Article:
Marchese Robinson, RL orcid.org/0000-0001-7648-8645, Geatches, D, Morris, C et al. (7 more authors) (2019) Evaluation of Force-Field Calculations of Lattice Energies on a Large Public Dataset, Assessment of Pharmaceutical Relevance, and Comparison to Density Functional Theory. Journal of Chemical Information and Modeling, 59 (11). pp. 4778-4792. ISSN 1549-9596

https://doi.org/10.1021/acs.jcim.9b00601

Evaluation of Force-Field Calculations of Lattice Energies on a Large Public Dataset, Assessment of Pharmaceutical Relevance and Comparison to Density Functional Theory

Richard L. Marchese Robinson,[a,] Dawn Geatches,[b,] Chris Morris,[b] Rebecca Mackenzie,[b] Andrew G.P. Maloney,[c] Kevin J. Roberts,[a] Alexandru Moldovan,[a] Ernest Chow,[d] Klimentina Pencheva,[d] Dinesh Ramesh Mirpuri Vatvani[c]*

a. Centre for Digital Design of Drug Products, School of Chemical and Process Engineering, University of Leeds, Leeds LS2 9JT, United Kingdom

b. Science and Technology Facilities Council, Daresbury Laboratory, Sci-Tech Daresbury, Warrington WA4 4AD, United Kingdom

c. Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, United Kingdom

d. Pfizer Worldwide R&D, Ramsgate Road, Sandwich CT13 9NJ, United Kingdom

*Corresponding author. E-mail address: dawn.geatches@stfc.ac.uk

Abstract

Crystal lattice energy is a key property affecting the ease of processing pharmaceutical materials during manufacturing, as well as product performance. We present an extensive comparison of 324 force-field protocols for calculating the lattice energies of single component, organic molecular crystals (further restricted to Z' less than or equal to one), corresponding to a wide variety of force-fields (DREIDING, Universal, CVFF, PCFF, COMPASS, COMPASSII), optimization routines and other variations which could be implemented as part of an automated workflow using the industry standard Materials Studio software. All calculations were validated using a large new dataset (SUB-BIG), which we make publicly available. This dataset comprises public domain sublimation data, from which estimated experimental lattice energies were derived, linked to 235 molecular crystals. Analysis of pharmaceutical relevance was performed according to two distinct methods based upon (A) public and (B) proprietary data. These identified overlapping subsets of SUB-BIG comprising (A) 172 and (B) 63 crystals, of putative pharmaceutical relevance, respectively. We recommend a protocol based on the COMPASSII force-field for lattice energy calculations of general organic or pharmaceutically relevant molecular crystals. This protocol was the most highly ranked prior to subsetting and was either the top ranking or amongst the top 15 protocols (top 5%) following subsetting of the dataset according to putative pharmaceutical relevance. Further analysis identified scenarios where the lattice energies calculated using the recommended force-field protocol should either be disregarded (values greater than or equal to zero and/or the messages generated by the automated workflow indicate extraneous atoms were added to the unit cell) or treated cautiously (values less than or equal to -249 kJ/mol), as they are likely to be inaccurate. Application of the recommended force-field protocol, coupled with these heuristic filtering criteria, achieved an RMSE around 17 kJ/mol

(MAD around 11 kJ/mol, Spearman's rank correlation coefficient of 0.88) across all 226 SUB-BIG structures retained after removing calculation failures and applying the filtering criteria. Across these 226 structures, the estimated experimental lattice energies ranged from -60 to -269 kJ/mol, with a standard deviation around 29 kJ/mol. The performance of the recommended protocol on pharmaceutically relevant crystals could be somewhat reduced, with an RMSE around 20 kJ/mol (MAD around 13 kJ/mol, Spearman's rank correlation coefficient of 0.76) obtained on 62 structures retained following filtering according to pharmaceutical relevance method B, for which the distribution of experimental values was similar. For a diverse set of 17 SUB-BIG entries, deemed pharmaceutically relevant according to method B, this recommended force-field protocol was compared to dispersion corrected density functional theory (DFT) calculations (PBE+TS). These calculations suggest that the recommended force-field protocol (RMSE around 15 kJ/mol) outperforms PBE+TS (RMSE around 37 kJ/mol), although it may not outperform more sophisticated DFT protocols and future studies should investigate this. Finally, further work is required to compare our recommended protocol to other lattice energy calculation protocols reported in the literature, as comparisons based upon previously reported smaller datasets indicated this protocol was outperformed by a number of other methods. The SUB-BIG dataset provides a basis for these future studies and could support protocol refinement.

Keywords

lattice energy; sublimation enthalpy; force-field; density functional theory; drug-likeness

Introduction

Being able to anticipate the different solid forms of drug candidates and, based upon knowledge of the solid form, determine their associated properties is crucial for assessing their developability.[1] In particular, being able to accurately predict the lattice energy, a measure of solid state cohesion defined more precisely below, serves a number of useful purposes in the context of pharmaceutical development.

The lattice energy may be related to the solid state contribution to the Gibbs free energy of solution and hence, in turn, related to thermodynamic solubility.[2–4] Whilst incorporating lattice energies calculated from experimental crystal structures may not necessarily produce more accurate quantitative structure-property relationship (QSPR) models of solubility than those based on molecular descriptors alone,[5,6] QSPR models based upon descriptors which can be explicitly related to the solid state or non-solid state contribution to solubility may provide greater physical insight.[2,3] In turn, this may support rational design of new active pharmaceutical ingredients (APIs) or solid forms with more desirable solubility. Control of pharmaceutical solubility affects the dissolution rate and bioavailability of the final drug product,[4] as well as supporting the design of manufacturing unit operations, such as cooling crystallization.[7] Lattice energy calculations also support crystal structure predictions, especially if they are accurate enough to distinguish between polymorphs, which supports polymorph screening and, hence, assessments of thermodynamic stability.[8–10]

Lattice energy calculations may also be used as an intermediate step for other analyses. Firstly, these calculations may be decomposed in terms of molecular synthons and support the design of

molecular crystals and crystallization processes.[9,11–14] Secondly, it is conceivable that crystal structure based lattice energy calculations could support assessment of other factors related to pharmaceutical manufacturing, such as mechanical properties and powder behavior during processing.[15] Using accurate lattice energy calculations as an intermediate step for these other analyses assumes that they correspond to accurate calculations of the constituent contributions, rather than being accurate due to cancellation of errors.

The lattice energy may be defined as the energy change upon bringing together static, infinitely separated molecules (or, equivalently, molecules in an ideal gas), in their lowest energy conformation, to form a static lattice for an ideal crystal.[8,16] The lattice energy is a contribution to the sublimation enthalpy, which can be obtained experimentally.[17] Under certain assumptions, notably assuming the sublimed crystal to be an ideal gas, the relationship between sublimation enthalpy and the lattice energy can be expressed as per equation (1).[17] (In equation (1), ΔH_{sub} denotes the sublimation enthalpy, LE the lattice energy, ΔE_{vib} the change in vibrational energy going from the solid state to the gas phase, R is the molar gas constant, T is the temperature in Kelvin, and x = 3.5 for a linear molecule and x = 4 for a non-linear molecule.) However, if further assumptions are made regarding the solid state and gas phase vibrational states, equation (1) may be further simplified, as per equation (2).[2] Since this equation implies all solids have the same, temperature independent, heat capacity, this is clearly a crude approximation.

$$\Delta H_{sub} = -LE + \Delta E_{vib} + xRT \quad (1)$$

$$\Delta H_{sub} = -LE - 2RT \quad (2)$$

In principle, if accurate values for ΔE_{vib} were available, the experimental lattice energy could be derived from the experimental sublimation enthalpy via equation (1). Alternatively, if accurate

solid and gas phase heat capacity values as a function of temperature were available – along with solid phase change enthalpies where applicable, allowing for the sublimation enthalpy to be corrected to zero Kelvin,[18,19] and the zero-point[17,19] ΔE_{vib} was either negligible or known, the experimental lattice energy could also be derived using equation (1). However, in the absence of this information, it is typically necessary to derive a more approximate estimate of experimental lattice energy using equation (2).

Hence, sublimation data may be used to evaluate lattice energy calculations, e.g. via comparing them to experimentally estimated lattice energies.[2,20–22] In general, physics based methods are employed for calculating lattice energies from crystal structures. These methods include those which calculate energies using molecular force-fields, *ab initio* approaches, such as density functional theory (DFT), and hybrid approaches.[23]

Molecular force-fields[24] calculate intramolecular energetic contributions in terms of parameterized terms reflecting the deviation of bond lengths and angles from local energy minima and intermolecular energetic contributions in terms of parameterized terms reflecting different kinds of contributions to intermolecular forces, with functional forms based upon knowledge of the relevant physics.[20,25–27] For example, the contribution to the intermolecular energy arising from the charge distribution of the isolated molecule might be approximated as the pairwise summation of atomic partial charges interacting via Coulomb's law.[20] The parameters assigned are contingent upon the "atom types" involved in the intermolecular or intramolecular interactions, where each "atom type" represents a given (set of) elements in a defined chemical environment.[20] For example, the force-field parameters might include different atomic partial charges, for calculating Coulombic interactions, for different atom types.[20] Sublimation data may be used to help parameterize / optimize force-field approaches for

calculating lattice energies from crystal structures.[26,27] However, the parameters used in these force-fields may be determined using a variety of calculated and experimental data and the parameters in existing force-fields may be refined using additional data.[20,28]

In general, DFT calculations[24,29,30] are based upon computing ground state electron densities, for a given solid state or molecular structure, and the corresponding electronic energy using a density functional which maps between them. Under certain simplifying assumptions, this density functional can be derived from first principles. A known weakness of many DFT density functionals is their failure to properly capture dispersion (van der Waals attraction) interactions in molecular solids. Hence, DFT calculations of lattice energies[17,22,31] commonly apply a dispersion correction to the energy calculated using the density functional. Depending upon the choice of density functional, dispersion correction or force-field, literature studies suggest DFT calculations may or may not outperform force-field calculations of lattice energies.[31] However, the computational overhead for DFT calculations of lattice energies is much higher than for force-field calculations.

For crystals comprising a single molecular component, the lattice energy may be computed as per equation (3).[22] In equation (3), E_c denotes the energy per mole of unit cell, Z the total number of molecules in the unit cell and E_g the energy per mole of isolated gas phase molecule.

$$LE = \frac{E_c}{Z} - E_g \quad (3)$$

These energetic terms may be computed using molecular force-fields or DFT calculations. In all cases, static structures are assumed. If it can further be assumed that the experimental crystal structure represents the lowest energy conformation of the solid state and that the lowest energy

gas phase geometry is approximately the same as the crystal structure molecular geometry,[32] a force-field calculation of lattice energy can be obtained via only computing the intermolecular contributions to E_c. These assumptions are made in the HABIT program[33] and its extensions (HABIT95,[34] HABIT98 and Visual HABIT),[14] as well as by other force-field protocols for calculating lattice energies.[26] However, even if changes in molecular structure in the gas phase are ignored, different lattice energy calculation protocols may also differ in the extent to which the input crystal structure geometry is optimized / relaxed to a local energy minimum with respect to the potential energy surface estimated from the calculation.[11,20] Whilst the lattice energy is defined with respect to the energy minimized lattice structure according to the true potential energy surface, it must be remembered that the calculations only provide an approximation of this potential energy surface. It is possible that the true energy minimized structure lies closer to the experimentally observed structure, allowing for some deviation due to thermal expansion and experimental error, i.e. complete relaxation of the crystal structure geometry may not yield the most accurate lattice energy calculation.

Hence, as well as the means of calculating energetic contributions (e.g. one out of many possible DFT methodologies or one out of many molecular force-fields), physics based calculations of lattice energies may also differ with respect to how the crystal and molecular structures are processed. Other variations, such as the choice of atomic partial charges to be assigned for force-field calculations, are also possible.

The primary goal of the present work was to thoroughly evaluate a range of possible protocols for computing the lattice energies of single component molecular crystals, using computationally inexpensive force-fields, which could be implemented via automated workflows using the industry standard Materials Studio software, so as to develop a recommended workflow for

automated lattice energy calculations that was appropriate for general organic and pharmaceutical materials.[35] To ensure a thorough evaluation was carried out, not only were a variety of possible force-fields selected, but a range of options, such as the optimization protocol, were also evaluated. To ensure a robust assessment of the performance of each protocol was performed, these protocols were evaluated using a new, large dataset, comprising molecular crystal structures linked to experimental lattice energy estimates derived from sublimation enthalpy data. We make this dataset publicly available for future evaluations of lattice energy calculation protocols. At the time we submitted this work for publication, this was the largest reported dataset for evaluations of lattice energy calculations from crystal structures and it remains the largest dataset used for evaluation of the methods for lattice energy calculation considered herein. Whilst an even larger dataset was reported by Gavezzotti and Chickos whilst we were revising our work for publication,[36] they evaluated different lattice energy calculation routines to the industry standard methods investigated here.

Since we were interested in the most suitable protocol for calculating the lattice energies of pharmaceutically relevant crystals, analyses were repeated on subsets of the data putatively deemed to be more pharmaceutically relevant, according to different criteria. We further develop heuristic recommendations as to when lattice energies calculated using the best force-field protocol should be treated with caution. Subsequently, a comparison was made between the best force-field protocol and a modestly sized, chemically diverse subset of the new dataset, of putative pharmaceutical relevance, and a dispersion corrected DFT lattice energy protocol which was implemented within the publicly available CASTEP software[37] when these studies were performed. Finally, we evaluated the best force-field protocol identified in our work using small datasets previously proposed in the literature for evaluating calculations of lattice energies /

sublimation enthalpies. These evaluations highlight the need for further comparative assessments of the different approaches, which would be supported by the large dataset we have made available.

Methods and Data

To assist the reader, a glossary of technical terms is provided in Supporting Information Table S1.

SUB-BIG Dataset

A dataset comprising 235 crystals, documented using their Cambridge Structural Database (CSD) refcodes,[38] linked to 434 experimentally derived sublimation enthalpies, along with the corresponding temperature, was curated. Each crystal corresponded to a unique chemical, i.e. none of the crystals was a polymorph or redetermination of any other dataset entry.[38] As is further discussed below, in the context of data quality, only a few sublimation enthalpies were confirmed to correspond to the same polymorph as the linked CSD refcode. Sublimation enthalpies reported at temperatures other than 298-298.15 Kelvin, i.e. 298 (0dp) Kelvin, were corrected to 298 Kelvin using the 3rd equation from Acree and Chickos.[39] (Where experimental solid state heat capacities at 298 Kelvin – or very close to this – were available, these were used to apply this correction. Otherwise, the group contribution estimate of Acree and Chickos was applied.)[39] This dataset was further processed to link each of the 235 crystals to an experimentally derived lattice energy estimate (see below). Here, this dataset is denoted "SUB-BIG" and is available in the Supporting Information as an Excel file (SUB-BIG_rev.3.xlsx), which links each sublimation enthalpy datapoint to the literature reference, the corresponding

CSD refcode and metadata required to assess the quality of this datapoint and the assignment (see below), along with the processed version of the dataset, as a CSV file (SUB-BIG_rev.3_FFevalInput.csv), linking each crystal structure to a single experimental estimate of lattice energy. Finally, the results of all force-field calculations, alongside the experimental lattice energy estimates and relevant metadata, including classification according to data quality and pharmaceutical relevance (see below), are also available from the Supporting Information as a single file (SUB-BIG_Integrated_Experi_Calc_Info.xlsx).

As was required by our implementation of the force-field lattice energy protocols evaluated in this work (see below), all crystal structures corresponded to a single molecular component, with at most a single molecule in the asymmetric unit, i.e. $Z' \leq 1$.[40] (Whilst most published organic crystal structures correspond to $Z' \leq 1$, the percentage of $Z' > 1$ crystals is rising over time, from 8% in 1970 to nearly 16% in 2012, with increasing interest in high Z' and co-crystal structures within the pharmaceutical industry.)[41] All crystal structures corresponded to organic molecular structures, albeit one entry (CSD refcode TPHBOR01) contained Boron, hence may be considered inorganic by some molecular modelling software programs.[42]

At the time this work was submitted for publication, the SUB-BIG dataset was, to the best of our knowledge, larger than all previously reported datasets integrating experimental sublimation enthalpies (or experimentally estimated lattice energies) and crystal structures.[2,9,17,19–22,25–28,31,43–52] (Larger datasets were previously reported for molecular QSPR studies of sublimation enthalpies, but those datasets were not integrated with crystal structures.[53,54]) However, in the course of revising our manuscript for publication, Gavezzotti and Chickos reported an even larger dataset than SUB-BIG.[36]

An expanded description of SUB-BIG is provided in Supporting Information section A.0.

Data Quality Assignments for SUB-BIG Dataset

Data quality labels were assigned to the sublimation enthalpies, at 298 Kelvin, for the SUB-BIG dataset and the linked crystal structure. These were assigned based upon a set of heuristics reflecting the confidence in the originally curated sublimation enthalpy value, i.e. the 298 Kelvin correction was not accounted for as this kind of correction is commonplace in the experimental literature, as well as both the inherent quality of the crystallographic data *and* the degree of confidence to which it could be linked to the sublimation enthalpy, given known polymorphism for the chemical in question. (In only six cases was it possible to determine that the experimentally tested polymorph corresponded to the linked crystal structure. However, only 110 datapoints, for which the experimentally tested polymorph could not be confirmed to correspond to the linked crystal structure, were linked to crystal structures for which polymorphism was observed via analysis of automatically identified polymorphs in the Best R-factor list[55] or, where necessary, manual inspection of all relevant members of the refcode family in version 5.38 of the CSD.[38] Moreover, previous studies indicate this would typically introduce a small degree of uncertainty into the experimentally estimated lattice energies for the corresponding crystal structures. Recent computational analysis suggests that most pairwise differences in lattice energies between polymorphs are typically less than 2 kJ/mol and are greater than 7.2 kJ/mol in only 5% of cases.[56] Earlier analysis of experimental data also suggested sublimation enthalpies of polymorphs typically differ by only a few percent.)[26] The quality labels assigned based upon the curated sublimation enthalpy and linked crystal structure were then combined to obtain an overall quality ranking (1 – 8) for the linked sublimation enthalpy at 298 Kelvin, with lower

values indicating higher confidence in the assignment of that value to the crystal structure and the crystallographic data.

Regarding the quality labels assigned based upon the curated sublimation enthalpy, data were deemed of higher quality if the experimental error, by which we mean any quantitative uncertainty estimate, and experimental method were both available and curated and if the experimental error did not exceed 4.9 kJ/mol. (It has been suggested that typical errors in experimental sublimation enthalpies are of the order of 4.9 kJ/mol.[22] However, as noted by Červinka et al.,[19] different approaches to estimating sublimation enthalpy are associated with different quantitative estimates of uncertainty. Moreover, different studies may estimate these uncertainties in different ways.)[19] If either the experimental method or error were not curated, or comments from the cited references suggested the datapoint might be considered unreliable, the datapoint was judged of uncertain quality, with the absence of a curated experimental error deemed to make the datapoint less reliable. If the experimental error exceeded 4.9 kJ/mol, the datapoint was deemed of lower quality.

Regarding the quality labels assigned to the linked crystal structures, these took account of whether the assignment was uncertain due to known polymorphism, the size of the R-factor and whether the crystal was missing hydrogens or contained disordered atoms, albeit the automated preparation of crystal structures performed in the current work was designed to take fix the latter problems and no crystal structures with other kinds of documented disorder were included in the dataset.

A full description of how these data quality labels and the final data quality ranks were assigned is provided in Supporting Information section A.1.

Derivation of Experimental Lattice Energy Estimates for SUB-BIG Dataset

The workflow described in Supporting Information section A.2 was applied to resolve all sublimation enthalpies values reported at 298 Kelvin, or corrected to 298 Kelvin (see above), associated with a given CSD refcode in SUB-BIG into a single (average) sublimation enthalpy value at 298 Kelvin. In brief, lower quality datapoints (see above) and statistical outliers were excluded prior to averaging. This average sublimation enthalpy was then converted into an approximate, given the necessary crude assumptions,[2,17] experimental estimate of lattice energy, using equation (2). The single data quality rank (see above) of the averaged sublimation enthalpies was assigned to the experimental estimate of lattice energy.

Parsing SUB-BIG Dataset for Calculations

Crystal structures were prepared for force-field and DFT calculations as summarized in **Figure 1**. Further details, including how SMILES for the corresponding molecular structures were generated for additional analyses, are provided in Supporting Information section A.3.

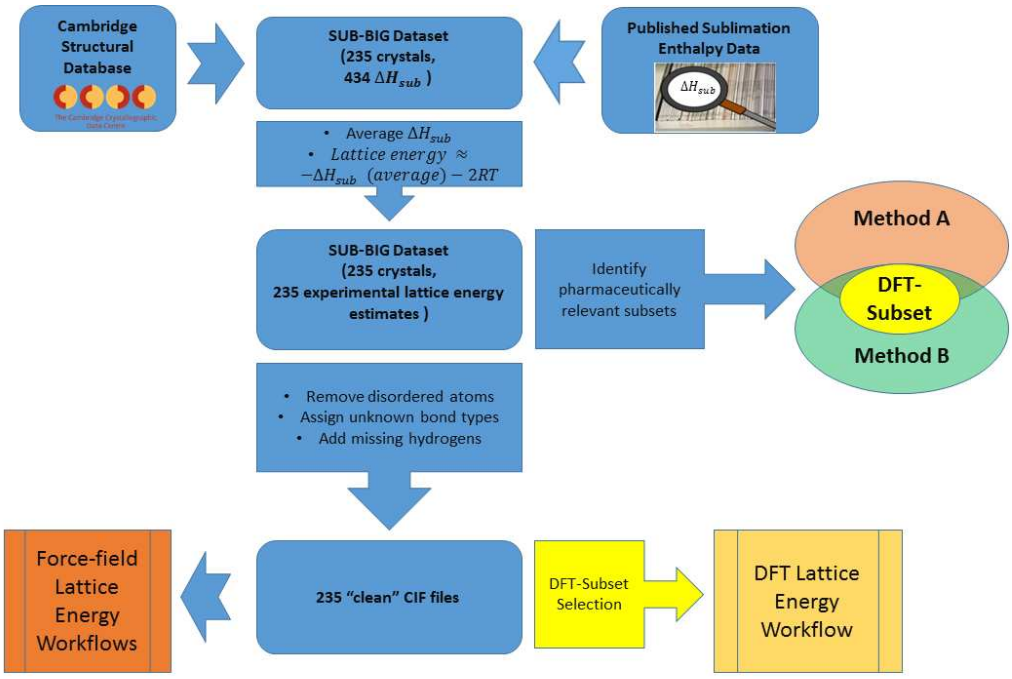

Figure 1. Preparation of SUB-BIG dataset for evaluation of different lattice energy calculation protocols using different subsets of the dataset. All dataset preparation steps, following curation of the original 235 crystal structures linked to 434 sublimation data points, were automated using publicly available Python scripts,[57] with "cleaning" of the crystal structures and filtering of Z' > 1[40] and multicomponent crystals – not applicable for SUB-BIG – carried out using the CSD Python API.[58]

Force-Field Calculations of Lattice Energy

Whilst 324 different protocols were evaluated, corresponding to variations in the details of each step, all protocols corresponded to the generic workflow in **Figure 2**. A variety of different options were explored: the choice of force-field (DREIDING,[59] Universal,[60] CVFF,[61] PCFF,[62] COMPASS,[20] COMPASSII);[28] the optimization routine for the crystal structure (if any) and extracted gas phase molecule; charges assigned; van der Waals cut-offs, Ewald summation accuracy and convergence

criteria ("quality").[63,64] Full details are given in Supporting Information section A.4. N.B. We note that some of these force-fields, such as COMPASSII, may be iteratively refined between different versions of Materials Studio, although earlier iterations can also be accessed, as explained in the documentation.[35] In the current work, we employed Materials Studio version 17.1.0.48 for all force-field calculations.

Our automated workflow for applying these protocols was only applicable to single component molecular crystals, further restricted to Z' <= 1.[40] We note that this does not reflect intrinsic limitations of Materials Studio,[35] but reflects the manner in which we implemented the calculations.[57] We further note that we only investigated local optimization of the gas phase geometry. Whilst this limits the computational overhead and is a reasonable approximation in many cases, since the crystal conformer is typically energetically similar to the gas phase global minimum, lattice energy is defined with respect to the global minimum[8] and this may sometimes be 25 kJ/mol lower than the crystal conformer.[32]

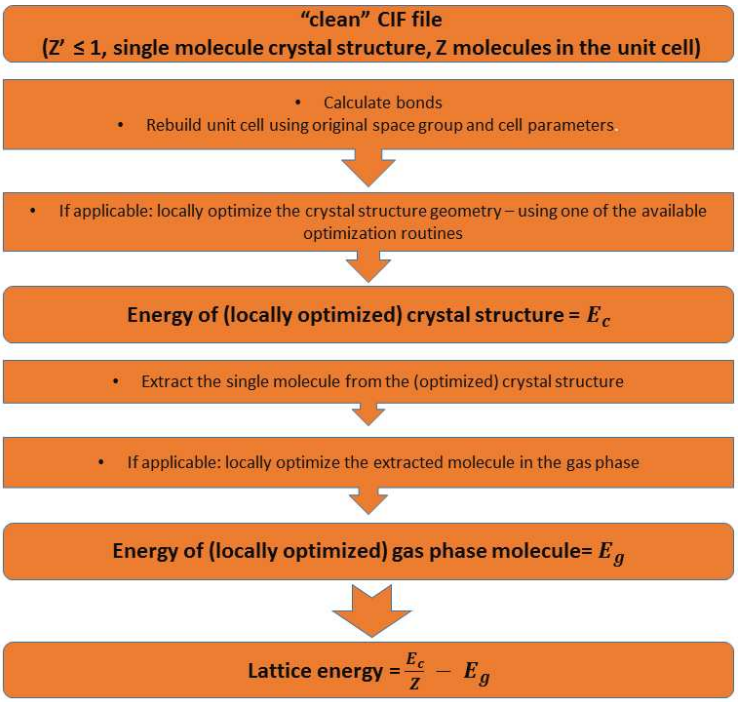

Figure 2. A generic workflow for the 324 force-field protocols evaluated using the SUB-BIG dataset and various subsets thereof. The variations giving rise to 324 protocols are described in Supporting Information section A.4. Importantly, due to the manner in which they were implemented as a script[57] for Materials Studio,[35] these protocols were not applicable to structures with Z' > 1[40] or more than one molecular component, e.g. co-crystals or salts.

Statistical Assessment of Lattice Energy Calculations

The predictive performance of the lattice energy calculations was assessed in terms of the root mean-squared error (RMSE), the coefficient of determination (R^2), mean absolute deviation (MAD), Pearson's correlation coefficient (r) and Spearman's rank correlation coefficient (ρ).[31,65] In addition, the 90% confidence intervals for the distributions of signed and unsigned errors obtained with the force-field protocols, for different scenarios, are provided in the Supporting Information (see below). For all force-field comparisons on the same sets of data, the protocols

were ranked in order of increasing RMSE, which is equivalent to ranking in order of decreasing R^2.[65] See Supporting Information section A.10 for further details.

Principles for Assessing Pharmaceutical Relevance

Two different approaches (A and B) were investigated for assessing the pharmaceutical relevance of entries in the SUB-BIG dataset. Importantly, for both approaches, pharmaceutical relevance was assessed with respect to whether the lattice energy – structure relationships observed within the subset of SUB-BIG classified as pharmaceutically relevant were expected to reflect the relationships observed for active pharmaceutical ingredients (APIs) or API candidates in drug development. (By lattice energy – structure relationships, we mean the molecular characteristics contributing to lattice energies and the manner in which they contribute. These could be different for pharmaceuticals as compared to general organic chemicals, e.g. due to the prevalence of particular molecular sub-structures or other changes which affect the kinds of intermolecular synthons formed and the strength of their contributions towards the lattice energy.) Any other considerations, such as reactivity and biological activity,[66] were not deemed relevant, as the aim here was to ensure that the evaluation of lattice energy calculation routines would be relevant for the kinds of chemicals being assessed during drug development.

Pharmaceutical Relevance Approach A

A Support Vector Machine (SVM) binary classification model,[67–69] employing a Tanimoto kernel[70] with molecular fingerprints chosen to best capture lattice energy – structure relationships, was trained to differentiate drugs (compounds found in DrugBank)[71–73] from non-drugs (not present in DrugBank). (A range of off-the-shelf molecular fingerprint types and parameters

controlling the computation of those fingerprints were evaluated in terms of the predictive ability of a Support Vector Regression[74] model for experimentally estimated lattice energies built using those fingerprints. The fingerprint which gave a model which was most predictive of experimentally estimated lattice energy was deemed most able to capture structural features of relevance to lattice energy. Full details regarding fingerprint selection are provided in Supporting Information section A.6.3.) Predicted drugs were deemed pharmaceutically relevant. A justification for this approach, in terms of its ability to differentiate chemicals exhibiting different lattice energy – structure relationships, is presented in Supporting Information section A6.4. A fuller explanation of how this method was implemented and its ability to correctly identify drug-like chemicals is presented in Supporting Information section A.6.6.

Pharmaceutical Relevance Approach B

SUB-BIG entries were deemed to be pharmaceutically irrelevant if their values for molecular weight and the nConf20 flexibility descriptor[75] were outside the typical range of values for a representative subset of the GlaxoSmithKline (GSK) in-house crystal structure database: molecular weight = 212 – 586, nConf20 = 0 - 37. The analysis which led to the selection of these descriptors and a complete explanation of method B are provided in Supporting Information sections A.6.5. and A.6.7. respectively.

Assessment of the Suitability of Pharmaceutical Relevance Approaches

Analyses, which are fully described in Supporting Information section A.6, provided partial evidence in support of different structure – lattice energy relationships for chemicals predicted to be drug-like vs. non-drug-like by method A. They more clearly indicated different structure –

lattice energy relationships when the available data were partitioned using the median value for the descriptors employed for method B, albeit different thresholds for these descriptors were employed to filter SUB-BIG entries.

Density Functional Theory Calculations of Lattice Energy and Vibrational Energy

The DFT lattice energy calculations are summarized in **Figure 3**, with full details in Supporting Information section A.7. The extent to which incorporating DFT calculations of vibrational energy improved the consistency between calculated and pseudo-experimental lattice energies, as per Reilly et al.,[22] was investigated for a few data points (see Supporting Information section A.8).

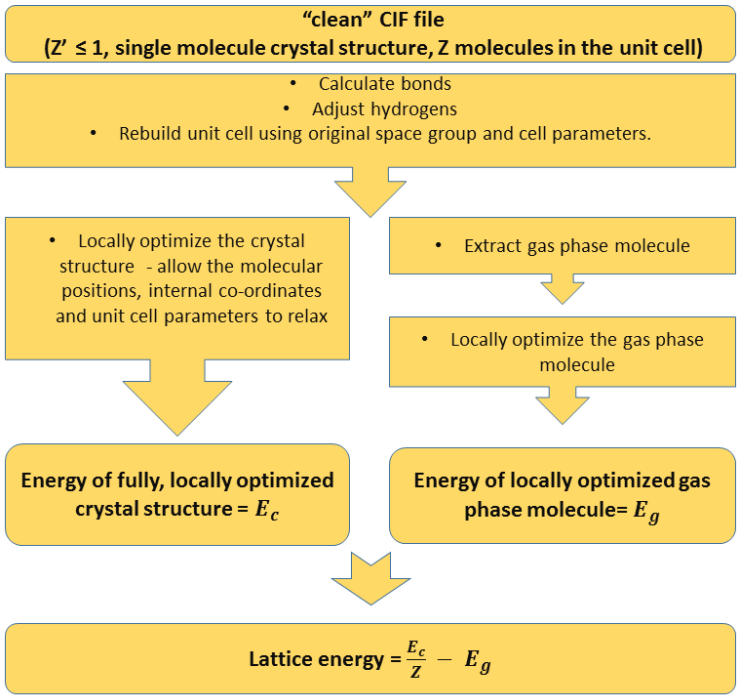

Figure 3. A simplified view of the DFT lattice energy calculation protocol evaluated in this work. (Full details are provided in Supporting Information section A.7.) This protocol was applied using the widely employed CASTEP software (standalone version 16.11).[37,76] Energies, both for

optimizations and single point calculations, were computed using the PBE density functional[77] in combination with the Tkatchenko and Scheffler[78] dispersion correction (PBE+TS).

DFT-Subset

Due to the much higher computational overhead of DFT lattice energy calculations, it was not practical to perform these calculations for all SUB-BIG entries. Hence, a chemically diverse subset of 17 crystal structures was selected from the SUB-BIG entries which were deemed pharmaceutically relevant according to method B. (Full details are provided in Supporting Information section A.9.) This subset is denoted the DFT-Subset (**Figure 4**). Their diversity is reflected in their nConf20 values, along with other descriptors calculated as per Supporting Information section A.6.7., molecular weights, density and whether they contained hydrogen bonds, according to the CSD Python API (*entry.crystal.hbonds()*).[58] Full details are provided in the Supporting Information ("SI_Electronic_Files\SUB-BIG_Dataset\DFT-Subset").

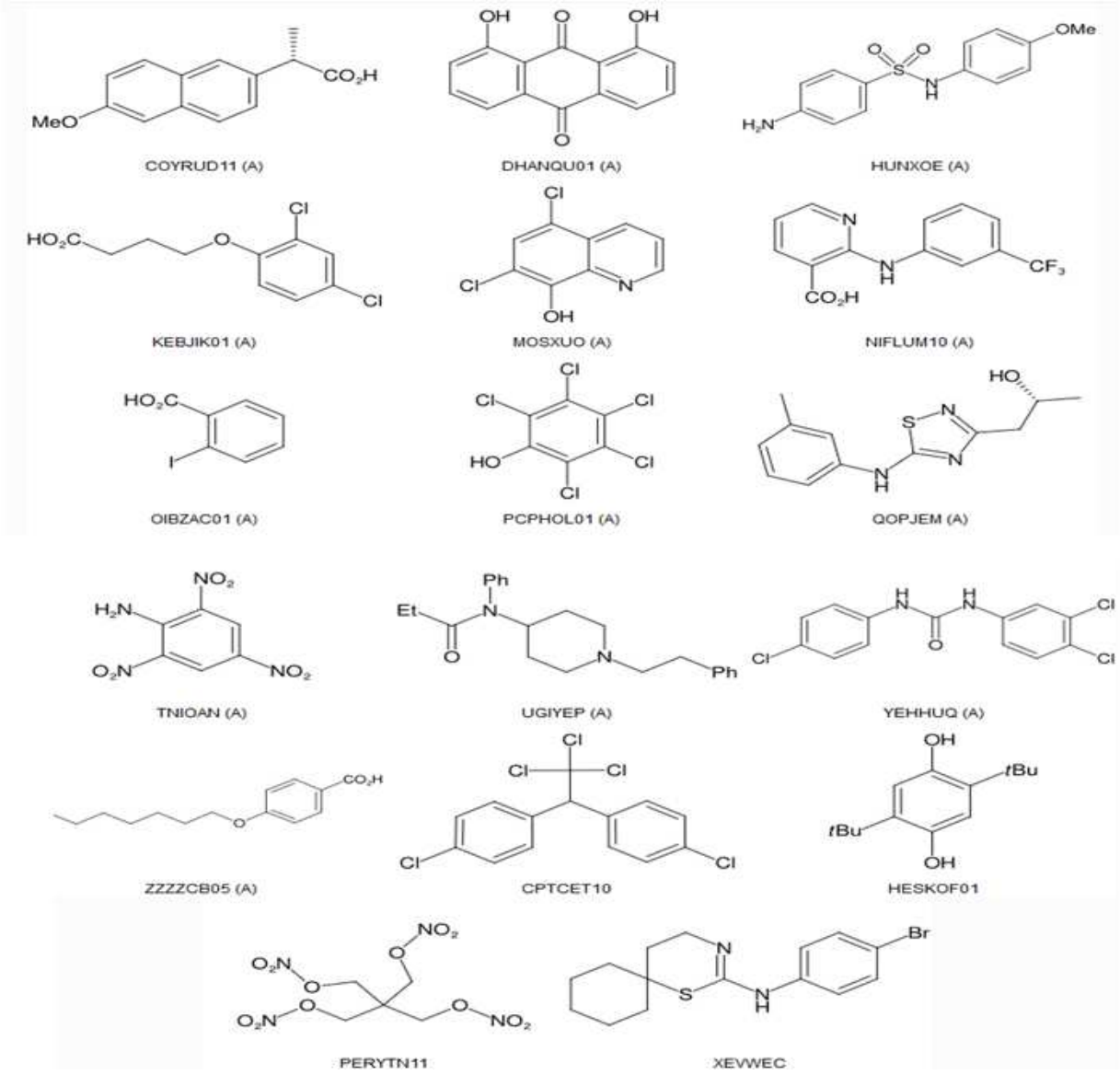

Figure 4. The molecular structures of all entries in the DFT-Subset, labelled by their corresponding CSD refcodes. All entries were deemed pharmaceutically relevant according to method B (based upon consideration of their molecular weight and flexibility). The 13 entries also deemed pharmaceutically relevant according to method A (based upon their similarity to marketed drugs) are annotated accordingly. These images were generated from the SMILES derived from the CSD refcodes (Supporting Information section A.3) using CDK Depict.[79]

Computational Details

A complete description of the versions of all software used for all calculations is provided in Supporting Information section B, along with detailed instructions required to reproduce all results. The scripts and Jupyter Notebooks which are required to reproduce these results have been made publicly available.[57]

Results and Discussion

The results of all force-field and DFT lattice energy calculations are provided in the Supporting Information: "SI_Electronic_Files". The key findings obtained from analysis of those predictions are presented below.

Comparative Performance of Force-Field Protocols and Selection of the Recommended Force-Field Protocol

It should be noted that various calculations failed, either because Materials Studio was unable to process the structure – for CSD refcode HEXDEC03 – or because the lattice energy calculation workflow failed at some point, for some of the force-field protocols. In the latter case, different protocols failed for different structures. (The one Boron containing structure – CSD refcode TPHBOR01 – could only be parsed when the Universal[60] or Dreiding force-fields[59] were employed. This may reflect the fact that these force-fields were designed to offer broad coverage of chemical space. The fact that the COMPASSII, COMPASS, PCFF and CVFF calculations failed reflects the more specific atom type definitions employed by these force-fields. When trying to run calculations using each of these force-fields for the TPHBOR01 structure, following

the bond calculation step in the Materials Studio GUI, Materials Studio reported that a force-field type could not be assigned. We note that the COMPASSII force-field does, at least, include an atom type for tetravalent Boron.[20,28] However, the Boron atom in TPHBOR01 is trivalent.) Since different protocols failed for different structures, only the subset of 210 SUB-BIG entries for which none of the calculations failed was used to compute performance statistics when comparing all 324 protocols.

The predictive performance statistics, across these 210 structures, for the five top ranking (lowest RMSE values) force-field protocols are presented in **Table 1**. (The y-scrambling analysis presented in Supporting Information section C.4 confirmed that the empirical performance of the best protocol was not an artefact of having evaluated a large number – 324 – of protocols.) All of these protocols involved the COMPASSII,[28] or COMPASS,[20] force-fields, force-field assigned charges and at least partial relaxation of the crystal structure. A fuller summary of the relative performance of the different force-field protocols is provided in **Figure 5**. It should be noted that the ranking of the protocols according to RMSE and Spearman's rank correlation coefficient is not always identical. Hence, some protocols may be better for screening crystals according to the strength of solid state interactions than for predicting accurate lattice energy values. Nonetheless, the top ranking protocol according to RMSE is ranked 3rd when the protocols are ranked according to Spearman's coefficient and the Spearman's coefficient of the new 1st ranked protocol is indistinguishable when the values are rounded to two decimal places.

Regarding the application of the calculation quality setting, we found that, across these 210 structures, the ultra-fine setting consistently increased the average calculation time, albeit typically by less than one second. However, it only improved the RMSE 101/162 times and all performance statistics were only consistently improved 69/162 times. In the case of the top

ranking protocol from **Table 1**, the corresponding protocol employing the medium calculation quality setting was worse in terms of all performance statistics across these 210 structures, albeit with an RMSE increase of only 1.12 kJ/mol and a corresponding increase in average computational time of less than one second. These findings reflect the tighter convergence limits for these calculations (see Table S5), as illustrated for some examples in Supporting Information section C.14.

Full performance statistics for all protocols, along with the 90% confidence intervals for the distributions of all signed and unsigned errors, are presented in an Excel file (SUB-BIG_PerfStatsAllFFs.xlsx) in the Supporting Information. In addition, the Supporting Information provides performance statistics obtained for all evaluated subsets of SUB-BIG (see the Excel files referred to under Supporting Information Sections C.1 – C.2, Tables S8-S9). When only the SUB-BIG entries considered pharmaceutically relevant according to method A are considered (see Figure S1), the protocol with the lowest RMSE value obtained without filtering SUB-BIG (top row of **Table 1**) remains the top ranking protocol. When only the SUB-BIG entries considered pharmaceutically relevant according to method B are considered (see Figure S2), that protocol is ranked 14th out of 324. The precise ranking, according to RMSE, of this protocol (see Figures S3 – S5) is further changed upon restricting consideration to those data judged of highest quality according to the heuristic data quality rankings assigned to every pair of crystal structure and estimated experimental lattice energy in SUB-BIG. (Since this appreciably reduces the size of the relevant subsets – see Figure S6, the rankings may not be as robust. The different rankings observed when filtering according to data quality may also, in some cases, partially reflect changes in coverage of chemical space as well. Similar molecular weight distributions – see Figure S6 – but appreciably different nConf20^{75} – Figure S7 –

distributions are observed when the entries deemed pharmaceutically relevant according to method B are filtered according to data quality.) Nonetheless, this protocol remains within the top 15 (top 5%) of all protocols for all evaluated subsets of SUB-BIG.

Table 1. Top ranking five of 324 force-field protocols. Evaluated on all 210 SUB-BIG entries for which none of the calculations failed. All protocols are ranked in order of increasing RMSE (decreasing R^2, the coefficient of determination for the predictions). All top ranking protocols involved relaxation (local optimization) of the molecular positions and internal co-ordinates within the crystal structure and force-field assigned charges. The details for the recommended protocol (lowest RMSE) are underlined.

Force-field	Unit cell parameters relaxed?	Gas phase molecule relaxed?	Quality	RMSE (kJ/ mol)	MAD (kJ/ mol)	R^2	r	ρ
<u>COMPASSII</u>	<u>FALSE</u>	<u>TRUE</u>	<u>Ultra-fine</u>	<u>17.36</u>	<u>11.23</u>	<u>0.62</u>	<u>0.82</u>	<u>0.87</u>
COMPASSII	TRUE	TRUE	Ultra-fine	17.84	10.98	0.60	0.81	0.85
COMPASS	FALSE	TRUE	Ultra-fine	18.13	11.90	0.59	0.80	0.84
COMPASSII	TRUE	TRUE	Medium	18.17	11.17	0.59	0.80	0.85
COMPASS	TRUE	TRUE	Ultra-fine	18.27	11.49	0.58	0.80	0.84

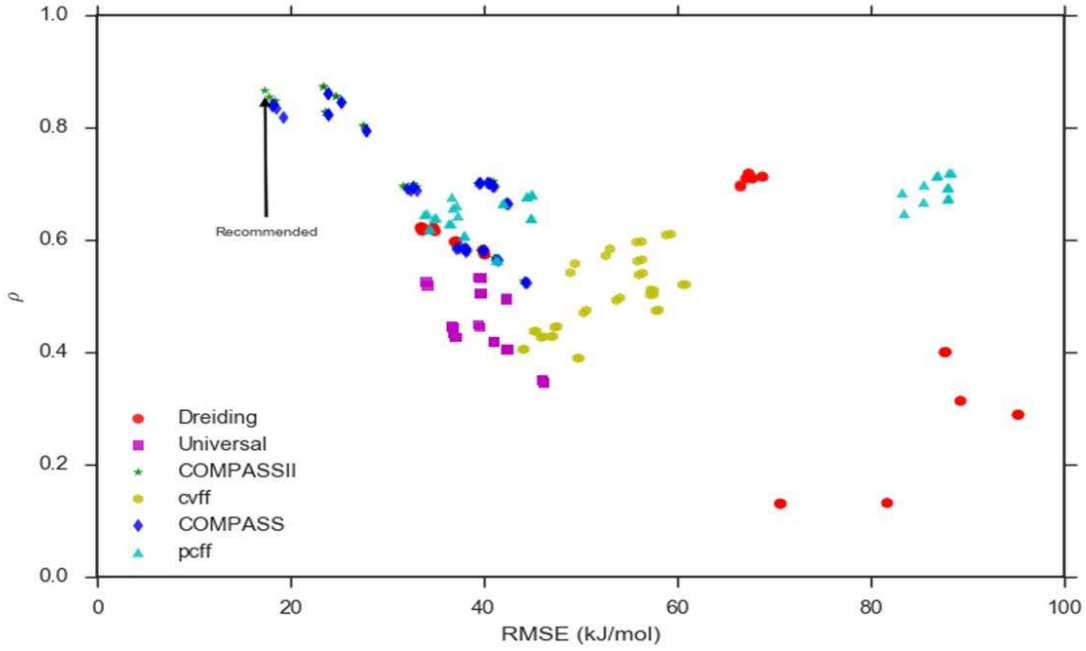

Figure 5. Summary of the relative performance of force-field protocols, corresponding to different force-fields and other protocol variations, on all 210 SUB-BIG entries for which all 324 protocols could compute a lattice energy. Protocols with negative Spearman rank correlation (ρ) or RMSE values greater than 100 kJ/mol are not shown. The recommended force-field protocol is highlighted using an arrow and corresponds to the top ranking protocol from **Table 1**.

Recommended Force-Field Protocol

Out of the force-field protocols evaluated in this work, we recommend the top ranking protocol (lowest RMSE) as assessed upon all SUB-BIG entries for which every force-field protocol was able to calculate a lattice energy (**Table 1**), be used for future calculations of lattice energies where the focus is obtaining the most accurate lattice energies. This was also amongst the top 15 (top 5%) of all protocols for all evaluated subsets of SUB-BIG (see the Excel files referred

to in Supporting Information Sections C.1-C.2, Tables S8-S9). Moreover, the rankings obtained using the larger set of data, i.e. without filtering SUB-BIG, can be expected to be more robust.

This recommended protocol corresponds to the following options: COMPASSII force-field,[28] with force-field assigned charges, partial relaxation of the crystal structure geometry (including the positions and internal co-ordinates of the molecules, but not the unit cell parameters), local optimization of the gas phase geometry and the ultra-fine quality setting for Materials Studio[35] calculations (see **Figure 6**).

However, it is not necessarily the case that the partial optimization of the crystal structure carried out by this protocol makes this most suitable for studies of intermolecular synthons or crystallization processes.[11–14] Indeed, Bernardes and Joseph previously found, based upon consideration of different force-field protocols, that the most suitable protocol for accurate lattice energy calculations was not most suitable for evaluating structural characteristics.[25] It is possible that, at least in part, this protocol gave the best empirical results, prior to subsetting the dataset, due to cancellation of errors involved in each step. Nonetheless, we found that allowing the unit cell parameters to relax – i.e. full local optimization – not only typically resulted in the calculated lattice energies becoming more negative and worse performance in terms of most statistical measures (**Table 1**), but could lead to significant distortion of the unit cell in some cases (Supporting Information section C.13). Hence, the constrained relaxation allowed for by the recommended protocol appears to be better for calculating lattice energies and may provide more realistic crystal structures for subsequent analysis.

However, crystal structure prediction[8,80] cannot rely upon fixed, experimental lattice parameters. Hence, the observation that the optimized structures obtained when the recommended protocol was modified to allow complete relaxation of the crystal structure, including the lattice parameters,

could differ significantly to the experimental structures in some cases (Supporting Information section C.13) may mean the force-field settings are not suitable for crystal structure prediction. However, this question is beyond the scope of the present study and requires further investigation.

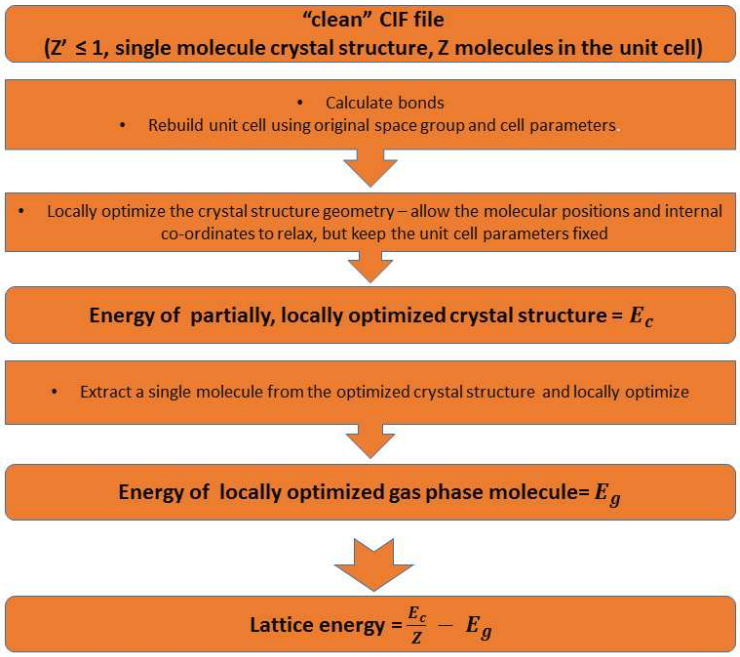

Figure 6. An overview of the recommended force-field protocol, applied to the "clean" crystal structures prepared via the automated workflow presented in **Figure 1**. All energies, both for optimizations and single point calculations, were computed using the COMPASSII force-field,[28] with force-field assigned charges and the ultra-fine quality setting for Materials Studio.[35]

Performance of the Recommended Force-Field Protocol Coupled with Heuristic Filtering Criteria

Analysis of six extreme prediction outliers (absolute errors of more than 100 kJ/mol with the recommended protocol, see Figures S12- S17) in SUB-BIG led to the development of the

following heuristic filtering criteria – designed to flag unreliable calculations obtained using the recommended force-field protocol. (N.B. These extreme outliers caused at least some of the force-field protocols to fail, hence weren't included in the 210 entries used to compare different protocols.) 1. The calculation cannot be trusted if the ostensible number of atoms in the unit cell, reported by the lattice energy code, is inconsistent with the product of the z-value and the number of atoms in the extracted gas phase molecule. (Typically, this was observed to reflect the erroneous addition of extra hydrogen atoms as part of the automated structure preparation steps from **Figure 1**.) 2. Positive or zero lattice energies are unphysical and should be discounted. 3. Negative values less than or equal to -249 kJ/mol are unusual and *may* be unreliable (only one SUB-BIG entry had an estimated experimental lattice energy more negative than this). In all cases, these entries appeared to correspond to problematic preparation of the crystal structures as part of the combined automated steps applied as per **Figure 1** and **Figure 2**. Other analyses were performed to see whether the prediction errors could be rationalized by consideration of differences in molecular weight (Figure S9) – since dispersion contributions were expected to increase with molecular weight and the handling of dispersion by the COMPASSII force-field[20,28] is imperfect,[31] flexibility[75] (Figure S10) – since the application of local optimization of the gas phase molecule would fail to properly handle scenarios where the energy of the global minimum gas phase conformer was significantly different to the crystal structure conformer,[32]and R-factor (Figure S11). However, no clear trends were observed.

The performance of the recommended protocol, coupled with the heuristic filtering criteria, on all applicable SUB-BIG entries, i.e. also excluding those entries for which the calculation failed, is shown in **Table 2**, alongside the performance on different subsets of SUB-BIG. (The structures for which the recommended protocol failed, i.e. only produced an error message rather

than a lattice energy, or produced results which were rejected by the heuristic filtering criteria, were a subset of the structures for which *any* of the 324 force-field protocols failed. Hence, the statistics in **Table 2** were computed on a larger set of 226 structures than the 210 structures used to compare different protocols, as per **Table 1**.) Regarding the variation in performance across different subsets, some of these statistics *might* be affected by artefacts due to variations in the distributions of the data. For example, a model which only successfully reproduced the lattice energies on average would have a lower RMSE if the data were distributed over a narrower range.[2,65](However, it should be noted that the force-field calculations reported in **Table 2** consistently obtain an RMSE less than the standard deviation in the experimental values, i.e. they perform better than a model which merely reproduces the average of the experimental values.[2] The corresponding standard deviations, from Excel 2013, of the experimental values are as follows: 28.57 kJ/mol prior to filtering, 25.22 kJ/mol following filtering according to method A, 29.41 kJ/mol following filtering according to method B, 21.57 kJ/mol for the DFT-Subset.) Conversely, *if* a model consistently produced residuals of a given magnitude, i.e. *if* RMSE was fixed, R^2 would be reduced if the data were distributed over a narrower range.[65] Nonetheless, in spite of the possibility that some of the variation in these statistics might be affected by artefacts, consideration of all statistics suggests that the performance of the recommended protocol is either comparable or somewhat, but not dramatically, worse when moving to subsets of higher putative pharmaceutical relevance, including the DFT-Subset (a diverse subset of entries deemed pharmaceutically relevant according to method B). All experimentally estimated and calculated lattice energies, using the recommended protocol coupled with the heuristic filtering criteria, for all SUB-BIG entries are shown graphically in **Figure 7**.

Additional statistics are presented in Supporting Information Table S11, following filtering of these subsets to only retain the highest quality datapoints (data quality ranking = 1), according to the heuristic data quality labels. Curiously, although all performance statistics improve when filtering the applicable SUB-BIG entries and the subset deemed pharmaceutically relevant according to method A, the performance statistics are often worse when filtering the subset deemed pharmaceutically relevant according to method B or the DFT-Subset, a subset of those deemed pharmaceutically relevant according to method B, according to the heuristic quality labels. However, the statistics calculated on these smaller subsets are arguably less robust and possibly reflect, in part, differences in coverage of chemical space as well as data quality (see Figure S7).

Finally, the distributions of the signed and absolute error distributions, when applying the recommended force-field protocol coupled with the heuristic filtering criteria, are summarized in Figures S18 – S25. These indicate that, depending upon the subset of the data considered, the absolute prediction error has at least a 95% chance of not exceeding 36.6 kJ/mol.

Table 2. Performance of the recommended force-field protocol, coupled with the heuristic filtering criteria, on different subsets of SUB-BIG, including comparison to DFT (PBE+TS) calculations on the DFT-Subset. These statistics were calculated following removal of three structures for which the recommended protocol couldn't generate any result and six structures excluded by the heuristic filtering criteria. Supporting Information Table S11 also presents statistics following filtering of these subsets according to the heuristic data quality labels.

Subset[a]	Number of crystal structures	Method	RMSE (kJ/mol)	MAD (kJ/mol)	R^2	R	ρ

All applicable SUB-BIG entries	226	Force-field[b]	17.17	11.20	0.64	0.83	0.88
Putative pharmaceutically relevant (Method A) subset	164	Force-field	16.83	11.59	0.55	0.81	0.86
Putative pharmaceutically relevant (Method B) subset	62	Force-field	19.80	12.64	0.55	0.75	0.76
DFT-Subset	17	Force-field	15.10	11.98	0.51	0.82	0.83
	17	DFT	37.14	29.00	-1.96	0.65	0.66

a. Entries for which the recommended force-field protocol failed or which did not pass the heuristic filtering criteria were removed. The filtering criteria removed any entries for which the force-field protocol calculated lattice energies which were positive, less than or equal to -249 kJ/mol and/or parsing of the messages generated by the Materials Studio Perl script indicated the automated structure preparation workflow had failed.

b. All force-field calculations refer to the recommended protocol: COMPASSII force-field, with force-field assigned charges, partial relaxation of the crystal structure geometry (including the positions and internal co-ordinates of the molecules, but not the unit cell parameters), local optimization of the gas phase geometry and the ultra-fine quality setting for Materials Studio.

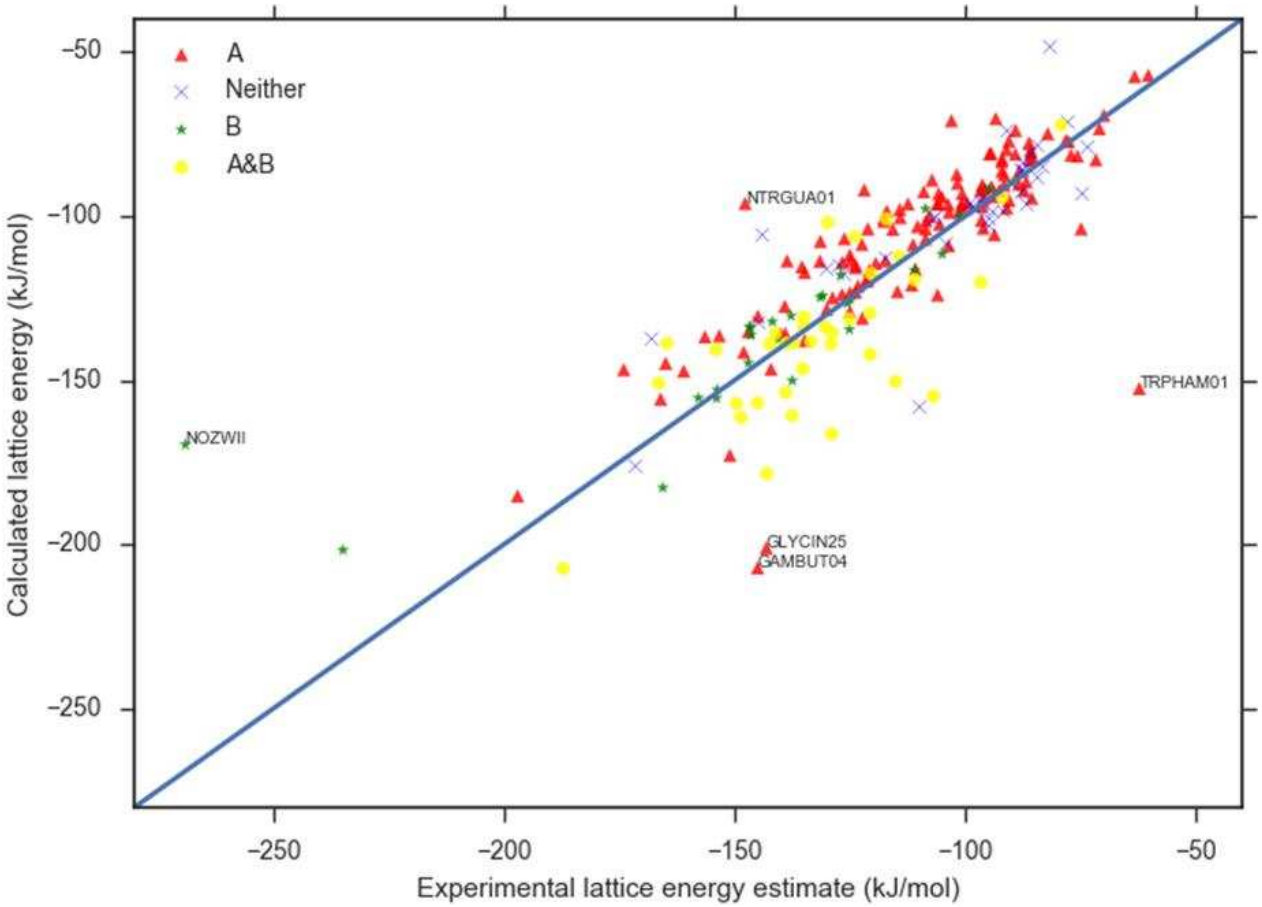

Figure 7. Comparison between experimental estimates of lattice energies and values calculated using the recommended force-field protocol, coupled with heuristic filtering criteria, for the SUB-BIG dataset. The identity line is shown. Different symbols and colors denote whether the corresponding entries were deemed pharmaceutically relevant according to method A, B, both, or neither. Of the significant outliers (absolute error >50 kJ/mol), labelled according to their CSD refcode, only NOZWII has significant experimental uncertainty (data quality rank = 8).

Comparison of Force-Field and DFT Lattice Energy Calculations

As can be seen from **Table 2**, the recommended force-field protocol clearly performs much better than the DFT protocol, in terms of all statistical measures. (As can be seen from Table S11,

this remains the case when the DFT-Subset is filtered to only retain the highest quality datapoints, according to the heuristic data quality labels.) The comparative performance of both protocols is graphically illustrated in **Figure 8**. It is apparent that the DFT protocol tends to significantly overestimate the magnitude of the lattice energy, a bias which is reflected in its RMSE value (37.14 kJmol) being substantially larger than the standard deviation of the estimated experimental lattice energies for the DFT-Subset (21.57 kJ/mol), i.e. the model yields worse absolute predictions than a model which merely reproduced the average experimental value. Still, the positive correlation coefficients are better than would be expected for this so-called null model.[2]

Finally, we note that the dispersion corrected DFT protocol was also outperformed by a variant of the recommended force-field in which the unit cell parameters were also relaxed (Supporting Information Table S12). This protocol is more directly comparable to the DFT protocol, since both protocols employed full local optimization of the unit cell geometry and gas phase molecule.

Our finding, that the dispersion corrected DFT protocol employed herein (PBE+TS),[77,78] was substantially worse at calculating lattice energies than the recommended force-field protocol, prompted recent enhancements to the CASTEP DFT code employed for this work. Specifically, this work resulted in the development of CASTEP's dispersion functionality to include an efficient many body dispersion functionality,[81] Grimme D3.[82] However, the newly developed CASTEP code was not available to us for the DFT calculations reported in the current work.

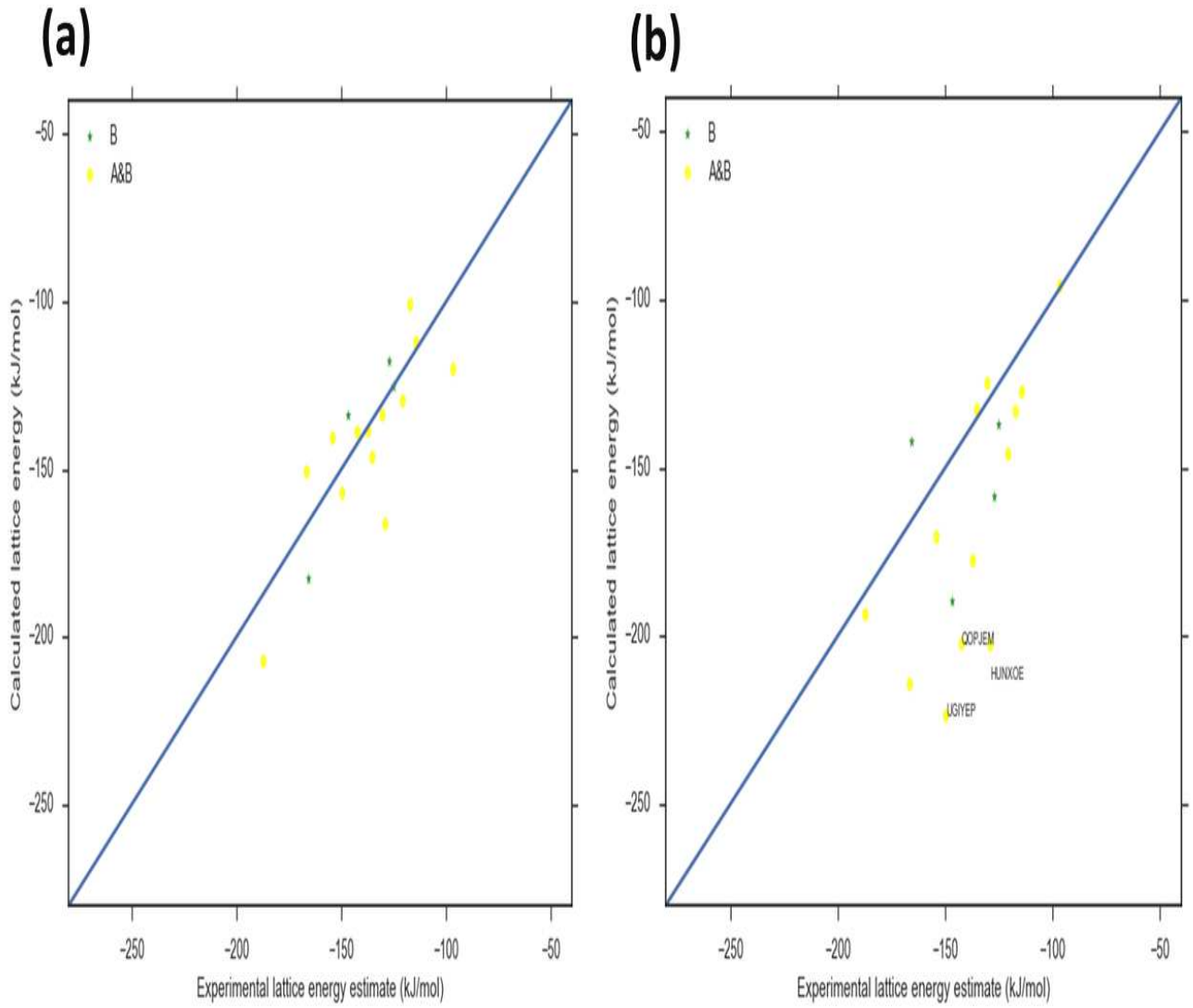

Figure 8. Comparison between experimental estimates of lattice energies and values calculated using (a) the recommended force-field protocol or (b) DFT for all 17 entries in the DFT-Subset. The identity line is shown for both plots. All entries were deemed pharmaceutically relevant according to method B and the results for those entries which were also deemed pharmaceutically relevant according to method A are denoted using yellow circles. Only with DFT predictions are significant outliers (absolute error >50 kJ/mol) observed. Of these outliers, labelled according to their CSD refcode, only UGIYEP has significant experimental uncertainty (data quality rank = 7). The molecular structures of these compounds can be seen in **Figure 4**.

Significance of Vibrational Contributions Calculated Using DFT

As can be seen in Supporting Information Section C.5 (Table S10), the magnitudes of the deviations between the DFT calculated lattice energies and the experimental (or pseudo-experimental) estimates of lattice energy consistently increased when pseudo-experimental lattice energies were derived using DFT computed vibrational calculations.

Comparison of Our Results to the Literature

In order to properly put our results in the context of previous studies which empirically assessed computational calculations of lattice energies or sublimation enthalpies, from crystal structures, we evaluated our recommended force-field protocol on a variety of the smaller datasets employed in previous studies: the X23 (23 crystals),[22,31] SUB-48 (48 crystals),[2] Bernardes et al. (23 crystals)[25] and original COMPASS II validation set (41 crystals).[28] We computed performance statistics on subsets of those datasets for which our recommended force-field protocol calculated lattice energies which passed the heuristic filtering criteria. These statistics were based upon experimental estimates of lattice energy reported within or which could be estimated from those earlier studies. Where possible, we also computed performance statistics for these subsets based upon the calculated lattice energies, obtained with a variety of computational methods, which were reported within or could be estimated from those earlier studies. Full details are presented in Supporting Information section C.10, with the performance statistics summarized in Table S13. These results clearly show that our recommended force-field protocol clearly gives worse performance, in terms of all statistics, than the best dispersion corrected DFT methods reported for the applicable subset of the X23 benchmark.[22] (However, these calculations have a much higher computational overhead than our recommended force-field protocol.) It is also outperformed, in

terms of most statistics, by the OPLS-AA force-field calculations for the applicable subset of the dataset of Bernardes et al.[25] However, its performance on the applicable subsets of the original COMPASS II validation set and the SUB-48 dataset is typically competitive, in terms of most statistics, to DMACRYS[2] and the COMPASS II sublimation enthalpy simulations of Sun et al.[28] (converted to estimated lattice energies herein) respectively.

Moreover, these comparisons are based upon much smaller datasets than the SUB-BIG dataset provided in our work and used to select our recommended force-field protocol (compare the numbers in Table 2 and Table S13). In addition, the subsets of these various datasets, for which the methods were compared, correspond to different distributions of key descriptors, which are expected to be related to different lattice energy – structure energy relationships, as shown in Figures S26 – S27.

We also note that a significantly higher R^2 (0.97) than that obtained on the SUB-BIG dataset using our recommended force-field protocol (0.64 on the 226 entries left after applying the heuristic filtering criteria) was reported more than two decades ago by Charlton et al.,[51] for comparisons of force-field lattice energy calculations using the HABIT[33] program without any structure optimization and atomic charges from semi-empirical quantum calculations. Whilst we were unable to derive their dataset to perform a direct comparison, we note that their dataset (62 crystal structures) was considerably smaller than the SUB-BIG dataset (235 crystal structures) employed for the current study.

Further work is required to more confidently assess the relative merits of these different lattice energy calculation protocols, via comparison using a larger dataset than in earlier studies. The SUB-BIG dataset reported herein, along with the putative pharmaceutical relevance classifications, could support these further studies, as could the even larger dataset reported, in the

course of revising this work for publication, by Gavezzotti and Chickos.[36] Indeed, integration of these datasets, especially if further supplemented with data clearly linked to multiple polymorphs of the same molecules,[25] could yield an enhanced dataset for future investigations.

Finally, we note that our study was concerned with evaluating calculations of lattice energies from crystal structures, focusing on force-field protocols which could be readily implemented using the industry standard Materials Studio software. Hence, models for lattice energy, or the related sublimation enthalpy, based purely on molecular structures were considered out of scope. In principle, calculations based upon crystal structures offer the ability to rank polymorph stability and support crystal structure prediction,[8] albeit the SUB-BIG dataset does not offer insights into this (see SUB-BIG Dataset under Methods and Data). However, the level of predictive performance obtained with the recommended force-field protocol, as estimated using the SUB-BIG dataset (**Table 2**), indicates it is far too inaccurate to capture the typically small energetic differences between polymorphs.[56,83] Indeed, it would be interesting to see how well methods based purely on molecular structure[53,54,84] compare to the force-field protocol recommended herein based upon the SUB-BIG dataset. For example, Cardozo reported a group contribution method with a standard deviation of 15 kJ/mol (experimental vs. calculated sublimation enthalpies) on an, unidentified, set of 115 organic compounds.[84] This is comparable to the value obtained (17 kJ/mol) using the recommended force-field protocol (estimated experimental vs. calculated lattice energies), coupled with the heuristic filtering criteria, on all applicable 226 SUB-BIG entries. However, whether these statistics are directly comparable in light of possible differences in data distributions[2,65] or whether the apparent performance would be comparable in terms of other statistics (**Table 2**) is an open question. It would be particularly interesting to see whether some

inexpensive calculations based upon molecular structure could offer comparable predictive performance to some crystal structure calculations, with reduced computational overhead.

Computational Overhead

The recommended force-field protocol takes approximately two seconds, on average, per crystal structure (based on the arithmetic mean timings for all DFT-Subset entries returned by the Perl *time()* function). In contrast, the dispersion corrected DFT lattice energy calculation took more than a day per crystal structure, on average, and the subsequent vibrational energy calculations took more than three days, on average, per crystal structure. (N.B. The different resources used to perform these calculations are documented in the Supporting Information (Section B).) Hence, also taking into account the reasonable correlation with experimental estimates of lattice energies (e.g. see **Table 2**), our recommended force-field protocol is suitable for screening in-house industrial databases of the order of hundreds (or maybe thousands) of crystal structures, as well as studying lattice energy structure relationships in larger datasets of tens of thousands of crystal structures. Indeed, recent work within industry[85] and ongoing work within the ADDoPT project[86] has investigated lattice energy structure relationships based upon COMPASSII force-field calculations for hundreds and tens of thousands of crystal structures respectively.

Scope for Improving the Recommended Force-Field Protocol

The high Spearman's rank correlation coefficient (0.88) obtained with the recommended protocol, coupled with the heuristic filtering criteria, across all applicable SUB-BIG dataset entries suggests this computationally inexpensive protocol is useful for screening large numbers of crystal structures according to their relative lattice energies. However, the corresponding error estimates

(RMSE around 17 kJ/mol) indicate that this method cannot be used to estimate relative stabilities of polymorphs.[56,83] Estimating the phase diagram requires Gibbs free energy calculations and very high-level estimates of the lattice energy, as recently demonstrated for methanol.[83]

There are several possible avenues which could be explored in future studies to improve the lattice energies calculated using the recommended force-field protocol: (1) improved description of the sublimation process; (2) improved description of the intermolecular and/or intramolecular interactions; (3) improved processing of the crystal structures; (4) Machine Learning modelling of the residuals. Regarding improved description of the sublimation process, the recommended protocol fails to take account of possible changes in tautomeric state[87] and does not fully allow for changes in conformation, which are sometimes significant,[32] during sublimation, i.e. only local optimization of the gas phase molecule extracted from the crystal is allowed. (In practice, experimental sublimation may also correspond to the formation of gas phase clusters, rather than isolated molecules.[19] However, it is arguable that this issue requires correction of the sublimation data,[19] rather than the calculation, given the definition of lattice energy with respect to infinitely separated gas phase molecules.)[8] Nonetheless, the poor correlation observed between absolute prediction errors and the conformational flexibility descriptor nConf20[75] (Supporting Information, Figure S10), suggests the failure to globally optimize the gas phase molecule was not a significant limiting factor here in most cases.

Regarding the need for improved description of the intermolecular and/or intramolecular interactions, it should be noted that, whilst the functional form of the underlying COMPASSII force-field is documented along with the parameterization routine, full details of the parameters were not reported in the accompanying publication.[28] Nonetheless, since this force-field employs a simple Lennard−Jones, pairwise additive, description of dispersion,[20,28] we speculate that the

incorporation of more sophisticated dispersion contributions, as recommended elsewhere,[31] could improve the calculations. We further speculate that the inclusion of atomic partial charges using quantum chemical methods might also yield better predictions.

Regarding processing of the crystal structures, it should be noted that our implementation of the force-field workflows assessed in this work is not capable of handling Z' > 1 structures.[40] It is also clear that the structure pre-processing (**Figure 1**, **Figure 6**) could cause significant errors in some cases (see Supporting Information section C.7.), albeit the heuristic filtering criteria should serve to identify when these errors occurred. Improvements to this automated workflow could avoid chemically incorrect representations of a few structures leading to serious errors in the calculated lattice energies.

Finally, recent work has found that Machine Learning may be used, with some success, to predict the difference between force-field and DFT calculations of lattice energies from knowledge of different crystal structures of the same molecule.[88] Similar approaches might be suitable to learn the error in the force-field calculations of lattice energy presented here. The experimental lattice energy estimates, coupled with the corresponding force-field calculated values, for the SUB-BIG dataset, which we make available in the Supporting Information ("SI_Electronic_Files"), provide a useful starting point for exploring this in future studies.

Conclusions

To the best of our knowledge, we report the most extensive evaluation of different force-field protocol variations for calculating lattice energies presented to date in the literature. Moreover, as well as evaluating a large variety of different force-field protocols, we empirically evaluated their performance on a much larger dataset than previously published studies, which supports

more robust conclusions. This SUB-BIG dataset, comprising 235 molecular crystals linked to sublimation enthalpy data, from which estimated experimental lattice energies were derived, is made publicly available for future investigations. Moreover, all force-field protocols, coupled with preparation of experimental crystal structures, were implemented using an automated workflow employing the Cambridge Structural Database Python Application Programming Interface and the industry-standard Materials Studio. We have made the scripts implementing that automated workflow publicly available.

Based upon our analysis, of relatively inexpensive force-field protocols implemented within the industry standard Materials Studio software, we recommend a force-field protocol based upon the COMPASSII force-field. This recommendation is based upon its having performed best on the unfiltered set of SUB-BIG for which all protocols could calculate a lattice energy, along with at least being amongst the top 15 ranking protocols (top 5% of 324 investigated protocols) following subsetting based upon putative pharmaceutical relevance according to two distinct methods.

We further recommend some heuristic criteria for filtering unphysical (values greater than or equal to zero and/or the messages generated by the automated workflow indicate extraneous atoms were added to the unit cell) or suspicious (values less than or equal to -249 kJ/mol) lattice energy values calculated using the recommended protocol. Application of the recommended force-field protocol, coupled with these heuristic filtering criteria, achieved an RMSE around 17 kJ/mol (MAD around 11 kJ/mol, Spearman's rank correlation coefficient of 0.88) across all 226 SUB-BIG structures retained after removing calculation failures and applying the filtering criteria. Across these 226 structures, the estimated experimental lattice energies ranged from -60 to -269 kJ/mol, with a standard deviation around 29 kJ/mol. The performance of the

recommended protocol on pharmaceutically relevant crystals could be somewhat reduced, with an RMSE around 20 kJ/mol (MAD around 13 kJ/mol, Spearman's rank correlation coefficient of 0.76) obtained on 62 structures retained following filtering according to pharmaceutical relevance method B, for which the distribution of experimental values was similar.

We determined that our recommended protocol outperformed the dispersion corrected DFT protocol investigated herein (PBE+TS), based on a diverse subset of SUB-BIG (17 molecular crystals) of putative pharmaceutical relevance (method B). This work prompted further development of the CASTEP DFT code, employed for this work, to improve its dispersion functionality.

Comparison of the predictive performance of our recommended force-field protocol, coupled with the heuristic filtering criteria, to the performance of other lattice energy calculation protocols reported in the literature for much smaller datasets suggested some other calculation routines could provide more reliable lattice energy estimates. However, some of these calculation routines have a significantly higher computational overhead, hence are less suitable for obtaining quick estimates of lattice energy or screening large databases. Moreover, further work is required to assess whether the relative performance of the methods would remain unchanged when assessed on larger datasets. In addition, there is scope for improving our recommended force-field protocol, building upon the results presented in our current work. The SUB-BIG dataset we make available here would provide a basis for future studies aimed at more robust comparison to other methods as well as improvements to our recommended force-field protocol.

Supporting Information

Supporting Information Available:

1. Additional details regarding the methods and data used to generate our results (section A), detailed instructions on how to reproduce our results using the code (Perl scripts, Python scripts, Jupyter Notebooks) which is made available on Zenodo[57] (section B), and the results (section C). (PDF)

2. A ZIP file containing a folder ("SI_Electronic_Files") populated with dataset files, intermediate files and results files, documented in a README_SI_Electronic_Files.xlsx file contained within this ZIP file, all of which are made available under the terms of the Creative Commons Attribution-NonCommercial 4.0 International license (https://creativecommons.org/licenses/by-nc/4.0/legalcode).

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. R.L.M.R. wrote the first draft of the manuscript and all Python scripts for analysis of lattice energy calculations and dataset calculations. R.L.M.R. also wrote the Perl script for calculation of force-field lattice energies with input from A.M. and D.G. Advice on the use of Materials Studio was provided by E.C. and K.P. The force-field protocols implemented were designed in consultation with K.P. and K.R. along with other scientists at the University of Leeds (see Acknowledgements). R.L.M.R., D.G. and A.G.P.M. wrote the Python workflow for generating CIF files required for lattice energy calculations. D.G. was responsible for all DFT calculations. Method A for pharmaceutically relevant subset selection was designed, implemented and analyzed by R.M. working with C.M. Method B was designed and implemented by R.L.M.R., with guidance from K.P., and was informed by analysis carried out by R.M. and C.M. The final version of the SUB-BIG dataset was prepared by R.L.M.R., building

upon the initial data curation and integration efforts of A.G.P.M., E.C., D.R.M.V. and other scientists at Pfizer and STFC (see Acknowledgements). Preparation of CIF files for some X23 dataset entries and assistance with analysis of polymorphism in the CSD was provided by A.G.P.M.

Funding Sources

We gratefully acknowledge the funding provided from the 'Advanced Digital Design of Pharmaceutical Therapeutics' (ADDoPT) project, funded by the UK's Advanced Manufacturing Supply Chain Initiative (AMSCI). (AMSCI grant no. 14060.) D.R.M.V thanks Pfizer for funding a PhD studentship. A.M. gratefully thanks EPSRC and Pfizer, for supporting a Postgraduate Scholarship through the Centre for Doctoral Training in Complex Particulate Products and Processes (CP3) program (grant number EP/L015285/1).

Notes

N/A

Acknowledgements

Dr. Christelle Gendrin (Airbus, formerly of Science and Technology Facilities Council) is thanked for work on preparing the SUB-BIG dataset, including PubChem, DrugBank, ChEBI queries and for initial work on searching the NIST Chemistry WebBook. Dr. Robert Docherty (Pfizer) is thanked for initial work on preparing the SUB-BIG dataset, as well as guiding this work. Dr. Jason Cole (Cambridge Crystallographic Data Centre) is thanked for assistance with data curation. Professor William Jones (University of Cambridge) and Dr Colin Groom (Cambridge Crystallographic Data Centre) are thanked for their advice concerning the initial data

curation that led to the dataset used in this study. Colin Edge (GlaxoSmithKline) is thanked for helpful discussions and for providing descriptor distribution statistics for a subset of the GlaxoSmithKline in-house crystal structures database. Jakub Janowiak (University of Leeds) is thanked for assistance with analysis of polymorphism in the CSD. Dr. John Mitchell (University of St. Andrews) and Dr. James McDonagh (IBM research UK) are thanked for providing a copy of the SUB-48 dataset in electronic form, which was integrated into the SUB-BIG dataset, and for helpful correspondence. Dr. Peter J. Linstrom (U.S. National Institute of Standards and Technology, NIST) is thanked for helpful correspondence. Dr. Robert Hammond, Dr. Ian Rosbottom and Dr. Jonathan Pickering (University of Leeds) are thanked for useful discussions regarding force-field protocols and/or lattice energy calculations using Materials Studio. Professor Richard Cooper (University of Oxford) and Dr. Jerome Wicker (Oxford Drug Design) are thanked for helpful correspondence, as well as for providing their nConf20 code and allowing us to modify and redistribute this. Dr. Jian-Jie Liang (Biovia) and Dr. Carsten Menke (Biovia) are thanked for guidance regarding the use of Materials Studio. Analyses of pharmaceutical relevance were, in part, carried out using resources made available by the University of Leeds. This work was undertaken on ARC2, part of the High Performance Computing facilities at the University of Leeds, UK. We acknowledge use of Hartree Centre resources in this work. The Science and Technology Facilities Council (STFC) Hartree Centre is a research collaborator in association with IBM providing High Performance Computing platforms funded by the UK's investment in e-Infrastructure. The Centre aims to develop and demonstrate next generation software, optimized to take advantage of the move towards exa-scale computing.

Abbreviations

DFT (density functional theory); RMSE (Root Mean-Squared Error); MAD (Mean Absolute Deviation)

References

(1) Huang, L.-F.; Tong, W.-Q. (Tony). Impact of Solid State Properties on Developability Assessment of Drug Candidates. *Adv. Drug Deliv. Rev.* **2004**, *56*, 321–334.

(2) McDonagh, J. L.; Palmer, D. S.; Mourik, T. van; Mitchell, J. B. O. Are the Sublimation Thermodynamics of Organic Molecules Predictable? *J. Chem. Inf. Model.* **2016**, *56*, 2162–2179.

(3) Docherty, R.; Pencheva, K.; Abramov, Y. A. Low Solubility in Drug Development: De-Convoluting the Relative Importance of Solvation and Crystal Packing. *J. Pharm. Pharmacol.* **2015**, *67*, 847–856.

(4) Skyner, R. E.; McDonagh, J. L.; Groom, C. R.; Mourik, T. van; Mitchell, J. B. O. A Review of Methods for the Calculation of Solution Free Energies and the Modelling of Systems in Solution. *Phys. Chem. Chem. Phys.* **2015**, *17*, 6174–6191.

(5) McDonagh, J. L.; Nath, N.; De Ferrari, L.; van Mourik, T.; Mitchell, J. B. O. Uniting Cheminformatics and Chemical Theory To Predict the Intrinsic Aqueous Solubility of Crystalline Druglike Molecules. *J. Chem. Inf. Model.* **2014**, *54*, 844–856.

(6) Marchese Robinson, R. L.; Roberts, K. J.; Martin, E. B. The Influence of Solid State Information and Descriptor Selection on Statistical Models of Temperature Dependent Aqueous Solubility. *J. Cheminformatics* **2018**, *10*, 44.

(7) Muller, F. L.; Fielding, M.; Black, S. A Practical Approach for Using Solubility to Design Cooling Crystallisations. *Org. Process Res. Dev.* **2009**, *13*, 1315–1321.

(8) Sarah L. Price. 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**.

(9) Červinka, C.; Fulem, M. State-of-the-Art Calculations of Sublimation Enthalpies for Selected Molecular Crystals and Their Computational Uncertainty. *J. Chem. Theory Comput.* **2017**, *13*, 2840–2850.

(10) Brandenburg, J. G.; Grimme, S. Organic Crystal Polymorphism: A Benchmark for Dispersion-Corrected Mean-Field Electronic Structure Methods. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **2016**, *72*, 502–513.

(11) Nguyen, T. T. H.; Rosbottom, I.; Marziano, I.; Hammond, R. B.; Roberts, K. J. Crystal Morphology and Interfacial Stability of RS-Ibuprofen in Relation to Its Molecular and Synthonic Structure. *Cryst. Growth Des.* **2017**, *17*, 3088–3099.

(12) Rosbottom, I.; Roberts, K. J. Crystal Growth and Morphology of Molecular Crystals. In *Engineering Crystallography: From Molecule to Crystal to Functional Form*; Roberts, K. J., Docherty, R., Tamura, R., Eds.; NATO Science for Peace and Security Series A: Chemistry and Biology; Springer Netherlands: Dordrecht, 2017; pp 109–131.
(13) Rosbottom, I.; Roberts, K. J.; Docherty, R. The Solid State, Surface and Morphological Properties of P-Aminobenzoic Acid in Terms of the Strength and Directionality of Its Intermolecular Synthons. *CrystEngComm* **2015**, *17*, 5768–5788.
(14) Pickering, J.; Hammond, R. B.; Ramachandran, V.; Soufian, M.; Roberts, K. J. Synthonic Engineering Modelling Tools for Product and Process Design. In *Engineering Crystallography: From Molecule to Crystal to Functional Form*; Roberts, K. J., Docherty, R., Tamura, R., Eds.; NATO Science for Peace and Security Series A: Chemistry and Biology; Springer Netherlands: Dordrecht, 2017; pp 155–176.
(15) Shariare, M. H.; Leusen, F. J. J.; de Matas, M.; York, P.; Anwar, J. Prediction of the Mechanical Behaviour of Crystalline Solids. *Pharm. Res.* **2012**, *29*, 319–331.
(16) Hammond, R. B. Modelling Route Map: From Molecule Through the Solution State to Crystals. In *Engineering Crystallography: From Molecule to Crystal to Functional Form*; Roberts, K. J., Docherty, R., Tamura, R., Eds.; NATO Science for Peace and Security Series A: Chemistry and Biology; Springer Netherlands: Dordrecht, 2017; pp 71–108.
(17) Otero-de-la-Roza, A.; Johnson, E. R. A Benchmark for Non-Covalent Interactions in Solids. *J. Chem. Phys.* **2012**, *137*, 54103.
(18) Chickos, J. S.; Hosseini, S.; Hesse, D. G.; Liebman, J. F. Heat Capacity Corrections to a Standard State: A Comparison of New and Some Literature Methods for Organic Liquids and Solids. *Struct. Chem.* **1993**, *4*, 271–278.
(19) Červinka, C.; Fulem, M.; Růžička, K. CCSD(T)/CBS Fragment-Based Calculations of Lattice Energy of Molecular Crystals. *J. Chem. Phys.* **2016**, *144*, 64505.
(20) Sun, H. COMPASS: An Ab Initio Force-Field Optimized for Condensed-Phase ApplicationsOverview with Details on Alkane and Benzene Compounds. *J. Phys. Chem. B* **1998**, *102*, 7338–7364.
(21) Osborn, J. C.; York, P. A Comparison of Sublimation Enthalpies with Lattice Energies Calculated Using Force Fields. *J. Mol. Struct.* **1999**, *474*, 43–47.
(22) Reilly, A. M.; Tkatchenko, A. Understanding the Role of Vibrations, Exact Exchange, and Many-Body van Der Waals Interactions in the Cohesive Properties of Molecular Crystals. *J. Chem. Phys.* **2013**, *139*, 24705.
(23) Sarah L. Price; Leslie, M.; A. Welch, G. W.; Habgood, M.; S. Price, L.; G. Karamertzanis, P.; Day, G. M. Modelling Organic Crystal Structures Using Distributed Multipole and Polarizability-Based Model Intermolecular Potentials. *Phys. Chem. Chem. Phys.* **2010**, *12*, 8478–8490.
(24) Jensen, F. *Introduction to Computational Chemistry*, 2nd ed.; WILEY: Chichester, England, 2007.
(25) Bernardes, C. E. S.; Joseph, A. Evaluation of the OPLS-AA Force Field for the Study of Structural and Energetic Aspects of Molecular Organic Crystals. *J. Phys. Chem. A* **2015**, *119*, 3023–3034.
(26) Filippini, G.; Gavezzotti, A. Empirical Intermolecular Potentials for Organic Crystals: The `6-Exp' Approximation Revisited. *Acta Crystallogr. B* **1993**, *49*, 868–880.

(27) Gavezzotti, A.; Filippini, G. Geometry of the Intermolecular X-H...Y (X, Y = N, O) Hydrogen Bond and the Calibration of Empirical Hydrogen-Bond Potentials. *J. Phys. Chem.* **1994**, *98*, 4831–4837.
(28) Sun, H.; Jin, Z.; Yang, C.; Akkermans, R. L. C.; Robertson, S. H.; Spenley, N. A.; Miller, S.; Todd, S. M. COMPASS II: Extended Coverage for Polymer and Drug-like Molecule Databases. *J. Mol. Model.* **2016**, *22*, 47.
(29) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys Rev* **1965**, *140*, A1133–A1138.
(30) Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys Rev* **1964**, *136*, 864–871.
(31) Nyman, J.; Pundyke, O. S.; Day, G. M. Accurate Force Fields and Methods for Modelling Organic Molecular Crystals at Finite Temperatures. *Phys. Chem. Chem. Phys.* **2016**, *18*, 15828–15837.
(32) Thompson, H. P.; Day, G. M. Which Conformations Make Stable Crystal Structures? Mapping Crystalline Molecular Geometries to the Conformational Energy Landscape. *Chem. Sci.* **2014**, *5*, 3173–3182.
(33) Clydesdale, G.; Docherty, R.; Roberts, K. J. HABIT - a Program for Predicting the Morphology of Molecular Crystals. *Comput. Phys. Commun.* **1991**, *64*, 311–328.
(34) Clydesdale, G.; Roberts, K. J.; Docherty, R. HABIT95 — a Program for Predicting the Morphology of Molecular Crystals as a Function of the Growth Environment. *J. Cryst. Growth* **1996**, *166*, 78–83.
(35) BIOVIA Materials Studio 2017 (17.1.0.48) http://accelrys.com/products/collaborative-science/biovia-materials-studio/ (accessed Jul 25, 2017).
(36) Gavezzotti, A.; Chickos, J. S. Sublimation Enthalpies Of Organic Compounds: A Very Large Database With A Match To Crystal Structure Determinations And A Comparison With Lattice Energies. *Cryst. Growth Des.* **2019**, Article In Press.
(37) CASTEP/Getting CASTEP http://www.castep.org/CASTEP/GettingCASTEP (accessed Sep 11, 2018).
(38) 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.
(39) Acree, W.; Chickos, J. S. Phase Transition Enthalpy Measurements of Organic and Organometallic Compounds. Sublimation, Vaporization and Fusion Enthalpies From 1880 to 2015. Part 1. C1 − C10. *J. Phys. Chem. Ref. Data* **2016**, *45*, 33101.
(40) Steed, J. Introduction | Z' http://zprime.co.uk/ (accessed Jul 25, 2017).
(41) Steed, K. M.; Steed, J. W. Packing Problems: High Z' Crystal Structures and Their Relationship to Cocrystals, Inclusion Compounds, and Polymorphism. *Chem. Rev.* **2015**, *115*, 2895–2933.
(42) standardiser module description https://wwwdev.ebi.ac.uk/chembl/extra/francis/standardiser/ (accessed Jul 25, 2017).
(43) Maschio, L.; Civalleri, B.; Ugliengo, P.; Gavezzotti, A. Intermolecular Interaction Energies in Molecular Crystals: Comparison and Agreement of Localized Møller–Plesset 2, Dispersion-Corrected Density Functional, and Classical Empirical Two-Body Calculations. *J. Phys. Chem. A* **2011**, *115*, 11179–11186.
(44) Fan, J.-Y.; Zheng, Z.-Y.; Su, Y.; Zhao, J.-J. Assessment of Dispersion Correction Methods within Density Functional Theory for Energetic Materials. *Mol. Simul.* **2017**, *43*, 568–574.
(45) Beran, G. J. O.; Nanda, K. Predicting Organic Crystal Lattice Energies with Chemical Accuracy. *J. Phys. Chem. Lett.* **2010**, *1*, 3480–3487.

(46) Zheng, Z.; Zhao, J.; Sun, Y.; Zhang, S. Structures and Lattice Energies of Molecular Crystals Using Density Functional Theory: Assessment of a Local Atomic Potential Approach. *Chem. Phys. Lett.* **2012**, *550*, 94–98.
(47) Carter, D. J.; Rohl, A. L. Benchmarking Calculated Lattice Parameters and Energies of Molecular Crystals Using van Der Waals Density Functionals. *J. Chem. Theory Comput.* **2014**, *10*, 3423–3437.
(48) Reilly, A. M.; Tkatchenko, A. Seamless and Accurate Modeling of Organic Molecular Materials. *J. Phys. Chem. Lett.* **2013**, *4*, 1028–1033.
(49) Jane Li, Z.; Ojala, W. H.; Grant, D. J. W. Molecular Modeling Study Of Chiral Drug Crystals: Lattice Energy Calculations. *J. Pharm. Sci.* **2001**, *90*, 1523–1539.
(50) Hagler, A. T.; Lifson, S.; Dauber, P. Consistent Force Field Studies of Intermolecular Forces in Hydrogen-Bonded Crystals. 2. A Benchmark for the Objective Comparison of Alternative Force Fields. *J. Am. Chem. Soc.* **1979**, *101*, 5122–5130.
(51) Charlton, M. H.; Docherty, R.; Hutchings, M. G. Quantitative Structure–sublimation Enthalpy Relationship Studied by Neural Networks, Theoretical Crystal Packing Calculations and Multilinear Regression Analysis. *J. Chem. Soc. Perkin Trans. 2* **1995**, *0*, 2023–2030.
(52) Thomas, S. P.; Spackman, P. R.; Jayatilaka, D.; Spackman, M. A. Accurate Lattice Energies for Molecular Crystals from Experimental Crystal Structures. *J. Chem. Theory Comput.* **2018**, *14*, 1614–1623.
(53) Salahinejad, M.; Le, T. C.; Winkler, D. A. Capturing the Crystal: Prediction of Enthalpy of Sublimation, Crystal Lattice Energy, and Melting Points of Organic Compounds. *J. Chem. Inf. Model.* **2013**, *53*, 223–229.
(54) Meftahi, N.; Walker, M. L.; Enciso, M.; Smith, B. J. Predicting the Enthalpy and Gibbs Energy of Sublimation by QSPR Modeling. *Sci. Rep.* **2018**, *8*, 9779.
(55) Streek, J. van de. Searching the Cambridge Structural Database for the `best' Representative of Each Unique Polymorph. *Acta Crystallogr. B* **2006**, *62*, 567–579.
(56) Nyman, J.; Day, G. M. Static and Lattice Vibrational Energy Differences between Polymorphs. *CrystEngComm* **2015**, *17*, 5154–5165.
(57) Marchese Robinson, R. L.; Mackenzie, R.; Geatches, D.; Maloney, A. G. P.; Moldovan, A.; Morris, C. Scripts and Jupyter Notebooks Required to Apply the Force-Field Calculations, Pharmaceutical Relevance Approaches and Analyses Reported in the Following Publication: "Evaluation of Force-Field Calculations of Lattice Energies on a Large Public Dataset, Assessment of Pharmaceutical Relevance and Comparison to density Functional Theory" (Version v4) http://dx.doi.org/10.5281/zenodo.3477986 (accessed Oct 9, 2019).
(58) CSD Python API (version 1.3.0). Quick primer to using the CSD Python API https://downloads.ccdc.cam.ac.uk/documentation/API/descriptive_docs/primer.html (accessed Jul 24, 2017).
(59) Mayo, S. L.; Olafson, B. D.; Goddard, W. A. DREIDING: A Generic Force Field for Molecular Simulations. *J. Phys. Chem.* **1990**, *94*, 8897–8909.
(60) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024–10035.
(61) Dauber-Osguthorpe, P.; Roberts, V. A.; Osguthorpe, D. J.; Wolff, J.; Genest, M.; Hagler, A. T. Structure and Energetics of Ligand Binding to Proteins: Escherichia Coli

Dihydrofolate Reductase-Trimethoprim, a Drug-Receptor System. *Proteins Struct. Funct. Bioinforma.* **1988**, *4*, 31–47.
(62) Sun, H. Force Field for Computation of Conformational Energies, Structures, and Vibrational Frequencies of Aromatic Polyesters. *J. Comput. Chem.* **1994**, *15*, 752–768.
(63) Ewald, P. P. Die Berechnung Optischer Und Elektrostatischer Gitterpotentiale. *Ann. Phys.* **1921**, *369*, 253–287.
(64) Tosi, M. P. Cohesion of Ionic Solids in the Born Model. In *Solid State Physics*; Seitz, F., Turnbull, D., Eds.; Academic Press, 1964; Vol. 16, pp 1–120.
(65) Alexander, D. L. J.; Tropsha, A.; Winkler, D. A. Beware of R2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *J. Chem. Inf. Model.* **2015**, *55*, 1316–1322.
(66) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the Chemical Beauty of Drugs. *Nat. Chem.* **2012**, *4*, 90–98.
(67) Muller, K.; Mika, S.; Ratsch, G.; Tsuda, K.; Scholkopf, B. An Introduction to Kernel-Based Learning Algorithms. *IEEE Trans. Neural Netw.* **2001**, *12*, 181–201.
(68) Hsu, C.-W.; Chang, C.-C.; Lin, C.-J. *A Practical Guide to Support Vector Classification*; Department of Computer Science, National Taiwan University: Taiwan, 2016.
(69) Mathematical Formulation of the NuSVC variant of Support Vector Classification http://scikit-learn.org/stable/modules/svm.html#mathematical-formulation (accessed Sep 28, 2018).
(70) Lind, P.; Maltseva, T. Support Vector Machines for the Estimation of Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1855–1859.
(71) DrugBank http://drugbank.ca/ (accessed Mar 16, 2011).
(72) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: A Knowledgebase for Drugs, Drug Actions and Drug Targets. *Nucleic Acids Res.* **2008**, *36*, D901–D906.
(73) Wishart, D. S. DrugBank: A Comprehensive Resource for in Silico Drug Discovery and Exploration. *Nucleic Acids Res.* **2006**, *34*, D668–D672.
(74) Smola, A. J.; Schölkopf, B. A Tutorial on Support Vector Regression. *Stat. Comput.* **2004**, *14*, 199–222.
(75) Wicker, J. G. P.; Cooper, R. I. Beyond Rotatable Bond Counts: Capturing 3D Conformational Flexibility in a Single Descriptor. *J. Chem. Inf. Model.* **2016**, *56*, 2347–2352.
(76) Clark, S. J.; Segall, M. D.; Pickard, C. J.; Hasnip, P. J.; Probert, M. I. J.; Refson, K.; Payne, M. C. First Principles Methods Using CASTEP. *Z Krist.* **2005**, *220*, 567–570.
(77) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
(78) 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*, 0730051–0730054.
(79) CDK Depict http://www.simolecule.com/cdkdepict/depict.html (accessed Sep 26, 2019).
(80) Reilly, A. M.; Cooper, R. I.; Adjiman, C. S.; Bhattacharya, S.; Boese, A. D.; Brandenburg, J. G.; Bygrave, P. J.; Bylsma, R.; Campbell, J. E.; Car, R.; et al. Report on the Sixth Blind Test of Organic Crystal Structure Prediction Methods. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **2016**, *72*, 439–459.

(81) Ambrosetti, A.; Reilly, A. M.; DiStasio, R. A.; Tkatchenko, A. Long-Range Correlation Energy Calculated from Coupled Atomic Response Functions. *J. Chem. Phys.* **2014**, *140*, 18A508.
(82) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.
(83) Červinka, C.; Beran, G. J. O. Ab Initio Prediction of the Polymorph Phase Diagram for Crystalline Methanol. *Chem. Sci.* **2018**, *9*, 4622–4629.
(84) Cardozo, R. L. Enthalpies of Combustion, Formation, Vaporization and Sublimation of Organics. *AIChE J.* **1991**, *37*, 290–298.
(85) Chow, E.; Docherty, R.; Lai, T.; Pencheva, K. De-Risking Early Stage Drug Development with a Bespoke Lattice Energy Predictive Model: A Materials Science Informatics Approach to Address Challenges Associated with a Diverse Chemical Space. *J. Pharm. Sci.* **2019**.
(86) ADDoPT Project - Advanced Digital Design of Pharmaceutical Therapeutics https://www.addopt.org/ (accessed Nov 22, 2018).
(87) Ismael, A.; Gómez-Zavaglia, A.; Borba, A.; Cristiano, M. L. S.; Fausto, R. Amino→Imino Tautomerization upon in Vacuo Sublimation of 2-Methyltetrazole-Saccharinate as Probed by Matrix Isolation Infrared Spectroscopy. *J. Phys. Chem. A* **2013**, *117*, 3190–3197.
(88) Musil, F.; De, S.; Yang, J.; Campbell, J. E.; Day, G. M.; Ceriotti, M. Machine Learning for the Structure–energy–property Landscapes of Molecular Crystals. *Chem. Sci.* **2018**, *9*, 1289–1300.

Table of Contents graphic