Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Recent applications of artificial intelligence-based site-specific integrases

Chang Wang[1,] ✉

[1] *Chemical Engineering and Technology School of Tianjin University, Tianjin, 300072, China*

ABSTRACT

The rapid development of artificial intelligence technology provides new tools to optimize the design and application of site-specific integrases and drive innovation in this field. In this study, a site-specific integrase generation model based on artificial intelligence was designed. The learning effect of the model to generate site-specific integrase is improved by mining sequence data of site-specific integrase with feature selection and discretization, and then using generative adversarial network as a framework to extract the detail information of protein sequences by using convolutional layer, and extracting the global features of sequences by using self-attention layer. In addition, to address the degradation problem during training, a residual structure module is constructed and spectral normalization is used to ensure training stability. Meanwhile, Gumbel Softmax Trick is used to solve the problem of non-returnable gradient of discrete data generated by the model. The sequence of the site-specific integrase generated by the model showed 92% identity with the training set, which has better sequence quality. In terms of amino acid composition, the Pearson value with the natural amino acid composition was greater than 0.8, and the two were highly correlated. The site-specific integrase can increase the expression of bax protein and decrease the expression of bax-2 protein and Ki67 protein in lung cancer patients, which is favorable for patient treatment. It can up-regulate the expression of ovarian STAR, CYP11A1, CTP19A1, and 3β-HSD genes and promote steroidogenesis in ewes. The alkane content of the group of strains incorporating site-specific integrase was 57.25%~63.00% lower than that of those without the enzyme in a high concentration of petroleum pollution environment.

Keywords: Artificial intelligence, Site-specific integrase, Generative adversarial network, Self-attention, Residual structure

✉ Corresponding author.
E-mail address: veronica1104@163.com (C. Wang).
Received 18 February 2024; Revised 10 April 2024; Accepted 25 December 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-427

1. Introduction

In transgenic research such as transgenic animal preparation and gene therapy, viral vectors and stochastic integration vectors are mostly used for gene transfection. Although these vectors are effective in gene transfection, they are faced with the shortcomings that the target genes can not be specifically integrated [1-4]. Although the later development of targeting vectors solved the problem of random integration of target genes to a certain extent, the targeting efficiency is not high in actual operation, and it increases the workload of researchers, which largely restricts the application of targeting vectors and the development of transgenic technology [5-8].

The emergence of site-specific integrase, which can promote the recombination of specific sites between 2 DNA strands to achieve targeted and efficient integration [9-10], has well solved the above problems. According to the different homology and catalytic mechanism, site-specific integrases can be divided into 2 categories, namely tyrosine integrases and serine integrases, and most of the commonly used integrase systems nowadays belong to these 2 categories, such as Cre/Loxp, FLP, P1 phage, and other integrase systems are tyrosine integrases [11-14]. φC31, R4, Tn4, etc. are serine integrases. Some studies have shown that the integration efficiency of tyrosine-like integrases is lower than that of serine-like integrases [15-16]. φC31 site-specific integrase, as a serine-like integrase, does not need to resort to external energy or cofactors when promoting site-specific recombination between DNA strands and not only realizes the site-specific integration, but also has a high integration efficiency [17-20].

Literature [21] highlights that λ site-specific recombination pathways have been receiving a lot of attention from experts in various fields and have proved to be useful as model systems that are both simple and complex enough to achieve a satisfactory level of understanding, and sufficiently attractive to attract a wider range of scientists to study the problem. Literature [22] devised a system based on split-internal serine integrase that has the potential to modulate site-specific recombination events and revealed split integrase as a potential, versatile and modifiable component. Literature [23] describes the characterization of phage-derived integrases and based on the efficient integration of Bxb1 integrase in mammalian cells, reveals that point mutation of Bxb1 target sites improves the efficiency of Bxb1-mediated integration at the Rosa26 locus in Chinese hamster ovary cells. Literature [24] developed a phage-assisted sequential evolutionary method, IntePACE, which efficiently improves the activity of PhiC31 and Bxb1 serine integrases and accelerates gene delivery therapy that can be more easily adapted to improve new integrases from nature. Literature [25] develops a toolkit that combines RNA-guided Cas9 and the site-specific integrase Bxb1, which will provide researchers with a valuable resource with applications in therapeutic gene integration and biopharmaceuticals. Literature [26] proposed an effective site-specific gene integration strategy that allows rapid and precise insertion of gene expression cassettes into specific sites in mammalian cells, achieving uniform transgene expression, and found that the H11 gene is a “safe harbor” site for gene knock-in in CHO-S cells. Literature [27] demonstrated that site-specific recombination (SSR) units are all compatible and interchangeable, and identified closely related prophages with different SSR units that control the rearrangement of the developmentally regulated kamA gene, revealing that SSR units are interchangeable components of MGEs. Literature [28] proposed a mathematical model that explains the directionality and regularity of the “forward” and “reverse” reactions, fits the kinetics of recombination observed in vitro by ϕC31 integrase, and defines the core features necessary for the directionality of the system. Characterization. Serine integrases were introduced in the literature [29], and a timeline with milestones describing representative achievements was presented, pointing out the applications of serine integrases in genome engineering, design and DNA assembly, and their potential for advancing interdisciplinary research.

The article designs a site-specific integrase generation model based on generative adversarial network for the application of artificial intelligence technology in enzyme engineering. Site-specific

integrase sequence features are extracted and discretized from the dataset, and the product sum of local regions is calculated using fractional stepwise convolution to improve the local information mining ability. The attention weight distribution of each position vector is obtained from the self-attention layer, and the global features are extracted. At the same time, the spectral normalization technique is introduced to limit the maximum feature value in the network to keep it in a fixed range to improve the training stability. The constructed residual module is then utilized to capture the information difference between input and output to prevent the model degradation phenomenon. Using this technique, corresponding site-specific integrases are generated in the fields of medicine, animal husbandry, and environment to study their practical application effects in popular fields.

2. Application of artificial intelligence techniques in enzyme engineering

2.1. Molecular design of auxiliary enzymes

In AI-assisted enzyme molecule design, various machine algorithms are used to construct the mapping linkage of "input properties - output properties" through the existing data. According to whether the trained experimental data have labeled information (e.g., one or more attributes of target proteins), machine learning [30] can be classified into supervised and unsupervised learning, and machine learning in the directed evolution of enzyme molecules generally belongs to supervised learning. Its main steps are after extracting protein feature attributes, machine learning algorithms will be delivered to learn and generate an objective function that can describe the data model, and virtual evolution of protein sequences, evaluate the efficacy through training and testing, and finally give the prediction results. There are three key points: building the sequence function model, molecular descriptors and machine algorithms. Molecular descriptors and algorithms have a greater impact on the accuracy of the prediction model, and are also the focus of AI strategy development.

2.2. Relevant databases and online platform support

In the era of artificial intelligence, databases and online platforms are indispensable resources for researchers, and they provide rich and powerful data support and efficient and fast analysis tools for the directed evolution of enzyme molecules. Databases contain a large amount of data on the function, structure and mutation effects of proteins. The online platform provides a full set of tools from feature extraction to prediction model construction, enabling researchers to analyze and predict protein sequences using artificial intelligence algorithms. The combined use of these databases and online platforms not only improves research efficiency, but also enhances the prediction of protein functions and properties through machine learning algorithms, helping researchers optimize enzyme design strategies.

3. Artificial intelligence-based modeling of site-specific integrase generation

3.1. Feature Extraction and Selection

3.1.1. Feature extraction. For the i st sample p of length L in the site-specific integrase dataset, the density of feature $f_{i,j}$ (the j th feature in the feature set) was calculated as Eq:

$$f_{i,j} = density_{p,f_{i,j}} = \frac{1}{L'}\sum_{k=1}^{L'} I(p_k) \tag{1}$$

where $f_{i,j}$ represents the value of the j rd feature taken in the i nd sample, and p_k represents the subsequence taken in the k th sliding window of sequence p. Considering that different sequence lengths will make the feature density calculation biased, this paper uses the sequence length to normalize the feature density calculation.

3.1.2. Feature selection. One-way analysis of variance (ANOVA) [31], also known as F-test, the larger the F-value, the better the discriminatory power of the corresponding feature. In this study, the F-value for each feature f (the jth feature in the feature set) is calculated as Eq:

$$F_j = \frac{\sum_{\mu=1}^{M} n_\mu (\bar{f}(\mu) - \bar{f})^2}{M-1} / \frac{\sum_{\mu=1}^{M} \sum_{i=1}^{n_\mu} (f_{i,j}(\mu) - \bar{f}(\mu))^2}{N-M} \tag{2}$$

where M and N denote the number of categories and the total number of samples, respectively. $\bar{f}(\mu)$ denotes the mean of the μth category of feature f. $\bar{f}$ is the mean value of feature f over the entire dataset. $f_{i,j}(\mu)$ represents the value of the feature taken by the jth feature of the ith sample and the category corresponding to the sample is μ. n represents the number of samples in the μth category.

By calculating the F-values of the features and sorting them in descending order of the F-values, the features with smaller distributional differences within classes and larger distributional differences between classes can be found, and here the ANOVA is used to measure the distributional differences of the features within classes as well as between classes.

3.1.3. Feature discretization. In this paper, we use feature discretization to transform continuous numerical features into discrete numerical features. The core of feature discretization is the creation of a many-to-one mapping function that maps the original continuous numeric data to the new discrete data. Discretization usually involves reducing the data from a large range of continuous data values to a smaller range of discrete values; therefore, discretization is often considered a data simplification mechanism. Discretization is easier to handle and closer to knowledge representation than continuous attributes.

Since the protein sequence features except for monopeptides are too sparse, in this paper, we directly divide the samples with 0 occurrences into class 0, which is represented by the number 0, and the samples with non-zero frequencies into class 1 and class 2, which represent the medium and high frequencies of amino acid occurrences, respectively.

In order to segment continuous data into non-overlapping discrete data for specific features, this study first sorts the samples in descending or ascending order by the frequency of occurrence of the features, and then finds the optimal segmentation point for the current feature. In fact, the finding process of each feature corresponding to the optimal segmentation point is independent of each other, so the optimal cutting point of each feature is solved separately in this paper.

The next section focuses on the discretization process of feature f. The frequency of occurrence of feature f in the samples is sorted in ascending order before solving for the optimal cut point of feature f. Samples with a frequency of occurrence of 0 are mapped to the symbol 0 and then these samples are excluded from the samples used to find the cut point. Taking the median (m) of the frequency of occurrence of non-zero samples as the optimal segmentation point, two non-overlapping segmentation intervals of feature f can be obtained by segmentation point division: (0,m]and(m,∞]. Then, this paper splits the continuous values into discrete symbols according to the partitioned intervals.

3.2. Generative Adversarial Network Modeling Framework

The training model is based on ProteinGAN [32] generative adversarial network model, the ProteinGAN model architecture is shown in Figure 1. The model consists of two parts: generator and discriminator. In practical applications, the model dimensions, optimizer parameters, and training hyperparameters were adjusted according to the dataset. The final model used for training has a total of 45192752 parameters that can be trained, of which 27691634 parameters can be trained by the generator and 26645127 parameters can be trained by the discriminator. The Adam (Adaptive

Moment Estimation) algorithm was used for optimization, with the first-order and second-order momentum set to 0.0 and 0.8, respectively. The model encodes the protein sequence in 21 dimensions using One-Hot coding, which includes 20 standard amino acids and a space indicating the beginning or end of the sequence. The input to the model generator is a random vector of length 128 obeying the normal distribution N (0, 0.5), and the output dimension is 512 × 21. The input to the model discriminator is a 512 × 21 One-Hot encoded protein sequence, and the output is a 1-dimensional scoring value, which is used to measure the authenticity of the input sequence.

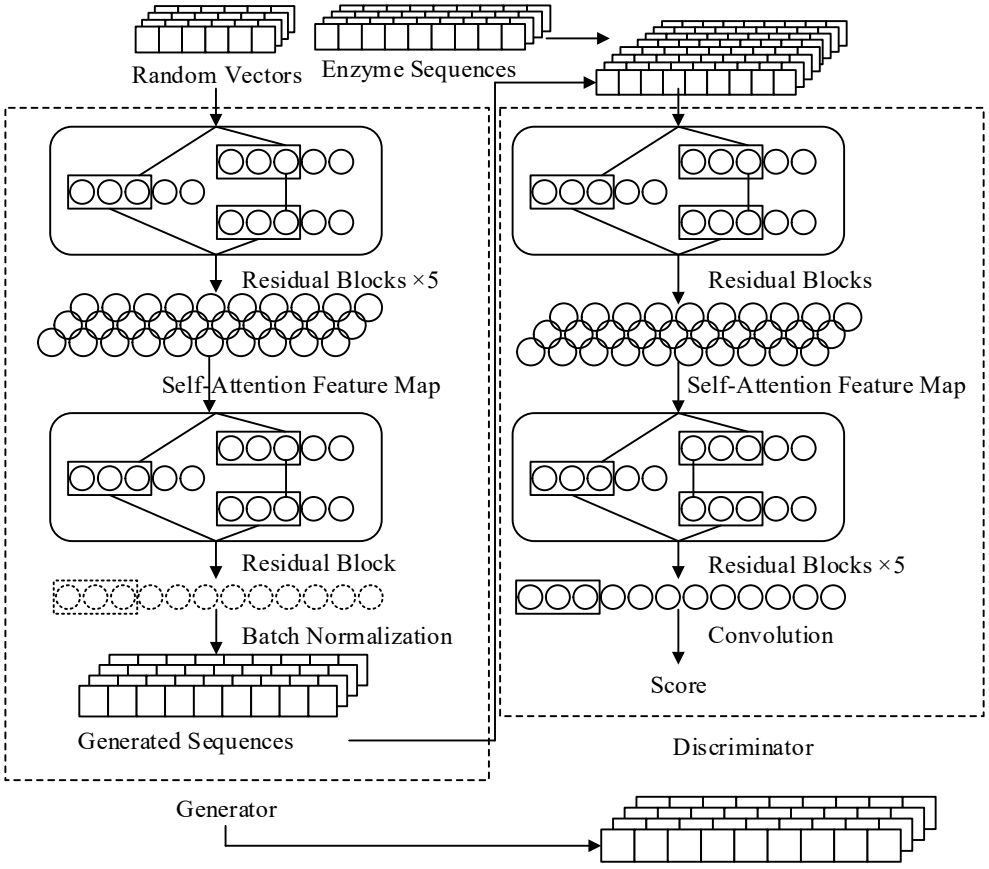

Fig. 1. Model Architecture Diagram

3.2.1. Convolutional layers. The model employs a convolutional layer to extract the localized detail information of the training protein sequence data. In this paper, DCGAN (Deep Convolutional Generative Adversarial Network) [33] is used to avoid the loss of local information during training. DCGAN uses stepwise convolution to downsample the image in the discriminator network, and fractional stepwise convolution to upsample the image in the generator network, and to prevent the loss of information, the pooling layer is removed from the generator part of the network.

Principle of Fractional Step Convolution: In a standard convolution operation, the convolution kernel slides over the input tensor and computes the sum of the products of the local regions to obtain the convolution result. As the network progresses through the layers, overly detailed positional information is lost. Whereas, fractional stepwise convolution is a method of inverse convolution commonly used in convolutional neural networks for up-sampling images. In fractional stepwise convolution, the input and output feature maps are of different sizes, and a complementary 0 operation is performed on the edges of the input feature map to control the size of the output feature map. The weights of the convolution kernel can be adjusted during training to maximize the quality and accuracy of the upsampling.

3.2.2. Self-attention layer. In addition to taking a convolutional layer for local detail feature extraction, the model also employs a self-attention layer for global feature extraction of the sequence. Specifically, firstly, for each position in the input feature map, three feature vectors are generated by three linear transformations: Query, Key and Value, where the weight matrices of the three linear transformations are learnable. For the Query vector of each position, an attention weight distribution is obtained by calculating the similarity with the Key vectors of all positions using dot product. After obtaining the attention weight distributions, they are multiplied with the Value vector to obtain a weighted Value vector. This weighted Value vector represents the importance of each position in the input sequence under attention, i.e., a weighted representation. If the multiple attention mechanism is used, the weighted Value vector is summed to obtain the final representation.

3.2.3. Spectral normalization. The model applies spectral normalization [34] to make training more stable. In the training of generative adversarial networks, particular attention needs to be paid to whether the network satisfies the Lipschitz continuity condition. Lipschitz continuity of a model means that the rate of change of a function within a certain interval does not exceed the product of the length of that interval and a constant. More specifically, if the function $f(x)$ satisfies the condition within the interval $[a,b]$:

$$\left|f\left(x_1\right)-f\left(x_2\right)\right| \leq L\left|x_1-x_2\right|\left(x_1, x_2 \in [a,b]\right) \tag{3}$$

where L is a constant, then the function $f(x)$ is said to be Lipschitz continuous in the interval $[a,b]$ and the constant L is called the Lipschitz constant of the function. In generative adversarial networks, if the output function of the discriminator network does not satisfy the Lipschitz continuity condition, it may lead to unstable distribution of samples produced by the training of unstable generator network or the discriminator network may not be able to classify the input samples correctly.

Spectral normalization technique is introduced to solve the Lipschitz problem in generative adversarial networks. Specifically, Spectral Normalization technique is introduced in each layer of the discriminator and generator networks to ensure that the network satisfies the Lipschitz continuity condition by restricting the maximum eigenvalue of the eigenvectors in each layer of the network, which keeps the maximum singular value of the weight matrices within a fixed range, thus making the training process more stable.

3.2.4. Residual Block Construction. The model introduces the residual module in order to prevent the model degradation problem. In deep neural networks, problems such as gradient explosion can make it difficult for backpropagation algorithms to train deep networks, and there will be a phenomenon where the performance on the training set is instead degraded, known as the model degradation problem. Residual network is a deep neural network structure that solves the gradient vanishing and network degradation problems in deep networks by introducing cross-layer connections and residual modules [35]. The residual module can capture significant differences between inputs and outputs by learning the residuals, thus reducing information loss and degradation. The structure of the generator and discriminator residual module in the model is shown in Fig. 2. The generator residual block consists of two transposed convolutional layers, one convolutional layer, and three filters of the same size and Leaky ReLU activation function. The discriminator residual block consists of three convolutional layers along with three filters of the same size and Leaky ReLU activation function.

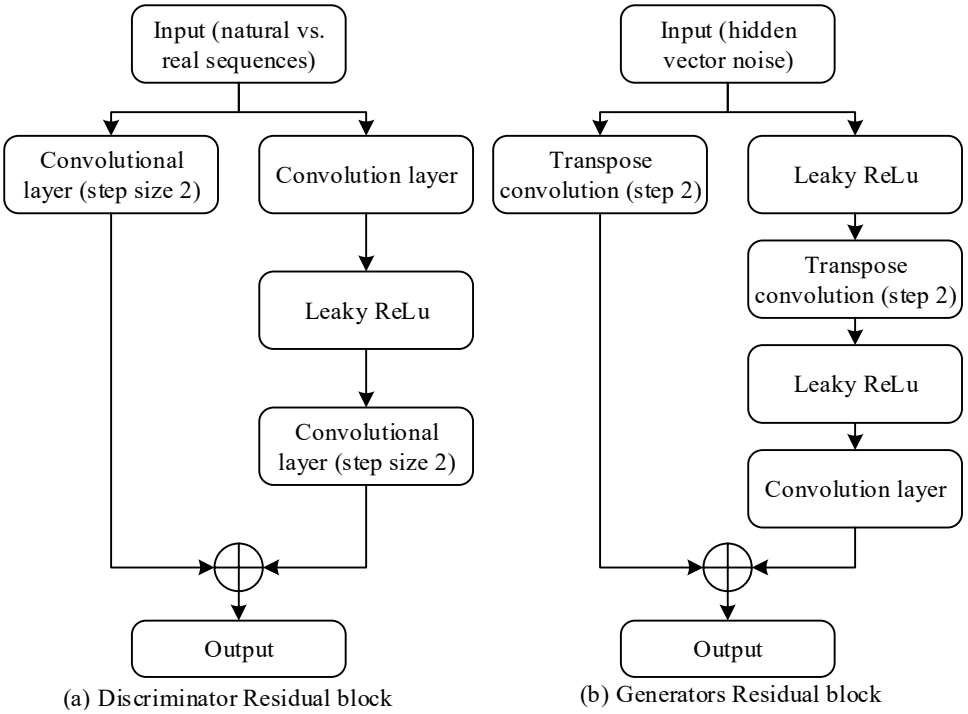

Fig. 2. The Architecture of Model Residual Blocks

3.2.5. Gumbel Softmax Trick. The model solves the problem of difficult backpropagation of the network gradient during training by using Gumbel Softmax Trick to turn the discrete output into a continuous probability distribution. Generative Adversarial Networks are often used to generate data with continuous pixel values such as images, and face difficulties when discrete data samples such as amino acid sequences need to be generated.

Gumbel Softmax Trick uses the Gumbel distribution for continuous sampling as a way to achieve discrete sampling over a probability distribution. Specifically, the Gumbel Softmax function is used in the generator network to map the output of the generator to a discrete distribution.The Gumbel Softmax function is denoted as:

$$y_i = soft\max\left(\frac{\log p_i + g_i}{\tau}\right) \tag{4}$$

where p_i is the probability of the original discrete distribution, g_i is the noise sampled from the Gumbel distribution, and τ is a parameter used to control the diversity of the generated samples. This formula can be viewed as replacing the way the discrete distribution is taken to be a Gumbel-Softmax distribution, allowing the model to be sampled and back-propagated using a continuous, differentiable approach, resulting in better generation of discrete data.

4. Artificial intelligence-based site-specific integrase application research

In this section, the artificial intelligence-based site-specific integrase generation model will be experimentally analyzed on the site-specific integrase dataset, and the site-specific integrase generated by the model will also be investigated for specific applications to verify the effectiveness of the proposed generation model.

4.1. Experimental environment

This experiment is based on two systems, Ubuntu18.04 and Windows 11 system, Windows system is mainly used for model training and Ubuntu18.04 is mainly used for testing. The specific configurations are as follows:

1) Windows experiment configuration:
System: Windows 11
Memory: 16GB
CPU: 12th Gen Intel(R) Core(TM) i5-12490F
CPU: NVIDIA GeForce RTX 2070
Python: 3.9.16

2) Ubuntu experimental environment configuration:
System: Ubuntu 18.04.1
CPU: Intel(R) Xeon(R) CPU E7-4820 v2@2.00GHz
CPU cores: 16
Memory: 16GB
Hard disk: 100GB
Clang/LLVVM: 11.0.0
AFL: 2.57b

4.2. Site-specific integrase dataset construction

The training data were searched in UniProt from January 5, 2024, and site-specific integrases were filtered by Taxonomy, totaling 15,215 sequences. Sequences with length greater than 600 amino acids and those containing non-standard amino acids were removed, and a total of 14,511 sequences were used for training, and the length distribution of site-specific integrases is shown in Figure 3. The length distribution of site-specific integrase sequences was within 580 amino acids, most of which were concentrated in the 420-490 amino acid interval (10,900 sequences). To avoid model pattern collapse, dynamic upsampling was used to balance the dataset. The sequences were first analyzed by clustering with the MMSeq2 tool, which clustered these sequences into 70% of the labeled clusters. Then, an up-sampling operation was performed on the sequences based on the number of sequences in the clusters. Data with less than 3 number of sequences in the clusters were used for validation (295 sequences) and the rest of the data were used for training (14216 sequences).

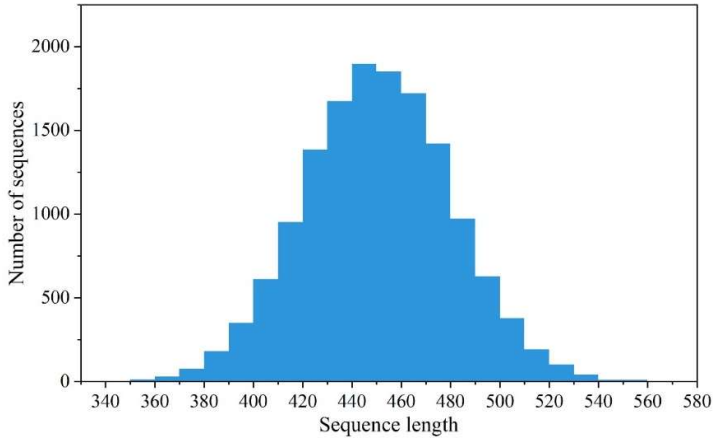

Fig. 3. Locus of specific integrase length distribution

4.3. Generating Sequence Consistency and Biological Evaluation

During the training process, 50 sequences are generated every 1000 steps for use with the training set and validation set using BLAST to compute sequence consistency, evaluating the current training status of the model based on the trend of change in the consistency of the generated sequences, and monitoring the trend of change in this variable using TensorBoard. Where BLAST sets BLOSUM45 matrix to calculate sequence consistency and uses Matplotlib library in Python to visualize the change process. Based on the training trend graph, it can be judged whether the model captures the distributional characteristics of the natural training set during the training process, so that the training strategy can be adjusted at any time and the model convergence can be evaluated.

The consistency trend during the training process of site-specific integrase is shown in Figure 4. From the data depicted in the figure, the generated sequence of the site-specific integrase eventually reached 92% consistency with the training set and 76% consistency with the validation set. Since only class II sequences were selected for its training, the training dataset is more conservative and the final generated sequences have a higher consistency with the training set. From the figure, it can be seen that the consistency between the generated sequences of site-specific integrase and the natural sequences all increased gradually during the training process, indicating that the model converged normally and the quality of its generated sequences increased gradually.

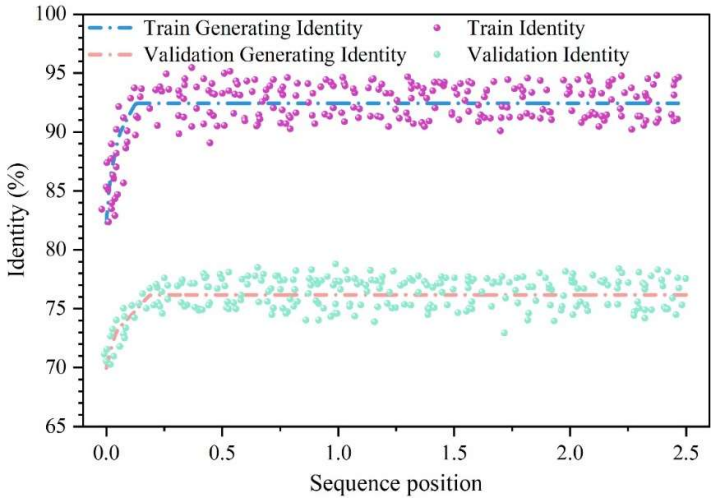

Fig. 4. The consistency trend in the specific integrase training process

Using the model to generate sequences that are equivalent to the natural sequences used for training, the amino acid composition of the generated sequences and the natural sequences (sequences originating from nature in the training set) were counted and compared to each other, and the Pearson correlation coefficients of the two were calculated. BioPython was used to count the amino acid species of the sequences, and the Pearson correlation coefficient was calculated using the machine learning library Scipy, with the following formula:

$$P=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \tag{5}$$

where X_i and Y_i are the ird observation of the two variables, $\bar{X}$ and $\bar{Y}$ are the mean values of the two variables, and n is the number of observations.

The calculation of amino acid groups with different attributes is shown in Table 1. Among the possible types of amino acids included in the amino acid group are: glycine (G), alanine (A), valine (V), leucine (L), isoleucine (I), proline (P), phenylalanine (F), tryptophan (W), methionine (M), serine (S),

threonine (T), asparagine (N), glutamine (Q), cysteine (C), tyrosine (Y), Lysine (K), Arginine (R), Histidine (H), Aspartic acid (D), Glutamic acid (E).

Table 1. The calculation of amino acid groups of different properties

Amino acid group	Computational mode
Hydrophilic	The proportion of Amino acid V, I, L, F, W, Y, M
Hydrophobic	The proportion of Amino acid S, T, H, N, Q, E, D, K, R
Aromatic	The proportion of Amino acid F, W, Y
Small	The proportion of Amino acid G, A, S
Positive	The proportion of Amino acid K, R, H
Negative	The proportion of Amino acid D, E
Aliphatic	The proportion of Amino acid V, I, L, M
Hydroxyl sulfur	The proportion of Amino acid S, C, T, M
Polar	The proportion of Amino acid S, T, C, M, N, Q
Charged	The proportion of Amino acid H, K, R, E, D

The amino acid composition of the site-specific integrase generation sequence and the natural sequence is shown in Figure 5. By analyzing the amino acid composition of the generated sequences by groups, it can be seen that the site-specific integrase generated sequences are highly correlated with the amino acid composition of the natural sequences. The Pearson value of site-specific integrase was calculated to be greater than 0.8. It can be seen that the model captured the amino acid composition of the natural sequences in the training set well.

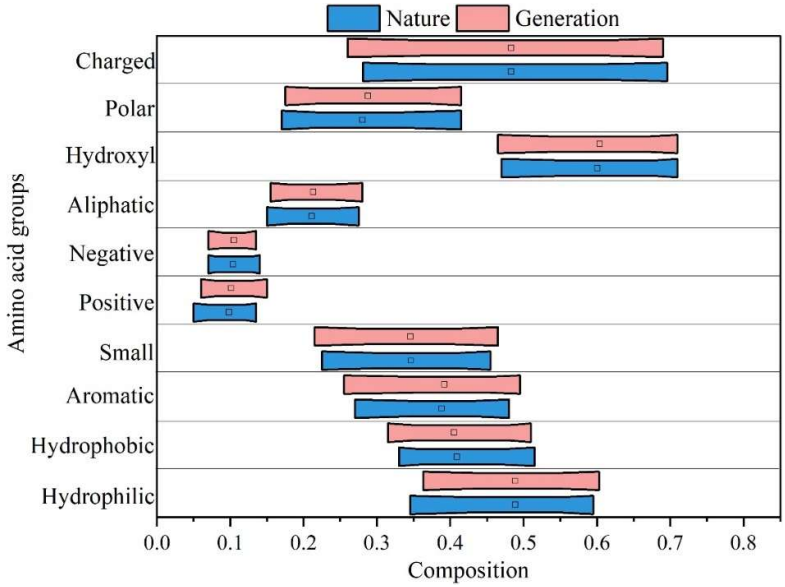

Fig. 5. Sequence amino acid composition comparison results

Sequences equivalent to nature were generated using the trained converged model and then the natural and generated sequences were clustered separately using the MMSeqs2 tool for both the natural and generated sequences with thresholds 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 0.9, 0.95 and 0.9 respectively. 0.9, 0.95 were clustered, visualized using Matplotlib, respectively, and the ratio of the number of both at different thresholds was calculated to evaluate their extension for the sequence space i.e. diversity. The diversity of model patterns was evaluated using t-SNE method visualization.

The results of the diversity comparison between the site-specific integrase generated sequences and the natural sequences of the training set are shown in Figure 6. As can be seen from the figure, under the condition of equal number of sequences, the number of clusters clustered by the generated sequences under each threshold is higher than that of the natural sequences, and the model in this paper greatly expands the sequence space. Among them, the highest number of clusters of site-specific integrase generating sequences is 4.7 times higher than that of nature. It indicates that the artificial intelligence technology is of great significance for the spatial expansion of functional proteins.

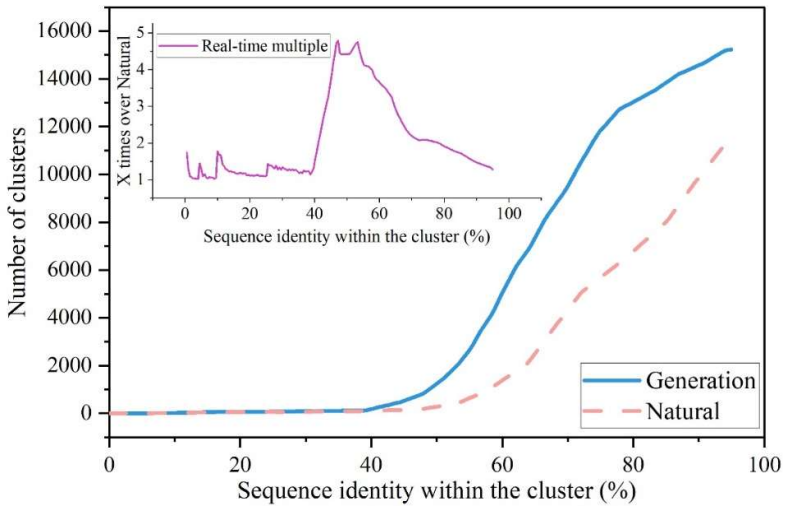

Fig. 6. Diversity of sequence and training set natural sequences

4.4. Applied Research and Analysis of Results

In this section, the site-specific integrase generation model constructed in this paper will be utilized in different fields to generate relevant enzymes applicable to the field and apply them to specific cases to evaluate the effectiveness of AI-based site-specific integrase in different fields.

4.4.1. Immune function and protein expression in lung cancer patients. The incidence rate of non-small cell lung cancer accounts for about 4/5 of lung cancer, with low long-term survival rate and poor prognosis, seriously threatening patients' life safety. Kang'ai injection is a proprietary Chinese medicine preparation, which has the efficacy of improving the body's immune function, detoxification and anticancer. In this study, we added site-specific integrase to Kang'ai injection to analyze the effects of site-specific integrase on the immune function and protein expression of non-small cell lung cancer patients.

Forty cases of non-small cell lung cancer patients were the study subjects, and they were divided into two groups of 20 cases each according to the method of random number table. In the control group (recorded as CK), there were 14 male cases and 6 female cases. Their ages ranged from 37 to 75 years, with an average age of (59.26±4.17) years. The experimental group (recorded T) had 16 male and 4 female cases. Age 38~77 years, mean age (60.08±3.21) years. All patients in this study signed informed consent. Before treatment, the differences in general information, immune function indexes and protein expression between the two groups were not statistically significant ($P > 0.05$) and were comparable.

The control group was given polyene paclitaxel, gemcitabine, and common kangai injection for treatment. The experimental group was treated by adding site-specific integrase on the basis of the control group. The method was as follows: 60 ml of Kangai injection (Changbaishan Pharmaceutical Co., Ltd., State Drug License Z20026868) was injected into 250 ml of sodium chloride solution with a solubility of 0.5% 1 d before chemotherapy, and then injected intravenously once/d. Both groups were

treated for 2 consecutive courses with 21 d as a course of treatment. Measurement indexes and their measurement methods are as follows:

1) Immune function: 3 ml of fasting venous blood was drawn from patients before treatment and after 2 courses of treatment, centrifuged and processed, and then the levels of T-cell subpopulations CD_4^+ and CD_8^+, which are indicators of immune function, were detected by using flow cytometry.

2) Protein expression: lung cancer specimens were collected, routinely embedded and sectioned, and the levels of bax protein, bax-2 protein and Ki67 protein were detected by immunohistochemical staining.

The results of the comparison of immune function indexes between the two groups of patients after treatment are shown in Figure 7. As can be seen from the figure, the average CD_4^+ score of the control group after treatment was 35.69±1.17 lower than that of the experimental group, which was 48.44±1.42, while on the contrary, the average CD_8^+ score of the control group was higher than that of the experimental group, and the p-value between the comparative indexes was less than 0.05, and the difference was statistically significant ($p<0.05$).

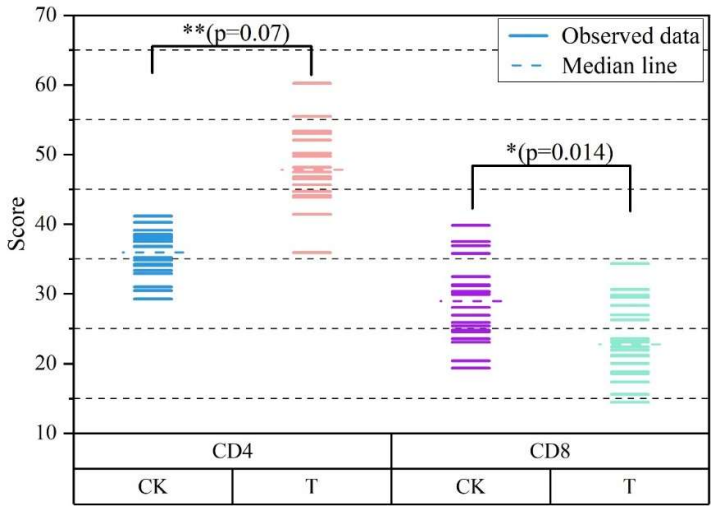

Fig. 7. The results of the treatment of immunological function were compared

Figures 8, 9, and 10 demonstrate the results of the comparison of the expression of bax protein, bax-2 protein, and Ki67 protein in the two groups of patients, respectively. The results of the protein expression study showed that the bax proteins of the patients in the CK group were all lower than those in the T group, while the bax-2 protein and Ki67 protein were generally higher than those in the observation group. Specifically, the average scores for bax protein, bax-2 protein, and Ki67 protein expression were 4.57, 5.40, and 33.1 in the CK group and 5.74, 4.48, and 26.45 in the T group, respectively. The p-values of bax protein, bax-2 protein, and Ki67 protein expression scores between the control group and the T group ranged from 0.015 to 0.033, which were all less than 0.05, and the difference was statistically significant ($P < 0.05$).

The above results show that the addition of site-specific integrase to Kang Ai injection on the basis of chemotherapy for non-small cell lung cancer patients can effectively improve the immune function of the body, reduce the side effects of therapy, and facilitate the treatment of patients' conditions.

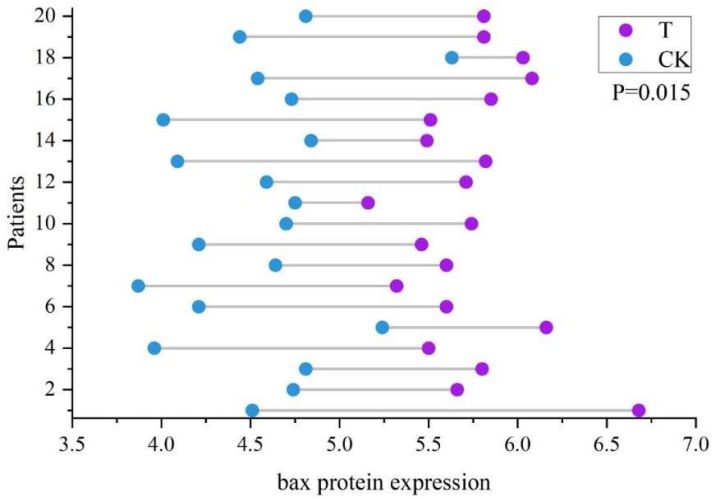

Fig. 8. The comparison of the expression of bax protein

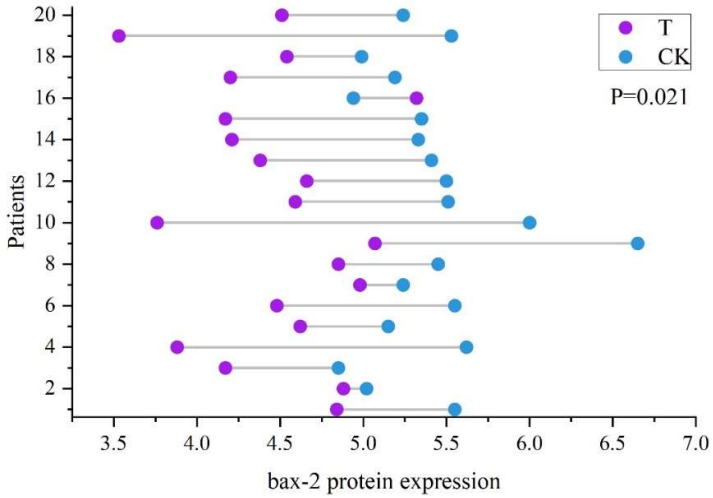

Fig. 9. The comparison of the expression of bax-2 protein

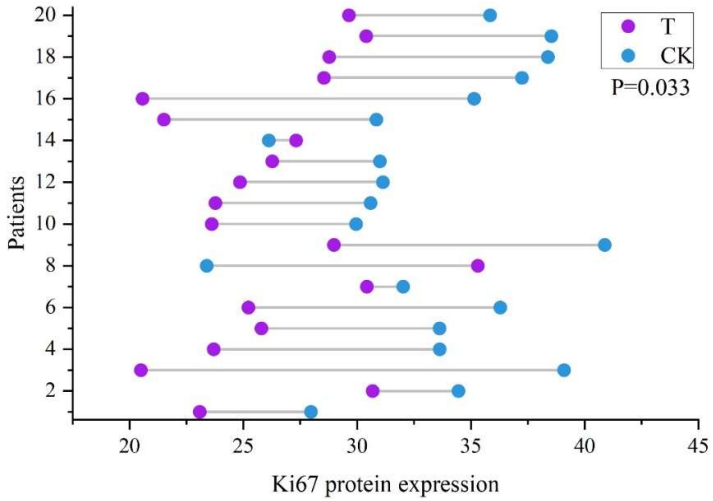

Fig. 10. The comparison of the expression of Ki67 protein

4.4.2. Ovarian steroidogenic gene expression in ewes. In this experiment, 40 ewes with an average body weight of (30±3.41) kg and a healthy body condition of about 2 weeks old were selected and

randomly divided into 2 groups: the control group (recorded as CK, fed with basal diet) and the experimental group (recorded as T, with an appropriate amount of site-specific integrase added to the basal diet). There were 3 replicates in each group, and 5 sheep in each replicate were kept in a separate pen and fed at 07:00 and 16:00 every day, respectively, and the pre-feeding period of the experiment was 7 d. The experimental group was divided into 2 groups: control group (recorded as CK) and experimental group (recorded as T).

In each group, 5 ewes were randomly selected for blood collection and serum separation, and the follicular development of ewes was observed and ovarian tissues were collected using laparoscopic + surgical method. According to the instructions of the kit, total RNA was extracted from ovarian tissues of the two groups using Trizol kit, and the quality of the RNA samples was examined, and the samples with D260nm/D280nm values ranging from 1.8 to 2.0 were selected for further analysis.

The expression of four genes related to ovarian steroidogenesis in ewes: STAR, CYP11A1, CTP19A1, and 3β-HSD, were counted. Each test was repeated three times, and the results of the effect of site-specific integrase on the expression of ovarian steroidogenesis-related genes in sheep were obtained as shown in Figure 11. It can be seen from the figure that compared with the control group, the expression of genes related to ovarian steroid production in ewe was increased under the intervention of site-specific integrase, and the relative mRNA expression levels of STAR, CYP11A1, CTP19A1 and 3β-HSD genes were 1.897, 1.844, 1.93 and 2.078, respectively, indicating that the expression of genes related to ovarian steroid production in ewe was up-regulated by site-specific integrase.

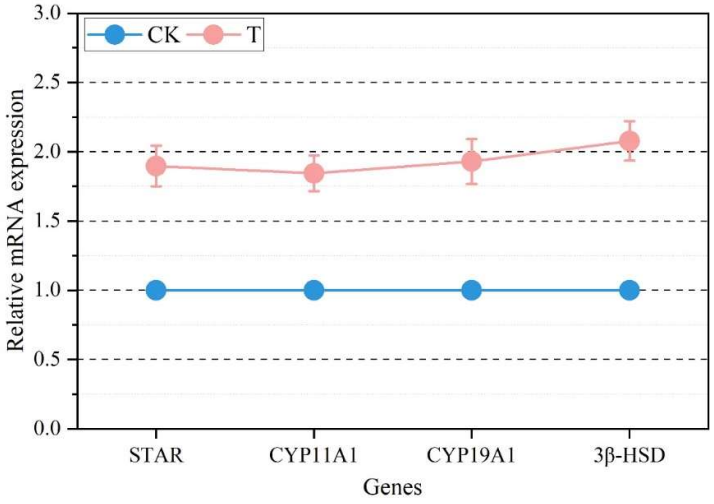

Fig. 11. The effects of the genetic expression of the genes of the ovarian steroid

4.4.3. Effectiveness of degradation of oil pollution in soil environment. In this section, two microbial strains for efficient degradation of oil pollution were purchased from a biological research institute and were noted as strain A and strain B. In order to assess the value of artificial intelligence-based site-specific integrase for environmental applications, the enzyme was used to optimize the metabolic pathways of the two microorganisms and to compare the degradation effects of the two microorganisms on oil pollution before and after the optimization.

For experimental analysis, soil samples were collected from four uncontaminated areas, which were distributed in different vegetation zones to ensure sample diversity. During the experiments, an appropriate amount of water was added to the soil samples to keep the water content consistent and eliminate the influence of water content differences on the experimental results to ensure accurate and reliable data. In order to simulate different levels of oil contamination, each sample was divided into three equal parts, and different amounts of oil were added, and the amount of oil added was precisely controlled to realize three groups of preset oil concentrations, namely high (18000mg/kg), medium (10000mg/kg), and low (2500mg/kg).

The oil-degrading bacteria incorporating site-specific integrase were added in equal amounts to the mudflat soils of the high, medium and low concentration groups and ensured to be well mixed.

Incubated soil samples were collected at predetermined time points of the experiment, i.e., 0, 2, 4, 7, 10, 14, 18, 23, and 28 d, respectively. In order to accurately monitor the residual levels of various types of alkanes (TPH) in the soil, a high performance gas chromatograph was used.

Figure 12 shows the results of the degradation experiments of the strains against the high concentration of petroleum pollution environment. The data show that in the initial stage of the degradation process, strain A alone as well as strain B exhibited a significant degradation rate. However, the addition of site-specific integrase to the strains resulted in faster degradation rates. This difference in degradation rate resulted in a 57.25% to 63.00% lower alkane content in the group of strains incorporating site-specific integrase compared to those without the enzyme at the end of the 28-d experimental cycle. The two strains incorporating site-specific integrase showed similar degradation of petroleum throughout the experimental period.

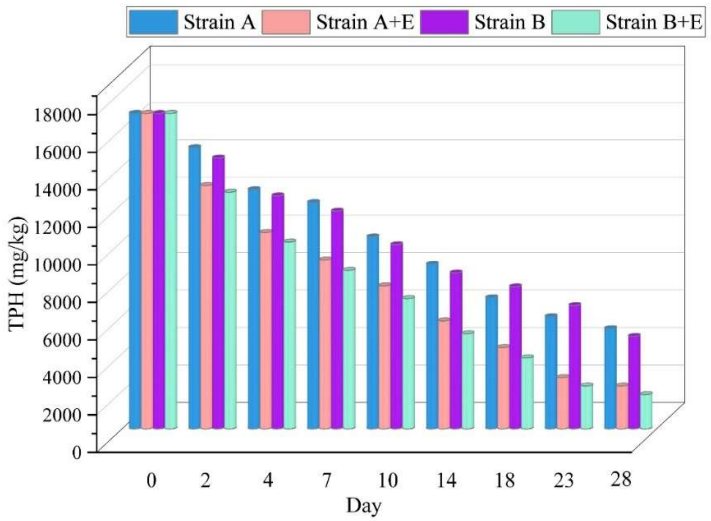

Fig. 12. Degradation of high concentrations of oil pollution at any time

Figure 13 shows the degradation effect of the strains under medium concentration of oil pollution. According to the experimental data, it can be found that the concentration of petroleum pollution was reduced from the initial 10,000 mg/kg to about 4,500 mg/kg for strains A and B after 28 d of treatment. And strains A and B incorporating site-specific integrase showed a higher degradation rate in the first four days. At the end of the 28d experimental cycle, it could be observed that the oil contamination concentration of strain A + enzyme and strain B + enzyme decreased to about 2000mg/kg, which was significantly lower than that of strain A and strain B.

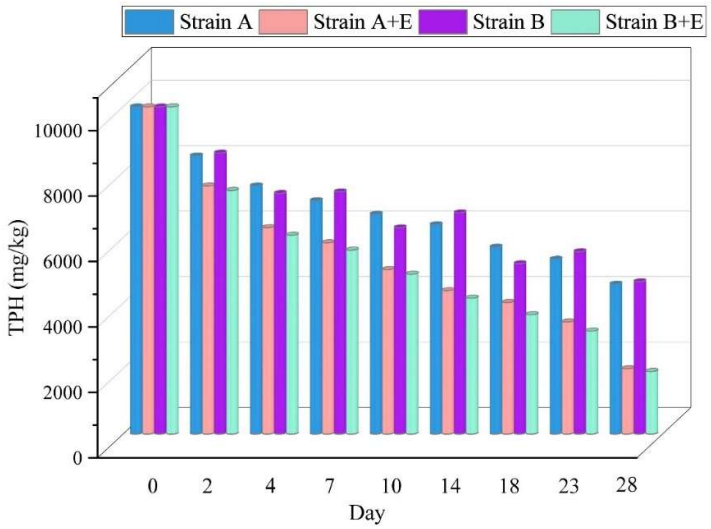

Fig. 13. Degradation of medium concentration of oil pollution at any time

Figure 14 demonstrates the degradation effect of the strains in the environment with low concentration of pollution. As can be seen from the figure, strain A and strain B, the concentration of TPH changed less after 14d and basically stabilized above and below 1050 mg/kg. While the strain with the addition of site-specific integrase still had obvious degradation effect on oil pollution after 14d, and the concentration of TPH finally decreased to about 400mg/ at 28d of the experiment.

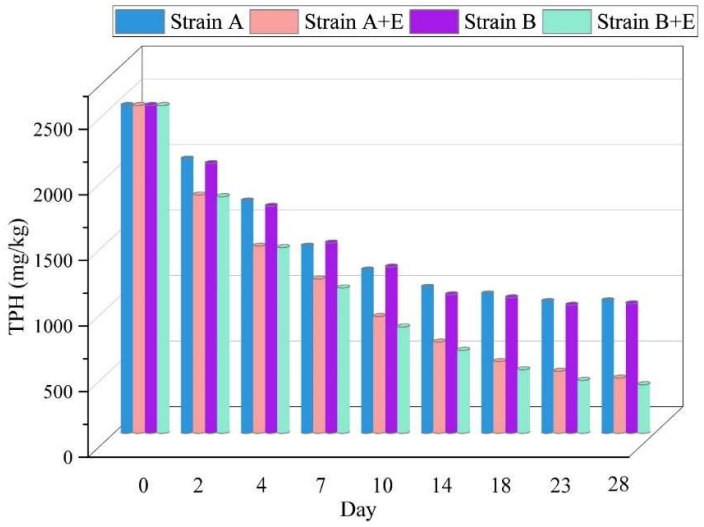

Fig. 14. Degradation of low concentration of oil pollution at any time

5. Conclusion

In this paper, we have mined the data features from the site-specific integrase dataset through feature extraction, selection and discretization, and used generative network models to generate site-specific integrases applicable to different domains. The model utilizes convolutional and self-attention layers to extract local and global features of the data by adjusting the convolutional kernel weights as well as calculating the attention weights of each vector, respectively. Spectral normalization technique is introduced during the training process to make the network satisfy the Lipschitz continuity condition and improve stability. Residual networks are utilized to reduce information loss and degradation. Also use Gumbel distribution for continuous sampling and realize discrete sampling on probability

distribution. The generated site-specific integrase is applied to specific studies such as protein expression and oil degradation to explore its application value.

1) The consistency of the AI-generated sequences with the training and validation sets is 92% and 76%, respectively, and the amino acid composition is highly correlated with the natural sequences, with a Pearson value greater than 0.8, and the AI technology extends the sequence space.

2) The site-specific integrase based on artificial intelligence can reduce CD4+ and increase CD8+ in lung cancer patients, and can affect the expression of their bax protein, bax-2 protein and other proteins.

3) The relative mRNA expression of STAR, CYP11A1, CTP19A1, and 3β-HSD genes in the ovary of ewes under the influence of site-specific integrase was 1.897, 1.844, 1.93, and 2.078, respectively, which was significantly increased.

4) Site-specific integrase promotes the degradation effect of the strain on oil in soil.

References

[1] Shakweer, W. M. E., Krivoruchko, A. Y., Dessouki, S. M., & Khattab, A. A. (2023). A review of transgenic animal techniques and their applications. Journal of Genetic Engineering and Biotechnology, 21(1), 55.

[2] Stepanenko, A. A., & Heng, H. H. (2017). Transient and stable vector transfection: Pitfalls, off-target effects, artifacts. Mutation Research/Reviews in Mutation Research, 773, 91-103.

[3] Shi, B., Xue, M., Wang, Y., Wang, Y., Li, D., Zhao, X., & Li, X. (2018). An improved method for increasing the efficiency of gene transfection and transduction. International journal of physiology, pathophysiology and pharmacology, 10(2), 95.

[4] Tan, S., Tao, Z., Loo, S., Su, L., Chen, X., & Ye, L. (2019). Non-viral vector based gene transfection with human induced pluripotent stem cells derived cardiomyocytes. Scientific reports, 9(1), 14404.

[5] Chong, Z. X., Yeap, S. K., & Ho, W. Y. (2021). Transfection types, methods and strategies: a technical review. PeerJ, 9, e11165.

[6] Salvi, V. R., & Pawar, P. (2019). Nanostructured lipid carriers (NLC) system: A novel drug targeting carrier. Journal of Drug Delivery Science and Technology, 51, 255-267.

[7] Rabea, S., Alanazi, F. K., Ashour, A. E., Salem-Bekhit, M. M., Yassin, A. S., Moneib, N. A., ... & Haq, N. (2020). Salmonella-innovative targeting carrier: Loading with doxorubicin for cancer treatment. Saudi Pharmaceutical Journal, 28(10), 1253-1262.

[8] Ahmad, M. M., Ali, A., Siddiqui, S., Kamaluddin, & Abdin, M. Z. (2017). Methods in transgenic technology. Plant biotechnology: principles and applications, 93-115.

[9] Olorunniji, F. J., Rosser, S. J., & Stark, W. M. (2016). Site-specific recombinases: molecular machines for the Genetic Revolution. Biochemical Journal, 473(6), 673-684.

[10] Guiziou, S., Maranas, C. J., Chu, J. C., & Nemhauser, J. L. (2023). An integrase toolbox to record gene-expression during plant development. Nature communications, 14(1), 1844.

[11] Stark, W. M. (2017). Making serine integrases work for us. Current opinion in microbiology, 38, 130-136.

[12] Merrick, C. A., Zhao, J., & Rosser, S. J. (2018). Serine integrases: advancing synthetic biology. ACS synthetic biology, 7(2), 299-310.

[13] Pandey, S., Gao, X. D., Krasnow, N. A., McElroy, A., Tao, Y. A., Duby, J. E., ... & Liu, D. R. (2024). Efficient site-specific integration of large genes in mammalian cells via continuously evolved recombinases and prime editing. Nature Biomedical Engineering, 1-18.

[14] Song, X., Li, Y., Xia, Y., Wang, G., Xiong, Z., Yang, Y., & Ai, L. (2024). Phage tyrosine integrase-mediated multi-sites genome engineering in Lacticaseibacillus casei. Food Bioscience, 58, 103694.

[15] Li, H., Sharp, R., Rutherford, K., Gupta, K., & Van Duyne, G. D. (2018). Serine integrase attP binding and specificity. Journal of molecular biology, 430(21), 4401-4418.

[16] Yoon, B., Kim, I., Nam, J. A., Chang, H. I., & Ha, C. H. (2016). In vivo and in vitro characterization of site-specific recombination of a novel serine integrase from the temperate phage EFC-1. Biochemical and biophysical research communications, 473(1), 336-341.

[17] Fan, H. F., Hsieh, T. S., Ma, C. H., & Jayaram, M. (2016). Single-molecule analysis of ϕC31 integrase-mediated site-specific recombination by tethered particle motion. Nucleic Acids Research, 44(22), 10804-10823.

[18] MacDonald, A. I., Baksh, A., Holland, A., Shin, H., Rice, P. A., Stark, W. M., & Olorunniji, F. J. (2024). Variable orthogonality of RDF–large serine integrase interactions within the ϕC31 family. bioRxiv.

[19] MacDonald, A. I., Baksh, A., Holland, A., Shin, H., Rice, P. A., Stark, W. M., & Olorunniji, F. J. (2024). Variable orthogonality of serine integrase interactions within the ϕC31 family. Scientific Reports, 14(1), 26280.

[20] Fogg, P. C., Younger, E., Fernando, B. D., Khaleel, T., Stark, W. M., & Smith, M. C. (2018). Recombination directionality factor gp3 binds ϕC31 integrase via the zinc domain, potentially affecting the trajectory of the coiled-coil motif. Nucleic Acids Research, 46(3), 1308-1320.

[21] Landy, A. (2015). The λ integrase site-specific recombination pathway. Mobile DNA III, 91-118.

[22] Olorunniji, F. J., Lawson-Williams, M., McPherson, A. L., Paget, J. E., Stark, W. M., & Rosser, S. J. (2019). Control of ϕC31 integrase-mediated site-specific recombination by protein trans-splicing. Nucleic Acids Research, 47(21), 11452-11460.

[23] Jusiak, B., Jagtap, K., Gaidukov, L., Duportet, X., Bandara, K., Chu, J., ... & Lu, T. K. (2019). Comparison of integrases identifies Bxb1-GA mutant as the most efficient site-specific integrase system in mammalian cells. ACS Synthetic Biology, 8(1), 16-24.

[24] Hew, B. E., Gupta, S., Sato, R., Waller, D. F., Stoytchev, I., Short, J. E., ... & Owens, J. B. (2024). Directed evolution of hyperactive integrases for site specific insertion of transgenes. Nucleic Acids Research, 52(14), e64-e64.

[25] Hosur, V., Low, B. E., & Wiles, M. V. (2022). Programmable RNA-guided large DNA transgenesis by CRISPR/Cas9 and site-specific integrase Bxb1. Frontiers in Bioengineering and Biotechnology, 10, 910151.

[26] Chi, X., Zheng, Q., Jiang, R., Chen-Tsai, R. Y., & Kong, L. J. (2019). A system for site-specific integration of transgenes in mammalian cells. PLoS one, 14(7), e0219842.

[27] Suzuki, S., Yoshikawa, M., Imamura, D., Abe, K., Eichenberger, P., & Sato, T. (2020). Compatibility of site-specific recombination units between mobile genetic elements. Iscience, 23(1).

[28] Pokhilko, A., Zhao, J., Stark, W. M., Colloms, S. D., & Ebenhöh, O. (2017). A simplified mathematical model of directional DNA site-specific recombination by serine integrases. Journal of The Royal Society Interface, 14(126), 20160618.

[29] Ba, F., Zhang, Y., Wang, L., Liu, W. Q., & Li, J. (2023). Applications of serine integrases in synthetic biology over the past decade. SynBio, 1(2), 172-189.

[30] Abu Dabous Saleh,Alzghoul Ahmad & Ibrahim Fakhariya. (2025). Intelligent condition prediction model for bridge infrastructure based on evaluating machine learning algorithms. Smart and Sustainable Built Environment(2),557-576.

[31] A. Somaiah,B. Anjaneya Prasad & N. Kishore Nath. (2025). Taguchi-Based Optimization and ANOVA Analysis of Drilling Parameters for Enhanced Hole Quality in GFRP Composites with MgO and TiO2 Nanofillers. Materials Circular Economy(1),5-5.

[32] Fangfang Tang,Mengyuan Ren,Xiaofan Li,Zhanglin Lin & Xiaofeng Yang. (2023). Generating Novel and Soluble Class II Fructose-1,6-Bisphosphate Aldolase with ProteinGAN. Catalysts(12).

[33] Chih Hsiung Chen,Kuang Yu Hsieh,Kuo En Huang & En Tsung Cheng. (2025). Using the Regression Slope of Training Loss to Optimize Chest X-ray Generation in Deep Convolutional Generative Adversarial Networks. Cureus(1),e77391.

[34] Shuqing Jia,Ronghua Zhang,Wenying Fu,Jinxun Le,Xuefeng Bai & Boyang Li. (2025). Image reconstruction of electromagnetic tomography based on generative adversarial network with spectral normalization and improved dung beetle optimization algorithm. The Review of scientific instruments(2).

[35] Yuanchao Xue,Yangsheng Hu,Yu Yao,Jie Huang,Haitao Wang & Jianfeng He. (2025). MSRMMP: Multi-scale residual module and multi-layer pseudo-supervision for weakly supervised segmentation of histopathological images. Medical Engineering and Physics104284-104284.