Perspective

Deep Learning for Molecular Thermodynamics

Hassaan Malik 1, Muhammad Umar Chaudhry 2 and Michal Jasinski 3,4,*

1 Department of Computer Science, University of Management and Technology, Lahore 54000, Pakistan
2 Department of Computer Science, MNS-University of Agriculture, Multan 60000, Pakistan
3 Department of Electrical Engineering Fundamentals, Faculty of Electrical Engineering, Wroclaw University of Science and Technology, 50-370 Wroclaw, Poland
4 Department of Electrical Power Engineering, Faculty of Electrical Engineering and Computer Science, VSB-Technical University of Ostrava, 708-00 Ostrava, Czech Republic
* Correspondence: michal.jasinski@pwr.edu.pl

Abstract: The methods used in chemical engineering are strongly reliant on having a solid grasp of the thermodynamic features of complex systems. It is difficult to define the behavior of ions and molecules in complex systems and to make reliable predictions about the thermodynamic features of complex systems across a wide range. Deep learning (DL), which can provide explanations for intricate interactions that are beyond the scope of traditional mathematical functions, would appear to be an effective solution to this problem. In this brief Perspective, we provide an overview of DL and review several of its possible applications within the realm of chemical engineering. DL approaches to anticipate the molecular thermodynamic characteristics of a broad range of systems based on the data that are already available are also described, with numerous cases serving as illustrations.

Keywords: deep learning; molecular thermodynamics; thermodynamic properties; artificial intelligence; forecasting; thermodynamics

1. Introduction

With the assistance of DL, one is now able to carry out research and make predictions relating to the molecular thermodynamics of intricate systems. The study of molecular thermodynamics comes with several responsibilities, one of which is the challenging duty of precisely predicting the thermodynamic properties of intricate systems. The thermodynamic properties of several systems have been accurately predicted using standard or semi-theoretical models [1]. It remains challenging to develop models that, when used within a standard theoretical framework, are capable of accurately predicting the thermodynamic properties of complex systems at the molecular level [2]. To that end, DL can facilitate the creation of such models. “Deep learning” describes how computers can “learn” from their own past mistakes [3]. The recent rise in popularity of DL methods can be linked to their victorious performance in competitions against human specialists in chess and autonomous car driving. Drug design [4,5], structure exploration [6,7], and molecular thermodynamics prediction [8] are just some examples of the material and chemical engineering domains where DL models have become increasingly popular due to their ability to accurately capture the complex interaction that occurs between variables. Recent research [8–10] suggests that DL may have at least three unique contributions to make to the field of molecular thermodynamics. (1) To begin, we make predictions about the thermodynamic parameters of a large group of different systems by applying DL approaches to the data that are currently accessible. (2) Another benefit of combining DL with molecular simulations is that it can significantly reduce the amount of time required to find new materials. A direct link between all-atom molecular dynamics simulations
and quantum physics is the third thing that we accomplish, and we do this by developing a DL force field. These three aspects are presented in Figure 1.

1. Predicting thermodynamics properties

![Diagram showing Deep Learning connected to Predicting thermodynamics properties and Designing Materials connected to Molecular Thermodynamics]

2. Designing Materials

3. Developing force field

**Figure 1.** Three aspects of DL in the study of molecular thermodynamics. (1) the development of many-body force fields for the purpose of simulating complex systems; (2) the estimation of thermodynamic parameters using molecular characterization; (3) the construction of materials using large-scale molecular simulations as an integral part of the process.

The enthusiasm that is currently surrounding DL has also taken hold of the research that is being conducted on molecular thermodynamics, and, as a direct result, we are experiencing significant growth in the number of activities and interests in this sector. This has been driven by the astonishing progress that has been made in the creation of DL algorithms and their availability (open access), and it has been accentuated by the extraordinary growth of financing options. Both factors have contributed to this development, and these aspects have played a role in the evolution of this phenomenon.

Given this background, several significant questions arise, including the following:

1. Why is everyone getting so worked up over this? Which DL methods apply to the study of molecular thermodynamics or have the potential to become applicable?
2. In the field of molecular thermodynamics research, what is new about the application of DL? Is DL going to revolutionize this industry?
3. Considering the recent excitement around DL, which perspective should be taken by academics working in molecular thermodynamics? “Keep calm and carry on!” or rather, “Hooray, and up she rises!”

Although our goal is not to provide an exhaustive analysis of the topic, we intend to contribute to the ongoing discussion around these issues.

DL refers to the practice of teaching machines how to solve problems by exposing them to data and allowing them to progressively improve themselves through the accumulation of experience. In this context, we are interested in the application of DL to research on molecular thermodynamics. The most notable application of DL is the prediction of molecular thermodynamics.

This perspective study is divided as follows: Section 2 presents the database, descriptors, and algorithms used for the prediction of molecular thermodynamics. In Section 3, DL in thermodynamic properties, force fields, molecular simulation, etc., is discussed. Section 4 presents the conclusion of the study.
2. Datasets’ Descriptions, Features’ Descriptors, and DL Algorithms

Predicting a system’s output using algorithms that have been trained on a collection of databases based on specific characteristics is what DL is all about (descriptors). Predicting the system’s output based on the given characteristics requires: (a) creating a database of relevant samples; (b) applying the defined characteristics to the samples; (c) deciding on an appropriate descriptor set and method; (d) teaching the algorithm to make sense of the data in the database; (e) making use of the trained model. Figure 2 is a visual representation of a typical procedure for the creation of DL models. Without data, labels, and computation, DL models cannot perform their tasks effectively [7–9]. There have been several in-depth studies [8–12] performed to better understand these three components, and research on them is progressing rapidly. Only a brief overview of them will be given here.

![Figure 2. Process of DL model.](image)

2.1. Datasets’ Descriptions

To train and test a DL model, a database is needed [13]. The experiment and computational findings can be stored in a database. Multiple papers [8–11] have been written about thermodynamic data archives. The contributions produced by these datasets are indeed very helpful in the creation of DL models. However, since DL has progressed at a dizzying pace, they are unable to meet the ever-increasing demand. Typically, researchers will need to create their databases to meet their individual needs. When creating a database, it is important to think about both the quantity and quality of the information included in it [14]. The efficacy of the DL model is proportional to the quality of the data used to train it [5]. Data cleansing, the act of removing wrong, irrelevant, or duplicate data and the formatting of unformatted data [7], is often required to obtain a qualified dataset. The quality of a DL model is also affected by the quantity and diversity of its input data.

2.2. Features’ Descriptors

When describing a system, we make use of its properties or descriptors [6,7]. In the DL process, these were chosen as inputs because of their centrality to the data. A conventional molecular descriptor can give either a qualitative or quantitative account of the system. Hydrophobicity is one characteristic that can be used to characterize molecules (1: yes, and 0: no). The molecular weight is a good illustration of a quantitative description. Another common kind of descriptor is the group contribution descriptor. Using this characterization, we can pinpoint the exact occurrences of each distinct fragment. There have been thousands upon thousands of terms coined up to this point. Having a plethora of adjectives to choose from highlights how crucial it is to make the right choice. In general, the selected descriptors should not contain highly linked terms and should be pertinent
Methods [1–8] for selecting descriptors have been developed to eliminate superfluous characteristics without losing sight of the most crucial ones. Recent advancements in natural language processing (NLP) have had an impact on a variety of innovative approaches to the generation of descriptors. However, crystalline structures and amorphous systems, both of which are instances of bulk systems, are described in a way that is fundamentally different from the way that molecules are amorphous systems, and they are classified as having a high degree of disorder. Smooth overlap of atomic positions (SOAP) [2,13,14] and atom-centered symmetry functions (ACSF) [15] are two of the several methods that have been created to characterize the local environment of each atom in a system.

2.3. DL Algorithms

Predicting the value of one variable in a system from the values of other variables is a common application of algorithms [9], which are mathematical models. In either a qualitative or quantitative fashion, the algorithm can link the descriptors to the results. This facilitates insight into the processes and reliable forecasting of future data. To successfully build a model for DL, one must give careful attention to the process of picking an algorithm that is suited for the task at hand. There is a possibility that this subfield of DL is one of the subfields that is advancing at the quickest rate [10]. Recent times have seen a substantial increase in the number of suggested new algorithms, in addition to modifications to already existing ones. These algorithms can be included in either the supervised or unsupervised learning categories [15]. Supervised learning includes scenarios such as learning from labeled data where both the inputs and outputs are known and training a model to predict new input values. These numbers might represent system variables or signify the category of unidentified chemicals in the context of applications in molecular thermodynamics (a classification challenge). Examples of common approaches include recurrent neural networks (RNN) [2], Gaussian process regression [3], and convolutional neural networks (CNN) [4]. Unsupervised learning approaches, on the other hand, make use of data that have not been labeled to train models that are capable of automatically classifying incoming data. The K-means clustering approach, the hierarchical clustering method, and the Gaussian mixture model are all examples of common methods [2,15,16].

3. Prediction of Thermodynamics Using DL Algorithms

The thermodynamic properties of the molecules at hand are essential to the successful completion of a wide variety of chemical processes, including reaction, separation, and purification. Quite a lot of information about thermodynamic properties has been collected by scientists. Still, it is next to impossible to compile information for every possible compound and combination of drugs. That is why a good predictive model is so important. Historically, predictive models were built on spreadsheets and were underpinned by mathematical functions generated from “hypothesis-driven” techniques. DL algorithms provide an alternative, more accurate method of forecasting thermodynamic characteristics. These methods are mechanism-agnostic and do not require you to know anything about the shape of the equation of state in advance. In response to a mountain of experimental data, scientists have compiled numerous comprehensive databases on thermodynamic properties. The Chemistry Webbook [16] maintained by the NIST and the Molecular Properties and Properties of Materials Research Database [17] maintained by the Department of Energy are two examples of databases that fall into this category. These datasets might be put to use in the construction of the DL foundation. Research on critical characteristics, phase shift enthalpy, and other physical factors has been improved by the combination of DL and these datasets [18,19]. We will look at two different scenarios to illustrate how DL works in practice. The first scenario shows a hybrid network (neural network and convolutional neural network) model that can assess the molecular density and viscosity of biofuels. Let us have a look at the second scenario, which will help us
better understand how the free energy of the solvation of organic solutes in typical solvents will manifest itself.

3.1. Density and Viscosity of the Liquid

Viscosity is the characteristic used to determine the relative thickness or thinness of fluid in the science of fluid dynamics. The average distance that may be anticipated to be covered by two particles is the distance that is used to define the density of a fluid. Although both viscosity and density are characteristics of the fluid, they do not correspond to each other in a one-to-one manner. In petrochemicals, aviation fuels, and other uses, the density and viscosity of liquids are essential thermodynamic parameters. Several different DL models, including convolutional neural networks (CNN) and neural networks (NN), were evaluated to see which one was the most accurate in predicting the density and viscosity of liquids. An input layer, many hidden layers, and an output layer were the three primary components that made up NNs [11]. Neurons, which are tiny computer processors and the main structural element of the brain, were abundant in each layer. These neurons take in information from all the neurons that are located below them, perform linear calculations and non-linear activation on that information, and then send the results back down to the cell that was responsible for generating the input information. Throughout the training, the weights and biases of each neuron were fine-tuned to ensure that it functioned to the best of its ability, thereby maximizing its potential.

3.2. Solvation-Free Energy

When it comes to preserving phase equilibrium, solvation is just as crucial as it is in a broad range of other chemical processes. This is because sustaining phase equilibrium is essential to the stability of a system. Researchers in the domains of chemical and biological engineering are interested in solvation-free energy because it is an extra property that is vital to thermodynamics. The collection and maintenance of solvation properties have been handled by several databases, including ESOL (Estimated Solubility) [17], Free-Solv (The Free Solvation Database) [18], and MNsol (The Minnesota Solvation Database) [19,20]. Lin et al. [7] developed a DL model for solvation-free energy in generic organic solvents and termed it Delfos [20]. MNsol served as the basis for this model. A solvation-free energy calculator may be found in this model. This indicated that a DL method was efficient in finding and isolating the key substructures of the compound while the process of solvation was taking place. Both 10-fold cross-validation and cluster cross-validation were utilized so that the predictive potential of Delfos could be assessed. This was performed to ensure that the model was as accurate as possible.

In another piece of research that investigates the relationship between viscosity and temperature for pure hydrocarbons, Masi et al. [9] employ DL to make predictions about the parameters of the empirical viscosity equation. They accessed the database that was kept by the NIST so that they might achieve their objective. After doing so, they retrieved from that database the information about the dynamic viscosity of a total of 261 distinct kinds of pure hydrocarbons. To begin making use of these data, the very first thing that needed to be performed was to regress a better version of the Andrade equation [4], which locates g(T) based purely on B and T0. Subsequently, a DL model was built to determine B and T0 using the molecular weight and 35 chemical descriptors. The various classes of hydrocarbons, including paraffin, naphthenes, and aromatics, were each denoted by one of only fifteen straightforward labels, and the isomers were shown using twenty distinct clusters. The B and T0 layers of the DL model each made use of a neural network that was composed of 18, 3, 36, and 10 layers, respectively. The Andrade equation was then applied to the data to calculate the hydrocarbon dynamic viscosity, g(T), using the predicted values. The method used in their study is shown in Figure 3.

DL may commonly lead to the finding of new and relevant data when it is used to make predictions of thermodynamic parameters. It is feasible that data fitting may aid researchers in understanding the underlying mechanism and giving more accurate and
focused data, but we should not expect it to completely replace an in-depth examination of the system. On the contrary, we should anticipate that it will assist researchers in acquiring a deeper comprehension of the process [9]. The inclusion of DL into databases of experimental chemistry offers a large amount of potential, and datasets need to be extended to include as many unique substructures as they possibly can inside the test set.

Figure 3. Visualization of the deep neural network model and training to predict $B_0$ and $T_0$, which were used to calculate the Andrade equation.

4. Development of Force Fields Using DL

To study the thermodynamics of complex systems, the DL force field can combine the precision of quantum mechanical computations with the speed of classical molecular descriptors. This ranges from relatively simple water [2,14] to more complicated organic and inorganic substances, all of which are studied using innovative force fields. During the training phase, quantum mechanical computations were used to compile a database of atomic coordinates and related parameters, such as system energy. This database was used to test the accuracy of the model. The local structural interrelationships of a significant number of atoms were extracted with the help of descriptors, and after that, they were incorporated into the DL force field [2]. With a set of attributes for a structure, it is feasible to infer the potential energy surface of the structure by utilizing DL methods. When it comes to explaining the complicated system interactions that take place, simulations that are run using traditional molecular descriptors are consequently not as dependable as the DL force field. Therefore, the study [4] used a machine-learned all-atom force field for Li–Si alloys and a machine-learned coarse-grained force field for water molecules. Both force fields were machine-learned and used to model the forces exerted by water molecules on an alloy.

As part of the research effort, a force field that operates on the principle of artificial intelligence has been produced [9]. This force field was developed for both crystalline and amorphous Li–Si alloys with Li/Si ratios from 0 to 4.2. In line with previous studies, models guided by chemical descriptors and a force field may be able to predict the volume
change that occurs during the early phases of lithiation. The study [4] also showed that the force field accurately predicted bulk densities, radial distribution functions, and the diffusivity of Li in amorphous Li–Si systems. This change in the amount of free energy that is accessible is accurately represented by the DL force field. It is difficult to depict the phase-change process that takes place over time due to the utilization of all-atom force fields in the simulations of molecular descriptors that are being utilized nowadays. Coarse-grained simulations of the process of water crystallization were carried out making use of the mW monatomic force field, as the findings described in the study [21] indicate. Because the CG model does not take hydrogen into explicit consideration, the process of ice nucleation can be sped up by a factor of ten or more, allowing for a faster formation of ice crystals. This is because hydrogen is not explicitly accounted for in the model. The Landau free energy model depicts the phasing characteristics of the phase-changing model in the most accurate way possible.

5. Integration of DL in Molecular Simulation

When large amounts of data are generated by molecular simulations, the DL process can be made simpler and finished more quickly. When the purpose of an experiment is to determine the essential characteristics of something, the stakes become significantly more important. Because it is difficult to conduct experiments on the vast number of different possible combinations of solutes and solvents, there are not a lot of datasets available on the topic of diffusion properties. The reason for this is that diffusion property is a very important topic. When high-throughput frameworks are utilized, simulations can produce datasets that are many orders of magnitude larger than the data that corresponds to the experiments that were carried out. These datasets may be utilized to reach inferences regarding the system that is being simulated. This is made possible because of the inherent capacity that high-throughput frameworks must produce datasets. In addition, the simulated dataset that is produced as a result of the high-throughput process has the potential to contain a great deal of variety if enhanced sampling methods such as active learning are utilized during the production of the dataset.

In the discipline of chemical engineering, alkanes are among the most valuable molecules that can be derived from petroleum. This is because they may be used to make other, more useful compounds. Before substances can be employed in the planning of chemical processes or the hunt for new compounds, their thermodynamic properties need to be analyzed. High-throughput simulation of force fields (HT-FFS) was made possible by a method presented in [22]. In combination with DL and NN, the HT-FFS computational framework is used to perform calculations and make predictions regarding the thermodynamic properties of alkanes. They determined the values of a chosen few features of 876 common alkanes by using molecular simulations and came up with a total of 49,044 state points for their calculations. The precision of the simulations was demonstrated by the fact that they were found to be accurate, which was made possible by the utilization of previously acquired experimental data from the NIST standard reference database [7]. Throughout the translation process, the OpenBabel package was responsible for selecting the structural descriptors that were used. There were a total of 25 distinct descriptors that have been included in the first descriptor list. Temperature and pressure are two descriptors that may be used to describe the thermodynamic state point. The thermodynamic state point can be characterized using a variety of other descriptors. When it comes to characterizing the state point, both of these descriptors contribute to the process in their unique ways. After then, a method called recursive feature elimination (RFE) [23] was applied to reduce the total number of descriptors, and then an SVM was implemented after that. For example, the amounts of tertiary and quaternary carbon are the qualities that are most likely to shift as a direct result of the effect that the created dispersion energy has on the system [7, 9, 24]. This is because these quantities are the ones that are most susceptible to change. This is because these quantities make up the majority of the
structure. The amount of methyl groups that are attached to the quaternary carbon, on the other hand, is the factor that has the greatest impact on the substance’s density.

6. Conclusions and Perspective

The significance of the potential that machines possess was illustrated in the previous cases: instruction in chemical engineering that focuses on predicting and analyzing the thermodynamics of molecular systems. Furthermore, benefits can be gained by combining DL with high-throughput simulations when designing materials. One more valuable skill enabled by the DL force field is the capacity to calculate thermodynamics in complicated systems. However, DL’s implementations in the field of chemical engineering are still in their early stages. This is mostly due to the unique requirements of the datasets employed in this field. Most prediction models in the fields of chemistry and materials science only use very few datasets. This is because it may be difficult and costly to ascertain the exact properties of chemicals. There are issues with the model’s extrapolation performance and its adaptability to fresh data. Therefore, the future applications of DL must create databases that have a large quantity of data diversity and dependability. We predict a rise in the number of applications that incorporate DL alongside simulation. One more thing to be concerned about is the interpretability of DL models. For instance, it is difficult to establish a direct connection between the weights of the many nodes in ANN and the factors that determine the properties of molecules. New algorithms that are being developed in computer science can be used as a resource in chemical engineering. Concepts such as active learning, CNN, and attention mechanisms have gained popularity among chemists in recent years. These cutting-edge methodologies open up new doors for the development of DL applications within the realm of molecular thermodynamics.

Although DL in molecular thermodynamics is not a novel concept, it has the potential to revolutionize the study of these properties if we can successfully integrate it with the wealth of databases from our area of chemical engineering. Therefore, if we were to select between the phrases “Hooray, and up she rises!” and “Keep calm and keep on!” we would choose the former. We would strive towards the “Hooray,” but not in an irrational manner. Instead, we would bear in mind where we came from and build on the information gained from previous generations.

DL has the potential to become an important future application as it combines the benefits of quantum physics with atomic and molecular dynamics. To completely realize the promise of DL, there are still obstacles that need to be conquered, such as element type constraints and the sensitivity of input configurations. The study of DL is becoming an increasingly valuable adjunct to the study of chemistry as a consequence of the various DL potentials that have been created and effectively applied to a range of systems. This is one of the reasons why DL has turned out to be a helpful supplement to the study of chemistry.

**Author Contributions:** Conceptualization, H.M.; methodology, M.U.C.; formal analysis, M.J.; investigation, H.M.; resources, M.U.C.; writing—original draft preparation, H.M.; writing—review and editing, M.U.C. and M.J.; visualization, H.M.; supervision, M.U.C.; project administration, M.J.; funding acquisition, M.J. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Wroclaw University of Science and Technology, K38W05D02 and SGS Grant from VSB—the Technical University of Ostrava under grand number SP2022/21.

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.
References


