Degree in Medicine and Surgery

BSc in Biomedical Engineering

Decoding the CHI3L1-RAGE Immunosuppressive Complex: Integrative Structural and Coevolutionary Insights into HER2+ Breast Cancer Resistance

Luca Maria Menga

Matriculation n. 02337

Lorenzo Profeta

Matriculation n. 01960

Advisor

Prof. Maria Rescigno

Co-advisors

Prof. Maria Laura Costantino, Dr. Luigi Angelo Scietti, Dr. Marina Mapelli, Dr. Crissy Lynette Tarver

Academic Year 2024-2025

Alle nostre mamme, Simona e Natalie

A tutte le persone importanti che abbiamo incontrato nella nostra vita

«Veniet tempus quo ista quae nunc latent in lucem dies extrahat et longioris aeui diligentia; ad inquisitionem tantorum aetas una non sufficit, ut tota caelo vacet [...] Veniet tempus quo posteri nostri tam aperta nos nescisse mirentur».

"*The time will come when diligent research over long periods will bring to light things which now lie hidden. A single lifetime, even though entirely devoted to the sky, would not be enough for the investigation of so vast a subject... There will come a time when our descendants will be amazed that we did not know things that are so plain to them."*

L. Annaei Senecae
Quaestiones Naturales, Liber VII *De Cometis* [25, 4–5]

Index

Abstract 5

1. Introduction 6

1.1 A Complex Biological System 6

1.1.1 Clinical Background 6

1.1.2 HER2+ Breast Cancer 9

1.1.3 NK cell Immunosuppression as a Novel Mechanism for Resistance in HER2+ Breast Cancer 15

1.1.4 Relevance of the CHI3L1-RAGE Axis in Other Pathologies 18

1.1.5 Significance of the CHI3L1-RAGE Axis and its Dysregulation in Pathology: A Node of Inflammation and Immune Surveillance 20

1.2 Conceptual Problem: Making Complexity Reachable 21

1.3 Biological Problem: Converting a Ligand-Receptor Checkpoint into a Therapeutic Handle 21

1.3.1 Understanding Structure as a Path to Intervention 21

1.3.2. RAGE: A Master Sensor and Transducer of Danger and Inflammation 22

1.3.3 CHI3L1 (YKL-40): An Immune Regulator with a Long History and Dark Side 31

1.3.4 Evidence That RAGE Is a Functional Receptor for CHI3L1 42

1.3.5 *CHI3L1: non-self, self, a mysterious motif* 43

1.3.6 Chito-oligosaccharides & caffeine, peptidoglycan, proteoglycans 45

1.4 Why a Clear Picture Is Still Missing 47

1.4.1 Multifunctional partners 47

1.4.2 Structural silence 47

1.4.3 Non-canonical affinity logic 47

1.5 Working Hypothesis and Experimental Awareness 47

1.6 Experimental Strategy and Planned Approach 49

1.7 Resolving the Nanoscopic World 50

1.7.1 From Sequence to Shape: Biology's Most Elegant Puzzle 50

1.7.2 The Folding Funnel: Physics Guides the Fold 51

1.7.3 The Funneling of an Idea 53

1.7.4 The Experimental Revelation: Structure by X-Ray Crystallography 53

1.7.5 Protein Crystallization and X-Ray Crystallography 54

1.7.6 The Structural Gap: A New Paradox 57

1.7.7 The Data-Driven Shift: An Evolutionary Rosetta Stone....59
1.7.8 The Co-evolutionary Clue....60
1.7.9 AlphaFold: Operationalizing the Folding Funnel....60
1.7.10 A Framework for Confidence: Interpreting the AlphaFold Output [269]....62
1.7.11 Impact and Synergy....64
1.7.12 The Computational Revolution: Folding by Co-evolution - Known Frontiers and Limitations....64
1.8 From Structural Insight to Therapeutic Rationale....65
2. Materials & Methods....66
2.1 Materials....66
2.1.1 Protein Expression and Purification: hCHI3L1-His and hRAGE(VC1C2)-His....66
2.1.2 Peptide-N-Glycosidase F Enzyme....70
2.1.3 Antibodies for Enzyme-Linked Immunosorbent Assays....71
2.1.4 Chito-oligosaccharides and Caffeine....71
2.1.5 Fluorescent 6FAM-Conjugated Peptide....72
2.2 Biochemical Studies....73
2.2.1 Protein Concentration Determination....73
2.2.2 Sodium Dodecyl Sulfate-Polyacrylamide Gel Electrophoresis....73
2.2.3 Size Exclusion Chromatography....75
2.2.4 Static Light Scattering....76
2.2.5 Deglycosylation....77
2.2.6 Enzyme-Linked Immunosorbent Assays....78
2.2.7 Mass Photometry....79
2.3 Structural Investigation Through X-Ray Crystallography....80
2.3.1 Crystallization Screening at Stanford Synchrotron Radiation Lightsource....80
2.3.2 Crystallization Screening at the European Institute of Oncology....83
2.4 AI Modeling....84
2.4.1 ColabFold....84
2.4.2 Code Availability....90
2.5 Protein Structures Analysis and Visualization....90
3. Results....91
3.1 Binding Studies....91
3.1.1 CHI3L1 Binds RAGE in the nM Affinity....91
3.1.2 Mildly Deglycosylated Proteins Display a Stronger Binding....92

3.1.3 Tetraacetyl-chitotetraose and Caffeine Do Not Inhibit CHI3L1-RAGE Binding ..93
3.2 Single Proteins Characterization 94
3.2.1 CHI3L1 Solubility Is Enhanced by High Ionic Strength 94
3.2.2 Either Protein Is Monomeric in (PBS, 0.5 M NaCl) Solution 95
3.2.3 N-Glycosylations on RAGE Are Highly Heterogeneous, Even More After Partial Deglycosylation 97
3.3 Protein Complex Characterization 100
3.3.1 CHI3L1 and RAGE Elute as a Single Peak in the SEC Column 100
3.3.2 The Gel-filtered Peak Contains Both CHI3L1 and RAGE 101
3.3.3 Progressive Dilution of the CHI3L1-RAGE Complex Determines its Step-wise Disassembly in a Concentration-dependent Fashion 102
3.4 Crystallography Studies 104
3.4.1 The Crystal Diffraction Spoke, but not Loud Enough to Be Understood 104
3.4.2 CHI3L1 Crystallizes in the PEGRx HT E7 Solution and Other 62 Conditions ... 105
3.5 Prediction of the CHI3L1-RAGE Interface Using AlphaFold Multimer v2.3 in ColabFold 108
3.5.1 Parameters Definition 109
3.5.2 An Efficient Approach: Optimizing Parameters Across 128 Seeds 113
3.5.3 AlphaFold Multimer v2.3 Predicts the Structure of the CHI3L1-RAGE VC1 Complex with an Accuracy of 88% for the Overall Fold and 92% for the Protein Interface 121
3.5.4 Predicted Binding Mode of RAGE VC1C2 to CHI3L1 Mirrors That of RAGE VC1 127
3.5.5 AlphaFold-Multimer v2.3 Is Also Capable of Predicting the Structure of the S100A8/A9-RAGE VC1 Complex 129
4. Discussion 133
4.1 The Architecture of an Innate Immunity Checkpoint 133
4.1.1 The Second Handshake: A Hypothesis for the Mysterious Motif 134
4.2 A Mechanism of Disruption 134
4.3 An Echo in the Proteome: A Checkpoint Forged in Deep Time 136
4.3.1 The Long Apprenticeship of CHI3L1 136
4.3.2 A Modern Receptor's Rapid Rise 139
4.3.3 The Evolutionary Convergence 144
4.4 When the Ancient Co-evolves with the New: Evolutionary Logic of the CHI3L1-RAGE Checkpoint 144

4.4.1 The Ancestral Tool: A Chitinase 144
4.4.2 The First Co-option: The Sensor 145
4.4.3 The Second Co-option: The Host Integrator 145
4.4.4 The Final Co-option: The Checkpoint 145
4.4.5 A Doubly-Constrained Groove 146
4.5 Limitations and Methodological Considerations 147
4.6 Future Perspectives 147
5. Conclusion 149
Acknowledgements 150
Bibliography 151
Appendix: Preliminary Analysis for ColabFold Prediction 163
1. Screening 128 seeds with default parameters and using custom templates (Colab's A100 GPU) 163
2. Parameters optimization on specific seeds 164

Abstract

Introduction: In HER2$^+$ breast cancer, resistance often evolves faster than intervention. Amid this complexity, we identified a structurally defined immune checkpoint: the interaction between chitinase-3-like protein 1 (CHI3L1) and the receptor for advanced glycation end-products (RAGE). This thesis rests on a central proposition: that structure *is* function, and both are outcomes of molecular co-evolution. Our aim was not only to characterize this interaction, but to resolve its architecture and logic as a basis for therapeutic intervention.

Methods: To that end, we designed an integrative strategy that reflects the multidimensional nature of such a system. Biophysical measurements quantified affinity and explored stoichiometry and glycosylations; crystallographic experiments, though failing to capture the complex, successfully resolved CHI3L1's apo structure. To close the gap, we implemented a deep exploration of AlphaFold-Multimer predictions, systematically optimized across seeds and alignments.

Results: The result was a high-confidence atomic model of the CHI3L1-RAGE complex (multi = 0.880; actifpTM = 0.920), revealing a chemically precise interface: the positively charged FG-loop of RAGE inserts into CHI3L1's conserved glycan-binding groove, stabilized by specific cation-π contacts. This interaction induces a conformational shift in Trp99, a gatekeeper residue already known to reorient upon glycan binding.

Interpretation: The model suggests a mechanistically elegant form of antagonism: CHI3L1 acts as a structural blocker of the oligomerization of RAGE required for downstream cytotoxic signaling. Rather than competing for a site, it locks the receptor in a non-signaling conformation. This architecture also exposes a deeper evolutionary pattern: RAGE, a modern innate immune sensor, appears to have co-opted an ancient microbial-recognition module for internal regulation. The dual selective pressure on CHI3L1's groove, binding both non-self glycans and self immune partners, explains its conservation across deep evolutionary time.

Conclusion: This work transforms a diffuse resistance mechanism into a structurally tractable target. It provides a template for rational inhibitor design, and a conceptual framework for decoding immune regulation through structure. Within complex biological systems, architecture can be the clearest language. Evolution has already written the code. Our task is to read it.

1. Introduction

1.1 A Complex Biological System

1.1.1 Clinical Background

Breast cancer remains an arduous challenge in oncology, both for its global prevalence and due to its biological heterogeneity, involving also complex modulation of the immune system. [1]

Epidemiology
With ~2.3 million new cases annually as of 2022, and estimates rising up to 3.2 million by 2050, breast cancer is the most common malignancy among women and the second most frequent cancer worldwide, after lung cancer. [2, 3] Despite advances in early detection and systemic therapy that have improved survival, breast cancer remains the second leading cause of cancer-related death in women with 670,000 deaths in 2022 (1.1 million by 2050).

In Italy, breast cancer is the most frequently diagnosed cancer among women, with 55,000 new cases annually and over 800,000 women affected, according to 2022 data. [4]

The burden is not confined to developed nations; incidence is rising globally, underlining the urgency of coordinated international research and healthcare strategies. [2] Figures 1a and 1b show the worldwide incidence and mortality per country in 2022, according to WHO. [5] In terms of deaths, breast cancer disproportionately affects individuals in low- and middle-income countries, with 5-year survival rates in high-income countries exceeding 90%, compared with 66% in India and 40% in South Africa. For this reason, the World Health Organization (WHO) established a Global Breast Cancer Initiative (GBCI) in 2021, guiding governments to strengthen early detection, timely diagnosis and comprehensive management. The goal of GBCI is to reduce breast cancer by 2.5% per year, which would save 2.5 million lives over 20 years. [6]

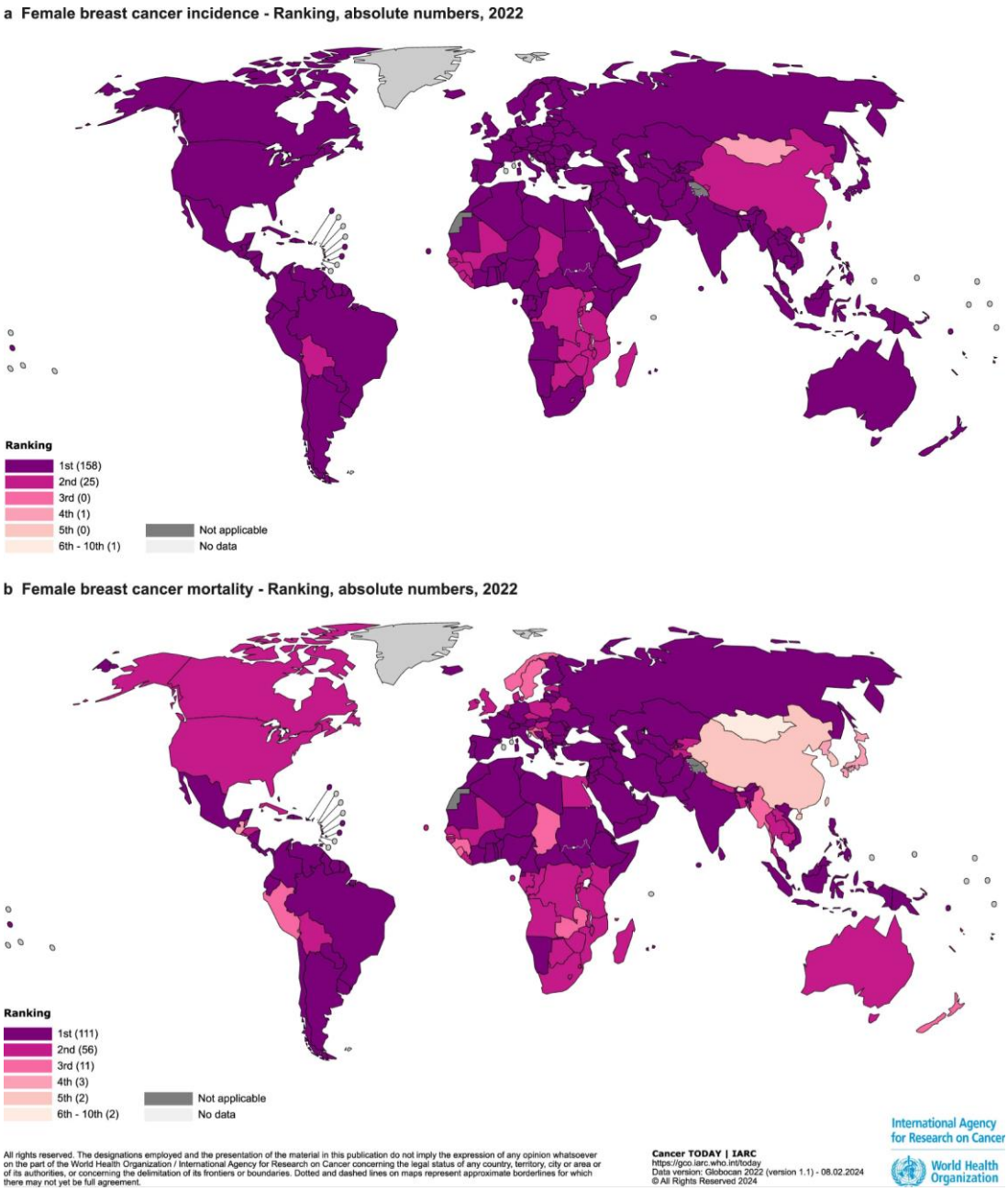


Figure 1. *Worldwide female breast cancer incidence and mortality.*
In 2022, female breast cancer was the most common type of malignancy in 158 countries and the deadliest in 111 of them.
Image reproduced from World Health Organization (WHO), International Agency for Research on Cancer. Copyright 2025 IARC/WHO.

Etiology

The etiology of breast cancer is multifactorial, meaning it involves multiple contributing factors, including genetic, hormonal, lifestyle, and environmental influences.

5-10% of cases can be attributed to inherited germline mutations, primarily referable to BRCA1 and BRCA2 genes, but also to TP53, PALB2 and CHEK2. [7, 8]

Prolonged exposure to endogenous estrogen represents another risk factor, as in early menarche, late menopause, nulliparity, or delayed first childbirth. At the same time, exogenous estrogen, usually in the form of hormonal replacement therapy, increases the risk in the long

term. Lifestyle factors are often modifiable: high body mass index in postmenopausal women, alcohol consumption, physical inactivity and poor diet. Protective factors include weight control, regular exercise and breastfeeding. [9]

An additional risk factor is environmental exposure to ionizing radiation and circadian rhythm disruption due to night shift work. [10, 11]

From a phenotypic standpoint, breast density at mammography has also been correlated to breast cancer incidence. [9]

Clinical-Molecular Classification and Treatment

Clinically, breast cancer is stratified into three distinct subtypes based on the expression of estrogen receptor (ER), progesterone receptor (PgR) and human epidermal growth factor receptor 2 (HER2). These subtypes, hormone receptor positive (ER+ and/or PgR+), HER2 positive, and triple-negative breast cancer (TNBC), are not merely diagnostic labels; they correspond to distinct biological identities shaped by different evolutionary paths, and thus predict clinical behavior and guide treatment decisions. [12]

At the same time, molecularly, breast cancer is stratified based on gene expression profiling into: Luminal A (ER+/PgR+, low Ki-67, HER2–), Luminal B (ER+, higher Ki-67, HER2+/–), HER2-enriched (HER2+, ER–/PR–), and Basal-like (expresses basal cytokeratins and high Ki-67). The clinical subsets contain each of these molecular subtypes, as depicted in Fig. 2. [13]

Hormone receptor positive (HR+) tumors are mainly divided into Luminal A and Luminal B molecular subtypes. Luminal A cancers are hormone-driven and slow-growing. These tumors are typically low-grade, with excellent prognosis. They originate from luminal epithelial cells and depend on estrogen signaling for survival, making them highly responsive to endocrine therapy. [14] However, resistance to anti-estrogen therapy in HR+ breast cancer, often driven by estrogen receptor 1 (ESR1) mutations emerging under treatment pressure, highlights the disease's adaptive nature. [15] On the other hand, Luminal B tumors also express hormone receptors but are more proliferative, often with higher Ki-67 or low progesterone expression. They are more aggressive than Luminal A and may require both endocrine and chemotherapy. Approximately 30% of luminal B tumors are HER2 positive, which require a distinct treatment approach involving HER2-targeted therapy. [16]

HER2-enriched cancers are defined by amplification of the HER2 gene, which drives growth through potent signaling cascades like PI3K/AKT and MAPK. HER2 typically requires HER3 to activate PI3K/Akt signaling, but at high expression levels, HER2 can bypass this dependency by amplifying its own functional output and acquiring progressive functional gains. [17]

While these tumors tend to be biologically aggressive, HER2-targeted therapies have significantly altered their clinical management and improved patient outcomes. [18]

Triple-Negative Breast Cancer (TNBC) lacks ER, PgR, and HER2 expression. It represents a biologically heterogeneous group, often arising from basal-like cells or BRCA1-deficient pathways. The clinical TNBC subtype is often used as a proxy for the molecular basal-like subtype, though they are not biologically identical: basal-like tumors are defined by specific gene expression profiles, including high levels of basal cytokeratins. Both subtypes are aggressive and more prevalent among women of African descent. [19] TNBCs are genomically unstable, highly proliferative, and often immune-infiltrated, making them responsive to chemotherapy and, increasingly, immunotherapy or PARP inhibitors. [20]

Treatment strategies for breast cancer also depend on disease stage. For early-stage disease, surgery remains the cornerstone, often supported by adjuvant or neoadjuvant systemic therapy to reduce recurrence. In advanced or metastatic settings, systemic treatments aim to prolong survival and maintain quality of life, even when cure is not achievable. [21]

In this study we will focus on the HER2+ breast cancer subtype, more specifically on its mechanisms of resistance.

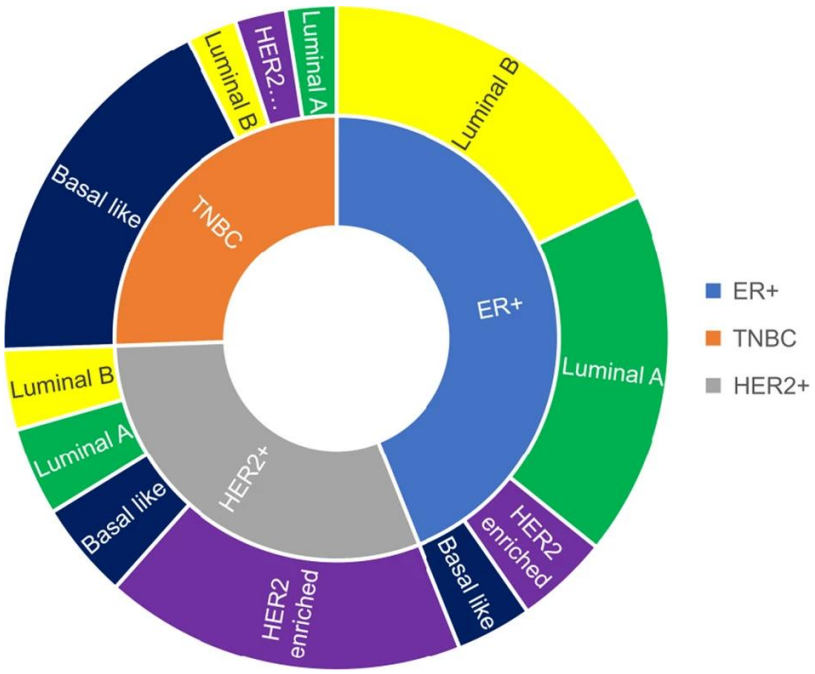


Figure 2. *Clinical and molecular subtypes of breast cancer.*
ER+: Endocrine receptor positive; HER2+: Human Epidermal Growth Factor Receptor 2 positive; TNBC: triple negative breast cancer.
Image reproduced from Zagami, Carey, 2022. [22]

1.1.2 HER2+ Breast Cancer

HER2+ breast cancer represents a clinical and biological subtype defined by an amplification of the ERBB2 or HER2 gene. To understand the relevance of this gene, we can look at its etymology. ERBB2 stands for *v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 2*, defining it as the human homologous of an oncogene found in the avian erythroblastosis virus, a retrovirus that causes tumors in birds. Scientists discovered in 1983-84 that this v-erbB oncogene encodes a truncated version of a normal avian gene (erbB) that drives cell proliferation and cancer. [23] On the other hand, HER2 refers directly to the protein expressed by the gene, the human epidermal growth factor receptor 2. This is a receptor tyrosine kinase belonging to the EGFR (epidermal growth factor receptor) family. [24] ERBB2 is located on chromosome 17q12 and regulates cell growth and survival. [25]

It turns out that HER2 is overexpressed and amplified in around 20% of breast cancers and gives increased aggressiveness. The cells become addicted to HER2-driven signaling via the PI3K/AKT (Phosphoinositide 3-Kinase/Albanian K strain rodents Transforming) and MAPK

(Mitogen-Activated Protein Kinase) pathways, proliferating without control, resisting apoptosis, and promoting metastases. [26]

The key characteristic of HER2 is that it is constitutively active when overexpressed, even in the absence of an external stimulus. [27] Differently from other EGFR family members, HER 2 lacks a ligand-binding site, and is normally complexed with inhibitory chaperones: HSP90, CDC37 and ERBIN. When it is overexpressed, HER2 saturates the chaperones and begins to homo- or hetero-dimerize, and its cytoplasmic tails auto-phosphorylate, triggering the signaling cascades. [28]

Even though HER2 homodimerization becomes significant at high expression levels, reducing the need for co-receptors, ligand-dependent heterodimerization with other ERBB family members, especially HER3, is still very potent and relevant. [17] With HER2, HER3 engages the PI3K/AKT signaling pathway. Instead, homodimeric HER2 can activate the MAPK pathway, via recruitment of adaptor proteins GRB2 and SOS, which activate in cascade RAS, RAF, MEK, and ERK. These pathways promote cellular proliferation and confer resistance to apoptosis.

By itself, HER2 is not strongly able to activate PI3K/AKT and therefore, especially in the early tumor stages, it relies more on heterodimerization with HER3. HER3, catalytically less active, contains multiple high-affinity binding motifs (YXXM) for the p85 subunit of PI3K, thus largely amplifying the pathway. However, at very high expression levels, HER2 is able to directly activate PI3K by self-phosphorylating weak-affinity sites like Y1139, and this pathway becomes functionally significant. In case of massive overexpression, the contribution of these weak sites is numerically overwhelming, even though just a minor percentage of HER2 molecules bind PI3K. [17] Actually, the more HER2 is overexpressed, the more the tumor may reduce its dependency on HER3. [29]

The HER2 and HER3 downstream signaling pathways can be visualized in Fig. 3.

Thus, the mechanism behind HER2's oncogenicity is quantitative, amplification- rather than mutation-based, and demonstrates that overexpression can endow weak signals with oncogenic potency. [17]

It becomes clear that HER2 is not just the key mechanism of the tumor for attack, but a mechanism of dependency and addiction for the tumor that can be targeted for therapy.

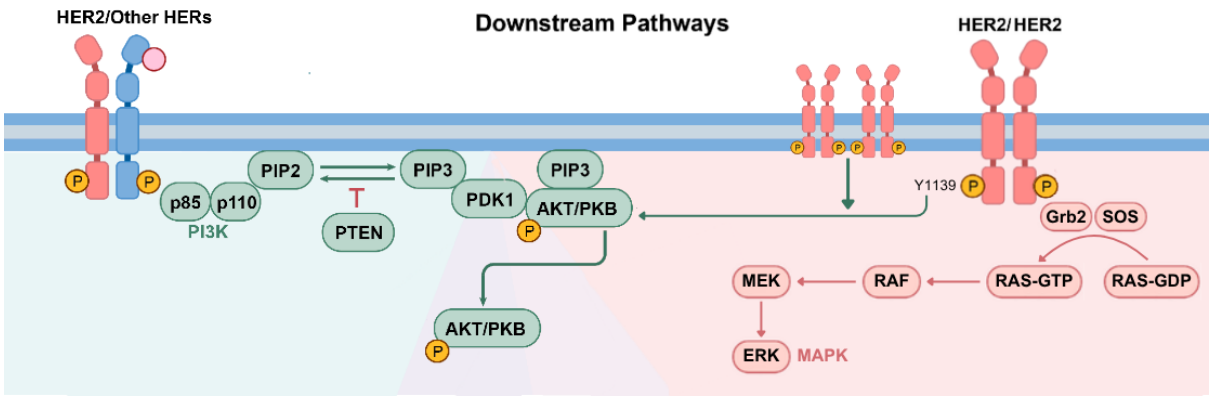


Figure 3. *HER2 downstream signalling pathways.*

In green, the PI3K/AKT pathway downstream to HER2 heterodimerization with other HERs (ligand-dependent) or to HER2 overexpression and homodimerization (ligand-independent) is visualized. In pink, the constitutively active (under HER2 overexpression) MAPK pathway downstream to HER2 homodimerization is visualized. The effects of these pathways include uncontrolled proliferation, resistance to apoptosis and promotion of metastasis.
Image from Zhong, Zhou, et al., 2024, has been adapted to create this figure. [30] Copyright 2024 by the authors. Licensee MDPI, Basel, Switzerland. [31]

HER2-Targeted Therapies: A Paradigm Shift in Tumor Treatment

And precisely HER2 was the target of trastuzumab, a monoclonal antibody developed in the early 1990s and approved by the FDA in 1998. [32] Its name follows the international monoclonal antibody naming convention: "tras-" (arbitrary prefix), "-tu-" (target: tumor), "-zu-" (humanized), "-mab" (monoclonal antibody). By binding to HER2, trastuzumab inhibits receptor homodimerization, induces internalization, and activates NK-cell-mediated antibody-dependent cellular cytotoxicity (ADCC). [33]

Clinically, trastuzumab was a breakthrough. In early-stage disease, it reduces recurrence and mortality by about a third when combined with adjuvant chemotherapy. [34] In metastatic settings, when associated with chemotherapy it prolongs median survival and progression-free survival, and reduces the risk of death by 20% when compared to chemotherapy alone. [35]

However, trastuzumab left a critical gap: while it binds domain IV of HER2, effectively inhibiting homodimerization, it does not block heterodimerization, particularly with HER3. In 1990, Genentech researchers isolated a monoclonal antibody from mice immunized with human HER2-overexpressing cells. [36] This antibody became later known as pertuzumab, following a similar nomenclature to trastuzumab, and was approved by FDA in 2012. [37] By binding domain II, pertuzumab inhibits ligand-induced heterodimerization with HER3 and other receptors of the ERBB family. [38] Furthermore, similarly to trastuzumab, also pertuzumab mediates ADCC. [39]

The additional and synergistic effect of pertuzumab was famously reported with the CLEOPATRA trial, which showed a jump in median overall survival of HER2+ metastatic breast cancer patients from 40.8 to 57.1 months, when complemented to trastuzumab and chemotherapeutic docetaxel. [40] In the neoadjuvant setting and in a similar combination, pertuzumab improves the pathological complete response rate. [41] Finally, also in early-stage HER2+ breast cancer, in the adjuvant setting, pertuzumab showed a reduction of 17% in the risk of death. [42]

Nonetheless, even the combination of trastuzumab, pertuzumab, and chemotherapy is not always curative. In the neoadjuvant setting, between 39 to 61% of patients fail to achieve a pathological complete response; in the adjuvant setting, approximately 6% still relapse within 3 years, rising to nearly 13% in 10 years; [43-45, 41] and in the metastatic setting, around 50% of patients experience disease progression or death within 1.5 years. [46] The majority of patients who initially respond to trastuzumab eventually develop resistance within a year, and half of those responding to the dual targeted therapy with chemotherapy acquire resistance in 20.2 months. [47, 48]

Besides resistance, HER2 heterogeneity or low expression, and difficulty to reach CNS metastases with the antibodies, due to the blood-brain barrier, limit the outcomes of this dual targeted treatment.

Thus, new therapeutic strategies were developed, leveraging targeted antibodies. In 2013, an antibody-drug conjugate (ADC), trastuzumab-derivative of maytansine 1 (T-DM1) or trastuzumab-emtansine, was approved by the FDA for metastatic HER2+ breast cancer patients, and later expanded in 2019 to the adjuvant setting. [49, 50] DM1 is a synthetic, ultra-potent microtubule inhibitory molecule derived from maytansine, a natural chemical weapon synthesized by endophytes (bacterial symbionts) in the vulnerable roots of the plants of the *Maytenus* species. This molecule is intended as a defense from herbivores or pathogens, and evolutionarily speaking is the result of a true ecological arms race, where the host and the endophyte develop a protection or deterrent strategy against predators, which in turn develop resistance to the toxins. It is no coincidence that various *Maytenus* plant species have been used in African, Chinese, and South American traditional medicine with potential effects against cancer, parasites, gastric ulcers, and inflammatory conditions. [51, 52] By chemically engineering this natural molecule and conjugating it with trastuzumab, T-DM1 now can provoke selective cytotoxicity to breast cancer cells, recognizing HER2.

Another instance of this synergy is trastuzumab deruxtecan (T-DXd), approved by the FDA in 2019 and now demonstrated superior to T-DM1 as second-line in metastatic breast cancer. [53, 54] Deruxtecan is a high-potency topoisomerase I inhibitor: it blocks the enzyme that during DNA replication and transcription resolves the torsional strain due to the opening of the double helix, by cutting one strand and later resealing it. By inhibiting this topoisomerase I, deruxtecan generates permanent breaks in DNA. Compared to T-DM1, T-Dxd has lysable linkers and has a greater drug-to-antibody ratio (8:1). [55] Furthermore, thanks to its membrane-permeability, it allows a bystander effect, killing also nearby HER2-low cancer cells.

Another reason for patient resistance to treatment with trastuzumab is the variability in immune system involvement, particularly in cases of polymorphisms in the Fcγ receptor (FcγR) IIIa, or CD16A, of NK cells. The number of immune cells in the tumor microenvironment could also modulate the ADCC activity, which is why combinations of anti-HER2 treatments and immunotherapies are being evaluated in several trials, with little success so far. [55] Moreover, engagement of immune cells that recognize trastuzumab induces downregulation of HER2. [56] For these reasons, in order to maximize ADCC activity, a new agent was developed and approved by the FDA in 2020, margetuximab. [57] Margetuximab is a chimeric (-xi-) antibody, containing the same antigen-binding region of the murine precursor of trastuzumab, and human IgG1 Fc regions engineered at five amino acid sites, in order to increase affinity for the NK cells' activating receptor FcγRIIIa and decrease affinity for the inhibitory FcγRIIb (or CD32B). [58, 55] The increased affinity is especially beneficial in case of low-affinity variants, such as the CD16A-158F allele, a co-dominant allele variant that may be carried by more than 80-85% of the population. [59, 58] Notably, margetuximab binds the 158F variant with higher affinity than trastuzumab binds the high-affinity 158V variant. [58] Margetuximab in vitro mediates ADCC more potently than trastuzumab, pertuzumab, or both. While direct comparisons of adaptive immunity responses to these treatments remain limited in the literature, by using data across independent studies margetuximab resulted in higher rates of

circulating HER2-specific antibodies produced by B cells, and in higher T cell-mediated responses. [58]

In 2020, a new cornerstone treatment was approved by the FDA, especially for the treatment of CNS HER2+ metastases, tucatinib. This small molecule inhibitor is a tyrosine kinase inhibitor (TKI), blocking the ATP-binding site of the receptor phosphorylating tails. But the peculiarity of this novel TKI is the high selectivity for HER2 while sparing the epidermal growth factor receptor (EGFR), thus limiting off-target toxicities such as rash and diarrhea, typically observed with other TKIs. [60] Furthermore, tucatinib is very effective at crossing the blood-brain barrier and, together with trastuzumab and capecitabine, improved the median CNS progression-free survival of patients with brain metastases to 9.9 months compared to the 4.2 months without it. The intracranial objective response rate was also 47.3% versus the control 20.0%. [61]

Other drugs are being developed and tested. Trastuzumab duocarmazine, for example, is an ADC whose payload is a DNA alkylating agent with high membrane permeability and strong bystander effect, and can be relevant for HER2-heterogeneous or HER2-low tumors, though approval is still pending. [62-64] A novel frontier is represented by bispecific antibodies, capable of targeting multiple HER2 domains. One such example is zanidatamab, which binds both the domain IV and the heterodimerization domain II of HER2, the same domains targeted by trastuzumab and pertuzumab, respectively, favoring receptor clustering and internalization. [65]

Finally, a novel concept was the integration of the bispecific architecture of zanidatamab with a microtubule-disrupting payload to create zanidatamab zovodotin, a first-in-class bispecific ADC. [66] Although early-phase trials showed encouraging activity, its further development was paused due to strategic market considerations. [67]

Resistance Mechanisms in HER2+ Breast Cancer

The CLEOPATRA trial showed a median duration of response of 20.2 months, meaning that half of the metastatic patients who initially responded to the trastuzumab/pertuzumab/docetaxel combination were still progressing in less than two years. [48] This implies that a big percentage of these tumors were able to develop resistance mechanisms against this synergistic and breakthrough therapeutic strategy.

In fact, breast cancer cells are able to enact, in their proliferative niche, a number of changes that impair the efficacy of our therapies. This astonishing adaptive behaviour can manifest both by rewiring their internal cellular signalling and by heavily influencing their surrounding microenvironment.

The remodeled cellular signalling aims to sustain proliferation bypassing the pathways inhibited by the therapeutic agents. Most of these mechanisms rely on a central pathway activated by both homo- and hetero-dimerization of HER2 and whose players are able to maintain proliferative signals even in presence of HER2 blockade: the PI3K/AKT (Phosphoinositide 3-Kinase/Albanian K strain rodents Transforming) pathway.

HER2+ breast cancer cells discovered various ways to exploit the PI3K/AKT boosting effect on cellular proliferation and apoptosis resistance. [30] The “simplest” strategy involves the upregulation of the other most relevant receptor of the ErbB family in this context: HER3. As previously mentioned, HER3, despite its impaired kinase function, possesses multiple binding

sites for p85 subunit of Phosphoinositide 3-Kinase (PI3K), which is a strong activator of the PI3K/AKT pathway. HER3 borrows HER2's strong kinase activity through heterodimerization in order to phosphorylate p85 and bind PI3K. Fortunately, new therapies are able to, at least partially, address this kind of adaptation via the addition of pertuzumab.

The HER2-independent recruitment of PI3K activity can be achieved also genetically, through gain or loss of functions. The activating mutations potentiate the PI3KCA gene, which encodes the catalytic subunit of PI3K, leading to persistent PI3K enzymatic activity and downstream unchecked AKT phosphorylation. PI3KCA mutations are the most common alteration in the PI3K pathway in HER2+ breast cancer, with a rate ranging from around 20% in HR–/ERBB2+ subtype to around 28% in HR+/ERBB2+ subtype. [68] Loss of functions, instead, hit PTEN, a tumor suppressor gene and negative regulator of the PI3K pathway that acts by dephosphorylation of PIP3 to PIP2, as shown in Fig. 3. PTEN loss in HER2+ breast cancer can happen through genetic deletions, epigenetic silencing, or post-transcriptional regulation.

Sometimes, the cellular remodelling is so profound that it ends up modifying the structure of the HER2 protein itself. Approximately 30% of these tumors start expressing p95-HER, a truncated form of the receptor lacking the extracellular domain required for trastuzumab binding. Although trastuzumab treatment might prevent further truncation, this isoform not only cannot be targeted, it also retains the intracellular kinase domain necessary for activating downstream PI3K/AKT signaling. This mechanism may still respond to small molecule tyrosine kinase inhibitors (TKIs) that target the intracellular domain. [30] However, the receptor hiding strategy is particularly effective in facilitating cancer cells escape, to the point that HER2+ breast cancer cells exploit it through the overexpression of MUC4, too. MUC4 is a transmembrane glycoprotein that, when overexpressed, has a dual resistance role: structurally, it masks the HER2 epitopes targeted by trastuzumab, preventing drug binding; [69] functionally, MUC4 contributes to epithelial–mesenchymal transition (EMT), promoting metastatic potential. [70]

Deeper into the cellular control center, the dysregulation of HER2+ breast cancer cells can affect even the cell cycle. In particular, the Cyclin D1-CDK4/6-Rb axis is a known driver of cell proliferation. Overexpression of Cyclin D1 or activation of CDK4/6 allows cells to bypass growth inhibition induced by HER2 blockade. [71] Preclinical and clinical data suggest that combining CDK4/6 inhibitors with HER2-targeted therapies may help restore cell cycle control and overcome resistance. [72, 73]

Beyond the direct alterations of the cellular state, HER2+ tumors exhibit several other mechanisms of therapeutic escape by means of modulating the tumor microenvironment. [84]

In this context, the role of the tumor immune microenvironment has drawn increasing attention, particularly the presence of tumor-infiltrating lymphocytes (TILs). The antibody-dependent cellular cytotoxicity (ADCC) mechanism that trastuzumab relies on depends on the immune effector cells in the tumor milieu. Tumors with poor immune infiltration or with features of immune evasion, such as PD-L1 overexpression, may resist trastuzumab through inadequate immune-mediated tumor clearance. This has stimulated interest in combining HER2-directed therapies with immunomodulatory agents to overcome this form of resistance.

HER2+ breast cancer cells have also developed other ingenious strategies to influence the response of the surrounding microenvironment to their proliferation. One of those is the

shedding of exosomal HER2 and other extracellular vesicles (EVs) from their membrane. By carrying functional HER2 receptors, the exosomes act as decoys, binding trastuzumab and sequestering it away from tumor cells. These EVs can even transmit resistance traits to nearby cells, for example, by transferring immune modulatory molecules as TGF-β1 (Transforming Growth Factor β 1) and PD-L1 they induce the characteristics of their source cells in drug-sensitive cells and mold a more therapy-resistant tumor microenvironment. [75]

The host, and their genetic factors, also play a role in the treatment response. Polymorphisms in Fc gamma receptors (FcγRs), particularly FcγRIIIa, influence the binding affinity between trastuzumab and immune cells. Low-affinity polymorphic variants reduce the efficiency of ADCC, diminishing therapeutic efficacy even when HER2 expression is high. Therefore, host genetics may partly determine inter-patient variability in response and support the development of next-generation antibodies engineered for enhanced immune engagement. [76]

The possible resistance mechanisms developed by these tumors are numerous and might also include the activation of alternative signaling pathways, such as that of the insulin-like growth factor I receptor (IGF-IR) or the deregulation of metabolism. [77, 78] Some studies have reported that pertuzumab, when used alone, can lead to resistance driven by activating mutations in the extracellular domain, EGFR–HER3 heterodimer formation, or microRNA-mediated regulation. [79-81]

1.1.3 NK cell Immunosuppression as a Novel Mechanism for Resistance in HER2+ Breast Cancer

From Null Lymphocytes to Natural Killers

To defend against diverse pathogens and maintain cellular balance, the immune systems of higher vertebrates evolved both innate and adaptive arms. Around 500 million years ago, the adaptive response, driven by B and T lymphocytes, emerged to complement innate defenses. [82] For much of the 20th century, these two cell types were believed to account for all lymphocytes. That view changed in the 1970s, when it was observed that naïve lymphocytes could kill antibody-coated target cells without prior sensitization. [83] This led to the discovery of a distinct population, neither B nor T cells but called “null” lymphocytes, with innate cytotoxic potential. In 1975, Keissling and colleagues identified these effectors as "natural killer" (NK) cells, due to their spontaneous ability to eliminate tumor cells. [84]

Natural killer (NK) cells are indeed innate lymphocytes able to eliminate virus-infected and transformed cells without prior antigen exposure. They account for approximately 4% of peripheral blood mononuclear cells (PBMCs) and reach up to 10% of resident lymphocytes in the lung tissue. [85] NK cells are classically defined by the expression of CD56 and absence of CD3.

Two main subsets exist in human blood: CD56dim CD16+ NK cells, and CD56bright CD16− NK cells. They complete their development in secondary lymphoid organs, where they emerge as CD56bright CD16− cells, highly capable of producing cytokines but with limited cytotoxic activity. These cells then circulate in the blood and, under the influence of IL-15, can further differentiate into CD56dim CD16+ NK cells, the predominant NK subset in peripheral blood with enhanced cytotoxicity and antibody-dependent killing, but no longer proliferative. [86]

This difference is exemplified by the fact that CD56 facilitates cell-to-cell contact and developmental synapses, while CD16 is the FcγRIIIa, the receptor that recognizes antibodies on targets. [87, 88]

In parallel to this blood-based maturation pathway, CD56bright NK cells can also migrate into most of the peripheral tissues, where they adapt in response to local environmental cues. These cells express molecules like CD69, CD103, and CD49a, which help anchor them within the tissue. These resident NK cells tend to be less cytotoxic but are potent producers of cytokines like IFN-γ. Interestingly, while humans show widespread tissue-resident NK populations, mice appear to rely on related but distinct innate lymphoid cells 1 (LC1s) in many peripheral tissues. [86]

The Cytotoxic Kiss of Death

Natural killer (NK) cells are equipped to eliminate abnormal cells through two main cytotoxic mechanisms: the release of lytic granules and the induction of apoptosis via death receptors. At the heart of both strategies lies the formation of a specialized immunological synapse, a highly organized interface that brings the NK cell into close contact with its target. Upon chemotactic attraction to the site and activation, NK cells first latch onto their target through adhesion molecules, creating a stable contact. Inside the NK cell, the cytoskeleton then reorganizes to direct the microtubule-organizing center (MTOC) and secretory lysosomes toward the synapse. Once aligned, perforin and granzyme are released directly into the narrow space between the cells. Perforin perforates the target cell membrane, creating transient pores through which granzymes enter to trigger a cascade of apoptotic signals. [89] In parallel, NK cells can express death-inducing ligands such as FasL (Fas Ligand) and TRAIL (TNF-Related Apoptosis-Inducing Ligand), which engage their respective receptors on the target cell surface, initiating apoptosis through extrinsic pathways. [90]

NK cells must first determine which cells to kill, a decision governed by a dynamic equilibrium between activating and inhibitory receptors. NK cells can recognize stressed, infected, or transformed cells through the loss of MHC class I molecules, the “missing-self” phenomenon. MHCI is normally sensed via the co-inhibitory receptors KIR (Killer Immunoglobulin-like Receptor) and NKG2A (Natural Killer Group 2 member A): when inhibitory signals weaken due to MHCI loss, activating inputs dominate.

In other cases, cancer cells upregulate certain activating ligands under cellular stress, such as MICA/B (MHC class I Chain-related proteins A and B) and ULBPs (Unique Long 16-Binding Proteins), detected by receptors like NKG2D. Other tumor- or infection-associated molecules are recognized by natural cytotoxicity receptors, such as NKp30, NKp44, NKp46.

A potent activator of NK cells is FcγRIIIa (CD16), expressed predominantly on the CD56dim subset. This receptor recognizes the constant region (Fc) of IgG antibodies, predominantly IgG1 in humans, and is responsible for antibody-dependent cellular cytotoxicity (ADCC). [91] The ADCC allows NK cells to kill targets opsonized with antibodies even without prior priming, an ability particularly relevant in cancer targeted therapies, where therapeutic antibodies like trastuzumab guide NK cells to HER2-expressing tumor cells.

Receptor for Advanced Glycation End-Products is a Central Mediator of NK-Cell Cytotoxicity during Trastuzumab Therapy

In the micro-environment of HER2-positive breast tumours, effective antibody-dependent cellular cytotoxicity relies on the ability of natural killer (NK) cells to build a fully functional immunological synapse. The membrane receptor for advanced glycation end-products, alias RAGE, is a pattern recognition receptor (PRR) highly expressed on the mature CD56dim subset: during natural cytotoxicity or ADCC, RAGE promotes this process and accumulates at the synaptic interface. [92, 93]

Binding by endogenous alarmins HMGB1 (High Mobility Group Box 1) and S100A8/A9 (commonly calprotectin) induces the short cytoplasmic tail of RAGE to couple to the formin adaptor Dia-1 (diaphanous homolog 1) and initiate downstream signaling. Specifically, while HMGB1 boosts the MAPK/NF-κB pathway and favors the chemotaxis of NK cells, S100A8/A9 activates NK cell cytotoxicity through JNK (c-Jun N-terminal kinase) phosphorylation. [94-96]

JNK in turn phosphorylates the microtubule-regulating protein Stathmin, stabilising the tubulin network and licensing the microtubule-organising centre (MTOC) to polarize towards the target cell. Lytic granules then converge along this stabilised track and discharge perforin and granzymes directly into the tumour cell. This sequence underlies both natural cytotoxicity and the high-affinity FcγRIIIa-mediated attack elicited by trastuzumab. In patients who respond to treatment, this RAGE-JNK-Stathmin axis remains intact and is even potentiated by the pro-inflammatory milieu, allowing NK cells to serve as the executioners of therapeutic antibody action. [93] Such a working mechanism for ADCC is illustrated in Fig. 4a.

How Trastuzumab-Resistant HER2$^+$ Breast Cancer Switches Off NK Cells

Resistance emerges when the tumour or the surrounding cells like cancer associated fibroblasts (CAFs) flood the microenvironment with chitinase-3-like protein 1 (CHI3L1), a chitinase that lost its enzymatic activity and evolved as a promoter of type 2 inflammation. [93, 97]

This secreted factor binds RAGE with nanomolar affinity, out-competing S100A8/A9 that operates in the micromolar range, and triggers rapid internalisation of the receptor. Deprived of surface RAGE, NK cells fail to activate JNK, their basal and calprotectin-induced phosphorylation curves collapsing in the presence of physiological (20 nM) concentrations of CHI3L1. The downstream consequence is a sharp fall in Stathmin abundance, destabilisation of the microtubule lattice and a stalled microtubule-organising centre that never reaches the synaptic cleft. Granules remain stranded, perforin stores dwindle, and tumor cell lysis drops precipitously. This effect can be replicated by means of a RAGE-blocking antibody or the JNK inhibitor SP600125. The mechanism of CHI3L1-induced resistance is illustrated in Fig. 4b.

In vivo, breast cancer xenografts engineered to secrete CHI3L1 grow aggressively and become completely refractory to trastuzumab, while circulating and splenic NK cells display reduced ex-vivo ADCC. Conversely, neutralising CHI3L1 restores JNK signalling, rescues MTOC polarization and, when combined with trastuzumab, eradicates CHI3L1-high HCC1569 tumours in humanised mice. CHI3L1 can be inhibited by a neutralizing antibody or a soluble RAGE (sRAGE), a secreted version of the receptor that scavenges ligands: high sRAGE correlates with better outcomes, but its levels are decreased in advanced cancer. [93]

Thus, CHI3L1 acts as an immune checkpoint that silences NK cells.

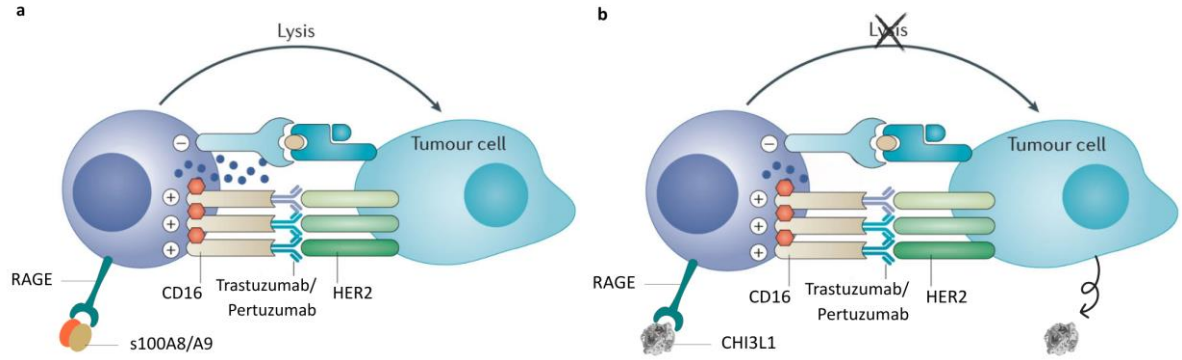


Figure 4. *Antibody-Dependent Cellular Cytotoxicity (ADCC).*
4a. NK cells recognize trastuzumab/pertuzumab antibodies opsonized on the target HER2 receptors overexpressed by tumor cells. In normal functioning conditions of the immune system, NK cells are able to lyse the tumor cells. The secretory granules polarization is facilitated by the binding of the inflammatory molecule s100A8/A9 on the receptor RAGE on the NK cell membrane.
4b. HER2+ breast cancer cells enact a mechanism of resistance to the trastuzumab/pertuzumab therapy by secreting, themselves or through the tumor microenvironment (e.g. CAFs), CHI3L1. This molecule competitively displaces s100A8/A9 from the RAGE receptor causing the paralysis of the cytotoxic machinery in NK cells. The effect of this phenomenon is the immunosuppression of NK cells with the loss of their ability to lyse targeted tumor cells, which can escape.
Image from Morvan. Lanier, 2015, has been adapted to create this figure. [98] Copyright 2016 Macmillan Publishers Limited.

1.1.4 Relevance of the CHI3L1-RAGE Axis in Other Pathologies

Inflammatory Bowel Disease and Colitis-Associated Carcinogenesis [99]
During intestinal inflammation the host epithelium is bathed in a fluctuating mixture of damage-associated proteins, among which CHI3L1 and S100A9 take centre stage for their opposing effects on cell survival and death.

In the acute phase of dextran-sulphate-sodium colitis, a widely used experimental model of inflammatory bowel disease (IBD), S100A9 is produced in abundance, binds RAGE on epithelial and immune cells, and amplifies its own transcription through an NF-κB positive feedback loop. NF-κB further fuels oxidative stress and antibacterial activity but also drives apoptotic loss of epithelial cells.

As inflammation persists, CHI3L1 expression in epithelial cells steadily rises and, thanks to its nanomolar affinity for the same RAGE ectodomain, progressively out-competes S100A9. This ligand switch arrests the S100A9-RAGE feedback circuit and creates a CHI3L1^high/S100A9^low mucosal milieu. In this context, STAT3 (Signal Transducer and Activator of Transcription 3), β-catenin and NF-κB are phosphorylated, microbicidal pressure vanishes, and intestinal epithelial cells shift from apoptosis to active proliferation.

The physiological dividend is faster wound restitution; the pathological price is the survival and expansion of mutated clones that seed colitis-associated cancer.

Consistent with this model, CHI3L1-deficient mice exposed to dextran-sulphate-sodium suffer deeper ulceration during acute colitis but show a markedly lower tumour burden during the chronic phase, while wild-type mice exhibit intense nuclear STAT3, β-catenin and p65 (a protein that belongs to the NF-κB transcription factor) staining within hyper-proliferative crypts.

Therefore, pharmacologically disrupting CHI3L1-RAGE signaling, via CHI3L1 blockade or CHI3L1-selective sRAGE decoys, may arrest the transition from reparative hyperplasia to carcinoma without compromising early antimicrobial defense mediated by S100A9.

Impairment of Hippocampal Repair in Multiple-Sclerosis-Associated Demyelination [100]

In the demyelinated hippocampus of multiple-sclerosis patients and correspondent mouse models, activated astrocytes secrete large amounts of CHI3L1, and its concentration rises and falls in parallel with demyelination and remyelination. This astrocytic secretion burst correlates strongly with memory loss, because CHI3L1 disables the very cellular programs that normally rebuild hippocampal circuits. Early in adult neurogenesis, the protein binds CRTH2 (Chemoattractant Receptor-homologous molecule expressed on Th2 cells) on neural stem cells and blocks their proliferation; but the critical disruption of cognitive function comes later, when CHI3L1 engages RAGE abundantly present on immature granule neurons.

Ligation of RAGE activates GSK3β (Glycogen Synthase Kinase 3 beta), which in turn accelerates β-catenin degradation, and thereby extinguishes transcriptional cues needed for dendritic elaboration, spine formation and synaptic integration of newborn neurons. The result is a hippocampus populated by structurally stunted cells unable to join functional networks. Such a pathology manifestly depresses performance in object-recognition and spatial-learning tasks.

Causality is underscored by genetics and pharmacology: astrocyte-specific CHI3L1 deletion, RAGE knock-down or antagonism, and downstream blockade of GSK3β or restoration of β-catenin each restore neurogenesis and rescue cognition. Together these findings position the CHI3L1-RAGE-β-catenin axis as a critical pathway through which neuroinflammation converts demyelination into lasting cognitive disability.

The main pathways downstream of RAGE upon binding with ligands are schematized in Fig. 5. Interestingly, these studies demonstrated a negative regulation of all these pathways by CHI3L1 binding of RAGE.

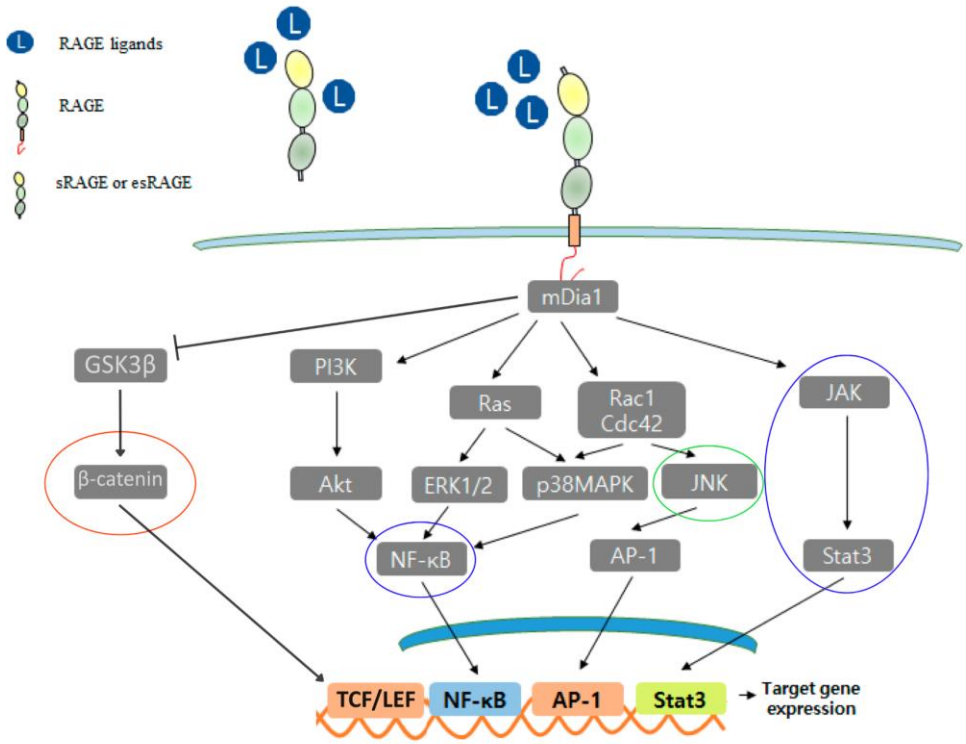


Figure 5. *Map of RAGE downstream signalling pathways upon ligand interaction.*
The signalling pathways activated by CHI3L1 in the three pathologies where the interaction with RAGE has been documented are displayed, as well as their target genes, and color coded as follows. In green, the JNK pathway, which is downregulated in HER2+ breast cancer. In blue, the JAK-Stat3 and the NF-κB pathways, which, along with β-catenin, are upregulated in colitis-associated cancer. Finally, in orange, only the β-catenin pathway, which is downregulated by degradation through GSK3β activation in multiple sclerosis-associated demyelination, while the usual effect on GSK3β by other ligands binding to RAGE is the inactivating tyrosine 216-dephosphorylation that in turn causes the stabilization of β-catenin.
Image from Kim, et al., 2021, has been adapted to create this figure. [101]

1.1.5 Significance of the CHI3L1-RAGE Axis and its Dysregulation in Pathology: A Node of Inflammation and Immune Surveillance

Given the involvement of the CHI3L1-RAGE axis in modulating inflammation and cell proliferation across disparate inflammatory landscapes, targeting it promises a unifying approach to limit malignant and degenerative evolutions of these pathologies.

In particular, in HER2+ breast cancer, disrupting the CHI3L1-RAGE interaction, while preserving the immunostimulatory S100A8/A9 signalling, could reverse the resistance mechanism developed by some HER2+ breast cancer to targeted therapy.

This thesis will focus on elucidating the molecular interaction between CHI3L1 and RAGE, as a physiologically and pathologically significant intersection between two evolutionarily distinct programs, type II inflammation and innate immune recognition, that converge through this ligand-receptor pairing.

1.2 Conceptual Problem: Making Complexity Reachable

The central challenge of this thesis is to find a way into a system that, by its nature, resists simple solutions. $HER2^+$ breast cancer is not just a disease of one gene or one cell type, it is a dynamic, adaptive ecosystem where tumour cells, immune cells, and signals from the surrounding environment are constantly shifting to protect the tumour's survival. Resistance to therapy is not a single event, but a process built into the nature of this complexity. The intuition behind this work is that, even in such a system, there may be key points, small, specific interactions, that carry much larger consequences. The CHI3L1-RAGE interaction appears to be one of these points. It is a molecular node that allows the tumour to silence the very immune cells that targeted therapy tries to activate. The deeper aim of this thesis is to explore whether such a small, precise interaction can be turned into a practical way to intervene, to ask whether identifying the right part of a complex system can allow us not to control everything, but to shift the balance in our favour. It is a way of turning the overwhelming into the workable.

1.3 Biological Problem: Converting a Ligand-Receptor Checkpoint into a Therapeutic Handle

1.3.1 Understanding Structure as a Path to Intervention

To act meaningfully within a complex system, we must first understand how its parts came to matter. The biological problem at the heart of this thesis is not just how CHI3L1 binds RAGE, but why this interaction exists at all, what evolutionary logic shaped it, and how that logic now enables tumour cells to manipulate immunity.

CHI3L1 and RAGE are not randomly paired: their interaction reflects millions of years of selective pressure aligning form and function. The way one molecule fits into the other, the specificity of their contact, the consequences of their binding, these are not detached features, but direct expressions of the biological roles they evolved to play. Structure *is* function, and function *is* history, written in molecular form.

In this work, we approach that question not from a single discipline, but through a combination of molecular biology, structural biophysics, immunology, and computational modelling. In particular, tools like AlphaFold, which embed within their architecture the evolutionary relationships between sequences, allow us to explore this interface as more than chemistry, as a coevolved solution to shared biological pressures. In doing so, we aim to understand not just how CHI3L1 and RAGE interact, but how intervening in that interaction can offer a practical way to re-open immune attack against $HER2^+$ tumours.

1.3.2. RAGE: A Master Sensor and Transducer of Danger and Inflammation

The AGER Gene: Genomic Setting and Transcriptional Control

The human AGER locus lies in the major-histocompatibility complex III on chromosome 6p21.3, embedded among genes that orchestrate innate and adaptive immunity. [102] This strategic placement foreshadows its function: AGER encodes the receptor for advanced glycation end-products (RAGE), a receptor for nonenzymatic glycosylated molecules and a pattern recognition receptor that translates biochemical stress into inflammation. [103, 92] The AGER gene encoding RAGE emerged only in mammals. Its closest relatives are immunoglobulin superfamily adhesion molecules (CAMs), such as ALCAM (Activated Leukocyte CAM), BCAM (Basal CAM), and MCAM (Melanoma CAM), which predated RAGE in metazoan evolution. [104] Early speculations during the discovery of RAGE hypothesized that it could represent a primitive T-cell immune receptor, suggesting that such molecules may have evolved from simple, non-rearranging cell-surface proteins, like adhesion molecules, in early vertebrates. [105]

Eleven exons separated by ten introns span the AGER locus; [106] when the entire cassette is transcribed and translated it yields a ~55 kDa single-pass membrane protein. [107] Under physiological conditions AGER transcripts are scarce in most organs, yet the lung is an exception. Alveolar type I epithelial cells, along with the vessel wall of alveolar capillaries, manufacture RAGE at high constitutive levels suggesting a role in maintaining the delicate air-blood interface. [108, 109]

Inflammatory stress overrides this quiet baseline. The shared PBX2/AGER promoter carries multiple κB sites; engagement of cell-surface RAGE by any of its diverse ligands activates NF-κB, which in turn binds the promoter and amplifies AGER transcription, establishing a positive feedback loop. [110, 111] Hyperglycaemia, hypoxia, oxidised lipids and sterile tissue injury converge on this pathway, explaining why RAGE becomes abundant in chronic metabolic and inflammatory disorders, especially those associated with vascular conditions, such as diabetes. [112]

AGER regulation is made still more intricate by alternative splicing. At least twenty human transcripts have been catalogued, many of which truncate or re-shuffle exons encoding the transmembrane helix or cytosolic tail. [113] Splicing therefore determines whether the translated product is membrane-anchored and signalling-competent, membrane-bound but signalling-silent, or secreted altogether. The resulting isoforms expand the functional spectrum of a single genetic locus, allowing the same gene to encode an alarm bell, a decoy, and a dominant-negative brake, often simultaneously and in tissue-specific proportions. [114, 115]

Genetic variants in AGER may influence its expression or function. At least 30 polymorphisms have been documented, including promoter variants that alter transcriptional activity and coding variants like Gly82Ser in the V-domain. [116, 117] The G82S polymorphism lies adjacent to an N-glycosylation motif and has been reported to affect glycosylation and ligand affinity. [117] Such variants have been associated with differential risk of cancer, diabetic complications and inflammatory diseases in some studies. [118, 102] Overall, the genetic and transcriptional landscape of RAGE reveals a receptor tightly controlled at the DNA/RNA level.

RAGE Protein Structure and Isoform Diversity

1) Full-Length RAGE: Modular Architecture of a Danger Receptor

Canonical RAGE begins as a 404-residue polypeptide with a cleavable signal peptide that directs it to the secretory pathway. [119] It is a type I transmembrane glycoprotein belonging to the immunoglobulin superfamily, characterized by a modular design that enables it to function as a broad-spectrum sensor of tissue damage and inflammation. It seems that RAGE evolved from a family of cell-adhesion molecules, and might still act as an adhesion molecule, in particular in the lung where it is highly expressed or under pathological conditions characterized by an increase of its protein levels; [104] later, its domains specialized for danger-signal detection.

The mature ectodomain comprises three immunoglobulin-like folds arranged linearly: an N-terminal V-type (Variable-set Ig-like) domain followed by C1 (Constant-1-set Ig-like) and C2 (Constant-2-set Ig-like) domains. V (residues 23-116) and C1 (residues 124-221) pack tightly to form a single rigid unit (VC1) whose central groove is ringed by basic residues and capped by a hydrophobic "crest" in proximity of the flexible C′D loop, an interface exploited by several acidic ligands such as S100 family proteins. The V-domain serves as the principal ligand binder. [119-124] Its structure is a characteristic β-sandwich stabilized by an intramolecular disulfide bond between Cys 38 and Cys 99. [114] Two asparagine-linked glycans project from the V-domain at Asn 25 and Asn 81; glycosylations can enhance affinity for certain S100 ligands, especially when asialo-carboxylated. [125]

The C2 domain (residues 227-317) exhibits a strongly acidic surface that faces the basic face of VC1 in higher-order oligomers, and rarely binds ligands directly. [114, 124] It is tethered to VC1 by a pliant linker and contributes two cysteines (C259, C301) that form inter-molecular disulfides, cementing a constitutive RAGE homodimer that is required for exit from the endoplasmic reticulum. [126]

Beyond the ectodomain lies a single-span transmembrane helix (residues 343–363), which likely supports receptor dimerization within the membrane, and a short intrinsically disordered cytoplasmic tail (residues 364–404). [120] Despite lacking enzymatic activity, this tail is functionally indispensable. It contains several functional motifs: a membrane-proximal basic segment, a central acidic stretch, and a conserved C-terminal region that serves as a docking platform for signal transducers. [127]

Ligand binding initiates receptor clustering through non-covalent V-V and C1-C1 interfaces, producing oligomers whose geometry positions the cytoplasmic tails roughly 100 Å apart, just the spacing needed to recruit the diaphanous-related formin DIAPH1 and to dock TIRAP-MyD88 (Toll/Interleukin-1 Receptor domain-containing Adaptor Protein - Myeloid Differentiation primary response gene 88), once Ser 391 in the tail has been phosphorylated by PKCζ. [127] Oligomerization is therefore the mechanical prerequisite for NF-κB and MAP-kinase signalling.

2) RAGE Glycosylation: Modulating Receptor Function Through Sugar Signatures

In RAGE, glycosylation is a molecular switch that affects ligand selectivity and signalling, but also stabilizes RAGE structurally. Two N-linked sites bracket the V-domain: Asn 25 is constitutively capped with complex, often asialo-carboxylated chains, whereas Asn 81 is

variably occupied by high-mannose, hybrid or unmodified residues and becomes fully glycosylated when the common Gly 82 to Ser polymorphism is present. [129, 130]

This asymmetry creates a glycan micro-heterogeneity axis: the rigid, acidic glycan at N25 stabilises the ectodomain, while the malleable N81 patch modulates access to the hydrophobic ligand cavity. [130] Functional studies confirm that engineering or selecting particular glycoforms rewires biology. Carboxylated (non sialic-acid–terminated) N-glycans, enriched at N25, serve as high-affinity docking pads for S100A12, enabling receptor clustering and downstream MAPK/NF-κB activation; the same motif gates binding of S100A8/A9 and potentiates HMGB1 signals in myeloid and endothelial cells. [125, 131-133]

Conversely, hypoglycosylation at N81, or its forced occupation in the G82S allele, tightens affinity for chemically diverse ligands: de-N-glycosylated or G82S RAGE binds glycolaldehyde-AGEs (advanced glycation end-products) with ∼10^3-fold lower Kd, amplifies VEGF transcription in endothelial cells, and heightens NF-κB and cytokine output after S100B stimulation. [131, 130] The same N81 glycan re-configuration underlies the higher affinity (Kd ≈ 45 nM) of G82S RAGE for amyloid-β_{42} ($A\beta_{42}$), providing a mechanistic link between this polymorphism and Alzheimer's risk, with RAGE mediating inflammatory response to $A\beta_{42}$. [130] Glycosylation also dictates the behaviour of the therapeutic decoy sRAGE: only mammalian-derived, bi-antennary N-glycoforms preserve anti-inflammatory potency in arterial-injury models, whereas insect-cell forms devoid of complex glycans are largely inert. [134] Collectively, RAGE glycosylation is remarkably plastic, spanning fixed complex, inducible high-mannose and ligand-sensitising carboxylated species, and this plasticity fine-tunes interactions with AGEs, S100/calgranulins, HMGB1 and amyloid-β.

Tissue-specific glycosylation patterns further diversify this responsiveness: for instance, neuroblastoma cells produce highly carboxylated RAGE (>90%), while HeLa (immortalized cell line from from a cervical carcinoma)-derived soluble RAGE shows minimal carboxylation (1-2%). [125, 135]

3) Soluble RAGE: Endogenous Decoys in the Circulation

Several splice events delete the transmembrane exon, producing proteins that are secreted rather than membrane-anchored. The principal form, RAGE_v1 (often called sRAGE, esRAGE or isoform 3), terminates in a unique 16-residue tail created by a frameshift. [113] Plasma also contains cRAGE, the VC1C2 ectodomain shed from full-length receptors by ADAM10 (a disintegrin and metalloproteinase domain-containing protein 10) or MMP-9 (matrix metalloproteinase-9). Both species retain the complete ligand-binding surface yet lack the signalling machinery, allowing them to mop up AGEs, S100s, HMGB1, amyloid-β, and other agonists. Their concentrations rise in many inflammatory states, sometimes as a gauge of disease burden, sometimes as a compensatory brake on runaway RAGE signalling. [136]

4) Dominant-Negative RAGE: Membrane-Bound but Mute

Another splice variant, RAGE_v20 (isoform 10), preserves the ectodomain and the transmembrane helix but truncates before the cytosolic tail. Anchored in the plasma membrane yet signalling-deficient, it competes with full-length receptors for ligand and for oligomeric partners, thereby dampening downstream pathways within the same cell. [114]

The balance between full-length, soluble, and dominant-negative molecules is therefore a key determinant of how vigorously a tissue responds to danger cues.

Species Variants and Experimental Implications

Murine RAGE mirrors the human receptor in overall layout, but mice express an additional splice product, mRAGE_v4 (isoform 3), that excises nine residues encompassing the ADAM10/MMP cleavage motif. As a result, this variant hardly sheds, remains confined to the plasma membrane, and is as abundant as full-length RAGE in healthy mouse lung. Humans lack an equivalent transcript, and human FL-RAGE is intrinsically less prone to shedding, illustrating why mouse data on soluble RAGE generation must be interpreted cautiously. Other conserved isoforms include mRAGE_v1 (secreted) and mRAGE_v20 (dominant-negative), paralleling their human counterparts. Sequence differences are modest, chiefly scattered surface substitutions, yet they can influence glycosylation and disulfide dynamics, reinforcing the need for structural validation whenever murine models are extrapolated to human disease. [137]

RAGE oligomerization is mediated by heparan sulfate

RAGE requires oligomerization to initiate intracellular signaling, and this process is critically mediated by heparan sulfate (HS). [137, 138] HS dodecasaccharides drive the formation of a stable RAGE hexamer, organized as a trimer of dimers with a 2:1 RAGE:HS stoichiometry. These hexamers were initially seen at the SEC, where in the presence of dodecasaccharides of HS, the sRAGE elution peak shifted from an estimated 41 kDa (more than its real mass probably due to the elongated shape) to around 200 kDa, as shown in Fig. 6.

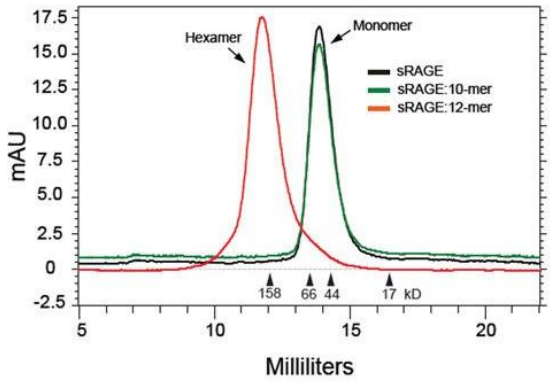


Figure 6. *Size exclusion chromatography elution peaks of sRAGE in presence or absence of heparan sulfate dodecasaccharides.*

These findings by Xu et al., 2014, were obtained using a Superdex Increase 10/300 GL column. Image reproduced from Xu et al., 2014. [139]

Crystallographic and small-angle X-ray scattering (SAXS) studies revealed the RAGE-HS oligomeric structure, with key interactions involving basic residues K39, K43, K44, R104, K107 (V domain) and R216, R218 (C1 domain), and a conserved hydrophobic dimer interface (V35, V78, F85, L86). The crystal structure of the RAGE V-C1 domain in the presence of (non-resolved) HS (PDB ID: 4IM8) confirmed the oligomer's architecture, while comparison

with monomeric structures (PDB IDs: 3CJJ, 3O3U) highlighted the conformational reorganization upon HS binding. Importantly, ligand-binding residues on the V-domain remain accessible in the oligomer, indicating that HS functions as a scaffold for receptor assembly rather than directly triggering signaling.

In the absence of oligomerization, such as when RAGE is expressed without HS or when key HS-binding residues are mutated, the receptor remains predominantly monomeric in solution and at the cell surface. These monomeric forms retain their ability to bind ligands like HMGB1 or S100B with comparable affinity, but are unable to propagate intracellular signals, as evidenced by loss of ERK1/2 phosphorylation and reduced NF-κB activation. This functional dependency on oligomerization has been shown in vivo, using knock-in mice bearing point mutations (R216A-R217H-R218A) that specifically disrupt HS binding without affecting ligand recognition. These mice failed to form RAGE oligomers and exhibited impaired osteoclast differentiation and reduced neutrophil-mediated liver injury, closely mirroring the phenotype of RAGE knockout mice.

The potentially relevant surface domains for RAGE oligomerization and the surface solvent electrostatics at physiological pH are shown in Fig. 7. Notably, one side of the molecule is extremely negatively charged while the other is relatively hydrophobic, potentially favoring the association with other monomers through hydrophobic interaction while exposing the charged surface to the ligands.

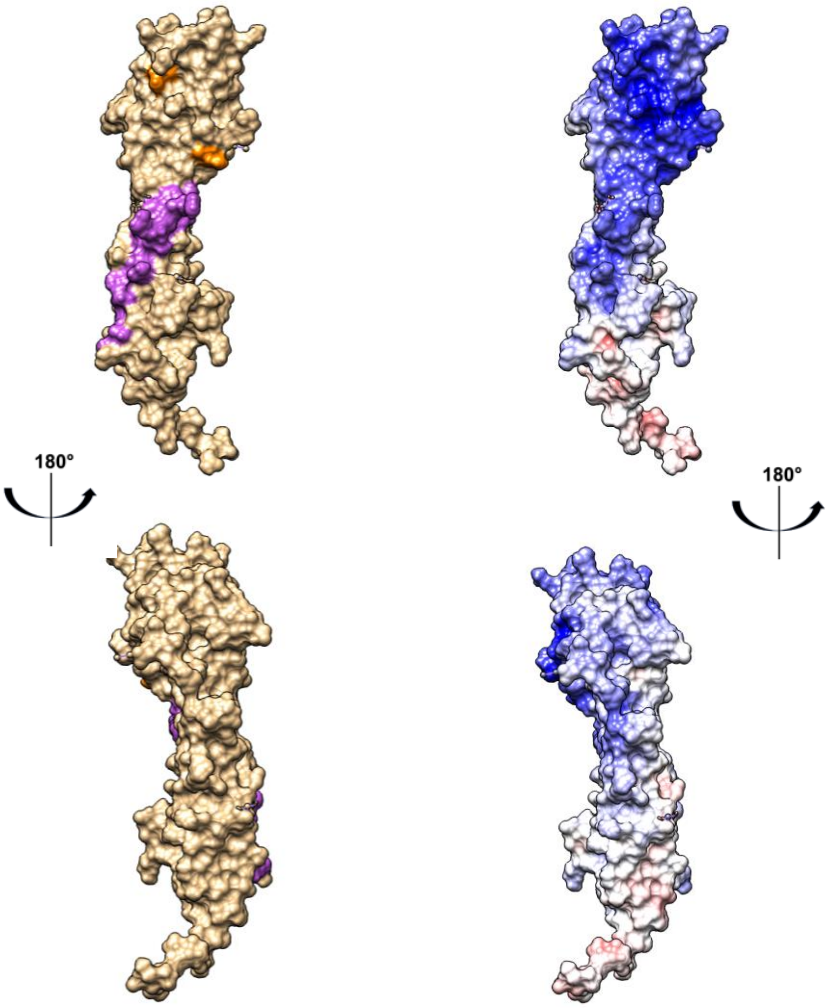


Fig. 7a (top). *Front view of the surface representation of the domains (left) and the solvent electrostatics at physiological pH of RAGE VC1 domains.*
The FG and GG' loops are represented in purple, while the site of N-glycosylation (Asn25 and Asn81) are shown in orange. From the electrostatic viewpoint this surface of VC1 is highly negative, in fact it is mostly represented in blue.
Fig. 7b (bottom). *Rear view of the surface representation of the domains (left) and the solvent electrostatics at physiological pH of RAGE VC1 domains.*
We can observe that this surface displays certain white areas on both domains, being relatively hydrophobic with respect to the other side of the receptor.

The Expansive Ligand Repertoire of RAGE

By virtue of its positively charged V-C1 tandem and an intrinsic tendency to oligomerize, [140, 141] RAGE operates as an innate immune sensor that samples the extracellular milieu for molecular signs of stress. Most of its ligands are either damage-associated molecular patterns (DAMPs) released from injured cells or neo-epitopes generated by chronic inflammation. [140, 142] Affinity for any single epitope is modest; high-avidity engagement arises only when ligands present those epitopes repeatedly (e.g. on an aggregate or multimer) and thereby nucleate higher-order RAGE clusters. [122] This pattern-recognition logic endows the receptor with extraordinary breadth while preserving selectivity for danger contexts.

The sections below trace how different ligand families exploit RAGE motifs to elicit distinct biological programmes.

1) Advanced Glycation End-Products: Metabolic By-products that Fuel Chronic Inflammation

Advanced glycation end-products (AGEs) arise when reducing sugars or oxidised lipids react non-enzymatically with lysine and arginine side-chains on long-lived proteins, an event accelerated in hyperglycaemia, renal failure and normal ageing. [143] Although the affinity of a single AGE adduct for RAGE is weak, serum albumin, collagen or low-density lipoprotein covered with many CML (Nε-carboxymethyl-lysine) and CEL (Nε-carboxyethyl-lysine) groups, AGEs that form on proteins during oxidative or glycation stress, bind avidly and cluster the receptor. [144, 92] Engagement of this multivalent surface switches on RAC (Ras-related C3 botulinum toxin substrate)-driven NADPH oxidase, sustains PKCζ-dependent NF-κB activation and drives expression of adhesion molecules, TGF-β and further RAGE itself. This is an inexorable feed-forward loop that underlies vascular and renal complications of diabetes. [145]

2) S100 Calcium-Binding Proteins: Versatile Alarmins with Oligomeric Tricks

Members of the S100 family share the EF-hand fold (helix E - Ca^{2+} binding loop - helix F motif) but differ markedly in quaternary structure and surface charge when Ca^{2+} and Zn^{2+} occupy their loops. [146] Those cations pry open a hydrophobic cleft on each monomer and trigger dimerization or higher oligomerization, an essential prerequisite for RAGE recognition. The receptor, for its part, offers a complementary hydrophobic patch flanked by basic residues (notably Lys 44/Arg 48 and the Trp 61-Val 63-Leu 64 triad on the C'D loop). [128, 147, 123]

S100A8/A9 (calprotectin) illustrates how the oligomer state dictates function. In the cytosol of neutrophils the heterodimer acts as a Ca^{2+} sensor, but once released and exposed to extracellular Ca^{2+} it assembles into a $(A8/A9)_2$ tetramer whose two A9 subunits dominate RAGE binding with mid-nanomolar affinity. [148] That interaction is amplified thirty-fold when the receptor bears asialo-carboxylated high-mannose glycans. [135, 149] This potently activates FAK (focal adhesion kinase) and MAP-kinases in breast-cancer cells, driving epithelial-mesenchymal transition and metastatic spread. [150] At higher, micromolar concentrations, as in septic shock, the same ligand-receptor pair flips to a cytotoxic mode, illustrating a concentration-encoded duality reminiscent of chemokines. [151]

S100B follows a similar principle but with different stoichiometry. The Ca^{2+}-bound dimer interacts with one RAGE V-domain on each S100B monomer ($K_d \approx 0.5$ μM). [121, 122] When S100B tetramerises, that structure might engage four RAGE molecules at once, stabilising an octameric signalling hub. [152] The importance of V-domain integrity is underscored by the observation that deletion of the C'D loop (isoform 2) abolishes S100B binding. [123]

Other S100s largely respect this theme. C3S-mutant S100A6 sandwiches two V-domains between the faces of its dimer; S100P uses a flexible linker centred on Phe^{44} to engage the same basic crest on RAGE V domain identified in the binding sites with S100B and S100A6, and its F44/Y88/F89 triad is indispensable for complex formation. [153, 154]

S100A12, which seems to be an homolog of murine S100A8, also displays a glycan-dependent high-affinity binding with RAGE. [125]

3) HMGB1: A Nuclear Architect Repurposed as a Danger-Associated Molecular Pattern

HMGB1 (high mobility group box 1, also known as amphoterin) shuttles between chromatin and the extracellular space. Amphoterin literally means "protein with dual properties", which reflects its dual function: inside the nucleus, it regulates chromatin and transcription; outside the cell, it acts as a pro-inflammatory cytokine or DAMP. [155]

Necrotic release or active secretion of HMGB1 exposes a highly acidic C-terminal tail (residues 150-183) that docks onto RAGE in a glycan-assisted manner. [156] On alignments, this region resembles the N-terminal part of S100 proteins, in particular S100A9. Indeed, similarly to other S100 proteins, also for HMGB1 the deglycosylation of the receptor or the competition with soluble carboxylated glycans markedly blunts binding. [157]

HMGB1 drives the Ras–MAPK–NF-κB signalling and cytokine production, but its immunological impact can vary. When in complex with complement component C1q, it reverts the inflammatory response: RAGE and the inhibitory receptor LAIR-1 (leukocyte-associated immunoglobulin-like receptor 1) form a ternary assembly with HMGB1 that re-programmes macrophages toward an M2 resolutive phenotype. [158] Instead, when HMGB1 is bound to extracellular DNA, as in sterile tissue damage, RAGE cooperates with TLR9 (toll-like receptor 9) and the AIM2 (absent in melanoma 2) inflammasome to produce interferon. [159]

4) Amyloid-β Peptides: Metabolic Stress in Neurodegeneration

In Alzheimer's disease, the 42-amino-acid amyloid-β peptide ($A\beta_{42}$), produced through sequential cleavage of APP (amyloid precursor protein) by β- and γ-secretases, tends to self-aggregate into oligomers and fibrils. These aggregates display repetitive, negatively charged surfaces that are readily recognised by RAGE on brain microvascular endothelial cells. Once bound, RAGE facilitates the transport of Aβ from the bloodstream into the brain parenchyma, thereby increasing amyloid accumulation in the central nervous system. [160]

Within the brain, Aβ engagement of RAGE on neurons and glial cells activates oxidative stress pathways and promotes tau phosphorylation, both hallmark processes in the progression of Alzheimer's disease. Circulating soluble sRAGE can sequester Aβ, preventing its polymerization and enhancing its clearance via the kidneys. However, conditions such as renal insufficiency, or the presence of the SARS-CoV-2 spike protein, which itself binds Aβ and disrupts its elimination, can indirectly amplify Aβ-RAGE interactions, worsening amyloid burden and inflammation. [136]

5) Lipids, Complement and Microbial Patterns

RAGE also binds non-protein ligands.

Lysophosphatidic acid (LPA) binds the ectodomain with micromolar affinity but, by leveraging receptor clustering, drives PKB-dependent proliferation and angiogenesis in lung and mammary tumours. [161]

The complement subcomponent C1q binds RAGE together with HMGB1 on macrophages promoting polarization. [158]

Finally, bacterial components such as lipopolysaccharide (LPS) and microbial nucleic acids do not typically bind RAGE directly, but instead form complexes with DAMPs like HMGB1.

These complexes are readily recognised by RAGE, enabling it to cooperate with Toll-like receptors and thereby sustain and amplify innate immune responses. [162]

The Molecular Pathways of RAGE Signalling and Its Cellular Repercussions

The extracellular life of RAGE begins inside the endoplasmic reticulum. Two cysteines in the C2 domain form an obligatory intermolecular disulphide bond; without this covalent handshake the receptor fails quality-control and is shunted to proteasomal destruction. [126] Most canonical ligands, such as AGEs, calcium-loaded S100 dimers and tetramers, HMGB1, are multivalent and therefore nucleate higher-order RAGE assemblies by bridging adjacent dimers through the juxtaposed V- and C1 domains. [163]

Upon ligand binding, the cytoplasmic tails of RAGE come together and form a scaffold that allows docking of the formin DIAPH1, specifically binding to an arginine-rich patch centered on residues R366/R367 in the RAGE cytoplasmic domain. [164]

DIAPH1 activates small Rho family GTPases Rac1 (Ras-related C3 botulinum toxin substrate 1) and Cdc42 (cell division control protein 42), which regulate actin cytoskeleton reorganization, lamellipodia formation, cell migration, and cell polarity. In parallel, through a burst of NADPH-oxidase–derived reactive oxygen species (ROS). These ROS serve not only as intracellular messengers but also as part of a self-amplifying damage signal that feeds back to further activate RAGE. [165]

Simultaneously, atypical protein kinase C ζ phosphorylates RAGE at Ser391, creating a docking site for the Toll-like receptor (TLR) adaptor proteins TIRAP and MyD88. This recruitment fixes the canonical TLR-IRAK-TRAF6 signaling module onto RAGE, leading to activation of the IKK (IκB kinase) complex and subsequent nuclear translocation of NF-κB. [166]

Concurrently, MAPK pathways are initiated: ERK1/2 promotes proliferation and survival, p38 modulates cytokine production, and JNK controls apoptotic or cytotoxic responses. [167]

Following activation, RAGE undergoes regulated proteolysis: RAGE downstream signalling events promote the upregulation of MMP9 and ADAM10, which cleave membrane-bound RAGE, causing the release of its ectodomain as the decoy cRAGE. [168] Then, γ-secretase cleaves the residual membrane-bound C-terminal fragments, generating an intracellular fragment that may then localize to the cytoplasm or nucleus, and promote signaling. [169, 170]

This sequential shedding links receptor activation to signal attenuation while producing soluble receptors capable of dampening inflammation at distant sites. [171]

At the transcriptional level, the RAGE pathway reinforces itself. NF-κB upregulates AGER and RAGE ligands too, including S100A8, S100A9, and HMGB1. [172, 173]

RAGE in Physiology and Pathology

RAGE has a homeostatic role in airway architecture: in fact, genetic ablation in mice leads to hyper-proliferative alveolar epithelium and a heightened susceptibility to fibrosis. [174] In the developing nervous system, low-nanomolar fluxes of HMGB1 or S100B engage neuronal RAGE to induce neurite extension and path-finding. [175] HMGB1-RAGE signalling orchestrates leucocyte chemotaxis and angiogenesis, accelerating tissue repair without overt inflammation. [176] These beneficial functions rely on tightly bounded, transient ligand

exposure. When ligand production becomes chronic, or when clearance fails, the same circuitry turns pathological.

Hyperglycaemia accelerates non-enzymatic glycation, flooding the vasculature with AGEs that lock endothelial and smooth-muscle RAGE into a state of persistent NF-κB activation. [177] The result is diabetic atherogenesis: oxidative stress, adhesion-molecule up-regulation and matrix metalloproteinase release that destabilise plaques and stiffen vessels. Similar mechanisms operate in the diabetic kidney, where podocyte RAGE drives TGF-β and collagen deposition, and in the failing myocardium, where S100A8/A9-RAGE interactions exacerbate post-infarction remodelling. [178, 179] In the central nervous system, RAGE accelerates synaptic loss in Alzheimer's disease. [180] In rheumatoid joints, S100A8/A9 and HMGB1 from infiltrating neutrophils activate synovial RAGE, fuelling cartilage destruction. [181] In colitic gut mucosa RAGE senses luminal β-lactoglobulin-induced AGEs and microbial byproducts, extending epithelial NF-κB activation long after the initial insult. [182] Airway RAGE, conversely, has a controversial role: despite its possible homeostatic role in the tissue, it is able to promote fibrotic remodelling in idiopathic pulmonary fibrosis and fuel chronic inflammation in asthma when exposed to S100A8/A9 or HMGB1. [183-185]

Many aggressive tumours, triple-negative and HER2-positive breast cancers, melanoma, pancreatic adenocarcinoma, up-regulate RAGE along with their own secretion of S100 proteins and HMGB1. [186-188, 147] This drives ERK- and PI3K-dependent proliferation, FAK-mediated migration and epithelial-to-mesenchymal transition, in an immunosuppressive microenvironment. Instead, in primary lung adenocarcinoma RAGE is often silenced, suggesting that in this case it may act as a tumour suppressor. [189]

1.3.3 CHI3L1 (YKL-40): An Immune Regulator with a Long History and Dark Side

Chitinase-3-like protein 1 (CHI3L1), alternatively known as YKL-40, HC-gp39 (human cartilage glycoprotein-39), or BRP-39 (breast regression protein-39, used for the mouse homologue), is a secreted glycoprotein that has attracted considerable attention for its roles in inflammation, tissue remodelling, and cancer. The name "YKL-40" originates from its three N-terminal amino acids, tyrosine (Y), lysine (K), and leucine (L), and its approximate molecular weight of 40 kilodaltons. [190] CHI3L1 is categorized within the glycoside hydrolase family 18 (GH18), which originated to hydrolyze β-1,4-glycosidic bonds in chitin, a structural polymer of N-acetylglucosamine found in arthropod exoskeletons and fungal cell walls. [191]

Origins and Evolution of the GH-18 Chito-lectin Sub-family

The glycoside hydrolase 18 family is an evolutionarily ancient group of enzymes present in archaea, bacteria, and eukaryotes. Although mammals do not produce or degrade chitin, the human genome still encodes eight GH18 genes. These fall into three major branches: chitobiases, which break down short chitin fragments; chitinases/chitolectins, which include both active enzymes and inactive chitin-binding proteins; and stabilin-1-interacting chitolectins, which likely function in immune regulation rather than carbohydrate metabolism.

Interestingly, the chitinase/chitolectin subgroup, the one CHI3L1 belongs to, shows a clear expansion in deuterostomes, a major branch of animals that includes vertebrates and echinoderms like sea urchins, suggesting an increasing functional importance in this lineage. However, evidence from earlier-diverging species shows that this gene group did not originate within deuterostomes.

In particular, a chitinase gene closely related to human chitinases and chitolectins was identified in *Hydractinia echinata*, a species of the cnidarian phylum, that includes jellyfishes, sea anemones and corals, and represents one of the most primitive animal lineages. Since cnidarians branched off before the evolutionary split between deuterostomes and protostomes (such as insects, worms, and mollusks), this finding indicates that the chitinase/chitolectin group was already present in the common ancestor of all animals. [192]

In contrast, the social amoeba *Dictyostelium discoideum*, which is even more distantly related to animals, lacks clear orthologs of this group. While it does possess several GH18-related genes, these are more closely related to chitobiases proteins and likely serve functions related to bacterial digestion rather than development or immunity. At the same time, since the eukaryotic ancestral genome included orthologs of the mammalian chitinase/chitolectin group, and the presence of chitinases in plants indicates it did, they likely have been lost in the *D. discoideum* lineage, rather than emerged later on.

Together, these findings suggest that chitinase/chitolectin genes are evolutionarily ancient but may have been lost in certain unicellular lineages like amoebozoans, where the GH18 family specialized for digestion, while retained, and later expanded, in animals, particularly in those with more complex immune systems. In support of this, human GH18 genes are clustered near the MHC paralogon on chromosome 1, a region derived from the ancestral MHC gene complex, and several members of this family are now known to participate in type 2 (Th2) immune responses. This points to a gradual shift from digestive to immunomodulatory roles for these proteins as animals evolved more specialized forms of immunity, possibly contributing to the emerging interface of innate and adaptive immunity during early vertebrate history. [192]

Gene Architecture, Promoter Logic & Alternative Splicing

The gene encoding CHI3L1 is located on human chromosome 1q32.1, embedded within the extended major-histocompatibility-complex paralogon, a genomic region enriched for genes involved in both innate and adaptive immunity.

The canonical human CHI3L1 locus spans ten exons. A GC-rich core promoter harboring a critical Sp1 (specificity protein 1) binding motif is essential for activating CHI3L1 transcription, particularly during monocyte-to-macrophage differentiation. Immediately upstream, a high-affinity NF-κB binding site integrates inflammatory signals; once secreted, CHI3L1 can engage RAGE to amplify NF-κB activity, creating a feedback loop that is particularly evident in chronic inflammatory disease. Additional regulatory input from STAT-3 (signal transducer and activator of transcription 3), AP-1 (activator protein 1), and C/EBPβ (CCAATT/enhancer-binding protein beta) further links CHI3L1 expression to cytokines such as IL-6 and IL-13, as well as to LPS and oxidative stress. [193]

Despite robust transcriptional control, CHI3L1 is conservative at the RNA-processing level, with only one reproducible alternative splice form documented: the ΔExon 8 variant, which removes codons 309–381. This isoform retains the N-terminal signal peptide but fails to

undergo secretion, instead accumulating intracellularly in cell types such as skeletal muscle and glioblastoma lines. The absence of the C-terminal β-barrel may underlie this retention. Given that the full-length protein is catalytically inactive, the ΔEx8 isoform is unlikely to possess enzymatic function and may instead serve as an intracellular decoy or regulatory protein, although its precise role remains to be fully elucidated.

CHI3L1 Architecture & Post-translational Ornamentation

CHI3L1 is a chitolectin. The etymology of the word *chitolectin* explains the main characteristic of the protein: the name derives from the Greek *chitōn* (χιτών), meaning "tunic" or "covering", and from the Latin *legō*, meaning "to gather, to collect", with reference to the protein's ability to bind chitin. A chitolectin differs from an actual chitinase because it is enzymatically inactive and, while it binds chitin, it cannot cleave it nor degrade it. However, it retains (and has refined) its ability to read glycans.

CHI3L1 is secreted as a 383-residue glycoprotein that emerges from the endoplasmic reticulum after cleavage of a 20-amino-acid hydrophobic signal peptide. The mature protein adopts a $(\beta/\alpha)_8$ TIM-barrel fold (TIM stands for triosephosphate isomerase, the protein in which this fold was first discovered) at the core of the protein, typical of the glycoside-hydrolase family 18. [194] Two disulfide bridges, C26-C51 and C300-C364, stabilize the N- and C-terminal regions of the TIM-barrel.

In this case, however, the usually conserved catalytic residues anchored to the barrel have diverged from the ancient ones: instead of the canonical DxDxE hydrolytic motif, we find a DxAxL. These point mutations, Asp 138 to Ala and Glu 140 to Leu, abolish hydrolysis yet preserve the 43 Å carbohydrate groove that traverses the C-face of the barrel. Here, nine subsites (numbered -6 to +3) accommodate oligosaccharides in a flexible and length-dependent manner, without cleaving them. The groove is lined, on the outside and inside, by three conserved cis-peptide bonds, which induce specific loop geometries. [194] The resolved crystal structures (and their corresponding PDB ID) of CHI3L1 alone and in complex with chito-oligosaccharides of different lengths are summarized in Table 1.

Aromatic clamps line the groove, promoting hydrophobic stacking interactions with the hydrophobic sides of the bound sugar rings. Among them, Trp99 and Trp352 additionally participate in ligand recognition by making hydrogen bonds. However, of all residues binding the sugars, only Trp31 and Trp352 are highly conserved. [194]

Trp99 acts as a dynamic gatekeeper to the groove, and undergoes a ligand-induced rotation that expands the cleft. This proper redirection allows to stabilize the chito-oligosaccharide, that bends and twists, in the groove. This accommodating movement by Trp99 is well depicted in Fig. 8. In true family 18 chitinases, Trp99 is already in the chitin-binding open position, but due to the different orientation, the CHI3L1 site is wider. [194] This may be the reason as to why human chitotriosidase, unlike CHI3L1, cannot bind insoluble chitin. Interestingly, the plant chitinase hevamine - found in *Hevea brasiliensis* or rubber tree, of the phylum *Streptophyta* - which acts as a defense endochitinase against pathogenic fungi and bacteria, but also insects, has a glycine in place of the tryptophan. [195] The glycine is much less bulkier, and ensures a completely open chitin-binding groove and full accessibility of the central subsites near the catalytic residues. [194]

Additional stabilizing interactions include Glu70, Arg263, Tyr141, and Asn100, which mediate specific hydrogen bonds with the sugar hydroxyls and N-acetyl groups. The α+β domain inserted after strand β7 also participates in ligand binding via Arg263.

Interestingly, binding specificity is not uniformly distributed across the groove. Long chito-oligosaccharides (e.g., $GlcNAc_5$) tend to engage the central subsites (–2 to +2), while shorter fragments like chitobiose preferentially associate with distal sites (e.g., –5 and –6). This spatial asymmetry suggests that CHI3L1 may distinguish between ligands based on size and flexibility, possibly supporting a dual-site binding mechanism, one distal and one central, though the functional consequence of this remains speculative. The binding site for chito-oligosaccharides can be schematically visualized in Fig. 9. The binding affinity (Kd) of the chito-oligosaccharides for CHI3L1 has been recently measured by ligand-induced intrinsic fluorescence changes (attributed to the various tryptophans involved in the binding); in particular, the Kd was 88.7 μM for tetraacetyl-chitotetraose, which contains 4 GlcNAc, 10.4 μM for pentaacetyl-chitopentaose, and 4 μM for hexaacetyl-chitohexaose, with 6 GlcNAc. [196]

Beyond chitin recognition, CHI3L1 also interacts with endogenous glycosaminoglycans, most notably heparan sulfate (HS). [197] This interaction is mediated not through the classical groove but via a distinct C-terminal KR-rich domain (residues 334-345), which is functionally indispensable for binding HS and facilitating Syndecan-1/integrin $\alpha_v\beta_3$-driven signalling pathways. [198, 197] Functional studies, including peptide competition and point mutagenesis, have definitively proven this site is indispensable for binding heparin and for promoting angiogenesis. [198] Interestingly, this same C-terminal neighborhood contains the domain (residues 325-338) that is essential for activating the Akt signaling pathway in colonic epithelial cells. Deletion of this segment or blockade with a targeted antibody abrogates Akt activation, highlighting its role as a critical signaling interface. [199]

In contrast, the GRRDKQH motif at residues 143-149, despite resembling canonical heparin-binding sequences, appears not to participate in fully-sulfated heparin binding according to crystallographic studies. [194]

Therefore, CHI3L1 is a versatile glycan reader, capable of sensing both pathogen-derived and endogenous glycans. The wide and accommodating hydrophobic cleft, as well as the HS-binding cationic patch, positioned on the surface exposed to the solvent and possibly strategically close to the groove itself, transform CHI3L1 from a former molecular scissors into a sensing and signaling molecule.

CHI3L1 has a single N-glycosylation site at Asn60; crystal electron density shows two β(1→4)-linked GlcNAc residues. [194]

CHI3L1 may form dimers that are resolved in SDS-PAGE gel-electrophoresis under non-reducing conditions and disappear in the presence of β-mercaptoethanol. [200]

Fig. 10 recapitulates visually the important domains of CHI3L1 on its crystal structure.

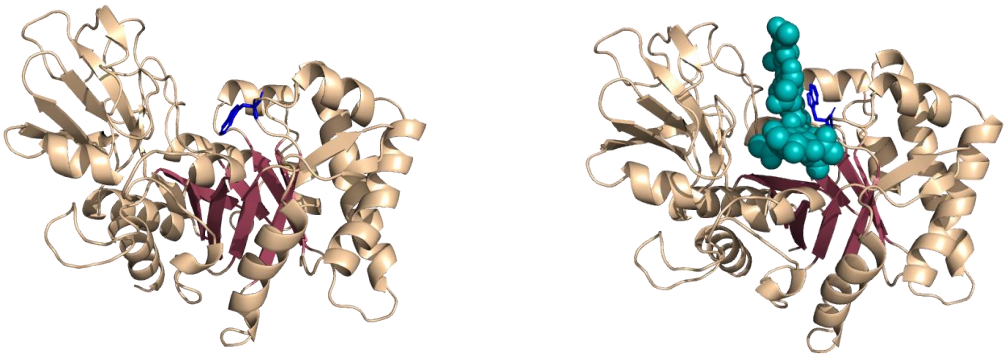


Figure 8. *Crystal structure of CHI3L1 alone (left) and when bound to hexaacetyl-chitotetraose (right)*. The first structure comes from the pdb 1NWR and clearly shows the residue Trp99 (in blue) directed toward the inward of the chitin-binding cleft when no chito-oligosaccharides are bound. The second structure comes from the pdb 1NWT and shows the transition of Trp99 to an open state when the cleft binds hexaacetyl-chitoheaxaose (6 GlcNAc), shown in cyan.

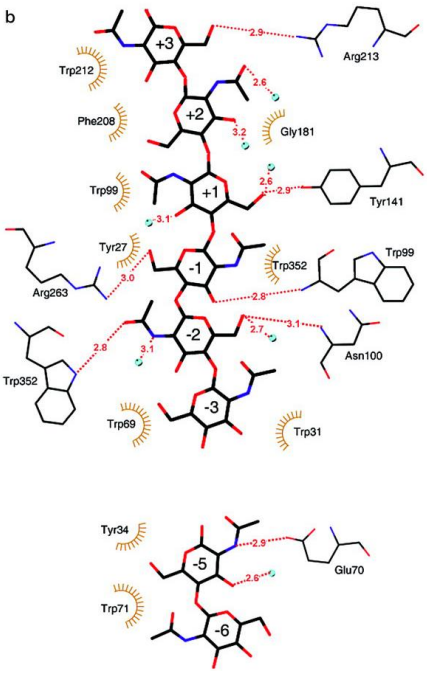


Figure 9. *Visualization of chito-oligosaccharides binding by CHI3L1*. Schematic of protein-chitin fragments interactions across subsites –6 to +3. The aromatic residues lining the carbohydrate-binding groove are shown in ball-and-stick representation. Carbohydrate and protein side chains are shown in thick and thin black lines, respectively. Hydrogen bonds (red dashed lines), hydrophobic contacts (brown half-circles), and water molecules (blue spheres) are indicated.
Image reproduced from Fusetti, et al., 2003. [201] Copyright 2025 Elsevier Inc.

Crystal Structure PDB ID	**Author(s)**	**N° of GlcNAc in the ligand**
1NWR	Fusetti et al., 2003	/
1NWS	Fusetti et al., 2003	2
1NWU	Fusetti et al., 2003	4

1NWT	Fusetti et al., 2003	5 and 6
1HJW	Houston et al., 2003	8

Table 1. *Summary of some crystal structures of CHI3L1 in complex with their chito-oligosaccharide ligands.* Fusetti, et al., 2003, solved the crystal structure of CHI3L1, alone and in complex with its main chito-oligosaccharide (COS) ligands. [194] The number of N-acetylglucosamine units in each structure goes from two to 6, the latter being the one with the highest affinity of the CHI3L1 carbohydrate binding groove. Houston et al., 2003, solved instead the crystal structure of CHI3L1 in complex with a chitin octamer. [202]

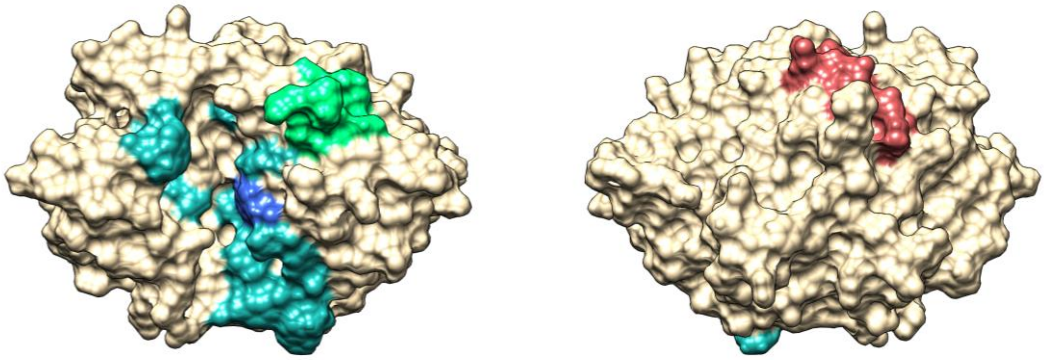


Figure 10. *Surface representation of the front view (left) and rear view (right) of CHI3L1 structure*. The crystal structure comes from the pdb 1NWR. The chitin-binding domain of CHI3L1 is shown in aquamarine, with the key residue Trp99 highlighted in blue. The putative heparin binding domain (W325-V338) is shown in green and the heparan sulfate binding motif (G122-H128) on the opposite side in red.

Cellular Sources and Inducible Regulation of CHI3L1 Expression

In normal tissues, CHI3L1 is constitutively or inducibly secreted by a remarkably broad spectrum of cell lineages: macrophages, neutrophils, stem and bone-derived cells, articular chondrocytes, fibroblast-like synoviocytes, endothelial cells, vascular smooth muscle cells, hepatic stellate cells, and mammary or other epithelial cells. [197] In the central nervous system, its principal source is reactive astrocytes, while activated glial cells and epithelial tissues also contribute under inflammatory or stress-related conditions. [204]

A wide variety of tumour cells, including those of breast, gastric and brain origin, as well as tumor-associated macrophages (TAMs) and stromal cells such as cancer-associated fibroblasts (CAFs), are robust producers in the tumour microenvironment. [205, 206] Additionally, differentiated immune effectors such as T lymphocytes and natural killer (NK) cells can express CHI3L1 in response to activation cues. [207, 93]

Transcription-translation is exquisitely sensitive to extracellular signals and cellular context. Matrix perturbation and miRNA targeting (e.g., miR-24) down-modulate expression, whereas a diverse cytokine milieu fine-tunes production. Interleukins IL-1β, IL-6, IL-13, IL-4, IL-17, IL-18, tumor necrosis factor-alpha (TNF-α), and interferon-gamma (IFN-γ) all act as strong inducers in cell-type-specific settings. In contrast, IL-1β combined with transforming growth factor-beta (TGF-β) represses expression in human chondrocytes, and TNF-α or basic fibroblast growth factor (bFGF) suppresses it in glioblastoma cells. Platelet-derived growth

factor (PDGF), IL-7, IL-11 and IL-12 appear neutral in cartilage. [208, 197] Additional inputs include insulin-like growth factors I and II, which upregulate CHI3L1 in guinea-pig but not human chondrocytes, and microbial stimuli such as lipopolysaccharide (LPS), which acts via NF-κB activation to induce CHI3L1 expression. [193] Hormonal signals, including vasopressin and parathyroid hormone-related peptide (PTHrP), further extend the regulatory repertoire. [209] A broad array of stressors, including hypoxia, ionizing radiation, oxidative stress, serum deprivation, confluence, and p53 blockade, markedly enhance CHI3L1 transcription in U87-MG (Uppsala 87 malignant glioma) cells. [210] Finally, its expression increases with age, likely reflecting "inflammaging," a low-grade, chronic inflammatory state characteristic of the elderly. [211, 212]

In summary, CHI3L1 functions as an acute and chronic sensor of the tissue microenvironment, integrating structural, cytokine, microbial, hormonal, and stress signals into a unified secretory response.

Inter-Species Conservation of CHI3L1

The same gene is known as YKL-40 in humans and BRP-39 in mice and is detectable throughout both prokaryotic and eukaryotic taxa and is conserved across a wide range of metazoans, reflecting strong evolutionary pressure on its structure and function. Human CHI3L1 shares 73.3% amino acid identity with its murine orthologue. Sequence conservation across other mammals remains similarly high, with homologies of 79.6% in rats, 96.6% in monkeys, and 83.8% in sheep. [213] Sequence divergence between the 383-residue human and 381-residue mouse proteins is predominantly confined to surface-exposed loops and does not disrupt the conserved $(\beta/\alpha)_8$ TIM-barrel fold.

Expression of CHI3L1 is not limited to experimental models; strong immunoreactivity and elevated circulating YKL-40 levels have been observed in naturally occurring canine tumors, where its expression correlates with disease progression. [214] Functional studies in CHI3L1-deficient mice reveal impaired IL-13–driven type 2 inflammation, attenuated fibrotic remodeling, and disrupted alternative macrophage activation, confirming a conserved immunomodulatory role. [215]

At the transcript level, humans express a shorter splice variant of CHI3L1 lacking exon 8, which results in a non-secreted cytoplasmic isoform. This variant has not been identified in mice, possibly suggesting a primate-specific mechanism of post-transcriptional regulation.

Receptors & Binding Partners: An Interaction Atlas

CHI3L1, despite lacking enzymatic chitinase activity, retains potent signalling functions. Much of its influence lies in its ability to engage with multiple cell-surface receptors and binding partners, triggering a range of downstream responses. Rather than acting through a single canonical receptor, CHI3L1 interacts with a series of surface molecules. These interactions form a dynamic signalling web that modulates inflammation, immunity, tissue remodelling, and oncogenesis, depending on context.

1) IL-13Rα2: A High-Affinity Gatekeeper of Survival Signalling

Among the best-characterized receptors for CHI3L1 is interleukin-13 receptor alpha 2 (IL-13Rα2). This receptor binds CHI3L1 with remarkably high affinity (dissociation constant ~12

pM), making it one of the most sensitive components of the CHI3L1 signalling network. In support of this, many of the identified CHI3L1 receptors interact and complex with IL-13Rα2, except for RAGE. [216]

In type 2 inflammation, IL-4 and IL-13 generally induce M2 polarization and are produced by Th2 cells. They bind the IL-13Rα1/ IL-4Rα heterodimer. [217] Therefore, IL13Rα2 was initially thought to act as a decoy receptor to inhibit response to IL-13. However, more recent studies have demonstrated that IL-13 also signals and regulates a variety of cellular and tissue responses via IL-13Rα2, possibly thanks to the presence of other ligands, such as CHI3L1. [216]

CHI3L1 and IL-13 do not compete for IL-13Rα2 binding and signaling, and they do not bind to identical locations on IL-13Rα2, but they form a multimeric complex to activate the MAPK/Erk, Akt, and Wnt/β-catenin. This interaction decreases apoptosis and inflammation but concomitantly increases invasion and metastasis of cancer cells. The trimeric binding is enhanced by TMEM219 (Transmembrane Protein 219), possibly acting as a dampening system on apoptotic signals, given that TMEM219 normally functions as a death receptor. Instead, Galectin-3 competes with TMEM219 for IL-13Rα2 binding, still in its trimeric complex, to increase apoptosis and M2 polarization, for example inducing fibrosis of the airway epithelium, as shown in Fig. 11. [197]

With respect to carcinogenesis, CHI3L1 secreted from M2 macrophages interacts with IL-13Rα2 on the plasma membrane of gastric and breast cancer cells, triggering the activation of the MAPK signaling pathway. Upon activation, Erk and JNK signaling enhance the expression of MMPs, which degrade the ECM in the tumor microenvironment, thereby facilitating metastasis. Accordingly, CHI3L1 also induces MMP-9 production by macrophages and enhances matrix degradation in triple-negative breast cancer mouse models. [197] Moreover, high IL-13Rα2 expression in breast cancer with brain metastasis predicted worse survival after metastasis diagnosis. IL-13Rα2 was indeed essential for cancer-cell survival by promoting proliferation. [218]

In the context of HER2+ breast cancer, blockade of CHI3L1 via ILRα2-Fc enhanced CAR-T cell therapy efficacy in mouse models. However, CHI3L1-mediated effects on IL-13Rα2 expressed in the microenvironment immune cells are probably intracellular, given that this receptor is located only intracellularly in NK and T cells. It was also observed that sRAGE competes with IL13Rα2 on CHI3L1 binding in competitive ELISA. [93]

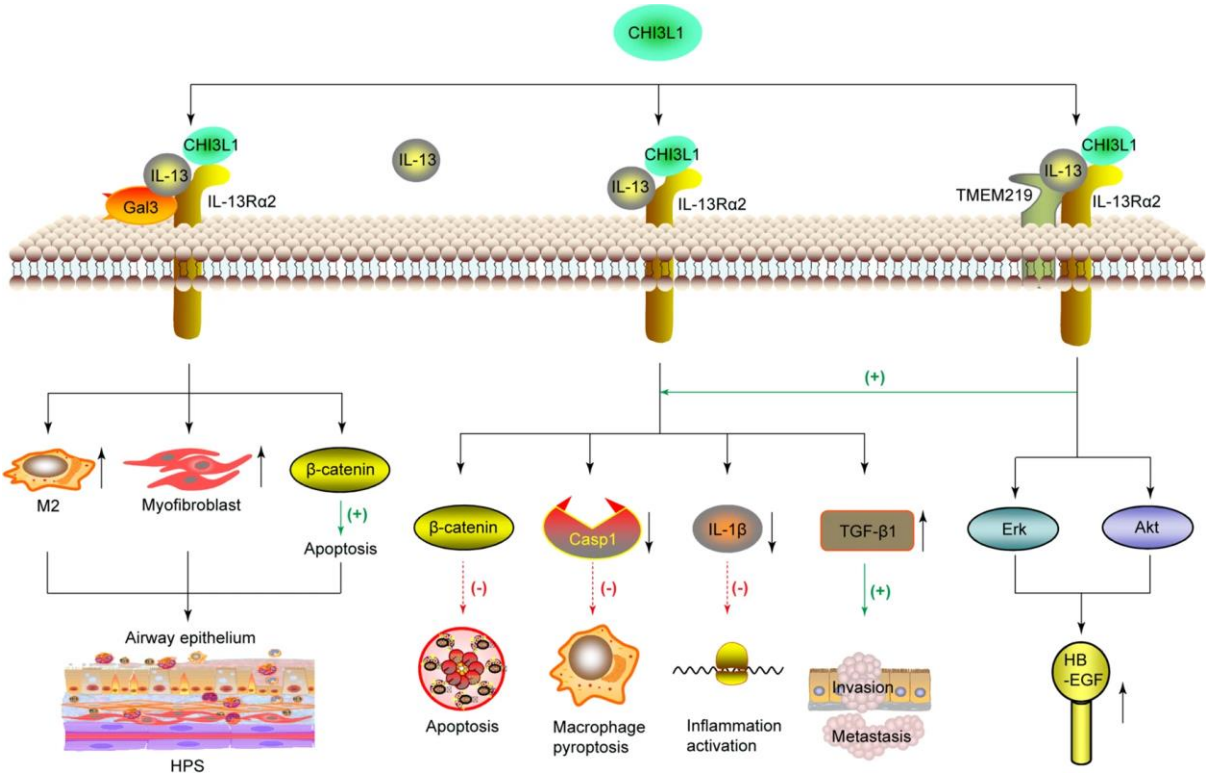


Figure 11. *Schematic drawing of the dynamic interaction between CHI3L1 and the receptor IL-13Rα2.* When the receptor binds both CHI3L1 and IL-13 in a trimeric complex, the downstream effects are increased invasion and metastasis, and diminished inflammation activation, macrophage pyroptosis, and apoptosis. This binding is potentiated by TMEM219 while its effects are reverted when Galectin-3 is present, resulting in apoptosis, M2 macrophage skewing and tissue remodelling.
Image reproduced from Zhao, et al., 2020. [219]

2) Galectin-3: A Molecular Switch with Dual Roles

Galectin-3 (Gal-3) is an intracellular and extracellular lectin that binds β-galactoside sugars, but its interaction with CHI3L1 occurs through a direct protein-protein interface that does not depend on carbohydrates. Gal-3 serves as a potent modulator of CHI3L1 function. When it binds CHI3L1, especially in the glioblastoma immune environment, the resulting complex activates a PI3K-Akt-mTOR signalling cascade that promotes the recruitment and polarization of tumour-associated macrophages (TAMs) into an M2-like, pro-tumour phenotype. This process is driven by a transcriptional programme involving NF-κB and CEBPβ. [220]

Gal-3 doesn't just activate downstream signalling, it also reshapes the receptor landscape. It competes with TMEM219 for IL-13Rα2 binding, thereby reducing CHI3L1–IL-13Rα2–TMEM219 signalling and instead favouring CHI3L1 engagement with CRTH2, a different receptor with distinct outcomes. [197]

Gal-3 function can itself be counteracted by galectin-3 binding protein (Gal3BP), a secreted antagonist that prevents Gal-3 from interacting with CHI3L1. Gal3BP can reduce TAM infiltration and restore anti-tumour immunity in animal models. [220]

Beyond cancer, the Gal-3-CHI3L1 axis is relevant in neurodegenerative disease. In Alzheimer's disease, Gal-3 and CHI3L1 are co-elevated in cerebrospinal fluid, and Gal-3 is primarily expressed by microglia around amyloid-beta plaques, suggesting a role in neuro-inflammation. [221]

3) CRTH2: A Driver of Fibrosis and Alternative Macrophage Activation

CRTH2, also known as the prostaglandin D2 receptor 2 (DP2), is another confirmed receptor for CHI3L1. It plays a role in promoting the differentiation of M2 macrophages, which are involved in tissue repair, fibrosis, and immunosuppression. [222]

In models of pulmonary fibrosis, CHI3L1-CRTH2 signalling has been shown to drive profibrotic responses, including the activation of fibroblasts and deposition of extracellular matrix. This pathway becomes especially dominant when Gal-3 is present, as Gal-3 diverts CHI3L1 away from IL-13Rα2 and toward CRTH2. Thus, the fibrotic versus survival-promoting roles of CHI3L1 can be dictated by the relative availability of Gal-3 and the expression of CRTH2. [223]

4) CD44: Supporting Cell Migration and Oncogenic Progression

CD44, a transmembrane glycoprotein best known as a hyaluronan receptor, is another cell-surface partner for CHI3L1. When CHI3L1 binds CD44 in a trimeric complex with IL13-Rα2, it activates Erk and Akt pathways, as well as β-catenin phosphorylation. These events promote cell proliferation, survival, migration, and invasion, traits that are especially relevant in tumour progression. In breast and gastric cancer models, CHI3L1-CD44 interaction has been linked to more aggressive disease behaviour. [197]

5) Glycan-Based and Matrix Interactions: Modulating the Extracellular Environment

In addition to classical receptor binding, CHI3L1 also interacts with components of the extracellular matrix and glycosaminoglycan molecules, further broadening its functional range.

Hyaluronic acid (HA) is an anionic, nonsulfated glycosaminoglycan distributed widely throughout the extracellular matrix of tissues; it is unique among glycosaminoglycans as it is non-sulfated. [224] CHI3L1 has two putative binding motifs (residues 147-155 and 369-377) for HA, though direct structural validation remains incomplete. CHI3L1-HA interactions are believed to contribute to extracellular matrix remodelling. [197]

Heparan sulfate (HS) is a glycosaminoglycan very closely related in structure to heparin, consisting of a variably sulfated repeating disaccharide unit. [225] It occurs in a proteoglycan in which two or three HS chains are attached in close proximity to the cell surface or the extracellular matrix proteins. [226] These interactions facilitate the clustering of syndecan-1 and integrin αvβ3 on the cell surface, which activates focal adhesion kinase (FAK) and Erk signalling. This promotes both VEGF-dependent and VEGF-independent angiogenesis, as well as tumour proliferation. [197]

Biological Functions of CHI3L1: Beyond Chitin Binding

CHI3L1 is best recognized as a master regulator of type 2 inflammation. [227, 97] It plays a role in skewing the immune response toward a Th2 phenotype, driving M2 macrophage and dendritic cell differentiation, while also signaling to limit tissue damage and favoring ECM remodeling.

CHI3L1 is not expressed in resting monocytes but is strongly induced during macrophage M2 differentiation through the transcription factor Sp1 (specificity protein 1). [228] SP1 expression is increased by RAGE signaling itself and, notably, Sp1 shows functional interference with

NF-κB and acts directly through a subset of its binding sites. [229] Since RAGE is until now the only known receptor of CHI3L1 to activate NF-κB, which can mediate continued CHI3L1 production, these findings suggest a positive feedback loop where CH3L1 is able to increase its expression by activating NF-κB via RAGE binding. [230]

The immune cytokine environment is a major driver of CHI3L1 expression. Proinflammatory cytokines such as IL-6, IL-13, IL-1β, IFN-γ, and TNF-α, as well as bacterial components like LPS, all induce its production. Hormonal inputs such as vasopressin and parathyroid hormone-related protein can also modulate its expression. In turn, CHI3L1 stimulates the secretion of pro-inflammatory and chemotactic molecules as CXCL8 (C-X-C motif chemokine ligand 8)/IL-8 and CCL2 (chemokine C-C motif ligand 2)/MCP-1 (monocyte chemoattractant protein-1), mediating macrophage recruitment and promoting angiogenesis. [231]

Interestingly, while CHI3L1 is generally associated with Th2 dominance, it is also expressed in Th17 cells and upregulated during diverse infections of mucosal tissues. Therefore, its role in modulating host-microbe interactions, reminiscent of its evolutionary function, suggests an immunoregulatory function that extends beyond strict T helper paradigms. Knockout studies in T cells have shown that the absence of CHI3L1 enhances Th1 responses and results in more effective control of tumor growth and metastasis. These findings support a model in which CHI3L1 tempers pro-inflammatory immunity while promoting tissue repair.

CHI3L1 in Pathology

Exaggerated levels of CHI3L1 exacerbate a variety of diseases and become especially relevant in chronic allergic diseases, fibrotic disorders, and cancer.

In asthma, CHI3L1 promotes eosinophilic infiltration, airway remodeling, and Th2-type inflammation, while in pulmonary fibrosis, it drives M2 macrophage polarization and fibroblast activation via CRTH2 signaling. [215, 223]

In the gut mucosa, elevated CHI3L1 are also observed in colitis and inflammatory bowel disease. CHI3L1-mediated modulation of host-microbiota interactions and its expression in the inflamed colonic mucosa promotes Escherichia coli and Salmonella typhimurium infection by enhancing the adhesion of these bacteria to intestinal epithelial cells (IECs). [232] In parallel, it contributes to both tissue repair and tumorigenesis in the gut; specifically, as previously mentioned, this is also achieved through the downregulation of the competitor on RAGE binding, S100A9. [99]

In neurodegenerative diseases such as Alzheimer's, CHI3L1 is produced by reactive astrocytes and co-expressed with galectin-3 around amyloid plaques, serving as a marker of neuroinflammation and disease progression. [221]

Most notably, in cancer, CHI3L1 seizes diverse strategies to help the tumor evade and promote the formation of a profoundly immunosuppressive microenvironment. The CAFs-derived CHI3L1 in breast cancer and the tumor cells-derived CHI3L1 in glioblastoma can induce M2-like macrophages infiltration. [233, 220] In lung cancer, the molecule is able repress tumoricidal genes in both Th1 and $CD8^+$ T cells, while in hepatocellular carcinoma it stimulates the secretion of the immunomodulatory cytokine TGF-β by tumor cells. [207, 234] In lung, pancreatic, and colorectal tumors, CHI3L1 stimulates production of immunosuppressive T-regs in the tumor microenvironment and can consequently be inhibited by CHI3L1 blockade. [235] As discussed, CHI3L1 can even impair the NK cell cytotoxicity in HER2+ breast cancer.

[93] Additionally, in HER2+ models, CHI3L1 supports tumor growth by enhancing fatty acid oxidation and cellular respiration, contributing to metabolic fitness. [236] By comparison, TNBC is characterized by high CHI3L1 expression in both tumor and stromal cells, where it facilitates immune evasion by inducing neutrophil extracellular traps (NETs), reprogramming macrophages toward an immunosuppressive phenotype, and suppressing cytotoxic T-cell responses via upregulation of CTLA-4. [237] Interestingly, chitin-mediated blockade of chitinase-like proteins reduces tumor immunosuppression, inhibits lymphatic metastasis and enhances anti-PD-1 efficacy in TNBC models. [238] In a potentially similar way, chito-oligosaccharides (COS) improve the efficacy of checkpoint inhibitors in mouse models of lung cancer. [239]

More in general, CHI3L1 levels rise in multiple solid tumours and correlate with metastatic spread, most probably due to its ability to favor angiogenesis, promote extracellular matrix degradation through MMP-9, and facilitate tumor cell survival. [197]

The results of a recent text-mining analysis suggested a number of diseases and cancer types potentially associated with CHI3L1, and they can be visualized in Fig. 12a and b.

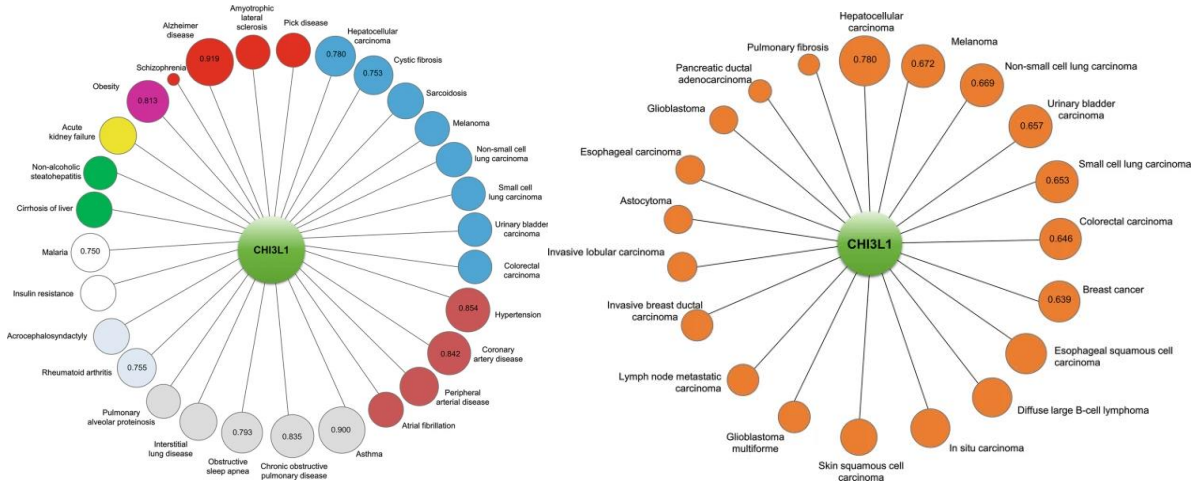


Fig. 12a. *Most statistically significant diseases resulting from a text mining analysis on CHI3L1.*
Fig. 12b. *Most statistically significant cancer types resulting from a text mining analysis on CHI3L1.*
Image reproduced from Yu, et al., 2024. [240]

1.3.4 Evidence That RAGE Is a Functional Receptor for CHI3L1

The receptor for advanced glycation end products (RAGE) has emerged as a bona fide signaling partner for CHI3L1 (YKL-40), extending its recognized repertoire beyond classical DAMPs such as S100 proteins, HMGB1, and AGEs. Multiple lines of evidence, ranging from biophysical assays to functional cell signaling and pathophysiological contexts, support this interaction as both specific and biologically consequential.

ELISA assays demonstrated a direct interaction between CHI3L1 and RAGE with a moderate-to-high affinity, with a reported dissociation constant Kd of approximately 16.65 nM. [93] This binding is not only specific but functionally competitive: CHI3L1 was shown to displace S100A9, a canonical RAGE ligand, in competitive binding assays. [99, 93] This

suggests overlapping, sterically adjacent or conformationally alternative binding interfaces on RAGE.

Moreover, further support comes from experiments employing soluble RAGE (sRAGE). It was found to compete with IL-13Rα2 for CHI3L1 binding in ELISA assays. [93] This indicates that CHI3L1 has sufficient binding affinity to interact with multiple receptors and that its interaction with RAGE is physiologically relevant and not merely incidental. The fact that sRAGE can sequester CHI3L1 adds an additional layer of regulation, suggesting that shifts in soluble versus membrane-bound RAGE isoform expression could modulate CHI3L1 availability and signaling bias in vivo.

1.3.5 *CHI3L1: non-self, self, a mysterious motif*

We have seen up to now that CHI3L1 possesses more than one relevant domain. As such, the literature has tried to clarify the roles of such domains, let first identifying them. But every time an answer seemed to manifest, through time it often transformed into a bigger question. To exemplify this, in 2024 two papers were published: one by Magnusdottir and colleagues, setting new words amidst a more than two decades old discussion regarding the position of the CHI3L1 heparin-binding site; another by Chen et al., proposing a remarkably fresh hypothesis on the molecule adaptation in mammals. [196, 241]

At the present, it is clear to us that CHI3L1 can bind (1) non-self carbohydrates, via the 43 Å groove resolved independently and in the same year by Fusetti et al. (2003) and Houston et al. (2003); and (2) self-glycans, via a site that can recognize heparin and heparan sulfate. [194, 202]

To understand why a protein should be able to recognize chitin via such a well-preserved groove lacking its original catalytic activity, and moreover in humans (or mammals, if we consider the homologs), who do not produce chitin, is not a simple task. One may hypothesize it could be for some sort of defense against fungi and nematodes. [202] For example, the murine Ym1 is a chito-lectin of the same glycoside-hydrolase 18 family, also with mutations in the ancestral enzymatic DXXDXDXE motif. Likewise, CHI3L1, Ym1 can still bind short GlcN oligomers. It is expressed by macrophages upon infection with the nematodal parasite *Trichinella spiralis* and acts as a chemotactic for eosinophils (in turn, CHI3L1 has recently been shown as a direct chemoattractant for tumor-associated neutrophils!). [237] However, compared to the human macrophage chitinase HCHT - against which CHI3L1 has a well-conserved groove in both sequence and structure - Ym1 does not have a well-defined cleft, is poorly conserved, and more negatively charged, indicating a possible divergence of function. The question remains open.

Last December (2024), Chen et al. had the intuition that chitin has a similar structure to peptidoglycan (Fig. 13). [241] Peptidoglycan is a heteropolymer found in the bacterial cell wall, and primarily relevant in Gram-positive bacteria, where it forms thick, multilayered walls. Chen proved that CHI3L1 can bind peptidoglycan and likely scaffolds Gram-positive bacteria of the gut microbiota, by anchoring them into the mucus. Deletion of CHI3L1 production from mice intestinal epithelial cells results in a dysbiotic gut microbiome, with significant loss in Gram-positive *Lactobacillus* and consequent - but reversible with supplementary *Lactobacillus* or fecal microbiota transplant - increased severity of DSS-induced colitis. Thus, CHI3L1 might

have evolved to keep hold of a healthy gut microbiota, sustaining homeostasis and protecting the host against inflammatory injury.

This means that CHI3L1 might have retained its original shape of the groove to interact with non-self carbohydrates in humans. But here comes the second layer of complexity, because CHI3L1 has developed a whole arsenal of interactions with self, too, as we saw above. It binds receptors and it also binds self carbohydrates, sometimes on receptors too. This is the case for heparan sulfate (HS), a proteoglycan on top of syndecan-1 (S1), and heparin.

The possible heparin binding site is not the in groove itself, because heparin is heavily covered with sulfates (SO_3^-) and carboxylates (COO^-), becoming the most negative bio-macromolecule known. The groove is a hydrophobic canion, sculpted to accommodate neutral or slightly polar sugars. For heparin, a site rich in positively charged amino acids, such as Lysines (K) and Arginines (R) is required.

There is one such site close to the C-terminus of the molecule, highlighted by residues 334-345. In 2017, Ngernyuang and colleagues demonstrated that this KR-rich candidate is responsible for heparin binding: an isolated peptide from this region successfully competed for heparin; single point mutations of its basic residues neutralized both heparin binding and, in vivo, downstream pro-angiogenic signaling; a neutralizing antibody recognized that exact site. [198] It was not the first time that this twelve-long amino acidic strand appeared in the literature. An overlapping segment (325-338) was identified in 2011 by Chen and Mizoguchi as indispensable for Akt downstream signaling in epithelial cells. [199]

However, what is even stranger is the presence of a second putative heparin binding site, whose relevance has been dismissed over the years by multiple pieces of evidence. Now, such site is the GRRDKQH motif at residues 143-149, perfectly matching the canonical XBBXBX heparin-binding consensus. However, Ngernyuang did not observe any competition by a peptide from this region, and mutations did not affect downstream signaling. Accordingly, in 2003, Fusetti had observed no binding of heparin at this place by soaking CHI3L1 crystals with HS or heparin-like molecules. Nevertheless, the question remains. Is this motif some evolutionary artifact? A piece that has lost its purpose? It is also remarkably close to the chitin-binding groove. Magnusdottir analyzed that the sequence changed conformation when tetra- (GlcNAc4, A4) to hexa- (GlcNAc6, A6) chito-oligosaccharides (COS) were bound to the groove (Fig. 14). COS had a negative allosteric effect on heparin binding, from 16.4% (A4) to 37.4% (A6) of maximal inhibition, adding another nuance to the story.

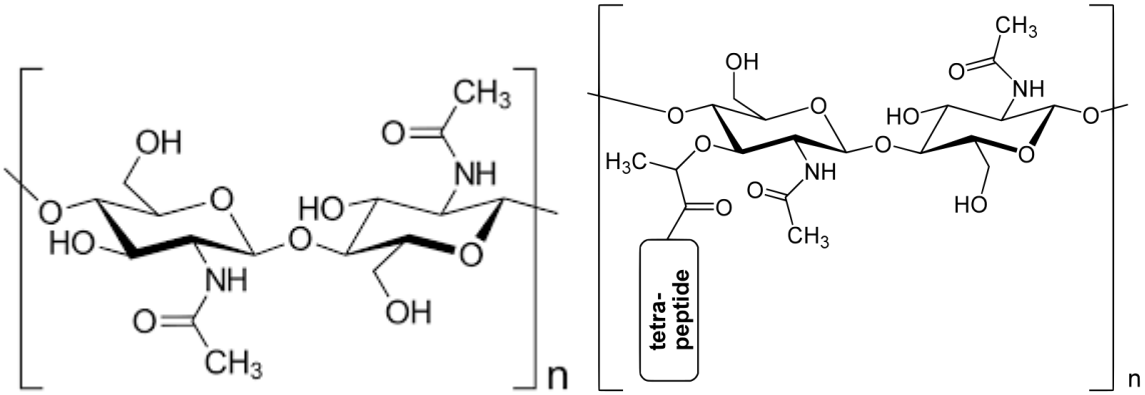


Figure 13. *Comparison of chitin and peptidoglycan repeating units*.

Left: Chemical structure of chitin, composed of β-(1→4)-linked N-acetylglucosamine (GlcNAc) units, forming long, linear homopolymers typical of fungal cell walls and arthropod exoskeletons.
Right: Repeating unit of peptidoglycan, a heteropolymer found in bacterial cell walls, consisting of alternating N-acetylglucosamine (GlcNAc) and N-acetylmuramic acid (MurNAc) residues connected via β-(1→4) linkages, with short peptide chains attached to MurNAc enabling cross-linking and mesh-like structural rigidity.

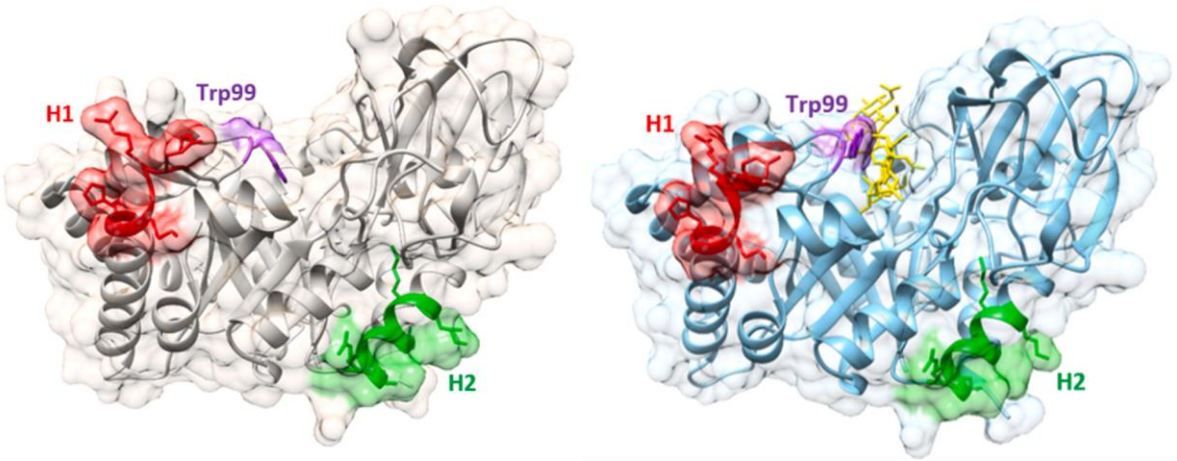


Figure 14. *Structural rearrangements in CHI3L1 upon chito-oligosaccharide binding: local Trp99 rotation and proximal allosteric modulation of the GRRDKQH motif.*
Ribbon and transparent surface rendering of CHI3L1 in its native (PDB: 1HJX, grey) and chitin-bound (PDB: 1HJW, blue) states, highlighting ligand-induced reorganization. The chitin octamer is shown in yellow, and Trp99 is marked in purple to emphasize its rotational shift upon ligand binding. The canonical heparin-binding motif GRRDKQH (residues 143-149) is shown in red, and the C-terminal KR-rich site (residues 334-345) in green. Notably, binding of the chitin oligosaccharide propagates structural changes toward the GRRDKQH loop, suggesting allosteric communication between the groove and this proximal, electropositive patch.
Image reproduced from Magnusdottir et al. [242] Copyright 2024 by the authors.

1.3.6 Chito-oligosaccharides & caffeine, peptidoglycan, proteoglycans

Chito-oligosaccharides (COS) are small homo- or heterooligomers of N-acetylglucosamine and D-glucosamine, derived from the degradation of chitin or chitosan. The literature is poor when it comes to mechanistically demonstrating the functional effects of their binding with CHI3L1. In vertebrates, COS serve as primers for new hyaluronic acid (HA) chains, and since both CHI3L1 and HA are implicated in tissue remodeling and inflammation, CHI3L1 could act as a sensor of extracellular matrix turnover. [194] What is also known is that COS have varying affinity for the groove of the protein, stronger for fully acetylated or for longer COS, yet still in the micromolar range. [243, 196] Long hexaacetyl-COS have proliferating effects on human chondrocytes which are counteracted by shorter triacetyl-COS, possibly competing for CHI3L1 binding site. [243]

Moreover, COS are known to be bioactive modulators of innate immunity. In macrophages, particularly chitobiose and chitohexaose enhance phagocytosis of pathogens, by activating NF-κB and TLR-4 pathways. [244] In differentiated human monocytes (THP-1), the same di- and

hexa- COS are anti-inflammatory, diminishing cytokine production in response to LPS stimulation. [245] In the gut of blue foxes, COS improve barrier integrity by upregulating tight junction proteins and secretory IgA, reduce intestinal inflammation by balancing pro- and anti-inflammatory cytokines, and increase microbial diversity, selectively enriching beneficial taxa such as *Lactobacillus agilis* and *L. murinus* while suppressing opportunistic pathogens like *Fusobacterium*. [246]

Chitin, by targeting CHI3L1 and CHI3L3 secreted by tumor-associated neutrophils in triple-negative breast cancer, reduces their expression and p-Stat3 signaling, disrupting immunosuppression and enhancing immunotherapy with anti-PD-1. [238]

A recent and compelling study by Cho and colleagues (2024) revealed that COS may modulate the CHI3L1-RAGE axis at a more fundamental level: by controlling the abundance of the RAGE receptor itself. [247] Their work provided direct cellular evidence that COS, particularly those of 1-3 kDa, act as potent scavengers of the inflammatory precursor dicarbonyl methylglyoxal. By neutralizing this upstream trigger of glycation stress, COS pre-treatment effectively prevents the subsequent upregulation of RAGE protein expression in intestinal epithelial cells.

Besides COS, also caffeines can bind CHI3L1 groove; caffeine is a natural pan-chitinase inhibitor. In intestinal epithelial cells, it downregulates CHI3L1 mRNA in a dose-dependent manner; inhibits CHI3L1-associated AKT activation and β-catenin nuclear translocation, both in vitro and in vivo; and reduces CHI3L1-mediated bacterial adhesion and invasion in early-colitis. [248, 232, 249] Furthermore, CHI3L1-mediated cell proliferation could be suppressed with caffeine treatment in a dose-dependent manner, controlling and retarding tumor formation. [248] This suggests that caffeine can also be utilized in IBD patients to control colitis-associated cancer development and prevention.

COS, caffeine, peptidoglycan, all recognize CHI3L1 hydrophobic groove, and they appear to modulate CHI3L1 action in several ways.

At the same time, proteoglycans like heparin and heparan sulfate (HS) on receptors such as syndecan-1 can contact the C-terminal KR-rich motif of CHI3L1 and modulate signaling from the side of the host. It may not be a coincidence that negative allosteric modulation exists between COS and heparin binding, possibly fine-tuning the interaction between non-self and self through CHI3L1 - depending on the microenvironment and on CHI3L1 levels, which are likely themselves integrated in the feedback.

In the gut, the dominant ligand is particulate peptidoglycan, engaging the chitin-binding groove to perform a homeostatic, scaffolding role. In a sterile tumor, however, the protein is free to bind cell-surface proteoglycans like heparan sulfate via its C-terminal hub, driving angiogenesis. Critically, these functions are allosterically coupled: the binding of microbial glycans in the groove *may* therefore inhibit the pro-angiogenic heparin interaction, suggesting a mechanism whereby the host's response is tuned by the presence of bacteria.

Of specific relevance, heparan sulfate is the functional mediator of RAGE hexamerization and if CHI3L1 recognizes HS too, close to the groove or further away from it, this may either be a modulatory addition to the interaction, or a way to alter the oligomerization state of the receptor upon binding and, possibly, decrease signaling by other ligands that require it. [250]

1.4 Why a Clear Picture Is Still Missing

Despite three decades of intense work on HER2-targeted therapy, the CHI3L1-RAGE pair remains an informational blind spot. Three factors converge to create this gap.

1.4.1 Multifunctional partners

RAGE is a pattern-recognition receptor whose activity depends on glycosylation, heparan-sulfate driven oligomerization and a changing roster of more than 30 ligands; CHI3L1 is an enzymatically silent chito-lectin endowed with two chemically distinct binding interfaces, a mysterious motif, and pleiotropic, context-dependent signalling. Both molecules therefore resemble hubs rather than linear pathways, making reductionist analysis very fragile.

1.4.2 Structural silence

To date no high-resolution structure (X-ray, cryo-EM or NMR) of the CHI3L-RAGE complex exists; the field relies on ELISA competition, pull-down and signalling read-outs that illuminate whether binding occurs but not how. This hinders rational mutagenesis, inhibitor design and, critically, any attempt to relate single-nucleotide polymorphisms or glycoforms to function.

1.4.3 Non-canonical affinity logic

CHI3L1 and RAGE at first appear unrelated: one evolved from a chitinase scaffold, the other from an immunoglobulin-like adhesion molecule. Only when their full context is considered (RAGE's heparan-sulfate scaffold, CHI3L1's KR-rich heparin patch, the shared dependence on S100/AGE inflammatory milieus) does their convergence become plausible. Mapping such emergent complementarity demands a systems-aware structural approach, not single-variable biochemistry alone.

1.5 Working Hypothesis and Experimental Awareness

We posit that CHI3L1 docks directly onto the RAGE signaling platform, and in doing so, disrupts its function through a multi-pronged mechanism. Our central hypothesis is that CHI3L1 engages the pre-formed, heparan sulfate (HS)-stabilized RAGE oligomer and that this interaction is sufficient to sterically or allosterically abrogate the pro-cytotoxic signaling initiated by other alarmins like S100A8/A9. This hypothesis rests on three converging lines of reasoning.

First, CHI3L1 and S100A8/A9 elicit diametrically opposed signaling outcomes despite competing for RAGE. Where S100A8/A9 activates pro-cytotoxic JNK signaling, CHI3L1 potently suppresses it. This functional antagonism strongly suggests they do not simply displace each other from an identical static site, but rather engage RAGE in fundamentally different ways. This is further underscored by their profoundly different electrostatic natures: S100A8/A9 is an acidic, negatively charged protein, whereas CHI3L1 carries a slight net positive charge at physiological pH (Fig. 15). Such chemical dissimilarity makes a shared binding epitope unlikely and points toward a more complex interaction, possibly involving allosterically coupled binding footprints.

Second, the functional disruption induced by CHI3L1 *could* directly involve modulating the RAGE oligomer itself. The assembly of RAGE into an HS-dependent oligomer is an absolute prerequisite for its downstream signaling, including the ERK (common to HMGB1, S100A8/A9, and S100B), NF-κB, and MAPK pathways. [138, 120] The fact that CHI3L1 binding can downregulate some of these pathways across different pathologies (JNK in HER2+ breast cancer and β-catenin in multiple sclerosis-associated demyelination, Fig. 5) points to a common mechanistic choke point. We hypothesize that CHI3L1 binding either sterically prevents the formation of a fully active signaling complex or actively disassembles the pre-formed oligomer.

Third, the molecular architecture of CHI3L1 is uniquely suited for a multi-contact engagement with the RAGE-HS platform. Its ancient chitin-recognition groove, being highly conserved and structurally stable, presents an ideal surface for a high-affinity protein-protein interaction. This suggests an elegant evolutionary shortcut: rather than evolving a new domain to sense microbial glycans directly, RAGE may have co-opted CHI3L1, recognizing the upregulation of this expert glycan-sensor as a reliable (due to conservation of the chitin-binding domain), amplified signal of an inflammatory event. At the same time, the C-terminal KR-rich heparin-binding domain of CHI3L1 lies on the opposite face than the groove, opening the possibility of a "two-point clamp" that engages both the RAGE protein and the surrounding HS scaffold, potentially even across adjacent receptor oligomers within membrane rafts. This leads also to consider a potential role of the mysterious 143-149 motif. While functionally inert to free heparin, this canonical, positively charged loop is positioned remarkably close to the groove and is known to be conformationally altered by ligand binding. It is perfectly placed to act as a secondary "latch," forming an electrostatic handshake with the negatively charged HS chains that decorate the exterior of the RAGE oligomer, thereby locking CHI3L1 onto the platform with high avidity.

Therefore, our experimental awareness must extend to several variables. The interaction could also be affected by the glycosylation state of RAGE, a known factor in ligand selectivity. Furthermore, small molecules that occupy CHI3L1's groove, such as chito-oligosaccharides or caffeine, must be tested not just as simple competitors, but as potential allosteric modulators that could fine-tune the CHI3L1-RAGE interaction itself.

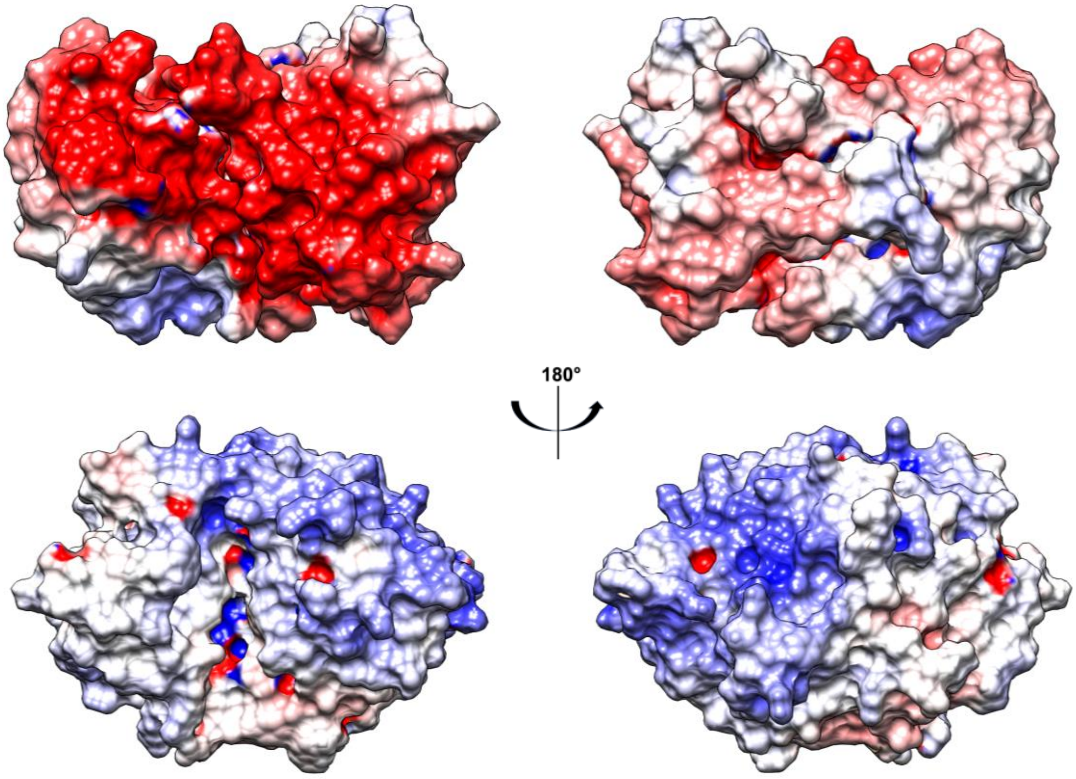


Figure 15. *Electrostatic surface potential of CHI3L1 (top) and S100A8/A9 heterodimer (bottom) at physiological pH (7.4).*
Electrostatic maps were generated using Chimera, PDB2PQR, and APBS tools, with coloring from red (−10 kT/e) to blue (+10 kT/e). The S100A8/A9 heterodimer is shown from two opposite orientations (A: front, B: back, rotated 180° along the Y-axis), highlighting its amphipathic charge distribution and acidic patches relevant for RAGE interaction. Similarly, CHI3L1 is shown from two opposite views (C: front, D: back, rotated 180° along the Y-axis). These surfaces illustrate the electrostatic complementarity potentially guiding receptor engagement and ligand competition at the RAGE interface.

1.6 Experimental Strategy and Planned Approach

To deconstruct the CHI3L1-RAGE axis, a complex biological checkpoint, requires an approach that mirrors its multifaceted nature. A single experimental technique is unlikely to capture the full picture of an interaction shaped by multiple domains, post-translational modifications, and dynamic, concentration-dependent equilibria. We have therefore designed a tripartite strategy that integrates biochemical characterization with parallel tracks of experimental and computational structure determination. This approach is built on the awareness that while an atomic-resolution structure is the ultimate goal, the path to obtaining and interpreting it is paved with foundational biophysical knowledge.

Our investigation begins with a rigorous biochemical and biophysical characterization of the system. We will first establish and quantify the binding affinity between CHI3L1 and the RAGE ectodomain using Enzyme-Linked Immunosorbent Assays (ELISAs). To test the

hypothesis that RAGE engages the conserved carbohydrate-binding groove of CHI3L1, competitive ELISAs will be performed using known groove ligands, specifically chito-oligosaccharides (tetra- and hexa-acetylated[1]) and caffeine. [251] The oligomeric state and homogeneity of the individual proteins and their complex will be assessed in solution using a combination of size-exclusion chromatography (SEC), static light scattering (SLS), and mass photometry. This suite of techniques allows for the determination of absolute molecular weight and possibly the stoichiometry, providing critical insights into the potential for concentration-dependent assembly or disassembly of the complex. Furthermore, given the reported significance of glycosylation, particularly for RAGE, we will employ enzymatic deglycosylation with PNGase F to probe the contribution of N-linked glycans to the stability and binding affinity of the interaction.

In parallel, we will pursue the experimental determination of the complex's three-dimensional structure via X-ray crystallography. This method was chosen as the most suitable for a complex of this predicted size (~75-80 kDa), which lies at the lower limit for routine structure determination by cryo-electron microscopy. Success in this high-risk, high-reward endeavor would provide an unambiguous, atomic-level map of the binding interface, definitively identifying the contact residues and the precise geometry of the interaction.

Recognizing the inherent challenges of protein crystallization, a third, complementary track will employ state-of-the-art computational modeling. We will leverage the predictive power of AlphaFold-Multimer to generate an *in silico* structural model of the CHI3L1-RAGE complex, and one of S100A8/A9-RAGE. This approach is uniquely powerful as it leverages the deep co-evolutionary information encoded in multiple sequence alignments, treating structure not merely as a static shape but as the physical manifestation of evolutionary history. By systematically optimizing model parameters and exploring the conformational space across numerous predictive seeds, we aim to produce a high-confidence structural hypothesis. This computational model serves a dual purpose: it provides invaluable insights in its own right and generates specific, testable predictions about the interface that can guide future experimental work, should crystallization prove intractable.

This integrated strategy is designed to maximize our ability to elucidate the molecular logic of the CHI3L1-RAGE interaction. The insights from each track are intended to inform and validate the others, creating a robust framework for converting a biological problem into a tractable structural and functional understanding. The foundations of our integrated approach are reviewed in the next chapter, "Resolving the Nanoscopic World", concerning both crystallography and computational folding.

1.7 Resolving the Nanoscopic World

1.7.1 From Sequence to Shape: Biology's Most Elegant Puzzle

When the first amino-acid sequences were read in the 1950s it became clear that the one-dimensional code of life somehow folds itself into unique three-dimensional machines.

[1] The hexaacetyl-chitohexaose will be tested after the present work, due to previous material unavailability.

Christian Anfinsen showed in 1961 that a denatured ribonuclease, given only water and time, regains full activity: the instructions are internal. A polypeptide chain can in principle assume a vast range of conformations, and only a few of these may be stable or metastable. Seven years later Cyrus Levinthal sharpened the paradox: if a 150-residue chain tried conformations at molecular speed it would still be searching after the heat-death of the universe (Fig. 16). [252] Folding therefore cannot be a blind hunt; it must follow a low-dimensional set of privileged routes. Half a century of experiment, theory and simulation had refined that intuition into the modern "energy-landscape" picture.

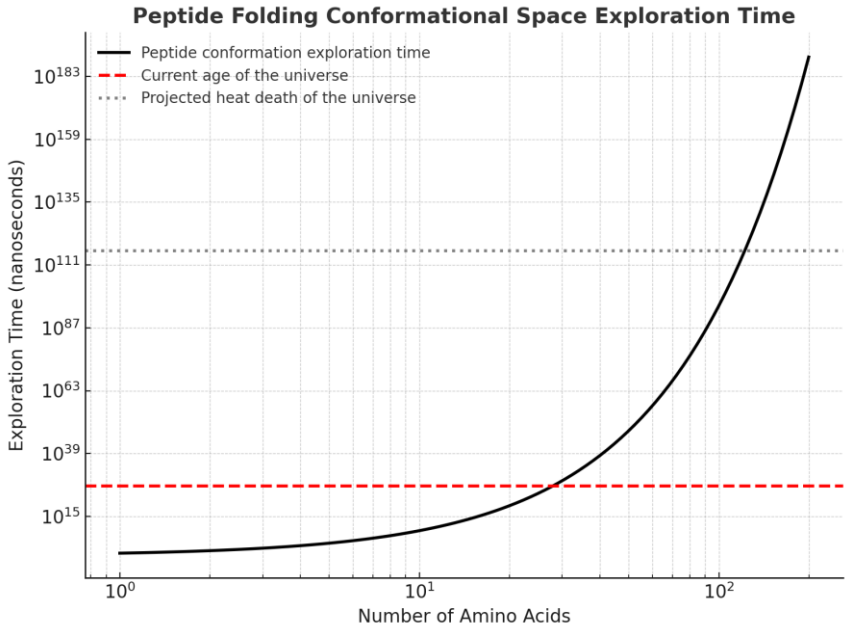


Figure 16. *Levinthal's Paradox.*
The fact that many proteins fold rapidly and reproducibly into their native structures, despite the astronomical number of possible conformations, is known as Levinthal's Paradox. If each torsion angle in a polypeptide chain assumes just three discrete states, a protein of 101 amino acids would have ~3^{100} (≈5×10^{47}) possible configurations. Even at a sampling rate of 10^{13} conformations per second (≈3×10^{20} per year), exhaustive search would require ~10^{27} years - far longer than the age of the universe. Yet most proteins fold in seconds. This striking discrepancy implies that folding must follow guided, energetically favorable pathways rather than random sampling.
Ironically, the original 1969 paper often cited in connection with this paradox has itself become difficult to locate, an observation that has been wryly noted as a paradox in its own right. [252]

1.7.2 The Folding Funnel: Physics Guides the Fold

Levinthal suggested that protein folding can be speeded and guided by the rapid formation of local, stable interactions serving as nucleation points in the folding process. In a more physical nuance, Frederic Richards, Robert Baldwin and, later, José Onuchic and Peter Wolynes reframed folding as diffusion down a funnel-shaped free-energy surface. [253-255] Three physical constraints carve that funnel: (1) hydrophobic residues are expelled from water, driving an early collapse; (2) backbone hydrogen bonds must be satisfied, stabilising secondary

structures such as α-helices and β-sheets; (3) hard-sphere packing fixes the final, tight core. The native state lies ~5-15 kcal·mol^{-1} below the average of mis-folded states. But the descent isn't smooth. The funnel is rugged: local energetic bumps from conflicting interactions can trap the chain. Folding is fast when the energy gap δE_s (stability of the native well) outweighs the roughness σ (variability across misfolded states). The critical ratio
$T_f/T_g \approx \delta E_s/\sigma > 1.3$ marks the border between funneled folding and kinetic arrest. T_f is the folding temperature, at which the free energy of the native state crosses below that of the ensemble of unfolded or misfolded states and the chain is more likely to be folded. Below the glass temperature T_g, the landscape is still rugged, but the protein no longer has enough thermal energy to escape local traps; above it, the chain can still rearrange. T_g marks the dynamic arrest: the destination is not wrong, the route becomes impassable. These parameters explain whether a chain *can* fold fast. Why *should* proteins funnel fast?

The *minimal frustration principle* holds that natural sequences have evolved to reduce conflicting interactions, carving an energy landscape in which the native fold is kinetically accessible (Fig. 17a-b). The funnel picture resolves the paradox.

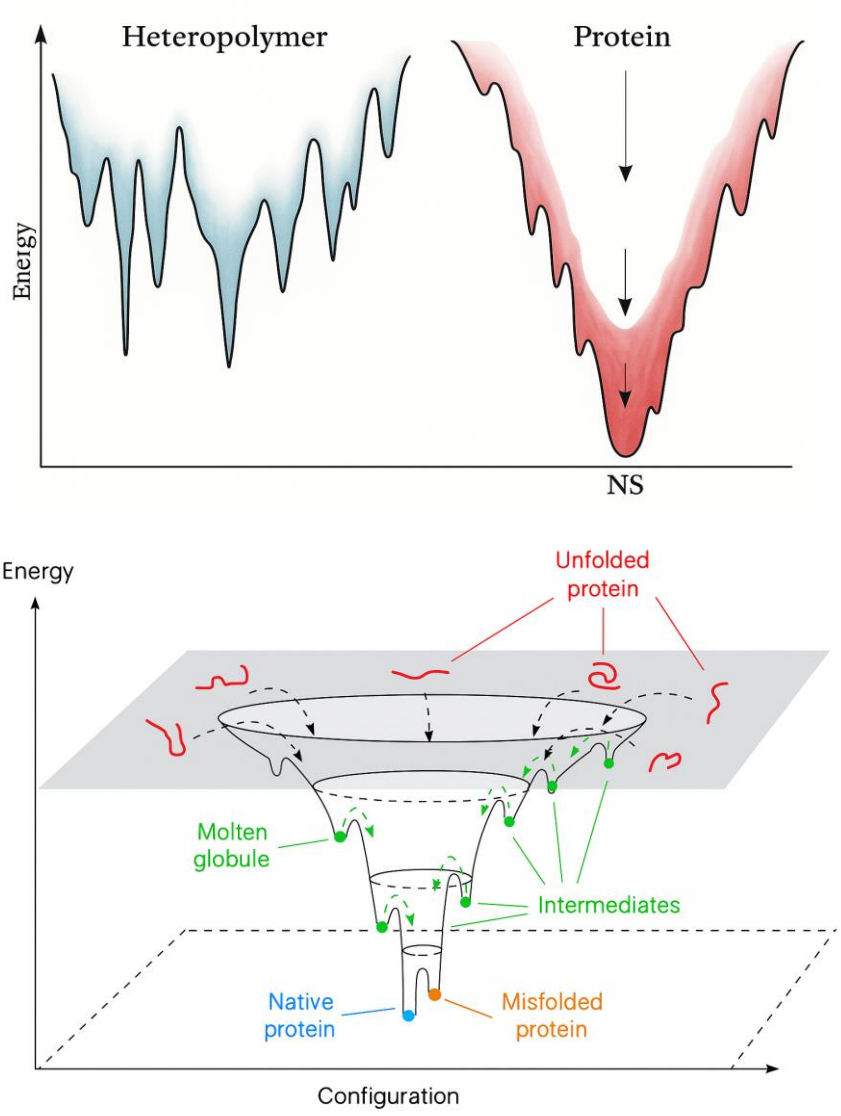


Figure 17a (top). *Sculpted by Evolution: The Funnel Landscape.*

Random heteropolymers (left) exhibit rugged, glass-like energy landscapes with many competing low-energy minima. Such landscapes trap the chain in misfolded states and hinder efficient folding. In contrast, natural proteins (right) have evolved smooth, funneled landscapes that direct the chain toward its native state (NS). This topography ensures both thermodynamic stability and rapid, reliable folding by minimizing energetic frustration and avoiding large kinetic barriers.
Image from Baldovin et al., 2018, has been adapted to create this figure. [256].
Figure 17b (bottom). *Energy Landscape of a Protein.*
The free energy of a protein is shown as a function of its conformational topology. The unfolded ensemble, broad and disordered at the top, descends into a funnel-shaped landscape with countless local minima. Most represent transient intermediates along productive folding routes. Some, like the molten globule, retain partial structure and native-like topology. Others are kinetic traps: misfolded states where the chain becomes irreversibly stuck. The native state lies at the bottom, representing the thermodynamically favoured and functionally active conformation.
Image from Radford et al., 2000, has been adapted to create this figure. [257]

1.7.3 The Funneling of an Idea

The funnel model did not appear fully formed. It crystallized slowly - from Pauling's helices and Kauzmann's core, through Levinthal's paradox, Ptitsyn's molten globules, and Creighton's intermediates, to the rugged energy landscapes of Bryngelson and Wolynes. By the late 1990s, folding was no longer a mystery, but a statistical physics of evolved landscapes.

1.7.4 The Experimental Revelation: Structure by X-Ray Crystallography

The energy landscape offers a solid explanation for the process of folding, the kinetic and thermodynamic ascent from chaos to order. Yet a map of the journey is not a portrait of the destination, and scientists still needed to visualize the final folded forms to clarify protein structure and function. Sequences hold the key, but the cipher was still unknown.

X-ray crystallography provided that crucial visualization. The method's foundation was laid in 1912 by William Lawrence Bragg, who showed that crystals diffract X-rays in a predictable manner, effectively turning diffraction patterns into atomic cartography. Yet globular proteins, unlike simple fibers, defied solution until the invention of heavy-atom methods at Cambridge. The central obstacle was the *phase problem*: while diffraction reveals the intensity of scattered X-rays, the phase information (essential for a 3D map) is lost. In 1953, Max Perutz cracked this problem with his ingenious solution of multiple isomorphous replacement. By soaking protein crystals with heavy atoms (like mercury), whose positions could be located, he created a set of reference points from which the missing phases could be calculated, finally unlocking the path to the structure.

The breakthrough came in 1958, when John Kendrew revealed the intricate fold of sperm-whale myoglobin. Perutz followed in 1960 with the far larger horse hemoglobin. Perutz and Kendrew worked closely together at the MRC Laboratory of Molecular Biology in Cambridge.

Their work was complementary, and they were jointly awarded the 1962 Nobel Prize in Chemistry. It was the first time a protein's form was seen.

1.7.5 Protein Crystallization and X-Ray Crystallography

The structural elucidation of biomolecular complexes at atomic resolution relies on X-ray crystallography, a powerful method whose success is fundamentally dependent on one critical prerequisite: the ability to produce high-quality, well-ordered crystals of the target molecule or molecular complex.

Principles of Crystallization

Protein crystallization is inherently empirical, governed by the principle of controlled supersaturation: proteins must be brought from a state of solubility to one where they self-organize into a periodic, crystalline lattice rather than precipitating randomly. This is typically achieved using vapor diffusion methods, particularly the sitting-drop technique. In this method, a small droplet containing a solution of the protein (or protein complex) is mixed with an equal volume of crystallization solution, sitting close to a larger reservoir containing only the crystallization solution. Researchers usually screen a great number (hundreds to thousands) of different crystallization solutions for one protein because each protein, and particularly each protein-protein complex, requires a distinct set of crystallization conditions; even proteins that have previously been crystallized in isolation may behave unpredictably when in complex. Typically, the crystallization solutions contain precipitants such as polyethylene glycols (PEGs), salts (e.g., NaCl), buffering agents (e.g., Tris), and occasionally volatile alcohols. Upon sealing the plate, water vapor begins to equilibrate between the drop and the reservoir. The vapor moves out of the protein drop into the reservoir because the crystallization solution is less concentrated in the former due to the mixing with the protein solution. Therefore, the drop slowly becomes more concentrated, causing the protein to reach supersaturation, the prerequisite for nucleation and crystal growth.

When a protein solution becomes supersaturated, that is, when it contains more solute than can be sustained under equilibrium conditions, it enters a thermodynamically unstable state. In this regime, protein molecules begin to leave the solution in search of a lower-energy, more ordered state: a crystal. In particular, following Gibb's equation ($\Delta G=\Delta H-T\Delta S<0$), the negative enthalpy ΔH must outweigh the entropic cost of ordering ($-T\Delta S$).

The first step in this process is nucleation, the formation of a small, transient, and ordered cluster of protein molecules. This step is energetically costly: the surface tension of tiny clusters tends to drive redissolution, making them inherently unstable. Only once such a nucleus exceeds a critical size does it become stable enough to act as a scaffold for further molecular addition. Because this event depends on rare and spontaneous local arrangements, nucleation is inherently stochastic, often representing the primary bottleneck in crystallization experiments. Days may pass without any visible change, and then suddenly, multiple crystals may appear within hours as nucleation cascades. In fact, the crystallization process is generally monitored over days to months, with temperature and incubation conditions carefully controlled to favor nucleation without inducing precipitation.

Following nucleation is the phase of crystal growth. Molecules from the surrounding supersaturated solution continue to integrate into the nascent lattice. Unlike small molecules, proteins are large, conformationally flexible, and chemically heterogeneous, which means their incorporation into a crystal requires highly specific alignment. Intermolecular hydrogen bonds, hydrophobic contacts, and electrostatic complementarity govern this alignment. When growth conditions are optimal, proteins organize layer by layer into a periodic and symmetric lattice, producing a crystal of sufficient size and order to diffract X-rays. In suboptimal conditions, however, molecules may aggregate into amorphous precipitates or form microcrystals too disordered for structural analysis.

Interestingly, the principles of protein crystallization reflect the universal physics of crystal formation across disciplines, from molecular biology to mineralogy. In nature, crystals form when energy conditions favor order over disorder, often over vastly different timescales and compositions. Quartz (SiO_2) precipitates from supersaturated silicate-rich fluids under slow cooling and pressure variation; ice crystals (snowflakes) nucleate on microscopic aerosols in cold, humid air; and gemstones like rubies and emeralds emerge deep underground from molten rock or hydrothermal fluids, growing over thousands to millions of years.

The key differences between biological and geological crystallization lie in scale and molecular complexity. Mineral crystals can persist in harsh environments and grow steadily around imperfections (heterogeneous nucleation sites like dust or grain boundaries). In contrast, protein crystals are fragile, hydrated, and highly sensitive to environmental conditions. Protein crystallographers must actively avoid impurities, which could interfere with lattice symmetry or introduce disorder. Yet both systems, mineral and biological, are governed by the same underlying principle: when energy favors symmetry, and time and conditions allow, order emerges spontaneously from chaos.

Crystal Diffraction

Once suitable crystals are obtained, they are subjected to X-ray diffraction analysis, a technique that probes the electron density within the crystal to infer the atomic structure of the molecule. This step requires an intense, highly collimated, and coherent source of X-rays, which we obtain from a synchrotron radiation facility.

At its core, a synchrotron accelerates electrons to relativistic speeds through a complex architecture of linear and circular accelerators. Electrons are first generated via thermionic emission from a cathode and pre-accelerated by a LINAC (linear accelerator). They are then injected into a booster ring, which increases their energy before transferring them into the storage ring, a larger circular structure maintained under high vacuum to preserve electron trajectories. The structure of the synchrotron is schematized in Fig. 18.

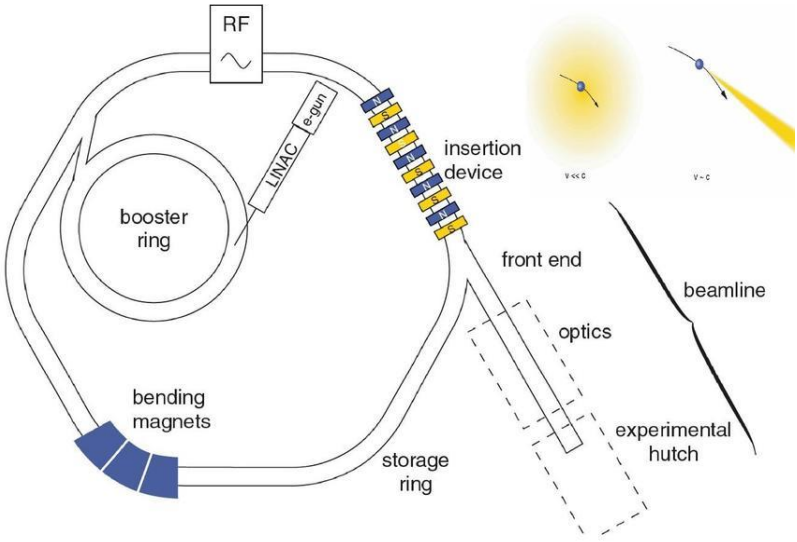


Figure 18. *Schematic drawing representing the essential components of a synchrotron facility.* Figure reproduced from Willmott, 2011. [258]

As the electrons travel through the storage ring, specialized magnetic elements force them to curve along the ring's path. This bending forces them to accelerate laterally, and as a consequence of classical electrodynamics, accelerated charges emit electromagnetic radiation. The resulting emission, known as synchrotron radiation, is exceptionally bright, collimated, and tunable across a wide spectral range, including infrared, ultraviolet, and most importantly, hard X-rays (with high energy and short wavelengths) used for crystallographic analysis. The exact properties of this radiation depend on the type of magnetic device employed at each beamline, which modifies the electron trajectory and thereby tunes the output intensity and spectral distribution. There are several major types of synchrotron radiation sources, illustrated in Fig. 19, each with distinct physical characteristics.

Bending magnets are the simplest form of insertion device, consisting of a single dipole magnet that gently bends the electron path. The radiation emitted is broad in spectrum and moderate in intensity, scaling linearly with the number of circulating electrons: $I \propto N$electrons. The spectrum is continuous, with a smooth distribution over photon energies ($\hbar\omega$), suitable for general applications but relatively low in brightness and spectral sharpness.

Wigglers consist of multiple alternating dipole magnets that force the electron beam into a large-amplitude sinusoidal path. Each deflection generates a pulse of radiation, and because these emissions do not coherently interfere (each pulse is slightly out of phase), the total intensity adds up across the poles: I ∝ N electrons × N poles. The resulting spectrum is still broad but significantly more intense than that from a bending magnet, making wigglers well-suited for experiments requiring high flux over a wide energy range.

Undulators are composed of many closely spaced, smaller dipoles that induce small-amplitude oscillations in the electron path. In this regime, the emitted radiation from each magnetic period interferes constructively (since they are in phase), producing a highly collimated and quasi-monochromatic beam with sharply defined spectral peaks. The brilliance is dramatically enhanced due to the coherence of the emitted waves, and the intensity scales with the square of the number of poles: I ∝ N electrons × (N poles)^2. As shown in the figure, the spectral distribution is no longer broad but exhibits distinct harmonic peaks, which can be

tuned by varying the magnet spacing or electron energy. This makes undulators ideal for macromolecular crystallography (like for protein crystals), where high intensity at specific wavelengths is needed.

Free Electron Lasers (FELs) go a step further, using high-gain undulator setups with extremely coherent, phase-aligned electron bunches to amplify emission through stimulated emission. This produces laser-like X-ray pulses with extraordinary brilliance and coherence. The intensity scaling is even steeper: I ∝ (N electrons)^2 × (N poles)^2. FELs generate near-instantaneous bursts of radiation with peak brightness orders of magnitude beyond conventional synchrotron sources. However, they are typically used for ultrafast time-resolved studies rather than standard protein crystallography.

In the end, each beamline leads to an experimental station, where X-rays are focused onto the protein crystal. As the X-rays strike the ordered lattice of atoms in the crystal, they are diffracted in a pattern specific to the spatial arrangement of electrons in the structure. By collecting these diffraction patterns at multiple angles (through crystal rotation), and applying the principles of Bragg's Law (that explains how and when X-rays are diffracted by the regular, repeating planes of atoms in a crystal) and Fourier transforms, a three-dimensional electron density map is reconstructed. From this map, the atomic model of the protein complex can be built, validated, and refined.

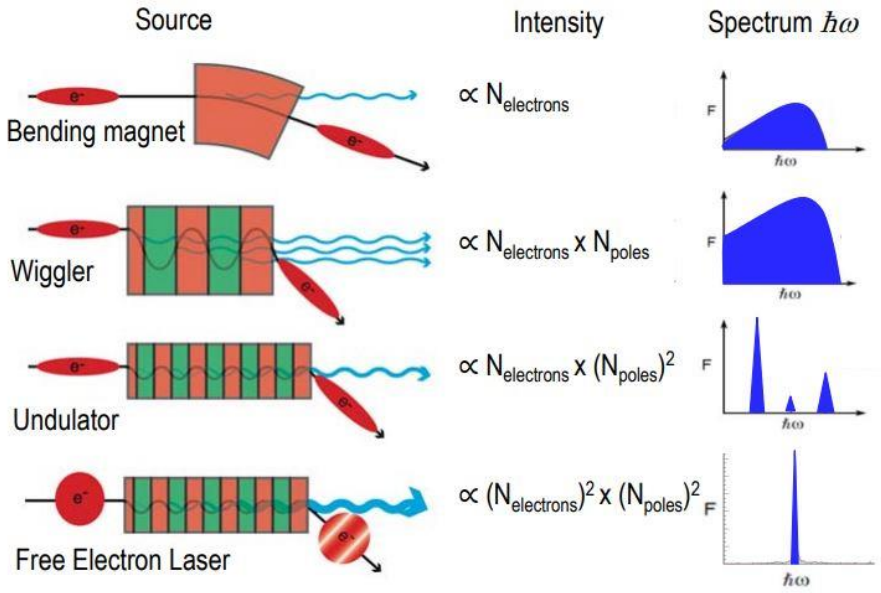


Figure 19. *Comparison of synchrotron radiation sources: Bending Magnets, Wigglers, Undulators and Free Electron Lasers.* Image reproduced from website. [259]

1.7.6 The Structural Gap: A New Paradox

For all its power, X-ray crystallography revealed a practical bottleneck. The process of producing high-quality crystals is laborious, uncertain, and sometimes impossible, particularly for large complexes or membrane-bound proteins. The advent of high-throughput DNA sequencing in the 2000s dramatically amplified this challenge. A profound asymmetry emerged: while the number of experimentally solved structures in the Protein Data Bank grew at a steady, linear-like pace, reaching 100,000 by 2014, the number of known protein sequences

in databases like UniProt exploded exponentially into the tens of millions. This created a vast and growing "structural gap" (Figs. 20 and 21).

Crucially, the discovery of new protein families (Pfam domains) began to plateau. The implication was clear: nature was not inventing endlessly new folds, but we were acquiring an exponentially deeper evolutionary record for each existing one. This created a vast and growing "structural gap," but also the very resource needed to bridge it. The fundamental question, therefore, returned to its origin: if the complete blueprint for the final 3D structure is encoded within the one-dimensional amino acid sequence, as Anfinsen's dogma holds, could we now, with this wealth of evolutionary data, finally learn to read it directly?

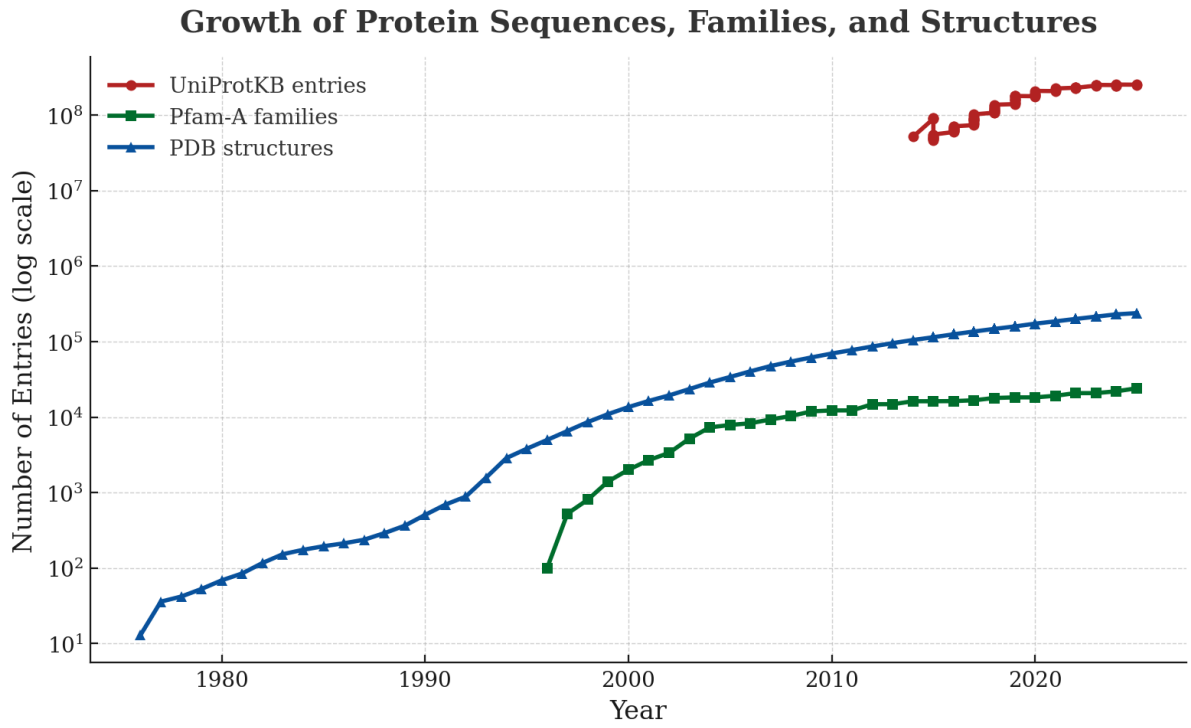


Figure 20. *Diverging Growth Rates of Protein Sequences, Structures, and Families Since 1996.*[2] Since its inception in 1996, the number of protein sequences in UniProtKB has grown exponentially (data not shown before 2014), driven by large-scale sequencing efforts. In contrast, the number of experimentally solved protein structures deposited in the Protein Data Bank (PDB) has increased at a linear sub-exponential rate. The number of curated protein families in Pfam has plateaued over the past decade, suggesting a saturation of the structural and functional diversity observed in nature.

[2] https://ftp.uniprot.org/pub/databases/uniprot/previous_releases/
https://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam37.4/
https://www.rcsb.org/stats/growth/growth-released-structures

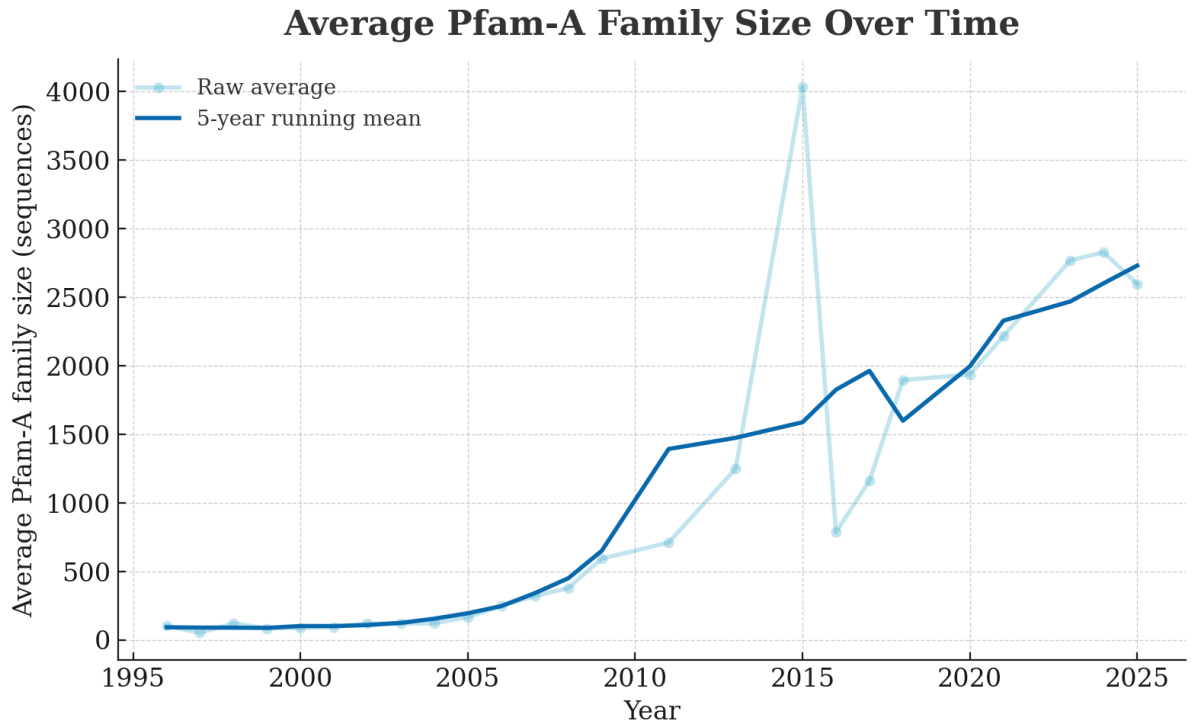


Figure 21. *Average Pfam-A Family Size Over Time.*[3]
The average number of sequences per Pfam-A family has increased steadily over the past three decades. This trend reflects the exponential growth in protein sequence databases such as UniProtKB, while the number of annotated protein families in Pfam is plateauing. As a result, each family now contains significantly more sequences, improving statistical power for multiple sequence alignments and enabling more accurate inference of protein function, structure, and evolutionary constraints. The data highlight a shift from expanding family definitions to deepening coverage within existing families.

1.7.7 The Data-Driven Shift: An Evolutionary Rosetta Stone

The path to a solution came not from simulating physics from first principles, but from leveraging the outcome of evolution's own grand experiment. The explosion of genome sequencing in the late 20th and early 21st centuries provided a new kind of resource: a massive library of protein sequences from millions of species. By collecting and aligning homologous sequences (versions of the same protein from different organisms) into a Multiple Sequence Alignment (MSA), researchers could create a rich evolutionary record. An MSA is more than a list of variations; it is a matrix of constraints, a historical account of which mutations were permitted and which were forbidden over millions of years of selective pressure to maintain a protein's fold and function.

[3] https://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam37.4/

1.7.8 The Co-evolutionary Clue

The critical insight, developed over decades, was that this evolutionary record contained a hidden cipher for 3D structure: co-evolution. The principle is simple yet profound. If two amino acid residues are distant in the linear sequence but make direct physical contact in the folded protein, they cannot evolve independently. A mutation in one residue that disrupts this contact (for example, by changing its size or charge) would destabilize the protein. To restore stability and function, a compensatory mutation must often occur at the second residue. Over evolutionary time, this leaves a distinct statistical fingerprint: a pattern of correlated mutations between pairs of residues that are physically interacting.

This principle was first demonstrated computationally by Göbel, Sander, and Valencia (1994), who showed that pairs of residues exhibiting high mutational correlation often lie close together in the 3D structure. [260] However, these early methods struggled to distinguish true, direct couplings from the much more numerous indirect correlations, where two residues (A and C) appear to co-vary only because both are coupled to a third, intervening residue (B). This challenge was overcome by a new generation of methods, including Direct Coupling Analysis (DCA) and Protein Sparse Inverse Covariance (PSICOV). These approaches moved beyond simple pairwise statistics, instead constructing a global statistical model of the entire alignment. By doing so, they could disentangle the network of dependencies and isolate the true, direct couplings from the transitive noise of indirect ones. [261] These refined co-evolutionary signals proved powerful enough to predict residue contacts within single proteins and, as demonstrated by Hopf et al. (2014), even between interacting proteins in a complex. [262] The statistical echo of evolutionary pressure rapidly became a reliable proxy for physical proximity.

1.7.9 AlphaFold: Operationalizing the Folding Funnel

The next leap came from treating the outputs of co-evolutionary analysis not as a final answer, but as a rich input for deep learning. As Arne Elofsson noted, a predicted contact map is, in essence, an image. [263] The field began applying the same convolutional neural networks that had revolutionized image recognition to learn the non-random, physically constrained patterns of protein contact maps. At CASP13 (Critical Assessment of Protein Structure Prediction, 13th round, 2018), DeepMind's first AlphaFold system demonstrated the power of this approach, shifting the prediction target from binary contacts to a more informative, probabilistic distribution of distances, a "distogram" (distance histogram). [264]

This set the stage for AlphaFold 2, a complete redesign that achieved unprecedented accuracy at CASP14 (2021). [265] Its architecture operationalizes the folding funnel through several key innovations. The core is the "Evoformer," a novel attention-based module that processes the full MSA directly to learn a geometric hypothesis. Such model implicitly learns the rules of protein structure by processing vast amounts of sequence and structural data. AlphaFold 2 apprehends the grammar of helices, sheets, and turns, as well as the syntactic laws of hydrophobic burial, hydrogen bonding and steric packing. Evolution itself has optimized the sequences for the funnel. AlphaFold 2 inherits the same funnel but “integrates over” it instead of simulating the full walk.

The work of Roney and Ovchinnikov provides a powerful physical interpretation of this process. [266] They propose that AlphaFold has learned and uses an *effective energy potential* for protein structures. The role of the MSA, then, is to solve the daunting global search problem posed by Levinthal's paradox. The co-evolutionary data act as a guide, directing the model to the correct basin on the folding energy landscape (Fig. 22). Once there, the system's "structure module", which ingeniously treats residues as a "gas of triangles" to avoid premature chain constraints, refines the hypothesis into an atomically precise structure. In essence, AlphaFold deciphers the evolutionary solution to the folding problem.

This revolution was extended to protein complexes with AlphaFold-Multimer in 2021. [267] This required solving the difficult *pairing problem*: correctly matching orthologous protein pairs from two separate MSAs while avoiding noise from incorrect paralog pairings. [262] By training the model on such paired data, AlphaFold-Multimer learns to use inter-protein co-evolutionary signals to predict complex interfaces with high accuracy.

The subsequent development of ColabFold democratized this power, replacing the time-consuming MSA generation step with the much faster MMseqs2. [268] The insight behind this tool's remarkable speed is a rapid pre-filtering of the vast sequence databases: instead of comparing entire proteins, it first identifies promising candidates based on shared, short amino acid 'words' (k-mers). This strategy dramatically narrows the search space before attempting more costly full alignments, making high-accuracy prediction accessible to the broader scientific community.

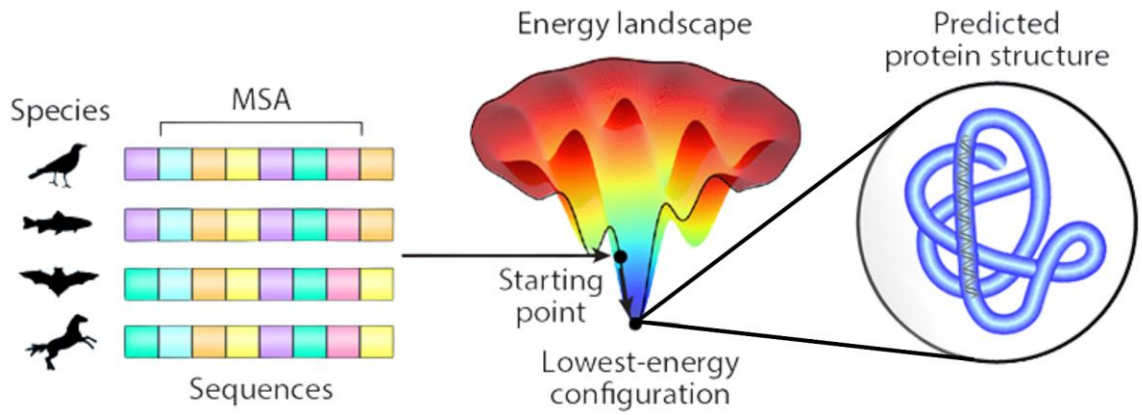


Figure 22. *AlphaFold's Inferred Physics: Using Evolution to Navigate the Energy Landscape.*
The figure illustrates the hypothesized mechanism by which AlphaFold predicts protein structure, as proposed by Roney and Ovchinnikov. This model posits a dual role for the network's learned intelligence. The process begins with a Multiple Sequence Alignment (MSA), a rich record of evolutionary constraints gathered from homologous sequences across different species. Importantly, this evolutionary data is not used to simulate the entire folding process from an unfolded state. Instead, it solves the vast global search problem posed by Levinthal's Paradox by providing a high-quality "Starting point", an initial structural hypothesis that is already located within the correct basin of the protein's folding energy landscape. From this vantage point, the network descends the landscape, not by simulating *ab initio* physics, but by following the gradients of its own implicitly learned 'effective

energy potential'. This potential approximates the true biophysical principles of folding, guiding the structure towards the stable, "Lowest-energy configuration". The final result of this guided local refinement is a single, highly accurate predicted protein structure. This dual strategy, using evolution for global navigation and learned physics for local refinement, is the key to AlphaFold's ability to identify the native fold from a universe of possibilities.
Image adapted from APS/Carin Cain. Source: American Physical Society.

1.7.10 A Framework for Confidence: Interpreting the AlphaFold Output [269]

A key feature that distinguishes AlphaFold from earlier methods is its ability to provide robust, per-residue estimates of its own accuracy. The model does not simply produce a structure; it annotates it with a multi-layered confidence framework, allowing researchers to critically assess which parts of a prediction are likely to be reliable. This framework operates at different scales, from the local environment of a single amino acid to the global arrangement of a multi-protein complex.

Local Structure Confidence: The pLDDT Score
At the most granular level, AlphaFold provides the predicted Local Distance Difference Test (pLDDT), a per-residue measure of local confidence from 0 to 100. This metric estimates the confidence in the local structure surrounding each amino acid. It is generated by the network performing a final self-consistency check, assessing whether the local environment of each residue (its bond angles, distances, and interactions) conforms to the highly refined patterns it has learned from thousands of experimental structures.

A pLDDT above 90 would be taken as the highest accuracy category, in which both the backbone and side chains are typically predicted with high accuracy. In contrast, a pLDDT above 70 usually corresponds to a correct backbone prediction with misplacement of some side chains.

The pLDDT score can vary significantly along a protein chain. This means AlphaFold2 can be very confident in the structure of some parts of the protein, but less confident in other regions. This gives users an indication of which parts of the predicted structure may be reliable and which are unlikely to be.

AlphaFold2 assigns low confidence scores (pLDDT < 50) to certain protein regions for two main reasons. First, the region may be intrinsically disordered or highly flexible, lacking a stable, well-defined structure. Second, the region may adopt a defined structure, but AlphaFold2 lacks sufficient information to predict it confidently. Typically, AlphaFold2 demonstrates high confidence in predicting globular domains due to their evolutionary conservation and structural stability. Conversely, it assigns low confidence to linkers between domains. Linker regions are usually less conserved, more variable, flexible, and inherently unstructured, limiting AlphaFold2's ability to predict a specific conformation. Most intrinsically disordered regions (IDRs) remain disordered under physiological conditions. However, in some cases, IDRs undergo binding-induced folding upon interaction with their native macromolecular partners. In these instances, AlphaFold2 may predict a structured, folded state with high pLDDT scores despite the region being disordered in isolation.

A high pLDDT score for all domains of a protein does not necessarily mean that AlphaFold2 is confident in the relative positions or orientations of those domains. pLDDT does not measure confidence at such large scales, so a different metric is required.

Relative Domain Confidence: The PAE Matrix

While pLDDT assesses local accuracy, it provides no information about the confidence in the relative orientation of different domains. This is addressed by the Predicted Aligned Error (PAE). The PAE is presented as a 2D plot where the value at position (x, y) estimates the expected positional error (in Ångströms) of residue X if the predicted and true structures were aligned on residue Y.

This metric essentially quantifies the model's confidence in the global framework of the protein by predicting the "wobble" between domains. Low PAE values between two domains indicate high confidence in their relative positions and packing, suggesting a rigid connection. Conversely, high PAE values between well-folded domains (which may themselves have high pLDDT scores) signify uncertainty in their arrangement, suggesting they may be connected by a flexible linker.

Global and Interface Confidence: pTM and ipTM

For predicting protein complexes, the AlphaFold-Multimer system introduces two additional global metrics derived from the PAE matrix. The predicted Template-Modeling (pTM) score provides a holistic judgment on the accuracy of the overall fold. A score above 0.5 suggests the global topology of the complex may be correct. pTM can be dominated by a single protein, if it is larger in the complex.

More critical for interaction studies is the interface predicted Template-Modeling (ipTM) score, which refines this concept by focusing the calculation exclusively on residue pairs that cross the protein-protein interface. It directly answers the question: "How confident is the model in the predicted binding mode?" An ipTM score greater than 0.8 typically indicates a high-confidence, high-quality prediction of the interaction. Such value assumes modelling with multiple recycling steps, so the process of prediction reaches a degree of convergence. In large-scale screenings for protein-protein interactions, often settings optimised for the speed of prediction are used, with very few or no recycling steps. In such cases ipTM thresholds as low as 0.3 have been used for initial screening; importantly though, all pairs of proteins with ipTM scores higher than 0.3 should be subjected to additional examination. ipTM may be more useful to users than pTM. This is because the quality of the prediction of the relative positions of the subunits and the quality of the whole complex prediction are highly interdependent: if the relative positions of the subunits are correct (as reflected in a high ipTM score), users can expect that the whole complex is also correct.

In practice, pTM and ipTM are often combined into a single ranking metric (multi-score = 0.8 × ipTM + 0.2 × pTM) to balance interface and overall fold quality. [267]

A known limitation of ipTM is that its value can be diluted by the presence of flexible regions flanking a well-defined interface. To address this, a refined metric, actifpTM (actual interface pTM), was developed to focus the confidence calculation only on the core, well-ordered interface residues, providing a more robust measure for interactions involving flexible components. [270]

In practice, overall confidence in predictions for multimers should be based on a combination of all the metrics, including both pTM and ipTM as well as pLDDT and PAE.

1.7.11 Impact and Synergy

AlphaFold has fundamentally transformed structural biology, but it does not replace experimentation. It is a *model*. It is a supremely powerful hypothesis-generation engine. The accuracy of its predictions, particularly for proteins with deep MSAs, is often comparable to low-resolution experimental data, but it can struggle with novel folds, orphan proteins, or dynamic systems. Its true power lies in synergy. Computational models can guide the design of crystallization or cryo-EM experiments, help interpret ambiguous density maps, and provide structural context for biochemical data. [271] Conversely, experimental structures provide the essential ground-truth data needed to train and validate these computational models. Our thesis operates at this modern intersection, employing both experimental characterization and computational prediction. We use biophysical methods to probe the CHI3L1-RAGE interaction and, in parallel, leverage the evolutionary logic embedded in AlphaFold to build a structural model, with each approach informing and strengthening the other.

1.7.12 The Computational Revolution: Folding by Co-evolution - Known Frontiers and Limitations

Despite its revolutionary impact, AlphaFold is not a panacea. Its predictive power is rooted in the patterns of static structures deposited in the PDB, and as Arne Elofsson noted in 2022, this foundation imposes several key limitations. By default, the model predicts a single, static conformation, and it does not reliably capture the dynamic ensembles or allosteric shifts that are fundamental to protein function. Similarly, it is insensitive to the effects of single-point mutations on stability ($\Delta\Delta G$). The field is advancing, however, and emerging techniques now seek to address these gaps. Methods that systematically subsample the input MSA have shown it is possible to generate alternative protein conformations. [272]

Another intrinsic, relevant limit of AlphaFold is its dependence on MSA depth; for orphan proteins with few or no homologs, the accuracy diminishes significantly. Furthermore, while the model excels at predicting globular domains, it struggles with orienting them in multi-domain proteins without a strong inter-domain signal, and it often misrepresents unstructured or disordered regions.

Overall, AlphaFold provides an unprecedented model of a protein's most probable static structure, but it is not inherently a model of its dynamics or thermodynamics.

Interestingly, as we will demonstrate in the Results section, the high-confidence model of the CHI3L1-RAGE complex generated in this work appears to capture a significant, binding-mediated conformational shift in a key interface residue. This finding suggests that while the general limitations hold, the co-evolutionary information specific to certain ligand-receptor systems may, in some cases, allow the model to implicitly learn and represent functionally critical structural rearrangements, even if minor.

1.8 From Structural Insight to Therapeutic Rationale

The work presented in this thesis aims to move beyond the phenomenological observation of the CHI3L1-RAGE interaction to provide the first detailed, molecular-level model of this critical immune checkpoint. By elucidating the specific atomic contacts and the geometry of the binding interface, this study provides the essential blueprint for the rational design of targeted therapeutics - such as small-molecule inhibitors or engineered protein decoys - capable of selectively disrupting this axis of immune suppression. More broadly, it serves as a template for an integrated strategy, combining biochemical characterization with advanced computational modeling, to dissect other complex molecular interactions that drive disease.

The ambition of this work is therefore twofold: to deliver the first molecular understanding of a fundamental innate immunity checkpoint, and in doing so, to demonstrate how the convergence of computational modeling and biochemical analysis can transform a complex biological problem into a tractable therapeutic target.

2. Materials & Methods

2.1 Materials

2.1.1 Protein Expression and Purification: hCHI3L1-His and hRAGE(VC1C2)-His

For this study, we outsourced the expression and purification of Histidine-tagged mammalian proteins from the company BioIntron (Shanghai). The expressed proteins were human CHI3L1 [UniProt accession ID: P36222, aa 22-383] and human RAGE extracellular domain [UniProt accession ID: Q15109, aa 23-342], that we named hecRAGE. Both were expressed in proprietary HEK293 cells by BioIntron, resulting in amounts of >20 mg per protein. A mammalian expression system was chosen in order to retain the proteins in their native glycosylation state, which we inferred from the literature might have a role, especially in RAGE binding to its ligands.

The protein sequences were obtained from Uniprot to design the plasmids. At the N-terminus of either protein, a human interleukin-10 secretion signal peptide was added (MHSSALLCCLVLLTGVRA). [273] This signal peptide was selected for its characteristic mammalian secretion features, including a strong hydrophobic core and a putative cleavage site conducive to processing by the endoplasmic reticulum (ER) signal peptidase in Human Embryonic Kidney 293 (HEK293) cells. This sequence ensures efficient co-translational translocation into the ER and subsequent secretion into the culture medium, enhancing the extracellular yield of the recombinant protein.

A 9x His-tag was selected for purification to avoid adding bulking portions to the proteins which may impair crystal formation. The tag was placed at the C-terminus because, from the literature, the N-terminal domains of both proteins were the most active in their biological interactions with ligands.

The two constructs details are shown in Table 2 and Fig. 23.

Protein	**UniProt Accession ID**	**Amino Acids**	**Sequences**	**Affinity Tag**
hCHI3L1	P36222	22-383	MHSSALLCCLVLLTGVRAYKLVC YYTSWSQYREGDGSCFPDALDRF LCTHIIYSFANISNDHIDTWEWND VTLYGMLNTLKNRNPNLKTLLSV GGWNFGSQRFSKIASNTQSRRTFI KSVPPFLRTHGFDGLDLAWLYPGR RDKQHFTTLIKEMKAEFIKEAQPG KKQLLLSAALSAGKVTIDSSYDIA KISQHLDFISIMTYDFHGAWRGTT GHHSPLFRGQEDASPDRFSNTDYA VGYMLRLGAPASKLVMGIPTFGRS FTLASSETGVGAPISGPGIPGRFTK	Histidine (x9 His)

			EAGTLAYYEICDFLRGATVHRILG QQVPYATKGNQWVGYDDQESVK SKVQYLKDRQLAGAMVWALDLD DFQGSFCGQDLRFPLTNAIKDALA ATHHHHHHHHH	
hecRAGE	Q15109	23-342	MHSSALLCCLVLLTGVRAAQNITA RIGEPLVLKCKGAPKKPPQRLEWK LNTGRTEAWKVLSPQGGGPWDSV ARVLPNGSLFLPAVGIQDEGIFRCQ AMNRNGKETKSNYRVRVYQIPGK PEIVDSASELTAGVPNKVGTCVSE GSYPAGTLSWHLDGKPLVPNEKG VSVKEQTRRHPETGLFTLQSELMV TPARGGDPRPTFSCSFSPGLPRHRA LRTAPIQPRVWEPVPLEEVQLVVE PEGGAVAPGGTVTLTCEVPAQPSP QIHWMKDGVPLPLPPSPVLILPEIG PQDQGTYSCVATHSSHGPQESRAV SISIIEPGEEGPTAGSVGGSGLGTLA HHHHHHHHH	Histidine (x9 His)

Table 2. *hCHI3L1-His and hecRAGE-His constructs.*
For each protein construct, the UniProt accession ID, the amino acid positions and sequence, as well as the affinity tag are specified. In the sequences, the starting Methionine residue is shown in red, the signal peptide is in dark cyan and the Histidine tag in dark orange.

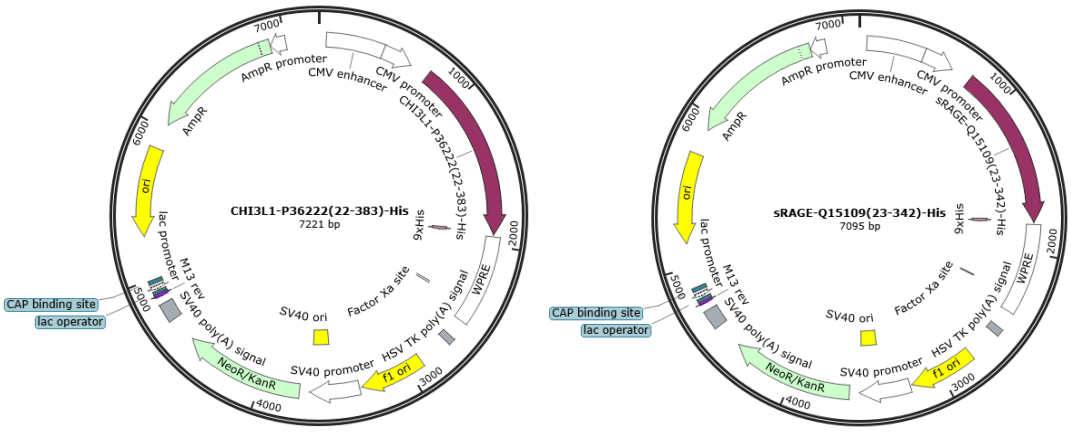


Figure 23. *Plasmid maps for hCHI3L1-His and hecRAGE-His expression in HEK293 cells*.
The figure shows the mammalian expression plasmids used for the production of recombinant human CHI3L1-His (left) and extracellular RAGE-His (right), both expressed in HEK293 cells. The coding sequences correspond to hCHI3L1 (P36222), amino acids 22–383, and hecRAGE (Q15109), amino acids 23–342, corresponding to the VC1C2 extracellular domains of RAGE. Both constructs feature an N-terminal signal peptide (from human IL-10) to direct secretion and a C-terminal 9x His-tag to facilitate purification via immobilized metal affinity chromatography (IMAC). Expression is driven by a human cytomegalovirus (CMV) immediate-early promoter/enhancer, ensuring high-level

transcription in mammalian cells. Additional regulatory elements include a woodchuck hepatitis virus post-transcriptional regulatory element (WPRE) to enhance mRNA stability and translation efficiency, and a polyadenylation signal (SV40 poly(A)) for proper transcript termination. The plasmids also contain a SV40 origin of replication for episomal replication in mammalian cells that express SV40 large T antigen, and a neomycin/kanamycin resistance gene (NeoR/KanR) for dual selection in both bacteria and eukaryotic cells. The bacterial backbone includes an ampicillin resistance gene (AmpR) and pUC origin of replication (ori) for propagation and selection in E. coli.

Quality Control of the Expressed Proteins

Quality control conducted by Biointron on hCHI3L1-His showed a purity of >95% at sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) under reducing conditions and of 96.241% by size exclusion chromatography - high performance liquid chromatography (SEC-HPLC).

hecRAGE-His had a purity >95% by SDS-PAGE under reducing conditions and of 90.050% by SEC-HPLC.

Both proteins displayed an endotoxin level at the Limulus Amebocyte lysate test of <1 EU/mg.

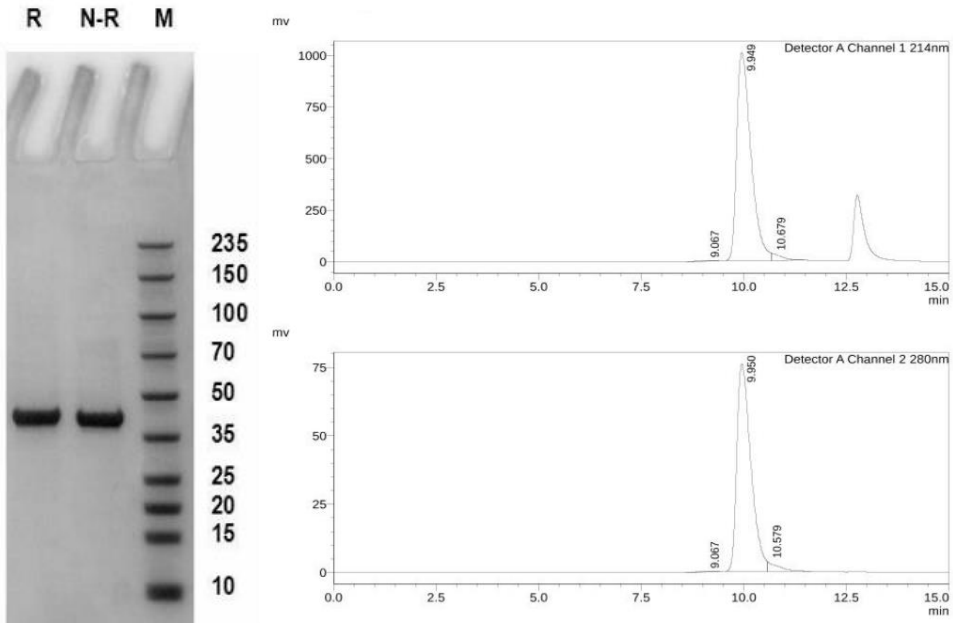


Figure 24. *Quality control results for hCHI3L1-His*.

The SDS-PAGE gel (left) under reducing (R) and non-reducing (N-R) conditions, alongside a molecular weight marker (M), demonstrates a single, well-defined band between 35–50 kDa, consistent with the expected molecular weight of hCHI3L1-His (41.7 kDa), not accounting for the contribution of glycosylations). [274] The similarity between the reducing and non-reducing lanes suggests the protein is monomeric and not forming disulfide-linked dimers or aggregates. The sharpness and lack of additional bands indicate high purity and low degradation or contamination.

The SEC-HPLC chromatograms (right) show absorbance at 214 nm (top) and 280 nm (bottom). The 214 nm signal reflects peptide bond content and total protein load, while 280 nm specifically detects aromatic amino acids, serving as a marker for protein content. A major single peak at approximately 10 minutes, with minimal minor peaks, confirms the monodispersity of the preparation and absence of significant aggregates or degradation products. The 214 nm and 280 nm profiles show close correspondence, which validates the integrity and purity of the expressed protein.

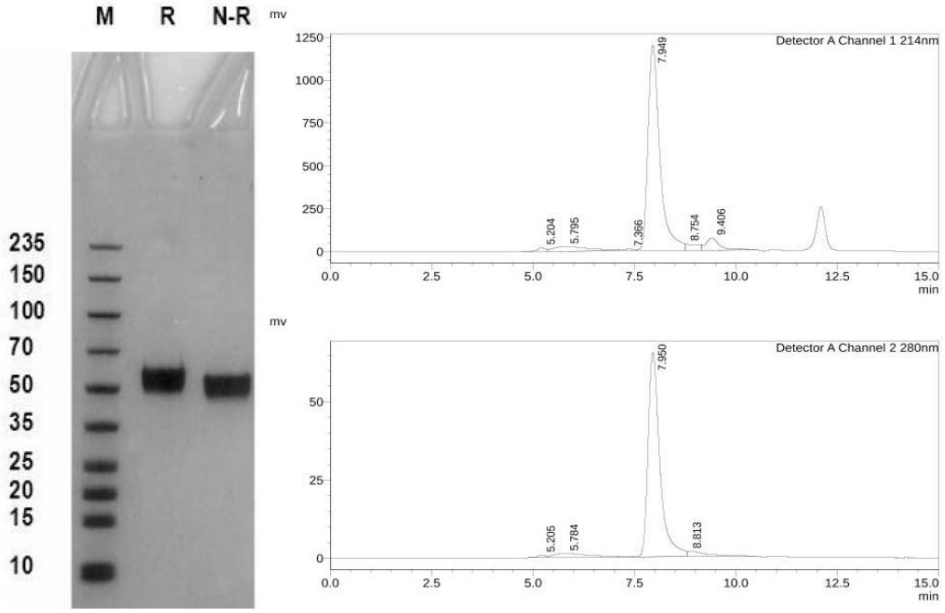


Figure 25. *Quality control results for hecRAGE-His*.

The SDS-PAGE gel (left) shows a prominent band between 50–70 kDa. Although the theoretical molecular weight of the unglycosylated hecRAGE-His protein used in this study is 35.2 kDa, the observed increase in apparent molecular weight is to be attributed to two reasons. [274] The first is that hRAGE has two glycosylation sites (Asn25 and Asn81), and previous studies in HEK293 cells-expressed hRAGE report full occupancy at Asn25 and partial or variable occupancy at Asn81, resulting in multiple glycoforms and corresponding band heterogeneity. [130] These modifications can contribute approximately 2, 7, or 9 kDa in additional mass to the fully unglycosylated species. The second reason is that the Ig-like β-sheet–rich structure of RAGE contributes to reduced SDS binding and incomplete denaturation, further increasing its apparent molecular weight by ~9 kDa. [276] The similar migration under both reducing (R) and non-reducing (N-R) conditions indicates that the protein is primarily monomeric and does not form disulfide-linked multimers.

The chromatograms from the SEC-HPLC (right) recorded at 214 nm (top) and 280 nm (bottom) show a dominant peak at ~7.9 minutes, corresponding to a monomeric protein population. The tight, symmetrical peak shape and minimal secondary peaks indicate a high degree of homogeneity and absence of aggregates. The 280 nm signal confirms the proteinaceous nature of the main peak, while the 214 nm signal reflects overall purity. Together, these data confirm that the hecRAGE-His preparation is highly pure, structurally intact, and exhibits expected glycosylation heterogeneity typical of mammalian-expressed RAGE.

hRAGE-Fc

For the enzyme-linked immunosorbent assays (ELISAs), a recombinant human RAGE protein fused to the Fc region of human IgG1 (hRAGE-Fc) was purchased from BioLegend. This chimera includes the extracellular portion of RAGE (amino acids Gln24-Ala344, based on UniProt accession ID BC020669) fused at the C-terminus to an Fc tag and a 6x His-tag. The protein was expressed in Chinese Hamster Ovary cells and formulated in phosphate-buffered saline (PBS), pH 7.2, without carrier protein. The 566 amino acid recombinant protein has a predicted molecular mass of approximately 61.8 kD and the DTT-reduced protein migrates at approximately 75 kD. The hRAGE-Fc construct is able to form a homodimer and retains biological activity, as demonstrated by its ability to bind immobilized S100A9 (EC_{50} = 0.8–3.2

μg/mL). The final product had a purity >90% as assessed by SDS-PAGE and an endotoxin level of <0.1 EU/μg, making it suitable for downstream binding assays and detection formats involving Fc receptor-mediated capture.

Protein Storage
All proteins were stored at -80 °C, aliquoted and stored at -20 °C for immediate use. All operations were conducted on ice to minimize protein denaturation, and repeated freeze/thaw cycles were avoided. The proteins were in phosphate buffer saline (PBS), pH 7.2 - 7.4. Stock solutions of hCHI3L1-His had a concentration of 1.03 mg/mL, stocks of hecRAGE-His of 1.42 mg/mL.

2.1.2 Peptide-N-Glycosidase F Enzyme

Peptide-N-Glycosidase F (PNGase F) from *Elizabethkingia meningoseptica* was used for enzymatic deglycosylation of N-linked glycans from glycoproteins. The enzyme (Sigma-Aldrich, P7367) was obtained in lyophilized form (≥95% purity by SDS-PAGE) and reconstituted at a working concentration of 500 U/mL in high-purity water from a starting quantity of 50 units, yielding a final buffer composition of approximately 5 mM potassium phosphate (pH 7.5). One Sigma unit of PNGase F activity is equal to 1 IUB milliunit. PNGase F hydrolyzes the amide bond between the innermost N-acetylglucosamine (GlcNAc) and the asparagine residue of glycoproteins (as shown in Fig. 26), thereby releasing a broad spectrum of N-linked glycans, including high-mannose, hybrid, and complex types. However, N-glycans containing an α(1→3)-linked fucose on the core GlcNAc are resistant to cleavage. The enzyme has a molecular weight of ~36 kDa and functions optimally at pH 8.6 and 37 °C.
The proteomics-grade enzyme is validated for use in glycoprotein deglycosylation workflows, including in-solution digestion, in-gel applications, and membrane-bound blots, and is compatible with downstream analysis such as SDS-PAGE or mass spectrometry (MALDI-TOF MS). Each lot is rigorously tested to confirm the absence of contaminating exoglycosidases (e.g., β-galactosidase, α/β-mannosidase) and proteases, ensuring specificity for N-linked oligosaccharides. Enzyme activity was defined as the amount required to release N-linked glycans from 1 nmol of denatured ribonuclease B per minute at 37 °C and pH 7.5, with results monitored by SDS-PAGE, and was reported to be of one unit. Reconstituted enzyme solutions were stored at –20 °C. PNGase F was employed in this study to assess the glycosylation state of hCHI3L1-His, hecRAGE-His and hRAGE-Fc by comparing molecular weight shifts before and after treatment via SDS-PAGE.

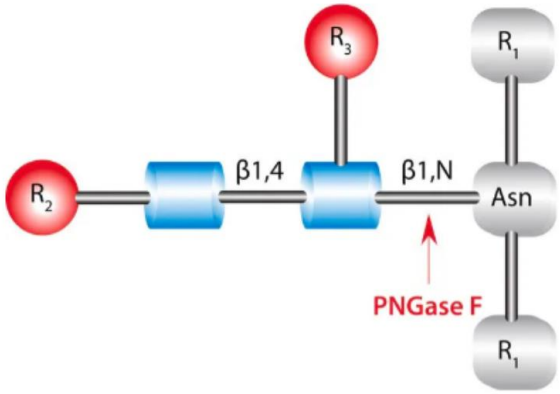


Figure 26. *Cleavage site and structural requirements for PNGase F*. [277]

2.1.3 Antibodies for Enzyme-Linked Immunosorbent Assays

To establish the Enzyme-Linked Immunosorbent Assays (ELISAs), we used different antibodies, depending on the specifics of each assay.
For the ligand-capture Fc-direct ELISAs, where the ligand was immobilized on the plate and its Fc-tagged receptor was allowed to bind, signal detection was achieved using an HRP-conjugated anti-human Fc antibody (donkey anti-human IgG Fc fragment-specific, Jackson ImmunoResearch, AB_2340482).
For the ligand-capture indirect detection ELISAs, in which either the ligand or the non–Fc-tagged receptor was immobilized, we introduced the corresponding binding partner, followed by an unlabeled primary antibody specific to that partner, and finally an Horseradish Peroxidase (HRP)-conjugated secondary antibody for signal generation. Specifically, for the detection of the receptor RAGE, we used a rabbit polyclonal anti-RAGE antibody (Abcam, ab37647), followed by an HRP-conjugated anti-rabbit IgG (Invitrogen, 31460); for detection of the ligand CHI3L1, we used a goat polyclonal anti-human CHI3L1 antibody (R&D Systems, AF2599), followed by an HRP-conjugated anti-goat IgG (Invitrogen, A15999).

2.1.4 Chito-oligosaccharides and Caffeine

The chito-oligosaccharides (COS) and caffeine used in the competitive ELISA assays were commercially sourced from two different suppliers.
Tetraacetyl-chitotetraose (O-CHI4), with the molecular formula $C_{32}H_{54}N_4O_{21}$ and a molecular weight of 830.4 Da, and hexaacetyl-chitohexaose (O-CHI6), with the formula $C_{48}H_{80}N_6O_{31}$ and a molecular weight of 1237.2 Da, were purchased from Neogen (Megazyme). The reported purities were >94% for O-CHI4 and >85% for O-CHI6.
Caffeine, with the molecular formula $C_8H_{10}N_4O_2$ and a molecular weight of 194.19 Da, was obtained from Fisher Scientific, with a reported purity of 99.7%.

2.1.5 Fluorescent 6FAM-Conjugated Peptide

For fluorescence polarization binding studies involving hCHI3L1-His, a custom synthetic peptide comprising the FG loop (Pro212-Ala219) and GG' loop (Thr222-Pro227) portions of the human RAGE C1-domain was designed based on AlphaFold2 predictions (ColabFold) of the protein complex interface. [140] The final selected sequence, S**PGLP**<u>**RHRA**</u>LR<u>**T**</u>**APIQP**R, included flanking regions to ensure structural flexibility and correct presentation of the loop. The RHRALRT portion is hypothesized to be the central mediator of the interactions, as per the model, via the positively charged group of Arginines and Histidine. Remarkably, the flanking unstructured regions contain a Proline at either extremity, signaling where the sequence in human RAGE transitions to secondary β-sheets.
To minimize potential steric hindrance during binding, the 6-carboxyfluorescein (6-FAM) fluorophore useful for the fluorescence polarization assays was attached at the C-terminus of the peptide through the attachment of a Lys(6-FAM).
The labeled peptide was synthesized by GenScript, yielding 4.5 mg with a purity of 95.2% by HPLC and a confirmed molecular weight of 2509.6 Da (theoretical: 2510.02 Da), verified by mass spectrometry (ESI-MS). Solubility testing demonstrated that the peptide was soluble up to 15 mg/mL in ultrapure water and saline, and up to 10 mg/mL in PBS (pH 7.4). The peptide was dissolved in 500 µL water. The fluorescent peptide was stored in aliquots at –20 °C and used freshly to avoid degradation or photobleaching. The peptide will be tested with fluorescent polarization assays to verify its ability to bind CHI3L1.

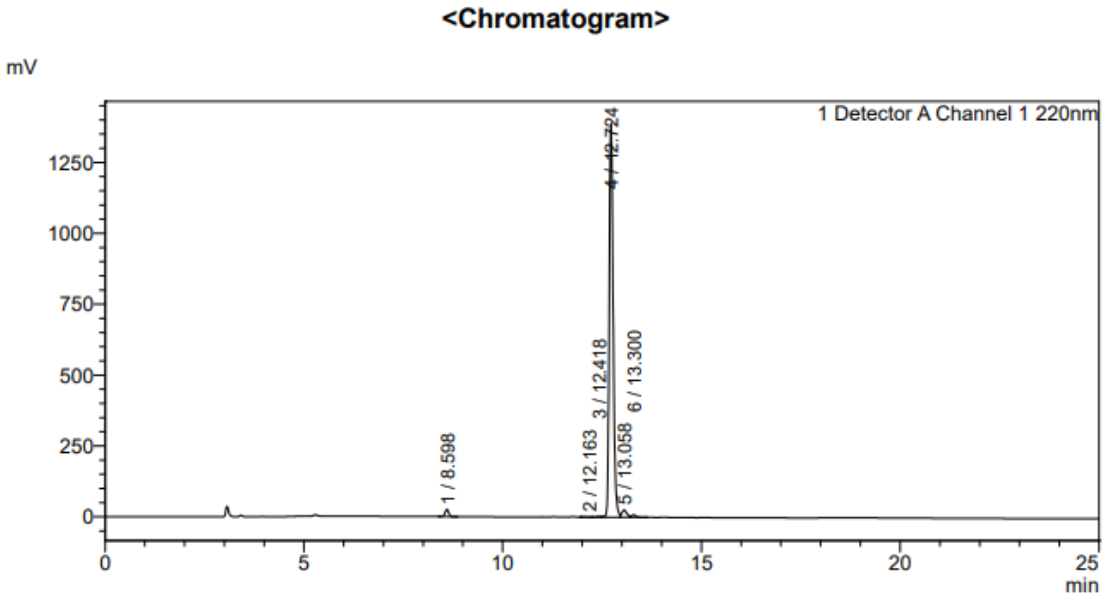


Figure 27. *HPLC chromatogram of the fluorescent FG-GG' loops peptide*.
The chromatogram, acquired at 220 nm, confirms the high purity of the synthesized fluorescent peptide, with a major peak at ~12.7 minutes accounting for >95% of the total area.

2.2 Biochemical Studies

2.2.1 Protein Concentration Determination

Protein concentrations were determined using a NanoDrop One spectrophotometer (Thermo Fisher Scientific) by measuring absorbance at 280 nm. This method relies on the intrinsic absorbance of aromatic amino acids, primarily tryptophan and tyrosine, within the protein sequence. For each recombinant protein, the theoretical extinction coefficient at 280 nm was calculated using the ExPASy ProtParam tool (https://web.expasy.org/protparam/), based on the full amino acid sequence and assuming all cysteine residues were in the oxidized state. Concentration was then calculated using Beer-Lambert's law ($A = \varepsilon \times c \times l$), where ε is the extinction coefficient ($M^{-1}\,cm^{-1}$), c is the concentration (M), and l is the path length (cm). The instrument was blanked with the same buffer as the sample, and all measurements were blank-corrected. This approach allowed rapid quantification of samples using a limited amount of material (2 µL).

For hCHI3L1-His, the extinction coefficient was 68,090 $M^{-1}\,cm^{-1}$ ($A_{0.1\%}$ = 1.632); for hecRAGE-His, 39,335 $M^{-1}\,cm^{-1}$ ($A_{0.1\%}$ = 1.117); and for hRAGE-Fc, 75,120 $M^{-1}\,cm^{-1}$ ($A_{0.1\%}$ = 1.230).

2.2.2 Sodium Dodecyl Sulfate-Polyacrylamide Gel Electrophoresis

Sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) is the most widely used technique for the qualitative analysis of protein mixtures. It involves the use of sodium dodecyl sulfate (SDS), an anionic detergent that binds to proteins, approximately one SDS molecule per two amino acids, disrupting almost all non-covalent interactions. This treatment unfolds the proteins into nearly linear chains and imparts a uniform negative charge proportional to their molecular mass. In addition, the presence of reducing agents such as dithiothreitol (DTT) or β-mercaptoethanol leads to the reduction of disulfide bonds between cysteine residues, ensuring complete denaturation. Protein separation is carried out using a polyacrylamide gel matrix. Polyacrylamide gels are the most commonly used support medium because they are chemically inert and tunable in porosity, which can be precisely controlled by adjusting the ratio of acrylamide to bisacrylamide, the crosslinking agent in the polymerization process. During electrophoresis, SDS-protein complexes migrate through the polyacrylamide gel (which also contains SDS) toward the anode. Separation is achieved based on molecular weight due to the molecular sieving effect of the gel: smaller proteins migrate faster through the pores, while larger proteins migrate more slowly. SDS-PAGE uses a discontinuous gel system, composed of a stacking gel, where wells are formed and samples are loaded, and a running (or separating) gel, which performs the actual molecular separation. The stacking gel serves to concentrate the protein components into a compact zone before they enter the running gel, where separation by size takes place. Samples must be pre-treated in a sample buffer that includes SDS, a reducing agent (e.g., β-mercaptoethanol), a density agent (such as sucrose or

glycerol), and a tracking dye (typically bromophenol blue) to visually monitor the progress of electrophoresis.

Protein samples were quantified using a NanoDrop spectrophotometer and diluted to 2 µg in a final volume of 41.25 µL of dH_2O per sample. They were then mixed with 13.75 µL of loading buffer, originally prepared by combining 100 µL of β-mercaptoethanol with 900 µL of 4× Laemmli buffer (Bio-Rad; 277.8 mM Tris-HCl pH 6.8, 44.4% v/v glycerol, 4.4% lithium dodecyl sulfate, 0.02% bromophenol blue). The samples were denatured by boiling at 95 °C for 5 minutes and centrifuged at 16000 G for 1 minute.

The gel was prepared at 10% or 12.5% of polyacrylamide concentration because the expected position on the molecular weight scale were between 35 and 50 kDa for hCHI3L1-His, 45 and 60 kDa for hecRAGE-His and around 75 kDa for hRAGE-Fc. The recipe of the different running gels used is shown in table 3.

Running gel solution at 10%		**Running gel solution at 12.5%**	
dH_2O	3.19 mL	dH_2O	4 mL
Tris pH 8.8	2.5 mL	Tris pH 8.8	2.5 mL
30% Acrylamide/bis	4.17 mL	30% Acrylamide/bis	3.33 mL
10% SDS	100 µL	10% SDS	100 µL
Ammonium persulfate (APS)	100 µL	Ammonium persulfate (APS)	100 µL
TEMED	10 µL	TEMED	10 µL
Total volume	10 mL	Total volume	10 mL

Table 3. *List of components for the running gels (10% and 12.5% polyacrylamide).*

The stacking gel solution was prepared at 4% of polyacrylamide concentration by mixing 3 mL of dH_2O, 1.25 mL of Tris pH 8.8, 0.67 mL of 30% Acrylamide/bis, 50 µLof 10% SDS, 50 µL of Ammonium persulfate (APS), and 5 µL of TEMED for a total volume of 5 mL.

7.5 mL of running gel solution were polymerized in the chamber and overlaid with a monomer solution of isopropanol to avoid bubbles, and subsequently 4 mL of stacking gel solution were polymerized on top of it. The wells-forming comb was inserted in the chamber whilst the gel was still in liquid form and, after polymerization, it was removed slowly and gently to form the wells.

A 750 mL solution of a fresh running buffer was prepared by mixing 75 mL of Tris-glycine/SDS 10x (250 mM Tris base, 1.92 M glycine, and 1% w/v SDS) with 675 mL of dH_2O.

The gel chamber was mounted in the Mini-PROTEAN Tetra Vertical Electrophoresis Cell by Bio-Rad and each protein sample was loaded separately in a lane of the gel by gentle pipetting. Lastly, 7 µL of the molecular weight marker PageRuler Prestained Protein Ladder (ThermoFisher; 10 to 180 kDa) were loaded in one of the lanes.

The electrophoresis apparatus was filled with the running buffer up to the appropriate line. The apparatus was attached to the power supply and run for 30 minutes at 110 V, up to the moment when the markers reached the end of the stacking gel. The voltage was then increased to 180 V to finish the run in about 45-50 minutes.

After electrophoresis, the gel was removed from the apparatus and the stacking gel and the wells were trimmed away. The remaining running gel was then carefully transferred into a tray with ultrapure water and shook gently on an orbital shaker at room temperature for 5-10 minutes to reduce background. In the following staining step, the gel was immersed in Bio-Rad Coomassie Brilliant Blue R-250 Staining Solution and incubated on the orbital shaker for 1 hour or until protein bands are clearly visible. To achieve destaining, the staining solution was decanted and the Bio-Rad Destaining Solution was added. The gel was again incubated on the shaker and the destaining solution was replaced every 20 minutes until the gel background was sufficiently clear (at least 1 hour). Finally, the gel was rinsed briefly in ultrapure water and transferred gently to the Bio-Rad GelDoc XR+ System to acquire the image, which was then post-processed using the software Image Lab.

This approach was used to assess protein homogeneity and complex formation following preparative SEC and to verify the efficiency of PNGase F-mediated deglycosylation, especially in h-RAGE glycoform analysis where shifts in apparent molecular weight were expected.

2.2.3 Size Exclusion Chromatography

Size-exclusion chromatography (SEC), also known as gel filtration, is a chromatographic technique that separates macromolecules based on their size as they pass through a porous resin. Larger molecules elute earlier because they are excluded from the internal pores of the resin beads, while smaller molecules are delayed due to their ability to penetrate the matrix. Elongated or branched molecules are also penalised in passing through the resin pores and then elutes as objects with apparent size larger than expected compared to their actual molecular weight.

SEC serves both as a purification step to separate the macromolecule of interest from other species present in the sample and as a biophysical tool to monitor the sample homogeneity. This method is often used to isolate reconstituted protein complexes from the single components that remained unbound, as most frequently these have smaller molecular weight than the reassembled object of interest.

In this work, SEC was employed for preparative purification of the hCHI3L1-His/hecRAGE-His reassembled complex.

All SEC runs were carried out at 4 °C on Fast Protein Liquid Chromatography (FPLC) ÄKTA systems (Cytiva), that allow to tightly control the flow of the chromatographic process, the injection of the sample to investigate in the SEC column, as well as to monitor and collect the distinct species that elutes from the resin. Proteins coming out from the column were in fact identified with a UV-cell detector by following the absorbance at 280 nm.

Commercial columns (Cytiva) Superdex 200 Increase 10/300 GL were used. Elution volumes were compared to molecular weight standards specific for each column and resin to estimate the apparent sizes of species.

For preparative purposes, large-scale purification of hCHI3L1-His/hecRAGE-His complex was performed on a column equilibrated in 50 mM HEPES pH 8.0, 300 mM NaCl. The chromatogram of the elution of molecular weight standards from the type of column used is shown in Fig. 28. Fractions corresponding to the complex peak were pooled, concentrated with 10 kDa cut-off Amicon Ultra centrifugal filters (Millipore) and used in downstream assays.

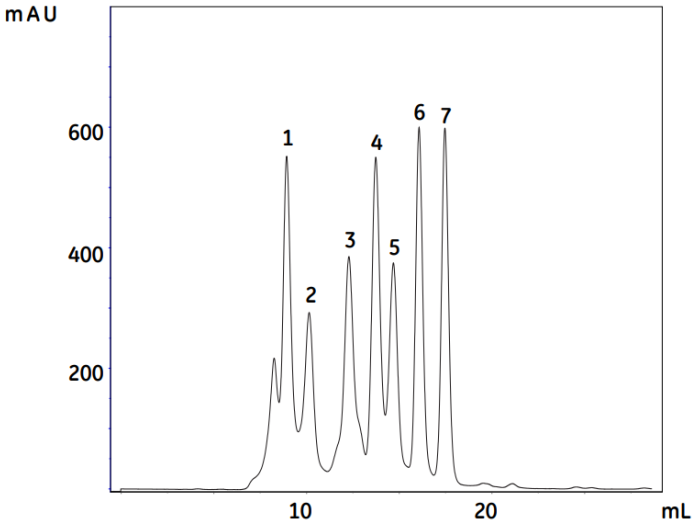


Figure 28. *Typical chromatogram from a function test of Superdex 200 Increase 10/300 GL using ÄKTApurifier.*
The peaks and their respective elution position of different reference proteins are shown. In particular the sample: 1 is thyroglobulin (MW 669 kDa) at 3 mg/mL, 2 is ferritin (MW 440 kDa) at 0.3 mg/mL, 3 is aldolase (MW 158 kDa) at 3 mg/mL, 4 is conalbumin (MW 75 kDa) at 3 mg/mL, 5 is carbonic anhydrase (MW 29 kDa) at 3 mg/mL, 6 is ribonuclease A (MW 13.7 kDa) at 3 mg/mL. The volume used was 100 µL, the eluent was 0.01 M phosphate buffer, 0.14 M NaCl, pH 7.4, the flow rate was 0.75 mL/min, run at room temperature with detection at 280 nm.
Image reproduced from the instructions for use sheet. [278] Copyright 2020 Cytiva.

2.2.4 Static Light Scattering

To determine the absolute molecular weights of the purified recombinant proteins, static light scattering (SLS) was performed using a triple-detection system combining refractive index (RI), ultraviolet absorbance (UV), and right-angle light scattering (RALS) placed downstream of SEC columns. SLS calculates molecular weight directly from the amount of light a molecule scatters in solution, which is proportional to its size, and allows precise measurement even for glycosylated or non-globular proteins.
hecRAGE-His and hCHI3L1-His were analyzed individually using a Malvern TDA-GPCMax system equipped with two in-tandem TSKgel G3000PWXL SEC columns (separation range 500-800.000 Da) maintained at 28 °C. The mobile phase consisted of PBS supplemented with NaCl up to 0.5 M to enhance protein stability and reduce aggregation or precipitation. This was particularly relevant for hCHI3L1-His, which in our hands appeared less stable at lower ionic strength.

75 µL of samples were injected at the stock concentration (specified in the Materials subsection), at a flow rate of 0.4 mL/min. Light scattering was measured at a fixed angle of 90° (RALS), and protein concentration was simultaneously determined by the RI detector. Although glycosylated proteins like hRAGE and hCHI3L1 may exhibit slightly different refractive index increments due to carbohydrate content or shape in the case of hRAGE, the standard protein dn/dc value of 0.185 mL/g was used for all analyses, which provides sufficiently accurate results, especially when comparing relative molecular weights or assessing monodispersity.

UV absorbance at 280 nm was recorded to support protein peak identification and assess signal integrity, but it was not used for molecular weight calculations. Instead, protein concentration was derived point-by-point from the RI signal, with the total protein amount represented by the area under the RI curve.

From the combination of RI and RALS signals, the software calculated key molecular parameters including the number-averaged molecular weight (Mn), the weight-averaged molecular weight (Mw), the Z-average molecular weight (Mz), the polydispersity index (Mw/Mn), and the most probable molecular weight (Mp). These values provided a detailed picture of sample monodispersity and the presence of any oligomeric species, or higher-order aggregates, helping confirm the solution-state behavior of each protein without relying on molecular weight ladders or assumptions of molecular shape.

2.2.5 Deglycosylation

Deglycosylation experiments were performed to assess the contribution of N-linked glycosylations to the interaction and complex stability between hCHI3L1 and hRAGE, as well as impact on protein molecular weights.

Both proteins are known to carry N-linked glycans: hCHI3L1 has a canonical N-glycosylation site at Asn60, and hRAGE at Asn25 and Asn81, in variable forms. To enzymatically remove these glycans, the endoglycosidase Peptide-N-Glycosidase F (PNGase F) was employed. PNGase F specifically cleaves between the innermost N-acetylglucosamine (GlcNAc) and asparagine residues of high mannose, hybrid, and complex N-linked glycans, converting glycosylated asparagine residues into aspartic acid.

Both denaturing and non-denaturing (native) deglycosylation protocols were tested to probe enzymatic efficiency and accessibility of glycosylation sites of hCHI3L1-His, hecRAGE-His and hRAGE-Fc. In denaturing conditions, glycan chains are fully exposed, maximizing enzyme activity, whereas in native conditions, glycan accessibility may be limited by the protein's tertiary structure. Small-scale tests were initially conducted to assess the reaction outcome.

In the denaturing protocol, proteins were fully denatured for maximal enzymatic accessibility. 10 µg of glycoprotein were mixed with 2 µL of 10× Glycoprotein Denaturing Buffer (5% SDS, 400 mM DTT) and adjusted with water to a total volume of 12 µL. This mixture was heated at 100 °C for 10 minutes to denature the protein, then immediately chilled on ice and briefly centrifuged. To this denatured mixture, 2 µL of 10× GlycoBuffer 2 (500 mM NaPi, pH 7.5), 2 µL of 10% NP-40, and 3 µL of water were added. PNGase F (1 µL of a 500 U/mL stock) was then added to reach a final volume of 20 µL, and the reaction was incubated at 37 °C for one

hour. The use of NP-40 is needed under these conditions to avoid SDS-mediated inhibition of the PNGase F enzyme.

For evaluating enzyme efficacy while preserving native protein folding, a non-denaturing protocol was implemented. The reaction contained 10 µg of protein in both the 20 µL and 50 µL final volumes, prepared with, 2 µL and 5 of 10× GlycoBuffer 2, respectively, and water as needed. PNGase F was added in amounts ranging from 2 to 4 µL, and the mixture was incubated at 37 °C for a duration of 6 and 24 hours. All the reaction conditions are shown in Table 4.

SDS-PAGE analysis was used to monitor shifts in protein mobility as an indirect measure of deglycosylation.

Denaturing

	Protein	10 ug
	10X Denaturing buffer	2 ul
	H2O	to 12 ul
	denature 10 min at 100C	
Add	10X Glycobuffer 2	2 ul
	10% NP40	2 ul
	H2O	3 ul
	PNGase F	1 ul
	total	20 ul
	incubated 1h, at 37C	

Native

	A	B	C
Protein	10 ug	10 ug	10 ug
10X Glycobuffer 2	2 ul	2 ul	5 ul
PNGase F	2 ul	4 ul	4 ul
H2O	to 20 ul	to 20 ul	to 50 ul
incubated 6h & 24h at 37C			

Table 4. *Deglycosylation reaction conditions for denaturing and native protocols*.

Reactions were scaled as needed for the protein quantity used in each test, maintaining PNGase F ≤10% of total volume to keep glycerol concentration under 5%, as recommended for activity.

2.2.6 Enzyme-Linked Immunosorbent Assays

To investigate the interaction between hCHI3L1 and hRAGE, Enzyme-Linked Immunosorbent Assays (ELISAs) were performed in both direct ligand-capture format with Fc-tag detection and indirect detection format using anti-hRAGE or anti-hCHI3L1 antibodies. Protocol parameters were refined over a series of iterative tests, optimizing protein concentrations, plate types, coating and blocking conditions, and detection strategies for signal sensitivity and reproducibility.

For the optimized assay, recombinant hCHI3L1 was diluted in PBS to a final concentration of 4 µg/mL and immobilized onto high-binding 96-well plates (Nunc MaxiSorp) by pipetting 50 µL/well and incubating overnight at room temperature (RT) under a lid to avoid evaporation. The following day, plates were washed once with PBS containing 0.05% (v/v) Tween-20 and blocked with 200 µL/well of 2% (w/v) BSA in PBS for 45 min at RT.

For direct Fc-tag detection ELISAs, wells were incubated with 50 µL of serially diluted (1:2) recombinant RAGE-Fc (ranging from 8 µg/mL to 0.125 µg/mL) in 1% PBS-BSA for 2 hr at RT. Detection was performed using an Horseradish Peroxidase (HRP)-conjugated anti-human Fc antibody (Jackson ImmunoResearch) diluted 1:5000 in PBS-BSA 1%, incubated for 1.5 hr at RT. Colorimetric detection was developed with 50 µL TMB (3,3',5,5' tetramethylbenzidine)

substrate for 15-20 min, and the reaction was stopped with 25 µL 1 M H_2SO_4 before absorbance was read at 450 nm.

For indirect detection ELISAs, binding was assessed using either hecRAGE-His or hCHI3L1-His as bait, followed by incubation with appropriate primary antibodies (anti-RAGE rabbit polyclonal or anti-CHI3L1 goat polyclonal, both at 1:2000 dilution), and secondary HRP-conjugated anti-rabbit or anti-goat antibodies (1:5000 dilution).

Competitive ELISAs were implemented by co-incubating on a hCHI3L1-His coating (4 µg/mL) hRAGE-Fc (at a constant 2 µg/mL concentration) and a putative inhibitory molecule in 25 µL + 25 µL mixtures. Caffeine, Tetra-Acetyl Chitotetraose, and Hexa-Acetyl Chitohexaose were tested as inhibitors, starting from 100 µg/mL and in 1:2 dilutions.

Plates were washed with PBS-Tween once after coating, and three times between all other steps, and care was taken to prevent wells from drying by preparing subsequent reagents in advance. All samples were typically run in triplicates, and signal development was monitored to fall within the linear detection range.

In order to collect the absorbance data on the plates, we used a CLARIOstar Plus (BMG Labtech). The data were analyzed using the software Prism and the four-parameter logistic (4PL) model was used to fit the data into a curve. The formula to apply this model is the following: $Y = Bottom + (Top - Bottom) / (1 + (IC_{50} / X)\text{^}HillSlope)$. This same fitted model was used also for any interpolation step. Instead, to estimate the Kd, we used the one-site total binding model, whose formula is:
$Y = Background + (B_{\max} \times X) / (K\_d + X)$.

2.2.7 Mass Photometry

To determine the molecular masses of hCHI3L1-His and hecRAGE-His, and of their putative complexes, as well as to explore their potential oligomeric states, we employed mass photometry using the Refeyn TwoMP instrument (Refeyn Ltd., Oxford, UK). Mass photometry is a label-free, single-molecule detection method based on interferometric scattering (iSCAT) microscopy, where a laser is focused on a glass coverslip. As protein molecules from the liquid sample land on the glass surface, they scatter the incoming light. This scattered light interferes with light reflected from the coverslip, creating tiny changes in the brightness, or contrast, of the image. Each landing event produces a contrast signal, and importantly, the size of that contrast is directly related to the mass of the molecule. Larger molecules scatter more light and therefore produce a stronger signal. Because of this, the mass of each molecule can be determined by comparing its signal to that of known protein standards with known molecular weight. This technique allows proteins and complexes to be measured in their native buffer without any labels or modifications, preserving their natural structure and interactions.

The samples used were first isolated via a large-scale preparative size-exclusion chromatography (SEC). They were prepared by first diluting the protein in the same buffer used during the SEC (50 mM Hepes pH 8.0, 300 mM NaCl) to a final working concentration between 5 and 80 nM. A clean sample carrier was assembled using high-precision glass coverslips and a silicone gasket.

Before sample addition, a reference movie was recorded with 18 µL of buffer alone to establish the optical baseline and check for spurious particle events. The diluted protein sample (2 µL) was then added to the buffer droplet and gently mixed by pipetting. The instrument's autofocus was engaged and adjusted to center on the reflection ring pattern, optimizing signal detection. Movies were captured for 60 seconds under ratiometric imaging mode, which provides real-time background subtraction and enhances sensitivity to individual landing events.

For molecular mass calibration, a standard curve was generated using proteins of known mass molecular weight (BSA, bovine thyroglobulin and apoferritin) within the instrument's working range of 30 kDa to 5 MDa megadaltons. The interferometric contrast from individual particle landing events was extracted using Refeyn's DiscoverMP software, which converts contrast values into molecular masses based on the calibration.

Data analysis was performed using the built-in software suite, where contrast peaks were compiled into histograms showing discrete mass populations, whose relative abundances were quantified by peak integration.

However, the application of MP to our samples was somewhat constrained by the instrument's lower detection limit of approximately 40 kDa. Since the molecular weights of our monomeric proteins are close to this threshold, we decided to complement the analysis with other techniques as well, such as Static Light Scattering. Indeed, at MP, the monomeric proteins' signals were near the edge of reliable detection, limiting the confidence in interpreting those peaks.

2.3 Structural Investigation Through X-Ray Crystallography

2.3.1 Crystallization Screening at Stanford Synchrotron Radiation Lightsource

The first crystallization experiments were conducted at the Stanford Synchrotron Radiation Lightsource (SSRL), SLAC National Accelerator Laboratory, Stanford University, Menlo Park, CA, USA.

The crystallization process was designed as a multi-stage iterative screening effort, integrating commercial screening kits, custom optimization, seeding strategies, and buffer exchange protocols to maximize the chance of obtaining high-quality protein crystals of the hCHI3L1-hRAGE complex. The crystallization strategy followed a classical vapor-diffusion approach, primarily using the sitting-drop format.

Manual Setups of Crystallization Screens

Manual crystallization trials were executed in 96-well plates, with each well containing a central reservoir (50-70 µL) surrounded by two lateral drop pits. The reservoir contained the crystallization solution, and the drops were composed of 1 µL of protein complex (at 1:1 or 1:2 ratio by concentration) mixed with 1 µL of reservoir solution. Upon sealing the plates, water vapor gradually equilibrated between the drop and reservoir, driving slow dehydration and, ideally, crystal formation.

hecRAGE-His was used at initial concentrations of 10.24 mg/mL or concentrated to 20.48 mg/mL (x2) using centrifugal filter units. hCHI3L1-His was supplied at 2.4 mg/mL and concentrated up to 10.24 mg/mL by ultrafiltration (x4.267). Protein complex PBS 1x solutions were prepared in two stoichiometric ratios to test potential biological binding configurations: a 1:1 molar ratio using equimolar amounts of hecRAGE-His and hCHI3L1-His, and a 2:1 molar ratio to explore the possibility of sRAGE homodimerization during binding. The incubation of the protein components prior to screens setup was performed at room temperature for approximately one hour. Manual crystallization plate setups were performed in a 4 °C cold room to limit evaporation. The first crystallization screens utilized the Shotgun I screen (SG1), which comprises a diverse array of solutions known for their historical success across various protein systems. Plates were incubated at two different temperatures, 4 °C and 16 °C, to account for the differing thermal sensitivities of the individual protein partners. Initial microscopy inspections were carried out on days 3, 5, and 7 post-set up.

Optimization Trials

Promising conditions identified in initial screens were selected for further optimization. Modified conditions were set up in new plates, testing different protein-to-reservoir ratios (1:1, 2:1, 1:2, and 2:2) and protein concentrations ranging from 4 to 8 mg/mL. Conditions were tested across multiple wells, with each row representing a different ratio and each column representing a different PEG concentration or salt concentration. In addition to sitting-drop setups, the hanging-drop vapor diffusion method was also evaluated as an alternative crystallization approach during the optimization phase.

Seeding and Morphological Refinement

To promote nucleation and improve crystal morphology, seeding techniques were employed. Needle-shaped crystals from earlier wells were harvested and mechanically fragmented using vortexing with a metal bead. These seed solutions were introduced into fresh crystallization drops using a fine syringe needle.

Buffer Exchange

To improve solubility and crystallization compatibility, alternative buffers to PBS were also tested. HEPES buffer at pH 7.5 (20 mM), supplemented with 250 mM NaCl, and Tris-HCl at pH 8.0 (25 mM), supplemented with 350 mM NaCl, were used to improve hCHI3L1-His solubility.

Buffer exchange was performed following two methods. Partial exchange involved stepwise dilution with the new buffer followed by reconcentration using centrifugal filters. Complete buffer exchange was achieved using dialysis with Slide-A-Lyzer cassettes (ThermoFisher). hCHI3L1-His was dialyzed using a 7 kDa cutoff cassette in 1 L of buffer with stirring, followed by a second fresh buffer replacement after 2 hours and overnight dialysis. hecRAGE-His was dialyzed in parallel using a 20 kDa cutoff cassette under identical conditions.

Following buffer exchange, new protein complex solutions were prepared using the freshly dialyzed samples. Protein concentrations were re-adjusted using centrifugal filters prior to mixing and crystallization setup, reaching 7.4 mg/mL in 200 µL for hCHI3L1-His and 9.6 mg/mL in 100 µL for hecRAGE-His in HEPES buffer. In the TRIS buffer condition, final

concentrations were 9.26 mg/mL in 250 µL for hCHI3L1-His and 23.76 mg/mL in 50 µL for hecRAGE-His.

Gryphon Robot Setups of Crystallization Screens

The second crystallization screening protocol was performed using a Gryphon liquid handling robot (Art Robbins Instruments), allowing for speed and precision in setting up nanoliter-scale drops. The robot consisted of a 96-channel syringe dispenser for crystallization-cocktail transfer and one Nano-nozzle non-contact dispenser for the protein sample. The method was based on a protocol kindly provided by PhD Crissy L. Tarver, and began with priming of the fluidics system using ultrapure water, followed by an automated wash cycle of the aspiration and dispensing tips.

Using 96-channel syringe aspiration, 100 µL of each reservoir solution (from commercial screening kits) was mixed and 85 µL aspirated per channel. 60 µL was dispensed into the reservoir wells of the sitting-drop vapor diffusion plate. Reservoir solutions were then dispensed into the three designated drop positions in volumes of 0.5, 0.25, and 0.75 µL, respectively.

Next, 175 µL of mixed ligand-receptor protein solution was aspirated into the nanodispenser, which then delivered 0.5, 0.75, and 0.25 µL of protein solution into the corresponding drops, with washing between the second and third dispenses. The unused protein was restituted back in the sample tube.

After drop setup, a final wash and purge cycle was executed to avoid cross-contamination between samples and prevent clogging.

Plates were spun at 950 rpm for 3 minutes to ensure centering of drops within the wells. The plates were then visually inspected using a stereomicroscope to confirm proper drop formation.

Final crystallization trials included five commercial sparse matrix kits: Index HT, PEGRx HT, Shotgun I (SG1), and HR2-133 (Hampton Research), tested in two buffer conditions: 25 mM Tris-HCl pH 8.0, 350 mM NaCl, and 20 mM HEPES pH 7.5, 250 mM NaCl. All plates were incubated at 16 °C for crystal growth by sitting-drop vapor diffusion and monitored with the microscope during the following weeks for crystal growth.

Crystal Harvesting and Diffraction Data Collection

Selected crystals were harvested using nylon loops and transferred to cryoprotectant solutions prepared by mixing reservoir solution with 25% glycerol in a 3:1 ratio. Each crystal was briefly dipped into the cryoprotectant and flash-frozen in liquid nitrogen. Mounted crystals were transferred to SSRL beamlines using automated robotic systems.

All diffraction experiments were performed at Beamline 12-2 of the Stanford Synchrotron Radiation Lightsource (SSRL), which provides a microfocused high-flux X-ray beam (especially useful given the small size of all the crystals examined) and supports fully automated data acquisition through the Blu-Ice interface. Blu-Ice is a software platform developed at SSRL for automated crystallographic data collection that integrates beamline hardware control, sample mounting, goniometer positioning, exposure setup and real-time feedback.

Diffraction data collection was carried out using beamline-specific parameters optimized for each crystal. For the single crystal that yielded a complete dataset, the beam size was set to 40

× 15 μm, with the detector placed at a distance of 350 mm and the beamstop positioned at 34.986 mm. Oscillation images were collected with a rotation increment (Δφ) of 0.2° and an exposure time of 0.2 seconds per frame. The total oscillation range varied between 180° and 360°, with the sample loop positioned at a φ angle of approximately 166.75°. Crystals were maintained at cryogenic temperatures (100 K) using a liquid nitrogen cryostream throughout the data collection process.

Diffraction Data Analysis

Diffraction data processing and validation were carried out using the Xtriage module from the Phenix software suite. The merged dataset was analyzed to determine the space group, unit cell parameters and overall data quality. Evaluation parameters included mean intensity, signal-to-noise ratio $\langle I/\sigma(I)\rangle$, and internal consistency indicators such as R-merge, R-meas and R-pim. These values were used to guide decisions for further phasing and model building.

The analysis workflow included quality control checkpoints to ensure the data were suitable for downstream structural determination. A completeness threshold of >99% was considered acceptable for initiating refinement and pursuing structure solution.

2.3.2 Crystallization Screening at the European Institute of Oncology

A further crystallization screening campaign was carried out at the European Institute of Oncology (IEO, Milan, Italy) in the effort to crystallize the hCHI3L1-hRAGE protein complex.

Ligand-Receptor Mixing and Buffer Exchange

The initial protein samples were prepared by mixing hCHI3L1-His (0.74 mg/mL = 17.7 μM) and hecRAGE-His (1.14 mg/mL = 32.6 μM) at a molar ratio of 3:1 in PBS in order to saturate the receptor, with concentrations of 0.74 mg/mL (17.7 μM, as estimated from the theoretical ε with reduced cysteines) for hCHI3L1-His and 1.14 mg/mL (32.6 μM) for hecRAGE-His. Upon mixing, the final concentrations of hCHI3L1-His and hecRAGE-His were approximately 14.5 μM and 4.8 μM (3:1), respectively. This mixture led to visible precipitation, which was removed by centrifugation. Sodium chloride was then added to the supernatant to a final concentration of 300 mM in an attempt to improve protein solubility. The clarified solution was concentrated to 0.5 mL, reaching an estimated concentration of 7.7 mg/mL for the hCHI3L1-His/hecRAGE-His putative complex, assuming a 1:1 stoichiometry (theoretical molecular weight: 76.9 kDa; calculated extinction coefficient: 107,425 $M^{-1}cm^{-1}$).

The putative protein complex was subsequently purified through a large scale preparative SEC using a Superdex 200 10/300 GL column equilibrated in a buffer containing 50 mM HEPES, pH 8.0, and 300 mM NaCl. Peak fractions B2-C4 of the column were analyzed by SDS-PAGE to verify the presence of both protein bands, and subsequently pooled and concentrated to approximately 16 mg/mL, although some further precipitation was observed at this stage. After clarification, the final protein solutions used for crystallization were at 13 mg/mL and 6.5 mg/mL, which were used for the top and bottom drop wells, respectively.

Setups of Crystallization Screens with Mosquito Robot
Crystallization screening in 96-wells MRC-2 plates (Hampton Research) were set up using a Mosquito robot (SPT Labtech), employing the sitting-drop vapor diffusion method. Drops were prepared by mixing 0.1 µL of protein solution with 0.1 µL of reservoir solution. Two different commercial screening kits were used: Index HT and PEGRx HT (Hampton Research). The plates were incubated at 4 °C and 20 °C and monitored with the microscope during the following months for crystal growth.

Crystal Diffraction
The most promising crystals formed in the screens were harvested, transferred to cryoprotectant solutions (reservoir solution supplemented by 20% glycerol) and flash-frozen in liquid nitrogen. The crystals were shipped in liquid nitrogen to the European Synchrotron Radiation Facility (ESRF), Grenoble, France, where diffraction data was collected at the fully automated ID30A-1/MASSIF-1 beamline and transmitted to us.

Analysis of Diffraction Data
Diffraction data were first processed directly at ESRF through the dedicated automatic beamline workflow. Specifically, a variety of pipelines for diffraction data indexing, integration, merging and scaling was run in parallel and the solution with the best parameters was then used for phasing attempts through molecular replacement. Models derived from automated search on the Protein DataBank (PDB ID: 1HJW and 8R42) were used and the successful solutions were carried to the final refinement stage.

To analyse and further validate the results of the automated processing, the datasets with the highest resolution were re-processed by hand using the CCP4 software suite. Indexed files from automated pre-processing were scaled and merged with Aimless. Molecular replacement was performed with PHASER-MR using the same 1HJW or 8R42 structures as models. Model improvement was carried out with iteration of structure visualization and manual adjustment (coot software) together with refinement with REFMAC5.

2.4 AI Modeling

2.4.1 ColabFold

Computational Strategy and Study Design
To explore the structural basis of interaction between human CHI3L1 and the soluble extracellular region of human RAGE, we employed AI-based protein structure prediction using ColabFold, an accelerated and accessible implementation of AlphaFold2. ColabFold is available at https://github.com/sokrypton/ColabFold. Its pipeline integrates MMseqs2 (Many-against-Many sequence searching), a software suite to search and cluster huge protein and nucleotide sequence sets for multiple sequence alignment (MSA) generation. ColabFold also includes tools for template retrieval, stochastic seed sampling, and multimeric modeling.

The strategy was divided into two principal phases. The first phase involved exploratory testing of various ColabFold-supported models, including AlphaFold2_mmseqs2 with and without templates, RosettaFold, and AlphaFold2_advanced_v2. The second phase consisted of systematic parameter optimization and seed-based prediction using AlphaFold-Multimer v2.3, ultimately selecting this model for high-resolution structure prediction.

Protein Sequences and Input Preparation

The protein interfaces selected for prediction included hCHI3L1 in complex with monomeric hRAGE-VC1C2, hRAGE-VC1, or hRAGE-C2, as well as the hS100A8/A9 heterodimer in complex with monomeric hRAGE-VC1. All input sequences were obtained from UniProt.

For hRAGE-VC1C2 (UniProt accession ID: Q15109), the C-terminal region was trimmed to exclude the final unstructured segment observed in the crystal structure of VC1C2 (PDB ID: 4LP5). In the case of hRAGE-VC1 (also Q15109), the sequence of the two domains was extended by 19 amino acids at the C-terminus to include an additional segment modeled in the crystal structure of VC1 (PDB ID: 3CJJ). These extensions and trimmings enabled better alignment and comparison of the 3D predictions with available receptor structures. For hRAGE-C2 (Q15109), we used the C2 domain as input sequence.

For hCHI3L1 (UniProt accession ID: P36222), the full-length sequence was used as input for ColabFold, and the predicted structure was compared to the crystal structure available under PDB ID: 1NWR.

For the hS100A8/A9 heterodimer, full-length sequences for both subunits were used, S100A8 (UniProt ID: P05109) and S100A9 (UniProt ID: P06702). The predicted complex was compared to the available reference crystal structure (PDB ID: 1XK4).

All the amino acids sequences are shown in Table 5.

The sequences were formatted as a single string with a colon ":" separator to distinguish the two or more chains, as required by ColabFold for multimeric modeling. The final input sequence string was directly passed into the notebook interface as plain text and processed for alignment, structural modeling, and visualization.

Protein	**UniProt Accession ID**	**Amino Acids**	**Input Sequences**
hCHI3L1	P36222	22-383	YKLVCYYTSWSQYREGDGSCFPDALDRFL CTHIIYSFANISNDHIDTWEWNDVTLYGML NTLKNRNPNLKTLLSVGGWNFGSQRFSKI ASNTQSRRTFIKSVPPFLRTHGFDGLDLAW LYPGRRDKQHFTTLIKEMKAEFIKEAQPGK KQLLLSAALSAGKVTIDSSYDIAKISQHLD FISIMTYDFHGAWRGTTGHHSPLFRGQED ASPDRFSNTDYAVGYMLRLGAPASKLVM GIPTFGRSFTLASSETGVGAPISGPGIPGRFT KEAGTLAYYEICDFLRGATVHRILGQQVP YATKGNQWVGYDDQESVKSKVQYLKDR QLAGAMVWALDLDDFQGSFCGQDLRFPL

			TNAIKDALAAT
hRAGE-VC1C2	Q15109	22-321	AQNITARIGEPLVLKCKGAPKKPPQRLEW KLNTGRTEAWKVLSPQGGGPWDSVARVL PNGSLFLPAVGIQDEGIFRCQAMNRNGKET KSNYRVRVYQIPGKPEIVDSASELTAGVPN KVGTCVSEGSYPAGTLSWHLDGKPLVPNE KGVSVKEQTRRHPETGLFTLQSELMVTPA RGGDPRPTFSCSFSPGLPRHRALRTAPIQPR VWEPVPLEEVQLVVEPEGGAVAPGGTVTL TCEVPAQPSPQIHWMKDGVPLPLPPSPVLI LPEIGPQDQGTYSCVATHSSHGPQESRAVS ISII
hRAGE-VC1	Q15109	23-240	AQNITARIGEPLVLKCKGAPKKPPQRLEW KLNTGRTEAWKVLSPQGGGPWDSVARVL PNGSLFLPAVGIQDEGIFRCQAMNRNGKET KSNYRVRVYQIPGKPEIVDSASELTAGVPN KVGTCVSEGSYPAGTLSWHLDGKPLVPNE KGVSVKEQTRRHPETGLFTLQSELMVTPA RGGDPRPTFSCSFSPGLPRHRALRTAPIQPR VWEPVPLEEVQL
hRAGE-C2	Q15109	227-317	PRVWEPVPLEEVQLVVEPEGGAVAPGGTV TLTCEVPAQPSPQIHWMKDGVPLPLPPSPV LILPEIGPQDQGTYSCVATHSSHGPQESRA VS
hS100A8	P05109	1-93	MLTELEKALNSIIDVYHKYSLIKGNFHAVY RDDLKKLLETECPQYIRKKGADVWFKELD INTDGAVNFQEFLILVIKMGVAAHKKSHEE SHKE
hS100A9	P06702	1-114	MTCKMSQLERNIETIINTFHQYSVKLGHPD TLNQGEFKELVRKDLQNFLKKENKNEKVI EHIMEDLDTNADKQLSFEEFIMLMARLTW ASHEKMHEGDEGPGHHHKPGLGEGTP

Table 5. *Amino acids input sequences for the proteins modeled in ColabFold.*

Model Selection and Configuration

The AlphaFold-Multimer v2.3 model was selected based on its improved performance in complex prediction tasks and was accessed via the AlphaFold2_advanced_v2 ColabFold notebook. This experimental notebook version allowed precise control over sequence alignment, parameter configuration, seed sampling, template selection, and recycling depth.

We edited the notebook Python 3.* code in order to perform multiple screenings and optimize the results of the predictions at high parameters depth, and testing multiple models. Predictions were carried out with a maximum of 24 recycles, which enabled AlphaFold to iteratively refine its predictions by feeding forward structural outputs. This recycling process continued until

structural convergence or until the set maximum number of iterations was reached. All predictions were executed using Google Colab Pro+ with A100 GPUs to ensure adequate computational capacity and runtime performance.

Multiple Sequence Alignment Construction

Multiple sequence alignments (MSAs) were generated using MMseqs2, a fast and scalable sequence search tool integrated into ColabFold. MMseqs2 utilized two primary databases, UniRef30_2302 and colabfold_envdb_202108, both available at https://colabfold.mmseqs.com/, to retrieve homologous sequences that informed evolutionary constraints.

MSA Filtering Parameters and Optimization

MSA filtering was applied to manage trade-offs between alignment quality and sequence diversity. It included a coverage threshold (cov), ranging between 0% and 100%, a maximum identity threshold (id) among the aligned sequences, either 90% or 100%, and a threshold of minimum identity (qid) to the query sequence, from 0% to 30%.

The optimal settings for each prediction were determined through an iterative process involving manual review, automated exploration of the parameters space, and univariate scans over 128 random seeds per condition.

Templates

Two strategies for template inclusion were explored: automatic retrieval of multiple templates for feature extraction using MMseqs2 from the pdb100_230517 dataset (available at https://colabfold.mmseqs.com/) and manual specification of custom templates. Templates were used to guide folding by aligning known structures with the input sequences. Custom PDB templates included 1NWR (for hCHI3L1), 3CJJ, 4LP5 or 4YBH (for hRAGE), and 1XK4 (for hS100A8/A9), selected based on structural relevance and residue coverage. These were integrated via the ColabFold notebook's custom template feature.

Seeds and Stochastic Parameters

To address the inherent stochasticity of AlphaFold’s neural network and obtain better results, structure prediction was repeated across 128 unique random seeds for each set of parameters. Each seed initializes the prediction process with different random values, potentially sampling distinct conformational states, especially in low-confidence regions.

This approach allowed us to identify seeds yielding high-confidence structures that would otherwise be inaccessible through single runs. Additionally, the interaction between stochastic settings such as masked language modeling and dropout regularization was systematically explored.

Model Confidence

All predicted structures were assessed using metrics computed by AlphaFold-Multimer v2.3. These included the predicted local distance difference test (pLDDT) for per-residue confidence, predicted aligned error (PAE) for interdomain alignment accuracy and predicted TM-score (pTM) for global fold quality. The interface-predicted TM-score (ipTM) was used

to evaluate the reliability of the inter-chain docking prediction. The combined multi score (0.8 × ipTM + 0.2 × pTM) was used to rank all models for downstream analysis. Additional metrics such as actifpTM (active interface pTM), which isolates well-ordered interface residues, were also computed.

Output Generation and Raw Files

The output generation pipeline involved a series of well-structured and hierarchical processes designed to convert raw input sequences into multiple sequence alignments (MSAs), filtered and formatted appropriately for structural prediction, via MMseqs2 databases and HH-suite tools.

Upon query submission, MMseqs2 is called through a wrapper function that handles the backend interaction with the online API at ColabFold. Here, the query is matched against large sequence databases including UniRef and environmental databases (e.g., BFD, MGnify, MetaEuk). The matching results are returned in the form of .a3m alignment files. When the environmental search mode is active (use_env = True), both uniref.a3m and bfd.mgnify30.metaeuk30.smag30.a3m files are returned. These represent hits from standard and environmental sources, respectively.

Once retrieved, these .a3m files are parsed into memory as MSA vectors, along with deletion matrices encoding gap information per sequence and position. All sequences are merged and grouped based on their per-chain coverage. The coverage assessment calculates how many residues in each segment of a sequence alignment are ungapped. For each chain, a binary label is assigned to each aligned sequence: a "1" indicates coverage of that chain, while a "0" indicates complete absence of coverage. Sequences that fall in between and show partial coverage are excluded from downstream processing. This labeling results in strings such as "10" or "01", denoting, for instance, good alignment only for the first or second chain, respectively. These labels are used to sort the sequences into separate files like 10.a3m, 01.a3m, and 11.a3m, in case of paired alignment.

Each a3m file is then filtered using hhfilter, a tool from HH-suite, with customizable parameters for identity threshold (id), query identity (qid), and coverage (cov). This step helps remove redundancy and retain diverse sequences relevant for structure prediction. The filtered files are written with a .out.a3m extension to distinguish them from the unfiltered versions. The output files and raw data are organized into a structured directory that reflects the logical steps of the pipeline. The top-level output folder (named based on the job title and hash of the input sequence) contains the primary output file msa.a3m, which is the final merged and filtered MSA used for AlphaFold modeling. It also contains an intermediate file msa_tmp.a3m, used during template alignment if templates are enabled. Visualization of sequence coverage and diversity is stored in msa_feats.png.

A subdirectory named msa/ holds the coverage-labeled hhfilter input and output alignment files (10.a3m, 01.a3m, 10.out.a3m, 01.out.a3m, etc.). Within this directory, an _env/ subfolder stores the raw database-derived .a3m files and other relevant files like pdb70.m8, which lists the results from HHsearch template matching (if enabled), and MMseqs2 cache files in out.tar.gz.

The final predicted structures and model metrics are saved in a pdb/ subdirectory. This includes the best-scoring model (best.pdb), its visualizations (best.png, best.pdf), and

numerical evaluation outputs in compressed .npz format (best.npz, all.npz) containing per-residue confidence scores (pLDDT), predicted aligned error (PAE) matrices, inter-chain and intra-chain interface metrics (if applicable), and tracking data for each model and recycle step.

All settings used for the job, including input options, model configuration, and runtime settings, are stored in settings.txt for reproducibility. A comprehensive log of the modeling run, including scores for each seed and model, is captured in log.txt. This entire output directory is compressed into a downloadable .zip archive at the end of the run, ensuring that all raw and processed files are retained in a format suitable for downstream analysis or archival. The complete output file tree is displayed in Fig. 29.

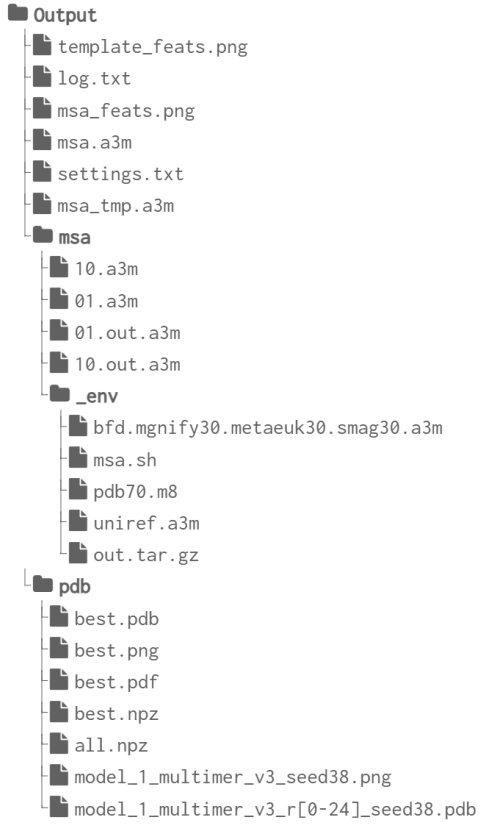


Figure 29. *Output file tree from ColabFold.* Diagram generated with folder2image.com.

Multimer Interface Characterization

Final model structures were analyzed for residue-level interactions using PLIP (Protein-Ligand Interaction Profiler), which identified hydrophobic contacts, hydrogen bonds, salt bridges, and π-cation interactions at the interface. [279] These analyses allowed us to verify the plausibility of the predicted interface and provided biochemical insights into the nature of the CHI3L1-RAGE binding mechanism.

Predicted complexes were visualized using PyMOL 3.1, with structural features mapped according to confidence scores and domain-specific residues annotated manually. Structural interface residues were cross-referenced with published experimental data and previous crystallographic models to assess biological plausibility.

2.4.2 Code Availability

All computational analyses were performed using custom Python scripts. The code was developed to automate AlphaFold parameters optimization, to perform analysis of the MSA and generate plots relevant to the study. To ensure reproducibility and transparency, all scripts used in this work are publicly available via a dedicated GitHub repository (https://github.com/LucaM00).

2.5 Protein Structures Analysis and Visualization

For protein structure analysis and visualization, we used PyMOL (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC). Electrostatic potential maps at physiological pH were generated using UCSF Chimera, which allowed for the calculation of solvent-accessible electrostatics and the creation of high-quality structural representations.

3. Results

3.1 Binding Studies

3.1.1 CHI3L1 Binds RAGE in the nM Affinity

A process of serial optimization conducted systematically on most of the parameters of the enzyme-linked immunosorbent assay (ELISA) assay revealed that CHI3L1 coats the plate in a nearly random fashion when the process is performed at 4 °C. Meanwhile, the optimal parameter for the coating of CHI3L1 to achieve reproducible results is room temperature. We also found that 4 μg/mL was its optimal coating concentration to achieve binding curves where the linear phase between the lower and upper plateaus was most clearly resolved. This concentration of h-CHI3L1-His corresponds to a molarity of 0.096 μM if we calculate it through the theoretical molecular weight (MW) of 41.7 kDa.

Upon hCHI3L1-His coating, hRAGE-Fc was titrated starting from a concentration of 8 μg/mL (corresponding to 0.227 μM if we consider the theoretical MW of 35.2 kDa), and the resulting absorbance values were plotted to generate a binding curve. A four-parameter logistic (4PL) model was used to fit the raw ELISA data and visualize the binding curve, as shown in Fig. 30. This empirical model is commonly used in ELISA analysis to describe sigmoidal responses without assuming a particular binding mechanism. To better define the upper plateau and improve the robustness of the Kd estimation, two additional high-concentration points (0.454 μM and 0.908 μM) were added. Their corresponding absorbance values were interpolated from the fitted 4PL curve.

Although the 4PL model provided an excellent empirical fit to the data, it is not designed to extract mechanistic parameters. Therefore, using the expanded dataset, the equilibrium dissociation constant (Kd) was estimated at 79.66 nM (95% CI: 46.81-139.20 nM) by applying a nonlinear regression with a one-site total binding model. This model is a mechanistic, saturable ligand-receptor binding model that assumes a 1:1 interaction between CHI3L1 and RAGE, following classic receptor-ligand kinetics and resulting in a hyperbolic binding curve. However, the stoichiometry and mechanism of this interaction are not fully characterized, and alternative possibilities, such as multivalency, cooperativity, or allosteric regulation, cannot be excluded. Thus, the reported Kd should be interpreted as an apparent affinity under the specific assay conditions.

Importantly, the estimated Kd falls within the nanomolar range, consistent with the value previously reported by Darwich et al., 2020 (16.65 nM). [93] The small discrepancy may reflect differences in experimental setup, protein constructs, or the inclusion of interpolated points in the present analysis. While the interpolation approach introduces a degree of model circularity, the resulting fit was consistent with the experimental trend and yielded a Kd in line with literature estimates.

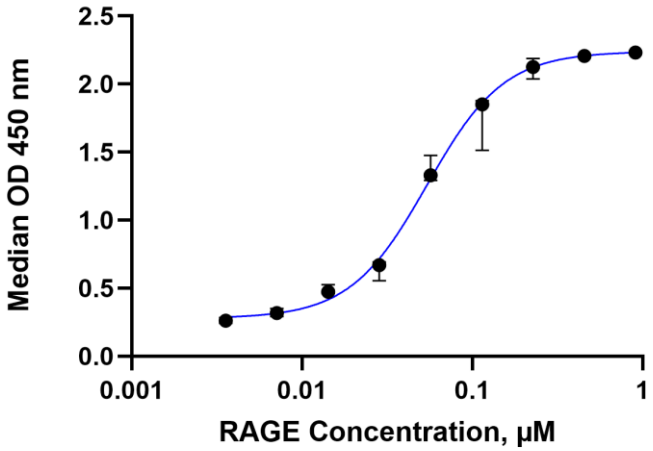


Fig. 30. *ELISA binding curve for the CHI3L1-RAGE interaction*.
The curve was obtained by coating the plate with a hCHI3L1-His concentration of 4 µg/mL (theoretical 0.096 µM), titrating hRAGE-Fc with 1:2 dilutions from an initial concentration of 8 µg/mL (theoretical 0.227 µM), fitting the data to a 4PL model and interpolating two upper points of concentration (0.454 µM and 0.908 µM). The x axis is displayed in logarithmic scale. The Kd was estimated to be 79.66 nM via the one-site total binding model, applied through Prism.

3.1.2 Mildly Deglycosylated Proteins Display a Stronger Binding

In order to test the ability of the two proteins to bind in (at least partial) absence of the glycans belonging to their N-glycosylation, we applied a deglycosylation protocol, described in the Materials and Methods section, to the samples of hCHI3L1-His and hRAGE-Fc. The enzyme used to attempt the deglycosylation of the proteins was the PNGase-F, which cleaves the bond between the innermost GlcNAc and asparagine in glycoproteins, releasing a wide range of N-linked glycans (high-mannose, hybrid, and complex). However, it cannot cleave N-glycans with core α(1→3)-linked fucose, which are rare in mammals. [280]

RAGE N-glycans, the most relevant ones in literature data, as explained in the Introduction section, are mainly of complex type but are poorly characterized. Therefore, we did not know what to expect from the deglycosylation reaction. For this reason, we tested two different durations of the deglycosylation reaction, 2 and 13 hours, and all the possible combinations of binding between the two protein species at various exposure times to deglycosylation. All the fitted curves (via the 4PL model) can be visualized in Fig. 31.

By comparing the binding curves, it appeared that, contrary to the literature-based expectations of relevance of the glycans for the binding, deglycosylation did not affect the binding strength at all. Instead, the proteins subjected to 2 hours of deglycosylation exhibited a significantly higher binding curve with respect to the control, with a p-value < 0.0001 at the F-test.

We must underline that we have no certainty about what deglycosylated species(s) populated the samples after the reactions, therefore we cannot infer any degree of involvement or not of these glycans in the molecular dynamics of the binding. In order to elucidate the effect of this

deglycosylation reaction on the molecular mass of the proteins, we performed an SDS-Page on various deglycosylated samples, explained in another subsection of the Results.

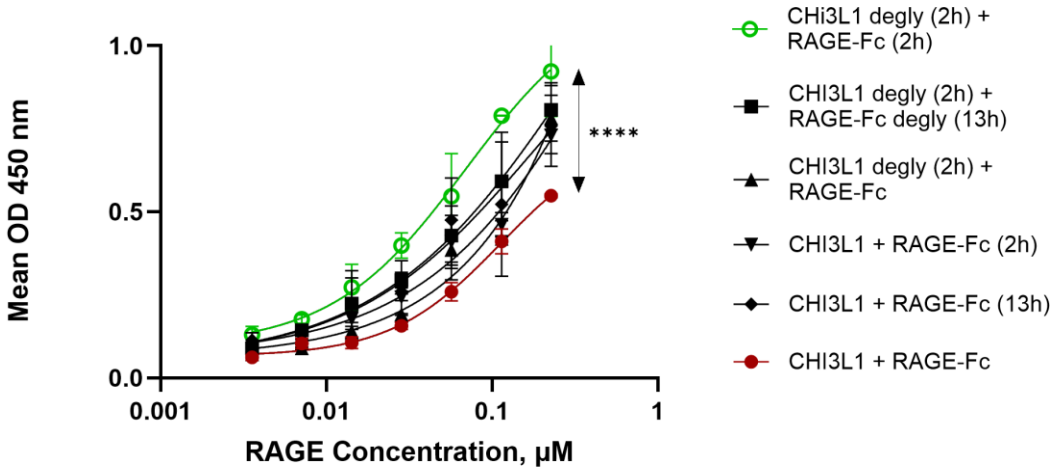


Figure 31. *ELISA curves for all the combinations of deglycosylated proteins*.
hCHI3L1-His, deglycosylated for 2 hours or not,w as coated at 4 µg/mL (theoretical 0.096 µM) and the titration of hRAGE-Fc, deglycosylated for 2 or 13 hours or not, started from 8 µg/mL (theoretical 0.227 µM). All the curves were fitted with a 4PL model and a global F-test comparison test was conducted between the green curve (both proteins deglycosylated for 2 hours) and the red curve (both natural), which revealed a statistically significant p-value < 0.0001, meaning that the two curves are globally different.

3.1.3 Tetraacetyl-chitotetraose and Caffeine Do Not Inhibit CHI3L1-RAGE Binding

Two ligands of the CHI3L1 carbohydrate-binding cleft were tested in a competitive ELISA with hRAGE-Fc to verify the hypothesis that these ligands occupy on CHI3L1 the same region that the extracellular domain of RAGE could employ to bind CHI3L1.

The two ligands were tetraacetyl-chitotetraose (composed of 4 GlcNAc units), which has a binding affinity for CHI3L1 of 88.7 µM, and the pan-chitinase inhibitor caffeine, whose IC50 for Chitinase-B1 is 469 µM, both measured with fluorescence analysis. [196, 281]

In the competitive ELISA, a polyclonal antibody against RAGE (whose details can be read in the Materials and Methods section) was used as a possible positive control of the inhibition of the CHI3L1-RAGE binding. The results of the assay are shown in Fig. 32. As expected from the relatively low Kd, neither of the two tested compounds, and surprisingly not even the antibody against RAGE, caused a significant reduction in the hCHI3L1-His-RAGE signal intensity of the binding. A fixed concentration of RAGE was selected and the tested molecules were titrated starting from a concentration of 100 µg/mL.

Only two ligands among various potential selected compounds were tested due to limited purchase availability. We will also test the hexaacetyl-chitohexaose (6 GlcNAc), whose Kd upon fluorescence analysis is 4 µM.

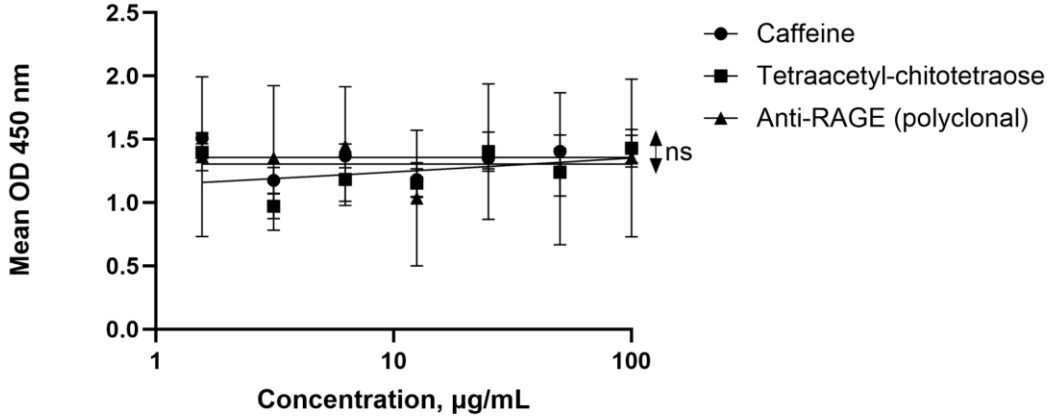


Figure 32. *Competitive ELISA between RAGE and potential inhibitor compounds for the binding with CHI3L1*.

The coating concentration for hCHI3L1-His was 4 µg/mL (theoretical 0.096 µM), and hRAGE-Fc was used at a fixed concentration, too. This was selected as 2 µg/mL (theoretical 0,0568 µM) based on previous ELISA results as the x-axis point the closest to the middle point of the linear range of the binding curve. All the three tested molecules, tetraacetyl-chitotetraose, caffeine and the polyclonal anti-RAGE were titrated with 1:2 dilutions from a starting concentration of 100 µg/mL. Neither of the three curves, fitted with a 4PL model, displayed statistically significant differences with each other at the F-test.

3.2 Single Proteins Characterization

3.2.1 CHI3L1 Solubility Is Enhanced by High Ionic Strength

In all the buffers used for CHI3L1 preparations (PBS, HEPES and Tris-HCl at varying concentrations), the protein exhibited a high tendency to precipitate in absence of NaCl addition. This phenomenon was observed especially during the process of concentrating the sample, which not only happened to be longer and difficult to achieve at low NaCl concentrations, but also caused visible precipitation. This finding confirms a similar low salt concentration behaviour by CHI3L1 already documented in literature. [282]

We tested different NaCl concentrations, finally finding 500 mM as the optimal minimal concentration to achieve a crystal-clear sample in any buffer avoiding visible precipitation.

For what concerns the preparative solutions for the crystallization plates, it is well known that high concentrations of NaCl can hinder protein crystallization by increasing protein solubility (and making it harder to reach supersaturation conditions), screening electrostatic interactions, and disrupting weak intermolecular contacts needed for crystal lattice formation. While moderate salt levels may stabilize proteins and reduce aggregation, excessive ionic strength often prevents the ordered assembly required for crystallization. Therefore, three slightly lower NaCl concentrations were selected, in order to find a trade-off between protein solubility maximization and crystals favorable conditions: namely, 300 mM for PBS; 250 mM for HEPES; and 350 mM for Tris-HCl buffer.

3.2.2 Either Protein Is Monomeric in (PBS, 0.5 M NaCl) Solution

To assess the oligomeric state and molecular homogeneity of hCHI3L1-His and hecRAGE-His, both proteins were subjected to static light scattering (SLS) analysis in PBS supplemented with 0.5 M NaCl. The experimental design allowed absolute molecular weight determination independent of protein standards or assumptions of molecular conformation, making it especially suitable for analyzing non-globular or glycosylated proteins.
The refractive index (RI) and right-angle light scattering (RALS) signals from both runs confirmed monodisperse and proteinaceous profiles.

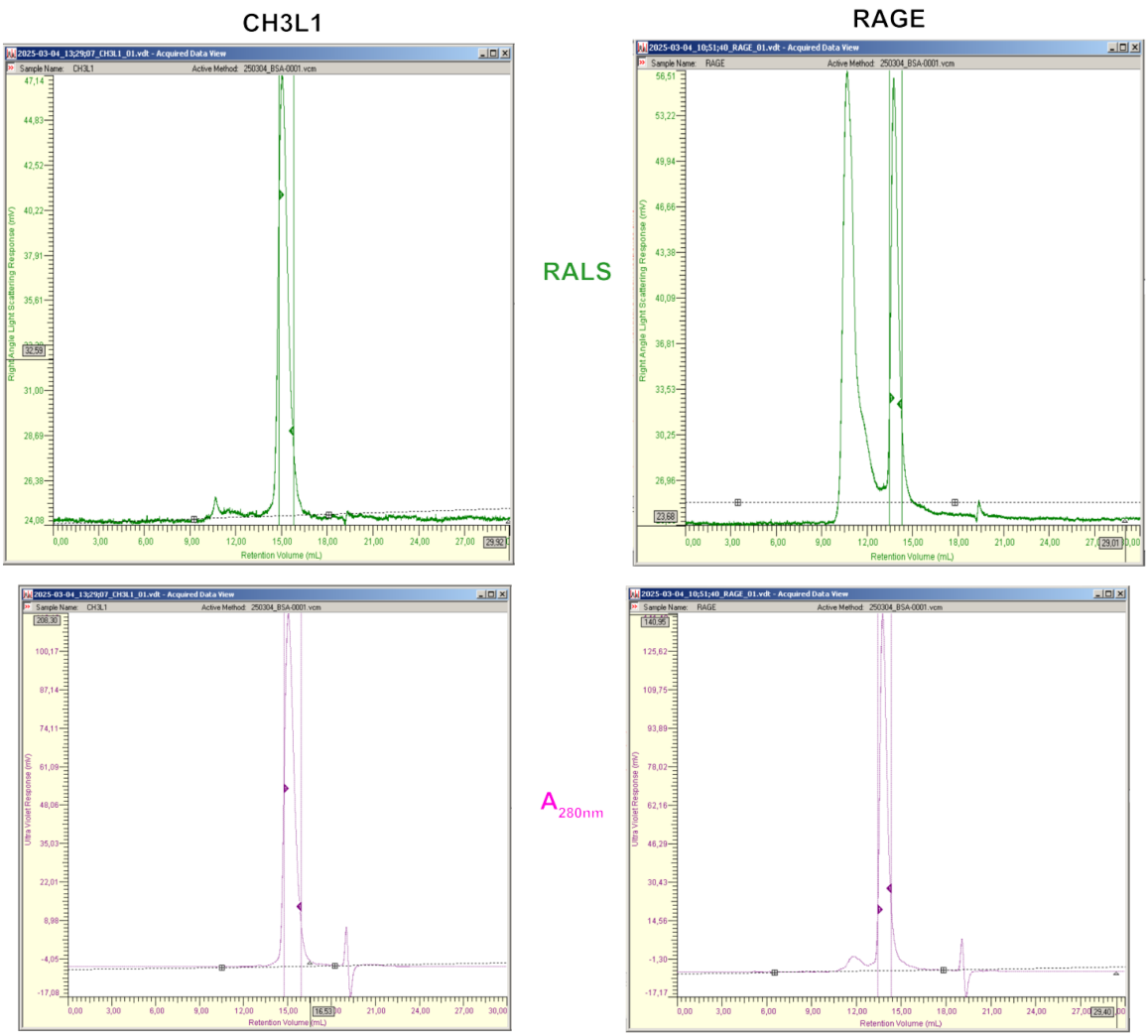
hCHI3L1-His
For hCHI3L1-His (Figure 33, left), a single, symmetrical peak was observed in the RI chromatogram, with a strong corresponding RALS signal co-eluting, indicating a well-defined species without evidence of aggregation or oligomerization. The weight-averaged molecular weight (Mw) calculated for this peak was 50.3 kDa, and the number-averaged molecular weight (Mn) was 44.5 kDa, resulting in a polydispersity index (Mw/Mn) of 1.1. This value is only slightly above 1, confirming a predominantly monomeric species with minimal size heterogeneity. The small offset from the theoretical monomeric mass of 41.7 kDa is attributed to expected N-linked glycosylation, consistent with the protein's eukaryotic expression origin.

hecRAGE-His
Similarly, the hecRAGE-His sample (Figure 33, right) exhibited a narrow RI peak with an overlapping RALS response, supporting the presence of a single, dominant species. Although an earlier-eluting peak was visible in the RALS trace, it was excluded from the integration range and molecular weight calculations. Its earlier elution volume, combined with strong RALS but weak UV signal, suggests it represents a minor population of large, light-scattering aggregates with low protein content. The Mw derived from the selected integration range was 36.6 kDa, closely matching the expected monomeric molecular weight of 35.2 kDa for the extracellular domain of RAGE, again factoring in potential glycosylation and slightly altered hydrodynamic behavior due to its β-sheet–rich Ig-like domains. The polydispersity index (Mw/Mn) was 1.2, slightly higher than that of hCHI3L1-His but still well within the range expected for a monomeric preparation. A polydispersity index of exactly 1.0 would indicate that all molecules in the sample have the same molecular weight, a perfectly monodisperse population; this is almost never the case in practice, especially for proteins.

Overall Comparison
Importantly, no secondary peaks or shoulders were detected in either RI or RALS channels for either protein, ruling out significant populations of dimers or larger aggregates under these buffer conditions. Moreover, the moderate difference between Mw and Mz (in either protein) is typical for glycosylated or slightly anisotropic proteins and does not indicate the presence of major significant aggregates. An anisotropic protein is one that does not have a spherical or symmetric shape, meaning its physical properties vary depending on the direction in which they are measured, as in the case of the elongated RAGE and the irregularly globular CHI3L1.

UV absorbance at 280 nm, while not used for concentration determination in the molecular weight calculation, showed co-eluting signals with RI and RALS for both samples, further validating that the peaks corresponded to folded, aromatic-residue-containing proteins rather than buffer contaminants or unfolded material.

Taken together, the SLS results strongly support that both hCHI3L1-His and hecRAGE-His are monomeric in solution when solubilized in PBS supplemented with 0.5 M NaCl. The good overlap of RI and RALS peaks, low polydispersity indices, and molecular weight estimates consistent with expected glycosylated monomers all confirm the absence of higher-order oligomers or aggregates. The graphs of the measured outputs of the SLS are depicted in Fig. 33.



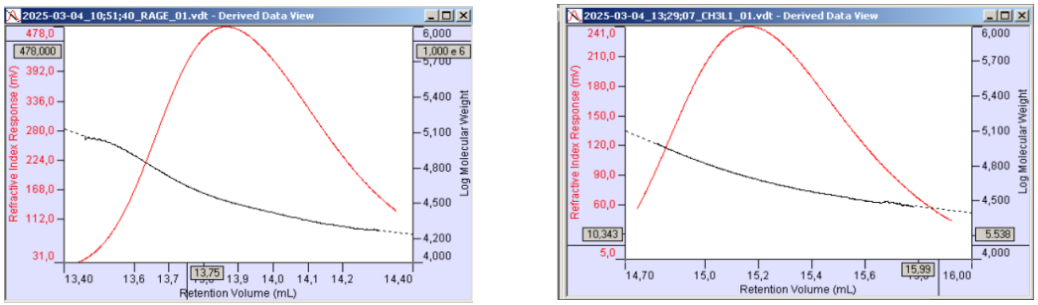


Figure 33. *Static Light Scattering results*.
On the left, CHI3L1, on the right, RAGE analysis. From bottom to top, the RALS curves, the 280 nm absorbance curves and the RI curves for the selected elution fraction, along with the corresponding log molecular weight curves. The samples used were hCH3L1-His at a concentration of about 1 mg/ml and hecRAGE-His at about 1.4 mg/ml. The SLS analysis was run in PBS at 0.5M NaCl (at room temperature). The output values obtained were: MWtheo = 41.7 kDa, MWobs = 50.3 kDa, Mw/Mn = 1.1 for CH3L1; MWtheo = 35.2 kDa, MWobs = 36.6 kDa, Mw/Mn = 1.2; and MWtheo = 35.2 kDa, MWobs = 36.6 kDa, Mw/Mn = 1.2 for RAGE.

3.2.3 N-Glycosylations on RAGE Are Highly Heterogeneous, Even More After Partial Deglycosylation

To assess the glycosylation status of the proteins hCHI3L1-His and hecRAGE-His, enzymatic deglycosylation assays were performed under both denaturing and native conditions with the use of PNGase-F and analyzed by SDS-PAGE, whose results are shown in Fig. 34.

Under denaturing conditions, CHI3L1 showed a clear and homogeneous shift in molecular weight following from around 40 kDa of the untreated condition to about 37 kDa after 1 hour of deglycosylation. This shift is consistent with the efficient removal of N-glycans, suggesting that CHI3L1 carries relatively uniform N-glycosylation patterns that are fully accessible to the enzyme when the protein is unfolded.

In contrast, RAGE displayed a more complex and heterogeneous response to deglycosylation. At baseline (t0), RAGE appeared as a diffuse smear across a range of molecular weights, the heaviest appearing around 60 kDa and the lightest slightly below 50 kDa, indicating a heterogeneous glycosylation profile. The slightly heavier appearance of the protein on gel than its theoretical molecular weight, independently of the glycosylation status, was expected due to previously reported contribution of the Ig-like β-sheet–rich structure of RAGE to reduced SDS binding and incomplete denaturation, increasing its apparent molecular weight by ~9 kDa. [276] After 1 hour of enzymatic treatment, the banding pattern shifted modestly downward but remained heterogeneous and diffuse, rather than collapsing into a sharp, deglycosylated band. This indicates only partial deglycosylation, likely due to structural constraints or variation in glycan types and occupancy across the protein population. The negative control showed no shift, confirming that the changes observed are due to enzymatic activity.

Under native conditions, additional RAGE samples were treated with the deglycosylation enzyme under three different conditions of varying enzyme and GlycoBuffer proportions. In particular, as further detailed in the Materials and Methods section, condition A contained 2

μL of PNGase-F and solutions B and C contained 5 μL of the enzyme. We stopped the reaction at two extended time points, 6 hours and 24 hours, to try to evaluate the maximal capability of deglycosylation of the reaction. Although the overall molecular weight range of RAGE did not exhibit a discrete downward shift under deglycosylating conditions, subtle but progressive changes in its banding pattern were observed. Specifically, one or two additional lower-migrating bands appeared over time, accompanied by a gradual redistribution of signal intensity from the higher molecular weight smear toward lighter bands. This pattern suggests that deglycosylation occurred in an incomplete, stepwise manner, affecting only a subset of glycosylation sites or glycoforms at each time point. The absence of a single, sharp deglycosylated band further confirms that RAGE exists as a diverse population of glycoforms, each potentially bearing different glycan structures or site occupancies, and that these are only partially accessible or susceptible to enzymatic removal under the tested conditions. The results of the reaction on RAGE conducted in parallel at 4 °C are not shown, but they did not exhibit any band shift with respect to the negative control.

These results on RAGE glycosylations show concordance with previous studies in HEK293 cells-expressed hRAGE that reported full occupancy at Asn25 and partial or variable occupancy at Asn81, resulting in multiple glycoforms and corresponding band heterogeneity. [130] These glycosylation states were estimated to contribute approximately 2, 7, or 9 kDa in additional mass relative to the fully unglycosylated species. However, our data newly reveal that the relative proportion of these species shifts progressively over time during deglycosylation with a redistribution of band intensity from heavier to lighter molecular weight forms, indicating a stepwise or sequential removal of specific glycans.

The native conditions for CHI3L1, instead, are shown in Fig. 35. These reactions, conducted both at 37 °C and 4 °C, resulted in a discrete and defined band that exhibited an extremely slight shift with respect to the negative control and to the denaturing conditions. This implies that the N-glycosylations are partly inaccessible to the PNGase-F when the protein is folded, and possibly buried in its structure.

Given the low efficiency with which RAGE N-glycosylations are removed by PNGase-F in the described conditions, the quantities of enzyme eventually needed to achieve complete deglycosylation and test this form in downstream assays have been deemed excessive. Therefore, we have not explored further the deglycosylation reaction and its effects on the two proteins.

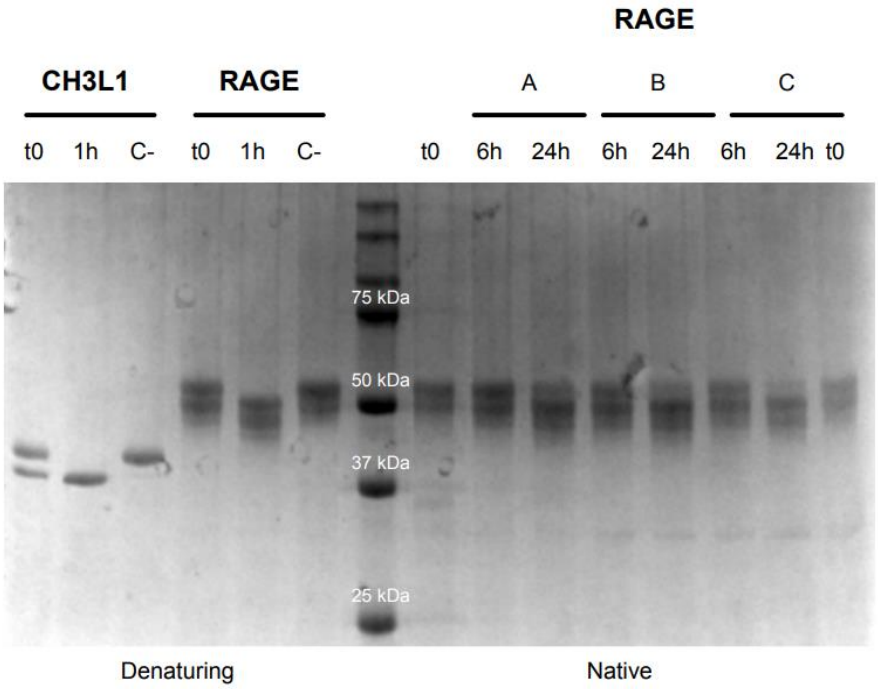


Figure 34. *SDS-Page of the (at least partially) deglycosylated samples of CHI3L1 and RAGE at various time points and conditions*.

On the left side of the gel, to the left of the molecular weight ladder, the results of the denaturing condition reactions for hCHI3L1-His and hec-RAGE-His are shown. For each protein three samples were tested in this condition, namely the negative control (not subjected to PNGase-F reaction), t0, which corresponds to the reaction immediately after addition of the enzyme, and after 1 hour of the reaction.

On the right, the results of the native condition reaction for hec-RAGE.His are shown. Three reaction proportions, each at 6 and 24 hours, were tested: condition A corresponds to 2 µL of PNGase-F and 2 µL of 10x Glycobuffer 2 per 10 µg of protein, B corresponds to 4 µL of enzyme and 2 µL of buffer per 10 µg of protein, and C to 5 µL of enzyme and 4 µL of buffer. Those were compared to t0 in native conditions.

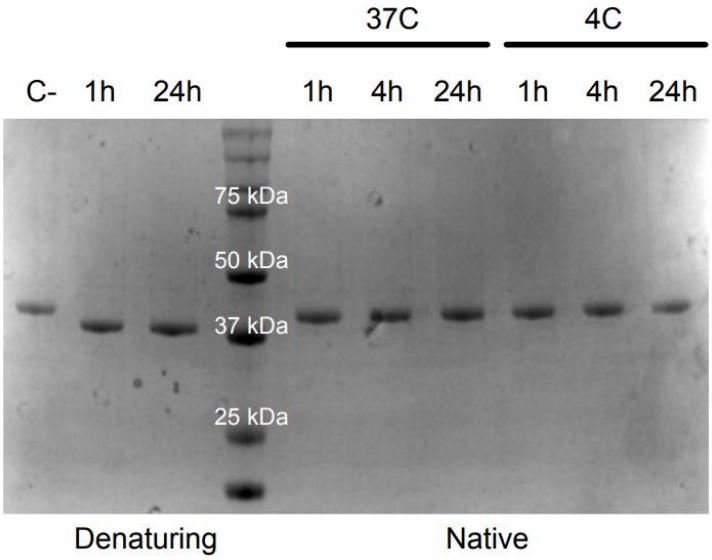


Figure 35. *SDS of the deglycosylated samples of CHI3L1 at 37 °C and 4 °C*.

On the right side of the gel, to the right of the molecular weight ladder, the results of the native condition reaction for hCHI3L1-His are shown. The reaction was conducted at 37 °C (as per the enzyme's manufacturer instructions) and at 4 °C, in order to test for potentially different behaviours of the PNGase-F with the variation of temperature. On the left, the denaturing conditions on the same protein are shown again.

3.3 Protein Complex Characterization

3.3.1 CHI3L1 and RAGE Elute as a Single Peak in the SEC Column

To assess the potential interaction and complex formation between hCHI3L1-His and hecRAGE-His, the two proteins were incubated together at room temperature at a 3:1 molar ratio in favor of the ligand CHI3L1, with the aim of saturating all the possible available receptor binding sites on RAGE. We observed visible precipitation in the mixed solution, which we attributed mainly to CHI3L1 due to the protein's previously noted tendency to aggregate even when handled alone, as well as its intentional molar excess in the mixture, which likely led to unbound and potentially insoluble CHI3L1. However, no further investigations on the actual nature of this precipitation were conducted. The resulting solution was then subjected to preparative large-scale size-exclusion chromatography (SEC) in ÄKTA at 4 °C using a Superdex 200 10/300 GL column equilibrated in 50 mM HEPES pH 8.0, 300 mM NaCl.

The elution profile, shown in Fig. 36, monitored at 280 nm, revealed a single, symmetrical peak eluting between approximately 13.5 mL and 15.5 mL, which roughly corresponds to 75 kDa from the reference protein samples run on the same column (discussed in the Materials and Methods section). This value approximately matches the sum of the theoretical MWs of the two proteins (35.2 kDa for RAGE and 41.7 kDa for CHI3L1) and the absence of multiple, well-resolved peaks suggests that the two proteins co-eluted as a single species. These findings are consistent with the formation of a stable heterocomplex with a 1:1 stoichiometry.

A very small baseline fluctuation between 12 and 13 mL was noted, which probably did not contain any significant species. However, the elution position exactly corresponds to the peak of the hexamer of sRAGE observed by Xu et al., 2014, by using the same type of Superdex column, in the presence of heparan sulfate dodecasaccharides (further explained in the Introduction section about RAGE). [139] The potential oligomerization states of RAGE is in fact something that should be possibly taken into consideration when performing similar experiments on this receptor.

To further investigate the composition of the eluted peak, fractions corresponding to the peak apex and shoulders, specifically from 12.6 mL to 15.6 mL, were collected for SDS-PAGE analysis. This subsequent step aimed to confirm the presence of both CHI3L1 and RAGE within the same elution volume and thereby support the hypothesis of complex formation.

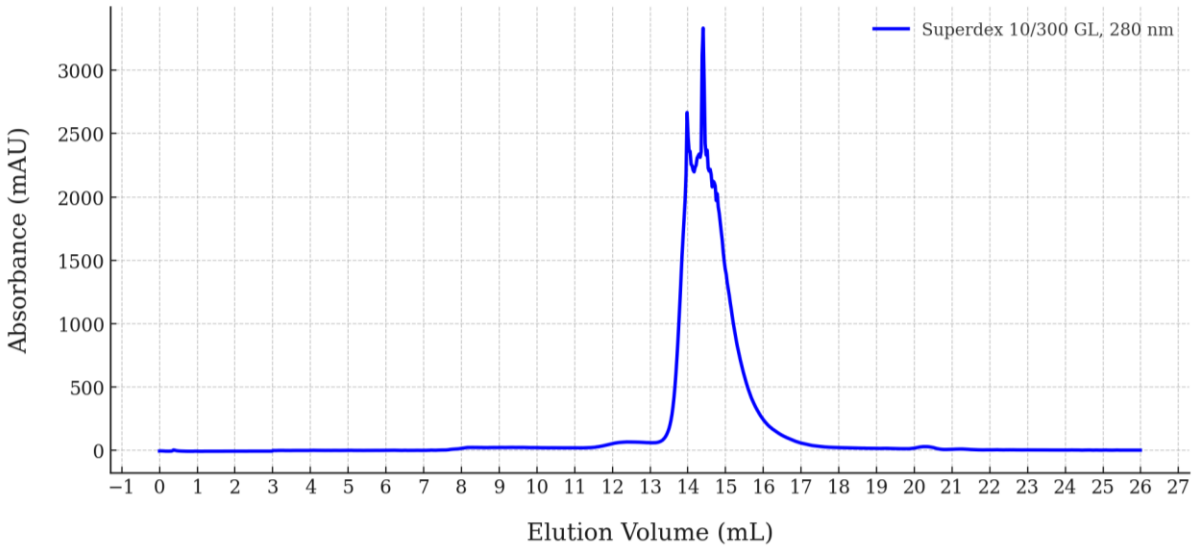


Figure 36. *Preparative SEC of the CHI3L1-RAGE complex on Superdex 200 10/300 GL column.* hCHI3L1-His (0.74 mg/mL, 17.7 µM, ε: 68,090 $M^{-1}cm^{-1}$) and hecRAGE-His (1.14 mg/mL, 32.6 µM, ε: 39,335 $M^{-1}cm^{-1}$) were mixed at a 3:1 molar ratio in favor of CHI3L1 and incubated at room temperature. Following visible precipitation, the sample was centrifuged, supplemented with NaCl to 300 mM, and concentrated to 0.5 mL (7.7 mg/mL total protein, assuming 1:1 complex). The mixture was loaded on a Superdex 200 10/300 GL column equilibrated in 50 mM HEPES pH 8.0, 300 mM NaCl, and eluted at 4 °C in ÄKTAmicro. The absorbance at 280 nm revealed a single, symmetrical peak, consistent with the presence of a stable CHI3L1-RAGE complex at around 75 kDa.

3.3.2 The Gel-filtered Peak Contains Both CHI3L1 and RAGE

To determine the molecular composition of the SEC peak obtained from the CHI3L1-RAGE mixture, SDS-PAGE analysis was performed on the collected fractions spanning the elution range from 12.6 mL to 15.6 mL (B2-C4 in the collecting 96-wells plate). We selected this range in order to include the main peak and any possible contribution of higher-order species eluting at smaller volumes, which could be populating the small baseline fluctuation between 12 and 13 mL of the SEC elution volume.

As shown clearly in the gel in Fig. 37, all fractions corresponding to the SEC peak contained two distinct bands, one at approximately 40 kDa and the other spanning from below 50 kDa to above this level, which align with the expected molecular weights of CHI3L1 and RAGE, respectively. Only one of these bands was alternatively present in the single-lane controls for CHI3L1 or RAGE alone, at the corresponding MW, but they were both visible in the mixed sample control, confirming the co-elution of both proteins.

The simultaneous presence of both proteins in the same fractions strongly supports the hypothesis that the SEC peak corresponds to a CHI3L1-RAGE complex. Although the band intensity estimation is made less accurate by the smearing of RAGE, the two bands intensities appear relatively balanced across most fractions, consistent with a 1:1 stoichiometry as predicted from the theoretical molecular weight of the complex and the SEC elution volume.

These results validate that the major peak obtained from preparative SEC contains both binding partners and is thus a suitable source for downstream analysis of the CHI3L1-RAGE complex, including structural studies.

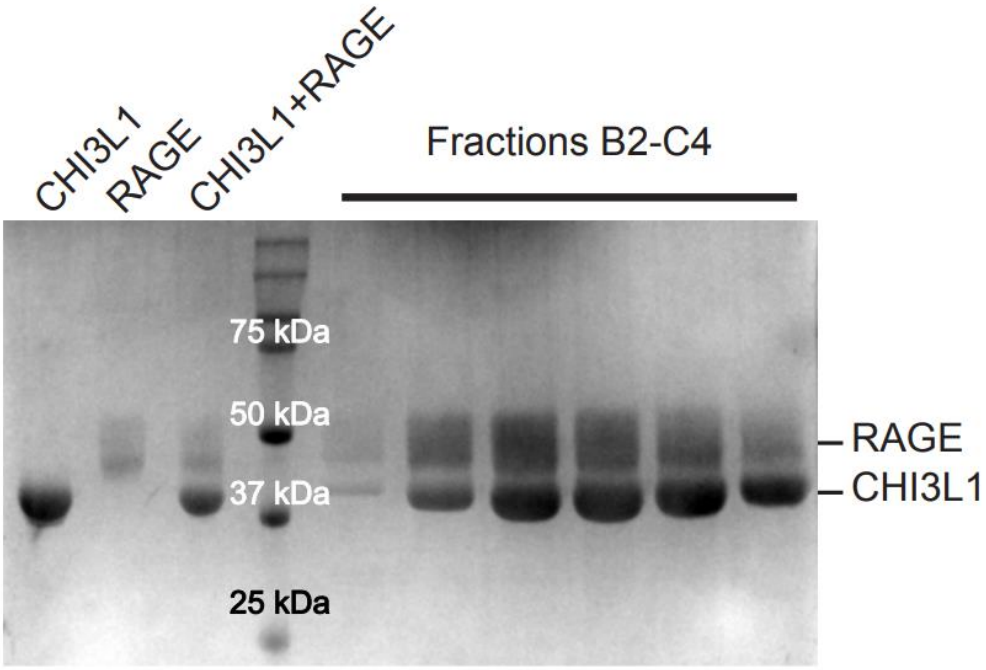


Figure 37. *SDS-PAGE analysis of SEC fractions from CHI3L1-RAGE complex formation*.
On the right side of the gel, to the right of the molecular weight ladder, the fractions B2 to C4 (12.6 mL to 15.6 mL) from the SEC peak analyzed by SDS-PAGE reveal the presence of two distinct bands corresponding to CHI3L1 and RAGE, confirming co-elution of both proteins. On the left, control samples containing CHI3L1 alone, RAGE alone, or a mixture of both proteins at concentrations comparable to those in the SEC experiment are shown, allowing identification of the respective bands.

3.3.3 Progressive Dilution of the CHI3L1-RAGE Complex Determines its Step-wise Disassembly in a Concentration-dependent Fashion

To assess the stability of the CHI3L1-RAGE complex and the molecular weight of the protein species contained in it, mass photometry was performed under increasingly dilute conditions on the protein mixture previously purified via preparative SEC (as described above). The results from the mass photometer are shown in Fig. 38.

When subjected to mass photometry at 80 nM total protein concentration, the dominant species was centered around 124 kDa, suggesting the formation of a higher-order assembly, potentially consistent with a 2:1 CHI3L1:RAGE complex. With a standard deviation of 31 kDa, this peak was relatively broad, indicating heterogeneity in the population and possibly the coexistence of other oligomeric states. Upon dilution to 40 nM, the major population shifted to ~90 kDa, consistent with a 1:1 CHI3L1-RAGE heterodimer. Notably, a shoulder at higher mass near 120 kDa was still detectable at this concentration, suggesting an equilibrium between different stoichiometries. As the complex was further diluted to 10 nM and 5 nM, the distributions progressively shifted toward lower molecular weights, with dominant species at 47 kDa and 43 kDa, respectively, values matching those of the monomeric CHI3L1 and free

RAGE. However, the limit of detection of the instrument is around 35-40 kDa, implying a lower reliability for the measures close to it.

This progressive redistribution from complex to monomeric species strongly supports a concentration-dependent dissociation mechanism and the non-covalent, reversible nature of the interaction. The abrupt transition between 40 nM and 10 nM aligns well with the dissociation constant reported by Darwich et al., 2020 (Kd ≈ 16.65 nM), since lowering the concentration below that value determines the complete disassembly of the complex. [93]

Interestingly, while SEC analysis of the same complex yielded a single peak with an apparent molecular weight near 75 kDa, suggestive of a predominant 1:1 complex, mass photometry reveals a more complex and dynamic picture. This discrepancy may arise from differences in the detection principles of the two methods: SEC separates species based on hydrodynamic volume and may not fully resolve closely sized oligomers, especially if they are in rapid equilibrium, while mass photometry directly measures molecular mass with high sensitivity to transient or low-abundance states. Moreover, some contribution from homotypic interactions, such as CHI3L1 or RAGE homo-dimerization,, which have been previously reported, cannot be excluded, particularly at higher concentrations. [200, 126, 163]

Such dynamic behavior emerging from the mass photometry analysis may be physiologically relevant, as it implies that the ligand-receptor complex stability can be modulated not only by affinity but also by stoichiometry and concentration-dependent assembly.

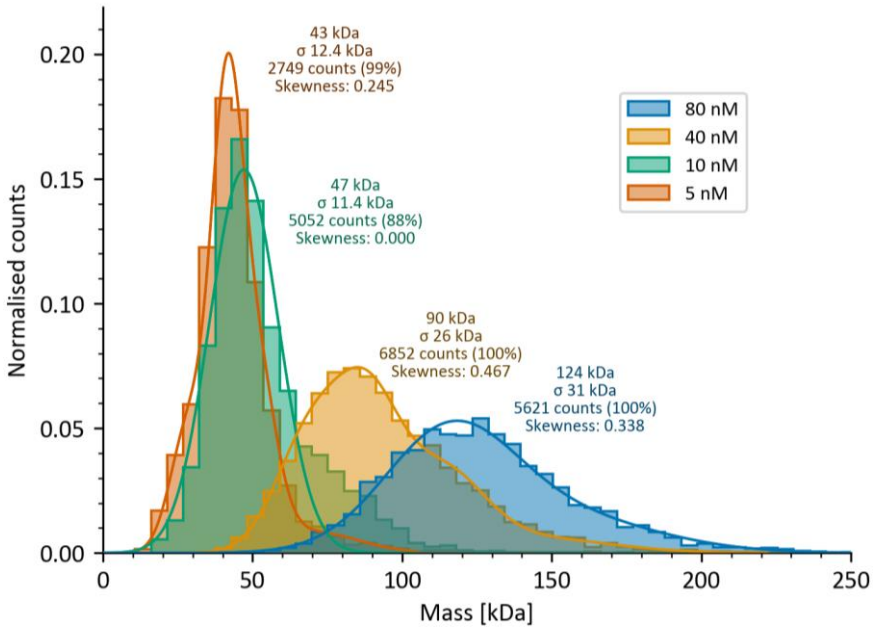


Figure 38. *Mass photometry reveals concentration-dependent disassembly of the CHI3L1–RAGE complex.*

Mass photometry analysis of the CHI3L1-RAGE complex with the Refeyn TwoMP instrument, previously purified via preparative SEC, was performed at decreasing total protein concentrations (80 nM to 5 nM). A higher-order species (~124 kDa) dominates at 80 nM, shifting to a 1:1 complex (~90 kDa) at 40 nM, and to monomers (~43–47 kDa) at 10 nM and 5 nM, indicating reversible, concentration-dependent disassembly.

3.4 Crystallography Studies

3.4.1 The Crystal Diffraction Spoke, but not Loud Enough to Be Understood

A first set of crystallization screens was attempted at SLAC, Stanford Synchrotron Radiation Lightsource (SSRL) facility. We screened for a total of 768 conditions, belonging to 4 different commercial sets in two protein buffers (further detailed in the Materials and Methods section). The hCHI3L1-His–hec-RAGE-His complex was prepared as described in the Materials and Methods section.

Protein crystals, which were numerous, spiky and small, appeared after some weeks in 4 of those tested conditions, with the most promising crystals growing in the well F10 of the Tris Index HT plate. This condition consisted of 0.2 M sodium chloride, 0.1 M BIS-TRIS pH 5.5, and 25% PEG 3350 incubated at 16 °C, with the protein complex mixture prepared in 25 mM Tris-HCl, pH 8.0, 350 mM NaCl. The crystal was mounted on the SSRL beamline BL12-2, where a complete diffraction dataset was successfully collected. A snapshot of the crystal mounted and ready to be diffracted is shown in Fig. 39a.

The diffraction image (Fig. 39b) showed clear Bragg reflections, which are bright spots that appear where X-rays diffracted by the crystal lattice constructively interfere. These extended to a modest resolution, with a well-defined central beam stop and visible resolution rings. Despite collecting 1800 frames over 360° of rotation, the final resolution did not extend into high-resolution shells, and diffraction quality dropped significantly beyond ~5–6 Å, as confirmed by the resolution plot shown in Fig. 39c. It is also worth noting that the R-factors from the dataset (R-merge = 0.45, R-meas = 0.472, R-pim = 0.128) indicated high redundancy but substantial noise, consistent with the visual observation of weak high-resolution spots. While the elevated R-merge and R-meas indicate significant variation among repeated measurements, the lower R-pim suggests that the high level of redundancy helped average out this random error, resulting in a dataset that, though noisy, retained a reasonable degree of internal consistency.

Indexing analysis determined the space group ($P\,2_12_12_1$), the unit cell dimensions (a (Å) 109.6 Å, b 121.16 Å, c 130.54 Å) and volume (~1.73×10^6 Å^3). While structure solution was not possible due to insufficient resolution, we compared these parameters to known structures. Interestingly, the cell dimensions and volume closely resemble those of CHI3L1 crystallized in PDB entry 1NWR, which has a similar orthorhombic cell and nearly identical volume (~1.74×10^6 Å^3). By contrast, known crystal forms of RAGE (PDBs 3CJJ, which contains only the VC1 domains, and 4LP5, containing all the three extracellular domains VC1C2) exhibit significantly smaller unit cells and different space groups, suggesting that RAGE is unlikely to be the dominant or ordered component of this crystal.

While no direct structural assignment can be made from the current data, these crystallographic comparisons suggest that CHI3L1 may have preferentially crystallized alone under these conditions. However, no evidence in this regard was available for us to further speculate on the nature of those crystals.

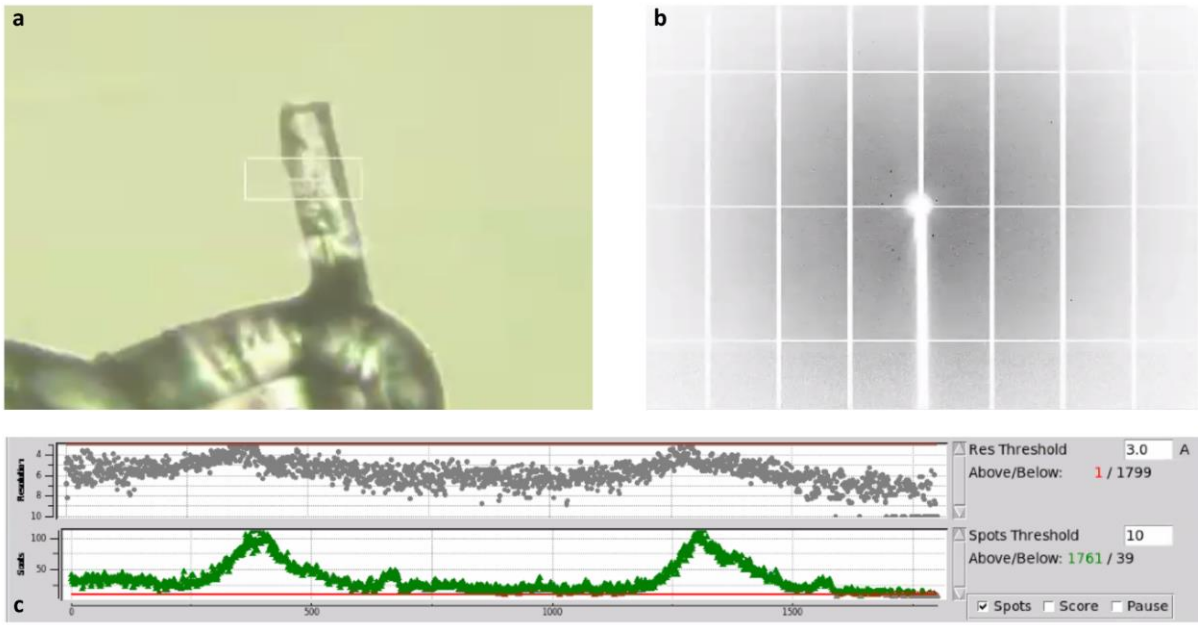


Figure 39a. *Snapshot of the crystal mounted on the goniometer at the beamline.* The white box represents the shape and position of the X-ray beam itself that will soon irradiate the crystal.

Figure 39b. *Diffraction image showing Bragg reflections distributed across resolution shells*. The image shows few discrete Bragg reflections (black points) arranged around the central beam stop, indicating the presence of an ordered crystal lattice. However, the reflections do not extend to the outermost resolution rings, suggesting that while the crystal diffracted well enough to collect a full dataset, the maximum resolution achieved was modest. The resolution is inversely proportional to the distance from the centre, the farther the points extend, the higher the resolution. The grid on the image serves as a reference for spatial frequency, and the absence of spots beyond the mid-range confirms that the data lacks high-resolution detail.

Fig. 39c. *Spot Resolution and Detection Over Data Collection*. The top scatter plot tracks the resolution of detected spots over the course of data collection. Most reflections fall above the red resolution threshold line (set at 3.0 Å), indicating that the majority of the data lies in the low-resolution range and that high-resolution reflections were scarce. The lower plot shows the number of Bragg spots detected per frame throughout the full 360° rotation. The resulting bell-shaped distribution confirms that the crystal diffracted consistently across the dataset, with no major radiation damage or mechanical issues during collection. However, the moderate spot count across frames and the absence of strong high-resolution data collectively indicate that while the dataset is complete, its overall resolution is insufficient for confident structure determination.

3.4.2 CHI3L1 Crystallizes in the PEGRx HT E7 Solution and Other 62 Conditions

Following the initial crystallization attempt at SLAC, a second screening campaign was conducted at the European Institute of Oncology (IEO) in Milan to optimize crystal quality and try to crystallize the CHI3L1-RAGE complex.

For this purpose, other than characterizing the protein complex, we performed the preparative steps described in the previous subsection. As mentioned, the fractions B2 to C4 from the gel-filtration peak, shown to contain both hCHI3L1-His and hecRAGE-His, were pooled, half

concentrated to an estimated value of 13 mg/ml and the other half to 6.5 mg/mL, and both solutions were used to set up crystallization trials.

Two commercial screens that had yielded initial hits at SLAC (PEGRx HT and Index HT) were selected, and plates were incubated in parallel at 4 °C and 20 °C. After several weeks of incubation, crystals had appeared under multiple conditions, detailed in Table 6. A total of 63 crystals from 11 different wells were harvested and shipped to the European Synchrotron Radiation Facility (ESRF) in Grenoble for X-ray diffraction analysis.

All datasets collected at ESRF revealed a single and consistent molecular identity: CHI3L1 alone. This finding was in line with previous suggestions that CHI3L1 may preferentially crystallize in isolation, even when co-purified with RAGE. In Table 7 we show the collected data from one of the representative datasets, in particular derived from a crystal grown in the condition E7 of the PEGRx HT screen, which is composed of 0.2 M ammonium acetate, 0.1 M sodium citrate tribasic dihydrate pH 5.5, 24% PEG 400, incubated at 4 °C. It was collected at 0.9655 Å wavelength, yielded crystals diffracting in the orthorhombic space group $I\,2_12_12_1$, with unit cell parameters a = 109.3 Å, b = 120.8 Å, c = 135.9 Å, approximately matching the cell dimensions of the crystal diffracted at SLAC.

The resulting representative CHI3L1 structure derived from this dataset is shown in Fig. 40a–c. It reveals the expected β-sandwich fold with excellent geometry and well-resolved side chains. In Fig. 40d, this structure is superposed with a previously published CHI3L1 structure (PDB ID: 8R42), which served as the search model for molecular replacement. This illustrates the structural consistency across preparations. Notably, the conserved N-glycosylation site and a bound inhibitor present, inside the CHI3L1 carbohydrate binding groove, in the 8R42 reference are highlighted in light blue and yellow, respectively. [283] Their absence in the new structure further supports that the protein crystallized here is unmodified CHI3L1, derived directly from the gel-filtration of the CHI3L1-RAGE complex, and is unlikely to be part of a stable heterocomplex under the conditions tested.

This finding may reflect either spontaneous dissociation of the complex, selective lattice incorporation of CHI3L1, or higher flexibility in RAGE that prevents its inclusion in the crystal lattice.

0,2 M Ammonium sulfate, 0,1 M BIS-TRIS pH 6,5, 25% w/v Polyethylene glycol 3,350
0,2 M Ammonium acetate, 0.1 M BIS-TRIS pH 6.5, 25 % w/v Polyethylene glycol 3,350
0.1 M MES monohydrate pH 6.0, 20% v/v Jeffamine® M-600® pH 7.0
0.1 M Tris pH 8.0, 30% v/v Jeffamine® M-600® pH 7.0
10% v/v 2-Propanol, 0.1 M Sodium citrate tribasic dihydrate pH 5.0, 26% v/v Polyethylene glycol 400
0.2 M Ammonium acetate, 0.1 M Sodium citrate tribasic dihydrate pH 5.5, 24% v/v Polyethylene glycol 400

0.2 M Potassium sodium tartrate tetrahydrate, 0.1 M BIS-TRIS pH 6.5, 10% w/v Polyethylene glycol 10,000
0.1 M Ammonium acetate, 0.1 M BIS-TRIS pH 5.5, 17% w/v Polyethylene glycol 10,000
0.1 M Sodium citrate tribasic dihydrate pH 5.0, 10% w/v Polyethylene glycol 6,000 (4 °C)
0.1 M Sodium citrate tribasic dihydrate pH 5.0, 10% w/v Polyethylene glycol 6,000 (20 °C)
0.1 M Sodium citrate tribasic dihydrate pH 5.5, 16% w/v Polyethylene glycol 8,000
0.1 M Sodium citrate tribasic dihydrate pH 5.0, 18% w/v Polyethylene glycol 20,000

Table 6. *The 11 different conditions in which the 63 crystals appeared.*
These conditions come from either PEGRx HT or Index HT screens. By comparison, CHI3L1 was previously crystallized alone in various conditions, among which there is a sodium citrate buffer pH 5.1 with PEG 8,000 (PDB ID: 1NWR) closely resembling some of these crystal-yielding solutions. [194] Notably, the starting protein solution was suspended in a buffer containing 1M NaCl (in 10 mM BES, pH 7.5). This elevated concentration of salt is generally considered unfavorable for crystal growing; however, it further confirms our observation that CHI3L1 displays a strong preference for buffers containing high salt concentrations, as previously mentioned.

Data collection and refinement statistics	
Space group	I 21 21 21
Wavelength (Å)	0.9655
Cell dimensions	
a, b, c (Å)	109.3, 120.8, 135.9
α, β, γ (°)	90.0, 90.0, 90.0
Resolution (Å)	90.3 (2.5)*
Rmerge	0.19 (1.45)*
CC1/2	1 (0.3)*
I /σI	11.3 (1.3)*
Completeness (%)	97.9 (85.4)*
Number of reflections	34518 (1764)*

*Values in parentheses are for the highest-resolution shell.

Table 7. *Data collection and refinement statistics for the representative crystal appeared in the E7 condition of the PEGRx HT screen*.
The dataset displays high completeness, strong signal-to-noise, and reliable refinement values.

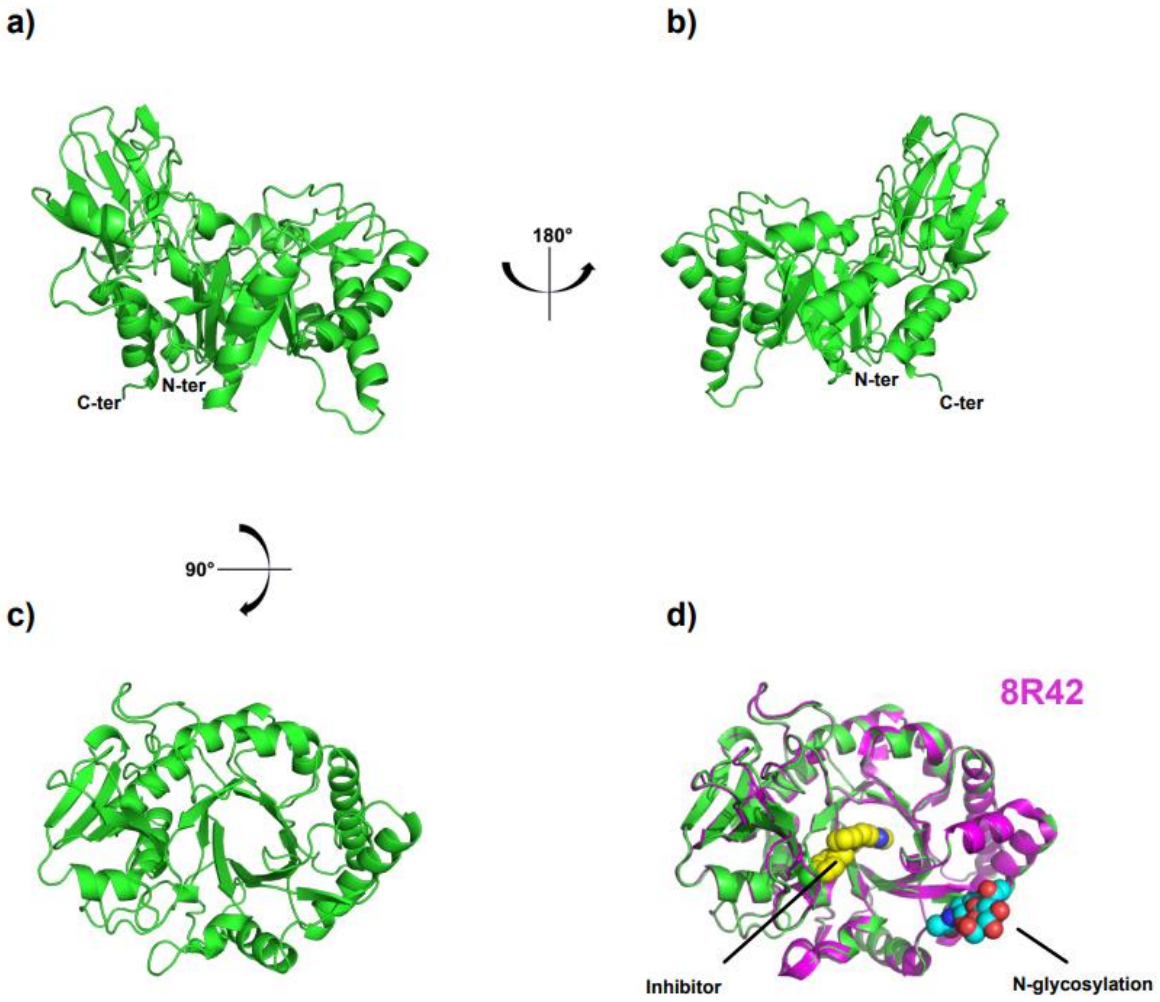


Figure 40a-c. *Cartoon representation of the crystallographic structure of CH3L1 solved from a dataset derived from crystals grown at IEO.*
The condition is the E7 of the PEGRx HT screen incubated at 4 °C.
Figure 40d. *Superposition of the structure as in (c) with the CH3L1 structure from PDB 8R42.*
The reference structure is shown in magenta, while the N-glycosylation of the protein and the bound inhibitor are highlighted in light blue and yellow, respectively.

3.5 Prediction of the CHI3L1-RAGE Interface Using AlphaFold Multimer v2.3 in ColabFold

In order to predict how CHI3L1 and RAGE interact, we first attempted to use the newest model through its web server, AlphaFold 3, which yielded extremely low score results for the CHI3L1-RAGE (VC1 domains) complex prediction. [284]
Therefore, we resorted to the ColabFold notebooks, accessible from their github. [285] All the predictions were carried out using *AlphaFold2_advanced_v2,* an experimental notebook including a pre-analysis, serial seed testing, and PDB number-guided template retrieval, other than direct access to the Python code to run the prediction. In this notebook, the latest model for complex prediction, and the one that we used, is AlphaFold-Multimer v2.3.

The following is a methodological Result section where we will demonstrate step-by-step our approach to master the prediction capabilities of the model. The section will culminate with the illustration of the final predicted interaction between the two proteins, CHI3L1 and RAGE (VC1 domains). The choice of this shortened input sequence for RAGE (specified in the Material and Methods section) was based on the literature finding that the VC1 domain is tightly packed together while the C2 domain is linked by a flexible loop that causes the two parts to move about with respect to each other. [126] We reasoned that this characteristic could hinder the model abilities to predict the binding, given that high mobility of the domains affects the accuracy of the prediction. Moreover, the V domain was also reported to be the one where most of RAGE ligands bind, meaning that this strategy could help us to globally maximize the chances of obtaining a high-fidelity predicted structure. [124]

3.5.1 Parameters Definition

Before running the prediction on the ColabFold notebook there are a number of parameters that can be tuned by the user, differently from the server web interface of AlphaFold 3, where the input freedom is very limited. The following is a list of parameters that we thoroughly analyzed in order to understand how to tune them in the most efficient way.

The MSA is the multiple sequence alignment with respect to the query sequence and it possesses several tunable parameters. The method we selected was the default one, the "mmseqs2", which is a fast sequence search tool to find homologous sequences to the query. The "pair_mode", instead, determines how homologous sequences are treated when constructing the multiple sequence alignment for protein complex modeling: unpaired mode treats each protein separately, allowing AlphaFold to use the widest range of evolutionary homologs without enforcing evolutionary coupling between the chains; paired mode forces homologs of the two proteins to be included only when they co-exist in the same genome, capturing evolutionary interactions but potentially reducing MSA diversity; paired_unpaired mode combines both approaches, allowing the model to use some co-evolutionary signals while maintaining a larger dataset of homologs.

In order to visually evaluate the effect of the pair mode, we can use the Coevolution Matrix. This matrix captures coevolutionary relationships between pairs of sequence positions, computed from the MSA using methods like covariance analysis. It serves as an input feature in AlphaFold2, helping predict which residues are likely close in space, and is extended in multimer to include all residues across chains. The matrices resulting from the three different pair modes are displayed in Fig. 41. As confirmed by the co-evolutionary analyses, RAGE, a pattern recognition receptor (PRR) with three globular extracellular domains, the outermost of which is an Ig-like domain homologous to the variable region of antibodies likely exhibits weak co-evolution both within its own residues and with those of individual binding partners. [119-121] This implies that RAGE maintains a structurally flexible binding interface, enabling recognition of diverse ligands.

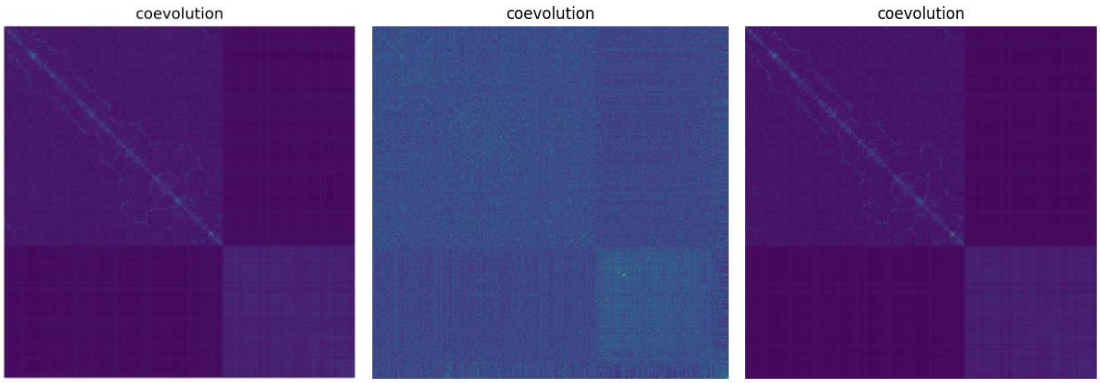


Figure 41. *Coevolution Matrix pre-analysis when using "unpaired_paired" (left) versus "paired" only (middle) versus "unpaired" only (right).*
CHI3L1 residues are represented in the upper left quadrant of each matrix, while RAGE-VC1 ones in the lower right quadrant; the quadrants in common represent the coevolution between the two proteins. The matrix is color coded with light color meaning strong signal and dark color weak. The coevolution signal is visible for CHI3L1, instead it is not strong at all for RAGE-VC1, and becomes even weaker when using paired mode. Even though paired mode is designed for complex co-evolution, it still affects the MSA of individual proteins because it filters the alignments. Initially, we chose the mode unpaired_paired in order to possibly account for both the approaches and considered that the coevolution signal did not show significant differences, apparently.

Then, there are the filtering options for the MSA sequences. Among these, the first is the parameter "cov", which stands for coverage threshold and determines the minimum percentage of a sequence's residues that must align with the query sequence for it to be included in the MSA. A higher cov (e.g., 75 or 90) ensures only sequences with more complete alignments are included, reducing noise from partial matches but potentially limiting the number of sequences. We initially chose cov as the default that was 75, meaning that only sequences covering at least 75% of the query protein were kept. Also the sequence coverage can be visualized for each cov parameter of choice, and in Fig. 42 we show the corresponding graph for cov 75.

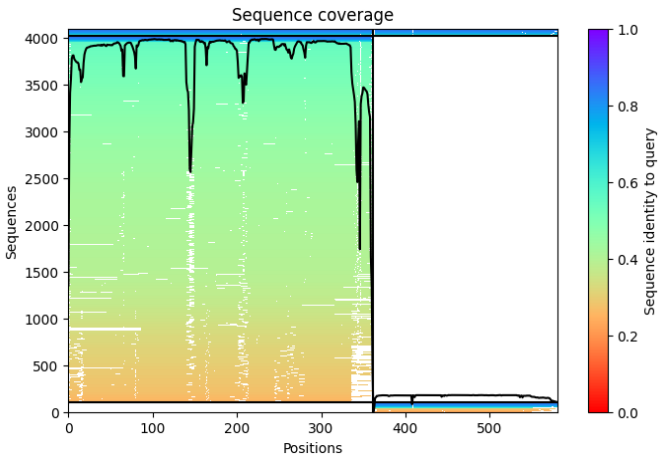


Figure 42. Sequence coverage graph with cov 75%.
CHI3L1 residues are represented in the left part of the graph, while RAGE-VC1 ones in the right part. All the sequences aligned by the MSA are shown stacked in the graph on the corresponding aligned residues of the query, and they are color coded based on the whole sequence identity to it. The black

line indicates the selected cov threshold. It is clear that CHI3L1 shows a good coverage over all the sequence, instead RAGE is almost not covered at all, further confirming our previous observation about its co-evolution and conservation.

The second filtering parameter is the "id", standing for identity threshold. It sets the maximum allowed sequence identity between any two sequences in the MSA. A lower id (e.g., 90) prevents overly similar sequences from dominating the MSA, promoting diversity to capture broader evolutionary relationships. Also in this case, we initially selected the default id of 90, which means that sequences that are >90% identical will be removed. Lastly, there is the "qid", the query identity threshold. It specifies the minimum percentage of identical residues a sequence must share with the query sequence to be included. A higher qid (e.g., 20 or 30) ensures included sequences are more similar to the query, improving relevance but possibly reducing evolutionary depth. The default qid was 0, meaning that there is no minimum threshold and all the sequences previously filtered are included.

Another relevant aspect to consider is the template**.** This refers to previously solved crystal structures similar to the query sequence and acts as a reference, nudging AlphaFold2's prediction to resemble the structure provided. We could finetune various parameters. First, the "rm_template_seq", which ensures that the prediction is more data-driven rather than just copying the template. We set it as True, since we observed that setting it as False significantly worsened the accuracy. Then, "propagate_to_copies", which applies changes to all identical chains in a multimer and is useful when predicting homomers (identical multimers). Propagation to copies was set as False because we did not want to model homomers. There is also "do_not_align" that prevents a sequence from being aligned in the MSA. We set it as False because, empirically, setting it True either did not change the results or slightly worsened them. The resulting template features extracted with this modality are depicted in Fig. 43.

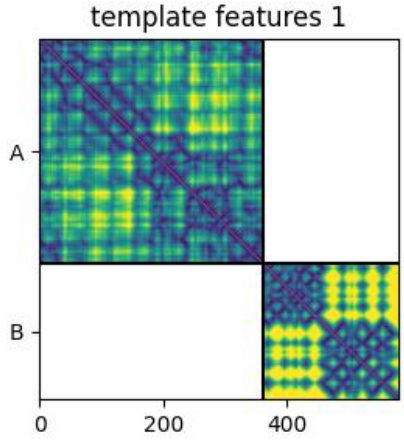


Figure 43. *Template features matrix for CHI3L1 residues (upper left) and RAGE-VC1 ones (lower right)*.

The matrix is color coded with dark color meaning strong signal and light color weak. We can observe that CHI3L1 displays strong features approximately across the whole protein, while RAGE-VC1 shows a clear distinction between the two domains. The V domain is represented in the left upper quadrant of the RAGE quadrant, while C1 in the lower right: the features inside each domain are strong but the signal for the interacting residues between the two domains is relatively weak, even though a pattern can still be appreciated.

An additional aspect is the MSA depth**.** It refers to the number of homologous sequences (evolutionary relatives) included in the Multiple Sequence Alignment. Since our co-evolutionary signal appears to be weak, at least for RAGE, AlphaFold tends to rely more on template structures for structural guidance, considering also that both CHI3L1 and RAGE-VC1 have been singularly crystallized several times. However, maximizing MSA depth is still beneficial because even weak evolutionary constraints can refine flexible regions, improve interface modeling, and enhance confidence estimates. Unlike cases where strong MSAs override templates, in this scenario, a deeper MSA provides additional refinement without compromising the structural guidance provided by the templates. For this reason, we set "num_msa" to 512 and "num_extra_msa" to 4096, the maximum for both. The number of msa controls the number of sequences kept in the primary MSA and these are directly used in structure prediction and fed into AlphaFold's Evoformer module; the number of extra msa specifies the number of additional (extra) sequences kept in the MSA for statistical signal extraction and these do not go into the attention layers but still help AlphaFold understand evolutionary constraints. To control how the model interprets the MSA sequences, we could modify the "cluster profile", too. Clustering the MSA by sequence similarity can help elicit a range of conformations, especially useful for predicting different conformational states of a protein. [286]

When using ColabFold Multimer v2.3 there is the possibility to choose among 5 models with slight variations in architecture and training data. We started by selecting the model 1, but will test differences in the models later on. The parameters of the model that can be tuned are: "use_mlm", which stands for masked language model, hides some amino acids in the MSA preventing overfitting to MSA and strengthening evolutionary learning when MSA is weak or diverse and we set it as True; "use_dropout", which refers to dropout regularization, randomly disables some neurons of the model reducing overfitting and improving structural diversity when running multiple seeds or modeling flexible proteins. The dropout regularization can drive AlphaFold2 to sample alternative conformations and/or different structure predictions. We set it as True.

AlphaFold2 uses random seeds to initialise its structure predictions. By changing the seed, you can sometimes guide AlphaFold2 towards a correct prediction. From a computational standpoint, using different random seeds introduces variability in the structure prediction outcomes. Normally, high-confidence parts of the structure will converge to the same conformation regardless of the random seed used. However, low-confidence portions may vary substantially. For example, when there are few to no sequences in the MSA, AlphaFold2 will struggle to predict the protein structure with high confidence. However, changing the random seeds will sometimes allow AlphaFold 2 to predict the structure despite this hurdle, although this is not guaranteed. Adding more recycles for the prediction usually also helps in this situation. For reference, AlphaFold 3 has a 60% failure rate for antibody and nanobody docking when using a single seed compared to the higher success rates when using 1000 seeds, as shown in Fig. 44. *AlphaFold2_advanced_v2* is the only notebook that allowed us to test many seeds contemporarily. [287] We set the number of seeds (num_seeds) as 128, the maximum, for each screening test. This significantly increased the computational cost but also dramatically augmented our chances of finding an accurate prediction.

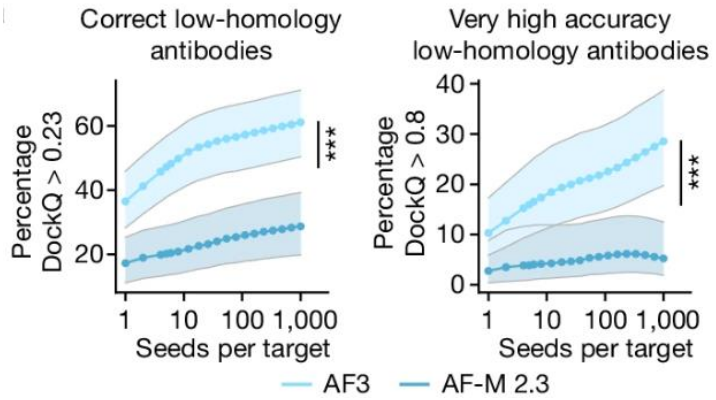


Figure 44. *Antibody prediction accuracy improves with the number of model seeds*.
Each point shows the average accuracy over 1,000 random samples (with replacement) drawn from 1,200 model seeds. This clearly shows that the accuracy increases with the number of seeds.
Image reproduced from Abramson et al., 2024. [288] Copyright 2025 Springer Nature Limited.

3.5.2 An Efficient Approach: Optimizing Parameters Across 128 Seeds

Local Optimization

From the preliminary analysis, both manual and automated, described in the Appendix we derived the following conclusions: some parameters are more impactful than others across all seeds; scanning at full depth all possible combinations of parameters and seeds is not computationally feasible. In other words, we cannot explore the full solution space of the models.

Here comes the need for an alternative approach, one that is more compute-efficient and yet systematic, by optimizing the model parameter-wise over all the 128 seeds. Ideally, these parameters should not show or have limited interaction between each other.

At the same time, the stochasticity of the seeds allows us to get a more accurate representation of the model performance by averaging its scores over 128 predictions.

Therefore, the following analysis aims at achieving a local optimum in the solution space of all possible models. This type of strategy can be visualized in Fig. 45.

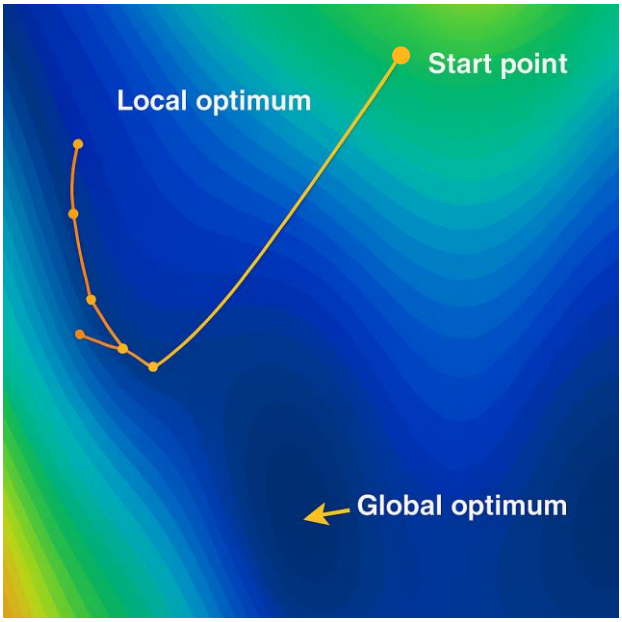


Figure 45. *Visual map of the strategy that enables to find the local optimum in the solution space.* Local optimization allows one to reach a good enough solution, even though not necessarily the best one, without exploring the full solution-space, which would be either too expensive or too elaborate.

Initializing a Fixed Set of Parameters

First, we identified a set of parameters that we could confidently say work better at specific values for our protein complex, when using AlphaFold multimer v.2.3 *model 1*. The notebook supplies users with other four models slightly varying in architecture or training strategy.

The need for a fixed set of parameters is explained by compute limits: testing every possible parameter value in the notebook would mean testing more than 40 different setups, while testing all of their combinations (even at low depth) would imply more than 7 millions runs times 128 seeds.

In order to select the fixed parameters, we relied on what we learned from the preliminary screens about the properties of our proteins relevant for AlphaFold predictions. This fixed set of parameters involved pair mode, template mode, cluster profile, do not align, remove template sequence, and number of msa and extra msa. For pair mode, *unpaired* was selected, given that our protein complex seems to be poorly co-evolved with a paired MSA, and we generally observed a worse result when using unpaired_paired as compared to unpaired only. This comparison is illustrated in Fig. 46.

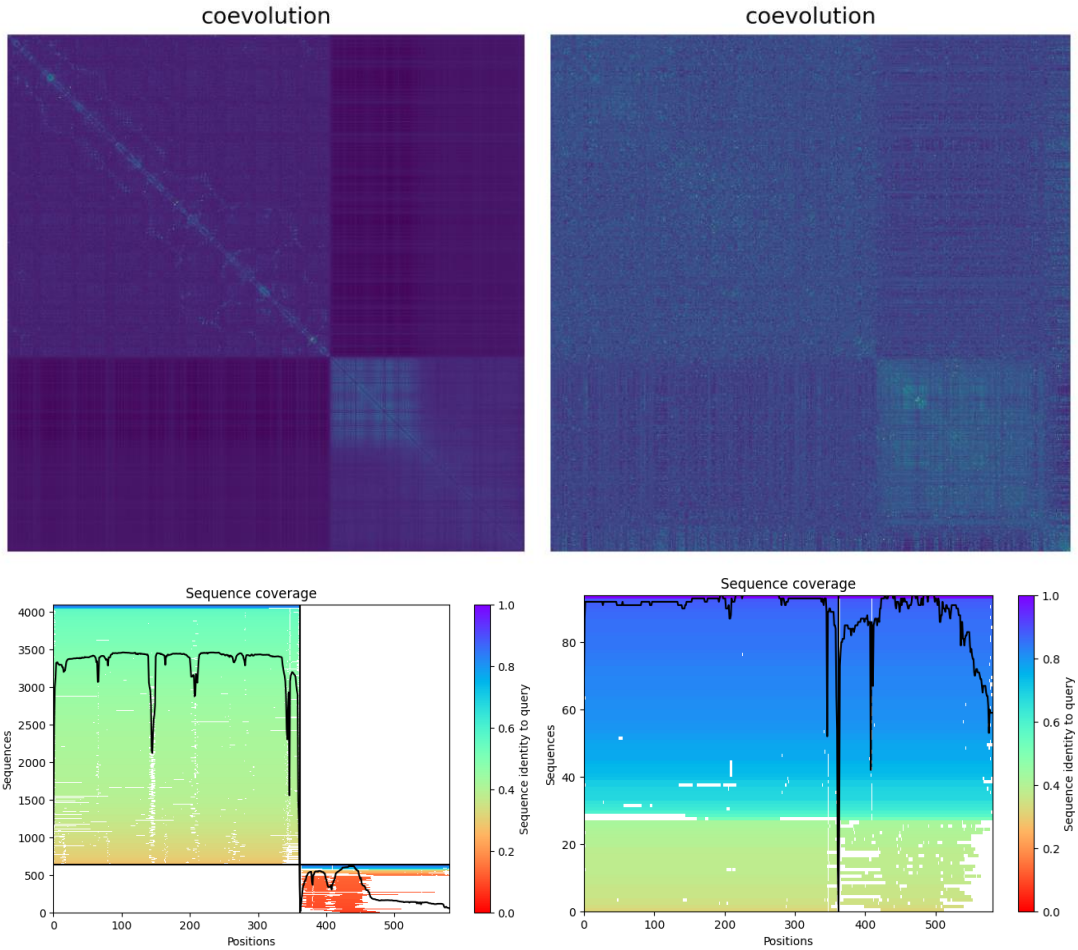


Figure 46. *Coevolution matrices and sequence coverage for unpaired (left) and paired MSA (right), both with coverage set at 25% (cov 25)*.

As highlighted in the Appendix when we tested cov 25, the unpaired matrix now shows a much more informative pattern with respect to when cov is set at 75. First of all, CHI3L1 signal is enriched with a clearer pattern around the diagonal, and RAGE V domain is now easily recognizable in the upper left quadrant of the RAGE quadrant. Moreover, we can observe a clear distinction in the signal coming interacting residues between the two proteins, where the V and C1 domains display different patterns of co-evolution with CHI3L1. The same cannot be said for the paired alignment, whose coevolution matrix is still weaker and noisier overall. This phenomenon can be explained by a low sequence number (<100) as compared to unpaired MSA (>4000 for CHI3L1 and >500 for RAGE-VC1).

For the template mode, *mmseqs2* appears to be slightly better than custom template (where the reference PDB IDs are defined by the user), possibly as it relies on a higher number of templates for a more thorough and accurate guide. However, improvement was minimal with respect to custom templates selected to cover all the residues of both proteins, and the latter in some cases still resulted in high peak values of the scores.

Do not align was kept as False, and remove template sequence was set as True.

Number of msa and extra msa were kept at their maximum of 512:4096, in order to maximize alignment. The cluster profile was set True due to the higher performance it demonstrated, since it helps deconvolve structural signals from different conformations, leading to more accurate and diverse structure predictions, as previously mentioned.

Univariate Optimization

For the sake of analyzing each variable parameter, we proceeded by progressively changing them, in the order of expected relevance. Specifically: cov was varied from 0 to 75, since we learned that the model performs better on our complex with a lower coverage threshold; id was 90 or 100; qid was 0 or 30, since we learned that most of the times the qid doesn't change the results when varying from 0 to 20. The stochastic variables, masked language model (MLM) and dropout regularization, needed to be tuned, too.

Before optimizing the prediction parameter-wise, the default tested setup was: id 90, qid 0, MLM and Dropout True.

In order to test for varying *coverage thresholds (cov),* we computed the scores of the predictions (in particular, we used the multi score, an overall metric detailed in the Introduction section) on 4 recycles and averaged on the 128 seeds for each value of cov, namely 0, 25, 50, and 75. The results of this analysis are shown in Fig. 47. We compared all the cov curves with each other through a paired T-test and obtained that cov 25 was the best coverage threshold for our complex among the tested values.

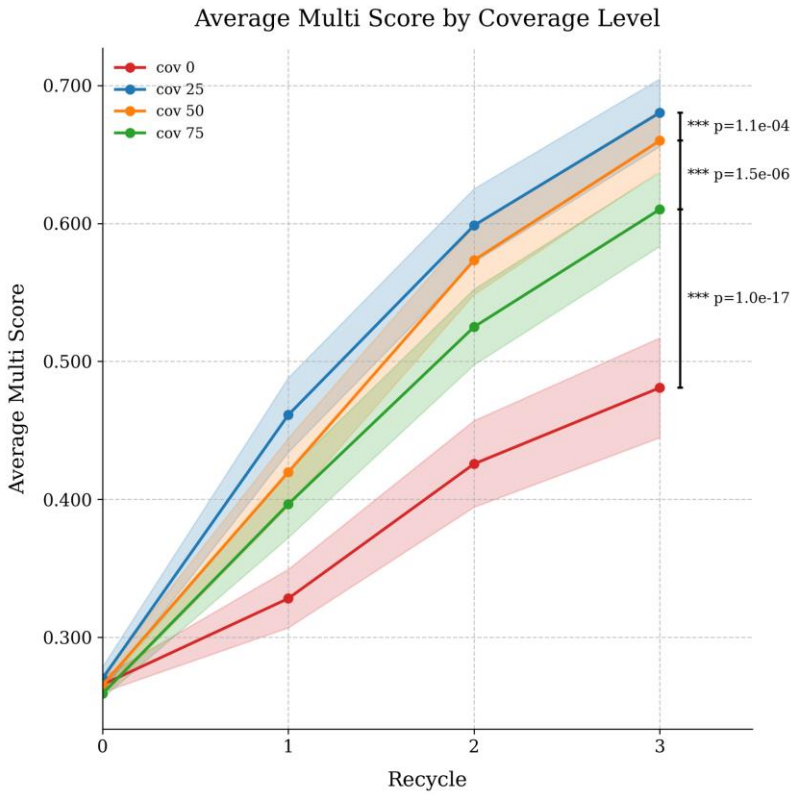


Figure 47. *Average multi scores over 128 seeds (from 0 to 127) across recycles for each coverage level.* Shaded areas represent confidence intervals. Black square brackets at the third recycle connect adjacent coverage levels, with significance symbols (* p<0.05, ** p<0.01, *** p<0.001) and p-values for T-test. The curve representing cov 25 has significantly higher scores across 4 recycles (0-3) than the others.

After selecting cov as 25, we proceeded with testing for varying *sequence identity (id).* By comparing the curves obtained analogously as before, id 90 resulted to be superior in scores than id 100, which instead would mean not filtering the MSA sequences in this step.. The results are shown in Fig. 48.

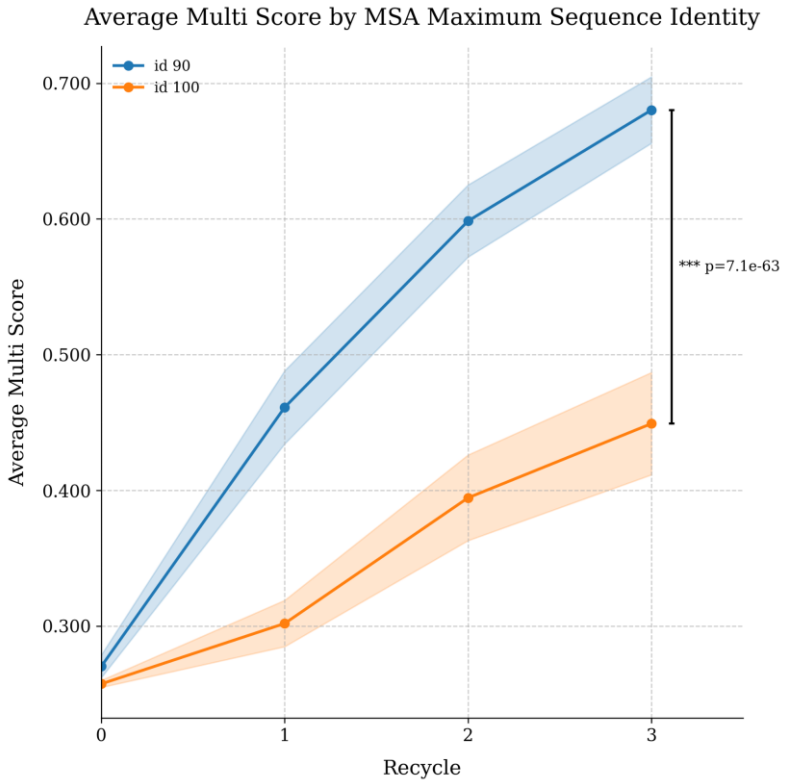


Figure 48. *Average multi scores over 128 seeds (from 0 to 127) across recycles for each sequence identity level.*
Shaded areas represent confidence intervals. Black square brackets at the third recycle connect adjacent coverage levels, with significance symbols (* p<0.05, ** p<0.01, *** p<0.001) and p-values for paired T-test. The curve representing id 90 has significantly higher scores across 4 recycles (0-3) than id 100.

Having chosen cov 25 and id 90, we had to test for varying *minimum query identity (qid).* By comparing the curves, qid 0, which means not filtering the MSA sequences in this step, slightly outperformed visually qid 30, but no statistically significant difference was detected by paired T-test., which instead would mean not filtering the MSA sequences in this step. The results are shown in Fig. 49.

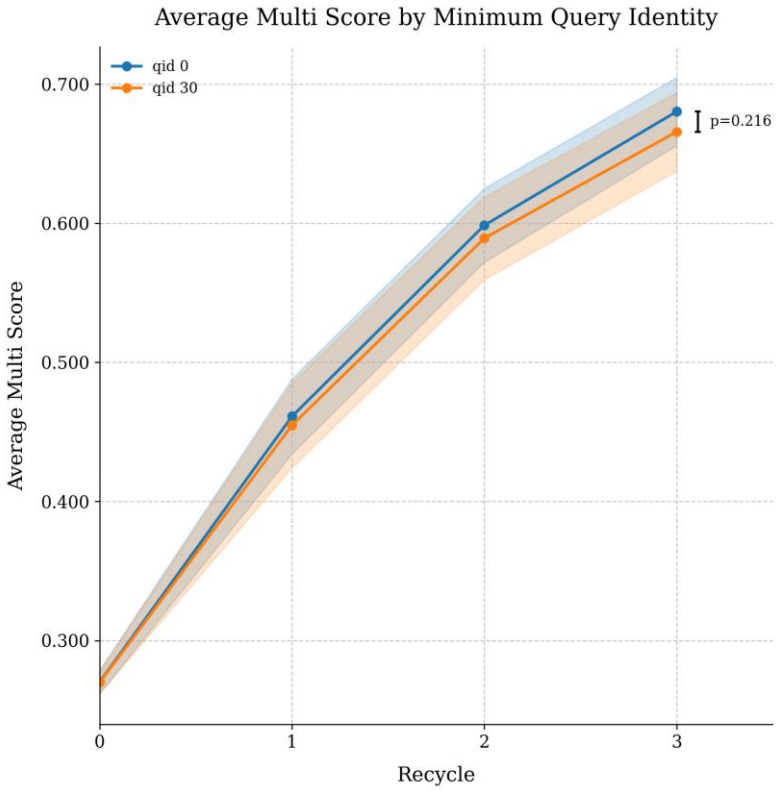


Figure 49. *Average multi scores over 128 seeds (from 0 to 127) across recycles for each query identity level.*
Shaded areas represent confidence intervals. Black square brackets at the third recycle connect adjacent coverage levels, with significance symbols (* p<0.05, ** p<0.01, *** p<0.001) and p-values for paired T-test. The curve representing qid 0 has slightly higher scores across 4 recycles (0-3) than qid 30, but not in a statistically significant way.

After having selected the filtering options for the MSA (cov 25, id 90, qid 0), we had to tune the stochastic variables, *MLM* and *dropout regularization*. We adopted the same strategy and we systematically analyzed all the four combinations of these two parameters: both True, one of them alternatively True, and both False. The resulting curves across 128 seeds and 4 recycles revealed that it is enough to set only the dropout as False (never disabling any neuron of the model) to obtain significantly higher prediction scores, independently of the value of MLM. Although no difference was detected between a True or False MLM, we selected it as False because the corresponding had the highest scores visually. The results of this analysis are shown in Fig. 50.

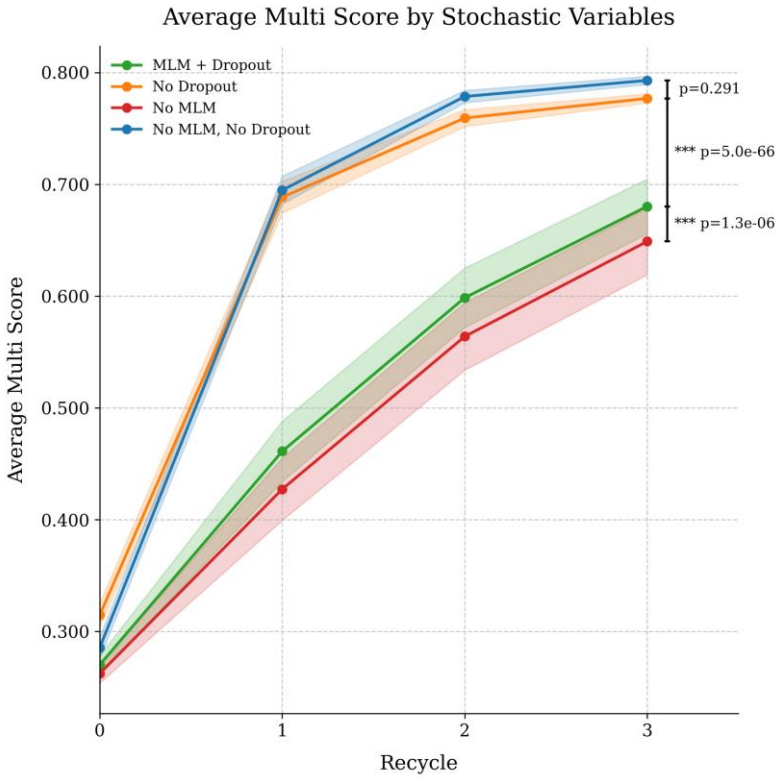


Figure 50. *Average multi scores over 128 seeds (from 0 to 127) across recycles for each combination of stochastic variables.*
Shaded areas represent confidence intervals. Black square brackets at the third recycle connect adjacent coverage levels, with significance symbols (* p<0.05, ** p<0.01, *** p<0.001) and p-values for paired T-test. The stochastic options MLM and Dropout Regularization impact the model accuracy curve, and they interact between each other. The curves representing Dropout Regularization False have significantly higher scores across 4 recycles (0-3) than the other two. In particular, setting also the MLM as False yields the highest curve visually.

For completeness, additional considerations were introduced about the systematic selection of the *pair mode* best *model* out of the five available in the notebook. Running again all the predictions, this time averaging over 32 seeds instead of 128, we confirmed previous observations on the superiority of unpaired mode over the unpaired_paired one and also the superiority of model 1 over all the others. Model 4, instead, appeared to be the second highest, while model 5 resulted to be the worst one performing on our query, both models in a statistically significant way. These findings are shown in Fig. 51.

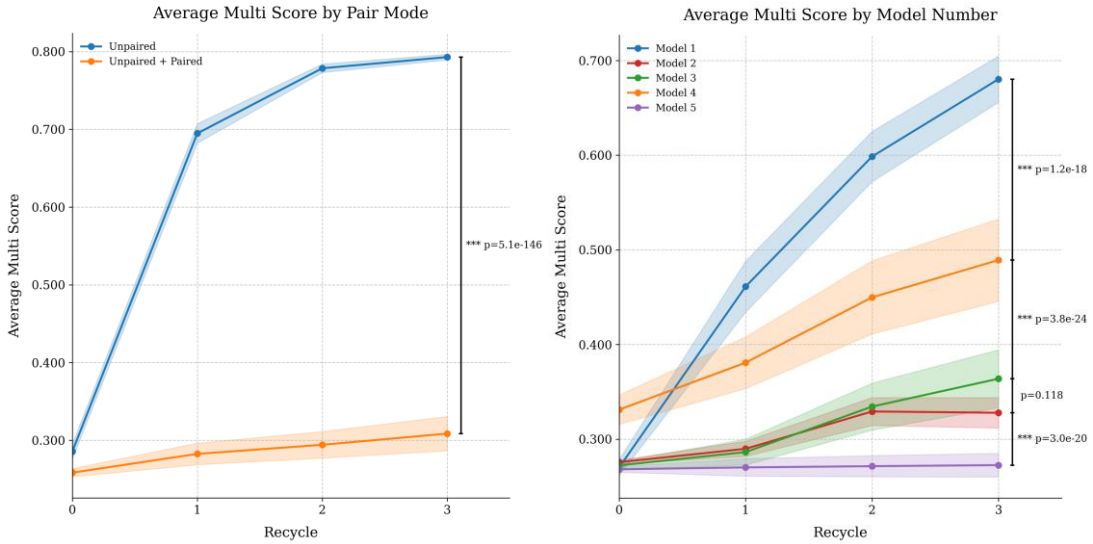


Figure 51a (left). *Average multi scores over 32 seeds (from 0 to 31) across recycles for unpaired and unpaired_paired modes.*

Shaded areas represent confidence intervals. Black square brackets at the third recycle connect adjacent coverage levels, with significance symbols (* p<0.05, ** p<0.01, *** p<0.001) and p-values for paired T-test. The curve representing unpaired mode has significantly higher scores (with a strikingly low p-value) across 4 recycles (0-3) than the unapired_paired mode.

Figure 51b (right). *Average multi scores over 32 seeds (from 0 to 31) across recycles for each different model architecture of Multimer v2.3.*

This test was conducted with MLM and Dropout Regularization set as True not to bias the analysis of all the models with the findings related to just model 1. However, the filtering parameters, considered their probable causative link more with the sequences than with the model architecture itself, were kept to the values that yielded the best results for model 1. Shaded areas represent confidence intervals. Black square brackets at the third recycle connect adjacent coverage levels, with significance symbols (* p<0.05, ** p<0.01, *** p<0.001) and p-values for paired T-test. The curve representing model 1 has slightly significantly higher scores across 4 recycles (0-3) than the other models. Interestingly, at the same time, model 4, despite being worse than model 1, significantly outperforms the remaining models. Also, model 5 is the statistically significant worst of all of them.

Based on these results, we chose to optimize the second- and third-best models in order to compare them against the optimized curve of model 1. Model 4 was selected as the second-best performer, and model 2 was chosen for third place, as there was no statistically significant difference between it and model 3.

The optimized parameters for model 2 were: cov 50, id 90, qid 0, no MLM and no dropout, unpaired mode. For model 4: cov 25, id 90, qid 0, no MLM and no dropout, unpaired mode. The results of the comparison of these optimized curves over 4 recycles with the best-performing model 1 are shown in Fig. 52. In the end, model 1 was still the statistically significant best. For this reason, the final prediction was carried out using model 1, optimized at cov 25, id 90, qid 0, no MLM and no dropout, unpaired mode.

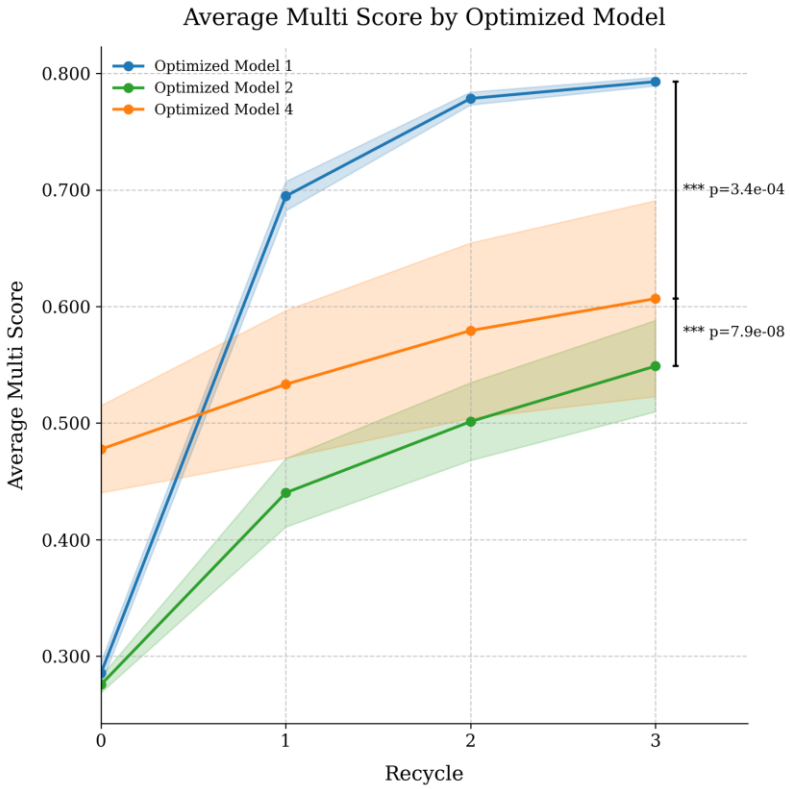


Figure 52. *Average multi scores across recycles for the model architectures of Multimer v2.3 number 1, 2 and 4 optimized for the local best set of parameters.*
Shaded areas represent confidence intervals. Black square brackets at the third recycle connect adjacent coverage levels, with significance symbols (* p<0.05, ** p<0.01, *** p<0.001) and p-values for paired T-test. The averaging of the scores was calculated over 128 seeds for model 1, while over 32 seeds for model 2 and 4, as demonstrated by the narrower confidence interval for the curve of model 1. The curve representing model 1 has significantly higher scores across 4 recycles (0-3) than the other optimized models. Therefore, it must be the model of choice for the final prediction.

3.5.3 AlphaFold Multimer v2.3 Predicts the Structure of the CHI3L1-RAGE VC1 Complex with an Accuracy of 88% for the Overall Fold and 92% for the Protein Interface

We ran the final predictions for the CHI3L1-RAGE (VC1) interaction of the optimized model 1 on all 128 seeds across 24 recycles (from 0 to 23).

On the whole, all scores reached a plateau at about 0.860 of i_pTM (interface predicted template modelling). The best prediction occured on seed 38, whose full log is shown in Fig. 53, and it achieved the following scores at recycle 22: pLDDT = 0.943, pTM = 0.920, i_pTM = 0.870, actifpTM = 0.92, and multi = 0.880.

Therefore, since the interface predicted template modelling (i_ptm) is 87%, so greater than 80% and even close to 90%, the resulting structure can be considered a confident high high-quality prediction. [289]

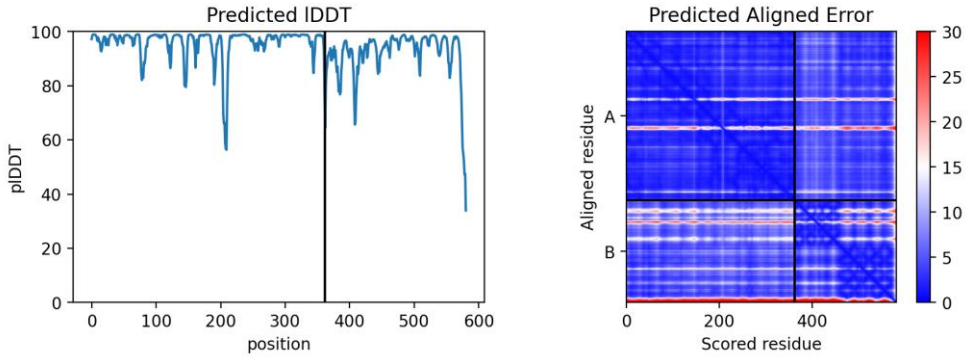
```
seed=38 model=model_1_multimer_v3 recycle=0 plddt=0.863 ptm=0.654 i_ptm=0.166 multi=0.264
seed=38 model=model_1_multimer_v3 recycle=1 plddt=0.919 ptm=0.869 i_ptm=0.760 multi=0.782 rmsd_tol=27.617
seed=38 model=model_1_multimer_v3 recycle=2 plddt=0.928 ptm=0.885 i_ptm=0.804 multi=0.821 rmsd_tol=1.346
seed=38 model=model_1_multimer_v3 recycle=3 plddt=0.931 ptm=0.889 i_ptm=0.813 multi=0.829 rmsd_tol=0.447
seed=38 model=model_1_multimer_v3 recycle=4 plddt=0.935 ptm=0.897 i_ptm=0.820 multi=0.836 rmsd_tol=0.330
seed=38 model=model_1_multimer_v3 recycle=5 plddt=0.937 ptm=0.902 i_ptm=0.836 multi=0.849 rmsd_tol=0.302
seed=38 model=model_1_multimer_v3 recycle=6 plddt=0.936 ptm=0.903 i_ptm=0.832 multi=0.846 rmsd_tol=0.127
seed=38 model=model_1_multimer_v3 recycle=7 plddt=0.937 ptm=0.904 i_ptm=0.837 multi=0.851 rmsd_tol=0.143
seed=38 model=model_1_multimer_v3 recycle=8 plddt=0.939 ptm=0.910 i_ptm=0.847 multi=0.860 rmsd_tol=0.079
seed=38 model=model_1_multimer_v3 recycle=9 plddt=0.939 ptm=0.910 i_ptm=0.850 multi=0.862 rmsd_tol=0.078
seed=38 model=model_1_multimer_v3 recycle=10 plddt=0.942 ptm=0.915 i_ptm=0.859 multi=0.870 rmsd_tol=0.047
seed=38 model=model_1_multimer_v3 recycle=11 plddt=0.942 ptm=0.916 i_ptm=0.861 multi=0.872 rmsd_tol=0.045
seed=38 model=model_1_multimer_v3 recycle=12 plddt=0.941 ptm=0.914 i_ptm=0.858 multi=0.869 rmsd_tol=0.033
seed=38 model=model_1_multimer_v3 recycle=13 plddt=0.941 ptm=0.914 i_ptm=0.857 multi=0.868 rmsd_tol=0.047
seed=38 model=model_1_multimer_v3 recycle=14 plddt=0.941 ptm=0.914 i_ptm=0.858 multi=0.869 rmsd_tol=0.051
seed=38 model=model_1_multimer_v3 recycle=15 plddt=0.942 ptm=0.917 i_ptm=0.864 multi=0.874 rmsd_tol=0.056
seed=38 model=model_1_multimer_v3 recycle=16 plddt=0.941 ptm=0.915 i_ptm=0.860 multi=0.871 rmsd_tol=0.059
seed=38 model=model_1_multimer_v3 recycle=17 plddt=0.942 ptm=0.916 i_ptm=0.860 multi=0.872 rmsd_tol=0.050
seed=38 model=model_1_multimer_v3 recycle=18 plddt=0.941 ptm=0.915 i_ptm=0.858 multi=0.869 rmsd_tol=0.057
seed=38 model=model_1_multimer_v3 recycle=19 plddt=0.940 ptm=0.914 i_ptm=0.859 multi=0.870 rmsd_tol=0.050
seed=38 model=model_1_multimer_v3 recycle=20 plddt=0.940 ptm=0.915 i_ptm=0.860 multi=0.871 rmsd_tol=0.051
seed=38 model=model_1_multimer_v3 recycle=21 plddt=0.941 ptm=0.915 i_ptm=0.861 multi=0.871 rmsd_tol=0.051
seed=38 model=model_1_multimer_v3 recycle=22 plddt=0.943 ptm=0.920 i_ptm=0.870 multi=0.880 rmsd_tol=0.067
seed=38 model=model_1_multimer_v3 recycle=23 plddt=0.941 ptm=0.915 i_ptm=0.858 multi=0.870 rmsd_tol=0.048
seed=38 model=model_1_multimer_v3 recycle=24 plddt=0.941 ptm=0.915 i_ptm=0.860 multi=0.871 rmsd_tol=0.043
```

Figure 53. *Log output of the best prediction (seed 38).*

The best scores are achieved at recycle 22 and are highlighted in red. We can clearly observe that the predicted structure shows convergence across 24 recycles, as the scores stabilize at high values and peak at one of the last recycles, while the rmsd_tol (that measures structural change with respect to the previous recycle) lowers. This is a further sign of good reliability of the prediction.

The per residue metrics associated with the best predicted complex display very high confidence, too. In particular, the predicted local distance difference test (pLDDT) and the predicted aligned error (PAE) are shown in Fig. 54.

The specific chain (CHI3L1 or RAGE VC1 taken singularly) predicted template modelling (cpTM), as well as the interface predicted template modelling (i_pTM) and the actual interface predicted template modelling (actifpTM, more accurate), are instead shown in Fig. 55. We can observe that the score for the actual interface is even higher than 90%.



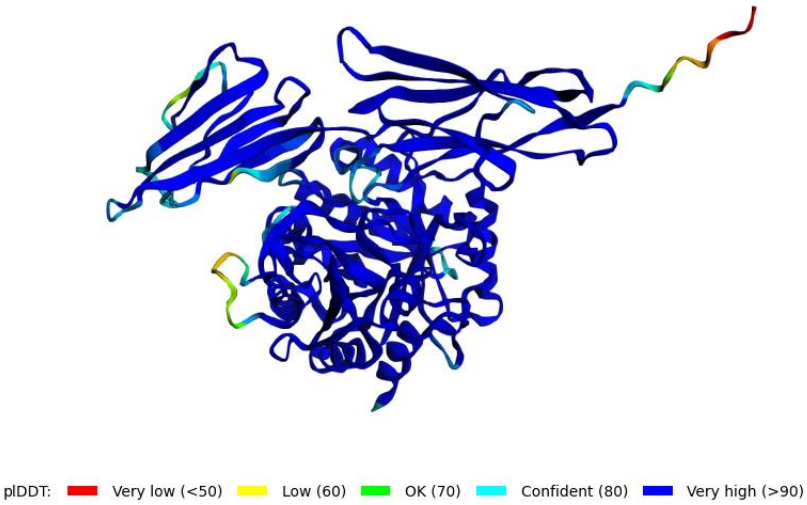


Figure 54. Per residue pLDDT and PAE for the best predicted CHI3L1-RAGE VC1 complex.
In the top left figure, the predicted LDDT per residue is shown. The PAE matrix (top right) shows small error in the relative position of the subunits, and none of the high error residues are present in the predicted interface. In the bottom representation, we can see that the complex colored by pLDDT is mostly very highly confident, except for some loops. The structure is oriented to show CHI3L1 at the bottom and the two domains of RAGE bound on top of it.

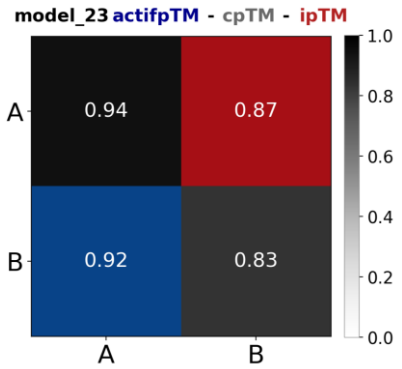


Figure 55. *Matrix of the per-chain and inter-chain scores of the best model in a comprehensive view.*
actifpTM = 0.92, chain-pTM for CHI3L1 = 0.83, chain-pTM for RAGE VC1 = 0.83, and i_pTM = 0.87.

We would like to underline that the other optimized Multimer v2.3 model architectures (2 and 4) also achieved very high scores, demonstrated convergence across 24 recycles and, most importantly, showed a very high degree of agreement in terms of predicted overall fold and interface of the two proteins, further confirming the goodness of the prediction.

Analysis of Amino Acids at the Interface of the CHI3L1-RAGE VC1 Complex Prediction
The AlphaFold-predicted model of the CHI3L1-RAGE VC1 complex reveals a specific and well-defined interaction surface. CHI3L1, as we know, is constituted by a β-barrel structure with a deep cleft or groove that typically functions as its chitin-binding site. In the model, this cleft is oriented directly toward the VC1 domains of RAGE, with a prominent interaction involving the FG loop of the C1 domain (residues Pro212 to Ala219, between the F and the G β strands) and the following first part of the GG' loop (residues Thr222 to Pro227, between the G and G' β strands). [140] This loop appears to insert into the central cleft of CHI3L1, as shown

in Fig. 56a, positioning itself snugly within it and forming multiple contacts that suggest a strong and specific mode of interaction.

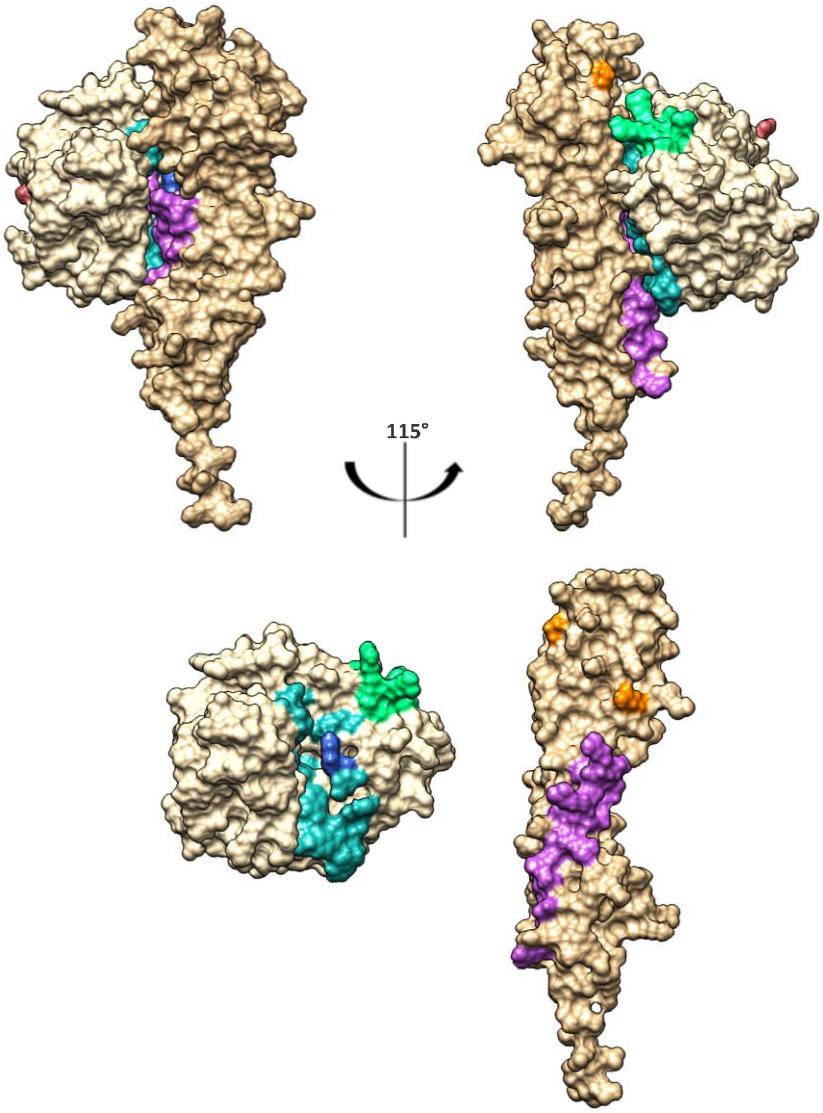


Figure 56a (top). *Surface representation of the front view (left) and rear view (right) of CHI3L1-RAGE VC1 predicted protein complex with open book representation of the two proteins*.
The FG and GG' loops of RAGE (P212-P227) are colored in purple. The sites of N-glycosylation of RAGE are shown in orange. The putative heparin binding domain (W325-V338) is shown in green and the heparan sulfate binding motif (G122-H128) on the opposite side in red. From the rear view, it is interesting to note that the N-glycosylation of RAGE present on Asn81, that can be glycosylated in variable forms, is positioned in the model exactly in correspondence of the putative heparin binding motif. Heparin is itself a glycosaminoglycan and it was hypothesized to bind on this site due to the presence of many lysine and arginine residues, which are positively charged and can welcome negatively charged glycans.
Figure 56b (bottom). *Surface representation of the front view CHI3L1-RAGE VC1 predicted protein complex with open book representation (at 115°) of the two proteins*.
The chitin-binding domain of CHI3L1 is shown in aquamarine, with Trp99 highlighted in blue. There is a correspondence between this site and the FG and GG' loops of RAGE.

Analyzing the interaction in more detail, we must consider that the FG loop of RAGE (sequence RHRALRT) is rich in positively charged arginine residues. In the predicted model, these residues form several cation-π interactions with aromatic side chains located within the chitin-binding groove of CHI3L1. Notably, Arg216 of RAGE interacts with Trp99 of CHI3L1, the residue known to play a key role at the entrance of the groove by regulating its conformational opening to accommodate chito-oligosaccharides, and this arginine is the deepest amino acid of the FG loop inside the groove. This specific cation-π interaction may be equally important in stabilizing the insertion of the FG loop into the CHI3L1 cleft. Furthermore, Arg218 forms a cation-π interaction with Trp31, and Arg221 interacts with Tyr34 at the outer edge of the groove. These interacting residues are shown in context inside the structure in Fig. 57. Interestingly, the same Arg216 and Arg218 of the FG loop are also known to interact with heparan sulfate, which is a driver of oligomerization for RAGE. [139]

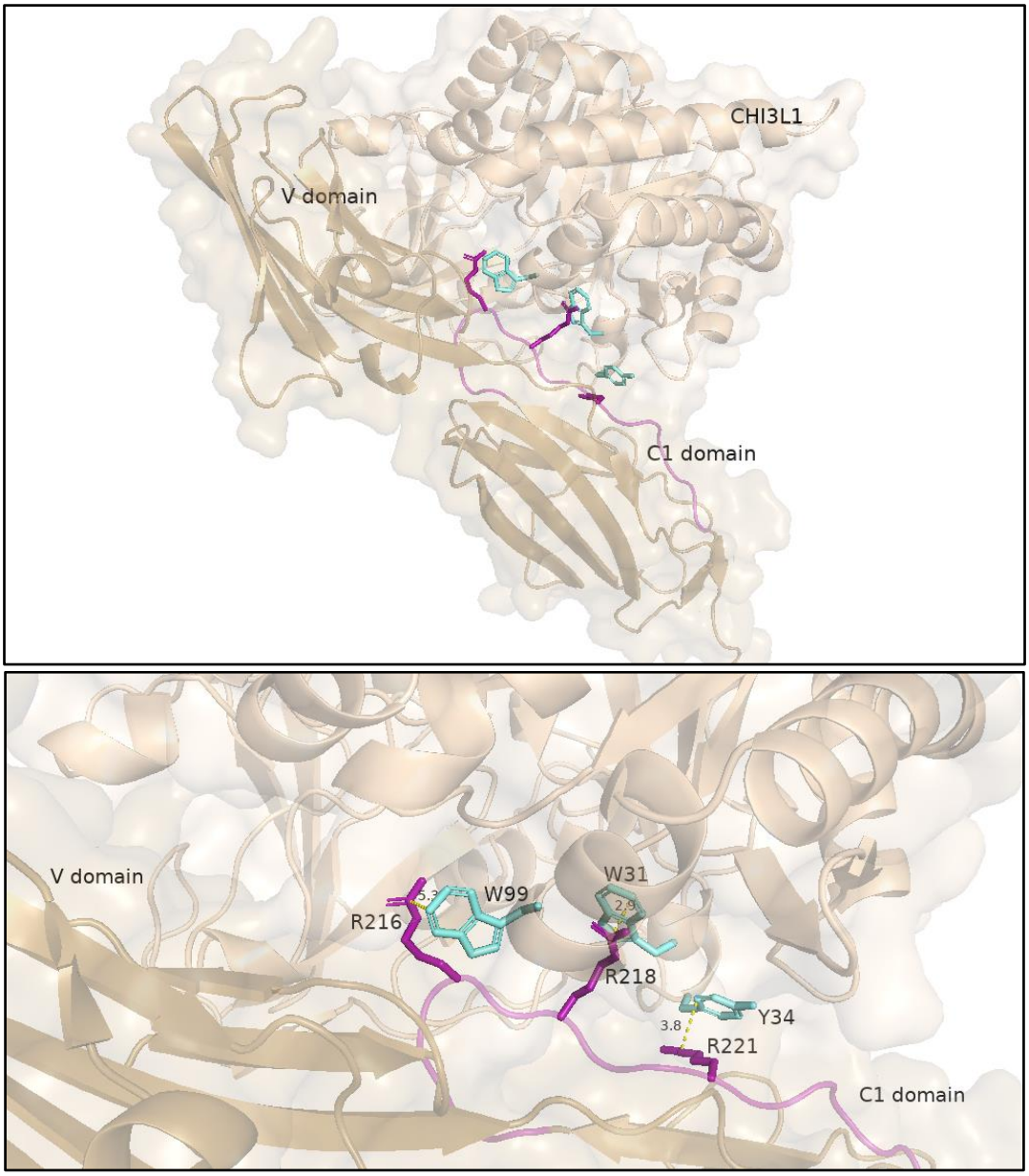


Figure 57. *Top-down view of the predicted CHI3L1-RAGE VC1 complex with the binding site highlighted (top) and zoomed-in view of the same interaction (right).*
In purple, the FG and the GG' loops of RAGE are shown, while in aquamarine the three main interacting residues from the CHI3L1 binding groove. The distance between the residues has been measured as 5.3 Å between Arg216 and Trp99, 2.9 Å between R218 and Trp31, and 3.8 Å between Arg221 and Tyr34.

Surprisingly, the prediction even reveals a binding-induced conformational shift of a specific amino acid, which is a notable outcome considering the rarity of such results in AlphaFold predictions. Specifically, the model captures the movement of Trp99 from a closed to an open conformation, an observation that aligns with previous reports describing this shift during the binding of chito-oligosaccharides within the CHI3L1 cleft and further detailed in the Introduction section about CHI3L1. [194] In our predicted complex, Trp99 adopts the open conformation to accommodate the FG loop of the RAGE C1 domain within the cleft. This structural feature not only supports the biological relevance of the predicted interaction but also highlights the potential of AlphaFold to capture functionally meaningful conformational changes. The result is illustrated in Fig. 58.

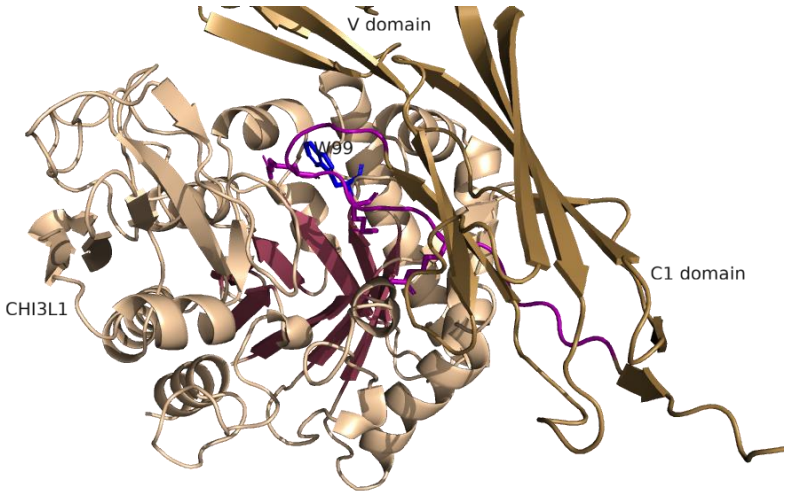


Figure 58. *Predicted complex of CHI3L1-RAGE VC1 showing the Trp99 (blue) shift to an open conformation in the interaction with RAGE to accommodate its FG loop (purple).*

It is also clear that the electrostatic charges of the two proteins might play a role in the complex formation. In fact, by visualizing the surface solvent electrostatics at physiological pH of the two monomers detached from the complex we could appreciate a very big positive patch on one side of RAGE VC1 domains. This feature was already well documented in literature and corresponds to the side that faces CHI3L1 in the complex predicted by the model. [140] More specifically, it contains a big part of the V domain and a part of the C1, including the FG loop. In fact, RAGE uses that side to bind heparan sulfate, too. This positively charged surface on RAGE VC1 seems to fit with two slightly negative areas on CHI3L1, one located precisely in the deepest point of the chitin-binding cleft and the other one just above. The open-book representation of this surface comparison is shown in Fig. 59.

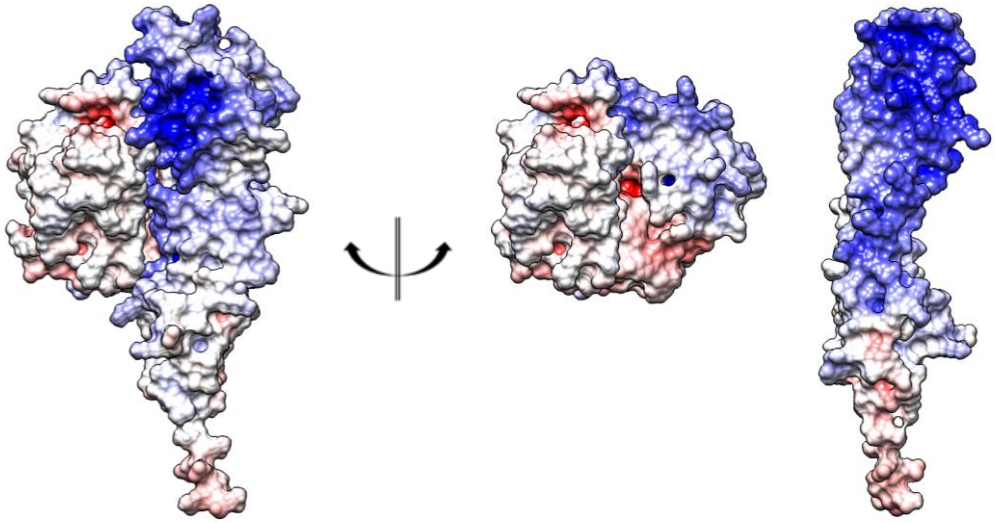


Figure 59. *Surface representation of the solvent electrostatics of the CHI3L1-RAGE complex with open-book representation.*

3.5.4 Predicted Binding Mode of RAGE VC1C2 to CHI3L1 Mirrors That of RAGE VC1

To fully confirm the reliability of the prediction for the complex involving the VC1 domains, we applied the same procedure described in the previous subsection to predict the interaction between CHI3L1 and the RAGE VC1C2 domains.

The parameters optimization procedure revealed that the only parameter that differs in the set used for the prediction with VC1C2 is id 100, which performed better than id 90.

The best scores for the final prediction were obtained in seed 11 at the recycle 20. Specifically, pLDDT = 0.928, pTM=0.837, i_pTM=0.845, actifpTM=0.92, and multi=0.843. All the outputs for this prediction are shown in Fig. 60.

Notably, not only are the scores very high, almost approaching the VC1 prediction accuracy that has the advantage of lacking a very flexible region between VC1 and C2, but the final predicted interface is almost identical to the one predicted for VC1. In fact, the actual interface predicted template modelling remained equal to the previous prediction. These findings are of paramount importance to strengthen even more the reliability of these predictions because, although we changed the starting domains of RAGE, the model placed the interaction exactly in the same location and with the same protein conformation. Moreover, also the prediction with VC1C2 showed convergence across recycles.

It is particularly interesting to analyze the features of the C2 domain. The coevolution matrix, in Fig. 61, shows that this domain is highly co-evolved within itself; however, the relative orientation of its residues with respect to both the other domains of RAGE and CHI3L1 is predicted with low confidence, as indicated by the predicted alignment error (PAE) matrix. This interpretation is further supported by the agreement matrix, which serves as a measure of predictive reliability: these residues were not informative, and likely even detrimental, for the structural prediction, despite exhibiting a strong internal coevolutionary signal. This finding, *a posteriori*, confirms the foresight of our initial decision to conduct the analysis using only the RAGE VC1 domains, which indeed yielded higher overall fold accuracy.

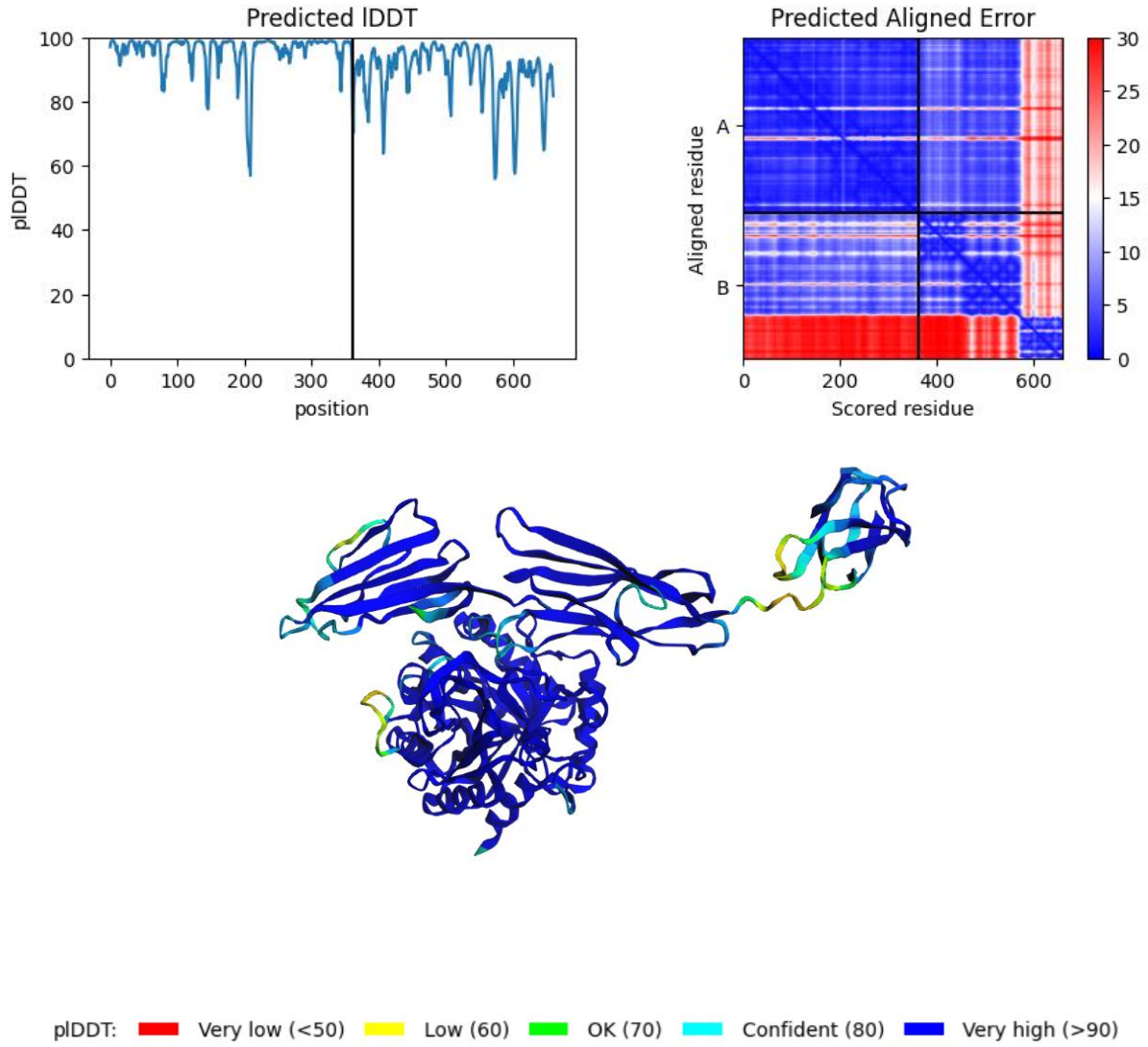


Figure 60. *Output of the best RAGE-VC1C2 complex prediction*.
At the top left, the pLDDT is plotted per residue, while at the bottom right, the predicted alignment error matrix is displayed.
At the bottom, the pLDDT is shown on the predicted structure, where the interaction is nearly identical to the one predicted in the CHI3L1-RAGE VC1 complex.

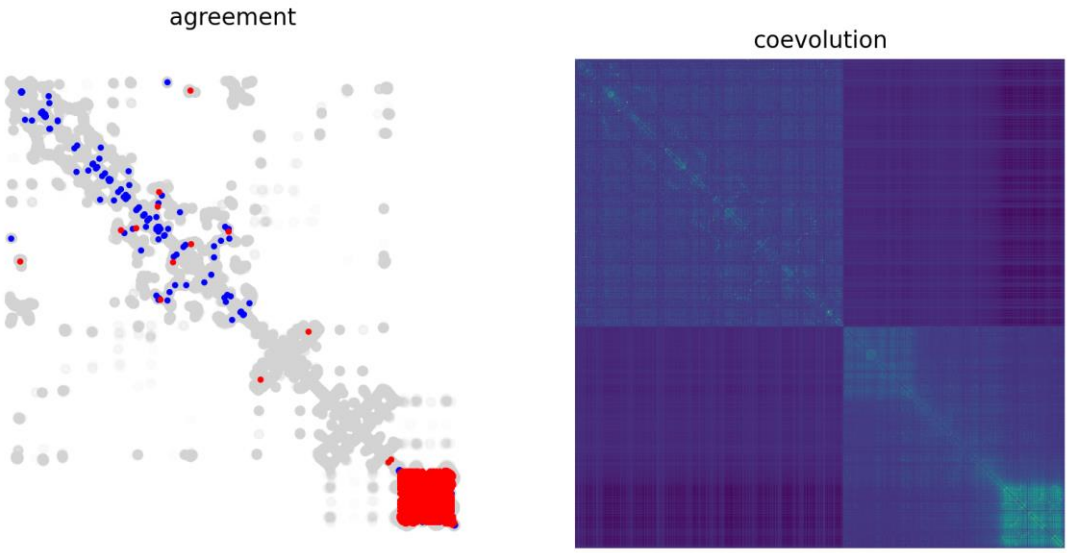


Figure 61. *Agreement matrix (left) and coevolution matrix (left) for the CHI3L1-RAGE VC1C2 complex.* It is clear that the C2 domain, despite being highly co-evolved within itself, supplies the model with coevolutionary predictions not matched by structural reality, meaning that it is likely not involved at all in the interaction with CHI3L1.

3.5.5 AlphaFold-Multimer v2.3 Is Also Capable of Predicting the Structure of the S100A8/A9-RAGE VC1 Complex

To better elucidate the molecular interactions landscape of the complex biologic system in which CHI3L1 and RAGE are involved, and to potentially find a way to selectively inhibit the binding of CHI3L1 to RAGE while preserving the beneficial interaction of the receptor with S100A8/A9, we applied the same strategy discussed above to also predict the structure of the S100A8/A9-RAGE VC1 complex.

For simplicity, we used the input sequences of the heterodimer of S100A8/A9, even though in physiological conditions and in the presence of calcium and zinc, these two proteins are able to also form a heterotetramer. As for the other predictions, the sequences are fully detailed in the Materials and Methods section.

The initial observation concerns the coevolution matrix of the S100A8/A9 heterodimer, shown in Fig. 62. The use of the unpaired_paired mode yielded better performance in this case, suggesting that an expected paired evolutionary pattern underlies the coevolution of these proteins. Accordingly, we computed the coevolution matrices using both the unpaired_paired and paired modes. Not coincidentally, these revealed distinct but complementary features: a strong intra-monomer coevolutionary signal in the unpaired_paired mode, and a clear inter-monomer coevolution pattern in the paired mode.

Regarding the final prediction, the optimized set of parameters was surprisingly different from the one used for CHI3L1 and RAGE. It consisted of cov 0, id 100, qid 0, MLM True, Dropout False, and unpaired_paired mode. The predicted complex displayed the following scores: pLDDT = 0.890, pTM = 0.824, ipTM = 0.783, multi = 0.792. The outputs of this prediction are shown in Fig. 63 and Fig. 64.

Therefore, the predicted structure for this complex had an overall good confidence. The only caveats of this analysis are: the multi score does not cross the 80 % threshold to be defined a confident high-quality predictions; the model did not show convergence across 24 recycles, while the best prediction has been achieved at recycle 3 and the following recycles determined an increase in the rmsd_tol metric and a decrease in the accuracy. This last issue could be related to the inherent complexity of modelling a trimeric complex, where the relative interaction of each protein with the others might affect the stability of the prediction.

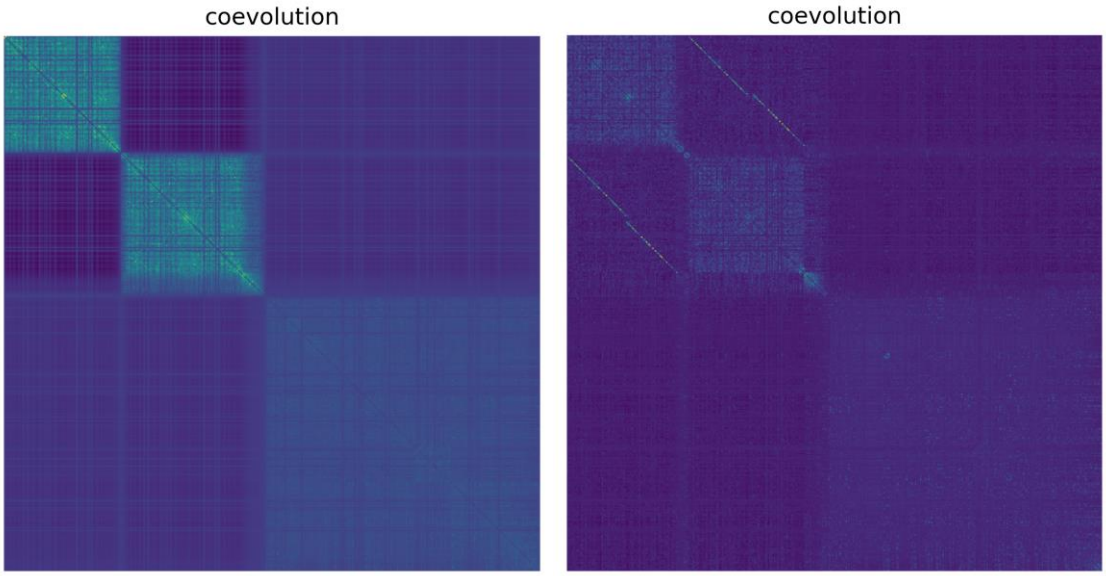
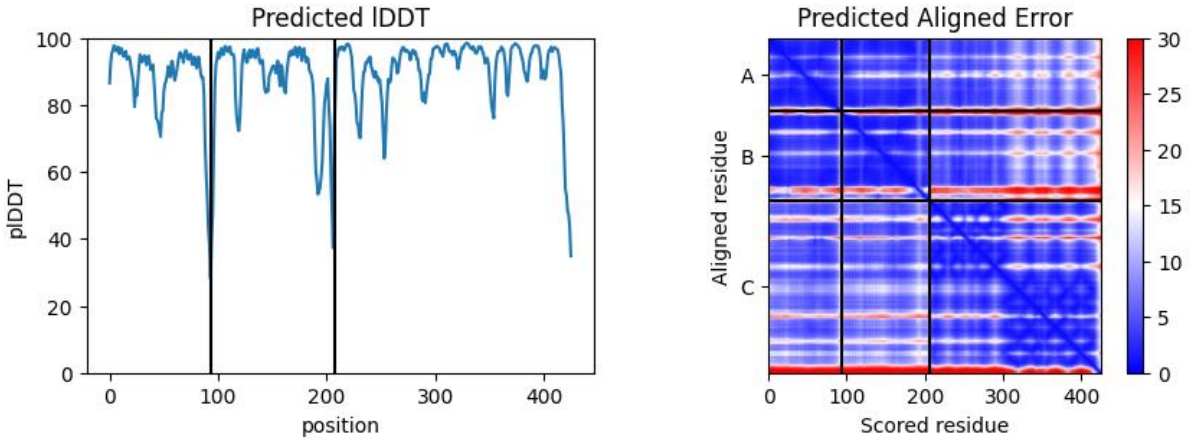


Figure 62. *Coevolution matrix of the S100A8/A9-RAGE VC1 complex with unpaired_paired mode (left) and paired mode (right)*.
S100A8 is displayed in the left upper quadrant, S100A9 in the middle one, and RAGE VC1 in the lower right quadrant of each matrix.



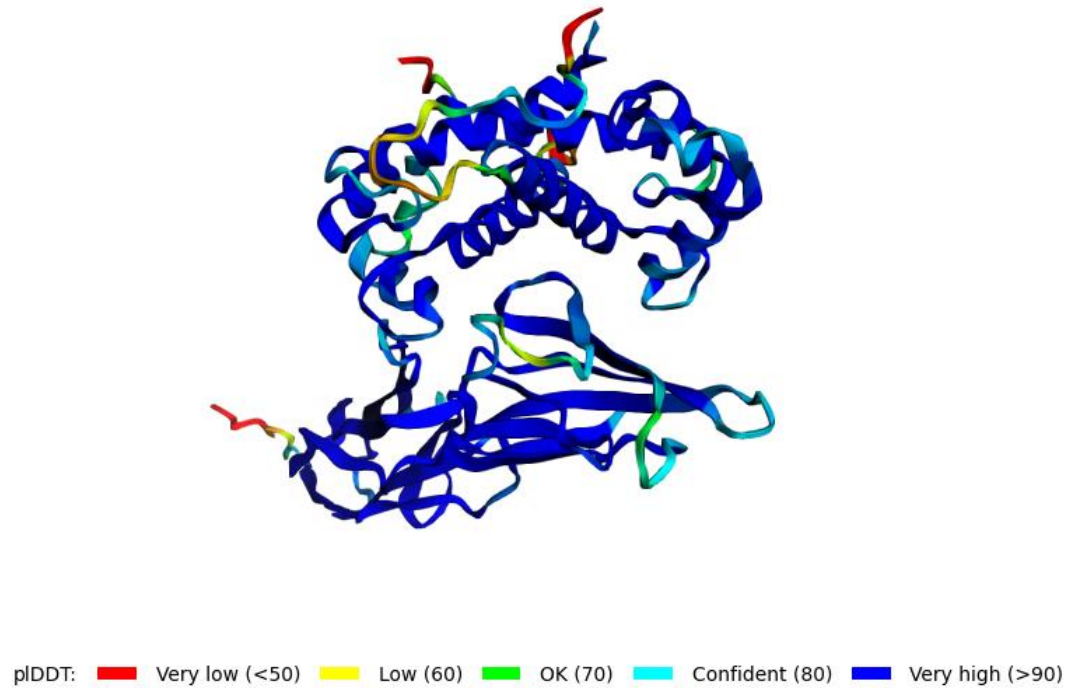


Figure 63. *Per residue pLDDT and PAE for the best predicted CHI3L1-RAGE VC1 complex.*
In the top left figure, the predicted LDDT per residue is shown. The PAE matrix (top right) shows small error in the relative position of the subunits, and none of the high error residues are present in the predicted interface. In the bottom representation, we can see that the complex colored by pLDDT is mostly highly confident, except for some small regions, which are not involved in the interaction, anyways. The structure is oriented to show the S100A8/A9 heterodimer at the top that binds the V domain of RAGE below it.

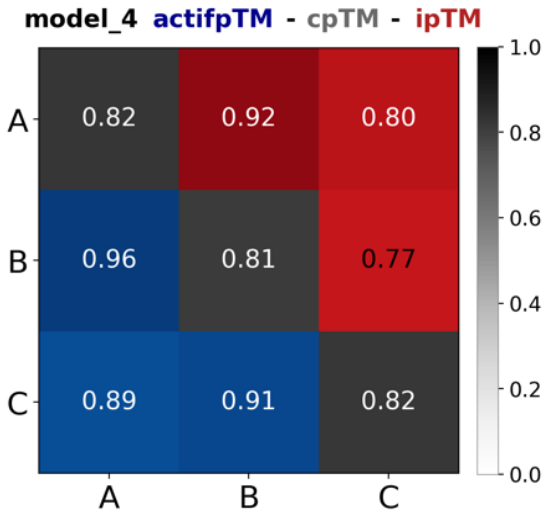


Figure 64. *Matrix of the per-chain and inter-chain scores of the best model in a comprehensive view*. Chain A represents S100A8, chain B S100A9, and chain C represents RAGE VC1. actifpTM (S100A8 with S100A9) = 0.96, actifpTM (S100A9 with RAGE VC1) = 0.89, actifpTM (S100A8 with RAGE VC1) = 0.91, chain-pTM for S100A8 = 0.82, chain-pTM for S100A9 = 0.81, chain-pTM for RAGE VC1 = 0.82, i_pTM (S100A8 with S100A9) = 0.92, i_pTM (S100A8 with RAGE VC1) = 0.80, i_pTM (S100A9 with RAGE VC1) = 0.77. Notably, all the metrics are above 80 %, except the interface predicted template modelling (i_pTM) for the S100A9-RAGE VC1 interaction, which is the one

modelled the worst. However, all the actual interface predicted template modelling values are above 80 %.

To investigate how the predicted RAGE VC1–S100A8/A9 complex might coexist within the ligand-binding landscape of RAGE domains alongside CHI3L1, we superimposed the structure obtained from the S100A8/A9 prediction onto that of the CHI3L1-bound complex. The results, presented in Fig. 65, provide valuable insights. The model positioned S100A8/A9 on the side opposite to CHI3L1, specifically near the C'D loop of the V domain, an area previously reported in the literature to mediate interactions with other S100 proteins, such as S100B. [123]

This structural superposition of CHI3L1, RAGE VC1, and S100A8/A9 raises questions about the compatibility of these two ligands with their common receptor, the observed competitive binding behavior, and the potential effects on the oligomeric states of the resulting complexes.

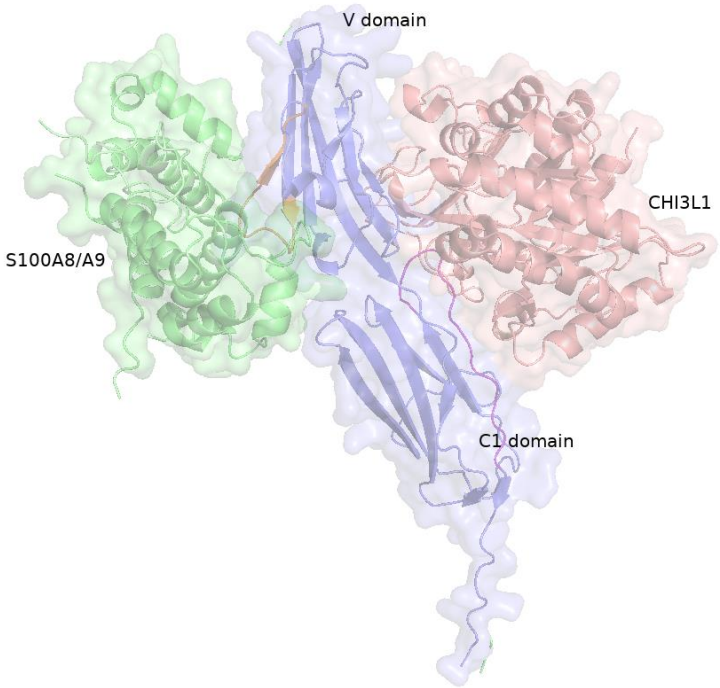


Figure 65. *Structure superposition of the prediction of the CHI3L1-RAGE VC1 complex with the prediction of the S100A8/A9-RAGE VC1 complex.*
RAGE VC1 is represented in blue, CHI3L1 in salmon and S100A8/A9 in lime. The two proteins are predicted to interact with the VC1 domains of RAGE on opposite sides of the receptor. In purple, we highlighted the FG and GG' loops of RAGE interacting with CHI3L1, and in orange the C'D loop (Asn54-Gln67) interacting with S100A8/A9.

4. Discussion

The emergence of therapeutic resistance remains a central obstacle in the management of HER2+ breast cancer. This thesis was predicated on the idea that even within a complex, adaptive system like a resistant tumor microenvironment, there exist specific molecular nodes whose disruption can yield an outsized therapeutic benefit. We identified the interaction between chitinase-3-like protein 1 (CHI3L1) and the receptor for advanced glycation end-products (RAGE) as one such node, a checkpoint through which tumors silence the cytotoxic machinery of Natural Killer (NK) cells, thereby evading antibody-based therapies like trastuzumab.

The work presented here sought to move beyond a phenomenological description of this checkpoint to provide the first detailed, molecular-level understanding of its architecture and evolutionary logic. By integrating biophysical characterization, state-of-the-art computational modeling, and co-evolutionary analysis, we have constructed a molecular blueprint of this critical immunosuppressive axis.

4.1 The Architecture of an Innate Immunity Checkpoint

The central achievement of this thesis is the generation of the first high-confidence, atomic-resolution model of the CHI3L1-RAGE complex. Our optimized AlphaFold-Multimer predictions, yielding a multi-score of 0.88 and an interface-focused actifpTM of 0.92, reveal a chemically and biologically plausible binding mode. The model demonstrates a specific, high-complementarity interaction involving the docking of the RAGE C1 domain's FG and GG' loops (residues 212-227) into the conserved 43 Å chitin-binding groove of CHI3L1.

This interface is stabilized by a network of specific interactions, most notably multiple cation-π contacts between the positively charged arginine residues of the RAGE loop and the aromatic tryptophan side chains lining the CHI3L1 groove. Our model captures a subtle but significant binding-induced rotation of Trp99 in CHI3L1, which acts as a gatekeeper to the binding cleft. (Fig. 58). This conformational shift from the protein's unbound state is a remarkable prediction by AlphaFold consistent with crystallographic studies of CHI3L1 in complexes with chito-pentaose, hexaose, and octaose (Figs. 8, 14). That an *in silico* model could sample such variability is a testament to the co-evolutionary cipher to decode structure *and* function hidden in the multiple sequence alignment (MSA).

Our structural findings are supported by biochemical data. The nanomolar affinity (apparent $K_d \approx 80$ nM) measured by ELISA is consistent with the extensive, high-complementarity interface revealed in the model. Furthermore, our competitive ELISA results, which showed that small-molecule groove binders like caffeine and short chito-tetraose could efficiently displace RAGE, are explained by this architecture. These smaller ligands likely occupy only a sub-pocket within the expansive groove, lacking the size and avidity to disrupt the far larger protein-protein interface. This observation aligns with the known biology of chito-oligosaccharides, where chain length dictates functional effect, and suggests that only longer, RAGE-mimicking ligands could serve as effective competitive inhibitors.

4.1.1 The Second Handshake: A Hypothesis for the Mysterious Motif

This model of the RAGE loop docking into the CHI3L1 groove resolves the primary binding question, but it also offers a compelling solution to a long-standing puzzle: the function of CHI3L1's mysterious motif, the GRRDKQH loop at residues 143-149. This loop is a canonical heparin-binding consensus sequence, yet shows no significant affinity for free heparin and seemed a possible evolutionary fossil.

We propose that this motif is not vestigial but acts as a context-dependent, modulatory latch. Its primary role is not to be essential for binding, but to fine-tune the interaction in a specific glycosylated environment. Our structural model (Fig. 56a) positions this positively charged loop perfectly to perform a "molecular handshake" with a structured polyanionic partner present in the native RAGE complex, such as the carboxylated glycans at Asn81 [125] or the external heparan sulfate scaffold.

This hypothesis must be reconciled with our biochemical finding that deglycosylated RAGE still binds CHI3L1 with high affinity in ELISA, and perhaps even slightly more strongly. This apparent contradiction is, in fact, highly informative. It strongly suggests that the primary, high-energy protein-protein interaction between the RAGE loop and the CHI3L1 groove is sufficient for binding and is independent of glycans.

The mysterious motif, therefore, is likely not an essential anchor but a secondary, regulatory interaction. In the crowded, dynamic environment of the cell surface, this latch could:

- increase specificity for this interaction;
- integrate the information on the glycosylation status of the receptor, that might be itself dependent on the status of the cell;
- modulate affinity. The slight increase in binding observed upon deglycosylation *in vitro* might reflect the removal of steric hindrance from bulky, flexible glycans in a simplified buffer system. On a structured cell surface, however, the same glycans could act as specific, positive anchors.

4.2 A Mechanism of Disruption

The predicted structure offers a direct and compelling mechanistic hypothesis for CHI3L1's immunosuppressive function: the disruption or alteration of RAGE oligomerization. RAGE signaling is not triggered by simple ligand binding; it is an absolute requirement that RAGE assembles into dimers or, for full activity, heparan sulfate (HS)-stabilized hexamers on the cell surface. This platform is essential for recruiting downstream adaptors and initiating signaling cascades.

Our model, as shown in Figs. 66 and 67, reveals a profound steric clash when superimposed onto the known architecture of a RAGE oligomer. When CHI3L1 docks onto one RAGE protomer, its molecular surface extends into the space required by the adjacent protomer in both the dimer and the HS-stabilized hexamer. It does not simply block the entire

oligomerization interface on one molecule; rather, it acts as a physical wedge, sterically preventing a second RAGE molecule from assuming its correct position in the complex. Furthermore, our model shows that CHI3L1 binds to the external face of the RAGE hexameric ring, precisely at the site where heparan sulfate chains are known to bind and stabilize the assembly.

By occupying this critical surface, monomeric CHI3L1 would act as a potent antagonist of oligomerization. This single, elegant mechanism unifies the seemingly disparate observations of CHI3L1's function across pathologies. The dismantling of the RAGE platform would abolish its ability to recruit DIAPH1, leading to the collapse of the JNK-Stathmin axis and the paralysis of NK cell cytotoxicity observed in $HER2^+$ breast cancer. The same disruption can explain the concomitant downregulation of the β-catenin pathway seen in multiple sclerosis. It also clarifies the nature of the competition with S100A8/A9. The contest is not for a single binding epitope, but for the conformational soul of RAGE itself. S100A8/A9, an oligomeric and acidic protein, promotes and stabilizes the active RAGE platform, while CHI3L1, a monomeric and basic protein, dismantles it. The balance between these two ligands in the tissue microenvironment dictates the life-or-death signaling status of the RAGE receptor.

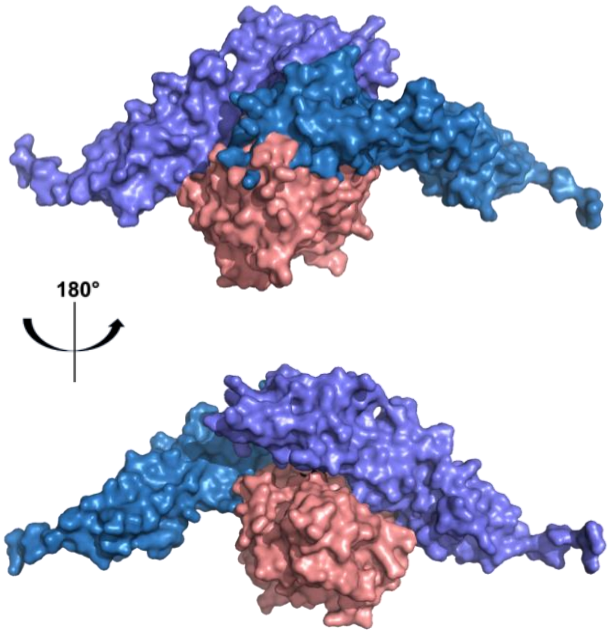


Figure 66. *RAGE VC1 homodimer superposition with CHI3L1 bound to RAGE as in the predicted complex from AlphaFold.*

To obtain this image, we generated the symmetry mates of the crystal structure of RAGE VC1 (PDB 3CJJ) and selected a homodimeric complex. We then superimposed the RAGE VC1 monomer of our AlphaFold predicted structure to one of these symmetric RAGE monomers (the one colored in slate in the Figure), obtaining the predicted mode of binding of CHI3L1 to the RAGE dimer. This led us to the observation that CHI3L1 (depicted in salmon) would clearly obstruct the formation of this dimer because it evidently clashes, especially from the front view (top), with the unbound RAGE monomer.

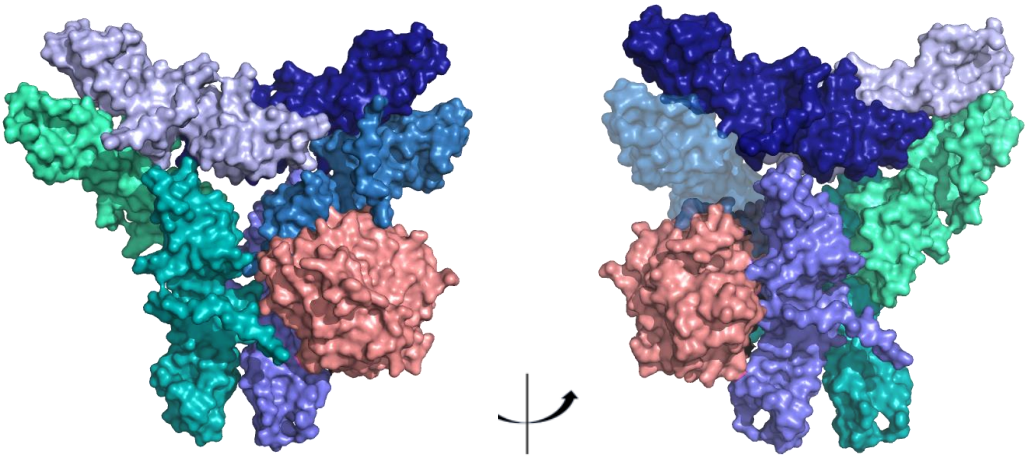


Figure 67. *RAGE VC1 hexamer in superposition with CHI3L1 bound to RAGE as in the predicted complex from AlphaFold.*
To obtain this image, we generated the symmetry mates of the crystal structure of hexameric RAGE VC1 (PDB 4IM8). We then superimposed the RAGE VC1 monomer of our AlphaFold predicted structure to one of these symmetric RAGE monomers (the one colored in slate in the Figure), obtaining the predicted mode of binding of CHI3L1 to the RAGE hexamer. Also in this case, the representation led us to the observation that CHI3L1 would obstruct the formation of this hexamer because it clashes with an unbound RAGE monomer adjacent to the bound one. It is especially visible from the half side view of the hexameric complex (right), where the semi-transparent RAGE monomer is the one clashed by the position of CHI3L1. These findings support the hypothesis that CHI3L1 (depicted in salmon) would inhibit the formation of higher order oligomers that are otherwise needed for the downstream signalling of the receptor RAGE.

4.3 An Echo in the Proteome: A Checkpoint Forged in Deep Time

The structural model of the CHI3L1-RAGE complex prompts a fundamental evolutionary question: how did an ancient chitin-binding protein and a modern immune receptor converge to form such a specific and consequential checkpoint? To answer this, we traced the evolutionary history of both proteins using taxonomic analysis of our deep multiple sequence alignments. When viewed through the lens of evolution, the structural model becomes even more fascinating.

From a curated dataset of 11,730 sequences (10,330 for CHI3L1 and 1,410 for RAGE), derived from MMseqs2 alignments used by AlphaFold, we uncovered a rich and intricate history, one that likely set the stage for the emergence of this modern-day immune checkpoint.

4.3.1 The Long Apprenticeship of CHI3L1

CHI3L1 belongs to the ancient glycoside hydrolase 18 family, with clear orthologs present in the common ancestor of all animals. By placing the organisms with aligned sequences along the time axis representing evolutionary divergence from Homo sapiens, as shown in Fig. 68,

we can trace hypothetical ancestral prototypes of the protein back to organisms that diverged as far back as microbial life in the Precambrian.

In particular, we identified several peaks of increasing similarity to our query CHI3L1, which correspond to points of divergence in the evolutionary tree. This allowed us to retrieve the most represented clades, which are monophyletic or natural groups composed of a common ancestor and all of its descendants, in each of these peaks through enrichment analysis. The organisms whose progenitors first diverged from the evolutionary tree did so during the Hadean Eon (4567–4031 million years ago, MYA[4]). Among these, the most represented are the bacterial phyla *Pseudomonadota*, *Actinomycetota*, *Bacillota*, *Bacteroidota*, and *Methanobacteriota*. This is supported by the taxonomic breakdown of all species included in the alignments, shown in Fig. 69 and color-coded by similarity to our query. Bacteria constitute approximately 30% of the entire pie chart, with the dominant phyla matching those listed above. This finding does not necessarily indicate that an ancestral CHI3L1 protein was present at the origin of life (approximately 4250 MYA), but rather suggests that the protein emerged in species whose last common ancestor with humans dates back to that time.

Progressing on the time axis, we encountered another point of divergence in the Mesoproterozoic period (1600 - 1000 MYA), where most of the enriched species belonged to the fungal phyla *Ascomycota* and *Basidiomycota*, and the plant *Streptophyta*. This must be strictly related to the emergence of chitin as a natural polymer produced by fungi and the need to cut it though chitinases for processing or defense.

A large burst of similarity to the human CHI3L1 appeared in organisms tracing back to the Neoproterozoic Era (1000 - 538.8 MYA), the intersection period between three billion years of pervasively microbial Precambrian life and the appearance of large multicellular organisms, along with supercontinental reconfiguration and the deepest glacial freeze. [290] As highlighted also by the taxonomic representation, the organisms started to show a greater similarity to our query CHI3L1, and the most enriched phyla that diverged at this time point were *Arthropoda*, *Mollusca*, *Nematoda* and *Cnidaria*, consistently with the utilization of chitin by these organisms for their external shells.

More recently, the protein evolved and adapted in clades belonging to chordates and vertebrates, starting from the class of *Actinopteri*, which includes most of the fishes, highly represented in the Silurian Period (443.1 - 419.62 MYA), and passing through the *Aves* (birds), the *Lepidosauria* (reptiles), and the *Amphibia,* all tracing back to the Carboniferous Period (358.86 - 298.9 MYA). As expected, the progressively increasing similarity reaches the highest point with the appearance of mammals that, from *Primates* to *Rodentia* and *Carnivora,* all showed high scores. The corresponding burst in our arrow of time is in the late Cretaceous (143.1 - 66 MYA), reportedly containing one of the three ecological diversification peaks of early mammals, specifically the upper cretaceous radiation. It was likely caused by a rebound in diversity from small insectivore ancestors after that the cretaceous terrestrial revolution (characterized by the rise of flowering plants) had caused a massive ecological bottleneck. [291] This period, as well as the preceding phase of vertebrate evolution, may be linked to the adaptation of CHI3L1 as an immunoregulatory molecule within the increasingly complex immune systems of these organisms.

[4] https://stratigraphy.org/chart

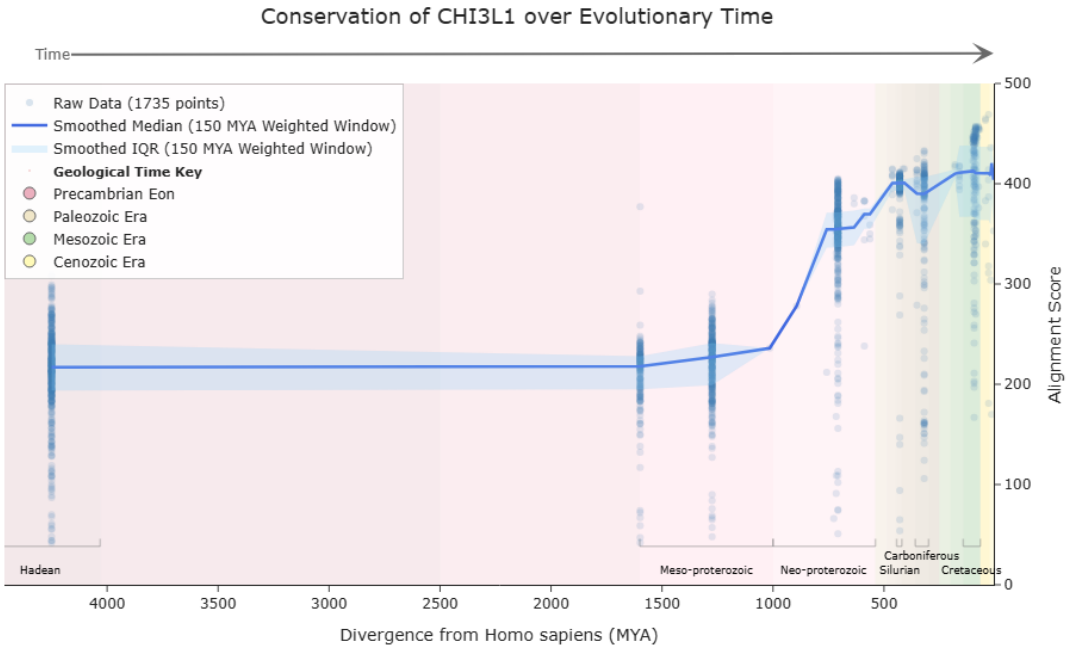
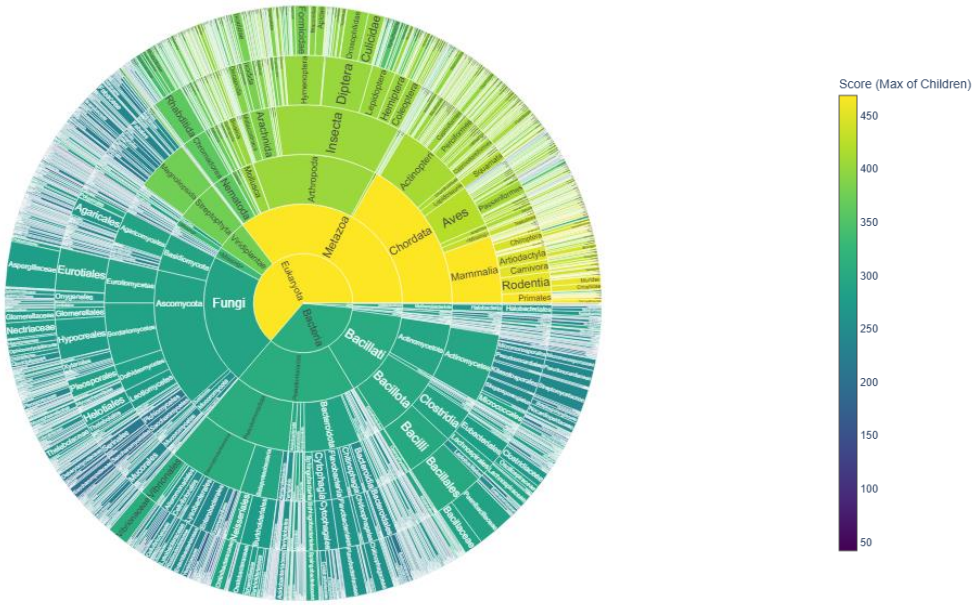


Figure 68. *Arrow of time representation of the divergence from Homo sapiens of the organisms whose sequences show similarity to human CHI3L1.*

For each species, the highest scoring sequence was retained and the evolutionary divergence time was calculated from TimeTree[5], then used as a scale of time in the plot (horizontal axis). On the vertical axis, instead, the alignment score is displayed. Therefore, the representation does not indicate the exact fossil age, but rather the time elapsed since the divergence of that organism's clade from the lineage that later evolved and specialized into *Homo sapiens*. We also conducted a clade-level enrichment analysis across geological periods using Fisher's exact test, identifying taxonomic groups that are significantly over-represented in specific evolutionary time windows based on divergence data. P-values were adjusted for multiple testing using the Benjamini-Hochberg FDR method.



[5] https://timetree.org/

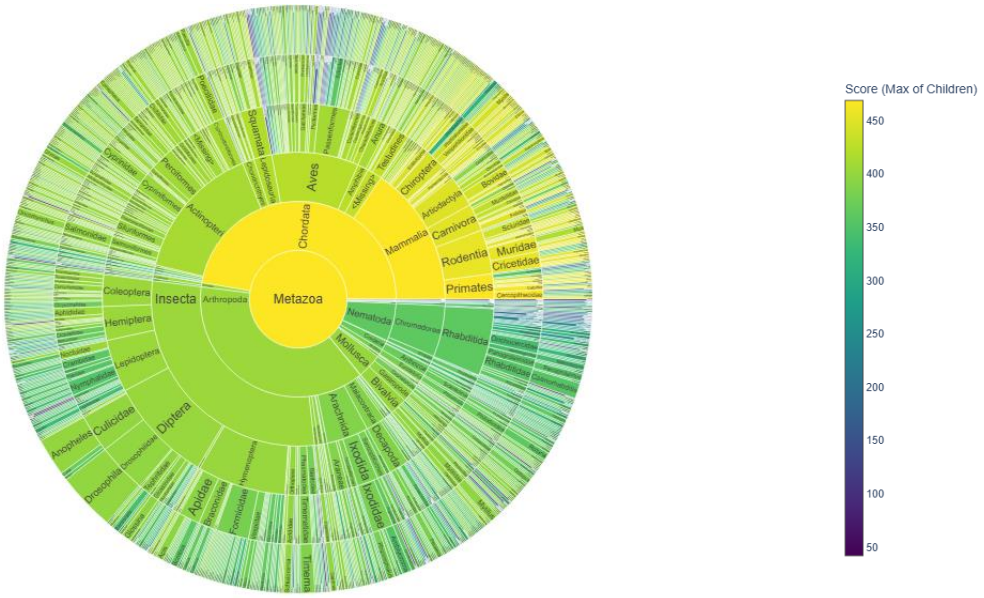


Figure 69. *Sunburst plot of the taxonomy of all the sequences aligned to CHI3L1*.
The taxonomy names were retrieved from NCBI and showed in eccentric fashion from domains at the center to species at the edge. The top plot shows that CHI3L1 aligns moderately well not only with *Metazoa*, but also with *Fungi* and *Bacteria*, both *Pseudomonadati* and *Bacillati* kingdoms. The bottom plot is focused on the *Metazoa* subkingdom and its constituents. We see that, as expected, the highest conservation is in the *Mammalia* class, but Aves, *Actinopteri*, *Insecta*, *Arachnida* and *Bivalvia* all show quite a high alignment score, color coded from blue to yellow.

4.3.2 A Modern Receptor's Rapid Rise

RAGE, in contrast, is a much more recent evolutionary invention, with the earliest extant organism that carries even a minimally similar sequence tracing its lineage back to an ancestor that diverged from Homo sapiens around 707.6 million years ago (Fig. 70). In fact, we observed that this receptor emerged mainly in mammals (Fig. 71). Only a slight conservation was observed in *Actinopteri*, *Aves* and *Arthropoda*, while the bacterial species that aligned to the query were extremely few, almost exclusively belonging to the *Pseudomonadota* phylum.

We know that RAGE appears to have evolved from an ancestral family of immunoglobulin-like cell adhesion molecules, repurposing this scaffold to become a master sensor of endogenous danger signals in the innate immune system. The associated period of evolutionary expansion may correspond to the peak of mammalian diversification during the Cretaceous Period, as illustrated in Fig. 70, although the timeline in our analysis does not align precisely with fossil record datings.

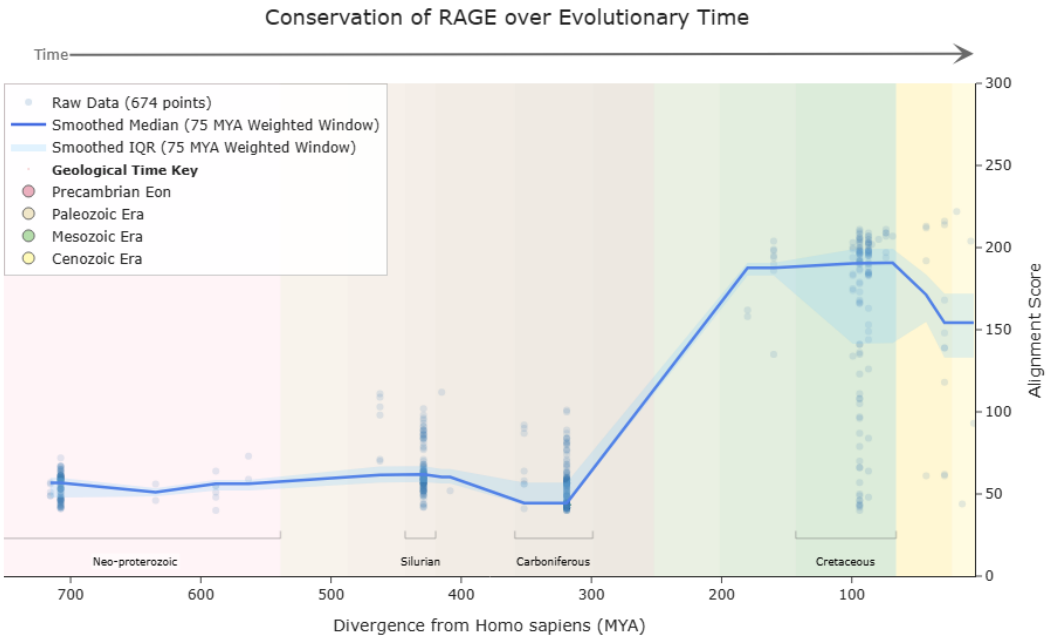
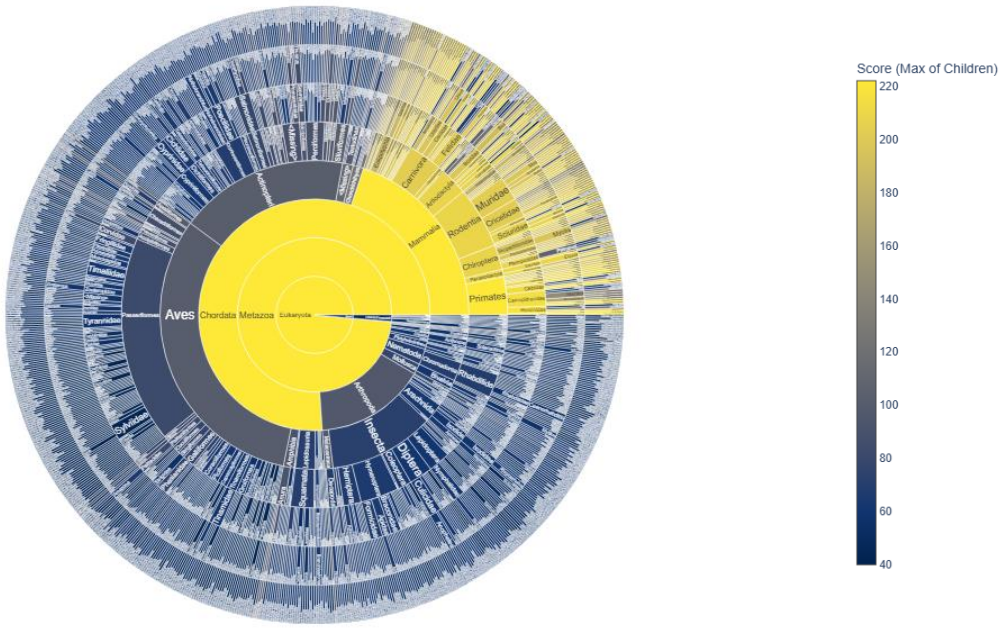


Figure 70. *Arrow of time representation of the divergence from Homo sapiens of the organisms whose sequences show similarity to human RAGE VC1.*

For each species, the highest scoring sequence was retained and the evolutionary divergence time was calculated from Time Tree, then used as a scale of time in the plot (horizontal axis). On the vertical axis, instead, the alignment score is displayed. Therefore, the representation does not indicate the exact fossil age, but rather the time elapsed since the divergence of that organism's clade from the lineage that later evolved and specialized into *Homo sapiens*. Clade-level enrichment across geological periods was performed using Fisher's exact test, with FDR correction to identify taxa over-represented in specific evolutionary time windows.



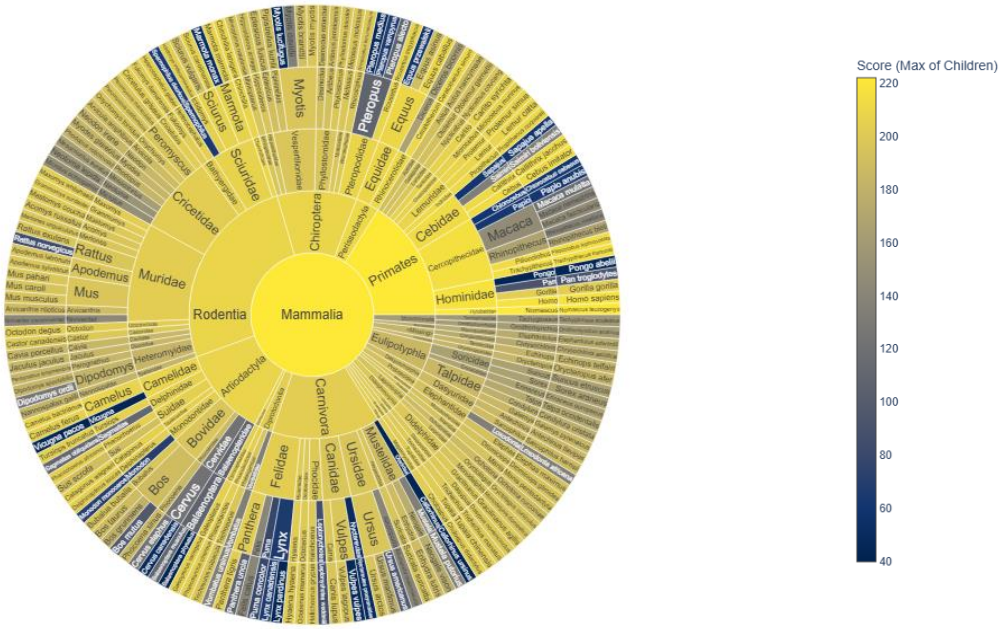


Figure 71. *Sunburst plot of the taxonomy of all the sequences aligned to RAGE VC1*.
The taxonomy names were retrieved from NCBI and showed in eccentric fashion from domains at the center to species at the edge. The top plot shows that *Mammalia* have a very high similarity, while only *Aves*, *Actinopteri* and *Arthropoda* retain a moderate similarity, while the rest of the life domains are basically absent from the alignment, meaning that the protein almost only evolved in *Metazoa*. The bottom plot is focused on the *Mammalia* and its constituents.

To further investigate the evolutionary tree of RAGE and its relationships with the history of how the domains changed, we computed and drew the phylogenetic tree (Fig. 72) containing the sequences coming from the MSA and realigned with each other with Clustal. Then, we selected the most informative nodes that would separate the sequences in major phylogenetic splits. The first node appeared to exactly separate ray-finned fishes (*Actinopteri*) from tetrapods (including mammals, reptiles and amphibians).

To analyze the two branches of this identified node, we calculated their coevolution matrices (Fig. 73), both with respect to the amino acids belonging to the query only, and also on all the alignment (including insertions). The first branch, identified as Child A and corresponding to ray-finned fishes, displays a dense co-evolution block across residues 22-120 that traces the β-strands of the V-domain and extends into the hinge (160-210), indicating that the V/C1 interface in ray-finned fishes remains under active compensatory evolution: multiple lineages adapted to very different osmolytes, pathogens, and temperatures. Coevolutionary methods, in fact, reveal exactly where the protein is still exploring sequence space: bright tiles correspond to positions that tolerate, and must compensate for, ongoing mutations, whereas dark tiles expose segments that have already crystallized under strong purifying selection. Strikingly, a second hotspot at residues 211-225 pinpoints the C1-domain FG loop. This loop, which binds endogenous glycans like heparan sulfate and that our AlphaFold multimer predicts to insert into the glycan-binding groove of CHI3L1, co-mutates both within itself and with the V-core, suggesting it was still being sculpted in early marine vertebrates.

By contrast, Child B, the tetrapod clade, is almost co-evolutionary silent: tetrapod sequences show near-identity in both the V-core and FG loop, consistent with strong purifying selection once the receptor functions became fixed in terrestrial vertebrates. This evolutionary constraint likely reflects the emergence of more specific immune roles for RAGE, coupled with increased endogenous AGE stress driven by higher oxygen tension, elevated body temperature, and greater mechanical and inflammatory load associated with life on land. The residual signals around residue 70 in Child B (but absent in ray-finned fish) may reflect adaptive tuning of RAGE's ability to multimerise in response to DAMPs. These insertions likely altered the electrostatics, hydrophobicity, or conformational dynamics of the N-terminal loop, influencing how RAGE monomers pack against each other or against multivalent ligands like S100A9.

Moreover, regarding the modularity of the receptor domains, the presence of cross-domain signal only in ray-finned fishes suggests RAGE in that clade relied more on allosteric crosstalk between V and C1, whereas tetrapods lock the hinge in one optimal conformation, which mirrors the structural data of human RAGE V-C1 interface being relatively rigid.

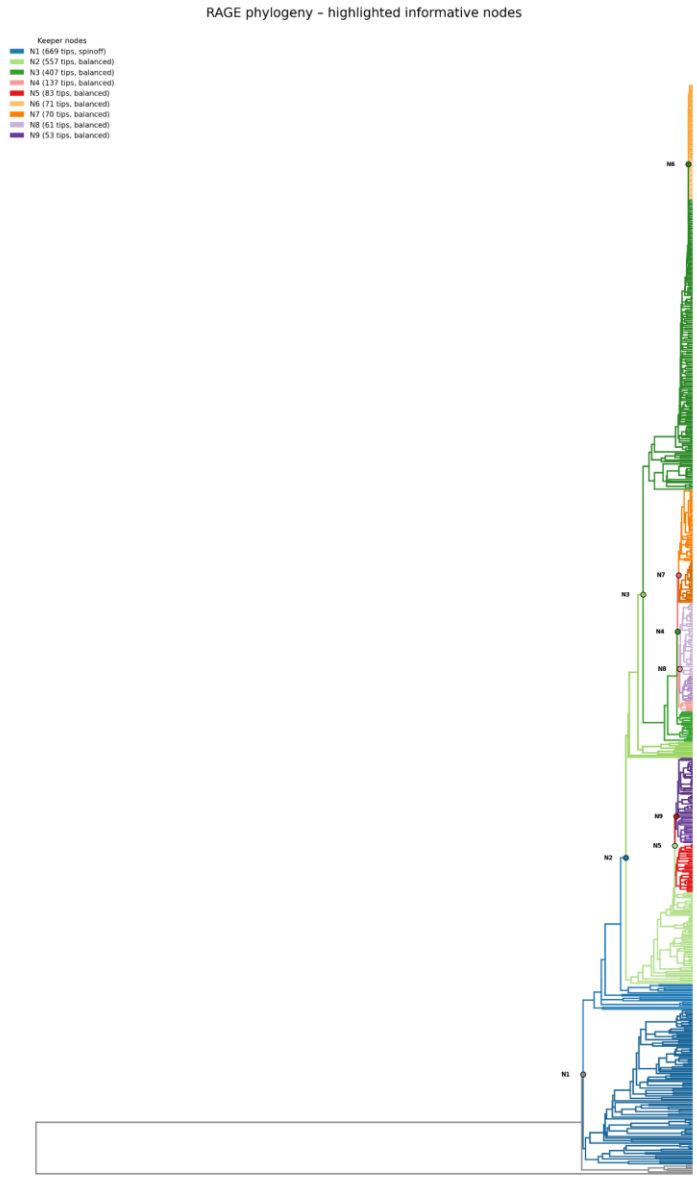


Figure 72. *Evolutionary phylogenetic tree of RAGE MSA sequences*.

The tree construction was based on the maximum-likelihood phylogeny, which finds the statistically best-supported hypothesis for how a set of sequences evolved, based on a model of how substitutions happen over time. To pinpoint evolutionarily informative splits we applied a balanced-divergence node picker: for every internal clade we required ≥ 50 sequences in total and both daughter branches to hold between 25 % and 75 % of those sequences. Node 1 passed these filters, separating 139 ray-finned fish sequences (Child A) from 417 tetrapod sequences (Child B).

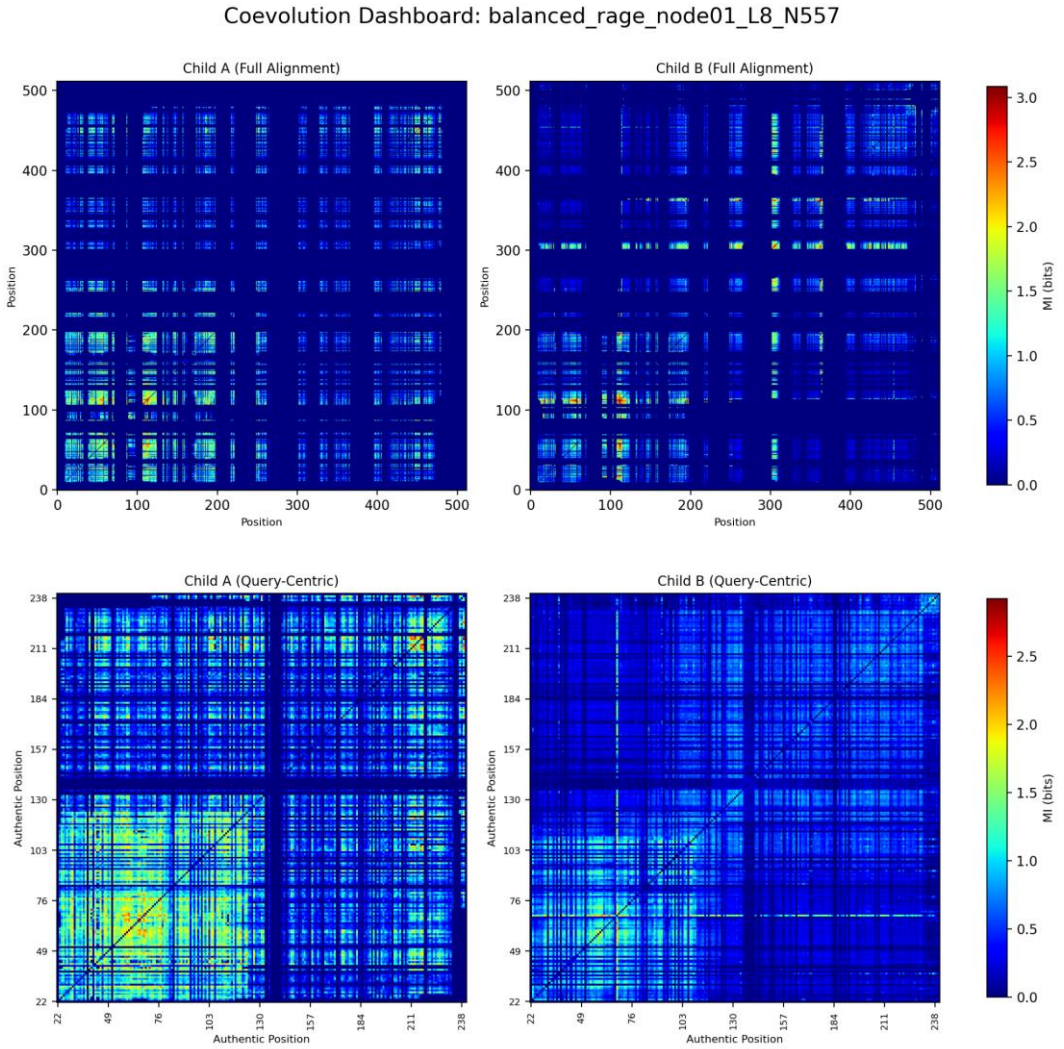


Figure 73. *Coevolutionary matrices of the children of node 1 based on mutual information as the numerical metric of residue-pair co-variation.*

To investigate how evolutionary constraints differ across the two first major branches of the RAGE phylogeny, we computed coevolution heat-maps for the two immediate children of node 1 of the tree. For each clade, we locally rebuilt the alignment with Clustal-Omega using only the sequences in that branch and calculated mutual information (MI) between all pairs of alignment columns. Two versions of the coevolution matrix were generated per child: one based on the full alignment space, including lineage-specific insertion and allowing for direct comparison across clades, and one “query-centric” version that included only those positions where the human RAGE query has a residue, thus excluding insertions and allowing for direct comparison with the domains of our reference protein. To reduce background noise arising from differences in overall conservation or sampling density, we applied Average Product Correction (APC) to all matrices. This correction subtracts a statistical baseline from each MI value, helping to isolate meaningful co-variation from generic variability.

In Child A (ray-finned fishes), the dense coevolution blocks across residues 22–120 (V domain) and residues 210-220 (FG loop) are confirmed in the full alignment, meaning that they are not an artifact of the query-centric alignment. In child B (tetrapods), instead the presence of a bright stripe around residue 300 that is not reflected in the query-centric alignment, might reflect a functional insertion in some of the sub-clades but not present in humans.

4.3.3 The Evolutionary Convergence

These two proteins, therefore, followed separate evolutionary paths for hundreds of millions of years before their functions became entangled. CHI3L1 is an ancient, repurposed scissors; RAGE is a modern, specialized receptor. This temporal mismatch is the key to understanding their interaction. RAGE did not evolve to recognize chitin. It evolved to recognize the protein that recognizes chitin. This is a story of evolutionary co-option: faced with the need for a mechanism to brake its own inflammatory signaling, the immune system repurposed CHI3L1, a protein already deeply embedded in inflammatory and tissue-remodeling pathways. The next section will explore the precise logic of this co-option cascade.

4.4 When the Ancient Co-evolves with the New: Evolutionary Logic of the CHI3L1-RAGE Checkpoint

The integrated biological, structural and taxonomic findings of this thesis present a fascinating evolutionary puzzle: *why* would a modern immune receptor like RAGE evolve to engage the ancient, 'non-self' glycan-binding groove of CHI3L1? What about the other 'self' recognizing site(s)? The answers lie not in two independent histories, but in a multi-step cascade of co-option and co-evolution, where an ancient molecular tool was repeatedly repurposed for new, more sophisticated functions, and became entangled in evolution with its new partners of interaction.

Such a process follows the powerful principle of *molecular economy*: evolution is "lazy"; it prefers to economically adapt existing tools rather than invent new ones from scratch. In this case, a checkpoint for innate immunity was created, at the intersection of two evolving and - at this exact contact point - coevolving worlds of both independent and mutual functional relevance.

4.4.1 The Ancestral Tool: A Chitinase

The story begins with the primordial CHI3L1 ancestor, a member of the ancient GH18 family whose function was to bind and degrade chitin, a classic innate immunity role. The 43 Å groove is, in its arrangement, this ancient tool.

4.4.2 The First Co-option: The Sensor

The pivotal first shift was the mutation of key catalytic residues, as shown in Fig. 74. This repurposed the protein from a scissors into a high-fidelity sensor. The wide groove could now still recognize its targets, but instead of degrading them, it became expert at sensing them, making it a *de facto* Pattern Recognition Receptor (PRR).

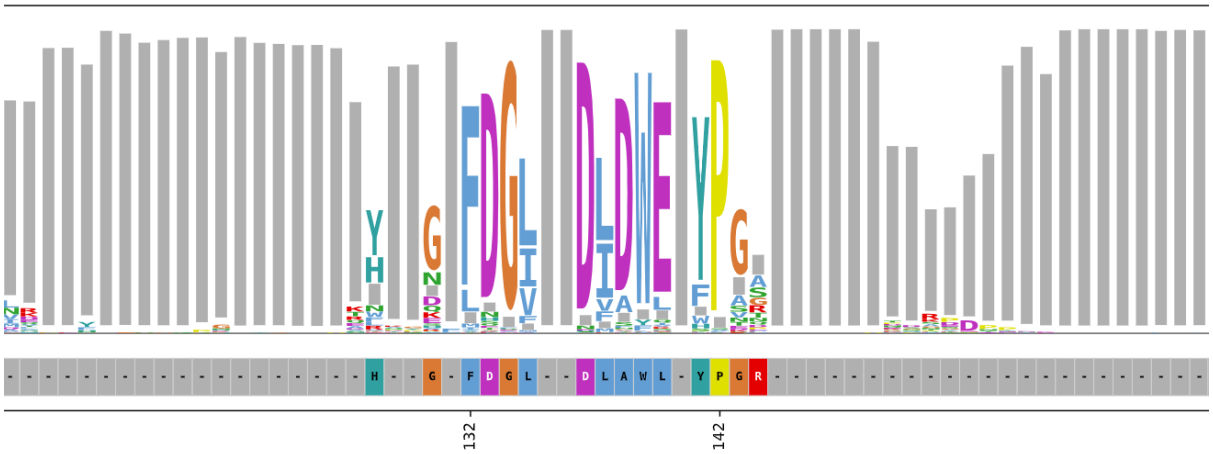


Figure 74. *Part of the consensus sequence of CHI3L1 computed on the aligned sequences from the MSA.* The consensus sequence displays the most conserved residues (color coded by function with ClustalX legend) as the biggest in each position, while gaps are represented by grey ticks. below the consensus, the original query sequence is displayed. We can observe that the catalytic residue, the DxDxE triad from position 136 to 140, is conserved across all sequences that pair with CHI3L1, presumably including many from the GH18 hydrolase family. [194] This site is instead mutated in two key residues (the second aspartic acid at position 138 substituted to an alanine and the glutamic acid at position 140 substituted to a leucine) in the original CHI3L1 query sequence, losing its catalytic activity.

4.4.3 The Second Co-option: The Host Integrator

Evolution layered on a second function by adding the C-terminal heparin-binding domain. This gave CHI3L1 the ability to interact and talk with host receptors decorated with proteoglycans, such as syndecan-1. Another heparin consensus motif appeared, and - despite its lack of affinity for heparin, as is - its allosteric relationship with the groove and its motif imply possible additional roles.

CHI3L1 was no longer a simple pattern detector but a sophisticated translator.

4.4.4 The Final Co-option: The Checkpoint

Both self and non-self sensing are useless unless the protein is in turn sensed by some other signaling molecule. Indeed, all co-options may have been parallel to each other, or may have originated through intermediate states and roles. Yet, the interaction with RAGE is the third and most elegant repurposing.

The immune system needed a mechanism to regulate the potent inflammatory signaling of RAGE in a context-dependent way, especially to avoid immune system overactivation in the presence of commensal microbes. The hard way would have been for RAGE to evolve its own domains to directly sense the context and read novel specific molecular patterns (i.e. carbohydrates) associated with tissue remodeling or microbial stress. But RAGE was already expert at detecting other, mostly proteomic signals; it was highly charged and would have had a hard time recognizing mildly polar sugars.

Instead of teaching RAGE a new language, evolution taught it how to recognize the translator (CHI3L1). By evolving a protein-protein interaction, RAGE effectively outsourced its sensing function, leveraging the pre-existing, high-fidelity detection machinery of CHI3L1. This is a classic evolutionary shortcut, where a complex function is achieved through a simple, efficient adaptation.

By evolving a flexible loop that lifts Trp99 and mimics a ligand for CHI3L1's ancient groove, RAGE learned to be regulated by this master sensor. This act created a dominant-negative, inhibitory interaction - an innate immunity checkpoint. When symbiotic healthy bacteria are present, such as in the gut, CHI3L1 sees them through peptidoglycan and inhibits the immune response and excessive inflammation. However, in a sterile environment like the tumor, CHI3L1 acts as a protumorigenic Type 2 immunomodulator.

4.4.5 A Doubly-Constrained Groove

This evolutionary model provides the most satisfying explanation for why the CHI3L1 groove is so highly conserved despite lacking enzymatic activity. Its structure is not under one, but two independent and mutually reinforcing selective pressures.

The old view held that the groove is conserved simply to bind microbial glycans. This is true, but incomplete; it fails to explain why RAGE would then need to interact with it. The new view, supported by our model, is that the groove's architecture is doubly-constrained.

First, it must retain its ancestral shape to bind non-self microbial glycans like chitin and peptidoglycan with high fidelity. Second, its same "ligand-bound" or "ligand-ready" conformation mediated by Trp99 is precisely what the RAGE loop has co-evolved to recognize. The proteins did not evolve in isolation but within the same inflammatory microenvironments, creating a reciprocal co-evolutionary pressure. Any mutation in the groove would now be selected against for compromising *both* its ability to sense microbes and its ability to regulate the host's immune response via RAGE. This dual constraint locks the groove's architecture in place, making it a remarkably stable and reliable signaling interface.

This dual function also explains the context-dependent nature of CHI3L1's activity, framing it as a molecular switch governed by the occupancy of its ancient groove. The state of this groove allosterically dictates heparan sulfate binding and therefore, possibly, function of the C-terminal hub, leading to two distinct and opposing biological outcomes:

- In a microbially rich environment (e.g., the gut): the groove is predominantly occupied by ligands like peptidoglycan. Based on the allosteric inhibition observed by Magnusdottir et al., this binding state would likely suppress the pro-angiogenic activity

of the C-terminal site[6]. Here, CHI3L1 acts as a homeostatic regulator, promoting tolerance to commensal flora while keeping its own tissue-proliferative functions in check.

- ➢ In a sterile, inflammatory environment (e.g., a tumor): the groove is unoccupied by microbial ligands. It is therefore free to engage the inhibitory loop of RAGE, executing its immunosuppressive checkpoint function. Simultaneously, the C-terminal hub is uninhibited and free to bind heparan sulfates on host cells, driving the pro-tumorigenic angiogenesis and proliferation program.

This paints a picture of CHI3L1 as a molecular palimpsest whose ancient history of sensing the microbial world remains inextricably linked to its modern, decisive role in the landscape of cancer and immunity.

4.5 Limitations and Methodological Considerations

This study, while providing significant new insights, has inherent limitations. Our structural conclusions are based on a high-confidence computational model, not an experimental structure. Despite extensive screening at two synchrotron facilities, we were unable to crystallize the CHI3L1-RAGE complex, obtaining crystals of CHI3L1 alone. This underscores the challenge of crystallizing transient or dynamic complexes and highlights the critical role that advanced modeling now plays in bridging these experimental gaps.

Furthermore, our investigation into the role of glycosylation was constrained by the partial efficacy of enzymatic deglycosylation under native conditions, particularly for RAGE with its heterogeneous glycoforms. While our ELISA results hint that partial deglycosylation might even enhance binding, a definitive conclusion requires a more systematic approach using proteins expressed in systems that allow for complete absence of glycosylation (e.g., prokaryotic cell lines or *in vitro* glycosylation).

4.6 Future Perspectives

The molecular blueprint of the CHI3L1-RAGE interface generated in this thesis provides a direct and rational path toward therapeutic development. The immediate next steps should focus on the experimental validation of our model and the translation of these structural insights into functional inhibitors.

The highest priority is the experimental validation of the predicted interface. This can be achieved through site-directed mutagenesis of key residues identified in our model (e.g., Arg216 on RAGE; Trp99 on CHI3L1). Abrogation of binding in surface plasmon resonance (SPR) or ELISA assays following these mutations would provide strong evidence for the

[6] This fascinating deduction is also testable.

predicted binding mode. Functionally, these mutations should be tested in NK cell co-culture assays to confirm that disrupting the interaction restores trastuzumab-mediated cytotoxicity.

Concurrently, screening for optimal buffers to crystallize the complex or attempting to solve its cryo-electron microscopy (Cryo-EM) structure remains a cardinal goal to provide definitive, high-resolution atomic coordinates.

Our model exposes the RAGE FG-loop as the critical recognition motif. This opens two avenues for inhibitor design. First, the development of peptidomimetics or small molecules that mimic this loop could act as direct competitive inhibitors. The fluorescently-labeled peptide synthesized in this work represents the first tool for developing a quantitative binding assay to screen for such compounds. Second, the structure provides a template for engineering therapeutic protein decoys. A soluble RAGE variant could be engineered with mutations that enhance its affinity for CHI3L1 while simultaneously ablating its binding sites for pro-inflammatory ligands like S100 proteins, creating a highly specific CHI3L1-trap.

The MSA analysis should also be expanded and deepened to better understand the specific evolution of single domains, especially with respect to each other in a protein and, even more importantly, with the interacting residues of the binding partner.

5. Conclusion

The central challenge of this thesis was to find a point of entry into the complexity of therapeutic resistance in $HER2^+$ breast cancer. We began with a hypothesis: that the interaction between CHI3L1 and RAGE represents a critical, yet poorly understood, checkpoint of innate immunity hijacked by tumors to suppress NK cell-mediated cytotoxicity.

By moving beyond the phenomenological description, this work has provided the first structural-level decoding of this innate immunity checkpoint, revealing not only its architecture but also the deep evolutionary logic that might govern its function.

Our integrated approach, converging biochemical analysis with advanced computational modeling, yielded a high-confidence atomic model of the CHI3L1-RAGE complex. This revealed a precise and elegant architecture: the modern RAGE receptor inserts a flexible loop into the ancient, glycan-binding groove of CHI3L1. This structural template provided a direct, physically plausible interpretation of the mechanism for the checkpoint's immunosuppressive action: the steric disruption of RAGE oligomerization. The model showed how monomeric CHI3L1, by docking into the RAGE VC1 domain, acts as a molecular blocker, preventing the assembly of the receptor platform required for pro-cytotoxic signaling. This single, elegant mechanism unifies the diverse, context-dependent effects of CHI3L1 across a range of pathologies and provides a satisfying explanation for its competitive antagonism with pro-inflammatory ligands like S100A8/A9.

Yet, the most captivating insight emerged when this structure was viewed through the lens of evolution. The model fuelled a remarkable story of co-option, where an ancient, catalytically inert chitin-binding groove was repurposed to become the docking site for a modern immune receptor. This theory, if confirmed, would resolve the long-standing puzzle of the groove's conservation, demonstrating that its structure is under a dual selective pressure: it must retain its ancestral ability to sense microbial patterns while simultaneously serving as the high-fidelity interface for its new, co-evolved regulatory partner, RAGE.

The structural and evolutionary understanding achieved here lays a direct and logical foundation for the development of novel therapeutics aimed at reawakening the immune system. The journey of this work, from a clinical problem of resistance to a molecular path for intervention, thus supports the synergy of modern computational and experimental science to turn complexity into clarity.

Acknowledgements

We would like to thank the many people who have supported us throughout this long and challenging journey.
First and foremost, we are deeply grateful to Professor Maria Rescigno, our mentor and guide in academic life. Without her trust, insight, and leadership, this project would never have come into existence.
We also sincerely thank our supervisor at the Politecnico di Milano, Professor Maria Laura Costantino, the first President of the MEDTEC program, of which we are proud to be part as its inaugural cohort.
A special thank-you goes to our co-supervisors, Dr. Luigi Angelo Scietti and Dr. Marina Mapelli. We would also like to include Dr. Giuseppe Ciossani, who, although not formally a co-supervisor, was one in practice. All three of them provided us with remarkable scientific and personal guidance, supporting us step by step through the entire experimental process.
We are especially grateful to Dr. Cissy Lynette Tarver, who welcomed us into her lab as if we were her own, dedicating time and care well beyond what a professional relationship would require. We extend our warmest thanks to researchers Dr. Aina Cohen and Dr. Clyde Smith, who received us with great kindness from the very beginning and openly shared their research work with us.
We would like to express our profound gratitude to Professor David Bermudes, who was our very first guide in a research laboratory. He has been a true mentor, both scientifically and personally, a special person to whom we owe a great deal and to whom we are deeply connected.
We would like to thank Dr. Giuseppe Penna, with whom we shared many scientific discussions that greatly enriched the cultural and experimental foundation of this thesis. We are equally grateful to Dr. Michela Lizier, a true cornerstone for us during our time in the Humanitas research lab. If we were able to complete any experiments, it was largely thanks to her guidance and constant support.
Finally, we thank all the members of Professor Rescigno's lab at Humanitas, who were always kind and welcoming to us throughout our time there.

Bibliography & Sitography

1. Xiong, X. et al. (2025) ‘Breast cancer: pathogenesis and treatments’, Signal Transduction and Targeted Therapy, 10(1). doi: 10.1038/s41392-024-02108-4

2. Kim, J. et al. (2025) ‘Global patterns and trends in breast cancer incidence and mortality across 185 countries’, Nature Medicine, 31(4), pp. 1154–1162. doi: 10.1038/s41591-025-03502-3

3. https://gco.iarc.who.int/media/globocan/factsheets/populations/900-world-fact-sheet.pdf, accessed June 2025

4. https://gco.iarc.who.int/media/globocan/factsheets/populations/380-italy-fact-sheet.pdf, accessed June 2025

5. International Agency for Research on Cancer, World Health Organization. Age-standardized rate (world) per 100 000, incidence, both sexes, in 2022. Cancer Today https://gco.iarc.fr/today/en/dataviz/maps-heatmap (2024). accessed June 2025

6. https://www.who.int/initiatives/global-breast-cancer-initiative, accessed June 2025

7. Jonsson, P. et al. (2019) ‘Tumour lineage shapes BRCA-mediated phenotypes’, Nature, 571(7766), pp. 576–579. doi: 10.1038/s41586-019-1382-1

8. Pal, M., Das, D. and Pandey, M. (2024) ‘Understanding genetic variations associated with familial breast cancer’, World Journal of Surgical Oncology, 22(1). doi: 10.1186/s12957-024-03553-9

9. García-Sancha, N., Corchado-Cobos, R. and Pérez-Losada, J. (2025) ‘Understanding Susceptibility to Breast Cancer: From Risk Factors to Prevention Strategies’, International Journal of Molecular Sciences, 26(7), p. 2993. doi: 10.3390/ijms26072993

10. Park, D.J. et al. (2022) ‘Assessment of risks for breast cancer in a flight attendant exposed to night shift work and cosmic ionizing radiation: a case report’, Annals of Occupational and Environmental Medicine, 34(1). doi: 10.35371/aoem.2022.34.e5

11. Haus, E.L. and Smolensky, M.H. (2013) ‘Shift work and cancer risk: Potential mechanistic roles of circadian disruption, light at night, and sleep deprivation’, Sleep Medicine Reviews, 17(4), pp. 273–284. doi: 10.1016/j.smrv.2012.08.003

12. Harbeck, N. et al. (2019) ‘Breast cancer’, Nature Reviews Disease Primers, 5(1). doi: 10.1038/s41572-019-0111-2

13. Zagami, P. and Carey, L.A. (2022) ‘Triple negative breast cancer: Pitfalls and progress’, npj Breast Cancer, 8(1). doi: 10.1038/s41523-022-00468-0

14. Nolan, E., Lindeman, G.J. and Visvader, J.E. (2023) ‘Deciphering breast cancer: from biology to the clinic’, Cell, 186(8), pp. 1708–1728. doi: 10.1016/j.cell.2023.01.040

15. Will, M. et al. (2023) ‘Therapeutic resistance to anti-oestrogen therapy in breast cancer’, Nature Reviews Cancer, 23(10), pp. 673–685. doi: 10.1038/s41568-023-00604-3

16. Cheang, M.C.U. et al. (2009) ‘Ki67 Index, HER2 Status, and Prognosis of Patients With Luminal B Breast Cancer’, JNCI: Journal of the National Cancer Institute, 101(10), pp. 736–750. doi: 10.1093/jnci/djp082

17. Ruiz-Saenz, A. et al. (2018) ‘HER2 Amplification in Tumors Activates PI3K/Akt Signaling Independent of HER3’, Cancer Research, 78(13), pp. 3645–3658. doi: 10.1158/0008-5472.CAN-18-0430

18. Swain, S.M., Shastry, M. and Hamilton, E. (2022) ‘Targeting HER2-positive breast cancer: advances and future directions’, Nature Reviews Drug Discovery, 22(2), pp. 101–126. doi: 10.1038/s41573-022-00579-0

19. Alluri, P. and Newman, L.A. (2014) ‘Basal-Like and Triple-Negative Breast Cancers’, Surgical Oncology Clinics of North America, 23(3), pp. 567–577. doi: 10.1016/j.soc.2014.03.003

20. Arnedos, M. et al. (2012) ‘Triple-negative breast cancer: are we making headway at least?’, Therapeutic Advances in Medical Oncology, 4(4), pp. 195–210. doi: 10.1177/1758834012444711

21. Wang, J. and Wu, S.-G. (2023) ‘Breast Cancer: An Overview of Current Therapeutic Strategies, Challenge, and Perspectives’, Breast Cancer: Targets and Therapy, Volume 15, pp. 721–730. doi: 10.2147/BCTT.S432526

22. Zagami, P. and Carey, L.A. (2022) ‘Triple negative breast cancer: Pitfalls and progress’, npj Breast Cancer, 8(1). doi: 10.1038/s41523-022-00468-0. License avail. at https://creativecommons.org/licenses/by/4.0/

23. Zabel, B.U. et al. (1984) ‘Cellular homologs of the avian erythroblastosis virus erb-A and erb-B genes are syntenic in mouse but asyntenic in man.’, Proceedings of the National Academy of Sciences, 81(15), pp. 4874–4878. doi: 10.1073/pnas.81.15.4874

24. Iqbal, Nida and Iqbal, Naveed (2014) ‘Human Epidermal Growth Factor Receptor 2 (HER2) in Cancers: Overexpression and Therapeutic Implications’, Molecular Biology International, 2014, pp. 1–9. doi: 10.1155/2014/852748

25. Brandt-Rauf, P.W., Pincus, M.R. and Carney, W.P. (1994) ‘The c-erbB-2 Protein in Oncogenesis: Molecular Structure to Molecular Epidemiology’, Critical Reviews™ in Oncogenesis, 5(2–3), pp. 313–329. doi: 10.1615/critrevoncog.v5.i2-3.100

26. Pan, L. et al. (2024) ‘HER2/PI3K/AKT pathway in HER2-positive breast cancer: A review’, Medicine, 103(24), p. e38508. doi: 10.1097/MD.0000000000038508

27. Sliwkowski, M.X. (2003) ‘Ready to partner’, Nature Structural & Molecular Biology, 10(3), pp. 158–159. doi: 10.1038/nsb0303-158

28. Santhanakrishnan, J., Meganathan, P. and Vedagiri, H. (2024) 'Structural biology of HER2/ERBB2 dimerization: mechanistic insights and differential roles in healthy versus cancerous cells', Exploration of Medicine, pp. 530–543. doi: 10.37349/emed.2024.00237

29. Moasser, M.M. (2022) 'Inactivating Amplified HER2: Challenges, Dilemmas, and Future Directions', Cancer Research, 82(16), pp. 2811–2820. doi: 10.1158/0008-5472.CAN-22-1121

30. Zhong, H. et al. (2024) 'The Biological Roles and Clinical Applications of the PI3K/AKT Pathway in Targeted Therapy Resistance in HER2-Positive Breast Cancer: A Comprehensive Review', International Journal of Molecular Sciences, 25(24), p. 13376. doi: 10.3390/ijms252413376

31. License available at https://creativecommons.org/licenses/by/4.0/

32. https://www.accessdata.fda.gov/drugsatfda_docs/label/1998/trasgen092598lb.pdf, accessed June 2025

33. Maadi, H. et al. (2021) 'Trastuzumab Mechanism of Action; 20 Years of Research to Unravel a Dilemma', Cancers, 13(14), p. 3540. doi: 10.3390/cancers13143540

34. Bradley, R. et al. (2021) 'Trastuzumab for early-stage, HER2-positive breast cancer: a meta-analysis of 13 864 women in seven randomised trials', The Lancet Oncology, 22(8), pp. 1139–1150. doi: 10.1016/S1470-2045(21)00288-6

35. Hudis, C.A. (2007) 'Trastuzumab — Mechanism of Action and Use in Clinical Practice', New England Journal of Medicine, 357(1), pp. 39–51. doi: 10.1056/nejmra043186

36. Fendly, B M et al. (1990) 'Characterization of murine monoclonal antibodies reactive to either the human epidermal growth factor receptor or HER2/neu gene product.' Cancer research vol. 50,5:1550-8

37. https://www.accessdata.fda.gov/drugsatfda_docs/label/2012/125409lbl.pdf, accessed June 2025

38. Nami, B., Maadi, H. and Wang, Z. (2018) 'Mechanisms Underlying the Action and Synergism of Trastuzumab and Pertuzumab in Targeting HER2-Positive Breast Cancer', Cancers, 10(10), p. 342. doi: 10.3390/cancers10100342

39. Phillips, G.D.L. et al. (2014) 'Dual Targeting of HER2-Positive Cancer with Trastuzumab Emtansine and Pertuzumab: Critical Role for Neuregulin Blockade in Antitumor Response to Combination Therapy', Clinical Cancer Research, 20(2), pp. 456–468. doi: 10.1158/1078-0432.ccr-13-0358

40. Swain, S.M. et al. (2020) 'Pertuzumab, trastuzumab, and docetaxel for HER2-positive metastatic breast cancer (CLEOPATRA): end-of-study results from a double-blind, randomised, placebo-controlled, phase 3 study', The Lancet Oncology, 21(4), pp. 519–530. doi: 10.1016/s1470-2045(19)30863-0

41. Gianni, L. et al. (2012) 'Efficacy and safety of neoadjuvant pertuzumab and trastuzumab in women with locally advanced, inflammatory, or early HER2-positive breast cancer (NeoSphere): a randomised multicentre, open-label, phase 2 trial', The Lancet Oncology, 13(1), pp. 25–32. doi: 10.1016/S1470-2045(11)70336-9

42. https://www.roche.com/media/releases/med-cor-2025-05-13, accessed 23 June 2025

43. Hurvitz, S.A. et al. (2018) 'Neoadjuvant trastuzumab, pertuzumab, and chemotherapy versus trastuzumab emtansine plus pertuzumab in patients with HER2-positive breast cancer (KRISTINE): a randomised, open-label, multicentre, phase 3 trial', The Lancet Oncology, 19(1), pp. 115–126. doi: 10.1016/S1470-2045(17)30716-7

44. von Minckwitz, G. et al. (2017) 'Adjuvant Pertuzumab and Trastuzumab in Early HER2-Positive Breast Cancer', New England Journal of Medicine, 377(2), pp. 122–131. doi: 10.1056/NEJMoa1703643

45. https://dailyreporter.esmo.org/esmo-breast-cancer-2025/latest-news/adding-adjuvant-pertuzumab-demonstrates-prolonged-survival-benefit-after-10-years-in-her2-breast-cancer, accessed 23 June 2025.

46. Baselga, J. et al. (2012) 'Pertuzumab plus Trastuzumab plus Docetaxel for Metastatic Breast Cancer', New England Journal of Medicine, 366(2), pp. 109–119. doi: 10.1056/NEJMoa1113216

47. Nahta, R., Yu, D., Hung, MC. et al. (2006) 'Mechanisms of Disease: understanding resistance to HER2-targeted therapy in human breast cancer.' Nat Rev Clin Oncol 3, 269–280.

48. Swain, S.M. et al. (2015) 'Pertuzumab, Trastuzumab, and Docetaxel in HER2-Positive Metastatic Breast Cancer', New England Journal of Medicine, 372(8), pp. 724–734. doi: 10.1056/NEJMoa1413513

49. https://www.accessdata.fda.gov/drugsatfda_docs/label/2013/125427lbl.pdf, accessed June 2025

50. https://www.fda.gov/drugs/resources-information-approved-drugs/fda-approves-ado-trastuzumab-emtansine-early-breast-cancer, accessed June 2025

51. Zeng, B. et al. (2019) 'Anticancer effect of the traditional Chinese medicine herb Maytenus compound via the EGFR/PI3K/AKT/GSK3β pathway.' Translational Cancer Research, 8(5), 2130-2140. doi:10.21037/tcr.2019.09.30

52. Clarice, C.V. et al. (2017) 'Pharmacological potential of Maytenus species and isolated constituents, especially tingenone, for treatment of painful inflammatory diseases' Revista Brasileira de Farmacognosia, Volume 27, Issue 4, p533-540, doi:10.1016/j.bjp.2017.02.006.

53. https://www.accessdata.fda.gov/drugsatfda_docs/label/2021/761139s011lbl.pdf, accessed June 2025

54. https://www.onclive.com/view/second-line-t-dxd-data-remain-strong-in-her2-metastatic-breast-cancer, accessed 23 June 2025

55. Giugliano, F. et al. (2023) 'Unlocking the Resistance to Anti-HER2 Treatments in Breast Cancer: The Issue of HER2 Spatial Distribution.' Cancers (Basel). 2023;15(5):1385. doi:10.3390/cancers15051385

56. Shi, Y. et al. (2014) 'Engagement of immune effector cells by trastuzumab induces HER2/ERBB2 downregulation in cancer cells through STAT1 activation', Breast Cancer Research, 16(2). doi: 10.1186/bcr3637

57. https://www.onclive.com/view/fda-approves-margetuximab-cmkb-plus-chemo-in-pretreated-metastatic-her2-breast-cancer, accessed 23 June 2025

58. Musolino, A. et al., (2022) 'Role of Fcγ receptors in HER2-targeted breast cancer therapy. J Immunother Cancer.' 10(1):e003171. doi:10.1136/jitc-2021-003171

59. Constantinides M., al. FCGR3A F158V alleles frequency differs in multiple myeloma patients from healthy population. Oncoimmunology. 2024;13(1):2388306. Published 2024 Aug 18. doi:10.1080/2162402X.2024.2388306

60. Conlin, A.K. et al. (2022) 'Cutaneous responses in HER2+ metastatic breast cancer: A retrospective case series of a Phase 1b study of Tucatinib, an Oral HER2-specific inhibitor in combination with Capecitabine and/or Trastuzumab in third-line or later treatment', Current Problems in Cancer: Case Reports, 7, p. 100170. doi: 10.1016/j.cpccr.2022.100170

61. Lin, N.U. et al. (2023) 'Tucatinib vs Placebo, Both in Combination With Trastuzumab and Capecitabine, for Previously Treated ERBB2 (HER2)-Positive Metastatic Breast Cancer in Patients With Brain Metastases: Updated Exploratory Analysis of the HER2CLIMB Randomized Clinical Trial.' JAMA Oncol. 9(2):197–205. doi:10.1001/jamaoncol.2022.5610

62. Xu, Z. et al. (2019) 'Novel HER2-Targeting Antibody-Drug Conjugates of Trastuzumab Beyond T-DM1 in Breast Cancer: Trastuzumab Deruxtecan(DS-8201a) and (Vic-)Trastuzumab Duocarmazine (SYD985)', European Journal of Medicinal Chemistry, 183, p. 111682. doi: 10.1016/j.ejmech.2019.111682

63. Banerji, U. et al. (2019) 'Trastuzumab duocarmazine in locally advanced and metastatic solid tumours and HER2-expressing breast cancer: a phase 1 dose-escalation and dose-expansion study', The Lancet Oncology, 20(8), pp. 1124–1135. doi: 10.1016/s1470-2045(19)30328-6

64. https://www.cancernetwork.com/view/fda-issues-crl-for-trastuzumab-duocarmazine-in-advanced-her2-breast-cancer, accessed 23 June 2025

65. Swain, S.Met al. (2023) 'Targeting HER2-positive breast cancer: advances and future directions.' Nat Rev Drug Discov 22, 101–126. doi: 10.1038/s41573-022-00579-0

66. Tolcher, A. et al. (2023) 'The evolving landscape of antibody-drug conjugates in gynecologic cancers' Cancer Treatment Reviews, Volume 116, 102546. doi: 10.1016/j.ctrv.2023.102546

67. https://www.globenewswire.com/en/news-release/2024/08/01/2923201/0/en/Zymeworks-Provides-Corporate-Update-and-Reports-Second-Quarter-2024-Financial-Results.html, accessed 23 June 2025

68. Cizkova, M. et al. (2012) 'PIK3CAmutation impact on survival in breast cancer patients and in ERα, PR and ERBB2-based subgroups', Breast Cancer Research, 14(1). doi: 10.1186/bcr3113

69. Price-Schiavi, S.A. et al. (2002) 'Rat Muc4 (sialomucin complex) reduces binding of anti-ErbB2 antibodies to tumor cell surfaces, a potential mechanism for herceptin resistance', International Journal of Cancer, 99(6), pp. 783–791. doi: 10.1002/ijc.10410

70. Dreyer, C.A. et al. (2022) 'The role of membrane mucin MUC4 in breast cancer metastasis', Endocrine-Related Cancer, 29(1), pp. R17–R32. doi: 10.1530/erc-21-0083

71. Goel, S. et al. (2016) 'Overcoming Therapeutic Resistance in HER2-Positive Breast Cancers with CDK4/6 Inhibitors', Cancer Cell, 29(3), pp. 255–269. doi: 10.1016/j.ccell.2016.02.006

72. Witkiewicz, A.K., Cox, D. and Knudsen, E.S. (2014) 'CDK4/6 inhibition provides a potent adjunct to Her2-targeted therapies in preclinical breast cancer models', Genes & Cancer, 5(7–8), pp. 261–272. doi: 10.18632/genesandcancer.24

73. Wang, X. et al. (2024) 'Recent progress of CDK4/6 inhibitors' current practice in breast cancer', Cancer Gene Therapy, 31(9), pp. 1283–1291. doi: 10.1038/s41417-024-00747-x

74. Hunter, F.W. et al. (2019) 'Mechanisms of resistance to trastuzumab emtansine (T-DM1) in HER2-positive breast cancer', British Journal of Cancer, 122(5), pp. 603–612. doi: 10.1038/s41416-019-0635-y

75. Dong, X. et al. (2020) 'Exosomes and breast cancer drug resistance', Cell Death & Disease, 11(11). doi: 10.1038/s41419-020-03189-z

76. Frampton, S. et al. (2024) 'Fc gamma receptors: Their evolution, genomic architecture, genetic variation, and impact on human disease', Immunological Reviews, 328(1), pp. 65–97. doi: 10.1111/imr.13401

77. Lu, Y. et al. (2001) 'Insulin-Like Growth Factor-I Receptor Signaling and Resistance to Trastuzumab (Herceptin)', JNCI Journal of the National Cancer Institute, 93(24), pp. 1852–1857. doi: 10.1093/jnci/93.24.1852

78. Zhao, Y. et al. (2011) ‘Overcoming Trastuzumab Resistance in Breast Cancer by Targeting Dysregulated Glucose Metabolism’, Cancer Research, 71(13), pp. 4585–4597. doi: 10.1158/0008-5472.can-11-0127

79. Zhang, Y. et al. (2019) ‘<p>Identification of an Activating Mutation in the Extracellular Domain of HER2 Conferring Resistance to Pertuzumab</p>’;, OncoTargets and Therapy, Volume 12, pp. 11597–11608. doi: 10.2147/ott.s232912

80. Knowlden, J.M. et al. (2003) ‘Elevated Levels of Epidermal Growth Factor Receptor/c-erbB2 Heterodimers Mediate an Autocrine Growth Regulatory Pathway in Tamoxifen-Resistant MCF-7 Cells’, Endocrinology, 144(3), pp. 1032–1044. doi: 10.1210/en.2002-220620

81. Wuerkenbieke, D. et al. (2015) ‘miRNA-150 downregulation promotes pertuzumab resistance in ovarian cancer cells via AKT activation’, Archives of Gynecology and Obstetrics, 292(5), pp. 1109–1116. doi: 10.1007/s00404-015-3742-x

82. Flajnik, M.F. and Kasahara, M. (2009) ‘Origin and evolution of the adaptive immune system: genetic events and selective pressures’, Nature Reviews Genetics, 11(1), pp. 47–59. doi: 10.1038/nrg2703

83. Greenberg, A.H. et al. (1973) ‘Antibody-dependent Cell-mediated Cytotoxicity due to a “Null” Lymphoid Cell’, Nature New Biology, 242(117), pp. 111–113. doi: 10.1038/newbio242111a0

84. Kiessling, R., Klein, E. and Wigzell, H. (1975) ‘Natural” killer cells in the mouse. I. Cytotoxic cells with specificity for mouse Moloney leukemia cells. Specificity and distribution according to genotype’, European Journal of Immunology, 5(2), pp. 112–117. doi: 10.1002/eji.1830050208

85. Grégoire, C. et al. (2007) ‘The trafficking of natural killer cells’, Immunological Reviews, 220(1), pp. 169–182. doi: 10.1111/j.1600-065X.2007.00563.x

86. Björkström, N.K., Strunz, B. and Ljunggren, H.-G. (2021) ‘Natural killer cells in antiviral immunity’, Nature Reviews Immunology, 22(2), pp. 112–123. doi: 10.1038/s41577-021-00558-3

87. Dalbeth, N. et al. (2004) ‘CD56bright NK Cells Are Enriched at Inflammatory Sites and Can Engage with Monocytes in a Reciprocal Program of Activation’, The Journal of Immunology, 173(10), pp. 6418–6426. doi: 10.4049/jimmunol.173.10.6418

88. Mace, E.M. et al. (2016) ‘Human NK cell development requires CD56-mediated motility and formation of the developmental synapse’, Nature Communications, 7(1). doi: 10.1038/ncomms12171

89. Ham, H., Medlyn, M. and Billadeau, D.D. (2022) ‘Locked and Loaded: Mechanisms Regulating Natural Killer Cell Lytic Granule Biogenesis and Release’, Frontiers in Immunology, 13. doi: 10.3389/fimmu.2022.871106

90. Coënon, L. et al. (2024) ‘Natural Killer cells at the frontline in the fight against cancer’, Cell Death & Disease, 15(8). doi: 10.1038/s41419-024-06976-0

91. Morvan, M.G. and Lanier, L.L. (2015) ‘NK cells and cancer: you can teach innate cells new tricks’, Nature Reviews Cancer, 16(1), pp. 7–19. doi: 10.1038/nrc.2015.5

92. Xie, J. et al. (2008) ‘Structural Basis for Pattern Recognition by the Receptor for Advanced Glycation End Products (RAGE)’, Journal of Biological Chemistry, 283(40), pp. 27255–27269. doi: 10.1074/jbc.m801622200

93. Darwich, A. et al. (2020) ‘The Role of CHI3L1 in Breast Cancer: a Driver of Immunosuppression on the Road of Tumor Progression’, https://hdl.handle.net/2434/789305, accessed June 2025

94. Yue, Q. et al. (2022) ‘Receptor for Advanced Glycation End Products (RAGE): A Pivotal Hub in Immune Diseases’, Molecules, 27(15), p. 4922. doi: 10.3390/molecules27154922

95. Parodi, M. et al. (2015) ‘Natural Killer (NK)/melanoma cell interaction induces NK-mediated release of chemotactic High Mobility Group Box-1 (HMGB1) capable of amplifying NK cell recruitment’, OncoImmunology, 4(12), p. e1052353. doi: 10.1080/2162402x.2015.1052353

96. Narumi, K. et al. (2015) ‘Proinflammatory Proteins S100A8/S100A9 Activate NK Cells via Interaction with RAGE’, The Journal of Immunology, 194(11), pp. 5539–5548. doi: 10.4049/jimmunol.1402301

97. Ahangari, F. et al. (2015) ‘Chitinase 3–like-1 Regulates Both Visceral Fat Accumulation and Asthma-like Th2 Inflammation’, American Journal of Respiratory and Critical Care Medicine, 191(7), pp. 746–757. doi: 10.1164/rccm.201405-0796OC

98. Morvan, M.G. and Lanier, L.L. (2015) ‘NK cells and cancer: you can teach innate cells new tricks’, Nature Reviews Cancer, 16(1), pp. 7–19. doi: 10.1038/nrc.2015.5

99. Low, D. et al. (2015) ‘Chitinase 3-like 1 induces survival and proliferation of intestinal epithelial cells during chronic inflammation and colitis-associated cancer by regulating S100A9’, Oncotarget, 6(34), pp. 36535–36550. doi: 10.18632/oncotarget.5440

100. Song, Y. et al. (2024) ‘Astrocyte-derived CHI3L1 signaling impairs neurogenesis and cognition in the demyelinated hippocampus’, Cell Reports, 43(5), p. 114226. doi: 10.1016/j.celrep.2024.114226

101. Kim, H.J., Jeong, M.S. and Jang, S.B. (2021) ‘Molecular Characteristics of RAGE and Advances in Small-Molecule Inhibitors’, International Journal of Molecular Sciences, 22(13), p. 6904. doi: 10.3390/ijms22136904

102. Mehta, R. et al. (2018) ‘Polymorphisms in the receptor for advanced glycation end-products (RAGE) gene and circulating RAGE levels as a susceptibility factor for non-alcoholic steatohepatitis (NASH)’, PLOS ONE. Edited by B.I. Hudson, 13(6), p. e0199294. doi: 10.1371/journal.pone.0199294

103. Vlassara, H. et al. (1987) ‘Advanced glycosylation endproducts on erythrocyte cell surface induce receptor-mediated phagocytosis by macrophages. A model for turnover of aging cells.’, The Journal of experimental medicine, 166(2), pp. 539–549. doi: 10.1084/jem.166.2.539

104. Sessa, L. et al. (2014) ‘The Receptor for Advanced Glycation End-Products (RAGE) Is Only Present in Mammals, and Belongs to a Family of Cell Adhesion Molecules (CAMs)’, PLoS ONE. Edited by B.I. Hudson, 9(1), p. e86903. doi: 10.1371/journal.pone.0086903

105. Du Pasquier, L, Chrétien, I (1996) “Why is CTX all the RAGE?.” Research in immunology vol. 147,4:261-6.

106. https://atlasgeneticsoncology.org/gene/594/ager-%28advanced-glycosylation-end-product-specific-receptor%29, accessed 23 June 2025

107. Zamoon, J., Madhu, D. and Ahmed, I. (2019) ‘Dynamic oligomerization of hRAGE’s transmembrane and cytoplasmic domains within SDS micelles’, International Journal of Biological Macromolecules, 130, pp. 10–18. doi: 10.1016/j.ijbiomac.2019.02.108

108. Brett, J. et al. (1993) ‘Survey of the distribution of a newly characterized receptor for advanced glycation end products in tissues.’ Am J Pathol; 143(6):1699-1712.

109. Shirasawa, M. et al. (2004) ‘Receptor for advanced glycation end-products is a marker of type I lung alveolar cells’, Genes to Cells, 9(2), pp. 165–174. doi: 10.1111/j.1356-9597.2004.00712.x

110. Hudson, B.I. et al. (2001) ‘Effects of Novel Polymorphisms in the RAGE Gene on Transcriptional Regulation and Their Association With Diabetic Retinopathy’, Diabetes, 50(6), pp. 1505–1511. doi: 10.2337/diabetes.50.6.1505

111. Cross, K. et al. (2024) ‘Role of the Receptor for Advanced Glycation End Products (RAGE) and Its Ligands in Inflammatory Responses’, Biomolecules, 14(12), p. 1550. doi: 10.3390/biom14121550

112. Kalea, A.Z., Schmidt, A.M. and Hudson, B.I. (2009) ‘RAGE: a novel biological and genetic marker for vascular disease’, Clinical Science, 116(8), pp. 621–637. doi: 10.1042/CS20080494

113. Hudson, B.I. et al. (2007) ‘Identification, classification, and expression ofRAGEgene splice variants’, The FASEB Journal, 22(5), pp. 1572–1580. doi: 10.1096/fj.07-9909com

114. Bongarzone, S. et al. (2017) ‘Targeting the Receptor for Advanced Glycation Endproducts (RAGE): A Medicinal Chemistry Perspective’, Journal of Medicinal Chemistry, 60(17), pp. 7213–7232. doi: 10.1021/acs.jmedchem.7b00058

115. Schlueter, C. et al. (2003) ‘Tissue-specific expression patterns of the RAGE receptor and its soluble forms—a result of regulated alternative splicing?’, Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression, 1630(1), pp. 1–6. doi: 10.1016/j.bbaexp.2003.08.008

116. Hudson, B. et al. (2002). ‘Glycation and diabetes: The RAGE connection.’ Current Science. 83.

117. Osawa, M. et al. (2007) ‘De-N-glycosylation or G82S mutation of RAGE sensitizes its interaction with advanced glycation endproducts’, Biochimica et Biophysica Acta (BBA) - General Subjects, 1770(10), pp. 1468–1474. doi: 10.1016/j.bbagen.2007.07.003

118. Zhao, D.C. et al. (2015) ‘Association between the receptor for advanced glycation end products gene polymorphisms and cancer risk: a systematic review and meta-analysis.’ J BUON;20(2):614-624.

119. Neeper, M. et al. (1992) ‘Cloning and expression of a cell surface receptor for advanced glycosylation end products of proteins.’ J Biol Chem;267(21):14998-15004

120. Zong, H. et al. (2010) ‘Homodimerization Is Essential for the Receptor for Advanced Glycation End Products (RAGE)-mediated Signal Transduction’, Journal of Biological Chemistry, 285(30), pp. 23137–23146. doi: 10.1074/jbc.M110.133827

121. Dattilo, B.M. et al. (2007) ‘The Extracellular Region of the Receptor for Advanced Glycation End Products Is Composed of Two Independent Structural Units’, Biochemistry, 46(23), pp. 6957–6970. doi: 10.1021/bi7003735

122. Xue, J. et al. (2016) ‘Change in the Molecular Dimension of a RAGE-Ligand Complex Triggers RAGE Signaling’, Structure, 24(9), pp. 1509–1522. doi: 10.1016/j.str.2016.06.021

123. Park, H. and Boyington, J.C. (2010) ‘The 1.5 Å Crystal Structure of Human Receptor for Advanced Glycation Endproducts (RAGE) Ectodomains Reveals Unique Features Determining Ligand Binding’, Journal of Biological Chemistry, 285(52), pp. 40762–40770. doi: 10.1074/jbc.m110.169276

124. Singh, H. and Agrawal, D.K. (2022) ‘Therapeutic potential of targeting the receptor for advanced glycation end products (RAGE) by small molecule inhibitors’, Drug Development Research, 83(6), pp. 1257–1269. doi: 10.1002/ddr.21971

125. Srikrishna, G. et al. (2010) ‘Carboxylated N-glycans on RAGE promote S100A12 binding and signaling’, Journal of Cellular Biochemistry, 110(3), pp. 645–659. doi: 10.1002/jcb.22575

126. Wei, W. et al. (2012) ‘Disulfide Bonds within the C2 Domain of RAGE Play Key Roles in Its Dimerization and Biogenesis’, PLoS ONE. Edited by M.P. Bachmann, 7(12), p. e50736. doi: 10.1371/journal.pone.0050736

127. Rai, V. et al. (2012) ‘Signal Transduction in Receptor for Advanced Glycation End Products (RAGE)’, Journal of Biological Chemistry, 287(7), pp. 5133–5144. doi: 10.1074/jbc.M111.277731

128. Yatime, L. et al. (2016) ‘The Structure of the RAGE:S100A6 Complex Reveals a Unique Mode of Homodimerization for S100 Proteins’, Structure, 24(12), pp. 2043–2052. doi: 10.1016/j.str.2016.09.011

129. C, R.Cathrine., Lukose, B. and Rani, P. (2020) ‘G82S RAGE polymorphism influences amyloid-RAGE interactions relevant in Alzheimer’s disease pathology’, PLOS ONE. Edited by E. Gallicchio, 15(10), p. e0225487. doi: 10.1371/journal.pone.0225487

130. Park, S.J. et al. (2011) ‘The G82S polymorphism promotes glycosylation of the receptor for advanced glycation end products (RAGE) at asparagine 81: comparison of wild-type rage with the G82S polymorphic variant.’ J Biol Chem. 2011;286(24):21384-21392. doi:10.1074/jbc.M111.241281

131. Srikrishna, G et al. (2009) ‘Endogenous damage-associated molecular pattern molecules at the crossroads of inflammation and cancer.’ Neoplasia;11(7):615-628. doi:10.1593/neo.09284

132. Sinha, P. et al. (2008) ‘Proinflammatory S100 proteins regulate the accumulation of myeloid-derived suppressor cells.’ J Immunol;181(7):4666-4675. doi:10.4049/jimmunol.181.7.4666

133. Srikrishna, G. et al. (2002) ‘N-Glycans on the receptor for advanced glycation end products influence amphoterin binding and neurite outgrowth’, Journal of Neurochemistry, 80(6), pp. 998–1008. doi: 10.1046/j.0022-3042.2002.00796.x

134. Augner, K. (2014) ‘Influence of nonenzymatic posttranslational modifications on constitution, oligomerization and receptor binding of S100A12. PLoS One;9(11):e113418. doi:10.1371/journal.pone.0113418

135. Turovskaya, O. et al. (2008) ‘RAGE, carboxylated glycans and S100A8/A9 play essential roles in colitis-associated carcinogenesis’, Carcinogenesis, 29(10), pp. 2035–2043. doi: 10.1093/carcin/bgn188

136. Curran, C.S. and Kopp, J.B. (2022) ‘RAGE pathway activation and function in chronic kidney disease and COVID-19’, Frontiers in Medicine, 9. doi: 10.3389/fmed.2022.970423

137. Peng, Y. et al. (2016) ‘Mouse RAGE Variant 4 Is a Dominant Membrane Receptor that Does Not Shed to Generate Soluble RAGE’, PLOS ONE. Edited by T. Fukai, 11(9), p. e0153657. doi: 10.1371/journal.pone.0153657

138. Li, M. et al. (2022) ‘Heparan sulfate-dependent RAGE oligomerization is indispensable for pathophysiological functions of RAGE’, eLife, 11. doi: 10.7554/eLife.71403

139. Xu, D. et al. (2013) ‘Stable RAGE-Heparan Sulfate Complexes Are Essential for Signal Transduction’, ACS Chemical Biology, 8(7), pp. 1611–1620. doi: 10.1021/cb4001553

140. Koch, M. et al. (2010) ‘Structural Basis for Ligand Recognition and Activation of RAGE’, Structure, 18(10), pp. 1342–1352. doi: 10.1016/j.str.2010.05.017

141. Moysa, A. et al. (2019) ‘Enhanced oligomerization of full-length RAGE by synergy of the interaction of its domains’, Scientific Reports, 9(1). doi: 10.1038/s41598-019-56993-9

142. Rojas, A. et al. (2024) ‘The RAGE Axis: A Relevant Inflammatory Hub in Human Diseases’, Biomolecules, 14(4), p. 412. doi: 10.3390/biom14040412

143. Ott, C. et al. (2014) ‘Role of advanced glycation end products in cellular signaling’, Redox Biology, 2, pp. 411–429. doi: 10.1016/j.redox.2013.12.016

144. Xue, J. et al. (2014) ‘The Receptor for Advanced Glycation End Products (RAGE) Specifically Recognizes Methylglyoxal-Derived AGEs’, Biochemistry, 53(20), pp. 3327–3335. doi: 10.1021/bi500046t

145. Yamagishi, S. and Matsui, T. (2015) ‘Role of receptor for advanced glycation end products (RAGE) in liver disease’, European Journal of Medical Research, 20(1). doi: 10.1186/s40001-015-0090-z

146. Donato, R. (2001) ‘S100: a multigenic family of calcium-modulated proteins of the EF-hand type with intracellular and extracellular functional roles’, The International Journal of Biochemistry & Cell Biology, 33(7), pp. 637–668. doi: 10.1016/S1357-2725(01)00046-2

147. Bresnick, A.R., Weber, D.J. and Zimmer, D.B. (2015) ‘S100 proteins in cancer’, Nature Reviews Cancer, 15(2), pp. 96–109. doi: 10.1038/nrc3893

148. Leukert, N. et al. (2006) ‘Calcium-dependent Tetramer Formation of S100A8 and S100A9 is Essential for Biological Activity’, Journal of Molecular Biology, 359(4), pp. 961–972. doi: 10.1016/j.jmb.2006.04.009

149. Srikrishna, G. et al. (2001) ‘Two Proteins Modulating Transendothelial Migration of Leukocytes Recognize Novel Carboxylated Glycans on Endothelial Cells’, The Journal of Immunology, 166(7), pp. 4678–4688. doi: 10.4049/jimmunol.166.7.4678

150. Rigiracciolo, D.C. et al. (2022) ‘Focal Adhesion Kinase (FAK)-Hippo/YAP transduction signaling mediates the stimulatory effects exerted by S100A8/A9-RAGE system in triple-negative breast cancer (TNBC)’, Journal of Experimental & Clinical Cancer Research, 41(1). doi: 10.1186/s13046-022-02396-0

151. Ghavami, S. et al. (2008) ‘S100A8/A9 at low concentration promotes tumor cell growth via RAGE ligation and MAP kinase-dependent pathway’, Journal of Leukocyte Biology, 83(6), pp. 1484–1492. doi: 10.1189/jlb.0607397

152. Moysa, A. et al. (2021) ‘A model of full-length RAGE in complex with S100B’, Structure, 29(9), pp. 989-1002.e6. doi: 10.1016/j.str.2021.04.002

153. Mohan, S.K., Gupta, A.A. and Yu, C. (2013) ‘Interaction of the S100A6 mutant (C3S) with the V domain of the receptor for advanced glycation end products (RAGE)’, Biochemical and Biophysical Research Communications, 434(2), pp. 328–333. doi: 10.1016/j.bbrc.2013.03.049

154. Penumutchu, S.R., Chou, R.-H. and Yu, C. (2014) ‘Structural Insights into Calcium-Bound S100P and the V Domain of the RAGE Complex’, PLoS ONE. Edited by A. Motta, 9(8), p. e103947. doi: 10.1371/journal.pone.0103947

155. Datta, S. et al. (2024) ‘High Mobility Group Box 1 (HMGB1): Molecular Signaling and Potential Therapeutic Strategies’, Cells, 13(23), p. 1946. doi: 10.3390/cells13231946

156. Park, J.S. et al. (2004) ‘Involvement of Toll-like Receptors 2 and 4 in Cellular Activation by High Mobility Group Box 1 Protein’, Journal of Biological Chemistry, 279(9), pp. 7370–7377. doi: 10.1074/jbc.M306793200

157. Huttunen, H.J. et al. (2002) ‘Receptor for Advanced Glycation End Products-binding COOH-terminal Motif of Amphoterin Inhibits Invasive Migration and Metastasis’ Cancer Res (2002) 62 (16): 4805–4811.

158. Son, M. et al. (2016) ‘C1q and HMGB1 reciprocally regulate human macrophage polarization’, Blood, 128(18), pp. 2218–2228. doi: 10.1182/blood-2016-05-719757

159. Giordano, L. et al. (2025) ‘Mitochondrial DNA signals driving immune responses: Why, How, Where?’, Cell Communication and Signaling, 23(1). doi: 10.1186/s12964-025-02042-0

160. Donahue, J.E. et al. (2006) ‘RAGE, LRP-1, and amyloid-beta protein in Alzheimer’s disease’, Acta Neuropathologica, 112(4), pp. 405–415. doi: 10.1007/s00401-006-0115-3

161. Ray, R. et al. (2020) ‘Lysophosphatidic acid-RAGE axis promotes lung and mammary oncogenesis via protein kinase B and regulating tumor microenvironment’, Cell Communication and Signaling, 18(1). doi: 10.1186/s12964-020-00666-y

162. Qin, Y.-H. et al. (2009) ‘HMGB1 Enhances the Proinflammatory Activity of Lipopolysaccharide by Promoting the Phosphorylation of MAPK p38 through Receptor for Advanced Glycation End Products’, The Journal of Immunology, 183(10), pp. 6244–6250. doi: 10.4049/jimmunol.0900390

163. Yatime, L. and Andersen, G.R. (2013) ‘Structural insights into the oligomerization mode of the human receptor for advanced glycation end-products’, FEBS Journal, 280(24), pp. 6556–6568. doi: 10.1111/febs.12556

164. Egaña-Gorroño, L. et al. (2020) ‘Receptor for Advanced Glycation End Products (RAGE) and Mechanisms and Therapeutic Opportunities in Diabetes and Cardiovascular Disease: Insights From Human Subjects and Animal Models’, Frontiers in Cardiovascular Medicine, 7. doi: 10.3389/fcvm.2020.00037

165. Dong, H. et al. (2022) ‘Pathophysiology of RAGE in inflammatory diseases’, Frontiers in Immunology, 13. doi: 10.3389/fimmu.2022.931473

166. Sakaguchi, M. et al. (2011) ‘TIRAP, an Adaptor Protein for TLR2/4, Transduces a Signal from RAGE Phosphorylated upon Ligand Binding’, PLoS ONE. Edited by D. Chandra, 6(8), p. e23132. doi: 10.1371/journal.pone.0023132

167. Lin, L. (2009) ‘RAGE signaling in inflammation and arterial aging’, Frontiers in Bioscience, Volume(14), p. 1403. doi: 10.2741/3315

168. Erusalimsky, J.D. (2021) ‘The use of the soluble receptor for advanced glycation-end products (sRAGE) as a potential biomarker of disease risk and adverse outcomes’, Redox Biology, 42, p. 101958. doi: 10.1016/j.redox.2021.101958

169. Braley, A. et al. (2016) ‘Regulation of Receptor for Advanced Glycation End Products (RAGE) Ectodomain Shedding and Its Role in Cell Function’, Journal of Biological Chemistry, 291(23), pp. 12057–12073. doi: 10.1074/jbc.M115.702399

170. Galichet, A., Weibel, M. and Heizmann, C.W. (2008) ‘Calcium-regulated intramembrane proteolysis of the RAGE receptor’, Biochemical and Biophysical Research Communications, 370(1), pp. 1–5. doi: 10.1016/j.bbrc.2008.02.163

171. Detzen, L. et al. ‘Soluble Forms of the Receptor for Advanced Glycation Endproducts (RAGE) in Periodontitis.’ Sci Rep 9, 8170 (2019). doi: 10.1038/s41598-019-44608-2

172. Eggers, K. et al. (2011) ‘RAGE-Dependent Regulation of Calcium-Binding Proteins S100A8 and S100A9 in Human THP-1’, Experimental and Clinical Endocrinology & Diabetes, 119(06), pp. 353–357. doi: 10.1055/s-0030-1268426

173. Cheng, M. et al. (2015) ‘HMGB1 Enhances the AGE-Induced Expression of CTGF and TGF-β via RAGE-Dependent Signaling in Renal Tubular Epithelial Cells’, American Journal of Nephrology, 41(3), pp. 257–266. doi: 10.1159/000381464

174. Baek, H. et al. (2021) ‘Reduced receptor for advanced glycation end products is associated with α-SMA expression in patients with idiopathic pulmonary fibrosis and mice’, Laboratory Animal Research, 37(1). doi: 10.1186/s42826-021-00105-0

175. Buhimschi, C.S. et al. (2009) ‘Characterization of RAGE, HMGB1, and S100β in Inflammation-Induced Preterm Birth and Fetal Tissue Injury’, The American Journal of Pathology, 175(3), pp. 958–975. doi: 10.2353/ajpath.2009.090156

176. Tirone, M. et al. (2017) ‘High mobility group box 1 orchestrates tissue regeneration via CXCR4’, Journal of Experimental Medicine, 215(1), pp. 303–318. doi: 10.1084/jem.20160217

177. Funk, S.D., Yurdagul, A. and Orr, A.W. (2012) ‘Hyperglycemia and Endothelial Dysfunction in Atherosclerosis: Lessons from Type 1 Diabetes’, International Journal of Vascular Medicine, 2012, pp. 1–19. doi: 10.1155/2012/569654

178. Wendt, T.M. et al. (2003) ‘RAGE Drives the Development of Glomerulosclerosis and Implicates Podocyte Activation in the Pathogenesis of Diabetic Nephropathy’, The American Journal of Pathology, 162(4), pp. 1123–1137. doi: 10.1016/S0002-9440(10)63909-0

179. Xu, Y. et al. (2024) ‘Unraveling the Mechanisms of S100A8/A9 in Myocardial Injury and Dysfunction’, Current Issues in Molecular Biology, 46(9), pp. 9707–9720. doi: 10.3390/cimb46090577

180. Yan, S.S. (2012) ‘RAGE is a key cellular target for A -induced perturbation in Alzheimer s disease’, Frontiers in Bioscience, S4(1), pp. 240–250. doi: 10.2741/265

181. Kang, K.Y., Woo, J.-W. and Park, S.-H. (2014) ‘S100A8/A9 as a biomarker for synovial inflammation and joint damage in patients with rheumatoid arthritis’, The Korean Journal of Internal Medicine, 29(1), p. 12. doi: 10.3904/kjim.2014.29.1.12

182. Shi, A. et al. (2023) ‘Long-term ingestion of β-lactoglobulin-bound AGEs induces colonic inflammation by modulating RAGE (TLR4)/MYD88/NF-κB signaling pathway and gut microbiota in mice’, Journal of Functional Foods, 107, p. 105690. doi: 10.1016/j.jff.2023.105690

183. Perkins, T.N. and Oury, T.D. (2021) ‘The perplexing role of RAGE in pulmonary fibrosis: causality or casualty?’, Therapeutic Advances in Respiratory Disease, 15. doi: 10.1177/17534666211016071

184. Halayko, A.J. and Ghavami, S. (2009) ‘S100A8/A9: a mediator of severe asthma pathogenesis and morbidity?This article is one of a selection of papers published in a special issue celebrating the 125th anniversary of the Faculty of Medicine at the University of Manitoba.’, Canadian Journal of Physiology and Pharmacology, 87(10), pp. 743–755. doi: 10.1139/Y09-054

185. Loh, Z. et al. (2020) ‘HMGB1 amplifies ILC2-induced type-2 inflammation and airway smooth muscle remodelling’, PLOS Pathogens. Edited by M.T. Heise, 16(7), p. e1008651. doi: 10.1371/journal.ppat.1008651

186. Logsdon, C. et al. (2007) ‘RAGE and RAGE Ligands in Cancer’, Current Molecular Medicine, 7(8), pp. 777–789. doi: 10.2174/156652407783220697

187. Leclerc, E. et al. (2009) ‘Binding of S100 proteins to RAGE: An update’, Biochimica et Biophysica Acta (BBA) - Molecular Cell Research, 1793(6), pp. 993–1007. doi: 10.1016/j.bbamcr.2008.11.016

188. Yang, Y. et al. (2020) ‘HMGB1 mediates lipopolysaccharide-induced inflammation via interacting with GPX4 in colon cancer cells’, Cancer Cell International, 20(1). doi: 10.1186/s12935-020-01289-6

189. Wu, S. et al. (2018) ‘RAGE may act as a tumour suppressor to regulate lung cancer development’, Gene, 651, pp. 86–93. doi: 10.1016/j.gene.2018.02.009

190. Johansen, J.S. et al. (1992) ‘Identification of proteins secreted by human osteoblastic cells in culture’, Journal of Bone and Mineral Research, 7(5), pp. 501–512. doi: 10.1002/jbmr.5650070506

191. Chen, W., Jiang, X. and Yang, Q. (2020) ‘Glycoside hydrolase family 18 chitinases: The known and the unknown’, Biotechnology Advances, 43, p. 107553. doi: 10.1016/j.biotechadv.2020.107553

192. Funkhouser, J.D. and Aronson, N.N. (2007) ‘Chitinase family GH18: evolutionary insights from the genomic history of a diverse protein family’, BMC Evolutionary Biology, 7(1). doi: 10.1186/1471-2148-7-96

193. https://www.researchgate.net/publication/247903069_The_molecular_mechanism_of_LPS-induced_CHI3L1_expression, doi: 10.3724/SP.J.1206.2008.00634

194. Fusetti, F. et al. (2003) ‘Crystal Structure and Carbohydrate-binding Properties of the Human Cartilage Glycoprotein-39’, Journal of Biological Chemistry, 278(39), pp. 37753–37760. doi: 10.1074/jbc.M303137200

195. Terwisscha van Scheltinga, A.C. et al. (1994) ‘Crystal structures of hevamine, a plant defence protein with chitinase and lysozyme activity, and its complex with an inhibitor’, Structure, 2(12), pp. 1181–1189. doi:10.1016/S0969-2126(94)00120-0

196. Magnusdottir, U. et al. (2025) ‘Heparin-binding of the human chitinase-like protein YKL-40 is allosterically modified by chitin oligosaccharides’, Biochemistry and Biophysics Reports, 41, p. 101908. doi: 10.1016/j.bbrep.2024.101908

197. Zhao, T. et al. (2020) ‘Chitinase-3 like-protein-1 function and its role in diseases.’ Sig Transduct Target Ther 5, 201. doi: 10.1038/s41392-020-00303-7

198. Ngernyuang, N. et al. (2018) ‘A Heparin Binding Motif Rich in Arginine and Lysine is the Functional Domain of YKL-40’, Neoplasia, 20(2), pp. 182–192. doi: 10.1016/j.neo.2017.11.011

199. Chen, C.C., et al. (2011) ‘Carbohydrate-binding motif in chitinase 3-like 1 (CHI3L1/YKL-40) specifically activates Akt signaling pathway in colonic epithelial cells.’ Clin Immunol;140(3):268-275. doi: 10.1016/j.clim.2011.04.007

200. Bigg, H.F. et al. (2006) ‘The Mammalian Chitinase-like Lectin, YKL-40, Binds Specifically to Type I Collagen and Modulates the Rate of Type I Collagen Fibril Formation’, Journal of Biological Chemistry, 281(30), pp. 21082–21095. doi: 10.1074/jbc.M601153200

201. Fusetti, F. et al. (2003) ‘Crystal Structure and Carbohydrate-binding Properties of the Human Cartilage Glycoprotein-39’, Journal of Biological Chemistry, 278(39), pp. 37753–37760. doi: 10.1074/jbc.M303137200, license available at https://creativecommons.org/licenses/by/4.0/

202. Houston, D.R. et al. (2003) ‘Structure and Ligand-induced Conformational Change of the 39-kDa Glycoprotein from Human Articular Chondrocytes’, Journal of Biological Chemistry, 278(32), pp. 30206–30212. doi: 10.1074/jbc.M303371200

203. Zhao, T. et al. (2020) ‘Chitinase-3 like-protein-1 function and its role in diseases’, Signal Transduction and Targeted Therapy, 5(1). doi: 10.1038/s41392-020-00303-7

204. Bonneh-Barkay, D. et al. (2010) ‘In vivo CHI3L1 (YKL-40) expression in astrocytes in acute and chronic neurological diseases’, Journal of Neuroinflammation, 7(1), p. 34. doi: 10.1186/1742-2094-7-34

205. Libreros, S. and Iragavarapu-Charyulu, V. (2015) ‘YKL-40/CHI3L1 drives inflammation on the road of tumor progression’, Journal of Leukocyte Biology, 98(6), pp. 931–936. doi: 10.1189/jlb.3VMR0415-142R

206. Geng, B. et al. (2018) ‘Chitinase 3-like 1-CD44 interaction promotes metastasis and epithelial-to-mesenchymal transition through β-catenin/Erk/Akt signaling in gastric cancer’, Journal of Experimental & Clinical Cancer Research, 37(1). doi: 10.1186/s13046-018-0876-2

207. Kim, D.-H. et al. (2018) ‘Regulation of chitinase-3-like-1 in T cell elicits Th1 and cytotoxic responses to inhibit lung metastasis’, Nature Communications, 9(1). doi: 10.1038/s41467-017-02731-6

208. Jin, T. et al. (2015) ‘The Role of MicroRNA, miR-24, and Its Target CHI3L1 in Osteomyelitis Caused by Staphylococcus aureus’, Journal of Cellular Biochemistry, 116(12), pp. 2804–2813. doi: 10.1002/jcb.25225

209. Petersson, M. et al. (2006) ‘Effects of arginine-vasopressin and parathyroid hormone-related protein (1–34) on cell proliferation and production of YKL-40 in cultured chondrocytes from patients with rheumatoid arthritis and osteoarthritis’, Osteoarthritis and Cartilage, 14(7), pp. 652–659. doi: 10.1016/j.joca.2006.01.003

210. Shi, M. et al. (2022) ‘Functional analysis of the short splicing variant encoded by CHI3L1/YKL-40 in glioblastoma’, Frontiers in Oncology, 12. doi: 10.3389/fonc.2022.910728

211. Sanfilippo, C. et al. (2019) ‘Sex difference in CHI3L1 expression levels in human brain aging and in Alzheimer’s disease’, Brain Research, 1720, p. 146305. doi: 10.1016/j.brainres.2019.146305

212. Nishimura, N. et al. (2021) ‘Chitinase 3-like 1 is a profibrogenic factor overexpressed in the aging liver and in patients with liver cirrhosis’, Proceedings of the National Academy of Sciences, 118(17). doi: 10.1073/pnas.2019633118

213. Yu, J.E. et al. (2024) ‘Significance of chitinase-3-like protein 1 in the pathogenesis of inflammatory diseases and cancer’, Experimental & Molecular Medicine, 56(1), pp. 1–18. doi: 10.1038/s12276-023-01131-9

214. Kuo, C.C. et al. (2024) ‘Evaluation of Serum YKL-40 in Canine Multicentric Lymphoma: Clinical and Diagnostic Implications.’ Animals (Basel);14(23):3391. doi: 10.3390/ani14233391

215. Lee, C.G. et al. (2009) ‘Role of breast regression protein 39 (BRP-39)/chitinase 3-like-1 in Th2 and IL-13–induced tissue responses and apoptosis’, Journal of Experimental Medicine, 206(5), pp. 1149–1166. doi: 10.1084/jem.20081271

216. He, C.H. et al. (2013) ‘Chitinase 3-like 1 Regulates Cellular and Tissue Responses via IL-13 Receptor α2’, Cell Reports, 4(4), pp. 830–841. doi: 10.1016/j.celrep.2013.07.032

217. Scott, T.E. et al. (2023) ‘IL-4 and IL-13 induce equivalent expression of traditional M2 markers and modulation of reactive oxygen species in human macrophages’, Scientific Reports, 13(1). doi: 10.1038/s41598-023-46237-2

218. Márquez-Ortiz, R.A. et al. (2021) ‘IL13Rα2 Promotes Proliferation and Outgrowth of Breast Cancer Brain Metastases’, Clinical Cancer Research, 27(22), pp. 6209–6221. doi: 10.1158/1078-0432.ccr-21-0361

219. Zhao, T. et al. (2020) ‘Chitinase-3 like-protein-1 function and its role in diseases’, Signal Transduction and Targeted Therapy, 5(1). doi: 10.1038/s41392-020-00303-7 license available at https://creativecommons.org/licenses/by/4.0/

220. Chen, A. et al. (2021) ‘Chitinase-3-like 1 protein complexes modulate macrophage-mediated immune suppression in glioblastoma’, Journal of Clinical Investigation, 131(16). doi: 10.1172/JCI147552

221. Boza-Serrano, A. et al. (2022) ‘Galectin-3 is elevated in CSF and is associated with Aβ deposits and tau aggregates in brain tissue in Alzheimer’s disease’, Acta Neuropathologica, 144(5), pp. 843–859. doi: 10.1007/s00401-022-02469-6

222. Cao, Y. et al. (2022) ‘CRTH2 Mediates Profibrotic Macrophage Differentiation and Promotes Lung Fibrosis’, American Journal of Respiratory Cell and Molecular Biology, 67(2), pp. 201–214. doi: 10.1165/rcmb.2021-0504oc

223. Zhou, Y. et al. (2018) ‘Galectin-3 Interacts with the CHI3L1 Axis and Contributes to Hermansky–Pudlak Syndrome Lung Disease’, The Journal of Immunology, 200(6), pp. 2140–2153. doi: 10.4049/jimmunol.1701442

224. Fraser, J.R.E., Laurent, T.C. and Laurent, U.B.G. (1997) ‘Hyaluronan: its nature, distribution, functions and turnover’, Journal of Internal Medicine, 242(1), pp. 27–33. doi: 10.1046/j.1365-2796.1997.00170.x

225. Gallagher, J.T. and Walker, A. (1985) ‘Molecular distinctions between heparan sulphate and heparin. Analysis of sulphation patterns indicates that heparan sulphate and heparin are separate families of N-sulphated polysaccharides’, Biochemical Journal, 230(3), pp. 665–674. doi: 10.1042/bj2300665

226. Iozzo, R.V. (1998) ‘MATRIX PROTEOGLYCANS: From Molecular Design to Cellular Function’, Annual Review of Biochemistry, 67(1), pp. 609–652. doi: 10.1146/annurev.biochem.67.1.609

227. Kwak, E.J. et al. (2019) ‘Chitinase 3-like 1 drives allergic skin inflammation via Th2 immunity and M2 macrophage activation’, Clinical & Experimental Allergy, 49(11), pp. 1464–1474. doi: 10.1111/cea.13478

228. Rehli, M. et al. (2003) ‘Transcriptional Regulation of CHI3L1, a Marker Gene for Late Stages of Macrophage Differentiation’, Journal of Biological Chemistry, 278(45), pp. 44058–44067. doi: 10.1074/jbc.M306792200

229. Hirano, F. et al. (1998) ‘Functional Interference of Sp1 and NF-κB through the Same DNA Binding Site’, Molecular and Cellular Biology, 18(3), pp. 1266–1274. doi: 10.1128/mcb.18.3.1266

230. Connolly, K. et al. (2022) ‘Potential role of chitinase-3-like protein 1 (CHI3L1/YKL-40) in neurodegeneration and Alzheimer’s disease’, Alzheimer’s & Dementia, 19(1), pp. 9–24. doi: 10.1002/alz.12612

231. Kawada, M. et al. (2011) ‘Chitinase 3-like 1 promotes macrophage recruitment and angiogenesis in colorectal cancer’, Oncogene, 31(26), pp. 3111–3123. doi: 10.1038/onc.2011.498

232. Mizoguchi, E. (2006) ‘Chitinase 3–Like-1 Exacerbates Intestinal Inflammation by Enhancing Bacterial Adhesion and Invasion in Colonic Epithelial Cells’, Gastroenterology, 130(2), pp. 398–411. doi: 10.1053/j.gastro.2005.12.007

233. Cohen, N. et al. (2017) ‘Fibroblasts drive an immunosuppressive and growth-promoting microenvironment in breast cancer via secretion of Chitinase 3-like 1’, Oncogene, 36(31), pp. 4457–4468. doi: 10.1038/onc.2017.65

234. Qiu, Q.-C. et al. (2018) ‘CHI3L1 promotes tumor progression by activating TGF-β signaling pathway in hepatocellular carcinoma’, Scientific Reports, 8(1). doi: 10.1038/s41598-018-33239-8

235. Yang, P.-S. et al. (2022) ‘Targeting protumor factor chitinase-3-like-1 secreted by Rab37 vesicles for cancer immunotherapy’, Theranostics, 12(1), pp. 340–361. doi: 10.7150/thno.65522

236. https://www.mcgill.ca/ose/files/ose/investigating_the_role_of_chi3l1_in_energy_reprogramming_in_her2_breast_cancer.pdf , accessed June 2025

237. Taifour, T. et al. (2023) ‘The tumor-derived cytokine Chi3l1 induces neutrophil extracellular traps that promote T cell exclusion in triple-negative breast cancer’, Immunity, 56(12), pp. 2755-2772.e8. doi: 10.1016/j.immuni.2023.11.002

238. Salembier, R. et al. (2024) ‘Chitin-mediated blockade of chitinase-like proteins reduces tumor immunosuppression, inhibits lymphatic metastasis and enhances anti-PD-1 efficacy in complementary TNBC models’, Breast Cancer Research, 26(1). doi: 10.1186/s13058-024-01815-8

239. Johansen, A.Z. et al. (2022) ‘Chitooligosaccharides Improve the Efficacy of Checkpoint Inhibitors in a Mouse Model of Lung Cancer’, Pharmaceutics, 14(5), p. 1046. doi: 10.3390/pharmaceutics14051046

240. Yu, J.E. et al. (2024) ‘Significance of chitinase-3-like protein 1 in the pathogenesis of inflammatory diseases and cancer’, Experimental & Molecular Medicine, 56(1), pp. 1–18. doi: 10.1038/s12276-023-01131-9 license available at http://creativecommons.org/licenses/by/4.0/

241. Chen, Y. et al. (2024) ‘Peptidoglycan-Chi3l1 interaction shapes gut microbiota in intestinal mucus layer’, eLife, 13. doi: 10.7554/eLife.92994

242. Magnusdottir, U. et al. (2025) ‘Heparin-binding of the human chitinase-like protein YKL-40 is allosterically modified by chitin oligosaccharides’, Biochemistry and Biophysics Reports, 41, p. 101908. doi: 10.1016/j.bbrep.2024.101908 license available at http://creativecommons.org/licenses/by/4.0/

243. Einarsson, J.M. et al. (2013) ‘Partially acetylated chitooligosaccharides bind to YKL-40 and stimulate growth of human osteoarthritic chondrocytes’, Biochemical and Biophysical Research Communications, 434(2), pp. 298–304. doi: 10.1016/j.bbrc.2013.02.122

244. Chen, J. et al. (2023) ‘Characterization of effects of chitooligosaccharide monomer addition on immunomodulatory activity in macrophages’, Food Research International, 163, p. 112268. doi: 10.1016/j.foodres.2022.112268

245. Jitprasertwong, P. et al. (2021) ‘Anti-inflammatory activity of soluble chito-oligosaccharides (CHOS) on VitD3-induced human THP-1 monocytes’, PLOS ONE. Edited by A.M. Abd El-Aty, 16(2), p. e0246381. doi: 10.1371/journal.pone.0246381

246. Wei, J. et al. (2024) ‘Chitooligosaccharides improves intestinal mucosal immunity and intestinal microbiota in blue foxes’, Frontiers in Immunology, 15. doi: 10.3389/fimmu.2024.1506991

247. Cho, C.H. et al. (2024) ‘Modulating intestinal health: Impact of chitooligosaccharide molecular weight on suppressing RAGE expression and inflammatory response in methylglyoxal-induced advanced glycation end-products’, International Journal of Biological Macromolecules, 269, p. 131927. doi: 10.1016/j.ijbiomac.2024.131927

248. Lee, I.-A. et al. (2013) ‘Oral caffeine administration ameliorates acute colitis by suppressing chitinase 3-like 1 expression in intestinal epithelial cells’, Journal of Gastroenterology, 49(8), pp. 1206–1216. doi: 10.1007/s00535-013-0865-3

249. Lee, I.-A. (2014) ‘Novel methylxanthine derivative-mediated anti-inflammatory effects in inflammatory bowel disease’, World Journal of Gastroenterology, 20(5), p. 1127. doi: 10.3748/wjg.v20.i5.1127

250. Xu, D. et al. (2013) ‘Stable RAGE-heparan sulfate complexes are essential for signal transduction.’ ACS Chem Biol;8(7):1611-1620. doi: 10.1021/cb4001553

251. https://www.megazyme.com/hexaacetyl-chitohexaose, accessed May 2025

252. https://web.archive.org/web/20110523080407/http://www-miller.ch.cam.ac.uk/levinthal/levinthal.html, accessed 23 June 2025.

253. Levinthal, Cyrus. “Are There Pathways For Protein Folding?”, Extrait du Journal de Chimie Physique, 1968, 65 no 1, p. 44.

254. Richards, Frederic M. “The Protein Folding Problem.” Scientific American, vol. 264, no. 1, 1991, pp. 54–65. JSTOR, http://www.jstor.org/stable/24936754. Accessed 24 June 2025.

255. Onuchic, J.N., Luthey-Schulten, Z. and Wolynes, P.G. (1997) ‘THEORY OF PROTEIN FOLDING: The Energy Landscape Perspective’, Annual Review of Physical Chemistry, 48(1), pp. 545–600. doi: 10.1146/annurev.physchem.48.1.545

256. Baldovin, M. et al. (2018) ‘The Role of Data in Model Building and Prediction: A Survey Through Examples’, Entropy, 20(10), p. 807. doi: 10.3390/e20100807

257. Radford, S. (2000) ‘Protein folding: progress made and promises ahead’, Trends in Biochemical Sciences, 25(12), pp. 611–618. doi: 10.1016/S0968-0004(00)01707-2

258. P. Willmott, An Introduction to Synchrotron Radiation: Techniques and Applications, Wiley-VCH, 2011.

259. https://www.nottingham.ac.uk/aspire-itn/aspire-blog/aspire-blogs-2020/synchrotron-radiation-and-synchrotron-light-sources.aspx, accessed 20 June 2025

260. Göbel, U. et al. (1994) ‘Correlated mutations and residue contacts in proteins’, Proteins: Structure, Function, and Bioinformatics, 18(4), pp. 309–317. doi: 10.1002/prot.340180402

261. de Juan, D., Pazos, F. and Valencia, A. (2013) ‘Emerging methods in protein co-evolution’, Nature Reviews Genetics, 14(4), pp. 249–261. doi: 10.1038/nrg3414

262. Hopf, T.A. et al. (2014) ‘Sequence co-evolution gives 3D contacts and structures of protein complexes’, eLife, 3. doi: 10.7554/eLife.03430

263. Arne Elofsson, Seminar - Predictions of protein-protein interactions using AlphaFold, Accademia delle Scienze dell’Istituto di Bologna, February 18th, 2022 (accessible at https://www.youtube.com/watch?v=PZfJRahqcds).

264. AlQuraishi, M. (2019) ‘AlphaFold at CASP13.’ Bioinformatics;35(22):4862-4865. doi:10.1093/bioinformatics/btz422

265. Jumper, J. et al. (2021) ‘Highly accurate protein structure prediction with AlphaFold’, Nature, 596(7873), pp. 583–589. doi: 10.1038/s41586-021-03819-2

266. Roney, J.P. and Ovchinnikov, S. (2022) ‘State-of-the-Art Estimation of Protein Model Accuracy Using AlphaFold’, Physical Review Letters, 129(23). doi: 10.1103/PhysRevLett.129.238101

267. Evans, R. et al. (2021) ‘Protein complex prediction with AlphaFold-Multimer’. doi: 10.1101/2021.10.04.463034

268. Mirdita, M. et al. (2022) ‘ColabFold: making protein folding accessible to all’, Nature Methods, 19(6), pp. 679–682. doi: 10.1038/s41592-022-01488-1

269. "Content is based on the EBI AlphaFold Online Course (accessed June 2025), unless otherwise indicated: https://www.ebi.ac.uk/training/online/courses/alphafold."

270. Varga, J.K., Ovchinnikov, S. and Schueler-Furman, O. (2025) ‘actifpTM: a refined confidence metric of AlphaFold2 predictions involving flexible regions’, Bioinformatics. Edited by A. Elofsson, 41(3). doi: 10.1093/bioinformatics/btaf107

271. Wang, W., Gong, Z. and Hendrickson, W.A. (2025) ‘AlphaFold-guided molecular replacement for solving challenging crystal structures’, Acta Crystallographica Section D Structural Biology, 81(1), pp. 4–21. doi: 10.1107/s2059798324011999

272. Monteiro da Silva, G. et al. (2024) ‘High-throughput prediction of protein conformational distributions with subsampled AlphaFold2’, Nature Communications, 15(1). doi: 10.1038/s41467-024-46715-9

273. https://www.signalpeptide.de/index.php, accessed June 2025. Signal Peptide Website: Advanced Search for “MHSSALLCCLVLLTGVRA”.

274. https://web.expasy.org/compute_pi/, accessed 23 June 2024

275. Park, S.J., Kleffmann, T. and Hessian, P.A. (2011) 'The G82S Polymorphism Promotes Glycosylation of the Receptor for Advanced Glycation End Products (RAGE) at Asparagine 81', Journal of Biological Chemistry, 286(24), pp. 21384–21392. doi: 10.1074/jbc.M111.241281

276. https://www.leadgenebio.com/human-rage-his-tag-e-coli-ldg054phe.html, accessed 23 June 2024

277. https://www.sigmaaldrich.com/deepweb/assets/sigmaaldrich/product/documents/361/652/p7367pis-mk.pdf, accessed June 2025

278. Superdex 200 Increase 5/150 GL and Superdex 200 Increase 10/300 GL Instructions for Use 29027271 AI

279. Accessible at https://plip-tool.biotec.tu-dresden.de/plip-web/plip/index (June 2025).

280. Vainauskas, S. et al. (2016) 'Profiling of core fucosylated N-glycans using a novel bacterial lectin that specifically recognizes α1,6 fucosylated chitobiose', Scientific Reports, 6(1). doi: 10.1038/srep34195

281. Rao, F.V. et al. (2005) 'Methylxanthine Drugs Are Chitinase Inhibitors: Investigation of Inhibition and Binding Modes', Chemistry & Biology, 12(9), pp. 973–980. doi: 10.1016/j.chembiol.2005.07.009

282. Hakala B.E., White C., Recklies A.D., Human cartilage gp-39, a major secretory product of articular chondrocytes and synovial cells, is a mammalian member of a chitinase protein family, J. Biol. Chem., 1993, 268, 25803-25810

283. Czestkowski, W. et al. (2024) 'Structure-Based Discovery of High-Affinity Small Molecule Ligands and Development of Tool Probes to Study the Role of Chitinase-3-Like Protein 1', Journal of Medicinal Chemistry, 67(5), pp. 3959–3985. doi: 10.1021/acs.jmedchem.3c02255

284. https://alphafoldserver.com/, accessed June 2025

285. https://github.com/sokrypton/ColabFold?tab=readme-ov-file, accessed June 2025

286. Wayment-Steele, H.K. et al. (2023) 'Predicting multiple conformations via sequence clustering and AlphaFold2', Nature, 625(7996), pp. 832–839. doi: 10.1038/s41586-023-06832-9

287. Abramson, J. et al. (2024) 'Accurate structure prediction of biomolecular interactions with AlphaFold 3', Nature, 630(8016), pp. 493–500. doi: 10.1038/s41586-024-07487-w

288. Abramson, J. et al. (2024) 'Accurate structure prediction of biomolecular interactions with AlphaFold 3', Nature, 630(8016), pp. 493–500. doi: 10.1038/s41586-024-07487-w license available at http://creativecommons.org/licenses/by/4.0/

289. https://www.ebi.ac.uk/training/online/courses/alphafold/inputs-and-outputs/evaluating-alphafolds-predicted-structures-using-confidence-scores/confidence-scores-in-alphafold-multimer/, accessed 23 June 2025

290. Butterfield, N.J. (2015) 'The Neoproterozoic.' Curr Biol;25(19):R859-R863. doi: 10.1016/j.cub.2015.07.021

291. Grossnickle, D.M. et al. (2019) 'Untangling the Multiple Ecological Radiations of Early Mammals' Trends in Ecology & Evolution', Volume 34, Issue 10, 936 - 949. doi: 10.1016/j.tree.2019.05.008

Appendix: Preliminary Analysis for ColabFold Prediction

1. Screening 128 seeds with default parameters and using custom templates (Colab's A100 GPU)

At the beginning, we attempted to tune the boolean parameters, specifically use_mlm, use_dropout, use_cluster_profile, while keeping the others as default, but they resulted in no significant changes. Therefore, we proceeded to screen multiple seeds (128), keeping these parameters as True so as to maximize the reliability across different predictions. Later, we would learn that the parameters tuning the MSA are likely the ones to determine the most important changes. However, with the default options we already found a few seeds to be significant.

We decided to screen custom templates first, being our coevolution signal (with the default MSA parameters) low, and wanting to rely more on existing structures. We also kept *unpaired_paired* as pair mode because, as shown above among the MSA parameters, the coevolution signal did not show significant changes with respect to *unpaired*. Later we would see that unpaired is a better choice for our complex. We screened three custom templates at low depth of recycles (0+3):

- 1nwr : 3cjj(A) -> seeds 18, 60, 90
- 1nwr : 4ybh(A) -> no seeds identified
- 1nwr : 4lp5(B) -> seeds 18, 28, (103)**.** We tested these seeds also on 1nwr : 4lp5(A) -> seeds 18, (103)

CHI3L1's PDB template	RAGE's PDB template	Chains (CHI3L1 : RAGE)	Seed	Recycle n° of best (/24)	pLDDT of best	pTM of best	iPTM of best	Multi of best
1nwr	3cjj	A : A	18	19	0.928	0.879	0.778	0.798
1nwr	3cjj	A : A	60	18	0.930	0.883	**0.785**	**0.804**
1nwr	3cjj	A : A	90	19	0.926	0.867	0.750	0.773
1nwr	4lp5	A : B	18	17	0.931	0.880	**0.782**	**0.801**
1nwr	4lp5	A : B	28	22	0.930	0.880	0.780	0.800
1nwr	4lp5	A : B	103	23	0.930	0.870	0.760	0.782
1nwr	4lp5	A : A	18	17	0.930	0.873	0.767	0.788
1nwr	4lp5	A : A	103	23	0.930	0.868	0.754	0.777

Supplementary Table 1.

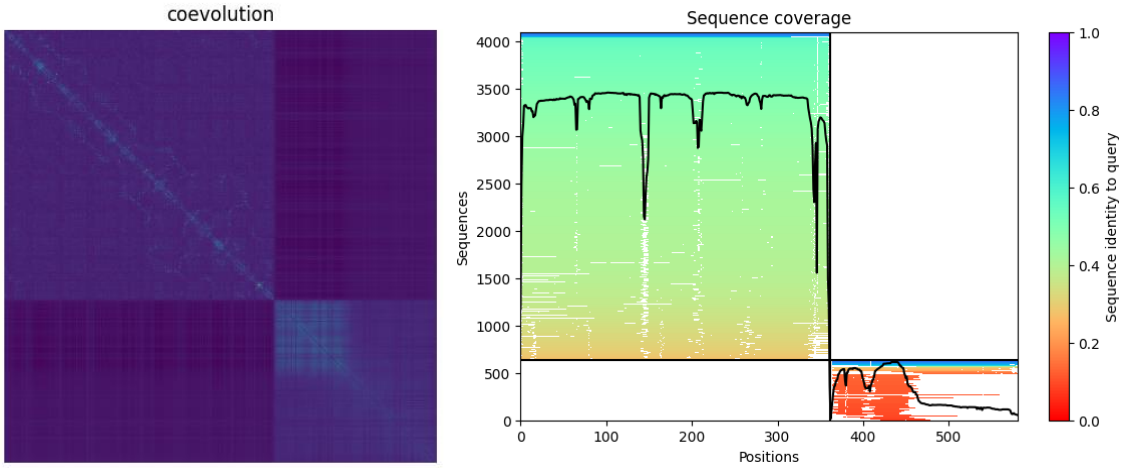
2. Parameters optimization on specific seeds

Manual optimization

We then focused on the template 3cjj and seed 60, being the one with the highest confidence, and we started to manually optimize some parameters. In particular, we found that:

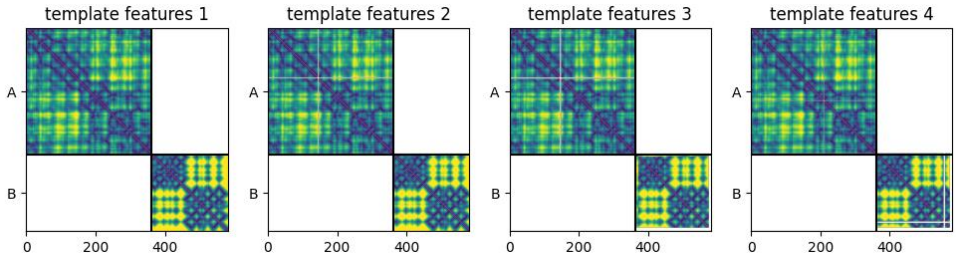
A. Unpaired instead of unpaired_paired -> intermediate new high (i_ptm = 0.812, multi = 0.829). This could be likely interpreted as the inclusion of paired sequences in "unpaired_paired" did not significantly contribute to prediction accuracy and narrowed the co-evolutionary signals. As shown below, the coverage signal for RAGE (right bottom quadrant) is now at least visible and shows some pattern of alignment, especially in the first part of the protein.

B. Unpaired with cov 50, 90 -> worse, with cov 25 -> intermediate new high (i_ptm = 0.842, multi = 0.855). Same prediction with unpaired_paired is extremely inaccurate. The improvement with low coverage threshold has likely happened thanks to increase in the sequence aligned with RAGE and a resulting stronger coevolution signal, as below.



Supplementary Figure 1. Coevolution signal with cov 25 (left): RAGE (lower right quadrant) has a visibly stronger signal, especially in the V domain. Sequence coverage signal with cov 25 (right): we can see that RAGE (right lower quadrant) has now a visible coverage and shows some patterns of alignment, especially in the V domain.

C. We also tested the mmseqs2 template, which does not use a pdb (as 3cjj) but a template automatically generated from the MSA -> new intermediate high (i_ptm = 0.843, multi = 0.856). The resulting template features appear as below.



[['1hjx', '3cjj'], ['1hjx', '4p2y'], ['1hjx', '3o3u'], ['1owq', '7lmw']]

Automated optimization

After having manually identified a satisfactory combination of parameters, we systematically screened all the combinations of parameters on the 3 seeds identified on the 3cjj template. The testing conditions were obtained by combining all the following parameters: cov [0, 25, 50, 75, 90, 99], id [90, 100], qid [0, 10, 15, 20, 30], pair_mode [unpaired, unpaired_paired]; for a total of 120 combinations for each seed (60 if pair_mode is fixed).
We used and adapted the following code to run the screening:

```
%%time
#@title run_alphafold
#@markdown Model options
model = "1" #@param ["1", "2", "3", "4", "5", "all"]
num_recycles = 24 #@param ["0", "1", "2", "3", "6", "12", "24"] {type:"raw"}
recycle_early_stop_tolerance = 0.0 #@param ["0.0", "0.5", "1.0"] {type:"raw"}
select_best_across_recycles = False #@param {type:"boolean"}
#@markdown Stochastic options
use_mlm = True #@param {type:"boolean"}
use_dropout = True #@param {type:"boolean"}
seeds = [60, 61, 62] #@param {type:"raw"} # List of specific seed numbers
#@markdown extras
show_images = True #@param {type:"boolean"}

run_opts = {
   "seeds": seeds,
   "use_mlm": use_mlm,
   "use_dropout": use_dropout,
   "num_recycles": num_recycles,
   "model": model,
   "use_initial_guess": use_initial_guess,
   "select_best_across_recycles": select_best_across_recycles,
   "recycle_early_stop_tolerance": recycle_early_stop_tolerance
}

# Decide which models to use
if model == "all":
   models = af._model_names
else:
   models = [af._model_names[int(model) - 1]]

# Set options
af.set_opt("mlm", replace_fraction=0.15 if use_mlm else 0.0)

# Define parameter lists for combinations
cov_list = [0, 25, 50, 75, 90, 99]
qid_list = [0, 10, 15, 20, 30]
id_list = [90, 100]
pair_mode_list = ["unpaired_paired", "unpaired"]

# Generate all possible combinations
from itertools import product
combinations = list(product(cov_list, qid_list, id_list, pair_mode_list))

# Loop over all combinations
for i, (cov, qid, id, pair_mode) in enumerate(combinations):
   # Create a unique identifier and subfolder for this combination
   combination_id = f"comb_{i}_cov{cov}_qid{qid}_id{id}_pair_{pair_mode}"
   sub_jobname = f"{jobname}/{combination_id}"
   os.makedirs(sub_jobname, exist_ok=True)
```

```
# Generate MSA with current parameters
msa, deletion_matrix = predict.get_msa(
    u_sequences, sub_jobname,
    mode=pair_mode,
    cov=cov, id=id, qid=qid, max_msa=4096,
    do_not_filter=do_not_filter,
    mmseqs2_fn=run_mmseqs2_wrapper,
    hhfilter_fn=run_hhfilter
)

# Update the model with the new MSA
af.set_msa(msa, deletion_matrix)

# Set output path for this combination
pdb_path = f"{sub_jobname}/pdb"
os.makedirs(pdb_path, exist_ok=True)

# Reset temporary storage for this combination
af._tmp = {
    "traj": {"seq": [], "xyz": [], "plddt": [], "pae": []},
    "log": [],
    "best": {}
}

# Initialize info list for this combination
info = []

# Run prediction
print(f"Running prediction for combination {combination_id}")
with open(f"{sub_jobname}/log.txt", "w") as handle:
    for seed in seeds:
        af.set_seed(seed)
        for model in models:
            recycle = 0
            af._inputs.pop("prev", None)
            stop_recycle = False
            prev_pos = None
            while recycle < num_recycles + 1:
                print_str = f"seed={seed} model={model} recycle={recycle} combination={combination_id}"
                af.predict(dropout=use_dropout, models=[model], verbose=False)
                af._inputs["prev"] = af.aux["prev"]
                if len(af._lengths) > 1:
                    af.aux["log"]["multi"] = 0.8 * af.aux["log"]["i_ptm"] + 0.2 * af.aux["log"]["ptm"]
                af.save_current_pdb(f"{pdb_path}/{model}_r{recycle}_seed{seed}.pdb")
                for k in print_key:
                    print_str += f" {k}={af.aux['log'][k]:.3f}"
                current_pos = af.aux["atom_positions"][:,1]
                if recycle > 0:
                    rmsd_tol = _np_rmsd(prev_pos, current_pos, use_jax=False)
                    if rmsd_tol < recycle_early_stop_tolerance:
                        stop_recycle = True
                    print_str += f" rmsd_tol={rmsd_tol:.3f}"
                prev_pos = current_pos
                print(print_str)
                handle.write(f"{print_str}\n")
                tag = f"{model}_r{recycle}_seed{seed}"
                if select_best_across_recycles:
                    info.append([tag, print_str, af.aux["log"][rank_by]])
                    if calc_extended_ptm:
                        extended_ptms = extended_ptm.get_chain_and_interface_metrics(
                            af.aux['debug']['outputs'], af._inputs['asym_id']
                        )
                        info[-1].extend([
                            extended_ptms['pairwise_iptm'],
                            extended_ptms['pairwise_actifptm'],
```

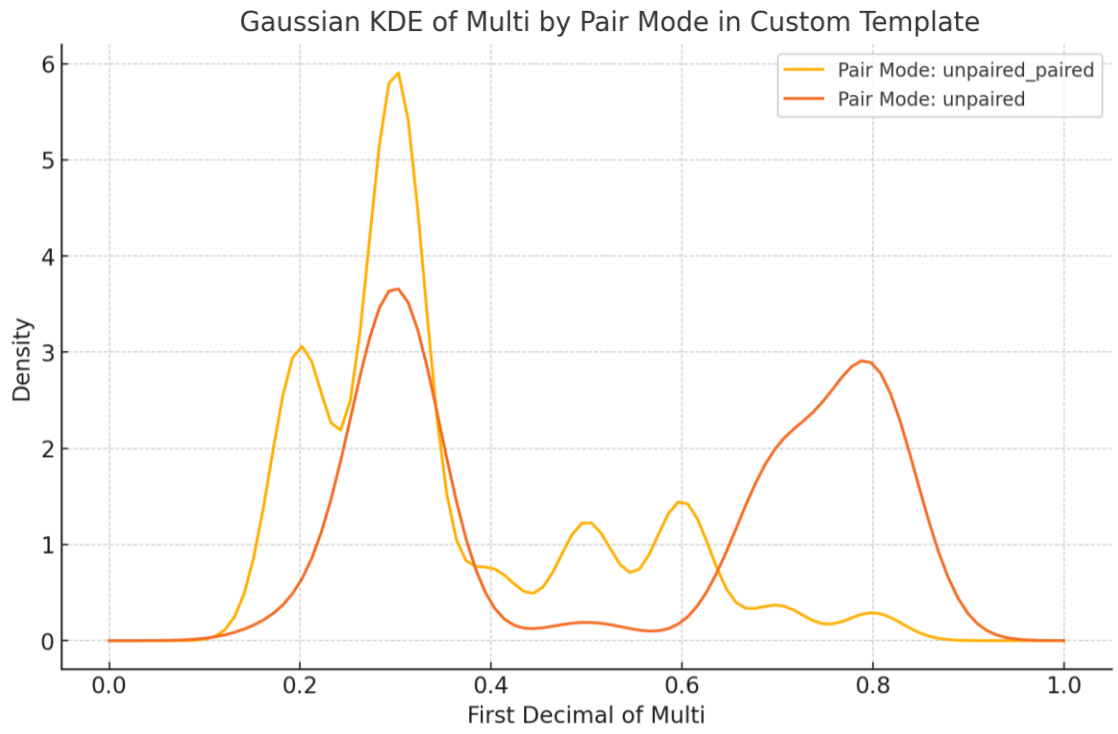
```
                    extended_ptms['per_chain_ptm']
                ])
            af._save_results(save_best=True, best_metric=rank_by, metric_higher_better=True, verbose=False)
            af._k += 1
        recycle += 1
        if stop_recycle:
            break
    if not select_best_across_recycles:
        info.append([tag, print_str, af.aux["log"][rank_by]])
        if calc_extended_ptm:
            extended_ptms = extended_ptm.get_chain_and_interface_metrics(
                af.aux['debug']['outputs'], af._inputs['asym_id']
            )
            info[-1].extend([
                extended_ptms['pairwise_iptm'],
                extended_ptms['pairwise_actifptm'],
                extended_ptms['per_chain_ptm']
            ])
        af._save_results(save_best=True, best_metric=rank_by, metric_higher_better=True, verbose=False)
        af._k += 1
    plot_3D(af.aux, Ls * copies, f"{pdb_path}/{model}_seed{seed}.pdf", show=show_images)
    predict.plot_confidence(af.aux["plddt"]*100, af.aux["pae"], Ls * copies)
    plt.savefig(f"{pdb_path}/{model}_seed{seed}.png", dpi=200, bbox_inches='tight')
    plt.close()

# Save best results for this combination
if info:
    rank = np.argsort([x[2] for x in info])[::-1][:5]
    print(f"best_tag={info[rank[0]][0]} {info[rank[0]][1]}")
    aux_best = af._tmp["best"]["aux"]
    af.save_pdb(f"{pdb_path}/best.pdb")
    np.savez_compressed(
        f"{pdb_path}/best.npz",
        plddt=aux_best["plddt"].astype(np.float16),
        pae=aux_best["pae"].astype(np.float16),
        tag=np.array(info[rank[0]][0]),
        metrics=np.array(info[rank[0]][1]),
        iptm_pairwise=np.array(info[rank[0]][3]) if len(info[rank[0]]) > 3 else np.array([]),
        actifptm_pairwise=np.array(info[rank[0]][4]) if len(info[rank[0]]) > 4 else np.array([]),
        cptm=np.array(info[rank[0]][5]) if len(info[rank[0]]) > 5 else np.array([])
    )
    np.savez_compressed(
        f"{pdb_path}/all.npz",
        plddt=np.array(af._tmp["traj"]["plddt"], dtype=np.float16),
        pae=np.array(af._tmp["traj"]["pae"], dtype=np.float16),
        tag=np.array([x[0] for x in info]),
        metrics=np.array([x[1] for x in info]),
        iptm_pairwise=np.array(info[rank[0]][3]) if len(info[rank[0]]) > 3 else np.array([]),
        actifptm_pairwise=np.array(info[rank[0]][4]) if len(info[rank[0]]) > 4 else np.array([]),
        cptm=np.array(info[rank[0]][5]) if len(info[rank[0]]) > 5 else np.array([])
    )
    plot_3D(aux_best, Ls * copies, f"{pdb_path}/best.pdf", show=False)
    predict.plot_confidence(aux_best["plddt"]*100, aux_best["pae"], Ls * copies)
    plt.savefig(f"{pdb_path}/best.png", dpi=200, bbox_inches='tight')
    plt.close()
```

Screening	Seeds	Depth (recycles)	Parameter combinations	Template	Learnings

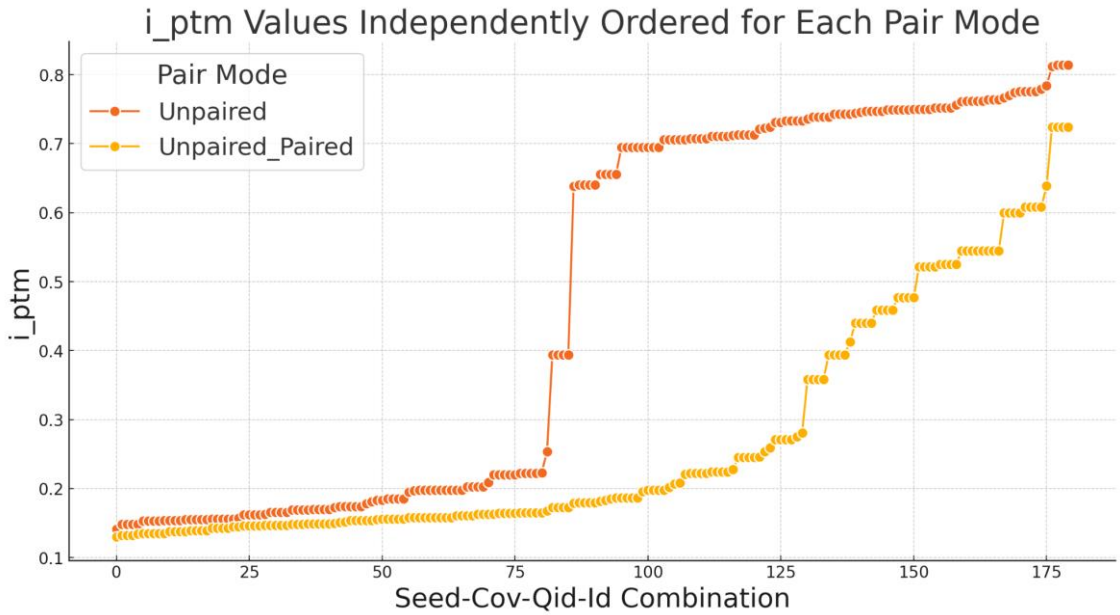
1) Optimizing the best seeds at low depth [360 combinations]	18, 60, 90	0 + 3	cov, id, qid, pair_mode	custom (1nwr:3cjj, A:A) (with cluster_profile)	qid gives equal results when between 0-15 if cov is 0 and id 100, and between 0-20 for all other combinations of cov and id + unpaired better than paired + cluster better than no cluster + mmseqs2 better than custom template
2) Optimizing the best seeds at low depth [180 combinations]	18, 60, 90	0 + 3	cov, id, qid, *unpaired*	mmseqs2 (without cluster_profile)	
3) Optimizing the best seeds at low depth [180 combinations]	18, 60, 90	0 + 3	cov, id, qid, *unpaired*	mmseqs2 (with cluster_profile)	

Supplementary Table 2. *Initial automated screening results*.

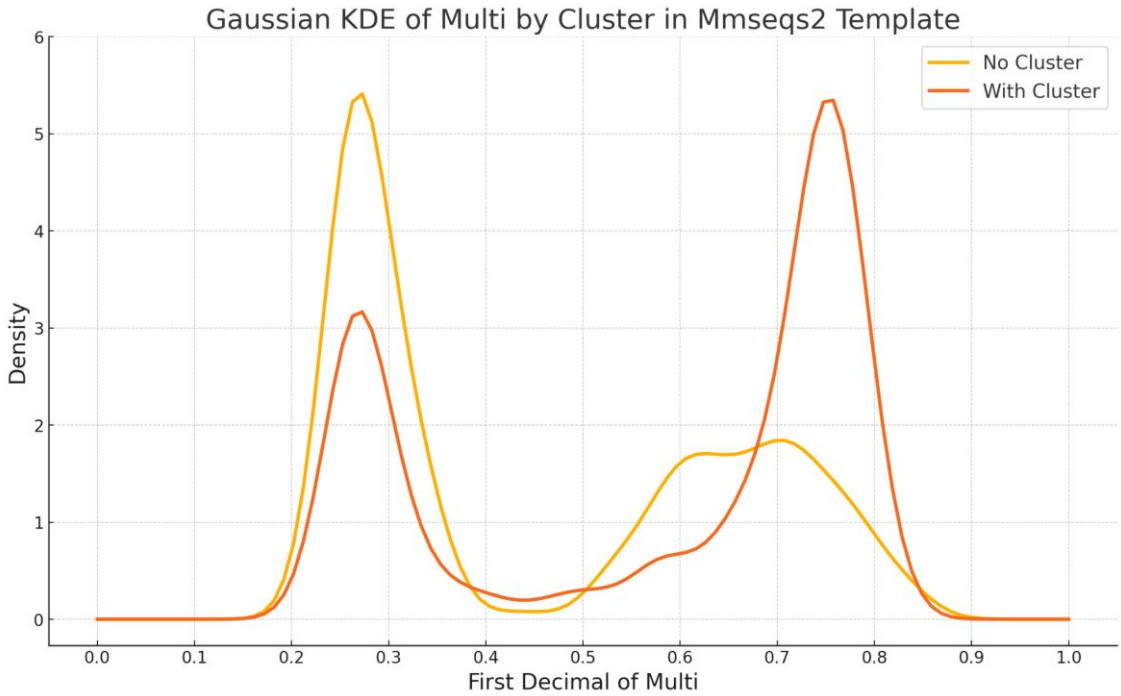


Supplementary Figure 2. On the custom template screening (1), we could evaluate the difference between unpaired and unpaired_paired. As we can see above, the unpaired mode better separated failure

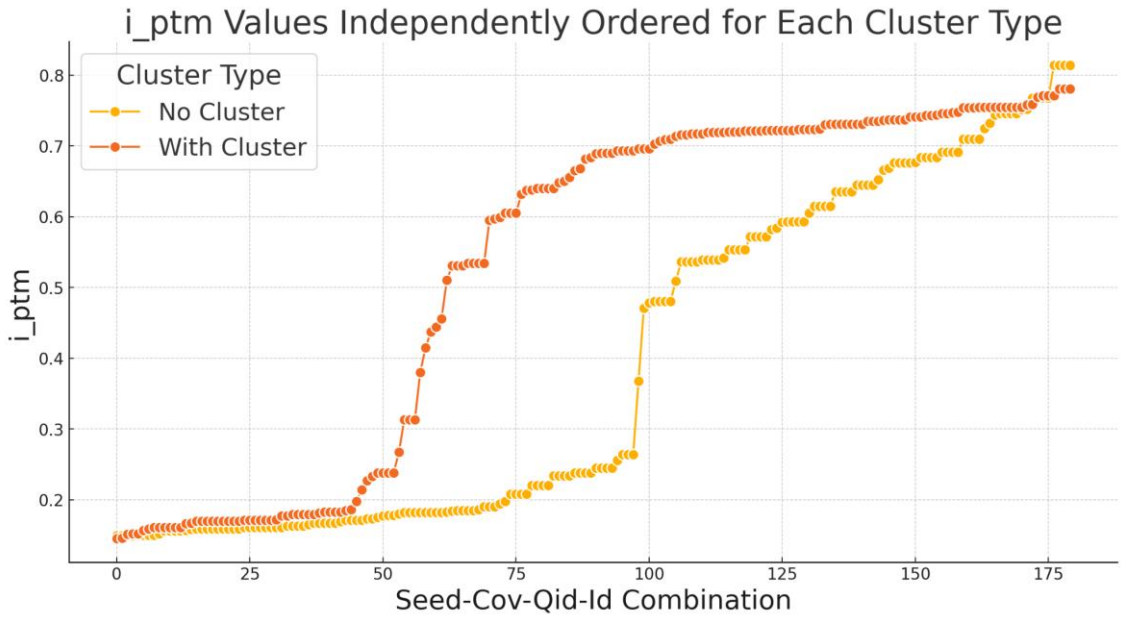
from success and achieved more frequently high multi scores, which were instead very rarely achieved with unpaired_paired mode, over the different combinations.



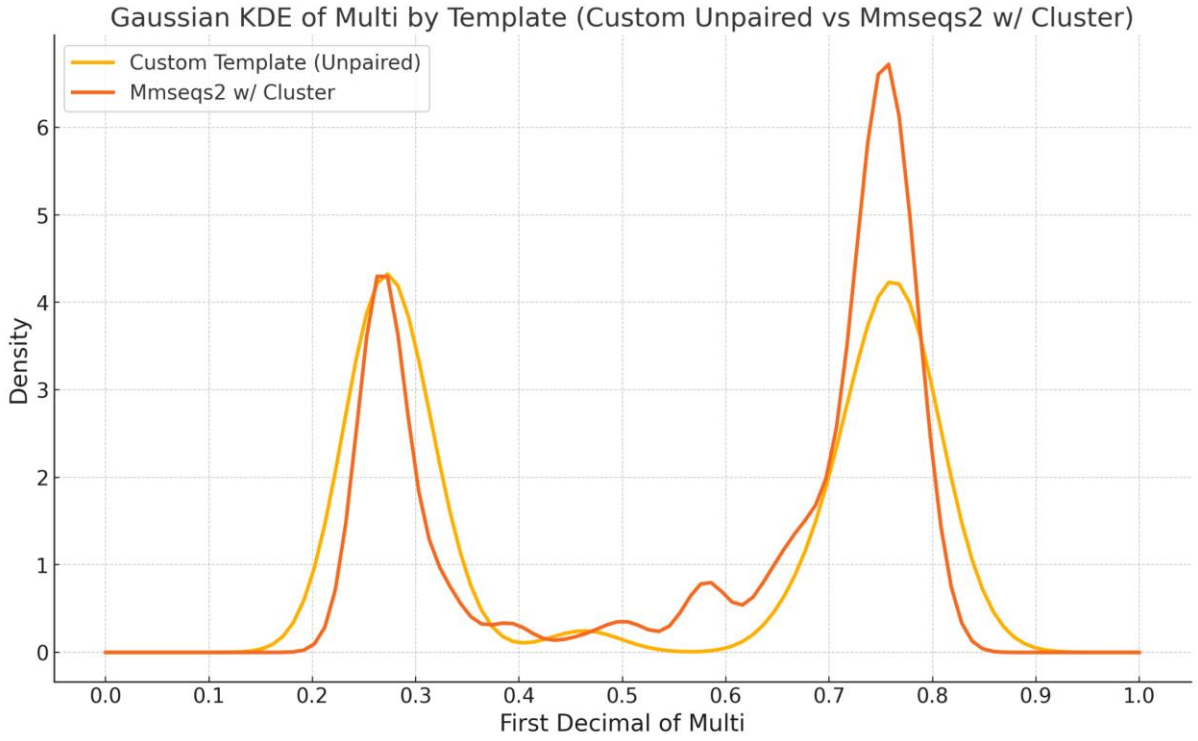
Supplementary Figure 3. The same screening (1) is shown above. By ordering based on the i_ptm value both the pair modes independently of the specific combination of parameters (displayed on the x axis, with each point representing one combination but not the same one with respect to the two pair modes), we can see that unpaired had consistently higher prediction accuracy on all the possible combinations.



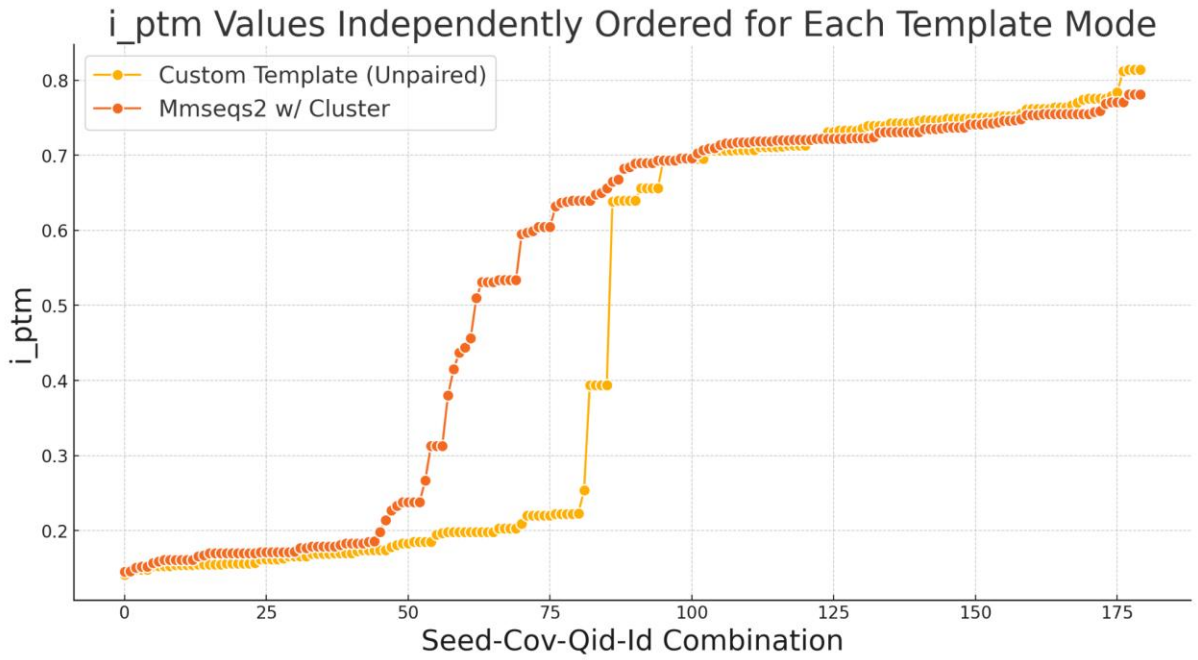
Supplementary Figure 4. On the mmseqs2 template screening (2 and 3), we could evaluate the difference between clustering or not of the MSA. As we can see above, the cluster mode better separated failure from success and achieved more frequently high multi scores, which were instead rarely achieved with unpaired_paired mode, over the different combinations.



Supplementary Figure 5. The same screenings (2 and 3) are shown above. By ordering based on the i_ptm value both the cluster mode independently of the specific combination of parameters (displayed on the x axis, with each point representing one combination but not the same one with respect to the two pair modes), we can see that cluster mode had almost always higher prediction accuracy on the possible combinations.



Supplementary Figure 6. *The two template modes evaluated above had a similar distribution of the multi scores, but the mmseqs2 with cluster achieved a higher percentage of high multi scores. However, the tail of the custom template curve at the highest end of the multi scores achieved bigger peaks.*



Supplementary Figure 7. *The same comparison on template modes is shown above. As before, we can see that the mmseqs2 with cluster achieved a higher number of high multi scores but some of the highest peaks in multi scores belong to the unpaired custom template.*

1) Full depth screening after removing duplicates and filtering for i_pTM > 0.2 in at least one recycle from screening 1 (template)	18, 60, 90	0 + 24	40 selected as promising among the unpaired mode	custom (1nwr:3cjj, A:A) (with cluster_ profile)	seed 90, **cov 25**, **id 90**, **qid 20** and seed 60, **cov 25**, **id 90**, **qid 20** were the highest with i_ptm 0.842 and multi 0.855
2) Full depth screening after removing duplicates and filtering for i_pTM > 0.2 in at least one recycle from screening 3 (mmseqs2)	18, 60, 90	0 + 24	24 selected as promising and belonging to both clustered and unclustered	mmseqs2 (with cluster_ profile)	seed 90, **cov 25**, **id 90**, **qid 20** reached **i_ptm 0,846** and **multi 0,859** -> new interm. high

Supplementary Table 3. *Final full depth screening results*.

The automated extensive screening, in the end, allowed us to obtain just a slight improvement (0.3% in i_ptm) with respect to the manual optimization of a single combination. These screenings were done without optimizing all possible parameters, too.
This prompted us to think about a less expensive and more efficient strategy, leveraging our learnings. This is presented in the Results section.