Applied Physics Letters PERSPECTIVE pubs.aip.org/aip/apl

Density functional theory and material databases in the era of machine learning

Cite as: Appl. Phys. Lett. **125**, 220502 (2024); doi: 10.1063/5.0235654
Submitted: 29 August 2024 · Accepted: 9 November 2024 ·
Published Online: 25 November 2024

Arti Kashyap[a]

AFFILIATIONS

School of Physical Sciences, IIT Mandi, HP 175005, India

[a]Author to whom correspondence should be addressed: arti@iitmandi.ac.in

ABSTRACT

This perspective article presents the density functional theory and traces its evolution. With the advancement in density functional theory-based computations and the efforts to collate the data generated through density functional theory, the field now has a good repository/database of materials and their properties. This repository, though not as substantial as generally used for machine learning, has nonetheless made it possible to combine density functional theory and machine learning. This article highlights current research challenges and presents an optimistic outlook for the future of "Density Functional Theory with Machine Learning" by discussing some specific examples.

I. INTRODUCTION

As of 2024, a quick search of one key phrase, "Density Functional Theory (DFT)," on the Web of Science (https://www.webofscience.com/) yields about 250 000 articles. There was only one article in the year 1965, and the number slowly increased by roughly 100 articles per year until 1990. The past few decades, however, have seen a major jump—the number of articles published on the properties of materials using DFT each year now total roughly 10 000 a year. Interestingly, even the number of review articles in this field has been about 250 per year for almost a decade now. Some bibliometric reviews that capture the data based on DFT in terms of publications, compounds studied, and properties[1,2] have also been published. How often one finds such a huge number of good review articles, bibliometric articles, and such a huge number of publications in different fields written with this frequency? A recent article by Kalita *et al.*[3] mentions that there are around 40 000 publications per year using density functional theory in various fields like materials science, chemistry, biology, and far beyond. This makes it very difficult to capture all the literature in the field. Therefore, this perspective follows an approach that highlights only the major milestones in the field that transformed the landscape of the studies performed using density functional theory. It does not capture every one of the numerous contributions of learned researchers who have enriched the field by their research using DFT or timely review articles. It puts an emphasis on the data generated by the density functional theory augmented by high throughput calculations and the databases created enabling the field to be benefited by the machine learning (ML) approach.

The article is structured into seven sections: Secs. II and III provide a brief overview of DFT and its advancements, Sec. IV discusses high-throughput calculations and data storage in databases, Sec. V introduces machine learning and its role in materials science, Sec. VI focuses on predicting formation energy and magnetic properties using ML, and Sec. VII presents the author's perspective on the future of the field. The discussion emphasizes key trends and contributions, without claiming to be an exhaustive review of the vast DFT literature.

II. DFT IN NUT SHELL

Recent advances in the study of materials using computational methods are the result of the theory given back in 1964 by Hohenberg and Kohn,[4] today known as density functional theory (DFT). The foundation of DFT is the development of quantum mechanics in the early 20th century, with ground breaking work by physicists such as Schrödinger,[5] Heisenberg,[6] Dirac,[7] Pauli,[8] and many more. The formation of quantum mechanics made understanding of the microscopic properties of the materials possible, but the major difficulties were in the solution of Schrödinger equation for the systems having electron–electron interaction. In the late 1920s and early 1930s, researchers tried to solve the many-electron system using some approximate methods, such as Hartree model[9] and the Thomas–Fermi model.[10,11] The Thomas–Fermi model used statistical mechanics and approximations to describe the electron–electron interactions. The Thomas–Fermi model was an important first step, but the method completely missed out the exchange energy of an atom as a consequence of the Pauli principle. Therefore, an exchange-energy functional was added to the

Thomas–Fermi model by Paul Dirac in 1928;[12] however, the theory remained inaccurate for most applications. The reason for this was that the error in the representation of kinetic energy as well as in the exchange energy, and the complete neglect of electron correlation. Hartree–Fock[13,14] method and its simplification by Slater[15] followed by the theorems by Hohenberg and Kohn[4] led to the wide acceptance of Hartree's ideas. Further. the mathematical and computational ease brought by the Kohn–Sham[16] equations marks the beginning of practical DFT calculations popular today. Now, DFT is the most robust theory used for computational studies of materials based on the central idea of expressing the total energy of a system in terms of the electron density rather than the wave functions of individual electrons.

However, the accuracy of any Kohn–Sham DFT calculation is limited by the approximation chosen for the exchange-correlation (XC) energy. Various exchange-correlation functionals have been developed to approximate this term. For many decades, researchers have developed many approximations for exchange-correlation functional using fundamental principles, empirical data, as well as a combination of both. Still, the development of more accurate exchange-correlation functionals is a major area of improvement in DFT. Various generations of functionals, from local density approximations (LDA)[17] to generalized gradient approximations (GGA),[18,19] have been developed. PW86 functional,[18] which apply real-space cutoffs to make GGAs and the PW91 functional[20] produced quite a precision for the binding energies. PBE[21] and BLYP[22] are the most widely used GGA to investigate materials today. In computational chemistry, Lee–Yang–Parr correlation[23] is the most used GGA, but the use of hybrid functionals has led to huge development in this area. It is to be emphasized here that the accuracy of hybrid functions is dependent on the trade-off between over-delocalization and underbinding, summarized nicely in the review by Janesko.[24] Discussion on hybrid functionals is avoided in this article so as not to diverge from the core theme of the article. Further, in solid-state computational physics, the use of hybrid functional is limited as aggregate properties in a generalized manner are captured by non-parameterized LDA and GGA quite nicely. It must be noted that the exchange-correlation functionals are the key components to calculate the material properties with high accuracy. Therefore, these functionals are being constantly improved, and functionals are developing to calculate different properties of material and enhance accuracy.

III. ADVANCEMENTS IN DFT AND PROMINENT CODES

The original DFT is based on two theorems and a few key approximations as described above. It is based on non-relativistic quantum mechanics where spin decouples completely from the spatial degrees of freedom and only enters in the orbital occupations. The non-relativistic approximation is valid where the electron velocities are much smaller than the speed of light, for example, for electrons far from the atoms. This is, however, no longer true for the electrons close to the nuclei or with non-zero angular momentum and even for valence electrons in heavy atoms where relativistic effects, particularly spin–orbit coupling is strong. Spin–orbit coupling is a significant relativistic effect where an electron's spin interacts with its orbital motion, leading to the splitting of energy levels. Therefore, for systems with relativistic effects, various approaches have evolved such as scalar relativistic,[25] fully relativistic,[26] spin–orbit coupling inclusion,[27] the use of relativistic pseudopotentials,[28] and hybrid functionals that include a portion of exact exchange in conjunction with relativistic corrections.[29–31] Another development in DFT is the non-collinear DFT, which addresses the spatial orientation of electron spins and is essential for studying complex magnetic structures.[32,33] Time-dependent DFT is an extension of DFT that is used to study excited-state properties, such as electronic excitations, optical properties, and photochemical reactions.[34] It provides a systematic way to calculate electronic excitation energies and transition probabilities. Linear response DFT extends DFT to calculate a wide range of properties systematically, including polarizabilities, dielectric constants, and NMR chemical shifts.[16] Reduced density matrix functionals[35] go beyond the electron density and aim to describe electron correlation more accurately. They are used for systems with strong electron–electron correlations and for achieving improved accuracy in the description of properties like dissociation energies. Time-dependent and real-time DFTs enable the study of dynamic processes, such as chemical reactions and electron transfer, using the DFT methods. Real-time DFT[36] allows for the simulation of molecular dynamics and electron dynamics. Fragmentation methods such as the divide-and-conquer method[37] and the fragment molecular orbital method[38] systematically enable the study of large and complex systems by dividing them into smaller, more manageable fragments. These milestones represent the evolution and continuous improvement of DFT, making it an indispensable tool in physics, materials science, theoretical and computational chemistry, and various other fields. Researchers continue to refine and expand DFT methods to address the challenges and applications.

While DFT continues to grow by incorporating various corrections to improve the accuracy, the computational power and techniques have also helped the field by making DFT realizable and faster even for complex systems. Several codes developed by different research groups at different parts of the world are now available to be used and adapted as per the interest of the researchers. Codes are also available commercially and provide good support and user experience of the software built to do DFT calculations. These codes are not only for calculating the basic electronic structure but are also combined with various formalisms to explore a wide range of material properties and phenomena. For example, electronic transport properties of a material can be computed by semiclassical or quantum or non-equilibrium Green's function (NEGF) methods using SIESTA-PRO, QuantumATK, Questaal, and Smeagol. Further, not only the codes to do the computation are now standardized, but many tools are also developed for visualization of input structure (the only input required for DFT) and for analyzing the output like band structure, density of states, and charge density.

IV. HIGH-THROUGHPUT DFT AND DATABASES

The Materials Genome Initiative (MGI) launched by the United States in 2011 (https://mgi.gov/) played a pivotal role in scaling up DFT computations. The objective of the initiative was to reduce the time and cost associated with the discovery and development of new materials. This contributed to the widespread adoption and advancement of high-throughput calculations in materials science. High-throughput computational methods systematically calculate the properties of a large number of materials by employing standardized computational workflows. These methods leverage advanced computational power, parallelization of code, and automation to efficiently navigate and explore extensive materials spaces. By automating the calculation process and optimizing the use of computational resources, high-throughput methods allow researchers to quickly screen and

identify materials with desirable properties, thereby accelerating the discovery and development of materials not known yet. The MGI also promoted the use of high-throughput DFT calculations to populate large databases with detailed information about materials' properties. Several databases have been created and continuously updated. Some of the large materials databases today are AFLOW,[39] Materials-project,[40] OQMD,[41] and NOMAD.[42] Some databases are also designed for specific purposes. C2DB[43] and JARVIS-DFT,[44] for instance, consist information of two-dimensional (2D) materials using high-through put DFT calculations while ICSD[45] is an experimental database for inorganic materials.

These databases store various information about the materials, such as basic information like material ID, chemical formula, crystal structure, lattice parameters, atomic positions, band structures, density of states, Fermi levels, formation energies, energy above hull (indicating stability), phase diagrams, elastic constants, bulk modulus, shear modulus, Poisson's ratio, magnetic moments, types of magnetic ordering, phonon dispersion curves, vibrational density of states, dielectric constants, refractive indices, and optical absorption spectra. The challenge remains in the accuracy and integrity of the data generated through high-throughput calculations. Also, high-throughput calculations require *a priori* knowledge about the materials, for example, metal or an insulator, and magnetic or nonmagnetic. The choice of the parameters, the accuracy settings, and the initial chosen type of each material may affect the data available in different datasets and it might not match. A good comparison of three major databases is presented in Ref. 46. As these databases are being used in ML models, it is highly important to know the accuracy and integrity of the dataset chosen from these databases.

V. DFT MEETS ML

The famous "Turing Test" laid the foundation for discussions on machine intelligence. It is a thought experiment or a conceptual exploration of what it means for a machine to be considered intelligent. This idea that a machine can exhibit intelligent behavior indistinguishable from a human was introduced by Alan Turing in 1950.[47] Another approach that attempts to model the brain's neural activity electronically was proposed by Minsky *et al.*[48] where they built the first artificial neural network called the SNARC (Stochastic Neural Analog Reinforcement Computer) in 1951. Soon after a self-learning checkers-playing program was reported as one of the first instances of a computer program, improving its performance through experience by Samuel *et al.*[49] Later the term "machine learning" was coined in his later work, published in 1959. Although there was limited activity in the field over the next couple of decades, the development of the back-propagation algorithm[50] its effectiveness in training multi-layer perceptron,[51] and its application in training convolutional neural networks to recognize handwritten ZIP codes[52] revitalized interest in the field during the 1980s. This development was further supported by many ML algorithms, based on variety of principles such as statistical methods (logistic regression,[53] Naïve Bayes,[54] support vector machine,[55] principal component analysis[56]), information theory (decision trees,[57] random forest[58]), biological inspiration (neural networks,[59] graph neural networks[60]), optimization techniques like gradient descent or linear programming, and some other principles. The availability of large amount of data in the past two decades has further contributed to the unprecedented rise of the field of ML in almost every domain of science. For example, the natural-language-related ML based on statistical speech recognition and statistical machine translation done with easily available large datasets has been very promising[61] in initial years of popularity of ML because of the datasets that could be used for training.

A key to understand and use ML effectively is to understand the interplay between data, algorithms, models, optimization processes, and finally the evaluations. The overall ambitious goal in any domain is to create models that generalize well to unknown data, providing accurate and reliable outputs. In the domain of study of material properties, the integration of ML techniques with DFT has been an active area of research in the recent times. Some key areas of advancement where the rigorous work has been done are designing machine learning potentials;[62,63] exploring the chemical reactions spaces;[64] image recognition in materials characterization;[65,66] prediction of crystal structures by streamlining the generation and evaluation of structures;[67] predicting material properties using sophisticated ML models, deep learning, neural networks as well as designing descriptors for better representation of the material.[68–70] While these examples illustrate the variety of ways materials science is enriched with ML, it is important to note that the earliest known use of ML in DFT dates back to the work of Tozer *et al.*,[71] who utilized neural networks to enhance the Zhao–Morrison–Parr (ZMP) method. It might be noted that constructing an exact universal functional for exchange-correlation part in DFT is an open-ended problem and as outlined in Sec. II, huge efforts are being done in this direction. Therefore, the paper by Tozer *et al.*[71] states: "*We subscribe to the view that it is very difficult to make progress in this direction and consider an alternative approach using the above ZMP potentials represented by a computational neural network (CNN).*"

The ZMP method is an inverse problem approach, where one starts with an exact electron density, often obtained from experiments or highly accurate calculations. From this exact density, the ZMP method reconstructs the Kohn–Sham potential, which includes the exchange-correlation component. By doing this, the method provides a direct link between the electron density and the exchange-correlation potential. In particular, this method and other inverse methods like Wu Yang method[72] allow to determine the effective potential of a Kohn–Sham system of non-interacting particles and thus also the exchange-correlation potential. Tozer *et al.*[71] computed potential for Ne, HF, N_2, H_2O, and N_2(1.5re), compared with popular analytic potentials, and explored the use of computational neural network for representing ZMP potential. They incorporated the neural network into a regular Kohn–Sham procedure.[16] Although the neural network-enhanced ZMP method showed improved accuracy for simple molecules (like Ne, HF, N_2, and H_2O), its ability to generalize to larger and more complex systems was not fully demonstrated. Training neural networks require significant data, and the ZMP method as primarily tested on small systems, limiting its applicability to more diverse and complex molecular or solid-state systems. However, this procedure worked specifically for the data chosen and transferability of approach remained a challenge.

However, in the past decade and more, many studies have been conducted to construct the exact exchange-correlation potential using ML and many review articles have captured the work done by various groups but not much progress has been achieved. A recent perspective article (2023)[73] summarizes the crux of the studies in this direction as follows:

"*We have demonstrated successful stories for these variants, and we encourage readers to read the respective original papers, as well as the open-source codes and examples provided. We hope that new generations of ML-DFT models will accurately construct the universal XC functional of DFT in the near future, revolutionizing the field of quantum chemistry, similar to how AlphaFold124 has transformed the field of structural biology.*"

The reason that constructing an exact universal functional for exchange-correlation using all possible routes remains an elusive goal is the inherent complexity of many-body electron interactions. These interactions are not captured in the ML models as the data that can improve the ML model are not good enough to train the model. However, the author carries an optimistic view as efforts in the direction of working with the better feature sets needed to train the model are showing promising results. For example, the recent work by Bystrom *et al.*[74] elucidates the roles of orbital-dependent and nonlocal features in learning the exchange energy and determines that both types of features provide vital and independently important information to the model. The work also demonstrates the scalability of their approach and presents a feasible path toward a universal exchange-correlation functional.

Another very recent article by Kulik *et al.*[75] discusses the use of ML in materials science and shares perspectives on current and future challenges in problems as diverse as the prediction of materials properties, the construction of force-fields, the development of exchange correlation functionals for density-functional theory, the solution of the many-body problem, and more. Many of the approaches are using material databases curated by different groups as listed in Table I. However, despite many improvements in the ML approaches, the universal behavior governing the physics of the materials is still not captured, neither with improved model nor with more data. Specific model with specific datasets can capture certain behavior but still fail when extended beyond.

To understand the fundamental roadblock in achieving the ambitious goal to create models that generalize well to unknown data and can reliably predict materials and their properties, one needs to understand the challenge in this computation. As well understood for ML, the results are as good as your training data. Therefore, this author's perspective about the use of ML with DFT revolves around the available data, manipulation, and generation of the data along with its use in a manner that ensures good training of the ML models. In the language of machine learning, the process of transforming raw data into a set of measurable and informative characteristics, called "features," is called feature extraction. The primary goal of feature extraction is to find the best suited features for training the ML model. Another closely related term is "descriptor," which is often domain-specific features tailored to capture the essential characteristics of data within a particular field.

One of the significant and initial descriptors in machine learning models for materials science is the Coulomb matrix.[76] The Coulomb matrix, an $N \times N$ matrix where N is the number of atoms in the molecule, combines the composition and geometry of a molecule. It represents the atomic structure of a molecule using Coulombic interactions. After this, a variety of descriptors have been developed to represent the properties and structures of molecules and materials for use in machine learning models, such as Smooth Overlap of Atomic Positions (SOAP),[77] Atom-Centred Symmetry Functions (ACSF),[78] and Crystal Graph Convolutional Neural Networks (CGCNN).[79] In fact, in the paper by Himanen *et al.*,[80] the collection of several descriptors in the DScribe software library are reported. DScribe has an easy-to-use python-interface and descriptors are compatible with general-purpose machine learning algorithms. They also discuss the technical details of using various descriptors and careful evaluation of the results obtained by taking the example of formation energy prediction for inorganic crystals and ionic charge prediction for organic molecules.

TABLE I. List of high throughput codes and databases from DFT computation.

HT-codes	URL
Pymatgen	http://pymatgen.org/
AiiDA	http://aiida.net/
AFLOW	http://aflowlib.org/src/aflowpi/index.html
FireWorks	https://materialsproject.github.io/fireworks/
Qmpy	https://static.oqmd.org/static/docs/index.html
AMP^2	https://github.com/MDIL-SNU/AMP2
ASE	https://wiki.fysik.dtu.dk/ase/
Atomate	https://atomate.org/
ICSD	http://www2.fiz-karlsruhe.de/icsd_home.html
COD	http://crystallography.net
Materials Project	https://materialsproject.org/
OQDM	http://oqmd.org/
CMR	https://cmr.fysik.dtu.dk/
OMDB	https://omdb.mathub.io/
MaterialsWeb	https://materialsweb.org/twodmaterials
JARVIS-DFT	https://ctcms.nist.gov/~knc6/JVASP.html
Materials Cloud	https://materialscloud.org/discover
C2DB	https://cmr.fysik.dtu.dk/c2db/c2db.html
AFLOWlib	http://aflowlib.org/
NOMAD	https://repository.nomad-coe.eu/
2DMatPedia	http://www.2dmatpedia.org/.
aNANt	http://anant.mrc. iisc.ac.in

VI. SUCCESS OR FAILURE?

As the number of publications in this domain are increasing day by day with more and more sophisticated ML models and larger and larger DFT datasets, it is good not to lose sight of the bigger picture at this juncture. ML models are excellent at recognizing patterns in data, but they do not inherently incorporate domain knowledge or physical laws. This may produce physically meaningless predictions if they are not carefully designed. Integrating domain expertise with ML algorithms, such as embedding physical constraints or using physics-informed ML models, seems to be the direction holding the promise to be able to get insights in materials understanding as well as predictions. This section reviews a few example properties and what has been achieved so far to get an overall understanding of the efforts in exploring these properties. This is done by concentrating on two critical aspects: (i) the formation energy of materials, which is essential to determine whether a material can be synthesized and is thermodynamically stable and (ii) the intrinsic magnetic properties, which arise from the complex interactions within the material and are crucial for

understanding and designing materials for magnetic and spintronic applications.

A. Formation energy

Before talking about ML models being used for the prediction of formation energies, one needs to understand that different databases might lead to different values/errors. This comparison is presented in Ref. 46, which effectively concludes that the formation energy is easily reproduced across different high-throughput DFT based databases, though this might not be the case with other properties such as magnetic properties. Now, starting with the datasets from any of these large databases, several studies have been done to demonstrate that various machine learning models can predict formation energy with good accuracy. Common to all studies is the focus on the feature set chosen carefully by designing better and better descriptors.

In Ref. 81, the authors refer to various contributions (Refs. 13–26, 31, and 39 given in Ref. 81) where different attributes (or features) are chosen for variety of properties and observed that most attribute sets were primarily based on statistics of the properties of constituent elements. From this, they concluded that a general-purpose attribute set should contain the statistics of a wide variety of elemental properties and created a bigger set of attributes that can be used for materials with any number of constituent elements. They used a set of 145 attributes calculated using Magpie, which is an open-source platform, and evaluated many possible ML models for each individual problem. Comparing other works and their own evaluation, the authors conclude that the ensembles of decision trees perform best with the chosen attribute set. They also partitioned the data by grouping the data set into chemically similar segments and training a separate model on each subset. This strategy resulted in quite an accurate prediction of formation energy with the datasets of 228 676 compounds from OQMD. These authors also used experimental dataset and predicted other properties as well, but we will restrict our discussion here to only formation energy.

In another example, formation energies are predicted using machine learning to investigate all possible elpasolites ($\sim 2 \times 10^6$ configurations) made up of the main-group elements.[82] The results have been used to identify the most strongly bound elpasolites as well as to investigate the energy and bonding trends at crystal structure sites, leading to a new "elpasolite order" of elements, consistent with the bonding physics in the elpasolite crystal structure. They identified and added 128 structures to the convex hull of the Material Project database. Their ML model is based on kernel ridge regression and constructed by focusing on the energy difference between the crystal energy and the sum of static, atom-type dependent, averaged atomic energy contributions. For training and evaluation, they have generated DFT formation energies for two datasets of elpasolites and included a Materials Project dataset consisting of $\sim 0.5 \times 10^3$ ternary crystals in the $ThCr_2Si_2$ prototype and made up of 84 different atom types. A mean absolute error (MAE) of 0.10 eV/atom is achieved with the larger dataset in this work, which is comparable to the DFT obtained values with estimated MAE of ~0.19 eV/atom compared to experiments, and with other groups, which report DFT errors on the order of 0.1 eV/atom for transition metal oxides and elemental solids (Refs. 30 and 31 of Ref. 82). MAE, which is the average of the absolute differences between predicted values and actual observed values, is a good evaluation criterion across many ML models as it is easy to understand and interpret and provides a straightforward measure of the average error in the same units as the data. The authors conclude that their crystal representation of elpasolite structures accounts for all the necessary degrees of freedom and hence the results are accurate. In other words, the simple descriptor based on energy gives good results with respect to the formation energy.

One of the fascinating works is based on neural networks by Xie *et al.*[79] By simply considering the connection of atoms in the crystal, they build convolutional neural networks (CNN) on top of crystal graphs (CG) generated from crystal structures. They trained the model using the database of a diverse set of inorganic crystals ranging from simple metals to complex minerals from the Materials Project. The best MAEs achieved in this work are 0.136 and 0.039 eV/atom, respectively, for two different convolution functions. As reported, 90% of the crystals were predicted within 0.3 and 0.08 eV/atom errors. The author also reports a comparison with other work having the MAE of the DFT calculation with respect to experimental measurements in the Open Quantum Materials Database to be 0.081–0.136 eV/atom.

The three distinct examples mentioned above use very different approaches and very different datasets, but in all cases, the formation energy is generally predicted very well. This demonstrates the effectiveness of ML models in predicting formation energies across different materials and methodologies. Similar observations are made in the work done by the author's own group for Fe-based[83] and Mn-based[84] datasets from the Materials project where formation energy comes out to be quite accurate provided the datasets are partitioned as per the magnetic ordering of the systems considered in the data. Some details of these results are given in the discussion of the magnetic properties.

B. Magnetic properties

DFT provides a powerful framework for calculating magnetic moments in materials. By using spin-polarized calculations, DFT can predict both local and total magnetic moments, which are essential for understanding and designing materials with specific magnetic properties. However, the accuracy of these predictions depends on the choice of exchange-correlation functional and the specific details of the material being studied. Nevertheless, all major DFT databases contain the data about magnetic properties. For example, Materials Project, AFLOW, and OQMD provide magnetic ordering information, total magnetic moment, and, sometimes, local magnetic moments for a wide range of materials. They include data derived from spin-polarized DFT calculations and sometimes results for multiple magnetic configurations to determine the ground state as well.

Despite the availability of DFT data for magnetic properties, not many papers talk specifically about the prediction of magnetic materials or materials with tailor made magnetic properties using ML. However, some results for specific systems are reported. Most recently (2024), Kaba *et al.*[85] reported the prediction of large magnetic moment materials using random forests and two state-of-the-art deep-learning graph convolutional algorithms, CGCNN, and MEGNet, on the Materials Project dataset. The study also reports good results with respect to the formation energy as discussed in Sec. VI A; however, for the prediction of magnetic materials, their sophisticated models based on deep learning also fail and predictions are not accurate when compared with the experimental data from the ICSD database.

The work by Pham *et al.*,[86] though mainly focused on the descriptor for better description of the materials, also reports the

magnetic properties using ML. Their study demonstrates that the more important factor for the prediction of the magnetic properties is the choice of the descriptor rather than the ML model. Their orbital field matrix (OFM) descriptor could accurately reproduce the DFT-calculated local magnetic moments of transition-metal sites in lanthanide metal and transition-metal based alloys. However, the prediction of new materials is still not reported.

At this point of time, the author would like to share her own experience in the search of magnetic materials using ML. The work done by the group is reported in Refs. 83, 84, 87, and 88. Understanding the importance of the descriptor, the main emphasis in these works is in choosing or designing the right descriptor. The OFM descriptor by Pham *et al.*[86] was found to be the most suitable for the desired purpose, and a slightly modified version of this descriptor, known as Hund's matrix descriptor, was proposed, which reduces the computational cost and gives results that are slightly better.[87] Apart from the carefully chosen descriptor, emphasis in these works is put on the right selection of the dataset. In Ref. 84, Mn-based magnetic materials are proposed using random forest ML model and a few materials out of over 4000 hypothetical structures generated through atom substitution in existing materials are reported to be of interest with respect to their magnetism. Chosen eight materials are subjected to DFT calculations. DFT validation revealed a strong correlation with the machine learning results when following the same magnetic configuration initialization (i.e., ferromagnetic) used in the Materials Project database. However, it is crucial to acknowledge that the true magnetic ground state may not always be reached with this specific initialization method. In Ref. 88, the chemical stability and magnetic moment of 2D MXene materials is predicted using data from the C2DB database. Interestingly, crystal graph-based descriptor is not useful in predicting magnetic moments for these 2D materials, although the results for the formation energy seem to be comparable with the other works mentioned above. A simple descriptor made from elemental properties was found to be useful in predicting the magnetic moment for the materials with an MAE of $0.498\mu_B$ per formula unit using the XGB regression algorithm.

Finally, a recent paper by Boulay *et al.*[89] used machine learning to detect the position, shape, and size of skyrmions, which are topologically stable, vortex-like spin configurations found in certain magnetic materials. Magnetic imaging is typically employed for detecting these, and a frequently used technique is magneto-optical Kerr microscopy. Conventional image analysis packages are used for the task, and this requires manual treatment. Authors report a convolutional neural network specifically designed to detect the position and shape of skyrmions in their measurements. The results of this study show that a well-trained network is a viable method of automating data preprocessing in magnetic microscopy. The approach is easily extendable to other spin structures and other magnetic imaging methods. Another earlier work by Nelson *et al.*[90] though the study does not include DFT databases, introduces a range of machine-learning models constructed by using experimental data for about 2500 known magnets to predict the Curie temperature of ferromagnets. The chemical composition of a compound was the only feature considered for determining Curie temperature. They reported that the best model can predict Curie temperatures with an accuracy of about 50 K. However, this work does not include any DFT data and ML models are solely trained with the experimental data. The results of these two contributions are not surprising. In case of skyrmions, it uses image identification, which is a well-established application of ML. It typically involves the use of convolutional neural networks (CNNs), a type of deep learning model particularly effective for processing and recognizing patterns in images, which is the model used in Ref. 89, and the results of classifying or identifying images are very accurate. In case of the Curie temperature prediction in Ref. 90, the data used for training are all experimentally verified, and great care is paid to collate and clean these data. Along with this, the model chosen captures the chemical composition quite nicely.

VII. PERSPECTIVE

Before presenting the conclusion/perspective of this study, it is worthwhile to understand few fundamental characteristics of the DFT based databases with respect to the two properties we considered in this article. DFT databases provide accurate and consistent formation energy values derived from *ab initio* calculations. These high-quality data points serve as a solid foundation for training ML models, allowing them to learn the underlying relationships effectively. On the other hand, the magnetic calculations are initialized in the ferromagnetic configuration in all the material databases. Starting from a ferromagnetic state is useful for standardization, but it does not accurately capture the true magnetic ground state for all materials. Materials that exhibit antiferromagnetism, ferrimagnetism, or more complex magnetic orders are either not available or not represented in these datasets accurately.

Now, formation energy is a well-defined thermodynamic quantity directly related to the stability of a compound. This clear definition makes it easier for ML models to capture and predict using input features. Many ML models use features that capture important physical and chemical characteristics, e.g., constituent elements, structural information, lattice parameters, atomic radii, electronegativity, and bonding environment, which are relevant to formation energy. Also, the formation energy data are widely available in DFT databases for various materials, providing the large dataset required for training robust ML models. The diversity and volume of data help the models generalize well across different types of materials. Finally, the data about formation energy in DFT databases are quite robust in terms of its accuracy and comparability across different databases. All these factors contribute to the success of ML models in accurately predicting formation energies using DFT data.

On the other hand, magnetic properties are often challenging to predict accurately using machine learning models trained on DFT datasets due to several factors. The most important factor is the complex magnetic interactions, such as exchange interaction, dipole–dipole interaction, electronic correlations, spin–orbit coupling, and crystal field effects, which are difficult to model accurately. Also, magnetic properties are highly sensitive to slight changes in atomic positions, bonding, and defects, making them harder to capture in a dataset. Another important factor is that the DFT calculations for magnetic properties may not be accurate enough compared to formation energies and high-quality experimental magnetic data for training is often limited and not easily available. As a result, the magnetic properties predicted by these initial conditions might not always reflect the actual magnetic behavior of the material in its true ground state. Magnetic properties also change with temperature, requiring models to capture not just ground-state properties but also thermal excitations, which add complexity.

With the help of only two critical aspects (one straightforward: formation energy and another complex: magnetic properties), this perspective presents the following view:

The promise of prediction of materials with the desired properties using machine learning models is quite fascinating and probably possible with fair accuracy, as one can see in the number of publications arising in this domain. However, the accuracy of predictions depends on the quality and representativeness of the training data. Quality here refers not only to complete, clean, and accurate data but also to the data fidelity, that is the correctness of the data. For example, if one is trying to predict the bandgap by training the ML model with DFT dataset created by doing simple DFT-LDA calculation without applying the proper treatment for necessary corrections to get the correct bandgap, no sophisticated ML model can provide good results as the data to train the model is not accurate and true. Those who are experienced in DFT also know that the same correction treatment does not work for all materials. Each material must be carefully checked for the accuracy of the calculations after doing the DFT computation. While high throughput calculations make it possible to do a large number of calculations, the accuracy of each material present in the database requires careful examination. One way forward for this can be that the databases should be created where DFT data and experimental data are put side by side wherever it is possible. A high-quality data from DFT with careful study of each material present in the dataset and experimental data should be gathered and used for training purpose. Second, a careful insight into the physics of the interactions should be included in the descriptor. A simple descriptor created from elemental properties may not be enough in general case but may be suitable for some specific cases as shown above. A complex descriptor designed based on the borrowed concepts from computer science used with celebrated machine learning or deep learning algorithms also might not work. The optimistic solution seems to be able to make the descriptor capture as much as physics it can and gather high-quality data even if it is small. Machine learning has benefited from a large amount of available data for physical objects like images as it makes the machine learn quickly but for the objects like materials where intricate details are hidden in their complex electronic structure, focusing on gathering more and more data, is not enough. Attention has to be paid to the "Quality of the data" and thoughtfully designed "Descriptors."

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Arti Kashyap: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Writing – original draft (equal).

DATA AVAILABILITY

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

REFERENCES

[1]R. Haunschild, A. Barth, and W. Marx, "Evolution of DFT studies in view of a scientometric perspective," J. Cheminform. **8**(1), 1–12 (2016).

[2]R. Haunschild, A. Barth, and B. French, "A comprehensive analysis of the history of DFT based on the bibliometric method RPYS," J. Cheminform. **11**(1), 72 (2019).

[3]B. Kalita, L. Li, R. J. Mccarty, and K. Burke, *Learning to Approximate Density Functionals* (ACS, 2012).

[4]P. Hohenberg and W. Kohn, "Inhomogeneous electron gas," Phys. Rev. **136**(3B), B864–871 (1964).

[5]E. Schrodinger, "An undulatory theory of the mechanics of atoms and molecules," Phys. Rev. **28**(6), 1049–1070 (1926).

[6]W. Heisenberg, "Über quantentheoretische Umdeutung kinematischer und mechanischer Beziehungen," Z. Phys. **33**(1), 879–893 (1925).

[7]P. A. M. Dirac, "Quantum mechanics of many-electron systems," Proc. R. Soc. London Ser. A **123**(792), 714–733 (1929).

[8]W. Pauli, "Über den Zusammenhang des Abschlusses der Elektronengruppen im Atom mit der Komplexstruktur der Spektren," Z. Phys. **31**(1), 765–783 (1925).

[9]R. D. Hartree, "The wave mechanics of an atom with a non-coulomb central field. Part I. Theory and methods," Ann. Phys. **24**, 89–110 (1928).

[10]E. Fermi, *Un metodo statistico per la determinazione di alcune proprietá dell 'atomo* (Endiconti Accademia Nazionale dei Lincei, 1927).

[11]L. H. Thomas, "The calculation of atomic fields," Math. Proc. Camb. Philos. Soc. **23**(5), 542–548 (1927).

[12]P. A. M. Dirac, "The quantum theory of the electron," Proc. R. Soc. A **117**, 610 (1928).

[13]D. R. Hartree and W. Hartree, "Self-consistent field, with exchange, for beryllium," Proc. R. Soc. A 150, 9 (1935).

[14]V. Fock, "Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems," Z. Phys. **61**(2), 126–148 (1930).

[15]J. C. Slater, "A simplification of the Hartree-Fock method," Phys. Rev. **81**, 385 (1951).

[16]W. Kohn and L. J. Sham, "Self-consistent equations including exchange and correlation effects," Phys. Rev. **140**, A1133 (1965).

[17]E. Gross and W. Kohn, "Local density-functional theory of frequency-dependent linear response," Phys. Rev. Lett. **55**, 2850 (1985).

[18]J. Perdew and W. Yue, "Accurate and simple density functional for the electronic exchange energy: Generalized gradient approximation," Phys. Rev. B **33**, 8800 (1986).

[19]J. Perdew and Y. Wang, "Accurate and simple analytic representation of the electron-gas correlation energy," Phys. Rev. B **45**, 13244 (1992).

[20]K. Burke, J. P. Perdew, and Y. Wang, "Derivation of a generalized gradient approximation: The PW91 density functional," in *Electronic Density Functional Theory: Recent Progress and New Directions*, edited by J. F. Dobson, G. Vignale, and M. P. Das, (Springer US, Boston, MA, 1998), pp. 81–111.

[21]J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," Phys. Rev. Lett. **77**, 3865 (1996).

[22]A. D. Becke, "Density-functional exchange-energy approximation with correct asymptotic behavior," Phys. Rev. A: Gen. Phys. **38**(6), 3098–3100 (1988).

[23]W. Yang, R. G. Parr, and C. Lee, "Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density," Phys. Rev. B: Condens. Matter **37**(2), 785–789 (1988).

[24]B. G. Janesko, "Replacing hybrid density functional theory: Motivation and recent advances," Chem. Soc. Rev. **50**(15), 8470–8495 (2021).

[25]D. Koelling and B. N. Harmon, "A technique for relativistic spin-polarised calculations," J. Phys. C: Solid State Phys. **10**, 3107 (1977).

[26]R. O. Jones and O. Gunnarsson, "The density functional formalism, its applications and prospects," Rev. Mod. Phys. **61**, 689–746 (1989).

[27]A. H. MacDonald, W. E. Pickett, and D. D. Koelling, "A linearised relativistic augmented-plane-wave method utilizing approximate pure spin basis functions," J. Phys. C: Solid State Phys. **13**, 2675 (1980).

[28]W. R. Wadt and P. J. Hay, "Ab initio effective core potentials for molecular calculations. Potentials for main group elements Na to Bi," J. Chem. Phys. **82**, 284 (1985).

[29]A. D. Becke, "Density-functional thermochemistry. III. The role of exact exchange," J. Chem. Phys. **98**(7), 5648–5652 (1993).

[30]W. Klopper, H. P. Lüthi, T. Brupbacher, and A. Bauder, "Ab initio computations close to the one-particle basis set limit on the weakly bound van der Waals complexes benzene–neon and benzene–argon," J. Chem. Phys. **101**(11), 9747–9754 (1994).

[31]C. Adamo and V. Barone, "Toward reliable density functional methods without adjustable parameters: The PBE0 model," J. Chem. Phys. **110**(13), 6158–6170 (1999).
[32]P. Novák, J. Kuneš, W. E. Pickett, W. Ku, and F. R. Wagner, "Self-interaction correction and contact hyperfine field," Phys. Rev. B **67**(14), 140403 (2003).
[33]D. Hobbs, G. Kresse, and J. Hafner, "Fully unconstrained noncollinear magnetism within the projector augmented-wave method," Phys. Rev. B **62**(17), 11556–11570 (2000).
[34]E. Runge and E. K. U. Gross, "Density-functional theory for time-dependent systems," Phys. Rev. Lett. **52**(12), 997–1000 (1984).
[35]T. L. Gilbert, "Hohenberg-Kohn theorem for nonlocal external potentials," Phys. Rev. B **12**(6), 2111–2120 (1975).
[36]K. Yabana and G. F. Bertsch, "Time-dependent local-density approximation in real time," Phys. Rev. B **54**(7), 4484–4487 (1996).
[37]W. Yang, "Direct calculation of electron density in density-functional theory," Phys. Rev. Lett. **66**(11), 1438–1441 (1991).
[38]E. Ikeo, T. Asada, T. Nakano, M. Uebasayi, and K. Kitaura, "Fragment molecular orbital method: An approximate computational method for large molecules," Chem. Phys. Lett. **313**, 701–706 (1999).
[39]S. Curtarolo, W. Setyawan, S. Wang *et al.*, "AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations," Comput. Mater. Sci. **58**, 227–235 (2012).
[40]A. Jain, S. P. Ong, G. Hautier, W. Chen *et al.*, "Commentary: The Materials Project: A materials genome approach to accelerating materials innovation," APL Mater. **1**(1), 011002 (2013).
[41]S. Kirklin, J. E. Saal, B. Meredig *et al.*, "The Open Quantum Materials Database (OQMD): Assessing the accuracy of DFT formation energies," npj Comput. Mater. **1**(1), 15010 (2015).
[42]C. Draxl and M. Scheffler, "NOMAD: The FAIR concept for big data-driven materials science," MRS Bull. **43**(9), 676–682 (2018).
[43]S. Haastrup1, M. Strange1, M. Pandey *et al.*, "The Computational 2D Materials Database: High-throughput modeling and discovery of atomically thin crystals," 2D Mater. **5**(4), 042002 (2018).
[44]K. Choudhary, K. F. Garrity, A. C. E. Reid *et al.*, "The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design," npj Comput. Mater. **6**(1), 173 (2020).
[45]A. Belsky, M. Hellenbrandt, V. L. Karen, and P. Luksch, "New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design," Acta Crystallogr. B **58**(3), 364–369 (2002).
[46]V. I. Hegde, C. K. H. Borg, Z. del Rosario *et al.*, "Quantifying uncertainty in high-throughput density functional theory: A comparison of AFLOW, Materials Project, and OQMD," Phys. Rev. Mater. **7**(5), 53805 (2023).
[47]A. M. TURING, "I.—Computing machinery and intelligence," Mind **LIX**(236), 433–460 (1950).
[48]M. Minsky, "Neural nets and the brain model problem," Ph.D. dissertation (Princeton University, 1951).
[49]A. L. Samuel, "Some studies in machine learning using the game of checkers," IBM J. Res. Dev. **3**(3), 210–229 (1959).
[50]D. E. Rumelhart and J. L. McClelland, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1 (MIT Press, Cambridge, MA, 1986).
[51]D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," Nature **323**(6088), 533–536 (1986).
[52]Y. LeCun, B. Boser, J. S. Denker *et al.*, "Backpropagation applied to handwritten zip code recognition," Neural Comput. **1**(4), 541–551 (1989).
[53]D. R. Cox, "The regression analysis of binary sequences," J. R. Stat. Soc. Ser. B (Methodol.) **20**(2), 215–242 (1958). http://www.jstor.org/stable/2983890.
[54]T. Bayes, "LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S," Philos. Trans. R. Soc. London **53**, 370–418 (1763).
[55]V. Vapnik, *The Nature of Statistical Learning Theory* (Springer, New York, 1995).
[56]K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," London, Edinburgh, Dublin Philos. Mag. J. Sci. **2**(11), 559–572 (1901).
[57]J. R. Quinlan, "Induction of decision trees," Mach. Learn. **1**(1), 81–106 (1986).
[58]T. K. Ho, "Random decision forests," in *Proceedings of the 3rd International Conference on Document Analysis and Recognition* (IEEE, 1995).
[59]J. Schmidhuber, "Deep learning in neural networks: An overview," Neural Netw. **61**, 85–117 (2015).
[60]F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," IEEE Trans. Neural Netw. **20**(1), 61–80 (2009).
[61]A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," IEEE Intell. Syst. **24**(2), 8–12 (2009).
[62]A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons," Phys. Rev. Lett. **104**(13), 136403 (2010).
[63]J. S. Smith, O. Isayev, and A. E. Roitberg, "ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost," Chem. Sci. **8**(4), 3192–3203 (2017).
[64]S. Stocker, G. Csányi, K. Reuter, and J. T. Margraf, "Machine learning in chemical reaction space," Nat. Commun. **11**(1), 5505 (2020).
[65]M. Botifoll, I. Pinto-Huguet, and J. Arbiol, "Machine learning in electron microscopy for advanced nanocharacterization: Current developments, available tools and future outlook," Nanoscale Horiz. **7**(12), 1427–1477 (2022).
[66]S. V. Kalinin, D. Mukherjee, K. Roccapriore *et al.*, "Machine learning for automated experimentation in scanning transmission electron microscopy," npj Comput. Mater. **9**(1), 227 (2023).
[67]M. Lu, S. Rao, H. Yue, J. Han, and J. Wang, "Recent advances in the application of machine learning to crystal behavior and crystallization process control," Cryst. Growth Des. **24**(12), 5374–5396 (2024).
[68]Y. Zhang, W. Xu, G. Liu, Z. Zhang, J. Zhu, and M. Li, "Bandgap prediction of two-dimensional materials using machine learning," PLoS One **16**(8), e0255637 (2021).
[69]F. Qayyum, M. A. Khan, D.-H. Kim, H. Ko, and G.-A. Ryu, "Explainable AI for material property prediction based on energy cloud: A Shapley-driven approach," Materials **16**(23), 7322 (2023).
[70]K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, "Machine learning for molecular and materials science," Nature **559**(7715), 547–555 (2018).
[71]D. J. Tozer, V. E. Ingamells, and N. C. Handy, "Exchange-correlation potentials," J. Chem. Phys. **105**(20), 9200–9213 (1996).
[72]Q. Wu and W. Yang, "A direct optimization method for calculating density functionals and exchange–correlation potentials from electron densities," J. Chem. Phys. **118**(6), 2498–2509 (2003).
[73]J. Wu, S.-M. Pun, X. Zheng, and G. Chen, "Construct exchange-correlation functional via machine learning," J. Chem. Phys. **159**(9), 090901 (2023).
[74]K. Bystrom and B. Kozinsky, "Nonlocal machine-learned exchange functional for molecules and solids," Phys. Rev. B **110**(7), 075130 (2024).
[75]H. J. Kulik, T. Hammerschmidt, J. Schmidt *et al.*, "Roadmap on machine learning in electronic structure," Electron. Struct. **4**(2), 023004 (2022).
[76]M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, "Fast and accurate modeling of molecular atomization energies with machine learning," Phys. Rev. Lett. **108**(5), 058301 (2012).
[77]A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," Phys. Rev. B **87**(18), 184115 (2013).
[78]J. Behler, "Atom-centered symmetry functions for constructing high-dimensional neural network potentials," J. Chem. Phys. **134**(7), 074106 (2011).
[79]T. Xie and J. C. Grossman, "Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties," Phys. Rev. Lett. **120**(14), 145301 (2018).
[80]L. Himanen, M. O. J. Jäger, E. V. Morooka *et al.*, "DScribe: Library of descriptors for machine learning in materials science," Comput. Phys. Commun.. **247**, 106949 (2020)..
[81]L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, "A general-purpose machine learning framework for predicting properties of inorganic materials," npj Comput. Mater. **2**(1), 16028 (2016).
[82]F. A. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, "Machine learning energies of 2 million elpasolite (ABC2D6) crystals," Phys. Rev. Lett. **117**(13), 135502 (2016).
[83]Y. Khatri, R. Sharma, A. Shah, and A. Kashyap, "Magnetization in iron based compounds: A machine learning model analysis," AIP Adv. **13**(2), 025318 (2023).

[84]Y. Khatri and A. Kashyap, "Advancing magnetic material discovery through machine learning: Unveiling new manganese-based materials," APL Mach. Learn. **1**(4), 046113 (2023).
[85]S.-O. Kaba, B. Groleau-Paré, M.-A. Gauthier, A.-M. S. Tremblay, S. Verret, and C. Gauvin-Ndiaye, "Prediction of large magnetic moment materials with graph neural networks and random forests," Phys. Rev. Mater. **7**(4), 044407 (2023).
[86]T. Lam Pham, H. Kino, K. Terakura *et al.*, "Machine learning reveals orbital interaction in materials," Sci. Technol. Adv. Mater. **18**(1), 756–765 (2017).
[87]Y. Khatri and A. Kashyap, "Hund's matrix: A structure-based descriptor to predict the magnetic properties " J. Magn. Magn. Mater. **597**, 172026 (2023).
[88]Y. Khatri, V. Atpadkar, A. Agarwal, and A. Kashyap, "Magnetic MXene: A machine-learning model with small data," IEEE Trans. Magn. **59**(11), 1 (2023).
[89]I. Labrie-Boulay, T. B. Winkler, D. Franzen, A. Romanova, H. Fangohr, and M. Kläui, "Machine-learning-based detection of spin structures," Phys. Rev. Appl. **21**(1), 014014 (2024).
[90]J. Nelson and S. Sanvito, "Predicting the Curie temperature of ferromagnets using machine learning," Phys. Rev. Mater. **3**(10), 104405 (2019).