biomolecules

MDPI

Article

Prediction of Protein Function from Tertiary Structure of the Active Site in Heme Proteins by Convolutional Neural Network

Hiroko X. Kondo [1,2,3,*], Hiroyuki Iizuka [4], Gen Masumoto [5], Yuichi Kabaya [1], Yusuke Kanematsu [2,6] and Yu Takano [2,*]

1 Faculty of Engineering, Kitami Institute of Technology, 165 Koen-cho, Kitami 090-8507, Japan
2 Graduate School of Information Sciences, Hiroshima City University, 3-4-1 Ozukahigashi Asaminamiku, Hiroshima 731-3194, Japan
3 Laboratory for Computational Molecular Design, RIKEN Center for Biosystems Dynamics Research, 6-2-3 Furuedai, Suita 565-0874, Japan
4 Graduate School of Information Science and Technology, Hokkaido University, Kita 14, Nishi 9, Kitaku, Sapporo 060-0814, Japan
5 Information Systems Division, RIKEN Information R&D and Strategy Headquarters, 2-1 Hirosawa, Wako 351-0198, Japan
6 Graduate School of Advanced Science and Engineering, Hiroshima University, 1-4-1 Kagamiyama, Higashi-Hiroshima 739-8527, Japan
* Correspondence: h_kondo@mail.kitami-it.ac.jp (H.X.K.); ytakano@hiroshima-cu.ac.jp (Y.T.); Tel.: +81-157-26-9401 (H.X.K.); +81-82-830-1825 (Y.T.)

Abstract: Structure–function relationships in proteins have been one of the crucial scientific topics in recent research. Heme proteins have diverse and pivotal biological functions. Therefore, clarifying their structure–function correlation is significant to understand their functional mechanism and is informative for various fields of science. In this study, we constructed convolutional neural network models for predicting protein functions from the tertiary structures of heme-binding sites (active sites) of heme proteins to examine the structure–function correlation. As a result, we succeeded in the classification of oxygen-binding protein (OB), oxidoreductase (OR), proteins with both functions (OB–OR), and electron transport protein (ET) with high accuracy. Although the misclassification rate for OR and ET was high, the rates between OB and ET and between OB and OR were almost zero, indicating that the prediction model works well between protein groups with quite different functions. However, predicting the function of proteins modified with amino acid mutation(s) remains a challenge. Our findings indicate a structure–function correlation in the active site of heme proteins. This study is expected to be applied to the prediction of more detailed protein functions such as catalytic reactions.

Keywords: structure–function correlation; active site conformation; convolutional neural network; machine learning

Citation: Kondo, H.X.; Iizuka, H.; Masumoto, G.; Kabaya, Y.; Kanematsu, Y.; Takano, Y. Prediction of Protein Function from Tertiary Structure of the Active Site in Heme Proteins by Convolutional Neural Network. *Biomolecules* **2023**, *13*, 137. https://doi.org/10.3390/biom13010137

Academic Editors: C. Martin Lawrence and Steven R. Van Doren

Received: 27 December 2022
Revised: 27 December 2022
Accepted: 7 January 2023
Published: 9 January 2023

1. Introduction

Proteins with metal cofactors and ions are called metal proteins, where a metal ion and its environment work as a catalytic active center. Because metal proteins enable biochemical reactions not possible with ordinary proteins, many researchers pay attention to them [1–5]. Heme proteins are the largest class of metal proteins and serve pivotal biological functions. Heme, a Fe–porphyrin complex, is an active center of heme proteins and expresses diverse functions such as an electron transport [6,7], a catalyst for various kinds of reactions [8,9], and an oxygen carrier [10,11]. Besides being an active center, it plays a role in the regulation of protein functions as a ligand [12,13] and in a source of Fe ions [14]. Some proteins bind to heme for transport or storage; these are referred to as hemophores [15]. The mechanism of heme protein functions has been a crucial scientific issue. The structural information

on heme proteins is increasing yearly [16], indicating a high level of scientific interest. However, few studies have comprehensively investigated heme proteins.

The key factors regulating the heme function are considered to be the axial ligand of heme, the side-chain orientation of the heme propionate, the types of heme, and the porphyrin distortion of heme. Because distal and proximal amino acids and chemical structures of heme are important factors in determining protein functions, their roles have been investigated [17–19]. However, these factors alone do not determine protein functions [20].

Both experimental and computational studies have shown a correlation between heme distortion and its chemical properties [21–24]. Our group discovered a heme distortion classified into oxygen-binding proteins and oxidoreductases by a combined analysis of machine learning and quantum chemical calculations [25]. Therefore, we focused on the contribution of heme distortion to the functional regulation of heme proteins. Heme complexed with its host protein exhibits a distorted conformation from its isolated structure [20], suggesting the regulation of the heme porphyrin structure by the protein environment around heme. From a simulation study for two oxygen carrier proteins, hemoglobin and myoglobin, it was suggested that the host protein environment affects heme distortion and controls chemical properties of heme relevant to the function of its host protein [26]. As a first step in clarifying such regulation of function of heme, we elucidated the correlations between the heme distortion and protein environment around heme, including proximal and distal amino acids using a machine learning method [27] and a convolutional neural network (CNN) [28]. Since the heme distortion correlates with its chemical properties, it is likely that it also correlates with protein function. Considering these results, we can expect to predict protein functions from the tertiary structures of heme-binding sites, including axial ligand(s).

In experimental studies, researchers are actively working on modifying the function of proteins by introducing amino acid mutations. Especially for myoglobin, which is an oxygen carrier, engineered proteins, such as peroxidase [29–33], exhibit enzymatic activity. These mutated sites are primarily located in heme-binding pockets (active sites) other than axial ligands. Thus, changes in the protein environment of a heme-binding site significantly affect protein functions.

In this study, we constructed a CNN model for predicting protein functions from the tertiary structures of the heme-binding sites of heme proteins, including proximal and distal amino acids. The CNN is a kind of deep neural network that is widely utilized in computer vision tasks, such as image classification [34,35]; it has also been applied to the classification of protein cavity structures [36]. We succeeded in predicting protein functions from the pocket structures of three functional groups of heme proteins. The prediction with our CNN model worked well between the groups with quite different functions. Analysis of the similarity of cavity shape among proteins with the same function suggests that there is no one-to-one correspondence between a protein function and a pocket structure. This study is expected to be applied to the prediction of more detailed protein functions such as catalytic reactions. This is the first step toward understanding structure–function relationships in the active sites of heme proteins.

2. Materials and Methods

2.1. Data Collection of Heme and Its Host Proteins

To collate the structural and functional information of heme proteins, we searched PDB entries containing the compound IDs (_chem_comp.id) of HEM, HEA, HEB, HEC, or HEO with a resolution of 2.0 Å or less using SQL in the PDBj Mine relational database [37] (https://pdbj.org/rdb/search, accessed on 6 October 2022). The PDBx/mmCIF files were downloaded from the Protein Data Bank Japan (PDBj) [38]. Structural information was extracted from the atom_site category of the PDBx/mmCIF file. We collected only one model for each PDB entry. When the occupancy value is <1.0 and pdbx_PDB_model_num is 1, the atom with the largest occupancy was selected from the atoms with the same

auth_seq_id and label_asym_id in the atom_site category. When the occupancy was 0.5, we chose the atoms with the label_alt_id of A. This selection was applied even to atoms with different auth_seq_id values in the atom_site category. After collecting the atomic coordinates, we excluded heme molecules missing one or more of the 25 heavy atoms forming the Fe–porphyrin skeleton (Figure 1). Consequently, 6866 heme molecules from 3206 unique PDB entries were obtained. The Bio.PDB package [39] for BioPython version 1.78 [40] was used to parse the mmCIF files.

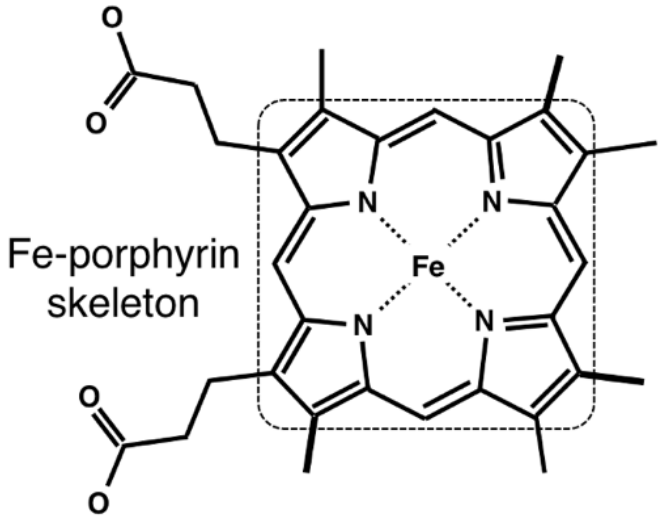

Figure 1. Chemical structure of heme. Fe–porphyrin skeleton is enclosed by a square-dotted line.

As a first step in elucidating the correlation between the tertiary structure of an active site and protein function, we used only the structures in which amino acids or water molecules were axially coordinated to heme. Here, 5185 samples were obtained. Axial ligands were defined as amino acid residues or other molecules, including one or more atoms within 3.1 Å of the heme iron atom. MDTraj library version 1.9.5 [41] was used to analyze the structural data. To reduce the redundancy of amino acid sequences, we excluded protein chains with sequence similarity higher than 99.99% using the PISCES server [42]. Finally, the samples in which the coverage of heme was less than 0.6 were excluded because the biological and asymmetric units were likely to differ. This nonredundant dataset was composed of 1234 samples and is referred as dataset_99. Although the oxidation state of Fe is closely related to the protein function, we did not consider it in this study because the aim of this study was an elucidation of the structure–function relationship in hemeproteins and a construction of functional predictor from the pocket structure for this purpose.

2.2. Assignment of Protein Function to Each Heme Sample

Information about protein function was assigned by the enzyme commission (EC) number and gene ontology (GO) associated with each entity in each PDB entry, as well as keywords and descriptions stored in each PDB entry. The EC number, GO, keywords, and description were collected by a SQL search from EC_number of sifts.pdb_chain_enzyme table, GOID of gene_ontology_pdbmlplus table, keywords of brief_summary table, and pdbx_description of entity table in PDBj Mine relational database (accessed on 6 October 2022), respectively. First, we assigned function(s) to each sample of the non-redundant dataset as follows:

(1) If the protein chain(s), including the axial ligand(s), had an EC number(s), the first digit of the EC number(s) was assigned.
(2) If case 1 did not apply and the protein chain(s) had GO associated with "oxygen-binding", "oxidoreductase activity", "electron transfer activity", "transcription", or

"heme transport" as the molecular function or biological process, one function was assigned in order from these functions.

(3) If cases 1 and 2 did not apply and the PDB entry had keywords associated with "hemophore", "electron transfer activity", "oxygen-binding", "oxidoreductase activity", "heme extraction", "signaling protein", "nitrophorin (NO transport)", or "heme transport", one function was assigned in order from these functions.

(4) If cases 1–3 did not apply and the PDB entry had a description of cytochrome p460, "oxidoreductase" was assigned.

(5) If cases 1–4 did not apply or there was no axial ligand, "unclassified" was assigned.

At this stage, 16 types of function labels, including multi-function combinations, were assigned. Next, we manually modified the function of dehaloperoxidase and myoglobin with oxidoreductase activity to "oxygen-binding and oxidoreductase" (dual-function). In this study, we only used the samples assigned "oxygen-binding", "oxidoreductase", "electron transfer", or "oxygen-binding and oxidoreductase" as protein functions. These protein functions are listed in SI (pdbid_function_list.csv).

2.3. CNN Model

Here, we constructed a CNN model whose input and output were the tertiary structure of the heme-binding pocket and the protein function, respectively. To use the non-uniform structural data of heme-binding site as input for the CNN model, we converted the data into uniform dimensional data. Then, we used voxel sets included in a cube-shaped inclusion region on the heme-binding site as an input (Figure 2). This inclusion region was defined as described below. First, we calculated a least-squares plane for CHA, CHB, CHC, and CHD atoms in the porphyrin ring of heme and defined it as the xy-plane. Then, we rotated the xy-plane such that the x-axis was parallel to the vector connecting CHA and CHC projected onto the least-squares plane and determined the z-axis to be perpendicular to the xy-plane and right-handed. Finally, the origin was translated to the barycenter of CHA, CHB, CHC, and CHD. The edge length of the inclusion region was set to 24 Å, which is identical to the value determined in our previous study [28]. For voxelization, we divided the space included in the inclusion region into the small cubic region (voxel) with an edge length of 1 Å. Using atomic coordinates of protein without heme and molecules other than proteins, we assigned 1 (occupied) or 0 (unoccupied) to each voxel depending on whether it was occupied by any atom or not, respectively. The input voxels were prepared for each atom of C, N, O, and S, and used as an input with four channels. For the detailed procedures for determining the inclusion region and voxelization, please refer to our previous study [28].

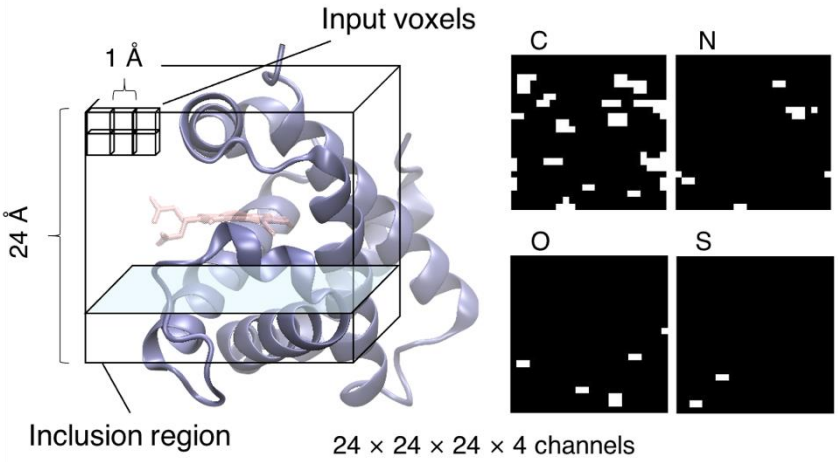

Figure 2. Schematic of the input of our CNN model. The protein backbone and heme molecule are represented as a blue cartoon and a licorice model colored in salmon, respectively. The input voxels were prepared for each atom of C, N, O, and S, as illustrated in the right panel. The coordinates of heme and molecules other than protein were not used in the voxel calculation.

The output of the CNN model is a class label of the protein function. Class labels are two- or three-dimensional, allowing multiple functions to be assigned to a single sample. The loss was calculated as binary cross-entropy between the observed (assigned function) and predicted class labels.

We constructed and trained all CNN models using PyTorch version 1.11.0 [43]. The parameters of our CNN model are shown in Table 1. These parameters are identical to those determined in our previous study [28] except the last layer. A brief demonstration regarding the method used in CNN is also described there. The network model was constructed so that the number of layers would not be too large, and the other hyperparameters were roughly tuned. The output dimension of each layer was determined by the number of output channels specified in Convolution layer, and the parameters such as the kernel size and/or stride of the Convolution and Pooling layers. These hyperparameters were set to those commonly used. We tried a couple of models with different hyperparameters for this study, which resulted in almost no effect on accuracy. For training, the stochastic gradient descent optimizer with a learning rate of 0.01 was used, and the batch size was set to 32. To verify the generalization performance of the model, five-fold cross-validation was performed. We did not separate the test and cross-validation datasets because of limited data. The detailed procedure of the cross-validation has been described in our previous study [28].

Table 1. Layers and parameters of our CNN model.

Layer	Function	Filter (Kernel)	Output Dimension (Channel × Depth × Width × Height)
1	Conv3d	2 × 2 × 2 with 0-padding	64 × 21 × 21 × 21
2	Conv3d	2 × 2 × 2 with 0-padding	128 × 22 × 22 × 22
3	BatchNorm3d	-	128 × 22 × 22 × 22
4	Conv3d	2 × 2 × 2 without padding	128 × 21 × 21 × 21
5	ReLU	-	128 × 21 × 21 × 21
6	BatchNorm3d	-	128 × 21 × 21 × 21
7	MaxPool3d	2 × 2 × 2 stride: 2 × 2 × 2	128 × 10 × 10 × 10
8	Full connection	-	128,000
9	Linear	-	128
10	ReLU	-	128
11	Dropout	0.4	128
12	Linear	-	64
13	BatchNorm1d	-	64
14	ReLU	-	64
15	Linear	-	2 or 3
16	Sigmoid	-	2 or 3

2.4. Analyses of Cavity of Heme-Binding Site

We computed the cavity shapes of heme-binding sites using POVME 3.0 [44]. With POVME, the cavity shape of a ligand-binding pocket can be represented as a bit vector, each element of which represents whether or not the respective grid is located in a ligand-binding cavity, 1 for a cavity and 0 for protein atoms. We refer to this bit vector as a "cavity vector" in the following. To compare the cavity shapes of various proteins, the region to be analyzed was limited to the vicinity of the heme molecule: the center and radius of the inclusion sphere (parameters for POVME) were set to the coordinates of the heme iron atom and 8.5 Å, respectively. We set the grid size to 1 Å and did not use the option for removing isolated points that were not contiguous with the specified region. The detailed

procedure for preparing the input protein coordinates has been described in our previous work [28].

3. Results and Discussion

3.1. Prediction of Protein Function from the Tertiary Structure of the Heme-Binding Pocket Using a CNN Model: Two-Label Classification

We constructed a CNN model to predict the function of proteins classified into the following three classes, namely, oxygen-binding protein (OB), oxidoreductase (OR), and proteins with both functions (OB–OR), from the tertiary structures of heme-binding pockets by using the dataset_99. The output of the CNN model is two-dimensional, with each label indicating whether each function (oxygen-binding or oxidoreductase) is retained, namely, (0, 1), (1, 0), and (1, 1) represent the OB, OR, and OB–OR classes, respectively. Only when the values of the two labels matched between the observed and predicted ones were the results considered true positives (TP). The obtained models were evaluated in terms of the score, S_{acc}, calculated as follows:

$$S_{\mathrm{acc}} = \frac{\sum_{c \in L} N_c^{\mathrm{TP}}}{\sum_{c \in L} N_c} \quad (1)$$

where L, N_c, and N_c^{TP} represent the labels of function, the number of samples belonging to class c, and the number of samples in class c that are TP as a result of prediction, respectively. In this analysis, L = {OB, OR, OB–OR}. N_{OB}, N_{OR}, and $N_{\mathrm{OB-OR}}$ for the test sets of five-fold cross-validation runs were 190, 312, and 35, respectively. The mean and standard deviation of the S_{acc} scores obtained from five-fold cross-validation was 0.959 ± 0.021, indicating high prediction accuracy.

We also calculated the confusion matrix **M** using the scikit-learn Python library [45] version 0.24.2 (Table 2). The non-diagonal element of a confusion matrix, $\mathbf{M}_{ij}$, represents the actual number of observations in class i but are predicted to be in class j. The confusion matrix of Table 2 was normalized, and each element has a mean value over five cross-validation runs. Although in two-label classification, the predicted value can also be (0, 0), which means that the sample is neither OB nor OR, there was no sample with a predicted value of (0, 0) in this analysis. Therefore, such a sample is omitted in Table 2. The protein function could be predicted with very high accuracy for the single-function proteins (OB and OR). However, protein function prediction was difficult for the dual-function proteins (OB–OR). We also calculated the mean values of accuracy, recall, precision, and specificity over the five-fold cross-validation runs for each class (Table S1). For the calculation of these indicators, we defined $\mathbf{M}_{ii}$ as TP, M_{ji} (j $\neq$ i) as false positive, $\mathbf{M}_{ij}$ (j $\neq$ i) as false negative, and $\mathbf{M}_{jk}$ ($j \neq i$, $k \neq i$) as true negative for class i. Whereas all indicators were high in the single-function proteins, only precision was high in the OB–OR. The latter means that samples that were predicted to be OB–OR were correct, but there were many samples belonging OB–OR that could not be correctly predicted. The dual-function proteins contain two types of proteins: dehaloperoxidase and myoglobin mutants. The ratios of TP in the samples included in the test sets of five cross-validation runs were 1.0 (9/9) for dehaloperoxidase and 0.423 (11/26) for myoglobin mutants with a dual-function (DF-Myoglobin). The low TP rate in the OB–OR class was due to the inaccuracy of the prediction of the function of DF-Myoglobins. Considering that the dataset_99 includes 116 samples with the description of "myoglobin" in PDB, 32 of which have dual functions, it is likely that the prediction was influenced by samples with similar pocket structures but a different function.

Table 2. Mean values and standard deviations of the normalized confusion matrices over five-fold cross-validation runs. Values in the parentheses represent the confusion matrix calculated with the combined data of the test sets of five-fold cross validation runs for two-label classification.

		Predicted Value		
		OB	**OR**	**OB–OR**
Observed Value	**OB [190]** †	0.985 ± 0.012 (187)	0.010 ± 0.012 (2)	0.005 ± 0.010 (1)
	OR [312] †	0.010 ± 0.008 (3)	0.990 ± 0.008 (309)	0.000 ± 0.000 (0)
	OB–OR [35] †	0.436 ± 0.248 (15)	0.000 ± 0.000 (0)	0.564 ± 0.248 (20)

† Values in the square brackets represent the sample numbers of each class.

Next, we examined in detail the samples with inaccurate function prediction. Fifteen of the twenty-one samples with inaccurate predictions were DF-Myoglobins, most of which were predicted to belong to the OB class. The samples other than DF-Myoglobins classified as OB are listed in Table 3. PDB ID of 3QZX [46] is protoglobin, which has highly distorted heme, suggesting that the pocket structure is different from those of other oxygen-binding proteins. For PDB IDs of 3QZX, 4XDI [47], and 6O0A [48], there was no sample with a similar amino acid sequence (similarity ≥ 0.7). The lack of sufficient training data may be the cause of prediction failure. For two cases (PDB ID of 2BK9 [49] and 3MVC [50]), the protein function assignment may be wrong, and the predicted results were correct (misassignment of protein function). Although the former is hexacoordinate hemoglobin, which is expected to function as oxidoreductase, it is unclear whether this protein exhibits enzymatic activity. The latter exhibits oxidoreductase activity and no affinity to the oxygen molecules, but OB was assigned as the protein function. There was a sample with OB as the class label (observed value) and OB–OR as the predicted value (PDB ID: 7CEZ). PDB ID of 7CEZ is myoglobin G5K/Q8K/A19K/V21K mutant. Its functional property is unknown because the paper is unpublished. This mutant may exhibit oxidoreductase activity, as we predicted. Considering these results, protein function assignment is one of the significant challenges in this type of research.

Table 3. List of the samples that failed to predict other than those classified as OB and DF-Myoglobins.

PDB ID	Protein Name	Observed Value	Predicted Value	Remark
2BK9	hemoglobin	OR	OB	misassignment
3MVC	GLB-6	OB	OR	misassignment
3QZX	protoglobin	OB	OR	-
4XDI	THB1 (truncated hemoglobin)	OR	OB	-
6O0A	flavohemoglobin	OR	OB	-
7CEZ	myoglobin (G5K/Q8K/A19K/V21K)	OB	OB–OR	detailed function unknown

3.2. Specification of Regions in Input Data Significant for Prediction

To determine the regions significant for predicting protein function, we examined the change in prediction scores when information about a specific region of input voxels was discarded. The model constructed in Section 3.1 was used for this analysis. Information was discarded in two ways. We refer to them as "outside discarding" and "inside discarding", which remove information from the outside (Figure 3a) and inside (center) (Figure 3b), respectively. First, two cubes were defined: the "outer cube" and the "inner cube". The vertex coordinates of the outer cube are (±12, ±12, ±12), being equivalent to the inclusion region of the CNN model. Let the vertex coordinates of the inner cube be (±(12–*r*), ±(12–*r*), ±(12–*r*)) on the "outside discarding" and be (±*r*, ±*r*, ±*r*) on "inside discarding". Then, the

sets of voxels in the outer and inner cubes are denoted as V_{outer} and V_{inner}, respectively. The voxels in V_{outer} but not in V_{inner} were replaced with 0 for "outside discarding" ($0 \leq r < 12$, Figure 3a), and those of V_{inner} were replaced with 0 for "inside discarding" ($0 \leq r < 12$, Figure 3b). In both cases, information is intact (not discarded) at $r = 0$.

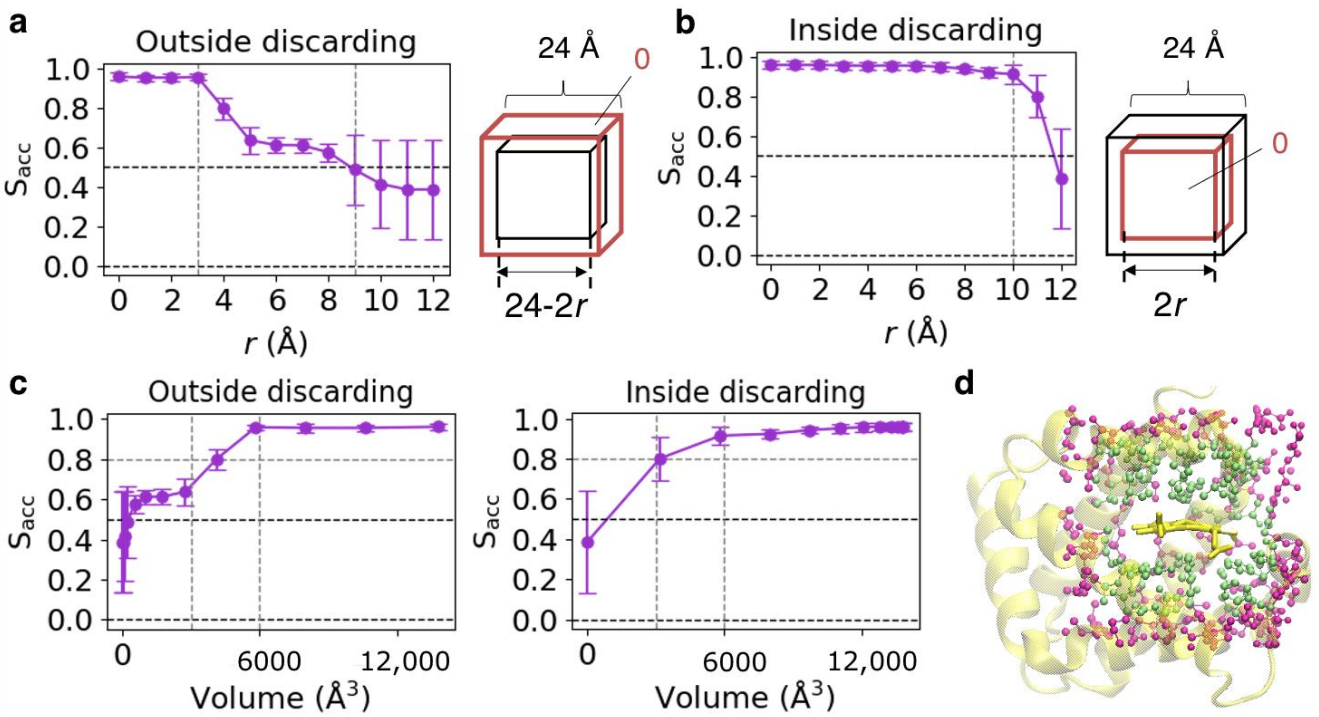

Figure 3. (**a**) Mean S_{acc} scores plotted against r, which is the distance between the faces of the outer (red) and inner (black) cubes presented in the right panel for "outside discarding." The error bar shows the standard deviation. The centers of the outer and inner cubes are identical, and their edges are parallel. (**b**) Mean S_{acc} scores plotted versus r for "inside discarding." (**c**) S_{acc} scores versus the volume of the region with the original information. (**d**) Atoms included in cube-shaped regions with an edge length of l are illustrated using the PDB entry of 1A00 as an example. The lime spheres and the combination of lime and magenta spheres represent l = 18 and 24, respectively. The main chain of the host protein is shown as a yellow cartoon, and heme as a yellow stick.

S_{acc} for "outside discarding" and "inside discarding" averaged over the test sets in the five-fold cross-validation runs are presented in the left panels of Figure 2a,b, respectively. Because the amount of information loss on the r value was different between "outside discarding" and "inside discarding" and nonlinear, S_{acc} scores were also plotted against the volume of the region with the original information (Figure 3c). Considering that the change in S_{acc} scores between the volumes of 3000 and 6000 Å^3 differed for "outside discarding" and "inside discarding," the score would depend on the region used for prediction. Whereas the scores dropped sharply when the value of r exceeded 3 Å, where the edge length of the inner cube was 18 Å and reached almost 0.5 r = 9 Å in "outside discarding", it did not significantly change between the values of r from 0 to 10 Å, where the edge length of the inner cube is 0–20 Å in "inside discarding." These results suggest that the prediction was performed using the information near the surface of the outer cube (input voxels). Examples of A_l (l = 18 and 24), which is an atom set included in the cube with edge lengths of l, are illustrated in Figure 3d using a PDB entry of 1A00. This may be one of the reasons why it was difficult to distinguish amino acid mutations in the heme-binding pocket of DF-Myoglobin. We also constructed a CNN model using smaller input voxels (edge length = 17 Å) as an input. However, almost the same result was obtained (the mean and standard deviation of S_{acc} score over five-fold cross-validation was 0.959 ± 0.024). The

confusion matrix is shown in Table S2. The modification of inputs may be required to incorporate information about the pocket surface into the prediction.

3.3. Prediction of Protein Function from the Tertiary Structure of the Heme-Binding Pocket Using a CNN Model: Three-Label Classification

We constructed a CNN model with three-dimensional output to predict the functions of proteins classified into the following four classes: OB, OR, OB–OR, and electron transport protein (ET) by using the dataset_99. Other classes were not assigned in this study. The output is three-dimensional, with each label indicating whether or not each function (oxygen-binding, oxidoreductase, or electron transfer) is retained, namely, (0, 1, 0), (1, 0, 0), (1, 1, 0), and (0, 0, 1) represent the OB, OR, OB–OR, and ET classes, respectively. Only when the values of the three labels matched between the observed and predicted ones were the results considered TP.

The number of samples belonging to OB, OR, OB–OR, and ET for the test sets of five-fold cross-validation were 193, 297, 36, and 371, respectively. The prediction accuracy was also reasonably high in the three-label classification, and the mean and standard deviation of the S_{acc} for L = {OB, OR, OB–OR, ET} in Equation (1) obtained from the five-fold cross-validation were 0.895 ± 0.031. As shown in the confusion matrix shown in Table 4, while the recall for the OB class was as high as that in the two-label classification, that for OR became lower and was nearly the same as that for ET. This may be because of the functional similarity between OR and ET. We also calculated the mean values of accuracy, recall, precision, and specificity over the five-fold cross-validation runs for each class (Table S3). Some of the samples that were erroneously predicted as ET despite being OR had a keyword associated with "electron transfer" in PDB. Notably, the low false recognition rates between OB and ET and between OB and OR, suggest a clear difference in the tertiary structures of their active sites. This indicates the structure–function relationships in the active sites of heme proteins. We expect the application of this method to the classification of a wider variety of protein functions in the future.

Table 4. Mean values and standard deviations of the normalized confusion matrices over five cross-validation runs. Values in the parentheses represent the confusion matrix calculated with the combined data of the test sets of five-fold cross validation runs for three-label classification.

		Predicted Value				
		OB	**OR**	**OB–OR**	**ET**	**Others** †
Observed Value	**OB [193]** ‡	0.973 ± 0.016 (188)	0.016 ± 0.013 (3)	0.006 ± 0.012 (1)	0.000 ± 0.000 (0)	0.005 ± 0.010 (1)
	OR [297] ‡	0.006 ± 0.007 (2)	0.907 ± 0.054 (268)	0.000 ± 0.000 (0)	0.084 ± 0.049 (26)	0.004 ± 0.007 (1)
	OB–OR [36] ‡	0.570 ± 0.296 (20)	0.000 ± 0.000 (0)	0.430 ± 0.296 (16)	0.000 ± 0.000 (0)	0.000 ± 0.000 (0)
	ET [371] ‡	0.000 ± 0.000 (0)	0.110 ± 0.019 (40)	0.000 ± 0.000 (0)	0.890 ± 0.019 (331)	0.000 ± 0.000 (0)

† "Others" represents the predicted value of (0, 0, 0). ‡ Values in the square brackets represent the sample numbers of each class.

3.4. Validation of Datasets Used for CNN Model Construction

To validate the dataset used for the CNN model construction in this study, we constructed CNN models using the additional datasets with different thresholds of the sequence similarity. Although a previous study, in which the heme-binding site was detected from the property of pocket cavity, adopted a threshold of 80% [36], a sufficient value of the threshold of sequence similarity is generally debatable [51]. Here, we used 25.00, 60.00, 80.00, and 99.99% as the threshold of sequence identity for nonredundant datasets. This is because thresholds of 25% were adopted for the prediction of secondary structure [52] and disorder region [53], and a threshold of 60% of the motif length was proposed for

the prediction of post-translational modifications [54]. Since these datasets included few samples of OB-OR, we removed the OB-OR samples from each dataset and carried out the classification of OB and OR (two-label and two-class classification). We referred to these datasets as dataset_25, dataset_60, dataset_80, and dataset_99_without_OB-OR, respectively, in the following. The mean S_{acc} scores over five-fold cross-validation runs were 0.923 ± 0.069, 0.934 ± 0.089, 0.974 ± 0.022 and 0.990 ± 0.011 for the dataset_25, dataset_60, dataset_80, and dataset_99_without_OB-OR, respectively. The mean values of accuracy, recall, precision, and specificity over five-fold cross-validation runs are listed in Table 5. Despite the bias in the sample numbers of each class, most indicators showed high values in both classes even in the dataset of dataset_25.

Table 5. Mean values and standard deviations of accuracy, precision, recall, and specificity obtained from two-label classification over the five-fold cross-validation runs for each class.

Dataset	Class Label	Accuracy	Recall	Precision	Specificity
Dataset_25	OB [9] †	0.923 ± 0.069	0.800 ± 0.400	0.653 ± 0.366	0.925 ± 0.074
	OR [55] †	0.923 ± 0.069	0.925 ± 0.074	0.983 ± 0.033	0.800 ± 0.400
Dataset_60	OB [36] †	0.934 ± 0.089	0.703 ± 0.381	0.979 ± 0.036 ‡	0.994 ± 0.013
	OR [192] †	0.934 ± 0.089	0.994 ± 0.013	0.937 ± 0.090	0.703 ± 0.381
Dataset_80	OB [59] †	0.974 ± 0.022	0.932 ± 0.097	0.896 ± 0.106	0.983 ± 0.015
	OR [239] †	0.974 ± 0.022	0.983 ± 0.015	0.987 ± 0.017	0.932 ± 0.097
Dataset_99_ without_OB-OR	OB [196] †	0.990 ± 0.011	0.995 ± 0.010	0.981 ± 0.026	0.987 ± 0.020
	OR [308] †	0.990 ± 0.011	0.987 ± 0.020	0.997 ± 0.007	0.995 ± 0.010

† Values in the square brackets represent the sample numbers of the test sets of each class. ‡ The results averaged over four runs of the five-fold cross-validation runs because both TP and FP were 0 in a run.

In addition, we performed the same analysis of Section 3.2 with the CNN model constructed by the dataset_25. As shown in Figure S1, the behaviors of both "outside discarding" and "inside discarding" are similar to those of the dataset_99, suggesting that both networks by the dataset_25 and dataset_99 may use similar features.

We also constructed a CNN model by using the dataset_25 for three-label classification, the same analysis as Section 3.3, and obtained the mean S_{acc} score of 0.767 ± 0.083. The number of samples belonging to OB, OR, and ET for the test sets of five-fold cross-validation were 15, 54, and 31, respectively. There was a sample that was erroneously classified as Others. The confusion matrix and values of accuracy, recall, precision, and specificity were listed in Tables S4 and S5. The slight decrease in the mean S_{acc} score compared with that of the dataset_99 would be mainly due to misclassification of OR. There was an increase in the number of cases where the OB was classified as OR and the OR was classified as ET. The small sample number may lead to a decrease in accuracy with an increase in class labels.

These results indicate that the presence of similar data does not unfairly increase accuracy, namely, the effect of a large value of the sequence identity is small. A similar kind of robustness to the sequence identity cutoff has been demonstrated for the performance of a structure-based graph convolution network model over the function prediction [55]. Therefore, we conclude that the sequence homology would have little impact on our problem.

3.5. Similarity of the Structures of Heme-Binding Pockets between Proteins with the Same Function

To estimate the similarity of cavity shapes of the heme-binding sites in proteins with the same function, we analyzed the variability of cavity shapes for each protein group using cavity vectors computed by POVME software. Let I be a set of samples of cavity shapes in a protein group. The mean distance from the barycenter for cavity vector $\boldsymbol{v}_i$ was

calculated for each protein group as an indicator of dispersion of a set of cavity vectors following the same procedure as our previous work [28], as follows:

$$N_I = |I|(\text{the number of samples of } I), \tag{2}$$

$$\mu_I = \frac{1}{N_I} \sum_{i \in I} v_i \tag{3}$$

$$\bar{d}_I = \frac{1}{N_I} \sum_{i \in I} \left|\left| v_i - \mu_I \right|\right| \tag{4}$$

where $||\ ||$ represents the L^2 norm.

The protein group identifier, number of samples, and $\bar{d}_I$ for each protein group calculated for the dataset_99 and dataset_25 are shown in Table 6. Results for the combined group of OB, OR, and OB–OR (referred to as "Combined" in the following), dehaloperoxidase, DF-Myoglobin, and myoglobin (OB) are also listed for comparison. For smaller $\bar{d}_I$ values, higher cavity shape similarity was expected in a protein group.

Table 6. Protein groups, sample numbers, and $\bar{d}_I$. The shaded row represents the protein group combined OB, OR, OB–OR, and ET.

Protein Group	Sample Number		$\bar{d}_I$	
	Dataset_99	Dataset_25	Dataset_99	Dataset_25
OB	241	16	12.70 (1.78) †	15.45 (2.23) †
OR	388	63	16.88 (2.45) †	15.86 (2.61) †
OB–OR	42	0	12.44 (2.41) †	-
ET	450	47	16.52 (2.88) †	16.07 (2.05) †
Combined	1121	126	16.80 (2.51) †	15.99 (2.34) †
Dehaloperoxidase	10	0	9.95 (1.21) †	-
DF-Myoglobin	32	0	11.20 (2.60) †	-
Myoglobin (OB)	82	1	9.99 (2.33) †	0.00 (0.00) †

† Values in parentheses represent the standard deviation of $||v_i - \mu_I||$.

As shown in Table 6, similar results were obtained for the dataset_99 and dataset_25. While $\bar{d}_I$ was slightly small in the OB and OB–OR classes for the result of the dataset_99, it was as high as that in the "Combined" group, including four protein groups for the OR and ET classes. For homologous protein groups, dehaloperoxidase and myoglobin, $\bar{d}_I$ was significantly smaller than that of the "Combined" group. The $\bar{d}_I$ of DF-Myoglobin was slightly larger than that of myoglobin, suggesting that the mutations in the active site change the cavity structures. This implies that the structure of an active site is not similar among proteins with the same function but varies significantly among protein groups. Considering the results of Section 3.1, the proteins with the same function have a common structural feature in spite of the difference in the overall cavity shapes.

4. Conclusions

In this study, we constructed a CNN model to predict protein functions from the tertiary structures of the active sites of heme proteins to examine the structure–function relationship. High S_{acc} scores (>0.95) were obtained by the CNN model for two-label classification for classifying OB, OR, and OB–OR. There were a few cases of false positives due to the misassignment of protein function, i.e., the predicted results were correct, resulting in the issue of improving the method of function assignment. In addition, the prediction of the function of engineered myoglobin (functionally modified mutants) remained a challenge. Because myoglobin is mostly an oxygen carrier, the difficulty in predicting the function of functionally modified mutants may be due to the lack of sufficient data. The analysis results of the similarity of cavity shape among proteins with the same function indicate that

there is no one-to-one correspondence between the protein function and pocket structure, suggesting that the proteins with the same function have a common structural feature in spite of the difference in the overall cavity shapes. Predicting the modified function of proteins with a single amino acid mutation may require some ingenuity.

We also constructed a CNN model for three-label classification to classify OB, OR, OB–OR, and ET. Although the overall accuracy was slightly lower than that of the two-label classification, the recall for OB was maintained at the same level as that for the two-label classification. The misclassification between OB and ET and between OB and OR is almost zero, indicating that the prediction works well between the groups with different functions. The application of this study to classification tasks with more labels is expected.

Overall, this study demonstrated the structure–function correlation in the active sites of heme proteins. In the future, we will attempt to construct a model to predict more detailed protein functions, such as catalytic reactions or function of proteins binding heme as a non-active center, such as hemophores. To improve the accuracy and robustness of the CNN model, we will attempt to increase the amount of structural data, improve the function assignment method, modify the input information, and so on. Since the protein dynamics are also important for protein function, we will also attempt to include them into the input to improve our CNN model in the future. Our previous study showed that AlphaFold2 [56], which is a deep learning algorithm for predicting the tertiary structure of proteins from the amino acid sequence, can accurately predict the structure of the heme-binding site in heme proteins [57]. If the challenge of predicting heme-binding sites from their amino acid sequences could be overcome, protein functions would be predicted using their amino acid sequences for heme proteins. We would like to attempt this challenge in the future.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/biom13010137/s1, Table S1: Confusion matrix resulted from the two-label classification with the edge length of inclusion region of 12.0 Å; Table S2: Confusion matrix resulted from the two-label classification with the edge length of inclusion region of 8.5 Å; Table S3: Mean values and standard deviations of accuracy, precision, recall, and specificity obtained from three-label classification; Figure S1: Plots of mean S_{acc} scores of the outside discarding and inside discarding for the CNN model by using the dataset_25; Table S4: Mean values and standard deviations of the normalized confusion matrices for three-label classification with the dataset_25; Table S5: Mean values and standard deviations of precision, recall, and specificity obtained from three-label classification by using the dataset_25.

Author Contributions: Conceptualization, H.X.K., H.I. and G.M.; Methodology, H.X.K., H.I., Y.K. (Yuichi Kabaya), G.M., Y.K. (Yusuke Kanematsu) and Y.T.; software, H.X.K.; investigation, H.X.K., H.I. and G.M.; resources, H.X.K.; data curation, H.X.K.; writing—original draft preparation, H.X.K.; writing—review and editing, H.X.K., H.I., G.M., Y.K. (Yuichi Kabaya), Y.K. (Yusuke Kanematsu) and Y.T.; visualization, H.X.K.; supervision, H.X.K., H.I. and G.M.; project administration, H.X.K., H.I. and G.M.; funding acquisition, H.X.K., H.I., Y.K. (Yusuke Kanematsu) and Y.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by FY 2021 KNIT Collaborative Research Fund from Kitami Institute of Technology and Hokkaido University, and a grant for the Basic Science Research Projects from the Sumitomo Foundation. We are grateful to the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) for a Grant-in-Aid for Scientific Research on Transformative Research Areas (A) "Hyper-Ordered Structures Science", 20H05883 and to the Japan Society for the Promotion of Science (JSPS), 19K06589, 19H02752, 22K06164, 22H04813, and 21K14741.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The atomic coordinates of heme proteins were downloaded from PDBj (https://pdbj.org/, accessed on 6 October 2022). For convenience, the list of PDB IDs and assigned protein functions is provided in Supplementary Materials.

Acknowledgments: Computations were performed at RIKEN Advanced Center for Computing and Communication (ACCC) and the Research Center for Computational Science, Okazaki, Japan. The present study was performed in part under the Collaborative Research Program of the Institute for Protein Research, Osaka University (CR-20-02, CR-21-02, and CR-22-02).

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

References

1. Solomon, E.I.; Szilagyi, R.K.; DeBeer George, S.; Basumallick, L. Electronic Structures of Metal Sites in Proteins and Models: Contributions to Function in Blue Copper Proteins. *Chem. Rev.* **2004**, *104*, 419–458. [CrossRef] [PubMed]
2. Tosha, T.; Nomura, T.; Nishida, T.; Saeki, N.; Okubayashi, K.; Yamagiwa, R.; Sugahara, M.; Nakane, T.; Yamashita, K.; Hirata, K.; et al. Capturing an Initial Intermediate during the P450nor Enzymatic Reaction Using Time-Resolved XFEL Crystallography and Caged-Substrate. *Nat. Commun.* **2017**, *8*, 1585. [CrossRef] [PubMed]
3. Nomura, T.; Kimura, T.; Kanematsu, Y.; Yamada, D.; Yamashita, K.; Hirata, K.; Ueno, G.; Murakami, H.; Hisano, T.; Yamagiwa, R.; et al. Short-Lived Intermediate in N_2O Generation by P450 NO Reductase Captured by Time-Resolved IR Spectroscopy and XFEL Crystallography. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2101481118. [CrossRef] [PubMed]
4. Saito, T.; Takano, Y. QM/MM Molecular Dynamics Simulations Revealed Catalytic Mechanism of Urease. *J. Phys. Chem. B* **2022**, *126*, 2087–2097. [CrossRef] [PubMed]
5. Shoji, M.; Murakawa, T.; Nakanishi, S.; Boero, M.; Shigeta, Y.; Hayashi, H.; Okajima, T. Molecular Mechanism of a Large Conformational Change of the Quinone Cofactor in the Semiquinone Intermediate of Bacterial Copper Amine Oxidase. *Chem. Sci.* **2022**, *13*, 10923–10938. [CrossRef]
6. Poulos, T.L. The Janus Nature of Heme. *Nat. Prod. Rep.* **2007**, *24*, 504–510. [CrossRef]
7. Louie, G.V.; Brayer, G.D. High-Resolution Refinement of Yeast Iso-1-Cytochrome c and Comparisons with Other Eukaryotic Cytochromes C. *J. Mol. Biol.* **1990**, *214*, 527–555. [CrossRef]
8. Shaik, S.; Kumar, D.; de Visser, S.P.; Altun, A.; Thiel, W. Theoretical Perspective on the Structure and Mechanism of Cytochrome P450 Enzymes. *Chem. Rev.* **2005**, *105*, 2279–2328. [CrossRef]
9. Ostermeier, C. Cytochrome c Oxidase. *Curr. Opin. Struct. Biol.* **1996**, *6*, 460–466. [CrossRef]
10. Perutz, M.F.; Rossmann, M.G.; Cullis, A.F.; Muirhead, H.; Will, G.; North, A.C.T. Structure of Hæmoglobin: A Three-Dimensional Fourier Synthesis at 5.5-Å. Resolution, Obtained by X-Ray Analysis. *Nature* **1960**, *185*, 416–422. [CrossRef]
11. Kendrew, J.C.; Dickerson, R.E.; Strandberg, B.E.; Hart, R.G.; Davies, D.R.; Phillips, D.C.; Shore, V.C. Structure of Myoglobin: A Three-Dimensional Fourier Synthesis at 2 Å. Resolution. *Nature* **1960**, *185*, 422–427. [CrossRef]
12. Faller, M.; Matsunaga, M.; Yin, S.; Loo, J.A.; Guo, F. Heme Is Involved in MicroRNA Processing. *Nat. Struct. Mol. Biol.* **2007**, *14*, 23–29. [CrossRef]
13. Sun, J.; Hoshino, H.; Takaku, K.; Nakajima, O.; Muto, A.; Suzuki, H.; Tashiro, S.; Takahashi, S.; Shibahara, S.; Alam, J.; et al. Hemoprotein Bach1 Regulates Enhancer Availability of Heme Oxygenase-1 Gene. *EMBO J.* **2002**, *21*, 5216–5224. [CrossRef]
14. Liu, H.-L.; Zhou, H.-N.; Xing, W.-M.; Zhao, J.-F.; Li, S.-X.; Huang, J.-F.; Bi, R.-C. 2.6 Å Resolution Crystal Structure of the Bacterioferritin from Azotobacter Vinelandii. *FEBS Lett.* **2004**, *573*, 93–98. [CrossRef]
15. Bateman, T.J.; Shah, M.; Ho, T.P.; Shin, H.E.; Pan, C.; Harris, G.; Fegan, J.E.; Islam, E.A.; Ahn, S.K.; Hooda, Y.; et al. A Slam-Dependent Hemophore Contributes to Heme Acquisition in the Bacterial Pathogen Acinetobacter Baumannii. *Nat. Commun.* **2021**, *12*, 6270. [CrossRef]
16. 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]
17. Rydberg, P.; Sigfridsson, E.; Ryde, U. On the Role of the Axial Ligand in Heme Proteins: A Theoretical Study. *JBIC J. Biol. Inorg. Chem.* **2004**, *9*, 203–223. [CrossRef]
18. Walker, F.A. Magnetic Spectroscopic (EPR, ESEEM, Mossbauer, MCD and NMR) Studies of Low-Spin Ferriheme Centers and Their Corresponding Heme Proteins. *Coord. Chem. Rev.* **1999**, *185–186*, 471–534. [CrossRef]
19. Takano, Y.; Nakamura, H. Density Functional Study of Roles of Porphyrin Ring in Electronic Structures of Heme. *Int. J. Quantum Chem.* **2009**, *109*, 3583–3591. [CrossRef]
20. Kondo, H.X.; Kanematsu, Y.; Masumoto, G.; Takano, Y. PyDISH: Database and Analysis Tools for Heme Porphyrin Distortion in Heme Proteins. *Database* **2020**, *2020*, baaa066. [CrossRef]
21. Bikiel, D.E.; Forti, F.; Boechi, L.; Nardini, M.; Luque, F.J.; Martí, M.A.; Estrin, D.A. Role of Heme Distortion on Oxygen Affinity in Heme Proteins: The Protoglobin Case. *J. Phys. Chem. B* **2010**, *114*, 8536–8543. [CrossRef] [PubMed]
22. Sun, Y.; Benabbas, A.; Zeng, W.; Kleingardner, J.G.; Bren, K.L.; Champion, P.M. Investigations of Heme Distortion, Low-Frequency Vibrational Excitations, and Electron Transfer in Cytochrome C. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 6570–6575. [CrossRef]
23. Imada, Y.; Nakamura, H.; Takano, Y. Density Functional Study of Porphyrin Distortion Effects on Redox Potential of Heme. *J. Comput. Chem.* **2018**, *39*, 143–150. [CrossRef] [PubMed]
24. Takano, Y.; Kondo, H.X.; Kanematsu, Y.; Imada, Y. Computational Study of Distortion Effect of Fe-Porphyrin Found as a Biological Active Site. *Jpn. J. Appl. Phys.* **2020**, *59*, 010502. [CrossRef]

25. Kanematsu, Y.; Kondo, H.X.; Imada, Y.; Takano, Y. Statistical and Quantum-Chemical Analysis of the Effect of Heme Porphyrin Distortion in Heme Proteins: Differences between Oxidoreductases and Oxygen Carrier Proteins. *Chem. Phys. Lett.* **2018**, *710*, 108–112. [CrossRef]
26. Kondo, H.X.; Takano, Y. Analysis of Fluctuation in the Heme-Binding Pocket and Heme Distortion in Hemoglobin and Myoglobin. *Life* **2022**, *12*, 210. [CrossRef]
27. Kondo, H.X.; Fujii, M.; Tanioka, T.; Kanematsu, Y.; Yoshida, T.; Takano, Y. Global Analysis of Heme Proteins Elucidates the Correlation between Heme Distortion and the Heme-Binding Pocket. *J. Chem. Inf. Model.* **2022**, *62*, 775–784. [CrossRef]
28. Kondo, H.X.; Iizuka, H.; Masumoto, G.; Kabaya, Y.; Kanematsu, Y.; Takano, Y. Elucidation of the Correlation between Heme Distortion and Tertiary Structure of the Heme-Binding Pocket Using a Convolutional Neural Network. *Biomolecules* **2022**, *12*, 1172. [CrossRef]
29. Liao, F.; Yuan, H.; Du, K.-J.; You, Y.; Gao, S.-Q.; Wen, G.-B.; Lin, Y.-W.; Tan, X. Distinct Roles of a Tyrosine-Associated Hydrogen-Bond Network in Fine-Tuning the Structure and Function of Heme Proteins: Two Cases Designed for Myoglobin. *Mol. Biosyst.* **2016**, *12*, 3139–3145. [CrossRef]
30. Watanabe, Y.; Nakajima, H.; Ueno, T. Reactivities of Oxo and Peroxo Intermediates Studied by Hemoprotein Mutants. *Acc. Chem. Res.* **2007**, *40*, 554–562. [CrossRef]
31. Du, J.-F.; Li, W.; Li, L.; Wen, G.-B.; Lin, Y.-W.; Tan, X. Regulating the Coordination State of a Heme Protein by a Designed Distal Hydrogen-Bonding Network. *ChemistryOpen* **2015**, *4*, 97–101. [CrossRef]
32. Zhang, P.; Yuan, H.; Xu, J.; Wang, X.-J.; Gao, S.-Q.; Tan, X.; Lin, Y.-W. A Catalytic Binding Site Together with a Distal Tyr in Myoglobin Affords Catalytic Efficiencies Similar to Natural Peroxidases. *ACS Catal.* **2020**, *10*, 891–896. [CrossRef]
33. Wang, C.; Lovelace, L.L.; Sun, S.; Dawson, J.H.; Lebioda, L. Structures of K42N and K42Y Sperm Whale Myoglobins Point to an Inhibitory Role of Distal Water in Peroxidase Activity. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2014**, *70*, 2833–2839. [CrossRef]
34. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*; Pereira, F., Burges, C.J., Bottou, L., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2012; Volume 25.
35. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
36. Pu, L.; Govindaraj, R.G.; Lemoine, J.M.; Wu, H.-C.; Brylinski, M. DeepDrug3D: Classification of Ligand-Binding Pockets in Proteins with a Convolutional Neural Network. *PLoS Comput. Biol.* **2019**, *15*, e1006718. [CrossRef]
37. Kinjo, A.R.; Yamashita, R.; Nakamura, H. PDBj Mine: Design and Implementation of Relational Database Interface for Protein Data Bank Japan. *Database* **2010**, *2010*, baq021. [CrossRef]
38. Kinjo, A.R.; Suzuki, H.; Yamashita, R.; Ikegawa, Y.; Kudou, T.; Igarashi, R.; Kengaku, Y.; Cho, H.; Standley, D.M.; Nakagawa, A.; et al. Protein Data Bank Japan (PDBj): Maintaining a Structural Data Archive and Resource Description Framework Format. *Nucleic Acids Res.* **2012**, *40*, D453–D460. [CrossRef]
39. Hamelryck, T.; Manderick, B. PDB File Parser and Structure Class Implemented in Python. *Bioinformatics* **2003**, *19*, 2308–2310. [CrossRef]
40. Cock, P.J.A.; Antao, T.; Chang, J.T.; Chapman, B.A.; Cox, C.J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; et al. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* **2009**, *25*, 1422–1423. [CrossRef]
41. McGibbon, R.T.; Beauchamp, K.A.; Harrigan, M.P.; Klein, C.; Swails, J.M.; Hernández, C.X.; Schwantes, C.R.; Wang, L.P.; Lane, T.J.; Pande, V.S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015**, *109*, 1528–1532. [CrossRef]
42. Wang, G.; Dunbrack, R.L. PISCES: A Protein Sequence Culling Server. *Bioinformatics* **2003**, *19*, 1589–1591. [CrossRef] [PubMed]
43. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*; Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2019; pp. 8024–8035.
44. Wagner, J.R.; Sørensen, J.; Hensley, N.; Wong, C.; Zhu, C.; Perison, T.; Amaro, R.E. POVME 3.0: Software for Mapping Binding Pocket Flexibility. *J. Chem. Theory Comput.* **2017**, *13*, 4584–4592. [CrossRef] [PubMed]
45. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
46. Pesce, A.; Tilleman, L.; Dewilde, S.; Ascenzi, P.; Coletta, M.; Ciaccio, C.; Bruno, S.; Moens, L.; Bolognesi, M.; Nardini, M. Structural Heterogeneity and Ligand Gating in Ferric Methanosarcina Acetivorans Protoglobin Mutants. *IUBMB Life* **2011**, *63*, 287–294. [CrossRef]
47. Rice, S.L.; Boucher, L.E.; Schlessman, J.L.; Preimesberger, M.R.; Bosch, J.; Lecomte, J.T.J. Structure of Chlamydomonas Reinhardtii THB1, a Group 1 Truncated Hemoglobin with a Rare Histidine–Lysine Heme Ligation. *Acta Crystallogr. Sect. F Struct. Biol. Commun.* **2015**, *71*, 718–725. [CrossRef]
48. Ianiri, G.; Coelho, M.A.; Ruchti, F.; Sparber, F.; McMahon, T.J.; Fu, C.; Bolejack, M.; Donovan, O.; Smutney, H.; Myler, P.; et al. HGT in the Human and Skin Commensal Malassezia: A Bacterially Derived Flavohemoglobin Is Required for NO Resistance and Host Interaction. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 15884–15894. [CrossRef]

49. de Sanctis, D.; Dewilde, S.; Vonrhein, C.; Pesce, A.; Moens, L.; Ascenzi, P.; Hankeln, T.; Burmester, T.; Ponassi, M.; Nardini, M.; et al. Bishistidyl Heme Hexacoordination, a Key Structural Property in Drosophila Melanogaster Hemoglobin. *J. Biol. Chem.* **2005**, *280*, 27222–27229. [CrossRef]
50. Yoon, J.; Herzik, M.A.; Winter, M.B.; Tran, R.; Olea, C.; Marletta, M.A. Structure and Properties of a Bis-Histidyl Ligated Globin from Caenorhabditis Elegans. *Biochemistry* **2010**, *49*, 5662–5670. [CrossRef]
51. Walsh, I.; Fishman, D.; Garcia-Gasulla, D.; Titma, T.; Pollastri, G.; Capriotti, E.; Casadio, R.; Capella-Gutierrez, S.; Cirillo, D.; Del Conte, A.; et al. DOME: Recommendations for Supervised Machine Learning Validation in Biology. *Nat. Methods* **2021**, *18*, 1122–1127. [CrossRef]
52. Rost, B. PHD: Predicting One-Dimensional Protein Structure by Profile-Based Neural Networks. *Methods Enzymol.* **1996**, *266*, 525–539. [CrossRef]
53. Ward, J.J.; Sodhi, J.S.; McGuffin, L.J.; Buxton, B.F.; Jones, D.T. Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. *J. Mol. Biol.* **2004**, *337*, 635–645. [CrossRef]
54. Schwartz, D.; Chou, M.F.; Church, G.M. Predicting Protein Post-Translational Modifications Using Meta-Analysis of Proteome Scale Data Sets. *Mol. Cell. Proteom.* **2009**, *8*, 365–379. [CrossRef]
55. Gligorijević, V.; Renfrew, P.D.; Kosciolek, T.; Leman, J.K.; Berenberg, D.; Vatanen, T.; Chandler, C.; Taylor, B.C.; Fisk, I.M.; Vlamakis, H.; et al. Structure-Based Protein Function Prediction Using Graph Convolutional Networks. *Nat. Commun.* **2021**, *12*, 3168. [CrossRef]
56. Tunyasuvunakool, K.; Adler, J.; Wu, Z.; Green, T.; Zielinski, M.; Žídek, A.; Bridgland, A.; Cowie, A.; Meyer, C.; Laydon, A.; et al. Highly Accurate Protein Structure Prediction for the Human Proteome. *Nature* **2021**, *596*, 590–596. [CrossRef]
57. Kondo, H.X.; Kanematsu, Y.; Takano, Y. Structure of Heme-Binding Pocket in Heme Protein Is Generally Rigid and Can Be Predicted by AlphaFold2. *Chem. Lett.* **2022**, *51*, 704–708. [CrossRef]