

COMMUNICATION

GALLOP: Accelerated molecular crystal structure determination from powder diffraction dataMark J. Spillman,^{*a} and Kenneth Shankland^bReceived 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

A combined local and global optimisation approach to crystal structure determination from powder diffraction data (SDPD) is presented. Using graphics processing units (GPUs) to accelerate the underpinning calculations, the speed and power of this approach is demonstrated with the solutions of two challenging crystal structures. In both cases, the frequency with which solutions were obtained was improved by an order of magnitude relative to DASH, a well-established SDPD program. With complex crystal structures increasingly being generated in polycrystalline form, this approach is a valuable step-forward in structure determination capabilities.

Introduction

Global Optimisation (GO) algorithms have been widely employed as a means of crystal structure determination from powder X-ray diffraction data (SDPD),^{1–5} with many programs^{6–8} adopting the approach pioneered by *DASH*,^{9–11} which combined simulated annealing with an efficient figure of merit to assess the agreement between observed and calculated diffraction data. Steady developments in such programs have maintained the relevance of SDPD, despite increasing competition from microcrystal X-ray diffraction, crystal structure prediction and, most recently, electron diffraction. However, recent work^{12, 13} has shown that even with state-of-the-art software and hardware, when the number of degrees of freedom to be determined by GO increases above ca. 25, the number of independent GO runs required to ensure a reasonable chance that one of them will return the correct crystal structure can become very large. For structures near the upper limits of SDPD's current capabilities, perhaps only 1 in 500 runs may return the correct structure. Whilst this low frequency of success can be offset by increasing the computational resources allocated to the problem via coarse-grained parallel processing^{14–17}, beyond a certain point, the computational cost becomes prohibitive. With particularly challenging crystal structures, investments of several CPU-months may be

required, with no guarantee of success. Previous work has shown that a “multi-start” local optimisation approach to SDPD is competitive with existing GO methods¹⁸. In a logical extension of that work, we present here a GPU-accelerated local optimisation and particle swarm approach (*GALLOP*) that significantly improves both the speed of, and frequency of success of, complex molecular crystal structure determination.

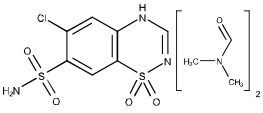
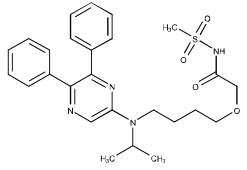
Background to the approach

GALLOP follows the broad strategy for SDPD laid out by *DASH*: powder diffraction intensities (and their associated covariances) are first extracted from the diffraction data by Pawley refinement, and the geometries of the molecular components to be located are described in a Z-matrix format. Solving the crystal structure is then a matter of determining the set of parameters that describe the position, orientation and conformation of these components within the unit cell, such that the agreement between calculated and extracted intensities (as judged by the well-established correlated integrated intensity χ^2 figure of merit)¹⁹ is good enough to indicate that the structure has been solved.

In *GALLOP*, as in *DASH*, maximising this agreement equates to minimising the χ^2 value.²⁰ This minimisation of χ^2 is carried out using a combination of (a) a local optimisation algorithm widely used in machine learning²¹ and (b) a particle swarm optimiser.²² First, a set (typically 1000) of putative crystal structures is initialised using randomly generated starting parameters. In this work, such a set is referred to as a *swarm*, whilst each individual crystal structure is referred to as a *particle*. The particles are optimised in parallel on a GPU for a fixed number of steps (typically 500) using the local optimisation algorithm. The optimised parameters and χ^2 values so obtained are then used as the input for a single step of particle swarm optimisation, which generates a new set of starting parameters for the local optimiser. The combination of the local optimisation steps followed by one particle swarm step make up a single *GALLOP* iteration. During SDPD, iterations continue until either a target value of χ^2 is achieved, a set number of iterations has been completed, or the user interrupts the program. Whilst *GALLOP* can run entirely on standard CPUs, it is designed and optimised

^a Nuclear Faculty, HMS Sultan, Gosport PO12 3BY, UK^b School of Pharmacy, University of Reading, Reading RG6 6AD, UK.Electronic Supplementary Information (ESI) available: *GALLOP* data files for selexipag and CT-DMF2. See DOI: 10.1039/x0xx00000x

Table 1 The crystal structures used in this study and their CSD refcodes

Compound	2D structure ²³	REFCODE	Ref
Chlorothiazide DMF (1:2) solvate [CT-DMF2]		NILSEH	24
Selexipag (form I)		VOHVIA	25

to run on GPUs and other hardware accelerators such as TPUs[†], which enable several thousand particles, and hence several independent swarms, to be optimised simultaneously.

Exemplars

Two published crystal structures, previously solved from PXRD data (Table 1), were chosen as examples of SDPD with reported low frequencies of success. The published structures were first validated by periodic dispersion-correct DFT (DFT-D) calculations, following the approach of van de Streek.^{26, 27} The PXRD data were then Pawley fitted using *DASH* (Table 2) and the resultant fit files used as input for *GALLOP*. Z-matrices for CT-DMF2 were generated from its Cambridge Structural Database²⁸ (CSD) entry, whilst those for selexipag were supplied by M. Husak (personal communication). *GALLOP* was run using a single cloud-based Nvidia Tesla V100 GPU with 16 GB of memory to perform the local optimisation. The GPU was accessed via Google Colaboratory,²⁹ a service that provides both free and paid-for access to GPU-equipped virtual machines.

Table 2 Crystallographic details of the structures used in this study

Parameter	Selexipag	CT-DMF2
Space group	<i>P2₁/c</i>	<i>P2₁/c</i>
<i>a</i> / Å	37.962	12.355
<i>b</i> / Å	6.110	8.560
<i>c</i> / Å	22.473	37.298
β / °	98.33	92.88
<i>V</i> / Å ³	5158	3940
<i>Z</i> '	2	2
λ / Å	0.39986	1.54056
$2\theta_{\max}$	10.800	36.993
<i>N</i> _{ref}	556	292
Resolution / Å	2.1245	2.4280
DoF _{position}	6	18
DoF _{orientation}	6	18
DoF _{torsion}	26	6
DoF _{total}	38	42
Published <i>DASH</i> frequency of success / %	0.5	1

*N*_{ref} = no. of reflections in Pawley fit; DoF = degrees of freedom. Frequency for CT-DMF2 comes from reference 12; that for selexipag from reference 25.

Each *GALLOP* iteration made use of 500 local optimisation steps, and a total of 100 *GALLOP* iterations were performed by each independent swarm. The size of each swarm was set to 1000 particles; hence each swarm performed 50 x 10⁶ χ^2 evaluations. Currently, as is the case in *DASH*, the χ^2 calculations involve only the non-hydrogen atoms of the structure. Runs, each consisting of *N* independent swarms, were carried out such that the GPU memory was close to fully utilised (ca. 14 GB in each case), which maximised the efficiency with which the χ^2 evaluations were executed. The maximum number of independent swarms that could be accommodated in the GPU memory varied with each structure (selexipag *N* = 9; CT-DMF2 *N* = 20) due to differences in the number of reflections, atoms and degrees of freedom. In order to obtain results from at least 100 independent swarms for each structure, the runs were repeated 12 times for selexipag and 5 times for CT-DMF2. All runs were allowed to go to completion and the best solutions found by each independent swarm were examined using the 'Crystal Packing Similarity' tool in Mercury.³⁰ Only those solutions that gave 15/15 molecules in common with the reference crystal structure (30% tolerances; H atoms ignored) were considered to be solved to a level of accuracy at which subsequent Rietveld refinement and/or DFT-D optimisation would be straightforward for someone familiar with refining molecular crystal structures against PXRD data. Henceforth, the term 'RMSD' refers to the 15 molecule root mean square deviation of 'Crystal Packing Similarity'. As reported elsewhere for CT-DMF2, there are three orientations of each -SO₂NH₂ group that give rise to similar χ^2 values, as the X-ray scattering power of -NH₂ is on a par with that of each O atom. Thus a CT-DMF2 solution might differ from NILSEH in the orientation of one or both of the -SO₂NH₂ groups but otherwise be in excellent agreement with NILSEH. In the current context, such a solution is considered to be correct.[‡] Thereafter, to facilitate a meaningful 'Crystal Packing Similarity' comparison with NILSEH, one or both -SO₂NH₂ groups in the *GALLOP* solution was rotated to match the orientation in NILSEH if required. Similar arguments apply to each -SO₂CH₃ group of selexipag.

Results and discussion

DFT-D validation of the published crystal structures returned RMSD values of 0.046 Å for CT-DMF2 and 0.115 Å for selexipag, the latter value in good agreement with the value of 0.129 Å published by Husak et al.²⁵ Assessment of *GALLOP* crystal structure solutions against the published crystal structures is, therefore, a valid measure of solution accuracy. As can be seen from Table 3, each crystal structure was solved quickly and accurately – the frequencies of success with which solutions are found are at least an order of magnitude higher than those reported for *DASH* (Table 2), even when the enhanced simulated annealing parameters in *DASH* are employed.¹² The low RMSD values for the solutions obtained demonstrate that *GALLOP* is returning solutions that lie very close to the published crystal structures, leaving little work to be done in the final refinement stage; see, for example, Figure 1, which shows the excellent agreement obtained for selexipag.

Table 3 Key indicators of *GALLOP* performance

Parameter	Selexipag	CT-DMF2
Swarm frequency of success / %	18	55
Run time / mins	116	104
Median solution time / mins	41	11
Best solution RMSD / Å	0.164	0.162
Average solution RMSD / Å	0.192	0.175
$P(\text{success} \text{run})$	0.84	> 0.99

Run times quoted are for one run of 9 and 20 swarms for selexipag and CT-DMF2 respectively. $P(\text{success} | \text{run})$ represents the probability of obtaining at least one correct solution in a single run comprising the number of swarms stated above.

In addition to the frequency with which the global minimum is located, it is also worth noting the speed with which solutions are returned; median solution times for these complex structures are of the order of tens of minutes, rather than tens of hours or days^{12, 25}. Figure 2 shows that most of the successful swarms required fewer than half of the allotted *GALLOP* iterations in order to locate the global minimum.

Conclusions and availability

The use of GPUs for processing diffraction data has been steadily increasing, with several examples of their use in powder diffraction.³¹⁻³⁴ In this work, we have combined a local optimiser, originally developed for machine learning applications and tuned to run on GPUs, with a particle swarm global optimiser to create a new and powerful method for solving molecular crystal structures from powder diffraction data. The two components of the *GALLOP* algorithm work in tandem; GPU-accelerated local optimisation provides rapid and efficient improvement of the χ^2 figure of merit for each particle,

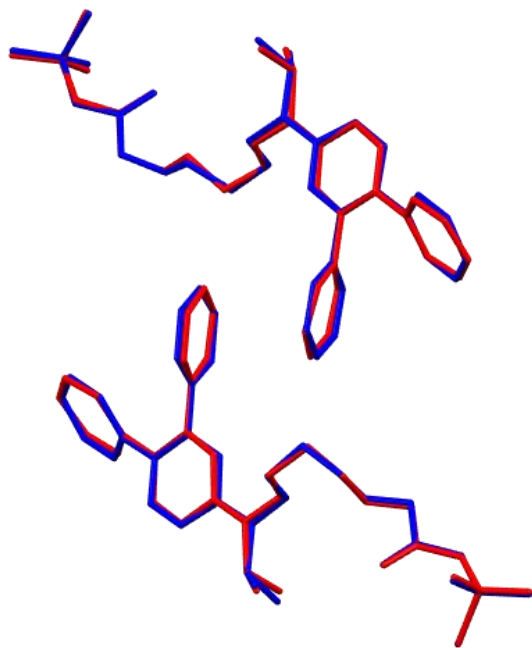


Fig. 1 The $Z' = 2$ asymmetric unit of selexipag, viewed down the b axis. The reference structure (CSD refcode VOHVIA) is shown in red, the *GALLOP* structure in blue. Hydrogen atoms have been omitted for clarity. RMSD = 0.164 Å.

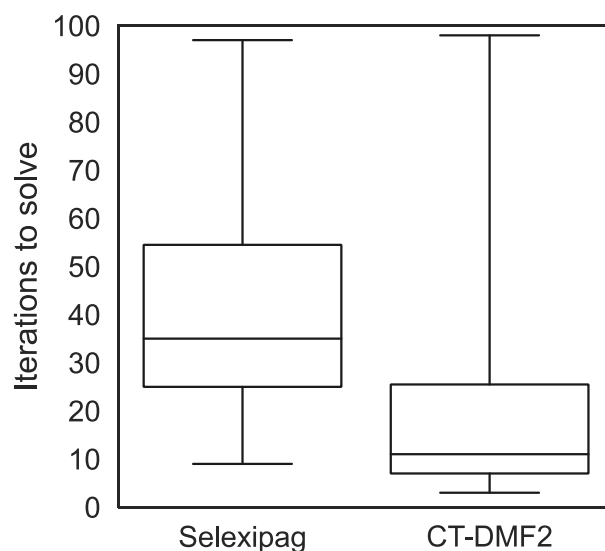


Fig. 2 A box plot of the number of iterations required to reach the global minimum for each structure.

whilst the particle swarm optimiser aggregates information from a large number of optimised particles, allowing it to initialise new starting points for the local optimiser in promising regions of the hypersurface.

The results presented here show that it considerably outperforms *DASH* in terms of speed and frequency of success for high-complexity problems that are increasingly typical of the polycrystalline materials now being generated, particularly by mechanochemistry. We envisage that the *GALLOP* approach will be of particular interest to crystallographers and material scientists working in high-throughput environments, where a rapid turnaround of results is of paramount importance. At present, our implementation of *GALLOP* is able to make use of diffraction data that has been Pawley fitted using *DASH* or *GSAS-II*.³⁵ Users interested in applying *GALLOP* to their own problems need not invest in dedicated GPU hardware in order to take advantage of this new approach; several cloud-based providers offer free or low-cost access to GPU-equipped virtual machines on which *GALLOP* can be rapidly deployed. To facilitate collaboration and further improvements to the approach, the full Python source code for *GALLOP* and instructions for its use are freely available at <https://www.github.com/mspillman/GALLOP>. The program can be operated through a versatile browser-based graphical user interface that provides control over the inputs, run setup parameters and allows visualisation of the best structure during each iteration, plus the ability to download the best CIFs during the runs. Alternatively, a Python API can also be used, which allows for seamless interaction with other libraries in the Python ecosystem.

Author Contributions

MJS: Conceptualization, investigation, methodology, project administration, software, validation, writing – original draft, review & editing; KS: methodology, validation, writing – original draft, review & editing

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We are grateful to Michal Husak (Uni. of Chem. & Tech., Prague) for selexipag data, and other files used for early GALLOP testing. Thanks are due to Elena Kabova (Uni. of Reading) for providing other diffraction data, Simon Mitchell (Uni. of Sussex) for helpful discussions regarding particle swarms, Jon Wright (ESRF) for useful correspondence regarding Pawley refinement software and Norman Shankland (CrystallografX Ltd) for many discussions relating to SDPD. We are grateful to the developers of the PyMatGen library, which made the development of GALLOP significantly easier. We are grateful to the UK Materials & Molecular Modelling Hub for computational resources, which is partially funded by EPSRC (EP/P020194/1 and EP/T022213/1), for DFT-D calculations.

Notes and references

† Tensor processing unit: Google's proprietary hardware for rapid training of neural networks. The latest TPU chips offer more on-board memory and performance compared to the latest Nvidia GPUs but are only publicly accessible via Google Cloud products.
‡ With a novel crystal structure determination, the actual orientation of a group such as $-SO_2NH_2$ can typically be deduced from one or more of the following: hydrogen-bonding considerations, Rietveld refinement and DFT-D crystal structure optimisation.

1. K. Shankland, W. I. F. David and T. Csoka, *Z. Kristallogr.*, 1997, **212**, 550-552.
2. O. Vallcorba, J. Rius, C. Frontera and C. Miravittles, *J. Appl. Crystallogr.*, 2012, **45**, 1270-1277.
3. Z. J. Feng, C. Dong, R. R. Jia, X. Di Deng, S. X. Cao and J. C. Zhang, *J. Appl. Crystallogr.*, 2009, **42**, 1189-1193.
4. G. W. Turner, E. Tedesco, K. D. M. Harris, R. L. Johnston and B. M. Kariuki, *Chem. Phys. Lett.*, 2000, **321**, 183-190.
5. W. I. F. David and K. Shankland, *Acta Crystallogr. Sect. A: Foundations and Advances*, 2008, **64**, 52-64.
6. V. Favre-Nicolin and R. Cerny, *J. Appl. Crystallogr.*, 2002, **35**, 734-743.
7. S. Pagola and P. W. Stephens, *J. Appl. Crystallogr.*, 2010, **43**, 370-376.
8. G. E. Engel, S. Wilke, O. Konig, K. D. M. Harris and F. J. J. Leusen, *J. Appl. Crystallogr.*, 1999, **32**, 1169-1179.
9. W. I. F. David, K. Shankland, J. van de Streek, E. Pidcock, W. D. S. Motherwell and J. C. Cole, *J. Appl. Crystallogr.*, 2006, **39**, 910-915.
10. A. J. Florence, N. Shankland, K. Shankland, W. I. F. David, E. Pidcock, X. L. Xu, A. Johnston, A. R. Kennedy, P. J. Cox, J. S. O. Evans, G. Steele, S. D. Cosgrove and C. S. Frampton, *J. Appl. Crystallogr.*, 2005, **38**, 249-259.
11. K. Shankland, L. McBride, W. I. F. David, N. Shankland and G. Steele, *J. Appl. Crystallogr.*, 2002, **35**, 443-454.
12. E. A. Kabova, J. C. Cole, O. Korb, M. Lopez-Ibanez, A. C. Williams and K. Shankland, *J. Appl. Crystallogr.*, 2017, **50**, 1411-1420.
13. E. A. Kabova, J. C. Cole, O. Korb, A. C. Williams and K. Shankland, *J. Appl. Crystallogr.*, 2017, **50**, 1421-1427.
14. T. A. N. Griffin, K. Shankland, J. V. van de Streek and J. Cole, *J. Appl. Crystallogr.*, 2009, **42**, 356-359.
15. T. A. N. Griffin, K. Shankland, J. V. van de Streek and J. Cole, *J. Appl. Crystallogr.*, 2009, **42**, 360-361.
16. M. J. Spillman, K. Shankland, A. C. Williams and J. C. Cole, *J. Appl. Crystallogr.*, 2015, **48**, 2033-2039.
17. J. Rohlicek, M. Husak and B. Kratochvil, *Acta Crystallogr. Sect. A: Foundations and Advances*, 2007, **63**, S242-S242.
18. K. Shankland, A. J. Markvardsen, C. Rowlatt, N. Shankland and W. I. F. David, *J. Appl. Crystallogr.*, 2010, **43**, 401-406.
19. W. I. F. David, *J. Appl. Crystallogr.*, 2004, **37**, 621-628.
20. W. I. F. David, K. Shankland and N. Shankland, *Chem. Commun.*, 1998, 931-932.
21. D. P. Kingma and J. Ba, *Computing Research Repository (CoRR)*, 2015, [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
22. J. Kennedy and R. Eberhart, in *Proc. of IEEE International Conference on Neural Networks*, 1995, vol. 4, pp. 1942-1948.
23. ChemAxon, MarvinSketch version 18.1, (<https://www.chemaxon.com>).
24. P. Fernandes, K. Shankland, A. J. Florence, N. Shankland and A. Johnston, *J. Pharm. Sci.*, 2007, **96**, 1192-1202.
25. M. Husak, A. Jegorov, J. Czernek, J. Rohlicek, S. Zizkova, P. Vraspir, P. Kolesa, A. Fitch and J. Brus, *Crystal Growth & Design*, 2019, **19**, 4625-4631.
26. J. van de Streek and M. A. Neumann, *Acta Crystallogr. Sect. B: Struct. Sci.*, 2014, **70**, 1020-1032.
27. J. van de Streek and M. A. Neumann, *Acta Crystallogr. Sect. B: Struct. Sci.*, 2010, **66**, 544-558.
28. C. R. Groom, I. J. Bruno, M. P. Lightfoot and S. C. Ward, *Acta Crystallogr. Sect. B: Struct. Sci.*, 2016, **72**, 171-179.
29. E. Bisong, in *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, Apress, Berkeley, CA, 2019, pp. 59-64.
30. C. F. Macrae, I. Sovago, S. J. Cottrell, P. T. A. Galek, P. McCabe, E. Pidcock, M. Platings, G. P. Shields, J. S. Stevens, M. Towler and P. A. Wood, *J. Appl. Crystallogr.*, 2020, **53**, 226-235.
31. I. Simecek, J. Rohlicek, T. Zahradnický and D. Langr, *J. Appl. Crystallogr.*, 2015, **48**, 166-170.
32. V. S. Neverov, *SoftwareX*, 2017, **6**, 63-68.
33. I. Simecek, O. Marik and M. Jelinek, *Rom. J. Inf. Sci. Technol.*, 2015, **18**, 182-196.
34. L. Gelisio, C. L. A. Ricardo, M. Leoni and P. Scardi, *J. Appl. Crystallogr.*, 2010, **43**, 647-653.
35. B. H. Toby and R. B. Von Dreele, *J. Appl. Crystallogr.*, 2013, **46**, 544-549.