CrystEngComm

COMMUNICATION

View Article Online
View Journal | View Issue

Cite this: *CrystEngComm*, 2024, **26**, 5845

Received 29th July 2024,
Accepted 24th September 2024

DOI: 10.1039/d4ce00752b

rsc.li/crystengcomm

Robust and efficient reranking in crystal structure prediction: a data driven method for real-life molecules†

Andrea Anelli, *[a] Hanno Dietrich, [b] Philipp Ectors,[c] Frank Stowasser,[a] Tristan Bereau, [d] Marcus Neumann[b] and Joost van den Ende[a]

We accelerate a key step in crystal structure prediction (CSP) using machine learning and report its robustness on a wide array of pharmaceutical molecules. The speedup achieved by our scheme allows for a scale-up in both the number of candidate drug molecules studied and the level of theory employed in their treatment, paving the way for tackling more complex crystal energy landscapes.

Finding the experimentally relevant solid forms of a molecule is a fundamental step in the development of a new small-molecule drug. The specific crystalline structure a candidate drug molecule can solidify into has a deep impact on its development track, as it directly affects its synthesis process and many of its bioavailability metrics, *e.g.* crystallization behavior, solubility, and dissolution rate.[1,2] The search for stable solid forms is usually carried out under various experimental conditions (*e.g.* solvents, temperature, pressure and external fields), requiring extensive work and availability of sufficient material.[3,4] However, even a thorough experimental screening can never completely rule out the possibility of the thermodynamically stable form appearing late in a drug development process under specific experimental conditions. The wider availability of computational resources, combined with advances in modelling molecular solids, has enabled *in silico* CSP to become a complementary route for solid form screening.

Successful CSP requires accurate energies and a thorough sampling of the crystal energy landscape. Traditionally, this means striking a trade-off between the accuracy needed to capture the relevant physical interactions and efficiency in being able to scan large pools of candidate structures.[5,6] For this reason, CSP is often carried out in a hierarchical fashion: starting from less accurate though fast force field measures of stabilities, one finds a promising subset of configurations on which more accurate calculations are performed. Since the seminal paper of Gavezzotti in 1994 (ref. 7) many developments around CSP have taken place as is, for example, reflected by the Sixth Blind Test on Organic Crystal Structure Prediction Methods[8] and its predecessors. An essential step to enable robust and efficient1 generation of crystal packings has been the introduction of tailor-made force fields (TMFF).[9] To gain precision in ranking the generated packings, the usage of dispersion corrected density functional theory (DFT-D) has been proven to be successful.[10] A more recent advance has been the incorporation of hybrid functionals, higher order dispersion corrections and approximations to the vibrational free energy[5] in order to estimate the effect of temperature on the relative stabilities of polymorphic forms. A correction of the conformational energy at the post-DFT level has been shown to be important in the highly polymorphic ROY case.[11] All these improvements have recently been combined and assessed against a carefully selected experimental benchmark.[12] While the accuracy of theoretical frameworks has increased, so has the computational footprint needed to leverage them. Atomistic machine learning has shown promise in accelerating the predictions of the stabilities of atomic[13–15] and molecular[16] materials at various levels of theory,[17–19] though, to the best of our knowledge, its impact on the efficiency of CSP applied to pharmaceutical compounds has not yet been reported.

In this communication, we present a scheme to integrate an atomistic machine learning model into an archetypical CSP workflow which bridges the generative step and the *ab initio* ranking process. By constructing an ML-reranker we enable a more efficient and modular (*e.g.* by substitution of different acquisition functions or ML models) screening procedure. The core of our approach is a regression model mapping the generated force field minima structures to the energies they would obtain upon minimization on a reference

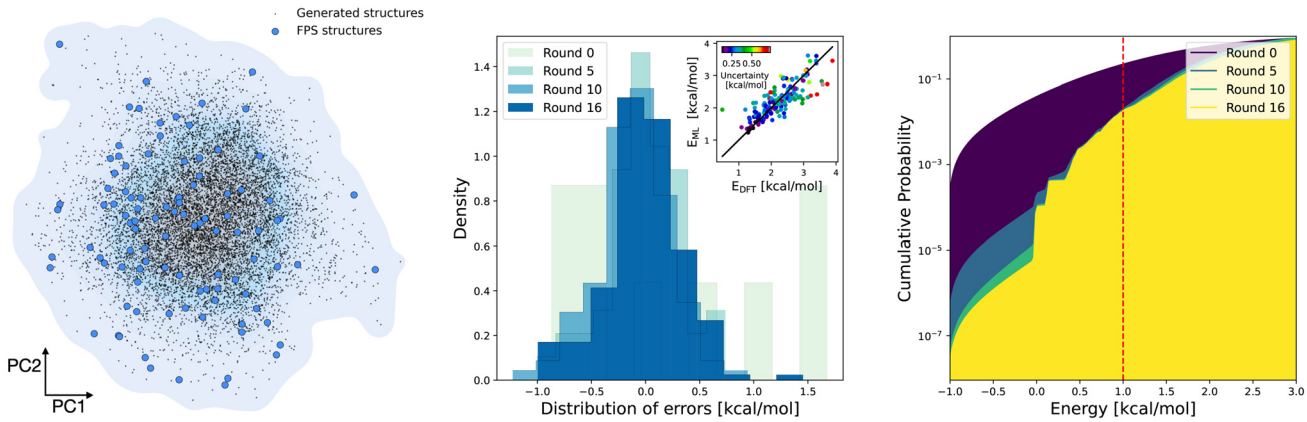
[a] *Roche Pharma Research and Early Development, Therapeutic Modalities, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd., Basel, Switzerland. E-mail: andrea.anelli@roche.com*
[b] *Avant-garde Materials Simulation Deutschland GmbH, Merzhausen, Germany*
[c] *Pharma Technical Development, F. Hoffmann-La Roche Ltd., Basel, Switzerland*
[d] *Institute for Theoretical Physics, Heidelberg University, Heidelberg, Germany*
† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d4ce00752b

ab initio potential energy surface. This minima predictor is built as a delta model baselined on force field energies and contains a measure of confidence indicating how likely the structures are to fall close to the target prediction.

We show that our approach accurately selects all the configurations which populate the low-energy DFT minima basins, and provides a reliable uncertainty estimate. In our approach, we focus on atom-centered representations. This characteristic allows our model to express confident predictions on a size extensive basis, making it reliable across the broad structural space of the search, and increasing Z' numbers, irrespective of the cell sizes.

In order to produce the crystal structure landscapes for our case studies, we use the GRACE 2.7.49 (ref. 20) software package. Initially, a tailor-made force field (TMFF) is constructed and is used to perform a thorough scan of the structure's crystalline configurations, acting as our generative step. The parameterization of each TMFF is carried out independently for each drug molecule, so as to adapt to the different molecular conformations, bonded and non-bonded interactions governing the system of choice. The CSP takes place by using the most commonly occurring space groups populating $Z' = 1, 2$. Following the generative step, we have access to a force field energy ranking of the different crystal structures. To refine the CSP results, we perform geometry optimizations on the number of structures needed to sample the population contained within a target energy window E_{window}. The reranking of low energy TMFF structures is obtained by performing PBE–DFT energy minimizations using the Neumann Perrin dispersion correction and the software package VASP[21,22] for the calculation of DFT energies and forces. The ranking obtained by this last set of optimizations is considered as the reranked landscape. This procedure aims at capturing the relevant minima of the crystal energy landscape and can be followed by a finite temperature correction step to account for entropic contributions and further *ab initio* minimizations at higher levels of theory. The objective of this work focuses on the reranking of the zero kelvin landscape, and as a result, no finite temperature sampling is considered in this manuscript.

To construct a predictor of the lattice energies we train a ridge regression model on average SOAP[23] power spectra per TMFF crystal packing. We chose this representation because of its wide success in the field of molecular crystal property predictions and transparency of its feature space.[24,25] The advantage of such a simple scheme is the absence of any hyperparameter besides the Tikhonov regularization term, which is optimized with an 8-fold cross-validation split at each training step. We additionally use a delta learning scheme,[26] predicting the differences between the TMFF and the DFT energies, to reduce the variance of the target property. Finally, by constructing a committee of models following Imbalzano *et al.*,[27] we introduce a confidence interval into each prediction. For details on the implementation of the ML model, we refer to the ESI† in Sections S1 and S2. The machine learning model is used to guide the selection of the structures that exhibit promise of being low in energy upon an *ab initio* minimization. After completing the generation step in the CSP routine, we follow an active learning approach to iteratively train a minima-to-minima machine learning model mapping the energy of the TMFF minima to their corresponding DFT minimized counterparts. The scheme follows the workflow proposed in Fig. 1. To initialize our model, we first minimize the number of candidate structures, N_{init}, covering the generated space uniformly using a farthest point sampling algorithm,[28] and training a committee of ridge regression models on them



(a) SOAP-PCA map of X_{gen} (b) Error distributions of the trained ML model (c) Convergence probability across each round

Fig. 1 Workflow of the ML-reranker: (a) principal components map of the generation pool (first two components) extracted on the soap vectors. The workflow is initialized by sampling N_{init} configuration from the TMFF-generated pool using furthest point sampling. Every selected crystal undergoes a geometric relaxation. (b) A committee of energy models is fitted on the sampled data and used to predict the stability (with their uncertainties) of the yet-to-sample generation pool structures. (c) Exploiting the available uncertainties, we calculate $p(E < E_{\text{window}})$: the cumulative probability function of having left a crystal with energy smaller than a desired window. When this value reaches the required convergence, the sampling stops. If not, new structures are sampled according to a desired acquisition function and go back to step (b). The different rounds depicted in the figure indicate how the models' accuracy and the reranking convergence improve over the sampling iterations.

Fig. 1a. We use this ensemble to predict the energy change the structures would undergo after a DFT minimization. The energy predicted is the mean of the ensemble distribution and the estimated uncertainty is its standard deviation.

The model is trained on the first 80% of the sampled structures, while the remaining 20% is used to evaluate its performance. The error distribution and accuracy of the trained models can be visualized independently as shown in Fig. 1b. To select a new configuration, we leverage an exploitation–exploration selection criteria following the expected improvement function proposed in ref. 29, and apply it to perform a batch selection of N_{batch} points at every iteration from the generation pool. The details on the implementation of the batch selectors are provided in ESI† Sections S2 and S3. Finally, we sum all the probabilities of each configuration of having energy E and estimate the probability of having left in the generation set (*i.e.* not sampled yet) a configuration with an energy lower than E_{best} + E_{window}, with E_{best} being the current lowest energy DFT structures, and E_{window} the target energy window to sample. To estimate this probability, we calculate the integral of the not-yet-sampled generation set's energy cumulative probability distribution until the desired window, as shown in Fig. 1c. This number provides us with the expected number of configurations left which could be below the window: N_{left}. We thus expect a total number of configurations defined as $N_{expected} = N_{left} + N_{found}$, where N_{found} is simply the number of structures found so far with DFT energies within the desired window. We introduce a convergence probability as $p_{conv} = N_{found}/N_{expected}$. If the probability is higher than a desired confidence $p_{threshold}$, we consider our TMFF basin exhausted, and the corresponding DFT landscape satisfactorily surveyed. A more detailed description of the convergence criteria is provided in ESI† S3.

The first molecule of our study is fentanyl. The proposed compound has 54 atoms and consists of 6 rotatable bonds, four chemical species (H, C, N, and O), two rings, and a molecular weight of 336.5 g mol^{-1}. The central plot of Fig. 2 shows the energy-density scatter plot for the reference CSP carried out with GRACE (in mint green). The low energy structure appearing as a ground state is competing with several other configurations within the target energy window, indicating a rich crystal landscape. The CSP workflow as implemented in GRACE found 10 342 low energy packings using a TMFF with a convergence threshold of 0.99 and 0.95 for structures having one (Z' = 1) and two (Z' = 2) independent molecules in their asymmetric units, respectively. The unpublished standard GRACE reranking algorithm converged the sampling of a window of 1 kcal mol^{-1} by optimizing a total of 2345 structures with PBE–DFT. To showcase the effectiveness of our minima to minima mapping, in the central plot in Fig. 2 we report the crystal structure landscape obtained by the 16th step of our approach, compared to the one obtained by a standard GRACE reranking. The comparison is performed by coloring each configuration found on the GRACE landscape depending on whether its originating force field configuration had been sampled by the ML-reranker or not. Circles in mint green indicate agreement between the two methods, while circles in light grey show a structure that only GRACE found. Notably, within this exercise, the ML-reranker did not find one configuration that GRACE did not sample – showcasing the robustness of the sampling implemented in GRACE. The two reranking datasets and the generation pool are available on Zenodo.[30]

The ML-reranker was initialized using N_{init} = 100 structures extracted from the TMFF generation pool, and using N_{batch} = 105 for each iteration. The exploitation energy window was set to E_{window} = 1 kcal mol^{-1} and target residual probability to $p_{threshold}$ = 0.99. Both GRACE and the proposed approach capture the ground state crystal structure, and sample in a comparable way the target energy window. The improvement of

N_{round}	$N_{sampled}$	p_{conv}
0 (FPS)	90	0.007
1	178	0.176
2	267	0.289
4	447	0.544
6	643	0.577
8	842	0.213
10	1051	0.948
12	1260	0.760
14	1469	0.966
15	1574	0.988
16	1678	0.993
GRACE	2523	0.99/0.95

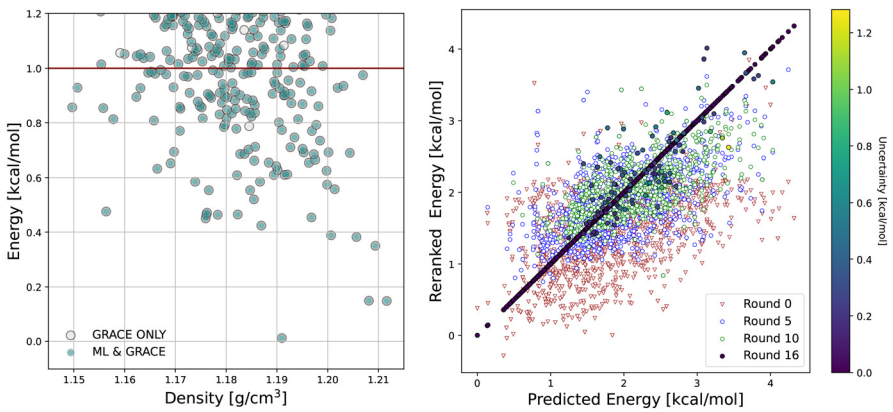


Fig. 2 An overview of the ML-reranker performances on fentanyl. On the left-hand side table we report the convergence of the algorithm metrics with respect to the number of iterations. $N_{sampled}$ indicates the number of structures minimized at N_{round} step. The central figure shows the mapping of the ML-reranker results to the grace crystal structure landscape. The coloring reflects whether the point's originating force field structure has been selected by grace only (white) or both (blue). The right-hand side plot shows how the model improves across the different iterations. The points' shape reflects their iteration round. The color of round 16 points' reflects the uncertainty in their predictions – with dark points along the diagonal indicating sampled structures (thus with "zero" uncertainty).

the model and its bias in selecting configurations exhibiting low energy is shown in the right scatter plot in Fig. 2, where a correlation plot between the TMFF sampled energies and their ML predictions is shown. The correlation plot shows low errors for points below the energy window – mostly lying on the diagonal (black points showing a zero error correspond to structures that have been sampled, thus having zero uncertainty in their predicted value). Out of 174 structures in the target energy window, only one has not been found by the ML-reranker in agreement with the value of $p_{\text{threshold}}$ configured to select 99% of the structures within the desired window contained in the initial generation set. When looking at the results from a data-efficiency perspective, this translates to 845 DFT minimization spared, corresponding to 34% savings on the reranking computational cost. From a wall-clock perspective, based on an estimated average optimization time per structure, this amounts to roughly 3000 node-hours using 96 cores.

To obtain a picture of the general applicability of our approach, we report the results on 17 compounds from Roche's portfolio. For each of these molecules, a complete reranking of TMFF structures using the standard GRACE reranker resulted in a subset of reranked configurations. These selections contain configurations spanning up to 10 kcal mol^{-1} above the ground state and constitute a smaller yet challenging dataset. Differently from the previous exercise, we will test the efficiency of our approach in extracting the lower energy structures with respect to their reranked sets. While this example is not directly predictive of the performance of the model on a realistic reranking scenario (as the reranked landscapes are smaller and more homogeneous), we show in Table 1 how this defines an upper bound for the performance gain that can be expected from this novel approach within this simplified context. The extended compound selection exercise serves the purpose of showing the applicability of our reranking approach to a wide variety of complex chemical spaces, with an average molecular weight of 500 g mol^{-1}, 7 rotatable bonds, and 6 rings. The summary of pharmacological metrics is shown in Fig. S4 in the ESI.† The performance of the ML-reranker is consistent with what we have observed in the fentanyl case. We remark that this metric serves the purpose of showing the ML-reranker robustness in sparse data regimes but does not constitute a fair comparison with GRACE's standard algorithm. GRACE had to select from all the TMFF crystal structures whereas the ML algorithm only selects from a smaller set of structures that were initially chosen by GRACE in the standard reranking and then replaced by their corresponding TMFFs for the computational exercise presented here. The parameters used across the API benchmark are kept constant at N_{init} = 100, N_{batch} = 100, E_{window} = 1 kcal mol^{-1} and residual probability of $p_{\text{threshold}}$ = 0.99. A summary of the results obtained across the whole set is shown in Table 1. This analysis shows how reliable the estimation of the convergence probability is throughout the set of different molecules. By targeting a 99% convergence across a total of 475 structures within all the windows, the ML-reranker fails to select 2 packings corresponding to an observed convergence probability of 99.6%. Further, our method reliably accelerates the sampling across the complex chemical space surveyed, as shown in the forward exercise – and the similar performances observed in the reranked dataset. Thirdly, we demonstrate how a sub 1 kcal mol^{-1} ML energy model can be trained efficiently to obtain reductions of minimization rounds at no additional cost.

Table 1 Summary of results on the reranking datasets from a selection of internal Roche compounds (ROX) and fentanyl set (FTN) for which a complete CSP has been carried out. The reranking pool of each molecule contains N_{tot} structures and is completely sampled by the ML-reranker (with a $p_{\text{threshold}}$ = 0.99) selecting N_{sampled} structures. The total number of configurations within the energy window is known, and equal to N_{below}, and the number of structures found by the ML-reranker at the end of the algorithm is N_{found}. In case a structure (or more) is missed, we report in E_{min} the lowest energy that has not been sampled in kcal mol^{-1}. Finally, we report the last iteration's model RMSE, calculated against the last 20% structures contained in the so-far sampled dataset and using the first 80% as training set

ID	N_{tot}	N_{sampled}	N_{below}	N_{found}	E_{min}	RMSE
RO1	3149	1110	12	12	—	0.71
RO2	905	706	26	26	—	0.83
RO3	1186	403	3	3	—	0.93
RO4	1606	504	3	3	—	0.85
RO5	1796	1110	21	21	—	0.68
RO6	1650	908	106	106	—	0.50
RO7	2535	1413	8	8	—	0.88
RO8	1678	908	5	5	—	0.66
RO9	2307	1514	3	3	—	1.01
RO10	856	807	7	7	—	0.62
RO11	857	807	16	16	—	0.84
RO12	1766	1009	9	9	—	0.85
RO13	2852	1413	12	11	0.98	0.72
RO14	1481	1110	5	5	—	0.86
RO15	1405	1009	1	1	—	0.97
RO16	1009	504	3	3	—	1.02
RO17	5034	908	95	94	0.76	0.43
FTN	2523	1110	140	140	—	0.37
Total	34 595	17 253	475	473	0.76	0.76

Extending the range of CSP to larger and larger molecules requires constant improvements of the various components forming a complete CSP workflow. In this work, we have investigated the performance of an ML-reranker based on standard ML approaches as an alternative to the unpublished GRACE reranker based on statistical correlations and detailed domain knowledge. The results of this investigation prove that this strategy offers substantial improvements in the CSP reranking exercise without incurring any additional cost, due to a simple efficient re-use of the data explored during the landscape explorations. The ML-reranker is an appealing component for anyone who wants to construct a CSP workflow from scratch and offers potential improvements in already honed architectures. The ML-reranker (i) reduces the number of *ab initio* crystal structure optimizations by the on-the-fly improvement of an ML surrogate model and (ii) offers an iterative sampling scheme with robust convergence behavior determined by a single threshold parameter.

Data availability

The data supporting this article have been included as part of the ESI.† Data for the fentanyl landscape, including the indexes of the structures chosen and their corresponding DFT and TMFF energies, are available on Zenodo at **https://doi.org/10.5281/zenodo.13362263**. The crystal structure landscapes of the Roche internal APIs and the code used in the manuscript are proprietary and as such they cannot be shared.

Author contributions

A. Anelli and J. van den Ende designed the research and wrote the manuscript. H. Dietrich and M. Neumann provided the implementation to GRACE and support to the benchmark analysis. T. Bereau and A. Anelli have designed the optimization workflow. J. van den Ende, F. Stowasser and P. Ectors have contributed to the calculation of the API dataset. A. Anelli has implemented the workflow and designed the algorithm.

Conflicts of interest

Marcus A. Neumann is the owner of Avant-garde Materials Simulation (AMS). Hanno Dietrich is an employee of AMS. AMS develops and distributes the GRACE code for crystal structure prediction.

Acknowledgements

A. Anelli acknowledges funding by the Roche Postdoctoral Fellowship (RPF) programme.

Notes and references

1 C. Taylor, M. Mulvee, D. Perenyi, M. Probert, G. Day and J. Steed, *J. Am. Chem. Soc.*, 2020, **142**, 16668–16680.
2 S. Datta and D. Grant, *Nat. Rev. Drug Discovery*, 2004, **3**, 42–57.
3 A. Newman, *Org. Process Res. Dev.*, 2013, **17**, 457–471.
4 J. Aaltonen, M. Alleso, S. Mirza, V. Koradia, K. C. Gordon and J. Rantanen, *Eur. J. Pharm. Biopharm.*, 2009, **71**, 23–37.
5 J. Hoja, *et al.*, *Sci. Adv.*, 2019, **5**, eaau3338.
6 G. M. Day, *Crystallogr. Rev.*, 2011, **17**, 3–52.
7 A. Gavezzotti, *Acc. Chem. Res.*, 1994, **27**, 309–314.
8 A. M. Reilly, *et al.*, *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.*, 2016, **72**, 439–459.
9 M. A. Neumann, *J. Phys. Chem. B*, 2008, **112**, 9810–9829.
10 M. Neumann, F. Leusen and J. Kendrick, *Angew. Chem., Int. Ed.*, 2008, **47**, 2427–2430.
11 G. J. Beran, *et al.*, *Chem. Sci.*, 2022, **13**, 1288–1297.
12 D. Firaha, Y. M. Liu, J. van de Streek, K. Sasikumar, H. Dietrich, J. Helfferich, L. Aerts, D. E. Braun, A. Broo and A. G. DiPasquale, *et al.*, *Nature*, 2023, **623**, 324–328.
13 N. Lopanitsyna, *et al.*, *Phys. Rev. Mater.*, 2021, **5**, 043802.
14 S. Batzner, A. Musaelian and B. Kozinsky, *Nat. Rev. Phys.*, 2023, 1–2.
15 V. Kapil, D. Kovacs, G. Csanyi and A. Michaelides, *Faraday Discuss.*, 2023, 50–68.
16 D. P. Kovacs, *et al.*, *J. Chem. Theory Comput.*, 2021, **17**, 7696–7711.
17 F. Brockherde, *et al.*, *Nat. Commun.*, 2017, **8**, 872.
18 A. Tirelli, *et al.*, *Phys. Rev. B*, 2022, **106**, L041105.
19 C. Zeni, A. Anelli, A. Glielmo, S. de Gironcoli and K. Rossi, *Digital Discovery*, 2024, **3**, 113–121.
20 Product of Avant-garde Materials Simulation Deutschland GmbH.
21 G. Kresse and J. Hafner, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1994, **49**, 14251.
22 G. Kresse and J. Furthmuller, *Comput. Mater. Sci.*, 1996, **6**, 15.
23 A. P. Bartok, R. Kondor and G. Csanyi, *Phys. Rev. B*, 2017, **96**, 019902.
24 F. Musil, S. De, J. Yang, J. E. Campbell, G. M. Day and M. Ceriotti, *Chem. Sci.*, 2018, **9**, 1289–1300.
25 J. P. Darby, J. R. Kermode and G. Csanyi, *npj Comput. Mater.*, 2022, **8**, 166.
26 R. Ramakrishnan, *et al.*, *J. Chem. Theory Comput.*, 2015, **11**, 2087–2096.
27 G. Imbalzano, *et al.*, *J. Chem. Phys.*, 2021, **154**, 074102.
28 G. Imbalzano, *et al.*, *J. Chem. Phys.*, 2018, **148**, 241730.
29 B. Mohr, K. Shmilovich, I. S. Kleinwachter, D. Schneider, A. L. Ferguson and T. Bereau, *Chem. Sci.*, 2022, **13**, 4498–4511.
30 A. Anelli, *et al.*, *Robust and Efficient Reranking in Crystal Structure Prediction: A Data Driven Method for Real-life Molecules*, 2024, DOI: **10.5281/zenodo.13362263**.