Efficient crystal structure prediction for structurally related molecules with accurate and transferable tailor-made force fields

Alessandra Mattei, Richard S. Hong, Hanno Dietrich, Dzmitry Firaha, Julian Helfferich, Yifei Michelle Liu, Kiran Sasikumar, Nathan S. Abraham, Rajni Miglani Bhardwaj, Marcus A. Neumann, Ahmad Y. Sheikh

a AbbVie Inc., Solid State Chemistry, Research & Development, 1 N Waukegan Road, North Chicago, IL 60064, United States
b Avant-garde Materials Simulation, GmbH, Alte Str. 2, 79249 Merzhausen, Germany
† Current address: JobRad GmbH, Heinrich-von-Stephan-Str. 13, 79100 Freiburg, Germany

Abstract

Crystal structure prediction (CSP) has been historically used to complement experimental solid form screening and applied to individual molecules in drug development. The fast development of algorithms and computing resources offers the opportunity to use CSP earlier and for a broader range of applications in the drug design cycle. This study presents a novel paradigm of CSP specifically designed for structurally related molecules, referred to as Quick-CSP. The approach prioritizes more accurate physics through robust and transferable tailor-made force fields (TMFFs), such that significant efficiency gains are achieved through the reduction of expensive ab initio calculations. The accuracy of the TMFF is increased by the introduction of electrostatic multipoles and the fragment-based force field parameterization scheme is demonstrated to be transferable for a family of chemically related molecules. The protocol is benchmarked with structurally related compounds from the Bromodomain and Extraterminal (BET) domain inhibitors series. A new convergence criterion is introduced that aims at performing only as many ab initio optimizations of crystal structures as required to locate the bottom of the crystal energy landscape within a user-defined accuracy. The overall approach provides significant cost savings ranging from three to eight-fold less than the Full-CSP workflow. The reported advancements expand the scope and utility of the underlying CSP building blocks as well as their novel reassembly to other applications earlier in the drug design cycle to guide molecule design and selection.
1 Introduction

Crystal structure prediction (CSP) of organic molecules aims to generate and rank the crystal structures with various three-dimensional (3-D) packing arrangements of molecules, starting from the chemical diagram alone. The field has made substantial progress since the challenge laid down in the seminal commentary by Dunitz in the early 2000s.\(^1\) This progress is best manifested by the results of the latest blind test organized by the Cambridge Crystallographic Data Center (CCDC),\(^2\) a sixth in the series where the predictions have become more accurate, despite increased chemical complexity and conformational flexibility.\(^2\)-\(^7\) Overall, much of the underlying research focus is directed towards accurately calculating the relative energies of all putative crystal packing arrangements. Hence, all CSP approaches are structured hierarchically, so that the overall workflow resembles a funnel of increasingly narrow criteria and accurate energy models. Relatively less demanding calculations based on force fields are used to generate a large number of candidate crystal structures, while more advanced first-principles computational methods are applied for the final ranking of the most promising crystal structures.

Tailor-made force fields (TMFFs) represented an initial breakthrough in the field, allowing sufficiently accurate calculations of the energy minima associated with crystal packings.\(^8\)-\(^9\) Currently, TMFFs and first-principles energy ranking approaches with dispersion-inclusive hybrid functionals\(^10\)-\(^11\) and many-body dispersion interactions\(^12\)-\(^13\) are the most successful methods for yielding reliable and accurate lattice energies.\(^14\)-\(^15\) More recently, molecular dynamics and enhanced molecular dynamics have also been rigorously utilized to probe free energy stabilities and to eliminate labile structures.\(^16\)-\(^18\) Such algorithms have been thoroughly benchmarked and widely applied to predict the crystal energy landscape of organic molecules, including complex chemical systems of pharmaceutical relevance. In parallel, modern machine learning potentials combined with state-of-the-art first-principles calculations have been demonstrated to approach the accuracy required for predictive stability rankings with affordable computational costs.\(^18\)-\(^19\) In addition, the unprecedented level of parallelization available through cloud computing has permitted the resource-intensive CSP calculations to be completed in weeks, if not days.\(^20\)

Driven by the fast development of algorithmic improvements and powerful computing resources, CSP has fueled considerable interest across the pharmaceutical industry to complement experimental solid form screening.\(^21\)-\(^24\) Contemporary prediction of crystal energy landscapes is
routinely used to assess and manage the robustness of the solid form selected for development. In the context of drug design, when CSP is performed earlier in the design cycle, crucial insights can be gained by understanding the structural determinants of key physicochemical properties inherently rooted in essential crystal structure features, such as conformational flexibility and intermolecular interactions. Beyond these downstream calculations, fundamental building blocks of the CSP workflow, such as force fields, enhanced Monte-Carlo simulations, and dispersion-corrected density functional theory (DFT-D), are also broadly applicable in the molecular design itself. The need for computational efficiency in computer aided drug design often leads to reliance on traditional transferable force fields, such as OPLS, AMBER, CHARMM, COMPASS, and GROMOS. While limited in accuracy by their underlying functional form, these generic force fields are parameterized with experimental and quantum mechanics data for small organic molecules and/or macromolecules, conferring good parameter transferability. On the other hand, TMFFs are parameterized on molecule-by-molecule basis and thus are more accurate. In addition to requiring more computational effort, they are obviously not transferable. Before the immense value of TMFFs can be realized for broader applications beyond CSP, improvements on both the transferability and computational efficiency are therefore needed. The limitations of standard CSP workflows even when executed at ever-increasing speeds leave many opportunities that, if addressed, can expand the scope and utility of the underlying building blocks and their novel reassemblies.

In this study, we present a new paradigm of CSP specifically designed for structurally related molecules, referred to as Quick-CSP, which improves the TMFF accuracy, broadens the concept of fragment-based parameterization to impart transferability, and significantly reduces the extent of ab initio calculations for unprecedented levels of overall efficiency and versatility. We demonstrate the performance of this novel approach on moderately flexible molecules in the pyrrolopyridone-based series of Bromodomain and Extraterminal (BET) domain inhibitors, where the implications of subtle chemical modifications on conformational flexibility, crystal packing, and the resulting aqueous crystalline solubility have been recently reported.
2 Results and Discussion

Figure 1 shows key differentiating features of the Quick-CSP workflow compared to the standard Full-CSP, as implemented in the GRACE\textsuperscript{41} software suite. Here the accuracy of the TMFF is improved by the introduction of electrostatic multipoles attached to local atomic coordinate systems. The fragment-based approach for force field parameterization is expanded to a family of chemically related molecules, resulting in higher accuracy and some level of transferability. Furthermore, the efficiency of the fragment-based approach systematically increases as the number of molecules under consideration increases and the approach is scalable to chemical series of virtually any size. In the Quick-CSP workflow, CSPs are run for all compounds in parallel, relaxing the convergence requirements. A key feature is that an estimated bottom energy of the crystal energy landscape is determined through a statistical analysis together with estimates of multiple crystal structures’ DFT-D energies and their associated error. This allows convergence with only a few, robustly selected crystal structures to be DFT-D optimized, ultimately reducing the computational cost.

![Figure 1. Comparison of Quick-CSP and Full-CSP workflows with the differences highlighted in blue text.](image-url)
The Quick-CSP approach was applied to BET bromodomain inhibitors, which have generated significant interest for studying various types of cancers, including acute myeloid leukemia (AML) and multiple myeloma (MM).\textsuperscript{42-44} Recent structure-activity-relationship (SAR) efforts have led to the identification of pyrrolopyridone core derivatives that demonstrated sub-micromolar activity in binding assays against BRD4 and substantial efficacy in several \textit{in vivo} mouse xenograft models.\textsuperscript{45} This family of molecules shows sufficient diversity in terms of both the features related to developability within the context of Rule of 5 (Ro5) framework and structural variance to challenge each step of the Quick-CSP and Full-CSP workflows.

2.1 Features of BET pyrrolopyridone core molecules

Eight molecules of the BET bromodomain inhibitors were selected in this study from the lead optimization efforts.\textsuperscript{45} As displayed in Table 1, these molecules feature a pyrrolopyridone ring that acts as a hydrogen bond acceptor and donor through the pyrrolopyridone carbonyl and the pyrrole amino proton, with the ability of forming a bidentate hydrogen bond. The hydrogen bonding groups and dominant molecular shape are conserved when small chemical modifications are made on the distal ether group or the central aryl ring and/or when the functional group attached to the central aryl ring (i.e., sulfonamide, sulfone, or reverse amide moieties) is substituted. The eight molecules show a moderate conformational flexibility with major torsion angles being around the terminal chain attached to the central aryl ring ($\Phi_1$), the distal cyclic ether group ($\Phi_2$), and the pyrrolopyridone ring ($\Phi_3$). The facile reorientation of the functional group attached to the central aryl ring, as measured by $\Phi_1$, is not sterically impeded by the distal ether or the pyrrolopyridone moieties, due to the strong cooperativity between neighboring torsion angles.\textsuperscript{29} The three key flexible torsion angles can however impact the degree of $\pi$-delocalization or conjugation throughout the molecule. Notably, the positioning of the sulfonamide moiety allows for some degree of conjugation through the central aryl ring and the pyrrolopyridone ring with a more planar conformation, as measured by $\Phi_1$ and $\Phi_3$, showing the maximum conjugation. Thus, the pyrrolopyridone-based molecules can adopt a variety of 3-D molecular conformations with different relative positions of the hydrogen bond acceptors and donors.\textsuperscript{29}

The range of 3-D molecular conformations can in turn result in different crystal packing arrangements and specific intermolecular interactions in the crystalline state. Electrostatic forces
between the polar and aromatic pyrrolopyridone core molecules are expected to play an important role in determining their crystal structures because of the large dipole moment (Table S1). Similarly, $\pi-\pi$ stacking intermolecular interactions, involving the distal ether aryl and/or the pyrrolopyridone rings, can contribute considerably to stabilize the crystal packing. While it is not always clear a priori, the intricacy of this family of molecules suggests the need of a TMFF that accounts for the electrostatics effects of features, such as $\pi$-electron density, on hydrogen bonding and aromatic $\pi$-stacking interactions to improve accuracy upfront in a CSP workflow.

**Table 1.** BET bromodomain inhibitors with chemical modifications on the pyrrolopyridone scaffold. The three flexible torsion angles are also indicated.
2.2 A more accurate TMFF method

The Quick-CSP workflow was run twice using the multipole TMFF in one case and point charge TMFF in the other. For each molecule, the accuracy of the backfitted TMFF was assessed by comparing the lattice energies of force field optimized and ab initio optimized crystal structures, the latter being taken from the eight Full-CSP runs. The root-mean-square deviation (σ), expressed in terms of an error per atom, between the TMFF and PBE-NP\textsuperscript{10} lattice energies was then calculated. The multipole TMFF shows systematically lower errors than the point charge TMFF across the pyrrolopyridone-based molecules (Figure 2). The multipole TMFF contains more parameters that allow for a better description of the electron density distribution, resulting in more accurate lattice energies. The large number of parameters can, however, make the fitting process more complicated and result in decreased accuracy if the fitting algorithm gets trapped in an unfavorable local minimum. The results demonstrate that GRACE’s force field parameter fitting algorithm transformed the increased complexity, through the introduction of electrostatic multipoles, into increased accuracy. In addition, switching point charge to multipole TMFFs reduces the force field error by 16.2% after backfitting and 14.2% before backfitting (Figure S1 and Table S2). The multipole TMFF shows the strongest improvements for the molecules containing a sulfonamide moiety (i.e., molecules 1–6). Figure 3a provides more detailed insights into the performance of the two TMFFs compared to the PBE-NP benchmark for molecule (2), which exhibits the largest difference in force field accuracies of the molecules with the sulfonamide moiety. The most striking feature of the correlation plot is an enhanced agreement of the multipole TMFF energies with PBE-NP calculations that results in a tighter crystal lattice energy distribution. As expected from the smaller σ-value, the multipole TMFF provides improved lattice energies of the predicted crystal structures, yielding a mean absolute error (MAE) of 1.21 kcal/mol. In contrast, there appears to be a significant scatter in the predictions of the relative energies computed by the point charge TMFF, leading to a large MAE of 1.57 kcal/mol.
Figure 2. Accuracy, as measured by $\sigma$-value, of backfitted point charge (grey) and multipole (blue) TMFFs for each pyrrolopyridone-based molecule. Reporting the $\sigma$-value per atom has the advantage of removing the dependency on the number of atoms.

Closer inspection of the crystal lattice energies revealed that the point charge TMFF can miscalculate the strength of hydrogen bonding interactions. The largest negative deviation relative to PBE-NP calculations occurred for the predicted crystal structure that shows a bidentate hydrogen bond between pyrrolopyridone rings and lacks intermolecular hydrogen bonding interactions involving the sulfonamide moiety (Figure 3a top inset). The latter structural feature leads to a large electron density on the electronegative sulfonamide group. The specific description of electrostatic effects by the lone pair electrons and the electron density distribution can make a significant difference to the ability of representing the directionality and strength of hydrogen bonding interactions. The error of the point charge TMFF may thereby originate from systematically underestimating the electrostatic interactions and overestimating the hydrogen bond strength, as the directionality of the intermolecular hydrogen bonding is not accounted. On the other hand, the multipole TMFFs account for the anisotropy of the electron density around the nitrogen and oxygen atoms of the sulfonamide group and considerably reduce the energy error.
Further assessment of the crystal lattice energy distribution shows that the multipole TMFF performs well for relatively planar molecular conformations that allow the \( \pi \)-delocalization or conjugation to extend to the central aryl ring via the nitrogen lone pair of the sulfonamide group. This is consistent with the electrostatic potential around conjugated systems, aromatic systems, and/or lone pairs being more realistically reproduced by electrostatic multipole models.\textsuperscript{47} In addition, as a result of the planar molecular conformation and extended \( \pi \)-delocalization, the predicted crystal structure with a large positive deviation displays aromatic \( \pi \)-stacking interactions that are known to be more correctly modeled by electrostatic multipole functionals (Figure 3a bottom inset).\textsuperscript{48} In such a case, the point charge TMFF error stems from the under-stabilization of the extended \( \pi \)-delocalization as well as the underestimation of the intermolecular aromatic \( \pi \)-stacking interactions present in the more planar conformation. Similar behavior occurred for the other molecules containing the sulfonamide moiety with a MAE range of 1.09–1.26 kcal/mol in the predictions of the relative energies computed by the multipole TMFF. A MAE range of 1.37–1.57 kcal/mol was obtained by the point charge model (Figures S2a-e).

For molecules (7) and (8) that contain the sulfonyl group and the reverse amide directly attached to the central aryl ring, respectively, both TMFFs yield crystal lattice energies that are in very good agreement with the reference PBE-NP method, as indicated by the low and comparable MAE values (Figure 3b and Figures S2f). The origin of these results may reflect the extent to which molecular conformations, intermolecular interactions, and their interplay contribute toward the stability of molecular organic crystals. The incorporation of a hydrogen bond acceptor directly attached to the central aryl ring alters the extent of the overall \( \pi \)-delocalization in the system. For instance, the molecular conformation of compound (7) in the crystal structure that shows a large positive energy error exhibits a \( \Phi_1 \) of about 90°, a position that completely “breaks” the conjugation between the sulfonyl group and the central aryl ring (Figure 3b inset). Accordingly, it decreases the overall \( \pi \)-delocalization of the system. Therefore, the subtle difference in \( \sigma \)-values between the two TMFFs may reflect the negligible contribution of the degree of electron delocalization to the accurate description of the molecular conformation.
Figure 3. Comparison of backfitted point charge (grey) and multipole (blue) TMFFs relative energies to the corresponding reference PBE-NP energies for a) molecule (2) and b) molecule (7). Insets show crystal structures with large deviations in TMFFs relative energies with respect to PBE-NP.

2.3 Fragment-based force field for structurally related molecules

Since many molecular fragments share force field parameters, the accuracy of these parameters depends on the transferability between fragments. To evaluate if the force field parameterization for a family of structurally related molecules results in decreased accuracy, the error per atom of
the combined TMFF with multipoles was compared to that of the TMFF derived for individual molecules. The $\sigma$-values obtained from the combined TMFF are slightly lower than those from the individual TMFFs (Figure 4), indicating that for a family of molecules the TMFF accuracy does not decrease due to the lack of transferability or the difficulty of fitting more parameters. On the contrary, the higher data-to-parameter ratio results in a better definition of force field parameters and/or a more robust fitting behavior. These results refer to the backfitted TMFF. On average, backfitting reduces the energy error by 5.6% and 4.7% in Full-CSP and Quick-CSP workflows, respectively (Table S3). The small size of these changes demonstrates that the TMFFs are well calibrated, even without backfitting, as such removing the backfitting in the Quick-CSP workflow may be acceptable to further reduce the central processing unit (CPU) time requirements. However, additional validation on other structurally related compounds is required to ensure similar performance.

The number of molecular fragments used for the individual and combined TMFFs is shown in Figure 5a. The individual force field generation for the eight molecules in the Full-CSP workflow used a total of 40 and 19 fragments for the force field parameterization of intramolecular interactions (bonded fragments) and additional torsion terms (torsion fragments), respectively. The combined force field generation in the Quick-CSP workflow employed 20 and 10 fragments for intramolecular interactions and torsion terms, respectively. The 2-fold reduction is reflected in the CPU time requirements for the force field parameterization. Generating the combined TMFF for the pyrrolopyridone core molecules indeed decreased the CPU timings by a factor of 1.9 compared to the cumulative time of the individual TMFFs (Figure 5b).

Most significantly, the combined TMFF does not only achieve an improved accuracy and speedup compared to the individual TMFF, it also provides transferability across the chemical series. This represents a significant advancement, because to date, each new compound required a newly and individually parameterized TMFF. A key factor underlying this success is that the transferability requirement shifts from the force field parameters themselves, as in traditional transferable force fields, to the shared molecular fragments and the methodology of generating force field parameters. To quantitatively assess transferability improvements, the accuracy of the combined TMFF was compared with that obtained by the COMPASS II force field,49 a newly parameterized, transferable force field. The combined TMFF for the pyrrolopyridone-based molecules
outperforms the COMPASS II force field in reproducing the lattice energies of putative crystal structures, giving a lower $\sigma$-value than that of the COMPASS II force field by a factor of 2 (Figure S3). This ability to employ an accurate TMFF that exhibits transferability within structurally related molecules represents a paradigm shift in computing crystal energy landscapes with unprecedented efficiency and accuracy. In addition, it opens the opportunity for CSP components to be leveraged for computer-guided molecule design in advance of molecular synthesis to make accurate yet computationally affordable predictions of physicochemical properties, including solubility, permeability, and potency.\textsuperscript{30,33,50}

\textbf{Figure 4.} Accuracy, as measured by $\sigma$-value, of the combined TMFF generated in the Quick-CSP workflow (blue), backfitted to each molecule, and the backfitted TMFF derived for each molecule in the Full-CSP workflow (green).
2.4 Crystal energy landscapes

Completeness of predicted structures within a given energy window and energy accuracy represent important metrics to judge the performance of a CSP calculation. For molecules (1), (2), and (6) with available experimental crystal structures or powder diffraction patterns, we calculated the energies of the top 20 ranked crystal structures from each Full-CSP at the PBE-NP and PBE(0)+MBD+F\text{vib} level of theory (Figure 6). Crystal energy landscapes of the three molecules and comparison to experimental crystal forms are provided in Figures S4 and S5, respectively. PBE(0)+MBD+F\text{vib} is currently considered the most accurate and affordable method with an error of the order of 0.25 kcal/mol.\textsuperscript{51} The root-mean-square deviations between the two methods are 0.85, 0.35, and 0.63 kcal/mol for the three molecules, respectively. Note that these values include not only the inaccuracy of the PBE-NP method, but also the neglect of zero-point vibrational energy and finite temperature enthalpy and entropy contributions. It is thus concluded that the PBE-NP method has an inherent inaccuracy of about 0.5 kcal/mol for the pyrrolopyridone core molecules. The Quick-CSP calculations were configured to determine the bottom energy of the crystal energy landscapes within an accuracy of 0.5 kcal/mol.
Figure 7 plots the energies from the Quick-CSP and Full-CSP at the PBE-NP level of theory. Across all eight molecules, most of the low-energy structures of the crystal energy landscape from Quick-CSP were identified and reproduced with high-fidelity compared to the Full-CSP method. In the Quick-CSP workflow, the expected value for the energy at the bottom of the landscape and its error bar were calculated and displayed in Figure 7. Table 2 shows that the difference between the actual bottom energy computed in the Full-CSP workflow and the estimate bottom energy calculated in the Quick-CSP workflow is smaller than the calculated standard deviation of the estimate for five molecules and is at the 1σ threshold for one more molecule. For a Gaussian distribution, one would expect 5.5 cases to fall inside one standard deviation on average. The results therefore demonstrate the ability to locate the bottom of the crystal energy landscape within a user-defined accuracy.

To further highlight the advantages of Quick-CSP over the established Full-CSP workflow, the computational resources employed for crystal structure generation and ranking were compared. The Quick-CSP workflow provides a substantial reduction in computational cost, associated with crystal structure generation and final energy ranking, as shown in Figure 8. The extent of cost savings however varied on a molecule-by-molecule basis due to the diversity of crystal packings that these molecules can adopt because of different spatial arrangements of substituents. It ranged from the lowest saving by a factor of 2.7 for molecule (3) to the highest by a factor of 7.8 for molecule (5). It is worth emphasizing that the extent of these savings, while variable and difficult to predict a priori, are significant when compared with the marginal gains achieved through a variety of conventional efficiency gain strategies in standard CSP workflows for these molecules.
Figure 6. Energies of a) molecule (1), b) molecule (2), and c) molecule (6) crystal structures as predicted at the PBE-NP and the PBE(0)+MBD+F_vib level of theory in the Full-CSP workflow. Experimentally observed crystal forms are highlighted in red, while all other predicted structures are in grey.
Figure 7. Absolute energies of crystal structures as predicted by the Full-CSP and Quick-CSP workflows. The PBE-NP optimized crystal structures are shown in grey, while crystal structures with an estimate of the PBE-NP energy are highlighted in orange. The estimate for the bottom of the energy window from the Quick-CSP workflow and its error bar are also displayed in black.

Table 2. Energy difference between the actual bottom energy from the Full-CSP workflow and the estimated bottom energy from the Quick-CSP workflow and the standard deviation ($\sigma$) of the estimated bottom energy.

<table>
<thead>
<tr>
<th>Compound ID</th>
<th>$\Delta$ Bottom Energy (kcal/mol)</th>
<th>$\sigma$ (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.099</td>
<td>0.272</td>
</tr>
<tr>
<td>2</td>
<td>0.494</td>
<td>0.493</td>
</tr>
<tr>
<td>3</td>
<td>-0.577</td>
<td>0.348</td>
</tr>
<tr>
<td>4</td>
<td>0.371</td>
<td>0.499</td>
</tr>
<tr>
<td>5</td>
<td>0.185</td>
<td>0.368</td>
</tr>
<tr>
<td>6</td>
<td>-0.110</td>
<td>0.427</td>
</tr>
<tr>
<td>7</td>
<td>0.337</td>
<td>0.414</td>
</tr>
<tr>
<td>8</td>
<td>0.601</td>
<td>0.489</td>
</tr>
</tbody>
</table>

a. $\sigma$ is expressed as error per molecule that is related to the error per atom by Gaussian error propagation.

Figure 8. Comparison of the computational cost, associated with the crystal-structure generation and reranking, between the Full-CSP (green) and Quick-CSP (blue) workflows for each pyrrolopyridone-based molecule. Numbers indicate the factor of computational cost savings obtained by the Quick-CSP workflow.
3. Conclusions

The BET case study shows that the first-of-a-kind Quick-CSP offers an attractive speed-accuracy compromise that allows for the earlier and broader use of the predicted crystal structures. The underlying design concept of Quick-CSP prioritizes more accurate physics earlier in the workflow through robust and transferable TMFFs. The TMFF, a necessary building block for modern CSP, can be generated cost-effectively for families of structurally related molecules by exploiting economy of scale, building on the fragment-based force field parameterization approach implemented in GRACE. For the BET family of molecules, it has been demonstrated that force field parameters are transferable between fragments resulting in improved accuracy when a combined TMFF is derived for the entire family rather than parameterizing individual TMFFs. The use of crystal structure reference data for the entire molecules in the fitting procedure, known as backfitting in the context of CSP, only results in a marginal improvement of the force field accuracy and may be omitted.

The completeness and accuracy requirements for the generated crystal energy landscapes are relaxed to the generation of representative crystal packings and the identification of the most stable crystal structure within a targeted accuracy. The ability to perform only as many ab initio crystal structure optimizations as required to locate the bottom of the crystal energy landscape affords a significant cost saving, ranging from a factor of three up to eight, compared to the Full-CSP workflow. Combining this cost and CPU time reduction with parallel execution of the individual CSPs with one molecule per asymmetric unit as part of the Quick-CSP workflow, crystal energy landscapes for all molecules of the family could be obtained within a couple of days. The described gains open the opportunity for CSP applications in early stages of the drug/material design cycle, where crystal packing can be utilized for various key physicochemical property prediction and the resulting structural insights can be fed into rational molecular design and developability assessment. At the same time, the highly sophisticated, accurate, and modular CSP building blocks can be utilized in other workflows related for instance to conformational scoring and molecular dynamics in the drug/material discovery cycle to guide molecule design and selection.
4. Computational Methods

All Full-CSP calculations were carried out using the computer program GRACE (version 3.0)\(^1\) on the Rescale cloud computing platform.\(^2\) The Quick-CSP workflow was run on in-house LINUX clusters. Details of the two hardware configurations and the translation from the CPU time consumption on the in-house LINUX cluster to that on Rescale are described in the supplementary material.

4.1 Ab initio methods

Most ab initio calculations in this study were carried out with the Vienna Ab Initio Simulation Package (VASP)\(^3-5\) and the Perdew-Burke-Ernzerhof density functional,\(^6-7\) which was augmented with a dispersion correction by Neumann & Perrin, referred to as PBE-NP.\(^8\) The method is also referred to as DFT-D in this work. VASP was used with a plane wave cutoff energy of 520 eV and a k-point spacing of roughly 0.07 Å\(^{-1}\).

Some low-energy crystal structures were further optimized at the PBE(0)+MBD+F\(_{\text{vib}}\) level of theory using the all-electron FHI-aims (Fritz Haber Institute ab initio molecular simulations) package for the calculation of ab initio energies and forces.\(^9\) Crystal structures are first optimized using the PBE-NP with the light basis set, the hybrid PBE(0) functional,\(^10\) and the many-body dispersion (MBD) model.\(^11-15\) In addition, the phonon energy (F\(_{\text{vib}}\)) was evaluated in the harmonic approximation with some additional corrections using PBE-NP with the light basis set.

In the force field generation procedure, isolated molecule reference data were generated with Turbomole 7.2 Rev 1.\(^16-18\) The PBE-NP/def2-SV(P) method with geometry counterpoise (gCP) correction\(^19\) for the def2-SV(P) basis set\(^20\) was used to perform coarse gas phase molecular optimization. Subsequently, the PBE-NP/aug-cc-pVDZ method with gCP for the aug-cc-pVDZ basis set\(^21\) was applied to compute the energies and forces used in the force field parameter fitting.

4.2 TMFF

TMFFs in GRACE feature atomic point charges calculated from bond increments, isotropic van der Waals interactions using the Leonard-Jones-9-6 potential, and anharmonic intramolecular bond, angle, inversion as well as torsion terms. For the van der Waals parameters, the full interaction matrix including off-diagonal terms is fitted. The description of Coulomb interactions was improved by the addition of electrostatic dipole and quadrupole moments, defined with respect to local atomic coordinate systems that move with each atom and its covalently bonded neighbors.
A detailed description of the reference data and the fitting process is published elsewhere. Instead of using predefined Force Field Atom Types (FFATs), GRACE constructs on-the-fly, as required, canonical names that describe local chemical environments. The canonical names take into account bond properties, such as bond order and aromaticity, and atomic properties, such as atomic number and formal charge. Primary FFATs consider the central atom and its bonds and are used to index the van der Waals parameters. Secondary FFATs consider the central atom, its bonds, its neighbors, and their associated bonds and are used to index all other parameters.

Large molecules and families of molecules are cut automatically into three types of smaller, chemically reasonable fragments. When bonds need to be cut, they are terminated with hydrogen atoms or methyl groups. The smallest fragments, called non-bonded fragments, are used to parameterize van der Waals interactions. The fragments of intermediate size, called bonded fragments, define most of the intramolecular and electrostatic force field parameters. The largest fragments, termed torsion fragments, are used to parameterize some torsion terms that are not covered by the bonded fragments.

The force field parameterization procedure occurs in three stages, from the smallest to the largest fragments. At each stage, force fields are generated first for the individual fragments with parameters available from a previous stage being kept constant. In the first stage, the individual force field parameterization is executed for all non-bonded fragments and all pairs of non-bonded fragments and stored in a database. The CPU time requirements reported in this work refer to a complete database (i.e., all force field parameterizations for non-bonded fragments and their pairs were fetched from the database). At the second stage, each bonded force field parameterization also includes a small amount of interaction data for the bonded fragment with the non-bonded fragments to achieve a balanced description of intra- and intermolecular interactions. The new force field parameters introduced for the individual fragments are combined by a Monte Carlo genetic algorithm procedure, which removes conflicts between parameters for the same interaction. Finally, a local parameter optimization is carried out in which all force field parameters are fitted to the reference data simultaneously.

In this work, force field errors with respect to ab initio reference data are expressed in terms of an energy error per atom, $\sigma$. The error per molecule, $\sigma_{\text{mol}}$, is related to the error per atom by Gaussian error propagation (i.e., $\sigma_{\text{mol}} = \sqrt{N} \times \sigma$ where $N$ is the number of atoms per molecule).
4.3 Full-CSP

To avoid force field mis-parameterization, each CSP started with a short and incomplete crystal structure generation from which a diverse set of crystal structures was selected for a training set and a test set. For both sets, PBE-NP lattice energy minimizations were carried out. The training set was added to the reference data for force field parameterization and all force field parameters were optimized against all the reference data. This force field parameter optimization is called backfitting, which generates an improved force field. The test set was used to assess the force field accuracy before and after backfitting. If the accuracy change was very large, the backfitting was repeated.

CSP was performed with one molecule per asymmetric unit ($Z'=1$) and consisted of three distinct steps. First, a Monte Carlo parallel tempering algorithm was used in conjunction with the TMFF to generate all relevant crystal packings in a certain energy window. The completeness of the crystal structure generation was estimated from the average number of times each crystal structure was generated and the fraction of unique generated structures. The crystal structure generation was stopped when 99% of all crystal structures had been generated. The energy window was determined by a user-defined target energy window (1 kcal/mol) plus n-times (usually n=4) the force field accuracy. In a second reranking step, force field optimized structures were subjected to ab initio minimization. Since ab initio optimizations of all generated structures would be computationally expensive, statistical correlations were built up between structural similarity, ab initio single point calculations on force field optimized structures and full ab initio optimizations. These statistical correlations were used to select crystal structures for further ab initio optimizations. The reranking always starts with 100 ab initio single point calculations and optimizations to initialize the statistical models. The details of these statistical models are rather intricate and go beyond the scope of this manuscript. The important point is that for each crystal structure there exists an ab initio energy estimation with known standard deviation. These energy estimations were used to calculate an estimated value for the number of structures in the user-defined target window. The reranking was terminated when the fraction of ab initio optimized structures in the target window divided by the estimated value was more than a certain threshold (typically 99%). Finally, all structures in the target energy window were further optimized with tighter convergence criteria.
4.4 Quick-CSP

First, a TMFF was parameterized for a family of chemically related molecules. Then, independent CSPs were executed simultaneously for all compounds of the family. Like in the Full-CSP workflow, backfitting was carried out. It is worth emphasizing that each CSP of the Quick-CSP workflow produces its own backfitted force fields (i.e., it tailors the transferable TMFF more towards the respective target molecule). The user-defined target energy window was 1 kcal/mol. In the crystal structure generation, a 98% convergence was targeted. During the reranking, statistical models were initialized with 100 ab initio calculations, but compared to the Full-CSP workflow a different convergence criterion was used. The reranking stopped when the bottom of the energy landscape was determined within a certain tolerance. From the ab initio energies and their corresponding estimates, 1000 random landscapes were generated by considering the error bars of the estimates and an average bottom energy with standard deviation was then computed. As an additional condition for convergence, the most stable predicted crystal structure must be ab initio optimized. For the crystal structures that were not ab initio optimized, the Quick-CSP workflow reports the estimated energies and the TMFF optimized crystal structures.

AUTHOR INFORMATION

Corresponding Authors
Ahmad Y. Sheikh, AbbVie Inc., North Chicago, IL 60064, United States
ORCID ID 0000-0002-5972-3938; ahmad.sheikh@abbvie.com; Tel: 847 938 1251
Marcus A. Neumann, Avant-garde Materials Simulation, Merzhausen, Germany
ORCID ID 0000-0003-2857-7488; marcus.neumann@avmatsim.eu

Authors ORCID ID
Alessandra Mattei: 0000-0002-5194-6563
Rajni M. Bhardwaj: 0000-0002-7402-9995
Richard S. Hong: 0000-0002-4214-6847
Nathan S. Abraham: 0000-0002-4707-909X
Hanno Dietrich 0000-0003-2622-438X
Dzmitry Firaha: 0000-0002-6387-0754
Yifei Michelle Liu: 0000-0002-4705-6664
Kiran Sasikumar: 0000-0001-5207-013X

Author Contributions and Acknowledgements

AbbVie sponsored and funded the study, helped conceive the overall design, performed Full-CSP and other benchmarking calculations, and provided the structures and experimental data reported herein. Avant-garde Materials Simulation designed, developed, and validated the Quick-CSP workflow and performed Quick-CSP calculations for the compounds reported here. All authors contributed to the analysis and interpretation of data, and in writing, reviewing, and approval of the final publication. A.M., A.Y.S., R.S.H., R.M.B., and N.S.A. are employees of AbbVie and may own AbbVie stock. M.A.N., H.D., D.F., Y.M.L., and K.S. are employees of Avant-garde Materials Simulation Deutschland GmbH (AMS) and have no additional conflict of interest to disclose. J.H. is an employee of JobRad GmbH and has no additional conflict of interest to disclose.

The authors wish to thank AbbVie colleagues Shailendra Bordawekar, Jeffrey Bien, James A. Williams and Catherine Srebalus Barnes for the sponsorship and support to execute the Quick-CSP development program.

References


(27) Neumann, M. A.; van de Streek, J. How many ritonavir cases are there still out there? *Faraday Discuss.* 2018, 211 (0), 441-458.


(29) Hong, R. S.; Mattei, A.; Sheikh, A. Y.; Bhardwaj, R. M.; Bellucci, M. A.; McDaniel, K. F.; Pierce, M. O.; Sun, G.; Li, S.; Wang, L.; Mondal, S.; Ji, J.; Borchardt, T. B. Novel physics-based ensemble modeling approach that utilizes 3D molecular conformation and packing to access aqueous thermodynamic solubility: A case study of orally available bromodomain and


(41) Neumann, M. A. GRACE software for crystal structure prediction of molecular compounds. *Avant-garde Materials Simulations Deutschland GmbH* **2011**.


