Supporting Information

Investigating Structural Biophysical Features for Antigen-Binding Fragment Crystallization via Machine Learning

Krishna Gopal Chattaraj, Joana Ferreira, Allan S. Myerson and Bernhardt L. Trout*

Department of Chemical Engineering; Massachusetts Institute of Technology; Cambridge, MA 02139, USA

Corresponding Author Email: *trout@mit.edu*

FAB PDB id:	6W4W, 6UMI, 6PE7, 6WG8, 6JC2, 6CR1, 6CNR, 5WUV, 5WHJ, 5NHW, 5I1C, 5GGQ, 5FHA, 5D7S, 5CMA, 4UB0, 4QXG, 4PUB, 4M6O, 4KAQ, 4JHA, 4HWE, 4G6K, 4G5Z, 4FQ2, 4FQ1, 4EOW, 4DN3, 3S34, 3NTC, 3NFS, 3HMW, 3GKW, 3GIZ, 3EO9, 3C08, 1T3F, 1MIM, 1L7I, 1BEY

TABLE S1: List of protein database identifiers for FABs used in the study.

Category	Feature	Subtypes	Description	NBH?
Structure	Hydrophobicity (Black-Mould)	-	Residue hydrophobicity	y
	protrusion index (CX)	Maximum	Maximum CX of any atom in the residue	y
		Average	Average CX over number of all atoms in the residue	y
Structure	Fractional exposure "or" relative exposure	-	ratio of the residue's SASA to the standard exposure of the residue in Ala-X-Ala	n
Structure	depth index (DPX)	Minimum	distance from the closest point on the Fab surface, calculated for each atom in a residue. Minimum is lowest value of any atom in the residue	y
		Maximum	Maximum DPX of any atom in the residue	y
		Average	Average DPX value of all atoms in the residue	y

TABLE S2: "Category" designates the overarching classification of each feature, delineated as either sequence-based or structural. Any affiliated subtypes for the features are enumerated under "Subtypes". Features computed exclusively for the residue, as well as in conjunction with neighboring residues within specified cutoff distances (5, 10, 15, 20, and 25 Å), are identified in the "Neighborhood (NBH)" column. In hydrophobicity analysis, we normalized parameter values by dividing them by the total number of residues within specific distances. During the calculation of the depth index, we modified the process slightly. We searched for nearby surface-exposed atoms within a 9 Å radius around any given atom. If a nearby surface-exposed atom was found, we calculated the distance. Otherwise, we assigned a high DPX value of 10 of that atom. All features were calculated using a custom TCL script through the VMD interface.

Category	Feature	Subtypes	Description	NBH?
Structure	Net Charge	Total	Residue net charge	y
		Exposed	Sum of partial charges of atoms with SASA > 0 Å	
	SAP	SAP	Measure of surface-exposed hydrophobicity of atoms within 5 Å	y
		SAP_10	Measure of surface-exposed hydrophobicity of atoms within 10 Å	y
		SAP_all_5	Measure of surface-exposed hydrophobicity of atoms within 5 Å	y
	SCM	SCM_postive	Measure of exposed positive charge within 10 Å	y
		SCM_negative	Measure of exposed negative charge within 10 Å	y
		SCM_all	Measure of exposed charge within 10 Å	y
Structure	SASA	All	SASA of all atoms in the residue	y
		Hydrophobic	SASA of hydrophobic residues	
		Hydrophilic	SASA of hydrophilic residues	
		Polar	SASA of polar residues	
		Aliphatic	SASA of aliphatic residues	
		Aromatic	SASA of aromatic residues	
		Charged	SASA of charged residues	
		Positive	SASA of positive residues	
		Negative	SASA of negative residues	
		Side Chain	Side chain of residues	
		Backbone	Backbone of residues	

TABLE S3: "Category" designates the overarching classification of each feature, delineated as either sequence-based or structural. Any affiliated subtypes for the features are enumerated under "Subtypes". Features computed exclusively for the residue, as well as in conjunction with neighboring residues within specified cutoff distances (5, 10, 15, 20, and 25 Å), are identified in the "Neighborhood (NBH)" column. SAP and SCM were derived from an in-house python script, while other calculations were conducted using a custom TCL script through the VMD interface.

Category	Feature	Subtypes	Description	NBH?
Structure	steric parameters (graph shape index)	–	residue steric parameter	y
	volume	–	residue volume	y
	isoelectric point	–	residue isoelectric point	y
	helix probability	–	residue helix probability	y
	sheet probability	–	residue sheet probability	y
	hydrophilicity	–	residue hydrophilicity	y
	side-chain residue size (SCRS)	–	residue size chain size	y
	polarity	–	residue polarity	y
	polarizability	–	residue polarizability	y
	(SASA) solvent-accessible surface area	–	residue sasa	y
	(NCN) net charge number residue	–	net charge number	y
	pI-pKa parameter residue	–	pI-pKa parameter	y

TABLE S4: "Category" designates the overarching classification of each feature, delineated as either sequence-based or structural. Any affiliated subtypes for the features are enumerated under "Subtypes". Features computed exclusively for the residue, as well as in conjunction with neighboring residues within specified cutoff distances (5, 10, 15, 20, and 25 Å), are identified in the "Neighborhood (NBH)" column. Here we have used various scales of different parameters used for protein-protein interaction. In our analysis, we employed a normalization technique for the parameter values, dividing each total values by the total number of residues within the respective distances. Utilizing these normalized values facilitates a more prominent comparison across specific regions of both classes. Features were derived from an in-house script.

Category	Feature	Subtypes	Description	NBH?
Structure	number	hydrophobic (ALA, ILE, LEU, MET, PHE, PRO, TRP, VAL)	number of hydrophobic residues	y
		hydrophilic (SER, THR, ASN, GLN, TYR, LYS, ARG, HIS, ASP, GLU)	number of hydrophilic residues	y
		polar (SER, THR, ASN, GLN, TYR)	number of polar residues	y
		aliphatic (calculated from VMD)	number of aliphatic residues	y
		aromatic (calculated from VMD)	number of aromatic residues	y
		charged (HSP, LYS, ARG, GLU, ASP)	number of charged residues	y
		positive (ARG, HSP, and LYS)	number of positive residues	y
		negative (ASP and GLU)	number of negative residues	y

TABLE S5: "Category" designates the overarching classification of each feature, delineated as either sequence-based or structural. Any affiliated subtypes for the features are enumerated under "Subtypes". Features computed exclusively for the residue, as well as in conjunction with neighboring residues within specified cutoff distances (5, 10, 15, 20, and 25 Å), are identified in the "Neighborhood (NBH)" column. Features were derived from an in-house script and also using a custom TCL script through the VMD interface.

Category	Feature	Subtypes	Description	NBH?
Structure	Number, SASA, sidechain SASA	Glycine (GLY)	A unique amino acid with a single H-atom as its side chain.	y
		Very Small Amino Acids	Include Alanine (ALA) and Valine (VAL).	y
		Small Amino Acids	Include Serine (SER), Threonine (THR), Isoleucine (ILE), Cysteine (CYS), and Leucine (LEU).	y
		Normal-Sized Amino Acids	Include Aspartic Acid (ASP), Asparagine (ASN), Glutamic Acid (GLU), Glutamine (GLN), and Methionine (MET).	y
		Long Amino Acids	Include Arginine (ARG) and Lysine (LYS).	y
		Proline (PRO)	A unique amino acid with a cyclic side chain.	y
		Polar Uncharged Amino Acids with OH group	Include Serine (SER) and Threonine (THR).	y
		Polar Uncharged Amino Acids with Amide group	Include Asparagine (ASN) and Glutamine (GLN).	y
		Non-Polar Sulfur-Containing Amino Acids	Include Methionine (MET) and Cysteine (CYS).	y

TABLE S6: "Category" designates the overarching classification of each feature, delineated as either sequence-based or structural. Any affiliated subtypes for the features are enumerated under "Subtypes". Features computed exclusively for the residue, as well as in conjunction with neighboring residues within specified cutoff distances (5, 10, 15, 20, and 25 Å), are identified in the "Neighborhood (NBH)" column. All features were calculated using a custom TCL script through the VMD interface.

Category	Feature	Description	NBH?
Structure	SASA × Hydrophobicity (BM)	The solvent accessible surface area of each residue is multiplied by its hydrophobicity score according to the BM scale.	y
	Hydrophobic SASA × Hydrophobicity (BM)	Only the SASA of hydrophobic residues is considered and multiplied by their respective hydrophobicity score (BM scale)	y
	Sidechain SASA × Hydrophobicity (BM)	The SASA of only the side chain of residues is considered and multiplied by their respective hydrophobicity score (BM scale)	y
	Relative Sidechain SASA × Hydrophobicity (BM)	The relative sidechain SASA (sidechain SASA divided by the maximum possible SASA) of each residue is multiplied by its hydrophobicity score (BM scale)	y
	Hydrophobic Relative Sidechain SASA × Hydrophobicity (BM)	The relative sidechain SASA of only the hydrophobic residues is considered and multiplied by their respective hydrophobicity score (BM scale)	y

TABLE S7: "Category" designates the overarching classification of each feature, delineated as either sequence-based or structural. Any affiliated subtypes for the features are enumerated under "Subtypes". Features computed exclusively for the residue, as well as in conjunction with neighboring residues within specified cutoff distances (5, 10, 15, 20, and 25 Å), are identified in the "Neighborhood (NBH)" column. Here, BM represents Black-Mould hydrophobicity scale and "×" represents multiplication. All features were calculated using a custom TCL script through the VMD interface.

Feature (structure)	SAP of i-th	Sidechain (SC) SASA of i & i+1		Sidechain (SC) SASA of i & i-1		Sidechain (SC) SASA of each pair of i & i_{nbh}		Residue type of i_{nbh}	d	SASA
		i	i+1	i	i-1	i	i_{nbh}			
SAP-adjacent-SC	≥ 0.15 & ≥ 0.20	5 & 10	5 & 10	5 & 10	5 & 10	5 & 10	10, 20, 30, 40, 50, & 75	any type, hydrophobic-polar-charged, hydrophobic-charged, charged, charged-polar, & Specific_hydrophobic-polar-charged	≤ 5 Å & ≤ 7 Å & ≤ 10 Å	≥ 50 $Å^2$

TABLE S8: The combinations of various conditions that are used to engineer **SAP_adjacent** (specifically, **SAP_adjacent_SC)** features by considering the SAP value of i-th residue and the sidechain-solvent accessible surface area (SC-SASA) values of both the i-th residue and its immediate neighbors, which include i+1, i-1, and i_{nbh}. To accurately represent and interpret each feature in the dataset, a distinct nomenclature has been devised. This nomenclature is formulated as **"SAP-adjacent-SC_(SAP_value) _ (SC-SASA of (i and i+1)) _ (SC-SASA of (i and i-1)) _ (SC-SASA of (i and i_{nbh})) _d_(residue-type)"**. For example, the feature denoted as **SAP-adjacent-SC_0.20_10-10_10-10_10-10_7_charged-polar** in the dataset specifically refers to a feature characterized by its unique conditions. The term 'specific hydrophobic' refers to a selected group of hydrophobic residues, which includes ILE, LEU, TRP, ALA, VAL, and PRO. 'd' represents the distance from the i-th residue used to determine the proximity of 'i_{nbh}' residues. SAP was derived from an in-house python script, while other calculations were conducted using a custom TCL script through the VMD interface.

Feature (structure)	SAP of i-th	Fractional exposure (FE) of i & i+1		Fractional exposure (FE) of i & i-1		Fractional exposure (FE) of each pair of i & i_{nbh}		Residue type of i_{nbh}	d	SASA
		i	i+1	i	i-1	i	i_{nbh}			
SAP-adjacent-FE	≥ 0.15	5	5	5	5	5	5, 10, 20, 30, & 40	any type, hydrophobic-polar-charged, hydrophobic-charged, & charged-polar	≤ 5 Å & ≤ 7 Å	≥ 50 Å^2

TABLE S9: The combinations of various conditions that are used to engineer **SAP_adjacent** (specifically, **SAP_adjacent_FE)** features by considering the SAP value of i-th residue and the fractional exposure values of both the i-th residue and its immediate neighbors, which include i+1, i-1, and i_{nbh}. The surface exposure of residues was assessed by computing their fractional exposure (FE), which is defined as the ratio of a residue's Solvent Accessible Surface Area (SASA) of sidechain to its standard sidechain exposure in an Ala-X-Ala tri-peptide configuration. To accurately represent and interpret each feature in the dataset, a distinct nomenclature has been devised. This nomenclature is formulated as **"SAP-adjacent-FE_(SAP_value) _ (FE of (i and i+1)) _ (FE of (i and i-1)) _ (FE of (i and i_{nbh})) _d_(residue-type)"**. For example, the feature denoted as **SAP-adjacent-FE_0.15_5%-5%_5%-5%_5%-5%_7_charged-polar** in the dataset specifically refers to a feature characterized by its unique conditions. 'd' represents the distance from the i-th residue used to determine the proximity of 'i_{nbh}' residues. SAP was derived from an in-house python script, while other calculations were conducted using a custom TCL script through the VMD interface.

Feature (structure)	**SAP of i-th**	**SASA of (i & i+1) and Sidechain (SC) SASA of (i & i+1)**				**SASA of (i & i+1) and Sidechain (SC) SASA of (i & i-1)**				**SASA of (i & i+1) and Sidechain (SC) SASA of each pair of (i & i_{nbh})**				**Residue type**	**d**
		SASA of i	SC-SASA of i	SASA of i+1	SC-SASA i+1	SASA of i	SC-SASA of i	SASA of i+1	SC-SASA i+1	SASA of i	SC-SASA of i	SASA of i_{nbh}	SC-SASA of i_{nbh}		
SAP-adjacent-overall	≥ 0.15	50	10	50	10	50	10	50	10	50	10	50	10	any type, hydrophobic-polar-charged, hydrophobic-charged, charged & charged-polar	**≤ 5 Å**

TABLE S10: The combinations of various conditions that are used to engineer **SAP_adjacent** (specifically, **SAP_adjacent_overall**) features by considering the SAP value of i-th residue and the SASA along with sidechain SASA values of both the i-th residue and its immediate neighbors, which include i+1, i-1, and i_{nbh}. To accurately represent and interpret each feature in the dataset, a distinct nomenclature has been devised. This nomenclature is formulated as **"SAP-adjacent-overall_(SAP_value) _ (SC-SASA of (i and i+1)) _ (SC-SASA of (i and i-1)) _ (SC-SASA of (i and i_{nbh})) _d_(residue-type)"**. For example, the feature denoted as **SAP-adjacent-overall_0.15_10-10_10-10_10-10_5_charged-polar** in the dataset specifically refers to a feature characterized by its unique conditions. 'd' represents the distance from the i-th residue used to determine the proximity of 'i_{nbh}' residue. SAP was derived from an in-house python script, while other calculations were conducted using a custom TCL script through the VMD interface.

Total Gain	
Feature names	**Importance**
sasa_all_0Å	0.438
SAP-adjacent-FE_0.15_5%-5%_5%-5%_5%-5%_7_polar-charged	0.040
sasa_hydrophilic_20Å	0.007
fractional_exposure	0.006
max_CX_0Å	0.005
exposed_charge_all_25Å	0.005
Cover	
Feature names	**Importance**
SAP-adjacent-FE_0.15_5%-5%_5%-5%_5%-5%_7_polar-charged	0.071
sasa_gly_0Å	0.021
sasa_all_0Å	0.020
SAP-adjacent-SC_0.15_10-10_10-10_10-10_7_polar-charged	0.018
sasa_proline_0Å	0.015
number_non_polar_sulfurous_15Å	0.011

TABLE S11: The top six descriptors for XGBoost are ranked by their 'Total Gain' and 'Cover,' showing their significant contributions to model accuracy and effectiveness.

<table>
<tr><td>XGBoost:

'n_estimators': [100, 200, 500],
'max_depth': [6, 10, 15],
'learning_rate': [0.01, 0.1, 0.3],
'subsample': [0.5, 0.8, 1],
'colsample_bytree': [0.3, 0.5, 0.8, 1],
'gamma': [0, 0.1, 0.2, 0.3, 0.4, 0.5],
'reg_alpha': [0, 0.1, 1, 10],
'reg_lambda': [1, 5, 10],
'scale_pos_weight': [1, num_neg / num_pos, num_neg / 2*num_pos, num_neg*2 / num_pos]

num_neg = number of negative (labeled as 0), and num_pos = number of positive samples in the dataset's target variable.</td><td>MLP:

'hidden_layer_sizes': [(50,), (100,), (150,), (200,), (50, 50), (100, 50), (50, 100), (100, 100), (150, 100), (100, 50, 30), (100, 100, 100)],
'activation': ['tanh', 'relu'],
'solver': ['sgd', 'adam'],
'alpha': [0.0001, 0.001, 0.01, 0.1, 1, 10],
'learning_rate_init': [0.001, 0.01, 0.1, 1, 10],
'early_stopping': [True, False].</td></tr>
<tr><td>RF:

'n_estimators': [10, 50, 100, 200, 300],
'max_depth': [None, 10, 50, 100],
'min_samples_split': [2, 5, 10],
'criterion': [entropy],
class_weight': [weights_1, weights_2, weights_3, weights_4, weights_5, weights_6]}

Calculate class counts
Neg and pos refer to number of negative (labeled as 0) and positive samples in the dataset, respectively.

# Define class weights
weights_1 = {0: 1, 1: neg / pos}
weights_2 = {0: 1, 1: 10}
weights_3 = {0: 1, 1: 2}
weights_4 = {0: 1, 1: neg / (2 * pos)}
weights_5 = 'balanced'
weights_6 = {0: 1, 1: 1}</td><td>KNN:

'n_neighbors': [3, 5, 7, 9, 11, 21, 31, 41, 51],
'weights': ['uniform', 'distance'],
'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute'],
'metric': ['euclidean', 'manhattan', 'minkowski'],
'p': [1, 2, 3].</td></tr>
<tr><td colspan="2">SVM:

'kernel': ['rbf', 'linear'], 'C': [0.1, 1, 10, 40, 50, 60, 100], 'gamma': [0.001, 0.01, 0.1, 1, 10], ‘class_weight': [None, 'balanced'].</td></tr>
</table>

TABLE S12: The ranges of hyperparameters used for various models in this study.

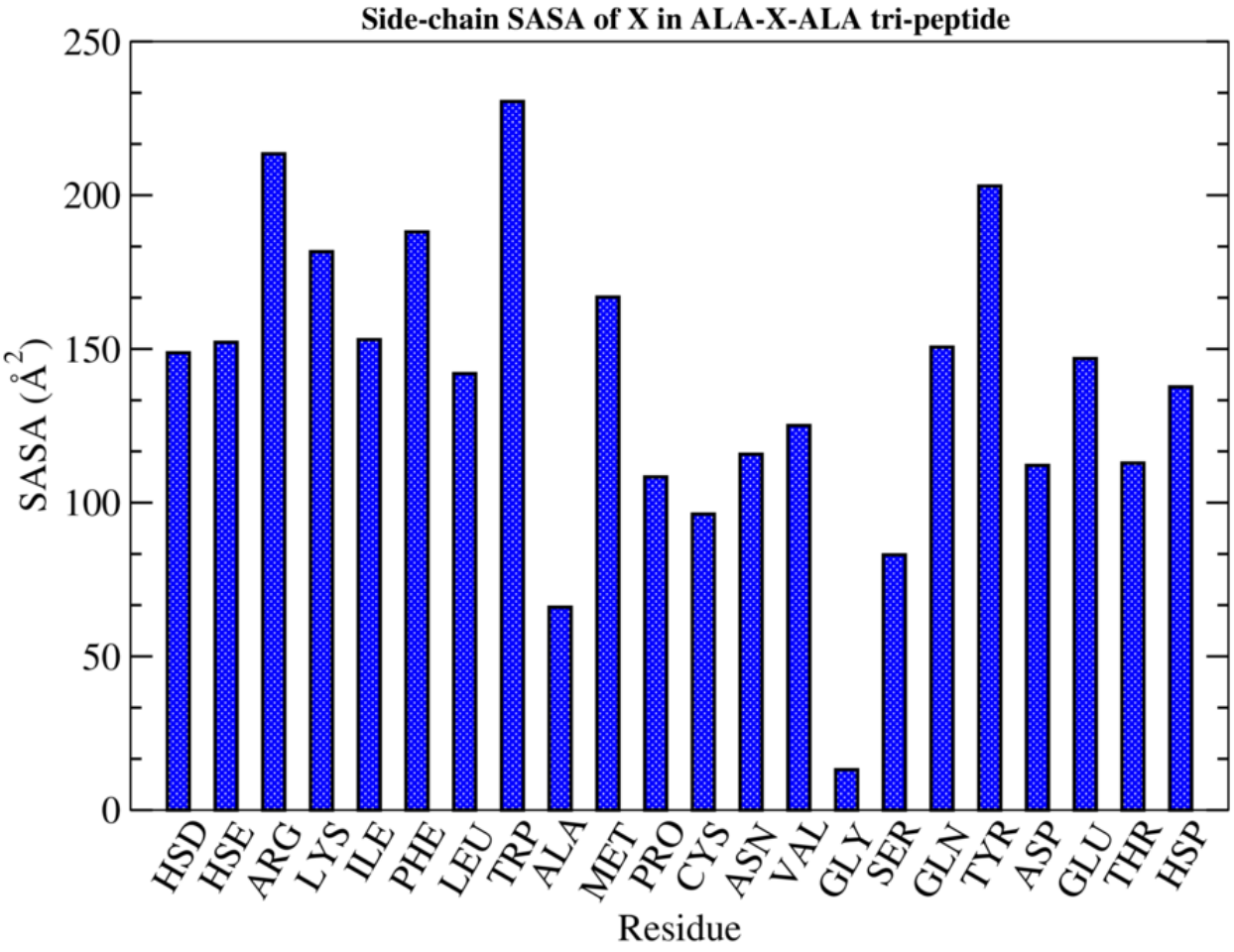

FIGURE S1. Sidechain SASA (Solvent Accessible Surface Area) of fully exposed amino acids (X) in ALA-X-ALA tripeptide. In this configuration, the 'X' residue is flanked by two small alanine (ALA) residues, allowing for maximum exposure of the 'X' sidechain, ideal for studying its intrinsic properties without hindrance from neighboring residues.

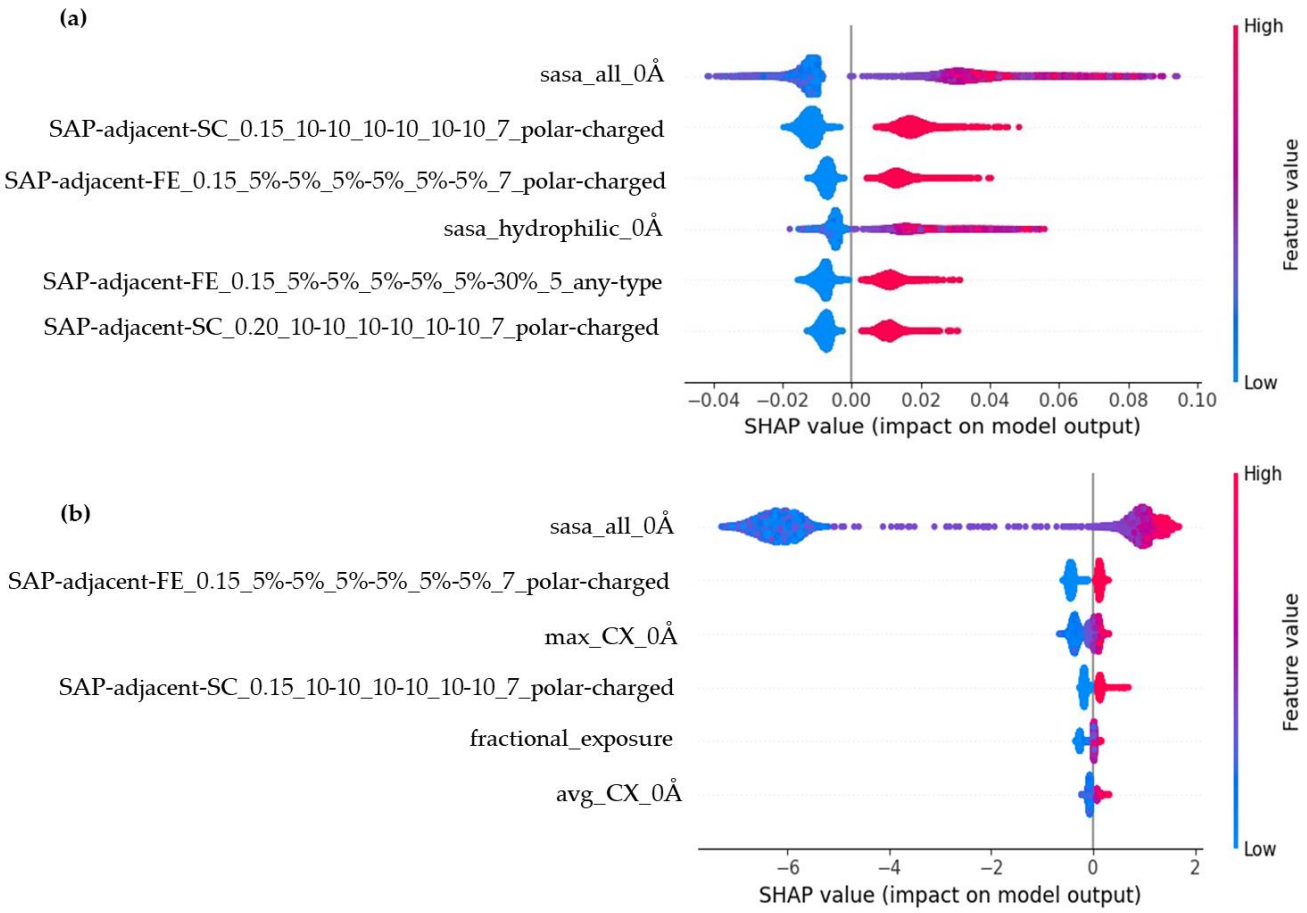

FIGURE S2. Summary plot of SHAP values depicting the impact of top six features on model predictions for both RF and XGBoost models, as shown in (a) and (b), respectively. The plot ranks feature by importance, measured by the magnitude of SHAP values. The summary plot indicates the relationship between the value of a feature and the impact on the prediction. Positive SHAP values indicate an increase in the likelihood of the target variable, while negative values suggest a decrease. Each point represents a SHAP value for a feature in an individual prediction, demonstrating how different features contribute to the model's decision-making process across the dataset. The color of each point on the graph represents the value of the corresponding feature, with red indicating high values and blue indicating low values.

Feature Name	Importance
SAP-adjacent-FE_0.15_5%-5%_5%-5%_5%-5%_7_polar-charged	0.4793256
sasa_all_0Å	0.07736088
sasa_gly_0Å	0.011598438
sasa_hydrophilic_20Å	0.006058179
SAP-adjacent-SC_0.15_10-10_10-10_10-10_7_polar-charged	0.004077815
fractional_exposure	0.00308073
sasa_all_back_20Å	0.002906246
max_CX_0Å	0.002819005
sasa_all_side_20Å	0.002777273
number_long_15Å	0.002430779
sasa_hydrophilic_0Å	0.00236033
sasa_all_back_25Å	0.002327172
avg_CX_0Å	0.002205659
target_sasa_multi_hydrophobicity_25Å	0.001985783
total_sasa_scale_0Å	0.001941003
total_polarizability_0Å	0.00183561
total_SCRS_0Å	0.001790181
total_volume_0Å	0.001705446
sasa_all_20Å	0.001689055
number_proline_5Å	0.001511506
sasa_non_polar_sulfur_containing_10Å	0.001480421
SAP-adjacent-SC_0.15_10-10_10-10_10-75_5_hydrophobic-charged-polar	0.001448556
number_hydrophilic_20Å	0.001371586
max_DPX_20Å	0.001347806
SAP-adjacent-FE_0.15_5%-5%_5%-5%_5%-20%_5_any-type	0.001334814
sasa_aliphatic_25Å	0.001310337
sasa_side_non_polar_sulfur_containing_15Å	0.00127796
sasa_non_polar_sulfur_containing_15Å	0.001277644
SAP-adjacent-SC_0.20_10-10_10-10_10-10_5_any-type	0.00127404
max_DPX_25Å	0.001250023
SAP-adjacent-overall_0.15_10-10_10-10_10-10_5_any-type	0.001242776
number_hydrophilic_15Å	0.001224435
number_polar_25Å	0.001222482
number_very_small_25Å	0.001217644
number_negative_20Å	0.001217383
sasa_aromatic_20Å	0.001209025
number_non_polar_sulfur_containing_15Å	0.001204407
sasa_side_non_polar_sulfur_containing_5Å	0.001170068
exposed_charge_all_25Å	0.001142949
sasa_side_non_polar_sulfur_containing_10Å	0.001140983

sasa_aromatic_15Å	0.001135104
sasa_polar_uncharged_with_amide_5Å	0.001130044
number_non_polar_sulfur_containing_10Å	0.001119722
number_gly_5Å	0.001108185
number_aromatic_25Å	0.001106855
sasa_side_polar_uncharged_with_hydroxyl_group_25Å	0.001106076
number_aliphatic_25Å	0.001090804
SAP-adjacent-overall_0.15_10-10_10-10_10-10_5_polar-charged	0.001088438
number_small_25Å	0.001082931
number_positive_15Å	0.001080008
SAP-adjacent-SC_0.15_10-10_10-10_10-10_7_hydrophobic-charged	0.001079795
SAP-adjacent-FE_0.15_5%-5%_5%-5%_5%-10%_5_any-type	0.001076425
sasa_aromatic_25Å	0.001067896
sasa_small_25Å	0.001067133
sasa_side_non_polar_sulfur_containing_25Å	0.001059395
number_negative_15Å	0.001057101
sasa_polar_uncharged_with_amide_20Å	0.0010523
number_gly_15Å	0.001048368
max_CX_25Å	0.00104088
sasa_negative_10Å	0.001040199
sasa_side_non_polar_sulfur_containing_20Å	0.001037681
total_helix_0Å	0.00103701
max_DPX_15Å	0.00103605
sasa_polar_uncharged_with_amide_0Å	0.001033852
sasa_side_polar_uncharged_with_amide_5Å	0.001033396
sasa_long_25Å	0.001031452
number_proline_15Å	0.001029027
scm_all_25Å	0.001025464
sasa_all_25Å	0.001024732
avg_DPX_20Å	0.00101887
sasa_non_polar_sulfur_containing_25Å	0.0010124
sasa_all_15Å	0.001010734
scm_pos_25Å	0.001006548
SAP-adjacent-overall_0.15_10-10_10-10_10-10_5_hydrophobic-charged-polar	0.001003101
sap_10_20Å	0.001002953

TABLE S13: The Top 75 descriptors and their "gain" importances for XGBoost model, listed in order from highest to lowest importance value.