Modeling the antibody response: from the structure of immunoglobulin - antigen complexes to the clonal complexity of heavy chain repertoires.

Simon Marillet

UNIVERSITÉ NICE SOPHIA ANTIPOLIS

ÉCOLE DOCTORALE STIC

SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

T H È S E

pour obtenir le titre de

Docteur en Sciences

de Université Côte d'Azur

Mention Informatique

présentée et soutenue par

Simon MARILLET

Modélisation de la réponse des anticorps: de la structure des complexes immunoglobuline - antigène à la complexité clonale des répertoires de chaines lourdes d'immunoglobulines.

Thèse dirigée par Frédéric CAZALS et co-dirigée par Pierre BOUDINOT

soutenue le 02/12/2016

Jury:

M. Paul Bates	Directeur de Recherche, Institut Francis Crick	Rapporteur
M. Alexandre Bonvin	Professeur, Université d'Utrecht	Rapporteur
M. Dominique Housset	Directeur de Recherche, CEA	Examinateur
Mme Véronique Braud	Directeur de Recherche, CNRS	Examinateur
M. Frédéric Cazals	Directeur de Recherche, INRIA	Directeur
M. Pierre Boudinot	Directeur de Recherche, INRA	Co-directeur

Contents

1 Introduction (Version Française) 1
1.1 Les anticorps au sein du système immunitaire. . . . 1
1.1.1 Le système immunitaire . . . 1
1.1.2 Génétique moléculaire des Ig: diversité et répertoires . . . 2
1.1.3 Structure et fonction des immunoglobulines . . . 3
1.1.4 Reconnaissance entre immunoglobulines et antigènes . . . 3
1.2 Modélisation des complexes Ig - Ag . . . 3
1.2.1 Notions clés de chimie physique . . . 3
1.2.2 Estimation de l'affinité à partir de données structurales . . . 4
1.3 Contributions . . . 4

2 Introduction (English Version) 7
2.1 The immune system . . . 7
2.1.1 The innate immune system . . . 8
2.1.2 The adaptive immune system . . . 11
2.2 Antibody diversity: molecular genetics and repertoires . . . 15
2.2.1 Molecular genetics of Ig gene diversification [Pau13, Chap. 6] . . . 15
2.2.2 Repertoires: potential and available . . . 19
2.3 Structure and function of immunoglobulins . . . 19
2.3.1 General structure of immunoglobulins . . . 19
2.3.2 Atomic structure of immunoglobulins . . . 23
2.3.3 Immunoglobulins - antigen recognition . . . 24
2.4 Modeling of Ig - Ag complexes. . . . 26
2.4.1 Key notions in physical chemistry . . . 26
2.4.2 Structural properties of Ig - Ag complexes . . . 31
2.4.3 Binding free energy estimation from structural data. . . . 32
2.4.4 Terms used by binding affinity models . . . 34
2.4.5 Ig - Ag specificity . . . 35
2.4.6 Review of the latest statistical models for binding affinity prediction. . . . 36
2.5 Contributions / Executive summary . . . 38
2.5.1 Binding affinity prediction . . . 38
2.5.2 Ig - Ag binding affinity and specificity. . . . 38
2.5.3 Comparison of antibody repertoires. . . . 39
2.5.4 Software: Binding affinity prediction modeling . . . 39

3 Protein - protein affinity prediction: High Resolution Crystal Structures Leverage Protein Binding Affinity Predictions 41
3.1 Introduction . . . 41
3.2 Estimating Affinities: Datasets and Parameters . . . 42
3.2.1 Datasets from the Structure Affinity Benchmark . . . 42
3.2.2 Parameters involved in Affinity Prediction Models . . . 43
3.2.3 Parameters Computation . . . 46
3.2.4 Statistical Methodology . . . 47

3.3 Results . . . 52
3.3.1 Specific Predictive Models . . . 52
3.3.2 Specific predictive Models yield Enhanced Correlations. . . . 52
3.3.3 . . . and Improved Predictions on a per Complex Basis . . . 53
3.3.4 Accounting for Interface Morphology and Packing Boosts Performances . . . 57
3.3.5 Performances using a k-nearest neighbors predictor . . . 59
3.4 Discussion and Outlook . . . 59

4 Novel structural parameters of Ig - Ag complexes yield a quantitative description of interaction specificity and binding affinity 61
4.1 Introduction . . . 61
4.2 Material and methods . . . 62
4.2.1 The dataset and data curation: the IMGT/3Dstructure-DB . . . 62
4.2.2 The binding affinity benchmark . . . 64
4.2.3 Voronoi interface models . . . 65
4.2.4 Predicting ligand types . . . 67
4.2.5 Predicting binding affinities . . . 68
4.2.6 Comparing the energetic contribution of interface atoms between CDRs . . . 69
4.3 Results . . . 71
4.3.1 Characteristics of the binding patch predict the ligand type . . . 71
4.3.2 Binding affinity predictions . . . 73
4.4 Discussion . . . 77

5 Comparison of immunoglobulin CDR3 repertoires. 81
5.1 Introduction . . . 81
5.2 Material and methods . . . 84
5.2.1 Dataset . . . 84
5.2.2 Characterization of the public and privates responses through repertoires comparison 84
5.2.3 Comparison of the earth mover distance and the Morisita-Horn distance . . . 88
5.2.4 Assessing how top clonotypes from a condition are shared between fish: TC search 88
5.2.5 Subsamples and their stability . . . 90
5.3 Results . . . 93
5.3.1 Characterization of the public and privates responses through repertoires comparison 93
5.3.2 Comparison of the earth mover distance and the Morisita-Horn distance . . . 95
5.3.3 Assessing how top clonotypes from a condition are shared between fish: TC search 98
5.3.4 Subsamples and their stability . . . 102
5.4 Discussion . . . 104

6 Software 107
6.1 Introduction . . . 107
6.2 Goals: Generating and evaluating predictive models for binding affinity . . . 107
6.3 General pre-requisites . . . 107
6.3.1 Binding affinity and dissociation free energy . . . 107
6.3.2 Key geometric constructions . . . 108
6.3.3 Associated variables . . . 109
6.3.4 Atoms' matching . . . 111
6.4 Using the programs to pre-process structural data . . . 111
6.4.1 Input: specifications and file types . . . 111
6.4.2 Output: specifications and file types . . . 112
6.5 Using the programs to select affinity prediction models . . . 113
6.5.1 Pre-requisites : Statistical Methods . . . 113
6.5.2 Input: Specifications and File Types . . . 115
6.5.3 Output: Specifications and File Types . . . 116
6.5.4 Model exploitation: predicting affinities . . . 116
6.6 Algorithms and Methods . . . 116

6.7 Programmer's Workflow 117
6.7.1 Pre-processing 117
6.7.2 Model selection 117
6.7.3 Model exploitation 117
6.8 External dependencies 117

7 Conclusion **119**
7.1 Discussion 119
7.2 Perspectives and future work 120

Appendix A Binding affinity models review **137**

Appendix B Protein - protein affinity prediction **143**

Appendix C Novel structural parameters of Ig - Ag complexes yield a quantitative description of interaction specificity and binding affinity **147**

Remerciements

Je souhaite tout d'abord remercier mes deux superviseurs, Pierre Boudinot et Frédéric Cazals, pour m'avoir fourni l'opportunité de réaliser cette thèse, pour leur disponibilité, ainsi que pour les discussions fructueuses que nous avons pu avoir tout au long de celle-ci. Le temps et l'énergie qu'ils ont engagé pour me faire partager leur savoir et leur expérience ont une grande valeur à mes yeux. Je souhaite également remercier les rapporteurs, M. Paul Bates et M. Alexandre Bonvin, pour leur relecture détaillée du manuscrit ainsi que leur commentaires constructifs. J'adresse également mes remerciements à Mme Véronique Braud et M. Dominique Housset pour leur participation en tant qu'examinateurs.

Je suis également reconnaissant à Marie-Paule Lefranc et Patrice Duroux (IMGT) qui ont partagé avec moi leur expertise des anticorps ansi que pour leur conseils sur l'utilisation de la base de données IMGT. Mes remerciements vont enfin à Sergei Grudinin (NANO-D) avec qui nous avons participé au challenge D3R, ainsi qu'à Luc Journeau (VIM INRA) qui s'est occupé de la gestion des données de séquençage.

Je suis également reconnaissant à l'INRA pour m'avoir octroyé la bourse Jeune Chercheurs qui a financé cette thèse ainsi, qu'à l'Inria pour les excellentes conditions de travail qui règnent au centre de Sophia-Antipolis.

J'ai également une pensée pour les collègues et doctorants de l'équipe – Deepesh, Alix, Romain, Dorian et Augustin – pour la chouette ambiance qu'ils ont fait régner durant ces trois années. J'ajoute un merci tout particulier à Tom qui aura su supporter les (nombreux) appels à l'aide lors de mes tentatives d'utilisation de notre bien-aimée librairie; j'ai nommé: la SBL.

Sur une note plus personnelle, je tiens également à remercier mes proches avec qui j'ai pu partager mes moments d'enthousiasme comme de doutes, et qui m'ont permis de m'aérer l'esprit quand les idées ou la motivation venaient à manquer.

List of Acronyms

ADCC: Antibody-Dependent Cell-mediated Cytotoxicity
AIC: Akaike's Information Criterion
ANSO: Average Normalized Shelling Order
APC: Antibody-Presenting Cell
BCR: B Cell Receptor
BSA: Buried surface area
CDC: Complement-Dependent Cytotoxicity
CDR: Complementarity-Determining Region
CLP: Common Lymphoid Progenitors
CLR: C-type Lectin Receptors
CRP: C-Reactive proteins
CSF: Colony-stimulating factor
CV: Cross-validation
DNA: Deoxyribonucleic acid
EEC: Enthalpy - Entropy compensation
EMD: Earth Mover Distance
FAB: Fragment Antigen-Binding
FR: Framework Region
GB: Generalized Born
HB: Hydrogen Bond
HSC: Hematopoietic Stem Cells
IFN: Interferon
IL: Interleukin
IPL: Internal Path Length
ITC: Isothermal Titration Calorimetry
IVWIPL: Inverse-Weighted Internal Path Length
KIR: Killer cell Inhibitory Receptors
LOO: Leave-one-out
LP: Linear Program
LRR: Leucine-Rich Repeats
MAC:Membrane Attack Complex
MARS: Multivariate Adaptive Regression Splines
MBL: Mannan-Binding Lectin
MCP: Monocyte Chemoattractant Protein
MHC: Major Histocompatibility Complex
MHD: Morisita-Horn Distance
MIP: Macrophage Inflammatory Protein
MLP: Multilineage Progenitor
NGS: Next generation Sequencing
NIS:Non-Interaction Surface
NK: Natural Killer (cells)
NLR: NOD-Like Receptors
NOD: Nucleotide-Binding Oligomerization Domain
PCC: Pearson's Correlation Coefficient
PDB: Protein Data Bank
PEL: Potential Energy Landscape
PMN: Polymorphonuclear Leukocytes
PRR: Pattern Recognition Receptors
RBF: Radial Basis Function
RIG: Retinoic acid-Inducible Gene
RLR: RIG-Like Receptor
RMSD: Root-Mean-Square Deviation
RMSE: Root-Mean-Square Error
RNA: RiboNucleic acid
RRS: Recombination Recognition Sequence
SAB: Structure Affinity Benchmark
SAM: Solvent Accessible Model
SAS: Solvent Accessible Surface
SASA: Solvent Accessible Surface Area
SBL: Structural Bioinformatics Library
SDRU: Specificity-Determining Residues Usage
SO: Shelling Order
SPR: Surface Plasmon Resonance
SVD: Sum of Volumes Differences
TC: Top Clonotype
TCR: T cell Receptor
TD: Thymus Dependent (Antigen)
TF: Tissue Factor
TGF: Transforming Growth Factor
TI: Thymus Independent (Antigen)
TLR: Toll-Like Receptors
TNF: Transforming Necrosis Factor
TRAF: TNF Receptor-Associated Factors
VC: Variable - Constant (pair)
VD: Volumes Differences
VEGF: Vascular Endothelial Growth Factor
VHSV: Viral Hemorrhagic Septicemia Virus

List of Figures

2.1 **Cell types involved in innate and acquired immunity.** Picture from [CS09, Chap. 2]. 7

2.2 **Illustration of some pattern recognition receptor families discussed in the text, with representative members.** LRR: leucine-rich repeats. Figure modified from [Pau13, Chap. 4]. 9

2.3 **Illustration of some cytokine and chemokine receptor families discussed in the text with representative members.** Figure from [CS09, Chap. 11]. 10

2.4 **Structure of a T cell receptor.** C: constant domain, V: variable domain. Complementarity determining regions (CDR) and framework regions (FR) colors follow the IMGT color scheme. Structure: entry IMGT-5CO7 from the IMGT/3Dstructure-DB. 13

2.5 **Ternary complex between the TCR, MHC and the antigenic peptide.** CDR and FR colors follow the IMGT color scheme. 14

2.6 **Interaction between a B cell and a T cell.** The antigen is processed by the B cell, and the resulting processed peptide is displayed on the MHC class II for recognition by the TCR. Interleukines secreted by the T cell control B cell differentiation. BCR: B cell receptor (Ig + Igα + Igβ); TCR: T cell receptor; MHC: major histocompatibility complex; IL(R): interleukine (receptor). Reproduced from en.wikipedia.org/wiki/CD154#/media/File:T-dependent_B_cell_activation.png (public domain). 15

2.7 **Schematic depiction of an Ig.** Colors for the V, D, J and C genes correspond to those in Figs 2.8, 2.9 and 2.10. Colors for the CDRs correspond to those in Fig. 2.13. A detailed view of the upper right portion of the figure is depicted in Fig. 2.14. 16

2.8 **Genomic locus of the κ light chain.** Adapted from [Pau13, Chap. 6]. 16

2.9 **Genomic locus of the λ light chain.**Adapted from [Pau13, Chap. 6]. 17

2.10 **Genomic locus of the heavy chain.** Adapted from [Pau13, Chap. 6]. 17

2.11 **Structure of an immunoglobulin (Ig).** 20

2.12 **Structure of the five Ig classes.** Figure from [CS09, Chap. 4] 21

2.13 **Structure of the Fab, with colored CDRs, and grey FRs.** 22

2.14 **Encoding of CDRs and FRs by the V, D and J genes.** The color scheme is the same as used in Fig. 2.7. 22

2.15 **Graphical representation of a V domain according to the IMGT numbering scheme.** Hatched circles correspond to positions which have been assigned a number but have no corresponding residue in this particular sequence. CDR anchors positions are displayed as squares. Conserved residues which always receive the same number are in (bright) red. VH CDR1, VH CDR2 and VH CDR3 are respectively colored (dark)red, orange and purple. Figure from the IMGT website `http://www.imgt.org/3Dstructure-DB/cgi/collier_perles.cgi?domcode=1AO7ED00&domdescr=V-BETA&domnum=1` 23

2.16 **Illustration of monovalent and multivalent Ags.** Figure from [CS09, Chap. 5]. . . 25

2.17 **Illustration of the phenomenon of enthalpy-entropy compensation.** $T\Delta S$ and ΔH are well correlated which makes their crossing point very sensitive to small changes in their slope. Moreover, the scale of changes for ΔG is much smaller (one order of magnitude) than that of the changes in $T\Delta S$ and ΔH. Figure courtesy of Alan Cooper. 28

2.18 **Molecular surface and solvent accessible model (SAM) of an atom and a molecule.** The surfaces on the left correspond to the van der Walls (vdW) radius of the atoms, and those on the right to the solvent accessible surface (SAS). The vdW representation of the atom was superimposed on the SAS representation in the upper right portion of the picture (darker, dotted circles). The yellow circles corresponds to solvent molecules as they rolls along the vdW surface to define the SAS. The yellow edges on the right molecule correspond to the boundaries of spherical polygons dividing the SAS. Structure: chain A from Protein Data Bank entry 3OJ3. 29

2.19 **Definition of the buried surface area (BSA).** Subtracting the solvent accessible surface from the complex (green) to that of the individual partners (blue) results in the buried surface area (red). 30

2.20 **Voronoi cell of a point surrounded by eight spheres.** Blue edges are the edges of the Voronoi cell and green circle arcs are the intersections of the balls. Figure from [CKL11] 30

2.21 **Voronoi interface between an Ig and a peptide.** Structure: entry 1GGI from the protein data bank. 31

2.22 **A one dimensional potential energy landscape (PEL)**. Local minima are represented by red dots with basins associated to their metastable states in green and purple respectively. The potential energy difference ΔE is distinct from ΔG because it only takes into account the local minima and not the rest of the basins. 32

2.23 **General workflow for building a statistical model of the binding affinity.** Input is boxed in green, output is boxed in purple. 33

3.1 **Resolution of the structures in the SAB.** The histogram and green kernel density estimation curve are for the whole SAB, the red curve is for the complexes and the blue curve is for the unbound partners. For the whole SAB: Minimum = 0.93 Å, median: = 2.13 Å, average = 2.19 Å, max = 3.5 Å. NB: the high resolution dataset SAB-A-HR retains only entries whose resolution is better than 2.5 Å for both the complex and the individual partners [ERS14]. 43

3.2 **The various datasets defined from the structure affinity benchmark (SAB), based on iRMSD between the unbound and bound structures.** 43

3.3 **Structural parameters used in this work. (A)** Labeling the atoms, illustration on a fictitious 2D complex. The binding patch on each partner consists of one layer of atoms ($\mathcal{I}$, colored solid balls), as identified by a Voronoi interface model [CPBJ06, LC10]. The non interface atoms ($\mathcal{I}^c$) are split into those which retain solvent accessibility (SASA > 0, dashed balls), and those which do not (SASA $= 0$, dotted balls) **(B)** Each interface atom is assigned an integer, its shelling order, equal to the smallest number of atoms traveled to reach an exposed non interface atom, *i.e.* an atom belonging to $\mathcal{I}^c$ and with SASA > 0 (in grey) [BGNC09]. **(C,D)** The volume of an atom is defined as the volume of the intersection between its ball in the solvent accessible model, and its Voronoi cell [CKL11], a quantity well defined even if the atom retains solvent accessibility. The packing of this atom is the inverse of this volume. Practically, interfaces and binding patches are computed with `Vorshell`[LC10], while atomic surface areas and volumes are computed with `Vorlume`[CKL11]. Both programs are available from the Structural Bioinformatics Library (SBL), see `http://sbl.inria.fr`. 44

3.4 **Running binding affinity predictions for a dataset $\mathcal{D}$ *i.e.* a subset of the structure affinity benchmark: graphical outline of the statistical methodology. (Templates)** From the pool of variables, templates are generated. **(Cross-validation)** Each template undergoes a number N_{XV} of repetitions of 5-fold cross-validation, yielding one binding affinity prediction per complex for each repetition. **(Statistics)** Various statistics are computed to assess the performances yielded by the predictive model associated to each template. **(Model selection)** Predictive models are compared, and the best ones selected. 47

3.6 **Comparison between two ways of computing the correlation for a given predictive model over multiple repetitions. Median of correlations**: for each of the N_{XV} repeats, compute the correlation between the predictions and experimental affinities. Take the median of these predictions for each complex. **Correlation of median of predictions**: compute a single prediction per complex as the median of all N_{XV} predictions. Compute the correlation between those predictions and the experimental affinities. The values of all predictive models tested on all datasets have been aggregated on this figure. The correlation between both methods is 0.997 with a median absolute difference of 0.005. Moreover, both measures are maximal for the same predictive model on all datasets . . . 49

3.5 **Predictive model complexity versus median correlation C_V between predicted and experimental values.** 51

3.7 **The hardness of predicting a binding affinity does not correlate with the flexibility of the complex.** x-axis: flexibility of the interface, expressed in terms of interface iRMSD; y-axis: median prediction error $e_i[T_l, \mathcal{D}]$ (Eq. (3.20)). 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. 56

3.8 57

4.1 **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. 63

4.2 **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). 66

4.3 **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 Fig. 4.2 is partitioned into contributions from these 14 regions. . 67

4.4 **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). 71

4.5 **Interaction specificity for Ig - Ag complexes: analysis and predictions.** Both analyses are based upon the average buried surface areas per atom (Equations (4.1) (4.2)): $\overline{\text{bsa}}_{\text{Ig}}$ 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). 72

4.6 **Classification rules characterizing the binding patch depending on the ligand types.** 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. 72

4.7 **Binding affinity analysis and predictions for Ig - Ag complexes. (4.7a) 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. **(4.7b) 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. **(4.7c) 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. 73

4.8 **Prediction error versus average distance of the 10 nearest-neighbors and the standard deviation of their affinity values.** 74

4.9 **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. 75

4.10 Buried Surface Area (A^2) of the VH and VL domains, and their respective CDR. 75

4.11 **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.** 76

4.12 **IVW-IPL of the CDR of VH (left panel) and VL (right panel).** 77

4.13 **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. 77

5.1 **Clonotype size distribution for all variable - constant gene pairs, and all fish.** Left: whole distribution, right: zoom on smaller clonotypes sizes. Blue line: clonotype size of 50. 86

5.2 **Probability for a clonotype of size m from a population of size n to be found at least once in a subsample of $s = 7000$ sequences.** 90

5.3 **IgM. (a,d,g) Distances distribution for fish within a condition (left) and between conditions (right)**. y-axis: earth mover distance distance; C: naive, E1: vaccinated, E4: vaccinated + infected. Each point represents a distance between two fish (jittered x-axis). **(b,e,h) Dendrogram built using Ward's method on the earth mover distance distance matrix. (c,f,i) Earth mover distance matrix.** Heatmap colors range from white (high values) to red (low values). 94

5.4 **IgT. (a,d,g) Distances distribution for fish within a condition (left) and between conditions (right)**. y-axis: earth mover distance distance; C: naive, E1: vaccinated, E4: vaccinated + infected. Each point represents a distance between two fish (jittered x-axis). **(b,e,h) Dendrogram built using Ward's method on the earth mover distance distance matrix. (c,f,i) Earth mover distance matrix.** Heatmap colors range from white (high values) to red (low values). 95

5.5 **IgM. (a,d,g) Distances distribution for fish within a condition (left) and between conditions (right)**. y-axis: Morisita-Horn distance; C: naive, E1: vaccinated, E4: vaccinated + infected. Each point represents a distance between two fish (jittered x-axis). **(b,e,h) Dendrogram built using Ward's method on the Morisita-Horn distance distance matrix. (c,f,i) Morisita-Horn distance matrix.** Heatmap colors range from white (high values) to red (low values). 97

5.6 **IgT. (a,d,g) Distances distribution for fish within a condition (left) and between conditions (right)**. y-axis: Morisita-Horn distance; C: naive, E1: vaccinated, E4: vaccinated + infected. Each point represents a distance between two fish (jittered x-axis). **(b,e,h) Dendrogram built using Ward's method on the Morisita-Horn distance distance matrix. (c,f,i) Morisita-Horn distance matrix.** Heatmap colors range from white (high values) to red (low values). 98

5.7 **Quantification of repertoire overlap using top clonotypes.** Color code: blue = naive (C), red = vaccinated (E1), green = infected (E4). Each line corresponds to a variable - constant gene pair. **(A) Number of top clonotypes with a given top clonotype count.** Each column corresponds to the condition of top clonotype sets. Abscissa: values correspond to top clonotypes counts (see Def. 3 with $\gamma = 0$). Ordinates: each bar corresponds to the average (over subsamplings) number of top clonotypes with a given clonotype count. The standard deviation is displayed with an error bar. **(B) Venn diagrams: number of top clonotypes shared between the top clonotype sets.** . 100

5.8 **Quantification of repertoire overlap using top clonotypes.** Color code: blue = naive (C), red = vaccinated (E1), green = infected (E4). Each line corresponds to a variable - constant gene pair. **(A) Number of top clonotypes with a given top clonotype count.** Each column corresponds to the condition of top clonotype sets. Abscissa: values correspond to top clonotypes counts (see Def. 3 with $\gamma = 2$). Ordinates: each bar corresponds to the average (over subsamplings) number of top clonotypes with a given clonotype count. The standard deviation is displayed with an error bar. **(B) Top clonotypes overlap table.** A cell in row x, column y with $x \neq y$ contains the number of clonotypes shared by the top clonotypes sets x and y. For $x = y$, it contains the number of clonotype shared by x and the two other clonotype sets. See Table 5.4. 101

5.9 **Estimated density of the distribution of the entropy H of the powerset vector (Def. 5) for IgM.** Left: VH4.1-Cμ, middle: VH5.1-Cμ, right: VH8.1-Cμ. Blue: condition C, Red, condition E1, green: condition E4. Ticks at the bottom show the maximum entropy for combinations of at least k fish, *i.e.* frequencies of occurence of (0.5, 0.5) for $k = 2$, (1/3, 1/3, 1/3) for $k = 3$ and so on. 103

5.10 **Estimated density of the distribution of the entropy H of the powerset vector (Def. 5) for IgT.** Left: VH4.1-Cτ, middle: VH5.4-Cτ, right: VH9.2-Cτ. Blue: condition C, Red, condition E1, green: condition E4. Ticks at the bottom show the maximum entropy for combinations of at least k fish, *i.e.* frequencies of occurence of (0.5, 0.5) for $k = 2$, (1/3, 1/3, 1/3) for $k = 3$ and so on. 103

5.11 **Entropy H of the powerset vector (Defs. 4 and 5) versus z, the number of fish combinations in which a top clonotype has been found.** The solid line shows the maximum entropy for a given z which is $-\ln(1/z)$. 104

C.1 **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. . . . 149

C.2 **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. 150

List of Tables

2.1 **Residue-based IMGT numbering for CDRs and FRs.** Table from `http://www.imgt.org/IMGTScientificChart/Nomenclature/IMGT-FRCDRdefinition.html` 23

2.2 **Comparison of recent works on general protein - protein binding affinity prediction** NB: all these results have been obtained on various subsets of the SAB or various datasets altogether. They also use various cross-validation procedures, which makes direct comparisons difficult. The terms used in the rightmost are defined as follows: PCC: Pearson's correlation coefficient. q^2: cross-validated R^2; q: cross-validated correlation coefficient; r^2_{pred}: correlation coefficient on an external test set; RMSE: root mean-squared error. ^{a}: One predictor uses all features but weights data points instead virtually using all variables. The union of the sets of variables used by the other predictors has 94 members. ^{b}: Linear regression optimized for correlation (*i.e.* not least squares). *: Correlation coefficient obtained on a mix of training and test sets. $^+$: q obtained on a reduced dataset consisting of high-resolution structures. 37

3.1 **Parameters used to estimate binding affinities.** Atomic level parameters: IVW-IPL, SVD_SO1, SVD_SOGT1, SVD_NI_B, SVD_NI_E, ATOM_SOLV, POLAR_SASA; Residue level parameters: NIS^{polar}, $\text{NIS}^{charged}$, ΔNIS^{polar}, $\Delta\text{NIS}^{charged}$; Interface level parameter: iRMSD. The acronyms read as follows (see text for details): **S**um of **V**olume **D**ifferences; **S**helling **O**rder; **I**nverse **V**olume **W**eighted; **I**nternal **P**ath **L**ength; **N**on **I**nteracting **B**uried/**E**xposed; **N**on **I**nteracting **S**urface; **S**olvent **A**ccessible **S**urface **A**rea; 45

3.2 **Binding affinities: correlations between predictions and measurements for all datasets and all specific predictive models. (Whole table)** Red values show the best results for a given category in the corresponding section, and cross-validated and classical correlation coefficients are treated separately in the second part of the table. **Bold** values in the second part of the table show the categories on which the variables selection was performed for a given predictive model. **(First section)** Previous work: values published and our replica (rep). Green cells correspond to the correlation coefficient of the predictive model over the train set (*i.e.* $\sqrt{R^2}$ from the linear regression). Purple cells correspond to either cross-validation results, prediction on a test set distinct from the trains set, or both. For the other cell colors, see details in Section 2.4.6. Yellow cells: discrepancies between original values and our replicas – we did not remove any complex. Orange cells: the cross-validation procedure made some test data information leak into the training set. Cyan cells: overlapping training and test sets. **(Second section)** Eight predictive models developed in this work. The value reported in a cell corresponds to the correlation between predictions and experimental values. For the $N_{XV} = 10000$ 5-fold cross-validation, the median of the predictions was used (Section 3.2.4). Purple lines show the cross-validated correlation coefficient, while green lines show the classical correlation coefficient (square root of the coefficient of determination). 54

3.3 **Datasets and their specific predictive models: performances in estimating the dissociation free energy ΔG_d.** Each predictive model (rows) was tested on each dataset (columns). A cell in the Table features the values of the affinity prediction ratio $p_{1.4}^{\text{error}}$, $p_{2.8}^{\text{error}}$ and $p_{4.2}^{\text{error}}$ respectively, see Eq. (3.23). For instance, Predictive Model 1, when evaluated on dataset SAB-A (139 complexes) predicted 47.48%, 78.42% and 92.09% of the complexes with a median absolute error below 1.4, 2.8 and 4.2 kcal/mol, respectively. Equivalently, these are the fractions of cases such that K_d is estimated within one, two and three orders of magnitude. **(Top part)** Previous work. Lines marked with Rep. (replica) were obtained using the values of the parameters provided in the SAB for [Jan14] and those provided by the authors (personal communication) for [KRF$^+$14], along with their respective protocols. Lines not marked with Rep. were obtained using the variables of the original models, within our setup. **(Bottom part)** Our predictive models. Bold values indicate when a predictive model was tested on its specific dataset. 55

3.4 **Validation of the models on an external test set.** The external test set from [KRF$^+$14, supplemental] was split using the same criteria as those used to define datasets from the structure affinity benchmark, yielding *external datasets*. Each linear model was trained using a specific template on the whole corresponding dataset and used to predict the corresponding external datasets. The first part of the table displays the external dataset size, Pearson's correlation coefficients and p-value for each predictive on its external dataset along with $p_{1.4}^{\text{error}}$, $p_{2.8}^{\text{error}}$ and $p_{4.2}^{\text{error}}$. The second part show the values from table the diagonal of tables 3.3 and 3.2 for comparison. 56

3.5 **Pearson correlation coefficients between the individual variables.** 58

3.6 **Parameters used by the best predictive model for a given dataset.** A *dataset* is a subset of the structure affinity benchmark. A *specific* predictive model is the predictive model which performed significantly better than all the others for a given dataset during model selection. The parameters are those from Table 3.1. Black dots mark variables used by statistically significant predictive models and white dots those used by other predictive models. The last column counts the number of statistically significant predictive models using a given parameter. Asterisks identify atomic level parameters. 58

4.1 **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)**. 62

4.2 **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 [LPR$^+$03]. 62

4.3 **Median BSA and median of BSA/BSA$_{\text{Ig}}$ 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. 70

4.4 **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. 72

5.1 **Dataset statistics.** Seq: Number of sequences resulting from sequencing Clono: corresponding number of clonotypes (distinct CDR3 amino-acid sequences). 85

5.2 **Amino-acid classes and conservative substitutions.** Amino acid in the same class define conservative substitutions. 86

5.3 **Amino-acid substitution matrix.** Conservative substitutions get a score of 0, and non-conservative ones get a score of -1. Conservative substitutions are defined in table 5.2. 87

5.4 Pattern of the overlap table containing the number of shared top clonotypes (in Fig. 5.8). 89

5.5 **Clonotype representation during subsampling.** Rows: minimum, first quartile, median, average, third quartile and maximum. n: number of sequences resulting from the sequencing, m^*: minimum size of a clonotype required for it to be found in a subsample of size $s = 10000$ from a population of size n with probability 0.99. 91

6.1 **Input files for the first step described in section 6.4.1.** . . . 112
6.2 **Output files for the step 2 described in section 6.4.1** . . . 113
6.3 **Input files for the third step described in section 6.5.2.** . . . 116
6.4 **Output files for the step 3 described in section 6.5.2** . . . 116

B.1 **Experimental affinities on a per complex basis: experimental measurements (ΔG_d) versus predictions ($\hat{g}_i$, Eq. 3.16). Predictions were generated with predictive Model 1 on dataset SAB-A using linear regression.** The median was taken over the NXV repetitions. Blue values indicate under-predicted complexes (63) and red indicate the over-predicted ones (76). A start denotes complexes with error in the top decile. 144
B.2 **Experimental affinities on a per complex basis: experimental measurements (ΔG_d) versus predictions ($\hat{g}_i$, Eq. 3.16). Predictions were generated with predictive Model 1 on dataset SAB-A with k-nearest neighbors regression.** The median was taken over the NXV repetitions. Blue values indicate under-predicted complexes (66) and red indicate the over-predicted ones (71). A start denotes complexes with error in the top decile. . . . 145
B.3 **Pearson correlation coefficients between the individual variables and the affinity.** 146
B.4 **Validation of the best overall model *i.e.* model 9 on an external test set.** See Table 3.4 for the statistics presented. . . . 146

C.1 **Main features of the Ig - Ag complexes found in the structure affinity benchmark.** {H,L}CDR len: length of the CDRs in residues. Numbers in the V{H,L} CDR 1,2,3 columns correspond to the first and last residue numbers in IMGT renumbered PDB files. . . . 148

Chapter 1

Introduction (Version Française)

1.1 Les anticorps au sein du système immunitaire.

1.1.1 Le système immunitaire

Le système immunitaire des vertébrés possède deux composantes: la composante innée et la composante adaptative. La première est formée d'éléments encodés tel quels dans le génome. Elle tient lieu de première ligne de défense et répond de façon rapide et générique à une large classe de pathogènes. Cependant, elle ne confère pas d'immunité à long terme. Le système immunitaire adaptatif en revanche, repose sur des gènes qui sont modifiés lors de la différentiation cellulaire, créant ainsi une grande diversité. La diversité des récepteurs ainsi créés donne lieu à des réponses spécifiques ainsi qu'à une mémoire immunitaire à l'origine de réponses secondaires rapides et efficaces lors de rencontres ultérieures avec le même pathogène.

La réponse innée est la première à entrer en jeu lors d'une infection par un pathogène et n'est pas spécifique d'un antigène (Ag). Elle est principalement impliquée dans les fonctions suivantes: recrutement des cellules immunitaires sur le site de l'infection, activation de mécanismes non spécifiques tels que la phagocytose, activation du complément, et activation du système immunitaire adaptatif.

Le système immunitaire adaptatif entre en jeu après la réponse innée dans le cas où elle n'aurait pas suffit à éliminer le pathogène. La réponse adaptative est strictement régulée par l'inflammation et la réponse innée. Elle est de plus caractérisée par la génération de cellules spécifiques de l'antigène, des cellules mémoires, et la sécrétion d'immunoglobulines. Deux principaux types cellulaires sont impliqués dans cette réponse: les lymphocytes B et T (ou cellules B et T) qui, contrairement, aux cellules impliquées dans la réponse innée expriment des récepteurs spécifiques de l'antigène générés de façon somatique.

Les lymphocytes B. Les lymphocytes B sont les seules cellules capable de synthétiser des immunoglobulines (Igs) et, selon leur stade de développement, les sécrètent ou les exposent à leur surface.

Le nombre de pathogènes qu'un individu est susceptible de rencontrer durant son existence est gigantesque. Pour cette raison, la reconnaissance spécifique des antigènes nécessite une diversité de récepteurs comparable. Le nombre de gènes d'Igs dans le génome d'un individu étant limité, le mode de production d'Igs aussi diverses que les antigènes potentiels à été une question majeure pour les immunologistes. La recombinaison des gènes d'Ig est à l'origine de cette diversité et implique que chaque clone de lymphocytes B exprime un récepteur différent.

Les Lymphocytes T. Les lymphocytes T exposent de récepteurs membranaires (TCR) qui, contrairement au Igs se lient uniquement à des antigènes exposés en combinaison avec le MHC.

La réponse primaire. La réponse primaire se déclenche après une première exposition à une antigène donné. Une première période de temps durant laquelle les cellules B et T contactent l'antigène, s'activent, communiquent et se différencient est suivie par une plus forte sécrétion d'Igs. Celle-ci atteint un plateau avant de de décliner lorsque l'infection a été éliminée.

Durant cette première phase l'interaction entre cellules B et T est essentielle pour l'activation et la différenciation des premières.

La réponse secondaire. La réponse secondaire à lieu lors de l'exposition d'un individu à un antigènes déjà rencontré auparavant. En raison de la mémoire immunologique due aux cellules B et T, la phase de latence avant activation est plus courte que lors de la réponse primaire. Une réponse secondaire peut avoir lieu des années après une première rencontre avec l'antigène.

L'importance des Igs: pourquoi les étudier?

Bien que de nombreuses protéines, et en particulier de nombreux récepteurs aient un importance centrale dans la réponse immunitaire, cette thèse se concentre sur les Igs pour les raisons suivantes.

Premièrement, les anticorps sont capable de se lier à une grande diversité de molécules: protéines, peptides, haptènes, sucres, lipides ou encore acides nucléiques. Une telle versatilité parant d'une diversité structurale réduite est extrêmement intéressante en terme de mécanismes de liaison. Deuxièmement, contrairement aux interactions faites par le TCR, celles des Igs ont lieu avec la forme native de l'antigène, et la pression de sélection s'applique au site de liaison dans son ensemble. Ceci permet un étude plus directe des mécanismes impliqués dans l'optimisation de l'affinité, en particulier dans le contexte du design de protéines de liaison. Enfin, les anticorps sont des outils important pour la biologie expérimentale (purification, fluorescence, assays) ainsi que des molécules thérapeutiques prometteuses.

1.1.2 Génétique moléculaire des Ig: diversité et répertoires

Les antigènes potentiels pouvant être rencontrés par le système immunitaire ne pouvant pas être prédits à l'avance, un grand nombre d'Igs au spécificités variées est nécessaire pour permettre une réponse spécifique à tous les antigènes possibles. Informellement, l'ensemble des séquences d'Igs pouvant être générées par un individu est nommé *répertoire*. Le nombre théorique d'Igs pouvant chez un individu est estimé entre 10^{15} et 10^{18} alors que le génome des mammifères contient de l'ordre de 10^4 à 10^5 gènes. Des mécanismes de diversification sont donc nécessaires pour expliquer la diversité du répertoire d'Igs.

Génétique moléculaire de la diversification des gènes d'Igs

Une Ig est formée de deux chaînes lourdes (H) et deux chaînes légères (L) dont le domaine variable (V) est le résultat d'un processus de recombinaison. Chaque domaine VL est codé par deux gènes V et J; et chaque domaine VH est codé par trois gènes V, D et J.

Recombinaison. La plupart des cellules du corps humain possèdent le même génome et se différencient en exprimant un sous-ensemble de leur gènes. La situation est similaire chez les cellules B matures, à part aux loci des gènes d'Igs, où elles suppriment ou déplacent les gènes V, D et J qu'elles ont hérité (lignée terminale), scellant ainsi la spécificité des Igs qu'elle synthétiseront.

Diversité jonctionnelle. Durant le processus de recombinaison, le raccordement des gènes V, D et J est imprécis, et un nombre variable de nucléotides peut-être supprimé. De plus des nucléotides peuvent êtres ajoutes aléatoirement à la jonction.

Hypermutation somatique et maturation d'affinité. L'hypermutation somatique à lieu durant l'expansion clonale des cellules B et consiste en une augmentation (jusqu'à un facteur 10000) des mutations, ciblées sur les gènes V, D et J. Ceci résulte en une population de cellules B aux affinités variées. Ce processus est suivi par une intense sélection des cellules B possédant une forte affinité pour l'antigène. Ceci a pour résultat une population de cellules B dont les Igs possèdent une forte concentration de mutations dans les régions en contact avec l'antigène.

1.1.3 Structure et fonction des immunoglobulines

Les immunoglobulines ou Igs sont des protéines composées de deux chaînes lourdes (H) et deux chaînes légères (L). Les deux chaînes lourdes et légères sont respectivement identiques au sein d'une même Ig. Chaque chaîne est composée d'un domaine variable (V) et d'un ou plusieurs domaines constants (C). Le *Fab* une la portion d'une Ig correspondant à au domaines V et au premier domaine C. Le *Fc* est la partie restante, composée de domaines C. Au sein du Fab, le fragment variable *Fv* est formé des domaines V et le fragment constant *Fc* est formé du premier domaine C.

Le domaine constant. Le domaine constant d'une Ig définit sa *classe*. La notion correspondante au niveau génétique est nommée *isotype*. Plusieurs Igs spécifiques d'un même antigène peuvent donc faire partie de différentes classes. Il existe 5 classes: A, D, G, E et M. Un changement de classe (partant toujours d'une classe M, la première à être synthétisée) est nommé *switch*.

Le domaine variable. Le domaine variable d'une chaîne d'Ig contient trois régions hypervariables nommées *CDR*s (régions déterminant la complémentarité) qui correspondent approximativement au site de liaison avec l'antigène.

En raison de cette variabilité, il est nécessaire de maintenir une structure suffisamment stable pour supporter n'importe quel combinaison de CDR, ce qui est possible grâce aux *FR*s (régions framework). Il existe quatre FR situées entre les CDRs au niveau de la séquence. Les CDRs 1, 2 et les FR 1, 2 et 3 sont codés par le gènes V tandis que le CDR3 est codé par les gènes V, D et J chez la chaîne H et V et J chez la chaîne L. Le FR4 est lui codé par le gène J. Pour ces raisons, ainsi que la jonction imprécise des gènes V, D et J, le CDR3 de la chaîne lourde est le principal site de diversité du domaine V.

1.1.4 Reconnaissance entre immunoglobulines et antigènes

La reconnaissance des antigènes par les Igs est centrale au succès de la réponse immunitaire adaptative. De plus, les Igs sont communément utilisées en tant que protéines de liaisons lors d'expériences telles que la purification de composés. Pour cette raison, les complexes Ig - Ag ont été l'objet de nombreuses études et une part non-négligeable d'entre elles s'est concentrée sur le site de liaison.

Propriétés physico-chimiques des sites de liaisons des Igs La large diversité d'Igs existantes a naturellement mené à l'analyse de la composition en acides-aminés du site de liaison, puis à celle des acides aminés contactant l'antigène lorsque suffisamment de données structurales furent obtenues.

En général les site de liaison est préférentiellement formé de résidus légèrement hydrophiles et aromatiques, en particulier Tyr et Trp. Excepté Arg, les résidus chargés sont en revanche sous-représentés.

Les interfaces de complexes Ig - Ag ont également un complémentarité physico-chimique plus élevée que celle de complexes protéine - protéine généraux.

1.2 Modélisation des complexes Ig - Ag

1.2.1 Notions clés de chimie physique

Lorsque deux molécules, nommées *partenaires* s'assemblent, formant ainsi un complexe, la force de cette association est nommée *affinité* ou *énergie libre de liaison* du complexe.

En raison de l'omniprésence des interactions entre molécules au sein de la plupart des processus biologiques, affinité de liaison est une notion centrale en biologie structurale. En effet, la stabilité d'un complexe est souvent critique pour que sa fonction soit menée à bien, un exemple du paradigme “structure - dynamique - fonction”.

L'estimation de l'affinité de liaison est donc centrale afin de comprendre comment les systèmes biologiques régulent la force avec laquelle différentes molécules interagissent. Être capable de faire des prédictions précises et fiables serait donc un pas majeur dans la direction de l'analyse d'interactomes. La prédiction d'affinité peut, de plus, être appliquée au design de protéines, au docking, ou encore à la

découverte de ligand (ligand discovery), avec application au design de médicament tels que les peptides thérapeutiques.

Définition formelle L'affinité de liaison est une quantité appartenant au domaine de la chimie et de la thermodynamique, et peut en tant que telle être quantifiée expérimentalement.

Elle peut être exprimée sous deux formes. Considérons deux partenaires A et B formant un complexe AB. La première forme dépend explicitement du ration K_d des concentration entre les partenaires isolés d'une part ([A] et [B]) et le complexe d'autre part ([AB]):

$$\Delta G = -RT \ln K_d/c^\circ = -RT \ln \frac{[A][B]}{[AB]} \tag{1.1}$$

R est la constante des gaz parfaits, T est la température absolue et c° est la concentration standard (1M). La seconde forme dépend de la variation de deux quantités thermodynamiques entre la forme liées et la forme non liée: la variation d'entropie et d'enthalpie.

$$\Delta G = \Delta H - T\Delta S. \tag{1.2}$$

ΔH est la variation d'enthalpie et ΔS est la variation d'entropie.

Détermination expérimentale De multiple méthodes permettant de mesurer l'affinité de liaison de deux partenaires ont été conçues telles que la calorimetrie par titration isothermale (ITC), la résonance plasmonique de surface (SPR) ou encore des méthodes basées sur la fluorescence.

1.2.2 Estimation de l'affinité à partir de données structurales

Deux principales familles de méthodes existent pour prédire l'affinité de liaison de deux partenaires.

La première, comprenant les méthodes d'échantillonnage, utilise des approximations mathématiques de phénomènes physiques dans le but de modéliser l'énergie potentielle d'un système. Ceci permet de construire un *paysage d'énergie potentielle* (PEL) assignant une valeur d'énergie potentielle à chaque état du système (par exemple une conformation des partenaires ainsi que des molécules de solvant environnantes). L'estimation de l'énergie libre de liaison peut ensuite être calculée en intégrant la fonction d'énergie définie par ce PEL. En pratique, cette tâche est cependant irréalisable en raison de la très grande dimension de cette fonction, et de nombreux travaux se basent sur l'échantillonnage du PEL afin d'obtenir une estimation.

La seconde famille, les modèles statistiques, se base sur des descripteurs de résolution plus grossière (pas nécessairement à l'échelle atomique) décrivant différent aspects de la liaison, et utilisent des données expérimentales afin d'ajuster les paramètres du modèle de façon à minimiser l'erreur entre l'estimation et la valeur réelle de l'affinité.

1.3 Contributions

Cette thèse étudie trois sujets relevant de la biologie structurale, de la génétique et de l'immunologie.

Premièrement, nous développons de nouveaux prédicteurs de l'affinité de liaison de complexes protéiques, produisant des résultats de niveau "état de l'art". La première étape réside dans le calcul de 12 variables modélisant la géométrie et les propriétés physico-chimiques des complexes. La seconde consiste à générer et évaluer des prédicteurs utilisant des sous ensembles de ces variables, de façon à identifier les plus performants. Le logiciel associé est distribué dans la Structural Bioinformatics Library.

Deuxièmement, nous proposons de nouvelles analyses de complexes Ig - Ag. D'une part nous concevons un classificateur distinguant les types de ligand des Ig. D'autre part, nous montrons que le modèle précédent prédit fidèlement l'affinité de complexes Ig - Ag. Enfin, nous quantifions la contribution des CDR3 de la chaine lourde à l'affinité de liaison, et montrons qu'il contribue significativement plus que les autres CDR.

Enfin, nous nous intéressons à la modélisation de la diversité des répertoires de chaîne lourde des Igs, à partir de données de séquençage de CDR3 issus des transcrits d'Igs, dans un modèle de vaccin chez le

poisson. Nous comparons les répertoires de deux individus en utilisant la “earth-mover distance”. En exploitant la correspondance entre clonotypes de deux répertoires, nous montrons qu'EMD révèle des informations inaccessibles aux méthodes basées sur les indices de diversité. Pour caractériser la notion de réponse immunitaire publique / privée, nous quantifions le chevauchement des clonotypes exprimés entre individus de la même ou de différentes conditions.

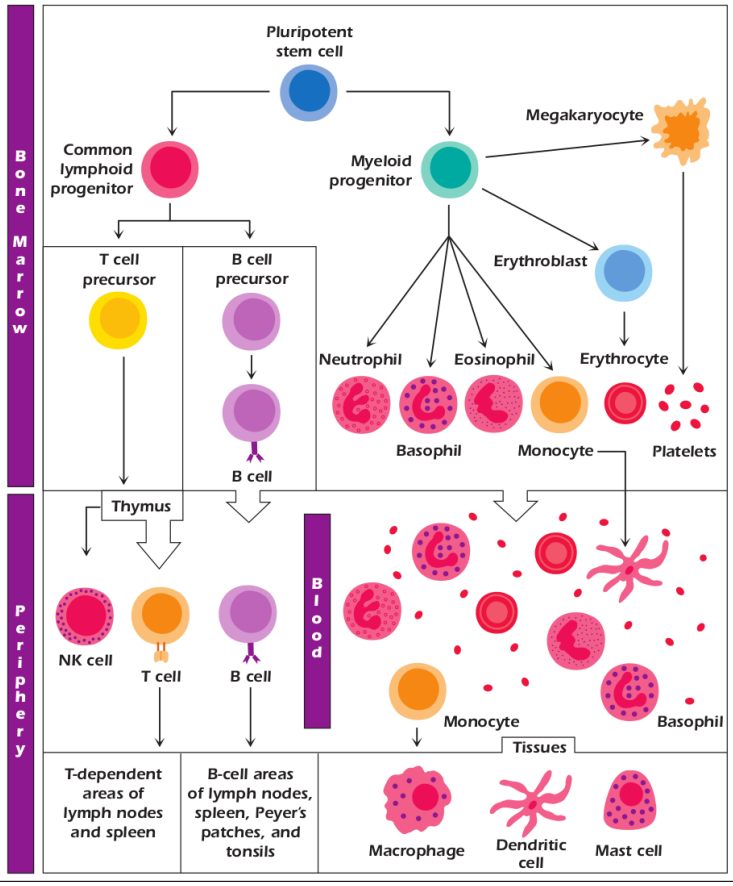
Chapter 2

Introduction (English Version)

2.1 The immune system

The immune system of vertebrates consists of two components: the innate and the adaptive components. The former is made of elements encoded as such in the genome, acts as a first line defense and responds quickly and generically to a broad class of pathogens but does not confer long-lasting immunity. The adaptive immune system on the other hand, is based on genes which are modified during the cell differentiation, creating a large diversity. This diversity of receptors allows specific responses and memory, hence quick and effective secondary responses during subsequent encounters with the same pathogen. The immunological memory is the basis of vaccination.

Figure 2.1 Cell types involved in innate and acquired immunity. Picture from [CS09, Chap. 2].



2.1.1 The innate immune system

The innate immune response is the first response to occur upon infection by pathogens and is not antigen-specific. Its main functions are: recruitment of immune cells to the site of infection, activation of nonspecific mechanisms such as phagocytosis, activation of the complement system, and activation of the adaptive immune system. We now describe the main cell types and molecules involved in the innate immune system.

Key components at the cellular level

The key cell types involved in innate immunity can be categorized as *polymorphonuclear leukocytes* (PMN, also called *granulocytes*), *macrophages*, *dendritic cells*, *natural killer cells*, and *mast cells* (Fig. 2.1).

Granulocytes [Pau13, Chap. 20]. Granulocytes are relatively short-lived cells with phagocytic activity and are the first cells to engage in the immune response. They can be further divided in three cell types shortly described hereafter.

Neutrophils, also called highly phagocytic polymorphonuclear cells, participate in both innate and acquired immunity, although, they have no specific antigen recognition abilities. They play a key role in the early stages of the inflammatory response during which they migrate from blood to the tissue where infection occurs. Neutrophils express a range of receptors among which Fc receptors used to bind *opsonized* (*i.e.* immunoglobulin-coated) pathogens.

Eosinophils are cytotoxic granulocytes which mainly play a role during allergic reactions. They bear the same Fc and complement receptors as neutrophils.

Basophils also play a role in allergic reactions and, along with mast cells, are the main responsible for the symptoms by releasing inflammatory mediators such as histamine. As neutrophils and eosinophils, they display Fc receptors.

Mast cells [Pau13, Chap. 20]. Along with basophils, mast cells are the main effectors during allergic reactions, and are causing symptoms through the release of inflammatory mediators. Their functions are similar to those of basophils, but they are tissue-based and not circulating cells. They participate in both innate and adaptive immunity but have no specific antigen recognition abilities.

Natural killer cells [Pau13, Chap. 17]. Natural killer cells are cytotoxic cells involved in the early phase of the immune response. They bear no antigen-specific receptors but instead induce lysis or apoptosis of cells which do not display MHC class I. Along with macrophages, they are one of the main effectors of cell-mediated immunity. They display killer cell inhibitory receptors (KIR) which bind to MHC class I.

Macrophages [Pau13, Chap. 19]. Macrophages are part of the *antigen presenting cells* (APC), which have the ability to process antigens and present them to antigen-specific T cells (section 2.1.2) although they have themselves no antigen-specific properties. They also play a role during *antibody-dependent cell-mediated cytotoxicity* (ADCC), during they become highly cytotoxic upon activation by *interferon* γ (IFN-γ) released by natural killer cells. They are one of the main effectors of cell-mediated immunity and engage in the response after granulocytes. They display *major histocompatibility complex* (MHC) class I and II receptors for antigen presentation to T cells, and Fc receptors for facilitated phagocytosis of opsonized pathogens and activation during ADCC.

Dendritic cells [Pau13, Chap. 16]. Like macrophages, dendritic cells are antigen presenting cells (APC). As such, they do not display antigen-specific receptors but can phagocyte, process, and present antigens to T cells displaying such receptors. They display MHC class I and II receptors and can secrete large amounts of α and β interferon in response to viruses. They are the only APC able to activate naive T cells.

Key components at the molecular level

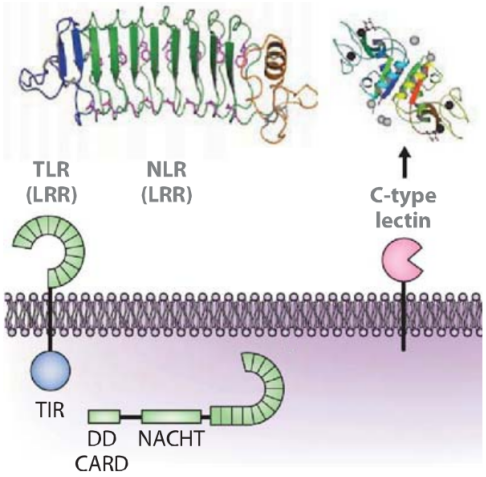
Receptors displayed at the surface of cells from the immune system play a major role during the innate immune response, acting as sensors to detect pathogens and communicate with other cells. For instance, receptors/ligand interactions are involved in antigen recognition and natural killer inhibition. After sensing by surface receptors, the information is transmitted through signaling pathways, leading to the activation or inhibition of effectors. Outside of the cell, secreted and seric factors are also key molecular components of immunity.

Membrane receptors of immune cells. Multiple types of membrane receptors can be found on the cells of the innate immune system. Among them, we focus on two groups: on one hand *pattern recognition receptors* (PRR) which recognize evolutionary conserved *pathogen-associated molecular patterns* (PAMPs), and on the other hand, cytokines and chemokines receptors [Jan89].

The first category can be split in several groups shortly described hereafter [Pau13, Chap. 15] (Fig. 2.2).

The family of *Toll-like receptors* (TLR) are transmembrane proteins which are able to recognize specific bacterial and viral molecular patterns, and activate the cell by which they are expressed upon binding. They can be either extracellular to recognize the surface molecules of pathogens, or intracellular to recognize viral RNA and DNA. *C-type lectin receptors* (CLR) are membrane-bound receptors which are involved in the recognition of fungal antigens and modulation of the innate immune response. The family of *f-Met-Leu-Phe receptors* are specific for formylated peptides found on the surface of bacteria. They are strongly expressed at the surface of granulocytes and mononuclear phagocytes. *RIG-I-like receptors* (RLR) are cytosolic RNA helicases, which specifically recognize viral and bacterial dsRNA. Finally, *NOD-like receptors* (NLR) are intracellular, cytoplasmic receptors which can sense various bacterial pathogenic molecules such as toxins or bacterial peptidoglycans. They are part of the inflammasome, a protein complex responsible for the activation of inflammatory processes.

Figure 2.2 Illustration of some pattern recognition receptor families discussed in the text, with representative members. LRR: leucine-rich repeats. Figure modified from [Pau13, Chap. 4].



Cytokines, chemokines and their receptors constitute the specialized communication system of immune cells. They can also be split in multiple groups described below (Fig. 2.3).

Members of the *class I cytokine receptors* family (also called *hematopoietin receptor* family) are the most numerous. These receptors consist of a cytokin-specific α subunit, and either a β or a γ signal transducing subunit possessing an cytoplasmic tail. However, there are exceptions to this structure such as the high affinity form of the IL-2 and IL-15 receptors which consist of trimers (α, β, γ).

The *class II cytokine receptors* family mainly bind interferons (IFN-α, IFN-β, IFN-γ) and, for this reason, is often called *interferon receptors* family. Its members also belong to the Immunoglobulin superfamily as they contain Ig-like domains.

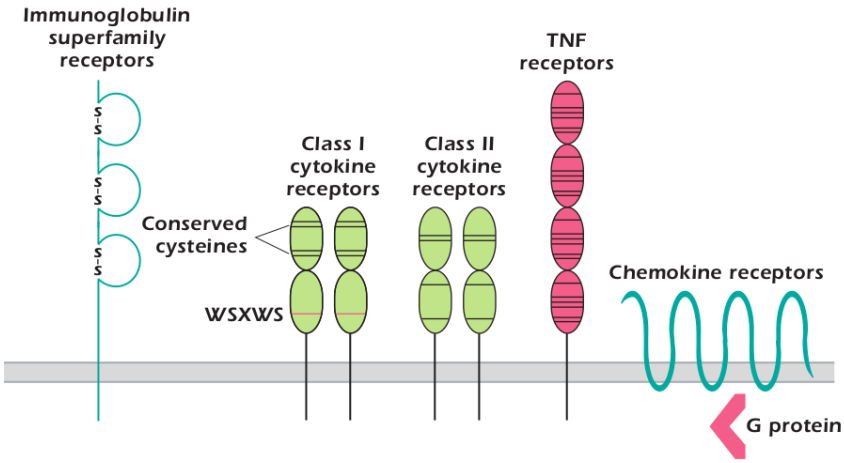
The *tumor necrosis factor (TNF) receptors* family can be split into death receptors, decoy receptors, and activating receptors differing by their intracellular domain. They mediate intracellular signals through *TNF receptor-associated factors* (TRAF).

The chemokine receptors family consists of G-protein-coupled receptors, which, interestingly, can bind some pathogens in addition to chemokines, sometimes leading to pathogen entry in the cell.

Finally, the immunoglobulin superfamily receptors consists of receptors which have one Ig-like domain.

Interestingly, multiple receptors (such as IL-2, IL-4, IL-7, IL-9, and IL-15) can share the same intracellular γ domain, which results in partial redundancy between their respective effects following activation.

Figure 2.3 Illustration of some cytokine and chemokine receptor families discussed in the text with representative members. Figure from [CS09, Chap. 11].



Specific signaling pathways of innate immunity. Signaling pathways are the link between pathogens or signaling molecules recognition by the receptors and the expression of effectors molecules such as interferon and interleukin. Ligand binding to PRRs or cytokine receptors results in a cascade of intracellular events usually involving multiple enzymes such as kinases and ubiquitases, which culminate in the activation of transcription factors and the expression of specific genes. Among the transcription factors involved, NF-κB is almost ubiquitous as it is found in TLR, RLR and NLR and multiple cytokine receptors pathways.

Secreted and seric factors. Communication between various cell types engaged in the immune response is mediated by various secreted factors. Such factors include cytokines (*e.g.* *interleukin* IL, *interferon* IFN and *tumor necrosis factor* TNF), *chemokines*, *acute phase proteins*, and the *complement system*.

Cytokines are soluble protein mediators which are produced by most cells of the immune system, and in particular T cells. Although they can be seen as the chemical messengers of the immune system, their action is not limited to its cellular components (pleiotropic properties), as they act on other cell types [Pau13, Chap. 25, 26, 27].

Cytokines can either be secreted or expressed as membrane proteins and can induce an effect at very low concentration (from 10^{-10} to 10^{-15} M). Each cytokine binds only its specific receptor, although there are counter-examples such as IL2, 4, 7, 9 and 15 which share a common domain recognized by all their respective receptors. The type of cytokines, their concentration and the amount of cytokine receptors expressed on the target cell determines how this cell is regulated. In particular, this allows for additive, synergistic or antagonistic effects between various cytokines.

Cytokines can have three modes of action depending on their reach: autocrine cytokines act on the cells which secreted them, and paracrine cytokines act on nearby cells. These are local effects as opposed to endocrine cytokines which act at a larger scale in a hormone-like fashion. Cytokines are involved in many processes such as promoting and ending the inflammatory response, inducing differentiation,

stimulating hematopoiesis, interfering with viral infection (IFN specifically), or initiating the acute phase response.

Chemokines are a specific sub-group of cytokines which have a chemotactic activity *i.e.* induce the migration of various cell types along their gradient in a process called chemotaxis. They also promote the production of adhesion molecules on target cells to facilitate the migration process [Pau13, Chap. 28].

Acute phase proteins are produced during the acute phase response, a systemic reaction occurring a few days after infection, and initiated by cytokine signaling. They are not produced by immune system cells but mainly by hepatocytes residing in the liver. Acute phase proteins include *C-reactive proteins* (CRP) and *mannan-binding lectins* (MBL) which are able to bind molecule on bacterial surfaces in an opsonin-like fashion [Pau13, Chap. 36] and to activate the complement system.

The *complement system* consists in a set of 30 circulated and membrane bound proteins, aimed at eliminating bacterial pathogens [Pau13, Chap. 36]. In the case of innate immunity, it is activated by the binding of opsonin C3b to the surface of pathogens, and leads to opsonization of the pathogen surface (for enhanced phagocytosis) by C3b and C5b, release of inflammation-promoting molecules (recruitment of phagocytes), and elimination of the pathogen by the membrane attack complex. In the case of acquired immunity, it is activated by Ig - Ag interactions.

In order to directly eliminate pathogens, the complement system initiates an enzymatic cascade which results in the *membrane attack complex* (MAC) being synthesized. This protein complex is then inserted in the cell walls surrounding the invading bacteria and results in their lysis.

2.1.2 The adaptive immune system

The adaptive immune system comes into play after the innate response if the pathogens could not be eliminated. It is tightly regulated and activated by the inflammation and the innate response. Moreover, it is characterized by the generation of antigen-specific cells, memory cells and immunoglobulin secretion. The main cells types involved in this response are B and T cells, which, as opposed to cells involved in the primary response, express somatically generated antigen-specific receptors.

B cells

B cells are the only cells able to synthesize *immunoglobulins* (Igs) (sections 2.2.1 and 2.3) and, depending on their developmental stage, can secrete them and display them on their membrane [GCJ56][Pau13, Chap. 8]. Depending on their location in the body, mature B cells may produce IgA, IgG or IgE after class switching (section 2.3.1, Fig. 2.12).

The number of pathogens an individual may encounter during its life is enormous. Therefore specific recognition of antigens requires a comparably large diversity of receptors. Since the number of Ig genes in the genome of an individual is limited, how Ig receptors as diverse as potential antigens can be produced has been a major question for immunologists. The recombination process of Ig genes described thereafter is responsible for this diversity and results in each B-cell clone expressing a different receptor.

B cells are produced from *hematopoietic stem cells* (HSC) which reside the in the bone marrow and further specialize in *multilineage progenitor* (MLP) and *common lymphoid progenitors* (CLP). At this point, the Ig genes are still in their germline state (section 2.2.1).

DJ_H recombination occurs at the pro-B cells stage. These cells show distinctive CD19 and CD10 membrane receptors which are still expressed at the next stage pre-B cell stage, however, the latter also display μ heavy chains with surrogate light chains λ_5 and VpreB. The μ, λ_5 and VpreB chains test whether the H chain is functional. If so, the L chain loci become accessible to the V(d)J-recombinase and V_L J_L recombination can start. If the H chain is not functional, the cell is deleted through apoptosis.

The μ heavy chain together with λ_5 and VpreB is non-covalently linked to other transmembrane proteins: Igα (CD79a) and Igβ (CD79b) which are bound together by a disulfide bond. Since μ chains only have a very short transmembrane segment, Igα/β are important for signaling and, when missing, block further development of the immune system. The μ heavy chains along with their surrogate light chains and Igα/β form the *pre-B cell receptor* (pre-BCR). V_H DJ_H recombination occurs at this stage and is followed by V_L J_L.

At the immature stage, B cells typically express the CD20 and CD19 membrane receptors. After recombination, light chains and μ chains pairing occur to form the IgM receptor which, together with Igα (CD79a) and Igβ (CD79b), forms the *B cell receptor* (BCR). It is also at that point that receptor editing may take place (section 2.2.1). Immature B cells then leave the bone marrow to move to the spleen and periphery where they become mature B cells. These display both IgM and IgD receptors and circulate through the lymph. As opposed to immature B cells, mature B cells can be activated by contact with antigens.

Upon contact with an antigen, mature B cells can differentiate into different subtypes. In particular, they can become memory B cells during the T cell dependent response. Although most of these have switched isotype (sections 2.3.1), *i.e.* do not display IgM nor IgD but instead Ig of other classes, some remain unswitched and express IgM, along with CD27 receptors which are often used as markers. Finally, they typically display high-affinity Ig since they have undergone affinity maturation, the process during which B cells undergo an accelerated selection driven by their affinity for an antigen (section 2.2.1). In some cases however, they have been observed to show no sign of affinity maturation or somatic hypermutation. Memory B cells have the ability to remain in circulation for a long time, to stay in a special niche in the bone marrow, and to quickly respond and secrete Ig upon secondary contact with the antigen.

Mature B cells can also differentiate into plasma cells after activation in the germinal center of lymph node and spleen. They synthesize and secrete Ig molecules of a single isotype and have no CD markers. A subset of plasma B cells is much long-lived and has the ability to remain alive after the infection. These memory plasma B cells keep secreting antibodies in the blood as opposed to "regular"memory B cells which rest until exposure to their specific antigen. As opposed to the latter they cannot further divide, *i.e.* they are terminally differentiated.

T cells

T cell belong to the lymphocytes (as B and NK cells) and are the only ones to display *T cell receptors* (TCRs).

T cell receptors. Due to the common origin of B and T cells, the TCR share some similarities with Igs and, in particular, the diversification mechanism of Igs and TCRs are similar and use the same set of enzymes. TCRs are made of two chains, (either α and β or γ and δ) each with a variable (V) and a constant (C) domain (Fig. 2.4). Most T cells display $\alpha\beta$ receptors but there also exists a subtype which bears $\gamma\delta$ receptors.

Similarly to Igs each V region contains three hypervariable loops called complementarity determining regions (CDR1, 2, 3) separated by framework regions (FR1, 2, 3, 4) . As for B cells, every T cell clone has a unique TCR (T cell receptor) which brings the number of unique TCR to $\sim 10^{18}$. Such a diversity is the result of rearrangements of the germline V(D)J genes encoding for the variable domains, a process similar to that occurring to Ig genes (section 2.2.1).

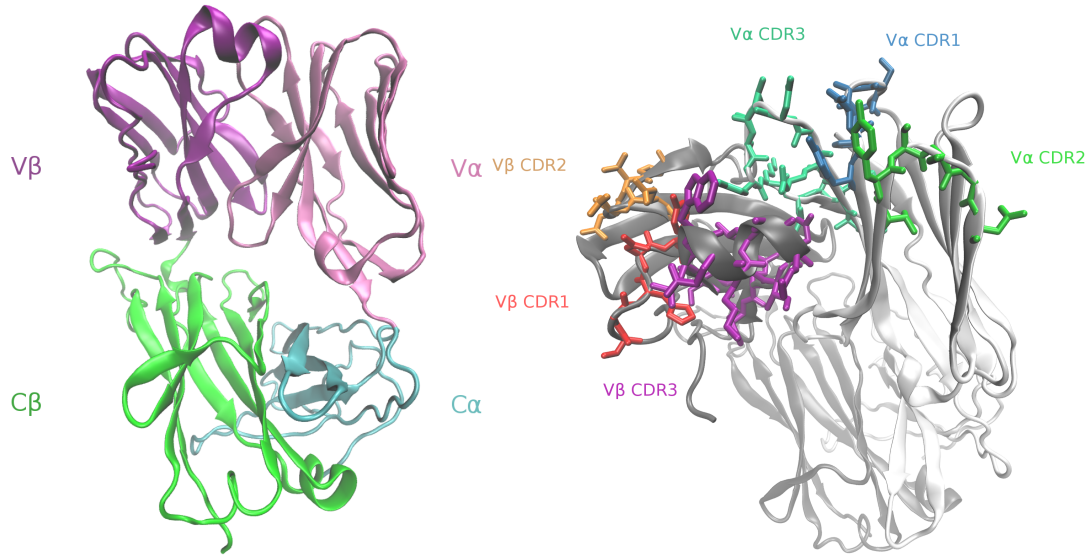
In contrast with Igs however, TCRs can only bind to antigens when presented at the surface of a host cell in combination with the MHC. The CDR3 is in contact with the antigen and is the most variable CDR in terms of sequence, whereas other CDRs mainly contact the MHC and are much more conserved.

Differentiation. T cells are produced from bone marrow-derived precursor cells which entered the thymus. During the next stage, rearrangement of β, γ, δ genes starts, and the decision for a cell to become $\gamma\delta$ or $\alpha\beta$ occurs. These precursors do not express CD4 or CD8 receptors which gives them the name of *double negative* cells. We now focus on the development of $\alpha\beta$ cells only.

After productive rearrangement of the β gene, chains β, pre-Tα, and receptors CD3 and ζ are expressed on the surface of the cell, forming the pre-TCR complex. This is the pre-T cell stage followed by the *double positive* stage, called this way because they express both CD4 and CD8 along with CD3 and $\alpha\beta$.

It is at this stage that T cells undergo thymic selection. This procedure consists of two phases and aim at obtaining cells which can only be activated by an antigen if it is associated to the MHC, but are not self-reactive. The first phase is the positive selection phase: double positive cells which have a high enough affinity for MHC presenting self antigen are selected ($< 10\%$). The second phase is the negative

Figure 2.4 Structure of a T cell receptor. C: constant domain, V: variable domain. Complementarity determining regions (CDR) and framework regions (FR) colors follow the IMGT color scheme. Structure: entry IMGT-5CO7 from the IMGT/3Dstructure-DB.



selection phase: double positive cells which have too high an affinity for MHC presenting self antigen are undergo apoptosis.

After thymic selection, the down-regulation of either CD4 or CD8 results in single positive cells. These naive cells can then join the bloodstream and circulate through the lymph.

Upon activation by an antigen, T cells differentiate into effector T cells. Namely, $CD4^+$ become helper T cells which synthesize cytokines and $CD8^+$ cells become cytotoxic T cells which can kill host cells infected by viruses. A small proportion of them becomes long-lived memory T cells, able to quickly multiply upon re-activation by their specific antigen.

The primary response

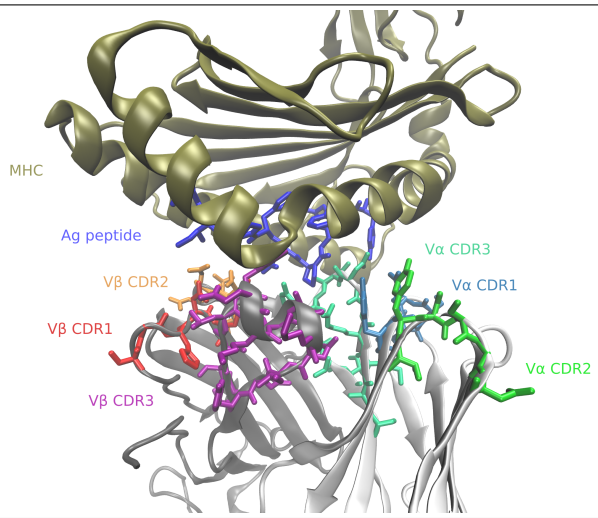
The primary response occurs after a first exposure to a particular antigen. A first period of time necessary for B and T cell to contact the antigen, get activated, communicate and differentiate is followed by an increasing secretion of Ig, which reaches a plateau before declining once the cause of infection has been eliminated. These events define four phases called *lag*, *exponential*, *steady state* and *declining phases*.

T and B cells. The first step of the primary adaptive is the activation of B and T cells. In the case of T cells, contact with the antigen occurs through a processed form displayed by antigen-presenting cells (APC); namely, T cell receptors contact the MHC-bound processed peptide (Fig. 2.5) presented at the surface of an APC. However this interaction is not sufficient for activation: other interactions between MHC and CD4, costimulator pairs, and adhesion molecules helps to further stabilize the interaction and enhance the activation of the T cell. Many intracellular events follow this binding event, leading to activation of an array of genes through various signaling pathways and resulting in synthesis of cytokines, clonal expansion and differentiation into effector cells (helper, cytotoxic and memory T cells). After the antigen has been eliminated, most of the large number of T cells generated at the clonal expansion step die, except for memory cells [Pau13, Chap. 14].

B cells can be activated in two ways: with the help of T cells for thymus dependent antigen (TD) or without their help for thymus independent antigens (TI), which are repetitive and therefore able to cross link the B cell receptors and activate the cell.

The B cell response to TI antigens mostly generates IgM (no isotype switch occurs), and does not results in differentiation to memory B cells. Notably, among TI antigens, some are able to activate

Figure 2.5 Ternary complex between the TCR, MHC and the antigenic peptide. CDR and FR colors follow the IMGT color scheme.



multiple B cell clones. On the other hand, TD antigens are mostly proteins and trigger the synthesis of high affinity antibodies. During the early phase of the response, IgM are produced, followed by other classes of Igs during the latter phase.

During the response to TD antigens, T cell - B cell interactions is of prime importance (Fig. 2.6). After a membrane bound Ig of a B cell binds to its specific antigen, the Ig - Ag complex is engulfed in the cells and processed to be displayed at the surface, bound to the MHC. B cells can then act as antigen-presenting cells (APCs) and display the antigen for recognition by T cells. The interaction activates both cells and leads to the formation of the germinal center where somatic hypermutation (section 2.2.1) and class switch recombination occurs (section 2.3.1). During this process, the cytokines expressed by T cells play a role in determining the class of the Igs which will results from the class switch [Pau13, Chap. 10].

The thymus-dependent antigen response also results in the differentiation of B cells into plasma and memory B cells.

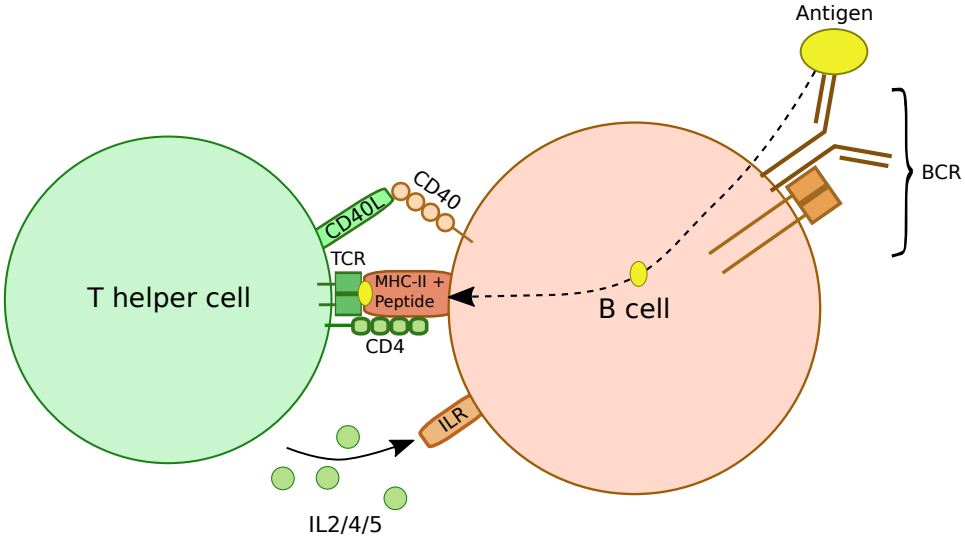
The secondary response

The secondary response occurs after exposure to an already encountered antigen. Thanks to the immunologic memory provided by memory B and T cells, the lag phase is much shorter than during the primary response. The ability for a secondary response to occur for a given antigen can persist for years.

B cells. The secondary response is faster than the primary response because memory B cells populations left from the first infection can respond to lower concentrations of their specific antigen than their naive counterparts. They also have the ability to quickly multiply and differentiate into plasma cells, leading to a copious secretion of antigen-specific antibodies. Since they already have undergone affinity maturation, the selection of B cells with highest affinity for an antigen (section 2.2.1), these antibodies are of much higher affinity than those of secreted during the first phase of the primary response [Pau13, Chap. 31]. As a result, both the speed of synthesis and the amount of high-affinity antibodies secreted are higher than during the primary response. Moreover, IgG appear at higher concentrations than IgM compared to the primary response because class switching has already occurred in most memory cells.

T cells. Similarly, memory T cells can be activated by lower levels of antigens and have a lower signaling threshold than their naive counterparts. This, coupled with a reservoir of antigen-specific T cells remaining from the primary response leads to a quicker an more effective response [Pau13, Chap. 31].

Figure 2.6 Interaction between a B cell and a T cell. The antigen is processed by the B cell, and the resulting processed peptide is displayed on the MHC class II for recognition by the TCR. Interleukines secreted by the T cell control B cell differentiation. BCR: B cell receptor (Ig + Igα + Igβ); TCR: T cell receptor; MHC: major histocompatibility complex; IL(R): interleukine (receptor). Reproduced from en.wikipedia.org/wiki/CD154#/media/File:T-dependent_B_cell_activation.png (public domain).



Importance of Ig and rationale for the choice of their study.

Although multiple proteins and, in particular, receptors such as TCR and PRR are of central importance during the immune response, the current work focuses the Igs for several reasons.

First, antibodies can bind to a wide range of molecules: proteins, peptides, haptens, sugars, lipids, despite a common structure. Such a versatility from a restricted structural diversity is extremely interesting in terms of binding mechanisms. Second, as opposed to the TCR, interactions occur with the antigen in its native form only, which means that the whole binding site of a specific Ig is involved in the interaction, and that the selection pressure applies to the whole binding site. This makes the study of mechanisms involved in tuning the binding affinity more straightforward, in particular when focusing on the design of binding molecules [LWT07]. Finally, antibodies are important tools for experimental biology (*e.g.* purification, fluorescence, assays), and promising therapeutic molecules (*e.g.* antibody based drugs for cancer, HIV).

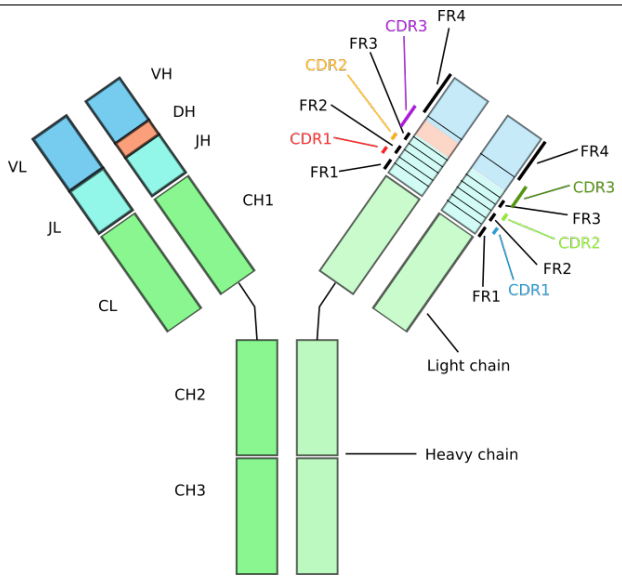
2.2 Antibody diversity: molecular genetics and repertoires

Because the potential antigens that will be encountered by the immune system cannot be predicted in advance, a large number of Igs with various specificities is critical for specific responses to all possible antigens to be possible. Informally, the set of Ig sequence that can be synthesized by an individual is called its Ig *repertoire*, or repertoire for short (Section 2.2.2). The potential, theoretical number (called potential repertoire) of Igs in an individual is estimated to range from 10^{15} to 10^{18} whereas the whole mammalian genome contains 10^4 to 10^5 genes ($\sim$20000 for the human). Therefore, specific diversification mechanism are needed to explain the diversity of the Ig repertoire [Ton83].

2.2.1 Molecular genetics of Ig gene diversification [Pau13, Chap. 6]

Igs are made of two heavy (H) and two light (L) chains (Fig. 2.7, section 2.3.1) whose V domain is resulting from a process of recombination. They are then paired to form the final receptor.

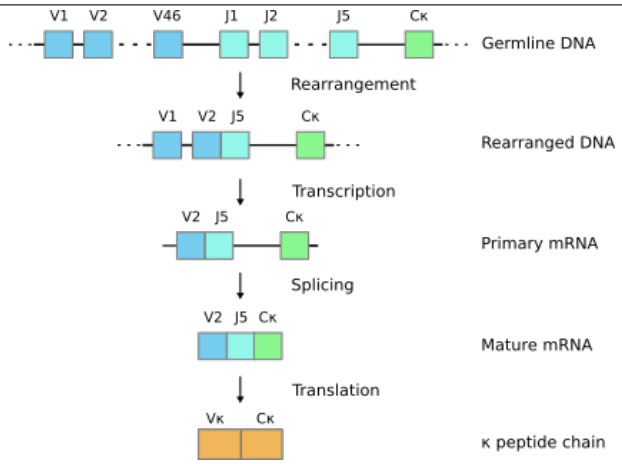
Figure 2.7 Schematic depiction of an Ig. Colors for the V, D, J and C genes correspond to those in Figs 2.8, 2.9 and 2.10. Colors for the CDRs correspond to those in Fig. 2.13. A detailed view of the upper right portion of the figure is depicted in Fig. 2.14.



Light chains locus. In mammals, two types of light chains exist: κ and λ. Both consist of two domains Vκ and Cκ, and Vλ and Cλ respectively [LL01]. The V domain of an L chain is coded by two gene segments, the V_L gene and J_L gene (or V_L segment and J_L segment) which recombine, and the C domain is coded by a single C_L gene.

Genes coding for κ chains are located on chromosome 2 in the human and 6 in the mouse. There are approximately 90 Vκ genes in the mouse and 40 in the human, separated by non-coding DNA. Five Jκ genes can be found downstream and a single Cκ gene is further downstream separated by an intron. This is true both for mouse and human although the third Jκ gene is not functional in mice (Fig. 2.8).

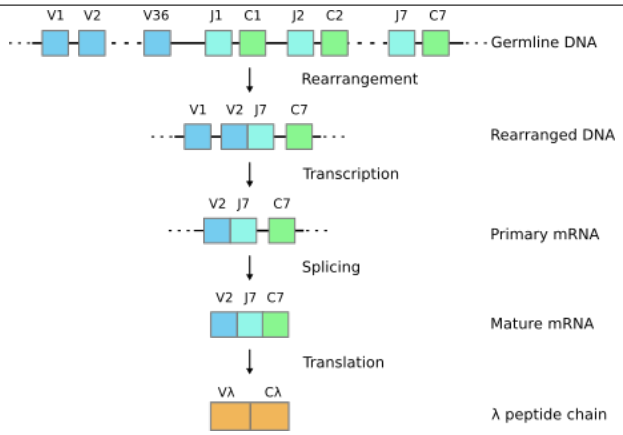
Figure 2.8 Genomic locus of the κ light chain. Adapted from [Pau13, Chap. 6].



In mice, the λ chain genes are located on chromosome 16 while they are located on chromosome 22 in the human. As opposed to κ chains, there are only 3 Vλ genes in inbred mice while there are around 30 in the human. Four Jλ genes are present in mice, one of them being non-functional while 4 to 5 functional genes are found in the human. An important difference between Jκ and Jλ genes is that each of the latter is followed by its own Cλ gene, the two of them making a single recombination units, thus

restricting the diversification potential of this locus (Fig. 2.9).

Figure 2.9 Genomic locus of the λ light chain. Adapted from [Pau13, Chap. 6].

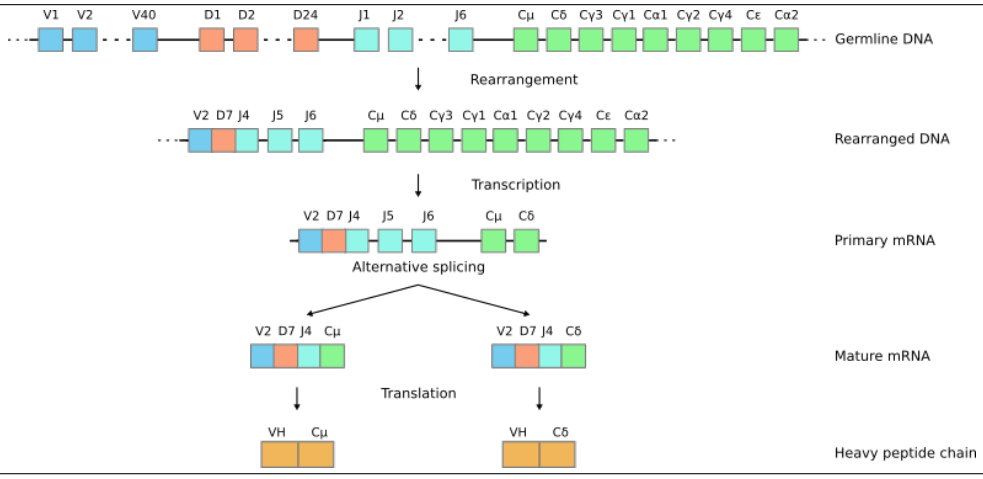


Heavy chain locus. In the mouse, genes coding for the H chain are located on chromosome 12 whereas they are on chromosome 14 in the human. Like L chains, H chains consist of a V and a C domain. However, the V domain is coded by three gene segments: V_H, D and J_H [LL01] which recombine.

In the mouse, around 100 V_H genes are classified into 16 families based on sequence similarity while approximately 40 genes form seven families in the human. There are $\sim$ 10 D genes in mice forming four families and they are approximately twice as numerous in the human. Interestingly, one reading frame is strongly preferred in the mouse whereas all three reading frames are used in the human. Humans and mice possess 4 and 6 J_H genes respectively all located downstream of the region containing the D genes.

Multiple C_H genes are further downstream such as Cμ, Cδ, Cγ and define the class of the resulting Ig (Fig. 2.10).

Figure 2.10 Genomic locus of the heavy chain. Adapted from [Pau13, Chap. 6].



Recombination. Most cells in the human body have the same genome and differentiate by expressing a subset of their genes. The situation is similar in mature B cells, except at the Ig genes loci where they discard or move part of their inherited (or *germline*) V, D and J genes during recombination, thereby sealing the specificity of the Igs they will be synthesizing [Ton83].

To make a complete Ig, a B cell has to produce one H and one L chain. The variable region of the H chain is typically synthesized by the recombination of one V_H, one D and one J_H gene. Similarly, each L

chain is synthesized from the recombination of one V_L and one J_L gene which form the variable domain and one C_L gene which forms the constant domain (Figs. 2.8, 2.9 and 2.10).

The recombination process, which slightly differs between L and H chains, is the following: The V_L and J_L genes are joined together by a V(D)J-recombinase (a collection of enzymes) which recognizes conserved *recombination recognition sequences* (RRS) located at the V, D and J genes. For H chains, two recombination events occur: the D and J_H gene are joined first, then the V_H gene is joined to DJ_H.

Junctional diversity. During the recombination events, the joining position of V_L and J_L for the L chain and of D and J_H, and V_H and DJ_H for the H chain are not accurate and a variable number of nucleotides can be deleted in the process. Moreover, random "N"-nucleotides and "P"-nucleotides can be added at the junction [LSJW$^+$84, LDB$^+$89]. N-addition, (for "non-germline" or "non-templated") is essentially random although it favors G and C. Palindromic "P"-nucleotides are added as follows: during recombination, the DNA is cut at the end of the coding segment and sealed to form a hairpin structure. This hairpin is subsequently cut open for the joining event to take place. However, it is not necessarily cut open at the same place it was sealed, which results in nicked extremities. Additional residues are added to fill the gaps, resulting in stretches of palindromic sequences, templated by the original sequence. Hence the junction of a given pair of segments can produce many different sequences.

Other diversification mechanisms

Random HL pairings. The diversity of Igs is also increased by the various possible combinations of H and L chains. They can be paired more or less arbitrarily, and B cells expressing receptors with compatible pairs are subsequently selected.

A given B cell produces one rearranged H chain and one rearranged L chain. Therefore, only one chromosome is used for each, either paternal or maternal, a process called allelic exclusion. Because of the random nature of the recombination and pairing process, there is a trial-and-error procedure which allows several Igs to be produced in case it is self-reactive or non-productive (*i.e.* contains stop-codons).

The H locus rearranges first. If no polypeptide chain could be synthesized after recombinations on one chromosome, the other is used, if it also fails, the cell enters apoptosis.

For the L chain, rearrangements generally occur in a ordered way for κ and λ chains. Namely, the following events occur: the κ chain of one chromosome is usually the first to be rearranged, expressed and subsequently tested. If it fails, the κ chain of the other chromosome is usually expressed and tested as well. If it also fails, the same process occurs for the λ chains on both chromosomes. If it also fails or if the L chain cannot pair properly with the available H chain, the cell enters apoptosis.

Somatic hypermutation and affinity maturation. Somatic hypermutation occurs during the B cell clonal expansion and consists in a targeted increase of the mutation rate in the V(D)J unit by up to 10000-fold [GJDH81]. As a result, point mutations lead to an array of B cells with varying affinities for the antigen. This process is followed by an intense selection of the B cells whose Ig binds the antigen with highest affinity. This results in a high concentration of selected mutations in the regions encoding the binding patch, *i.e.* CDR1, 2 and 3. This process only happens in the germinal centers of the lymph node and spleen upon encounter with an Ag and activation by helper T cells. It is a key component of affinity maturation, the process during which the population of B cells goes through stages of expansion and stringent selection driven by their affinity for an antigen.

Receptor editing. Because of the random nature of recombinations during B cell development, it may occur that some Ig have a strong affinity to self-antigens. Such a phenomenon is at the root of auto-immmune diseases and is strongly selected against. Schematically, when the Ig from the first successful rearrangement binds to a self-antigen, either its cell is deleted, or a second rearrangement occurs, leading to the replacement of the membrane Ig. Unused gene segments which were still in the germline DNA can thus be expressed and the first used gene segments are removed or down-regulated [RELW93]

2.2.2 Repertoires: potential and available

Antibody repertoire roughly correspond to the set all possible Ig sequences found in an individual or a species. However, this can refer to two related but distinct concepts: the potential and available repertoire. As introduced by Niels Jerne [Jer72],

> When an antigen confronts the immune system, it impinges upon a repertoire of available lymphocytes [...] We must distinguish between the potential repertoire of specificities that could arise given the genetic constitution of the zygote from which the animal develops, and the available repertoire embodied in the cells that can respond to antigens at a given moment in the life of the animal.

Potential. The potential repertoire of an individual can be defined as the set of V domain sequences that can be synthesized by its rearrangement machinery given its genetic constitution.

A rough estimate of the number H and L chains can be computed as follows: Assuming 40 different Vκ, 30 Vλ, 50 V_H, 5 Jκ, 4 Jλ, 6 J_H and 20 D genes (approximate values for the human), there are 200, 120 and 6000 Lκ, Lλ and H chains. Note that the latter number is likely underestimated because D genes can be expressed in three reading frames, although one is usually favored. Moreover, this calculation does not take into account junctional diversity, where nucleotides can be added or removed between V, D and J genes, further underestimating the total diversity.

H and L chains also can be paired arbitrarily, therefore, adding the random pairing of L and H chains to the previous estimation results in $(200 + 120) \cdot 6000 \approx 2 \cdot 10^6$ different Igs. This diversity is greatly increased during affinity maturation which acts on the whole V-region.

Available. As opposed to the size of the potential repertoire, which can be theoretically derived with a combinatorial calculation from the number of genes, the size of the available repertoire cannot be estimated easily. In effect, the available repertoire varies between individuals and with time: it essentially results from both the history of infections which occurred to an individual and from the mechanistic constraints of the rearrangements. Moreover, the repertoire resulting from the response to the same antigen differs wildly between individuals independently of their genetic makeup [Jer72]. Therefore, one has to resort to experimental methods to characterize the available repertoire of a given individual at a given time.

Recent years have seen the fast development and spread of high throughput sequencing methods (or next generation sequencing, NGS) resulting in the ability to obtain massive amounts of nucleic-acid sequences. In particular, they opened the way for whole transcriptome sequencing which could be used to obtain all Ig RNA sequences from individuals at a given time [WJW$^+$09, JWP$^+$11, JHW$^+$13, KLS$^+$14].

Analyzing data obtained thanks to NGS offers invaluable insights on various aspects of the Ig repertoires such as clonotype and isotype frequencies, CDR3 diversity[SMFC$^+$13] and overall structure [MWBC10] [BMFDP$^+$08].

2.3 Structure and function of immunoglobulins

2.3.1 General structure of immunoglobulins

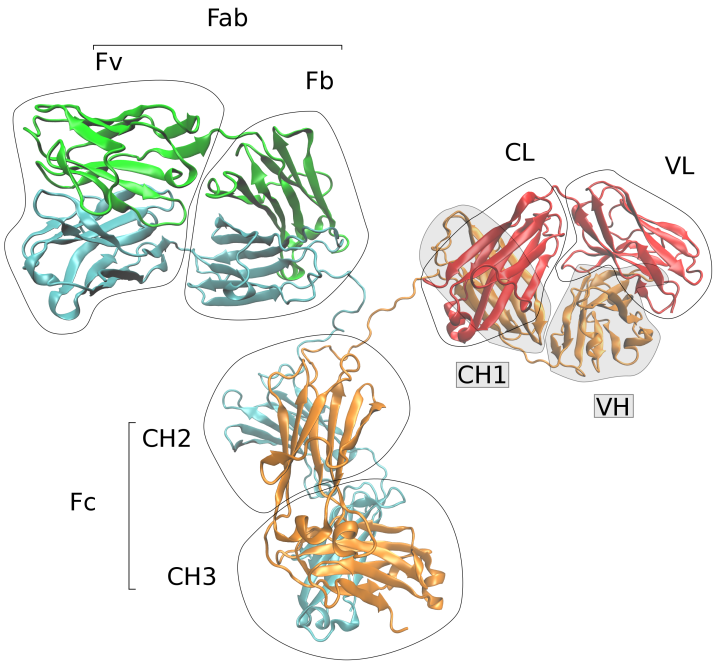
Immunoglobulins or *Igs* are proteins consisting of two heavy (H) and two light (L) chains. The H and L chains, respectively, are identical within an Ig molecule. Each chain is divided into a variable and a constant region. The variable region correspond to a V_H or V_L (for variable) domain, and the constant region to one or several C_H or C_L (for constant) domains [LL01][ECG$^+$69] (Fig. 2.11).

The *Fab* consists in the V_H and V_L domains along with the first C_H and the only C_L domain. It has been historically defined after digestion of an Ig by papain. The *Fc* is the other part produced by the digestion, *i.e.* the remaining C_H domains.

Within the *Fab*, the variable fragment *Fv* consists of the V_H and V_L domains, and the constant fragment *Fb* in the first C_H and C_L domains.

The site of interaction between an Ig and an antigen (Ag) is called *paratope* on the Ig side and *epitope* on the Ag side. Therefore, the same monospecific Ig can bind to distinct antigens (Ags) provided they have the same epitope.

Figure 2.11 Structure of an immunoglobulin (Ig).



Constant domain: classes and isotypes

The *class* of an Ig is defined by its constant region, and the corresponding notion at the gene level, i.e. the C gene variant, is called an *isotype*. This implies that, Igs with the same Ag specificity can be of different classes. In particular, the first class expressed during B cell differentiation is always IgM. The genes encoding the variable region of an Ig, which define its the specificity, can be later combined to other C genes during a genomic recombination event called "switch". This leads to the replacement of the Cμ domains by other C domains in the protein. Different classes have different effector functions and are important in various aspects of the immune response.

In humans, there are 5 classes of Ig [Pau13, Chap. 5] (Fig. 2.12). IgMs are monomeric when bound to the membrane and pentameric when secreted (which leads to high avidity). They are mainly found during early primary response, before class switching occurs or during thymus independent response (section 2.1.2). The pentameric form of the secreted IgMs makes them very efficient at activating the complement. This property, along with the early synthesis make them the most important class of Ig during the early immune response.

IgDs are monomeric and mainly found in membrane form. As opposed to other classes which are produced after a recombination during class switching, IgDs only differ from IgMs because of a differential splicing pattern. They are found in serum at very low level, and are mostly found at the surface of mature B cells in conjunction with IgMs. Their function has not been clearly elucidated.

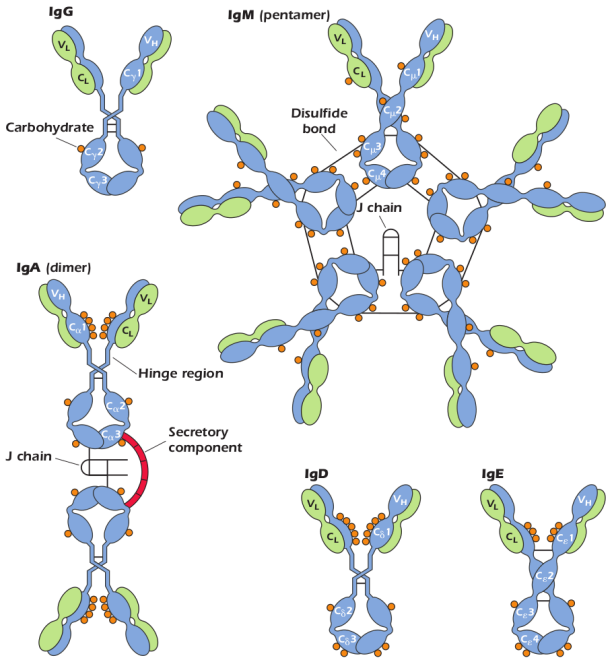
IgGs are monomeric and are the most common Igs found in blood. They are the result of IgMs undergoing class switching and can undergo affinity maturation (section 2.2.1). They can be further split in four subclasses: IgG1, IgG2, IgG3, IgG4. IgGs cause agglutination of insoluble antigens, and can also precipitate soluble multivalent antigens so that the resulting insoluble Ig - Ag complex can be phagocytized. They also have the ability to coat pathogens in a process called opsonization so that phagocytic cells bearing receptors for the Fc portion of the Ig can more readily phagocytize the antigen. Natural killer cells which are also bearing Fc receptors, take advantage of the IgG coating to destroy the

cell in a process called antibody dependent cell-mediated cytotoxicity (ADCC). Finally, IgGs have the ability to activate the complement and neutralize viruses and toxins.

IgAs are mostly found in secretions (*e.g.* tears, saliva, mucus) in dimeric form, and also in the plasma where they are monomeric. Mucosal IgAs are one of the most important Igs during respiratory or gastrointestinal infections due to their location. IgAs can also trigger agglutination and possess viral neutralization properties.

IgEs are typically associated with allergies and immunity to parasites. They are found in very small concentrations because of their high affinity to the FcϵRI receptors found on the membrane of mast cells and basophils.

Figure 2.12 Structure of the five Ig classes. Figure from [CS09, Chap. 4]



Variable domain

CDRs and FRs. The variable regions of H and L chains contain three hypervariable regions called *CDRs* for complementarity determining regions roughly corresponding to the binding site of the Ig. Such variability has first been characterized from the sequence [KTWF$^+$92] although current methods to delimit CDRs also use structural information [LL01].

Because of this variability however, there is a strong need to maintain a structure stable enough to support any complementary V domain, which is possible through the high conservation of so-called *framework* regions (FRs). There are four FRs which are located between the CDRs along the sequence. Because of sequence conservation maintaining a beta barrel conformation of the domain, different FR1 cluster V_H genes in *families* and in *clans* along with FR3 (Fig. 2.13).

CDR1, CDR2 and FR1, FR2, FR3 are all coded by the V gene segment while the CDR4 is coded by the J gene segment. In contrast, CDR3, is coded by the junction between the V_L and J_L genes in the L chain and by the junction between the V_H, D and J_H genes in the H chain (Fig. 2.14). Because of this and the imprecise joining of the gene segments together during recombination (section 2.2.1), VH CDR3 is the major site of diversity in the V domain.

V domain residue numbering schemes. The first numbering scheme for the residues of the V domain was designed by Kabat and colleagues [WK70, KTWB76] and was based on protein sequence data only. In short, the variability at every site is computed after alignment of the V domain sequences,

Figure 2.13 Structure of the Fab, with colored CDRs, and grey FRs.

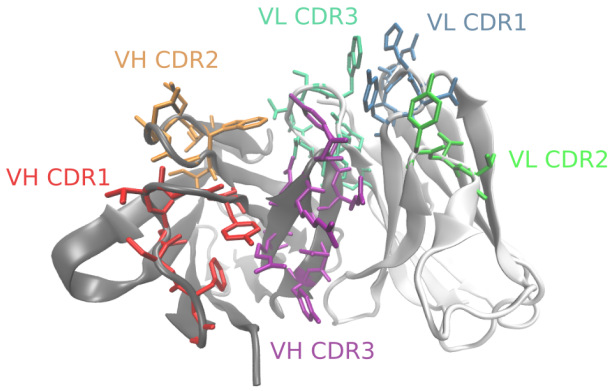
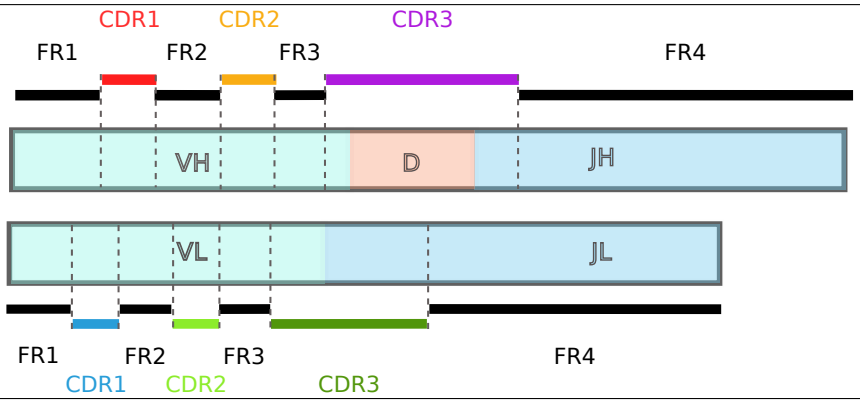


Figure 2.14 Encoding of CDRs and FRs by the V, D and J genes. The color scheme is the same as used in Fig. 2.7.



and the regions of high variability are defined as the CDRs, whereas the most conserved ones are defined as FRs. The main drawback of this numbering scheme is that only a limited number of insertions can be numbered in a standard way, and that some positions of longer CDRs (especially VH CDR3) cannot be assigned a number.

With the advent of crystallographic structures of Igs, this numbering was revised by Chothia et al [CL87] to fit the structural locations the insertion positions in CDRs. In particular, the positions of VH CDR1 and VL CDR1 were updated.

More structural data led to the revision of the positions for the latter CDR [CLT+89]; a change which was reverted after more structures became available [ALLC97].

The IMGT numbering scheme [Lef99, LPR+03] (Table 2.1) was then introduced with the following properties:

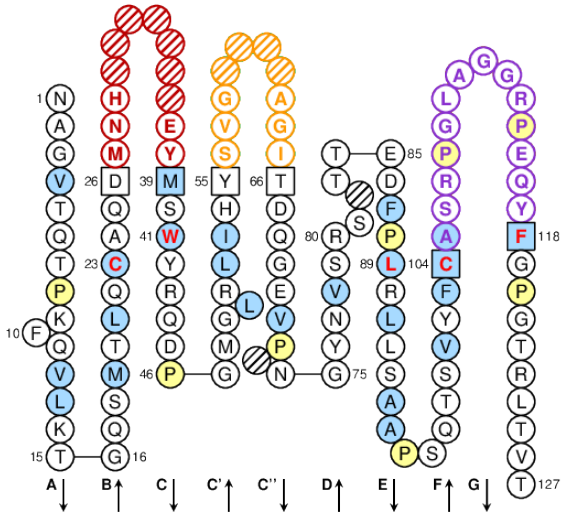
- it remains consistent across heavy and light V domain of Igs, and V domains of TCRs as well as for other Ig domains of the V types found in other proteins. It has also been extended to C domains to cover all structures within the Ig superfamily.
- it corrects a drawback of the previous numbering schemes which have to use insertion letters when a CDR is longer than commonly found (*e.g.* 60A). However, new structures of Igs with very long VH CDR3 must still use insertion letters.
- it always places conserved residues at the same position as well as hydrophobic residues of the FRs (23, 41, 89, 104 and 118; red residues in Fig 2.15).
- it places insertions in CDRs symmetrically around a central position (not in the original scheme but it was added in subsequent versions). This can be seen for VH CDR1 and VH CDR2 on Fig. 2.15 where missing positions are at the center of the loop..

These properties allow the comparison of sequences without necessarily resorting to alignments. An example of the graphical representation of a V domain based on this numbering (denoted "collier de perles" is shown in Fig. 2.15).

Table 2.1 Residue-based IMGT numbering for CDRs and FRs. Table from `http://www.imgt.org/IMGTScientificChart/Nomenclature/IMGT-FRCDRdefinition.html`

FR1	CDR1	FR2	CDR2	FR3	CDR3 germline (rearranged)	FR4
1-26	27 - 38	39 - 55	56 - 65	66 - 104	105 - 116 (117)	118 - 129

Figure 2.15 Graphical representation of a V domain according to the IMGT numbering scheme. Hatched circles correspond to positions which have been assigned a number but have no corresponding residue in this particular sequence. CDR anchors positions are displayed as squares. Conserved residues which always receive the same number are in (bright) red. VH CDR1, VH CDR2 and VH CDR3 are respectively colored (dark)red, orange and purple. Figure from the IMGT website `http://www.imgt.org/3Dstructure-DB/cgi/collier_perles.cgi?domcode=1AO7ED00&domdescr=V-BETA&domnum=1`



The Aho numbering scheme [HP01] was designed following a different approach: the structures of the V domains of Igs are structurally aligned and numbered following this alignment. This scheme is rather similar to the updated IMGT numbering scheme but has three main differences. First CDR1 is divided in two parts following the observation that a residue in the middle of the corresponding loop assumes a specific and conserved conformation. Each part has an insertion position, which results in two insertion positions for CDR1 as opposed to one for the IMGT numbering scheme. Second, CDR2 also has two insertion positions, the second added in order to account for the distinct conformation assumed by Vα from the TCR compared to Igs. Finally, longer CDR3 can be numbered without adding insertion letters compared to the IMGT numbering scheme.

Despite their drawbacks, Kabat / Chothia numbering schemes are still used for analysis of Igs sequences. For this reason, a last update to the Kabat / Chothia numbering [AM08] was made, mainly correcting the positions of insertion in FRs.

2.3.2 Atomic structure of immunoglobulins

Among studies of Igs at the atomic level, many have focused on the recombining site, and on CDRs in particular. Structurally, CDRs typically form loops of variable lengths, and, despite their wide variability, all but HCDR3 have been shown to form classes of "similar" conformations coined *canonical structures* [CL87, CMCT11, CLT$^+$89, MTR$^+$98, NLD11]. Only 10 of these classes describe most of the

human and mouse sequences. Moreover, canonical structures of the V_H domain are under influence of the FRs, which result in a correlation between canonical structures and families defined by FRs.

Because of the contribution of all three V_H, D and J_H genes segments and random or templated nucleotide additions (section 2.2.1), VH CDR3 is the most variable CDR. This explains why canonical structures have remained more elusive and why, despite their classification in kinked, extrakinked and extended and their associated predictors, the prediction of their conformation is still difficult [SKN99, SKN96, KKGE06, KSKN08].

Interestingly, one study considered all CDR conformations at once instead of only focusing on single CDRs [VMLOA95], highlighting the fact that some combinations are multi-specific, while others are specific of an antigen type.

However, the relevance of canonical conformations for the prediction of the 3D structures of CDRs of Igs was questioned [CD11], since general loop prediction methods matched (and in some cases outperformed) the prediction performances of methods exploiting specific rules associated with canonical conformations of CDRs.

In the case of the TCR, the situation is different since binding to the MHC imposes strong constraints on the conformation of its binding site. This makes it a more favorable target for the prediction of canonical conformations which originally prompted its transposition to Igs.

2.3.3 Immunoglobulins - antigen recognition

The recognition of antigens by Igs is central for the success of the adaptive immune response. Moreover, Igs are commonly used during experiments as binding molecules for *e.g.* purification. It is therefore not surprising that Ig - Ag complexes have been well-studied and that a fair part of these studies has focused the Fab and Ag binding site.

Physico-chemical properties of Ig binding sites

The huge diversity of Ig specificity naturally prompted the analysis the amino-acid composition of the binding site, and, when enough crystallographic structures became available, of the contacting residues.

In general, it has been observed that the Ig binding site has a preference for slightly hydrophilic residues and aromatic residues, specifically Tyr and Trp [SM02]. Apart, from Arg, charged residues are usually under-represented [CBM03]. In particular, the strong preference for Tyrosine has been explained by both over-representation of Tyr codons in germline genes, and preferential genes rearrangement and affinity maturation [ILIS02][SM02].

Several works using restricted sets of amino-acids encoding the solvent-accessible CDR positions have shown that nanomolar affinities can be reached using phages display using only Tyr and Ser [FLC$^+$05]. Moreover, Tyr-rich binding sites tend to result in a higher specificity, whereas Arg-rich ones lead to non specific binding [BZF$^+$08].

Finally, still using a restricted amino-acid alphabet Tyr has been shown to occur more often at antigen contacting positions than Ala, Asp and Ser [FFS04][FBKS06]. This somewhat agrees with the observation that in natural Ig, Arg, Asn, Asp, His, Ser, Thr and Tyr are over-represented at the antigen-contacting positions, whereas, Cys, Pro, Gln, Glu and hydrophobic residues are under-represented [RSWA12].

It has been noted that there is a usually stronger physico-chemical complementarity between the Ig and Ag side of the interface than between general protein - protein partners. Namely, hydrophobic patches, polar residues of opposite charge and electron donor/acceptors face each other [SM02].

Flexibility and influence on the binding

The structure of the Fab suggests that FRs, which are well structured should be rigid, while CDRs which consists of loosely structured loops should be flexible. This intuition is not always verified as large conformational changes, beyond individual CDRs, and rigidification of the CDR loops in bound form have been observed. This has important implications for the binding affinity as we discuss below.

In general, only small structural rearrangements occur upon binding, such as concerted movements of the CDRs of less than 3Å (lRMSD). On a larger scale, slight shifts in the relative orientation of the VL and VH domains ($<$ 3 degrees) are also common. However, larger conformation changes such a as

CDR loop displacement up to 7 Å, have been described for certain complexes involving DNA or peptide antigens [SM02].

Structural flexibility can also play an important role during affinity maturation [MSSR00]. In particular, comparing mature and naive Ig from the same lineage, several works found that the difference did not lie in their binding mode but in the rigidity of VH CDR3. In other words, the binding contact were similar, but VH CDR3 of the mature Ig was rigidified compared to the naive one, and pre-configured to its binding configuration, thereby reducing the entropic cost of binding [JSY+04, ZOT+06, WSJ11, SXK+13].

In another instance, the naive Ig was shown to go through a large conformational change upon binding (4.6° change in the angle between H and L chains), while the mature Ig does not (0.4° change for the same angle) [WPW+97].

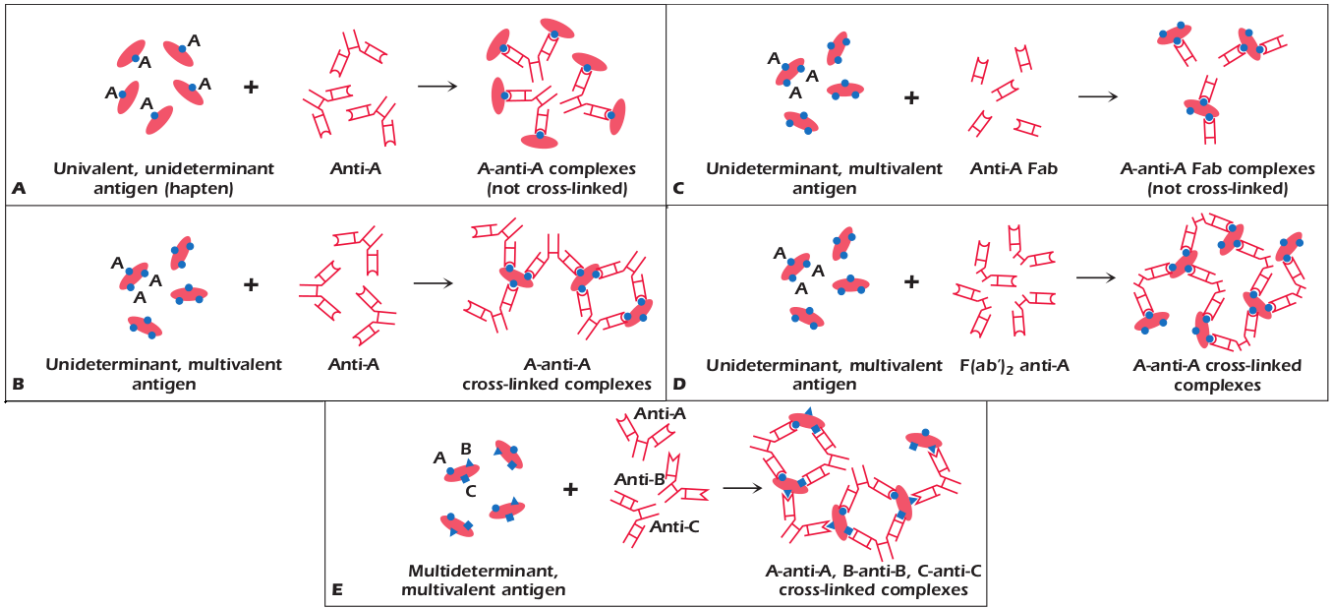
Role of the constant region

Although studying the V domains seems natural considering that they contains the Fab and binding site, several studies have found that constant regions also influence the binding affinity. Another relevant concept in that context is that of *avidity* or *functional affinity*. As opposed to affinity which is related to the monovalent binding of a Fab to an Ag, avidity refers to the strength of (positively or negatively cooperative) interactions between one or more Fab with a complex (multivalent) antigen (Fig 2.16).

Several studies have investigated the avidity of different subclasses of IgG (*i.e.* with the same V region) to the same antigen, showing that various subtypes bound to the same Ag with varying avidities [MTS+96]. In two cases, these results were shown to depend depending on the antigen density [CSG+93] [CRGG94], hinting at cooperative effects.

Two other papers have also shown a difference in affinity between Ig with C1 domains coming from different classes [PHCB+96][PMD+00], and also subclasses of IgG [TFFFC07].

Figure 2.16 Illustration of monovalent and multivalent Ags. Figure from [CS09, Chap. 5].



2.4 Modeling of Ig - Ag complexes.

2.4.1 Key notions in physical chemistry

How strongly molecules interact: a primer on binding free energy

When two molecules, called *partners* in the sequel, assemble to form a complex, the strength of this association is called the *affinity* or *binding free energy* of the complex.

Because interactions between molecules are at the heart of most if not all biological processes, binding affinity is a central notion in structural biology. In effect, the stability of a complex is often critical for a function to be carried out, which is yet another instance of the "structure - dynamics - function" paradigm. For instance, the affinity between the TCR and the corresponding MHC/peptide complex must be high enough for the interaction to activate the T cell hosting the TCR, but this interaction should be short-lived enough so that the T cell can go away, and other clones can bind to the same MHC and become activated. On the contrary, the affinity between an Ig and its antigen must be very high because dissociation is not desirable neither for secreted Ig, which should stick to the pathogen until eliminated, nor for membrane-expressed Ig which should be engulfed in the cell and processed. For general protein-protein complexes, affinities measured by dissociation constants (K_d, see next paragraph) span 11 orders of magnitude, a range illustrating the diversity of biological processes and the various binding modes inherent to them [JBC08].

Estimating binding free energies is therefore central in order to understand how biological systems regulate the strength of the association of molecular partners, and the ability to do so reliably and accurately would be a major step toward a functional analysis of interactomes [Bon10]. Moreover, binding affinity prediction can be applied to various tasks such as protein design, docking or ligand discovery with application in medicine and in particular drug design [CeCG07, GZ07], such as therapeutic peptides [RVK08] and Igs [LWT07].

Remark. When referring to the binding free energy, we are not being completely accurate. The relevant quantity is actually the change in free energy between the complex and the unbound partners. This is clear when considering the delta notation for the binding free energy ΔG. In the sequel we will keep abusing terminology, and refer to this variation by the terms affinity, binding affinity or binding free energy indifferently.

Formal definition. The binding free energy is a well defined quantity in the realm of chemistry and thermodynamics and, as such, can be experimentally quantified. It is therefore amenable to (bio-)physical modeling.

In particular it can be expressed in two forms: for two partners A and B forming a complex AB, the first one depends explicitly on the ratio K_d (called the dissociation constant) of the concentration of the isolated partners ([A] and [B]) and of the complex ([Ab]):

$$\Delta G = -RT \ln K_d/c^\circ = -RT \ln \frac{[A][B]}{[AB]} \tag{2.1}$$

where R is the gas constant, T is the absolute temperature and c° is the standard concentration (1M). The second form depend on the change in two classical thermodynamic quantities between bound and unbound form: the variation of enthalpy and entropy.

$$\Delta G = \Delta H - T\Delta S. \tag{2.2}$$

Here ΔH is the change in enthalpy, T is the absolute temperature, and ΔS is the change in entropy.

Experimental determination. A large number of experimental methods have been designed to quantify the binding free energy of two compounds. Assuming one seeks to measure the binding affinity of a complex involving two proteins called partner A and partner B, we now quickly describe three of the most commonly used methods.

Isothermal titration calorimetry (ITC) measures the amount of heat taken or released by a reaction. For this, the temperature of a solution containing the partner A is monitored while a solution of the

partner B is being added. A heater is used to keep this solution at the same temperature as a control solution where only buffer (*i.e.* no partner B) is added. The free energy can be deduced from the curve of the amount of energy used as a function of the amount of second partner being added [CJI68].

Sometimes, one of the partners cannot be put in a solution. In this case, surface plasmon resonance (SPR) can be used to quantify the binding free energy [LNL83]. Partner A is adsorbed on a surface while partner B is in a solution which flows over the surface. This surface is the floor of a so-called flow-cell whose bottom part is a thin layer of refractive metal. The binding of free-floating molecules to the adsorbed ones is monitored by looking at the refracting index of this refractive layer. The change in this refractive index upon binding and unbinding of partner B can then be used to calculate ΔG.

Fluorescence-based methods are typically used to measure high affinity interactions. A binding assay is prepared where partner A is in solution with an fluorophore-labeled partner other than B. Partner B is then added progressively, displacing the labeled partner, which can be monitored thanks to its fluorescent properties.

Such methods typically yield error between 0.1 and 0.25 kcal / mol [GZ07, KMH$^+$11, CM13]. However, experimental conditions and in particular solute concentration, temperature, ionic strength, or pH, can have a major impact on the measurements, up to 2.3kcal/mol (a factor of 48 on K_d) [KMH$^+$11].

Enthalpy - entropy compensation

One of the classical postulates of thermodynamics is that a reaction is favorable if its ΔG is negative. It is clear that, for a complex at a constant temperature, this can be achieved either by lowering the enthalpy or by increasing the entropy (Eq. 2.2).

In terms of molecules binding this means more atomic interactions in the first case, or reduced loss of freedom to move in the second case. More precisely, more/stronger atomic interactions can be obtained through a better surface complementarity and a tighter packing at interface, whereas more freedom to move is related to several size and timescales. Namely the atomic (vibrational entropy), residue and domain (conformational entropy), and global (rotational and translational) scales have been considered.

The previous conditions seem hardly compatible since a tighter packing, for instance, will likely reduce the vibrational entropy, resulting in an entropic penalty. This phenomenon has been observed, noticing that ΔH and $T\Delta S$ are correlated. Namely, increasing one will increase the other, up to the point where very small or very large values for both quantities end up making a small difference in the resulting ΔG. This can be observed on plots where the curves for ΔH and $T\Delta S$ are very close to each other (Fig. 2.17). The closer they are, the more difficult it is to identify their crossing point, which is the point at which the reaction become favorable. This phenomenon is called the enthalpy-entropy compensation (EEC) [MABM10, Dun95] and is one reason why estimating the binding affinity is difficult: small errors on either term not compensated in the other can lead to large errors on the affinity predictions.

Binding free energy calculation: the statistical thermodynamics approach

The previous formal definitions (Eq. 2.1 and 2.2) gives a fairly limited amount of information about the physical meaning of K_d and ΔG, in particular when considering molecular conformations.

From the statistical thermodynamics point of view [GZ07], and ignoring solvent for simplicity, the chemical potential μ_m of a molecule with internal coordinates (*i.e.* conformation) r_m at concentration C_m is:

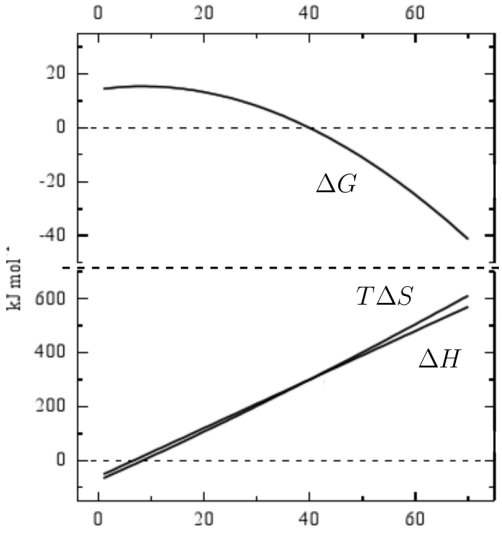
$$\mu_m = -\frac{1}{\beta} \ln \left(\frac{8\pi^2}{C_m} \int e^{-\beta U(r_m)} dr_m \right) \tag{2.3}$$

where $U(r_m)$ is the potential energy of conformation r_m, $\beta = 1/RT$, R is the gas constant and T is the absolute temperature. The integral term is analog to summing over potential energies of all possible conformations, giving higher values to lower (*i.e.* more favorable) energies. The factor $8\pi^2/C_m$ is due to the fact that overall rotations do not affect the integral since it is over internal coordinates.

Consider the binding free energy for two partners A and B as the free energy change of putting the complex in solution (μ_{AB}) plus that of removing isolated partners from a solution ($-\mu_A - \mu_B$):

$$\Delta G = \mu_{AB} - \mu_A - \mu_B \tag{2.4}$$

Figure 2.17 Illustration of the phenomenon of enthalpy-entropy compensation. $T\Delta S$ and ΔH are well correlated which makes their crossing point very sensitive to small changes in their slope. Moreover, the scale of changes for ΔG is much smaller (one order of magnitude) than that of the changes in $T\Delta S$ and ΔH. Figure courtesy of Alan Cooper.



Equation 2.3 can then be rewritten as follows:

$$\Delta G = -\frac{1}{\beta}\ln\left(\frac{1}{8\pi^2}\frac{C_A C_B}{C_{AB}}\frac{\int e^{-\beta U(r_{AB})}dr_{AB}}{\int\left(e^{-\beta U(r_A)}dr_A\right)\left(\int e^{-\beta U(r_B)}dr_B\right)}\right) \tag{2.5}$$

Calculating a binding free energy therefore necessitates to integrate potential energies over the configuration space of the complex and both partners. As discussed in section 2.4.3, Eq. 2.5 makes it look deceptively simple.

Geometric models of protein - protein complexes

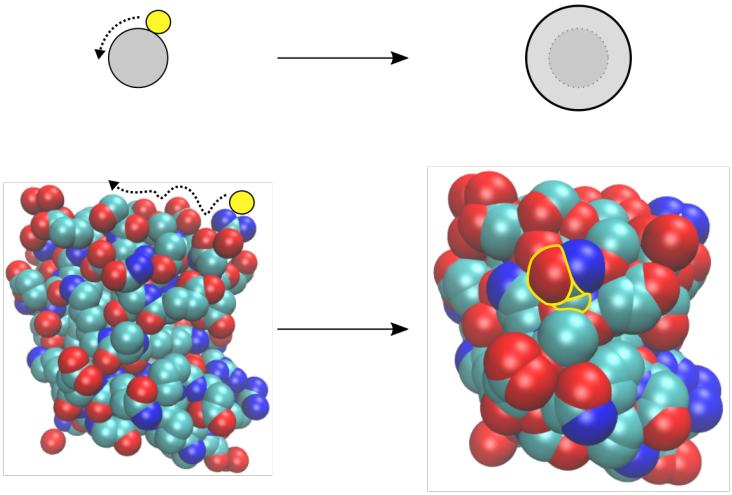
Non-parametric descriptors. Analysis of protein - protein complexes has originally relied on descriptive quantities related to non-bonded interactions found at interfaces. Such quantities, *e.g.* the number of van der Waals bonds, salt bridges and hydrogen bonds are easy to compute and only rely on the distance between the center of atoms. Such an approach require thresholds to be defined in order to assess whether a given interaction occurs between two given atoms. Such thresholds have not always been used consistently [KJ13] and must therefore be taken into account when considering the resulting analyses. Such quantities do not require the definition of an *a priori* model of the protein hence the name *non-parametric*; in effect, only the atomic coordinates, *i.e.* the coordinates of the center of atoms, are needed, which are provided by crystallographic data.

The definition of interface between two molecules with this approach is also based on a threshold distance. Namely, pairs of atoms belonging to different partners and which are closer than a certain threshold (typically 5Å) are considered to be at interface. This approach has the drawback to classify too many atoms as interfacial in convex regions [Caz10].

Parametric descriptors In order to gain more insights in the factors contributing to the association of molecular partners, it has been considered useful to extend such descriptive quantities to take into account the *shape* of molecules. The notion of shape is particularly relevant to define the concepts of molecular surface and related solvent accessible area (SAS), and that of molecular volume. These concepts pave to way to compute useful quantities such as the buried surface area (BSA), surface complementarity measures, surface curvature, and atomic packing, which are detailed shortly after.

As opposed to the non-parametric descriptors described above, some choices have to be made in order to obtain a useful representation of the shape of a molecule from the data at hand. In particular since a molecule is a collection of atoms, one has to define a model for single atoms, along with a way to combine these in multiple-atom models. The classical geometric model of an atom is that of a ball whose radius is equal to its van der Walls (vdW) radius. Extending this idea, the concept of *solvent accessible model* (SAM) was developed in order to assess the accessibility of the atoms to solvent molecules of various sizes [LR71]. The SAM of an atom is a ball whose radius is its vdW radius augmented by the radius of a solvent molecule. For water molecules the typical value for that radius is 1.4Å. This model allows a molecular complex to be described as a union of balls, and to apply the tools provided by geometry to their analysis (Figure 2.18). It has become central in the analysis of protein - protein complexes [CJ75, BCRJ04b].

Figure 2.18 Molecular surface and solvent accessible model (SAM) of an atom and a molecule. The surfaces on the left correspond to the van der Walls (vdW) radius of the atoms, and those on the right to the solvent accessible surface (SAS). The vdW representation of the atom was superimposed on the SAS representation in the upper right portion of the picture (darker, dotted circles). The yellow circles corresponds to solvent molecules as they rolls along the vdW surface to define the SAS. The yellow edges on the right molecule correspond to the boundaries of spherical polygons dividing the SAS. Structure: chain A from Protein Data Bank entry 3OJ3.



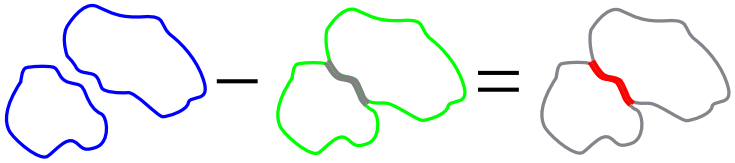
In particular, the SAM is used to define the notion of *solvent accessible surface* (SAS) and *buried surface area* (BSA). The surface of the union of balls has an area which is the sum of spherical polygons defined by their intersection. The surface of two balls intersect in a *circle arc*, and those of three sphere intersect in a point. These arcs and points define spherical polygons whose area can be computed (Figure 2.18. When these balls are the SAM of atoms, the resulting area is the *solvent accessible surface area* (SASA) and represents how much of the molecular surface can interact with solvent molecules. More intuitively, this definition is essentially equivalent to the surface described by the center of a ball, whose radius is that of the solvent molecule, as it rolls along the vdW surface of the molecule. This surface can be used to estimate solvation properties of proteins and to compute surface complementarity measures on interfaces.

The buried surface area of a complex AB is then simply defined as the SASA of both isolated partners from which the SASA of the complex is subtracted:

$$\mathrm{BSA}(AB) = \mathrm{SASA}(A) + \mathrm{SASA}(B) - \mathrm{SASA}(AB)$$

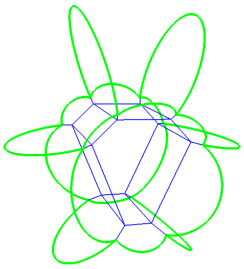
It is the amount of SASA that is lost upon association assuming that both proteins remain perfectly rigid (Fig. 2.19). Although this is rarely true in practice [CGR$^+$13], the BSA has proved to be a valuable tool for the analysis of protein - protein complexes [CJ75, BCRJ04b]. It was also the first parameter to provide a good approximation of dissociation free energies, albeit for rigid association only [HL92, Jan14].

Figure 2.19 Definition of the buried surface area (BSA). Subtracting the solvent accessible surface from the complex (green) to that of the individual partners (blue) results in the buried surface area (red).



The SAM is also relevant for computing the *atomic volume* of molecules. The volume of an union of balls can be computed by defining the *Voronoi diagram* of the balls and computing its volume [Ric74]. Such a Voronoi diagram is a 3-dimensional polyhedron made of *Voronoi cells*, where each cell is associated to a ball, and is the set of points whose nearest neighbor is the center of the corresponding ball (Fig. 2.20). One issue is that the volume of Voronoi cells at the boundary of the diagram is infinite, which was first tackled by adding an artificial layer of solvent [Ric74]. This introduces uncertainties in the computation of the volume, which can be avoided by computing the volume of the *restriction* of Voronoi cells. A restriction is the intersection of the Voronoi cell and the corresponding ball, *i.e.* the total volume is bounded by that of the ball. An algorithm to compute the certified volume of such restrictions has since been published [CKL11]. When applied to a SAM, the volume of balls describes to which extent the atoms are *packed* together. This was used to show that the interface of proteins is tightly packed, *i.e.* similarly to proteins interior and amino-acid crystals. In particular, packing is a proxy for the number of neighbors of a given atom: a tight packing (low volume) indicates many vdW interactions with its neighbors.

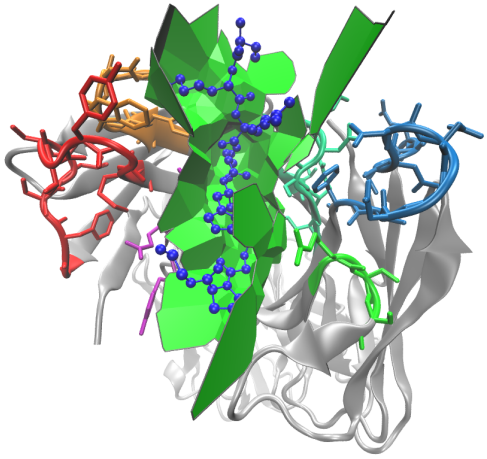
Figure 2.20 Voronoi cell of a point surrounded by eight spheres. Blue edges are the edges of the Voronoi cell and green circle arcs are the intersections of the balls. Figure from [CKL11]



Finally, two approaches have been described to define the interface atoms of a complex. First, atoms losing solvent accessibility (or a large enough proportion of it) are classified as interfacial. This approach has two drawbacks: it requires the definition of a threshold on the amount of SASA lost by an atom, and it can miss interfacial atoms. In effect, the SAM of a *buried* atom (*i.e.* with a SASA of 0Å) may intersect that of another atom on the other partner. Although it makes sense to consider this atom to be at interface for this reason, it would not be classified as such because it has no SAS to be lost.

An alternative interface model, also based on the SAM, was thus designed to correct for these drawbacks [CPBJ06, Caz10]. This approach is based on the α-complex of the SAM of the atoms. This construction is closely related to the Voronoi diagram, except that Voronoi cells cannot extend outside of their respective ball. This means that only the Voronoi cells of atoms whose SAM representation intersect have a (2D-)face in common. It is then possible to keep only facets separating atomic which belong to different partners. These atoms are defined to be interface atoms and the collection of Voronoi facets is the *Voronoi interface* of the complex (Fig. 2.21). Quantities such as the number of connected component and curvature of this Voronoi interface can then be computed. Importantly, this construction can accommodate the presence of interface solvent molecules.

Figure 2.21 Voronoi interface between an Ig and a peptide. Structure: entry 1GGI from the protein data bank.



2.4.2 Structural properties of Ig - Ag complexes

Global properties: typical quantities and associated values

Most studied Ig - Ag complexes involve protein antigens, for which the buried surface area (BSA) can range between 1200 and 2300 Å^2 [SM02, RBCJ05], which is comparable to general protein-protein complexes [BCRJ04b, CJ02].

In terms of thermodynamics, the B cell differentiation and expansion process results in Ig having high affinity for their antigens. In practice, the typical affinity range is from 10^{-7} to 10^{-9} M (section 2.4.1). The theoretical maximum has been predicted to be around 10^{-10}, and such an affinity has been reached using antigen display methods [SM02]. For comparison, K_d values for general protein-protein complexes in the structure affinity benchmark [KMH$^+$11] range from $6.35 \cdot 10^{-4}$ to $2.4 \cdot 10^{-14}$.

Finally, is interesting to note that although Igs need to contact the antigen only, as opposed to TCRs which must contact both the conserved MHC and the variable antigenic peptide, their conserved FRs are also involved at the interface, sometimes contributing up to ∼15% to the BSA [SM02].

Geometrical and topological properties: on shape of the binding site and complementarity

Physico-chemical properties of the interface have been classically used to characterize Ig - Ag complexes (Section 2.3.3). However, geometrical and topological features depict another relevant aspect of molecular interfaces.

A strong shape complementarity between the Ig and Ag sides has been observed, protruding parts of one partner being buried into depressions of the other. This complementarity is however weaker than in the case of homodimers and protease - inhibitor complexes, most probably because, contrary to the latter, no co-evolution occurs between an Ig and its Ag [SM02].

The topography of the binding site has been especially studied in regard to ligand types. Namely, planar shapes are usually related with larger protein antigens, binding sites with a groove are usually found for peptide antigens, and cavity-like ones for haptens. This cavity is often located between the heavy and light chains, deeper than for other antigens [Alm04]. Finally, large antigens usually interact with the edge of the binding sites and the most apical portion of the CDRs [RSWA12]. Intuitively, small ligands tend to be trapped inside the binding site whereas large ones tend to be resting on it.

The role of water molecules at interface has been discovered later, when crystallographic structures of high enough resolution were resolved. It was then observed that stable, or crystallographic, water molecules can engage in hydrogen bonding with both partners, and, by filling the gaps between them, increase their surface complementarity [BBB$^+$94, SM02]. Interestingly, although the force of a van der Walls interaction is weak compared to that of an hydrogen bond , the much larger number of the former compared to the latter results in the van der Walls component dominating the overall effect.

2.4.3 Binding free energy estimation from structural data.

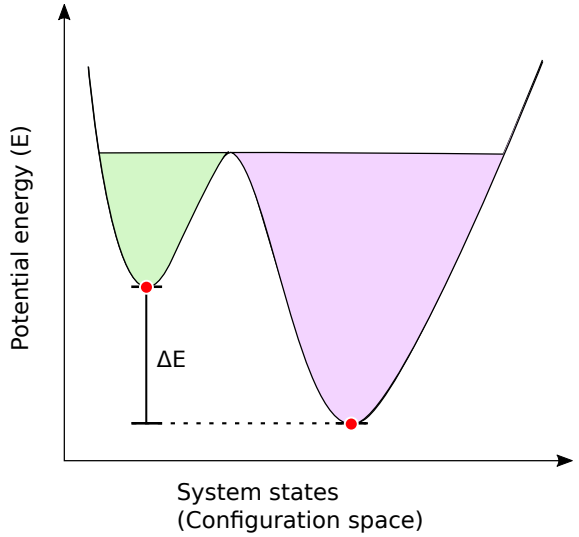
Two main families of models to estimate binding affinities from structural data. First, sampling methods use mathematical approximations of physical phenomena (called physical potentials) to model the potential energy of the system. This results in a *potential energy landscape* (PEL, Fig. 2.22) which assigns a potential energy to a state of the system (*i.e.* conformations of the complex and the surrounding solvent). Estimation of the binding free energy is then done by various methods described hereafter. Second, statistical models instead rely on descriptors of higher-level (*i.e.* not necessarily on the atomic scale) aspects of the binding and use experimental data to fit the model parameters so that the error between estimated and actual free energy is minimized.

Sampling methods: from potential energy to free energy

The computation of the binding free energy relies on the computation of the free energy of the bound state and the unbound state. However, the previous sentence is deceiving: there is not a single bound/unbound state but instead a set of configurations associated to the corresponding state, also called *metastable state*. These are typically basins in the PEL, *i.e.* configuration of the system which, when minimized reach the same local minimum (Fig. 2.22). The computation of the free energy of metastable states relies on integration of the potential energy function, via Boltzmann's factors, over states in these metastable states (Eq. 2.5). Because there is no closed form for these integrals, numerical integration must be used. However, even then it is very difficult to obtain good approximations; in effect, the contribution of a state to the total free energy is related to the negative exponential of its potential energy (Eq. 2.5). It follows that only states with a low potential energy or which are part of a large metastable state will contribute to the estimation. Incidentally, these states make a extremely small part of the configuration space of the metastable state which is sampled to obtain the PEL. Therefore, the sheer dimensionality of this configuration space ($3N$, where N is the number of atoms of the partners, to which the number of atoms of solvent molecules must be added when dealing with explicit solvent) makes it very difficult to sample those states, which contribute the most to the estimation of the integral.

For this reason many different methods have been designed to overcome the difficulty of directly computing the integral over initial and final metastable states. Such methods include thermodynamic integration and potential of mean force methods.

Figure 2.22 A one dimensional potential energy landscape (PEL). Local minima are represented by red dots with basins associated to their metastable states in green and purple respectively. The potential energy difference ΔE is distinct from ΔG because it only takes into account the local minima and not the rest of the basins.



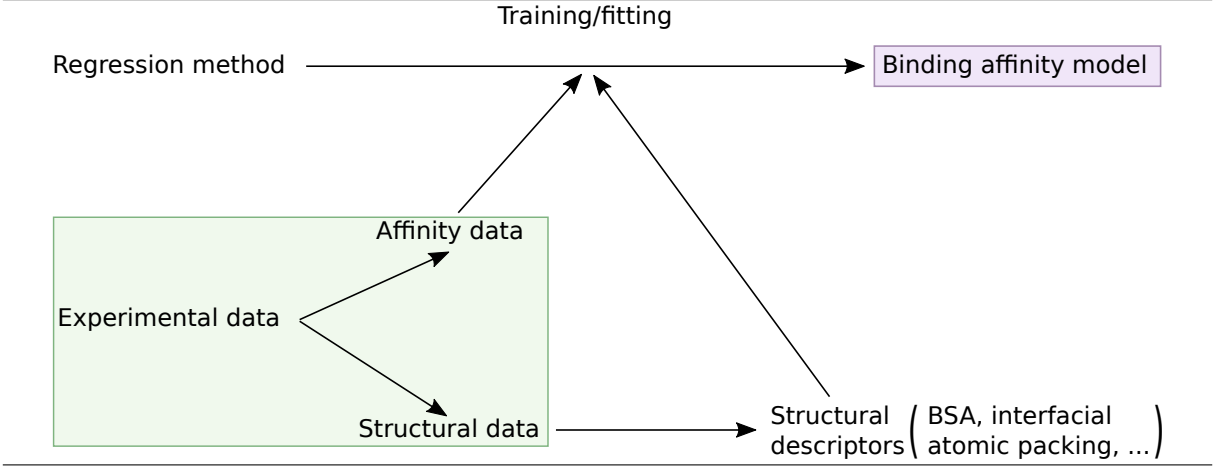
Sampling methods can in principle be very accurate, but they often involve manual tuning. For

instance finding a relevant subset of the reaction coordinates, also called collective variables, in order to make the integration practical is a essential but non-trivial task. Moreover, they require massive amounts of computing power: a molecular dynamics simulation essentially integrates Netwon's equations of motion numerically for each atom. It is therefore necessary for the step of integration to be very small in order to get an acceptable error. The large number of time steps required to get a long enough simulation, multiplied by the number of atoms, and further multiplied by the number of replicates used for averaging of the trajectories makes the number of computation extremely large. On the other hand, sampling methods must sample a very high dimensional space (3^N dimensions, where N is the number of atoms). Since the number of samples required to get a constant density in a space is growing exponentially with this space dimension, a substantial amount of processing power is required to get an acceptable density of samples.

Statistical models: high-level description of binding affinity

Statistical models take a very different (some may say non-physical) approach. There are three pre-requisite for statistical models to be used (Fig 2.23). First, *descriptors* relevant to the binding process must be designed and computed. These are the *independent variables* (also called features) of the model and can be computed from various experimental sources such as crystal structures or nuclear magnetic resonance data. Second, experimental data for which the values of the *dependent variable* to be predicted is known must be available for a number of cases (called the *training set*). In our case, the values are experimental dissociation free energies. The training set are necessary to fit the model parameters. Third, a *regression method* using the independent variables to predict the dependent variable must be selected. This includes ordinary or regularized least squares regression, regression trees, k-nearest neighbors regression, and many others. A regression method along with its parameters is called the *model*. The fitting, also called training is a strategy to adjust the model parameters such that the error between actual and predicted values of the dependent variable is minimized.

Figure 2.23 General workflow for building a statistical model of the binding affinity. Input is boxed in green, output is boxed in purple.



For the first point, several approaches to designing dependent variables have been described in the literature, which can be grouped in three broad families.

First, the simple, ad-hoc approach consists in using a handful of descriptors designed to be intuitively relevant. This usually results in simplistic models with limited prediction abilities. For instance, following the observation that the *buried surface area* (BSA) and binding affinity are correlated [CJ75][BCRJ04a], one of the first models for binding affinity used the polar and apolar BSA [HL92] to build a linear model for affinity. Other models taking into account solvation [KRF+14], interface flexibility [Jan14] and contacts between residues of various types [VB15] have also been described (Section 2.4.6).

Second, the complex, exhaustive-like approach consists in using as many variables as available and feed them to a regularized regression method such as penalized least-squares (*e.g.* LASSO [Tib96]).

Statistical potentials are part of such models as they estimate the distribution of all pairs of amino-acids at a given distance [YGHW13]. Moal and colleagues [MAB11] take another approach by using many descriptors related to various phenomena (*e.g.* surface, solvent, dynamics, electrostatics). The use of regularization is central since models with so many parameters will very likely overfit the training data.

A third class, including our approach (Chapter 3) and that of [ERS14] considers an alternative approach using a restricted set of descriptors to be selected according to how well they predict the binding affinity. We review the most recent statistical models in more details in Section 2.4.6.

Considering the choice of the regression method, one should keep in mind that it is very unlikely that the binding affinity is the result of a process which can be described by a probabilistic model. This means that whatever regression methods will be chosen, it will only be a statistical approximation of the actual process. With this in mind, instead of trying to choose the regression methods leading to the best performances on a given (and necessarily restricted) dataset, it is wiser to choose a simple model (following Occam's razor) or one for which the data has to satisfy the fewest hypotheses (*e.g.* linearity). This approach increases the chances to obtain a model which is able to generalize i.e. to predict new unseen data, as opposed to an model which has been overfitted and can only accurately model its training data. In particular, the use of cross-validation is necessary to ensure that no overfitting is happening but must be complemented by checks on an external test set [GT02].

2.4.4 Terms used by binding affinity models

Both sampling-based and statistical methods share a common point: they require some descriptors derived from structural data. In the first case, it is necessary to build the potential energy function used during sampling. In the second these descriptors are the independent variables used by the regression method. These can either account for the enthalpy term, the entropy term or both.

Force fields / empirical energy functions: atomic-level description of comlexes

To ease the description of these approaches, we introduce the notions of system and state. The *system* consists of the partners and, depending on the representation chosen, the solvent molecules. A *state* of this system is a configuration which can be parametrized in various ways, the most intuitive being the atomic coordinates of the partners (and potentially solvent).

Force fields or empirical energy functions seek to model various types of interactions and physical phenomena at the atomic level in order to calculate the potential energy E of states of the system. The terms which compose a force field are usually categorized as bonded (bond length, bond angle and dihedral angle) and non-bonded (van der Waals (vdW), H-bonds (HB), electrostatics, solvation).

The resulting energy is often a weighted sum of the various terms. For instance:

$$E = E_{\text{bonded}} + E_{\text{non-bonded}} \tag{2.6}$$

with E_{bonded} and $E_{\text{non-bonded}}$ defined as follows:

$$E_{\text{bonded}} = \alpha E_{\text{bond length}} + \beta E_{\text{bond angle}} + \gamma E_{\text{dihedral}} \tag{2.7}$$

$$E_{\text{non-bonded}} = \delta E_{vdW} + \epsilon E_{elec} + \zeta E_{HB} \tag{2.8}$$

For the bonded term, $E_{\text{bond length}}$ is usually a sum over all pairs of atoms of a distance-dependent Morse potential (or a quadratic approximation thereof), $E_{\text{bond angle}}$ is usually a sum over all bond angles of a quadratic function of the angle, and E_{dihedral} is the sum over all dihedral angles of a periodic function of the angle.

For the non-bonded terms, E_{vdW} is usually approximated by the sum over all pairs of atoms of the distance-dependent Lennard-Jones potential, E_{HB} is usually approximated by a sum over pairs of electron donors/acceptors of a distance-dependent potential such as the Morse potential and E_{elec} is usually approximated by the sum of the coulomb potential over all pair of charged atoms.

The weights α to ζ for each term are usually fitted on experimental data for which potential energies are easy to compute *e.g.* small organic molecules.

Force fields typically account for the enthalpic component of the free energy while the entropy is estimated from properties of the resulting potential energy landscape. Such force fields include CHARMM [BBO+83], AMBER [CCB+95] and GROMOS [GB87].

Phenomenological descriptors: high-level description of complexes

For statistical models, phenomenological descriptors are used except in the case of statistical potentials. They are properties corresponding to higher-level concepts. As opposed to force fields, they do not describe every interaction or physical phenomenon at an atomic level but instead aggregate them in biophysically meaningful features. Most of these parameters, describe the morphology of the interface (size, shape, packing properties) and its biochemistry (salt bridges, solvation, hydrogen bonds) [JT96, LCJ99, BGNC09, MDBC12].

A typical example is the buried surface area (BSA), the surface between both partners of the complex which is not solvent accessible [CJ75][BCRJ04a]. The BSA is an approximation of the actual solvent accessible surface which gets buried upon binding, in the limit of rigid association. In practice, it can be very well approximated by the number of interface atoms. The BSA has been divided in polar and apolar constituents to account for their different effect on the binding affinity [HL92]. The internal path length (IPL) has also been proposed as an improvement to this classical quantity [BGNC09]. It takes into account the shape of the interface and note only its size. In particular, it results in higher values for isotropic (round) interface and lower values for anisotropic (stretched) ones.

Other interface properties have been used such as salt bridges, hydrogen bonds and cavities [ERS14], residue conservation [GC05], and hot spots [MFR07]. Hot spots are residues which contribute strongly to the overall binding energy [CW95]. They are experimentally defined as residues whose mutation to alanine results in a variation in ΔG ($\Delta\Delta G$) greater than a cutoff (typically 1-2 kcal/mol). Hot spots can also be predicted using computational alanine scanning, which is similar to experimental alanine scanning except that the free energy change is estimated using force fields [KB02] or classification methods [DGW+13, WLZC12].

Phenomenological descriptors are, however not limited to the interface. For instance, the percentage of polar and charged residue at the non-interacting surface (NIS) have been used for prediction in [KRF+14], and their relationship with solvation investigated in [VKB15].

2.4.5 Ig - Ag specificity

Specificity is another important aspect of the interaction between two molecules. As opposed to the binding affinity, specificity has no formal definition and is not a well-defined thermodynamic quantity. In this respect, it is difficult to discuss specificity without defining precisely what is meant beforehand. A tentative and broad definition of specificity could be:

> How strongly a partner A binds to a partner B compared to other partners B'.

Two facts are immediately obvious in this definition. First, specificity is linked to binding affinity, and, more specifically, binding affinity relative to other partners. Second, "other partners" is extremely vague and could for instance refer to "all hypothetical partners", "all partners from the same class" (be it a superfamily, family or fold) or "all partners differing by a restricted set of mutations". In particular, an antibody which binds to a handful of related antigens could be deemed specific as the number of potential antigen is enormous.

A related notion is that of *ligand type specificity*, *i.e.* whether a molecule binds to a protein, peptide, hapten, nucleic acid, lipid, or sugar. This notion is straightforward to define assuming a criterion to discriminate between proteins and peptides has been specified.

The factors contributing to the specificity of interaction between Ig and Ag have been investigated in several works. Xu and colleagues [XD00] studied the impact of VH CDR3 on the specificity of Igs by studying the responses of mice with a single VH gene to various antigens (protein and haptens). After showing that most V genes are in their germline state, they show that all Ig only bind to the antigen against which they were raised. Showing that differences in VH CDR3 only can determine the affinity of an Ig.

Two related works sought to determine how the shape of the binding site determines the type of ligand [MMT96][LLZ+06]. To this end, they classified recombining sites into different groups corresponding to concave (also called cave-like), moderately concave (crater-like), ridged (canyon-like) and flat (plain-like) recombining sites. Each category is preferentially bound by an antigen type; namely haptens for concave, peptide/carbohydrate/nucleic acid for intermediate and ridged, and proteins for ridged and flat. However, the assignment of ligand types to a given group using this description is far from unambiguous.

The influence of the association of heavy and light chain on the ligand type has also been studied [CMT11]. The authors showed that a set of five residues at the interface between the H and L chains leads to two modes of interactions between them. This affects the shape of the binding site leading to one mode of interaction favoring small antigens, and the other favoring larger antigens.

Finally, two related works [Alm04, RSWA12] have 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. In the first study, the authors show that the number and the location on the recombining site of SDRU were different for Igs binding to proteins, peptides and haptens. In the second study, they also show that the residue distribution at these sites varies with the ligand types.

2.4.6 Review of the latest statistical models for binding affinity prediction.

Multiple studies have sought to predict the affinity of protein-protein complexes. Most, and force field-based methods in particular due to their sheer computational cost, have focused on a limited number of complexes. The following Table 2.2 briefly reviews the latest works which aim at predicting the binding affinity of general protein-protein complexes on a larger scale. These are described in more details in Appendix A.

Reference	Dataset(s)	Model type	#Variables	Feature selection	Cross-validation	Main result
[MAB11]	SAB, seven entries discarded	Unweighted average of the predictions from 4 predictors	94 (200)[a]	Built in the predictors	Leave-one-out	0.77 (q), 1.67 kcal/mol (cross-validated RMSE), 0.55 (PCC*)
[VHPW12]	SAB	Linear regression[b]	9	Various combinations of up to 9 variables are tested	Leave-one-out	0.63 (q), 2.25 kcal/mol (cross-validated RMSE)
[TLY12]	SAB	Partial least squares linear regression	4	Genetic algorithm	Monte-carlo	0.815 (R^2), 0.722 (q^2), 0.693 (r^2_{pred})
[Jan14]	SAB, various entries discarded	Least-squares linear regression	2	None	None	0.32 (PCC), 3.0 kcal/mol (RMSE)
[YGHW13]	DOCKGROUND database + docking decoys	Knowledge-based statistical potential	1092	None	None	0.63 (PCC)
[KRF+14]	SAB, one entry discarded	Least-squares linear regression	3	None	4-fold	0.48 (q)
[ERS14]	SAB, ten complexes discarded	Least-squares linear regression	4	All combinations of up to 7 variables among 13	Leave-one out	0.57 (q)
[VB15]	SAB, 19 entries discarded	Least-squares linear regression	6	Stepwise based on Aikake's information criterion (AIC)	4-fold	0.67 (q), 0.73 (PCC), 1.89 (RMSE, training set)
[MBC15]	SAB, 5 entries discarded	Least-squares regression	2	All combinations of up to 5 variables among 12	Repeated 5-fold cross-validation	0.48, 0.72+ (q)

Table 2.2: **Comparison of recent works on general protein - protein binding affinity prediction** NB: all these results have been obtained on various subsets of the SAB or various datasets altogether. They also use various cross-validation procedures, which makes direct comparisons difficult. The terms used in the rightmost are defined as follows: PCC: Pearson's correlation coefficient. q^2: cross-validated R^2; q: cross-validated correlation coefficient; r^2_{pred}: correlation coefficient on an external test set; RMSE: root mean-squared error. [a]: One predictor uses all features but weights data points instead virtually using all variables. The union of the sets of variables used by the other predictors has 94 members. [b]: Linear regression optimized for correlation (*i.e.* not least squares). *: Correlation coefficient obtained on a mix of training and test sets. +: q obtained on a reduced dataset consisting of high-resolution structures.

2.5 Contributions / Executive summary

The main contributions of this thesis belong to three domains: the prediction of the binding affinity of general protein - protein complexes, the quantitative description of the interaction specificity and binding affinity of Ig - Ag complexes, and the global comparison of immunological repertoires.

2.5.1 Binding affinity prediction

We make a stride towards a better understanding of three core questions related to binding affinity predictions.

The first one relates to the variables and models best suited to perform such predictions. We introduce sparse models relying on 12 variables aiming at capturing enthalpic and entropic changes upon binding.

These models use least-squares linear regression and use subsets of the original pool of variables called *templates*. The best performing templates are selected using a statistical machinery working in two steps. First, they are evaluated using repeated 5-fold cross-validation. This results in a distribution of performance metrics associated to each model (the median absolute error in this case). These distributions are then used by a procedure based on the Kruskal - Wallis test to obtain a set of templates resulting in significantly better performance. The corresponding models are used to estimate binding affinities on a per complex basis, from which an assessment at the dataset level is obtained by reporting the fraction of cases for which K_d is estimated within one, two and three orders of magnitude.

The variables used by these models describe surface areas, packing properties, and their variations at the atomic level, and solvation both at the residue and atomic scale. In particular, one encodes interface size, morphology and packing at once, four encode atomic packing variations upon binding at various locations (interface vs non-interface and buried vs surface), four encode residue-level solvation properties and their variations upon binding, two encode atomic-level solvation, and one encodes interface flexibility upon binding.

Using these variables, we identify *specific models* for subsets of the SAB considered by previous studies, whose performances match or outperform those previously published, in particular for flexible and high resolution cases. Each specific model is also challenged on its non-specific datasets, to highlight the relevance of its variables in handling features specific from these datasets. In particular, this analysis singles out a novel variable, encoding the morphology and the packing properties of the interface, namely properties reminiscent of enthalpy and entropy.

The second question relates to a key difficulty in predicting affinities, namely flexibility. In previous work, flexible cases have been described as the most challenging ones. Using our models, we show that flexibility and prediction hardness do not correlate, and that for flexible cases, a performance almost matching that of the whole SAB can be achieved.

The third one pertains to the quality of predictions. For the whole dataset, we present a model predicting K_d within one and two orders of magnitude for 48% and 79% of cases, respectively. These statistics jump to 62% and 89% respectively, for the subset of the SAB consisting of high resolution structures, a marked improvement over previous work, also stressing the dependence of energies on atomic details.

2.5.2 Ig - Ag binding affinity and specificity.

The difficulty of understanding molecular recognition between proteins in general and antibody - antigens in particular is well known [SM02]. In this thesis, we present novel quantitative analyses for interfaces of Ig - Ag complexes. Using the annotated IMGT/3Dstructure-DB [EKL10], 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, we show how to unambiguously distinguish ligand types using a simple model. Namely, two variables encoding the average BSA per interface atom are computed: one on the Ig side, the other

on the Ag side. These can then be using as input to a classification tree to obtain a cross-validated classification error of 9.6%.

We also show how variables selected for their ability to predict binding affinities in protein - protein complexes (Chapter 3) can also be used to predict binding affinity of Ig - Ag complexes with unprecedented accuracy (median absolute error of 0.878 kcal/mol).

Finally, we develop quantitative models for the contribution of VH CDR3 to binding affinity and interaction specificity. In particular, we assess the respective contribution of CDRs to the binding energy using a descriptor which combines the position at interface and packing of the CDR. We show how this description differs from those using CDR length or BSA only. These results allow us to bridge the gap between various observations (canonical backbone conformations, mutagenesis data, affinity measurements), and to explain the emergence of function from a combination of structural and dynamical properties.

2.5.3 Comparison of antibody repertoires.

The recent advent of high throughput sequencing of RNA in immunology calls for rigorous methods for the analysis of the resulting data. In particular, the sequencing depth provided by the Illumina sequencing technology gives access to sequences with low abundance in the population of transcripts to be sequenced. This raises new questions as characteristics of the immune response can be observed in smaller clonotypes.

We thus present a description of the B cell response to the rhabdovirus VHSV (Viral Hemorrhagic Septicemia Virus) in a fish vaccination model by analyzing the evolution of the VH CDR3 repertoire upon vaccination and challenge. We focus on six pairs of variable and constant genes (VC pairs).

We first characterize the structure, (*i.e.* the heterogeneity from fish to fish) of public, intermediate and private responses in which the VC pairs engage. To this end, we compare the repertoires of pairs of fish using the earth mover distance (EMD). The EMD takes into account both sequence similarity and clonotype size and returns a single value assessing the global similarity of two repertoires. We also compare it to the Morisita-Horn distance, a widely used method with similar purposes, and show that EMD describes responses in finer details.

We also quantify the overlap of large (top) clonotypes between naive, vaccinated and re-infected fish. We consider two levels of detail by looking for both identical sequences and sequences from the same similarity class. This allows us to identify four distinct repertoire behaviors upon vaccination and re-infection among VC pairs. In the first case, naive fish have distinct top clonotypes and responses upon challenge remain distinct; this indicates a private response. In the second case, naive fish share many top clonotypes and more similarity is observed upon challenge; this indicates a public response. In the third case, naive fish also share many top clonotypes but these become distinct upon challenge; this indicates another private response. In the last case, naive fish share many top clonotypes and respond weakly to the challenge.

Finally, we study the representation of small clonotypes in subsamples and quantify the resulting variability when searching for top clonontypes in these subsamples. In particular, that for small clonotype shared among many fish are very sensitive to sampling effects.

2.5.4 Software: Binding affinity prediction modeling

The various programs written during our work on binding affinity prediction were deemed to be of general interest. For this reason, we provide them as a specific-purpose package of the structural bioinformatics library (SBL: `http://sbl.inria.fr/`).

Succinctly, this packages allows one to build and select models of binding affinity prediction in three steps. The first step consists in running various binaries from the SBL, organize their output, parse it and use it to compute various structural descriptors to be used by the models. Because the output is in a standard format (XML), this step can be performed by any other software preferred by the user. The second step consists in generating and evaluating various models using the previously defined descriptors, and reporting which are performing significantly better than the others. The final step consist in using the resulting model(s) to predict new data.

This package, written in python, provides classes to be used in applications, along with scripts to quickly run analyses. This will allow both users with limited knowledge and advanced users to build and select models for binding affinity.

Chapter 3

Protein - protein affinity prediction: High Resolution Crystal Structures Leverage Protein Binding Affinity Predictions

3.1 Introduction

Deciphering the dynamics of macromolecular interactions in general , and those of proteins in particular, is a major challenge as they determine virtually all processes in living organisms. If structural models of complexes shed light on interactions at the atomic level, the formation of a complex and its stability are explained by its binding affinity. Estimating affinities is thus a central step while modeling biological systems, both in the scope of basic research and applications (namely for biological experiments and drug design). Affinities measured by dissociation constants (K_d) span 11 orders of magnitude, a range illustrating the diversity of biological processes and the various binding modes inherent to them [JBC08].

From a modeling perspective, the estimation of affinities relies on structure-based modeling, to bridge the gap between 3D atomic coordinates and thermodynamics. More precisely, consider two species A and B forming a complex AB. The aforementioned dissociation constant K_d is defined by $K_d = [A][B]/[AB]$, and the corresponding dissociation free energy ΔG, in the $c^\circ = 1M$ standard state satisfies

$$\Delta G = -RT \ln K_d/c^\circ = \Delta H - T\Delta S. \tag{3.1}$$

It also illustrates *enthalpy - entropy* compensation phenomenon [MABM10, Dun95], which stipulates that a favorable enthalpic change upon association is accompanied by an entropic penalty (Section 2.4.1).

In theory, estimating a dissociation free energy can be done using free energy calculations methods such as thermodynamic integration, umbrella sampling, or potential of mean forces [FS02, Chi14]. While in principle highly accurate, these methods are extremely demanding in terms of sampling, at the expense of high computational requirements to generate appropriate sampling. They are not suitable to large scale studies, which motivated the development of estimation methods focusing on relevant phenomena.

For large scale protein binding affinity studies, prediction models may be classified into two classes. The first class consists of models using a small number of variables aiming at explaining intuitively important components of the affinity. The second class consists of models using machine learning techniques to select relevant variables among a large set of potential candidates. While using a large number of variables helps to provide a detailed account of chemical properties of amino-acids and atoms. Yet parameterizing such complex models is prone to overfitting, especially given the scarcity of structural data, so that performances on external datasets are often limited. The latest published models from both classes are summarized in Section 2.4.6 and detailed in Appendix A.

Contributions In this work, we make a stride towards a better understanding of three core questions related to binding affinity predictions. The first one relates to the variables and models best suited to perform such predictions. We introduce sparse models relying on 12 variables aiming at capturing enthalpic and entropic changes upon binding. The variables used by these models describe surface areas, packing properties, and their variations at the atomic level, and solvation both at the residue and atomic scale. Using these variables, we identify *specific models* for subsets of the SAB considered by previous studies, whose performances match or outperform those previously published, in particular for flexible and high resolution cases. Each specific model is also challenged on its non-specific datasets, to highlight the relevance of its variables in handling features specific from these datasets. In particular, this analysis singles out a novel variable, encoding the morphology and the packing properties of the interface, namely properties reminiscent of enthalpy and entropy.

The second question relates to a key difficulty in predicting affinities, namely flexibility. In previous work, flexible cases have been described as the most challenging ones. Using our models, we show that flexibility and prediction hardness do not correlate, and that for flexible cases, a performance almost matching that of the whole SAB can be achieved.

The third one pertains to the quality of predictions. For the whole dataset, we present a model predicting K_d within one and two orders of magnitude for 48% and 79% of cases, respectively. These statistics jump to 62% and 89% respectively, for the subset of the SAB consisting of high resolution structures, a marked improvement over previous work, also stressing the dependence of energies on atomic details.

3.2 Estimating Affinities: Datasets and Parameters

3.2.1 Datasets from the Structure Affinity Benchmark

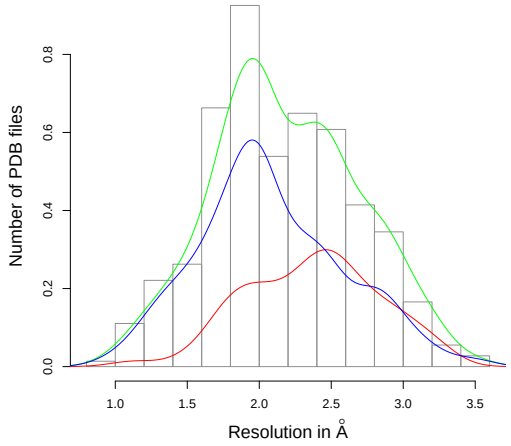
We use the structure-affinity benchmark [KMH$^+$11] (SAB, denoted SAB-A), providing 144 cases with crystal structures for the partners and the complex, as well as an experimentally measured dissociation free energy ΔG . Following previous work, we extract seven *datasets* using a flexibility criterion, and one dataset of high resolution structures (Fig. 3.1). These datasets are (Fig. 3.2):

- SAB-A (139 complexes): all complexes.
- SAB-$R_{1.0}$ (68 complexes): (focus on ridigity, strict threshold) complexes characterized by iRMSD $<$ 1Å [MAB11] ([KRF$^+$14] and [YGHW13] used iRMSD $\leq$ 1Å, 69 complexes).
- SAB-$R_{1.1}$ (78 complexes): (focus on ridigity, intermediate threshold) complexes characterized by iRMSD $<$ 1.1Å [Jan14].
- SAB-$R_{1.5}$ (105 complexes): (focus on ridigity, relaxed threshold) complexes characterized by iRMSD $\leq$ 1.5Å [KRF$^+$14].
- SAB-I (27 complexes): (intermediate complexes) complexes characterized by $1.1\mathring{A} \leq$ iRMSD $\leq$ 1.5Å [Jan14].
- SAB-F_1 (70 complexes): (focus on flexibility, relaxed threshold) complexes characterized by iRMSD $>$ 1Å [YGHW13] ([MAB11] used iRMSD $\geq$ 1Å, 71 complexes)
- SAB-$F_{1.5}$ (34 complexes): (focus on flexibility, strict threshold) complexes with iRMSD $>$ 1.5Å [Jan14][KRF$^+$14].

To which we add:

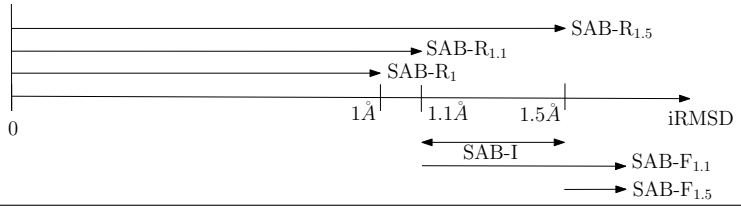
- SAB-A-HR (37 complexes): high resolution complexes from [ERS14]

Figure 3.1 Resolution of the structures in the SAB. The histogram and green kernel density estimation curve are for the whole SAB, the red curve is for the complexes and the blue curve is for the unbound partners. For the whole SAB: Minimum = 0.93 Å, median: = 2.13 Å, average = 2.19 Å, max = 3.5 Å. NB: the high resolution dataset SAB-A-HR retains only entries whose resolution is better than 2.5 Å for both the complex and the individual partners [ERS14].



Curation. To exploit variation of structural parameters between the unbound and bound form, we establish a one-to-one correspondence between the atoms of a partner (from bound to unbound). To cope with cases involving missing residues or atoms, we proceed in two stages. First, we perform an alignment and map residues of the bound and unbound chains. Second, we map atoms of paired residues. We then retain the cases for which at least 80% of atoms are paired. This procedure ruled out two cases, namely 1E6J (78%) and 1ZLI (76%). We also removed three cases (1IQD, 1NSN, 1UUG) for which an upper bound on K_d instead of a proper value is provided for a total of 139 complexes.

Figure 3.2 The various datasets defined from the structure affinity benchmark (SAB), based on iRMSD between the unbound and bound structures.



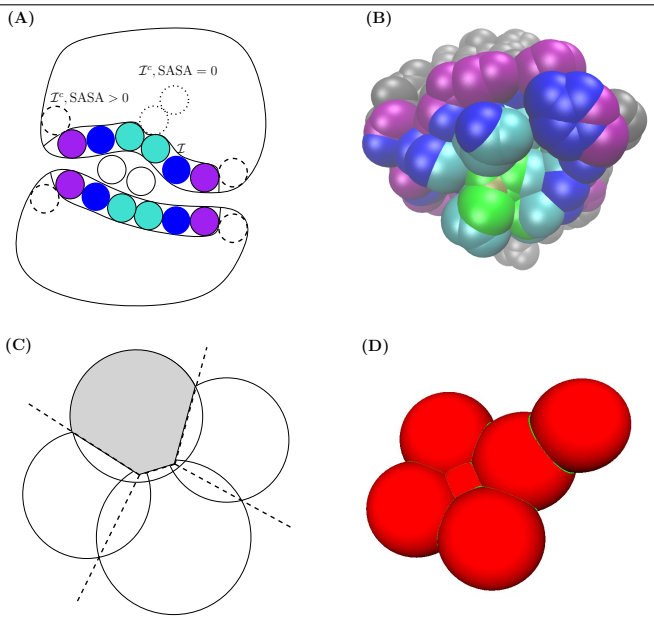
3.2.2 Parameters involved in Affinity Prediction Models

In the sequel, having presented key geometric constructions associated with solvent accessible models of the partners and of the complex, we define parameters meant to capture information on enthalpic and entropic contributions associated with complex formation (Fig. 3.3 and Table 3.1).

Key Geometric Constructions

Surface areas. The solvent accessible surface area (SASA for short) of a solvent accessible model is the sum of the surface areas exposed by the individual atoms. Upon complex formation, the *buried surface area* (BSA) is the surface area of the partners buried at the interface, namely the SASA lost by the

Figure 3.3 Structural parameters used in this work. (A) Labeling the atoms, illustration on a fictitious 2D complex. The binding patch on each partner consists of one layer of atoms ($\mathcal{I}$, colored solid balls), as identified by a Voronoi interface model [CPBJ06, LC10]. The non interface atoms ($\mathcal{I}^c$) are split into those which retain solvent accessibility (SASA > 0, dashed balls), and those which do not (SASA $= 0$, dotted balls) **(B)** Each interface atom is assigned an integer, its shelling order, equal to the smallest number of atoms traveled to reach an exposed non interface atom, *i.e.* an atom belonging to $\mathcal{I}^c$ and with SASA > 0 (in grey) [BGNC09]. **(C,D)** The volume of an atom is defined as the volume of the intersection between its ball in the solvent accessible model, and its Voronoi cell [CKL11], a quantity well defined even if the atom retains solvent accessibility. The packing of this atom is the inverse of this volume. Practically, interfaces and binding patches are computed with `Vorshell`[LC10], while atomic surface areas and volumes are computed with `Vorlume`[CKL11]. Both programs are available from the Structural Bioinformatics Library (SBL), see `http://sbl.inria.fr`.



individual atoms. This quantity has long been known as the simplest and most descriptive parameter of specific protein interfaces [BCRJ04a].

Voronoi interfaces and their shelling order (SO). In describing a protein - protein interface, various parameters are of interest beyond the mere list of atoms, namely its shape (*e.g.* elongated vs isotropic), its partition into a core and a rim, its curvature, or its number of patches. A parameter free *Voronoi interface model* encapsulating all these parameters into a single construction, the α-complex derived from the Voronoi (power) diagram of the atoms, has been proposed [CPBJ06, LC10]. In a nutshell, define the *restriction* of an atom as the intersection between its ball in the solvent accessible model and its cell in the Voronoi diagram. The Voronoi interface identifies pairs of neighboring restrictions, such that each pair involves either two different partners or a partner and the interfacial solvent. The atoms found in at least one such pair are denoted $\mathcal{I}$ and their complement $\mathcal{I}^C$. This Voronoi-based model was instrumental to show that the interface may involve atoms which do not lose solvent accessibility, and also to stress the role of water mediated contacts[CPBJ06]. We note in passing that the exposed atoms in the set $\mathcal{I}^C$ form the *non interacting surface* (NIS) [KRF$^+$14].

Consider the BSA, and more specifically the atoms of one partner contributing to the BSA. The exposed surface of the atoms contributing to the BSA define a *binding patch* (patch for short) [BGNC09]. The *shelling order* (SO) of an atom from a patch is its least distance, counted in integer steps, to the nearest atom from the NIS. That is, the atoms on the border of the patch have a SO of 1 and the remaining ones have a $SO > 1$ (Fig. 3.3(B)). Thus, the SO generalizes core-rim models [JBC08], since the rim corresponds to $SO = 1$, and the core to $SO > 1$.

Table 3.1 Parameters used to estimate binding affinities. Atomic level parameters: IVW-IPL, SVD_SO1, SVD_SOGT1, SVD_NI_B, SVD_NI_E, ATOM_SOLV, POLAR_SASA; Residue level parameters: NIS^{polar}, $\text{NIS}^{charged}$, ΔNIS^{polar}, $\Delta\text{NIS}^{charged}$; Interface level parameter: iRMSD. The acronyms read as follows (see text for details): **S**um of **V**olume **D**ifferences; **S**helling **O**rder; **I**nverse **V**olume **W**eighted; **I**nternal **P**ath **L**ength; **N**on **I**nteracting **B**uried/**E**xposed; **N**on **I**nteracting **S**urface; **S**olvent **A**ccessible **S**urface **A**rea;

$$\Delta\text{-vol}(a) = \text{volume\_bound}(a) - \text{volume\_unbound}(a). \tag{3.2}$$

$$\text{IVW-IPL} = \sum_{a\in\mathcal{I}} \frac{\text{SO}(a)}{\text{volume\_bound}(a)} \tag{3.3}$$

$$\text{SVD\_SO1} = \sum_{a\in\mathcal{I}, SO(a)=1} \Delta\text{-vol}(a) \tag{3.4}$$

$$\text{SVD\_SOGT1} = \sum_{a\in\mathcal{I}, SO(a)>1} \Delta\text{-vol}(a) \tag{3.5}$$

$$\text{SVD\_NI\_B} = \sum_{a\in\mathcal{I}^C, \text{SASA}(a)=0} \Delta\text{-vol}(a) \tag{3.6}$$

$$\text{SVD\_NI\_E} = \sum_{a\in\mathcal{I}^C, \text{SASA}(a)>0} \Delta\text{-vol}(a) \tag{3.7}$$

$$\text{NIS}^{polar} = \frac{\#\text{solvent accessible polar residues}}{\#\text{solvent accessible residues}} \tag{3.8}$$

$$\text{NIS}^{charged} = \frac{\#\text{solvent accessible charged residues}}{\#\text{solvent accessible residues}} \tag{3.9}$$

$$\Delta\text{NIS}^{polar} = \text{NIS}^{polar}_{bound} - \text{NIS}^{polar}_{unbound} \tag{3.10}$$

$$\Delta\text{NIS}^{charged} = \text{NIS}^{charged}_{bound} - \text{NIS}^{charged}_{unbound} \tag{3.11}$$

$$\text{ATOM\_SOLV} = \sum_{a\in\mathcal{I}^C} SASA(a)\cdot\sigma(a) \tag{3.12}$$

$$\text{POLAR\_SASA} = \sum_{a\in\mathcal{I}^C \text{ and } \sigma(a)<0} SASA(a) \tag{3.13}$$

$$\text{iRMSD} = \text{Interface RMSD} \tag{3.14}$$

Atomic packing properties. Early models to assess atomic packing properties resorted to the volume of Voronoi cells [GR01], preferably using the power diagram of the atoms instead of the Euclidean Voronoi diagram [BY98], since different atomic radii are accommodated. However, the Voronoi cell of an atom located on the convex hull of the protein (or complex) is unbounded. To avoid boundary effects, we focus in the sequel on the aforementioned atomic restrictions, whose volume can be computed accurately [CKL11]. That is, denoting volume_bound(a) (resp. volume_unbound(a)) the volume of the Voronoi restriction of an atom a in the bound form (resp. unbound form), the difference between these quantities defines the volume variation of this atom (Eq. (3.2)).

Partners: Enthalpic Contributions

Local interactions. The BSA alone does not account for the interface geometry, as the same surface area may be obtained for by morphologies as diverse as a perfectly isotropic patch, or a long and skinny patch, letting alone curvature. The obliviousness to interface morphology is intuitively detrimental, since morphology relates to the cooperativity of phenomena inherent to non-bonded interactions. To take into account such morphological features, a weighted average of atomic shelling orders, called the *internal path length* (IPL) was defined from the shelling order [BGNC09] [1]. The IPL has been shown to improve the analysis of correlations between interface morphology against conserved residues and interfacial solvent dynamics [BGNC09].

In terms of binding energies, a limitation of IPL is that the SO of an atom does not account for the atomic environment of this atom–that is two atoms with identical SO may be located in a dense and loose environments respectively. This is detrimental since a dense packing is likely to favor local interactions, in particular van der Waals interactions. Since a packed interface is more likely to result in a high affinity,

[1] To be precise, $\text{IPL} = \sum_{a\in\mathcal{I}} \text{SO}(a)$. Note that replacing the SO of each atom by one results in the number of interface atoms, which is known to correlate with BSA for rigid cases [KMH+11].

the shelling order is weighted by the inverse of the volume, yielding the *inverse volume-weighted internal path length* (Eq. (3.3)).

Partners: Entropic Contributions

Assessing entropic variations requires taking several components into account, in particular configurational entropy and vibrational entropy. Large conformational changes yielding structured elements correspond to entropic penalties, and can be assessed using the interface root mean square deviation (iRMSD). In the sequel, we refine this measure using atomic packing properties.

Packing properties. A closely packed environment yields favorable interactions by increasing the number of neighbors. But it also entails an entropic penalty for that atom, illustrating the classical enthalpy - entropy compensation, which holds in particular for biological systems involving weak interactions [Dun95, CM13]. We therefore use our atomic volumes and their variations upon binding (Eq. (3.2)) to model both the interaction energy and the entropic changes upon binding.

To model entropic changes, we resort to volume variations. We do so by considering four categories of atoms. For interface atoms, we define two groups, those found on the rim ($\mathcal{I}, SO = 1$), retaining solvent accessibility, and the remaining ones ($\mathcal{I}, SO > 1$). Likewise, for the set of non interface atoms, we distinguish between those retaining solvent accessibility ($\mathcal{I}^C$ and SASA > 0 in the complex), and those which do not ($\mathcal{I}^C$ and SASA $= 0$ in the complex). Adding up volume variations for these four categories of atoms yields the following four *Sum of Volumes Differences (SVD)* parameters, namely SVD_SO1 ($\mathcal{I}, SO(a) = 1$; Eq. (3.4)), SVD_SOGT1 ($\mathcal{I}, SO(a) > 1$; Eq. (3.5)), SVD_NI_B ($\mathcal{I}^C$, SASA$(a) = 0$; Eq. (3.6)), SVD_NI_E ($\mathcal{I}^C$, SASA$(a) > 0$; Eq. (3.7)).

Solvent Interactions and Electrostatics

The interaction between a protein molecule and water molecules is complex. In particular, the exposition to the solvent of non polar groups hinders the ability of water molecules to engage into hydrogen bonding, yielding an entropic loss for such water molecules. To account for these effects, we use the fractions of charged and polar a.a. on the non interacting surface [KRF+14], respectively denoted NIS^{polar} (Eq. (3.8)) and NIS^{charged} (Eq. (3.9)). We also use the variation of these quantities to account for conformational changes upon binding, yielding the quantities ΔNIS^{polar} (Eq. (3.10)) and ΔNIS^{charged} (Eq. (3.11)).

To challenge a.a. terms with their atomic counterparts and see which ones are best suited to perform affinity predictions, we also included the atomic solvation energy from Eisenberg et al [EWY89], describing the free energies of transfer from 1-octanol to water per surface unit (Å^2). The corresponding variable, ATOM_SOLV, is a weighted sum of atomic solvent accessible surface areas (Eq. (3.12)), and may be seen as the atomic-scale counterparts of NIS^{charged} and NIS^{polar}.

Finally, we include an intermediate-grained description of the non-interacting surface which consists in the atomic-wise polar area of the complex. The corresponding term, POLAR_SASA (Eq. (3.13)), is also a weighed sum of exposed areas.

3.2.3 Parameters Computation

To compute the atoms at interface along with their shelling order, packing and volume, we use the application `sbl-vorshell-bp-ABW-atomic.exe` from the Structural Biology Library (SBL) [**?**], see `http://sbl.inria.fr`. Contacts mediated by water molecules are included because crystallographic water molecules are biologically relevant [RBCJ05].

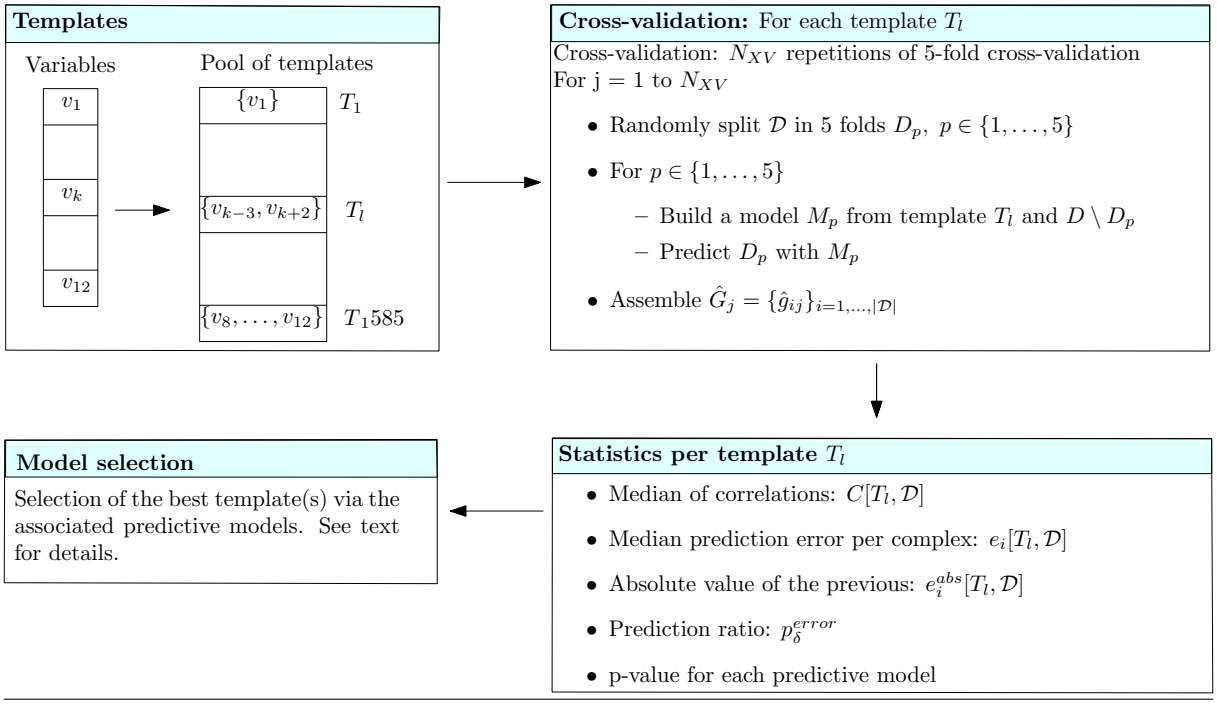
To compute the solvent accessible atoms of the molecules, we use the application `sbl-vorlume-pdb.exe` software [CKL11], also from the SBL [**?**]. In that case, water molecules are not considered since they contribute to the protein surface solvation as much the bulk solvent.

3.2.4 Statistical Methodology

In the sequel, we explain how to predict ΔG of complexes from a dataset $\mathcal{D}$. Estimation is performed on a per complex basis, from which performances at the whole dataset level will be derived. Our predictions rely on three related concepts defined precisely hereafter (see also Fig. 3.4):

- Template: a fixed set of variables from $\mathcal{V}$,
- Model: a linear model consisting in a template plus the associated coefficients. As we shall see, such models are associated with cross-validation folds.
- Predictive model for $\mathcal{D}$: the machinery returning one binding affinity estimate $\hat{g}_i$ per complex from $\mathcal{D}$, using N_{XV} repetitions of the k-fold cross validation.

Figure 3.4 Running binding affinity predictions for a dataset $\mathcal{D}$ *i.e.* a subset of the structure affinity benchmark: graphical outline of the statistical methodology. (Templates) From the pool of variables, templates are generated. **(Cross-validation)** Each template undergoes a number N_{XV} of repetitions of 5-fold cross-validation, yielding one binding affinity prediction per complex for each repetition. **(Statistics)** Various statistics are computed to assess the performances yielded by the predictive model associated to each template. **(Model selection)** Predictive models are compared, and the best ones selected.



Templates. Denote $\mathcal{V}$ the pool of twelve variables specified by Eq. (3.3) to (3.13) (Table 3.1), plus the iRMSD defined in the SAB. Let a *template* be a set of variables, *i.e.* a subset of $\mathcal{V}$. To define parsimonious templates from the set $\mathcal{V}$, we generate subsets of $\mathcal{V}$ involving up to at most five variables–an upper bound dictated by the fact that beyond five variable, the performance of the corresponding best predictive model starts to decrease (Fig. 3.5). This defines a pool of templates $\mathcal{T} = \{T_1, \ldots, T_{1585}\}$ [2].

Cross-validation. In the following a *model* is associated to both a template $T_l \in \mathcal{T}$ and a dataset $\mathcal{D}$ from the SAB. More precisely, a model refers to a linear model, *i.e.* the variables of the template plus the associated coefficients.

[2]Since we have 12 variables, one has $\sum_{k=1}^{5} \binom{12}{k} = 1585$.

Practically, models are defined during k-fold cross-validation (with $k = 5$), and a number N_{XV} (=10000) of repetitions (Fig. 3.4). Consider one repetition, which thus consists of splitting at random $\mathcal{D}$ into 5 subsets called folds. For one fold, a linear model associated with T_l is trained on 4/5 of the dataset $\mathcal{D}$, and predictions are run on the remaining 1/5 of complexes. Processing the five folds yields one repetition of the cross validation procedure, resulting in one prediction $\hat{g}_{ij}$ for the ΔG_i of each complex. The set of all predictions in one repeat, say the jth one, is denoted

$$\hat{G}_j = \{\hat{g}_{ij}\}_{i=1,\dots,|\mathcal{D}|}. \tag{3.15}$$

Note again that these predictions stem from k linear models associated with T_l, namely one per fold.

Computing correlation and prediction errors for repeated cross-validation For a given predictive model, our validation protocol results in N_{XV} predictions for each complex. This can be seen as a $139 \times N_{XV}$ matrix $\hat{G}$ where each entry $\hat{g}_{ij}$ is the prediction for complex i obtained at repetition j. From the experimental values ΔG, there are therefore two ways to get a single value for the correlation and prediction error per complex.

As a first option, one can agglomerate all N_{XV} predictions into a single value by taking their median:

$$\hat{g}_i = \text{median}_j\ \hat{g}_{ij}. \tag{3.16}$$

Then it is straightforward to compute the correlation between $\{\Delta G_i\}$ and $\{\hat{g}_i\}$:

$$C[T_l, \mathcal{D}] = \text{Corr}(\{\Delta G_i\}, \{\hat{g}_i\}) \tag{3.17}$$

and the prediction error for complex i:

$$e_i[T_l, \mathcal{D}] = \Delta G_i - \hat{g}_i \tag{3.18}$$

As a second option, one can take the median of the correlations (resp. prediction errors) over the repetitions. Let $Corr_j$ be the correlation coefficient associated with repetition j, *i.e.* the correlation between $\{\Delta G_i\}$ and $\{\hat{g}_{ij}\}$ for a given j. This results in Eqs. 3.19 and 3.20.

$$C[T_l, \mathcal{D}] = \text{median}_j\ Corr_j \tag{3.19}$$

$$e_i[T_l, \mathcal{D}] = \text{median}_j(\Delta G_i - \hat{g}_{ij}) \tag{3.20}$$

We choose the second method because it makes more sense to us to compute median over statistics than over predictions. Moreover, for a given complex i, the ordering of values $\hat{g}_{ij}$ and $\Delta G_i - \hat{g}_{ij}$ is the same. We therefore have that $\text{median}_j(\Delta G_i - \hat{g}_{ij}) = \Delta G_i - \hat{g}_i$. For the correlation, the two methods give very similar values and give the highest value to the same predictive model. (Fig. 3.6).

Statistics per template. Considering one cross-validation repetition, we define the correlation $Corr_j$ as the correlation between the experimental values $\{\Delta G\}$ and the predictions $\hat{G}_j$. An overall assessment of the template T_l using the N_{XV} repetitions is obtained by the *median of correlations* (Eq. 3.19). For a complex, we define the binding affinity *prediction* $\hat{g}_i$ (Eq. 3.16 as the median across repetitions. Likewise, the *median prediction error* is defined by Eqs 3.20 and 3.22.

$$e_i \equiv e_i[T_l, \mathcal{D}] = \text{median}_j(\Delta G_i - \hat{g}_{ij}), \tag{3.21}$$

and the *median absolute prediction error* by:

$$e_i^{\text{abs}} \equiv e_i^{\text{abs}}[T_l, \mathcal{D}] = \text{median}_j(|\Delta G_i - \hat{g}_{ij}|). \tag{3.22}$$

Using this latter value, we define the *prediction ratio* p_δ^{error} as the percentage of cases such that the dissociation free energy is off by a specified amount δ:

$$p_\delta^{\text{error}} = \%\text{cases in } \mathcal{D} \text{ such that } e_i^{\text{abs}}[T_l, \mathcal{D}] \leq \delta. \tag{3.23}$$

In particular, setting δ to 1.4, 2.8 and 4.2 kcal/mol in the previous equation yields cases whose K_d is approximated within one, two and three orders of magnitude respectively.

Finally, a permutation test yields a p-value for each predictive model [PS10]. In a nutshell, the rationale consists of generating randomized datasets by shuffling their ΔG values. Then, one computes a performance criterion for each such dataset, from which the p-value is inferred (Algorithm 1).

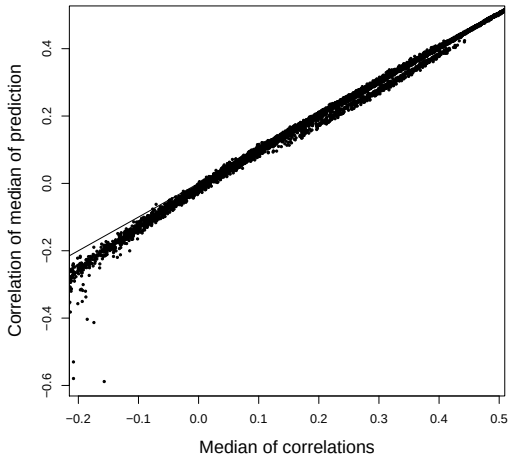
Model selection. Define the best predictive model as the one maximizing the median correlation $C[T_l, \mathcal{D}]$ (Eq. (3.19)), called the *performance criterion* for short in the sequel.

We wish to single out the best predictive models, *i.e.* those that cannot be statistically distinguished from the best predictive model, as just defined.

To single out such models, observe that to compare two predictive models M_{T_l} and $M_{T_{l'}}$, a univariate two-sample test suffices to check whether the two sets of performances (one per model) obtained for the N_{XV} repetitions come from the same distribution (the null hypothesis H0), or whether one dominates the other. In an analogous spirit and since we are handling a pool of predictive models $\mathcal{T}$, we wish to identify within $\mathcal{T}$ a subset of predictive models whose distribution cannot be distinguished from the best predictive model. To this end, we decompose the predictive models as $\mathcal{T} = \mathcal{T}_1 \cup \mathcal{T}_2$ such that (i) the best predictive model is in $\mathcal{T}_1$, (ii) in comparing two predictive models from $\mathcal{T}_1$, one does not reject H0, and (iii) in comparing one predictive model from $\mathcal{T}_1$ against one predictive model from $\mathcal{T}_2$, one rejects H0. The predictive models in $\mathcal{T}_1$ are called the *specific models* for the dataset $\mathcal{D}$. The corresponding procedure is based on the Kruskal-Wallis test (Algorithm 2). The p-value threshold is set to $\alpha = 0.01$.

We also use the eight datasets to define the *best overall predictive model*. To this end, we sorted the models using the aforementioned performance criterion and took the model with lowest median rank among all datasets. This yields the predictive model 9 in the sequel.

Figure 3.6 Comparison between two ways of computing the correlation for a given predictive model over multiple repetitions. Median of correlations: for each of the N_{XV} repeats, compute the correlation between the predictions and experimental affinities. Take the median of these predictions for each complex. **Correlation of median of predictions**: compute a single prediction per complex as the median of all N_{XV} predictions. Compute the correlation between those predictions and the experimental affinities. The values of all predictive models tested on all datasets have been aggregated on this figure. The correlation between both methods is 0.997 with a median absolute difference of 0.005. Moreover, both measures are maximal for the same predictive model on all datasets



Algorithms

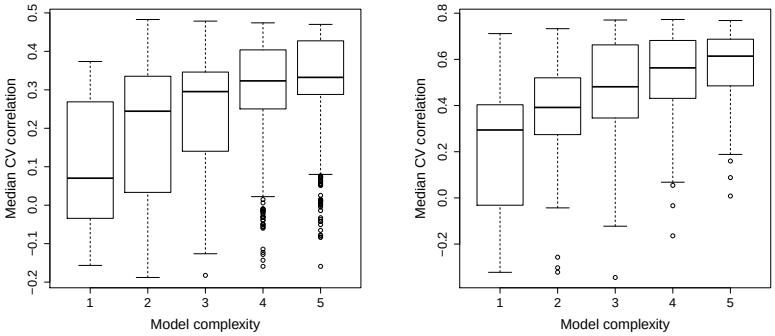
Algorithm 1 Computing a permutation p-value for a binding affinity predictive model specified by a template T_l. The p-value is based on a permutation test [PS10], which uses the prediction performances obtained on random datasets, each such dataset being obtained by permuting the dependent variable (*i.e.* the affinity) over the dataset.

Require: $\mathcal{D}$: dataset; T_l: a template; p_{T_l}: a performance criterion for Tx_l; $N_{\text{perm.}}$: number of repeats
  for $q \in \{1 \dots N_{\text{perm.}}\}$ **do**
    Randomly permute the dependent variable in $\mathcal{D}$ (here the affinity) to obtain $\mathcal{D}_q^{perm}$
    Perform 5-fold cross-validation of linear models using the variables in T_l on $\mathcal{D}_q^{perm}$
    Store the performance criterion in $p_{T_l}^{perm}$
  Report the approximate p-value for T_l to be $\frac{B+1}{N_{\text{perm.}}+1}$, with B the number of elements in $p_{T_l}^{perm}$ which are more extreme than p_{T_l}.

Algorithm 2 Model selection: identifying specific predictive models for a dataset $\mathcal{D}$. The algorithm returns the index of the last predictive model which cannot be distinguished from the best ones given their performance criterion distribution. It is assumed that the predictive models are sorted in non-decreasing order by their median performance criterion. In short, the algorithm executes a binary search, shrinking the interval by its end when there is a significant difference between the predictive models in the interval, and expanding it by its end when there is no difference. All shrink/expand events are applied at the end of the interval to only keep the best predictive models in the final set. Storing the smallest upper bound encountered so far and stopping when it is equal to the upper bound ensures that the algorithm finishes.

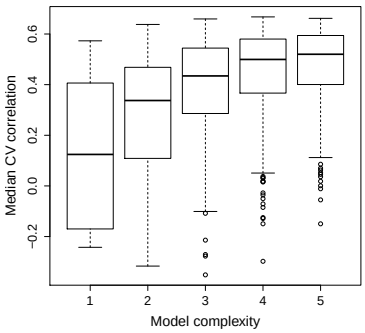
Require: $P = \{\mathcal{P}_{T_l},\ l \in \{1 \dots\ 1585\}\}$: the set of distributions of the performance criterion for each template T_l, sorted by non-decreasing median value; cutoff: a cutoff for the p-value of the Kruskal - Wallis test.
  start:= 0
  end := $|P|$
  while TRUE **do**
    $P_{min} := \{\mathcal{P}_{T_l},\ i \in \{\text{start}, \dots, \text{end}\}\}$
    Perform Kruskal - Wallis' test on P_{min}. Store the p-value in p.
    if $p <$ cutoff **then**
      `## Shrink toward best predictive models`
      end := $\lceil |P_{min}|/2 \rceil$
    else
      `## Expand toward worse predictive models`
      tmp := start
      start := end
      end := end + $\lfloor$(end − start)/2$\rfloor$
    if end $> |P|$ **then**
      `## All predictive models are equivalent`
      return($|P|$)
    if start = end **then**
      `## the shrinking / expanding process has converged`
      if $p \geq$ cutoff **then**
        `## The final pivot is part of the similar distributions`
        return(end)
      else
        `## The final pivot is part of the outliers`
        return(end −1)
    if end = 1 **then**
      `## Only one remains (after a sequence of shrinkings only)`
      return(end)

Figure 3.5 Predictive model complexity versus median correlation C_V between predicted and experimental values.

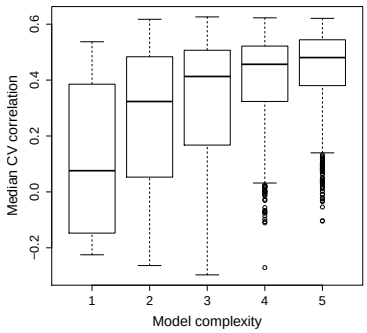


SAB-A

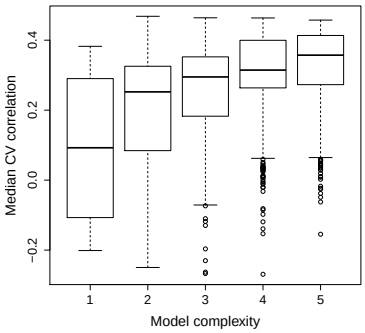
SAB-A-HR



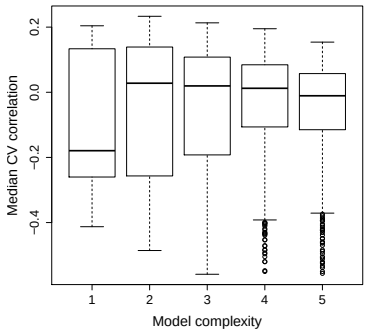
SAB-$R_{1.0}$



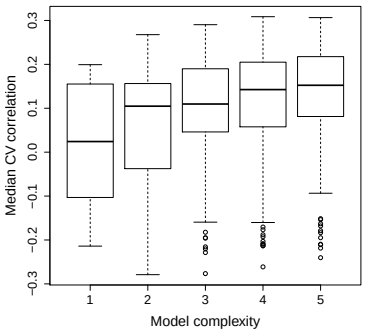
SAB-$R_{1.1}$



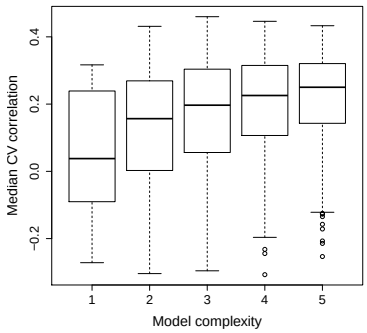
SAB-$R_{1.5}$



SAB-I



SAB-F_1



SAB-$F_{1.5}$

3.3 Results

3.3.1 Specific Predictive Models

Upon applying the methods described in Section 3.2.4, a single best specific predictive model was obtained for each dataset.

▶ **Predictive Model: 1. Obtained for dataset(s): SAB-A** 2 variables, p-value ≤ 0.0001.

$$\Delta G = \alpha + \beta \cdot \text{IVW-IPL} + \gamma \cdot \text{NIS}^{charged} \tag{3.24}$$

▶ **Predictive Model: 2. Obtained for dataset(s): SAB-A-HR** 4 variables, p-value ≤ 0.0001.

$$\Delta G = \alpha + \beta \cdot \text{SVD\_SOGT1} + \gamma \cdot \text{SVD\_NI\_B} + \epsilon \cdot \text{NIS}^{charged} + \zeta \cdot \text{ATOM\_SOLV} \tag{3.25}$$

▶ **Predictive Model: 3. Obtained for dataset(s): SAB-R$_{1.0}$** 4 variables, p-value ≤ 0.0001.

$$\Delta G = \alpha + \beta \cdot \text{iRMSD} + \gamma \cdot \text{IVW-IPL} + \delta \cdot \text{SVD\_SO1} + \epsilon \cdot \text{NIS}^{charged} \tag{3.26}$$

▶ **Predictive Model: 4. Obtained for dataset(s): SAB-R$_{1.1}$** 3 variables, p-value ≤ 0.0001.

$$\Delta G = \alpha + \beta \cdot \text{IVW-IPL} + \gamma \cdot \text{SVD\_SO1} + \delta \cdot \text{NIS}^{charged} \tag{3.27}$$

▶ **Predictive Model: 5. Obtained for dataset(s): SAB-R$_{1.5}$** 2 variables, p-value ≤ 0.0001.

$$\Delta G = \alpha + \beta \cdot \text{IVW-IPL} + \gamma \cdot \text{NIS}^{polar} \tag{3.28}$$

▶ **Predictive Model: 6. Obtained for dataset(s): SAB-I** 2 variables, p-value ≤ 0.090.

$$\Delta G = \alpha + \beta \cdot \text{SVD\_NI\_B} + \gamma \cdot \Delta\text{NIS}^{polar} \tag{3.29}$$

▶ **Predictive Model: 7. Obtained for dataset(s): SAB-F$_1$** 4 variables, p-value ≤ 0.0091.

$$\Delta G = \alpha + \beta \cdot \text{IVW-IPL} + \gamma \cdot \text{SVD\_NI\_E} + \delta \cdot \text{NIS}^{charged} + \epsilon \cdot \text{ATOM\_SOLV} \tag{3.30}$$

▶ **Predictive Model: 8. Obtained for dataset(s): SAB-F$_{1.5}$** 3 variables, p-value ≤ 0.0054.

$$\Delta G = \alpha + \beta \cdot \text{IVW-IPL} + \gamma \cdot \text{SVD\_SO1} + \delta \cdot \text{ATOM\_SOLV} \tag{3.31}$$

▶ **Predictive Model: 9. Obtained for dataset(s): All datasets** 3 variables, p-value ≤ 0.0001 for SAB-A, SAB-A$_h r$, SAB-R$_{1.0}$, SAB-R$_{1.1}$, SAB-R$_{1.5}$; ≤ 0.6949 for SAB-I; ≤ 0.0158 for SAB-F$_1$; $\leq$,0.0089 for SAB-F$_{1.5}$.

$$\Delta G = \alpha + \beta \cdot \text{IVW-IPL} + \gamma \cdot \text{SVD\_SO1} + \delta \cdot \text{NIS}^{charged} \tag{3.32}$$

3.3.2 Specific predictive Models yield Enhanced Correlations...

Recall that a *dataset* can be the SAB or a subset of the SAB defined by bounds on the iRMSD or the resolution of complexes and partners. In the sequel, we analyze the performances of predictive models, as defined in the previous section.

Interestingly, a single predictive model is significantly better than the others for all datasets. These predictive models are all statistically significant with a p-value smaller than 0.01, except for the one associated with the dataset SAB-I, therefore omitted from subsequent analysis.

In terms of correlations between estimates and ΔG (Table 3.2, 5-fold cross validation), our specific predictive models outperform previous works in 5/8 cases. For two of the three remaining cases, the top correlation is provided by the complex model from [MAB11], which we estimated to use 94 variables

(Section 2.4.6). For the remaining one, [ERS14] provides the best results with a seven variables model. Unfortunately, the corresponding variables are not specified.

In terms of correlation values themselves, three facts emerge. First, the predictive model specific of the high resolution model dataset yields a remarkable correlation of 0.77. Second, for flexible datasets, satisfactory performances are observed, which is unexpected since such cases are generally considered as the most challenging ones for affinity predictions. In particular, for flexible cases characterized by an interface iRMSD larger than 1.5Å, a correlation of 0.46 is obtained, a value comparable to that of the whole dataset, namely 0.48. Finally, the best overall predictive model, when challenged by individual datasets, shows performances comparable to those of their specific predictive models with maximum drop in correlation of 0.06. This is a clear assessment of its robustness.

3.3.3 . . . and Improved Predictions on a per Complex Basis

The correlation between predictions and ΔG provides a global performance assessment of a predictive model for a dataset. To gain insights at the individual complex level, we use the individual predictions $\hat{g}_{ij}$. Using these individual predictions, we compute the prediction ratio (Eq. 3.23) for $\delta = 1.4, 2.8$ and 4.2 kcal/mol, respectively, yielding the fraction of cases for which K_d is predicted within one, two and three orders of magnitude. Three striking facts emerge from Table 3.3.

First, the merits of our specific predictive models as well as those of the best overall predictive model clearly emerge. As a quantitative measure, we collect the min and max prediction ratios for the aforementioned three values of δ, yielding a three pairs min-max percentages. For our best overall predictive model, one gets 44-57%, 74-86% and 91-95% within one, two and three orders of magnitude. In contrast, the intervals for [KRF$^+$14] are 46-51%, 68-83% and 85-95%, and those for [Jan14] are 22-44%, 57-73% and 85-93%. Collecting now the min and max prediction ratios of the specific predictive models on their specific datasets, one gets 46-62%, 78-89%, 85-97%. Thus, for the whole SAB, both the specific predictive model and the best overall predictive model yield improved performances.

Second, the prediction ratios of the predictive model specific of the high resolution dataset turn out to be 62%, 89% and 97% within one, two and three orders of magnitude, an outstanding performance.

Third, concerning the flexible datasets, considered as the most challenging ones in previous studies, predictive model 7 (dataset SAB-F$_1$) and predictive model 8 (dataset SAB-F$_{1.5}$) reach performances comparable to those obtained on the whole SAB, namely $p_{1.4}^{\text{error}}$ values of 50% and 50% respectively, instead of 47%. This shows that the difficulty of predicting binding affinity for flexible interfaces can be circumvented by the right choice of variables. This observation is also backed up by the lack of correlation between the interface flexibility and the prediction error (Fig. 3.7). We also note in passing that this conclusion is based on the analysis of the prediction ratios of Eq. (3.23), rather than that of the correlation coefficients of Eq. (3.19) (Table 3.2). Correlation coefficients are indeed global indicators of the dependency between two random variables, and do not assess the predictive performances on a per-complex basis.

Specific cases. Inspecting extreme cases is informative (named cases, Fig. 3.7). The individual predictions $\hat{g}_i$ from Eq. 3.16 are provided in Appendix Table B.1.

On the one hand, the affinity of three complexes with sub-picomolar affinity (1EMV, 1BRS, 1DFJ) is significantly under-estimated (Fig. 3.7). These three complexes involve an inhibitor taking the place of a cognate nucleic acid. Such complexes typically involve strong electrostatic interactions [PDCR10, Jan14], which are overlooked by our models. It could also be the case that such complexes manage to limit the entropic loss upon binding, possibly by transferring the dynamics of interfacial atoms to the protein's non interacting atoms.

On the other hand, predictions are excellent for several flexible cases, in particular 1F6M and 2I9B (Fig. 3.7). Complex 1F6M consists of a thioredoxin reductase in flavin-reducing conformation with its substrate. The reductase switches between bound and unbound conformations using a hinge-like motion. Complex 2I9B consists of a urokinase plasminogen activator receptor and its associated ligand. There is a global conformational change of the receptor upon binding (RMSD 2.657 Å) but no obvious hinge motion. It is the only complex with an iRMSD greater than 3 Å and a prediction error below 1.4 kcal/mol.

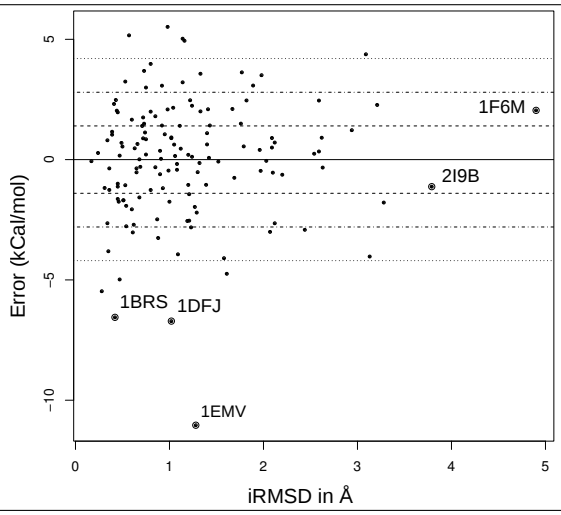
Table 3.2 Binding affinities: correlations between predictions and measurements for all datasets and all specific predictive models. (Whole table) Red values show the best results for a given category in the corresponding section, and cross-validated and classical correlation coefficients are treated separately in the second part of the table. **Bold** values in the second part of the table show the categories on which the variables selection was performed for a given predictive model. **(First section)** Previous work: values published and our replica (rep). Green cells correspond to the correlation coefficient of the predictive model over the train set (*i.e.* $\sqrt{R^2}$ from the linear regression). Purple cells correspond to either cross-validation results, prediction on a test set distinct from the trains set, or both. For the other cell colors, see details in Section 2.4.6. Yellow cells: discrepancies between original values and our replicas – we did not remove any complex. Orange cells: the cross-validation procedure made some test data information leak into the training set. Cyan cells: overlapping training and test sets. **(Second section)** Eight predictive models developed in this work. The value reported in a cell corresponds to the correlation between predictions and experimental values. For the $N_{XV} = 10000$ 5-fold cross-validation, the median of the predictions was used (Section 3.2.4). Purple lines show the cross-validated correlation coefficient, while green lines show the classical correlation coefficient (square root of the coefficient of determination).

Predictive Model	#param	SAB-A	SAB-A-HR	SAB-$R_{1.0}$	SAB-$R_{1.1}$	SAB-$R_{1.5}$	SAB-I	SAB-F_1	SAB-$F_{1.5}$
[MAB11]	94 (200) (Section A)	0.55	-	0.7	-	-	-	0.36	-
[Jan14] rep	2	0.20	-	-	0.55 (0.62)	-	0.07 (0.38)	-	0.13
[YGHW13]	∼1000 (Section A)	0.39	-	0.63	-	-	-	0.24	-
[KRF+14] rep	3	0.48	-	0.58	-	0.54	-	-	0.34
[ERS14]	7	0.57	0.71	-	-	-	-	-	-
Predictive Model 1: selected for datasets SAB-A									
N_{XV}· 5-fold CV	2	**0.48**	0.72	0.64	0.62	0.46	-0.39	0.27	0.39
LOO CV	2	**0.48**	0.72	0.63	0.61	0.46	-0.57	0.24	0.35
No CV	2	**0.52**	0.76	0.67	0.64	0.51	0.20	0.37	0.59
Predictive Model 2: selected for datasets SAB-A-HR									
N_{XV}· 5-fold CV	4	0.44	**0.77**	0.52	0.51	0.40	-0.08	0.24	0.31
LOO CV	4	0.44	**0.77**	0.51	0.51	0.39	-0.16	0.22	0.27
No CV	4	0.50	**0.83**	0.62	0.60	0.49	0.40	0.40	0.59
Predictive Model 3: selected for datasets SAB-$R_{1.0}$									
N_{XV}· 5-fold CV	4	0.47	0.71	**0.67**	0.62	0.46	-0.21	0.24	0.42
LOO CV	4	0.46	0.71	**0.66**	0.62	0.45	-0.31	0.23	0.39
No CV	4	0.52	0.76	**0.72**	0.67	0.53	0.40	0.41	0.65
Predictive Model 4 and 9: selected for datasets SAB-$R_{1.1}$									
N_{XV}· 5-fold CV	3	0.47	0.71	0.64	**0.63**	0.46	-0.21	0.28	0.44
LOO CV	3	0.47	0.71	0.63	**0.62**	0.45	-0.30	0.26	0.41
No CV	3	0.52	0.76	0.68	**0.67**	0.51	0.35	0.41	0.65
Predictive Model 5: selected for datasets SAB-$R_{1.5}$									
N_{XV}· 5-fold CV	2	0.44	0.64	0.63	0.60	**0.47**	-0.27	0.25	0.31
LOO CV	2	0.43	0.63	0.62	0.59	**0.46**	-0.54	0.22	0.28
No CV	2	0.47	0.68	0.66	0.62	**0.5**	0.12	0.32	0.43
Predictive Model 6: selected for datasets SAB-I									
N_{XV}· 5-fold CV	2	0.04	0.39	0.00	0.01	0.14	**0.23**	-0.11	-0.15
LOO CV	2	-0.04	0.07	0.03	0.02	0.07	**0.17**	-0.10	-0.18
No CV	2	0.15	0.50	0.18	0.18	0.23	**0.43**	0.12	0.15
Predictive Model 7: selected for datasets SAB-F_1									
N_{XV}· 5-fold CV	4	0.46	0.69	0.59	0.57	0.43	-0.19	**0.31**	0.33
LOO CV	4	0.46	0.69	0.59	0.56	0.43	-0.28	**0.29**	0.28
No CV	4	0.52	0.77	0.67	0.65	0.51	0.34	**0.44**	0.60
Predictive Model 8: selected for datasets SAB-$F_{1.5}$									
N_{XV}· 5-fold CV	3	0.35	0.46	0.59	0.53	0.36	-0.04	0.20	**0.46**
LOO CV	3	0.34	0.43	0.58	0.52	0.35	-0.12	0.18	**0.45**
No CV	3	0.40	0.55	0.64	0.59	0.42	0.29	0.33	**0.59**

Table 3.3 Datasets and their specific predictive models: performances in estimating the dissociation free energy ΔG_d. Each predictive model (rows) was tested on each dataset (columns). A cell in the Table features the values of the affinity prediction ratio $p_{1.4}^{\text{error}}$, $p_{2.8}^{\text{error}}$ and $p_{4.2}^{\text{error}}$ respectively, see Eq. (3.23). For instance, Predictive Model 1, when evaluated on dataset SAB-A (139 complexes) predicted 47.48%, 78.42% and 92.09% of the complexes with a median absolute error below 1.4, 2.8 and 4.2 kcal/mol, respectively. Equivalently, these are the fractions of cases such that K_d is estimated within one, two and three orders of magnitude. **(Top part)** Previous work. Lines marked with Rep. (replica) were obtained using the values of the parameters provided in the SAB for [Jan14] and those provided by the authors (personal communication) for [KRF$^+$14], along with their respective protocols. Lines not marked with Rep. were obtained using the variables of the original models, within our setup. **(Bottom part)** Our predictive models. Bold values indicate when a predictive model was tested on its specific dataset.

	SAB-A	SAB-A-HR	SAB-$R_{1.0}$	SAB-$R_{1.1}$	SAB-$R_{1.5}$	SAB-I	SAB-F_1	SAB-$F_{1.5}$
(1) [Jan14, Janin, rep]	29.63, 60.74, 77.78	-	-	37.33, 76.00, 92.00	-	37.04, 62.96, 85.19	-	6.06, 24.24, 39.39
(2) [Jan14, Janin]	39.57, 64.75, 89.21	44.12, 72.06, 92.65	37.18, 73.08, 91.03	39.05, 71.43, 90.48	39.57, 64.75, 89.21	37.14, 71.43, 88.57	32.35, 67.65, 85.29	21.62, 56.76, 86.49
(3) [KRF$^+$14, Kastritis, rep]	46.85, 76.92, 90.21	-	41.67, 80.56, 90.28	-	48.62, 81.65, 91.74	-	-	41.18, 61.76, 85.29
(4) [KRF$^+$14, Kastritis]	46.76, 75.54, 88.49	51.35, 75.68, 94.59	45.59, 80.88, 91.18	48.72, 80.77, 93.59	47.62, 82.86, 91.43	46.76, 75.54, 88.49	50.00, 72.86, 85.71	47.06, 67.65, 85.29
(5) Predictive Model 1	**47.48, 78.42, 92.09**	56.76, 86.49, 94.59	54.41, 77.94, 92.65	53.85, 80.77, 91.03	47.62, 79.05, 91.43	44.44, 62.96, 85.19	44.29, 77.14, 92.86	47.06, 70.59, 97.06
(6) Predictive Model 2	47.48, 77.70, 90.65	**62.16, 89.19, 97.30**	41.18, 76.47, 89.71	44.87, 79.49, 88.46	40.95, 76.19, 90.48	29.63, 70.37, 85.19	47.14, 77.14, 88.57	55.88, 73.53, 91.18
(7) Predictive Model 3	48.92, 78.42, 91.37	54.05, 83.78, 94.59	**51.47, 82.35, 92.65**	55.13, 79.49, 91.03	45.71, 80.95, 91.43	37.04, 70.37, 88.89	48.57, 74.29, 91.43	52.94, 79.41, 91.18
(8) Predictive Models 4 and 9	48.20, 79.14, 91.37	51.35, 86.49, 94.59	57.35, 79.41, 91.18	**55.13, 79.49, 91.03**	43.81, 77.14, 91.43	40.74, 66.67, 88.89	48.57, 74.29, 92.86	52.94, 79.41, 91.18
(9) Predictive Model 5	42.45, 76.98, 89.93	56.76, 81.08, 89.19	57.35, 79.41, 89.71	55.13, 80.77, 88.46	**45.71, 80.00, 89.52**	40.74, 74.07, 88.89	47.14, 75.71, 88.57	41.18, 70.59, 85.29
(10) Predictive Model 6	37.41, 64.03, 87.05	37.84, 64.86, 83.78	36.76, 55.88, 88.24	34.62, 56.41, 87.18	37.14, 66.67, 85.71	**44.44, 70.37, 88.89**	35.71, 68.57, 87.14	32.35, 67.65, 88.24
(11) Predictive Model 7	48.92, 79.14, 91.37	59.46, 83.78, 91.89	54.41, 77.94, 91.18	52.56, 78.21, 91.03	46.67, 79.05, 91.43	37.04, 66.67, 85.19	**50.00, 78.57, 90.00**	50.00, 73.53, 94.12
(12) Predictive Model 8	38.85, 74.82, 89.93	32.43, 70.27, 89.19	48.53, 79.41, 88.24	44.87, 78.21, 87.18	39.05, 74.29, 89.52	33.33, 74.07, 88.89	47.14, 72.86, 90.00	**50.00, 79.41, 85.29**

Figure 3.7 The hardness of predicting a binding affinity does not correlate with the flexibility of the complex. x-axis: flexibility of the interface, expressed in terms of interface iRMSD; y-axis: median prediction error $e_i[T_l, \mathcal{D}]$ (Eq. (3.20)). 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.



Classically for complexes with large interfaces, affinity predictions based on the BSA often result in overestimates. Beyond a certain interface size, the affinity no longer increases as much with the interface size, a behavior which could be related to a non-uniform atomic packing at the interface [Jan14]. However, the packing distribution of large interfaces matches that of the remaining ones (Fig. 3.8a), and no correlation is observed between the quality of individual predictions and interface size (Fig. 3.8b). Thus, packing heterogeneity may not account for mild to poor prediction performances in that context.

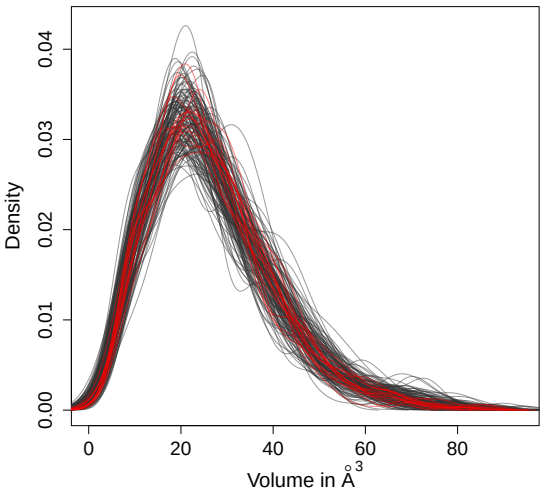
Validation on external datasets. Cross validation results obtained on datasets of small size should be interpreted with care [GT02], and checks on external datasets are a must [MFR14]. We therefore ran predictions on an external dataset (Table 3.4), from which two striking facts emerge.

The correlations observed compare to those obtained with cross-validation, with a maximum drop of 0.11 excluding predictive models 2 and 8. For the latter two predictive models, the drop reaches 0.33 and 0.25 respectively, a fact likely related to the small size of their training datasets. Second, the proportions $p_{1.4}^{\text{error}}$, $p_{2.8}^{\text{error}}$ and $p_{4.2}^{\text{error}}$ are smaller than their cross-validated counterparts, by a factor 1.4 ($p_{4.2}^{\text{error}}$, predictive model 8) to 9.5 ($p_{1.4}^{\text{error}}$, predictive model 7). Therefore, on this external dataset, despite being good predictors on a global level, as assessed by the correlation coefficient, our predictive models do not always perform robustly on a per complex basis.

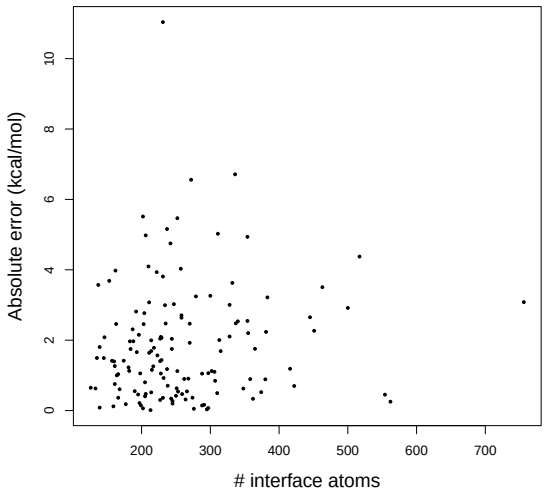
Table 3.4 Validation of the models on an external test set. The external test set from [KRF+14, supplemental] was split using the same criteria as those used to define datasets from the structure affinity benchmark, yielding *external datasets*. Each linear model was trained using a specific template on the whole corresponding dataset and used to predict the corresponding external datasets. The first part of the table displays the external dataset size, Pearson's correlation coefficients and p-value for each predictive on its external dataset along with $p_{1.4}^{\text{error}}$, $p_{2.8}^{\text{error}}$ and $p_{4.2}^{\text{error}}$. The second part show the values from table the diagonal of tables 3.3 and 3.2 for comparison.

	Predictive Model 1 SAB-A	Predictive Model 2 SAB-A-HR	Predictive Model 3 SAB-$R_{1.0}$	Predictive Model4 SAB-$R_{1.1}$	Predictive Model 5 SAB-$R_{1.5}$	Predictive Model 6 SAB-I	Predictive Model 7 SAB-F_1	Predictive Model 8 SAB-$F_{1.5}$
dataset size	51	24	13	16	23	7	38	28
p-value	0.0004	0.0295	0.0022	0.0392	0.0170	0.6034	0.0043	0.2753
correlation	0.47	0.44	0.77	0.52	0.49	0.24	0.45	0.21
$p_{1.4}^{\text{error}}$, $p_{2.8}^{\text{error}}$, $p_{4.2}^{\text{error}}$	11.76, 33.33, 47.06	20.83, 25.00, 41.67	30.77, 30.77, 46.15	31.25, 43.75, 43.75	26.09, 34.78, 43.48	0.00, 14.29, 14.29	5.26, 26.32, 52.63	17.86, 46.43, 60.71
median corr.	0.48	0.77	0.67	0.63	0.47	0.23	0.31	0.46
$p_{1.4}^{\text{error}}$, $p_{2.8}^{\text{error}}$, $p_{4.2}^{\text{error}}$	47.48, 78.42, 92.09	62.16, 89.19, 97.30	51.47, 82.35, 92.65	55.13, 79.49, 91.03	45.71, 80.00, 89.52	44.44, 70.37, 88.89	50.00, 78.57, 90.00	50.00, 79.41, 85.29

Figure 3.8



(a) Distribution of atomic volumes *i.e.* volumes of Voronoi restrictions for interface atoms. Red curves denote complexes whose interface lies in the top decile in terms of size (i.e, more than 354 interface atoms).



(b) The quality of individual predictions, assessed by $e_i[T_l]$, does not correlate with the interface size.

3.3.4 Accounting for Interface Morphology and Packing Boosts Performances

The performances of our predictive models owe to the new variables introduced in this study (Table 3.6). The variable selected most often is IVW-IPL (6/8 cases), stressing the role of the interface size (in terms of buried surface area), but also of atomic packing properties. The second variable selected most often is $\text{NIS}^{charged}$ (5/8 cases), highlighting the role of solvent interactions [Jan14]. Two other variables selected for 3/8 datasets, respectively represent volume variation at the interface rim (SVD_SO1), and solvation properties of the complex at the atomic scale (ATOM_SOLV). Interestingly, inspecting these four variables reveals a correlation between IVW-IPL and SVD_SOGT1 (Table 3.5), so that these variables might be used interchangeably. The same observation holds for $\text{NIS}^{charged}$ and NIS^{polar}.

Of particular interest in this context is our best overall predictive model. This predictive model uses variables IVW-IPL, SVD_SO1 and $\text{NIS}^{charged}$ and is therefore equivalent to predictive model 4. Not surprisingly, these variables form the top three of variables selected most often by the specific predictive models (Table 3.6). Its performances, are similar to those of specific predictive models on their own datasets (Table 3.3). Interestingly, it is a better predictor of flexible complexes than predictive model 1. Finally, its results on external datasets (Tables 3.4 and B.4) show that it is outperformed by specific predictive models for four datasets, and outperforms them for two (not considering predictive model 6).

Table 3.5 Pearson correlation coefficients between the individual variables.

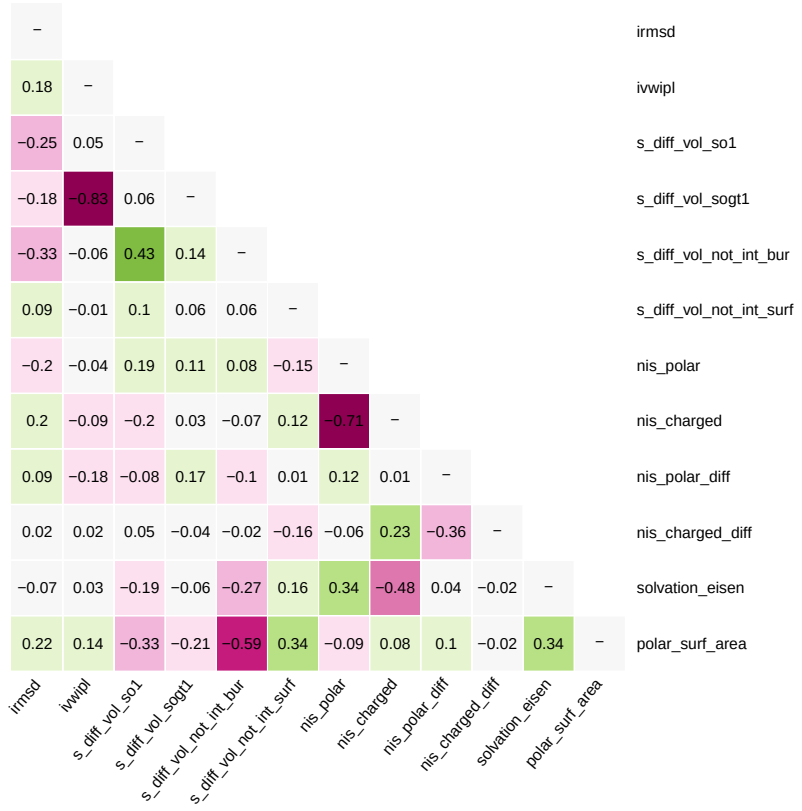


Table 3.6 Parameters used by the best predictive model for a given dataset. A *dataset* is a subset of the structure affinity benchmark. A *specific* predictive model is the predictive model which performed significantly better than all the others for a given dataset during model selection. The parameters are those from Table 3.1. Black dots mark variables used by statistically significant predictive models and white dots those used by other predictive models. The last column counts the number of statistically significant predictive models using a given parameter. Asterisks identify atomic level parameters.

	Predictive Model 1 SAB-A	Predictive Model 2 SAB-A-HR	Predictive Model 3 SAB-$R_{1.0}$	Predictive Models 4 and 9 SAB-$R_{1.1}$	Predictive Model 5 SAB-$R_{1.5}$	Predictive Model 6 SAB-I	Predictive Model 7 SAB-F_1	Predictive Model 8 SAB-$F_{1.5}$	Counts
iRMSD			●						1
IVW-IPL*	●		●	●	●		●	●	6
SVD_SO1*			●	●				●	3
SVD_SOGT1*		●							1
SVD_NI_E*							●		1
SVD_NI_B*		●				○			1 (2)
NIS^{polar}					●				1
$NIS^{charged}$	●	●	●	●			●		5
ΔNIS^{polar}						○			(1)
$\Delta NIS^{charged}$									
ATOM_SOLV*		●					●	●	3
POLAR_SASA*									

3.3.5 Performances using a k-nearest neighbors predictor

In order to see how the statistical method used to predict affinity values could influence the quality of predictions, we run the repeated cross-validation procedure on model 1 using k-nearest neighbors (knn) regression with $k = 10$ instead of least-squares linear regression. For this, complexes are represented as vectors in the space of parameters. In our case, each complex is represented by a 2D vector consisting in its IVW-IPL and $\text{NIS}^{charged}$ values. This allows the euclidean distance between vectors to be computed which can then be used to compute the nearest neighbors of a given complex. To predict a given complex, the affinity values of its k nearest neighbors are simply averaged.

This results in a Pearson correlation coefficient of 0.39 on the whole SAB (versus 0.48 for the linear model). In term of errors, this methods results in $p_{1.4}^{\text{error}} = 39.42$ (versus 47.48), $p_{2.8}^{\text{error}} = 72.99$ (versus 78.42) and $p_{4.2}^{\text{error}} = 88.32$ (versus 92.09).

Despite the fact that it does not assume linearity, the knn method does not improve the accuracy of the predictions.

3.4 Discussion and Outlook

This work develops sparse binding affinity predictions models, which shed new light on the hardness of affinity prediction, and improve prediction quality using variables coding enthalpic and entropic variations upon binding.

On the hardness of affinity predictions. Flexible datasets have been reported as the most challenging ones in previous studies. However, as shown here, the segregation of flexible versus rigid appears partially founded, with some easy to predict flexible complexes, and some hard to predict rigid cases. This observation is not completely surprising, since conformational changes alone tell little, in particular, on entropic changes upon binding. It also hints at the possibility of improving the quality of predictions for cases with small conformational changes upon binding, as molecular dynamics simulations in the intermediate time range may provide good estimates for the entropic penalties in those cases.

On the quality of predictions. A key achievement of this study is the quality of predictions, assessed in terms of absolute error or equivalently accuracy on K_d. To summarize, two values may be put forward, namely the fraction of cases for which K_d is predicted within one and two orders of magnitude. For the best overall predictive model, these fractions, corresponding to the whole SAB, are 48% and 79%, respectively. For the predictive model specific of high resolution complexes, these fractions are 62% and 89%. These numbers clearly advance the state-of-the-art, and call for two comments.

First, our models do not take into account the pH, whose change by two units may alter K_d by a factor ten or more. Given this specificity, they second the goal set in [Jan14], namely that of approximating K_d within two orders of magnitude.

Second, the high performance obtained for high resolution structures recalls the short range nature of selected forces–van der Waals interactions in particular, and stresses the dependence of energies on atomic details. From a quantitative standpoint, from Cruickshank's formula, the typical precision on atomic coordinates at a resolution of say 2.5Å lies in the range $[0.2, 0.4]$ Å [Cru99, Blo02]. At such a resolution, which is the worst used in the high resolution dataset (Fig. 3.1), the inter-atomic distance between non covalently bonded atoms located nearby in 3D space [CPBJ06] may already be spoiled by a factor circa $\sim 1/4$ (say $2 \times 0.3\mathring{A}/2.5\mathring{A}$). The situation deteriorates with the resolution, with a potential significant impact on the atomic scale parameters listed in Table 3.1. Therefore, the incidence of resolution on prediction performance should not come as a surprise. In a more general perspective, this observation is reminiscent of the role of molecular shape in determining motions [LM05], and also on the importance of packing properties in protein structure [Cha03].

One generic predictive model versus several specific predictive models. The diversity of the specific predictive models may be seen as a weakness or a strength. For the former viewpoint, one may argue that thermodynamics call for a unified model. For the latter one, given the intrinsic complexity

of the problem (recall that the binding affinity is inherently coupled to a thermodynamic equilibrium), and the paucity of the dataset, it is clearly beneficial to exploit specific features of datasets. Moreover, specific predictive models are of practical interest since to predict the affinity of a complex performing a specific biological function, one may use a dataset of complexes related to that function. Further arguments to choose between these two interpretations will likely emerge upon populating the structure affinity benchmark.

On key parameters. Our predictive models preferably use parameter IVW-IPL, and then $\text{NIS}^{charged}$. The former, introduced in this work, combines the overall shape of the interface and involves atomic packing properties. It is reminiscent of cooperativity phenomena observed for weak interactions [BGNC09]. The latter, $\text{NIS}^{charged}$, encodes the electrostatic properties of the non interacting surface, as recently investigated [VKB15]. The following top scorers represent volume variation at the interface rim (SVD_SO1), and solvation properties of the complex at the atomic scale (ATOM_SOLV). Among these four variables, two describe surface properties at different scales (atomic for ATOM_SOLV, and at residue level for $\text{NIS}^{charged}$), and two encode interface properties, one static for the whole interface (IVW-IPL), and one dynamic for the outer layer of the interface (SVD_SO1).

Remarkably, these parameters are simple ones, derived from the Voronoi diagram of the solvent accessible models of the three structures involved (two for the partners, one for the complex). From a computational standpoint, processing a structure of say up to 10,000 atoms takes a handfull of seconds on a desktop computer [CKL11].

Outlook. Estimating binding affinities is a central endeavor to understand protein - protein interactions. Strikingly, the predictive models and variables presented here yield a prediction accuracy of 2.8 kcal/mol per complex in 79% of cases for the whole SAB, and in 89% of cases for high resolution complexes. This represent a significant progress over previous methods. Since our methods inherently exploit static properties of crystal structures, improving results even further calls for developments in two directions. On the one hand, unveiling dynamical properties of the partners and the associated complex, by sampling and modeling the associated (potential, free) energy landscapes will undoubtedly yield enhanced predictions [Wal03]. Along the way, a central problem to be addressed is that of the potential energy model best suited, since, as shown in this work, coarse grain descriptors can match or surpass the performances of detailed chemical ones. In this respect, our ability to accurately sample [Wal03, Chi14] and compare [CDM$^+$15] sampled energy landscapes should prove critical. On the other hand, a weakness shared by our method and previous ones is the absence of terms taking into account the pH and the ionic strength – a limitation actually accounting for the poor performances observed on complexes involving significant electrostatic interactions. For such cases, incorporating terms accounting for counter-ion condensation seems critical, yet, controlling the enthalpy - entropy balance within such models remain challenging [PDCR10, Sch99].

The affinity prediction problem is also of special interest from the machine learning perspective. Affinity prediction is indeed modeled here a particular instance of a problem known as regression [GK02]. In this setting, the data is assumed to be generated by a process and applied some random noise. The most important attribute of regressors is their *consistency*, *i.e.* their ability to converge toward the true model given data accounting for the whole space. However, for a regressor to achieve consistency, the data must satisfy some assumptions. For instance it should be well distributed over the space of possible data points. In our case, this means that the dataset should evenly represent all possible protein-protein complexes. This is most probably not the case for the SAB. The availability of larger datasets will also ease the model selection problem, undertaken by complete enumeration over the parameter set in this work. In principle, sparse least square models can be obtained using regularization techniques [Tib96]. However, the inherent randomization used by cross-validation makes model selection unstable for small datasets, making such methods hard to use at this stage. For these reasons, sparse specific models using with relevant variables, as developed in this work, appear as a privileged solution to estimate binding affinities.

Chapter 4

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

4.1 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. 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 [BN98]. 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) [LL01].

Analysis of Ig - Ag complexes. The prominent role played in Ag binding by CDRs has prompted the analysis of CDR-specific statistics related to their conformation (both individually and all at once) and contribution to Ag binding (Sections 2.3 and 2.4.5).

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 and the complex (Ig - Ag). It is generally measured by the dissociation constant K_d (= [Ig]·[Ag]/[Ig - Ag]) of this equilibrium. Equivalently, it is expressed by the corresponding dissociation free energy ΔG (Section 2.4.1). Predicting binding affinities from structural data is a notoriously challenging problem for protein complexes in general [KMH$^+$11, MBC15], and for Ig - Ag complexes in particular

[LWT07].

Contributions In this work, we present novel quantitative analyses for interfaces of Ig - Ag complexes. Using the annotated IMGT/3Dstructure-DB [EKL10], 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.

4.2 Material and methods

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

Ig - Ag complexes

We use the Ig - Ag complexes from the IMGT/3Dstructure-DB (`http://www.imgt.org/3Dstructure-DB/` [EKL10]), corresponding to the category *IG/Ag* for *IMGT complex type*. Only IMGT-PDB files are kept. This dataset features 1602 files. Each file is processed in order to identify *canonical complexes* involving one heavy chain, one light chain, and one ligand (paragraph *Infering canonical complexes*). A total of 1275 canonical complexes are thus extracted, of which 554 non-redundant complexes (paragraph *Removing redundancies from IMGT/3Dstructure-DB*). 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. The distribution of complexes among these categories is displayed in Table 4.1. The distribution of ligand sizes is shown in Fig. 4.1. In total, 489 complexes are retained after filtering for missing data, inconsistencies, redundancy, ligand type and species.

Table 4.1 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).

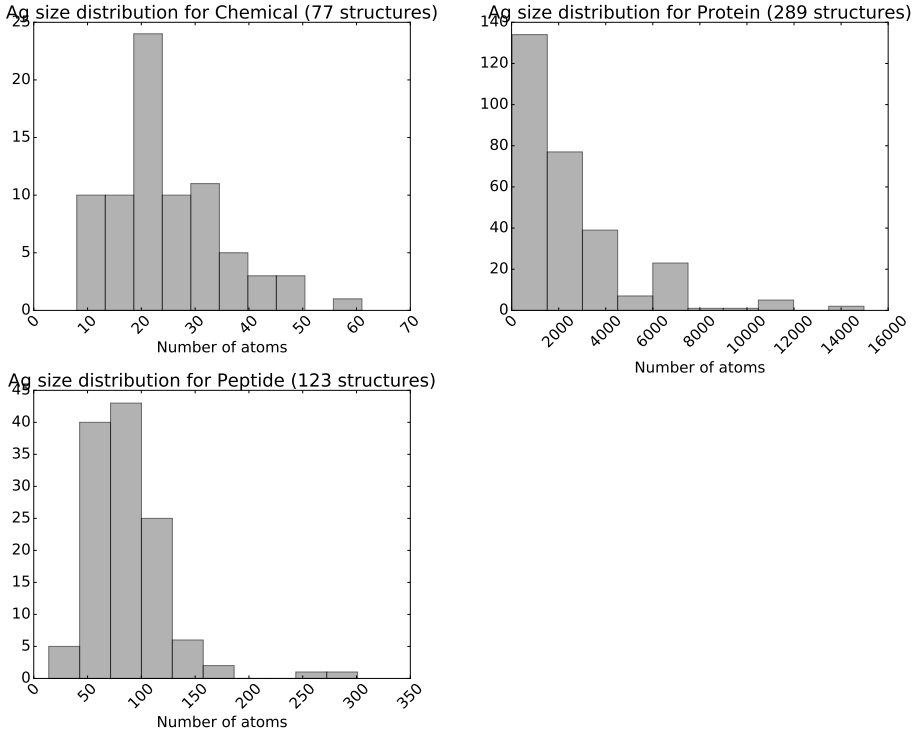
	Mouse	Human	Other	total
Peptide	80	32	11	123
Protein	168	91	30	289
Chemical	65	7	5	77
total	313	130	46	489

CDR and FR limits of the VH and VL domains are according to the IMGT unique numbering [LPR$^+$03] (Table 4.2). 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.

Table 4.2 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 [LPR$^+$03].

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

Figure 4.1 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.



Inferring canonical complexes

Canonical complexes. A *canonical configuration* from an 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

Overall, 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 4.2.3, 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).

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.

4.2.2 The binding affinity benchmark

Our affinity predictions exploit the structure affinity benchmark (SAB) [KMH$^+$11], 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 (Appendix Table C.1), and 125 non-Ig - Ag cases defining the training set.

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.

Having learned a statistical model from the latter, we predict affinities for Ig - Ag complexes of the former.

4.2.3 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 [BCRJ04a] and the associated Voronoi based interface model (Fig. 4.2 and [LC10]).

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 [LL01]. 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 [LPR$^+$03] (Table 4.2). 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 (Fig. 4.3).

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 [Caz10], and has been shown to provide information on binding modes [CPBJ06].

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 [GR01, BCRJ04a]. A convenient construction to study SAM is the Voronoi (power) diagram defined by the atoms [GR01]. 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 [GR01].

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 [LCJ99]. BSA has been shown to

Figure 4.2 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).

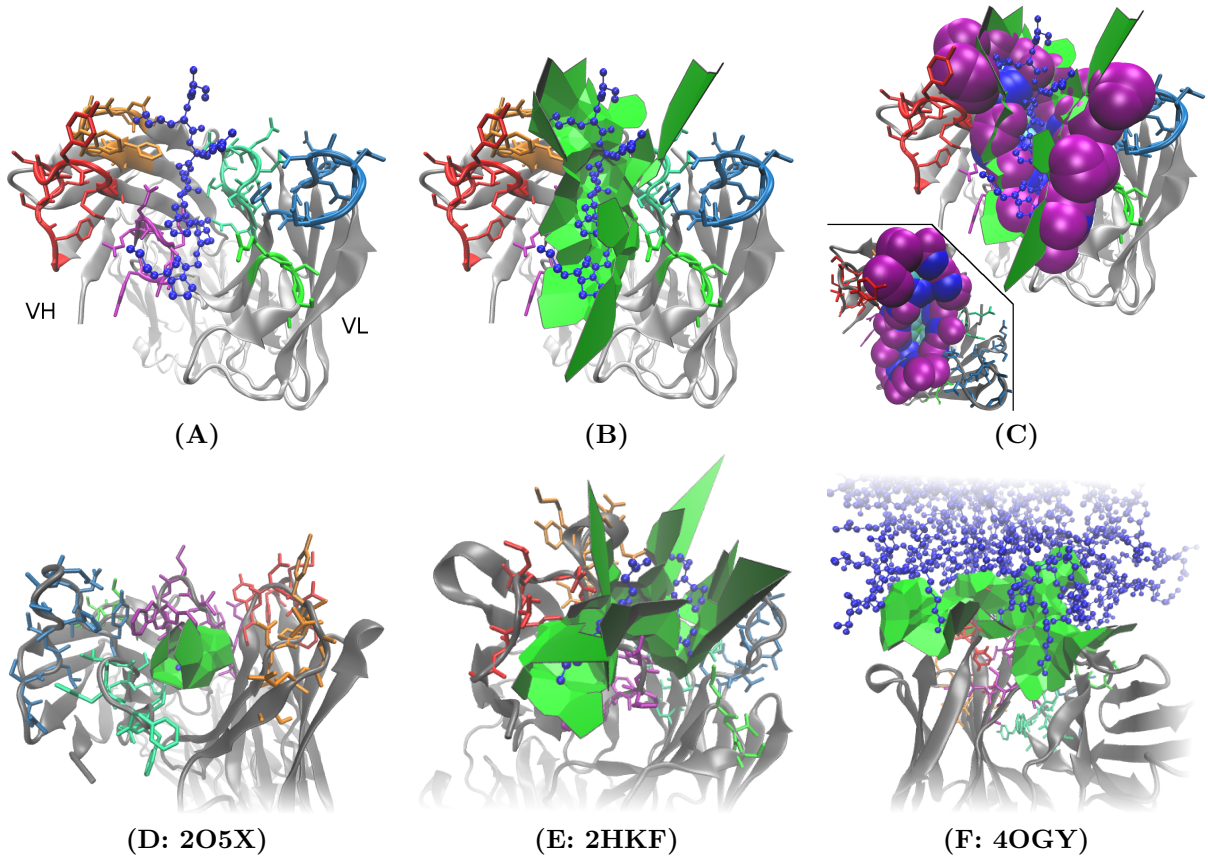
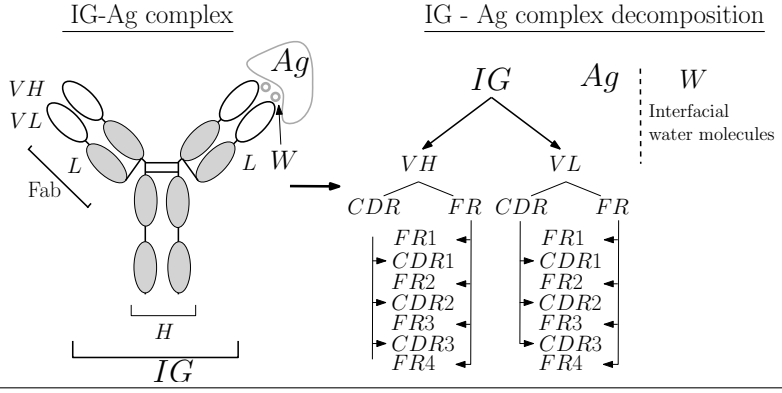


exhibit remarkable correlations with various biophysical quantities [JBC08], and notably dissociation free energies for complexes involving moderate flexibility [MDBC12].

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

Figure 4.3 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 Fig. 4.2 is partitioned into contributions from these 14 regions.



loosing solvent accessibility [CPBJ06]. 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 [BGNC09]. This information generalizes the core-rim model [LCJ99], and has been shown to provide state-of-the-art correlations with solvent dynamics, conservation of amino acids [BGNC09], and dissociation free energies [MDBC12]. 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 $\mathcal{I}_{\rm Ig}$ contributed by the Ig, and the atoms $\mathcal{I}_{\rm Ag}$ contributed by the Ag, so that $\mathcal{I} = \mathcal{I}_{\rm Ig} \cup \mathcal{I}_{\rm Ag}$. It follows that the number of interface atoms $|\mathcal{I}|$ is the sum of those contributed by the Ig and the Ag respectively, namely $|\mathcal{I}| = |\mathcal{I}_{\rm Ig}| + |\mathcal{I}_{\rm Ag}|$. Similarly, we charge the Buried Surface Area (BSA) to the Ig and Ag respectively, so that $\mathrm{BSA} = \mathrm{BSA}_{\rm Ig} + \mathrm{BSA}_{\rm Ag}$. These quantities yield the average BSA per interface atom on Ig and Ag side:

$$\overline{\mathrm{bsa}}_{\rm Ig} = \frac{\mathrm{BSA}_{\rm Ig}}{|\mathcal{I}_{\rm Ig}|}, \tag{4.1}$$

$$\overline{\mathrm{bsa}}_{\rm Ag} = \frac{\mathrm{BSA}_{\rm Ag}}{|\mathcal{I}_{\rm Ag}|}. \tag{4.2}$$

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 4.3). (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.

4.2.4 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: $\overline{\mathrm{bsa}}_{\rm Ig}$ and $\overline{\mathrm{bsa}}_{\rm Ag}$ 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. 4.5 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 (from the R package `rpart`) 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 the average confusion matrix(Table 4.4) and both the overall and per class error rates 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.

4.2.5 Predicting binding affinities

Relevant variables. The affinity prediction problem was recently revisited and posed as a sparse linear model estimation problem [MBC15], 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 [MBC15].

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 $\text{SO}(a)$ and $\text{Vol}(a)$ be the shelling order and Vol_bound the volume of atom a in the complex (see Section 4.2.3).

The, IVW-IPL is defined as follows:

$$\text{IVW-IPL} = \sum_{a \in \mathcal{I}} \frac{\text{SO}(a)}{\text{Vol\_bound}(a)} \tag{4.3}$$

On the one hand, the shelling order refines so-called core-rim models [LCJ99]. 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 ($\text{NIS}^{charged}$) 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 [KRF+14].

Statistical methodology. We estimate binding affinities using k nearest neighbors regression (knn) [GK02, BD15], 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. 4.7a). 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.

We assess the quality of our predictions by varying the value k. From a theoretical standpoint [GK02], it is known that k must be super-logarithmic and sub-linear in the number of cases processed. Since $log(144) \approx 5$, we explore the range $k \in 5, \ldots, 25$ (Fig. 4.7b). 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.

Practically, the variables used by the regression method were computed using the `binding affinity prediction` package from the structural bioinformatics library (SBL, `sbl.inria.fr`). For the fitting, we use the scikit-learn library [PVG+11], namely the `neighbors` package for knn regression. regression.

4.2.6 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. (4.3)) into the contributions of CDR1 + CDR2 and CDR3. We also compute the *average normalized shelling order* (or ANSO for short) for each CDR

$$\text{ANSO} = \frac{1}{|A|} \sum_{a \in A} \frac{\text{SO}(a)}{\text{Vol\_bound}(a)}, \tag{4.4}$$

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.

Table 4.3 Median BSA and median of BSA/BSA$_{Ig}$ 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.

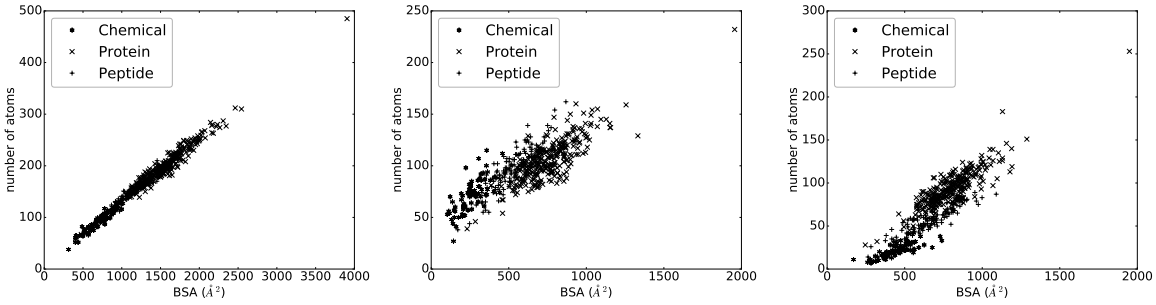
	Human			Mouse			Other		
	Chemical	Peptide	Protein	Chemical	Peptide	Protein	Chemical	Peptide	Protein
Ig	143.7 (100.0%)	645.7 (100.0%)	812.0 (100.0%)	254.6 (100.0%)	486.6 (100.0%)	703.9 (100.0%)	275.0 (100.0%)	546.6 (100.0%)	767.4 (100.0%)
VH	89.8 (57.0%)	418.3 (66.2%)	569.9 (67.7%)	154.8 (59.9%)	287.0 (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%)

4.3 Results

4.3.1 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 [JBC08]. 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 (Fig. 4.4).

Figure 4.4 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 $|\mathcal{I}_{\mathbf{Ag}}|$.**

To further investigate this observation, we compute the average BSA per interface atom for both the Ig and Ag (Eqs. (4.1) and (4.2)). 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. 4.5). 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 (Fig. 4.2(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 4.2.4, Fig. 4.5 and Fig. 4.6).

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 4.4), 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.

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 error rates per ligand type are the following: chemical: 5%; peptide: 20%; protein: 7%; and the overall median error rate is 9.6%, 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 (Fig. 4.6): $\overline{\text{bsa}}_{\text{Ag}} \geq 14.3 \Rightarrow$ 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.

Overall, our classifier is able to accurately predicts ligand types, despite the fact that the data is unbalanced.

Figure 4.5 Interaction specificity for Ig - Ag complexes: analysis and predictions. Both analyses are based upon the average buried surface areas per atom (Equations (4.1) (4.2)): $\overline{\text{bsa}}_{\text{Ig}}$ 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).

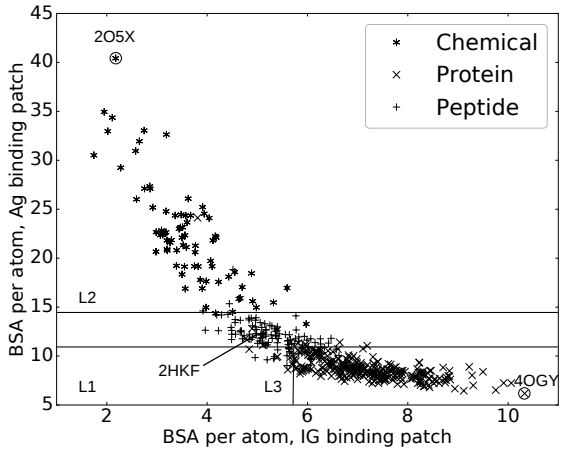


Figure 4.6 Classification rules characterizing the binding patch depending on the ligand types. The classification rules are: $\overline{\text{bsa}}_{\text{Ag}} \geq 14.3 \Rightarrow$ chemical ligand; $10.7 \leq \overline{\text{bsa}}_{\text{Ag}} < 14.3 \Rightarrow$ peptide ligand; $\overline{\text{bsa}}_{\text{Ag}} < 10.7$ 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.

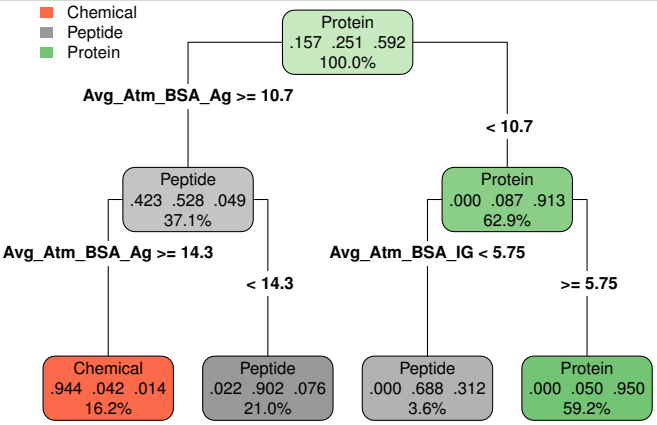


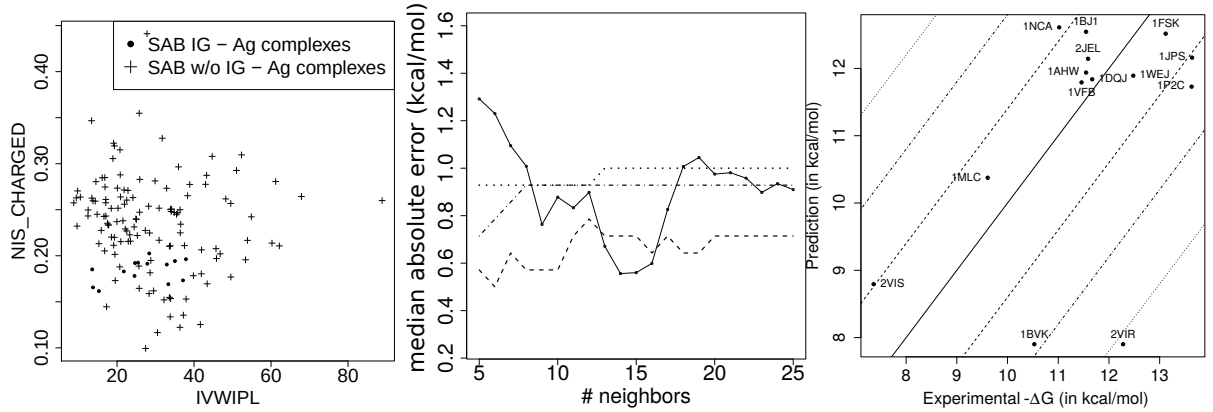
Table 4.4 Average confusion matrix for ligand type prediction. Results obtained by running 5-fold cross-validation 1000 times. Each repetition results in a confusion matrix which is averaged–e.g. on average 4.6 chemicals out of 77 are predicted as peptides.

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

4.3.2 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. 4.7c). 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 [Jan14]. They are also informative from a biological standpoint, as an affinity enhancement of two orders of magnitude is typically observed during affinity maturation.

Figure 4.7 Binding affinity analysis and predictions for Ig - Ag complexes. (4.7a) 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. **(4.7b) 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. **(4.7c) 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.

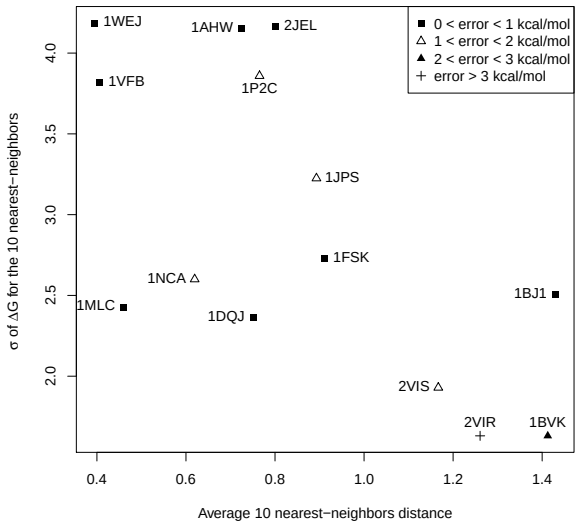


In order to compare these results to what could be expected from a null model, we take the average binding affinity of the training dataset (10.78 kcal/mol ±2.84) as prediction for all complexes. Note in passing that this is equivalent to using knn regression with k=125. This results in a median absolute 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 (Fig. 4.8). 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

Figure 4.8 Prediction error versus average distance of the 10 nearest-neighbors and the standard deviation of their affinity values.



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 [XRK+16], and is based on the work from Vangone *et al* [VB15].

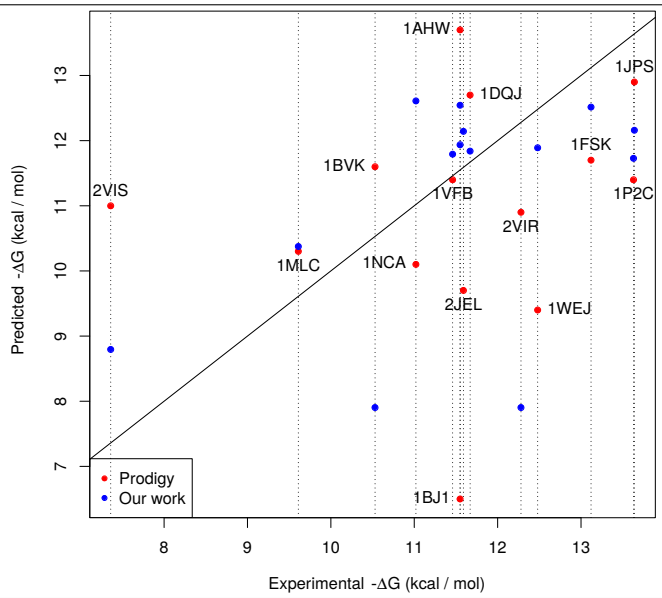
The accuracy of PRODIGY is lower than that of the current study with median absolute errors of 1.4 versus 0.878 kcal/mol respectively. For reference, we also provide the root mean squared errors (2.226 versus 1.676 kcal/mol), Pearson's correlation coefficients (0.149 versus 0.488) and Spearman's correlation coefficients (0.238, 0.291).

Interestingly, our method is successful at predicting similar affinities (Fig. 4.9) for five complexes (1AHW, 1DQJ, 1VFB, 2JEL, 1BJ1) for who PRODIGY predicts widely varying values.

CDRs: lengths and BSA. It has been observed that CDR lengths differ between different antigen types [CBM03, RSWA12], 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 investigated 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 (Appendix Figs. C.1 and C.2). 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 [CBM03]. 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) [CBM03].

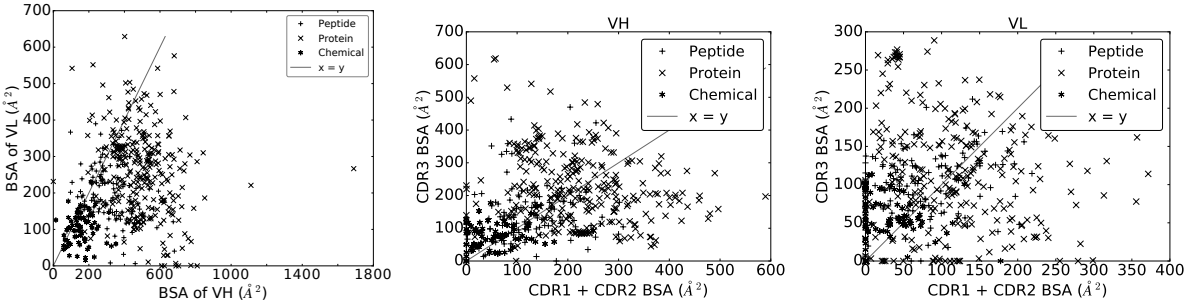
Respective contributions of the CDRs to the interface, for VH and VL domains. In an Ig - Ag complex, it is generally believed that VH contributes more to the recognition than VL. With a BSA

Figure 4.9 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.



of VH strictly larger than that of VL for 430/489 complexes (∼86%) (Fig. 4.10a), 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. 4.10b and 4.10c). 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 ∼46% of the complexes for VH, and for ∼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.

Figure 4.10 Buried Surface Area (A^2) of the VH and VL domains, and their respective CDR.



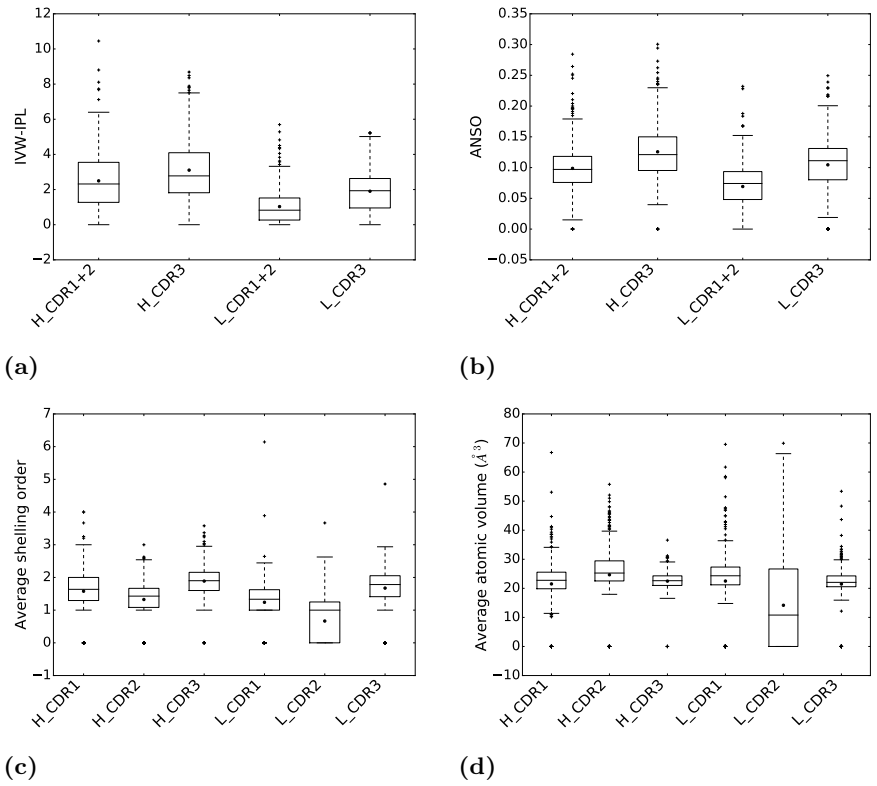
(a) VH versus VL **(b)** VH domain. **(c)** VL domain.

To assess the contributions of CDRs to binding energy, we compute both their IVW-IPL and ANSO (Eq. (4.3) and (4.4)) for all complexes (Fig. 4.11a and 4.11b). 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 (Fig. 4.12). 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 = $2.480 \cdot 10^{-37}$).

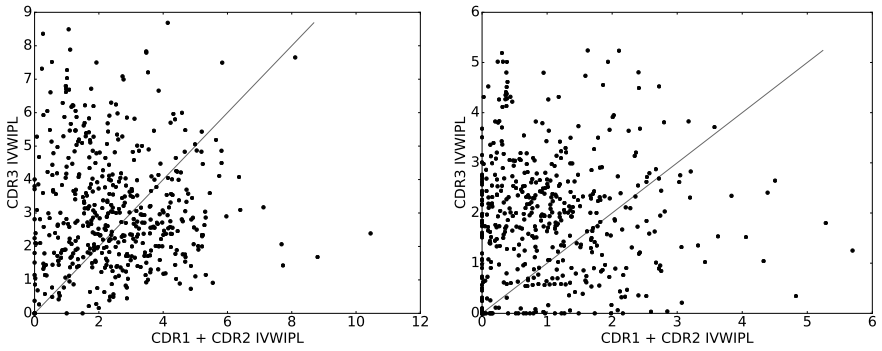
Thus, as opposed to the results obtained when considering the BSA, the sum of contributions to the binding affinity of CDR1 and CDR2 is significantly lower than that of CDR3 for both VH and VL.

Figure 4.11 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.



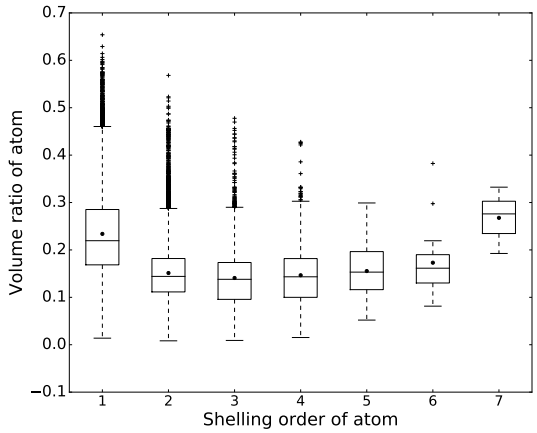
For both chains, the difference in ANSO can be imputed to two facts. First the average shelling order (Section 4.2.3) for atoms of the CDR3 is higher than those of CDR 1 and 2 (Fig. 4.11c). Second, their average atomic volume is lower (Fig. 4.11d). Both are related since the shelling order and the atomic volume are negatively correlated (Fig. 4.13).

Figure 4.12 IVW-IPL of the CDR of VH (left panel) and VL (right panel).



(a) **(b)**

Figure 4.13 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.



4.4 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 [BCRJ04a]. 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 [CMT11, LLZ$^+$06], 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 [MBC15]. 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. They are also more accurate than those returned by one of the state-of-the-art prediction method. 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 [LW13]. 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 [MSSR00, RTVC04, CeCG07, SXK$^+$13]. A useful proxy for dynamics is the length of VH CDR3, and difficulties were observed to define canonical conformations for VH CDR3 [CL87, ALLC97, SKN96, SKN99, KSKN08, NLD11] 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 4.2.5) 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 [XD00] 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 [FWE$^+$11] 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 [LC88, SZWR06], and more generally the constant regions which have been shown to influence the avidity [CSG$^+$93, CRGG94, MTS$^+$96, PMD$^+$00], and are involved in Ig effector properties, such as ADCC or CDC [GSF$^+$95]. 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.

Chapter 5

Comparison of immunoglobulin CDR3 repertoires.

5.1 Introduction

Adaptive immunity relies on specific responses to pathogens from a set of pre-existing receptors, among which Igs whose diversity come from random recombinations of germline Ig genes. Ag recognition leads to the activation of B cells and differentiation in Ig-secreting plasma and memory B cells. The latter typically allow a fast and specific secondary response upon re-encounter with a pathogen.

Typical secondary responses in mammals are fast and lead to copious secretion of specific Igs, which have usually undergone affinity maturation. In fish however, variations occur as slow primary responses, but also faster secondary responses, have been observed. In particular, fish seem to lack the ability to undergo efficient affinity maturation although the enzymatic machinery needed to produce hypermutated Igs is present and hypermutation has been observed. In particular high-affinity Igs take a long time to appear but have been observed without re-exposure to the Ag.

For fish, little is known about the characteristics of the B cell response during infection and, in particular after a second exposure to with the same pathogen. Repertoire sequencing appears as an efficient approach to get more insights into this, and to understand better the mechanisms of protection afforded by vaccination.

Recent years have seen the fast development and spread of high throughput sequencing methods resulting in the ability to obtain massive amounts of nucleic-acid sequences. In particular, these have allowed detailed descriptions of Ig diversity in various contexts, which we now shortly review.

Multiple studies focus on VDJ gene usage. In particular the seminal work of Weinstein *et al* [WJW$^+$09] set the basis for high throughput sequencing data analysis of Ig repertoires. This work and subsequent ones [JWP$^+$11, JHW$^+$13] largely focus on the analysis of VDJ genes usage, reporting descriptive statistics about VDJ combinations (counts, distributions), as well as analyses about their diversity (rarefaction curves and entropy). These VDJ-based analyses are complemented by descriptive statistics of CDR3 length, charge and quantify junctional diversity in [KLS$^+$14]. In terms of repertoire comparison, correlations between the VDJ combinations count vectors of pairs of individuals are used [WJW$^+$09, JWP$^+$11] sometimes along with PCA [KLS$^+$14]. Although VDJ usage analysis is valuable by itself, all these methods remain limited to a very coarse-grained description of the repertoire and remains oblivious to the effect of mutations since differing sequences can be assigned the same VDJ combination.

On the other hand, other studies such as the one by Vollmers *et al* [VSW$^+$13] use methods coming from the field of ecology to analyze Ig sequencing data without being restricted to VDJ analysis. In particular, the authors use the capture - recapture method to find that sequences shared between sequencing replicates correspond to sequences from activated B cells. They also use this method to estimate the effective size of the repertoire. For capture - recapture analysis however, only exact matches between sequences can be considered which allows the comparison of replicates but not of different fish. Quantification of the amount of identical clonotypes between multiple fish is also performed in [WJW$^+$09].

Other typical methods for repertoire analysis borrow from ecology. For instance diversity indices such as clonality, Shannon, Piélou and Hill indexes are commonly used by immunologists. In terms of repertoire comparisons, set overlap measures such as Sørensen and Jacquard indices are often used. More sophisticated methods such as the Morisita-Horn index and the normalized expected species shared (NESS) were defined in the context of ecology and allow to take into account the size of clonotypes. All these methods however consider that non-identical sequences are dissimilar without quantifying this dissimilarity, which very stringent and does not take into account lineages or convergent evolution of sequences.

This problem is often tackled by clustering or aggregation of sequences in order to get a finer level of detail than allowed by VDJ usage, while keeping the notion of lineages. However, the sequence comparison methods used for this task have remained generally limited to thresholding on the number of mutations [BHE11], or to various criteria depending on the study. For instance, lineages are defined as follows in [JWP+11, JHW+13]: sequences from the same VJ combination are allowed to cluster together if they have junction boundaries varying by at most one nucleotide, and if they share more than 80% similarity in the VDJ junction region. Single-linkage hierarchical clustering is then applied using this rule and each cluster corresponds to a lineage. In [VSW+13], the definition is the following: A lineage starts with a seed sequence. All sequences using the same V and J segment, of equal length, and whose junctional regions (non-templated nucleotides and D segments) share 90% similarity with all sequences in the lineage (initially only the seed) are added to the lineage. The process is repeated until no sequence is added to the lineage. In [BHE11], four level of sequence similarity are defined corresponding to one substitution, one insertion/deletion, both, and more mutations. The variations resulting from these different methods and thresholds may influence the resulting conclusions and complicate the comparison of the studies. Only [WJW+09] uses alignment scores along with clustering, although this is only used to define consensus sequences and mutation rates in D-segments.

Overall, all these methods focus on various aspects of repertoires such as VDJ usage, number of shared sequences or lineages or repertoire effective size.

In order to get a more global understanding of repertoires, two studies have taken a modeling approach.

First, a study by Mora *et al* [MWBC10] presents a very different approach consisting in a probabilistic model for the sequences of D gene-encoded regions. This model assigns any sequence a probability which is consistent with the frequencies of single and pairs of residues found in the data. The authors then analyze the resulting probability distribution to assess repertoire diversity (percentiles, rank versus probability). To assess the similarity of various fish repertoires, they compute the amount of information a given sequence gives about the identity of the fish to which it belongs (mutual information, MI), and compare it with the "fish entropy" computed on the probability for a fish to be found in a subgroups of different sizes. Finally, the authors report the sequences with highest probability along with their neighbors in terms of sequence similarity. Using a probabilistic model allows the measure of diversity to be more general than that computed on raw data because a sequence which is absent from the sequencing will be given a low probability in the model, essentially correcting for the sampling effects induced by the sequencing. Moreover the comparison of MI to fish entropy allows the comparison of repertoires of multiple fish without being limited to pairwise comparisons. Finally, neighbors of the sequences with high probability found by their model may not be found in the sequencing data but may represent past or future large clonotypes. The drawback of this approach is that the resulting probability distribution approximates the repertoire of a given fish in the limit of an infinite number of sequences.

Another study [BHE11] focuses on the network structure of the repertoire. For this, each vertex in the graph corresponds to a clonotype (and its associated size), and two clonotypes are linked by an edge if they differ by one mutation, one deletion or both. The authors find that some fish have little structure, with essentially isolated vertices whereas others have a very structured graph with large clusters of similar sequences. Noticing that clonotypes found in clusters are often larger than others, they relate this to clonal expansion and conclude that a cluster is the equivalent of a response to an antigen. The authors also perform VJ genes usage analysis and find that fish with structured networks show a skewed VJ genes usage. Modeling repertoires with graphs allows for an intuitive global representation. However, it is essentially qualitative, and classical graph quantities (such as clustering coefficients) are not necessarily meaningful in the context of immunology. Moreover, this does not allow for a straightforward comparison of repertoires.

For a review of repertoire sequencing and analysis techniques, we refer the reader to the [BBHLE12].

In this work, we analyze CDR3 sequences obtained through Illumina deep sequencing to study how the splenic Ig repertoires of rainbow trouts evolve upon primary and secondary infection by the rhabdovirus VHSV (Viral Hemorrhagic Septicemia Virus). We focus on the two isotypes involved in responses, IgM and IgT. To carry out this analysis we present a new method, based on the so-called *earth mover distance*, allowing the comparison of repertoires while retaining sequence similarity information. We also analyze repertoire overlap while taking into account both sequence identity (exact matches) and similarity classes (approximate similarity).

Contributions This work makes four main contributions. First, we characterize public and private responses by comparing fish repertoires using the earth-mover distance (EMD). Second we show that EMD describes responses in finer details than MHD by adding sequence similarity information to diversity. Third, we single out four different repertoire behaviors upon vaccination and challenge, by searching for shared large clonotypes in naive, vaccinated and vaccinated + infected fish. Fourth, we show how sampling sequences affects the representation of small clonotypes in the final dataset and how this affects the search for large clonotypes in subsampled repertoires.

From a methodological standpoint, the last two contributions call for one comment related to sequencing data. Sequencing essentially amounts to randomly sampling among the set of RNA expressed by an individual at a given time, although it cannot be assessed with certainty whether this process is truly random. Moreover, many experimental factors come into play which results in variable sequence counts for different fish. This is especially problematic for the methodology used in the third contribution where we seek to quantify the number of fish containing a given clonotype. In effect, such a clonotype is much more likely to be found in fish whose sequencing yielded many sequences, although this is not related to biological features of their repertoires. Moreover, using frequencies is not an option in that case since we are not looking at counts but at presence / absence. To bring all fish on an equal footing while preserving the clonotypes relative frequencies, we resort to *subsampling*. This is the *in silico* equivalent of the sampling process occurring during sequencing, except that the number of sequences obtained in the end is controlled and equal in all fish.

5.2 Material and methods

5.2.1 Dataset

Raw data and subsamplings. The dataset consists of protein sequences of VH CDR3. These are translation of the transcripts obtained from the spleen of 12 fish through Illumina sequencing. These fish belong to 3 *conditions*: naive (C); vaccinated (E1) against VHSV and analyzed 5 months latter; vaccinated and infected (E4) by VHSV 5 months post-vaccination, analyzed one month later.

For each fish, 6 sets of sequences have been sequenced, and we refer to these as *variable - constant pairs*, or VC pairs for short. The term "constant" refers to the constant region gene defining the class of the Ig. There are two such classes: IgM (with corresponding Cμ gene) and IgT (with corresponding Cτ gene). IgM is the main class of fish central response. IgT is specialized in mucosal immunity and, in this respect, is similar to IgA in humans. The term "variable" refers here to the VH gene, encoding the part of the variable domain of the heavy chain that is not encoded by the D and J segments or by the N diversity,*i.e.* roughly all the region encoded upstream VH CDR3. Six such genes in total have been sequenced:

- VH4.1-Cμ which was shown to engage in the private response [CJP$^+$13]
- VH5.1-Cμ which engages in the public response [CJP$^+$13]
- VH8.1-Cμ which, according to CDR3 spectratyping data, does not contribute significantly to the response
- VH4.1-Cτ which also engages in the private response [CJP$^+$13]
- VH5.4-Cτ which, according to CDR3 spectratyping data, is thought to be responsive only after challenge
- VH9.2-Cτ which, according to CDR3 spectratyping data, does not contribute significantly to the response

The total number of sequences obtained for each VC pair varies between 2805 and 1,574,557 (Table 5.1). Distinct sequences in each sample (one VC pair in one fish) are called *clonotypes* (thus defined by a V, a C, a J, and a CDR3 sequence). The "size" of a clonotype is the number of identical sequences found in the dataset. Barcoding systems used during the preparation of libraries ensure that sequences coming from a given molecule of RNA are counted only once. The distribution of clonotype sizes is given in Fig. 5.1.

Variable-constant (VC) gene pairs, top clonotypes (TCs). Consider the six aforementioned CV gene pairs. We define:

Definition. 1. For each VC pair $X \in \{\text{VH4.1-C}\mu, \text{VH5.1-C}\mu, \text{VH8.1-C}\mu, \text{VH4.1-C}\tau, \text{VH5.4-C}\tau, \text{VH9.2-C}\tau\}$ and condition $Y \in \{\text{C}, \text{E1}, \text{E4}\}$, we define the *top clonotype set* (TCS for short) $\text{TCS}_{\text{X}}^{\text{Y}}$ as the set consisting of the 50 largest clonotypes of the VC pair X collected over all fish of condition Y.

The elements of a top clonotype set $\text{TCS}_{\text{X}}^{\text{Y}}$ are called *top clonotypes* (TCs). Each such TC is represented by its a.a. sequence.

The six VC pairs and the three conditions yield 18 such TC sets. For instance, $\text{TCS}^{\text{C}}_{\text{VH4.1-C}\mu}$ is the TC set containing the 50 largest clonotypes for gene VH4.1-Cμ from the four fish in condition C. A TC set contains at most 200 TC since identical clonotypes can occur in several fish.

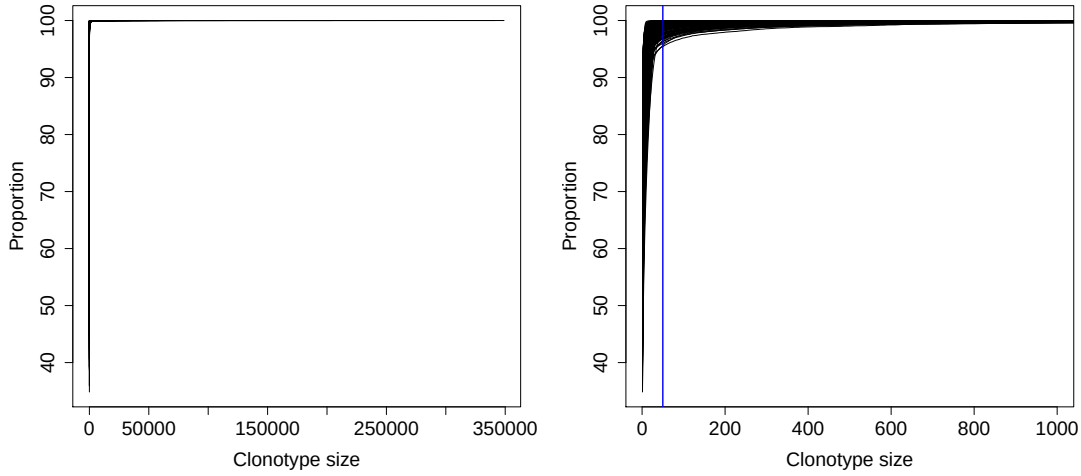
5.2.2 Characterization of the public and privates responses through repertoires comparison

We use the earth mover distance to compare pairs of repertoires. For this, two ingredients from each repertoire are needed: clonotypes with associated sizes and dissimilarity measures between two clonotypes. The first are straightforwardly obtained from sequencing data while the second are computed as defined below. We consider the 50 largest clonotypes for each fish.

Table 5.1 Dataset statistics. Seq: Number of sequences resulting from sequencing Clono: corresponding number of clonotypes (distinct CDR3 amino-acid sequences).

			C				E1				E4			
			1-1	1-4	1-5	1-7	3A1-R2	3A2	3A5	3B1	3A5	3A6	3A7	3A8
Cμ	VH4.1	Seq	32736	95939	83164	31409	2805	112355	74572	117823	37256	28492	53155	96160
		Clono	7473	22257	18973	5930	1448	32211	10669	13264	9436	7622	4425	28456
	VH5.1	Seq	577663	613067	929904	963886	15563	921995	714576	524406	298369	215815	331231	851702
		Clono	18074	32353	38053	16177	4406	43855	31215	28228	20238	14330	13339	54200
	VH8.1	Seq	1433801	1173399	1051013	568890	1237992	1254137	1073323	700591	973907	752533	616156	1574557
		Clono	45341	98138	62942	33478	165808	191558	84593	93342	113558	107693	52210	186471
Cτ	VH4.1	Seq	109753	225237	82671	179862	146393	172899	87082	269453	15050	17229	21285	374037
		Clono	44480	43092	41322	53515	18309	24038	23276	32891	6101	5026	8934	44435
	VH5.4	Seq	173694	311182	148688	245387	165575	438303	139405	142081	27818	20918	24812	333620
		Clono	52384	46346	58083	56646	22680	48104	39801	29328	9897	7732	11600	55517
	VH9.2	Seq	492250	696328	737376	641418	949909	733742	1006634	1157872	174338	194028	262638	905198
		Clono	126987	101372	181798	123584	106075	84234	179359	157148	47838	45926	85215	148981

Figure 5.1 Clonotype size distribution for all variable - constant gene pairs, and all fish. Left: whole distribution, right: zoom on smaller clonotypes sizes. Blue line: clonotype size of 50.



Definition of conservative substitutions. The selection pressure acting on Ig genes is different from that acting on general protein-encoding genes. For this reason, using regular substitution matrices (such as BLOSUM and PAM) to define conservative substitutions would likely give a biased view of the similarity between two protein sequences. Importantly, when two CDR3 sequences display a medium to low similarity, they are unlikely to target the same epitope. To take this into account as well as possible, the amino-acid classes defined in Table 5.2, specify which substitutions are conservative Ig protein sequences.

Table 5.2 Amino-acid classes and conservative substitutions. Amino acid in the same class define conservative substitutions.

Class name	Amino-acid code letter code
Tiny	A G S C N D T
Charged/acidic	Q N E K R D H
Aromatic	F W Y H
Aliphatic	I L V
Met	M
Pro	P

Sequence alignments. In order to get a *dissimilarity measure* between two amino-acid sequences, we align them using the Gotoh algorithm [Got93], namely a global alignment with affine gap costs and free ends. The substitution matrix is built using the conservative substitution rules defined from the amino-acid classes in Table 5.2. The gap opening score is set to -2 and the gap extension score is set to -1. The whole substitution matrix is given in Table 5.3.

The rationale for this non-standard matrix is that it assigns identical (and similar) sequences a score of 0, whereas dissimilar sequences get negative scores. We can then take the negative of this score to get a dissimilarity measure between sequences. In contrast, the identity matrix, giving 1 to conservative substitutions and 0 to non-conservative ones, would give different scores to pairs of identical sequences, depending on their length. The resulting dissimilarity measure is used when computing the earth mover distance(Eq. 5.3).

Practically, we use the `T_Alignment_engine_sequences_seqan` class from the Structural Bioinformatics Library (SBL, `sbl.inria.fr`) which is powered by the Seqan library [DWRR08].

Table 5.3 Amino-acid substitution matrix. Conservative substitutions get a score of 0, and non-conservative ones get a score of -1. Conservative substitutions are defined in table 5.2.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0	0	0	-1	-1	0	-1	-1	-1	-1	-1	0	-1	-1	-1	0	0	-1	-1	-1
C	0	0	0	-1	-1	0	-1	-1	-1	-1	-1	0	-1	-1	-1	0	0	-1	-1	-1
D	0	0	0	0	-1	0	0	-1	0	-1	-1	0	-1	0	0	0	0	-1	-1	-1
E	-1	-1	0	0	-1	-1	0	-1	0	-1	-1	0	-1	0	0	-1	-1	-1	-1	-1
F	-1	-1	-1	-1	0	-1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0
G	0	0	0	-1	-1	0	-1	-1	-1	-1	-1	0	-1	-1	-1	0	0	-1	-1	-1
H	-1	-1	0	0	0	-1	0	-1	0	-1	-1	0	-1	0	0	-1	-1	-1	0	0
I	-1	-1	-1	-1	-1	-1	-1	0	-1	0	-1	-1	-1	-1	-1	-1	-1	0	-1	-1
K	-1	-1	0	0	-1	-1	0	-1	0	-1	-1	0	-1	0	0	-1	-1	-1	-1	-1
L	-1	-1	-1	-1	-1	-1	-1	0	-1	0	-1	-1	-1	-1	-1	-1	-1	0	-1	-1
M	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1
N	0	0	0	0	-1	0	0	-1	0	-1	-1	0	-1	0	0	0	0	-1	-1	-1
P	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	-1
Q	-1	-1	0	0	-1	-1	0	-1	0	-1	-1	0	-1	0	0	-1	-1	-1	-1	-1
R	-1	-1	0	0	-1	-1	0	-1	0	-1	-1	0	-1	0	0	-1	-1	-1	-1	-1
S	0	0	0	-1	-1	0	-1	-1	-1	-1	-1	0	-1	-1	-1	0	0	-1	-1	-1
T	0	0	0	-1	-1	0	-1	-1	-1	-1	-1	0	-1	-1	-1	0	0	-1	-1	-1
V	-1	-1	-1	-1	-1	-1	-1	0	-1	0	-1	-1	-1	-1	-1	-1	-1	0	-1	-1
W	-1	-1	-1	-1	0	-1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0
Y	-1	-1	-1	-1	0	-1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0

Comparisons of VH CDR3 amino-acid repertoires using the earth mover distance. We model the repertoire of a fish as a distribution over sequences. The actual underlying probability distribution cannot be inferred directly from the data for two reasons: first the maximum length of the CDR is unknown; second and most importantly, the number of possible sequences is too large for the distribution to be parametrized by the data at hand. This precludes using classical information-theoretic measures such as the Jensen-Shannon divergence to compare repertoires. Therefore, we compare repertoires using the so-called earth-mover distance (EMD) [RTG00], also known as the Mallows distance in statistics [LB01], a particular Wasserstein metric used on optimal transportation theory [Vil03].

Assume that repertoire R is represented as a set of n sequences and their normalized counts, that is $R = \{(s_i, w_i)\}_{i=1,\dots,n}$, with $\sum_i w_i = 1$. Likewise, consider a second repertoire $R' = \{(s'_j, w'_j)\}_{j=1,\dots,n'}$ whose w'_i also sum to 1. For the sake of exposure, the quantities s_i are called the supply nodes and the weights w_i the supplies. Likewise, the s'_j are the demand nodes, and the weights w'_j the demands.

Also assume that we are given a dissimilarity measure $d(s_i, s'_j)$ to compare two sequences. Practically, we compute a sequence alignment between them and use the resulting score (see the previous paragraph). Note that this is not a metric since the scores returned by the alignment are only dissimilarity scores without the properties of metrics.

Using the previous ingredients, to compare the repertoires R and R', the EMD solves for the quantities $\{f_{ij}\}$, called *flows*, minimizing the following linear expression

$$C_{\text{EMD}} = \sum_{i=1,\dots,n, j=1,\dots,n'} f_{ij}\ d(s_i, s'_j), \tag{5.1}$$

under two constraints respectively expressing that a supply node exports all its mass, and that a demand notes has its demand satisfied:

$$\forall i : \sum_j f_{ij} = 1, \forall j : \sum_i f_{ij} = 1. \tag{5.2}$$

The previous two equations define a linear program (LP), that is a linear functional (Eq. (5.1)) minimized under linear constraints (Eq. (5.2)). This LP is known as a transportation problem in operations research. In solving it, exactly $n+n'-1$ flow variables f_{ij}, out of $n \times n'$ are involved in the definition of the optimum, as is easily shown using a transportation tableau [HL77, Fer00]. These variables define the *transport plan*, that is the amount of mass transported from sequence s_i to sequence s'_j. Note that the transport cost associated with f_{ij} is $f_{ij}\ d(s_i, s'_j)$.

Using the flows f_{ij}, one defines the earth mover distance as

$$d_{\text{EMD}} = C_{\text{EMD}} / \sum_{ij} f_{ij} \tag{5.3}$$

Since the w_i sum to 1, then $\sum_{ij} f_{ij} = 1$ and $d_{\text{EMD}} = C_{\text{EMD}}$.

Remark 1. In the particular case where $\sum_i w_i = \sum_j w_j$ and $d(.,.)$ is a metric, d_{EMD} is also a metric [LB01]. In that case, d_{EMD} is symmetric ($d_{\text{EMD}}(R, R') = d_{\text{EMD}}(R', R)$) and satisfies the triangle inequality.

5.2.3 Comparison of the earth mover distance and the Morisita-Horn distance

Both the earth mover distance (EMD) and the Morisita-Horn distance (MHD) can be used to compute distances between repertoires. The latter being commonly used, we benchmark the results from EMD against it.

The Morisita-Horn distance. The Morisita-Horn overlap index (MH for short) is a classical measure of the overlap between two samples classically used by ecologists and immunologists. Assuming a sample consists of *species* with an associated number of occurrences in the population. MH equals 0 if no species are common to the two samples, and 1 if both samples share the same species in the same proportions. In the context of repertoires, species correspond to clonotypes.

Consider two samples $\mathcal{A}$ and $\mathcal{B}$, *i.e.* two sets of sequences resulting from sequencing or processing thereof. The *size* $A = |\mathcal{A}|$ of the population $\mathcal{A}$ (and equivalently for $\mathcal{B}$) is the total number of distinct clonotypes in $\mathcal{A}$. Also $S = |\mathcal{A} \cup \mathcal{B}|$ is the total number of distinct clonotypes in both samples. The respective sizes of the clonotypes (*i.e.* the number of identical sequences corresponding to this clonotype) in $\mathcal{A}$ (resp. $\mathcal{B}$) are $\{a_1, \ldots, a_A\}$ (resp. $\{b_1, \ldots, b_B\}$). The Morisita-Horn overlap index is then defined follows:

$$MH(\mathcal{A}, \mathcal{B}) = 2 \cdot \frac{\sum_{i=1}^{S} \frac{a_i}{A} \cdot \frac{b_i}{B}}{\frac{\sum_{i=1}^{A} a_i^2}{A^2} + \frac{\sum_{i=1}^{B} b_i^2}{B^2}} \tag{5.4}$$

Equivalently, when working with clonotype frequencies $f(\mathcal{A})_i = \frac{a_i}{A}$ and $f(\mathcal{B})_i = \frac{b_i}{B}$:

$$MH(\mathcal{A}, \mathcal{B}) = 2 \cdot \frac{\sum_{i=1}^{S} f(\mathcal{A})_i \cdot f(\mathcal{B})_i}{\sum_{i=1}^{S} [f(\mathcal{A})_i^2 + f(\mathcal{B})_i^2]} \tag{5.5}$$

Since MH is always between 0 and 1, it can be transformed in a distance-like measure by subtracting it from 1:

$$MHD(\mathcal{A}, \mathcal{B}) = 1 - MH(\mathcal{A}, \mathcal{B}) \tag{5.6}$$

Distance matrix comparisons. For comparison purposes, we compute EMD and MHD distances between all pairs of fish for a given VC pair, resulting in 12 pairwise distance matrices in total (6 for each distance measure). We then compare both distance measures by studying the corresponding heatmaps (Figs. 5.3(c,f,i) and 5.4(c,f,i)). We also compare hierarchical clustering induced by these matrices. The dendrograms displayed in Figs. 5.3(b,e,h) and 5.4(b,e,h) are built using R with Ward's agglomeration criterion (minimization of total within cluster variance) on the (EMD or MHD) distance matrix (Sections 5.2.2 and 5.2.3).

5.2.4 Assessing how top clonotypes from a condition are shared between fish: TC search

Definition of similarity class. The notion of CDR3 lineages has been defined in mammals, and refers to sequences originating from the same B cell clone. Because of the somatic hypermutation process taking place during affinity maturation, CDR3 sequences from heavy chains of the same lineage may differ but are assumed to differ less that sequences synthesized from different V, D and J genes (because they target the same epitope with increased affinity along affinity maturation process). This notion allows one to follow the evolution of a clone during the response.

In fish, somatic hypermutation does not occur as much as in mammals and there is therefore no clear associated notion of lineages. However, because of the selection process taking place during challenge by an antigen, convergent evolution of CDR3 sequences is expected to take place and is indeed observed in public response to the virus [CJP+13]. It is therefore useful to define a notion of *similarity class*, in fish, analogous to that of lineages in mammals.

For this, we translate nucleotide sequences into amino-acid sequences, and, using the conservative substitution rules (Table 5.2) define the following *conservative substitution distance*.

Definition. 2. The distance $d_{cs}(p, q)$ between two sequences p and q, is equal to the number of conservative substitutions between p and q. In particular, sequences with differing length or with non-conservative mutations get a distance of ∞.

This distance is used to define the top clonotype counts (Def. 3).

Search and subsampling for TCs. We seek to quantify the overlap between TC sets and the clonotypes expressed by fish from all conditions. However, we face the following issue: the sequencing did not result in similar number of sequences for every fish. It is therefore necessary to perform a subsampling in order to bring the number of sequences on which the search for TCs is performed to the same level in every fish. In particular, as noticed in section 5.2.1, using frequencies is not an option. We proceed as follows: $s = 10000$ sequences are randomly drawn in every fish, and the presence / absence of a TC is assessed on these subsamples.

The results discussed thereafter are averages over M $(= 5000)$ subsamplings.

Definition. 3. Consider a subsample for each fish from condition Y, and a top clonotype t (Def. 1). Also consider a non-negative integer $\gamma = (0, 2)$. The *top clonotype count* $\mathrm{TCC}^{\gamma}_{\mathrm{Y}}$ at threshold γ of t is defined as the number of fish whose subsample contains a clonotype q such that $d_{cs}(t, q) \leq \gamma$ (Def. 2).

Equivalently[1], for subsamples $\mathcal{S}_x,\ x \in \{1, 2, 3, 4\}$ from four fish in condition Y:

$$\mathrm{TCC}^{\gamma}_{\mathrm{Y}}(t) = |\{\mathcal{S}_x \mid \exists q \in \mathcal{S}_x\ ,\ d_{cs}(t, q) \leq \gamma\}| \tag{5.7}$$

Prosaically, $\mathrm{TCC}^{\gamma}_{Y}$ is the number of fish from condition Y containing a given top clonotype, up to some distance tolerance γ. Note again that a TC count is defined with respect to a subsample, which imposes to assess the stability of TC counts amidst repeated subsamplings – we use M $(= 5000)$ of them.

Remark 2. Since the sequencing of VH4.1-Cμ 3A1-R2 only resulted in 2805 sequences, it is excluded form the subsampling step. It is however included in the TC sets.

Making Venn diagrams with a distance threshold. Venn diagrams in Fig. 5.7 display the number of identical sequences between top clonotype sets. However, when allowing a threshold on the sequence similarity as when $\gamma = 2$ (Fig. 5.8), making Venn diagrams is not possible since the relation is no longer symmetrical. (That is, the number of *neighbors* of a sequence $s \in S$ in a set T may not match the reverse.) We therefore proceed as follows: let TCS_A, TCS_B and TCS_C be the top clonotype sets corresponding to three conditions. Then we define the following counts:

$$A \to B = |\{x \in \mathrm{TCS}_A \mid \exists y \in \mathrm{TCS}_B,\ d_{cs}(x, y) \leq \gamma\}| \tag{5.8}$$

Note that in the previous equation, one uses are reference set a top clonotype TCS_A, while in the similar equation Eq. (5.7), one uses are reference sets the subsamplings $\mathcal{S}_x$.

$$A \to B, C = |\{x \in \mathrm{TCS}_A \mid \exists y \in \mathrm{TCS}_B,\ d_{cs}(x, y) \leq \gamma \text{ and } \exists z \in \mathrm{TCS}_C,\ d_{cs}(x, z) \leq \gamma\}| \tag{5.9}$$

With these notations, and A, B, C replaced by C, E1, E4 alternatively, the tables displayed in Fig. 5.8 follow the pattern shown in table 5.4.

Table 5.4 Pattern of the overlap table containing the number of shared top clonotypes (in Fig. 5.8).

$C \to E1, E4$	$C \to E1$	$C \to E4$
$E1 \to C$	$E1 \to C, E4$	$E1 \to E4$
$E4 \to E1$	$E4 \to E1$	$E4 \to C, E1$

[1] The size of a finite set X is denoted $|X|$.

5.2.5 Subsamples and their stability

Since sequencing did not result in comparable sequence counts for all fish (Table 5.1), we subsample the original data to bring these numbers to the same level. On one hand, large clonotypes are certain to be part of the resulting subsample (provided that it is not too small compared to the original population). On the other hand, some small clonotypes are going to be in the sample, whereas others of similar size will not, and this by chance alone.

We therefore study how the size of the clonotypes relate to their probability to be part of a subsample, how this translate to variability across subsamples, and how this may affect analyses using subsampled data.

Subsampling and urn models

Let $\mathcal{N}$ be the set of sequences resulting from the sequencing and $\mathcal{S}$ a random sample taken from $\mathcal{N}$. The size s ($= |\mathcal{S}|$) of $\mathcal{S}$, and size n ($= |\mathcal{N}|$) of $\mathcal{N}$ are their total number of sequences. The size m of a clonotype is the corresponding number of sequences in $\mathcal{N}$.

Thus, the probability for a clonotype of size m in the original population to be found k times in $\mathcal{S}$ is

$$P_k(n, m, s) = \frac{\binom{m}{k}\binom{n-m}{s-k}}{\binom{n}{s}} \tag{5.10}$$

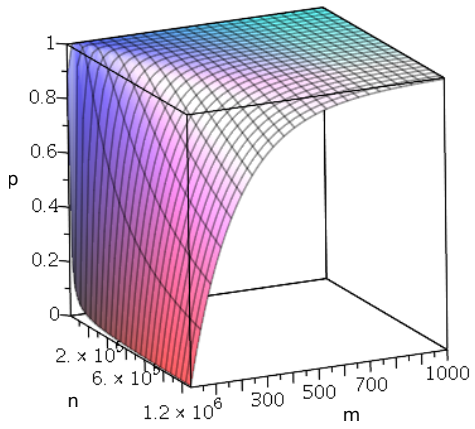
The probability for a clonotype of size m in the original population to be found in $\mathcal{S}$ at least once is therefore

$$P_{>0}(n, m, s) = 1 - \frac{\binom{n-m}{s}}{\binom{n}{s}} \tag{5.11}$$

Guaranteeing the presence of a clonotype with high probability. For sequencing data consisting of n sequences, we use Eq. (5.11) to compute the minimum size m^* a clone must have to be picked in a subsample of size s with probability p. We do so by solving numerically the equation $P_{>0}(n, m, s) = p$ for m (with Maple's function `fsolve`).

For instance, consider a population of $n = 5 \cdot 10^5$ sequences from which a subsample of $s = 7000$ sequences is drawn. Then the minimum size of a clonotype to be drawn with probability $p = 0.9$ is $m \approx 163$. This means that if a given clonotype of size $m < 163$ (in the original population) is not found in the subsample, it has a $> 10\%$ chance to actually be in the original population anyway. For $s = 7000$, this probability is computed for various m (1 to 1000) and n (7917 to 1290714 as in the dataset). (Fig. 5.2).

Figure 5.2 Probability for a clonotype of size m from a population of size n to be found at least once in a subsample of $s = 7000$ sequences.



For each sub-dataset, i.e. every gene from every fish in every condition (6 x 4 x 3 = 72 in total), with $s = 10000$ and $p = 0.99$, Table 5.5 shows statistics about n, m^* and the number and proportion of clonotypes of size greater than m^*.

Table 5.5 Clonotype representation during subsampling. Rows: minimum, first quartile, median, average, third quartile and maximum. n: number of sequences resulting from the sequencing, m^*: minimum size of a clonotype required for it to be found in a subsample of size $s = 10000$ from a population of size n with probability 0.99.

	n	m^*	# clonotypes $\geq m^*$	% clonotypes $\geq m^*$
Min	15050	5.0	1102	0.08
1st Qu	96105	45.5	6014	1.13
Median	266046	122.0	13980	8.32
Mean	443868	205.4	18804	13.83
3rd Qu	734650	337.0	28310	26.02
Max	1574557	723.0	58082	49.98

In the worst case, clonotypes of size < 723 may be absent from the subsample and still be present in the original population with a 1% probability. Avoiding these sampling effects by removing the clonotypes of size smaller than m^* can lead to a drastic reduction in size of the dataset. For instance, as little as 8% of the total clonotypes remain in the worst case. We therefore study the impact of the sampling effects on the presence or absence of sequences in the sample when small clonotypes are not filtered out.

Assessing the stability of top clonotypes

We investigate to which extent the previous findings affect actual results when looking for top clonotypes (TCs) in subsamples. For this purpose, we repeat the TC search protocol defined in Section 5.2.2. However, instead of recording TC counts, we record in which fish a TC has been found. Moreover, only exact matches between clonotypes are considered, i.e. we set $\gamma = 0$. This procedure is repeated M ($= 5000$) times.

Consider the set of fish in which a TC has been found. This set can be of size 0 to 12 since there are 12 fish in total. The total number of such sets is given by the powerset of the set of fish. This set has cardinality $\sum_{i=0}^{12} \binom{12}{i} = 2^{12} = 4096$. Note that the powerset contains the empty set because we search for a TC in subsamples. It is therefore possible (though unlikely) that a TC from a given fish will not found in a subsample from the very same fish. We define:

Definition. 4. The *powerset vector* $V(t) = \{v_1, \ldots, v_{4096}\}$ of a top clonotype t is defined as the vector of size 4096 whose i-th entry counts the number of times a subset of fish has been found to contain this top clonotype, upon M subsamplings.

The extreme cases give the rationale of the powerset vector:

- if the same set of fish is found M times, then, the inspected TC is stable and *abundant* in these fish. In that case, the powerset vector has a unique non-zero entry, equal to M.
- if on the opposite each subsampling yields a different set of fish, then the studied TC is not well represented in these fish.

We assess this stability using the entropy encoded in the powerset vector:

Definition. 5. Consider the powerset vector $V(t)$ of a top clonotype t, built over M ($= 5000$) repetitions of the sampling/subsampling process. Let $f_i = v_i/M$ be the frequency associated to the count v_i. We define the *stability of the clonotype* with the following *shrinkage* estimator of the Shannon entropy [HS09]:

$$H(V(TC)) = -\sum_{i=1}^{4096} \theta_i \ln \theta_i \tag{5.12}$$

with

$$\theta_i = \lambda^* \frac{1}{4096} + (1 - \lambda^* f_i) \tag{5.13}$$

where $\lambda \in [0, 1]$ is the shrinkage intensity parameter whose theoretical optimum λ^* is computed as follows:

$$\lambda^* = \frac{1 - \sum_{i=1}^{4096} f_i^2}{(4096 - 1) \sum_{i=1}^{4096} (\frac{1}{4096} - f_i)^2} \tag{5.14}$$

We now shortly discuss the rationale for choosing this estimator instead of the simple maximum likelihood estimator (MLE):

$$H_{MLE}(V(TC)) = -\sum_{i=1}^{4096} f_i \ln f_i. \tag{5.15}$$

The terms f_i are frequencies *i.e.* maximum likelihood estimators of the underlying probabilities p_i. Accordingly, Eq. 5.15 is a MLE of the entropy. This estimator (also called plugin estimator) is known to be negatively biased, *i.e.* to systematically underestimate the entropy. This bias becomes large when the vector V is sparse (*i.e.* has many terms equal to 0), which is the case here with only 5000 observations for a vector of size 4096. Although it has been shown that no unbiased estimator of entropy exists [Pan03], multiple studies have sought better entropy estimators than the MLE. The shrinkage estimator essentially replaces the MLE estimators f_i of p_i by a version θ_i which is regularized toward the maximum entropy target $1/M$. The regularization intensity λ^* is set in a data-dependent manner. The maximum entropy equal to $-\ln(1/4096) \approx 8.32$ occurs for the uniform distribution ($f_i = 1/4096$) in which case $H_{MLE} = H$.

For each VC pair and each condition, this process assigns each TC an associated entropy H across M subsamples. It is worth noticing that the entropy of a given clonotype is influenced by two factors:

- the number of fish containing it,
- and the size of this clonotype in these fish.

For instance, a clonotype found in two fish will have a low entropy if it is large in both fish because it will always be found in both fish in the subsamples. It will have a greater entropy (with a maximum value of $-\ln(1/4) \approx 1.39$) if it is small because it will alternatively appear in either fish, no fish, or both. On the other hand, a clonotype found in many fish will have a low entropy if it is large in all of them because it will always be found in the same combination of fish in the subsamples. On the other hand, it will have a much larger entropy if it is small because it will appear in many different fish combinations across subsamples. For visualization purposes, and since the process described above results in 18 tables of up to 200 rows (3 TC sets for 6 VC pairs, 200 TC for each), we present an aggregate representation. For each TC (Def. 1), we compute the previous entropy, and estimate the corresponding density (Figs 5.9 and 5.10). We also report a scatterplot of the entropy H versus the number z of of non-zero elements in V, *i.e.* the number of combinations in which a TC has been found (Fig. 5.11).

5.3 Results

5.3.1 Characterization of the public and privates responses through repertoires comparison

To test the validity of our approach to compare repertoires, we consider a model in which public and private responses have been identified and characterized. The public response is defined for a VC pair and is a response where multiple clonotypes are shared between most fish with identical genetic background (as is the case in this study). The private response, in contrast, is a response where most clonotypes are found in a small subset of the fish population.

The response of rainbow trout to VHSV has been studied using deep sequencing in a prior work [CJP+13]. In particular, VH5.1-Cμ was shown to engage in public response, VH4.1-Cμ and VH4.1-Cτ were shown to participate in private responses. In the present dataset, we study VH8.1-Cμ, VH5.4-Cτ and VH9.2-Cτ as examples of VC pairs involved in weaker responses to the pathogen.

We now analyze the comparisons obtained on the 50 largest clonotypes of each fish.

Characterizing the public response. Considering VH5.1-Cμ (public response), the distance between fish in condition C are larger than those in condition E1, which are in turn larger than those in condition E4 (Fig. 5.3d(left part)). It also appears that the distances go below 0.7 and 0.5 respectively for VH5 in conditions E1 and E4 only compared to other genes and conditions, indicating an increased similarity upon vaccination and infection not only relatively to condition C but also in an absolute sense.

Comparing intra versus inter conditions, for conditions E1 and E4, the former distances are smaller than the latter between all pairs of conditions except for one outlier in E1. This translates as separated clusters of individuals from the same condition (Fig. 5.3f) The infection therefore leads to a convergence of the public response.

Finally, distances between C and E1 are greater on average than those between E1 and E4, and smaller than those between C and E4. Therefore, upon infection and re-infection, the public response repertoires become increasingly distinct from the naive repertoire.

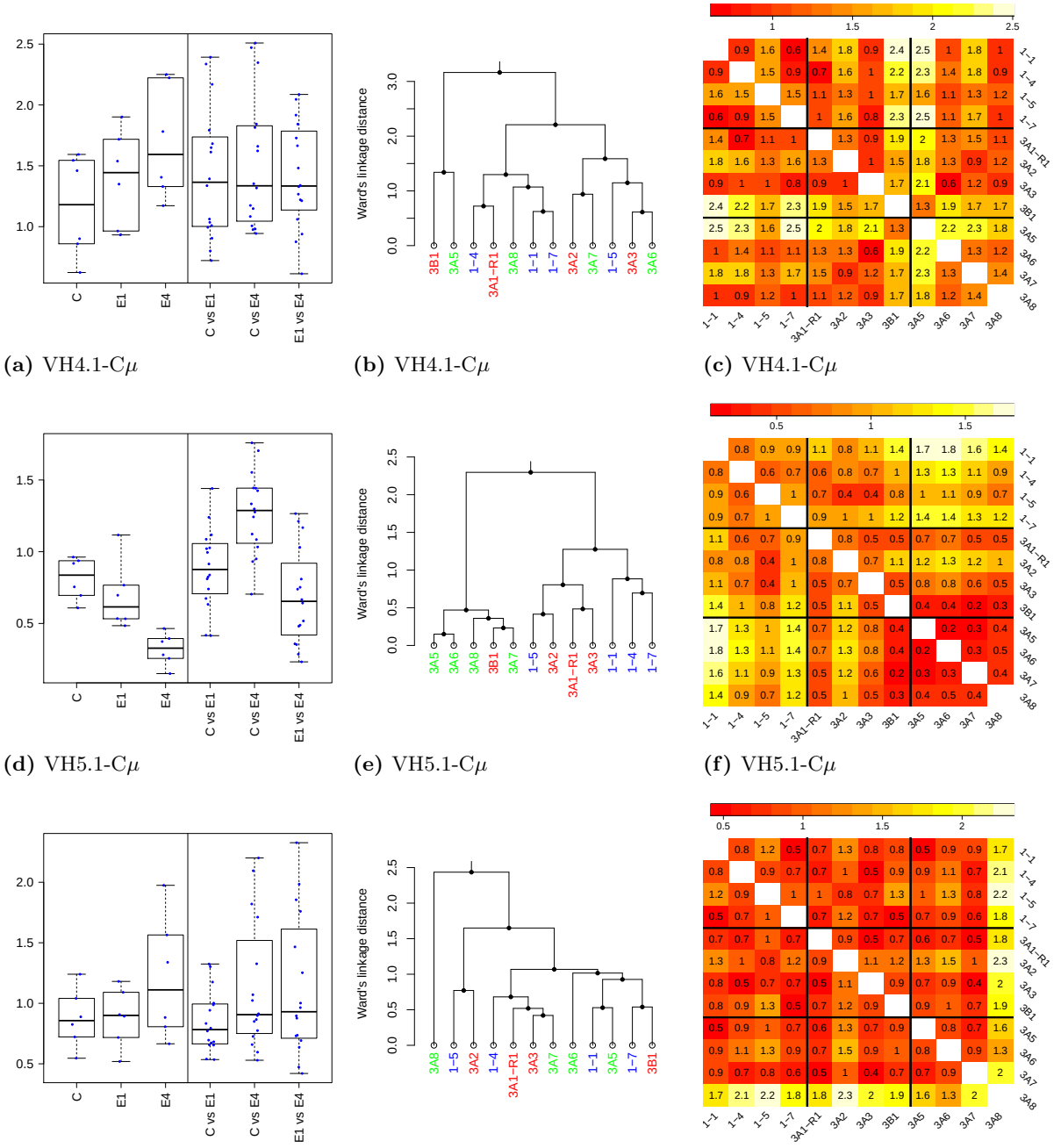
Concluding, the repertoire of naive fish are more different than those of fish exposed to the same pathogen. Moreover, memory responses yield even smaller distances (albeit less so). VH5 is therefore able to quantify the similarity of the (public) response to the pathogen either within or between conditions.

Characterizing the private response. The results for VH4.1-Cμ (Fig. 5.4a) are typical of a private response. Intra condition distances are expected to vary little between conditions since vaccination and infection trigger disjoint responses in all fish, which is what we observe. The slight increase in intra distance when moving from C to E1 and to E4 is likely due to the appearance of few large clonotypes, dissimilar between fish. Inter conditions distances do not vary as expected since every fish develops a different response.

The results for VH4.1-Cτ, suggest an intermediate situation between public and private responses. Inter condition distances show that a component is common between fish in E1 and E2 as the distance between them tend to be lower than the distances of either with C. However, distances from C to E1 and C to E2 are similar indicating a private component to the response of VH4.1-Cτ. This is clear in Fig. 5.4a, where apart from one infected fish, all naive fish are clustered together.

Other VC pairs The results for VH8.1-Cμ, VH5.4-Cτ and VH9.2-Cτ are less clear which is likely due to their weak response to infection. For the former, should a response occurs, it is mostly private as can be assessed by its similarities with VH4.1-Cμ.

Figure 5.3 IgM. (a,d,g) Distances distribution for fish within a condition (left) and between conditions (right). y-axis: earth mover distance distance; C: naive, E1: vaccinated, E4: vaccinated + infected. Each point represents a distance between two fish (jittered x-axis). **(b,e,h) Dendrogram built using Ward's method on the earth mover distance distance matrix. (c,f,i) Earth mover distance matrix.** Heatmap colors range from white (high values) to red (low values).

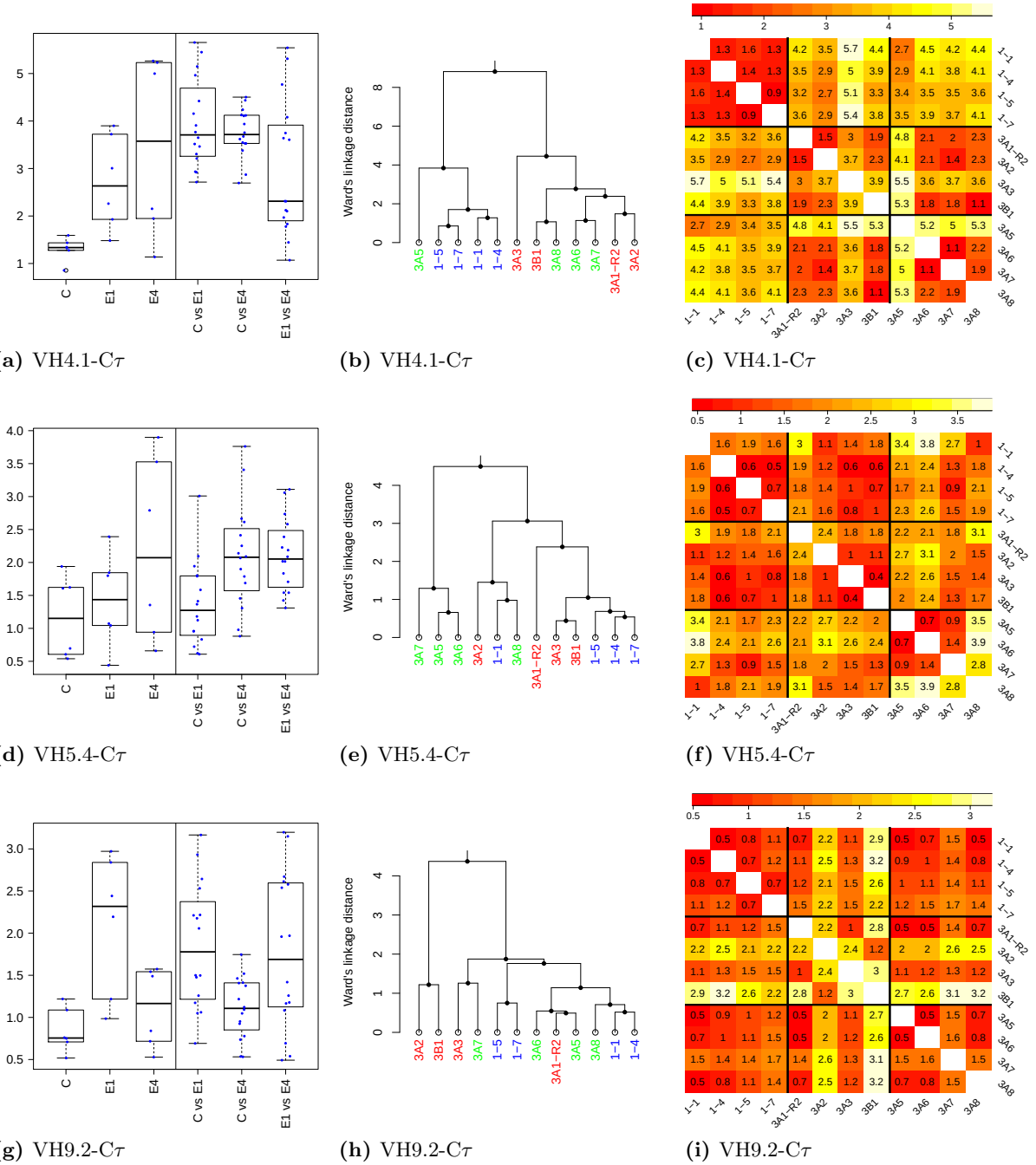


(a) VH4.1-Cμ **(b)** VH4.1-Cμ **(c)** VH4.1-Cμ

(d) VH5.1-Cμ **(e)** VH5.1-Cμ **(f)** VH5.1-Cμ

(g) VH8.1-Cμ **(h)** VH8.1-Cμ **(i)** VH8.1-Cμ

Figure 5.4 IgT. (a,d,g) Distances distribution for fish within a condition (left) and between conditions (right). y-axis: earth mover distance distance; C: naive, E1: vaccinated, E4: vaccinated + infected. Each point represents a distance between two fish (jittered x-axis). **(b,e,h) Dendrogram built using Ward's method on the earth mover distance distance matrix. (c,f,i) Earth mover distance matrix.** Heatmap colors range from white (high values) to red (low values).



(a) VH4.1-Cτ **(b)** VH4.1-Cτ **(c)** VH4.1-Cτ

(d) VH5.4-Cτ **(e)** VH5.4-Cτ **(f)** VH5.4-Cτ

(g) VH9.2-Cτ **(h)** VH9.2-Cτ **(i)** VH9.2-Cτ

5.3.2 Comparison of the earth mover distance and the Morisita-Horn distance

The Morisita-Horn overlap index is a classical quantity used to assess the amount of overlap between individuals while taking into account the diversity of their respective repertoires. As such, it does not

take into account the similarity between sequences.

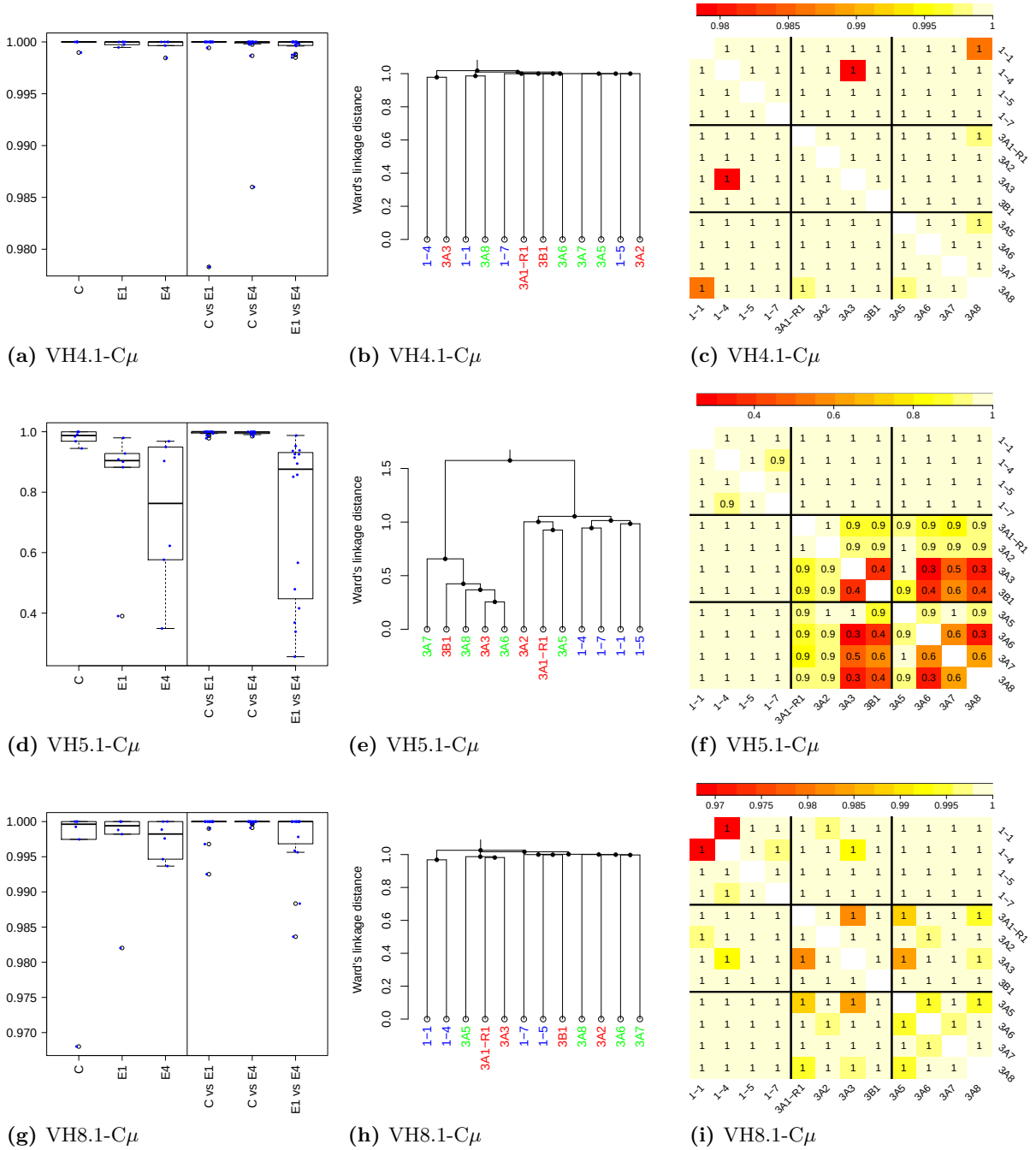
We perform the previous analysis with the distance associated to MH, called MHD, (Figs. 5.5 and 5.6) instead of the EMD and compare it to the results of the previous section.

For VH4.1-Cμ, the results are very different. In particular, the MHD if very close to 1 between all fish whereas we can see that vaccination and infection increase the within-condition EMD. For VH5.1-Cμ, both distances show that the response has a public component although EMD shows a more progressive convergence when moving to vaccination and then infection. Moreover, both methods show some ability to distinguish naive from vaccinated/infected fish, although the separation is more clear for EMD. For VH8.1-Cμ, both methods indicate no obvious changes upon infection although EMD, identify one infected fish as very different from all others.

For VH4.1-Cτ, both methods identify a fish-specific specialization of repertoires upon challenge. MHD detects no difference in intra condition distances, whereas EMD shows a progressive divergence upon vaccination then infection. Both methods are somewhat able to distinguish naive from vaccinated/infected fish, but this separation is clearer for MHD. For VH5.4-Cτ and VH9.2-Cτ, MHD and EMD give qualitatively comparable results.

Overall, EMD is able to capture finer details when comparing the repertoire of two fish. Namely, it can identify progressive convergence/divergence upon vaccination and subsequent infection.

Figure 5.5 IgM. (a,d,g) Distances distribution for fish within a condition (left) and between conditions (right). y-axis: Morisita-Horn distance; C: naive, E1: vaccinated, E4: vaccinated + infected. Each point represents a distance between two fish (jittered x-axis). **(b,e,h) Dendrogram built using Ward's method on the Morisita-Horn distance distance matrix. (c,f,i) Morisita-Horn distance matrix.** Heatmap colors range from white (high values) to red (low values).

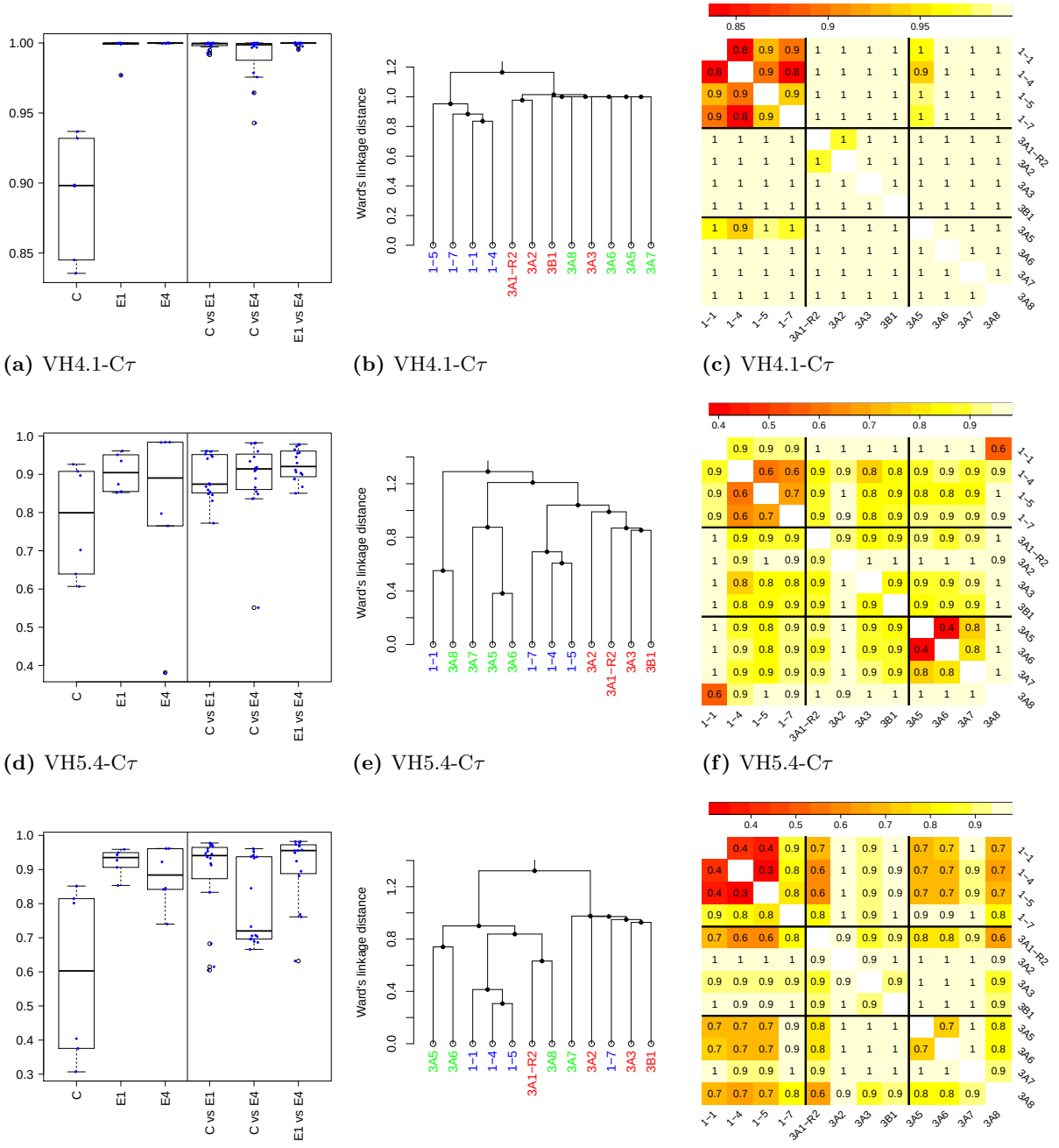


(a) VH4.1-Cμ **(b)** VH4.1-Cμ **(c)** VH4.1-Cμ

(d) VH5.1-Cμ **(e)** VH5.1-Cμ **(f)** VH5.1-Cμ

(g) VH8.1-Cμ **(h)** VH8.1-Cμ **(i)** VH8.1-Cμ

Figure 5.6 IgT. (a,d,g) Distances distribution for fish within a condition (left) and between conditions (right). y-axis: Morisita-Horn distance; C: naive, E1: vaccinated, E4: vaccinated + infected. Each point represents a distance between two fish (jittered x-axis). **(b,e,h) Dendrogram built using Ward's method on the Morisita-Horn distance distance matrix. (c,f,i) Morisita-Horn distance matrix.** Heatmap colors range from white (high values) to red (low values).



(a) VH4.1-Cτ **(b)** VH4.1-Cτ **(c)** VH4.1-Cτ

(d) VH5.4-Cτ **(e)** VH5.4-Cτ **(f)** VH5.4-Cτ

(g) VH9.2-Cτ **(h)** VH9.2-Cτ **(i)** VH9.2-Cτ

5.3.3 Assessing how top clonotypes from a condition are shared between fish: TC search

Focusing on the largest clonontypes which are likely to reflect the response to the virus, it is natural to ask whether such a large clonotype from a fish from one condition is shared by fish from the same or

other conditions. For this, we define top clonotype sets (TCSs) (Def. 1) for each VC pair and condition. Each top clonotype (TC) in the corresponding TCS is assigned a top clonotype count for each condition: $\mathrm{TCC}^{\gamma}_{\mathrm{C}}$, $\mathrm{TCC}^{\gamma}_{\mathrm{E1}}$ and $\mathrm{TCC}^{\gamma}_{\mathrm{E4}}$ (Def. 3). We discuss two sets of results: first for $\gamma = 0$ (Fig. 5.7), then when $\gamma = 2$ (Fig. 5.8).

Case $\gamma = 0$. The difference between the public and private response is clear in Fig. 5.7. Namely, Very few TCs from $\mathrm{TCS}^{\mathrm{E1}}_{\mathrm{VH4.1\text{-}C}\mu}$ and $\mathrm{TCS}^{\mathrm{E4}}_{\mathrm{VH4.1\text{-}C}\mu}$ are shared. The same is true for $\mathrm{TCS}^{\mathrm{E1}}_{\mathrm{VH4.1\text{-}C}\tau}$ and $\mathrm{TCS}^{\mathrm{E4}}_{\mathrm{VH4.1\text{-}C}\tau}$. Although few TCs are found in most fish, these are found in naive fish as often as in vaccinated / infected fish. The difference between VH4.1-Cμ and VH4.1-Cτ, is that, contrary to $\mathrm{TCS}^{\mathrm{C}}_{\mathrm{VH4.1\text{-}C}\mu}$, $\mathrm{TCS}^{\mathrm{C}}_{\mathrm{VH4.1\text{-}C}\tau}$ shares multiple TCs with several fish from all conditions, suggesting a partially common naive repertoire.

For VH5.1-Cμ, we see that 10-20 TCs from $\mathrm{TCS}^{\mathrm{E1}}_{\mathrm{VH5.1\text{-}C}\mu}$ and $\mathrm{TCS}^{\mathrm{E4}}_{\mathrm{VH5.1\text{-}C}\mu}$ are shared by most fish. Importantly, they are only shared between vaccinated or infected fish, which corresponds to the public response.

The barplots for VH8.1-Cμ look very similar to those of VH4.1-Cμ. Namely, the small TC counts for TCs in $\mathrm{TCS}^{\mathrm{C}}_{\mathrm{VH8.1\text{-}C}\mu}$ indicate distinct naive repertoires. However, it cannot be assessed from this data only whether VH8.1-Cμ responds to the challenge. Thus, assuming VH8.1-Cμ does not respond to the challenge, the small TC counts for TCs from $\mathrm{TCS}^{\mathrm{E1}}_{\mathrm{VH5.1\text{-}C}\mu}$ and $\mathrm{TCS}^{\mathrm{E4}}_{\mathrm{VH8.1\text{-}C}\mu}$ are simply a consequence of distinct naive repertoires remaining distinct since they do not respond. Assuming it does respond to the challenge, it would indicate a private response as for VH4.1-Cμ.

In the case of VH5.4-Cτ and VH9.2-Cτ, a large proportion of TCs are shared by fish from all condition, suggesting a common naive repertoire. Upon vaccination and infection however, few TCs from $\mathrm{TCS}^{\mathrm{E1}}_{\mathrm{VH5.4\text{-}C}\tau}$, $\mathrm{TCS}^{\mathrm{E4}}_{\mathrm{VH5.4\text{-}C}\tau}$, $\mathrm{TCS}^{\mathrm{E1}}_{\mathrm{VH9.2\text{-}C}\tau}$, $\mathrm{TCS}^{\mathrm{E4}}_{\mathrm{VH9.2\text{-}C}\tau}$ are only found in a single fish, indicating a potential private component. However, because of the low response of both VC pairs to infection, it is difficult to draw further conclusions.

Case $\gamma = 2$. When allowing up to two conservative substitutions, the picture changes and reveals additional details (Fig. 5.8).

Comparing VH4.1-Cμ and VH4.1-Cτ, we see that, for all conditions, the former shares fewer TCs with other fish than the latter. However in this case, these TCs are shared by fish from all conditions. This suggests than clonotypes in VH4.1-Cτ are more similar than those in VH4.1-Cμ.

In the case of VH5.1-Cμ, we see that ~ 100 TCs from $\mathrm{TCS}^{\mathrm{E1}}_{\mathrm{VH5.1\text{-}C}\mu}$ and $\mathrm{TCS}^{\mathrm{E4}}_{\mathrm{VH5.1\text{-}C}\mu}$ are shared by most fish from either condition. Interestingly, these are shared by naive fish as well, and the situation is the same for $\mathrm{TCS}^{\mathrm{C}}_{\mathrm{VH5.1\text{-}C}\mu}$. Out of the clonotypes which are similar to ~ 100 top clonotypes from $\mathrm{TCS}^{\mathrm{C}}_{\mathrm{VH5.1\text{-}C}\mu}$, 45 are top clonotypes which also belong to $\mathrm{TCS}^{\mathrm{E1}}_{\mathrm{VH5.1\text{-}C}\mu}$ and $\mathrm{TCS}^{\mathrm{E4}}_{\mathrm{VH5.1\text{-}C}\mu}$ (Fig. 5.8(B)). Similarly, out of the clonotypes which are similar to ~ 100 top clonotypes from $\mathrm{TCS}^{\mathrm{E1}}_{\mathrm{VH5.1\text{-}C}\mu}$, 48 are top clonotypes which also belong to $\mathrm{TCS}^{\mathrm{C}}_{\mathrm{VH5.1\text{-}C}\mu}$ and $\mathrm{TCS}^{\mathrm{E4}}_{\mathrm{VH5.1\text{-}C}\mu}$ (Fig. 5.8(B)). Finally, out of the clonotypes which are similar to ~ 100 top clonotypes from $\mathrm{TCS}^{\mathrm{E4}}_{\mathrm{VH5.1\text{-}C}\mu}$, 64 are top clonotypes which also belong to $\mathrm{TCS}^{\mathrm{C}}_{\mathrm{VH5.1\text{-}C}\mu}$ and $\mathrm{TCS}^{\mathrm{E1}}_{\mathrm{VH5.1\text{-}C}\mu}$ (Fig. 5.8(B)). Therefore, many top clonotypes involved in the public response are very similar to large clonotypes from naive fish. They are not identical however as this number is only 6 for $\gamma = 0$ (Fig. 5.7, Venn diagram)

As in the previous paragraph, the weak response of VH8.1-Cμ does not allow us to draw conclusions using these data alone.

For VH5.4-Cτ and VH9.2-Cτ, many TCs are shared by all fish from all conditions, reinforcing the thesis of a lack of response. Few TCs from $\mathrm{TCS}^{\mathrm{E1}}_{\mathrm{VH5.4\text{-}C}\tau}$ and $\mathrm{TCS}^{\mathrm{E1}}_{\mathrm{VH9.2\text{-}C}\tau}$ occur in a single fish as in Fig. 5.7. This is no longer the case for $\mathrm{TCS}^{\mathrm{E4}}_{\mathrm{VH5.4\text{-}C}\tau}$ and $\mathrm{TCS}^{\mathrm{E4}}_{\mathrm{VH9.2\text{-}C}\tau}$ however, meaning that this TCs were very similar to other TCs shared between all fish.

Overall, four clear types of behaviour are observed: 1) all fish have distinct clonotypes from the start and evolve diverging responses upon challenge (VH4.1-Cμ); 2) All fish share a large enough ($\sim$50%) set of clonotypes and evolve converging responses upon challenge (VH5.1-Cμ; 3) All fish share a large enough ($\sim$50%) set of clonotypes and evolve diverging responses anyway upon challenge (VH4.1-Cτ); 4) Naive fish share many TCs and this does not change upon infection because of the weak response of the VC pairs (VH5.4-Cτ and VH9.2-Cτ).

Figure 5.7 Quantification of repertoire overlap using top clonotypes. Color code: blue = naive (C), red = vaccinated (E1), green = infected (E4). Each line corresponds to a variable - constant gene pair. **(A) Number of top clonotypes with a given top clonotype count.** Each column corresponds to the condition of top clonotype sets. Abscissa: values correspond to top clonotypes counts (see Def. 3 with $\gamma = 0$). Ordinates: each bar corresponds to the average (over subsamplings) number of top clonotypes with a given clonotype count. The standard deviation is displayed with an error bar. **(B) Venn diagrams: number of top clonotypes shared between the top clonotype sets.**

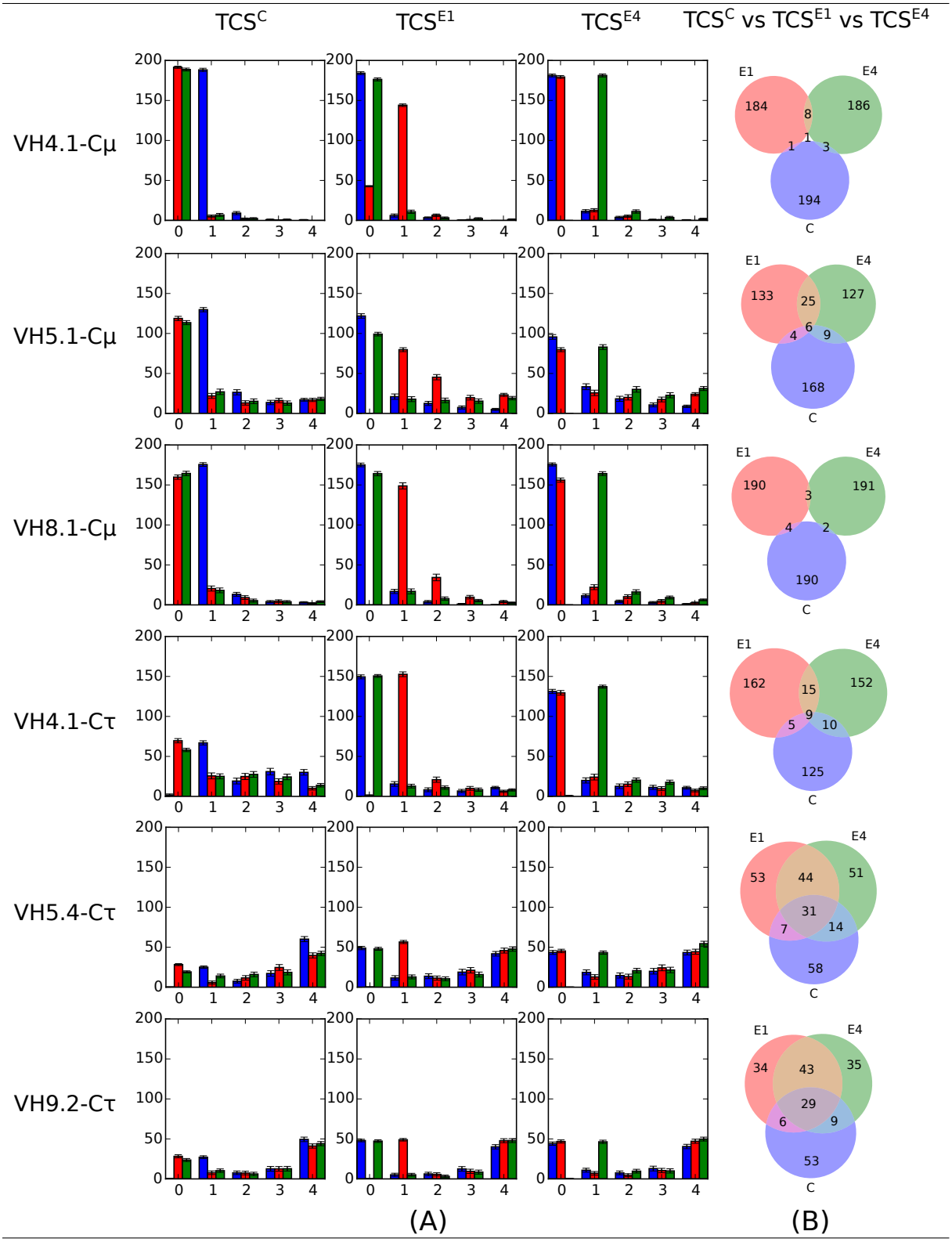
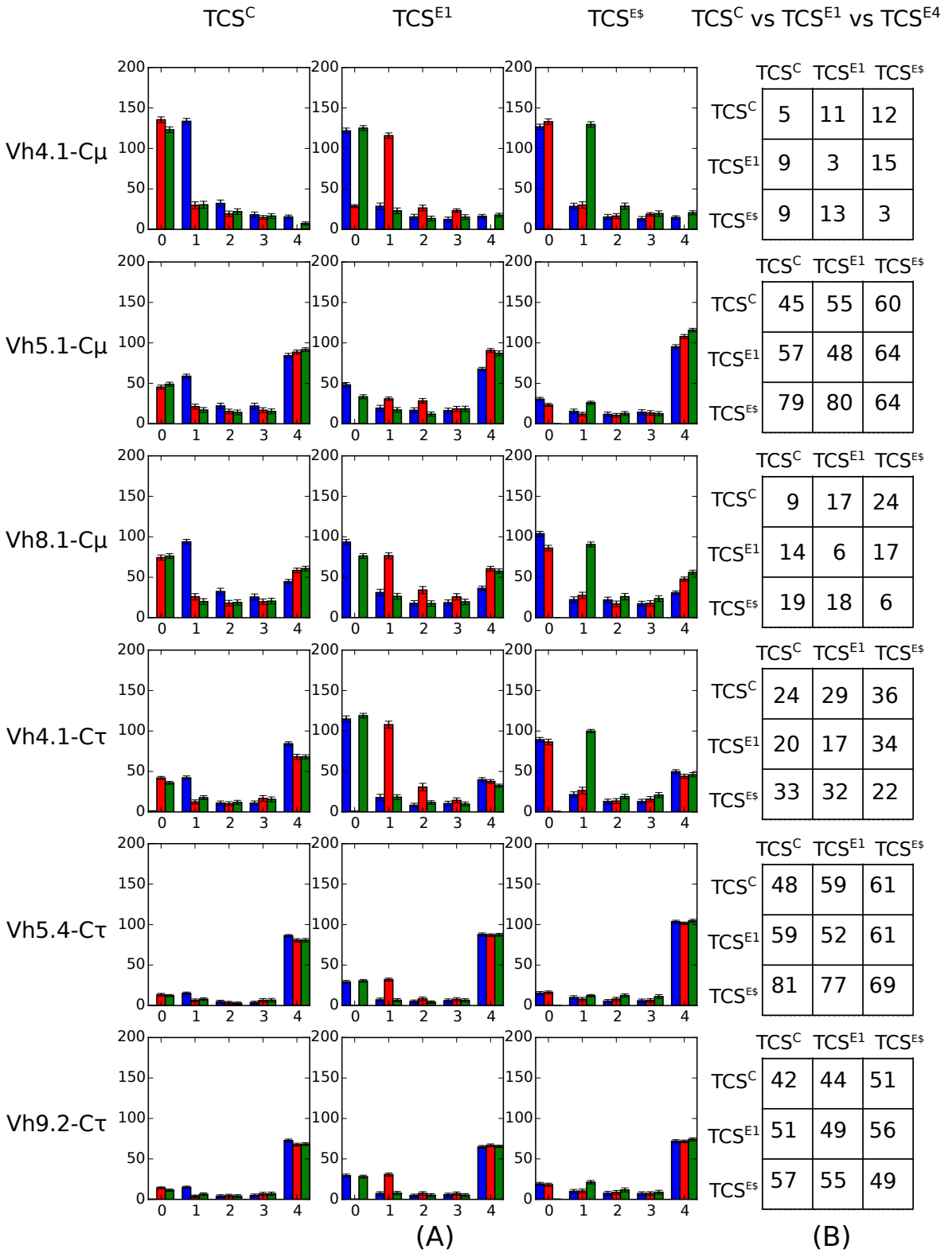


Figure 5.8 Quantification of repertoire overlap using top clonotypes. Color code: blue = naive (C), red = vaccinated (E1), green = infected (E4). Each line corresponds to a variable - constant gene pair. **(A) Number of top clonotypes with a given top clonotype count.** Each column corresponds to the condition of top clonotype sets. Abscissa: values correspond to top clonotypes counts (see Def. 3 with $\gamma = 2$). Ordinates: each bar corresponds to the average (over subsamplings) number of top clonotypes with a given clonotype count. The standard deviation is displayed with an error bar. **(B) Top clonotypes overlap table.** A cell in row x, column y with $x \neq y$ contains the number of clonotypes shared by the top clonotypes sets x and y. For $x = y$, it contains the number of clonotype shared by x and the two other clonotype sets. See Table 5.4.



	TCS^C	TCS^E1	TCS^E$
TCS^C	5	11	12
TCS^E1	9	3	15
TCS^E$	9	13	3

	TCS^C	TCS^E1	TCS^E$
TCS^C	45	55	60
TCS^E1	57	48	64
TCS^E$	79	80	64

	TCS^C	TCS^E1	TCS^E$
TCS^C	9	17	24
TCS^E1	14	6	17
TCS^E$	19	18	6

	TCS^C	TCS^E1	TCS^E$
TCS^C	24	29	36
TCS^E1	20	17	34
TCS^E$	33	32	22

	TCS^C	TCS^E1	TCS^E$
TCS^C	48	59	61
TCS^E1	59	52	61
TCS^E$	81	77	69

	TCS^C	TCS^E1	TCS^E$
TCS^C	42	44	51
TCS^E1	51	49	56
TCS^E$	57	55	49

5.3.4 Subsamples and their stability

Aside from the previous analysis, we now study how subsampling can affect the TC search for small clonotypes. Although it is not directly relevant for the previous results, because only the *number* of fish in which TCs are found matters, it will be important to consider when looking for more details, namely *in which* they are found.

To find out to which extent sampling effects on small clonotypes affect our results when performing TC search (section 5.3.3), we look at how the set of fish in which a TC is found changes between subsamples. The procedure described in Section 5.2.5 assigns a entropy value H to each TC (Def. 5). For each top clonotype set, the estimated densities of the distributions of H (Def. 5) are plotted (Figs 5.9 and 5.10).

For VH4.1-Cμ, most TCs have an associated entropy less than 1.39 ($\approx -\ln(1/4)$), indicating that they randomly appear in at least four combinations of fish across subsamples.

For VH8.1-Cμ, the curves for conditions C (blue) and E4 (green) show two peaks, centered at ~ 0.14 and ~ 1. These corresponds to two subpopulations of TCs, the first with very low entropy indicating they are found in at least two combinations of fish ($-\ln(1/2) \approx 0.69$), and the other with a somewhat higher entropy corresponding to at least four fish ($1.39 \approx -\ln(1/4)$). The curve for condition E1 spans approximately the same entropy range as those of C and E4 with only one peak around 0.14. Over all conditions, most TCs have entropy values less than 1.6 ($\approx -\ln(1/5)$) corresponding to at least five different fish combinations.

For VH4.1-Cτ, the curve for conditions E1 (red) shows two peaks, one centered at ~ 0.1 and another around 0.9. The curve for E4 (green) shows only one peak centered around 0.9 as well. The first peak (E1 only) corresponds to TCs with low entropy indicating they are found in at least two combinations of fish ($-\ln(1/2) \approx 0.69$). The other peak (E1 and E4) spreads until entropy values of ~ 2 corresponding to at last eight ($2.01 \approx -\ln(1/8)$) combinations of fish. For condition C (blue), to broad peaks are observed the first one extending between 0 and ~ 2.5 and the second between ~ 2.5 and ~ 6.2. The first peak corresponds to TCs with low to high variability; to get an idea, TCs with an entropy of 2.48 ($\approx -\ln(1/12)$) can be found in at least 12 different fish combinations. The second peak corresponds to TCs with very high variability since TCs with an entropy of 6.2 ($\approx -log(1/492)$) can be found in at least 492 fish combinations. Interestingly, TCs from the corresponding TC set $\text{TCS}^{\text{C}}_{\text{VH4.1-C}\tau}$ are shared between multiple fish (Figs. 5.7 and 5.8)

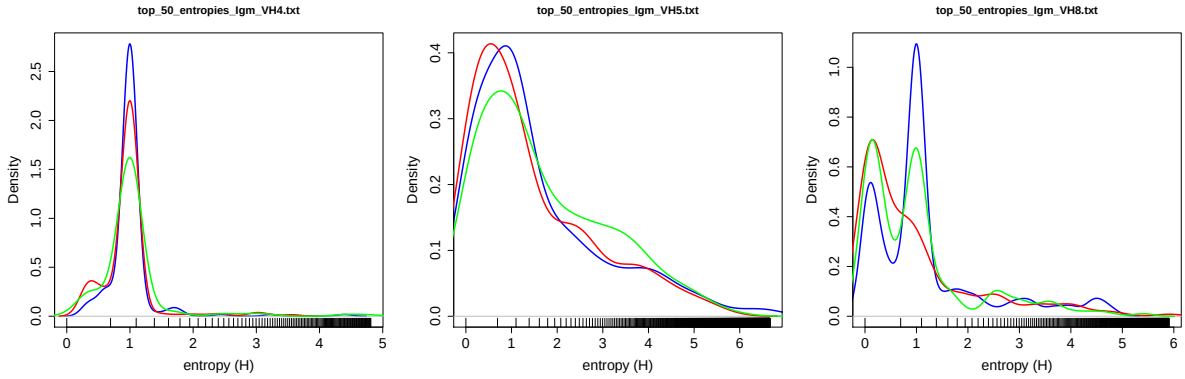
For VH5.1-Cμ, VH5.4-Cτ and VH9.2-Cτ a large proportion of TCs have associated entropies greater than 2 ($\approx -\ln(1/8)$) independently from the condition. This corresponds to random occurrences in more than eight combinations of fish across subsamples. As for VH4.1-Cτ, notice how this matches with the proportion of shared clonontypes (Figs. 5.7 and 5.8)

Overall, we see that high entropy correlates with the fact that TCs are shared among several fish. This suggests that clonotypes contributing to the bars in Figs. 5.7 and 5.8 for $\text{TCS}^{\text{C}}_{\text{VH4.1-C}\tau}$ and VH5.1-Cμ, VH5.4-Cτ and VH9.2-Cτ (for all conditions) are not the same for each subsample. Repeated subsamplings are therefore required to get an accurate picture of clonotypes occurrences in fish.

Interestingly, this variation is not observed when performing TC search as seen from the small error bars in Figs. 5.7 and 5.8. This may be due to the fact that multiple TCs occurring in different fish over multiple subsamples cancel each other out when aggregating the resulting number of TCs found in a given number of fish. For instance, assuming TC1 is found in fish A in subsample 1 and in fish B in subsample 2, this will be canceled by another TC2 found in fish B first, then in fish A.

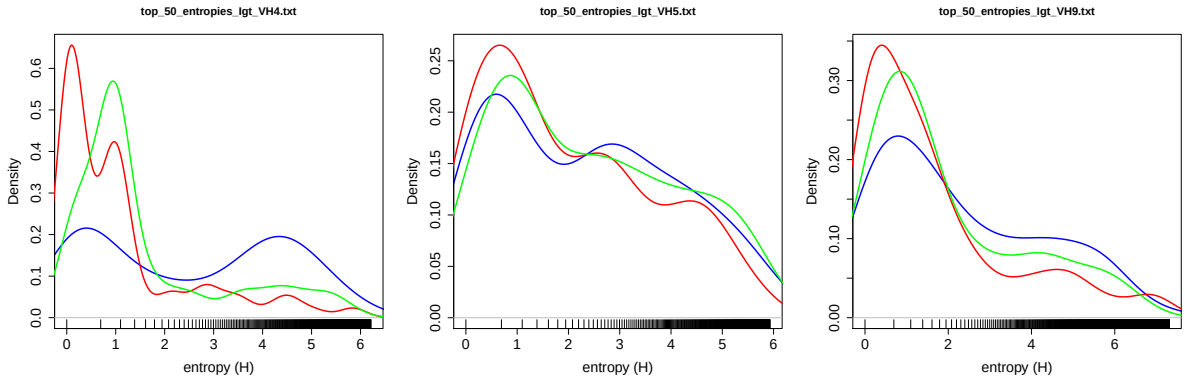
As an aside, Fig. 5.11 shows that for a given number z of non-zero elements in a powerset vector V (*i.e.* the number of combinations of fish in which a given TC is found), the entropy is often lower than the maximum entropy $-\ln(1/z)$. This indicates that for a TC found in z fish, it is found more often in a subset of these fish than in others subsets.

Figure 5.9 Estimated density of the distribution of the entropy H of the powerset vector (Def. 5) for IgM. Left: VH4.1-Cμ, middle: VH5.1-Cμ, right: VH8.1-Cμ. Blue: condition C, Red, condition E1, green: condition E4. Ticks at the bottom show the maximum entropy for combinations of at least k fish, *i.e.* frequencies of occurence of (0.5, 0.5) for $k = 2$, (1/3, 1/3, 1/3) for $k = 3$ and so on.



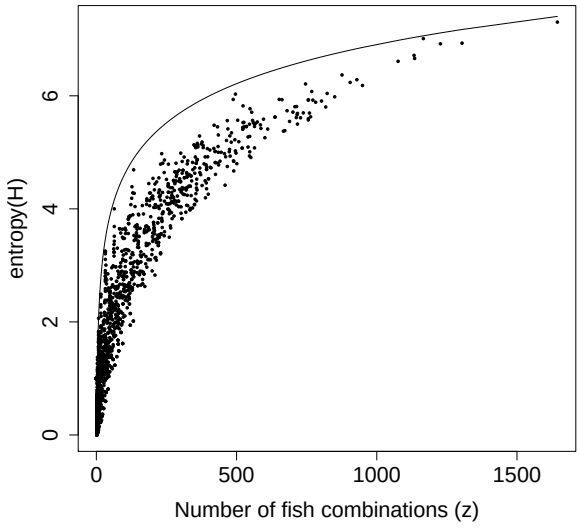
(a) VH4.1-Cμ **(b)** VH5.1-Cμ **(c)** VH8.1-Cμ

Figure 5.10 Estimated density of the distribution of the entropy H of the powerset vector (Def. 5) for IgT. Left: VH4.1-Cτ, middle: VH5.4-Cτ, right: VH9.2-Cτ. Blue: condition C, Red, condition E1, green: condition E4. Ticks at the bottom show the maximum entropy for combinations of at least k fish, *i.e.* frequencies of occurence of (0.5, 0.5) for $k = 2$, (1/3, 1/3, 1/3) for $k = 3$ and so on.



(a) VH4.1-Cτ **(b)** VH5.4-Cτ **(c)** VH9.2-Cτ

Figure 5.11 Entropy H of the powerset vector (Defs. 4 and 5) versus z, the number of fish combinations in which a top clonotype has been found. The solid line shows the maximum entropy for a given z which is $-\ln(1/z)$.



5.4 Discussion

With the advent of high throughput sequencing in immunology, rigorous and standardized methods become necessary to undertake the analysis of resulting data. We take a step in this direction focusing on specific questions about B cell response to a viral challenge. Namely, the previous work [CJP+13] based on 454 pyrosequencing allowed the identification of VC pairs engaging in public and private responses. The Illumina technology used for this study gives a sequencing depth which is one to two orders of magnitude larger. This opens new questions as we are no longer being restricted to data concerning the largest clonotypes, and requires new methods to answer these questions.

We therefore present a description of the B cell response to VHSV in a fish vaccination model by analyzing the evolution of the VH CDR3 repertoire upon vaccination and challenge. We take advantage of the wealth of information provided by Illumina Ig RNA sequencing data to quantify clonotypes size and diversity.

This information is then used to characterize the structure, (*i.e.* the heterogeneity from fish to fish) of different types of responses (public, intermediate and private) in which VC pairs engage. For this, we use the earth mover distance (EMD) as a mean to compare the repertoires of two fish. The point this approach is that, both sequence similarity and clonotypes size naturally fit into the EMD framework. The resulting EMD distance is global in the sense that a single value quantifies differences between whole repertoires, but remains sensitive to the similarity between sequences when doing so. Moreover, we show that it describes responses in finer details than the Morisita-Horn distance which is often used to assess repertoire divergence.

We also quantify the extent to which large clonotypes are shared between naive, vaccinated and re-infected fish. Searching for both identical sequences and sequences from the same similarity class provides us with two levels of details. This allows us to identify four distinct repertoire behaviors upon vaccination and re-infection among VC pairs. First when all naive fish have distinct TCs and evolve diverging responses upon challenge, indicating that a VC pair engages in a private response; second when naive fish share many TCs and evolve converging responses upon challenge, indicating that a VC pair engages in a public response; third when they also share many TCs but evolve diverging responses upon challenge, *i.e.* the corresponding VC pair engages in a private response; Finally, when naive fish share many TCs and this does not change upon challenge because of the weak response of the corresponding

VC pairs. Therefore, the fact control fish share many TC does not imply that the response of this VC combination is public.

Finally, we quantify how random subsampling affects the representation of small clonotypes in the resulting subsamples and show that multiple repetitions are necessary when looking for clonotype occurrences in fish.

Although our analyses are informative at the sequence level, transposing our conclusions to the structural and functional realm remains a challenge. In particular, small changes in the amino-acid sequence can sometimes have strong repercussions on the overall function of a protein. This is even more pronounced for molecular recognition, as surface complementarity both in the geometric and physico-chemical sense plays a key role. Further works should therefore investigate how sequence similarity of CDR3 reflects in their ability to bind and neutralize pathogens. This would allow a better description of the response at the functional level.

Chapter 6

Software

6.1 Introduction

The various programs written during our work on binding affinity prediction were deemed to be of general interest. For this reason, we provide them as a specific-purpose package of the structural bioinformatics library (SBL: `http://sbl.inria.fr/`). Succinctly, this packages allows one to build and select models of binding affinity prediction using either quantities computed using executables from the SBL or user-defined ones. These can then be fed to the statistical machinery in charge =of evaluating and selecting significantly best models which can then be used to predict new data. This package, written in python, provides classes to be used in applications, along with scripts to quickly run analyses. This will allow both users with limited knowledge and advanced users to build and select models for binding affinity.

6.2 Goals: Generating and evaluating predictive models for binding affinity

This package provides tools to (i) select binding affinity prediction models from atomic coordinates, typically crystal structures of the partners and/or of the complex, and to (ii) exploit such models to perform affinity predictions.

At the selection stage, given a set of predefined variables coding structural and biochemical properties of the partners and of the complex, the approach consists of building sparse regression models which are subsequently evaluated and ranked using repeated cross-validation [MBC15] .

At the exploitation stage, the variables selected are used to build a predictive model on a training dataset, and predict the affinity on a test dataset.

More precisely, the approach runs through 3 steps:

- **Pre-processing**: a set of complexes is pre-processed, so as to compute the variables used in the model selection and/or exploitation.
- **Model selection**: sparse regressors using subsets of the previous variables are built, and the best ones selected.
- **Model exploitation**: given a set of variables selected from a given dataset, affinity predictions on another dataset can be performed.

6.3 General pre-requisites

6.3.1 Binding affinity and dissociation free energy

Consider two species A and B forming a complex AB. The dissociation is defined by $K_d = [A][B]/[AB]$, namely the ratio between the concentration of individual partners and bound partners.

The strength of this association is quantified by the dissociation free energy ΔG_d, which, in the $c^\circ = 1M$ standard state, satisfies with R Boltzmann's constant and T the temperature:

$$\Delta G_d = -RT \ln K_d/c^\circ = \Delta H - T\Delta S \tag{6.1}$$

This equation shows that ΔG_d has two components respectively coding the enthalpic (ΔH), and entropic (ΔS) changes upon binding. As stated above, we wish to estimate these quantities from atomic coordinates, typically crystal structures of the partners and/or the complex.

6.3.2 Key geometric constructions

Parameters The parameters used in this package are based on the following key geometric constructions:

- **Solvent accessible surface and its area**. The solvent accessible surface of a molecular model is the boundary of its constituting balls. The associated surface area (SASA for short) is the sum of the surface areas exposed by the individual atoms. Upon complex formation, the *buried surface area* (BSA) is the surface area of the partners buried at the interface, namely the SASA lost by the individual atoms. The solvent accessible surface areas are computed using `sbl-vorlume-pdb.exe`.

- **Buried surface and binding patches**. Upon formation of a complex, the BSA is the surface area of the partners which gets buried. Also, the exposed surface of the atoms contributing to the BSA define a *binding patch* (patch for short) [BGNC09] . The BSA and binding patches are computed using `sbl-vorshell-bp-ABW-atomic.exe`.

- **Atomic packing**. Consider the Voronoi (power) diagram of an atomic model, using its solvent accessible representation. The *restriction* of an atom as the intersection between its ball in the solvent accessible model and its cell in the Voronoi diagram. We denote volume_bound(a) (resp. volume_unbound(a)) the volume of the Voronoi restriction of an atom a in the bound form (resp. unbound form). The volumes are computed using `sbl-vorlume-pdb.exe`.

- **Shelling order (SO).** The shelling order of an atom from a binding patch is its least distance, counted in integer steps, to the nearest atom from the non-interacting surface (NIS, atoms which are not located at the interface). That is, the atoms on the border of the patch have a SO of 1 and the remaining ones have a $SO > 1$ Thus, the SO generalizes core-rim models [JBC08] , since the rim corresponds to $SO = 1$, and the core to $SO > 1$. The shelling order of interface atoms is computed using `sbl-vorshell-bp-ABW-atomic.exe`.

Using the previous notions, we define four categories of atoms of a complex:

- For interface atoms, denoted $\mathcal{I}$, we define two groups, those found on the rim ($\mathcal{I}, SO = 1$), retaining solvent accessibility, and the remaining ones ($\mathcal{I}, SO > 1$).

- For the set of non interface atoms, denoted $\mathcal{I}^C$, we distinguish between those retaining solvent accessibility ($\mathcal{I}^C$ and SASA > 0 in the complex), and those which do not ($\mathcal{I}^C$ and SASA $= 0$ in the complex).

Biophysical rationale. Before defining our parameters, we raise several simple observations underlying their design:

- **Generalizing the BSA.** The BSA is known to exhibit remarkable correlations with various biophysical quantities of protein complexes [JBC08] . However, it does not account for the interface geometry, as the same surface area may be observed for morphologies as diverse as a perfectly isotropic patch, or a long and skinny patch, letting alone curvature. The obliviousness to interface morphology is intuitively detrimental, since morphology relates to the cooperativity of phenomena inherent to non-bonded interactions. As we shall see below, we define a parameter generalizing the BSA by taking into account the SO of interface atoms and their packing properties.

- **Packing.** A closely packed environment yields favorable interactions by increasing the number of neighbors. But it also entails an entropic penalty for that atom, illustrating the classical enthalpy - entropy compensation, which holds in particular for biological systems involving weak interactions [Dun95] [CM13] . We therefore use atomic volumes and their variations upon binding (Eq. (6.3.3)) to model both the interaction energy and the entropic changes upon binding.

- **Entropy.** Assessing entropic variations requires taking several components into account, in particular configurational entropy and vibrational entropy. Large conformational changes yielding structured elements correspond to entropic penalties, and can be assessed using the root mean square deviation of interface atoms (iRMSD). In the sequel, we refine this measure using atomic packing properties. Indeed, packing properties are intuitively related to the vibrational entropy of atoms.

6.3.3 Associated variables

Using the quantities defined in section 6.3, we define the following variables:

Interface terms.

- IVW-IPL, see Eq. (6.3.3): the sum over interface atoms of their shelling order normalized by their packing. Note that since a packed interface is more likely to result in a high affinity, the shelling order is weighted by the inverse of the volume, yielding the *inverse volume-weighted internal path length* .

- iRMSD = Interface RMSD : the least RMSD for interface atoms.

Packing terms. Consider the volume variation of Δ-vol(a), see Eq. (6.3.3). As recalled above, packing properties are important in several respects. Adding up volume variations yields the following four *sum of volumes differences (VD)* parameters:

- SVD_SO1 ($\mathcal{I}, SO(a) = 1$; Eq. (6.3.3)): sum of VD for the rim interface atoms.

- SVD_SOGT1 ($\mathcal{I}, SO(a) > 1$; Eq. (6.3.3)): sum of VD for interface atoms in the interface core.

- SVD_NI_B ($\mathcal{I}^C, \text{SASA}(a) = 0$; Eq. (6.3.3)): sum of VD for buried non interface atoms.

- SVD_NI_E ($\mathcal{I}^C, \text{SASA}(a) > 0$; Eq. (6.3.3)): sum of VD for exposed non interface atoms.

Solvent interactions and electrostatics. The interaction between a protein molecule and water molecules is complex. In particular, the exposition to the solvent of non-polar groups hinders the ability of water molecules to engage into hydrogen bonding, yielding an entropic loss for such water molecules. We define the following terms:

- NIS^{polar}, see Eq. (6.3.3): the fraction of charged a.a. on the non interacting surface, as defined in [KRF$^+$14] .

- $\text{NIS}^{charged}$, see Eq. (6.3.3): the fraction of polar a.a. on the non interacting surface [KRF$^+$14] .

- POLAR_SASA, see (Eq. (6.3.3): an intermediate-grained description of the non-interacting surface which consists in the atomic-wise polar area of the complex. The corresponding term, POLAR_SASA (Eq. (6.3.3)), is a weighed sum of exposed areas.

- ΔNIS^{polar}, see Eq. (6.3.3): variation of NIS^{polar}, to account for conformational changes upon binding,

- $\Delta\text{NIS}^{charged}$, see Eq. (6.3.3): variation of $\text{NIS}^{charged}$, to account for conformational changes upon binding,

- ATOM_SOLV, see Eq. (6.3.3): To challenge amino-acid terms with their atomic counterparts and see which ones are best suited to perform affinity predictions, we also included the atomic solvation energy from Eisenberg et al [EWY89] , describing the free energies of transfer from 1-octanol to water per surface unit (Å ^{2}). The corresponding variable, ATOM_SOLV, is a weighted sum of atomic solvent accessible surface areas (Eq. (6.3.3)), and may be seen as the atomic-scale counterparts of NIS^{charged} and NIS^{polar}.

Parameters used to estimate binding affinities. There are three groups of parameters (see the next equations for their definition):

- atomic level parameters : IVW-IPL, SVD_SO1, SVD_SOGT1, SVD_NI_B, SVD_NI_E, ATOM_SOLV and POLAR_SASA.
- residue level parameters : NIS^{polar}, NIS^{charged}, ΔNIS^{polar} and ΔNIS^{charged}.
- interface level parameter : iRMSD.

The acronyms read as follows (see text for details):

- **S**um of **V**olume **D**ifferences;
- **S**helling **O**rder;
- **I**nverse **V**olume **W**eighted;
- **I**nternal **P**ath **L**ength;
- **N**on **I**nteracting **B**uried/**E**xposed;
- **N**on **I**nteracting **S**urface;
- **S**olvent **A**ccessible **S**urface **A**rea;

$$\Delta\text{-vol}(a) = \text{volume\_bound}(a) - \text{volume\_unbound}(a)$$

$$\text{IVW-IPL} = \sum_{a\in\mathcal{I}} \frac{\text{SO}(a)}{\text{volume\_bound}(a)}$$

$$\text{SVD\_SO1} = \sum_{a\in\mathcal{I}, SO(a)=1} \Delta\text{-vol}(a)$$

$$\text{SVD\_SOGT1} = \sum_{a\in\mathcal{I}, SO(a)>1} \Delta\text{-vol}(a)$$

$$\text{SVD\_NI\_B} = \sum_{a\in\mathcal{I}^C, \text{SASA}(a)=0} \Delta\text{-vol}(a)$$

$$\text{SVD\_NI\_E} = \sum_{a\in\mathcal{I}^C, \text{SASA}(a)>0} \Delta\text{-vol}(a)$$

$$\text{NIS}^{polar} = \frac{\#\text{solvent accessible polar residues}}{\#\text{solvent accessible residues}}$$

$$\text{NIS}^{charged} = \frac{\#\text{solvent accessible charged residues}}{\#\text{solvent accessible residues}}$$

$$\Delta\text{NIS}^{polar} = \text{NIS}^{polar}_{bound} - \text{NIS}^{polar}_{unbound}$$

$$\Delta\text{NIS}^{charged} = \text{NIS}^{charged}_{bound} - \text{NIS}^{charged}_{unbound}$$

$$\text{ATOM\_SOLV} = \sum_{a\in\mathcal{I}^C} SASA(a)\cdot\sigma(a)$$

$$\text{POLAR\_SASA} = \sum_{a\in\mathcal{I}^C \text{ and } \sigma(a)<0} SASA(a)$$

$$\text{iRMSD} = \text{Interface RMSD}$$

6.3.4 Atoms' matching

For variables describing a change between the bound and unbound forms of the partners *e.g.* SVD_SO1 or $\Delta\text{NIS}^{charged}$, it is necessary to match the atoms of the bound partners to those of the unbound partners. For this, the amino-acid sequences of the partners are extracted from the PDB files and aligned. Then, the atoms of every pair of corresponding residues are matched using their name.

Since missing residues and missing atoms are common in crystal structures, the proportion of matched atoms is computed as to investigate potential wrong matchings, and discard entries for which the alignment is not good enough.

6.4 Using the programs to pre-process structural data

In this section we explain how to compute the variables presented in section 6.3.3.

This step consists of two sub-steps:

- `sbl-bap-step-1-run-applications.py`: Variables describing geometrical and physico-chemical are computed, using various constructions provided in the SBL. Two settings are supported:
  - Crystal structures of the complex and the unbound partners are given,
  - Only the crystal structure of the complex is given.
- `sbl-bap-step-2-compile-molecular-data.py`: Information from the individual runs is assembled. The result may be seen as a matrix of complexes, each defined by its variables.

In section 6.4.1, we explain how other variables can be used.

6.4.1 Input: specifications and file types

The input of the pre-processing step consists in:

- An XML file complying with the following format:
  - each entry should be contained in an <`entry`> element
  - the PDB ID of the complex (tag <`complex_pdb`>)
  - optionally (flag `-b` not used) the PDB IDs of the unbound partners (tags <`UnboundPDBA`> and <`UnboundPDBA`>),
  - the chains of the complex (tags <`chainsA`>, <`chainsB`>)
  - optionally (flag `-b` not used) the chains of the unbound partners (tags <`chainsUA`> and <`chainsUB`>)
  - the experimental ΔG (tag <`dG`>)
  - optionally the iRMSD (tag <`I-RMSD`>)
- A path to the directory containing the corresponding PDB files.
- A file containing atomic solvation parameters

Step 1. The first step of the workflow is performed by the script `sbl-bap-step-1-run-applications.py` which is called as follows:

```
sbl-bap-step-1-run-applications.py -i data/example_data.xml -s data/eisenberg_solvation_parameters.tx
```

The main options of the program `sbl-bap-step-1-run-applications.py` are:

- `(-i, --ids-and-chains-file-path) string`: input XML file containing the PDB ids and chains of the complexes and unbound partners (required)

Table 6.1 Input files for the first step described in section 6.4.1.

File Name	Description
eisenberg_solvation_parameters.txt	Eisenberg Solvation Parameters
example_data.xml	PDB ids and chains of complexes and unbound partners
1A19.pdb	Example PDB file in the input directory of option `-p`

- `(-p, --pdb-dir-path) string`: directory containing the PDB files to be analyzed (required)
- `(-n, --nb-process) int`: number of processes on which to dispatch the execution
- `(-b, --bound-only) bool(=False)`: toggle computations on the bound structures (complexes) only *i.e.* not on the unbound partners
- `(-c, --contacts-area) bool(=False)`: toggle the computation of the area of the Voronoi facets between atoms and residues (uses sbl-vorlume-pdb.exe)
- `(-x, --executables-path) string(=$SBL_BIN_DIR)`: path to the directory containing the executables.
- `(-v, --verbose) bool(= False)`: toggle verbose output

Step 2. The second step of the workflow is performed by the script `sbl-bap-step-2-compile-molecular-data.py` which is called as follows:

```
sbl-bap-step-2-compile-molecular-data.py -f data/example_data.xml -d results -o results/affinity_datase
```

The main options of the program `sbl-bap-step-2-compile-molecular-data.py` are:

- `(-f, sab-file-path) string`: path to the XML file containing the affinity (and optionally iRMSD) data for the entries (required)
- `(-o, --output-file-name) string`: name of the output file
- `(-b, --bound-only): bool(=False)`: if set, the program does not try to fetch data about the unbound partners (to be used in conjunction with the -b option from step 1)
- `(-d, --data-directory) string`: path to the directory containing the data (defaults to the current directory)
- `(-v, --verbose) bool(= False)`: toggle verbose output

6.4.2 Output: specifications and file types

Step 1. Practically, this first step consists in running `sbl-match-PDB-residues-and-atoms.exe` , `sbl-vorlume-pdb.exe` and `sbl-vorshell-bp-ABW-atomic.exe` on various entries of the dataset. For each entry, a directory named after the following scheme is created in the current working directory: `<PDB_ID>_<chains_A>_<chains_B>`

Each directory is structured as follows:

```
1A2K_AB_C
|-- 1OUN_A_A_atom_matchings.txt
|-- 1OUN_A_A_residue_matchings.txt
|--  1OUN_AB
| |-- sbl-vorlume_eisenberg_solvation_parameters__log.txt
| |-- sbl-vorlume_eisenberg_solvation_parameters__surface_volumes.xml
|-- 1OUN_B_A_atom_matchings.txt
|-- 1OUN_B_A_residue_matchings.txt
```

Table 6.2 Output files for the step 2 described in section 6.4.1

File Name	Description
affinity_dataset.xml	Compiled results of runs of step 1

```
|-- 1QG4_A
| |-- sbl-vorlume_eisenberg_solvation_parameters__log.txt
| |--  sbl-vorlume_eisenberg_solvation_parameters__surface_volumes.xml
|-- 1QG4_C_A_atom_matchings.txt
|-- 1QG4_C_A_residue_matchings.txt
|-- sbl-vorlume_eisenberg_solvation_parameters__log.txt
|-- sbl-vorlume_eisenberg_solvation_parameters__surface_volumes.xml
|-- sbl-vorshell-bp-ABW-atomic__AB_ball_shelling_forest.dot
|-- sbl-vorshell-bp-ABW-atomic__AB_ball_shelling_forest.xml
|-- sbl-vorshell-bp-ABW-atomic__AB_packing.xml
|-- sbl-vorshell-bp-ABW-atomic__BA_ball_shelling_forest.dot
|-- sbl-vorshell-bp-ABW-atomic__BA_ball_shelling_forest.xml
|-- sbl-vorshell-bp-ABW-atomic__BA_packing.xml
|-- sbl-vorshell-bp-ABW-atomic__interface_AB.pdb
|-- sbl-vorshell-bp-ABW-atomic__interface_BA.pdb
|-- sbl-vorshell-bp-ABW-atomic__log.txt
|-- sbl-vorshell-bp-ABW-atomic__surface_volumes.xml
```

This example is for entry 1A2K with partners AB and C. All `*_matchings.txt` files contain the matchings between atoms or residues of the bound and unbound structures. Directories `1OUN_AB` and `1QG4_A` are for the unbound partners. All `sbl-*` files contain the results of the SBL executables. If verbose output is toggled (`-v` flag), various information will be output on the standard and error output.

Step 2.

6.5 Using the programs to select affinity prediction models

In this section, we explain how to select a binding affinity model maximizing a performance criterion, using the script `sbl-bap-step-3-models-analysis.py`. Two points worth noticing are:

- **Performance criterion:** the default criterion used is the correlation between predictions and experimental values. But any other criterion of the kind such as the median absolute error, root mean squared error... is eligible.
- **Variables used:** the default variables used are those introduced in section 6.3.2. But variables stemming from different analysis can be incorporated to the model section, provided that they comply with the input format specified in section 6.5.2. In both cases, we plainly refer to the *pool of variables* in the sequel.

6.5.1 Pre-requisites : Statistical Methods

In the sequel, we explain how to predict $\Delta G_d^{exp_i}$ of complexes from a dataset $\mathcal{D}$. Estimation is performed on a per complex basis, from which performances at the whole dataset level are derived.

Our predictions rely on three related concepts defined precisely hereafter.

- Template: a fixed set of variables from $\mathcal{V}$,
- Model: a regression model consisting in a template plus the associated coefficients. As we shall see, such models are associated with cross-validation folds.
- Predictive model for $\mathcal{D}$: the machinery returning one binding affinity estimate $\hat{g}_i$ per complex from $\mathcal{D}$, using N_{XV} repetitions of the k-fold cross validation.

Templates. Denote $\mathcal{V}$ the pool of variables. Let a *template* be a set of variables, *i.e.* a subset of $\mathcal{V}$. To define parsimonious templates from the set $\mathcal{V}$, we generate subsets of $\mathcal{V}$ involving up to at most M variables. This defines a pool of templates $\mathcal{T} = T_1, \ldots, T_N$.

Cross-validation. In the following a *model* is associated to both a template $T_l \in \mathcal{T}$ and a dataset $\mathcal{D}$. More precisely, a model refers to a regression model, *i.e.* the variables of the template plus the associated coefficients.

Practically, models are defined during k-fold cross-validation, and a number of N_{XV} of repetitions. Consider one repetition, which thus consists of splitting at random $\mathcal{D}$ into M subsets called folds. For one fold, a regression model associated with T_l is trained on (k-1)/k of the dataset $\mathcal{D}$, and predictions are run on the remaining 1/k of complexes. Processing the M folds yields one repetition of the cross validation procedure, resulting in one prediction $\hat{g}_{ij}$ for the $\Delta G_d^{exp_i}$ of each complex. The set of all predictions in one repeat, say the jth one, is denoted

$$\hat{G}_j = \{\hat{g}_{ij}\}_{i=1,\ldots,|\mathcal{D}|}. \tag{6.2}$$

Note again that these predictions stem from k regression models associated with T_l, namely one per fold.

Statistics per template. Considering one cross-validation repetition j, we define the correlation $Corr_j$ as the correlation between the experimental values $\Delta G_d^{exp_i}$ and the predictions $\hat{G}_j$. An overall assessment of the template T_l using the N_{XV} repetitions is obtained by the following *median of correlations*

$$C[T_l, \mathcal{D}] = \text{median}_j \; Corr_j. \tag{6.3}$$

For a complex i, we define the binding affinity *prediction* $\hat{g}_i$ as the median across repetitions *i.e.*

$$\hat{g}_i = \text{median}_j \; \hat{g}_{ij}. \tag{6.4}$$

The *median prediction error* is defined by

$$e_i \equiv e_i[T_l, \mathcal{D}] = \text{median}_j(\Delta G_d^{exp_i} - \hat{g}_{ij}) \tag{6.5}$$

and the *median absolute prediction error* by:

$$e_i^{\text{abs}} \equiv e_i^{\text{abs}}[T_l, \mathcal{D}] = \text{median}_j(|\Delta G_d^{exp_i} - \hat{g}_{ij}|). \tag{6.6}$$

Using this latter value, we define the *prediction ratio* p_δ^{error} as the percentage of cases such that the dissociation free energy is off by a specified amount δ:

$$p_\delta^{\text{error}} = \%\text{ cases in } \mathcal{D} \text{ such that } e_i^{\text{abs}}[T_l, \mathcal{D}] \leq \delta. \tag{6.7}$$

In particular, setting δ to 1.4, 2.8 and 4.2 kcal/mol in the previous equation yields cases whose K_d is approximated within one, two and three orders of magnitude respectively.

Finally, a permutation test yields a p-value for each predictive model [PS10] .

In a nutshell, the rationale consists of generating randomized datasets by shuffling their $\Delta G_d^{exp_i}$ values. Then, one computes the performance criterion for each such dataset, from which the p-value is inferred (see Algorithm∼6.6).

Model selection. Define the best predictive model as the one maximizing performance criterion. We wish to single out the best predictive models, *i.e.* those that cannot be statistically distinguished from the best predictive model. More precisely, consider a pool of predictive models $\mathcal{T}$, out of which we wish to identify a subset of predictive models whose distribution cannot be distinguished from that of the best predictive model. To this end, let H0 be the null hypothesis stating that two predictive models yield identical performance distributions. We decompose the predictive models as $\mathcal{T} = \mathcal{T}_1 \cup \mathcal{T}_2$ such that :

- (i) the best predictive model is in $\mathcal{T}_1$,
- (ii) in comparing two predictive models from $\mathcal{T}_1$, one does not reject H0, and

- (iii) in comparing one predictive model from $\mathcal{T}_1$ against one predictive model from $\mathcal{T}_2$, one rejects H0.

The predictive models in $\mathcal{T}_1$ are called the *specific models* for the dataset $\mathcal{D}$. The corresponding procedure is based on the Kruskal-Wallis test, see [MBC15] .

The p-value threshold for this test is set to $\alpha = 0.01$.

6.5.2 Input: Specifications and File Types

Model generation and selection is performed by the script `sbl-bap-step-3-models-analysis.py` which is executed as follows :

```
# Quick example: maximum 2-variables models, 10 repetitions and p-value permutations,
# flexible complexes only (irmsd > 1.5 A), run on a single process
 sbl-bap-step-3-models-analysis.py -f affinity_data.txt -m 2 -n 1 -r
 10 -p 10 -l 1.5 -o flex_quick

# Heavier example: maximum 4-variables models, 100 repetitions and
# 1000 p-value permutations, all complexes, run on three
# processes. Performs k-nearest neighbors regression with 10 neighbors
# sbl-bap-step-3-models-analysis.py -f affinity_data.txt -m 4 -n 3 -r
# 100 -p 1000 -k 10 -o all_heavier
```

The main options of the program `sbl-bap-step-3-models-analysis.py` are:

- `(-f, --data-file-path) string`: path to the file resulting from running `sbl-bap-step-2-compile-molecular-da`
- `(-m, --max-order) int`: maximum order (*i.e.* number of variables) to be included in a model (required)
- `(-a, --matching-atoms-filter) float(= 0.8)`: minimum proportion of atoms matched from a complex structure to its unbound partners structures required for an entry to be included
- `(-n, -nb-process) int(= 1)`: number of processes on which to dispatch the execution
- `(-d, --nb-folds) int(= 5)`: number of folds of the cross-validation
- `(-r, --nb-repeats) int(= 1000)`: number of repetitions of the cross-validation procedure
- `(-p, --nb-pval-permutations) int(= 1000)`: number of permutations used to compute the p-value of a model
- `(-u, --irmsd-upper-bound) float`: maximum iRMSD allowed for an entry to be included
- `(-l, --irmsd-lower-bound) float`: minimum iRMSD allowed for an entry to be included
- `(-i, --inclusive-bounds) bool(= False)`: set if the previous bounds specified by the -i and -l flags should be inclusive instead of strict
- `(-o, --output-file-prefix) string`: prefix for the output file
- `(-k, --knn) bool(= False)`: toggle knn regression instead of linear regression with the given number of neighbors

The XML file specified by option `-f` should comply with the following format:

- Each entry should be contained in an <*entry*> element
- Each entry should contain the same elements
- Each element within an entry should have an attribute *type* set to one of the following values:

Table 6.3 Input files for the third step described in section 6.5.2.

File Name	Description
affinity_dataset.xml	Compiled data generated from `sbl-bap-step-2-compile-molecular-data.py`

Table 6.4 Output files for the step 3 described in section 6.5.2

File Name	Description
example_correlation_based_results.txt	File 1
example_error_based_results.txt	File 2
example_irmsd_vs_error.pdf	Scatter plot of the prediction error

- "info": for general information (*e.g.* PDB ID, chains, ...) which will not be used during the model selection
- "dep_var": for the dependent variable, *i.e.* the one to be predicted
- "feature": for the variables to be used during the regression

6.5.3 Output: Specifications and File Types

This step outputs three files:

- Files 1 and 2 are text files with the following structure:
  - A line giving the (0-based) index of the last significantly best model. All models ranked higher are statistically equally good, and all models ranked lower are statistically worse.
  - A table with models as rows and various metrics associated (error, correlation, p-values, ...)

  In the first one, the index and the ranks are computed on the cross-validated median absolute error of the model. In the second they are computed on the median cross-validation correlation coefficient.
- File 3: an optional PDF file output only when the iRMSD is provided. The file contains a scatter plot of the prediction error (Eq. 6.5) made by the best predictive model (correlation-wise) for each complex against its iRMSD.

6.5.4 Model exploitation: predicting affinities

Once a set of variable has been selected during the previous step, it can be used to estimate a model. Recall that a model is a function returning an estimate for binding affinity of a complex. For this, one needs a training set and a regressor, *i.e.* an object able to use the training set to predict the affinity of a new dataset (for instance, linear least-squares fitting, k nearest neighbors, regression trees, ...). An example of the procedure to train such a regressor on the whole Structure Affinity Benchmark and predict its own entries (that is to perform predictions without cross-validation) is given in a demo file (`sbl-binding-affinity-model-exploitation-example.py`).

6.6 Algorithms and Methods

We now describe two algorithms used by this package. The first simply computes an approximate (upper bound on the) permutation p-value for a given statistic [PS10] . The pseudo-code is given on Algorithm 6.6.

The second algorithm is used to divide a pool of models into two groups, namely $\mathcal{T} = \mathcal{T}_1 \cup \mathcal{T}_2$, as discussed in the section 6.5. The corresponding procedure is based on the Kruskal-Wallis test, and can be found in [MBC15] .

Algorithm 3 Computing a permutation p-value for a binding affinity predictive model specified by a template T_l. The p-value is based on a permutation test, which uses the prediction performances obtained on random datasets, each such dataset being obtained by permuting the dependent variable (*i.e.* the affinity) over the dataset.

Require: $\mathcal{D}$: dataset; T_l: a template; p_{T_l}: a performance criterion for Tx_l; $N_{\text{perm.}}$: number of repetitions

for $q \in \{1 \dots N_{\text{perm.}}\}$ **do**
  Randomly permute the dependent variable in $\mathcal{D}$ (here the affinity) to obtain $\mathcal{D}_q^{perm}$
  Perform 5-fold cross-validation of regression models using the variables in T_l on $\mathcal{D}_q^{perm}$
  Store the performance criterion in $p_{T_l}^{perm}$
Report the approximate p-value for T_l to be $\frac{B+1}{N_{\text{perm.}}+1}$, with B the number of elements in $p_{T_l}^{perm}$ which are more extreme than p_{T_l}.

6.7 Programmer's Workflow

6.7.1 Pre-processing

We describe successively the 2 steps of section 6.4.

Step 1. The script `sbl-bap-step-1-run-applications.py` runs applications from the SBL on the bound and unbound partners.

The matching between atoms (section 6.3.4) is done using the python module SBL::PDB_complex_to_unbound_partners_ which is a wrapper around the executable `sbl-match-PDB-residues-and-atoms.exe` . This executable performs the matching using the class `SBL::CSB::T_Alignment_engine_sequences_seqan`, from package `Alignment_engines`, which itself uses the `Seqan` library to perform the sequence alignment, and subsequently match residue and atoms based on the alignment found.

In order to compute shelling orders and packing properties, the executables `sbl-vorlume-pdb.exe` and `sbl-vorshell-bp-ABW-atomic.exe` are also run on the complex, and `sbl-vorlume-pdb.exe` is run on the unbound partners.

Step 2. The script `sbl-bap-step-2-compile-molecular-data.py` compiles the previous results and computes the variables listed in section 6.3.3 for each entry of the dataset.

Data structures to store the properties of entities such as atoms, residues, chains and files are grouped in the python module SBL::Protein_complex_analysis_molecular_data_structures . To fill these data structures it also provides a class used to traverse the directory structure and parse the files resulting from step 1.

6.7.2 Model selection

The machinery used to build, evaluate and rank various models using the previously computed variables can be found in the python module `SBL::Combinations_of_variables_model_evaluation::Combinations_of_variables`

6.7.3 Model exploitation

Once a set of variables has been selected, model fitting on a training set and prediction of a new dataset can be done using any software or language. An example script is given using python and scikit-learn in file `sbl-bap-step-3-models-analysis.py`

6.8 External dependencies

The seqan library [DWRR08] (BSD, `Seqan`) is necessary when computing the variables which encode the variations between bound and unbound partners.

Chapter 7

Conclusion

7.1 Discussion

During the present thesis, we focused on various structural aspects of Ig - Ag complexes and the relationship with their function.

We first sought to understand the factors affecting the binding affinity between a protein and a ligand. The limited amount of curated data available for this task prompted us to analysis general protein-protein complexes instead of Ig - Ag complexes only. In particular the structure affinity benchmark was well-suited for this task.

The results we obtained, along with other published works, raise two comments. First, affinity prediction methods based on the SAB may have reached a plateau. In effect, it seems unlikely that models with few parameters could yield significantly better predictions than [MBC15] and [VB15] on this dataset since both get close to the theoretical limit. In particular, it is unrealistic to expect predictions with errors smaller than 1 kcal/mol without explicitly taking into account various experimental conditions such as pH and ionic strength. Even so, experimental errors both on coordinates of atoms and affinity measurement would make this task very difficult if not impossible. Increasing the complexity of the models is not a viable path either because of the risk of overfitting due to the restricted size of this dataset. An increase in the number and diversity of high-resolution structures of complexes, provided with binding affinity data for a range of compatible temperatures should foster the search for a unified model of the binding affinity along with a thermodynamic justification for this model.

Second, incorporating dynamics in binding affinity models seems now unavoidable as it is the only way to account for entropy-driven binding. On the way to modeling dynamics, a first step would be to compute and validate estimates of the energy of the unbound state to be included in the calculation of the free energy. The development of automatic and fast sampling strategies of energy landscapes and analysis thereof should be key in both cases.

We then moved on to a detailed study of Ig - Ag complexes, focusing on their interfaces. We sought to find a quantitative description of the parameters determining the affinity and specificity of Igs.

The very good prediction accuracy of the previous model of affinity obtained on Ig - Ag complexes when trained on other protein-protein complexes leads us to emit two hypotheses. Either the very small number of Ig - Ag complexes available for this analysis was similar to other protein-protein complexes by chance alone, either Ig - Ag share the same modes of interaction with general protein-protein complexes. In the second case, the design of ligand binding molecules could be made easier by taking advantage of the framework provided by Igs. An increase in the number of structures of Ig - Ag complexes should hopefully settle the question.

Considering ligand types specificity, the simple two descriptors we found show that the ligand type imposes unambiguous constraints on the binding site of the Ig. However, being able to find new, possibly related, descriptors when the structure of the Ig only is available would make this result more valuable.

Finally, we undertook the description and characterization of the B cell response of rainbow trout to viral infection from Ig RNA sequencing data.

We first introduced a new method for the global comparison of repertoires based on sequence alignments. Computing the earth mover distance (EMD) using sequence similarity information and clonotypes counts allowed us to characterize various types of responses among variable - constant genes (VC) pairs. In particular our results agreed with identified VC pairs contributing to public intermediate and private responses. We also showed that EMD reveals finer details than population diversity-based quantities, such as the Morisita-Horn distance, by taking into account sequence similarity.

In a second time, we quantified the degree to which large clonotypes are shared between fish from different conditions (naive, vaccinated, vaccinated + infected). We looked for large (top) clonotypes found in fish from a given condition in other fish at two levels of details. First by considering a clonotype to be found only if it was identical to at least one sequence in the fish, thereby quantifying exact overlaps. Second by allowing two conservative amino-acid substitutions in order to account for convergent evolution to the same antigen. This allowed us to identify four distinct repertoire behaviors upon vaccination and infection, related to public, private and absence of response.

Finally, we investigated how random subsampling could affect the representation of small clonotypes in the resulting subsample, both at a theoretical and at a practical level. In the first case, we derived bounds on the minimum size of a clonotype to be picked with a given probability. In the second, we quantified the variation with which a given clonotype was found in specific combinations of fish during multiple subsampling rounds .

7.2 Perspectives and future work

We see three main extensions to our work on binding affinity prediction. First, the design of new variables and application of our statistical methodology to protein - ligand complexes. Because of applications to drug discovery, the potential affinity data available should be much larger than for protein-protein complexes. In terms of structure, pose predictions software will hopefully be mature enough to allow the output to be directly used by our method (see also next paragraph). A quick experiment (participation in the D3R grand challenge) with current parameters, and using both experimental and pose-predicted structures, resulted in poor predictions, although this may be due to the values to be predicted – which were IC50 and Ki instead of K_d [GKE$^+$16]. Therefore, investigating how binding an inhibition relate with respect to structure and dynamics may be another interesting research direction.

Second, assessing the robustness of the method on homology and docking models. The main drawback of the current approach is that the training data consists only of medium to high affinity native complexes. The lack of negative examples, or decoys, (*i.e.* non-interacting proteins) implies that the predictor will not be able to correctly quantify the affinity of complexes with unnatural configurations. This is a mandatory step to bridge the gap between affinity prediction models and scoring functions, since the latter must be able to discriminate native from non-native complexes.

Third, extending the method to predict the contribution of individual residues ($\Delta\Delta G$) or sets of residues. In particular being able to predict hot spots, modules (or hot regions), and to correctly model the (absence of) cooperativity between them [RRA$^+$05, KMR$^+$05] makes for a very interesting challenge and could potentially yield novel approaches for improved affinity prediction.

The obvious extension to the analysis of Ig - Ag complexes, would be to consider the whole Ig instead of the interface only. In particular, being able to estimate the avidity of Igs for a multivalent Ag requires taking into account constant domains. Moreover, antibody flexibility may be required in order to access cluttered epitopes as it has been speculated to be the case for broadly neutralizing antibodies targeting the stem of hemagglutinin of the influenza virus. Modeling this flexibility would therefore be necessary to understand the binding mechanisms of such antibodies. The very limited amount of structures of whole Igs is, however, a limiting factor in both cases.

The analysis of repertoires we carried out is based on sequence data and is therefore limited to this aspect. Transposing our conclusions to the structural, dynamical and functional realms would thus be essential in order to better describe the immune response.

In particular, small changes in the amino-acid sequence can sometimes have strong repercussions on the overall function of a protein. In this respect developing a similarity measure between sequences while

taking structural aspects into account would be a big step forward. Although this task is part of the long-standing problem of *in silico* structure determination, focusing on a subset of short, well characterized, sequences such as VH CDR3 may be a start.

The dependency to small changes in the sequence is even more pronounced for molecular recognition, as surface complementarity between partners, both in the geometric and physico-chemical sense, plays a key role. This of course takes us to the field of docking and affinity prediction which are still not mature enough to perform accurate large scale predictions, especially using models instead of experimental data as would be needed when starting from sequencing data. Further works could however investigate how sequence similarity of CDR3 reflects in their ability to bind and neutralize pathogens.

Bibliography

[ALLC97] 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.

[Alm04] 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.

[AM08] KR Abhinandan and Andrew CR Martin. Analysis and improvements to kabat and structurally correct numbering of antibody variable domains. *Molecular immunology*, 45(14):3832–3839, 2008.

[BBB+94] 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.

[BBHLE12] Jennifer Benichou, Rotem Ben-Hamo, Yoram Louzoun, and Sol Efroni. Rep-seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology*, 135(3):183–191, 2012.

[BBO+83] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, and M. Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4:187–217, 1983.

[BCRJ04a] 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.

[BCRJ04b] Ranjit Prasad Bahadur, Pinak Chakrabarti, Francis Rodier, and Joël Janin. A dissection of specific and non-specific protein–protein interfaces. *Journal of molecular biology*, 336(4):943–955, 2004.

[BD15] G. Biau and L. Devroye. *Lectures on the nearest neighbor method*. Springer, 2015.

[BGNC09] 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.

[BHE11] Rotem Ben-Hamo and Sol Efroni. The whole-organism heavy chain b cell repertoire from zebrafish self-organizes into distinct network features. *BMC systems biology*, 5(1):1, 2011.

[Blo02] D. Blow. *Outline of crystallography for biologists*. Oxford University Press, 2002.

[BMFDP+08] Pierre Boudinot, Maria Encarnita Marriotti-Ferrandiz, Louis Du Pasquier, Abdenour Benmansour, Pierre-André Cazenave, and Adrien Six. New perspectives for large-scale repertoire analysis of immune receptors. *Molecular immunology*, 45(9):2437–2445, 2008.

[BN98] 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.

[Bon10] L. Bonetta. Protein-protein interactions: Interactome under construction. *Nature*, 468(7325):851–854, 2010.

[BY98] J.-D. Boissonnat and M. Yvinec. *Algorithmic geometry*. Cambridge University Press, UK, 1998. Translated by H. Brönnimann.

[BZF+08] S. Birtalan, Y. Zhang, F.A. Fellouse, L. Shao, G. Schaefer, and S.S. Sidhu. The intrinsic contributions of tyrosine, serine, glycine and arginine to the affinity and specificity of antibodies. *Journal of molecular biology*, 377(5):1518–1528, 2008.

[Caz10] 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.

[CBM03] 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.

[CCB+95] W. Cornell, P. Cieplak, C. Bayly, I. Gould, K. Merz, D. Ferguson, D. Spellmeyer, T. Fox, J. Caldwell, and P. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, 1995.

[CD11] Y. Choi and C.M. Deane. Predicting antibody complementarity determining region structures without classification. *Molecular Biosystems*, 7(12):3327–3334, 2011.

[CDM+15] F. Cazals, T. Dreyfus, D. Mazauric, A. Roth, and C.H. Robert. Conformational ensembles and sampled energy landscapes: Analysis and comparison. *J. of Computational Chemistry*, 36(16):1213–1231, 2015.

[CeCG07] C.A. Chia-en, W. Chen, and M.K. Gilson. Ligand configurational entropy and protein binding. *PNAS*, 104(5):1534–1539, 2007.

[CGR+13] Devlina Chakravarty, Mainak Guharoy, Charles H. Robert, Pinak Chakrabarti, and Joel Janin. Reassessing buried surface areas in protein-protein complexes. *Protein Sci*, 22:1453–57, Aug 2013.

[Cha03] T.C. Chalikian. Volumetric properties of proteins. *Annual review of biophysics and biomolecular structure*, 32(1):207–235, 2003.

[Chi14] C. Chipot. Frontiers in free-energy calculations of biological systems. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 4(1):71–89, 2014.

[CJ75] C. Chothia and J. Janin. Principles of protein-protein recognition. *Nature*, 256:705–708, 1975.

[CJ02] P. Chakrabarti and J. Janin. Dissecting protein-protein recognition sites. *Proteins*, 47(3):334–43, 2002.

[CJI68] James J Christensen, H Dee Johnston, and Reed M Izatt. An isothermal titration calorimeter. *Review of Scientific Instruments*, 39(9):1356–1359, 1968.

[CJP+13] R. Castro, L. Journeau, H.P. Pham, O. Bouchez, V. Giudicelli, M-P. Lefranc, E. Quillet, A. Benmansour, F. Cazals, A. Six, S. Fillatreau, O. Sunyer, and P. Boudinot. Teleost fish mount complex clonal IgM and IgT responses in spleen upon systemic viral infection. *PLOS Pathogens*, 9(1):e1003098, 2013.

[CKL11] F. Cazals, H. Kanhere, and S. Loriot. Computing the volume of union of balls: a certified algorithm. *ACM Transactions on Mathematical Software*, 38(1):1–20, 2011.

[CL87] C. Chothia and A.M. Lesk. Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Bio*, 196(4), 1987.

[CLT+89] 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.

[CM13] John D Chodera and David L Mobley. Entropy-enthalpy compensation: role and ramifications in biomolecular ligand recognition and design. *Biophysics*, 42:121–142, 2013.

[CMCT11] 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.

[CMT11] 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.

[CPBJ06] F. Cazals, F. Proust, R. Bahadur, and J. Janin. Revisiting the Voronoi description of protein-protein interfaces. *Protein Science*, 15(9):2082–2092, 2006.

[CRGG94] 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.

[Cru99] DWJ. Cruickshank. Remarks about protein structure precision. *Acta Crystallographica Section D: Biological Crystallography*, 55(3):583–601, 1999.

[CS09] R. Coico and G. Sunshine. *Immunology: a short course*. John Wiley & Sons, 2009.

[CSG+93] 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.

[CW95] T Clackson and JA Wells. A hot spot of binding energy in a hormone-receptor interface. *Science*, 267(5196):383–386, 1995.

[DGW+13] Lei Deng, Jihong Guan, Xiaoming Wei, Yuan Yi, Qiangfeng Cliff Zhang, and Shuigeng Zhou. Boosting prediction performance of protein–protein interaction hot spots by using structural neighborhood properties. *Journal of Computational Biology*, 20(11):878–891, 2013.

[Dun95] J. Dunitz. Win some, lose some: enthalpy-entropy compensation in weak intermolecular interactions. *Chemistry & biology*, 2(11):709–712, 1995.

[DWRR08] Andreas Döring, David Weese, Tobias Rausch, and Knut Reinert. Seqan an efficient, generic c++ library for sequence analysis. *BMC bioinformatics*, 9(1):11, 2008.

[ECG+69] Gerald M Edelman, Bruce A Cunningham, W Einar Gall, Paul D Gottlieb, Urs Rutishauser, and Myron J Waxdal. The covalent structure of an entire γg immunoglobulin molecule. *Proceedings of the National Academy of Sciences*, 63(1):78–85, 1969.

[EKL10] 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.

[ERS14] A. Erijman, E. Rosenthal, and J.M. Shifman. How structure defines affinity in protein-protein interaction. *PLOS one*, 9(10), 2014.

[EWY89] David Eisenberg, Morgan Wesson, and Mason Yamashita. Interpretation of protein folding and binding with atomic solvation parameters. *Chem. Scr. A*, 29:217–221, 1989.

[FBKS06] F.A. Fellouse, P.A. Barthelemy, R.F. Kelley, and S.S. Sidhu. Tyrosine plays a dominant functional role in the paratope of a synthetic antibody derived from a four amino acid code. *Journal of molecular biology*, 357(1):100–114, 2006.

[Fer00] Thomas S Ferguson. Linear programming: A concise introduction. *UCLA [online] http://www. math. ucla. edu/˜ tom/LP. pdf*, 2000.

[FFS04] C. Wiesmann F.A. Fellouse and S.S. Sidhu. Synthetic antibodies from a four-amino-acid code: a dominant role for tyrosine in antigen recognition. *PNAS*, 101(34):12467–12472, 2004.

[FLC+05] F.A. Fellouse, B. Li, D.M. Compaan, A.A. Peden, S.G. Hymowitz, and S.S. Sidhu. Molecular recognition by a binary code. *Journal of molecular biology*, 348(5):1153–1162, 2005.

[FS02] D. Frenkel and B. Smit. *Understanding molecular simulation*. Academic Press, 2002.

[FWE+11] 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.

[GB87] W.F. Van Gunsteren and H.J.C. Berendsen. Groningen molecular simulation (GROMOS). *Library manual, Biomos, Groningen, The Netherlands*, pages 1–221, 1987.

[GC05] M. Guharoy and P. Chakrabarti. Conservation and relative importance of residues across protein-protein interfaces. *PNAS*, 102(43):15447–15452, Oct 2005.

[GCJ56] Bruce Glick, Timothy S Chang, and R George Jaap. The bursa of fabricius and antibody production. *Poultry Science*, 35(1):224–225, 1956.

[GJDH81] Patricia J Gearhart, Nelson D Johnson, Richard Douglas, and Leroy Hood. Igg antibodies to phosphorylcholine exhibit more diversity than their igm counterparts. 1981.

[GK02] L. Györfi and A. Krzyzak. *A distribution-free theory of nonparametric regression*. Springer, 2002.

[GKE+16] S. Grudinin, M. Kadukova, A. Eisenbarth, S. Marillet, and F. Cazals. Predicting binding poses and affinities for protein - ligand complexes in the 2015 D3R grand challenge using a physical model with a statistical parameter estimation. *J. of Computer-Aided Molecular Design*, NA(NA):NA, 2016.

[Got93] Osamu Gotoh. Optimal alignment between groups of sequences and its application to multiple sequence alignment. *Computer applications in the biosciences: CABIOS*, 9(3):361–370, 1993.

[GR01] 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.

[GSF+95] 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.

[GT02] A. Golbraikh and A. Tropsha. Beware of q2! *Journal of Molecular Graphics and Modelling*, 20(4):269–276, 2002.

[GZ07] M.K. Gilson and H-X. Zhou. Calculation of protein-ligand binding affinities. *Annual review of biophysics and biomolecular structure*, 36(1):21, 2007.

[HL77] F. Hillier and G. Lieberman. *Introduction to mathematical programming*. McGraw-Hill, 1977.

[HL92] N. Horton and M. Lewis. Calculation of the free energy of association for protein complexes. *Protein Science*, 1(1):169–181, 1992.

[HP01] Annemarie Honegger and Andreas PluÈckthun. Yet another numbering scheme for immunoglobulin variable domains: an automatic modeling and analysis tool. *Journal of molecular biology*, 309(3):657–670, 2001.

[HS09] Jean Hausser and Korbinian Strimmer. Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research*, 10(Jul):1469–1484, 2009.

[ILIS02] I. Ivanov, J-M. Link, G. Ippolito, and H.H. Schroeder. Constraints on hydropathicity and sequence composition of HCDR3 are conserved across evolution. *The antibodies*, 7:43–67, 2002.

[Jan89] Charles A Janeway. Approaching the asymptote? evolution and revolution in immunology. In *Cold Spring Harbor symposia on quantitative biology*, volume 54, pages 1–13. Cold Spring Harbor Laboratory Press, 1989.

[Jan14] J. Janin. A minimal model of protein–protein binding affinities. *Protein Science*, 23(12):1813–1817, 2014.

[JBC08] J. Janin, R. P. Bahadur, and P. Chakrabarti. Protein-protein interaction and quaternary structure. *Quarterly reviews of biophysics*, 41(2):133–180, 2008.

[Jer72] NK Jerne. What precedes clonal selection. In *Ontogeny of Acquired Immunity. A Ciba Foundation Symposium*, pages 1–15, 1972.

[JHW+13] Ning Jiang, Jiankui He, Joshua A Weinstein, Lolita Penland, Sanae Sasaki, Xiao-Song He, Cornelia L Dekker, Nai-Ying Zheng, Min Huang, Meghan Sullivan, et al. Lineage structure of the human antibody repertoire in response to influenza vaccination. *Science translational medicine*, 5(171):171ra19–171ra19, 2013.

[JSY+04] Ralph Jimenez, Georgina Salazar, Jun Yin, Taiha Joo, and Floyd E Romesberg. Protein dynamics and the immunological evolution of molecular recognition. *PNAS*, 101(11):3803–3808, 2004.

[JT96] S. Jones and JM Thornton. Principles of protein-protein interactions. *PNAS*, 93(1):13–20, 1996.

[JWP+11] Ning Jiang, Joshua A Weinstein, Lolita Penland, Richard A White, Daniel S Fisher, and Stephen R Quake. Determinism and stochasticity during maturation of the zebrafish antibody repertoire. *PNAS*, 108(13):5348–5353, 2011.

[KB02] Tanja Kortemme and David Baker. A simple physical model for binding energy hot spots in protein–protein complexes. *Proceedings of the National Academy of Sciences*, 99(22):14116–14121, 2002.

[KJ13] P. L. Kastritis and Bonvin A. M. J. J. On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *Journal of Royal Society Interface*, 10:20120835, 2013.

[KKGE06] 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.

[KLS+14] J. Kaplinsky, A. Li, A. Sun, M. Coffre, S.B. Koralov, and R. Arnaout. Antibody repertoire deep sequencing reveals antigen-independent selection in maturing b cells. *PNAS*, 111(25):E2622–E2629, 2014.

[KMH+11] 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.

[KMR+05] Ozlem Keskin, Buyong Ma, Kristina Rogale, K Gunasekaran, and Ruth Nussinov. Protein–protein interactions: organization, cooperativity and mapping in a bottom-up systems biology approach. *Physical biology*, 2(2):S24, 2005.

[KRF+14] 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.

[KSKN08] 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.

[KTWB76] Elvin Abraham Kabat, Tai Te Wu, and Howard Bilofsky. *Sequences of Immunoglobulin Chains: Tabulation and Analysis of Amino Acid Sequences of Precursors, V-regions, C-regions, J-chain and-microglobulins*. US Dept. of Health, Education, and Welfare, Public Health Service, National Institutes of Health, 1976.

[KTWF+92] Elvin A Kabat, Tai Te Wu, Carl Foeller, Harold M Perry, and Kay S Gottesman. *Sequences of proteins of immunological interest*. DIANE publishing, 1992.

[LB01] E. Levina and P. Bickel. The earth mover's distance is the mallows distance: Some insights from statistics. In *IEEE ICCV*, volume 2, pages 251–256. IEEE, 2001.

[LC88] A. Lesk and C. Chothia. Elbow motion in the immunoglobulins involves a molecular ball-and-socket joint. *Nature*, 8(335):188–90, 1988.

[LC10] S. Loriot and F. Cazals. Modeling macro-molecular interfaces with Intervor. *Bioinformatics*, 26(7):964–965, 2010.

[LCJ99] L. Lo Conte, C. Chothia, and J. Janin. The atomic structure of protein-protein recognition sites. *JMB*, 285(5):2177–2198, 1999.

[LDB+89] Juan J Lafaille, Amy DeCloux, Marc Bonneville, Yohtaroh Takagaki, and Susumu Tonegawa. Junctional sequences of t cell receptor $\gamma\delta$ genes: implications for $\gamma\delta$ t cell lineages and for a novel intermediate of v-(d)-j joining. *Cell*, 59(5):859–870, 1989.

[Lef99] M-P. Lefranc. The IMGT Unique Numbering for Immunoglobulins, T-Cell Receptors and Ig-Like Domains. *The Immunologist*, 7(4):132–136, 1999.

[LL01] M-P. Lefranc and G. Lefranc. *The immunoglobulin FactsBook*. Academic Press, 2001.

[LLZ+06] 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.

[LM05] M. Lu and J. Ma. The role of shape in determining molecular motions. *Biophysical journal*, 89(4):2395–2401, 2005.

[LNL83] Bo Liedberg, Claes Nylander, and Ingemar Lunström. Surface plasmon resonance for gas detection and biosensing. *Sensors and actuators*, 4:299–304, 1983.

[LPR+03] 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.

[LR71] B. Lee and FM Richards. The interpretation of protein structures: Estimation of static accessibility* 1. *Journal of Molecular Biology*, 55(3):379–380, 1971.

[LSJW+84] Nathaniel R Landau, Thomas P St John, Irving L Weissman, Susan C Wolf, Allen E Silverstone, and David Baltimore. Cloning of terminal transferase cdna by antibody screening. *Proceedings of the National Academy of Sciences*, 81(18):5836–5840, 1984.

[LW13] M.F. Lensink and S.J. Wodak. Docking, scoring, and affinity prediction in CAPRI. *Proteins: Structure, Function, and Bioinformatics*, 81(12):2082–2095, 2013.

[LWT07] 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.

[MAB11] I.H. Moal, R. Agius, and P.A. Bates. Protein–protein binding affinity prediction on a diverse set of structures. *Bioinformatics*, 27(21):3002–3009, 2011.

[MABM10] 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.

[MBC15] 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.

[MDBC12] 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.

[MFR07] I. Moreira, P. Fernandes, and M.J. Ramos. Hot spots – a review of the protein–protein interface determinant amino-acid residues. *Proteins: Structure, Function, and Bioinformatics*, 68(4):803–812, 2007.

[MFR14] I. Moal and J. Fernández-Recio. Comment on *protein-protein binding affinity prediction from amino acid sequence. Bioinformatics (Oxford, England)*, 2014.

[MMT96] 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.

[MSSR00] 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.

[MTR+98] 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.

[MTS+96] 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.

[MWBC10] Thierry Mora, Aleksandra M Walczak, William Bialek, and Curtis G Callan. Maximum entropy models for antibody diversity. *Proceedings of the National Academy of Sciences*, 107(12):5405–5410, 2010.

[NLD11] 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.

[Pan03] Liam Paninski. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, 2003.

[Pau13] W.E. Paul. *Fundamental Immunology (7th Ed.)*. Lippincott Williams and Wilkins, Wolters and Kluwer, 2013.

[PDCR10] P. Privalov, A. Dragan, and C. Crane-Robinson. Interpreting protein/dna interactions: distinguishing specific from non-specific and electrostatic from non-electrostatic components. *Nucleic acids research*, pages 2483–2491, 2010.

[PHCB+96] 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.

[PMD+00] 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.

[PS10] B. Phipson and G.K. Smyth. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1), 2010.

[PVG+11] 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.

[RBCJ05] F. Rodier, R.P. Bahadur, P. Chakrabarti, and J. Janin. Hydration of protein - protein interfaces. *Proteins*, 60(1):36–45, 2005.

[RELW93] Marko Z Radic, Jan Erikson, S Litwin, and Martin Weigert. B lymphocytes may escape tolerance by revising their antigen receptors. *The Journal of experimental medicine*, 177(4):1165–1173, 1993.

[Ric74] Frederic M Richards. The interpretation of protein structures: total volume, group volume distributions and packing density. *Journal of molecular biology*, 82(1):1–14, 1974.

[RRA+05] D. Reichmann, O. Rahat, S. Albeck, R. Meged, O. Dym, and G. Schreiber. The modular architecture of protein-protein binding interfaces. *PNAS*, 102(1):57–62, 2005.

[RSWA12] 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.

[RTG00] Y. Rubner, C. Tomasi, and L.J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[RTVC04] D. Rajamani, S. Thiel, S. Vajda, and C.J. Camacho. Anchor residues in protein-protein interactions. *PNAS*, 101(31):11287–11292, 2004.

[RVK08] G. Subba Rao, R. Vijayakrishnan, and M. Kumar. Structure-based design of a novel class of potent inhibitors of inha, the enoyl acyl carrier protein reductase from mycobacterium tuberculosis: A computer modelling approach. *Chemical biology & drug design*, 72(5):444–449, 2008.

[Sch99] H. Schiessel. Counterion condensation on flexible polyelectrolytes: dependence on ionic strength and chain concentration. *Macromolecules*, 32(17):5673–5680, 1999.

[SKN96] H. Shirai, A. Kidera, and H. Nakamura. Structural classification of CDR-H3 in antibodies. *FEBS letters*, 399(1):1–8, 1996.

[SKN99] J. Shirai, A. Kidera, and H. Nakamura. H3-rules: identification of CDR-H3 structures in antibodies. *FEBS letters*, 455(1):188–197, 1999.

[SM02] E.J. Sundberg and R.A. Mariuzza. Molecular recognition in antibody-antigen complexes. *Advances in protein chemistry*, 61:119–160, 2002.

[SMFC+13] Adrien Six, Encarnita Mariotti-Ferrandiz, Wahiba Chaara, Susana Magadan, Hang-Phuong Pham, Marie-Paule Lefranc, Thierry Mora, Véronique Thomas-Vaslin, Aleksandra M Walczak, and Pierre Boudinot. The past, present, and future of immune repertoire biology–the rise of next-generation repertoire analysis. *Frontiers in immunology*, 4:413, 2013.

[SXK+13] 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.

[SZWR06] 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.

[TFFFC07] 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.

[Tib96] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[TLY12] Feifei Tian, Yonggang Lv, and Li Yang. Structure-based prediction of protein–protein binding affinity with consideration of allosteric effect. *Amino Acids*, 43(2):531–543, 2012.

[Ton83] S. Tonegawa. Somatic generation of antibody diversity. *Nature*, 302(5909):575–581, 1983.

[VB15] A. Vangone and A. Bonvin. Contacts-based prediction of binding affinity in protein–protein complexes. *eLife*, 4:e07454, 2015.

[VHPW12] Thom Vreven, Howook Hwang, Brian G Pierce, and Zhiping Weng. Prediction of protein–protein binding free energies. *Protein Science*, 21(3):396–404, 2012.

[Vil03] C. Villani. *Topics in optimal transportation*. Number 58. AMS, 2003.

[VKB15] K.M. Visscher, P.L. Kastritis, and A. Bonvin. Non-interacting surface solvation and dynamics in protein–protein interactions. *Proteins*, 83:445–458., 2015.

[VMLOA95] 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.

[VSW+13] Christopher Vollmers, Rene V Sit, Joshua A Weinstein, Cornelia L Dekker, and Stephen R Quake. Genetic measurement of memory b-cell recall using antibody repertoire sequencing. *Proceedings of the National Academy of Sciences*, 110(33):13463–13468, 2013.

[Wal03] D. J. Wales. *Energy Landscapes*. Cambridge University Press, 2003.

[WJW+09] J.A. Weinstein, N. Jiang, R.A. White, D.S. Fisher, and S.R. Quake. High-throughput sequencing of the zebrafish antibody repertoire. *Science*, 324(5928):807–810, 2009.

[WK70] T.T. Wu and E.A. Kabat. An analysis of the sequences of the variable regions of bence jones proteins and myeloma light chains and their implications for antibody complementarity. *The Journal of experimental medicine*, 132(2):211–250, 1970.

[WLZC12] Lin Wang, Zhi-Ping Liu, Xiang-Sun Zhang, and Luonan Chen. Prediction of hot spots in protein interfaces using a random forest model with hybrid features. *Protein Engineering Design and Selection*, page gzr066, 2012.

[WPW+97] Gary J Wedemayer, Phillip A Patten, Leo H Wang, Peter G Schultz, and Raymond C Stevens. Structural insights into the evolution of an antibody combining site. *Science*, 276(5319):1665–1669, 1997.

[WSJ11] Sergio E Wong, Ben D Sellers, and Matthew P Jacobson. Effects of somatic mutations on cdr loop flexibility during affinity maturation. *Proteins: Structure, Function, and Bioinformatics*, 79(3):821–829, 2011.

[XD00] 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.

[XRK+16] 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.

[YGHW13] Z. Yand, L. Guo, L. Hu, and J. Wang. Specificity and affinity quantification of protein - protein interactions. *Bioinformatics*, 29(9):1127–1133, 2013.

[ZOT+06] Jörg Zimmermann, Erin L Oakman, Ian F Thorpe, Xinghua Shi, Paul Abbyad, Charles L Brooks, Steven G Boxer, and Floyd E Romesberg. Antibody evolution constrains conformational heterogeneity by tailoring protein dynamics. *PNAS*, 103(37):13722–13727, 2006.

Communications

Journal papers

Simon Marillet, Pierre Boudinot, Frédéric Cazals. "High-resolution crystal structures leverage protein binding affinity predictions." *Proteins: Structure, Function, and Bioinformatics*, 84.1, (2016): 9-20.

Sergei Grudinin, Maria Kadukova,Andreas Eisenbarth, Simon Marillet, Frédéric Cazals. "Predicting binding poses and affinities for protein - ligand complexes in the 2015 D3R Grand Challenge using a physical model with a statistical parameter estimation". *Journal of computer-aided molecular design*, 30(9), 791-804.

Simon Marillet, Marie-Paule Lefranc, Pierre Boudinot, Frédéric Cazals. "Novel structural parameters of IG - Ag complexes yield a quantitative description of interaction specificity and binding affinity." *Submitted to Frontiers in Immunology.*

Simon Marillet, Pierre Boudinot, Frédéric Cazals. "Repertoire analysis of naive and infected fishes" *Manuscript in preparation.*

Appendices

Appendix A

Review of the latest statistical models for binding affinity prediction.

Multiple studies have sought to predict the affinity of protein-protein complexes. Most, and force field-based methods in particular due to their sheer computational cost, have focused on a limited number of complexes. This section briefly reviews the latest works which aim at predicting the binding affinity of general protein-protein complexes on a larger scale. To ease comparison, we describe each approach in three parts: the dataset(s) used for training and testing, the type of prediction model used, the variables used, and the statistical methodology for variable selection, training and validation.

Protein–protein binding affinity prediction on a diverse set of structures [MAB11]

This work aims at building a model from a comprehensive set of variables describing various structural aspects of a protein complex and the each individual partners. Four statistical methods are then combined to obtain the predictions.

Datasets. This work uses the structure affinity benchmark (SAB) [KMH$^+$11]. Seven complexes are discarded from the original benchmark: three because only the upper bound of the affinity is known (1UUG, 1IQD and 1NSN) and four because some features needed by the models are missing (1DE4, 1M10, 1NCA and 1NB5).

Types of models. Affinity values are predicted as the un-weighted average of the output of four different regression methods (random forest, multivariate adaptive regression splines, M5' regression trees and radial basis function interpolation). Each regression methods is fed a total of 200 different and possibly correlated features. All regression methods are able to perform feature selection to some extent and therefore, the used number of parameter is smaller than 200. In effect, M5' trees use 84 variables, random forest uses 19 variables, MARS uses 10 variables and RBF weights all variable equally and therefore virtually uses all 200 variables. The union of the three first sets contains 94 variables.

Variables. The variables fall into 7 categories:

- Statistical potentials (both atomistic and coarse-grained)
- Solvation and electrostatics (using force fields)
- Entropy terms (translational, rotational, vibrational)
- Contact potentials (H-bonds, π-π interactions, Van der Waals, salt bridges)

- Interface properties (BSA, polarity, geometrical features, surface complementarity)
- Change between bound and unbound states for all of the above
- All of the above computed on an ensemble of structures generated with CONCOORD.

Statistical procedures. The models are trained on a subset of 57 complexes with further validated affinity values (validated set). The predictions are tested on the training dataset using leave-one-out cross-validation. The prediction is further challenged on 80 complexes (test set). However, the reported results do not include the correlation between predictions and affinity on the test set alone. Instead, the correlations are reported for the test set augmented with the train set.

Prediction of protein–protein binding free energies [VHPW12]

This work introduces a model for affinity. Many terms from existing force-fields such a ZRANK, ZDOCK, Rosetta, pyDock and AffinityScore1.0 were screened. The results are stable for various subsets of the data, and in particular subsets defined by interface flexibility (iRMSD), and pH during experimental determination.

Dataset. The whole structure affinity benchmark was used (144 complexes).

Type of model. This study uses a linear model optimized for correlation with the experimental affinities (i.e. not least-squares).

Variables. Nine terms are used by this model: hydrogen bonds (from Rosetta), attractive and repulsive long-range electrostatics (from ZRANK), solvent loss upon binding (from Rosetta), hydrophobic surface loss upon binding, a residue-based docking contact potential (253 parameters), the number of loops and helices at interface, and number of atoms appearing at interface upon binding.

Statistical procedures. Various combinations of many terms are tested until a combination of 9 terms cannot be improved by the addition of another term.

Leave-one-out cross-validation is used to evaluate the models.

Structure-based prediction of protein–protein binding affinity with consideration of allosteric effect [TLY12]

This study build three model to assess possible allosteric effects upon association of two proteins. In particular, the model ignoring conformational changes upon binding (*i.e.* assuming the isolated partners have the exact same conformation than in the complex) performs worse than one that does.

Dataset. The whole structure affinity benchmark is used (144 complexes).

Type of model. This study uses partial least-squares regression.

Variables. Four variables are used to describe the interaction of each pair of amino acid. Namely, electrostatics (coulomb potential), van der Walls interactions (Lennard-Jones potential), hydrogen bonds (angle-weighted Lennard-Jones-like 8-6 function), and the hydrophobic effect (distance-dependent potential based on Eisenberg solvation parameters and exposed surface area).

For each pair of amino-acid, each potential is summed over all such pairs, resulting in 840 (= 4 * 20 * 21 / 2) terms. The differences between the two states state on these terms are used in the model.

These states can be 1) unbound, 2) allosteric intermediates (*i.e.* partners in bound conformation which have been separated), 3) bound. This results in 3 models of 840 variables each.

Statistical procedures. A Genetic algorithm is used for selecting the most important variables. In the end 378 are selected and subsequently used in the resulting model (based on the difference between states 1 and 3) out of 840.

Cross-validation, an external test-set and monte-carlo cross-validation are used to assess the performance of the model.

A minimal model of protein–protein binding affinities [Jan14]

This work aims at building a baseline model for affinity prediction using only two variables already computed and included in the SAB.

Datasets. The set of rigid complexes (with iRMSD < 1.1) minus six complexes (1UUG, 2PTC, 1BRS, 2BTF, 1Z0K and 1S1Q) is used to fit the model. In SAB-I (see Section 3.2.1), four more complexes are also removed: 1EMV, 1KXP, 1AKJ and 1WQ1.

Model. A linear model is fitted on the data using least-square regression.

Variables. This model uses two variables: iRMSD, which is the least RMSD of interface atoms between the bound and unbound conformations of the partners, and the buried surface area.

Statistical procedures. No cross-validation is involved. The correlation between fitted and experimental values is computed on various subset of the SAB and the whole SAB. The results are therefore optimistic, in particular on rigid complexes.

Remarks. In various datasets, complexes which were badly predicted by the model are removed as outliers. This leads to artificially high correlation coefficients. This is denoted by yellow cells in Table 3.2.

Specificity and affinity quantification of protein - protein interactions [YGHW13]

This study first aims at creating a scoring function for docking using statistical parameters. The correlation between the score of a complex and its affinity is then computed.

Dataset. The original full dataset consists in 3045 complexes extracted from DOCKGROUND, and the training dataset consists of half of it. Moreover, docking decoys are used as negative examples. The test set is SAB with 1UUG, 1IQD, 1NSN, 1DE4, 1M10, 1NCA and 1NB5 removed.

Model. The model is based on a scoring function optimized to discriminate between native complexes and decoys. This function uses knowledge-based statistical potentials derived from the training set *i.e.* the probability of a given pair of atoms interacting in a given radius compared to that same probability for non-interacting atoms.

Variables. The equivalent of variables for that model are the distance-dependent atom-pair potentials. These are based on observed and expected frequencies of occurrences of atom pairs. From 12 atom types, 78 different pairs occur, and this is computed for 14 different radii, leading to 1092 parameters

Statistical procedures. No cross-validation is used since the training and test are assumed to be disjoint. It is worth noting however that 24 complexes from the SAB are also part of the 3045 original complexes. Since 1UUG was removed that results in at most 23 complexes shared between the training and test sets.

Proteins Feel More Than They See: Fine-Tuning of Binding Affinity by Properties of the Non-Interacting Surface [KRF$^+$14]

This study introduces two new variables describing the non-interface surface (NIS) in addition to the classical BSA.

Dataset. Only one complex is removed from the SAB, namely 2OZA because "its BSA was extraordinary large and detected as an outlier using the standard Grubbs' test"

Model. A linear model is fitted on a rigid subset of the SAB using least-square regression. This training set is defined by complexes with an iRMSD $\leq$ 1Å.

Variables. Three variables are used: the BSA which accounts for interface properties, NIS^{charged} and NIS^{polar} which are respectively the proportion of charged and polar residue at the surface of the complex (outside the interface).

Statistical procedures. 4-fold cross-validation is performed to assess the performances of the model on the training set.

Remarks, see Section 3.3. The cross-validation procedure used in this work is a less strict strategy than the standard cross validation, where the intersection between the data used to train and test is void. Namely, during the 4-fold cross-validation, the coefficients of the four models trained on their respective folds are averaged to get a single model. The correlation coefficient of the prediction of that model with the actual values is then reported. Therefore, through averaging of the coefficients, information about the whole dataset is used for training a model that is tested on the very same dataset, leading to overfitting. This is denoted by an orange cell in Table 3.2.

Moreover, dataset SAB-I is a superset of the training set. Namely, the model is trained on all complexes with iRMSD $<$ 1Å and tested on complexes with iRMSD $<$ 1.5Å. This is another instance of overfitting. This is denoted by a cyan cell in Table 3.2.

How structure defines affinity in protein-protein interaction [ERS14]

Dataset. This paper uses the SAB from which ten complexes are filtered out, namely: 1BJ1, 1F34, 1JIW, 1JMO, 1S1Q, 1XD3, 2J0T, 2TGP, 1NVU and 2OZA. It also builds a second dataset consisting of high-resolution entries, *i.e.* complexes for which both the individual partners and the bound structure have a resolution lower than 2.5Å.

Model. A linear model is fitted on the data using least-square regression.

Variables. The variables used consist in intra and inter-chain hydrogen bond potentials, geometric complementarity (Van der Waals interactions), volume of cavities at the surface (large enough to contain water molecules), iRMSD of interface atoms, C-α and side-chains χ1 and χ2 dihedral angles, alanine-scanning defined hotspots, interface amino-acid propensities and electrostatics (Coulomb). In total the combinations of 13 variables are studied. The authors mention that adding more than four variables does not significantly improve the results, but the reported correlation coefficient seem to be highest for models of 7 or more variables (figure 5 of the paper). Which variables are actually selected by this procedure is not reported.

Statistical procedures. The reported correlation coefficients are computed using leave-one-out cross-validation.

Contacts-based prediction of binding affinity in protein–protein complexes [VB15]

This study build upon [KRF$^+$14] and introduces new variables accounting for the number of contacts between amino-acids of different classes; namely charged, polar and apolar.

Dataset. The dataset used is the SAB with 19 cases discarded; three because only the upper bound of the affinity is known, and 16 because of missing residues at interface.

Model. A linear model is fitted on the data using least-squares regression.

Variables. The selected model uses the following variables: $\text{ICs}_{charged/charged}$, $\text{ICs}_{charged/apolar}$, $\text{ICs}_{polar/polar}$, $\text{ICs}_{polar/apolar}$, NIS^{apolar}, $\text{NIS}^{charged}$. The variables NIS^{apolar} and $\text{NIS}^{charged}$ have been defined in [KRF$^+$14]. ICs are the number of inter-residue contacts at interface (defined with a cutoff of 4Å) between charged/charged, charged/apolar, polar/polar and polar/apolar residues respectively.

Statistical procedures. The Akaike's Information Criterion (AIC) stepwise selection method (backward and forward) is used for variable selection. The performance of resulting models is then evaluated using four-fold cross-validation.

Appendix B

Protein - protein affinity prediction: High Resolution Crystal Structures Leverage Protein Binding Affinity Predictions

Table B.1 lists the individual predictions, obtained from Eq. (3.16).

Table B.1 Experimental affinities on a per complex basis: experimental measurements (ΔG_d) versus predictions ($\hat{g}_i$, Eq. 3.16). Predictions were generated with predictive Model 1 on dataset SAB-A using linear regression. The median was taken over the NXV repetitions. Blue values indicate under-predicted complexes (63) and red indicate the over-predicted ones (76). A start denotes complexes with error in the top decile.

PDB ID	Measured	Predicted	PDB ID	Measured	Predicted	PDB ID	Measured	Predicted
1A2K	9.31	10.72	1I4D	7.46	9.55	1XU1	11.18	10.66
1ACB	13.05	12.63	1IB1	9.76	10.65	1YVB	11.17	9.48
1AHW	11.55	11.25	1IBR	12.07	12.32	1Z0K	6.98	10.22
1AK4	6.43	10.00	1IJK	10.42	8.85	1ZHI	9.08	9.09
1AKJ	5.32	10.34 *	1J2J	8.13	8.6	1ZM4	8.03	9.26
1ATN	12.07	10.29	1JIW	15.55	12.55	2A9K	10.25	9.93
1AVX	12.50	12.66	1JMO	9.47	11.74	2ABZ	11.67	11.06
1AVZ	6.55	10.24	1JPS	13.64	11.95	2AJF	10.63	10.27
1AY7	13.23	10.46	1JTG	12.82	13.52	2AQ3	6.71	9.17
1B6C	8.94	9.35	1JWH	11.14	9.17	2B42	12.11	13.86
1BJ1	11.55	12.09	1K5D	12.77	10.22	2B4J	10.86	10.4
1BRS	17.32	10.76 *	1KAC	10.68	11.74	2BTF	7.69	10.68
1BUH	9.70	9.91	1KKL	10.02	9.39	2C0L	9.82	10.73
1BVK	10.53	10.65	1KLU	7.28	9.75	2FJU	7.20	9.35
1BVN	15.06	12.58	1KTZ	8.92	9.95	2GOX	12.08	10.01
1CBW	10.75	11.88	1KXP	12.34	12.79	2HLE	10.09	11.19
1DE4	9.78	10.12	1KXQ	11.54	12.43	2HQS	10.15	13.37
1DFJ	18.05	11.34 *	1LFD	7.79	8.34	2HRK	10.98	10.93
1DQJ	11.67	12.52	1M10	11.24	10.7	2I25	12.28	10.84
1E4K	7.87	10.33	1MAH	14.51	11.49	2I9B	12.93	11.81
1E6E	8.28	10.28	1MLC	9.61	11.27	2J0T	13.34	10.53
1E96	7.42	8.82	1MQ8	7.53	9.02	2JEL	11.59	11.53
1EAW	14.06	12.13	1NB5	13.86	9.77 *	2MTA	7.42	9.73
1EER	15.59	12.67	1NCA	11.02	11.29	2NYZ	12.69	13.19
1EFN	10.12	10.48	1NVU	7.43	10.94	2O3B	15.68	11.65 *
1EMV	18.58	7.54 *	1NVU	7.80	12.18 *	2OOB	5.66	7.47
1EWY	7.43	9.42	1NW9	11.19	10.72	2OOR	10.65	10.72
1EZU	13.77	11.23	1OC0	12.28	10.53	2OUL	11.96	10.9
1F34	14.19	13	1OPH	11.32	11.52	2OZA	11.73	14.81
1F6M	7.60	9.64	1P2C	13.63	11.88	2PCB	6.82	8.79
1FC2	10.43	9.68	1PPE	15.56	12.92	2PCC	7.91	9.07
1FFW	8.09	9.51	1PVH	9.52	10.32	2PTC	18.04	12.57 *
1FLE	12.28	13.21	1PXV	12.97	12.63	2SIC	13.84	13.47
1FQJ	9.79	9.82	1QA9	7.16	8.65	2SNI	15.96	12.15
1FSK	13.12	11.48	1R0R	14.17	13.05	2TGP	7.54	12.7 *
1GCQ	6.51	9.59	1R6Q	8.84	10.94	2UUY	11.26	13.3
1GL1	13.23	12.18	1RLB	8.18	8.83	2VDB	13.40	8.42 *
1GLA	6.76	8.84	1RV6	13.86	9.93	2VIR	12.28	11.02
1GPW	11.32	10.8	1S1Q	4.29	9.8 *	2VIS	7.36	11.34 *
1GRN	9.03	11.5	1T6B	13.10	10.39	2WPT	10.67	5.92 *
1GXD	11.30	10.25	1US7	8.09	8.24	3BP8	11.44	10.44
1H1V	10.20	10.83	1VFB	11.46	12.36	3BZD	9.57	9.39
1H9D	9.18	9.03	1WDW	12.72	10.52	3CPH	8.84	9.55
1HCF	13.08	9.82	1WEJ	12.48	11.3	3SGB	14.51	13.25
1HE8	7.37	8.78	1WQ1	6.62	11.56 *	4CPA	11.32	11.23
1HIA	10.76	11.39	1XD3	8.90	11.14			
1I2M	15.83	13.18	1XQS	7.08	10.71			

Table B.2 Experimental affinities on a per complex basis: experimental measurements (ΔG_d) versus predictions ($\hat{g}_i$, Eq. 3.16). Predictions were generated with predictive Model 1 on dataset SAB-A with k-nearest neighbors regression. The median was taken over the NXV repetitions. Blue values indicate under-predicted complexes (66) and red indicate the over-predicted ones (71). A start denotes complexes with error in the top decile.

PDB ID	Measured	Predicted	PDB ID	Measured	Predicted	PDB ID	Measured	Predicted
1A2K	9.31	10.43	1I4D	7.46	10.85	1XU1	11.18	10.70
1ACB	13.05	12.25	1IB1	9.76	13.36	1YVB	11.17	9.75
1AHW	11.55	12.58	1IBR	12.07	11.63	1Z0K	6.98	12.96
1AK4	6.43	10.00	1IJK	10.42	9.33	1ZHI	9.08	10.01
1AKJ	5.32	11.79	1J2J	8.13	9.11	1ZM4	8.03	9.09
1ATN	12.07	11.57	1JIW	15.55	12.33	2A9K	10.25	12.04
1AVX	12.50	12.44	1JMO	9.47	11.82	2ABZ	11.67	9.51
1AVZ	6.55	8.8	1JPS	13.64	12.82	2AJF	10.63	9.23
1AY7	13.23	11.78	1JTG	12.82	11.31	2AQ3	6.71	8.88
1B6C	8.94	9.89	1JWH	11.14	8.56	2B42	12.11	NA
1BJ1	11.55	12.42	1K5D	12.77	12.00	2B4J	10.86	10.49
1BRS	17.32	12.72	1KAC	10.68	8.90	2BTF	7.69	12.46
1BUH	9.70	9.35	1KKL	10.02	8.40	2C0L	9.82	13.33
1BVK	10.53	8.87	1KLU	7.28	10.68	2FJU	7.20	9.29
1BVN	15.06	11.22	1KTZ	8.92	8.40	2GOX	12.08	9.20
1CBW	10.75	11.65	1KXP	12.34	11.05	2HLE	10.09	11.48
1DE4	9.78	9.7	1KXQ	11.54	11.96	2HQS	10.15	NA
1DFJ	18.05	11.32	1LFD	7.79	10.16	2HRK	10.98	11.92
1DQJ	11.67	12.08	1M10	11.24	10.07	2I25	12.28	12.99
1E4K	7.87	9.01	1MAH	14.51	12.45	2I9B	12.93	12.25
1E6E	8.28	12.99	1MLC	9.61	10.18	2J0T	13.34	10.58
1E96	7.42	9.44	1MQ8	7.53	8.96	2JEL	11.59	11.19
1EAW	14.06	12.53	1NB5	13.86	9.68	2MTA	7.42	9.15
1EER	15.59	11.19	1NCA	11.02	12.74	2NYZ	12.69	11.25
1EFN	10.12	9.57	1NVU	7.43	12.43	2O3B	15.68	12.74
1EMV	18.58	11.38	1NVU	7.80	11.94	2OOB	5.66	9.06
1EWY	7.43	9.37	1NW9	11.19	13.23	2OOR	10.65	12.56
1EZU	13.77	11.69	1OC0	12.28	9.69	2OUL	11.96	11.82
1F34	14.19	12.97	1OPH	11.32	12.34	2OZA	11.73	10.98
1F6M	7.60	9.75	1P2C	13.63	11.77	2PCB	6.82	8.83
1FC2	10.43	9.14	1PPE	15.56	12.62	2PCC	7.91	9.07
1FFW	8.09	8.85	1PVH	9.52	9.75	2PTC	18.04	12.64
1FLE	12.28	11.44	1PXV	12.97	11.97	2SIC	13.84	11.79
1FQJ	9.79	11.75	1QA9	7.16	9.35	2SNI	15.96	12.09
1FSK	13.12	12.07	1R0R	14.17	12.13	2TGP	7.54	12.27
1GCQ	6.51	9.69	1R6Q	8.84	12.68	2UUY	11.26	11.86
1GL1	13.23	12.68	1RLB	8.18	8.87	2VDB	13.40	9.29
1GLA	6.76	8.71	1RV6	13.86	8.90	2VIR	12.28	8.57
1GPW	11.32	13.22	1S1Q	4.29	10.51	2VIS	7.36	8.76
1GRN	9.03	12.64	1T6B	13.10	11.34	2WPT	10.67	13.22
1GXD	11.30	11.81	1US7	8.09	9.91	3BP8	11.44	9.09
1H1V	10.20	12.38	1VFB	11.46	11.80	3BZD	9.57	8.88
1H9D	9.18	11.71	1WDW	12.72	12.69	3CPH	8.84	11.67
1HCF	13.08	11.07	1WEJ	12.48	12.18	3SGB	14.51	12.94
1HE8	7.37	9.07	1WQ1	6.62	11.24	4CPA	11.32	8.93
1HIA	10.76	11.93	1XD3	8.90	11.91			
1I2M	15.83	11.03	1XQS	7.08	12.62			

Table B.3 Pearson correlation coefficients between the individual variables and the affinity.

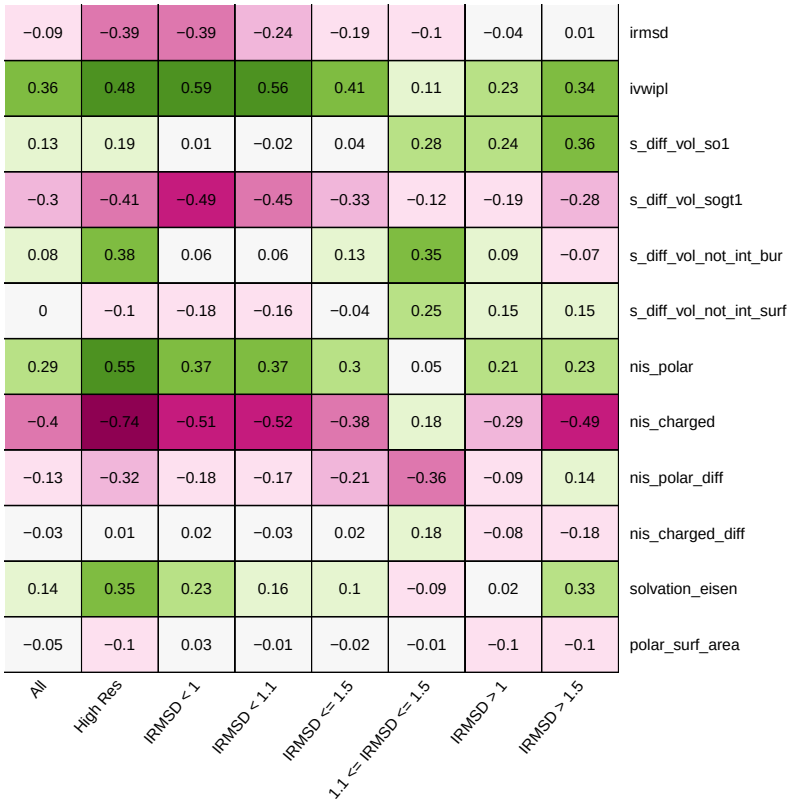


Table B.4 Validation of the best overall model *i.e.* model 9 on an external test set. See Table 3.4 for the statistics presented.

Dataset	SAB-A	SAB-A-HR	SAB-$R_{1.0}$	SAB-$R_{1.1}$	SAB-$R_{1.5}$	SAB-I	SAB-F_1	SAB-$F_{1.5}$
dataset size	51	24	13	16	23	7	38	28
p-value	0.0003	0.0727	0.0471	0.0392	0.0565	0.3673	0.0190	0.0155
correlation	0.48	0.37	0.56	0.52	0.40	0.40	0.38	0.45
$p_{1.4}^{\text{error}}$, $p_{2.8}^{\text{error}}$, $p_{4.2}^{\text{error}}$	11.76, 33.33, 49.02	20.83, 37.50, 45.83	30.77, 46.15, 46.15	31.25, 43.75, 43.75	21.74, 34.78, 39.13	14.29, 14.29, 28.57	13.16, 39.47, 50.00	17.86, 39.29, 60.71
median corr.	0.47	0.71	0.64	0.63	0.46	-0.24	0.27	0.42
$p_{1.4}^{\text{error}}$, $p_{2.8}^{\text{error}}$, $p_{4.2}^{\text{error}}$	48.2, 79.14, 91.37	51.35, 86.49, 94.59	57.35, 79.41, 91.18	55.13, 79.49, 91.03	43.81, 77.14, 91.43	40.74, 66.67, 88.89	51.35, 86.49, 94.59	52.94, 79.41, 91.18

Appendix C

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

Table C.1 Main features of the Ig - Ag complexes found in the structure affinity benchmark. {H,L}CDR len: length of the CDRs in residues. Numbers in the V{H,L} CDR 1,2,3 columns correspond to the first and last residue numbers in IMGT renumbered PDB files.

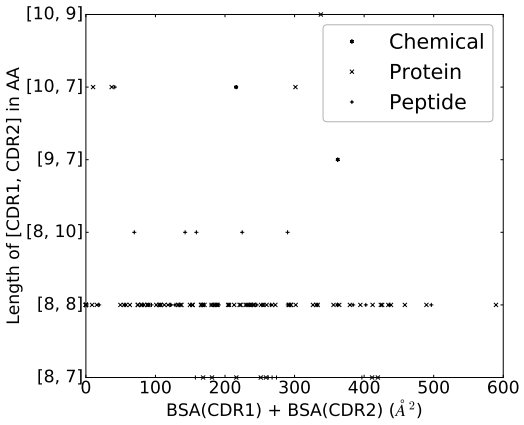
PDB ID	Ig H	Ig L	Ag	Ag type	Species	VH V and J gene	VL V and J gene
1AHW	B	A	C	Protein	Mus musculus (house mouse)	IGHV14-1*02 IGHJ2*01	IGKV14-111*01 IGKJ2*01
1BJ1	H	L	WV	Protein	Humanized (humanized)	IGHV7-4-1*02 IGHJ2*01	IGKV1-33*01 IGKJ1*01
1BVK	E	D	F	Protein	Humanized (humanized)	IGHV4-59*01 IGHJ4*03	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 len	VL CDR len	Ag size (#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);

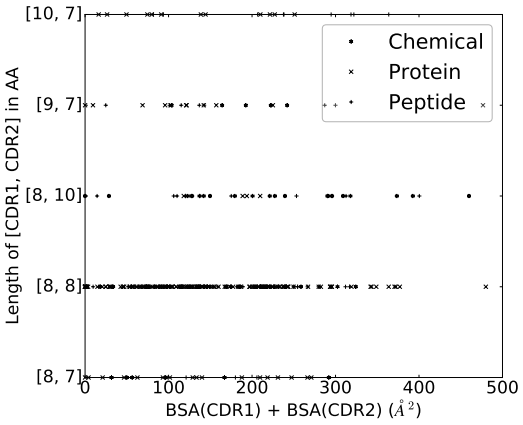
PDB ID	Ig name	Resolution	VH CDR1	VH CDR2	VH CDR3	VL CDR1	VL CDR2	VL CDR3
1AHW	AB-GAMMA-1_KAPPA	3.0	27 38	56 65	105 117	27 38	56 65	105 117
1BJ1	AB-GAMMA-1_KAPPA	2.4	27 38	56 65	105 117	27 38	56 65	105 117
1BVK	V-HEAVY_KAPPA	2.7	27 38	56 65	105 117	27 38	56 65	105 117
1DQJ	AB-GAMMA-2A_KAPPA	2.0	27 38	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)

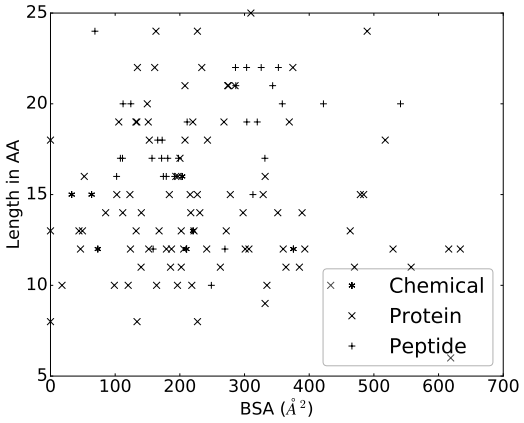
Figure C.1 Human and mouse VH CDR length versus BSA. The [CDR1. CDR2] length are characteristic of the different *Homo sapiens* and *Mus musculus* VH subgroups. There are highly varying levels of BSA for CDR of the same length. The information given by the length of a CDR is therefore not sufficient to infer its contribution to the interface.



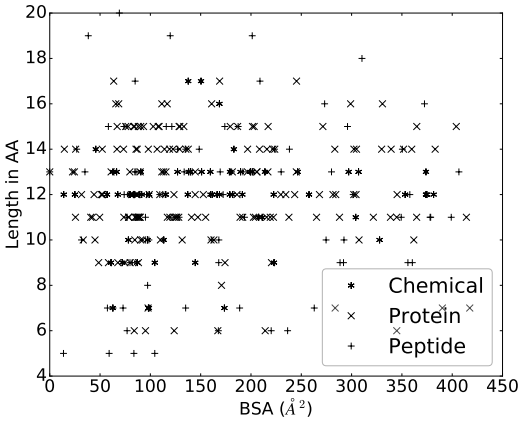
(a) Human [VH CDR1. VH CDR2]. Five complexes are discarded because of aberrant VH CDR1 and VH CDR2 lengths



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

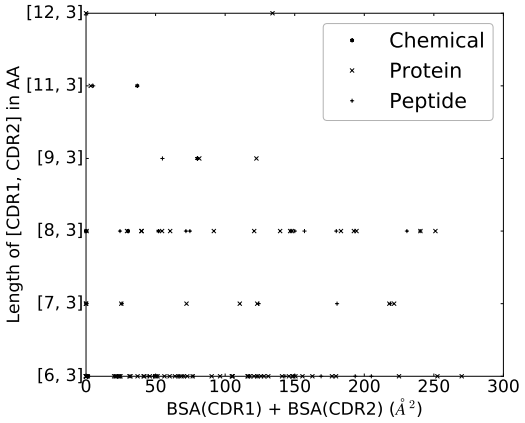


(c) Human VH CDR3. Twelve complexes are discarded because of aberrant VL CDR1 and VL CDR2 lengths

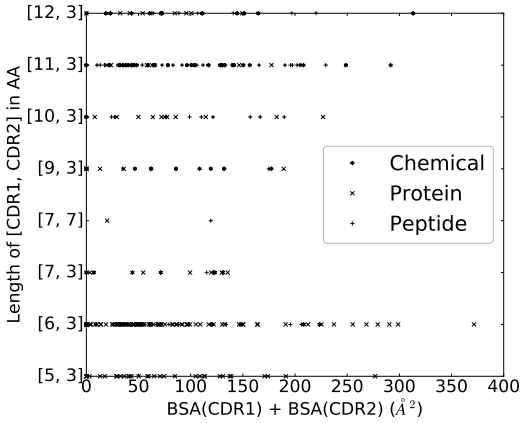


(d) Mouse VH CDR3.

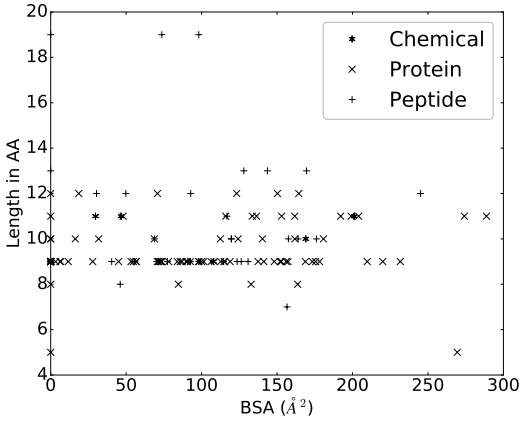
Figure C.2 Human and mouse VL CDR length versus BSA. The human [CDR1.CDR2] lengths [6.3] characterize both V-kappa and V-lambda. The other lengths characterize either V-kappa ([7.3], [11.3] and [12.3]) or V-lambda ([8.3] and [9.3]). The mouse [CDR1.CDR2] lengths [7.7] and [9.3] characterize V-lambda. 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.



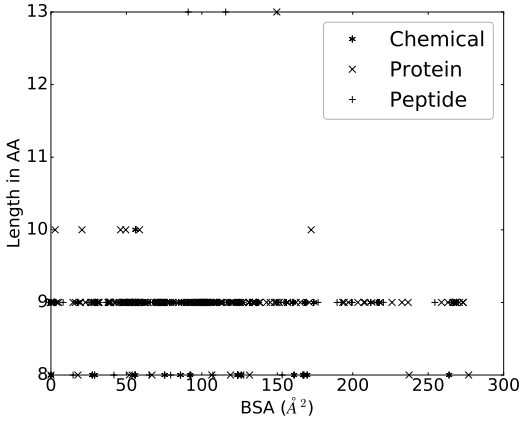
(a) Human [VL CDR1. VL CDR2].



(b) Mouse VL CDR1 and VL CDR2.



(c) VL CDR3, Human.



(d) VL CDR3, Mouse.

Abstract

This thesis investigates three topics at the cross-roads of structural biology, genetics and immunology.

First, we develop a pipeline to design and select binding affinity predictors for protein complexes, yielding state-of-the art results. The first step is the design and computation of 12 different variables accounting for geometric and physico-chemical properties of the complexes. The second step is the generation and evaluation of models using subsets of these variables, followed by the selection of the best performing ones. The corresponding software is distributed within the Structural Bioinformatics Library.

Second, we provide an analysis of the interface properties of Ig - Ag complexes. In particular, we design a classifier using two descriptors, which is able to distinguish ligand types. We also apply the previous binding affinity prediction model to Ig - Ag complexes and obtain accurate predictions. We then develop a quantitative model for the contribution of VH CDR3 to the binding affinity and interaction specificity, and show that it contributes significantly more than other CDRs.

Third, we model the diversity of VH CDR3 repertoires from Ig RNA sequencing data in a fish vaccination model. We analyze repertoires from three conditions: naive, vaccinated and vaccinated + infected fish. Comparison of the repertoires of two individuals uses the earth-mover distance (EMD). By exploiting a mapping between the clonotypes of the repertoires, we show that EMD reveals information beyond classical methods based on diversity indexes. To characterize the notion of public / private immune response, we quantify the overlap of clonotypes between individuals of the same or different conditions.

Cette thèse étudie trois sujets relevant de la biologie structurale, de la génétique et de l'immunologie.

Premièrement, nous développons de nouveaux prédicteurs de l'affinité de liaison de complexes protéiques, produisant des résultats de niveau "état de l'art". Nous calculons d'abord 12 variables modélisant diverses propriétés structurales des complexes. Nous générons et évaluons des estimateurs utilisant des sous ensembles de ces variables, de façon à identifier les plus performants. Le logiciel associé est distribué dans la Structural Bioinformatics Library.

Deuxièmement, nous proposons de nouvelles analyses de complexes Ig - Ag. D'une part nous concevons un classificateur distinguant les types de ligand des Ig. D'autre part, nous montrons que le modèle précédent prédit fidèlement l'affinité de complexes Ig - Ag. Enfin, nous quantifions la contribution des CDR3 de la chaine lourde à l'affinité de liaison, et montrons qu'il contribue significativement plus que les autres CDR.

Enfin, nous nous intéressons à la modélisation de la diversité des répertoires de chaîne lourde des Igs, à partir de données de séquençage de CDR3, dans un modèle de vaccin chez le poisson. Nous comparons les répertoires de deux individus en utilisant la "earth-mover distance", laquelle exploite la correspondance entre clonotypes de deux répertoires, révélant ainsi des informations inaccessibles aux méthodes basées sur les indices de diversité. Pour caractériser la notion de réponse immunitaire publique / privée, nous quantifions le chevauchement des clonotypes exprimés entre individus de la même ou de différentes conditions.