Supplementary Information (SI) for Digital Discovery.
This journal is © The Royal Society of Chemistry 2024

Supporting Information

Powder X-Ray Diffraction Assisted Evolutionary Algorithm for Crystal Structure Prediction

Stefano Racioppi, Alberto Otero De la Roza, Samad Hajinazar, Eva Zurek*

Table of Contents

S1. Estimation of the fitness value for a candidate structure

Equation 1 (main text) provided a formula to calculate the fitness of a candidate structure. This expression, with a slight reformulation, can be used to obtain insight into the values of the multi-objective optimization weight that can guide the search more effectively towards the desired regions of the energy landscape. The following presents an example of this analysis, applied to our results for the TiO_2 search.

Assuming that the enthalpy and PXRD similarity index of the s^{th} structure are given by H_s and S_s, respectively, the fitness is calculated from:

$$f_s = f_H + f_S = (1-w)\left(\frac{H_{max} - H_s}{H_{max} - H_{min}}\right) + w\left(\frac{S_{max} - S_s}{S_{max} - S_{min}}\right)$$

where w is the weight for the similarity index. H_{max} and H_{min} are the enthalpies of the highest and lowest enthalpy structures, whereas S_{max} and S_{min} are the similarity indices of the structures that afford the worst/best match with the PXRD.

For the enthalpy, we can write

$$H_{max} = H_{min} + \Delta H$$

which can be used to reformulate the enthalpy's share to the total fitness as:

$$f_H = (1-w)\left(\frac{\Delta H + H_{min} - H_s}{\Delta H + H_{min} - H_{min}}\right) = (1-w)\left(\frac{\Delta H - [H_s - H_{min}]}{\Delta H}\right)$$

Assuming that our target H_s is above the minimum enthalpy in the pool by an absolute value of E_s eV/atom, then the enthalpy's contribution to the fitness value can be expressed as:

$$f_H = (1-w)\left(\frac{\Delta H - E_S}{\Delta H}\right) = (1-w)\left(1.0 - \frac{E_S}{\Delta H}\right)$$

A similar result can be obtained for the PXRD objective's contribution to the fitness. Considering that $S_{min} = 0$ and $S_{max} = 1$, S_s is already the absolute value above the minimum of the objective. Hence:

$$f_S = w\left(\frac{S_{max} - S_s}{S_{max} - S_{min}}\right) = w\left(\frac{\Delta S - S_s}{\Delta S}\right) = w(1.0 - S_s)$$

Combining these results, the structure's fitness can be expressed as:

$$f_s = \left(1.0 - \frac{E_S}{\Delta H}\right) - w\left(S_s - \frac{E_S}{\Delta H}\right)$$

Given the range of enthalpies explored in the search, similarity index, and enthalpy of a candidate structure, one can use the above expression to estimate its fitness as a function of the PXRD objective's weight.

As an example, in a typical search for the TiO_2 polymorphs, the enthalpies relative to Anatase [meV/atom] and dissimilarity index for Anatase, Brookite, and Rutile were computed to be:

Structure	E_s	S_s
Anatase	0.0	0.31
Brookite	13.5	0.01
Rutile	26.7	0.91

Then, considering that in this search we found ΔH = 12.87 eV/atom and using $\Delta S =$ 1, the above formula results in the following values for the fitness of these phases as a function of the optimization weight as reported in Figure 3 in the main text.

weight	**Anatase**	**Brookite**	**Rutile**
0	1	0.997	0.994
0.1	0.9637	0.9969	0.9061
0.2	0.9274	0.9968	0.8182
0.3	0.8911	0.9967	0.7304
0.4	0.8548	0.9967	0.6425
0.5	0.8185	0.9966	0.5546
0.6	0.7822	0.9965	0.4667

0.7	0.7459	0.9964	0.3789
0.8	0.7096	0.9963	0.291
0.9	0.6733	0.9962	0.2031
1	0.637	0.9961	0.1152

It should be noted that the fitness values represent the suitability of the candidate structures in terms of the considered objectives in the global optimization. XtalOpt employs the calculated fitness values, along with the user-specified parent pool size to determine a structure's selection probability (to be a parent for the next generation), as described fully in Reference 1.

S2. Na-*hP4*

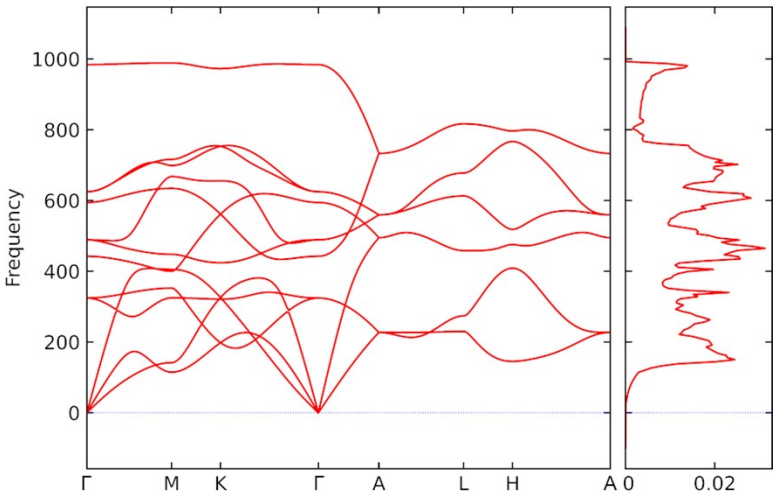

Figure S1. Phonon spectrum of Na-*hP4** calculated with VASP, version 6.4.2.[2,3], using the PBE[4,5] exchange-correlation funcitonal and a plane-wave basis set with an energy cutoff of 1200 eV. The k-point mesh generated using the Γ-centered Monkhorst−Pack scheme,[6] was such that the number of divisions along each reciprocal lattice vector was selected so that the product of this number with the real lattice constant was greater than or equal to 60 Å. The Tetrahedron method was adopted[7].

S3 New Predicted Phases of $CaCO_3$

A new *Pca2₁* structure was generated by our PXRD assisted structure search. The interesting aspect of this structure is the small peak at ~29° in 2*θ* (Figure S2), which was consistently observed in multiple experiments[8–11] but often assigned to calcite impurities.[9] However, the evident peak at ~35° would rule out *Pca2₁* as a possible candidate, since it is not observed experimentally. An additional, relatively stable, *P2₁/c* phase was predicted as well, which turns out to be a distortion of Meyer's *Pnma* structure.

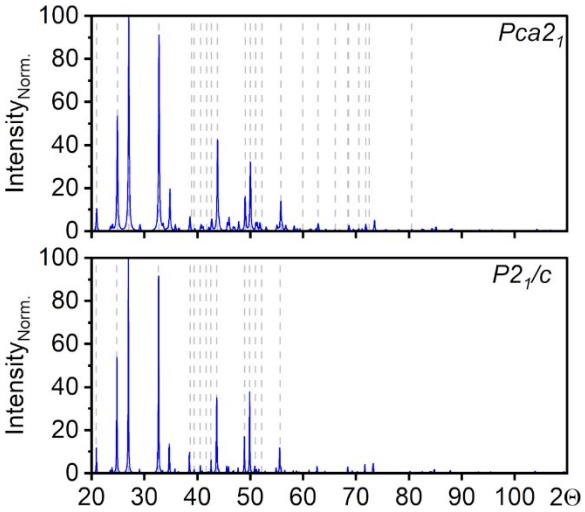

Figure S2. Simulated PXRD pattern for the predicted *Pca2₁* (top) and *P2₁/c* (bottom) phases (Cu *K*α radiation), after the VC-GPWDF refinement over the experimental diffraction data (dashed grey lines) of Le Bail's[9] (top) and DuPont's[8] (bottom) data.

S4 Experimental PXRD Patterns

List of the experimental PXRD pattern used in the present work.

Na-hP4[12]

2θ	***I***
37.0	40
43.3	100
62.9	80
63.4	80
75.1	10
79.2	10

Vaterite Le Bail[9]

2θ	***I***
20.953	19.6
24.883	62.1
27.05	100.0
32.753	92.6
38.871	6.2
39.381	4.9
40.672	4.8
41.774	2.9
42.673	11.2
43.822	51.3
49.049	18.3
50.023	47.9

51.039	4.1
52.204	1.3
55.767	16.1
59.898	2.7
62.842	3.6
66.111	1.0
68.609	3.8
68.455	1.5
70.533	1.3
71.840	5.9
72.496	7.2
80.532	1.3

Vaterite Dupont[8]

2θ	***I***
20.866	20
24.775	43
26.940	81
32.646	100
38.669	6
39.373	1
40.555	6
41.640	2
42.565	17
43.697	64
48.923	25
49.906	84
50.924	5
52.165	2
55.644	22

S5 Script

Here we report the script used to generate the similarity index value used in the *multi-objective* search. It first generate the input file (pxrd.cri) for critic2,[13] which is then run taking the list of peaks listed in pxrd.peaks. The output (pxrd.cro), is then parsed to extract the similarity index and print it in pxrd.dat, which will be read by XtalOpt.

```
###
echo "
TRICK GAUCOMP CONTCAR pxrd.peaks global maxfeval 5000
" > pxrd.cri

critic2 pxrd.cri pxrd.cro

grep 'DIFF' pxrd.cro | awk '{printf "%.8f\n", $4}' | tail -1  > pxrd.dat
###
```

S5 References

1 S. Hajinazar and E. Zurek, 2024, arXiv:2405.02138.

2 G. Kresse and J. Hafner, *Phys. Rev. B*, 1994, **49**, 14251–14269.

3 G. Kresse and D. Joubert, *Phys. Rev. B - Condens. Matter Mater. Phys.*, 1999, **59**, 1758–1775.

4 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.

5 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.

6 H. J. Monkhorst and J. D. Pack, *Phys. Rev. B - Condens. Matter Mater. Phys.*, 1976, **13**, 5188–5192.

7 P. E. Blöchl, O. Jepsen and O. K. Andersen, *Phys. Rev. B*, 1994, **49**, 16223–16233.

8 L. Dupont, F. Portemer and M. Figlarz, *J. Mater. Chem.*, 1997, **7**, 797–800.

9 A. Le Bail, S. Ouhenia and D. Chateigner, *Powder Diffr.*, 2011, **26**, 16–21.

10 E. Mugnaioli, I. Andrusenko, T. Schüler, N. Loges, R. E. Dinnebier, M. Panthöfer, W. Tremel and U. Kolb, *Angew. Chemie - Int. Ed.*, 2012, **51**, 7041–7045.

11 X. San, J. Hu, M. Chen, H. Niu, P. J. M. Smeets, C. D. Malliakas, J. Deng, K. Koo, R. dos Reis, V. P. Dravid and X. Hu, *Nat. Commun.*, 2023, **14**, 7858.

12 D. N. Polsin, A. Lazicki, X. Gong, S. J. Burns, F. Coppari, L. E. Hansen, B. J. Henderson, M. F. Huff, M. I. McMahon, M. Millot, R. Paul, R. F. Smith, J. H. Eggert, G. W. Collins and J. R. Rygg, *Nat. Commun.*, 2022, **13**, 2534.

13 A. Otero-De-La-Roza, E. R. Johnson and V. Luaña, *Comput. Phys. Commun.*, 2014, **185**, 1007–1018.