nature portfolio

Peer Review File

A Robust Crystal Structure Prediction Method to Support Small Molecule Drug Development with Large Scale Validation and Blind Study

Corresponding Author: Dr Lingle Wang

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
In the manuscript entitled 'A Robust Crystal Structure Prediction Method to Support Small Molecule Drug Development with Large Scale Validation and Prospective Studies' describes a new approach to crystal structure prediction of organic molecular crystals with Z' = 1 and tests it on 65 various molecules, some of them being quite complex ones. It is nice to see a CSP study on such a large molecular set and it does seem that the approach is rather successful. However, I feel the Authors are overoptimistic in presenting their approach and do not openly discuss the limitations of their approach, apart from mentioning that for now it deals with Z' = 1 structures. The second major issue is a lack of the detailed explanation of some of the steps of this new approach – at times it really is very vogue, or at least it was not understandable to me. Below are some specific points I would like to raise.
Examples of the work overoptimism and/or lack of enough precision:
1. Page 11: 'The conformer generation protocol reliably samples conformations within an RMSD 0.40 Å' – this is an example of a potential limitations I see in the designed approach to CSP. The reported RMSD value from molecular overlay is rather high and such discrepancies can easily lead to missing some important crystal structures in the predictions. As none of the calculations were actually blind tests, it is at this stage impossible to judge the severity of this problem, but this is definitely the field for improvements and the potential source of difficulties in the presented approach and the Authors should be clear on that in the manuscript.
2. On what grounds is the 'unprecedented accuracy' claim based on? I do not feel this is justified by the presented data. To claim that the Authors would have to test one of the existing CSP approaches using the same set of molecules. Or at least show how much better their approach is in several cases. Meanwhile what they show is that the current approach is comparable at best.
3. Page 15: 'This approach differs from the tailor-made force-fields (TMFF) used commonly in other CSP approaches' - TMFF is used in one approach and that is in GRACE, so it is far from being 'commonly used in other approaches'
4. What do you mean by: '[This is]…the first use of machine learning force fields in a hierarchical crystal energy ranking.'? Several groups before used machine learning to improve the energy ranking from CSP (e.g. https://doi.org/10.1021/acs.jctc.9b00038, https://doi.org/10.1039/C7SC04665K, https://doi.org/10.1021/acs.jpca.0c05006) , maybe in a bit different capacity, but still it is not that the Authors use ML to improve the estimated energies of the crystals from CSP for the first time.
5. For cocaine and in several other places the Authors say that the energy difference of 4.18 kJ/mol (1kcal/mol) between the most stable structure and the next best one indicate 'low likelihood of finding different Z'=1 polymorphs'. At the same time, for example for rotigotine but also for some other molecules, the Authors themselves find a P212121 structure higher than the most stable one by more than 4.18 kJ/mol and there they suggest it is potentially observable form. So which is it? The same argument cannot be used when it is convenient for the narrative and ignored at other places. Many polymorphs have energy difference higher than 4.18 kJ/mol, especially polymorphs which are kinetically favoured. So this energy difference is not showing a low likelihood in my opinion.
6. A general comment to the two prospective applications: while it is nice to see this pharmaceutical industry perspective on the usefulness of the CSP calculations, without knowing the details of the experiments this part is of limited meaning. For example, in the second case it is not clear how typical the crystallization conditions that led to the discovery of new polymorphs were and whether they would be attempted regardless of the CSP results. Without such details it is impossible to judge how helpful CSP actually was here.

7. For galunisertib, the Authors said that in the previous study ‘form V [was] identified as the most stable form’ – this is not true. It was form VII that was found the most stable one among the experimentally observed forms, and among the Z’=1 polymorphs forms VIII, IX and X were found to have lower energy than form V. Also further the Authors say that contrarily to the previous studies, in their work form IX was predicted to be more stable than form V. – in the previous work this was also the case.
8. The Authors are saying that their method used less resources than the previous ones, but they account only for some part of the process and list only CPU time. What about GPU? If you want to have that kind of comparison to show the superiority of this approach, you need to be able to compare not only one stage (e.g. crystal generation stage), but the whole process and listing all the resources used.

Specific questions to the lack of details of the approach or explanation of how it actually works:
a. What is the way of selecting the conformations tested in CSP? What is the molecular cluster size while generating possible crystal packings? How do you decide that? How do you know they are low energy clusters - are they optimized somehow on the fly? How the distance between molecules within the cluster is decided? how do you determine a step size? How many packings are generated and how do you now it is enough? – these are the all important details of the approach that can show how it is comparable to other approaches. It is also not clear at what stage conformational flexibility is allowed in the generated crystals – is it at the stage of energy ranking only? If yes, then this is also a serious limitation of the approach. I understand that you must start somewhere, I just wish the Authors are open and honest about the potential limitations.
b. Page 15: ‘Fragments with low energy according to the OPLS4 force field are combined’ – what is this low energy at this point?
Some more trivial issues and specific questions to the studied systems:
- For olanzapine, did the Authors compared the potential structures of form III found in P21/c and Pbca space groups? What is the RMSD for 15-molecule cluster? P21/c is a subgroup of Pbca, so it often happens for CSP calculations to find the same or very similar structures in a subgroup of lower symmetry.
- Page 10: ‘During the IND-enabling stage of the drug development, experimental polymorph screening identified several polymorphs of the API, including one pure form (form XI) and a few multicomponent crystals (hydrates and solvates).’ - I strongly oppose naming solvate/hydrates or any multicomponent crystals "polymorphs" of the drug! It is more appropriate to say 'crystalline forms', the term 'polymorphs' should be reserved for crystals with the same composition.
- It’s a pity the Authors are not showing in the Supporting Information the CSP landscapes for all studied molecules.
- Molecules on the grey field in Figure 1 are hardly visible
- Caption to Figure 6 reports different value of the energy difference than the text.
- It might not be clear to the readers what is ‘IND-enabling stage’ or ‘ASU’

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The authors should fully address the following issues before further consideration for publication in any journal:

1) The results of drug formulation projects presented in figure 6 and the surrounding discussion cannot be verified or reproduced by others and should therefore be removed. The chemical structures and the experimentally observed crystal structures of these compounds are not provided. This means there is no way for other researchers to reproduce or verify the results presented in the manuscripts. In theory, the authors could have fabricated these plots and the whole story about the success of their methods in drug development and nobody would be able to disprove their claims. Results that cannot be verified do not belong in peer reviewed articles. If the authors cannot make the compounds and their known crystal structures publicly available they should delete these results, the surrounding discussion, and all unverifiable claims of success in pharmaceutical development projects. “Prospective studies” should also be removed from the title.

2) The authors make several false/ unsubstantiated claims of accuracy and novelty in the abstract, the conclusion, and other places throughout the manuscript. All of these claims should be corrected or deleted. Only the most egregious examples are quoted here:

2a) The authors claim multiple time that their method has “unprecedented accuracy and efficiency”. This claim can only be substantiated by direct comparison to other methods in a blind test setting. Therefore, all such claims should be deleted. Avant-garde Materials Simulation has a long-standing track record of success in CSP blind tests and newcomer Xtal-Pi has had comparable results in the 7th CSP blind test. Until Schrodinger put their methods to the test it remains to be seen how their performance compares to their prominent competitors. The CSP community is weary of people making unsubstantiated claims regarding the performance of their methods. For example, Krzysztof Szalewicz claimed in Nature Communications 13, 3095 (2022) to have developed a reliable CSP method, however in the 7th CSP blind test his team had a 0% success rate and failed to generate even a single structure.

2b) The authors falsely claim to be the first to use machine learned potentials for CSP. In fact, others have already used ANI potentials for CSP in J. Chem. Theory Comput. 19, 24, 9388 (2023). Moreover, multiple teams used machine learned potentials in the 7th CSP bind test and at least one of the authors who was on one of those teams, should be well aware of that.

2c) The authors repeatedly describe their structure generation method as novel and imply that it is superior to existing methods. However, their method is currently limited to the 22 most common space groups (a fact that is only revealed in the SI and swept under rug in the main text) and to Z'=1. Considering that some of the structure generation methods developed by others do not have these limitations, the authors cannot claim superiority. They should provide a balanced description of their method and its limitations and avoid unsubstantiated claims.

2d) In the conclusion the authors claim to have a "pretrained, transferable MLFF that accurately predicts the relative stability of different polymorphs without requiring extensive DFT calculations". This claim is not supported by the data presented in the manuscript. To the contrary, in the SI it is shown that their pretrained model ranks the experimentally observed structures of some compounds above 400 and even above 1000 in a couple of cases. The system specific models still rank some targets above 200. No benchmarks are presented of the MLFF geometry, energy, and ranking compared to DFT. In light of this, it seems that the statement quoted above wildly exaggerates the capabilities of the MLFF. The authors should provide an accurate description and acknowledge the limitations of their methods in the main text.

2e) Also in relation to the previous comment, the authors make a song and dance of how the GRACE tailor-made force field ranked the crystal structure of MK-8876 as 2290th. They compare the GRACE force field ranking to their own *DFT* ranking of the structures of MK-2022 and MK-8876 as 1st and 10th. This comparison seems misleading considering that the authors' pretrained MLFF ranks the structure of MK-2022 as ~900 according to Figure S1 and the MLFF ranking of MK-8876 is not even shown. To be fair, the data shown in the SI for the performance of the MLFF indicates that some structures could be similarly missed by the authors' workflow if discarded at an early stage based on the MLFF ranking. This issue with the MLFF should be acknowledged and discussed in the text.

3) The authors seem to take the DFT ranking of putative polymorphs as the ground truth. They repeatedly claim that their method "not only reproduces all the experimentally known polymorphs, but also suggests new low energy polymorphs yet to be discovered by experiment that might pose potential risks to development of the currently known forms of these compounds". They completely ignore the so-called over-prediction problem, the tendency of CSP methods to predict multiple low-energy polymorphs that are not in fact observed in reality [e.g., CrystEngComm 23, 5575 (2021); Cryst. Growth Des. 20, 6847 (2020)]. There should be a more nuanced discussion of whether these predictions are to be believed in the introduction and throughout the manuscript, in particular in cases such as MK-8876 and Galunisertib, where multiple putative polymorphs are predicted to be more stable than experimentally observed structures.

4) An analysis of whether the correct ordering of experimentally observed polymorphs is obtained is missing from the manuscript. The analysis provided in the SI for ROY and Galunisertib should be moved to the main text. Similar analysis should be performed in all cases where experimental data on relative energies exist. Furthermore, in many cases, even if there is no experimental information on the energy differences between polymorphs, there is information on the order of stability. An analysis should be presented of whether the correct ordering of stability is reproduced (in Figure 3 or as a separate figure). This should be accompanied by a rigorous discussion of the reliability of DFT methods for ranking.

5) It is unclear why only ROY, Galunisertib, Rotigotine, MK-8876, MK-2022, and LY-156735 were selected for discussion. It seems based on Figure 3 that there are several other interesting cases such as FuROY, oxalyl dihydrazide, axitinib, bicalutamide, diflunisal, and flufenaic acid to name a few. The authors should expand the detailed discussion of interesting cases. If this makes the manuscript too long for a Communication then it might be more appropriate for it to be in a more specialized journal. In addition, it would be nice to see the energy vs density plots of all targets with multiple polymorphs in the SI.

6) It seems like all or most of the information provided in the inappropriately named "Discussion" sub-section belongs in the SI. This and the removal of Figure 6 and associate discussion should make room for additional case studies of interest.

7) Issues with figures:
7a) Half of Figure 1 is a graphic with no scientific content. This should be removed
7b) In all figures showing relative energies, ranking, and energy vs density plots it should be clarified in the figure caption and in the discussion of the figure which method was used to produce the data shown
7c) In Figure 3c,d is there a significance to the colors? Colors could be used e.g., to indicate whether the correct stability ordering was obtained (see item 4 above)
7d) Only one structure is shown for Diflunisal in Figure 3. Where are the other three?

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
Review of the paper "A Robust Crystal Structure Prediction Method to Support Small Molecule Drug Development with Large Scale Validation and Prospective Studies"

The paper is well written and has a narrative which is easy to read also for non-experts in the field. I have some remarks on the novelty of the findings in this paper and some suggestions on clarifications needed to be added.

In the Introduction the authors write: “However, this process can be time consuming and may miss some important low energy polymorphs due to its inability to exhaust all crystallization conditions.”
As presented, the reader has not yet been introduced why a low energy polymorph is of interest. That should be clarified and exemplified.

It is also stated “Computational polymorph prediction can complement experiments to de-risk unexpected polymorphic changes during drug development”
A few references with clear industrial perspective are missing. For example Firaha et al in Nature (https://doi.org/10.1038/s41586-023-06587-3), Sun et al (https://doi.org/10.1021/acs.cgd.0c01266) , Broo et al (https://doi.org/10.1107/S2052520616006831). The recently published paper in Nature as given above, advances the area of CSP to another scientific level. In that study free energies are used to risk assess polymorph landscape for example, which is missing in this paper.

The present paper shows a wider scope than previously addressed (65 API:s) and also a novel method for crystal packing searching, in combination with ML-FF reranking. Half of the set has only one experimentally known crystal structure while the other half has more than one. Molecules were restricted to Z´=1 only. Nowadays it is standard in the industry to handle CSP for such systems. More interest lies in the more challenging cases of Z´=2 or multicomponent systems such as solvents or cocrystals. Thus, the scope of the present study is not very novel.

In the section "Summary of results"
Non-trivial duplicates are discussed and how they were removed. It would have been interesting to link this to disordered structures, where a small rotation or ring-flip at typical temperatures for API storage may lead to a disordered structure which is favored by configurational entropy, i.e. favored free energy due to increased entropy. It may be that the non-trivial duplicates actually are of interest and importance. See for example Putra et al (https://doi.org/10.1021/acs.cgd.1c01124)

It is also stated "For a few molecules, the relative energy is less than 0.5 kcal/mol, which is likely within the error bars of the DFT calculations"
How has the error bars been evaluated? No mentioning of that is found in the paper, so this is very vaguely described. A recent good validation paper is in https://doi.org/10.1038/s41586-023-06587-3)

Some unclarities in the Cocaine example:
r2SCAN-D3 is mentioned to have a different energy gap than PBE-D3, but it is not discussed in the paper why r2SCAN-D3 is the preferred method by any validation. In the supporting information this is discussed but it would be good to inform the reader already in the main text. Now it just appears without any rationale.

The section "Prospective applications on two drug formulation"
These are nice stories to tell, but it seems the authors are unaware this is standard procedure already in several pharmaceutical companies. This makes the study less novel.

In the Discussion section:
Some of the discussion on conformer generation and Tables better sits in the Supporting Information. Now that part of the manuscripts becomes quite heavy to read.

Section of Methods:
I was surprised to see that the crystal packing part uses rigid conformations. Most up-to-date software uses conformations free to relax in the packing stage. Only a small deviation in dihedral angles may give a significant impact on the packing arrangement. It would be good to discuss any comparisons and why rigid conformers has been chosen. Computer time is one part obviously but some thoughts on the selection would be informative.

Color labels in Figure 3c,d is not given, so context is unclear.

The paper also does not discuss the importance of adding temperature effects on top of the 0K CSP landscape. This has in several papers been shown to be of very big importance, already referenced in this review.

The presented method is a nice step forward for the software developers, and with impressive validation statistics. However, the lack of novelty and the fact that the protocol is missing some key parts such as free energy contributions does not merit a publication in Nature Communications. I would suggest publication in a journal with a more focused readership, such as Crystal Growth&Design or Molecular Pharmaceutics.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I am satisfied with the majority of the corrections the Authors have made to the manuscript with a small exception of referring, again, to other CSP approaches which use molecule-specific force fields in their calculations. In the reply the Authors said it is not only the case of GRACE, but that also DMACRYS and CrystalPredictor use these. This is not exactly right. First, DMACRSY is not a CSP code, it uses force fields and atomic multipoles for geometry optimization of crystals, but it is not capable by itself to conduct CSP calculations. Instead, it is used by other CSP codes (e.g. CrystalOptimizer and mol-CSPy Day's code) to optimize crystal geometries and/or to calculate their energies. Secondly, these approaches also use pretrained force fields (mostly FIT or Williams99) and not molecule-specific ones. I again encourage the Authors to be more precise in their reference to other CSP methods. I'd also like to point out that if the Authors are aiming to introduce a new approach to CSP it is advisable to know what are the basis of the other available codes, and not only that of their direct commercial competitors (GRACE, XTalPi).

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
In the revised version of the manuscript the authors have addressed most of my comments adequately.

I do have one major criticism remaining. I must insist that the authors remove the results of STC-1 and STC-2 from Figure 7 and all the related discussion from the text. I understand that some IP protections apply in this case, however RESULTS THAT CANNOT BE VERIFIED AND REPRODUCED BY OTHERS SIMPLY DO NOT BELONG IN THE SCIENTIFIC LITERATURE. With the addition of the Flupyradifurone molecule removing the other two would not take away from the impact of this manuscript. The authors may publish their work on STC-1 and STC-2 in the future when the IP protections are no longer needed and they can reveal the molecules and their crystal structures.

An optional change is to move Table 1 and the detailed discussion of computer time usage to the SI and only summarize it briefly in the main text

Additional minor comments:

In the discussion of the usage of machine learned force fields for molecular crystal structure prediction it would be appropriate to cite the 7th CSP blind test and discuss the usage of MLFFs by participants in both stages of the blind test

Because one of the authors' selling points is their new approach to crystal structure generation, it would be appropriate to explain the difference between their method and the methods used by others with references to other structure generation methods

In relation to the performance of DFT and importance of free energy it would be appropriate to cite: J. Chem. Theory Comput. 2022, 18, 4456–4471

In relation to ROY it would be appropriate to cite: CrystEngComm 2019, 21, 2080–2088

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The authors have carefully replied to most of the comments outlined for the initial version of the manuscript. I am pleased that several aspects of the concerns have been covered, such as introduction of a "blind test" molecule for transparency. In addition, free energies have been computed for the final rankings of polymorphs.

Although the manuscript has significantly improved, my major concern is still valid. This is a big study with some good benchmarking results. The major novelty claimed by the authors lies in the sampling method. The speedup examples comparing to other CSP protocols does not take into account that those studies were performed 5-6 years back as well so a direct comparison in cpu hours is questionable. The authors exemplify their methods on "...prospective applications on two drug formulations". Today this is routine daily work in the pharmaceutical industry, and has been covered in many publications. The authors themselves reply to one of the comments as: "We agree with the review that the applications domain covered in our paper may not be entirely novel. However, our sampling method represents a significant advancement." . But this does not merit a publication in Nature Communications. I recommend to publish in a more specialized journal.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I am happy with the provided revision and recommend publication of the work in Nature Communications.

(Remarks on code availability)

Reviewer 1

In the manuscript entitled 'A Robust Crystal Structure Prediction Method to Support Small Molecule Drug Development with Large Scale Validation and Prospective Studies' describes a new approach to crystal structure prediction of organic molecular crystals with Z' = 1 and tests it on 65 various molecules, some of them being quite complex ones. It is nice to see a CSP study on such a large molecular set and it does seem that the approach is rather successful. However, I feel the Authors are overoptimistic in presenting their approach and do not openly discuss the limitations of their approach, apart from mentioning that for now it deals with Z' = 1 structures. The second major issue is a lack of the detailed explanation of some of the steps of this new approach – at times it really is very vogue, or at least it was not understandable to me. Below are some specific points I would like to raise.

Examples of the work overoptimism and/or lack of enough precision:
1. Page 11: 'The conformer generation protocol reliably samples conformations within an RMSD 0.40 Å' – this is an example of a potential limitations I see in the designed approach to CSP. The reported RMSD value from molecular overlay is rather high and such discrepancies can easily lead to missing some important crystal structures in the predictions. As none of the calculations were actually blind tests, it is at this stage impossible to judge the severity of this problem, but this is definitely the field for improvements and the potential source of difficulties in the presented approach and the Authors should be clear on that in the manuscript.

We thank the reviewer for pointing out the importance of conformer sampling for successful crystal structure sampling. The reported RMSD of up to 0.4 Å seems high but does not seem to be a problem in practice for the following reasons:

A): The initial conformer pool only serves as a starting point for the rigid-conformer packing pattern search. During the MD, QRNN, and DFT optimizations, all degrees of freedom including the conformers in the candidate crystal structures are fully flexible to relax.
B): The 0.40 Å RMSD the reviewer quoted is the upper bound of the results on a very large set of more than 400 cases. For most cases, the RMSD is much smaller, in the range of 0.2-0.3 Å.
C): Even for the small number of cases where the best match RMSD is close to 0.4 Å, there are usually many conformers with RMSD in the range of 0.4-0.6 Å. We agree with the reviewer that sometimes a conformer with a RMSD of 0.4 Å resulting from the part of the molecule forming key interactions in the crystal structure may lead to the correct structure missed in the sampling. However, in practice there are always some conformers with seemingly high RMSD but does not materially affect the packing pattern.
D): We could relax the conformers during the packing parameter search step. Doing that will certainly increase the compute cost by a lot but it is not clear whether it will always improve the accuracy. Our test results on a large dataset indicates the current protocol achieves a good accuracy while balancing the compute cost.

2. On what grounds is the 'unprecedented accuracy' claim based on? I do not feel this is justified by the presented data. To claim that the Authors would have to test one of the existing CSP approaches using the same set of molecules. Or at least show how much better their approach is in several cases. Meanwhile what they show is that the current approach is comparable at best.

Our claim of accuracy is based on the following evidence. First, we have provided, in a single publication, results of successful predictions on 65 molecules encompassing 135 experimentally known polymorphs. No other single publication in the field of CSP offers such a diverse and comprehensive set of successful results. Second, we present specific comparisons where available, and show examples like Olanzapine, MK-2022, MK-8876, where our results are more accurate than the previously reported studies.

Still, we acknowledge the point by the reviewer that the statement of “unprecedented accuracy” needs some further evidence to fully support. In the revised manuscript, we removed the statement of “unprecedented accuracy” in all places of the paper.

3. Page 15: ‘This approach differs from the tailor-made force-fields (TMFF) used commonly in other CSP approaches’ - TMFF is used in one approach and that is in GRACE, so it is far from being 'commonly used in other approaches'

We thank the reviewer for pointing out the potential confusion regarding the term of TMFF. We intended to use the term "tailor-made force-field" (TMFF) as a generic term to describe customized force fields tailored to specific molecules or sets of molecules, rather than referring exclusively to GRACE. As referenced in the original TMFF publication[10.1021/jp710575h], the term is used in a general sense to describe such customized force fields. We acknowledge that the phrase TMFF is closely associated with GRACE, although similar protocols are also employed by many other CSP methods.

Examples of methods that create models of atomic interactions specific for an input molecule are DMACRYS [10.1039/C004164E], CrystalOptimizer [10.1021/ct100597e], and Xtalpi’s CSP force-field [10.1021/acs.cgd.8b01098].

To avoid the confusion, we rephrased the sentence to the following in the revised manuscript.

“This approach differs from other CSP approaches where customized parameters are trained specifically on one molecule or a set of molecules.[24,55] The resulting parameters from our method are transferable to different chemistries and for different applications, whereas the customized parameters used in other CSP methods only have limited scope of application.”

4. What do you mean by: ‘[This is]…the first use of machine learning force fields in a hierarchical crystal energy ranking.’? Several groups before used machine learning to improve

the energy ranking from CSP (e.g. https://doi.org/10.1021/acs.jctc.9b00038, https://doi.org/10.1039/C7SC04665K, https://doi.org/10.1021/acs.jpca.0c05006) , maybe in a bit different capacity, but still it is not that the Authors use ML to improve the estimated energies of the crystals from CSP for the first time.

We did not intend to claim we are the first workers to explore the use of a MLFF in the area of crystal structure prediction and we apologize for the confusion. To our knowledge, we are the first publicly available manuscript to use an MLFF ranking as an intermediary stage in a hierarchical ranking strategy involving a classical force field, MLFF and finally periodic DFT ranking. This approach has been validated in a thorough set of molecules with significant flexibility. To clarify how our work differs from the existing literature we have added a clarifying paragraph in the Supplementary Information with several citations of prior work.

Still, we acknowledge the technical differences between our methods and previous work may be subtle, and removed the potentially confusing statement about the novelty in the abstract and main text.

The abstract has been modified to read as follows

"The method combines a novel systematic crystal packing search algorithm and the use of machine learning force fields in a hierarchical crystal energy ranking."

5. For cocaine and in several other places the Authors say that the energy difference of 4.18 kJ/mol (1kcal/mol) between the most stable structure and the next best one indicate 'low likelihood]of finding different Z'=1 polymorphs'. At the same time, for example for rotigotine but also for some other molecules, the Authors themselves find a P212121 structure higher than the most stable one by more than 4.18 kJ/mol and there they suggest it is potentially observable form. So which is it? The same argument cannot be used when it is convenient for the narrative and ignored at other places. Many polymorphs have energy difference higher than 4.18 kJ/mol, especially polymorphs which are kinetically favoured. So this energy difference is not showing a low likelihood in my opinion.

We thank the reviewer for pointing out the inconsistencies in our discussions of cocaine versus other molecules. We fully agree with the reviewer that the energy difference of 4.18 kJ/mol between the most stable structure and the next best one for cocaine could rule out the possibility of other polymorphs. It does indicate though that there might not be other more stable polymorphs. In the revised manuscript, the discussion has been modified to read as follows:

"This further indicates the low likelihood of observing any additional more stable Z'=1 polymorph of cocaine."

6. A general comment to the two prospective applications: while it is nice to see this pharmaceutical industry perspective on the usefulness of the CSP calculations, without knowing

the details of the experiments this part is of limited meaning. For example, in the second case it is not clear how typical the crystallization conditions that led to the discovery of new polymorphs were and whether they would be attempted regardless of the CSP results. Without such details it is impossible to judge how helpful CSP actually was here.

We thank the reviewer for the excellent point. In the received manuscript, we made it clear that the additional crystal screening experiments were motivated by the CSP, results, with the following sentence:

"Indeed, additional crystal screening experiments motivated by our calculations identified a few other crystalline forms that were found to interconvert with form XXVI under different process conditions."

In addition, we included an additional CSP result from a blinded study on an agrochemical molecule developed by Bayer complete with all necessary information for reproducibility of our blindly computed CSP landscape.

7. For galunisertib, the Authors said that in the previous study 'form V [was] identified as the most stable form' – this is not true. It was form VII that was found the most stable one among the experimentally observed forms, and among the Z'=1 polymorphs forms VIII, IX and X were found to have lower energy than form V. Also further the Authors say that contrarily to the previous studies, in their work form IX was predicted to be more stable than form V. – in the previous work this was also the case.

We thank the reviewer for the suggestion to re-examine the publication describing the experimental solid form landscape of Galunisertib [10.1021/jacs.9b06634]. The first mention of stability is in the Results in the Experimental Solid-Form Landscape section. At the beginning of the third paragraph, it reads, "Forms IV and VI, the most stable of the neat polymorphs, were crystallized directly, and frequently concomitantly, from a variety of solvents." The next mention of experimental stability is in the Thermodynamic Stability section.

> "***Thermodynamic Stability.*** *The thermal properties of GAL forms I−VII were examined by DSC using fast heating (50 ° C min−1) to minimize decomposition. The melting points of these seven polymorphs spanned a range of more than 125 ° C, with form VI appearing to be the most stable (highest melting) form at high temperatures (SI Figure S17). Melting enthalpies were reliably measured for forms II−VII, for which the heats of fusion varied by up to 4 kJ mol−1 (Figure 6). The measured enthalpies of desolvate forms II and III were consistently lower than those of forms IV−VII, which were within the margin of error of one another (SI Table S12)."*

The referenced Figure 6 is reproduced below. In this figure, The experimentally determined relative melting enthalpy of form V seems to be below that of form VI and VII which are followed by form IV, whereas form IV is stated as being the most stable form at room temperature in the figure caption. Our assertion that Form V is the most stable is based on the raw experimental data in Figure 6 presented in the referenced paper. In the same figure, Form VII is reported to be the most stable experimentally known form by the PBE+NP periodic DFT method. It's possible that our assertion that form V is the most stable was confused as an assertion about computational results, whereas we were referring to the experimental form with the largest melting enthalpy.

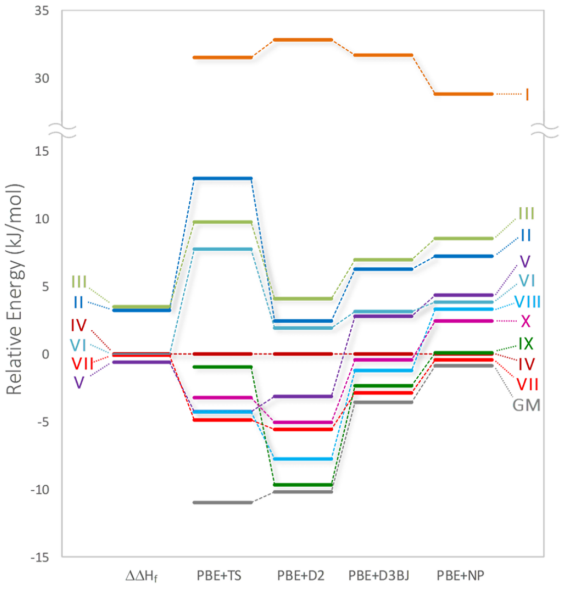

Figure 6. Experimental relative enthalpies of the **GAL** polymorphs contrasted with relative lattice energies of the crystal forms and the CSP-generated GM structure, calculated using various periodic DFT+D methods. All polymorphs are ranked in energy relative to the room temperature stable polymorph, form IV.

We revised this sentence in our manuscript for clarity. It now reads as follows.

"For Galunisertib, solid form screening experiments produced many solvates and ten anhydrous forms (form I – form X), with forms IV-VII found to be the most stable, having experimentally measured enthalpies within the margin of error of one another."

In addition, we have improved the clarity of the analysis of our results in comparisons to the results of others. This paragraph now reads as follows.

"For Galunisertib, forms VIII, IX, X, were predicted to be more stable than forms V and VI according to the r^{2}SCAN-D3 functional. Our second most stable candidate structure is in the $P2_12_12_1$ space group, is yet to be observed experimentally, and matches the structure reported as the global minimum in prior computed CSP landscapes.[38] Using the conformer energy correction, the order of stabilities of the Z'=1 experimentally known forms remains largely

unchanged (Figure S7), in agreement with prior computed results.[39] Adding room temperature free energy calculations (Figure S18(e) in SI) indicates that form VIII becomes less stable with the increase of temperature and forms IX and X remain more stable than forms V and VI across the computed temperature range."

8. The Authors are saying that their method used less resources than the previous ones, but they account only for some part of the process and list only CPU time. What about GPU? If you want to have that kind of comparison to show the superiority of this approach, you need to be able to compare not only one stage (e.g. crystal generation stage), but the whole process and listing all the resources used.

We have provided a complete account of both CPU and GPU usage in Table 3 of our submitted manuscript where one GPU hour is converted to 10 CPU hours.

In the revised manuscript, in addition to the caption of table 1, we also made it clear in the discussion that our comparison included both CPU and GPU costs, which now reads as follows.

"In Table 1, we list the compute cost for each step of the polymorph prediction workflow for a representative set of molecules with increasing complexities. Our CSP workflow is significantly more computationally efficient than CPU costs reported previously by other approaches. To facilitate comparisons, we convert our GPU costs to CPU costs using a 1:10 conversion. For PF-998245, our CPU cost is 22.6K hours, whereas approximately 200K hours was reported in Ref 12. For rotigotine, our CPU cost is 26.5K hours whereas 125K CPU hours was reported in Ref 28."

Specific questions to the lack of details of the approach or explanation of how it actually works:
a. What is the way of selecting the conformations tested in CSP? What is the molecular cluster size while generating possible crystal packings? How do you decide that? How do you know they are low energy clusters - are they optimized somehow on the fly? How the distance between molecules within the cluster is decided? how do you determine a step size? How many packings are generated and how do you now it is enough? – these are the all important details of the approach that can show how it is comparable to other approaches. It is also not clear at what stage conformational flexibility is allowed in the generated crystals – is it at the stage of energy ranking only? If yes, then this is also a serious limitation of the approach. I understand that you must start somewhere, I just wish the Authors are open and honest about the potential limitations.

We thank the reviewer for the interest in all details of the method. In the revised manuscript, we provide as much details as possible about the method, including the criteria for conformer selection, how the intermediate clusters are constructed and cluster sizes for possible packing search. An example pseudocode describing the packing search method is also available in the

SI. We also clearly describe in what stages the conformers are fully relaxed, and the limitations of the method as compared to the stochastic approach employed by other CSP packages.

Some more trivial issues and specific questions to the studied systems:
- For olanzapine, did the Authors compared the potential structures of form III found in P21/c and Pbca space groups? What is the RMSD for 15-molecule cluster? P21/c is a subgroup of Pbca, so it often happens for CSP calculations to find the same or very similar structures in a subgroup of lower symmetry.

We thank the reviewer for the excellent question. The deduplication and clustering performed in the periodic DFT steps compares all structures regardless of space group assignment. The $P2_1/c$ structure differs significantly from the Pbca structure. As discussed in the manuscript and shown in Figure 5, "Both the putative form III* in the Pbca space group and the newly discovered $P2_1/c$ structure have identical 2D layers as the experimental form II, but the packing along the third dimension differs, similar to the two forms of Aspirin.[22]" $RMSD_n$ calculation between these two structures finds 21/34 of the molecular centers of masses are within comparison tolerances but with an overall RMSD of 2.47. The number of molecules with the centers of masses matching is large because the lattice parameters of the two structures are close, but nevertheless the structures are significantly distinct from one another. We also further improved Figures 5(i) and 5(j) in the revised manuscript to highlight the difference between these two structures.

We would also like to remind the reviewer that we have already made all of the low energy crystal structures available. We welcome the reviewer to confirm this comparison for themselves.

- Page 10: 'During the IND-enabling stage of the drug development, experimental polymorph screening identified several polymorphs of the API, including one pure form (form XI) and a few multicomponent crystals (hydrates and solvates).' - I strongly oppose naming solvate/hydrates or any multicomponent crystals "polymorphs" of the drug! It is more appropriate to say 'crystalline forms', the term 'polymorphs' should be reserved for crystals with the same composition.

We thank the reviewer for the suggestion, and have changed this to now read as follows.

"During the Investigational New Drug (IND) enabling stage of the drug development, experimental crystallization screening identified several crystalline forms of the API, including one pure form (form XI) and a few multicomponent crystals (hydrates and solvates)."

- It's a pity the Authors are not showing in the Supporting Information the CSP landscapes for all studied molecules.

We deposited all the predicted structures along with their relative energies in the github repository shared in the SI. We feel it is better to provide the raw data for any reader to analyze and validate the low-energy crystal structure landscapes produced by our CSP method.

- Molecules on the grey field in Figure 1 are hardly visible

We thank the reviewer for pointing this out. We improved Figure 1 in the revised manuscript.

- Caption to Figure 6 reports different value of the energy difference than the text.

The manuscript text reported that the energy gap is “at least 0.8 kcal/mol” and the Figure 6 caption reports the value is “about 0.9 kcal/mol”. We have updated the text to match the Figure 6 caption.

- It might not be clear to the readers what is ‘IND-enabling stage’ or ‘ASU’

We explicitly defined IND in the resized manuscript.

Reviewer 2

The authors should fully address the following issues before further consideration for publication in any journal:

1) The results of drug formulation projects presented in figure 6 and the surrounding discussion cannot be verified or reproduced by others and should therefore be removed. The chemical structures and the experimentally observed crystal structures of these compounds are not provided. This means there is no way for other researchers to reproduce or verify the results presented in the manuscripts. In theory, the authors could have fabricated these plots and the whole story about the success of their methods in drug development and nobody would be able to disprove their claims. Results that cannot be verified do not belong in peer reviewed articles. If the authors cannot make the compounds and their known crystal structures publicly available they should delete these results, the surrounding discussion, and all unverifiable claims of success in pharmaceutical development projects. “Prospective studies” should also be removed from the title.

We appreciate the reviewer’s concerns regarding the verifiability and reproducibility of the results presented and the surrounding discussion. Ensuring that scientific findings can be independently verified is crucial for the integrity of research. In the revised manuscript, we have included additional results from a blinded CSP study on an agrochemical molecule developed by Bayer complete with all necessary information for reproducibility, and added two Bayer Scientists on the author list. Furthermore, it is a common practice in the field of drug discovery to withhold IP-sensitive information while still providing relevant data to support the scientific narrative. While we understand the need for data transparency, we are bound by constraints surrounding IP protection that limit the extent of information we can share. We believe that our manuscript strikes an appropriate balance by providing essential details without compromising proprietary information.

2) The authors make several false/ unsubstantiated claims of accuracy and novelty in the abstract, the conclusion, and other places throughout the manuscript. All of these claims should be corrected or deleted. Only the most egregious examples are quoted here:

2a) The authors claim multiple time that their method has “unprecedented accuracy and efficiency”. This claim can only be substantiated by direct comparison to other methods in a blind test setting. Therefore, all such claims should be deleted. Avant-garde Materials Simulation has a long-standing track record of success in CSP blind tests and newcomer Xtal-Pi has had comparable results in the 7th CSP blind test. Until Schrodinger put their methods to the test it remains to be seen how their performance compares to their prominent competitors. The CSP community is weary of people making unsubstantiated claims regarding the performance of their methods. For example, Krzysztof Szalewicz claimed in Nature Communications 13, 3095 (2022) to have developed a reliable CSP method, however in the 7th CSP blind test his team had a 0% success rate and failed to generate even a single structure.

We appreciate the reviewer's concern regarding the substantiation of claims about our method's accuracy and efficiency. Please refer to our response to the second comment of reviewer 1. We removed statements that require further evidence to fully support in the revised version of the manuscript.

2b) The authors falsely claim to be the first to use machine learned potentials for CSP. In fact, others have already used ANI potentials for CSP in J. Chem. Theory Comput. 19, 24, 9388 (2023). Moreover, multiple teams used machine learned potentials in the 7th CSP bind test and at least one of the authors who was on one of those teams, should be well aware of that.

We apologize for our potentially confusing statement about the novel use of MLFF in our work. Please see the response to the third comment of reviewer 1 who shares this criticism. We removed this statement that might cause confusion among the readers.

2c) The authors repeatedly describe their structure generation method as novel and imply that it is superior to existing methods. However, their method is currently limited to the 22 most common space groups (a fact that is only revealed in the SI and swept under rug in the main text) and to Z'=1. Considering that some of the structure generation methods developed by others do not have these limitations, the authors cannot claim superiority. They should provide a balanced description of their method and its limitations and avoid unsubstantiated claims.

We regret that the original manuscript may have caused confusion. Our novelty statement is not about the scope of the application domain; rather it is about a completely new way to address structure generation, a challenging problem, with a method that appears to have some advantage. It is a common practice in the CSP community to only sample the most frequently populated space groups. The methods we compared with all sampled the limited space groups as detailed in these papers. [10.1021/jacs.9b06634, 10.1021/acs.cgd.8b01098, 10.1021/acs.cgd.0c00422, 10.1039/c8fd00069g]. In addition, even without the superior performance, we think the community would benefit from the presented broad verification of our new way to sample crystal structures. Furthermore, we also clarified the limitations of our method in the revised manuscript:

"The systematic and hierarchical crystal packing approach (see Method Summary section) provides a significant efficiency improvement as compared to the stochastic approach employed by other CSP methods. However, it requires a lot of effort to implement as the symmetry operations for each space group are different thus the intermediate cluster construction is different for each space group. Our current sampling covers the top 22 most common space groups, and it will take some significant efforts to cover all 230 space groups."

2d) In the conclusion the authors claim to have a "pretrained, transferable MLFF that accurately predicts the relative stability of different polymorphs without requiring extensive DFT calculations". This claim is not supported by the data presented in the manuscript. To the contrary, in the SI it is shown that their pretrained model ranks the experimentally observed

structures of some compounds above 400 and even above 1000 in a couple of cases. The system specific models still rank some targets above 200. No benchmarks are presented of the MLFF geometry, energy, and ranking compared to DFT. In light of this, it seems that the statement quoted above wildly exaggerates the capabilities of the MLFF. The authors should provide an accurate description and acknowledge the limitations of their methods in the main text.

We use the MLFF as an intermediate stage in the filtering of structures from the packing search, not as a final ranking method. As such, when we stated that the method "accurately predicts the relative stability of different polymorphs" we were implying that the MLFF serves well in this capacity. We did not state that our MLFF is of high enough accuracy to replace the role of periodic DFT. We have clarified this in the main text:

"Third, our method leverages the power of a pretrained, transferable MLFF which predicts the relative stability of different polymorphs with sufficient accuracy to require only a practical number of DFT calculations, and the pretrained MLFF can be further customized on a per molecule basis with molecule specific DFT calculations to further improve the reliability and decrease the number of expensive periodic DFT calculations. Considering the well documented high accuracy requirements of successful crystal structure ranking, neither of the current MLFF models used in this work are yet of sufficient accuracy to completely remove the need for periodic DFT refinement."

We intend to report on our refinement procedure in detail in the near future and will show explicit correlation with DFT in that manuscript. For this manuscript we feel it is sufficient to demonstrate the model's capability to filter structures that are ultimately ranked with DFT.

2e) Also in relation to the previous comment, the authors make a song and dance of how the GRACE tailor-made force field ranked the crystal structure of MK-8876 as 2290th. They compare the GRACE force field ranking to their own *DFT* ranking of the structures of MK-2022 and MK-8876 as 1st and 10th. This comparison seems misleading considering that the authors' pretrained MLFF ranks the structure of MK-2022 as ~900 according to Figure S1 and the MLFF ranking of MK-8876 is not even shown. To be fair, the data shown in the SI for the performance of the MLFF indicates that some structures could be similarly missed by the authors' workflow if discarded at an early stage based on the MLFF ranking. This issue with the MLFF should be acknowledged and discussed in the text.

We have stated the results of previous studies on the systems reported in this work. We simply compared our results with a previous study on MK-8876, the same as we did for other molecules when previous CSP studies were available; we disagree with the reviewer that it is a "song and dance" to do so. The MLFF ranking results for MK-8876 were not shown in detail because the transferable MLFF model ranked this system well (rank was 39 with 237 structures in a 6 kcal/mol energy window) and it did not satisfy the selection criterion for demonstration of the fine tuning approach. We also added a figure in SI reporting the rankings for all 65 test systems using the pre-trained MLFF model in the revised manuscript. We agree that with a pretrained model it will

be possible to miss structures and we state this in the supplementary material as “Even within the druglike space, it may be unreasonable to assume one could pretrain an empirical model to have uniform accuracy across even this relatively limited domain due to the well known large size of the chemical space.” This is the motivation for the fine tuning procedure which is discussed.

3) The authors seem to take the DFT ranking of putative polymorphs as the ground truth. They repeatedly claim that their method “not only reproduces all the experimentally known polymorphs, but also suggests new low energy polymorphs yet to be discovered by experiment that might pose potential risks to development of the currently known forms of these compounds”. They completely ignore the so-called over-prediction problem, the tendency of CSP methods to predict multiple low-energy polymorphs that are not in fact observed in reality [e.g., CrystEngComm 23, 5575 (2021); Cryst. Growth Des. 20, 6847 (2020)]. There should be a more nuanced discussion of whether these predictions are to be believed in the introduction and throughout the manuscript, in particular in cases such as MK-8876 and Galunisertib, where multiple putative polymorphs are predicted to be more stable than experimentally observed structures.

We thank the reviewer for the excellent point. We made several changes in the revised manuscript to address the concern. First, we cited the paper the reviewer pointed out and discussed the overprediction problem in the results summary section:

“Upon careful inspection of the top ranked predicted structures, we found that some of them adopt very similar conformers and packing patterns. These structures correspond to different local minima of the quantum chemical potential energy surface at 0K, but may interconvert at room temperature since the corresponding energy barriers might be comparable to thermal fluctuation.[20,21] This could be related to disordered structures[22] and has been proposed to be one of the common reasons for the well known over-prediction problem in CSP calculations.[23] To remove this type of non-trivial duplicate from the static landscapes, we clustered similar structures (with $RMSD_{15}$ better than 1.2 Å) into a single representative structure with lowest energy among the cluster, similar to what has been done in earlier studies.[21,23] The rankings of the best matched structures after clustering are shown in Figure 3(a) (red bars). This improved the rankings, for example, for MK-8876, Target V, and naproxen.”

Second, we performed entropy calculations of all the molecules that are discussed extensively in the revised manuscript, and the entropy calculations helped rule out some of the low energy structures at 0K.

Third, we weakened the statement that “new low energy polymorphs yet to be discovered by experiment” in the revised manuscript. For example, in the discussion of GSK269984B, the revised manuscript now reads as follows.

“Although the experimental structure was predicted to have higher energy than a few other predicted structures at 0K, its relative stability improved substantially with increased temperature. (Figure S18(c) in SI.) At room temperature, the rank I predicted structure and the experimental

structure have a free energy difference of about 0.3 kcal/mol while all other predicted structures have higher free energies. These results highlight the important role of entropy in accurately evaluating the relative stabilities of crystal structures at room temperature."

4) An analysis of whether the correct ordering of experimentally observed polymorphs is obtained is missing from the manuscript. The analysis provided in the SI for ROY and Galunisertib should be moved to the main text. Similar analysis should be performed in all cases where experimental data on relative energies exist. Furthermore, in many cases, even if there is no experimental information on the energy differences between polymorphs, there is information on the order of stability. An analysis should be presented of whether the correct ordering of stability is reproduced (in Figure 3 or as a separate figure). This should be accompanied by a rigorous discussion of the reliability of DFT methods for ranking.

We thank the reviewer for their feedback that we have missed an opportunity to discuss whether the correct ordering of the experimentally observed polymorphs is obtained using the best periodic DFT calculations performed.

In Figure 4 of the revised manuscript, we graphically display the experimental stabilities for each compound where available. We also discussed the comparison of the calculated versus experimental stabilities in the revised manuscript.

We also revised the last paragraph of the discussion section regarding the DFT accuracy:

"Through careful comparison of relative stabilities across different DFT functionals for the set of molecules studied, we found the relative energies between different DFT functionals can vary by ~0.5 kcal/mol or even up to 1.0 kcal/mol in some cases, in agreement with the error estimates from previous studies.[7] Future work on more accurate DFT methods, particularly DFT methods that more accurately describe the conformational energies could further improve the reliability of the ranking among crystals with competitive stabilities. The temperature dependent free energy calculations also improved the relative stabilities as compared to experiment in many cases. We hope the large number of predicted solid form landscapes we have provided, including experimental relative energy rankings where available, can aid in the investigation and further development of more accurate energy ranking methods, including DFT functional selection, conformer energy correction protocols, and temperature dependent free energy evaluations."

5) It is unclear why only ROY, Galunisertib, Rotigotine, MK-8876, MK-2022, and LY-156735 were selected for discussion. It seems based on Figure 3 that there are several other interesting cases such as FuROY, oxalyl dihydrazide, axitinib, bicalutamide, diflunisal, and flufenaic acid to name a few. The authors should expand the detailed discussion of interesting cases. If this makes the manuscript too long for a Communication then it might be more appropriate for it to be in a more specialized journal. In addition, it would be nice to see the energy vs density plots of all targets with multiple polymorphs in the SI.

We are happy that the reviewer is interested in seeing an extended discussion on more case studies. We acknowledge almost all of the molecules in our dataset are worth detailed discussions–they are either drug molecules, or molecules that have been extensively studied by other CSP methods, which speaks to the broad interest and potential impact of our paper. Since this paper is intended for the introduction of a new method, due to the space limitations, we have to draw a line for how many molecules we can discuss in detail in one paper. In the revised manuscript, we included detailed comparison between experimental and calculated stabilities for all tier 3 molecules with more than one Z'=1 experimentally known structure. These include Bicalutamide, XXIII, Axitinib and AZD8329 in addition to what we have discussed in detail in the original manuscript.

6) It seems like all or most of the information provided in the inappropriately named "Discussion" sub-section belongs in the SI. This and the removal of Figure 6 and associate discussion should make room for additional case studies of interest.

Following the reviewer's suggestion, we shortened the discussion section and moved some of the information to the SI in the revised manuscript.

7) Issues with figures:
7a) Half of Figure 1 is a graphic with no scientific content. This should be removed

The reviewer did not point out which half of figure 1 does not have scientific content. But no matter which half the reviewer was referring to, we do not agree with that assessment. As stated in the caption of figure 1, the left half describes our novel packing search method using a divide-and-conquer strategy. The example structure in the P1 space group was sampled from a single conformer, to a cluster of 3 molecules corresponding to 1D crystal, to a cluster of 9 molecules corresponding to 2D crystal, to a cluster of 18 molecules corresponding to 3D crystal. We think this is useful for the readers to have a high level understanding about our unique sampling method. The right half of the figure describes our energy ranking method integrating a multi-stage energy relaxation and filtering of increasing accuracy and cost. This includes molecular dynamics simulations using the classical force field, structure optimization and reranking using a machine learning force field (MLFF) with long range electrostatic and dispersion interactions, and periodic DFT calculations for ranking the final shortlist. We do not think any part of this figure should be removed. In the revised manuscript, we provide a better representation of the molecular clusters in the first half of the figure to further highlight the novel sampling strategy.

7b) In all figures showing relative energies, ranking, and energy vs density plots it should be clarified in the figure caption and in the discussion of the figure which method was used to produce the data shown

We thank the reviewer for their suggestion. We have updated the figure captions and relevant discussion sections to indicate the computational method used.

7c) In Figure 3c,d is there a significance to the colors? Colors could be used e.g., to indicate whether the correct stability ordering was obtained (see item 4 above)

We thank the reviewer for their suggestion. We have changed the color labels in the referenced figure to display the experimental stabilities for each compound, where available.

“The color codes are used to indicate experimental relative stability ordering if available. The experimental relative stability ordering denoted as: blue > red > green > orange > brown > yellow > gold. The black lines indicate experimental relative stability is unknown.”

7d) Only one structure is shown for Diflunisal in Figure 3. Where are the other three?

We thank the reviewer for pointing out this important detail regarding Diflunisal. To our knowledge, there are two solved homomolecular Z'=1 forms of Diflusinal form I and form V and there's one solved homomolecular Z'=2 form in the CSD. There are also many other forms of diflunisal noted in the literature but whose full atomic coordinates remain unsolved.

We limited the discussion and analysis of the paper to only Z’=1 polymorphs. Since the form V is the metastable form having much higher energy, the ranking is much higher than most other experimental structures studied in the manuscript. In the original manuscript, the scale of figure 3 was not adjusted to include the metastable form V of Diflunisal. In the revised manuscript, we adjusted the scales in updated figure 4 to include both Z’=1 polymorphs of Diflunisal.

Reviewer 3

Review of the paper “A Robust Crystal Structure Prediction Method to Support Small Molecule Drug Development with Large Scale Validation and Prospective Studies”

The paper is well written and has a narrative which is easy to read also for non-experts in the field. I have some remarks on the novelty of the findings in this paper and some suggestions on clarifications needed to be added.

In the Introduction the authors write: “However, this process can be time consuming and may miss some important low energy polymorphs due to its inability to exhaust all crystallization conditions.”
As presented, the reader has not yet been introduced why a low energy polymorph is of interest. That should be clarified and exemplified.

We thank the reviewer for the great suggestion. In the revised manuscript, we first explained why later appearing polymorphs pose significant challenges for drug development with examples, and then introduce the limitations of experimental polymorph screening.

It is also stated “Computational polymorph prediction can complement experiments to de-risk unexpected polymorphic changes during drug development”
A few references with clear industrial perspective are missing. For example Firaha et al in Nature (https://doi.org/10.1038/s41586-023-06587-3), Sun et al (https://doi.org/10.1021/acs.cgd.0c01266) , Broo et al (https://doi.org/10.1107/S2052520616006831). The recently published paper in Nature as given above, advances the area of CSP to another scientific level. In that study free energies are used to risk assess polymorph landscape for example, which is missing in this paper.

We thank the reviewer for pointing out these additional references that are relevant to the discussion. We cited all these papers in the revised manuscript.

The present paper shows a wider scope than previously addressed (65 API:s) and also a novel method for crystal packing searching, in combination with ML-FF reranking. Half of the set has only one experimentally known crystal structure while the other half has more than one. Molecules were restricted to Z´=1 only. Nowadays it is standard in the industry to handle CSP for such systems. More interest lies in the more challenging cases of Z´=2 or multicomponent systems such as solvents or cocrystals. Thus, the scope of the present study is not very novel.

We agree with the review that the applications domain covered in our paper may not be entirely novel. However, our sampling method represents a significant advancement. Specifically, our approach has the potential to outperform existing methods by up to 5X for the same application. In addition, we believe there is significant impact in presenting the low-energy crystal structure landscapes across 65 molecules, with energies from periodic DFT calculations. This breadth of

publically available results for molecules with many rotatable bonds is not available in any other single previous publication, representing a valuable resource for the scientific community. In addition, we added a sentence to the end of the discussion to support this: “We hope the large number of predicted solid form landscapes we have provided, including experimental relative energy rankings where available, can aid in the investigation and further development of more accurate energy ranking methods, including DFT functional selection, conformer energy correction protocols, and temperature dependent free energy evaluations.”

In the section "Summary of results"
Non-trivial duplicates are discussed and how they were removed. It would have been interesting to link this to disordered structures, where a small rotation or ring-flip at typical temperatures for API storage may lead to a disordered structure which is favored by configurational entropy, i.e. favored free energy due to increased entropy. It may be that the non-trivial duplicates actually are of interest and importance. See for example Putra et al (https://doi.org/10.1021/acs.cgd.1c01124)

We thank the reviewer for sharing their insights on this matter. We cited the above paper and added the following sentence in the revised manuscript when discussing this matter:

“This could be related to disordered structures[22] and has been proposed to be one of the common reasons for the well known over-prediction problem in CSP calculations.[23]”

It is also stated "For a few molecules, the relative energy is less than 0.5 kcal/mol, which is likely within the error bars of the DFT calculations"
How has the error bars been evaluated? No mentioning of that is found in the paper, so this is very vaguely described. A recent good validation paper is in https://doi.org/10.1038/s41586-023-06587-3)

We thank the reviewer for the reference regarding the DFT error. In the revised manuscript, we have the following paragraph in the discussion section regarding the DFT error and cited above referenced paper.

“Through careful comparison of relative stabilities across different DFT functionals for the set of molecules studied, we found the relative energies between different DFT functionals can vary by ~0.5 kcal/mol or even up to 1.0 kcal/mol in some cases, in agreement with the error estimates from previous studies.[7] Future work on more accurate DFT methods, particularly DFT methods that more accurately describe the conformational energies could further improve the reliability of the ranking among crystals with competitive stabilities. The temperature dependent free energy calculations also improved the relative stabilities as compared to experiment in many cases.”

Some unclarities in the Cocaine example:
r2SCAN-D3 is mentioned to have a different energy gap than PBE-D3, but it is not discussed in the paper why r2SCAN-D3 is the preferred method by any validation. In the supporting

information this is discussed but it would be good to inform the reader already in the main text. Now it just appears without any rationale.

We thank the reviewer for the suggestion. In the revised manuscript, we moved this discussion about why r^2SCAN is more accurate than PBE-D3 in the main text.

The section "Prospective applications on two drug formulation"
These are nice stories to tell, but it seems the authors are unaware this is standard procedure already in several pharmaceutical companies. This makes the study less novel.

We agree with the reviewer that similar prospective applications of CSP might be a common practice in some pharmaceutical companies. The purpose of this section is not to demonstrate a novel use case, but to further substantiate the accuracy of our new method as clearly described in the text:

“To further substantiate the reliability and accuracy of our polymorph prediction method, we recently applied it blindly on an insecticide developed by Bayer Crop Science, and prospectively in several of our active drug development projects. We present here results from one blind study and two prospective projects. ”

In the Discussion section:
Some of the discussion on conformer generation and Tables better sits in the Supporting Information. Now that part of the manuscripts becomes quite heavy to read.

We thank the reviewer for the suggestion, and moved these tables into SI in the revised manuscript.

Section of Methods:
I was surprised to see that the crystal packing part uses rigid conformations. Most up-to-date software uses conformations free to relax in the packing stage. Only a small deviation in dihedral angles may give a significant impact on the packing arrangement. It would be good to discuss any comparisons and why rigid conformers has been chosen. Computer time is one part obviously but some thoughts on the selection would be informative.

We thank the reviewer for the interest in the details of the method. Please refer to our response to the first reviewer’s first comment.

Color labels in Figure 3c,d is not given, so context is unclear.

We thank the reviewer for their suggestion. We have changed the color labels in Figure 3 to display the experimental stabilities for each compound, where available.

The paper also does not discuss the importance of adding temperature effects on top of the 0K CSP landscape. This has in several papers been shown to be of very big importance, already referenced in this review.

Following the reviewer's suggestion, in the revised manuscript, we included the temperature dependence of the relative stabilities through free energy calculations by using established methods. The results are included in the SI and discussed in the main article.

The presented method is a nice step forward for the software developers, and with impressive validation statistics. However, the lack of novelty and the fact that the protocol is missing some key parts such as free energy contributions does not merit a publication in Nature Communications. I would suggest publication in a journal with a more focused readership, such as Crystal Growth&Design or Molecular Pharmaceutics.

We thank the reviewers again for the general comment. As stated in response to the third comment, while we agree the application domains covered in our paper may not be entirely novel, our sampling method represents a significant advancement that the entire community could benefit from.. Specifically, our approach has the potential to outperform existing methods by up to 5X for the same application. In addition, we believe there is significant impact in presenting the low-energy crystal structure landscapes across 65 molecules, with energies from periodic DFT calculations. This breadth of publically available results for molecules with many rotatable bonds is not available in any other single previous publication, representing a valuable resource for the scientific community. In addition, we performed entropy calculations and obtained the temperature dependent relative stabilities for all the structures that are explicitly discussed in the revised paper.

Reviewer 1

I am satisfied with the majority of the corrections the Authors have made to the manuscript with a small exception of referring, again, to other CSP approaches which use molecule-specific force fields in their calculations. In the reply the Authors said it is not only the case of GRACE, but that also DMACRYS and CrystalPredictor use these. This is not exactly right. First, DMACRSY is not a CSP code, it uses force fields and atomic multipoles for geometry optimization of crystals, but it is not capable by itself to conduct CSP calculations. Instead, it is used by other CSP codes (e.g. CrystalOptimizer and mol-CSPy Day's code) to optimize crystal geometries and/or to calculate their energies. Secondly, these approaches also use pretrained force fields (mostly FIT or Williams99) and not molecule-specific ones. I again encourage the Authors to be more precise in their reference to other CSP methods. I'd also like to point out that if the Authors are aiming to introduce a new approach to CSP it is advisable to know what are the basis of the other available codes, and not only that of their direct commercial competitors (GRACE, XTalPi).

We are happy that the reviewer is generally satisfied with the previous round of revision. Regarding the reference to other molecule-specific force field, first of all, we already removed the term of tailor-made force field with the following text in the manuscript to avoid confusion.

"This approach differs from other CSP approaches where customized parameters are trained specifically on one molecule or a set of molecules.[24,55] The resulting parameters from our method are transferable to different chemistries and for different applications, whereas the customized parameters used in other CSP methods only have limited scope of application."

Second, we are aware that DMACRYS is not a CSP code. In our reply, we just wanted to use DMACRYS as an example of a method that generates customized parameters (including the conformer dependent atomic partial charges fitted to QM) for one or a set of input molecules.

Third, while our paper was under review, results of the 7th CSP Blind Test were published. We expanded the discussion of the difference between our method and other methods used in the 7th blind test, not just GRACE and Xtalpi.

"The systematic and hierarchical crystal packing approach (see Method Summary section) provides a significant efficiency improvement as compared to the stochastic approach employed by other CSP methods. Most groups participating in the 7th CSP Blind Test used methods that parameterize the search space via the six unit cell lattice parameters, the six parameters for ASU location and orientation, and the conformational degrees of freedom, and search these degrees of freedom simultaneously.[11,12] These methods require substantial computational effort to adequately cover the search space due to the size and complexity of the parameter space.

Alternative methods included making perturbations to existing crystal structures and choosing to accept or reject the new candidate using a Monte Carlo scheme as well as the generative adversarial network of Group 10. To manage computational cost, many participants chose to terminate their program's execution after a certain number of structures or generation attempts were made. Some participants included convergence criteria by monitoring repeated generation of the lowest energy structures or by monitoring the number of generation attempts without finding new low-energy candidates. In contrast, the method presented here efficiently covers the search space for molecular crystals using a divide-and-conquer strategy that partitions the parameter space into subspaces based on space group symmetries, eliminating the need for the convergence checks required by stochastic approaches.

We have presented the use of QRNN, a machine learned force field (MLFF), to perform structural relaxation and energy ranking of molecular crystals in our CSP workflow prior to DFT. Innovations in the successful use of machine learning methods were also demonstrated by participants in the 7th CSP Blind Test.[11,12] MLFFs were developed and utilized in the CSP protocols of groups 12, 15, 16, and 23. In the second part of the seventh blind test, Group 16 demonstrated energy ranking results with system specific MLFFs comparable to accurate DFT methods. Additionally, groups 10 and 20 made use of machine learning to reduce the number of structures required to be evaluated by costly methods, such as DFT. Future developments of QRNN in our CSP workflow will include improving molecule specific fine-tuning protocols to further reduce the cost of DFT evaluations."

Reviewer 2

In the revised version of the manuscript the authors have addressed most of my comments adequately.

I do have one major criticism remaining. I must insist that the authors remove the results of STC-1 and STC-2 from Figure 7 and all the related discussion from the text. I understand that some IP protections apply in this case, however RESULTS THAT CANNOT BE VERIFIED AND REPRODUCED BY OTHERS SIMPLY DO NOT BELONG IN THE SCIENTIFIC LITERATURE. With the addition of the Flupyradifurone molecule removing the other two would not take away from the impact of this manuscript. The authors may publish their work on STC-1 and STC-2 in the future when the IP protections are no longer needed and they can reveal the molecules and their crystal structures.

We thank the reviewer for the careful review and are pleased that the reviewer is generally satisfied with the previous round of revision. Per the reviewer's suggestion, in the revised manuscript, we removed the results of STC-1 and STC-2 from Figure 7 and all related discussion from the text.

An optional change is to move Table 1 and the detailed discussion of computer time usage to the SI and only summarize it briefly in the main text

We thank the reviewer for the suggestion. Since Table 1 and the discussion of compute time do not take too much space and this is an optional change, we think it is useful to keep this information in the main text for the reader to have some idea about the scope of validation of the method.

Additional minor comments:

In the discussion of the usage of machine learned force fields for molecular crystal structure prediction it would be appropriate to cite the 7th CSP blind test and discuss the usage of MLFFs by participants in both stages of the blind test

Because one of the authors' selling points is their new approach to crystal structure generation, it would be appropriate to explain the difference between their method and the methods used by others with references to other structure generation methods

In relation to the performance of DFT and importance of free energy it would be appropriate to cite: J. Chem. Theory Comput. 2022, 18, 4456−4471

In relation to ROY it would be appropriate to cite: CrystEngComm 2019, 21, 2080–2088

We thank the reviewer for the suggestion. In the revised manuscript, we included our CSP results on compound XXXI from the seventh blind test, and compared our results with that from participants of the blind test. We also expanded the discussion of the difference between our method and others used in the 7th blind test (see reply to Reviewer 1's comments for details). The above references are also cited as reference 59 and 42 respectively.

Reviewer 3

The authors have carefully replied to most of the comments outlined for the initial version of the manuscript. I am pleased that several aspects of the concerns have been covered, such as introduction of a "blind test" molecule for transparency. In addition, free energies have been computed for the final rankings of polymorphs.

Although the manuscript has significantly improved, my major concern is still valid. This is a big study with some good benchmarking results. The major novelty claimed by the authors lies in

the sampling method. The speedup examples comparing to other CSP protocols does not take into account that those studies were performed 5-6 years back as well so a direct comparison in cpu hours is questionable. The authors exemplify their methods on "...prospective applications on two drug formulations". Today this is routine daily work in the pharmaceutical industry, and has been covered in many publications. The authors themselves reply to one of the comments as: "We agree with the review that the applications domain covered in our paper may not be entirely novel. However, our sampling method represents a significant advancement." . But this does not merit a publication in Nature Communications. I recommend to publish in a more specialized journal.

We thank the reviewer for the feedback. In the revised manuscript, we included our CSP results on compound XXXI from the seventh blind test, and compared our results with that from participants of the blind test (see reply to Reviewer 1's comments for details). This study represents the state of the art of the field, and we think this comparison can further substantiate the conclusions of the paper. It should be noted that direct comparisons of CPU costs for CSP methods is difficult due to sparsity of published results, particularly that most CSP papers consist of only a small number of examples. We hope the large scale benchmark with detailed compute cost information reported in this paper will help facilitate the comparison of the performance among different CSP methods.

Regarding the two prospective studies, we removed the results of STC-1 and STC-2 from Figure 7 and all related discussion in the revised manuscript.

"While this paper was under review, reports of the seventh blind test results of crystal structure prediction organized by the CCDC were published.[11,12] Compound XXXI is the only molecule in this blind test that is relevant for agrochemical and pharmaceutical development and has experimental forms with only Z'=1 structures. Three experimental structures of XXXI were determined. However, Form C is a channel-type solvate containing unresolved solvent and therefore falls outside the scope of prediction. Form A contains two disorder components resulting from the flipping of the fluorinated ring, and form B was determined by competitive slurry experiments as the most stable form below 55°C with form A becoming more stable for temperatures above 55°C.

The predicted energy landscape of XXXI at 0K from our calculations and the temperature dependent relative stability between forms A and B are reported in Figure 7 (a), (b) and (c). Our calculations correctly predicted form B and the two disordered forms of form A among the top 25 predicted structures. The temperature dependent relative stability calculations indicate that with increased temperature form A, with its two disordered structures, become more stable as compared to form B, consistent with the experimental findings that form A is more stable at temperatures higher than 55°C. However, the relative free energy difference between the two forms at room temperature is small (about 0.4 kcal/mol), within the errors of the calculations, and therefore it can not be predicted with confidence which form is more stable at room temperature. It is also interesting to note that the relative lattice energy difference (stability at 0K reported in Figure 7 (a)) among these three structures from our calculations are consistent with

the results from all groups participating in seventh blind test that used similar DFT calculations to rank order the structures (including groups 3, 20), except for group 10 who reported a different order than others.[12]"

"For XXXI, our computational cost is 93K CPU hours and 200 GPU hours, whereas most participants in the seventh blind test who successfully predicted the experimental structures used significantly more CPU hours. It should be noted that our calculations were limited to Z'=1 structure predictions, whereas these participants might have spent a lot of effort for higher Z' structure predictions, so these CPU hours might not be directly comparable."