uc3m | Universidad Carlos III de Madrid

University Degree in Biomedical Engineering
Academic Year (2025-2026)

Bachelor Thesis

"Development of Computational Workflows for Identification of Antibody-Recognized Epitopes"

Irene Sánchez Martín

Carlos Óscar Sorzano
Leonor Kremer Barón
Cristina Quílez López

Leganés, 2026

RESUMEN

El objetivo de este Trabajo de Fin de Grado es diseñar y validar un flujo de trabajo computacional para la identificación de epítopos (sitios específicos del antígeno al que se une un determinado anticuerpo), combinando herramientas de predicción estructural y análisis de interfaz.

El flujo de trabajo desarrollado permite generar modelos tridimensionales de complejos antígeno-anticuerpo usando tres predictores de estructura basados en inteligencia artificial (AlphaFold, Boltz-1 y Chai-1), produciendo un total de 15 modelos por anticuerpo, y posteriormente integrarlos en un entorno reproducible (Scipion) para identificar las regiones de contacto entre el antígeno y el anticuerpo. A partir de estos modelos, se aplica una estrategia de consenso multimodelo que, para cada residuo del antígeno, contabiliza en cuántos de los 15 modelos aparece como residuo de contacto con el anticuerpo, destacando los residuos con mayor probabilidad de pertenecer al epítopo. Los resultados se visualizan en el entorno estructural interactivo PyMOL.

El flujo de trabajo se ha aplicado a cinco anticuerpos monoclonales (mAbs) específicos del receptor CCR9 humano (91R, 92R, 115, 3C3 y Ma-10), validando los resultados obtenidos frente a los epítopos experimentalmente caracterizados en la literatura y proponiendo un epítopo candidato para el anticuerpo Ma-10, cuyo epítopo no ha sido publicado hasta la fecha.

Palabras clave: Bioinformática, predicción estructural, Scipion, flujo de trabajo computacional, anticuerpos, epítopos.

ABSTRACT

The aim of this Bachelor's Thesis is to design and validate a computational workflow for the identification of epitopes (specific regions of an antigen recognized by a given antibody), combining structural prediction tools and interface analysis methods.

The developed workflow generates three-dimensional models of antigen-antibody complexes using three artificial intelligence-based structure predictors (AlphaFold, Boltz-1, and Chai-1), producing a total of 15 models per antibody, and integrates them within a reproducible environment (Scipion) to identify contact regions between the antigen and the antibody. From these models, a multi-model consensus strategy is applied that, for each antigen residue, counts in how many of the 15 models it appears as a contact residue with the antibody, highlighting the residues most likely to belong to the epitope. Results are visualized using the PyMOL interactive structural environment.

The workflow has been applied to five monoclonal antibodies (mAbs) specific to the human CCR9 receptor (91R, 92R, 115, 3C3, and Ma-10), validating the obtained results against experimentally characterized epitopes reported in the literature, and proposing a candidate epitope for antibody Ma-10, whose binding region has not been experimentally determined to date.

Keywords: Bioinformatics, structural prediction, Scipion, computational workflow, antibodies, epitopes.

DEDICATION

En primer lugar, quiero expresar mi gratitud al Centro Nacional de Biotecnología (CNB-CSIC) y en particular a la Unidad de Biocomputación (Laboratorio B13), por darme la oportunidad de continuar colaborando con ellos después de mis prácticas para desarrollar este TFG. Esta experiencia me ha permitido no solo aprender sobre bioinformática estructural y biología computacional, sino también he obtenido una formación práctica que no habría podido obtener en otro entorno. El nivel científico del equipo y el gran ambiente de trabajo han sido muy enrequecedores.

Agradezco especialmente a Carlos Óscar Sorzano su dirección, dedicación y guía a lo largo de todo el proceso. A pesar de tener una agenda muy ocupada, siempre encontró un momento para orientarme y resolver mis dudas.

Quiero también agradecer a Leonor Kremer su generosidad y disponibilidad aun no siendo mi tutora oficial. Su conocimiento y sus ideas me ayudaron mucho en el desarrollo de este trabajo. Asimismo, me proporciono material del cual saque mucho provecho. Me exigió mucho y hoy lo agradezco profundamente.

A Cristina Quílez, mi tutora de la UC3M, por ayudarme a asegurar la coherencia y estructura del trabajo.

Por último, gracias a mis padres, por su apoyo incondicional durante estos meses. Aunque desconocen el área, siempre me escuchan con entusiasmo y mucho interés, incluso cuando es la decimoquinta vez que les explico lo mismo. Y, por supuesto, a mis abuelos a Loli y a Rolandito, que no fallan en llamar para estar al tanto de las novedades. Esa ilusión ha sido una gran motivación.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Definition
AbAgIoU	Antibody-antigen intersection over union
AF	AlphaFold
AI	Artificial intelligence
AUC	Area under the curve
CAR-T	Chimeric antigen receptor T cell
CCL25	C-C motif chemokine ligand 25 (TECK)
CCR9	C-C chemokine receptor type 9
CDR	Complementarity-determining region
CT	C-terminal domain
EC1/2/3	Extracellular loop 1/2/3
ECL	Extracellular loop
GPCR	G protein-coupled receptor
HDX-MS	Hydrogen-deuterium exchange mass spectrometry
ICL	Intracellular loop
ipTM	Interface predicted template modeling score
IVA	Impuesto sobre el valor añadido
mAb	Monoclonal antibody
NT	N-terminal domain
PCC	Pearson correlation coefficient
PDB	Protein Data Bank
pLDDT	Predicted local distance difference test
pTM	Predicted template modeling score
RMSD	Root mean square deviation
ROI	Region of interest
scFv	Single-chain variable fragment
SMI	Salario mínimo interprofesional
SPR	Surface plasmon resonance
T-ALL	T-cell acute lymphoblastic leukemia
TECK	Thymus-expressed chemokine (CCL25)
TM	Transmembrane helix
VH	Variable heavy chain
VL	Variable light chain
WSL2	Windows Subsystem for Linux 2

1. INTRODUCTION

1.1. Scientific context and motivation

Monoclonal antibodies (mAbs) are now the largest and fastest-growing class of approved therapeutics with applications from autoimmune disease to cancer. In oncology, CCR9-directed therapies (including monoclonal antibodies and CAR-T constructs) have drawn growing interest due to the selective expression of CCR9 on the T cells implicated in these diseases, making anti-CCR9 antibodies useful tools for both identifying and eliminating these cell populations. A fundamental step in the development of an antibody therapy is the identification of the epitope recognized by that antibody. This is the specific region of the target protein the antibody recognizes. Experimental epitope mapping is reliable but really slow and costly, especially for membrane proteins. On the other hand, computational prediction offers a faster and more cost-effective alternative that can, at a later stage, be confirmed by experimental methods with less time effort.

CCR9 is a G protein-coupled receptor (GPCR) with seven transmembrane domains, mainly expressed on intestinal T cells and thymocytes. In this bachelor thesis, we will focus on the longer isoform CCR9A (UniProt P51686-1) which has 369 amino acids. The shorter form, CCR9B, does not include the first 12 aminoacids of the N-terminal domain. It starts at the methionine corresponding to position 13 of the longer isoform.

Its 48-residue extracellular N-terminal domain is the main target for antibody recognition. Four antibodies with experimentally characterized epitopes are described in the literature: 91R and 92R [1], [2], both recognizing PNMADD (residues 11–16); 115 [3], recognizing MEDYVNF (25–31); and 3C3 [4], recognizing FTDFYC (33–38). A fifth antibody, Ma-10, has been characterized but its epitope remains unknown [5], [6].

This thesis designs and validates a computational workflow for epitope identification on CCR9, using three AI-based structure predictors and a multi-model consensus strategy implemented in Scipion. The four characterized antibodies are used to validate the workflow. Additionaly, a proposed epitope for Ma-10 is given being this the main contribution of this work.

1.2. Objectives

The main objective of this thesis is to design and validate a reproducible computational workflow for the identification of antibody-recognized epitopes on the CCR9 receptor. This goal breaks down into the following:

1. To review the experimental epitopes and variable heavy (VH) and light (VL) chain

sequences of the four characterized anti-CCR9 mAbs (91R, 92R, 115, and 3C3) reported in the literature.

2. To design and implement a reproducible computational pipeline using three AI-based structure predictors (AlphaFold, Boltz-1, and Chai-1) for generating the 3D complex models. Then, the Scipion-Chem platform is used for epitope prediction.

3. To apply this workflow to all five mAbs, with a deep analysis for 92R and Ma-10 and then a comparative overview for the other three: 91R, 115, and 3C3.

4. To propose a candidate binding epitope for Ma-10, the only antibody in this study without a published experimental characterization.

5. To evaluate the accuracy of the predictions by comparing them against experimentally reported epitopes and to assess the individual contribution of each predictor (AlphaFold, Boltz-1, and Chai-1) to the final output.

6. To document the methodological challenges encountered during the workflow development, including parameter selection and voting strategy decisions.

7. To evaluate the limitations of the current approach and propose a future scope based on the results obtained.

1.3. Methodological approach

The workflow follows three main steps. First, CCR9–single-chain variable fragment (scFv) complexes were generated using three independent AI-based predictors (AlphaFold, Boltz-1, and Chai-1), producing five models per predictor and 15 per antibody in total. Afterwards, these models were imported into Scipion and analyzed using Scipion-Chem. Specifically the *ProtDefineStructROI* protocol which identifies CCR9A residues at the antibody-antigen interface based on configurable distance and depth parameters, and *ProtROIVoting*, which combines these contacts across all 15 models to create an aminoacid-frequency score. Finally, four different result sets were obtained including the visualization in PyMOL to examine the predicted binding interface. These were compared with the experimentally reported epitopes. Full details of each step, including all parameters chosen, are provided in Chapter 3.

1.4. Key concepts and definitions

Antigen: Any molecule (typically a protein) that the adaptive immune system recognizes as foreign. In this thesis, the antigen is the extracellular domain of CCR9.

Antibody (immunoglobulin): A Y-shaped glycoprotein produced by B cells that specifically binds a target antigen with high precision. It consists of two heavy chains

and two light chains linked by disulfide bonds; the tips of the Y carry the variable regions that determine binding specificity.

Variable region (V-region): The region of an antibody that differs between antibodies and mediates antigen binding. The VH and VL regions together form the antigen-binding site.

Epitope: The specific region of an antigen bound by an antibody. Epitopes can be linear (a contiguous amino acid sequence) or conformational (residues that are close in 3D space but not necessarily adjacent in the sequence).

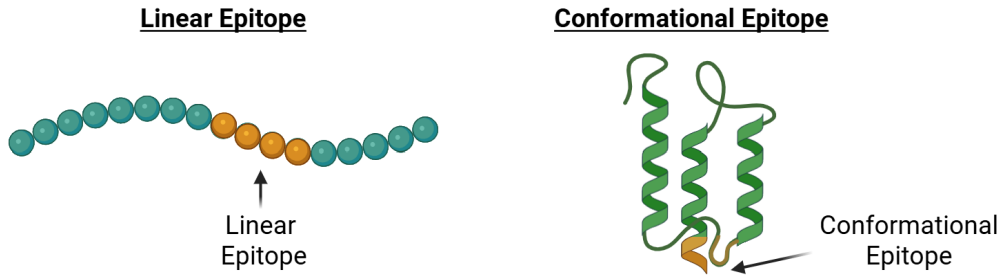


Fig. 1.1. Schematic comparison of linear and conformational epitopes. A linear epitope consists of a contiguous sequence of amino acid residues. A conformational epitope is formed by residues that are separated in the primary sequence but spatially close in the folded 3D structure. Own elaboration. Created with BioRender.com

Paratope: The region of the antibody, located in the variable domains, that physically contacts the epitope on the antigen.

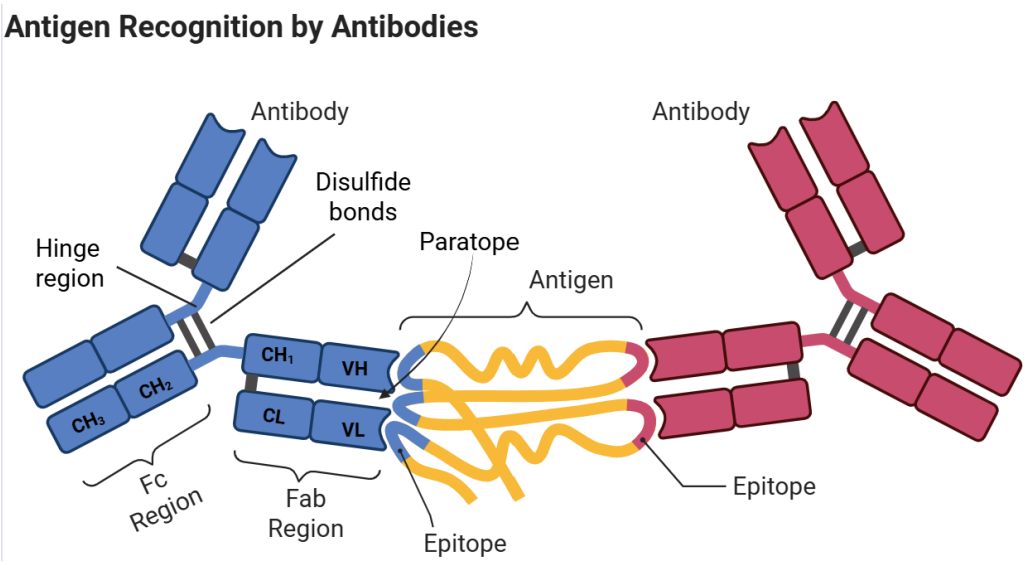


Fig. 1.2. Schematic representation of antigen recognition by an antibody. The Fab region contains the variable domains VH and VL, which together form the paratope that binds the epitope on the antigen. The Fc region contains the constant domains CH2 and CH3. Own elaboration. Created with BioRender.com

GPCR (G protein-coupled receptor): A large family of membrane receptors with seven transmembrane α-helices. Ligand binding on the extracellular side activates intracellular G proteins and downstream signaling. CCR9 belongs to this family.

Structural ROI (Region of Interest): A region of a protein structure selected for analysis based on geometric criteria. Here, ROIs identify residues at the antigen–antibody interface in a defined Ångström distance threshold.

RMSD (Root Mean Square Deviation): A standard metric for structural similarity between two protein models. It measures the average distance between equivalent atoms after optimal superposition. The lower the value, the more similar structures.

2. STATE OF THE ART

2.1. The CCR9 receptor

Chemokine receptors are a subfamily of G protein-coupled receptors (GPCRs) and their primary function is to direct the migration of leukocytes in response to chemotactic gradients. The GPCRs structure is conserved in the family and it consists of seven transmembrane α-helices connected by alternating extracellular and intracellular loops, an extracellular N-terminal domain and an intracellular C-terminal domain. When ligand binding occurs, GPCRs undergo conformational changes that activate intracellular heterotrimeric G proteins. This then triggers downstream signaling cascades that are involved in cell migration, proliferation and survival.

The CC chemokine receptor 9 (CCR9), originally identified and characterized as GPR-9-6 [4], is a member of this family that plays a key role in T cell trafficking. It is selectively expressed on intestinal homing T lymphocytes, mucosal lymphocytes and thymocytes. All of these migrate towards thymus-expressed chemokine (TECK, also known as CCL25) [4]. Beyond its physiological function, CCR9 has become a really relevant therapeutical target in T cell acute lymphoblastic leukemia (T-ALL), a hematologic malignancy where leukemic cells expressing CCR9 benefit from signaling from their receptor to maintain proliferation and survival [1], [3], [5].

In this thesis we have worked exclusively with the isoform CCR9A (UniProt accession: P51686-1), which has 369 amino acids and a calculated molecular mass of 42,016 Da [7]. The shorter isoform, CCR9B, does not include the first 12 amino acids of the N-terminal domain of CCR9A. It starts at the methionine corresponding to position 13 of the longer isoform.

Table 2.1. TOPOLOGICAL DOMAIN ORGANIZATION OF CCR9A ISOFORM (UniProt P51686-1, 369 aa).

Region	Residues	Topology
N-terminal domain	1–48	Extracellular
TM1	49–74	Transmembrane
ICL1	75–85	Intracellular
TM2	86–109	Transmembrane
ECL1	110–120	Extracellular
TM3	121–150	Transmembrane
ICL2	151–159	Intracellular
TM4	160–185	Transmembrane
ECL2	186–208	Extracellular
TM5	209–243	Transmembrane
ICL3	244–248	Intracellular
TM6	249–283	Transmembrane
ECL3	284–290	Extracellular
TM7	291–321	Transmembrane
C-terminal domain	322–369	Intracellular

The N-terminal extracellular domain (residues 1–48, pink) is the region targeted by the five antibodies studied in this thesis. Of note, there are minor differences in TM boundary between different sources (e.g., GPCRdb places the start of TM1 at residue 43) but all experimentally characterized epitopes (residues 11–38) fall within the N-terminal domain under any annotation. TM: transmembrane helix; ECL: extracellular loop; ICL: intracellular loop. Own elaboration based on UniProt [7].

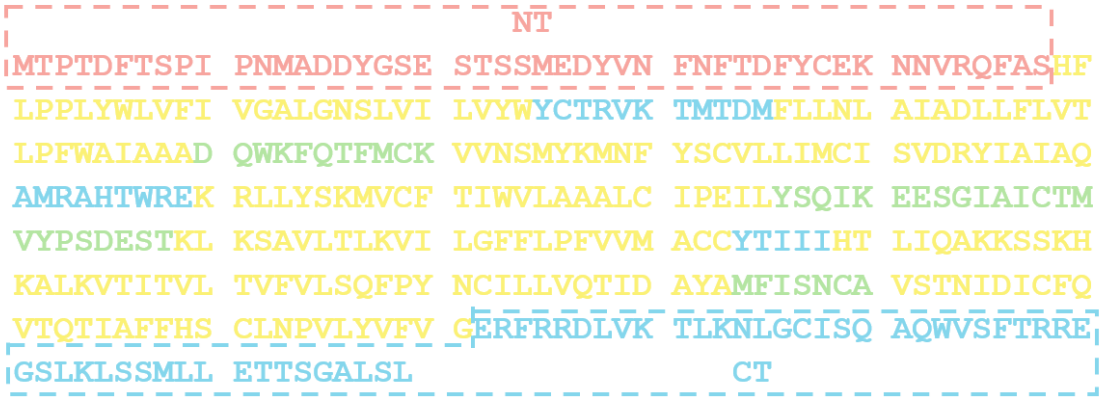


Fig. 2.1. Linear sequence of CCR9A isoform (369 aa) with topological domains with the same color-code and according to UniProt annotation [7]: N-terminal (NT) extracellular domain (pink, residues 1–48), transmembrane helices (yellow), extracellular loops (green) and intracellular regions including the C-terminal (CT) domain (blue). The first dashed boxes indicates the NT domain targeted by the antibodies studied in this thesis. The second one indicates the CT. Own elaboration.

The receptor contains seven transmembrane helices that traverse the plasma membrane connected by three extracellular loops (EC1, EC2, EC3) and three intracellular loops. The extracellular N-terminal domain (residues 1–48) is especially relevant from an immunological point of view because it constitutes the primary surface-exposed region accessible to antibody recognition without cell permeabilization needed. Its amino acid sequence is:

MTPTDFTSPIPNMADDYGSESTSSMEDYVNFNFTDFYCEKNNVRQFAS

In this 48-residue region, three linear epitopes have been experimentally mapped for the antibodies characterized in this thesis: **PNMADD** (residues 11–16), **MEDYVNF** (residues 25–31) and **FTDFYC** (residues 33–38). The spatial clustering of three non-overlapping epitopes in such a short domain shows the remarkable antigenic richness of the CCR9 N-terminus and its potential as a target for therapeutic antibody development.

All antibodies studied in this thesis bind epitopes located in the N-terminal domain. As expected for antibodies targeting membrane receptor surface epitopes, they do not need to enter the cell to bind their target.

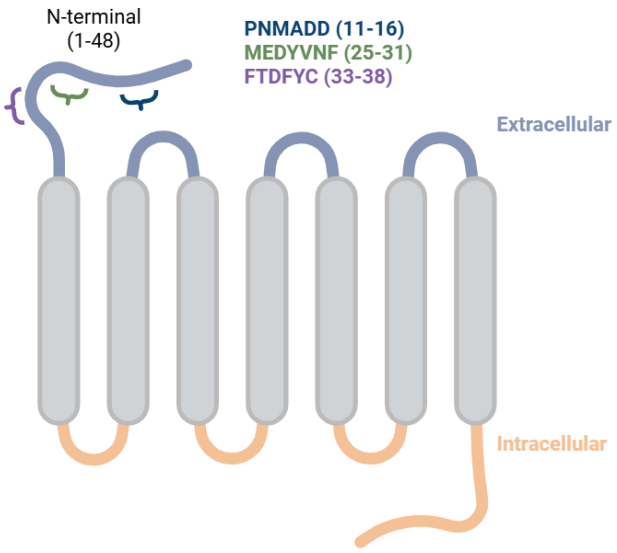


Fig. 2.2. Schematic representation of the CCR9A receptor showing the seven transmembrane domains, extracellular N-terminal domain (residues 1–48), and the three experimentally characterized epitopes: PNMADD (residues 11–16), MEDYVNF (25–31) and FTDFYC (33–38). Own elaboration. Figure created with BioRender [8].

2.2. Monoclonal antibodies against CCR9

So far, four of the monoclonal antibodies studied here against human CCR9 with experimentally characterized epitopes have been reported in the scientific literature: 91R, 92R, 115, and 3C3. A fifth antibody, Ma-10, has been described in a recently published patent, but its exact binding epitope on CCR9 has not been reported in the literature. This section describes the origin, characterization and structural properties of each antibody. Specifically, the epitope sequences and variable region data used in the computational analyses of this thesis.

2.2.1. mAbs 91R and 92R

The monoclonal antibodies 91R (IgG2b) and 92R (IgG2a) were developed by the Kremer L. group at the Centro Nacional de Biotecnología (CNB-CSIC, Madrid) through genetic immunization of mice [1], [2]. The immunization strategy consisted in the intramuscular administration of a mammalian expression plasmid that encoded the full-length human CCR9A sequence (residues 1–369), using a gene gun approach. This DNA immunization strategy was not an arbitrary choice. In fact, it was chosen because transmembrane proteins are difficult to work with using conventional protein immunization due to their seven transmembrane topology and highly hydrophobic nature. DNA immunization bypasses this whole problem.

It was found that both antibodies recognized the same linear epitope present in the

N-terminal extracellular domain of CCR9: the hexapeptide **PNMADD** corresponding to residues 11–16 of CCR9A. The epitope was determined using Pepscan analysis. In this analysis, 180 overlapping 12-amino acid peptides covering the complete CCR9A sequence were synthesized on a cellulose membrane and analyzed with each antibody independently. Four of the seven peptides spanning residues 1–24 obtained positive signals for both 91R and 92R mAbs, and the minimal consensus sequence present in all positive peptides was identified as PNMADD (aa 11–16) [2].

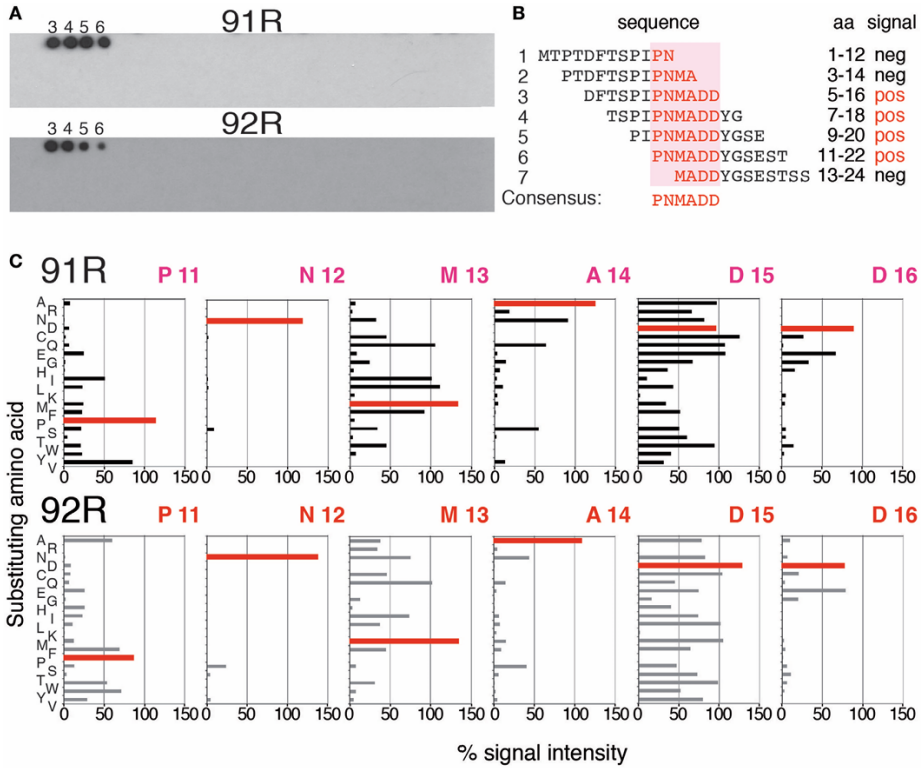


Fig. 2.3. Identification of the critical amino acid sequence recognized by 91R and 92R by Pepscan analysis. **(A)** Pepscan membrane showing positive signals for peptides 3–6 for both antibodies. **(B)** Peptide sequences and binding signals. The minimal consensus sequence common to all positive peptides is PNMADD (aa 11–16). **(C)** Substitution analysis in which each position of PNMADD was individually replaced by all other 19 amino acids. For both antibodies, N12 is the single most critical residue: no substitution at this position is tolerated. Reproduced from [2].

A further characterization by substitution Pepscan in which each amino acid situated in the epitope was individually replaced by each of the other 19 standard amino acids. This revealed that residues N12 and A14 are the most critical for high-affinity binding, because their substitution produces a remarkable reduction in signal intensity. This finding provides structural insight into the binding interface and highlights the specificity of the antibody-epitope interaction [2].

Although they bind to the same epitope and have same chain length, 91R and 92R differ in their variable region sequences. Sequence alignment of their VH and VL chains using the Clustal program reveals differences both in the framework regions and in the

complementarity-determining regions (CDRs), particularly in CDR1 and CDR3 of the heavy chain and CDR1 of the light chain [2]. The VH chains are 117 amino acids long and the VL (kappa) chains are 113 amino acids long in both cases. Full sequences are provided in Appendix B.

Binding kinetics were characterized using surface plasmon resonance (SPR), with a synthetic peptide as antigen, that corresponded to CCR9A residues 2–22 immobilized on a CM5 biosensor chip. Both antibodies bind with high apparent affinity in the low nanomolar range, tested at concentrations from 0.41 to 33.3 nM. Competitive SPR experiments confirmed that the epitope region recognized by both antibodies overlaps with residues 10–30 of CCR9A and that peptides spanning residues 2–22 effectively block CCR9-antibody binding [2].

The antitumor potential of both antibodies has been demonstrated *in vivo* using xenograft mouse models. Chamorro et al. [1] showed that 91R significantly inhibits the growth of CCR9-expressing human leukemia cells in immunodeficient mice, while Somovilla-Crespo et al. [2] confirmed similar antitumor effects for 92R in NSG mouse xenograft models.

For the computational studies in this thesis, the amino acid sequences of the variable regions of both antibodies were used to construct single-chain variable fragments (scFv) in the VL–linker–VH orientation, using the flexible $(G_4S)_3$ linker sequence (GGGGSGGGGSGGGGS). The scFv of 92R was provided by the Kremer group and served as the reference construction format.

2.2.2. mAb 115

mAb 115 is a mouse monoclonal antibody of the IgG1 isotype (clone #115) developed by the Menéndez P. group at the Institut de Recerca contra la Leucèmia Josep Carreras (Barcelona). It was characterized in the context of a CAR-T cell therapy study targeting CCR9 and CD1a for the treatment of T-ALL [3].

The epitope of mAb 115 was determined by Pepscan analysis of the N-terminal domain of CCR9A (aa 1–48), using 16 overlapping peptides approximately 10 amino acids in length. Peptides 9 through 11, covering residues 22–34, gave positive signals and the minimal consensus sequence was identified as **MEDYVNF**, corresponding to residues 25–31 of CCR9A [3].

The VH chain of 115 has 112 amino acids and the VL (kappa) chain is 113 amino acids long. Full sequences are provided in Appendix B.

Notably, Tirado et al. also describe the humanization of 115 for reducing its immunogenicity for potential clinical use, generating two humanized variants (CCR9_H1 and CCR9_H2). This is done by grafting the CDR regions onto human germline frameworks IGHV1-3*01 (VH) and IGKV2D-29*02 (VL) [3]. The fact that humanization was carried out shows the therapeutic interest in 115 and CCR9-targeting antibodies in general.

2.2.3. mAb 3C3

mAb 3C3 is a mouse monoclonal antibody originally described by Zabel et al. in 1999 as part of the molecular characterization of CCR9 (back then known as GPR-9-6) as a chemokine receptor selectively expressed on intestinal homing T lymphocytes [4]. It was developed at LeukoSite Inc. (Cambridge, MA), a biotechnology company that was later acquired by Millennium Pharmaceuticals and afterwards by Takeda Pharmaceutical.

The epitope recognized by 3C3 is located close to the end of the N-terminal extracellular domain: **FTDFYC**, corresponding to residues 33–38 of CCR9A [4]. This epitope contains cysteine involved in potential disulfide bridge formation in the N-terminal region, which may be relevant to the conformational accessibility of this binding site.

The VH chain of 3C3 is 117 amino acids long and the VL (kappa) chain has 112 amino acids. Full sequences are provided in Appendix B.

mAb 3C3 was also used in the original Zabel et al. study to demonstrate cell-surface expression of CCR9 on non-permeabilized lymphocytes, confirming that its epitope is genuinely extracellular and accessible without membrane disruption [4].

2.2.4. mAb Ma-10

The fifth antibody included in this bachelor thesis, referred to as Ma-10, is an anti-human CCR9 antibody described by Maciocia et al. in the context of anti-CCR9 chimeric antigen receptor (CAR-T) cell therapy for T-ALL [5]. The variable region amino acid sequences were extracted from a US patent application filed by the same group (US Patent Application #20240374727, issued November 2024) [6].

Unlike the other four antibodies in this thesis, the exact amino acid sequence of the epitope recognized by Ma-10 on CCR9 has not been reported in any published study or patent document. The patent states only that *"the CCR9 binding domain may selectively bind to an epitope within the N-terminal domain of CCR9"* [6], without further specification of the binding region. This constitutes the central challenge and the principal contribution of this thesis: to use a computational workflow to propose a candidate binding epitope for Ma-10, based on structural modeling of the CCR9–Ma-10 complex and residue contact analysis.

The VH and VL sequences of Ma-10 were obtained from the patent. During the analysis, some errors in the sequence were detected. Specifically, a tryptophan (W) in the VL that was an artifact of PDF and should be two valines (VV) and a discrepancy in the VL CDR1 region. The corrected sequences were checked with the sequences provided in the published Blood 2022 paper [5]. The scFv of Ma-10 was constructed in the VH–linker–VL orientation, as specified in the patent, using the $(G_4S)_3$ linker. Full sequences are provided in Appendix B.

2.2.5. Summary of epitopes and antibody properties

Table 2.2 provides a comparative overview of the five antibodies studied in this thesis. All experimentally characterized epitopes map to the 48-residue extracellular N-terminal domain of CCR9A covering three different and non-overlapping linear regions:

MTPTDFTSP-I-**PNMADD**-YGSESTSS-**MEDYVNF**-N-**FTDFYC**-EKNNVRQFAS

Table 2.2. SUMMARY OF THE MAIN CHARACTERISTICS OF THE FIVE MONOCLONAL ANTIBODIES AGAINST CCR9 STUDIED IN THIS THESIS.

mAb	Isotype	Origin	Epitope (aa)	Sequence	Reference
91R	IgG2b (mouse)	Kremer L., CNB-CSIC	11–16	PNMADD	[1]
92R	IgG2a (mouse)	Kremer L., CNB-CSIC	11–16	PNMADD	[2]
115	IgG1 (mouse)	Menéndez P., IJC	25–31	MEDYVNF	[3]
3C3	mouse	LeukoSite Inc.	33–38	FTDFYC	[4]
Ma-10	mouse	Maciocia, UCL	unknown	—	[5], [6]

Own elaboration.

2.3. Computational epitope prediction: state of the art

The identification of antibody-recognized epitopes is a central challenge in structural immunology with direct implications for vaccine design, the development of therapeutic antibodies and the understanding of immune responses.The experimental techniques such as Pepscan, hydrogen-deuterium exchange mass spectrometry (HDX-MS), mutagenesis studies and X-ray crystallography or cryo-EM of antibody-antigen complexes still remain as the standard reference for epitope mapping but nowadays they are very work demanding, expensive and not always possible. This is especially the case for membrane proteins such as GPCRs, whose structural characterization presents well-known technical difficulties.

Computational approaches to epitope prediction have therefore received growing attention with the goal of identifying and prioritizing candidate binding regions before committing to expensive experimental validation. These methods can be classified into two main categories: *sequence-based* methods (targeting linear epitopes) and *structure-based* methods (targeting conformational epitopes) as illustrated in (Figure 1.1).

Sequence-based methods analyze the primary amino acid sequence of the antigen to identify regions with physicochemical properties associated with antigenicity, such as high solvent exposure, hydrophilicity, flexibility and propensity of forming turns. Classical tools include BepiPred and its neural network-based successors, which have been trained on datasets of known linear epitopes. They are computationally efficient and broadly accessible, however, these methods are fundamentally limited to linear epitopes and cannot capture the three-dimensional context of the antibody-antigen interface.

On the other hand, structure-based methods benefits from the three-dimensional protein structures to identify surface-exposed areas with geometric and physicochemical complementarity with the antibody paratopes. These approaches include tools based on solvent accessibility calculations, electrostatic potential surface analysis and docking simulations. Nonetheless, their applicability has historically been constrained because, as mentioned before, obtaining the 3D protein structure is really expensive and complicated.

The release of AlphaFold2 [9] in 2021 gave an incredible shift in structural biology, making possible the accurate prediction of protein three-dimensional structures directly from amino acid sequences with unprecedented accuracy. It extended to multimeric and antibody-antigen complex prediction initially through AlphaFold-Multimer and more recently through AlphaFold3 via the AlphaFold Server [10] has greatly expanded the scope of computational epitope analysis. Additionally, alternative structure prediction tools have appeared that also support multimeric complex prediction for antibody-antigen systems. Among them, Boltz-1 [11] and Chai-1 [12] provide complementary structural hypotheses that, when integrated in a multi-model consensus framework, can increase prediction robustness. All these three predictors are used in this thesis.

The use of multiple structural models generated by different prediction algorithms or with different random seeds in a consensus is particularly powerful because it reduces the impact of individual model errors and biases and highlights interface residues that are consistently predicted across models. This multi-model integration approach is implemented in this study in the Scipion-Chem platform [13] through the *ProtROIVoting* protocol, which aggregates residue contact frequencies across all input models and produces a consensus prediction of the antibody-antigen interface. This protocol is a fundamental step in the computational workflow developed in this thesis.

The evaluation and comparison of computational methods for antibody-antigen interface prediction remains an active research area. Recent studies have proposed new evaluation metrics such as the interface intersection-over-union (AbAgIoU), which quantifies the overlap between predicted and experimentally observed interface residues. They have compared the performance of different structure predictors and scoring functions across large datasets of antibody-antigen complexes.

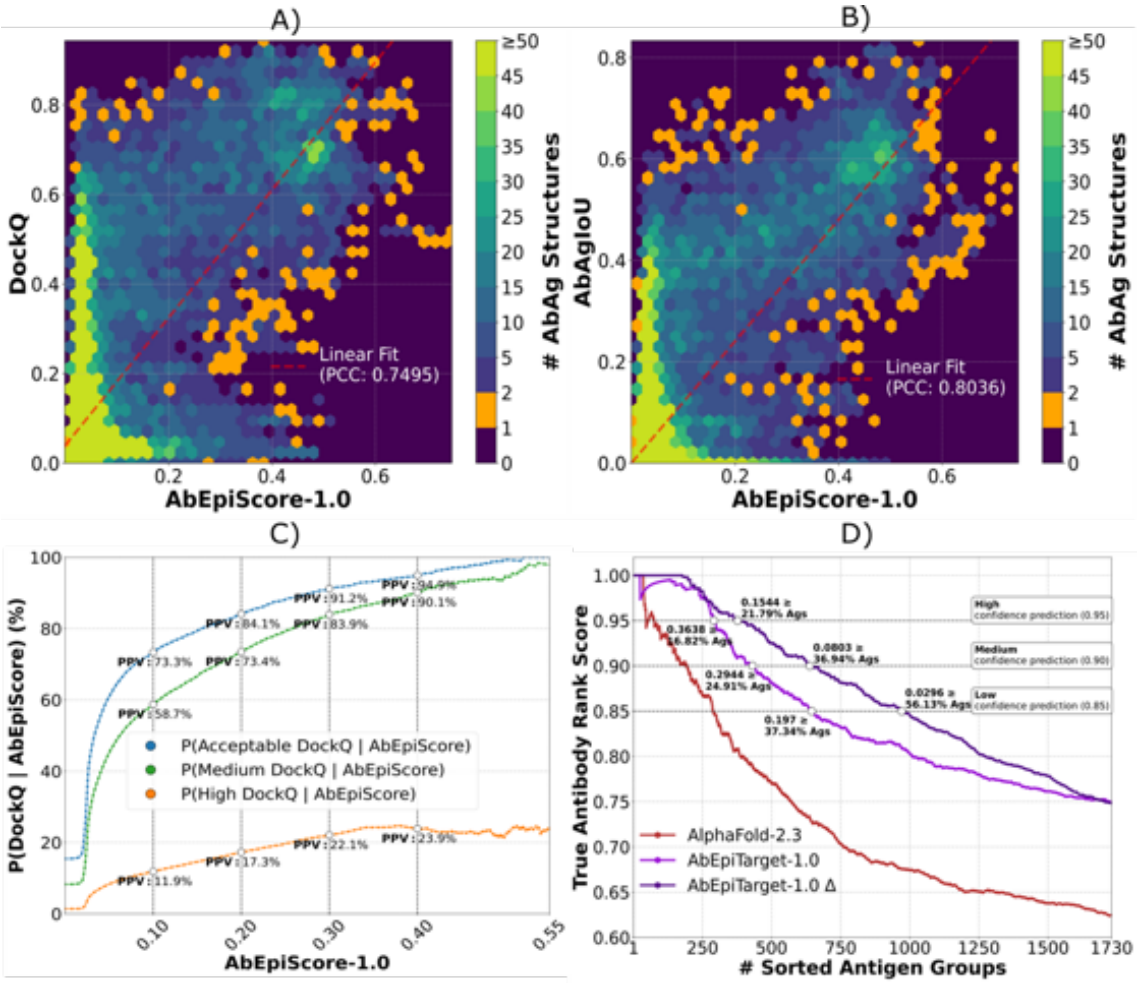


Fig. 2.4. Performance of AbEpiTope-1.0 on antibody-antigen interface prediction. Left: correlation between AbEpiScore-1.0 and DockQ scores across antibody-antigen structures (PCC = 0.7495). Right: correlation between AbEpiScore-1.0 and AbAgIoU (PCC = 0.8036). Bottom left: probability of acceptable, medium, and high DockQ given AbEpiScore-1.0. Bottom right: True Antibody Rank Score for AlphaFold-2.3 and AbEpiTarget-1.0 across antigen groups. Adapted from Clifford et al. [14].

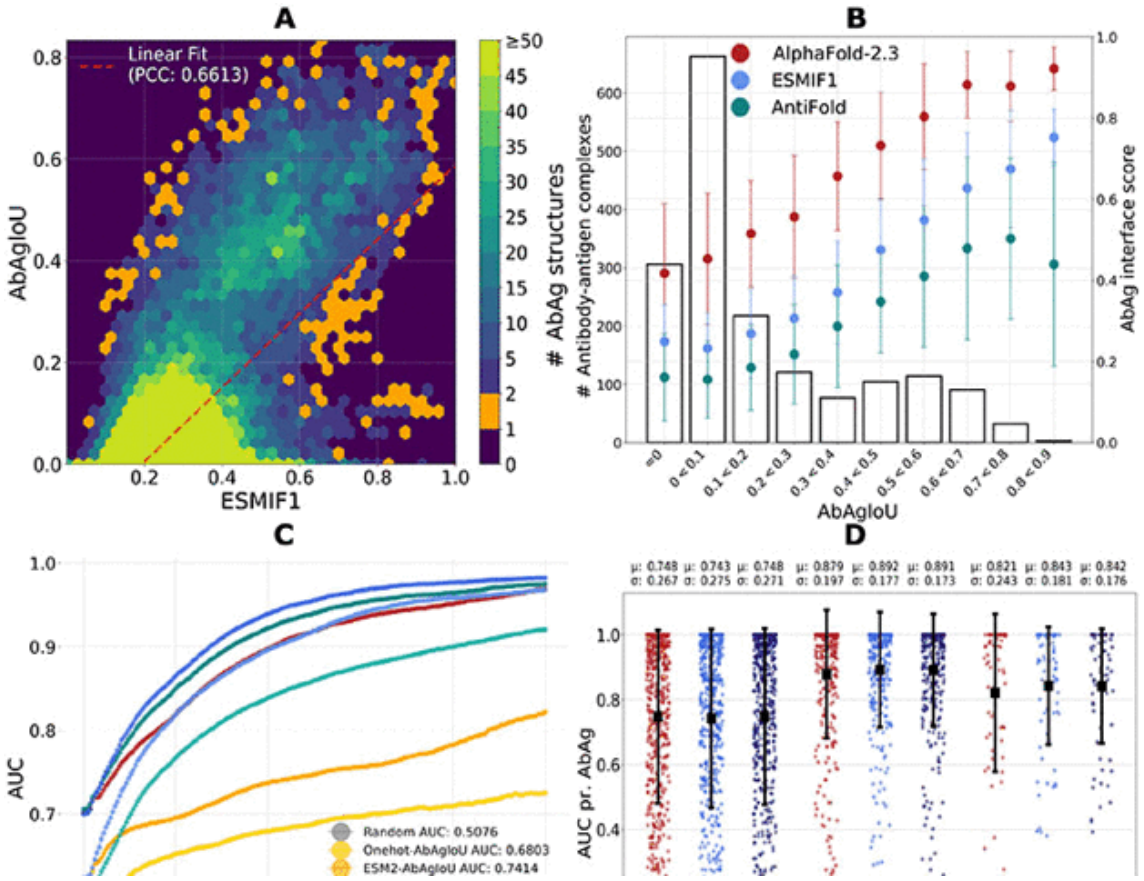


Fig. 2.5. Benchmarking of structure-based interface scoring methods on antibody-antigen complexes. (A) Correlation between ESMIF1 scores and AbAgIoU (PCC = 0.6613). (B) Comparison of AlphaFold-2.3, ESMIF1, and AntiFold interface scores across AbAgIoU bins. (C) AUC curves for random, one-hot, and ESM2-based AbAgIoU prediction models. (D) Per-antigen AUC distributions for antibody-antigen pair prediction. Adapted from Clifford et al. [14].

These benchmarking efforts highlight both the progress achieved in recent years and the challenges that still remain unsolved. In particular, for antigens with flexible or disordered regions such as the N-terminal domain of CCR9.

The workflow developed in this thesis does not rely on a single prediction tool. Instead, as mentioned before, it integrates the structural outputs of three independent predictors (AlphaFold, Boltz, and Chai-1) within a standardized analysis framework (ScipionChem), with the goal of benefiting from the strength of complementing each tool while also mitigating and minimizing individual limitations. The specific parameters and protocols used are described in detail in Chapter 3.

3. MATERIALS AND METHODS

This chapter describes the complete computational workflow developed in this thesis for the identification of antibody-recognized epitopes on the human CCR9 receptor. The workflow follows four main stages: (1) Obtaining antibody variable region sequences and construction of single-chain variable fragments (scFv); (2) 3D structure prediction of the CCR9–scFv complexes using three independent AI-based predictors; (3) Identification and integration of antibody–antigen interface residues using the Scipion-Chem platform; (4) Structural visualization and analysis of the results. The full workflow is schematically summarized in Figure 3.1.

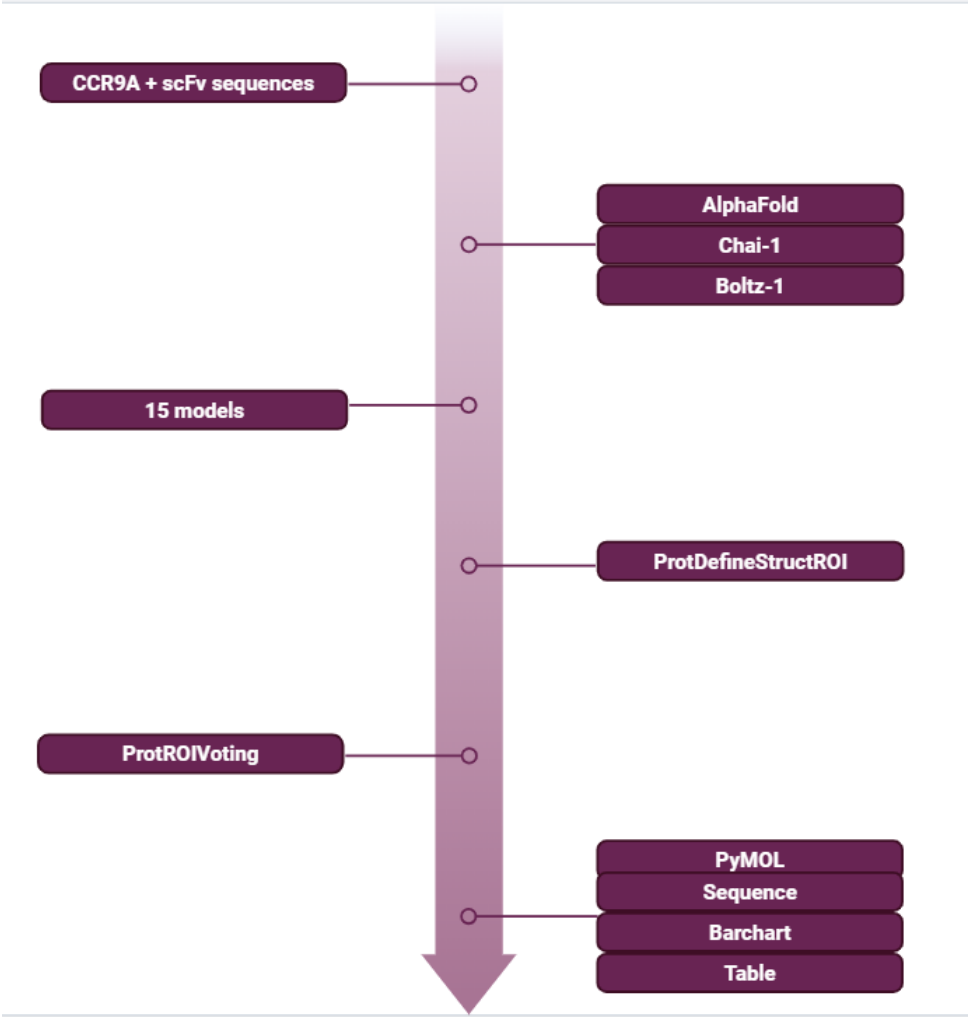


Fig. 3.1. Schematic overview of the computational workflow developed in this thesis. Starting from CCR9A and scFv sequences, three AI-based structure predictors (AlphaFold, Boltz-1, and Chai-1) generate 15 structural models per antibody. These are analyzed using the Scipion-Chem plugin (*ProtDefineStructROI* and *ProtROIVoting*) to produce contact frequency scores, visualized in PyMOL. Own elaboration. Figure created with BioRender [8].

3.1. Computational environment

All computational analyses were performed on a personal laptop running Windows 11 (Intel Core Ultra 7 155H, 16 GB RAM) with Ubuntu 22.04 via Windows Subsystem for Linux 2 (WSL2). Structure predictions were generated through web-based servers: the AlphaFold Server (https://alphafoldserver.com) for AlphaFold and the Tamarin.bio platform (https://tamarin.bio) for both Boltz-1 and Chai-1. Structural analysis and consensus integration were performed using Scipion 3.8.1 with the pwchem plugin v1.0.1 running under Python 3.8.12. Three-dimensional structural visualization was carried out in PyMOL 2.5.0 (Open-Source) launched directly from the Scipion environment.

The predicted structures were imported into Scipion using the *Import Structure Predictions* protocol. This import workflow was developed by a collaborating student (B. Pueche) as part of the Scipion-chem biofold plugin, and allowed the batch import of all five models from each predictor simultaneously from the downloaded compressed archive files, replacing a manual one-by-one import approach used initially.

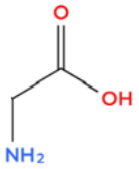
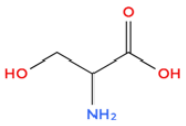
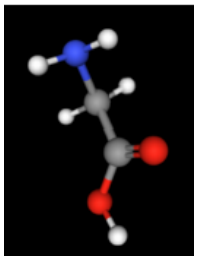
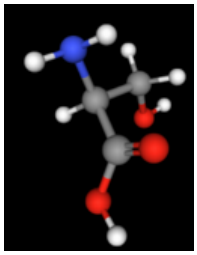
The *ProtDefineStructROI* protocol, available in the pwchem plugin, was used for interface residue identification. The *ProtROIVoting* protocol, developed as part of this thesis and contributed to the pwchem plugin, was used for multi-model consensus integration. The VH and VL sequences of all antibodies were obtained from published sources and also provided by Dr. Leonor Kremer (CNB-CSIC). All scFv sequences were assembled manually by concatenating the VH and VL sequences with the $(G_4S)_3$ linker in the appropriate orientation.

3.2. Antibody sequences and scFv construction

The variable region sequences of the five anti-CCR9 monoclonal antibodies (91R, 92R, 115, 3C3, and Ma-10) were obtained from the published literature and patent sources described in Chapter 2. For each antibody, only the variable heavy chain (VH) and variable light chain (VL) sequences were used. Constant regions and signal peptides were excluded as these do not contribute to antigen binding and are not part of the mature antigen-binding fragment.

For the creation of the structure prediction of the antibody–antigen complex, each antibody was reformatted as a single-chain variable fragment (scFv). An scFv is a fusion protein consisting of the VH and VL domains connected by a flexible linker peptide which allows the two domains to fold independently and still maintain antigen-binding activity as a single polypeptide chain. The $(G_4S)_3$ linker (sequence: GGGGSGGGGSGGGGS) was used in all cases. This is because glycine (G) and serine (S) are particularly well-suited for linker design: glycine lacks a side chain, giving the backbone maximum rotational freedom, while serine adds mild hydrophilicity that keeps the linker soluble. Together they produce a flexible, non-immunogenic spacer.

Table 3.1. STRUCTURAL REPRESENTATION OF GLYCINE (G) AND SERINE (S), THE TWO AMINO ACIDS COMPRISING THE $(G_4S)_3$ FLEXIBLE LINKER.

	Glycine (G)	**Serine (S)**
2D structural formula		
3D ball-and-stick model		

Glycine, the smallest amino acid, lacks a side chain providing maximum backbone flexibility. Serine contributes mild hydrophilicity that keeps the linker soluble. Together they form a flexible, non-immunogenic spacer suitable for scFv construction. Own elaboration.

The orientation of the scFv (the order of the VH and VL domains) was determined based on available experimental information for each antibody. For 92R, the VL–linker–VH orientation was used, as specified in the published scFv construct provided by the Kremer group [2]. On the other hand, for Ma-10, the VH–linker–VL orientation was used, as indicated in the patent sequence listing [6]. In theory, both scFv conformations should be considered functionally equivalent when using a sufficiently long flexible linker like the one used in this study $(G_4S)_3$. The complete scFv sequences used in this thesis are provided in Appendix B.

As stated previously, the target antigen used in all structure predictions was the full-length CCR9A isoform (UniProt accession P51686-1, 369 amino acids) [7].

3.3. Three-dimensional structure prediction of CCR9–scFv complexes

For each of the five antibodies, the three-dimensional structure of the CCR9–scFv complex was predicted using three independent AI-based structure prediction tools: the AlphaFold Server [10], Boltz-1 [11], [15] and Chai-1 [12], [15]. Each predictor received as input the two polypeptide sequences forming the complex: the full-length CCR9A sequence and the corresponding scFv sequence as separate chains.

No additional constraints, templates or prior structural information were provided to any of the predictors. Each tool was used with its default parameters for multimeric complex prediction. Each predictor generates five ranked structural models per run, ranked

by an internal confidence metric (predicted local distance difference test, pLDDT, for AlphaFold; equivalent metrics for Boltz-1 and Chai-1). All five models from each predictor were saved for the future analysis, resulting in a total of 15 structural models per antibody (5 models × 3 predictors).

The outputs of each prediction run were downloaded as compressed archive files (.zip) containing the predicted PDB coordinate files.These files were used directly as input to the Scipion-Chem analysis platform described in the following section.

3.3.1. AlphaFold Server

The AlphaFold Server (https://alphafoldserver.com) provides access to AlphaFold's multimeric complex prediction capabilities through a web interface [10]. Once the two protein sequences were given as inputs, the server returned five predicted models per submission, ranked by the predicted template modeling score (pTM) and interface pTM (ipTM).

3.3.2. Boltz-1

Boltz-1 is an open-source biomolecular structure prediction model that supports the prediction of protein–protein complexes [15]. Predictions were submitted using the Boltz-1 online interface, providing the two protein sequences as input. Five structural models were obtained per antibody, ranked by the model's internal confidence scoring.

3.3.3. Chai-1

Chai-1 is a multi-modal foundation model for molecular structure prediction that supports protein–protein complex modeling [15]. Predictions were submitted through the Chai-1 web server, providing CCR9A and the scFv sequences as input chains. Five structural models were obtained per antibody.

3.4. Interface residue identification and multi-model consensus: Scipion-Chem

The structural models generated by the three predictors were analyzed using the Scipion-Chem plugin within the Scipion framework [13]. Scipion is a workflow-oriented platform for structural biology that enables the integration and standardized analysis of outputs from multiple software tools in a reproducible environment. The Scipion-Chem plugin extends this framework with tools for the analysis of protein–ligand and protein–protein interfaces, including the identification of binding site residues and multi-model consensus analysis.

The analysis pipeline within Scipion-Chem consisted of two main protocols applied sequentially: *ProtDefineStructROI* for interface residue identification and *ProtROIVoting* for multi-model consensus integration.

3.4.1. Import of structural predictions

The 15 structural models per antibody (organized as three sets of five models from AlphaFold, Boltz-1 and Chai-1 respectively) were imported into Scipion using three independent *Import Structure Predictions* protocols (one per predictor). Each import protocol received the corresponding compressed archive file (.zip) as input and generated a set of AtomStruct objects within the Scipion project, one per model.

3.4.2. ProtDefineStructROI: interface residue identification

The *ProtDefineStructROI* protocol was applied to each of the 15 structural models individually, resulting in 15 independent protocol executions per antibody. This protocol identifies the residues of a defined chain (chain of interest) that lie within a specified distance threshold of a second interacting chain, defining a Region of Interest (ROI) corresponding to the protein–protein interface.

The following parameters were used consistently across all executions:

- **Extract ROIs from:** Protein-Protein Interface
- **Chain of interest:** Chain A (CCR9A, the antigen)
- **Chain of interest 2:** Chain B (the scFv, the antibody)
- **Interface distance threshold:** 3.0 Å
- **Maximum distance between pocket points:** 2.0 Å
- **Map coordinates to surface:** Yes
- **Maximum atom depth:** 3.0 Å

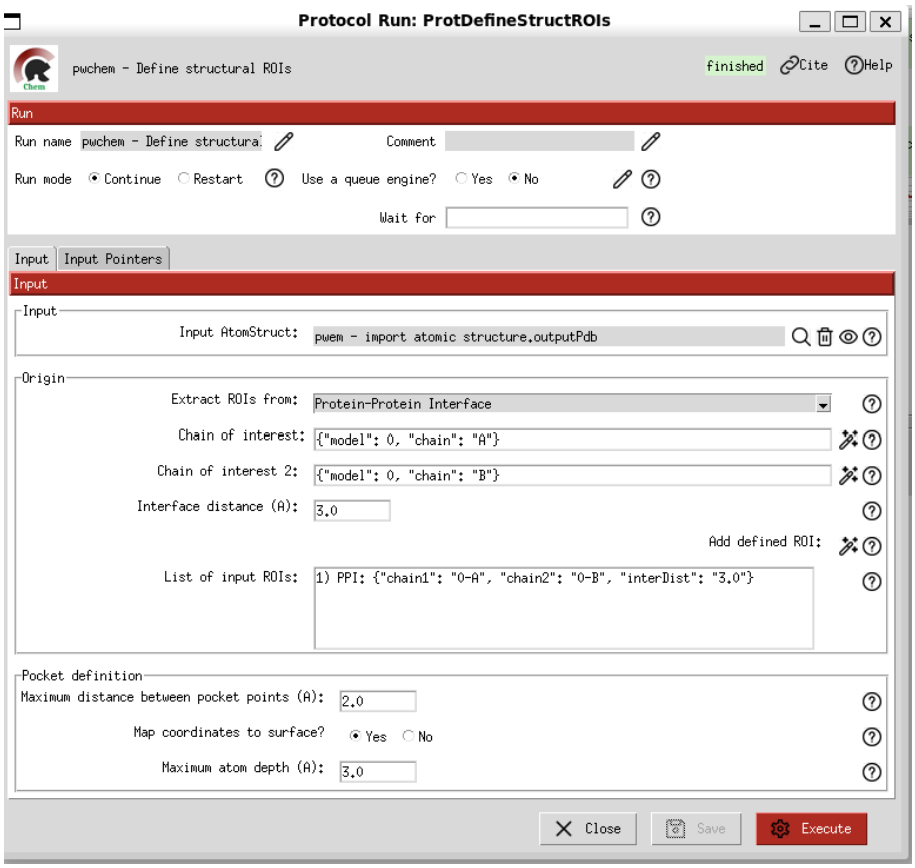


Fig. 3.2. Parameter configuration of the *ProtDefineStructROI* protocol in Scipion-Chem for a representative model. Chain A (CCR9A) is defined as the chain of interest and Chain B (scFv) as the interacting chain. Interface distance threshold: 3.0 Å; maximum distance between pocket points: 2.0 Å; maximum atom depth: 3.0 Å. These settings were kept identical across all 15 executions per antibody. Own elaboration.

The interface distance threshold of 3.0 Å defines the maximum distance between any atom of a CCR9A residue and any atom of the scFv for that CCR9A residue to be considered part of the binding interface. This threshold is consistent with the typical range used in the structural biology literature to define direct inter-residue contacts at protein–protein interfaces. The output of each *ProtDefineStructROI* execution is a list of contacts from both chain A (CCR9A) and chain B (scFv). However, only the CCR9 residues are relevant for the epitope identification and are the only ones passed as input to the next protocol. Chain B contacts are ignored.

3.4.3. ProtROIVoting: multi-model consensus integration

The 15 ROI outputs generated by the *ProtDefineStructROI* executions were integrated using the *ProtROIVoting* protocol. This protocol implements a multi-model consensus strategy: for each residue of CCR9A, it counts in how many of the 15 structural models that residue was identified as an interface contact. The result is a frequency score for each residue, ranging from 0 (not identified as a contact in any model) to 15 (identified as a

contact in all 15 models).

This consensus approach is designed to reduce the impact of model-specific errors or structural biases introduced by individual predictors. Residues that are consistently identified as interface contacts across multiple models and multiple predictors are considered much more reliable candidates for the binding epitope than residues identified in only one or a few models.

The *ProtROIVoting* protocol generates the following outputs for each antibody:

- A **frequency table** listing each CCR9A residue and its associated contact frequency across the 15 models. There is also a column showing the percentage. This one is normalized to the most frequent residue, meaning 100% corresponds to the highest frequency rather than to a 15/15 score.
- A **PyMOL session with a white-to-red color gradient**, in which each CCR9A residue is colored according to its frequency score (white = frequency 0; dark red = maximum frequency).
- A **bar chart** displaying contact frequency as a function of CCR9A residue position (x-axis: residue position 1–369; y-axis: contact frequency 0–15).
- A **sequence view** with residues colored according to the same red gradient as before, allowing direct identification of high-frequency contact regions along the 369 amino acid sequence.

This complete pipeline was applied independently to each of the five anti-CCR9 antibodies studied in this thesis. Figures 3.3 and 3.4 show the Scipion project view and summary panel for a representative run (Ma-10).

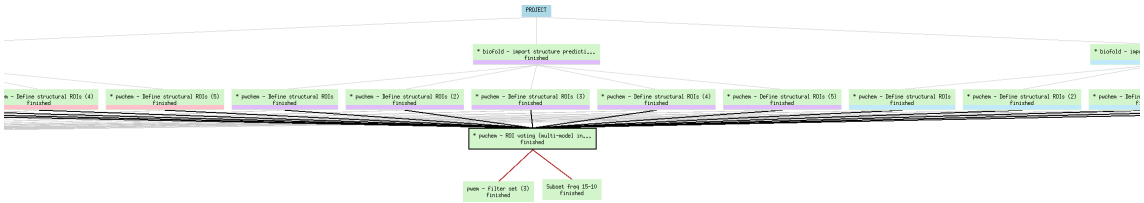


Fig. 3.3. Scipion project view of the analysis workflow. The first row shows the *Import Structure predictions* (3 of them). The second row shows the *ProtDefineStructROI* executions (5 per each). All 15 connect to the central *ProtROIVoting* protocol. Not all 15 runs are fully visible at this zoom level. Filters or subset selection can be then optionally applied in the final step. Own elaboration.

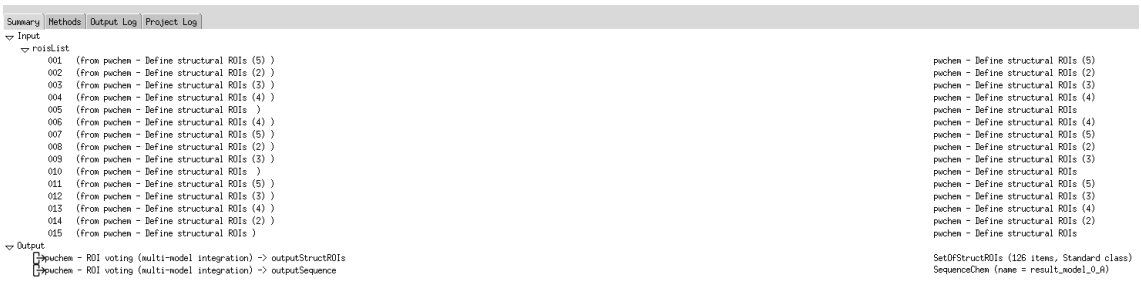


Fig. 3.4. Summary panel of the *ProtROIVoting* protocol, listing all 15 input runs (001–015) and the output: a set of structural ROIs showing the amino acid frequency scores. Own elaboration.

3.5. Structural visualization

All structural figures in Chapter 4 were generated from the *ProtROIVoting* output. The protocol produces as as described in Section 3.4.3 several visualization outputs all in the Scipion environment. Three-dimensional structural figures were generated in PyMOL [16], launched directly from the Scipion environment. The sequence was displayed in an interactive HTML also launched by the Scipion environment.

4. RESULTS

This chapter presents the results of applying the computational workflow described in Chapter 3 to the five anti-CCR9 monoclonal antibodies studied in this thesis. For each antibody, the workflow produced 15 predicted CCR9–scFv complex structures (5 models per predictor $\times$ 3 predictors) and a multi-model consensus analysis of interface contact frequencies across the full CCR9A sequence (369 residues). Contact frequency scores are reported as absolute counts out of 15 models; percentages are given relative to the maximum frequency obtained.

The results are organized as follows. Section 4.1 summarizes the results for antibodies 91R, 115, and 3C3, whose epitopes are experimentally known and serve as a validation set; full figures for these antibodies are provided in Appendix A. The other two antibodies are studied rigorously. Section 4.2 provides an in-depth analysis of antibody 92R. Section 4.3 presents the results for Ma-10, where the computational prediction constitutes the main original contribution of this work. Section 4.4 offers a comparative overview across all five antibodies.

One observation applies to all five cases without exception: the high-frequency contact residues are concentrated in the extracellular N-terminal domain of CCR9A (residues 1–48) with no substantial signal in the transmembrane helices or intracellular regions. This is precisely what one would expect for antibodies that recognize cell-surface epitopes on intact, non-permeabilized cells and it provides an initial global validation that the workflow is correctly localizing the antibody-binding surface of the receptor. As described in Section 3.3, no prior biological information was provided to any of the predictors: the full-length CCR9A sequence (369 amino acids) was used as input without restricting the search space in any way. The fact that the workflow consistently recovers signal in the N-terminal domain under these unbiased conditions strengthens the validity of the approach.

4.1. Validation results: antibodies 91R, 115, and 3C3

Antibodies 91R, 115, and 3C3 were used as a validation set because their epitopes on CCR9A have been experimentally determined by Pepscan analysis (Chapter 2). The complete set of figures and frequency tables for these three antibodies is provided in Appendix A. Here we summarize the key findings and compare the computational results with the experimental data.

Antibody 91R. The workflow identified 117 contact residues, with a maximum frequency of 12/15 at F31. Three residues located in the experimental epitope PNMADD (11–16): 'A14, D15, and D16' appear at frequencies of 9, 10 and 9 respectively, confirming recovery of the known binding region. Furthermore, the predicted window (residues

14–37) extends into regions overlapping with the experimentally characterized epitopes of other antibodies in this study: MEDYVNF (residues 25–31, 115) and FTDFYC (residues 33–38, 3C3).This suggests that the workflow is indeed capturing genuine surface-exposed contacts.

Antibody 115. The workflow identified 111 contact residues, with a maximum frequency of 14/15 at N30. Every residue in the experimental epitope MEDYVNF (25–31) appears in the high-frequency table (≥8) being this great results.

Antibody 3C3. The workflow identified 125 contact residues, with a maximum frequency of 12/15 at both F31 and N32. Residues F33 (11/15) and T34 (8/15) confirm recovery of the experimental epitope FTDFYC (33–38). A secondary peak at T190 (10/15) reflects the spatial proximity of ECL2 in the predicted 3D models; the structural context of these contacts can be seen clearly illustrated in Figure A.7 (see Appendix).

Table 4.1. COMPARISON BETWEEN EXPERIMENTALLY DETERMINED EPITOPES AND COMPUTATIONALLY PREDICTED HIGH-FREQUENCY CONTACT REGIONS FOR THE THREE VALIDATION ANTIBODIES.

mAb	Experimental epitope	Cutoff	Epitope residues recovered	Max freq.
91R	PNMADD (11–16)	≥8/15	A14, D15, D16 (3/6)	12/15
115	MEDYVNF (25–31)	≥8/15	M25–F31 (7/7)	14/15
3C3	FTDFYC (33–38)	≥8/15	F33, T34 (2/6)	12/15

Own elaboration.

Across all three antibodies, the workflow correctly places the binding surface in the N-terminal extracellular domain with barely false positives in transmembrane or intracellular regions. This pattern is discussed further in Chapter 5.

4.2. In-depth analysis: antibody 92R

Antibody 92R shares its experimental epitope with 91R: PNMADD (residues 11–16), with N12 and A14 identified as the most critical positions [2]. They have, however, different variable region sequences, particularly in CDR1 and CDR3 of the heavy chain and CDR1 of the light chain. It served as the primary reference antibody for the in-depth analysis, as its scFv construct was provided directly by the Kremer group in a validated orientation (VL–linker–VH).

4.2.1. Interface residue frequency analysis

The workflow identified 122 contact residues across the full CCR9A sequence, with a maximum frequency of 11 out of 15 models at residues F31 and F33. A threshold of ≥6 was chosen to include N12 and P11, the two residues experimentally identified as most critical for 92R binding [2]. Table 4.2 lists all residues at or above this threshold.

Table 4.2. CCR9A RESIDUES WITH CONTACT FREQUENCY ≥6 OUT OF 15 MODELS FOR ANTIBODY 92R, RANKED BY FREQUENCY.

Position	Residue	Frequency (/15)
31	F	11
33	F	11
13	M	10
16	D	10
17	Y	10
20	E	10
32	N	10
15	D	9
21	S	9
30	N	9
34	T	9
14	A	8
19	S	8
35	D	8
37	Y	7
11	P	6
12	N	6

Within the same frequency, residues are listed in sequence order. The lower threshold was chosen to include N12 and P11, experimentally identified as the most critical binding residues. Own elaboration.

In the following, all results obtained for this antibody are presented.

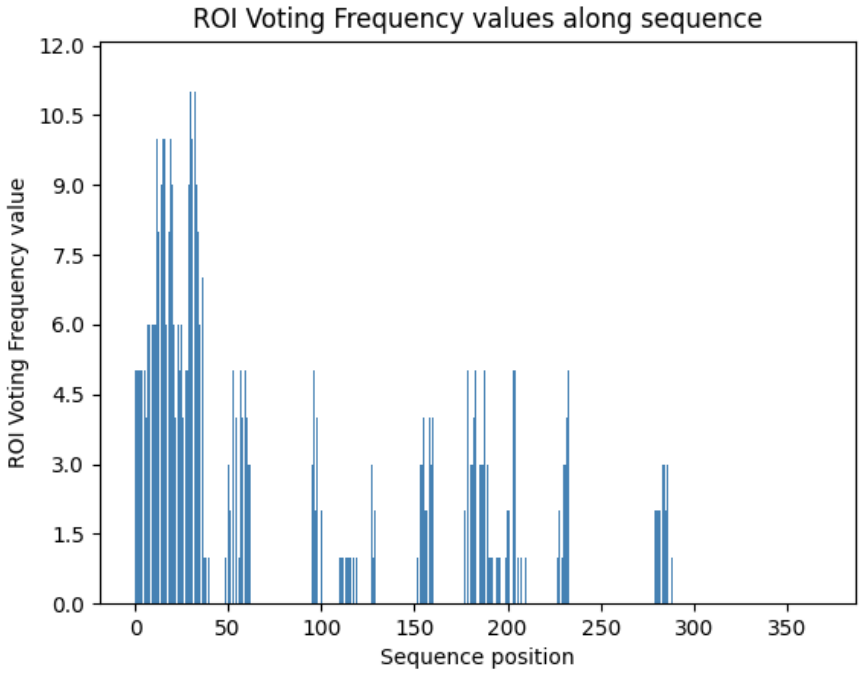


Fig. 4.1. Contact frequency of each CCR9A residue across 15 predicted complex structures (5 AlphaFold + 5 Boltz-1 + 5 Chai-1) for antibody 92R. Signal concentrates in the N-terminal domain (residues 1–48), with the highest-frequency residues spanning positions 11–37. Maximum frequency: 11/15. Own elaboration using Scipion-Chem [13].

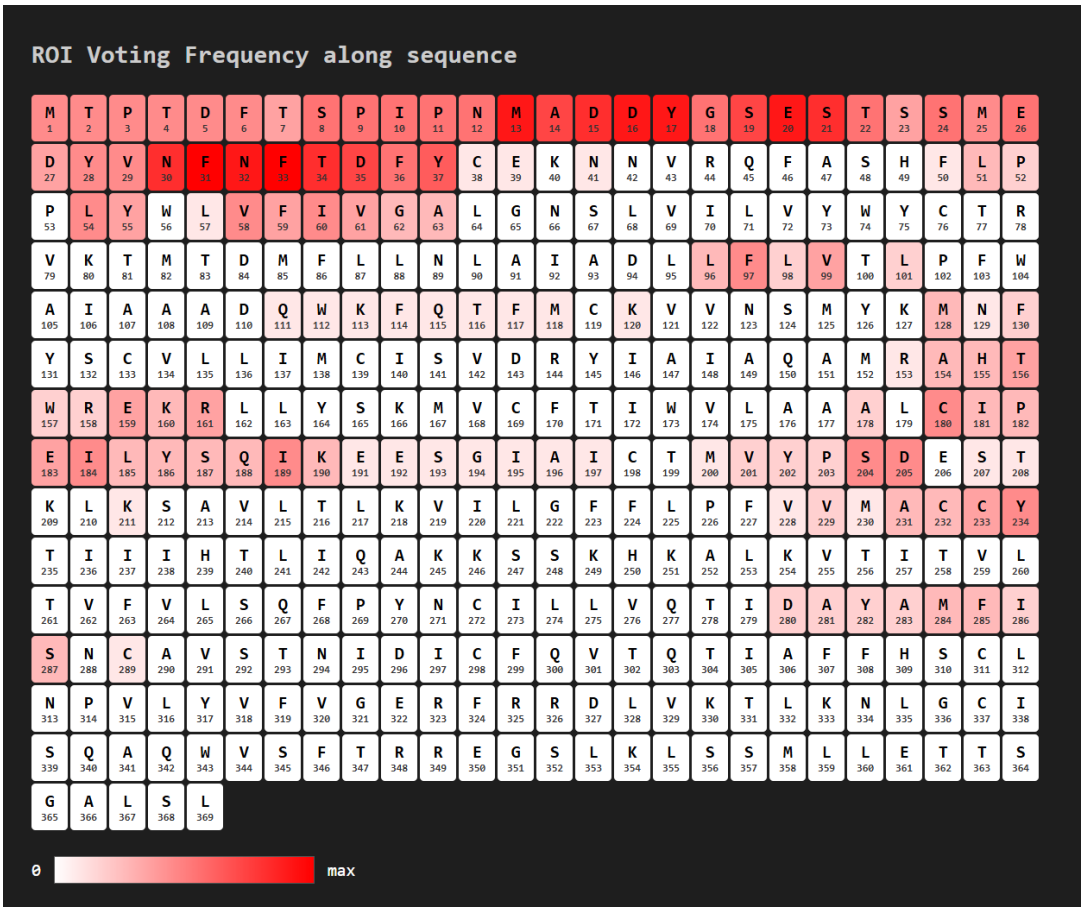


Fig. 4.2. ROI Voting Frequency mapped along the CCR9A sequence for antibody 92R. Color gradient from white (frequency = 0) to dark red (maximum = 11). The experimentally characterized epitope PNMADD (aa 11–16) falls within the high-frequency region. Own elaboration using Scipion-Chem [13].

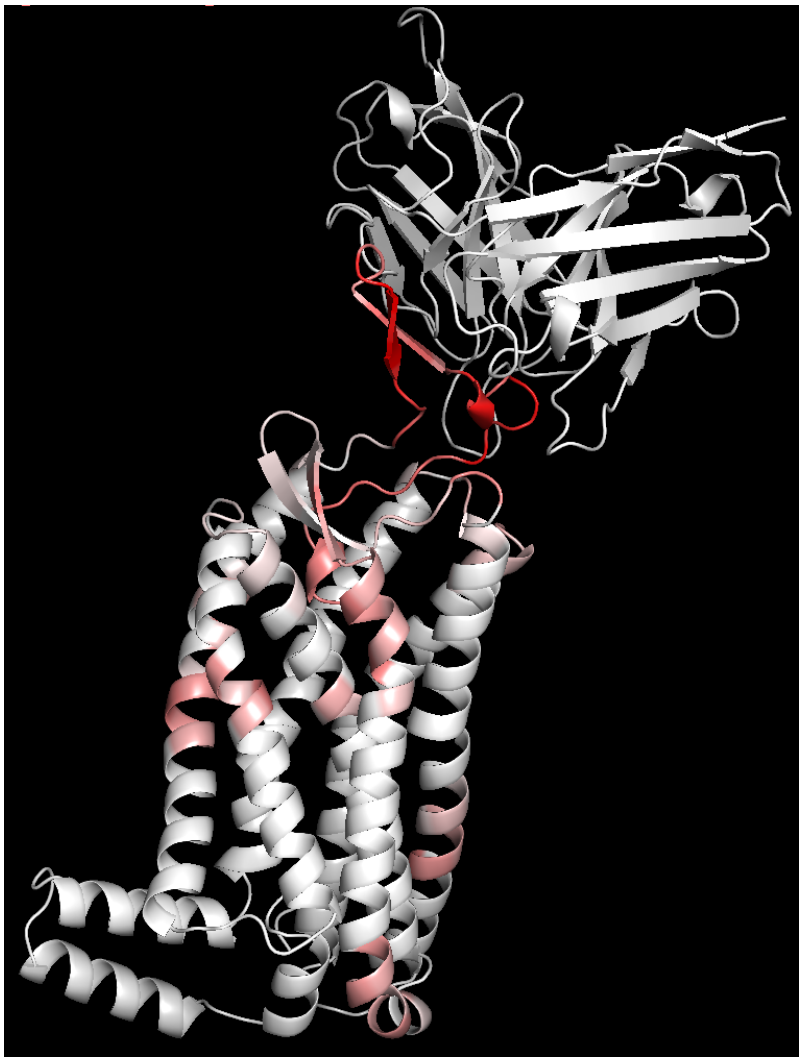
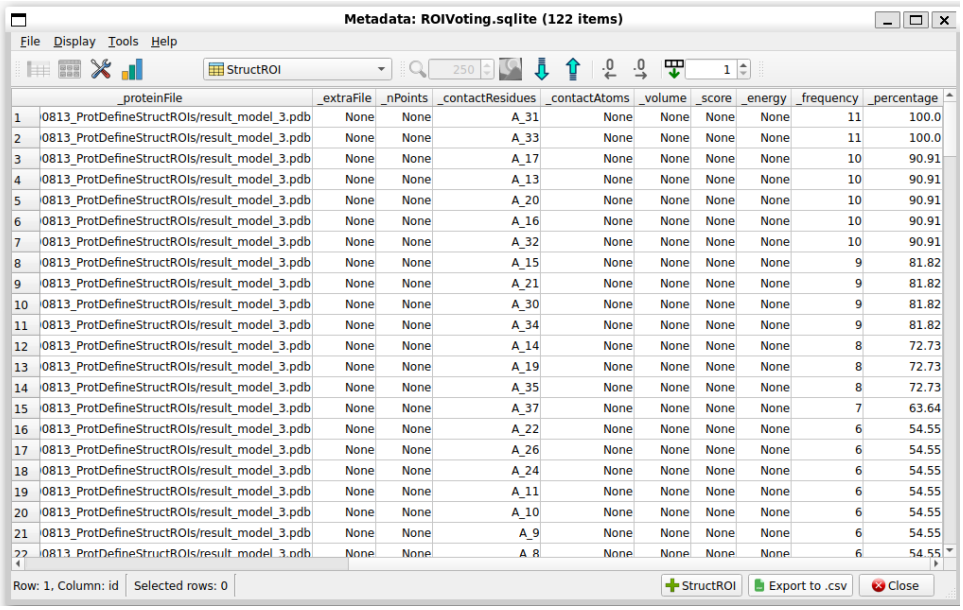


Fig. 4.3. PyMOL white-to-red gradient view of CCR9A for antibody 92R. Color gradient from white (frequency = 0) to red (maximum = 11).The highest contact residues (F31, F33; 11/15) appear in dark red. Signal is distributed across residues 11-27, with the transmembrane helices and intracellular regions remaining white. Own elaboration using ScipionChem [13] and PyMOL [16].



	_proteinFile	_extraFile	_nPoints	_contactResidues	_contactAtoms	_volume	_score	_energy	_frequency	_percentage
1	0813_ProtDefineStructROIs/result_model_3.pdb	None	None	A_31	None	None	None	None	11	100.0
2	0813_ProtDefineStructROIs/result_model_3.pdb	None	None	A_33	None	None	None	None	11	100.0
3	0813_ProtDefineStructROIs/result_model_3.pdb	None	None	A_17	None	None	None	None	10	90.91
4	0813_ProtDefineStructROIs/result_model_3.pdb	None	None	A_13	None	None	None	None	10	90.91
5	0813_ProtDefineStructROIs/result_model_3.pdb	None	None	A_20	None	None	None	None	10	90.91
6	0813_ProtDefineStructROIs/result_model_3.pdb	None	None	A_16	None	None	None	None	10	90.91
7	0813_ProtDefineStructROIs/result_model_3.pdb	None	None	A_32	None	None	None	None	10	90.91
8	0813_ProtDefineStructROIs/result_model_3.pdb	None	None	A_15	None	None	None	None	9	81.82
9	0813_ProtDefineStructROIs/result_model_3.pdb	None	None	A_21	None	None	None	None	9	81.82
10	0813_ProtDefineStructROIs/result_model_3.pdb	None	None	A_30	None	None	None	None	9	81.82
11	0813_ProtDefineStructROIs/result_model_3.pdb	None	None	A_34	None	None	None	None	9	81.82
12	0813_ProtDefineStructROIs/result_model_3.pdb	None	None	A_14	None	None	None	None	8	72.73
13	0813_ProtDefineStructROIs/result_model_3.pdb	None	None	A_19	None	None	None	None	8	72.73
14	0813_ProtDefineStructROIs/result_model_3.pdb	None	None	A_35	None	None	None	None	8	72.73
15	0813_ProtDefineStructROIs/result_model_3.pdb	None	None	A_37	None	None	None	None	7	63.64
16	0813_ProtDefineStructROIs/result_model_3.pdb	None	None	A_22	None	None	None	None	6	54.55
17	0813_ProtDefineStructROIs/result_model_3.pdb	None	None	A_26	None	None	None	None	6	54.55
18	0813_ProtDefineStructROIs/result_model_3.pdb	None	None	A_24	None	None	None	None	6	54.55
19	0813_ProtDefineStructROIs/result_model_3.pdb	None	None	A_11	None	None	None	None	6	54.55
20	0813_ProtDefineStructROIs/result_model_3.pdb	None	None	A_10	None	None	None	None	6	54.55
21	0813_ProtDefineStructROIs/result_model_3.pdb	None	None	A_9	None	None	None	None	6	54.55
22	0813_ProtDefineStructROIs/result_model_3.pdb	None	None	A_8	None	None	None	None	6	54.55

Fig. 4.4. Contact frequency table for antibody 92R showing CCR9A residues ranked by frequency score. Own elaboration using Scipion-Chem [13].

4.2.2. Summary and comparison with experimental epitope

The multi-model consensus places the predicted binding interface of 92R in a continuous region spanning residues 11–37 of the CCR9A N-terminal domain. Four of the six residues in the experimental epitope PNMADD: 'M13 (10/15), A14 (8/15), D15 (9/15), D16 (10/15)' appear at high frequency. The two residues experimentally identified as most critical, N12 and A14 [2], are both recovered: A14 at 8/15 and N12 at 6/15. P11 appears at frequency 6/15.

The highest-frequency residues, F31 and F33 (both at 11/15), fall outside the minimal PNMADD epitope but within the adjacent MEDYVNF and FTDFYC regions. This point is discussed further in Chapter 5.

4.2.3. Per-predictor analysis

Additionaly, a study looking individually the contribution of each model has been done. Undeniably, the three predictors do not agree on where the interface is and the disagreement is worth studying. Boltz-1 and Chai-1 both find the experimental epitope. All five Boltz-1 models place the contact region in the N-terminal domain, with residues 11–20 appearing in every model including P11 and N12 which are the two positions experimentally identified as most critical for 92R binding [2]. Chai-1 similarly obtains residues 13–35, recovering M13, A14, D15 and D16 in four of its five models. Neither tool produces meaningful signal outside the N-terminus. AlphaFold is the outlier. None of its five models detect any contact in the 11–16 region. All five instead converge on residues 30–37 and a broad cluster around positions 50–63 and 150–190 which correspond to trans-

membrane helices and intracellular loops (not the extracellular surface where 92R binds). AlphaFold appears to fold the disordered N-terminal tail away from the antibody and dock the scFv against a completely different part of the receptor. This was not anticipated and has no experimental support. The consequence for the consensus is concrete. F31 and F33 (both 11/15) rank at the top not because all three predictors agree, but because Boltz-1 and Chai-1 contact them from the right binding mode while AlphaFold contacts them incidentally from a different one. P11 and N12 (6/15 each) are recovered exclusively by Boltz-1 and Chai-1 and would not appear in the results at all if AlphaFold were the only predictor used. The multi-model consensus is really beneficial because it compensates for this. Undeniably, the per-predictor breakdown is greatly important to understand where the signal actually comes from and which predictors are driving the result.

4.3. In-depth analysis: antibody Ma-10

Ma-10 is the only antibody in this study without an experimentally characterized epitope. The patent describing it states only that it binds within the N-terminal domain of CCR9 [6]. Unlike the validation antibodies, there is no experimental reference against which to compare the prediction. The computational result presented here constitutes a genuine novel contribution: a candidate epitope proposed for the first time.

4.3.1. Interface residue frequency analysis

The workflow identified 126 contact residues, with a maximum frequency of 15 out of 15 models at residue F33. Ma-10 is the only antibody from all of them where any residue reaches full consensus. Table 4.3 lists residues with frequency ≥8.

Table 4.3. CCR9A RESIDUES WITH CONTACT FREQUENCY ≥8 OUT OF 15 MODELS FOR ANTIBODY MA-10, RANKED BY FREQUENCY.

Position	**Residue**	**Frequency (/15)**
33	F	15
31	F	13
32	N	13
29	V	11
27	D	10
30	D	10
35	N	10
2	Y	9
28	T	9
1	M	8
24	S	8
25	M	8
26	E	8

Own elaboration.

In the following, all results obtained for this antibody are presented.

The bar chart shows a sharp, well-defined peak centered on residues 25–35, with M1 and T2 appearing as a secondary signal at frequencies 8 and 9 respectively.

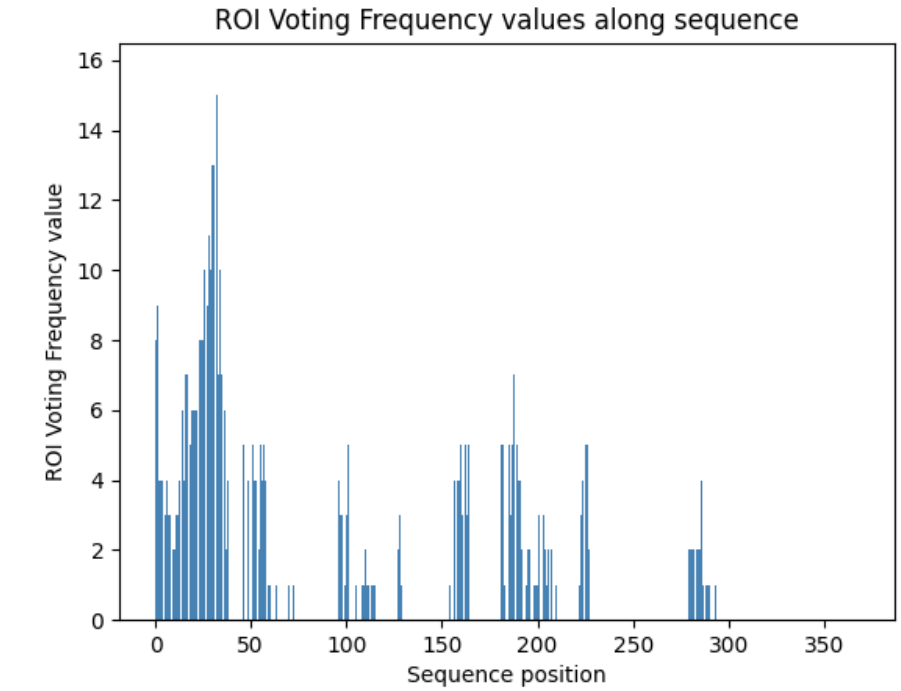


Fig. 4.5. Contact frequency of each CCR9A residue across 15 predicted complex structures for antibody Ma-10. F33 reaches the maximum possible frequency of 15/15 (the only perfect-consensus residue across all five antibodies). The primary high-frequency cluster spans residues 24–35. Own elaboration using Scipion-Chem [13].

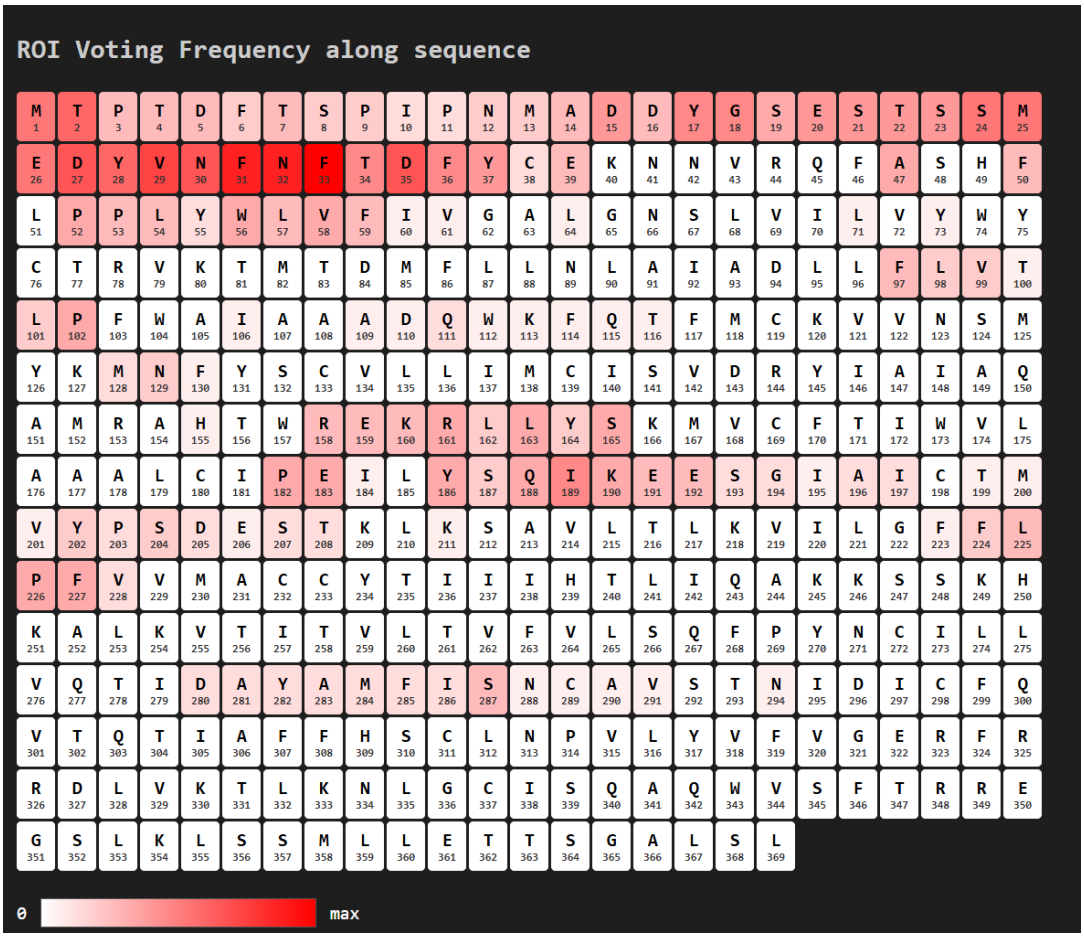


Fig. 4.6. ROI Voting Frequency mapped along the CCR9A sequence for antibody Ma-10. Color gradient from white (frequency = 0) to dark red (maximum = 15). F33 is the darkest red residue. The high-frequency cluster spans approximately residues 24–35, with a secondary signal at M1–T2. Own elaboration using Scipion-Chem [13].

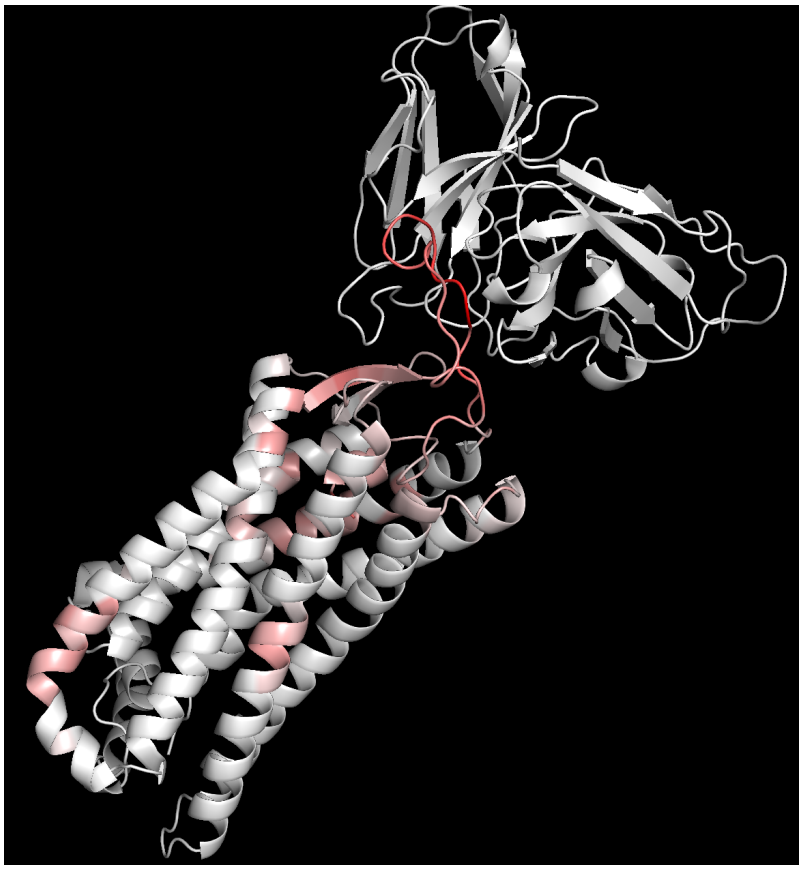
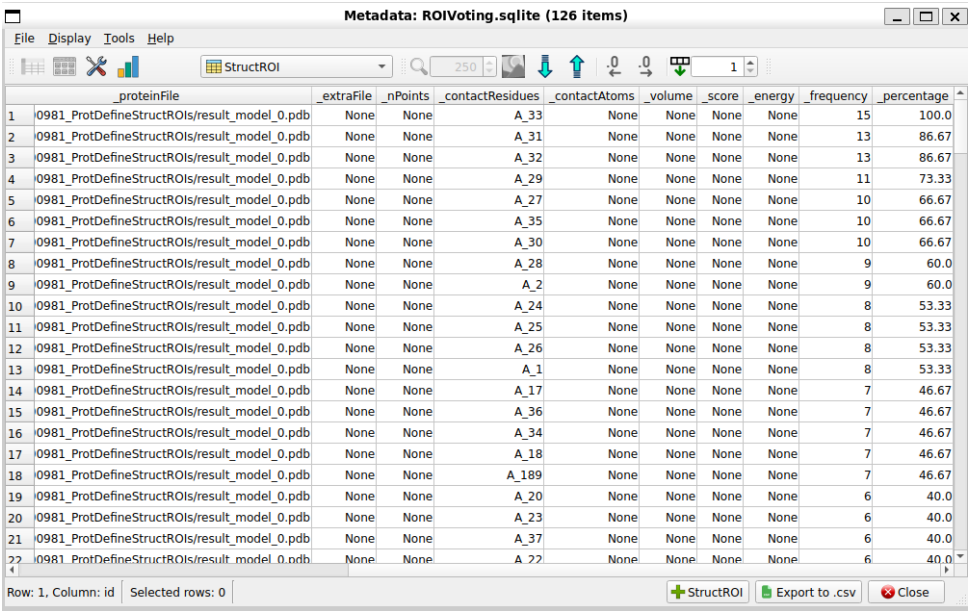


Fig. 4.7. PyMOL white-to-red gradient view of CCR9A for antibody Ma-10. F33 (15/15) appears as the most intensely colored residue. The high-frequency cluster (residues 24-35) is sharply defined in the NT domain. Own elaboration using Scipion-Chem [13] and PyMOL [16].



	_proteinFile	_extraFile	_nPoints	_contactResidues	_contactAtoms	_volume	_score	_energy	_frequency	_percentage
1	0981_ProtDefineStructROIs/result_model_0.pdb	None	None	A_33	None	None	None	None	15	100.0
2	0981_ProtDefineStructROIs/result_model_0.pdb	None	None	A_31	None	None	None	None	13	86.67
3	0981_ProtDefineStructROIs/result_model_0.pdb	None	None	A_32	None	None	None	None	13	86.67
4	0981_ProtDefineStructROIs/result_model_0.pdb	None	None	A_29	None	None	None	None	11	73.33
5	0981_ProtDefineStructROIs/result_model_0.pdb	None	None	A_27	None	None	None	None	10	66.67
6	0981_ProtDefineStructROIs/result_model_0.pdb	None	None	A_35	None	None	None	None	10	66.67
7	0981_ProtDefineStructROIs/result_model_0.pdb	None	None	A_30	None	None	None	None	10	66.67
8	0981_ProtDefineStructROIs/result_model_0.pdb	None	None	A_28	None	None	None	None	9	60.0
9	0981_ProtDefineStructROIs/result_model_0.pdb	None	None	A_2	None	None	None	None	9	60.0
10	0981_ProtDefineStructROIs/result_model_0.pdb	None	None	A_24	None	None	None	None	8	53.33
11	0981_ProtDefineStructROIs/result_model_0.pdb	None	None	A_25	None	None	None	None	8	53.33
12	0981_ProtDefineStructROIs/result_model_0.pdb	None	None	A_26	None	None	None	None	8	53.33
13	0981_ProtDefineStructROIs/result_model_0.pdb	None	None	A_1	None	None	None	None	8	53.33
14	0981_ProtDefineStructROIs/result_model_0.pdb	None	None	A_17	None	None	None	None	7	46.67
15	0981_ProtDefineStructROIs/result_model_0.pdb	None	None	A_36	None	None	None	None	7	46.67
16	0981_ProtDefineStructROIs/result_model_0.pdb	None	None	A_34	None	None	None	None	7	46.67
17	0981_ProtDefineStructROIs/result_model_0.pdb	None	None	A_18	None	None	None	None	7	46.67
18	0981_ProtDefineStructROIs/result_model_0.pdb	None	None	A_189	None	None	None	None	7	46.67
19	0981_ProtDefineStructROIs/result_model_0.pdb	None	None	A_20	None	None	None	None	6	40.0
20	0981_ProtDefineStructROIs/result_model_0.pdb	None	None	A_23	None	None	None	None	6	40.0
21	0981_ProtDefineStructROIs/result_model_0.pdb	None	None	A_37	None	None	None	None	6	40.0
22	0981_ProtDefineStructROIs/result_model_0.pdb	None	None	A_22	None	None	None	None	6	40.0

Fig. 4.8. Contact frequency table for antibody Ma-10 showing CCR9A residues ranked by frequency score. Own elaboration using Scipion-Chem [13].

4.3.2. Proposed candidate epitope for Ma-10

The multi-model consensus points to a binding interface centered on residues 25–35 of CCR9A, with F33 as the most consistently predicted contact residue. The corresponding sequence is `MEDYVNFNFTD` (aa 25–35).

This region bridges the experimentally characterized epitopes of 115 (MEDYVNF, 25–31) and 3C3 (FTDFYC, 33–38), with F33 being the first aminoacid from FTDFYC sequence. The workflow proposes **FNFTD** (residues 31–35) as the candidate epitope, with F33 as the predicted anchor residue. Once experimental data becomes available, comparing them against this prediction will be the real test of whether the workflow got it right. Pepscan analysis of the CCR9A N-terminal peptides would be the most direct way to test this.

4.3.3. Per-predictor analysis

For Ma-10, the three predictors reach a much stronger consensus than they do for 92R, which makes the results easier to interpret and harder to ignore. F33 is the one residue that appears in all 15 models, across all three predictors (5/5 AlphaFold, 5/5 Boltz-1, 5/5 Chai-1). That kind of unanimous agreement is unique and clearly suggests F33 is not a coincidence or an artifact of any of the tools. F31 and F32 follow at 13/15 each, again with contributions from all three predictors. The core of the proposed epitope :residues 31–35 is supported by at least three predictors in every case, with F33 being the only perfectly conserved contact across the entire dataset. The M1 and T2 signal is also really important to study because it constitutes the initiation of the N-terminal. M1 (8/15) and T2 (9/15) appear frequently enough to stand out, but the breakdown reveals they come almost entirely from AlphaFold (4/5 M1 and 5/5 T2) and Chai-1 (4/5 each). Boltz-1 detects neither in any of its five models. This predictor-specific pattern could be an artifact: the free N-terminal methionine is intrinsically disordered and AlphaFold and Chai-1 appear to fold it back toward the interface in some models while Boltz-1 does not. The absence of this signal in Boltz-1 is enough to know that M1 and T2 are not amino acids with robust consensus prediction. AlphaFold is again the outlier in terms of interface location. Most of its N-terminal contacts are limited to residues 31–35 and 47, while the bulk of its signal falls in transmembrane and intracellular regions (50–65, 97–102,155–165, 182–189). Boltz-1 and Chai-1 both concentrate strongly on the N-terminal domain, with over 80% of their contacts falling within residues 1–48. The proposed epitope FNFTD (31–35) is therefore obtained mainly by Boltz-1 and Chai-1, with AlphaFold contributing partial support at F31, N32 and F33 but diverging elsewhere.

Table 4.4. PER-PREDICTOR CONTACT FREQUENCIES FOR KEY CCR9A RESIDUES IN THE MA-10 ANALYSIS.

Residue	Total (/15)	AF (/5)	Boltz (/5)	Chai (/5)	Note
M1	8	4	0	4	N-terminal artifact?
T2	9	5	0	4	N-terminal artifact?
F31	13	3	5	5	Proposed epitope
N32	13	5	3	5	Proposed epitope
F33	15	5	5	5	Anchor residue
N34	7	0	4	3	Proposed epitope
F35	10	3	4	3	Proposed epitope

Values show the number of models (out of 5) in which each residue was identified as an interface contact. Own elaboration.

4.4. Comparative overview

Table 4.5 summarizes the key metrics across all five antibodies.

Table 4.5. SUMMARY OF MULTI-MODEL CONSENSUS RESULTS FOR ALL FIVE ANTI-CCR9 ANTIBODIES.

mAb	Max freq.	Total contacts	Peak residue	Cutoff	Experimental epitope
91R	12/15	117	F31	≥8	PNMADD (11–16)
92R	11/15	122	F31, F33	≥6	PNMADD (11–16)
115	14/15	111	N30	≥8	MEDYVNF (25–31)
3C3	12/15	125	F31, N32	≥8	FTDFYC (33–38)
Ma-10	15/15	126	F33	≥8	unknown

Own elaboration.

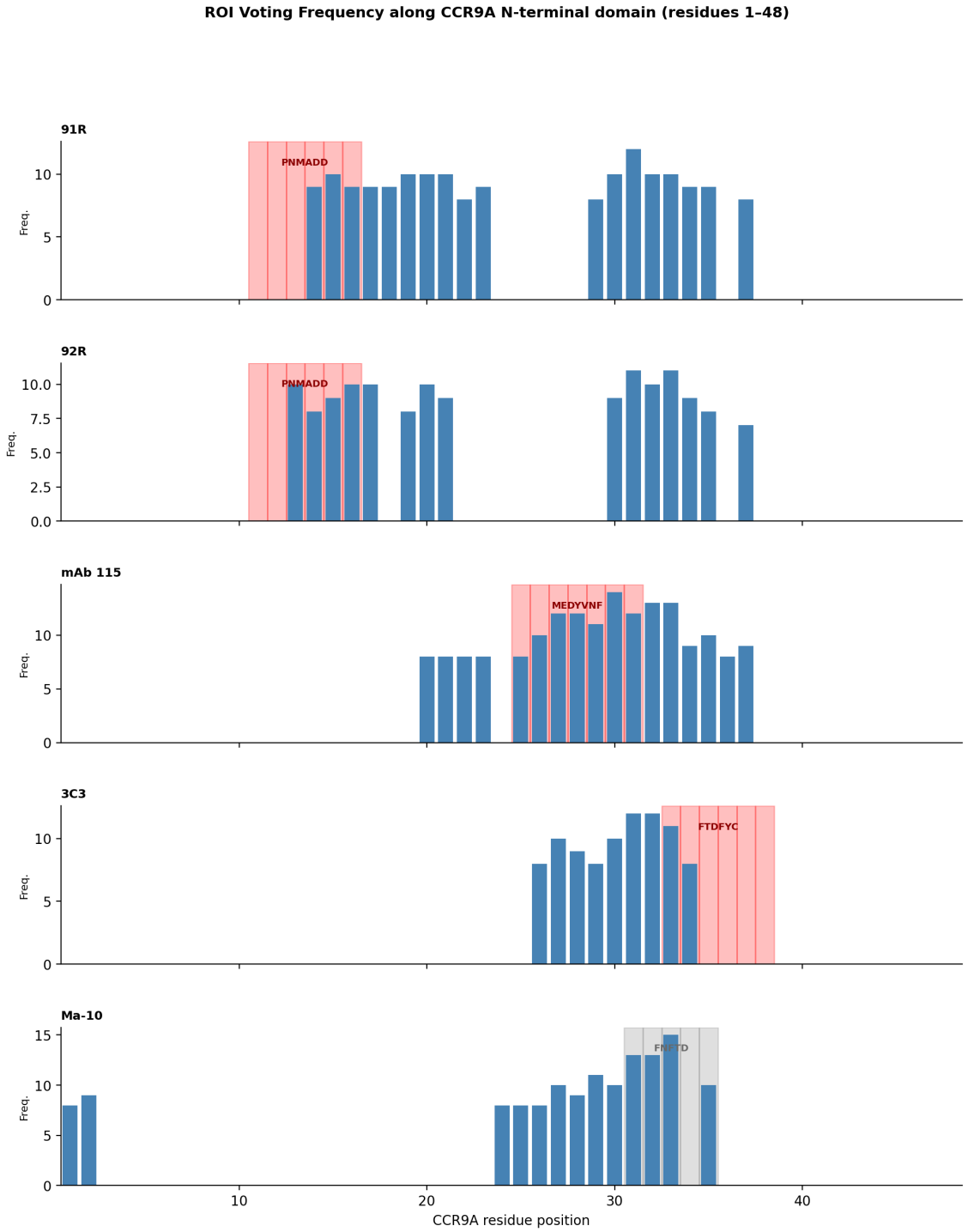


Fig. 4.9. ROI Voting Frequency along the CCR9A N-terminal domain (residues 1–48) for all five anti-CCR9 antibodies. Red shading indicates the experimentally characterized epitope region; grey shading indicates the computationally proposed candidate epitope for Ma-10 (FTDFYC, aa 33–38). No experimental epitope data is available for Ma-10. Own elaboration based on Scipion-Chem [13] output.

Residues in the window 27–33 appear at high frequency across all five antibodies regardless of their known epitope. Ma-10 stands out as the only antibody to reach full consensus at any residue (F33, 15/15), and it has the sharpest frequency peak of the group. 115 shows the strongest agreement with its experimental epitope. For 91R and 3C3, the workflow correctly identifies the binding region but predicts a broader interface than the minimal Pepscan epitope. This is discussed in Chapter 5.

5. DISCUSSION

5.1. Accuracy and limitations

5.1.1. Overall performance of the workflow

The results presented in Chapter 4 show that the multi-model consensus workflow correctly localizes the antibody-binding surface to the extracellular N-terminal domain of CCR9A in all five cases with no high-frequency false positives in transmembrane or intracellular regions. This is a remarkable result: the full-length CCR9A sequence (369 residues) was used as input in all predictions and the workflow had no prior information about the location of the epitopes. The fact that the signal consistently concentrates in residues 1–48 confirms that the structure predictors are generating biologically realistic complex geometries and that the contact frequency aggregation implemented in *ProtROIVoting* is an effective strategy for filtering consistent interface signals from an heterogeneous set of structural models.

From each antibody we can obtain valuable information. Agreement with experimental epitopes varies across the three validation antibodies. For 115, the agreement is nearly perfect because all seven residues of MEDYVNF (25–31) are recovered at high frequency, with N30 reaching 14 out of 15 models. For 91R and 3C3, the main epitope residues are present but the predicted window is broader than the minimal Pepscan sequence. The fact that we are obtaining computationally a broader region for the epitope sequence is expected. Pepscan identifies the shortest peptide sufficient for antibody binding, on the other hand, the structural models capture the full interface geometry of the folded complex, which includes neighboring residues that contribute to binding stability without being strictly required.

The overall performance of the workflow is consistent with the difficulty of the task as reported in the literature. Clifford et al. [14] showed that even optimized scoring methods achieve AUC values around 0.74 for antibody-antigen interface prediction on large benchmark datasets. The multi-model consensus strategy used here does not rely on a trained classifier but still correctly recovered the binding region in all four validation cases with no substantial false positives outside the N-terminal domain.

These findings support the hypothesis stated in the introduction: computational epitope prediction can serve as a faster and more cost-effective alternative to experimental mapping. In particular, for membrane proteins like CCR9 where structural characterization are technically challenging.

5.1.2. Output sequence assignment in ProtROIVoting

One technical issue encountered during the analysis was the incorrect assignment of the output sequence in the *ProtROIVoting* protocol. All the study is revolved around he CCR9A sequence where the epitope must be found. However, in some runs, the consensus frequency scores were mapped onto the scFv sequence (~250 residues), producing results that were biologically meaningless (the frequency values were being assigned to antibody esidues instead of antigen residues). The issue was identified by verifying the length of the output sequence before interpreting any results and fixed by re-running the protocol and after confirming that chain A in the structural predictions corresponded to CCR9A and matched the chain of interest defined in *ProtDefineStructROI*. This highlights the importance of systematically checking output sequence identity as a quality control step when applying multi-model consensus workflows to protein–protein complexes.

5.1.3. Parameter selection in ProtDefineStructROI

An important challenge encountered during the development of the workflow was the choice of parameters for the *ProtDefineStructROI* protocol, particularly the interface distance threshold (set at 3.0 Å) and the maximum atom depth (3.0 Å). These values are critical because they define which residues are counted as interface contacts in each structural model and evidently different parameter combinations give different contact lists. The values used in this thesis were chosen based on standard practice in the structural biology literature for defining direct inter-residue contacts. On top of that, an extensive analysis repeating the workflow across a range of threshold values and comparing the resulting frequency distributions took place to decide the final values that were going to be used across the five antibodies.

5.1.4. From pocket-based to model-based voting

An important methodological decision made during the development of this workflow was the switch from pocket-based to model-based voting in the *ProtROIVoting* protocol. Initially, the voting was performed per pocket: a given residue could be counted multiple times in the same structural model if it appeared in more than one identified pocket. This meant that the maximum possible frequency for a residue was not the number of models (15) which made the frequency scores difficult to interpret biologically. By switching to model-based voting in which each residue contributes at most one count per model, regardless of how many pockets it appears in, the frequency score is bounded to 15 and acquires a direct biological interpretation. This change greatly improved the interpretability of the results and is the approach used throughout this thesis.

5.1.5. Contact frequency cutoffs

The choice of frequency cutoff for defining high-confidence interface residues initially was the same for all of the antibodies. Nonethless, it was more appropriate for each antibody to analyze the results and see which one was the most appopriate based on the shape of the frequency distribution. For four of the five antibodies a threshold of ≥8 out of 15 models was used; for 92R the threshold was lowered to ≥6 to include N12 and P11 which form part of the experimental epitope sequence being the most critical for binding [2]. For a future scope, a more objective way to choose this cutoff would be to define a statistical threshold (instead of manually) although the cutoff values used in this thesis were selected based on the observed frequency distributions and are consistent.

An open question that arises from this analysis is what the exact criteria is that separates a residue that is *functionally important* from one that is *structurally important*. A residue that is consistently within 3 Å of the antibody CDRs across models may always be spatially close to the interface without the atoms of that receptor residue being able to form chemical bonds with the antibody atoms. This means it shapes the binding geometry without being strictly required for recognition. Alanine scanning mutagenesis, discussed in Section 5.2 would be the most direct way to address this distinction.

5.1.6. Secondary contacts outside the N-terminal domain

For the antibody 3C3, even though it is not an antibody studied in depth in this thesis, it is worth mentioning the secondary contact signal found at residue T190 which falls within extracellular loop 2 (EC2) of CCR9A (approximately residues 186–210, between TM4 and TM5). In the three-dimensional models, EC2 folds spatially close to the N-terminal domain, bringing T190 very close with the primary contact region. This may reflects a genuine secondary binding contact or a modeling artifact of the flexible loop geometry and it is unclear.

For antibody Ma-10, residues M1 and T2 appear at frequencies of 8 and 9 out of 15 models respectively. These are the first two amino acids of the CCR9A sequence and their appearance as predicted contacts may reflect a modeling artifact: the free N-terminal end of the polypeptide chain is intrinsically flexible and may fold back onto the interface in some models, generating spurious contacts.

5.1.7. The proposed epitope for Ma-10

The computational prediction for Ma-10 places the binding interface in the region spanning residues 25–35 of CCR9A, with F33 as the anchor residue at full consensus (15/15 models). This region bridges the experimentally characterized epitopes of 115 (MEDYVNF, 25–31) and 3C3 (FTDFYC, 33–38) and the proposed minimal epitope FNFTD (31–35) is located in between them and partially overlapping both.

Two observations support the credibility of this prediction. First, F33 reaches 15/15 which is unlikely to result from random model variability. Second, the overall frequency profile for Ma-10 is the sharpest and most localized of the five antibodies, suggesting that the structural models are placing the scFv in a consistent orientation.

Having said this, the prediction must be interpreted with caution. The N-terminal domain of CCR9A has no fixed structure in solution so the models are just picking one conformation out of the many possible ones. Additionaly, The fact that residues 27–33 appear at high frequency across all five antibodies including those whose experimental epitope is not located there (91R, 92R) suggests that this segment may be systematically over-represented in the structural models, possibly because it adopts a particularly accessible conformation in the predicted complexes. This does not invalidate the Ma-10 prediction, but it means that experimental confirmation is essential before having certain coclusions.

5.1.8. Contribution of individual predictors

The per-predictor analysis for 92R and Ma-10 is fundamentall for understanding and seeing what results come from each model and obtain some conclusions.

For antibody 92R, AlphaFold produced outlier predictions in all five models, placing contacts in the transmembrane and intracellular regions of CCR9A rather than the N-terminal extracellular domain. This is likely a consequence of how AlphaFold handles flexible, disordered N-terminal regions. On the other hand, Boltz-1 and Chai-1 did not show this behavior. Boltz-1 obtained P11 and N12 residues (most relevant amino acids from epitope PNMADD) in every one of its five models. Chai-1 covered the residues M13–D16, which are also within the experimental epitope. Two out of three predictors converging on the correct region without any structural information given is reassuring.

For Ma-10, the per-predictor breakdown makes the case for F33 straightforwardly: 5/5 AlphaFold, 5/5 Boltz-1, and 5/5 Chai-1. No other residue in the dataset reached this level of agreement across all three predictors which gives the F33 contact unique credibility.

The M1/T2 signal in the Ma-10 data is also worth discussing, precisely because the multi-predictor setup is what shows that it could potentially be an artifact. In the aggregate frequency plot, both residues appear at high frequency. Broken down by predictor, however, the pattern is inconsistent: AlphaFold and Chai-1 both flag them, while Boltz-1 gives them a score of zero. A residue that is only seen by two predictors and biologically it is located in the tip of the protein should not be given the same importance as one that appears in all 15 models. The multi-predictor approach makes this distinction visible in a way that a single-predictor analysis would not.

5.2. Future perspectives

Several directions could extend and refine the work presented in this thesis.

Alanine scanning mutagenesis. A natural next step for validating the predicted epitopes and particularly the Ma-10 candidate would be computational alanine scanning. Two variants of this approach could be applied: single alanine scanning, replacing one residue at a time or double alanine scanning, replacing two consecutive residues simultaneously. Both methods can provide complementary results. This would only be applied to the amino acids above the cutoff frequencies. This reduces greatly the work but it is still very computational demanding. For each antibody there would be a certain number of repetitions for example for 91R it would mean 18 repetitions of the full workflow, while for 3C3 only 9. In each case, the residues predicted to be at the interface are individually substituted by alanine and the workflow is repeated for each one. Residues whose substitution substantially disrupts the predicted interface (reflected in lower contact frequencies or altered binding geometry) would be identified as functionally important. This was proposed by the experimental group (L. Kremer, CNB-CSIC) as a computational complement to substitution Pepscan, the experimental technique used to identify critical residues for 91R and 92R [2]. The approach is methodologically straightforward but computationally demanding.

Species-comparative N-terminal domain analysis. An interesting direction would be to analyze in silico the potential binding of the five antibodies using only the N-terminal amino acid sequence (residues 1–48) of mCCR9 and hCCR9 as antigen to compare results between species. It would be useful to start just with the NT domain alone because it would be computationally much easier to follow. If the results were informative, the analysis could then be repeated with the full receptor. Comparing the binding results of the five mAbs to human CCR9 Vs mouse CCR9 could indicate whether the residues that appear frequently in the predictions but were not identified experimentally are relevant because they help shape the three-dimensional structure that positions the minimal epitope correctly or whether they simply reflect species-specific differences with no functional role in antibody recognition. A sequence alignment of the human and mouse CCR9 N-terminal domains would provide a straightforward first indication of which positions are conserved and which differ between species.

Antibody sequence optimization. Starting from a known antibody sequence, each position in the variable region is systematically mutated to all other amino acids and the effect on binding score is evaluated for each variant. This strategy, analogous to directed evolution but guided by structure prediction, could be applied to the five anti-CCR9 antibodies studied here to identify sequence variants with improved predicted affinity for their respective epitopes. Protocols implementing this approach are already available within the Scipion-Chem platform and are currently being developed in the B13 department at CNB-CSIC making this a really realistic extension of the present work.

Per-predictor analysis and consensus refinement. As noted in the limitations section, the relative contribution of AlphaFold, Boltz-1, and Chai-1 to the final consensus has not been characterized in detail. Doing a systematic per-predictor breakdown would clarify whether the three tools do agree on the same interface residues or if the consensus is driven mainly by one or two of the predictor. This analysis could also inform a weighted voting scheme, in which models from better-performing predictors receive higher weight in the final frequency count instead of all having the same value.

More biologically grounded scoring functions. The current workflow uses a simple contact frequency count as its interface score (either '1' or '0'). More sophisticated approaches like for example, weighting contacts by the predicted confidence of the structural model (pLDDT or ipTM scores) or incorporating physicochemical complementarity metrics at the interface could improve the resolution of the prediction and reduce the impact of low-confidence models on the final result. Some of these scoring functions are already available in the Scipion-Chem platform making this also a really realistic extension of the existing workflow.

Experimental validation of the Ma-10 epitope. The most direct extension of this work is experimental: Pepscan analysis of the CCR9A N-terminal peptides using the Ma-10 antibody, following the same protocol applied to 91R, 92R, 115, and 3C3, would confirm or disprove the proposed FNFTD (31–35) candidate epitope. Substitution Pepscan at the predicted critical positions would further identify the anchor residues within the epitope.

6. CONCLUSIONS

The following conclusions can be obtained from the work presented in this thesis:

1. A reproducible computational workflow for the identification of antibody-recognized epitopes on the CCR9 receptor has been designed and implemented within the Scipion-Chem platform, integrating three independent AI-based structure predictors (AlphaFold, Boltz-1 and Chai-1) and a multi-model consensus strategy based on residue contact frequency aggregation. This fulfils Objective 2.

2. The workflow was validated against three anti-CCR9 monoclonal antibodies with experimentally characterized epitopes (91R, 115, and 3C3). In all three cases, the high-frequency contact residues were correctly localized to the extracellular N-terminal domain of CCR9A, with no substantial false positives in transmembrane or intracellular regions. Agreement with the experimental epitope was strongest for 115 where all seven residues of MEDYVNF (25–31) were obtained at high contact frequency. This fulfils Objectives 3 and 5.

3. For antibody 92R, the in-depth analysis recovered the experimental epitope PNMADD (11–16), including the two experimentally critical residues N12 and A14. The predicted high-frequency window (residues 11–37) is broader than the minimal Pepscan epitope. However, this makes sense due to the full-length structural models. This fulfils Objectives 3 and 5.

4. For antibody Ma-10, whose binding epitope has not been experimentally determined, the workflow predicts a binding interface centered on residues 25–35 of the CCR9A N-terminal domain, with F33 as the anchor residue reaching full consensus across all 15 structural models (15/15). The proposed minimal candidate epitope is FNFTD (residues 31–35). This fulfils Objective 4.

5. The switch from pocket-based to model-based voting in the *ProtROIVoting* protocol was a key methodological decision that improved the biological interpretability of the frequency scores, bounding the maximum possible frequency to the number of structural models used. This fulfils Objective 6.

6. The main limitations of the current workflow are the subjective nature of the frequency cutoff and the inherent difficulty of modeling the disordered N-terminal domain of CCR9A. In section 5.2 the main idea is to work on these issues. This fulfils Objective 7.

7. The proposed candidate epitope for Ma-10 requires experimental validation, most directly by Pepscan analysis of the CCR9A N-terminal peptides. Computational alanine scanning of the predicted high-frequency residues represents a further step

that could potentially prioritize positions for experimental mutagenesis. This fulfils Objective 7.

7. SOCIO-ECONOMIC IMPACT, ETHICAL IMPLICATIONS AND EXPENDITURE

7.1. Socio-economic impact

Therapeutic antibodies are now the largest and fastest-growing segment of the pharmaceutical market, with the global antibody therapy market valued at $317.4 billion in 2025 and projected to reach $1.01 trillion by 2035 [17]. Getting a new antibody from target identification to clinical approval can take over a decade and a significant portion of that time is spent on epitope characterization figuring out exactly which region of the target protein the antibody binds. Methods like Pepscan, HDX-MS or cryo-EM work well, but they are slow and expensive. In particular, for membrane proteins like GPCRs.

CCR9 is a clinically relevant target in two disease areas with a high medical need that is not yet covered. In T-cell acute lymphoblastic leukemia (T-ALL), adult patients face a five-year overall survival rate of approximately 50%, which drops below 10% in relapsed or refractory disease [18]. CCR9-directed CAR-T therapies are under active investigation as a strategy to improve these outcomes [3], [5]. In inflammatory bowel disease, CCR9 antagonism has been explored as a therapeutic strategy given its role in intestinal T cell trafficking [4]. The workflow developed here is directly applicable to antibody development programs in both areas.

The economic burden of epitope mapping is considerable. Experimental techniques like HDX-MS or cryo-EM structural determination of antibody–antigen complexes require specialized instrumentation, trained personnel and significant reagent costs, making them inaccessible to many academic research groups. Computational approaches like the one developed in this thesis can reduce the number of experimental iterations required, potentially saving significant resources per antibody candidate in early-stage development and allowing smaller laboratories with less resources prioritize the most promising candidates before making any expensive experimental validation.

The workflow developed in this thesis fixes that problem. Three AI-based structure predictors, a multi-model consensus strategy, publicly free available tools, no cost. Applied to CCR9, a receptor relevant in both T-cell leukemia and inflammatory bowel disease, it produced predictions consistent with the known experimental epitopes for four of the five antibodies tested and proposed a candidate for the fifth. Beyond CCR9, the workflow is generalizable to any antibody–antigen system where the antigen sequence is known. This is particularly relevant for membrane proteins which represent more than 60% of current drug targets despite being structurally undercharacterized due to the technical difficulty of their experimental study [19], [20]. The open-source nature of all tools used (AlphaFold, Boltz-1, Chai-1, Scipion) means the workflow is accessible to research

groups without access to expensive commercial platforms, making it accessible also to smaller academic laboratories and biotechnology startups.

By applying these tools: the experimental iterations needed to find an epitope are reduced, candidates can be prioritized earlier and CAR-T design can be informed better. Cutting even one round of experimental validation saves real time and money. From an environmental perspective, computational approaches reduce the need for animal experiments and wet lab consumables. Computational approach can help identify the most promising antibody candidates before committing to in vivo validation, reducing animal use in line with the 3R principles (Replacement, Reduction, Refinement) established in Directive 2010/63/EU on the protection of animals used for scientific purposes.

7.2. Ethical implications

For the completion of this study no animals, human subjects, or biological samples were involved at any stage. Everything used is publicly available sequences, published data, and open-access tools.

There is also a practical argument for this kind of work beyond the science. In a publicly funded institution like the CNB-CSIC, every round of experimental screening that can be avoided or better targeted is money that will go somewhere else. Computational epitope prediction does not replace the experiments but it can make them a lot cheaper to plan.

Having said that, it is also important to be aware of the limitations of this technique. The AI predictors it relies on are not fully interpretable and they have not been validated across all protein classes. The results presented here are hypotheses. They are meant to guide experiments but not replace them. So, when using this kind of tools, this must be taken into account.

7.3. Budget and expenditure

Table 7.1 breaks down the estimated costs of this thesis. Personnel costs are estimated using the 2025 Spanish minimum wage as a reference (7.40 €/hour gross) for student work, and the 30% employer social security contribution is included. Supervisor and tutor hours are estimated at 35 €/hour consistent with the salary range of a senior researcher or associate professor in Spain. All software was open-source; no licenses were needed.

Table 7.1. ESTIMATED BUDGET FOR THIS THESIS.[3]

Item	Details	Cost (€)
Personnel		
Practical work (330 h)[1]	7.40 €/h gross	2,442
Writing and analysis (30 h)[1]	7.40 €/h gross	222
Carlos Oscar Sorzano (tutor, 30 h)	35 €/h gross	1,050
Leonor Kremer (supervisor, 30 h)	35 €/h gross	1,050
Cristina Quílez López (tutora UC3M, 10 h)	35 €/h gross	350
Social security (30%)		1,5344
Personnel subtotal		**6,648**
Equipment (3 months amortization)		
LG PC Intel Core Ultra 7 155H	1,500 €, 4-year life	93
Microsoft Surface Pro 9	1,200 €, 4-year life	75
Equipment subtotal		**168**
Travel		
Commuting (66 days, 36 km/day)	~7 L/100 km, 1.65 €/L	274
Travel subtotal		**274**
Other		
Electricity	360 h, 65 W, 0.15 €/kWh	4
Software licenses	All open-source	0
Indirect costs (20% personnel)[2]	Infrastructure, admin, etc.	900
Other subtotal		**900**
TOTAL		**7,590**

Own elaboration.

[1]The Spanish minimum interprofessional wage (SMI) for 2025 is 1,184 €/month, equivalent to approximately 7.40 €/hour for a standard 40-hour week (a research technician salary at a CSIC-affiliated center would be closer to 12–13 €/hour). In practice, this work was not remunerated; personnel costs are included for budgetary completeness only.

[2]Most indirect costs would still be present regardless of this thesis. The marginal personal consumption (electricity for monitor or laptop use, bathroom, etc.) is minimal.

[3]All figures include IVA, social security contributions and other applicable taxes where relevant.

Personnel is by far the biggest cost. Equipment and travel are minor. The zero software cost is worth mentioning. A similar analysis using commercial tools like the

Schrödinger Suite could easily add several thousand euros in licensing fees alone.

BIBLIOGRAPHY

[1] S. Chamorro et al., "Antitumor effects of a monoclonal antibody to human CCR9 in leukemia cell xenografts," *mAbs*, vol. 6, 2014. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC4171004/

[2] B. Somovilla-Crespo et al., "92R monoclonal antibody inhibits human CCR9+ leukemia cells growth in NSG mice xenografts," *Frontiers in Immunology*, vol. 9, p. 77, 2018. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC5797297/

[3] N. Tirado, K. Fidyt, M. J. Mansilla, et al., "CAR-T cells targeting CCR9 and CD1a for the treatment of T cell acute lymphoblastic leukemia," *Journal of Hematology & Oncology*, vol. 18, p. 69, 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC12220514/

[4] B. A. Zabel et al., "Human G protein-coupled receptor GPR-9-6/CC chemokine receptor 9 is selectively expressed on intestinal homing T lymphocytes, mucosal lymphocytes, and thymocytes," *The Journal of Experimental Medicine*, vol. 190, no. 9, pp. 1241–1256, 1999. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC2195678/

[5] P. M. Maciocia et al., "Anti-CCR9 chimeric antigen receptor T cells for T cell acute lymphoblastic leukemia," *Blood*, vol. 140, no. 1, pp. 25–37, 2022. DOI: 10.1182/blood.2021013648

[6] P. M. Maciocia and M. A. Pule, *Anti-CCR9 antibodies and chimeric antigen receptors*, US Patent Application #20240374727, Filed by University College London. Available at: https://patentimages.storage.googleapis.com/f6/71/f3/c5c0bbccebe0f5/US20240374727A1.pdf, 2024.

[7] UniProt Consortium. "C-c chemokine receptor type 9 (CCR9) [P51686]," Accessed: Feb. 5, 2026. [Online]. Available: https://www.uniprot.org/uniprotkb/P51686/entry

[8] BioRender. "BioRender," Accessed: Mar. 1, 2026. [Online]. Available: https://www.biorender.com

[9] J. Jumper et al., "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021. DOI: 10.1038/s41586-021-03819-2

[10] Google DeepMind. "AlphaFold server," Accessed: Feb. 1, 2025. [Online]. Available: https://alphafoldserver.com/

[11] MIT Jameel Clinic. "Boltz-1: Democratizing biomolecular interaction modeling," Accessed: Feb. 7, 2025. [Online]. Available: https://github.com/jwohlwend/boltz

[12] Chai Discovery. "Chai-1: A multi-modal foundation model for molecular structure prediction," Accessed: Feb. 7, 2025. [Online]. Available: https://www.chaidiscovery.com

[13] Scipion-CHEM team. "Scipion-chem documentation," Accessed: Jan. 3, 2025. [Online]. Available: https://scipion-chem.github.io/docs/index.html

[14] J. N. Clifford et al., "AbEpiTope-1.0: Improved antibody target prediction by use of alphafold and inverse folding," *Science Advances*, vol. 11, 2025. DOI: 10.1126/sciadv.adu1823

[15] Tamarin Bio. "Tamarin.bio: Protein structure prediction platform," Accessed: Feb. 7, 2025. [Online]. Available: https://tamarin.bio

[16] Schrödinger, LLC, *The PyMOL molecular graphics system, version 2.5*, 2021.

[17] Global Market Insights Inc. "Antibody therapy market size and share 2026-2035," Accessed: Jun. 8, 2026. [Online]. Available: https://www.gminsights.com/industry-analysis/antibody-therapy-market

[18] D. I. Marks and C. Rowntree, "Management of adults with T-cell lymphoblastic leukemia," *Blood*, vol. 129, no. 9, pp. 1134–1142, 2017. DOI: 10.1182/blood-2016-07-692608

[19] J. P. Overington, B. Al-Lazikani, and A. L. Hopkins, "How many drug targets are there?" *Nature Reviews Drug Discovery*, vol. 5, pp. 993–996, 2006. DOI: 10.1038/nrd2199

[20] H. Yin and A. D. Flynn, "Drugging membrane protein interactions," *Annual Review of Biomedical Engineering*, vol. 18, pp. 51–76, 2016. DOI: 10.1146/annurev-bioeng-092115-025322

A. VALIDATION ANTIBODY RESULTS

This appendix contains the complete set of figures and frequency tables for the three validation antibodies (91R, 3C3, and 115).

A.1. Antibody 91R

Table A.1. CCR9A RESIDUES WITH CONTACT FREQUENCY ≥8 OUT OF 15 MODELS FOR ANTIBODY 91R, RANKED BY FREQUENCY.

Position	Residue	Frequency (/15)
31	F	12
15	D	10
19	S	10
20	E	10
21	S	10
30	N	10
32	N	10
33	F	10
14	A	9
16	D	9
17	Y	9
18	G	9
23	S	9
34	T	9
35	D	9
22	T	8
29	V	8
37	Y	8

Within the same frequency, residues are listed in sequence order. Own elaboration.

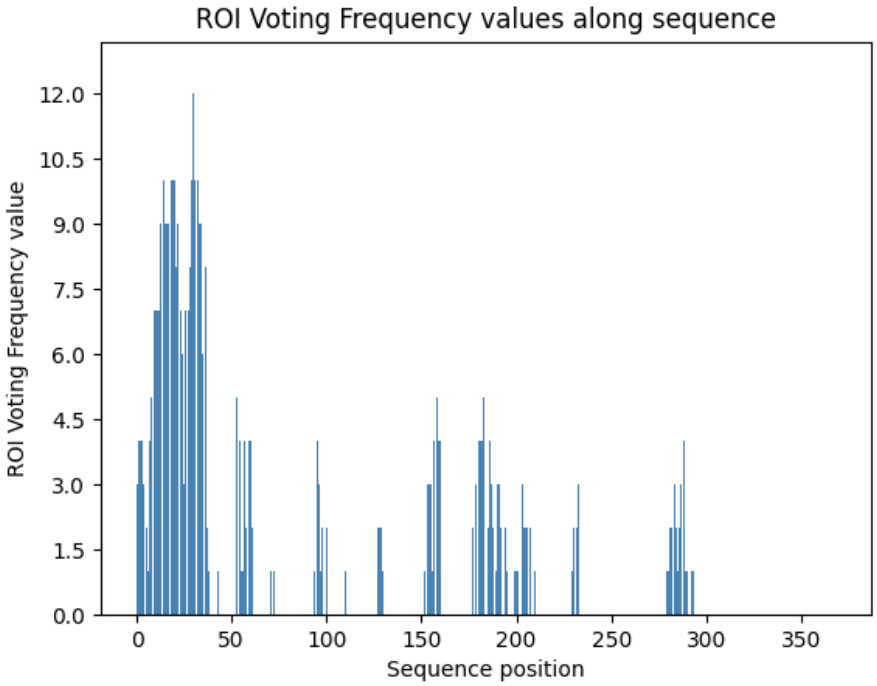


Fig. A.1. Contact frequency of each CCR9A residue across 15 predicted complex structures for antibody 91R. Signal concentrates in the N-terminal domain (residues 1–48), with maximum frequency 12/15 at F31. Own elaboration using Scipion-Chem [13].

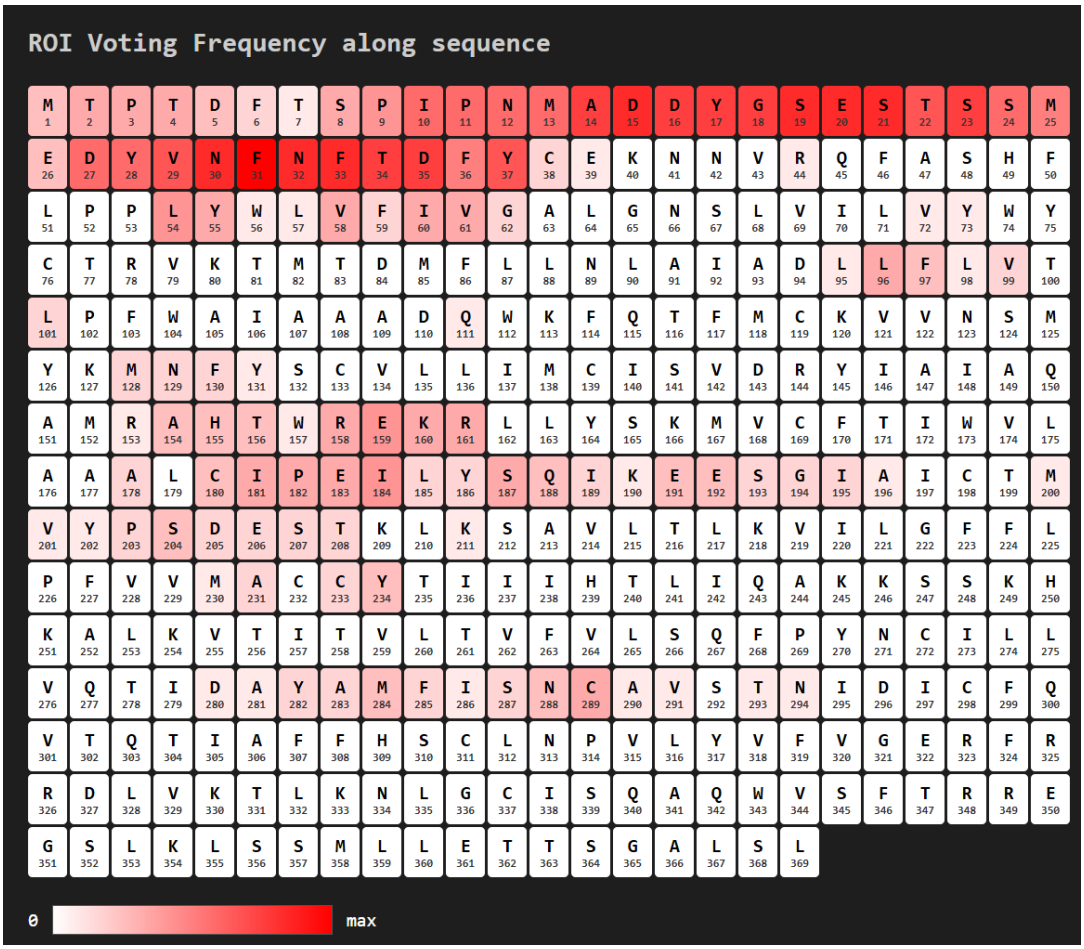


Fig. A.2. ROI Voting Frequency mapped along the CCR9A sequence for antibody 91R. Color gradient from white (frequency = 0) to red (maximum = 12). The known epitope PNMADD (aa 11–16) falls within the high-frequency region. Own elaboration using Scipion-Chem [13].

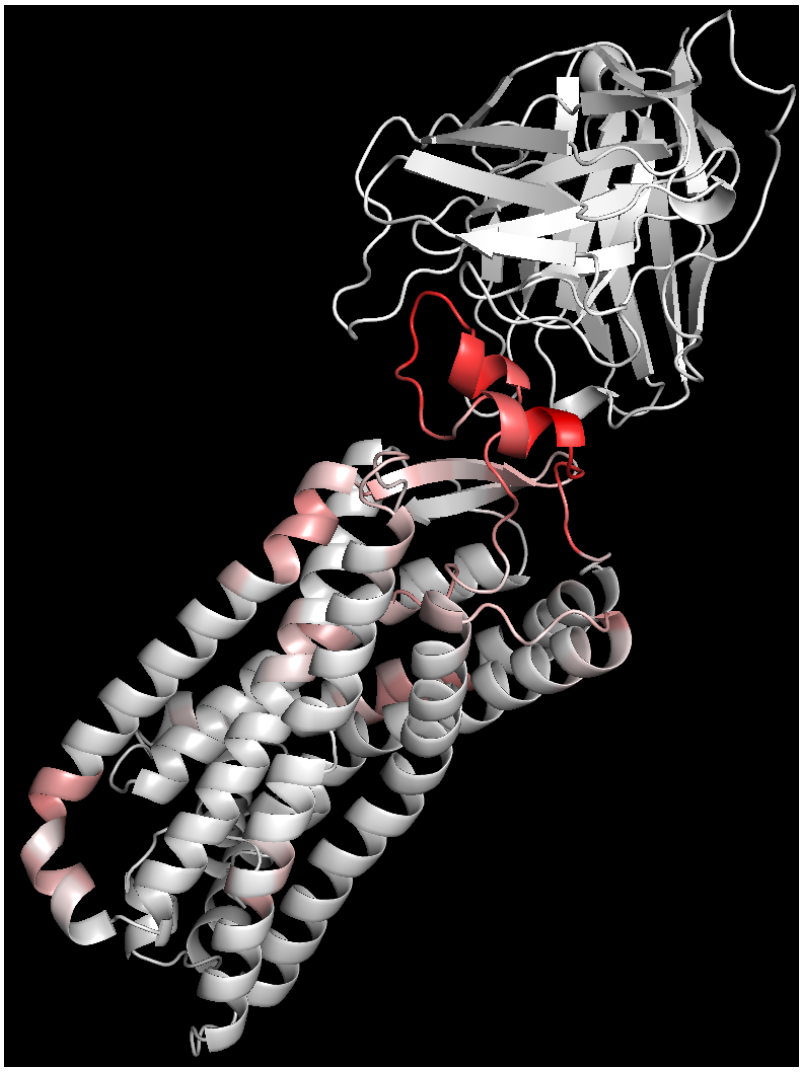
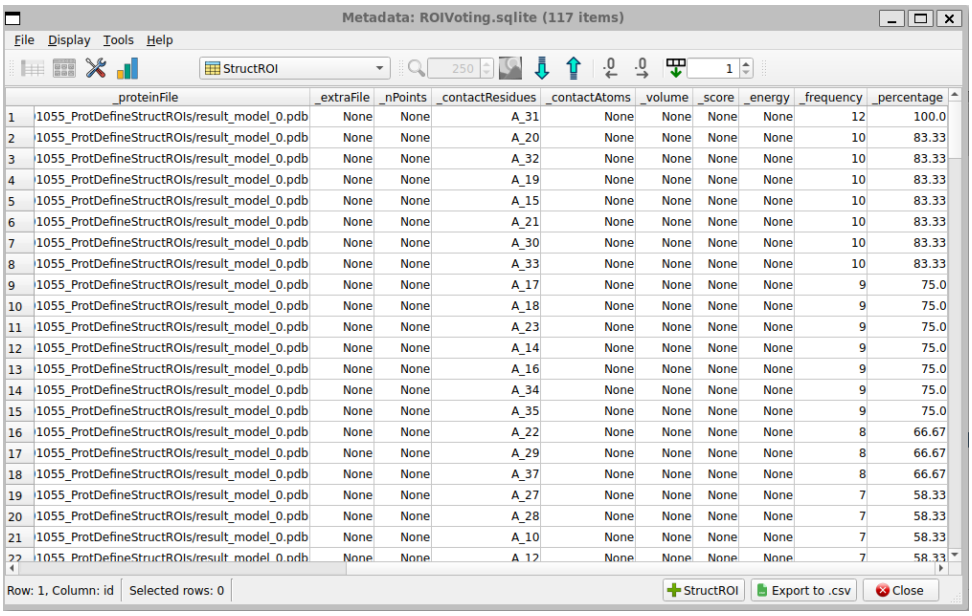


Fig. A.3. PyMOL white-to-red gradient view of CCR9A for antibody 91R. Color gradient from white (frequency = 0) to red (maximum = 12). The highest-frequency contact residue (F31; 12/15) appear in dark red in the extracellular NT domain. Compared to 92R, the high-frequency signal is more concentrated at F31, with moderate coloring extending across residues 14–37. Transmembrane and intracellular regions remain white. Own elaboration using Scipion-Chem [13] and PyMOL [16].



	_proteinFile	_extraFile	_nPoints	_contactResidues	_contactAtoms	_volume	_score	_energy	_frequency	_percentage
1	1055_ProtDefineStructROIs/result_model_0.pdb	None	None	A_31	None	None	None	None	12	100.0
2	1055_ProtDefineStructROIs/result_model_0.pdb	None	None	A_20	None	None	None	None	10	83.33
3	1055_ProtDefineStructROIs/result_model_0.pdb	None	None	A_32	None	None	None	None	10	83.33
4	1055_ProtDefineStructROIs/result_model_0.pdb	None	None	A_19	None	None	None	None	10	83.33
5	1055_ProtDefineStructROIs/result_model_0.pdb	None	None	A_15	None	None	None	None	10	83.33
6	1055_ProtDefineStructROIs/result_model_0.pdb	None	None	A_21	None	None	None	None	10	83.33
7	1055_ProtDefineStructROIs/result_model_0.pdb	None	None	A_30	None	None	None	None	10	83.33
8	1055_ProtDefineStructROIs/result_model_0.pdb	None	None	A_33	None	None	None	None	10	83.33
9	1055_ProtDefineStructROIs/result_model_0.pdb	None	None	A_17	None	None	None	None	9	75.0
10	1055_ProtDefineStructROIs/result_model_0.pdb	None	None	A_18	None	None	None	None	9	75.0
11	1055_ProtDefineStructROIs/result_model_0.pdb	None	None	A_23	None	None	None	None	9	75.0
12	1055_ProtDefineStructROIs/result_model_0.pdb	None	None	A_14	None	None	None	None	9	75.0
13	1055_ProtDefineStructROIs/result_model_0.pdb	None	None	A_16	None	None	None	None	9	75.0
14	1055_ProtDefineStructROIs/result_model_0.pdb	None	None	A_34	None	None	None	None	9	75.0
15	1055_ProtDefineStructROIs/result_model_0.pdb	None	None	A_35	None	None	None	None	9	75.0
16	1055_ProtDefineStructROIs/result_model_0.pdb	None	None	A_22	None	None	None	None	8	66.67
17	1055_ProtDefineStructROIs/result_model_0.pdb	None	None	A_29	None	None	None	None	8	66.67
18	1055_ProtDefineStructROIs/result_model_0.pdb	None	None	A_37	None	None	None	None	8	66.67
19	1055_ProtDefineStructROIs/result_model_0.pdb	None	None	A_27	None	None	None	None	7	58.33
20	1055_ProtDefineStructROIs/result_model_0.pdb	None	None	A_28	None	None	None	None	7	58.33
21	1055_ProtDefineStructROIs/result_model_0.pdb	None	None	A_10	None	None	None	None	7	58.33
22	1055_ProtDefineStructROIs/result_model_0.pdb	None	None	A_12	None	None	None	None	7	58.33

Fig. A.4. Contact frequency table for antibody 91R showing CCR9A residues ranked by frequency score. Own elaboration using Scipion-Chem [13].

A.2. Antibody 3C3

Table A.2. CCR9A RESIDUES WITH CONTACT FREQUENCY ≥8 OUT OF 15 MODELS FOR ANTIBODY 3C3, RANKED BY FREQUENCY.

Position	Residue	Frequency (/15)
31	F	12
32	N	12
33	F	11
27	D	10
30	N	10
190	T	10
28	Y	9
26	E	8
29	V	8
34	T	8

Within the same frequency, residues are listed in sequence order. Own elaboration.

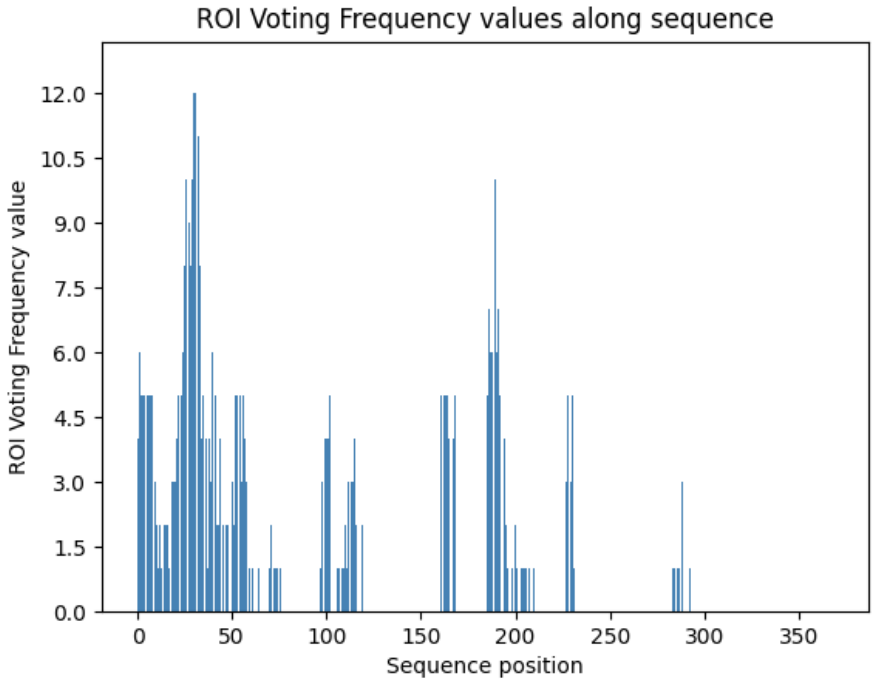


Fig. A.5. Contact frequency of each CCR9A residue across 15 predicted complex structures for antibody 3C3. Maximum frequency 12/15 at F31 and N32. Secondary peak at T190 reflects the spatial proximity of extracellular loop 2 in the predicted models. Own elaboration using Scipion-Chem [13].

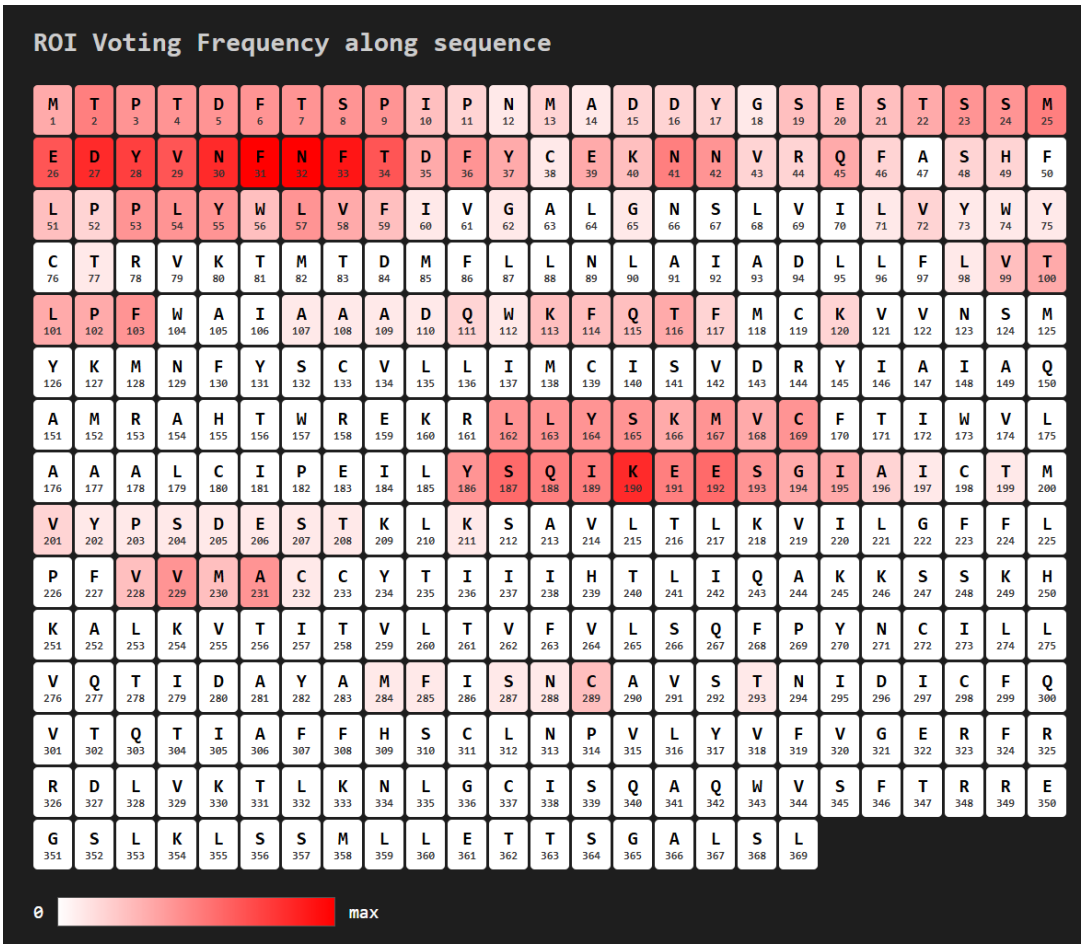


Fig. A.6. ROI Voting Frequency mapped along the CCR9A sequence for antibody 3C3. Color gradient from white (frequency = 0) to red (maximum = 12). The known epitope FTDFYC (aa 33–38) falls within the high-frequency region; T190 (EC2) appears as a secondary contact. Own elaboration using Scipion-Chem [13].

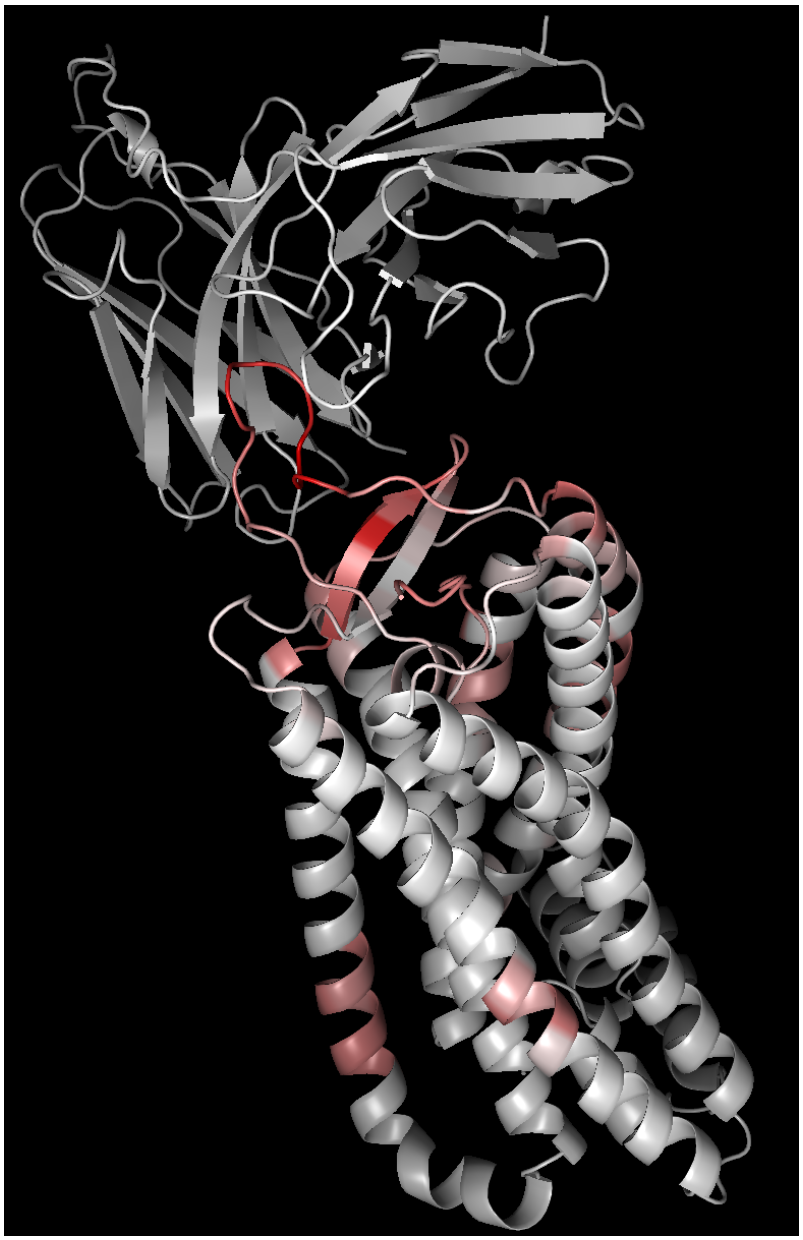
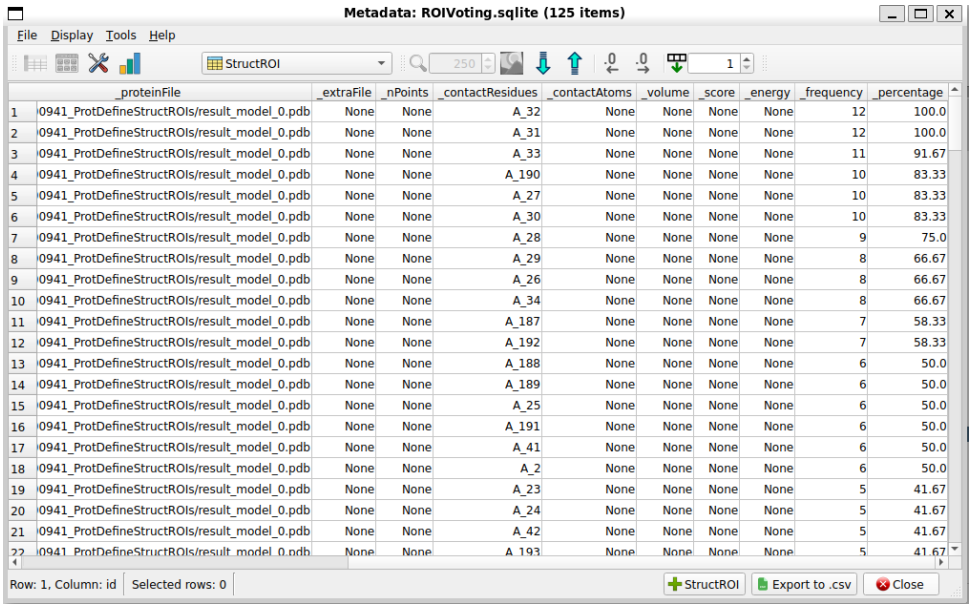


Fig. A.7. PyMOL representation of CCR9A (grey) with high-frequency contact residues for antibody 3C3 highlighted in red. The N-terminal cluster (F31, N32, F33) and the secondary ECL2 peak (T190) are both visible. The antibody variable domains are shown in dark grey. Own elaboration using Scipion-Chem [13] and PyMOL [16].



	_proteinFile	_extraFile	_nPoints	_contactResidues	_contactAtoms	_volume	_score	_energy	_frequency	_percentage
1	0941_ProtDefineStructROIs/result_model_0.pdb	None	None	A_32	None	None	None	None	12	100.0
2	0941_ProtDefineStructROIs/result_model_0.pdb	None	None	A_31	None	None	None	None	12	100.0
3	0941_ProtDefineStructROIs/result_model_0.pdb	None	None	A_33	None	None	None	None	11	91.67
4	0941_ProtDefineStructROIs/result_model_0.pdb	None	None	A_190	None	None	None	None	10	83.33
5	0941_ProtDefineStructROIs/result_model_0.pdb	None	None	A_27	None	None	None	None	10	83.33
6	0941_ProtDefineStructROIs/result_model_0.pdb	None	None	A_30	None	None	None	None	10	83.33
7	0941_ProtDefineStructROIs/result_model_0.pdb	None	None	A_28	None	None	None	None	9	75.0
8	0941_ProtDefineStructROIs/result_model_0.pdb	None	None	A_29	None	None	None	None	8	66.67
9	0941_ProtDefineStructROIs/result_model_0.pdb	None	None	A_26	None	None	None	None	8	66.67
10	0941_ProtDefineStructROIs/result_model_0.pdb	None	None	A_34	None	None	None	None	8	66.67
11	0941_ProtDefineStructROIs/result_model_0.pdb	None	None	A_187	None	None	None	None	7	58.33
12	0941_ProtDefineStructROIs/result_model_0.pdb	None	None	A_192	None	None	None	None	7	58.33
13	0941_ProtDefineStructROIs/result_model_0.pdb	None	None	A_188	None	None	None	None	6	50.0
14	0941_ProtDefineStructROIs/result_model_0.pdb	None	None	A_189	None	None	None	None	6	50.0
15	0941_ProtDefineStructROIs/result_model_0.pdb	None	None	A_25	None	None	None	None	6	50.0
16	0941_ProtDefineStructROIs/result_model_0.pdb	None	None	A_191	None	None	None	None	6	50.0
17	0941_ProtDefineStructROIs/result_model_0.pdb	None	None	A_41	None	None	None	None	6	50.0
18	0941_ProtDefineStructROIs/result_model_0.pdb	None	None	A_2	None	None	None	None	6	50.0
19	0941_ProtDefineStructROIs/result_model_0.pdb	None	None	A_23	None	None	None	None	5	41.67
20	0941_ProtDefineStructROIs/result_model_0.pdb	None	None	A_24	None	None	None	None	5	41.67
21	0941_ProtDefineStructROIs/result_model_0.pdb	None	None	A_42	None	None	None	None	5	41.67
22	0941_ProtDefineStructROIs/result_model_0.pdb	None	None	A_193	None	None	None	None	5	41.67

Fig. A.8. Contact frequency table for antibody 3C3 showing CCR9A residues ranked by frequency score. Own elaboration using Scipion-Chem [13].

A.3. Antibody 115

Table A.3. CCR9A RESIDUES WITH CONTACT FREQUENCY ≥8 OUT OF 15 MODELS FOR ANTIBODY 115, RANKED BY FREQUENCY.

Position	Residue	Frequency (/15)
30	N	14
32	N	13
33	F	13
27	D	12
28	Y	12
31	F	12
29	V	11
26	E	10
35	D	10
34	T	9
37	Y	9
20	E	8
21	S	8
22	T	8
23	S	8
25	M	8
36	F	8

Within the same frequency, residues are listed in sequence order. Own elaboration.

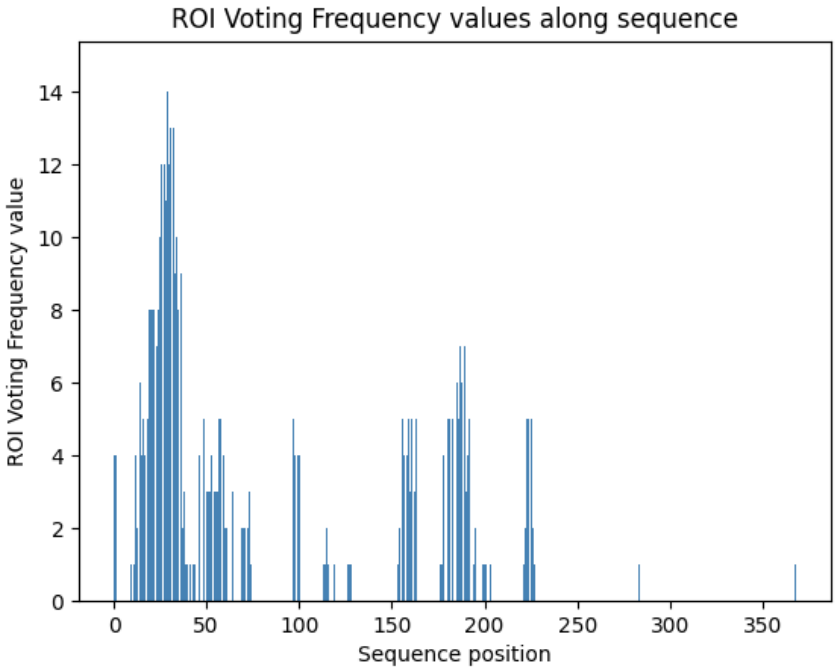


Fig. A.9. Contact frequency of each CCR9A residue across 15 predicted complex structures for antibody 115. Maximum frequency 14/15 at N30. The high-frequency region spans residues 20–37, encompassing the experimentally characterized epitope MEDYVNF (aa 25–31). Own elaboration using Scipion-Chem [13].

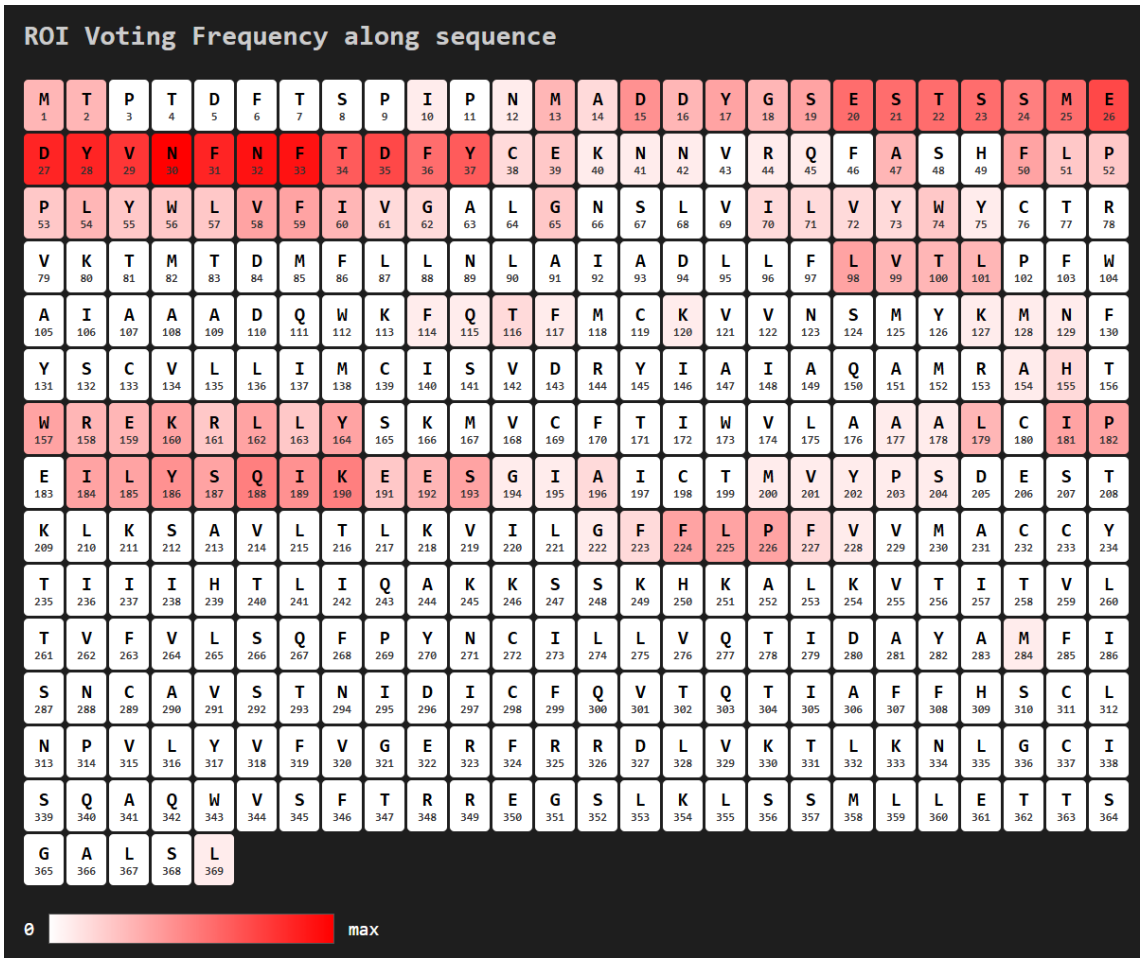


Fig. A.10. ROI Voting Frequency mapped along the CCR9A sequence for antibody 115. Color gradient from white (frequency = 0) to red (maximum = 14). The known epitope MEDYVNF (aa 25–31) falls within the high-frequency region. Own elaboration using Scipion-Chem [13].

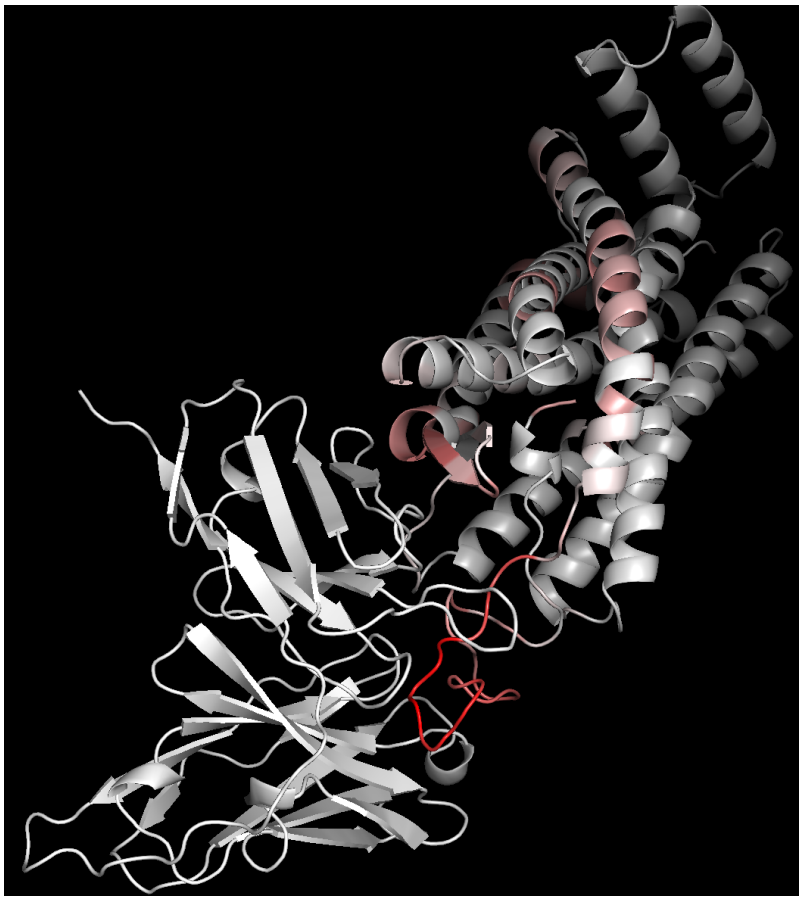
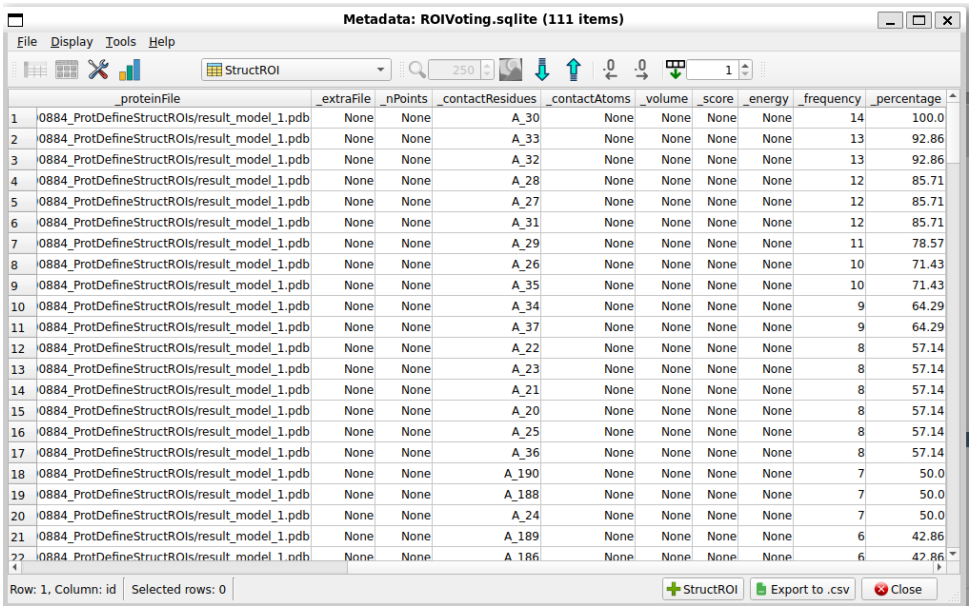


Fig. A.11. PyMOL white-to-red gradient view of CCR9A for antibody 115. The high-frequency contact region is concentrated in the extracellular N-terminal domain, with N30 (14/15) as the most intensely colored residue. The signal is notably clean, with minimal coloring outside the NT domain and no substantial contact predicted in the transmembrane helices or intracellular regions. Own elaboration using Scipion-Chem [13] and PyMOL [16].



	_proteinFile	_extraFile	_nPoints	_contactResidues	_contactAtoms	_volume	_score	_energy	_frequency	_percentage
1	0884_ProtDefineStructROIs/result_model_1.pdb	None	None	A_30	None	None	None	None	14	100.0
2	0884_ProtDefineStructROIs/result_model_1.pdb	None	None	A_33	None	None	None	None	13	92.86
3	0884_ProtDefineStructROIs/result_model_1.pdb	None	None	A_32	None	None	None	None	13	92.86
4	0884_ProtDefineStructROIs/result_model_1.pdb	None	None	A_28	None	None	None	None	12	85.71
5	0884_ProtDefineStructROIs/result_model_1.pdb	None	None	A_27	None	None	None	None	12	85.71
6	0884_ProtDefineStructROIs/result_model_1.pdb	None	None	A_31	None	None	None	None	12	85.71
7	0884_ProtDefineStructROIs/result_model_1.pdb	None	None	A_29	None	None	None	None	11	78.57
8	0884_ProtDefineStructROIs/result_model_1.pdb	None	None	A_26	None	None	None	None	10	71.43
9	0884_ProtDefineStructROIs/result_model_1.pdb	None	None	A_35	None	None	None	None	10	71.43
10	0884_ProtDefineStructROIs/result_model_1.pdb	None	None	A_34	None	None	None	None	9	64.29
11	0884_ProtDefineStructROIs/result_model_1.pdb	None	None	A_37	None	None	None	None	9	64.29
12	0884_ProtDefineStructROIs/result_model_1.pdb	None	None	A_22	None	None	None	None	8	57.14
13	0884_ProtDefineStructROIs/result_model_1.pdb	None	None	A_23	None	None	None	None	8	57.14
14	0884_ProtDefineStructROIs/result_model_1.pdb	None	None	A_21	None	None	None	None	8	57.14
15	0884_ProtDefineStructROIs/result_model_1.pdb	None	None	A_20	None	None	None	None	8	57.14
16	0884_ProtDefineStructROIs/result_model_1.pdb	None	None	A_25	None	None	None	None	8	57.14
17	0884_ProtDefineStructROIs/result_model_1.pdb	None	None	A_36	None	None	None	None	8	57.14
18	0884_ProtDefineStructROIs/result_model_1.pdb	None	None	A_190	None	None	None	None	7	50.0
19	0884_ProtDefineStructROIs/result_model_1.pdb	None	None	A_188	None	None	None	None	7	50.0
20	0884_ProtDefineStructROIs/result_model_1.pdb	None	None	A_24	None	None	None	None	7	50.0
21	0884_ProtDefineStructROIs/result_model_1.pdb	None	None	A_189	None	None	None	None	6	42.86
22	0884_ProtDefineStructROIs/result_model_1.pdb	None	None	A_186	None	None	None	None	6	42.86

Fig. A.12. Contact frequency table for antibody 115 showing CCR9A residues ranked by frequency score. Own elaboration using Scipion-Chem [13].

B. ANTIBODY SEQUENCES

This appendix provides the complete scFv sequences used in the structural predictions for all five anti-CCR9 antibodies. All constructs use the flexible $(G_4S)_3$ linker (GGGGSGGGGSGGGGS). The orientation used in the predictions is indicated for each antibody.

B.1. Antibody 92R

Orientation used in predictions: VL–linker–VH (scFv 1). 245 amino acids.

```
DVVMTQTPLSLPVSLGDQASISCRSSQSLVHSNGNTYLNWCLQRPGQSPKSLIYKVSNRF
SGVPDRFSGSGSGTDFTLKISRVEAEDLGVYFCSQSTHFPRTFGGGTKLEIKRGGGGSGGG
GSGGGGSEVKLEESGGGLVQPGGSMKLSCVASGFTFNKFWMNWVRQSPEKGLEWVAEIRLK
SNNYATHYAESVKGRFTISRDDSKSSVYLQMNNLRAEDTGIYYCASDGWFAYWGQGTLVTVSA
```

B.2. Antibody 91R

Orientation used in predictions: VL–linker–VH (scFv 1). 245 amino acids.

```
DVVMTQTPLSLPVSLGDQTSISCRSSQSLLHSNGNTYVQWYLRKPGQSPKLLIYKVSNRF
PGVPDRFSGSGSGTDFTFKISRVEAEDLGVYFCAQSTHVPRTFGGGTKLEIKRGGGGSGGG
GSGGGGSEVKLEDSGGGLVQPGRSMKLSCVASGFTFSNFWMNWVRQSPEKGLEWVAEIRLK
SNNYATHYAESVKGRFTISRDDSKSSVYLQMNNLRTEDTGIYYCTSDGWFAYWGQGTLVTVSA
```

B.3. Antibody 115

Orientation used in predictions: VL–linker–VH (scFv 1). 240 amino acids.

```
DVVMTQTPLSLTVSLGDQASISCRSSQSLVHSNGKTYLQWYLQKPGQSPKLLIYKVSNRF
SGVPDRFSGSGSGTDFTLKISRVEAEDLGVYFCAQSTHVWTFGGGTKLEIKRAGGGGSGGG
GSGGGGSQVQLKQSGPELVKPGTSVRVSCTASGYSFTDYIIYWVKQSHGRSLEWIGYIDPN
NYNTRYSQKFKGKATLTVDKSSTSAFMHLNSLTSDDSAVYYCARDVYWGQGTTLTVSS
```

B.4. Antibody 3C3

Orientation used in predictions: VL–linker–VH (scFv 1). 244 amino acids.

DVLMTQTPLSLPVSLGDQASISCRSSQSIVHSNGNTYLEWYLQKPGQSLKLLIYKVSNRF
SGVPDRFSGSGSGTDFTLKINRVEAEDLGVYYCFQGSLVPPTFGGGTKLEIKGGGGSGGGG
SGGGGSEVKLEESGGGLVQPGGSMKLSCVASGFTFNKFWMNWVRQSPEKGLEWVAEIRLK
DVQLQESGPDLVKPSQSLSLTCTVTGYSITSAYTWHWVRQFPGNRLEWMGYIHYSGRTSY
NPSLKSRISFTRDASRNQFFLQLNSVTTEDTATYYCAINRYYYFDVWGAGTTVTVSS

B.5. Antibody Ma-10

Orientation used in predictions: VH–linker–VL (scFv 2, corrected). 239 amino acids.

QVQLKESGPGLVQPSQTLSLTCTVSGFSVASYDMHWVRLPPGKGLEWMGIIWANGNTHYN
SGLKSRLSISRDTSKSQVFLKMNSLQTEDTAIYFCTRGGFAYWGQGTLVTVSSGGGGSGG
GGSGGGGSDVVMTQTPVSLSVSLGGQVSISCRSSQSLVHNNGNTYLSWYLQKPGQSPQLL
IYKVSNRFSGVSDRFSGSGSGTDFTLKISRVEPDDLGVYYCGQGTQYPTFGGGTKLELK

C. USE OF ARTIFICIAL INTELLIGENCE TOOLS

AI writing and coding assistants were used at various points during this thesis. Claude Sonnet 4.6 (Anthropic) helped with three main things: improving the clarity of written sections in English, writing and debugging Python scripts for file processing tasks and with technical issues with Scipion and LaTeX.

All scientific content: results, interpretations, and conclusions are the author's own work. The AI tools were used as assistants, not as sources of scientific content.

DECLARACIÓN DE USO DE INTELIGENCIA ARTIFICIAL GENERATIVA (IAG) EN EL TRABAJO DE FIN DE GRADO (TFG)

He usado IAG en mi TFG

Marca lo que corresponda:

SI	**NO**

Si has marcado SI, completa las siguientes 3 partes de este documento:

Parte 1: declaración sobre comportamiento legal, ético y responsable

Ten presente que el uso de IAG conlleva unos riesgos y puede generar una serie de consecuencias académicas graves: la evaluación de tu TFG puede verse comprometida si el uso de la IAG comporta la utilización de datos de carácter confidencial, materiales protegidos por derechos de autoría, o datos de carácter personal, y se hace sin cumplir las condiciones exigidas en cada caso (autorización de los interesados, autorización de los titulares, seguimiento de las instrucciones de la Universidad).

Pregunta	
1. En mi interacción con herramientas de IAG he facilitado **datos de carácter confidencial** contando siempre con la debida autorización de los interesados. La confidencialidad abarca toda información que una persona u organización desea proteger por razones legales, comerciales, de privacidad o estratégicas (como patentes o secretos comerciales).	
SÍ, he usado estos datos con la autorización de los interesados	NO, no he usado datos de carácter confidencial
2. En mi interacción con herramientas de IAG he facilitado **materiales protegidos por derechos de autoría** contando siempre con la autorización de los respectivos titulares.	
SÍ, he usado estos materiales con autorización de los titulares de derechos de autor; o bien sin ella porque se ajustan a una de las excepciones o límites que permite la ley: ● obra en dominio público ● obra licenciada (licencias Creative Commons)	NO, no he usado materiales protegidos por derechos de autoría

• uso de fragmentos con fines de investigación (derecho de cita)	
3. En mi interacción con herramientas de IAG he facilitado **datos de carácter personal** con la debida autorización de los interesados.	
SÍ, he usado estos datos con autorización de los interesados y conforme a las instrucciones contenidas en la guía aprobada por la Universidad	NO, no he usado datos de carácter personal
4. Mi utilización de la herramienta de IAG ha **respetado sus términos de uso**, así como los principios éticos esenciales, no orientándola de manera maliciosa a obtener un resultado inapropiado para el trabajo presentado, es decir, que produzca una impresión o conocimiento contrario a la realidad de los resultados obtenidos, que suplante mi propio trabajo o que pueda resultar en un perjuicio para las personas.	
SI	NO

Parte 2: declaración de uso técnico

Utiliza el siguiente modelo de declaración tantas veces como sea necesario, a fin de reflejar todos los tipos de iteración que has tenido con herramientas de IAG. Incluye un ejemplo por cada tipo de uso realizado donde se indique: *[Añade un ejemplo].*

Declaro haber hecho uso del sistema de IAG (Nombre del sistema/herramienta IAG y versión: ChatGPT, Gemini, Copilot…) **para:**

Documentación y redacción:

- *Soporte a la reflexión en relación con el desarrollo del trabajo: proceso iterativo de análisis de alternativas y enfoques utilizando la IAG*

SÍ. Por ejemplo: he pedido que me explique conceptos de biología computacional y me ha ayudado a estructurar algunas secciones del TFG.

- *Revisión o reescritura de párrafos redactados previamente*

He solicitado correcciones gramaticales y de estilo en secciones redactadas previamente en inglés. Por ejemplo: he pedido que se revisara la redacción de párrafos del State of the Art manteniendo el contenido científico original.

- *Búsqueda de información o respuesta a preguntas concretas*

He consultado terminología técnica en bioinformática e inglés científico. Por ejemplo: he preguntado el significado de términos como "proteinogenic amino acids" o sinónimos de palabras concretas.

- *Búsqueda de bibliografía*
- *Resumen de bibliografía consultada*
- *Traducción de texto consultados*

Desarrollar contenido específico

Se ha hecho uso de IAG como herramienta de soporte para el desarrollo del contenido específico del TFG, incluyendo:

- *Asistencia en el desarrollo de líneas de código (programación)*

He pedido ayuda para escribir scripts en Python para generación de figuras y procesamiento de datos y para resolver errores de compilación en LaTeX. Por ejemplo: he solicitado ayuda para corregir un script de matplotlib que generaba las figuras comparativas de los cinco anticuerpos.

- *Generación de esquemas, imágenes, audios o vídeos*

No se uso.

- *Procesos de optimización*

No se uso.

- *Tratamiento de datos: recogida, análisis, cruce de datos…*

No se uso.

- *Inspiración de ideas en el proceso creativo*

No se uso.

- *Otros usos vinculados a la generación de puntos concretos del desarrollo específico del trabajo*

Si crees necesario incluirlo, añade, en cada uno de los casos:

empleando el *prompt* …
(escribe la petición realizada a la IAG)
teniendo como interacción …
(describe qué interacción realizaste con la IAG tras la respuesta del prompt)

Ejemplo: ***Declaro*** *haber hecho uso del sistema de IAG* ***Chat GPT X.x*** *para* ***buscar información*** *empleando* ***el prompt: "Dime un ejemplo concreto que ilustre la aplicación de la propuesta urbanística de la Ciudad de los 15 minutos en España"*** *teniendo como interacción* ***la proposición del ejemplo concreto del distrito de Chamartín que se ha reflejado en la memoria****.*

Parte 3: reflexión sobre utilidad

Aporta una valoración personal (formato libre) sobre las fortalezas y debilidades que has identificado en el uso de herramientas de IAG en el desarrollo de tu trabajo. Menciona si te ha servido en el proceso de aprendizaje, o en el desarrollo o en la extracción de conclusiones de tu trabajo.

Claude ha sido útil principalmente como asistente de escritura y programación. Me ha ayudado a mejorar la claridad del inglés científico y a resolver errores técnicos de forma más rápida. No lo he usado para generar contenido científico ya que los resultados, interpretaciones e ideas son completamente míos.

D. REGULATORY FRAMEWORK

This thesis is entirely computational. No human subjects, biological samples, patient data, or personal information were involved at any stage. Everything is based on publicly available protein sequences, published antibody data and open-access prediction tools.

Given that, most of the regulatory frameworks relevant to biomedical research do not apply here. For completeness, the table below lists the main ones and explains why each falls outside the scope of this work.

Regulation	**Applicability**
Declaration of Helsinki (2024)	Not applicable. No human subjects involved.
GDPR 2016/679 and Ley Orgánica 03/2018	Not applicable. No personal data collected or processed.
Ley 41/2002 de autonomía del paciente	Not applicable. No clinical data or patient records used.
Ley 14/2007 de Investigación Biomédica	Not applicable. No biological samples or human-derived material used.
Reglamento (UE) 2017/745 and RD 192/2023	Not applicable. No medical device developed.

No IRB approval was required for this work.

E. CODE REPOSITORY

The scripts developed during this thesis are available at:

https://github.com/Irenesanch/scipion-chem

The repository includes Python scripts for PDB file processing, Scipion project files, and the scFv sequences for all five anti-CCR9 antibodies.