Proceedings of the 4th International Conference on Materials Chemistry and Environmental Engineering
DOI: 10.54254/2755-2721/58/20240723

The significance of materials informatics on material science

Siyu Gong

University of Waterloo 200 University Avenue West, Waterloo, ON, Canada

s2gong@uwaterloo.ca

Abstract. In 2011, U.S. President Barack Obama proposed the Material Genome Project. Want to high high-speed and low-cost methods to develop material science, which promoted the rapid development of materials informatics. Materials informatics is an interdisciplinary field that employs the principles of informatics to enhance the comprehension, utilization, and advancement of materials within the realm of materials science. Materials informatics can take multiple approaches and influence many aspects of new material development. Material informatics will cause great changes in the material industry and promote the rapid development of materials. A variety of materials informatics schemes have been developed to analyze materials, and there have been many successful cases worth applying. This article introduces the basic concepts and main research fields of materials informatics and describes four main steps in materials informatics: data collection, data processing, the establishment of a database, construction of a model. The main applications of materials informatics include simulation and prediction of material properties, development of new materials, material optimization and performance improvement. Thinking about the difficult problems in the field of materials informatics for application.

Keywords: Materials Informatics, Machine Learning, Database.

1. Introduction

At present, the traditional way to develop new materials is to accumulate experience through experimental design and simulation. Scientists develop new materials through experimentation and laboratory manipulation. Improve the properties of materials by accumulating experience.

However, the traditional research and development mode is time-consuming and costly. Lab workers need to conduct a large number of experiments to obtain data about the properties of materials which makes the development way of materials at a slow pace. This also needs a large number of experimental funds to support their experiment. At the same time, the subjectivity and uncertainty of traditional experiments will affect the results. Error in the experimental operation steps will lead to error in the results, and these factors may lead to the unreliability of the research and obtain wrong results of the experiment. The traditional research and development mode of materials has lots of disadvantages such as long time, high cost and uncertain results. Therefore, new research and development models are needed to improve the efficiency of material research and development. The concept of material informatics was thus proposed. In the early 1980s, the field of materials science began to try to apply computer technology to process large amounts of data about materials. In 1999, at the conference named "Materials informatics- Effective Data Management New Materials Discovery" held in the United States,

DOI: 10.54254/2755-2721/58/20240723

Professor John first proposed the concept of materials informatics which opened the curtain of materials informatics [1]. Since then, there have been more and more studies on materials informatics. In 2007, Krishna Rajan authored an article for the journal Materials Today, wherein he proposed the notion of regarding Materials Informatics as a distinct subfield within the domain of materials science and engineering [2]. With the rapid development of computer technology, computer simulation technology has been widely used. Materials informatics has therefore made great progress. In 2014 Dey et al. use material informatics to predict the bandgaps of over 200 new chalcopyrite compounds for previously untested chemistries [3]. In 2015 Sarkisov et al. came up with computational structure characterization tools for the era of material informatics [4]. In 2018 Jiang et al. Found a materials informatics approach to Ni-based single crystal superalloys lattice misfit prediction [5]. In recent years, the development of materials informatics is getting faster. More and more scholars have begun to pay attention to the role of material informatics. Materials informatics plays a vital role in modern materials science and engineering. It can help lab workers better understand the properties and behavior of materials. Materials informatics makes it more efficient to predict the property structure and function of materials. It greatly reduces the time and cost of material development. The increasing computing power and storage capacity of computers has laid a more solid foundation for the development of materials informatics. Computers provide a wealth of data processing and analysis tools so that a large number of data can be more easily managed and analyzed. At the same time, the gradual maturity of machine learning has also promoted the emergence and development of material informatics. Machine learning can help build fast and accurate material property prediction models, which can provide important support and guidance for material science research and material design. Material informatics brings new research methods and tools to the field of materials science. However, as a new subject, material informatics still needs further research and development in the application of related technologies and methods.

2. Procedures of material informatics

The essence of material informatics is the integrated design of materials and the construction of a material database platform, as well as big data analysis in the field of materials. Combined with the experimental data and process data of materials to form a large data set of materials. Through the calculation of high flux materials to get a lot of material theory data. Using database technology for management, and data mining methods for analysis and prediction to summarize and form new knowledge. Material informatics is to explore the "genes" that determine the structure-property relationship of materials and promote the rapid development of new materials. Materials informatics is mainly divided into the following steps.

2.1. Data collection

Material informatics is the effective management of material data. In the process of material development, the available data mainly includes the data in academic papers, the data in published patents, and the evaluation results of physical properties of materials collected during research and development. Material informatics is a new data-driven discipline, which requires a large amount of raw data to support the modeling and prediction of materials. The accuracy and reliability of data should be ensured during data collection, so as to provide better data support for the following prediction and database construction.

2.2. Data processing

Because the experimenters and methods of recording various data are different. A large number of raw data obtained through experiments and calculations are stored in different data formats. Material staff need to carry on the preliminary statistical analysis of large amounts of data. To develop a more robust performance prediction model, it is imperative to comprehend the structure and importance of the data, and undertake data preprocessing prior to modeling to guarantee the accuracy and coherence of the data. For example, delete data noise, outliers, missing values, duplicate data, etc. The data can be used for the next step after the overall decluttering and Statisticaling.

DOI: 10.54254/2755-2721/58/20240723

2.3. Establishment of database

The core issue of materials informatics is the establishment of the database. Material data is the basis of material information technology and material genetic engineering. To effectively manage and analyze the large amount of data generated by the combined experiment, it is necessary to establish the corresponding material database.

In 2011, the United States proposed the development of the "Materials Genome Initiative" (MGI) project, which has three major elements: database, high-throughput calculation method and high-throughput experimental method, to provide technical support for accelerating the intelligent design of materials. The role and status of the material database become more prominent: on the one hand, a material database can provide huge amounts of data storage space for high-throughput experiments and high flux calculation results; On the other hand, material databases provide parameters for high-throughput calculations, or guide material design by mining knowledge models in databases [6].

With the implementation of the MGI program, databases with the concept of promoting the rapid development of materials science and materials data sharing have been emerging. The continuous attention, integration and improvement of materials science databases in the world has greatly reduced the difficulty of obtaining material data in the technology of material big data.

More and more data resources are available, and these databases include Springer Materials, a comprehensive database for finding and characterizing materials, Total Material, which is a database of figures and tools; ICSD, the world's largest database of inorganic crystal structures, CSD, the Cambridge Crystal Structure Database, Calculating Materials Data from First Principles of the Materials Project, and Aflow, which has data on more than 100 million material properties. In recent years, great breakthroughs have been made in material databases, which can better simulate the characteristics of materials [7].

2.4. Construction of model

The current algorithms and models in materials informatics are mainly machine learning. Deep learning is a more general term for machine learning. It is a data analytics method that uses artificial intelligence (AI) to define distinct associations between input and output datasets from in vitro studies in order to investigate the rules underlying datasets. [8]. Common machine learning-related problems in Materials informatics can be roughly divided into supervised learning and unsupervised learning. Depending on whether the predicted values are continuous or discrete, they can be divided into regression and classification tasks. In addition, machine learning methods commonly used also include feature selection and dimensionality reduction [7].

Commonly used machine learning models mainly include Linear Regression which can process linear features, Support Vector Machines which can be processing nonlinear features, and Deep Neural Networks which can perform nonlinear function fitting feature extraction. Recursive or Recurrent Neural Networks can be used to establish functional relationships between structure and material properties, etc.

3. Applications

3.1. Simulation and prediction of material properties

The demand for the design and development of new materials has surpassed the capabilities of traditional analysis methods relying on theory and experimentation. Consequently, the integration of machine learning and materials genomics has emerged as a promising approach in the pursuit of high-performance materials. This interdisciplinary approach has not only facilitated the research and development of advanced materials but has also propelled the progress of materials science as a whole. Machine learning using artificial intelligence (AI) by defining the experiment of the relationship between input and output data sets in exploring the law behind the data set. It has lately emerged as a key tool for artificial intelligence and has been applied to solve direct problems with unknown features (composition, experimental circumstances, etc.) by predicting material attributes. Using this method to

DOI: 10.54254/2755-2721/58/20240723

predict the properties of materials is particularly efficient for the development of new materials. As shown in Figure 1, the material can be predicted by machine learning based on the composition of the material.

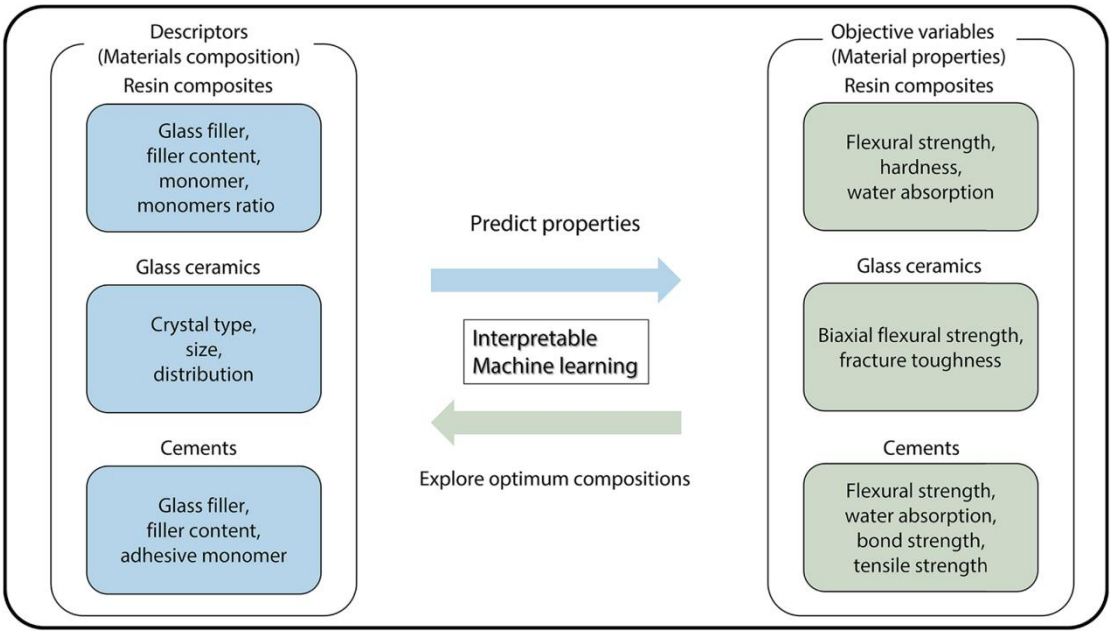

Figure 1. Schematic illustration of the materials informatics approach.

The traditional ways to measure mechanical properties of the composite are use the numerical and experimental methods to measure, but because of cost, time and effort, it hindered the innovation of the material. Predicting various properties of materials can be done by developing machine learning-aided models. The internal composition of the material will be different with the different manufacturing processes, and the internal composition of the material also determines the performance of the material. Li et al. successfully developed machine learning and material information methods for predicting the lateral mechanical properties of microporous unidirectional CFRP composites through the steps of random microstructure generation, evaluation of lateral mechanical properties, statistical representation of microstructure, dimension reduction and microstructure extraction-property connection [9].

3.2. Development of new materials

In addition to predicting material properties, obtaining new material structures is also an important work in materials research. Structural design is divided into forward and reverse two kinds. Forward design is to select a reasonable structure directly from the batch structure. Reverse material design refers to the determination of material structure in reverse from the target properties [10]. Yamaguchi et al. use machine learning methods to explore the best ingredients to achieve the desired properties of dental materials. The invention of new dental materials can be explored through four steps: data preparation, regression model development, model evaluation, and best descriptor search [8]. Jiang et al designed the transmitted and reflected color radiation coolers through material informatics methods to achieve radiation cooling. It can be well applied to energy-saving buildings, electrical equipment, etc. [11].

3.3. Material optimization and performance improvement

Material informatics can provide a large amount of information about material properties, parameters and past academic research results through the search and calculation of databases and models. This information can be used for material selection, design and improvement to help researchers better understand and optimize material properties. The simulation and optimization of materials can be realized by theoretical calculation of their structure and properties through computer simulation

DOI: 10.54254/2755-2721/58/20240723

technology. The thermodynamic properties of materials can be obtained by molecular dynamics simulation. DFT can calculate the electron band structure, atomic position and other information. After simulating the materials, the materials with excellent properties are selected. The properties and strength of the material are improved by changing the composition structure and other parameters of the material. The use of machine learning to optimize the properties of existing materials will greatly reduce the time, and clear the direction of optimization, reducing scientific research costs. Zhao illustrates that the development of pesticides is not an easy process and requires a lot of money. A new pesticide needs to be tested on many different compounds and takes a lot of time. Traditional testing methods are no longer suitable for the development of pesticides. With the development of the computer, the informatics-related solution for pesticide molecular design and optimization of progress greatly reduces the time cost [12].

4. Impacts and challenges of materials informatics

4.1. Impacts

Materials informatics leverages big data analytics and machine learning algorithms. Data mining and modeling are used to predict the properties and properties of materials. To better guide material synthesis and design. The use of materials informatics in the development of materials can reduce the research cost and time, and improve the research efficiency. More use of materials informatics computational simulation and other technologies can optimize the preparation and processing of materials, improve the performance and quality of materials, and reduce the waste of resources and environmental pollution.

The requirements of materials informatics accelerate the construction and improvement of materials databases. The material database collects and organizes a large number of material-related data, including the composition, structure and performance of materials. This provides a platform for researchers to share and access materials and data and reduces the workload of repeated experiments and data collection, which greatly improves the efficiency of research. At the same time, these data can be used not only to verify the accuracy of the experimental results but also to decrease the repeatability and improve the reliability of the experimental results.

4.2. Challenges

Materials informatics has developed at the speed of nearly light in recent years. However, as a kind of new science, there are still many challenges and problems that need to be solved. The lack of appropriate theoretical models to describe and explain the behavior of some materials is a huge challenge to the modeling ability of materials informatics. Most of the properties of materials are nonlinear, and the laws are always difficult to find. Since 2006, machine learning has made breakthrough progress in the field of new material discovery and creation, which is a new reform of the traditional new material research and development model. Machine learning is of great innovative significance for material science research and the creation of new materials. But when results conflict with knowledge extracted from traditional, interpretable models and methods, should materials scientists trust predictions made by machine learning models with new knowledge?

Due to the particularity of different materials, the applicability and accuracy of the model for each material are different, which will lead to certain result errors. The collection and management of the large amount of data required for databases remains difficult. The quality of data is difficult to ensure, and there is no standardization requirement for data. The database needs to be updated and maintained regularly, and an effective data management mechanism needs to be established.

At the same time, material informatics is an interdisciplinary subject, which not only requires both statistical knowledge and computer knowledge but also professional knowledge of materials. This is a great interdisciplinary integration, which requires different disciplines to strengthen cooperation and communication.

DOI: 10.54254/2755-2721/58/20240723

5. Conclusion

The big data analysis of materials is the main component of material informatics. It is an advanced method from multiple material information databases, knowledge and forecasting rules are extracted using a sophisticated data mining technique. It leverages techniques such as big data, machine learning, data mining, and model prediction. Can process and analyze a large amount of material data. The correlation between material properties and behavior is extracted and simulated by a machine to form a complete analysis of a material. As a new emerging discipline, materials informatics has brought great development and challenges to the materials industry. Material informatics establishes a database by systematically collecting and summarizing material data information. And build machine learning models to predict different properties of materials. It will greatly speed up the development of new materials 4So as to achieve the ultimate goal of shortening the material development cycle and reducing the material research and development cost. Material informatics will become an extremely important method in future material research and development. However, due to the short history of materials informatics, there are still many imperfections. For example, the accuracy of data obtained from experiments or literature cannot be verified. The source of the data is highly uncertain. At the same time, how to achieve high-quality integration of data will become a huge challenge. The relationship between the properties and composition of most materials is complex. Machine learning models can make mistakes in handling complex tasks. For example, incorrect algorithm selection, improper feature selection, or too high model complexity can lead to inaccurate results. At present, there is a lack of checking the correctness of machine model results. How to improve the integrity of the database and the accuracy of material informatics prediction will be the main problem still waiting to be solved in the next few years. It is necessary for researchers to deeply study and develop material models and algorithms based on multi-scale to achieve accurate prediction and optimization of material properties. Improving the requirements for raw data to ensure the accuracy of data in the database. There is a need to select the appropriate machine model algorithm and optimize the relevant parameters to improve the accuracy of the algorithm. Rigorous validation and evaluation are required after optimizing the model to ensure its accuracy. And based on the existing algorithm to continue to learn and improve. Materials informatics will have broad application prospects and development space in the future, and create a huge progress in materials science.

References

[1] John R Rodgers. Materials Informatics- Effective Data Management for New Materials Discovery [M]: Boston :knowledge Press, 1999 .

[2] Krishna Rajan, Materials informatics, Materials Today,Volume 15, Issue 11,2012,Page 470, ISSN 1369-702.

[3] Dey, P. P., Bible, J., Datta, S., Broderick, S., Jasinski, J. B., Sunkara, M. K., Menon, M., & Rajan, K. (2014). Informatics-aided bandgap engineering for solar materials. Computational Materials Science, 83, 185–195.

[4] Sarkisov, L., & Kim, J. (2015). Computational structure characterization tools for the era of material informatics. Chemical Engineering Science, 121, 322–330.

[5] Jiang, X., Yin, H., Zhang, C., Zhang, R., Zhang, K., Deng, Z., Liu, G., & Qu, X. (2018). An materials informatics approach to Ni-based single crystal superalloys lattice misfit prediction. Computational Materials Science, 143, 295–300.

[6] Yang Li, Su Hang, Chai Feng, Luo Xiaobing, & Duan Linna. (2019). Application status of material database and data mining technology. Progress of materials in China, 38(7), 11.

[7] Niu Chengcheng, Li Shaobo, Hu Jianjun ,Dan Yabo, Cao Zhuo & Li Xaing. (2020).Overview of machine learning applications in materials informatics. Material guide(23),23100-23108.

[8] Yamaguchi, S., Li, H., & Imazato, S. (2023). Materials informatics for developing new restorative dental materials: A narrative review. Frontiers in Dental Medicine, 4.

DOI: 10.54254/2755-2721/58/20240723

[9] Li, M., Zhang, H., Li, S., Zhu, W., & Ke, Y. (2022). Machine learning and materials informatics approaches for predicting transverse mechanical properties of unidirectional CFRP composites with microvoids. Materials & Design, 224, 111340.

[10] Guo Jialong, Wang Zongguo, Wang Yangang, Zhao Xunshan, Su Yanjing & Liu Zhiwei.(2021). Overview of material research and development methods based on computer technology. Frontiers of data and computing (02),120-132.

[11] Guo, J., Ju, S., Lee, Y., Gunay, A. A., & Shiomi, J. (2022). Photonic design for color compatible radiative cooling accelerated by materials informatics. International Journal of Heat and Mass Transfer, 195, 123193.

[12] Zhao, W., Huang, Y., & Hao, G. (2022). Pesticide informatics expands the opportunity for structure-based molecular design and optimization. Advanced Agrochem, 1(2), 139–147.