MACHINE LEARNING COUPLED WITH DENSITY FUNCTIONAL THEORY: A FUTURISTIC DUO

Abstract

Since, the beginning of the human age man is interested in develop- ing machines. From the stone age to era of artificial intelligence we have come a long way. Now we have Machine learning algorithms i.e. the computer programs which can improve themselves through self learning. Density functional theory which accquired a large attention of the chemists as well as physicists during last decades, is now cou- pled with ML which can help in improving the exchange correlation functionals which were the major concerns as far as the accuracy of the DFT is concernerd. In this chapter a brief introduction of the DFT and ML is provided with their combined application in different fields like heterogeneous catalysis, material property prediction etc.

Keywords: Machine learning (ML), Density Functional Theory (DFT), Artificial Neural Networks (ANN), environmental remediation

.

Author

Ravi Kumar
Department of Chemistry and Center for advanced studies,
Panjab University,
Sec. 14, Chandigarh, Chandigarh,India.
ravikumar2167@gmail.com

I. INTRODUCTION

All branches of chemistry are united by the principles and ideas found in theoretical chemistry[1]. The systematisation, elaboration, and hierarchy-building of chemical laws, principles, and norms occur within the context of theoretical chemistry. The idea of the connectivity of a molecular system's structure and properties holds a fundamental position in theoretical chemistry. It makes use of mathematical and physical techniques to correlate, comprehend, and fore- cast the thermodynamic and kinetic properties of chemical systems as well as their structures and dynamics. In its broadest meaning, it refers to the expla- nation of chemical processes using theoretical physics techniques. In contrast to theoretical physics, theoretical chemistry frequently employs semi-empirical and empirical methodologies in addition to approximate mathematical tech- niques in order to account for the great complexity of chemical systems [2].In recent years, it has mostly focused on quantum chemistry, or using quantum mechanics to solve chemical-related problems. Molecular dynamics, statistical thermodynamics, theories of electrolyte solutions, reaction networks, polymerization, catalysis, molecular magnetism, and spectroscopy are further important components.

The study of chemical structure and the study of chemical dynamics might be considered to be the two main branches of contemporary theoreti- cal chemistry. Studies of electronic structure, potential energy surfaces, force fields, vibrational-rotational motion, and the equilibrium characteristics of condensed-phase systems and macromolecules are all included in the first cate-gory. Bimolecular kinetics, the collision theory of reactions and energy transfer, unimolecular rate theory, metastable states, condensed-phase dynamics, and macromolecular dynamics are all parts of chemical dynamics. Theoretical chemistry can be divided into several branches as discussed:

1. **Quantum Chemistry**: It is the application of basic interactions from quan- tum mechanics to chemical and physico-chemical issues. The most often modelled properties include those that are spectroscopic and magnetic.

2. **Computational Chemistry:** is the use of scientific computing to chemistry, using approximation techniques including force field methods, semiempirical approaches (like PM3), density functional theory, and Hartree-Fock and post-Hartree-Fock. The most often predicted property is molecular shape. Additionally, computers are able to capture and Fourier transform infrared data into frequency information in addition to predicting vibrational spectra and vibronic coupling. The projected form is supported by a comparison to predicted vibrations.

3. **Molecular Modelling:** molecular structure modelling techniques that don't always use quantum mechanics. Examples include drug design, com- binatorial chemistry, protein-protein docking, and molecular docking. The motivating force behind this graphical technique is the fitting of shape and electric potential.

4. **Molecular Dynamics (MD):** applying classical mechanics to model the movement of an assemblage of atoms and molecules' nuclei. Van der Waals forces and temperature both influence how molecules in an ensemble rearrange one another.

5. **Molecular Mechanics (MM):** Utilising potentials, MM models the energy surfaces of intra- and intermolecular interactions. The latter are often parameterized using computations from the beginning.

6. **Mathematical Chemistry:** Using mathematical techniques, the molecule structure is discussed and predicted without necessarily referencing quantum mechanics. With the help of the mathematical field of topology, scientists can forecast the characteristics of flexible bodies with finite sizes, such as clusters.

7. **Theoretical Chemical Kinetics:** Study of the related differential equations in the theory of the dynamical systems connected to reactive substances, the activated complex.

8. **Cheminformatics:** Often called chemo in formatics, is the application of computer and informational tools to crop information in order to address chemistry-related issues.

9. **Chemical Engineering:** The practise of conducting research and development on industrial processes using chemistry. This makes it possible to develop and enhance both new and existing items as well as manufacturing techniques.

The inhomogeneous electron gas, which consists of a collection of interacting point electrons travelling quantum mechanically in the potential field of a collection of atomic nuclei, is the most common theoretical repre- sentation of solid-state and/or molecular systems. which, according to the Born-Oppenheimer approximation, are static. The employment of approxi- mation approaches, including the most fundamental ones—the independent electron approximation, the Hartree theory, and the Hartree-Fock theory—is typically necessary to solve such models. Density Functional Theory (DFT), is an alternative strategy that has, over the past 30 years or more, been more and more popular for solving these issues [3]. This approach has the dual benefits of being computationally straightforward and capable of handling a variety of issues with a high level of precision.

Since the beginning of time, humans have used a variety of instruments to complete various jobs more quickly. The inventiveness of the human mind produced a variety of machines. These devices made life easier for humans by allowing them to fulfil a variety of demands, such as travel, industry, and com- puting. And among these, there is machine learning. Sometimes, even after viewing the data, we are unable to evaluate or extrapolate the information. We then use Machine Learning (ML) in that situation. The availability of a large number of data sets has increased demand for ML. ML is used in many industries to retrieve pertinent data. Learning from the data is the goal of ML. How to make robots learn on their own without being explicitly programmed has been the subject of numerous studies. Numerous mathematicians and pro- grammers use a variety of techniques to solve this problem, which involves large amounts of data.

As early as the late 20*th* century, ML was being used in catalysis. One of the first studies to employ neural networks to establish a connection between a catalyst's physicochemical parameters and catalytic performance [4]. Fol- lowing Himmelblau's review of the use of Artificial Neural Networks (ANN) in the field of Chemical engineering, ANN has found use in a number of cat- alytic processes, including the steam reforming and dry reforming of CH_4, water-gas shift reactions, and the epoxidation of large olefins [5]. Large-scale simulations and their analysis could be accelerated to previously unachievable scales with the use of artificial intelligence and robust data analysis. For the analysis of such intricate data sets, ML is a fast expanding area. In the area of electronic structure simulations, where DFT assumes the significant role of being the most popular electronic structure approach, it has recently gained traction. As a result, DFT computations place a heavy burden on global academic high-performance computing systems. Simulating larger systems is made possible by accelerating these with ML, which can also lower the amount of resources needed. Consequently, the fusion of DFT and ML has the potential to significantly advance applications involving electronic structures, such as the identification of new chemical reaction routes and in silico materials [6]. Cal- culations from the electronic structure theory support experimental research in chemistry and material science by enabling a quantum-level understanding of matter. They are crucial in addressing complex scientific and technological issues. Modern, high-performance computational resources have in turn made it possible to do extensive simulations of electrical structures. However, the ever-growing need for precise first-principles data makes even the most effec- tive simulation software impractical. On the other hand, a number of research domains have seen a sharp increase in the usage of data-driven ML techniques. These techniques are becoming more and more important as they are used to expedite, replace, or enhance conventional electronic structure theory tech- niques. DFT is frequently the foundation of electronic structure computational workflows. While DFT offers a practical compromise between computing cost and accuracy, combining it with ML can result in huge speedups.

II. MACHINE LEARNING: A TOOL FOR PREDICTION

The issue of creating computers that learn automatically through use is addressed by ML. The convergence of computer science and statistics, as well as the foundation of AI and data science, make it one of the technical domains with the fastest growth rates today. ML has advanced recently as a result of the creation of new learning theories and algorithms as well as the con- tinual explosion in the accessibility of online data and low-cost processing. Science, technology, and business have all adopted data-intensive ML tech- niques, which has increased the use of evidence in decision-making in numerous fields, such as marketing, manufacturing, healthcare, and financial modelling. A ML algorithm's learning system is composed of three primary components.

1. **A Decision-Making Process:** ML algorithms are typically used to create a prediction or classify data. The algorithm will generate an estimate about a pattern in the data based on some input data, which may be labelled or unlabeled.

2. **An Error Function:** It measures the accuracy of the model's prediction. If there are known examples, a comparison can be made to judge the model's accuracy.

3. **A Model Optimisation Process:** If the model can match the training set's data points more accurately, weights are modified to lessen the difference between the model estimate and the known example. This ”evaluate and optimise” procedure will be repeated by the algorithm, with weights being updated automatically, until a predetermined level of accuracy is reached.
   Many ML algorithms are frequently employed [7–9]. These consist of:

4. **Neural Networks:** With a vast number of connected processing nodes, neural networks mimic how the human brain functions. Natural language trans- lation, picture identification, speech recognition, and image generation are just a few of the applications that benefit from neural networks' aptitudefor pattern detection.

5. **Linear Regression:** Based on a linear connection between various variables, the linear regression process is used to forecast numerical values. The method might be applied, for instance, to forecast housing values based on local historical data.

6. **Logistic Regression:** The supervised learning process known as logistic regression uses categorical response variables, such as ”yes/no” responses to questions, to produce predictions. Applications for it include sorting spam and performing quality control on a production line.

7. **Clustering:** Data can be grouped using clustering, which uses unsupervised learning to find patterns in the data. Data scientists can benefit from com- puters' ability to spot distinctions between data points that humans have missed.

8. **Decision Trees:** Decision trees can be used to categorise data into groups as well as forecast numerical values (regression). A tree diagram can be used to show the branching sequence of connected decisions used in decision trees. In contrast to the neural network's ”black box,” decision trees are simple to validate and audit, which is one of their benefits.

9. **Random Forest:** By merging the outcomes from various decision trees, the machine learning algorithm predicts a value or category in a random forest.

The hundreds of simulations required for some heterogeneous catalysis scenarios cannot be done due to the high cost of DFT calculations. In these strange situations, physics-based atomistic potentials have been employed for a long time, but they typically lack the necessary accuracy. As a result, there has been an increase in interest in using ML to create atomistic potentials using DFT calculations [10]. The potential energy of a system of atoms is cal- culated using these Machine-learned Atomistic Potentials (MLPs), which are mathematical operations. They estimate interaction energies with improved numerical efficiency and speed while keeping high accuracy of results since theyare typically trained on data from QM Modelling techniques like DFT [11]. ML has had considerable success in a range of applications over the past few

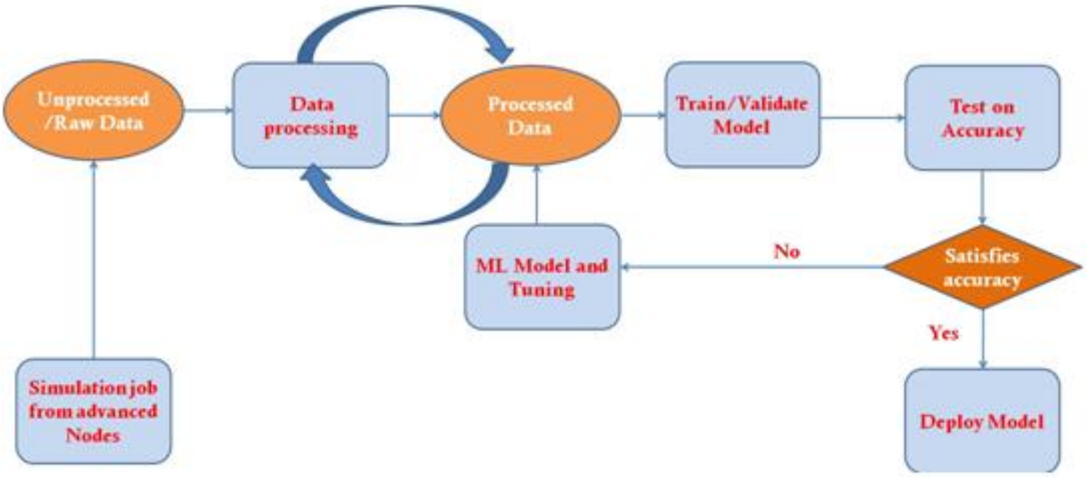

Figure 1: Flow chart representation of a Machine learning process

years, including bandgap property predictions [12], elastic moduli [13], stability analyses of crystals [14, 15], and molecular force-field estimations [16]. The figure given below shows the schematic representation of how a ML algorithm works.

III. DFT: A GUIDE TO CHEMISTS

Electronic structure computations frequently use the DFT approach because it strikes an ideal mix between acceptable accuracy and affordable computational cost. The electronic density, or n(r), is the main quantity. The Hohenberg-Kohn theorems [17] ensure that the electronic density and the external potential, such as the electron-ion potential V_{ei}(r), correspond one to one. This means that any desirable quality may be identified as a function of density. The Kohn-Sham density-functional theory (KS-DFT), the current industry stan- dard for computing electronic structure, has seen significant development in recent years [18]. Early DFT approximations relied entirely on uniform elec- tron gas models without any empirical characteristics. Incorporating density gradients and a few of these techniques resulted in significant gains. parameters from fits to atomic data, usually one or two. Even greater advancements were made possible by accurate interchange and fitting to molecular data, such as experimental heats of formation. However, because there are so many possible chemical reactions, this unlocked a Pandora’s Box of possibilities. As a result, there has been a recent surge in the number of DFT approximations that have been empirically fitted to hundreds or thousands of chemical reference data [18]. DFT has been used in a wide variety of applications over the years. This variety developed because it is possible to predict the molecule and crystal structures as well as the forces acting on the atomic nuclei when they are not in their equilibrium positions by understanding the electronic ground-state energy as a function of the locations of the atomic nuclei. DFT is currently used frequently to solve issues in atomic and molecular physics, such as the

Table 1: Some major historical developments in the field of computational chemistry.

Year	Event
1927	First theoretical calculations using valence bond theory.
1940	Revolution in computer technology.
1950	First semi-empirical calculations were performed.
1950	The first configuration interaction calculations using GTO.
1956	First HF calculations were done using STO basis set.
1959	The first polyatomic calculations using Gaussian orbital.
1960	First calculation with larger basis set and study of minimal basis set.
1964	Huckel method calculations using LCAO approach.
1970	Use of Computer programs: ATMOL, Gaussian, IBMOL, and POLYAYTOM.
1971	First bibliography of ab initio calculations.
1973	Development of molecular mechanics, such as MM2 force field.

calculation of ionisation potentials and vibration spectra, the study of chem- ical reactions, the structure of biomolecules, and the nature of active sites in catalysts, as well as issues in condensed matter physics, such as lattice struc- tures, phase transitions in solids, and liquid metals.46 DFT turns become a crucial tool while researching bigger compounds. With typical quantum chem- istry methods, the computational work required increases exponentially as the number of electrons involved increases, whereas with DFT it increases roughly as the third power of this number. In actuality, this means that although CI can only be used for systems with a few atoms, DFT may be applied to molecules with hundreds of atoms. It may be impossible to simply solve the noninteracting problem for a complex molecule, thus several techniques are employed to make the issue more manageable from a computational stand- point. similar to the popular pseudo-potential method, which avoids repeatedly recalculating the wave functions of the inactive core electrons. Systems for delivering pharmaceuticals to specific bodily parts are known as drug delivery systems. These systems frequently include delivery components consisting of biodegradable and bioabsorbable polymers. The application of computational material science to the design and development of drug delivery materials has greatly benefited by the introduction of DFT. DFT and other computer tech- niques are used to circumvent laborious empirical procedures [19]. Evolution of the methodological advancements and computer programs in the field of computational chemistry is tabulated below:

DFT is one of the most popular computational techniques for heteroge- neous catalytic activity prediction. It is well known for its distinctive capacity to accurately simulate the structure of atoms, molecules, crystals, surfaces, and their interactions [20]. However, the accuracy of the DFT calculations is dependent upon the accuracy of the exchange-correlation functional employed for the estimation of the exchange-correlation energy, which are still based on certain approximations. ML can help in improving of those functionals which in order can result in the results predicted by the DFT. The upcoming sections of the chapter discuss the various aspects of collaborating DFT with ML in various fields of research.

IV. DFT COUPLED WITH ML: THE UNBEATABLE DUO

The accuracy of the total DFT simulation is mostly determined by the precision of the selected exchange-correlation approximation. Exchange-correlation functional constructs, but fitting to either experimental or high-level computa- tions data is also a typical practise. These have been used for a very long period, particularly in quantum chemistry [21]. It follows naturally that similar fits can be performed using ML techniques[22]. Here, the electronic density is used as the input quantity for a NN that approximates the exact exchange-correlation energy for a system of finite size. Automatic differentiation is used to determine the exchange-correlation potential. Since standard exchange-correlation func- tionals typically lack nonlocal effects, the resulting model includes them with a minor processing burden compared to accurate calculations. Similar methods are used in [23], which introduces NeuralXC, a framework for building such NN exchangecorrelation functionals, and [24], which uses CNNs to estimate the exchange-correlation energy from the density by utilising convolutions of CNN architecture to model electronic density. They accurately recreate the B3LYP functional exchange. This is crucial because it shows how NNs may derive use- ful information from electronic density as the B3LYP functional accomplishes accurate exchange by evaluating the KS orbitals rather than density. Build- ing ML functionals is not the only method for enhancing exchange-correlation energies. Exchange-correlation functionals are just one type of density func- tional that can be treated by ML. The accuracy of computationally effective OF-DFT simulations largely depends on the kinetic energy functional approx- imation selected. It goes without saying that ML can help in finding a good functional, and this is a topic of ongoing research [6, 25–27]. A topical introduc- tion has been provided by Li et. al [28]. Specified restrictions are employed to confine ML functionals to improve performance on this task [27]. A slightly dif- ferent approach is taken in Reference [29] , which teaches a density functional for the entire total energy. In contrast to a direct mapping of atomic locations to total energies, it is found [30] that learning electronic density from atomic positions and then performing a secondary mapping from electronic density to total energy produces findings that are more accurate. Since the KS-DFT energy functional contains terms that are just implicit functionals of density, a different mapping is required. Similar research is done, but this time it focuses on the exchange-correlation energy rather than the total energy [31]. However, the results were contrary to those reported by Brockherde et al [30] i.e., the exchange-correlation energy is supplied more precisely for a direct mapping. These results may be justified by the fact that indirect mappings via the den- sity are more advantageous in extrapolative contexts whereas direct mappings perform better elsewhere, according to recent research [32]. However, a study based on energy is not the only option available in electronic structure theory. Use of Hamiltonians and ML is done to compute various quantities of inter- est [33, 34]. Similar to this, GPR is used to fit potentials on DFT data, which can be utilised to enhance computations at lower levels of theory (such density functional tight-binding) [35]. Last but not least, ML can be utilised to speed up the numerical treatment of DFT by lowering the computational burden of solving the KS equations [36] or by locating effective, adaptive basis sets [37]. In the upcoming subsections the applications of combined DFT-ML methodsis discussed in various fields.

1. **DFT-ML Approaches for Lattice Parameters Calculations:** Modern materials research relies on computational high-throughput investi- gations to find novel materials. These investigations are mostly carried out inside the KS-DFT) in solid-state physics [38]. Although the many-body Schrödinger equation is precisely described by DFT, it actually depends on approximations for the exchange-correlation energy. To describe solids' elastic and electrical properties, one must be familiar with their crystal structure. To specifically predict the electrical characteristics of materials that have not yet been synthesised, their precise prediction is crucial. The PBE approximation and its version PBEsol are most frequently used as exchange-correlation functionals in DFT calculations of lattice parameters. They are effective at describing the properties of materials, although they don't always match experiments' accuracy levels. We suggest a crystal structure optimisation method that is computationally efficient and based on interpretable ML. It is showed that, as a result, PBE- and PBEsolstructure accuracy can be consid- erably improved. To describe solids' elastic and electrical properties, one must be familiar with their crystal structure. To specifically predict the electrical characteristics of materials that have not yet been synthesised, their precise prediction is crucial. The PBE approximation and its version PBEsol are most frequently used as exchange-correlation functionals in DFT calculations of lattice parameters. They are effective at describing the properties of materials, although they don't always match experiments' accuracy levels. We suggest a crystal structure optimisation method that is computationally efficient and based on interpretable ML. It is found that as a result, PBE- and PBEsol- structure accuracy can be considerably improved. We examine how well these functionals predict lattice parameters and demonstrate how to improve their accuracy using ML. The Inorganic Crystal Structure Database's experimental crystal structures that have been matched with PBE-optimized structures kept in the materials project database make up our data set. PBEsol com- putations were addes to these data as a complement. We show that using straightforward, comprehensible ML models can significantly increase the a posteriori accuracy and precision of PBE/PBEsol volume estimates. These models can increase PBE unit cell volumes to equal PBEsol calculations in accuracy and decrease the latter's error relative to experiment by 35%. The implicit correction of finite temperature effects without phonon computations is another advantage of our method [15].

   The accuracy of the quick DU8+ hybrid density functional theory/parametric calculations of nuclear magnetic resonance spectra is significantly improved by ML, enabling high-throughput in silico validation and revision of com- plex alkaloids and other natural products. 35 structures of the almost 170 alkaloids studied are altered using the DU8ML approach, the next-generation ML-augmented DU8 method [39].

It is essential to create new quaternary semiconductor materials with supe-rior qualities in order to speed up the application of quaternary opto-electronic materials in the field of luminescence. However, standard trial-and-error tech- niques are sometimes time-consuming and ineffective when dealing with a wide range of different quaternary semiconductors. The band-gaps of 2180 quaternary semiconductors, the majority of which were underdeveloped but environmentally friendly, were predicted here using a combination of ML and DFT calculations. The model using the random forest technique had an evalu- ation coefficient (R2) of up to 0.93 in machine learning. Ag_2InGaS_4, $AgZn_2In_4$, Ag_2ZnSnS_4, and $AgZn_2GaS_4$ are four new quaternary semiconductors with direct band-gaps that were chosen from the ML model. The four quaternary semiconductors were then further examined and their electronic structures were confirmed by DFT calculations, which showed that they had direct band- gaps, a tiny effective mass, a large exciton binding energy, and Stokes shift. Our computation has a definite reference value for the research of luminous materials and devices and could greatly speed up the discovery of innovative opto-electronic semiconductors [40].

2. **DFT-ML Approaches in Thermoelectric Materials:** Thermoelectrics are employed in a variety of specialised technologies, including wine coolers, hiking stoves with mobile phone chargers, and radioisotope ther- moelectric (TE) generators that power things like the Curiosity Mars rover. Through waste heat recovery, thermoelectrics could also help to lower global greenhouse gas emissions, but their current contribution is constrained by the meagre efficacy of devices [41]. Another drawback is the presence of rare or hazardous elements in some cutting-edge TE materials [42]. Thus, the search for novel TE materials has attracted considerable scientific attention lately [43].Recent years have seen a significant increase in the use of high-throughput screening based on first-principle calculations in the hunt for novel TE materi- als [44, 45]. Many investigations employ straightforward models or estimations of lattice thermal conductivity (K_l) and concentrate on electrical characteris- tics. This is due, in part, to the high computational cost of computing K_l. The expense results from the need to get third-order force constants from several supercell-based DFT computations in order to account for the phonon-phonon interactions resulting from the anharmonicity of the lattice vibrations [46]. For predicting K_l, ML techniques are increasingly used in addition to first- principles calculations [47]. Additionally, pre-trained ML models may be made accessible through practical web-based apps [48].

3. **DFT and ML in Heterogenous Catalysis:** Heterogeneous work in catalysis is being influenced by ML. Nowadays, the search for optimum catalysts in huge combinatory spaces, such as bimetals, is accelerated by combining ML with first-principle calculation methods. The identification of catalyst output characteristics in huge data sets uses inter- atomic capacity gained from ML for accurate and quick catalyst modelling. The trend of increasing global energy demands and the desire to protect the environment have made it necessary to look for **alternatives**

to fuels based on petroleum. In order to produce hydrogen gas (H_2) as a primary product or as a product in combination with other gases, such as carbon monoxide (CO), as in the case of syngas, research has been conducted into better ways of reforming various compounds that contain hydrogen. An alternate renewable energy source is hydrogen. Nevertheless, for the onboard use of hydrogen, hydrogen storage is still a key research field [49]. Up until around 20 years ago, the majority of research on catalytic reforming was experimental [50]. Since then, computational tools, especially DFT, have been applied to the investi- gation of the electronic structure of molecules, materials, and processes. DFT calculations provide information about heterogeneous catalyst systems that is challenging to get through experimental means. The calculations describe the reactivity patterns and the characteristics of the transition states of molecules reacting at solid surfaces. In a real high-pressure, high-temperature process, a catalyst's functioning condition is characterised using DFT calculations. DFT calculations shed light on the kinetics, activity, and reaction mechanism of catalysts in catalytic reactions. The hydrogenation of furfural occurs through either a hydroxyalkyl or an alkoxide intermediate, according to the results of DFT computations for the reaction path on a Cu/SiO2 catalyst at 230-290

C [51]. The predicted energy barriers are of the same order as in experiments. The accuracy of DFT computing in predicting experimental outcomes is very good. A comparison of DFT results and actual results in the semi-hydrogenation of acetylene (SHA) in an ethyleneconcentrated stream reveals that the PtCu bimetallic catalyst imitates the DFT projections in regard to the electronic structures. This was further supported by the catalytic activity and X-ray photoelectron spectroscopy [52].

To enable real-time manufacturing in accordance with feasibility studies, modelling the production in any reformation processes is necessary. Computa- tional and artificial intelligence have been used in studies of energy engineering for a while [53]. Dataset patterns that are hidden from view can be studied using ML, and these patterns can then be used to model an objective variable [54]. Computer-based method research demonstrates that computational intel- ligence approaches can make use of a sizable amount of data that is computed using DFT to achieve a high accuracy level [54].

Based on the photocatalytic degradation of lignin model compounds, a collection of lignin-model-compound-photocatalysis conditions and outcomes was created. Using input variables that describe the lignin model compound, the cleavage dissociation energy and molecular dipole moment, a traditional ML model was able to accurately predict selectivity during the photocatalytic cleavage of lignin bonds. We initially predicted the photocatalytic breakage of the C-C bonds in lignin using the K-nearest neighbour (K-NN), nave Bayes (NB), support vector machine (SVM), logistic regression (LR), and random forest (RF) algorithms, with accuracy and precision assessed. The classifica- tion prediction performance of the RF model was strong, with a prediction accuracy of 0.99. Additionally, by integrating DFT and ML, molecular yields were successfully predicted by taking

C-C bond cleavage during photocatalytic lignin degradation into account. The total feature relevance was found to be 39.22%, with the reaction circumstances being found to be more significant than the characteristics of the reactants during the photocatalytic breakage of lignin C-C linkages. The risk of a C-C bond breaking is reduced when the num- ber of methoxy groups attached to benzene rings in lignin increases, increasing the bond dissociation energy. Additionally, it was discovered that the bandgap width, followed by specific surface area, which is more significant than pore volume and pore size, is the most crucial catalyst property [55]. Using the artificial neural network (ANN) and the AdaBoost (AB) algorithms to simulate the syngas composition, Adeniyi et al. examined the steam reforming of biomass bio-oil. He came to the conclusion that the two algorithms produced R2¿0.999 results, and the product selectivity result showed significant data variability capture. Overall, for the system under consideration, the ANN pre- dictions outperformed the AB predictions in terms of accuracy [56]. In their analysis of the CH_4 reaction process over a Ni/TiO_2 catalyst using DFT, Yang et al. noted good carbon tolerance. He focused on the effects of Ni/TiO_2 as a catalyst on the generation of hydrogen and carbon monoxide while also inves- tigating six potential paths for methane reforming processes. The observed results imply that the predominant pathway for methane (CH_4) reformation is the reaction between carbon and the lattice oxygen to produce CO [57].

4. **Correction of Adsorption Energies:** For the best and most accurate findings, ML is sometimes employed to con- firm DFT computed data[58]. This method is used by Okamoto to determine the PtRu bimetal’s ideal composition in order to lower the energy required for carbon monoxide adsorption. He first determines the adsorption energies. on variously configured PtRu (111) bimetallic slabs, followed by the application of multiple regression analysis for the data mining-ML model. The estimated DFT data for the identical alloy composition validates that the carbon monox- ide adsorption energies on PtRu [59] were correctly predicted. For the purpose of correcting system-entrained Alchemical perturbation density functional theory (AP-DFT) prediction errors for diverse carbon and oxygen-based adsor- bates, three independent support vector regression machine learning models were trained using ML [60]. Toyao et al. devised a number of ML techniques to determine the binding energies of methane-related molecules on metal alloys [61]. Methane activation on metal-organic framework catalysts must be guided by structural principles, and Rose et al.’s study of the structure-activity rela- tionship using DFT simulations supports this. The investigation was successful in demonstrating the inverse correlation between the metal-oxi site’s forma- tion energy and the activation energy of methane-x [62]. The problem of CO_2 emissions has led to the development of carbon capture and storage (CCS), especially bio-route CCS, which uses activated carbon derived from biomass to trap CO_2. N-doped BAC adsorbents’ well-known multi-nitrogen functional groups and microstructure characteristics can work together to favour CO_2 physisorption. The numerous physicochemical characteristics of N-doped BAC were subjected to ML modelling in this case as a challenge to identify the as-yet-unidentified mechanism of CO_2 capture. To calculate the in operando effectiveness of microstructural and N-functionality groups at six situations of pressures ranging from 0.15 to 1 bar at

ambient temperature and cryogenic temperatures, a radial basis function neural network (RBF-NN) was used. A number of different training algorithms were used, and trainbr showed the low- est mean absolute percent error (MAPE), which was 3.5%. RBF-NN calculates the CO_2 capture of BACs as solid adsorbents with an R2 range of 0.97-0.99. Additionally, the generalisation evaluation of RBF-NN noticed errors, which tolerates 0.5–6% MAPE in 50–80% of all data sets. An alternate survey sen- sitivity analysis reveals the significance of several features, including nitrogen content (N%), oxidized-N, and graphitic-N as nitrogen functional groups, spe- cific surface area (SSA), micropore volume (%Vmic), average pore diameter (AVD), and nitrogen content (N%). The physiochemical characteristics of N- doped ACs were improved using a genetic algorithm (GA). It suggested a value of 9.2 mmol g^{-1} at 1 pressure and 273 K as the ideal CO_2 capture. The exemplified BACs' shapes and adsorption energies with CO2 molecules were optimised using a combination of the GA and DFT [63].

V. POSSIBILITIES AND CONSTRAINTS

Building intricate workflows integrating DFT and ML presents a number of technological challenges that must be solved. Any data-driven strategy is, first and foremost, wholly dependent on data. With regard to ab initio simulations, Data can be exchanged digitally and is expensive to obtain. In light of this, it makes logical sense to reuse precalculated data and models in databases. For instance, JARVIS contains ready-to-use ML models trained on DFT data in addition to characteristics for almost 40 000 different types of materials. The database is continuously growing, which makes it easier to create new workflows. Similarly, connected workflows make it easier to acquire additional training data [62]. The decisions made in relation to the ML techniques are another crucial factor. DFT computations and data can be used with a wide variety of ML approaches. Similar to how there are numerous methods for encoding data in terms of descriptors. It's imperative to do research [64–66] to determine which of these methods is superior to others. According to Ref. [66], a shallow DNN outperformed all other NN types in terms of predicting molecular characteristics (CNN, DNN, single-layer NN). Similar comparisons of several compositional information encoding schemes for a GB model were made in Ref. [64] . None of the methodologies under investigation could pro- vide models that could accurately predict the stability of novel materials. In fact, research is still being done to generate novel descriptors and methods to express chemical information for specialised use cases [67–72]. In the quest for functional groups that may capture CO_2, for instance, Ref. [67] introduces per- sistent picture descriptors based on ideas from applied mathematics to depict chemical structures. In order to compare these descriptors to other widely used descriptors, various ML models were trained on data encoded by these descrip- tors. SOAP [73], the Coulomb matrix [74], BoB [75], FCHL [76], and ACE [77, 78] are common descriptors. This list is not exhaustive, as many ML procedures represent materials with customised descriptors or just readily available chemical data. These publications provide examples of the usefulness of these descriptors. References [74–76] learn and forecast atomization energies using their respective descriptors and KRR. In contrast, Ref. [73] builds and com- pares GAPs using SOAP descriptors, and Ref. [77] builds IAPs based on ACE. To deploy integrated

ML-DFT workflows in future applications, transferabil- ity and uncertainty quantification must be taken into consideration. Correctly handling such factors lessens the requirement for model retraining. The trans- ferability of ML models can be used, as Ref. [79] recommends, however there are some restrictions, as shown for modest volume changes in the liquid phase. In Ref. [80], the topic of uncertainty quantification, it is discussed how to determine whether a prediction can be believed or whether more information is required to strengthen the model.

VI. CONCLUSION AND FUTURE PERSPECTIVE

As can be seen, there is a rapid increase in research interest in ML-DFT techniques, and techniques that directly tackle the electrical structure problem are becoming more popular. The number of publications in this field of study is increasing, therefore further growth is anticipated. Future research may focus on issues like active learning strategies and uncertainty quantification. By doing so, they might increase the usefulness of machine learning in computa- tional chemistry and material science. New applications, such as large-scale automated materials discovery, multi-scale modelling of materials, and digital twins of complex systems, will become feasible with ever-improving ML-DFT models. Machine learning and the artificial intelligence is the future and the future can be great if it holds the hand of the present i.e. DFT firmly.

VII. ACKNOWLEDGMENTS

Acknowledgments are not compulsory. Where included they should be brief.
Grant or contribution numbers may be acknowledged.
Please refer to Journal-level guidance for any specific requirements.

REFERENCES

[1] Primas, H.: Chemistry, Quantum Mechanics and Reductionism: Perspec- tives in Theoretical Chemistry vol. 24. Springer, ??? (2013)

[2] Ma, X.: Development of computational chemistry and application of com- putational methods. In: Journal of Physics: Conference Series, vol. 2386, p. 012005 (2022). IOP Publishing

[3] Mardirossian, N., Head-Gordon, M.: Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. Molecular physics **115**(19), 2315–2372 (2017)

[4] Goldsmith, B.R., Esterhuizen, J., Liu, J.-X., Bartel, C.J., Sutton, C.: Machine learning for heterogeneous catalyst design and discovery (2018)

[5] Erdem Günay, M., Yıldırım, R.: Recent advances in knowledge discov- ery for heterogeneous catalysis using machine learning. Catalysis Reviews **63**(1), 120–164 (2021)

[6] Fiedler, L., Shah, K., Bussmann, M., Cangi, A.: Deep dive into machine learning density functional theory for materials science and chemistry. Physical Review Materials **6**(4), 040301 (2022)

[7] Mahesh, B.: Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet] **9**(1), 381–386 (2020)

[8] Bonaccorso, G.: Machine Learning Algorithms. Packt Publishing Ltd, ???(2017)

[9] Ray, S.: A quick review of machine learning algorithms. In: 2019 Inter- national Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), pp. 35–39 (2019). IEEE

[10] Behler, J.: Constructing high-dimensional neural network potentials: a tutorial review. International Journal of Quantum Chemistry **115**(16), 1032–1050 (2015)

[11] Botu, V., Batra, R., Chapman, J., Ramprasad, R.: Machine learning force fields: construction, validation, and outlook. The Journal of Physical Chemistry C **121**(1), 511–522 (2017)

[12] Patra, A., Batra, R., Chandrasekaran, A., Kim, C., Huan, T.D., Ram-prasad, R.: A multi-fidelity information-fusion approach to machine learn and predict polymer bandgap. Computational Materials Science **172**, 109286 (2020)

[13] De Jong, M., Chen, W., Notestine, R., Persson, K., Ceder, G., Jain, A., Asta, M., Gamst, A.: A statistical learning framework for materials science: application to elastic moduli of k-nary inorganic polycrystalline compounds. Scientific reports **6**(1), 34256 (2016)

[14] Ye, W., Chen, C., Wang, Z., Chu, I.-H., Ong, S.P.: Deep neural networks for accurate predictions of garnet stability. arXiv preprint arXiv:1712.01908 (2017)

[15] Hussein, R., Schmidt, J., Barros, T., Marques, M.A., Botti, S.: Machine- learning correction to density-functional crystal structure optimization.MRS Bulletin **47**(8), 765–771 (2022)

[16] Pathrudkar, S., Yu, H.M., Ghosh, S., Banerjee, A.S.: Machine learn-ing based prediction of the electronic structure of quasi-one-dimensional materials under strain. Physical Review B **105**(19), 195141 (2022)

[17] Hohenberg, P., Kohn, W.: Inhomogeneous electron gas. Physical review **136**(3B), 864 (1964)

[18] Becke, A.D.: Density-functional theory vs density-functional fits. The Journal of Chemical Physics **156**(21) (2022)

[19] Adekoya, O.C., Adekoya, G.J., Sadiku, E.R., Hamam, Y., Ray, S.S.: Application of dft calculations in designing polymer-based drug delivery systems: An overview. Pharmaceutics **14**(9), 1972 (2022)

[20] Argaman, N., Makov, G.: Density functional theory: An introduction. American Journal of Physics **68**(1), 69–79 (2000)

[21] Becke, A.D.: Density-functional exchange-energy approximation with correct asymptotic behavior. Physical review A **38**(6), 3098 (1988)

[22] Schmidt, J., Benavides-Riveros, C.L., Marques, M.A.: Machine learn- ing the physical nonlocal exchange–correlation functional of density- functional theory. The journal of physical chemistry letters **10**(20), 6425–6431 (2019)

[23] Dick, S., Fernandez-Serra, M.: Machine learning accurate exchange and correlation functionals of the electronic density. Nature communications **11**(1), 3509 (2020)

[24] Lei, X., Medford, A.J.: Design and analysis of machine learning exchange- correlation functionals via rotationally invariant convolutional descrip- tors. Physical Review Materials **3**(6), 063801 (2019)

[25] Meyer, R., Weichselbaum, M., Hauser, A.W.: Machine learning approaches toward orbital-free density functional theory: Simultaneous training on the kinetic energy density functional and its functional deriva- tive. Journal of chemical theory and computation **16**(9), 5685–5694 (2020)

[26] Snyder, J.C., Rupp, M., Hansen, K., Müller, K.-R., Burke, K.: Find- ing density functionals with machine learning. Physical review letters **108**(25), 253002 (2012)

[27] Hollingsworth, J., Li, L., Baker, T.E., Burke, K.: Can exact conditions improve machine-learned density functionals? The Journal of chemical physics **148**(24) (2018)

[28] Li, L., Snyder, J.C., Pelaschier, I.M., Huang, J., Niranjan, U.-N., Dun-can, P., Rupp, M., Müller, K.-R., Burke, K.: Understanding machine- learned density functionals. International Journal of Quantum Chemistry **116**(11), 819–833 (2016)

[29] Li, L., Baker, T.E., White, S.R., Burke, K., *et al.*: Pure density func-tional for strong correlation and the thermodynamic limit from machine learning. Physical Review B **94**(24), 245129 (2016)

[30] Brockherde, F., Vogt, L., Li, L., Tuckerman, M.E., Burke, K., Müller, K.- R.: Bypassing the kohn-sham equations with machine learning. Nature communications **8**(1), 872 (2017)

[31] Grisafi, A., Fabrizio, A., Meyer, B., Wilkins, D.M., Corminboeuf, C., Ceri-otti, M.: Transferable machine-learning model of the electron density. ACScentral science **5**(1), 57–64 (2018)

[32] Lewis, A.M., Grisafi, A., Ceriotti, M., Rossi, M.: Learning electron densities in the condensed phase. Journal of Chemical Theory andComputation **17**(11), 7203–7214 (2021)

[33] Hegde, G., Bowen, R.C.: Machine-learned approximations to density functional theory hamiltonians. Scientific reports **7**(1), 42669 (2017)

[34] Ferreira, A.R.: Chemical bonding in metallic glasses from machine learn- ing and crystal orbital hamilton population. Physical Review Materials**4**(11), 113603 (2020)

[35] Panosetti, C., Engelmann, A., Nemec, L., Reuter, K., Margraf, J.T.: Learning to use the force: Fitting repulsive potentials in density-functional tight-binding with gaussian process regression.

Journal of chemical theoryand computation **16**(4), 2181–2191 (2020)

[36] Ku, J., Kamath, A., Carrington Jr, T., Manzhos, S.: Machine learn- ing optimization of the collocation point set for solving the kohn–sham equation. The Journal of Physical Chemistry A **123**(49), 10631–10642(2019)

[37] Schutt, O., VandeVondele, J.: Machine learning adaptive basis sets for effi- cient large scale density functional theory simulation. Journal of chemical theory and computation **14**(8), 4168–4175 (2018)

[38] Lee, C., Yang, W., Parr, R.G.: Development of the colle-salvetti correlation-energy formula into a functional of the electron density. Physical review B **37**(2), 785 (1988)

[39] Novitskiy, I.M., Kutateladze, A.G.: Du8ml: machine learning-augmented density functional theory nuclear magnetic resonance computations for high-throughput in silico solution structure validation and revision of complex alkaloids. The Journal of Organic Chemistry **87**(7), 4818–4828 (2022)

[40] Gao, M., Cai, B., Liu, G., Xu, L., Zhang, S., Zeng, H.: Machine learning and density functional theory simulation of the electronic structural prop- erties for novel quaternary semiconductors. Physical Chemistry Chemical Physics **25**(13), 9123–9130 (2023)

[41] Champier, D.: Thermoelectric generators: A review of applications. Energy Conversion and Management **140**, 167–181 (2017)

[42] Gutiérrez Moreno, J.J., Cao, J., Fronzi, M., Assadi, M.H.N.: A review of recent progress in thermoelectric materials through computational methods. Materials for Renewable and Sustainable Energy **9**, 1–22 (2020)

[43] Recatala-Gomez, J., Suwardi, A., Nandhakumar, I., Abutaha, A., Hip- palgaonkar, K.: Toward accelerated thermoelectric materials and process discovery. ACS Applied Energy Materials **3**(3), 2240–2257 (2020)

[44] Berland, K., Shulumba, N., Hellman, O., Persson, C., Løvvik, O.M.: Ther- moelectric transport trends in group 4 half-heusler alloys. Journal of Applied Physics **126**(14) (2019)

[45] Choudhary, K., Garrity, K.F., Tavazza, F.: Data-driven discovery of 3d and 2d thermoelectric materials. Journal of Physics: Condensed Matter **32**(47), 475501 (2020)

[46] Zhou, F., Nielson, W., Xia, Y., Ozoliņš, V., *et al.*: Lattice anharmonic- ity and thermal conductivity from compressive sensing of first-principles calculations. Physical review letters **113**(18), 185501 (2014)

[47] Wang, X., Zeng, S., Wang, Z., Ni, J.: Identification of crystalline materials with ultra-low thermal conductivity based on machine learning study. The Journal of Physical Chemistry C **124**(16), 8488–8495 (2020)

[48] Gaultois, M.W., Oliynyk, A.O., Mar, A., Sparks, T.D., Mulholland, G.J., Meredig, B.: A recommendation engine for suggesting unexpected thermoelectric chemistries. arXiv preprint arXiv:1502.07635 (2015)

[49] Li, X., Lim, K.H.: Dft study of steam reforming of formaldehyde on cu, pdzn, and ir. ChemCatChem **4**(9), 1311–1320 (2012)

[50] Schlögl, R.: Heterogeneous catalysis. Angewandte Chemie International Edition **54**(11), 3465–3520 (2015)

[51] Sitthisa, S., Sooknoi, T., Ma, Y., Balbuena, P.B., Resasco, D.E.: Kinetics and mechanism of hydrogenation of furfural on cu/sio2 catalysts. Journal of catalysis **277**(1), 1–13 (2011)

[52] Ayodele, O.B., Cai, R., Wang, J., Ziouani, Y., Liang, Z., Spadaro, M.C., Kovnir, K., Arbiol, J., Akola, J., Palmer, R.E., *et al.*: Synergistic computational–experimental discovery of highly selective ptcu nanoclus- ter catalysts for acetylene semihydrogenation. ACS Catalysis **10**(1), 451–457 (2019)

[53] Faizollahzadeh Ardabili, S., Najafi, B., Shamshirband, S., Minaei Bid- goli, B., Deo, R.C., Chau, K.-w.: Computational intelligence approach for modeling hydrogen production: A review. Engineering Applications of Computational Fluid Mechanics **12**(1), 438–458 (2018)

[54] O'Leary, D.E.: Artificial intelligence and big data. IEEE intelligent systems **28**(2), 96–99 (2013)

[55] Zhang, T., Wu, C., Xing, Z., Zhang, J., Wang, S., Feng, X., Zhu, J., Lu,X., Mu, L.: Machine learning prediction of photocatalytic lignin cleav- age of c–c bonds based on density functional theory. Materials Today Sustainability **20**, 100256 (2022)

[56] Adeniyi, A.G., Ighalo, J.O., Marques, G.: Utilisation of machine learning algorithms for the prediction of syngas composition from biomass bio- oil steam reforming. International Journal of Sustainable Energy **40**(4), 310–325 (2021)

[57] Yang, W., Wang, Z., Tan, W., Peng, R., Wu, X., Lu, Y.: First principles study on methane

reforming over ni/tio2 (110) surface in solid oxide fuel cells under dry and wet atmospheres. SCIENCE CHINA-MATERIALS **63**(3), 364–374 (2020)

[58] Schleder, G.R., Padilha, A.C., Acosta, C.M., Costa, M., Fazzio, A.: Fromdft to machine learning: recent approaches to materials science–a review. Journal of Physics: Materials **2**(3), 032001 (2019)

[59] Okamoto, Y.: Finding optimum compositions of catalysts using ab initio calculations and data mining. Chemical physics letters **395**(4-6), 279–284 (2004)

[60] Griego, C.D., Zhao, L., Saravanan, K., Keith, J.A.: Machine learning corrected alchemical perturbation density functional theory for catalysis applications. AIChE Journal **66**(12), 17041 (2020)

[61] Toyao, T., Suzuki, K., Kikuchi, S., Takakusagi, S., Shimizu, K.-i., Taki- gawa, I.: Toward effective utilization of methane: machine learning prediction of adsorption energies on metal alloys. The Journal of Physical Chemistry C **122**(15), 8315–8326 (2018)

[62] Nandy, A., Zhu, J., Janet, J.P., Duan, C., Getman, R.B., Kulik, H.J.: Machine learning accelerates the discovery of design rules and exceptions in stable metal–oxo intermediate formation. Acs Catalysis **9**(9), 8243– 8255 (2019)

[63] Rahimi, M., Abbaspour-Fard, M.H., Rohani, A., Yuksel Orhan, O., Li, X.: Modeling and optimizing n/o-enriched bio-derived adsorbents for co2 capture: machine learning and dft calculation approaches. Industrial & Engineering Chemistry Research **61**(30), 10670–10688 (2022)

[64] Bartel, C.J., Trewartha, A., Wang, Q., Dunn, A., Jain, A., Ceder, G.: A critical examination of compound stability predictions from machine- learned formation energies. npj computational materials **6**(1), 97 (2020)

[65] Agarwal, S., Mehta, S., Joshi, K.: Understanding the ml black box with simple descriptors to predict cluster–adsorbate interaction energy. New Journal of Chemistry **44**(20), 8545–8553 (2020)

[66] Hou, F., Wu, Z., Hu, Z., Xiao, Z., Wang, L., Zhang, X., Li, G.: Comparison study on the prediction of multiple molecular properties by various neu- ral networks. The Journal of Physical Chemistry A **122**(46), 9128–9134 (2018)

[67] Townsend, J., Micucci, C.P., Hymel, J.H., Maroulas, V., Vogiatzis, K.D.: Representation of molecular structures with persistent homology for machine learning applications in chemistry. Nature communications **11**(1), 3230 (2020)

[68] Deimel, M., Reuter, K., Andersen, M.: Active site representation in first- principles microkinetic models: data-enhanced computational screening for improved methanation catalysts. Acs Catalysis **10**(22), 13729–13736 (2020)

[69] Laghuvarapu, S., Pathak, Y., Priyakumar, U.D.: Band nn: A deep learn- ing framework for energy prediction and geometry optimization of organic small molecules. Journal of computational chemistry **41**(8), 790–799 (2020)

[70] Dickel, D., Francis, D., Barrett, C.: Neural network aided development of a semi-empirical interatomic potential for titanium. Computational Materials Science **171**, 109157 (2020)

[71] Gusarov, S., Stoyanov, S.R., Siahrostami, S.: Development of fukui func-tion based descriptors for a machine learning study of co2 reduction. The Journal of Physical Chemistry C **124**(18), 10079–10084 (2020)

[72] Collins, E.M., Raghavachari, K.: Effective molecular descriptors for chem- ical accuracy at dft cost: Fragmentation, error-cancellation, and machine learning. Journal of Chemical Theory and Computation **16**(8), 4938–4950 (2020)

[73] Bartȏk, A.P., Kondor, R., Csȃnyi, G.: On representing chemical environ- ments. Physical Review B **87**(18), 184115 (2013)

[74] Rupp, M., Tkatchenko, A., Müller, K.-R., Von Lilienfeld, O.A.: Fast and accurate modeling of molecular atomization energies with machine learning. Physical review letters **108**(5), 058301 (2012)

[75] Hansen, K., Biegler, F., Ramakrishnan, R., Pronobis, W., Von Lilien-feld, O.A., Muller, K.-R., Tkatchenko, A.: Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. The journal of physical chemistry letters **6**(12), 2326–2331(2015)

[76] Faber, F.A., Christensen, A.S., Huang, B., Von Lilienfeld, O.A.: Alchemi- cal and structural distribution based representation for universal quantum machine learning. The Journal of chemical physics **148**(24) (2018)

[77] Drautz, R.: Erratum: Atomic cluster expansion for accurate and trans- ferable interatomic potentials [phys. rev. b 99, 014104 (2019)]. Physical Review B **100**(24), 249901 (2019)

[78] Lysogorskiy, Y., Oord, C.v.d., Bochkarev, A., Menon, S., Rinaldi, M., Hammerschmidt, T., Mrovec, M., Thompson, A., Csânyi, G., Ortner, C., *et al.*: Performant implementation of the atomic cluster expansion (pace) and application to copper and silicon. npj computational materials **7**(1),97 (2021)

[79] Tamura, R., Lin, J., Miyazaki, T.: Machine learning forces trained by gaussian process in liquid states: transferability to temperature and pressure. Journal of the Physical Society of Japan **88**(4), 044601 (2019)

[80] Peterson, A.A., Christensen, R., Khorshidi, A.: Addressing uncertainty in atomistic machine learning. Physical Chemistry Chemical Physics **19**(18), 10978–10985 (2017)