

Cite this: *Chem. Sci.*, 2021, 12, 830

Can we predict materials that can be synthesised?

Filip T. Szczypiński,  Steven Bennett  and Kim E. Jelfs *

The discovery of materials is an important element in the development of new technologies and abilities that can help humanity tackle many challenges. Materials discovery is frustratingly slow, with the large time and resource cost often providing only small gains in property performance. Furthermore, researchers are unwilling to take large risks that they will only know the outcome of months or years later. Computation is playing an increasing role in allowing rapid screening of large numbers of materials from vast search space to identify promising candidates for laboratory synthesis and testing. However, there is a problem, in that many materials computationally predicted to have encouraging properties cannot be readily realised in the lab. This minireview looks at how we can tackle the problem of confirming that hypothetical materials are synthetically realisable, through consideration of all the stages of the materials discovery process, from obtaining the components, reacting them to a material in the correct structure, through to processing into a desired form. In an ideal world, a material prediction would come with an associated 'recipe' for the successful laboratory preparation of the material. We discuss the opportunity to thus prevent wasted effort in experimental discovery programmes, including those using automation, to accelerate the discovery of novel materials.

Received 6th August 2020
Accepted 25th November 2020

DOI: 10.1039/d0sc04321d

rsc.li/chemical-science

1 Introduction

Advances in materials and technology have always accompanied human progress. However, the fast growth of industry is a double-edged sword that has created many new global challenges, such as resource scarcity, waste, and pollution. The discovery of new, cheap, and sustainable materials with tailored properties can help us address humanity's current and future

challenges. Materials can be simple systems made from single molecules or a small number of (inorganic) elements, or more complex systems processed into a variety of forms, including crystals, amorphous structures, thin films, and devices with multiple materials assembled together. Materials discovery is frustratingly slow, with frequent incremental improvements and much rarer leaps to new materials classes with ground-breaking properties.

There is considerable discussion of the potential for automation to revolutionise materials science so that new materials can enter the market sooner than the current 40 year timescale

Department of Chemistry, Imperial College London, Molecular Sciences Research Hub, White City Campus, Wood Lane, London, W12 0BZ, UK. E-mail: k.jelfs@imperial.ac.uk



Filip Szczypiński obtained his Master's degree in Natural Sciences in 2014 from the University of Cambridge, where he investigated the synthesis and host-guest binding inside large metal-organic cages. For his PhD, he studied synthetic information molecules under the supervision of Prof. Chris Hunter. Currently, Filip is working with Dr Kim Jelfs at Imperial College London on the

modelling of new porous organic molecular materials, where he combines his synthetic experience with computational tools to predict new materials that can be synthesised.



Steven Bennett obtained his Master's degree in Chemistry in 2018 from University College London (UCL). In October 2018, he joined the Jelfs Group at Imperial College London to work on the computational discovery of synthetically accessible functional materials, namely porous organic cages. This work is supported by a PhD studentship from the Leverhulme Trust via the Leverhulme Centre for

Functional Materials, which aims to develop his interest in sustainable, advanced material development.



from laboratory discovery to industrialisation.¹ The design and synthesis processes in materials discovery can be accelerated with autonomous workflows, such as the use of data repositories, automation, and parallelisation.² Recently, a mobile robotic chemist was used to autonomously explore a large chemical space in search for an improved hydrogen-production catalyst.³ Such an autonomous screening strategy was employed as many materials are impossible to design rationally due to their extremely high multi-scale complexity. Great efforts and resources are wasted on the synthesis of systems that do not yield materials with interesting properties. The chemical space of drug-like organic molecules alone, which are in theory potential material building blocks, is estimated to be between 10^{23} and 10^{60} possible compounds.^{4,5} Moreover, most materials are not built from organic molecules alone and for inorganic materials, the number of potential elemental compositions and stoichiometries is practically infinite, but many are not stable, or their preparation conditions are not known. Even with automation, it is not possible to adequately search the chemical space of materials, and many material syntheses or property characterisation techniques are not simply automatable due to complex multi-step protocols or specialist offline equipment being required.

In recent years, we are increasingly turning to computation to assist in accelerating material discovery across broad classes of materials and applications. The Holy Grail of materials prediction is inverse design (see Fig. 1), where appropriate materials and their components are designed based upon knowledge of only the desired material properties. The capability to achieve inverse design would be equivalent to the retrosynthetic analysis that can be carried out for molecular organic synthesis. The unpredictability of component assembly into the final material form means that we cannot reliably provide retrosynthetic pathways to optimal materials from first principles. Thus, rather than *design* materials, we must typically *screen* viable materials that have targeted properties. A computational screening process usually starts with large libraries of precursors, predicts how they will assemble into a material and finally calculates their properties using computational chemistry tools or data-driven models such as machine learning. The

significant challenge of the structure prediction stage means that this is often skipped, or assumptions are made based on commonly occurring structural motifs. The computed material properties of the candidates can be used to select the experimental target for laboratory synthesis and testing.

Computational screening is still more time- and resource-efficient than a synthetic screening. Machine learning is also beginning to influence and accelerate the structure–property prediction, but the lack of sufficient training data inhibits generation of predictive models, especially for organic materials.⁶ Molecules can be encoded for machine learning in numerous ways: from simple molecular formulae that do not convey any connectivity, through 2D chemical graphs, to full spatial coordinate descriptors at a single molecule level. Beyond representations of the material's components, it is often important to encode the broader environment of the solid-state packing and the development of representations in general is an active area of research. Which specific prediction task is being tackled will heavily influence the choice of representation and the correct selection can be critical to any potential prediction success.

A key question remains: *can screened or designed materials actually be synthesised, and how?* As Jansen and Schön argue, we are never designing materials but merely searching for thermodynamically viable minima on the potential energy landscape.⁷ It is far easier to predict materials with good properties than it is to predict materials with good properties that can be synthetically realised, rather than remaining “hypothetical”. Furthermore, even if synthetically realised in solution, many materials need to be processed into other forms, such as thin films and membranes. It is additionally challenging to predict whether the calculated properties are achieved in the material's final form.

This minireview aims to highlight the major challenges in the prediction of materials that can actually be synthesised in the laboratory. We try to follow an experimentalists' thought process – what precursors to use, are they available, how to combine them to form a material, what is the desired form – and discuss the recent computational advances that can guide or complement each of those stages.



Kim Jelfs is a Reader in Computational Materials Chemistry and Royal Society University Research Fellow at the Department of Chemistry at Imperial College London, UK. Kim carried out her PhD modelling zeolite crystal growth at UCL, before working as a post-doctoral researcher at the University of Liverpool with Prof. Andy Cooper FRS. Kim began her independent research at Imperial College in 2013 and her research

focuses upon the use of computer simulations to assist in the discovery of supramolecular materials.

2 Harnessing the power of the literature

Most researchers currently embark on a new material discovery project by exploring the primary form of scholarly communication: published scientific articles. A corresponding literature-based data extraction should also play an important role in automated and autonomous materials discovery. Albeit fragmented into multiple research items, most synthetic procedures are highly prescriptive, and hence in theory suitable for automated extraction using natural language processing methods.^{8,9} Extraction of chemical data and its automated association with the relevant chemical entities as well as details of the physical measurements (“metadata”) allow for easy creation of massive chemical databases.¹⁰ Generated databases



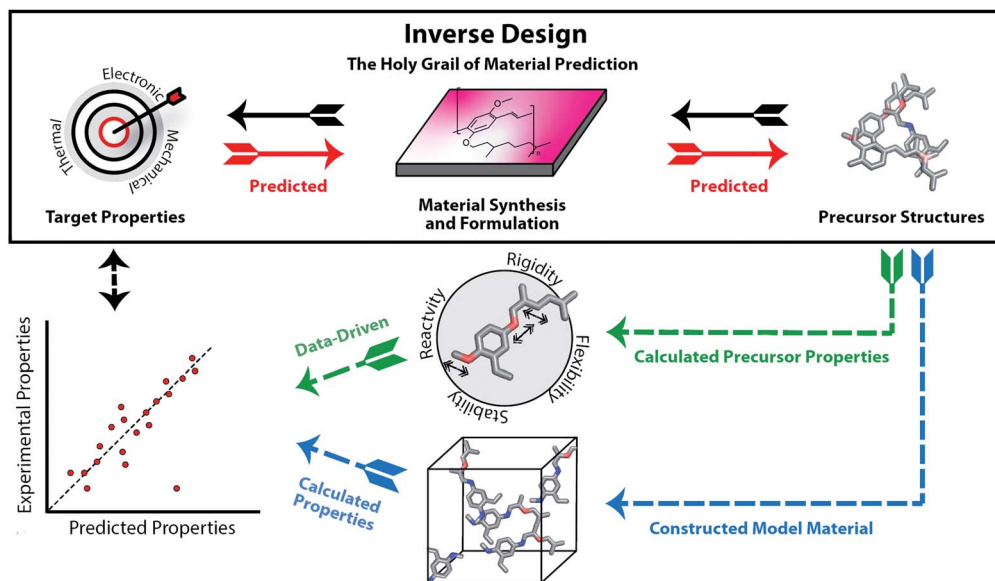


Fig. 1 Inverse design (top), as opposed to a typical computational-experimental discovery workflow (below). Ideally, the precursor structures could be designed directly from the desired properties, considering the synthesis and formulation of the material. Instead, typically researchers start from databases of precursors used to construct molecules *in silico* and calculate the properties of the modelled material using quantum chemistry or machine learning techniques (blue route). Sometimes, data-driven approaches can be used to predict the material properties directly from the precursor structures or properties (green route). The main question stays the same: can both the precursors and the material in its correct form to achieve the desired properties actually be synthesised?

can be subsequently enriched by simulations in order to create new data used for materials prediction. Such a workflow has recently been successfully applied in the prediction of dye pairs suitable for solar-cell applications.¹¹ Natural language processing has been used on a corpus of inorganic materials synthesis literature containing the synthesis conditions for various metal oxides. In conjunction with machine learning, those data allowed the prediction of synthesis outcomes in unseen materials systems.^{12,13} The approaches described above require manual labelling of large datasets and hard-coding complicated language grammar rules, making their implementation extremely labour-intensive. Unsupervised and semi-supervised machine learning have been recently used to by-pass that problem, allowing one to reconstruct not only flowcharts of possible synthetic procedures, but also to capture latent chemical knowledge such as the prediction of promising material candidates years prior their publication.^{14–16}

However, the quality of the datasets and the predictive power of literature-based models can only be as good as the original literature data. Many technical details and conditions are often omitted from reported procedures, which rely on the experience of the experimentalist and hidden instrument settings. Many reactions and processes might indeed be extremely sensitive to changes in room temperature and ambient humidity throughout the year, let alone different countries. Standardisation of experimental procedures and the resulting measurements across the fields of chemistry and material science would greatly simplify the generation of data-driven models and their transferability across various fields. Furthermore, the notable absence of unsuccessful experiments from

the scientific literature causes bias in the datasets and thus reduces the reliability of the resulting knowledge. With the advent of pre-print servers and journals focusing on scientific validity instead of perceived research impact, we anticipate that in the near future we will see more accurate computational models trained on a wider range of publicly available experimental results.

3 Accessibility and availability of components

The first experimental step of bottom-up material synthesis always involves reacting some type of precursor components. Inorganic materials are typically obtained from commercially available or relatively simple elemental compounds, alloys, and salts. While some elements might be scarce, precursors for most naturally occurring elements are readily available and thermodynamically-driven reactions are relatively insensitive to the exact nature of the inorganic building blocks. For those reasons, tailored precursors are rarely used in inorganic materials discovery, especially when using automated synthesis platforms. The vast synthesis search space for inorganic materials is further discussed in Section 4.

For organic and metal-organic materials, the synthesis of the organic precursors might be the most time and resource-heavy part of the development process, and small changes in the structure of the organic building blocks might lead to drastic changes in their reactivity. There are millions of commercially available, albeit not necessarily cheap, organic compounds and a practically infinite space of possible



molecules that can be enumerated. This provides a great advantage when searching for desired properties in organic materials. However, the synthesis of the organic precursors might be prohibitively difficult or even impossible, thus yielding “ideal” materials “unsynthesisable”. Many libraries of organic precursors contain either potentially biologically active molecules (often of high complexity and not viable for large scale material synthesis, *e.g.* the ZINC database)^{17,18} or algorithmically- and machine-learning-generated structures through artificial transformations (hence not necessarily synthetically feasible at all, *e.g.* the GDB family).¹⁹

It is therefore crucial to consider precursor accessibility for material synthesis when assessing whether the material can be experimentally realised. Precursor synthesizability can be used as a filter to remove unsynthesisable materials, but even better, it could provide synthetic chemists with a synthetic protocol (or a retrosynthetic analysis) from commercially available building blocks to the final material. To address this issue, there have been attempts to provide metrics of the synthetic accessibility for organic compounds, albeit focused on their synthesis for drug discovery. These metrics broadly fit into four separate categories: (a) focusing on cheminformatic features related to fragment complexity, (b) built using the available reaction data reported in the chemical literature, (c) based on expert chemists' scoring, and (d) following retrosynthetic planning tools (see Fig. 2).

Early attempts at modelling the synthetic accessibility of organic molecules focused on a quantitative analysis of the underlying complexity of a molecule based on which structural motifs were present.^{20,21} However, the assumption that complexity corresponds to synthetic accessibility is not always valid. To mitigate the issues associated with this complexity-based approach, analyses of molecular fragments found in large compound libraries were incorporated into some scoring methods. Fragments that commonly appeared in commercially available compound databases were considered more synthetically accessible than those that appeared infrequently.^{22–24} Separately, the chemical intuition of expert chemists has been modelled statistically to elucidate the key molecular features that influence a chemist's decision making process.^{25,26}

More recently, extensive reaction databases have paved the way for data-driven synthetic accessibility scores and automated retrosynthesis planning tools to emerge. The synthetic complexity score by Coley *et al.* used the Reaxys databases, comprised of 12 million reactant–product pairs, to train a neural network on a pairwise ranking task.²⁷ Molecules commonly seen as products of reactions in this database were ranked as more synthetically complex than reactants. One significant development in retrosynthesis planning tools utilised the Monte Carlo tree search algorithm, in combination with a neural network to generate viable retrosynthetic disconnections.²⁸ Natural language processing techniques, such as the transformer neural network architecture,²⁹ in addition to reinforcement learning techniques,³⁰ have also been shown to be effective algorithms for performing a retrosynthetic analysis. However, non-ML techniques such as identifying pathways to similar compounds and hard-coding reaction rules

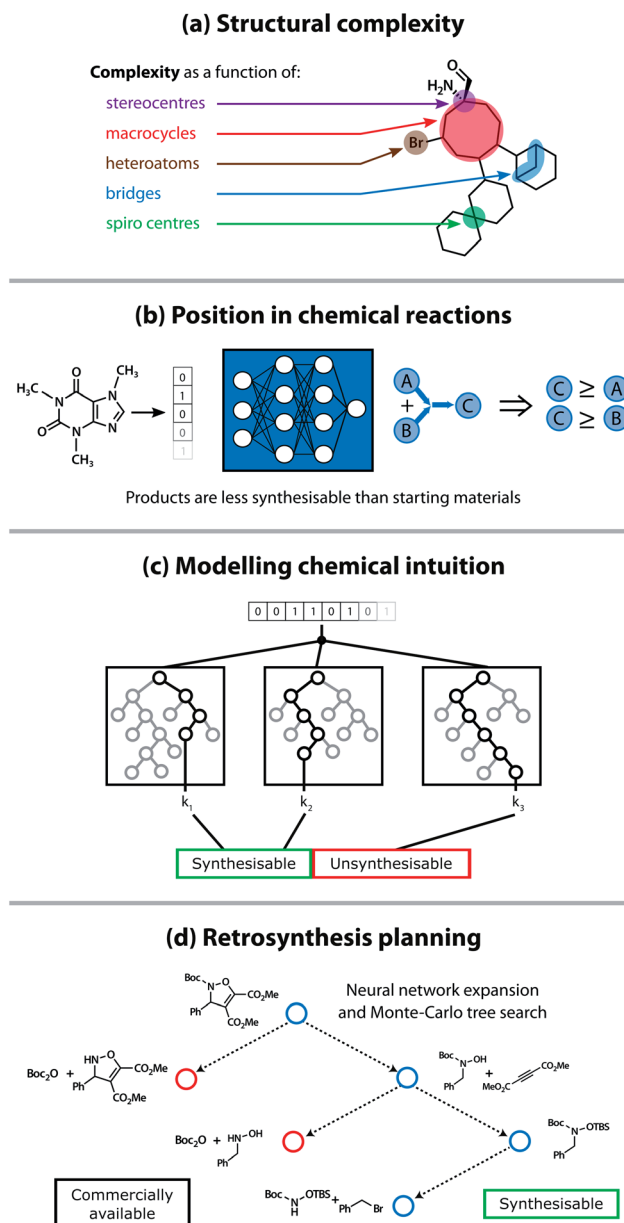


Fig. 2 Four approaches used to calculate the synthetic accessibility of organic compounds: (a) as function of the number of complex structural motifs present in the molecule; (b) deep neural networks trained on extensive reaction databases (reported and proprietary); (c) modelling the decision making process of expert chemists (the “chemical intuition”); (d) by identifying viable reaction pathways using automated retrosynthesis planning tools.

have also performed well.^{31,32} While computer-assisted retrosynthesis methods are now frequently able to find routes to desirable precursors, it is often time-consuming to fully explore the generated reaction tree and many state-of-the-art models have limited or commercial-only access, making them somewhat unsuitable for high-throughput screening methods.

None of the aforementioned synthetic accessibility scores are perfect, but they can greatly simplify the materials discovery process when acting as a filter for unreasonable precursors.



Material synthesis often requires precursors that are cheap and accessible on a large scale, and these heuristic methods could be used to identify precursors that fit these criteria, although not originally developed for this purpose. These methods have been shown to increase the number of experimentally viable candidates when incorporated into objective functions of generative algorithms.³³ Such methods may in the future be incorporated into prediction workflows, providing a full synthetic scheme from simple and commercially available building blocks to final materials.

4 Material synthesis

There are two major materials synthesis routes: direct synthesis of a framework material, *e.g.* inorganic oxides, zeolites, metal-organic frameworks (MOFs), covalent organic frameworks (COFs) – be they crystalline or amorphous – or solid-phase formulation of molecular materials (*e.g.* porous organic cages). Challenges for framework materials are: do the reactive groups react in the way we want them to, can we predict the solid-phase structure of the resulting material, and is that the form that gives the desired properties? In the case of molecular materials, the corresponding challenges are the following: do the reactive groups react in the way that we would want them to create discrete molecular units and can we predict the structure of those units, (*e.g.* the correct topology – for organic cages formed from a condensation reaction between a tritopic and a ditopic building blocks, there are six commonly observed cage topologies)³⁴ and can we predict the packing of those units in the solid-phase? The two prediction steps are crucial in this case, as the properties of the resulting material will depend on the intrinsic properties of the discrete units and the extrinsic properties originating from solid-phase packing.

The formulation of solid-state materials carries other challenges than those present in synthetic chemistry. While molecular reactions commonly happen at homogeneous equilibria with all reactants present at finite amount, formulation into the solid-state always involves heterogeneous equilibrium and may often proceed until one or more of the condensed phases fully disappear. The crystal (or amorphous) solid-state structure can provide chemists with valuable insight into many properties of the target materials. Therefore, solid-phase structure prediction is invaluable in predicting new materials. Solid-phase structure prediction is a global optimisation problem, where one is trying to generate the thermodynamic structure based on the chemical composition of the material.³⁵ The two major considerations are the efficient exploration of the multidimensional energy landscape, and then the correct ranking of the relative energies of the resulting spatial arrangements.³⁶ We shall focus on crystal structure prediction (CSP) of crystalline materials, but significant advances have also been made in amorphous structure prediction.^{37,38}

While it is tempting to derive a common theory unifying solid-phase structure prediction for organic and inorganic materials, the challenges in those fields are different and hence they must advance in parallel. We will discuss the different

prediction aspects related to inorganic and organic (and hybrid metal-organic) materials below, as outlined in Fig. 3.

Inorganic materials

Inorganic materials are most often mixtures of metals or their oxides with a practically infinite continuum of possible elemental compositions.³⁹ Since availability of the inorganic precursors is generally not the limiting factor, the composition possibilities are endless and structure prediction is thus challenging. Indeed, not all possible binary systems were even partially experimentally studied under normal conditions. When looking at the more complex ternary and quaternary structures, the vast majority of the possible compound space remains unexplored.⁴⁰

Inorganic CSP in its simplest form is based on structure sampling followed by (normally) *ab initio* energy calculations.^{41–43} Other methods used to effectively explore the configurational space include simulated annealing,^{44,45} basin hopping,⁴⁶ minima hopping,⁴⁷ metadynamics,⁴⁸ and evolutionary techniques.^{36,49–51} Such CSP techniques require high-level *ab initio* calculations for reliable relative energy ranking to correctly identify local minima and predict, at least theoretically, the existence of (meta)stable inorganic structures. Various further simplifications have been implemented, such as the application of common formulae per unit cell or the use of robust building blocks like inorganic anions.⁵² Complementary CSP approaches to the computationally expensive *ab initio* methods harvest the large amount of data gathered in the Inorganic Crystal Structure Database⁵³ and the Cambridge Structural Database.⁵⁴ Recently, Hegde and Aykol *et al.* encoded thermodynamic stability of inorganic materials as a phase diagram network. Such a graph-theoretic approach allows one to uncover unknown relationships between materials, such as the discovery of novel materials as absences of expected nodes in thus derived networks.⁵⁵

Machine learning can be applied directly to crystallographic data “allowing prediction of novel structures”^{56–59} or to the *ab initio* energy landscape,^{38,60–62} thus providing much faster yet accurate potentials used for structure prediction. In particular, generative machine learning models allow one to generate materials with target properties rather than predict the properties of candidate materials. Such approaches have been used to generate novel structures of binary and ternary oxides.^{63,64} Furthermore, data-mining approaches focus on the extrapolation from the structures that have already been explored and might not be useful for prediction of completely new classes of materials.

Mixed oxides are of high interest due to their applications in the molecular electronics and energy materials. A major issue in the prediction of such materials is the relative stability of possible phases, as unstable phases are by default not synthesisable. Buckeridge *et al.* reported software that can be used to test the thermodynamic stability of multi-ternary materials.⁵⁶ Their algorithm significantly speeds up the analysis of the thermodynamic stability of a material, which is normally a very slow process in the case of ternary and quaternary systems.



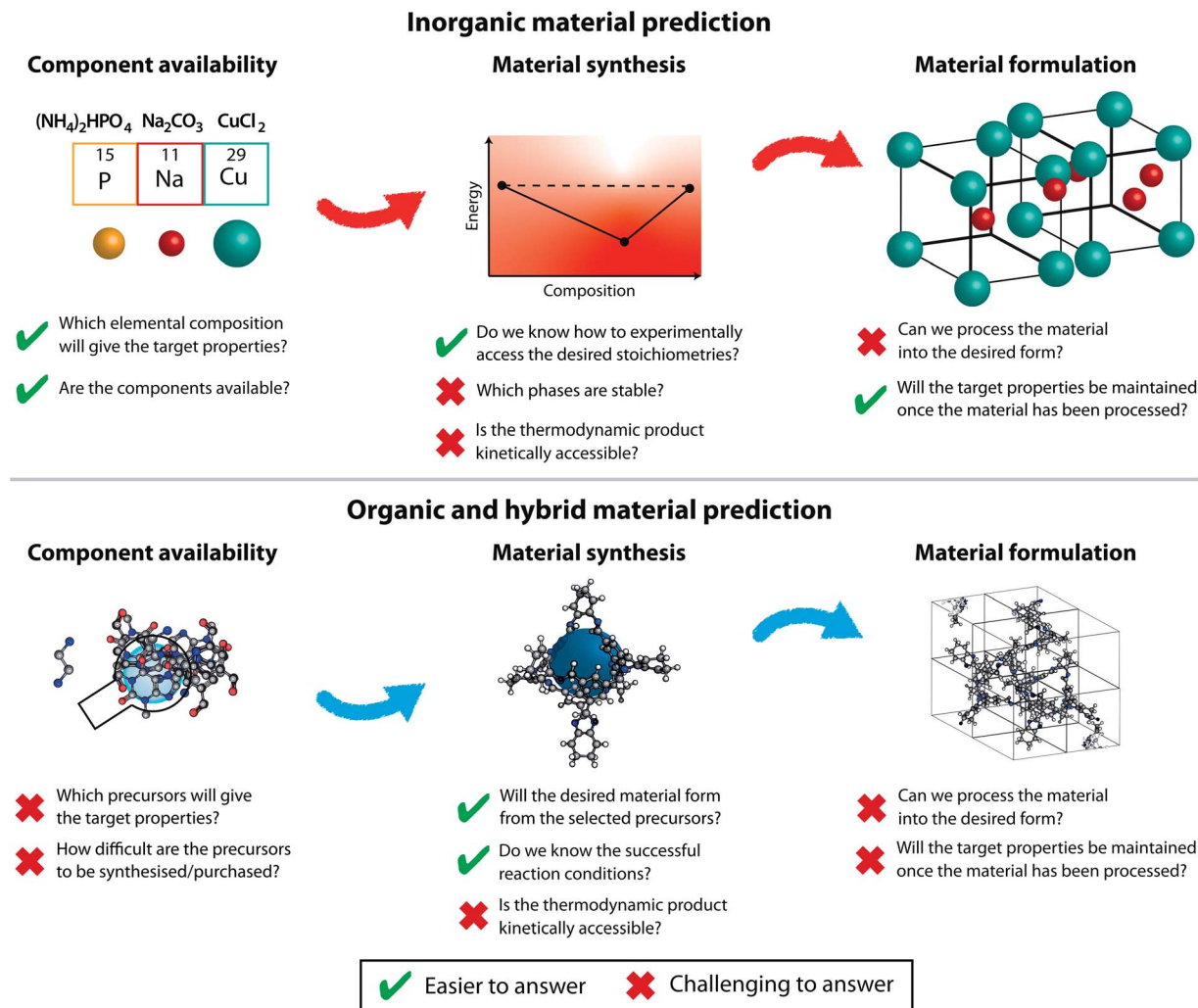


Fig. 3 The challenges to be considered at various stages of prediction for organic materials (top) and inorganic materials (bottom). There are additional considerations at the device level that are not considered in this figure.

Density functional theory calculations have also been successfully used to obtain reaction enthalpies for complex oxides with large unit cells and doped oxides from binary oxides, which correctly identified the structures obtained experimentally.^{65–67}

Porous materials find application in molecular separations, catalysis and sorption. The most industrially widespread microporous inorganic materials are zeolites and zeo-types: crystalline structures formed by XO_4 tetrahedra (where $\text{X} = \text{Si}, \text{Al}, \text{P}, \text{etc.}$). All possible zeolitic frameworks have been enumerated mathematically, or built bottom-up from secondary building units, providing the basis for the prediction of novel materials.^{68,69} These materials have then been assessed for their relative thermodynamic stability, which would correlate with their ease of synthetic access.⁷⁰ To access hypothetical structures, organic templates have also been successfully computationally screened.⁷¹

In summary, accurate prediction of the relative stability of different phases in inorganic materials can help researchers filter out non-viable candidates, hence allowing for the identification of promising materials that can actually be synthesised.

However, two different yet similar phases can both be locally stable and the desired form might not be obtained upon first attempt, as the relative thermodynamic stability does not imply that the structure can actually be accessed since other factors, such as kinetic barriers or unknown synthetic routes, might inhibit their experimental realisation. Looking on the bright side, thermodynamic stability at least focuses chemists' efforts in hopeful areas.

Organic and hybrid materials

If the organic precursors for organic and hybrid materials are readily available and stable (which is a prediction challenge in itself, as discussed in Section 3), the synthesis of the material is typically a result of a relatively simple condensation reaction between building blocks. Generally, a single topology of the product is anticipated, but researchers should always expect formation of beautiful and unexpected structures, especially if slight variations from the perfect precursor geometry are present.^{72,73} We have recently enumerated common topologies



unsuccessful synthetic attempts will provide more balanced datasets, which will improve accuracy of computational models and speed up the discovery process. Close discussions and feedback loops between computational predictions and experimental outcomes can direct modelling efforts towards more fruitful and trustworthy results. Those more relevant models will in turn further stimulate experimental material discovery. If we can get better at predicting materials that can be quickly synthetically realised, it will overcome a major bottleneck in the discovery process, offering the potential to accelerate materials discovery beyond the current 40 year timescale.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We thank Dr Rebecca Greenaway for useful discussions. K. E. J. thanks the Royal Society for a Royal Society University Research Fellowship. K. E. J. and F. T. S. thank the Leverhulme Trust for a Leverhulme Trust Research Project Grant. S. B. thanks the Leverhulme Research Centre for Functional Materials Design for a Ph.D. studentship. We acknowledge funding from the European Research Council under FP7 (CoMMaD, ERC Grant No. 758370).

Notes and references

- 1 K. Alberi, M. B. Nardelli, A. Zakutayev, L. Mitás, S. Curtarolo, A. Jain, M. Fornari, N. Marzari, I. Takeuchi, M. L. Green, M. Kanatzidis, M. F. Toney, S. Butenko, B. Meredig, S. Lany, U. Kattner, A. Davydov, E. S. Toberer, V. Stevanovic, A. Walsh, N. G. Park, A. Aspuru-Guzik, D. P. Tabor, J. Nelson, J. Murphy, A. Setlur, J. Gregoire, H. Li, R. Xiao, A. Ludwig, L. W. Martin, A. M. Rappe, S. H. Wei and J. Perkins, The 2019 materials by design roadmap, *J. Phys. D: Appl. Phys.*, 2019, **52**, 013001.
- 2 H. S. Stein and J. M. Gregoire, Progress and prospects for accelerating materials science with automated and autonomous workflows, *Chem. Sci.*, 2019, **10**, 9640–9649.
- 3 B. Burger, P. Maffettone, V. Gusev, C. Aitchison, Y. Bai, W. Xiaoyan, R. S. Sprick and A. I. Cooper, A Mobile Robotic Researcher, *Nature*, 2020, **583**(7815), 237–241.
- 4 P. G. Polishchuk, T. I. Madzhidov and A. Varnek, Estimation of the size of drug-like chemical space based on GDB-17 data, *J. Comput.-Aided Mol. Des.*, 2013, **27**, 675–679.
- 5 P. Ertl, Cheminformatics analysis of organic substituents: Identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 374–380.
- 6 S. Bennett, A. Tarzia, M. A. Zwijnenburg and K. E. Jelfs, *Artificial Intelligence Applied to the Prediction of Organic Materials*, Royal Society of Chemistry, Cambridge, 2020, ch. 12.
- 7 M. Jansen and J. C. Schön, Rational development of new materials — putting the cart before the horse?, *Nat. Mater.*, 2004, **3**, 838.
- 8 L. Hawizy, D. M. Jessop, N. Adams and P. Murray-Rust, ChemicalTagger: A tool for semantic text-mining in chemistry, *J. Cheminf.*, 2011, **3**, 1–13.
- 9 D. M. Lowe and R. A. Sayle, LeadMine: A grammar and dictionary driven approach to entity recognition, *J. Cheminf.*, 2015, **7**, 1–9.
- 10 M. C. Swain and J. M. Cole, ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature, *J. Chem. Inf. Model.*, 2016, **56**, 1894–1904.
- 11 J. M. Cole, A Design-to-Device Pipeline for Data-Driven Materials Discovery, *Acc. Chem. Res.*, 2020, **53**, 599–610.
- 12 E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder and E. Olivetti, Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning, *Chem. Mater.*, 2017, **29**, 9436–9444.
- 13 E. Kim, Z. Jensen, A. Van Grootel, K. Huang, M. Staib, S. Mysore, H. S. Chang, E. Strubell, A. McCallum, S. Jegelka and E. Olivetti, Inorganic Materials Synthesis Planning with Literature-Trained Neural Networks, *J. Chem. Inf. Model.*, 2020, **60**, 1194–1201.
- 14 H. Huo, Z. Rong, O. Kononova, W. Sun, T. Botari, T. He, V. Tshitoyan and G. Ceder, Semi-supervised machine-learning classification of materials synthesis procedures, *npj Comput. Mater.*, 2019, **5**, 1–7.
- 15 L. Weston, V. Tshitoyan, J. Dagdelen, O. Kononova, A. Trewartha, K. A. Persson, G. Ceder and A. Jain, Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction from the Materials Science Literature, *J. Chem. Inf. Model.*, 2019, **59**, 3692–3702.
- 16 V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder and A. Jain, Unsupervised word embeddings capture latent knowledge from materials science literature, *Nature*, 2019, **571**, 95–98.
- 17 J. J. Irwin and B. K. Shoichet, ZINC - A free database of commercially available compounds for virtual screening, *J. Chem. Inf. Model.*, 2005, **45**, 177–182.
- 18 T. Sterling and J. J. Irwin, ZINC 15 - Ligand Discovery for Everyone, *J. Chem. Inf. Model.*, 2015, **55**, 2324–2337.
- 19 L. Ruddigkeit, R. Van Deursen, L. C. Blum and J. L. Reymond, Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17, *J. Chem. Inf. Model.*, 2012, **52**, 2864–2875.
- 20 S. H. Bertz, The First General Index of Molecular Complexity, *J. Am. Chem. Soc.*, 1981, **103**, 3599–3601.
- 21 H. W. Whitlock, On the structure of total synthesis of complex natural products, *J. Org. Chem.*, 1998, **63**, 7982–7989.
- 22 P. Ertl and A. Schuffenhauer, Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions, *J. Cheminf.*, 2009, **1**, 1–11.



- in synthesis of metal-organic frameworks, *Nat. Commun.*, 2019, **10**, 1–7.
- 81 S. L. Price, Predicting crystal structures of organic compounds, *Chem. Soc. Rev.*, 2014, **43**, 2098–2111.
- 82 A. J. Cruz-Cabeza, S. M. Reutzler-Edens and J. Bernstein, Facts and fictions about polymorphism, *Chem. Soc. Rev.*, 2015, **44**, 8619–8635.
- 83 J. Yang, S. De, J. E. Campbell, S. Li, M. Ceriotti and G. M. Day, Large-Scale Computational Screening of Molecular Organic Semiconductors Using Crystal Structure Prediction, *Chem. Mater.*, 2018, **30**, 4361–4371.
- 84 A. G. Slater and A. I. Cooper, Function-led design of new porous materials, *Science*, 2015, **348**, aaa8075.
- 85 T. Hasell and A. I. Cooper, Porous organic cages: Soluble, modular and molecular pores, *Nat. Rev. Mater.*, 2016, **1**(9), 16053.
- 86 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 and G. M. Day, Functional materials discovery using energy-structure-function maps, *Nature*, 2017, **543**, 657–664.
- 87 R. L. Greenaway, V. Santolini, A. Pulido, M. A. Little, B. M. Alston, M. E. Briggs, G. M. Day, A. I. Cooper and K. E. Jelfs, From Concept to Crystals *via* Prediction: Multi-Component Organic Cage Pots by Social Self-Sorting, *Angew. Chem., Int. Ed.*, 2019, **58**, 16275–16281.
- 88 L. Bouéssel Du Bourg, A. U. Ortiz, A. Boutin and F. X. Coudert, Thermal and mechanical stability of zeolitic imidazolate frameworks polymorphs, *APL Mater.*, 2014, **2**(12), 124110.
- 89 Y. G. Chung, J. Camp, M. Haranczyk, B. J. Sikora, W. Bury, V. Krungleviciute, T. Yildirim, O. K. Farha, D. S. Sholl and R. Q. Snurr, Computation-Ready, Experimental Metal–Organic Frameworks: A Tool To Enable High-Throughput Screening of Nanoporous Crystals, *Chem. Mater.*, 2014, **26**, 6185–6192.
- 90 Y. G. Chung, E. Haldoupis, B. J. Bucior, M. Haranczyk, S. Lee, H. Zhang, K. D. Vogiatzis, M. Milisavljevic, S. Ling, J. S. Camp, B. Slater, J. I. Siepmann, D. S. Sholl and R. Q. Snurr, Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: CoRE MOF 2019, *J. Chem. Eng. Data*, 2019, **64**, 5985–5998.
- 91 P. Z. Moghadam, A. Li, S. B. Wiggin, A. Tao, A. G. P. Maloney, P. A. Wood, S. C. Ward and D. Fairen-Jimenez, Development of a Cambridge Structural Database Subset: A Collection of Metal–Organic Frameworks for Past, Present, and Future, *Chem. Mater.*, 2017, **29**, 2618–2625.

