Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Chemical Engineering

Search methods for inorganic materials crystal structure prediction

Xiangyu Yin and Chrysanthos E Gounaris

Crystal structure prediction (CSP) is the problem of determining the most stable crystalline arrangements of materials given their chemical compositions. In general, CSP methodologies include two algorithmic steps, namely a method for assessing material stability of any given design, and a search algorithm for exploring the design space. For inorganic crystals, in particular, the most critical aspect is to develop an effective search algorithm. This paper summarizes previous research and discusses recent progress in search methods developed for inorganic CSP. Empirical methods, guided-sampling algorithms, and more recent data-driven approaches are discussed. Additionally, we describe a mathematical optimization-based search paradigm that has been recently introduced as an alternative CSP approach. A semiconductor nanowire design approach is then presented to illustrate this paradigm.

Address
Department of Chemical Engineering, Carnegie Mellon University, United States

Corresponding author: Gounaris, Chrysanthos E (gounaris@cmu.edu)

Current Opinion in Chemical Engineering 2021, **XX**:xx-yy

This review comes from a themed issue on **Frontiers in chemical engineering; chemical product design — II**

Edited by **Rafiqul Gani**, **Lei Zhang** and **Chrysanthos Gounaris**

https://doi.org/10.1016/j.coche.2021.100726

Introduction

Crystal structure prediction (CSP) pertains to identifying the most stable structures of given classes of crystalline materials. The stability metric for CSP is often defined via a complex energy function, such as a potential energy surface, while in certain cases, other more straightforward metrics such as the structure's cohesive energy may be used instead. CSP is fundamentally important for materials research and application, as the predicted most stable crystal structure is often used to guide the experimental synthesis and other theoretical studies. Indeed, even with solely the crystal structure information at hand, we can calculate many physical and chemical properties of crystalline materials using *ab initio* or machine learning methods [1]. In addition, CSP plays a critical role in crystal engineering and, more broadly, in the inverse design of crystalline materials, enabling us to design materials with desired properties. Clearly, the ability to perform accurate, efficient, and robust CSP is key to next-generation technologies in energy transformation and storage, catalysis, and quantum computing [2••].

There are two major pieces in CSP methodologies, namely the means to assess the stability of a given structure, and the process to search over the space of possible structures. First, using some appropriate stability assessment method (a.k.a. stability *ranking*), we seek to obtain an estimate of relative stability between all candidate structures in terms of a defined stability metric. The structure search step then explores the structural design space and identifies the structures that correspond to the optima of the stability metric function. Whereas the stability ranking problem is usually tackled with existing computational chemistry tools for predicting material properties, the structure search step often involves the development of custom-built algorithmic procedures to numerically zero in the most promising structures.

Historically, there is a parallelism between inorganic and organic CSP research. The stability ranking problem has attracted more research attention in the organic CSP community, while the inorganic CSP community has focused more efforts on structure search algorithms. This could be partially explained by the difference between *ab initio* computational tools utilized by those two communities. Whereas organic materials' properties can often be obtained from fast but less accurate Monte-Carlo simulations or molecular dynamics calculations, computational studies for inorganic materials usually involve accurate but more expensive quantum-mechanics calculations. Thus, the organic CSP community has focused more on improving the underlying force fields for more accurate stability ranking, while the inorganic CSP community has devoted more time on advanced search methods to reduce computational cost and/or function evaluations required for the structure search problem. Furthermore, organic crystalline materials can usually be expressed via modular building blocks (e.g. molecules), resulting in much smaller structural design spaces than then case of inorganic materials, where often the structure needs to be expressed at the atom level. In fact, enumerative search algorithms are often viable means for CSP of certain classes of organic materials. On the other hand, the search space of inorganic CSP is generally astronomically large.

For example, a target periodic unit cell with K atoms will possess at least $3K + 3$ degrees of freedom[1], even without considering the types of atoms. It is estimated that there are at least 10^K potential structure candidates for such a system, as approximated by Oganov and Glass [3•]. The exponential increase of design space with respect to system size makes the inorganic CSP search problem a very challenging combinatorial optimization problem, which necessitates systematic algorithms to explore and reduce the design space efficiently. In this review, we will specifically focus on inorganic materials CSP search methods. For organic CSP search methodologies, we refer the readers to Bowskill *et al.* [4••] for a comprehensive overview.

Screening all candidate structures is inefficient and costly. Two major categories of systematic search algorithms have thus been developed for inorganic CSP, namely the guided sampling approaches and the data-driven approaches. The earliest data-driven methods were developed in the form of empirical rules that exploited analogies between structures. Later on, many meta-heuristics based sampling algorithms (e.g. simulated annealing, genetic algorithms, particle swarm optimization) were proposed and employed to solve inorganic CSP search problems. More recently, data-driven methods have again become the research hotspot given all the advancements in machine learning and deep learning algorithms. Additionally, hybrid frameworks have been developed where meta-heuristics and data-driven approaches are combined. A brief summary of CSP search methods for inorganic materials is provided in Table 1, while more detailed discussions along with references are provided in the remainder of the manuscript. More specifically, in the section 'Guided-sampling methods', we discuss guided sampling search methods previously developed for inorganic CSP, including random sampling, simulated annealing, and evolutionary algorithms. Next, in the section 'Data-driven methods' we discuss recent progress in applying advanced data-driven techniques and models in inorganic CSP. Finally, in the section 'Mathematical optimization', we introduce a mathematical optimization-based materials design paradigm that has been recently developed as an alternative CSP approach for inorganic materials. We illustrate this paradigm using a case study of designing stable semiconductor nanowires.

Guided-sampling methods

As mentioned earlier, the design space of an inorganic CSP problem is prohibitively large for enumerative screening. To reduce the computational cost, a straightforward idea is to sample candidate structures from the design space iteratively, hoping that one of the sampled structures would be the optimally stable one. Such guided sampling methods have been practiced in various material systems and proved to provide sufficiently good solutions in many cases. Two key methodological steps are involved in designing a sampling-based search method. The first step pertains to sampling from the design space (i.e. generating new structure candidates), and it is usually performed by a material class-specific heuristic that is developed to that purpose. The second step is tasked with guiding the sampling process and involves starting, stopping and tuning the search. For this, meta-heuristic frameworks are often deployed. In the following, we will discuss several popular generic meta-heuristics in the literature that can couple with materials-specific heuristics for inorganic CSP.

The simplest approach is *random sampling*. This term is used to describe a random walk-like procedure where the sampling is purely random, storing the best solution encountered in the process thus far. Although primitive, random sampling has proven to be a sufficiently good algorithm in a series of practical applications, as illustrated by Pickard and Needs [5–8]. Their methodology, called the *ab initio random structure searching* (AIRSS), is summarized and described systematically in [9•]. We note that random sampling is more widely used in the organic CSP community for reasons discussed earlier. For example, purely random, grid-based, or more involved quasi-random sampling have all been shown to perform well in blind tests of organic CSP [10,11].

Inorganic CSP search problems often require more efficient search methods. To that end, several nature-inspired meta-heuristic frameworks have been successfully utilized for inorganic CSP. *Simulated annealing* (SA) is one of the earliest developed and widely applied sampling algorithms. It was inspired by the physical annealing process and first proposed by Kirkpatrick *et al.* [12] to address classic combinatorial optimization problems. It was then soon utilized for framework materials CSP [13] and dense-packed materials CSP [14]. Later, [15] carried out a series of SA-based inorganic CSP studies for different material systems, with their methodology systematically described in [16]. The cost functions used in those studies are usually empirical or theoretically derived potential energy surfaces (PES) that are highly non-convex and have lots of local minima, which reduces the search efficiency of SA.

Finnila *et al.* [17] proposed quantum annealing, which performs SA on the quantum mechanics characteristics of nanoclusters to avoid local minima. Reinaudi *et al.* [18,19] imposed symmetry restrictions to decrease the number of local minima. Another technique to improve the efficiency of the search process is the *basin-hopping* technique, which transforms the PES surface into a set of basins without changing the global minimum [20,21]. Mellot-Draznieks *et al.* [22–25] proposed the *automated*

[1] Six lattice parameters plus the shift-invariant xyz-coordinates of the atoms, that is, $6 + 3(K - 1)$ degrees of freedom.

Table 1

Summary of search methods for inorganic materials CSP

Guided-sampling methods
- Use of metaheuristic algorithms to guide the sampling of candidate structures
- Require a well-defined sampling heuristic and fitness function (i.e. search objective)

Random sampling
- Samples structures randomly or quasi-randomly
- Simple but only sufficient for problems with relatively limited search space

Simulated annealing
- Trades-off search intensity and diversity by imitating the physical annealing process
- Efficient for fitness functions with relatively small number of local optima

Hopping methods
- Transform the fitness function into basins via local structure relaxation
- Useful for high dimensional fitness functions that lend themselves to an efficient local optimization routine

Evolutionary algorithms
- Maintain and evolve a population of structures imitating natural evolution processes
- State-of-the-art for various applications, but may be intractable when stability assessment is computationally expensive

Data-driven methods
- Predict crystal properties/types/structures from existing structural information
- Require considerable amount of data

Empirical rules
- Predict with empirical principles/diagrams obtained via observation and/or simple data mining
- Simple and fast, but with relatively low precision

ML-based structure-function relationships
- Develop inexpensive ML models of structure-function relationships for use in CSP search
- Efficient and accurate when combined with proper structural representation

Metaheuristic-ML hybrid methods
- Accelerate the guided sampling by learning ML models on the fly
- Useful for problems with expensive stability assessment computation

Crystal structure classification
- Classify and select materials macro-structure types directly from data
- Unable to predict non-existing structure types

Generative models
- Learn a representation of structures and directly reconstruct stable candidates
- Require an efficient and invertible structure representation

Mathematical optimization
- Formulates the problem as a mathematical optimization model and solves for its global optima
- Rigorous and flexible, but with relatively higher computational cost

Nonlinear programming
- Models with continuous variables, incorporating nonlinear constraints, when applicable
- Compatible with most algebraic structure-function relationships

Mixed-integer linear programming
- Models problem with a mix of continuous and integer variables (including logical/Boolean variables), but utilizes only linear constraints
- Requires more modeling effort, but shows better numerical tractability than previous method

assembly of secondary building units (AASBU) that transforms the design space from atoms to secondary building units. This method performs especially well in terms of reducing the computational cost for CSP of framework structures.

Geodecker [26•] proposed the *minima hopping* (MH) algorithm that introduces a history list that contains all previously visited minima to avoid exploring the same regions. While MH has similar characteristics as basin-hopping, it has a different theoretical basis and is

expected to climb out of a local minimum must faster than the latter. This method has been utilized extensively in non-periodic systems such as nanoclusters [27–29]. Amsler and Geodecker [30] extended this method to be able to deal with unit cell-based periodic systems. Later, researchers further extended MH to handle surfaces [31], porous structures [32], two-dimensional materials [33,34], and grain boundaries [35]. We note that, over time, the cost function in MH has evolved from theoretical and/or empirical PES to density functional theory (DFT) calculated PES, following the major trend in computational chemistry.

A large category of meta-heuristic search frameworks is that of evolutionary algorithms, with the most popular variant being *genetic algorithms* (GA), which are inspired by biological evolution [36]. Unlike the other aforementioned algorithms, which consider (and gradually improve) a single solution at a time, the main characteristic of GA is that it maintains a collection (a.k.a. *population*) of solutions. To the best of our knowledge, the first deployment of GA for addressing an inorganic CSP search problem was proposed by Smith [37]. Bush *et al.* [38] later combined GA with a standard local energy minimization routine. Deaven and Ho [39] described the system with atomic coordinates and applied advanced mating and mutation techniques, proving that GA outperforms SA significantly. Many of the ideas in that paper, such as a real-space representation of structures, spatial heredity, and the use of local optimization, profoundly influenced many evolutionary algorithms developed later. Johnston [40] has utilized GA extensively to study nanoclusters and nanoparticles. Woodley *et al.* [41,42] proposed a multistage GA method and further improved their method by imposing geometry constraints [43]. More recent studies have focused on improving, extending, and efficiently implementing GA algorithms. Lloyd *et al.* [44] explored several strategies to improve the efficiency of the GA algorithm for nanoalloy cluster CSP, such as advanced initialization, predating, and mutation schemes. Glass *et al.* [45] implemented a code called the *universal structure predictor: evolutionary xtallography* (USPEX) with advanced features, such as local optimization, spatial heredity, lattice mutation, and others. They tested the USPEX code on numerous systems and observed a high success rate with some examples discussed in [3•]. Trimarchi and Zunger [46] designed an evolutionary algorithm to predict both the lattice geometry and the atomic configuration of a crystalline material, referred to as the *global space-group optimization* (GSGO) problem. They further extended their algorithm to be able to predict the stoichiometries at the same time [47]. Froltsov and Reuter [48] discussed how the size and mutation schemes in a GA algorithm would affect its robustness concerning algorithmic parameters in the context of nanocluster CSP. Woodley and Catlow [49] implemented and compared the Darwinian and Lamarckian evolving schemes in GA algorithms for CSP. Lonie and Zurek [50] implemented *XTALOPT*, another GA algorithm in which they developed a new periodic displacement operator and used mixed operators to eliminate duplicate structures and improve search efficiency.

Particle swarm optimization (PSO) is yet another evolutionary meta-heuristic framework that has been successfully utilized in inorganic CSP search problems. First proposed by Kennedy and Eberhart [51], it was inspired by patterns of a flying flock of birds. Wang *et al.* [52••] implemented such an algorithm for CSP in their *crystal structure analysis by particle swarm optimization* (CALYPSO) code. For this, they designed a unique scheme to eliminate similar structures as well as imposed symmetry-breaking constraints to improve the search efficiency. Readers can find a detailed description in Wang *et al.* [53]. Later, Lyakhov *et al.* [54] reported new developments of the USPEX as well as developed a version of PSO algorithm based on their core methodologies, showing that that USPEX strongly outperforms PSO. Meanwhile, Wu *et al.* [55] have developed a new GA framework where they combined the classical potentials with DFT calculations to achieve an efficient GA-based CSP with DFT accuracy. The research on meta-heuristic sampling search methods is still highly active, with new algorithms and applications reported every year.

Data-driven methods

All the previously mentioned guided sampling methods rely on either less accurate algebraic structure-function relationships (e.g. empirical equations) or computationally expensive oracles (e.g. DFT calculations). At the same time, a meta-heuristic search algorithm is highly dependent on the problem settings (e.g. the definition of the design space, cost function), making it difficult to modify and transfer to other problems. Data-driven methods have been proposed as an alternative approach for CSP to address these issues. Compared with meta-heuristic search methods, data-driven methods can learn the implicit rules and constraints governing stability, based on a large number of known stable crystal structures, to accelerate the exploration of the design space [56].

The practice of data-driven CSP dates back to the use of empirical rules derived from experimental as well as theoretical observations. The earliest results include Pauling's rules for determining the structures of ionic compounds [57,58], which guided some of the earliest attempts to justify and guide the structure search. Villars [59–62] built multiple three-dimensional stable phase diagrams from data analysis to predict lattice geometry of thousands of binary and ternary compounds. Later, more elaborate phase diagrams were developed, such as the *Miedema rules* for predicting compound forming and

the *Pettifor maps* for predicting binary, pseudobinary, and ternary compounds [63,64].

With advances in computational chemistry algorithms and the ready access to faster computers, the availability of crystal structures and their calculated properties have grown exponentially. Large online databases have also been established to encourage standardization and collaboration within the community. Important online databases for inorganic crystals include the *International Crystal Structure Database* (ICSD) [65], the *Materials Project* (MP) [66], the *Open Quantum Materials Database* (OQMD) [67], and the *Atomic-FLOW for materials discovery* (Aflow) [68]. Schön [69] discussed how those databases could be employed to assist CSP studies. With such vast quantity of available data, advanced data science and machine learning (ML) techniques have found ample potential for application in inorganic CSP studies.

One crucial aspect of data-driven CSP studies is to learn appropriate structure-function relationships. Early attempts have used simple data mining techniques, such as perceptron [70] and regression [71]. More advanced ML architectures have since been utilized and resulted in more accurate structure-function relationships. Special research interest has been focused on the representation of crystalline material structures, as it will significantly affect the speed and accuracy of the machine-learned model. Behler and Parrinello [72] represented the crystal systems with all atomic positions and built a neural network model to learn the DFT potential energy surfaces. Meredig *et al.* [73] developed an ensemble tree-based formation energy prediction model, which takes atomic properties of constituent elements as inputs, reducing the computational cost by six orders of magnitude in certain cases. Schütt *et al.* [74] proposed a new representation that is suitable for periodic solids based on partial radial distribution functions. Faber *et al.* [75] focused on ML models of formation energies of solids, investigating the performance of various crystal structure representations, including Ewald sum matrices and generalized Coulomb matrices. Isayev *et al.* [76] proposed property-labeled materials fragments as representations of inorganic crystal materials, which requires minimal structural inputs while preserving a high prediction accuracy. Schmidt *et al.* [77] implemented ridge regression, random forests, extremely randomized trees, and neural network models for predicting the stability of perovskites based solely on properties of constituent elements. They found that those ML techniques speed up the computation by at least five-fold without degradation of accuracy. Xie and Grossman [1] proposed to use the connections between atoms (i.e. connectivity matrix) of the crystal as an interpretable universal representation of crystalline materials in their *crystal graph convolutional neural networks* (CGCNN) framework, and they showed promising predictions for various properties including formation energies. Zhou *et al.* [78] developed the *Atom2Vec* code that learns properties of atoms as feature vectors for other ML models. In addition, Chen *et al.* [79] constructed a topology-based ML model using a low dimensional representation of crystal structures derived from persistent homology methods.

Generally speaking, ML models perform better when incorporating structural information. The proper level of abstraction depends on the materials system as well as the ML model used in each case (e.g. supervised versus unsupervised, classification versus regression) and what type of data (e.g. features, labels) are available. For a more detailed discussion, we refer the reader to comprehensive reviews on this topic [80,81]. Today, ML techniques have been well-accepted and are common practice to accelerate and/or outright replace expensive first-principles calculations, as suggested by Ward and Wolverton [82] and Schmidt *et al.* [83]. Following this trend, popular meta-heuristic search frameworks have also incorporated ML techniques. Tong *et al.* [84•] have combined an ML potential with their *CALYPSO* code to pre-construct the ML potential and replace DFT, as well as to train the ML potential on the fly during the search. Deringer *et al.* [85•] have also incorporated ML-based interatomic potentials in their *AIRSS* code to reduce the computational cost. Podryabinkin *et al.* [86••] proposed to train ML interatomic potentials with the USPEX evolutionary algorithm. Jennings *et al.* [87] developed *MLaGA*, an ML-accelerated GA framework for nanoparticle CSP and reported a 50-fold reduction in computational cost. It is worth noting that the use of ML-derived structure-function relationships has the potential to accelerate the meta-heuristic search significantly. However, given the combinatorially large design space of inorganic crystals, the number of structure-function evaluations may still pose tractability challenges in many cases. Moreover, the search still suffers from inherent limitations of meta-heuristics, such as getting stuck at a local minimum.

Ideally, data-driven CSP should locate the optimal structure(s) directly in the design space given information about stable crystal structures. This idea roots in the earliest studies of structure analogy, such as with the Pettifor maps mentioned earlier, where unknown stable structures are inferred from known stable ones. Structure analogy is essentially a classification problem that predicts the crystal's macro-structure types given particular conditions (e.g. type of elements and compositions). Advanced ML techniques have been developed to systematically tackle such structure analogy type CSP problems. Curtarolo *et al.* [71] developed a data mining workflow for predicting binary alloy structure types using principal component analysis (PCA) and partial least squares (PLS). Fischer *et al.* [88] developed an informatics-based structure suggestion model for predicting ground state structure types of a large range of

intermetallics, which was later extended to ternary oxides by Hautier *et al.* [89]. Balachandran *et al.* [90] utilized decision tree and support vector machines (SVM) to classify a multitude of wide band gap AB compounds as well as RM intermetallics. Pilania *et al.* [91] built an SVM based classifier to predict the formability of a given ABX_3 halide composition in the perovskite crystal structure. Oliynyk *et al.* [92] applied PLS discriminant analysis (PLS-DA) and SVM towards the CSP of binary AB compounds. Oliynyk *et al.* [93] further developed a method that utilizes cluster resolution feature selection (CR-FS) and SVM classification for the CSP of equiatomic ternary compositions based only on the identity of their constituent elements. Yamashita *et al.* [94] extended the traditional classification problem by generating the pool of possible structure types on-the-fly and selecting the best class with Bayesian optimization (BO) iteratively. Takahashi and Takahashi [95] used a random forest classification model to predict the crystal structure of alloys and oxides. Liang *et al.* [96••] developed the *Crystal Structure Prediction Network* (CRYSPNet), a predictor of the applicable Bravais lattice, space group, and lattice parameters of an inorganic material based only on the latter's chemical composition.

Although significant progress has been made in using data-driven methods to classify/select candidate structure types, it is essentially a simplified version of the original CSP problem, where the ultimate aim is to obtain the atomic-level stable structure of crystal materials. To fully realize such a goal, the ML model should reconstruct the crystal structure from the representation of the material. Recently, generative ML architectures have attracted research interest and been utilized as a novel approach for CSP. Generative models (GMs) for CSP are unsupervised ML models that learn a low dimensional representation from a high dimensional structural design space and generate new structures using knowledge embedded in the latent space. The key to successful CSP with GMs relies on an efficient and invertible representation of the crystal design space, preferably with a one-to-one mapping between the representation and the structure design space. Nouira *et al.* [97•] first applied a generative adversarial network (GAN) architecture for CSP and developed the *CrystalGAN* code, which generates ternary stable crystallographic structures from observed binary structures. Noh *et al.* [98•] proposed a variational autoencoder (VAE) based crystal structure generator with 3D image-based invertible input representation. Hoffman *et al.* [99] developed a general-purpose VAE model based on a 3D atomic density representation. Kim *et al.* [100] built a GAN for CSP that utilizes a representation consisting of unit cell parameters and fractional atomic coordinates. Evidently, GMs have found broad application in organic CSP [101]. In fact, GMs can achieve more than CSP as they could be applied towards the general inverse design of materials by adding a target function as a *condition* [102,103].

Mathematical optimization

The search for the most stable crystal structure in inorganic CSP studies is a global optimization problem with a vast search space. As illustrated in the previous sections, meta-heuristic search algorithms and data-driven methods have been efficient approaches for tackling certain CSP instances. However, outside the field of CSP, a widely accepted paradigm for solving global optimization problems is mathematical optimization (MO), which has not yet been widely applied in this context. An MO model expresses the optimization problem via an objective and a set of constraints, and solves this model with well-established algorithms (e.g. combinatorial and/or spatial branch-and-bound) that can return mathematically proven global optimal solutions. The unique advantage of MO solvers that distinguishes them from meta-heuristic search methods and data-driven methods is that it is possible to obtain information about the quality of the solution; that is, a certificate of whether the solution is globally optimal, or else how far it is estimated to be from a possible global optimum (a.k.a. the *optimality gap)*, which can vary significantly depending on the specific problem.

The earliest attempt to apply formal MO tools to inorganic CSP problems is the nanocluster study by Maranas and Floudas [104]. In this work, the authors first transformed the non-convex potential energy function to the difference between two convex functions via standard model transformation techniques often applied in the MO field. Then they formulated the search for the minimum of the potential energy surface as an optimization problem for which they then developed a decomposition type algorithm to find the globally optimal solution. Although this algorithm could elegantly find the global optimal solution, it was only efficient for very small problems (i.e. particle size less than 24). The computational cost of MO methods is generally higher than meta-heuristic methods and data-driven methods, which is one of the main reasons why MO has not been widely applied in inorganic CSP search problems to-date. In contrast, due to the generally smaller design space of certain organic materials, there has been a lot more MO studies in the field of organic CSP [105–111,112•]. We note that MO methods have also been extensively used in computer-aided molecule/solvent design problems (see, e.g. [113••,114,115•]). Another major obstacle to applying MO to the inorganic CSP problem is that MO generally uses predefined algebraic form objectives and constraints, which are not naturally compatible with first-principle calculations or machined-learned models.

Despite the limitations stated above, we believe that MO can be a viable alternative approach for inorganic CSP search problems with potential unique advantages. From a practical point of view, the computational cost of MO methods is constantly decreasing with many major recent

advancements of optimization algorithms. Furthermore, we can incorporate non-algebraic structure-function relationships, such as *ab initio* calculations and machine-learned models, into an MO framework through surrogate modeling techniques. We can also apply specialized modeling tricks (e.g. indicator constraints) to encode non-algebraic structure-function information (e.g. scenarios, conformations, active sites). As previously stated, the unique advantage of MO is the ability to determine the solution's optimality. Notably, since MO carries out a systematic search over the entire design space, it has a high potential to result in non-intuitive designs and expose previously undiscovered trends. Furthermore, MO is a relatively flexible paradigm that readily allows for modifications to the objective and/or constraints considered in the model, enabling us to go beyond the core CSP problem (i.e. stability) and explore the design space in various ways.

In recent years, we have developed a crystal materials design paradigm that is based on mixed-integer linear programming (MILP), a popular subclass of MO. We simplify the design space of a crystal material into a set of discrete locations that can be occupied by building blocks (e.g. atoms, secondary building units, molecules). In a nutshell, we start with a canvas of possible locations that may be occupied by a building block, and introduce binary decision variables to encode the design choice of whether or not to place a building block of a certain type in each location. With this definition, a crystal materials design essentially has a one-to-one mapping to a set of binary decision variables. We can then formulate a structure-function relationship that links material properties and decision variables within an MILP model that can be solved with powerful existing numerical methods. We have applied this design paradigm to a variety of material systems and applications. Hanselman and Gounaris [116••] first introduced the paradigm and demonstrated its application in designing two-dimensional periodic catalytic surface patterns using a coordination number based structure-function relationship. This work was later extended into a multi-objective optimization framework to explicitly account for the stability of the catalytic surfaces [117]. In addition, a conformation-based model was used as the basis for a perovskite design framework, in which first principle calculations were incorporated via advanced regression techniques [118,119•]. Finally, Isenberg *et al.* [120] applied an MILP-based paradigm to non-periodic nanocluster systems and showed how to design highly cohesive monometallic nanoclusters. Yin *et al.* [121] further extended the methodologies to bimetallic nanoclusters, demonstrating also how MILP could be combined with metaheuristic search algorithms. The concepts and methodologies associated with the MO paradigm can also be applied on one-dimensional material systems. We shall illustrate this in the below case study, where we will apply this paradigm in the context of a nanowires CSP problem.

Case study: Semiconductor nanowires CSP

Semiconductor nanowires (NWs) exhibit unique physical properties due to their nanoscale sizes and reduced dimensionality. They are predicted to play a key role as fundamental building blocks in next-generation optical, electronic and catalytic devices and systems. Crystal structures are crucial in semiconductor NW research as they affect the quantum and electronic properties of the NWs. Interestingly, NWs can form different crystal structures from their bulk counterparts and exhibit phase changes across orientations and sizes. In our case study, we will illustrate how mathematical optimization could handle the semiconductor NWs CSP problem. To formulate a CSP optimization model, one needs to define a proper optimization objective as a stability indicator. Following the literature in this field, we choose the per-atom cohesive energy (E_{coh}) as the design objective for our optimization model. The cohesive energy is defined as the average energy difference between infinitely separated neutral metal atoms and the crystalline nanostructures formed by those atoms [122]. It measures the strength of interatomic bonding between atoms, and is thus often used to indicate the stability of nanomaterials. In the following, we will introduce the structure-function relationship we used for determining the cohesive energy of any given NW structure. Subsequently, we will illustrate how to formulate a mathematical optimization model that accounts for this structure-function relationship. Finally, we will briefly discuss our computational experiments and results.

Structure-function relationship. The first step of applying MO to materials design problems is to identify the structure-function relationship, that is, a relationship between a nanostructure and its functionality, for which in this case study we will consider E_{coh} to be the functionality of interest. In particular, we will approximate E_{coh} with the sum of pairwise potential energies, averaged by the number of atoms in the system, as illustrated in Equation 1, where N refers to the number of atoms and B is the complete set of bonds. We will adopt the Khor-Das Sarma (KDS) empirical potential energy [123], which is a function of the interatomic distance r_{ij} between two atoms i and j. We highlight that the KDS potential function has been applied to various elemental and compound semiconductor systems and showed good agreement with experimental/*ab initio* calculation results (see, e.g. [124–127]). In Equation 2, A, B, α, β, γ, θ, and λ are all empirical parameters that depend on the semiconductor type. The symbol $r_{i,\min}$ denotes the distance between atom i and its nearest neighbor, while Z_i is the effective coordination number of atom i. This structure-function relationship applies not only to elemental NWs, but also

to compound NWs when considering perfect lattice geometries, where each lattice point is associated with a fixed type of atom:

$$E_{\text{coh}} = \frac{1}{N} \sum_{(i-j)\in B} E_{\text{coh}}^{ij} \tag{1}$$

$$E_{\text{coh}}^{ij} = Ae^{-\beta(r_{ij}-r_{i,\min})^{\gamma}} \left[\frac{Be^{-\lambda r_{ij}}}{Z_i^{\alpha}} - e^{-\theta r_{ij}} \right] \tag{2}$$

As mentioned previously, we will define our design space as a set of discrete locations for building blocks, which are simply atoms in this case study. Under this assumption, any distance r_{ij} would attain a value from a discrete set of values associated with the lattice type of choice. The possible values of r_{ij} will be dependent on which layer n of atom's i neighbors does atom j belong to. Given this, the value of $r_{i,\min}$ will be a fixed parameter that depends on the lattice constant. Thus, the structure-relationship can be discretized and reformulated into a summation of individual atom contributions as:

$$E_{\text{coh}} = \frac{1}{N} \sum_{n} \sum_{i} f_i^n(CN_i^n) \tag{3}$$

$$f_i^n(CN_i^n) = p^n (CN_i^n)^{1-\alpha} - q^n CN_i^n \tag{4}$$

$$p^n = ABe^{(\alpha-1)\beta(r^n-r_{\min})^{\gamma}-\lambda r^n} \tag{5}$$

$$q^n = Ae^{-\beta(r^n-r_{\min})^{\gamma}-\theta r^n} \tag{6}$$

where CN_i^n is the coordination number of atom i at the nth layer of neighbors, while p^n and q^n are suitably congregated parameters. This simplified structure-function relationship can then serve as the objective of an optimization model.

Mathematical modeling. Let our NW design canvas be I, consisting of a set of discrete locations from a perfect lattice without defects. For each location $i \in I$, we denote with $N_i^n \subset I$ all locations that are nth layer neighbors of i, taking into account the applicable periodicities, as dictated by the lattice type. The set of all neighboring layers considered is denoted as L. We then introduce binary decision variables y_i to indicate the occupancy of an atom (of the corresponding elemental type) at each location i. Specifically, $y_i = 1$ encodes the existence of an atom at canvas location i, while $y_i = 0$ means that location i does not contain an atom. Thus, a set of decision variables y_i represents a unique NW design within the specified canvas. Additionally, we define integer variables cn_i^n to represent the coordination number of i at n-th layer neighbors. Finally, auxiliary continuous variables v_i^n are utilized for encoding atom at location i's contribution to the overall cohesion. The search for the most stable NW design can now be equivalently viewed as the identification of a set of decision variable values that are optimal with respect to the objective and that satisfy a number of constraints, as presented in the below optimization model:

$$\max_{y_i, cn_i^n, v_i^n} \quad \frac{1}{N} \sum_{n\in L} \sum_{i\in I} v_i^n \tag{7}$$

$$\text{s.t.} \quad \sum_{i\in I} y_i = N \quad \forall i \in I \tag{8}$$

$$y_i \Rightarrow \{cn_i^n = \sum_{j\in N_i^n} y_j\} \quad \forall i \in I \quad \forall n \in L \tag{9}$$

$$y_i \Rightarrow \{cn_i^n \geq CN_{\min}^n\} \quad \forall i \in I \quad \forall n \in L \tag{10}$$

$$\neg y_i \Rightarrow \{cn_i^n \leq 0\} \quad \forall i \in I \quad \forall n \in L \tag{11}$$

$$v_i^n = f_i^n(cn_i^n) \quad \forall i \in I \, \forall n \in L \tag{12}$$

$$y_i \in \{0, 1\} \quad \forall i \in I \tag{13}$$

$$cn_i^n \in \{1, 2, 3, \ldots, CN_{\max}^n\} \quad \forall i \in I \, \forall n \in L \tag{14}$$

$$0 \leq v_i^n \leq V_{\max}^n \quad \forall i \in I \quad \forall n \in L \tag{15}$$

Equation 7 is the objective function, which implements the cohesive energy function from Equation 3. Here, values of $f_i^n(CN_i^n)$ are encoded by auxiliary variables v_i^n. Equation 8 defines the number of atoms to be occupied in each periodic repetition of the canvas as being N, which is an integer parameter input to this model. Equations 9–11 define auxiliary variables cn_i^n to represent the number of layer-n neighbors of each location i. Note that this variable shall attain the value of zero whenever there is no atom occupying location i, while when i is occupied,

it will be required to be at least greater than some predefined number, $CN^n_{\min}$, to avoid extremely low coordinated atoms. Equation 12 define auxiliary variables v^n_i to be equal to $f^n_i(CN^n_i)$, where the latter is taken as in Equation 4. Finally, Equations 13–15 declare the domains of all variables, with $CN^n_{\max}$ and $V^n_{\max}$ being the upper bounds of variables cn^n_i and v^n_i, respectively.

The above optimization model belongs to a class of optimization problems called mixed-integer nonlinear program (MINLP) and can be readily solved with state-of-the-art global optimization solvers. Since the only nonlinear components in this optimization is the function $f^n_i(cn^n_i)$, which depends on a discrete integer variable cn^n_i, we could further simplify this model by replacing Equation 12 using standard piecewise linear (PWL) reformulations that result in mixed-integer linear constraints. Importantly, the new PWL-reformulated optimization will be a mixed-integer linear program (MILP) that is typically more tractable, as it is amenable to be solved with well-established commercial codes (e.g. CPLEX [128]). Note that, due to the discrete nature of cn^n_i, the PWL formulation will be able to precisely represent $f^n_i(cn^n_i)$ values at those integer points. For details on PWL reformulations, readers are referred to our previous work on designing bimetallic nanoclusters [121].

We note that the curvature of the function $f^n_i(cn^n_i)$ depends on the values of parameters of α and p^n. In cases when $f^n_i(cn^n_i)$ is a concave function, we could capitalize on additional model simplifications via direct linearization without the need for PWL reformulations. This is accomplished by modeling $f^n_i(cn^n_i)$ as a set of secant lines passing through points corresponding to integer values of cn^n_i. Specifically, Equation 12 can be replaced with $v^n_i \le s^n_p cn^n_i + t^n_p$, for all $i \in I$, $n \in L$, and $p \in \{1, 2, 3, \ldots, CN\max\}$, where p are the indices of secant lines, while s^n_p and t^n_p are respectively the slopes and intercepts of those lines. As we are maximizing the cohesive energy, the optimizer will choose the exact value on its corresponding secant line, as it is the maximally attainable value permitted by the inequality. For details on how to treat a concave function under a maximization setting, readers are referred to our previous work on designing monometallic nanoclusters [120].

The above mentioned modeling and reformulation procedures are popular techniques practiced in the process system engineering community in contexts of process design. However, we recognize that familiarity with such techniques might not be as prevalent in the broader CSP research community, which represents a significant barrier to adopting an MO paradigm for CSP. To bridge the knowledge gap and simplify the MO modeling and implementation process, we have developed a Python toolkit called *MatOpt* to automate many aspects of this process. The *MatOpt* toolkit[2] uses materials research-inspired syntax and hides the mathematical modeling and numerical optimization details from its users. It is distributed as part of the IDAES-PSE [129] package[3] and is freely available. Along with the software distribution, we have provided several Jupyter notebook examples[4] to demonstrate various use cases for this toolkit. In this case study, we will use *MatOpt* to instantiate and solve NW design optimization problems in accordance to the models described above.

Computational study: setup. For our computational study, we shall focus on III-V compound semiconductor NWs, which constitute an important class of NW systems. They show unique properties, including controllable bandgap, high carrier mobility, great mechanical flexibility, and large surface-to-volume ratio, making them good candidates for next-generation electronics, photonics, and sensors. Contrary to the fact that bulk III-V systems usually adopt the cubic zinc blende lattice geometry, III-V NWs have been found to also exhibit a hexagonal wurtzite (WZ) geometry, depending on their size, orientation, and synthesis conditions. Research efforts have been devoted to controllable synthesis and tailoring of pure phase III-V NWs as well as advanced device/system design based on heterostructures of mixed III–V NW phases [130,131]. To that end, the ability to predict the stable crystal structures of III–V NWs would benefit both experimental synthesis and computational design studies in this field.

More specifically, we choose to illustrate our methodology with InAs NWs, a representative III–V system. The main characteristic of a NW's structure is its large length-to-width ratio, which we achieve by defining a canvas along a predefined direction. In particular, the subsets with neighboring locations are defined taking into account that the canvas repeats periodically along the growth orientation. Given the WZ lattice geometry, we consider three common orientations, namely 0001, $1\bar{1}00$ and $11\bar{2}0$ in hexagonal Bravais-Miller indices.

For the periodic length of the canvas along the growth orientation, we have chosen a multiple of the smallest periodic unit length for each orientation. For example, the smallest periodic unit length for the 0001 orientation of a WZ lattice is $3.771 \times$ IAD, where IAD is the crystal's interatomic distance, and we specifically use four times this value as the periodic length of our NW, that is,

[2] https://idaes-pse.readthedocs.io/en/stable/user_guide/modeling_extensions/matopt/index.html.

[3] https://idaes-pse.readthedocs.io/en/stable/getting_started/index.html.

[4] https://idaes.github.io/examples-pse/latest/Examples/MatOpt/index.html.

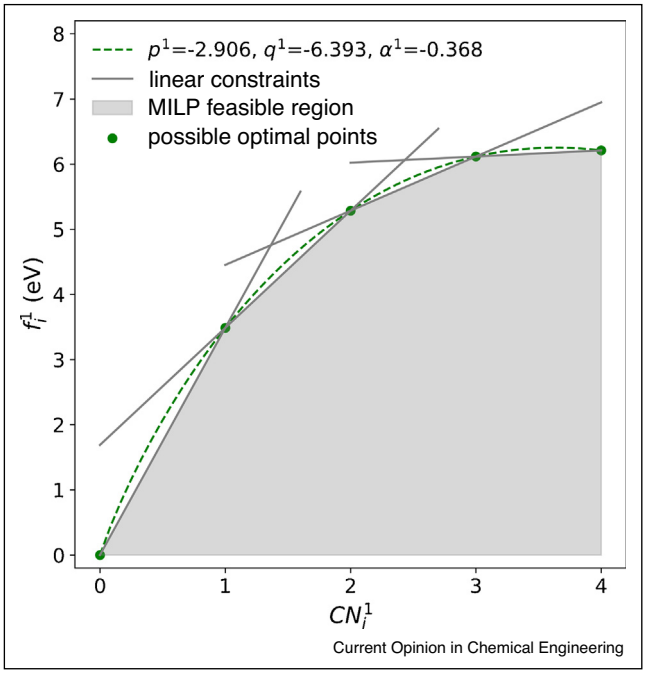
13.353 × IAD. Furthermore, the canvas should extend radially outwards from the growth orientation axis of the NW. We do so by layering atomic locations, following the applicable lattice geometry. A question arises as to how many layers of atom locations should be specified. As demonstrated in our previous works, we need to carefully decide the size of the canvas for an optimal trade-off between model tractability and solution optimality. Too small a canvas may lead to a suboptimal design limited by the canvas boundary. At the same time, too large a canvas will cause numerical tractability issues and we might not be able to get the optimal solution within the time limit. In this study, for each given size N, we will iteratively increase the canvas size (i.e. increase layers of locations around the axis) and solve the optimization problem until no atoms in the resulting design are on the outermost layer of atom locations (i.e. the canvas boundary). To eliminate the translational symmetries (that are perpendicular to the axis), as well as to encourage a solution that distributes along the axis, we fix the occupancy of the 'core' of the NW (i.e. atom locations on or near the axis) as well as we enforce a radial growth of atoms using constraints similar to those described in [116••].

The parameters of the KDS potential function are taken as $A = -709.003$ eV, $B = 1.978$, $\alpha = -0.368$, $\beta = 12.028$, $\gamma = 3.202$, $\theta = 1.794$ Å^{-1}, $\lambda = 2.355$ Å^{-1}, as suggested in the literature for InAs NWs [132]. The lattice constants of the ideal WZ geometry are calculated from the cubic bulk lattice constant, $a_{\text{cubic}} = 6.058$ Å, via simple geometric conversion; that is, $a = \frac{\sqrt{2}}{2} a_{\text{cubic}} = 4.284$ Å and $c = \frac{2\sqrt{3}}{3} a_{\text{cubic}} = 6.996$ Å. The InAs NW's IAD associated with the calculated lattice constants is IAD = 2.624 Å. Using the above information, we then calculate the discrete distances r^n and $r_{\min}$, as well as the intermediate model parameters p^n and q^n according to Equations 5 and 6, respectively. An important parameter to determine is the range of n, which indexes over the neighboring layers considered in the model (set L). For this, we calculated and compared the value range of the function $f_i^n(cn_i^n)$ for different n and found that, in the InAs NW system of interest, the first layer contribution dominates the overall value. For example, when comparing the first layer and the second layer, we found that $\frac{f_i^1(cn_i^1)}{f_i^2(cn_i^2)} > 10^{12}$. Thus, we can simplify the optimization model further by considering only the first layer of neighbors around each location in our canvas. In regards to parameters of the function $f_i^1(cn_i^1)$, we used the values $p^1 = -2.906$, $q^1 = -6.393$, $CN^1_{\max} = 4$, $CN^1_{\min} = 2$, $V^1_{\max} = 6.212$, and $\alpha = -0.368$. With this set of parameters, the function is concave and, as mentioned previously, it can be expressed without loss of accuracy using several linear constraints (see Figure 1), reducing our optimization model to an MILP model.

Computational study: Implementation and results. We utilized the *MatOpt* toolkit [133] to facilitate the implementation of our NW design model. The pertinent code can be found online in the form of a Jupyter notebook.[5] For detailed descriptions of *MatOpt* concepts and functionalities, we refer the users to the online documentation. It suffices to say that the toolkit shall generate appropriate Pyomo model objects that are readily solvable by any MILP solver accessible via the Pyomo modeling library [134,135]. In our case, we instruct *MatOpt* to invoke the well-established commercial solver CPLEX [128].

For each selected orientation, we run the optimization model for various settings of N and obtain optimal (i.e. relative optimality gap less than 0.5%) solutions in each case. The optimal objective values are presented in Figure 2, noting that in order to facilitate a fair comparison across all orientations, the results are displayed as a function of the number of atoms per NW length equal to the lattice's IAD. In all cases, we observe that the per-atom cohesive energy exhibits an increasing trend, asymptotically approaching the bulk value as the NW becomes wider. This suggests that InAs NWs are less stable than their bulk counterparts, which agrees with the literature. Furthermore, for a given NW size, the 0001 orientation is more advantageous in terms of cohesive

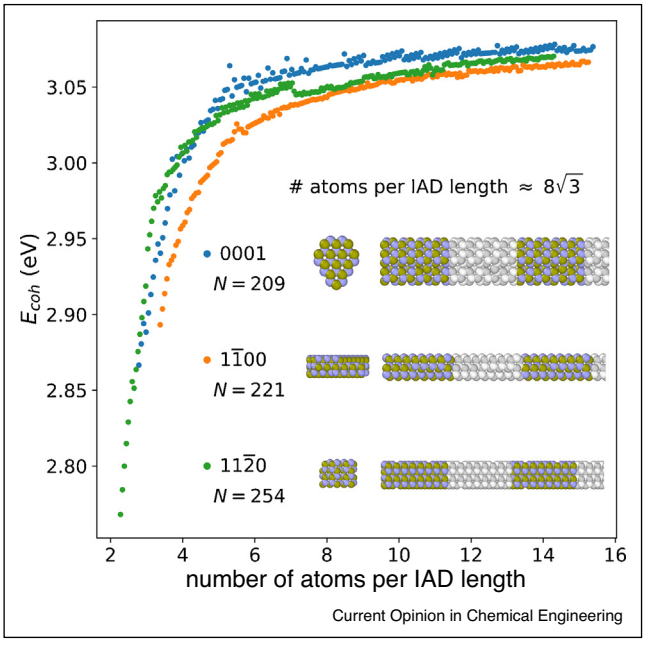
Figure 1



Modeling the structure-function relationship with a set of linear constraints (secant lines).

[5] https://github.com/IDAES/examples-pse/blob/main/src/Examples/MatOpt/nanowire_design.ipynb.

Figure 2



Optimal cohesive energy values at different NW sizes, along with example optimal structures, for three different orientations of WZ lattice geometry. To aid in the visual illustration of the example structures, a cross-sectional view as well as a side view are provided. The purple color represents In atoms, while the yellow color represents the As atoms. Every other periodic repetition of the canvas is gray-scaled so as to better convey the overall NW shape.

energy than the other two studied orientations, which agrees with both computational and experimental results in the literature [136,137]. Aside from such general trends that can be inferred from our computational results, the obtained solutions are useful inasmuch as they can serve as model NW structures to guide further research effort. In this way, the MO-based NW design optimization model has the potential to be a simple and efficient tool in assisting semiconductor NW CSP research, complementing the other approaches. On that note, Figure 2 also illustrates some example solutions that can be obtained via the approach used in our study.

Conclusions

Owing to the inherent combinatorial complexity of an inorganic material's design space, an efficient search algorithm is the key to a successful CSP methodology. This paper reviewed past studies as well as recent progress on this topic. Guided-sampling methods and data-driven methods constitute the majority of the methodologies developed to date. The main difference between the two approaches is that guided-sampling methods make self-improving predictions iteratively based on an explicit structure-function relationship, while data-driven methods learn the optimal result from data directly. In addition, we reviewed a mathematical optimization-based materials design framework, which can serve as an alternative approach for CSP. We then presented an example application of this framework to design highly cohesive semiconductor nanowires, demonstrating the concepts and procedures of the approach as well as the toolkit we have developed to accelerate and automate this process. We opine that there is currently no dominant search method for inorganic materials CSP and that, given today's explosion of new materials discovery, the many diverse search paradigms currently in existence position the field well for a bright future.

Disclaimer

This paper was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Conflict of interest statement

In spite of his role as a Guest Editor of this special issue, Chrysanthos E Gounaris had no involvement in the peer-review of this article and had no access to information regarding its peer-review. Full responsibility for the editorial process for this article was delegated to Professors Rafiqul Gani and Lei Zhang.

Acknowledgments

We graciously acknowledge funding from the U.S. Department of Energy, Office of Fossil Energy's Crosscutting Research Program through the Institute for the Design of Advanced Energy Systems.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- • of special interest

1. Xie T, Grossman JC: **Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties**. *Phys Rev Lett* 2018, **120**:145301.

2. •• Oganov AR, Pickard CJ, Zhu Q, Needs RJ: **Structure prediction drives materials discovery**. *Nat Rev Mater* 2019, **4**:331-348
This paper reviews CSP methods and presents their potential for discovery of various novel materials.

3. • Oganov AR, Glass CW: **Crystal structure prediction using ab initio evolutionary techniques: principles and applications**. *J Chem Phys* 2006, **124**:244704

This paper introduces the *universal structure predictor: evolutionary xtallography* (USPEX), a state-of-the-art evolutionary algorithm-based CSP methodology.

4. •• Bowskill DH, Sugden IJ, Konstantinopoulos S, Adjiman CS, Pantelides CC: **Crystal structure prediction methods for organic molecules: State of the art**. *Annu Rev Chem Biomol Eng* 2021, **12**

This article systematically reviews state-of-the-art CSP methodologies for organic molecules as well as provides insights for future directions.

5. Pickard CJ, Needs RJ: **High-pressure phases of silane**. *Phys Rev Lett* 2006, **97**:045504.

6. Pickard CJ, Needs RJ: **Structure of phase III of solid hydrogen**. *Nat Phys* 2007, **3**:473-476.

7. Pickard CJ, Needs RJ: **Highly compressed ammonia forms an ionic crystal**. *Nat Mater* 2008, **7**:775-779.

8. Pickard CJ, Needs RJ: **Aluminium at terapascal pressures**. *Nat Mater* 2010, **9**:624-627.

9. • Pickard CJ, Needs RJ: **Ab initio random structure searching**. *J Phys: Condens Matter* 2011, **23**:053201

This article describes the simple yet efficient *ab initio random structure searching* (AIRSS) CSP metaheuristic and summarizes its applications for various material systems.

10. Day GM, Cooper TG, Cruz-Cabeza AJ, Hejczyk KE, Ammon HL, Boerrigter SXM, Tan JS, Raffaele G, Valle D, Venuti E, Jose J *et al.*: **Significant progress in predicting the crystal structures of small organic molecules-a report on the fourth blind test**. *Acta Crystallogr Sect B: Struct Sci* 2009, **65**:107-125.

11. Bardwell DA, Adjiman CS, Arnautova YA, Bartashevich E, Boerrigter SXM, Braun DE, Cruz-Cabeza AJ, Day GM, Della Valle RG, Desiraju GR *et al.*: **Towards crystal structure prediction of complex organic compounds — a report on the fifth blind test**. *Acta Crystallogr Sect B: Struct Sci* 2011, **67**:535-551.

12. Kirkpatrick S, Daniel Gelatt C, Vecchi MP: **Optimization by simulated annealing**. *Science* 1983, **220**:671-680.

13. Deem MW, Newsam JM: **Determination of 4-connected framework crystal structures by simulated annealing**. *Nature* 1989, **342**:260-262.

14. Pannetier J, Bassas-Alsina J, Rodriguez-Carvajal J, Caignaert V: **Prediction of crystal structures from crystal chemistry rules by simulated annealing**. *Nature* 1990, **346**:343-345.

15. Schön JC, Jansen M: **Determination of candidate structures for Lennard-Jones-crystals through cell optimisation**. *Berich Bunsengesellsch Phys Chem* 1994, **98**:1541-1544.

16. Christian Schön J, Jansen M: **First step towards planning of syntheses in solid-state chemistry: determination of promising structure candidates by global optimization**. *Angew Chem* 1996, **35**:1286-1304.

17. Finnila AB, Gomez MA, Sebenik C, Stenson C, Doll JD: **Quantum annealing: a new method for minimizing multidimensional functions**. *Chem Phys Lett* 1994, **219**:343-348.

18. Reinaudi L, Carbonio RE, Leiva EPM: **Inclusion of symmetry for the enhanced determination of crystalline structures from powder diffraction data using simulated annealing**. *Chem Commun* 1998:255-256.

19. Reinaudi L, Leiva EPM, Carbonio RE: **Simulated annealing prediction of the crystal structure of ternary inorganic compounds using symmetry restrictions**. *Dalton Trans* 2000, **23**:4258-4262.

20. Wales DJ, Doye JPK: **Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms**. *J Phys Chem A* 1997, **101**:5111-5116.

21. Wales DJ, Scheraga HA: **Global optimization of clusters, crystals, and biomolecules**. *Science* 1999, **285**:1368-1372.

22. Mellot-Draznieks C, Newsam JM, Gorman AM, Freeman CM, Férey G: **De novo prediction of inorganic structures developed through automated assembly of secondary building units (aasbu method)**. *Angew Chem* 2000, **39**:2270-2275.

23. Mellot-Draznieks C, Girard S, Férey G, Christian Schön J, Cancarevic Z, Jansen M: **Computational design and prediction of interesting not-yet-synthesized structures of inorganic materials by using building unit concepts**. *Chemistry* 2002, **8**:4102-4113.

24. Mellot-Draznieks C, Girard S, Férey G: **Novel inorganic frameworks constructed from double-four-ring (d4r) units: computational design, structures, and lattice energies of silicate, aluminophosphate, and gallophosphate candidates**. *J Am Chem Soc* 2002, **124**:15326-15335.

25. Mellot-Draznieks C, Dutour J, Férey G: **Hybrid organic-inorganic frameworks: routes for computational design and structure prediction**. *Angew Chem* 2004, **116**:6450-6456.

26. • Goedecker S: **Minima hopping: an efficient search method for the global minimum of the potential energy surface of complex molecular systems**. *J Chem Phys* 2004, **120**:9911-9917

This article describes *minima hopping*, a novel trajectory-based CSP metaheuristic that improves the search efficiency by avoiding re-visiting known parts of the configurational space.

27. Goedecker S, Hellmann W, Lenosky T: **Global minimum determination of the born-oppenheimer surface within density functional theory**. *Phys Rev Lett* 2005, **95**:055501.

28. Hellmann W, Hennig RG, Goedecker S, Umrigar CJ, Delley B, Lenosky T: **Questioning the existence of a unique ground-state structure for si clusters**. *Phys Rev B* 2007, **75**:085411.

29. Bao K, Goedecker S, Koga K, Lançon F, Neelov A: **Structure of large gold clusters obtained by global optimization using the minima hopping method**. *Phys Rev B* 2009, **79**:041405.

30. Amsler M, Goedecker S: **Crystal structure prediction using the minima hopping method**. *J Chem Phys* 2010, **133**:224104.

31. Amsler M, Botti S, Marques MAL, Goedecker S: **Conducting boron sheets formed by the reconstruction of the α-boron (111) surface**. *Phys Rev Lett* 2013, **111**:136101.

32. Amsler M, Botti S, Marques MAL, Lenosky TJ, Goedecker S: **Low-density silicon allotropes for photovoltaic applications**. *Phys Rev B* 2015, **92**:014101.

33. Borlido P, Steigemann C, Lathiotakis NN, Marques MAL, Botti S: **Structural prediction of two-dimensional materials under strain**. *2D Mater* 2017, **4**:045009.

34. Borlido P, Huran AW, Marques MAL, Botti S: **Structural prediction of stabilized atomically thin tin layers**. *NPJ 2D Mater Appl* 2019, **3**:1-5.

35. Sun L, Marques MAL, Botti S: **Direct insight into the structure-property relation of interfaces from constrained crystal structure prediction**. *Nat Commun* 2021, **12**:1-10.

36. Holland JH: **Genetic algorithms**. *Sci Am* 1992, **267**:66-73.

37. Smith RW: **Energy minimization in binary alloy models via genetic algorithms**. *Comput Phys Commun* 1992, **71**:134-146.

38. Bush TS, Rchard C, Catlow A, Battle PD: **Evolutionary programming techniques for predicting inorganic crystal structures**. *J Mater Chem* 1995, **5**:1269-1272.

39. Deaven DM, Ho K-M: **Molecular geometry optimization with a genetic algorithm**. *Phys Rev Lett* 1995, **75**:288.

40. Johnston RL: **Evolving better nanoparticles: genetic algorithms for optimising cluster geometries**. *Dalton Trans* 2003, **22**:4193-4207.

41. Woodley SM, Sokol AA, Richard C, Catlow A: **Structure prediction of inorganic nanoparticles with predefined architecture using a genetic algorithm**. *Zeit Anorganisch Allg Chem* 2004, **630**:2343-2353.

42. Woodley SM: **Prediction of crystal structures using evolutionary algorithms and related techniques**. *Appl Evol Comput Chem* 2004:95-132.

43. Woodley SM: **Engineering microporous architectures: combining evolutionary algorithms with predefined exclusion zones**. *Phys Chem Chem Phys* 2007, **9**:1070-1077.

44. Lloyd LD, Johnston RL, Salhi S: **Strategies for increasing the efficiency of a genetic algorithm for the structural optimization of nanoalloy clusters**. *J Comput Chem* 2005, **26**:1069-1078.

45. Glass CW, Oganov AR, Hansen N: **Uspex-evolutionary crystal structure prediction**. *Comput Phys Commun* 2006, **175**:713-720.

46. Trimarchi G, Zunger A: **Global space-group optimization problem: finding the stablest crystal structure without constraints**. *Phys Rev B* 2007, **75**:104113.

47. Trimarchi G, Freeman AJ, Zunger A: **Predicting stable stoichiometries of compounds via evolutionary global space-group optimization**. *Phys Rev B* 2009, **80**:092101.

48. Froltsov VA, Reuter K: **Robustness of 'cut and splice' genetic algorithms in the structural optimization of atomic clusters**. *Chem Phys Lett* 2009, **473**:363-366.

49. Woodley SM, Catlow CRA: **Structure prediction of titania phases: implementation of darwinian versus lamarckian concepts in an evolutionary algorithm**. *Comput Mater Sci* 2009, **45**:84-95.

50. Lonie DC, Zurek E: **Xtalopt: an open-source evolutionary algorithm for crystal structure prediction**. *Comput Phys Commun* 2011, **182**:372-387.

51. Kennedy J, Eberhart R: **Particle swarm optimization**. In *Proceedings of ICNN'95-International Conference on Neural Networks, vol 4*. IEEE; 1995:1942-1948.

52. •• Wang Y, Lv J, Zhu L, Ma Y: **Crystal structure prediction via particle-swarm optimization**. *Phys Rev B* 2010, **82**:094116
This paper was the first to propose the application of a particle-swarm optimization algorithm to crystal structure prediction.

53. Wang Y, Lv J, Zhu L, Ma Y: **CALYPSO: a method for crystal structure prediction**. *Comput Phys Commun* 2012, **183**:2063-2070.

54. Lyakhov AO, Oganov AR, Stokes HT, Zhu Q: **New developments in evolutionary structure prediction algorithm USPEX**. *Comput Phys Commun* 2013, **184**:1172-1182.

55. Wu SQ, Ji M, Wang C-Z, Nguyen MC, Zhao X, Umemoto K, Wentzcovitch RM, Ho K-M: **An adaptive genetic algorithm for crystal structure prediction**. *J Phys: Condens Matter* 2013, **26**:035402.

56. Hu J, Yang W, Dong R, Yuxin L, Li X, Li S, Siriwardane EMD: **Contact map based crystal structure prediction using global optimization**. *CrystEngComm* 2021, **23**:1765-1776.

57. Pauling L: **The principles determining the structure of complex ionic crystals**. *J Am Chem Soc* 1929, **51**:1010-1026.

58. Pauling L *et al.*: *The Nature of the Chemical Bond, vol 260*. Ithaca, NY: Cornell University Press; 1960.

59. Villars P: **A three-dimensional structural stability diagram for 998 binary ab intermetallic compounds**. *J Less Common Metals* 1983, **92**:215-238.

60. Villars P: **A three-dimensional structural stability diagram for 1011 binary ab2 intermetallic compounds: II**. *J Less Common Metals* 1984, **99**:33-43.

61. Villars P: **A semiempirical approach to the prediction of compound formation for 3486 binary alloy systems**. *J Less Common Metals* 1985, **109**:93-115.

62. Villars P: **A semiempirical approach to the prediction of compound formation for 96446 ternary alloy systems: II**. *J Less Common Metals* 1986, **119**:175-188.

63. Pettifor DG: **The structures of binary compounds. I. phenomenological structure maps**. *J Phys C: Solid State Phys* 1986, **19**:285.

64. Pettifor DG: **Structure maps for. pseudobinary and ternary phases**. *Mater Sci Technol* 1988, **4**:675-691.

65. Bergerhoff G, Brown ID, Allen F *et al.*: **Crystallographic databases**. *Int Union Crystallogr* 1987, **360**:77-95.

66. Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G, Persson KA: **The Materials Project: a materials genome approach to accelerating materials innovation**. *APL Mater* 2013, **1**:011002 http://dx.doi.org/10.1063/1.4812323 2166532X.

67. Saal JE, Kirklin S, Aykol M, Meredig B, Wolverton C: **Materials design and discovery with high-throughput density functional theory: the open quantum materials database (oqmd)**. *J Miner Metals Mater Soc* 2013, **65**:1501-1509.

68. Curtarolo S, Setyawan W, Hart GLW, Jahnatek M, Chepulskii RV, Taylor RH, Wang S, Xue J, Yang K, Levy O *et al.*: **Aflow: an automatic framework for high-throughput materials discovery**. *Comput Mater Sci* 2012, **58**:218-226.

69. Christian Schön J: **How can databases assist with the prediction of chemical compounds?** *Zeit Anorg Allg Chem* 2014, **640**:2717-2726.

70. Hofmann DWM, Apostolakis J: **Crystal structure prediction by data mining**. *J Mol Struct* 2003, **647**:17-39.

71. Curtarolo S, Morgan D, Persson K, Rodgers J, Ceder G: **Predicting crystal structures with data mining of quantum calculations**. *Phys Rev Lett* 2003, **91**:135503.

72. Behler J, Parrinello M: **Generalized neural-network representation of high-dimensional potential-energy surfaces**. *Phys Rev Lett* 2007, **98**:146401.

73. Meredig B, Agrawal A, Kirklin S, Saal JE, Doak JW, Thompson A, Zhang K, Choudhary A, Wolverton C: **Combinatorial screening for new materials in unconstrained composition space with machine learning**. *Phys Rev B* 2014, **89**:094104.

74. Schütt KT, Glawe H, Brockherde F, Sanna A, Müller K-R, Gross EKU: **How to represent crystal structures for machine learning: towards fast prediction of electronic properties**. *Phys Rev B* 2014, **89**:205118.

75. Faber F, Lindmaa A, Anatole von Lilienfeld O, Armiento R: **Crystal structure representations for machine learning models of formation energies**. *Int J Quant Chem* 2015, **115**:1094-1101.

76. Isayev O, Oses C, Toher C, Gossett E, Curtarolo S, Tropsha A: **Universal fragment descriptors for predicting properties of inorganic crystals**. *Nat Commun* 2017, **8**:1-12.

77. Schmidt J, Shi J, Borlido P, Chen L, Botti S, Marques MAL: **Predicting the thermodynamic stability of solids combining density functional theory and machine learning**. *Chem Mater* 2017, **29**:5090-5103.

78. Zhou Q, Tang P, Liu S, Pan J, Yan Q, Zhang S-C: **Learning atoms for materials discovery**. *Proc Natl Acad Sci U S A* 2018, **115**:E6411-E6417.

79. Chen X, Chen D, Weng M, Jiang Y, Wei G-W, Pan F: **Topology-based machine learning strategy for cluster structure prediction**. *J Phys Chem Lett* 2020, **11**:4392-4401.

80. Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A, Kim C: **Machine learning in materials informatics: recent applications and prospects**. *NPJ Comput Mater* 2017, **3**:1-13.

81. Haghighatlari M, Li J, Heidar-Zadeh F, Liu Y, Guan X, Head-Gordon T: **Learning to make chemical predictions: the interplay of feature representation, data, and machine learning methods**. *Chemistry* 2020.

82. Ward L, Wolverton C: **Atomistic calculations and materials informatics: a review**. *Curr Opin Solid State Mater Sci* 2017, **21**:167-176.

83. Schmidt J, Marques MRG, Botti S, Marques MAL: **Recent advances and applications of machine learning in solid-state materials science**. *NPJ Comput Mater* 2019, **5**:1-36.

84. • Tong Q, Xue L, Lv J, Wang Y, Ma Y: **Accelerating calypso structure prediction by data-driven learning of a potential energy surface**. *Faraday Discuss* 2018, **211**:31-43

This article proposes to accelerate the CALYPSO code with a state-of-the-art machine learning potential and develops two different acceleration schemes.

85. Deringer VL, Pickard CJ, Csányi G: **Data-driven learning of total**
• **and local energies in elemental boron**. *Phys Rev Lett* 2018, **120**:156001
This paper develops a methodology to combine machine learning with *random structure searching* and applies the method for the systematic construction of boron interatomic potential.

86. Podryabinkin EV, Tikhonov EV, Shapeev AV, Oganov AR:
•• **Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning**. *Phys Rev B* 2019, **99**:064114
This article introduces a methodology that actively learns interatomic potentials and combines it with the USPEX code for efficient and accurate CSP.

87. Jennings PC, Lysgaard S, Hummelshøj JS, Vegge T, Bligaard T: **Genetic algorithms for computational materials discovery accelerated by machine learning**. *NPJ Comput Mater* 2019, **5**:1-6.

88. Fischer CC, Tibbetts KJ, Morgan D, Ceder G: **Predicting crystal structure by merging data mining with quantum mechanics**. *Nat Mater* 2006, **5**:641-646.

89. Hautier G, Fischer CC, Jain A, Mueller T, Ceder G: **Finding nature's missing ternary oxide compounds using machine learning and density functional theory**. *Chem Mater.* 2010, **22**:3762-3767.

90. Balachandran PV, Theiler J, Rondinelli JM, Lookman T: **Materials prediction via classification learning**. *Sci Rep* 2015, **5**:1-16.

91. Pilania G, Balachandran PV, Kim C, Lookman T: **Finding new perovskite halides via machine learning**. *Front Mater* 2016, **3**:19.

92. Oliynyk AO, Adutwum LA, Harynuk JJ, Mar A: **Classifying crystal structures of binary compounds ab through cluster resolution feature selection and support vector machine analysis**. *Chem Mater* 2016, **28**:6672-6681.

93. Oliynyk AO, Adutwum LA, Rudyk BW, Pisavadia H, Lotfi S, Hlukhyy V, Harynuk JJ, Mar A, Brgoch J: **Disentangling structural confusion through machine learning: structure prediction and polymorphism of equiatomic ternary phases abc**. *J Am Chem Soc* 2017, **139**:17870-17881.

94. Yamashita T, Sato N, Kino H, Miyake T, Tsuda K, Oguchi T: **Crystal structure prediction accelerated by Bayesian optimization**. *Phys Rev Mater* 2018, **2**:013803.

95. Takahashi K, Takahashi L: **Creating machine learning-driven material recipes based on crystal structure**. *J Phys Chem Lett* 2019, **10**:283-288.

96. Liang H, Stanev V, Gilad Kusne A, Takeuchi I: **Cryspnet: crystal**
•• **structure predictions via neural networks**. *Phys Rev Mater* 2020, **4**:123802
The authors develop the *Crystal Structure Prediction Network* (CRYSPNet) model that predicts an inorganic material's Bravais lattice, space group, and lattice parameters based only on its chemical composition.

97. Nouira A, Sokolovska N, Crivello J-C: *CrystalGAN: Learning to*
• *Discover Crystallographic Structures With Generative Adversarial Networks*. 2018arXiv:1810.11203
This paper develops CrystalGAN, a generative adversarial network (GAN) architecture that generates new stable crystal structures with increased domain complexity from known stable structures.

98. Noh J, Kim J, Stein HS, Sanchez-Lengeling B, Gregoire JM,
• Aspuru-Guzik A, Jung Y: **Inverse design of solid-state materials via a continuous representation**. *Matter* 2019, **1**:1370-1384
This paper introduces the inverse materials design framework iMatGen that utilizes an efficient image-based representation in a variational autoencoder model.

99. Hoffmann J, Maestrati L, Sawada Y, Tang J, Sellier JM, Bengio Y: *Data-Driven Approach to Encoding and Decoding 3-D Crystal Structures*. 2019arXiv:1909.00949.

100. Kim S, Noh J, Gu GH, Aspuru-Guzik A, Jung Y: **Generative adversarial networks for crystal structure prediction**. *ACS Central Sci* 2020, **6**:1412-1420.

101. Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel TD, Adams RP, Aspuru-Guzik A: **Automatic chemical design using a data-driven continuous representation of molecules**. *ACS Central Sci* 2018, **4**:268-276.

102. Kim B, Lee S, Kim J: **Inverse design of porous materials using artificial neural networks**. *Sci Adv* 2020, **6**:eaax9324.

103. Ren Z, Noh J, Tian S, Oviedo F, Xing G, Liang Q, Aberle A, Liu Y, Li Q, Jayavelu S *et al.*: *Inverse Design of Crystals Using Generalized Invertible Crystallographic Representation*. 2020arXiv:2005.07609.

104. Maranas CD, Floudas CA: **A global optimization approach for Lennard-Jones microclusters**. *J Chem Phys* 1992, **97**:7667-7678.

105. Karamertzanis PG, Pantelides CC: **Ab initio crystal structure prediction. I. Rigid molecules**. *J Comput Chem* 2005, **26**:304-324.

106. Karamertzanis PG, Price SL: **Energy minimization of crystal structures containing flexible molecules**. *J Chem Theory Comput* 2006, **2**:1184-1199.

107. Karamertzanis PG, Pantelides CC: **Ab initio crystal structure prediction. II. Flexible molecules**. *Mol Phys* 2007, **105**:273-291.

108. Issa N, Karamertzanis PG, Welch GWA, Price SL: **Can the formation of pharmaceutical cocrystals be computationally predicted? I. comparison of lattice energies**. *Cryst Growth Des* 2009, **9**:442-453.

109. Karamertzanis PG, Kazantsev AV, Issa N, Welch GWA, Adjiman CS, Pantelides CC, Price SL: **Can the formation of pharmaceutical cocrystals be computationally predicted? 2. Crystal structure prediction**. *J Chem Theory Comput* 2009, **5**:1432-1448.

110. Kazantsev AV, Karamertzanis PG, Adjiman CS, Pantelides CC, Price SL, Galek PTA, Day GM, Cruz-Cabeza AJ: **Successful prediction of a model pharmaceutical in the fifth blind test of crystal structure prediction**. *Int J Pharmaceut* 2011, **418**:168-178.

111. Kazantsev AV, Karamertzanis PG, Adjiman CS, Pantelides CC: **Efficient handling of molecular flexibility in lattice energy minimization of organic crystals**. *J Chem Theory Comput* 2011, **7**:1998-2016.

112. Pantelides CC, Adjiman CS, Kazantsev AV: **General**
• **computational algorithms for ab initio crystal structure prediction for organic molecules**. *Prediction and Calculation of Crystal Structures* 2014:25-58
This book chapter reviews current status, recent progress, limitations and opportunities for the CSP of organic molecules.

113. Zhang L, Cignitti S, Gani R: **Generic mathematical programming**
•• **formulation and solution for computer-aided molecular design**. *Comput Chem Eng* 2015, **78**:79-84
This paper proposes a generic mathematical programming formulation for computer-aided molecular design (CAMD) problems.

114. Samudra AP, Sahinidis NV: **Optimization-based framework for computer-aided molecular design**. *AIChE J* 2013, **59**:3686-3701.

115. Liu Q, Zhang L, Liu L, Du J, Tula AK, Eden M, Gani R: **OptCAMD:**
• **an optimization-based framework and tool for molecular and mixture product design**. *Comput Chem Eng* 2019, **124**:285-301
This paper introduces the molecular-mixture design toolbox OptCAMD that integrates and streamlines the preliminary design, mathematical optimization-based CAMD, and product evaluation/verification steps of mixture design problem.

116. Hanselman CL, Gounaris CE: **A mathematical optimization**
•• **framework for the design of nanopatterned surfaces**. *AIChE J* 2016, **62**:3250-3263
This article is the first to introduce a mathematical optimization-based nanomaterials design paradigm, demonstrating it with a catalytic surface design problem.

117. Hanselman CL, Zhong W, Tran K, Ulissi ZW, Gounaris CE: **Optimization-based design of active and stable nanostructured surfaces**. *J Phys Chem C* 2019, **123**:29209-29218.

118. Hanselman CL, Alfonso DR, Lekse JW, Matranga C, Miller DC, Gounaris CE *et al.*: **Design of doped perovskite oxygen carriers using mathematical optimization**. *Computer Aided Chemical Engineering, vol 44*. Elsevier; 2018:2461-2466.

119. • Hanselman CL, Alfonso DR, Lekse JW, Matranga C, Miller DC, Gounaris CE *et al.*: **A framework for optimizing oxygen vacancy formation in doped perovskites**. *Comput Chem Eng* 2019, **126**:168-177

This article demonstrated the use of mathematical optimization for the design of doped perovskites that can be highly potent oxygen carriers for chemical looping combustion.

120. Isenberg NM, Taylor MG, Yan Z, Hanselman CL, Mpourmpakis G, Gounaris CE: **Identification of optimally stable nanocluster geometries via mathematical optimization and density-functional theory**. *Mol Syst Des Eng* 2020, **5**:232-244.

121. Yin X, Isenberg NM, Hanselman CL, Dean JR, Mpourmpakis G, Gounaris CE: **Designing stable bimetallic nanoclusters via an iterative two-step optimization approach**. *Mol Syst Des Eng* 2021, **6**:545-557.

122. Sutton AP: *Electronic Structure of Materials*. Clarendon Press; 1993.

123. Khor KE, Das Sarma S: **Proposed universal interatomic potential for elemental tetrahedrally bonded semiconductors**. *Phys Rev B* 1988, **38**:3318.

124. Ito T, Khor KE, Sarma DS: **Empirical potential-based Si-Ge interatomic potential and its application to superlattice stability**. *Phys Rev B* 1989, **40**:9715.

125. Kodiyalam S, Khor KE, Bartelt NC, Williams ED, Das Sarma S: **Energetics of vicinal Si(111) steps using empirical potentials**. *Phys Rev B* 1995, **51**:5200.

126. Joe H, Akiyama T, Nakamura K, Kanisawa K, Ito T: **An empirical potential approach to the structural stability of inas stacking-fault tetrahedron in InAs/GaAs(111)**. *J Cryst Growth* 2007, **301**:837-840.

127. Hasegawa Y, Akiyama T, Pradipto AM, Nakamura K, Ito T: **Empirical interatomic potential approach to the stability of graphitic structure in BAlN and BGaN alloys**. *J Cryst Growth* 2018, **504**:13-16.

128. IBM Corporation: *IBM ILOG CPLEX Optimization Studio V12.9.0*. 2019.

129. National Energy Technology Laboratory, Institute for the design of advanced energy systems (IDAES). www.idaes.org/download/.

130. Zheng H, Wang J, Huang JY, Wang J, Zhang Z, Mao SX: **Dynamic process of phase transition from wurtzite to zinc blende structure in inas nanowires**. *Nano Lett* 2013, **13**:6023-6027.

131. Lehmann S, Wallentin J, Mårtensson EK, Ek M, Deppert K, Dick KA, Borgstrom MT: **Simultaneous growth of pure wurtzite and zinc blende nanowires**. *Nano Lett* 2019, **19**:2723-2730.

132. Ito T: **Recent progress in computer-aided materials design for compound semiconductors**. *J Appl Phys* 1995, **77**:4845-4886.

133. Hanselman CL, Yin X, Miller DC, Gounaris CE: *Matopt: A Python Package for Nanomaterials Design Using Discrete Optimization*. 2021. (under review).

134. Hart WE, Watson J-P, Woodruff DL: **Pyomo: modeling and solving mathematical programs in python**. *Math Program Comput* 2011, **3**:219-260.

135. Bynum ML, Hackebeil GA, Hart WE, Laird CD, Nicholson BL, Siirola JD, Watson J-P, Woodruff DL: *Pyomo-Optimization Modeling in Python, vol 67*. 3rd edition. Springer Science & Business Media; 2021.

136. Galicka M, Bukała M, Buczko R, Kacman P: **Modelling the structure of gaas and inas nanowires**. *J Phys: Condens Matter* 2008, **20**:454226.

137. Bukała M, Galicka M, Buczko R, Kacman P, Shtrikman H, Popovitz-Biro R, Kretinin A, Heiblum M: **What determines the crystal structure of nanowires?** *AIP Conference Proceedings, vol 1199*. American Institute of Physics; 2010:349-350.