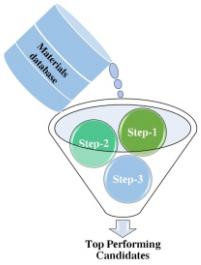
Machine Learning

13

Machine Learning and Data Science for Accelerated Materials Discovery

Srinivasu Kancharlapalli*
Chemistry Division, Bhabha Atomic Research Centre, Mumbai 400085, INDIA
Homi Bhabha National Institute, Mumbai 400094, INDIA



High-throughput screening of materials database for identifying top performing materials

ABSTRACT

Invention of advanced functional materials plays a great role in technological advancements and industrial revolution which ultimately improves living standards of mankind. Traditional materials discovery through experimental, theoretical and computational studies makes the process of materials discovery very expensive and time consuming. With the tremendous increase in available open materials database and advanced algorithms along the exponential growth in computational infrastructure, data science and machine learning became the fourth pillar of the materials discovery. In this article, current state of the materials informatics and challenges are discussed along with few important studies in designing advanced materials using machine learning and high-throughput screening techniques.

KEYWORDS: *Data science, Accelerated materials discovery, Machine learning, Metal organic frameworks*

Introduction

In recent years, revolution in artificial intelligence (AI) and big data have shown potential applications in accelerating the discovery of new molecules and materials[1]. Beyond the traditional methods of materials exploration using the trial and error experiments, theoretical and computational studies, data driven materials discovery is emerging as the fourth paradigm of material science which can improve the pace of materials innovation[2,3]. With the intention of accelerating the discovery of new materials, "Material Genome Initiative" (MGI) project was launched by the USA[4]. In machine learning (ML), a subclass of AI, machine extracts the knowledge from data (of materials) through mapping the structure-property relations which can be quite complex and beyond human intelligence in most of the cases and the knowledge gained can be applied for future predictions. One of the early attempts to use data for materials informatics was the development of CALculation of PHAse Diagrams (CALPHAD) to calculate the phase diagrams of alloys using the computed data of phase diagrams[5]. Data can be considered as the key component for material informatics and the amount of open-source data of materials has been rapidly increasing over the years. A large number of open materials databases like Inorganic Crystallographic Structural Database (ICSD)[6], Cambridge Crystallographic Data Centre (CCDC)[7], AFLOW[8], NOMAD[9], Materials Project[10], MARVEL NCCR[11], etc. composed of both experimental and computational data are openly available. Apart from the existing database, with different possible chemical compositions and structures, the chemical space of materials is virtually unlimited and many more new materials can be explored. In addition to the easily accessible open data resources, explosive growth in the computational infrastructure along with the development of efficient algorithms like deep learning methods accelerated the field of data driven materials innovation.

*Author for Correspondence: Srinivasu Kancharlapalli
E-mail: ksvasu@barc.gov.in

Challenges in ML for Materials Science

Fig. 1 depicts the typical supervised ML model trained using labelled data to predict the material properties. Major components of such ML model are (a) Defining a problem (b) Data acquisition and selecting appropriate feature space, (c) Data processing or Exploratory Data Analysis (EDA) and (d) Training and validating the model using a suitable algorithm. Though many open-source materials databases are available, data is composed of different categories and data of each category is relatively limited when compared to other fields of data science. In most of the experimental data, studies were conducted at different experimental conditions and hence the data depends on various control parameters like temperature, time, humidity, raw chemicals used, etc. Once the data is selected, next key challenge is to select appropriate set of features (fingerprints) of the materials to map with the target property. Open-source libraries like Pymatgen[12], Matminer[13], Atomic Simulation Environment (ASE)[14], DScribe[15], etc. are highly useful for extracting different site, bond and global (lattice) features of molecule and materials. EDA includes verifying any outliers, imputing the missing data, encoding the object type parameters to numeric type, checking for any duplicate copies in the data, etc. Once the data is ready, selecting a particular algorithm for a given problem is another challenge and it should consider different factors like size of data, feature space, complexity of problem etc. If a too complex (high variance) model like deep learning algorithm is selected with limited data points, it can lead to over fitting. Interpretability of the trained model is another important factor to understand the features that attribute the most to overall prediction[16]. Accuracy of the model can be further tuned using the hyper-parameter tuning methods like random search cross validation and grid search cross validation. Other than

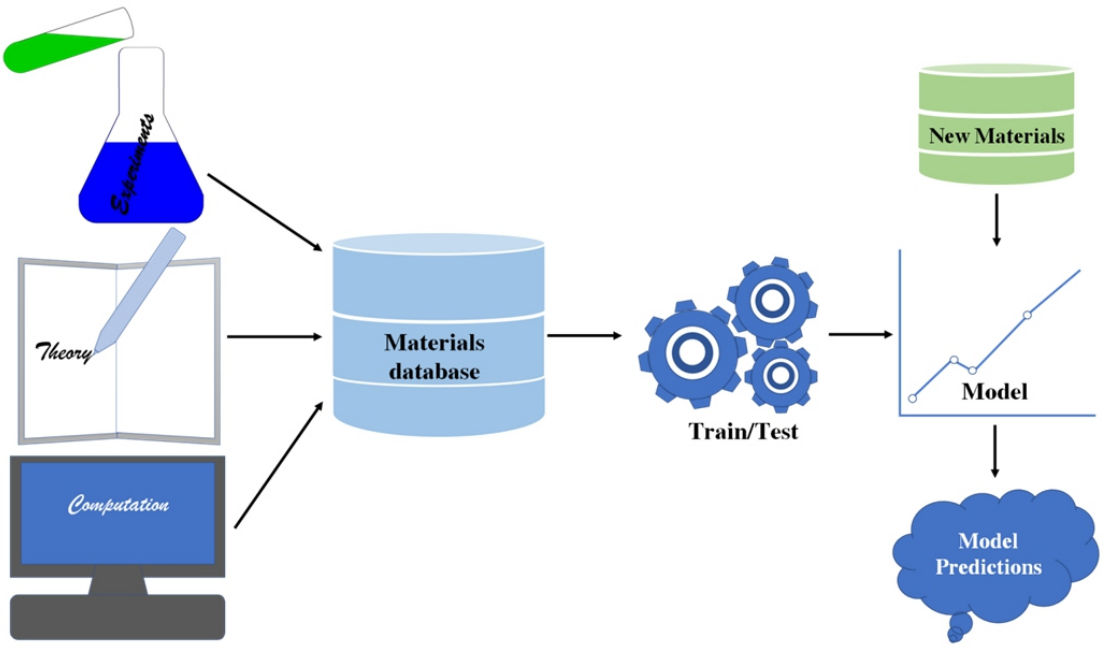


Fig.1: Schematic representation of training a supervised machine learning model for materials property prediction.

training ML models for predicting properties of materials, High-Throughput Screening (HTS) techniques are shown to be potential tool to identify top materials for a particular application from large database of materials which is schematically shown in Fig.2. Recently, self-driving experimental laboratories were developed where robots can perform autonomous experiments very precisely to discover advanced functional materials.

Materials for Energy related Applications

With ever increasing energy demands and adverse environmental effects of burning fossil fuels, great attention has been given to the clean and renewable energy technologies like solar, wind, hydrogen, nuclear energy, etc.[17] Progress in these new and advanced energy sectors is highly dependent on the design and development of advanced functional materials to withstand specific conditions like high temperature, corrosive conditions, high pressure, high energy radiation, etc. where data driven materials discovery has shown potential applications. Band gap of materials is an important property for designing materials for optical and electronic applications and the conventional Density Functional Theory (DFT) methods are inefficient in producing the accurate results whereas the hybrid functional methods which can provide reasonably good results are highly expensive. Zhuo et al.[18] developed a ML model to predict the band gap in inorganic solids where, support vector classification model was first used to classify metal and semiconductors followed by a support vector regressor to predict the band gaps. Kim et al.[19] trained a ML model using the Least Absolute Shrinkage and Selection Operator (LASSO) methods for predicting the dielectric breakdown strength in perovskite materials. Rajan et al.[20] trained a Gaussian process regressor model for predicting the band gaps in two-dimensional (2D) transition metal carbides and nitrides, MXenes. Through the ML guided DFT studies, Sendek et al.[21] discovered many crystalline solid materials with high Li ion conductivity at room temperature which is highly important in designing efficient Li-ion batteries. Using the DFT results on 104 graphene-supported single atoms catalysts, Lin et al.[22] trained random forest ML model to predict the overpotentials associates with the oxygen reduction reaction, oxygen and hydrogen evolution reactions over the selected catalysts. The trained model was used to predict the catalytic activity of other 260 graphene supported single atoms catalysts. Using the first principles based HTS study, Wu et al.[23] screened nitride and oxynitride compounds to identify the novel water splitting photocatalyst. Through the DFT based HTS study, Greeley et al.[24] screened 700 binary alloy surfaces for hydrogen evolution reaction and identified BiPt with better activity as compared to Pt and same was synthesized and the experimental results also shown improved catalytic activity compared to pure Pt surface. Developing advanced and safe fuel materials for nuclear reactors is one of the challenging problems in nuclear energy. Predicting properties of nuclear materials under operational and accidental conditions is a challenging task and difficult to carry out experiments where ML has shown potential applications. Kobayashi et al.[25] constructed a machine learning potential for thorium dioxide fuel materials using the first principles molecular dynamic studies with limited number of atoms and the developed potential was used to simulate the high temperature thermodynamic properties of ThO_2. Using the available experimental data, Jin et al.[26] trained a machine learning model to predict the radiation induced void swelling in different steels.

Porous Materials for Adsorption/Separation of Gases

Designing porous materials for adsorption and separation of gas mixtures is another important area for energy, environment and many other industrial gas separation applications. Metal organic frameworks (MOFs) are reported to have potential applications in separation/storage of various important gases[27]. CO_2 capture is an important technique which can be installed at stationary emission points to restrict the CO_2 levels in the atmosphere as the conventional capture through aqueous amine scrubbing is an energy intensive process. Li et al.[28] screened a database of 5109 MOFs for CO_2 capture from wet flue gas mixture through grand canonical Monte Carlo (GCMC) simulations using framework charges calculated from the extended charge equilibration (EQeq) method. Comparison of the CO_2/H_2O selectivity in the top

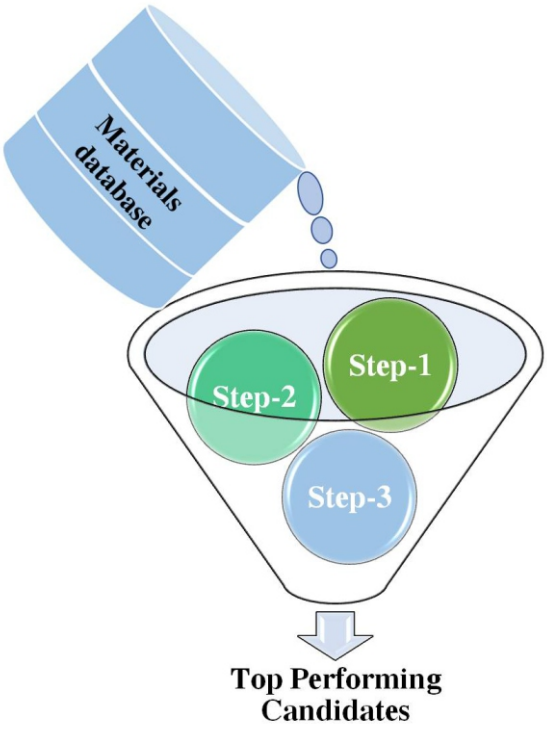


Fig.2: Schematic representation for multi step high-throughput screening of materials database to identify the top performing materials for a particular application .

15 MOFs calculated using the two different charge methods, EQeq and DFT based Repeating Electrostatic Potential Extracted ATomic (REPEAT) revealed that the CO_2/H_2O selectivity values using EQeq charges were overestimated and the K_H (Henry's constant) of H_2O is more sensitive to the charge method than that of CO_2 and N_2. Though the DFT based atomic charges are accurate as compared to empirical methods like EQeq, it is not practical to use for screening large database of MOFs.

To answer this issue, we trained a Random Forest based ML model to predict the atomic charges of MOF atoms using a limited yet meaningful set of features representing both the atom site properties and the local bonding environment and the atomic charges calculated using the Density Derived Electrostatic and Chemical (DDEC6) method[29]. The trained model predicts accurate atomic charges in MOFs with R^2 value of 0.9952 and a mean absolute error of 0.019 at a fraction of the computational cost of DFT. In another interesting study, Boyd et al.[30] screened a library of 325,000 hypothetically generated MOFs and proposed that MOFs with parallel aromatic rings separated by around 7 Å as effective for CO_2 capture in presence of water vapor. It was also experimentally validated by synthesizing such MOFs with optimal CO_2 binding environment which have shown minimal influence of water on the CO_2 capture capacity. Simon et al.[31] screened a database of 670,000 porous materials for Xe/Kr separation and identified aluminophosphate zeolite analogue and a calcium based coordination network as the two most selective materials. HTS of 137,000 hypothetical MOFs for Xe/Kr separation by Sikora et al.[32] concluded that MOFs with pores just enough to fit a xenon and having tubular morphology of uniform width are ideal for Xe/Kr separation.

In summary, this review article elaborates the importance and future scope of data driven materials discovery. Major steps involved in a typical supervised machine learning model for materials property prediction have been discussed with different challenges associated with them. Few data driven materials discovery reports especially for energy related materials have been discussed.

References

[1] Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A., "Machine learning for molecular and materials science." Nature 559 (7715), (2018):547-555.

[2] Hey, A. J.; Tansley, S.; Tolle, K. M., The fourth paradigm: data-intensive scientific discovery. Microsoft research Redmond, WA: 2009; Vol. 1.

[3] Pilania, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R., "Accelerating materials property predictions using machine learning." Sci. Rep.3 (1), (2013):2810.

[4] https://www.mgi.gov. https://www.mgi.gov

[5] Saunders, N.; Miodownik, A. P., CALPHAD (calculation of phase diagrams): a comprehensive guide. Elsevier: 1998.

[6] https://icsd.products.fiz-karlsruhe.de/.

[7] https://www.ccdc.cam.ac.uk/.

[8] Curtarolo, S.; Setyawan, W.; Hart, G. L.; Jahnatek, M.; Chepulskii, R. V.; Taylor, R. H.; Wang, S.; Xue, J.; Yang, K.; Levy, O., "AFLOW: An automatic framework for high-throughput materials discovery." Comput. Mater. Sci.58, (2012):218-226.

[9] Zacharias, N.; Monet, D. G.; Levine, S. E.; Urban, S. E.; Gaume, R.; Wycoff, G. L. In The Naval Observatory merged astrometric dataset (NOMAD), American Astronomical Society Meeting Abstracts, 2004; p 48.15.

[10] Materials Project, https://materialsproject.org.

[11] Materials Cloud, https://www.materialscloud.org.

[12] Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G., "Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis." Comput. Mater. Sci. 68, (2013): 314-319.

[13] Ward, L.; Dunn, A.; Faghaninia, A.; Zimmermann, N. E.; Bajaj, S.; Wang, Q.; Montoya, J.; Chen, J.; Bystrom, K.; Dylla, M., "Matminer: An open source toolkit for materials data mining." Comput. Mater. Sci.152, (2018):60-69.

[14] Larsen, A. H.; Mortensen, J. J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dułak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C., "The atomic simulation environment—a Python library for working with atoms." J. Phys.: Condens. Matter 29 (27), (2017): 273002.

[15] Himanen, L.; Jäger, M. O.; Morooka, E. V.; Canova, F. F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S., "DScribe: Library of descriptors for machine learning in materials science." Comput. Phys. Commun. 247, (2020):106949.

[16] James, G.; Witten, D.; Hastie, T.; Tibshirani, R., An introduction to statistical learning. Springer: 2013; Vol. 112.

[17] Armaroli, N.; Balzani, V., The Future of Energy Supply: "Challenges and Opportunities." Angew. Chem. Int. Ed.46 (1-2), (2006):52-66.

[18] Zhuo, Y.; Mansouri Tehrani, A.; Brgoch, J., "Predicting the Band Gaps of Inorganic Solids by Machine Learning." J. Phys. Chem. Lett. 9 (7), (2018):1668-1673.

[19] Kim, C.; Pilania, G.; Ramprasad, R., "Machine Learning Assisted Predictions of Intrinsic Dielectric Breakdown Strength of ABX3 Perovskites." J. Phys. Chem. C 120 (27), (2016):14575-14580.

[20] Rajan, A. C.; Mishra, A.; Satsangi, S.; Vaish, R.; Mizuseki, H.; Lee, K.-R.; Singh, A. K., "Machine-Learning-Assisted Accurate Band Gap Predictions of Functionalized MXene." Chem. Mater.30 (12),(2018):4031-4038.

[21] Sendek, A. D.; Cubuk, E. D.; Antoniuk, E. R.; Cheon, G.; Cui, Y.; Reed, E. J., "Machine Learning-Assisted Discovery of Solid Li-Ion Conducting Materials." Chem. Mater.31 (2), (2019):342-352.

[21] Lin, S.; Xu, H.; Wang, Y.; Zeng, X. C.; Chen, Z., "Directly predicting limiting potentials from easily obtainable physical properties of graphene-supported single-atom electrocatalysts by machine learning." J. Mater. Chem. A8 (11), (2020):5663-5670.

[23] Wu, Y.; Lazic, P.; Hautier, G.; Persson, K.; Ceder, G., "First principles high throughput screening of oxynitrides for water-splitting photocatalysts." Energy Environ. Sci.6 (1), (2013):157-168.

[24] Greeley, J.; Jaramillo, T. F.; Bonde, J.; Chorkendorff, I.; Nørskov, J. K., "Computational high-throughput screening of electrocatalytic materials for hydrogen evolution." Nat. Mater.5 (11), (2006):909-913.

[25] Kobayashi, K.; Okumura, M.; Nakamura, H.; Itakura, M.; Machida, M.; Cooper, M. W. D., "Machine learning molecular dynamics simulations toward exploration of high-temperature properties of nuclear fuel materials: case study of thorium dioxide." Sci. Rep.12 (1), (2022):9808.

[26] Jin, M.; Cao, P.; Short, M. P., "Predicting the onset of void swelling in irradiated metals with machine learning." J. Nucl. Mater.523, (2019):189-197.

[27] Zhou, H.-C.; Long, J. R.; Yaghi, O. M., "Introduction to Metal–Organic Frameworks." Chem. Rev.112 (2), (2012):673-674.

[28] Li, S.; Chung, Y. G.; Snurr, R. Q., "High-Throughput Screening of Metal–Organic Frameworks for CO2 Capture in the Presence of Water." Langmuir 32 (40), (2016):10368-10376.

[29] Kancharlapalli, S.; Gopalan, A.; Haranczyk, M.; Snurr, R. Q., "Fast and Accurate Machine Learning Strategy for Calculating Partial Atomic Charges in Metal–Organic Frameworks." J. Chem. Theory Comput.17 (5), (2021):3052-3064.

[30] Boyd, P. G.; Chidambaram, A.; García-Díez, E.; Ireland, C. P.; Daff, T. D.; Bounds, R.; Gładysiak, A.; Schouwink, P.; Moosavi, S. M.; Maroto-Valer, M. M.; Reimer, J. A.; Navarro, J. A. R.; Woo, T. K.; Garcia, S.; Stylianou, K. C.; Smit, B., "Data-driven design of metal–organic frameworks for wet flue gas CO^2 capture."Nature 576 (7786), (2019):253-256.

[31] Simon, C. M.; Mercado, R.; Schnell, S. K.; Smit, B.; Haranczyk, M., "What Are the Best Materials To Separate a Xenon/Krypton Mixture?" Chem. Mater.27 (12), (2015): 4459-4475.

[32] Sikora, B. J.; Wilmer, C. E.; Greenfield, M. L.; Snurr, R. Q., "Thermodynamic analysis of Xe/Kr selectivity in over 137,000 hypothetical metal–organic frameworks." Chem. Sci.3 (7), (2012):2217-2223.