HOKKAIDO UNIVERSITY

| | |
|---|---|
| Title | Exploring The Interplay Between Scoring Functions and Physico-chemical Properties in Antibody-antigen Docking |
| Author(s) | Ratnayake, Sangeetha; Martinelli, Axel; Endo, Toshinori et al. |
| Citation | IPSJ Transactions on Bioinformatics, 17, 10-17<br>https://doi.org/10.2197/ipsjtbio.17.10 |
| Issue Date | 2024-02 |
| Doc URL | https://hdl.handle.net/2115/96098 |
| Rights | Notice for the use of this material The copyright of this material is retained by the Information Processing Society of Japan (IPSJ). This material is published on this web site with the agreement of the author (s) and the IPSJ. Please be complied with Copyright Law of Japan and the Code of Ethics of the IPSJ if any users wish to reproduce, make derivative work, distribute or make available to the public any part or whole thereof. All Rights Reserved, Copyright (C) Information Processing Society of Japan. Comments are welcome. Mail to address editj@ipsj.or.jp, please. |
| Type | journal article |
| File Information | IPSJ Trans. Bioinf.17_10-17.pdf |

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

**IPSJ Transactions on Bioinformatics Vol.17 10–17 (Feb. 2024)**

[DOI: 10.2197/ipsjtbio.17.10]

**Original Paper**

# Exploring The Interplay Between Scoring Functions and Physico-chemical Properties in Antibody-antigen Docking

Sangeetha Ratnayake[1,a)] Axel Martinelli[2] Toshinori Endo[3] Naoki Osada[3]

Received: August 31, 2023, Revised: October 10, 2023,
Accepted: November 26, 2023

***Abstract:*** The advent of antibody therapy has brought about a change in the treatment of diseases. The efficacy of antibody modeling relies on the intricate atomic interactions between antibodies and antigens. Traditional methods for determining antibody structures, such as X-ray crystallography, are costly and time-consuming. Computational docking offers a faster and more cost-effective approach to obtaining complex antibody and antigen complexes even in challenging scenarios. Rosetta, a widely employed software for protein structure modeling, incorporates a scoring function specifically tailored for modeling antibody-antigen interactions. The unique characteristics of the antibody-antigen interface can result in inaccurate predictions. Therefore, it is essential to understand the existing scoring function and the behavior of the antibody-antigen interface. In this study, we evaluated specific parameters within Rosetta-derived scoring functions, with a particular focus on the energy landscape of the structures they generated. We found that performance in antibody-antigen docking simulations could be enhanced by omitting parameters related to solvation. Also, we delved into the physico-chemical properties of antibody-antigen interfaces, paying special attention to the complementarity-determining regions and epitopes. Our exploration helped identify certain parameters that significantly influence docking simulation performance. These insights pave the way for the creation of more accurate scoring functions tailored for specific antibody-antigen interactions.

***Keywords:*** antibody-antigen complexes, CDR loops, epitope, physico-chemical properties, scoring function

## 1. Introduction

Biological medicines, particularly antibody-based therapy, have brought a transformation in the pharmacological treatment of various disorders. As vital components of the immune system, antibodies possess the ability to identify and eliminate foreign entities such as bacteria and viruses. Understanding the mechanisms underlying the interaction between antibodies and antigens has become a focal point of research in antibody-based treatments [1], [2], [3].

Experimental techniques such as X-ray crystallography are commonly employed to uncover the structures of antibody-antigen (Ab-Ag) complexes. However, these techniques have limitations in terms of throughput, cost, and applicability to a wide range of proteins. Computational docking is a technique used in computational biology and drug discovery to predict the binding orientation and affinity of a small molecule (ligand) with a target biomolecule, such as a protein or nucleic acid, and significantly reduce the time and cost for predicting Ab-Ag binding structures. The objective of docking is to evaluate the interaction between the ligand and biomolecule, enabling the prediction of their three-dimensional structure upon binding [4]. This enables the prediction of the binding affinity and interaction mode between the two molecules. One widely used method for protein-protein binding prediction is the optimization of a "Scoring Function" which is an algorithm designed for rigid and well-characterized protein structures. These scoring functions, found in various systems [5], [6], [7], [8], [9], [10], [11], [12], are derived to approximate the system's typical chemical potentials. They are incorporated into several docking software, as provided in the supplementary document S1, which lists commonly used computational docking software with the implemented scoring function and included parameters. These scoring functions are developed based on known protein structures. In this process, it is often assumed that protein-protein interactions (PPI) are primarily governed by hydrophobic interactions [5], [7], [8].

However, many studies have highlighted the presence of peculiarities in Ab-Ag interfaces that can result in inaccurate predictions [1], [2]. One such peculiarity is the enrichment of aromatic residues, such as tyrosine (Tyr), tryptophan (Trp), and phenylalanine (Phe), in the antibody complementarity-determining regions (CDR) loops responsible for direct antigen interaction [1], [2]. In addition, these CDR loops exhibit enrichment of hydrophilic residues with short side chains, such as serine (Ser), threonine (Thr), asparagine (Asn), and aspartic acid (Asp), alongside aro-

[1] Division of Bioinformatics and Bioengineering, Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Hokkaido 060–0814, Japan
[2] BigOmics Analytics, Via Serafino Balestra 12, 6900 Lugano, Switzerland
[3] Faculty of Information Science and Technology, Hokkaido University, Sapporo, Hokkaido 060–0814, Japan
[a)] sangeetha.ratnayake@gmail.com

© 2024 Information Processing Society of Japan

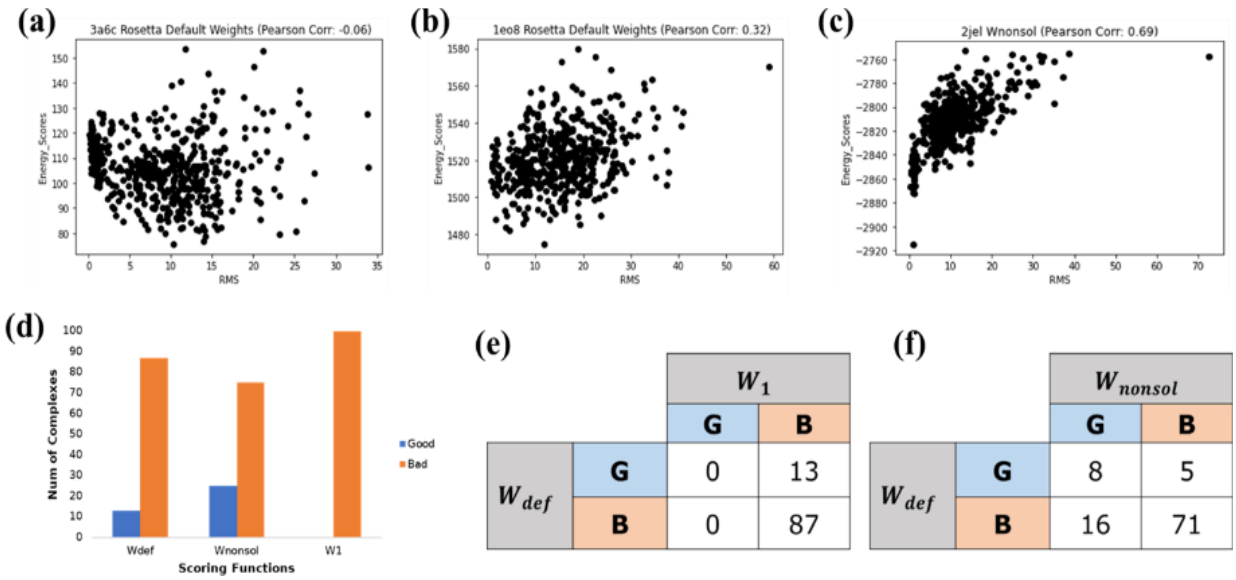



**Fig. 1** Goodness of the decoy distributions. a) and b) examples of bad decoy distributions. c) An example of good funnel-like decoy distribution in scoring function. d) Distribution of good (blue) and bad (orange) funnels in all three scoring functions. e) and f) Counts of funnel-like shape improvement of two example structures respectively.

matic side chains [13], [14], [15]. The significance of Tyr and Ser residues in the antibody side chain has been extensively studied, with aromatic interactions involving aromatic residues in CDR. Tyr is more prevalent than typical hydrophobic protein-protein interfaces at the interfaces of the Ab-Ag complex [16].

The roles of short hydrophilic side chains in Ab-Ag interactions have yet to be implemented into the Ab-Ag-specific scoring functions. Due to the unique nature of Ab-Ag interactions, specialized scoring functions are required to thoroughly investigate and understand their complexities.

In this study, we utilized the Rosetta software, which encompasses a range of packages and frameworks for protein modeling and docking, to investigate the impact of different scoring functions on the prediction of Ab-Ag binding structures [17]. Rosetta has played a significant role in advancing computational biology, enabling achievements in de novo protein design, enzyme design, ligand docking, and the prediction of macromolecular complexes [17]. For Ab-Ag docking, we employed the Rosetta Antibody docking scoring function, a linear equation with predefined weights based on positive interactions between the docking partners. The scoring function operates on the principle that energetically favorable structures, exhibiting lower energy, are more likely to represent correct biological configurations [17]. It incorporates parameters related to electrostatic interactions, van der Waals interactions, and molecular dynamic properties to calculate the energy of the complex.

In the context of docking simulations, a "decoy," also referred to as a docking decoy or scoring benchmarking set, represents an alternative binding pose or conformation of a ligand molecule within a protein binding site [23]. Docking simulations involve exploring multiple potential conformations, orientations, and interactions to identify the most favorable binding pose. While docking simulations, Rosetta generates multiple decoys, which are subsequently plotted on a distribution graph showcasing the energy score versus the Root Mean Square Deviation (RMSD) values, commonly known as the decoy distribution.

A practical criterion commonly used by researchers to assess scoring function accuracy is the shape of the decoy distribution [18], [19]. Despite being a widely adopted and established software for Ab-Ag docking and analysis, Rosetta is not immaculate and may exhibit limitations when applied to antibodies [20]. The accuracy of Rosetta's predictions is dependent on factors such as the quality of input structures, the complexity of the Ab-Ag interface, and the limitations of the scoring function [21], [22]. Additionally, Rosetta's performance can be affected by the presence of non-specific interactions, such as hydrophobic interactions, which are not explicitly accounted for in the scoring function. This can lead to false-positive predictions and inaccurate assessments of binding orientation and energetics.

In this study, we improved the accuracy of the RosettaAb Docking Scoring Function for Ab-Ag complexes through three analyses.

First, we assessed the distribution of decoy shapes using various parameters and criteria. This helped us understand how effective the scoring function is at distinguishing between favorable and unfavorable poses. We identified the most important parameters that affect the scoring function's ability to find high-quality docking positions.

Second, we studied the physical and chemical properties of the interface between antigens and antibodies, focusing on specific regions called CDR loops and epitopes. Our goal was to better understand how these regions influence the docking process. This analysis revealed the chemical makeup of CDR loops and epitopes in binding interactions and their roles in antibodies and antigens.

Lastly, we explored the connection between certain physical and chemical properties and the accuracy of the decoy distribution. We aimed to identify the properties that significantly contribute to the correct representation of docking positions. This investigation helped us uncover the key factors that affect the ac-

© 2024 Information Processing Society of Japan

curate prediction of docking orientations.

These analyses provide valuable insights for refining the RosettaAb Docking Scoring Function and improving its ability to predict precise docking positions for Ab-Ag complexes.

## 2. Methods

### 2.1 Input Data Preparation

We retrieved the Ab-Ag bound structures from SabDab Database [24] and created a dataset of 100 bound Ab-Ag complexes with a structural resolution of 5Å or less and a length of antigen sequence is 25 amino acids or more (Supplementary document S2). The data for CDR loops (L1, L2, L3, H1, H2, H3 fragments) is also retrieved from the SabDab Database [24]. The dataset consists only of antigens that are peptides or proteins, and the antibody complexes consist of human or human-like protein sequences. Our dataset further consists of high-quality structures with fewer than 10% missing residues and an R-factor of less than 0.4.

### 2.2 Ab-Ag Docking and Generating Decoys

To generate decoys, we used RosettaAb Docking which is the Ab-Ag interface specified program of computational docking [16]. For the selected 100 Ab-Ag bound complexes, we generated 500 decoys of each complex, and decoy distributions were created for each docking run. Then the total energy score was calculated using the Rosetta scoring function.

### 2.3 Energy Scoring with Customized Weights

The Rosetta scoring function *ref2015* is a linear equation with pre-assigned weights, which are defined on the positive interaction between the docking partners with the assumption that the correct biological structure is energetically favored and has, therefore, the lowest energy [17]. The Rosetta energy function approximates the conformational energy of biomolecules. This amount, denoted by the symbol $\Delta E_{total}$, is derived from a linear combination of energy terms $E_i$ calculated for the $i$th parameter as a function of geometric degrees of freedom, $\theta$, chemical identities, $aa$, and a weight on each term, $w$ as shown below,

$$\Delta E_{total} = \sum_i w_i E_i (\theta_i, aa_i) \tag{1}$$

Rosetta Default scoring function has 20 default parameters which represent different free energy. We developed three sets of weights derived from the Rosetta scoring function: 1) The Rosetta default weights ($W_{def}$) were considered the baseline of weights; 2) all parametric weights were set to 1 ($W_1$); and 3) Rosetta energy scores were computed by omitting the solvation-based parameters [(*fa_sol*, *fa_intra_rep*, *fa_intra_sol_xover4*, and *lk_ball_wtd*)] from the Rosetta default scoring function ($W_{nonsol}$). Using the three functions, the scoring energy for each Ab-Ag complex's generated decoys was then computed (Supplementary document S3).

The docking funnel's shape was evaluated using either the position of the decoy with the smallest RMSD and Pearson's correlation coefficient. Decoys with the smallest RMSD indicate how similar they are to the original complex. If these lowest RMSD decoys are also the ones with the lowest energy among all the other decoys, the docking funnel shape is considered good, and if this assumption is not met, the funnel is considered bad. In the criteria using Pearson's correlation coefficient, we labeled the funnel shape good when the value is equal or greater than 0.4, and bad when the value is smaller than 0.4.

### 2.4 Extracting Antibody CDR Loops and Antigen Features

We retrieved the information of epitopes of antigens from The Immune Epitope Database (IEDB). IEDB is a resource that is freely available to the public and is funded by NIAID (www.iedb.org) [25]. It compiles experimental data on antibodies and T cell epitopes that have been researched in humans, non-human primates, and other animal species. These studies were conducted in the context of infectious disease, allergy, autoimmunity, and transplantation. The IEDB antibody epitope prediction was carried out with BepiPred-2.0 [26], which is a sequential B-cell epitope predictor. The BepiPred-2.0 server can determine B-cell epitopes from a protein sequence by utilizing a Random Forest algorithm that has been trained on epitopes and non-epitope amino acids that have been determined from crystal structures. Following that, a sequential prediction smoothing will be carried out. The residues that have scores that are higher than the threshold (the default value is 0.5) are thought to be part of an epitope.

The hydrophobic and hydrophilic residue distribution of each CDR loop and the epitope is calculated using Kyte and Doolittle's (1982) hydrophobicity scale [27]. We used the Zimmerman, Eliezer, and Simha (1968) polarity index [28] to determine the distribution of high- and low-polar amino acids in CDR loops and the Ag epitope to comprehend the polarity of the Ab-Ag interface.

Understanding the surface area of the Ab-Ag binding pocket is crucial. Therefore, we mathematically determined the area of the interacting interface.

$$A_{Ab-Ag} = A_s A_f \tag{2}$$

We determined the surface area of the Ab-Ag interface ($A_{Ab-Ag}$) by multiplying the surface area accessible ($A_s$) to water in the unfolded position of each CDR loop, as described by Miller et al. (1987) [29], and the fraction of accessible area lost ($A_f$) [30] when a peptide folds. In this calculation, we assumed that the area lost when the peptide folds contribute 100% to the Ab-Ag binding. Antibody CDR loops are typically enriched with aromatic residues [31]. We calculated the distribution of aromatic residues in the CDR loops of each Ab-Ag complex. The table containing calculations of the physicochemical properties of amino acids can be found in supplementary document S4.

### 2.5 Statistical Analysis

The Generalized Linear Model (GLM) is a statistical model that provides flexibility in analyzing non-normally distributed data. It allows for the modeling of response variables with different distributions and diverse relationships. In our study, we employed the GLM function from the statsmodels.api Python library to identify explanatory variables that strongly correlate with the response variable, which is the accuracy of the decoy distribu-

© 2024 Information Processing Society of Japan

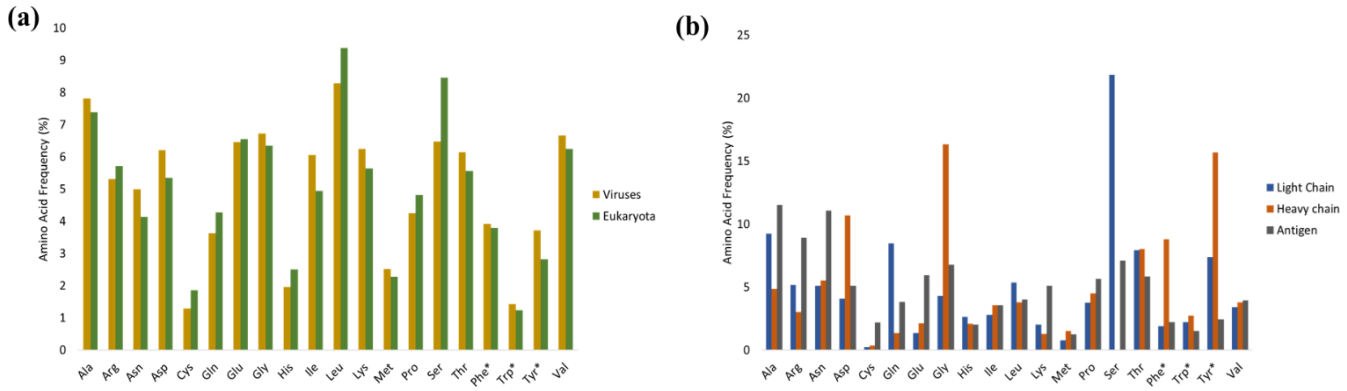



**Fig. 2** Amino Acids Frequency comparison. a) Amino acid frequency encoded in the genome of eukaryotes and viruses [41]. Yellow: Viruses and green: Eukaryotes. b) Amino acid frequencies of antibody CDR loops and Antigen epitope. Blue, orange, and gray bars reflect the frequency of residues on CDR loops (heavy chain, and light chain) and Ag, respectively. * Aromatic amino acids.

tions represented by Pearson's correlation coefficient of the $W_{def}$ scoring function and the $W_{nonsol}$ scoring ($W_1$ is ignored because all the funnels were bad).

The dataset used as input comprises the percentage occurrence of various explanatory variables: hydrophilicity, polarity, presence of aromatic residues, and surface area of each CDR loop and antigen epitope. The response variable reflects the outcome of the decoy distribution for each scoring function, categorized as either good (1) or bad (0) in a binary format.

## 3. Results

### 3.1 Docking Scoring Functions and Docking Funnels

The performance of a scoring function on an Ab-Ag complex can be explained in terms of the distribution of the docking decoys. We evaluated three different scoring functions: a function with the Rosetta default weights ($W_{def}$), a function with all changeable weights set to 1 ($W_1$), and a function with Rosetta default weights except for the solvation-based weights set to 0 ($W_{nonsol}$). For each scoring function, the respective docking funnels were created by plotting the generated decoys' RMSD versus energy score. As depicted in **Fig. 1**, certain decoy distributions clearly display a funnel-like structure with low decoy energies, whereas others lack any unambiguous structure. When studying the accuracy of the decoy distribution we identified four characteristics that distinguish a good decoy distribution from a poor one.

Theoretically, a good distribution consists of decoys displaying both the lowest RMSD and the lowest energy scores. Figure 1 (a) and 1 (b) represent two distinctive representations of relatively bad decoy distribution which lack a funnel-like decoy distribution shape. Figure 1 (c) is an example of a good distribution. The decoy distribution in Fig. 1 (c) is an excellent example of a "funnel" with a distinct form and low energy and RMSD decoys at its tip.

Namely, $W_{def}$, $W_1$, and $W_{nonsol}$, we counted the number of good and bad decoy distributions docking funnels (Fig. 1 (e) and 1 (f)). We revealed the impact of different scoring functions on the decoy distributions for Ab-Ag complexes. Specifically, we observed that the $W_1$ scoring function yielded 100 percent poor decoy distributions, whereas the $W_{nonsol}$ scoring function achieved a success rate of 25 percent (Supplementary document S5). To assess the relationship between accurate decoy distribution (a funnel-like distribution) and the scoring function, we conducted a chi-square test of independence. The obtained chi-square value of 11.544 exceeded the critical value of 3.841, rejecting the null hypothesis that the decoy distribution's accuracy is unrelated to the exclusion of solvation parameters from the default Rosetta scoring function ($p < 0.05$). Excluding solvation parameters from the scoring function significantly influences the outcome, resulting in a more funnel-like decoy distribution. The analysis of three scoring functions highlights the significance of considering solvation effects to enhance protein-protein docking accuracy.

### 3.2 Amino Acid Composition in the Ab-Ag Interface

The amino acid properties of the 100 Ab-Ag complexes in our dataset were analyzed, and the findings are depicted in **Fig. 2**. A comparison between the amino acid frequencies observed in our study and those in the genome revealed three notable discoveries.

Firstly, Ser emerged as the most prevalent amino acid in antibodies, particularly in the light chains CDR loops, constituting 21.8% of the total composition. Interestingly, Ser was completely absent in the heavy chains of antibodies. It was consistently present in the light chains of all 100 complexes but not observed in the heavy chains.

Secondly, despite being generally infrequent in eukaryotes, our study demonstrated a strong presence of Tyr in antibody-heavy chain CDR loops, accounting for 15.6% of the total composition. Tyr residues were found in the light chain fragments of 97 complexes, the heavy chain fragments of 97 complexes, and 84 epitopes in our dataset. Notably, Tyr residues were more abundant in the heavy chain CDR loops compared to the epitope regions and the light chain.

Lastly, aromatic amino acids such as Phe and Thr exhibited higher frequencies in the heavy chain CDR loops compared to their relatively low occurrence in eukaryotic genomes.

### 3.3 Significance of Physico-chemical Properties

To assess the impact of different explanatory variables on the

© 2024 Information Processing Society of Japan

| Parameter | | Wdef | | Wnonsol | |
|---|---|---|---|---|---|
| | | Coefficient | P | Coefficient | P |
| | constant | 3.832 | 0.076 | -0.983 | 0.745 |
| Hydrophilicity | L1 | -0.003 | 0.609 | -0.010 | 0.266 |
| | L2 | -0.004 | 0.307 | 0.005 | 0.356 |
| | L3 | -0.005 | 0.210 | -0.008 | 0.093 |
| | H1 | 0.004 | 0.478 | 0.003 | 0.670 |
| | H2 | 0.000 | 1.000 | 0.006 | 0.250 |
| | H3 | 0.000 | 0.904 | -0.001 | 0.793 |
| | Ag | -0.014 | 0.086 | 0.002 | 0.852 |
| Polarity | L1 | -0.002 | 0.732 | 0.001 | 0.916 |
| | L2 | -0.006 | 0.114 | 0.008 | 0.104 |
| | L3 | 0.007 | 0.076 | -0.002 | 0.692 |
| | H1 | -0.001 | 0.732 | -0.006 | 0.272 |
| | H2 | -0.004 | 0.228 | 0.001 | 0.804 |
| | H3 | 0.002 | 0.620 | 0.001 | 0.766 |
| | Ag | -0.003 | 0.666 | -0.011 | 0.287 |
| Non-aromatic residues | L1 | -0.016 | 0.259 | -0.006 | 0.756 |
| | L2 | -0.011 | 0.128 | 0.003 | 0.777 |
| | L3 | 0.005 | 0.218 | 0.015 | 0.016* |
| | H1 | -0.003 | 0.578 | -0.013 | 0.044* |
| | H2 | 0.002 | 0.751 | 0.007 | 0.300 |
| | H3 | 0.006 | 0.106 | 0.003 | 0.545 |
| | Ag | 0.012 | 0.273 | 0.010 | 0.514 |
| Ab-Ag Surface Area | L1 | 0.000 | 0.719 | 0.000 | 0.635 |
| | L2 | -0.001 | 0.047* | 0.000 | 0.954 |
| | L3 | 0.000 | 0.456 | 0.000 | 0.752 |
| | H1 | 0.000 | 0.361 | 0.000 | 0.816 |
| | H2 | 0.001 | 0.116 | 0.001 | 0.098 |
| | H3 | 0.000 | 0.896 | 0.000 | 0.756 |
| | Ag | 0.000 | 0.827 | 0.000 | 0.465 |

*** Significant**

**Fig. 3** Significance of independent parameters in Generalized Linear Regression (GLM). 1.0 < (P > |z|) < 0.05 Green, and yellow colors reflect the positive and negative correlation of significant parameters respectively.

accuracy of the decoy distribution, logistic regression with a generalized linear model (GLM) was performed on both the $W_{def}$, and $W_{nonsol}$ scoring functions. **Figure 3** illustrates the results, indicating that three variables demonstrated significant associations ($p < 0.05$) with the predictive power of decoy distribution. These variables include the presence of non-aromatic residues in the L3 and H1 regions (0.016 and 0.044, respectively) in the $W_{nonsol}$ scoring function, as well as the surface area of the Ab-Ag binding interface in the L2 region (0.047) in the $W_{def}$ scoring function [Supplementary documents S6–S7].

## 4. Discussion

Ab-Ag interactions play a crucial role in various biological processes, including immune response modulation, disease diagnosis, and therapeutic interventions. Understanding the physicochemical properties underlying these interactions is essential for evaluating their specificity, affinity, and overall stability. This article aims to delve deeply into the exploration of these physicochemical properties involved in Ab-Ag interactions, shedding light on their significance in the realm of drug discovery and design.

### 4.1 Evaluation of Docking Scoring Functions

This research represents an initial effort to assess the efficacy of Rosetta scoring functions in handling Ab-Ag complexes, employing both quantitative and qualitative measures through the practical use of decoy distributions, a common practice among scientists. The performance of scoring functions in Ab-Ag complex docking was evaluated based on the distribution of docking decoys. The analysis of three scoring functions ($W_{def}$, $W_1$, and $W_{nonsol}$) revealed that the removal of solvation parameters had a significant impact on the quality of docking performance, highlighting the need to account for solvation effects to improve the accuracy of Ab-Ag docking predictions.

While the results indicate that excluding solvation parameters affects the quality of docking performance, they do not provide a comprehensive understanding of all factors that influence the precision of docking scoring functions. Other factors influencing docking accuracy, such as hidden scoring terms, energy functions, and additional factors, were not directly addressed in this study. There is a need for additional research and investigation into the broader landscape of docking scoring functions and their overall performance in various docking scenarios, but it is beyond the goal of this research.

### 4.2 Influence of Physico-chemical Properties in CDR Loops and Epitopes on Docking

In analyzing 100 datasets comprising Ab-Ag bound complexes, we identified a conspicuous predominance of Tyr residues in the heavy chain CDR loops of antibodies, particularly enriched in the paratope regions. Conversely, Ser residues were completely absent from these heavy chain CDR loops, and Cys residues were notably sparse in both heavy and light chain CDR loops.

Previous studies have underscored the integral roles of Tyr, Ser, and Cys in antibodies [36], [37], [38], [39], [40]; however, they have predominantly done so within the confines of controlled environments, involving mutations in the epitope interacting regions of nonhuman immunoglobulins [36], [37], [38], and have not provided CDR location-specific insights into amino acid composition. Furthermore, these investigations were performed on unbound immunoglobulins, leaving the stability of Ab-Ag interactions uncertain and suggesting that the molecular recognition mechanisms of naïve or unbound antibodies differ from those of co-evolved or affinity-matured complexes [41], [42], [43], [44], [45], [46], [47].

Our research conclusively demonstrates the significant correlation of Tyr, Ser, and Cys with high-affinity, specific in-situ Ab-Ag complexes. The distinct positioning of these amino acids within the CDR loops of engaged Ab-Ag pairs is pivotal, highlighting an area that requires meticulous attention in the refinement of docking scoring algorithms for enhanced precision.

While the analysis provides valuable insights into the distribution and prevalence of specific amino acids within Ab-Ag complexes, there are limitations to consider. The results are confined to the specific dataset and conditions of the study and may not fully represent the entire population of Ab-Ag complexes or account for potential variations due to experimental setups, antibody types, antigen characteristics, or other factors. The observed variations in amino acid composition do not necessarily provide mechanistic insights into the functional implications or reasons for their presence or absence in CDR loops. Therefore, further research is necessary to explore the broader context, underlying mechanisms, functional implications, and generalizability of these findings.

### 4.3 Correlation between Physico-chemical Properties and Decoy Distribution

Through GLM analysis, we identified three critical variables impacting the predictive accuracy of decoy distribution within CDR segments: 1) the prevalence of non-aromatic residues in

© 2024 Information Processing Society of Japan

L3; 2) the scarcity of non-aromatic residues in H1; and 3) a smaller Ab-Ag binding interface in the L2 region. These findings highlight how the scoring function's accuracy fluctuates based on the abundance of non-aromatic residues and the interaction's surface area. Notably, the influence of aromatic residues persists even in scoring functions devoid of solvation parameters. While omitting solvation parameters can enhance Rosetta's performance with Ab-Ag structures, this alone doesn't fully negate the effects of irregular amino acids. Therefore, the development of scoring functions specialized for Ab-Ag interfaces is essential for more precise future Ab-Ag structural predictions.

The study offers important insights but is limited by its specific dataset and conditions, not fully encompassing the diversity of Ab-Ag complexes. A comprehensive understanding of the mechanisms dictating amino acid presence in CDR loops is missing, underscoring the need for broader research. This should include how amino acids influence the Rosetta scoring function, deeper exploration of Ab-Ag interactions, and enhancements in bio medicinal prediction models for wider applicability.

## 5. Conclusion

Progress in computational modeling, experimental methodologies, and high-throughput screenings enhances our grasp of the physicochemical intricacies involved in Ab-Ag interactions. The notable performance enhancement achieved by omitting the solvation parameter from the Rosetta scoring function highlighted the critical role this parameter plays. Nonetheless, it remains a crucial factor, necessitating a comprehensive evaluation of scoring functions with careful consideration of solvation impacts to accurately predict docking outcomes in Ab-Ag engagements. A deeper exploration of physicochemical characteristics can elucidate these interactions more clearly, facilitating the development of more refined scoring functions.

The physicochemical attributes of viable chemical substitutes are pivotal in evaluating Ab-Ag relations. Elements like hydrophilicity, hydrophobicity, polar versus nonpolar interactions, and the presence of particular amino acid residues contribute valuable information regarding the stability, specificity, and efficacy of Ab-Ag complexes. The synthesis of these methodologies propels the sophistication of computational techniques used in forecasting and enhancing Ab-Ag interactions, which is instrumental in the realms of drug innovation and formulation. Such insights hold substantial promise in propelling advancements in drug development, potentially leading to more effective therapeutic strategies and improved patient prognoses.

However, further research is imperative to unravel the elaborate nature of these interactions and their ramifications for therapeutic antibody production. Capitalizing on this understanding could elevate the precision and potency of antibody-centric treatments, culminating in more targeted and effective remedies for various ailments.

## 6. Abbreviations

**Ab:** Antibody
**Ag:** Antigen
**CDR:** Complementarity-determining region
**ELISA:** Enzyme-Linked Immunosorbent Assay
**Fab:** Fragment Antigen Binding
**ITC:** Isothermal Titration Calorimetry
**LK:** Lazaridis-Karplus
**PPI:** Protein-protein Interactions
**RMSD:** Root Mean Squared Deviation
**SPR:** Surface Plasmon Resonance

**Acknowledgments** We thank Prof. Luca Varani and Dr. Luca Simonelli from the Laboratory of Structural Biology, Institute for research in biomedicine, Università Della Svizzera Italiana, Via Vela 6, 6500 Bellinzona, Switzerland for providing the knowledge and the guidance from the beginning of this study.

### References

[1] Kringelum, J.V., Nielsen, M., Padkjær, S.B. and Lund, O.: Structural analysis of B-cell epitopes in antibody: Protein complexes, *Mol. Immunol.*, Vol.53, pp.24–34 (2013).
[2] Ramaraj, T., Angel, T., Dratz, E.A., Jesaitis, A.J. and Mumey, B.: Antigen-antibody interface properties: Composition, residue interactions, and features of 53 non-redundant structures, *Biochim. Biophys. Acta*, Vol.1824, pp.520–532 (2012).
[3] Mian, I.S., Bradwell, A.R. and Olson, A.J.: Structure, function and properties of antibody binding sites, *J. Mol. Biol.*, Vol.217, pp.133–151 (1991).
[4] Morris, G.M. and Lim-Wilby, M.: Molecular docking, *Methods Mol. Biol.*, Vol.443, pp.365–382 (2008).
[5] Kramer, B., Rarey, M. and Lengauer, T.: Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking, *Proteins*, Vol.37, pp.228–241 (1999).
[6] Böhm, H.-J.: The development of a simple empirical scoring function to estimate the binding constant for a protein–ligand complex of known three-dimensional structure, *J. Comput. Aided Mol. Design*, Vol.8, pp.243–256 (1994).
[7] Verdonk, M.L., Cole, J.C., Hartshorn, M.J., Murray, C.W. and Taylor, R.D.: Improved protein-ligand docking using GOLD, *Proteins*, Vol.52, pp.609–623 (2003).
[8] Halgren, T.A., Murphy, R.B., Friesner, R.A., Beard, H.S., Frye, L.L., Pollard, W.T. and Banks, J.L.: Glide: A new approach for rapid, accurate docking and scoring, 2. Enrichment factors in database screening, *J. Med. Chem.*, Vol.47, pp.1750–1759 (2004).
[9] Hill, A.D. and Reilly, P.J.: Scoring functions for AutoDock, *Methods Mol. Biol.*, Vol.1273, pp.467–474, DOI: 10.1007/978-1-4939-2343-4_27 (2015).
[10] Trott, O. and Olson, A.J.: AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading, *J. Comput. Chem.*, Vol.31, No.2, pp.455–461, DOI: 10.1002/jcc.21334 (2010).
[11] Macindoe, G., Mavridis, L., Venkatraman, V., Devignes, M.-D. and Ritchie, D.W.: HexServer: An FFT-based protein docking server powered by graphics processors, *Nucleic Acids Res.*, Vol.38, pp.W445–W449, DOI: 10.1093/nar/gkq311 (2010).
[12] Tobi, D.: Designing coarse grained-and atom based-potentials for protein-protein docking, *BMC Struct. Biol.*, Vol.10, No.40, DOI: 10.1186/1472-6807-10-40 (2010).
[13] Mian, I.S., Bradwell, A.R. and Olson, A.J.: Structure, function and properties of antibody binding sites, *J. Mol. Biol.*, Vol.217, pp.133–151 (1991).
[14] Ofran, Y., Schlessinger, A. and Rost, B.: Automated identification of complementarity determining regions (CDRs) reveals peculiar characteristics of CDRs and B cell epitopes, *J. Immunol.*, Vol.181, pp.6230–6235 (2008).
[15] Yu, C.M. et al.: Rationalization and design of the complementarity determining region sequences in an antibody-antigen recognition interface, *PLoS ONE*, Vol.7, e33340 (2012).
[16] Peng, H.P., Hsu, H.J., Yu, C.M., et al.: Antibody CDR amino acids underlying the functionality of antibody repertoires in recognizing diverse protein antigens, *Sci. Rep.*, Vol.12, 12555, DOI: 10.1038/s41598-022-16841-9 (2022).
[17] Koehler Leman, J., Weitzner, B.D., Renfrew, P.D., Lewis, S.M., Moretti, R., Watkins, A.M., Mulligan, V.K., Lyskov, S., Adolf-Bryfogle, J., Labonte, J.W., Krys, J., Bystroff, C., Schief, W., Gront, D., SchuelerFurman, O., Baker, D., Bradley, P., Dunbrack, R., Kortemme, T., Leaver-Fay, A., Strauss, C.E.M., Meiler, J., Kuhlman, B., Gray, J.J. and Bonneau, R.: Better together: Elements of successful scientific software development in a distributed collaborative commu-

© 2024 Information Processing Society of Japan

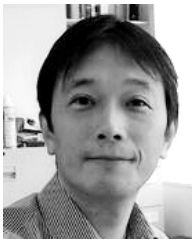
nity, *PLoS Comput. Biol.*, Vol.16, e1007507 (2020).
[18] Kellogg, E.H., Leaver-Fay, A. and Baker, D.: Role of conformational sampling in computing mutation-induced changes in protein structure and stability, *Proteins*, Vol.79, No.3, pp.830–838, DOI: 10.1002/prot.22921 (2011).
[19] Chaudhury, S., Berrondo, M., Weitzner, B.D., Muthu, P., Bergman, H. and Gray, J.J.: Benchmarking and analysis of protein docking performance in Rosetta v3.2, *PLoS One*, Vol.6, No.8, e22477, DOI: 10.1371/journal.pone.0022477 (2011).
[20] Alford, R.F., Leaver-Fay, A., Jeliazkov, J.R., O'Meara, M.J., DiMaio, F.P., Park, H., Shapovalov, M.V., Renfrew, P.D., Mulligan, V.K., Kappel, K., Labonte, J.W., Pacella, M.S., Bonneau, R., Bradley, P., Dunbrack, R.L. Jr., Das, R., Baker, D., Kuhlman, B., Kortemme, T. and Gray, J.J.: The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design, *J. Chem. Theory Comput.*, Vol.13, No.6, pp.3031–3048, DOI: 10.1021/acs.jctc.7b00125 (2017).
[21] Schoeder, C.T., Schmitz, S., Adolf-Bryfogle, J., Sevy, A.M., Finn, J.A., Sauer, M.F., et al.: Modeling Immunity with Rosetta: Methods for Antibody and Antigen Design, *Biochemistry*, Vol.60, No.11, pp.825–846 (2021).
[22] Bender, B.J., Cisneros, A., 3rd, Duran, A.M., Finn, J.A., Fu, D., Lokits, A.D., Mueller, B.K., Sangha, A.K., Sauer, M.F., Sevy, A.M., Sliwoski, G., Sheehan, J.H., DiMaio, F., Meiler, J. and Moretti, R.: Protocols for Molecular Modeling with Rosetta3 and RosettaScripts, *Biochemistry 2016*, Vol.55, pp.4748–4763 (2016).
[23] Kotthoff, I., Kundrotas, P.J. and Vakser, I.A.: Dockground scoring benchmarks for protein docking, *Proteins*, Vol.90, No.6, pp.1259–1266, DOI: 10.1002/prot.26306 (2022).
[24] Schoeder, C.T., Schmitz, S., Adolf-Bryfogle, J., Sevy, A.M., Finn, J.A., Sauer, M.F., Bozhanova, N.G., Mueller, B.K., Sangha, A.K., Bonet, J., Sheehan, J.H., Kuenze, G., Marlow, B., Smith, S.T., Woods, H., Bender, B.J., Martina, C.E., del Alamo, D., Kodali, P., Gulsevin, A., Schief, W.R., Correia, B.E., Crowe Jr., J.E., Meiler, J. and Moretti, R.: Modeling Immunity with Rosetta: Methods for Antibody and Antigen Design, *Biochemistry 2021*, Vol.60, No.11, pp.825–846, DOI: 10.1021/acs.biochem.0c00912 (2021).
[25] Schneider, C., Raybould, M.I.J. and Deane, C.M.: SAbDab in the age of biotherapeutics: Updates including SAbDab-nano, the nanobody structure tracker, *Nucleic Acids Research*, Vol.50, No.D1, pp.D1368–D1372, DOI: 10.1093/nar/gkab1050 (2022).
[26] Vita, R., Mahajan, S., Overton, J.A., Dhanda, S.K., Martini, S., Cantrell, J.R., Wheeler, D.K., Sette, A. and Peters, B.: The Immune Epitope Database (IEDB): 2018 update, *Nucleic Acids Res.*, DOI: 10.1093/nar/gky1006.
[27] Jespersen, M.C., Peters, B., Nielsen, M. and Marcatili, P.: BepiPred-2.0: Improving sequence-based B-cell epitope prediction using conformational epitopes, *Nucleic Acids Res.*, Vol.45, No.W1, pp.W24–W29, DOI: 10.1093/nar/gkx346.
[28] Kyte, J. and Doolittle, R.F.: A simple method for displaying the hydropathic character of a protein, *J. Mol. Biol.*, Vol.157, No.1, pp.105–132, DOI: 10.1016/0022-2836(82)90515-0 (1982).
[29] Miller, S., Janin, J., Lesk, A.M. and Chothia, C.: Interior and surface of monomeric proteins, *J. Mol. Biol.*, Vol.196, No.3, pp.641–656, DOI: 10.1016/0022-2836(87)90038-6 (1987).
[30] Zimmerman, J.M., Eliezer, N. and Simha, R.: The characterization of amino acid sequences in proteins by statistical methods, *J. Theor. Biol.*, Vol.21, No.2, pp.170–201, DOI: 10.1016/0022-5193(68)90069-6 (1968).
[31] Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H. and Zehfus, M.H.: Hydrophobicity of amino acid residues in globular proteins, *Science*, Vol.229, No.4716, pp.834–838, DOI: 10.1126/science.4023714 (1985).
[32] Kuo, H.T., Yeh, J.Z., Wu, P.H., Jiang, C.M. and Wu, M.C.: Application of immunomagnetic particles to enzyme-linked immunosorbent assay (ELISA) for improvement of detection sensitivity of HCG, *J. Immunoassay Immunochem.*, Vol.33, No.4, pp.377–387 (2012).
[33] Zeng, S., Baillargeat, D., Ho, H.P. and Yong, K.T.: Nanomaterials enhanced surface plasmon resonance for biological and chemical sensing applications, *Chem. Soc. Rev.*, Vol.43, No.10, pp.3426–3452, DOI: 10.1039/c3cs60479a (2014).
[34] Grolier, J.-P.E. and del Río, J.M.: Isothermal titration calorimetry: A thermodynamic interpretation of measurements, *The Journal of Chemical Thermodynamics*, Vol.55, pp.193–202, DOI: 10.1016/j.jct.2012.05.018 (2012).
[35] Kozlowski, L.P.: Proteome-pI 2.0: Proteome isoelectric point database update, *Nucleic Acids Res.*, Vol.50, No.D1, pp.D1535–D1540, DOI: 10.1093/nar/gkab944 (2022).
[36] Padlan, E.A.: On the nature of antibody combining sites: Unusual structural features that may confer on these sites an enhanced capacity for binding ligands, *Proteins*, Vol.7, pp.112–124, DOI: 10.1002/prot.340070203 (1990).
[37] Tsumoto, K., Ogasahara, K., Ueda, Y., Watanabe, K., Yutani, K. and Kumagai, I.: Role of Tyr residues in the contact region of antilysozyme monoclonal antibody HyHEL10 for antigen binding, *J. Biol. Chem.*, Vol.270, pp.18551–18557 (1995).
[38] Shiroishi, M., Tsumoto, K., Tanaka, Y., Yokota, A., Nakanishi, T., Kondo, H. and Kumagai, I.: Structural consequences of mutations in interfacial Tyr residues of a protein antigen-antibody complex, The case of HyHEL-10-HEL, *J. Biol. Chem.*, Vol.282, pp.6783–6791 (2007).
[39] Steinmaurer, A, Wimmer, I., Berger, T., Rommer, P.S. and Sellner, J.: Bruton's Tyrosine Kinase Inhibition in the Treatment of Preclinical Models and Multiple Sclerosis, *Curr. Pharm. Des.*, Vol.28, No.6, pp.437–444, DOI: 10.2174/1381612827666210701152934 (2022).
[40] Koide, S. and Sidhu, S.S.: The Importance of Being Tyrosine: Lessons in Molecular Recognition from Minimalist Synthetic Binding Proteins, *ACS Chemical Biology*, Vol.4, No.5, pp.325–334, DOI: 10.1021/cb800314v (2009).
[41] Hagihara, Y. and Saerens, D.: Engineering disulfide bonds within an antibody, *Biochimica Et Biophysica Acta* (*BBA*) - *Proteins and Proteomics*, Vol.1844, No.11, pp.2016–2023, DOI: 10.1016/j.bbapap.2014.07.005 (2014).
[42] Birtalan, S., Zhang, Y., Fellouse, F.A., Shao, L., Schaefer, G. and Sidhu, S.S.: The Intrinsic Contributions of Tyrosine, Serine, Glycine and Arginine to the Affinity and Specificity of Antibodies, *Journal of Molecular Biology*, Vol.377, No.5, pp.1518–1528, DOI: 10.1016/j.jmb.2008.01.093 (2008).
[43] Fellouse, F.A., Wiesmann, C. and Sidhu, S.S.: Synthetic antibodies from a four-amino-acid code: A dominant role for tyrosine in antigen recognition, *Proc. National Academy of Sciences*, Vol.101, No.34, pp.12467–12472, DOI: 10.1073/pnas.0401786101 (2004). Structure, Function and Properties of Antibody Binding Sites – ScienceDirect, DOI: 10.1016/0022-2836(91)90617-F
[44] Martins, J.N., Lima, J.C. and Basílio, N.: Selective Recognition of Amino Acids and Peptides by Small Supramolecular Receptors, *Molecules*, Vol.2021, No.26, p.106, DOI: 10.3390/molecules26010106 (2021).
[45] Janin, J.: Wet and dry interfaces: The role of solvent in protein-protein and protein-DNA recognition, *Structure*, pp.R277–R279 (1999).
[46] Ma, H., Ó'Fágáin, C. and O'Kennedy, R.: Antibody stability: A key to performance - Analysis, influences and improvement, *Biochimie*, Vol.177, pp.213–225, DOI: 10.1016/j.biochi.2020.08.019 (2020).
[47] Lazaridis, T. and Karplus, M.: Effective energy function for proteins in solution, *Proteins*, Vol.35, No.2, pp.133–152, DOI: 10.1002/(sici)1097-0134(19990501)35:2<133::aid-prot1>3.0.co;2-n (1999).

**Sangeetha Ratnayake** is a dedicated researcher in the field of computational biology and bioinformatics. She is currently pursuing her Ph.D. at Hokkaido University, Japan, focusing on the design of computational docking scoring functions tailored to antibody-antigen interfaces. Sangeetha has a strong foundation in data analysis, pattern recognition, and prediction modeling using Python and R. Her research areas are computational and structural bioinformatics, genomics, prediction modeling, bigdata analysis and machine learning. Her commitment to research-driven approaches and her expertise in computational biology make her a valuable contributor to the field.

© 2024 Information Processing Society of Japan

**Axel Martinelli** is a highly accomplished scientist and Chief Scientific Officer (CSO) at BigOmics. With expertise in genomics and bioinformatics, they have made significant contributions to the study of parasite genomics, particularly malaria. Axel's research, including papers like "Rapid identification of genes controlling virulence and immunity in malaria parasites," showcases their deep understanding of the genetic factors underlying diseases. They hold advanced degrees in Bioinformatics, Biology, and Immunology from renowned universities. Axel's commitment to continuous learning is demonstrated through their completion of courses in Operations Management, Financial Accounting, and more. Their expertise and dedication position them as a respected figure in scientific research and leadership.

**Toshinori Endo** received his Ph.D. in Genetics from The Graduate University for Advanced Studies, SOKENDAI in 1996. He experienced post-docs at National Institute of Genetics from 1996, Biolabs, Harvard University from 1997, and Genome Sciences Center, RIKEN from 1999. He was a lecturer at the Medical Research Institute, Tokyo Medical and Dental University from 2000 to 2004. He became a full professor of bioinformatics at Graduate School of Information Science Technology, Hokkaido University in 2004. His major interests are molecular evolution, prediction, and design of gene functions.



**Naoki Osada** is an associate professor at Hokkaido University, Japan. He received his Ph.D. in 2002 at the University of Tokyo for his work on genomics and the molecular evolution of primates. He worked as a post-doctoral researcher at the University of Tokyo from 2002 to 2003, and the University of Chicago from 2003 to 2004. During that time, he studied bioinformatics and population genetics for speciation. He worked at the National Institute of Biomedical Innovation from 2005 to 2010 as a Scientific Researcher. From 2010 to 2015, he worked at the National Institute of Genetics as an assistant professor. He is currently working on a variety of research topics in molecular evolution and population genetics, including the study of human/primate evolution and gene expression evolution.

(Communicated by *Tetsuo Shibuya*)

© 2024 Information Processing Society of Japan