



# Deep Learning for Green Chemistry: An AI-Enabled Pathway for Biodegradability Prediction and Organic Material Discovery

Dela Quarme Gbadago<sup>1,2</sup> · Gyuyeong Hwang<sup>1</sup> · Kihwan Lee<sup>3</sup> · Sungwon Hwang<sup>1,3</sup> 

Received: 29 February 2024 / Revised: 22 May 2024 / Accepted: 29 May 2024 / Published online: 12 June 2024  
© The Author(s), under exclusive licence to Korean Institute of Chemical Engineers, Seoul, Korea 2024

## Abstract

The increasing global demand for eco-friendly products is driving innovation in sustainable chemical synthesis, particularly the development of biodegradable substances. Herein, a novel method utilizing artificial intelligence (AI) to predict the biodegradability of organic compounds is presented, overcoming the limitations of traditional prediction methods that rely on laborious and costly density functional theory (DFT) calculations. We propose leveraging readily available molecular formulas and structures represented by simplified molecular-input line-entry system (SMILES) notation and molecular images to develop an effective AI-based prediction model using state-of-the-art machine learning techniques, including deep convolutional neural networks (CNN) and long-short term memory (LSTM) learning algorithms, capable of extracting meaningful molecular features and spatiotemporal relationships. The model is further enhanced with reinforcement learning (RL) to better predict and discover new biodegradable materials by rewarding the system for identifying unique and biodegradable compounds. The combined CNN-LSTM model achieved an 87.2% prediction accuracy, outperforming CNN- (75.4%) and LSTM-only (79.3%) models. The RL-assisted generator model produced approximately 60% valid SMILES structures, with over 80% being unique to the training dataset, demonstrating the model's capability to generate novel compounds with potential for practical application in sustainable chemistry. The model was extended to develop novel electrolytes with desired molecular weight distribution.

**Keywords** Biodegradability · SMILES · Green chemistry

## Introduction

Throughout scientific development, humanity has produced an abundance of organic compounds, many of which are utilized once and then discarded. The yearly production of plastic has reached an astonishing 450 million tons, with 340 million tons being generated as waste [1]. Regrettably, these organic compounds exhibit remarkable resistance to natural decomposition, leading to their persistence in the environment and posing significant threats to human well-being and ecosystems [2]. Consequently, assessing the biodegradability

of organic compounds has been increasingly regarded as crucial in recent times. Following the European Registration, Evaluation, Authorization, and Restriction of Chemicals (REACH) regulation, companies engaged in the manufacturing or importing of chemicals exceeding 1 ton per year are mandated to provide detailed information regarding the biodegradability of their compounds [3]. To evaluate biodegradability, standardized test methods published by prestigious organizations such as the Organization for Economic Co-operation and Development (OECD) [4] and Japan's Ministry of International Trade and Industry (MITI) [5] are primarily employed. In addition to assessing the biodegradability of existing compounds, the significance of discovering novel biodegradable organic compounds is also growing. However, searching for potential candidates within the entire compound space is nearly impossible due to its vast scale, estimated to range from  $10^{23}$  to  $10^{60}$ . Predicting new molecules through calculations, synthesizing them, and testing their physical properties is time-consuming. As a result, only

✉ Sungwon Hwang  
sungwon.hwang@inha.ac.kr

<sup>1</sup> Graduate School of Chemistry and Chemical Engineering, Inha University, Incheon 22212, Republic of Korea

<sup>2</sup> 3D Convergence Center, Inha University, Incheon 22212, Republic of Korea

<sup>3</sup> Department of Smart Digital Engineering, Inha University, Incheon 22212, Republic of Korea

approximately  $10^8$  compounds have been synthesized thus far [6].

Utilizing generative models for discovering new molecules alleviates these challenges. Unlike conventional methods, generative models operate through inverse modeling. This means that new molecules are generated based on desired properties, offering a more efficient approach to exploration. Different methods have been devised to enable the incorporation of complex molecular structures into neural networks. One prevalent approach is using a Simplified Molecular Input Line Entry System (SMILES) [7], which converts molecules into a one-dimensional text array following a specific set of rules. Due to its effectiveness, SMILES is widely employed in many molecular generation models. Recently, the use of generative models for chemical substance discovery has been actively researched [8]. Early generative models were developed by combining recurrent neural networks (RNNs) and reinforcement learning [9]. However, to overcome the limitations of these models, various types of generative models have been developed. Chiu et al. [10] proposed a method for predicting the hydrolysis rate by utilizing not only the SMILES representation but also the partial charge of the molecule as inputs to the autoencoder. Wang et al. [11] addressed the challenge of balancing desirable properties and novelty in molecular design. They developed a model that interprets the ligand-receptor structure by taking the molecular 3D structure as an input. Arús-Pous et al. [12] divided the existing dataset into subsets with desired molecular scaffolds to devise a strategy to create molecules with specific characteristics without using reinforcement learning. Cao et al. [13] conducted research on avoiding the computationally expensive likelihood-matching process. They used generative adversarial networks (GANs) with graphs as inputs. Tang et al. [14] employed a Support Vector Machine (SVM) classifier to enhance the prediction accuracy and overcome the limitations of linear regression when predicting the biodegradability of large molecules. Dollar et al. [15] attempted to introduce the attention mechanism, commonly used in translation tasks, into variational autoencoders (VAE) for de novo molecular design. While several studies have been conducted in this area, there is a notable lack of research on generative models for discovering biodegradable organic compounds. The main challenge lies in training a model due to the severe insufficiency of the biodegradability database. In contrast to the readily available abundance of information, such as LogP, which can be easily accessed through methods like RDKit, the resources for biodegradability data remain scarce. As a response to this issue, a study was carried out by Lunghini et al. [16] to construct a substantial database by integrating various biodegradability data. Additionally, given the complex mechanisms determining the biodegradation rate, numerous

models employing the Quantitative Structure–Activity Relationship (QSAR) method are being explored to classify compounds into biodegradable and non-biodegradable substances [17, 18]. However, these models are imperfect, mainly due to their limited applicability scope.

Furthermore, like the previous examples, much research has focused on enhancing prediction performance by altering the generative model. However, a limited body of research is dedicated to improving the prediction model. Particularly in the case of biodegradability, accessing sufficient databases for training remains challenging, and a well-defined mathematical and quantitative method for determining the biodegradability of newly synthesized molecules has yet to be established. Given these constraints, a viable approach for biodegradability prediction involves enabling the neural network to learn molecular features.

Recent progress in deep learning have led to advanced approaches that effectively combine long short-term memory (LSTM) networks and convolutional neural networks (CNNs) networks to enhance the analysis of spatiotemporal data. For instance, Barros et al. developed a hybrid CNN-LSTM model specifically for the classification of lung ultrasound videos in COVID-19 diagnosis, leveraging CNNs for extracting spatial features and LSTMs for capturing temporal dependencies, demonstrating high efficiency in handling the spatiotemporal dynamics akin to those found in chemical compound analysis via SMILES representations and molecular images [19]. Similarly, Dang et al. [20] explored the potential of a hybrid 1D-CNN-LSTM architecture in structural health monitoring. Their research underscores the model's capability to integrate local and temporal feature extraction, making it particularly relevant for applications such as biodegradability prediction where both structural integrity and sequential reactions are pivotal. Parallel to the discussion on hybrid models, the debate between the utility of LSTM and Transformer models continues, especially in fields requiring the processing of sequential data. In the context of electronic trading, Bilokon and Qiu compared these models, finding that while Transformers excel in certain types of sequence prediction, LSTMs offer superior performance in more complex scenarios such as predicting price movements, highlighting their robustness and applicability in financial markets [21]. Further extending the capabilities of LSTMs, Tsunami and Taki introduced the Sequencer, a novel LSTM-based architecture for image classification that competes against the typically dominant Vision Transformers. Their results illustrate the LSTM's capability to effectively model long-range dependencies, thereby affirming its competitiveness in tasks traditionally reserved for Transformers [22]. Additionally, Zeyer et al. provided a comparative analysis of Transformer and LSTM models in automatic speech recognition, emphasizing that despite Transformers' faster

training and stability, they are prone to overfitting. This comparison stresses the necessity to tailor the choice of model to specific data characteristics and task requirements to optimize performance and ensure generalization [23].

Therefore, in this study, we introduce an integrated methodology that significantly advances the field of biodegradability prediction and material discovery. This innovative approach combines deep learning techniques, generative models, and reinforcement learning to address the complex task of efficiently identifying novel biodegradable organic compounds. Our research establishes a robust data preparation pipeline, utilizing SMILES notations for versatile compound representation and employing data augmentation techniques to enhance dataset diversity. The proposed prediction model adopts a hybrid architecture, leveraging LSTM networks and CNNs, effectively handling sequential data and spatial patterns to provide highly accurate biodegradability predictions. By adopting a stack-augmented RNN for molecular trajectory generation within a reinforcement learning framework, our generator model empowers the exploration of intricate chemical spaces, facilitating the discovery of environmentally friendly materials. Furthermore, our research incorporates a reward mechanism that quantifies the value of molecular structures based on biodegradability, thus ensuring the alignment of the learning process with environmentally conscious objectives. We also employ a systematic grid search for hyperparameter optimization, guaranteeing that model configurations are finely tuned for optimal predictive accuracy. Here, the LSTM and CNN models discussed in “[Methodology](#)” were initially optimized as standalone models using the grid search method described in “[Hyperparameter Optimization](#)” albeit with different parameter spaces. These model results were then used to validate the hybrid model in Fig. 4.

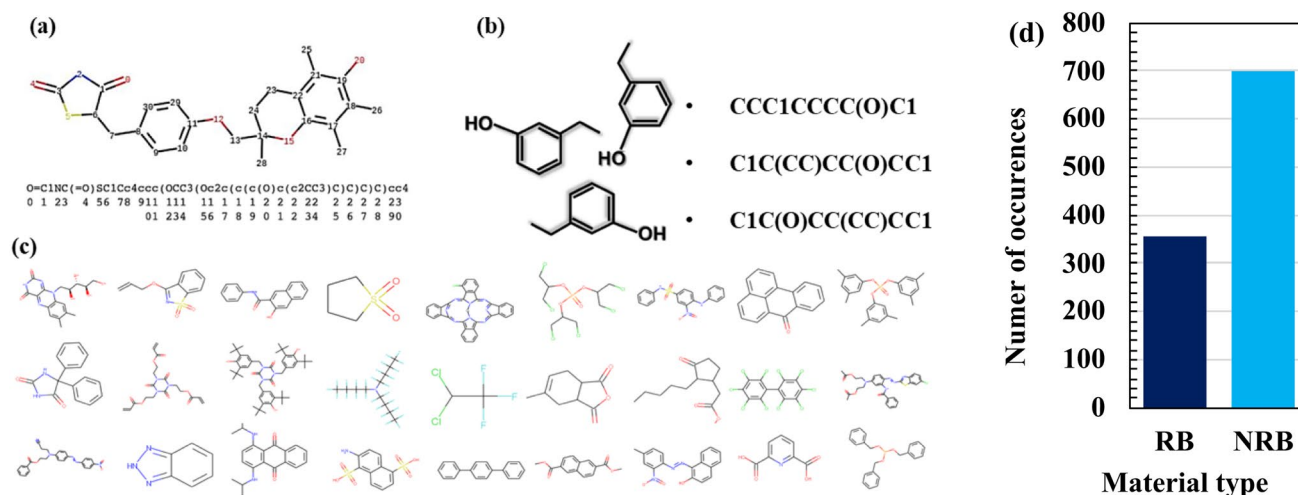
The rest of the study is structured as follows. “[Methodology](#)” describes the algorithms and procedures implemented in this work. “[Results and Discussion](#)” presents the simulation results, comparative analysis, and discussion of findings. The study is concluded in “[Conclusion](#)”, wherein an overview of the contributions of this study and its applications are presented.

## Methodology

This section comprehensively describes the solution strategies and algorithms adopted in executing the study. The data processing methods, prediction models, optimization steps, and generator models are discussed.

### Data Preparation and Processing

The rapid advancement of computing has opened new avenues for predicting and exploring the biodegradability of organic compounds. Existing methods often require laborious and computationally expensive DFT calculations, hindering their scalability and efficiency. This research aims to develop an AI-driven model that leverages molecular formulas and structures for efficient biodegradability prediction. To represent a large number of compounds, we employ simple and independent nomenclatures (SMILES) that are easy for computers to understand. These nomenclatures allow us to effectively encode and process the chemical structures of compounds in the AI models. The SMILES notation allows flexibility in representing molecules by specifying the connectivity of atoms through their bonds. Different starting atoms or bond connectivity result in distinct SMILES strings, enabling multiple valid representations for the same compound. The SMILES compounds are also converted into structural images for subsequent training. An example of compounds, their respective SMILES notations, and structural images is depicted in Fig. 1. A diverse dataset of 1055 organic compounds with known readily biodegradable (RB) materials (355 species) and non-readily biodegradable materials (700 species) are obtained from Kamel et al. [24]. A detailed description of the data is therefore available as referenced. The dataset was shuffled to ensure a random distribution, and subsequently divided into specific segments for training and validation. The SMILES strings were converted into canonical forms, ensuring a standardized representation of each chemical compound. Additionally, random permutations of atomic indices were generated to augment the dataset, providing diverse representations of the same chemical structures. A tokenization procedure was applied to the SMILES strings to separate them into individual atomic symbols and other special characters. The set of unique tokens obtained was mapped to corresponding indices, creating a consistent format for subsequent training. The length of the tokenized SMILES strings in the dataset was evaluated, and the maximum length was determined, allowing for the consistent handling of SMILES strings of varying lengths. The dataset was further processed to generate input–output pairs suitable for training LSTM networks, involving randomizing the SMILES strings and converting them into a tensor format. A conversion process was



**Fig. 1** **a** Representation of SMILES Tokenization, **b** different SMILES representations of 3-Ethylphenol, **c** samples of compounds used during the model training, and **d** distribution of materials in the dataset

implemented to transform characters or strings into corresponding tensor formats. This facilitated the handling of data within the deep learning framework.

Training the model with different SMILES representations and images of the same compound at each iteration can enhance the model's generalizability as the dataset increases. This approach allows the model to learn diverse representations of the same compound, capturing various aspects of its chemical structure and visual characteristics. The training process benefits from the increased variability in the data, enabling the model to better generalize and make accurate predictions on unseen compounds. This technique promotes robustness and adaptability in the model's learning process, ultimately improving its performance in biodegradability prediction and material discovery.

## Prediction Model Building

In this study, a hybrid approach leveraging two distinct deep learning architectures, namely LSTM networks and CNN, was developed to tackle the predictive task encompassing the analysis of chemical structures. LSTM networks are efficient at processing time series and textual data, which are essential in extracting features in organic compounds. They excel in recognizing long-term dependencies and patterns within sequential data, such as chemical structures and physical properties, which are crucial for predicting biodegradability. LSTM's ability to retain and utilize historical information allows for accurate biodegradability predictions by learning from molecular descriptors and their effects over time. CNNs are effective in biodegradability prediction by extracting features from image data of chemical structures. Training

on these structures and their biodegradability labels, CNNs identify local patterns and spatial relationships key to assessing biodegradation potential. Convolutional layers use filters to capture significant features at different scales, enabling CNNs to forecast the biodegradability of previously unseen compounds with enhanced precision.

The combined architecture synthesizes the inherent strengths of both LSTM and CNN models, facilitating the interpretation of complex patterns within data represented through both sequences and images. The LSTM component, constructed as a two-layer model accepting inputs of dimension 165 (representing the maximum length of token sequences that reflect the SMILES strings in the dataset), was employed for its ability to handle sequential data, reflecting the sequential nature of chemical information in SMILES strings. An embedding layer was incorporated with an optimized output dimension of 12, effectively reducing dimensionality and capturing semantic relationships, represented by:

$$e(x) = W_e \cdot x + b_e, \quad (1)$$

where  $x$  represents the input,  $W_e$  represents the embedding matrix, and  $b_e$  represents the bias.

The LSTM layer, consisting of 256 units, provides the network's memory function, capturing long-term dependencies and patterns over time, making it highly relevant for analyzing the chemical structure of organic compounds and their biodegradability. The layer can be mathematically represented as [25–27]:

$$\begin{aligned}
 f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\
 i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\
 o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\
 c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \\
 h_t &= o_t \odot \tanh(c_t)
 \end{aligned} \quad (2)$$

where  $f_t$ ,  $i_t$ , and  $o_t$  are the forget, input, and output gates,  $c_t$  is the cell state,  $h_t$  is the hidden state,  $\sigma$  is the sigmoid activation function, and  $\odot$  represents elementwise multiplication.

Subsequent layers included a dropout layer with a rate of 0.3, to prevent overfitting, and a dense layer with 35 units employing a hyperbolic tangent activation function and He normal initialization, enhancing the network's ability to capture non-linear relationships, all of which were obtained by grid search optimization as described in the subsequent section. The output layer is defined as:

$$y = \tanh(W_d \cdot h + b_d), \quad (3)$$

where  $y$  is the output vector or tensor,  $W_d$  is the weight matrix connecting the previous layer's outputs  $h$  to the current layer's inputs,  $h$  is the input vector or tensor from the previous layer, and  $b_d$  is the bias vector added to the weighted sum before applying the activation function.

Conversely, the CNN model was adopted for its effectiveness in analyzing spatial patterns within images, pertinent to the  $300 \times 300$  images with three channels used in this study. The model initiated with a Conv2D layer composed of 6 filters of size  $3 \times 3$  and strides of  $4 \times 4$ , represented as:

$$Y_{ij} = \sum_{m,n} X_{i+m,j+n} \cdot K_{mn} + b, \quad (4)$$

where  $Y_{ij}$  is the output feature map at position  $(i,j)$ ,  $X_{i+m,j+n}$  are input values at relative positions,  $K_{mn}$  are convolutional filter weights, and  $b$  is the bias term.

Followed by batch normalization and ReLU activation to accelerate training and introduce non-linearity:

$$Y_{\text{normalized}} = \frac{Y - \mu}{\sqrt{\sigma^2 + \epsilon}} \quad (5)$$

$$Y_{\text{ReLU}} = \max(0, Y_{\text{normalized}}).$$

Subsequent max-pooling layers reduced dimensionality and emphasized salient features, while the sequence concluded with a flattening step, a dropout layer with a rate of 0.3, and a dense layer of 50 neurons with ReLU activation [28, 29] and He normal initialization, further contributing to robust feature extraction. These hyperparameters were also obtained via the grid search optimization.

The outputs from the LSTM and CNN models were concatenated, capitalizing on their synergistic strengths, followed by two dense layers with 40 and 2 units, respectively. The latter employed a SoftMax activation function [30], enabling probabilistic interpretation of the model's predictions:

$$y = \text{SoftMax}(W \cdot [h_{\text{LSTM}}, h_{\text{CNN}}] + b). \quad (6)$$

The combined model was compiled with the Adam optimizer [31–33] at a learning rate of 0.0001 and categorical cross-entropy loss function [34, 35], optimizing for multi-class classification performance:

$$L = -\frac{1}{N} \sum_{i=1}^N y_i \log(\hat{y}_i), \quad (7)$$

where  $y_i$  denotes the true label,  $\hat{y}_i$  denotes the predicted label for each sample in the batch of size  $N$ .

Model parameters were saved and loaded from the disk, enhancing reproducibility, and allowing for further utility. For training, an iterative tokenization procedure was applied to the training and validation datasets across a sequence of times, aligning with the sequential nature of the data. The combined model was fit for 2000 epochs with a batch size of 10, balancing the trade-off between computational efficiency and convergence stability. Following training, the model underwent evaluation on a test dataset, and various functionalities were deployed, including saving, loading best models, and executing predictions with the optimally performing model. Additionally, a series of utility functions were employed to perform crucial tasks such as validation of SMILES strings, generation, and canonical conversion of specific SMILES strings, pairwise similarity computation, prediction using generated SMILES strings, simple moving average calculation, reward calculation, and similarity and canonical checks on generated strings. These functions not only enriched the model's interpretive capability but also facilitated a more nuanced assessment and interpretation of predictions concerning biodegradable and non-biodegradable chemical structures. Collectively, the integrated methodology provided a robust framework for predictive analysis, merging sequence understanding with spatial pattern recognition and supporting comprehensive validation and interpretive analysis.

## Generator Model Building

Developing novel biodegradable materials is crucial in modern materials science, contributing to sustainable development and environmental protection. In this research, a methodology is constructed leveraging reinforcement

learning (RL), uniquely suited to this task due to its ability to explore and optimize complex, high-dimensional spaces. The RL model consists of three primary components: the generator, predictor, and reward function, each with distinct implications. (1) **Generator**: utilized for generating molecular trajectories, the generator, adopted from Popova et al. [36, 37] is the core of the explorative aspect of the RL framework. It symbolizes the ability to propose new molecular structures in the search space, allowing the discovery of potentially novel biodegradable materials. The generator model is a stack-augmented RNN developed using PyTorch. It consists of an Embedding layer to translate the input  $x$  into continuous space,  $e(x)$ , facilitating the nuanced processing of molecular structures and understanding complex relations within the molecules. The gated recurrent unit (GRU) is employed, whose update and reset gates are governed by:

$$\begin{aligned} r_t &= \sigma(W_r \cdot [h_t - 1, x_t] + b_r) \\ z_t &= \sigma(W_z \cdot [h_t - 1, x_t] + b_z), \end{aligned} \quad (8)$$

and its hidden state by:

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t, \quad (9)$$

where  $\sigma$  is the sigmoid function, to enhance the handling of sequential data and SMILES representations, vital for capturing temporal dependencies in molecular design.  $r_t$  and  $z_t$  denote the reset and update gates at time  $t$ ,  $h_{t-1}$  is the previous hidden state,  $x_t$  is the current input,  $W_r$ ,  $W_z$ ,  $b_r$ , and  $b_z$  are the weight matrices and bias terms.

An innovative feature of this model is the stack augmentation mechanism, which is central to generating diverse and complex molecular trajectories. The stack operation equations, governed by push, pop, and no-op controls, enable flexible and intelligent manipulation of the stack structure. The decoder, coupled with LogSoftmax activation, translates the GRU's output and ensures normalization, fundamental for accurate prediction and selection of the next molecular character.

$$y_t = \text{LogSoftmax}(W_o \cdot h_t + b_o), \quad (10)$$

where  $W_o$  and  $b_o$  are the weight matrix and bias term of the output layer, and  $y_t$  is the predicted output at time  $t$ .

The training and evaluation functions encapsulate the learning process, which is essential for adapting the model to generate desired molecular structures. The loss is computed using the Cross-Entropy Loss function. Additionally, various utilities, including changing the learning rate and handling stack operations, enhance the flexibility and efficiency of the model. (2) **Predictor**: previously described in “[Prediction Model Building](#)”, this component evaluates the generated trajectories, functioning as the evaluative

mechanism within the RL environment. It serves as the scientific bridge between the mathematical formulations of RL and the physical properties of molecules, providing tangible feedback based on generated molecular structures. (3) **Reward Function**: computing the reward based on the generated sequence of molecular structures, the reward function plays a critical role in guiding the learning process. By quantifying the value of each structure in terms of biodegradability, ensures that the learning process aligns with the ultimate scientific goal of the research. Herein, a high reward is assigned if the generated material is not in the training data and is biodegradable. This allows the weights for the newly generated model to be updated. The reward is expressed as  $R(s, a)$ , where  $s$  denotes the state, and  $a$  denotes the action taken.

The policy gradient method is applied, vital for continuous, high-dimensional action spaces common in molecular design. This method maximizes the expected cumulative reward [9, 36, 38], emphasizing the trajectories that lead to the most promising materials, according to the following equation:

$$\nabla_{\theta} J(\theta) = E_{\pi_{\theta}} [\nabla_{\theta} \log \pi_{\theta}(a|s) Q^{\pi}(s, a)], \quad (11)$$

where  $\theta$  is the policy parameter,  $\pi$  represents the policy, and  $Q_{\pi}$  is the action-value function.

Using gradient clipping ensures stable and robust convergence by avoiding the exploding gradient problem. The clipped gradient [38] can be represented as:

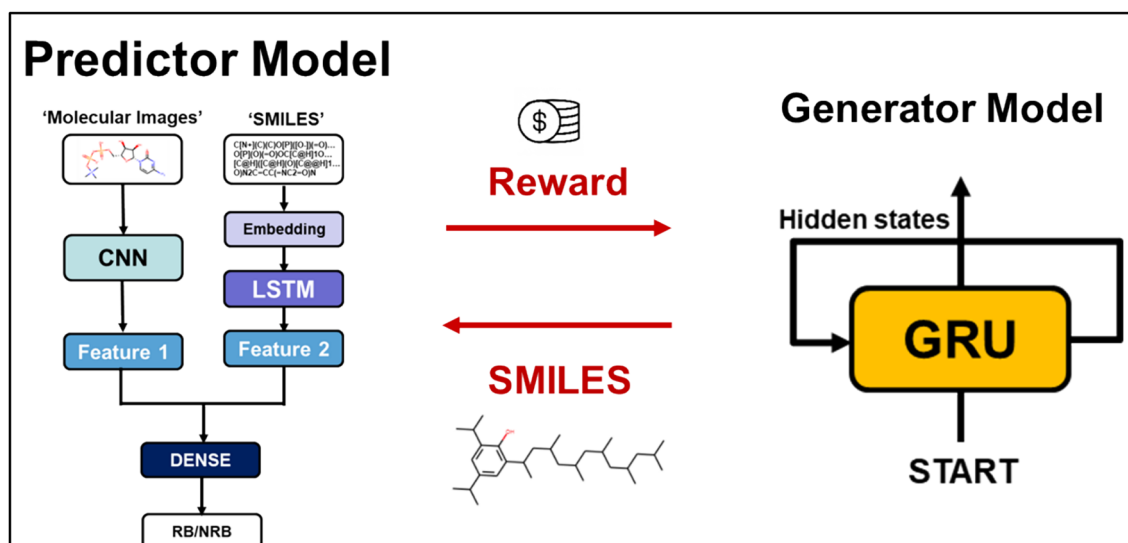
$$\nabla_{\text{clipped}} = \min \left( \nabla, \frac{\nabla}{\|\nabla\|} \times \text{threshold} \right). \quad (12)$$

The iterative process, involving policy replay and updates, illustrates RL's dynamism. The update rule can be expressed using the Bellman equation [39–41]:

$$Q(s, a) \leftarrow (1 - \alpha)Q(s, a) + \alpha(r + \gamma \max_{a'} Q(s', a')), \quad (13)$$

where  $\alpha$  is the learning rate, and  $\gamma$  is the discount factor.

Furthermore, evaluating the generated SMILES strings' validity and canonicity ensures that the generated molecular structures are not only novel but also chemically accurate and practically feasible. Lastly, converting valid canonical SMILES into canonical form, and the subsequent visualization, encapsulates the synthesis of theoretical findings with practical applications, bridging computational discoveries with real-world chemical representations. The overall schematic representation of the proposed solution strategy is presented in Fig. 2. It is worth noting that the LSTM and CNN models described herein were initially optimized as standalone models using the grid search method described in “[Hyperparameter Optimization](#)” albeit with different



**Fig. 2** Graphical representation of the proposed novel AI model for material discovery and prediction

parameter spaces as specified in the descriptions. These model results were then used to validate the hybrid model in Fig. 4.

### Hyperparameter Optimization

In this study, we employed grid search to optimize the hyperparameters for our CNN, LSTM, and hybrid CNN–LSTM models, prioritizing its well-reported efficiency in scenarios with well-defined and constrained parameter spaces. Grid search is widely recognized for its simplicity and effectiveness, making it a robust choice [42], thereby eliminating the possibility of obtaining sub-optimal models generally obtained via the conventional trial-and-error approach to model finetuning [28, 29]. Numerous algorithms, such as GA [28] and Bayesian [43] as well as derivative-free optimization algorithms, such as NOMAD and DIRECT [44] highlight a spectrum of robust

methods available for hyperparameter tuning. Although these algorithms are noted for their performance in several scenarios, our selection of grid search was driven by its adequacy for our specific application needs and computational limitations. Recent empirical evaluations, such as those by Alibrahim and Ludwig [45], have demonstrated that grid search remains competitive with more complex methods like Bayesian optimization and GAs, particularly when computational resources are a consideration. These authors found that the performance of grid search effectively bridged the gap between these advanced techniques, affirming its suitability for our research needs. Additionally, the study by Ogunsanya et al. [46] has revealed the application of grid search in additive manufacturing processes, further proving its versatility and relevance across different fields. This underscores the method's practical applicability and relevance in diverse research scenarios. Thus, the grid search methodology represents a fundamental algorithmic approach for hyperparameter tuning [47]. Herein, we partition the domain of the hyperparameters

**Table 1** Parameters used in the hyperparameter optimization

Type	Range or candidates
Number of epochs <sup>†</sup>	1–10,000
Number of neurons in the first layer <sup>†</sup>	1–500
Number of neurons in the second layer <sup>†</sup>	1–100
Activation functions <sup>†</sup>	Sigmoid, Hyperbolic tangent function, ReLU, leaky ReLU, ELU, and SELU
Learning rates <sup>†</sup>	0.1, 0.05, 0.04, 0.03, 0.02, 0.01, 0.007, 0.005, 0.003, and 0.001
Loss functions <sup>†</sup>	MSE and MAE
Optimizer <sup>†</sup>	AdaGrad, Adam, AdaMax, and Nadam

<sup>†</sup> [50, 51], <sup>‡</sup>[28, 52, 53]

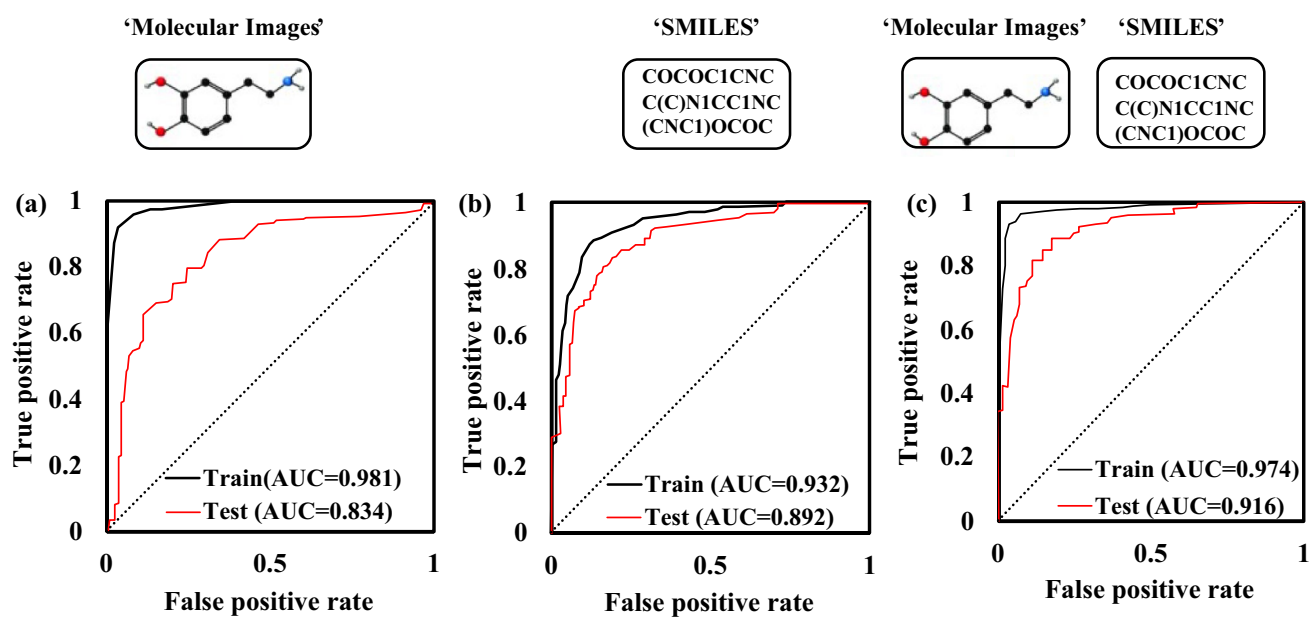


Fig. 3 Comparison plots between a the CNN-only model, b the LSTM-only model, and c the LSTM-CNN integrated model results

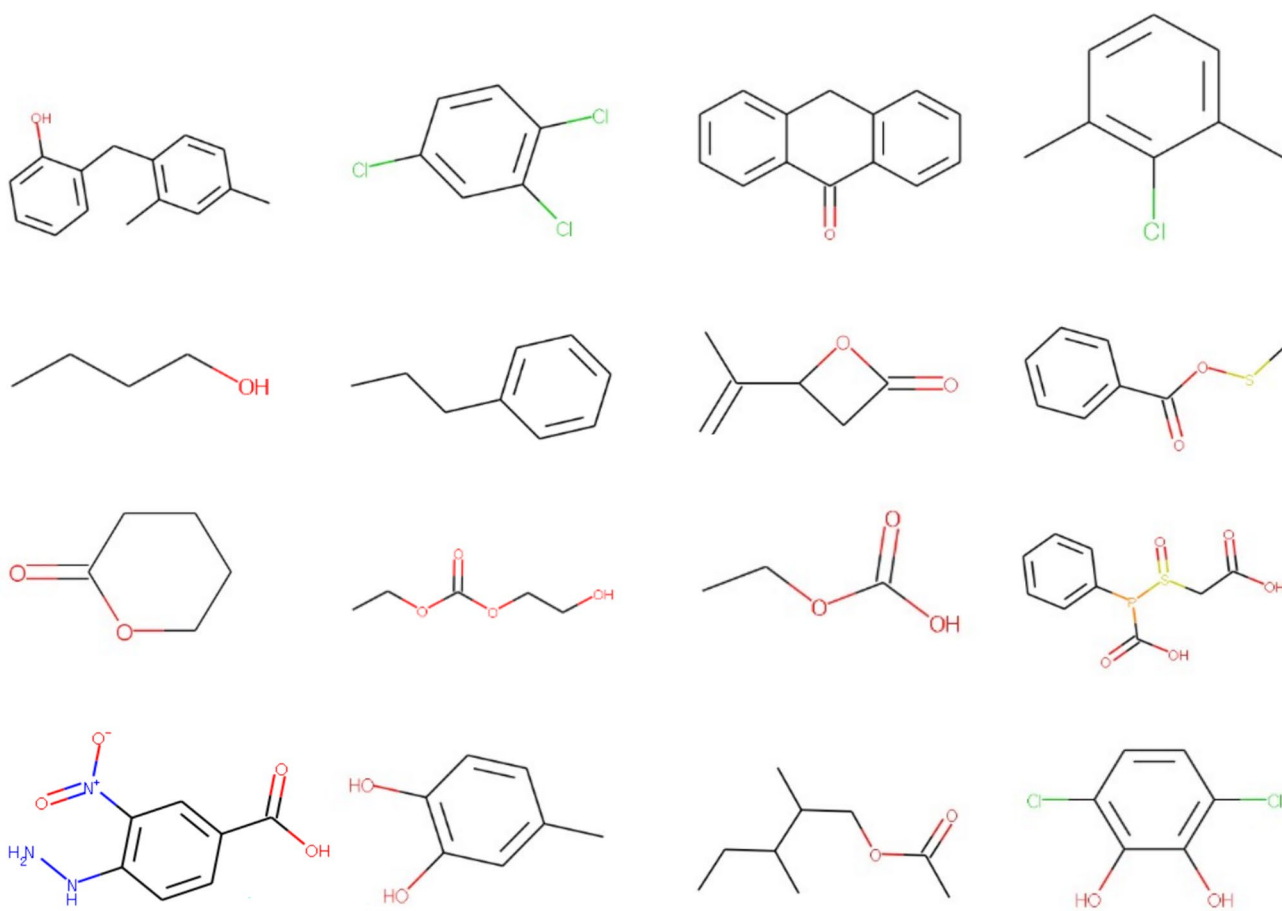


Fig. 4 Generated novel biodegradable materials from the GRU



into a discretized grid. Next, we systematically explore all possible permutations of values within this grid while concurrently evaluating various performance metrics through cross-validation. The grid point that yields the highest average value during cross-validation represents the optimal configuration of hyperparameters. Grid search is an algorithm that comprehensively explores all possible combinations, thereby enabling the identification of the optimal point within the given domain [48]. The significant limitation lies in its notably sluggish learning rate. Performing a comprehensive exploration of all spatial configurations necessitates a substantial amount of time. Acknowledging that each point within the grid necessitates k-fold cross-validation, a process that entails k-training iterations [49]. Thus, optimizing the hyperparameters of a model using this methodology can present significant intricacies and costs. However, exploring the synergistic effects of hyperparameters in pursuing optimal performance is practical, where a grid search is a superior approach in this endeavor. The range of hyperparameters explored is presented in Table 1.

## Results and Discussion

This section presents and discusses the findings from our proposed model and its comparison with previous models. Testing of the model on the electrolyte dataset to establish its generalizability and novel material discovery potential is also presented.

### Model Validation

In Fig. 3, we demonstrated the prediction capability of our integrated model in terms of training and validation datasets and compared them with the CNN- and LSTM-only models. It is worth noting that these models were optimized using the grid search method presented in “Hyperparameter Optimization”. The proposed model exhibited a very high accuracy of 87.2% compared to 75.4% and 79.3% for the CNN and LSTM-only models, ascribed to the proposed model’s ability to learn both spatial and temporal dependencies in the SMILES data, enhancing its capability to efficiently predict the biodegradability of the organic materials. The CNN-only model significantly overfits the model, as revealed in its higher training accuracy (AUC) of 0.981, albeit with a much lower test accuracy of 0.834 AUC. The LSTM-only model, on the other hand, outperformed the CNN-only model in terms of generalizability, achieving a test score of 0.892 AUC, but underperformed the integrated model. The accuracy of the proposed integrated model can be further improved by increasing the number of molecules in the

dataset, which is only 1055 in this case. It is worth noting that the hyperparameters employed in these models were optimized using the grid method elaborated in “Hyperparameter Optimization”.

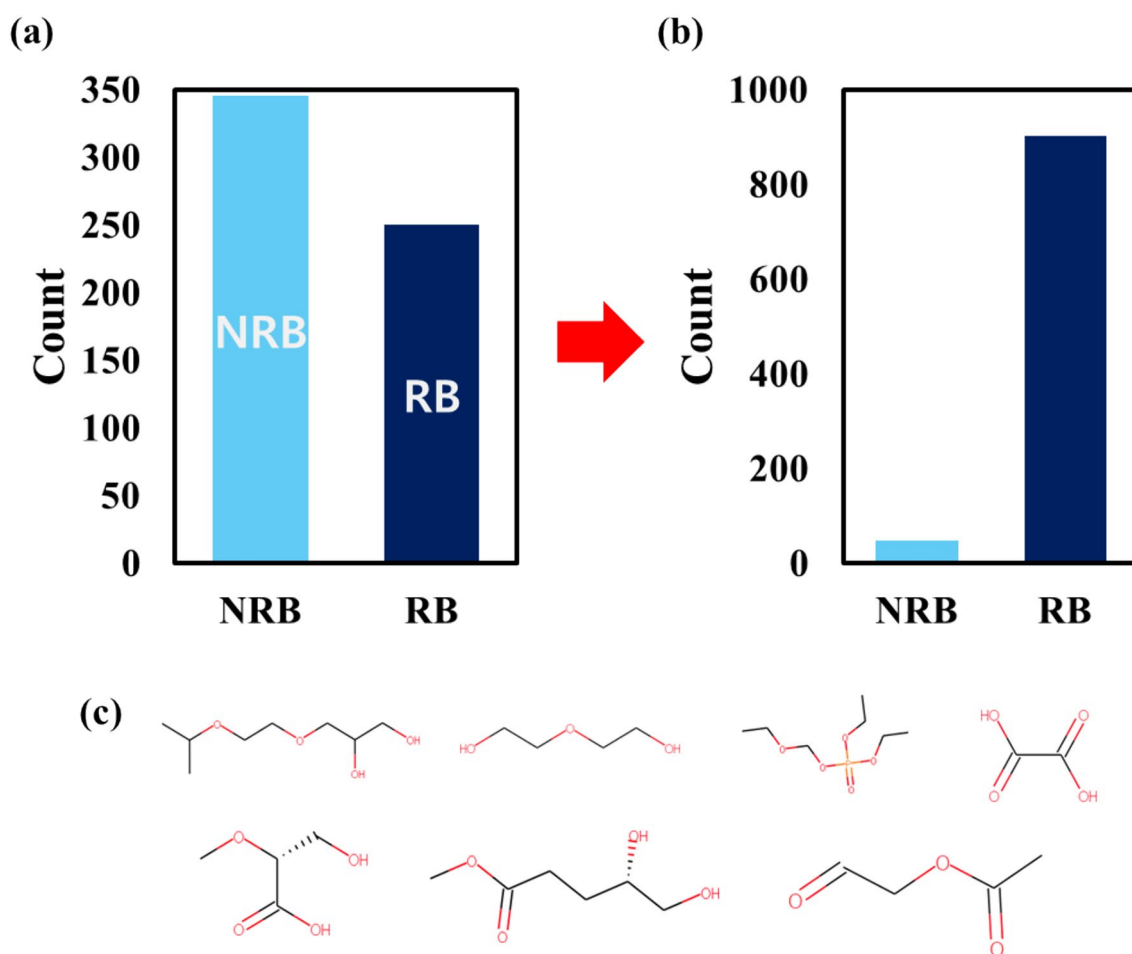
### Novel Material Discovery

Gated recurrent unit (GRU), which has a similar structure to LSTM, enables faster SMILES generation [54]. The GRU predicts the next character based on the current input character and hidden state. This allows for the generation of new compounds or materials. To train the GRU model, ‘<’ is added at the beginning and ‘>’ at the end of each SMILES sequence. This modification ensures that the model learns to recognize the start and end points of the SMILES sequences during training. This is demonstrated in this example: <[O-][N+](c1c(Cl)ccc([N+](O-)=O)c1)=O>.

From our results, approximately 60% of the generated SMILES are valid, indicating that most of them adhere to the structural rules governing the SMILES notation. In addition, more than 80% of these generated SMILES are distinct from the compounds in the training dataset, highlighting their novel characteristics and exploratory potential. Examples of the generated materials are presented in Fig. 4. Nevertheless, only about 40% of the generated SMILES were biodegradable (Fig. 5a). Hence, the biodegradable material generation ability of the model must be enhanced.

Further, the biodegradability of compounds generated in this step is analyzed based on their structural attributes, as demonstrated through various studies. Chlorinated aromatic compounds like Compound 1–2 (where 1 represents the row and 2 represents the column) typically exhibit low biodegradability due to the stability of C–Cl bonds, requiring anaerobic conditions for reductive dechlorination [55]. Conversely, aliphatic compounds such as Compound 2–1 are more biodegradable, as microbial enzymes effectively cleave carbon–carbon bonds [56]. Ester-containing compounds, including Compound 2–4, generally facilitate hydrolytic cleavage, enhancing microbial degradation [57]. However, fused aromatic rings, as seen in Compound 1–3, pose significant biodegradation challenges, requiring specific microbial pathways for effective degradation, indicative of their persistence in the environment [58]. These findings reveal that the generated compounds exhibit varying biodegradability potentials with some not likely degradable, necessitating algorithmic augmentation to enhance the generation capability of the model towards readily biodegradable compounds.

Therefore, by leveraging the developed biodegradability prediction model and the GRU-based SMILES generator, we can harness the power of RL to explore and discover novel organic compounds with inherent biodegradable properties.

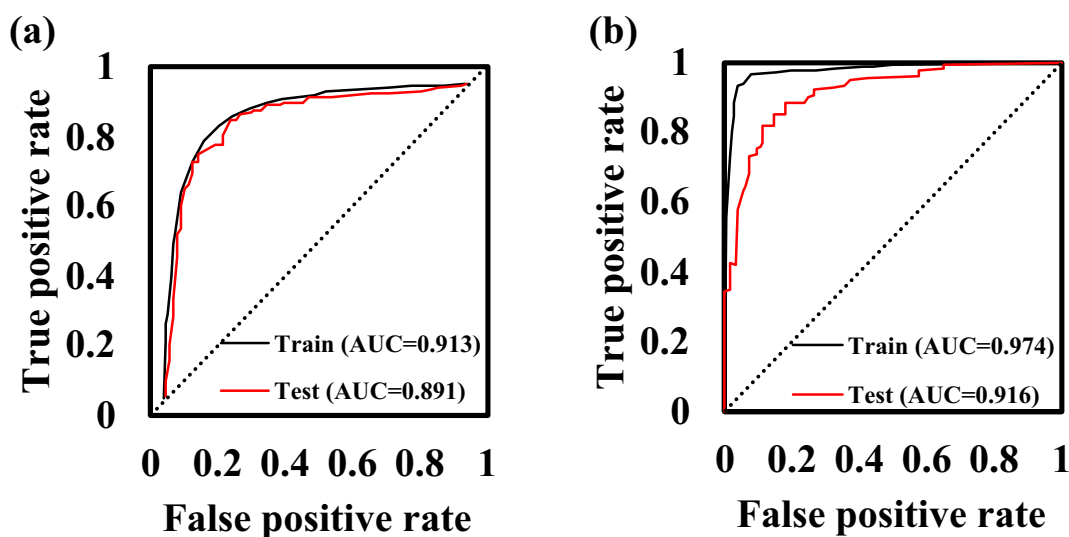


**Fig. 5** Comparison between **a** GRU-only generative model results and **b** GRU-RL generative model results. **c** Newly discovered materials with diverse functional groups

Here, a high reward is assigned if the generated material is not in the training data and is biodegradable. This allows the weights of the newly generated model to be updated. Therefore, we next present the results of the final generator model integrated with RL. The generative model has been successfully trained through the RL to discover more biodegradable compounds. Compared to the GRU-based generative model without RL, the final model could discover about 95% of biodegradable materials (Fig. 5a and b), of which 42% is not present in the original training dataset, demonstrating the novel material discovery capability of our model. By incorporating constraints on the similarity between specific functional groups/atoms and the generated compounds, we could generate diverse materials while preserving specific functional group/atom characteristics (Fig. 5c). Thus, by integrating the generator model with RL, the biodegradability potential of the discovered compounds improved, revealing several key characteristics. Aliphatic compounds with hydroxyl groups, such as Compound 1–1, typically exhibit increased biodegradability due to their

hydrophilicity, which aids microbial uptake [56]. Ether linkages in compounds like Compound 1–2 can introduce biodegradation pathways, though the rate depends on microbial communities and environmental conditions [59]. Phosphate ester functionalities in Compound 1–3 and ester groups in Compounds 2–1 and 2–2 offer susceptibility to hydrolytic cleavage, enhancing biodegradability [57, 60]. Additionally, compounds with dicarboxylic acid groups, such as Compound 1–4, generally show high biodegradability due to the susceptibility of carboxylic acids to microbial metabolism [59]. Also, compounds with aldehyde and ester functional groups, as in Compound 7, also have the potential for microbial degradation through various enzymatic pathways [56]. These findings highlight the potential of the proposed model in generating novel biodegradable materials across these compounds, reflecting its applicability in material science and other similar areas with limited data.

Next, we compared our model with the state-of-the-art model proposed by Popova et al. [36] for De novo drug



**Fig. 6** Comparison of **a** state-of-the-art model [36] results with our **b** our proposed model

design and obtained superior prediction results in terms of ROC, RMSE, and MAPE. Notably, our model significantly outperformed the state-of-the-art model with a training AUC of 0.974 and a testing AUC of 0.916 compared to 0.913 and 0.891, respectively. This result is ascribed to the proposed model's ability to learn both spatial and temporal dependencies in the trained data set using the CNN-LSTM integrated model. It is worth mentioning that the CNN component of the proposed model leads to overfitting (wider gap between the train and test results) and thus was carefully optimized to yield the expected synergistic effect. In terms of computational time, both models achieve nearly the same training time of 2 h and 20 min, indicating that no computing burden is incurred by the proposed model albeit with better accuracy (Fig. 6).

In summary, our methodology employed a hybrid architecture combining CNN and LSTM networks to process both spatial and sequential data from molecular structures. This approach contrasts with the simpler LSTM model employed by [36] which focuses exclusively on SMILES strings. Our model enhances robustness and generalizability through advanced data processing and augmentation techniques that efficiently manage diverse SMILES strings and structural images. Notably, despite using a significantly smaller dataset of only 1055 entries compared to the 1.5 million used in previous studies, our model demonstrates superior performance. This is achieved through our novel use of data augmentation and the synergistic capabilities of the CNN-LSTM architecture, demonstrating our model's effectiveness in chemical applications where large datasets are typically unavailable.

## Conclusion

This study presents a comprehensive analysis of our proposed integrated model for biodegradability prediction and novel material discovery. The model's predictive capabilities were validated, demonstrating superior performance compared to CNN- and LSTM-only models. The integrated model achieved an impressive 87.2% AUC, showcasing its ability to learn spatial and temporal dependencies in SMILES data. Our novel material discovery approach, utilizing a GRU-based SMILES generator within a reinforcement learning framework, showed significant potential. Around 60% of the generated SMILES were valid, and over 80% were distinct from the training dataset, indicating their novelty. Moreover, through RL, we enhanced the model's ability to generate biodegradable materials, with approximately 95% being biodegradable, including 42% not present in the original training dataset. Furthermore, we compared our model to a state-of-the-art model proposed for De novo drug design and achieved superior results in terms of ROC, highlighting the model's potential in diverse applications. Expanding the scope of our research to the design of novel electrolytes by employing large-scale molecular data, we developed a novel electrolyte with specific properties like low viscosity, high conductivity, and cost-effectiveness, contributing to the advancement of organic materials and their applications. Our integrated model has shown exceptional promise in biodegradability prediction, material discovery, and electrolyte design. The complexities of synthesizing new chemical compounds generated by our hybrid model, alongside significant costs, and stringent regulatory requirements, pose substantial challenges to the experimental testing of our theoretical predictions. In response, our future work will implement DFT

as an intermediate step to computationally validate and refine our predictions of compound biodegradability. This method will enable us to explore the feasibility of synthesis and assess molecular behavior without the immediate resource-intensive requirements of laboratory experiments. Through this approach, we aim to bridge the gap between theoretical research and practical application, advancing towards reliable biodegradability assessments while mitigating the challenges of direct empirical testing. Future work could also further enhance the model's capabilities and explore its applications in various material discovery fields. This research represents a significant step towards leveraging artificial intelligence for material discovery and design in today's dynamic scientific landscape.

**Acknowledgements** This research was supported by Inha University Research.

**Funding** This research was supported by Inha University (Grant 70317).

## References

1. F. Wu, M. Misra, A.K. Mohanty, Challenges and new opportunities on barrier performance of biodegradable polymers for sustainable packaging. *Prog. Polym. Sci.* **117**, 101395 (2021). <https://doi.org/10.1016/j.progpolymsci.2021.101395>
2. R. Grace, Closing the circle: reshaping how products are conceived and made. *Plast. Eng.* **73**, 8–11 (2017). <https://doi.org/10.1002/j.1941-9635.2017.tb01670.x>
3. F. Allen, J. Gasparro, J. Swaney, M. Phelan, J. Gillespie, Directive 2004/38/EC of the European Parliament and of the Council of 29 April 2004, Immigration Law Handbook (2023) 2253-C79P212. <https://doi.org/10.1093/oso/9780192896292.003.0079>
4. T. No, 301: Ready biodegradability. OECD (1992). <https://doi.org/10.1787/9789264070349-en>
5. P.A. Vanrolleghen, K.J. Keesman. Identification of biodegradation models under model and data uncertainty, *Water Sci. Technol.* (1996). [https://doi.org/10.1016/0273-1223\(96\)00192-8](https://doi.org/10.1016/0273-1223(96)00192-8)
6. P.G. Polishchuk, T.I. Madzhidov, A. Varnek, Estimation of the size of drug-like chemical space based on GDB-17 data. *J. Comput. Aided Mol. Des.* **27**, 675–679 (2013). <https://doi.org/10.1007/s10822-013-9672-4>
7. D. Weininger, SMILES, a chemical language and information system. I. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988). <https://doi.org/10.1021/ci00057a005>
8. C. Bilodeau, W. Jin, T. Jaakkola, R. Barzilay, K.F. Jensen, Generative models for molecular discovery: recent advances and challenges. *WIREs Comput. Mol. Sci.* (2022). <https://doi.org/10.1002/wcms.1608>
9. M. Olivecrona, T. Blaschke, O. Engkvist, H. Chen, Molecular de novo design through deep reinforcement learning. *J. Cheminform.* **9**, 48 (2017). <https://doi.org/10.1186/s13321-017-0235-x>
10. P.-H. Chiu, Y.-L. Yang, H.-K. Tsao, Y.-J. Sheng, Deep learning for predictions of hydrolysis rates and conditional molecular design of esters. *J. Taiwan Inst. Chem. Eng.* **126**, 1–13 (2021). <https://doi.org/10.1016/j.jtice.2021.06.045>
11. M. Wang, C.-Y. Hsieh, J. Wang, D. Wang, G. Weng, C. Shen, X. Yao, Z. Bing, H. Li, D. Cao, T. Hou, RELATION: a deep generative model for structure-based de novo drug design. *J. Med. Chem.* **65**, 9478–9492 (2022). <https://doi.org/10.1021/acs.jmedchem.2c00732>
12. J. Arús-Pous, A. Patronov, E.J. Bjerrum, C. Tyrchan, J.-L. Raymond, H. Chen, O. Engkvist, SMILES-based deep generative scaffold decorator for de-novo drug design. *J. Cheminform.* **12**, 38 (2020). <https://doi.org/10.1186/s13321-020-00441-8>
13. N. De Cao, T. Kipf, MolGAN: An implicit generative model for small molecular graphs, ArXiv abs/1805.1 (2018) null. <https://www.semanticscholar.org/paper/def1049b5aae96c8e1eab0ca58d77ac9c2f0e3e9>
14. W. Tang, Y. Li, Y. Yu, Z. Wang, T. Xu, J. Chen, J. Lin, X. Li, Development of models predicting biodegradation rate rating with multiple linear regression and support vector machine algorithms. *Chemosphere* **253**, 126666 (2020). <https://doi.org/10.1016/j.chemosphere.2020.126666>
15. O. Dollar, N. Joshi, D.A.C. Beck, J. Pfandtner, Attention-based generative models for de novo molecular design. *Chem. Sci.* **12**, 8362–8372 (2021). <https://doi.org/10.1039/d1sc01050f>
16. F. Lunghini, G. Marcou, P. Gantzer, P. Azam, D. Horvath, E. Van Miert, A. Varnek, Modelling of ready biodegradability based on combined public and industrial data sources. *SAR QSAR Environ. Res.* **31**, 171–186 (2019). <https://doi.org/10.1080/1062936x.2019.1697360>
17. W.F.C. Rocha, D.A. Sheen, Classification of biodegradable materials using QSAR modelling with uncertainty estimation. *SAR QSAR Environ. Res.* **27**, 799–811 (2016). <https://doi.org/10.1080/1062936x.2016.1238010>
18. K. Acharya, D. Werner, J. Doling, M. Barycki, P. Meynet, W. Mrozik, O. Komolafe, T. Puzyn, R.J. Davenport, A quantitative structure-biodegradation relationship (QSBR) approach to predict biodegradation rates of aromatic chemicals. *Water Res.* **157**, 181–190 (2019). <https://doi.org/10.1016/j.watres.2019.03.086>
19. B. Barros, P. Lacerda, C. Albuquerque, A. Conci, Pulmonary COVID-19: learning spatiotemporal features combining CNN and LSTM networks for lung ultrasound video classification. *Sensors (Basel)* (2021). <https://doi.org/10.3390/S21165486>
20. H.V. Dang, H. Tran-Ngoc, T.V. Nguyen, T. Bui-Tien, G. De Roeck, H.X. Nguyen, Data-driven structural health monitoring using feature fusion and hybrid deep learning. *IEEE Trans. Autom. Sci. Eng.* **18**, 2087–2103 (2021). <https://doi.org/10.1109/TASE.2020.3034401>
21. P. Bilokon, Y. Qiu, Transformers versus LSTMs for electronic trading. *SSRN Electron. J.* (2023). <https://doi.org/10.2139/ssrn.4577922>
22. Y. Tatsunami, M. Taki, Sequencer: Deep LSTM for Image Classification, *Adv Neural Inf Process Syst* **35** (2022). <https://arxiv.org/abs/2205.01972v4>. Accessed 29 Apr 2024
23. A. Zeyer, P. Bahar, K. Irie, R. Schluter, H. Ney, A Comparison of Transformer and LSTM Encoder Decoder Models for ASR, 2019 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019 - Proceedings (2019) 8–15. <https://doi.org/10.1109/ASRU46091.2019.9004025>
24. R.T.B.D.T.R. Mansouri Kamel, V. Consonni, QSAR biodegradation, (2013)
25. P. Dey, S.K. Chaulya, S. Kumar, Hybrid CNN-LSTM and IoT-based coal mine hazards monitoring and prediction system. *Process. Saf. Environ. Prot.* **152**, 249–263 (2021). <https://doi.org/10.1016/j.psep.2021.06.005>
26. Y. Zhao, Improvement and application of multi-layer LSTM Algorithm based on spatial-temporal correlation. *Ingénierie Des Systèmes d'Inf.* **25** (2020) null. <https://doi.org/10.18280/isi.250107>
27. C. Ding, G. Wang, X. Zhang, Q. Liu, X. Liu, A hybrid CNN-LSTM model for predicting PM2.5 in Beijing based on

- spatiotemporal correlation. *Environ. Ecol. Stat.* **28**, 503–522 (2021). <https://doi.org/10.1007/s10651-021-00501-8>
28. D.Q. Gbadago, J. Moon, M. Kim, S. Hwang, A unified framework for the mathematical modelling, predictive analysis, and optimization of reaction systems using computational fluid dynamics, deep neural network and genetic algorithm: a case of butadiene synthesis. *Chem. Eng. J.* **409**, 128163 (2021). <https://doi.org/10.1016/j.cej.2020.128163>
  29. J. Moon, D.Q. Gbadago, G. Hwang, D. Lee, S. Hwang, Software platform for high-fidelity-data-based artificial neural network modeling and process optimization in chemical engineering. *Comput. Chem. Eng.* **158**, 107637 (2022). <https://doi.org/10.1016/j.compchemeng.2021.107637>
  30. P. Dey, K. Saurabh, C. Kumar, D. Pandit, S.K. Chaulya, S. Ray, G.M. Prasad, S.K. Mandal, t-SNE and variational auto-encoder with a bi-LSTM neural network-based model for prediction of gas concentration in a sealed-off area of underground coal mines. *Soft. Comput.* **25**, 14183–14207 (2021). <https://doi.org/10.1007/s00500-021-06261-8>
  31. W. Wang, A Pre-trained Conditional Transformer for Target-specific De Novo Molecular Generation, (2022). <https://www.semanticscholar.org/paper/ed9763062daec0eec7ceb65e822360e340c75605>
  32. X. Yang, Z. Zhang, An attention-based domain spatial-temporal meta-learning (ADST-ML) approach for PM<sub>2.5</sub> concentration dynamics prediction. *Urban Clim.* (2023). <https://doi.org/10.1016/j.uclim.2022.101363>
  33. N. Xu, X. Wang, X. Meng, H. Chang, Gas concentration prediction based on IWOA-LSTM-CEEMDAN residual correction model. *Sensors (Basel)* (2022). <https://doi.org/10.3390/s22124412>
  34. L. Pingyang, N. Chen, M. Shanjun, L. Mei, LSTM based encoder-decoder for short-term predictions of gas concentration using multi-sensor fusion. *Process. Saf. Environ. Prot.* **137**, 93–105 (2020). <https://doi.org/10.1016/j.psep.2020.02.021>
  35. K. Kumari, P. Dey, C. Kumar, D. Pandit, S. Mishra, V. Kisku, S.K. Chaulya, S. Ray, G.M. Prasad, UMAP and LSTM based fire status and explosibility prediction for sealed-off area in underground coal mine. *Process. Saf. Environ. Prot.* **146**, 837–852 (2021). <https://doi.org/10.1016/j.psep.2020.12.019>
  36. M. Popova, O. Isayev, A. Tropsha, Deep reinforcement learning for de novo drug design. *Sci. Adv.* **4**, eaap7885 (2018). <https://doi.org/10.1126/sciadv.aap7885>
  37. M. Popova, M. Shvets, J.B. Oliva, O. Isayev, MolecularRNN: Generating realistic molecular graphs with optimized properties, *ArXiv abs/1905.1* (2019). <https://www.semanticscholar.org/paper/3ccd291c8848c73ca34152e27c3ec296cfe838d0>
  38. Z. Zhou, S. Kearnes, L. Li, R. Zare, P.F. Riley, Optimization of molecules via deep reinforcement learning. *Sci. Rep.* (2018). <https://doi.org/10.1038/s41598-019-47148-x>
  39. Bellman-consistent Pessimism for Offline Reinforcement Learning | OpenReview, (n.d.). <https://openreview.net/forum?id=e8WWUBeafM>. Accessed 10 Oct 2023
  40. B. O'donoghue, I. Osband, R. Munos, V. Mnih, The uncertainty bellman equation and exploration, (2018)
  41. Y. Fei, Z. Yang, Y. Chen, Z. Wang, Exponential Bellman Equation and Improved Regret Bounds for Risk-Sensitive Reinforcement Learning, (n.d.)
  42. J. Bergstra, J.B. Ca, Y.B. Ca, Random search for hyper-parameter optimization Yoshua Bengio. *J. Mach. Learn. Res.* **13**, 281–305 (2012). <http://scikit-learn.sourceforge.net>. Accessed 29 Apr 2024
  43. J. Ma, Y. Ding, J.C.P. Cheng, F. Jiang, V. Gan, Z. Xu, A Lag-FLSTM deep learning network based on Bayesian optimization for multi-sequential-variant PM<sub>2.5</sub> prediction. *Sustain. Cities Soc.* **60**, 102237 (2020). <https://doi.org/10.1016/j.scs.2020.102237>
  44. M. Kim, A. Han, J. Lee, S. Cho, I. Moon, J. Na, Comparison of derivative-free optimization: energy optimization of steam methane reforming process. *Int. J. Energy Res.* (2023). <https://doi.org/10.1155/2023/8868540>
  45. H. Alibrahim, S.A. Ludwig, Hyperparameter Optimization: Comparing Genetic Algorithm against Grid Search and Bayesian Optimization, 2021 IEEE Congress on Evolutionary Computation, CEC 2021 - Proceedings (2021) 1551–1559. <https://doi.org/10.1109/CEC45853.2021.9504761>
  46. M. Ogunsanya, J. Isichei, S. Desai, Grid search hyperparameter tuning in additive manufacturing processes. *Manuf Lett* **35**, 1031–1042 (2023). <https://doi.org/10.1016/j.mfglet.2023.08.056>
  47. S.M. LaValle, M.S. Branicky, S.R. Lindemann, On the relationship between classical grid search and probabilistic roadmaps. *Int. J. Robot. Res.* **23**, 673–692 (2004). <https://doi.org/10.1177/0278364904045481>
  48. P. Liashchynskiy, P. Liashchynskiy, Grid search, random search, genetic algorithm: a big comparison for NAS, (2019). <https://arxiv.org/abs/1912.06059v1>. Accessed 11 Oct 2023
  49. F.J. Pontes, G.F. Amorim, P.P. Balestrassi, A.P. Paiva, J.R. Ferreira, Design of experiments and focused grid search for neural network parameter optimization. *Neurocomputing* **186**, 22–34 (2016). <https://doi.org/10.1016/j.neucom.2015.12.061>
  50. R.Y. Acharya, N.F. Charlot, M.M. Alam, F. Ganji, D. Gauthier, D. Forte, Chaogate parameter optimization using bayesian optimization and genetic algorithm, Proceedings - International Symposium on Quality Electronic Design, ISQED 2021-April (2021) 426–431. <https://doi.org/10.1109/ISQED51717.2021.9424355>
  51. H. Alibrahim, S.A. Ludwig, Hyperparameter Optimization: Comparing Genetic Algorithm against Grid Search and Bayesian Optimization, IEEE Congress on Evolutionary Computation (CEC) (2021) 1551–1559. <https://doi.org/10.1109/cec45853.2021.9504761>
  52. Y. Shin, Z. Kim, J. Yu, G. Kim, S. Hwang, Development of NO<sub>x</sub> reduction system utilizing artificial neural network (ANN) and genetic algorithm (GA). *J. Clean. Prod.* **232**, 1418–1429 (2019). <https://doi.org/10.1016/j.jclepro.2019.05.276>
  53. F. Mohammadi, M.R. Samaei, A. Azhdarpoor, H. Teiri, A. Badeenezhad, S. Rostami, Modelling and optimizing pyrene removal from the soil by phytoremediation using response surface methodology, artificial neural networks, and genetic algorithm. *Chemosphere* **237**, 124486 (2019). <https://doi.org/10.1016/j.chemosphere.2019.124486>
  54. B. Athiwaratkun, J.W. Stokes, Malware classification with LSTM and GRU language models and a character-level CNN, ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings (2017) 2482–2486. <https://doi.org/10.1109/ICASSP.2017.7952603>
  55. J. Palau, R. Yu, S. Hatijah Mortan, O. Shouakar-Stash, M. Rosell, D.L. Freedman, C. Sbarbati, S. Fiorenza, R. Aravena, E. Marco-Urrea, M. Elsner, A. Soler, D. Hunkeler, Distinct dual C-Cl isotope fractionation patterns during anaerobic biodegradation of 1,2-dichloroethane: potential to characterize microbial degradation in the field. *Environ. Sci. Technol.* **51**, 2685–2694 (2017). [https://doi.org/10.1021/ACS.EST.6B04998/ASSET/IMAGES/LARGE/ES-2016-04998R\\_0003.JPEG](https://doi.org/10.1021/ACS.EST.6B04998/ASSET/IMAGES/LARGE/ES-2016-04998R_0003.JPEG)
  56. R.S. Boethling, M. Alexander, Effect of concentration of organic chemicals on their biodegradation by natural microbial communities. *Appl. Environ. Microbiol.* **37**, 1211–1216 (1979). <https://doi.org/10.1128/AEM.37.6.1211-1216.1979>
  57. A.A. Toropov, A.P. Toropova, A. Lombardo, A. Roncaglioni, N. De Brita, G. Stella, E. Benfenati, CORAL: the prediction of biodegradation of organic compounds with optimal SMILES-based descriptors. *Cent. Eur. J. Chem.* **10**, 1042–1048 (2012). <https://doi.org/10.2478/S11532-012-0031-4/MACHINEREADEABLECITAT ION/RIS>

58. X.L. Wang, R.X. Zhuo, L.J. Liu, Synthesis and characterization of novel biodegradable poly (carbonate-co-phosphate)s. *Polym. Int.* **50**, 1175–1179 (2001). <https://doi.org/10.1002/PI.711>
59. R.S. Boethling, Application of molecular topology to quantitative structure-biodegradability relationships. *Environ. Toxicol. Chem.* **5**, 797–806 (1986). <https://doi.org/10.1002/ETC.5620050904>
60. J. Xu, Z.L. Liu, R.X. Zhuo, Synthesis and properties of biodegradable copolymers of 9-phenyl-2, 4, 8, 10-tetraoxaspiro-[5, 5] undecane-3-one and ethylene ethyl phosphate. *Chin. Chem. Lett.* **17**, 1365–1368 (2006). Accessed 29 Apr 2024

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.