Application of machine learning in the quantitative analysis of surface characteristics of highly abundant cytoplasmic proteins: Toward AI-based biomimetics

Jooa Moon , Guanghao Hu , Tomohiro Hayashi *

Posted Date: 15 December 2023

doi: 10.20944/preprints202312.1187.v1

Keywords: bioinformatics; machine learning; protein surfaces; surface engineering

Article

Application of machine learning in the quantitative analysis of surface characteristics of highly abundant cytoplasmic proteins: Toward AI-based biomimetics

Jooa Moon , Guanghao Hu , Tomohiro Hayashi *

Posted Date: 15 December 2023

doi: 10.20944/preprints202312.1187.v1

Keywords: bioinformatics; machine learning; protein surfaces; surface engineering

Article

Application of Machine Learning in the Quantitative Analysis of Surface Characteristics of Highly Abundant Cytoplasmic Proteins: Toward AI-Based Biomimetics

Joo A Moon [1], Guanghao Hu [1], and Tomohiro Hayashi [1,2,*]

[1] Department of Materials Science and Engineering, School of Materials and Chemical Technology, Tokyo Institute of Technology, 4259 Nagatsuta-cho, Midori-ku, Yokohama, Kanagawa 226-8502, Japan

[2] The Institute for Solid State Physics, The University of Tokyo, 5-1-5, Kashiwanoha, Kashiwa, Chiba 277-0882, Japan

Abstract: Proteins in the crowded environment of human cells have often been studied regarding nonspecific interactions, misfolding, and aggregation, which may cause cellular malfunction and disease. Specifically, proteins with high abundance are more susceptible to these issues due to the law of mass action. Therefore, the surfaces of highly abundant cytoplasmic (HAC) proteins directly exposed to the environment can exhibit specific physicochemical, structural, and geometrical characteristics that reduce nonspecific interactions and adapt to the environment. However, the quantitative relationships between the overall surface descriptors still need clarification. Here, we used machine learning to identify HAC proteins using hydrophobicity, charge, roughness, secondary structures, and B-factor from the protein surfaces and quantify the contribution of each descriptor. First, several supervised learning algorithms were compared to solve binary classification problems for the surfaces of HAC and extracellular proteins. Then, logistic regression was adopted for the feature importance analysis of descriptors considering model performance (80.2% accuracy and 87.6% AUC) and interpretability. The HAC proteins showed positive correlations with negatively and positively charged areas but negative correlations with hydrophobicity, B-factor, beta structure proportion, roughness, and disordered regions proportion. Finally, the details of each descriptor could be explained concerning adaptative surface strategies of HAC proteins to regulate nonspecific interactions, protein folding, flexibility, stability, and adsorption. This study presented a novel approach using various surface descriptors to identify HAC proteins and provided quantitative design rules for the surfaces well-suited to human cellular crowded environments.

Keywords: bioinformatics; machine learning; protein surfaces; surface engineering

1. Introduction

The intracellular space of living organisms is highly crowded with macromolecules, which can occupy up to nearly one-third of the entire cellular volume.[1] The resulting highly crowded environment poses challenges of nonspecific interactions, critically influencing issues such as protein folding, stability, and adsorption.[2-4] In human cells, these issues are specifically crucial since the intracellular proteins that fail to fold correctly into their native shapes tend to aggregate and cause cellular malfunction and death, resulting in detrimental pathological consequences.[5] In particular, cytoplasmic proteins with high abundance, i.e., highly expressed proteins, are more likely to encounter nonspecific interactions due to the law of mass action.[6] Thus, highly abundant cytoplasmic (HAC) proteins must exhibit certain physicochemical, structural, and geometrical characteristics to adapt to the environment and mitigate the issues. Eventually, intracellular proteins, especially highly abundant ones, are expected to share particular characteristics differentiated from

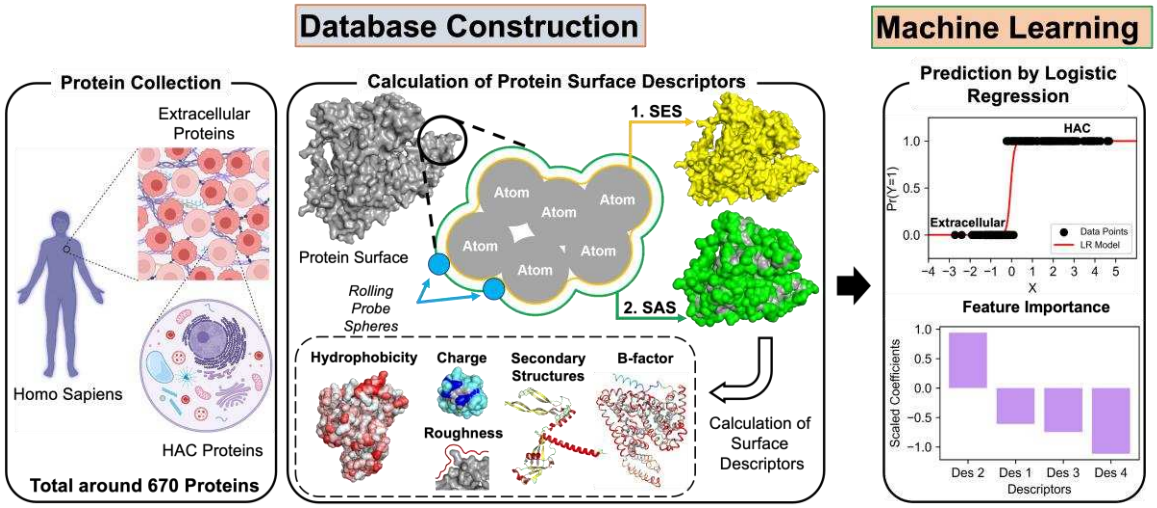
 doi:10.20944/preprints202312.1187.v1

extracellular proteins, which often experience less crowded environments,[4,7] to ensure proper cellular function in such a highly crowded environment.

Previously, computational approaches have aided in the characterization of intracellular proteomes, with various techniques targeting different regions of proteins, including global regions,[8] surface regions,[9,10] or both regions[11]. Notably, the surface regions of proteins are essential for studying protein characteristics since the regions are directly exposed to the external environment and potential partners and thus reflect various properties.[4,10] While there have been several works on using the frequency of surface residues, [9,10] there is a lack of research revealing quantitative relationships of specific physicochemical, structural, and geometrical descriptors, which can have different scales for characterizing the surfaces of the HAC protein.

To address this issue, we use interpretable machine learning (ML)-based approach to characterize the surfaces of HAC proteins by quantifying the contribution of the surface descriptors. Over the past few decades, ML techniques have been increasingly applied to predict protein-protein interaction,[12] protein-ligand molecular docking,[13] protein subcellular localization,[14] and protein 3D structure prediction.[15] Despite significant advances in these areas, identifying protein surface characteristics using only a few representative physicochemical, structural, and geometrical descriptors remains challenging. This is the first study focusing on this specific task, revealing quantitative relationships between the surface descriptors. By understanding the surface rules of HAC proteins in human cells through interpretable ML, this study will enable the development of efficient drug delivery systems, such as deepening our knowledge of the interactions between therapeutic nanoparticles and proteins [16]

In this study, we aim to distinguish the surfaces of HAC proteins from those of extracellular proteins using binary classification algorithms. We extracted surface physicochemical, structural, and geometrical descriptors from protein surfaces to build a database and apply ML (Figure 1). As a first step of the database construction, we collected around 330 3D protein structures each for human HAC and extracellular proteins. Then, various descriptors on protein surfaces, such as hydrophobicity, charged area, roughness, B-factor, and the proportions of protein structures, were calculated for the collected 3D protein structures. Then, several supervised ML algorithms: K-Nearest Neighbor (KNN), Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVM) were used to solve the binary classification of extracellular and HAC proteins. Based on excellent performance and high model interpretability, we selected the LR algorithm to explain the importance of each descriptor quantitatively. Namely, this study answers the following questions: (1) Can surface characteristics of HAC proteins be identified with several physicochemical, structural, and geometrical descriptors? and (2) Which descriptor contributes to the crowded environment-adaptive surface in human cells to what extent? The LR model used in our study enabled the identification of HAC proteins, and coefficients from LR represented the importance of each descriptor.



 doi:10.20944/preprints202312.1187.v1

Figure 1. Schematic representation of key processes in functional prediction and quantitative analysis of surface physicochemical, structural, and geometrical descriptors on protein surface. HAC: Highly abundant cytoplasmic; SES: Solvent-excluded surface; SAS: Solvent accessible surface.

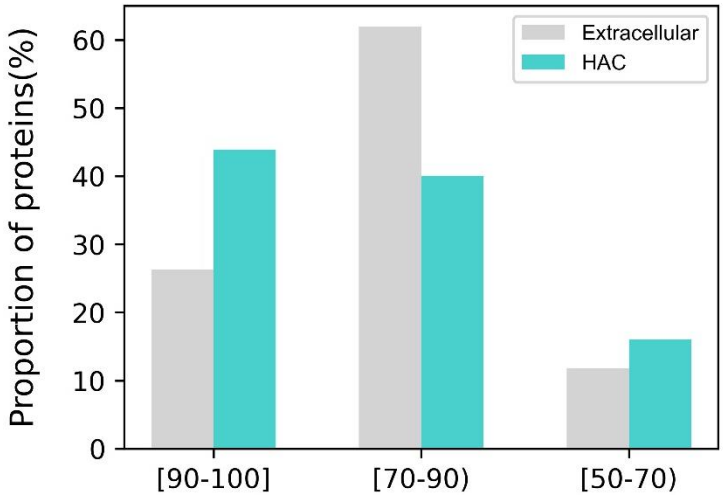
2. Methodology

2.1. Protein Sample Collection

The datasets consist of two types of human proteins: human cytoplasmic proteins with high abundance and extracellular proteins. First, we collected cytoplasmic proteins with the highest abundance level from the PaxDb database, a collection of experimental data on protein abundance.[17] The cytoplasmic proteins that were also tagged extracellular keywords (e.g., secreted, extracellular matrix, and extracellular space) in Uniprot were eliminated. Then, proteins in extracellular environments determined with experimental assay were collected (GO ID: 5615).[18] Finally, 331 human extracellular proteins and 337 HAC proteins within the sequence length range of 100 to 700 were collected for analysis.

The 3D structures of a total of 668 proteins were collected through the Alphafold ver2.0 (Alphafold2) (https://alphafold.ebi.ac.uk/) protein structure prediction model.[15,19] Alphafold2 3D models provide entire protein structures, allowing for comprehensive surface analysis, unlike partial structures often found in experimental Protein Data Bank (PDB) files from X-ray crystallography. Alphafold2 is known to be the top-ranked prediction model with a median global distance test score of 92.4 across all targets and 87.0 on the challenging free modeling category in the 14th CASP assessment (https://predictioncenter.org/casp14/zscores_final.cgi). Additionally, in most cases, Alphafold2's structural prediction accuracy has reached experimental accuracy.[15]

Even though the overall predictability of Alphafold2 is exceptional, not all predicted structures are suitable for the analysis. Every residue from Alphafold2 3D protein structure is given with per-residue metric, which reflects the structural model confidence called predicted local distance difference test (pLDDT), scaling from 0 to 100. The pLDDT evaluates how well the predicted model agrees with experimental data using local distance difference test Cα.[20] pLDDT > 90 is considered a high-accuracy cut-off, and pLDDT> 70 can be regarded as generally correct backbone prediction.[21] When the pLDDT is lower than 50, the predicted region is expected to be intrinsically disordered.[22] However, a low pLDDT score in Alphafold2 results from high residue flexibility and dynamic structure rather than 'low confidence.'[23] Also, since disordered regions of proteins are involved in molecular recognition and hydrophobic interactions, it is essential to include the regions for the analysis.[24] Considering the potential interpretability difficulty from intrinsically disordered proteins, we set our cut-off value as average pLDDT>50 for a whole protein structure. Finally, we ensured that over 80% of extracellular and HAC proteins had average pLDDT values over 70 (Figure 2).



 doi:10.20944/preprints202312.1187.v1

Figure 2. Average predicted local distance difference test value distribution of collected 668 proteins.

2.2. Calculation of Surface Descriptors

Previous studies have introduced several definitions of protein surfaces, each with different characteristics. Among them, we adopted solvent-accessible Surface (SAS) and solvent-excluded surface (SES) for calculating other descriptors (Figure 1).[25] The SAS was calculated by rolling probe spheres that have an equivalent size with water molecules. We used SAS for the residue-based analysis: we assumed that a specific residue in a protein could have a maximum SAS when its neighboring amino acids are Glycines. (i.e., having Gly-residue-Gly structure) When the proportion of an actual SAS for a residue to a maximum SAS was higher than or equal to 30%, the residue was defined as surface residue. Another protein surface used in the analysis is SES, also called the Connolly surface.[26] The surface moves inward from the SAS by a distance identical to the probe sphere radius (Figure 1). Lewis et al. uncovered that this continuous and functional surface is particularly useful in calculating the protein surface roughness. Then, protein surface descriptors representing various physicochemical, structural, and geometrical descriptors were calculated (Table 1) based on the two surface types. All the descriptors were computed using Python 3.9.12.

Table 1. Descriptors used in this work to explain protein surface characteristics.

Category	Variables (Descriptors)	Definition	Analyzed Surface
Hydrophobicity	*s_phobic_avg*	Average Surface Hydrophobicity	Solvent accessible surface
Charge	*s_pos_area*	Fraction of Positively Charged Surface Area	
	s_neg_area	Fraction of Negatively Charged Surface Area	
	s_charge_avg	Fraction of Total Charged Surface Area	
Protein Structure	*s_ah*	Proportion of Surface Alpha Helix	
	s_bs	Proportion of Surface Beta Structures	
	s_do	Proportion of Surface Disordered Regions	
	s_sf	Structure Surface Exposure Degree	
Flexibility	*norm_s_b*	Average Normalized Surface B-factors	
Geometry	*FD*	Average Protein Surface Roughness	Solvent excluded surface

Surface hydrophobicity, charge, secondary structures, and overall morphology of proteins are critical parameters for protein structures. The normalized consensus hydrophobicity scale

 doi:10.20944/preprints202312.1187.v1

quantitatively measured the average protein surface hydrophobicity.[27] Surface charge-related descriptors were collected by calculating the fraction of SAS of negatively charged and positively charged amino acids under physiological conditions (pH=7). Each surface amino acid contributing to the secondary structure was directly extracted by Pymol (http://www.pymol.org) to calculate the surface proportion of each secondary structure. The surface exposure degree was defined by the SAS divided by the volume of protein.

The B-factor, which is also called the Debye-Waller factor, indicates the thermal motion-induced attenuation of X-ray scattering or coherent neutron scattering.[28,29] Equation (1) defines B-factor:

, (1)

where u (Å) denotes the mean displacement of a scattering center. The B-factor is used to interpret properties such as thermostability, flexibility, internal motion, and binding of proteins.[30-34] In Alphafold2 models, the B-factor columns are replaced by pLDDT values, which can provide insights into structural flexibility.[23] We converted pLDDT values into pseudo B-factors since pLDDT values and original B-factor show a reverse relationship. The pLDDT values were first converted into root-mean-square deviation (RMSD) using the following empirical formula (Equation (2)):

$$\Delta = 1.5\exp[4(0.5 - pLDDT)], \quad (2)$$

where Δ denotes error estimates. pLDDT values were transformed into the scale of 0-1 from 0-100.[35-37] Then, the converted pseudo B-factor is expressed as Equation (3) after substituting the converted error estimates into eq 1, considering the root-mean-square positional variation in three dimensions.

$$B = \frac{8\pi^2\Delta^2}{3} \quad (3)$$

The converted pseudo-B-factors were calculated for each residue in proteins. However, in the case of X-ray analysis, low resolution leads to high B-factors around 100-200, and such high values of B-factors are not recommended to make specific conclusions.[38] Therefore, only surface residues with RMSD smaller than or equal to 1.5 (almost equivalent to $B \leq 60$) were included for analyzing surface B-factors. Finally, B-factors were normalized as Equation (4) since the non-normalized B-factor does not represent an absolute quantity, therefore, cannot be used to compare different protein structures.[39]

$$B_{norm} = \frac{B - \langle B \rangle}{\sigma} \quad (4)$$

$\langle B \rangle$ denotes the average B-factor in the whole protein structure, and σ implies the standard deviation. Then, the mean value of normalized surface B-factors in a protein was used to characterize the protein surface.

Surface roughness, which can be quantitatively characterized by the fractal dimension (*FD*), was calculated to identify the surface structural irregularity (Equation (5)).[26]

$$FD = 2 - \frac{dlog(A_s)}{dlog(R)} \quad (5)$$

A_s and R represent the molecular surface area and rolling probe radius, respectively. *FD* falls within the range of 2 to 3, having the smoothest surface on 2 and having the roughest surface on 3. As for the calculation of A_s, we calculated SES using 3V calculator (http://3vee.molmovdb.org).[40] Then, Equation (5) was transformed into Equation (6) for the convenience of calculation.

$$D_i = \frac{(\log A_{ses})_i - (\log A_{ses})_{i-1}}{(\log R)_i - (\log R)_{i-1}}, \; FD = \frac{1}{N}\sum_{n=1}^{N} D_n \quad (6)$$

i refers to a probe radius starting from 1.2, in the range of 1.0 to 3.6, with the interval of 0.2 (1.0,1.2,1.4, 1.6 … 3.6, N (Number of sets)=13). i-1 refers to the previous step of i (i-1 starts from 1.0). $(log\, A_{ses})_i$ indicates the log value of solvent excluded surface area under the probe radius i. The range of probe radius is suitable for the analysis since the probe sizes are sensitive to specific interactions between residues, reflecting the size of water molecules and side chains.[26] Finally, the mean value of all the calculated D_i represents the *FD*.

 doi:10.20944/preprints202312.1187.v1

2.3. Application of Machine Learning

The logistic Regression (LR) model, a regression model for binary classification problems, shows its chief advantage in providing high model interpretability. An odds ratio of each independent variable enables quantitatively evaluating its contribution to dependent variables. Surface descriptors were given as independent continuous variables, and <HAC:1, Extracellular:0> tags were provided as dependent dichotomous variables in the models. Then, Equation (7) represents the probability of being HAC proteins under the given independent variables[41]:

$$P(y = 1|x_1, x2, \dots, xi) = \frac{exp[f(X_i,\beta_i)]}{1+exp[f(X_i,\beta_i)]} = \frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_iX_i}}{1+e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_iX_i}} \quad (7)$$

P, x_i, β_i denote the probability of being a HAC protein, a surface descriptor, and an accompanying beta coefficient. LR uses a method of maximum likelihood to estimate β_i and the odds ratio corresponds to $exp[\beta_i]$. Then, logistic transformation, which converts the non-linear relationship into the original linear regression equation, is applied as eq 8.

$$ln\ P\ = ln\frac{exp[f(X_i,\beta_i)]}{1+exp[f(X_i,\beta_i)]} = ln\frac{P}{1-P} = \beta_0 + \sum \beta_i X_i = \beta_0 + \beta_1X_1 + \beta_2X_2 + \cdots + \beta_iX_i \quad (8)$$

A positive β_i indicates that the increase in x_i leads to the stochastic increase in the probability of being HAC proteins. Conversely, A negative β_i means that an increase in x_i results in the stochastic decrease for the probability of being HAC proteins.

As a parametric model, LR requires several statistical assumptions to perform well.[41] Thus, several data preprocessing steps were conducted, including checking the multicollinearity of surface descriptors, deletion of strongly influential outliers, and data scaling to meet the assumptions and enhance the model performance. Pearson correlation (PC) analysis, a statistical test that measures the linear association between two variables, was conducted to limit the multicollinearity problem. Also, Cook's distance from the statsmodels module in Python was calculated for leverage and residual values analysis. Conclusively, 1.03% of proteins turned out to be highly influential and outliers simultaneously and were eliminated from the dataset. Finally, the surface descriptors were standardized with StandardScaler function in Python sci-kit learn library for data scaling.

Upon constructing the LR model, several popular supervised learning algorithms for classifications: K-Nearest Neighbor (KNN), Random Forest (RF), and Support Vector Machine (SVM), were used to compare the performance of different models. All the algorithms were performed through Scikit-learn Package in Python 3.9.12. The hyperparameters for each algorithm were optimized using GridSearch cross-validation (CV), where every parameter combination is tested to evaluate ML models. 5-fold cross-validation was used to avoid overfitting to the test set. Before constructing machine learning models, datasets were randomly divided into a training set (80%) and a test set (20%), maintaining the original ratio of the target class. Then, the performance of different models was assessed by the predictive indicators: the classification accuracy and the area under the curve of receiver operating characteristic (AUC-ROC) curve. We repeated five times randomly splitting training and test sets to avoid sampling bias and overfitting, then reported the mean accuracy of each model. We selected the final ML model, LR, for the feature importance analysis considering high accuracy and model interpretability. Finally, each descriptor's significance and importance were explained with statistical analysis.

4. Results and Discussion

4.1. Pearson Correlation (PC) Analysis

First, PC analysis for all the descriptors in the train set was conducted before applying machine learning. Table 2 shows PC coefficients between independent variables, i.e., surface descriptors and dependent variables (where HAC is tagged as 1 and extracellular as 0). A PC coefficient ranges from -1 to 1, showing a perfectly negative correlation at -1 and a perfectly positive correlation at 1. A PC coefficient of 0 represents the absence of linear correlation. As a result, all the relationships between

 doi:10.20944/preprints202312.1187.v1

each surface descriptor and dependent variables were significant at 0.05 ($p<0.05$) except for structure surface exposure degree (*s_sf*) (Table 2).

Table 2. Pearson Correlation (PC) Coefficients between independent and dependent variables from a train set.

Surface Descriptors		PC Coefficients
Hydrophobicity	*s_phobic_avg*	-0.472 **
Charge	*s_pos_area*	0.401 **
	s_neg_area	0.239 **
	s_charge_avg	0.142 **
Protein Structures	*s_ah*	0.206 **
	s_bs	-0.228 **
	s_do	-0.102 *
	s_sf	-0.023
Flexbility	*norm_s_b*	-0.225 **
Geometry	*FD*	-0.106 *

** p-value<0.01 * p-value<0.05.

As shown in Figure 3, two descriptors, surface alpha helix proportion (*s_ah*) and the proportion of total charged surface area (*s_charge_avg*), were highly linearly correlated to the descriptors in their categories: protein structures and charge, respectively. Therefore, the descriptors were eliminated from the descriptor pool, considering that they showed the highest linear correlation with other descriptors in their category. According to the above results, we excluded three descriptors using PC analysis: *s_sf*, *s_ah*, and *s_charge_avg* from the initial pool of ten surface descriptors, only applying seven for machine learning.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 December 2023 doi:10.20944/preprints202312.1187.v1

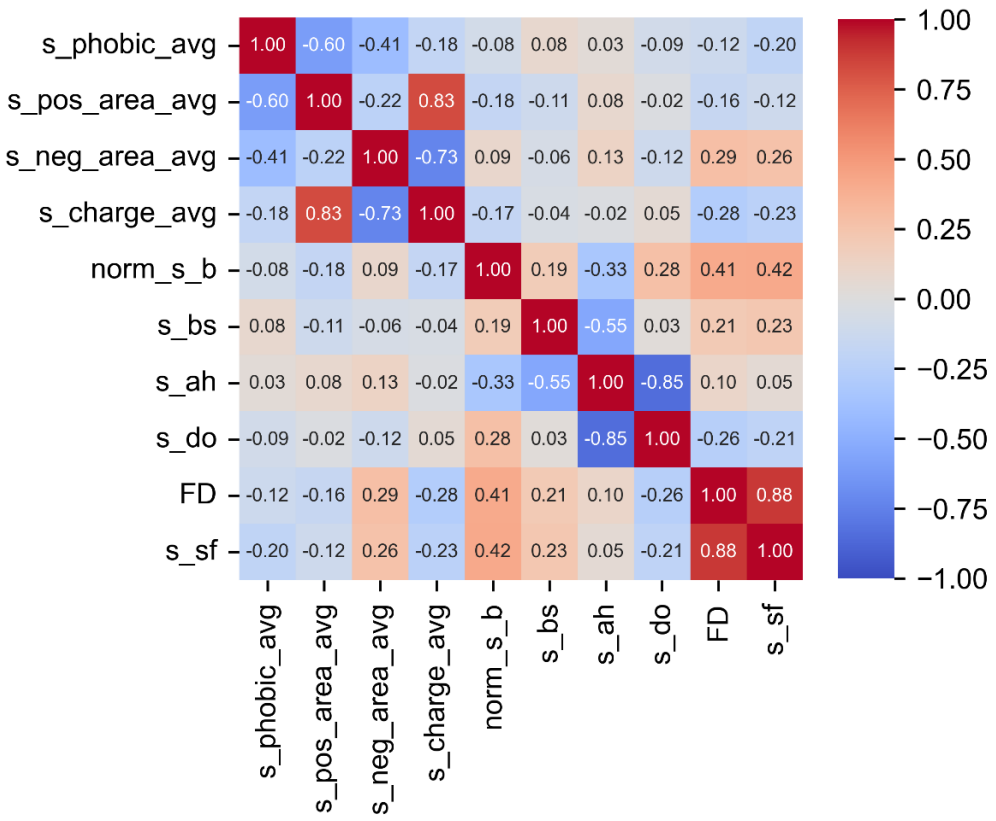


Figure 3. PC coefficients between the descriptors from a train set.

4.2. Comparison of Supervised Machine Learning Algorithms for Binary Classification Problem

The performance of different machine learning algorithms for the binary classification problem (KNN, LR, RF, and SVM) was compared using identical training and test data sets. The performance of each model was evaluated using accuracy and AUC-ROC graphs. The models were compared with the 5-times of randomly splitting train and test sets to avoid the effect of fluctuation in the results (Figure 4a). As a result, all the algorithms showed excellent and similar performance, exhibiting 79.7%,80.2%,79.3%, and 80.2% accuracy for KNN, LR, RF, and SVM, respectively. The ROC curves for the algorithms were also in nearly identical and impartial shapes (Figure 4b). The algorithms also demonstrated comparable AUC scores, with the LR exhibiting the highest AUC score (87.6%), albeit not significantly outperforming the other algorithms (87.5%,87.3%, and 87.1% for KNN, RF, and SVM, respectively). After comprehensively considering prediction performance and interpretability, we chose LR for the feature importance analysis of the surface descriptors.

 doi:10.20944/preprints202312.1187.v1

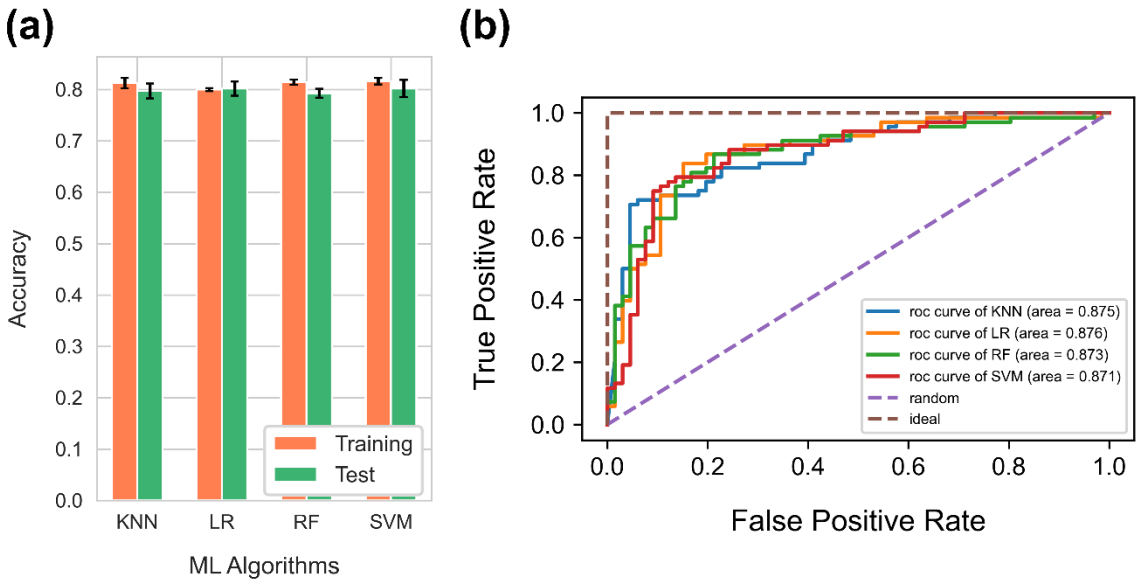


Figure 4. (a) Comparison of the performance of different ML algorithms. KNN: K-Nearest Neighbor; RF: Random Forest; LR: Logistic Regression; and SVM: Support Vector Machine (b) ROC curves for four machine learning algorithms from a single-shot trial. Hyperparameters used to tune each model in a single-shot trial is described in Table S1.

4.3. Results of Logistic Regression Analysis

Table 3 and Figure 5 show the influence of each surface descriptor on the logistic regression analysis. The coefficients and standard errors of the descriptors were calculated based on the mean values from five randomly split training sets. Table 3 showed that all the surface descriptors were statistically significant at 0.05 ($p<0.05$). The sign of the coefficient for each descriptor determines its influence on the probability of the protein being classified as an HAC protein: a positive coefficient suggests that an increase in the descriptor value increases the likelihood of the protein being classified as an HAC protein. In contrast, a negative coefficient indicates that an increase in the descriptor value decreases the probability of the protein being classified as an HAC protein. Two descriptors related to surface charge had positive coefficients in the model: negatively charged surface area (*s_neg_area*) and positively charged surface area (*s_pos_area*).

On the other hand, the other descriptors: surface hydrophobicity (*s_phobic_avg*), normalized surface B-factor (*norm_s_b*), the proportion of surface beta structures (*s_bs*), surface roughness (*FD*), and the proportion of surface disordered regions (*s_do*) exhibited negative coefficients. Moreover, the odds ratio, which is the exponentiated coefficient of a descriptor, along with its 95% confidence interval (C.I.), can aid in interpreting each coefficient by providing information on the probability of being a HAC protein.[41] All the statistical summaries of each descriptor were provided in Table S2. The following sections will provide further statistical details for each descriptor, including their relationships with several issues related to crowded cellular environments and nonspecific interactions.

Table 3. Results of Logistic Regression Analysis for each surface descriptor.

Logistic Regression Analysis							
Descriptors	β	S.E.	z-value	Significant Level	Odds ratio	Exp(β) 95% C.I.	
						Min	Min
s_phobic_avg	-0.807	0.045	-17.913	<0.001	0.446	0.408	0.487

 doi:10.20944/preprints202312.1187.v1

s_pos_area_avg	0.617	0.051	12.016	<0.001	1.853	1.675	2.049
s_neg_area_avg	0.622	0.047	13.112	<0.001	1.862	1.697	2.043
norm_s_b	-0.408	0.029	-13.972	<0.001	0.665	0.628	0.704
s_bs	-0.286	0.036	-7.992	<0.001	0.751	0.700	0.806
s_do	-0.138	0.050	-2.738	<0.05	0.872	0.790	0.962
FD	-0.265	0.024	-11.211	<0.001	0.767	0.733	0.804

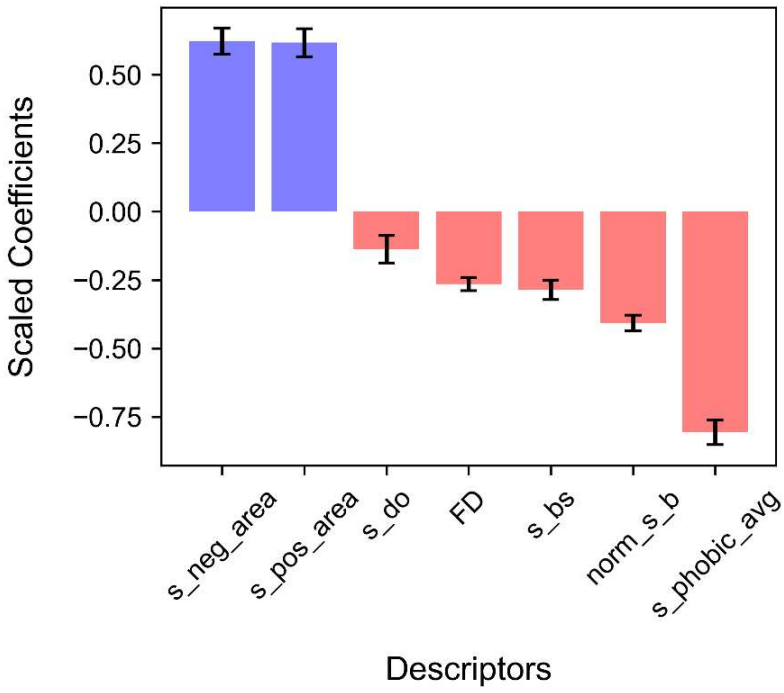


Figure 5. Importance of surface descriptors toward classifying proteins into HAC and extracellular proteins. Error bars denote the standard deviation from five times of randomly splitting training and test sets to prevent sampling bias and overfitting.

4.4. Proper Folding of HAC Proteins Can Be Achieved with Low Surface Hydrophobicity and Secondary Structure Compositions

Our findings corroborate that HAC proteins adopt a protein folding strategy limiting the nonspecific interactions in crowded environments. A protein entropically prefers a compactly folded state over an unfolded or expanded state in macromolecular crowded environments.[42-44] Especially, hydrophobic interaction plays a central role in protein folding, clustering non-polar residues in the protein core to form globular structures.[45] On the other hand, polar residues are often exposed to the protein surface, restricting hydrophobic interactions involved in molecular recognition. We observed that the surfaces of HAC proteins exhibited lower hydrophobicity and well-folded states with a lower proportion of disordered regions (Figure 6a,b).

Surface hydrophobicity, as quantitatively measured using the normalized consensus hydrophobicity scale proposed by Eisenberg et al. (Figure 6c),[27] had the highest influence (*s_phobic_avg* = -0.807) among all the surface descriptors (Figure 5). With the considerably high population of highly hydrophilic aspartic acid (D) and arginine (R), we assume that the significantly high surface hydrophilicity on HAC proteins mainly derives from the remarkable scarcity of leucine (L) and notably abundant lysine (K) and glutamic acid (E) (Figure 6d). Our observations of the high population of K and E on the HAC protein surfaces are consistent with the findings of White et al.[9]

 doi:10.20944/preprints202312.1187.v1

Their study demonstrated that molecular chaperones, which require non-adhesive surfaces for reversible interactions with multiple proteins, have a higher abundance of E and K, that possess strong water-binding properties and weak associations with surrounding amino acids. Here, we suggest that highly hydrophobic L also plays a vital role in forming hydrophilic surfaces. While the proportion of L is similar in buried regions of both protein types, there is a significant contrast on the surface region, where HAC proteins are strikingly lacking L compared to extracellular proteins (Figure 6d). Hence, the HAC proteins can have a stable hydrophobic core and exhibit higher surface hydrophilicity.

The negative coefficients (s_bs= -0.286 and s_do= -0.138) obtained from Figure 5 indicated that the HAC proteins generally exhibited higher proportions of alpha-helices and lower proportions of beta structures and disordered regions than those of extracellular proteins in both surface and buried regions (Figure 6b). This trend of surface secondary structures aligns with the global secondary structures of cytoplasmic proteins proposed by Loos et al., which revealed that cytoplasmic proteins are globally more enriched in alpha-helices and show lower frequency of beta structures and disordered regions.[8] Furthermore, the surface trend in the two well-folded structures: alpha-helices and beta structures can be supported by the previous study of Bhattacharjee and Biswas, which suggested that beta sheets are highly hydrophobic and buried in the core of proteins. In contrast, long polar residues contribute to the formation of alpha helices.[46] The lower proportions of disordered regions of HAC proteins can be explained by the nonspecific interaction propensity of its innate flexibility. The study of Nishizawa et al. highlighted the engagement of disordered regions on nonspecific interaction, observing the nonspecific ATP-protein interactions in intrinsically disordered proteins and flexible regions.[47] The study used NMR spectroscopy and molecular dynamics simulations to capture concentration-dependent noncovalent interactions between ATP and disparate proteins. As a result, the interaction was notably distinct in the intrinsically disordered proteins (α-synuclein) and flexible regions (loops or termini). Our findings on the hydrophobicity and secondary structures on the surfaces of HAC proteins support the protein folding strategy for environmental adaptation in crowded environments.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 December 2023 doi:10.20944/preprints202312.1187.v1

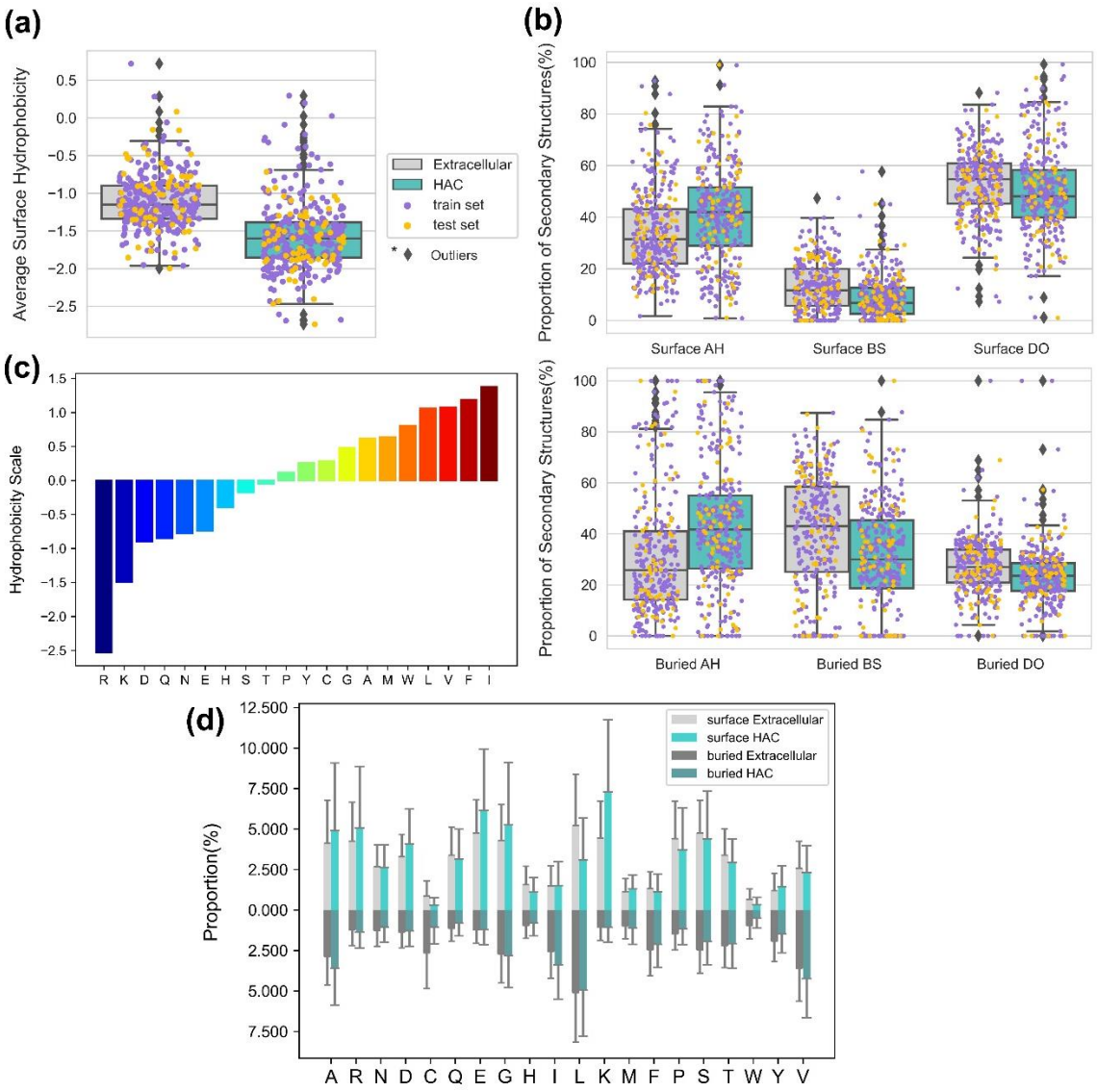


Figure 6. (a) Boxplots of average surface hydrophobicity of extracellular and HAC proteins (b) Boxplots of the proportion of secondary structures of extracellular and HAC proteins (AH: Alpha helix; BS: Beta structure; and DO: Disordered region) in the surface and buried regions (c) Hydrophobicity scale of 20 kinds of amino acids (d) Proportion of amino acids in the surface and buried regions of proteins.

4.5. The HAC Proteins Are Emphasized with Surface Rigidity and an Extreme Range of net Surface Charge

The HAC proteins should perform different structural surface characteristics to function correctly in a crowded environment. For instance, proteins in cellular environments are expected to have better thermostability with higher melting temperatures due to the crowding effect.[3] Previous studies have shown that increased thermostability is often accompanied by decreased overall flexibility of proteins.[30,31] Also, protein solubility, which indicates the characteristic of a protein to maintain its intact state, is an essential issue for protein stability to avoid aggregation, which refers to protein binding accompanying irreversible conformation change.[48] Here, we plotted the distributions of surface pseudo-B-factors and the distributions of surface charges to understand the surface flexibility and stability of HAC proteins (Figure 7).

A pseudo-B-factor gets higher as protein structures show more considerable flexibility.[23] We obtained two insights from Figure 7a: (1) the surfaces of HAC proteins tend to have lower flexibility than extracellular proteins, and (2) the lower flexibility on the surfaces of HAC proteins is

 doi:10.20944/preprints202312.1187.v1

emphasized as the analyzed domain is shifted from buried regions to surface regions. The lower flexibility on the surfaces of HAC may be supported by the recent findings on the direct relationship between protein intracellular abundance and thermal stability, which is often observed with reduced flexibility.[49,50] The findings showed that the protein interface stability was positively correlated with the protein abundance, enabling the prevention of misinteractions. At the same time, abundant intracellular proteins with high thermostability were less prone to aggregation or local unfolding. Thus, we suggest that the surfaces of HAC proteins reflect reduced flexibility to be adaptive in crowded environments.

Two charge-related descriptors with positive coefficients contributed to the model with nearly equivalent scales (*s_neg_area* = 0.622, *s_pos_area* = 0.617) (Figure 5). Our findings show that the richness of both negatively charged and positively charged areas is significant on the surfaces of HAC proteins compared to extracellular proteins (Figure 7b). To further understand the charge distribution on protein surfaces, we plotted the net surface charge distribution of extracellular and HAC proteins using the rearranged Henderson-Hasselbalch equation (more details, see Table S3) (Figure 7c).[51] In nature, it is known that zwitterionic surfaces with evenly distributed positively and negatively charged residues help resist nonspecific interactions with stronger hydrostatic repulsion fields.[4] Our data showed the more extreme range of net surface charge in HAC proteins. We assume that the results come from the complex considerations of aggregation and solubility. For instance, Ryan et al. elucidated that increased protein solubility is strongly correlated with negative surface charge, explained by the water-binding properties of E and D.[52] Also, positively charged amino acids like K and R have effectively inhibited aggregation by weakening protein-protein interactions.[53] To sum up, our results showed a higher charged surface and extreme net charge range on the surfaces of HAC proteins, and we assume that this is the result of complex behaviors of HAC proteins for the adaptation in crowded environment.

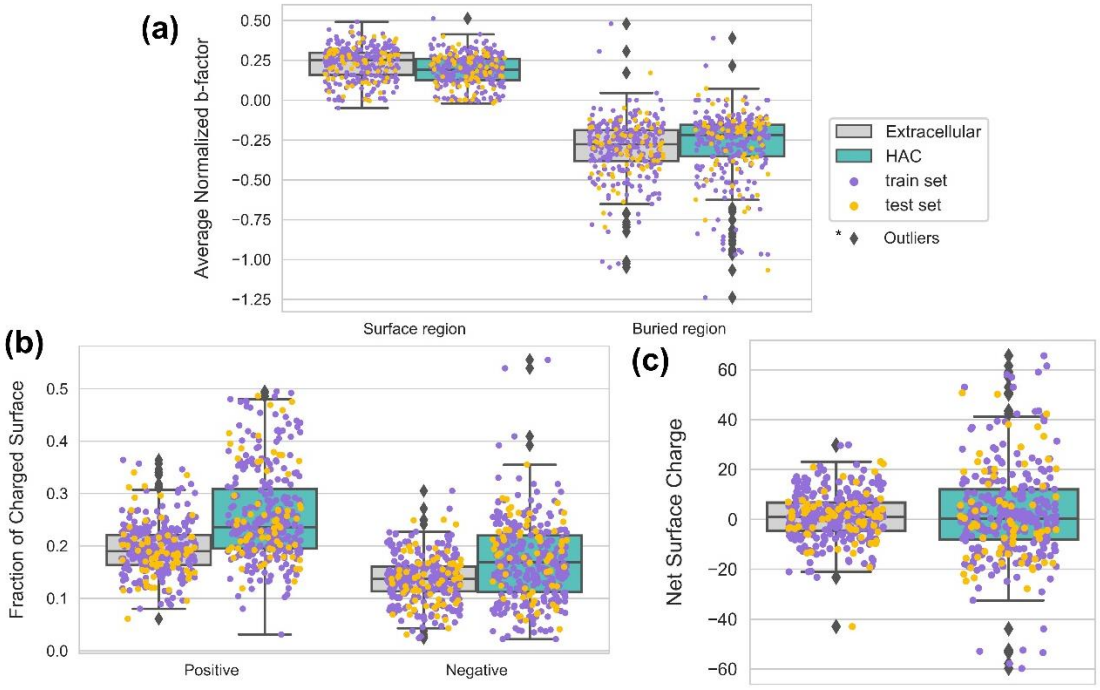


Figure 7. Boxplots of (a) average surface B-factors on surface and buried regions, (b) fraction of positively charged and negatively charged surface area on extracellular and HAC proteins, and (c) net surface charge of extracellular and HAC Proteins.

4.6. The Smoother Surface of HAC Proteins May Modulate Molecular Adsorption

As mentioned, molecular crowding and protein abundance are crucial for studying nonspecific interactions. We hypothesized that the surface geometry of HAC proteins should take strategies

 doi:10.20944/preprints202312.1187.v1

minimizing molecular adsorption and nonspecific interactions. Surface roughness is a critical parameter to describe surface geometry. Indeed, nano-scale surface roughness had a significant influence on protein-protein interactions.[54,55] Also, surface homogeneity and low surface roughness were found on the surface of streptavidin, which is known to have exceptionally strong specific binding with biotin and exhibits low nonspecific binding.[56] Here, we calculated the surface roughness of proteins using *FD*, which can represent the degree of surface irregularity.[26] *FD* shows the lowest value for a completely smooth surface (*FD*=2). In contrast, it has the highest value for the roughest protein surface (*FD*=3). With *FD* of all proteins ranging from 2.044 to 2.372, we observed subtle but discernable distinctions between extracellular and HAC proteins (Figure 8).

The HAC proteins exhibited smoother surfaces in general, which can be inferred by the large population of Alanine, which has the shortest residue chain length among 20 amino acids (Figure 6d). To add, among four types of aromatic amino acids (Tryptophan, Phenylalanine, Tyrosine, and Histidine) that can have higher van der Waals volumes, three of them (Tryptophan, Phenylalanine, and Histidine) were more abundant on the surfaces of extracellular proteins. Considering that protein surface roughness is necessary upon binding with small molecules[57], we suggest that the smoother surface of an HAC protein can be a strategy for minimizing small molecules-induced nonspecific interactions. However, further investigation will be necessary to substantiate our assumptions.

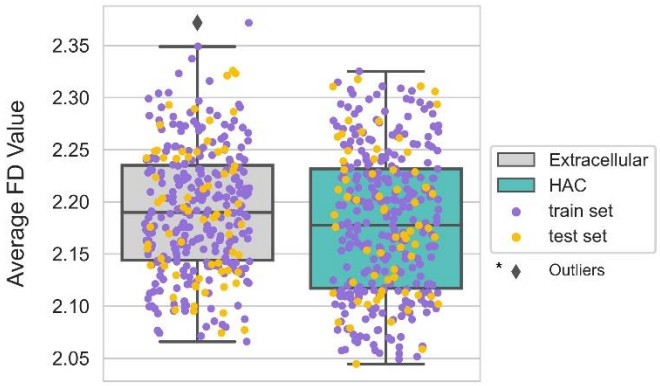


Figure 8. Boxplots of surface roughness of extracellular and HAC proteins.

5. Summary and Conclusions

In this study, we utilized surface physicochemical, structural, and geometrical descriptors to identify HAC proteins with ML and quantitatively analyzed the surface characteristics. We first solved binary classification for HAC and extracellular proteins using several supervised ML algorithms (KNN, LR, RF, and SVM). Then, LR was chosen for the descriptors' final feature importance analysis, considering both excellent model performance (80.2% accuracy, 87.6% AUC) and high model interpretability. The charge-related descriptors showed positive correlations, while hydrophobicity, B-factor, beta structures proportion, roughness, and disordered regions proportion exhibited negative correlations to the HAC proteins in the importance analysis of descriptors.

We also found that the E, K, and L populations and well-folded secondary structures on the HAC protein surfaces played vital roles in their hydrophilicity and compactly folded structures. Also, we observed limited protein flexibility and extreme net charge from the surfaces of HAC proteins, which the previous studies on the adaptation of cytoplasmic proteins in crowded environments can explain. Finally, we suggest that smoother surfaces of proteins can be critical in minimizing the nonspecific adsorption of small molecules. Our results indicate that several surface descriptors can be employed to identify, quantify, and explain the protein surface characteristics in a crowded cellular environment.

Our findings on the quantitative analysis of the descriptors could facilitate the design of surfaces well-adapted to crowded environments, such as nonspecific interaction-resistant surfaces with

 doi:10.20944/preprints202312.1187.v1

selectivity to target materials.[58-63] One example of the application is the design of immunosensors, where the nonspecific adsorption of various biomolecules causes background noise and critically impairs sensitivity.[64] Another field highlighting the importance of nonspecific interaction-resistant surfaces is reducing protein corona on nanoparticles.[65] When nanoparticles first come into contact with biological fluid, proteins attach to their surfaces and form a protein layer, i.e., protein corona. Since protein corona causes direct impacts on the performance of nanoparticles, the new strategy – applying nonspecific interaction-resistant surface – of nanoparticles should aim to reduce or slow protein corona formation.

Acknowledgments: We acknowledge Ms. Kazue Taki for her help in arranging this project. This work was supported by JSPS KAKENHI Grant Number (JP23H04059 and JP 22H045302). This work was performed under the Research Program for CORE lab of "Five-star Alliance" in "NJRC Mater. & Dev.".

References

1. Ellis, R.J. Macromolecular crowding: obvious but underappreciated. *Trends Biochem Sci* **2001**, *26*, 597-604, doi:10.1016/s0968-0004(01)01938-7.
2. Barbieri, L.; Luchinat, E.; Banci, L. Protein interaction patterns in different cellular environments are revealed by in-cell NMR. *Sci Rep* **2015**, *5*, 14456, doi:10.1038/srep14456.
3. Despa, F.; Orgill, D.P.; Lee, R.C. Molecular crowding effects on protein stability. *Ann N Y Acad Sci* **2005**, *1066*, 54-66, doi:10.1196/annals.1363.005.
4. Frutiger, A.; Tanno, A.; Hwu, S.; Tiefenauer, R.F.; Vörös, J.; Nakatsuka, N. Nonspecific Binding-Fundamental Concepts and Consequences for Biosensing Applications. *Chem Rev* **2021**, *121*, 8095-8160, doi:10.1021/acs.chemrev.1c00044.
5. Siddiqui, G.A.; Naeem, A. Connecting the Dots: Macromolecular Crowding and Protein Aggregation. *J Fluoresc* **2023**, *33*, 1-11, doi:10.1007/s10895-022-03082-2.
6. Levy, E.D.; Michnick, S.W.; Landry, C.R. Protein abundance is key to distinguish promiscuous from functional phosphorylation based on evolutionary information. *Philos Trans R Soc Lond B Biol Sci* **2012**, *367*, 2594-2606, doi:10.1098/rstb.2012.0078.
7. van den Berg, B.; Ellis, R.J.; Dobson, C.M. Effects of macromolecular crowding on protein folding and aggregation. *EMBO J* **1999**, *18*, 6927-6933, doi:10.1093/emboj/18.24.6927.
8. Loos, M.S.; Ramakrishnan, R.; Vranken, W.; Tsirigotaki, A.; Tsare, E.P.; Zorzini, V.; Geyter, J.; Yuan, B.; Tsamardinos, I.; Klappa, M., et al. Structural Basis of the Subcellular Topology Landscape of. *Front Microbiol* **2019**, *10*, 1670, doi:10.3389/fmicb.2019.01670.
9. White, A.D.; Nowinski, A.K.; Huang, W.; Keefe, A.J.; Sun, F.; Jiang, S. Decoding nonspecific interactions from nature. *Chemical Science* **2012**, *3*, 3488-3494, doi:10.1039/C2SC21135A.
10. Levy, E.D.; De, S.; Teichmann, S.A. Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. *Proc Natl Acad Sci U S A* **2012**, *109*, 20461-20466, doi:10.1073/pnas.1209312109.
11. Mer, A.S.; Andrade-Navarro, M.A. A novel approach for protein subcellular location prediction using amino acid exposure. *BMC Bioinformatics* **2013**, *14*, 342, doi:10.1186/1471-2105-14-342.
12. Casadio, R.; Martelli, P.L.; Savojardo, C. Machine learning solutions for predicting protein–protein interactions. *WIREs Computational Molecular Science* **2022**, *12*, e1618, doi:https://doi.org/10.1002/wcms.1618.
13. Crampon, K.; Giorkallos, A.; Deldossi, M.; Baud, S.; Steffenel, L.A. Machine-learning methods for ligand–protein molecular docking. *Drug Discovery Today* **2022**, *27*, 151-164, doi:https://doi.org/10.1016/j.drudis.2021.09.007.
14. Zhang, T.; Gu, J.; Wang, Z.; Wu, C.; Liang, Y.; Shi, X. Protein Subcellular Localization Prediction Model Based on Graph Convolutional Network. *Interdiscip Sci* **2022**, *14*, 937-946, doi:10.1007/s12539-022-00529-9.
15. 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, doi:10.1038/s41586-021-03819-2.
16. Fam, S.Y.; Chee, C.F.; Yong, C.Y.; Ho, K.L.; Mariatulqabtiah, A.R.; Tan, W.S. Stealth Coating of Nanoparticles in Drug-Delivery Systems. *Nanomaterials (Basel)* **2020**, *10*, doi:10.3390/nano10040787.
17. Wang, M.; Weiss, M.; Simonovic, M.; Haertinger, G.; Schrimpf, S.P.; Hengartner, M.O.; von Mering, C. PaxDb, a database of protein abundance averages across all three domains of life. *Mol Cell Proteomics* **2012**, *11*, 492-500, doi:10.1074/mcp.O111.014704.
18. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T., et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **2000**, *25*, 25-29, doi:10.1038/75556.
19. Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A., et al. AlphaFold Protein Structure Database: massively expanding the structural coverage

of protein-sequence space with high-accuracy models. *Nucleic Acids Res* **2022**, *50*, D439-D444, doi:10.1093/nar/gkab1061.
20. Mariani, V.; Biasini, M.; Barbato, A.; Schwede, T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **2013**, *29*, 2722-2728, doi:10.1093/bioinformatics/btt473.
21. Tunyasuvunakool, K.; Adler, J.; Wu, Z.; Green, T.; Zielinski, M.; Žídek, A.; Bridgland, A.; Cowie, A.; Meyer, C.; Laydon, A., et al. Highly accurate protein structure prediction for the human proteome. *Nature* **2021**, *596*, 590-596, doi:10.1038/s41586-021-03828-1.
22. Ruff, K.M.; Pappu, R.V. AlphaFold and Implications for Intrinsically Disordered Proteins. *J Mol Biol* **2021**, *433*, 167208, doi:10.1016/j.jmb.2021.167208.
23. Guo, H.B.; Perminov, A.; Bekele, S.; Kedziora, G.; Farajollahi, S.; Varaljay, V.; Hinkle, K.; Molinero, V.; Meister, K.; Hung, C., et al. AlphaFold2 models indicate that protein sequence determines both structure and dynamics. *Sci Rep* **2022**, *12*, 10696, doi:10.1038/s41598-022-14382-9.
24. Morris, O.M.; Torpey, J.H.; Isaacson, R.L. Intrinsically disordered proteins: modes of binding with emphasis on disordered domains. *Open Biology* **2021**, *11*, 210222, doi:doi:10.1098/rsob.210222.
25. Maglic, J.B.; Lavendomme, R.: an easy-to-use program for analyzing cavities, volumes and surface areas of chemical structures. *J Appl Crystallogr* **2022**, *55*, 1033-1044, doi:10.1107/S1600576722004988.
26. Lewis, M.; Rees, D.C. Fractal surfaces of proteins. *Science* **1985**, *230*, 1163-1165, doi:10.1126/science.4071040.
27. Eisenberg, D.; Schwarz, E.; Komaromy, M.; Wall, R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J Mol Biol* **1984**, *179*, 125-142, doi:10.1016/0022-2836(84)90309-7.
28. Debye, P. Interferenz von Röntgenstrahlen und Wärmebewegung. *Annalen der Physik* **1913**, *348*, 49-92, doi:https://doi.org/10.1002/andp.19133480105.
29. Trueblood, K.N.; Burgi, H.-B.; Burzlaff, H.; Dunitz, J.D.; Gramaccioli, C.M.; Schulz, H.H.; Shmueli, U.; Abrahams, S.C. Atomic Dispacement Parameter Nomenclature. Report of a Subcommittee on Atomic Displacement Parameter Nomenclature. *Acta Crystallographica Section A* **1996**, *52*, 770-781, doi:doi:10.1107/S0108767396005697.
30. Parthasarathy, S.; Murthy, M.R. Protein thermal stability: insights from atomic displacement parameters (B values). *Protein Eng* **2000**, *13*, 9-13, doi:10.1093/protein/13.1.9.
31. Vihinen, M. Relationship of protein flexibility to thermostability. *Protein Eng* **1987**, *1*, 477-480, doi:10.1093/protein/1.6.477.
32. Yuan, Z.; Zhao, J.; Wang, Z.X. Flexibility analysis of enzyme active sites by crystallographic temperature factors. *Protein Eng* **2003**, *16*, 109-114, doi:10.1093/proeng/gzg014.
33. Blaisse, M.R.; Fu, B.; Chang, M.C.Y. Structural and Biochemical Studies of Substrate Selectivity in Ascaris suum Thiolases. *Biochemistry* **2018**, *57*, 3155-3166, doi:10.1021/acs.biochem.7b01123.
34. Liu, Q.; Li, Z.; Li, J. Use B-factor related features for accurate classification between protein binding interfaces and crystal packing contacts. *BMC Bioinformatics* **2014**, *15 Suppl 16*, S3, doi:10.1186/1471-2105-15-S16-S3.
35. Oeffner, R.D.; Croll, T.I.; Millán, C.; Poon, B.K.; Schlicksup, C.J.; Read, R.J.; Terwilliger, T.C. Putting AlphaFold models to work with phenix.process_predicted_model and ISOLDE. *Acta Crystallogr D Struct Biol* **2022**, *78*, 1303-1314, doi:10.1107/S2059798322010026.
36. 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, doi:10.1126/science.abj8754.
37. Hiranuma, N.; Park, H.; Baek, M.; Anishchenko, I.; Dauparas, J.; Baker, D. Improved protein structure refinement guided by deep learning based accuracy estimation. *Nat Commun* **2021**, *12*, 1340, doi:10.1038/s41467-021-21511-x.
38. Sun, Z.; Liu, Q.; Qu, G.; Feng, Y.; Reetz, M.T. Utility of B-Factors in Protein Science: Interpreting Rigidity, Flexibility, and Internal Motion and Engineering Thermostability. *Chem Rev* **2019**, *119*, 1626-1665, doi:10.1021/acs.chemrev.8b00290.
39. Schlessinger, A.; Rost, B. Protein flexibility and rigidity predicted from sequence. *Proteins* **2005**, *61*, 115-126, doi:10.1002/prot.20587.
40. Voss, N.R.; Gerstein, M. 3V: cavity, channel and cleft volume calculator and extractor. *Nucleic Acids Res* **2010**, *38*, W555-562, doi:10.1093/nar/gkq395.
41. Hosmer Jr, D.W.; Lemeshow, S.; Sturdivant, R.X. *Applied logistic regression*; John Wiley & Sons: 2013; Vol. 398.
42. Gomez, D.; Huber, K.; Klumpp, S. On Protein Folding in Crowded Conditions. *J Phys Chem Lett* **2019**, *10*, 7650-7656, doi:10.1021/acs.jpclett.9b02642.
43. Tokuriki, N.; Kinjo, M.; Negi, S.; Hoshino, M.; Goto, Y.; Urabe, I.; Yomo, T. Protein folding by the effects of macromolecular crowding. *Protein Science* **2004**, *13*, 125-133, doi:https://doi.org/10.1110/ps.03288104.
44. Minton, A.P. Excluded volume as a determinant of macromolecular structure and reactivity. *Biopolymers* **1981**, *20*, 2093-2120, doi:https://doi.org/10.1002/bip.1981.360201006.

45. Tang, S.; Li, J.; Huang, G.; Yan, L. Predicting Protein Surface Property with its Surface Hydrophobicity. *Protein Pept Lett* **2021**, *28*, 938-944, doi:10.2174/0929866528666210222160603.
46. Bhattacharjee, N.; Biswas, P. Structural patterns in alpha helices and beta sheets in globular proteins. *Protein Pept Lett* **2009**, *16*, 953-960, doi:10.2174/092986609788923239.
47. Nishizawa, M.; Walinda, E.; Morimoto, D.; Kohn, B.; Scheler, U.; Shirakawa, M.; Sugase, K. Effects of Weak Nonspecific Interactions with ATP on Proteins. *J Am Chem Soc* **2021**, *143*, 11982-11993, doi:10.1021/jacs.0c13118.
48. Vihinen, M. Solubility of proteins. *ADMET DMPK* **2020**, *8*, 391-399, doi:10.5599/admet.831.
49. He, Y.M.; Ma, B.G. Abundance and Temperature Dependency of Protein-Protein Interaction Revealed by Interface Structure Analysis and Stability Evolution. *Sci Rep* **2016**, *6*, 26737, doi:10.1038/srep26737.
50. Leuenberger, P.; Ganscha, S.; Kahraman, A.; Cappelletti, V.; Boersema, P.J.; von Mering, C.; Claassen, M.; Picotti, P. Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability. *Science* **2017**, *355*, doi:10.1126/science.aai7825.
51. Moore, D.S. Amino acid and peptide net charges: A simple calculational procedure. *Biochemical Education* **1985**, *13*, 10-11, doi:https://doi.org/10.1016/0307-4412(85)90114-1.
52. Kramer, Ryan M.; Shende, Varad R.; Motl, N.; Pace, C.N.; Scholtz, J.M. Toward a Molecular Understanding of Protein Solubility: Increased Negative Surface Charge Correlates with Increased Solubility. *Biophysical Journal* **2012**, *102*, 1907-1915, doi:https://doi.org/10.1016/j.bpj.2012.01.060.
53. Wang, W.; Nema, S.; Teagarden, D. Protein aggregation—Pathways and influencing factors. *International Journal of Pharmaceutics* **2010**, *390*, 89-99, doi:https://doi.org/10.1016/j.ijpharm.2010.02.025.
54. Rechendorff, K.; Hovgaard, M.B.; Foss, M.; Zhdanov, V.P.; Besenbacher, F. Enhancement of protein adsorption induced by surface roughness. *Langmuir* **2006**, *22*, 10885-10888, doi:10.1021/la0621923.
55. Scopelliti, P.E.; Borgonovo, A.; Indrieri, M.; Giorgetti, L.; Bongiorno, G.; Carbone, R.; Podestà, A.; Milani, P. The effect of surface nanometre-scale morphology on protein adsorption. *PLoS One* **2010**, *5*, e11862, doi:10.1371/journal.pone.0011862.
56. Ettelt, V.; Ekat, K.; Kämmerer, P.W.; Kreikemeyer, B.; Epple, M.; Veith, M. Streptavidin-coated surfaces suppress bacterial colonization by inhibiting non-specific protein adsorption. *J Biomed Mater Res A* **2018**, *106*, 758-768, doi:10.1002/jbm.a.36276.
57. Pettit, F.K.; Bowie, J.U. Protein surface roughness and small molecular binding sites. *J Mol Biol* **1999**, *285*, 1377-1382, doi:10.1006/jmbi.1998.2411.
58. Chang, R.Y.S.; Mondarte, E.A.Q.; Palai, D.; Sekine, T.; Kashiwazaki, A.; Murakami, D.; Tanaka, M.; Hayashi, T. Protein- and Cell-Resistance of Zwitterionic Peptide-Based Self-Assembled Monolayers: Anti-Biofouling Tests and Surface Force Analysis. *Front Chem* **2021**, *9*, 748017, doi:ARTN 74801710.3389/fchem.2021.748017.
59. Hayashi, T.; Sano, K.; Shiba, K.; Iwahori, K.; Yamashita, I.; Hara, M. Critical amino acid residues for the specific binding of the Ti-recognizing recombinant ferritin with oxide surfaces of titanium and silicon. *Langmuir* **2009**, *25*, 10901-10906, doi:10.1021/la901242q.
60. Hayashi, T.; Sano, K.; Shiba, K.; Kumashiro, Y.; Iwahori, K.; Yamashita, I.; Hara, M. Mechanism underlying specificity of proteins targeting inorganic materials. *Nano Lett* **2006**, *6*, 515-519, doi:10.1021/nl060050n.
61. Kim, S.O.; Jackman, J.A.; Mochizuki, M.; Yoon, B.K.; Hayashi, T.; Cho, N.J. Correlating single-molecule and ensemble-average measurements of peptide adsorption onto different inorganic materials. *Phys Chem Chem Phys* **2016**, *18*, 14454-14459, doi:10.1039/c6cp01168c.
62. Mochizuki, M.; Oguchi, M.; Kim, S.O.; Jackman, J.A.; Ogawa, T.; Lkhamsuren, G.; Cho, N.J.; Hayashi, T. Quantitative Evaluation of Peptide-Material Interactions by a Force Mapping Method: Guidelines for Surface Modification. *Langmuir* **2015**, *31*, 8006-8012, doi:10.1021/acs.langmuir.5b01691.
63. Yamashita, K.; Kirimura, H.; Okuda, M.; Nishio, K.; Sano, K.I.; Shiba, K.; Hayashi, T.; Hara, M.; Mishima, Y. Selective nanoscale positioning of ferritin and nanoparticles by means of target-specific peptides. *Small* **2006**, *2*, 1148-1152, doi:10.1002/smll.200600220.
64. Wen, W.; Yan, X.; Zhu, C.; Du, D.; Lin, Y. Recent Advances in Electrochemical Immunosensors. *Anal Chem* **2017**, *89*, 138-156, doi:10.1021/acs.analchem.6b04281.
65. Rampado, R.; Crotti, S.; Caliceti, P.; Pucciarelli, S.; Agostini, M. Recent Advances in Understanding the Protein Corona of Nanoparticles and in the Formulation of "Stealthy" Nanomaterials. *Front Bioeng Biotechnol* **2020**, *8*, 166, doi:10.3389/fbioe.2020.00166.