biology

MDPI

Article

Different Recognition of Protein Features Depending on Deep Learning Models: A Case Study of Aromatic Decarboxylase UbiD

Naoki Watanabe [1,†], Yuki Kuriya [1,†], Masahiro Murata [2], Masaki Yamamoto [1], Masayuki Shimizu [3] and Michihiro Araki [1,2,4,5,*]

[1] Artificial Intelligence Center for Health and Biomedical Research, National Institutes of Biomedical Innovation, Health and Nutrition, 3-17 Senrioka-shinmachi, Settsu 566-0002, Japan; n.watanabe@nibiohn.go.jp (N.W.); kuriya@nibiohn.go.jp (Y.K.); m.yamamoto@nibiohn.go.jp (M.Y.)
[2] Graduate School of Science, Technology and Innovation, Kobe University, 1-1 Rokkodai, Nada-Ku, Kobe 657-8501, Japan; murata.masahiro.8r@people.kobe-u.ac.jp
[3] Bacchus Bio Innovation Co., Ltd., 6-3-7 Minatojima minami-machi, Kobe 650-0047, Japan; m_shimizu@b2i.co.jp
[4] Graduate School of Medicine, Kyoto University, 54 Shogoin-Kawahara-cho, Sakyo-ku, Kyoto 606-8507, Japan
[5] National Cerebral and Cardiovascular Center, 6-1 Kishibe-Shinmachi, Suita 564-8565, Japan
* Correspondence: araki@nibiohn.go.jp; Tel.: +81-6-6384-1573
† These authors contributed equally to the work.

Simple Summary: Various protein sequences are registered in biological databases, and hundreds of the sequences have recently been sequenced by way of next-generation sequencing, and then the number of sequences with unknown functions is explosively increasing. To efficiently determine the annotations, new feature extraction of protein sequences that is different from existing knowledge is required. Deep learning can extract various features based on training data. Many studies have reported deep learning models with high accuracy for predicting protein annotations; however, in the reports, which amino acid sites in protein are important for the prediction of the annotations have not been discussed among multiple deep learning models. Here, 3 deep learning models for the prediction of the proteins included in a protein family were analyzed using an explainable artificial intelligence method to explore important protein features. The models regarded different sites as important for each model, and all models also recognize different amino acids from the secondary structure, conserved regions and active sites as important features. These results suggest that the models can interpret protein sequences through different perspectives from existing knowledge.

Abstract: The number of unannotated protein sequences is explosively increasing due to genome sequence technology. A more comprehensive understanding of protein functions for protein annotation requires the discovery of new features that cannot be captured from conventional methods. Deep learning can extract important features from input data and predict protein functions based on the features. Here, protein feature vectors generated by 3 deep learning models are analyzed using Integrated Gradients to explore important features of amino acid sites. As a case study, prediction and feature extraction models for UbiD enzymes were built using these models. The important amino acid residues extracted from the models were different from secondary structures, conserved regions and active sites of known UbiD information. Interestingly, the different amino acid residues within UbiD sequences were regarded as important factors depending on the type of models and sequences. The Transformer models focused on more specific regions than the other models. These results suggest that each deep learning model understands protein features with different aspects from existing knowledge and has the potential to discover new laws of protein functions. This study will help to extract new protein features for the other protein annotations.

check for updates

Citation: Watanabe, N.; Kuriya, Y.; Murata, M.; Yamamoto, M.; Shimizu, M.; Araki, M. Different Recognition of Protein Features Depending on Deep Learning Models: A Case Study of Aromatic Decarboxylase UbiD. *Biology* **2023**, *12*, 795. https://doi.org/10.3390/biology12060795

Academic Editor: Lin Lu

Received: 21 April 2023
Revised: 17 May 2023
Accepted: 29 May 2023
Published: 31 May 2023

Keywords: deep learning; protein feature; feature extraction; explainable artificial intelligence; integrated gradients; protein annotation

1. Introduction

Protein sequence information is registered in various biological databases [1,2]. Various protein sequences are being sequenced by next-generation sequencing technology [3]. The number of unannotated sequences registered in the databases is explosively increasing, such as putative proteins, hypothetical proteins, and uncharacterized proteins. Therefore, in addition to efficiently assigning the annotations for a large number of proteins, extracting new protein features that differ from existing knowledge is required.

Deep learning automatically can learn and extract various features of input data, and the higher the model performances are, the more valid the training data are. Therefore, the utilization of deep learning is expected to discover important features and classify data based on the features [4]. Several studies have reported deep learning models for predicting protein functions [5,6], protein structures [7–12], multi-domain protein structures [13,14], protein subcellular localization [15,16], enzyme commission numbers [6,17–19], and products in organic synthesis [20,21]. Each model has been evaluated and compared using multiple performance evaluation parameters in machine learning tasks. Although the evaluation of model prediction accuracy is important, these studies have not sufficiently discussed which features of input data influence prediction accuracy and have not evaluated the detailed difference of the results among multiple models.

Most of the deep learning models cannot interpret prediction results without the other methods. However, deep learning has the potential to recognize extensive and new protein features that are different from existing knowledge, such as secondary structures, conserved residues, ligand binding sites, and active sites, because the models achieve more adequate prediction accuracy than previous machine learning. Model interpretability helps to know how the model reaches the results and to quantify prediction reliability [22,23]. Several studies have recently reported integrated gradients (IG) and Shapley additive explanations (SHAP), included in explainable artificial intelligence methods [24,25], to interpret prediction models and to explore important features for prediction [6,26,27]. However, the previous reports using integrated gradients have not discussed the exploration of important features and the difference of the features among multiple deep learning models.

Here, several deep learning models derived from enzyme sequences were developed to extract protein features for each amino acid residue and then to explore the validity of the features in comparison to previously reported information. As a case study, UbiD enzymes, one of the decarboxylases which biosynthesize various aromatic compounds [28–35], were used. To extract new UbiD features, prediction and feature extraction models for UbiD were built using convolutional neural network (CNN), CNN-based autoencoder (CNN-AE), and Transformer [36–38]. The important protein features between these models were explored by analyzing prediction scores and feature vectors derived from the models using clustering and IG (Figure 1). As a result, UbiD features could also be extracted from the different residues from the existing knowledge by these models, and the features were varied for each model and sequence, and only the Transformer model characterized a few amino acid residues as important UbiD features. The results indicate that each deep learning model extracts different protein features from each amino acid and recognizes each sequence as different. In short, the analysis of protein features using multiple explainable deep learning is required to more deeply understand proteins.

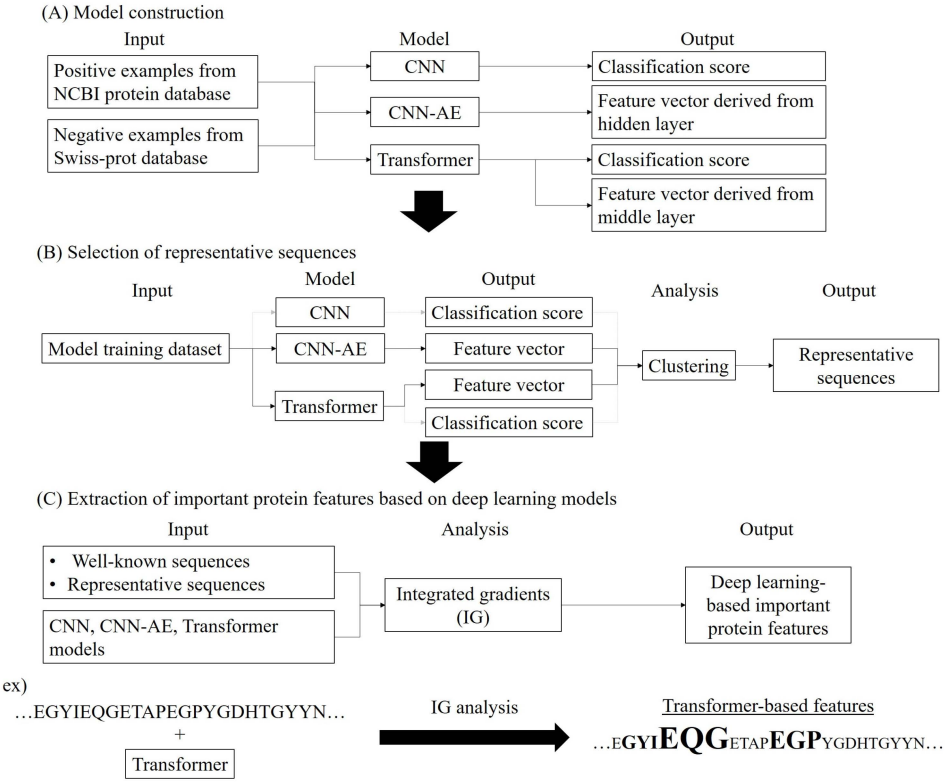

Figure 1. Workflow of methodology. (**A**) Model construction and model information, (**B**) Selection of representative sequences using deep learning and clustering, (**C**) Extraction of important protein features based on deep learning using Integrated gradients.

2. Materials and Methods

2.1. Dataset Construction

2.1.1. Positive Data

25,294 UbiD sequences were collected as positive data from National Center for Biotechnology Information (NCBI) Protein database [1] on 31 July 2019 by searching UbiD as a keyword. The enzyme sequences that were duplicated or included non-canonical amino acids were removed. The length of amino acid residues was limited from 400 to 700 because the length of 25,135 UbiD sequences was in the range. The sequences were clustered at 95 % identity using CD-HIT [39] to remove sequence redundancy and then were split into training, validation, and test data based on the number of sequences included in each cluster.

The 3 sequences were randomly extracted from all sequences included in a cluster and were split into each data when the number of sequences was 3 or more. When the number was 2, one sequence was added to training data, and the other was added to either validation or test data. In the rest of the cases, the sequences collected from all clusters, in which the number of sequences was 1, were randomly split into training, validation, and test data, at an approximate ratio of 8:1:1.

2.1.2. Negative Data

All protein sequences registered in Swiss-Prot [2] were collected as negative data on 26 April 2019. The negative data was the protein sequences except for UbiD. Some sequences were also removed in the same way as positive data construction. The highly similar negative sequences to positive sequences were omitted at 1.0×10^{-10} E-value using BLAST+ 2.7.1 [40]. The rest of the sequences were clustered at 90% identity using CD-HIT, and then only a single enzyme sequence from each cluster was included.

Artificial negative data were built to prevent deep learning models from judging as positive using only a few amino acids in a specific position. In total, 400 sequences whose same dipeptide amino acids continue were generated (AA . . . AA, AC . . . AC, . . .). The length of the upper negative data and the artificial sequence was randomly determined from 400 to 700, such as the construction of positive data. All negative data were randomly split into training, validation, and test data, at an approximate ratio of 8:1:1. Total amounts of positive and negative data are shown in Table 1.

Table 1. UbiD datasets for CNN, CNN-AE, and Transformer models.

Dataset Category	Training	Validation	Test
Positive data	1593	646	645
Negative data	62,476	8168	8167

2.2. Model Construction and Evaluation

CNN model, which predicts whether or not input protein sequences are target enzymes, was built. CNN-AE model to output feature vectors derived from enzyme sequences was built. The autoencoder model transformed input protein sequences to low dimensional feature vectors and outputted similar sequences to input data. Finally, Transformer model was built for prediction of target enzymes, such as CNN model, and for output of feature vectors, such as CNN-AE model (Figure 1A). The prediction scores and feature vectors outputted from the models were evaluated using clustering and IG (Figure 1B,C).

The architecture of CNN prediction and CNN-AE feature extraction models are shown in Supplementary Figures S1 and S2. The 3 hidden layers were used in CNN model, and self-attention was inserted next to the second hidden layer. The 5 hidden layers were used in encoder and decoder of CNN-AE, respectively, and 200 dimensional feature vectors were outputted. Self-attention was inserted next to the first hidden layer of the encoder and the fourth layer of the decoder. One-hot matrices transformed from amino acid sequences were inputted to both models. The sequences whose number of amino acids was less than 700 were transformed to matrices using zero-padding. CNN-AE model was built using only positive data.

Transformer model, which predicted target enzymes and extracted features derived from sequence information, was built using the encoder of Transformer [38], as shown in Supplementary Figure S3. Enzyme sequences were transformed to the tokens using 3-gram model. The special tokens (<CLS>, <EOS>) were used at the beginning and end of each token. <pad> tokens were added up to 700 tokens for the sequences whose number of amino acids was less than 700. In total, 64 dimensional vectors in Extract layer (Figure S3) were used to analyze feature vectors. A binary cross-entropy loss function was used to train all models.

CNN and Transformer models were trained using several batches, including only positive or only negative sequences to prevent overfitting due to imbalanced data. The positive and negative batches were separately built using random sampling without replacement, and then the models were learned for each batch in turn. If the number of the sequences that could be extracted was less than that of batch size in batch construction, the following batches were rebuilt using the first data. CNN model was trained until 4000 steps, while Transformer model was trained until 1000 steps.

CNN and Transformer models were evaluated using test data. Accuracy (ACC), AUC, F_1 score, and Matthews correlation coefficient (MCC) were used to evaluate the prediction models. CNN-AE model was evaluated using Match rate between input sequences and output sequences, given by the following:

$$\text{Match rate} = \frac{Matched\ number\ of\ amino\ acids\ in\ sequence}{Sequence\ length}, \quad (1)$$

The CNN-AE model using the epochs, where the number of sequences with Match rates 0.9 and more, and the average of Match rates were highest, were used in the following analysis. The CNN and Transformer models using the epochs, where all evaluation parameters were highest, were used. All models were built by Tensorflow version 2.1.0 [41].

2.3. Case Study

UbiD enzymes were used to explore important enzyme feature vectors between the deep learning models. The enzymes catalyze the decarboxylation reactions included in the ubiquinone biosynthesis pathway, which were identified in *Escherichia coli* for the first time [28,29,42]. Usual UbiD enzymes act on para-hydroxybenzoic acid-type substrates, while the other UbiD family enzymes catalyze the reversible reactions to synthesize various aromatic compounds such as protocatechuic acid and vanillic acid [30–35]. Therefore, the analysis of UbiD family enzyme features is expected to expand the diversity of aromatic compounds, which can be biosynthesized using engineered microbes. The *E. coli* UbiD secondary structures, conserved residues, ligand binding, and active sites [29,42] are shown in Supplementary Sheet S1. N175 and E241 residues of *E. coli* UbiD are Mn^{2+} binding sites, I178 to R180 residues, R192 to L194, R197 residues, and G198 residue are prenylated flavin mononucleotide binding sites, and D290 residue is an active site.

2.4. Clustering and Integrated Gradients Analyses

The feature vectors of positive data were extracted from CNN-AE and Transformer models and were clustered by k-means algorithm. A single sequence from each cluster, whose feature vector was closest to the cluster centroid, was selected as representative sequence. Then, the representative sequences in all clusters were analyzed using IG. Moreover, these sequences were compared to the UbiD enzyme of *E. coli* (*E. coli* UbiD) registered in Swiss-Prot (sp | P0AAB4 | UBID_ECOLI) to compare the clustering method based on deep learning models to sequence similarity method. In the evaluation, the distances of feature vectors and bitscore of BLASTp were calculated for *E. coli* UbiD and each representative sequence.

Integrated gradients algorithm [24] is used to evaluate the important variables that machine learning models contribute to determining the prediction results. Therefore, in this study, the algorithm was applied to explore where region of the amino acid sequence each deep learning model grasped as important UbiD features in prediction, which are similar to the secondary structure and the important functional sites in the known annotations. The features based on CNN and Transformer models were extracted by IG analysis because these models were binary classification models that can find UbiD sequences from input proteins. On the other hand, the CNN-AE-based UbiD features were obtained from hidden layers in which UbiD features were included. Absolute values of IG were calculated for each amino acid residue in *E. coli* UbiD and the representative sequences using Tensorflow (Figure 1C), and the amino acid residues with high IG values were regarded as important features for the predictions. The IG values of output scores for input sequences were calculated in CNN model, while the IG values of feature vectors for input sequences were calculated in CNN-AE model. In Transformer model, both IG values were calculated. The IG values, multiple sequence alignments between *E. coli* UbiD and the representative sequences, and secondary structures of *E. coli* UbiD were visualized using Jalview 2.11.1.4 [43]. Multiple alignment sequences were built using MAFFT version 7 [44]. In this study, Xeon E5-2609 v4 1.7 GHz, memory 32 GB (Intel, Santa Clara, CA, USA), NVIDIA Quadro GP100 16 GB × 2 (Nvidia Corporation, Santa Clara, CA, USA) running CentOS version 7.4 was used.

3. Results

3.1. Model Training and Evaluation

The loss function curves for training and validation in CNN and CNN-AE models are shown in Supplementary Figures S4 and S5, respectively. The loss values for training

were almost the same as the loss values for validation for both models. The matching loss values suggest that the models do not tend to overfit. Test results of the CNN model were calculated for each epoch (Supplementary Table S1). The optimized CNN model was built using 4000 epochs, where 4 evaluation parameters were the highest. On the other hand, the CNN-AE model was evaluated using the Match rate. The CNN-AE model in the 2000 epochs, whose number of sequences with Match rates 0.9 and over and the average of Match rates were highest, was selected as the optimized model (Supplementary Figure S6).

The loss function curve for validation in the Transformer model decreased with matching the curve for training (Supplementary Figure S7), indicating that overfitting does not occur. Test results of the Transformer model are shown in Supplementary Table S1. The Transformer model was predicted with high accuracy in all epochs, and the test results were best in epoch 1000 according to all evaluation parameters. The model in the epochs was used in the following analysis.

3.2. Model Interpretation

UbiD feature vectors derived from CNN-AE and Transformer models were separated into 7 clusters using the k-means algorithm. From each cluster, a single UbiD sequence whose feature vector was closest to its cluster centroid was selected as the representative sequence (Supplementary Table S2). Supplementary Table S3 shows the results of feature vector distance and BLASTp bitscore between *E. coli* UbiD and each representative sequence. The higher the bitscore was, the more similar the sequence was to *E. coli* UbiD. However, feature vector distance did not seem to relate to bitscore (Supplementary Table S3).

To explore what features of UbiD sequences the deep learning models learned, these models were analyzed using IG. The IG values of *E. coli* UbiD sequence for CNN, CNN-AE, and Transformer models were calculated for all amino acid residues, as shown in Figure 2 [45,46], and the IG values of all representative UbiD sequences were calculated (Supplementary Figures S8–S12). The IG values of all representative UbiD sequences were compared to secondary structural information, conserved residues, ligand binding and active sites of *E. coli* UbiD [29,42], and the IG values of *E. coli* UbiD. Supplementary Figure S12 shows the predicted structures by ESMFold [12] and the residues with higher IG values of *E. coli* UbiD and 3 representative sequences.

The IG values of the conserved V29 residue, the conserved P357 and P423 residues in α-helix and conserved P216 residue in β-strand were higher for conserved the V29 residue, the conserved P357 and P423 residues in α-helix, and conserved P216 residue in β-strand in *E. coli* UbiD using CNN model. On the other hand, the P48, P61, and P152 residues, which are not known as the regions of secondary structures and the conserved residues, exhibited important features for the prediction. The IG values of the same residues were not necessarily high for *E. coli* UbiD and each representative sequence, although the results of more than half of the representative sequences were similar for the P48, P61, and P357 residues. Proline residues tended to be high IG values for most of the *E. coli* UbiD and representative sequences in only the CNN model.

CNN-AE model regarded more amino acid residues of *E. coli* UbiD as important factors than other models. The IG values of conserved P234 residue, the similar (semi-conserved) Q132 and R380 residues in the α-helix, and the similar I134 and L183 residues in β-strand were high in the sequence. Moreover, the M382 residue included in the α-helix and L429 and L63 residues included in the β-strand were regarded as important amino acids. CNN-AE model also identified P61, M4, and K5 residues, which were not included in secondary structures and were not conserved. More kinds of amino acids with high IG values appeared for *E. coli* UbiD and each representative sequence. The IG values of the active sites with the substrates and binding sites with prenylated flavin mononucleotide and Mn^{2+} by *E. coli* UbiD were not high in all representative sequences using CNN and CNN-AE models.

Figure 2. Multiple sequence alignment for E. coli UbiD (UBID_ECOLI) and representative UbiD sequences (Table S2) and IG results of *E. coli* UbiD derived from each deep learning model. The secondary structures of *E. coli* UbiD registered as 2IDB [45] in the Protein Data bank [46] are shown below the alignment, helix, and sheet structures are displayed as red tubes and green arrows, respectively. Bar charts show the IG values that are normalized between 0 and 1. Transformers 1 and 2 represent IG values derived from feature vectors and prediction scores, respectively.

The number of residues with high IG values for *E. coli* UbiD using the Transformer model was so smaller than those in CNN-type models. The IG results derived from prediction scores were almost the same as the results derived from feature vectors. The conserved E285 residue included in β-strand, the conserved G286 and P287 residues, and E278, Q279, and G280 residues included in β-strand exhibited high IG values. The amino acid region between 278 and 280 residues was the highest value. The high IG residues of *E. coli* UbiD were so different from those of the representative sequences in comparison to CNN-type models. Moreover, the Transformer model regarded the different consecutive amino acid residues for each representative sequence as important amino acids.

In the Transformer model, the IG values of the E285 to P287 conserved residues of *E. coli* UbiD were high, and the E285 was a putative active site in *Pseudomonas aeruginosa* [29,42]. EJZ42036.1 and WP_058074034.1 results showed the same tendency. Moreover, in AKC32612.1, the IG values of Y242 adjacent to the Mn^{2+} binding site, the 241E, were almost 1. The transformer model tended to extract UbiD features from the other residues except for annotated functional residue in the other representative sequences. In all models, the correlation coefficients between the IG values of each residue of all UbiD sequences and sequence conservation [42] were almost 0 (Supplementary Table S5).

4. Discussion

Functional annotations for protein sequences are required to understand cell functions and to search novel enzymes for target compound productions. However, annotating sequence functions is insufficient due to the increase in the number of unannotated proteins. Therefore, in this study, comprehensive features for accurately annotating enzyme sequences were explored by analyzing enzyme feature vectors derived from various deep-learning methods and IG values of each amino acid residue.

CNN, CNN-AE, and Transformer models for UbiD enzyme prediction and feature extraction were built and evaluated using multiple evaluation parameters. The validation results indicate that all current models do not occur overfitting because the loss values decreased as the training proceeds. CNN model was improved by increasing the number of training steps according to all parameter values, then the CNN model in the last 4000 steps was used. Then, the Transformer model predicted the enzymes with high accuracy, and test results showed constant prediction accuracy in all epochs. The Transformer model in 1000 steps where all parameter values were highest was selected, although the model in lower steps exhibited sufficient accuracy and seemed to be optimized. Moreover, the CNN-AE model in 2000 epochs generated output sequences with 0.9 and more Match rates, which were the almost same as input UbiD sequences, and therefore the model can learn sufficient UbiD sequence features.

To analyze the enzyme features derived from each model, the feature vectors built from the hidden layer of CNN-AE and Transformer model were clustered using the k-means algorithm, and 7 representative UbiD sequences were selected by each model. The distances between *E. coli* UbiD and the representative sequence feature vectors did not seem to relate to BLAST bitscores based on sequence identity. The results indicate that CNN-AE and Transformer models grasp different features from the conventional method. Moreover, the Transformer model enables us to deeply understand slight differences in each sequence because the distances derived from the Transformer model varied depending on the combinations of 2 sequences among UbiD sequences than the CNN-AE model (Supplementary Tables S3 and S4).

Next, what features and amino acid residues of UbiD sequences were regarded as important for each model were explored using IG. The UbiD amino acid residues regarded as important features were different for each model. The important residues consisted of not only the secondary structures, the conserved regions, the ligand binding and the active sites but also the other regions. CNN-based models did not extract the features from active and cofactor binding sites. The results suggest that the models learn the new important enzyme feature information, which is not included in the protein database [1,2,47] and

previously annotated information. Moreover, the CNN model identified the same kind of amino acid as important according to the results of the most of UbiD sequences, while the CNN-AE model showed the high IG values of the extensive residues for UbiD sequences. This is because the autoencoder model learns to ensure that the outputs match the input enzyme sequences.

The results using the Transformer model were surprisingly quite different from those of CNN-type models. The residues, which UbiD features were extracted from, were different depending on the sequences, and the residues were not necessarily important functional sites. The number of the important residues for each UbiD sequence was much smaller, although the most of residues were included in secondary structure and conserved amino acids. Therefore, the Transformer model focuses on more specific residues than CNN-type models and can extract more different enzyme features from the existing annotations for each sequence. Moreover, the Transformer model results that the different amino acid regions for each sequence showed high IG values are consistent with the results that the variance of distances between the 2 UbiD sequences was larger in comparison to CNN-type models. According to the results of the correlations between IG values and UbiD conservation, all deep learning models also extract the features that are different from the important conserved residues. In the future, the optimizations of the training datasets, especially for negative data and model structures of each model are also required for building more accurate models and extracting more higher quality features from protein sequences, such as ablation study [48,49]. Moreover, to apply the analysis to determine annotations for protein sequences, more extensive tests using the other protein families are needed because the amino acids with important features are determined depending on the models and sequences.

5. Conclusions

In this study, deep learning models were built using specific enzyme sequences included in one of the protein families, and the feature vectors derived from the models were analyzed using IG. As a result, the models regarded not only the amino acid residues included in not only the secondary structures, the conserved regions, the ligand binding and the active sites but also the other regions as important features. Therefore, the analysis can grasp multiple enzyme features that are different from previously reported information. Moreover, these models extracted different features from the sequences for each model and recognized each sequence with different features, even for similar sequences. These results show that building and evaluating models using multiple deep learning methods are more important to extract various protein features, which will be the basis of new knowledge because the recognitions of protein features are more different among each method. This method will help to interpret protein sequences through different perspectives from existing knowledge and to discover new features and motifs for unannotated protein sequences.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/biology12060795/s1, Figure S1: CNN model architecture for predicting target enzymes; Figure S2: CNN-AE model architecture for extracting target enzyme sequence features; Figure S3: Architecture of transformer model for predicting target enzymes and extracting target enzyme sequence features; Figure S4: Training for positive samples (blue line), training for negative samples and validation loss curves of CNN model for 4000 steps; Figure S5: Training and validation loss curves of CNN-AE model for 2000 epochs; Figure S6: Histogram of match rates between output sequences and input sequences using CNN-AE model in 2000 epochs; Figure S7: Training for positive samples, training for negative samples and validation loss curves of Transformer model for 1000 steps; Figure S8: Multiple sequence alignment and IG results of representative UbiD sequences derived from classification scores using CNN model; Figure S9: Multiple sequence alignment and IG results of representative UbiD sequences derived from feature vectors using CNN-AE model; Figure S10: Multiple sequence alignment and IG results of representative UbiD sequences derived from feature vectors using Transformer model; Figure S11: Multiple sequence alignment and IG results of representative UbiD sequences derived from classification scores using Transformer

model; Figure S12: UbiD structures using ESMFold structure prediction and IG results; Table S1: Test evaluations for CNN and Transformer models; Table S2: Representative sequences selected from each cluster derived from clustering by feature vectors of CNN-AE and Transformer models; Table S3: Euclidean distances of feature vectors and bitscores using BLASTp; Table S4: All Euclidean distances of feature vectors between 2 sequences for representative sequences derived from CNN-AE and Transformer models; Table S5: The results of the correlations between IG scores for each model and sequence conservation of *E. coli* UbiD; Sheet S1: *E. coli* UbiD secondary structures, conserved residues, ligand binding, active sites, and IG results for all UbiD sequences using each model.

Author Contributions: Conceptualization, N.W., Y.K., M.M., M.Y. and M.A.; methodology, M.M. and M.Y.; software, M.M. and M.Y.; validation, N.W., Y.K., M.M. and M.Y.; formal analysis, M.M., M.Y.; investigation, N.W., Y.K., M.M. and M.Y.; resources, M.M., M.Y. and M.S.; data curation, M.M., M.Y. and M.S.; writing—original draft preparation, N.W. and Y.K.; writing—review and editing, N.W., Y.K., M.M., M.Y. and M.A.; visualization, N.W., Y.K. and M.Y.; supervision, M.A.; project administration, M.A.; funding acquisition, M.A. All authors have read and agreed to the published version of the manuscript.

Funding: This article was based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO), project code P20011, Japan Science and Technology Agency: COI-NEX, grant number JPMJPF2018 and Japan Society for the Promotion of Science (JSPS), Grant-in-Aid for Scientific Research (B), grant number JP22501554.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data and source codes in this study are freely available at https://drive.google.com/drive/folders/1c_O5FDqXylDLx55e9duj-GbBPedPhWmA (accessed on 28 May 2023). The Python3 source codes can be utilized to build and evaluate 3 deep learning models, analyze models and extract protein features using IG.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Agarwala, R.; Barrett, T.; Beck, J.; Benson, D.A.; Bollin, C.; Bolton, E.; Bourexis, D.; Brister, J.R.; Bryant, S.H.; Canese, K.; et al. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2018**, *46*, D8–D13. [CrossRef]
2. Bateman, A.; Martin, M.J.; Orchard, S.; Magrane, M.; Agivetova, R.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E.H.; Britto, R.; Bursteinas, B.; et al. UniProt: The Universal Protein Knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480–D489. [CrossRef]
3. Goodwin, S.; McPherson, J.D.; McCombie, W.R. Coming of Age: Ten Years of next-Generation Sequencing Technologies. *Nat. Rev. Genet.* **2016**, *17*, 333–351. [CrossRef]
4. Sengupta, S.; Basak, S.; Saikia, P.; Paul, S.; Tsalavoutis, V.; Atiah, F.; Ravi, V.; Peters, A. A Review of Deep Learning with Special Emphasis on Architectures, Applications and Recent Trends. *Knowl. Based Syst.* **2020**, *194*, 105596. [CrossRef]
5. Kulmanov, M.; Hoehndorf, R.; Cowen, L. DeepGOPlus: Improved Protein Function Prediction from Sequence. *Bioinformatics* **2020**, *36*, 422–429. [CrossRef]
6. Strodthoff, N.; Wagner, P.; Wenzel, M.; Samek, W. UDSMProt: Universal deep sequence models for protein classification. *Bioinformatics* **2020**, *36*, 2401–2409. [CrossRef] [PubMed]
7. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef] [PubMed]
8. Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G.R.; Wang, J.; Cong, Q.; Kinch, L.N.; Dustin Schaeffer, R.; et al. Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network. *Science* **2021**, *373*, 871–876. [CrossRef] [PubMed]
9. Xu, J.; McPartlon, M.; Li, J. Improved protein structure prediction by deep learning irrespective of co-evolution information. *Nat. Mach. Intell.* **2021**, *3*, 601–609. [CrossRef]
10. Jing, X.; Xu, J. Fast and effective protein model refinement using deep graph neural networks. *Nat. Comput. Sci.* **2021**, *1*, 462–469. [CrossRef]
11. Greener, J.G.; Kandathil, S.M.; Jones, D.T. Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. *Nat. Commun.* **2019**, *10*, 3977. [CrossRef]
12. Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **2023**, *379*, 1123–1130. [CrossRef]
13. Zhou, X.; Hu, J.; Zhang, C.; Zhang, G.; Zhang, Y. Assembling multidomain protein structures through analogous global structural alignments. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 15930–15938. [CrossRef] [PubMed]

14. Zheng, W.; Wuyun, Q.; Zhou, X.; Li, Y.; Freddolino, P.L.; Zhang, Y. LOMETS3: Integrating deep learning and profile alignment for advanced protein template recognition and function annotation. *Nucleic Acids Res.* **2022**, *50*, W454–W464. [CrossRef]
15. Almagro Armenteros, J.J.; Sønderby, C.K.; Sønderby, S.K.; Nielsen, H.; Winther, O. DeepLoc: Prediction of Protein Subcellular Localization Using Deep Learning. *Bioinformatics* **2017**, *33*, 3387–3395. [CrossRef] [PubMed]
16. Wang, F.; Wei, L. Multi-scale deep learning for the imbalanced multi-label protein subcellular localization prediction based on immunohistochemistry images. *Bioinformatics* **2022**, *38*, 2602–2611. [CrossRef] [PubMed]
17. Ryu, J.Y.; Kim, H.U.; Lee, S.Y. Deep Learning Enables High-Quality and High-Throughput Prediction of Enzyme Commission Numbers. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 13996–14001. [CrossRef]
18. Li, Y.; Wang, S.; Umarov, R.; Xie, B.; Fan, M.; Li, L.; Gao, X. DEEPre: Sequence-Based Enzyme EC Number Prediction by Deep Learning. *Bioinformatics* **2018**, *34*, 760–769. [CrossRef]
19. Nallapareddy, M.V.; Dwivedula, R. ABLE: Attention Based Learning for Enzyme Classification. *Comput. Biol. Chem.* **2021**, *94*, 1–10. [CrossRef]
20. Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C.A.; Bekas, C.; Lee, A.A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019**, *5*, 1572–1583. [CrossRef]
21. Ucak, U.V.; Ashyrmamatov, I.; Ko, J.; Lee, J. Retrosynthetic reaction pathway prediction through neural machine translation of atomic environments. *Nat. Commun.* **2022**, *13*, 1186. [CrossRef]
22. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
23. Jiménez-Luna, J.; Grisoni, F.; Schneider, G. Drug Discovery with Explainable Artificial Intelligence. *Nat. Mach. Intell.* **2020**, *2*, 573–584. [CrossRef]
24. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017.
25. Lundberg, S.M.; Lee, S. A Unified Approach to Interpreting Model Predictions. In Proceedings of the Advances in Neural Information Processing Systems 30, Long Beach, CA, USA, 4–9 December 2017.
26. Jha, A.; Aicher, J.K.; Gazzara, M.R.; Singh, D.; Barash, Y.; Barash, Y. Enhanced Integrated Gradients: Improving Interpretability of Deep Learning Models Using Splicing Codes as a Case Study. *Genome Biol.* **2020**, *21*, 149. [CrossRef] [PubMed]
27. Lin, Y.; Pan, X.; Shen, H. Bin. LncLocator 2.0: A Cell-Line-Specific Subcellular Localization Predictor for Long Non-Coding RNAs with Interpretable Deep Learning. *Bioinformatics* **2021**, *37*, 2308–2316. [CrossRef] [PubMed]
28. Junghare, M.; Spiteller, D.; Schink, B. Anaerobic Degradation of Xenobiotic Isophthalate by the Fermenting Bacterium Syntrophorhabdus Aromaticivorans. *ISME J.* **2019**, *13*, 1252–1268. [CrossRef]
29. Marshall, S.A.; Fisher, K.; Cheallaigh, A.N.; White, M.D.; Payne, K.A.P.; Parker, D.A.; Rigby, S.E.J.; Leys, D. Oxidative maturation and structural characterization of prenylated FMN binding by UbiD, a decarboxylase involved in bacterial ubiquinone biosynthesis. *J. Biol. Chem.* **2017**, *292*, 4623–4637. [CrossRef]
30. Weber, C.; Brückner, C.; Weinreb, S.; Lehr, C.; Essl, C.; Boles, E. Biosynthesis of cis,cis-muconic acid and its aromatic precursors, catechol and protocatechuic acid, from renewable feedstocks by saccharomyces cerevisiae. *Appl. Environ. Microbiol.* **2012**, *78*, 8421–8430. [CrossRef]
31. Yoshida, T.; Inami, Y.; Matsui, T.; Nagasawa, T. Regioselective Carboxylation of Catechol by 3,4-Dihydroxybenzoate Decarboxylase of Enterobacter Cloacae, P. *Biotechnol. Lett.* **2010**, *32*, 701–705. [CrossRef]
32. Álvarez-Rodríguez, M.L.; Belloch, C.; Villa, M.; Uruburu, F.; Larriba, G.; Coque, J.J.R. Degradation of Vanillic Acid and Production of Guaiacol by Microorganisms Isolated from Cork Samples. *FEMS Microbiol. Lett.* **2003**, *220*, 49–55. [CrossRef]
33. Dhar, A.; Lee, K.S.; Dhar, K.; Rosazza, J.P.N. *Nocardia* Sp. Vanillic Acid Decarboxylase. *Enzym. Microb. Technol.* **2007**, *41*, 271–277. [CrossRef]
34. He, Z.; Wiegel, J. Purification and characterization of an oxygen-sensitive, reversible 3,4-dihydroxybenzoate decarboxylase from Clostridium hydroxybenzoicum. *J. Bacteriol.* **1996**, *178*, 3539–3543. [CrossRef] [PubMed]
35. Matsui, T.; Yoshida, T.; Hayashi, T.; Nagasawa, T. Purification, characterization, and gene cloning of 4-hydroxybenzoate decarboxylase of Enterobacter cloacae P240. *Arch. Microbiol.* **2006**, *186*, 21–29. [CrossRef] [PubMed]
36. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems 25, Lake Tahoe, NV, USA, 3–8 December 2012.
37. Baldi, P. Autoencoders, Unsupervised Learning, and Deep Architectures. In Proceedings of the ICML Workshop on Unsupervised and Transfer Learning, Bellevue, WA, USA, 2 July 2011.
38. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30, Long Beach, CA, USA, 4–9 December 2017.
39. Li, W.; Godzik, A. Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences. *Bioinformatics* **2006**, *22*, 1658–1659. [CrossRef] [PubMed]
40. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and Applications. *BMC Bioinf.* **2009**, *10*, 421. [CrossRef] [PubMed]
41. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv* **2016**, arXiv:1603.04467. [CrossRef]

42. Jacewicz, A.; Izumi, A.; Brunner, K.; Schnell, R.; Schneider, G. Structural Insights into the UbiD Protein Family from the Crystal Structure of PA0254 from Pseudomonas Aeruginosa. *PLoS ONE* **2013**, *8*, e63161. [CrossRef]
43. Waterhouse, A.M.; Procter, J.B.; Martin, D.M.A.; Clamp, M.; Barton, G.J. Jalview Version 2-A Multiple Sequence Alignment Editor and Analysis Workbench. *Bioinformatics* **2009**, *25*, 1189–1191. [CrossRef]
44. Katoh, K.; Rozewicki, J.; Yamada, K.D. MAFFT Online Service: Multiple Sequence Alignment, Interactive Sequence Choice and Visualization. *Brief. Bioinform.* **2019**, *20*, 1160–1166. [CrossRef]
45. Zhou, W.; Forouhar, F.; Seetharaman, J.; Fang, Y.; Xiao, R.; Cunningham, K.; Ma, L.-C.; Chen, C.X.; Acton, T.B.; Montelione, G.T.; et al. Crystal Structure of 3-octaprenyl-4-hydroxybenzoate decarboxylase (UbiD) from Escherichia coli, Northeast Structural Genomics Target ER459. 2006. Available online: https://www.wwpdb.org/pdb?id=pdb_00002idb (accessed on 28 May 2023).
46. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [CrossRef]
47. Blum, M.; Chang, H.Y.; Chuguransky, S.; Grego, T.; Kandasaamy, S.; Mitchell, A.; Nuka, G.; Paysan-Lafosse, T.; Qureshi, M.; Raj, S.; et al. The InterPro Protein Families and Domains Database: 20 Years On. *Nucleic Acids Res.* **2021**, *49*, D344–D354. [CrossRef] [PubMed]
48. Zheng, K.; Zhang, X.L.; Wang, L.; You, Z.H.; Ji, B.Y.; Liang, X.; Li, Z.-W. SPRDA: A link prediction approach based on the structural perturbation to infer disease-associated Piwi-interacting RNAs. *Brief Bioinform.* **2023**, *24*, bbac498. [CrossRef] [PubMed]
49. Zhang, H.Y.; Wang, L.; You, Z.H.; Hu, L.; Zhao, B.W.; Li, Z.W.; Li, Y.-M. iGRLCDA: Identifying circRNA–disease association based on graph representation learning. *Brief Bioinform.* **2022**, *23*, bbac083. [CrossRef] [PubMed]