MECE: a method for enhancing the catalytic efficiency of glycoside hydrolase based on deep neural networks and molecular evolution

Jian Tian (tianjian@caas.cn)  
Biotechnology Research Institute, Chinese Academy of Agricultural Sciences, Beijing  
https://orcid.org/0000-0002-9997-6518

Feifei Guan  
Biotechnology Research Institute, Chinese Academy of Agricultural Sciences, Beijing

Hanqing Liu  
Biotechnology Research Institute, Chinese Academy of Agricultural Sciences, Beijing

Tuoyu Liu  
Biotechnology Research Institute, Chinese Academy of Agricultural Sciences, Beijing

Lixin Yang  
Biotechnology Research Institute, Chinese Academy of Agricultural Sciences, Beijing

Xiaoqing Liu  
Biotechnology Research Institute

Huiying Luo  
Institute of Animal Sciences  https://orcid.org/0000-0002-4596-2639

Ningfeng Wu  
Biotechnology Research Institute, Chinese Academy of Agricultural Sciences, Beijing

Bin Yao  
Chinese Academy of Agricultural Sciences

Huoqing Huang  
Northwest A&F University  https://orcid.org/0000-0003-3197-6631

Article

**Keywords:** MECE, Deep learning, Catalytic efficiency, Glycoside hydrolases, Mutants

**Posted Date:** September 15th, 2022

**DOI:** https://doi.org/10.21203/rs.3.rs-1988173/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

High efficiency glycoside hydrolases (GH) are in high demand for numerous industrial applications. This study demonstrates the use of a deep neural network and molecular evolution (MECE) platform for predicting catalysis-enhancing mutations in GHs. The MECE platform integrates a deep learning model (DeepGH), trained with 119 GH family protein sequences from the CAZy database. Ten-fold cross-validated DeepGH models showed 96.73% predictive accuracy. MECE also includes a quantitative mutation design component that uses Grad-CAM with homologous protein sequences to identify key features for mutation in the target GH. Validation of the MECE platform with chitosanase CHIS1754 and glucoamylase GA51, resulted in generation of CHIS1754-MUT7, harboring seven amino acid conversions, and GA51-MUT5, carrying five residue conversions. The $k_{cat}/K_m$ of CHIS1754-MUT7 was 18.08-fold higher than CHIS1754, while GA51-MUT5 was 7.64-fold greater than that of GA51. This resource can facilitate the rational design of catalytically efficient enzymes for a broad range of applications.

Introduction

Protein engineering is essential to improve the enzyme function beyond that of their in vivo constraints, such as increasing catalytic activity, broadening substrate specificity, enhancing thermodynamic stability, or increasing its expression levels and solubility. Among these demands, high catalytic activity is an essential requirement for industrial applications.

Glycoside hydrolases (GHs) act on various glycosides or oligosaccharides to hydrolyze glycosidic bonds, and these proteins, such as lysozyme, amylase, chitinase, chitosanase and glucoamylase, are widely used in industry, medicine, agriculture, and food production. There are two main approaches to improving the catalytic efficiency of glycoside hydrolases, namely directed evolution and rational design. The former approach is often a random mutagenesis-driven method, especially when no structure–function relationship data are available for a given protein. For example, through error prone PCR, Chen et al. generated a library of 1200 variants of the *Lentinula edodes* glycoside hydrolase LXYL-P1-2, two of which showed 17% and 47% higher catalytic efficiency towards β-xylosyl-10-deacetyltaxol. By contrast, rational design broadly can entail either structure-based in silico prediction or data-driven protein design.

Structure-based in silico prediction can help to reduce library size (e.g., smart libraries) by identifying and directly designing site-directed mutations. However, this method greatly depends on the accuracy of the protein structure, dynamics, and a deep understanding of the relationship between structure and function. In 2021, structural analysis of *Bacillus circulans* xylanase by Min et al. identified the key site R49. Subsequent saturation mutagenesis yielded the R49N variant, with 7.51-fold higher catalytic efficiency. Although these methods have been successfully applied to enhance the catalytic activity of numerous enzymes, the tremendous size of the protein sequence space and a lack of an effective high-throughput screening system continue to pose a challenge for enzyme development.

More recently, data-driven approaches have facilitated protein engineering through increased access to large experimental protein databanks, high throughput sequencing, high-throughput screening (HTS)
methods, and advances in artificial intelligence algorithms. These approaches use sequence-fitness relationships to train machine learning models to perform \textit{in silico} screening and variant selection for a defined or combinatorial sequence space. For example, the ProSAR platform was developed to capture information contained in sequence-activity data followed by directed evolution to obtain mutants with high catalytic efficiency. The DeepEC tool is a deep learning-based computational framework that predicts Enzyme Commission (EC) numbers for protein sequences with high precision in a high-throughput manner. The Michaelis constant ($K_m$), which can be difficult and time-consuming to determine experimentally, can also be predicted by machine learning or deep learning methods. Additionally, site-directed mutations that enhance catalytic efficiency can be selected from the preferred mutations based on natural sequence evolution. The success of the above approaches shows that machine learning methods can effectively recognize sequence or structural features related to protein function from large protein datasets.

The CAZymes Database is a specialized database dedicated to the display and analysis of genetic, structural and biochemical information for carbohydrate-active enzymes, including nearly all glycoside hydrolases. As of September 2020, the database contained 854,279 glycoside hydrolases. The recent emergence of the deep learning framework and its application in large-scale protein data reveals the potential of this approach for addressing classic challenges in protein engineering. Based on the premise that the classification features of glycoside hydrolases are related to their catalytic efficiency, this study investigates which features are relevant to engineering improved catalytic efficiency and how to extract them.

Here, we describe the MECE platform (http://www.elabcaas.cn/pird/mece), a method for enhancing the catalytic efficiency of glycoside hydrolase proteins using a deep neural network and molecular evolution. This platform uses a deep neural network model trained with 119 different glycoside hydrolase family proteins (DeepGH) and Gradient-weighted Class Activation Mapping (Grad-CAM) to visualize features of the last convolutional neural network (CNN) layer in the DeepGH model. This platform includes a mutation score index based on extracted features from protein evolution data. The MECE strategy was then validated through application in two typical glycoside hydrolases (chitosanase CHIS1754 and glucoamylase GA51). The resulting mutants exhibited markedly enhanced catalytic efficiency. This method thus provides a valuable resource for rationally designed, high efficiency enzymes and proteins, and with the potential for broad application in other proteins or functions.

Results

Overview of the MECE prediction workflow

In order to design a robust workflow for engineering mutations that improve the catalytic efficiency of enzymes, we integrated functional feature extraction with evolutionary analysis-based mutation selection to establish the MECE platform. Based on their high demand across numerous industrial applications, we
selected glycoside hydrolases (GHs) as an enzyme family to explore and modify through a deep learning predictive platform. To this end, we downloaded all Gene IDs for known and predicted GHs from the Carbohydrate-Active enZYmes (CAZy) Database, as well as their corresponding protein sequences from NCBI. The protein sequence data was cleaned by removing the aberrantly short or long sequences and any sequences that belonged to multiple families. After removal of redundant sequences using CD-HIT software, the resulting datasets were used to train a DNN-based model for GH classification, or DeepGH (Fig. 1A). Next, we used the DeepGH model to reclassify the protein sequences collected from NCBI (Non-redundant protein sequence, nr), including all homologous sequences returned for the GH amino acid query sequence. All of the homologous sequences were then reclassified by the DeepGH model, and proteins that did not belong to the same family as the query sequence were removed. The informative features of each sequence that belonged to the same family as the query were extracted separately and visualized using Grad-CAM. All of these features were then used to build a functional feature matrix for the query sequence ($M_e$) (Fig. 1B). This strategy was used to identify functionally relevant sites for targeted mutations in chitosanase CHIS1754 and glycosylase GA51, which were experimentally validated and tested for changes in catalytic efficiency (Fig. 1C). To further and wider use of the MECE, we developed a website (http://www.elabcaas.cn/pird/mece). Thus, the MECE platform facilitates the design of GH variant enzymes based on co-evolutionary relationships among functionally relevant amino acid residues to improve protein function under different conditions.

**High accuracy classification of carbohydrate-active enzymes with the DeepGH predictive model**

In order to identify the typical characteristics of each GH family and classify them based on their respective amino acid sequences, we constructed a DNN model (DeepGH). To build this model, 854,279 GH GenelIDs and protein sequences were downloaded from CAZy and NCBI, respectively, and finally divided into 119 families (Fig. 2A). Protein sequence data was cleaned by removing truncated or excessively long sequence, sequences misannotated as belonging to multiple GH families, as well as redundant sequences predicted using the CD-HIT tool (Fig. 2B). The 119 remaining GH families (Table S3) were used to train the predictive model and extract functional features. One protein from each GH family was extracted as an independent test sample, and the resulting training and validation datasets were comprised of sequences ranging in similarity from 40–100% (5% interval).

Ten-fold cross-validation was used to train and evaluate the performance of the DeepGH models. Datasets with high similarity generally showed higher predictive accuracy, and overall, both the predicted performance in validation experiments and independent test data revealed that the threshold of 65% similarity was identified as the lowest similarity threshold for accurate classification by by DeepGH (Fig. 2C and 2D). Therefore, a 65% similarity threshold was used to construct training and validation datasets in subsequent analyses. Further ten-fold cross-validation with ROC (Receiver Operating Characteristic Curve) analysis using the validation dataset with 65% sequence similarity resulted in an average ACC (Accuracy) of 94.96%, and AUC (Area Under Curve) of 99.96%, 95.60% precision, 95.00%
Recall, and an F1-score of 95.00% (Fig. 2C). In the independent dataset, the average prediction accuracy of this model was 82.69% (Fig. 2D), which was slightly lower than that in the validation dataset possibly due to the imbalanced distribution of reclassified glycoside hydrolases in some families (Figure S2).

Finally, we optimized the size and number of the convolution kernels in DeepGH. In the optimized model, ten-fold cross-validation showed that ACC reached 96.73%, AUC was 99.97%, Precision was 97.00%, Recall was 97.00%, the F1-score was 97.00%, and the average prediction accuracy in this independent test increased to 90.76% (Figure S3 and Table S4). Additionally, ROC and PRC analysis of other machine learning methods (Bayes, Decision Tree, Logistic Regression and Random Forest) to train the predictive models indicated that the DeepGH model could provide higher accuracy, specificity, and sensitivity than the other four machine learning approaches (Fig. 2E and 2F). In addition, DeepGH could extract the vital features of the different GH families and discriminate the family of the GH target protein.

**DeepGH recognition of GH families and functional residues**

Upon completion of the DeepGH model, it was used to analyze the functional residues that contributed to the classification layer. Grad-CAM was used to extract the weight of each residue in the last CNN layer. The characteristic weights of all amino acids in this independent test dataset were then calculated and presented at 0.2 intervals, with amino acids that had a characteristic weight of 0 accounting for 27.5% of all amino acids, those with weights between 0-0.2 accounting for 61.0%, and amino acids with characteristic weights of 0.2–0.4, 0.4–0.6, 0.6–0.8, and 0.8-1.0 together comprising the remaining 11.5% (Fig. 3A). These results indicated that few residues in the protein provided relatively large contributions to the protein classification. Further, we analyzed the distribution of each amino acid with different feature weights. As shown in Fig. 3A, the weights of aspartic acid (D) and tyrosine (Y) residues accounted for a higher percentage ranged from 0.2 to 1.0 than that of zero, respectively, possibly due to their common function in catalytic or substrate binding sites in the protein. Obviously, glycine (G) had the highest proportion (except 0-0.2) among the weighted residues, and its proportion increased with increasing weight. Besides, G residues could be observed in the extracted features generally in close proximity to the active center, possibly because it could provide the flexibility necessary for the substrate binding pocket and the enzyme active sites to change conformation, which might help to improve the catalytic efficiency (Fig. 3C and 3D). Therefore, the method can help to extract those important functional residues in the protein.

To more clearly examine the classification features of GHs extracted by DeepGH, we selected a β-galactosidase (GH2, 1f4a), a glucoamylase (GH15, 1ayx), and a lysozyme (GH24, 1k28) in the independent test for analysis. The Grad-CAM toolkit was used to visualize the last layer of the convolutional neural network (CNN) generated by the trained DeepGH model. The total extracted features in the upper CNN layer for these three proteins are shown in Figure S4-S6. Conserved motif analysis was used to identify the two most functionally relevant sequence motifs (> 10 residues long). High scoring amino acids in these motifs were considered likely to contribute to the DeepGH-based classification layer and strongly affect protein function. At the same time, we obtained the advanced structures of these three proteins from the PDB database and annotated the active sites. Comprehensive analysis showed
that the high scoring motifs were generally located spatially close to the protein active center (Fig. 3B-D). These collective results showed that the trained DeepGH model could recognize the family of a GH protein, and could also identify its essential functional residues.

**Modification of chitosanase CHIS1754 by MECE to improve its catalytic efficiency**

To determine which residues in the target glycoside hydrolase were related to catalytic efficiency, the evolutionary and functional characteristics were identified by weighing the functional amino acid features identified in motif analysis with their relative conservation within their respective GH family (determined by pairwise alignment of wild-type and homologous proteins). In this study, the chitosanase CHIS1754, a GH46 family member, was chosen for proof-of-concept mutation analysis to validate the role of the identified residues in protein function. The functional amino acid features extracted by DeepGH were shown in Figure S7. Blastp was used to obtain sequences homologous to CHIS1754, and after removing redundant sequences, a dataset containing 563 chitosanase sequences was obtained. The feature matrix ($M_i$) of each sequence in the dataset was calculated, which was aligned with CHIS1754 respectively. Based on the alignment results, all sequence feature matrices were summed to obtain the feature matrix ($M_e$) related to the chitosanase function. $F_j$ was defined as the fold relationship between the mutation site $P_{j,\text{max}}$ with the highest importance score in each row of the matrix $M_e$ and the wild-type site $P_{j,\text{wt}}$ (Fig. 4A Table S5), with higher scores suggesting greater probability of improved catalytic efficiency. This analysis identified 9 residues with $F_j \geq 20$ and 17 residues with $F_j \geq 10$ in CHIS1754 (Table 1), which thus represented the strongest candidate mutations for increased activity. The top nine ($F_j \geq 20$) mutants were then selected for experimental verification. First, we used a heat map to visualize which amino acids were most likely to improve catalytic efficiency at these nine sites, with glutamate (E) and proline (P) emerging as obviously strong candidates at residues 76 and 213, respectively (Fig. 4B, Figure S8, Table 1).
Subsequently, we expressed and purified all of the above nine single residue conversion variants, as well as one variant harboring all nine mutations, in *E. coli* (Figure S9) and quantified their activity for comparison with the wild type. Using 1% water-soluble chitosan (degree of deacetylation > 92%) as substrate, eight of the single point mutants showed increased activity over that of WT at 55°C and pH5.5, *in vitro*, with the exception of L157E, which showed a ~35.71% reduction in activity. Among the variants, K76E showed the highest activity, approximately 7.00-fold that of WT. These results suggested that MECE could provide 88.89% predictive accuracy in mutation selection (Fig. 4C). Furthermore, the CHIS1754-MUT9 variant showed 3.14-fold higher activity than WT, 40.81% lower than K76E, suggesting that some of the nine mutations might have antagonistic effects. Co-evolutionary analysis to clarify the interactions between these mutated residues suggested that H142R negatively affected almost all other residue conversions, while Q104L conflicted with three other sites (E47D, K76E, and H142R) (Fig. 4D).

<table>
<thead>
<tr>
<th>NO.</th>
<th>sites</th>
<th>Wild-type</th>
<th>mutants</th>
<th>Fj</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>213</td>
<td>S</td>
<td>P</td>
<td>220.68</td>
</tr>
<tr>
<td>2</td>
<td>76</td>
<td>K</td>
<td>E</td>
<td>118.61</td>
</tr>
<tr>
<td>3</td>
<td>122</td>
<td>H</td>
<td>R</td>
<td>55.00</td>
</tr>
<tr>
<td>4</td>
<td>104</td>
<td>Q</td>
<td>L</td>
<td>45.18</td>
</tr>
<tr>
<td>5</td>
<td>157</td>
<td>L</td>
<td>E</td>
<td>44.31</td>
</tr>
<tr>
<td>6</td>
<td>160</td>
<td>S</td>
<td>R</td>
<td>32.05</td>
</tr>
<tr>
<td>7</td>
<td>119</td>
<td>D</td>
<td>P</td>
<td>26.12</td>
</tr>
<tr>
<td>8</td>
<td>47</td>
<td>E</td>
<td>D</td>
<td>24.74</td>
</tr>
<tr>
<td>9</td>
<td>142</td>
<td>H</td>
<td>R</td>
<td>20.79</td>
</tr>
<tr>
<td>10</td>
<td>161</td>
<td>M</td>
<td>V</td>
<td>15.10</td>
</tr>
<tr>
<td>11</td>
<td>235</td>
<td>L</td>
<td>K</td>
<td>14.59</td>
</tr>
<tr>
<td>12</td>
<td>92</td>
<td>T</td>
<td>C</td>
<td>12.69</td>
</tr>
<tr>
<td>13</td>
<td>204</td>
<td>H</td>
<td>K</td>
<td>12.60</td>
</tr>
<tr>
<td>14</td>
<td>209</td>
<td>H</td>
<td>K</td>
<td>11.56</td>
</tr>
<tr>
<td>15</td>
<td>208</td>
<td>E</td>
<td>A</td>
<td>11.30</td>
</tr>
<tr>
<td>16</td>
<td>221</td>
<td>E</td>
<td>A</td>
<td>10.58</td>
</tr>
<tr>
<td>17</td>
<td>168</td>
<td>D</td>
<td>R</td>
<td>10.45</td>
</tr>
</tbody>
</table>
Based on these findings, we next generated three multi-point mutants from the nine-mutation MUT9, including CHIS1754-MUT8, CHIS1754-MUT7, and CHIS1754-MUT6. More specifically, CHIS1754-MUT8 carried all mutations except L157E, which reduced enzymatic activity; CHIS1754-MUT7 carried all mutations in CHIS1754-MUT8 except for H142R, and CHIS1754-MUT6 carried all mutations in the MUT7 variant except for Q104L, which also had deleterious effects that masked the benefits of other mutations. After heterologous expression and purification, \textit{in vitro} activity assays, under the same conditions as above, showed that the CHIS1754-MUT7 variant had 14.53 times higher activity than that of WT, and 2.00-fold higher than that of the K76E mutant (Fig. 4E). Comparison of reaction kinetics among CHIS1754, K76E, and CHIS1754-MUT7 showed a $k_{cat}/K_m$ of 9.95 ± 1.69 mL mg$^{-1}$·s$^{-1}$ for CHIS1754, while K76E was approximately 10-fold higher at 109.14 ± 23.27 mL mg$^{-1}$·s$^{-1}$, while the $k_{cat}/K_m$ of CHIS1754-MUT7 was 189.89 ± 19.30 mL mg$^{-1}$·s$^{-1}$, or roughly 18.08-fold higher than WT (Table 2). Further structural modeling of CHIS1754 revealed that these seven mutation sites were distant from the catalytic center at E62 and E80 (Fig. 4F), and thus possibly affected chitosanase activity through conformational changes at the whole protein level. This likelihood was supported by significantly increased RMSF values at positions 126–129, 238–243, 267–272, but decreased at positions 62–67, 102–114 and 193–199 (Fig. 4G), in the higher efficiency CHIS1754-MUT7 variant. These results indicated that mutants predicted by the MECE platform could indeed increase enzymatic activity.

### Table 2

Kinetic parameters of CHIS1754 and GA51

<table>
<thead>
<tr>
<th>Chitosanase/Glycosylase</th>
<th>CHIS1754</th>
<th>K76E-CHIS1754</th>
<th>CHIS1754-MUT7</th>
<th>GA51</th>
<th>GA51-A557S</th>
<th>GA51-MUT5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K_m$ (mg·mL$^{-1}$)</td>
<td>2.51 ± 0.83</td>
<td>1.43 ± 0.48</td>
<td>1.22 ± 0.22</td>
<td>0.98 ± 0.25</td>
<td>0.63 ± 0.07</td>
<td>0.94 ± 0.38</td>
</tr>
<tr>
<td>$k_{cat}$ (s$^{-1}$)</td>
<td>24.98 ± 7.82</td>
<td>158.85 ± 38.37</td>
<td>231.67 ± 37.40</td>
<td>145.15 ± 12.77</td>
<td>388.41 ± 9.39</td>
<td>1203.32 ± 169.92</td>
</tr>
<tr>
<td>$V_{max}$ (µg·min$^{-1}$·mL$^{-1}$)</td>
<td>17.79 ± 2.78</td>
<td>27.74 ± 3.41</td>
<td>31.72 ± 2.56</td>
<td>6.89 ± 0.59</td>
<td>43.11 ± 1.20</td>
<td>7.94 ± 1.20</td>
</tr>
<tr>
<td>$k_{cat}/K_m$ (mL·mg$^{-1}$·s$^{-1}$)</td>
<td>9.95 ± 1.69</td>
<td>109.14 ± 23.27</td>
<td>189.89 ± 19.30</td>
<td>148.11 ± 12.77</td>
<td>612.34 ± 48.85</td>
<td>1279.78 ± 210.56</td>
</tr>
</tbody>
</table>

### Modification of glucoamylase GA51 by MECE to improve its catalytic efficiency

To further validate the accuracy MECE in predicting effective mutations, the GH15 family member glucoamylase GA51 was also tested. The functional amino acid features extracted by DeepGH were shown in Figure S10. The $F_j$ values of all amino acids in GA51 were also calculated (Fig. 5A, Table S6), which identified 5 residues with $F_j \geq 20$ and 9 residues with $F_j \geq 10$ in GA51 (Table 3). To experimentally verify the effects of the top five sites with $F_j \geq 20$, a heat map was again generated to identify the most
likely amino acids that could improve catalytic efficiency at each of the five sites (Fig. 5B, Figure S11, Table 3).

Table 3

<table>
<thead>
<tr>
<th>NO.</th>
<th>sites</th>
<th>Wild-type</th>
<th>mutants</th>
<th>$F_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>304</td>
<td>A</td>
<td>S</td>
<td>42.0</td>
</tr>
<tr>
<td>2</td>
<td>590</td>
<td>G</td>
<td>S</td>
<td>28.5</td>
</tr>
<tr>
<td>3</td>
<td>557</td>
<td>A</td>
<td>S</td>
<td>28.3</td>
</tr>
<tr>
<td>4</td>
<td>576</td>
<td>L</td>
<td>T</td>
<td>21.1</td>
</tr>
<tr>
<td>5</td>
<td>397</td>
<td>I</td>
<td>V</td>
<td>21.0</td>
</tr>
<tr>
<td>6</td>
<td>207</td>
<td>N</td>
<td>A</td>
<td>13.9</td>
</tr>
<tr>
<td>7</td>
<td>606</td>
<td>A</td>
<td>S</td>
<td>13.9</td>
</tr>
<tr>
<td>8</td>
<td>563</td>
<td>D</td>
<td>N</td>
<td>11.0</td>
</tr>
<tr>
<td>9</td>
<td>571</td>
<td>D</td>
<td>S</td>
<td>11.0</td>
</tr>
</tbody>
</table>

The WT GA51, each of the five single point mutants, and variant carrying all five residue conversions were expressed and purified in *P. pastoris* GS115 (Figure S12). Evaluation of catalytic activity *in vitro* using soluble starch and colorimetric detection of reducing sugars showed that each of the nine single mutation variants exhibited at least two-fold greater specific activity than WT, with the A557S variant showing the greatest increase at 26.71 ± 0.15 U/mg, approximately 2.87 times that of WT (Fig. 5C), while its $k_{cat}/K_m$ was 612.34 ± 48.85 mL mg$^{-1}$ s$^{-1}$, 3.13 times higher than WT. By contrast, the GA51-MUT5 variant had a specific activity of 54.30 ± 5.11 U/mg, 4.84 times greater than WT, and a $k_{cat}/K_m$ of 1279.78 ± 210.56 mL mg$^{-1}$ s$^{-1}$, representing a 7.64-fold increase over WT (Table 2).

Co-evolutionary analysis showed that there were no conflicts among the five converted amino acids (Figure S13). In addition, three-dimensional structural analysis of GA51-MUT5 revealed that G590, L576, and A557 were located in the substrate binding domain, while L576 and A557 served as binding sites for the starch substrate (Fig. 5D). Calculation of RMSF values indicated that these mutations resulted in substantial fluctuations, which could affect substrate binding, and consequently, catalytic efficiency (Fig. 5E). These cumulative results indicated that the MECE platform could be used to predict mutations that effectively enhance the catalytic efficiency of GH family proteins.

**Discussion**

In this study, we established a data-driven approach, MECE, to design mutants with enhanced catalytic efficiency. The MECE platform combines a deep learning-based classification model for glycoside
hydrolases with a component for selection of key features in the last convolutional neural network (CNN) layer, and a mutation score indicator based on features extracted from protein evolutionary data. Our results are consistent with earlier reports that describe machine learning approaches for classifying different proteins based on the sequences of a large number of proteins. Here, features potentially related to protein function were recognized in the DNN models. Then, protein evolutionary data were incorporated to establish the mutation score indicator. When applied to two typical glycoside hydrolases (chitosanase CHIS1754 and glucoamylase GA51), the catalytic efficiency of the resulting variants was substantially higher than that of the wild type. These experimental results thus demonstrated the predictive power, computational efficiency, and low experimental cost of the MECE platform. Moreover, this method is suitable for application with most glycoside hydrolase families, and can adapted to other protein families.

Glycoside hydrolases are particularly important due to their wide range current and prospective applications. In recent years, the number of GHs in the CAZy database has exploded. Their classification is generally based on blastp similarity with other proteins, and functional annotation mainly relies on literature searches, assessment of the reliability of the data, and its storage/integration into the database. Without careful curation, the number of invalid sequences, including redundant, truncated, or low-quality sequences, can proliferate within the database. Based on the assumption that the features used for classification are indeed related to catalytic activity, in this study MECE was used to reorganize the glycoside hydrolases in CAZy and extracted the relevant classification features, consequently providing an effective guide for GH analysis.

There are also websites and software tools for predicting the domains, families and functional sites of enzymes, such as prosite (https://prosite.expasy.org/) and MEME (https://meme-suite.org/meme/), which can predict information such as conserved structural domains or active centers of enzymatic proteins, but cannot learn from them, as MECE does, the specific amino acid residues associated with catalytic efficiency.

Deep learning methods have been used in other studies to reclassify or functionally annotate enzymes, among other uses. For example, DeepFRI, developed by Gligorijević, was trained on protein structures from the PDB and SWISS-MODEL and could predict both GO terms and protein EC numbers by extracting local sequences and global structural features. Although the structural features of proteins can determine a wide range of functions, and may be very useful for classifying unknown proteins, this approach does not help with further engineering its functionality. Even for proteins that are correctly classified, functional predictions are approximated based on their family, and prediction of key amino acids is extremely limited. The power of MECE lies in its ability to directly identify the key amino acids related to catalytic function based on sequence features, and its prediction of residue conversions that will most effectively enhance catalytic efficiency.

**Methods**
Construction of a glycoside hydrolase database

This study focused on the glycoside hydrolases, of which there are 169 families in the CAZy database, including the GH0 family, which contains unclassified sequences, and 7 families that contain no reference sequences. For the remaining 162 families, we obtained the corresponding GenBank numbers through the CAZy website (http://www.cazy.org/) and downloaded the corresponding amino acid sequences using the Batch Entrez port and Biopython toolkit provided by the National Center for Biotechnology Information database (NCBI, https://www.ncbi.nlm.nih.gov/). Next, sequences that were too long or short, existed in several families at the same time, or were redundant were removed. We set the amino acid sequence length to $\geq 200$ and $\leq 1000$ to eliminate nonfunctional proteins and irregular sequences, respectively. After the above processing, 327,302 sequences remained.

To further improve the generalization ability of the model, the sequences were de-duplicated using CD-HIT. A series of thresholds were set from 40–100% with a gradient at 5% intervals to determine the optimal threshold for judging sequence similarity. Finally, the families in which the number of sequences was greater than or equal to 10 at a sequence similarity threshold of 40% were chosen, and a total of 119 families remained.

The above processed sequences were further divided into training, validation, and test sets, respectively. The test set refers to a dataset that is not involved in the training process of the model at all. In this study, the hmmerscan, a method calculated using hidden Markov models, was used to confirm that the sequences with the highest scores from each family were chosen as the test set, for 119 in total. The sequences not chosen for the test set were used as the training and validation sets.

Deep neural network model construction

A deep neural network (DNN) architecture was used. Specifically, a convolutional layer + pooling layer + convolutional layer + pooling layer + fully connected layer architecture was employed; the detailed architecture of the model was shown in Figure S1. The final number of epochs was 30, the batch_size was 128, the num_classes was 119, and the max_seq_len was 735 that was the average length plus one standard deviation of the length of all sequences in the dataset. To improve the generalization performance of the model, we also added a Dropout layer. In addition, ReLU was used as the activation function, Categorical Cross Entropy Loss as the loss function, and Adam as the optimization algorithm. A 10-fold cross-validation strategy was used to train and evaluate the performance of the model, and the data in the metrics were calculated using the keras package. Two hyperparameters, the convolutional kernel size (kernel_size) and the number of convolutional kernels (filters), were tuned and compared using a grid search approach. The codes for the above CNN architecture with PYTHON v.2.7 (compatible with v.3.5) environments are available from GitHub (https://github.com/BRITian/MECE).

Visualization of family characteristics
The Gradient-weighted Class Activation Mapping (Grad-CAM) method was used to perform feature extraction on the already constructed 119-family convolutional neural network (CNN) model to resolve each glycoside hydrolase and its functionally relevant key loci and associated features. The principle was to use the gradient information accompanying the convolutional network at the last layer of the CNN and weight the corresponding gradient to construct the heat maps of model classification to understand the importance of each neuron for a decision. The ggseqlogo tool was used in the R software to visualize heat maps on sequences to facilitate visual evaluation and representation of analytical features.

**Weighted analysis of family and evolutionary characteristics**

To improve the positivity of the designed mutant sites, we analyzed the evolutionary features of the target protein (CHIS1745 and GA51). The specific methods were as follows. First, the homologous sequences were collected using the online NCBI Blast tool with 60% < identity < 100%, 50% < coverage < 100%, and E-value < e⁻⁵. Then CD-HIT was used to remove the redundant sequences with similarity > 90%. Next, the remaining sequences were submitted to the DeepGH model to remove sequences that did not belong to the GH46 or GH15 families. The Grad-CAM method was used to calculate the feature matrix \( M_i \) of each sequence that was retained, and the matrix consists of \( l \times n \) elements arranged as \( P_{ab} \in M_i \) \((a = 1, 2, \ldots, l, b = 1, 2, \ldots, n)\) in a rectangular table of \( l \) rows and \( n \) columns, where \( l \) was the length of the i-th sequence and \( n \) took a value in the range of 1–20, representing the 20 amino acids.

\[
M_i = \begin{bmatrix}
P_{11} & \cdots & P_{1n} \\
\vdots & \ddots & \vdots \\
P_{l1} & \cdots & P_{ln}
\end{bmatrix}
\]

Next, a pairwise local alignment of each sequence homologous to the wild type was performed using the Water tool, with gap open set to 10 and gap extend to 0.5. Based on the results of sequence alignment, the functionally relevant evolutionary feature matrix \( M_e \) of the wild type was obtained by summing all sequence feature matrices using the wild type as the standard. The number of rows of this matrix was the sequence length of wild type and the number of columns was 20 (i.e., 20 amino acids).

\[
M_e = \begin{bmatrix}
P_{1,1} & \cdots & P_{1,n} \\
\vdots & \ddots & \vdots \\
P_{l1} & \cdots & P_{l,n}
\end{bmatrix}
\]

The difference between the mutant site \( P_{j,\text{max}} \) with the highest importance score in each row of the \( M_e \) and the wild-type site \( P_{j,\text{wt}} \) was compared, and the ploidy relationship \( (F_j) \) between \( P_{j,\text{max}} \) and \( P_{j,\text{wt}} \) was calculated using the division method. A decimal (\( \alpha \)) was added to the values of all elements to avoid a division of 0. The final formula was:
The sites with $F_j \geq 20$-fold were selected as single point mutants. The loci with $F_j \geq 20$-fold were selected according to the ploidy size to design a multipoint mutant, which was synthesized by General Bio (Anhui) Co., Ltd.

**Construction of mutants**

The two-step PCR mutagenesis strategy \(^{27}\) was used to construct the mutants. The plasmids pET28a::\textit{chis1754} and Ppic9\gamma::\textit{GA51} were used as the respective templates; the primers are listed in Table S1 and S2. The PCR reaction was as follows: preheat at 95°C for 5 min; followed by 32 cycles of 95°C for 30 s, 54°C (chitosanase) or 58°C (glucoamylase) for 30 s, and 72°C for 45 s (chitosanase) or 1 min (glucoamylase), followed by a final extension at 72°C for 10 min. The target fragments were purified by the DNA Gel Extraction Kit (TianGen, Beijing, China) and used as the primers for the second round of PCR amplification, in which the above plasmids were used as templates. The second-round PCR reaction was as follows: 95°C for 5 min; followed by 32 cycles of 95°C for 30 s and 72°C for 4.5 min (chitosanase) or 5.5 min (glucoamylase); followed by a final extension at 72°C for 10 min. Residual template plasmids in the reaction system were eliminated using \textit{DpnI} (NEB, Ipswich, United Kingdom), then the PCR products were purified and recovered using a purification and recovery kit (TianGen, Beijing, China). Finally, the fragments were transformed into \textit{Escherichia coli} Top10 competent cells using standard procedures. Single clones were verified by PCR and sequenced by TSINGKE Biotechnology (Beijing, China). Multipoint mutants were synthesized by General Bio (Anhui) Co., Ltd.

**Heterologous expression and purification of chitosanase in \textit{E. coli}**

The successfully constructed plasmids were transformed into \textit{E. coli} BL21 (DE3) for heterologous expression. Single clones were picked and inoculated into 3 mL LB medium containing 50 µg/mL kanamycin, then cultured at 37°C and 200 rpm for 12 h to prepare the seed fluid. Then, 2 mL seed fluid was inoculated into 200 mL LB medium containing 50 µg/mL kanamycin and the mixture was cultivated at 37°C and 200 rpm for 6–8 h until reaching an OD\textsubscript{600} of 0.6–0.8. Then, 0.5 mM (final concentration) of isopropyl-β-d-thiogalactopyranoside (IPTG) was added to the mixture followed by incubation at 16°C and 200 rpm for 18–20 h to induce expression of the target protein. The cells were collected by centrifugation at 4°C and 8000 rpm for 10 min. Then the pellet was resuspended the in 20 mL Tris-HCl (20 mM, pH 7.2), followed by sonication on ice for 20 min. Next, the mixture was centrifuged at 4°C and 8000 rpm for 30 min and the supernatant containing the target protein was retained. The whole mixture was immediately loaded onto a Ni-NTA column (GE, Boston, MA, USA), and the impure proteins were washed successively with 10 mL NTA-0, NTA-20, and NTA-40 (50 mM NaCl, 50 mM Tris, and 0, 20, and 40 mM imidazole, pH 7.2, respectively). The target recombinant protein was finally eluted with 10 mL NTA-200 (50 mM NaCl, 50 mM Tris, and 200 mM imidazole, pH 7.2). The purified protein was dialyzed against Tris-HCl (pH 7.2) for 12 h, concentrated with PEG 8000, and quantified by a BCA protein quantification kit (TIANGEN, Beijing, China).
Heterologous expression and purification of glucoamylase in *Pichia pastoris*

The successfully constructed recombinant plasmids were linearized with *BglII*, then enriched by isopropanol precipitation. Then, 30 µg the linearized plasmids were transformed into competent yeast GS115 (His–Mut+) cells and spread on MD solid medium (20.0 g/L glucose, 100 mL/L 10× YNB, 2 mL/L 500× biotin, and 20 g/L agar), then cultured at 28°C for 2 days. Next, single clones were picked and cultured in 3 mL YPD complete medium (20.0 g/L tryptone, 10.0 g/L yeast extract, and 20.0 g/L glucose) at 28°C and 200 rpm for another 2 days. Then, 100 µL of the above seed solution was transferred to 10 mL BMGY medium (20.0 g/L tryptone, 10.0 g/L yeast extract, 10 mL/L glycerol, 100 mL/L 10× YNB, 2 mL/L 500× biotin, and 0.5% methanol) and incubated at 28°C and 200 rpm for 2 days. The yeast pellet was collected by centrifugation under sterile conditions, resuspended in 5 mL BMMY medium (20.0 g/L tryptone, 10.0 g/L yeast extract, 100 mL/L 10× YNB, and 2 mL/L 500× biotin), and incubated at 28°C and 200 rpm for 72 h while supplemented with 0.5% (final concentration) methanol at 24 h intervals to induce target protein expression. Finally, the mixture was centrifuged at 4°C and 5000 rpm for 5 min, and the supernatant (the crude enzyme solution containing the target protein) was collected.

The crude enzyme solution was concentrated using a 1 KDa membrane, dialyzed in 10 mM citric acid-phosphate buffer for 12 h, and then purified by AKATA through an anion exchange column. Briefly, the column was equilibrated with liquid A (citrate-phosphate buffer, 10 mM, pH 6.5) and liquid B (NaCl, 1 M, pH 6.5) at a 3:2 ratio, the proteins were loaded into the upper sample cup, the different proteins were separated by elution with a concentration gradient of NaCl, and the solution containing the target proteins was collected. Pure proteins were concentrated using ultrafiltration tubes (Merck-Millipore, Germany) and quantified using a BCA protein quantification kit.

**Determination of chitosanase activity**

The chitosanase activity was measured using the 3,5-dinitrosalicylic acid solution method (DNS). The reaction system consisted of 500 µL 50 mM acetic acid–sodium acetate buffer (pH 5.5), 450 µL 1% water-soluble chitosan (degree of deacetylation > 92%), and 50 µL chitosanase. The mixture was incubated at corresponding temperatures for 15 min and stopped by adding 750 µL DNS, followed by boiling for 5 min to develop the color. Next, the mixture was immersed in an ice water bath for 2 min, then centrifuged at 7200 × g for 1 min to obtain the supernatant. Then, 200 µL supernatant was used to measure the absorbance at 520 nm using the SpectraMax M2 microplate reader (Molecular Devices, Silicon Valley, USA). One unit of chitosanase activity was set as the amount of chitosanase required to produce 1 µM/min reducing sugar (calculated as glucosamine).

**Determination of glucoamylase activity**

The DNS assay was used with some modifications to measure the activity of glucoamylase GA51. The reaction substrate, 2% soluble starch, was prepared by dissolving 2 g starch in 100 mL water, then boiling for 10 min for gelatinization. The reaction system consisted of 450 µL 2% soluble starch, 450 µL citric acid–disodium hydrogen phosphate buffer (0.2 M, pH 3.5), and 100 µL glucoamylase (1 mg/mL). The
mixture was reacted at 65°C for 30 min, followed by addition of 1.5 mL DNS to terminate the reaction, and boiling for 5 min to develop the color. Finally, after cooling, the absorbance at 540 nm was measured as described above. One unit of glucoamylase activity was defined as the amount of glucoamylase required to catalyze the production of 1 µM/min reducing sugar (calculated as glucose) under optimal reaction conditions.

## Determination of kinetic parameters of chitosanase and glucoamylase

The chitosan substrate was formulated to various concentrations (0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.25, 1.5, 1.75, 2, 3, 4, or 5 mg/mL) and then the chitosanase activity was measured at pH 5.5 and 55°C according to the corresponding system. The soluble starch substrate was set to 0.2, 0.3, 0.5, 0.8, 1, 1.3, 1.8, 2, 2.5, 3.5, 5, or 7.5 mg/mL, then the activity of glucoamylase was measured at pH 3.5 and 55°C according to the corresponding system. Finally, the Mie equation curves were fitted with the substrate concentration and enzymatic reaction rate as the horizontal and vertical coordinates to calculate $K_m$, $k_{cat}$, and $V_{max}$.

## Data availability

All the data generated or analyzed during this study are included in this published article (and its Supplementary Information files) or available from the authors upon request.

## Declarations

### Acknowledgements

We thank Prof. Yunliu Fan for her valuable suggestions on this research. The authors thank Dr Jaie Woodard for their careful revising of the manuscript and many useful comments.

### Author information

## Authors and Affiliations

**Biotechnology Research Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, China.**

Feifei Guan, Hanqing Liu, Tuoyu Liu, Lixin Yang, Xiaoqing Liu, Ningfeng Wu and Jian Tian

**Institute of Animal Science, Chinese Academy of Agricultural Sciences, Beijing 100193, China.**

Huiying Luo, Bin Yao, Huoqing Huang and Jian Tian

## Contributions

Feifei Guan and Jian Tian conceived and designed the project. Feifei Guan, Hanqing Liu, Tuoyu Liu and Lixin Yang performed and analyzed the experimental data. Feifei Guan, Hanqing Liu and Tuoyu Liu wrote
the manuscript with input from all authors and discussed the manuscript with Huoqing Huang and Jian Tian. Huoqing Huang and Jian Tian supervised the work. Xiaoqing Liu, Ningfeng Wu, Huiying Luo and Bin Yao analyzed the experimental data and provided some suggestions.

**Corresponding authors**

Correspondence to Huoqing Huang or Jian Tian.

**Ethics declarations**

Competing interests

The authors declare no competing interests

**FUNDING**

This work was supported by the National Key R&D Program of China (Grant No. 2021YFC2100300), the Central Public-interest Scientific Institution Basal Research Fund [Grant numbers Y2019XK19, 1610392021008 and 1610392020001].

*Conflict of interest statement.* None declared.

**References**


**Figures**
Computational workflow for enhancing the catalytic efficiency of glycoside hydrolase proteins based on deep neural networks and molecular evolution. (A) Data collection and DeepGH model construction. (B) DeepGH combined with molecular evolution to design mutations to enhance the catalytic efficiency of target enzymes. (C) Experimental validation and mechanism analysis.
A deep neural network model (DeepGH) that can classify glycoside hydrolases according to their sequences. (A) The sequence distribution of each family of glycoside hydrolases. The figure only shows families with a proportion greater than or equal to 0.6%; the remaining families are classified into the 'others' category. GH, glycoside hydrolase. (B) The number of sequences under different sequence similarity thresholds. (C) Prediction accuracy of DeepGH at different sequence similarity thresholds based on the results of 10-fold cross-validation. (D) Prediction accuracy of DeepGH at different sequence similarity thresholds based on the results of independent tests. (E, F) Receiver operating characteristic (ROC) curve (E) and Precision-Recall curve (PRC) (F) of DeepGH output for predicting the glycoside hydrolase families. ROC and PRC were calculated from classifying glycoside hydrolases based on the results of 10-fold cross-validation.
Figure 3

Extraction and visualization of amino acid features related to catalytic efficiency. (A) Distribution and ratio analysis of all amino acids encoding glycoside hydrolases in the independent test. (B–D) The location of the top two motifs in β-galactosidase (GH2, 1f4a), glucoamylase (GH15, 1ayx), and lysozyme (GH24, 1k28). The residues were colored using a gradient color scheme matching the feature weight: residues with higher scores highlighted in red and lower scores highlighted in blue. The activity sites are marked in yellow.
Figure 4

**Enhancing the catalytic efficiency of chitosanase CHIS1754 using MECE.** (A) The $F_j$ of all positions of CHIS1754; (B) The feature score ($P_{ja}$) of CHIS1754 at nine positions with $F_j \geq 20$, "j" represented the residue number and “a” represented the 20 amino acids; (C) Detection of the activities of single-point mutants and multi-point mutant CHIS1754-MUT9; (D) Coevolutionary analysis between single-point mutations; (E) Detection of the activities of multi-point mutants and single-point mutant K76E. (F) Structural analysis of CHIS1754-MUT7 by Chimera. All residues are colored using a gradient color scheme matching the $F_j$ value: residues with higher $F_j$ values highlighted in red and lower values highlighted in blue. The activity sites are marked in black. (G) Simulation analysis of CHIS1754-MUT7 dynamics.
Figure 5

**Enhancing the catalytic efficiency of glycosylase GA51 using MECE.** (A) The $F_j$ of all positions of GA51; (B) The feature score ($P_{j,a}$) of GA51 at nine positions with $F_j \geq 20$, "j" represented the residue number and "a" represented the 20 amino acids; (C) Detection of the activities of single-point mutants and multi-point mutant GA51-MUT5. (D) Structural analysis of GA51-MUT5 by Chimera. All residues are colored using a gradient color scheme matching the $F_j$ value: residues with higher $F_j$ values highlighted in red and lower values highlighted in blue. The activity sites are marked in black. (E) Simulation analysis of GA51-MUT5 dynamics.

**Supplementary Files**

This is a list of supplementary files associated with this preprint. Click to download.
- rs.pdf
- Suppdata.docx
- TableS5.xls
- TableS6.xls