Expert Systems With Applications 63 (2016) 222–230

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Collaborative crystal structure prediction

Sangheum Hwang[a,1], Jiho Yoo[a], Chanhee Lee[a], Sang Hyun Lee[b,*]

[a] Samsung Advanced Institute of Technology (SAIT), Suwon, 16678, Korea
[b] Department of Electrical Engineering, Sejong University, Seoul, 05006, Korea

ARTICLE INFO

Article history:
Received 26 April 2016
Revised 27 June 2016
Accepted 1 July 2016
Available online 5 July 2016

Keywords:
Crystal structure prediction
High-throughput computational screening
Collaborative filtering
Side information

ABSTRACT

The prediction of crystal structures is one of the most essential challenges in designing novel functional materials. A data-driven prediction technique that uses the database of known crystal structures and substitutes ions among materials of known crystal structures to concoct new crystal structures has been proposed. This technique has been applied to generate crystal-structure candidates for the purpose of first-principles-calculation-based high-throughput computational screening. However, this technique has a functional limitation that the ion substitution tendencies are available only for typical ions such that their associated crystal structures appear in well-known materials. To overcome such a limitation, this work introduces an idea of collaborative filtering to the calculation of the ionic substitution tendencies. Based on this idea, we develop symmetric matrix factorization (SMF) method to model underlying substitution conditions. In addition, we present a symmetric matrix co-factorization (SMCF) method to incorporate additional knowledge pertaining to chemical properties in estimating the substitution tendencies among ions with extremely small amount of previous knowledge in the database. The performance of the prediction is investigated along with existing techniques through *in silico* experiments using real crystal-structure database. The numerical results show that the proposed SMF- and SMCF-based prediction outperform existing techniques in terms of the prediction accuracy.

1. Introduction

The prediction of the atomic-level crystal structure is a challenging fundamental task in the discovery of new functional materials. To assess its difficulty and importance, Maddox (1988) has published an article to question the predictability of crystal structures for first-principles calculation using the knowledge of chemical compositions only. However, recent advances in computational resource and quantum mechanics theory, such as density functional theory (Hohenberg & Kohn, 1964), have made a remarkable progress in modeling of the ground-state and finite temperature behavior (e.g., 0˜K energy) for a given crystal structure. The prediction of the crystal structure can be cast as an optimization problem: The objective function is the ground-state energy obtained by first-principles calculations, and unknown variables correspond to three-dimensional coordinates of atoms in a unit cell. From this mathematical framework, the prediction of the crystal structure is carried out by comparing a set of candidate structures in terms of estimated energy and choosing the lowest-energy structure as the ground-state configuration.

However, it still remains very difficult to predict the crystal structure and its associated three-dimensional coordinates of all atoms in a computational way for several reasons: First, the energy landscape associated with structure configurations has a great number of local minima, which prevent finding the global minimum, or the ground-state structure configuration, using a simple gradient search. Second, there are infinitely many configurations of three-dimensional coordinates for atoms. Thus, the solution space is so huge that the entire solution space cannot be explored. Third, first-principles calculations for evaluating the objective function, i.e., the ground-state energy, still require considerable amount of time and computational efforts.

To tackle such issues, several meta-heuristic approaches, such as simulated annealing, genetic algorithm variants, random sampling, have been developed (Woodley & Catlow, 2008). Most of them deal with the solution space exploration using a simple objective function which avoids computationally demanding first-principles calculations. Although those approaches have lead a noticeable progress in the crystal structure prediction, they fail to provide the prediction structures quickly because they rely basically on solving an extremely complicated optimization problem. In particular, most of them are not suitable for high-throughput

* Corresponding author.
E-mail addresses: shwang@lunit.io (S. Hwang), jiho.mw.yoo@samsung.com (J. Yoo), xfingers@samsung.com (C. Lee), sanghyunlee@sejong.ac.kr (S.H. Lee).
[1] He is currently working for Lunit Inc., Seoul, Korea.

http://dx.doi.org/10.1016/j.eswa.2016.07.001
0957-4174/

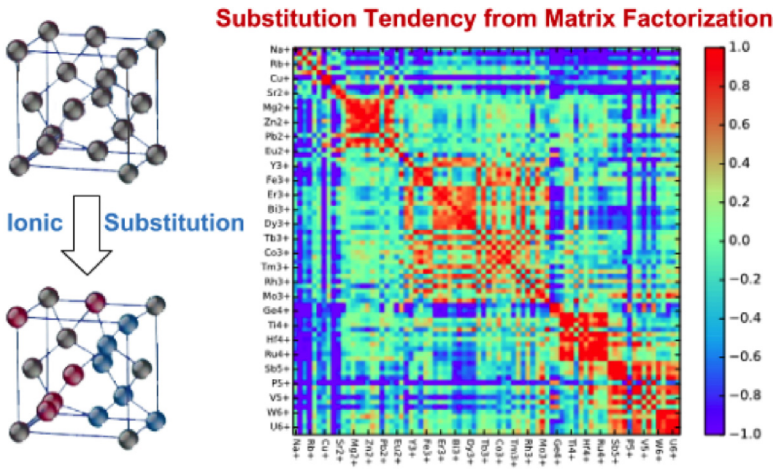

Fig. 1. Schematic diagram of the ionic substitution algorithm. Based on an origin structure, the most probable structure which maximizes the substitution probabilities is generated.

computational screening (HTCS) which chooses promising candidates for functional materials among a huge number of candidates (Jain et al., 2011). The computational infrastructure for HTCS includes the structure database, the computational resource, and the job management. The automation of those tasks in the data processing allows the computational screening of a massive amount of candidate materials. The benefits of this framework have been demonstrated for last few decades, in various studies, e.g., finding Li-ion battery anode materials (Kirklin, Meredig, & Wolverton, 2013) and radiation detector materials (Ortiz, Eriksson, & Klintenberg, 2009).

To find crystal structures of high stability without quantum-mechanics calculations, novel structure prediction techniques have been introduced using data mining techniques (Fischer, Tibbetts, Morgan, & Ceder, 2006; Hautier, Fischer, Ehrlacher, Jain, & Ceder, 2006). Those techniques share a common underlying idea of constructing a probability model using the structural information from known crystal structures in order to quantify the stability of particular structural configurations (see Fig. 1). To be more specific, the probability model of a given chemical composition enables to seek feasible structural configurations from the set of existing structures and to choose the most probable one. Those techniques have proved useful in the discovery of novel chemical materials. For example, Wu, Lazic, Hautier, Persson, and Ceder (2013) has employed the ionic substitution algorithm (Hautier et al., 2006) to find new functional water-splitting photocatalysts. However, this method cannot estimate the substitution probability among ions where no substitution is found in the data set. Those missing values severely undermine the prediction performance.

To improve the predictive accuracy of the ionic substitution algorithm (Hautier et al., 2006), this paper introduces machine learning techniques used frequently for collaborative prediction (Bobadilla, Ortega, Hernando, & Gutierrez, 2013; Su & Khoshgoftaar, 2009): symmetric matrix factorization (SMF) and symmetric matrix co-factorization (SMCF). Those approaches utilize a low-rank approximation (via matrix factorization) for the data set organized in a matrix form with an individual element corresponding to the substitution probability between two ions. Meanwhile, some missing elements are inevitable outcomes for this data matrix. To handle them, several simple low-rank matrices are found such that all filled elements of the data matrix are equal to the corresponding elements of the product of those simple matrices. Thus, the data matrix is reconstructed from the product of those component matrices. In this step, those elements in the reconstructed matrix at the position of missing elements are filled. Furthermore, the weighted matrix factorization (i.e., SMF) model is extended to the weighted matrix co-factorization (i.e., SMCF) model to incorporate the prior-knowledge matrix along into the reconstructed data matrix.

2. Ionic substitution approach

Chemical experimentalists frequently make attempts to replace chemical elements (i.e., ions) in a known chemical compound with another to develop a new idea. It basically originates from the assumption that, if two elements have similar atomic-level properties, the new structure obtained from the substitution still retains a similar level of stability. The ionic substitution algorithm (Hautier et al., 2006) incorporates this idea by learning how likely the substitution of certain ions in a compound results in another compound with the same crystal structure *in silico* from the knowledge based on the experimental database. To this end, the likelihood of the substitution is characterized by a probabilistic model developed in the way presented in the sequel.

2.1. Substitution probabilistic model

A compound comprised of n different ions is represented in an n-dimensional vector as $\mathbf{x} = \{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$ where $x^{(i)}$ denotes an i-th ionic species. The ionic substitution probability between two compounds $\mathbf{x}$ and $\mathbf{y}$ with the same crystal structure is defined by:

$$p(\mathbf{x}, \mathbf{y}; \boldsymbol{\lambda}) = \frac{1}{Z}\exp\left[\sum_{i=1}^{n}\sum_{(a,b)}\lambda_{ab}^{(i)} f_{ab}^{(i)}\left(x^{(i)}, y^{(i)}\right)\right] = \frac{1}{Z}\prod_{i=1}^{n}\exp\left[\sum_{(a,b)}\lambda_{ab}^{(i)} f_{ab}^{(i)}\left(x^{(i)}, y^{(i)}\right)\right], \quad (1)$$

where $f_{ab}^{(i)}\left(x^{(i)}, y^{(i)}\right)$ is a binary feature function that yields one, if two ions $x^{(i)} = a$ and $y^{(i)} = b$ in position i are switched, and zero otherwise. A positive parameter $\lambda_{ab}^{(i)}$ denotes the relative frequency of the corresponding substitution and $\boldsymbol{\lambda}$ denotes the vector of model parameters. Also, Z is a normalization constant (or a partition function from science's perspectives) given by: