

Novel structural parameters of Ig - Ag complexes yield a quantitative description of interaction specificity and binding affinity

Simon Marillet and Marie-Paule Lefranc and Pierre Boudinot and Frédéric Cazals

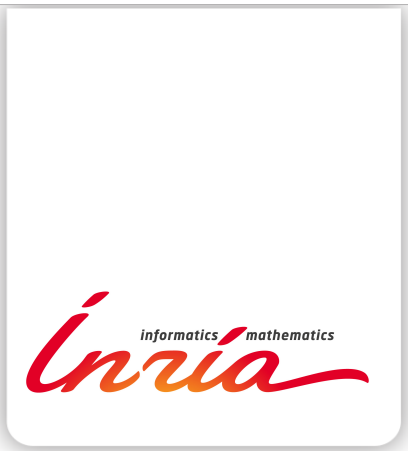
RESEARCH REPORT

N° 8963

October 2016

Project-Team Algorithms-Biology-Structure

ISSN 0249-6399 ISRN INRIA/RR--8963--FR+ENG



Novel structural parameters of Ig - Ag complexes yield a quantitative description of interaction specificity and binding affinity

Simon Marillet* and Marie-Paule Lefranc† and Pierre Boudinot‡ and Frédéric Cazals §

Project-Team Algorithms-Biology-Structure

Research Report n° 8963 — October 2016 — 40 pages

Abstract: Antibody-antigen complexes challenge our understanding, as analyses to date failed to unveil the key determinants of binding affinity and interaction specificity. We partially fill this gap based on novel quantitative analyses using two standardized databases, the IMGT/3Dstructure-DB and the structure affinity benchmark.
First, we introduce a statistical analysis of interfaces which enables the classification of ligand types (protein, peptide, chemical; cross-validated classification error of 9.6%), and yield binding affinity predictions of unprecedented accuracy (median absolute error of 0.878 kcal/mol). Second, we exploit the contributions made by CDRs in terms of position at the interface and atomic packing properties to show that in general, VH CDR3 and VL CDR3 make dominant contributions to the binding affinity, a fact also shown to be consistent with the enthalpy - entropy compensation associated with pre-configuration of CDR3.
Our work suggests that the affinity prediction problem could be solved from databases of high resolution crystal structures of complexes with known affinity.

Key-words: Antibody - antigen complex; interface ; affinity

* Inria and INRA (Unité de Virologie et Immunologie Moléculaires), France
† IMGT, IGH, CNRS, France
‡ INRA, Unité de Virologie et Immunologie Moléculaires, Jouy-en-Josas, France
§ Inria, France

RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

De nouveaux paramètres pour une description quantitative de la spécificité d'interaction de l'affinité de liaison de complexes Ig - Ag

Résumé : Les complexes anticorps - antigène défient notre compréhension, les analyses publiées à ce jour n'ayant pas pu déterminer les éléments clés expliquant l'affinité de liaison et la spécificité d'interaction. Nous comblons partiellement cette lacune en nous basant sur des nouvelles analyses quantitatives utilisant deux bases de données standardisées, IMGT/3Dstructure-DB et le *structure affinity benchmark*.

Premièrement, nous présentons des statistiques sur les interfaces permettant de distinguer les types de ligands (protéine, peptide, composé chimique; erreur de classification cross-validée de 9.6%), et d'obtenir des prédictions d'affinité de liaison d'une précision sans précédent (erreur médiane absolue de 0.878 kcal/mol). Deuxièmement, nous exploitons les contribution des CDR en termes de position a l'interface et de propriétés de packing, et montrons qu'en général, les contributions de VH CDR3 et VL CDR3 a l'affinité de liaison sont prédominantes. Ce résultat est cohérent avec la compensation entre enthalpie et entropie associée a la pré-configuration du CDR3.

Notre travail suggère que le problème de prédiction d'affinité pourrait être résolu par l'utilisation de bases de données de structures cristallographiques à haute résolution dont l'affinité serait connue.

Mots-clés : Complexe anticorps - antigène ; interface ; affinité; positions relatives des CDR

1 Introduction

Immunoglobulins and the immune response. Adaptive immunity is based on antigen (Ag)-specific lymphocyte responses. Upon specific recognition of an antigenic epitope by a given receptor unique to a lymphocyte, this cell gets activated and proliferates, leading to a clonal expansion. B lymphocytes thus recognize antigens through membrane-bound immunoglobulins (Ig) expressed at their surface. Seric Igs can opsonize bacteria and facilitate their uptake by phagocytes, or neutralize viruses thus preventing recognition by their receptor or fusion with the target cell. Immunoglobulins fundamentally consist of two identical heavy (H) chains and two identical light (L) chains, each H chain being bound to an L chain. The antigen-binding site is located at the top of the paired VH and VL, and generally overlaps the two V domains. It mainly consists of three flexible loops on each V domain, called complementarity determining regions (CDR1-3). The diversity of antibodies is concentrated in the CDRs.

From the structural standpoint, the functional relevance of an Ig depends on its binding affinity for the targeted antigen and the specificity of the interactions, which provides the basis of immune memory and vaccination. The affinity sets the strength of the interaction. For the membrane-bound Ig, it determines if enough aggregation of surface Igs and Ig co-receptors occurs, so that a sufficient signal can be sent to the cell to induce activation and proliferation [4]. For secreted Ig, once bound to the target, pathogens or host infected / tumoral cells, the affinity sets the efficiency of Ig-mediated pathogen opsonisation and/or neutralization, or Ig effector properties (antibody-dependent cell-cytotoxicity or ADCC, complement-dependent cytotoxicity or CDC) [36].

Ig - Ag complexes and underlying genetic mechanisms. The prominent role played in Ag binding by CDRs has prompted the analysis of CDR-specific statistics. Using a handful of crystallographic structures, canonical conformations *i.e.* commonly occurring backbone CDR conformations were first reported [14] and subsequently updated [15, 1], using 300 non-redundant Ig structures in the latest work [51]. Moving from individual CDR to all CDRs, correlations between canonical conformations were further studied [64], highlighting the fact that some combinations are multi-specific, while others are specific of an antigen type. The VH CDR3 is the most variable and was therefore the focus of several studies [58, 49, 59] which defined and updated sequence-based rules to predict its conformations. More recently, these studies have been refined, based on a larger number of structures (of the order of hundreds instead of tens). For VL CDR3, new canonical conformations were proposed [33], and for VH CDR3, previous rules were updated and complemented [31, 32]. Distinguishing lambda versus kappa chains, it has been shown that canonical conformations from the former are more diverse than those from the latter in the human and the mouse [10]. However, the relevance of canonical conformations for the prediction of the 3D structures of CDRs was questioned [13], since general loop prediction methods matched (or even outperformed) the prediction performances of methods exploiting specific rules associated with canonical conformations of CDRs. In parallel, two related works [2, 55] studied the differential CDR lengths and Specificity-Determining Residues Usage (SDRU, proportion of Ig amino-acids at a given CDR position which contact the antigen) between ligand types. However, these analyses do not allow antigen type predictions. To assess the role of individual CDRs, it has also been established that except in the case of bacterial carbohydrates, diversity in VH CDR3 alone can result in primary responses specific to the antigen [65]. Structural and genetic aspects of individual CDRs in natural and artificial antibody repertoires are reviewed in [21].

Ig - Ag complexes and thermodynamics. The analysis of Ig - Ag complexes can also be posed from the thermodynamics standpoint. Specifically, the binding affinity is a thermodynamic quantity describing the chemical equilibrium associated with the two partners (Ig and Ag in our case) and the complex (Ig - Ag, denoted IG/Ag in the IMGT nomenclature [20]). It is generally measured by the dissociation constant K_d (= [Ig] $\cdot$ [Ag]/[Ig - Ag]) of this equilibrium. Equivalently, it is expressed by the corresponding dissociation free energy $\Delta G_d = -RT \ln K_d/c^\circ = \Delta H - T\Delta S$, in the $c^\circ = 1M$ standard state, with T the temperature and R the gas constant. Thus, by nature, the affinity has an enthalpic component (ΔH) qualifying the interaction energy, but also an entropic component ($T\Delta S$) qualifying the loss of dynamical properties upon complex formation (intuitively, the formation of the Ig - Ag complex indeed restricts the degrees of freedom of both partners). These two competing interests illustrate the enthalpy - entropy compensation phenomenon [48, 19], which stipulates that a favorable enthalpic change upon association is accompanied by an entropic penalty. Predicting binding affinities from structural data requires to quantify this compensation, and is therefore a notoriously challenging problem, for protein complexes in general [29, 46], and for Ig - Ag complexes in particular [40].

To model the enthalpic component, various parameters have been proposed. Most of these parameters, which describe the morphology of the interface (size, shape, packing properties) and its biochemistry (salt bridges, solvation, hydrogen bonds), were estimated from crystal structures of complexes [28, 41, 7, 44]. More recently, it has also been shown that non-interacting atoms play an important role, intuitively related to solvent interactions [30]. Such approaches have been applied to Ig - Ag complexes [61], stressing in particular the role of interfacial solvent [5], biochemical properties of Igs as a function of epitopes [50], or the correlation between interface curvature and ligand size [43].

To model the entropic component, the conformational and vibrational properties of the partners must be captured. It has indeed been shown that the preconfiguration of the binding site may yield a decreased entropic loss, hence an enhanced binding affinity [45, 56, 12, 57]. It has also been shown that a preconfiguration of the variable domains can be induced by the constant domain 1 (CH1) of the heavy chain [53, 62], suggesting that the isotype switching commonly occurring during B cell differentiation may affect the affinity through changes in the dynamic properties of the Ig. Parallel to binding affinity, the notion of functional affinity or avidity which takes into account the (possibly negative) cooperativity between monomers of an antibody is highly relevant *in-vivo*. In that context, constant regions have been shown to influence the avidity [17, 18, 47, 54]. Likewise, an intact ball-and-socket joint between VH and CH1 domains has been shown to affect antibody neutralizing activity [34].

Contributions. The difficulty of understanding molecular recognition between proteins in general and antibody - antigens in particular is well known [61]. In this work, we present novel quantitative analyses for interfaces of Ig - Ag complexes. Using the annotated IMGT/3Dstructure-DB [20], the interface between the Ig chains and the Ag is determined using a Voronoi based model for each complex, and decomposed into contributions from CDR, framework (FR) and atoms outside the V-region. This interface allows dissecting the interface into contributions made by CDRs, in terms of position of their atoms at the interface, and of packing properties of these atoms. Using these parameters, we show how to unambiguously distinguish ligand types and predict binding affinity with unprecedented accuracy. We also develop quantitative models for the contribution of VH CDR3 to binding affinity and interaction specificity, bridging the gap between various observations (canonical backbone conformations, mutagenesis data, affinity measurements), and explaining the emergence of function from a combination of structural and dynamical properties.

2 Material and methods

2.1 Voronoi interface models

Given a macro-molecular complex, an interface model is a structural model of the atoms accounting for the interactions, ideally encompassing its enthalpic (*i.e.* interaction energy) and entropic (*i.e.*, dynamic) dimensions. In the sequel, we model complexes and their interfaces using solvent accessible models [3] and the associated Voronoi based interface model (Figure S3 and [42]).

Solvent accessible models and Voronoi interfaces. The *solvent accessible model* (SAM) of a set of atoms is a model where each atom is represented by a ball whose radius is the van der Waals radius expanded by the radius $r_w = 1.4$Å of a water probe accounting for a continuous solvation layer [23, 3]. A convenient construction to study SAM is the Voronoi (power) diagram defined by the atoms [23]. In particular, the Voronoi diagram induces a partition of the molecular volume, obtained by computing for each atom its *Voronoi restriction*, namely the intersection between its atomic ball and its Voronoi region. The volume of this restriction, also called atomic volume, is a direct measure of the atomic packing [23].

The *exposed surface* of a SAM consists of the boundary of the union of balls defining the SAM. This surface consists of spherical polygons, delimited by circle arcs (every such arc is located on the intersection circle of two atoms), themselves delimited by points (each such point is found at the intersection of three atoms). When two molecules assemble to form a complex, the *buried surface area* (BSA) is the portion of the exposed surface of both partners which gets buried [41]. BSA has been shown to exhibit remarkable correlations with various biophysical quantities [27], and notably dissociation free energies for complexes involving moderate flexibility [44].

Consider the SAM of a complex whose partners are denoted A and B, and also involving interfacial water molecules W. Two atoms are in *contact* provided that their Voronoi restrictions are neighbors. Pairs of type (A,B) define the AB interface, namely direct contacts between the partners. Focusing on water molecules W sandwiched between the partners, pairs (A,W) and (B,W) correspond to water mediated interactions. It can be shown that all atoms from the partners identified this way form a superset of atoms loosing solvent accessibility [9]. The *binding patch* of a partner consists of its interface atoms. The atoms of the binding patch can be assigned an integer called its *shelling order*, which is a measure of the distance of this atom to the boundary of the patch it belongs to [7]. This information generalizes the core-rim model [41], and has been shown to provide state-of-the-art correlations with solvent dynamics, conservation of amino acids [7], and dissociation free energies [44]. All tools to compute the parameters just discussed are available within the Structural Bioinformatics Library at `http://sbl.inria.fr` > Applications > Space Filling Models.

Application to Ig - Ag complexes. For an Ig - Ag complex, we partition the set $\mathcal{I}$ of interface atoms just defined into the atoms $\mathcal{I}_{\mathrm{Ig}}$ contributed by the Ig, and the atoms $\mathcal{I}_{\mathrm{Ag}}$ contributed by the Ag, so that $\mathcal{I} = \mathcal{I}_{\mathrm{Ig}} \cup \mathcal{I}_{\mathrm{Ag}}$. It follows that the number of interface atoms $|\mathcal{I}|$ is the sum of those contributed by the Ig and the Ag respectively, namely $|\mathcal{I}| = |\mathcal{I}_{\mathrm{Ig}}| + |\mathcal{I}_{\mathrm{Ag}}|$. Similarly, we charge the Buried Surface Area (BSA) to the Ig and Ag respectively, so that $\mathrm{BSA} = \mathrm{BSA}_{\mathrm{Ig}} + \mathrm{BSA}_{\mathrm{Ag}}$. These quantities yield the average BSA per interface atom on Ig and Ag side:

$$\overline{\mathrm{bsa}}_{\mathrm{Ig}} = \frac{\mathrm{BSA}_{\mathrm{Ig}}}{|\mathcal{I}_{\mathrm{Ig}}|}, \tag{1}$$

$$\overline{\mathrm{bsa}}_{\mathrm{Ag}} = \frac{\mathrm{BSA}_{\mathrm{Ag}}}{|\mathcal{I}_{\mathrm{Ag}}|}. \tag{2}$$

The previous analysis can be generalized to accommodate the structure of Fabs, by decomposing the variable domains of each chain (VH and VL) into to three complementarity determining regions (CDRs) and four framework regions (FRs), resulting in 14 Voronoi interfaces. Practically, we focus on contacts made by the six CDRs, those made by framework regions being negligible (Table SV). (Details of the method used at `http://sbl.inria.fr/doc/Space_filling_model_interface-user-manual.html`.) In doing so, a buried surface area is defined for each CDR.

2.2 The dataset and data curation: the IMGT/3Dstructure-DB

We use the Ig - Ag complexes from the IMGT/3Dstructure-DB (`http://www.imgt.org/3Dstructure-DB/` [20]), corresponding to the category *IG/Ag* for *IMGT complex type*. Each such complex is processed in order to identify *canonical complexes* involving one heavy chain, one light chain, and one ligand. Upon inspecting such cases, two decisions are made. First, on the antigen side, we retain three types only (peptide, protein, chemical), due to the scarcity of cases involving other types. Moreover, we also remove complexes involving multiple ligands types. For the same reason, regarding species, complexes are assigned to three classes: human, mouse and other. In total, 489 complexes are retained after filtering for missing data, inconsistencies, redundancy, ligand type and speciess. The detailed processing methodology is described in the supplemental section 6.1. The main features of the complexes used are also summarized in the Supplemental file *dataset-IMGT-details.csv*.

CDR and FR limits of the VH and VL domains are according to the IMGT unique numbering [37] (Table SII). Practically, we use the following notations: CDR1-IMGT of VH is written VH CDR1 and FR3-IMGT of VL is written VL FR3. Other CDRs and FRs follow the same scheme.

2.3 The binding affinity benchmark

Our affinity predictions exploit the structure affinity benchmark (SAB) [29], a manually curated dataset containing 144 cases, each described by three crystal structures (of the unbound partners and of the complex) and the experimentally measured binding affinity in controlled conditions. In this work, we split the SAB into two sets: 14 Ig - Ag cases defining the test set (Table SIII), and 125 non-Ig - Ag cases defining the training set. Five complexes (among which 3 Ig - Ag) were removed from the SAB because only an upper bound on their K_d was provided, or had too many missing atoms. Having learned a statistical model from the latter, we predict affinities for Ig - Ag complexes of the former. See details in the Supplemental section.

2.4 Predicting ligand types

Antigens in the dataset are categorized as chemical, peptide and protein. Predicting the ligand type therefore requires to build a 3-class predictor.

Relevant variables. In order to predict ligand types , we represent each complex by two variables: $\overline{\text{bsa}}_{\text{Ig}}$ and $\overline{\text{bsa}}_{\text{Ag}}$ which are the average BSA per atom for atoms on the Ig and the Ag side respectively. These variables define the two-dimensional space displayed in Fig. 1 where each point represents a complex. A classifier *i.e.* a method predicting the antigen type from the parameters $\overline{\text{bsa}}_{\text{Ag}}$ and $\overline{\text{bsa}}_{\text{Ig}}$ is then trained on this data. Practically, we use a decision tree partitioning the space into rectangular regions, each corresponding to a ligand type.

Statistical methodology. Since the performance of classifiers tested on the training data is overestimated and leads to classifiers with poor generalization abilities (overfitting), various schemes have been devised to obtain an estimate of the generalization error.

We use the k-fold cross-validation where the dataset is randomly divided in k subsets of equal size, and $k-1$ subsets are alternatively used to classify the remaining one. At the end of this procedure, each sample has been predicted and the proportion of misclassified samples can be computed. Here k is set to 5. Since the partition into training and test data used during this procedure is inherently random and may lead to non-representative results for a single run, we report median errors over 1000 cross-validation runs.

In order to size the expected performance of a random classifier, we use a simple permutation test. Basically, complexes are randomly predicted by permuting the ligand types in the original data set and assigning the result of the permutation to each complex. This procedure maintains the number of complexes per ligand type. Median errors over 10000 random permutations are reported.

Ligand redundancy. In total, there are 465 distinct ligands out of 489 complexes, with the most represented ones appearing at most 3 times. Overfitting due to Ag redundancy in the dataset is therefore not an issue.

2.5 Predicting binding affinities

Relevant variables. The affinity prediction problem was recently revisited and posed as a sparse linear model estimation problem [46], stressing the importance of two variables. These two variables turn out to be the most informative ones when estimating binding affinities, in the sense where they get selected most often amidst a pool of variables modeling relevant biophysical properties [46].

The first one, the inverse volume-weighted internal path length (IVW-IPL), encodes the size and morphology of the interface and takes atomic packing into account. Let $\mathcal{I}$ be the set of interface atoms in a complex. Let $\mathrm{SO}(a)$ and $\mathrm{Vol}(a)$ be the shelling order and Vol_bound the volume of atom a in the complex (see section 2.1). The, IVW-IPL is defined as follows:

$$\text{IVW-IPL} = \sum_{a \in \mathcal{I}} \frac{\mathrm{SO}(a)}{\mathrm{Vol_bound}(a)} \tag{3}$$

On the one hand, the shelling order refines so-called core-rim models [41]. Borrowing to the notion of cooperative effects involving non-bonded weak interactions, an isotropic or disk-like interface is indeed expected to be more stable than an elongated one—even if their surface areas match. On the other hand, the atomic packing encodes the local density of neighbors of a given atom, and thus provides a measure for local interactions (hydrogen bonds, van der Waals interactions). Note that packing is a subtle quantity related to the enthalpy - entropy compensation discussed in Introduction, as its properties strike a balance between enthalpy (a large number of neighbors favors interactions) and entropy (too small of a packing is detrimental for dynamics yielding an entropic penalty).

The second variable ($\mathrm{NIS}^{charged}$) is the fraction of charged residues on the non-interacting surface (NIS, *i.e.* the exposed surface of the Ig and of the Ag not involved in the interface). The NIS is meant to encode electrostatic properties and solvent interactions [30].

Statistical methodology. We estimate binding affinities using k nearest neighbors regression (knn) [25, 6], a non-parametric regression strategy which does not require any a priori on the

mathematical model for the response variable estimated – as opposed to linear regression for instance. This strategy is a two step strategy. As a pre-processing step, we compute the parameters IVW-IPL and $\text{NIS}^{charged}$ for the training set (125 cases), yielding a point cloud P in the two dimensional space defined by IVW-IPL and $\text{NIS}^{charged}$. (Fig. 2a). To estimate the affinity of a complex q (an Ig - Ag case), we proceed in two steps. First, the k nearest neighbors of q in P are sought, with k a predefined number. Second, the affinity of q is estimated by averaging those of its k nearest neighbors. (Practically, the scikit-learn library [52] was used, namely the `neighbors` package for knn regression.)

We assess the quality of our predictions by varying the value k. From a theoretical standpoint [25], it is known that k must be super-logarithmic and sub-linear in the number of cases processed. Since $log(144) \sim 5$, we explore the range $k \in 5, \ldots, 25$ (Fig. 2b and Figure S9). The results discussed in the main text correspond to $k = 10$.

In order to assess the impact of the distance to nearest neighbors and of the consistency of their affinity values on the accuracy of the predictions, we compute the average distance d_i between each Ig - Ag complex i and its $k = 10$ nearest neighbors in the training set (*i.e.* those used to estimate its binding affinity using k-nearest neighbor regression). We also compute the standard deviation of the affinity values σ_i of these 10 nearest neighbors. These are compared to the absolute error $|e_i|$ ($= |\text{experimental_affinity}_i - \text{predicted_affinity}_i|$) of the prediction on complex i.

2.6 Comparing the energetic contribution of interface atoms between CDRs

To assess the respective energetic contributions of CDRs to binding affinity, we dissect the IVW-IPL (Eq. (3)) into the contributions of CDR1 + CDR2 and CDR3. We also compute the *average normalized shelling order* (or ANSO for short) for each CDR

$$\text{ANSO} = \frac{1}{|A|} \sum_{a \in A} \frac{\text{SO}(a)}{\text{Vol_bound}(a)}, \tag{4}$$

with A is the set of interface atoms of the CDR and the size of this set is $|A|$. The distribution of IVW-IPL and ANSO between CDR1 + 2 and CDR3 within the same chain are then compared using a Wilcoxon signed-rank test.

3 Results

3.1 Characteristics of the binding patch predict the ligand type

Atomic solvent accessibility asymmetry is a signature for the ligand type. A classical and informative variable describing a protein - protein interface is the buried surface area (BSA), which is known to correlate to the number of interface atoms [27]. In our case, a Pearson coefficient equal to 0.99 is obtained. However, this value drops down to 0.82 and 0.89 respectively for the Ig and the Ag sides, a fact owing to the shape complementarity between the binding patches on the Ig and Ag sides (Figure S4). To further investigate this observation, we compute the average BSA per interface atom for both the Ig and Ag (Eqs. (1) and (2)). Strikingly, the ligand type has a strong impact on these quantities: complexes involving a chemical ligand have a higher average BSA per atom at the Ag side of the interface ($\overline{\text{bsa}}_{\text{Ag}}$) than those involving a peptide ligand which in turn have a higher $\overline{\text{bsa}}_{\text{Ag}}$ than those involving a protein ligand (Fig. 1).

Note that $\overline{\text{bsa}}_{\text{Ag}}$ and $\overline{\text{bsa}}_{\text{Ig}}$ can be seen as proxies for curvature of the Ag and Ig binding patches, hence their strong inverse correlation due to the complementarity between binding patches on the Ig and Ag sides (Figure S3(D, E, F)). This inverse correlation is rather intuitive for small ligands, but may not be trivial for bigger antigens. Our contribution corroborates this fact for a whole set of structures.

To further exploit the ability of the parameters $\overline{\text{bsa}}_{\text{Ag}}$ and $\overline{\text{bsa}}_{\text{Ig}}$ to characterize interfaces as a function of the ligand type, we build a decision tree classifier (section 2.4, Fig. 1 and Figure S5).

The median cross-validated error over all classes is 9.6% over 1000 repetitions whereas the permutation test resulted in a median error of 56%. More precisely, the median cross-validated error rates per class are 5%, 19% and 7% for chemical, peptides and proteins. The higher error rate for peptides is mostly due to the classifier predicting proteins instead of peptides (Table SIV), which is not unexpected as the criterion to classify polypeptides as peptides or proteins is not standardized. For comparison, the permutation test resulted in error rates of 84% for chemicals, 75% for peptides, and 41% for proteins; clearly showing the influence of the number of complexes per class on the accuracy of the prediction. Overall, our classifier is able to accurately predicts ligand types, despite the fact that the data is unbalanced.

3.2 Binding affinity predictions

Our k-nearest neighbors based model predicts 8 (57.14%), 13 (92.86 %) and 13 of the dissociation constants K_d within one, two and three orders of magnitude respectively, with a median absolute error of 0.878 kcal/mol, which corresponds in a ratio for K_d equal to 4.4 (Fig. 2c). In terms of correlation coefficients, one gets 0.488 (Pearson) and 0.291 (Spearman). These results are very good, as predicting K_d within one order of magnitude is essentially the best one can hope for without modeling subtle effects such as the pH in particular [26]. They are also informative from a biological standpoint, as an affinity enhancement of two orders of magnitude is typically observed during affinity maturation. In order to compare these results to what could be expected from a null model, we take the average binding affinity of the training dataset (10.78 kcal/mol $\pm$2.84) as prediction for all complexes. This results in a median absolute error of 1.03 kcal/mol, or equivalently, in a ratio for K_d equal to 5.7. The previous conclusions must therefore be mitigated, since a simple null model can show good, albeit less so, performances as well.

In order to rationalize the varying accuracy of predictions depending on the complex, we compute the average distance d_i between each Ig - Ag complex i and its 10 nearest neighbors in the training set. We also compute the standard deviation of the affinity values of these 10 nearest neighbors σ_i (Figure S12). Both d_i and σ_i are weakly correlated to the absolute prediction error $|e_i|$ with Pearson's correlation coefficients of 0.57 and -0.57 respectively. Both coefficients are (weakly) significantly different from zero with p-values of 0.0312 and 0.03316 respectively. The correlation between $|e_i|$ and d_i/σ_i is higher however with a Pearson correlation coefficient equal to 0.72 and a p-value of 0.00363. This suggests that good binding affinity prediction can be obtained provided that sufficiently similar complexes are in the training set and that their affinity values are consistent with each other. Interestingly, this property also accounts for the good performances of the null model.

The success of the affinity prediction owes to two important properties of the learning set (non-Ig - Ag complexes) and the training set (Ig - Ag complexes). First, Ig - Ag complexes fall in a reduced region of the space defined by the two parameters IVW-IPL and $\text{NIS}^{charged}$ of the model, *i.e.* they are similar from the point of view of the model. Second, the Ig - Ag complexes fall in a region which is well represented in the training set (*i.e.*, the rest of the SAB). This means that in the space of the two parameters of the model, Ig - Ag complexes are similar to the

other protein - protein complexes of the SAB. In order to predict the binding affinity of Ig - Ag complexes with protein ligands, our model therefore takes advantage of the fact that they are similar both to each other and similar to other protein - protein complexes.

Comparison with the PRODIGY server. In order to see how our approach fares against the state of the art, we compare our results against the PRODIGY server. The PRODIGY server is one of the most recent tools for affinity prediction [66], and is based on the work from Vangone *et al* [63].

The accuracy of PRODIGY is lower than that of the current study with median absolute errors of 1.4 versus 0.878 kcal/mol respectively. For reference, we also provide the root mean squared errors (2.226 versus 1.676 kcal/mol), Pearson's correlation coefficients (0.149 versus 0.488) and Spearman's correlation coefficients (0.238 versus 0.291). Interestingly, our method is successful at predicting similar affinities (Fig. 11) for five complexes (1AHW, 1DQJ, 1VFB, 2JEL, 1BJ1) for which PRODIGY predicts widely varying values.

CDRs: lengths and BSA. It has been observed that CDR lengths differ between different antigen types [16, 55], a finding suggesting that CDR lengths influence the binding site to accommodate the ligand. We therefore undertook the characterization of this relationship in the IMGT/3Dstructure-DB. Since all the atoms of a CDR may not contribute to the interface, we investigate the correlation between the length of a CDR and its contribution to the BSA. As CDR1 and CDR2 are both encoded by V genes we study them together and subsequently investigate the relationship between [CDR1 . CDR2] pairs and BSA on the one hand, and CDR3 and BSA on the other hand. We observe that CDRs of a given length can display widely varying levels of BSA (Figure S6 and Figure S7). These results indicate that CDR lengths must be complemented to fully describe the involvement of a CDR in the interaction with the Ag. This is backed up by the very limited ability of neural networks trained on sequence data only to predict the ligand type bound by an Ig in [16]. An error rate of 54% is indeed observed, to be compared to a baseline of 75% for a random predictor on four classes (protein, hapten, nucleotide and viral protein) [16].

Respective contributions of the CDRs to the interface, for VH and VL domains. In an Ig - Ag complex, it is generally believed that VH contributes more to the recognition than VL. With a BSA of VH strictly larger than that of VL for 430/489 complexes ($\sim$86%) (Fig. 3a), our analyses support this idea. To refine this view, we split the BSA into contributions by the CDRs within a V-domain, observing a great deal of variation across the dataset, independent from the ligand type (Figs. 3b and 3c). A general observation is that the sum of contributions of CDR1 and CDR2 essentially matches that of CDR3 for both VH and VL. Consider the sum of the BSA of CDR1 and CDR2 on one hand, and the BSA of CDR3 on the other hand. The first quantity is larger than the second one for $\sim$46% of the complexes for VH, and for $\sim$40% of the complexes for VL. Moreover, a Wilcoxon signed-rank test does not find a significant difference between them for VH (two-sided p-value = 0.1460), but does for VL (two-sided p-value = 0.0001), indicating that the contribution of CDR3 in terms of BSA and relative to other CDRs from the same chain is higher for the light chain than for the heavy chain.

To assess the contributions of CDRs to binding energy, we compute both their IVW-IPL and ANSO (Eq. (3) and (4)) for all complexes (Fig. 4a and 4b). We then compare the distributions of these two quantities for CDR1 + 2 and CDR3 in the same chain, using a Wilcoxon signed-rank test at significance level $\alpha = 0.01$. Consider the sum of the IVW-IPL of CDR1 and CDR2 on one hand, and the IVW-IPL of CDR3 on the other hand. The first quantity is larger than the

second one for ∼41% of the complexes for VH, and for ∼27% of the complexes VL (Figure S8). Wilcoxon signed-rank tests find significant differences between them for both VH (two-sided p-value = $6.404 \cdot 10^{-7}$), and VL (two-sided p-value = $7.217 \cdot 10^{-30}$). Removing the dependence on the number of atoms, *i.e.* comparing the ANSO distribution computed on both CDR1 and CDR2 on the one hand and CDR3 on the other hand, leads to significant differences as well for VH (two-sided p-value = $6.221 \cdot 10^{-30}$), and VL (two-sided p-value = $2.480 \cdot 10^{-37}$).

Thus, as opposed to the results obtained when considering the BSA, the sum of contributions to the binding affinity of CDR1 and CDR2 is significantly lower than that of CDR3 for both VH and VL.

For both chains, the difference in ANSO can be imputed to two facts. First the average shelling order (Section 2.1) for atoms of the CDR3 is higher than those of CDR 1 and 2 (Fig. 4c). Second, their average atomic volume is lower (Fig. 4d). Both are related since the shelling order and the atomic volume are negatively correlated (Figure S8c).

4 Discussion

In this work, we provide a precise quantitative description of Ig - Ag interfaces, leading to an accurate classification of ligand types and to accurate binding affinity predictions. We also quantify the contributions made by CDRs at interface both in terms of surface area and binding energy, and we show that VH CDR3 is the main factor determining binding affinity and interaction specificity. While these facts were previously known from a qualitative standpoint, the task of designing quantitative models supporting them had remained elusive, with insights focused on specific conformations. Instead, our models provide quantitative estimates illustrating the relationship between structure, dynamics and affinity of Ig - Ag complexes.

Enhanced specificity and affinity descriptions from global interface statistics. The buried surface area (BSA) of a protein complex has long been known to be a simple and informative descriptor of interfaces [3]. We refine this statistic by computing the average BSA contributed by interfacial atoms from the Ig (statistic $\overline{\text{bsa}}_{\text{Ig}}$) and the Ag (statistic $\overline{\text{bsa}}_{\text{Ag}}$). These quantities turn out to be clear a signature of the ligand type, a property which can further be exploited for classification purposes. While the classification of Ig - Ag interfaces into classes depending on structural features has already been addressed [11, 35], our parameters are the first ones yielding such a clear separation between specific antigen types.

To complement this analysis, we perform binding affinity predictions for 14 Ig - Ag complexes, based on structural parameters encoding enthalpic and entropic quantities [46]. Our predictions of K_d are accurate within two orders of magnitude for all but one complex and within one order of magnitude for 8 of them. Interestingly, these results stress the relevance of the overall approach, which exploits structural and functional similarities between the test set (the Ig - Ag complexes) and the training set (the SAB deprived from the Ig - Ag complexes). In fact, the high accuracy of our predictions shows that the binding affinity prediction problem could be partially solved using large databases of Ig - Ag complexes with binding affinity measurements.

Our results on specificity analysis and affinity predictions are of immediate practical relevance in the context of Ig design and Ig - Ag docking. Docking is the problem of predicting the pose (*i.e.*, the static structure) and the affinity of a complex from the unbound partners [38]. The latter problem is harder than the former, another embodiment of the role of dynamics in the emergence of function. Our parameters are of high interest for both problems. At the pose prediction stage, they provide filters to check that putative Ig - Ag complexes proposed by docking algorithms comply with our classification rules, as a function of the ligand type. In

a similar spirit, these parameters are of direct relevance to predict the ligand type from the structure of the Ig VH+VL domains. At the affinity prediction stage, assuming a good quality (*i.e.*, resolution) putative structure for the complex, reliable affinity predictions can be made.

These results also call for extensions, in particular to handle different ligand types (peptides, haptens). Since the quality of predictions owes in particular to a good coverage of the region of the model space targeted by predictions, this extension is likely to be successful assuming a database—identical in spirit to the SAB, providing sufficiently many cases to learn from. From a formal standpoint, we also envision progress on the analysis of the correctness of affinity predictions, based on two ingredients. The first one is the accuracy of estimators for thermodynamic quantities, using parameters such as those used in this work. The second one is the mathematical convergence of regressors, in particular those based on nearest neighbors, as used in this work.

Bridging the gap between structure, dynamics and function. Our findings show that global structural parameters perform remarkably well to predict affinity and specificity, which are notions formally defined in the realm of thermodynamics. It is therefore instrumental to understand which features of CDRs explain the relevance of our parameters. In other words, it appears important to consider at once the role of the six CDRs for most antibody specificities.

If the molecules studied were perfectly rigid, local interactions (hydrogen bonds and van der Waals interactions) would play a prominent role in the formation of the Ig - Ag complex, and the comparable BSA contributed by CDR1+2 vs CDR3 would hint at commensurable contributions from all CDRs. This purely enthalpic view is however insufficient, as preconfiguration/prerigidification of the binding site may yield a decreased entropic loss upon complex formation, hence a enhanced binding affinity [45, 56, 12, 57]. A useful proxy for dynamics is the length of VH CDR3, and difficulties were observed to define canonical conformations for VH CDR3 [14, 1, 58, 59, 32, 51] as opposed to the other CDRs. Indeed, accurate sequence-based conformation predictions are limited to the base or *torso* of the VH CDR3. In this work, we code the enthalpy - entropy compensation (see discussion in section 2.5) using packing properties via our parameters IVW-IPL and ANSO. This leads to two important observations: first, independently of the number of interface atoms, VH CDR3 contributes significantly more to the binding energy than VH CDR1 and VH CDR2 combined; second, interface atoms in VH CDR3 are more closely packed than in other CDRs in the heavy chain. The latter point implies that it is important to minimize the entropic penalty entailed upon binding, which can be achieved by preformation *i.e.* the CDR is in bound conformation prior to the binding event. Interestingly the authors of [65] come to the conclusion that VH CDR3 is responsible for the specificity of the interaction whereas the other CDRs account for its stability. We provide a quantitative view on this property, based on our parameters IVW-IPL and ANSO.

Summarizing, the genetic variability of VH CDR3 is complemented structurally by its dynamic nature to make it the main factor involved in the determination of the specificity and increase of affinity of an Ig for an Ag. It should be stressed that, although this observation can be used as a guide during the design of Ig, it is by no means necessary, as tight binders can be designed *de novo* without any CDR – see [22] for an example involving the stem of influenza virus hemagglutinin.

Naturally, one should also expand our analysis at the whole Ig level, as various structural features of Igs influence their efficacy in the immune response. These include the ball-and-socket joint relating VL and VH, the CL and CH1 constant domains [39, 60], and more generally the constant regions which have been shown to influence the avidity [17, 18, 47, 54], and are involved in Ig effector properties, such as ADCC or CDC [24]. A quantitative assessment of the role of these features requires going beyond the Ig - Ag interface level, with a clear focus on the dynamics

of the whole Ig protein. Again, the identification of the most relevant degrees of freedom in such regions may pave the way to efficient simulation and design strategies.

5 Figures

Figure 1

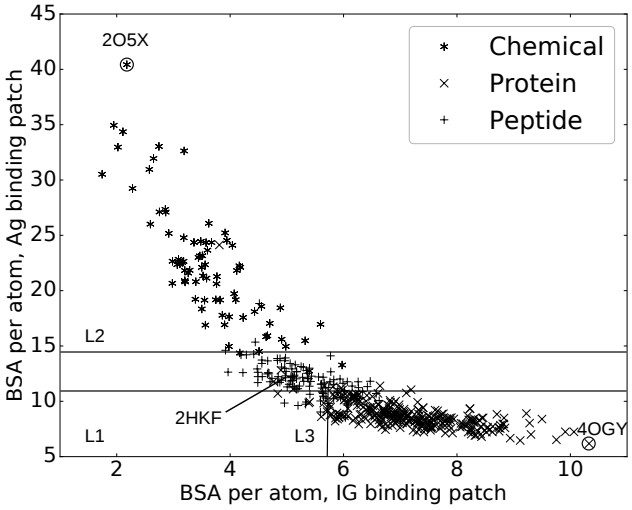
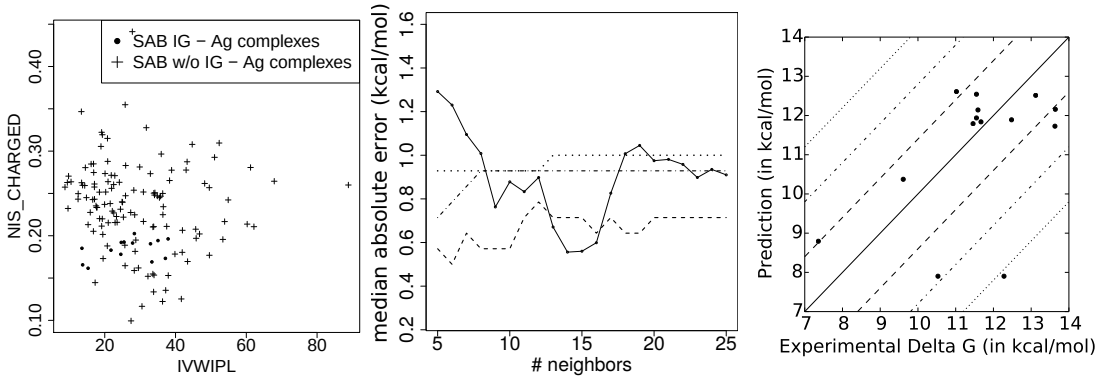
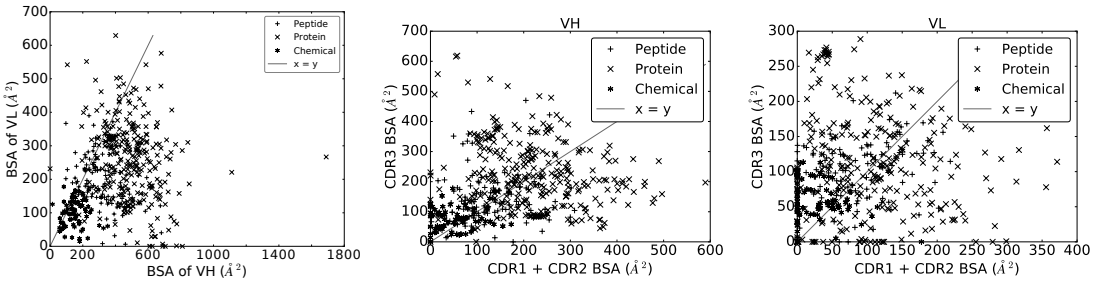


Figure 2



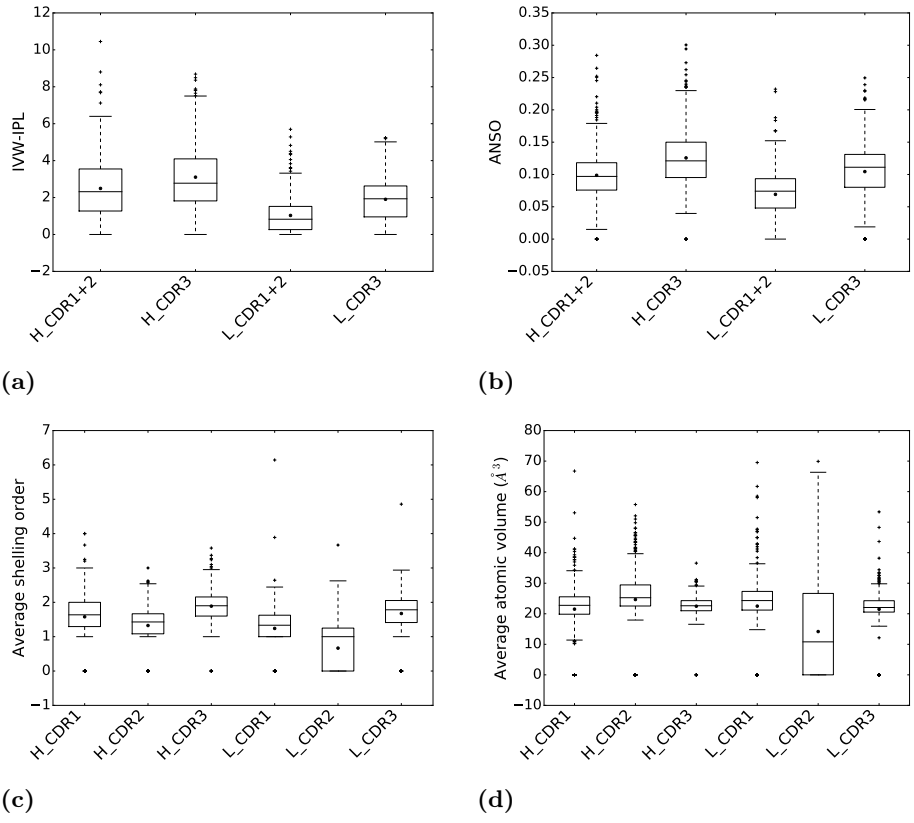
(a) **(b)** **(c)**

Figure 3



(a) VH versus VL **(b)** VH domain. **(c)** VL domain.

Figure 4



(a) **(b)** **(c)** **(d)**

Figure legends

Figure 1. Interaction specificity for Ig - Ag complexes: analysis and predictions. Both analyses are based upon the average buried surface areas per atom (Equations (1) (2)): $\overline{\text{bsa}}_{\text{Ig}}$ versus $\overline{\text{bsa}}_{\text{Ag}}$. Scatter plot as a function of the ligand type. The three lines (L1, L2 and L3) show the partition defined by the decision tree, separating the ligand types (see main text).

The points labeled 2O5X, 2HKF and 4OGY correspond to complexes displayed in Figure S3.

Figure 2. Binding affinity analysis and predictions for Ig - Ag complexes. (2a) Complexes in the two-parameter space of the model. The model uses two variables (see main text): IVWIPL: Inverse volume weighted internal path length; NIS_CHARGED: proportion of charged residue on the non-interacting solvent-accessible surface. **(2b) Stability of affinity prediction.** Performance of the k nearest neighbors estimates when varying the number of neighbors k. Solid line: median absolute error (kcal/mol); dashed, dot-dashed, dotted lines: proportion of predictions with error below 1, 2 and 3 orders of magnitude respectively. **(2c) Predicted versus experimental affinities for Ig - Ag complexes.** Dashed, dash-dotted and dotted lines respectively show errors of ± 1.4, ± 2.8, ± 4.2 kcal/mol, corresponding to K_d approximated within one, two and three orders of magnitude.

Figure 3 Buried Surface Area (A^2) of the VH and VL domains, and their respective CDRs.

Figure 4. Comparison of CDRs in terms of (a) inverse volume-weighted internal path length (IVW-IPL), (b) average normalized shelling order (ANSO), (c) average shelling order, and (d) average atomic volumes.

Acknowledgments. Patrice Duroux is acknowledged for his help with the IMGT/3Dstructure-DB database.

References

[1] B. Al-Lazikani, A.M. Lesk, and C. Chothia. Standard conformations for the canonical structures of immunoglobulins. *Journal of molecular biology*, 273(4):927–948, 1997.

[2] J.C Almagro. Identification of differences in the specificity-determining residues of antibodies that recognize antigens of different size: implications for the rational design of antibody repertoires. *Journal of Molecular Recognition*, 17(2):132–143, 2004.

[3] R. Bahadur, P. Chakrabarti, F. Rodier, and J. Janin. A dissection of specific and non-specific protein–protein interfaces. *JMB*, 336(4):943–955, 2004.

[4] F.D. Batista and M.S. Neuberger. Affinity dependence of the b cell response to antigen: a threshold, a ceiling, and the importance of off-rate. *Immunity*, 8(6):751–759, 1998.

[5] T.N. Bhat, G.A. Bentley, G. Boulot, M.I. Greene, D. Tello, W. Dall'Acqua, H. Souchon, F.P. Schwarz, R.A. Mariuzza, and R.J. Poljak. Bound water molecules and conformational stabilization help mediate an antigen-antibody association. *PNAS*, 91:1089–1093, 1994.

[6] G. Biau and L. Devroye. *Lectures on the nearest neighbor method*. Springer, 2015.

[7] B. Bouvier, R. Grunberg, M. Nilgès, and F. Cazals. Shelling the Voronoi interface of protein-protein complexes reveals patterns of residue conservation, dynamics and composition. *Proteins: structure, function, and bioinformatics*, 76(3):677–692, 2009.

[8] F. Cazals. Revisiting the Voronoi description of protein-protein interfaces: Algorithms. In T. Dijkstra, E. Tsivtsivadze, E. Marchiori, and T. Heskes, editors, *International Conference on Pattern Recognition in Bioinformatics*, pages 419–430, Nijmegen, the Netherlands, 2010. Lecture Notes in Bioinformatics 6282.

[9] F. Cazals, F. Proust, R. Bahadur, and J. Janin. Revisiting the Voronoi description of protein-protein interfaces. *Protein Science*, 15(9):2082–2092, 2006.

[10] A. Chailyan, P. Marcatili, D. Cirillo, and A. Tramontano. Structural repertoire of immunoglobulin λ light chains. *Proteins: Structure, Function, and Bioinformatics*, 79(5):1513–1524, 2011.

[11] A. Chailyan, P. Marcatili, and A. Tramontano. The association of heavy and light chain variable domains in antibodies: implications for antigen specificity. *FEBS Journal*, 278(16):2858–2866, 2011.

[12] C.A. Chia-en, W. Chen, and M.K. Gilson. Ligand configurational entropy and protein binding. *PNAS*, 104(5):1534–1539, 2007.

[13] Y. Choi and C.M. Deane. Predicting antibody complementarity determining region structures without classification. *Molecular Biosystems*, 7(12):3327–3334, 2011.

[14] C. Chothia and A.M. Lesk. Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Bio*, 196(4), 1987.

[15] C. Chothia, A.M. Lesk, A. Tramontano, M. Levitt, S.J. Smith-Gill, G. Air, S. Sheriff, E.A. Padlan, D. Davies, W.R. Tulip, et al. Conformations of immunoglobulin hypervariable regions. *Nature*, 342(6252):877–883, 1989.

[16] A.V.J. Collis, A.P. Brouwer, and A.C.R. Martin. Analysis of the antigen combining site: correlations between length and sequence composition of the hypervariable loops and the nature of the antigen. *Journal of molecular biology*, 325(2):337–354, 2003.

[17] L.J. Cooper, A.R. Shikhman, D.D Glass, D. Kangisser, M.W. Cunningham, and N.S. Greenspan. Role of heavy chain constant domains in antibody-antigen interaction. apparent specificity differences among streptococcal IgG antibodies expressing identical variable domains. *The Journal of Immunology*, 150(6):2231–2242, 1993.

[18] L.J.N Cooper, D. Robertson, R. Granzow, and N.S. Greenspan. Variable domain-identical antibodies exhibit IgG subclass-related differences in affinity and kinetic constants as determined by surface plasmon resonance. *Molecular immunology*, 31(8):577–584, 1994.

[19] J. Dunitz. Win some, lose some: enthalpy-entropy compensation in weak intermolecular interactions. *Chemistry & biology*, 2(11):709–712, 1995.

[20] F. Ehrenmann, Q. Kaas, and M-P. Lefranc. IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucl. Acids Res.*, 38:D301–307, 2010.

[21] W.J.J Finlay and J.C. Almagro. Natural and man-made v-gene repertoires for antibody discovery. *Frontiers in immunology*, 3, 2012.

[22] S.J. Fleishman, T.A. Whitehead, D.C. Ekiert, C. Dreyfus, J.E. Corn, E.M. Strauch, I.A. Wilson, and D. Baker. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science*, 332(6031):816–821, 2011.

[23] M. Gerstein and F.M. Richards. Protein geometry: volumes, areas, and distances. In M. G. Rossmann and E. Arnold, editors, *The international tables for crystallography (Vol F, Chap. 22)*, pages 531–539. Springer, 2001.

[24] L.W. Guddat, L. Shan, Z-C. Fan, K.N. Andersen, R. Rosauer, D.S. Linthicum, and A.B. Edmundson. Intramolecular signaling upon complexation. *The FASEB journal*, 9(1):101–106, 1995.

[25] L. Györfi and A. Krzyzak. *A distribution-free theory of nonparametric regression*. Springer, 2002.

[26] J. Janin. A minimal model of protein–protein binding affinities. *Protein Science*, 23(12):1813–1817, 2014.

[27] J. Janin, R. P. Bahadur, and P. Chakrabarti. Protein-protein interaction and quaternary structure. *Quarterly reviews of biophysics*, 41(2):133–180, 2008.

[28] S. Jones and JM Thornton. Principles of protein-protein interactions. *PNAS*, 93(1):13–20, 1996.

[29] P.L. Kastritis, I.H. Moal, H. Hwang, Z. Weng, P.A. Bates, A. Bonvin, and J. Janin. A structure-based benchmark for protein-protein binding affinity. *Protein Science*, 20:482–491, 2011.

[30] P.L. Kastritis, J.P.G.L.M. Rodrigues, G.E. Folkers, R. Boelens, and A.M.J.J. Bonvin. Proteins feel more than they see: Fine-tuning of binding affinity by properties of the non-interacting surface. *J.M.B.*, 426:2632–2652, 2014.

[31] O.V. Koliasnikov, M.O. Kiral, V.G. Grigorenko, and A.M. Egorov. Antibody CDR H3 modeling rules: extension for the case of absence of Arg H94 and Asp H101. *Journal of bioinformatics and computational biology*, 4(02):415–424, 2006.

[32] D. Kuroda, H. Shirai, M. Kobori, and H. Nakamura. Structural classification of CDR-H3 revisited: A lesson in antibody modeling. *Proteins: Structure, Function, and Bioinformatics*, 73(3):608–620, 2008.

[33] D. Kuroda, H. Shirai, M. Kobori, and H. Nakamura. Systematic classification of CDR-L3 in antibodies: Implications of the light chain subtypes and the VL–VH interface. *Proteins: Structure, Function, and Bioinformatics*, 75(1):139–146, 2009.

[34] N.F. Landolfi, A.B. Thakur, H. Fu, M. Vásquez, C. Queen, and N. Tsurushita. The integrity of the ball-and-socket joint between V and C domains is essential for complete activity of a humanized antibody. *The Journal of Immunology*, 166(3):1748–1754, 2001.

[35] M. Lee, P. Lloyd, X. Zhang, J.M. Schallhorn, K. Sugimoto, A.G. Leach, G. Sapiro, and K.N. Houk. Shapes of antibody binding sites: qualitative and quantitative analyses based on a geomorphic classification scheme. *The Journal of organic chemistry*, 71(14):5082–5092, 2006.

[36] M-P. Lefranc and G. Lefranc. *The immunoglobulin FactsBook*. Academic Press, 2001.

[37] M-P. Lefranc, C. Pommié, M. Ruiz, V. Giudicelli, E. Foulquier, L. Truong, V. Thouvenin-Contet, and G. Lefranc. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Developmental & Comparative Immunology*, 27(1):55–77, 2003.

[38] M.F. Lensink and S.J. Wodak. Docking, scoring, and affinity prediction in CAPRI. *Proteins: Structure, Function, and Bioinformatics*, 81(12):2082–2095, 2013.

[39] A. Lesk and C. Chothia. Elbow motion in the immunoglobulins involves a molecular ball-and-socket joint. *Nature*, 8(335):188–90, 1988.

[40] S.M. Lippow, K.D. Wittrup, and B. Tidor. Computational design of antibody-affinity improvement beyond in vivo maturation. *Nature biotechnology*, 25(10):1171–1176, 2007.

[41] L. Lo Conte, C. Chothia, and J. Janin. The atomic structure of protein-protein recognition sites. *JMB*, 285(5):2177–2198, 1999.

[42] S. Loriot and F. Cazals. Modeling macro-molecular interfaces with Intervor. *Bioinformatics*, 26(7):964–965, 2010.

[43] R.M. MacCallum, A.C.R. Martin, and J.M. Thornton. Antibody-antigen interactions: contact analysis and binding site topography. *Journal of molecular biology*, 262(5):732–745, 1996.

[44] N. Malod-Dognin, A. Bansal, and F. Cazals. Characterizing the morphology of protein binding patches. *Proteins: structure, function, and bioinformatics*, 80(12):2652–2665, 2012.

[45] V. Manivel, N.C. Sahoo, D.M. Salunke, and K.V.S Rao. Maturation of an antibody response is governed by modulations in flexibility of the antigen-combining site. *Immunity*, 13(5):611–620, 2000.

[46] S. Marillet, P. Boudinot, and F. Cazals. High resolution crystal structures leverage protein binding affinity predictions. *Proteins: structure, function, and bioinformatics*, 1(84):9–20, 2015.

[47] N McCloskey, MW Turner, P Steffner, R Owens, and D Goldblatt. Human constant regions influence the antibody binding characteristics of mouse-human chimeric IgG subclasses. *Immunology*, 88(2):169–173, 1996.

[48] G. Meng, N. Arkus, M.P. Brenner, and V.N. Manoharan. The free-energy landscape of clusters of attractive hard spheres. *Science*, 327(5965):560–563, 2010.

[49] V. Morea, A. Tramontano, M. Rustici, C. Chothia, and A.M. Lesk. Conformations of the third hypervariable region in the VH domain of immunoglobulins. *Journal of molecular biology*, 275(2):269–294, 1998.

[50] Deepak T Nair, Kavita Singh, Zaved Siddiqui, Bishnu P Nayak, Kanury VS Rao, and Dinakar M Salunke. Epitope recognition by diverse antibodies suggests conformational convergence in an antibody response. *The Journal of Immunology*, 168(5):2371–2382, 2002.

[51] B. North, A. Lehmann, and R.L. Dunbrack. A new clustering of antibody CDR loop conformations. *Journal of molecular biology*, 406(2):228–256, 2011.

[52] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[53] O. Pritsch, G. Hudry-Clergeon, M. Buckle, Y. Pétillot, J-P. Bouvet, J. Gagnon, and G. Dighiero. Can immunoglobulin CH1 constant region domain modulate antigen binding affinity of antibodies? *Journal of Clinical Investigation*, 98(10):2235, 1996.

[54] O. Pritsch, C. Magnac, G. Dumas, J-P. Bouvet, P. Alzari, and G. Dighiero. Can isotype switch modulate antigen-binding affinity and influence clonal selection? *European journal of immunology*, 30(12):3387–3395, 2000.

[55] G. Raghunathan, J. Smart, J. Williams, and J-C. Almagro. Antigen-binding site anatomy and somatic mutations in antibodies that recognize different types of antigens. *Journal of Molecular Recognition*, 25(3):103–113, 2012.

[56] D. Rajamani, S. Thiel, S. Vajda, and C.J. Camacho. Anchor residues in protein-protein interactions. *PNAS*, 101(31):11287–11292, 2004.

[57] A. Schmidt, H. Xu, A. Khan, T. O'Donnell, S. Khurana, L. King, J. Manischewitz, H. Golding, P. Suphaphiphat, A. Carfi, E. Settembre, P. Dormitzer, T. Kepler, R. Zhang, A. Moody, B. Haynes, H-X. Liao, D. Shaw, and S. Harrison. Preconfiguration of the antigen-binding site during affinity maturation of a broadly neutralizing influenza virus antibody. *PNAS*, 110(1):264–269, 2013.

[58] H. Shirai, A. Kidera, and H. Nakamura. Structural classification of CDR-H3 in antibodies. *FEBS letters*, 399(1):1–8, 1996.

[59] J. Shirai, A. Kidera, and H. Nakamura. H3-rules: identification of CDR-H3 structures in antibodies. *FEBS letters*, 455(1):188–197, 1999.

[60] R.L. Stanfield, A. Zemla, I.A Wilson, and B. Rupp. Antibody elbow angles are influenced by their light chain class. *Journal of molecular biology*, 357(5):1566–1574, 2006.

[61] E.J. Sundberg and R.A. Mariuzza. Molecular recognition in antibody-antigen complexes. *Advances in protein chemistry*, 61:119–160, 2002.

[62] M. Torres, N. Fernández-Fuentes, A. Fiser, and A. Casadevall. The immunoglobulin heavy chain constant region affects kinetic and thermodynamic parameters of antibody variable region interactions with antigen. *Journal of Biological Chemistry*, 282(18):13917–13927, 2007.

[63] A. Vangone and A. Bonvin. Contacts-based prediction of binding affinity in protein–protein complexes. *eLife*, 4:e07454, 2015.

[64] E. Vargas-Madrazo, F. Lara-Ochoa, and J.C. Almagro. Canonical structure repertoire of the antigen-binding site of immunoglobulins suggests strong geometrical restrictions associated to the mechanism of immune recognition. *Journal of molecular biology*, 254(3):497–504, 1995.

[65] J.L Xu and M.M. Davis. Diversity in the CDR3 region of VH is sufficient for most antibody specificities. *Immunity*, 13(1):37–45, 2000.

[66] Li C Xue, João PGLM Rodrigues, Panagiotis L Kastritis, Alexandre MJJ Bonvin, and Anna Vangone. Prodigy: a web server for predicting the binding affinity of protein-protein complexes. *Bioinformatics*, page btw514, 2016.

6 Supplementary information

6.1 The IMGT dataset

6.1.1 Ig - Ag complexes

We use the Ig - Ag complexes from the IMGT/3Dstructure-DB (`http://www.imgt.org/3Dstructure-DB/` [20]), corresponding to the category *IG/Ag* for *IMGT complex type*. Only IMGT-PDB files are kept.

This dataset features 1602 files. Each such complex is processed in order to identify canonical complexes involving one heavy chain, one light chain, and one ligand (Section S6.1.2). A total of 1275 canonical complexes are thus extracted, of which 554 non-redundant complexes. After further filtering on the ligand types (keeping only complexes with a single ligand type among protein, peptide and chemical), 489 complexes remain.

Fig. S 1 Size of the antigens (number of atoms) Two large peptides (IMGT-PDB file 3W11 chain E, 2301 atoms, and IMGT-PDB file 4R4N chain I, 5172 atoms) are not displayed for readability. See section 2.2 for the data curation methods.

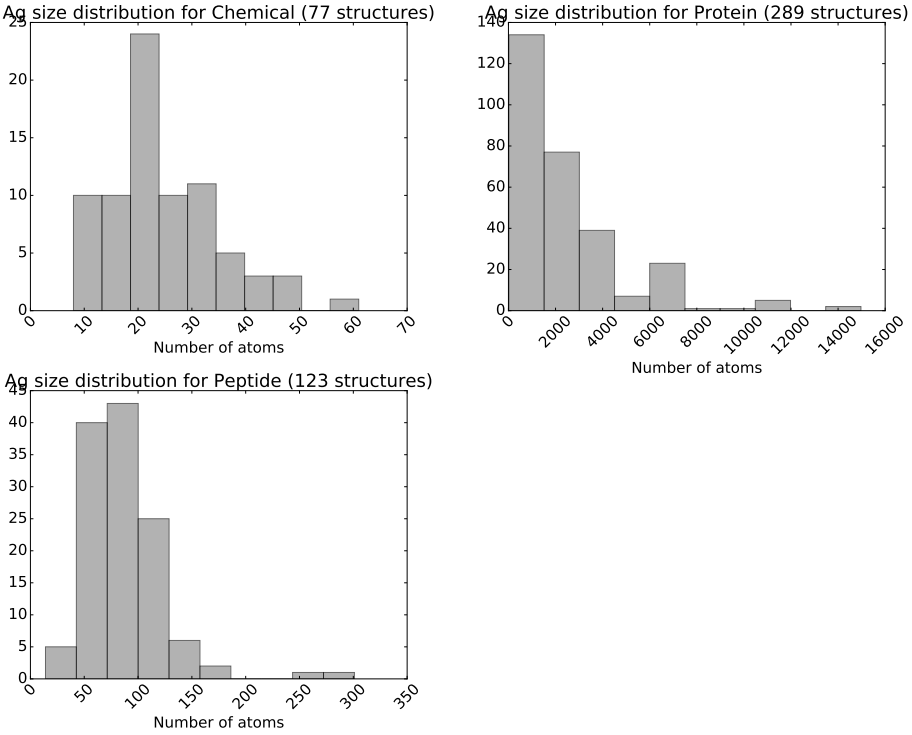


Table S I Summary of the number of Ig - Ag complexes in each class of species / ligand type. The dataset includes VH (V-domains of heavy chains)and VL comprising V-KAPPA (V domains of kappa chains) and V-LAMBDA (V domains of lambda chains. See section 2.2 for the data curation methods.

	Mouse	Human	Other	total
Peptide	80	32	11	123
Protein	168	91	30	289
Chemical	65	7	5	77
total	313	130	46	489

Table S II Amino acid positions associated with each IMGT label defining the decomposition of a V-domain into seven regions Positions of the complementarity determining regions (CDRs) using the IMGT numbering scheme [37]. See section 2.2 for the data curation methods.

Region	FR1	CDR1	FR2	CDR2	FR3	CDR3	FR4
start-stop	1 - 26	27 - 38	39 - 55	56 - 65	66 - 104	105 - 117	118 - 128

6.1.2 Inferring canonical complexes

Canonical complexes. A canonical configuration for a IMGT/3Dstructure-DB IMGT-PDB file is as follows: *one H chain, one L chain, one ligand.* A non canonical configuration may occur for different reasons:

- The asymmetric unit of the crystal structure contains two or more Fabs.
- Several molecules have co-cristallized with the Ig - Ag complex.
- Two Ig chains, H and L, and one Ag chain are found but the Ig chains are not annotated as forming a receptor in the IMGT 410 section.
- An Ig receptor is annotated as containing more than two chains.
- The ligand is a multi-chain protein

The following issues are faced:

- A file may not be canonical *i.e.* there might be several complexes in a single file.
- There might be some issues with the numbering of the chains.
- There might be missing data (residues, chains information, labels)
- Several complexes might be similar and bias the results.
- Some molecules annotated as ligand may actually be buffer molecules (*e.g.* glycerol)
- Some purification proteins remain (*e.g.* protein L, A or G) and do not engage in specific contacts with the Ig

Using the executable `sbl-intervor-ABW-atomic.exe` from the structural bioinformatics library (SBL, `sbl.inria.fr`), which implements the Voronoi interface model presented in section 2.1, we proceed in two steps. First, we infer the chains *pairings* in every file which does not contain a canonical complex. For this, we compute the interfaces between all pairs of chains. We then group L and H chains in pairs for which the number of atoms at the interface is the highest. We then assign the ligand(s) chains to the HL pairs if they make contacts with either chain.

Note that in the case where an Ag is in contact with several Ig, it will be assigned to both Ig.

Finally, all buffer molecules and Ig purification proteins (namely protein L, A and G) whose annotated name satisfy the regexp `"immunoglobulin g-binding | protein[ ]+[gl]($|\s|\') | glycerol | 2-Amino-2-Hydroxymethyl-Propane-1,3-Diol | tris | 2-(N-Morpholino)-Ethanesulfonic Acid"` are removed from the files because they are not representative of Ig - Ag interactions.

Crystal contacts. The previous automatic detection raises the problem of crystal contacts, since complexes reported might be false positives.

They could potentially be ruled out by using a cutoff such as the minimal number of atoms at an interface to be considered significant, however, there might also be few contacts between a Fab and a small ligand. It would therefore be necessary to study the distribution of the number of atoms at the interface for different classes of ligands to set a specific cutoff.

To circumvent this issue, we currently exclude from the analysis complexes in which the ligand does not make at least one contact with the variable domain (CDR or FR).

6.1.3 Removing redundancies from IMGT/3Dstructure-DB

Redundant complexes may come from two sources: the same complex may be found in the same asymmetric crystal unit, or it may be found in two different IMGT-PDB files.

We therefore need to remove the redundancy of the dataset to avoid biasing the statistics. For this, we need to consider similarities at the interface level. Once all complexes are extracted from the database, we need to compare the interfaces of all pairs of complexes, group complexes having a similar interface, and keep one representative complex for each group.

We rely on a quick method based upon IMGT labels. Consider triplets formed by the IMGT labels of both Ig chains and the Ag chain (*e.g.* `(VH-CH1, L-KAPPA, Capsid protein C)`). We record triplets which have already been included in the analysis and exclude complexes which have the same triplet.

6.2 The binding affinity benchmark dataset

Test set. The SAB contains 17 Ig - Ag cases (PDB IDs: 1AHW, 1BJ1, 1BVK, 1DQJ, 1E6J, 1FSK, 1IQD, 1JPS, 1MLC, 1NCA, 1NSN, 1P2C, 1VFB, 1WEJ, 2JEL, 2VIR and 2VIS). Their K_d was determined at temperatures ranging between 20 and 25 °C or reported as ambient/room temperature. The temperature was not reported in one case. The pH during measurements ranged between 7 and 7.5 except in one case where it was 4.8 (1BJ1). It was not reported in five cases, and for two it is likely to have been 7.4 (BIAcore standard). All the Igs are either murine or humanized monoclonal Igs raised against their antigen *in vivo* or *in vitro*, with K_d ranging from $4 \cdot 10^{-6}$ to 10^{-10} kcal/mol (or equivalently, ΔG_d ranging from 7.36 to 13.64 kcal/mol). Out of these 17 cases, 1IQD and 1NSN are discarded as only an upper bound on their K_d is provided in the SAB. Furthermore, 1E6J is also discarded because too many atoms could not be matched between the bound and unbound structures. The 14 remaining cases only involve protein ligands. Among them, five are hen egg lysozymes (HEL), two are a tissue factors(TF), two

are hemagglutinins (HA), and the remaining ones are birch pollen allergen (Bet v 1), cytochrome c (Cytc), HPr protein, neuraminidase (NA) and vascular endothelial growth factor (VEGF). We note that the iRMSD and the total RMSD between the bound and unbound form of the Igs are always smaller than 1.24Å and 0.95Å respectively. That is, the 14 cases are essentially rigid cases.

Training set. The rest of the SAB is used to train the model and is called *training set* in the sequel. 1ZLI is removed from the training set because too many atoms could not be matched between the bound and unbound structures and 1UUG is also removed because only an upper bound on its K_d is provided.

Table S III Main features of the Ig - Ag complexes found in the structure affinity benchmark. CDR bounds correspond to the first and last residue numbers in IMGT renumbered PDB files.

PDB ID	Ig H chain	Ig L chain	Ag chain(s)	Ag type	Species	VH V and J gene	VL V and J gene
1AHW	B	A	C	Protein	Mus musculus (house mouse)	IGHV14-1*02 IGHJ2*01	IGKV14-111*01 IGKJ2*01
1BJ1	H	L	WV	Protein	Humanized (humanized)	IGHV7-4-1*02 IGHJ2*01	IGKV1-33*01 IGKJ1*01
1BVK	E	D	F	Protein	Humanized (humanized)	IGHV4-59*01 IGHJ4*03	IGKV1-27*01 IGKJ1*01
1DQJ	B	A	C	Protein	Mus musculus (house mouse)	IGHV3-8*02 IGHJ6*03	IGKV5-43*01 IGKJ1*02
1FSK	C	B	A	Protein	Mus musculus (house mouse)	IGHV1-61*01 IGHJ3*01	IGKV6-20*01 IGKJ1*02
1JPS	H	L	T	Protein	Homo sapiens (human)	IGHV3-66*04 IGHJ4*03	IGKV1-39*01 IGKJ1*01
1MLC	B	A	E	Protein	Mus musculus (house mouse)	IGHV1-9*01 IGHJ2*01	IGKV5-43*01 IGKJ2*01
1NCA	H	L	N	Protein	Mus musculus (house mouse)	IGHV9-3*03 IGHJ2*01	IGKV6-25*01 IGKJ1*01
1P2C	B	A	C	Protein	Mus musculus (house mouse)	IGHV1-9*01 IGHJ4*01	IGKV5-43*01 IGKJ1*01
1VFB	B	A	C	Protein	Mus musculus (house mouse)	IGHV2-6-7*01 IGHJ2*01	IGKV12-41*02 IGKJ2*01
1WEJ	H	L	F	Protein	Mus musculus (house mouse)	IGHV14-3*02 IGHJ2*01	IGKV12-41*02 IGKJ1*01
2JEL	H	L	P	Protein	Mus musculus (house mouse)	IGHV1-67*01 IGHJ1*01	IGKV1-117*01 IGKJ1*02
2VIR	B	A	C	Protein	Mus musculus (house mouse)	IGHV2-9*02 IGHJ4*01	IGLV1*01 IGLJ1*01
2VIS	B	A	C	Protein	Mus musculus (house mouse)	IGHV2-9*02 IGHJ4*01	IGLV1*01 IGLJ1*01

PDB ID	VH CDR lengths (1, 2, 3)	VL CDR lengths (1, 2, 3)	Ag size (number of atoms)	Ag name
1AHW	8 8 10	6 3 9	1612	Thromboplastin (synonym: tissue factor, TF, coagulation factor
1BJ1	8 8 16	6 3 9	1522	VEGF (Vascular endothelial growth factor A)
1BVK	8 7 10	6 3 9	1001	Lysozyme C [hen egg white] (HEL) EC:3.2.1.17
1DQJ	8 7 7	6 3 9	1007	Lysozyme C [hen egg white] (HEL) EC:3.2.1.17
1FSK	8 8 11	6 3 9	1230	Major birch pollen allergen Bet v1
1JPS	8 8 10	6 3 9	1611	Tissue Factor
1MLC	8 8 9	6 3 9	1001	Lysozyme C [hen egg white] (HEL) EC:3.2.1.17
1NCA	8 8 13	6 3 9	3075	Neuraminidase [influenza virus, A/Tern strain, N9 subtype]
1P2C	8 8 9	6 3 9	1001	Lysozyme C [hen egg white] (HEL) EC:3.2.1.17
1VFB	8 7 10	6 3 9	1265	Lysozyme C [hen egg white] (HEL) EC:3.2.1.17
1WEJ	8 8 10	6 3 9	826	Cytochrome c [horse]
2JEL	8 8 11	11 3 9	640	Histidine-containing protein of the phosphoenolpyruvate: sugar
2VIR	8 7 16	9 3 9	2075	Hemagglutinin HA1 [influenza virus]; residues: 28-328
2VIS	8 7 16	9 3 9	2076	Hemagglutinin HA1 [influenza virus] T131I (escape mutant);

PDB ID	Ig name	Resolution	VH CDR1 bounds	VH CDR2 bounds	VH CDR3 bounds	VL CDR1 bounds	VL CDR2 bounds	VL CDR3 bounds
1AHW	AB-GAMMA-1_KAPPA	3.0	27 38	56 65	105 117	27 38	56 65	105 117
1BJ1	AB-GAMMA-1_KAPPA	2.4	27 38	56 65	105 117	27 38	56 65	105 117
1BVK	V-HEAVY_KAPPA	2.7	27 38	56 65	105 117	27 38	56 65	105 117
1DQJ	AB-GAMMA-2A_KAPPA	2.0	27 38	56 65	105 117	27 38	56 65	105 117
1FSK	AB-GAMMA-1_KAPPA	2.9	27 38	56 65	105 117	27 38	56 65	105 117
1JPS	AB-GAMMA-1_KAPPA	1.85	27 38	56 65	105 117	27 38	56 65	105 117
1MLC	FAB-GAMMA-1_KAPPA	2.5	27 38	56 65	105 117	27 38	56 65	105 117
1NCA	AB-GAMMA-2A_KAPPA	2.5	27 38	56 65	105 117	27 38	56 65	105 117
1P2C	FAB-GAMMA-1_KAPPA	2.0	27 38	56 65	105 117	27 38	56 65	105 117
1VFB	FV-HEAVY_KAPPA	1.8	27 38	56 65	105 117	27 38	56 65	105 117
1WEJ	AB-GAMMA-1_KAPPA	1.8	27 38	56 65	105 117	27 38	56 65	105 117
2JEL	AB-GAMMA-1_KAPPA	2.5	27 38	56 65	105 117	27 38	56 65	105 117
2VIR	AB-GAMMA-1_LAMBDA	3.25	27 38	56 65	105 117	27 38	56 65	105 117
2VIS	AB-GAMMA-1_LAMBDA	3.25	27 38	56 65	105 117	27 38	56 65	105 117

PDB ID	K_d (M)	ΔG (kcal/mol)	iRMSD (Å)	Method	pH
1AHW	$3.40 \cdot 10^{-9}$	-11.55	0.69	Competitive Inhibition assay	not stated
1BJ1	$3.40 \cdot 10^{-9}$	-11.55	0.5	SPR	4.8
1BVK	$1.40 \cdot 10^{-8}$	-10.53	1.24	Stopped-flow inhibition	7
1DQJ	$2.80 \cdot 10^{-9}$	-11.67	0.75	SPR	7.5
1FSK	$2.40 \cdot 10^{-10}$	-13.12	0.45	SPR	7.4
1JPS	$1.00 \cdot 10^{-10}$	-13.64	0.51	SPR	7.2
1MLC	$9.10 \cdot 10^{-8}$	-9.61	0.6	SPR	7.4
1NCA	$8.30 \cdot 10^{-9}$	-11.02	0.24	Fluorescence inhibition assay	7.2
1P2C	$1.02 \cdot 10^{-10}$	-13.63	0.46	SPR	not stated
1VFB	$3.70 \cdot 10^{-9}$	-11.46	1.02	ITC	7.1
1WEJ	$7.14 \cdot 10^{-10}$	-12.48	0.31	Spectroscopic inhibition assay	not stated
2JEL	$2.80 \cdot 10^{-9}$	-11.59	0.17	Fluorescence inhibition assay	7.2
2VIR	$1.00 \cdot 10^{-9}$	-12.28	0.8	SPR	not stated (BIAcore standard:7.4)
2VIS	$4.00 \cdot 10^{-6}$	-7.36	0.8	SPR	not stated (BIAcore standard: 7.4)

6.3 Hierarchical Voronoi interface models

Consider a complex where partner A is an Ig, and partner B an antigen. We wish to accommodate the hierarchical structure of the Fab [36]. We focus on the variable domains of the heavy and light chains, denoted VH and VL respectively, and decompose each of them into seven regions, namely three Complementarity Determining Regions (CDRs), and the four Framework Regions (FRs) flanking them [37] (Table SII). For example, a V domain is decomposed as FR1+CDR1+FR2+CDR2+FR3+CDR3+FR4.

Consider the partition of the variable domains VH and VL induced by the previous 14 labels. For the sake of conciseness and since we focus on interfaces involving the variable domains only, the domains VH and VL are plainly denoted H and L. Using these notations, we partition the $IGAg$ interface as follows:

- Hierarchical bicolor interface (no water): $IGAg = (L \cup H)Ag = LAg \cup HAg$

- Hierarchical mediated interface (water mediated only): $IGW - AgW = (LW - AgW) \cup (HW - AgW)$

- Hierarchical tricolor interface (both): $IGAgW = IGAg \cup (IGW - AgW)$

Analogously, the partition of the H (or L) V-domain into seven CDR and FR regions induces a partition of the HAg (or LAg) interface (Figure S2).

The Voronoi facets associated to pairs of type (A, B) define the *bicolor* interface $A-B$ (bicolor since there are two partners); those associated to pairs of type (A, W) and (B, W) define the mediated interface $AW - BW$, since interactions between A and B are mediated by W(ater) molecules; finally, the union of the bicolor and mediated interface define the *tricolor interface* ABW. Geometrically, this interface is a polyhedron separating the partners. The curvature of this polyhedron is easily computed [8], and has been shown to provide information on binding modes [9].

Fig. S 2 Decomposition of an Ig - Ag complex. The Ig (or the Fab fragment) is decomposed into heavy (H) and light (L) chains (one H and one L per Fab) whose variable domains only (VH and VL) are of interest in this study. These domains are further decomposed into three complementarity determining regions (CDRs) and four framework regions (FRs). The Voronoi interface of Figure S3 is partitioned into contributions from these 14 regions.

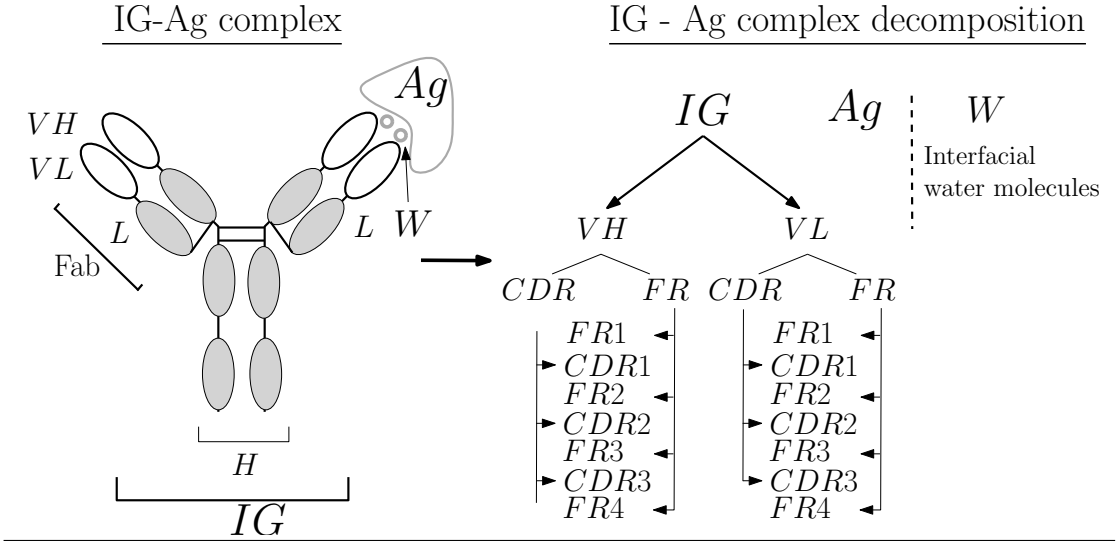
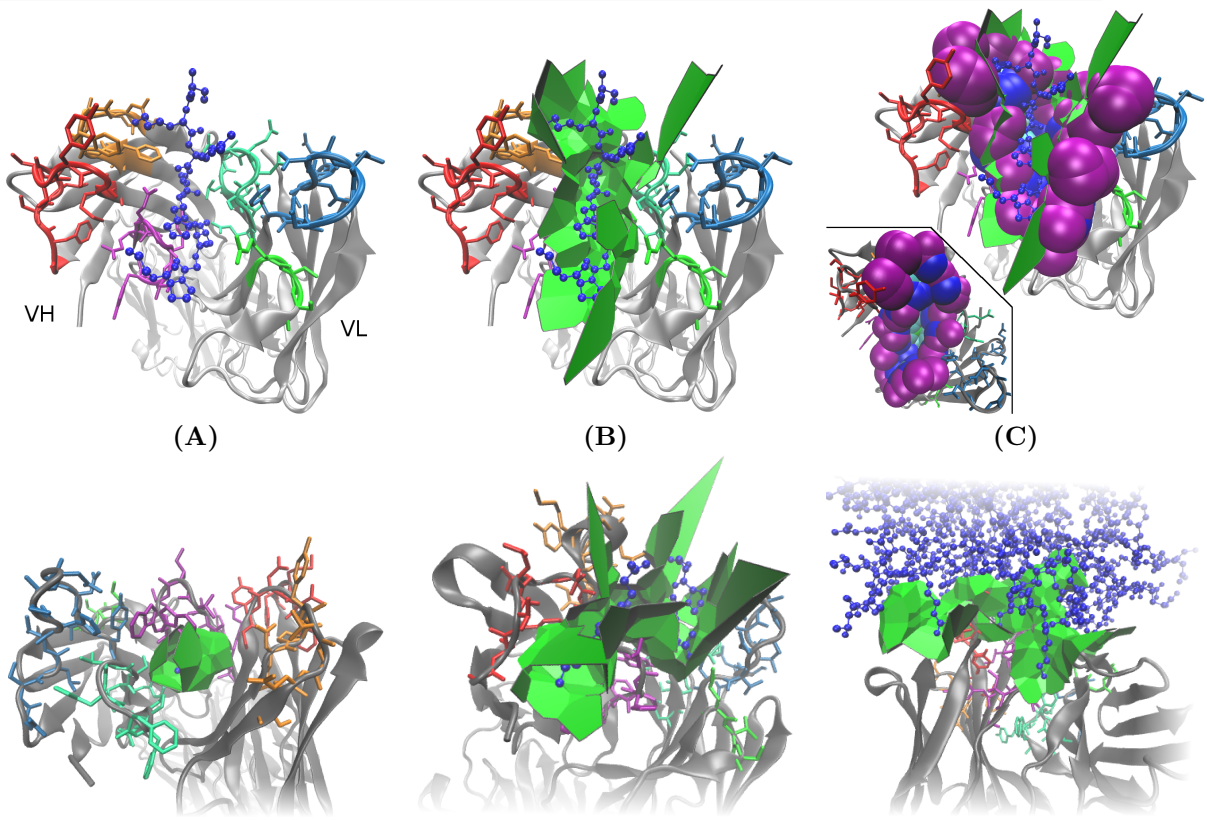


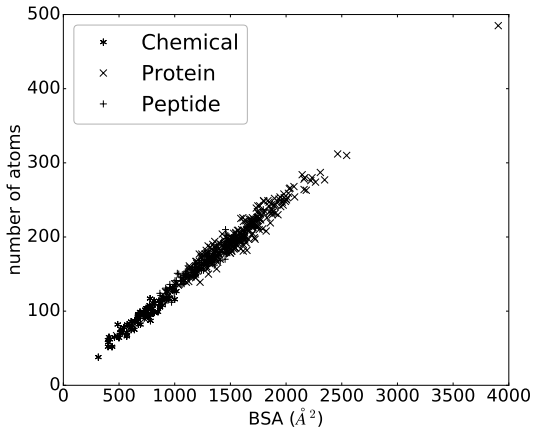
Fig. S 3 Voronoi interface model of an Immunoglobulin - Antigen (Ig - Ag) complex, defined from the solvent accessible model of the crystallographic complex. The Ig consists of H and L chains, with here the VH and VL domains shown in grey (cartoon representation), while the Ag consists of the chain in blue (CPK representation). **(A)** Ig - Ag complex, with the six complementarity determining regions (CDRs) colored using the IMGT conventions (VH CDR1: red, VH CDR2: orange, VH CDR3: purple, VL CDR1: blue, VL CDR2: green, VL CDR3: green-blue). **(B)** The Voronoi interface is a polyedral model separating the partners, whose parameters (area, curvature) convey information about the binding modes. **(C)** Each face of the Voronoi interface involves two interacting atoms, either from the partners or the interfacial water molecules sandwiched between them. The *buried surface area* (BSA) on each partner (by the second partner and interfacial water) is of prime interest to describe the interface. For the Ig, the BSA can be charged to the CDRs and framework regions (FRs). **(C, inset)** The interface atoms of a partner define its binding patch, which can be shelled into concentric shells (from the outside to the core), defining a distance to the patch boundary. The binding patch on the Ig side is shown from above (inset) where purple, blue and cyan identify atoms with shelling order 1, 2 and 3 respectively. **(D, E, F)** Voronoi interface of three complexes in (a) to illustrate different types: convex on the Ig side (small chemical ligand), saddle-like (peptide ligand), concave on the Ig side (protein ligand).



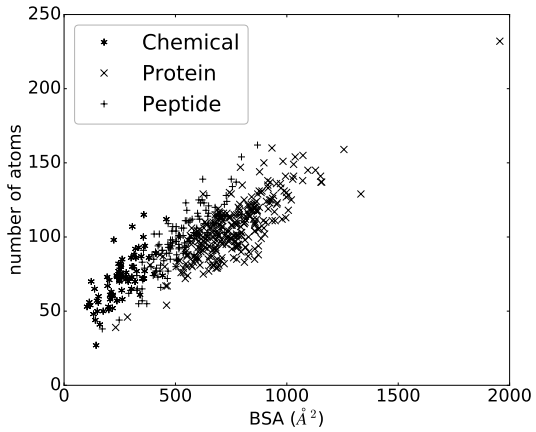
6.4 Characteristics of the binding patch predict the ligand type

Solvent accessibility properties of binding patches exhibit a broken symmetry between the Ig and Ag side. In Figure S4, there is a strong correlation between BSA and $|\mathcal{I}|$, a very-well known fact. However,this is less obvious when considering only Ig atoms or Ag atoms. This shows how the shape complementarity between binding patches results in an overall balance between the BSA and the number of atoms at interface.

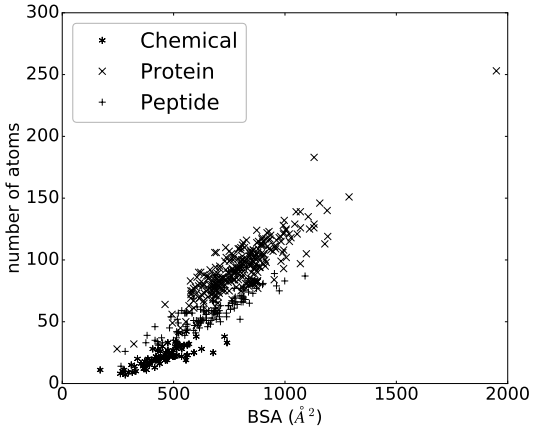
Fig. S 4 Buried Surface Area versus number of interface atoms: whole interface, Ig side, Ag side. The well-known strong correlation between BSA() and $|\mathcal{I}|$ (panel (a)) gets weaker when considering the Ig (panel (b)) and the Ag sides (panel (c)) separately. The Pearson coefficients obtained are equal to 0.99, 0.82 and 0.89 in cases (a,b,c).



(a) BSA versus $|\mathcal{I}|$.



(b) $\mathbf{BSA_{Ig}}$ versus $|\mathcal{I}_{\mathbf{Ig}}|$.



(c) $\mathbf{BSA_{Ag}}$ versus $|\mathcal{I}_{\mathbf{Ag}}|$.

Solvent accessibility asymmetry is a signature for the ligand type. We feed two descriptors ($\overline{\text{bsa}}_{\text{Ag}}$ and $\overline{\text{bsa}}_{\text{Ig}}$) to a classifier in order predict the ligand type of a complex. These descriptors are computed using the `sbl-intervor-ABW-atomic.exe` binary from the structural bioinformatics library (SBL, `sbl.inria.fr`). The classifier is a decision tree from the R package

`rpart`.

We first compute the in-sample classification error. This error is obtained by classifying the training set and results in optimistic error rates. We therefore compute a *cross-validated* error rate. To this end, we randomly divide the dataset in five subsets of equal size and alternatively use four subsets to classify the fifth. At the end of this procedure, each sample has been predicted and the proportion of misclassified samples can be computed. Since the data is partitioned randomly, we repeat the above procedure 1000 times and compute the average confusion matrix(Table SIV) and both the overall and per class error rates.

The resulting median and average error rates per ligand type are the following: chemical: 5%, 6%; peptide: 19%, 19% ; protein: 7%, 7%. The overall median and average error rates are 9.6% and 9.7% respectively.

Since the data is not balanced, *i.e.* some ligand types are over-represented compared to others we check whether keeping a balanced proportion of classes in each fold would yield differing results. The resulting median and average error rates per ligand type are the following: chemical: 5%, 6%; peptide: 20%, 20% ; protein: 7%, 6%; and the overall median and average error rates are 9.6% and 9.7% respectively, which is essentially similar to the non-balanced cross-validation.

The classification rules resulting from the decision tree run on the whole dataset (i.e. no-cross-validation) are the following (Figure S5) : $\overline{\text{bsa}}_{\text{Ag}} \geq 14.3 \Rightarrow$ chemical ligand; $10.7 \leq \overline{\text{bsa}}_{\text{Ag}} < 14.3 \Rightarrow$ peptide ligand; $\overline{\text{bsa}}_{\text{Ag}} < 10.7$ AND $\overline{\text{bsa}}_{\text{Ig}} < 5.75 \Rightarrow$ peptide ligand; $\overline{\text{bsa}}_{\text{Ag}} < 10.7$ AND $\overline{\text{bsa}}_{\text{Ig}} \geq 5.75 \Rightarrow$ protein ligand.

Fig. S 5 Classification rules characterizing the binding patch depending on the ligand types. See supplemental text for details. The classification rules are: $\overline{\text{bsa}}_{\text{Ag}} \geq 14.3 \Rightarrow$ chemical ligand; $10.7 \leq \overline{\text{bsa}}_{\text{Ag}} < 14.3 \Rightarrow$ peptide ligand; $\overline{\text{bsa}}_{\text{Ag}} < 10.7$ AND $\overline{\text{bsa}}_{\text{Ig}} < 5.75 \Rightarrow$ peptide ligand; $\overline{\text{bsa}}_{\text{Ag}} < 10.7$ AND $\overline{\text{bsa}}_{\text{Ig}} \geq 5.75 \Rightarrow$ protein ligand. The three lines of a box read as follows: tow row: majority ligand type (chemical, peptide, protein); middle row: fraction for the three classes; bottom row: percentage of the whole dataset.

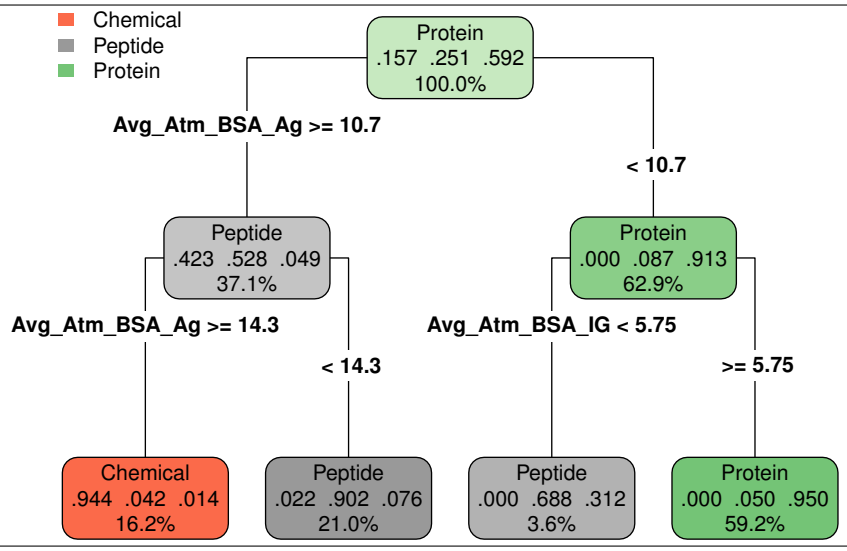


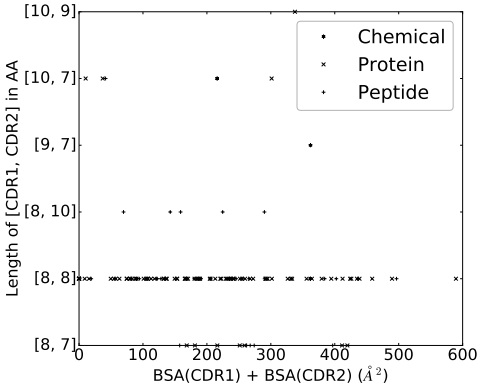
Table S IV Average confusion matrix for ligand type prediction. Results obtained by running 5-fold cross-validation 1000 times. Each repetition results in a confusion matrix which is averaged–e.g. on average 4.6 chemicals out of 77 are predicted as peptides.

Predicted \ Actual	Chemical	Peptide	Protein
Chemical	72.4	3.0	1.0
Peptide	4.6	99.1	17.9
Protein	0.0	20.9	270.1

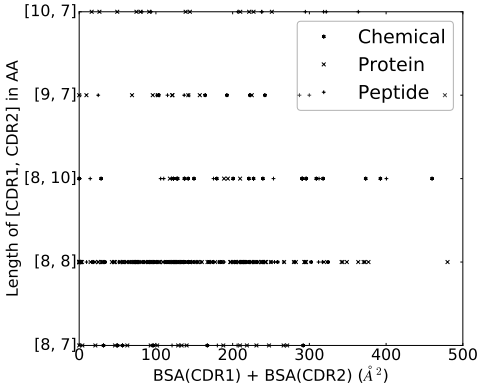
Table S V Median BSA and median of BSA/BSA$_{\mathbf{Ig}}$ per species and per ligand type. Median BSA contributed to the interface by different parts of the Ig, for various ligand types and species. Percentages relative to the BSA of the whole Ig are included in parentheses.

	Human			Mouse			Other		
	Chemical	Peptide	Protein	Chemical	Peptide	Protein	Chemical	Peptide	Protein
Ig	143.7 (100.0%)	645.7 (100.0%)	812.0 (100.0%)	254.6 (100.0%)	486.6 (100.0%)	703.9 (100.0%)	275.0 (100.0%)	546.6 (100.0%)	767.4 (100.0%)
VH	89.8 (57.0%)	418.3 (66.2%)	569.9 (67.7%)	154.8 (59.9%)	287.0 (58.4%)	471.1 (64.3%)	141.1 (54.4%)	307.5 (54.0%)	462.8 (63.7%)
VH CDR	49.4 (39.6%)	381.3 (56.9%)	510.7 (62.1%)	128.3 (46.6%)	236.8 (49.5%)	434.9 (57.4%)	116.2 (42.3%)	241.3 (50.1%)	420.0 (58.9%)
VH CDR1	10.3 (7.5%)	63.0 (11.1%)	86.2 (9.5%)	19.0 (8.5%)	46.5 (10.0%)	101.7 (14.7%)	23.6 (7.7%)	37.6 (8.2%)	54.5 (8.0%)
VH CDR2	0.0 (0.0%)	76.3 (12.6%)	139.0 (17.4%)	13.0 (5.0%)	61.4 (13.2%)	99.8 (14.4%)	0.0 (0.0%)	55.9 (13.8%)	148.9 (20.5%)
VH CDR3	44.0 (26.9%)	180.3 (29.3%)	233.9 (30.4%)	76.9 (29.4%)	117.4 (24.1%)	208.7 (29.1%)	97.3 (35.4%)	108.5 (20.7%)	193.2 (26.1%)
VH FR	25.8 (11.7%)	24.5 (4.7%)	30.6 (3.4%)	24.0 (7.8%)	42.6 (8.4%)	28.8 (4.3%)	15.0 (6.9%)	38.2 (7.1%)	48.7 (7.6%)
VH FR1	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)
VH FR2	8.1 (5.3%)	6.8 (1.6%)	0.0 (0.0%)	11.9 (3.8%)	22.5 (4.4%)	6.6 (0.9%)	11.3 (3.7%)	15.5 (2.9%)	0.1 (0.0%)
VH FR3	0.0 (0.0%)	15.5 (2.1%)	18.1 (2.5%)	0.0 (0.0%)	10.8 (2.3%)	15.7 (2.4%)	0.0 (0.0%)	11.5 (2.1%)	29.2 (3.5%)
VH FR4	0.1 (0.1%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)
VH OTHER	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)
VL	71.2 (43.0%)	185.6 (33.8%)	261.6 (32.3%)	104.2 (40.1%)	190.6 (41.6%)	276.3 (35.7%)	132.8 (45.6%)	203.7 (46.0%)	292.9 (36.3%)
VL CDR	46.3 (30.2%)	162.9 (33.4%)	211.1 (26.9%)	92.5 (37.4%)	175.3 (38.2%)	224.5 (30.4%)	113.6 (41.3%)	196.4 (37.5%)	247.2 (32.6%)
VL CDR1	0.0 (0.0%)	46.5 (9.2%)	74.7 (8.3%)	32.4 (13.3%)	76.6 (15.7%)	52.5 (7.9%)	13.5 (4.9%)	49.5 (11.7%)	86.1 (12.8%)
VL CDR2	0.0 (0.0%)	0.0 (0.0%)	3.9 (0.5%)	0.0 (0.0%)	0.0 (0.0%)	10.0 (1.5%)	0.0 (0.0%)	0.0 (0.0%)	8.7 (0.9%)
VL CDR3	45.0 (23.7%)	114.9 (17.9%)	98.1 (12.1%)	52.9 (22.4%)	101.7 (20.6%)	103.3 (16.0%)	94.4 (34.2%)	124.6 (23.2%)	111.4 (14.5%)
VL FR	25.0 (16.3%)	0.0 (0.0%)	11.1 (1.3%)	0.0 (0.0%)	0.0 (0.0%)	6.8 (0.9%)	3.7 (1.2%)	16.8 (2.8%)	2.8 (0.4%)
VL FR1	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)
VL FR2	15.4 (9.5%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	3.7 (1.2%)	16.2 (2.7%)	0.0 (0.0%)
VL FR3	0.0 (0.0%)	0.0 (0.0%)	0.9 (0.1%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)
VL FR4	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)
VL OTHER	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)

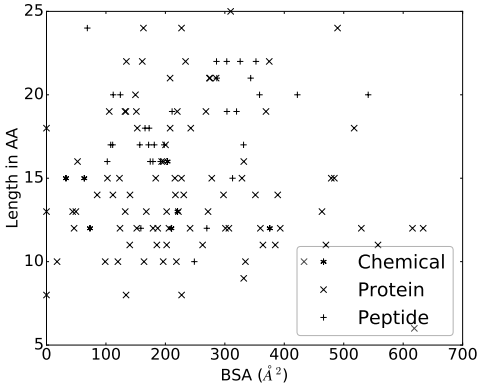
Fig. S 6 Human and mouse VH CDR length versus BSA. The [CDR1. CDR2] length are characteristic of the different *Homo sapiens* and *Mus musculus* VH subgroups. There are highly varying levels of BSA for CDR of the same length. The information given by the length of a CDR is therefore not sufficient to infer its contribution to the interface.



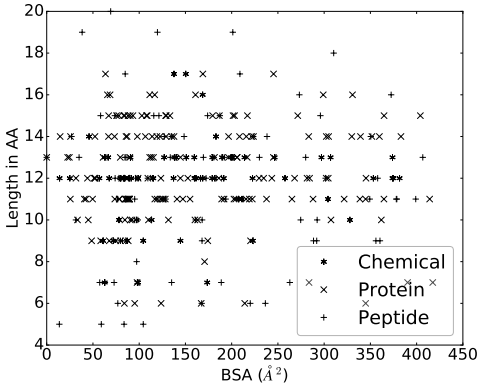
(a) Human [VH CDR1. VH CDR2]. Five complexes are discarded because of aberrant VH CDR1 and VH CDR2 lengths



(b) Mouse [VH CDR1. VH CDR2].

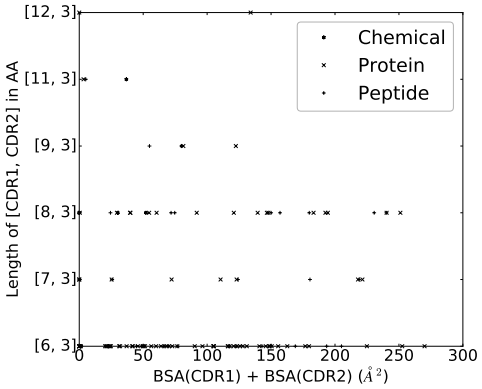


(c) Human VH CDR3. Twelve complexes are discarded because of aberrant VL CDR1 and VL CDR2 lengths

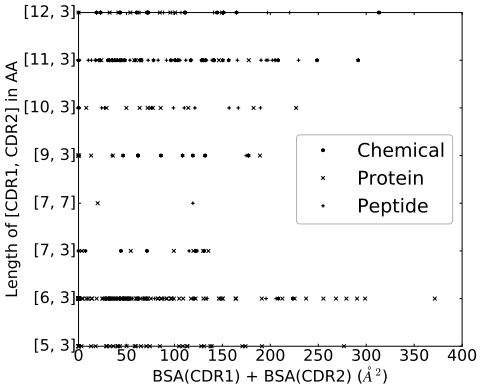


(d) Mouse VH CDR3.

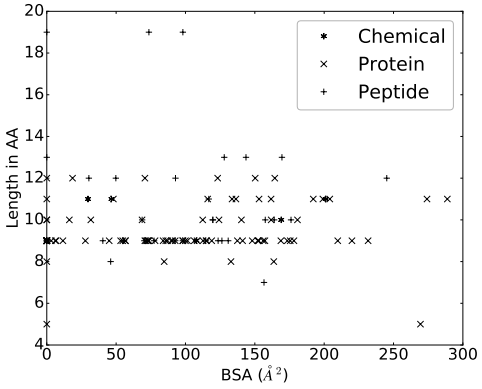
Fig. S 7 Human and mouse VL CDR length versus BSA. The human [CDR1.CDR2] lengths [6.3] characterize both V-kappa and V-lambda. The other lengths characterize either V-kappa ([7.3], [11.3] and [12.3]) or V-lambda ([8.3] and [9.3]). The mouse [CDR1.CDR2] lengths [7.7] and [9.3] characterize V-lambda. The other lengths characterize V-kappa. There are highly varying levels of BSA for CDR of the same length. The information given by the length of a CDR is therefore not sufficient to infer its contribution to the interface.



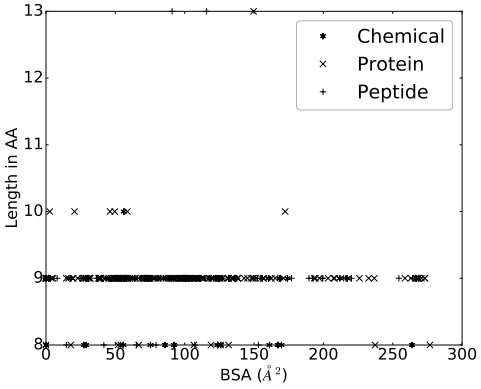
(a) Human [VL CDR1. VL CDR2].



(b) Mouse VL CDR1 and VL CDR2.



(c) VL CDR3, Human.

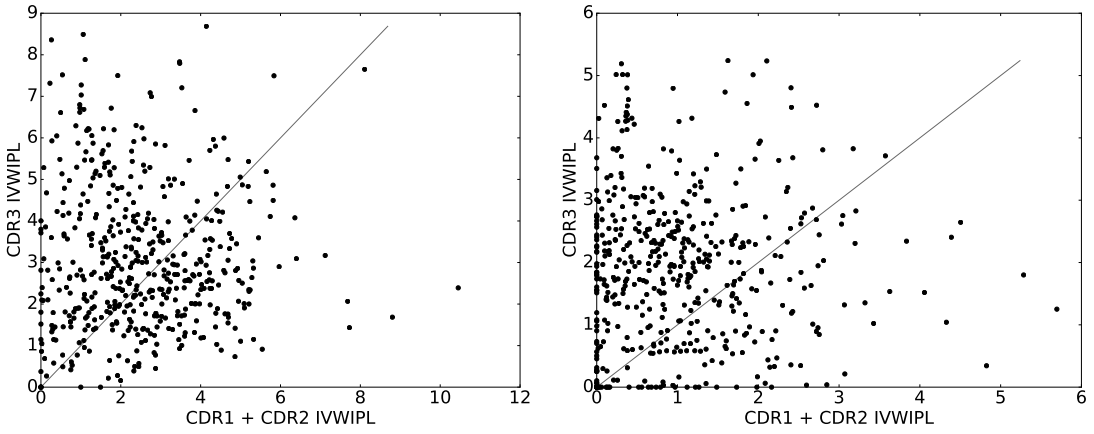


(d) VL CDR3, Mouse.

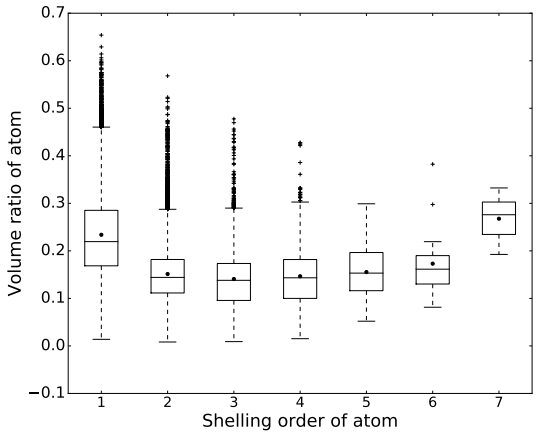
Complexes with free VH CDR1 or VH CDR2 are not uncommon since they occur for 45/489 (∼9%) and 50/489 (∼10%) complexes, respectively (3b). In contrast, this is rare for VH CDR3 (6/489 occurrences, ∼1%).

On the other hand, 242/489 (∼49%) complexes involve a free VL CDR2 (3c), a fact to be interpreted in the context of a lesser length variability and, as we shall see, a location on the side of the domain. 70/489 (∼14%) complexes have an free VL CDR1 and 34/489 (∼7%) complexes have an free VL CDR3 which is more than five time as much as VH CDR3.

Fig. S 8 a) and b): IVW-IPL of the CDR of VH and VL respectively. c): Variation of the atomic volume as a function of the shelling order. Atoms with a higher shelling order tend to be more packed. The rise after shelling order 4 is likely due to a much smaller number of atoms since 1) interfaces with deeply buried atoms are rare, 2) only a limited number of atoms can be deeply buried in an interface.



(a) **(b)**



(c)

6.5 Binding affinity predictions

Ig - Ag complexes. The structure affinity benchmark (SAB) [29] contains 17 Ig - Ag cases (PDB IDs: 1AHW, 1BJ1, 1BVK, 1DQJ, 1E6J, 1FSK, 1IQD, 1JPS, 1MLC, 1NCA, 1NSN, 1P2C, 1VFB, 1WEJ, 2JEL, 2VIR and 2VIS). However, 1IQD and 1NSN are discarded as only an upper bound on their K_d is provided in the SAB. Furthermore, 1E6J is also discarded because too many atoms could not be matched between the bound and unbound structures. Finally 1ZLI is removed from the training set because too many atoms could not be matched between the bound and unbound structures and 1UUG is also removed from the training set because of only an upper bound on its K_d is provided.

Note that this leaves 14 Ig - Ag cases to predict, using a learning set involving $144-2-3-14 = 125$ complexes.

Affinity estimation as a regression problem. In this work, we estimate binding affinities using k nearest neighbors regression (knn) [25, 6], a non parametric strategy which does not require any a priori on the mathematical model for the response variable estimated – as opposed to say linear regression.

To describe knn regression, which is a two step strategy, recall that we model complexes using two parameters denoted IVW-IPL and $\text{NIS}^{charged}$ (see main text); these variables define a parameter space denoted $\mathcal{C}$. As a pre-processing step, we compute the parameters IVW-IPL and $\text{NIS}^{charged}$ for the training set (125 cases), yielding a point cloud P in $\mathcal{C}$ (Fig. 2, right panel). To estimate the affinity a complex q, we proceed in two steps. First, the k nearest neighbors of q in P are sought, with k a predefined number. Second, the affinity of q is estimated by averaging those of its k nearest neighbors.

We assess the quality of our predictions in two ways:

- First, by varying the value k. From a theoretical standpoint [25], it is known that k must be super-logarithmic and sub-linear in the number of cases processed. Since $log(144) \sim 5$, we explore the range $k \in 5, \ldots, 25$ (Fig. 2, right panel). Practically, the results discussed in the main text correspond to $k = 10$.

- Second, by comparing the estimates against those yielded by the linear model using the same variables. This model corresponds to the one defined in [46]. See Table SVI for this comparison.

Practically, we compute the variables used by the regression method using the `binding affinity prediction` package from the structural bioinformatics library (SBL, `sbl.inria.fr`). For the fitting, we use the scikit-learn library [52], namely the `neighbors` package for knn regression, and the `linear_model` package for the least-squares linear regression.

Results. The median absolute error does not vary much between 8 and 12 neighbors (from 1.008 to 0.898, Figure S9). Moreover, the error for 10 neighbors is close to the median error for the total range considered (0.878 compared to 0.910 kcal/mol).

Statistics for the KNN predictions: Pearson's correlation coefficient: 0.488 (p-value = 0.077). Median absolute error: 0.878. Number of complexes with errors below 1, 2, and 3 orders of magnitude respectively: 8, 13, 3.

Statistics for predictions: obtained with the linear model from [46]: Pearson's correlation coefficient: 0.326 (p-value = 0.255). Median absolute error: 1.017. Number of complexes with errors below 1, 2, and 3 orders of magnitude respectively: 9, 13, 14.

Fig. S 9 Stability of affinity prediction using k nearest neighbors estimates, when varying the number of neighbors k.

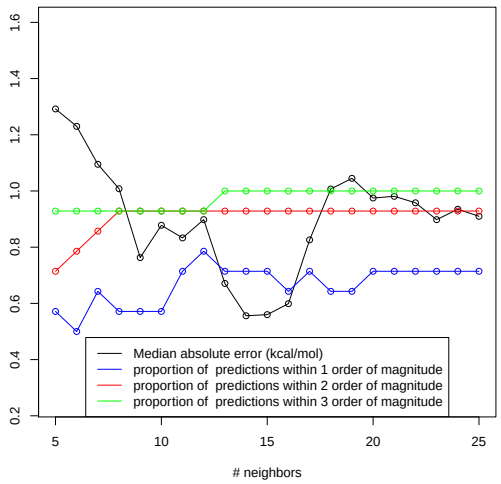


Table S VI Binding affinities: experimental values against those obtained with two regression methods. KNN predictions: obtained using the k-nearest neighbors regressor, as explained in the text; linear model predictions: obtained with the linear model introduced in [46].

PDB ID	Experimental ΔG_d	KNN predictions	linear model predictions
1AHW	11.55	11.938	11.296
1BJ1	11.55	12.543	12.093
1BVK	10.53	7.902	10.699
1DQJ	11.67	11.839	12.522
1FSK	13.12	12.515	11.526
1JPS	13.64	12.159	11.943
1MLC	9.61	10.374	11.293
1NCA	11.02	12.61	11.326
1P2C	13.63	11.728	11.882
1VFB	11.46	11.792	12.354
1WEJ	12.48	11.891	11.338
2JEL	11.59	12.142	11.580
2VIR	12.28	7.902	11.120
2VIS	7.36	8.794	11.306

Fig. S 10 Binding affinity analysis and predictions for Ig - Ag complexes. Predictions performed by the linear model from [46]. Dashed, dash-dotted and dotted lines respectively show errors of ±1.4, ±2.8, ±4.2 kcal/mol, corresponding to K_d approximated within one, two and three orders of magnitude.

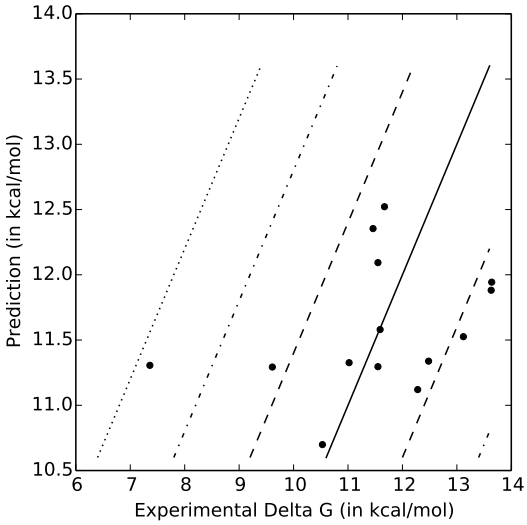


Fig. S 11 Comparison between this work and the PRODIGY server. The vertical dashed lines materialize the experimental values of the complexes. Labels are positioned next to the corresponding red dot.

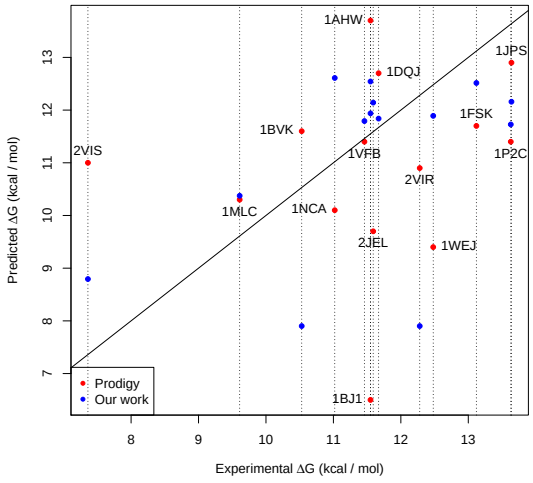
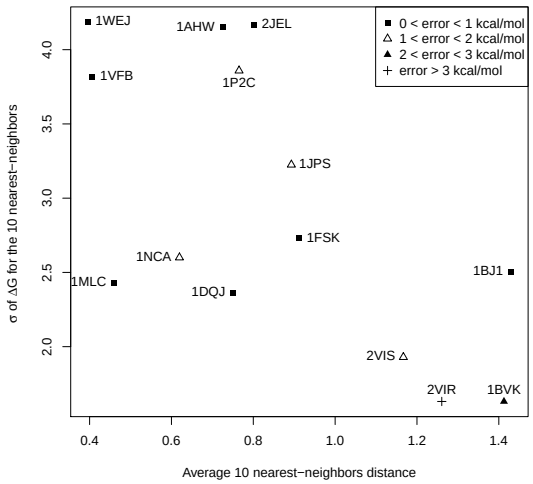


Fig. S 12 Prediction error versus average distance of the 10 nearest-neighbors and the standard deviation of their affinity values.



Contents

1 Introduction **3**

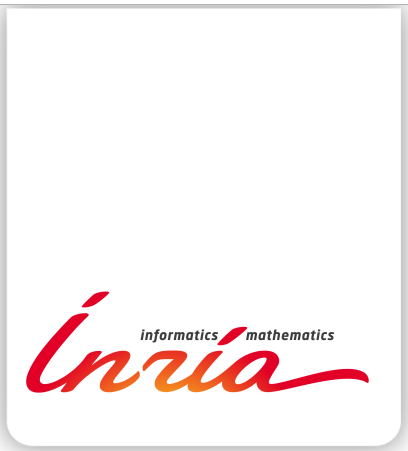
2 Material and methods **5**
2.1 Voronoi interface models 5
2.2 The dataset and data curation: the IMGT/3Dstructure-DB 6
2.3 The binding affinity benchmark 6
2.4 Predicting ligand types 6
2.5 Predicting binding affinities 7
2.6 Comparing the energetic contribution of interface atoms between CDRs 8

3 Results **8**
3.1 Characteristics of the binding patch predict the ligand type 8
3.2 Binding affinity predictions 9

4 Discussion **11**

5 Figures **14**

6 Supplementary information **21**
6.1 The IMGT dataset 21
6.1.1 Ig - Ag complexes 21
6.1.2 Inferring canonical complexes 22
6.1.3 Removing redundancies from IMGT/3Dstructure-DB 23
6.2 The binding affinity benchmark dataset 23
6.3 Hierarchical Voronoi interface models 26
6.4 Characteristics of the binding patch predict the ligand type 29
6.5 Binding affinity predictions 35



RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399