Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery

Latent Space Encoding of Molecular Crystal Structure

Project Report

Project Dates: 05/07/2021 - 30/09/2021

University of Southampton

Project Student: Wong King Ming Alex, University of Southampton

Supervised by: Professor Graeme Day, University of Southampton

Report Date: 30/09/2021

AI3SD-Intern-Series:Report-9_Wong

Latent Space Encoding of Molecular Crystal Structure
AI3SD-Intern-Series:Report-9_Wong
Report Date: 30/09/2021
DOI: 10.5258/SOTON/AI3SD0150
Published by University of Southampton

Network: Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery
This Network+ is EPSRC Funded under Grant No: EP/S000356/1

Principal Investigator: *Professor Jeremy Frey*
Co-Investigator: *Professor Mahesan Niranjan*
Network+ Coordinator: *Dr Samantha Kanza*

Contents

1 Project Details

Title	Latent Space Encoding of Molecular Crystal Structure
Project Reference	AI3SD-FundingCall3_010
Supervisor Institution	Unviersity of Southampton
Project Dates	05/07/2021 - 30/09/2021

2 Project Team

2.1 Project Student

Name and Title	Mr Wong King Ming Alex
Employer name / University Department Name	University of Southampton Computational Systems Chemistry
Work Email	kmw1g19@soton.ac.uk
Website Link	N/A -

2.2 Project Supervisor

Name and Title	Professor Graeme Day
Employer name / University Department Name	University of Southampton
Work Email	G.M.Day@soton.ac.uk
Website Link	https://www.southampton.ac.uk/chemistry/about/staff/gmd1a11.page

2.3 Researchers & Collaborators

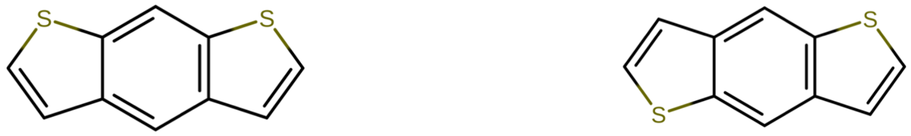
Thanks to Miss Rebecca J Clements, the Day Group and AI3SD for funding and training and Iridis HPC

3 Lay Summary

The computational methods for predicting crystal structures have been developed with one major application being the screening of molecules for polymorphism, which is the ability to crystallise in multiple crystal structures. A second application is the design of materials with targetted properties which depend strongly on solid state structure.

To help guiding the synthesis of candidate materials, atomic-scale modelling is used to calculate the stable polymorphs of a molecule and to predict its properties. Current methods for crystal structure prediction (CSP) sample the energy landscape for local minima by generating and lattice energy minimising trial crystal structures.

The energy surface describes the thermodynamic stability of the material as a function of chemical composition and the relative positions of atoms within the extended solid. In this project, we are taking the first steps to investigating the possibility of deep generative methods for crystal structure generation. This idea underpins the field of crystal structure prediction (CSP), whose aim is to enumerate and rank the possible crystal structures available in a molecule. As a first step, we have investigated the intrinsic dimensionality of the structural landscape of two small molecules, where ensembles of crystal structures have been generated by traditional methods of sampling the energy landscape. The results will inform future work, where the structure of neural networks will need to reflect the intrinsic dimensionality of the space of structures that need to be generated.



(a) syn-BDT

(b) anti-BDT

Figure 1: Chemical diagrams of the molecules studied.

4 Aims and Objectives

Learning to calculate descriptors for set of crystal structures and their storage in a suitable format for use in training of machine learning methods with the analysis of the resulting latent space and comparison to existing methods for measuring crystal structure similarity. Lastly, learn the background and packages available for training of variational autoencoder for compression of crystal structures into a latent space

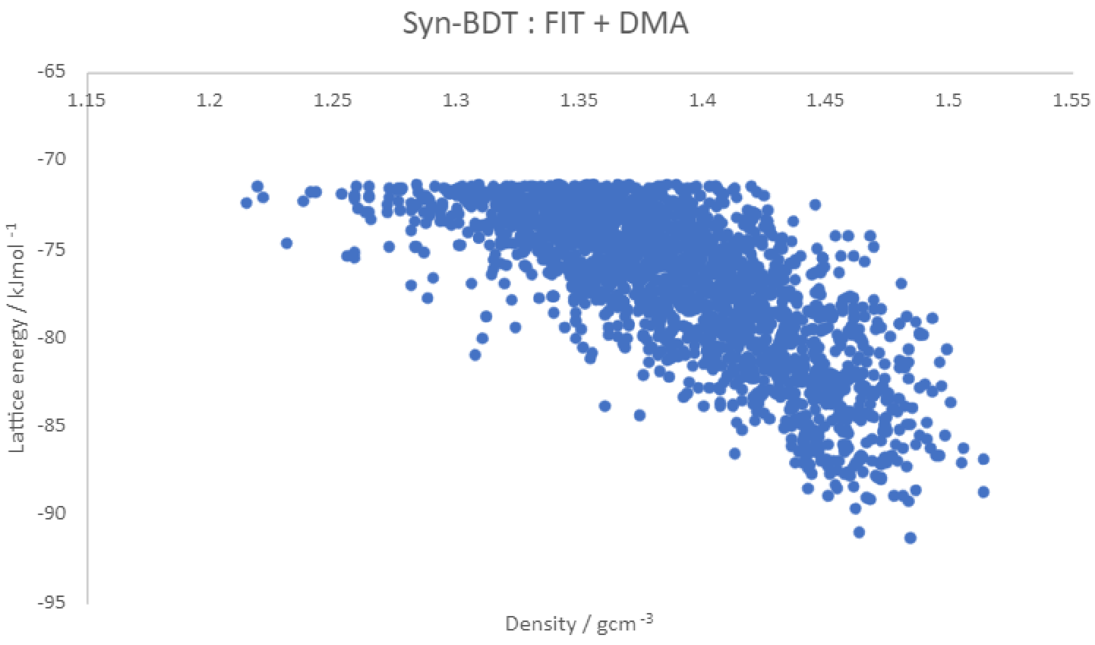
5 Methodology

5.1 Data Origin

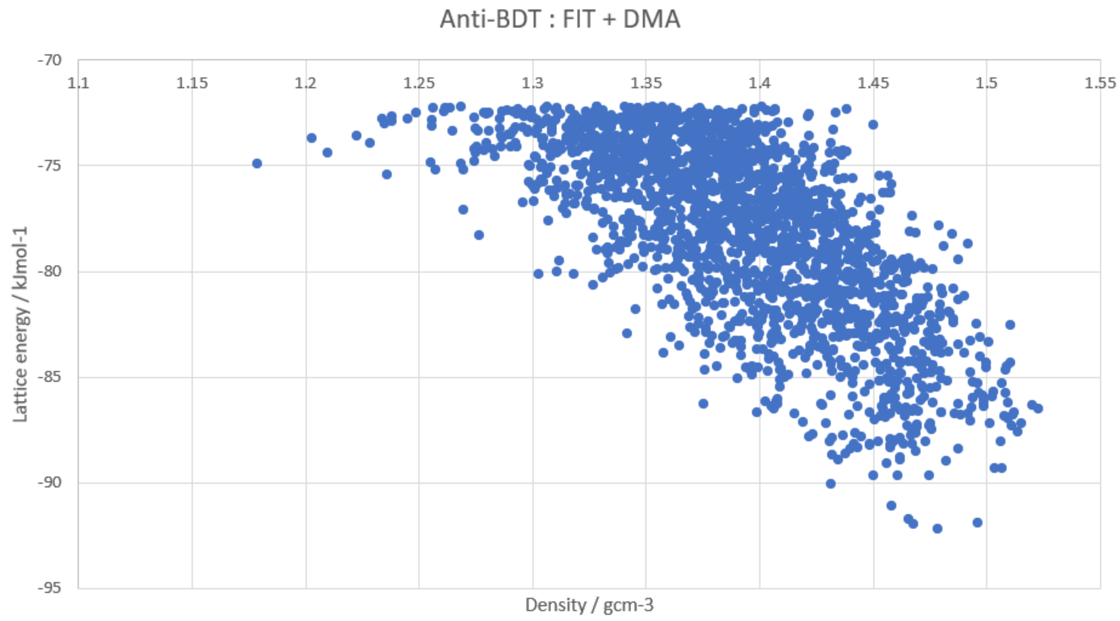
Crystal structure prediction was performed for two molecules, syn and anti-benzodithiphene(BDT), using the quasi-random sampling of the energy surface using the Global Lattice Energy Explorer software [1]. Within the molecules, there are types of symmetry element are preferred for molecular packing and over 95% of the structures belong to either the triclinic, monoclinic, or orthorhombic. The space group with the low-symmetry crystal systems results fewer symmetry operators which makes them simpler to understand than high-symmetry space groups, so we have chosen to generate the 10 most frequently observed space groups for organic molecules: P21/c, P212121, P-1, P21, Pbca, C2/c, Pna21, Cc, Pca21 and C2 in both syn and anti-BDT.

Energy window in molecule	Total number of eigenvalues
10 kJ/mol anti-BDT	315
20 kJ/mol anti-BDT	1301
10 kJ/mol syn-BDT	267
10 kJ/mol syn-BDT	1349

Table 1: Number of eigenvalues in different energy windows



(a) CSP data for syn-BDT in lowest 20 kJ/mol



(b) CSP data for anti-BDT in lowest 20 kJ/mol

5.2 Descriptors

In terms of the descriptors, we have used the atom centred symmetry functions proposed by Behler and Parrinello [2] to convert the atomistic structure of each predicted crystal structure into a suitable data type to carry on further analysis. We used the modified version developed for the ANI-1 force field[3] to generate the lattice energies which only takes into account of the intermolecular forces but not covalent bonds. It successfully describes the atomic environment with a series of radial and angular functions built up from the distribution of neighbouring atoms.

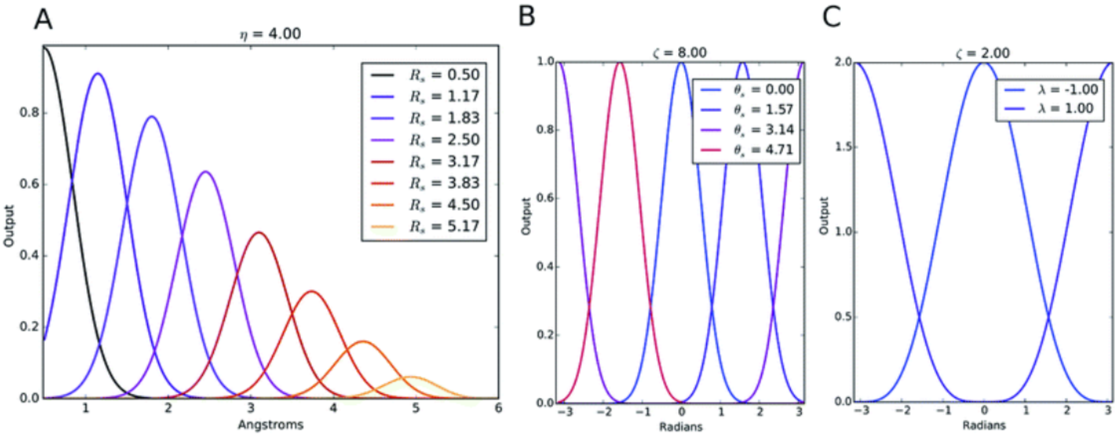


Figure 3: Examples of the symmetry functions with different parameter sets. (A) Radial symmetry functions, (B) modified angular symmetry functions and (C) the original Behler and Parrinello angular symmetry functions. These figures all depict the use of multiple shifting parameters for each function, while keeping the other parameters constant.

$$G_m^R = \sum_{j\neq i}^{all\ atoms} e^{-\eta(R_ij-R_s)^2}\ f_c\ (R_ij) \tag{1}$$

$$G_m^A = \ 2^{1-\zeta} \sum_{j,k\neq i}^{all\ atoms} (1+cos(\theta_{ijk}-\theta_s))^{\zeta} \times e^{-\eta((R_ij-R_s)/2-R_s)^2}\ f_c\ (R_ij) f_c\ (R_ik) \tag{2}$$

5.3 Principal Component Analysis

Pca was performed separately on the symmetry function description of crystal structures in each space group and on the entire set of low energy crystal structures. It has reduced the dimensionality to the projected data points in order to obtain the valuable data.

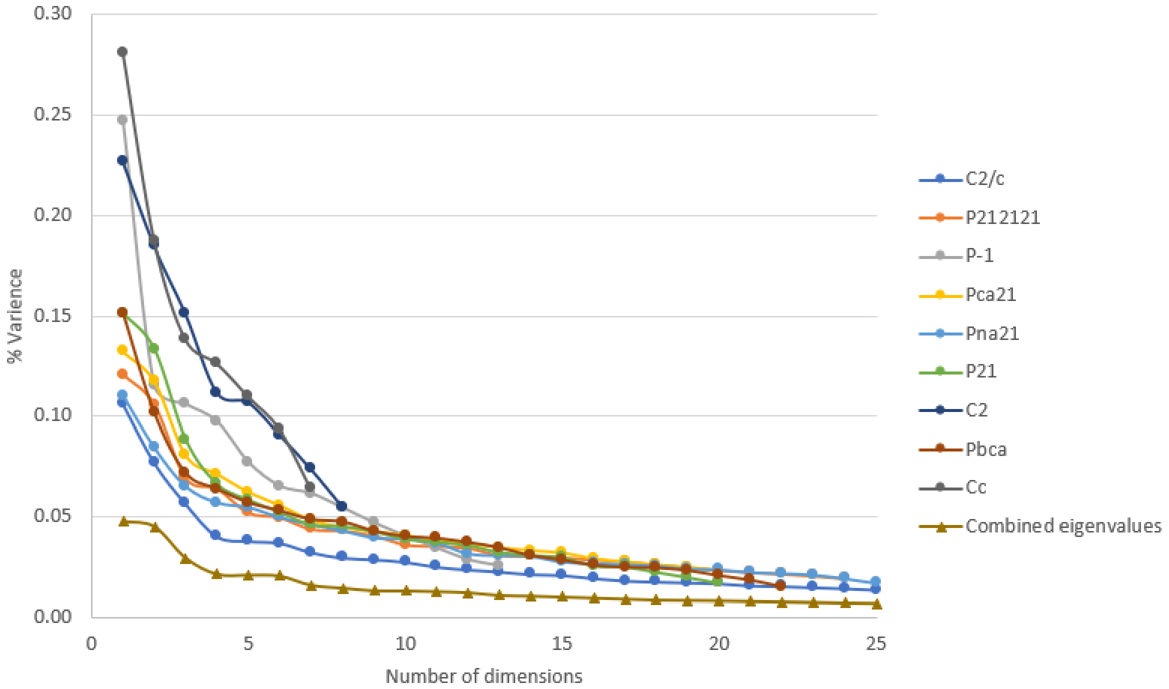
6 Results

6.1 Comparison between individual space groups and combined eigenvalues in lowest 10 and 20 kJ/mol for both molecules

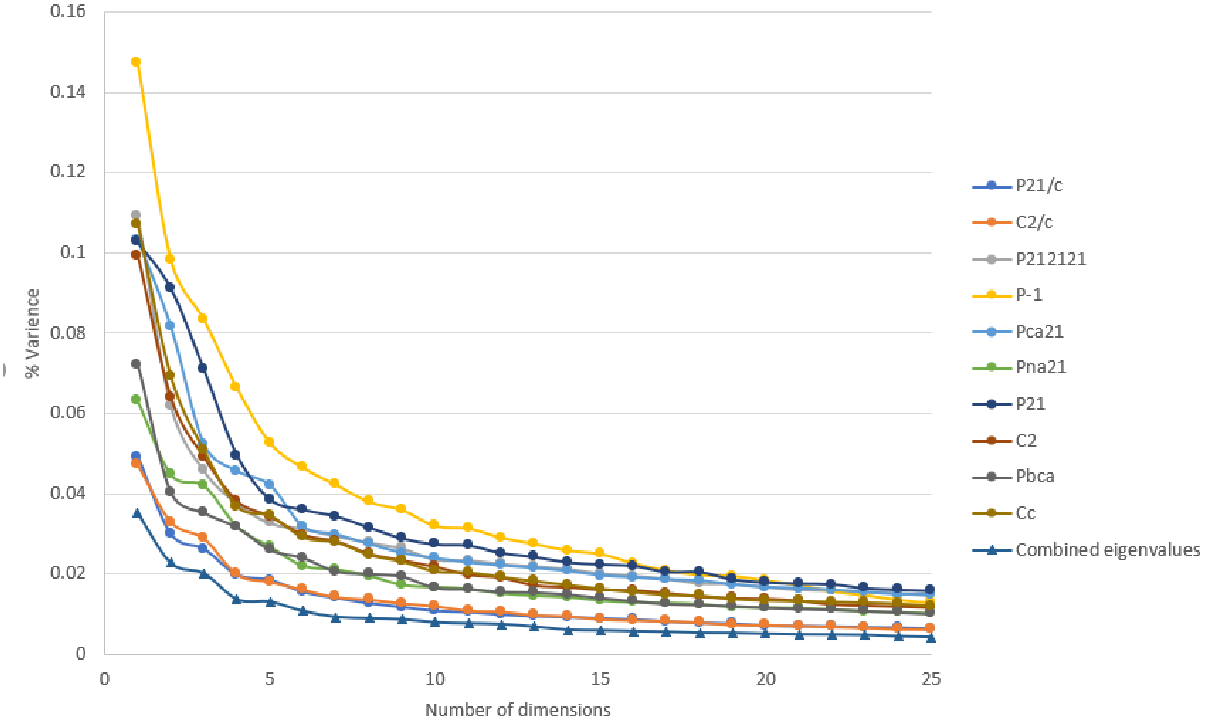
To compare between the actual degree of freedom and the dimensionality through the PCA graph, we can look at the elbow point which is the point of inflection in different space groups.

According to 4b: the point of inflection of C2 was more close to 5 and its actual degree of freedom is 9. The difference between the results can be explained as PCA assumes the regrees of freedom will correspond to linear combinations of the descriptor elements(atomic symmetry functions). However, the relationship is probably non-linear, so the principal components from PCA are not properly capturing the relationships between descriptors and degree of freedom.

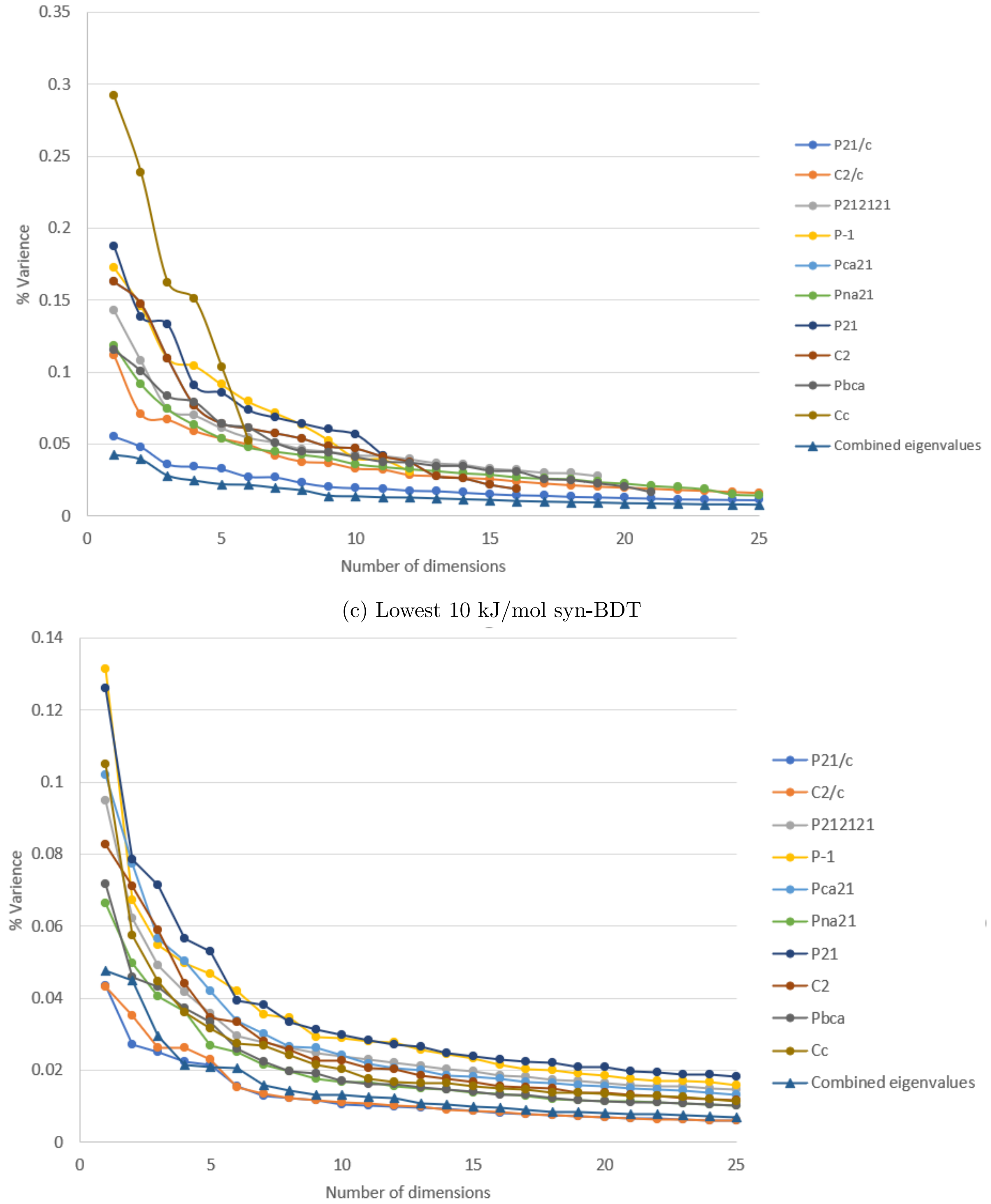
Furthermore, the descriptors have limit which is the description that they provide is too local as they only describe the close contacts around individual atoms in the crystal structure.



(a) Lowest 10 kJ/mol anti-BDT



(b) Lowest 20 kJ/mol anti-BDT



(c) Lowest 10 kJ/mol syn-BDT

(d) Lowest 20 kJ/mol syn-BDT

Figure 4: Plot of eigenvalues for pca fpr anti-BDT and syn-BDT in 10 most common space group under lowest 10 and 20 energy windows

Space group	Actual degree of freedom
P21/c	10
C2/c	10
P212121	9
P-1	12
Pca21	8
Pna21	8
P21	9
C2	9
Pbca	9
Cc	8

Table 2: Actual degree of freedom across 10 most common space group

Space Group	L10 syn-BDT	L20 syn-BDT	L10 anti-BDT	L20 anti-BDT
P21/c	58	200	59	186
C2/c	19	194	29	186
P212121	12	33	15	32
P-1	7	22	7	15
Pca21	9	31	12	29
Pna21	16	65	17	68
P21	7	20	12	24
C2	9	39	5	42
Pbca	13	60	13	66
Cc	3	45	4	39
Combined	260	534	144	511

Table 3: Total number of Eigenvalues adds up to 0.8 across 10 most common space group in the lowest 10 and 20 kJ/mol energy windows

6.2 Comparison between Pca1 and Pca2

6.2.1 Space group 14

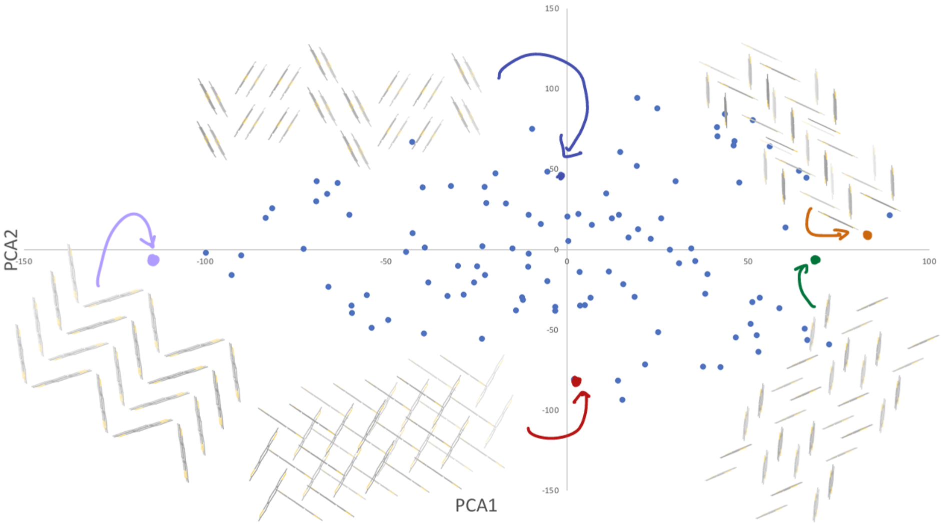


Figure 5: PCA plot for anti-BDT between principle component 1 and principle component 2 to show the landscape structure similarity

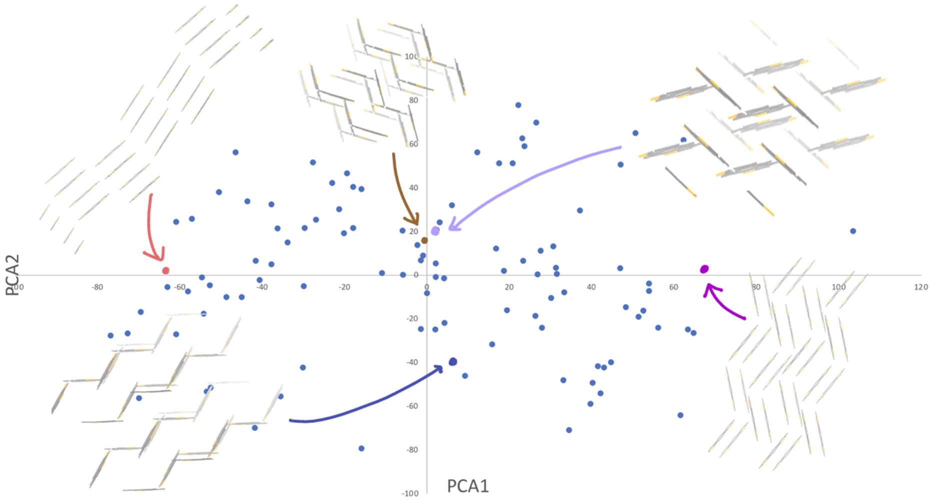


Figure 6: PCA plot for syn-BDT between principle component 1 and principle component 2 to show the landscape structure similarity

6.2.2 Space group 15

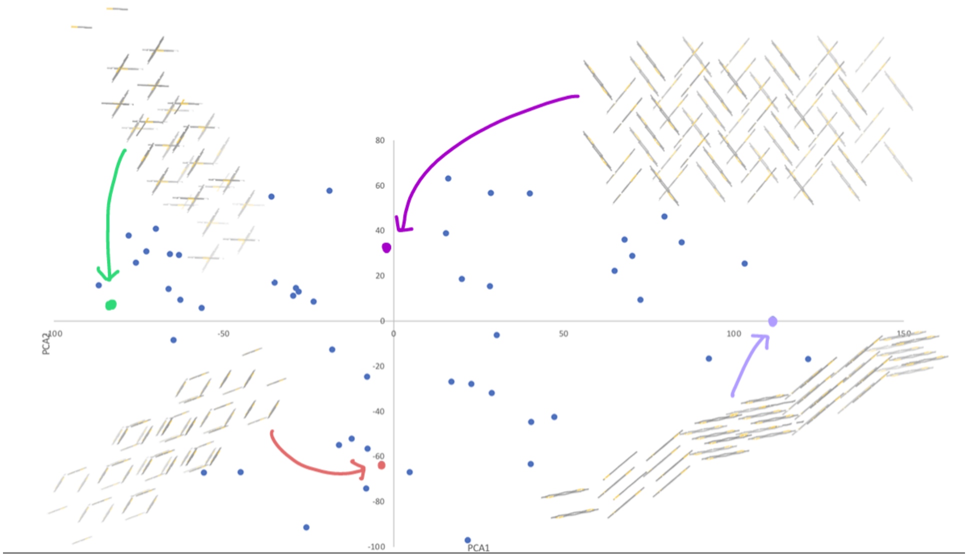


Figure 7: PCA plot for anti-BDT between principle component 1 and principle component 2 to show the landscape structure similarity

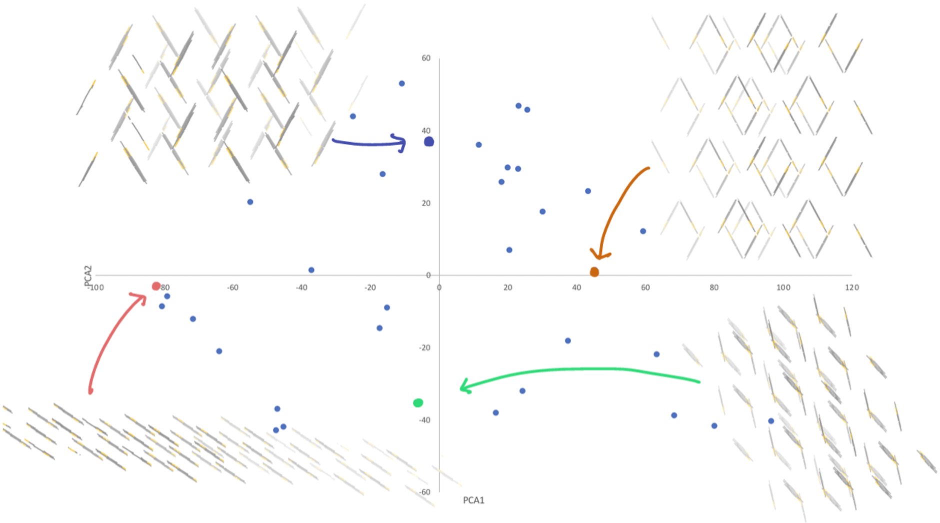


Figure 8: PCA plot for syn-BDT between principle component 1 and principle component 2 to show the landscape structure similarity

7 Conclusions & Future Work

An immediate area for future work is to continue the investigation of the intrinsic dimensionality of crystal structure landscapes. A more complete understanding is both of academic interest and is required prior to attempts to build generative models of crystal structures. Work here will involve investigations of local dimensionality of sets of related structures, of non-linear dimensionality reduction Iincluding neural networks, see belo) and of the dependence of the analysis on the crystal structure descriptor.

The next step is to build a neural network model to discover the correlations for the intrinsic dimensionality across 10 most common space group for syn-BDT and anti-BDT.

An encoder is the first part of the network which takes in the input and produces a lower dimensional encoding, coupled to a decoder, which reproduces the original input. The bottleneck is the lower dimensional hidden layer where the encoding is produced. It consists of lower number of nodes and gives the dimensionality of the encoding of the input, and a decoder recreates back the input and make sure there is not any loss of information.

We will investigate whether the latent space described by the low dimensionality encoding maintains structural similarity that we recognise as similar, in terms of the arrangement of molecules and the interaction between them, positioned nearby in the original landscape. If structural similarity in the original structures is maintained as closeness in the latent space, it should be feasible to sample from and explore the latent space of new structures. Thus, the next steps would be to investigate whether sampling of the latent space, followed by reconstruction via the decoder network, results is realistic crystal structures. In this context, this means structures that lie close to points from the complete crystal structure landscape.

Furthermore, by testing different network structures and dimensions of the latent space, we could learn the effective dimensionality of the structural space occupied by feasible crystal structures which has practical implications for the development of CSP method as the design of global optimisation algorithms depends on the dimensionality of the configuration space

8 Outputs, Data & Software Links

Data generated as part of the project includes crystal structure prediction output for two molecules. This will form part of a larger dataset that will be deposited with the University of Southampton and assigned a DOI on eprints.soton.ac.uk. This data will include further, related molecules and their predicted crystal structures.

There are no software outputs as yet from this work.

References

(1) Case, D. H.; Campbell, J. E.; Bygrave, P. J.; Day, G. M. *Journal of Chemical Theory and Computation* **2016**, *12*, 910–924.

(2) Behler, J.; Parrinello, M. *Physical Review Letters* **2007**, *98*, DOI: 10.1103/PhysRevLett.98.146401.

(3) Smith, J. S.; Isayev, O.; Roitberg, A. E. *Chemical Science* **2017**, *8*, 3192–3203.