leveraged this resource to train a classifier only on zinc-binding sites, using a deep learning (DL) approach.[25] DL is becoming increasingly popular in structural bioinformatics, not only for the prediction of 3D protein structures[26,27] but also to support the analysis of experimental structures.[28] Two relevant recent examples are the identification of water interaction sites[29] and, even more relevant to this work, the classification of enzymatic vs non-enzymatic metal ions in proteins.[30] Our neural classifier was able to identify physiological sites for both zinc- and iron-binding sites, which are both transition metal ions, with very good accuracy. The analysis of the relative importance of the different features in driving the performance of the neural network suggested general properties of physiological MBSs.

In summary, the present work provides the community (i) with an extensive, organized dataset of annotated physiological/adventitious metal sites, which can be reused in other structural bioinformatics studies of metalloproteins, as well as (ii) with a freely available tool enabling non-experts to analyze new MBSs. Our analysis pinpointed some crucial properties defining the profile of physiological sites, which may be generally relevant at least for transition metal ions.

■ METHODS

Preparation of the Dataset. MetalPDB[1] contains information for all metalloproteins archived in the Protein Data Bank.[31] For each metalloprotein, all its MBSs are automatically extracted according to the following procedure: for each metal atom in the structure, the non-hydrogen atoms at a distance smaller than 3.0 Å are identified as its donor atoms (shown in red in Supporting Figure S1), i.e., the atoms that bind directly to the metal. The protein residues or small molecules that contain at least one donor atom are the metal ligands (shown in cyan in Supporting Figure S1) and constitute the first coordination sphere of the metal ion. The full MBS (called minimal functional site in ref 2) is obtained by including any other residue or chemical species having at least one atom within 5.0 Å from a metal ligand (shown in pink in Supporting Figure S1). MBSs describe the local structural environment around a metal ion or metal cofactor and do not depend on the overall macromolecular structure.

We started from the clusters of "equivalent sites" already available in MetalPDB.[1] Two sites are equivalent if (i) they are found in PDB chains with the same structure (based on Pfam domain composition or on the sequence identity between the two chains being ≥50%) and, after structural superposition of the PDB chains, they (ii) are superimposed with the same metal atoms in the same positions. By construction, each cluster contains MBSs with a specific metal ion, i.e., metal-substituted sites are assigned to distinct clusters. This clustering procedure, which is similar to what is done in ref 30, allows redundancy to be removed from the dataset (with the exception of proteins having multiple MBS, as described below).

We made functional annotations available through the public MetalPDB interface for several of the above sites.[2] We have been producing functional annotations of several of these sites through a manual protocol since shortly after the first release of the database in 2012.[1] At the end of 2017, the annotations available through the public MetalPDB interface covered 17% of all zinc sites and 86% of all iron sites in MetalPDB.[2] In the subsequent years, we have continued to extend the annotation of all metal sites through the same manual protocol in order to assemble the present dataset. All annotations are based on the analysis of relevant scientific literature, as follows: if the relevant article(s) describe the function of the metal (zinc or iron in this work), then the site is annotated as physiological; if no role is described for the metal but the experimental section reports that it was present in the purification or crystallization buffer, then the site is annotated as adventitious; in all other cases, the metal is annotated as "unknown role", and that site was not used in this work. We further discarded all sites with no donor atoms from the protein. The large majority of the annotations for zinc and iron sites were motivated by and carried out in specific projects or collaborations.[32−35]

First, we randomly extracted one site for each cluster containing either zinc or individual iron ions (except heme sites). Then, for each selected site, we extracted the protein chain to which the site belongs and used it as an input to PROMOTIF[36] and NACCESS[37] to respectively calculate secondary structures and solvent accessibility (without taking into account the presence of the metal ion) at the residue level. A multiple sequence alignment (MSA) was generated for each protein using HHblits v. 3[38] from the HH-suite[39] to search the UniClust30 v. 2018-8 database[40] with the parameters "-diff inf -id 99 -cov 50 -n 3". This corresponds to the first stage of the DeepMSA protocol;[41] we did not perform the additional stages of the procedure, suggested when the depth of the MSA is relatively low (for $N_f < 128$), as initial tests on a subset of sites did now show any appreciable improvement in the final performance. For each residue of the protein chain, we included the following groups of features: (i) the profile of sequence conservation (fractional occurrence of each of the 20 amino acids), where the i-th row of the position-specific frequency matrix (PSFM) derived from the MSA was used as the representation of the ith position in the sequence; (ii) absolute and relative solvent accessibility; (iii) role in the MBS (2 for metal ligands; 1 for all other MBS residues; 0 for all other residues in the protein) (iv) secondary structure (helix/sheet/turn/other). This defined a set of 29 features for each residue in the protein (Supporting Table S1), leading to an input matrix of size $L \times 29$ for each MBS, where L is the length of the protein chain harboring the site. The PSFM contains the frequency of each of the 20 amino acids at all the positions of the MSA where the majority of the sequences do not have gaps.

Note that proteins harboring more than one MBS will give rise to as many clusters as the number of MBSs. The definition of MBS in MetalPDB mandates that independent MBSs must have all distinct metal-binding protein residues. Therefore, even if a given chain may appear more than once in the dataset, the features describing the location (group i above) of its various MBSs and, consequently, the importance that the neural network gives to all the features along the sequence, will differ for the different sites.

The list of sites with their annotation is available as Supporting Information; the latter includes also the values of the features for all sites.

Neural Classifier. Figure 1 shows the building blocks of the neural architecture and how they interact with each other. The input is a sequence S comprising L data points (x_t). L is the length of the protein sequence, x_t represents the set of the n input features for the t-th residue, as described in the previous section. Thus, the input is an $n \times L$ matrix. After processing all the blocks, the classifier returns a prediction of the type of MBS that is contained in the protein structure.
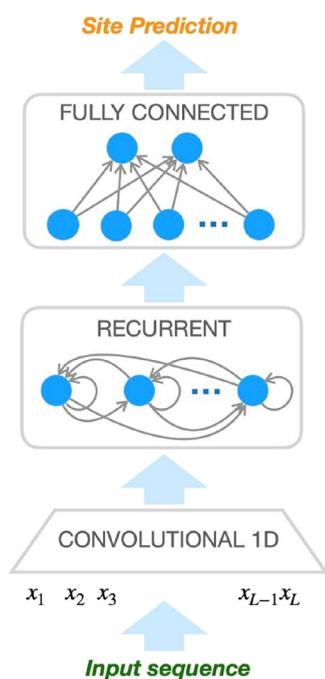
Figure 1. Scheme of the classifier. The network is composed by three modules. The convolutional module processes the input data, and its outcome is then fed to the recurrent module; finally, the fully connected module generates the estimated class probabilities for the input site.

Since we know that the role of the residues in the protein is strongly influenced by their neighborhood, in the first step, the classifier extracts information at a local level. The module Convolutional 1D generates a new representation of the sequence, S', where the value of each position is a function of its neighborhood in S. This is done by performing a convolutional operation on S using a convolutional filter of size w (k_w), corresponding to the size of the neighborhood that we take into account. Thus, $S' = \{x_t'; t = 1, ..., L\}$ where $x_t' = f_{CONV}(x_t, k_w)$. As we are in a machine learning context, this module learns to extract local information from the sequence, which is done by learning the convolutional function f_{CONV} and, more specifically, the values of the convolutional filter k_w.

A suitable network model to deal with sequential data, such as S', is the recurrent neural network (RNN). In RNNs, the model creates a representation h_t for the t-th residue of the chain, which depends on the representation generated for the preceding residue, so $h_t = f_{RNN}(x_t, h_{(t-1)})$. This implies that the representation of each residue depends on all the preceding ones. This property of the RNN makes it a suitable tool for extracting a global representation of the input sequence. The RNN is fed with the output of the convolutional 1D block, S', and generates a sequence of residue representations $h_1, h_2, ..., h_L$. In practice, we are only interested in the last representation, h_L, which constitutes the global representation of the whole sequence and, thus, is the outcome of the RNN module.

At this point, we have an array h_L representing the input sequence S, whose size does not depend on the length L any longer. h_L is the input to the final layer of the classifier (fully connected), whose aim is to associate one of the two classes to the input sequence S (now represented by h_L). This is achieved by approximating a probability distribution over the classes, given the input, as the outcome of the layer, $y =$

$[P(\text{physiological}|S), P(\text{adventitious}|S)]$. The resulting predicted class is the one having the highest probability value as computed by a linear feedforward layer with two output neurons.

All the layers described above are neural blocks, whose functions are learned as in the classical learning process using the backpropagation algorithm.[42] A complete description of the network architecture, all related parameters, hyperparameters, and training strategy is available in the Supporting Information.

Evaluation of Feature Importance. $S_i = (x_{i,1}, x_{i,2}, ..., x_{i, L})$ is the i-th row of S, i.e., the ensemble of the values of the ith feature for all the residues in the sequence. We define a perturbed version of S_i as $S_i^p = S_i + \alpha_i g$, where α_i is the magnitude of the domain of the ith feature and $g \in \mathbb{R}^L$ is an array of Gaussian noise $N(0,1)$ samples. We then define: $Y_{targ} = \{y_1, ..., y_N\}$, $Y = \{y_1, y_2, ..., y_N\}$, $Y_{noise(i)} = \{y_{1, noise(i)}, y_{2, noise(i)}, ..., y_{N, noise(i)}\}$ where Y_{targ} contains the target values, Y contains the predictions with the original input dataset (S_i), and $Y_{noise(i)}$ contains the predictions generated after the perturbation of the ith feature for all residues (S_i^p). If $Acc() \in [0,1]$ is the accuracy of the prediction, we define the importance of the ith feature as $I_i = Acc(Y_{targ}, Y) - Acc(Y_{targ}, Y_{noise(i)})$. The more sensitive the model is to variations of the ith feature, the greater is the impact on the accuracy of the predictions and, hence, the value of I_i. These statistics were calculated on the test sets of the 10-fold cross-validation procedure (Supporting Information).

■ RESULTS

Zinc Dataset and Features. We annotated one representative site for each zinc−protein family in MetalPDB.[2] These sites were taken from structures solved using any technique; for X-ray and cryoEM structures, we did not use a resolution filter. Our dataset of zinc-binding sites comprised 1944 physiological sites and 3352 adventitious sites.

Our work started from the hypothesis that it is possible to pinpoint a restricted number of key structural properties that, together with sequence information, determine the nature (physiological or adventitious) of the metal-binding site (MBS hereafter). We thus aimed at building a classifier to predict the type of MBS, and, in line with our hypothesis, understand/discover whether a sequential representation of a well-defined set of features was sufficient to successfully accomplish this task. We adopted a combination of features focused on the structural properties and features encompassing the entire protein sequence. The protein structure provides information on which protein residues constitute the first coordination sphere of the metal ion and its local environment (defined in the binding role features) and on the secondary structure and solvent accessibility of all protein residues, whereas multiple sequence alignments (MSA) quantify the conservation of the protein chain harboring the site. The MBS is a relatively small portion of the entire 3D structure, and its associated features can act as a weight for the different parts of the sequence from the binding role perspective, even though we did not explicitly instruct the classifier to do so.

To train the neural network, we used all zinc(II) sites in the dataset; the contribution of the physiological sites to the cost function of the classifier was scaled up by a 1.7 factor to account for the imbalance with respect to adventitious sites. The training involved a k-fold cross-validation[43] procedure (with k = 10) for which the dataset was divided in 10

subdatasets, commonly referred to as folds. Each fold is expected to be representative, from a data distribution perspective, of the whole dataset. Out of these 10 subdatasets, one is kept as the holdout dataset and used as the test set, one is used as the validation set, and the remaining eight folds are used to optimize the network parameters. In practice, the parameters are adjusted to optimize the classification of the validation set. In our configuration, the average sequence similarity between each fold and all other folds was 22.5% ± 0.1% (which thus corresponds to the average similarity between any given test set and the corresponding training set). After optimization, the performance of the model for a given fold selection is computed based on the classification of the test set, which has not been used until this point of the procedure and, thus, is completely new to the classifier. The procedure is repeated k times by rotating the role of all folds (training, validation, and test).

Performance of the Zinc(II) Neural Classifier. We obtained an average accuracy (fraction of all sites that were correctly identified over the 10 test sets) of 89.9% ± 1.3%. The sequence similarity of the test sets with respect to the corresponding training sets was sufficiently low to ensure that the observed performance was not biased by the occurrence of close homologues in the different groups. To understand how the model behaves on the two classes (physiological or adventitious in this case), we computed the confusion matrix (Table 1) from which a number of performance metrics can be

Table 1. Confusion Matrix of the Performances for the Test Sets Averaged over the 10-Fold Cross Validation Procedure[a]

	estimated physiological	estimated adventitious
real physiological	1615 TP	329 FN
real adventitious	208 FP	3144 TN

[a]Each row corresponds to the data points belonging to a certain class ("real" class, corresponding to physiological/adventitious zinc(II) sites in this work), whereas the columns show how the model classified the points ("estimated" class).

Table 2. Performance Metrics Derived from the Results of Table 1

metric	value	formula	meaning
PPV	0.886	TP/(TP + FP)	fraction of positive predictions that are correct
recall, TPR	0.831	TP/(TP + FN)	fraction of all positive sites that are correctly classified
NPV	0.905	TN/(TN + FN)	fraction of negative predictions that are correct
specificity, TNR	0.938	TN/(TN + FP)	fraction of all negative sites that are correctly classified
FDR	0.114	1-PPV	fraction of positive predictions that are wrong
MCC	0.780		Matthews' correlation coefficient

derived (Table 2). The MCC value, a measure of overall performance that is not particularly sensitive to the different sizes of the positive and negative datasets,[44] of 0.780 indicates that the neural classifier has a balanced performance. More in detail, our classifier appears to be slightly better at identifying adventitious than physiological sites. The predictions of

physiological sites had an 11.4% error (false discovery rate, FDR). This error rate is somewhat higher than the misannotation rate, which we estimate to be about 5% by double checking a random selection of 100 sites.

Binary classifiers assign each data point to a class based on their computed score with respect to a threshold β (typically β = 0.5 for classifiers using the 0−1 score range as done here). Thus, for different values of β, we have different associated confusion matrices. The ROC curve is the set of pairs $[1\text{-}TNR(\beta), TPR(\beta)]$ obtained by varying the threshold β.[45] In practice, the ROC curve plots TPR as a function of 1-TNR. The ideal classifier should have 1-TNR = 0 and TPR = 1 (corresponding to the top-left corner of the plot). The area under the ROC curve (AUC) quantifies the performance of the tool; the larger the AUC, which ranges from 0 to 1, the better. The average ROC curve over the 10-fold cross validation procedure (Supporting Figure S2) shows that the behavior of our classifier was very similar for all the folds and clearly distant from a random classifier. The AUC was 0.940 ± 0.006.

For any given site, the score assigned by the neural classifier to each of the two output classes can be regarded as the estimated probability that the site belongs to either class. The absolute value of the difference between the two scores is thus a measure of the imbalance with which the classifier predicts a site to be physiological or adventitious. We refer to this difference as the "confidence" of the prediction. As shown in Figure 2, the large majority of zinc(II) sites (88%) are
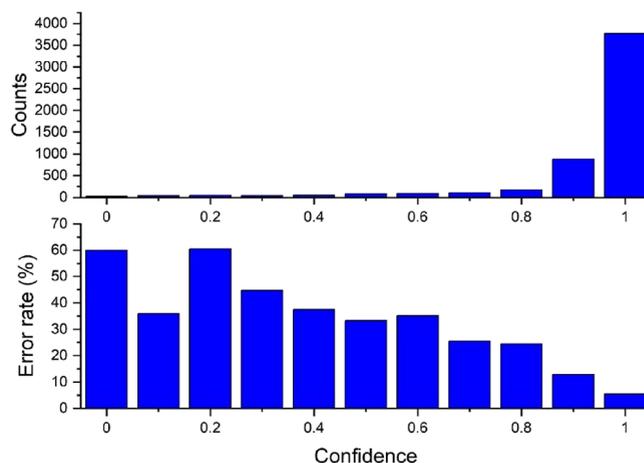
Figure 2. (Top) Number of predictions for zinc(II) sites with a given confidence (absolute value of the difference between the score of the positive and of the negative classes). (Bottom) Error rate of the neural classifier in each confidence range. The data have been computed using 0.1 bins.

classified with a confidence higher than 0.85. The error rate of the neural classifier is also strongly dependent on the confidence of the prediction. Indeed, low-confidence predictions have error rates between 30 and 60%, whereas the error rate for predictions having a confidence between 0.85 and 0.95 is 13% and the rate when the confidence is higher than 0.95 is as low as 5.4%. By grouping together all 4652 predictions with a confidence higher than 0.85, we have an error rate of 6.9%.

In our protocol, we used the information from the asymmetric unit of the crystal structure without attempting a reconstruction of the structure based on symmetry informa-

tion. Furthermore, for MBSs at a protein−protein interface, we computed the features using only a single chain even if both chains were present in the asymmetric unit. This choice was due to the difficulty of properly building deep multiple sequence alignments for pairs of interacting proteins. Nevertheless, the performance of the neural classifier for interfacial sites was marginally lower than for all other sites.

Evaluation of the Results. Since our classifier is a neural network, it is not possible to rationalize its decisions.[46,47] Nevertheless, we tried to obtain some insight by perturbing the input features in order to evaluate the corresponding impact on the predictions, thus revealing the importance of each of them from the network perspective. We did this by adding Gaussian noise to one feature at the time in the test sets (thus degrading the quality of the input to the network) and observing the decrease of classification accuracy (Figure 3). The most
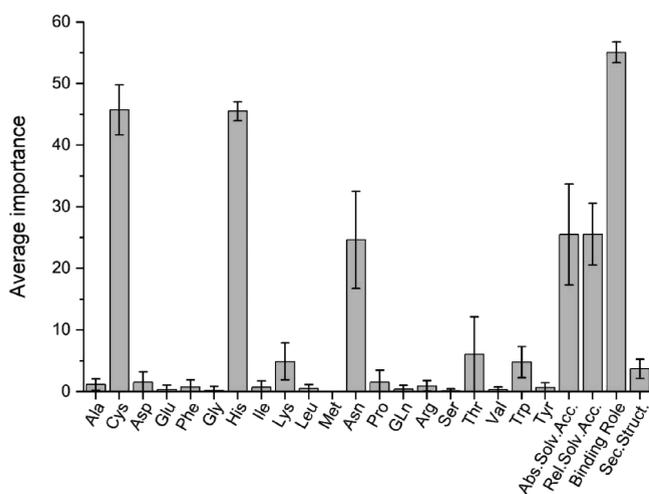
Figure 3. Importance of the input features. The plot shows the decrease in classification accuracy caused by the perturbation of the input features of the test sets, measured by the importance parameter (see Methods) averaged over the ten folds. The 20 amino acids were perturbed individually. Features describing the binding role of the residues and their secondary structure were merged.

important feature is the identification of which amino acids belong to the first and second coordination spheres of the metal (binding role). This is followed closely by information on the conservation profiles of Cys and His, essentially at the same level. Other significant impacts are those of solvent accessibility, either absolute, Abs.Solv.Acc., or relative, Rel.Solv.Acc.[37] The conservation profile of Asn has some importance. Notably, information on secondary structure elements has a negligible impact on the performance of the classifier (Figure 3).

We further analyzed the above results by training the same neural classifier with a reduced set of input features. For this experiment, we used the following groups: (i) the sequence conservation of Cys, His, Asn, and Thr (the four most relevant residues in Figure 3); (ii) the three binding role features; (iii) the two solvent accessibility features; and (iv) various combinations of (i), (ii) and (iii). The results confirmed the prominent relevance of the binding role information, closely followed by the conservation of four selected amino acids (Table 3). This analysis deemphasized the importance of solvent accessibility. However, it should be noted that we did not perform a search for the best NN architecture for each

Table 3. Average Accuracy over the Test Sets of the 10 Fold Cross-Validation Procedure for Neural Networks Trained with a Subset of the 29 Features

features	test accuracy	standard deviation
conservation of Cys, His, Asn, and Thr	80.9	1.5
binding role	84.5	1.5
solvent accessibility	64.0	4.9
conservation of C, H, N, T plus binding role	86.6	1.6
conservation of C, H, N, T plus solvent accessibility	75.7	3.3
binding role plus solvent accessibility	84.2	1.9

combination of input features, making the results of Table 3 a lower limit for the performance of the corresponding optimal network architecture.

Figure 4 shows a visualization of the input data in the form of a reduced-dimensionality plot of the data representation learned by the network (i.e., the h_L vector generated by the RNN module), colored based on the input class (Figure 4A) and predicted class (Figure 4B). The neural classifier achieved a clear separation between the predicted physiological and adventitious sites. Thus, from the machine perspective, the large majority of the points belonging to a given class are closer to each other than to the points of the other class. This separation is a very good match to the distribution of the real data (Figure 4A). This is the result of a complex interplay among all the features analyzed in the previous sections. Nevertheless, it is possible to relate the data representation learned by the network to the experimentally determined properties of the MBSs, also based on the information on feature importance. For example, by highlighting the accessibility of the sites, we observed that low-accessibility sites lie in the region that comprises the majority of the physiological sites (Figure 4C).

Application of the Zinc(II) Neural Classifier to Non-Heme Mononuclear Iron Sites. We applied the zinc(II) neural classifier to 451 non-heme mononuclear iron sites taken from MetalPDB in order to investigate whether the differences between the coordination chemistry of zinc and iron would significantly affect the performance. We did not take into account heme proteins because these cofactors have highly specific characteristics. Iron(II) and iron(III) sites were used without distinction. The results were good, with 78.6% of the physiological sites and 78.3% of the adventitious sites correctly identified (Table 4). The error rate for positive predictions was 10.8%, which is practically the same as for the zinc sites. Seventy-three percent of the predictions had a confidence value of 0.85 or more.

Experimental Determinants of the Performance of the Neural Classifier. Figure 5 shows the distribution of the features highlighted in Figure 3 for the residues of the zinc(II) MBSs in the whole dataset. First, we inspected the binding role of residues in the MBS, which was the most important group of features. It can be immediately observed that physiological and adventitious sites differ in the distributions of the number of metal-binding amino acids ("metal ligands"), as well as for their average solvent accessibility (Figure 5A,B). Adventitious sites tend to have a lower number of metal ligands, which are more solvent-exposed than in the case of physiological sites. However, the distribution of solvent accessibility values for the latter sites is relatively wide. When looking at the whole MBS, physiological sites involve a larger number of residues than
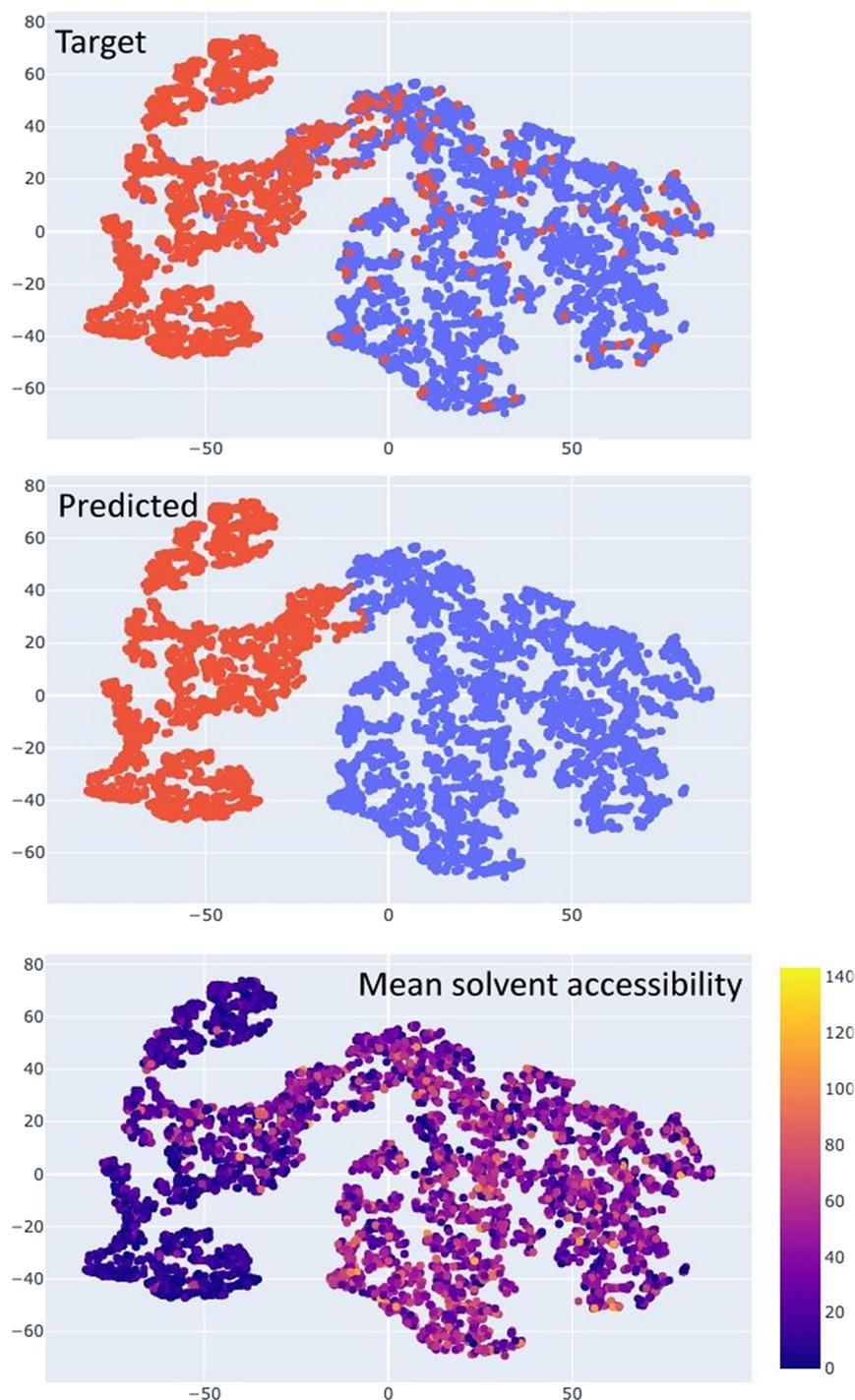
Figure 4. Data visualization generated with the TSNE dimensionality reduction algorithm. This algorithm produces a representation in an arbitrary 2D space of the distance between the points in the original multidimensional space of the data representation of the neural network. The points (red: physiological sites; blue: adventitious sites) are colored according to the (top) known class and (center) predicted class. In the bottom panel, the points are colored based on the average absolute solvent accessibility of the protein residues providing the donor atoms to the zinc(II) ion(s) regardless of their classification.

Table 4. Confusion Matrix for the Classification of Non-Heme Mononuclear Iron Sites by the Zinc(II) Neural Classifier

	estimated P	estimated N
real P	246 TP	67 FN
real N	30 FP	108 TN

adventitious sites. Per construction of the MetalPDB database, the MBS is the ensemble of the atoms in the metal ligands and any other atom belonging to a chemical species within 5 Å from a ligand;[1] the MBS describes the local 3D environment around the metal cofactor independently of the protein fold. The average number of residues in the MBS is 22.3 vs 12.5 for physiological and adventitious sites, respectively (Figure 5C). Notably, the amino acids in the second coordination sphere
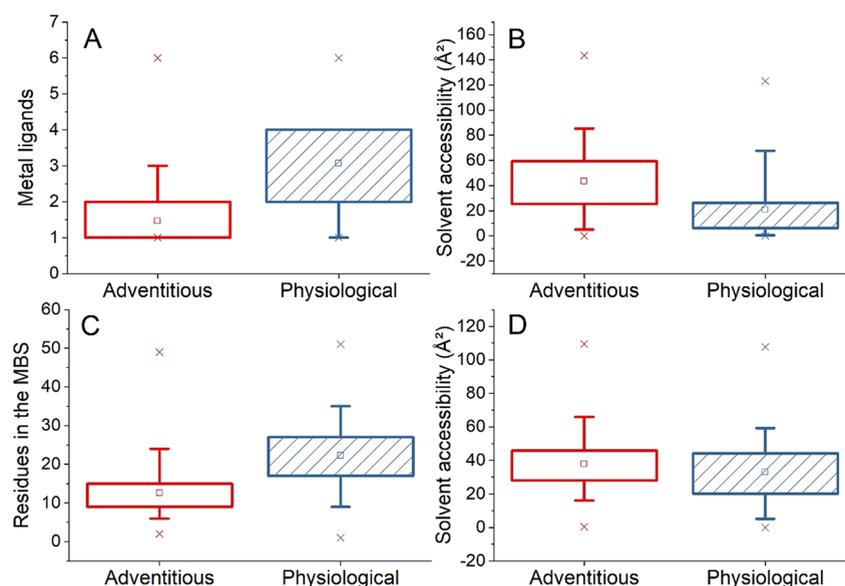
Figure 5. Comparison of value ranges (adventitious vs physiological) for a selection of the features defined for all zinc-binding sites. (A) Number of amino acids binding the metal ("metal ligands"). (B) Average absolute solvent accessibility of the metal ligands. (C) Number of residues in the site. (D) Average absolute solvent accessibility of the residues in the second coordination sphere. Red empty boxes: adventitious sites; blue hatched boxes: physiological sites. Box plot setup: the box goes from the 25th to the 75th percentile (1st and 3rd quartile, respectively); whiskers are at the 5th and 95th percentile; the minimum and maximum values are shown by crosses; the square in the box corresponds to the mean value.

tend to have the same solvent exposure in both groups of sites (Figure 5D), with values similar to those of the metal ligands in adventitious sites. As a result, the average solvent accessibility of the whole MBS is alike for both types of sites (not shown).

■ DISCUSSION

We trained a convolutional-recurrent deep neural network[48−50] to classify physiological and adventitious zinc(II)-binding sites (Supporting Figure S3) in proteins using a consolidated machine learning approach.[25] Adventitious sites are the result of the experimental conditions under which the protein sample was prepared and crystallized and are expected not to be populated in vivo. We constructed the dataset for training and testing the neural classifier using the manually annotated entries in the MetalPDB database.[2] Sequence and structure comparison methods were used to remove redundancy from the dataset. Our zinc(II) neural classifier achieved a satisfactory performance, quantified by a MCC value of 0.780, which is in line with the performance of other DL applications in structural biology,[29,30] and a AUC of 0.94. Only about 11% of the sites classified as physiological were incorrectly assigned, whereas the error rate for the classification of adventitious sites was slightly better (9.6%). The network outputs two independently estimated scores for each site, practically corresponding to the probability that the site is adventitious and the probability that the site is physiological. The absolute value of the difference between the two scores can be taken as a measure of the confidence that the classifier has in its assignment of a specific site to either class. In fact, the error rate for predictions with a confidence of 0.85 or greater is somewhat smaller than for lower confidence values (7% vs 30−50%). A lower confidence for a given data point reflects the fact that the point is located close to the surface separating physiological and adventitious sites (Figure 4), which indeed corresponds to a lower reliability of the prediction. The neural classifier featured a satisfactory performance also for non-heme

mononuclear iron sites (78% accuracy), which had not been used at any stage of its development. This is remarkable given the negligible protein sequence similarity between the two groups and the different coordination preferences of zinc(II) vs iron. For example, in our datasets, the majority of zinc(II) sites have a coordination of 4, whereas the most common coordination number for the analyzed iron sites is 6 (e.g. Supporting Figure S3D). The results on iron sites constitute both a further validation of the implemented approach and an intriguing outcome of our work.

The satisfactory performance of the neural classifier allows it to be used as a tool to validate novel metal sites binding zinc(II) or mononuclear iron ions, which are both transition metals. Beyond this achievement, by looking at how the different features affected the classification, we obtained insight on the chemical properties of physiological and adventitious sites. This process was eased by the relatively straightforward definition of the features we input to the network. Figure 3 shows that pinning down the metal ligands as well as the protein residues in the MBS constitutes the most important input to achieve a high-quality classification. Using only these features to train the network results in an accuracy of 84.5% (Table 3). From the structural viewpoint, the next most important property is the solvent accessibility of the metal ligands. At the sequence level, the conservation pattern of Cys and His residues are crucial. Information about conservation of Cys and His, along with Asp and Glu, has been extensively used for the sequence-based prediction of the occurrence of metal-binding sites.[51−54] Intriguingly, information on Asp and Glu did not have a significant impact on the performance of the neural classifier, whereas the conservation of Asn played some role. The latter finding is difficult to rationalize: the only indication we have is that Asn is about 1.6 times more common in the second sphere of physiological sites than adventitious sites. Based on the above hints, we analyzed in detail the experimental dataset and were able to define some specific trends that can constitute rules for the identification of

physiological sites. These trends support the anecdotal evidence in the scientific literature that adventitious sites occur at the protein surface.[23,24,55] A likely reason for this is the fact that the protein is fully folded before the metal ion is captured e.g., from the crystallization buffer.

In detail, the protein provides less metal ligands in adventitious sites than in physiological sites; in addition, the ligands in the former sites show higher solvent accessibility in agreement with the surface location of adventitious sites (Figure 5). Consequently, the first coordination sphere of adventitious sites is more likely to involve water molecules or other non-proteinaceous ligands and has higher local B-factors, potentially hampering the detection of all metal ligands.[23] Another noteworthy difference among physiological and adventitious sites, which is also dependent on their different location within the protein structure, is in the number of residues in the site. According to the MetalPDB protocol,[1] the MBS comprises the metal ligands together with any protein residue having at least one heavy atom within 5 Å from any atom of the ligands. Indeed, using only the information on which residues are in the MBS (binding role) allows the classifier to achieve an accuracy of nearly 85% (Table 3). As they are located deeper within the structure, physiological MBSs involve, on average, nearly twice the number of residues than adventitious MBSs.

To evaluate a possible usage scenario for our classifier, we selected 14 different zinc sites from protein structures released in 2022. For 11 of these (five physiological and six adventitious sites), we obtained predictions with a confidence greater than 0.50, which were all correct (Supporting Table S2). Whereas all the physiological sites had some sequence similarity to structures already contained in the dataset, the adventitious sites were all novel. An interesting example is that of a zinc(II) ion found in the active site of 1,2-β-mannobiose phosphorylase from *Thermoanaerobacter* sp. X-514 (PDB entry7FIS), bound to two protein residues and to the phosphate moiety of a molecule of the substrate mannose-1-phosphate.[56] Enzyme activity assays show that the zinc ion is a result of crystallization conditions and is not required for catalysis. In agreement with this, the neural classifier predicted this MBS to be adventitious.

CONCLUSIONS

We trained a deep neural network to classify zinc(II)-binding sites in the 3D structures of proteins as physiological or adventitious. In addition to achieving a very good performance for such sites, the classifier also had a remarkable accuracy for non-heme mononuclear iron sites. Using the hints provided by the analysis of feature importance, we managed to pinpoint some simple structural features that can be used as rules to distinguish physiological and adventitious sites. MBSs involving 20 protein residues or more (as computed by MetalPDB, Supporting Figure S1) are extremely likely to be physiological as well as sites with four metal ligands or more provided by the protein chain. In evaluating this aspect, one needs to be cautious of the inclusion of additional amino acids provided by sequence tags inserted, e.g., to facilitate protein purification (such as poly-His tags), which of course are not physiologically relevant. Another important parameter is the solvent accessibility of the metal ligands, although it is not practical in this case to define a reliable cut-off value because of the width of the corresponding distributions. Nevertheless, metal ligands in physiological MBSs tend to have low solvent

accessibility. The above rules should apply at least to "simple" (mononuclear) sites harboring transition metal ions. Notably, the classification of complex metal cofactors, such as polymetallic clusters or organometallic cofactors, should be more straightforward than what we accomplished here, mostly because these sites are very unlikely to assemble in the absence of specific biosynthetic systems or of finely tuned chemical conditions during sample preparation.

The present classifier is freely available to the scientific community as a stand-alone tool (see Data and Software Availability) to enable the annotation of zinc and iron sites in the newly determined 3D structures of metalloproteins. This allows researchers to keep pace with the ever-increasing throughput of structural biology projects and makes functional analysis of metal sites possible even to non-experts in bioinorganic chemistry.

■ ASSOCIATED CONTENT

ⓈI Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.2c00522.

> (Supporting Figure S1) Construction of MBSs; (Supporting Figure S2) ROC curve for the DL classification of zinc sites; (Supporting Figure S3) examples of successful predictions by the neural classifier; (Supporting Table S1) list of the 29 features representing each residue of the input sequence; (Supplementary Table S2) classification of a set of zinc sites extracted from PDB structures released in 2022; preparation and training of the neural network (PDF)

> Table listing all physiological zinc sites used in this work (XLSX)

> Table listing all adventitious zinc sites used in this work (XLSX)

> Table listing all physiological iron sites used in this work (XLSX)

> Table listing all adventitious iron sites used in this work (XLSX)

> Pickle files for physiological and adventitious sites (ZIP)

■ AUTHOR INFORMATION

Corresponding Authors

Claudia Andreini − *Consorzio Interuniversitario di Risonanze Magnetiche di Metallo Proteine, 50019 Sesto Fiorentino, Italy; Magnetic Resonance Center (CERM), University of Florence, 50019 Sesto Fiorentino, Italy; Department of Chemistry, University of Florence, 50019 Sesto Fiorentino, Italy*; Email: andreini@cerm.unifi.it

Antonio Rosato − *Consorzio Interuniversitario di Risonanze Magnetiche di Metallo Proteine, 50019 Sesto Fiorentino, Italy; Magnetic Resonance Center (CERM), University of Florence, 50019 Sesto Fiorentino, Italy; Department of Chemistry, University of Florence, 50019 Sesto Fiorentino, Italy;* ⓞ orcid.org/0000-0001-6172-0368; Email: rosato@cerm.unifi.it

Authors

Vincenzo Laveglia − *Consorzio Interuniversitario di Risonanze Magnetiche di Metallo Proteine, 50019 Sesto Fiorentino, Italy*

Andrea Giachetti − *Consorzio Interuniversitario di Risonanze Magnetiche di Metallo Proteine, 50019 Sesto Fiorentino, Italy*

Davide Sala − *Consorzio Interuniversitario di Risonanze Magnetiche di Metallo Proteine, 50019 Sesto Fiorentino, Italy; Institute for Drug Discovery, Leipzig University, 04103 Leipzig, Germany; Magnetic Resonance Center (CERM), University of Florence, 50019 Sesto Fiorentino, Italy;* orcid.org/0000-0002-3900-0011

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jcim.2c00522

Notes

The authors declare no competing financial interest.

All data are available as Supporting Information. The code for the classifier is available from "GitHub - cerm-cirmmp/ MBSDL: Metal Binding Sites Deep Learning".

■ ACKNOWLEDGMENTS

■ REFERENCES

(1) Andreini, C.; Cavallaro, G.; Lorenzini, S.; Rosato, A. MetalPDB: a database of metal sites in biological macromolecular structures. *Nucleic Acids Res.* **2013**, *41*, D312−D319.

(2) Putignano, V.; Rosato, A.; Banci, L.; Andreini, C. MetalPDB in 2018: a database of metal sites in biological macromolecular structures. *Nucleic Acids Res.* **2018**, *46*, D459−D464.

(3) Andreini, C.; Bertini, I.; Cavallaro, G.; Holliday, G. L.; Thornton, J. M. Metal-MACiE: a database of metals involved in biological catalysis. *Bioinformatics* **2009**, *25*, 2088−2089.

(4) Waldron, K. J.; Rutherford, J. C.; Ford, D.; Robinson, N. J. Metalloproteins and metal sensing. *Nature* **2009**, *460*, 823−830.

(5) Bertini, I.; Gray, H. B.; Stiefel, E. I.; Valentine, J. S. *Biological Inorganic Chemistry*; University Science Books, 2006.

(6) Frausto da Silva, J. J. R.; Williams, R. J. P. *The biological chemistry of the elements: the inorganic chemistry of life*; Oxford University Press, 2001.

(7) Valasatava, Y.; Rosato, A.; Furnham, N.; Thornton, J. M.; Andreini, C. To what extent do structural changes in catalytic metal sites affect enzyme function? *J. Inorg. Biochem.* **2018**, *179*, 40−53.

(8) Ben-David, M.; Soskine, M.; Dubovetskyi, A.; Cherukuri, K.-P.; Dym, O.; Sussman, J. L.; Liao, Q.; Szeler, K.; Kamerlin, S. C. L.; Tawfik, D. S. Enzyme Evolution: An Epistatic Ratchet versus a Smooth Reversible Transition. *Mol. Biol. Evol.* **2019**, *37*, 1133−1147.

(9) Berman, H.; Henrick, K.; Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.* **2003**, *10*, 980.

(10) Handing, K. B.; Niedzialkowska, E.; Shabalin, I. G.; Kuhn, M. L.; Zheng, H.; Minor, W. Characterizing metal-binding sites in proteins with X-ray crystallography. *Nat. Protoc.* **2018**, *13*, 1062−1090.

(11) Grime, G. W.; Zeldin, O. B.; Snell, M. E.; Lowe, E. D.; Hunt, J. F.; Montelione, G. T.; Tong, L.; Snell, E. H.; Garman, E. F. High-Throughput PIXE as an Essential Quantitative Assay for Accurate Metalloprotein Structural Analysis: Development and Application. *J. Am. Chem. Soc.* **2020**, *142*, 185−197.

(12) Zheng, H.; Chordia, M. D.; Cooper, D. R.; Chruszcz, M.; Muller, P.; Sheldrick, G. M.; Minor, W. Validation of metal-binding sites in macromolecular structures with the CheckMyMetal web server. *Nat. Protoc.* **2014**, *9*, 156−170.

(13) Touw, W. G.; van Beusekom, B.; Evers, J. M. G.; Vriend, G.; Joosten, R. P. Validation and correction of Zn-CysxHisy complexes. *Acta Crystallogr., Sect. D.* **2016**, *72*, 1110−1118.

(14) Yao, S.; Flight, R. M.; Rouchka, E. C.; Moseley, H. N. B. Aberrant coordination geometries discovered in the most abundant metalloproteins. *Proteins: Struct., Funct., Bioinf.* **2017**, *85*, 885−907.

(15) Zheng, H.; Cooper, D. R.; Porebski, P. J.; Shabalin, I. G.; Handing, K. B.; Minor, W. CheckMyMetal: a macromolecular metal-binding validation tool. *Acta Crystallogr., Sect. D.* **2017**, *73*, 223−233.

(16) Alberts, I. L.; Nadassy, K.; Wodak, S. J. Analysis of zinc binding sites in protein crystal structures. *Protein Sci.* **1998**, *7*, 1700−1716.

(17) Rulisek, L.; Vondrasek, J. Coordination geometries of selected transition metal ions (Co2+, Ni2+, Cu2+, Zn2+, Cd2+, and Hg2+) in metalloproteins. *J. Inorg. Biochem.* **1998**, *71*, 115−127.

(18) Tamames, B.; Sousa, S. F.; Tamames, J.; Fernandes, P. A.; Ramos, M. J. Analysis of zinc-ligand bond lengths in metalloproteins: Trends and patterns. *Proteins: Struct., Funct., Bioinf.* **2007**, *69*, 466−475.

(19) Maret, W.; Li, Y. Coordination dynamics of zinc in proteins. *Chem. Rev.* **2009**, *109*, 4682−4707.

(20) Zheng, H.; Chruszcz, M.; Lasota, P.; Lebioda, L.; Minor, W. Data mining of metal ion environments present in protein structures. *J. Inorg. Biochem.* **2008**, *102*, 1765−1776.

(21) Abriata, L. A. Analysis of copper-ligand bond lengths in X-ray structures of different types of copper sites in proteins. *Acta Crystallogr., Sect. D* **2012**, *68*, 1223−1231.

(22) Zheng, H.; Langner, K. M.; Shields, G. P.; Hou, J.; Kowiel, M.; Allen, F. H.; Murshudov, G.; Minor, W. Data mining of iron(II) and iron(III) bond-valence parameters, and their relevance for macromolecular crystallography. *Acta Crystallogr., Sect. D.* **2017**, *73*, 316−325.

(23) Laitaoja, M.; Valjakka, J.; Janis, J. Zinc coordination spheres in protein structures. *Inorg. Chem.* **2013**, *52*, 10983−10991.

(24) Yao, S.; Moseley, H. N. B. Finding High-Quality Metal Ion-Centric Regions Across the Worldwide Protein Data Bank. *Molecules* **2019**, *24*, 3179.

(25) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436−444.

(26) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583−589.

(27) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871−876.

(28) Bonk, B. M.; Weis, J. W.; Tidor, B. Machine Learning Identifies Chemical Characteristics That Promote Enzyme Catalysis. *J. Am. Chem. Soc.* **2019**, *141*, 4108−4118.

(29) Zaucha, J.; Softley, C. A.; Sattler, M.; Frishman, D.; Popowicz, G. M. Deep learning model predicts water interaction sites on the surface of proteins using limited-resolution data. *Chem. Commun.* **2020**, *56*, 15454−15457.

(30) Feehan, R.; Franklin, M. W.; Slusky, J. S. G. Machine learning differentiates enzymatic and non-enzymatic metals in proteins. *Nat. Commun.* **2021**, *12*, 3712.

(31) Rose, P. W.; Prlic, A.; Altunkaya, A.; Bi, C.; Bradley, A. R.; Christie, C. H.; Costanzo, L. D.; Duarte, J. M.; Dutta, S.; Feng, Z.; et al. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.* **2017**, *45*, D271−D281.

(32) Holliday, G. L.; Andreini, C.; Fischer, C.; Rahman, S. A.; Almonacid, D. E.; Williams, S. T.; Pearson, W. R. MACiE: exploring the diversity of biochemical reactions. *Nucleic Acids Res.* **2012**, *40*, D783−D789.

(33) Buracco, S.; Peracino, B.; Andreini, C.; Bracco, E.; Bozzaro, S. Differential Effects of Iron, Zinc, and Copper on Dictyostelium discoideum Cell Growth and Resistance to Legionella pneumophila. *Front. Cell. Infect. Microbiol.* **2018**, *7*, 536.

(34) Wang, Y.; Weisenhorn, E.; MacDiarmid, C. W.; Andreini, C.; Bucci, M.; Taggart, J.; Banci, L.; Russell, J.; Coon, J. J.; Eide, D. J. The

Journal of Chemical Information and Modeling
pubs.acs.org/jcim
Article

cellular economy of the Saccharomyces cerevisiae zinc proteome. *Metallomics* **2018**, *10*, 1755−1776.

(35) Wang, J.; Lonergan, Z. R.; Gonzalez-Gutierrez, G.; Nairn, B. L.; Maxwell, C. N.; Zhang, Y.; Andreini, C.; Karty, J. A.; Chazin, W. J.; Trinidad, J. C.; et al. Multi-metal Restriction by Calprotectin Impacts De Novo Flavin Biosynthesis in Acinetobacter baumannii. *Cell Chem. Biol.* **2019**, *26*, 745−755.e747.

(36) Hutchinson, E. G.; Thornton, J. M. PROMOTIF–a program to identify and analyze structural motifs in proteins. *Protein Sci.* **1996**, *5*, 212−220.

(37) NACCESS; *Department of Biochemistry and Molecular Biology*; University College, London., 1993. (accessed.

(38) Remmert, M.; Biegert, A.; Hauser, A.; Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **2012**, *9*, 173−175.

(39) Steinegger, M.; Meier, M.; Mirdita, M.; Vöhringer, H.; Haunsberger, S. J.; Söding, J. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinf.* **2019**, *20*, 473.

(40) Mirdita, M.; von den Driesch, L.; Galiez, C.; Martin, M. J.; Söding, J.; Steinegger, M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.* **2017**, *45*, D170−D176.

(41) Zhang, C.; Zheng, W.; Mortuza, S. M.; Li, Y.; Zhang, Y. DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics* **2020**, *36*, 2105−2112.

(42) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533−536.

(43) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer, 2009.

(44) Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**, *405*, 442−451.

(45) Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861−874.

(46) Ribeiro, M. T.; Singh, S.; Guestrin, C. ″Why Should I Trust You?″: Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, USA; 2016.

(47) Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* **2018**, *51*, 93.

(48) Elman, J. L. Finding Structure in Time. *Cognit. Sci.* **1990**, *14*, 179−211.

(49) Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278−2324.

(50) Cho, K.; Merrienboer, B. v.; Gülçehre, Ç.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder Decoder for Statistical Machine Translation. In *EMNLP*, 2014.

(51) Zhao, W.; Xu, M.; Liang, Z.; Ding, B.; Niu, L.; Liu, H.; Teng, M. Structure-based de novo prediction of zinc-binding sites in proteins of unknown function. *Bioinformatics* **2011**, *27*, 1262−1268.

(52) Levy, R.; Edelman, M.; Sobolev, V. Prediction of 3D metal binding sites from translated gene sequences based on remote-homology templates. *Proteins: Struct., Funct., Bioinf.* **2009**, *76*, 365−374.

(53) Haberal, İ.; Oğul, H. Prediction of Protein Metal Binding Sites Using Deep Neural Networks. *Mol. Inf.* **2019**, *38*, No. e1800169.

(54) Sanchez-Aparicio, J. E.; Tiessler-Sala, L.; Velasco-Carneros, L.; Roldan-Martin, L.; Sciortino, G.; Marechal, J. D. BioMetAll: Identifying Metal-Binding Sites in Proteins from Backbone Preorganization. *J. Chem. Inf. Model.* **2021**, *61*, 311−323.

(55) Tamames, J. A.; Ramos, M. J. Metals in proteins: cluster analysis studies. *J. Mol. Model.* **2011**, *17*, 429−442.

(56) Dai, L.; Chang, Z.; Yang, J.; Liu, W.; Yang, Y.; Chen, C.-C.; Zhang, L.; Huang, J.-W.; Sun, Y.; Guo, R.-T. Structural investigation of a thermostable 1,2-β-mannobiose phosphorylase from Thermoanaerobacter sp. X-514. *Biochem. Biophys. Res. Commun.* **2021**, *579*, 54−61.