ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

Predictive Models and Impact of Interfacial Contacts and Amino Acids on Protein–Protein Binding Affinity

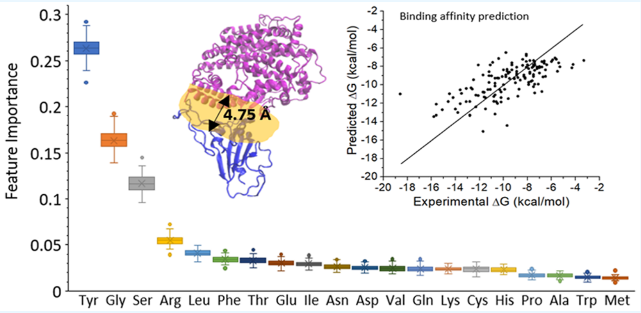
Carey Huang Yi, Mitchell Lee Taylor, Jesse Ziebarth, and Yongmei Wang*

Cite This: *ACS Omega* 2024, 9, 3454–3468

Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: Protein–protein interactions (PPIs) play a central role in nearly all cellular processes. The strength of the binding in a PPI is characterized by the binding affinity (BA) and is a key factor in controlling protein–protein complex formation and defining the structure–function relationship. Despite advancements in understanding protein–protein binding, much remains unknown about the interfacial region and its association with BA. New models are needed to predict BA with improved accuracy for therapeutic design. Here, we use machine learning approaches to examine how well different types of interfacial contacts can be used to predict experimentally determined BA and to reveal the impact of the specific amino acids at the binding interface on BA. We create a series of multivariate linear regression models incorporating different contact features at both residue and atomic levels and examine how different methods of identifying and characterizing these properties impact the performance of these models. Particularly, we introduce a new and simple approach to predict BA based on the quantities of specific amino acids at the protein–protein interface. We found that the numbers of specific amino acids at the protein–protein interface were correlated with BA. We show that the interfacial numbers of amino acids can be used to produce models with consistently good performance across different data sets, indicating the importance of the identities of interfacial amino acids in underlying BA. When trained on a diverse set of complexes from two benchmark data sets, the best performing BA model was generated with an explicit linear equation involving six amino acids. Tyrosine, in particular, was identified as the key amino acid in controlling BA, as it had the strongest correlation with BA and was consistently identified as the most important amino acid in feature importance studies. Glycine and serine were identified as the next two most important amino acids in predicting BA. The results from this study further our understanding of PPIs and can be used to make improved predictions of BA, giving them implications for drug design and screening in the pharmaceutical industry.



1. INTRODUCTION

Protein–protein interactions (PPIs) play a central role in nearly all cellular processes, from DNA replication to biomolecule transport and body immune response.[1−4] Abnormal protein–protein interactions are associated with many diseases such as cancer, neurodegenerative diseases, and infectious diseases.[5−7] Characterizing PPIs is therefore crucial for understanding biological processes and designing therapeutic drugs. PPIs are complicated processes, and it is the binding affinity (BA) of the interactions that determines whether proteins form stable or transient complexes and perform functions.[8] Although crystallography techniques are able to provide structural understanding of PPIs at the atomic level, these techniques do not directly measure BA.

BA is measured as the change in free energy (ΔG) during formation of a complex, where a negative value of a large magnitude indicates a strong interaction. ΔG is related to the dissociation constant (K_d) by the formula $\Delta G = RT \ln(K_d)$ where R is the ideal gas constant, T is the temperature, and ΔG is the BA. BA can be measured in terms of dissociation constant (K_d) using experimental approaches such as surface plasmon resonance (SPR) and isothermal titration calorimetry (ITC).[9] In drug discovery, therapeutic design, and the development of inhibitors, predictions of the BA of proposed interactions are crucial for the identification of leading candidates.[10−12] For example, computational approaches, such as molecular docking, can reveal numerous theoretically possible protein–protein complexes within cells, and BA prediction is needed to evaluate which of these proposed structures are likely to be stable. However, the complexity of PPIs and the large diversity of protein structures and functions makes it difficult to predict BA and understand the features of the complex that impact BA.

Received: September 13, 2023
Revised: December 11, 2023
Accepted: December 14, 2023
Published: January 11, 2024

ACS Publications

https://doi.org/10.1021/acsomega.3c06996

During the past two decades, three general classes of computational approaches have been used to predict BA:[13,14] (1) physics-based ab initio methods such as free energy perturbation and thermodynamics integration;[15,16] (2) knowledge-based statistical potentials such as DFIRE[17,18] or other empirical scoring functions such as Kdeep;[19] and (3) machine learning (ML)-based approaches based on structural and chemical properties of the protein−protein complexes[20−29] or their amino acid sequences.[30−34] Particularly, ML has become a sought-after approach for developing predictive models due to its high power of uncovering insights in large volumes of data.[35] The most common features in predictive ML models with three-dimensional (3D) structures are the contacts at the complex interface, a method exemplified by Vangone and Bonvin with the development of the PRODIGY model and Web server in mid 2010s.[21−23] Compared to other ML models, PRODIGY is simple and interpretable as it employs a linear regression (LR) algorithm, using the number of interfacial contacts (ICs) of polarity-classified residues and noninteracting surface (NIS) residue properties as the major features. The PRODIGY model outperformed many preceding models with a Pearson's correlation coefficient (R) of 0.73 when trained on a data set of 81 complexes. But, a recent study showed that PRODIGY's performance on a different BA database was much lower (R = 0.31),[27] suggesting the need for further investigation of contact-based features and predictive models.

In this work, we compare a series of LR models involving residue/atomic contacts and find that several different approaches to defining interfacial contacts have roughly similar performance when predicting experimental binding affinity. The highest performing of these models is based on a novel set of features with the quantitative information on the identities of the amino acids at the binding interface. As of today, there have been no studies that use the identities of specific amino acids to build models for BA prediction. Substantial studies have helped us to have a good understanding of the composition and role of specific amino acids at the complex interface for protein−protein binding.[36−43] However, translation of the information on the specific amino acids at a binding interface into a BA prediction model has not been achieved. Here, we report a quantitative association of the interfacial number (INs) of amino acids (AA) with BA and use this knowledge to develop an ML-based model for BA prediction. We found a group of high-impact amino acids, led by tyrosine, that had an association with BA comparable to the most impactful residue or atomic contact features. We also show that this AA-INs model, which uses only six amino acids, outperforms other classes of contact-based LR models on benchmark data sets, indicating the importance of the identities of amino acids at protein interfaces in underlying the strength of protein−protein BA.

2. METHODS

2.1. Features and Their Calculation. The contact features that we investigated included residue-based interfacial contacts (residue ICs) where a contact was defined between two residues if any of their heavy atoms were within a defined cutoff, atom-based interfacial contacts (atomic ICs) where any atoms that were within the cutoff were counted (not limited to one contact per residue), and hydrogen bonds (HB) where we used a cutoff of 3.5 Å between any N and O atoms. The residue ICs were further divided into different types based on charge and polarity of the amino acids. We used the following classifications: charged (Arg, Asp, Glu, His, Lys), polar (Asn, Gln, Ser, Thr), and apolar (Ala, Cys, Gly, Ile, Leu, Met, Phe, Pro, Trp, Tyr, Val). Atomic ICs were divided into three types based on the polarity of atoms, where C is considered as nonpolar and N, O, S as polar. We also considered the noninteracting surface (NIS) of the complex as it has been shown to contribute to BA.[44] NIS features were calculated as the percentage of surface area contributed by the polar, apolar, and charged residues over the total NIS area.[21]

Additionally, we considered the hydrophobicity of the contacting residues by defining a hydrophobicity score (HS) for each complex. HS was defined according to

$$\mathrm{HS} = \sum \left(1 - \frac{|\mathrm{KD}_i - \mathrm{KD}_j|}{4.5}\right) \tag{1}$$

where the KD values for the two contacting residues, ith and jth, are the hydropathy indexes from the Kyte-Doolittle scale[45] and the score is summed over all pairs of contacting residues in the complex. The division by 4.5 and subtraction from 1 ensures that the score for each residue pair ranges between −1 and 1. With this definition, contacts between residues with similar hydrophobicity will contribute positive values near +1 to HS, while contacts between unlike residues have negative contributions to HS.

Thus, a total of 14 residue and atomic features were investigated and were indexed sequentially as (0) residue ICs_charged-charged, (1) residue ICs_charged-polar, (2) residue ICs_charged-apolar, (3) residue ICs_polar-polar, (4) residue ICs_polar-apolar, (5) residue ICs_apolar-apolar, (6) HB, (7) atomic ICs_polar-polar, (8) atomic ICs_polar-apolar, (9) atomic ICs_apolar-apolar, (10) %NIS_polar, (11) %NIS_apolar, (12) %NIS_charged, and (13) HS. Residues ICs, atomic ICs, HB, and HS were calculated with Python, utilizing the Bio.PDB package within the Biopython library to extract the features from PDB files.[46] %NIS was calculated using NACESS with the same configurations used in the PRODIGY model by Vangone and Bonvin.[21,47]

Additionally, the interfacial number (IN) of each amino acid (AA), AA-IN, for each complex was counted using the distance cutoff optimized by residue contacts. The AA-INs of the 20 standard amino acids were then used as features to develop LR models and investigate the impact of specific amino acids on BA.

The scripts for calculating these features and using the final AA-INs model based on six amino acids for predicting BA of any protein complex with 3D structures are available at https://github.com/kagrat17/AAIN_Predictor. The code to calculate % NIS for the contact-based models was taken from the PRODIGY model by Vangone and Bonvin.[22] The distance cutoff for the residue and atomic contacts was optimized by investigating the Pearson's correlation between the predicted ΔG and experimental ΔG at different distances.

2.2. Data Sets. To build structure-based models for predicting protein−protein BA, a diverse and reliable data set of solved protein−protein structures with reliable affinity data is required. Vangone and Bonvin built a data set by "cleaning" the structure-based protein−protein BA benchmark[48] and used it to develop the PRODIGY model and Web server.[21,22] This data set, referred to as the PRODIGY data set, contains 81 complexes with diverse functions and a wide range of experimental BA (ΔG from −18.6 to −4.3 kcal/mol). However, a later study by Romero-Molina et al.[27] showed

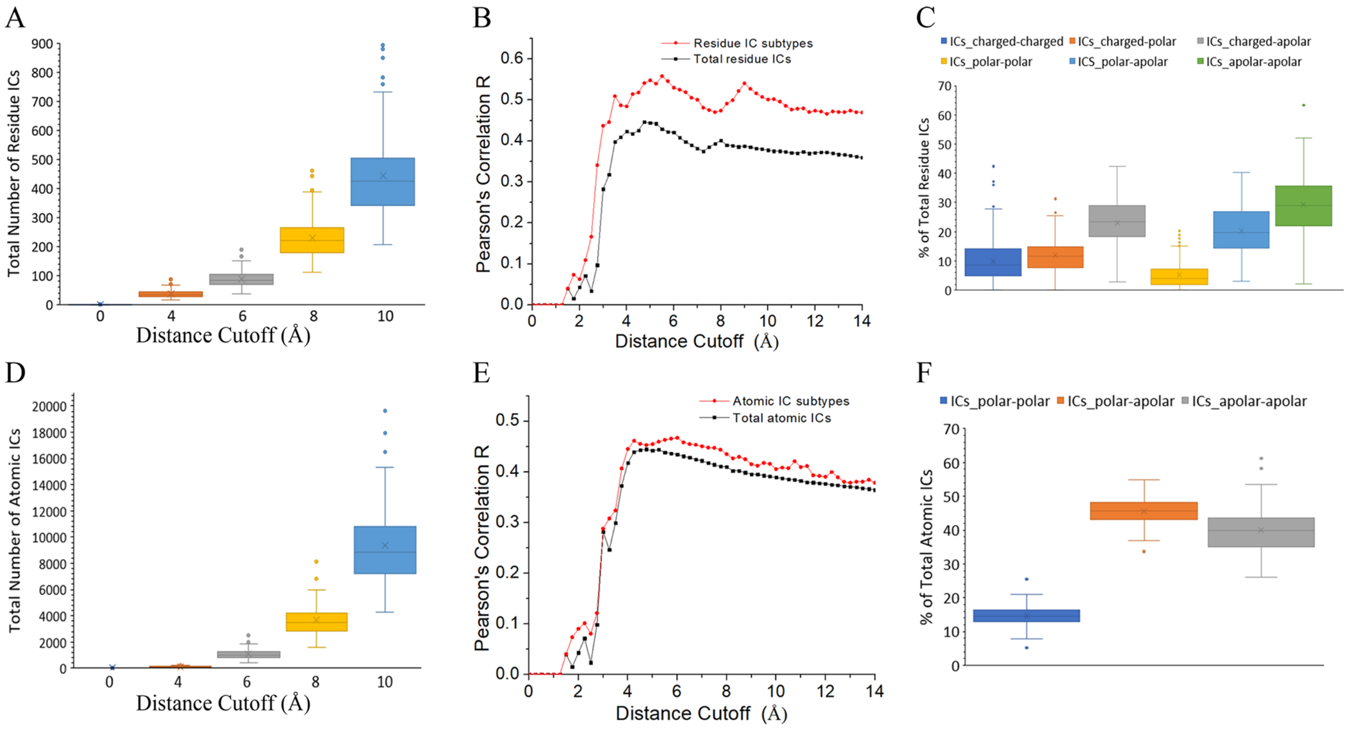


Figure 1. Dependence of the number of ICs and predictive performance of LR models on the distance cutoff and polarity-classified IC subtypes. (A) The total number of residue ICs at different distance cutoffs. (B) Comparison of the correlation between predicted and experimental DG at different distance cutoffs for the LR models generated with the features of total number of residue ICs (black curve) and the six residue IC subtypes (red curve). (C) Proportion of residue ICs subtypes in the total number of residue ICs at 4.75 Å distance cutoff. (D) The total number of atomic ICs at different distance cutoffs. (E) Comparison of the correlation between predicted and experimental DG at different distance cutoffs for the LR models generated with the features of total number of atomic ICs (black curve) and the three atomic IC subtypes (red curve). (F) Proportion of atomic ICs subtypes in the total number of atomic ICs at 4.75 Å distance cutoff.

that PRODIGY did poorly on a different data set that contained 90 complexes gathered from the PDBbind database (v.2020).[49] PRODIGY gave a Pearsons' correlation R of 0.31 on this PDBbind data set in contrast to R of 0.73 on the PRODIGY data set. In this work, we use both data sets as the resource of protein–protein complexes to develop improved models for BA prediction.

It has been shown by Vangone and Bonvin that the reliability of a data set depends on the experimental methods used to determine K_d of the complexes.[21] The SPR, ITC, stopped-flow fluorimetry, and spectroscopic methods were shown to be reliable as the experimental ΔG measured by these methods gave reasonable correlations with residue ICs. Other methods, including inhibition assays and fluorescence spectrophotometry, gave BA with low correlations with residue ICs, possibly because they were indirect methods that were less reliable.[50−52] Complexes with K_d measured with these methods were thus removed by Vangone and Bonvin from the initial benchmark data set. We conducted the same assessment for the PDBbind data set, and the results showed that the experimental ΔG from these other methods indeed gave low correlations with the number of residue ICs and atomic ICs (Table S1). Thus, we excluded these complexes from the data set. We also excluded two other complexes from the PDBbind data set, 2FTL and 2JGZ. The experimental K_d for the complex 2FTL was measured at 100 K rather than room temperature. The complex 2JGZ only had a lower bound K_d (K_d > 1 mM). After removing these complexes, we had two final data sets of size 81 and 60, whose compositions are shown in Table S2. We further examined whether there were significant differences between the two data sets. Specifically, we examined whether the BA distributions of their complexes were significantly different. The two-sample Kolmogorov–Smirnov test (testing whether two samples came from the same distribution) gave a p-value of 0.022, indicating that the distribution of BA differs between the two sets. This can also be seen by the BA histograms shown in Figure S1A. Thus, we decided to combine the two data sets (total of 141 complexes) to make a large, diverse, and reliable data set that can better represent protein–protein complexes. This combined data set has a wide range of BA, with experimental ΔG from −3.3 to −18.6 kcal/mol (Figure S1B).

2.3. Machine Learning Methods.

The statsmodels and sklearn python libraries were used to build and investigate BA models. We used different combinations of features to construct multivariate linear regressions and evaluated them using the coefficient of determination R^2, Pearson's product-moment correlation coefficient R (which is the square root of R^2 for linear regressions), and the Akaike Information Criterion (AIC). In addition, root-mean-square error (RMSE) was calculated to evaluate the average difference between the predicted and experimental ΔG. We also built random forest (RF) models to rank the features of the models by their impurity-based feature importance. Cross-validation was also performed to evaluate our models by partitioning the data set into four subsets, training on 75% of the data (training set) and validating on the other 25% of the data (validation set) and repeating until each of the 4 subsamples was used as the validation set. Such 4-fold cross-validation was repeated 10 times, which gave mean test R with standard deviation.

Additionally, we also trained Support Vector Machine (SVM) and AdaBoost models on the combined data set to compare the AA-INs model's performance with more advanced ML techniques. For the SVM model, the Support Vector Classifier from scikit learn was used along with the Radial Basis Function (RBF) kernel, which was the only standard kernel found to produce similar results as the Random Forest and AdaBoost models. For the AdaBoost model, we used the AdaBoostRegressor from scikit-learn with 20 estimators and a learning rate of 0.1.

3. RESULTS

3.1. Sensitivity of Contact-Based Prediction of BA on the Distance Cutoff Used to Define Contacts.

The simplest method of using protein−protein contacts to predict binding affinity is by creating a linear model using only the total number of interfacial contacts (ICs) of the residues from the two protein components. The total number of ICs increased rapidly with the distance cutoff used to define the contact (Figure 1A). To examine whether the model performance was sensitive to the cutoff used to calculate the number of ICs, we built LR models with the total number of ICs calculated at different cutoffs and examined the Pearson's correlation R between predicted and experimental ΔG of these models. The value of R and, thus, the predictive ability of the number of ICs, increased as the cutoff distance increased above 3 Å, reaching a relative plateau from ∼4 to 6 Å and then gradually decreasing as the cutoff continued to increase (Figure 1B, black curve). The maximum value of R occurred at a cutoff of 4.75 Å (R = 0.45), but the correlation was not particularly sensitive to the distance cutoff as distances of between ∼4 and 6 Å provided similar performance. Thus, a cutoff of 4.75 Å was used to calculate residue ICs for further studies.

Based on the polarity of residues, Vangone and Bonvin classified the residue ICs into six subtypes, ICs_charged-charged, ICs_charged-polar, ICs_charged-apolar, ICs_polar-polar, ICs_polar-apolar, and ICs_apolar-apolar, and showed that an LR model including subtype ICs performed better than the model using total ICs.[21] To examine whether this applies to a larger and more diverse data set, we examined the relationship between the distance cutoff used to define a contact and the correlation between experimental ΔG and ΔG predicted by an LR model with these six features using the combined PRODIGY and PDBbind data set. The general behavior of this relationship was similar to that found using only the total number of ICs, as R plateaued for cutoff distances between ∼4 to 6 Å and reached a maximum R of 0.54 at the same cutoff distance 4.75 Å (Figure 1B, red curve). However, in agreement with Vangone and Bonvin's results, we found that the model performance improved when splitting the total contact number into the numbers of IC subtypes, with R values increasing by about 25%. Among the six ICs subtypes, the percentage of residue ICs_apolar-apolar was highest, accounting for nearly 30% (Figure 1C). The percentage of ICs_polar-polar was lowest, only about 5%.

While the previous discussion has focused on contacts between the residues in protein−protein complexes, these residues actually interact with each other through atoms. Thus, we also investigated the correlation between the number of atom−atom contacts and binding affinity to compare residue-based and atomic-based approaches for defining protein−protein contacts. Defining contacts on an atomic basis greatly increased the number of contacts that were identified in the complexes, as there was over an order of magnitude increase in atomic contacts in the complexes compared with residue contacts (Figure 1D). The relationship between the cutoff distance used to define a contact and the correlation between the number of atomic contacts and the experimental ΔG (Figure 1E, black curve) was similar to that found for residue ICs, showing a plateau around 5 Å. The peak R for atomic ICs occurs at the same cutoff (4.75 Å) as that found for residue ICs, and the R value at the peak was similar for the two approaches (R = 0.44 for atomic ICs and R = 0.45 for residue ICs). Similar to how residue contacts can be classified based on the classes of the interacting residue, the atomic ICs can be grouped into three types (polar-polar, polar-apolar, and apolar-apolar) based on the polarity of atoms, where C is considered as nonpolar and N, O, S as polar. Different from the residue contacts, we found that the correlation of the LR models based on the three atomic subtypes was only slightly improved at each distance cutoff compared to those using the total atomic ICs (Figure 1E, red curve), with R only reaching 0.45 at the optimal distance cutoff, a value only slightly above that of the LR model with the total atomic ICs. Among the three atomic ICs subtypes, the highest contribution to the total atomic ICs was from atomic ICs_polar-apolar, with an average of 45.5% (Figure 1F). This was followed by atomic ICs_apolar-apolar (40%). The atomic ICs_polar-polar had the lowest contribution, accounting for only 14.5% of total atomic ICs.

3.2. Correlation between Residue/Atomic Features and Experimental BA.

To further investigate and compare the relationship between residue/atomic ICs and BA, we calculated the Pearson's correlation R between individual features and the experimental ΔG (Table 1). The majority of these features had a negative correlation with ΔG and thus will lead to stronger BA. Several individual features were able to produce correlations with experimental ΔG comparable to total residue or total atomic ICs. Specifically, atomic ICs_polar-apolar, atomic ICs_polar-polar, residue ICs_polar-apolar, and HS all had R values of −0.4 or less, near the R ∼ −0.45 achieved by the total residue or atomic ICs. The strong

Table 1. Pearson's Correlation R between the Experimental ΔG and Residue ICs, Atomic ICs, HB, HS, and %NIS Features

Residue-ICs	R	p-value
Residue ICs_charged−charged	−0.12	0.078
Residue ICs_charged-polar	−0.21	0.006
Residue ICs_charged-apolar	−0.39	<0.0001
Residue ICs_polar−polar	−0.15	0.038
Residue ICs_polar−apolar	−0.44	<0.0001
Residue ICs_apolar−apolar	−0.19	0.012
Total residue ICs	−0.45	<0.0001
HS	−0.41	<0.0001
Atomic ICs		
Atomic ICs_polar−polar	−0.44	<0.0001
Atomic ICs_polar−apolar	−0.45	<0.0001
Atomic ICs_apolar−apolar	−0.37	<0.0001
Total atomic ICs	−0.45	<0.0001
HB	−0.33	<0.0001
%NIS		
%NIS_polar	−0.35	<0.0001
%NIS_apolar	0.09	0.144
%NIS_charged	0.30	0.0002

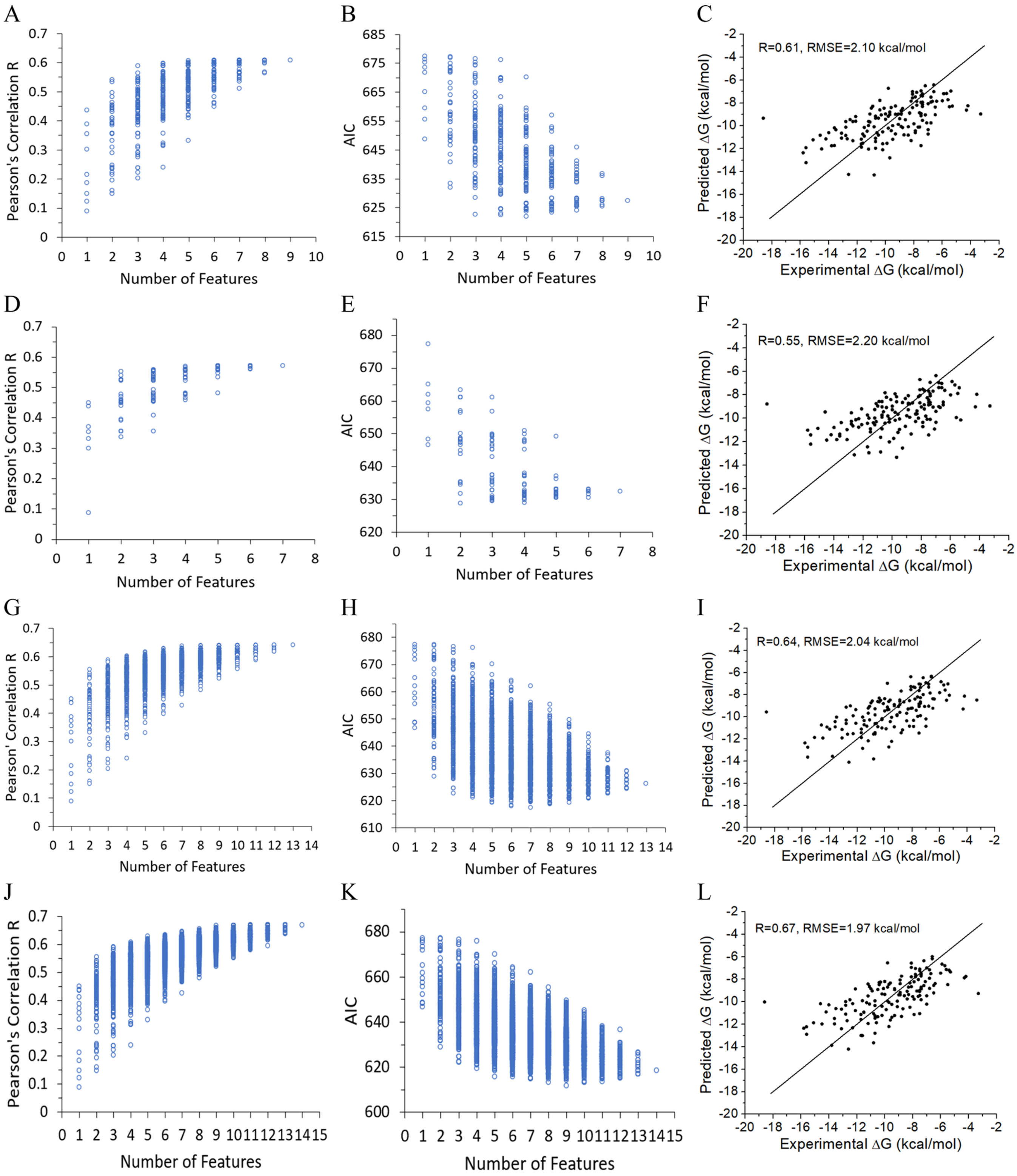


Figure 2. Comparison of the performance of different groups of contact-based LR models. (A–C) Residue ICs/NIS models. (D–F) Atomic ICs/NIS models. (G–I) Residue/atomic ICs/NIS models. (J–L) Residue/atomic ICs/HS/NIS models. (A, D, G, J) Pearson's correlation R as a function of the number of features used to generate the models. (B, E, H, K) AIC as a function of the number of features used to generate the models. (C, F, I, L) Scatter plots between the predicted and experimental ΔG for the minimum AIC models. The straight lines are the function $y = x$.

correlation between experimental ΔG and residue ICs polar-apolar was previously reported by Vangone and Bonvin with the PRODIGY data set.[19] Other features with relatively strong correlations with experimental ΔG included residue ICs charged-apolar, atomic ICs apolar-apolar, %NIS_polar, and HB. %NIS_charged had a relatively strong positive correlation with ΔG (R = 0.30), indicating that increasing the percentage of NIS from charged residues would lead to weaker BA.

One somewhat surprising result from this analysis is that contacts between "like" residues or "like" atoms (e.g., polar−polar residue contacts or apolar−apolar atomic contacts) had weaker correlations with BA than "unlike" contacts. For example, polar−apolar residue and atomic contacts had the strongest correlation with experimental ΔG in their respective categories, indicating that these "unlike" interactions were the best at predicting BA. To further investigate "like" and "unlike" contacts, we calculated the expected percentage (average from 141 complexes) of residue pairs in contacts based on the percentages of each residue class in contacts and compared these expected percentages with their actual values (Table S3). All "like" residue pairs (charged-charged, polar-polar, and apolar-apolar) had higher percentages than their expected values by ∼10%. Thus, there does seem to be a preference for the formation of "like" contact pairs at the protein interface, but this preference does not translate to these contacts being better predictors of binding affinity.

3.3. Comparison of Residue, Atomic, and Combined Feature Sets in Creating LR Models of BA.

To more fully explore the contact-based features and predictive models, we built and evaluated a series of LR models with different feature combinations. First, we used the set of features used in the development of the PRODIGY model to build residue ICs/NIS models to examine how these models performed on the expanded data set used in this work that contained the 81 complexes from the PRODIGY data set and additional 60 complexes from PDBbind. Specifically, the features used here included six types of residue ICs (Feature Index 0,1,2,3,4,5) and three %NIS (Feature Index 10,11,12). We trained a total of 511 LR models using all possible combinations of these nine features on our data set. The performance of these models was evaluated by both Pearson's correlation R (Figure 2A) and AIC (Figure 2B). The AIC criteria was used to correct for overfitting and gave more weight to simpler models. The results showed that increasing the number of the features from 1 to 3 led to a rapid increase of the model performance. The maximal R ($R_{\max}$) increased by 0.15 and the minimal AIC ($\mathrm{AIC}_{\min}$) decreased by 26.1 from single-feature models to three-feature models. However, as the number of features included in the model increased above 5, the $R_{\max}$ had negligible changes (remaining near 0.61), while the $\mathrm{AIC}_{\min}$ increased as AIC penalized the increasing complexity of these models. This residue ICs/NIS model with the lowest AIC (AIC = 622.0) gave R of 0.61 ($p < 0.0001$) and RMSE of 2.10 kcal/mol (Figure 2C). It included the following 5 features: residue ICs_charged-charged, residue ICs_charged-apolar, residue ICs_polar-polar, residue ICs_polar-apolar, and % NIS_polar. While this model had the lowest AIC, many models performed relatively well, with 153 of the 511 models giving $R > 0.55$.

Next, we built a set of LR models based on atomic contacts to compare with residue-based models. These atomic contact-based models included all possible combinations of 7 features, HB, atomic ICs_polar-polar, atomic ICs_polar-apolar, atomic ICs_apolar-apolar features, and the three %NIS features (Feature Index 6,7,8,9,10,11, 12), resulting in a total of 127 LR models. The changes of R and AIC as a function of feature number for these atomic IC/NIS models had similar trends as residue ICs/NIS models, as $R_{\max}$ and AIC values soon reached a plateau as more than two features were added to models (Figure 2D,E). The $R_{\max}$ increased from 0.45 for single-feature model to 0.55 for two-feature model, and only increased to 0.57 for the all-feature model. The model with the minimum AIC (AIC = 628.8) included two features, ICs_polar−apolar and %NIS_polar and gave $R = 0.55$ ($p < 0.0001$) and RMSE = 2.20 kcal/mol (Figure 2F). Again, many atomic ICs/NIS models (>20%) performed similarly to the minimum AIC model.

Then, we examined whether combining residue-based and atomic-based contact features of protein complexes would improve predictions of their experimental ΔG. The combination of residue contacts, atomic contacts, and %NIS features resulted in a total of 13 features and 8191 possible models. Both R and AIC improved significantly as more features were added until 7 features, with $R_{\max}$ remaining at 0.64 for models with 7 to 13 features (Figure 2G,H). The model with the lowest AIC (AIC = 617.3) included a total of 7 features and gave $R = 0.64$ ($p < 0.0001$) and RMSE = 2.04 kcal/mol (Figure 2I). The 7 features in this model included a mix of atomic ICs, residue ICs, and %NIS, which were residue ICs_charged-polar, residue ICs_charged-apolar, residue ICs_polar-polar, residue ICs_polar-apolar, HB, atomic ICs_polar-polar, and %NIS_polar. Similarly, many residue/atomic ICs/NIS performed relatively well, with ∼20% of the models giving $R > 0.60$.

One additional way to characterize residue interactions is to use hydropathy indices of residues. Hydropathy index is a number representing the hydrophobic or hydrophilic properties of its side chain. The larger the number is, the more hydrophobic the amino acid is. Several hydrophobicity scales have been published. We used the commonly used Kyte-Doolittle scale[45] to calculate the hydrophobicity score, HS, for each protein−protein complex. Adding HS to the residue ICs, atomic ICs, and %NIS gives a total of 14 features, resulting in 16,383 possible feature combinations to create models. R reached the maximum of 0.67 for models with 9 to 14 features, and the AIC reached minimum of 611.7 at 9 features (Figure 2J,K). About 30% of the residue/atomic ICs/HS/NIS models gave $R > 0.60$. The minimum AIC model used 9 features: residue ICs_charged-polar, residue ICs_charged-apolar, residue ICs_polar-polar, residue ICs_polar-apolar, residue ICs_apolar-apolar, HB, atomic ICs_polar-polar, % NIS_polar, and HS, with R of 0.67 ($p < 0.0001$) and RMSE of 1.97 kcal/mol (Figure 2L).

The minimum AIC models for the four different groups of features were tested by 4-fold cross-validation and compared with the all-feature model for each class (Table 2). In all cases, the minimum AIC models performed slightly better than the all-feature models during cross-validation. For example, the test R of the minimum AIC residue/atomic ICs/HS/NIS model was 0.61, whereas the test R of the all-feature model was 0.58. The four different feature groups (i.e., residue ICs/NIS, atomic

Table 2. Test R for the Cross-Validation of Minimum AIC and All-Feature Models[a]

	Residue ICs/NIS features	Atomic ICs/NIS features	Residue/atomic ICs/NIS features	Residue/atomic ICs/HS/NIS features
Minimum AIC model	0.55 (0.01)	0.54 (0.02)	0.60 (0.02)	0.61 (0.02)
All-feature model	0.54 (0.02)	0.51 (0.03)	0.53 (0.04)	0.58 (0.02)

[a] Data are presented as the mean with standard deviation from 10× trials.

Table 3. Coefficients of Features in the Linear Equations Predicting ΔG for Minimum AIC Models

	Residue ICs/NIS model		Atomic ICs/NIS model		Residue/atomic ICs/NIS model		Residue/atomic ICs/HS/NIS model	
Feature	Coefficient	*p*-value	Coefficient	*p*-value	Coefficient	*p*-value	Coefficient	*p*-value
Residue ICs_charged-charged	−0.0735	$p = 0.107$						
Residue ICs_charged-polar					0.1106	$p = 0.092$	0.1564	$p = 0.019$
Residue ICs_charged-apolar	−0.1262	$p = 0.002$			−0.1024	$p = 0.014$	−0.1240	$p = 0.004$
Residue ICs_polar-polar	0.1456	$p = 0.154$			0.1811	$p = 0.042$	0.2689	$p = 0.004$
Residue ICs_polar-apolar	−0.1528	$p = 0.001$			−0.0947	$p = 0.032$	−0.1171	$p = 0.013$
Residue ICs_apolar-apolar							0.0601	$p = 0.071$
HB					0.0893	$p = 0.117$	0.1232	$p = 0.031$
Atomic ICs_polar-polar					−0.0680	$p = 0.003$	−0.0588	$p = 0.010$
Atomic ICs_polar-apolar			−0.0181	$p = 0.000$				
Atomic ICs_apolar-apolar								
% NIS_polar	−0.1329	$p = 0.000$	−0.1242	$p = 0.000$	−0.1241	$p = 0.000$	−0.1198	$p = 0.000$
% NIS_apolar								
% NIS_charged								
HS							−0.1102	$p = 0.004$
Constant	−1.796	$p = 0.000$	−2.393	$p = 0.000$	−1.925	$p = 0.000$	−2.131	$p = 0.000$

ICs/NIS, residue/atomic ICs/NIS, and residue/atomic ICs/HS/NIS) provided roughly similar results, as the R values ranged between 0.55 and 0.61. Thus, increasing from the two features included in the atomic ICs/NIS model to the nine-feature atomic/residue ICs/HS/NIS model only increased R from 0.54 to 0.61. The coefficients for the linear equations of the minimum AIC models and the all-feature models with their *p*-values are shown in Table 3 and Table S4, respectively. Generally, the minimum AIC model picked up features that had $p < 0.05$ in the all-feature models.

3.4. Feature Importance Studies Revealed the Impact of Different Residue/Atomic Properties on BA.

To decipher interfacial factors that impact BA, we conducted feature importance studies with several approaches. First, we examined the feature composition in the minimum AIC models for different feature combinations (Figure S2). We then calculated the frequency of feature occurrence in the top models (Figure 3A−D). We chose a minimal number of the top models such that each feature gave a different frequency of occurrence in these models. This was top 105 for the residue ICs/NIS models, top 50 for the atomic ICs/NIS models, top 100 for the residue/atomic ICs/NIS models, and top 200 for the residue/atomic ICs/HS/NIS models. In addition, we predicted the feature importance by RF using the all 14-feature model (Figure 3E). For convenience, the feature index was used in all the plots.

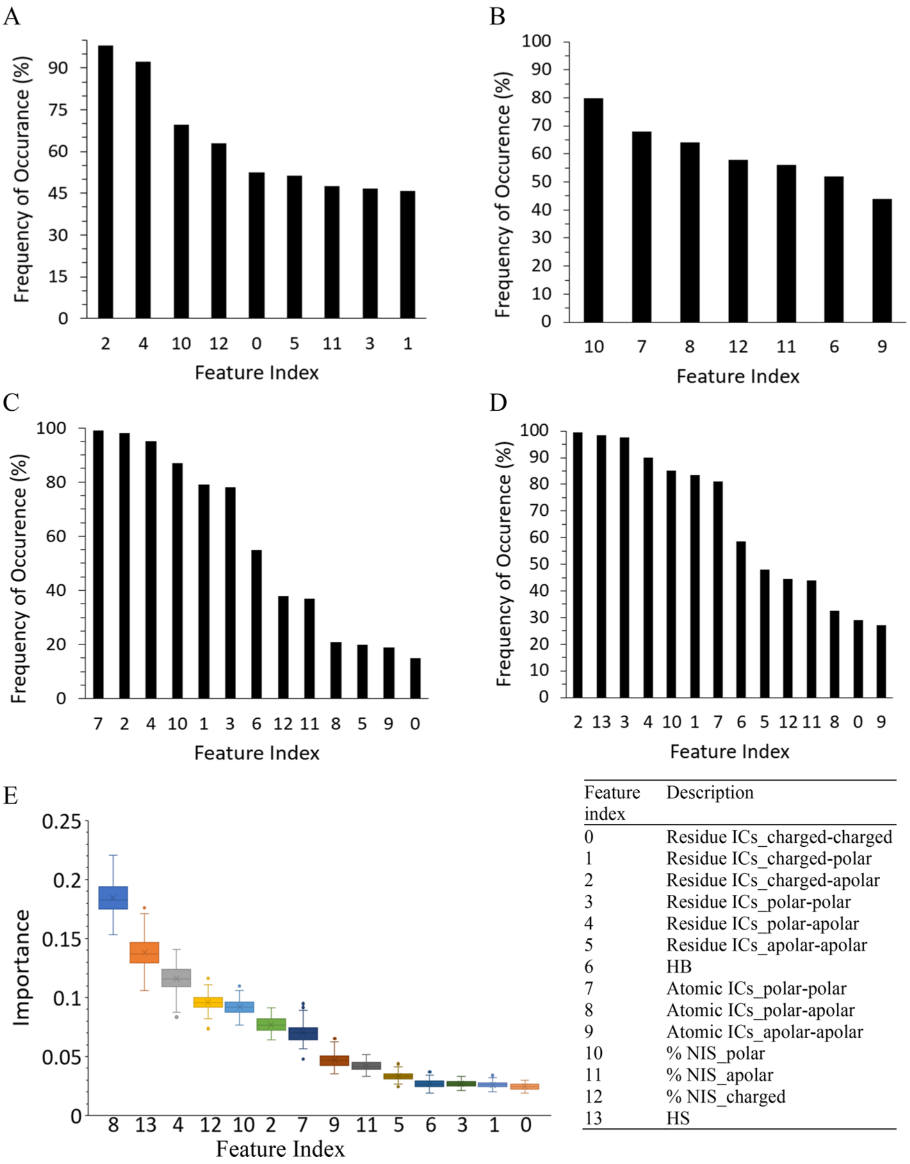
Due to the different features used and the different methods used to identify important features, there were, not surprisingly, large differences in the features that were identified as most important for each feature set. Additionally, there are high correlations between many of the contact features (Table S5). For example, the atomic ICs_polar-polar had high correlation with atomic ICs_polar-apolar (R = 0.93), residue ICs_charged-polar (R = 0.74), and hydrogen bond (R = 0.90). Therefore, the addition of these three features to a model that contained atomic ICs polar-polar would involve redundant information and would not be likely to significantly improve the model. However, a couple of consistent results were identified. First, HS was identified as the second most important feature in both cases when it was included in the model, indicating that including the hydropathy index of contacting residues contained valuable information for predicting experimental ΔG. On the other hand, HB was not identified as one of the five most important features in any of the models and was always ranked behind atomic polar-polar contacts in importance. While HB was included in the minimum AIC model for two feature sets (Table 3), it had a lower *p*-value than atomic polar-polar contacts in both models. Finally, in agreement with our previous discussion on the R values for single-feature LR models, feature importance between “like” pairs of residues and contacts did not seem to consistently be more important features than “unlike” contacts. For example, residue contacts between apolar and either charged or polar residues (Feature Index 2 and 4) were usually identified as being the most important class of residue contacts.

3.5. Interfacial Numbers of Amino Acids, Enrichment, and Depletion.

Inspired by our analysis of the importance of contact features (e.g., the high importance of polar−apolar residue and atomic contacts) and recent results that have shown the importance of specific residues such as tyrosine at the binding interface in protein−protein complexes,[36] we further investigated the amino acids that were found at the binding interface. Specifically, we determined the enrichment or depletion of amino acids at the binding interface compared to their presence in the entire complex (Figure 4 and Table S6). First, we calculated the number and proportion of each amino acid in the entire complex. Next, we calculated the number of each amino acid at the binding interface, AA-IN, and their proportions at the optimal cutoff of 4.75 Å. An enrichment factor was then calculated using the ratio of these two percentages (% at interface/% in entire complex), where enrichment factors >1 indicate that an amino acid had an abundance at the interface that was greater than expected by its abundance in the entire complex.

The results showed that leucine was the most abundant amino acid in protein−protein complexes, accounting for 8.9% of all amino acids. The least abundant amino acid was tryptophan, which only accounted for 1.6%. These results agree with those by others using large data sets (Swissprot and TrEMBL) that showed leucine was the most abundant amino acid whereas tryptophan and cystine are the least abundant,[41] indicating that the amino acid percentages in these protein−protein complexes in our data set are similar to those found in all proteins.

https://doi.org/10.1021/acsomega.3c06996



Feature index	Description
0	Residue ICs_charged-charged
1	Residue ICs_charged-polar
2	Residue ICs_charged-apolar
3	Residue ICs_polar-polar
4	Residue ICs_polar-apolar
5	Residue ICs_apolar-apolar
6	HB
7	Atomic ICs_polar-polar
8	Atomic ICs_polar-apolar
9	Atomic ICs_apolar-apolar
10	% NIS_polar
11	% NIS_apolar
12	% NIS_charged
13	HS

Figure 3. Feature importance of residue ICs, atomic ICs, HB, HS, and %NIS by frequency of occurrence in top models (A–D) and RF (E). (A) Residue ICs/NIS models. (B) Atomic ICs/NIS models. (C) Residue/atomic ICs/NIS models. (D) Residue/atomic ICs/HS/NIS models. (E) Feature importance predicted by RF using the all-feature model.

Of these residues at the binding interface, leucine remained the most abundant amino acid (8.0%), with Leu-IN reaching as high as 51 for the 5YR0 complex. However, the enrichment factor of leucine was <1, as it was more prevalent in the entire complex than it was in the contacts. Tyrosine was the second most abundant amino acid at the binding interface, accounting for 7.5% of all amino acids at the interface. Compared to its abundance in the entire complex (3.8%), the abundance of tyrosine at the interface was doubled, making it the amino acid with the highest enrichment factor. For an individual complex, the highest percentage of tyrosine at the interface was 28.3% (2AJF). Other remarkably enriched amino acids at the contact interface were arginine and tryptophan, which had enrichment factors near 1.65. Alanine had the lowest enrichment factor among the 20 amino acids. It was the fifth most abundant amino acid in the entire complex, but it became the 15th at contact interface.

3.6. AA-INs Are Correlated with Experimental BA. To examine the association of specific amino acids with BA, we determined whether the numbers of individual amino acids at the interface (AA-INs) in each complex were correlated with the experimental ΔG of the complex (Table 4). The R values of the correlations between ΔG and the INs of several amino acids were comparable to the best performing residue and atomic ICs features. Specifically, the INs of tyrosine had a lower R value (R = −0.49, p < 0.0001) than any of the previously discussed contact features, while glycine (R = −0.45, p < 0.0001) and serine (R = −0.44, p < 0.0001) were able to perform as well as any of the residue and atomic ICs features when predicting experimental ΔG. Tyrosine has been shown to play a dominant role for protein−protein binding due to its unique physicochemical properties that make it effective at mediating molecular recognition.[38] Thus, simply using the number of tyrosine residues involved in contacts at the interface of a protein−protein complex is better able to predict the BA of the complex than other features including the total number of atomic or residue ICs or the number of any one of the IC contact subclasses. We also investigated the

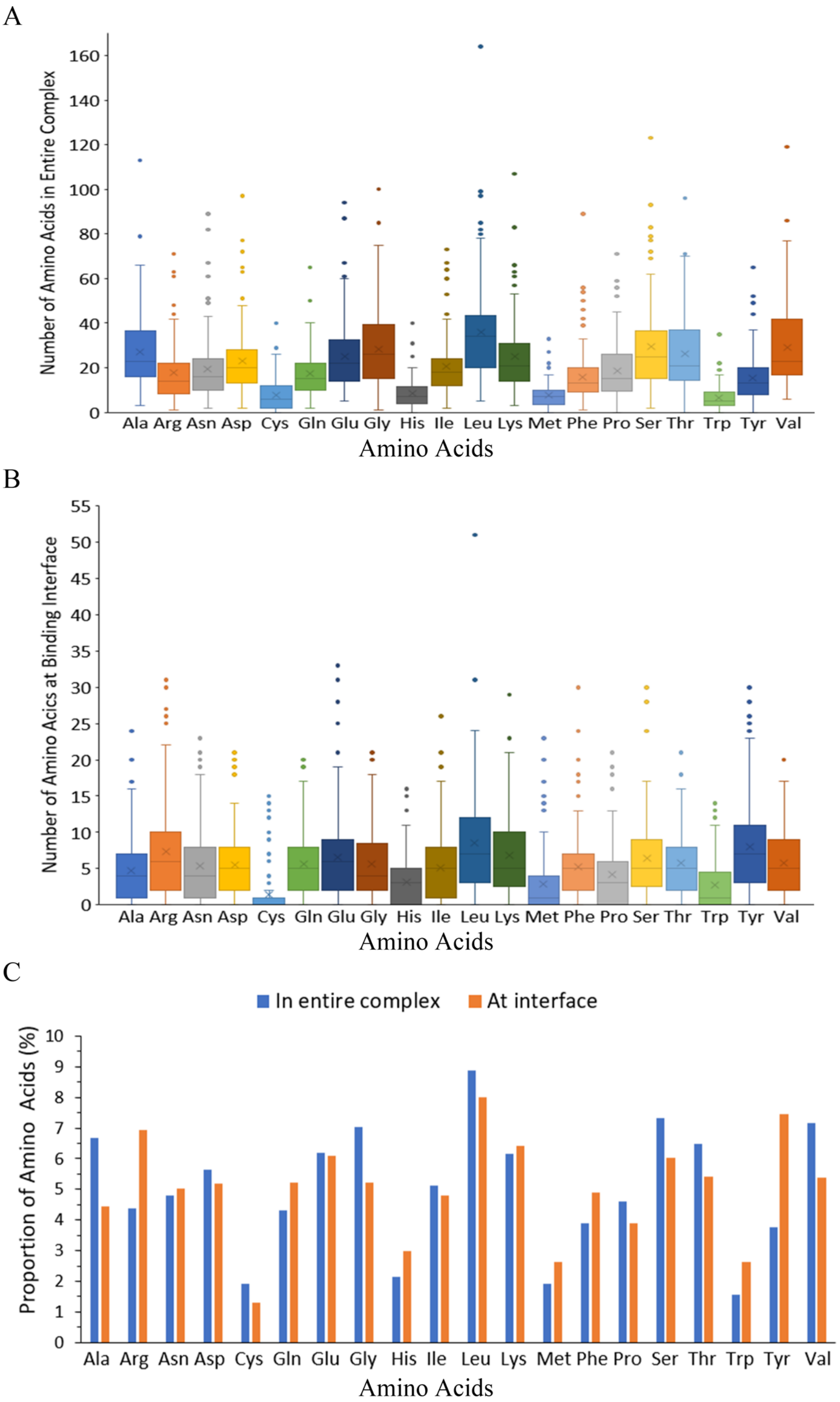


Figure 4. Abundance of the amino acids in the entire protein−protein complex and at the binding interface. (A) Box plots of the number of each amino acid in the entire protein−protein complex in the data set. (B) Box plots of the number of each amino acid involved at the binding interface (AA-IN) of protein−protein complex in the data set. (C) Comparison of the proportion of amino acids at the interface and in the entire complex.

correlation between ΔG and enrichment factor (Table S7). However, these correlations were lower than those of AA-INs, and thus we decided to focus our investigation on AA-INs.

3.7. AA-INs-Based Models Predict BA with Accuracy Comparable to Contact-Based Models.

To investigate the feasibility of specific amino acids for BA prediction, we used the numbers of each amino acid at the binding interface, AA-INs, of the 20 amino acids to build LR models predicting BA and understand the relationship between specific amino acids and BA. Using the AA-INs of the 20 amino acids, we built 1,048,575 LR models for all possible combinations of 20 features. Due to the huge number of models, we selected the top 2000 models to examine the R and AIC as a function of feature combination. These top 2000 models included models

Table 4. Pearson's Correlation R between AA-INs and Experimental ΔG

AA-INs	R	p-value	AA-INs	R	p-value
Tyr-IN	−0.49	<0.0001	Cys-IN	−0.13	0.062
Gly-IN	−0.45	<0.0001	Pro-IN	−0.13	0.062
Ser-IN	−0.44	<0.0001	His-IN	−0.08	0.173
Arg-IN	−0.27	0.0006	Leu-IN	−0.05	0.278
Asn-IN	−0.25	0.001	Lys-IN	−0.04	0.319
Thr-IN	−0.24	0.002	Met-IN	0.03	0.362
Asp-IN	−0.21	0.006	Ala-IN	0.03	0.362
Trp-IN	−0.21	0.006	Phe-IN	0.04	0.319
Glu-IN	−0.16	0.029	Gln-IN	0.11	0.097
Val-IN	−0.15	0.038	Ile-IN	0.11	0.097

with the number of features between 4 and 13 (Figure 5A,B). Similar to the results for the residue and atomic IC models above, the R_{max} increased as the number of features were added, but AIC reached a minimum of 600.9 when six features were used. This minimum AIC model used the following six amino acids: tyrosine, glycine, serine, arginine, valine, and isoleucine with a simple linear equation as follows.

$$\Delta G_{pred} = -0.1535\text{Tyr-IN} - 0.1288\text{Gly-IN} - 0.0840\text{Ser-IN} - 0.0805\text{Arg-IN} - 0.0684\text{Val-IN} + 0.0731\text{Ile-IN} - 6.46 \quad (2)$$

This model gave $R = 0.68$ ($p < 0.0001$) and RMSE = 1.94 kcal/mL, with test $R = 0.63$ in cross-validation (Figure 5C and Table S8). This model outperformed all of the models above that used a combination or a subset of residue ICs, atomic ICs, HS, and NIS features, while using only six features, as compared to the nine features used in the residue/atomic ICs/HS/NIS model. To examine whether the model can be further improved by adding residue and atomic ICs and the high-impacting NIS and HS features, we made sets of models combining AA-INs with these features. The best-performing model gave an R of 0.71 and a test R of 0.67 after cross-validation, indicating that combining AA-INs with these other features does not greatly improve predictions.

The six amino acids included in the minimum AIC AA-INs model included four apolar (tyrosine, glycine, valine, isoleucine), one polar (serine), and one charged (arginine) residue. Thus, differences in the relative importance of the different amino acids may help explain the lack of agreement between residue similarity and feature importance that we observed when analyzing LR models based on residue ICs. For example, a residue IC that contained an "unlike" combination of residues such as tyrosine and threonine would be more likely to have a strong impact on binding affinity than a "like" combination of residues with INs that are not correlated with BA. To support this conclusion, we investigated the number of times that tyrosine was in contact with each of the other amino acids and found that, of the 5 most common tyrosine contact partners, 2 were polar (asparagine and threonine) and 2 were charged (lysine and arginine). These "unlike" contacts that included tyrosine would increase the relative importance of polar-apolar and charged-apolar ICs, making these features more important, in general, when analyzing the residue contact-based LR models.

3.8. Feature Importance Studies Identified a Group of Amino Acids with High Impact on BA.

To investigate the importance of the 20 amino acids in BA prediction, we examined how the features were used to build the top 1000 models via the frequency of feature occurrence in these models. The results showed that there were six amino acids that were used with much higher frequency than others in the top 1000 models, tyrosine, glycine, arginine, serine, isoleucine, and valine (Figure 6A). These top six amino acids were the same six amino acids that were used to generate the minimum AIC model. Particularly, tyrosine, glycine, and arginine were used in all the top 1000 models.

Additionally, feature importance was studied with one-parameter LR modeling and RF. The one-parameter LR ranked tyrosine as the leading amino acid ($R = 0.49$), followed by glycine ($R = 0.45$) and serine ($R = 0.44$) (Figure 6B). These three amino acids performed much better than the rest of the amino acids ($R < 0.3$) in BA prediction. Arginine was ranked the fourth important feature with $R = 0.27$. However, the other two top amino acids given by their frequency of occurrence in the top 1000 models, isoleucine and valine, only gave R of 0.11 and 0.15, respectively. The amino acid that gave lowest R was methionine, with R of only 0.027. Methionine was also the least frequently used amino acid in the top 1000 models. The RF ranked the same top four amino acids in the same order as LR (Figure 6C).

A further way to identify important features is to examine the sign and amplitude of the coefficients of the features in the

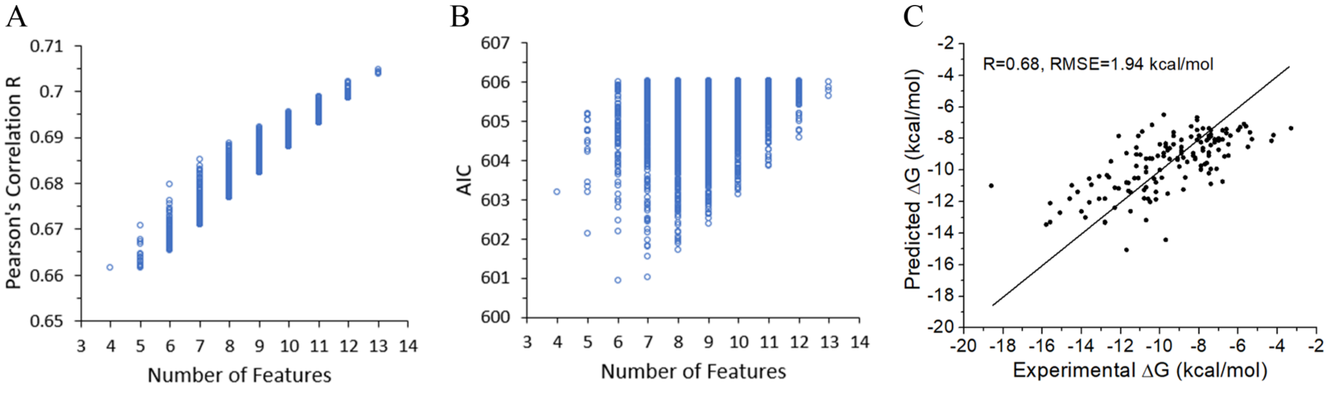


Figure 5. Performance of the AA-INs models for BA prediction. (A) Pearson's correlation R and (B) AIC as a function of the number of features used to generate the top 2000 models. (C) Scatter plot for the minimum AIC AA-INs model between predicted and experimental ΔG. The straight line is the function $y = x$.

https://doi.org/10.1021/acsomega.3c06996

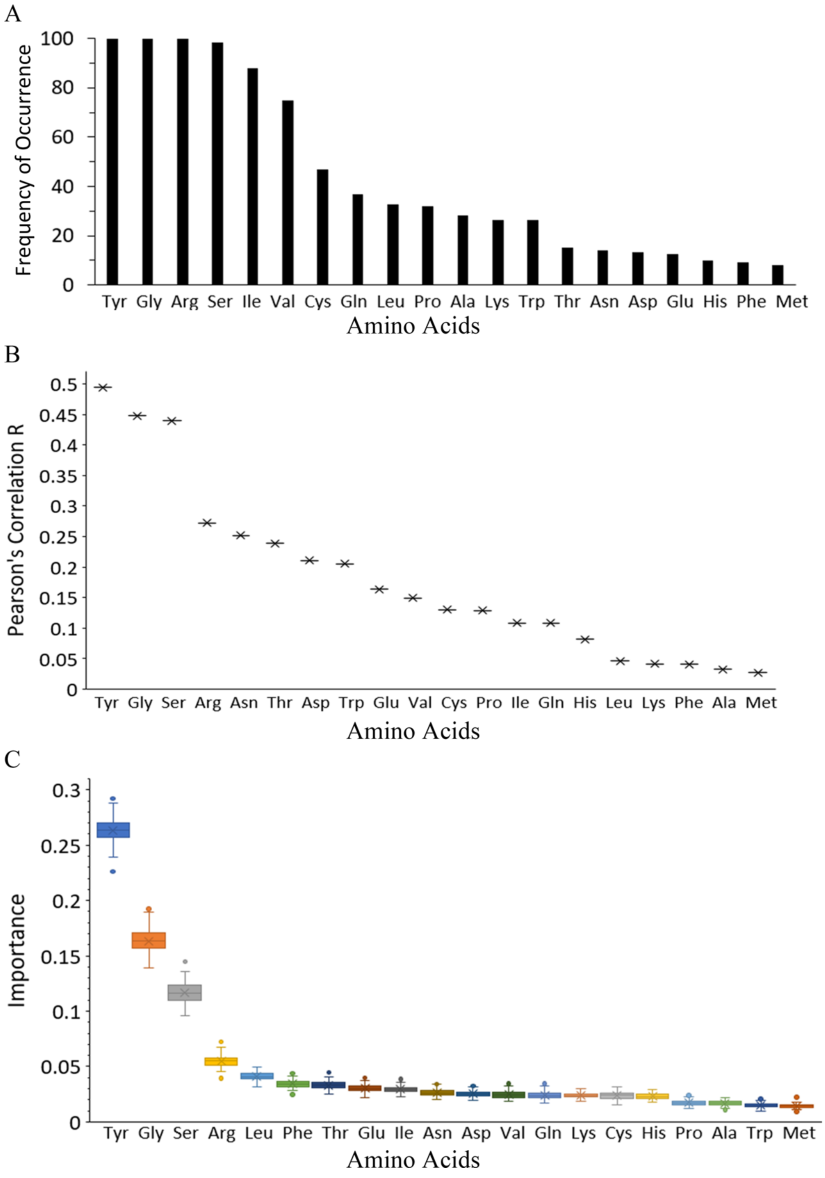


Figure 6. Feature importance of AA-INs by different approaches. (A) Frequency of occurrence in the top 1000 AA-INs models. (B, C) Feature importance of the 20 amino acids predicted by one-parameter LR (B) and RF (C).

linear equations of the LR models. The coefficients in the LR models contain information on the importance of each feature on BA; however, the values cannot be directly used to infer their importance yet. To evaluate the feature importance based on the coefficients in the LR model, we need to standardize each feature by scaling each to have zero mean and standard deviation of one. We also need to calculate the variance inflation factor (VIF) for each feature in the models to examine the multicollinearity among the features. The results showed that the VIF values for the coefficients in the minimum AIC model was very low for all the amino acids (<5.0), indicating that the model did not suffer from multicollinearity and the regression results were reliable (Table 5). In fact, the AA-INs had low correlation between each other (Table S9). The result showed that the coefficients of tyrosine, glycine, serine, and arginine were negative with $p < 0.05$, suggesting that increasing the INs of these amino acids leads to increased BA. The impact to BA follows the order of tyrosine, glycine, and serine, which

Table 5. Coefficients in the Linear Equation Predicting ΔG for the AA-INs Model with Minimum AIC before and after Feature Standardization

AA-INs	VIF	Coefficient before standardization	Coefficient after standardization	*p*-value
Tyr-IN	1.21	−0.1535	−1.0180	*p* = 0.000
Gly-IN	1.27	−0.1288	−0.6530	*p* = 0.001
Ser-IN	1.36	−0.0840	−0.4636	*p* = 0.019
Arg-IN	1.07	−0.0805	−0.5147	*p* = 0.004
Val-IN	1.09	−0.0684	−1.0180	*p* = 0.082
Ile-IN	1.04	0.0731	0.3462	*p* = 0.045
Constant	1.00	−6.46	−9.5582	*p* = 0.000

agrees with the results from the correlation between AA-INs with experimental ΔG.

To further investigate the feature importance, we analyzed the correlation between AA-INs and experimental BA for Set 1 and Set 2 separately. We found that the top 3 amino acids observed from the above studies with Set 3 (the combined Set 1 and Set 2) were also the top 3 amino acids across both data sets (Tables S12 and S13). However, arginine does not correlate with both data sets with consistent accuracy. Thus, tyrosine, glycine, and serine were identified as the top players in controlling BA.

3.9. Comparative Studies Show That the AA-INs Models Have More Consistent Performance Across Data Sets than Other Contact Models.

As we previously mentioned, the performance of the PRODIGY LR model on the PDBbind data set has been shown to be worse than its performance on the original PRODIGY data. To examine if the models generated in this work suffered from the same limitation, we split the data used here into three sets: the PRODIGY data set (Set 1), the PDBbind data set (Set 2), and the combined data set (Set 3). We then examined the performance of the minimum AIC models for each of the five feature sets (residue ICs/NIS, atomic ICs/NIS, residue/atomic ICs/NIS, residue/atomic ICs/HS/NIS, and AA-INs) on Set 1 and Set 2 when trained on the combined data set (Set 3). Similar to previous results using the PRODIGY model,[25] the R of the correlation between predicted and experimental ΔG decreased greatly on the PDBbind data set for the residue ICs/NIS model (which is highly similar to the PRODIGY model) and the atomic ICs/NIS (Table 6). The combined models, residue/atomic ICs/NIS and residue/atomic ICs/HS/NIS models, had a less extreme drop in performance, while the AA-INs had the most similar R across Sets 1 and 2.

Table 6. Comparison of R Values Calculated for the PRODIGY Dataset (Set 1) and PDBbind Dataset (Set 2) Using Minimum AIC Models Trained on the Combined Dataset

Model	Number of features used in the model	Set 1	Set 2
Residue ICs/NIS	5	0.69	0.54
Atomic ICs/NIS	2	0.61	0.49
Residue/atomic ICs/NIS	7	0.69	0.59
Residue/atomic ICs/HS/NIS	9	0.71	0.60
AA-INs	6	0.69	0.65

To supplement these results, we trained each of the 5 groups of models on Set 1, Set 2, and Set 3 individually and calculated correlations between experimental and predicted ΔG for the original sets and after 4-fold cross-validation for the minimum AIC models (Table 7). These results mirrored those discussed for Table 6, with the AA-INs model having similar performance across the three data sets. Thus, the AA-INs-based LR models were more robust to the training data set than the other contact-based models. Unlike the residue contact models that do not have the identities of the amino acid, the AA-INs model directly includes the amino acids involved in contacts when making predictions.

It is worth mentioning that our simple AA-INs model outperforms many other models. In the previous work, Vangone and Bonvin used 79 complexes from the PRODIGY data set and compared the performance of the PRODIGY model with 15 previously published models such as DFIRE, CP_PIE, and FIREDOCK that were built with different methods (e.g., global surface, buried surface area, composite scoring function).[21] They showed that the PRODIGY model outperformed all these models, suggesting the same for our AA-INs model as it has comparable performance as the PRODIGY model on the PRODIGY data set. Our AA-INs model also outperforms ISLAND and is comparable to PPI-Affinity, two recently reported top models, in making predictions of the PRODIGY data set, as the amino acid sequence-based ISLAND had R = 0.38 and the contact molecular descriptor-based PPI-Affinity had R = 0.62.[27] To evaluate whether the AA-INs model gives better performance using more advanced ML techniques, we trained the combined data set with SVM and AdaBoost models. However, neither of these two more complex models outperformed our simpler AA-INs model, with R^2 = 0.326 for SVM and R^2 = 0.305 for AdaBoost.

Inspired by the good performance of the AA-INs model, we further investigated whether intensive property based on AA-INs could predict BA with similar accuracy to the extensive AA-INs model. To do this, we divided the AA-INs for each amino acid by the total number of AA-INs for each complex and used them as features to train LR and RF models. However, models based on these AA-INs percentages were much worse than extensive AA-INs. For example, R = 0.24 was obtained for predictions of experimental BA using the LR model with all AA-INs percentages. However, the feature importance studies still show that the top three amino acids are tyrosine, glycine, and serine (Figure S3).

4. DISCUSSION

Despite advancements in understanding PPIs in the past two decades, understanding of how BA is controlled by physical/chemical parameters and how these parameters can be used to

Table 7. Comparison of the Performance of the Minimum AIC Models Using the PRODIGY Dataset (Set 1), PDBbind Dataset (Set 2), and the Combined Dataset (Set 3)[a]

	Set 1		Set 2		Set 3	
Model	R	Test R	R	Test R	R	Test R
Residue ICs/NIS	0.72	0.63 (0.04)	0.54	0.53 (0.07)	0.61	0.55 (0.01)
Atomic ICs/NIS	0.68	0.64 (0.02)	0.50	0.49 (0.04)	0.55	0.54 (0.02)
Residue/atomic ICs/NIS	0.73	0.63 (0.06)	0.62	0.52 (0.05)	0.64	0.60 (0.02)
Residue/atomic ICs/HS/NIS	0.74	0.68 (0.03)	0.62	0.45 (0.07)	0.67	0.61 (0.02)
AA-INs	0.74	0.66 (0.03)	0.71	0.61 (0.04)	0.68	0.63 (0.02)

[a] R is the Pearson's correlation coefficient of predictions after training on the respective complete datasets. Test R is Pearson's correlation coefficient (mean with standard deviation) of the test set after 10× fourfold cross-validation.

make predictions about complex binding affinities remains a challenge. One key set of features that has been commonly used to build models that predict binding affinity from a complex's structure are the numbers and identities of the protein−protein contacts that hold the complex together. However, inconsistent performance of contact-based models indicates that there is a need for further investigations into the relationship between binding affinity and the contacts in protein−protein complexes. To better understand this relationship, we compiled and studied a data set of 141 protein−protein complexes with diverse functions and a wide range of BAs and used it to build and test a series of LR models with a range of contact features. These features include both residue-based contact features that have been commonly used in previous BA prediction models and other features (atomic contacts, hydrogen bonding, hydrophobic indices, and the identities of the amino acid at the binding interface) that have received little or no attention in previous models.

In general, we found that the different sets of features provided roughly similar performance in predicting experimental binding affinities. First, a large number of single features produced a Pearson's correlation R between experimental and predicted BA of around −0.45. The single features that were able to achieve this performance were from all the major contact feature groups investigated here and included total residue ICs and polar/apolar residue ICs; total, polar/apolar, and polar/polar atomic ICs, Tyr-IN, Gly-IN, and Ser-IN. Second, the performance of both minimum AIC and all-feature models for the different investigated feature sets were roughly comparable, with the models for the different feature sets providing R values between 0.55 to 0.65 after 4-fold cross-validation. Additionally, continuing to add new features to the models typically resulted in no or very slight improvements in performance. For example, despite the relatively large single-feature R values for all the atomics-based contact classes, the minimum AIC atomic ICs model contained only two features (%NIS_polar and polar-apolar ICs). Thus, adding polar-polar and apolar-apolar atomic contacts to this two-feature model was not able to improve the model, despite the high single-feature R of atomic polar-polar and apolar-apolar contacts. One possible exception to this trend was that combining atomic and residue ICs resulted in a modest increase in performance, and the future development of models containing both atomic and residue ICs would be of interest. Taken together, our results indicate that currently used contact features alone may not be able to make highly accurate predictions of experimental BA, as the R between contact feature-based predicted and experimental BA seems to have an upper limit of ∼0.7. Improving models beyond this limit will likely require the addition of noncontact features or novel ways of describing the contacts at the binding interface, such as the AA-INs introduced here.

One of the key results of this study was the determination that the numbers of amino acids in protein−protein contacts (AA-INs) could be used to produce models that have more consistent performance across data sets than models built with standard contact features. Indeed, the AA-INs model had the highest correlation with BA of any of the model groups investigated here, and, in particular, it had the most consistent performance across the two different benchmark data sets (Table 7). Using a diverse set of 141 complexes from two benchmark data sets, we generated a best-performing AA-INs model with a simple and explicit linear equation involving six amino acids (tyrosine, glycine, serine, arginine, valine, and isoleucine).

We identified three top amino acids that underlie the protein−protein binding strength, which were tyrosine, glycine, and serine. We showed that tyrosine played the leading role in predicting BA, with the highest correlation of all amino acids between INs and ΔG (R = −0.49). Its impact surpassed that of the top polarity-classified residue ICs, polar-apolar residues (R = −0.44), and charged-apolar residues (R = −0.39). Using the interfacial number of tyrosine alone, we can generate an LR model to predict BA with R of nearly 0.5. Combing tyrosine with one of the other amino acids used in the minimal AIC AA-INs model (eq 2), R as high as 0.59 (with glycine or serine) was reached (Tables S10 and S11). Tyrosine is known to have unique physicochemical properties that make it the most effective amino acid in mediating molecular recognition.[38] Tyrosine is a large and apolar amino acid with a hydroxyl group. Its unique side chain makes it well-suited for antibodies to make productive contacts with antigen.[53] Tyrosine is the second most-enriched amino acid at the binding interface (Figure 4C and Table S6). While tyrosine is large, glycine is small and serine is relatively small. These small residues may provide space and flexibility for tyrosine to mediate molecular contacts.[38] Thus, tyrosine, glycine, and serine seem to intrinsically work together to mediate molecular recognition. In fact, tyrosine, glycine, and serine have been combined to make highly specific synthetic antibodies with high affinity.[54] Tyrosine together with arginine and tryptophan are the three amino acids with the highest enrichment at the interface (1.99× for tyrosine, 1.65× for tryptophan, and 1.61× for arginine) (Table S6). These three amino acids have been shown to appear in hot spots of the complex interface with a frequency over 10% (21% for tryptophane, 13% for arginine, and 12% for tyrosine).[36] However, the interfacial number of tryptophan has low correlation with BA. Thus, tryptophan was not found to contribute to interface energetics in our analysis although it is highly enriched at interface and shares many attributes with tyrosine.

In summary, there have been decades of major effort at anatomizing protein−protein interfaces to better understand PPIs and develop therapeutic drugs. The results from this study further our understanding of PPIs by unraveling a quantitative link between the abundance of specific amino acids at the interface and the strength of PPIs. Thus, interfacial amino acids-based features, such as AA-INs, should be considered in future attempts to build ML models to predict BA for drug design and screening in the pharmaceutical industry or for other applications.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.3c06996.

> Additional information on the data sets, feature compositions in the minimal AIC contact based models, expected and actural proportion of residue paris in the protein−protein complexes, Coefficients of features in the linear equations predicting ΔG for of all-feature models, Pearson's correlation R among the 14 features, proportion (%) of each amino acid in the entire complex and at interface as well as the enrichment factor, Pearson's correlation R between AA enrichment factor

and experimental ΔG, Test R for the 4-fold cross-validation of the AA-INs model with minimal AIC, Pearson's correlation R between AA-INs features, Pearson's correlation R and AICs for the single-feature and two-feature AA-INs model using two different features from Tyr-IN, Gly-IN, Ser-IN, Arg-IN, Val-IN, and Ile-IN, Pearson's correlation R between AA-INs and experimental ΔG for Set 1 and Set 2, calculated features for the contact models and AA-INs model, and predicted ΔG using these models with minimal AIC (PDF)

Data from contact models and AA-INs model (XLSX)

AUTHOR INFORMATION

Corresponding Author

Yongmei Wang − *Department of Chemistry, The University of Memphis, Memphis, Tennessee 38152, United States;* orcid.org/0000-0002-7418-9489; Email: ywang@memphis.edu

Authors

Carey Huang Yi − *Department of Chemistry, The University of Memphis, Memphis, Tennessee 38152, United States*

Mitchell Lee Taylor − *Department of Chemistry, The University of Memphis, Memphis, Tennessee 38152, United States;* orcid.org/0000-0002-6997-7800

Jesse Ziebarth − *Department of Chemistry, The University of Memphis, Memphis, Tennessee 38152, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c06996

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Kaya Marascalco for the help on compiling the dataset. We also thank Dr. Qianyi Chen for discussions on the RF algorithm and the calculation of the number of hydrogen bonds.

REFERENCES

(1) Jones, S.; Thornton, J. M. Principles of Protein-Protein Interactions. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 13−20.

(2) Nooren, I. M. A.; Thornton, J. M. Structural Characterization and Functional Significance of Transient Protein−Protein Interactions. *J. Mol. Biol.* **2003**, *325*, 991−1018.

(3) Keskin, O.; Nussinov, R. Similar Binding Sites and Different Partners: Implications to Shared Proteins in Cellular Pathways. *Structure* **2007**, *15*, 341−354.

(4) Morris, R.; Black, K. A.; Stollar, E. J. Uncovering Protein Function: From Classification to Complexes. *Essays Biochem.* **2022**, *66*, 255−285.

(5) Ryan, D.; Matthews, J. Protein−Protein Interactions in Human Disease. *Curr. Opin. Struct. Biol.* **2005**, *15*, 441−446.

(6) Gonzalez, M. W.; Kann, M. G. Chapter 4: Protein Interactions and Disease. *PLoS Comput. Biol.* **2012**, *8*, No. e1002819.

(7) Calabrese, G.; Molzahn, C.; Mayor, T. Protein Interaction Networks in Neurodegenerative Diseases: From Physiological Function to Aggregation. *J. Biol. Chem.* **2022**, *298*, No. 102062.

(8) Perkins, J. R.; Diboun, I.; Dessailly, B. H.; Lees, J. G.; Orengo, C. Transient Protein-Protein Interactions: Structural, Functional, and Network Properties. *Structure* **2010**, *18*, 1233−1243.

(9) Kastritis, P. L.; Bonvin, A. M. J. J. On the Binding Affinity of Macromolecular Interactions: Daring to Ask Why Proteins Interact. *J. R. Soc. Interface* **2013**, *10*, No. 20120835.

(10) Thafar, M.; Raies, A. B.; Albaradei, S.; Essack, M.; Bajic, V. B. Comparison Study of Computational Prediction Tools for Drug-Target Binding Affinities. *Front. Chem.* **2019**, *7*, 782.

(11) Paggi, J. M.; Belk, J. A.; Hollingsworth, S. A.; Villanueva, N.; Powers, A. S.; Clark, M. J.; Chemparathy, A. G.; Tynan, J. E.; Lau, T. K.; Sunahara, R. K.; Dror, R. O. Leveraging nonstructural data to predict structures and affinities of protein-ligand complexes. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118*, No. e2112621118.

(12) Thafar, M. A.; Alshahrani, M.; Albaradei, S.; Gojobori, T.; Essack, M.; Gao, X. Affinity2Vec: drug-target binding affinity prediction through representation learning, graph mining, and machine learning. *Sci. Rep.* **2022**, *12*, 4751.

(13) Keskin, O.; Tuncbag, N.; Gursoy, A. Predicting Protein−Protein Interactions from the Molecular to the Proteome Level. *Chem. Rev.* **2016**, *116*, 4884−4909.

(14) Siebenmorgen, T.; Zacharias, M. Computational Prediction of Protein−Protein Binding Affinities. *WIREs Comput. Mol. Sci.* **2020**, *10*, No. e1448.

(15) Wang, L.; Chambers, J.; Abel, R. Protein-Ligand Binding Free Energy Calculations with FEP. *Methods Mol. Biol.* **2019**, *2022*, 201−232.

(16) Zou, J.; Tian, C.; Simmerling, C. Blinded Prediction of Protein−Ligand Binding Affinity Using Amber Thermodynamic Integration for the 2018 D3R Grand Challenge 4. *J. Comput. Aided Mol. Des.* **2019**, *33*, 1021−1029.

(17) Liu, S.; Zhang, C.; Zhou, H.; Zhou, Y. A Physical Reference State Unifies the Structure-Derived Potential of Mean Force for Protein Folding and Binding. *Proteins: Struct. Funct. Bioinform.* **2004**, *56*, 93−101.

(18) Zhang, C.; Liu, S.; Zhu, Q.; Zhou, Y. A Knowledge-Based Energy Function for Protein−Ligand, Protein−Protein, and Protein−DNA Complexes. *J. Med. Chem.* **2005**, *48*, 2325−2335.

(19) Jiménez, J.; Škalič, M.; Martínez-Rosell, G.; De Fabritiis, G. Kdeep: Protein−Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *J. Chem. Inf. Model.* **2018**, *58*, 287−296.

(20) Moal, I. H.; Agius, R.; Bates, P. A. Protein−Protein Binding Affinity Prediction on a Diverse Set of Structures. *Bioinformatics* **2011**, *27*, 3002−3009.

(21) Vangone, A.; Bonvin, A. M. Contacts-Based Prediction of Binding Affinity in Protein−Protein Complexes. *eLife* **2015**, *4*, No. e07454.

(22) Xue, L. C.; Rodrigues, J. P.; Kastritis, P. L.; Bonvin, A. M.; Vangone, A. PRODIGY: A Web Server for Predicting the Binding Affinity of Protein−Protein Complexes. *Bioinformatics* **2016**, *32*, No. 3676.

(23) Vangone, A.; Bonvin, A. PRODIGY: A Contact-Based Predictor of Binding Affinity in Protein-Protein Complexes. *Bio-protocol.* **2017**, *7* (3), No. e2124.

(24) Li, S.; Wan, F.; Shu, H.; Jiang, T.; Zhao, D.; Zeng, J. MONN: A Multi-Objective Neural Network for Predicting Compound-Protein Interactions and Affinities. *Cell Syst.* **2020**, *10*, 308−322.

(25) Wang, M.; Cang, Z.; Wei, G.-W. A Topology-Based Network Tree for the Prediction of Protein−Protein Binding Affinity Changes Following Mutation. *Nat. Mach. Intell.* **2020**, *2*, 116−123.

(26) Liu, X.; Luo, Y.; Li, P.; Song, S.; Peng, J. Deep Geometric Representations for Modeling Effects of Mutations on Protein-Protein Binding Affinity. *PLoS Comput. Biol.* **2021**, *17*, No. e1009284.

(27) Romero-Molina, S.; Ruiz-Blanco, Y. B.; Mieres-Perez, J.; Harms, M.; Münch, J.; Ehrmann, M.; Sanchez-Garcia, E. PPI-Affinity: A Web Tool for the Prediction and Optimization of Protein−Peptide and Protein−Protein Binding Affinity. *J. Proteome Res.* **2022**, *21*, 1829−1841.

(28) Yang, Y. X.; Wang, P.; Zhu, B. T. Importance of Interface and Surface Areas in Protein-Protein Binding Affinity Prediction: A Machine Learning Analysis Based on Linear Regression and Artificial Neural Network. *Biophy. Chem.* **2022**, *283*, No. 106762.

(29) Wee, J.; Xia, K. Persistent Spectral Based Ensemble Learning (PerSpect-EL) for Protein−Protein Binding Affinity Prediction. *Brief. Bioinformatics* **2022**, *23*, No. bbac024.

(30) Yugandhar, K.; Gromiha, M. M. Protein−Protein Binding Affinity Prediction from Amino Acid Sequence. *Bioinformatics* **2014**, *30*, 3583−3589.

(31) Abbasi, W. A.; Asif, A.; Ben-Hur, A.; Minhas, F. U. A. A. Learning Protein Binding Affinity Using Privileged Information. *BMC Bioinformatics* **2018**, *19*, 425.

(32) Chen, M.; Ju, C.J.-T.; Zhou, G.; Chen, X.; Zhang, T.; Chang, K.-W.; Zaniolo, C.; Wang, W. Multifaceted Protein−Protein Interaction Prediction Based on Siamese Residual RCNN. *Bioinformatics* **2019**, *35*, i305−i314.

(33) Abbasi, W. A.; Yaseen, A.; Hassan, F. U.; Andleeb, S.; Minhas, F. U. A. A. ISLAND: In-Silico Proteins Binding Affinity Prediction Using Sequence Information. *BioData Mining* **2020**, *13*, 20.

(34) Rube, H. T.; Rastogi, C.; Feng, S.; Kribelbauer, J. F.; Li, A.; Becerra, B.; Melo, L. A. N.; Do, B. V.; Li, X.; Adam, H. H.; et al. Prediction of Protein−Ligand Binding Affinity from Sequencing Data with Interpretable Machine Learning. *Nat. Biotechnol.* **2022**, *40*, 1520−1527.

(35) Guo, Z.; Yamaguchi, R. Machine Learning Methods for Protein-Protein Binding Affinity Prediction in Protein Design. *Front. Bioinform.* **2022**, *2*, No. 1065703.

(36) Bogan, A. A.; Thorn, K. S. Anatomy of Hot Spots in Protein Interfaces. *J. Mol. Biol.* **1998**, *280*, 1−9.

(37) Lo Conte, L.; Chothia, C.; Janin, J. The Atomic Structure of Protein-Protein Recognition Sites. *J. Mol. Biol.* **1999**, *285*, 2177−2198.

(38) Koide, S.; Sidhu, S. S. The Importance of Being Tyrosine: Lessons in Molecular Recognition from Minimalist Synthetic Binding Proteins. *ACS Chem. Biol.* **2009**, *4*, 325−334.

(39) Kuo, H.-C.; Lin, J.-C.; Ong, P.-L.; Huang, J.-P. Discovering Amino Acid Patterns on Binding Sites in Protein Complexes. *Bioinformation* **2011**, *6*, 10−14.

(40) Talavera, D.; Robertson, D. L.; Lovell, S. C. Characterization of Protein-Protein Interaction Interfaces from a Single Species. *PLoS One* **2011**, *6*, No. e21053.

(41) Krick, T.; Verstraete, N.; Alonso, L. G.; Shub, D. A.; Ferreiro, D. U.; Shub, M.; Sanchez, I. E. Amino Acid Metabolism Conflicts with Protein Diversity. *Mol. Biol. Evol.* **2014**, *31*, 2905−2912.

(42) Erijman, A.; Rosenthal, E.; Shifman, J. M. How Structure Defines Affinity in Protein-Protein Interactions. *PLoS One* **2014**, *9*, No. e110085.

(43) Jayashree, S.; Murugavel, P.; Sowdhamini, R.; Srinivasan, N. Interface Residues of Transient Protein-Protein Complexes Have Extensive Intra-Protein Interactions Apart from Inter-Protein Interactions. *Biol. Direct.* **2019**, *14*, 1.

(44) Kastritis, P. L.; Rodrigues, J.P.G.L.M.; Folkers, G. E.; Boelens, R.; Bonvin, A. M. J. J. Proteins Feel More Than They See: Fine-Tuning of Binding Affinity by Properties of the Non-Interacting Surface. *J. Mol. Biol.* **2014**, *426*, 2632−2652.

(45) Kyte, J.; Doolittle, R. F. A Simple Method for Displaying the Hydropathic Character of a Protein. *J. Mol. Biol.* **1982**, *157*, 105−132.

(46) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; et al. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* **2009**, *25*, 1422−1423.

(47) Hubbard, S.; Thornton, J. *Naccess,* V2.1.1, Computer Program; Department of Biochemistry Molecular Biology, University College London, 1993.

(48) Kastritis, P. L.; Moal, I. H.; Hwang, H.; Weng, Z.; Bates, P. A.; Bonvin, A. M. J. J.; Janin, J. A Structure-Based Benchmark for Protein−Protein Binding Affinity. *Protein Sci.* **2011**, *20*, 482−491.

(49) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein−Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* **2004**, *47*, 2977−2980.

(50) Lazareno, S.; Birdsall, N. J. M. Estimation of Competitive Antagonist Affinity from Functional Inhibition Curves Using the Gaddum, Schild and Cheng-Prusoíf Equations. *Br. J. Pharmacol.* **1993**, *109*, 1110−1119.

(51) Wilkinson, K. D. Quantitative Analysis of Protein-Protein Interactions. *Method. Mol. Biol. 261* **2004**, *261*, 15−31.

(52) Masi, A.; Cicchi, R.; Carloni, A.; Pavone, F. S.; Arcangeli, A. Optical Methods in the Study of Protein-Protein Interactions. *Adv. Exp. Med. Biol.* **2010**, *674*, 33−42.

(53) Fellouse, F. A.; Wiesmann, C.; Sidhu, S. S. Synthetic antibodies from a four-amino-acid code: a dominant role for tyrosine in antigen recognition. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 12467−12472.

(54) Birtalan, S.; Zhang, Y.; Fellouse, F. A.; Shao, L.; Schaefer, G.; Sidhu, S. S. The intrinsic contributions of tyrosine, serine, glycine and arginine to the affinity and specificity of antibodies. *J. Mol. Biol.* **2008**, *377*, 1518−1528.