Roles and opportunities for machine learning in organic molecular crystal structure prediction and its applications

Rebecca J. Clements[†], Joshua Dickman[†], Jay Johal[†], Jennie Martin [†], Joseph Glover and Graeme M. Day*

[†] These authors contributed equally.

* G.M.Day@soton.ac.uk

Keywords: machine learning; crystallographic structure; crystal; simulation; polymorphism

Abstract

The field of crystal structure prediction (CSP) has changed dramatically over the past decade and methods now exist that will strongly influence the way that new materials are discovered, in areas such as pharmaceutical materials and the discovery of new, functional molecular materials with targeted properties. Machine learning (ML) methods, which are being applied in many areas of chemistry, are starting to be explored for CSP. This overview will discuss the areas where ML is expected to have the greatest impact on CSP and its applications: improving the evaluation of energies; analyzing the landscapes of predicted structures and for the identification of promising molecules for a target property.

Introduction

The principal goal of CSP is to calculate the potential crystal structures of any given material from its chemical composition. Most CSP methods involve a search for local minima on the energy surface defined by the structural variables describing a crystal structure (unit cell dimensions, molecular positions and orientations), and ranking of these minima by their calculated energies. [1] The difficulty of CSP is highlighted by the occurrence of polymorphism, where a molecule can exist in one of several crystalline phases that may possess different physical and chemical properties. It has been shown that most polymorphs of organic molecules are separated by less than 2 kJ/mol in lattice energy, [2] thus highlighting the need for accurate computational methods to correctly predict the energy ranking of structures. The large number of possible structures involved in CSP for a given molecule means that accurate energies must be achieved at as low a computational cost as possible. Despite these challenges, considerable progress has been achieved in the past few decades and it is now possible to predict the crystal structure landscape of even quite complex systems, including multi-component crystals (salts, co-crystals and solvates), [3] [4] and pharmaceuticals that can adopt many molecular conformations. [5] [6] [7] Therefore, CSP is becoming more widely used, for example in the pharmaceutical industry to complement experimental polymorph screening, [8] and to prioritize candidates for synthesis for functional materials discovery. [9] Predicting the likely crystal structures of a molecule is crucial for predicting many properties that depend on the relative arrangement of

molecules in a crystal: a few examples include mechanical properties, porosity, the kinetics of dissolution and electron and hole mobilities in organic semiconductors.

The results of CSP can provide crucial information for experimental design. For example, predictions were able to show that the formation of caffeine-benzoic acid co-crystals is thermodynamically favourable, leading to the design of a seeding experiment which ultimately enabled the synthesis of the elusive crystal structure. [10] More recently, new polymorphs of iproniazid [11] and dalcetrapib [12] have been discovered from crystallizations under pressure which were guided from the results of CSP that discovered thermodynamically stable structures with high densities. CSP has also been applied in the area of porous molecular materials, helping to understand and modify the crystalline properties of porous organic cages [13] [14] and guiding the discovery of extrinsically porous molecular crystals for gas storage and molecular separations [9].

In recent years, ML methods have been employed across many areas of chemistry and are starting to be explored for CSP. [15] [16] [17] This includes their use for making highly accurate predictions of the relative energies of crystals, and in the analysis of CSP landscapes where the high dimensionality of the structural space can be simplified using ML algorithms.

This overview highlights the areas where ML methods are expected to have greatest impact on CSP methods and their applications. While many of the same challenges exist for other types of crystalline materials, we focus on organic molecular CSP.

Machine learning of the relative stabilities of putative structures

A requirement of CSP is the evaluation of accurate energies for large numbers of computer-generated crystal structures. A large-scale computational study has shown that free energy differences are < 2 kJ/mol in over half of observed polymorphs for small organic molecules, exceeding 6.4 kJ/mol in only 5 % of cases [2]. These differences are usually dominated by the lattice energy – the energy of the static arrangement of molecules in a crystal – with differences in entropy due to lattice vibrations usually being smaller than lattice energy differences [2]. Thus, temperature effects, including thermal expansion, are usually treated as a minor energetic contribution; it has been estimated that about 1 in 5 polymorph pairs swap their order of stability between 0 K and their melting point [18]. Therefore, the focus of CSP has largely been on obtaining accurate lattice energies.

Traditionally, the choice has been between force fields, which describe the interactions between atoms using physically-motivated functional forms, and more expensive quantum mechanical electronic structure methods (typically, solid state density functional theory, DFT). Many of the available methods have been benchmarked against the X23 benchmark set of measured sublimation enthalpies of a set of small organic molecules, [19] [20] [21], showing that the best force fields have mean absolute errors of 9 kJ/mol, while errors in the best DFT methods are about half this magnitude. Therefore, the fact that CSP is ever successful, given the small energy separations between polymorphs and predicted crystal structures, relies on cancellation of errors. For force fields, in particular, much of the errors are systematic, so do not affect the energetic *ranking* of predicted structures. Nevertheless, the increased accuracy of DFT methods is often necessary, particularly where polarization or charge-transfer interactions make important contributions to intermolecular interactions, or where changes in molecular geometry are significant. However, DFT energy calculations can be 10^3 – 10^5 times more computationally expensive than force fields [17], even for

small molecules. Thus, ML has been investigated as a means to achieve DFT-quality energies in CSP studies at more affordable computational costs.

The main requirement of an ML model is the ability to model the nonlinear relationship between energy and geometric descriptors of the crystal structure, which are usually represented by their local atomic environments, such as in the smooth overlap of atomic positions (SOAP) [22] and atom-centred symmetry function (ACSFs) [23] approaches. The use of a kernel, or covariance, matrix describing similarity between atomic configurations in Gaussian process regression (GPR) models is popular in chemical applications of ML, as are neural networks. While ML models have been thoroughly tested for accurately predicting the energies and properties of inorganic crystal structures, [24] and small organic molecules, [25] applications to organic molecular crystal structures are more challenging, in part due to the number of atoms involved. [26]

In the field of inorganic structure prediction, Tong et al. [27] demonstrated that a GPR model could be trained to high-level DFT calculations on-the-fly during a structure search for predicting boron clusters, saving an estimated 1-2 orders of magnitude in computational cost compared to full DFT calculations, and suggested that their work could apply to periodic systems. Deringer et al. [28] used GPR to train a potential for CSP of elemental phosphorus, initially training on DFT calculated energies of randomly generated structures, and refining the potential during the structure search, so that it could eventually identify complex structures whose size puts them out-of-reach of DFT calculations.

In the area of organic molecular CSP, several approaches have shown how applying ML allows for the use of quantum mechanics methods, more affordably than running such high-level calculations on all predicted crystal structures of a molecule. As a first demonstration in this area, Musil et al. [15] showed that GPR using the SOAP description of structural similarity could predict DFT lattice energies of pentacene CSP structures with less than 1 kJ/mol error, although the errors are higher for chemically more complex molecules. A Δ-ML approach, which learns the difference in energy between lower (force field) and higher (DFT) levels led to lower errors and more uniform performance for different molecules. The approach has been extended by Egorova et al., who developed a multilevel ML approach to correct the relative stabilities of predicted structures, [17] further reducing the required amount of the most computationally expensive, high-level energy calculations.

Other work includes training on a finite molecular cluster from the crystal structure at the target level using the many-body expansion of the lattice energy, instead of using periodic calculations. McDonagh et al. explored ML models for learning individual two-body (i.e. dimer) corrections to force field-calculated energies while keeping the long-range interactions at low level to reduce the cost. [29] This approach allows more accurate quantum chemistry models, such as correlated wavefunction methods, which are currently unaffordable for calculations on periodic structures. A similar approach was described by Wengert et al. [30], who included larger molecular clusters and reported that the use of ML reduces the computing time taken for 10 000 crystal structures of a small organic molecule from 30 million CPU hours to 80 000 CPU hours.

This dramatic reduction in cost means that ML models can be used to assess more computationally demanding free energy differences of crystal structures and include contributions from lattice vibrations, which are known to be important for polymorph relative stabilities [2] [18], and nuclear

quantum effects. Kapil and Engel [31] demonstrated such an approach, training a neural network model that was used for calculating free energy differences between crystal structures of benzene, glycine and succinic acid.

ML-enabled analysis of structural landscapes

An aim of CSP is to produce a set of all energetically feasible crystal packings of the studied system. These structure sets are rich with information on structure-function relationships, and are analyzed to identify which structures are most likely to be experimentally realized. Key barriers in such analysis include uncertainty in the predictions themselves and the high dimensionality of CSP landscapes, which creates a challenge for visualizing the distribution of predicted structures. While the dimensionality of the energy surface depends on the symmetry and number of molecules in the crystallographic unit cell, as an illustrative example, a structure with four rigid molecules in the unit cell is defined by up to 30 degrees of freedom: the unit cell dimensions, along with the orientations and positions of each molecule within the unit cell.

Reducing this dimensionality can help identify structure-function relationships. A common solution is to visualize the structures in only a few dimensions; for example, CSP results are often presented as a plot of relative energies against densities of the predicted structures. Chemical intuition can also be applied to also classify structures by the presence of certain interactions (e.g. hydrogen bonds) [32] or packing features. [33] The information loss and bias in analyzing sets of predicted structures may be minimized by applying dimensionality reduction methods to identify features that capture the greatest structural variation across the set of structures, and to form a lower dimensional representation in which similarity and dissimilarity of structures is preserved. As with ML for energies of crystal structures, dimensionality reduction relies on descriptors of each structure; again, such descriptors usually describe the local environment of atoms, such as ACSFs and SOAP. As a part of their work on learning energies, Egorova et al. [17] performed principal component analysis (PCA) of the sets of predicted crystal structures of a series of small molecules, each described using ACSFs. This work found that a small number of principal components (linear combinations of ACSFs) capture most of the variability across predicted structures, demonstrating that stable crystal structures tend to be found in a lower-dimensional manifold of the full dimensionality of the energy surface.

A related approach, useful for classification of structures, is clustering - an unsupervised ML method that optimizes the separation data points into clusters of similar points and defines each point by just one descriptor – its cluster index. As an example, Musil et al. [15] combined the non-linear dimensionality reduction technique, sketch-map [34], with clustering methods to produce 2-dimensional mappings of crystal structure landscapes of pentacene (**Error! Reference source not found.**a) and two azapentacenes proposed as promising organic semiconductors. The reduced mapping for pentacene showed clear groupings of structures which, when classified using hierarchical density-based clustering [35], reproduced results from heuristic classification of the structures, according to the known structural classes of the crystal structures of polyaromatic hydrocarbons (sheet-like, herringbone, etc). This demonstration that known structural classes can be identified algorithmically supports the application of these methods for analysis of CSP results. The approach was applied to analyze the combined set of predicted crystal structures of 28 molecules in a single clustered mapping [36] (**Error! Reference source not found.**b), which revealed relationships between molecular structure, preferred crystal packing and electron mobility. These findings demonstrate

potential for ML to accelerate structure classification and navigation of the combined space of molecular structures, crystal structure and materials properties. Thus, similar approaches have been applied in exploration for porous molecular crystals. [37] [38]

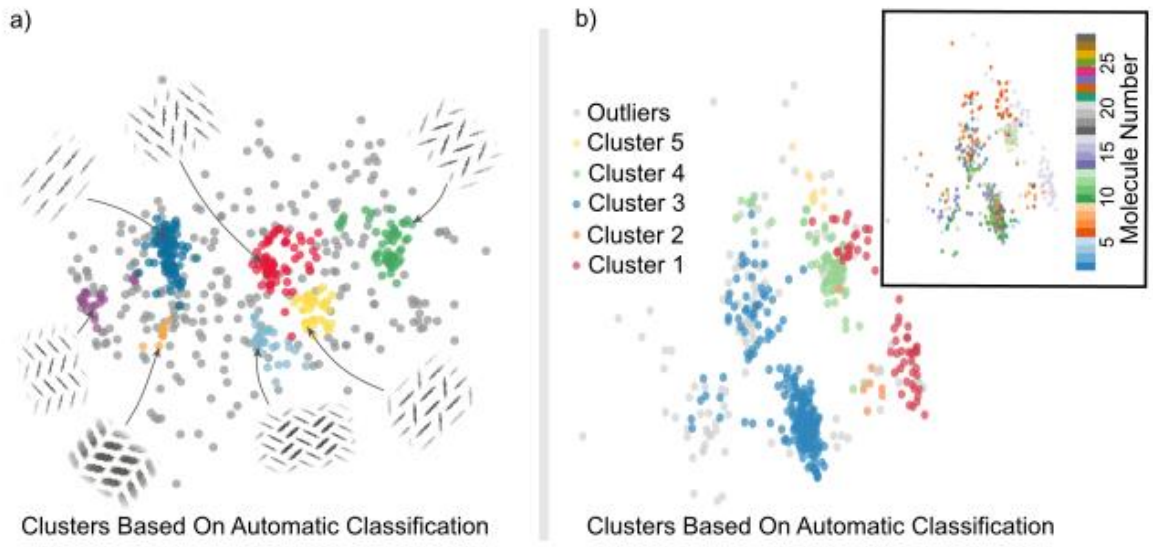

Figure 1: a) Reduced mapping of the CSP structures of pentacene, colored by cluster (inset structures show the crystal packing in representative structures from each cluster). Reproduced from [15] with permission from the Royal Society of Chemistry b) CSP structures of 28 pyrrole azaphenacenes combined on one map. The inset image in b) shows the same mapping colored by the molecule in the structure – indicating the ability of multiple similar molecules to adopt each packing type. Adapted with permission from [34]. Copyright 2018 American Chemical Society [36].

These studies also highlighted several challenges. For instance, not all structure sets can be effectively clustered, as was found for a second azapentacene in [15], and the mappings can be sensitive to choices in the representation; for example, the cutoff distance around each atom in SOAP influences the relative importance of inter- and intra-molecular similarity when comparing crystal structures of different molecules [36] Crystal structure representations can sometimes be constructed with a particular application in mind: Moosavi et al. [38] showed that representations capturing the topological features of pores within predicted crystal structures lead to mappings that group structures with similar calculated methane deliverable capacities.

The same need to choose ‘the right tool for the job’ applies to dimensionality reduction algorithms. Direct comparison of algorithms for dimensionality reduction has been underexplored in the context of analysing crystal structure landscapes. In their CSP study of pyrrole azaphenacenes, Yang et al. observed important differences in the distribution of points in the mappings produced using four

different dimensionality reduction algorithms [39], demonstrating that researcher expertise is still required in algorithm selection.

While the examples discussed so far have focussed on using unsupervised ML for visualization and the identification of structure-function relationships, a related method called the Generalized Convex Hull (GCH) attempts to identify which predicted crystal structures should be synthesizable. A conventional convex hull (CH) examines the energy of a material with respect to stoichiometry or some structural variable, such as molar volume. Only structures on the CH are considered thermodynamically stable. This analysis, however, relies on intuitively chosen features. The GCH algorithm [40] uses dimensionality reduction, *via* kernel PCA [41], to select data-driven coordinates for CH construction. In this way, the GCH identifies structures that are low in energy or extremal in geometry in some respect and could, therefore, be stabilizable. Uncertainty in the structural features and energies is also addressed. The GCH samples the hull points probabilistically across many iterations in which the data points are randomised within boundaries determined by a machine-learned estimation of their uncertainties. The approach was demonstrated for the identification of crystalline phases of hydrogen from CSP at high pressure and identified magnetically stabilizable phases of oxygen. From the perspective of applying CSP for the discovery of functional molecular materials, the GCH was demonstrated to identify predicted crystal structures of pentacene that could be stabilized by chemical modification.

Chemical space exploration

To be used as an effective method for discovering materials with targeted properties, CSP must be combined with methods for proposing promising molecules. Exhaustive searches of possible candidates are prohibitively expensive; as an example, for small drug-like molecules a calculated search space of up to 10^{60} possible molecules is estimated to exist. [42] [43] In contrast, the largest CSP studies to date have assessed groups of 10-30 molecules to identify the candidates with the best predicted properties. [9] [44] [36] Due to the gulf between the scale of CSP that is currently affordable and the size of chemical space, more targeted methods, such as data-driven techniques, are required to focus effort on the best candidate molecules. [45] Data-driven methods to generate molecules, which have mainly applied in the area of drug discovery, have also been demonstrated for functional materials discovery.

Molecular Representations Data driven chemical space exploration requires molecules to be represented in a computer readable manner. [22] Ideal representations are invertible, mapping only to specific molecular structures, and invariant to symmetry operations.

Molecular graphs are a common method of representing structures as bonds and atoms within a molecule can be represented as the edges and vertices of a graph. A popular method converts a 2D molecular graph into a string of ascii characters called simplified molecular-input line-entry system (SMILES) strings. These are invertible, but not unique - one molecular structure can map to multiple SMILES strings. [46] Suggested improvements, all canonical, include the unique, non-standardized canonical SMILES [47], InChi – a standardized string identifier [48] – and SELFIES, which always represent valid molecules [49].

High-throughput Virtual Screening (HTVS)

The conceptually simplest approach for finding high-performing molecules is HTVS, where molecular datasets are tested for a targeted property via computational predictions. HTVS is often performed using a funnelling method (**Error! Reference source not found.**a), to reduce cost whilst allowing properties to be determined more accurately for the later candidate pools. Quantitative structure-property relationship and ML models for property prediction have been applied as steps in the funnelling strategy. [45] [50] HTVS can use generative models or existing chemical databases, such as ZINC [51], the CSD [52] [53] and the Harvard Clean Energy Project [54], to build the initial populations of compounds to screen. Whilst CSP has not been applied in truly high-throughput studies, improvements in efficiency of the methods have made it possible to perform CSP, followed by property prediction for the sets of low-energy predicted crystal structures, for up to about 30 small molecules. [36] The properties of molecular materials often depend on intrinsic molecular properties, as well as properties that emerge due to the way that molecules are arranged in the solid. HTVS can take advantage of this: applying initial filtering based on calculated properties of isolated molecules and predicting crystal structures of only the molecules with the most promising properties. With further improvements in methods, coupled with increasingly available high-performance computing, CSP and property predictions should soon be possible on hundreds of candidate molecules on sufficiently short timescales to be useful in guiding experiments.

Generative Modelling

Generative Neural Network Models (GNNs): An advantage of sequence-based descriptors, such as SMILES, is that recurrent NNs can be trained to generate new descriptor sequences corresponding to new molecules. This approach has been demonstrated for molecular discovery of drug-like and other small organic molecules. [55] [56]

Other NN approaches involve training a generator to sample latent space for candidate molecules (**Error! Reference source not found.**b). Two main methods for this are generative adversarial networks (GANs) [57] and variational autoencoders (VAEs). [58]. VAEs were demonstrated for molecular design by Gómez-Bombarelli and co-workers [59], training the model on molecules from the ZINC and QM9 datasets. They demonstrated that the autoencoder can be jointly trained to predict molecules, along with their properties (such as drug-likeness and synthetic accessibility). Similarly, GANs have been demonstrated for the discovery of drug-like molecules and organic photovoltaic molecules. [60] [61] A workflow can be envisioned of performing CSP on the optimized generated molecules to predict their material properties.

As VAEs and GANs typically work on single molecules, CSP could have a similar role here as in HTVS: molecules are generated with optimized properties, followed by prediction of crystal structures and the resulting properties of the materials. Their direct application to the generation of crystal structures has been demonstrated for simple inorganic crystals, [62] [63] but is hindered by the challenge of representing three-dimensional crystal structures in a continuous latent space. We envisage further challenges for organic molecular crystals: because of the small energy differences between alternative crystal packings, minor changes in molecular structure can often lead to energy re-rankings of proposed crystal packings and, therefore, the experimental observation of completely different crystal structures, introducing discontinuities in the relationship between molecule and solid-state properties.

Evolutionary Algorithms (EAs): Issues with previously discussed methods include the computational cost of training the generator and the large amount of training data required to learn from. EAs are one alternative for generating new promising molecules (**Error! Reference source not found.**c). EAs are an optimization method inspired by evolution where members of an initial population undergo genetic operations and fitness evaluations to create successive generations, with the fitter candidates more likely to contribute to the next generation. In the area of molecular materials discovery, EAs have been applied to discovery of organic semiconductors, [44] and porous organic cages [64], where the properties to be optimized were charge carrier mobilities and persistent porosity, respectively. EAs can be efficient for exploration, requiring calculations on a small fraction of possible molecules during optimization to the best molecules. This opens the possibility for CSP within the fitness evaluation itself, if these methods can be made sufficiently fast for application to hundreds of molecules.

Chemical space networks (CSNs) can be used to monitor the progress of chemical space exploration campaigns [65], where edges of a graph denote 'morphing relationships' used to generate one molecule from another. One can trace a path from each generated molecule back to the initial species, where edges contain information on the operations used. CSNs are powerful tools for the visualization of molecular sets which can reveal potential design rules, as shown by Kunkel et al. [66].

Regardless of the approach used to generate molecules for assessment using CSP, the outcome is a set of molecules, each of which has an associated ensemble of predicted crystal structures. Thus, prioritization of molecules must consider the relevant properties (e.g. charge carrier mobility or porosity) of multiple crystal structures for each molecule. Some methods have been proposed, considering weighted averaged properties over low energy crystal structures to assess the likelihood that a molecule will lead to a crystal structure with the desired properties, [33] [36] along with measures of property variation amongst low energy crystal structures to assess uncertainty and risk. [44]

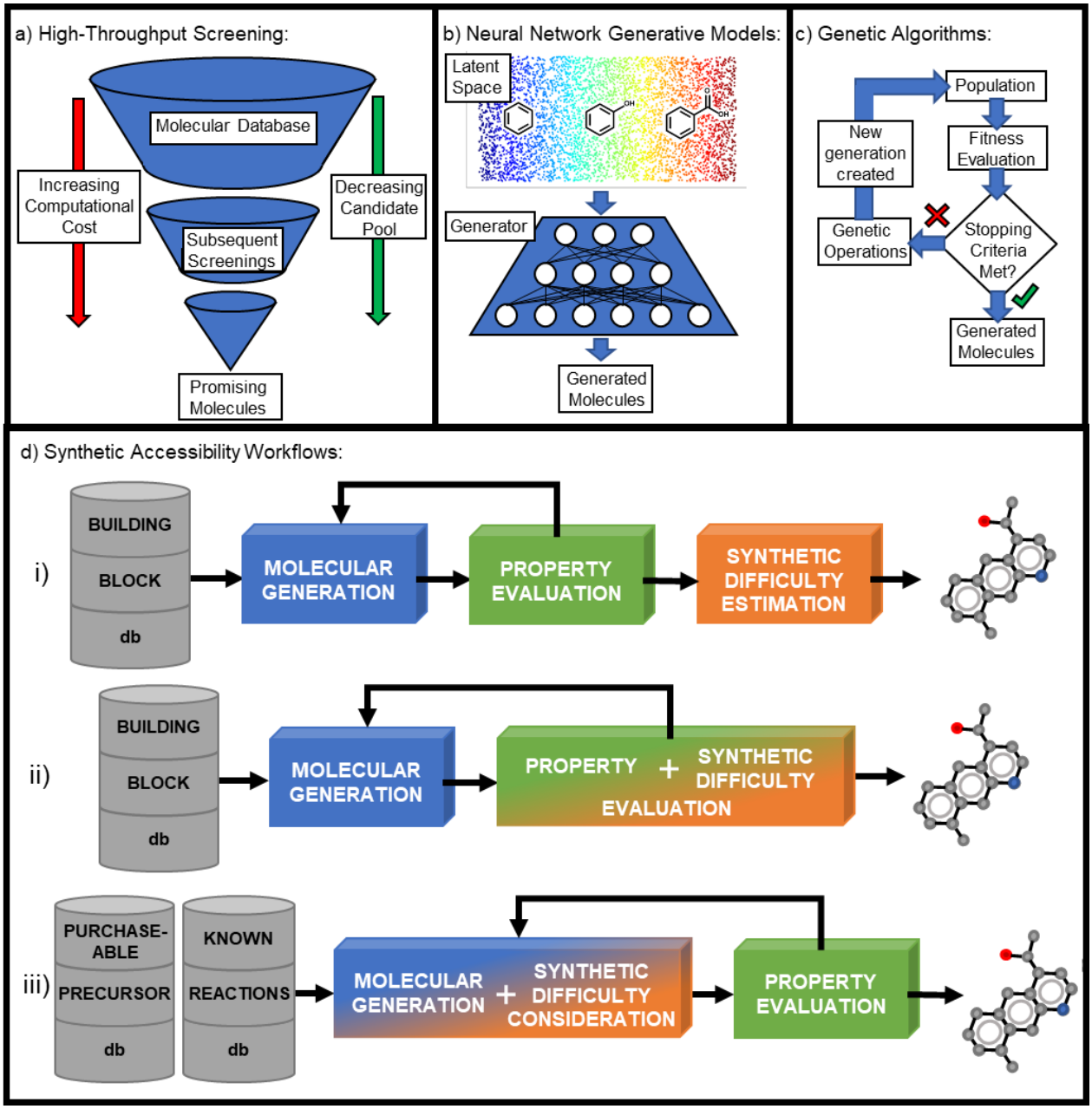

Figure 2 - Schematic representations of three approaches for exploring chemical space for new molecules with targeted properties: (a) high-throughput virtual screening; (b) GNNs (c) evolutionary algorithms and d) the inclusion of synthetic accessibility in chemical space exploration workflow. (Part d adapted with permission from *J. Chem. Inf. Model.* 2020, 60, 5714. Copyright 2020 American Chemical Society.)

Synthetic accessibility

The emerging methods for computational exploration of chemical space are exciting steps on the path to materials discovery, but experimental realization is vital; this requires feasible synthetic pathways to candidate species. When synthetic accessibility (SA) is not considered by a molecular generative model, candidates may be too challenging to synthesize in the lab. Approaches to biasing the generation of molecules towards synthetically accessible species have been discussed in detail by Gao and Coley [67].

Evaluation of SA with a post-hoc filter (**Error! Reference source not found.**Figure 2di) is the simplest approach: molecular generation is not biased and filtering is applied to the generated species after exploration is complete. Alternatively, heuristic biases can be used to guide molecular generation as part of the optimization function (Figure 2dii). However typical rapid scoring functions for SA (e.g. Ertl SAScore, [68] SYBA [69]) focus on molecular complexity; *structurally* complex candidates are not always *synthetically* complex, given a reasonable set of starting materials and reactions. Instead, Computer-aided synthesis planning (CASP) such as AiZynthFinder [70], involves prediction of full retrosynthetic pathways to a given molecule. While the process can be computationally expensive, it effectively mimics the process that chemists undertake. Information such as the number of synthetic steps or the availability and cost of precursors can be included in the fitness evaluation of candidates.

A third possibility is to directly influence the generation of molecules (Figure 2diii**Error! Reference source not found.**). Imposing explicit constraints on the building blocks and molecular transformations of a chemical space exploration campaign limits the proportion of the space that can be assessed, but should produce more synthetically accessible candidates. The Synthetically Accessible Virtual Inventory (SAVI) [71] works in this style, where 53 known single-step, two-reactant reactions were applied to 150,000 readily available precursors, generating a 1.75 billion dataset of which 1.09 billion compounds scored highly for synthetic accessibility.

As with the field of computer generation of molecules as a whole, synthetic accessibility prediction methods have focused on drug-like targets. These methods might need significant modification to account for the different types of molecular targets and the scale of synthesis required in materials discovery. Bennett et. al. [72] developed a binary classification model for the synthetic difficulty of porous organic cage precursors, learning the responses from experienced synthetic chemists to the question *"Can you make 1g of the compound in under 5 steps?".* Limiting the number of reaction steps works to reduce the overall yield loss during synthesis. While this limits access to species up to five synthetic steps away from available starting materials, their model was able to find precursors for promising porous materials with easier synthetic requirements.

Outlook

CSP methods can guide and accelerate materials discovery as research in this area shifts from an interesting academic challenge to applied studies [8]. While much prior progress has built on developments of traditional simulation methods – algorithms for exploring multi-dimensional energy landscapes, and models for calculating accurate lattice energies – data-driven, ML methods could lead to further exciting advances. These include acceleration of CSP through machine learning of accurate energies, methods for visualizing and interpreting the outcomes of large CSP datasets, and approaches to chemical space exploration to identify the best molecules to explore using CSP. Thus, the area will continue to benefit from close collaborations across chemistry, mathematics and computing science.

Acknowledgements

We thank the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement No 856405, and the Leverhulme Trust via the Leverhulme Research Centre for Functional Materials Design, for funding.

Conflicts of interest. On behalf of all authors, the corresponding author states that there is no conflict of interest

Data availability statement. No data was generated for this review.

References

[1] S. M. Woodley, G. M. Day, R. Catlow, *Phil. Trans. R. Soc. A.* (2020)*,* 378, 20190600.

[2] J. Nyman, G. M. Day, *CrystEngComm* (2015)*,* 17, 5154.

[3] D. E. Braun, Kahlenberg, V., U. J. Griesser, *Cryst.Growth Des.* (2017)*,* 17, 4347.

[4] H. C. S. Chan, J. Kendrick, M. A. Neumann, F. J. J. Leusen, *CrystEngComm* (2013)*,* 15, 3799.

[5] A. V. Kazantsev, P. G. Karamertzanis, C. S. Adjiman, Pantelides, C. C., S. L. Price, P. T. A. Galek, G. M. Day, A. J. Cruz-Cabeza, *Int. J. Pharm.*(2011)*,* 418, 168.

[6] R. M. Bhardwaj, L. S. Price, S. L. Price, S. M. Reutzel-Edens, G. J. Miller, I. D. H. Oswald, B. F. Johnston, A. J. Florence, *Cryst.Growth Des.* (2013)*,* 13, 1602.

[7] R. M. Bhardwaj, J. A. McMahon, J. Nyman, L. S. Price, S. Konar, I. D. H. Oswald, C. R. Pulham, S. L. Price, S. M. Reutzel-Edens, *J. Amer. Chem. Soc.* (2019)*,* 141, 13887.

[8] J. Nyman, S. M. Reutzel-Edens, *Faraday Discussions* (2018)*,* 211, 459.

[9] A. Pulido, L. Chen, T. Kaczorowski, D. Holden, M. A. Little, S. Y. Chong, B. J. Slater, D. P. McMahon, B. Bonillo, C. J. Stackhouse, A. Stephenson, C. M. Kane, R. Clowes, T. Hasell, A. I. Cooper, G. M. Day, *Nature* (2017)*,* 543, 657.

[10] D.-K. Bučar, G. M. Day, I. I. Halasz, G. G. Z. Zhang, J. R. G. Sander, D. G. Reid, L. R. MacGillivray, Duer, M. J., W. Jones, *Chemical Science* (2013)*,* 4, 4417.

[11] C. R. Taylor, M. T. Mulvee, D. S. Perenyi, M. R. Probert, G. M. Day, J. W. Steed, *J. Amer. Chem. Soc.* (2020)*,* 142, 16668.

[12] M. A. Neumann, J. van de Streek, F. P. A. Fabbiani, P. Hidber, O. Grassmann, *Nature Communications* (2015), 6, 7793.

[13] J. T. A. Jones, T. Hasell, X. Wu, J. Bacsa, K. E. Jelfs, M. Schmidtmann, S. Y. Chong, D. J. Adams, A. Trewin, F. Schiffman, F. Cora, B. Slater, A. Steiner, G. M. Day, A. I. Cooper, *Nature* (2011), 474, 367.

[14] A. G. Slater, P. S. Reiss, A. Pulido, M. A. Little, D. L. Holden, L. Chen, S. Y. Chong, B. M. Alston, R. Clowes, M. Haranczyk, M. E. Briggs, T. Hasell, G. M. Day, A. I. Cooper, *ACS Central Science* (2017), 3, 734.

[15] F. Musil, S. De, J. Yang, J. E. Campbell, G. M. Day, M. Ceriotti, *Chemical Science* (2018), 9, 1289.

[16] E. V. Podryabinkin, E. V. Tikhonov, A. V. Shapeev, A. R. Oganov, *Physical Review B* (2019), 99, 064114.

[17] O. Egorova, R. Hafizi, D. C. Woods, G. M. Day, *J. Phys. Chem. A* (2020), 124, 8065.

[18] J. Nyman, G. M. Day, *Phys. Chem. Chem. Phys.* (2016), 18, 31132.

[19] A. Otero-de-la-Roza, E. R. Johnson, *J. Chem. Phys.* (2012), 137, 054103.

[20] A. M. Reilly, A. Tkatchenko, *J. Chem. Phys.* (2013), 139, 024705.

[21] J. Nyman, O. S. Pundyke, G. M. Day, *Physical Chemistry Chemical Physics* (2016), 18, 15828.

[22] A. P. Bartók, R. Kondor, G. Csányi, *Phys. Rev. B* (2013), 87, 184115.

[23] J. Behler, *J. Chem. Phys.* (2011), 134, 074106.

[24] C. Chen, W. Ye, Y. Zuo, C. Zheng, S. P. Ong, *Chem. Mater.* (2019), 31, 3564.

[25] S. De, A. P. Bartók, G. Csányic, M. Ceriotti, *Phys. Chem. Chem. Phys.* (2016), 18, 13754.

[26] B. Olsthoorn, R. M. Geilhufe, S. S. Borysov, A. V. Balatsky, *Advanced Quantum Technologies* (2019), 2, 1900023.

[27] Q. Tong, L. Xue, J. Lv, Y. Wang, Y. Ma, *Faraday Discuss.* (2018), 211, 31.

[28] V. L. Deringer, D. M. Proserpio, G. Csányi, C. J. Pickard, *Faraday Discuss.* (2018), 211, 45.

[29] D. McDonagh, C.-K. Skylaris, G. M. Day, *J. Chem. Theory Comput.* (2019), 15, 2743.

[30] S. Wengert, G. Csányi, K. Reuter, J. T. Margraf, *Chemical Science* (2021), 12, 4536.

[31] V. Kapil, E. A. Engel, *Proc. Natl. Acad. Sci. U.S.A.* (2022), 119, e2111769119.

[32] D. E. Braun, H. Oberacher, K. Arnhard, M. Orlovac, U. J. Griessera, *CrystEngComm* (2016), 18, 4053.

[33] J. E. Campbell, J. Yang, G. M. Day, *J. Mater. Chem. C* (2017), 5, 7574.

[34] M. Ceriotti, G. A. Tribello, M. Parrinello, *Proc. Natl. Acad. Sci. U.S.A.* (2011), 108, 13023.

[35] R. J. G. B. Campello, D. Moulavi, A. Zimek, J. Sander, *ACM Transactions on Knowledge Discovery from Data* (2015), 10, 1.

[36] J. Yang, S. De, J. E. Campbell, S. Li, M. Ceriotti, G. M. Day, *Chem. Materials* (2018), 30, 4361.

[37] C. Zhao, L. Chen, Y. Che, Z. Pang, X. Wu, Y. Lu, H. Liu, G. M. Day, A. I. Cooper, *Nature Communications* (2021), 12, 817.

[38] S. M. Moosavi, H. Xu, L. Chen, A. I. Cooper, B. Smit, *Chemical Science* (2020), 11, 5423.

[39] J. Yang, N. Li, S. Li, *CrystEngComm* (2019), 21, 6173.

[40] A. Anelli, E. A. Engel, C. J. Pickard, M. Ceriotti, *Phys. Rev. Materials* (2018), 2, 103804.

[41] B. Schölkopf, A. Smola, K.-R. Müller, *Neural Computation* (1998), 10, 1299.

[42] J.-L. Reymond, R. van Deursen, L. C. Bluma, L. Ruddigkeit, *Med. Chem. Commun.* (2010), 1, 30.

[43] P. G. Polishchuk, T. I. Madzhidov, A. Varnek, *J. Comput.-Aided Mol. Des.* (2016), 27, 675.

[44] C. Cheng, G. M. Day, *Chemical Science* (2020), 11, 4922.

[45] Ö. H. Omar, M. del Cueto, T. Nematiaram, A. Troisi, *J. Mater. Chem. C* (2021), 9, 13557.

[46] D. Weininger, *J. Chem. Inf. Comput. Sci.* (1988), 28, 31.

[47] D. Weininger, A. Weininger, J. L. Weininger, *J. Chem. Inf. Comput. Sci.* (1989), 29, 97.

[48] S. R. Heller, A. McNaught, I. Pletnev, S. Stein, D. Tchekhovskoi, *J. Cheminformatics* (2015), 7, 23.

[49] M. Krenn, F. Florian Häse, A. Nigam, P. Friederich, A. Aspuru-Guzik, *Machine Learning: Science and Technology* (2020), 1, 045024.

[50] E. O. Pyzer-Knapp, C. Suh, R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, A. Aspuru-Guzik, *Annu. Rev. Mater. Res.* (2015), 45, 195.

[51] J. J. Irwin, B. K. Shoichet, *J. Chem. Inf. Model.* (2005), 45, 177.

[52] C. R. Groom, I. J. Bruno, M. Lightfoot, S. C. Ward, *Acta Cryst.* (2016), B72, 171.

[53] Ö. H. Omar, T. Nematiaram, A. Troisi, D. Padula, *Scientific Data* (2022), 9, 54.

[54] J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway, A. Aspuru-Guzik, *J. Phys. Chem. Lett.* (2011), 2, 2241.

[55] M. H. S. Segler, T. Kogej, C. Tyrchan, M. P. Waller, *ACS Cent. Sci.* (2018), 4, 120.

[56] K. Kim, S. Kang, J. Yoo, Y. Kwon, Y. Nam, D. Lee, I. Kim, Y.-S. Choi, Y. Jung, S. Kim, W.-J. Son, J. Son, H. S. Lee, S. Kim, J. Shin, S. Hwang, *npj Computational Materials* (2018), 4, 67.

[57] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, N. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, *arxiv:*1406.2661v1 (2014).

[58] D. P. Kingma, M. Welling, *arxiv:*1312.6114v10 (2014).

[59] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, A. Aspuru-Guzik, *ACS Cent. Sci*. (2018), 4, 268

[60] B. Sanchez-Lengeling, C. Outeiral, G. L. Guimaraes, A. Aspuru-Guzik, *chemRxiv:* https://doi.org/10.26434/chemrxiv.5309668.v3 (2017).

[61] N. De Cao and T. Kipf, *arxiv,* p. https://doi.org/10.48550/arXiv.1805.11973, 2018.

[62] A. Nouira, N. Sokolovska, J.-C. Crivello, *arXiv:*1810.11203 (2018).

[63] S. Kim, J. Noh, A. Aspuru-Guzik, Y. Jung, *ACS Cent. Sci.*(2020)*,* 6, 1412.

[64] E. Berardo, L. Turcani, M. Miklitz, K. E. Jelfs, *Chemical Science* (2018)*,* 9, 8513.

[65] C. Kunkel, J. T. Margraf, K. Chen, H. Oberhofer K. Reuter, *Nature Communications* (2021)*,* 12, 2422.

[66] C. Kunkel, C. Schober, H. Oberhofer, K. Reuter, *J. Mol. Model.* (2019)*,* 25, 87.

[67] W. Gao, C. W. Coley, *J. Chem. Inf. Model.* (2020)*,* 60, 5714.

[68] P. Ertl, A. Schuffenhauer, *J Cheminform* (2009)*,* 1, article number 8,.

[69] M. Voršilák, M. Kolář, I. Čmelo, D. Svozil, *J Cheminform* (2020)*,* 12, 35.

[70] S. Genheden, A. Thakkar, V. Chadimová, J.-L. Reymond, O. Engkvist, E. Bjerrum, *J. Cheminformatics* (2020)*,* 12, 70.

[71] H. Patel, W. Ihlenfeldt, P. Judson, Y. S. Moroz, Y. Pevzner, M. L. Peach, V. Delannée, N. I. Tarasova, M. C. Nicklaus, *Sci Data* (2020)*,* 7, 384.

[72] S. Bennett, F. T. Szczypiński, L. Turcani, M. E. G. R. L. Briggs, K. E. Jelfs, *J. Chem. Inf. Model.* (2021)*,* 61, 4342.