nature portfolio

Peer Review File

Rapid prediction of molecular crystal structures using simple topological and physical descriptors

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

This is a very exciting paper, which could revolutionize our understanding of the principles underlying organic crystal structures and the ability to predict them. The results are comparably accurate (in terms of 20 molecule cluster overlays, RMSD20) to current crystal structure prediction (CSP) methods of lattice energy minimization, which are probably not much larger than the variation of crystal structures with temperature. The choice of the first two molecules in the last blind test is good, as such a choice is not biased by the results. The other two systems are aspirin and ROY (in the SI). However, the abstract should mention that van der Waals volumes and CSD derived ranges of atom-atom contact distances (c.f. van der Waals radii) were used. This approach may be a very clever mathematical method of applying Kitaigorodski's principle1 that structures are close packed. If this code could be used to by-pass the structure generation step in CSP, and provide a reasonable sized set of plausibly close packed crystal structures for evaluation by the best methods of predicting relative thermodynamics of organic crystals, this would be a game changer. This paper is worthy of publication in Nature, once it has been revised to be more comprehensible to the crystallographic and CSP community.

I hope that the authors intend to make the Crystal Math code, including the underlying CSD derived distributions in SI) available to the community. Obviously, they will be doing further testing and development first. I am unable to say the results/code are reproducible on the basis of the information given.

My concern about presentation is for the maths to be more intelligible to the CSP community:

P3 Does it take 4 parameters to determine the orientation of the molecular fragment, and so the total number of parameters to determine given is for each space group?

P3 the definition of nmax, which appears to be typically 5, is both the maximum value for any component, and ???(I don't know the notation for the first other condition) I think the second condition implies that one component is zero. (A statement in the SI about the relationship to (h,k,l) would be welcome.) Fig A1 is very convincing, though restricted to CHO compounds.

P4 the difference between triclinic and orthorhombic and monoclinic cells is unclear – does "each pair of fragments in the reference molecule" also apply to rigid molecules? Fig A2 is also convincing. It seems that the third principle is just to define the molecular shape of the rigid fragment, using basis functions (Zernike order parameters ZOPs) that could readily be mistaken for the hydrogenic orbitals, unless you look carefully at equation 12. This is clearly a function that is often zero. I found the following paragraphs hard to follow, even after looking at the SI, with the key term ZZP being defined in terms of zeros of ZOPs. However, I wondered if the whole method was relying on the rigid molecular fragments in organic chemistry having approximately fairly high symmetry and so the shape is readily expanded in terms of spherical harmonics. The use of the atoms with the highest electropositivity and electronegativity is also perhaps relying on these forming hydrogen bonds or being highly repulsive. It would be helpful to have diagrams showing how fragment in the 3 molecules in the m/s were represented in terms of the basis functions, and showing some crystal structures defining the ZZPs to demonstrate the basis of Crystal Math approach.

The chemistry of intermolecular forces is implicit in the two dominant terms in the objective function, both from the use of atoms with highest electropositivity/negativity to the elimination of unphysical contacts and the somewhat undefined final step of "close intermolecular contacts and the van der Waals free volume are minimized". The SI tables S.1-S.6 show the 95% confidence intervals for all close contacts in the most common space groups, for structures composed of C, H, O atoms and for structures composed of C, H, N and O atoms, but it is not clear whether the optimization is more than just elimination structures with implausibly short intermolecular contacts. The treatment of flexible molecules is potentially a game changer in CSP as current CSP generation methods scale very badly with number of flexible torsion angles trying to cover the possible conformation space. In CrystalMath flexible molecules are divided into unphysical fragments, with the fragments having a N or C atom in common. Round 3 of the procedure, where these junction atoms are sufficiently close (how close does this distance have to be to zero or do the distributions of this distance have quite sharp, well separated peaks?) eliminates most of the structures, giving the impressive reduction to almost only the observed polymorphs. (It is rather larger in the case of ROY (in SI) , where 10 polymorphs are denoted as known, but then, how many more polymorphs of ROY remain undiscovered?2) CSP methods for rigid molecules, such the close-packing based code MOLPAK3 are very quick for rigid molecules, but needed a good electrostatic model to get relative energy rankings accurate.

The discussion does appear to show that the known organic crystal structures largely adhere to the principles used in Crystal Math, implying that this should be an efficient way of generating plausible candidates for crystal structures. There will be exceptions, such as desolvated solvates. The extent to which the rules of CrystalMath are a new framework does need to consider whether it is mainly applying the close-packing principle in a way that appears very effective for the most common spacegroups for organic molecules. The MOLPAK analysis of common packing types may well conform to these principles.

(1) Kitaigorodski, A. I. Molecular Crystals and Molecules; Academic Press, 1973.
(2) Beran, G. J. O.; Sugden, I. J.; Greenwell, C.; Bowskill, D. H.; Pantelides, C. C.; Adjiman, C. S. How many more polymorphs of ROY remain undiscovered. Chemical Science 2022, 13 (5), 1288-1297, Article. DOI: 10.1039/d1sc06074k.
(3) Holden, J. R.; Du, Z. Y.; Ammon, H. L. Prediction of Possible Crystal-Structures For C-, H-, N-, O- and F-Containing Organic Compounds. Journal of Computational Chemistry 1993, 14 (4), 422-437.

Reviewer #2 (Remarks to the Author):

The manuscript presents a geometric approach to predicting molecular crystal structures, as a fast alternative to methods that rely on energy calculations and optimisations. The methods is described in detail and results are shown for 3 molecules (+ another in the SI, which is not mentioned in the text).

The results seem impressive considering the computing times often applied to crystal structure prediction, as many methods in this field have moved towards using solid state density functional theory calculations for the final optmisation and energy ranking of crystal structures.

Results are first presented for aspirin, which has three known polymorphs. Crystal structre prediction studies based on energy calculations have prevously been performed for aspirin, predicting the structure of form II before it was identified experimentally. The results presented here show forms I and II as the only surviving predicted crystal structures after their procedure. I have a few comments:

- in the rigid-molecule search, how was the geometry of the molecule determined? Previous work showed that forms I and II do not adopt the lowest energy conformer (https://doi.org/10.1021/cg049922u), which seems to be consistent across different levels of theory for the conformational energy calculation. It would have been a fairer comparison to traditional CSP methods if the procedure had been run using at least the two lowest energy conformers of the molecule. The text should, at least, describe how the rigid molecular geometry was obtained and why the lower energy conformer was ruled out.

- Considering that Z'=2 predictions were performed for a more complex molecule later in the paper, it would have been nice to see Z'=2 calculations performed for aspirin to assess prediction of the third polymorph (https://doi.org/10.1021/acs.cgd.7b00673). Given the emphasis in the manuscript on computational efficiency (that calculations can be perfomred in under a day on a desktop), this is an odd omission.
- I would like to see more discussion of the step 3 filtering of structures. This seems to be very important. As stated in the caption to figure 3: "In both searches, we found at rounds 1 and 2 structures that have lower vdWFV and/or lower combined cost function compared to the predicted structures. However, these structures are discarded because they have unphysical between neighboring molecules and/or unphysical intremolecular distances between the atoms in the unit cell." Could the authors make the low cost function structures that were filtered out after step 2 available to readers, along with more discussion of the filtering step, which seems crucial to the success of the method?
- In the flexible-molecule search for aspirin, figure 3 shows forms I and II showing up in the same place in vdW free volume/cost function. However, some structues that were found in the rigid-molecule search are absent. Cold the authors comment on why this search does not located some of the round 1 and round 2 structures that were found in the rigid-molecule search?

Molecule XXII (previous blind test molecule). The reslts end up with 8 surviving crystal structures. The structure with the lowest free volume matches the experimentally known crystal structure. Do the authors suggest free volume as the final descriminator between predicted crystal structures? If the purpose of the method is to avoid energy calculations, then we need some method to rank structures: if we didn't know the true result in this case, what would we have done with these 8 structures? Much of the difficulty in crystal structure prediction is in distinguising structures at the final stage, where energies are very close.

Similar comments could be made for the ROY results which are provided in the SI, but not discussed in the manscript: the Z'=1 polymorphs are all generated, but there are also quite a few other structures. Low vdW free volume is seen in the observed polymorphs, but there are some observed polymorphs that have higher free volume than quite a few unobserved predicted structures. There should be more discussion of wat one would do with such results if CSP was being applied to this molecule in a predictive sense (before the polymorphs are known).

Without a final cost function to rank the predicted structures, this is an impressive structure generation method, but not a complete structure prediction method, which requires a final method of ranking the generated structures.

Some minor comments:

"P − 1" should have the bar over the 1 in a couple places in the manuscript.

Table 1 and text ... "high electropositivity or electronegativity" are discussed quite a bit. Define these. What is "high"?

"Recent studies have revealed that in more than 50% of structures in the CCDC, energy differences between pairs of polymorphs are smaller than ~2 kJ/mol, while only about 5% have energy differences larger than ~7 kJ/mol [17]." I don't believe that reference 17, which is a review chapter, is the original source of this information. The authors should cite the original study, which is coincidentally reference 17 in the chapter.

Reviewer #3 (Remarks to the Author):

-Crystal structure prediction (CSP), the problem of finding the most stable arrangement of atoms given only the chemical composition, has long remained a major unsolved scientific problem. Six Blind Tests took place in CCDC until 2016, to explore the ability of different methods to find the corresponding crystal structure.
-This manuscript presents a mathematical tool built taking into account crystal packing from a large CCDC data for crystal organic compounds including C, N, O.
- Nowadays there are many approaches used to search for possible crystal packing arrangements of unknown crystals using global optimization algorithms. The results presented in this communication are valuable considering the low computational resources, low cpu time and one desk computer, involved to produce acceptable results, in comparison with those derived from the first-principles calculations.
-In this manuscript the authors designed a topological approach to molecular crystal structure prediction, named CrystalMath. They made an examination of a database of nearly 240,000 organic crystal structures with one and two molecules in the asymmetric units, composed by C, O and H atoms in the Cambridge Structural Database three general principles which give, in their criteria, "a new framework for understanding how molecules pack into three-dimensional crystal structures". Those topological principles and the resultant derived mathematical equations are presented in detail for monoclinic, orthorhombic and triclinic cells. The method does not include any energy, free energy nor other thermodynamic evaluation for the resulting structures. In summary, it consists in the generation of a large pool of random structures for a target compound, and the "optimization" of the structures, consists in the application of the three CrystalMath principles and the corresponding topological filters.

They chose 3 molecules to show the performance of CrystalMath, rigid and flexible aspirin; targets XXII and XXIII from the 6th CCDC blind test.
They found polymorphs with RMSD_20 in the order of 0.080 angstroms and 0.234 angstroms from their respective experimental structures for rigid aspirin, 0.103 angstroms and 0.177 angstroms for flexible aspirin, and of 0.251 angstroms angstroms for compound XXII and similar values for compound XXIII from the CCDC 6th BLIND TEST.
The larger computation time corresponding to compound XXIII is 26 cpu hours to complete on a mid-range laptop (Intel(R) Core(TM) i7-10750H CPU @ 2.60GHz, 16 GB RAM. Surprising.

-My question to the authors is:
Since recent studies have revealed that in more than 50% of structures in the CCDC, energy differences between pairs of polymorphs are smaller than~2 kJ/mol, while only about 5% have energy differences larger than ~7 kJ/-mol, as they underline in the text, and they do not report any energy calculation, it would be mandatory:
1- Report, please, the energy differences between aspirin polymorphs, between the 3 to 10 better structures respect the experimental ones. The comparison is valuable to have an insight into the degree of precision of the CrysMath
2- Report, please, also the same comparison for Targets XXII and XXIII from the 6th CCDC blind test, to infer something else about degree precision of CrysMath.
I consider that those degrees of precision must be checked. There are many CODES to make the

evaluation with a DFT-D method. I suppose they must have access to one of them. Quantum Espresso code, by the way, is free access.
To my criteria, the method is useful to get a list of several possible crystal structures of a compound and might be considered a useful tool for proper subsequent first principles refinement calculations, considering also the importance of temperature pressure conditions present in crystallization.
It might be useful also when no single crystals are available and the experimental structure must be determined by other techniques like X-ray powder diffraction or solid-state NMR, to have a reliable list of even several hundred possible candidate structures can be extremely valuable.
It would be important that, in a near future, the authors check performance and degree of precision of the method in different kinds of crystals, co crystals, chiral compounds.

Responses to Reviewers' Comments

General comments

In the time between the original submission of the manuscript and the current revision, new developments improved the performance of CrystalMath. As a result, we felt it was necessary to revise the results section to be up-to-date with the latest developments. The main differences are related to the optimization of the structures for close contacts and the filtering for unnatural close contacts. The new developments are discussed in detail in section "The CrystalMath Protocol".

Reviewer #1

This is a very exciting paper, which could revolutionize our understanding of the principles underlying organic crystal structures and the ability to predict them. The results are comparably accurate (in terms of 20 molecule cluster overlays, RMSD20) to current crystal structure prediction (CSP) methods of lattice energy minimization, which are probably not much larger than the variation of crystal structures with temperature. The choice of the first two molecules in the last blind test is good, as such a choice is not biased by the results. The other two systems are aspirin and ROY (in the SI). However, the abstract should mention that van der Waals volumes and CSD derived ranges of atom-atom contact distances (c.f. van der Waals radii) were used. This approach may be a very clever mathematical method of applying Kitaigorodski's principle1 that structures are close packed. If this code could be used to by-pass the structure generation step in CSP, and provide a reasonable sized set of plausibly close packed crystal structures for evaluation by the best methods of predicting relative thermodynamics of organic crystals, this would be a game changer. This paper is worthy of publication in Nature, once it has been revised to be more comprehensible to the crystallographic and CSP community.

I hope that the authors intend to make the Crystal Math code, including the underlying CSD derived distributions in SI) available to the community. Obviously, they will be doing further testing and development first. I am unable to say the results/code are reproducible on the basis of the information given.

My concern about presentation is for the maths to be more intelligible to the CSP community:

- P3: Does it take 4 parameters to determine the orientation of the molecular fragment, and so the total number of parameters to determine given is for each space group?

  The orientation of a fragment is determined by the components of the unit vector $\hat{\mathbf{k}} = (k_x, k_y, k_z)$, subject to the constraint $|\mathbf{k}| = 1$ and the orientation angle ω, so it takes 3 parameters to determine the orientation. The orientation for the fragments in the reference molecule is the same for all space groups. The space group is determined by the crystal system (triclinic, monoclinic, orthorhombic) and symmetry operations.

- P3: The definition of $n_{\max}$, which appears to be typically 5, is both the maximum value for any component, and ???(I don't know the notation for the first other condition) I think the second condition implies that one component is zero. (A statement in the SI about the relationship to (h,k,l) would be welcome.) Fig A1 is very convincing, though restricted to CHO compounds.

The first condition means that at least one component is equal to $n_{\max}$ and the second that at least one of the components is zero. In the revised manuscript, we have replaced the mathematical notation with an explicit explanation of the required conditions on the vector components of $\mathbf{n}_c$.

- P4: The difference between triclinic and orthorhombic and monoclinic cells is unclear – does “each pair of fragments in the reference molecule” also apply to rigid molecules? Fig A2 is also convincing. (more detail here).

  The statement “each pair of fragments in the reference molecule” can be applied to rigid molecules. The updated section for the search of the target XXII structures demonstrates that when applying this principle to non-flat rigid molecules, the approach can generate not only the correct structure but also the correct geometry for a puckered molecule such as the target XXII compound.

- It seems that the third principle is just to define the molecular shape of the rigid fragment, using basis functions (Zernike order parameters ZOPs) that could readily be mistaken for the hydrogenic orbitals, unless you look carefully at equation 12. This is clearly a function that is often zero. I found the following paragraphs hard to follow, even after looking at the SI, with the key term ZZP being defined in terms of zeros of ZOPs. However, I wondered if the whole method was relying on the rigid molecular fragments in organic chemistry having approximately fairly high symmetry and so the shape is readily expanded in terms of spherical harmonics. The use of the atoms with the highest electropositivity and electronegativity is also perhaps relying on these forming hydrogen bonds or being highly repulsive. It would be helpful to have diagrams showing how fragment in the 3 molecules in the m/s were represented in terms of the basis functions, and showing some crystal structures defining the ZZPs to demonstrate the basis of Crystal Math approach.

  We want to be clear about the purpose of the third principle because it is an extremely important one in CrystalMath. The purpose of the ZZPs is not to define the shape of rigid molecules or rigid molecule fragments. The ZZPs determine the optimal molecular positions within the unit cell for a given space group. Specifically, for a given atom in the reference molecule of a molecular crystal, the ZZPs are calculated following these steps: (1) apply the $Z-1$ symmetry operations to generate the symmetric atoms in the unit cell; (2) define the center of mass (centroid) for the arrangement of these Z atoms; (3) calculate the positions of the Z atoms relative to the centroid; (4) calculate the ZOPs. We use the zeros of the ZOPs (which we call the ZZPs), which are actually planes within the unit cell to determine the optimal positions for the fragments/molecules the molecules in a unit cell by demanding the highly electropositive/electronegative atoms to be as close as possible to the ZZPs. We have added a simple diagram in the SI (Fig. S.1) illustrating how certain positions of atoms in a unit cell adhere to the ZZPs. We hope this clarifies the role of the ZZPs in the CrystalMath workflow.

- The chemistry of intermolecular forces is implicit in the two dominant terms in the objective function, both from the use of atoms with highest electropositivity/negativity to the elimination of unphysical contacts and the somewhat undefined final step of "close intermolecular contacts and the van der Waals free volume are minimized". The SI tables S.1-S.6 show the 95% confidence intervals for all close contacts in the most common space groups, for structures composed of C, H, O atoms and for structures composed of C, H, N and O atoms, but it is not clear whether the optimization is more than just elimination structures with implausibly short intermolecular contacts.

  The reviewer is correct that some simple physics is implicit in parts of the CrystalMath protocol. We now note this in the first sentence of the Discussion section. To address the reviewer's specific question, the optimization of the close contacts involves a translation of the reference molecule in the unit cell to minimize the updated cost function in eq. (16) so that the lengths of the contacts adhere to the distributions observed in the database. The filtering or elimination of structures having unphysical close contacts is a subsequent step for which the optimization failed to generate an appropriate contact distribution. The "CrystalMath Protocol" section in the paper was revised to accommodate latest developments and describe the process with additional detail.

- The treatment of flexible molecules is potentially a game changer in CSP as current CSP generation methods scale very badly with number of flexible torsion angles trying to cover the possible conformation space. In CrystalMath flexible molecules are divided into unphysical fragments, with the fragments having a N or C atom in common. Round 3 of the procedure, where these junction atoms are sufficiently close (how close does this distance have to be to zero or do the distributions of this distance have quite sharp, well separated peaks?) eliminates most of the structures, giving the impressive reduction to almost only the observed polymorphs. (It is rather larger in the case of ROY (in SI) , where 10 polymorphs are denoted as known, but then, how many more polymorphs of ROY remain undiscovered? 2) CSP methods for rigid molecules, such the close-packing based code MOLPAK3 are very quick for rigid molecules, but needed a good electrostatic model to get relative energy rankings accurate.

  In CrystalMath, flexible molecules are divided into fragments that are joined to their common atom(s) of any species. This process takes place in the very first step of structure generation where conformers are generated and placed in the unit cell. During the process the first fragment is placed in the unit cell and additional fragments are translated so that the common atoms between the consecutive fragments coincide. When the fragments have one common atom, this process generates the conformations directly with no additional steps required (apart from the check for unnatural intramolecular contacts). For fragments with two or more common atoms (as in the case of the target XXII compound), the second fragment is translated to join the fragments in at of the common atoms. The position of the second atom is taken as the average of the positions in the two fragments and a check for the bond lengths is applied to discard unnatural conformations.

- The discussion does appear to show that the known organic crystal structures largely adhere to the principles used in Crystal Math, implying that this should be an efficient way of generating plausible candidates for crystal structures. There will be exceptions, such as desolvated solvates. The extent to which the rules of CrystalMath are a new framework does need to consider whether it is mainly applying the close-packing principle in a way that appears very effective for the most common spacegroups for organic molecules. The MOLPAK analysis of common packing types may well conform to these principles.

Future work will involve continued testing CrystalMath on challenging cases and extending is capabilities to more complex cases such as desolvated solvates, cocrystals, and so forth. Only a detailed comparison to MOLPAK can answer the reviewer's comment. However, we believe CrystalMath is doing more than simply applying close-packing principles, as it can correctly predict molecular conformations, orientations, and locations in the unit cell according to very specific principles that have more to say about crystal structures than can packing principles alone.

Reviewer #2

The manuscript presents a geometric approach to predicting molecular crystal structures, as a fast alternative to methods that rely on energy calculations and optimizations. The methods is described in detail and results are shown for 3 molecules (+ another in the SI, which is not mentioned in the text).

The results seem impressive considering the computing times often applied to crystal structure prediction, as many methods in this field have moved towards using solid state density functional theory calculations for the final optimization and energy ranking of crystal structures. Results are first presented for aspirin, which has three known polymorphs. Crystal structure prediction studies based on energy calculations have previously been performed for aspirin, predicting the structure of form II before it was identified experimentally. The results presented here show forms I and II as the only surviving predicted crystal structures after their procedure. I have a few comments:

- In the rigid-molecule search, how was the geometry of the molecule determined? Previous work showed that forms I and II do not adopt the lowest energy conformer (https://doi.org/10.1021/cg049922u), which seems to be consistent across different levels of theory for the conformational energy calculation. It would have been a fairer comparison to traditional CSP methods if the procedure had been run using at least the two lowest energy conformers of the molecule. The text should, at least, describe how the rigid molecular geometry was obtained and why the lower energy conformer was ruled out.

  The conformation was taken directly from the known experimental structures of aspirin. Since aspirin is a flexible molecule, a rigorous CSP search that tests the CrystalMath protocol should follow the flexible search protocol, which is our second search presented. This first rigid search is for demonstration purposes only. This is now clarified in the respective section.

- Considering that $Z' = 2$ predictions were performed for a more complex molecule later in the paper, it would have been nice to see $Z' = 2$ calculations performed for aspirin to assess prediction of the third polymorph (https://doi.org/10.1021/acs.cgd.7b00673). Given the emphasis in the manuscript on computational efficiency (that calculations can be performed in under a day on a desktop), this is an odd omission.

We agree with the reviewer and have now included the results for the $Z' = 2$ structures in both our rigid and fully flexible searches.

- I would like to see more discussion of the step 3 filtering of structures. This seems to be very important. As stated in the caption to figure 3: "In both searches, we found at rounds 1 and 2 structures that have lower vdWFV and/or lower combined cost function compared to the predicted structures. However, these structures are discarded because they have unphysical between neighboring molecules and/or unphysical intermolecular distances between the atoms in the unit cell." Could the authors make the low cost function structures that were filtered out after step 2 available to readers, along with more discussion of the filtering step, which seems crucial to the success of the method?

  The "CrystalMath Protocol" section was revised to include latest developments and provide additional details on the filtering process. A figure was added in the SI section 4, to illustrate the close contacts filtering and the acceptance/rejection of structures on that basis. In addition, per the reviewer's request, we have made the low vdW free-volume structures discarded throughout the CrystalMath protocol available in the supporting information.

- In the flexible-molecule search for aspirin, figure 3 shows forms I and II showing up in the same place in vdW free volume/cost function. However, some structures that were found in the rigid-molecule search are absent. Cold the authors comment on why this search does not located some of the round 1 and round 2 structures that were found in the rigid-molecule search?

  A comment is added in the updated figure 3, that explains the differences between the two landscapes. Briefly, the differences seen in these two landscapes starts with the fact that for a given inertial eigenvector set, the orientation of the aspirin molecule in the rigid search is different from the orientation in the fragment based approach even if the conformations are similar. A rigid search conformation will generally not match conformations obtained in a flexible search and, as a result, the structures with a rigid conformation exhibit a different adherence to the ZZPs and different contact distributions.

- Molecule XXII (previous blind test molecule). The results end up with 8 surviving crystal structures. The structure with the lowest free volume matches the experimentally known crystal structure. Do the authors suggest free volume as the final discriminator between predicted crystal structures? If the purpose of the method is to avoid energy calculations, then we need some method to rank structures: if we didn't know the true result in this case, what would we have done with these 8 structures? Much of the difficulty in crystal structure prediction is in distinguising structures at the final stage, where energies are very close.

  For the target XXII molecule, the updated protocol improved the accuracy of the results by allowing only one structure to pass all the filters towards the final pool. However, as discussed in the "Conclusion and next steps" section, we understand the importance of having an accurate ranking scheme, and we are working to improve our ranking procedure.

- Similar comments could be made for the ROY results which are provided in the SI, but not discussed in the manscript: the Z'=1 polymorphs are all generated, but there are also quite a few other structures. Low vdW free volume is seen in the observed polymorphs, but there are some observed polymorphs that have higher free volume than quite a few unobserved predicted structures. There should be more discussion of what one would do with such results if CSP was being applied to this molecule in a predictive sense (before the polymorphs are known).

  We appreciate the reviewer's comments on our presentation of ROY in the SI. Because of the relatively high number of ROY polymorphs that are already known and the possibility of more exotic structures that may yet to be discovered, we have decided that a presentation of ROY in the SI does not do the problem justice. Rather, it presents a rather unique opportunity for CrystalMath, which we would prefer to reserve for a separate study to be published on its own. For this reason, we have decided to remove ROY from the SI altogether and reserve for future work and a future publication a deeper dive into ROY using the CrystalMath protocol. We will use the opportunity of a deeper study of ROY to address more fully the question raised by the reviewer concerning the fate of structures of low vdW free volume.

- Without a final cost function to rank the predicted structures, this is an impressive structure generation method, but not a complete structure prediction method, which requires a final method of ranking the generated structures.

- Some minor comments:

  - "$P-1$" should have the bar over the 1 in a couple places in the manuscript.

    Corrected as requested.

  - Table 1 and text ... "high electropositivity or electronegativity" are discussed quite a bit. Define these. What is "high"?

    The relevant section has been revised to provide more detail on the electronegativity of the atoms. In the updated protocol, all non-hydrogen atoms are used to determine the optimal positions fro the molecules. Highly electronegative atoms (O, N, Cl, etc) are given priority through the electronegativity parameter χ_λ.

  - "Recent studies have revealed that in more than 50% of structures in the CCDC, energy differences between pairs of polymorphs are smaller than ∼2 kJ/mol, while only about 5% have energy differences larger than ∼7 kJ/mol [17]." I don't believe that reference 17, which is a review chapter, is the original source of this information. The authors should cite the original study, which is coincidentally reference 17 in the chapter.

    We thank the reviewer for pointing this out. The citation was corrected to represent the original source.

Reviewer #3

Crystal structure prediction (CSP), the problem of finding the most stable arrangement of atoms given only the chemical composition, has long remained a major unsolved scientific problem. Six Blind Tests took place in CCDC until 2016, to explore the ability of different methods to find the corresponding crystal structure.

This manuscript presents a mathematical tool built taking into account crystal packing from a large CCDC data for crystal organic compounds including C, N, O.

Nowadays there are many approaches used to search for possible crystal packing arrangements of unknown crystals using global optimization algorithms. The results presented in this communication are valuable considering the low computational resources, low cpu time and one desk computer, involved to produce acceptable results, in comparison with those derived from the first-principles calculations.

In this manuscript the authors designed a topological approach to molecular crystal structure prediction, named CrystalMath. They made an examination of a database of nearly 240,000 organic crystal structures with one and two molecules in the asymmetric units, composed by C, O and H atoms in the Cambridge Structural Database three general principles which give, in their criteria, "a new framework for understanding how molecules pack into three-dimensional crystal structures".

Those topological principles and the resultant derived mathematical equations are presented in detail for monoclinic, orthorhombic and triclinic cells. The method does not include any energy, free energy nor other thermodynamic evaluation for the resulting structures. In summary, it consists in the generation of a large pool of random structures for a target compound, and the "optimization" of the structures, consists in the application of the three CrystalMath principles and the corresponding topological filters.

They chose 3 molecules to show the performance of CrystalMath, rigid and flexible aspirin; targets XXII and XXIII from the 6th CCDC blind test.

They found polymorphs with RMSD_20 in the order of 0.080 angstroms and 0.234 angstroms from their respective experimental structures for rigid aspirin, 0.103 angstroms and 0.177 angstroms for flexible aspirin, and of 0.251 angstroms angstroms for compound XXII and similar values for compound XXIII from the CCDC 6th BLIND TEST.

The larger computation time corresponding to compound XXIII is 26 cpu hours to complete on a mid-range laptop (Intel(R) Core(TM) i7-10750H CPU @ 2.60GHz, 16 GB RAM. Surprising.

My question to the authors is: Since recent studies have revealed that in more than 50% of structures in the CCDC, energy differences between pairs of polymorphs are smaller than ∼2 kJ/mol, while only about 5% have energy differences larger than ∼7 kJ/-mol, as they underline in the text, and they do not report any energy calculation, it would be mandatory:

- Report, please, the energy differences between aspirin polymorphs, between the 3 to 10 better structures respect the experimental ones. The comparison is valuable to have an insight into the degree of precision of the CrysMath.

  In the updated CrystalMath protocol, the search of aspirin yielded only 3 structures in the final pool, that correspond to the three known experimental structures of aspirin. Structures in the intermediate pools are not compared against the experimental since the vast majority of them include unphysical close contacts and have very high vdW free volume, rendering them as upnhysical structures. To validate the cost function ranking, as requested by the reviewer, we report a two different energy calculations of the predicted aspirin structures: (a) one performed using the UNI intermolecular potentials implemented in the CSD python API, and (b) the relative energies extracted from single-point DFT PBE0+MBD energy calculations. Both shows that the combination of the vdWFV-C_{ICC} provides a sufficient ranking scheme, consistent with experimental measurements.

- Report, please, also the same comparison for Targets XXII and XXIII from the 6th CCDC blind test, to infer something else about degree precision of CrysMath.

- I consider that those degrees of precision must be checked. There are many CODES to make the evaluation with a DFT-D method. I suppose they must have access to one of them. Quantum Espresso code, by the way, is free access.

  For the updated target XXII prediction, we identified only one structure in the final pool, so we feel it is not necessary to provide energy calculations. For the target XXIII structure, as requested by the reviewer, we performed DFT PBE0+MBD energy calculations, similar to the case of aspirin, to compare the cost function ranking against the energy ranking of the predicted structures. The results of the comparison are described in the revised manuscript.

To my criteria, the method is useful to get a list of several possible crystal structures of a compound and might be considered a useful tool for proper subsequent first principles refinement calculations, considering also the importance of temperature pressure conditions present in crystallization.

It might be useful also when no single crystals are available and the experimental structure must be determined by other techniques like X-ray powder diffraction or solid-state NMR, to have a reliable list of even several hundred possible candidate structures can be extremely valuable.

It would be important that, in a near future, the authors check performance and degree of precision of the method in different kinds of crystals, co crystals, chiral compounds.

We agree with the reviewer on this point. Future work will consist in improving our ranking methodology and considering a wider variety of crystal systems such as the reviewer suggested (co=crystals, hydrates, chiral compounds,...). As to having longer lists of more candidates, these can be generated within CrystalMath by loosening the criteria for adherence to the CrystalMath principles.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The authors' response satisfies most of my concerns. The authors have not addressed my comment that "the abstract should mention that van der Waals volumes and CSD derived ranges of atom-atom contact distances (c.f. van der Waals radii) were used" but do admit that this is the case in their comment that "the lengths of the contacts adhere to the distributions observed in the database.". The authors "believe CrystalMath is doing more than simply applying close-packing principles," and I would agree that it is using the limitations of translational symmetry on the close packing and also they are using the most electropositive atoms. So just adding a line "and some physical principles" to the first line of the discussion seems an inadequate acknowledgement of the extent that this methodology relies on the CSD distributions and the proximity of highly electropositive/electronegative atoms. The fragment geometries are also derived from the CSD (p10). Thus there should be some mention of the CSD in the abstract and ideally more to make the CrystalMath process sound physically reasonable to a crystal engineer. The removal of ROY from the SI and the discussion of the further work that needs developing in Crystal Math for large flexible molecules is welcome. The results are certainly impressive and warrant publication in Nature.
I am happy for publication to proceed, but there are some minor points that need attention:
The figure numbering and referencing needs checking. Fig 1 contains figs A1, A2, B, C, D1, D2, but the caption implies it contains C1-C3 and D1-D3. Furthermore the caption states that this is for C, H and O atoms. However text in the discussion p12 has Figure 1 and 2 coming from distributions involving C, H, N, O, F, Cl, Br and I and Fig 2 does not contain any distributions. P12 also refers to Fig 1 C1-C2). The reference to Fig 1 in the paragraph that goes over p7 and 8 needs to be more specific.
P8 COMPAC (or COMPACK?) requires a reference.
P11 Are "base structures" conformations
P12 dot should be not

Reviewer #2 (Remarks to the Author):

Some of the original review comments have been addressed. Below are my comments on the revised manuscript.

The authors now present results for three molecules (aspirin and blind test molecules XXII and XXIII. In all cases, the crytal structure prediction using CrystalMath seems to be perfect: all known polymorphs are predicted and only the structures corresponding to known polymorphs are predicted. The case (ROY) from the first version of the manuscript, where the known polymorphs were predicted, but also many other crystal structures, some of which had better calculated free volumnes than known polymorphs, has been removed. According to the authors: "We appreciate the reviewer's comments on our presentation of ROY in the SI. Because of the relatively high number of ROY polymorphs that are already known and the possibility of more exotic structures that may yet to be discovered, we have decided that a presentation of ROY in the SI does not do the problem justice." I can see that the authors would like more space to discuss the ROY results. However, I find the deletion of these results worrying. i) A large number of polymorphs is good as a test system. ii) Why is there a greater possibility of "exotic structures" of ROY than the other systems studies? It could be argued that ROY has been studied so thoroughly that its polymorphism is now well characterised. iii) It leaves a manuscript presenting apparently perfect results, which will certainly create a high impact impression of the method. It is left to the authors to follow this up with a presentation of other results where this is not the case, including ROY. The field of crystal structure prediction has seen this before: metadynamics was presented as solving CSP 20 years ago, https://doi.org/10.1002/anie.200462760, and that paper is highly cited. However, the method is never used for crystal structure prediction because it did not work as "sold" in follow-up studies. This is a reason for the blind tests of crystal structure prediction.

I do not think that this case is the same: the results for ROY from the earlier version were still impressive, but present a different picture from those that have been included in the revised version. I would encourage the authors to include the ROY results in the main manuscript and discuss them thoroughly, rather than delete them, to give the readers a better complete picture of the overall results of the method. I do not think that the authors intend to mislead readers with this decision, but the result will be that it does exactly this. With the removal of these results, I am slightly uncomfortable with endorsing publication.

Regarding aspirin results, my earlier comments had asked for more information about filtering of structures. The reply is as follows: "The ”CrystalMath Protocol” section was revised to include latest developments and provide additional details on the filtering process. A figure was added in the SI section 4, to illustrate the close contacts filtering and the acceptance/rejection of structures on that basis. In addition, per the reviewer’s request, we have made the low vdW free-volume structures discarded throughout the CrystalMath protocol available in the supporting information." I do not find section 4 of the SI to be very informative. The caption to the figure S6 states "Structures can be discarded on the basis of having unnatural short contacts (C1) and/or on the basis of hydrogen bond absence in cases where their existence is required (C2)." Where are the criteria for the expected presence or absense of hydrogen bonds between specific acceptor and donor atoms provided? This is not a straightforward decision, as real molecules will often have competing acceptor and donor atoms. Aspects of the filtering are still opaque to me.

About energy calculations: why would you use the UNI potential, which only assesses intermolecular interactions, when the molecules studied here are all flexible, so that an assessment of relative energies requires intermolecular and intramolecular energy differences?

Reviewer #3 (Remarks to the Author):

I have carefully revised this new version of the manuscript "Topological Crystal Structure Prediction" by Galanakis and Tuckerman.

I consider that the authors have taken into account my questions and doubts about the original version submitted to Nature Communications.

They made a very good revision considering the three referees’ reports. This new version is improved respect the original one, and makes a very valuable contribution to the difficult task of predicting crystal structures
The manuscript is worthy of publication in Nature Communications

Responses to Reviewers' Comments

Reviewer #1

The authors' response satisfies most of my concerns. The authors have not addressed my comment that "the abstract should mention that van der Waals volumes and CSD derived ranges of atom-atom contact distances (c.f. van der Waals radii) were used" but do admit that this is the case in their comment that "the lengths of the contacts adhere to the distributions observed in the database.". The authors "believe CrystalMath is doing more than simply applying close-packing principles," and I would agree that it is using the limitations of translational symmetry on the close packing and also they are using the most electropositive atoms. So just adding a line "and some physical principles" to the first line of the discussion seems an inadequate acknowledgement of the extent that this methodology relies on the CSD distributions and the proximity of highly electropositive/electronegative atoms. The fragment geometries are also derived from the CSD (p10). Thus there should be some mention of the CSD in the abstract and ideally more to make the CrystalMath process sound physically reasonable to a crystal engineer.

We have revised the abstract based on the reviewer's comments to acknowledge the importance of the CSD database in formulating the present CrystalMath protocol.

The removal of ROY from the SI and the discussion of the further work that needs developing in Crystal Math for large flexible molecules is welcome. The results are certainly impressive and warrant publication in Nature. I am happy for publication to proceed, but there are some minor points that need attention:

- *The figure numbering and referencing needs checking. Fig 1 contains figs A1, A2, B, C, D1, D2, but the caption implies it contains C1-C3 and D1-D3.*
- *Furthermore the caption states that this is for C, H and O atoms. However text in the discussion p12 has Figure 1 and 2 coming from distributions involving C, H, N, O, F, Cl, Br and I and Fig 2 does not contain any distributions.*
- *P12 also refers to Fig 1 C1-C2). The reference to Fig 1 in the paragraph that goes over p7 and 8 needs to be more specific.*
- *P8 COMPAC (or COMPACK?) requires a reference.*
- *P11 Are "base structures" conformations*
- *P12 dot should be not*

The manuscript has been revised to address the reviewer's comments. For the comment regarding the base structures, we modified the manuscript in page 7 to provide a clear definition of the base structures.

Reviewer #2

Some of the original review comments have been addressed. Below are my comments on the revised manuscript.

The authors now present results for three molecules (aspirin and blind test molecules XXII and XXIII. In all cases, the crytal structure prediction using CrystalMath seems to be perfect: all known polymorphs are predicted and only the structures corresponding to known polymorphs are predicted. The case (ROY) from the first version of the manuscript, where the known polymorphs were predicted, but also many other crystal structures, some of which had better calculated free volumnes than known polymorphs, has been removed. According to the authors: "We appreciate the reviewer's comments on our presentation of ROY in the SI. Because of the relatively high number of ROY polymorphs that are already known and the possibility of more exotic structures that may yet to be discovered, we have decided that a presentation of ROY in the SI does not do the problem justice." I can see that the authors would like more space to discuss the ROY results. However, I find the deletion of these results worrying. i) A large number of polymorphs is good as a test system. ii) Why is there a greater possibility of "exotic structures" of ROY than the other systems studies? It could be argued that ROY has been studied so thoroughly that its polymorphism is now well characterised. iii) It leaves a manuscript presenting apparently perfect results, which will certainly create a high impact impression of the method. It is left to the authors to follow this up with a presentation of other results where this is not the case, including ROY. The field of crystal structure prediction has seen this before: metadynamics was presented as solving CSP 20 years ago, https://doi.org/10.1002/anie.200462760, and that paper is highly cited. However, the method is never used for crystal structure prediction because it did not work as "sold" in follow-up studies. This is a reason for the blind tests of crystal structure prediction.

I do not think that this case is the same: the results for ROY from the earlier version were still impressive, but present a different picture from those that have been included in the revised version. I would encourage the authors to include the ROY results in the main manuscript and discuss them thoroughly, rather than delete them, to give the readers a better complete picture of the overall results of the method. I do not think that the authors intend to mislead readers with this decision, but the result will be that it does exactly this. With the removal of these results, I am slightly uncomfortable with endorsing publication.

We have incorporated the updated results for ROY generated using the revised CrystalMath protocol applied to all other structures in the manuscript, into the main text. Compared to the original submission, the results for ROY are improved due to the implementation of more accurate topological filters in the updated CrystalMath protocol. This refinement has significantly reduced the number of possible ROY structures in the final pool from 48 to 19.

Regarding aspirin results, my earlier comments had asked for more information about filtering of structures. The reply is as follows: "The "CrystalMath Protocol" section was revised to include latest developments and provide additional details on the filtering process. A figure was added in the SI section 4, to illustrate the close contacts filtering and the acceptance/rejection of structures on that basis. In addition, per the reviewer's request, we have made the low vdW free-volume structures discarded throughout the CrystalMath protocol available in the supporting information."

I do not find section 4 of the SI to be very informative. The caption to the figure S6 states "Structures can be discarded on the basis of having unnatural short contacts (C1) and/or on the basis of hydrogen bond absence in cases where their existence is required (C2)." Where are the criteria for the expected presence or absense of hydrogen bonds between specific acceptor and donor atoms provided? This is not a straightforward decision, as real molecules will often have competing acceptor and donor atoms. Aspects of the filtering are still opaque to me.

A new, more informative, version of Fig. S.6 was created detailing the process of filtering and accepting/rejecting structures on the basis of the close contacts distribution. For the aspirin molecule, given the presence of the carboxyl group and several oxygen atoms in the aspirin molecule, it is highly unlikely for aspirin to form a crystal with no hydrogen bonds. It is logical argument to discard generated crystal with no hydrogen bonds between the neighbouring molecules. However, CrystalMath, allows the generation of hydrogen bonds between all possible combination of donor/acceptor pairs. For the other molecules in presented in the current manuscript, no hydrogen bond criterion was applied, and structures with no hydrogen bonds were allowed to be generated and accepted in the various filtering steps.

About energy calculations: why would you use the UNI potential, which only assesses intermolecular interactions, when the molecules studied here are all flexible, so that an assessment of relative energies requires intermolecular and intramolecular energy differences?

We acknowledge the limitations of the UNI potential in accurately calculating energy differences, as it disregards intramolecular energy contributions. This makes it unsuitable for the energetic ranking of crystal structures that exhibit polymorphism. Nevertheless, we chose to use the UNI potential to benchmark the cost function ranking against a simple alternative with non-system-specific force field parameters. This approach allowed us to conduct a broad comparison against the custom cost function used in CrystalMath protocol, utilizing the diverse range of structures available in the CSD database.

Reviewer #3

I have carefully revised this new version of the manuscript "Topological Crystal Structure Prediction" by Galanakis and Tuckerman.

I consider that the authors have taken into account my questions and doubts about the original version submitted to Nature Communications.

They made a very good revision considering the three referees' reports. This new version is improved respect the original one, and makes a very valuable contribution to the difficult task of predicting crystal structures.

The manuscript is worthy of publication in Nature Communications.

We appreciate all of the Reviewer's efforts and the positive comment on our latest revision.

REVIEWERS' COMMENTS

Reviewer #2 (Remarks to the Author):

Having read through the revised manuscript and comments from the authors, I am satisfied that the authors have addressed the concerns from my earlier reviews.