Thomas Minotto*, Philippe A. Robert, Ingrid Hobæk Haff and Geir K. Sandve

Assessing the feasibility of statistical inference using synthetic antibody-antigen datasets

https://doi.org/10.1515/sagmb-2023-0027
Received June 23, 2023; accepted March 13, 2024; published online April 3, 2024

Abstract: Simulation frameworks are useful to stress-test predictive models when data is scarce, or to assert model sensitivity to specific data distributions. Such frameworks often need to recapitulate several layers of data complexity, including emergent properties that arise implicitly from the interaction between simulation components. Antibody-antigen binding is a complex mechanism by which an antibody sequence wraps itself around an antigen with high affinity. In this study, we use a synthetic simulation framework for antibody-antigen folding and binding on a 3D lattice that include full details on the spatial conformation of both molecules. We investigate how emergent properties arise in this framework, in particular the physical proximity of amino acids, their presence on the binding interface, or the binding status of a sequence, and relate that to the individual and pairwise contributions of amino acids in statistical models for binding prediction. We show that weights learnt from a simple logistic regression model align with some but not all features of amino acids involved in the binding, and that predictive sequence binding patterns can be enriched. In particular, main effects correlated with the capacity of a sequence to bind any antigen, while statistical interactions were related to sequence specificity.

Keywords: antibody-antigen binding; synthetic data; emergent properties; logistic regression; main effects; statistical interactions

1 Introduction

Characterizing how specific properties emerge from a set of simple rules, and reversely how complex behaviour can be decomposed into a set of simple rules, is of high importance in the understanding and accurate modeling of natural phenomena. Emergence can be defined as "the global behaviours or patterns that arise through 'self-organization' and that could not have otherwise been characterized *a priori*" (Chavali et al. 2008). There have been various attempts at characterizing emergent properties in simulated data, for instance how individual fish behaviour affects the general properties of a fish school (Parrish et al. 2002; Viscido et al. 2004), the effect of thermodynamic parameters on climate models used to study convection (Raymond and Fuchs-Stone 2021), or how parameters in models based on differential equations lead to group-level properties such as resilience or diversity in ecological modelling (van den Berg et al. 2022), to name a few. In the domain of immunology, Chavali et al. have described how emergent properties in a population of T-cells such as activation and proliferation arise from the agent-based modeling and cellular automata approaches.

***Corresponding author: Thomas Minotto**, Department of Mathematics, University of Oslo, Oslo, Norway, E-mail: thomamin@uio.no
Philippe A. Robert, Department of Immunology, University of Oslo and Oslo University Hospital, Oslo, Norway; and Departmemt of Biomedicine, University of Basel, Basel, Switzerland, E-mail: philippe.robert@ens-lyon.org
Ingrid Hobæk Haff, Department of Mathematics, University of Oslo, Oslo, Norway, E-mail: ingrihaf@math.uio.no
Geir K. Sandve, Department of Informatics, University of Oslo, Oslo, Norway, E-mail: geirksa@ifi.uio.no

Protein function is often mediated by binding to other proteins, and is intimately linked with 3D structure, which is a deterministic emergent property of the sequence. Understanding how sequence translates to function is still very challenging and most recent tools succeed at predicting protein structure, yet, predicting protein-protein binding still needs improvement (Yin et al. 2022). A key example is the recognition of pathogens by antibodies produced by the immune system (immune receptors). These receptors are generated with a gigantic sequence diversity and can selectively bind their target with high affinity (antigens) (Greiff et al. 2017).

In the binding of immune receptors to antigens, the binding landscape of an antibody (which antigen targets it binds) is mainly determined by a stretch of about 15 amino acids long on average, called CDRH3 (Complementarity Determining Region 3 of the heavy chain). Interactions between the CDRH3 and the antigen are governed by various physico-chemical rules, and the resulting behaviour of the whole chain of molecules is complex and challenging to model.

Experimentally, high-throughput sequencing methods provide large scale sequencing of binding antibodies to predefined antigens of interest, without knowing the structural properties of their binding (i.e. the emergent properties responsible for their biological function). This is a problem because improving protein binding or designing therapeutical compounds that modulate binding requires to know how the proteins bind (i.e. which amino acids are involved in the interaction between the proteins). Starting from large-scale data linking sequence of proteins to their matching epitopes, one would like to infer the hidden structural properties of their binding (statistical inference). Due to the extremely small size of ground truth datasets (for example, the AbDb database has around 1200 non redundant antibody-antigen structures (Ferdous and Martin 2018)), applying statistical inference tools to sequence datasets without knowing the structural binding for many proteins (the ground truth) is prone to fail because there is no target to achieve. However, generation of synthetic datasets can provide access to ground truth structural properties, provided the simulation framework reproduces enough levels of complexity of structural binding.

In this article, we work on simulated data with unconstrained size from a model of simplified antibody-antigen structures (Robert et al. 2022). There, individual amino acids are positioned on a 3D lattice, and the energy of interaction between neighboring amino acids is computed from an empirical potential energy function (Miyazawa–Jernigan contact potential, Miyazawa and Jernigan (1999)) that depends on their nature. Following these rules, a given antibody chain wraps itself around a target antigen so as to minimize the total potential energy, and structures with the lowest potential energy are considered binding structures.

In this setting, the set of simple rules is the folding possibilities in the 3D lattice, together with the potential energy for each configuration, and the emergent property is the capacity of an antibody sequence to reach a conformation that will bind an antigen with high affinity in the simulation model. The computation of the binding energy is based on a sub-slide of 11 amino acids of the receptor, which may follow almost any consecutive path through the lattice, meaning that the number of possible docking structures ranges in the millions for a given amino acid chain. It is thus not possible to infer *a priori* the optimal structure and binding potential, hence the relevance of seeing them as emergent properties. In terms of complexity, these synthetic datasets lie between real-world data where we do not know the underlying truth (Greiff et al. 2020), and more simple simulated data where manually implanted statistical interactions or motifs are the only determinants of binding, used to simulate data (Kanduri et al. 2022; Pavlović et al. 2021) or assumed to analyse experimental data (Glanville et al. 2017; Ostmeyer et al. 2019). This yields non-trivial emergent properties that can be studied in the detail as we have access to all the underlying processes.

In this study, we propose to investigate the contribution of the type and position of individual amino acids, and statistical interactions between pairs of them at different distances, and see how this low-level information can be related to higher-level information on binding. The contribution of individual amino acids is assessed from a logistic regression model of the main effects, and the contribution of pairs of amino acids is revealed using the statistical interaction detection method NID (Neural Interaction Detection, Tsang et al. 2018). We distinguish between two types of observable features: sequence, that is the observed amino acid composition, and structure, that is their arrangement in space. For the structural information, we focus on the binding degree of the amino acids, whether they are part of the paratope, their proximity in space, and whether they have a

common bond on the epitope. We assert whether information retrieved by the statistical methods at the lower-level of the sequence can be linked to the structural information at an intermediate and hidden level in this simulated framework, and at a higher level to the binding information with stickiness and specificity of the sequences.

Altogether, we showcase the usefulness of using synthetic datasets with controllable properties and known ground-truth in the development of statistical inference tools to identify emergent properties from protein sequences datasets, that are directly applicable to experimental datasets when they become available.

2 Materials and methods

The model for synthetic data takes the sequence of an antibody and the structure of an antigen, and calculates their energetically optimal binding structure. For most antigens, their structure is already known and can be mapped into the 3D lattice of the model, while the antibodies are extremely diverse and flexible and different wrapping conformations must be tested to find the optimal structure. We use the murine data from the synthetic dataset, available in the NIRD research data archive (Robert et al. 2021). Binding energies of antibody-antigen structures have been computed for 159 antigens and 6.9 millions antibodies per antigen, annotated with their 3D binding structure to the respective antigen. The top 1 % affinity sequences are defined as binding antibody sequences, while non-binding sequences are the top 1 %–5 % affinity sequences, and sequences with less affinity are considered noise. The sequences outside the top 5 % affinity can also be considered as non-binders, but including them in the data brings more heterogeneity. For this reason, we exclude sequences outside the top 5 % of binding affinity when fitting statistical models, as they are not binding at all and may represent unspecific stickyness. This reduces the noise in the training data and helps to identify better which amino acids or groups of them are responsible for binding. Further, in the original paper (Robert et al. 2022), authors have shown that the top 1 % and 5 % sequences seem to contain the key information for predicting binding specificity. We only include these sequences when measuring high-level properties related to binding that are described below. Immune sequences are represented as chains of amino acids that correspond to the CDRH3 part of the receptor and can be of size 11 or longer. The binding energy of the antibody-antigen complex is computed based on an optimal slice of 11 amino acids of the antibody. The dataset contains detailed information on the antibody-antigen structures. In particular, we use the full amino acid composition of the 11-mer slice of the antibody, the binding degree of each antibody amino acid, defined as the number of bonds it has on the epitope, the presence on the paratope-epitope interface of the amino acids from both chains, defined here as the region where amino acids of one chain are in direct proximity with amino acids of the other chain, the proximity in the 3D lattice within the antibody chain and the binding energy of the structure. In this model, a bond is defined as two amino acids coming from each chain that are at a distance 1 from each other in the lattice, i.e. there can be no other amino acid in between. Distances greater than 1 do not form bonds.

We focus by default on binding and non-binding antibody sequences, discarding other sequences with less affinity unless mentioned otherwise. The statistical methods require that amino acid composition is one-hot encoded into a binary array with 220 covariates (11 positions times 20 possible amino acids).

From a processed data of around 20,000 binders and 80,000 non-binders to a given antigen (exact numbers vary slightly by antigen), we proceed to a simple analysis of individual amino acids (main effects). For this, we fit an unpenalised logistic regression model where the predictor is the one-hot encoded immune sequence and the response variable is its binding status (1 for binding, 0 for non-binding). From the fitted model, we study which covariates have the highest and lowest regression coefficients, and link this to the binding degree and potential presence on the paratope of the corresponding amino acids.

For each sequence that contains the amino acid of interest (conditioning), we retrieve this information to compute its mean binding degree and probability of being on the paratope. More precisely, if we denote by X the immune sequence, M the motif (amino acid of interest), D_M its binding degree, P_M the event that M belongs to the paratope, these two quantities can be denoted as $\mathbb{E}_X(D_M|M \in X)$ and $\mathbb{P}_X(P_M|M \in X)$, and are computed empirically by

$$\mathbb{E}_X(D_M|M \in X) = \frac{\sum_{X, M\in X} D_M}{|\{X, s.t. M \in X\}|}, \quad \mathbb{P}_X(P_M|M \in X) = \frac{\sum_{X, M\in X} \mathbb{1}_{M\in\text{paratope}}}{|\{X, s.t. M \in X\}|}.$$

This yields a conditional expectation and a conditional probability, respectively.

We then study statistical interactions within pairs of amino acids, and see whether they correspond to particular structural properties of the pairs. These interactions must be understood in the statistical sense that the paired effect on the response is different from the sum of marginal effects (see for example Sorokina et al. (2008) for more details on statistical interactions). To detect statistical interactions, we use the detection method NID (Neural Interaction Detection, Tsang et al. 2018). It consists in first fitting a fully connected feed-forward neural network to the data, then studying the fitted weights of the network to search for interactions. Here, weights arriving at the first hidden layer correspond to the strength of the statistical interactions at this layer, and weights going towards the other layers correspond to their actual influence on the network. Both types of information are combined to give the total strength of the statistical interactions in the model. For instance, for given covariates x_1 and x_2 corresponding to 2 amino acids at different positions in the chain, the NID method looks at the weights from these input covariates to the units of the first hidden layer.

If these particular weights are high, then x_1 and x_2 are likely to interact in the network, and this is further measured by studying the weights from the first hidden layer to the rest of the network. The input data is the one-hot encoded sequences with their binding labels, and interactions are returned as a list of pairs of covariates, ordered by decreasing strength. The method is configured with 3 hidden layers with 100, 60 and 20 neurons each, and binary cross entropy is used as the loss function. The learning rate is set to 10^{-2}, the L1 constant to $5e-4$. We do not use a main effect net, and we use equal sizes for the training, validation and test sets during the learning process. This is based on authors usage in their paper.

We chose this detection method because it is adapted to binary and categorical data, it has a reasonable running time of around a few minutes for our dataset dimensions, and gathers information on several thousands potential interacting pairs. For each antigen, we compare statistical interactions to structural information on the corresponding pairs of amino acids. This includes the average sum of their binding degrees (bindingDegree), the probability that both are part of the paratope (bothOnParatope), the probability that they are in proximity in space when they are not direct neighbours on the sequence (intraAntibodyInteraction), the probability that they have a common bond on the epitope (sameEpitopeTarget) and the probability that they both have a binding degree of D or more, with $D = 2$ or $D = 3$ (bindingDegree_DandMore).

Again, we condition the expectancy and the probabilities on the sequences that contain the motif of interest. We study whether these quantities increase with the strength of statistical interactions.

In addition to statistical interaction strength, we also measure positional dependencies within pairs of amino acids $M = (i, j)$,

$$\frac{\Delta_{i,j} + \Delta_{j,i}}{2}\text{, where } \Delta_{i,j} = \mathbb{P}(X_i = 1 | X_j = 1, Y = 1) - \mathbb{P}(X_i = 1 | X_j = 1, Y = 0)$$

with X_i and X_j binary covariates from the one-hot encoding and Y the binding status, and see how it relates to the structural information.

Further, we study a couple of high-level properties related to binding, namely affinity, stickiness and specificity, and link these to both main effects and statistical interaction contributions in a logistic regression model for binding prediction. To define these properties, we use the binding energy of antibody-antigen structures, where the lowest energies for a given antigen correspond to antibodies that bind it the most. For an antibody AB, its affinity $AF_{AB \to AG}$ to a given antigen AG is defined as 0 for the worst binder (highest binding energy), 1 for the best binder (lowest binding energy), and intermediate values at a constant interval from each other for the remaining antibodies ordered by decreasing energy. This way, the top 1 % affinity sequences will correspond to sequences with an affinity score above 0.99. For each antigen AG, we can compute the affinity of all related antibody sequences. It is also possible to use a different scaling for the binding affinity that is not based on constant intervals. For instance, we can use a proportionality rule tuned to transform the original energy values into affinity values between 0.99 and 1 for the binders, and the rest of the sequences then have affinity values typically between 0.90 and 0.99. This kind of rule allows to take into account the distance in binding affinity and highlight the binders. We verify in Supplementary Figure S12 that the conclusions drawn from the first definition of the affinity are not different from what we obtain with this other definition. To avoid biases due to different binding energy distribution among antigens, we focus on a sub-group of 80 antigens from the dataset with similar binding energy profiles, defined as those antigens where the top 1 % affinity sequences have a maximum energy between −110 and −90. A histogram of these energies is displayed in Supplementary Figure S1.

Then, we define the stickiness of an antibody sequence AB as the average value of affinity $\overline{AF_{AB}}$ over the antigens in which it appears. However, when we cross information from different antigens, binders to one antigen are not necessarily binders to another antigen, and the same yields for other sequences in the top 5 % affinity, hence a lack of information if using only these sequences. To reduce a bit that effect, we keep binders as is in the data, but we replace non-binders (with affinity between 0.95 and 0.99) by a random selection of 500,000 sequences with affinity ranging from 0 to 0.99, among the remaining antibodies of the total 6.9 million. This enables to have sequences that are more likely to appear with different antigen, and obtain more robust values for the average. In addition, we will focus only on those sequences that appear in at least 10 different antigens. Affinity, stickiness and specificity values are computed on this modified dataset. Using the stickiness, we define the specificity of an antibody sequence to a particular antigen as

$$SP_{AB \to AG} = AF_{AB \to AG} - \overline{AF_{AB}}.$$

There, a positive value corresponds to an antibody specific to the antigen, meaning that it binds better to this antigen than to a random antigen. Limiting values are 1 and −1.

To relate the different measures to the probability of binding, we use the log-odds of a logistic regression model for binding prediction. This model can be written as

$$\log\left(\frac{P(Y=1|X=x)}{1-P(Y=1|X=x)}\right) = \beta_0 + \sum_{i \in \text{main effects}} \beta_i x_i + \sum_{(j,k) \in \text{interactions}} \beta_{j,k} x_j \cdot x_k$$

where $P(Y = 1|X = x)$ is the probability that a given sequence x binds (label Y has value 1), β_0 is an intercept term, the β_is are model parameters, the x_is are the values of the binary covariates X_i, and $x_j \cdot x_k$ represents a statistical interaction between two binary covariates X_j and X_k. The logarithm of the ratio of probabilities is called the log-odds, and if it has value above (below) 0 it indicates that sequence X is more likely to (not) bind. By studying the values of $\sum_{i \in \text{main effects}} \beta_i x_i$ and $\sum_{(j,k) \in \text{interactions}} \beta_{j,k} x_j \cdot x_k$ for different sequences, we can observe whether the main effects or the interactions are most responsible for the binding.

3 Results

3.1 Correlation between main effects from logistic regression and binding properties

We study 5 antigens from the dataset, 1ADQ_A, 1FBI_X, 1FNS_A, 1FSK_A and 1H0D_C, and for each of them we select binding and non-binding sequences from the top 5 % affinity. After removing those that correspond to the same 11-mer slice (around 40 % of the total), we select half of them as a training set for our analysis and focus only on their 11-mer slice. This constitutes a dataset of around 100,000 sequences, of which 20 % are binders.

A logistic regression model of binding prediction with only the main effects is fitted to the 11-mer slices by maximum likelihood without regularisation, using the function `glm()` in *R*. After one-hot encoding into 220 binary covariates, the columns corresponding to amino acid M are removed to set a reference and avoid collinearity. This amino acid was selected because it does not appear too often in the sequences, but other amino acids could have been selected as well. Not setting a reference this way would mean that after one-hot encoding the sum of all binary covariates for a given position is always 1, creating several situations of collinearity. We investigated briefly the impact of setting as a reference another amino acid, for instance N or P, and found that numerical values were slightly different, but the overall trends remained the same (results not shown). To avoid numerical issues and undefined coefficients, we also remove columns where the proportion of 0 s or 1 s is below 0.1 %, or that have a correlation over 0.7 in absolute value with another column, but these are rare events concerning only a few columns for each antigen.

For the structural information we compute from all sequences X the average binding degree $\mathbb{E}(D_M|M \in X)$ of each amino acid at each position M. This yields a real number between 0 and 5. We also compute the probability of belonging to the paratope $\mathbb{P}(P_M|M \in X)$. These two quantities are plotted in Figure 1, against the estimated coefficients of the logistic regression model, and for five different antigens.

Given the high ratio of observations to covariates, and removal of potentially detrimental covariates, coefficient values can be trusted to approximate their true values well. The standard errors on estimated coefficients were below 0.62 in all 5 fitted models. For antigens 1ADQ_A, 1FSK_A and 1H0D_C, amino acids with a positive regression coefficient tend to have a binding degree over 2, while it is below 2 for amino acids with a negative regression coefficient. For the other antigens, a minimum degree of 3 is observed, although the trend is less clear.

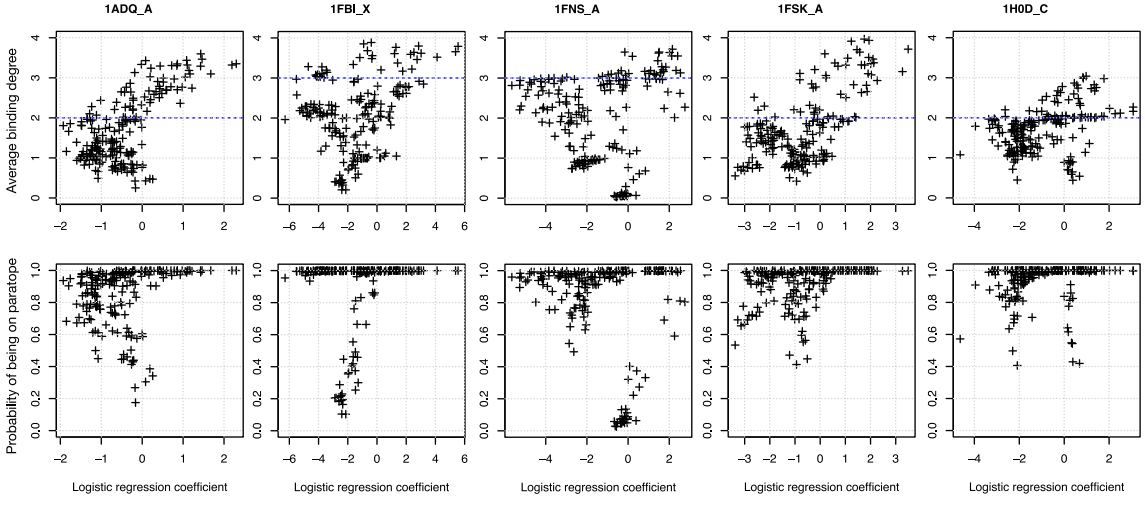

Figure 1: Average binding degree and probability of being on the paratope, against estimated regression coefficients of the amino acids. High positive coefficients are associated with a binding degree above 2 or 3, and an amino acid that is almost always on the paratope.

When it happens, this suggests that amino acids with the highest binding degrees impact binding positively when they are present. Besides, a higher average binding degree might indicate stronger conditions for binding.

Concerning the probability of an amino acid to be on the paratope, for the same 3 antigens mentioned above, a negative regression coefficient or one that is around zero indicates an amino acid that might be present on the paratope but not always, while a positive coefficient indicates a presence on the paratope with probability very close to 1. It suggests that all positive contributions to the binding come from amino acids of the paratope, while negative contributions come from amino acids on non-binding parts of the antibody chain as well.

3.2 Correlation between statistical interactions and binding properties

For interaction detection, we use the one-hot encoded 11-mer slices of binding and non-binding sequences associated with one antigen, together with their binding status, as input data. On the contrary to logistic regression, all amino acid columns for each position can be provided here, we do not need to set one letter as a reference. The method outputs a list of interacting pairs of amino acids, each coming with a value of statistical interaction strength. For ease of visualisation, we take the log in base 10 of this value, and we use box plots. Results are presented in Figure 2 for antigen 1ADQ_A, and in the Supplementary Figures S2–S5 for the other antigens. The detection takes about 2 min per dataset.

Again, the binding degree is an important parameter. Those pairs with the strongest statistical interaction effects correspond to amino acids whose sum of binding degrees is higher than that of most other pairs.

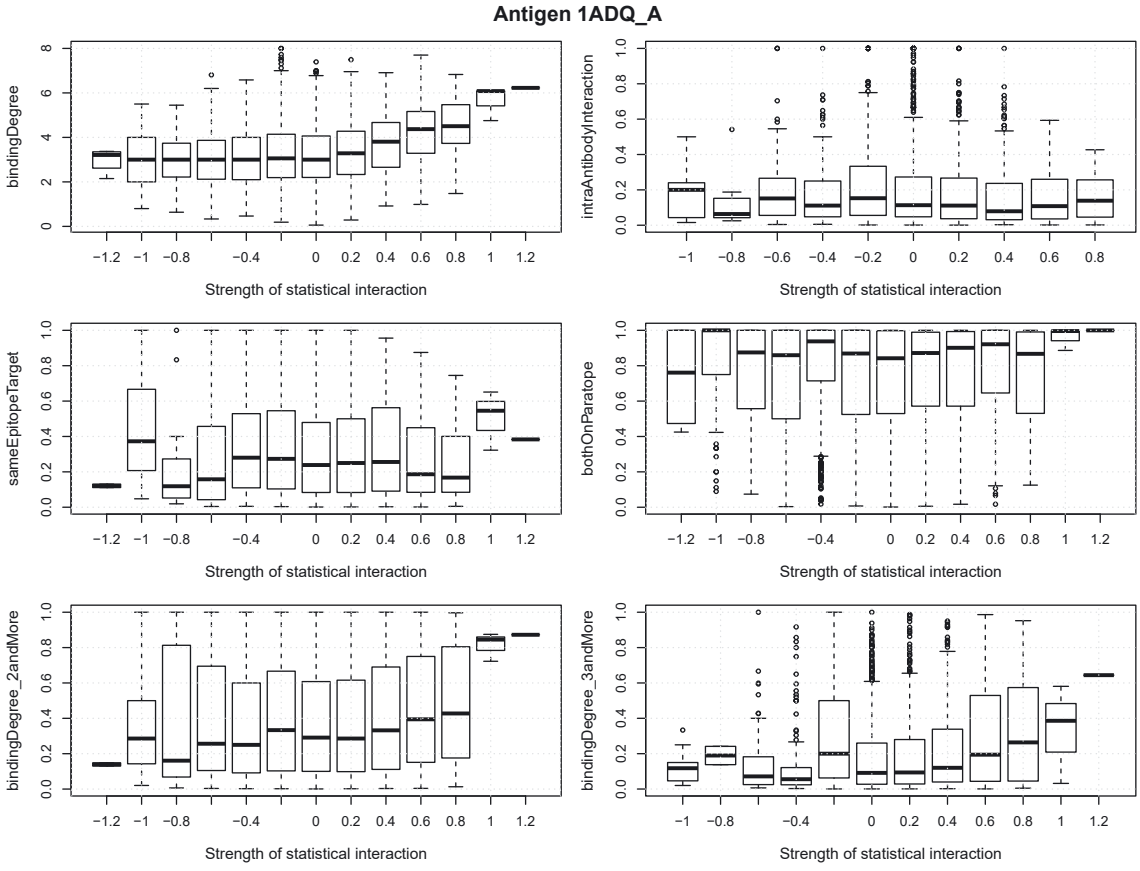

Figure 2: Structural information (clockwise from top left: average sum of the binding degrees, proximity on the antibody, presence on the paratope, binding degree 3 or more, binding degree 2 or more, same epitope target) against the log in base 10 of the strength of statistical interactions, for amino acid pairs. Antigen 1ADQ_A. We use box plots for visualisation.

For antigen 1ADQ_A, most pairs have a sum of binding degrees around 3, but those pairs with the strongest value of statistical interaction have a sum around 6. That phenomenon is more or less noticeable depending on the antigen. Further, for all antigens, the highest statistical interaction strength is associated with amino acids that are both on the paratope with probability 1, while weaker interactions are associated with amino acids on both binding and non-binding parts of the antibody chain. This agrees with what was already observed for the main effects contributing positively and negatively to the binding. Noticeably, the strongest interacting amino acids correspond to pairs that are rarely in direct contact on the chain due to folding for antigen 1ADQ_A, never for antigens 1FBI_X, 1FNS_A, 1FSK_A, and with probability mostly below 0.4 for antigen 1H0D_C. This means that the statistical interaction that favours binding most does not correspond to an easily observable phenomenon on the antibody chain alone, but rather to something that is happening on the whole antibody-antigen structure.

To be sure that patterns that we find do not correspond to motifs that are extremely rare, we also plot the same figure restricted to amino acid pairs that appear in at least 1 % of the sequences, and it lead to the same results (see Supplementary Figure S6 for antigen 1ADQ_A). Further, we compute the structural information against the positional dependency within each pair, $(\Delta_{i,j} + \Delta_{j,i})/2$ as defined in the Methods sections, instead of the statistical interaction strength. This is shown in the Supplementary Figure S7. The scale unit of the x-axis now goes from -1 to $+1$, while the y-axis remains the same. As positional dependencies are easier to compute and simpler to understand than statistical interactions, yet a close concept, the idea here is to confirm the observed results in a simpler setting. The trends for the first antigen are slightly less pronounced compared to the statistical interactions, but the general aspect does not contradict the earlier conclusions.

3.3 Motifs revealed by statistical methods improve identification of binding sequences

As another interpretation of the insight brought by the statistical methods, we compute for a given antigen the conditional probability that a sequence is a binder given that it has a particular motif M, when this motif comes from either the coefficients with highest values in a logistic regression model with main effects, or from the top statistical interactions detected by the NID method, on this antigen. We further divide this by the total proportion of binders to the antigen, to quantify independence of the two events, binding and having the motif:

$$\frac{\mathbb{P}(\text{X is binder}|\text{X has motif M})}{\mathbb{P}(\text{X is binder})},$$

where these probabilities are computed empirically. A value of 1 here indicates independence. We do this when the motif is a pair of amino acids, a 3-mer and a 4-mer. For example, for antigen 1ADQ_A, the top regression coefficients are “2L”, “2F”, “2I”, “3F”, “11F”, “11L”, “2V”, “11I”, “6L”, ordered by decreasing value. We create motifs by associating the highest possible main effect with the next highest that is on a different position, so the motifs created are “2L_3F”, “2L_3F_11F” and “2L_3F_11F_6L”. We proceed similarly for statistical interactions, the top ones detected are “2L_4L”, “2L_4G”, “2L_6L”, “4L_6L”, “4G_6L”, “2L_10L” and the corresponding motifs are “2L_4L”, “2L_4L_6L” and “2L_4L_6L_10L”. Working with motif “2L_3F” means that we focus on all sequences that have L at position 2 and F at position 3, the rest of the sequences does not matter, and we verify whether these sequences bind more often than a randomly taken sequence. In Figure 3, we present summary statistics on the ratios, computed on all 159 antigens. Motifs are extracted from models fitted on the train set (50 % of the sequences), and the probabilities are computed empirically on the remaining sequences.

For a 2-mer motif, most ratio values lie between 0.5 and 3 for logistic regression, and 1.5 and 3 for NID. For the latter method it means that the main discovered motif increases the probability of binding by a significant amount, hence its relevance, but for the former method there are some antigens where the motif does not bring better insight. The upper range of numerical values increases further when we select a longer motif: for a 3-mer, values reach as high as 4.3 for logistic regression and 4 for NID, and for a 4-mer they reach as high as 5. Again, most values for the NID method are above 1.5 with a few values below 1, while there is a significant amount below 1 for logistic regression, meaning that sequences with the chosen motif bind less often than in average.

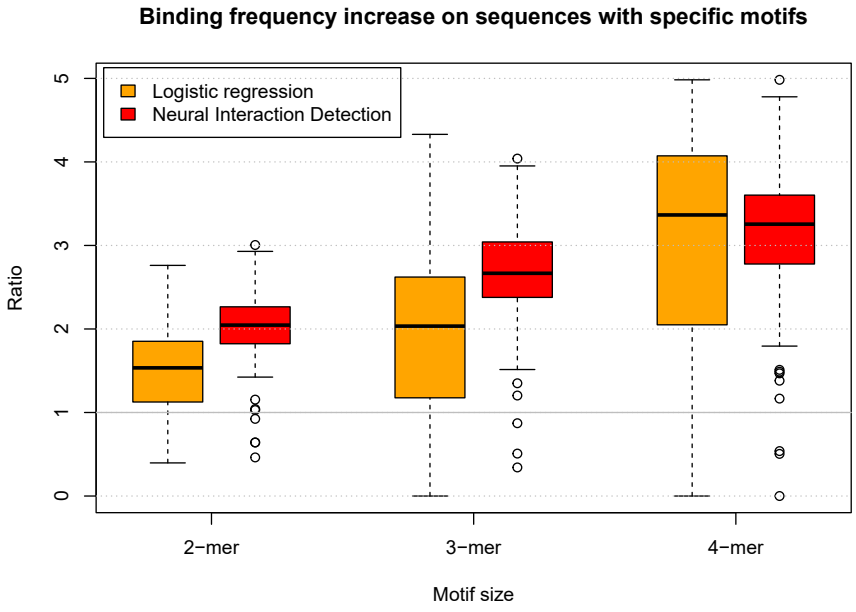

Figure 3: Proportion of sequences with a specific motif that bind, divided by the total proportion of binders for a given antigen. Values for all 159 antigens are gathered into box plots. The motif is taken from either the logistic regression coefficients with the highest values, or the top statistical interactions detected by the NID method. Values above 1 indicate that sequences with the motif bind more often than the rest.

Mean values for this plot are 1.5, 1.9, 3.0 for logistic regression, 2.0, 2.6, 3.1 for NID, for a 2-mer, 3-mer, 4-mer, respectively. In this setting, the frequency of binding for a random sequence is always around 0.2, so we can reach a theoretical maximum binding frequency increase of 5.

A two sample *t*-test that the increase with logistic regression is not less than the increase with NID, gives *p*-values less than 2.2e-16 for a 2-mer motif, 2.6e-12 for a 3-mer and 9.7e-2 for a 4-mer. This indicates that the motif found by NID brings a larger frequency increase and is more relevant, even though there is slightly less certainty for the last case. Thus, statistical interactions are relevant in determining which sequences may bind to the antigen.

3.4 Asserting which statistical features are shared for different antigens

We plot in Figure 4 the statistical information computed on one antigen, against the statistical information computed on another antigen, for all pairs of antigens among the 5. This is in order to see whether patterns that give a high specificity to one antigen also give high specificity to other antigens. We proceed for both the main effects and the statistical interactions.

Notably, we notice that for all pairs of antigens, most individual amino acids have a very similar effect in terms of the value and sign of the estimated logistic regression coefficient. For example, for antigens 1FNS_A and 1FSK_A the estimated coefficients have a Pearson correlation of 0.86, and for antigens 1FBI_X and 1FNS_A the correlation is 0.84. Other pairs of antigens have lower correlations, with a minimum at 0.71, but for most amino acids, high positive coefficients in one antigen correspond to high positive coefficients in another antigen, and the same yields for low negative coefficients. This means that some amino acids are sticky independently of the antigen, while others are more reluctant to bind.

Regarding statistical interactions, pairs of amino acids seem to affect different antigens in different ways, meaning that motifs are more specific to a particular antigen. It is more difficult to see a trend here, as most values of statistical interaction strengths are close to 0. Going to log scale here did not lead to greater insight (maximum observed Pearson correlation of 0.25, not shown). The pairs of interest are the ones with the highest

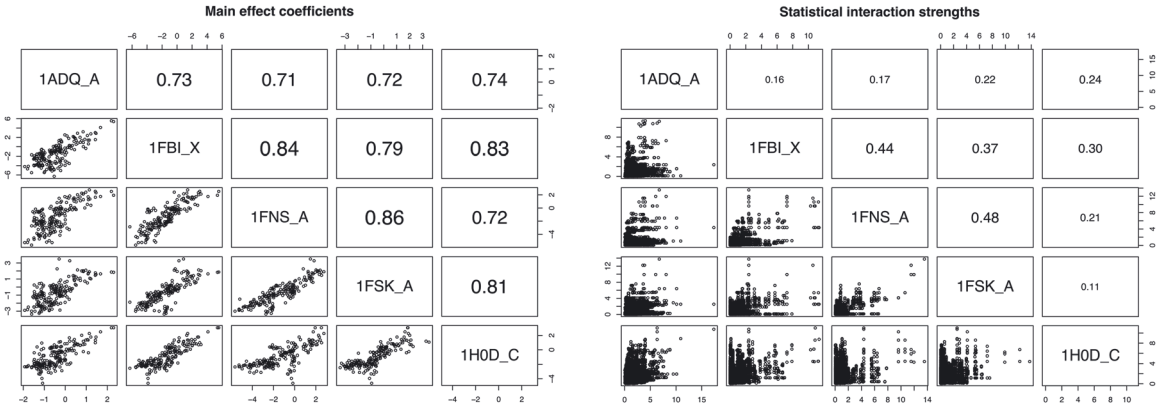

Figure 4: Pair plots of main effect coefficients (left) and statistical interaction strengths (right), estimated on one antigen against the same quantities estimated on another antigen, by amino acids and positions, with the corresponding Pearson correlations in the upper parts.

statistical interaction strength, but they are different in most antigens, except for 1FNS_A and 1FSK_A that share a small group of 4 pairs with the highest statistical interactions, resulting in a Pearson correlation of 0.48. These 2 antigens had also very similar regression coefficients for the main effects.

More results for a different group of 5 antigens, namely 2R0K_A, 3MJ9_A, 4IJ3_A, 4ZFO_F and 5KN5_C, are present in the Supplementary Figure S8. We study these antigens to verify that previous results obtained on the first five antigens are not a mere consequence of similar molecular conformations. For these different antigens, cross-specificity of individual amino acid effects is also important and reaches correlation values as high as 0.91 for antigens 4IJ3_A and 4ZFO_F, although with a minimum at 0.50 for antigens 2R0K_A and 5KN5_C, below the lowest value of 0.71 for the first group. Cross-specificity of pairs of amino acids is very similar as well, with a maximum observed correlation of 0.45 for antigens 3MJ9_A and 4ZFO_F.

3.5 Main effects and statistical interactions contributions to stickiness and specificity

On a higher level of the binding process, one important point is whether an antibody sequence binds a target antigen because it is generally sticky to many antigens, or because it is specific to this one in particular. As mentioned in the Methods section, we work now with a dataset that contains all the binders of each antigen, and a selection of 500 000 sequences from outside its top 1 % affinity. As before, we remove duplicated slices and select half of them as a training set. This constitutes for each antigen a dataset of around 220,000 sequences, of which about 10 % are binders. The non-binders are still defined as the top 1 %–5 % affinity sequences. We compute the stickiness of binder and non-binder sequences related to ten different antigens in Figure 5. A highlight on stickiness values above 0.95 for this figure is also present in the Supplementary Figure S9.

Results show that most binder antibodies correspond to sequences that already have high stickiness values, above 0.99. Concerning the non-binders, they span on stickiness values lower than the binders, with most values between 0.8 and 0.99. Outlier values of both binder and non-binder sequences are above 0.55, except for antigen 1ADQ_A with a minimum at 0.43, and do not bring more insight on the difference between the two sets. For further intuition of what it means to be sticky, we also display in the Supplementary Figure S10 the number of antigens an antibody binds to, that increases as a function of its stickiness value. Some antibodies with a stickiness between 0.99 and 1 can bind as many as 40 different antigens. Thus, it appears a relevant measure of the binding power of a particular sequence.

For specificity, on the 80 antigens selected for their energy profile, we find values that are at most 0.734, for antigen 4RGM_S, and below 0.5 for most other antigens (see Supplementary Figure S11 for extrema and quantiles

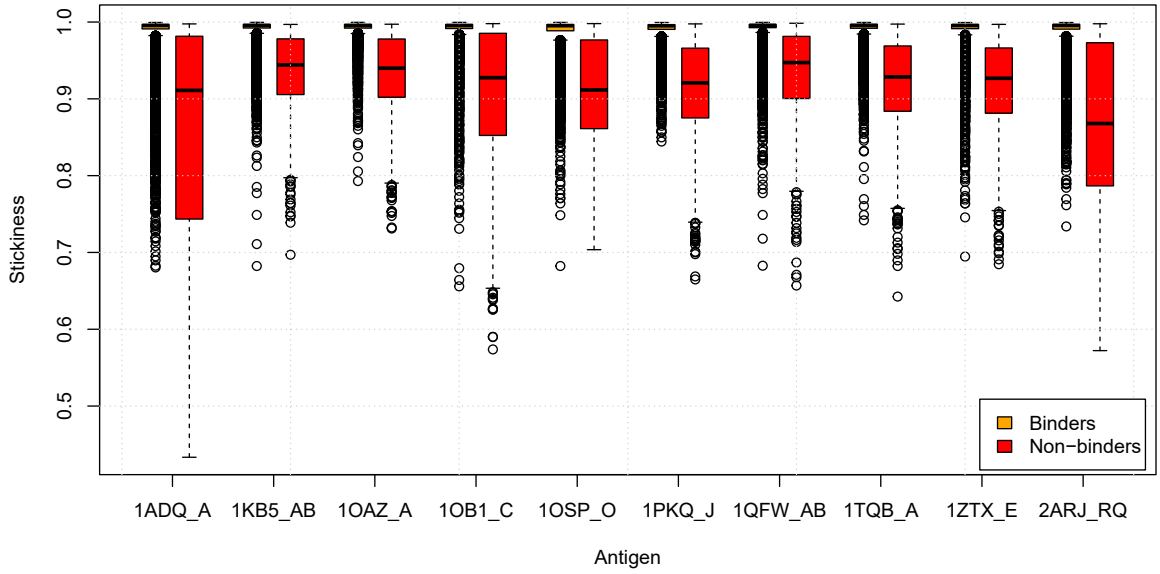

Figure 5: Stickiness of binder and non-binder antibodies related to different antigens.

of specificity values for the different antigens). As the affinity is the sum of specificity and stickiness, antibodies binding one antigen (affinity 0.99 or more) already have stickiness values above 0.25, which restricts the set of potential binders, and confirms the intuitions on stickiness. Concerning the negative range however, values reach as low as −0.999 for antigen 4KXZ_A, meaning that some sequences are very badly suited to this antigen even if they have a high stickiness on average. Thus, binding several antigens is not a sufficient condition to bind one specific antigen, but it is necessary to have a minimum stickiness in order to be a binder, and sequences that bind are in most cases the ones with the highest stickiness.

We then study which of the main effects or the statistical interactions are responsible for the specificity and stickiness of the sequences. For this, we fit a penalised logistic regression model for binding prediction to half of the top 5 % affinity sequences (binders and non-binders) of antigen 1ADQ_A, with all main effects and the top 2000 interactions found by the NID method on this antigen as the predictors. After removing duplicated slides, the data set contained 105 133 sequences of which 23 % were binders. The reason why we did not take all 22,000 possible interactions is that R's memory cannot handle more than a few thousands interactions with our data dimensions. We use a ridge penalisation with penalty value 0.01 (found with 10-fold cross-validation) to fit the model with interactions. Once this model is fitted, we compute the log-odds of main effects and statistical interactions on sequences related to the same antigen 1ADQ_A for which we have computed values of affinity, stickiness and specificity (all affinity values possible). Here, the train and test sets are not on the same domain, and this extrapolation brings some uncertainties, but the results were less noisy than when using the same domain for both training and testing. We do not include the intercept in any of the log-odds. Results are presented in Figure 6.

The contribution of statistical interactions spans a larger range than that of main effects (−10 to +10 against −3 to +3), maybe because there are many more of the former. Interestingly, there is no clear trend between affinity or stickiness and the log-odds of main effects or statistical interactions for the top affinity sequences. However for sequences with lower affinity, both log-odds increase as a function of affinity and stickiness, even though the logistic regression model was not trained on these sequences. It is especially true between the log-odds of main effects and stickiness, for which the Pearson correlation was 0.82. Hence, the patterns learned can be related to high-level properties of the data. Specificity could not be related to either of the log-odds. We also display the contributions with a different scaling for the affinity in Supplementary Figure S12. This scaling highlights the binders, and numerical values for stickiness and specificity are modified. There, we

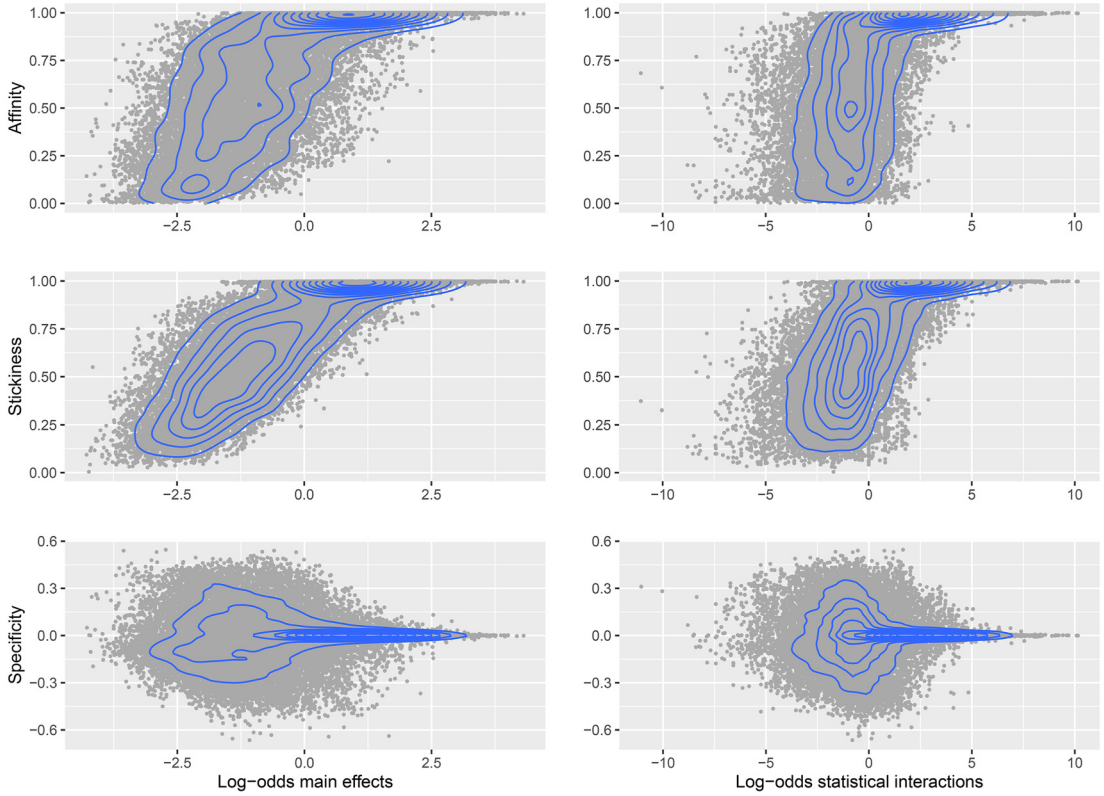

Figure 6: Specificity, stickiness and affinity, plotted against contributions of main effects and statistical interactions in a logistic regression model for binding prediction. Contour lines for the density are represented in blue. Computed on antigen 1ADQ_A.

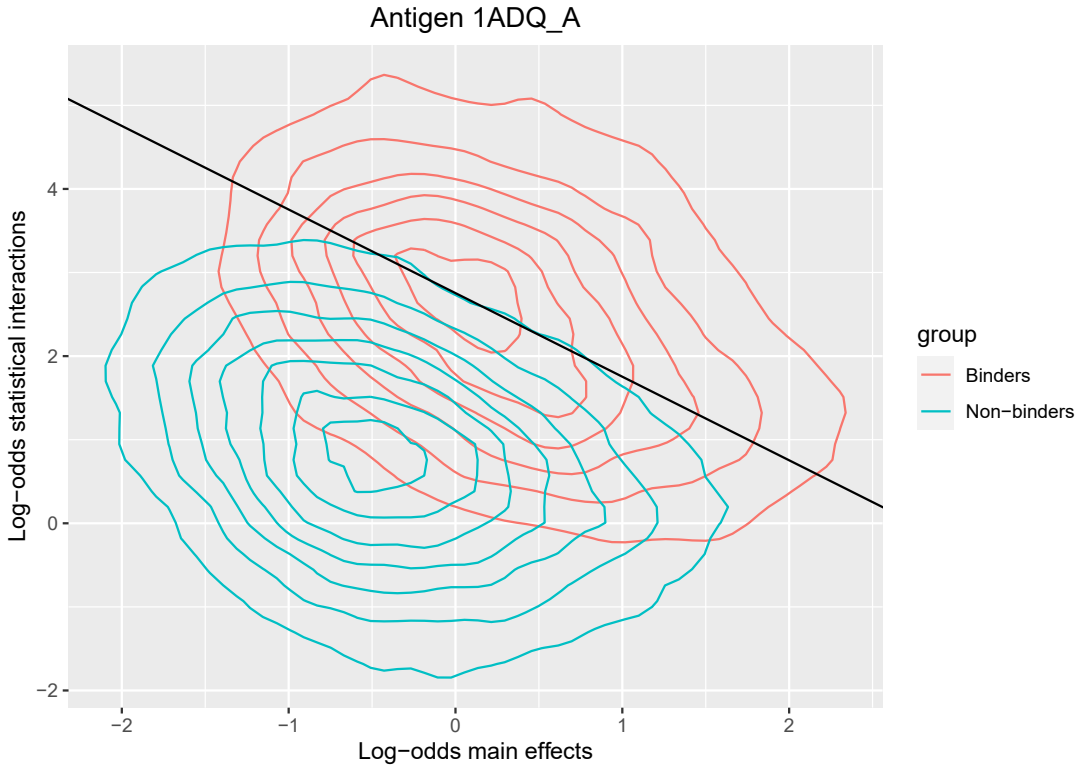

Figure 7: Contour plot of the log-odds of main effects and statistical interactions in sequences, one point is a sequence. The predicted binding probability exceeds 0.5 above the black line.

observe the same trends as in Figure 6, indicating that results are consistent across different ways of defining the affinity.

We also plot the contributions of main effects and statistical interactions among the binders and non-binders antibody sequences in Figure 7, still without the intercept term. There, the difference between binders and non-binders seems more driven by interactions than by main effects. Indeed, the center of the density moves from $(-0.5, 0.5)$ to $(0, 2.5)$ between non-binders and binders, where the first coordinate indicates log-odds of main effects, and the second one log-odds of interactions. One limitation of these results is that introducing statistical interactions might change the main effects in the regression model, making their contribution decrease, which can further vary if the model contains more or less interactions.

4 Discussion

In this manuscript, we studied data from a simulation where amino acids are placed on a 3D lattice, and a potential energy function is used to compute the energy of antibody-antigen structures, the structures with the lowest energy being considered as binding structures. Amino acid antibody sequences of length 11 were encoded into 220 binary covariates using one-hot encoding, and their binding to several different antigens was explored. We focused on the impact of both main effects (individual covariates) and statistical interactions (pairs of covariates) on the binding. The strength of main effects was obtained from coefficients of a logistic regression model fitted on binding and non-binding sequences to a particular antigen, while the strength of statistical interactions was computed from a neural network trained on the same data. These statistical insights were related to various levels of information on the binding process.

In a first step, we focused on the structural information, at an intermediate level of the binding process. In particular, the binding degree of an amino acid of the antibody, defined as the number of amino acids from the antigen it is bonded to, could be related to numerical values of statistical measures. Amino acids with a positive regression coefficient had a higher average binding degree than most other amino acids, usually 2 to 4 bonds, while the binding degree of those with a negative coefficient was below a certain threshold of 2 or 3 bonds. Pairs with the highest statistical interaction were also the ones whose constituting amino acids had a higher binding degree compared to the rest of the sequence. Thus, information on the binding degree may help to improve the accuracy of models for predicting binding. This was done by Robert et al. for the same simulated data. For real data however, such a concept is not as straightforward to define, and positional information is rarely accessible.

We also found it useful to have knowledge of the paratope composition, as results showed that amino acids with a high positive contribution to binding are very often found there, either those with high positive regression coefficients, or those forming statistical interactions with the highest strengths. Such results were in accordance with e.g. those of Ostmeyer et al. who state that the largest contribution to binding should come from amino acids in direct contact with the epitope. Yet, we did not know whether non-binding amino acids, outside the paratope area, were important or not. It might be that a low stickiness amino acid helps the two neighbor amino acids to be binders for instance. Other structural information, such as whether amino acids had the same epitope target, or where neighbors on the chain, could not straightforwardly be related to these effects.

Further, sequences that contained a motif made of top main effects or top interacting pairs typically had a higher probability of binding than a randomly sampled sequence. For the motifs found by the Neural Interaction Detection method, that increase was by a factor around 2 for a 2-mer motif, and around 3 for a 4-mer motif, for most antigens. Values based on logistic regression coefficients were slightly lower. They varied by antigen and in a few of the 159 antigens we even observed a decrease in the probability of binding when selecting by motif. Notably, the NID method had a running time in the same range as logistic regression with main effects, and both are faster than logistic regression with k-mers, that is used for example by Glanville et al. Thus, these two methods constitute interesting alternatives to k-mer based methods for discovering patterns that are important for binding.

Cross-antigen analysis revealed that amino acids responsible for binding to one antigen could also explain binding to another antigen, suggesting that amino acids could be more or less inherently sticky. It was not

possible to find a similar pattern for pairs of amino acids, except for a few, suggesting that motifs were more specific to a particular antigen.

In a second step, we focused on information more directly related to the binding capacity of antibody sequences, namely their stickiness and specificity. We found that most binding antibody sequences had a high value of stickiness, and were binders to several different antigens at the same time. This suggests that binding was not specific to certain antigens.

Finally, based on the log-odds of a logistic regression model for binding prediction, we found that the contribution of main effects increased when stickiness was also increasing, opening up possibilities to explain the stickiness of a given sequence in more details by its amino acid composition.

This 3D model of antibody-antigen structures is a major simplification of the true underlying reality, that greatly reduces the diversity of interactions between amino acids of both chains, the possible forms that they might take, as well as the environment and the rest of the antibody chain. Hence, conclusions drawn from this analysis cannot be extrapolated directly to infer the ground truth in real data, but they can serve as a basis for initial insights, and help in the understanding of the different steps of the binding process.

In this study, we have shown that a number of emergent properties common to different antigens could be linked to the role of individual amino acids in a logistic regression model (main effects). These properties concerned the binding degree (number of bond amino acids on the epitope), the presence or not on the paratope, as well as the stickiness of a sequence defined as its capacity to bind any antigen in general. Further, statistical interactions, that we expected would also play a role in the binding process, were shown to be important for binding prediction and contain some information related to the specificity of sequences to particular antigens. However, there was less certainty about the relation between statistical interactions and structural properties. Altogether, this showed the feasibility of using statistical methods to decipher some of the structural properties of antibody-antigen binding.

Acknowledgment: We thank Milena Pavlović for useful discussion.
Research ethics: Not applicable.
Author contributions: The authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Competing interests: The authors state no conflict of interest.
Research funding: This research was funded by the UiO:LifeScience Convergence Environment Immunolingo.
http://dx.doi.org/10.13039/501100009566.
Data availability: Not applicable.

References

Chavali, A.K., Gianchandani, E.P., Tung, K.S., Lawrence, M.B., Peirce, S.M., and Papin, J.A. (2008). Characterizing emergent properties of immunological systems with multi-cellular rule-based computational modeling. *Trends Immunol.* 29: 589—599.

Ferdous, S. and Martin, A.C.R. (2018). AbDb: antibody structure database — a database of PDB-derived antibody structures. *Database* 2018: 9.

Glanville, J., Huang, H., Nau, A., Hatton, O., Wagar, L.E., Rubelt, F., Ji, X., Han, A., Krams, S.M., Pettus, C., et al. (2017). Identifying specificity groups in the T cell receptor repertoire. *Nature* 547: 94—98.

Greiff, V., Menzel, U., Miho, E., Weber, C., Riedel, R., Cook, S., Valai, A., Lopes, T., Radbruch, A., Winkler, T.H., et al. (2017). Systems analysis reveals high genetic and antigen-driven predetermination of antibody repertoires throughout B cell development. *Cell Rep.* 19: 1467—1478.

Greiff, V., Yaari, G., and Cowell, L.G. (2020). Mining adaptive immune receptor repertoires for biological and clinical information using machine learning. *Curr. Opin. Syst. Biol.* 24: 109—119.

Kanduri, C., Pavlović, M., Scheffer, L., Motwani, K., Chernigovskaya, M., Greiff, V., and Sandve, G.K. (2022). Profiling the baseline performance and limits of machine learning models for adaptive immune receptor repertoire classification. *GigaScience* 11: giac046.

Miyazawa, S. and Jernigan, R.L. (1999). An empirical energy potential with a reference state for protein fold and sequence recognition. *Proteins* 36: 357—369. .

Ostmeyer, J., Christley, S., Toby, I.T., and Cowell, L.G. (2019). Biophysicochemical motifs in T-cell receptor sequences distinguish repertoires from tumor-infiltrating lymphocyte and adjacent healthy tissue. *Cancer Res.* 79: 1671—1680.

Parrish, J.K., Viscido, S.V., and Grünbaum, D. (2002). Self-organized fish schools: an examination of emergent properties. *Biol. Bull.* 202: 296—305.

Pavlović, M., Scheffer, L., Motwani, K., Kanduri, C., Kompova, R., Vazov, N., Waagan, K., Bernal, F.L.M., Costa, A.A., Corrie, B., et al. (2021). The immuneML ecosystem for machine learning analysis of adaptive immune receptor repertoires. *Nat. Mach. Intell.* 3: 936—944.

Raymond, D.J. and Fuchs-Stone, Z. (2021). Emergent properties of convection in OTREC and PREDICT. J. Geophys. Res. Atmos. 126: 1—19.

Robert, P.A., Akbar, R., and Greiff, V. (2021). Absolut! in silico antibody — antigen binding database. *Nird Res. Data Arch.*.

Robert, P.A., Akbar, R., Frank, R., Pavlović, M., Widrich, M., Snapkov, I., Slabodkin, A., Chernigovskaya, M., Scheffer, L., Smorodina, E., et al. (2022). Unconstrained generation of synthetic antibody—antigen structures to guide machine learning methodology for antibody specificity prediction. *Nat. Comput. Sci.* 2: 845—865.

Sorokina, D., Caruana, R., Riedewald, M., and Fink, D. (2008). Detecting statistical interactions with additive groves of trees. In: *Proceedings of the 25th international conference on machine learning — ICML '08*'. ACM Press, Helsinki, Finland.

Tsang, M., Cheng, D., and Liu, Y. (2018). Detecting statistical interactions from neural network weights. In: *International conference on learning representations 2018*, arXiv:1705.04977.

van den Berg, N.I., Machado, D., Santos, S., Rocha, I., Chacón, J., Harcombe, W., Mitri, S., and Patil, K.R. (2022). Ecological modelling approaches for predicting emergent properties in microbial communities. *Nat. Ecol. Evol.* 6: 855—865.

Viscido, S., Parrish, J., and Grünbaum, D. (2004). Individual behavior and emergent properties of fish schools: a comparison of observation and theory. *Mar. Ecol. Prog. Ser.* 273: 239—249.

Yin, R., Feng, B.Y., Varshney, A., and Pierce, B.G. (2022). Benchmarking alphafold for protein complex modeling reveals accuracy determinants. *Protein Sci.* 31: e4379.

Supplementary Material: This article contains supplementary material (https://doi.org/10.1515/sagmb-2023-0027).