Bioresource Technology 388 (2023) 129758

Contents lists available at ScienceDirect

Bioresource Technology

journal homepage: www.elsevier.com/locate/biortech

A new paradigm in lignocellulolytic enzyme cocktail optimization: Free from expert-level prior knowledge and experimental datasets

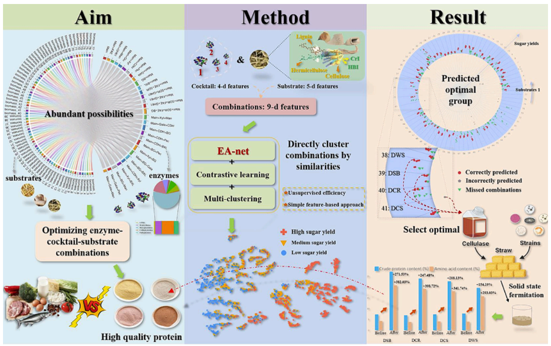
Le Gao [a,1], Zhuohang Yu [b,1], Shengjie Wang [a], Yuejie Hou [b], Shouchang Zhang [b], Chichun Zhou [b,*], Xin Wu [a,*]

[a] Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, National Technology Innovation Center of Synthetic Biology, Tianjin 300308, China
[b] School of Engineering, Dali University, Dali, Yunnan 671003, China

HIGHLIGHTS

- A deep-learning model could predict optimal enzyme-cocktail-substrate interaction.
- The model needs no pre-labeled datasets, with easily obtainable features.
- The model eliminating the reliance on expert-level prior knowledge of mechanism.
- The model shows high precision of 91.98% in tailor-made lignocellulolytic enzyme.
- The method has good applications in artificial proteins biosynthesis from straw.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Tailor-made lignocellulolytic enzyme cocktails
Enzyme-cocktail-substrate interactions
Deep-learning method
Without prior knowledge
Artificial proteins biosynthesis

ABSTRACT

Effectively pairing diverse lignocellulolytic enzyme cocktails with intricately structured lignocellulosic substrates is an enduring challenge for science and technology. To date, extensive trial-and-error remains the primary approach and no deep-learning methods were developed to address it due to limited experimental data and incomplete expert-level knowledge of enzyme-cocktail-substrate structure-dynamics-function relationships. Here, a novel model is developed to tackle this issue in efficient, cost-effective, and high-throughput manners. It needs no pre-labeled datasets, instead utilizing simple features, eliminating the reliance on expert-level prior knowledge of reaction mechanisms. Experimentally optimal combinations were found within predicted ranges of tailor-made combinations with precision of 91.98%, covering 80.00% of overall top-100. Practical tests demonstrated its effectiveness in narrowing down potential optimal combinations, speeding up targeted screening, and enabling efficient degradation of lignocellulosic biomass. The method has good applications in artificial proteins biosynthesis from low-value lignocellulosic straw, providing alternative solutions for biomass biorefining challenges in complex enzyme-cocktail-substrate interactions.

[*] Corresponding authors.
E-mail addresses: zhouchichun@dali.edu.cn (C. Zhou), wuxin@tib.cas.cn (X. Wu).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.biortech.2023.129758
Received 4 August 2023; Received in revised form 5 September 2023; Accepted 7 September 2023
Available online 17 September 2023
0960-8524/© 2023 Elsevier Ltd. All rights reserved.

1. Introduction

The global lignocellulosic biomass production amounts to approximately 200 billion metric tons annually (Ashokkumar et al., 2022). To achieve food supply sustainability and reduce dependence on natural resources, the effective utilization of abundant agricultural waste straw biomass for food or fuel production is of paramount importance. However, the conversion of carbohydrates in biomass into fermentable sugars poses a significant challenge for lignocellulose-based biorefining (Ragauskas et al., 2006). The low digestibility of lignocellulosic biomass and the high cost of recombinant enzymes have impeded the progress in this field (Liu and Qu, 2021).

To advance lignocellulose-based biorefining, it is crucial to develop specifically tailored multienzyme mixtures capable of efficiently degrading specific lignocellulosic materials (Binod et al., 2019). Lignocellulosic biomass is composed of cellulose, hemicelluloses, and lignin, each possessing distinct characteristics due to their composition, interlinkage, and distribution within the plant cell wall (Binod et al., 2019). Due to the complexity of biomass, the degradation of lignocellulosic biomass in natural living systems requires a complex multienzyme system, including not only cellulases but also other auxiliary enzymes such as xylanases and peroxidases, among others (Hu et al., 2011; Kumar and Wyman, 2014; Murashima et al., 2003; Selig et al., 2008; Zhang et al., 2011). Notable companies such as Novozymes and Genencor have made significant strides in creating more efficient and cost-effective enzymes for lignocellulose hydrolysis. However, these advanced products have been tailored to specific substrates and may not be universally applicable (Wang et al., 2019). Furthermore, the deconstruction of lignocellulose involves over 30 classes of enzymes produced by microorganisms, resulting in a vast number of potential enzyme combinations (Liu et al., 2013). As a result, optimizing enzyme cocktails for different substrates remains a significant challenge (Du et al., 2020) To date, the development of a systematic approach to match optimal lignocellulolytic enzymes to specific substrates has primarily relied on time-consuming, labor-intensive, and expensive experimental methods, which are often not economically viable according to techno-economic analyses. Therefore, there is an urgent need for high-throughput methods to select optimal lignocellulolytic enzyme cocktails for the utilization of specific biomass resources.

The currently available supervised methods can accurately predict enzyme properties using sequence and structure features, relying on prelabeled training sets (Li et al., 2022). However, employing conventional supervised methods to predict properties of enzyme-cocktail-substrate combinations, which involve complex features of enzymes and substrates as input (Mazurenko et al., 2020), may encounter a number of challenges. Firstly, constructing such a model requires profound expert-level knowledge of reaction mechanisms (Li et al., 2022; Mazurenko et al., 2020; Smith et al., 2020), including structure-dynamics-function relationships in macromolecular assemblies of enzyme cocktails and substrates. However, the current understanding of complex macromolecular structure matching, including interactions between enzyme mixtures and the intricate structure of lignocellulosic substrates, remains limited (Jabbour et al., 2013). Additionally, models that utilize high-dimensional sequence and structure information as input, except for those designed with great delicacy, often require a larger pre-labeled training set (Mazurenko et al., 2020) to achieve optimal performance and overcome difficulties such as overfitting, which can arise due to high dimensionality (Indyk and motwani, 2000). Moreover, acquiring sequence and structure information from experiments involves additional analyses and calculations, resulting in a time-consuming and costly process needed to prepare a substantial pre-labeled training set. When dealing with specific tasks, utilizing existing training sets may pose various challenges, including domain transfer difficulties, feature gaps, and class imbalances (Mazurenko et al., 2020). Although deep learning holds promise solutions for these biorefining challenges, no deep-learning methods have been reported to specifically address this problem (effectively pairing diverse lignocellulolytic enzyme cocktails with intricately structured lignocellulosic materials) to date.

In this work, a novel deep-learning method that can tackle this issue by adopting a distinct approach from traditional supervised methods is developed. It based on two guiding ideas: (1) Enzyme-cocktail-substrate combinations with high degradation rates share common characteristics, which are already embedded in the features of enzyme cocktails and substrates. Therefore, instead of initially predicting their properties and then classifying them based on those predictions, one can directly clusters these combinations into groups based on their similarities, so that they can be extracted from simple features, including indicators of substrate composition and structural properties, as well as simple indices of the role played by each enzyme in the cocktails. (2) Unlike the conventional supervised learning methods that rely on large labeled datasets to guide the learning of important features, one can leverage unsupervised strategies to guide the model in discarding irrelevant information, retaining crucial information, and assigning higher weights to key factors.

Based on the simple ideas mentioned above, three key techniques were developed and introduced: (1) The EA-net is designed to capture the intricate interactions between substrates and enzyme cocktails using only simple features. (2) The contrastive learning strategy (Chen et al., 2020) is modified as a fully unsupervised strategy and is combined with EA-net to enable the model to learn the similarities and differences between samples, thereby discarding useless features, retaining useful ones, and strengthening important ones. (3) Finally, a multi-clustering strategy (Zhou et al., 2022) is introduced to enhance the purity of each clustered group.

The proposed method clusters enzyme-cocktail-substrate combinations into high-purity groups and provides a category consisting of predicted optimal combinations. It avoids the dilemma of limited experimental datasets and incomplete prior expert-level knowledge of structure-dynamics-function relationships crucial for the reaction mechanisms. The optimal group contains combinations with higher sugar yields, achieving a precision of 91.98%, covering 80.00% of the overall optimal 100 combinations and 89.81% of the optimal combinations for each substrate. It serves as a valuable guide for subsequent applications showcasing its effectiveness in speeding up targeted optimal enzyme-cocktail-substrate combinations screening. Finally, repeated experiments and extensive comparisons were conducted to prove its effectiveness and robustness.

By following the key ideas and leveraging the proposed key techniques, an unprecedented method was offered to expedite the conversion of biomass into fermentable sugars by efficiently identifying optimal enzyme-cocktail-substrate combinations. This method eliminates the reliance on experimental datasets and expert-level knowledge, resulting in a significant advancement, novel insights and alternative solutions for biorefining challenges beyond enzyme-cocktail-substrate interactions, with potential applications in the realm of macromolecular assemblies. Its user-friendly nature and cost-effectiveness ensure accessibility to a wide range of researchers and industrial engineers, thereby fostering innovation and progress in related fields.

2. Materials and methods

In this section, the main methods, including the criteria of the data, the construction of the unsupervised deep learning framework, and the application of the selected optimal enzyme-cocktail-substrate combinations are introduced.

2.1. Lignocellulosic biomass materials

The *77* biomass samples were obtained from various accessions collected in China, covering a diverse range of biomass samples (see Supplementary Materials). The raw or pretreated samples were collected, dried, ground, passed through a 60-mesh screen and stored in

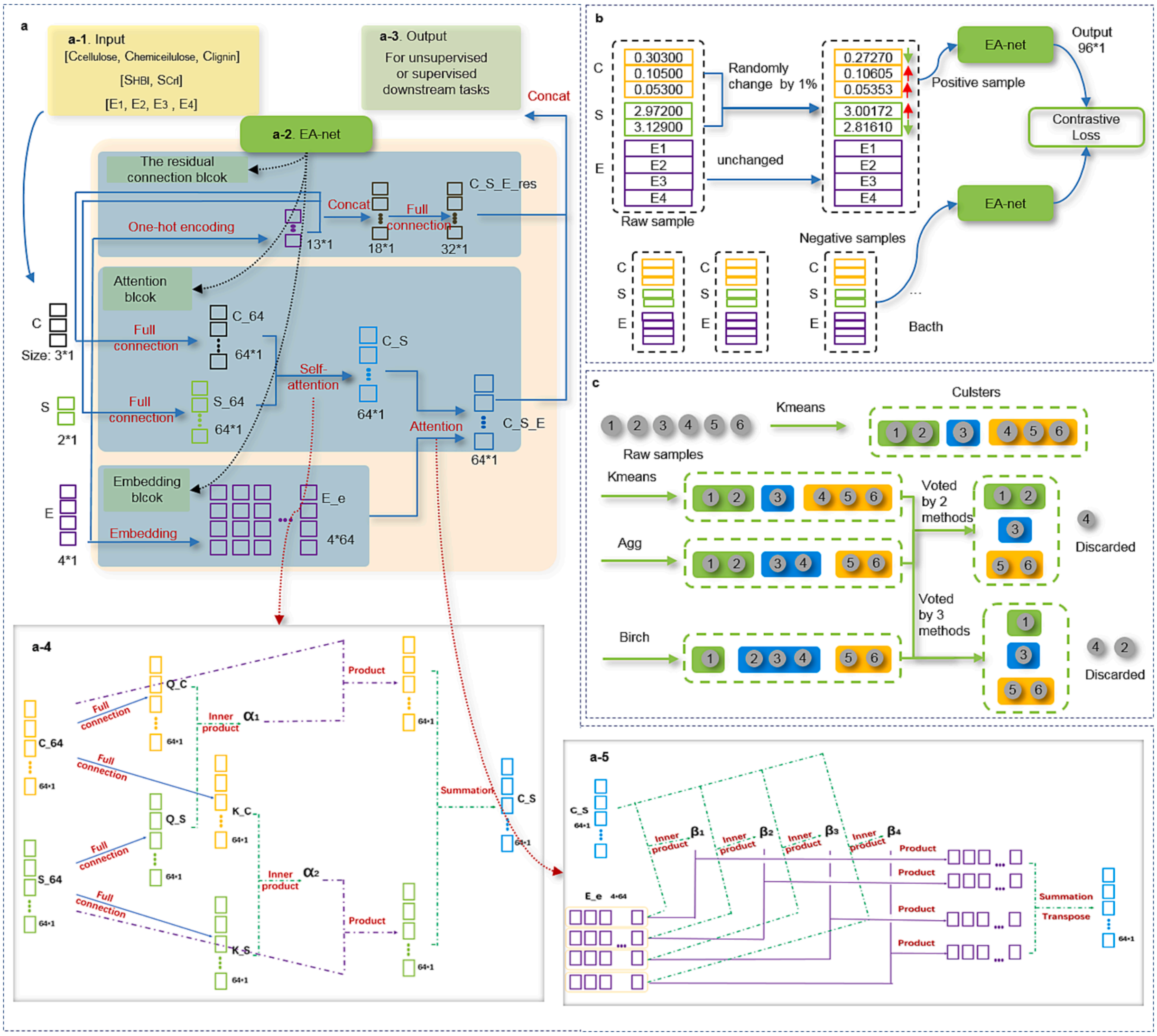


Fig. 1. Illustration of the proposed method. a, The structure of EA-net. a-1, The input consists of 9 simple features. a-2, The proposed EA-net mainly consists of attention and embedding blocks. The residual connection block, which increases the precision of EA-net by ~ 2%, is also illustrated. a-3, The output is versatile, and can include both supervised and unsupervised downstream tasks. **b,** Illustration of the applied contrastive learning strategy. **c,** Illustration of the multi-clustering strategy.

a dry container until use.

2.2. Analysis of plant cell wall components

The cellulose and hemicellulose contents of lignocellulosic biomass were quantitatively analyzed according to the NREL Laboratory Analytical Procedures using a two-step acid method (Sluiter et al., 2008) (see Supplementary Materials). The cellulose and hemicellulose contents were calculated (Sluiter et al., 2008), where factors of 0.90 and 0.88 reflect the weight loss in the conversion of glucose to glucan and xylose to xylan, respectively. The lignin content was determined according to Chinese standard methods (Gao et al., 2018).

2.3. FT-IR spectroscopy and x-ray diffraction (XRD) analysis

The sample was lyophilized at −20 °C for 24 h in a vacuum dryer (FD-IC-50, Beijing). Infrared spectra were recorded on an FT-IR 710 infrared spectrophotometer (Nicolet, Madison, WI). A total of 100 scans with a 2 cm^{-1} resolution were signal-averaged and stored. The scanned wave number range was 4000–400 cm^{-1}. The ratio of absorbance at 4000–2995 cm^{-1} to that at 1337 cm^{-1} of C-OH in-plane stretching was introduced as an empirical criterion of HBI (Gao et al., 2018) (see Supplementary Materials). The crystallinity of samples was examined by XRD measurements performed on a Bruker D8 Advance Diffractometer using Cu Kα radiation (λ = 0.1541 nm) at 30 kV and 30 mA, according to the method of Gao et al. (Gao et al., 2018).

2.4. Enzymatic preparations

T. reesei A2H (China General Microbiological Culture Collection Center; CGMCC 21470) is a lignocellulolytic enzyme-hyperproducing mutant strain obtained by chemical mutagenesis in our laboratory. The medium used for lignocellulolytic enzyme production was prepared in accordance with a previous study. The linear sgRNA construct and pCas9 expression plasmid with a codon-optimzied Cas9 expression cassetted were used to co-transform *Aspergillus niger* protoplasts. The following three DNA fragments were used to co-transform the protoplasts: (i) donor DNA, consisting of 500 bp 5′and 3′ homologous flanks around the target site; (ii) the Cas9 expression plasmid, and (iii) the sgRNA expression cassette. The transformation efficiency of NHEJ-deficient strains can reach 100.00%. The novel lignocellulolytic enzymatic preparation was a cocktail with cellulase from *T. reesei* and

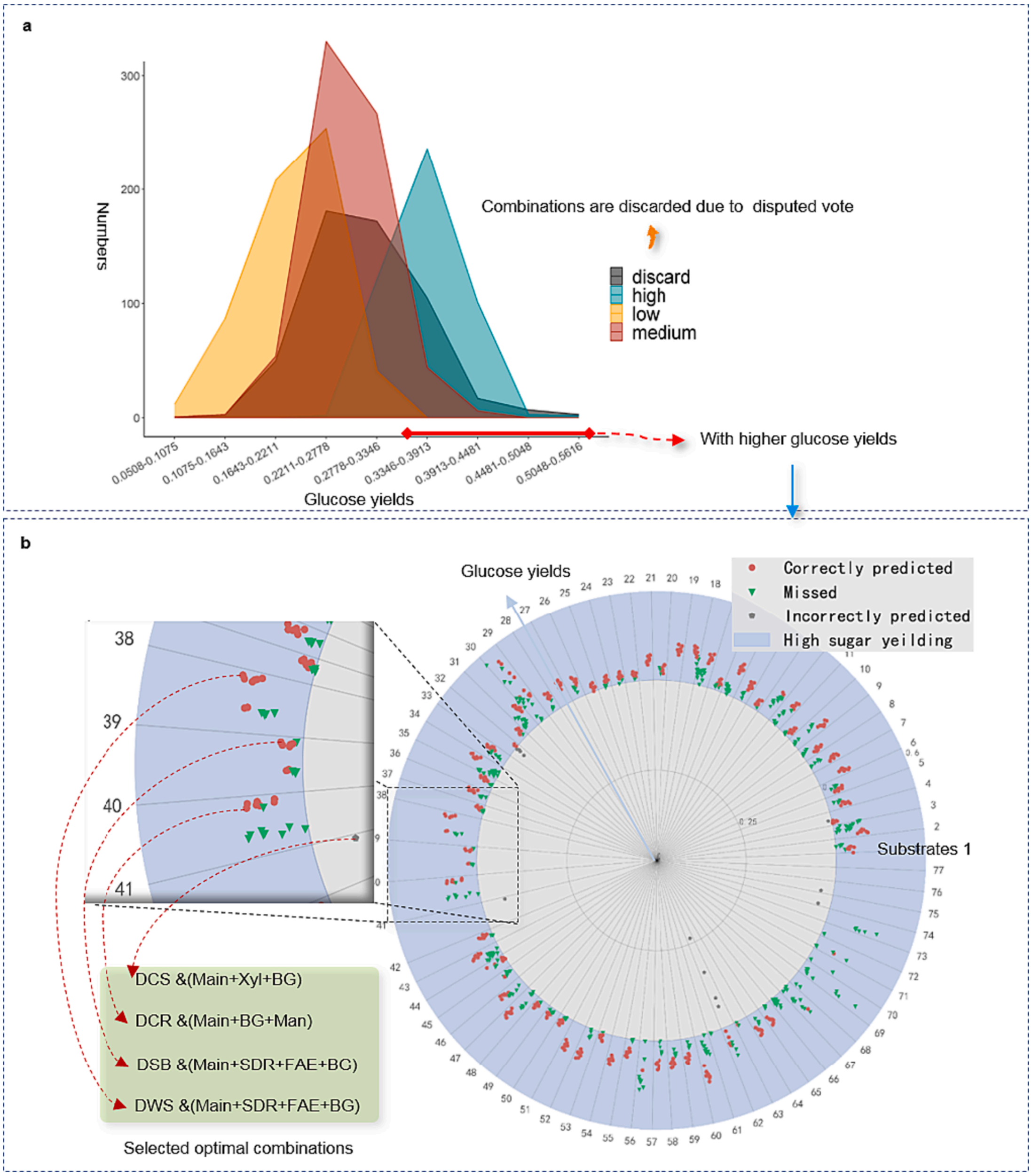


Fig. 2. The illustration of the main result. a, The histograms of sugar yields of the three groups clustered by the proposed proposed method are shown in the figure. The precisions of each group are improved by generating an additional group that contains combinations due to disputed votes. **b,** A radar chart with angles representing the substrates showing the predicted optimal group including those correctly predicted and missed. Although the algorithm missed some combinations with high sugar yields, it successfully predicted the combination of higher sugar yields for most of the substrates (the experiments with the highest precision among 6 repeats is presented here).

auxiliary enzymes from *A. niger* at a ratio of 8.5:1 (v/v) (see Supplementary Materials).

2.5. Analysis of biomass enzymatic digestibility

The biomass was subjected to enzymatic hydrolysis by cellulase at 50 °C for 72 h in triplicate (see Supplementary Materials). Hydrolysis experiments were conducted in 50 mL Erlenmeyer flasks with a total working volume of 20 mL while maintaining a substrate concentration of 5% (w/v). The enzyme loading was 20 FPU/g substrate. The reaction mixtures were supplemented with 0.5% NaN_3 to prevent microbial contamination. The samples were removed at regular intervals, and the supernatant was analyzed by high-performance liquid chromatography (HPLC) (Shimadzu, Kyoto, Japan) with a refractive index detector (Shimadzu) on an Aminex HPX-87H column (Bio-Rad, Hercules, CA, USA).

2.6. The EA-net model

In this section, the proposed EA-net which was used as the fundamental backbone network in the proposed method is introduced. As the algorithm is mainly based on two commonly used skills in deep learning, namely, the Embedding (Chowdhary and Chowdhary, 2020) and Attention (Vaswani et al., 2017) operations, it is named EA-net. The algorithm can effectively learn the nonlinear relations within the combination of biomass substrates and enzyme systems. Here, the input, structure, and output of EA-net is introduced, see Fig. 1.

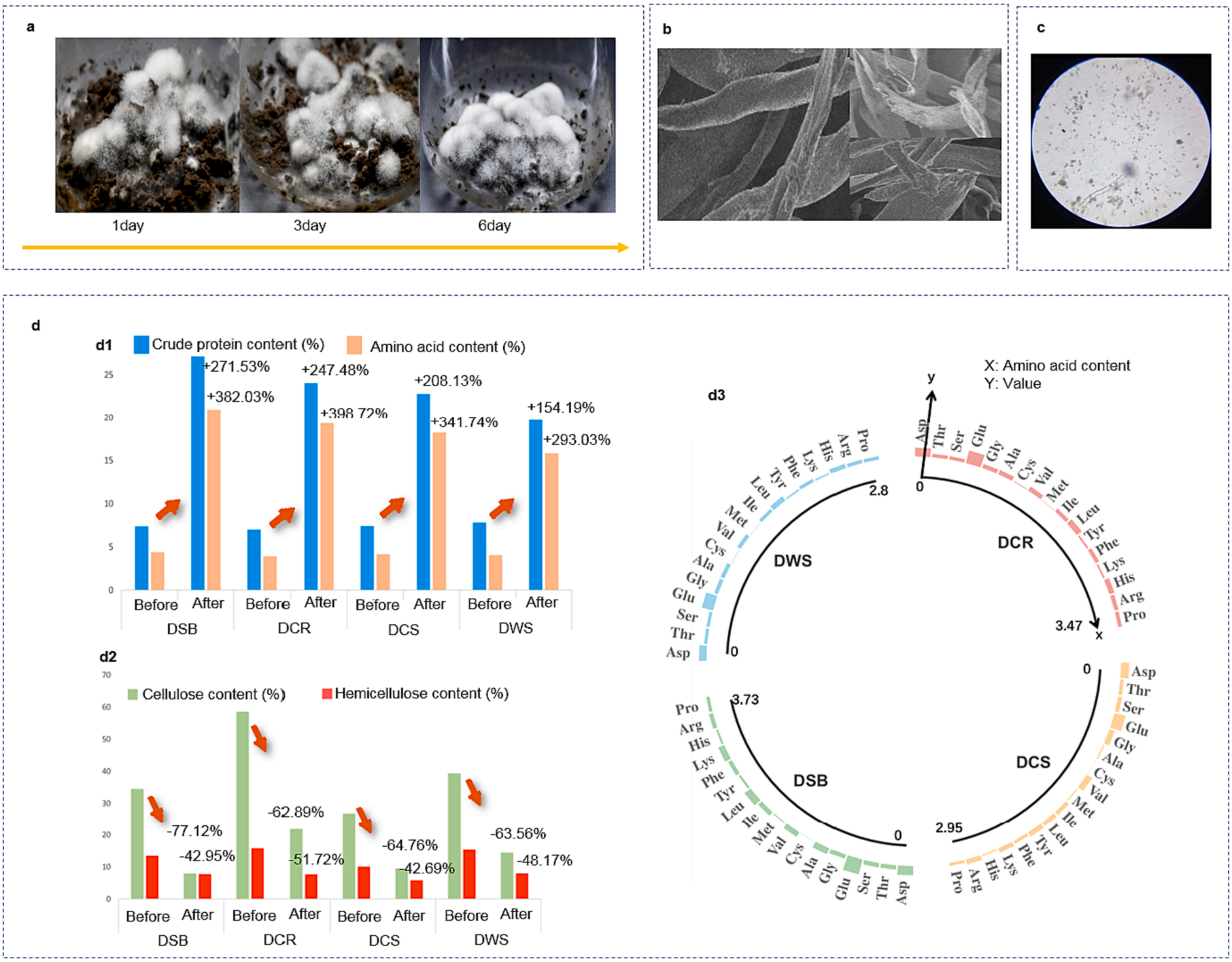


Fig. 3. Enhancing the nutritional value of non-food biomass. a, Solid fermentation process of agricultural biomass. **b,** SEM analysis of biomass before and after digestion with lignocellulolytic enzymes. **c,** Microscopic observation of solid fermentation. **d,** The results of the application. d-1, Comparison of the crude protein content before and after fermentation. d-2, Comparison of the cellulose and hemicellulose content before and after fermentation. d-3, Amino acid profile of protein fermented from agricultural biomass.

The input of EA-net

Firstly, the biomass substrates were characterized based on the proportions of components and the intricate structure type. Here, it considered the three canonical lignocellulose components, i.e. cellulose, hemicelluloses, and lignin. The intricate structure type is further indexed by HBI and CrI. Secondly, the method numbered each enzyme, and considered multi-enzyme mixtures containing no more than four kinds of enzymes (see Supplementary Materials). For those containing less than four enzymes, the missing enzymes were numbered as −1. Enzymes were selected from a pool of 12 distinct enzymes at a specific ratio. The enzymes from *T. reesei* constitute the main enzyme system, while the auxiliary system from *A. niger*. For example, if the multi-enzyme mixtures contain only “Main” and “FAE”, then, they are labeled as [1, 2, −1, −1].

As a result, the input of the EA-net is a combination of 2-, 3-, and 4-dimensional features corresponding to the proportion of components, the intricate structure type, and the multi-enzyme mixtures, respectively (Fig. 1a-1).

The structure of EA-net

In this section, the structure of EA-net is introduced, as shown in Fig. 1a-2. EA-net essentially consists of two parts, the embedding block and the attention block. Thus, the embedding operation and the attention operation are the keys to understanding the EA-net algorithm.

The embedding block. The embedding operation is essential to substitute the one-dimensional raw data with a multi-dimensional vector. This procedure can help the model “understand” the hidden meaning of the raw data. For example, the embedding operation helps the network understand the meaning of a word in the natural language processing (NLP) task (Chowdhary and Chowdhary, 2020). In the embedding block of the EA-net, the 4-dimensional input features, or vectors, representing the multi-enzyme mixtures, are embedded in vectors with the shape of: **[4,64]**, where 64 is the embedding size. For the sake of clarity, the embedded vectors are labeled as **E_e**. This procedure aims to make the network “understand” the properties of the enzyme.

The attention block: self-attention and attention operations. The attention operation (Vaswani et al., 2017) is usually used to find the implicit law where two sets of data interact with each other. In order to discover the implicit law based on which the biomass substrates interact with the enzyme systems, two attention operations are used in EA-net. Here, a detailed description of EA-net is given, starting from the inputs, **C**, **S**, and **E**, and ending at the output.

Firstly, the method transformed both the 2- and 3-dimensional features **C** and **S**, representing components and intricate structures, into 64-

dimensional features through dense connections, where $y_{out}\ (x) = \sigma\ (Wx_{input} + b)$, where x_{input} is the input vector, y_{out} is the output vector, W and b the weight and bias, and $\sigma(x)$ the activation function. Here, the activation function is a hyperbolic tangent function (tanh). For clarity, those two 64-dimensional features were named **C_64** and **S_64**.

Secondly, the method applied the self-attention operation (Vaswani et al., 2017) to modeling the relationship between variables. The modified self-attention operation (Vaswani et al., 2017) was applied on **C_64** and **S_64** yielding a 64-dimensional feature, named **C_S** for convenience. In the self-attention operation, the vector **C_64** is transformed into two 64-dimensional vectors denoted by **K_C** and **Q_C** using two independent full connections. Accordingly, 64 is the attention size, and the vector is **S_64**. Two 64-dimensional vectors obtained from the **S_64** are denoted by **K_S** and **Q_S**. The inner product between **Q_C** and **Q_S** gives the weight $\boldsymbol{\alpha_1}$ and that between **K_C** and **K_S** gives the weight $\boldsymbol{\alpha_2}$. The final result **C_S** of the self-attention operation is obtained by the weighted summation of **C_64** and **S_64** using the weights $\boldsymbol{\alpha_1}$ and $\boldsymbol{\alpha_2}$. This operation is used to deduce the law based on which the components interact with the intricate structure type (see Supplementary Materials).

Thirdly, the second attention operation is applied to **C_S** and the embedded vector of **E_e**. The second attention operation is used to deduce the law based on which the biomass substrates interact with the enzyme systems. In this attention operation, the inner products between **C_S** and the four 64-dimentional components of **E_e** give four weights, named $\boldsymbol{\beta_1}$, $\boldsymbol{\beta_2}$, $\boldsymbol{\beta_3}$, and $\boldsymbol{\beta_4}$ (see Supplementary Materials).

Finally, the output of the attention operation, a 64-dimensional vector named **C_S_E**, is the weighted summation of the four 64-dimensional components of **E_e** with weights $\boldsymbol{\beta_1}$, $\boldsymbol{\beta_2}$, $\boldsymbol{\beta_3}$, and $\boldsymbol{\beta_4}$.

In addition to the embedding and attention blocks, there is a residual connection in EA-net that can improve its behavior to some extent (over 2% in terms of precision). Therefore, the approach also introduced the residual connection. **C**, **E**, and **S** are transformed into a 32-dimentional feature, **C_S_E_res**, through a full connection layer, where **E** is converted to a vector of size 13 by a one-hot encoding operation, as shown in Fig. 1a-2.

The output of EA-net.

EA-net is a backbone structure that can be used for different downstream tasks. Therefore, the output of EA-net depends on the downstream task. Here, the downstream task is contrastive learning, so that the output of the EA-net is the concatenation of the **C_S_E_res** and the **C_S_E**, giving a 96-dimensional vector which will be considered as the representation of the sub-sample and support the downstream clustering task, see Fig. 1a-3.

EA-net can also be used to train the supervised task if a pre-labeled training set is available. In the supervised task, the output is a 3-d vector, $\mathbf{P_{predict}} = [\mathbf{P_{high}}, \mathbf{P_{medium}}, \mathbf{P_{low}}]$, representing the probability of each category. Firstly, the **C_S_E_res** and the **C_S_E** vectors are mapped onto 3D vectors by independent dense connections, and the output is the weighted summation of these two 3D vectors. Finally, the softmax operation is applied to the 3D vectors to give a predicted probability.

2.7. The modified contrastive learning approach

In this section, how EA-net is trained using the unsupervised contrastive learning strategy is described. The main idea of contrastive learning is to obtain a new representation of each sample so that each sample is close to positive samples and far from negative samples. Therefore, the positive and negative samples are the key concepts of contrastive learning. Conventionally, contrastive learning is a semi-supervised learning method (Chen et al., 2020), in which positive samples are pre-labeled by human operators. Here, the contrastive learning paradigm was extended into a fully unsupervised learning strategy by considering the sub-sample with the 5 features changing randomly by 1.00% as the positive sample. The negative samples are other samples except for the raw sample and its positive sample. Then, the contrastive learning algorithm is applied to obtain the final encoded features of each sub-sample. The contrastive loss is expressed as $loss = -\sum_i \log\left[\exp(S_{i,i}/\tau)/\left[\sum_{k\neq i}\exp(S_{i,k}/\tau) + \exp(S_{i,i}/\tau)\right]\right]$, where τ is a parameter and $S_{i,k}$ are negative sub-samples of x_i, the i-th sample in the batch, and $S_{i,i}$ is a positive sub-sample of x_i (Chen et al., 2020). Fig. 1b illustrates the contrastive learning strategy. The parameters embedded in EA-net and the contrastive learning strategy are adjusted simultaneously to minimize the contrastive loss.

In addition, EA-net can also be trained using a supervised approach if a pre-labeled training set is available. The loss function is the cross entropy between the predicted and real probabilities, i.e. $\mathbf{P_{predict}} = [\mathbf{P_{high}}, \mathbf{P_{medium}}, \mathbf{P_{low}}]$ and $\mathbf{p_{true}} = [\mathbf{p_{high}}, \mathbf{p_{medium}}, \mathbf{p_{low}}]$, respectively. Therefore, $loss = \sum_{i=1}^{3} p_i \ln\left(e^{P_i}/\sum_{k=1}^{3} e^{P_k}\right)$, where $\boldsymbol{P_i}$ and $\boldsymbol{p_i}$ are the predicted and real sugar production ratios, respectively.

The total number of parameters need for training is 18,820 (e.g., the embedding block has 845 parameters and the attention block has 16,896 parameters). The updating of the parameters is implemented using the TensorFlow framework for deep learning designed by Google. During the training, the batchsize is 256, and the learning rate is 0.00001.

2.8. The multi-clustering strategy

In this method, the main aim of EA-net and the contrastive learning strategy is to provide the final encoding, or representation, of each sub-sample. Then, three clustering methods, the k-means (Zhou et al., 2022), the agglomerative clustering algorithm (AGG) (Marcó et al., 2002), and balance iterative reducing and clustering using hierarchies (BIRCH) (Marcó et al., 2002), are used to group 2,310 sub-samples into 3 clusters, after which the final groups are obtained by voting. The multi-clustering strategy was proposed in previous work (Marcó et al., 2002) and it is reported that by voting on such different clustering methods, one can obtain clustering results with high purity at the cost of only rejecting some sub-samples with disputed vitiation. This shows that one can obtain more robust results by using the multi-clustering strategy. An illustration of the multi-clustering strategy is shown in Fig. 1c.

2.9. Application of optimal lignocellulolytic enzymes in the fermentation of protein from agricultural biomass

The biomass materials were hydrolyzed using 3.00% of optimal lignocellulolytic enzymes matched for the specific biomass at 50 °C for 24 h. At the same time, mycelia of *A. niger* were activated and grown in liquid seed medium in a rotary shaker at 28 °C and 150 rpm for 24 h. Subsequently, the analogous method was used to incubate *Candida utilis* for 24 h. For the solid-state fermentation, a 10.0% inoculum of *A. niger* was spread over the biomass which was maintained at 65.0–70.0% moisture at 30 °C for 2 days. After 48 h of fermentation, a 12.0% inoculum of *C. utilis* was added and the fermentation was continued for 3 days. The solid-state fermentation medium was composed of 100 g biomass powder with 4.00% $(NH_4)_2SO_4$, 0.20% $MgSO_4{\cdot}7H_2O$, and 1.53% KH_2PO_4, without further pH adjustment. Crude protein in the fermentation product was analyzed using the Kjeldahl method (Marcó et al., 2002). A conversion factor of 6.25 was used to calculate the theoretical protein content form the nitrogen content.

2.10. Brief summary of the database preparation for the proposed method

A dataset consisting of 2,310 enzyme-cocktail-substrate combinations without labels was utilized. This dataset included all possible combinations between 77 biomass sources and 30 lignocellulolytic enzyme cocktails (see Supplementary Materials). Each combination is characterized by five simple features to describe the biomass. These features encompass biomass compositions such as cellulose, hemicellulose, and lignin, as well as structural features like crystallinity intensity

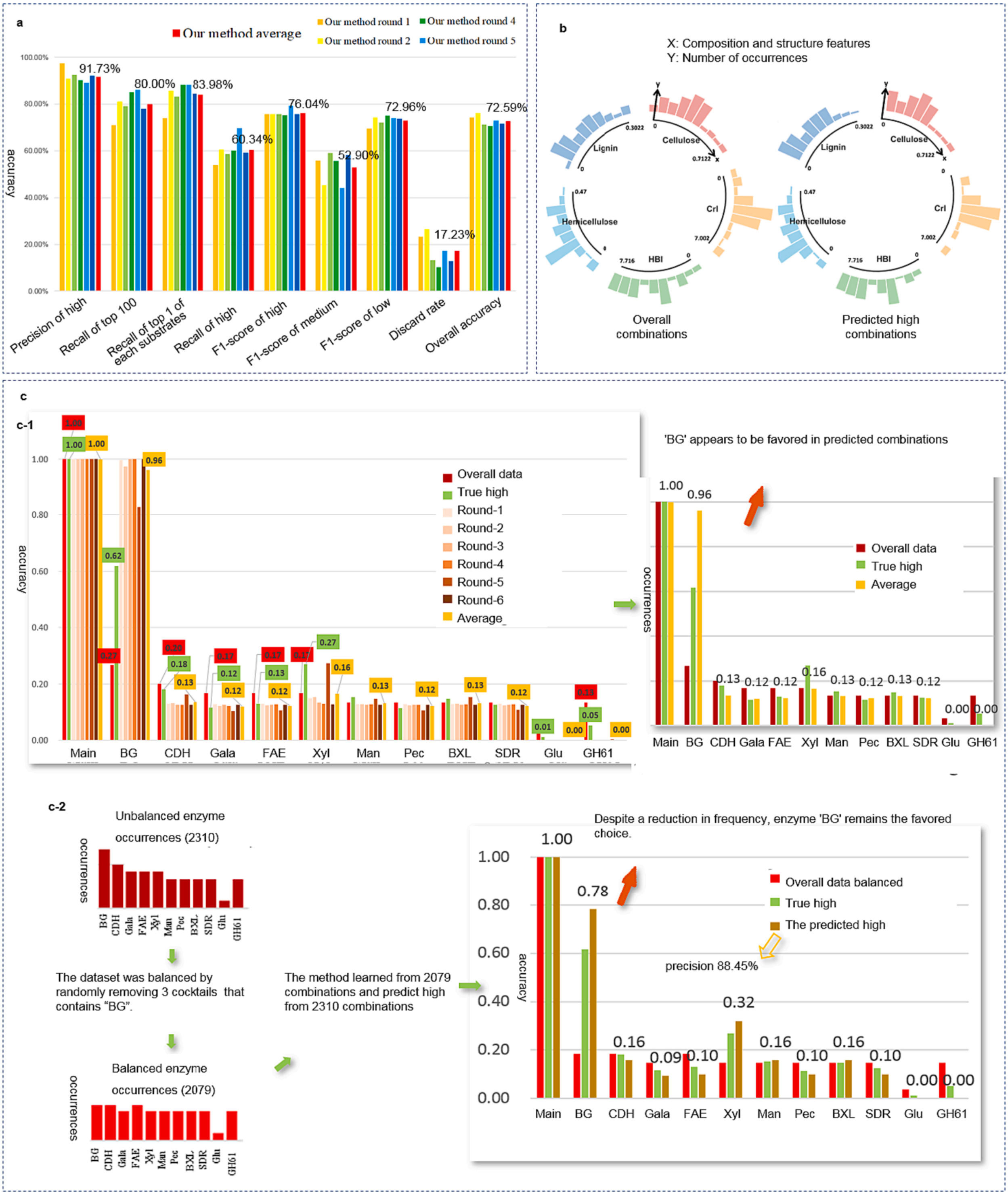


Fig. 4. The evaluation and analysis of the proposed method. a, Validation of proposed method through repeated experiments. The results demonstrated the robustness of the proposed method in 6 repeated experiments. **b,** Feature-focused analysis of predicted highly effective combinations. **c,** Enzyme-focused analysis of the predicted highly effective combinations. A comparison of the enzyme occurrences in the overall raw dataset and the predicted combinations with high sugar yields. c-1, Occurrence of enzymes in predicted highly effective combinations from 6 repeated experiments. c-2, Occurrence of enzymes in predicted highly effective combinations learned from datasets with a balanced enzyme distribution.

(CrI) and hydrogen-bond intensity (HBI). The enzyme cocktail is composed of a maximum of four enzymes, selected from a pool of 12 distinct enzymes. This selection is done at a specific ratio. The enzymes from *T. reesei* constitute the main enzyme system, while the auxiliary system from *A. niger* is also included. To represent the enzyme composition, each enzyme is assigned a simple index ranging from 1 to 12, with the missing enzyme indicated by −1 (see Supplementary Materials).

2.11. Brief summary of the framework of the proposed method

The 2,310 label-free enzyme-cocktail-substrate combinations were clustered into three distinct groups using the proposed method (see

Supplementary Materials). This method comprises three main steps. Firstly, the approach employed the proposed EA-net, which mainly consists of the embedding and attention blocks, to model the complex relationship between composition and structure features of biomasses and enzyme cocktail components, utilizing only simple feature inputs. Secondly, the approach enhanced the contrastive learning strategy by transforming it from a semi-supervised strategy to a fully unsupervised strategy through a modification in the selection of positive sub-samples, see method section. As a consequence, the method can combine the EA-net to efficiently identify and emphasize the critical information that distinguishes different enzyme-cocktail-substrate combinations, enabling us to focus on the most relevant features for clustering. Finally, to enhance the purity of each clustered group, the approach employed a voting strategy known as the multi-clustering strategy to eliminate combinations that receive disputed votes, further refining the clustering results. To find the optimal group, more than 10 combinations from each of the three groups were randomly selected, followed by experimental validation of the fermentable sugar/glucose yields.

3. Results and discussions

3.1. Practical application of the predicted optimal combination

Based on the experimental results, the glucose yields of the combinations were categorized as low (<0.25), medium (<=0.25 and < 0.30), or high (>=0.30). Interestingly, it observed that combinations belonging to one specific group consistently produce high glucose yields, surpassing the threshold of 0.3. Threshold configuration was based on the fact that the majority of samples demonstrated a cellulose conversion efficiency greater than 80%. This group was referred to as the predicted optimal group, as it demonstrates superior performance in terms of sugar production, see Fig. 2a.

The predicted optimal combinations obtained from the predicted optimal group were further applied to the degradation of four specific agricultural biomass materials: delignified wheat straw (DWS), delignified cotton stalks (DCS), delignified sugarcane bagasse (DSB), and delignified corncob residues (DCR), as shown in Fig. 2b. From the predicted optimal group, one optimal combination was identified for each of the four substrates, including DSB & (Main + SDR + FAE + BG), DCR & (Main + BG + Man), DCS & (Main + Xyl + BG), and DWS & (Main + SDR + FAE + BG). Subsequently, these four lignocellulolytic enzyme cocktails were utilized in solid-state fermentation of agricultural biomass over a period of six days (Fig. 3a). *A. niger* and *C. utilis* are two microorganisms that can utilize fermentable sugars derived from agricultural biomass for their rapid growth during solid-state fermentation, leading to the accumulation of high-quality single-cell protein (Fig. 3b–d). The results showed a significant increase in the crude protein content of the agricultural straws, with values ranging from 19.70 to 27.27%, compared to the initial content of 6.97–7.75%, as depicted in Fig. 3d-1. This increase in crude protein content corresponds to an actual protein content increase from 3.87 to 4.34% to 15.80–20.92% after fermentation. Moreover, the resulting fermented straws exhibited a balanced nutritional profile, rich in eight essential amino acids, as shown in Fig. 3d-3. These outcomes surpassed the results reported in previous studies. For example, Sun et al. observed a crude protein increase from 6.74 to 9.51% during corn straw silage fermentation (Sun et al., 2019), while Zhang et al. achieved a protein content in corn straw of 67.00 g/kg using combined fermentation with *Geotrichum candidum* and *Phanerochaete chrysosporium* (Zhang et al., 2006). Additionally, the selected optimal enzyme cocktail approach demonstrated efficient cellulose degradation, with efficiencies of 62.89–77.12%, as well as hemicellulose degradation efficiencies of 42.69–51.72%, as displayed in Fig. 3d-2. These results highlight the promising potential of the proposed approach in utilizing available agricultural biomass and enhancing its nutritional value for non-food purposes. Using low-priced agricultural straw and a low-energy solid-state fermentation system results in higher efficiency of protein production compared to soy (38.60%), fish (17.80%), meat (21.20%) and whole milk (3.28%) (Salazar-López et al., 2022).

By 2055, it is anticipated that the world's population will grow from 7.4 billion to 10 billion, leading to a 50% surge in global food demand (Heleniak 2021). One promising solution for reducing dependence on natural resources and promoting sustainable food production involves the conversion of lignocellulosic biomass, such as agricultural straw, into protein-rich food sources for livestock and poultry (Xu et al., 2023). However, technological bottlenecks have hampered the biotransformation of agricultural straw due to the indigestibility of lignocellulosic biomass. To achieve more economical conversion of lignocellulosic biomass into fuels and food, researchers should further focus on the production of low-cost and effective lignocellulolytic enzyme cocktails. However, the current approaches for matching optimal enzymes to specific substrates rely on time-consuming, labor-intensive, and expensive experimental methods. Despite the recent advancements in the application of artificial intelligence in biotechnological research, such as enzyme classification annotation (Bileschi et al., 2022), enzyme function prediction (Zou et al., 2019), antimicrobial identification (Ma et al., 2022), and evaluation of substrate specificity (Mou et al., 2021), existing methods have encountered challenges in learning the properties of enzyme-cocktail-substrate combinations due to limited experimental datasets (pre-labeled training sets) and the incomplete understanding of reaction mechanisms, such as structure-dynamics-function relationships in macromolecular assemblies consisting of enzyme cocktails and substrates (Jabbour et al., 2013).

3.2. Evaluation of the proposed method

The performance of the proposed model was evaluated by comparing its predicted optimal group with the experimentally determined high-yield combinations. The experiments were repeated six times, and on average, the predicted optimal group contained 552 matching combinations. Out of these, 91.98% were accurately predicted, amounting to an average of 506 combinations with high sugar yields. In terms of sub-sample recall rates, the model achieved an 80.00% recall rate for the top 100 sugar yields. Additionally, for each substrate, the recall rate for sub-samples with optimal sugar yield was 89.81%, with an average of 64.66 out of 72 sub-samples being accurately recalled. It is worth noting that among the 77 substrates, only 72 had high sugar yields. Overall, the recall rate for sub-samples with high sugar yields was 60.23%, with 506 out of 840 sub-samples being accurately recalled out of a total of 2,310. The evaluation of the predicted optimal groups is illustrated in Fig. 4a.

Instead of using traditional supervised methods, this study developed an easy and effective method to optimize enzyme cocktails for specific substrates by adopting a unique and distinct approach that bypasses challenges faced by traditional supervised methods. The key idea is to cluster the macromolecular assemblies based on their similarities without considering the specified physical and chemical properties. The key contents can be summarized as follows: (1) An unsupervised-learning approach that eliminates the need for costly pre-labeled datasets obtained from experiments was proposed. Despite this, the proposed method outperforms supervised-learning methods in predicting optimal lignocellulolytic enzyme cocktails. (2) It utilized simple and easily obtainable features as inputs, such as simple fingerprints of substrates and a numerical representation of the enzyme cocktail. By focusing on these features instead of high-dimensional sequence and structure information, it achieved high prediction accuracy without relying on expert-level prior knowledge of reaction mechanisms. (3) It demonstrated the applicability of the proposed method in the context of the solid-state fermentation of agricultural straw for protein production, showcasing its promising potential in accelerating the selection of optimal enzyme-cocktail-substrate combinations. Notably, this approach narrows down the range of possible auxiliary cocktails, enabling a more efficient and targeted selection process.

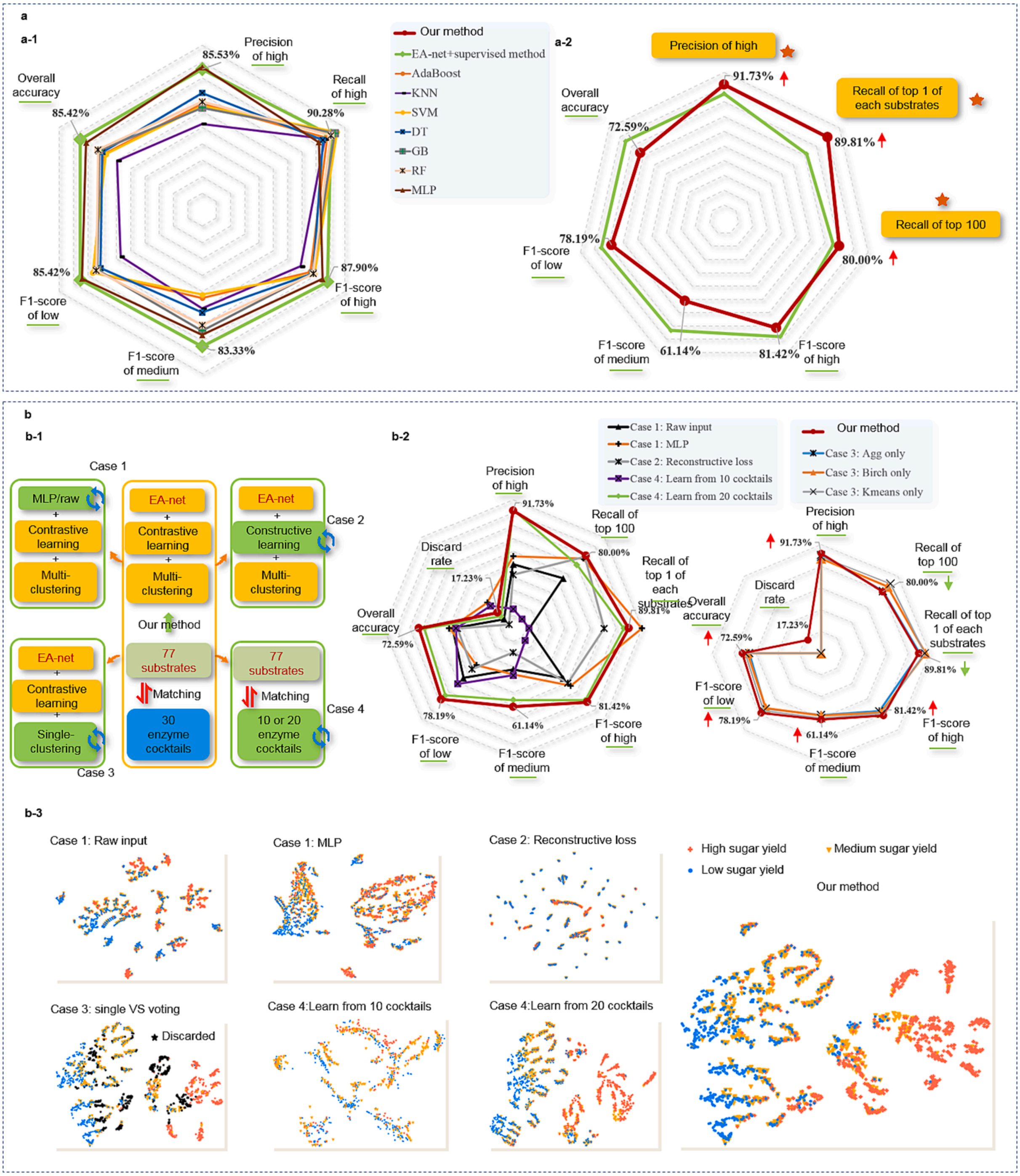


Fig. 5. The effectiveness and robustness of the proposed method. a, Comparison between the supervised and unsupervised methods. a-1, Comparison between the EA-net based supervised method and existing supervised methods. A radar graph showing that the EA-net based supervised method outperformed exiting supervised methods, including AdaBoost (AB), k-neighbors (KNN), support vector machine (SVM), decision tree (DT), gradient boost (GB), random forest (RF), and multi-layer perception (MLP). a-2, Comparison between the EA-net based supervised and unsupervised methods. The results showed that the EA-net based unsupervised method outperformed the EA-net based supervised method in the prediction the optimal combinations. **b,** The results of ablation experiments. b-1, Illustrations of the cases of removal experiments where three components of the proposed method were substituted by the alternative methods and the dataset was also altered. b-2, The results of different removal experiments. b-3, The t-SNE visualization graphs of each case, further demonstrating the effectiveness of the proposed method.

3.3. Analysis of the substrate features and enzyme cocktail compositions

Further analysis of the substrate features and enzyme cocktail compositions of the predicted optimal combinations indicated that the proposed method was agnostic towards substrate features (Fig. 4b). Additionally, it observed a tendency for the proposed method to group combinations containing the enzyme “BG” as having a high efficiency. This observation could be attributed to the imbalanced occurrence of the “BG” enzyme in the raw dataset. Out of the 2310 combinations, 616 contained this enzyme, resulting in the highest occurrence probability of 26.66% among the 11 auxiliary enzymes, which had an average occurrence probability of 15.00% (Fig. 4c-1). To investigate the impact of this imbalanced enzyme distribution on the proposed method, additional experiments were conducted. The method was trained using a balanced dataset by randomly removing 231 combinations that contained the "BG" enzyme. The results demonstrated that the proposed method was capable of learning optimal enzyme-cocktail-substrate combinations from different datasets, albeit to varying degrees, e.g., the precision of predicting high combination by learning on a balanced but smaller dataset is 88.43% (Fig. 4c-2). Without a pre-labeled training set, the model deduced that compositions of natural lignocellulolytic cocktails such as that of *T. reesei* are usually imbalanced and need the addition of "BG" (Gusakov et al., 2007; Waghmare et al., 2021). Notably, the method inferred that enzyme combinations including "BG" were associated with higher degradation rates, which was consistently observed across four datasets that varied in the frequency of this particular enzyme (see Supplementary Materials). These findings highlight the robustness and adaptability of the proposed method in learning from datasets with imbalanced distributions of enzymes. By effectively capturing the underlying patterns and relationships, the proposed method can provide accurate predictions and insights for optimal enzyme cocktails, even in scenarios with varying enzyme frequencies.

3.4. The effectiveness and robustness of the proposed method

Further assessments of the effectiveness and robustness of the proposed method confirmed that the EA-net neural network is versatile and can serve as a backbone for both supervised and unsupervised methods. To demonstrate its effectiveness, a supervised approach based on EA-net was developed, whereby the network was trained on 69 out of the 77 substrate samples, with the remaining 8 used for testing. The resulting training and test sets consisted of 2,070 and 240 sub-samples, respectively. The unsupervised contrastive learning loss function (Fig. 1b) was replaced with a cross-entropy loss function that uses known labels. Compared to the currently available supervised methods, the EA-net achieved greater precision in almost every category and had the highest overall accuracy of 85.42% when applied to the test set (Fig. 5a-1). However, the accuracy of the optimal matching combinations, 85.42%, derived from the supervised method was lower than that of the proposed unsupervised method, 91.73% (Fig. 5a-2, see Supplementary Materials).

The proposed unsupervised learning method leverages EA-net to encode 9-dimensional sub-samples (comprising 5 substrate features and 4 cocktail features) into higher-dimensional vectors. By doing so, EA-net is able to identify and emphasize the crucial characteristics that set apart different enzyme-cocktail-substrate combinations. To demonstrate the effectiveness of EA-net, alternative methods were also tested, such as encoding each sub-sample using a multi-layer perception (MLP) network or using the raw 9-dimensional vector directly. The contrastive learning strategy is crucial for discovering similarities and differences between combinations during the encoding process. The effectiveness of this strategy was illustrated by replacing the contrastive learning loss function with a reconstructive loss function (Fig. 1) that aims to reconstruct each sub-sample. Furthermore, a multi-clustering strategy, or a voting approach, was employed to enhance the robustness of the clustering outcomes, taking into account the results obtained using three clustering algorithms. Voting improves the precision of each group at the expense of potentially missing some optimal combinations due to disputed votes, and the higher precision can aid in identifying the optimal group through experimental verification. Its effectiveness is shown by comparing the results of single- and multi-clustering strategies. Fig. 5 shows the results that illustrate the superiority of the proposed unsupervised learning method.

Compared to supervised learning methods that predict the properties of combinations based on sequence information and then classify samples based on the predictions (see Supplementary Materials), the approach directly clustered combinations based on the superficial features of the samples that reflect their similarity without knowing the detailed properties of the combinations such as the degradation rate. Therefore, the proposed approach is highly workable (only simple and easily obtainable features were utilized) and efficient (no need for costly pre-labeled training sets). However, since the model needs to compare the differences between samples in the given dataset, the proposed method, like supervised learning methods, also relies on the given dataset. In addition, the robustness of the proposed method was investigated across different datasets by training it on alternative datasets that consist of only 770 sub-samples (77 substrates and 10 enzyme cocktails) or 1,540 sub-samples (77 substrates and 20 cocktails), in addition to the primary dataset of 2,310 sub-samples, and then generating clusters of sub-samples containing superior matches. However, training with 770 sub-samples with precision of prediction high 15.00% proved unsatisfactory due to the insufficient combinations of substrates and enzymes resulting from the uneven and scarce distribution of the 10 cocktails, which prevented EA-net from acquiring knowledge about enzyme properties. On the other hand, when 1,540 sub-samples containing 20 enzyme cocktails were utilized for learning, the outcome improved significantly (Fig. 5b-2). The results of the experiments conducted on other datasets are presented, including two experiments the influences of data sizes and four examining the influences of imbalanced enzyme distribution (see Supplementary Materials). It demonstrates that the proposed method, similar to the supervised method, relies on the provided dataset and has the ability to learn optimal combinations from different datasets. This indicates that the method is adaptable and can effectively learn from different datasets, albeit to varying degrees. The robustness of the method in terms of its reliance on different datasets will be further investigated in future studies. This will provide additional insights into the generalizability and reliability of the method in real-world applications.

4. Conclusions

The proposed method offers a means to expedite the conversion of biomass into fermentable sugars by efficiently identifying optimal enzyme-cocktail-substrate combinations, thus enabling us to effectively leverage the available agricultural biomass and enhance the nutritional value of non-food biomass. This breakthrough opens up new possibilities for the wider adoption of deep learning techniques in biomass biorefining technologies by eliminating the reliance on experimental datasets and expert-level prior knowledge of structure-dynamics-function based reaction mechanisms. Finally, the user-friendly interface and cost-effectiveness make proposed method accessible to a broader range of researchers and industrial engineers, thus fostering innovation and progress in biomass biorefining.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgements

This work was supported by the National key R&D program of China (2022YFC2105002); Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDA28030301); National Natural Science Foundation of China (62106033); Yunnan Fundamental Research Project (202001AU070020); Jilin Province and Chinese Academy of Sciences Science and Technology Cooperation High-tech Industrialization Fund Project (2022SYHZ0015, 2022SYHZ0019).

Author contributions statement.

Experimental work: L.G; Formal analysis: S.J.W; Writing original draft: Z.H.Y and L.G; Development of the deep learning model: C.C.Z, Z. H.Y, Y.J.H, and S.C.Z; Analyzing the results: Z.H.Y, Y.J.H, S.C.Z and C.C. Z; Reviewing the manuscript: C.C.Z and X.W. All authors have read and agreed to the published version of the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.biortech.2023.129758.

References

Ashokkumar, V., Venkatkarthick, R., Jayashree, S., Chuetor, S., Dharmaraj, S., Kumar, G., Chen, W.H., Ngamcharussrivichai, C., 2022. Recent advances in lignocellulosic biomass for biofuels and value-added bioproducts-a critical review. Bioresour. Technol. 344, 126195.

Bileschi, M.L., Belanger, D., Bryant, D.H., Sanderson, T., Carter, B., Sculley, D., Bateman, A., DePristo, M.A., Colwell, L.J., 2022. Using deep learning to annotate the protein universe. Nat. Biotechnol. 40 (6), 932–937.

Binod, P., Gnansounou, E., Sindhu, R., Pandey, A., 2019. Enzymes for second generation biofuels: recent developments and future perspectives. Bioresour. Technol. Rep. 5, 317–325.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations. In Int. conf. mach. learning. 1597–1607.

Chowdhary, K., Chowdhary, K.R., 2020. Natural language processing. Fundamentals of artificial intelligence. 603–649.

Du, J., Liang, J., Gao, X., Liu, G., Qu, Y., 2020. Optimization of an artificial cellulase cocktail for high-solids enzymatic hydrolysis of cellulosic materials with different pretreatment methods. Bioresour. Technol. 295, 122272.

Gao, L., Chen, S., Zhang, D., 2018. Neural network prediction of corn stover saccharification based on its structural features. Biomed. Res, Int, p. 9167508.

Gusakov, A.V., Salanovich, T.N., Antonov, A.I., Ustinov, B.B., Okunev, O.N., Burlingame, R., Emalfarb, M., Baez, M., Sinitsyn, A.P., 2007. Design of highly efficient cellulase mixtures for enzymatic hydrolysis of cellulose. Biotechnol. Bioeng. 97 (5), 1028–1038.

Heleniak, T., 2021. The future of the Arctic populations. Polar Geogr. 44 (2), 136–152.

Hu, J., Arantes, V., Saddler, J.N., 2011. The enhancement of enzymatic hydrolysis of lignocellulosic substrates by the addition of accessory enzymes such as xylanase: is it an additive or synergistic effect? Biotechnol. Biofuels 4, 36.

Indyk, P., Motwani, R., 2000. Approximate nearest neighbors: towards removing the curse of dimensionality. Theory of Computing. 1, 604–613.

Jabbour, D., Borrusch, M.S., Banerjee, G., Walton, J.D., 2013. Enhancement of fermentable sugar yields by α-xylosidase supplementation of commercial cellulases. Biotechnol. Biofuels 6 (1), 58.

Kumar, R., Wyman, C.E., 2014. Strong cellulase inhibition by Mannan polysaccharides in cellulose conversion to sugars. Biotechnol. Bioeng. 111 (7), 1341–1353.

Li, F., Yuan, L., Lu, H., Li, G., Chen, Y., Engqvist, M.K., Kerkhoven, E.J., Nielsen, J., 2022. Deep learning-based k_{cat} prediction enables improved enzyme-constrained model reconstruction. Nat. Catal. 5 (8), 662–672.

Liu, G., Qin, Y., Li, Z., Qu, Y., 2013. Development of highly efficient, low-cost lignocellulolytic enzyme systems in the post-genomic era. Biotechnol. Adv. 31 (6), 962–975.

Liu, G., Qu, Y., 2021. Integrated engineering of enzymes and microorganisms for improving the efficiency of industrial lignocellulose deconstruction. Eng. Microbiol. 1, 100005.

Ma, Y., Guo, Z., Xia, B., Zhang, Y., Liu, X., Yu, Y., Tang, N., Tong, X., Wang, M., Ye, X., Feng, J., Chen, Y., Wang, J., 2022. Identification of antimicrobial peptides from the human gut microbiome using deep learning. Nat. Biotechnol. 40 (6), 921–931.

Marcó, A., Rubio, R., Compañó, R., Casals, I., 2002. Comparison of the Kjeldahl method and a combustion method for total nitrogen determination in animal feed. Talanta 57 (5), 1019–1026.

Mazurenko, S., Prokop, Z., Damborsky, J., 2020. Machine learning in enzyme engineering. ACS Catal. 10 (2), 1210–1223.

Mou, Z., Eakes, J., Cooper, C.J., Foster, C.M., Standaert, R.F., Podar, M., Doktycz, M.J., Parks, J.M., 2021. Machine learning-based prediction of enzyme substrate scope: application to bacterial nitrilases. Proteins 89 (3), 336–347.

Murashima, K., Kosugi, A., Doi, R.H., 2003. Synergistic effects of cellulosomal xylanase and cellulases from *Clostridium cellulovorans* on plant cell wall degradation. J. Bacteriol. 185 (5), 1518–1524.

Ragauskas, A.J., Williams, C.K., Davison, B.H., Britovsek, G., Cairney, J., Eckert, C.A., Frederick Jr, W.J., Hallett, J.P., Leak, D.J., Liotta, C.L., Mielenz, J.R., Murphy, R., Templer, R., Tschaplinski, T., 2006. The path forward for biofuels and biomaterials. Science 311 (5760), 484–489.

Salazar-López, N.J., Barco-Mendoza, G.A., Zuñiga-Martínez, B.S., Domínguez-Avila, J.A., Robles-Sánchez, R.M., Villegas Ochoa, M.A., González-Aguilar, G.A., 2022. Single-cell protein production as a strategy to reincorporate food was te and agro by-products back into the processing chain. Bioengineering 9, 623.

Selig, M.J., Knoshaug, E.P., Adney, W.S., Himmel, M.E., Decker, S.R., 2008. Synergistic enhancement of cellobiohydrolase performance on pretreated corn stover by addition of xylanase and esterase activities. Bioresour. Technol. 99 (11), 4997–5005.

Sluiter, A., Hames, B., Ruiz, R., Scarlata, C., Sluiter, J., Templeton, D., Crocker, D.L.A.P., 2008. Determination of structural carbohydrates and lignin in biomass. Lab. anal. proced. 1617 (1), 1–16.

Smith, A., Keane, A., Dumesic, J.A., Huber, G.W., Zavala, V.M., 2020. A machine learning framework for the analysis and prediction of catalytic activity from experimental data. Appl. Catal. B-Environ. 263, 118257.

Sun, L., Wang, Z., Gentu, G., Jia, Y., Hou, M., Cai, Y., 2019. Changes in microbial population and chemical composition of corn stover during field exposure and effects on silage fermentation and *in vitro* digestibility. Asian-Australas. J. Anim. Sci. 32 (6), 815–825.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Proces. Syst. 30.

Waghmare, P.R., Waghmare, P.P., Gao, L., Sun, W., Qin, Y., Liu, G., Qu, Y., 2021. Efficient constitutive expression of cellulolytic enzymes in *Penicillium oxalicum* for improved efficiency of lignocellulose degradation. J. Microbiol. Biotechnol. 31 (5), 740–746.

Wang, J., Chen, X., Chio, C., Yang, C., Su, E., Jin, Y., Cao, F., Qin, W., 2019. Delignification overmatches hemicellulose removal for improving hydrolysis of wheat straw using the enzyme cocktail from *Aspergillus niger*. Bioresour. Technol. 274, 459–467.

Xu, X., Zhang, W., You, C., Fan, C., Ji, W., Park, J.T., Kwak, J., Chen, H., Zhang, Y.P.J., Ma, Y., 2023. Biosynthesis of artificial starch and microbial protein from agricultural residue. Sci. Bull. 68 (2), 214–223.

Zhang, Y., Lin, S.M., Zhu, Y.J., Liu, C.J., Dong, Y., Li, F.F., Wu, G.F., Wang, H.Y., Zhang, J.H., 2006. Protoplast fusion between *Geotrichum candidium* and *Phanerochaete chrysosporium* to produce fusants for corn stover fermentation. Biotechnol. Lett 28 (17), 1351–1359.

Zhang, J., Tuomainen, P., Siika-Aho, M., Viikari, L., 2011. Comparison of the synergistic action of two thermostable xylanases from GH families 10 and 11 with thermostable cellulases in lignocellulose hydrolysis. Bioresour. Technol. 102 (19), 9090–9095.

Zhou, C., Gu, Y., Fang, G., Lin, Z., 2022. Automatic morphological classification of galaxies: convolutional autoencoder and bagging-based multiclustering model. Astron. J. 163 (2), 86.

Zou, Z., Tian, S., Gao, X., Li, Y., 2019. MlDEEPre: multi-functional enzyme function prediction with hierarchical multi-label deep learning. Front. Genet. 9, 714.