

Reliable crystal structure predictions from first principles

Rahul Nikhar¹ & Krzysztof Szalewicz^{*1}

¹*Department of Physics and Astronomy, University of Delaware, Newark, Delaware 19716*

Abstract

An inexpensive and reliable method for molecular crystal structure predictions (CSPs) has been developed. The new CSP protocol starts from a two-dimensional graph of crystal's monomer(s) and utilizes no experimental information. Using results of quantum mechanical calculations for molecular dimers, an accurate two-body, rigid-monomer *ab initio*-based force field (aiFF) for the crystal is developed. Since CSPs with aiFFs are essentially as expensive as with empirical FFs, tens of thousands of plausible polymorphs generated by the crystal packing procedures can be optimized. Here we show the robustness of this protocol which found the experimental crystal within the 20 most stable predicted polymorphs for each of the 15 investigated molecules. The ranking was further refined by performing periodic density-functional theory (DFT) plus dispersion correction (pDFT+D) calculations for these 20 top-ranked polymorphs, resulting in the experimental crystal ranked as number one for all the systems studied (and the second polymorph, if known, ranked in the top few). Alternatively, the polymorphs generated can be used to improve aiFFs, which also leads to rank one predictions. The proposed CSP protocol should result in aiFFs replacing empirical FFs in CSP research.

Introduction

Properties of crystalline solids depend critically on the polymorphic form of a given substance and many crystals can exist in several such forms [1, 2]. The knowledge of possible stable polymorphic forms of a crystal is of particular importance for pharmaceutical industry [3]. If a polymorph different from the one obtained in laboratories crystallizes during manufacturing of a drug, it will have different physicochemical properties and may lead to undesirable therapeutic effects, two examples are ritonavir [4, 5] and rotigotine [6–8]. Thus, in the drug development process, one would like to know if the polymorph used is thermodynamically the most stable form in ambient conditions. In defense industry, developments of energetic materials are costly and highly dangerous [9, 10] and *a priori* knowledge of crystal structure of notional materials would allow accelerated screening of such materials. Also semiconductor industry can benefit from such knowledge [11]. CSP methods answer these needs by finding a set of most stable crystalline polymorphs of a given molecule starting from its two-dimensional diagram and not using any experimental information.

Reliable CSPs for molecular crystals starting from the knowledge of only two-dimensional diagrams of monomer(s) were nearly impossible for a long time. In 1988, Maddox [12] described failure of CSPs as a continuing scandal in the physical sciences and stated that in general even simplest crystalline solids posed great chal-

lenge. In mid 1990s, Gavezzotti [13] asked the fundamental question: ‘Are crystal structures predictable?’, and his answer was ‘No’. In response to this criticism, the Cambridge Crystallographic Data Centre (CCDC) conducted a series of “blind” tests [14–19] by providing only two-dimensional diagrams of monomers of crystals that have been measured but not published and asking research groups to submit their predictions, with the results of the first test published in 2000. While the field has advanced significantly since the first test, the results of the last, 6th test [19] are still not completely satisfactory. The participating groups achieved the success rate between 13% and 57% (not including polymorphs C and E of system XXIII), where success means that the experimental polymorph was found among polymorphs on two lists containing 100 polymorphs each.

One should remark here that predictions of crystals structures are actually a difficult problem for physical science, opposite to what Maddox [12] implied. The reason is the high dimensionality of the conformational and crystallographic space resulting in thousands of plausible polymorphs produced by sampling of this space within a relatively narrow window of lattice energies and densities. The energetic distances between consecutive polymorphs ordered by lattice energy are of the order of 1 kJ/mol at the low-energy end, which requires accuracies nearly impossible to achieve by empirical FFs. Also, for experimentally observed polymorphs, the differences between their computed lattice energies are of the same order [20]. While there are several variants of CSPs, including a re-

*email: szalewicz@udel.edu

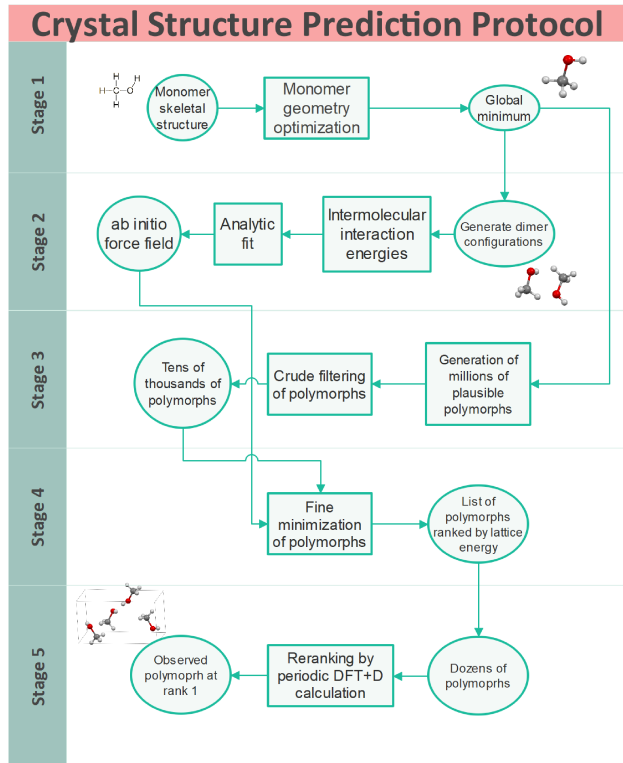


FIG. 1: Overview of aiFF-based CSP protocol.

Stage 1: monomer energy minimization to find the equilibrium geometry. Stage 2: *ab initio* calculations of dimer intermolecular interaction energies followed by fitting an analytic form of aiFF to these data. Stage 3: generation of millions of plausible packing arrangements of polymorphs by sampling different space groups, orientations of monomers, and unit cell parameters, followed by a reduction of this set to tens of thousands of polymorphs using density criteria or crude lattice energy minimizations with simple FFs. Stage 4: fine minimization with aiFFs for all polymorphs in the reduced set. Stage 5: refinement of the ranking via pDFT+D calculations on a couple dozen top-ranked polymorphs from Stage 4.

cent use of deep neural networks [21], the majority of groups participating in the 6th blind test used some form of FFs, mostly of empirical character. The most successful CSP protocol consisting of polymorph-space sampling plus lattice-energy minimization has been developed by Neumann *et al.* [22, 23]. This protocol uses a tailor-made FF which is obtained by refining parameters of an empirical FF to reproduce as close as possible pDFT+D lattice energies (and their derivatives). The initial polymorphs for pDFT+D calculations are obtained using the empirical FF. The method is included in the commer-

cial software package GRACE (Generation Ranking and Characterization Engine) [24], but some of its details are not available. Recent reviews of the field of CSPs can be found in Refs. 25–31.

In the present work, a CSP protocol is proposed based entirely on first principles, i.e., not utilizing any experimental information. Since the main characteristic of this method is the use of aiFFs, we will refer to it as the CSP(aiFF) protocol. This protocol consists of several stages shown in Fig. 1. While aiFFs have been used in CSPs for some time [19, 32–34], such predictions were taking a long time (several months at the minimum), required huge amounts of human effort, and were possible for monomers with up to about 20 atoms. In the present work, four recent developments are combined to dramatically reduce costs and increase predictability of such CSPs: (a) The development of a very effective variant [35] of symmetry-adapted perturbation theory (SAPT) [36] for *ab initio* calculations of interaction energies; (b) The creation of autoPES [37, 38]: an automatic, effective, and reliable method for generation of potential energy surfaces (PESs) with minimal human involvement; (c) Enabling the use such aiFFs in the lattice-energy minimization stage of CSPs, a part of the present work; (d) The application of pDFT+D for a final refinement of polymorph rankings. Stage 3 of Fig. 1 can produce even millions of polymorphs at low costs and past experience indicates that the experimentally relevant polymorphs are almost always among them. Thus, the essence of CSP protocols is to filter all relevant low lattice energy polymorphs out of this set. In the past few years, it has been demonstrated by several groups that pDFT+D geometry optimization of polymorphs places the experimental polymorph ranked within the top few, often as number one [19, 39–41]. However, such calculations are so expensive that they can be afforded for only a hundred or so polymorphs. In contrast, if an FF is used in Stage 4, tens of thousands polymorphs can be optimized. This FF has to be sufficiently accurate not to miss any important polymorphs. Thus, both the *ab initio* method and the fit to interaction energies computed using this method must have sufficiently small uncertainties. In calculations of dimer interaction energies, the variant of SAPT used by us (see Methods) is nearly as accurate [35, 42, 43] as the coupled cluster method with single, double, and noniterative triple excitations, CCSD(T), the “gold-standard” method of electronic structure theory, but is significantly less expensive. To prevent loss of accuracy due to fitting,

the form of the fitting function has to be significantly more involved than those of empirical FFs that are typically built from Lenard-Jones plus Coulomb potentials, see Methods. The extended form can fit *ab initio* data with uncertainties of about 1 kJ/mol, which we will show to be sufficient for reliable CSPs. Such form has never been used in lattice energy minimizations and we had to modify CSP software to make it possible. Finally, to make Stage 2 affordable, the number of *ab initio* grid points needed to fit an aiFF has to be reasonably small. The autoPES method [37, 38] reduces this number by two orders of magnitude compared to typical surface-fitting approaches, reducing in this way the development costs by the same ratio. It also reduces amount of human involvement almost to zero as the whole process is completely automated. We show below that the proposed protocol found the experimental crystal ranked as number one for all 15 molecules studied (and the second polymorph, if known, ranked in the top few).

Results and Discussion

Performance of CSP(aiFF) protocol. To assess the performance of our method, we carried out CSPs for 15 molecules including several systems from the CCDC blind tests [14–19] (denoted by roman numerals), as well as for methanol, benzene, nitromethane, 5,5'-dinitro-2H,2'H-3,3'-bi-1,2,4-triazole (DNBT), 1-3-5-trinitrobenzene (TNB), deferiprone, and fluorouracil. The molecular graphs are shown in Supplementary Figure 1. The results are summarized in Table I. An extended version of this table is available in Supplementary Table 1. The CSP(aiFF) protocol ranked the experimental polymorph as number 1 in 5 cases, as number 2-6 in 7 cases, and as numbers 9, 9, and 16. We have also included a second experimentally identified polymorph in the cases of system I, benzene, and deferiprone, denoted as "Poly2" in Table I, and these are ranked as numbers 8, 4, and 8, respectively. After pDFT+D calculations on top-ranked 20 polymorphs of each crystal, without any further geometry optimization, an experimental crystal became ranked as number 1 in each case. For deferiprone, it was Poly2 that became the rank 1 polymorph, while Poly1 remained at rank 2. For system I and benzene, Poly2 changed rank from 8 to 2 and from 4 to 3, respectively. RMSD₂₀'s between the calculated and experimental crystals vary between 0.09 and 0.67 Å, below the CCDC threshold of 0.8 Å. Also densities and cell parameters, shown in Supplementary Table 1, agree very closely. Supplementary Figure 2 displays the percent de-

TABLE I: **CSPs from SAPT(DFT)-based aiFFs minimizations followed by pDFT+D fixed-geometry calculations.** SG: predicted space group of the crystal (SG is the same for experimental and predicted polymorphs); Rank: rank of the experimental polymorph after minimizations and after pDFT+D calculations; RMSD₂₀: root mean square deviation (in Å) between the experimental crystal and the calculated polymorph for 20 overlapping molecules (heavy atoms only); RMSE: root mean square error (in kJ/mol) of the fit for negative interaction energies.

System	SG	Rank	RMSD ₂₀	RMSE
I _{Poly1}	<i>P2₁/c</i>	2/1	0.09	0.6
I _{Poly2}	<i>Pbca</i>	8/2	0.32	0.6
II	<i>P2₁/n</i>	1/1	0.59	1.3
IV	<i>P2₁/c</i>	2/1	0.24	0.63
VIII	<i>C2/c</i>	4/1	0.28	1.1
XII	<i>Pbca</i>	9/1	0.53	0.84
XIII	<i>P2₁/c</i>	4/1	0.45	1.1
XVI	<i>Pbca</i>	16/1	0.29	1.0
XXII	<i>P2₁/n</i>	1/1	0.15	1.4
Methanol	<i>P2₁2₁2₁</i>	6/1	0.4	0.92
Benzenep _{Poly1}	<i>Pbca</i>	1/1	0.16	0.59
Benzenep _{Poly2}	<i>P2₁/c</i>	4/3	0.4	0.59
Nitromethane	<i>P2₁2₁2₁</i>	1/1	0.27	0.74
DNBT	<i>P2₁/c</i>	1/1	0.58	1.56
TNB	<i>P2₁/c</i>	3/1	0.67	1.28
Deferiprone _{Poly1}	<i>Pbca</i>	2/2	0.28	0.71
Deferiprone _{Poly2}	<i>P2₁/c</i>	8/1	0.24	0.71
Fluorouracil	<i>P2₁/c</i>	9/1	0.61	1.06

viations between the calculated and experimental lattice parameters. The average errors for the cell parameters *a*, *b*, *c*, and β amount to 4.3%, 2.6%, 4.3%, and 2.4%, respectively. Such level of predictivity is unprecedented for a completely first-principles CSP protocol. The overlaps of the experimental polymorphs with the closest calculated ones are shown in Fig. 2. This figure allows intuitive appreciation how close these structures are. This exceptional performance of CSP(aiFF) has been achieved despite the investigated systems exhibiting typical difficulties due to closeness of polymorphs' lattice energies and despite using rigid-monomer approximation. The lattice energy vs. density landscapes from the aiFF minimizations for systems IV and XXII are shown in Supplementary Figure 3. Analogous graphs for the other systems look similar. The lowest-energy 100 polymorphs spread the range of about 5 kJ/mol for systems I, XII, XIII, benzene, and nitromethane, about 10 kJ/mol for systems II, IV, VIII, XVI, XXII, methanol, TNB, deferiprone, and fluorouracil, and about 20 kJ/mol for DNBT. At the low-energy end, the energy differences between consecutive polymorphs are less than 1 kJ/mol, i.e., comparable to the RMSEs of the fits over all dimer configurations with negative interaction energies, shown in Table I.

Performance of a simplified aiFF form. The use

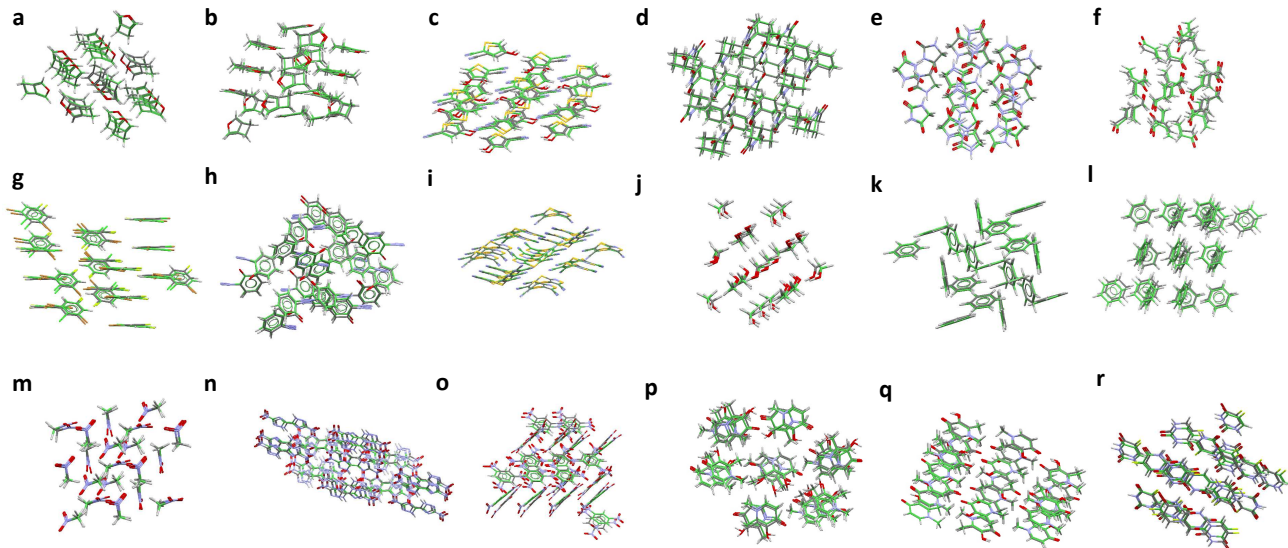


FIG. 2: **Overlaps of crystal structures.** Overlap of the experimental crystal structure (element-specific colors) with the closest calculated crystal structure (green) using SAPT(DFT)-based aiFFs for systems: (a) and (b) I, (c) II, (d) IV, (e) VIII, (f) XII, (g) XIII, (h) XVI, (i) XXII, (j) methanol, (k) and (l) benzene, (m) nitromethane, (n) DNBT, (o) TNB, (p) and (q) Deferiprone, (r) Fluorouracil.

of the extended functional form of aiFFs in the lattice energy minimizations instead of the simpler exp-6-1 form (not including a polynomial in front of exponential, damping functions, etc., see Methods) used in some empirical FFs leads to enormous improvements in rankings. To quantify such improvements, we performed lattice energy minimizations with the exp-6-1 form of aiFFs, fitted using the same level of theory as in the case of the extended form, for systems I, II, IV, and XXII, achieving rankings of 138, 2231, 49, and 60, respectively, while the rankings of the extended aiFF form for these systems are 1 or 2, see Table I. The main reason for this improvement is that RMSEs for negative interaction energies are from 2.3 to 5.3 times smaller in the latter case (these ratios are correlated with the number of fit parameters: e.g., 30 and 270 for the exp-6-1 and the extended form, respectively, in the case of system IV).

Performance of an empirical FF. In order to quantify better the predictive power of our approach, calculations analogous to those described above have been performed with an empirical FF. We have chosen the W99 FF [44] with point charges computed by us using the

CHELPG method [45]. For the 18 experimental polymorphs considered, the W99+charges FF found 33% of them at rank 10 or better, while the analogous result for aiFF (without the pDFT+D step) is 94%. This amounts to a qualitative difference for technological applications. For more details on CSPs with the W99+charges FF, see Supplementary Tables 1 and 2.

Alternative CSP(aiFF) protocol. One may ask why pDFT+D calculations are needed to improve the rankings, while several comparisons on benchmark interaction energies, see, e.g., Refs. 42 and 43, show that SAPT(DFT) is nearly as accurate as CCSD(T) and more accurate than DFT+D methods. The main reason is that what is used in CSPs are aiFFs, and they include additional uncertainties due to fitting. Although the average fit error for negative interaction energies is only ~ 1 kJ/mol, errors may be larger at some configurations. If a configuration with a too negative interaction energy is important for a polymorph, this polymorph may become overly stable and therefore too highly ranked. Two other possible reasons, basis set size and neglect of many-body effects in CSP(aiFF), are discussed in Supplementary In-

formation and found unlikely to be a reason. To improve the predictions from Stage 4, we have developed an alternative version of our method, alt-CSP(aiFF). After executing the CSP(aiFF) protocol less the pDFT+D stage, the geometries of 20 polymorphs with the lowest lattice energies are examined and consecutive nearest neighbour dimers identified. Then SAPT calculations are performed for these dimers, the aiFF is refitted, and lattice minimizations for the 20 polymorphs are performed with the new aiFF. This procedure is iterated until the energies of the 5x5x5 clusters extracted from each polymorph computed in two ways: just from the aiFF and in a hybrid way, replacing the aiFF interaction energies by the available SAPT ones, are the same to within some threshold. We have applied alt-CSP(aiFF) to two of the worst ranking crystals from Table I: system XVI (rank 16) and fluorouracil (rank 9). In each case, alt-CSP(aiFF) resulted in the experimental polymorph at rank 1, while RMSD₂₀ was reduced from 0.29 to 0.15 Å and from 0.61 to 0.42 Å, respectively. Thus, alt-CSP(aiFF) can be used without the pDFT+D stage. However, the additional *ab initio* calculations are about as expensive as the pDFT+D ones, so there is no gain in terms of efficiency.

Cost comparisons. The method proposed not only is highly reliable, as shown above, but also is very efficient compared to alternative ways of combining FF-based CSPs with pDFT+D calculations. To demonstrate this, we show in Fig. 3 the costs of three possible CSP strategies in terms of single-core wall times on the example of system I. Note that this type of calculations are typically performed on hundreds of cores, so the actual wall time is just a couple hours for Strategy 1, the approach proposed here. The majority of time for Strategy 1, 7 core-days, is spent for the development of an aiFF and most of this time is used to compute SAPT(DFT) interaction energies for 706 dimer configurations, with very little time spent on fitting these energies. The next stage, the packing and minimization (PACK+MIN) of hundreds of thousands of crystals, requires only less than a third of a day. The final stage, pDFT+D calculations for the top 20 polymorphs at aiFF geometries, requires approximately one day. Hypothetical Strategy 2 differs from Strategy 1 by the use of an empirical FF in the PACK+MIN stage and by performing pDFT+D calculations for 100 polymorphs with reoptimization of geometries (this number of polymorphs was chosen as a trade-off between success rate and computational costs). The time required for the

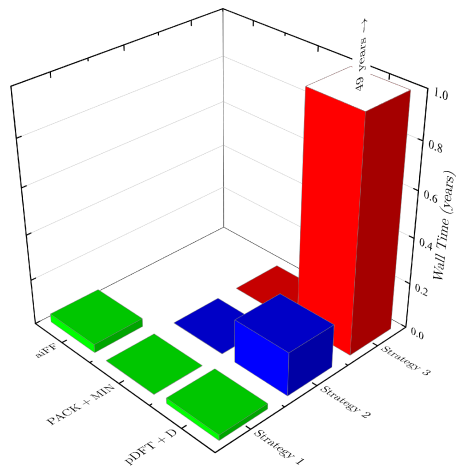


FIG. 3: Computational cost of the considered CSP protocols. Total wall times required for system I CSPs on a single core of the Intel E5-2670 processor using different strategies. Rows “aiFF”, “PACK+MIN”, and “pDFT+D” denote times of an aiFF development, packing and minimization with an aiFF, and of periodic DFT+D calculations.

latter stage would be 70 core-days, so Strategy 2 is about an order of magnitude more expensive than Strategy 1. Moreover, if the W99+charges FF were used, the success rate of Strategy 2 on the set of 18 polymorphs examined here would be 72% (see Supplementary Table 2), while the success rate of Strategy 1 is 100% already with 16 top-ranked polymorphs. All the PACK+MIN bars appear to be of about the same height for aiFF and for the empirical FF. This is because the calculation of the lattice energy is only about two times more expensive in the former case. Hypothetical Strategy 3 performs pDFT+D calculations with geometry optimization for all 25,500 polymorphs produced by PACK+MIN. This strategy would have a very high reliability (since practice indicates that the experimental polymorphs are almost always included in such a large pool of candidate structures), but it would be extremely costly, 49 single-core years, and hence not practical (although possible if a few thousands cores were used). With the use of an empirical FF, one can set the number of polymorphs included in the pDFT+D stage anywhere between 100 and 25,000, systematically increasing costs and reliability relative to Strategy 2. However, with W99+charges and our set of polymorphs, the success rate would remain at 72% until the number of polymorphs is at least 589. For Strategies 2 and 3, the PACK+MIN stage can be replaced by any

other protocol producing the required number of candidate polymorphs, with insignificant effects on the total timings.

Neglected effects. Since aiFFs are sums of two-body interactions, they neglect the many-body effects mentioned earlier and discussed in Supplementary Information. While we show that these effects are not critical in CSPs for the crystals considered here, they may be significant for some other crystals [46–48]. The most important many-body effect, the many-body polarization, can be accounted for using polarizable aiFFs that can be developed using autoPES, but are not yet implemented in our CSP codes. In Supplementary Information, we also explain why the relatively small basis set that we used is adequate for CSPs. A much more important neglected effect is flexibility of monomers. Although the monomers considered by us were assumed to be rigid, the proposed CSP(aiFF) protocol can be applied to monomers with soft degrees of freedom. Such monomers may be significantly deformed in crystals compared to their equilibrium structures in gas phase. The recent version of autoPES [38] has the capability of computing interaction energies accounting for all or selected intramonomer degrees of freedom and most CSP codes can perform packing and minimization including all degrees of freedom, therefore such predictions can be made still completely from first principles. However, costs of such calculations increase steeply with the total number of degrees of freedom. One way around this problem is to assume separation of inter- and intramonomer degrees of freedom in Stage 2, as it has been done in all biomolecular FFs and in all FFs used in flexible-monomers CSPs. Since our aiFFs depend only of separations between atoms of different monomers, interaction energies can be computed for arbitrary monomer configurations. Such “flexibilized” intermonomer FF can replace the intermonomer component of current empirical FFs, while the intramonomer component can be kept unchanged. One can expect that such a replacement should lead to improved predictions in flexible-monomer CSPs.

Other effects neglected by the present version of CSP(aiFF) are thermal and entropic ones, as the results presented by us correspond to 0 K temperature. For some crystals, these effects can change the rankings of polymorphs, as pointed out by Brandenburg and Grimme [39] and recently investigated extensively by Hoja *et al.* [41]. The thermal and entropic effects can be routinely computed using pDFT+D, although such calculations are

several times more expensive than pDFT+D calculations with static geometries. As a test, we have computed both effects for the 5 lowest lattice energy polymorphs of system XXII, leading to no change of rankings.

Concluding remarks. The first-principles CSP(aiFF) method developed here was applied to crystals of 15 rigid molecules with 18 known experimental polymorphs. When aiFFs are applied in CSPs for crystals of these molecules, 17 or 94% the polymorphs are ranked in the range 1–10, while the remaining one has rank 16. For comparison, analogous CSPs with the empirical W99+charges FF ranks only 33% of polymorphs in the range 1–10, 3 experimental polymorphs are not found within 568 or more generated ones, and for two molecules predictions were not possible due to missing atom types. The ability of CSP(aiFF) to minimize tens of thousands polymorphs is its key advantage over alternative approaches which have to use low-accuracy methods at this stage, often erroneously leading to discarding of correct structures. Upon a subsequent reranking of the top 20 polymorphs with pDFT+D calculations at fixed aiFF geometries, for all 15 molecules an experimental polymorph became ranked as number 1, while the second polymorphs became ranked as numbers 2, 2, and 3. The pDFT+D step can be omitted if aiFFs are iteratively improved by performing *ab initio* calculations on dimers extracted from crystals predicted with the previous iteration of an aiFF [the alt-CSP(aiFF) protocol]. The proposed CSP protocol not only shows ultimate predictive power for the systems tested, but is also inexpensive compared to other highly predictive approaches. On about a hundred cores, complete predictions for any of the systems investigated here take less than a day, including the aiFF generation. The CSP(aiFF) protocol requires a minimal human involvement, consisting only of input preparation for autoPES, UPACK, and pDFT+D calculations, and includes only free software with open source codes. Limitations of the current implementation of the CSP(aiFF) methodology have been discussed, in particular the neglect of many-body interactions and the rigid-monomer approximation. Although the test set included only homogeneous crystals, there are no reasons to doubt that the method will work equally well for cocrystals including salts since the quality of aiFFs does not depend on dimers being homogeneous or heterogeneous (of course, for two-component cocrystals, three PESs have to be developed). Also, while the largest of the test molecules included 22 atoms, the method should

apply equally well to larger molecules since the relative accuracy of SAPT(DFT) does not change with system size [35]. Of course, calculations will be more expensive as the size increases, but molecules with about 100 atoms are within reach. The effectiveness of the proposed CSP protocol is due to the use of the SAPT(DFT) method which is computationally efficient relative to other accurate electronic structure methods and due to the use of the autoPES method for fitting aiFFs since this method not only cuts the costs of such fits by orders of magnitude, but also reduces human effort of this most difficult to automate step almost to zero. An important element of the CSP(aiFF) protocol is that it replaces simple potential forms used in all earlier CSP protocols by an extended form capable of fitting *ab initio* interaction energies with significantly decreased uncertainties. An advantage of the proposed protocol is that it constitutes a complete first-principles procedure for investigating crystal structures and properties. Such a protocol should work equally well for any type of monomer, in contrast to the protocols using empirical FFs, which are expected to work well only for systems similar to those used in fitting such FFs. We believe that the overall effect of the proposed CSP protocol will be that the field of CSPs will move from the use of empirical FFs to aiFFs. This should increase reliability of predictions and therefore, while CSPs have played so far at the best advisory role in technology developments, they may become a leading element in developments of novel crystalline materials. More generally, aiFFs can be used in several types of computational material design.

Methods

Monomer geometry minimization. In Stage 1, monomer geometries were optimized using ORCA [49, 50] with the PBE [51] functional and D3 correction [52] in the aug-cc-pVTZ [53] basis set.

***Ab initio* calculations of interaction energies.** To make the CSP(aiFF) protocol practical, aiFFs have to be constructed in Stage 2 at reasonably low costs, but at the same time with small uncertainties, for monomers with dozens of atoms. This requires first that the *ab initio* method used to compute intermolecular interaction energies is inexpensive and accurate. It appears that the best current choice for such calculations is SAPT [36, 54], an *ab initio* method that computes interaction energies directly, starting from isolated monomers and imposing the correct electron permutational sym-

metry. We applied the SAPT variant based on DFT, SAPT(DFT) [55, 56], see Ref. 35 for a recent review of this method. SAPT(DFT) and CCSD(T) calculations scale as $\mathcal{O}(n^5)$ and $\mathcal{O}(n^7)$ with system size, respectively, where n is the number of electrons, and for dimers with a couple dozens of atoms, SAPT(DFT) calculations are about two orders of magnitude less expensive than CCSD(T) calculations. The recently developed new SAPT(DFT) algorithms and effective computer codes [35, 42] can be used to compute thousands of grid points for dimers with ~ 100 -atom monomers using reasonable computer resources and being able to achieve this in a few days if a sufficient number of computer cores are available.

The details of calculations of SAPT(DFT) [55–58] first- and second-order interaction energies are as follows. We used the density-fitting version [35, 59, 60] in the SAPT2020 [61] codes interfaced with the ORCA package [49, 50] for calculations on monomers. The PBE [51] functional was used in DFT calculations applying the gradient-regulated asymptotic correction (GRAC) [62, 63]. The aug-cc-pVDZ [53] basis set plus a set of 3s3p2d2f midbond functions (default of autoPES) was used in the monomer-centered plus basis set (MC⁺BS) format [64]. The terms accounting for higher-order induction and exchange-induction effects, denoted as $\delta E_{\text{int,resp}}^{\text{HF}}$ and obtained as a difference between Hartree-Fock (HF) interaction energies and the sum of appropriate SAPT(HF) first- and second-order corrections in their response (resp) versions, was included for all systems except system XIII, benzene, DNBT, and TNB. We use a short-hand notation for SAPT interaction energy components: “indx” is the sum of the second-order induction and exchange-induction components, as well as of the $\delta E_{\text{int,resp}}^{\text{HF}}$ contribution, “dispx” is the sum of the dispersion and exchange-dispersion components, “elst” is the electrostatic component, and “exch” is the first-order exchange component. Relative importance of attractive components is illustrated in Supplementary Figure 4.

Generation of aiFFs. In all past CSPs, only simple FFs have been used at the lattice-energy minimization stage. The two most often used forms are the Lennard-Jones 12-6-1 potential: $A_{12}/r^{12} - C_6/r^6 + q_a q_b/r$, and the Buckingham exp-6-1 potential: $Ae^{-\beta r} - C_6/r^6 + q_a q_b/r$, where r is an atom-atom distance and A_{12} , A , β , C_6 , q_a , and q_b are adjustable parameters. SAPT(DFT)-based aiFFs have been used in CSPs, but always with the exp-6-1 potential form in the minimization stage. This form

is not pliable enough to fit well *ab initio* data, leading to uncertainties of a few kJ/mol, too large for reliable CSPs. In contrast, the extended form used by us in the

CSP(aiFF) protocol can fit *ab initio* data with uncertainties of about 1 kJ/mol, which we show to be sufficient for reliable CSPs. This functional form is [37, 38]

$$V = \sum_{a \in A, b \in B} \left\{ \left[1 + \sum_{i=1,2} a_i^{ab}(r_{ab})^i \right] e^{\alpha^{ab} - \beta^{ab} r_{ab}} + \frac{A_{12}^{ab}}{(r_{ab})^{12}} - \sum_{n=6,8} f_n(\delta_n^{ab}, r_{ab}) \frac{C_n^{ab}}{(r_{ab})^n} + f_1(\delta_1^{ab}, r_{ab}) \frac{q_a q_b}{r_{ab}} \right\} \quad (1)$$

where a (b) goes over the sets of atoms in monomer A (B), respectively, α^{ab} , β^{ab} , a_i^{ab} , A_{12}^{ab} are repulsion-energy parameters, C_n^{ab} are long-range dispersion plus induction energy parameters, q_x , $x = a, b$, are atomic partial charges, δ_n^{ab} are damping parameters, and f_n are the Tang-Toennies [65] damping functions: $f_n(\delta, r) = 1 - e^{-\delta r} \sum_{m=0}^n (\delta r)^m / m!$ Long-range interaction energies were computed using an *ab initio*-distributed approach. The damping parameters in the dispersion plus induction term were fitted separately to the sum of all close-range second-order components plus $\delta E_{\text{int,resp}}^{\text{HF}}$, while δ_1^{ab} were fitted to electrostatic energies. All PESs developed here are two-body, 6-dimensional PESs, i.e., assume rigid monomers. The aiFFs were constructed as sums of these two-body PESs. One should add that the extended form of FFs given by Eq. (1) has been used in some published CSPs, but only in molecular dynamics (MD) simulations that can replace the pDFT+D calculations of Stage 5. Note that MD calculations are about as expensive as pDFT+D ones and significantly more expensive than the minimizations of Stage 4. Graphs showing SAPT(DFT) interaction energy components and their fits as functions of the distance R between the centers of mass of monomers are included in Supplementary Figure 5. One can see in particular that the *ab initio* electrostatic energies are reproduced very well for R 's larger than the van der Waals minimum distance R_{vdW} despite using only damped charge-charge interactions, i.e., omitting higher multipolar terms. While the use of the latter terms in empirical FFs improves the predictions compared to the use of charges only [66, 67], our results show that higher-rank multipoles are not needed if the electrostatic function includes damping and is fitted to *ab initio* electrostatic energies. The worsening of the agreement with *ab initio* values seen for $R < R_{\text{vdW}}$ is inevitable and is due to the charge-overlap effects that are not proportional to inverse powers of R [36]. These effects are accounted for in the overall fit by the first term in Eq. (1). This is why

the total fitted and *ab initio* interaction energies are in excellent agreement for all R .

Crystal packing and lattice-energy minimization.

Since none of the available CSP packages is capable of using the form of aiFFs given by Eq. (1), we have modified two such packages: MOLPAK [68] and UPACK [69] to be applied in Stages 3 and 4. MOLPAK uses the concept of coordination geometry and by default searches in 26 space groups: $P1$, $P\bar{1}$, $P2$, Pm , Pc , $P2_1$, $P2/c$, $P2_1/m$, $P2/m$, $P2_1/c$, Cc , $C2$, $C2/c$, $Pnn2$, $Pba2$, $Pnc2$, $P22_1$, $Pmn2_1$, $Pma2$, $P2_12_12$, $P2_12_12_1$, $Pca2_1$, $Pna2_1$, $Pnma$, $Fdd2$, $Pbcn$, and $Pbca$. It generates polymorphs on a grid in three-dimensional search space by systematically varying the orientation of the central molecule in steps of 10° . This generation is performed in all 51 coordination geometries. The packing in the unit cell is controlled by a simple repulsive $1/r^{12}$ interaction between atoms: the molecules are brought together until an energy threshold is reached. This step provides an initial set of 6,859 angle combinations $\times$ 51 coordination geometries = 349,809 hypothetical polymorphs. From this set, 25,500 densest polymorphs, 500 from each coordination geometry, are minimized using the program WMIN [70]. The default functional form of the FF in WMIN is exp-6-1. We have modified this code to include FFs of the form of Eq. (1).

UPACK generates random crystal structures in 13 default space groups: $C2$, $C2/c$, Cc , $P1$, $P\bar{1}$, $P2_1$, $P2_1/c$, $P2_12_12_1$, $Pbca$, Pc , $Pbcn$, $Pca2_1$, and $Pna2_1$. It can use any 12-6-1 potential and we selected the OPLS-AA FF [71]. The packing stage is divided in UPACK into two steps. In the first step, only 500 reasonable structures per symmetry group are randomly generated in an unrestricted way and are then used to estimate cell dimensions. In the second step, the random generation is performed in a restricted coordinate space using this cell estimate. Most of the generated structures are immediately rejected using the criterion that atom-atom 12-6-1 interactions are not allowed to be larger than 2000

kJ/mol for any pair. Such generations plus energy criterion testings continue until 5000 polymorphs per symmetry group, i.e., the total of 65,000 polymorphs are found. This second step involves also a rough optimization of lattice energies. The resulting list is subjected to clustering [72] to remove duplicates. Clustering reduces the pool significantly. For example, for system XXII it is reduced to 13,014 polymorphs.

In Stage 4 of CSP(aiFF) realized with UPACK, all the polymorphs from the reduced set are minimized with tight thresholds. We have modified UPACK to be able to use FFs of the form of Eq. (1). We found that it is advantageous to perform Stage 4 first with the OPLS-AA FF, i.e., using the original UPACK path including clustering, and then minimize the reduced set using aiFF. The procedure was chosen not to save time, although it does result in minor savings, but to avoid minimizations ending up in “holes” of an FF, i.e., unphysical minima at very short intermonomer separations. By construction, 12-6-1 FFs do not have any holes, while exp-6-1 and our extended-form FFs almost always have holes (although behind about 100 kJ/mol barriers, one of constraints of the autoPES fitting). We found that aiFF minimizations starting from the OPLS-minimized structures almost never end up in holes. We could have easily avoided the use of OPLS by fitting a 12-6-1 FF to the *ab initio* data.

The two CSP packages modified by us produced almost identical predictions for cases where we used both. MOLPAK was used for systems I, II, XII, XXII, nitromethane, and benzene. UPACK was used for the remaining systems, as well as for system I, II, and XXII treated also by MOLPAK. For these three systems, rankings of the experimental crystal by the two packages were identical.

PLATON [73] was used for checking missed symmetries [74] and for space group transformations from non-standard setting to standard setting by assigning the target crystal the proper space group and cell parameters, leading to the data in Table I. For example, for system II both MOLPAK and UPACK predicted the experimental crystal in $P2_1/c$ symmetry, and PLATON transformed it to $P2_1/n$ symmetry.

pDFT+D calculations. In Stage 5, periodic single-point DFT+D lattice energy calculations, i.e., without geometry optimizations, were performed for the 20 top-ranked polymorphs from aiFF minimizations using the PBE [51] functional with pseudopotentials [75] plus the D3 dispersion correction [52] with the Becke-Johnson (BJ) damping [76, 77]. We used Quantum ESPRESSO

(QE) [78, 79] codes, with the plane-wave kinetic energy cutoffs of 340 and 3061 eV for the wave functions and charge densities, respectively.

The zero-point vibrational energy (ZPVE) and thermal effects were calculated within the harmonic approximation using Phonopy 2.8.1 [80] and VASP 5.4.4 [81–85] with the same DFT+D approach as applied in QE. The projector augmented-wave pseudopotentials [86, 87] were used. For the relaxation of the crystal, a cutoff of 1000 eV for the plane-wave basis set was used. The relaxation was stopped if the total energy change between two steps for electronic and ionic motions were smaller than 10^{-5} and $0.5 \cdot 10^{-2}$ eV, respectively. Phonon calculations were performed at the Γ -point using a supercell of at least 10 Å length in each direction. Similarly to the relaxation step, a cutoff of 1000 eV for the plane-wave basis set and a convergence threshold of 10^{-8} eV were used in the total energy calculation. Next, ZPVE and thermal effects were calculated on a mesh of $8 \times 8 \times 8$ using the dynamical matrix built from the force constants of the displaced atoms in the supercell.

Data availability

The data that support the findings of this study are included within the Article and Supplementary Information. In particular, the .zip file contains coordinates and energies of all computed data points, parameters of the fits, and the crystallographic information files for a set of top-ranked polymorphs.

Code availability

The codes used for electronic structure calculations, fitting, CSPs, and pDFT+D calculations: SAPT, ORCA, autoPES (part of the SAPT package), MOLPAK, UPACK, Quantum Espresso, and VASP are available on the web and the links are provided in references of the main paper and the Supplementary Information. A patch to UPACK is available on the SAPT web site. A FORTRAN program computing the fitted potentials is included in the Supplementary_Data_1.zip file.

References

- [1] Cruz-Cabeza, A. J., Reutzel-Edens, S. M. & Bernstein, J. Facts and fictions about polymorphism. *Chem. Soc. Rev.* **44**, 8619–8635 (2015).
- [2] Bučar, D., Lancaster, R. W. & Bernstein, J. Disappearing polymorphs revisited. *Angew. Chem. Int. Ed.* **54**, 6972–6993 (2015).
- [3] Hilfiker, R. & von Raumer, M. (eds.) *Polymorphism in the Pharmaceutical Industry: Solid Form and Drug Development* (Wiley-VCH, Weinheim, Germany, 2019).

- [4] Bauer, J. *et al.* Ritonavir: An extraordinary example of conformational polymorphism. *Pharm. Res.* **18**, 859–866 (2001).
- [5] Chemburkar, S. R. *et al.* Dealing with the impact of ritonavir polymorphs on the late stages of bulk drug process development. *Org. Process Res. Dev.* **4**, 413–417 (2000).
- [6] Waknine, Y. Rotigotine patch recalled due to drug crystallization. *Medscape* 9 April, 2008.
- [7] Rietveld, I. B. & Ceolin, R. Rotigotine: Unexpected polymorphism with predictable overall monotropic behavior. *J. Pharm. Sci.* **104**, 4117–4122 (2015).
- [8] Mortazavi, M. *et al.* Computational polymorph screening reveals late-appearing and poorly-soluble form of rotigotine. *Comm. Chem.* **2**, 70 (2019).
- [9] Badgular, D. M., Talawar, M. B., Asthana, S. N. & Mahulikar, P. P. Advances in science and technology of modern energetic materials: An overview. *J. Hazard. Mat.* **151**, 289–305 (2008).
- [10] Zhang, C. Origins of the energy and safety of energetic materials and of the energy & safety contradiction. *Propellants Explos. Pyrotech.* **43**, 855–856 (2018).
- [11] Jurchescu, O. D. *et al.* Effects of polymorphism on charge transport in organic semiconductors. *Phys. Rev. B* **80**, 085201 (2009).
- [12] Maddox, J. Crystals from first principles. *Nature* **335**, 201 (1988).
- [13] Gavezzotti, A. Are crystal structures predictable? *Acc. Chem. Res.* **27**, 309–314 (1994).
- [14] Lommerse, J. P. M. *et al.* A test of crystal structure prediction of small organic molecules. *Acta Cryst. B* **56**, 697–714 (2000).
- [15] Motherwell, W. D. S. *et al.* Crystal structure prediction of small organic molecules: A second blind test. *Acta Cryst. B* **58**, 647–661 (2002).
- [16] Day, G. M. *et al.* A third blind test of crystal structure prediction. *Acta Cryst. B* **61**, 511–527 (2005).
- [17] Day, G. M. *et al.* Significant progress in predicting the crystal structures of small organic molecules - A report on the fourth blind test. *Acta Cryst. B* **65**, 107–125 (2009).
- [18] Bardwell, D. A. *et al.* Towards crystal structure prediction of complex organic compounds - A report on the fifth blind test. *Acta Cryst. B* **67**, 535–551 (2011).
- [19] Reilly, A. M. *et al.* Report on the sixth blind test of organic crystal-structure prediction methods. *Acta Cryst. B* **72**, 439–459 (2016).
- [20] Nyman, J. & Day, G. M. Static and lattice vibrational energy differences between polymorphs. *CrystEngComm* **17**, 5154–5165 (2015).
- [21] Ryan, K., Lengyel, J. & Shatruck, M. Crystal structure prediction via deep learning. *J. Am. Chem. Soc.* **140**, 10158–10168 (2018).
- [22] Neumann, M. A., Leusen, F. J. J. & Kendrick, J. A major advance in crystal structure prediction. *Angew. Chem. Int. Ed.* **47**, 1–5 (2008).
- [23] Neumann, M. A. Tailor-made force fields for crystal-structure prediction. *J. Phys. Chem. B* **112**, 9810–9829 (2008).
- [24] Neumann, M. A. GRACE; Avant-garde Materials Simulation: St-Germain-en-Laye, France (2008). URL <https://www.avmatsim.eu/>.
- [25] Day, G. M. Current approaches to predicting molecular organic crystal structures. *Crystallogr. Rev.* **17**, 3–52 (2011).
- [26] Price, S. L. Predicting crystal structures of organic compounds. *Chem. Soc. Rev.* **43**, 2098–2111 (2014).
- [27] Szalewicz, K. Determination of structure and properties of molecular crystals from first principles. *Acc. Chem. Res.* **47**, 3266–3274 (2014).
- [28] Hoja, J., Reilly, A. M. & Tkatchenko, A. First-principles modeling of molecular crystals: structures and stabilities, temperature and pressure. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **7**, e1294 (2017).
- [29] Oganov, A. R. Crystal structure prediction: Reflections on present status and challenges. *Faraday Discuss. R. Soc. Chem.* **211**, 643–660 (2018).
- [30] Price, S. L. & Brandenburg, J. G. Molecular crystal structure prediction. In de-la Roza, A. O. & DiLabio, G. A. (eds.) *Non-Covalent Interactions in Quantum Chemistry and Physics*, 333–363 (Elsevier, 2017).
- [31] Price, S. L. Is zeroth order crystal structure prediction (csp_0) coming to maturity? What should we aim for in an ideal crystal structure prediction code? *Faraday Discuss. R. Soc. Chem.* **211**, 9–30 (2018).
- [32] Podeszwa, R., Bukowski, R., Rice, B. M. & Szalewicz, K. Potential energy surface for cyclotrimethylene trinitramine dimer from symmetry-adapted perturbation theory. *Phys. Chem. Chem. Phys.* **9**, 5561–5569 (2007).
- [33] Podeszwa, R., Rice, B. M. & Szalewicz, K. On predicting structure of molecular crystals from first principles. *Phys. Rev. Lett.* **101**, 115503 (2008).
- [34] Misquitta, A. J., Welch, G. W. A., Stone, A. J. & Price, S. L. A first principles prediction of the crystal structure of C₆Br₂ClFH₂. *Chem. Phys. Lett.* **456**, 105–109 (2008).
- [35] Garcia, J., Podeszwa, R. & Szalewicz, K. SAPT codes for calculations of intermolecular interaction energies. *J. Chem. Phys.* **152**, 184109 (2020).
- [36] Jeziorski, B., Moszyński, R. & Szalewicz, K. Perturbation theory approach to intermolecular potential energy surfaces of van der Waals complexes. *Chem. Rev.* **94**, 1887–1930 (1994).
- [37] Metz, M. P., Piszczatowski, K. & Szalewicz, K. Automatic generation of intermolecular potential energy surfaces. *J. Chem. Theory Comput.* **12**, 5895–5919 (2016).
- [38] Metz, M. P. & Szalewicz, K. Automatic generation of flexible-monomer intermolecular potential energy surfaces. *J. Chem. Theory Comput.* **16**, 2317–2339 (2020).

- [39] Brandenburg, J. G. & Grimme, S. Organic crystal polymorphism: A benchmark for dispersion-corrected mean-field electronic structure methods. *Cryst. Acta B* **52**, 502–513 (2016).
- [40] Whittleton, S. R., de-la Roza, A. O. & Johnson, E. R. Exchange-hole dipole dispersion model for accurate energy ranking in molecular crystal structure prediction. *J. Chem. Theory Comput.* **13**, 441–450 (2017).
- [41] Hoja, J. *et al.* Reliable and practical computational description of molecular crystal polymorphs. *Sci. Adv.* **5**, eaau3338 (2019).
- [42] Garcia, J. & Szalewicz, K. *Ab Initio* extended Hartree-Fock plus dispersion method applied to dimers with hundreds of atoms. *J. Phys. Chem. A* **124**, 1196–1203 (2020).
- [43] Taylor, D. C. *et al.* Blind test of density-functional-based methods on intermolecular interaction energies. *J. Chem. Phys.* **145**, 124105 (2016).
- [44] Williams, D. E. Improved intermolecular force field for molecules containing H, C, N, and O atoms, with application to nucleoside and peptide crystals. *J. Comput. Chem.* **22**, 1154–1166 (2001).
- [45] Breneman, C. M. & Wiberg, K. B. Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis. *J. Comp. Chem.* **11**, 361 (1990).
- [46] Welch, G. W. A., Karamertzanis, P. G., Misquitta, A. J., Stone, A. J. & Price, S. L. Is the induction energy important for modeling organic crystals? *J. Chem. Theory Comput.* **4**, 522–532 (2008).
- [47] Karamertzanis, P. G. *et al.* Modeling the interplay of inter- and intramolecular hydrogen bonding in conformational polymorphs. *J. Chem. Phys.* **128**, 244708 (2008).
- [48] Greenwell, C. *et al.* Overcoming the difficulties of predicting conformational polymorph energetics in molecular crystals via correlated wavefunction methods. *Chem. Sci.* **11**, 2200–2214 (2020).
- [49] Neese, F. The ORCA program system. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2**, 73–78 (2012).
- [50] Neese, F. *ORCA, An Ab Initio, DFT, and Semiempirical electronic structure package*. with contributions from U. Becker, *et al.* URL <https://orcaforum.kofo.mpg.de>.
- [51] Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
- [52] Grimme, S., Antony, J., Elrich, 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.* **132**, 154104 (2010).
- [53] Kendall, R. A., Dunning, Jr., T. H. & Harrison, R. J. Electron-affinities of the 1st-row atoms revisited - systematic basis-sets and wave-functions. *J. Chem. Phys.* **96**, 6796–6806 (1992).
- [54] Szalewicz, K. Symmetry-adapted perturbation theory of intermolecular forces. *Wiley Interdisc. Rev.-Comp. Mol. Sci.* **2**, 254–272 (2012).
- [55] Misquitta, A. J., Podeszwa, R., Jeziorski, B. & Szalewicz, K. Intermolecular potentials based on symmetry-adapted perturbation theory including dispersion energies from time-dependent density functional calculations. *J. Chem. Phys.* **123**, 214103 (2005).
- [56] Hesselmann, A., Jansen, G. & Schütz, M. Density-functional theory-symmetry-adapted intermolecular perturbation theory with density fitting: A new efficient method to study intermolecular interaction energies. *J. Chem. Phys.* **122**, 014103 (2005).
- [57] Misquitta, A. J., Jeziorski, B. & Szalewicz, K. Dispersion energy from density-functional theory description of monomers. *Phys. Rev. Lett.* **91**, 033201 (2003).
- [58] Hesselmann, A. & Jansen, G. Intermolecular dispersion energies from time-dependent density functional theory. *Chem. Phys. Lett.* **367**, 778–784 (2003).
- [59] Bukowski, R., Podeszwa, R. & Szalewicz, K. Efficient calculations of coupled Kohn-Sham dynamic susceptibility functions and dispersion energies with density fitting. *Chem. Phys. Lett.* **414**, 111–116 (2005).
- [60] Podeszwa, R., Bukowski, R. & Szalewicz, K. Density fitting methods in symmetry-adapted perturbation theory based on Kohn-Sham description of monomers. *J. Chem. Theory Comput.* **2**, 400–412 (2006).
- [61] Bukowski, R. *et al.* SAPT2020: An *ab initio* program for many-body symmetry-adapted perturbation theory calculations of intermolecular interaction energies. University of Delaware and University of Warsaw (2020). URL <http://www.physics.udel.edu/~szalewic/SAPT/SAPT.html>.
- [62] Grüning, M., Gritsenko, O. V., van Gisbergen, S. J. A. & Baerends, E. J. Shape corrections to exchange-correlation potentials by gradient-regulated seamless connection of model potentials for inner and outer region. *J. Chem. Phys.* **114**, 652–660 (2001).
- [63] Cencek, W. & Szalewicz, K. On asymptotic behavior of density functional theory. *J. Chem. Phys.* **139**, 024104–(1:27) (2013). Erratum: **140**, 149902 (2014).
- [64] Williams, H. L., Mas, E. M., Szalewicz, K. & Jeziorski, B. On the effectiveness of monomer-, dimer-, and bond-centered basis functions in calculations of intermolecular interaction energies. *J. Chem. Phys.* **103**, 7374–7391 (1995).
- [65] Tang, K. T. & Toennies, J. P. An improved simple-model for the van der Waals potential based on universal damping functions for the dispersion coefficients. *J. Chem. Phys.* **80**, 3726–3741 (1984).
- [66] Day, G. M., Motherwell, W. D. S. & Jones, W. Beyond the isotropic atom model in crystal structure prediction of rigid molecules: Atomic multipoles versus point

- charges. *Cryst. Growth Des.* **5**, 1023–1033 (2005).
- [67] Price, S. L. Computational prediction of organic crystal structures and polymorphism. *Int. Rev. Phys. Chem.* **27**, 541–568 (2008).
- [68] Holden, J. R., Du, Z. & Ammon, H. L. Prediction of possible crystal structures for C-, H-, N-, O-, and F-containing organic compounds. *J. Comput. Chem.* **14**, 422–437 (1993).
- [69] van Eijck, B. P. & Kroon, J. UPACK program package for crystal structure prediction: Force fields and crystal structure generation for small carbohydrate molecules. *J. Comput. Chem.* **20**, 799–812 (1999).
- [70] Busing, W. R. Report ORNL-5747. Oak Ridge National Laboratory, Oak Ridge, TN (1981).
- [71] Jorgensen, W. L. & Tirado-Rives, J. The OPLS [Optimized Potentials for Liquid Simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **110**, 1657–1666 (1988).
- [72] Van Eijck, B. P. & Kroon, J. Fast clustering of equivalent structures in crystal structure prediction. *J. Comput. Chem.* **18**, 1036–1042 (1997).
- [73] Spek, A. L. Single-crystal structure validation with the program PLATON. *J. Appl. Crystallogr.* **36**, 7–13 (2003).
- [74] Spek, A. L. Structure validation in chemical crystallography. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **65**, 148–155 (2009).
- [75] Rappe, A. M., Rabe, K. M., Kaxiras, E. & Joannopoulos, J. D. Optimized pseudopotentials. *Phys. Rev. B: Condens. Matter* **41**, 1227 (1990).
- [76] Becke, A. D. & Johnson, E. R. Exchange-hole dipole moment and the dispersion interaction. *J. Chem. Phys.* **122**, 154104–(1:5) (2005).
- [77] Grimme, S., Ehrlich, S. & Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comp. Chem.* **32**, 1456–1465 (2011).
- [78] Giannozzi, P. *et al.* Quantum ESPRESSO: A modular and open-source software project for quantum simulations of materials. *J. Phys.: Condens. Matter* **21**, 395502 (2009). URL <http://www.quantum-espresso.org>.
- [79] Giannozzi, P. *et al.* Advanced capabilities for materials modelling with Quantum ESPRESSO. *J. Phys.: Condens. Matter* **29**, 465901 (2017).
- [80] Togo, A. & Tanaka, I. First principles phonon calculations in materials science. *Scr. Mater.* **108**, 1–5 (2015).
- [81] Kresse, G. & Hafner, J. *Ab initio* molecular dynamics for liquid metals. *Phys. Rev. B* **47**, 558–561 (1993).
- [82] Kresse, G. & Hafner, J. *Ab initio* molecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in germanium. *Phys. Rev. B: Condens. Matter* **49**, 14251–14269 (1994).
- [83] Kresse, G. & Furthmüller, J. Efficiency of *ab-initio* total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci* **6**, 15–50 (1996).
- [84] Kresse, G. & Furthmüller, J. Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set. *Phys. Rev. B: Condens. Matter* **54**, 11169–11186 (1996).
- [85] Kresse, G. & *et al.* VASP: Vienna *ab initio* simulation package (2021). URL <http://www.vasp.at>.
- [86] Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B: Condens. Matter* **50**, 17953–17979 (1994).
- [87] Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B: Condens. Matter* **59**, 1758–1775 (1999).

Acknowledgements. This work was supported by the U.S. Army Research Laboratory and Army Research Office (Grant No. W911NF-19-1-0117) and the NSF (Grant No. CHE-1900551). We thank Rafal Podeszwa for comments on the manuscript.

Author Contributions. R.N. and K.S. designed the method. R.N. coded it and performed numerical calculations. Both authors analyzed the results, wrote the manuscript, and revised it.

Additional Information. **Supplementary information** The online version contains supplementary material available at

Competing Interests. The authors declare no competing interests.

Correspondence. Correspondence and requests for materials should be addressed to Krzysztof Szalewicz (email: szalewicz@udel.edu).

TABLE I: **CSPs from SAPT(DFT)-based aiFFs minimizations followed by pDFT+D fixed-geometry calculations.** SG: predicted space group of the crystal (SG is the same for experimental and predicted polymorphs); Rank: rank of the experimental polymorph after minimizations and after pDFT+D calculations; RMSD₂₀: root mean square deviation (in Å) between the experimental crystal and the calculated polymorph for 20 overlapping molecules (heavy atoms only); RMSE: root mean square error (in kJ/mol) of the fit for negative interaction energies.

System	SG	Rank	RMSD ₂₀	RMSE
I _{Poly1}	<i>P2₁/c</i>	2/1	0.09	0.6
I _{Poly2}	<i>Pbca</i>	8/2	0.32	0.6
II	<i>P2₁/n</i>	1/1	0.59	1.3
IV	<i>P2₁/c</i>	2/1	0.24	0.63
VIII	<i>C2/c</i>	4/1	0.28	1.1
XII	<i>Pbca</i>	9/1	0.53	0.84
XIII	<i>P2₁/c</i>	4/1	0.45	1.1
XVI	<i>Pbca</i>	16/1	0.29	1.0
XXII	<i>P2₁/n</i>	1/1	0.15	1.4
Methanol	<i>P2₁2₁2₁</i>	6/1	0.4	0.92
Benzene _{Poly1}	<i>Pbca</i>	1/1	0.16	0.59
Benzene _{Poly2}	<i>P2₁/c</i>	4/3	0.4	0.59
Nitromethane	<i>P2₁2₁2₁</i>	1/1	0.27	0.74
DNBT	<i>P2₁/c</i>	1/1	0.58	1.56
TNB	<i>P2₁/c</i>	3/1	0.67	1.28
Deferiprone _{Poly1}	<i>Pbca</i>	2/2	0.28	0.71
Deferiprone _{Poly2}	<i>P2₁/c</i>	8/1	0.24	0.71
Fluorouracil	<i>P2₁/c</i>	9/1	0.61	1.06

FIG. 1: **Overview of aiFF-based CSP protocol.**

Stage 1: monomer energy minimization to find the equilibrium geometry. Stage 2: *ab initio* calculations of dimer intermolecular interaction energies followed by fitting an analytic form of aiFF to these data. Stage 3: generation of millions of plausible packing arrangements of polymorphs by sampling different space groups, orientations of monomers, and unit cell parameters, followed by a reduction of this set to tens of thousands of polymorphs using density criteria or crude lattice energy minimizations with simple FFs. Stage 4: fine minimization with aiFFs for all polymorphs in the reduced set. Stage 5: refinement of the ranking via pDFT+D calculations on a couple dozen top-ranked polymorphs from Stage 4.

FIG. 2: **Overlaps of crystal structures.** Overlap of the experimental crystal structure (element-specific colors) with the closest calculated crystal structure (green) using SAPT(DFT)-based aiFFs for systems: (a) and (b) I, (c) II, (d) IV, (e) VIII, (f) XII, (g) XIII, (h) XVI, (i) XXII, (j) methanol, (k) and (l) benzene, (m) nitromethane, (n) DNBT, (o) TNB, (p) and (q) Deferiprone, (r) Fluorouracil.

FIG. 3: Computational cost of the considered CSP protocols. Total wall times required for system I CSPs on a single core of the Intel E5-2670 processor using different strategies. Rows “aiFF”, “PACK+MIN”, and “pDFT+D” denote times of an aiFF development, packing and minimization with an aiFF, and of periodic DFT+D calculations.