubility point, when the solution is in equilibrium with the solid.

$$\mu_{solute}^{solid} = \mu_{solute}^{solution}$$

Solid particles precipitate in concentrations higher than the solubility point because the solid phase becomes more stable in these conditions. In principle, we can predict at which concentration a molecule precipitates in solution if we calculate the chemical potentials of the components:

$$\mu_i = \left( \frac{\partial A}{\partial N_i} \right)_{V,T,N_\neq i} = \left( \frac{\partial G}{\partial N_i} \right)_{P,T,N_\neq i}$$

where $\mu_i$ is the chemical potential of component $i$; $A$ is the Helmholtz free energy; $G$ is the Gibbs free energy; $N_i, i \neq i$ is the number of molecules of each component in the mixture; $V$ is the volume of the system; $T$ its temperature; and $P$ its pressure. Calculations from systems under constant $V$ and $T$ yield $A$; $G$ is obtained from simulations under constant $P$ and $T$ conditions. In order to estimate the chemical potential of one component in solution and in its molecular solid, however, we need to know the absolute free energy of the system in these states. We calcul-
lated absolute free energies using alchemical free energy calculations.

Alchemical free energy calculations can be used to calculate absolute free energies

The absolute free energy of a system can be determined if we know its partition function \( Q \), a function that connects microscopic properties of the system with macroscopic thermodynamic quantities. Unfortunately, it is very hard to calculate the absolute free energy of real systems because we don’t know their partition functions. Free energy calculations allow us to bypass this problem, but require at least two states: a reference state whose free energy can be analytically or numerically found, and a final state of interest [36, 37]. We chose to calculate the free energy difference using alchemical free energy calculations, a method in which we simulate a series of nonphysical intermediates between the end states [38].

Each intermediate state in the alchemical path is described by a Hamiltonian \( H(q, p; \lambda) \), i.e., the energy of the state as a function of atomic positions \( q \), momenta \( p \) and a coupling parameter \( \lambda \):

\[
H(q, p; \lambda) = f(\lambda)H_{\text{initial}}(q, p; \lambda) + g(\lambda)H_{\text{final}}(q, p; \lambda)
\]

where \( H_{\text{initial}} \) and \( H_{\text{final}} \) respectively are the Hamiltonians of the initial and the final state; and \( f(\lambda) \) and \( g(\lambda) \) are functions used to mix the Hamiltonians and are usually set such that \( H = H_{\text{initial}} \) at \( \lambda = 0 \) and \( H = H_{\text{final}} \) at \( \lambda = 1 \).

A variety of different estimators can be used to analyze alchemical free energy calculations, and have different strengths and weaknesses as well as different data requirements. Here, we employ several different estimators we introduce briefly in the following.

One way to calculate the free energy difference (\( \Delta A \)) between the end states is Thermodynamic Integration (TI) [39]:

\[
\Delta A = \sum_{\lambda=0}^{\lambda=1} \frac{1}{\partial H/\partial \lambda}(\lambda) \text{d}\lambda
\]

in which a set of discrete \( \lambda \) values correspond to states along the alchemical path. \( \beta \) means that we are have to calculate the ensemble average of the derivative between the brackets. TI performs as well as more efficient methods if the integrand is smooth, but breaks down if this condition is not satisfied [40, 41, 42].

An alternate free energy estimation method computes \( \Delta A \) directly via:

\[
\Delta A = -\frac{1}{\beta} \ln \langle e^{-\beta(H_{\text{final}} - H_{\text{initial}})} \rangle_{\text{initial}}
\]

where the ensemble average is calculated over the configurations of the initial state, and \( \beta \) is the reciprocal of \( k_B T \), the product between the Boltzmann constant and the absolute temperature. We call this approach exponential averaging [43] (EXP).

Most free energy calculations involve many intermediates associated with the coupling parameter \( \lambda \), allowing simulation of intermediate states in between the two end states of interest. The free energy change between the end points of a path defined by \( N \) intermediates is:

\[
\Delta A = \sum_{n=1}^{N-1} \Delta A_{n-n+1}
\]

where \( \Delta A_{n-n+1} \) is the free energy difference between \( n \)-th and the \( n+1 \)-th intermediate states. Eq. 5 can be used to calculate the free energy difference between each adjacent pair of states and yields the exact result at the limit of very large samples, but it is inefficient for most applications [38].

The Bennett acceptance ratio [44] (BAR) provides an estimator which is superior for most purposes. It calculates the free energy difference between the \( n \)-th and the \( (n+1) \)-th states from the following relationship:

\[
\left\langle \frac{1}{1 + \frac{N_{n+1}}{N_n} e^{\beta(\Delta H_{n+1,n} - \Delta A)}} \right\rangle_n = \left\langle \frac{1}{1 + \frac{N_{n+1}}{N_n} e^{\beta(\Delta H_{n+1,n} + \Delta A)}} \right\rangle_{n+1}
\]

where \( N_n \) and \( N_{n+1} \) respectively are the number of statistically independent samples in states \( n \) and \( n+1 \), and \( \Delta H_{n,n+1} = -\Delta H_{n+1,n} \) are the Hamiltonian differences between \( n \) and \( n+1 \). BAR is more efficient than EXP [45, 46] and minimizes the free energy uncertainty [44]. Multistate Bennett acceptance ratio [41] (MBAR) is an extension of BAR that takes in consideration the degree of configuration space overlap between a given state and all other states in the transformation, while BAR only uses the information of neighboring states. MBAR and BAR perform similarly when the spacing between the intermediate states is moderate, but MBAR is the most well-performing free energy estimator [42].

The absolute free energy of a solid is calculated using an ideal system as reference

In this work, we seek to predict solubilities of molecular solids. Part of this problem requires predicting the free energy or chemical potential of the solid. One way this has been attempted in the past is via the Einstein Crystal Method (ECM) which calculates the absolute free energy of a solid using an Einstein crystal as a reference state. In this method, the crystal lattice is made of atoms restrained to their positions by a harmonic potential; additionally, the center of mass of the system is held fixed [47].

In the ECM, and in this work, the absolute free energy of the molecular solid is found by designing a path where force field terms are progressively turned on, and the harmonic potential position restraints are turned off. The fixed center of mass is important to avoid a quasidivergence issue when calculating the free energy term of releasing the system from the harmonic position restraints, but the contribution of the fixed center of mass needs to be included in the cycle to obtain the correct absolute free energy for the system (Fig. 1(a)) [47, 48, 49].

In ECM, the free energy is calculated by:

\[
A_{\text{id}} = A_{\text{FCM}}^{\text{EC}} + \Delta A_{\text{EC} \rightarrow \text{IEC}} + \Delta A_{\text{IEC} \rightarrow \text{SFCM}} + \Delta A_{\text{release C M}}
\]
where $A_{EC_{FCM}}$ is the free energy of the Einstein crystal (EC) with a fixed center of mass (FCM); $\Delta A_{EC\rightarrow IEC}$ is the free energy difference between the Einstein crystal (EC) and the interacting Einstein crystal (IEC), i.e., the free energy difference in a transformation where the force field is progressively turned on throughout the calculation path. $\Delta A_{IEC\rightarrow SFCM}$ is the free energy difference between the IEC and the solid with a fixed center of mass (SFCM), i.e., turning off the harmonic restraints; and $\Delta A_{release\,CM}$ is the free energy of release of the center of mass (CM).

ECM can be difficult to implement because of the need for a fixed center of mass, so our work here is based on an alternative approach which is easier to implement. When particles move in ECM, the lattice needs to be moved because the center of mass is fixed [47, 50, 48]. Our method of choice, the Einstein Molecule Method (EMM, see Fig. 1(b)), fixes a single atom in the lattice instead of the center of mass and is more easily implemented than ECM because of the relative difficulty of introducing center of mass restraints into existing simulation packages[50, 22, 48, 51, 52]. EMM has been used to predict phase diagrams of TIP4P and SPC/E water models [48], free energies of ice polymorphs, solid methanol and toy systems [49, 52], and the solubilities of potassium and sodium chlorides [22, 51].

In EMM, the free energy of a solid is:

$$A_{solid} = A^EM + \Delta A_{EM\rightarrow IEM} + \Delta A_{IEM\rightarrow solid}$$

where $A^EM$ is the free energy of the ideal Einstein molecule; $\Delta A_{id\rightarrow IEM}$ is the free energy difference between the ideal Einstein molecule and the interacting Einstein molecule (i.e., turning on the force field); and $\Delta A_{IEM\rightarrow solid}$ is the free energy difference between the interacting Einstein molecule and the solid (i.e., turning off the harmonic restraints). The advantage of EMM over ECM is the absence of the need to calculate a free energy term associated with releasing the fixed reference point [48].

Here, as per equation 9, we compute the free energy of the solid by combining the absolute free energy of the ideal Einstein molecule with two terms that we calculate via alchemical free energy calculations — $\Delta A_{EM\rightarrow IEM}$ and $\Delta A_{IEM\rightarrow solid}$; these involve alchemically changing the interactions in the system. Numerical integration of Eq. 10 allows the calculation of the ideal term, $A^EM$ [52]:

$$A^EM = -\frac{1}{\beta} \ln Q_{EM} = \frac{1}{\beta} \ln \frac{N A^3}{V} - \frac{1}{\beta} \ln \int e^{-\beta U_{EM,1}(\Omega_1)} d\Omega_1$$

$$- \frac{(N-1)}{\beta} \ln \int \frac{1}{\Omega_2} e^{-\beta U_{EM,2}(r_2,\Omega_2)} dr_2 d\Omega_2$$

The chemical potential of a component of a solution can be calculated using free energy calculations.

Another critical component of computing the solubility of a compound is estimating the chemical potential of a solute in solution, since the solubility point is the concentration at which the chemical potentials of compound in the two phases are equal.
The chemical potential of a component \( i \) in solution, \( \mu_i \), has an ideal and an excess component:

\[
\mu_i = -\frac{1}{\beta} \ln q_i + \frac{1}{\beta} \ln \frac{\Lambda^3 N_i}{V} - \frac{1}{\beta} \ln \left( e^{-\beta[U(N_i+1)−U(N_i)]}\right)_{\text{initial}}
\]

where \( q_i \) is the internal partition function of a single molecule of the solute, \( U(N_i) \) is the potential energy of the system with \( N_i \) particles, \( \Lambda \) is the de Broglie thermal wavelength, and \( V \) is the system’s volume [53]. \( \beta \) means that the term was obtained from an ensemble average over the configurations of the initial state (see Eq. 5). The first two terms of the equation above correspond to the ideal component of \( \mu_i \); the last one, \( \mu_i^{\text{ex}} \), corresponds to the excess component of \( \mu_i \), and is associated with all non-ideal interactions of the extra component \( i \) with the solution (i.e. physical interactions that differ from those given by the ideal gas law). We obtained excess chemical potentials from solvation free energy calculations; the solute molecule is inserted in the solution by progressively turning on its interactions with the surrounding environment [24, 54, 28].

The challenge associated with the calculation of \( \mu_i \) is the calculation of the standard chemical potential of \( i \), \( \mu_i^0 \), the first term of equation 11. \( q_i \), the internal partition function, includes the rotation, vibrational, electronic and nuclear partition functions of a single molecule [53] and is unknown. Here, we found a way of calculating \( \mu_i^0 \) without the knowledge of \( q_i \) by alchemically transforming a single solute molecule into a single Einstein molecule, whose absolute free energy we know how to calculate [50, 48, 49].

**Methods**

**Systems under study**

Here, we chose three systems to study: An argon crystal for some small initial tests, \( \alpha \)-methanol to help establish our protocol, and acetylsalicylic acid (ASA) as our main object of study. ASA is a known anti-inflammatory whose most stable polymorph, form I [57], has an aqueous solubility of approximately 0.038 % mole fraction at 298 K [58]. We also used \( \alpha \)-methanol at 150 K and a toy face-centered cubic (fcc) argon crystal [59] to help us find an optimal protocol to calculate the absolute free energy of a molecular solid. \( \alpha \)-methanol was chosen because it had been used before in a study which applied the EMM to calculate the absolute free energy of the solid [52].

All simulations were run in GROMACS 4.6.7 [60, 61, 62, 63]. With one exception, all simulations used the General Amber force field (GAFF) version 1.7 with AM1-BCC charges [64, 65]; the exception was \( \alpha \)-methanol, because we ran these simulations using the input files – coordinates and force field parameters – provided by Aragonés et al., who used an united atom version of the OPLS force field [52].

We simulated all solids and liquids using 5 ns Langevin dynamics simulations. ASA, \( \alpha \)-methanol, and argon were simulated respectively at 298.15 K, 150.0 K, and 4.0 K. Our simulations had the same length as the simulations run by Aragonés et al. All solid state simulations were run in NVT conditions. Liquid state simulations were run in NPT conditions; pressure was kept constant at 101.335 kPa using the Parrinello-Rahman barostat [66]. We used the TIP3P water model [67] for all our liquid state simulations. More simulation details and example input files with full details can be found in the Supporting Information.

**Calculation of the absolute free energy of molecular crystals**

The absolute free energies of the solids were calculated from trajectories of simulation boxes with 64 ASA molecules, 100 OPLS methanol molecules, and 864 argon atoms with periodic boundary conditions. ASA’s unit cell was obtained from Mercury CSD 3.8 [68] and the fcc argon crystal was obtained from the literature [59]. Simulation box sizes were chosen to be approximately between 2 nm and 3 nm to ensure that box sizes were large enough that atoms and their periodic copies were not within cut-off distance of one another. \( \alpha \)-methanol’s crystal was obtained from the Supporting information of Aragonés et al.[52] We used Amber14’s ambertools [69, 70, 71, 72] and ParmEd [73] to generate the ASA’s and argon’s solid state input files. All atoms but one were subjected to harmonic constraints in the \( x \), \( y \), and \( z \) coordinates. A single atom was kept fixed in space to act as the reference point for the calculations, as explained in the Introduction.

Monte Carlo integration yielded \( A_{EM} \), the free energy of the Einstein molecule, as it was previously done for \( \alpha \)-methanol in the literature[52]. \( \Delta A_{d→EM} \) and \( \Delta A_{EM→solid} \) were estimated using Thermodynamic Integration (TI) [39] and the Multistate Bennett Accep-
tance Ratio (MBAR) [74]. We used force constants of 4000 kJ/molÅ² to restrain atoms to their lattice positions in acetylsalicylic acid and argon simulations because it allowed us to use a reasonable time step of 1.0 fs in all simulations. α-methanol simulations used the same force constant that had been previously used by Aragonès et al. [52].

We used alchemical free energy calculations to obtain the difference in free energy between the reference Einstein molecule and the solid. This step was divided in two parts: (a) the force field parameters are alchemically turned on, and (b) the harmonic constraints are turned off.

Here, we deviate from earlier work which calculated the absolute free energy of a solid using EMM by introducing additional intermediate states to improve accuracy, along with using a superior free energy estimator.

For the calculation of $\Delta A^{id \rightarrow IEM}$, we found it was crucial to introduce intermediate states; we also switched to using the MBAR estimator. The original EMM calculation of the absolute free energy of a solid [48, 49, 50, 22, 51, 52] estimated $\Delta A^{id \rightarrow IEM}$ using exponential averaging (EXP) with just two states: the Einstein molecule (EM) and the interacting Einstein molecule (IEM) [48, 49, 50, 22, 51, 52, 21, 55]. As EXP is known to have convergence issues and biases [40, 38, 41, 45], we switched to the superior MBAR free energy estimator[74]. Additionally, when we did so, we found that overlap of states (as measured by the overlap matrix [75]) was insufficient so we created a series of intermediate states connecting both ends of the transformation.

For $\Delta A^{IEM \rightarrow solid}$, the original work used Thermodynamic Integration (TI) [39]. Here, we replaced TI with MBAR as our analysis method of choice. Generally, the literature shows that TI performs as well as more efficient methods like BAR and MBAR when the integrand is smooth [40, 38, 41], but it is sensitive to the choice and number of intermediate states [76]. MBAR is the most consistently well-performing free energy estimator [42] and exploits the overlap between states more thoroughly than its predecessor, the Bennett Acceptance Ratio (BAR) estimator [74]. Here, we chose to compare performance of MBAR and TI for calculation of $\Delta A^{IEM \rightarrow solid}$ for ASA and α-methanol; we also applied EXP as a comparison in the latter case only.

**Chemical potential calculations**

The chemical potential of a pure solid is its molar free energy:

$$\mu = \frac{A}{N} \quad (12)$$

where $N$ is the number of molecules in the solid, and $A$ its Helmholtz free energy.

The chemical potential of a substance $i$ in water is defined as the derivative of the free energy of the system with respect to the composition:

$$\mu_i = \left( \frac{\partial G}{\partial N_i} \right)_{P,T,N_{H_2O}} \quad (13)$$

where $G$ is the Gibbs free energy, and $N_i$ is the number of molecules of $i$ in solution; $P$, $T$, and $N_{H_2O}$ are the pressure, absolute temperature, and number of water molecules in solution, and are kept constant in the calculation.

One important aspect to discuss is reason why we chose to calculate the Helmholtz free energy for the solid and Gibbs free energies for each solution. Solid state simulations with position restraints required running under constant temperature and constant volume conditions due to software limitations, therefore we were able to calculate $A$ for the solids.. At constant pressure, both kinds of free energy are related by:

$$\Delta G = \Delta A + P \Delta V \quad (14)$$

Since solids are much less susceptible to volume changes than liquids, it is reasonable to consider that $P \Delta V$ is negligible and $\Delta G \approx \Delta A$. For instance, the difference in volume between the experimental ASA crystal structure and the simulation box after a constant pressure equilibration stage is 0.14 nm³. The $P \Delta V$ term – i.e., the free energy difference discounting possible structure relaxation effects – would be much smaller than the simulation error.

As we explain in more detail in the Results section, successful absolute free energy calculations for molecular solids require a pathway involving a large number of alchemical intermediate states. The calculation of the absolute free energies of α-methanol at 150 K and acetylsalicylic acid required 600 states. Since GROMACS reads each $\lambda$ until its fourth decimal place and the states need to be spaced more closely together as the harmonic restraints are turned off (See Supporting Information), we decided to split each free energy calculation in sets of 100 states.

Liquid state simulation boxes were generated using the SolvationToolkit [77], a Python package that uses packmol [78], OpenMolTools (v0.6.7) [79] and OpenEye Python Toolkits [80, 81, 82]. Excess chemical potentials were obtained with the same solvation free energy protocol used in previous studies [28]: Starting from a fully interacting system, we progressively decouple the interactions of a single solute molecule with the remaining of the system, which allows us to calculate the free energy difference between a solute molecule in vacuum and in solution, i.e., the solvation free energy.

We also used alchemical free energy calculations using a single Einstein molecule as a reference state to estimate the standard chemical potential of a substance, $\mu_i^0$:

$$\mu_i^0 = \mu_i^{ideal} - (\mu_i^{FF off} + \mu_i^{restraining}) \quad (15)$$

where $\mu_i^{FF off}$ and $\mu_i^{restraining}$ respectively are the chemical potential associated with turning off the force field and chemical potential of restraining the atoms of the molecule to their lattice positions. $\mu_i^{ideal}$ is calculated using the Monte Carlo integration procedure that we used to calculate $A^{EM}$ to a single molecule.

**Results**
Table 1. Absolute free energy components for α-methanol at 150 K, in $k_B T$.

| Literature [52] | Our replica |
|-----------------|-------------|
| $A_{EM}$        | 29.05       |
| $\Delta A_{id \rightarrow EM}$ | $-41.27(1)$ |
| $\Delta A_{IEM \rightarrow solid}$ | $-17.33(3)$ |

Chemical potential of molecular solids

The first step to predict aqueous solubilities with the aid of absolute free energy calculations was the assessment of the methodologies we chose to use. Since our method is the same one used by Aragonès et al. [52] and we wanted to be sure that we could reproduce previous results, we ran simulations for α-methanol at 150 K and estimated the free energies of solids using MBAR. Turning off the harmonic restraints was the challenging step. Our MBAR calculation of $\Delta A_{IEM \rightarrow solid}$ for α-methanol using 18 intermediate states yielded $-18(3) k_B T$ while our TI result was $-18.421(5) k_B T$ and the literature result was $-17.33(3) k_B T$ using 17 states [52]. The MBAR error was unusually high (3 k$B T$) which is usually a signal of overlap problems or other serious concerns.

MBAR is a free energy estimation method that minimizes the free energy variance and considers the overlap between a given state and all the others in the transformation path [41], which means that high uncertainties ($\pm 3 k_B T$) suggest the presence of problems in the transformation’s path. TI’s uncertainty estimates are much lower, but we believe this is an artifact. Error analysis for TI simply does not work the same way and does not give insight into whether exploration of phase space is adequate, unlike MBAR. Specifically, uncertainty estimates from TI usually factor in only the uncertainty in the integrand at each sampled lambda value and could potentially also factor in the smoothness of the integrand (i.e. numerical integration error) but do nothing to factor in whether the integrand will in fact vary smoothly in between lambda points; usually no data is available on this. BAR and MBAR, in contrast, factor in information about how well the intermediate states overlap in phase space and reflect high uncertainties when phase space overlap is poor. In our experience, usually TI would suffer from similar problems if additional intermediate states were added, but uncertainties in TI typically do not reflect this, as is the case here. Thus, the high uncertainty of the MBAR value indicates a sampling/convergence problem which warrants further exploration.

To explore the high uncertainty of our MBAR free energy estimates, we examined the degree of overlap the intermediate states had with each other. Phase space overlap analysis [83, 84, 85] quantifies the probability that any given configuration of an intermediate state can be found in other states. A good rule of thumb for designing a set of free energy calculations spanning between two states is to ensure that the states along the path have significant overlap with their neighbors as shown in Fig. 2. More overlap improves the quality of the MBAR free energy estimation: Fig. 2(b) represents a set of restraining simulations where the free energy uncertainty can potentially be accurately estimated using BAR and MBAR; Fig. 2(a) shows a case where it cannot. In our case we find that the α-methanol simulation using 18 intermediate states does not have adequate overlap (Fig. 3) – specifically, the states $4 \leq \lambda_i \leq 17$ do not have overlapping configurations with other states, which explains the 3 k$B T$ uncertainty in our MBAR estimate.

Since prior work had appeared to do this estimation successfully [52], we were uncertain why we were encountering such overlap problems, so we studied an even simpler system. We calculated $\Delta A_{IEM \rightarrow solid}$ of fcc argon at 4 K with 18 states as in our α-methanol free energy estimation. MBAR yielded an error estimate of infinity while TI estimated $\Delta A_{IEM \rightarrow solid}$ to be $-1666.5(8) k_B T$ which, as we show below, is incorrect. This path resulted phase space overlap diagram without overlap between the states after state number 2 (Fig. 4). Apparently as the harmonic potential that holds atoms in their lattice positions tends to zero, atoms become rather mobile, dramatically decreasing phase space overlap and leading to poor free energy estimates.

To improve phase space overlap, we introduced more intermediate states along the path for removing the restraints (see Fig 2). We chose to break down the simulation in smaller parts, adding a significant amount of states near the point where the harmonic restraints are approximately zero. The MBAR estimate of $\Delta A_{IEM \rightarrow solid}$ for fcc argon is $-1016.0(2) k_B T$ using 300 states. TI’s corresponding value was $-1017.1(1) k_B T$, differing by far from the (incorrect) value of $-1666.5(8) k_B T$ obtained above with fewer states. Phase space overlap diagrams showed significant improvement in the configuration overlap between the states (Supporting Information). Thus, increasing the number of states was an effective strategy, and we used it in all subsequent calculations.

Even though our α-methanol results were similar to results previously published by other authors [52], we need to emphasize that reliable free energies resulted from simulations with a large number of intermediate states, as can be seen in Table 1. Despite its conceptual simplicity, calculating the components of the absolute free energy of a solid to a point where there is significant phase space
Figure 2. Phase space overlap between the states in a thermodynamic path for removing restraints with \( \lambda \). \( \Gamma \) represents the phase space that contains all the configurations for all the states in the path. \( \lambda_0 \) and \( \lambda_1 \) (left) or \( \lambda_N \) (right) represent the end states along the path, each shaded region represents a state in phase space and the red lines represent the configurations visited by the simulation run in the \( \lambda_0 \) state. The restrained state is a subset of the unrestrained one. (a) and (b) represent simulations with different numbers of intermediate states along the path between a fully restrained state (\( \lambda_1 \) (a) or \( \lambda_N \) (b)) and an unrestrained state (\( \lambda_0 \)). In (a), the simulation (red) only visits very few configurations consistent with the restrained state – i.e, there is poor phase space overlap – indicating a need for more intermediate states, otherwise any free energy estimates will be subject to very high uncertainties; in (b) there is still almost no overlap between the simulation and states consistent with \( \lambda_N \), but there is overlap with the next shaded region, \( \lambda_1 \), indicating the potential for overlap and accurate free energy estimates. Thus simulations run in each shaded region are more likely to have a bigger phase space overlap with \( \lambda_N \) than simulations run in \( \lambda_0 \).

The overlap between the intermediate states is computationally demanding. A 900-atom OPLS \( \alpha \)-methanol system required 40 states to calculate \( \Delta A_{id\rightarrow IEM} \), and 600 states for \( \Delta A_{IEM\rightarrow solid} \).

We chose these intermediate states in advance, and these ultimately led to free energy errors smaller than 0.1 \( k_B T \); the estimated TI and MBAR values differed by no more than 0.3 \( k_B T \). Our results for ASA using an optimal number of states can be seen in Table 2. The MBAR chemical potential of ASA at 298.15 K equals to \(-220.67(3) k_B T \).

### Table 2. Absolute free energy components for polymorph I of acetylsalicylic acid (ASA) at 298.15 K, in \( k_B T \).

| Acetylsalicylic Acid | \( A_{EM} \) | \( \Delta A_{id\rightarrow IEM} \) | \( \Delta A_{IEM\rightarrow solid} \) |
|---------------------|--------------|-------------------------|-------------------------|
|                     | 48.047       | \(-167.316(1) \) (TI, 118 states) | \(-101.656(2) \) (TI, 600 states) |
|                     |              | \(-167.07(3) \) (MBAR, 118 states) | \(-101.644(2) \) (MBAR, 600 states) |

The computational cost of calculating \( A^{ASA} \) was high; each state required a separate simulation (of a 1344-atom ASA system), with 718 states in total. Simulations typically required 11 hours on a single CPU, so the calculation of a single absolute free energy of a molecular solid required approximately 7898 CPU-hours.

**Chemical potential of solutions and the solubility of GAFF acetylsalicylic acid in TIP3P water.**

Equation 11 states that the absolute chemical potential of a solution is determined by three quantities: \( \mu_0^i \), the standard chemical potential; \( \mu_{ex}^i \), the excess chemical potential of the component at a concentration of \( \chi \); and a volume-dependent ideal gas component of \( k_B T \cdot \ln (\Lambda_i^3 \cdot (N_{ASA}/V)_{solution}) \). \( \mu_{ASA}^0 \) only required information regarding the internal structure of the molecule [53], thus we estimated \( \mu_{ASA}^0 \) by alchemically transforming a single solute molecule into a single Einstein molecule (Table 3), whose absolute free energy we know how to calculate. We used the same number of states that we chose for the solid state simulations and we found that \( \mu_{ASA}^0 \) is equal to \(-150.7(2) k_B T \), as discussed in the last subsection of the Methods section.

Concentrations, volumes and excess chemical potentials can be seen in Table 4. We obtained the excess chemical potentials from solvation free energy calculations [24, 54, 28]. Volumes were obtained from the state in the alchemical path where the solute was fully coupled...
to the rest of the system.

The experimental aqueous solubility of ASA is approximately 0.038% in water at 298 K [58], but our model predicts that ASA is effectively insoluble in water (Figure 5). While all-atom simulations can yield solubility estimates given adequate simulation time and a correct method, the computed solubility will be that dictated by the underlying energy model or force field, and will not necessarily match experiment. Here, we use GAFF, a general-purpose force field with known limitations [86, 71, 87, 28]; apparently, here, the right answer for the force field is not correct. Perhaps this is because of limitations in describing the solid state, as the force field is parameterized for liquid state simulations. Indeed, classical fixed charge force fields have shown severe limitations for polymorph prediction for these reasons [31, 34, 33, 5, 35]. Also, point partial atomic charges regularly used in molecular dynamics do not describe electrostatic interactions in a solid particularly well [88]. In the case of the ASA crystal, it is possible that its hydrogen bonds and π-stacking interactions add layers of complexity that are not properly described by GAFF.

**Discussion and Conclusions**

Despite its theoretical rigor, solubility prediction from absolute free energy calculations is a difficult task: it is computationally expensive and, at least in the present approach, requires many different steps and a great deal of care. Here, we attempted to develop and test a general approach to compute the solubility of molecular solids by adapting the Einstein Molecule Method (EMM) to tackle this problem, as discussed above.

To tune our methodology, we initially decided to repro-

![Figure 3](image1.png)

**Figure 3.** Phase space overlap between the states in the path between IEM and the α-methanol solid. The sum of all the elements in a row should yield 1.0, a probability of 100%. A good free energy estimate is obtained when the states along the alchemical path contain configurations that can be found in other intermediate states. Here, however, the phase space overlap is non-zero, which results in non-zero off-diagonal elements. Here, however, the phase space overlap plot shows that there is no overlap between the states \( \lambda_i \), \( 4 \leq i \leq 17 \) indicating poor free energy estimates will result.

![Figure 4](image2.png)

**Figure 4.** Phase space overlap between the states in the path between IEM and the fcc argon solid. A good free energy estimate is obtained when the states along the alchemical path contain configurations that can be found in other intermediate states. Here, however, the phase space overlap diagram shows that there is no overlap between the states \( \lambda_i \), \( 3 \leq i \leq 17 \), which explains the poor quality of the free energy result.

| Table 3. Standard chemical potential of acetylsalicylic acid (ASA) at 298.15 K, in \( \mkappa T \). |
|----------------|----------------|
| **Acetylsalicylic Acid** |             |
| \( \mu_{\text{ASA}} \) | 9.3 (MBAR 600 states) |
| \( \mu_{\text{FF off}} \) | 65.7409(9) (MBAR 118 states) |
| \( \mu_{\text{Rest training}} \) | 94.3(2) (MBAR 600 states) |

![Figure 5](image3.png)

**Figure 5.** Chemical potentials of ASA, solid and solution in different concentrations, with respect to mole fraction.
duce the absolute free energy of solid α-methanol, one of methanol’s polymorphs, at 150 K using EMM before doing the same calculations for our compound of choice, ASA. We verified that the free energy differences between the Einstein molecule and the interactive Einstein molecule ($\Delta A_{\text{EM-HEM}}$) and between the latter state and the solid ($\Delta A_{\text{EM-solid}}$) were more reliably estimated with the multistate Bennett acceptance ratio (MBAR). The absolute free energy of the crystal (as computed for united-atom OPLS α-methanol) agreed with results found in the literature, which suggested that we were on the right path. We did, however, require a very large number of intermediate alchemical states to obtain accurate free energy estimates, making these simulations fairly computationally demanding.

We then chose to calculate the solubility of ASA due to its pharmacological importance and due to its relative complexity in comparison to previous molecular solids whose absolute free energies have been computed via EMM previously [52]. As for α-methanol, this calculation required a large number of intermediate alchemical states and considerable computational cost – approximately 8000 CPU hours for a single absolute free energy calculation for the molecular solid, even with the crystal structure as input. Perhaps the number of intermediate states could be further optimized, but clearly a large number of intermediate simulations was required and thus considerable computational cost. Despite all of this, we still could not reproduce the experimental aqueous solubility of acetylsalicylic acid (ASA); experimentally it is modestly soluble, whereas our work would suggest it is essentially completely insoluble in water, likely due to force field limitations.

The solubility of naphthalene was recently estimated using a similar methodology, the Extended Einstein Crystal Method [23], but with additional approximations. Specifically, since naphthalene molecules interact very weakly with each other in the crystal lattice and with water molecules in solution, the differences between the internal partition function of a naphthalene molecule in the solid and in the solution were assumed to be negligible. This allowed the authors to drop some complexities in treatment of the solution-phase part of the calculation. However, that approach is only suitable for compounds that are only very weakly interacting in solution and in the crystal. ASA, in contrast, is a molecule that interacts strongly with other ASA molecules in its crystal lattice and with water molecules in solution via hydrogen bonds. For instance, an important crystalline feature that is not necessarily present in solution is the dimer structure, with two ASA molecules bound together via hydrogen bonds between the carboxylic acid groups. Differences between the internal partition functions of the molecule in the solid ($q_{\text{ASA}}^{\text{solid}}$) and in solution ($q_{\text{ASA}}^{\text{solution}}$) would probably not be negligible in this scenario, thus a more general approach is needed for treatment of such cases. Our work here provides one attempt in that direction.

Overall, the present approach seems to have significant limitations – most notably that the computational expense is considerable, and the resulting estimated solubility is quite inaccurate. Perhaps both of these may be surmountable; GPU-based free energy calculations can be dramatically faster potentially making an 8000 CPU-hour calculation be 80 GPU hours which would amount to overnight on 8 GPUs, and perhaps this could be optimized via changes to simulation time and number of intermediate states. And with better force fields, perhaps accuracy could be improved; the AMOEBA-based approach of Schnieders shows considerable promise [15]. Alternatively, other approaches may be of interest. Solubility has been predicted by simulations using pseudocritical paths (i.e., paths were molecular crystals are transformed in tractable Einstein crystal-like states between the ending states of the transformation [89, 90, 91, 92]) and a single experimental reference point [91]), and with the aid of a thermodynamic cycle formed by the molecular crystal, the molecule in vacuum, and the solvated molecule [15].

We believe the time has come for routine physical methods for estimation of solubility, even if improved force fields prove necessary before results have significant ac-

### Table 4. Simulation data for solutions of acetyl salicylic acid in water in different concentrations.

| Molar fraction (%) | Volume (nm$^3$) | # solute molecules | # solvent molecules | $\mu^\text{sol}$ ($k_B T$) |
|--------------------|-----------------|--------------------|---------------------|--------------------------|
| 2.000 e-03         | 3035.99(5)      | 2                  | 99998               | −16.80(5)               |
| 6.666 e-03         | 911.17(2)       | 2                  | 30002               | −15.88(4)               |
| 7.999 e-03         | 759.33(1)       | 2                  | 25000               | −15.51(5)               |
| 9.998 e-03         | 911.45(3)       | 3                  | 30003               | −15.65(4)               |
| 9.999 e-03         | 607.59(2)       | 2                  | 20000               | −15.47(5)               |
| 1.3330 e-02        | 911.72(2)       | 4                  | 30004               | −15.77(4)               |
| 1.3332 e-02        | 455.84(2)       | 2                  | 15000               | −15.61(4)               |
| 1.666 e-02         | 912.00(3)       | 5                  | 30005               | −15.96(5)               |
| 1.9992 e-02        | 912.27(2)       | 6                  | 30006               | −15.78(4)               |
| 1.9996 e-02        | 304.01(1)       | 2                  | 10000               | −15.62(5)               |
| 3.998 e-02         | 152.25(1)       | 2                  | 5000                | −15.41(6)               |
| 1.996 e-01         | 30.835(7)       | 2                  | 1000                | −16.37(5)               |
| 2.991 e-01         | 31.069(3)       | 3                  | 1000                | −16.40(6)               |
| 3.984 e-01         | 31.309(7)       | 4                  | 1000                | −16.62(6)               |
| 4.975 e-01         | 31.547(3)       | 5                  | 1000                | −17.1(1)                |
curacy for application to biomolecular design problems.

Author contributions
Competing interests

DLM is a member of the Scientific Advisory Board for OpenEye Scientific Software.

Grant information

DLM and GDRM appreciate the financial support from the National Science Foundation (CHE 1352608), and computing support from the UCI GreenPlanet cluster, supported in part by NSF Grant CHE-0840513. GDRM appreciates support from the Brazilian agency CAPES - Science without Borders program (BEX 3932-13-3)

Acknowledgements
The authors also would like to thank Dr. Gaetano Calabrò (OpenEye Software), Prof. Michael Shirts (University of Colorado, Boulder), Dr. Eric Dybeck (Pfizer), and Prof. Michael Schnieders (University of Iowa) for fruitful discussions on the project.

References
[1] Madhu Pudipeddi and Abu T.M. Serajuddin. Trends in solubility of polymorphs. J. Pharm. Sci., 94(5):929–939, May 2005. ISSN 1520-6017. doi: 10.1002/jps.20302.
[2] J. Bauer, S. Spanton, R. Henry, J. Quick, W. Dziki, W. Porter, and J. Morris. Ritonavir: An extraordinary example of conformational polymorphism. Pharm. Res., 18(6):859–866, June 2001. ISSN 0724-8741.
[3] W. D. S. Motherwell, H. L. Ammon, J. D. Dunizt, A. Dzyabchenko, P. Eik, A. Gavezottzi, D. W. M. Hofmann, F. J. J. Leusen, J. van de Streek, F. P. A. Fabbiani, P. Hidber, and O. Grassmann. Combined crystal structure prediction and high-pressure crystallization in rational pharmaceutical polymorph screening. Nature Communications, 6:7793, July 2015. ISSN 2041-1723. doi: 10.1038/ncomms8793.
[4] Gregory J. O. Beran. Modeling Polymorphic Molecular Crystals with Electronic Structure Theory. Chem. Rev., 116(9):5567–5613, May 2016. ISSN 0009-2665. doi: 10.1021/acs.chemrev.5b00648.
[5] Antonio Llinás, Robert C. Glen, and Jonathan M. Goodman. Solubility Challenge: Can You Predict Solubilities of 32 Molecules Using a Database of 100 Reliable Measurements? J. Chem. Inf. Model., 48(7):1289–1303, July 2008. ISSN 1549-9596. doi: 10.1021/ci080058v.
[6] Doris E. Braun, Jennifer A. McMahon, Lien H. Koztecki, Sarah L. Price, and Susan M. Reutzel-Edens. Contrasting Polymorphism of Related Small Molecule Drugs Correlated and Guided by the Computed Crystal Energy Landscape. Crystal Growth & Design, 14(4):2056–2072, April 2014. ISSN 1528-7483. doi: 10.1021/cg500185h.
[7] Aurora J. Cruz-Cabeza and Joel Bernstein. Conformational Polymorphism. Chem. Rev., 114(4):2170–2191, February 2014. ISSN 0009-2665. doi: 10.1021/cr400249d.
[8] M. A. Neumann, J. van de Streek, F. P. A. Fabbiani, P. Hidber, and O. Grassmann. Combined crystal structure prediction and high-pressure crystallization in rational pharmaceutical polymorph screening. Nature Communications, 6:7793, July 2015. ISSN 2041-1723. doi: 10.1038/ncomms8793.
[21] Mauro Ferrario, Giovanni Cicotti, Eckhard Spohr, Thierry Cartailler, and Pierre Turq. Solubility of KF in water by molecular dynamics using the Kirkwood integration method. *J. Chem. Phys.*, 117(10):4947–4953, August 2002. ISSN 0021-9606. doi: 10.1063/1.1498820.

[22] E. Sanz and C. Vega. Solubility of KF and NaCl in water by molecular simulation. *J. Chem. Phys.*, 126(1):014507, January 2007. ISSN 0021-9606, 1089-7690. doi: 10.1063/1.2397683.

[23] Lunna Li, Tim Totton, and Daan Frenkel. Computational methodology for solubility prediction: Application to the sparingly soluble solutes. *J. Chem. Phys.*, 146(21):214110, June 2017. ISSN 0021-9606. doi: 10.1063/1.4983754.

[24] Pavel V Klimovich and David L. Mobley. Predicting hydration free energies using all-atom molecular dynamics simulations and multiple starting conformations. *J. Comput. Aided Mol. Des.*, 24(4):307–316, 2010. ISSN 0920-654X. doi: 10.1007/s10822-010-9343-7.

[25] Devleena Shivakumar, Joshua Williams, Yujie Wu, Wolfgang Damm, John Shelley, and Woody Sherman. Prediction of Absolute Solvation Free Energies using Molecular Dynamics Free Energy Perturbation and the OPLS Force Field. *J. Chem. Theory Comput.*, 6(5):1509–1519, May 2010. ISSN 1549-9618. doi: 10.1021/ct900203w.

[26] Devleena Shivakumar, Edward Harder, Wolfgang Damm, Richard A. Friesner, and Woody Sherman. Improving the Prediction of Absolute Solvation Free Energies Using the Next Generation OPLS Force Field. *J. Chem. Theory Comput.*, 8(8):2553–2558, August 2012. ISSN 1549-9618. doi: 10.1021/ct300203w.

[27] R. E. Skyner, J. L. McDonagh, C. R. Groom, T. van Mourik, and J. B. O. Mitchell. A review of methods for the calculation of solution free energies and the modelling of systems in solution. *Phys. Chem. Chem. Phys.*, 17(9):6174–6191, February 2015. ISSN 1463-9084. doi: 10.1039/C5CP00288E.

[28] Guilherme Duarte Ramos Matos, Daisy Y. Kyu, Hannes H. Loeffler, John D. Chodera, Michael R. Shirts, and David L. Mobley. Approaches for Calculating Solution Free Energies and Enthalpies Demonstrated with an Update of the FreeSolv Database. *J. Chem. Eng. Data*, 62(5):1559–1569, May 2017. ISSN 0021-9568. doi: 10.1021/acs.jced.7b00104.

[29] Eliot Boulanger, Lei Huang, Chetan Rupakheti, Alexander D. Mackerrill, and Benoit Roux. Optimized Lennard-Jones Parameters for Druglike Small Molecules. *J. Chem. Theory Comput.*, April 2018. ISSN 1549-9618. doi: 10.1021/acs.jctc.7b00172.

[30] Sarah L. Price. The computational prediction of pharmaceutical crystal structures and polymorphism. *Adv. Drug Deliv. Rev.*, 56(3):301 – 319, 2004. ISSN 0169-409X. doi: https://doi.org/10.1016/j.addr.2003.10.006.

[31] G. M. Day, W. D. S. Motherwell, H. L. Ammon, S. X. M. Boerrigter, R. G. Della Valle, E. Venuti, A. Dzyabchenko, J. D. Dunitz, B. Schweizer, B. P. van Eijck, P. Erk, J. C. Facelli, V. E. Bazzetra, M. B. Ferraro, D. W. M. Hofmann, F. J. J. Leusen, C. Liang, C. C. Pantelides, P. G. Karamterzanis, S. L. Price, T. C. Lewis, H. Nowell, A. Torrisi, A. S. A. Arntauda, M. U. Schmidt, and P. Verwer. A third blind test of crystal structure prediction. *Acta Cryst. B*, 61(5):511–527, October 2005. ISSN 0108-7681. doi: 10.1107/S0108768105015653.

[32] Scott M. Woodley and Richard Catlow. Crystal structure prediction from first principles. *Nature Materials*, 7(12):937, December 2008. ISSN 1476-4660. doi: 10.1038/nmat2321.

[33] G. M. Day, T. G. Cooper, A. J. Cruz-Cabeza, K. E. Hejczyk, H. L. Ammon, S. X. M. Boerrigter, J. S. Tan, R. G. Della Valle, E. Venuti, J. Jose, S. R. Gadre, G. R. Desiraju, T. S. Thakur, B. P. van Eijck, J. C. Facelli, V. E. Bazzetra, M. B. Ferraro, D. W. M. Hofmann, M. A. Neumann, F. J. J. Leusen, J. Kendrick, S. L. Price, A. J. Misquitta, P. G. Karamterzanis, G. W. A. Welch, H. A. Scheraga, Y. A. Arntauda, M. U. Schmidt, J. Van de Streek, A. K. Wolf, and B. Schweizer. Significant progress in predicting the crystal structures of small organic molecules – a report on the fourth blind test. *Acta Cryst. B*, 65(2):107–125, April 2009. ISSN 0001-7861. doi: 10.1107/S0108768109004066.

[34] Sarah (Sally) L. Price. Computed Crystal Energy Landscapes for Understanding and Predicting Organic Crystal Structures and Polymorphism. *Acc. Chem. Res.*, 42(1):117–126, January 2009. ISSN 0001-4842. doi: 10.1021/ar000147.

[35] A. M. Reilly, R. I. Cooper, S. C. Adjiman, S. Bhattacharya, A. D. Boese, J. G. Brandenburg, P. J. Bygrave, R. Bylsma, J. E. Campbell, R. Car, D. H. Case, R. Chadha, J. C. Cole, K. Cosburn, H. M. Cuppen, F. Curtis, G. M. Day, R. A. DiTasso Jr, A. Dzyabchenko, B. P. van Eijck, D. M. Elking, J. A. van den Ende, J. C. Facelli, M. B. Ferraro, L. Fusti-Molnar, C.-A. Gatsiou, T. S. Gee, R. de Gelder, L. M. Ghiringhelli, H. Goto, S. Grimme, R. Guo, D. W. M. Hofmann, J. Hoja, R. K. Hylton, L. Iuzzolino, W. Jankiewicz, D. T. de Jong, J. Kendrick, N. J. J. D. Kerk, H.-Y. Ko, L. N. Kuleshova, X. Li, S. Lohani, F. J. J. Leusen, A. M. Lund, J. Ly, Y. Ma, N. Marom, A. E. Masunov, P. McCabe, D. P. McMahon, H. Meekees, M. P. Metz, A. J. Misquitta, S. Mohamed, B. Monsserrat, R. J. Needs, M. A. Neumann, N. Nyman, S. Obata, H. Oberhofer, A. R. Oganov, A. M. Orendt, G. I. Pagola, C. C. Pantelides, C. J. Pickard, R. Podeszwa, L. S. Price, S. L. Price, A. Pulido, M. G. Read, K. Reuter, J. A. van den Ende, C. A. Gatsiou, T. S. Gee, R. de Gelder, L. M. Ghiringhelli, H. Goto, S. Grimme, R. Guo, D. W. M. Hofmann, J. Hoja, R. K. Hylton, L. Iuzzolino, W. Jankiewicz, D. T. de Jong, J. Kendrick, N. J. J. D. Kerk, H.-Y. Ko, L. N. Kuleshova, X. Li, S. Lohani, F. J. J. Leusen, A. M. Lund, J. Ly, Y. Ma, N. Marom, A. E. Masunov, P. McCabe, D. P. McMahon, H. Meekees, M. P. Metz, A. J. Misquitta, S. Mohamed, B. Monsserrat, R. J. Needs, M. A. Neumann, N. Nyman, S. Obata, H. Oberhofer, A. R. Oganov, A. M. Orendt, G. I. Pagola, C. C. Pantelides, C. J. Pickard, R. Podeszwa, L. S. Price, S. L. Price, A. Pulido, M. G. Read, K. Reuter, E. Schneider, C. Schober, G. P. Shields, P. Singh, I. J. Sugden, K. Szalewicz, C. R. Taylor, A. Tkatchenko, M. E. Tuckerman, F. Vacarro, M. Vasileiadis, A. Vázquez-Mayagoitia, L. Vogt, Y. Wang, R. E. Watson, G. A. de Wijz, J. Yang,
Q. Zhu, and C. R. Groom. Report on the sixth blind test of organic crystal structure prediction methods. Acta Cryst. B, 72(4):439–459, August 2016. ISSN 0021-9526. doi: 10.1107/S0252520616007447.

[36] Cristophe Chipot and Andrew Pohorille. Free Energy Calculations Theory and Applications in Chemistry and Biology. Springer, 2007. ISBN 978-3-540-38447-2.

[37] Cristophe Chipot. Frontiers in free-energy calculations of biological systems. WIREs Comput Mol Sci, 4(1):71–89, January 2014. ISSN 1759-0884. doi: 10.1002/wcms.1157.

[38] Michael R. Shirts, David L. Mobley, and John D. Chodera. Alchemical Free Energy Calculations: Ready for Prime Time? In D. C. Spellmeyer and R. Wheeler, editors, Annual Reports in Computational Chemistry, volume 3, pages 41–59. Elsevier, January 2007. doi: 10.1016/S1574-1400(07)30004-6.

[39] John G. Kirkwood. Statistical Mechanics of Fluid Mixtures. The Journal of Chemical Physics, 3(5):300–313, May 1935. ISSN 0021-9606, 1089-7690. doi: 10.1063/1.1749657.

[40] E Marty Ytreberg, Robert H. Swendsen, and Daniel M. Zuckerman. Comparison of free energy methods for molecular systems. The Journal of Chemical Physics, 125(18):184114, November 2006. ISSN 0021-9606, 1089-7690. doi: 10.1063/1.2378907.

[41] Michael R. Shirts and Vijay S. Pande. Comparison of efficiency and bias of free energies computed by exponential averaging, the Bennett acceptance ratio, and thermodynamic integration. J. Chem. Phys., 122(14):144107, April 2005. ISSN 0021-9606. doi: 10.1063/1.1873592.

[42] Himanshu Paliwal and Michael R. Shirts. A Benchmark Test Set for Alchemical Free Energy Transformations and Its Use to Quantify Error in Common Free Energy Methods. J. Chem. Theory Comput., 7(12):4115–4134, 2011. doi: 10.1021/ct200395s.

[43] Robert W. Zwanzig. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. The Journal of Chemical Physics, 72(8):1420–1426, August 1954. ISSN 0021-9606, 1089-7690. doi: 10.1063/1.1740409.

[44] Charles H Bennett. Efficient estimation of free energy differences from Monte Carlo data. J. Comp. Phys., 22(2):245–268, October 1976. ISSN 0021-9991. doi: 10.1016/0021-9991(76)90078-4.

[45] Di Wu and David A. Kofke. Asymmetric bias in free-energy perturbation measurements using two Hamiltonian-based models. Phys. Rev. E, 70(6 Pt 2):066702, December 2004. ISSN 1539-3755. doi: 10.1103/PhysRevE.70.066702.

[46] Di Wu and David A. Kofke. Phase-space overlap measures. I. Fail-safe bias detection in free energies calculated by molecular simulation. Journal of Chemical Physics, 123(5), 2005. ISSN 00219606. doi: 10.1063/1.1992483.

[47] D. Frenkel and A. J. C. Ladd. New Monte Carlo method to compute the free energy of arbitrary solids. Application to the fcc and hcp phases of hard spheres. J. Chem. Phys., 81(7):3188–3193, 1984. ISSN 00219606. doi: 10.1063/1.448024.

[48] C Vega, E Sanz, J L. F. Abascal, and E G Noya. Determination of phase diagrams via computer simulation: Methodology and applications to water, electrolytes and proteins. Journal of Physics: Condensed Matter, 20(15):153101, April 2008. ISSN 0953-8984, 1361-648X. doi: 10.1088/0953-8984/20/15/153101.

[49] E. G. Noya, M. M. Conde, and C. Vega. Computing the free energy of molecular solids by the Einstein molecule approach: Ices XIII and XIV, hard-dumbbells and a patchy model of proteins. J. Chem. Phys., 129(10):104704, September 2008. ISSN 1089-7690. doi: 10.1063/1.2971188.

[50] Carlos Vega and Eva G. Noya. Revisiting the Frenkel-Ladd method to compute the free energy of solids: The Einstein molecule approach. The Journal of Chemical Physics, 127(15):154113, October 2007. ISSN 0021-9606, 1089-7690. doi: 10.1063/1.2790426.

[51] J. L. Aragones, E. Sanz, and C. Vega. Solubility of NaCl in water by molecular simulation revisited. J. Chem. Phys., 136(24):244508, June 2012. ISSN 0021-9606, 1089-7690. doi: 10.1063/1.4728163.

[52] J. L. Aragones, E. G. Noya, C. Valeriani, and C. Vega. Free energy calculations for molecular solids using GROMACS. The Journal of Chemical Physics, 139(3):034104, July 2013. ISSN 0021-9606, 1089-7690. doi: 10.1063/1.4812362.

[53] Arieh Ben-Naim. Molecular Theory of Solutions. Oxford University Press, Oxford, 2006. ISBN 0-19-929969-2.

[54] David L. Mobley and J. Peter Guthrie. FreeSol: A database of experimental and calculated hydration free energies, with input files. Journal of Computer-Aided Molecular Design, 28(7):711–720, July 2014. ISSN 0920-654X, 1573-4951. doi: 10.1007/s10822-014-9747-x.

[55] A. L. Benavides, J. L. Aragones, and C. Vega. Consensus on the solubility of NaCl in water from computer simulations using the chemical potential route. J. Chem. Phys., 144(12):124504, March 2016. ISSN 0021-9606, 1089-7690. doi: 10.1063/1.4943780.

[56] Jooyeon Park, Ian Nessler, Brian McClain, Dainius Maciukevicius, Jonas Baltrusaitis, and Michael J. Schnieders. Absolute Organic Crystal Thermodynamics: Growth of the Asymmetric Unit into a Crystal via Alchemy. J. Chem. Theory Comput., 10(7):2781–2791, July 2014. ISSN 1549-9618. doi: 10.1021/ct500180m.

[57] Y. Kim, K. Machida, T. Taga, and K. Osaki. Structure determination and packing analysis of aspirin crystal. Chem. Pharm. Bull., 33(7):2641–2647, July 1985. ISSN 0009-2363.

[58] Samuel H. Yalkowsky, Yan He, and Parijat Jain. Handbook of Aqueous Solubility Data. CRC Press, second edition edition, 2010. ISBN 978-1-4398-0246-5.

[59] D. G. Henshaw. Atomic Distribution in Liquid and Solid Neon and Solid Argon by Neutron Diffraction. Phys. Rev., 111(6):1470–1475, September 1958. doi: 10.1103/PhysRev.111.1470.
[60] H. J. C. Berendsen, D. van der Spoel, and R. van Drunen. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Comm.*, 91(1):43–56, September 1995. ISSN 0010-4655. doi: 10.1016/0010-4655(95)00042-E.

[61] David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E. Mark, and Herman J. C. Berendsen. GROMACS: Fast, flexible, and free. *J. Comput. Chem.*, 26(16):1701–1718, December 2005. ISSN 1096-987X. doi: 10.1002/jcc.20291.

[62] Berk Hess, Carsten Kutzner, David van der Spoel, and Erik Lindahl. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.*, 4(3):435–447, March 2008. ISSN 1549-9618. doi: 10.1021/ct070301q.

[63] Sander Pronk, Szilárd Páll, Roland Schulz, Per Larsson, Pär Bjelkmar, Rossen Apostolov, Michael R. Shirts, Jeremy C. Smith, Peter M. Kasson, David van der Spoel, Berk Hess, and Erik Lindahl. GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29(7):845–854, April 2013. ISSN 1367-4803. doi: 10.1093/bioinformatics/btt055.

[64] Araz Jakalian, Bruce L. Bush, David B. Jack, and Christopher I. Bayly. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *J. Comput. Chem.*, 21(2):132–146, January 2000. ISSN 1096-987X. doi: 10.1002/(SICI)1096-987X(20000130)21:2<132::AID-JCC5>3.0.CO;2-P.

[65] Araz Jakalian, David B. Jack, and Christopher I. Bayly. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.*, 23(16):1623–1641, December 2002. ISSN 1096-987X. doi: 10.1002/jcc.10128.

[66] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.*, 52(12):7182, 1981. ISSN 00218979. doi: 10.1063/1.328693.

[67] William L. Jorgensen, Jayaram Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926–935, July 1983. ISSN 0021-9606. doi: 10.1063/1.445869.

[68] C. F. Macrae, P. R. Edgington, P. McCabe, E. Piccock, G. P. Shields, R. Taylor, M. Towler, and J. van de Streek. Mercury: Visualization and analysis of crystal structures. *J. Appl. Cryst.*, 39(3):453–457, June 2006. ISSN 0021-8898. doi: 10.1107/S002188980600731X.

[69] David A Case, Thomas E Cheatham III, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M Merz Jr., Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J Woods. The Amber biomolecular simulation programs. *J. Comput. Chem.*, 26(16):1668–1688, December 2005.

[70] Romelia Salomon-Ferrer, David A Case, and Ross C Walker. An overview of the Amber biomolecular simulation package. *WIREs Comput. Mol. Sci.*, 3(2):198–210, September 2012.

[71] Thomas E Cheatham and David A. Case. Twenty-five years of nucleic acid simulations. *Biopolymers*, 99(12):969–977, December 2013. ISSN 1097-0282. doi: 10.1002/bip.22331.

[72] D. A. Case, V. Babin, J. T. Berryman, R. M. Betz, Q. Cai, D. S. Cerutti, T. E. Cheatham, T. A. Darden, R. E. Duke, H. Gohlke, A. W. Goetz, S. Gasarov, N. Homeyer, P Janowsky, J. Kaus, I. Kolossváry, A. Kovalenko, T. S. Lee, L. Grand, T. Luchko, R. Luo, B. Madej, K. M. Merz, F. Pecelli, D. R. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C. L. Simmerling, W. Smith, J. Swails, Walker, J. Wang, R. M. Wolf, X. Wu, and P. A. Kollman. Amber 14, 2014.

[73] Jason Swails, Carlos Hernandez, David L Mobley, Hai Nguyen, Lee-Ping Wang, and Pawel Janowski. Parmed, (accessed October 9, 2015).

[74] Michael R. Shirts and John D. Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.*, 129(12):124105, September 2008. ISSN 0021-9606. doi: 10.1063/1.2978177.

[75] Pavel V. Klimovich and David L. Mobley. A Python tool to set up relative free energy calculations in GROMACS. *J. Comput. Aided Mol. Des.*, 29(11):1007–1014, November 2015. ISSN 0920-654X, 1573-4951. doi: 10.1007/s10822-015-9873-0.

[76] Anita de Ruiter, Stefan Boresch, and Chris Oostenbrink. Comparison of thermodynamic integration and Bennett acceptance ratio for calculating relative protein-ligand binding free energies. *J. Comp. Chem.*, 34(12):1024–1034, May 2013. ISSN 1096-987X. doi: 10.1002/jcc.23229.

[77] Caitlin C. Bannan, Gaetano Calabró, Daisy Y. Kyu, and David L. Mobley. Calculating Partition Coefficients of Small Molecules in Octanol/Water and Cyclohexane/Water. *J. Chem. Theory Comp.*, 12(6):4015–4024, 2016. doi: 10.1021/acs.jctc.6b00449.

[78] L. Martínez, R. Andrade, E. G. Birgin, and J. M. Martínez. PACKMOL: A package for building initial configurations for molecular dynamics simulations. *J. Comp. Chem.*, 30(13):2157–2164, October 2009. ISSN 1096-987X. doi: 10.1002/jcc.21224.

[79] Kyle A Beauchamp, Ariën S Rustenburg, Andrea Rizzi, Julie M Behr, Guilherme Matos, Lee-Ping Wang, Robert T McGibbon, David L Mobley, and John D Chodera. OpenMolTools.

[80] Inc. OpenEye Scientific Software. OEChem Toolkit, 2010. (accessed June 16, 2015).

[81] Inc. OpenEye Scientific Software. QUACPC 1.7.0.2.

[82] Paul C D Hawkins and Anthony Nicholls. Conformer generation with OMEGA: Learning from the data set and the analysis of failures. *J. Chem. Inf. Model.*, 52(11):2919–2936, November 2012.

[83] Daan Frenkel and Berend Smit. *Understanding Molecular Simulation: From Algorithms to Applications*, volume 1. Academic Press Inc. (London) Ltd., 1996. ISBN 978-0-12-267351-1.
Andrew Pohorille, Christopher Jarzynski, and Christophe Chipot. Good Practices in Free-Energy Calculations. *J. Phys. Chem. B*, 114(32):10235–10253, August 2010. ISSN 1520-6106. doi: 10.1021/jp102971x.

Pavel V. Klimovich, Michael R. Shirts, and David L. Mobley. Guidelines for the analysis of free energy calculations. *Journal of Computer-Aided Molecular Design*, 29(5):397–411, May 2015. ISSN 0920-654X, 1573-4951. doi: 10.1007/s10822-015-9840-9.

David L. Mobley, Elise Dumont, John D. Chodera, and Ken A. Dill. Comparison of charge models for fixed-charge force fields: Small-molecule hydration free energies in explicit solvent. *J. Phys. Chem. B*, 111(9):2242–2254, March 2007. ISSN 1520-6106. doi: 10.1021/jp0667442.

Christopher J. Fennell, Karisa L. Wymer, and David L. Mobley. A Fixed-Charge Model for Alcohol Polarization in the Condensed Phase, and Its Role in Small Molecule Hydration. *J. Phys. Chem. B*, 118(24):6438–6446, June 2014. ISSN 1520-6106. doi: 10.1021/jp411529h.

Christian Kramer, Alexander Spinn, and Klaus R. Liedl. Charge Anisotropy: Where Atomic Multipoles Matter Most. *J. Chem. Theory Comput.*, 10(10):4488–4496, October 2014. ISSN 1549-9618. doi: 10.1021/ct5005565.

D. M. Eike, J. F. Brennecke, and E. J. Maginn. Toward a robust and general molecular simulation method for computing solid-liquid coexistence. *J. Chem. Phys.*, 122(1):14115, January 2005. ISSN 0021-9606. doi: 10.1063/1.1823371.

D. M. Eike and E. J. Maginn. Atomic simulation of solid-liquid coexistence for molecular systems: Application to triazole and benzene. *J. Chem. Phys.*, 124(16):164503, April 2006. ISSN 0021-9606. doi: 10.1063/1.2188400.

A. S. Paluch, S. Jayaraman, J. K. Shah, and E. J. Maginn. A method for computing the solubility limit of solids: Application to sodium chloride in water and alcohols. *J. Chem. Phys.*, 133(12):124504, September 2010. ISSN 1089-7690. doi: 10.1063/1.3478539.

Eric C. Dybeck, Natalie P. Schieber, and Michael R. Shirts. Effects of a More Accurate Polarizable Hamiltonian on Polymorph Free Energies Computed Efficiently by Reweighting Point-Charge Potentials. *J. Chem. Theory Comput.*, 12(8):3491–3505, August 2016. ISSN 1549-9618. doi: 10.1021/acs.jctc.6b00397.
Supporting Information

Challenges of the Use of Atomistic Simulations to Predict Solubilities of Drug-like Molecules

Guilherme Duarte Ramos Matos\textsuperscript{1} and David L. Mobley\textsuperscript{2}
\textsuperscript{1}Department of Chemistry, University of California, Irvine
\textsuperscript{2}Departments of Pharmaceutical Sciences and Chemistry, University of California, Irvine

I. SIMULATION DETAILS

The following are GROMACS 4.6.7 simulation input parameters, as are the MDP files with full details which are deposited in the Supporting Information.

**General information**

\begin{itemize}
\item Friction coefficient = \frac{\text{mass}_{\text{particle}}}{\tau}, \tau_t = 2.0 \text{ ps}.
\item Parrinello-Rahman barostat (when applicable): \tau_p = 10 \text{ ps} and compressibility = 4.5 \cdot 10^{-5} \text{ bar}^{-1}.
\end{itemize}

**Electrostatics (solid)**

\begin{itemize}
\item PME cut-off: 1.0 nm.
\item PME order: 4
\item Fourier spacing = 0.10 nm
\item we used the same parameters as the Aragonés et al. in solid state simulations. Additional details can be found in the MDP files deposited with this paper.
\end{itemize}

**Electrostatics (solution)**

\begin{itemize}
\item PME cut-off: 1.2 nm.
\item PME order: 6
\item Fourier spacing = 0.10 nm
\item we used the same parameters as the Aragonés et al. for solid state simulations. Additional details can be found in the MDP files deposited with this paper.
\end{itemize}

**vdW interactions**

\begin{itemize}
\item Cut-off: 1.0 nm
\item Switch at 0.9 nm
\item DispCorr = AllEnerPres
\item additional details can be found in the MDP files deposited with this paper.
\end{itemize}

Solution simulation files were generated using the SolvationToolkit module found at https://github.com/MobleyLab/SolvationToolkit. Solvation Toolkit relies on openmoltools, mdtraj, packmol, ParmEd, and OpenEye tools. As noted in the main body of the text, AM1-BCC charges were assigned with OpenEye’s quacpac python module; we used openmoltools to drive this process. Specific source code used for charging is available at https://github.com/choderalab/openmoltools/blob/v0.6.7/openmoltools/openeye.py#L13. The code generates molecular conformations prior to charging, as was recommended at http://docs.eyesopen.com/toolkits/cookbook/python/modeling/am1-bcc.html.

II. SUPPORTING DETAILS

The \texttt{DA\_ideal\_to\_IEM.csv} file containing the elements of the phase space overlap matrix of a $\Delta A_{EM\rightarrow IEM}$ estimated from an alchemical path of 118 states can be found with this Supporting Information.

DISCLOSURE STATEMENT

DLM is a member of the Scientific Advisory Board for OpenEye Scientific Software. JDC is a member of the Scientific Advisory Board for Schrödinger, LLC.

* dmobley@mobleylab.org
