International Journal of Chemical and Biochemical Sciences
(ISSN 2226-9614)

Journal Home page: www.iscientific.org/Journal.html

The Utilization of a Deep Learning Approach to the Synthesis of Molecules and the Prediction of their Properties

Jaskirat Singh[1], Padmapriya G[2], Vaishali Singh[3], Rakesh Kumar Dwivedi[4], Sanskriti Tiwari[5]

[1]Centre of Research Impact and Outcome, Chitkara University, Rajpura, Punjab, India
[2]Assistant Professor, Department of Chemistry, School of Sciences, JAIN (Deemed-to-be University), Karnataka, India
[3]Assistant Professor, Maharishi School of Engineering & Technology, Maharishi University of Information Technology, Uttar Pradesh, India
[4]Professor, College of Computing Science and Information Technology, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India
[5]Assistant Professor, Department of Microbiology, Parul University, PO Limda, Vadodara, Gujarat, India

Abstract

The synthesis of molecules involves integrating atoms to produce compounds. Molecular properties are determined by structure and bonding, which influence physical and chemical features, determine reactivity, and establish functioning in various kinds of applications. Molecular synthesis provides problems in terms of accuracy, efficiency, and environmental effect, which affect qualities including stability, toxicity, as well as reactivity. In this research, we proposed an innovative technique of hybrid gradient-based optimization with a multi-layer recurrent neural network (HGO-MLRNN) for predicting the molecular property using deep learning (DL). To begin with, Drug Bank and CHEMBL datasets were collected for this study. In addition, we utilize the min-max normalization for data pre-processing, and Principal component analysis (PCA) is employed for feature extraction. As a result, to examine the efficiency of molecule property prediction, we leverage metrics like mean squared error (MSE), root mean square error (RMSE), and mean absolute error (MAE) for existing and proposed methods. Our proposed HGO-MLRNN method attains MSE (0.0201), RMSE (0.1231), and MAE (0.0612) which provides better results compared to the existing methods. Predictive property models and efficient synthesis methods combined improve molecular design, making it easier to produce useful molecules and promoting developments in a number of scientific domains.

Keywords: Molecule, Property, Synthesis, Multi-Layer Recurrent Neural Network, HGO

Full length article **Corresponding Author*, e-mail: jaskirat.singh.orp@chitkara.edu.in

1. Introduction

Molecules are basic units of matter, which form the foundation of all chemical compounds. Molecules are made up of atoms bonded together; exhibit the various properties and features that originate from the specific arrangement of their component elements [1]. These entities perform an important role in several natural processes and synthetic applications, serving as the fundamental elements for the various ranges of compounds and materials found in our surrounding environment [2]. Based on the simplest diatomic molecules to complicated macromolecules like deoxyribonucleic acid (DNA) and proteins, the molecule constitutes an immense and complicated surface that underlies the intricacy of organisms and the substances humans encounter across in our everyday lives [3]. Further, advancements in molecular research have led to significant developments in sectors including medicine as well as materials science, and nanotechnologies, revealing the profound influence that an in-depth understanding of molecules could have on scientific and technological advancements [4]. Molecular syntheses are the process of mixing atoms to produce new molecules. It is a significant process in biochemistry subsequently, it allows for the design as well as synthesis of an extensive range of compounds, including pharmaceuticals and materials [5]. Controlled manipulation of chemical processes allows for the synthesis of a wider range of molecules, therefore contributing to scientific progress and technological innovation [6]. Molecules, the basic building units of matter,

play an important role in defining the characteristics of substances in the physical world [7]. The microscopic entities are made up of atoms that have been chemically bonded together to produce a variety of compounds with distinct properties [8]. The molecular properties delve into the complexities in which these molecular structures interact, determining material behavior and properties. At the core of molecular characteristics are components such as molecule shape, size, as well as composition, each contribute to the entire behavior of the substances [9]. The type and strength of intermolecular interactions between molecules also significantly influence their chemical and physical properties [10]. Exploring these features offers valuable insight into phenomena like as solubility, melting temperatures, and reactivity, which are critical for understanding the larger consequences of molecular interactions [11]. This research into the characteristics of molecules extends beyond the laboratory, which influences the comprehension of natural processes, industrial utilizes, and the complex workings of biochemical structures [12]. Complexity, resource-intensive techniques, possible environmental effects, and limited predictability are the difficulties in synthesizing molecules with desired properties. The main goal of this research is to efficiently synthesize the molecules and predict their properties using DL, improving materials and discovering drugs through cutting-edge computational techniques. The study [13] proposed a graph convolution algorithm that exhibits consistent performance comparable with, or surpassing, algorithms employing constant molecular variables, in addition to prior implementations of graph neural networks on proprietary as well as public datasets. The research [14] explained that several Simplified Molecular-Input Line-Entry Systems (SMILES) are encoded for each molecule. It is an automatic augmenting data technique utilized in molecular property prediction that mitigates the issue of over fitting that arises from the limited data size present in molecule property forecasting datasets. The article [15] described the molecular transformers generate predictions through the process of assuming connections among the presence as well as absence of molecular structures in the given dataset's reactants, reagents, as well as products. It forecasts intricate chemical alterations without the need for customized principles. The paper [16] examined the development of the trained Bidirectional Encoder Representation from Transformers (BERT) to gather molecular sub-structural data, which can be utilized for forecasting molecular properties. They introduce molecular representation with bidirectional encoder representation from transformers (Mol-BERT), an innovative end-to-end DL system that integrates a trained BERT algorithm specifically designed for molecular characteristic forecast in an efficient molecule representation.

2. Materials and Methods

The HGO-MLRNN technique is employed for predicting the molecular property. In the beginning stage, the dataset were gathered. The data collection was divided into two processes; they are training, and testing. The training process includes data pre-processing, feature extraction, molecular property prediction, and molecular synthesis. The testing process includes results. The second stage is to preprocess the data were done by min-max normalization. After that process, the PCA technique is used for the feature extraction process. Following that, the extracted data was utilized by the HGO-MLRNN method to predict the molecular property. The entire flow of this molecular property prediction is shown in (Fig.1).

2.1. Dataset

The CHEMBL [17] is an open data resource that makes use of the Drug Bank and CHEMBL databases to provide detailed data on the functional, binding as well as ADMET characteristics of a wide range of drug like bioactive chemicals. This important data is handpicked, chosen and standardized from reputable literature sources to improve its quality and suitability for use in many chemical biology and drug development research fields. As the CHEMBL database has 5200 protein targets and 5.4 million bioactivity values for more than a million compounds. They identified molecules using the SMILES string format, which was created with both grammatical consistency and machine friendliness in mind. The characters used in SMILES representations that represent atoms, bonds and chemical structures. The molecule length ranges from 35 to 75 and the SMILES molecule is encoded in a one pass, providing every character in a 53-dimensional vector of zeros. We split the dataset into two types, namely: training (80%) and testing (20%).

2.2. Preprocessing using Min-max normalization

Pre-processing in molecular property prediction improves data representation for computing efficiency by using methods including reducing dimensionality, component determination, as well as scalability to improve model training with prediction speed. Min-max normalization improves effectiveness in predicting molecular properties. By converting the component values into a standardized range, preventing numeric instabilities, and increasing speed convergence throughout model development. Such streamlined preprocessing allows for quicker and more effective development, which leads to better prediction performance. Min-max normalization, also referred as variation normalization, corresponds to a linear adjustment of the original information, where max represents the maximum while min represents the minimum of sample information. Using the counterfactual identification technique, the dimension of a normal characteristic is zero. The value of an anomalous characteristic represents a positive integer. Therefore, it must be normalized according to a natural integer. Normalizing the data is an important stage, where every value must be stretched to a suitable range. This approach assists to reduce large discrepancies in characteristics:

$$\Psi_{j,i} = Round\left[\left(\frac{Z_{ji}-\min(Z_i)}{max(Z_i)-\min(Z_i)}\right) * M\right] \qquad (1)$$

Where $\Psi_{j,i}$ denotes a normalized value from Z_{ji} within the range of $0\ to\ M$ in integer representation, min (Z_i) is the minimal value that defines the i^{th} characteristic, while max (Z_i) represents a maximal value that describes the i^{th} characteristic.

2.3. Feature extraction using Principal component analysis (PCA)

Feature extraction in molecular property prediction simplifies data by extracting important information from molecular structures, increasing computational effectiveness and allowing for more accurate predictions with less computational materials. PCA improves molecular property forecasting through the extraction of essential properties from significant information. It decreases dimensionality and captures the most important variance. In terms of molecular properties, PCA improves model effectiveness by concentrating on essential molecule characterizations, resulting in more accurate predictions with less computing complexity. PCA provides one of the most extensively used methods for reducing information dimensionality. Its primary goal is to correlate information samples between highly dimensional regions to a lower dimensional region utilizing an orthogonal matrix. The objective function for I_{PCA}could be properly defined as equation (2).

$$I_{PCA}(Z) = \max_{O} \sum_{j=1}^{m} ||z_j - \bar{z}||^2 \quad (2)$$

$$p.s. O^S O = 1$$

Equation (2) might be reduced to the corresponding trace representation after a straightforward algebraic modification, that is, equation (3).

$$I_{PCA}(O) = \max_{O} \sum_{j=1}^{m} ||O^S(w_j - \bar{w})||^2$$
$$= \max_{O} sq\{O^S(w_j - \bar{w})(w_j - \bar{w})^S O\} \quad (3)$$
$$= \max_{O} sq\{O^S D O\}$$
$$p.s. O^S O = 1$$

The covariance matrices represent $D = \sum_{j=1}^{m}(w_j - \bar{w})((w_j - \bar{w})^S$, and $\bar{w} = \frac{1}{m}\sum_{j=1}^{m} w_j$. While $sq(*)$signifies matrix $*$ trace, which is the sum of its principal diagonal components of * matrix.

2.4. Hybrid Gradient-Based optimization with multi-layer recurrent neural network (HGO-MLRNN)

HGO-MLRNN combines hybrid gradient-based optimization and multi-layer recurrent neural networks to create molecules and predict their properties, developing applications in discovering drugs and materials research.

2.4.1. Multi-layer recurrent neural network (MLRNN)

MLRNN predicts molecule characteristics through discovering the complicated connections in molecular structures, which helps with synthesizing optimization and property estimate. The MLRNN design leverages earlier information $(s-1)$ to create result information for the present time (s). The information provided and it transmitted through a hidden layer for development. There is an interface to keep the prior information from the hidden component in the context of the element. The equation is provided through the following in Equation (4-5):

$$g_s = \varphi_g(V_{jm} w_s + U_g g_{s-1} + a_g) \quad (4)$$

$$z_s = \varphi_z(X_{out} g_s + a_z) \quad (5)$$

Both vectors $g_{s-1}\ and g_s$represent the hidden layers from earlier and present times, correspondingly. The activation performed for the hidden as well as output layers are $\varphi_g\ and\ \varphi_z$, accordingly. Where V_{jm} represents the weighted matrices among the hidden layers as well as input. Where U_{jm}represents the weighted matrices among the hidden layers.Both vectors $a_g\ and\ a_z$ indicate biased in the hidden and output layers respectively, and X_{out} represents the weighted matrices among the hidden and output layers. Standard activation operators such as Rectified Linear Unit (ReLU), sigmoid, as well as tanh might have slow converging speeds. As a result, alternative non-linear operations, such as bipolar-sigmoid, as well as power-sigmoid activation processes were employed in developing MLRNN. The equations are described as follows in Equation (6-7):

Bipolar-sigmoid activation processes:

$$\varphi(w) = \frac{1-f^{-\varepsilon w}}{1+f^{-\varepsilon w}} \quad (6)$$

Where$\varepsilon > 2$.

Power-sigmoid activation processes:

$$\varphi(w) = \begin{cases} \frac{(1-f^{-\varepsilon w})(1+f^{-\varepsilon})}{(1-f^{-\varepsilon})(1+f^{-\varepsilon w})} |w| < 1 \\ w^b |w| < 1 \end{cases} \quad (7)$$

Where$\varepsilon > 2\ and\ b \geq 3$.

2.4.2. Hybrid Gradient-Based Optimization (HGO)

HGO improves molecule synthesis efficiencies by using optimized structures and gradient-based approaches to predict properties accurately and optimize specifically. The HGO relies on 2 operators for updating the solutions; every one of the molecules has its own function. The initial function represents a Gradient Search Rule (GSR), that is utilized toward enhance the discovery, and a second function serves as the Local Escape Operators (LEO) that are utilized to improve the exploitation capability. The initial step in HGO is to create a molecule $W\ with\ M$solutions that are arbitrarily generated employing the equation that follows:

$$w_j = w_{min} + rand \times (w_{max} - w_{min}), j = 1,2,..,M \quad (8)$$

The searching region limitations are$w_{min} and w_{max}$, and $randrand \in [0,1]$ indicates an arbitrary number. The value of fitness for every solution is subsequently calculated and the optimal solution is selected.The GSR with direction movements (DM) is used to modify the outcomes $\left(w_j^{js}, j = 1,2,..,M\right)$in the direction$(w_a - w_j^{js}).w_a$. Describes the most optimal solution. This update procedure is

accomplished by calculating novel three solutions as $w1_j^{js}, w2_j^{js}, and\ w3_j^{js}$

$$w1_j^{js} = w_j^{js} - GSR + rand \times \rho_1 \times \left(w_a - w_j^{js}\right) \quad (9)$$

Equation (10) uses $\rho 1$ to enhance the balance among exploiting and exploring throughout the optimizing procedure. It is described as in Equation (10)

$$\rho_1 = 2 \times rand\ \alpha - \alpha \quad (10)$$

Where

$$\alpha = |\beta \times \sin(3\pi/2 + \sin(\beta \times 3\pi/2))|$$

$$\beta = \beta_{min} + (\beta_{max} - \beta_{min}) \times (1 - (Js/Max_{Js})^3)^2$$

Where $\beta_{min} = 0.2\ and\ \beta_{max} = 1.2$.

Whereas Max_{Js} represents entire iterations. The GSR is described in the following manner in Equation (11).

$$GSR = \times randn \times \rho_2 \times (2 \times \Delta w \times w_j^{js})/(zo_s - zr_s + \varepsilon) \quad (11)$$

With

$$\Delta w = rand(1{:}M) \times |((w_a - w_{q1}^{js}) + \delta)/2|$$

$$\delta = 2 \times rand \times (|(w_{q1}^{js} + w_{q2}^{js} + w_{q3}^{js} + w_{q4}^{js})/4 - w_s^{js}|)$$

Where, $rand(1{:}M)$ is a randomized vector with Mdimension $(q1, q2, q3, and\ q4)$ referring to picked integers from$[1, M]$. Equation (10) defines the formulation for ρ_2. Equations (12) and (13) update the positions$zo_j\ and\ zr_j$, respectively.

$$zo_s = rand \times \frac{w_t + w_s}{2} + rand \times \Delta w \quad (12)$$

$$zr_s = rand \times \frac{w_t + w_s^{js}}{2} - rand \times \Delta w \quad (13)$$

With

$$w_t = w_j^{js} - randn \times \rho_1 \times (2 \times \Delta w \times w_j^{js})/(w_a - w_{worst} + \epsilon) \quad (14)$$

$$w2_j^{js} = w_a - GSR + rand \times \rho_2 \times (w_{q1}^{js} - w_{q2}^{js}) \quad (15)$$

$$w3_j^{js} = w_j^{js} - \rho_1 \times (w1_j^{js} - w2_j^{js}) \quad (16)$$

Lastly, depending on the coordinates $w1_j^{js}, w2_j^{js}\ and\ w3_j^{js}$, a novel solution at iterations $Js + 1$ are discovered in Equation (17):

$$w_j^{js+1} = q_b \times \left(q_a \times w1_j^{js} + (1 - q_a) \times w2_j^{js}\right) + (1 - q_b) \times w3_j^{js} \quad (17)$$

Where $q_a\ and\ q_b$ represent two arbitrary integers. Furthermore, the LEO is used to increase the exploiting capability of GBO. For modifying the solution w_j^{js} based on the possibility oq, use the equation provided below in Equation (18):

$$w_j^{js+1} = \begin{cases} w_j^{js} + e_1 + X_1 + e_2 \times \rho_1 \times X_3 + v_2 \times X_2/2 & oq < 0.5 \\ w_a + e_1 + X_1 + e_2 \times \rho_1 \times X_3 + v_2 \times X_2/2 & otherwise \end{cases} \quad (18)$$

$$X_1 = \left(v_1 \times w_a - v_2 \times w_l^{js}\right),$$

$$X_2 = \left(w_{q1}^{js} - w_{q2}^{js}\right),$$

$$X_3 = (v_3 \times \left(w2_j^{js} - w1_j^{js}\right))$$

Within equation (18), $e_1 \in [-1,1]\ and\ e_2$represent regular and standard randomized integers, correspondingly. $u_1, u_2\ and\ u_3$ are three randomly generated integers specified as in Equation (19a-c)

$$u_1 = K_1 \times 2 \times rand + (1 - K_1) \quad (19\ a)$$

$$u_2 = K_1 \times rand + (1 - K_1) \quad (19b)$$

$$u_3 = K_1 \times rand + (1 - K_1) \quad (19c)$$

Where K_1denotes binary integer$(i.e., allocated\ to\ 0\ or\ 1)$. Thus, the novel solution is derived utilizing the subsequent equation (20):

$$w_l^{js} = K_2 \times w_o^{js} + (1 - K_2) \times w_{rand} \quad (20)$$

Where $K_2$is equivalent to $K - 1$ then w_o^{js} denotes a selected solutions for W, whereas w_{rand} signifies an arbitrary solution produced employing equation (8).The primary phases of the HGO technique are shown in Algorithm 1.

Algorithm 1: Hybrid Gradient-Based Optimizer (HGO)

Initializing the variables of HGO: ϵ, oq, Max_{Js} Maximum. Iteration number, M: Molecule size. Initialize at random the molecule of M vectors using Equation (8). Estimate the position of every single vector using the fitness factor fit. Establish the worst and best solutions:

$$w_{best}, w_{worst}$$

Let $It = 1$. While $Js \leq Max_{Js} do$ for each vector $w_j^{Js} do$.
Select 4 integers arbitrarily ranging from $[1..M]\ such\ that: q1 \neq q2 \neq q3 \neq q4$. Upgrade the positions of the vector w_j^{Js+1} using Equation (17). Estimate the qualities of the vector w_j^{Js+1} using the fitness factor end for if $rand < oq$ then. Upgrade the positions of w_j^{Js+1} using the I branch of Equation (18) else. Upgrade the positions of w_j^{Js+1} using the II branch of Equation (18).
end if Establish the worst and best solutions: w_{best}, w_{worst} $Js = Js + 1$. end while Return the best solution w_{best}

2.5. Molecular synthesis

Molecular synthesis for molecular property prediction employs computational techniques to design and generate distinctive compounds with the required properties. Using modern modeling methods, it forecasts molecular behaviors depending on composition, structure, as well as environmental variables. Through analyzing the interactions between molecules, quantum mechanics, as well as electronic structure, it develops molecular compounds that have been optimized for particular functions. This method allows for quick exploration of a large chemical dimension, which accelerates material discoveries as well as drugs development. By utilizing DL with statistical techniques, it provides accurate predictions about properties including reactivity, solubility, as well as toxicity, facilitating focused molecular development for a variety of sectors, like materials research, pharmaceuticals and then environmental engineering.

3. Results and discussion

Utilizing the Adam optimizer with a 64-batch batch dimension and a starting learning rate of 0.0001. These parameters have been implemented using Py Torch (version 1.12.0) with a Deep Graph Library. Results for molecule synthesis include assessing computational models' accuracy, efficiency, and predictive capacity in producing molecules and anticipating their properties, which is essential for progressing drug development and materials research. In this article, we collected the existing methods for molecular property prediction like “gated recurrent unit (GRU) [18], long short-term memory (LSTM) [18], as well as one-dimensional convolutional gated recurrent unit neural network (1D-CNN-GRU) [18]”. “Mean absolute error (MAE), Mean squared error (MSE) and Root mean square error (RMSE)” are those metrics used to evaluate the performance of existing and proposed techniques.

3.1. Mean squared error (MSE)

MSE quantifies an average squared variance between predicted and real results. Within the framework of creating molecules and predicting their properties, MSE measures the model's efficiency. Table 1 and Fig.2demonstrate the MSE values and graphical representation. LSTM [18] (0.0287), GRU [18] (0.0292), and 1D-CNN-GRU [18] (0.023) are the MSE values for existing methods. When compared to the proposed and existing methods our proposed method HGO-MLRNN achieves (0.0201) superior results.

3.2. Root mean square error (RMSE)

RMSE improves efficiency in molecule synthesis by offering a succinct metric for evaluating prediction accuracy. It simplifies model assessment, leading individuals toward more reliable and productive approaches to molecule development and property prediction. Table 2 and Fig.3 represents the values and graphical representation for RMSE. The existing method values for RMSE are LSTM [18] (0.1696), GRU [18] (0.1709) and 1D-CNN-GRU [18] (0.1517). Our proposed HGO-MLRNN method (with a value of 0.1231) offers a better RMSE efficiency compared to the existing methods.

3.3. Mean absolute error (MAE)

MAE calculates the average absolute distinction among anticipated and actual values. In molecular synthesis, it measures the accuracy of property predictions and evaluates the model performance in computation circumstances. Table 3 and Fig.4 shows the MAE values and graphical representation. The existing methods attain LSTM [18] (0.0862), GRU [18] (0.0863), and 1D-CNN-GRU [18] (0.0693). In comparison to both proposed and existing methods, our proposed HGO-MLRNN method establishes a superior outcome (0.0612). Fig.5 demonstrates the entire actual and predicted values of HGO-MLRNN.

Table 1. Values for MSE

Method	MSE
LSTM [18]	0.0287
GRU [18]	0.0292
1D-CNN-GRU [18]	0.023
HGO-MLRNN [Proposed]	0.0201

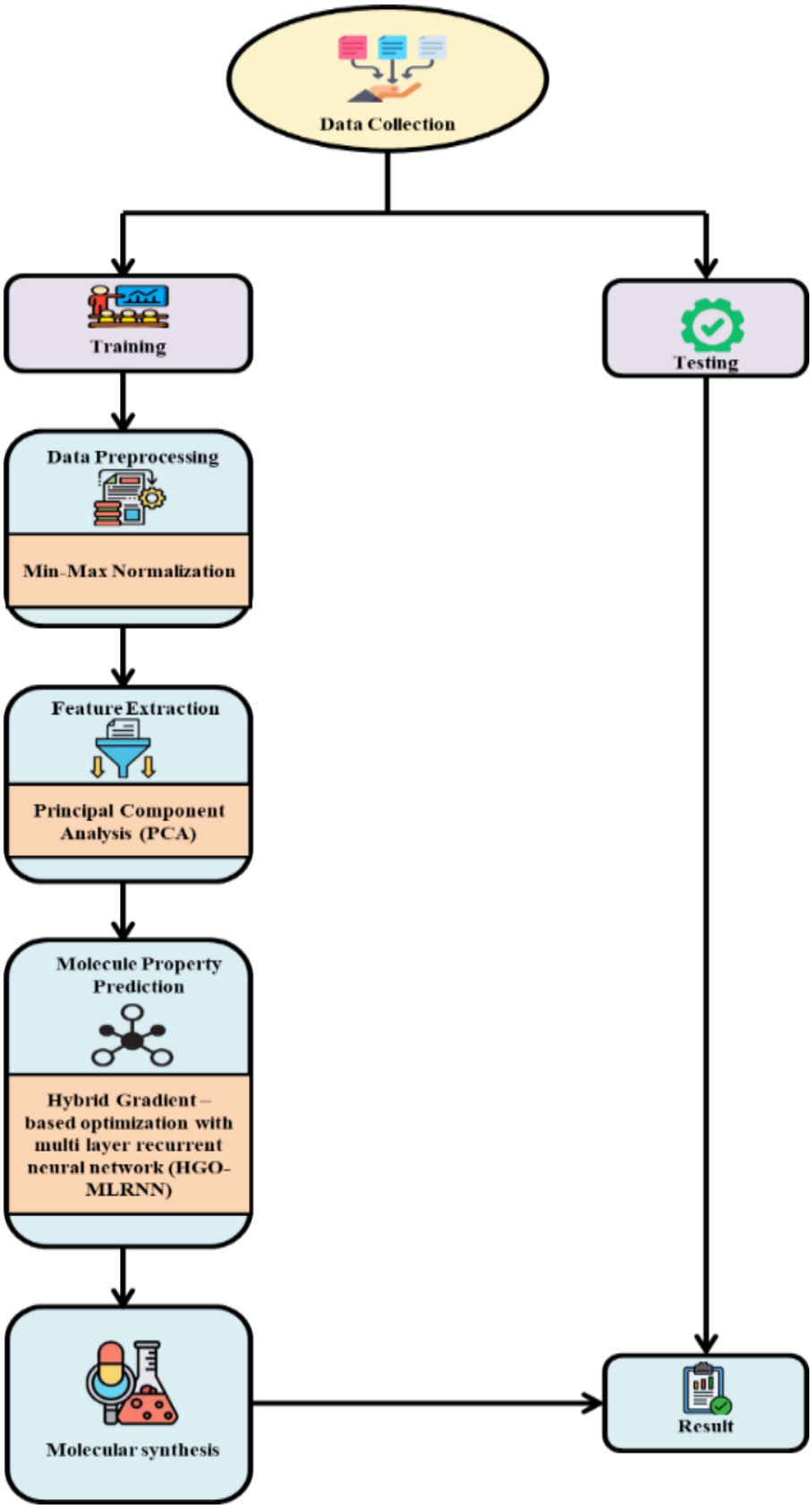

Figure 1. Overview of this research

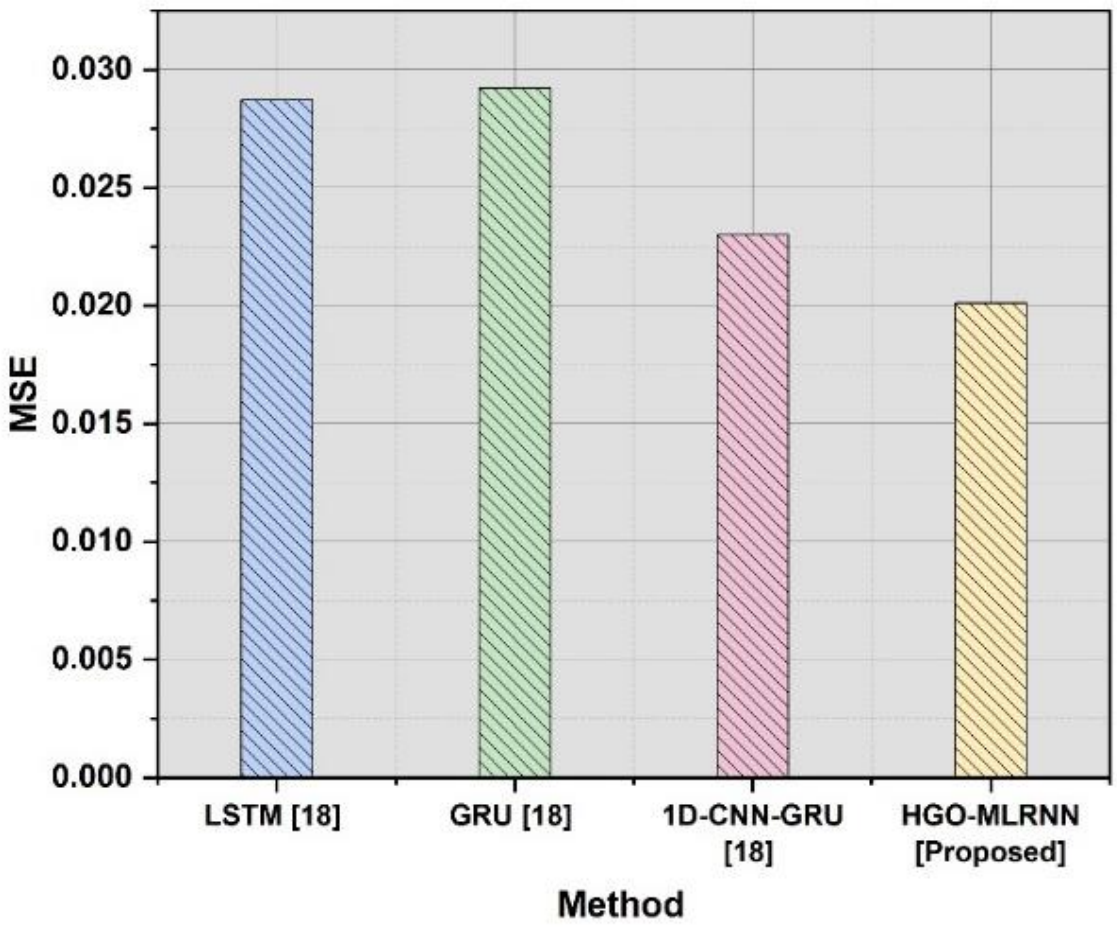

Figure 2. Graphical representation of MSE

Table 2. Values for RMSE

Method	RMSE
LSTM [18]	0.1696
GRU [18]	0.1709
1D-CNN-GRU [18]	0.1517
HGO-MLRNN [Proposed]	0.1231

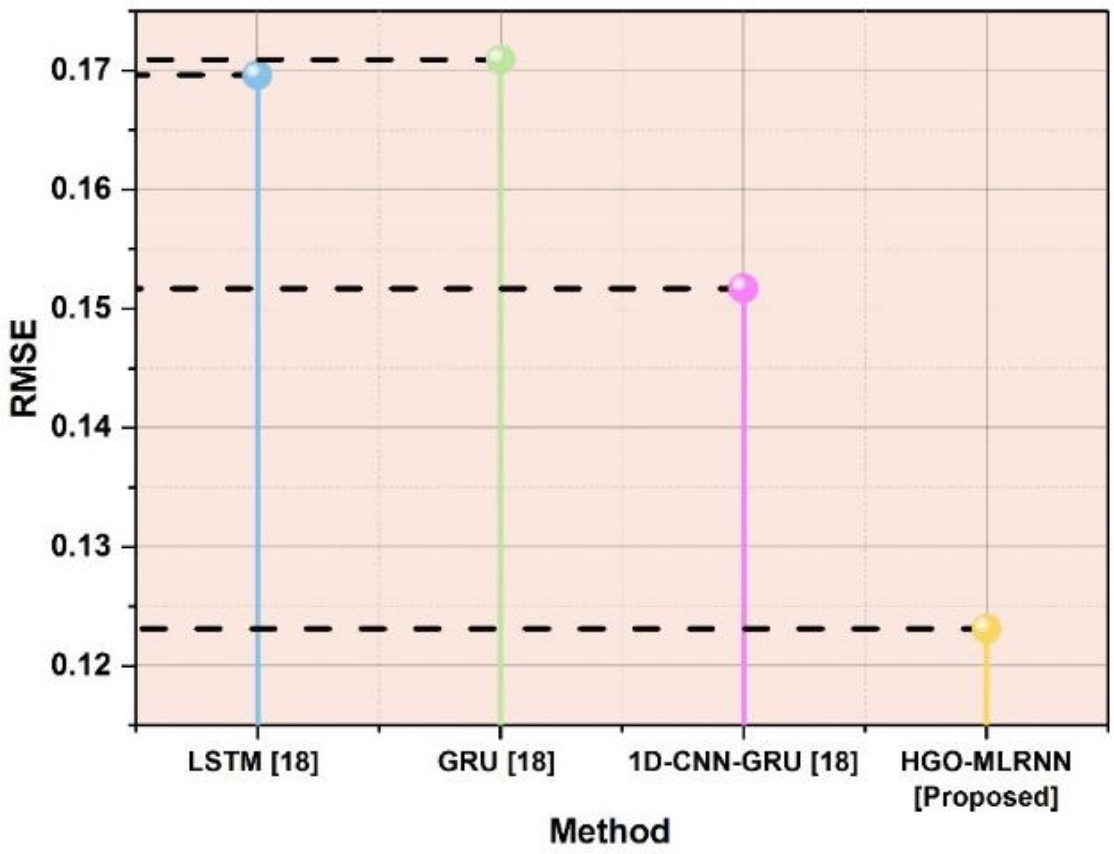

Figure 3. Graphical representation of RMSE

Table 3. Values for MAE

Method	MAE
LSTM [18]	0.0862
GRU [18]	0.0863
1D-CNN-GRU [18]	0.0693
HGO-MLRNN [Proposed]	0.0612

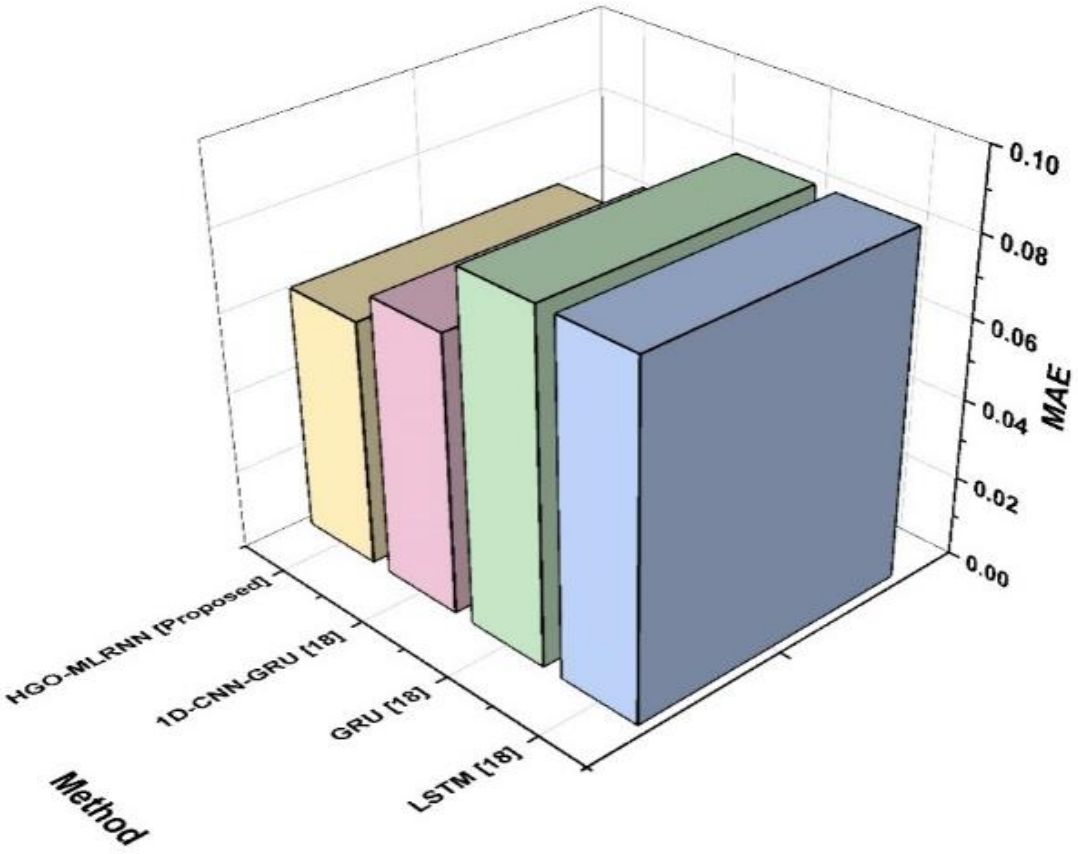

Figure 4. Graphical representation of MAE

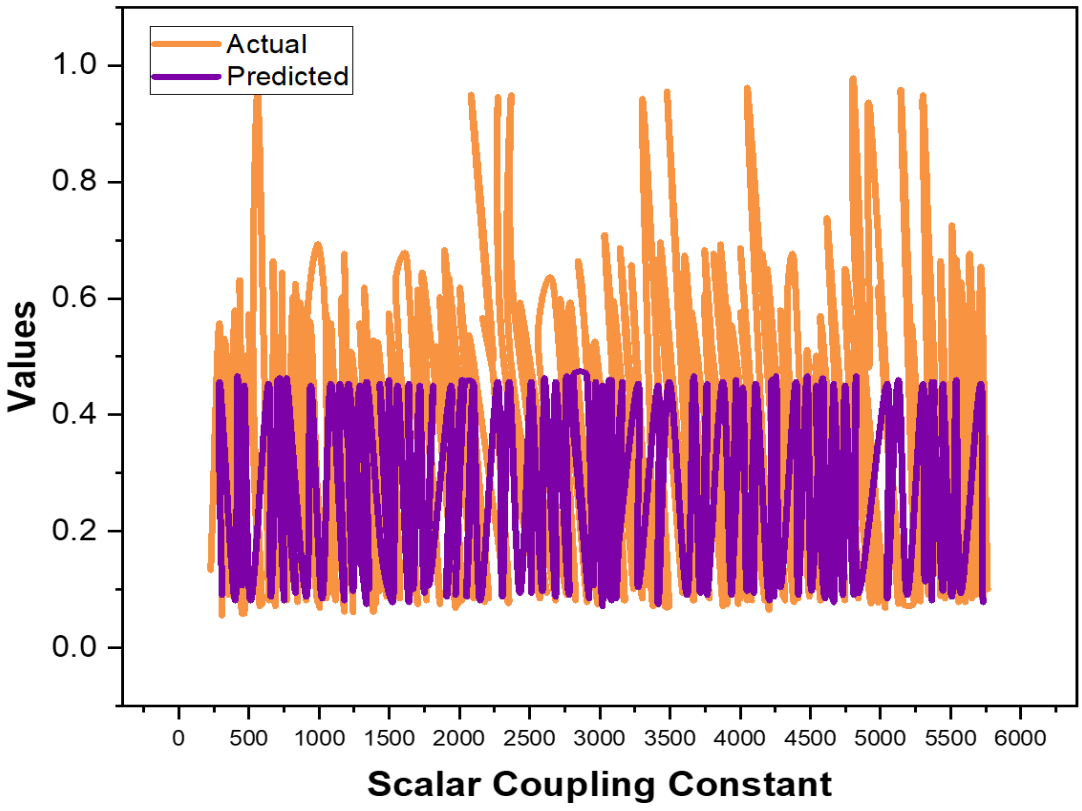

Figure 5. Actual and predicted values of proposed (HGO-MLRNN)

4. Conclusions

The research emphasizes the essential importance of molecules as matter-building components, which influence natural processes and technological progress. The focus on molecule synthesis and properties highlights their importance in a variety of domains, including medicine and materials research. The described Hybrid Gradient-Based optimization with multi-layer recurrent neural network (HGO-MLRNN) methodology for predicting molecular properties shows superior results when compared to existing methods. The result is evidenced by lower MSE, RMSE, and MAE. The implementation of DL and computational approaches in molecular synthesis shows the potential to advance the discovery of drugs and materials research. Considering problems in the synthesis of molecules with desired properties, the suggested research seeks to handle these difficulties, highlighting the importance of molecular information for scientific and technological advancement. The results of this study emphasize the prospective influence of HGO-MLRNN in enhancing molecular property prediction and contribution to the larger field of molecular studies and their applications. Challenges in molecular synthesis and predicting properties include accuracy as well as computational complexities, and various chemical interactions, which restrict prediction ability. Improve molecule synthesis using artificial intelligence (AI) for efficient discovery of drugs, materials structure and properties prediction, hence improving scientific innovations and applications.

References

[1] S. Mena-Hernando, E.M. Pérez. (2019). Mechanically interlocked materials. Rotaxanes and catenanes beyond the small molecule. Chemical Society Reviews. 48 (19) 5016-5032.

[2] S.A. Sandford, M. Nuevo, P.P. Bera, T.J. Lee. (2020). Prebiotic astrochemistry and the formation of molecules of astrobiological interest in interstellar clouds and protostellar disks. Chemical reviews. 120 (11) 4616-4659.

[3] A.S. Pillai, G.K. Hochberg, J.W. Thornton. (2022). Simple mechanisms for the evolution of protein complexity. Protein Science. 31 (11) e4449.

[4] K. Ariga. (2021). Progress in molecular nanoarchitectonics and materials nanoarchitectonics. Molecules. 26 (6) 1621.

[5] K. Murugesan, T. Senthamarai, V.G. Chandrashekhar, K. Natte, P.C. Kamer, M. Beller, R.V. Jagadeesh. (2020). Catalytic reductive aminations using molecular hydrogen for synthesis of different kinds of amines. Chemical Society Reviews. 49 (17) 6273-6328.

[6] M. Arabi, A. Ostovan, J. Li, X. Wang, Z. Zhang, J. Choo, L. Chen. (2021). Molecular imprinting: green perspectives and strategies. Advanced Materials. 33 (30) 2100543.

[7] F. Musil, A. Grisafi, A.P. Bartók, C. Ortner, G. Csányi, M. Ceriotti. (2021). Physics-inspired structural representations for molecules and materials. Chemical Reviews. 121 (16) 9759-9815.

[8] L. Zhao, S. Pan, N. Holzmann, P. Schwerdtfeger, G. Frenking. (2019). Chemical bonding and bonding models of main-group compounds. Chemical reviews. 119 (14) 8781-8845.

[9] D.T. Gentekos, R.J. Sifri, B.P. Fors. (2019). Controlling polymer properties through the shape of the molecular-weight distribution. Nature Reviews Materials. 4 (12) 761-774.

[10] Q. Zhou, J. Ma, S. Dong, X. Li, G. Cui. (2019). Intermolecular chemistry in solid polymer electrolytes for high-energy-density lithium batteries. Advanced Materials. 31 (50) 1902029.

[11] A. Kovács, E.C. Neyts, I. Cornet, M. Wijnants, P. Billen. (2020). Modeling the physicochemical properties of natural deep eutectic solvents. ChemSusChem. 13 (15) 3789-3804.

[12] C. Chen, Y. Kuang, S. Zhu, I. Burgert, T. Keplinger, A. Gong, L. Hu. (2020). Structure–property–function relationships of natural and engineered wood. Nature Reviews Materials. 5 (9) 642-666.

[13] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, R. Barzilay. (2019). Analyzing learned molecular representations for property prediction. Journal of chemical information and modeling. 59 (8) 3370-3388.

[14] C. Li, J. Feng, S. Liu, J. Yao. (2022). A novel molecular representation learning for molecular property prediction with a multiple SMILES-based augmentation. Computational Intelligence and Neuroscience, 2022.

[15] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C.A. Hunter, C. Bekas, A.A. Lee. (2019). Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. ACS central science. 5 (9) 1572-1583.

[16] J. Li, X. Jiang. (2021). Mol-BERT: an effective molecular representation with BERT for molecular property prediction. Wireless Communications and Mobile Computing, 2021. 1-7.

[17] M. Aly, A.S. Alotaibi. (2023). Molecular Property Prediction of Modified Gedunin Using Machine Learning. Molecules. 28 (3) 1125.

[18] D.O. Oyewola, E.G. Dada, O. Emebo, O.O. Oluwagbemi. (2022). Using Deep 1D Convolutional Grated Recurrent Unit Neural Network to Optimize Quantum Molecular Properties and Predict Intramolecular Coupling Constants of Molecules of Potential Health Medications and Other Generic Molecules. Applied Sciences. 12 (14) 7228.