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

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

The Utilization of Machine Learning in the Forecasting of Chemical Reactions and Chemical Synthesis

Nayana Borah[1], Trapty Agarwal[2], Rupal Gupta[3], Anuradha Rohinkar[4]

[1]*Department of Life Sciences, School of Sciences, JAIN (Deemed-to-be University), Karnataka, India*

[2]*Maharishi School of Engineering & Technology, Maharishi University of Information Technology, Uttar Pradesh, India*

[3]*College of Computing Science and Information Technology, TeerthankerMahaveer University, Moradabad, Uttar Pradesh, India*

[4]*Department of Biochemistry, Parul University, PO Limda, Vadodara, Gujarat, India*

Abstract

Chemical synthesis is a process that involves designing and fabricating molecules or compounds using regulated chemical processes. It is essential for creating medications, materials and modifying molecule structures according to desired criteria. Modern chemistry, based on chemical reactions and synthesis, has a significant impact on various scientific and industrial fields.To predict chemical reactions and chemical synthesis; we established an upgraded slap swarm optimization with bidirectional long and short-term memory (USSO-Bi-LSTM). This addresses the challenges in machine learning (ML) concerning the prediction of chemical reactions and synthesis. Our approach seeks to mitigate the lack of data that enhances a model adaption to the intricate structure of chemical space, and lessen issues related to USSO-Bi-LSTM extrapolation beyond training data. By improving data accessibility and refining the model structures, this hybrid approach seeks to increase their resilience to the intricacies of chemical reactions and promote an excellent knowledge of chemical processes.We collected theUnited States Patent and Trademark Office (USPTO) dataset and extracted features from chemical quantum classifiers (CQC).In our study, we employed an evaluation metrics such as True Negative Rate (TNR), Accuracy, and Root Mean Square Error (RMSE). The results specify that our proposed method outperforms than other existing approaches, including Long Short-Term Memory (LSTM), Siamese networks and, Artificial Neural Networks (ANN) representing superior performance across these metrics.ML combines chemical reaction and synthesis prediction for precise navigation, efficient optimization, and innovation in various fields of science and industry.

Keywords: Chemical reaction, upgraded salp swarm optimization (USSO), bidirectional long short-term memory (Bi-LSTM), machine learning (ML)

Full-length article **Corresponding Author*, e-mail: b.nayana@jainuniversity.ac.in

1. Introduction

Chemical reactions and synthesis drive the generation and alteration of substances, the central pillars of chemistry. Chemical reactions are the backbone of many processes in nature and human endeavors, from the delicate tango of atoms to the laborious assembly of large molecules [1]. Orchestrating these reactions to create new molecules with desired qualities is the heart of chemical synthesis, which had a profound impact on industries as diverse as materials science and medicines [2]. All chemical reactions include the breaking and mending of chemical bonds between atoms as their primary mechanism. Both the burning of a hydrocarbon and the development of a new medicine under this type of change [3]. Our understanding of the molecular world and tap into its revolutionary potential, we must be able to predict and manipulate these interactions. To develop and optimize chemical processes, chemists have historically relied on empirical knowledge, intuition, and years of experience [4]. The introduction of ML to the field of chemistry, however, marks a significant transition in our modern, information-driven world. This confluence provides the prospect of expediting the discovery and optimization of chemical processes by using enormous databases, computer algorithms, and predictive models [5]. This study aims to improve chemical synthesis by utilizing ML to enhance precision and efficiency in designing and fabricating molecules or compounds through controlled and optimized chemical processes.

The study [6] provided a data-mining-based, interpretable impurity prediction methodology for massive chemical reaction datasets. Python and RDK it were used to construct a 14-step process that used Reaxysdata.The study [7] examined the challenge of predicting the yield of a reaction, which helps chemist to choose beneficial reactions

with less exploration in a unique chemical region. As a first step towards overcoming the difficulty, they proposed a few-shot yield prediction-specific attention-based random forest model called MetaRF. The study [8] employed a specific feature allows the network to select which data to remove and save, which facilitates handling of data that is necessary for modelling biological-chemical interactions, that may contribute to the development of AGEs. The research [9] provided Graph2Edits, a comprehensive framework for retrosynthesis prediction Motivated by the rigid, arrow-pointing formalism of chemical reactions. To forecast the alterations to the product graph, Graph2Edits use an auto-regressive graph neural network, and then it generates the transformation's intermediates and final products in the expected sequence.

The research [10] discussed the several approaches to CRN construction and analysis that may be taken to achieve a wide range of scientific goals. They examine the ML methods already used to CRNs and outline upcoming CRN-ML strategies, describing the technological and scientific obstacles that must be fulfilled. The study [11] provided ChemCrow, an LLM chemical agent trained to perform operations in organic synthesis, drug discovery, and materials design. ChemCrow improves LLM chemical performance by including 18 tools developed by domain experts, and it enables the emergence of new capabilities. The study [12] presented the SolvBERT approach, which uses the SMILES model of the combination to deduce the solute and solvent. Unsupervised learning was used to pre-train SolvBERT on a substantial collection of computational salvation-free energies.

2. Methodology

The study proposed USSO-Bi-LSTM to have significant potential for optimizing human chemical interactions, improving prediction accuracy and convergence speed, and enhancing knowledge and control of chemical reactions in humans. We collected the USPTO data and extracted features from quantum chemical descriptors. Fig 1 shows the block diagram of our proposal.

2.1. Data collection

The research made use of Janssen's Reactlake response database, which contains datasets from the USPTO, Reaxys, Pistachio, and other sources. The data were subjected to standardization, aromaticity correction, functional group translation, and the use of RxnMapper and CGRtools software. The final dataset consisted of 15.5 million unique reactions, with a bias towards the reactions with yields higher than 5%. The primary purpose of the dataset for masked language modeling (MLM), which looks for the probability distribution of each word in relation to the data that surrounds it. A sub-selection of 750,000 solutions from the Janssen ELN dataset was fine-tuned, revealing the inherent separation and role-specific linkage of the entities [13].

2.2. Feature extraction using chemical quantum classifiers

In different molecular contexts, the selected quantum chemical descriptors represented the atoms' of electrical configurations and chemical reactivity. In the screening stage of polar processes, a combination of vectors pertaining toorbitalsc^{atm} and atoms $c^{atm}$are used to create a total vector c^{tot} of the descriptors for each atom in the reactants.

$$c^{tot} = c^{atm} \oplus c^{orb} \quad (1)$$

Specifically, c^{atm} stores data on electron densities and molecule structures, while c^{orb} stores data on atomic and molecular orbitals. Atoms that act as donors and acceptors use the vector separately. The total vector c^{tot} is defined in the ranking step for both donor and acceptor atoms by utilizingc^{tot}.

$$C^{tot} = c^{tot}_{donor} \oplus c^{tot}_{acceptor} \oplus (c^{tot}_{donor} - c^{tot}_{acceptor}) \quad (2)$$

When solving equation (1) for radical reactions, only $c^{atm}$was considered. The (Y)-number, the (R_{atom})charge, and the condensed Fukui functions 38 are used in quantum mechanics (c^{atm}). For every atom in the Fukui function, the condensed version provides a local index, which may be used to comprehend atomic features that associated with electron donation and withdrawal. This is made up of three separate signs which is explained by equation (3) and (4-5):

$$e_B^+ = R_B(M+1) - R_B(M), \quad (3)$$

$$e_B^- = R_B(M) - R_B(M-1) \quad (4)$$

$$e_B^0 = \tfrac{1}{2}\{R_B(M+1) - R_B(M-1)\}$$

Atomic charge (R_B) and total number of electrons (M) in a molecule (B) are used in this context. The study of natural bond orbitals accomplished qualitative analysis. In addition to the first and second neighbor atomic charges, this set of descriptors contains the highest and lowest values of the condensed Fukui functions, as given by Equation (6-12)

$$\max \Delta_B^m R = \max\{R_B - R_A\}(A\epsilon mthneighboredatomfromB), \quad (6)$$

$$\max \Delta_B^m R = \min\{R_B - R_A\}(A\epsilon mthneighboredatomfromB), \quad (7)$$

$$\max \Delta_B^m e^W = \max\{R_B^W - R_B^W\}(A\epsilon mthneighboredatomfromB), \quad (8)$$

$$\max \Delta_B^m e^W = \min\{R_B^W - R_B^W\}(A\epsilon mthneighboredatomfromB) \quad (9)$$

To the extent that W is positive, negative, or zero in equations (6) through (9). $c^{atm}$was based on the nuclear magnetic shielding constant, and the steric factor both provide details on the three-dimensional molecule structuresc^{atm}. To get the steric factor, one uses the formula

$$T_B = \sum_A q_{WB} \exp(-c_{AB}), \quad (10)$$

Where $c_{AB}$is the distance between B and A and q_{WB} is the Vander radius for atom$A$. c^{atm}issymbolized by

$$c^{atm} = (Y, R_{atom}, e^{+}, e^{-}, e^{0}, max\Delta^{1}R_{atom}, \dots.) \quad (11)$$

The highest occupied MO, HOMO-1, and HOMO-2, and the lowest unoccupied MO, LUMO+2 andLUMO+1, forc^{orb}, is calculated using orbital energies $(\varepsilon)$and MO coefficients. While an atom's orbital energy doesn't affect the orbital's value, it's expected to be a big deal when defining the electron donation or acceptance features of various compounds. The quantity of electrons for eache^{0}, was extracted from the Mulliken population research. The direct manifestation of c^{orb} in equation (12)

$$c^{orb} = (\epsilon_{HOMO,}D_{1THOMO,}R_{1T}, D_{2THOMO,}R_{2T}, \dots.) \quad (12)$$

The atomic number, a discrete quantity was excluded from the learning system to maintain high prediction accuracy. The sparse descriptors, which include null values for 99% of AO data points were eliminated. Atomic number-related descriptors were removed during atomic pair data production due to the limited possible element combinations.

2.3. Bidirectional long short-term memory (Bi-LSTM)

According to the paragraph on Bi-LSTM networks, to apply them to biological processes is to make use of Bi-LSTM's strengths in dealing with problems like fading gradients and long-term dependencies. As a result of its adaptive multiplicative gates, the input gate (IT), the output gate (OT), and the forget gate (FT), the Bi-LSTM unit is able to control its memory state selectively. This specific feature allows the network to select which data to remove and save, which facilitates handling of data that is necessary for modeling biological-chemical interactions. Fig 2 depicts the structure of Bi-LSTM. The bidirectional Bi-LSTM boosts the network's capability to collect information from previous and succeeding levels with its mix of forward and backward hidden layers. For applications involving chemical interactions in the human body, this bidirectional technique works better than traditional Bi-LSTM in sequential modeling. The decision-making process of the Bi-LSTM is guided by a sigmoid activation function, which given a value between -1 and +1; this line with the study's emphasis on chemical processes. The network's ability to process and store information pertinent to the intricate dynamics of chemical events inside the human body is highlighted by this activation function, which helps in deciding that cell state data to erase in Equation (13-15).

$$J_s = \sigma(X_s W_s + Q_j G_{s-1} + a_j) \quad (13)$$

$$E_s = \sigma(X_e W_e + Q_e G_{s-1} + a_e) \quad (14)$$

$$P_t = \sigma(X_p W_s + Q_p G_{s-1} + a_p) \quad (15)$$

Where $a_j, a_e,$ and$a_p$represent the bias gates, $X_j,$ and $X_e W_e$ stand for the input weights, and$Q_j, Q_e,$ and Q_p represent the recurrent weights. $E_s$represents the sigmoid activation function, the previous block output isG_{s-1}, and the current input isY_s.The computed modulated new memory W_s is calculated as in Equation (16):

$$Y_s = \tanh(X_s W_s + Q_s G_{s-1} + a_s) \quad (16)$$

Where the hyperbolic tangent function is denoted by $tanh(\cdot)$, and $W_s$and Q_s stand for the input weight and recurring weight, respectively.

N_s, the present memory cell, is calculated as in Equation (17):

$$N_s = J_s \odot Y_s + E_s \odot G_{s-1}) \quad (17)$$

Where $\odot$ denotes the action of multiplying elements one by one, and N_{s-1}represents the value of the memory cell.As the LSTM unit's output, we get the hidden stateP_s, which is represented in equation (18):

$$G_s = P_s \odot \tanh(N_s) \quad (18)$$

Lastly, the Bi-LSTM networksof the output signal is applied to the classification module to complete the classification task.

2.4. Upgraded salp swarm optimization (USSO)

The USSO was developed to address various optimization problems. Salps, which are a part of the Salpidae family, resemble jellyfish in weight, movement, and tissues. They engage in a swarm activity called salp chain, they benefit from their ability to use rapid harmonic changes for mobility and eating. The USSO mathematical model was based on this study to assess salp chains in optimization problems. The first step in USSO is to create a leader and a follower from half of the population. There is a specific function for each salp at the front of the chain, and the people following behind them are known as followers. The search space size and the number of variables in the issue are both represented by n, and we locate the salps in an n-dimensional space. The salps in this swarm are looking for food. With this formula, we can update the sales leader.

$$w_s^1 = \begin{cases} E_i + d_1\big((ub_i - lb_i) \times d_2 + lb_i\big)d_3 \le 0 \\ E_i - d_1\big((ub_i - lb_i) \times d_2 + lb_i\big)d_3 > 0 \end{cases} \quad (19)$$

With $E_i$standing for the food supply and ub_i and lb_i denoting the upper and lower boundaries, w_s^1represents the function of the ruler in the ith dimension. The search space is maintained by creating d_2 and d_3 from the interval$[0, 1]$. Since d_1 is the most critical coefficient in this algorithm and that is essential to maintain the equilibrium between the program's exploration and exploitation stages in Equation (20-21).

$$d_1 = 2f^{-\left(\frac{4s}{S_{max}}\right)^2} \quad (20)$$

The maximum number of iterations is S_{max}, and S. represents the current iteration. After the leader's position is changed, the SSA uses the following equation to update the locations of the followers:

$$w_i^j = \frac{1}{2}\left(w_i^j + w_i^{j-1}\right) \quad (21)$$

When i is more than 1, the i^{th} follower position in the j^{th} dimension is denoted by w_i^j in Algorithm 1, which sets out the latter phases of the SSA. Algorithm 1 represents the steps of USSO.

Algorithm 1: USSO

Initialize bpopulation w.
Repeat
Determine the goal functional for every possible outcome w_j
Revise the optimal salp $(E = W^{a})$
Updated 1 *using Eq. (20)*
for $j = 1$*: M* ***do***
if $j == 1$ ***then***
modernize the position of USSO
else
Utilize to modify a location Eq. (21)
End if
End for
pending $(s < s_{max})$
return, E

3. Evaluation Metrics

Our suggested method USSO-Bi-LSTM, an upgraded slap swarm optimization model, is designed to forecast chemical reactions and synthesis. The python tool is used to simulate the result. The findings indicate that our proposed USSO-Bi-LSTM technique performs better than current methods, such as Siamese networks [15], LSTM [15], and ANN [14], with a significant advantage of these criteria. We used assessment criteria including RMSE (1.112), Accuracy (98.89), and TNR (90.3) in our research.

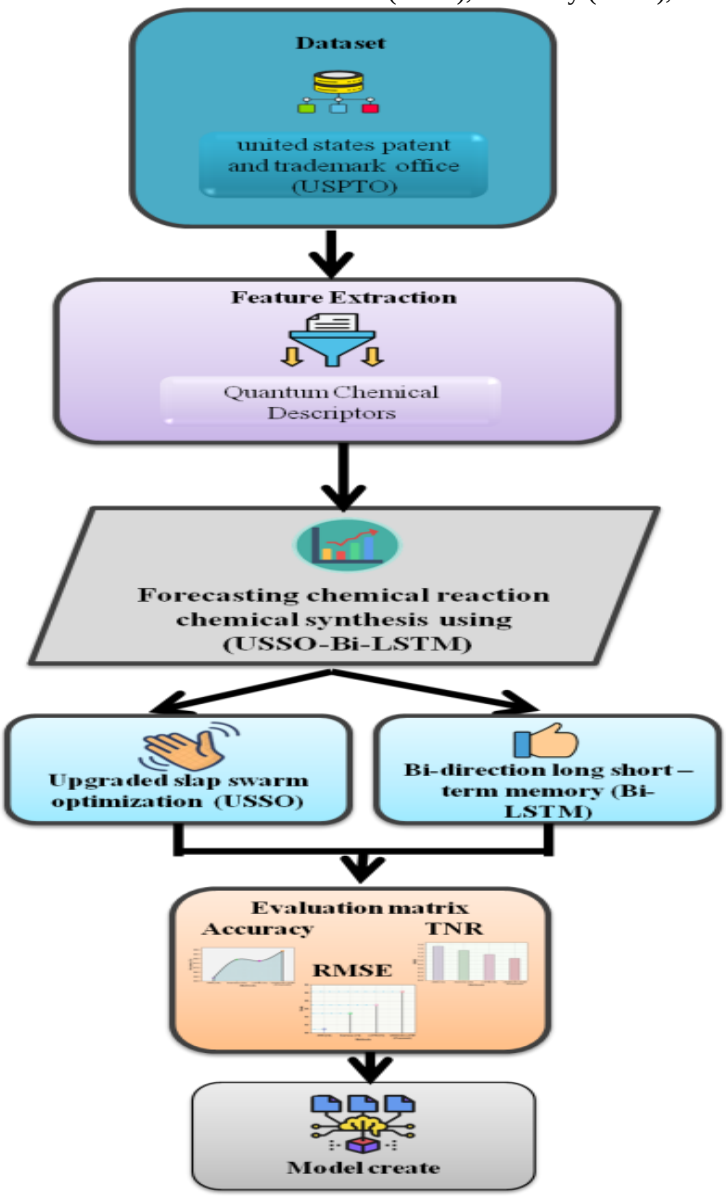

Figure 1. Flow of our proposed method (Source: Author)

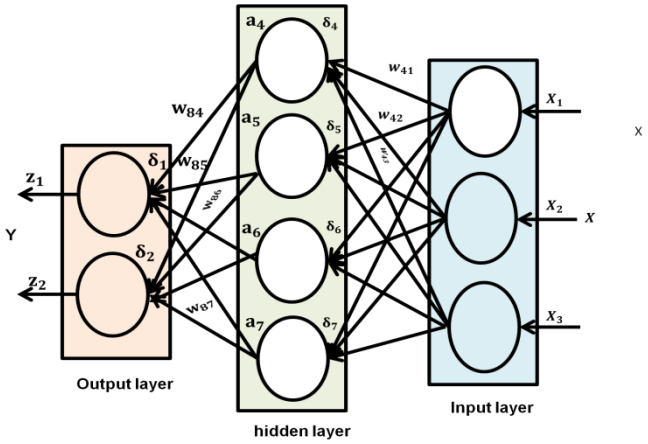

Figure 2. Structure of Bi-LSTM(Source: Author)

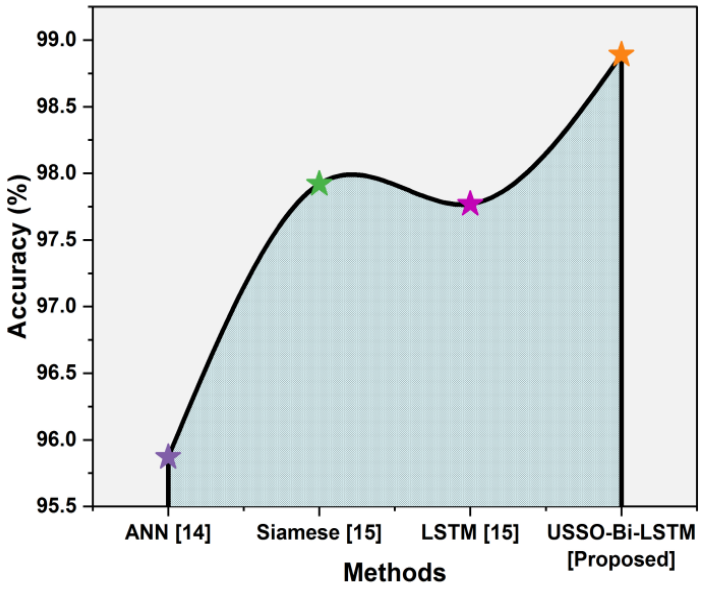

Figure 3. Graphical outcomes for accuracy (Source: Author)

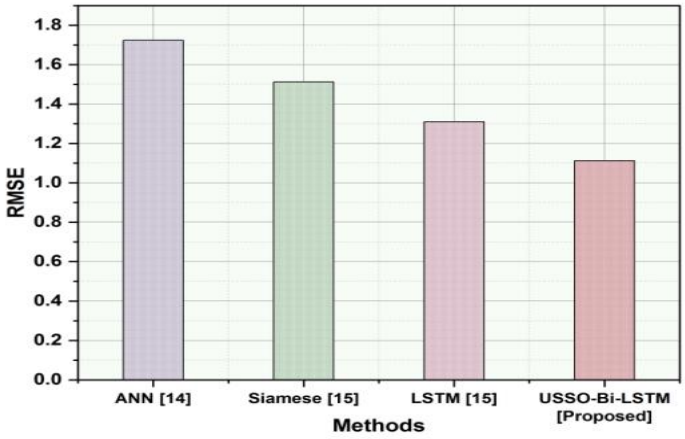

Figure 4. Graphical outcomes for RMSE (Source: Author)

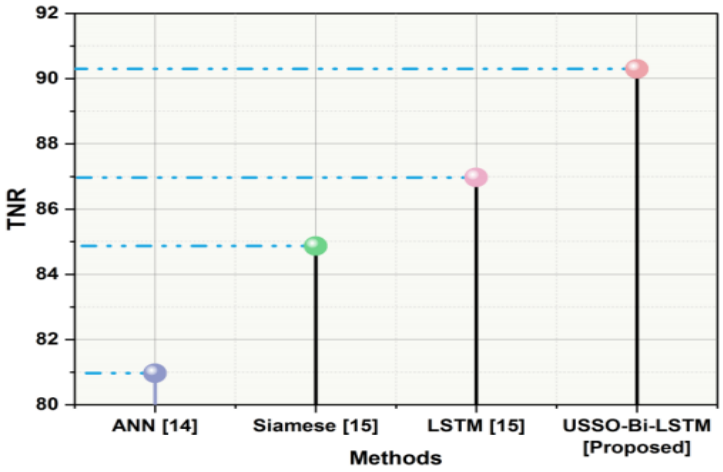

Figure 5. Graphical outcomes for TNR (Source: Author)

Table 1. Numerical Outcomes of Accuracy (Source: Author)

Methods	Accuracy (%)
ANN [14]	95.87
Siamese [15]	97.92
LSTM [15]	97.77
USSO-Bi-LSTM [Proposed]	98.89

Table 2. Numerical outcomes of RMSE (Source: Author)

Methods	RMSE
ANN [14]	1.723
Siamese [15]	1.512
LSTM [15]	1.31
USSO-Bi-LSTM [Proposed]	1.112

Table 3. Numerical outcomes of TNR (Source: Author)

Methods	TNR
ANN [14]	80.97
Siamese [15]	84.87
LSTM [15]	86.97
USSO-Bi-LSTM [Proposed]	90.3

3.1. Accuracy

Accuracy refers to the precision and dependability of a model in predicting, explaining, or interpreting complex biochemical interactions in the human body. It is crucial for understanding chemical events like food metabolism and neurotransmitter release, leading to improvements in pharmacology and tailored treatments. Computational models, including algorithms and machine learning techniques, are used for precision.Fig 3 shows that the graphical results for accuracy, Table 1 displays the numerical results of the accuracy, and our suggested USSO-Bi-LSTM approach (98.89) beats state-of-the-art algorithms like ANN (95.87), LSTM (97.77%), and Siamese networks (97.92%).

3.2. Root Mean Square Error(RMSE)

Data analysts' professionals utilize RMSE to evaluate the efficacy of their prediction models. Other metrics, such as enzyme kinetics, reaction rates, and concentration changes, are more relevant; it must portray the intricacies of biochemical interactions in live creatures. The RMSE data are shown in Fig 4. Compared to the current Siamese networks (1.512), LSTM (1.31), and ANN (1.723) approaches, our proposed USSO-Bi-LSTM (1.112) approach obtains better RMSE. Table 2 shows the results of the RMSE.

3.3. True Negative Rate (TNR)

The TNR is a statistical measure that has no direct relevance to the chemical processes occurring in the human body, rather it is utilized in ML for problems involving binary categorization. Indicating the model's ability to identify non-existent responses, it evaluates the model's sensitivity to chemical states or interactions. The visual outcomes of the TNR is shown in Fig 5. To beat the existing ANN (80.97), LSTM (86.97) and Siamese networks (84.87), we provide the USSO-Bi-LSTM (90.3) approach, and the numerical findings of TNR is shown in Table 3.

4. Conclusion

The field of chemistry has seen a fundamental shift with the use of ML to chemical synthesis and reaction predictions. ML algorithms have significantly sped research and development by analyzing enormous datasets, identifying trends, and forecasting reaction times. This innovation opens up new possibilities for discovering new reactions and modifying reaction parameters in addition to increase the efficiency of chemical synthesis. The combination of human expertise and ML capabilities could lead to creative and sustainable chemical synthesis. As we move into the era of intelligent chemistry, one powerful tool that can help to understand chemical interactions better is the application of ML. With notable advantages in evaluation metrics including TNR (90.3), Accuracy (98.89), and RMSE(1.112), the suggested USSO-Bi-LSTM strategy beats out current approaches of Siamese networks, LSTM, and ANN, exhibiting higher performance in a number of domains. The development of advanced ML models that can handle an intricate reaction mechanisms that integrate to multiple chemical data sources, and allow real-time prediction for synthesis processes that are accelerated and optimized holds the key to the future potential for chemical reaction and synthesis prediction.

References

[1] S. Khan, M. Falahati, W.C. Cho, Vahdani, Y. Siddique, R. Sharifi, M. Jaragh-Alhadad, L.A. Haghighat, S. Zhang, X. Ten, T.L. Hagen, Q. Bai. (2023). Core-shell inorganic NPs@ MOF nanostructures for targeted drug delivery and multimodal imaging-guided combination tumor treatment. Advances in Colloid and Interface Science. 103007.

[2] A. Rayhan. (2023). Accelerating Drug Discovery and Material Design: Unleashing AI's Potential for Optimizing Molecular Structures and Properties.30888.

[3] R.O. Kareem, N. Bulut, O. Kaygili. (2024). Hydroxyapatite Biomaterials: A Comprehensive Review of their Properties, Structures, Medical Applications, and Fabrication Methods. Journal of Chemical Reviews. 6(1): 1-26.

[4] P.N. Shiammala, N.K.B. Duraimutharasan, B.V. aseeharan, A.S. Alothaim, E.S. Al-Malki, B. Snekaa, S.Z. Safi, S.K. Singh, D. Velmurugan, C. Selvaraj. (2023). Exploring the Artificial Intelligence and Machine Learning Models in the Context of Drug Design Difficulties and Future Potential for the Pharmaceutical Sectors. Methods.

[5] G.S. Priyanga, G. Pransu, H. Krishna, T. Thomas. (2023). Discovery of Novel Photocatalysts Using Machine Learning Approach. In Machine Learning for Advanced Functional Materials. Singapore: Springer Nature Singapore. 233-261.

[6] A. Arun, Z. Guo, S. Sung, A.A. Lapkin. (2023). Reaction impurity prediction using a data mining approach. Chemistry-Methods. 202200062.

[7] K. Chen, G. Chen, J. Li, Y. Huang, E. Wang, T. Hou, P.A. Heng. (2023). MetaRF: attention-based random forest for reaction yield prediction with a few trails. Journal of Cheminformatics, 15(1): 1-12.

[8] H. Yang, X. Bai, B. Feng, Q. Wang, L. Meng, F. Wang, Y. Wang. (2023). Application of Molecular Transformer approach for predicting the potential reactions to generate advanced glycation end products in infant formula. Food Chemistry. 407: 135143.

[9] W. Zhong, Z. Yang, C.Y.C. Chen. (2023). Retrosynthesis prediction using an end-to-end graph generative architecture for molecular graph editing. Nature Communications. 14(1): 3009.

[10] M. Wen, E.W.C. Spotte-Smith, S.M. Blau, M.J. McDermott, A.S. Krishnapriyan, K.A. Persson. (2023). Chemical reaction networks and opportunities for machine learning. Nature Computational Science. 3(1): 12-24.

[11] A.M. Bran, S. Cox, A.D. White, P.S. chwaller. (2023). ChemCrow: Augmenting Large-language Models with Chemistry Tools. 2304: 05376.

[12] J. Yu, C. Zhang, Y. Cheng, Y.F. Yang, Y.B. She, F. Liu, W. Su, A.Su. (2023). SolvBERT for solvation free energy and solubility prediction: a demonstration of an NLP model for predicting the properties of molecular complexes. Digital Discovery. 2(2): 409-421.

[13] P. Neves, K. McClure, J. Verhoeven, N. Dyubankova, R. Nugmanov, A. Gedich, S. Menon,

Z. Shi, J.K. Wegner. (2023). Global reactivity models are impactful in industrial synthesis applications. Journal of Cheminformatics. 15(1): 1-11.

[14] C. Chi, G. Janiga, D. Thévenin. (2021). On-the-fly artificial neural network for chemical kinetics in direct numerical simulations of premixed combustion. Combustion and Flame. 226: 467-477.

[15] S. Jiang, Z. Zhang, H. Zhao, J. Li, Y. Yang, B.L. Lu, N. Xia. (2021). When SMILES smiles, practicality judgment and yield prediction of chemical reaction via deep chemical language processing. IEEE Access. 9:.85071-85083.