 doi:10.20944/preprints202208.0029.v1

Article

Elucidation of the Correlation between Heme Distortion and Tertiary Structure of the Heme-Binding Pocket Using A 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 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, Tel.: +81-157-26-9401 (H.X.K); ykanem@hiroshima-u.ac.jp, Tel.: +81-82-424-7726 (Y.K.); ytakano@hiroshima-cu.ac.jp, Tel.: +81-82-830-1825 (Y.T.)

Abstract: Heme proteins serve diverse and pivotal biological functions. Therefore, clarifying the mechanisms of these diverse functions of heme is a crucial scientific topic. Distortion of heme porphyrin is one of the key factors regulating the chemical properties of heme. Here, we constructed convolutional neural network models for predicting heme distortion from the tertiary structure of the heme-binding pocket to examine their correlation. For saddling, ruffling, doming, and waving distortions, the experimental structure and predicted values were closely correlated. Furthermore, we assessed the correlation between the cavity shape and molecular structure of heme and demonstrated that hemes in protein pockets with similar structures exhibit near-identical structures, indicating the regulation of heme distortion through the protein environment. These findings indicate that the tertiary structure of the heme-binding pocket regulates the distortion of heme porphyrin, thereby controlling the chemical properties of heme relevant to the protein function; this implies a structure–function correlation in heme proteins.

Keywords: heme distortion; pocket conformation; convolutional neural network; machine learning

1. Introduction

Heme proteins are a group of proteins that bind heme(s)—a complex of iron and porphyrin—to serve diverse and important biological functions. The roles of heme in heme proteins are diverse; for instance, heme acts as an electron carrier[1,2], an active site for enzymes, such as oxidoreductases[3,4], an oxygen storage molecule[5,6], a ligand for proteins[7,8], and an iron storage molecule[9]. Hemophore proteins bind heme for its transport or storage[10]. Heme is classified into several types according to its peripheral groups (Figure 1), and the most common heme types are heme *b* and *c*[11,12]. Other major heme types include heme *a* and *o*, in addition to a few minor types. Presumably, the key factors regulating the diverse functions of heme include the distortion of heme porphyrin, axial ligands of heme, types of heme, and orientation of the propionate side chains. Heme distortion is correlated with the chemical properties of heme, such as redox potential and oxygen affinity[13].

 doi:10.20944/preprints202208.0029.v1

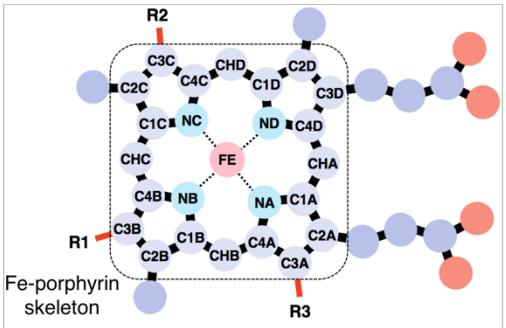
Normal-coordinate structural decomposition (NSD)[14] is one of the most common methods for estimating heme porphyrin distortion. In NSD, displacement from the equilibrium structure—or distortion—is represented as a linear combination of the vibrational modes of heme porphyrin. Among these, the three lowest vibrational modes of heme, namely saddling, ruffling, and doming (out-of-plane distortion), and the breathing mode (in-plane distortion) are closely correlated with its chemical properties. Bikiel et al.[15] clarified that the out-of-plane distortion tends to marginally decrease the binding affinity of heme for oxygen, while the breathing mode tends to decrease or increase it significantly. In a study on cytochrome c_{551}, Sun et al.[16] suggested a significant role of ruffling distortion in redox control. In a systematic study, Imada et al.[17] examined the association between saddling and ruffling distortions and redox potential and indicated that saddling distortion increases the redox potential of heme, while ruffling distortion exhibits the opposite tendency. In another study, a novel distortion correlated with the chemical properties of heme was elucidated. Kanematsu et al.[18] analyzed the molecular structures of hemes in oxidoreductases and oxygen-binding proteins and successfully discovered a distortion correlated with both redox potential and oxygen affinity.

Regulation of heme distortion through the host protein environment is one of our research interests. Heme in its host protein exhibits various degrees of distortion from the isolated structure[12]. Our simulation study revealed that doming distortions in the oxygenated and deoxygenated states differ between hemoglobin and myoglobin, suggesting that the molecular structure of heme is affected by its protein environment, which controls the chemical properties of heme relevant to the protein function[19]. Some studies have reported the structural rigidity of heme-binding pockets. In addition, studies on protein structures in the apo (heme-unbound) and holo (heme-bound) states have shown that most apo–holo pairs exhibit small structural differences[20,21]. Using Brownian dynamic simulations, Sacquin-Mora et al.[22] showed that residues in the heme-binding site must be tightly anchored to realize biological functions, except for those flexible for protein function.

Our recent study using machine learning indicated a correlation between the amino acid composition of the heme-binding pocket and heme distortion along the three lowest vibrational modes[23]. Here, we investigated the correlation between the tertiary structure of the heme-binding pocket and distortion of heme by predicting the latter from the former using a convolutional neural network (CNN). CNN is a deep learning method that has enabled breakthroughs in various computer vision tasks, such as image classification[24,25].

To this end, in the present study, we constructed a CNN model and trained it to predict heme distortion from the structure of the heme-binding pocket. We obtained high correlation coefficients for saddling, ruffling, doming, and waving(y) distortions, suggesting an association between heme-binding pocket structure and heme distortion for these vibrational modes. Furthermore, we revealed that hemes in protein pockets with similar structures exhibit near-identical structures. These results suggest that the protein environment of the heme-binding pocket regulates the molecular structure of heme, thereby controlling the chemical properties of heme relevant to protein function. This is a significant finding of the structure–function correlation in heme proteins and will be conducive for designing protein functions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 August 2022 doi:10.20944/preprints202208.0029.v1



	R1	R2	R3
heme *a*	hydroxyethyl-farnesyl group	–CH=CH_2	–CH=O
heme *b*	–CH=CH_2	–CH=CH_2	–CH_3
heme *c*	–$CH(CH_3)$–Cys	–$CH(CH_3)$–Cys	–CH_3
heme *o*	hydroxyethyl-farnesyl group	–CH=CH_2	–CH_3

Figure 1. Chemical structures of hemes *a*, *b*, *c*, and *o*.

2. Materials and Methods

2.1. Data collation on heme proteins and dataset preparation for deep learning

Structural information on heme proteins was extracted from the PDBx/mmCIF files downloaded from the Protein Data Bank Japan (PDBj)[26]. Briefly, we collated PDB entries, including compound IDs (_chem_comp.id), of HEM, HEA, HEB, HEC, and HEO, and their structures at a resolution ≤ 2.0 Å via SQL search in PDBj Mine relational database[27] (https://pdbj.org/rdb/search). Hemes with missing data in the coordinates of 25 atoms that form the Fe–porphyrin skeleton (Figure 1, upper panel) were excluded. Consequently, 6,677 heme samples from 3,121 unique PDB entries were selected (as of January 4, 2022). The Bio.PDB package[28] for BioPython[29] and MDTraj library[30] were used to analyze the structural data. The type of each heme molecule was determined based on peripheral groups, and the type was considered "unknown" when atoms were missing from the structural data of a heme. Protein function was classified based on structural keywords stored in the PDB entry. Details of heme protein data collection are described in our previous studies[12,23].

To eliminate the effect of binding of small molecules on the pocket structure, we collected heme molecules whose axial ligands were composed of regular amino acids only as the whole dataset. At this stage, 3,843 samples were extracted. The axial ligands were defined as the residues or molecules with atom(s) within 3.1 Å from the Fe atom of heme. To reduce sequence redundancy in the whole dataset, we excluded protein chains with the same amino acid sequence using the PISCES server[31], yielding a nonredundant dataset. The nonredundant dataset contained 939 samples. Since even a slight difference in the amino acid sequence can affect the tertiary structure of the heme-binding pocket and distortion of heme, the threshold for sequence similarity was set to 99.99%.

The distortion of heme porphyrin was estimated using NSD[14], which is a common method for evaluating heme conformation. As mentioned earlier, NSD represents porphyrin distortion as a linear combination of distortions along the vibrational modes of heme. We calculated the equilibrium structure and vibrational modes of the Fe–porphyrin molecule using the PBE0 hybrid functional[32] with 6-31G(d) basis sets[33–35] and used these to estimate heme distortion. Details of calculation are described elsewhere[12]. Only 12 vibrational modes described by Bikiel et al.[15] were considered.

 doi:10.20944/preprints202208.0029.v1

2.2 CNN model

We converted the non-uniform protein structural data into uniform dimensional data for use as input for the CNN model. Although a set of voxels is a candidate to represent the tertiary structures of protein pockets, determining the pocket area is a problem. As shown in Figure 2a, sets of voxels in a cube-shaped inclusion region centered on the heme-binding pocket were used as input for the CNN model in the present study. The location of the cube was defined as follows. First, a least-squares plane was calculated for four atoms in the Fe–porphyrin skeleton of heme, namely CHA, CHB, CHC, and CHD (the correspondence between atom positions and names is presented in Figure 1), and defined as the xy-plane. The xy-plane was rotated such that the x-axis was parallel to the vector connecting CHA and CHC projected the least-squares plane (Figure 2b). Then, the z-axis was determined to be perpendicular to the x- and y-axes and right-handed, and the origin was determined as the mean coordinate of the four atoms: CHA, CHB, CHC, and CHD. The cube was placed such that each edge of the cube was parallel to the x-, y-, and z-axes, and the center was at the origin (0, 0, 0). The edge length was set to 17, 20, and 24 Å to examine the effect of inclusion region size on prediction. Next, we demonstrated voxelization of the inclusion region. Using a protein structure without heme and other molecules, we generated a cubic grid with 1 Å spacing, computed whether each area was occupied by any atom, and assigned 0 (unoccupied) or 1 (occupied) to each grid. The occupied region by each atom was defined as the region within a sphere whose radius is half the length of the Van der Waals radius of the atoms C:1.70 Å, N:1.55 Å, O:1.52 Å, and S:1.80 Å. The voxels were calculated for each atom (C, N, O, and S), and the generated data were used as an input with four channels (right panels of Figure 2a). The output was the distortion of heme porphyrin along the 12 vibrational modes (12 dimensions) or each vibrational mode (one dimension). Loss was calculated as the mean-square error between the observed (experimentally determined) and predicted values.

All CNN models were constructed and trained using PyTorch[36]. The model used in the present study is described in Figure 2c and Table 1. The dimensions of data presented in Figures 2a and 2c are for the case in which the edge length in the inclusion region was 20 Å. Here, we briefly demonstrate the method commonly used in CNNs: convolution, batch normalization, activation function, pooling, and dropout. The convolutional layer, as exemplified by Conv3d in PyTorch, is the main building block of a CNN and plays a role in the extraction of local features. It selects a dot product between the values of the input voxels and filter weights. The hyperparameters of convolution include the number of output channels (number of filters), kernel size of filters, stride (number of voxels that move a filter in each step), and presence or absence of padding (adding voxels outside the input voxels). Batch normalization, as exemplified by BatchNorm3d or BatchNorm1d in PyTorch, is a method for standardizing the inputs over mini-batches to stabilize and accelerate training by reducing the internal covariate shift. An activation function adds nonlinearity to the output and helps the neural network to learn complex patterns. Rectified linear unit (ReLU), sigmoid, and hyperbolic tangent functions are common activation functions. We used ReLU function in the present study. Pooling is a technique used to reduce feature dimensions. Max pooling, as exemplified by MaxPool3d in PyTorch, is the most commonly used pooling method. It selects the maximum value in each kernel of a feature map and generates a down-sampled feature map. The hyperparameters of max pooling include the kernel size of filters and stride. Finally, feature maps in the CNN were fully connected. Specifically, the weighted sum of outputs was computed from previous layers to obtain a specific output. A dropout layer is often added to avoid over-learning. Outputs of a randomly selected set of neurons were ignored during training. The probability of ignoring nodes is specified by a hyperparameter.

To verify the generalization performance of the model, five-fold cross-validation was performed for each vibrational mode. We did not isolate a test dataset from a cross-

 doi:10.20944/preprints202208.0029.v1

validation dataset because of limited data. The non-redundant dataset was split into five subsets after shuffling the samples. The following steps were performed for each subset:

1. The subset was split into validation and test datasets at a ratio of 0.2:0.8.
2. The model was trained using the remaining four subsets for 300 epochs.
3. The model with the minimum value of loss, calculated as the mean-square error, in the validation dataset was selected.
4. The resulting model was validated on the test dataset.

Adam optimizer[37] with a learning rate of 0.01 was used for training. Batch size was set to 32.

2.3 Clustering and principal component analyses of heme-binding pockets

We analyzed the three-dimensional shapes of heme-binding pockets (cavity) using POVME 3.0[38]. In POVME, the cavity shape of a ligand-binding protein structure can be quantified as a bit vector, each element of which indicates whether the respective grid point belongs to the ligand-binding pocket. The protein structures complexed with heme were superimposed on five atoms in heme: FE, NA, NB, NC, and ND. The coordinates of the missing atoms for proteins were generated using the AMBER LEaP program included in AmberTools version 19.0[39]. The grid structure of the cavity was computed using only protein coordinates (i.e., heme and other molecules were removed). Parameters for POVME calculation were as follows: the center and radius of the inclusion sphere were set to the coordinates of the Fe atom of heme and 8.5 Å, respectively. This radius was determined according to the molecular size of heme. The distance between Fe and oxygen atoms of propionates was approximately 8.5 Å. The Tanimoto score implemented in POVME 3.0 was used to estimate the similarity between pairs of heme-binding pockets. Hierarchical clustering and principal component analysis (PCA)[40] of cavity shapes were performed using POVME 3.0. The number of clusters was set to 35. We examined three cases of the number of clusters (15, 25, and 35) and obtained the most preferable results (many eigenvectors correlated with heme distortions) for 35.

2.4 Alignment of amino acid sequences of heme proteins

We downloaded the amino acid sequences of the target heme proteins as FASTA files from PDBj (as of January 4, 2022) and extracted the sequences of 2,867 protein chains in the whole dataset. Clustering was performed for the obtained sequence data using Cd-Hit[41], and threshold of sequence similarity was set to 90%.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 August 2022 doi:10.20944/preprints202208.0029.v1

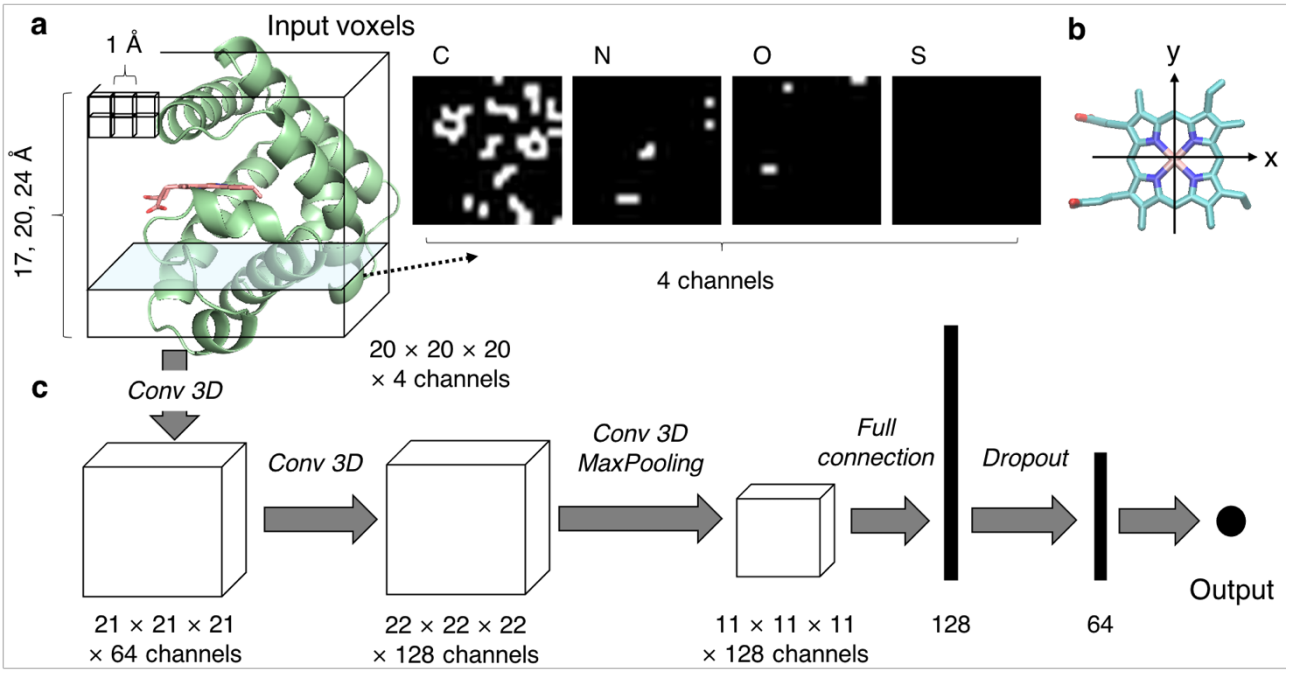


Figure 2. CNN model used in the present study. (**a**) A schematic diagram of input voxels. The protein backbone is represented as a green cartoon, and the heme molecule is shown as the licorice model colored in salmon. The input voxels were calculated for each atom (C, N, O, or S), as illustrated in the right panel. The heme molecule(s) were excluded in the voxel calculation. (**b**) A diagram of determination of x- and y-axes based on the coordinates of heme for the calculation of input voxels. The heme molecule is represented as the licorice model, and the atoms used for the determination of the axes are shown by dotted circles. (**c**) Layers included in the developed CNN model are shown.

Table 1. The layers and parameters of the CNN model used in this study.

Layer	Function	Kernel/Filter	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	-	128000
9	Linear	-	128
10	ReLU	-	128
11	Dropout	0.4	128
12	Linear	-	64
13	BatchNorm1d	-	64
14	ReLU	-	64
15	Linear	-	1 (or 12)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 August 2022 doi:10.20944/preprints202208.0029.v1

3. Results and Discussion

3.1. Prediction of heme distortion from the tertiary structure of the heme-binding pocket using a CNN model

We constructed a model to simultaneously predict the magnitude of distortions along the 12 vibrational modes. The edge length of the input voxel was set to 20 Å. The obtained models were assessed based on the R^2 score calculated as follows:

$$R^2 = 1 - \frac{\sum_i (p_i^{\text{observed}} - p_i^{\text{predicted}})^2}{\sum_i (p_i^{\text{observed}} - \bar{p}^{\text{observed}})^2}, \quad (1)$$

where p_i^{observed} and $p_i^{\text{predicted}}$ are the distortions of i^{th} heme molecule in the PDB structures and as predicted by the CNN model, respectively, and $\bar{p}^{\text{observed}}$ is the mean of heme distortion in the PDB structures in the test dataset. The R^2 score is a measure used to evaluate how well the model fits the regression, and its values ranges from $-\infty$ to 1. A relatively strong correlation (correlation coefficient ≥ 0.6) was observed between the predicted and true values for saddling, ruffling, doming, and waving(y) distortions. Detailed prediction results are presented in Table S1, and the plot of observed and predicted values is shown in Figure S1 using results from the model with the maximum R^2 score among the five cross-validation runs as an example.

To examine the effect of different edge lengths of input voxels on the prediction result, we constructed models using the input voxels with edge lengths of 17, 20, and 24 Å (an example is shown in Figure 3a) for each of these four vibrational modes and calculated the corresponding R^2 score. The means and standard deviations of R^2 scores of the five cross-validation runs are shown in Figure 3b. Except for the waving(y) mode, changes in R^2 score due to differences in the edge length of the input were very small, suggesting that information on the structure of the heme-binding pocket near the surface is sufficient to predict heme distortion. In our previous study examining the correlation between the composition of amino acid residues in the heme-binding pocket and heme distortion[23], no correlation was detected for the waving(y) mode, as opposed to that for the first three vibrational modes. This might be because more detailed information on the tertiary structure of the pocket enabled us to predict even a small conformational difference.

Next, we focused on the three vibrational modes correlated with the redox potential[17] and oxygen affinity[15] of heme: the saddling, ruffling, and doming modes. The input edge length was set to 24 Å because high R^2 scores were obtained for all three vibrational modes. The means and standard deviations of R^2 scores and the root-mean-square errors (RMSEs) of the five cross-validation runs are presented in Table 2, and the corresponding correlation coefficients are presented in Table S2. Although the variation in scores among the cross-validation runs was higher for the doming distortion than for the other two distortions, we noted a strong correlation between the observed and predicted values for all three modes. In particular, high correlation coefficients were obtained for the saddling distortion, regardless of the combination of the test and training datasets; the minimum value of the correlation coefficient was 0.77.

Table 2. The results of the prediction by the input voxels with the edge length of 24 Å. The mean value and standard deviation of R^2 score, and RMSE values are listed.

Vibrational mode	Saddling	Ruffling	Doming
R^2 score	0.62 ± 0.05	0.50 ± 0.09	0.46 ± 0.15
(max., min. values)	(0.70, 0.55)	(0.65, 0.39)	(0.70, 0.25)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 August 2022 doi:10.20944/preprints202208.0029.v1

RMSE	0.21 ± 0.02	0.31 ± 0.04	0.16 ± 0.03
(min., max. values)	(0.20, 0.24)	(0.25, 0.37)	(0.11, 0.20)

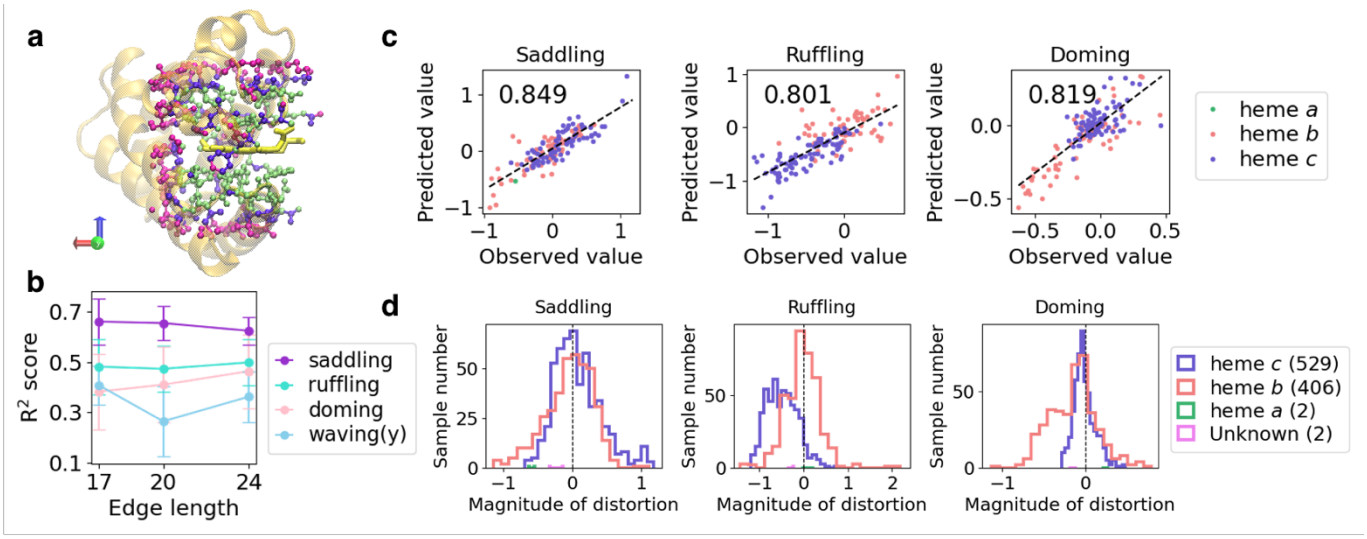


Figure 3. (**a**) Atoms in the inclusion region with the edge length of 17 (lime), 20.0 (violet), and 24.0 (magenta) Å, as exemplified by PDB ID: 1mba. The whole protein structure and heme molecule are shown as the orange cartoon and the yellow licorice model, respectively. (**b**) Plot of R^2 scores averaged over five cross-validation runs *versus* the edge length of the input voxels for each heme distortion. (**c**) Correlation between the predicted and observed values in the test dataset of the best model among five cross-validation runs for each heme distortion. Values on the upper left of each panel represent correlation coefficients. Slate-blue, light-coral, and sea-green points indicate heme *c*, *b*, and *a*, respectively. (**d**) Distribution of saddling, ruffling, and doming distortions for each heme type in the non-redundant dataset.

To examine the effect of heme type on prediction, the RMSE values for the test datasets in each cross-validation run were calculated for each heme type (Table 3). Regarding the correlation between the protein environment and heme type, only heme *c* forms covalent bonds with its host protein, causing distortion along the ruffling mode[42]. The prediction results for each heme type are shown as color-coded points in Figure 3c, and the histograms of each distortion for each heme type are presented in the lower panels of Figure 3d. Heme *c* tends to be distorted toward the ruffling mode. For ruffling and doming distortions, the RMSE values for heme *c* were almost half of those for heme *b*, suggesting a strong effect of the protein environment on heme distortion. Further, we analyzed the effect of protein function on prediction. However, the results were not sufficiently simple to observe differences in the degree of distortion for each protein function (Table S3 and Figure S2).

Table 3. The mean values and standard deviations of RMSE between the observed and predicted values for each heme type.

Vibrational mode/ Protein function	Saddling	Ruffling	Doming
heme *c* (85.8 ± 2.7) †	0.20 ± 0.01	0.22 ± 0.02	0.11 ± 0.01
heme *b* (64.2 ± 3.0)	0.22 ± 0.02	0.41 ± 0.07	0.22 ± 0.06

† Values in parentheses represent the means of the sample numbers in the test set for five cross-validation runs.

3.2 Differences in the importance of information included in subsets of input data

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 August 2022 doi:10.20944/preprints202208.0029.v1

To specify a region important for predicting heme distortions, we discarded the information of a specific region of input voxels and computed prediction scores using the model described in section 3.1 (edge length of input = 24 Å). Information was discarded in two ways: removing information from the outside (or "outside discarding," Figure 4a) and from the inside (center) (or "inside discarding," Figure 4b). First, we defined two cubes: "outer cube" formed of vertices with coordinates of (±12, ±12, ±12) and "inner cube" (right panels in Figures 4a-b). Let the coordinates of vertices of the inner cube on the "outside discarding" and "inside discarding" be (±(12-r), ±(12-r), ±(12-r)) and (±r, ±r, ±r), respectively. Then, we denote the sets of voxels in the outer and inner cubes as V_{outer} and V_{inner}, respectively. For "outside discarding," the elements of V_{outer}-V_{inner} (a set of elements in V_{outer} but not in V_{inner}) were replaced by 0 ($0 \le r < 12$, Figure 4a), that is, the information was removed from the outside of the input voxels. For the "inside discarding," the elements of V_{inner} were replaced by 0 ($0 \le r < 12$, Figure 4b), that is, the information was removed from the inside. Since V_{outer} is equivalent to the input voxels used to train the CNN model, the information is intact when $r = 0$ in both cases.

Mean R^2 scores obtained from predictions for each test dataset in the five cross-validation runs are shown in the left panels of Figures 4a-b. Because change in the amount of information loss for a change in r was not linear and differed between "outside discarding" and "inside discarding," we also plotted the resulting R^2 scores against the volume of the region where the information remained (Figure 4c). As shown in Figure 4c, the change in R^2 scores was not correlated with the amount of information but depended on region included in the input for prediction. With "outside discarding" (Figure 4a), the scores started decreasing significantly at r = 4–6 Å, where the edge length of the inner cube was 16–12 Å, reaching almost 0 at r = 7 Å, where the edge length of the inner cube was 10 Å. Meanwhile, for "inside discarding" (Figure 4b), the scores did not largely change at r = 4 Å, where the edge length of the inner cube was 8 Å, but decreased slowly at r = 5 Å, where the edge length of the inner cube was 10 Å. Based on these results, information from an inclusion region with the edge length of 8–16 Å is essential, while that from an inclusion region with the edge length of 8 Å is non-essential, and A_l is a set of atoms included in the cubic region with edge lengths of 2l. Examples of A_l (l = 4, 5, 6, and 7) are illustrated in Figure 4d using PDB ID 1mba[43]. From these results, a cubic region with the edge length (2l) of <8 Å contains very few protein atoms; therefore, the structure of the pocket surface is considered to be important for prediction.

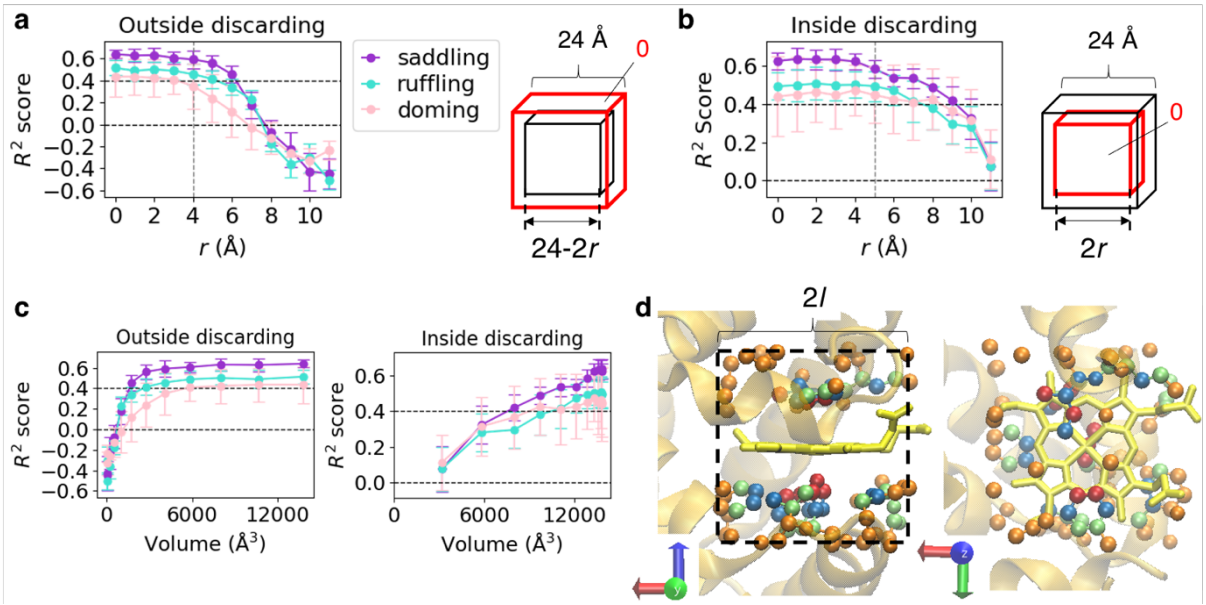


Figure 4. (**a**) Mean R^2 scores for each vibrational mode *versus* r, which represents the distance between the faces of the red and black cubes illustrated in the right panel. The error bar shows

 doi:10.20944/preprints202208.0029.v1

standard deviation. The red and black cubes have the identical center, and their edges are parallel. The red cube is equivalent to the inclusion region used as the CNN input. Voxel values in the region between the red and black cubes were replaced by 0. (**b**) Mean R^2 scores for each vibrational mode versus r. Voxel values in the red cube were replaced by 0. The coloring method is the same as that in (**a**). (**c**) R^2 scores identical to those in (**a**) and (**b**) *versus* the volume of region including the original information. The coloring method is the same as that in (**a**). (**d**) Atoms included in cube-shaped regions with the edge length of $2l$ are illustrated using PDB ID 1mba as an example. The dark-red, lime, marine-blue, and orange spheres represent l = 4, 5, 6, and 7, respectively. The backbone of the host protein is represented as an orange cartoon and heme as a yellow licorice model.

Furthermore, we examined the impact of separation of atomic species in the input on prediction. The CNN model shown in Figure 2c was trained and validated on a dataset with one-channel inputs (only the input dimension was different from the model in Table 1). The one-channel input was generated by calculating the logical sum (OR) of the four-channel inputs; therefore, the difference in atomic species was not considered. The results of five-fold cross-validation are presented in Table 4. The R^2 score decreased in the ruffling mode, whereas no large difference was noted in the saddling and doming modes, suggesting that the steric effect was dominant for the latter two distortions.

Table 4. Results of prediction by the model which takes voxels with one-channel as an input.

Vibrational mode	Saddling	Ruffling	Doming
R^2 score	0.63 ± 0.07	0.39 ± 0.10	0.43 ± 0.17
(max., min. values)	(0.72, 0.53)	(0.52, 0.24)	(0.68, 0.17)
RMSE	0.21 ± 0.02	0.34 ± 0.02	0.16 ± 0.03
(min., max. values)	(0.19, 0.25)	(0.31, 0.37)	(0.12, 0.21)

3.3 Similarity of the structure of heme-binding pockets and hemes

To elucidate the association between the shape of the heme-binding pocket and heme distortion, we evaluated the similarity of cavity shapes, which is a structural property of the region surrounded by the protein, for pairs of protein chains. Since we considered only the structure in the vicinity of the target heme, the cavity shapes of hemes binding to a unique pocket varied in the present study. The cavity shape of the i^{th} sample was represented as a bit vector using POVME, referred to as cavity vector $\boldsymbol{v}_i$, and the similarity score between the i^{th} and j^{th} samples was calculated as the Tanimoto score between $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$. The Tanimoto score ranges from 0 to 1, with 1 indicating identical shapes. Because the number of combinations of protein chains was very large for analysis, the pairs were randomly sampled without replacement from the whole or non-redundant dataset. The similarity score was plotted against the root-mean-square deviation (RMSD) of the heavy atoms of the heme Fe–porphyrin skeleton (Figure 5a). The pairs with high similarity scores showed small RMSD values for heme, indicating that hemes exhibit similar structures in protein pockets of similar structures. In addition, some pairs with low similarity scores showed small RMSD values for heme, indicating the lack of one-to-one correspondence between cavity shape and heme distortion.

To elucidate the simple correlation between cavity shape and heme distortion, we performed hierarchical clustering of cavity shapes for the whole dataset, followed by PCA of cavity shapes in each cluster (i.e., we conducted PCA for a group of samples with similar cavity shapes). In some clusters, we obtained eigenvectors correlated to heme distortion. Two examples with high correlation coefficients are shown in Figure 5b (clusters 9 and 11). In cluster 9, PC1 values of cavity shapes were correlated with doming distortion, with a correlation coefficient of 0.84. In cluster 11, PC1 values of cavity shapes were correlated with the saddling distortion, with a correlation coefficient of 0.99. In these examples, a difference along eigenvector led to a large difference in heme distortion, as shown in Figure 5b. The corresponding eigenvectors for clusters 9 and 11 are shown in

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 August 2022 doi:10.20944/preprints202208.0029.v1

Figures 5c and 5d, respectively. In cluster 9, the area corresponding to the element of the eigenvector with a relatively large value surrounded the Fe atom and was distributed at periphery of the heme molecule. Meanwhile, in cluster 11, this area was distributed only at the periphery of the heme molecule. Therefore, the cavity shape of the periphery of heme may be important for saddling distortion, whereas protein structure surrounding the Fe atom may be significant for doming distortion. Incidentally, we could not obtain features correlated with heme distortion using PCA for all samples in the whole dataset. Therefore, heme distortion is regulated by even a slight difference in cavity shape, and it is smaller than the difference in structures between all protein chains in the whole dataset (differences between clusters would be preferentially detected using PCA).

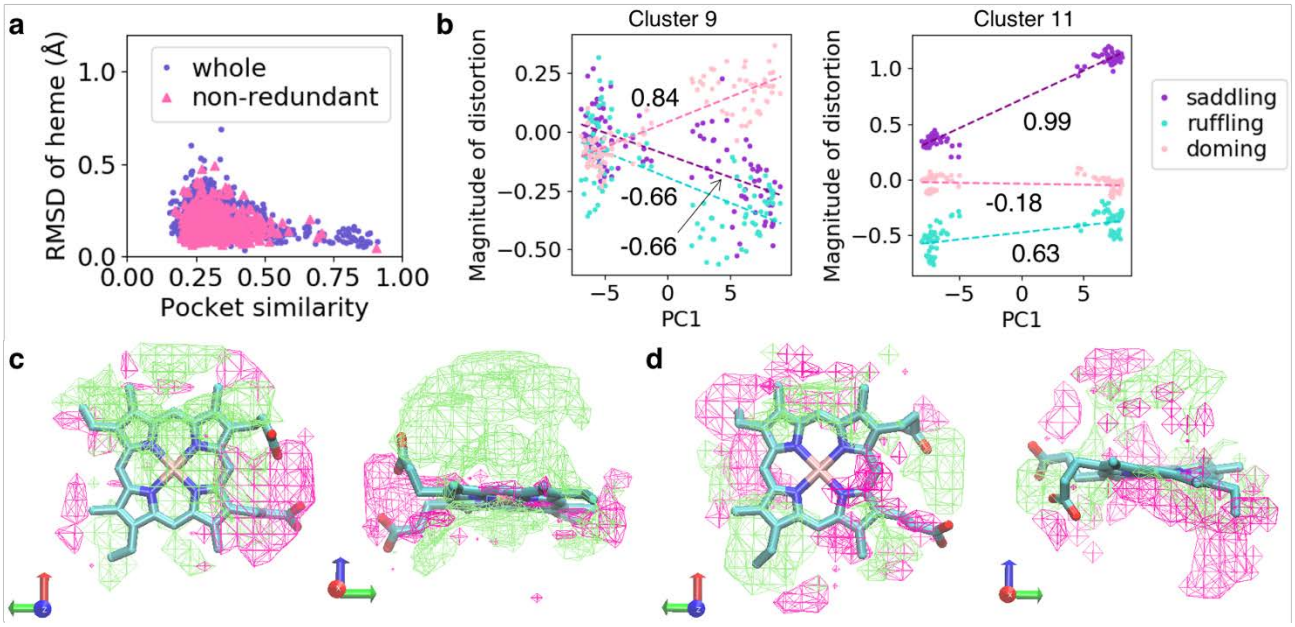


Figure 5. (**a**) Plot of similarity scores of cavity shapes *versus* RMSD of heme for the pairs of protein chains in the whole and non-redundant datasets. (**b**) Plot of PC1 values of cavity shapes *versus* the magnitude of distortion of heme in clusters 9 (left panel) and 11 (right panel). Dashed lines colored in the dark-orchid, pink, and turquoise are linear regression lines for saddling, ruffling, and doming distortions, respectively. Values in the graph are correlation coefficients calculated from linear regression analysis. (**c**, **d**) First eigenvectors obtained from PCA for clusters 9 (**c**) and 11 (**d**). Lime and magenta mesh surfaces represent the isosurfaces of +0.25 and -0.25. Structures with large PC1 values would have the cavity containing lime area but not the magenta area. Heme is represented as the licorice model. Left and right panels show the same vector viewed from different directions.

3.4 Similarity of the structures of heme-binding pockets between protein chains with similar amino acid sequences

To estimate the correlation between the amino acid sequences and cavity shapes of the pocket, we analyzed the variability of cavity shapes among homologous protein chains. By clustering protein chains in the whole dataset according to amino acid sequence, 2,867 protein chains were classified into 399 clusters. From these clusters, we selected 10 clusters in the order of the number of protein chains in a cluster. Let I be a set of samples of cavity shapes in a cluster (the number of protein chains does not correspond to the number of heme-binding pockets because of the existence of multi-heme proteins). To estimate the dispersion of cavity shapes, we calculated the mean distance from the barycenter for cavity vector $\boldsymbol{v}_i$ in each cluster as follows:

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

$$\boldsymbol{\mu}_I = \frac{1}{N_I} \sum_{i \in I} \boldsymbol{v}_i, \tag{3}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 August 2022 doi:10.20944/preprints202208.0029.v1

$$\bar{d}_I = \frac{1}{N_I} \sum_{i \in I} \|\boldsymbol{v}_i - \boldsymbol{\mu}_I\|, \quad (4)$$

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

The number of protein chains, number of samples, $\bar{d}_I$, and protein names for each cluster are presented in Table 5. Results for the whole dataset (3,843 samples) are included at the bottom of the table for reference. For smaller $\bar{d}_I$ values, higher similarity was expected for cavity shapes in a cluster.

For 6 of the 10 clusters, the mean distance ($\bar{d}_I$) was smaller than half for the whole dataset ($\bar{d}_I^{\text{whole}}$), while for 2 of them (total eight clusters), the value was <60% of $\bar{d}_I^{\text{whole}}$, indicating that pocket structures are similar between protein chains with near-identical amino acid sequences. The former six clusters in Table 5, whose indices are 1, 2, 3, 4, 5, and 8, include nitric oxide synthase[44], bacterioferritin[45], and hemoglobin α and β chains[46,47]. Bacterioferritin functions as an iron storage molecule or an oxidoreductase and is composed of 12 homo-dimers (i.e., 24-mer protein). Since some PDB structures only include coordinates of the asymmetric unit, resulting in a split of heme-binding pockets[48], we excluded samples with a heme coverage of <0.6. The latter clusters ($\bar{d}_I \leq 0.6 \times \bar{d}_I^{\text{whole}}$) with indices of 6 and 9 included cytochrome *c* oxidase[49] and cytochrome *c*[2], respectively. Therefore, these may be important to maintain the microstructure of the heme-binding pocket for redox control. For cluster 7, which included dehaloperoxidases[50], $\bar{d}_I$ was slightly larger. This protein harbors a globin-like fold and functions as an oxygen storage molecule, similar to hemoglobin and peroxidase. The conformational flexibility of distal histidine increases in the deoxygenated state[51], which may explain the slightly large $\bar{d}_I$ value. Meanwhile, the $\bar{d}_I$ value of cluster 10 was much larger than that of the other clusters. This cluster included eight-heme nitrite reductase. This enzyme possesses eight heme-binding sites, of which three are in the N-terminal domain and the remainder are in the catalytic C-terminal domain[52]. As shown in Figure 6, structural differences in these eight pockets may explain the large $\bar{d}_I$ value.

Table 5. The cluster indices, sample numbers, $\bar{d}_I$, and protein names of each cluster. The shaded lines represent the clusters with large $\bar{d}_I$.

Cluster index	Sample number	$\bar{d}_I$	Protein name
1	407 (407)†	7.55	Nitric-oxide synthase
2	146 (146)	8.43	Hemoglobin (beta chain)
3	133 (95)	5.46	Bacterioferritin
4	103 (103)	7.72	Hemoglobin (alpha chain)
5	99 (99)	8.18	Nitric oxide synthase
6	64 (81)	8.94	Cytochrome *c* oxidase subunit 1
7	55 (55)	11.14	Dehaloperoxidase
8	50 (50)	6.41	Nitric oxide synthase oxygenase
9	47 (47)	9.84	Cytochrome *c*
10	46 (321)	14.46	Eight-heme nitrite reductase
whole dataset	3843	17.27	-

† Values in parentheses represent the number of heme-binding pocket samples.

 doi:10.20944/preprints202208.0029.v1

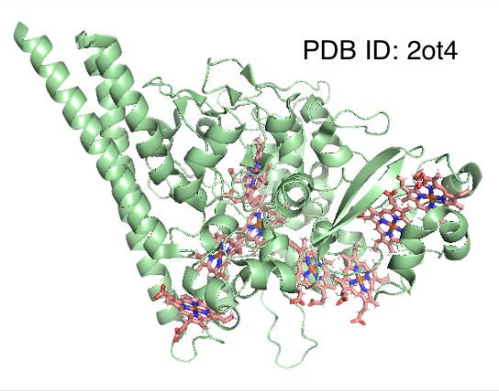


Figure 6. Structure of eight-heme nitrite reductase. The protein backbone is represented as a green cartoon, and hemes are shown as stick models.

4. Conclusions

In the present study, we constructed a CNN model to predict heme distortion from the tertiary structure of the heme-binding pocket to examine the correlation between them. The correlation between heme-binding pocket structure and heme distortion suggests that the protein environment affects the distortion of heme and regulates its chemical properties. High R^2 scores were obtained from prediction using the CNN model for saddling, ruffling, doming, and waving(y) distortions. In our previous study[23], no correlation was indicated for waving(y) distortion, as opposed to that for the remaining three distortions. This may be because detailed information on the tertiary structures of heme-binding pockets enabled us to predict even small conformational differences. These results of prediction based on partial information of heme-binding pocket suggests that the structural information of the pocket surface is significant for the prediction of heme distortion, and the steric effect is dominant, particularly in the saddling and doming modes.

Furthermore, we examined the correlation between the shape of cavity and molecular structure of heme and showed that hemes in protein pockets with similar structures exhibit near-identical structures. Therefore, heme distortion may be regulated by the protein environment. Finally, we estimated the correlation between the amino acid sequences and cavity shapes of heme-binding sites. The variability of cavity shapes was compared among clusters of protein chains with 90% or higher sequence similarity. We selected 10 clusters with a large number of samples and found that eight of them showed a mean distance ($\bar{d}_I$) of <60% of that for the whole dataset. Therefore, pocket structures are similar among protein chains with near-identical amino acid sequences.

Overall, the tertiary structure of the heme-binding pocket is determined by the amino acid sequence of protein chain, and it regulates the molecular structure of heme, thereby controlling its chemical properties, as relevant to the protein function. In the future, by predicting the location of the heme-binding site, the function of heme proteins may be predicted based on the amino acid sequence of the protein.

Supplementary Materials: The following supporting information can be downloaded at www.mdpi.com/xxx/s1, Table S1: Prediction results of heme distortions along the 12 vibrational modes; Figure S1: Correlation between the predicted and observed values; Table S2: Prediction results from input voxels with the edge length of 24 Å; Table S3: Mean values and standard deviations of RMSE for each protein function; Figure S2: Distribution of saddling, ruffling, and doming distortions for each protein function.

Author Contributions: Conceptualization, H.X.K, H.I., and G.M.; Methodology, H.X.K., H.I., Y.Kb., G.M., Y.Kn., 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.Kb, Y.Kn., 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., and Y.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by FY 2020 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 and 19H02752. The APC was funded by a Grant-in-Aid for Scientific Research on Transformative Research Areas (A), 20H05883.

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/). Our collated data on hemes are available in PyDISH: https://pydish.bio.info.hiroshima-cu.ac.jp/. For convenience, the list of PDB IDs is provided in the Supplementary Materials.

Acknowledgments: Computations were performed at the Research Center for Computational Science, Okazaki, Japan, and RIKEN Advanced Center for Computing and Communication (ACCC). The present study was performed in part under the Collaborative Research Program of the Institute for Protein Research, Osaka University (CR-18-02, CR-19-02, CR-20-02, and CR-21-02).

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

References

1. Poulos, T. L. The Janus Nature of Heme. *Nat. Prod. Rep.* **2007**, *24* (3), 504–510. https://doi.org/10.1039/B604195G.
2. 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* (2), 527–555. https://doi.org/10.1016/0022-2836(90)90197-T.
3. 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* (6), 2279–2328. https://doi.org/10.1021/cr030722j.
4. Ostermeier, C. Cytochrome c Oxidase. *Curr. Opin. Struct. Biol.* **1996**, *6* (4), 460–466. https://doi.org/10.1016/S0959-440X(96)80110-2.
5. Perutz, M. F.; Rossmann, M. G.; Cullis, A. F.; Muirhead, Hi.; 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* (4711), 416–422. https://doi.org/10.1038/185416a0.
6. 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* (4711), 422–427. https://doi.org/10.1038/185422a0.
7. Faller, M.; Matsunaga, M.; Yin, S.; Loo, J. A.; Guo, F. Heme Is Involved in MicroRNA Processing. *Nat. Struct. Mol. Biol.* **2007**, *14* (1), 23–29. https://doi.org/10.1038/nsmb1182.
8. Sun, J.; Hoshino, H.; Takaku, K.; Nakajima, O.; Muto, A.; Suzuki, H.; Tashiro, S.; Takahashi, S.; Shibahara, S.; Alam, J.; Taketo, M. M.; Yamamoto, M.; Igarashi, K. Hemoprotein Bach1 Regulates Enhancer Availability of Heme Oxygenase-1 Gene. *EMBO J.* **2002**, *21* (19), 5216–5224. https://doi.org/10.1093/emboj/cdf516.
9. 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* (1–3), 93–98. https://doi.org/10.1016/j.febslet.2004.07.054.
10. 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.; Gray-Owen, S. D.; Chen, W.; Moraes, T. F. A Slam-Dependent Hemophore Contributes to Heme Acquisition in the Bacterial Pathogen Acinetobacter Baumannii. *Nat. Commun.* **2021**, *12* (1), 6270. https://doi.org/10.1038/s41467-021-26545-9.
11. Reedy, C. J.; Elvekrog, M. M.; Gibney, B. R. Development of a Heme Protein Structure Electrochemical Function Database.

Nucleic Acids Res. **2007**, *36* (Database), D307–D313. https://doi.org/10.1093/nar/gkm814.

12. 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. https://doi.org/10.1093/database/baaa066.
13. 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* (1), 010502. https://doi.org/10.7567/1347-4065/ab62b9.
14. Jentzen, W.; Song, X. Z.; Shelnutt, J. A. Structural Characterization of Synthetic and Protein-Bound Porphyrins in Terms of the Lowest-Frequency Normal Coordinates of the Macrocycle. *J. Phys. Chem. B* **1997**, *101* (9), 1684–1699. https://doi.org/10.1021/jp963142h.
15. 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* (25), 8536–8543. https://doi.org/10.1021/jp102135p.
16. 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. U. S. A.* **2014**, *111* (18), 6570–6575. https://doi.org/10.1073/pnas.1322274111.
17. Imada, Y.; Nakamura, H.; Takano, Y. Density Functional Study of Porphyrin Distortion Effects on Redox Potential of Heme. *J. Comput. Chem.* **2018**, *39* (3), 143–150. https://doi.org/10.1002/jcc.25058.
18. 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. https://doi.org/10.1016/j.cplett.2018.08.071.
19. Kondo, H. X.; Takano, Y. Analysis of Fluctuation in the Heme-Binding Pocket and Heme Distortion in Hemoglobin and Myoglobin. *Life* **2022**, *12* (2), 210. https://doi.org/10.3390/life12020210.
20. Li, T.; Bonkovsky, H. L.; Guo, J. Structural Analysis of Heme Proteins: Implications for Design and Prediction. *BMC Struct. Biol.* **2011**, *11* (1), 13. https://doi.org/10.1186/1472-6807-11-13.
21. 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**.
22. Sacquin-Mora, S.; Lavery, R. Investigating the Local Flexibility of Functional Residues in Hemoproteins. *Biophys. J.* **2006**, *90* (8), 2706–2717. https://doi.org/10.1529/biophysj.105.074997.
23. 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* (4), 775–784. https://doi.org/10.1021/acs.jcim.1c01315.
24. 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., 2012; Vol. 25.
25. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*; 2015. https://doi.org/https://doi.org/10.48550/arXiv.1409.1556.
26. Kinjo, A. R.; Suzuki, H.; Yamashita, R.; Ikegawa, Y.; Kudou, T.; Igarashi, R.; Kengaku, Y.; Cho, H.; Standley, D. M.; Nakagawa, A.; Nakamura, H. Protein Data Bank Japan (PDBj): Maintaining a Structural Data Archive and Resource Description Framework Format. *Nucleic Acids Res.* **2012**, *40* (D1), D453–D460. https://doi.org/10.1093/nar/gkr811.
27. 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. https://doi.org/10.1093/database/baq021.
28. Hamelryck, T.; Manderick, B. PDB File Parser and Structure Class Implemented in Python. *Bioinformatics* **2003**, *19* (17), 2308–2310. https://doi.org/10.1093/BIOINFORMATICS/BTG299.
29. 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.; De Hoon, M. J. L. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* **2009**, *25* (11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163.

30. 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* (8), 1528–1532. https://doi.org/10.1016/j.bpj.2015.08.015.
31. Wang, G.; Dunbrack, R. L. PISCES: A Protein Sequence Culling Server. *Bioinformatics* **2003**, *19* (12), 1589–1591. https://doi.org/10.1093/bioinformatics/btg224.
32. Adamo, C.; Barone, V. Toward Reliable Density Functional Methods without Adjustable Parameters: The PBE0 Model. *J. Chem. Phys.* **1999**, *110* (13), 6158–6170. https://doi.org/10.1063/1.478522.
33. Ditchfield, R.; Hehre, W. J.; Pople, J. A. Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules. *J. Chem. Phys.* **1971**, *54* (2), 724–728. https://doi.org/10.1063/1.1674902.
34. Hariharan, P. C.; Pople, J. A. The Influence of Polarization Functions on Molecular Orbital Hydrogenation Energies. *Theor. Chim. Acta* **1973**, *28* (3), 213–222. https://doi.org/10.1007/BF00533485.
35. Rassolov, V. A.; Pople, J. A.; Ratner, M. A.; Windus, T. L. 6-31G* Basis Set for Atoms K through Zn. *J. Chem. Phys.* **1998**, *109* (4), 1223–1229. https://doi.org/10.1063/1.476673.
36. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; Chintala, S. 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., 2019; pp 8024–8035.
37. Kingma, P. D.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, No. 1412.6980. https://doi.org/10.48550/arXiv.1412.6980.
38. 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* (9), 4584–4592. https://doi.org/10.1021/acs.jctc.7b00500.
39. Case, D. A.; Ben-Shalom, I. Y.; Brozell, S. .; Cerutti, D. S.; Cheatham, T. E.; Cruzeiro, III, V. W. D.; Darden, T. A.; Duke, R. E.; Ghoreishi, D.; Giambasu, G.; Giese, T.; Gilson, M. K.; Gohlke, H.; Goetz, A. W.; Greene, D.; Harris, R.; Homeyer, N.; Huang, Y.; Izadi, S.; Kovalenko, A.; Krasny, R.; Kurtzman, T.; Lee, T. S.; LeGrand, S.; Li, P.; C., L.; Liu, J.; Luchko, T.; Luo, R.; Man, V.; Mermelstein, D. J.; Merz, K. M.; Miao, Y.; Monard, G.; Nguyen, C.; Nguyen, H.; Onufriev, A.; Pan, F.; Qi, R.; Roe, D. R.; Roitberg, A.; Sagui, C.; Schott-Verdugo, S.; Shen, J.; Simmerling, C. L.; Smith, J.; Swails, J.; Walker, R. C.; Wang, J.; Wei, H.; Wilson, L.; Wolf, R. M.; Wu, X.; Xiao, L.; Xiong, Y.; York, D. M.; Kollman, P. A. AMBER 2019. University of California: San Francisco 2019.
40. Jolliffe, I. T. Principal Component Analysis, Second Edition. *Encycl. Stat. Behav. Sci.* **2002**, *30* (3), 487. https://doi.org/10.2307/1270093.
41. Li, W.; Godzik, A. Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences. *Bioinformatics* **2006**, *22* (13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158.
42. Shelnutt, J. A.; Song, X. Z.; Ma, J. G.; Jia, S. L.; Jentzen, W.; Medforth, C. J. Nonplanar Porphyrins and Their Significance in Proteins. *Chem. Soc. Rev.* **1998**. https://doi.org/10.1039/a827031z.
43. Bolognesi, M.; Onesti, S.; Gatti, G.; Coda, A.; Ascenzi, P.; Brunori, M. Aplysia Limacina Myoglobin. *J. Mol. Biol.* **1989**, *205* (3), 529–544. https://doi.org/10.1016/0022-2836(89)90224-6.
44. Li, H.; Shimizu, H.; Flinspach, M.; Jamal, J.; Yang, W.; Xian, M.; Cai, T.; Wen, E. Z.; Jia, Q.; Wang, P. G.; Poulos, T. L. The Novel Binding Mode of N -Alkyl- N ‘-Hydroxyguanidine to Neuronal Nitric Oxide Synthase Provides Mechanistic Insights into NO Biosynthesis. *Biochemistry* **2002**, *41* (47), 13868–13875. https://doi.org/10.1021/bi020417c.

45. Yao, H.; Wang, Y.; Lovell, S.; Kumar, R.; Ruvinsky, A. M.; Battaile, K. P.; Vakser, I. A.; Rivera, M. The Structure of the BfrB–Bfd Complex Reveals Protein–Protein Interactions Enabling Iron Release from Bacterioferritin. *J. Am. Chem. Soc.* **2012**, *134* (32), 13470–13481. https://doi.org/10.1021/ja305180n.
46. Hui, H. L.; Kavanaugh, J. S.; Doyle, M. L.; Wierzba, A.; Rogers, P. H.; Arnone, A.; Holt, J. M.; Ackers, G. K.; Noble, R. W. Structural and Functional Properties of Human Hemoglobins Reassembled after Synthesis in Escherichia Coli ,. *Biochemistry* **1999**, *38* (3), 1040–1049. https://doi.org/10.1021/bi981986g.
47. Kavanaugh, J. S.; Rogers, P. H.; Arnone, A. High-Resolution x-Ray Study of Deoxy Recombinant Human Hemoglobins Synthesized from .Beta.-Globins Having Mutated Amino Termini. *Biochemistry* **1992**, *31* (36), 8640–8647. https://doi.org/10.1021/bi00151a034.
48. Wang, Y.; Yao, H.; Cheng, Y.; Lovell, S.; Battaile, K. P.; Midaugh, C. R.; Rivera, M. Characterization of the Bacterioferritin/Bacterioferritin Associated Ferredoxin Protein–Protein Interaction in Solution and Determination of Binding Energy Hot Spots. *Biochemistry* **2015**, *54* (40), 6162–6175. https://doi.org/10.1021/acs.biochem.5b00937.
49. Tsukihara, T.; Shimokata, K.; Katayama, Y.; Shimada, H.; Muramoto, K.; Aoyama, H.; Mochizuki, M.; Shinzawa-Itoh, K.; Yamashita, E.; Yao, M.; Ishimura, Y.; Yoshikawa, S. The Low-Spin Heme of Cytochrome c Oxidase as the Driving Element of the Proton-Pumping Process. *Proc. Natl. Acad. Sci.* **2003**, *100* (26), 15304–15309. https://doi.org/10.1073/pnas.2635097100.
50. LaCount, M. W.; Zhang, E.; Chen, Y. P.; Han, K.; Whitton, M. M.; Lincoln, D. E.; Woodin, S. A.; Lebioda, L. The Crystal Structure and Amino Acid Sequence of Dehaloperoxidase from Amphitrite Ornata Indicate Common Ancestry with Globins. *J. Biol. Chem.* **2000**, *275* (25), 18712–18716. https://doi.org/10.1074/jbc.M001194200.
51. Chen, Z.; de Serrano, V.; Betts, L.; Franzen, S. Distal Histidine Conformational Flexibility in Dehaloperoxidase from Amphitrite Ornata. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2009**, *65* (1), 34–40. https://doi.org/10.1107/S0907444908036548.
52. Polyakov, K. M.; Boyko, K. M.; Tikhonova, T. V.; Slutsky, A.; Antipov, A. N.; Zvyagilskaya, R. A.; Popov, A. N.; Bourenkov, G. P.; Lamzin, V. S.; Popov, V. O. High-Resolution Structural Analysis of a Novel Octaheme Cytochrome c Nitrite Reductase from the Haloalkaliphilic Bacterium Thioalkalivibrio Nitratireducens. *J. Mol. Biol.* **2009**, *389* (5), 846–862. https://doi.org/10.1016/j.jmb.2009.04.037.