A universal residue numbering scheme for the Immunoglobulin-fold (Ig-fold) to study Ig-Proteomes and Ig-Interactomes

Caesar Tawfeeq[1], Jiyao Wang[2], Umesh Khaniya[3], Thomas Madej[2], James Song[2], Ravinder Abrol[1*], and Philippe Youkharibache[3*].

[1]Department of Chemistry and Biochemistry, California State University Northridge, CA 91330, USA
[2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA
[3]Cancer Data Science Laboratory, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA.

[*]Corresponding authors. e-mail address: abrol@csun.edu; philippe.youkharibache@nih.gov

ABSTRACT 3
KEYWORDS 3
INTRODUCTION 3
Proteins containing the Ig fold in the Human proteome 3
The Ig fold 6
The classical Ig-domains: IgV, IgI, IgC1 and IgC2 6
Sequence patterns of classical Ig-domains 7
Known topostructural variants of the Ig-fold beyond the classical Ig-domains 8
RESULTS and DISCUSSION 10
Ig, Ig-like and Ig-extended domains 10
Structural invariance and topological variability of the Ig-fold 10
The irreducible structural signature of the Ig-fold 10
Qualifying the variability of domains sharing the Ig-fold 11
Towards positional structural bioinformatics 12
Positional bioinformatics 12
Requirements for a numbering scheme describing the Ig-fold 12
Topostructural variants reference dataset 13
The original Kabat numbering and its variants for IgV domains 13
Definition of the IgStrand numbering 14
Defining positional structural anchors “ij50” in the Ig-fold beta strands 15
Defining canonical and additional beta strand numbers (i and j) 15
IgStrand Structural Anchors 16
Ig-extended domains: Ig-fold topostructural variable regions 20
Arrestin 21
Lamin 22
Transcription factors Ig-domains 23
Ig-domains in viruses 25
Ig or not Ig? Blurring (or swapping) the limits of the Ig-fold 26
Three levels of organized complexity: Ig domains, chains, assemblies 28
Self assembly of Ig-domains 28

The canonical quaternary antibody IgVH-IgVL and IgCH1-IgCL interactomes 29
Comparison of IgV-IgV quaternary interfaces beyond antibodies 31
Comparison of more complex Ig-chains interfaces 33
Pseudo-quaternary structures in multi-domain chains 33
Example of Ig-chains forming intrachain pseudo-quaternary Ig-Ig interfaces 33
Multi Ig-domains intrachain and interchain quaternary interfaces 35
METHODS & DATASETS 37
Ig-domains Nomenclature and Coloring scheme 37
iCn3D Ig domain detection 38
Set up the Ig topostructural variant templates 38
Detect Ig domains 38
Assign Ig reference numbers 39
How to use iCn3D to assign Ig reference numbers 39
CONCLUSION 39
ACKNOWLEDGMENTS 41
REFERENCES 41

ABSTRACT

The Immunoglobulin fold (Ig-fold) is found in proteins from all domains of life and represents the most populous fold in the human genome, with current estimates ranging from 2 to 3% of protein coding regions. That proportion is much higher in the surfaceome where Ig and Ig-like domains orchestrate cell-cell recognition, adhesion and signaling. The ability of Ig-domains to reliably fold and self-assemble through highly specific interfaces represents a remarkable property of these domains, making them key elements of molecular interaction systems: the immune system, the nervous system, the vascular system and the muscular system. We define a universal residue numbering scheme, common to all domains sharing the Ig-fold in order to study the wide spectrum of Ig-domain variants constituting the Ig-proteome and Ig-Ig interactomes at the heart of these *systems*. The "IgStrand numbering scheme" enables the identification of Ig structural proteomes and interactomes in and between any species, and comparative structural, functional, and evolutionary analyses. We review how Ig-domains are classified today as topological and structural variants and highlight the *"Ig-fold irreducible structural signature"* shared by all of them. The IgStrand numbering scheme lays the foundation for the systematic annotation of structural proteomes by detecting and accurately labeling Ig-, Ig-like and Ig-extended domains in proteins, which are poorly annotated in current databases and opens the door to accurate machine learning. Importantly, it sheds light on the robust *Ig protein folding algorithm* used by nature to form beta sandwich supersecondary structures. The numbering scheme powers an algorithm implemented in the interactive structural analysis software iCn3D to systematically recognize Ig-domains, annotate them and perform detailed analyses comparing any domain sharing the Ig-fold in sequence, topology and structure, regardless of their diverse topologies or origin. The scheme provides a robust fold detection and labeling mechanism that reveals unsuspected structural homologies among protein structures beyond currently identified Ig- and Ig-like domain variants. Indeed, multiple folds classified independently contain a common structural signature, in particular jelly-rolls. Examples of folds that harbor an "Ig-extended" architecture are given. Applications in protein engineering around the Ig-architecture are straightforward based on the universal numbering.

KEYWORDS

IgSF, Ig-fold, Ig-domain, Ig-like, Ig-extended, PPI, Proteome, Interactomes, beta barrels, protein folding, beta sandwich proteins.

INTRODUCTION

Proteins containing the Ig fold in the Human proteome

Protein domains sharing the Ig-fold are often referred to as belonging to the Ig Superfamily (IgSF). Ig-domains have been identified as **the most populous superfamily in the human genome, with an initial estimate of over 2% of coding genes** (Lander et al. 2001). The Ig-fold is found in all domains of life, yet nowhere as pervasively as in vertebrates where it is essential in multiple intracellular, extracellular *and intercellular* communication functions. As such, it is a key element of the immune system, the nervous system, the vascular system and the muscular system. Figure 1 shows a sample of protein families containing Ig domains, yet it would take a very large encyclopedia to cover the reach of Ig-fold in all life forms or even just in humans. To date, it has been found in numerous proteins in a myriad of topological variants, in all subcellular compartments from the nucleus to the cytoplasm to extracellular regions. It can be considered as **a wunderkind of evolution**. The question we are addressing is: **what makes it such a superfold?** How can we uncover the reasons for its success, powering so many diverse biological functions?

How can we study its folding across so many variant forms, so many functions, in so many unicellular and multicellular species, in so many diverse tertiary chain arrangements or quaternary molecular complexes?

Ig-like folds have been found in viruses, bacteria, archaea and eukaryotes (Halaby and Mornon 1998; Farré et al. 2017; Fraser et al. 2006)(Bateman, Eddy, and Chothia 1996), and in that latter kingdom from unicellular to multicellular organisms (Louis Du Pasquier, Zucchetti, and De Santis 2004). They can be found in the nucleus or the nucleus surface (Lamin, POM210, P53, NFκB, PTEN), in the cytoplasm (β-Arrestins, PKC, PI3K, PLC, Synaptotagmin), and on the cell surface (CD4, CD8, CTLA-4, TCR, NCAM, ICAM), and even as secretable proteins such as antibodies and transthyretin. It is fair to say that the Ig-fold is a wunderkind of structural evolution, irrespective of divergent or convergent evolutionary paths. To aim at understanding key elements governing the folding of the Ig-fold in its multiple domain variants, tertiary architectures of chains composed of Ig domains and their quaternary assemblies, **we seek a way to compare and analyze all protein domains that share the Ig-fold, their tertiary chains and quaternary assemblies.**

Immunoglobulin (Ig) or Ig-like domains expressed on the surface of cells are of particular interest, as they orchestrate cell-cell recognition and adhesion to trigger further cell signaling and gene expression through receptor/ligand interactions. A remarkable aspect of the Ig-based cell communication system is the symmetry between receptors and ligands, *both* made of Ig- or Ig-like domains, making the Ig-fold a unique and special element of the immune system from which it inherits its name (Halaby and Mornon 1998; Watson et al. 2005; Louis Du Pasquier 2004; L. Du Pasquier 2000; Cooper and Alder 2006; Naeim et al. 2018; Janes et al. 2023); the nervous system (Lawrence Shapiro, Love, and Colman 2007; Honig and Shapiro 2020; Tong and Zou 2019)(Zinn and Özkan 2017; Cortés, Pak, and Özkan 2023; S. Cheng, Park, et al. 2019; Xu et al. 2018; Boni et al. 2022), the vascular system (Shibuya 2008, 2013)(Larrivée et al. 2009), and the muscular system (Labeit and Kolmerer 1995; H. Li et al. 2002; Eckels et al. 2018), both skeletal and cardiac (Benian and Mayans 2015; Giganti et al. 2018; Z. Wang and Raunser 2023). This systemic omnipresence in complex organisms ***underscores the importance of the Ig-fold in health and disease.*** Indeed, a myriad of diseases are rooted in Ig-domain dysfunctions that range from cancer and autoimmune diseases to muscular diseases (Dittmer et al. 2014; Otey et al. 2009), to vascular diseases (Shibuya 2013), to amyloidosis (Brumshtein et al. 2018; Wilkinson et al. 2023). As science and medicine start exploring the immune-muscle system interface, the Ig-domains appear central, with skeletal muscles expressing immunomodulatory molecules such as ICAM-1, ICOSL, CD80 or PD-L1 (Afzali et al. 2018; Nelke et al. 2019). It is also the case in the immune-nervous systems interface (Steinman 2004; Morimoto and Nakajima 2019; Dantzer 2018; Ren and Dubner 2010; Elenkov et al. 2000; Salvador, de Lima, and Kipnis 2021; Mancusi and Monje 2023) and the neuroimmune cardiovascular interface where, for example, the parallel role of semaphorins in neurons and immune cells has become an intense field of research (Mohanta et al. 2022; Yin et al. 2022; Mizuno, Nakanishi, and Kumanogoh 2023).

*To enable the systematic and parallel study of **the Ig-folded elements** of the molecular and cellular systems they articulate, we need a means to bring them into a common frame of reference in sequence, topology and structure. In the following, we define the IgStrand universal reference numbering scheme to identify exactly and identically any **position in sequence and structure** of a protein domain exhibiting an Ig-fold.*

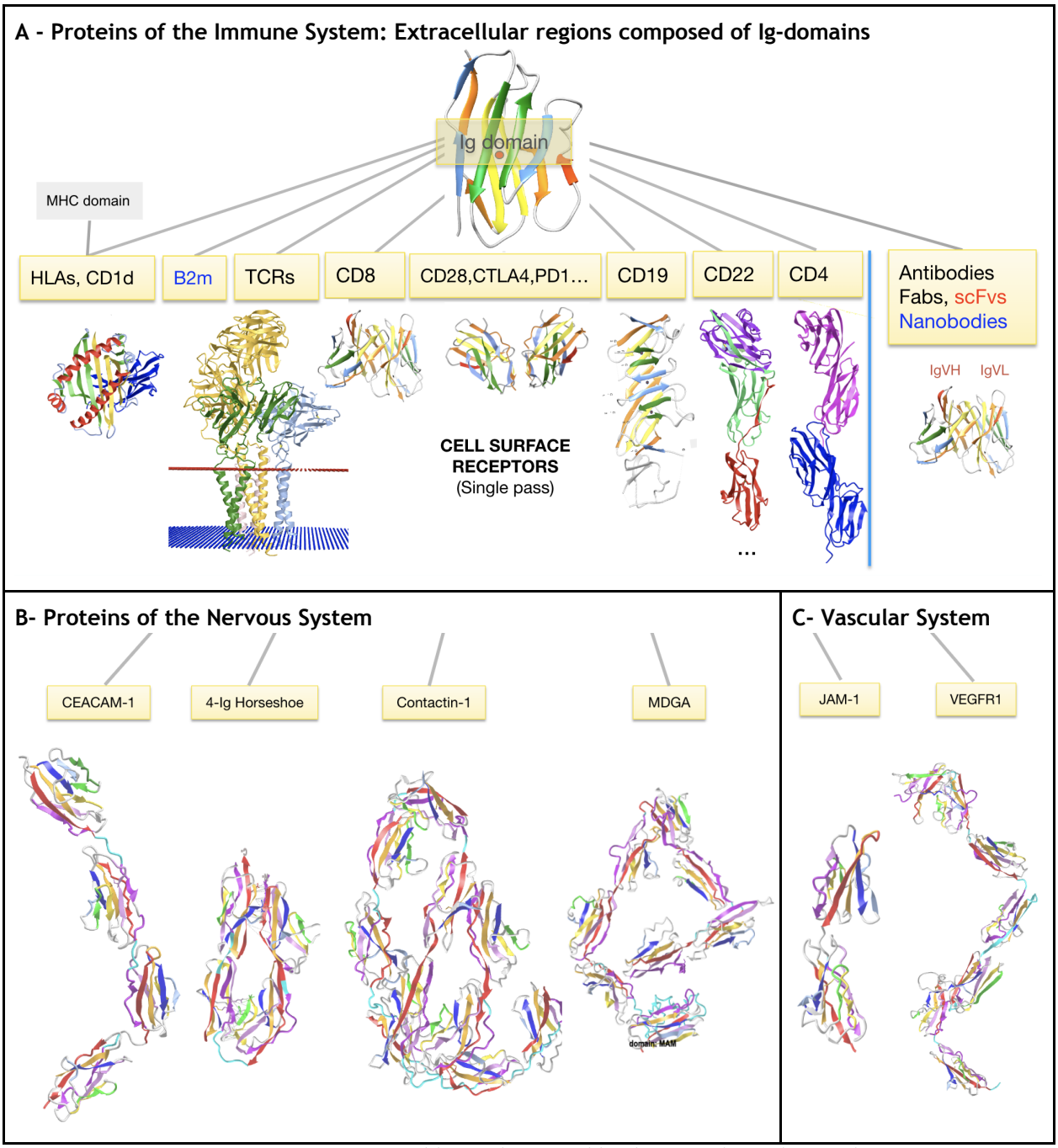


Figure 1: Diverse proteins containing a variable number of Ig-like domains chained in tandem. A) immune system proteins: TCRs, MHCs, antibodies and many other cell surface receptors extracellular regions contain Ig-domains (CD8, CD4, CD22, CD28, PD1, CTLA4 etc.). **B) Nervous system** proteins: CEACAMs, contactins, that like sidekick or DSCAM proteins contain a 4-Ig horseshoe "superdomain" at the N-terminus (see section on Ig-Ig interfaces). MDGA proteins (MAM domain-containing glycosylphosphatidylinositol anchor) contain 7 Ig-domains followed by a MAM domain that has a jelly roll fold with similarities with the Ig-fold. **C) Vascular system** proteins: JAM proteins contain 2 Ig-domains in tandem, and VEGFR act as receptors for VEGF (Vascular endothelial growth factor) proteins. 3D models can be visualized in iCn3D links: Contactin1 https://structure.ncbi.nlm.nih.gov/icn3d/share.html?imRwYoJKmCpCaXYm7&t=Q12860; CEACAM1 https://structure.ncbi.nlm.nih.gov/icn3d/share.html?g89JayoAbbjN7Yg18&t=P13688; VCAM1 https://structure.ncbi.nlm.nih.gov/icn3d/share.html?9edsvYBJ43ZyPv9j8&t=P19320; VEGFR1

https://structure.ncbi.nlm.nih.gov/icn3d/share.html?SbL83DUuwccuJWtz8&t=Q8QHL3; JAM1
https://structure.ncbi.nlm.nih.gov/icn3d/share.html?6mGp5wdHxcxV4Cv98&t=1NBQ

Many cell surface proteins of the nervous system are composed of Ig-domains chained in tandem in their extracellular region and are involved in cell adhesion (CAMs: NCAM, CEACAM, L1CAM, MDGA, OBCAM, SYnCAM, etc.). Some, such as contactins, Sidekick proteins, DSCAM and others, form a horseshoe "superdomain" with the four N-terminal Ig domains of the extracellular Ig-chain, and contain FN3 domains in the C-terminal membrane proximal region of the extracellular Ig-chain. Cell adhesion Molecules in the vascular system are also built out of Ig-domains (ICAM, VCAM, PECAM, JAM, etc.). For example, Vascular Endothelial Growth Factor Receptor (VEGFR) is a Receptor-type Tyrosine Kinase (RTK) consisting of a ligand-binding region with seven Ig-like domains as its extracellular region. In vertebrates it plays essential roles in the regulation of angiogenesis and lymphangiogenesis. In the muscular system Obscurin and Titin (Benian and Mayans 2015) are giant macromolecules composed for the latter of over 200 Ig-like domains.

The Ig fold

The classical Ig-domains: IgV, IgI, IgC1 and IgC2

The Immunoglobulin fold is both one and many in topology and structure, leading to multiple domain variants. It is usually described as a beta sheet sandwich barrel. It can contain 9 strands (A, B, C, C', C'', D, E, F, G) with the A strand sometimes splitting into A/A' strands and can have even more strands in Ig-extended cases as described later. As we shall see, there are multiple variants in topology and structure, but four main variant types, named V-set (IgV), C1-set (IgC1), C2-set (IgC2), and I-set (IgI) (see Figure 2) (Lesk and Chothia 1982; Harpaz and Chothia 1994; Alan F. Williams and Barclay 1988) have been studied extensively in the human immune system in particular. Among them, and since its original structural determination (Poljak et al. 1973b; Saul, Amzel, and Poljak 1978; Poljak et al. 1974; Poljak 1975; P. M. Alzari, Lascombe, and Poljak 1988; Pedro M. Alzari 1998), the Immunoglobulin variable domain has been the most studied, as it is found at the N terminus of antibodies in their Heavy and Light chains (VH and VL domains) and on T-cell Receptors (TCRs), where it is responsible for binding antigens and neoantigens respectively, hence its immense importance in immunity. In addition, immune cells are harnessed with a multitude of cell surface receptors composed of Ig domains chained in tandem starting with IgV domains at their N termini followed by IgC2 or IgC1 domains in most cases (Díaz-Ramos, Engel, and Bastos 2011; Bausch-Fluck et al. 2018; Verschueren et al. 2020; Shilts et al. 2022; Naeim et al. 2018). Beyond the four classical variants, sub-variants have also been further classified (V1, V2, I1, I2, I3) to account for small structural variations. However, the Ig-fold can be seen in many more topological, structural and functional variants.

I-set domains are commonly found in various cell adhesion molecules, such as vascular cell adhesion molecule (VCAM), intercellular adhesion molecule (ICAM), neural cell adhesion molecule (NCAM), mucosal addressin cell adhesion molecule (MADCAM), and junction adhesion molecules (JAM) (Harpaz and Chothia 1994). Additionally, the I-set domains are present in diverse protein families, including tyrosine kinase receptors, the hemolymph protein hemolin, muscle proteins like titin, telokin, and twitchin, the neuronal adhesion molecule axonin-1, the signaling molecule semaphorin 4D involved in axonal guidance, immune function, and angiogenesis. V-set and I-set Ig domains are particularly important in forming ectodomains of cell surface receptors, with secondary and tertiary structures that closely resemble each other, and phylogenetic analyses frequently group them together. The striking similarity between V-set and I-set Ig

domains shows their shared evolutionary heritage and suggests conserved function in cell surface receptors.

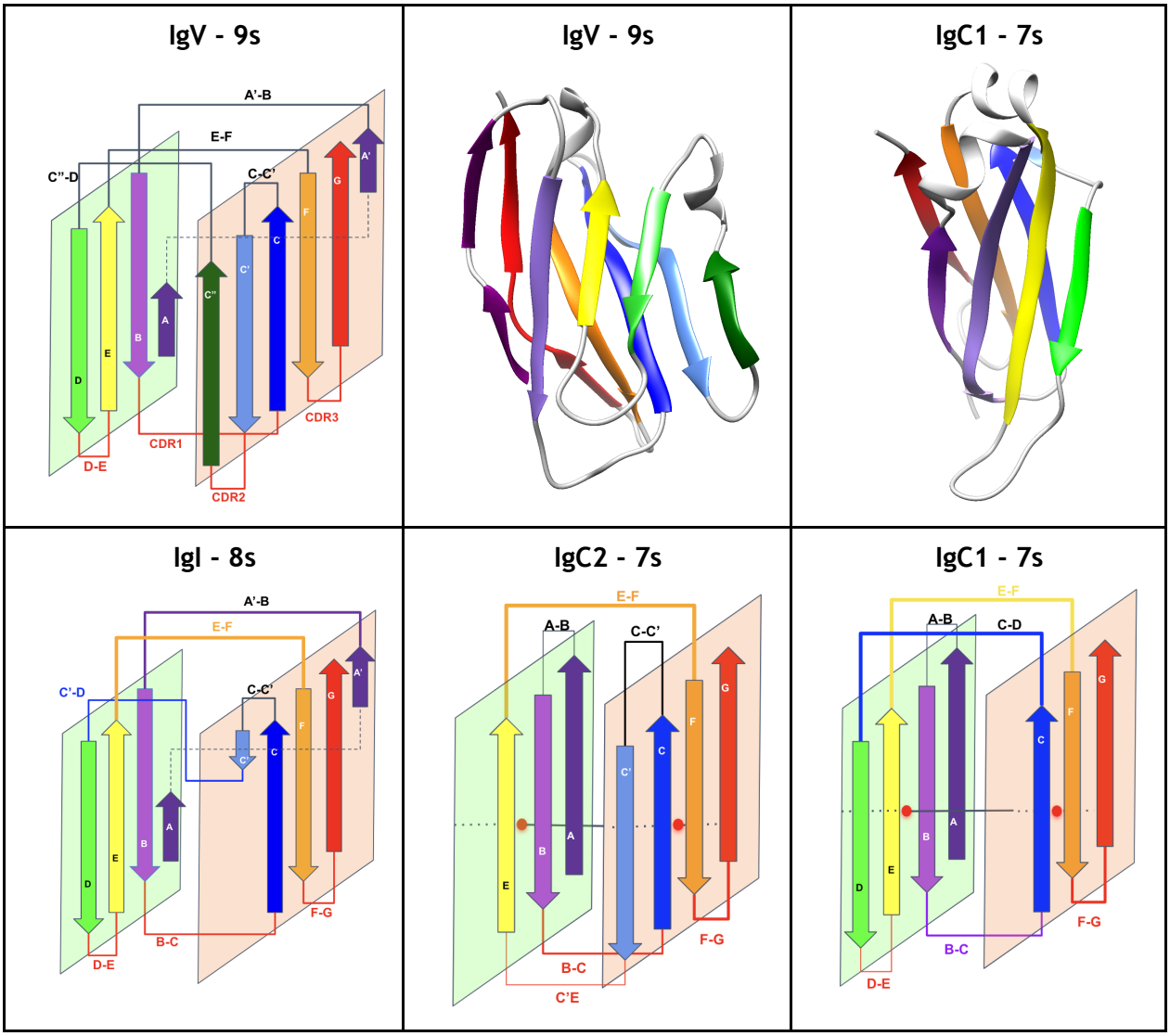


Figure 2 - Canonical Ig domain variants: Schematic and Ribbon representations. The IgV domain contains 9strands (AA')BCC'C"DEFG (9s) according to the nomenclature as displayed, The IgI domain variant contains 8 strands (AA')BCC'DEFG , IgC1 and IgC2 contain both 7 strand, ABCDEFG and ABCC'EFG resp.. The A strand splits as A and A' between the two sheets of the beta sandwich in IgI and IgV domains, as displayed. Some IgV domains can exhibit A' strand only. The IgC2 can also present an A strand split. (see text for details)

Sequence patterns of classical Ig-domains

The four classical topological variants of the Ig-fold (IgV, IgC1, IgC2, and IgI) are highly conserved in sequence and structure. They have been extensively studied and classified and are the object of a curated database (Marchler-Bauer et al. 2003; Lu et al. 2020) (Figure S1). **Figure 3** presents a graphical view of their strand-aligned sequence patterns, where one can appreciate the level of conservation of CCW(L) key residues in the core BFCE strands (Youkharibache 2019; Loerger, Du, and Linthicum 1999) despite topological variations. **Figure 3** also presents FN3 and Cadherins sequence logos (aligned to the four classical Ig topologies) representing two major Ig-fold variants with different sequence patterns, except

for a conserved Tyrosine in the F strand (called the Y corner, see later discussion of this figure) and other less conserved residues present across the BCEFG strands. Structural conservation despite a lack of sequence conservation is a hallmark of the Ig-fold variants in topology and structure, which can point to both divergent and convergent evolution (L. Shapiro et al. 1995). Despite the strand topologies observed for the four canonical Ig-domains that have been used extensively in annotating Ig-domains (Marchler-Bauer et al. 2003; Lu et al. 2020) (see Figure S1), some observed Ig-domains evade the classification using the V, I, C1, C2 types. Indeed, many more Ig-fold variants have been identified, such as the IgE set (early Ig domains) and domains that share the fold but present significant diversity in sequence, topology and structure.

Known topostructural variants of the Ig-fold beyond the classical Ig-domains

The first immunoglobulin variable domain homodimer X-ray structure, the Bence-Jones protein (Epp et al. 1974, 1975), was part of the very first set of three structures deposited to the PDB [PDBid 1REI]. To date, the protein data bank (PDB) provides 16,661 files containing Ig-like structures (as determined by IgStrand) representing 7.5% or the largest body of experimentally determined structures available on any domain (https://www.rcsb.org/) with over 25,000 Ig|Ig domain interfaces represented. The largest portion of experimental structures contain antibodies exhibiting the canonical VH|VL and CH1|CL interfaces. But, overall, this body of data covers multiple types of different Ig domain quaternary interfaces. Over the years, the growing body of experimental structures in the PDB has shown an ever-growing set of Ig-domain variants from a structural and topological standpoint. We can refer to the diverse set of Ig-domain types as either *topostructural* variants or *isotypes*.

Three taxonomy databases classify Ig-domain structural variants, beyond the common V/I/C1/C2-set nomenclature: **1) SCOP** (Murzin et al. 1995; Chandonia et al. 2022) classifies the Immunoglobulin fold under its hierarchical nomenclature as immunoglobulin-like beta sandwich b.1 containing 33 superfamilies, where b.1.1, for example, contains the main 4 variants V-set, C1-set, C2-set, I-set. Beyond these domains others that vary in topology are classified under b.2 (Common fold of diphtheria toxin/transcription factors/cytochrome f) for 10 superfamilies, b.3 (Prealbumin) with 8 superfamilies, b.7 (C2-domain like) with an additional 5 superfamilies, for **a total of 56 Ig and Ig-like topostructural variants currently delineated**. **2) CATH** (Orengo et al. 1997; Sillitoe et al. 2021) classifies Immunoglobulin beta sandwiches as **2.60.40** in their CATH hierarchical nomenclature as containing 330 superfamilies, with 45422 domains in the PDB database. **3) ECOD** (H. Cheng et al. 2014, 2015) classifies Immunoglobulin beta sandwiches in their XHTF hierarchical nomenclature, which emphasizes distant evolutionary relationships, groups superfamilies in 42 "H" homologous categories under the "X" level immunoglobulin-like and an extra immunoglobulin related category containing 8 superfamilies such as Superoxide dismutase or Prealbumin that show some "T" topology variations.

An understanding of the **Ig-fold plasticity** leading to numerous tertiary and quaternary architectural complexes is required to decipher the myriad of molecular mechanisms underlying their biological functions, to design and develop therapies to correct dysfunctions, or to engineer new functions. In the following sections, we consider this body of experimental data on the Ig-fold to define a common, universal, residue numbering system. It will enable the parallel analysis of Ig-domain variants in terms of their diversity as well as the analysis of higher levels of complexity found in quaternary assemblies, from for example TCR/CD3 or BCR complexes and/or their co-receptors such as CD4, CD8 or CD79 at the surface of cells, but also from Ig-like domains such as Lamins and Arrestins.

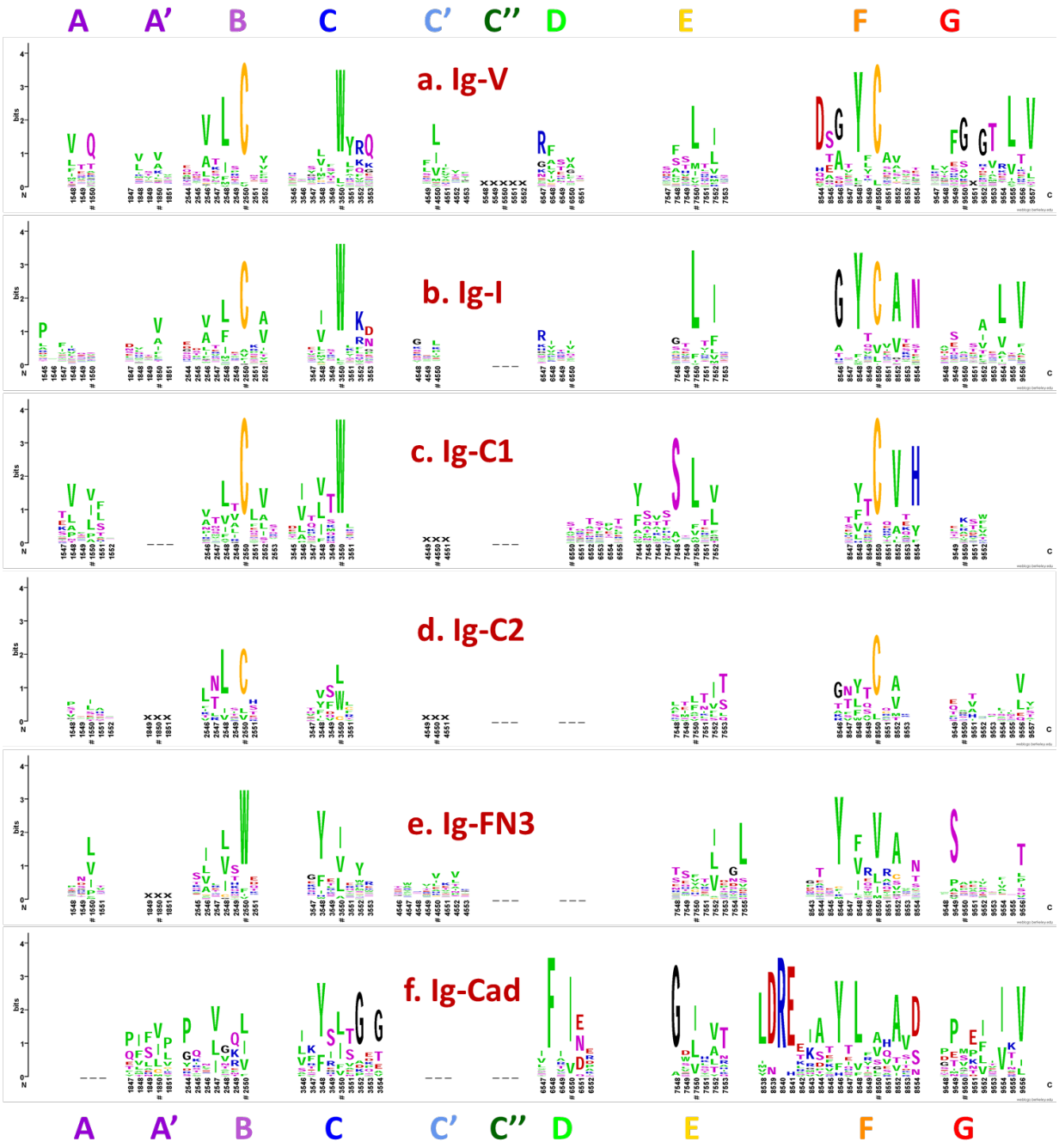


Figure 3: Sequence patterns of Ig-domain variants in WebLogo format. a) The **IgV-set** with most commonly an (AA')BCC'C"DEFG topology (9 strands)**.** There is however some diversity of IgV domains. Some IgV domains do not present a spilt AA' strand, just an A' strand as in CD4 or only an A strand as in ICOS. Some are missing the C" altogether as in PD-L2 or VNAR domains in sharks. **b)** The **IgI-set** presents a **(**AA')BCC'DEFG topology (8 strands) highly similar to IgV domains**. c)** The **IgC1-set** exhibits an ABCDEFG topology (7 strands), with usually two hallmarks: a straight A strand, even if some strand breaks can be observed, and a D strand. Some IgC1 domains can also exhibit a small non-conserved C' strand (denoted by XXX) as in IgI or IgC2 domains**. d)** The IgC2-set exhibits an ABCC'EFG topology (7 strands). IgC2 domain can exhibit an AA' strand split with a non-conserved A' strand (denoted by XXX) as in LILRs. It also has a C' strand with no sequence conservation (denoted by XXX). IgC2 differs from IgC1 domains in not presenting a D strand**.** The first four canonical Ig-domains IgV, IgI, IgC1 and IgC2 present high similarity in sequence, especially the Cysteines in strands B and F forming a Cys-Cys bridge and a Tryptophan in strand C flanking

the Cys-bridge. Each possess shared specific sequence patterns, for example the Tyrosine corner in strand F (see later in text) or some specific to each type such as R and N residues in IgI strand D and F respectively, residues Q, RF and D in IgV's strand A, D and F respectively, etc. One should note that the G-strand sequence pattern in the WebLogo results from an over representation of antibody domains in the dataset. **e)** The **Ig-FN3 set** presents an ABCC'EFG topology (7 strands). Like any other Ig-domain variant some FN3 domains can exhibit an AA' strand split**. f)** The **Ig-Cadherin set** exhibits an A'BCDEFG topology (7 strands), although while sharing Ig-fold it exhibits a very different sequence pattern and may result from convergent evolution (L. Shapiro et al. 1995). The A' strand in Cadherins corresponds to the A strand designation in the literature (J. Vendome et al. 2011). Classical cadherins can exhibit an A*/A strand split where its most N-terminal segment, called the "A* strand", provides an adhesive mechanism between cells by swapping between N-terminal (EC1) domains (J. Brash et al. 2012). Both FN3 and Cadherins, however, show key Tyrosine residues conserved in strands C and F, where the Tyr in F strand might correspond to the Tyr corner of the F strand in the previous four canonical domains. (See supplements for multiple sequence alignments used in producing sequence logos).

RESULTS and DISCUSSION

Ig, Ig-like and Ig-extended domains

All structurally-conserved folds and superfamilies exhibit sequence variations. Among them, some superfolds that are highly successful in evolution, such as 7-transmembrane G protein coupled receptors (GPCRs), kinases, Src homology 3 (SH3) domains, and Oligonucleotide/oligosaccharide-binding (OB) folds, can withstand very low sequence conservation to fold with an invariable framework topology, where structural insertions and deletions are localized in loops. **The Ig-fold is unique in that it can also vary its topology**. The topological variation of the Ig-fold is in itself a **paradox** that can be resolved by admitting that insertions can themselves form strands that extend a beta sandwich fold core, and that some strands around that core can swap between the two sheets of the beta sandwich. Variations in sequence, strand topology and loops offer an immense structurally and chemically diverse repertoire to the Ig-fold that enables it to engage in a spectrum of molecular interactions spanning multiple biological functions (both mechanical and biochemical).

The difficulty is in capturing the immense plasticity of the Ig-fold and to qualify, let alone quantify, the variability in sequence, topology, and structure. We first review the elements of plasticity in the Ig-fold and then define a universal residue numbering scheme that can be used across all topostructural variants seen in nature.

Structural invariance and topological variability of the Ig-fold

The irreducible structural signature of the Ig-fold

The plasticity observed across all topostructural variants of the Ig-fold (Poljak et al. 1973a, 1974; Poljak 1975) is significant. However, they all exhibit a structural signature composed of four central beta-strands B-C and E-F using the classical Ig-domain nomenclature (Lesk and Chothia 1982; A. F. Williams and Barclay 1988; Bork, Holm, and Sander 1994). These 4-strands form the core central beta sandwich supersecondary structure (4s-SSS) as two straddling 2-beta strands motifs B-C and E-F intertwined in a pseudo-symmetric arrangement, and are supplemented by a 5th G strand to give a 5-strands supersecondary structure signature for the Ig-fold (5s-SSS) (see Figure 4). The two substructures B-C and E-F, straddle the two sheets of the sandwich through the BC loop (CDR1) and the EF loop barrel exhibiting a C2 pseudosymmetry (Youkharibache 2019); the G strand at the C terminus forms the FG loop, known as CDR3 (complementary determining region 3) in antibodies' variable

domains. Two to five lateral strands, comprising A/A' at the N terminus and C'/C"/D between C and E strands, define and differentiate the four classes of canonical Ig-domains (Figure 3 and Figure 4): IgC1, IgC2, IgI and IgV (Lesk and Chothia 1982; Bork, Holm, and Sander 1994).

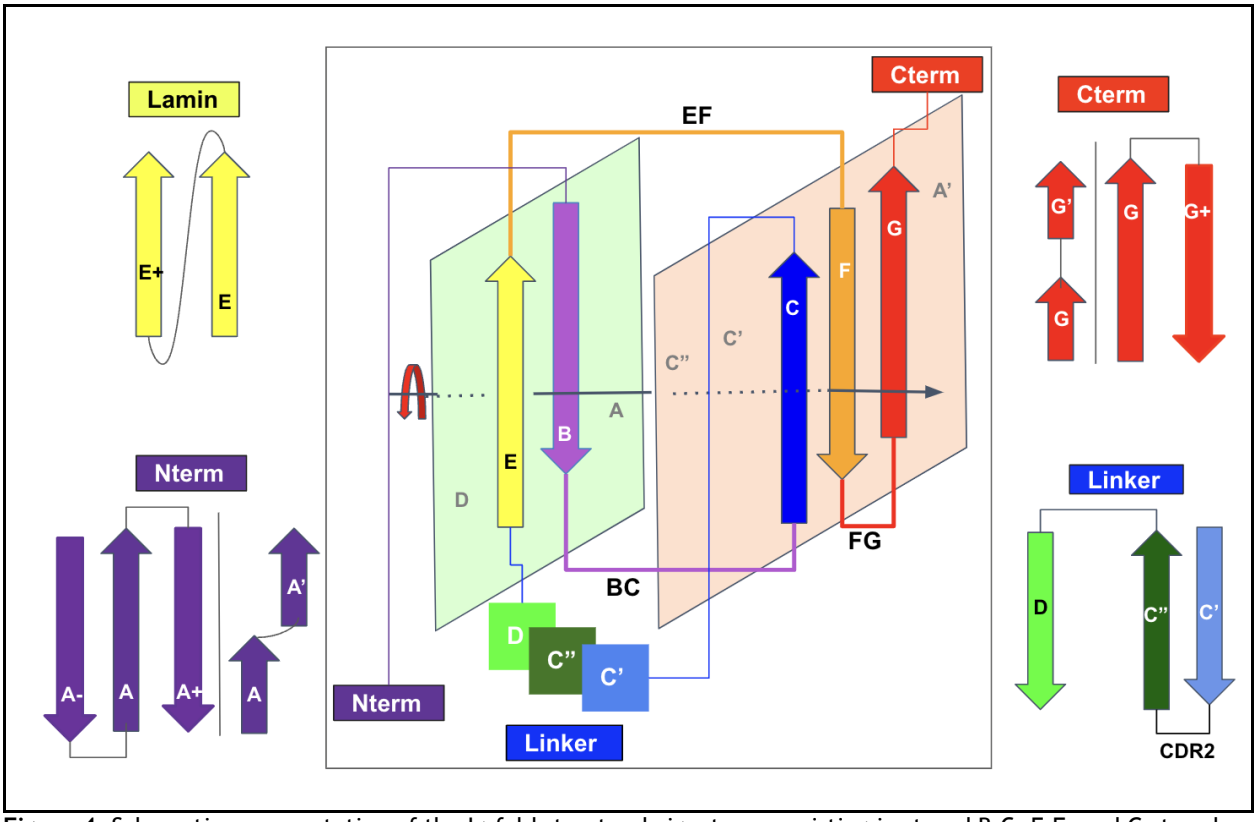


Figure 4. Schematic representation of the Ig-fold structural signature consisting in strand B-C, E-F, and G strands arranged as displaced in the central panel, and main regions of lateral variability: NTerm (strand A), Cterm (strand G) and Linker (Strands between C and D), as well as an example of a less known lateral strand addition. (See text)

Interestingly the central 4 strands (BCEF, 4s-SSS) represent a universal supersecondary structure motif common not just to the Ig-fold, but also to the jelly-roll fold as originally defined by Jane Richardson (Richardson 1981), suggesting that this central supersecondary structure may act as a stable folding nucleus for beta-sandwich folds to accommodate a myriad of topostructural variants. Jelly rolls represent, after the Ig-fold, the second largest taxonomic group in structural classifications of SCOP, CATH, and ECOD, and these two groups represent the overwhelming majority of beta sandwich structures that in turn, as a class, represent 30% of all folds with beta secondary structures (Edwards, Abeln, and Deane 2013).

Qualifying the variability of domains sharing the Ig-fold

While the Ig-fold has been recognized in numerous topostructural variants, detailed comparisons across all of them are difficult: in addition to variable sequence identity and variable lengths of the strands and loops as in any structural family, they exhibit different tilt angles between the two beta-sheets of the sandwich, and variable topologies with different number of strands in the well-known V/C1/C2/I sets. In some variants, additional strands may appear at the N-terminus to the A strand, interspaced structurally between the A and G strands, or at the C-terminus to the G strand, both on Sheet 1 (light green plane in Figure 4) and/or Sheet 2 (light orange plane in Figure 4), and in parallel or antiparallel with the adjacent strands. Variations within some strands can result in Ig variants with striking differences as some strands can form bulges, and/or can split with loop insertions, and split strands, or overall strands may swap

between Sheet 1 and Sheet 2. This is a basis for the well-known A strand split with its first half on Sheet 1 (ABED) and its second half (called A') on Sheet 2 (A'GFCC') in IgV and IgI domains. Some IgV domains can also swap the C" strand at the tertiary level between Sheet 1 and Sheet 2, as observed in crystal structures of TCR IgV or CTLA-4 IgV domains. That C" strand may also become disordered as in PD1, or be absent altogether as in PD-L2 or VNAR domains. In addition, beta bulges are common in beta strands A and G. Therefore, it is not surprising that when continuously numbering an Ig-domain, a specific structurally-conserved position, common to all members of the fold, gets a different residue number in different domains, even with simple variations. Considering the immense plasticity in strand topology, defining a universal residue numbering system becomes an important challenge.

Towards positional structural bioinformatics

Positional bioinformatics

Since the very beginning, the field of bioinformatics has been focused on sequence alignment to compare the *relative* sequence identity between two or more protein sequences at given *positions*. However, even for highly similar sequences, the ***sequence numbers*** of corresponding (mapped) positions in alignments vary, due to insertions/deletions accumulated during evolution. A reference numbering system common to all members of a multiple sequence alignment (MSA), can enable *positional* comparisons in sequence, structure, and function. Since sequence datasets and especially structural datasets are becoming larger, the ability to compare similar positions of residues in sequence, topology, tertiary and quaternary structure becomes a necessity to capture position-based interaction patterns that underlie biological functions.

Examples of reference residue numbering systems in proteins are not legions for proteins: two main ones have been defined for domains with a fixed topology: immunoglobulin variable domains (Kabat numbering) and for GPCRs (Ballesteros-Weinstein numbering). In the realm of RNA, studies on the ribosome benefit from a reference numbering, using E.Coli as a reference species to enable comparisons and delineating evolution across species (Petrov et al. 2014, 2015; Kushner, Petrov, and Dao Duc 2023). But the most impressive use of a reference numbering is in the field of genomics to annotate DNA base pair positions. Genome reference numbering highlights a transition from comparative to positional bioinformatics. In fact, one may wonder where genomics would be today if there was no common reference numbering for the community to share annotations on the human genome, and how difficult it would be to discuss mutations, transcription factor binding sites, ORFs, introns and exons positions in genes, or any other positional feature. One can argue that a standardized nomenclature and a reference numbering system allows the scientific and medical community to annotate, analyze, compare, and coalesce information in a consistent manner on a biomolecular family of importance. Establishing a reference residue numbering scheme for proteins that share the Ig-fold is needed despite its inherent difficulty in representing the many topological and structural variations, and this is exactly what we seek: a universal numbering scheme that enables comparisons in structurally-conserved residue positions across all topological variants of Ig-domains.

Requirements for a numbering scheme describing the Ig-fold

Ig-domain families offer an elaborate diversity in structure, topology, sequence and function. Focusing on Ig-domains and seeking a universal residue numbering scheme adaptable to all Ig-domain topological variants, we aim at enabling parallel analyses of **tertiary domains, chains**, and their quaternary **assemblies**. Ig domains form a myriad of Ig-Ig complexes with diverse geometries that can, in some cases, be considered as **"quaternary topologies"** and even **"quaternary folds"**, as for example in

antibody Fab, IgV and IgC1 dimers of heavy and light chains. While we expect to learn evolutionary relationships from tertiary domains and chains made of Ig-domains in tandem, we should also expect to learn about co-evolutionary relationships between receptors and ligands made of and interacting with Ig-domains, given their remarkable property of self-assembly as in antibodies heavy and light chains, or in cell surface receptors-ligand associations in either cis or trans orientations. Ig-domains assemble in tandem in chains and in quaternary complexes, which brings three levels of complexity (see below). From a domain level reference numbering scheme one can expect the ability to then compare Ig domain quaternary complexes.

A universal numbering scheme common to all Ig, Ig-like and Ig-extended domains is intended to enable direct comparisons between domains and their associations across very large datasets and opens the door to proteome scale analysis. While a numbering scheme may be used for machine learning, we seek one that is human readable, following examples such as Kabat numbering on IgVH and IgVL domains of antibodies or Ballesteros-Weinstein numbering in GPCRs (see later).

Topostructural variants reference dataset

The structural diversity of Ig-domains is captured by classifications such as ECOD, CATH and SCOP as was described earlier. These classifications rely on overall pattern matching between domains, using structural alignments. They offer distinct variant classes in structure and topology, albeit in slightly different ways. The drawbacks of such detailed classifications are that i) they separate topostructural variants in silos and ii) they do not offer a way to compare and relate topologies. This makes the study of tertiary plasticity across them difficult. To establish a universal numbering scheme for the Ig-fold, we first analyzed a dataset of Ig-domain superfamilies, classified as b.1 in SCOP, and containing domains from plant, fungi, archaea, bacteria, and viruses to capture and evaluate structural diversity across life. We then selected representatives of cytoplasmic proteins, nuclear proteins, and cell surface receptors in the immune, nervous, vascular and muscular systems from the animal kingdom favoring a human origin. Table 2 lists a selected reference set of Ig-domains that will be used as “templates". While biased towards the human species, the selected Ig-domain representatives are structurally fit for the identification and labeling of Ig-domains in all domains of life. In fact, while the common numbering scheme defined on this b.1 dataset (Table 2) was initially aimed at covering Ig- and Ig-like domains, we found that it could be easily adapted to cover Ig-extended domains, i.e., topostructural variants classified in various fold categories such as SCOP b.2 (like p53), b.3 (like prealbumin), b.7 (like C2-domain) by allowing topological insertions of strands, instead of introducing more templates covering these variants, although the need for high precision may require more templates in the future, and other algorithmic improvements.

The original Kabat numbering and its variants for IgV domains

The idea of a domain-level positional analysis, originally in sequence, is due to Elvin A. Kabat in late 1960s (Kabat and Others 1968; T. T. Wu and Kabat 1970; Elvin A. Kabat 1992) (E. A. Kabat, Wu, and Bilofsky 1976) as he started compiling and tabulating a large number of antibody sequences to form the Kabat database (Johnson and Wu 2000), *observing that **positions in antibody sequences** could be associated with invariant vs. highly variable residue types in certain segments of their sequence*, named Framework (FR) vs. hypervariable regions (CDR: Complementary Determining Regions) (E. A. Kabat, Wu, and Bilofsky 1976). Kabat’s delineation of functional subdomains (Framework) and hypervariable regions (CDR loops) responsible for binding antigens, was later confirmed when the first Bence-Jones protein (an IgV_L dimer) (T. T. Wu and Kabat 1970; Epp et al. 1975) and antibody structures were determined (Saul, Amzel, and Poljak 1978; Chothia et al. 1986; P. M. Alzari, Lascombe, and Poljak 1988). This was one of

the first pieces of evidence of the close relation between sequence, structure and function of antibody proteins. The "Kabat numbering" scheme was assigning positional numbers in sequence common to all antibody IgV domain light chains (IgV_L) and heavy chains (IgV_H) (independently), i.e., each residue in the antibody domain is assigned a unique number based on its position in sequence, allowing an easy reference and comparison on specific locations within antibody IgV_L and IgV_H domains.

As more sequences and structures accumulated some shortcomings of the initial Kabat numbering appeared and a number of improvements to the numbering scheme were proposed: the Chothia numbering (Chothia and Lesk 1987), and the improved Chothia, also named Martin numbering (Abhinandan and Martin 2008) to improve Kabat's numbers based on structure. Another limitation of the Kabat numbering is the use of different position numbers in VH and VL domains. Improvements and extensions were later proposed to describe antibody and TCR Ig domains with the IMGT numbering scheme (Lefranc et al. 2003, 2005; Lefranc 2011). Variations on the latter have also been proposed as yet another numbering scheme (AHo) (Honegger and Plückthun 2001). While Kabat numbering has been used for a long time, only recently were tools developed to map Kabat and other numbers to antibody variable domains (Dunbar and Deane 2016).

1	Filename	Structural Classification	Protein	Template	SCOPe	UniProt	Domain # (PDB)	Domain # (Protein)	PDB ID	Chain ID	RefSeq	Species
2	B2Microglobulin_7phrL_human_C1.pdb	C1	**B2-Microglobulin**	Ig	b.1.1.2	C1	1	1	7phr	L	P61769	Human
3	BTLA_2aw2A_human_Iset.pdb	I-set	**BTLA**	Ig	b.1.1.4	V	1	1	2aw2	A	Q7Z6A9	Human
4	CD19_6al5A_human_C2orV-n1.pdb	CD19	**CD19**	Ig	Unk.	C2	1	1	6al5	A	P15391	Human
5	CD2_1hnfA_human_C2-n2.pdb	C2	**CD2**	Ig	b.1.1.3	C2	2	2	1hnf	A	P06729	Human
6	CD2_1hnfA_human_V-n1.pdb	V	**CD2**	Ig	b.1.1.1	V	1	1	1hnf	A	P06729	Human
7	CD28_1yjdC_human_V.pdb	V	**CD28**	Ig	b.1.1.1	V	1	1	1yjd	C	P10747	Human
8	CD3d_6jxrd_human_C1.pdb	C1	**CD3d**	Ig	b.1.1.4	Ig-like	1	1	6jxr	d	P04234	Human
9	CD3e_6jxrf_human_C1.pdb	C1	**CD3e**	Ig	b.1.1.4	Ig-like	1	1	6jxr	f	P07766	Human
10	CD3g_6jxrg_human_C2.pdb	C2	**CD3g**	Ig	b.1.1.4	Ig-like	1	1	6jxr	g	P09693	Human
11	CD8a_1cd8A_human_V.pdb	V	**CD8a**	Ig	b.1.1.1	V	1	1	2atp	A	P01731	Mouse
12	Contactin1_3s97C_human_Iset-n2.pdb	I-set	**Contactin1**	Ig	b.1.1.0	C2	2	3	3s97	C	Q12860	Human
13	FAB-HEAVY_5esv_C1-n2.pdb	C1	**FAB-CH1**	Ig	b.1.1.2	C1	2	2	5esv	C	5ESV_C	Human
14	FAB-HEAVY_5esv_V-n1.pdb	V	**FAB-VH**	Ig	b.1.1.1	V	1	1	5esv	C	5ESV_C	Human
15	FAB-LIGHT_5esv_C1-n2.pdb	C1	**FAB-CL1**	Ig	b.1.1.2	C1	2	2	5esv	D	5ESV_D	Human
16	FAB-LIGHT_5esv_V-n1.pdb	V	**FAB-VL**	Ig	b.1.1.1	V	1	1	5esv	D	5ESV_D	Human
17	ICOS_6x4gA_human_V.pdb	V	**CD278 (ICOS)**	Ig	b.1.1.1	V	1	1	6x4g	A	Q9Y6W8	Human
18	JAM1_1nbqA_human_Iset-n2.pdb	I-set	**JAM1**	Ig	b.1.1.4	V	2	2	1nbq	A	Q9Y624	Human
19	LAG3_7tzgD_human_C2-n2.pdb	C1	**LAG3**	Ig	Unk.	C2	2	2	7tzg	D	P18627	Human
20	LAG3_7tzgD_human_V-n1.pdb	V	**LAG3**	Ig	Unk.	V	1	1	7tzg	D	P18627	Human
21	MHCIa_7phrH_human_C1.pdb	C1	**MHCI alpha**	Ig	b.1.1.2	C1	1	1	7phr	H	P04439	Human
22	Palladin_2dm3A_human_Iset-n1.pdb	I-set	**Palladin**	Ig	b.1.1.4	C2	1	1	2dm3	A	Q8WX93	Human
23	PD1_4zqkB_human_V.pdb	V	**CD279 (PD1)**	Ig	b.1.1.1	V	1	1	4zqk	B	Q15116	Human
24	PDL1_4z18B_human_V-n1.pdb	V	**CD274 (PDL1)**	Ig	b.1.1.1	V	1	1	4z18	B	Q9NZQ7	Human
25	Siglec3_5j0bB_human_C1-n2.pdb	C1	**Siglec-3 CD33**	Ig	Unk.	C2	2	2	5j0b	B	P20138	Human
26	Titin_4uowM_human_Iset-n152.pdb	I-set	**Titin**	Ig	Unk.	Ig-like	1	152 (M10)	4uow	M	Q5VST9	Human
27	VNAR_1t6vN_shark_V.pdb	V	**VNAR**	Ig	b.1.1.1	Ig-like	1	1	1t6v	N	1T6V_N	Shark
28	C3_2qkiD_human_n1.pdb	FN3-like	**C3 Complement**	Ig-like	b.1.29.2	Unknown	1	1	2qki	D	P01024	Human
29	Contactin1_2ee2A_human_FN3-n9.pdb	FN3	**Contactin1**	Ig-like	b.1.2.0	FN3	1	9	2ee2	A	Q12860	Human
30	ECadherin_4zt1A_human_n2.pdb	Cadherin	**E-Cadherin**	Ig-like	b.1.6.1	Cadherin	1	2	4zt1	A	P12830	Human
31	GHR_1axiB_human_FN3-n1.pdb	FN3	**GHR**	Ig-like	b.1.2.1	FN3	1	1	1axi	B	P10912	Human
32	IL6Rb_1bquB_human_FN3-n2.pdb	FN3	**CD130 (IL6R-beta)**	Ig-like	b.1.2.1	FN3	1	2	1bqu	B	P40189	Human
33	IL6Rb_1bquB_human_FN3-n3.pdb	FN3	**CD130 (IL6R-beta)**	Ig-like	b.1.2.1	FN3	2	3	1bqu	B	P40189	Human
34	InsulinR_8guyE_human_FN3-n1.pdb	FN3	**Insulin Receptor**	Ig-like	b.1.2.1	FN3	1	1	8guy	E	P06213	Human
35	InsulinR_8guyE_human_FN3-n2.pdb	FN3	**Insulin Receptor**	Ig-like	b.1.2.1	FN3	2	2	8guy	E	P06213	Human
36	NaCaExchanger_2fwuA_dog_n2.pdb	FN3-like	**NaCa Exchanger**	Ig-like	b.1.27.1	Unknown	1	2	2fwu	A	P23685	Dog
37	NaKATPaseTransporterBeta_2zxeB_spurdogshark.pdb	Unknown	**NaK ATPase Transporter-Beta**	Ig-like	b.1.32.1	Ig-like	1	1	2zxe	B	P05027	Spurdog shark
38	ORF7a_1xakA_virus.pdb	Unknown	**ORF7a**	Ig-like	b.1.24.1	Unknown	1	1	1xak	A	P59635	Virus
39	RBPJ_6py8C_human_Unk-n1.pdb	Unknown	**RBPJ**	Ig-like	Unk.	Unknown	1	1	6py8	C	Q06330	Human
40	RBPJ_6py8C_human_Unk-n2.pdb	Unknown	**RBPJ**	Ig-like	Unk.	Unknown	2	2	6py8	C	Q06330	Human
41	Sidekick2_1wf5A_human_FN3-n7.pdb	FN3	**Sidekick-2**	Ig-like	b.1.2.1	FN3	1	7	1wf5	A	Q58EX2	Human
42	TP47_1o75A_bacteria.pdb	Unknown	**Tp47 Lipoprotein**	Ig-like	b.1.20.1	Unknown	1	1	1o75	A	P29723	Bacteria
43	ASF1A_2iijA_human.pdb	FN3-like	**ASF-1A Histone Chaperone**	Ig-extended	b.1.22.1	Unknown	1	1	2iij	A	Q9Y294	Human
44	BArrestin1_4jqiA_rat_n1.pdb	FN3-like	**B-Arrestin-1**	Ig-extended	b.1.18.11	Unknown	1	1	4jqi	A	P29066	Rat
45	CoAtomerGamma1_1r4xA_human.pdb	Unknown	**CoAtomer-Gamma-1**	Ig-extended	b.1.10.3	Unknown	1	1	1r4x	A	Q9Y678	Human
46	CuZnSuperoxideDismutase_1hl5C_human.pdb	FN3-like	**CuZn Superoxidase Dismutase**	Ig-extended	b.1.8.2	Unknown	1	1	1hl5	C	P00441	Human
47	Endo-14-BetaXylanase10A_1i8aA_bacteria_n4.pdb	Unknown	**Endo-1,4-Beta--Xylanase-A**	Ig-extended	b.1.9.2	Unknown	1	4	1i8a	A	Q60037	Bacteria
48	IsdA_2iteA_bacteria.pdb	Unknown	**IsdA NEAT domain**	Ig-extended	b.1.28.1	Unknown	1	1	2ite	A	Q2FZE9	Bacteria
49	LaminAC_1ifrA_human.pdb	Unknown	**Lamin-A/C**	Ig-extended	b.1.16.1	Unknown	1	1	1ifr	A	P02545	Human
50	MPT63_1lmiA_bacteria.pdb	FN3-like	**MPT63**	Ig-extended	b.1.19.1	Unknown	1	1	1lmi	A	P9WIP1	Bacteria
51	TEAD1_3kysC_human.pdb	FN3-like	**TEAD1 (TEF1)**	Ig-extended	b.1.18.26	Unknown	1	1	3kys	C	P28347	Human
52	TP34_2o6cA_bacteria.pdb	Unknown	**TP34**	Ig-extended	b.1.33.1	Unknown	1	1	2o6c	A	P19478	Bacteria
53	VISTA_6oilA_human_V.pdb	V	**VISTA (VSIR)**	Ig-extended	b.1.1.1	V	1	1	6oil	A	Q9H7M9	Human

Table 2 - Reference set of diverse topological and structural variants of Ig domains from the SCOP b.1 Superfamily. (see supplement file SF7 for the corresponding spreadsheet). It includes surface receptors in the immune, nervous, and vascular systems, cytoplasmic proteins and enzymes, transcription factors and nuclear proteins.

Definition of the IgStrand numbering

The IgStrand numbering scheme (IgStRAnD acronym for Ig Strand Residue Anchor Dependent) is inspired by Kabat's original numbering for IgVH and IgVL domains, as well as the Ballesteros-Weinstein (BW) GPCR numbering scheme (Ballesteros and Weinstein 1995). The latter uses numbers centered on secondary structure elements (SSE) and their most conserved residues, considered as sequence "**anchors**", assigning

them a number 50, then counting both positively and negatively in each SSE, where each SSE is assigned a number i=1,2,...7 for 7-helical membrane proteins. For example, 2.50 would be "helix 2, number 50" for the most conserved residue in TM2, an aspartate, in class A GPCRs. The residue that appears before this conserved aspartate in sequence would be numbered 2.49 and the one that appears after it would be numbered 2.51, and so on. The numbers can be considered hierarchical, but also simply decimal numbers (i.e. 2.50 or 250 in this case).

Similarly, the IgStrand numbering scheme defines anchors with a number 50, but *these are defined structurally* and are not based solely on sequence conservation to account for the immense variability in sequence in topostructural variants. Besides, this scheme uses 4 digits to number residues in an Ig domain (ij50). The numbers are hierarchical but also decimal to ensure sequence continuity. The first digit "i" (1000th place) defines the strand number ranging from 1 to 9 (i.e, 1jxx-9jxx) defining the most common Ig topologies such as the Immunoglobulin variable domain. The second digit "j" (100th place) is set to 5 for the main canonical strands (A, B, C, C', C", D, E, F, G). The structurally conserved anchor positions in those strands would be 1550, 2550, 3550, 4550, 5550, 6550, 7550, 8550, and 9550.

Defining positional structural anchors "ij50" in the Ig-fold beta strands

To define structurally conserved anchor positions, a reference set of Ig domains providing topological diversity were structurally aligned to the IgVH domain, which has been extensively studied. This alignment (Figure 3) confirmed known conserved residues and uncovered the conservation of specific residue positions within each strand and their network of interactions defining the interior of the Ig-folds (see Methods). The most structurally conserved position in each strand was selected as the residue anchor for its respective strand. All anchors selected are in-facing residues as they form an interconnected network of backbone and sidechain interactions, crucial to the folding and stability of the Ig-fold. Anchors chosen structurally coincide with key conserved residues in sequence for well known superfamilies such as Immunoglobulins (SCOP b.1.1.1), but are otherwise highly variable across all topostructural variants (Figure 3). Figure 5 and Figure 6 highlight highly conserved residues in variable and constant domains of antibodies, TCRs and other immune receptors. The specific residue position chosen as anchor in each of the classical 7-9 strands (A, B, C, C', C", D, E, F, G), as well as in the characteristic A strand split (see later), forming a so-called A' strand in IgV and IgI domains, is given a number "50" regardless of the Ig variant type, the lengths of the strands, the loops, angle between the two beta-sheets, or sequence variability.

Defining canonical and additional beta strand numbers (i and j)

The classical 7-9 strands for canonical Ig domains A, B, C, C', C", D, E, F, G are given the number i=1,2,3,4,5,6,7,8,9 respectively (1000's) (see Figure 5A). The second digit is fixed to j=5. To allow for strand insertions, the second digit j can be numbered negatively or positively from it, the special case of A' strand getting j=8 (see Figure 5A). The third and fourth digits correspond to the residue number in a given strand and beyond, counting both positively and negatively from the anchor residue number 50, as in BW numbering for GPCRs, so a strand of 7 residues for example could run from ij47 to ij53 around the anchor ij50. The numbers will then run further than the strand in both directions to accommodate strand length variability as well as loop numbering from and to the next strand. In this scheme, the N-terminal half of the loops inherit the continuous numbering forward from the previous strand and the C-terminal half of the loops inherit the continuous numbering backward from the following strand (see Methods). The N and C term residues, before the first and the last strand respectively, inherit the numbering going backward and forward respectively from those strands. So, all residues in the Ig domain are accounted for in this numbering scheme. This numbering does not currently explicitly consider the sheet in which a

strand lies, as in some alternative and hierarchical numbering schemes that we have considered. The sheet information would be a useful future extension to the Igstrand numbering scheme.

Anchor selection started with the disulfide bridge cysteines on the B strand (igs# 2550) and the F strand (igs# 8550), with respect to which the rest of the anchors were subsequently selected. The GFCC' anchors form an extended stretch of backbone-backbone interactions, which includes the conserved tryptophan on the C strand (igs# 3550), while C" anchor is shifted negatively by one residue with respect to interstrand backbone interactions on the GFCC'C" sheet. The A strand backbone neighbor of the B strand cysteine was selected as the A strand anchor (igs# 1550), which is often a glutamate or glutamine that interacts with the backbone of the F strand cysteine (igs# 8550) in variable domains. Due to the rotation of the two beta-sheets in the Ig domain beta-sandwich, the anchors are shifted by two residue positions negatively for the D strand (igs# 6550) and positively for the E strand (igs# 7550). Particularly, this igs# 7550 is highly conserved as leucine and stabilizes the C strand tryptophan (igs# 3550) and may play a critical role in folding. In a special case, since the A' is an extension of the A strand that is located at the edges of the Ig domain beta-sandwich, the most C-terminal inward-facing residue was selected as the A' anchor position (igs# 1850), which predominantly provides hydrophobic contacts that stabilize the EF loop and the core of the Ig domain.

IgStrand Structural Anchors

The structural anchors provide a network of backbone interactions along the two sheets of the beta sandwich as well as non-bonded core stabilizing hydrophobic interactions, and in the case of canonical domains the formation of covalent cysteine bridges.

The A strand anchor position (igs# 1550) (see Figure 5A: i=1,j=5) is typically occupied by residues E/Q in IgVH, IgVL, and many other IgV domains, while in others it may be a hydrophobic residue. E/Q at position igs# 1550 provides a polar sidechain interaction with the backbone of F strand Cysteine igs# 8550, stabilizing the classical BF disulfide bond. **The A strand** represents a structurally variable region, even within canonical Ig domains. **The A' strand**, as it is named conventionally, may be pictured as an A strand that splits, with the second half (A') swapping from Sheet 1 (ABED) to Sheet 2 (A'GFCC'C") in IgV or IgI domains. This gives that strand an additional A' anchor (igs# 1850) on the edge of the strand before the A'B loop, indirectly connected to the hydrophobic core of the Ig domain through a network of interactions. The split A/A' strand provides a strong connection between the two sheets of the sandwich at the N-terminus. The A' strand anchor residue (igs# 1850) is in-facing, and the prior in-facing residue in the strand (igs# 1848) forms a parallel backbone H-bond with the residue (igs# 9554) on the G strand, itself two residues downstream of the G strand residue (igs# 9552) that is adjacent to and forms a backbone pair with the residue (igs# 8548) on the F strand, which is two residues upstream of the F strand anchor (8550) (Figure 5G and Figure 6). We opted to assign the A' strand an anchor position because many Ig domains lack the A strand but contain only an A' strand, such as CD4, CD19, and many others. This A' anchor position is structurally important for closing the ends of the two beta-sheets by providing a hydrophobic contact and rendering the core as solvent-inaccessible.

As we have seen, the core structural signature of an Ig-fold is composed of the four central beta-strands **B,C,E,F**, forming two loops straddling the two sheets of the sandwich barrel: BC loop and EF loop. The corresponding anchors igs # 2550, 3550, 8550 and to a lesser extent 7550 in strands B, C, F, and E respectively (see Figure 5A) are highly conserved positions occupied by the residues CCW(L) (Youkharibache 2019), respectively. The highly conserved Cysteine anchors igs# 2550 and igs# 8550 form a disulfide bond in most canonical Ig domains (Figure 5C and Figure 6), apart from some exceptions such as CD2. Another highly conserved residue is Trp on the C strand (igs# 3550) flanking the Cysteine bridge,

which plays an important role in stabilizing the core of canonical Ig-folds. This Trp can also form hydrogen bond interactions through its indole ring amine group with residues on the E strand either directly with polar sidechains or indirectly through a water molecule with backbone atoms (PDB ID: 5DK3).

Strand Name	Strand# (i) 1000's	(j) 100's	Anchor & Key 10's	Residue type (IgVL/VHs)
A-	1	4	1450	
A	1	5	1550	E/Q
A+	1	6	1650	
A'	1	8	1850	
B	2	5	2550	C
C	3	5	3550 3552	W R
C'	4	5	4550	M/V/L
C"	5	5	5550	
D	6	5	6550	
E	7	5	7550	L
F	8	5	8544 8548 8550 8552	D Y (*) C R
G	9	5	9549 **9550** 9553 9555	F,W G T V
G+	9	6	9650	

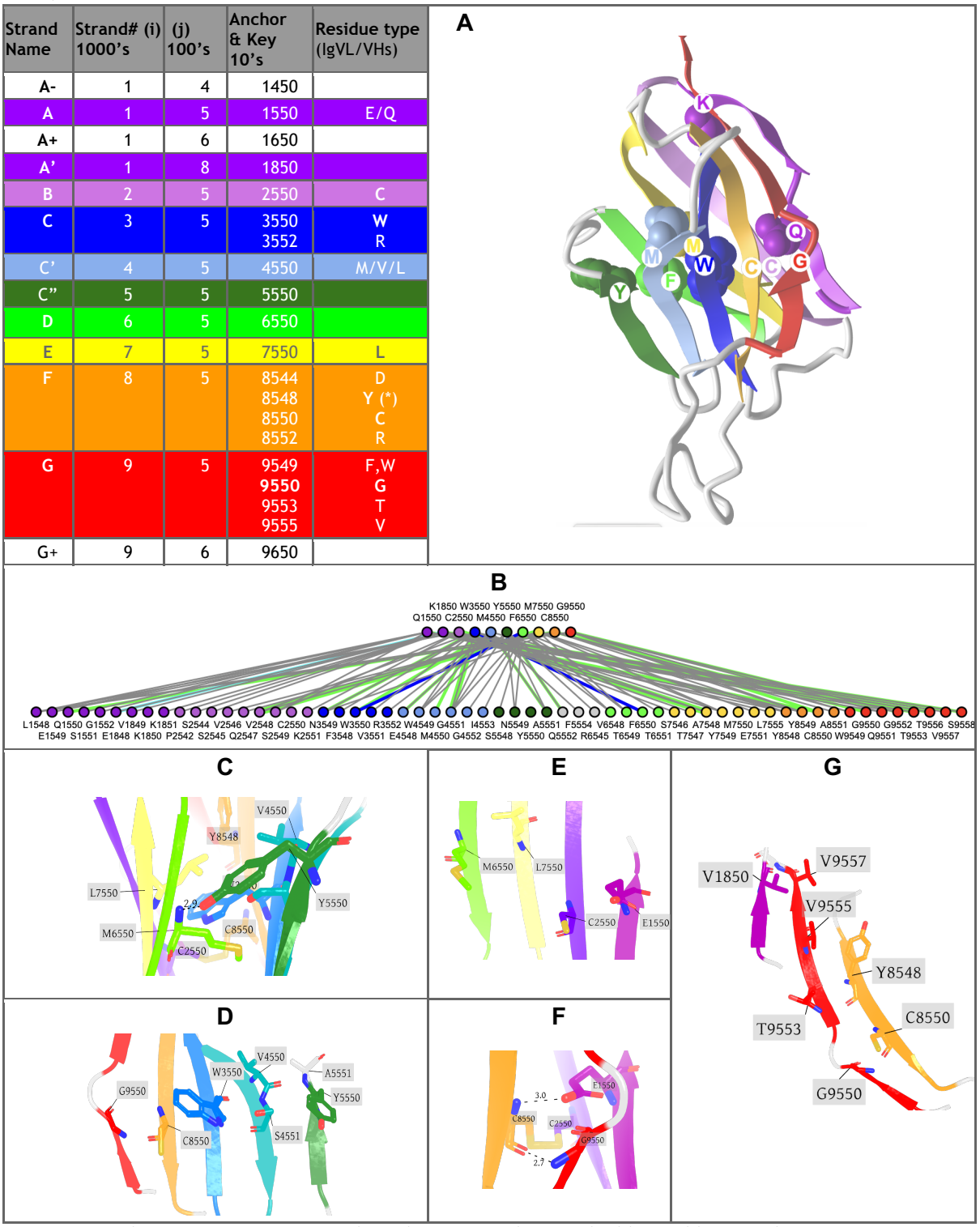


Figure 5 - Anchor positions. A) IgStrand numbers for anchors and additional key residues in important IgV domains Definitions for Classical Strands (1000's), Non-classical Strands (100's) with Anchor numbers (50s). Residue

types are shown for well-known IgV examples as in antibodies where sheet 1 is composed of strands ABED and sheet 2 (A')GFCC'(C"). The A strand anchor residue is given the number 1550, where it is the first strand (A) on Sheet 1. A' strand anchor is given the number 1850 it is the second half of the A strand swaps to Sheet 2 adjacent to the G strand in parallel. The B strand anchor is 2550. C strand anchor is 3550 and is on Sheet 2. C' strand anchor is 4550 and is on Sheet 2. C" strand anchor is 5550 on Sheet 2, but it can swap to Sheet 1. D strand anchor is 6550. The E strand anchor is 7550. The F strand anchor is 8550. The G strand anchor is 9550. Additional, non-classical strands can also appear in Ig-extended domains. This is taken into account by the second digit, for example the anchor of the strand appearing before the A strand (A- strand) will have the IgStrand number 1450, while the anchor of a strand A+ after the A strand will be 1650, and for a G+ strand appearing after a G strand similarly 9650. The anchors form a network. **B)** shows the extensive residue interaction network between the anchors core and residues in all strands of the fold. **C)** shows sidechain and backbone representations of all anchors with respect to the overall fold. **D)** shows the positions and backbone connections of the anchors in the GFCC'C" Sheet 2. **E)** shows D 6550 and E 7550 strand anchors vs. A 1550 and B 2550 anchors on the ABED Sheet 1. **F)** shows the interaction of A strand anchor 1550 with the backbone of 8550, forming the signature bulge in the G strand of IgVs. **G)** shows the position of 1850 in the A' strand with respect to the F and G strand residues. (PDBid used 5ESV H heavy chain variable domain)

In domains with a **C' strand** (igV, IgI and IgC2) the anchor position (igs# 4550) is adjacent to the anchor on the C strand (igs# 3550) and is typically occupied by a hydrophobic residue such as Leucine. In most IgVs, a beta bulge on C' strand provides two consecutive in-facing residues due to trans backbone configuration of igs# 4550. Both in-facing residues form H-bond backbone interactions with the igs# 3550 backbone atoms on the C strand. Igs# 3550 also accepts a hydrogen bond from the backbone of igs# 5551 on the C" strand. The first of these two residues is chosen as the anchor since some Ig domains such as VNARs and IgIs have a very short C' strand that ends at igs# 4550 precisely before forming the C'D loop. IgC2 variants, which classically have a relatively long C' strand and no C", lack the two in-facing consecutive residues signature seen in most IgVs, and instead only have one in-facing residue that donates and accepts backbone hydrogen bonds with igs# 3550. Igs# 4550 provides additional stability to the hydrophobic core of the Ig-fold by forming vdW interactions with igs# 3550 and igs# 7550.
Igs# 5550, present predominantly in IgVs, lies one residue upstream from igs# 5551 that forms backbone hydrogen bond interaction with igs# 4550. The reason for this is that as additional strands extend away from the BCEF core, the residues on these strands that continue the backbone interactions from core anchors shift away from the hydrophobic core of the Ig-fold due to the angle between the two beta-sheets. Oftentimes, these additional strands break off early and form short strands, as seen with C". Based on this, we selected the anchor of the C" to be an in-facing residue that interacts through its sidechain with core anchors, and this position appears to be conserved within subclasses of Ig-domains. Igs# 5550 is typically a Tyr in VH domains whose sidechain hydroxyl group interacts with the backbone of igs# 6550 on the D strand, providing a bulky sidechain that protects the hydrophobic core as well as a polar interaction that stabilizes the intersheet distance (PDB: 5ESV). In VL domains, Arg appears to be highly conserved at the igs# 5550 position.

Igs# 6550 and 7550 are adjacent to one another and face each other through the alpha carbons of their backbones. Igs# 6550 is less conserved at the sequence level and the residue here can vary depending on the overall structure of the Ig-fold and the type of in-facing residues on the C' and C" strands, whereas igs# 7550 is highly conserved and is typically a Leucine in classical Ig-folds that stabilizes the Tryptophan igs# 3550 and may potentially be important for the folding of an Ig-domain. Both igs# 6550 and 7550 are shifted two positions away from the backbone network of interactions spanning from igs# 2550 on the B strand. The reason for this positional shift is similar to that of the C" (core interactions) strand.

The F strand is the most conserved strand in sequence and structure across all Ig-domain variants (see Figure 3, Figure 5 and Figure 6), and contains what may be the most conserved residue of all Ig-fold

variants: a Tyrosine at position igs# 8548 that has been called the Tyrosine corner (Hemmingsen et al. 1994; Hamill et al. 2000). This Tyrosine is an in-facing residue that usually points its sidechain towards the EF loop where it typically interacts with the backbone of an Aspartate residue in the small helix of this loop (PDB ID: 5esv). It is present in most canonical Ig-domains IgVs, IgC1s, IgC2s, but can also be seen in other Ig-like domain variants such as many FN3, cadherin, and ß-Arrestin domains (Figure 3). Noticeably, in these domains, this Tyrosine appears at position igs# 8546 and further away from the core as compared with classical Ig-folds. Some other residues Phe, Val and Ile can sometimes appear at that position instead of the Tyrosine (see Figure S2).

The G strand anchor residue (igs# 9550) is adjacent to the conserved Cysteine residue (igs# 8550) on the F strand. The igs# 9550 residue is typically glycine in antibody VH and VL Ig domains, whose backbone amino group is a hydrogen bond donor to the carbonyl group of the Cysteine igs#8550. The carbonyl group of the Glycine igs# 9550 is kinked away from the F strand, forming the structurally conserved bulge in the F strand of VH and VL domains. Interestingly, the amino group of the Cysteine igs#8550 is stabilized by an interaction with the sidechain of the A strand (igs# 1550) E/Q anchor residue. In antibody CH and CL domains, lysine appears at the igs# 9550 position which now both donates and accepts backbone hydrogen bonds with the Cysteine igs# 8550, hence there is no bulge in the G strand.

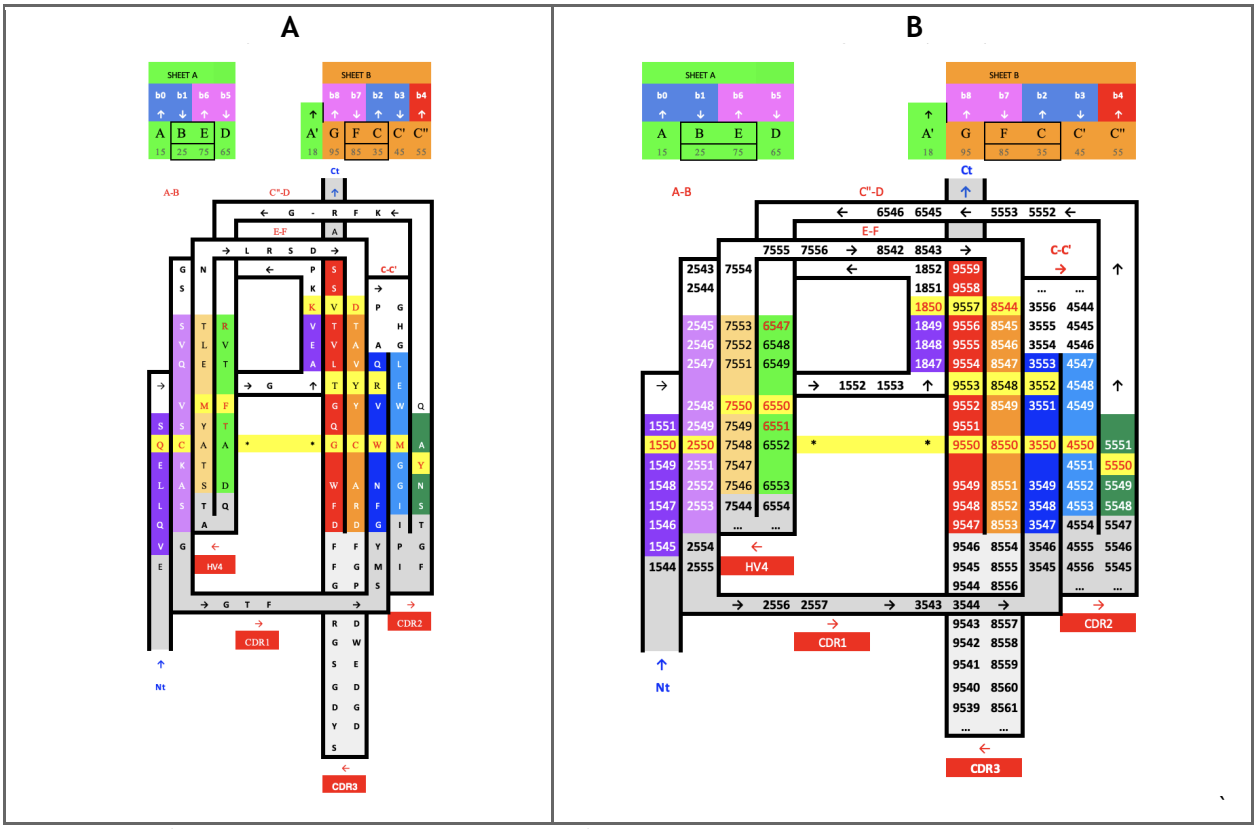


Figure 6 - IgV domain Proteomap A) Sequence/Topology Map (PDBid 1RHH) of the VH domain show both the the sequence and topology simultaneously https://structure.ncbi.nlm.nih.gov/icn3d/share.html?DTtCtLWevzwS2XkJ8&t=1RHH. **B)** IgStrand numbers corresponding to IgV domains with the AA'BCC'C"DEFG topology

Ig-extended domains: Ig-fold topostructural variable regions

Ig-fold canonical variants of **Ig- and Ig-like domains** vary in the following regions: **The linker region between C and E strands** differentiates the canonical immunoglobulin domain variants IgC1, IgC2, IgI, and IgV; **the N-terminus A strand** offers significant variability, with a continuous A strand in IgC1 domains, a split A/A' strand in IgI, IgV, and some IgC2 domains, or in some IgVs such as CD4 only a single A' strand is found; **the C-terminus invariant G-strand** often offers a strand split in the middle but remains on the same GFC(C') sheet in Ig-domains to form an FG loop called CDR3 in antibodies. This leads to a 5-strand supersecondary structure (5s-SSS) forming the irreducible Ig-fold structural signature (Figure 4).

Besides canonical Ig-domains topologies, **Ig-extended** domains exhibit more variations: **N-terminus strand additions,** beyond the classical A/A' strand split, can present additional strands that we name A+ or A- that can be positioned after or before in sequence and in structure on either sheet; and **C-terminus strand additions** that we name G+, G++, etc., can extend the Ig-fold on either sheet of the beta sandwich. **Strand insertions** also occur in a variety of positions in the two sheets of the beta sandwich. As mentioned above, insertions can lead to additional strands. An exhaustive survey is beyond the scope of the current paper, but we will see a number of examples in the following sections. The most surprising insertions leading to extended sheets may be stemmed in the central BC or EF loops leading to multiple variants: **insertion of a sequence between E and F** can lead to the formation of an additional E+ strand. It is characteristic of Lamins, extending the ABE sheet as ABEE+ similarly to the ABED sheet in canonical Ig-domains IgC1/I/V where the E+ strand is positioned as a D strand in Ig-domains but in reverse strand directions (see Figure 2 and Figure 4). This introduces an asymmetry with the formation of an E+F straddling loop instead of a canonical EF loop in the 4s-SSS; similarly, **insertion of a sequence between B and C** can form**,** as in orthopoxviruses immune evasion proteins for example, a beta hairpin extending the GFC sheet in the same positions as the canonical C" and C' in IgV domains in reverse order. It introduces an asymmetry vs. the canonical BC straddling loop in the 4s-SSS and what may look like a permutation between strands C' and D. This type of Ig-domain extension is not shown in Figure 2, but can be seen later in the example of the monkeypox virus M2 protein in Figure 11. A number of Ig-extended examples are presented hereafter.

Finally, **strand permutations** can also bring an additional level of combinatorial plasticity to the Ig-fold and help relate other sandwich folds to the Ig-fold itself. Circular permutations in sequence leading to structurally homologous structures are common in proteins. In C2 domains for example the A strand at the N terminus can be circularly permuted at the C terminus (Figure 7). Apparent simultaneous insertions/deletions as in the case of the Transthyretin/Prealbumin fold (SCOP b.7) of strands can also be seen as a pseudo-circular permutation where a G+G++ at the C terminus can be seen as replacing a possible A-A harpin at the N-terminus. (Figure 7).
In summary, numerous domain variants can be classified as Ig-extended when considering strand insertions, permutations and N/C extensions, in beta sandwich sheets, but also deletions

except for the four central B,C,E,F strands (4s-SSS). This can occur with N-terminal and C-terminal strands.

The Ig-fold extensions are reviewed below through a number of examples found in nuclear, cytoplasmic and extracellular regions of cells in a diverse set of biological functions.

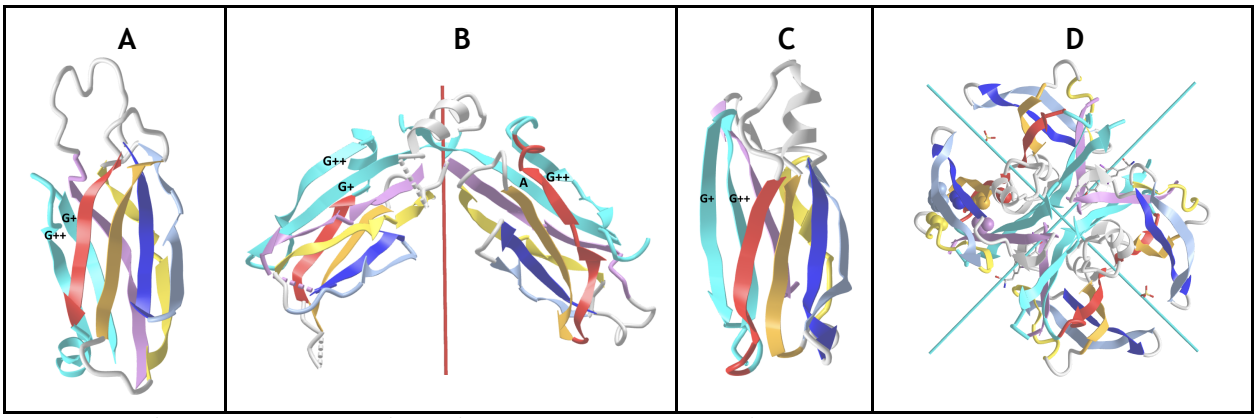


Figure 7 - Ig domain extension and circular permutations A) C2A domain - This type II C2 domain has a G+G++ hairpin extension (shown in cyan) and no A. The G++ is a circular permutation of the A strand of a Type I C2 domain. **B) Synaptotagmin 2 tandem domains C2A - C2B** (PDBid: 4P42) in a C2 pseudo-symmetric head to head tandem arrangement . The C2B domains the A strand circularly permuted w.r.t. G++ strand in C2A (A and G+ in cyan) https://structure.ncbi.nlm.nih.gov/icn3d/share.html?T2NCEnWVHaoW2AE9A&t=4P42. **C) Transthyretyin (PDBid 4P42) domain. Similarly it shows a G+G++ extension but in a reverse order vs. C2A forming a parallel G++ strand to the B strand** https://structure.ncbi.nlm.nih.gov/icn3d/share.html?wgNVJar5M1ZBwiuL9&t=2ROX. **D) Transthyretin tetramer with a D2 symmetry** formed by the G+G++ hairpin.

Arrestin

Arrestin is particularly interesting as it is composed of two extended Ig-domains in tandem with an unusual head-to-head arrangement, as compared to the classical tail to head arrangement found in Ig-chains (in the N to C direction of Ig chains). The Arrestin domain (Aubry, Guetta, and Klein 2009) is a good example of an Ig-fold variant with an A- strand extension at the N-term that precedes that A strand, as well as an additional E+ strand following the E strand to form an A-ABEE+ sheet facing a GFCC' sheet. The tandem Arrestin results from the G strand of the Arrestin_N domain pipelined to the A- strand of the Arrestin_C domain. This A- strand of the C-terminus domain acts like an inversion strand for the C-terminal Ig domain and results in a head-to-head pseudo-symmetric structure (Figure 8A). **This shows the profound change in chain architecture obtained through the Ig domain A- strand extension.**

It is important to note that the "head" we consider here is formed by the CC' and EF loops. When compared to canonical Ig domains, Arrestins have significantly extended loops that are key to their function in binding GPCRs. In the Arrestin_N domain, the Finger loop corresponds to an extended CC' loop that engages the cavity between the transmembrane helices of activated G-protein-coupled receptors (Vishnivetskiy et al. 2021), and the Middle loop corresponds to an extended EF loop (see Figure 8B). These two loops correspond respectively to what could be called the anti-CDR3 (CC') and the anti-CDR1 loop (EF) that are positioned pseudo-symmetrically to FG (CDR3) and BC (CDR1) loops in antibodies (Youkharibache 2019).

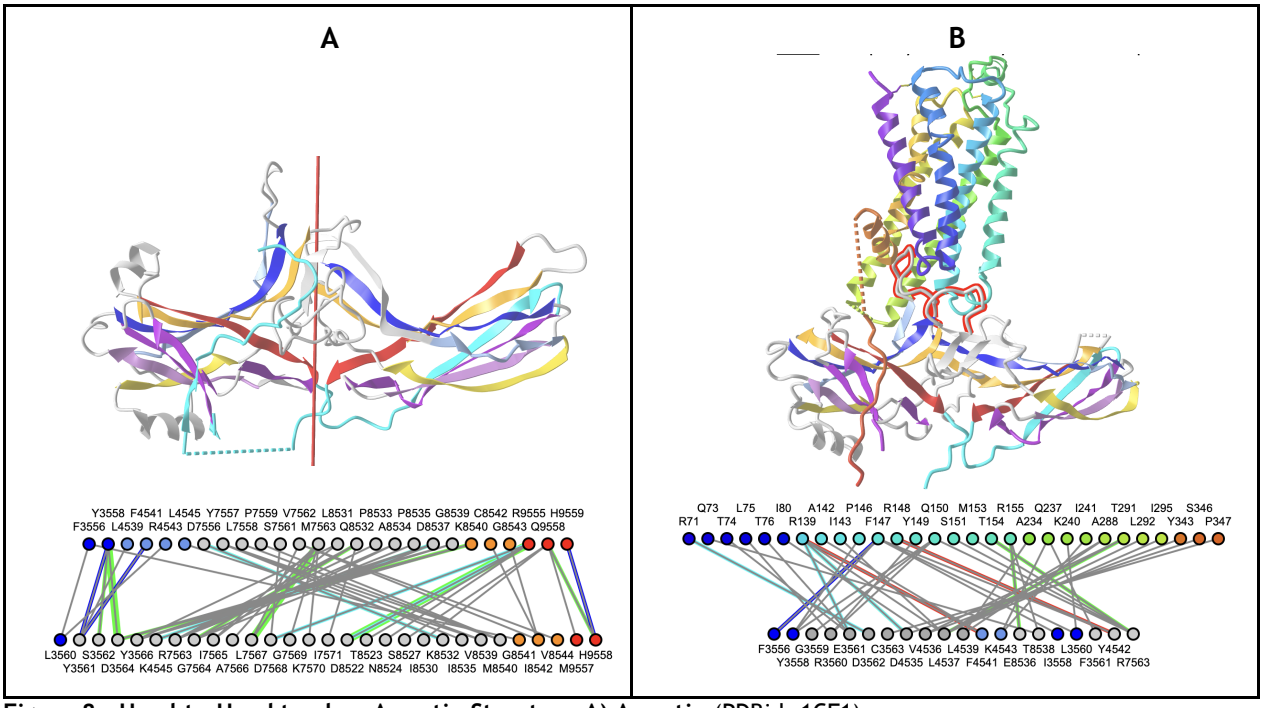


Figure 8 - Head-to-Head tandem Arrestin Structure A) Arrestin (PDBid: 1CF1) https://structure.ncbi.nlm.nih.gov/icn3d/share.html?1LKCe6fyS1fzacVL7&t=1CF1 **B) Arrestin and GPCR 7-TM domain interacting in a active conformation** (PDBid: 6TKO). The Arrestin_N domain's Finger loop (CC') and Middle loop (extended EF loop) interact with the transmembrane helices of activated G-protein-coupled receptors. The additional A- strand (see text) is shown in cyan to highlight its role in enabling a head to head configuration between the two Arrestin Ig domains. https://structure.ncbi.nlm.nih.gov/icn3d/share.html?He7U1h5mUbipbVW58&t=6TKO.

Lamin

Lamins are thought to be present only in the animal kingdom. They are Intermediate filament proteins that make up the nuclear lamina and are involved in diverse functions such as chromatin organization, gene regulation, and cell differentiation (Dittmer and Misteli 2011). There are three Lamin genes in the human genome: Lamin B1, Lamin B2, and Prelamin A/C. The latter undergoes post-translational processing into Prelamin A or Prelamin C as the two major forms. Human Lamins contain an Ig-like domain in the tail region, the LTD (Lamin Tail Domain) exhibiting several unique structural features with respect to a typical Ig domain.

Lamin Tail domains (LTD) contain 9 strands **A^{-}ABCC'EE^{+}FG** that differ in part from canonical IgV

or IgC domains (Figure 9). They form a beta sandwich with two sheets **with a highly unusual Sheet1 formed by ABEE⁺,** lacking the antiparallel D strand (in the usual ABED Sheet1 encountered in IgV and especially IgC domains) but containing an E⁺ strand instead, following the E strand and running parallel. **Sheet2** on the other hand is formed as **A⁻GFCC' strands where the G strand is flanked by the additional A- strand in antiparallel instead of the parallel A' strand found in IgI/IgV domains (A'GFCC')**. Topologically, apart from the E⁺ strand insertion, it is similar to FN3 domains that also possess no D strand and form a **C'E cross-ß sheet loop**. However, the peculiar insertion of an E⁺ strand gives rise to a **unique E⁺F cross-ß sheet loop**, instead of the canonical EF loop, that forms a beta strand and antiparallel 2-stranded beta sheet with the BC loop (CDR1) also forming a beta strand, a very unique structural region of Lamins, implicated in Lamins specific functions and where mutations cause a number of diseases (Krimm et al. 2002; Dittmer and Misteli 2011). The Ig hallmark of the Ig-fold consists in the four central beta-strands B,C,E,F, forming two loops straddling the two sheets of the barrel BC (CDR1) and EF. Lamin's E+ insert displaces the latter as a **E⁺F** loop, a notable difference w.r.t. other Igs that may give it specific structure-function specificity; but conserving the GFCC' sheet with CC' and FG(CDR3) loops as in classical IgV antibodies.

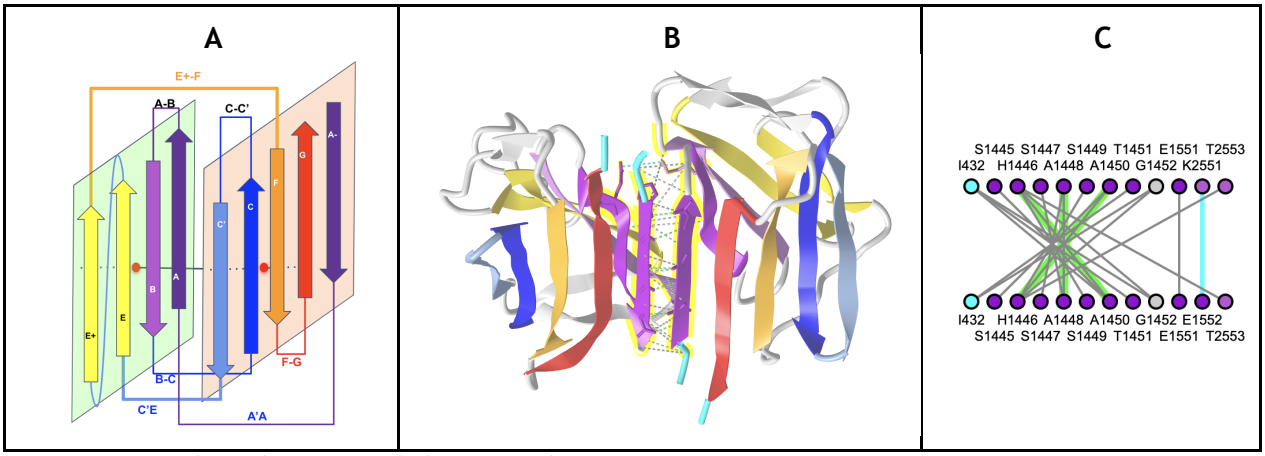


Figure 9 - Example of the Ig-extended Lamin Tail Domain (LTD). A) Schematic LTD strand topology forming sheets ABED+ and A-GFCC'. **B) Pseudo symmetric LTD dimer** brings the two A-GFCC' as one in antiparallel through the additional A- strand at the N terminus. **C) LTD dimerization interface** using the IgStrand numbers for the A- strand (i=1; j=4 running from 1445 to 1452) and a few residues on strand A and B. Note that the current IgStrand numbering does not cover residues before 1445 (based on current template - see Table 2)
https://structure.ncbi.nlm.nih.gov/icn3d/share.html?BRks581vEqzUpENs9&t=7DTG

The presence of the additional A⁻ strands on the N-terminus sandwich barrel edge offers a side-by-side dimerization interface, observed in experimental structures of individual Lamin Tail domains (LTD) (PDB: 3UMN;7DTG in Lamin B1) (Ruan et al. 2012; Ahn et al. 2021) with the formation of backbone hydrogen bonding of the A- strands of each monomer in antiparallel.

Transcription factors Ig-domains

Some transcription factor domains exhibit Ig-like and Ig-extended domains. This is the case for tumor suppressor p53 (Figure 10A) that possess an N-terminal A+ strand extension between A and A' and an insertion that forms two extra strands, one contributing to the ABE

strand in position of a D strand and one contributing to the GFCC' sheet in a C" position before the C' strand (see Figure 10). Similarly, the Rel Homology Domain (RHD) is found in eukaryotic transcription factors, in particular in NF-κB and NFAT. It exhibits an extension forming two strands as in p53 between C and C'. Its N-terminus however is similar to IgV domains forming a split AA' strand. It is followed by another Ig-C2 like domain, the IPT (Immunoglobulin, Plexin, Transcription factor) domain, also known as TIG (Transcription factor ImmunoGlobulin). In the p53 DNA binding domain or the NFAT or NF-kB Rel homology domain the linker region between C and E strands is inverted as compared to the canonical IgV domains: if we name strands according to their positions in the latter, the topology runs C>DC"C'>E as opposed to C>C'C"D>E, leading to a permuted IgV-like domain.

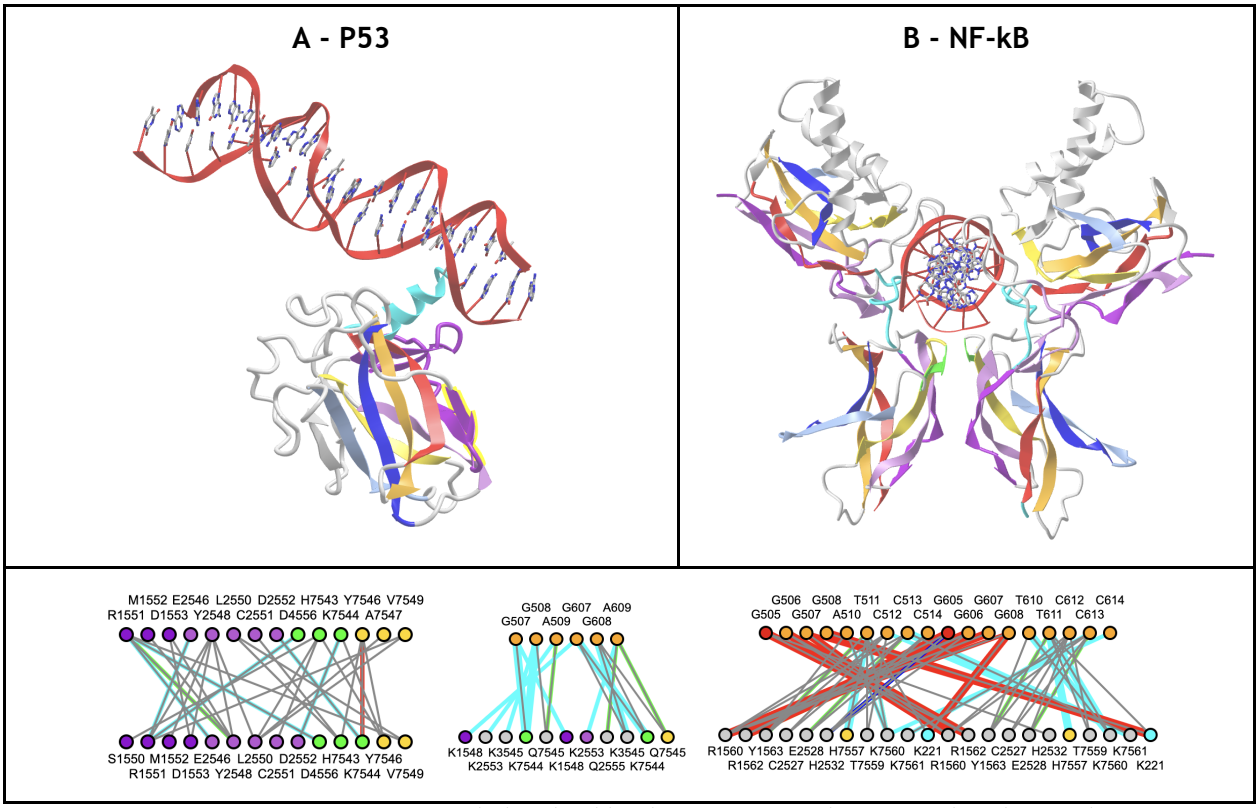


Figure 10 - Transcription factors Ig-extended and Ig-like domains. A) NF-kB RHD and NF-kB IPT dimer. The Rel Homology Domain (RHD) in tandem with the Immunoglobulin-Plexin-Transcription factor (IPT) domain form a dimer that binds DNA https://structure.ncbi.nlm.nih.gov/icn3d/share.html?NVxjekmSMxWnrs5P7&t=1A3Q. The RHD exhibits an insertion between strands C and C' extending the ABE strand in position of a D strand and one contributing to the GFCC' sheet in a C" position before the C' strand in that order. The linker region between C and E strands is therefore inverted as compared to the canonical IgV domains. If we name strands according to the IgV canonical strand names, the topology runs C>DC"C'>E in RHD as opposed to C>C'C"D>E in IgV. Overall the topology can be seen as a permuted IgV-like topology, forming two sheets GFCC'C" and ABED as in IgV domains. It is followed by the IPT (Immunoglobulin, Plexin, Transcription factor) domain, also known as TIG (Transcription factor ImmunoGlobulin), with an IgC2 topology extended by an unusual small D strand and a split AA' N-Terminal strand. The two IPT domains dimerize through their ABED sheet, in parallel. The two chains bind DNA pseudo-symmetrically. **B) P53 DNA Binding Domain** possesses, as in NF-kB RHD, a permuted IgV-like topology https://structure.ncbi.nlm.nih.gov/icn3d/share.html?RiTYSGVRTWCA2AyZ8&t=1TUP with an additional N-terminal strand extension before and between A and A'. **C) Residue Interactions** in the IPT-IPT dimer interface, **the two IPT domains** - DNA interface, and the **two RHD domains** - DNA interface.

Ig-domains in viruses

Ig-like domains are found in numerous viruses. This has been largely described in the literature (Farré et al. 2017; Maginnis 2018). Figure 11A shows the example of the murine leukemia virus' RBD (Receptor Binding Domain) (PDBid: 1AOL), responsible for binding the murine PIT2 receptor, that exhibits an extended Ig-domain framework with receptor binding regions called VRA and VRB correspond to loops BC (CDR1) and DE (HV4) in canonical Ig-domains. The Ig-domain is extended at the N-terminus with an additional A+ strand hydrogen bonded to the G-strand in antiparallel (insertion between A' and G) extending the GFCC'C" sheet (Barnett et al. 2003; Fass et al. 1997) to A+A'GFCC'C" of an otherwise IgV topology. While Ig-related genes have been acquired by viruses from the ancient genomes, it is interesting to note that integration in the human genome of an estimated 8% of sequences of retroviral origin (HERV) that are remnants of ancient retroviral endogenization. Among them, the endogenous retroviral envelope EnvP(b)1 protein is coded by a gene found in human and other primate genomes, and traced to an event estimated between 40 and 71 million years ago (McCarthy et al. 2020; Hogan and Johnson 2023). The receptor binding domains (RBD) of human EnvP(b)1 structure (PDBid: 6W5Y) exhibits an IgV topology (A'GFCC'C" - ABED) that defines structural similarities with extant leukemia viruses' envelopes, the main difference lying in the presence of the additional A- strand in the latter, pointing to a deletion or addition of that strand in EnvP(b)1 during its evolution.

Viral pathogens have evolved sophisticated mechanisms to evade the host immune system, this is the case for example of orthopoxviruses (Senkevich et al. 2021; Mutz et al. 2023) that contain poxvirus immune evasion proteins (PIE) (Nelson et al. 2015). PIE domains can be considered Ig-extended. For example, the SECRET domain (smallpox virus-encoded chemokine receptor) represents a family of viral CKBPs (chemokine-binding proteins) that modulate the chemokine network in host response in binding chemokines to inhibit their activities (Xue et al. 2011). The recently determined monkeypox virus M2 protein structure (Figure 11B) shows how it interacts with the immune system (S. Yang et al. 2023), with an exquisite structural interface between the Ig-extended MPXV M2 protein and the extracellular variable Ig-domain of CD80 (Figure 11B) that interrupts the natural interactions of human B7.1/2 with CD28 and CTLA4 and subverts T cell activation mediated by B7.1/2 costimulatory signals.

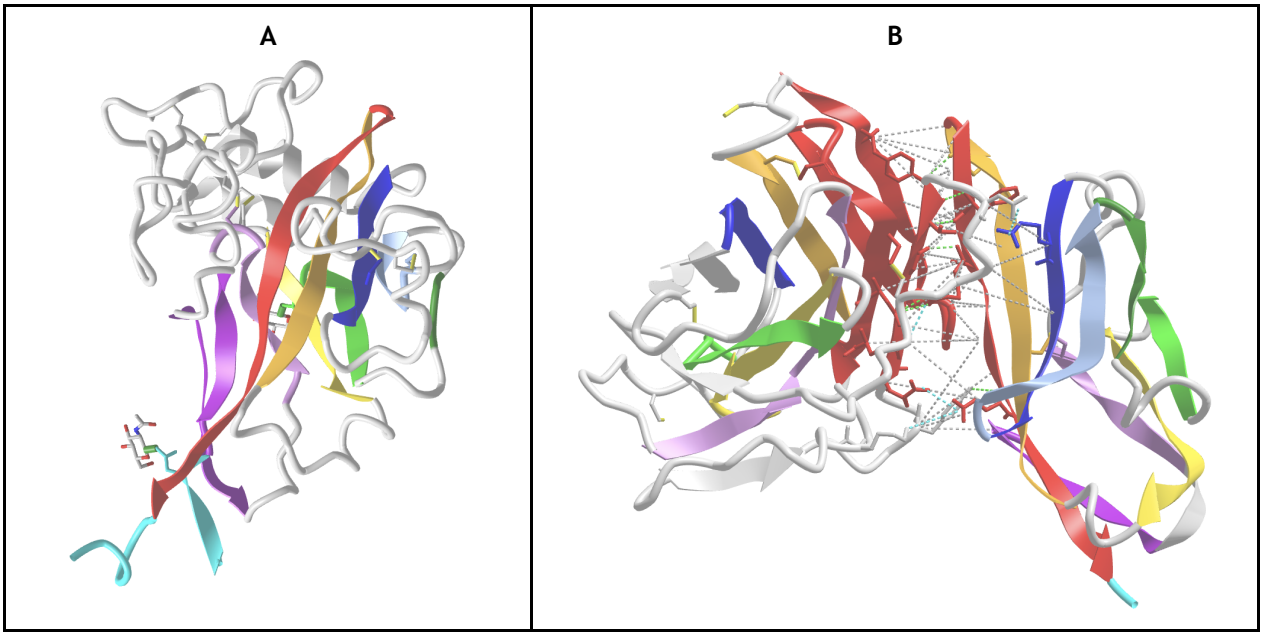


Figure 11 - A) The Murine Leukaemia viruses envelope RBD (PDB id 1AOL) SCOP b.20 shows an extended Ig-domain framework with receptor binding regions called VRA and VRB correspond to loops BC (CDR1) and DE (HV4) in canonical Ig-domains. The Ig-domain is extended at the N-terminus with a A'- strand hydrogen bonded to the G-strand in antiparallel and forming a A'-A' hairpin extending the A'GFCC'C" to A'A-GFCC'C" where the A'- in inserted between the A' and G strand https://structure.ncbi.nlm.nih.gov/icn3d/share.html?hA7Ftj6A2QEMoCSV7&t=1AOL**.** The leukemia virus envelopes can be compared to the Feline Leukemia virus RBD as well as the endo-retrovirus EnvP(b)1 https://structure.ncbi.nlm.nih.gov/icn3d/share.html?oFgS8acCM7AW1hZDA&t=6W5Y,1LCS,1AOL**. B) The monkeypox virus protein M2 binding CD80.** (PDBid: 8HXA). This beta sandwich is similar to the SECRET domain found in other orthopoxviruses binding chemokines. This domain is also called PIE (Poxvirus immune evasion). In this case, the M2 protein modulates T cell co-stimulation in binding CD80. It is missing the A strand altogether at the N terminus. It also presents an insertion of two strands between B and C forming a hairpin C--C-, that take the place of C" and C' (in reverse order vs. a canonical IgV domain), extending Sheet 2. The C-terminal extension is composed of 3 additional strands to the G-strand, first a G+ intercalating strand forming a GG+ hairpin and a parallel beta sheet interface to strand F, extending the GG+FCC' followed by a G++G+++ hairpin extending Sheet1 as G+++G++BED, the latter G+++ strand replacing the missing A strand, similarly to a circular permutation as found in type II C2 domains. The M2 domain binds the G-strand (red) of CD80 laterally through its C-terminus G-strands extension (red), mainly G++/G+++ , to modulate T cell co-stimulation. https://structure.ncbi.nlm.nih.gov/icn3d/share.html?cbid7J2RCJhMNRca8&t=8HXA (S. Yang et al. 2023)

Ig or not Ig? Blurring (or swapping) the limits of the Ig-fold

While our IgStrand numbering scheme is aimed at representing all domain variants of an Ig-fold, it can apprehend similarities with other beta sandwiches, using the same strand nomenclature.

Any Ig-domain variant contains the full 5-strand core supersecondary structure formed by B-C, E-F, plus either an A, A' or split A/A' strand at the N-terminus, a central linker between C-E strands with one or more strands among C',C",D, and the G-strand at the C-terminus and we defined Ig-extended domains as encompassing the Ig-fold irreducible signature with additional strands as explained earlier (Figure 2).

Strand swaps can also be considered and can relate diverse variants of Ig-like vs. beta sandwiches classified as different folds, such as jelly rolls (Richardson 1981). Beta sandwiches are estimated to represent overall 30% of all folds composed of beta strands (Edwards, Abeln, and Deane 2013). However, among these, after the Ig-fold, Jelly-rolls represent the second major beta sandwich fold. In terms of numbers, taxonomy databases classify dozens of beta sandwich folds, for example CATH currently counts 44 of them representing 536 superfamilies for a total number of 60,890 domains known experimentally. The Ig fold alone accounts for 330 superfamilies (62%) and 45,422 domains, while the Jelly-roll accounts for 121 superfamilies (22%) and 11,770 domains, the two combined representing 84% of beta sandwich superfamilies, dwarfing the other 42 folds combined that represent only 85 superfamilies (16%) and 3,698 domains known experimentally. These numbers can now be re-examined with predicted structural proteomes becoming available, and work is underway on the Taxonomy of the Alphafold database (Schaeffer et al. 2024; Lau et al. 2024). Focusing on the Ig-fold and Jelly Roll fold, it is clear that evolution has favored these two main folds to perform a myriad of biological functions. One can relate the two folds topologies by considering a conceptual

concerted strand swap of the A and G strands at the N- and C-terminus, respectively (Figure 12).

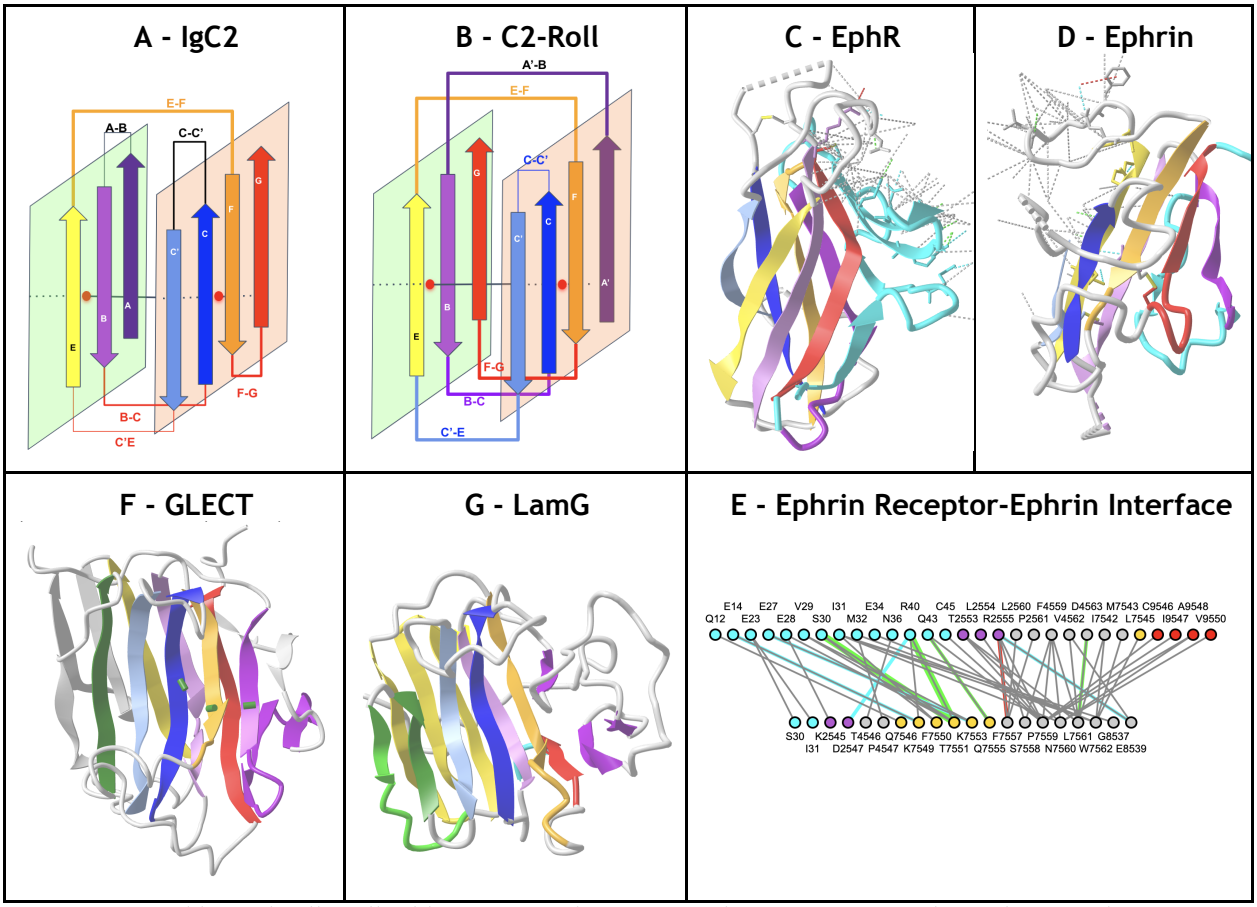


Figure 12. - Ig-like and Jelly-roll folds. A) IgC2 schematic topology B) Conceptual A and G strands swap comparing Jelly-roll to Ig-like transforming AB and GF loops into straddling loops between Sheets 1 and 2 of the sandwich changing from ABE and GFCC' in an IgC2 domain to BGE and AFCC' sheet respectively in what could therefore be called a C2-roll domain, using the Ig strands nomenclature. **C) The Ephrin Receptor** (PDBid: 3GXU), a **Galactose-binding domain-like fold** classified as a jelly-roll in CATH, and ECOD and SCOP (**b.18)** represents a good example of an extended "C2-roll" domain where the N-terminus provides an additional set of strands, extending both sheets and providing the bulk of the interface its ligand Ephrin. **D) Ephrin** exhibits an Ig-extended **Cupredoxin-like** beta sandwich fold (SCOP **b.6**). The G-strand is forming, as in many IgV domains, an FG loop and a parallel A' strand (A'GFCC') but it is preceded by an N-terminus extension (in cyan) providing additional strands A- on Sheet 2 to form A-A'GFCC' and a strand A-- on Sheet 1 to form the A--BE, the interacting surface to the Ephrin Receptor https://structure.ncbi.nlm.nih.gov/icn3d/share.html?kRY8dJJvV4aM7fFx6&t=3GXU. **E) The Ephrin Receptor-Ephrin interactome. F) Galectin (GLECT) domain** is a galactose-binding lectin (PDBid 4AGV) that represents **yet another Jelly-roll variant topology,** classified as a **Concanavalin A-like** (SCOP **b.29**). GLECT exhibits the A and G strand swap between the sandwich sheets as compared to an IgV domain, extending one sheet with an additional C''' strand (in grey) after C" to form a sheet AFCC'C"C''', and with a two strand E+E++ β-hairpin insertion (in grey) after E to form the GBEE+E++ extended sheet.
https://structure.ncbi.nlm.nih.gov/icn3d/share.html?rYBVVeHiBh2TkPVM6&t=4AGV. **G) Laminin G-like (LamG) domain is** also classified as a **Concanavalin A-like jelly roll** with a two strands extension at the N-terminus extending the first sheet A--A-GBEE+E++ as compared as to GLECT
https://structure.ncbi.nlm.nih.gov/icn3d/share.html?bmYk9ciuwBQb2td47&t=5DZE.

Three levels of organized complexity: Ig domains, chains, assemblies

We can consider three levels of complexity: at the tertiary domain, tertiary chain, and quaternary assembly. Many protein chains contain Ig- or Ig-like domains. When considering the Ig-fold plasticity in light of extreme diversity of molecular surfaces of interaction they can form, it is no surprise, but awe, that the hierarchical complexity of tertiary Ig-chains and their quaternary assemblies can orchestrate complex systems in vertebrate organisms. 3D structures start to emerge for proteins such as Pom210 with Ig-chains counting 17 Ig-domains in tandem in the context of the full nuclear pore receptor complex and allow to visualize conformational plasticity of Ig-chains (Y. Yang et al. 2023; Mosalaganti et al. 2022). In the cell surface proteome, or surfaceome, a high number of cell surface receptors responsible for cell-cell interactions at the heart of the immune system, the nervous system, the vascular system (see examples in Figure 1), the extracellular regions (ectodomain) of single pass cell surface proteins form Ig-chains that can contain from one to tens of Ig domains, and even in hundreds in the muscular system. CD8 for example contains one extracellular Ig domain, CD4 contains four of them, CD22 has seven, cell adhesion proteins such as DSCAM (see later) can contain up to 16 Ig-like domains, including FN3 domains. Much longer chains, such as the giant protein Titin can contain hundreds of chained Ig and FN3 domains (Kenny, Liston, and Higgins 1999). Considering alternative splicing of Ig-domain containing genes can produce a myriad of proteins with variable numbers of Ig domains, adding more complexity due to the already large number of Ig domain variants, a bewildering number of Ig-domain combinations in Ig-chains (see for example (W. Wu et al. 2012; S.-A. Li et al. 2016)), making the Ig-proteome very large and diverse.

Self assembly of Ig-domains

Considering the omnipresence of Ig-domains among cell surface receptors, especially in the immune system and the nervous system, one is naturally led to consider their role in cell communication. This is a vast subject, but we only want to focus on the contacting mechanisms of these molecules at the surface of cells that may lead to attractive or repulsive actions involved in synapse formation. In neuronal synaptogenesis, the primary cells involved are neurons, which form synaptic connections with other neurons or target cells (S. Cheng, Ashley, et al. 2019; Xu et al. 2018; Kulahin et al. 2011; Cortés, Pak, and Özkan 2023). In immunological synaptogenesis, various immune cells such as T cells, B cells, dendritic cells, and macrophages, form contacts with each other or with non-immune cells such as APCs or target cells.

A striking feature of Ig domains at the surface of cells is their pairing, in both cis and trans. External residues on the surface of Ig domains themselves represent a 3D scaffold for Ig-recognition. Yet that recognition, both attractive and repulsive at the heart of self- vs. non-self, is based on co-evolved codes between external residues between Ig-receptors and Ig-ligands (see for example (Harrison et al. 2012)). The combinatorics of 3D Ig-pairing for a given Ig-domain variant type such as IgV domains, found in so many cell surface receptor-ligands interfaces, is staggering. Multi-Ig-domain chains bring even greater combinatorial capabilities (Zipursky, Wojtowicz, and Hattori 2006; Wojtowicz et al. 2007; Sergeeva et al. 2020; Sanes and Zipursky 2020; Honig and Shapiro 2020). When one considers the diversity of Ig-domain topostructural variant types and/or Ig-chain constructs with alternative splicing, the conceptual pairwise interactome numbers become impossible to grasp.

While a myriad of distinct Ig-Ig interactomes discriminate self- vs. non-self between cells, a number of Ig-Ig interactomes have been determined experimentally leading to the observation of conserved Ig-Ig pairwise interaction interface structures that can in some cases be seen as “quaternary folds”. This is the case of IgV-IgV canonical interfaces found in antibodies, TCRs but also in many cell surface

receptor ligands of the immune system or the nervous system, some being common to both such as nectins. An initial survey of the Protein Data Bank (PDB) detected the presence of at least one Ig-domain in 16,661 protein structures that included 28,795 pairs of Ig-domains in contact, with 24,222 with high confidence, as identified through the IgStrand algorithm (see methods).

The IgStrand based Ig-detection algorithm is highly reliable in detecting domains with an Ig-fold; but it does not accurately label the variant types (this is a function of the structural fit with templates (Table 2)). Work is underway to improve this accuracy and determine the diverse classes of interfaces (or quaternary folds) yet an early analysis using CDDs tend to show that Ig domains pair according to their types, i.e., IgV-IgV, IgI-IgI, IgC1-IgC1, IgC2-IgC2.

The canonical quaternary antibody IgVH-IgVL and IgCH1-IgCL interactomes

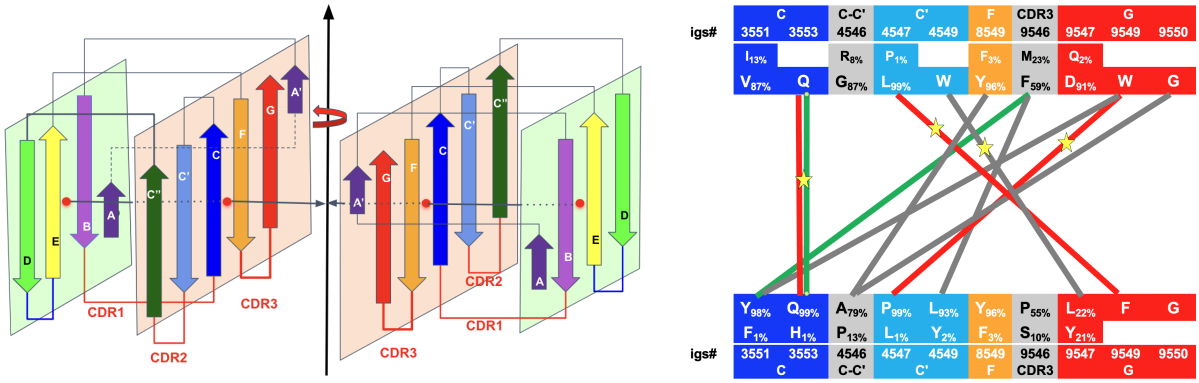
Antibody heavy and light chains assemble through both i) VH and VL domains using the [A']GFCC'[C"] sheet in parallel, and ii) CH1 and CL domains using the ABED sheet in antiparallel, respectively (see schematic Figure 13). The paired variable and constant domains form interfaces VH|VL and CH1|CL interfaces that can be considered canonical. The affinity and specificity of an antibody towards an antigen results from diverse factors among which the pairing and dynamics of VH|VL and CH1|CL domain quaternary interfaces and their relative variable-constant two-domain interfaces and elbow angle (Fernández-Quintero et al. 2020; Stanfield et al. 2006; Lesk and Chothia 1988). The dynamics of the interface can be measured by the elbow angle, which is the angle formed by the pseudo-2-fold axis connecting CH1 to CL and VH to VL. This angle is dependent on various combinations of germline heavy and light chains (Fernández-Quintero et al. 2021, 2023). The residues on the VH-VL interface can influence the VH-VL packing geometry and the characterization of VH-VL and CH1-CL interfaces is important for understanding the function of antibodies. Here, the IgStrand numbering scheme is used to analyze interactions between VH|VL and CH1|CL Ig-domains.

Ig-Ig domain interfaces can be compared across a dataset of any size and across any Ig-fold variants using the Ig-Strand numbers. This is naturally true for variable and constant domains of antibodies. Given the importance of antibodies, we cover in this section some elements of the interfaces between the heavy and light chains that constitute the vast majority of antibody fragment structures. To do so, we selected a dataset covering **258 Fab antibody fragments binding the SARS-CoV-2 spike protein**, contained in 165 PDB structures available in the PDB to survey antibody interfaces (see supplement for details and alignment of the dataset, based in IgStrand numbering).

The analysis identifies 361 unique pairs of interacting residues in the VH|VL interface. Among these, **9 pairs of the VH|VL interface between the GFCC' sheets** are common to at least least 200 antibodies (78%): (igs# **4547-9549**), (igs# **9549-4547**), (igs# **3553-3553**), (igs# **4549-9547**), (igs# 9546-3551), (igs# 9550-4546), (igs# 9546-4549), (igs# 9549-3551), and (igs# 8549-4546), with the first four common to 90% of the dataset and the first three involving symmetric positions (Figure 13A). **The VH and VL domains** present a pseudo symmetric parallel interface. As seen in Figure 13A, the side chains of the highly conserved Glutamine residue in the C-strand igs# **3553** in both VH and VL face each other and form two hydrogen bonds. This residue and dimeric interaction, positioned on the symmetry axis, is also found in homodimers such as VL|VL or non-antibodies IgV domain dimers such as CD8aa. Mutations at that position (Q>E) also present one H-bond, while mutation to a hydrophobic residue Valine disrupts the hydrogen bond interaction pattern, leading to more compact hydrophobic packing (N. Wang et al. 2009).The highly conserved C-strand Tyrosine igs# **3551** in the VL domain forms hydrogen bond interactions with the Phenylalanine igs# **9546** in the CDR3 loop of the VH domain in 149 out of 224 structures where this

pair exits, and Van der Waals interactions with the Tryptophan igs# **9549** in the G strand of the VH domain in 211 structures.

A - IgVH-IgVL parallel



B - IgCH1-IgCL antiparallel

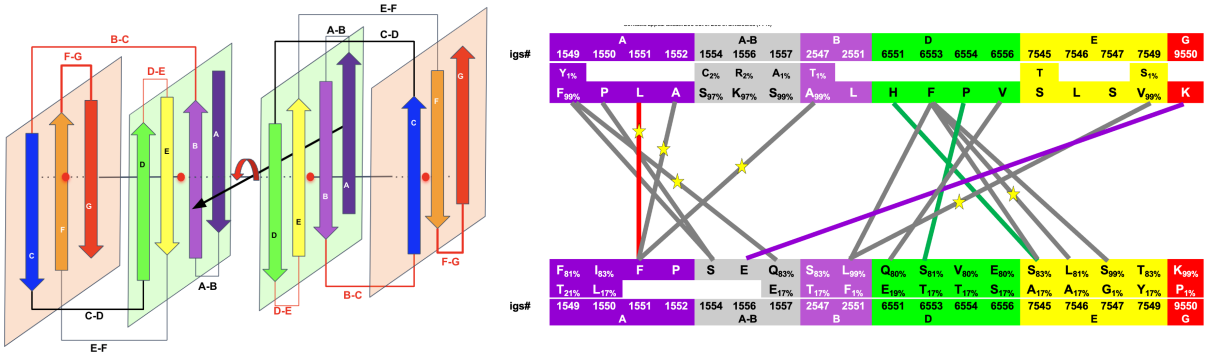


Figure 13 - Schematic representation of Ig-domains dimerization patterns and contact network A) VH|VL schematic parallel interface and conserved interactions B) CH1|CL schematic antiparallel interface and conserved interactions (see text). The VH|VL and CH1|CL interfaces from a Fab (PDBid:7N4I) can be visualized and analyzed in the iCn3D link: https://structure.ncbi.nlm.nih.gov/icn3d/share.html?4JEvtKi1cdPGpW6a6 . On the left a schematic representation of the domain interface with the C2 pseudo symmetry axis. On the right the residue interaction network. Residue subscripts indicate the percentage of occurrence of that residue at that position if less than 100% in the dataset. Ig strands are color-coded according to the ICN3D IgStrand scheme (See Methods). Residue interaction lines between H and L domains are color-coded in red for symmetric contacts, green and blue line for hydrogen bonds and ionic interactions, respectively, according to the dominant interaction type; otherwise, grey indicates Van der Waals interactions.The highly common interactions, present in at least 90% of the dataset, are indicated by a star.

The analysis identifies 213 unique interacting residue pairs CH1-CL with **14 pairs of the CH1-CL in common between the ABED sheets**: (igs# **1551-1551**); (igs# **2547-1551**), (igs# **1552-1551**), (igs# **1549-1557**), (igs# **7549-2551**), (igs# **6553-7546**), (igs# 6553-7547), (igs# 6553-2551), (igs# 6553-7545), (igs# 6551-7545), igs#(6554-6553), (igs# 1550-1554), (igs# 1549-1554), and (igs# 6556-6551), and one pair involving the conserved Lysine residue anchor in the G-strand (igs# **9550**-1556), and where the first six common pairs are present in at least 90% of the dataset and the first at symmetric positions (Figure 13B). **The CH1 and CL domains** also present a pseudo symmetric interface but in antiparallel (see Figure 13B).

This is a canonical interface between CH1 and CL domains in Fabs. As seen in Figure 13B, among the 14 interacting most common residue pairs, the conserved Lysine igs# **9550** is the G strand anchor of CH1 that forms an ionic interaction with the highly conserved Glutamine igs# **1556** from the A-B loop of the CL domain. The side chain of the highly conserved CH1 residues igs# **6551** and igs #**6554** form hydrogen bonds with CL residues igs # **7545** and igs# **6553** in over 140 structures of the dataset, respectively (see Figure 13B). The Phenylalanine residue igs# **1551** in the CL domain A strand makes 3 highly conserved positional hydrophobic contacts with three highly conserved residues in the CH1 domain A and B strands: the Alanine igs# **1552** , the Leucine igs# **1551**, its pseudo-symmetric cognate , and the Alanine igs# **2547** in the B strand. These hydrophobic interactions that give stability to the CH1-CL dimer (Fernández-Quintero et al. 2022) are conserved in at least 90% of the domains. The pair (igs# 1551-1551) is positioned in vis a vis vs. the CH1-CL axis of symmetry. The **IgStrand based sequence alignment** and VH|VL and CH1|CL interactomes for the full dataset are available in Supplements.

One should note that the canonical "quaternary fold" for IgV-IgV paired domains corresponds to a C2-symmetric parallel 8 stranded beta barrel (GFCC')*2 (see schematic Figure 13A), where the A' and C" strands are, in most cases, not forming close interactions in dimerization, and for IgC1-IgC1 a C2-symmetric antiparallel 8 stranded quaternary beta barrel (ABED)*2 (Figure 13B).

Comparison of IgV-IgV quaternary interfaces beyond antibodies

While many different types of interfaces between Ig-domains are possible and do exist, variable domains tend to bind using their (A')GFCC'(C") sheet in parallel, considered as the canonical (parallel) IgV-IgV interface used by antibodies and TCRs IgV domains as well as other cell surface receptors. Immune cell surface receptors such as CD8a/b interacting in cis use this canonical interface as a natural interface orientation. Canonical parallel interfaces between IgV domains, similar to VH|VL, are found in many protein pairs. If one compares for example a VH|VL (PDBid: 7N4I), a VL|VL interface found in Bence-Jones proteins (PDBid: 1REI), and CD8aa (PDBid: 1CD8), the Igstrand numbering identifies nine residue pairs common to all three interactomes, with the conserved symmetrical contacts **(igs# 3553-3553), (igs# 3551-9549) (igs# 3551-9549) and (igs# 9549-3551), (igs# 4547-9549) and (igs# 9549-4547)**. Overall VH|VL and VL|VL interfaces share 15 common igs# pairs, while VH|VL and CD8aa share 18, and VL|VL, and CD8aa share 11 common igs# pairs.
https://structure.ncbi.nlm.nih.gov/icn3d/share.html?4SPQw7jDoGU1tZtbA

The Nectin family also presents this parallel interface while interacting in trans (Harrison et al. 2012; Honig and Shapiro 2020) as well as with cell surface receptors TIGIT (Stengel et al. 2012) and CD96 (Deuss et al. 2018) (see Figure 14). Although the IgV-IgV interfaces use the same GFCC' sheet to bind in either parallel or antiparallel pseudo-symmetric fashion (Youkharibache 2021) between cell surface receptors and ligands, they use a different and highly specific code of recognition between their respective out-facing residues. The IgStrand numbering is aimed at capturing similarities and differences at any scale to residue interaction networks in the most intimate details across the universe of Ig-domains, irrespective of their topological variations. One should note that many other Ig-Ig interfaces do exist, even when using one sheet or the other, but it is beyond the scope of this paper to cover all these interface variants. The IgStrand numbering is precisely designed to capture the diversity of Ig-Ig interfaces as it can capture the diversity of Ig-fold variants themselves. The IgStrand numbering scheme opens the door to positional structural bioinformatics of Ig domains across one or more predicted structural proteomes. Additionally, for known assemblies in the PDB database today and increasingly available through cryo electron tomography, and hopefully through accurate predictions in the future, the IgStrand scheme enables Ig-interactomes analysis.

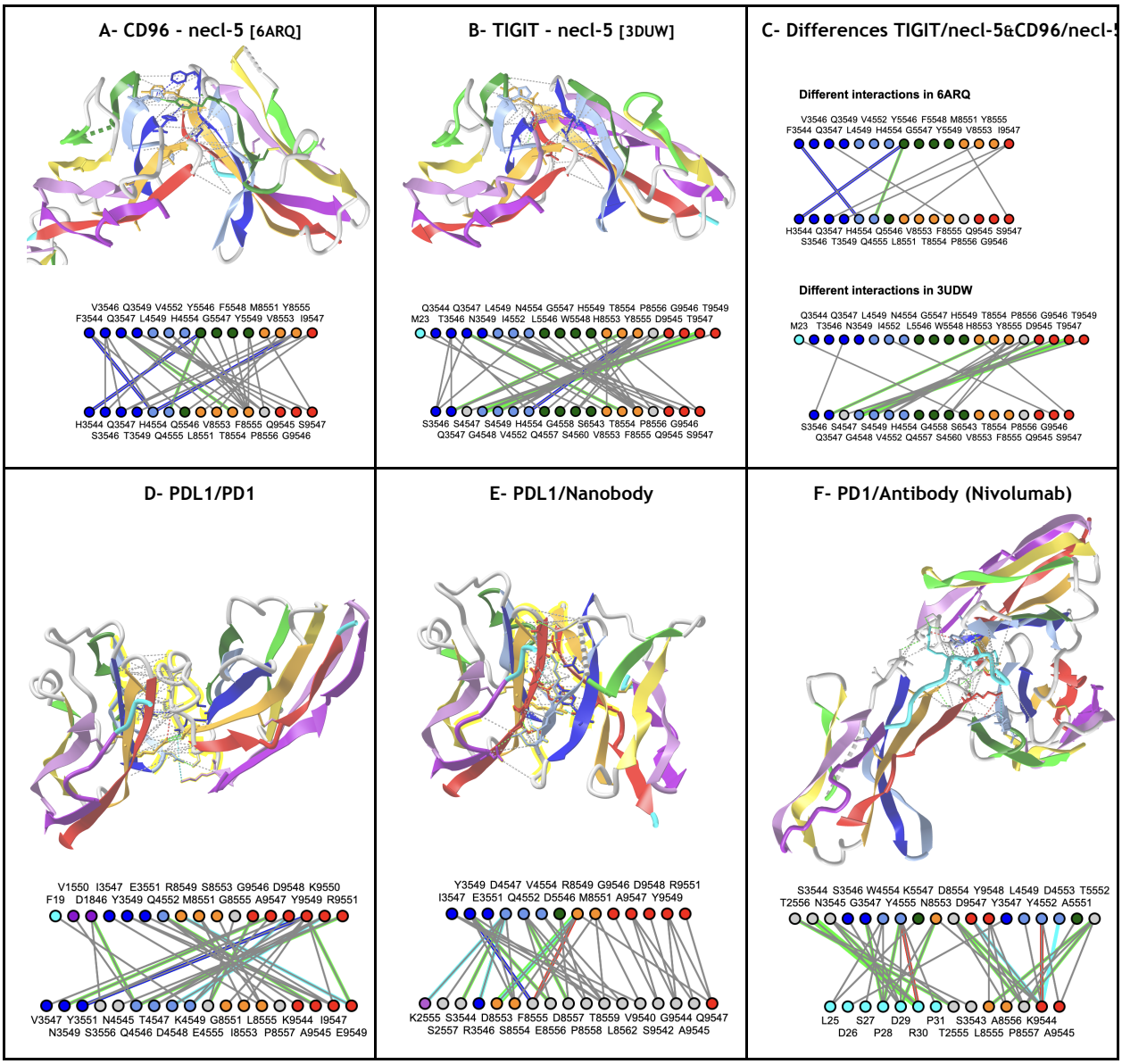


Figure 14- Comparing IgV-IgV interfaces A-B) CD96 vs. TIGIT interacting with nectin-like protein-5 (necl-5) (PDBid: 6ARQ, 3UDW) using their N-terminal IgV domain shown side by side with their residue interaction network, using IgStrand numbering. **C) Different interactions in the residue interaction network in CD96 and in** https://structure.ncbi.nlm.nih.gov/icn3d/share.html?HppETGxGj1yjQ7md7&t=6ARQ,3UDW . **D-E) PD-L1 interacting with PD1 (PDBid 4ZQK) vs. PDL1 interacting with a nanobody (PDBid 5JDS)** targeting the same epitope/surface on the PDL1 GFCC' sheet, using an elaborate FG (CDR3) loop from igs# 8553 to igs# 9547. **(**https://structure.ncbi.nlm.nih.gov/icn3d/share.html?ehLCiHmy953yyGMeA&t=4ZQK,5JDS . **F) PD1/Nivolumab (PDBid: 5WT9) interaction.** Nivolumab binds to a PD1 epitope composed of the FG loop residues as well as an N-terminal loop that is considered outside of the PD1 domain in IgStrand numbering, hence the residues (in cyan)are not numbered and retain their PDB numbers. This points to the need to consider Ig domain extensions. The interface that can be compared to the PD1/PDL1 interface in D) https://structure.ncbi.nlm.nih.gov/icn3d/share.html?UhSpS48KgM7dsMUD8&t=4ZQK,5WT9

Comparison of more complex Ig-chains interfaces

Pseudo-quaternary structures in multi-domain chains

Ig domains can assemble in tandem to form short to long Ig-chains in, mostly, a head to tail arrangement. As mentioned earlier, cell surface receptors containing Ig-domains represent the most populous proteins in the human surfaceome (Verschueren et al. 2020; Bausch-Fluck et al. 2018; Díaz-Ramos, Engel, and Bastos 2011). The most well-known are of course the antibodies with chains composed of an IgV domain at the N terminus followed by one (light chains, TCR chains) or more IgC1 domains (heavy chains) forming both heterotypic (light/heavy chains) or homotypic (heavy/heavy) interfaces for CH domains following CH1 and even in CH1 domains in Sharks' heavy chain only antibodies. There are also a multitude of cell surface receptors in the immune system such as B7 receptors and their ligands, or in the nervous system. Ig-chains can contain a diversity of Ig isotypes, mainly IgV, IgC1, IgC2, IgI, FN3 or as a separate type, cadherins. These *isotypes* tend to compose chains in well defined sequential sets. For example, in vertebrate immune receptors we find numerous chains with an N-terminus IgV followed by one or more IgC1 domains. IgVs can also be followed by C2 domains. Many cell surface receptors also contain a number of FN3 domains, usually after a set of IgV and IgC domains in the membrane proximal regions. The chain lengths encountered can range from a single IgV domain, a couple IgV-IgC domains, or up to 16 or so IgV, IgC, and FN3 domains.

Multiple cell surface adhesion receptors are formed from chains of IgI domains.These chains can be very long, from tens to hundreds of domains, as in Titin (Kenny, Liston, and Higgins 1999). Although we do not have many structures of long chains available in the PDB, the few available exhibit structural plasticity. It is certain that most long chains offer tremendous plasticity with dynamic flexibility in complexes they form, as in the case of obscurin (Mauriello et al. 2023) or the Pom210 chains within the nuclear pore complex (Mosalaganti et al. 2022; Y. Yang et al. 2023). In the numerous cell surface receptors these Ig-domains in sequence can exhibit local rigidity or flexibility in the linkers or "hinges" that connect them and their residue contacts, and give a chain both structure and plasticity. Some domains may form specific Ig-intrachain interfaces, similarly to observed in quaternary Ig-domain assemblies. Long chains therefore can present a **pseudo-quaternary** structure. Chain geometries and flexibilities depend on the length of linkers/hinges and domain contacts. We do not have experimental structures of long Ig-chains but structure predictions of proteomes (Tunyasuvunakool et al. 2021) may possibly give us sequential Ig-domains pairing, since these tertiary intrachain Ig-Ig interfaces have coevolved within a protein chain.

Example of Ig-chains forming intrachain pseudo-quaternary Ig-Ig interfaces

IgV domains tend to assemble using their GFCC' sheets in quaternary interfaces. It is most commonly observed in the PDB as we have seen within antibodies but also in so many cell surface receptors between their N-terminal IgV domains, but this is not the only mode of interaction. The formation of IgV-IgV interfaces depends on the linker length and flexibility between Ig domains. This is obvious as a short linker geometrically forbids chained domain dimerization, long vs. short linkers are commonly used in scFv construction that artificially link IgVH and IgVL domains, where short linkers, between 3 and 5 residues, sometimes called Winter linkers (Holliger, Prospero, and Winter 1993) inhibit intrachain dimer formation to favor interchain dimeric formation between VH and VL domains. Experimentally known natural structures involving IgV domains chained in tandem forming IgV-IgV interfaces are not common in known structures of Ig-chains. Two examples of N-terminal double IgV domains, forming a superdomain with an intrachain IgV-IgV interface resembling quaternary interfaces are presented in Figure 15A for CD226 (H. Wang et al. 2019) and in Figure 15C for Vcbp3 (Hernández Prada et al. 2006).

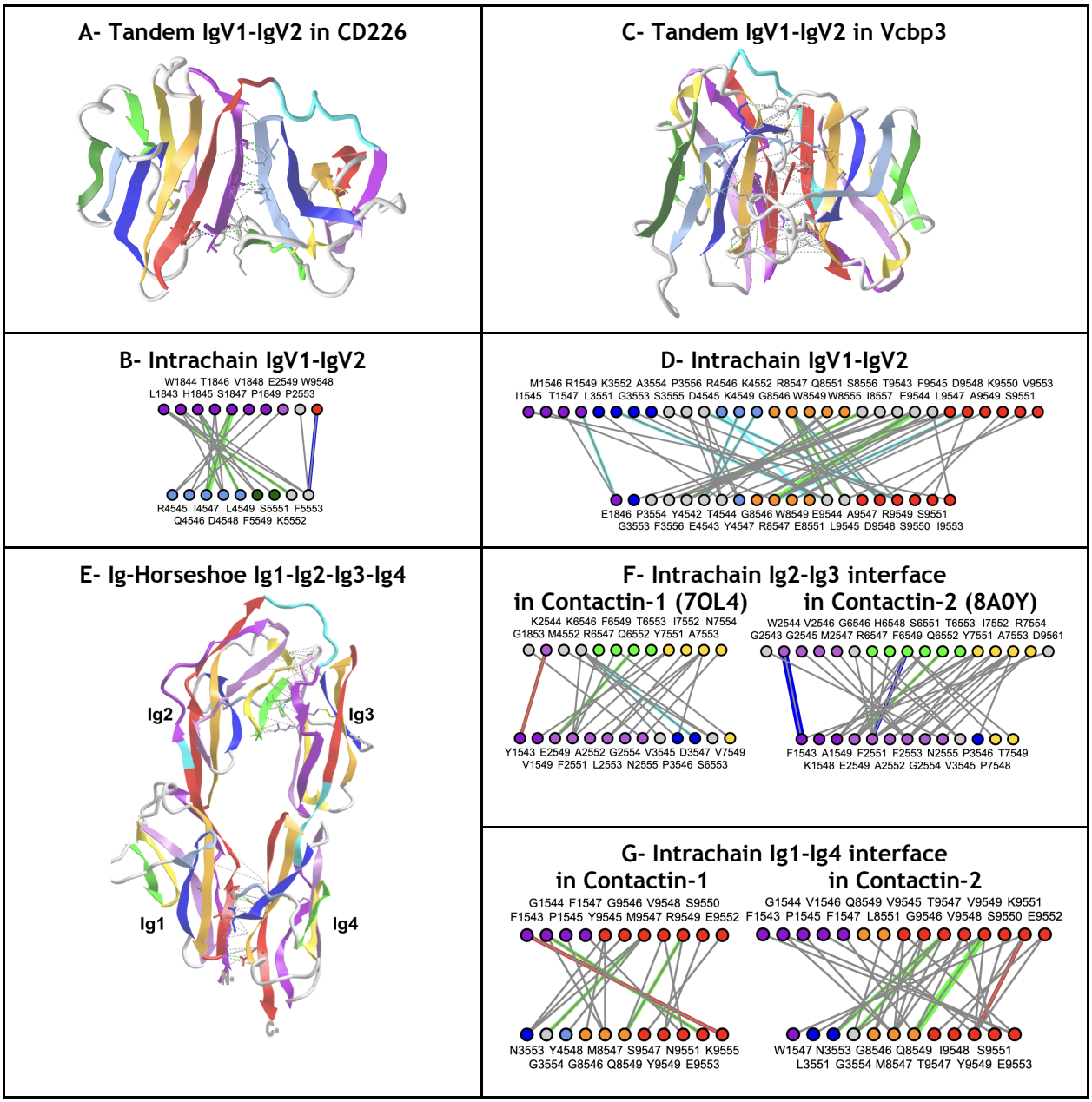


Figure 15 - Multi Ig-domains Intrachain pseudo quaternary interfaces. A-B) Tandem IgV domains in human CD226 (DNAM-1). (PDBid 6ISB) The two IgV domains in tandem with a long linker form an antiparallel interface involving a (tertiary) antiparallel strand zipper between IgV1 (strand A') and IgV2 (strand C') as well as a few residues in the G and C" resp. https://structure.ncbi.nlm.nih.gov/icn3d/share.html?UM2jEBErouevRpGR8&t=6ISB. **C-D) Tandem IgV domains of Vcbp3.** (PDBid 2FBO) The variable region-containing chitin-binding protein-3 (VCBP) is an immune-type molecule found in amphioxus (Branchiostoma floridae). The two IgV domains in tandem form an antiparallel interface involving the GFCC' sheets exhibiting C2 pseudosymmetry, partially resembling an inverted IgV-IgV quaternary interface https://structure.ncbi.nlm.nih.gov/icn3d/share.html?kyVRzncH2ZGzKbQL9&t=2FBO **E-F-G) N-terminal Ig-Horseshoe domain** formed by 4 IgI domains in tandem with a long linker between Ig2 and Ig3 forming an interface Ig2-Ig3 using partially the ABED sheets and an Ig1-Ig4 interface using in part the GFC sheet and the A strand. The 4 domains are highly superimposable (RMSD 4.20/TM-0.77 using TM-Align) when comparing structures of contactin-1 (PDBid 7OL4) and contactin-2 (PDBid 8A0Y) https://structure.ncbi.nlm.nih.gov/icn3d/share.html?iyZJT4PXqFaWnQhP8&t=7OL4,8A0Y.

The four N-terminal Ig domains are arranged in a highly stable horseshoe conformation resulting from an anti-parallel interaction between domains Ig1–Ig4 and Ig2–Ig3, made possible through a long flexible linker between Ig2 and Ig3 (see Figure 15E). When compared across all diverse proteins, the horseshoe superdomain presents a degree of plasticity (see Figure 15E and Figure 15F comparing Contactin-1 to Contactin-2, and Figure S3 comparing to DSCAM). The Ig1-Ig4 domain interface shows a significant degree of structural and interactome conservation and may be the main driver in the horseshoe superdomain folding.

Multi Ig-domains intrachain and interchain quaternary interfaces

Multiple adhesion protein extracellular chains orchestrate cell-cell communication using a variety of Ig domains, FN3 domains, or cadherin domains, all sharing the Ig-fold (Honig and Shapiro 2020; Sanes and Zipursky 2020). For these proteins, cell adhesion equates to Ig-chains binding. Sidekick proteins (Goodman et al. 2016), contactin-2, also called Axonin-1/TAG-1 (Freigang et al. 2000), and other Ig-like molecules form a horseshoe "superdomain" with the four N-terminal Ig domains of the extracellular Ig-chain. These 4-Ig horseshoe superdomains self-interact in either homophilic or heterophilic modes. One mode of quaternary interaction is using a G-strand zipper mechanism between the Ig2 domains, as seen in a structure of contactin-2 in homophilic contact (Chataigner et al. 2024; Sen 2024) and a structure of contactin-1 in heterophilic contact with neurofascin-155 (Chataigner et al. 2022) (Figure 16).

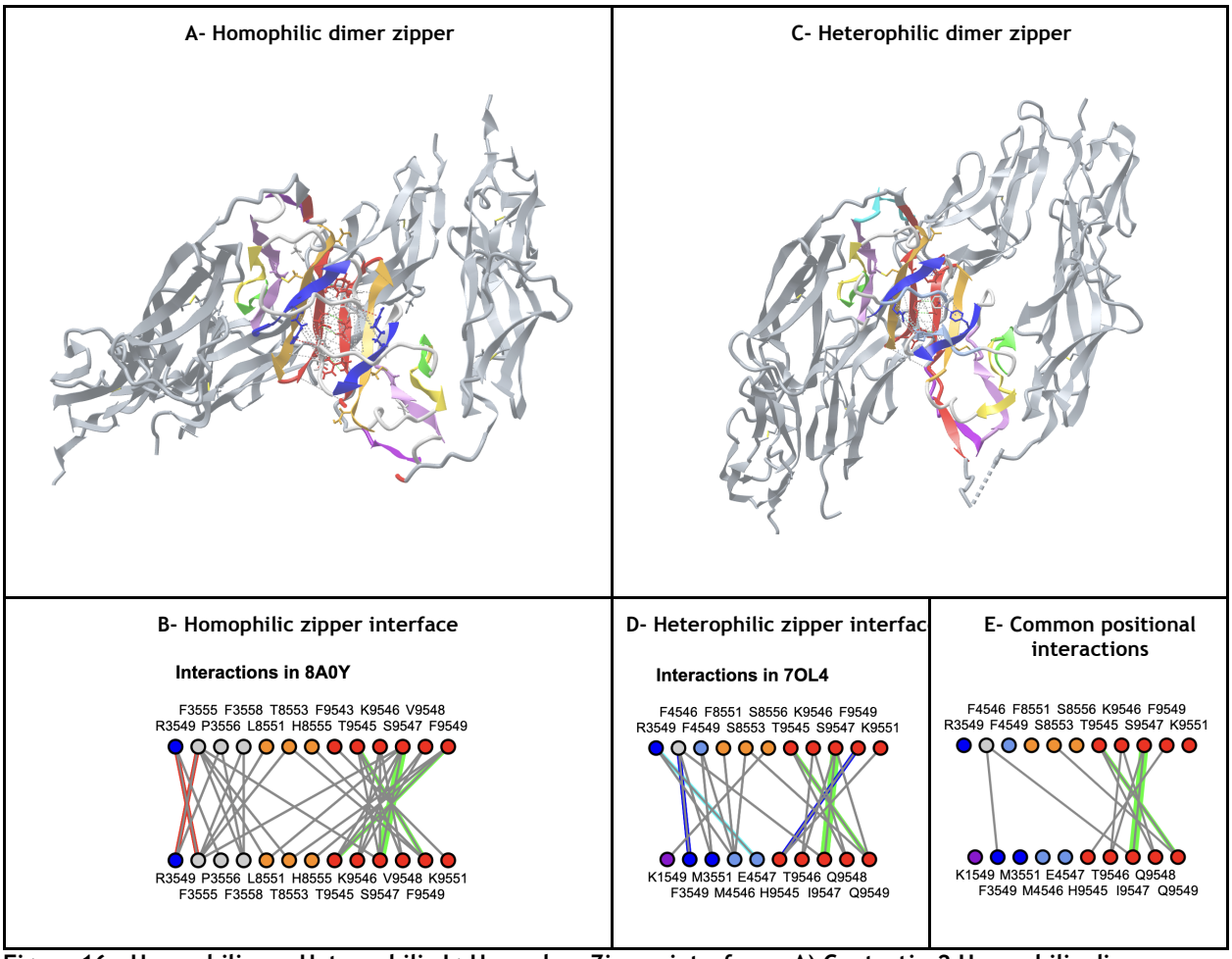


Figure 16 - Homophilic vs. Heterophilic Ig-Horseshoe Zipper interface. A) Contactin-2 Homophilic dimer interface (PDBid 8a0y). Contactin-2 extracellular region contains 6 N-terminal Ig domains followed by 4 FN3 domains. The four N-terminus Ig-domains Ig1-4) form the Ig-Horseshoe. The Ig2 domain forms a dimer through a

G-strand antiparallel zipper interface. **B) Contactin-2 Homophilic Horseshoe Zipper interactome** showing the G-strand zipper (red) as well as residues in the F and C strands of the GFCC' sheet **C) Contactin-1 - neurofascin-155 Heterophilic dimer interface** (PDBid 7OL4). Contactin-1 is highly homologous to Contactin-2 and interacts with **neurofascin-155. Both interact through** their N-terminal Ig-Horseshoe substructure, also using the Ig2 domain forming a G-strand antiparallel zipper interface. The dimer complexes show nonetheless significant plasticity https://structure.ncbi.nlm.nih.gov/icn3d/share.html?ZnVaH3FLKjYL4Eod8&t=7OL4,8A0Y. **D) Contactin-1 - neurofascin-155 Heterophilic Horseshoe Zipper interactome** showing the G-strand zipper (red) as well as residues in the F, C and C strands of the GFCC' sheet. **E) Conserved positional interaction.** The G-strand zipper interface is conserved in terms of its positional residue network. However One should note that this does not mean that all the atom level residue interactions are the same. For example while the symmetric pair igs # 9547-9547 involves a conserved antiparallel backbone-backbone interaction in both the homophilic (S9547-S9547) or heterophilic (S9547-I9547) pair, the symmetric pair T9545-F9549 in the homophilic dimer (PDBid 8A0Y) forms a side chain(T)-backbone(F) HBond, replaced in the heterophilic dimer (PDBid 7OL4) by a backbone-backbone HBond in the T9545-Q9549 pair while the pseudo symmetric H9545-F9549 pair is reduced to a vdW interaction.

Other interfaces besides G-strand zippers also exist between horseshoe superdomains (Chataigner et al. 2024; Sen 2024). The case of Drosophila DSCAM (Down syndrome cell adhesion molecule) is particularly interesting in this regard, as a very large family of immunoglobulin Ig-protein produced through alternative splicing that play an important role in neural development in wiring the nervous system (Zipursky, Wojtowicz, and Hattori 2006; Wojtowicz et al. 2007). DSCAMs use a horseshoe superdomain at the N-terminus (Ig1-4) to form a first half-S shape structure using the double Ig1-Ig4 and Ig2-Ig3 intradomain interfaces, and a second half-S shaped through a third Ig5-Ig6 intrachain interface (Sawaya et al., 2008). This S-shaped structure of the 8 N-terminal Ig-domains gives the chain the ability to self-interact through three domains: Ig2, Ig3 and Ig7 simultaneously. To add to the exquisiteness of the homophilic quaternary interface Ig2-Ig2, Ig3-Ig3, and Ig7-Ig7, Nature orchestrates alternative splicing of exons producing Ig2 and Ig3, and Ig7 domains to specifically recognize self- from non-self and assemble accordingly. The three Ig2-Ig2, Ig3-Ig3 and Ig7-Ig7 domain interfaces form distinct antiparallel Ig-Ig interfaces and lie on their common C2-symmetry axis (Sawaya et al., 2008) (see Figure 17A). Figure 17B, Figure 17C, and Figure 17D dissect the interactomes of the homophilic quaternary recognition of the three domain pairs.

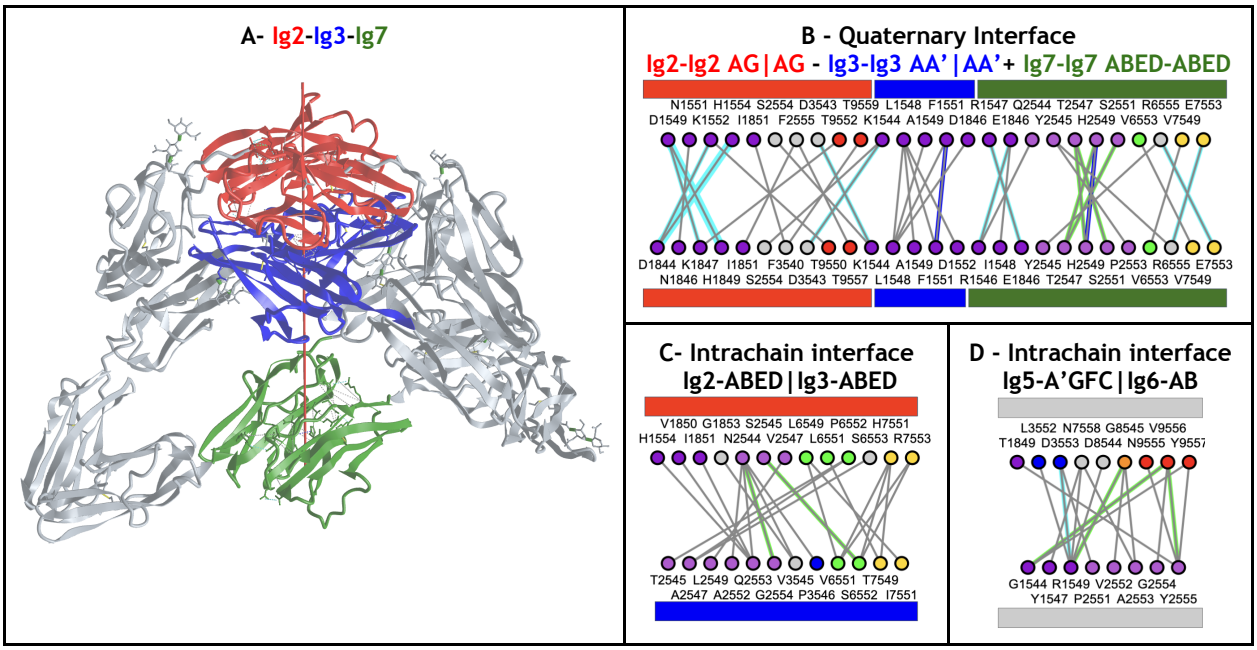


Figure 17 A) https://structure.ncbi.nlm.nih.gov/icn3d/share.html?6swEjWBjXWhufUycA&t=3DMK **. B) Ig2-Ig2, Ig3-Ig3 and Ig7-Ig7 homophilic interfaces** using the AG strands, AA' strands and ABED sheet, respectively. **C) Ig2-Ig3 intrachain Horseshoe interface** using the ABED Sheet. **D) Ig5-Ig6 intrachain Horseshoe interface** using the A'GFC Sheet in Ig6 vs. the A and B strands in Ig6.

METHODS & DATASETS

Ig-domains Nomenclature and Coloring scheme

The classical Ig domain nomenclature defines 7 to 9 strands for IgV/I/C1/C2 domains named alphabetically for the 7-strands constant domain IgC1 and adds C' and C" for 9-strands variable domain IgV found in antibodies (Lesk and Chothia 1982; Chothia and Lesk 1987; A. F. Williams and Barclay 1988), The 7-strand IgC2 domain loses the D strand but gains a C' strand w.r.t the IgC1 domain. This can be also seen as a swap of a strand D in sheet ABED to the GFCC' sheet, while the 8-strands IgI domain possesses both C' and D strands. This is how currently the classical four domains have been described and in the following we keep the strand letter names as nomenclature. We use a 7/9 rainbow color spectrum for visualization of strands, from violet to red (Figure 18) for visualization.

A - Ig Strand color spectrum

A	B	C	C'	C"	D	E	F	G	Loops	Ext.
Dark Violet	Violet	Dark Blue	Light Blue	Dark Green	Light Green	Yellow	Orange	Red	Grey	Cyan
9400D3	BA55D3	0000FF	6495ED	006400	00FF00	FFD700	FF8C00	FF0000	CCCCCC	00FFFF

B - Sequence of Ig Strands colored

C - Ig Protodomain color spectrum

A	B	C	C'	C"	D	E	F	G	Loops	Ext.
Dark Blue	Dark Green	Yellow	Orange	Red	Dark Blue	Dark Green	Yellow	Orange	Grey	Cyan

Figure 18 – Ig Strand color spectrum used for ABCC'C"DEFG strands in Ig-domains. A) Ig Strand extended rainbow color spectrum for 9 strands – 9 colors as indicated. **B) Colored Ig strands according to the Ig Strand Rainbow Spectrum**. Additional and split strands at the N-terminus A',A-,A+,… use the same color as A: dark violet, and similarly additional strands at the C-terminus G+,G++, … use the same color as for G: red. Other inserted strands in Ig-extended domains can appear in white or cyan. **C) Ig Protodomain reduced rainbow color spectrum** 4 strands colors blue, green, yellow, orange for ABCC' and DEFG and a fith color red for the C" strand, if present. Using this color scheme a sheet ABED will be Blue green green blue; a sheet GFCC' will be orange yellow yellow orange. In iCn3D the command used is "color ig strand" and "color ig protodomain" respectively (lowercase). Loops are in Grey between strands. Extensions are in Cyan, and insertions in White. The hexadecimal RGB color codes used in iCn3D are indicated (In this paper, we use the pure yellow code (FFFF00) instead).

Residue Interactions color codes: Interactomes use the following colors for interaction types **Green**: H-Bonds; **Cyan**: Salt Bridge/Ionic; **Grey**: contacts (Van der Waals); **Magenta**: Halogen Bonds; **Red**: π-Cation; **Blue**: π-Stacking.

Contacts are displayed as dotted lines in 3D and traced from Ca to Ca between residues while other types of interactions are atom specific.

iCn3D Ig domain detection

iCn3D (J. Wang et al. 2020, 2022), pronounced "I see in 3D" is an open-source web-based software program to visualize and analyze molecular structures and interactions. The majority of pictures in this paper and all 3D links are from iCn3D. These links allow the 3D visualization of the proteins described in the paper and their interactions. They integrate the software itself so that the reader can himself or herself visualize, analyze, compare the proteins and their interactions in any way desired. The Ig detection, labeling and coloring algorithm using the IgStrand numbering, termed in short "IgStrand algorithm" is briefly described below and can be used on any protein whose structure is known or predicted.

Set up the Ig topostructural variant templates

We use 55 Ig-containing structures as Ig templates to define the strands and loops (see Table 2). The strands in the templates are pairwise aligned using TM-align and the output TM-scores are used to cluster with MEGA11 (Figure S4). We grouped the templates into 16 clusters from the clustering. Each template was assigned as one of the following Ig types: IgV, IgC1, IgC2, IgI, IgE, IgFN3, IgFN3-like, and other Ig. Each Ig-domain template (Table 2) has assigned reference numbers for the strands. Figure 18 shows a simplified Ig strand diagram with the central "anchor" residues on each strand labeled. In each strand, the central "anchor" residue is numbered as ij50. Reference numbers are continuous before and after the anchor, e.g., ij48, ij49, ij50, ij51, ij52, etc., similar to a convention used in GPCRs (Ballesteros and Weinstein 1995). (See main section for details). Each strand is assigned a 1000's number from 1 to 9 for strands A to G, and a color (Figure 18). The conventional strand letter name is also added for convenience in reporting, although redundant with the 1000s digit (see Figure 18 for details). A 100's number is set to 5 for a canonical strand. For example, the anchor of the A strand is #1550, 100's numbers 1-4 and 6-9 are reserved for inserted strands before or after a strand x5xx. For example, A- will be assigned numbers 14xx, A-- 13xx etc., and G+ 96xx, G++ 97xx etc. The A strand can split in IgV and IgI domains: in that case the conventional name A is used for the first half strand on the A-BED sheet with assigned 15xx numbers, and the conventional name A' for the second half strand on the A'GFCC'C" with assigned 18xx numbers.

Detect Ig domains

We detect Ig domains using structural alignment with the TM-align program (Zhang and Skolnick 2005). We first split each chain of the input structure into 3D domains, each of which then goes through two steps of alignment to find the best template. The first step is to align the 3D domain with each representative of the 16 clusters to find the top five best alignments. If the best alignment has a TM score larger than 0.85, the templates in that cluster are chosen for the second step alignment. Otherwise, all templates in the top five clusters are chosen. The second step is to align the 3D domain with the chosen templates to find the template with the highest

TM score. If the highest TM score is larger than 0.4, an Ig domain is detected in the 3D domain. The two-step alignment is more efficient than the one-step alignment to align each of the 55 templates since the number of total templates may increase in the future. After an Ig domain is detected in a protein chain, the rest of residues in the protein chain iteratively detect additional Ig domains, until no more Ig domains are detected.

Assign Ig reference numbers

When the detected Ig domain is aligned with the best template, key residues, i.e. “anchors” (see above), are matched to the template. The matched anchor residues are assigned with the reference numbers ij50 (e.g., 1550 for the anchor residue of strand A). The reference numbers in the beta-sheet are then numbered continuously before and after the anchor residue. For coils/loops between two Ig strands, the residues in the first half of the coil are assigned reference numbers continuously from the preceding strand, residues in the second half of the coil are assigned reference numbers continuously to the following strand. If a residue is not in an Ig domain, its reference number is undefined and it will be colored in cyan. Strand insertions are numbered continuously. A strand anchoring mechanism for inserted strands in Ig-extended domains is being developed for future releases.

How to use iCn3D to assign Ig reference numbers

Once you load structures into iCn3D at https://www.ncbi.nlm.nih.gov/Structure/icn3d/, you can assign IgStrand reference numbers via the menu "Analysis > Ref. Number > Show Ig for Selection", or via the menu "Analysis > Seq. & Annotations" and click the checkbox "Ig Domains" in the "Annotations" section. The reference numbers can be exported via the menu "File > Save File > Reference Numbers".
If you want to process a list of structures, you can download iCn3D Node.js script "refnum.js" at https://github.com/ncbi/icn3d/tree/master/icn3dnode. This script runs TM-align locally using the program at https://github.com/ncbi/icn3d/tree/master/icn3dnode/tmalign-icn3dnode and the templates at https://github.com/ncbi/icn3d/tree/master/icn3dnode/refpdb.

CONCLUSION

The importance of the Ig-fold cannot be overstated. The first human genome project identified the Ig-fold as the most populous in human genes. The proteins coded by these genes can be found in any region of the cell from the nucleus to the cytoplasm to the extracellular regions of transmembrane proteins. Indeed, a very large number of genes code for cell surface proteins involved in cell adhesion (CAMs) and the development and function of the nervous system, the immune system, the vascular system and the muscular system. This systemic aspect is of particular interest: the Ig-fold in its diverse variants is a **self-interactor** and, precisely, these “systems” use Ig-domains as essential molecular building blocks for bilateral cell recognition with high precision, synapse formation, and communication.

The IgStrand numbering scheme enables **positional comparisons** at the tertiary and quaternary level. Any position in an Ig- domain involved in folding or protein binding can be analyzed in multiple contexts. The IgStrand universal numbering scheme can therefore be used in **comparative analysis** of protein domains exhibiting an Ig-fold and protein chains containing multiple Ig-domains (Ig-chains) as well as Ig-Ig

domain interfaces, at any scale. It is suited to study the residues responsible for forming the fold to better understand folding of Ig-domains across the protein universe exhibiting a beta sandwich architecture. A preliminary survey of the human structural proteome using the IgStrand algorithm, considering Ig-, Ig-like and Ig-extended domains, leads to estimates ca. 10% of human genes contain Ig-folded proteins rather than 2-3% as originally estimated (to be published). The study of the Ig-proteomes across the increasing number of related species' proteomes may also lead to a better understanding of the evolution of genes containing the Ig-fold, especially in the development of the nervous system and immune system.

The formation of synapses through structurally similar cell surface receptors and ligands points to molecular recognition surfaces encoded in Ig-domains surface residues. The IgStrand scheme can be used to precisely study residues facing out and possibly decode Ig-Ig interactomes. As mentioned above, an initial survey of the PDB database leads to a dataset of over 25,000 Ig-Ig domain interfaces that can now all be analyzed in parallel to better understand both the folds and their molecular interactions, towards an understanding of co-evolution of receptors ligand pairs and their codes on recognition (work in progress).

We expect the parallel analysis of protein families containing Ig-, Ig-like, or Ig-extended domains to reveal potential clues in biology, not just on evolution that led to complex multicellular organisms with an immune system, a nervous system, a vascular system, and a muscular system, but on protein folding. Beyond research on evolution, folding, and molecular assemblies involving the Ig-fold, the IgStrand scheme can be used in assisting protein engineering, and we can foresee a number of direct applications for example in antibody and single Ig-domain (nanobody) design, and CAR T-cell therapies, especially when when both target binding moieties and antigen targets themselves consist of Ig-domains or Ig-chains. Indeed, it is important to realize that a multitude of antigen targets in immunology are extracellular Ig-chains with one, two (CD19) or more (CD22) of single pass transmembrane receptors (Naeim et al. 2018; Moreno-Cortes et al. 2021).

We have implemented the IgStrand numbering scheme in iCn3D (J. Wang et al. 2020) to enable interactive analysis. An Ig-fold recognition algorithm based on IgStrand enables the accurate identification of Ig-domains across any protein structure. We refer to it in short as the IgStrand algorithm. The software can also be used in batch mode (J. Wang et al. 2022), enabling large datasets analyses. As mentioned above an initial survey of over 20,000 genes and 82,518 proteins of the structural human proteome (Acids research and 2021 2021; Tunyasuvunakool et al. 2021) indicates the presence of the Ig-fold in a much larger proportion than the original estimates. Improvements are still needed to accurately label each and every type of topostructural variant. Once we reach that stage, we can envision the study of evolution of all Ig-domains and proteins across structural proteomes.

The Ig-fold is unique in its ability to sustain topological variations and we have focused on developing a universal numbering scheme that can likewise sustain these variations. We deliberately chose to make the IgStrand numbering human as well as machine readable as interactive analysis is central to our approach. In the future we envision extensions of IgStrand with a hierarchical numbering, to manage long Ig-chains and their interactions, as well as explicit alternative positioning of strands in beta-sandwich sheets to extend the nomenclature to jelly rolls and other beta-sandwich topologies. Finally, loop regions present challenges in universal numbering that need further innovation. The IgStrand scheme was a necessary step towards positional structural bioinformatics of Ig-proteomes and Ig-interactomes.

other important folds would benefit from the development of their own universal reference numbering, while offering less topological challenges, to provide a means to full scale positional structural bioinformatics.

ACKNOWLEDGMENTS

This work was supported in part by the Center for Cancer Research at the National Cancer Institute; the National Center for Biotechnology Information of the National Library of Medicine (NLM), National Institutes of Health. The authors thank Dr. Aron Marchler-Bauer for useful discussions.

REFERENCES

Abhinandan, K. R., and Andrew C. R. Martin. 2008. "Analysis and Improvements to Kabat and Structurally Correct Numbering of Antibody Variable Domains." *Molecular Immunology* 45 (14): 3832–39. https://doi.org/10.1016/j.molimm.2008.05.022.

Acids research, Nucleic, and 2021. 2021. "UniProt: The Universal Protein Knowledgebase in 2021." *Nucleic Acids Research* 49 (D1): D480–89. https://academic.oup.com/nar/article-abstract/49/D1/D480/6006196.

Afzali, Ali Maisam, Thomas Müntefering, Heinz Wiendl, Sven G. Meuth, and Tobias Ruck. 2018. "Skeletal Muscle Cells Actively Shape (auto)immune Responses." *Autoimmunity Reviews* 17 (5): 518–29. https://doi.org/10.1016/j.autrev.2017.12.005.

Ahn, Jinsook, Jinwook Lee, Soyeon Jeong, So-Mi Kang, Bum-Joon Park, and Nam-Chul Ha. 2021. "Beta-Strand-Mediated Dimeric Formation of the Ig-like Domains of Human Lamin A/C and B1." *Biochemical and Biophysical Research Communications* 550 (April): 191–96. https://doi.org/10.1016/j.bbrc.2021.02.102.

Alzari, Pedro M. 1998. "Domains, Immunoglobulin-Type." In *Encyclopedia of Immunology (Second Edition)*, edited by Peter J. Delves, 775–78. Oxford: Elsevier. https://doi.org/10.1006/rwei.1999.0204.

Alzari, P. M., M. B. Lascombe, and R. J. Poljak. 1988. "Three-Dimensional Structure of Antibodies." *Annual Review of Immunology* 6: 555–80. https://doi.org/10.1146/annurev.iy.06.040188.003011.

Aubry, Laurence, Dorian Guetta, and Gérard Klein. 2009. "The Arrestin Fold: Variations on a Theme." *Current Genomics* 10 (2): 133–42. https://doi.org/10.2174/138920209787847014.

Ballesteros, Juan A., and Harel Weinstein. 1995. "[19] Integrated Methods for the Construction of Three-Dimensional Models and Computational Probing of Structure-Function Relations in G Protein-Coupled Receptors." In *Methods in Neurosciences*, edited by Stuart C. Sealfon, 25:366–428. Academic Press. https://doi.org/10.1016/S1043-9471(05)80049-7.

Barnett, Anna L., David L. Wensel, Weihua Li, Deborah Fass, and James M. Cunningham. 2003. "Structure and Mechanism of a Coreceptor for Infection by a Pathogenic Feline Retrovirus." *Journal of Virology* 77 (4): 2717–29. https://doi.org/10.1128/jvi.77.4.2717-2729.2003.

Bateman, A., S. R. Eddy, and C. Chothia. 1996. "Members of the Immunoglobulin Superfamily in Bacteria." *Protein Science: A Publication of the Protein Society* 5 (9): 1939–41. https://doi.org/10.1002/pro.5560050923.

Bausch-Fluck, Damaris, Ulrich Goldmann, Sebastian Müller, Marc van Oostrum, Maik Müller, Olga T. Schubert, and Bernd Wollscheid. 2018. "The in Silico Human Surfaceome." *Proceedings of the National Academy of Sciences of the United States of America*, October. https://doi.org/10.1073/pnas.1808790115.

Benian, Guy M., and Olga Mayans. 2015. "Titin and Obscurin: Giants Holding Hands and Discovery of a New Ig Domain Subset." *Journal of Molecular Biology*.

https://doi.org/10.1016/j.jmb.2014.12.017.

Boni, Nadir, Lawrence Shapiro, Barry Honig, Yinghao Wu, and Rotem Rubinstein. 2022. "On the Formation of Ordered Protein Assemblies in Cell-Cell Interfaces." *Proceedings of the National Academy of Sciences of the United States of America* 119 (34): e2206175119. https://doi.org/10.1073/pnas.2206175119.

Bork, P., L. Holm, and C. Sander. 1994. "The Immunoglobulin Fold. Structural Classification, Sequence Patterns and Common Core." *Journal of Molecular Biology* 242 (4): 309–20. https://doi.org/10.1006/jmbi.1994.1582.

Brasch, Julia, Oliver J. Harrison, Barry Honig, and Lawrence Shapiro. 2012. "Thinking Outside the Cell: How Cadherins Drive Adhesion." *Trends in Cell Biology* 22 (6): 299–310. https://doi.org/10.1016/j.tcb.2012.03.004.

Brumshtein, Boris, Shannon R. Esswein, Michael R. Sawaya, Gregory Rosenberg, Alan T. Ly, Meytal Landau, and David S. Eisenberg. 2018. "Identification of Two Principal Amyloid-Driving Segments in Variable Domains of Ig Light Chains in Systemic Light-Chain Amyloidosis." *The Journal of Biological Chemistry* 293 (51): 19659–71. https://doi.org/10.1074/jbc.RA118.004142.

Chandonia, John-Marc, Lindsey Guan, Shiangyi Lin, Changhua Yu, Naomi K. Fox, and Steven E. Brenner. 2022. "SCOPe: Improvements to the Structural Classification of Proteins - Extended Database to Facilitate Variant Interpretation and Machine Learning." *Nucleic Acids Research* 50 (D1): D553–59. https://doi.org/10.1093/nar/gkab1054.

Chataigner, Lucas M. P., Christos Gogou, Maurits A. den Boer, Cátia P. Frias, Dominique M. E. Thies-Weesie, Joke C. M. Granneman, Albert J. R. Heck, Dimphna H. Meijer, and Bert J. C. Janssen. 2022. "Structural Insights into the Contactin 1 - Neurofascin 155 Adhesion Complex." *Nature Communications* 13 (1): 6607. https://doi.org/10.1038/s41467-022-34302-9.

Chataigner, Lucas M. P., Lena Thärichen, J. Wouter Beugelink, Joke C. M. Granneman, Nadia J. Mokiem, Joost Snijder, Friedrich Förster, and Bert J. C. Janssen. 2024. "Contactin 2 Homophilic Adhesion Structure and Conformational Plasticity." *Structure* 32 (1): 60–73.e5. https://doi.org/10.1016/j.str.2023.10.012.

Cheng, Hua, Yuxing Liao, R. Dustin Schaeffer, and Nick V. Grishin. 2015. "Manual Classification Strategies in the ECOD Database." *Proteins* 83 (7): 1238–51. https://doi.org/10.1002/prot.24818.

Cheng, Hua, R. Dustin Schaeffer, Yuxing Liao, Lisa N. Kinch, Jimin Pei, Shuoyong Shi, Bong-Hyun Kim, and Nick V. Grishin. 2014. "ECOD: An Evolutionary Classification of Protein Domains." *PLoS Computational Biology* 10 (12): e1003926. https://doi.org/10.1371/journal.pcbi.1003926.

Cheng, Shouqiang, James Ashley, Justyna D. Kurleto, Meike Lobb-Rabe, Yeonhee Jenny Park, Robert A. Carrillo, and Engin Özkan. 2019. "Molecular Basis of Synaptic Specificity by Immunoglobulin Superfamily Receptors in Drosophila." *eLife* 8 (January). https://doi.org/10.7554/eLife.41028.

Cheng, Shouqiang, Yeonwoo Park, Justyna D. Kurleto, Mili Jeon, Kai Zinn, Joseph W. Thornton, and Engin Özkan. 2019. "Family of Neural Wiring Receptors in Bilaterians Defined by Phylogenetic, Biochemical, and Structural Evidence." *Proceedings of the National Academy of Sciences of the United States of America* 116 (20): 9837–42. https://doi.org/10.1073/pnas.1818631116.

Chothia, C., and A. M. Lesk. 1987. "Canonical Structures for the Hypervariable Regions of Immunoglobulins." *Journal of Molecular Biology* 196 (4): 901–17. https://www.ncbi.nlm.nih.gov/pubmed/3681981.

Chothia, C., A. M. Lesk, M. Levitt, A. G. Amit, R. A. Mariuzza, S. E. Phillips, and R. J. Poljak. 1986. "The Predicted Structure of Immunoglobulin D1.3 and Its Comparison with the Crystal Structure."

Science 233 (4765): 755–58. https://doi.org/10.1126/science.3090684.

Cooper, Max D., and Matthew N. Alder. 2006. "The Evolution of Adaptive Immune Systems." *Cell* 124 (4): 815–22. https://doi.org/10.1016/j.cell.2006.02.001.

Cortés, Elena, Joseph S. Pak, and Engin Özkan. 2023. "Structure and Evolution of Neuronal Wiring Receptors and Ligands." *Developmental Dynamics: An Official Publication of the American Association of Anatomists* 252 (1): 27–60. https://doi.org/10.1002/dvdy.512.

Dantzer, Robert. 2018. "Neuroimmune Interactions: From the Brain to the Immune System and Vice Versa." *Physiological Reviews* 98 (1): 477–504. https://doi.org/10.1152/physrev.00039.2016.

Deuss, Felix A., Gabrielle M. Watson, Zhihui Fu, Jamie Rossjohn, and Richard Berry. 2018. "Structural Basis for CD96 Immune Receptor Recognition of Nectin-like Protein-5, CD155." *Structure* , November. https://doi.org/10.1016/j.str.2018.10.023.

Díaz-Ramos, M. Carmen, Pablo Engel, and Ricardo Bastos. 2011. "Towards a Comprehensive Human Cell-Surface Immunome Database." *Immunology Letters* 134 (2): 183–87. https://doi.org/10.1016/j.imlet.2010.09.016.

Dittmer, Travis A., and Tom Misteli. 2011. "The Lamin Protein Family." *Genome Biology* 12 (5): 222. https://doi.org/10.1186/gb-2011-12-5-222.

Dittmer, Travis A., Nidhi Sahni, Nard Kubben, David E. Hill, Marc Vidal, Rebecca C. Burgess, Vassilis Roukos, and Tom Misteli. 2014. "Systematic Identification of Pathological Lamin A Interactors." *Molecular Biology of the Cell* 25 (9): 1493–1510. https://doi.org/10.1091/mbc.E14-02-0733.

Dunbar, James, and Charlotte M. Deane. 2016. "ANARCI: Antigen Receptor Numbering and Receptor Classification." *Bioinformatics* 32 (2): 298–300. https://doi.org/10.1093/bioinformatics/btv552.

Du Pasquier, L. 2000. "The Phylogenetic Origin of Antigen-Specific Receptors." In *Origin and Evolution of the Vertebrate Immune System*, edited by Louis Du Pasquier and Gary W. Litman, 159–85. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-59674-2_8.

Du Pasquier, Louis. 2004. "Speculations on the Origin of the Vertebrate Immune System." *Immunology Letters* 92 (1-2): 3–9. https://doi.org/10.1016/j.imlet.2003.10.012.

Du Pasquier, Louis, Ivana Zucchetti, and Rosaria De Santis. 2004. "Immunoglobulin Superfamily Receptors in Protochordates: Before RAG Time." *Immunological Reviews* 198 (April): 233–48. https://doi.org/10.1111/j.0105-2896.2004.00122.x.

Eckels, Edward C., Rafael Tapia-Rojo, Jamie Andrés Rivas-Pardo, and Julio M. Fernández. 2018. "The Work of Titin Protein Folding as a Major Driver in Muscle Contraction." *Annual Review of Physiology* 80 (February): 327–51. https://doi.org/10.1146/annurev-physiol-021317-121254.

Edwards, Hannah, Sanne Abeln, and Charlotte M. Deane. 2013. "Exploring Fold Space Preferences of New-Born and Ancient Protein Superfamilies." *PLoS Computational Biology* 9 (11): e1003325. https://doi.org/10.1371/journal.pcbi.1003325.

Elenkov, Ilia J., Ronald L. Wilder, George P. Chrousos, and E. Sylvester Vizi. 2000. "The Sympathetic Nerve—An Integrative Interface between Two Supersystems: The Brain and the Immune System." *Pharmacological Reviews* 52 (4): 595–638. https://pharmrev.aspetjournals.org/content/52/4/595.full?casa_token=DMn-o8sffZMAAAAA:5QDq4ZSsAbDj6sPPNgwMXMixO1BSZb24HlOUb4GKH6MqCBxuEXDlYz8v-4HdnE1ewGBwaUc0w-w.

Epp, O., P. Colman, H. Fehlhammer, W. Bode, M. Schiffer, R. Huber, and W. Palm. 1974. "Crystal and

Molecular Structure of a Dimer Composed of the Variable Portions of the Bence-Jones Protein REI." *European Journal of Biochemistry / FEBS* 45 (2): 513–24. https://www.ncbi.nlm.nih.gov/pubmed/4851220.

Epp, O., E. E. Lattman, M. Schiffer, R. Huber, and W. Palm. 1975. "The Molecular Structure of a Dimer Composed of the Variable Portions of the Bence-Jones Protein REI Refined at 2.0-A Resolution." *Biochemistry* 14 (22): 4943–52. https://doi.org/10.1021/bi00693a025.

Farré, Domènec, Pablo Martínez-Vicente, Pablo Engel, and Ana Angulo. 2017. "Immunoglobulin Superfamily Members Encoded by Viruses and Their Multiple Roles in Immune Evasion." *European Journal of Immunology* 47 (5): 780–96. https://doi.org/10.1002/eji.201746984.

Fass, D., R. A. Davey, C. A. Hamson, P. S. Kim, J. M. Cunningham, and J. M. Berger. 1997. "Structure of a Murine Leukemia Virus Receptor-Binding Glycoprotein at 2.0 Angstrom Resolution." *Science* 277 (5332): 1662–66. https://doi.org/10.1126/science.277.5332.1662.

Fernández-Quintero, Monica L., Katharina B. Kroell, Lisa M. Bacher, Johannes R. Loeffler, Patrick K. Quoika, Guy Georges, Alexander Bujotzek, Hubert Kettenberger, and Klaus R. Liedl. 2021. "Germline-Dependent Antibody Paratope States and Pairing Specific VH-VL Interface Dynamics." *Frontiers in Immunology* 12 (August): 675655. https://doi.org/10.3389/fimmu.2021.675655.

Fernández-Quintero, Monica L., Katharina B. Kroell, Martin C. Heiss, Johannes R. Loeffler, Patrick K. Quoika, Franz Waibl, Alexander Bujotzek, Ekkehard Moessner, Guy Georges, and Klaus R. Liedl. 2020. "Surprisingly Fast Interface and Elbow Angle Dynamics of Antigen-Binding Fragments." *Frontiers in Molecular Biosciences* 7 (November): 609088. https://doi.org/10.3389/fmolb.2020.609088.

Fernández-Quintero, Monica L., Nancy D. Pomarici, Anna-Lena M. Fischer, Valentin J. Hoerschinger, Katharina B. Kroell, Jakob R. Riccabona, Anna S. Kamenik, et al. 2023. "Structure and Dynamics Guiding Design of Antibody Therapeutics and Vaccines." *Antibodies* 12 (4): 67. https://doi.org/10.3390/antib12040067.

Fernández-Quintero, Monica L., Patrick K. Quoika, Florian S. Wedl, Clarissa A. Seidler, Katharina B. Kroell, Johannes R. Loeffler, Nancy D. Pomarici, et al. 2022. "Comparing Antibody Interfaces to Inform Rational Design of New Antibody Formats." *Frontiers in Molecular Biosciences* 9 (January): 812750. https://doi.org/10.3389/fmolb.2022.812750.

Fraser, James S., Zhou Yu, Karen L. Maxwell, and Alan R. Davidson. 2006. "Ig-like Domains on Bacteriophages: A Tale of Promiscuity and Deceit." *Journal of Molecular Biology* 359 (2): 496–507. https://doi.org/10.1016/j.jmb.2006.03.043.

Freigang, Jörg, Karl Proba, Lukas Leder, Kay Diederichs, Peter Sonderegger, and Wolfram Welte. 2000. "The Crystal Structure of the Ligand Binding Module of Axonin-1/TAG-1 Suggests a Zipper Mechanism for Neural Cell Adhesion." *Cell* 101 (4): 425–33. https://doi.org/10.1016/S0092-8674(00)80852-1.

Giganti, David, Kevin Yan, Carmen L. Badilla, Julio M. Fernandez, and Jorge Alegre-Cebollada. 2018. "Disulfide Isomerization Reactions in Titin Immunoglobulin Domains Enable a Mode of Protein Elasticity." *Nature Communications* 9 (1): 185. https://doi.org/10.1038/s41467-017-02528-7.

Goodman, Kerry M., Masahito Yamagata, Xiangshu Jin, Seetha Mannepalli, Phinikoula S. Katsamba, Göran Ahlsén, Alina P. Sergeeva, Barry Honig, Joshua R. Sanes, and Lawrence Shapiro. 2016. "Molecular Basis of Sidekick-Mediated Cell-Cell Adhesion and Specificity." *eLife* 5 (September). https://doi.org/10.7554/eLife.19058.

Halaby, D. M., and J. P. Mornon. 1998. "The Immunoglobulin Superfamily: An Insight on Its Tissular, Species, and Functional Diversity." *Journal of Molecular Evolution* 46 (4): 389–400.

https://doi.org/10.1007/pl00006318.

Hamill, S. J., E. Cota, C. Chothia, and J. Clarke. 2000. "Conservation of Folding and Stability within a Protein Family: The Tyrosine Corner as an Evolutionary Cul-de-Sac." *Journal of Molecular Biology* 295 (3): 641–49. https://doi.org/10.1006/jmbi.1999.3360.

Harpaz, Y., and C. Chothia. 1994. "Many of the Immunoglobulin Superfamily Domains in Cell Adhesion Molecules and Surface Receptors Belong to a New Structural Set Which Is close to That Containing Variable Domains." *Journal of Molecular Biology* 238 (4): 528–39. https://doi.org/10.1006/jmbi.1994.1312.

Harrison, Oliver J., Jeremie Vendome, Julia Brasch, Xiangshu Jin, Soonjin Hong, Phinikoula S. Katsamba, Goran Ahlsen, et al. 2012. "Nectin Ectodomain Structures Reveal a Canonical Adhesive Interface." *Nature Structural & Molecular Biology* 19 (9): 906–15. https://doi.org/10.1038/nsmb.2366.

Hemmingsen, J. M., K. M. Gernert, J. S. Richardson, and D. C. Richardson. 1994. "The Tyrosine Corner: A Feature of Most Greek Key Beta-Barrel Proteins." *Protein Science: A Publication of the Protein Society* 3 (11): 1927–37. https://doi.org/10.1002/pro.5560031104.

Hernández Prada, José A., Robert N. Haire, Marc Allaire, Jean Jakoncic, Vivian Stojanoff, John P. Cannon, Gary W. Litman, and David A. Ostrov. 2006. "Ancient Evolutionary Origin of Diversified Variable Regions Demonstrated by Crystal Structures of an Immune-Type Receptor in Amphioxus." *Nature Immunology* 7 (8): 875–82. https://doi.org/10.1038/ni1359.

Hogan, Victoria, and Welkin E. Johnson. 2023. "Unique Structure and Distinctive Properties of the Ancient and Ubiquitous Gamma-Type Envelope Glycoprotein." *Viruses* 15 (2). https://doi.org/10.3390/v15020274.

Holliger, P., T. Prospero, and G. Winter. 1993. "'Diabodies': Small Bivalent and Bispecific Antibody Fragments." *Proceedings of the National Academy of Sciences of the United States of America* 90 (14): 6444–48. https://doi.org/10.1073/pnas.90.14.6444.

Honegger, A., and A. Plückthun. 2001. "Yet Another Numbering Scheme for Immunoglobulin Variable Domains: An Automatic Modeling and Analysis Tool." *Journal of Molecular Biology* 309 (3): 657–70. https://doi.org/10.1006/jmbi.2001.4662.

Honig, Barry, and Lawrence Shapiro. 2020. "Adhesion Protein Structure, Molecular Affinities, and Principles of Cell-Cell Recognition." *Cell* 181 (3): 520–35. https://doi.org/10.1016/j.cell.2020.04.010.

Ioerger, T. R., C. Du, and D. S. Linthicum. 1999. "Conservation of Cys-Cys Trp Structural Triads and Their Geometry in the Protein Domains of Immunoglobulin Superfamily Members." *Molecular Immunology* 36 (6): 373–86. https://doi.org/10.1016/s0161-5890(99)00032-2.

Janes, Morgan E., Allison Kinlein, Martin F. Flajnik, Louis Du Pasquier, and Yuko Ohta. 2023. "Genomic View of the Origins of Cell-Mediated Immunity." *Immunogenetics* 75 (6): 479–93. https://doi.org/10.1007/s00251-023-01319-3.

Johnson, G., and T. T. Wu. 2000. "Kabat Database and Its Applications: 30 Years after the First Variability Plot." *Nucleic Acids Research* 28 (1): 214–18. https://doi.org/10.1093/nar/28.1.214.

Kabat, E. A., T. T. Wu, and H. Bilofsky. 1976. "Attempts to Locate Residues in Complementarity-Determining Regions of Antibody Combining Sites That Make Contact with Antigen." *Proceedings of the National Academy of Sciences of the United States of America* 73 (2): 617–19. https://doi.org/10.1073/pnas.73.2.617.

Kabat, Elvin A. 1992. *Sequences of Proteins of Immunological Interest*. DIANE Publishing. https://play.google.com/store/books/details?id=3jMvZYW2ZtwC.

Kabat, and Others. 1968. "Structural Concepts in Immunology and Immunochemistry." *Structural Concepts in Immunology and Immunochemistry*. https://www.cabdirect.org/cabdirect/abstract/19690404395.

Kenny, P. A., E. M. Liston, and D. G. Higgins. 1999. "Molecular Evolution of Immunoglobulin and Fibronectin Domains in Titin and Related Muscle Proteins." *Gene* 232 (1): 11–23. https://doi.org/10.1016/s0378-1119(99)00122-5.

Krimm, Isabelle, Cecilia Ostlund, Bernard Gilquin, Joël Couprie, Paul Hossenlopp, Jean-Paul Mornon, Gisèle Bonne, Jean-Claude Courvalin, Howard J. Worman, and Sophie Zinn-Justin. 2002. "The Ig-like Structure of the C-Terminal Domain of Lamin A/C, Mutated in Muscular Dystrophies, Cardiomyopathy, and Partial Lipodystrophy." *Structure* 10 (6): 811–23. https://doi.org/10.1016/s0969-2126(02)00777-3.

Kulahin, Nikolaj, Ole Kristensen, Kim K. Rasmussen, Lars Olsen, Patrik Rydberg, Bente Vestergaard, Jette S. Kastrup, et al. 2011. "Structural Model and Trans-Interaction of the Entire Ectodomain of the Olfactory Cell Adhesion Molecule." *Structure* 19 (2): 203–11. https://doi.org/10.1016/j.str.2010.12.014.

Kushner, Artem, Anton S. Petrov, and Khanh Dao Duc. 2023. "RiboXYZ: A Comprehensive Database for Visualizing and Analyzing Ribosome Structures." *Nucleic Acids Research* 51 (D1): D509–16. https://doi.org/10.1093/nar/gkac939.

Labeit, S., and B. Kolmerer. 1995. "Titins: Giant Proteins in Charge of Muscle Ultrastructure and Elasticity." *Science* 270 (5234): 293–96. https://doi.org/10.1126/science.270.5234.293.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, et al. 2001. "Initial Sequencing and Analysis of the Human Genome." *Nature* 409 (6822): 860–921. https://doi.org/10.1038/35057062.

Larrivée, Bruno, Catarina Freitas, Steven Suchting, Isabelle Brunet, and Anne Eichmann. 2009. "Guidance of Vascular Development: Lessons from the Nervous System." *Circulation Research* 104 (4): 428–41. https://doi.org/10.1161/CIRCRESAHA.108.188144.

Lau, A. M., N. Bordin, S. M. Kandathil, I. Sillitoe, V. P. Waman, J. Wells, C. A. Orengo, and D. T. Jones. 2024. "Exploring Structural Diversity across the Protein Universe with The Encyclopedia of Domains." *bioRxiv*. https://doi.org/10.1101/2024.03.18.585509.

Lefranc, Marie-Paule. 2011. "IMGT Unique Numbering for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF." *Cold Spring Harbor Protocols* 2011 (6): 633–42. https://doi.org/10.1101/pdb.ip85.

Lefranc, Marie-Paule, Christelle Pommié, Quentin Kaas, Elodie Duprat, Nathalie Bosc, Delphine Guiraudou, Christelle Jean, et al. 2005. "IMGT Unique Numbering for Immunoglobulin and T Cell Receptor Constant Domains and Ig Superfamily C-like Domains." *Developmental and Comparative Immunology* 29 (3): 185–203. https://doi.org/10.1016/j.dci.2004.07.003.

Lefranc, Marie-Paule, Christelle Pommié, Manuel Ruiz, Véronique Giudicelli, Elodie Foulquier, Lisa Truong, Valérie Thouvenin-Contet, and Gérard Lefranc. 2003. "IMGT Unique Numbering for Immunoglobulin and T Cell Receptor Variable Domains and Ig Superfamily V-like Domains." *Developmental and Comparative Immunology* 27 (1): 55–77. https://www.ncbi.nlm.nih.gov/pubmed/12477501.

Lesk, A. M., and C. Chothia. 1982. "Evolution of Proteins Formed by Beta-Sheets. II. The Core of the

Immunoglobulin Domains." *Journal of Molecular Biology* 160 (2): 325–42. https://doi.org/10.1016/0022-2836(82)90179-6.

———. 1988. "Elbow Motion in the Immunoglobulins Involves a Molecular Ball-and-Socket Joint." *Nature* 335 (6186): 188–90. https://doi.org/10.1038/335188a0.

Li, Hongbin, Wolfgang A. Linke, Andres F. Oberhauser, Mariano Carrion-Vazquez, Jason G. Kerkvliet, Hui Lu, Piotr E. Marszalek, and Julio M. Fernandez. 2002. "Reverse Engineering of the Giant Muscle Protein Titin." *Nature* 418 (6901): 998–1002. https://doi.org/10.1038/nature00938.

Li, Shu-Ang, Linna Cheng, Yamei Yu, Jia-Huai Wang, and Qiang Chen. 2016. "Structural Basis of Dscam1 Homodimerization: Insights into Context Constraint for Protein Recognition." *Science Advances* 2 (5): e1501118. https://doi.org/10.1126/sciadv.1501118.

Lu, Shennan, Jiyao Wang, Farideh Chitsaz, Myra K. Derbyshire, Renata C. Geer, Noreen R. Gonzales, Marc Gwadz, et al. 2020. "CDD/SPARCLE: The Conserved Domain Database in 2020." *Nucleic Acids Research* 48 (D1): D265–68. https://doi.org/10.1093/nar/gkz991.

Maginnis, Melissa S. 2018. "Virus-Receptor Interactions: The Key to Cellular Invasion." *Journal of Molecular Biology* 430 (17): 2590–2611. https://doi.org/10.1016/j.jmb.2018.06.024.

Mancusi, Rebecca, and Michelle Monje. 2023. "The Neuroscience of Cancer." *Nature* 618 (7965): 467–79. https://doi.org/10.1038/s41586-023-05968-y.

Marchler-Bauer, Aron, John B. Anderson, Carol DeWeese-Scott, Natalie D. Fedorova, Lewis Y. Geer, Siqian He, David I. Hurwitz, et al. 2003. "CDD: A Curated Entrez Database of Conserved Domain Alignments." *Nucleic Acids Research* 31 (1): 383–87. https://www.ncbi.nlm.nih.gov/pubmed/12520028.

Mauriello, Gianna E., Grace E. Moncure, Roujon A. Nowzari, Callie J. Miller, and Nathan T. Wright. 2023. "The N-Terminus of Obscurin Is Flexible in Solution." *Proteins* 91 (4): 485–96. https://doi.org/10.1002/prot.26442.

McCarthy, Kevin R., Joseph L. Timpona, Simon Jenni, Louis-Marie Bloyet, Vesna Brusic, Welkin E. Johnson, Sean P. J. Whelan, and Lindsey R. Robinson-McCarthy. 2020. "Structure of the Receptor Binding Domain of EnvP(b)1, an Endogenous Retroviral Envelope Protein Expressed in Human Tissues." *mBio* 11 (6). https://doi.org/10.1128/mBio.02772-20.

Mizuno, Yumiko, Yoshimitsu Nakanishi, and Atsushi Kumanogoh. 2023. "Pathophysiological Functions of Semaphorins in the Sympathetic Nervous System." *Inflammation and Regeneration* 43 (1): 30. https://doi.org/10.1186/s41232-023-00281-7.

Mohanta, Sarajo K., Li Peng, Yuanfang Li, Shu Lu, Ting Sun, Lorenzo Carnevale, Marialuisa Perrotta, et al. 2022. "Neuroimmune Cardiovascular Interfaces Control Atherosclerosis." *Nature* 605 (7908): 152–59. https://doi.org/10.1038/s41586-022-04673-6.

Moreno-Cortes, E., J. V. Forero-Forero, P. A. Lengerke-Diaz, and J. E. Castro. 2021. "Chimeric Antigen Receptor T Cell Therapy in Oncology - Pipeline at a Glance: Analysis of the ClinicalTrials.gov Database." *Critical Reviews in Oncology/hematology* 159 (March): 103239. https://doi.org/10.1016/j.critrevonc.2021.103239.

Morimoto, Keiko, and Kazunori Nakajima. 2019. "Role of the Immune System in the Development of the Central Nervous System." *Frontiers in Neuroscience* 13 (September): 916. https://doi.org/10.3389/fnins.2019.00916.

Mosalaganti, Shyamal, Agnieszka Obarska-Kosinska, Marc Siggel, Reiya Taniguchi, Beata Turoňová, Christian E. Zimmerli, Katarzyna Buczak, et al. 2022. "AI-Based Structure Prediction Empowers

Integrative Structural Analysis of Human Nuclear Pores." *Science* 376 (6598): eabm9506. https://doi.org/10.1126/science.abm9506.

Murzin, A. G., S. E. Brenner, T. Hubbard, and C. Chothia. 1995. "SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures." *Journal of Molecular Biology* 247 (4): 536–40. https://doi.org/10.1006/jmbi.1995.0159.

Mutz, Pascal, Wolfgang Resch, Guilhem Faure, Tatiana G. Senkevich, Eugene V. Koonin, and Bernard Moss. 2023. "Exaptation of Inactivated Host Enzymes for Structural Roles in Orthopoxviruses and Novel Folds of Virus Proteins Revealed by Protein Structure Modeling." *mBio* 14 (2): e0040823. https://doi.org/10.1128/mbio.00408-23.

Naeim, Faramarz, P. Nagesh Rao, Sophie X. Song, and Ryan T. Phan. 2018. "Chapter 2 - Principles of Immunophenotyping." In *Atlas of Hematopathology (Second Edition)*, edited by Faramarz Naeim, P. Nagesh Rao, Sophie X. Song, and Ryan T. Phan, 29–56. Academic Press. https://doi.org/10.1016/B978-0-12-809843-1.00002-4.

Nelke, Christopher, Rainer Dziewas, Jens Minnerup, Sven G. Meuth, and Tobias Ruck. 2019. "Skeletal Muscle as Potential Central Link between Sarcopenia and Immune Senescence." *EBioMedicine* 49 (November): 381–88. https://doi.org/10.1016/j.ebiom.2019.10.034.

Nelson, Christopher A., Megan L. Epperson, Sukrit Singh, Jabari I. Elliott, and Daved H. Fremont. 2015. "Structural Conservation and Functional Diversity of the Poxvirus Immune Evasion (PIE) Domain Superfamily." *Viruses* 7 (9): 4878–98. https://doi.org/10.3390/v7092848.

Orengo, C. A., A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. 1997. "CATH--a Hierarchic Classification of Protein Domain Structures." *Structure* 5 (8): 1093–1108. https://www.ncbi.nlm.nih.gov/pubmed/9309224.

Otey, Carol A., Richard Dixon, Christianna Stack, and Silvia M. Goicoechea. 2009. "Cytoplasmic Ig-Domain Proteins: Cytoskeletal Regulators with a Role in Human Disease." *Cell Motility and the Cytoskeleton* 66 (8): 618–34. https://doi.org/10.1002/cm.20385.

Petrov, Anton S., Chad R. Bernier, Chiaolong Hsiao, Ashlyn M. Norris, Nicholas a. Kovacs, Chris C. Waterbury, Victor G. Stepanov, et al. 2014. "Evolution of the Ribosome at Atomic Resolution." *Proceedings of the National Academy of Sciences of the United States of America* 111 (28): 10251–56. https://doi.org/10.1073/pnas.1407205111.

Petrov, Anton S., Burak Gulen, Ashlyn M. Norris, Nicholas A. Kovacs, Chad R. Bernier, Kathryn A. Lanier, George E. Fox, et al. 2015. "History of the Ribosome and the Origin of Translation." *Proceedings of the National Academy of Sciences of the United States of America* 112 (50): 15396–401. https://doi.org/10.1073/pnas.1509761112.

Poljak, R. J. 1975. "Three-Dimensional Structure, Function and Genetic Control of Immunoglobulins." *Nature* 256 (5516): 373–76. https://doi.org/10.1038/256373a0.

Poljak, R. J., L. M. Amzel, H. P. Avey, B. L. Chen, R. P. Phizackerley, and F. Saul. 1973a. "Three-Dimensional Structure of the Fab′ Fragment of a Human Immunoglobulin at 2.8-Å Resolution." *Proceedings of the National Academy of Sciences* 70 (12): 3305–10. https://doi.org/10.1073/pnas.70.12.3305.

———. 1973b. "Three-Dimensional Structure of the Fab' Fragment of a Human Immunoglobulin at 2,8-A Resolution." *Proceedings of the National Academy of Sciences of the United States of America* 70 (12): 3305–10. https://www.ncbi.nlm.nih.gov/pubmed/4519624.

Poljak, R. J., L. M. Amzel, B. L. Chen, R. P. Phizackerley, and F. Saul. 1974. "The Three-Dimensional Structure of the Fab' Fragment of a Human Myeloma Immunoglobulin at 2.0-Angstrom Resolution."

Proceedings of the National Academy of Sciences of the United States of America 71 (9): 3440–44. https://doi.org/10.1073/pnas.71.9.3440.

Ren, Ke, and Ronald Dubner. 2010. "Interactions between the Immune and Nervous Systems in Pain." *Nature Medicine* 16 (11): 1267–76. https://doi.org/10.1038/nm.2234.

Richardson, J. S. 1981. "The Anatomy and Taxonomy of Protein Structure." *Advances in Protein Chemistry* 34: 167–339. https://doi.org/10.1016/s0065-3233(08)60520-3.

Ruan, Jianbin, Chao Xu, Chuanbing Bian, Robert Lam, Jia-Pey Wang, Joanna Kania, Jinrong Min, and Jianye Zang. 2012. "Crystal Structures of the Coil 2B Fragment and the Globular Tail Domain of Human Lamin B1." *FEBS Letters* 586 (4): 314–18. https://doi.org/10.1016/j.febslet.2012.01.007.

Salvador, Andrea Francesca, Kalil Alves de Lima, and Jonathan Kipnis. 2021. "Neuromodulation by the Immune System: A Focus on Cytokines." *Nature Reviews. Immunology* 21 (8): 526–41. https://doi.org/10.1038/s41577-021-00508-z.

Sanes, Joshua R., and S. Lawrence Zipursky. 2020. "Synaptic Specificity, Recognition Molecules, and Assembly of Neural Circuits." *Cell* 181 (3): 536–56. https://doi.org/10.1016/j.cell.2020.04.008.

Saul, F. A., L. M. Amzel, and R. J. Poljak. 1978. "Preliminary Refinement and Structural Analysis of the Fab Fragment from Human Immunoglobulin New at 2.0 A Resolution." *The Journal of Biological Chemistry* 253 (2): 585–97. https://www.ncbi.nlm.nih.gov/pubmed/618887.

Schaeffer, R. Dustin, Jing Zhang, Kirill E. Medvedev, Lisa N. Kinch, Qian Cong, and Nick V. Grishin. 2024. "ECOD Domain Classification of 48 Whole Proteomes from AlphaFold Structure Database Using DPAM2." *PLoS Computational Biology* 20 (2): e1011586. https://doi.org/10.1371/journal.pcbi.1011586.

Senkevich, Tatiana G., Natalya Yutin, Yuri I. Wolf, Eugene V. Koonin, and Bernard Moss. 2021. "Ancient Gene Capture and Recent Gene Loss Shape the Evolution of Orthopoxvirus-Host Interaction Genes." *mBio* 12 (4): e0149521. https://doi.org/10.1128/mBio.01495-21.

Sen, Mehmet. 2024. "Interplay between Intrinsic Flexibility and Sugar Coating in Contactin-2 Homodimerization." *Structure* 32 (1): 3–5. https://doi.org/10.1016/j.str.2023.12.005.

Sergeeva, Alina P., Phinikoula S. Katsamba, Filip Cosmanescu, Joshua J. Brewer, Goran Ahlsen, Seetha Mannepalli, Lawrence Shapiro, and Barry Honig. 2020. "DIP/Dpr Interactions and the Evolutionary Design of Specificity in Protein Families." *Nature Communications* 11 (1): 2125. https://doi.org/10.1038/s41467-020-15981-8.

Shapiro, Lawrence, James Love, and David R. Colman. 2007. "Adhesion Molecules in the Nervous System: Structural Insights into Function and Diversity." *Annual Review of Neuroscience* 30: 451–74. https://doi.org/10.1146/annurev.neuro.29.051605.113034.

Shapiro, L., P. D. Kwong, A. M. Fannon, D. R. Colman, and W. A. Hendrickson. 1995. "Considerations on the Folding Topology and Evolutionary Origin of Cadherin Domains." *Proceedings of the National Academy of Sciences of the United States of America* 92 (15): 6793–97. https://www.ncbi.nlm.nih.gov/pubmed/7624321.

Shibuya, Masabumi. 2008. "Vascular Endothelial Growth Factor-Dependent and -Independent Regulation of Angiogenesis." *BMB Reports* 41 (4): 278–86. https://doi.org/10.5483/bmbrep.2008.41.4.278.

———. 2013. "VEGFR and Type-V RTK Activation and Signaling." *Cold Spring Harbor Perspectives in Biology* 5 (10): a009092. https://doi.org/10.1101/cshperspect.a009092.

Shilts, Jarrod, Yannik Severin, Francis Galaway, Nicole Müller-Sienerth, Zheng-Shan Chong, Sophie Pritchard, Sarah Teichmann, Roser Vento-Tormo, Berend Snijder, and Gavin J. Wright. 2022. "A Physical Wiring Diagram for the Human Immune System." *Nature* 608 (7922): 397–404. https://doi.org/10.1038/s41586-022-05028-x.

Sillitoe, Ian, Nicola Bordin, Natalie Dawson, Vaishali P. Waman, Paul Ashford, Harry M. Scholes, Camilla S. M. Pang, et al. 2021. "CATH: Increased Structural Coverage of Functional Space." *Nucleic Acids Research* 49 (D1): D266–73. https://doi.org/10.1093/nar/gkaa1079.

Stanfield, Robyn L., Adam Zemla, Ian A. Wilson, and Bernhard Rupp. 2006. "Antibody Elbow Angles Are Influenced by Their Light Chain Class." *Journal of Molecular Biology* 357 (5): 1566–74. https://doi.org/10.1016/j.jmb.2006.01.023.

Steinman, Lawrence. 2004. "Elaborate Interactions between the Immune and Nervous Systems." *Nature Immunology* 5 (6): 575–81. https://doi.org/10.1038/ni1078.

Stengel, Katharina F., Kristin Harden-Bowles, Xin Yu, Lionel Rouge, Jianping Yin, Laëtitia Comps-Agrar, Christian Wiesmann, J. Fernando Bazan, Dan L. Eaton, and Jane L. Grogan. 2012. "Structure of TIGIT Immunoreceptor Bound to Poliovirus Receptor Reveals a Cell-Cell Adhesion and Signaling Mechanism That Requires Cis-Trans Receptor Clustering." *Proceedings of the National Academy of Sciences of the United States of America* 109 (14): 5399–5404. https://doi.org/10.1073/pnas.1120606109.

Tong, Xiajing, and Yan Zou. 2019. "Cell Adhesion Molecules." In *Advances in Membrane Proteins: Building, Signaling and Malfunction*, edited by Yu Cao, 67–83. Singapore: Springer Singapore. https://doi.org/10.1007/978-981-13-9077-7_4.

Tunyasuvunakool, Kathryn, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, et al. 2021. "Highly Accurate Protein Structure Prediction for the Human Proteome." *Nature*, July. https://doi.org/10.1038/s41586-021-03828-1.

Vendome, Jeremie, Shoshana Posy, Xiangshu Jin, Fabiana Bahna, Goran Ahlsen, Lawrence Shapiro, and Barry Honig. 2011. "Molecular Design Principles Underlying β-Strand Swapping in the Adhesive Dimerization of Cadherins." *Nature Structural & Molecular Biology* 18 (6): 693–700. https://doi.org/10.1038/nsmb.2051.

Verschueren, Erik, Bushra Husain, Kobe Yuen, Yi Sun, Sairupa Paduchuri, Yasin Senbabaoglu, Isabelle Lehoux, et al. 2020. "The Immunoglobulin Superfamily Receptome Defines Cancer-Relevant Networks Associated with Clinical Outcome." *Cell* 182 (2): 329–44.e19. https://doi.org/10.1016/j.cell.2020.06.007.

Vishnivetskiy, Sergey A., Elizabeth K. Huh, Eugenia V. Gurevich, and Vsevolod V. Gurevich. 2021. "The Finger Loop as an Activation Sensor in Arrestin." *Journal of Neurochemistry* 157 (4): 1138–52. https://doi.org/10.1111/jnc.15232.

Wang, Han, Jianxun Qi, Shuijun Zhang, Yan Li, Shuguang Tan, and George F. Gao. 2019. "Binding Mode of the Side-by-Side Two-IgV Molecule CD226/DNAM-1 to Its Ligand CD155/Necl-5." *Proceedings of the National Academy of Sciences of the United States of America* 116 (3): 988–96. https://doi.org/10.1073/pnas.1815716116.

Wang, Jiyao, Philippe Youkharibache, Aron Marchler-Bauer, Christopher Lanczycki, Dachuan Zhang, Shennan Lu, Thomas Madej, et al. 2022. "iCn3D: From Web-Based 3D Viewer to Structural Analysis Tool in Batch Mode." *Frontiers in Molecular Biosciences* 9 (February): 831740. https://doi.org/10.3389/fmolb.2022.831740.

Wang, Jiyao, Philippe Youkharibache, Dachuan Zhang, Christopher J. Lanczycki, Renata C. Geer, Thomas Madej, Lon Phan, et al. 2020. "iCn3D, a Web-Based 3D Viewer for Sharing 1D/2D/3D

Representations of Biomolecular Structures." *Bioinformatics* 36 (1): 131–35. https://doi.org/10.1093/bioinformatics/btz502.

Wang, Norman, William F. Smith, Brian R. Miller, Dikran Aivazian, Alexey A. Lugovskoy, Mitchell E. Reff, Scott M. Glaser, Lisa J. Croner, and Stephen J. Demarest. 2009. "Conserved Amino Acid Networks Involved in Antibody Variable Domain Interactions." *Proteins* 76 (1): 99–114. https://doi.org/10.1002/prot.22319.

Wang, Zhexin, and Stefan Raunser. 2023. "Structural Biochemistry of Muscle Contraction." *Annual Review of Biochemistry* 92 (June): 411–33. https://doi.org/10.1146/annurev-biochem-052521-042909.

Watson, Fiona L., Roland Püttmann-Holgado, Franziska Thomas, David L. Lamar, Michael Hughes, Masahiro Kondo, Vivienne I. Rebel, and Dietmar Schmucker. 2005. "Extensive Diversity of Ig-Superfamily Proteins in the Immune System of Insects." *Science* 309 (5742): 1874–78. https://doi.org/10.1126/science.1116887.

Wilkinson, Martin, Rodrigo U. Gallardo, Roberto Maya Martinez, Nicolas Guthertz, Masatomo So, Liam D. Aubrey, Sheena E. Radford, and Neil A. Ranson. 2023. "Disease-Relevant β2-Microglobulin Variants Share a Common Amyloid Fold." *Nature Communications* 14 (1): 1190. https://doi.org/10.1038/s41467-023-36791-8.

Williams, A. F., and A. N. Barclay. 1988. "The Immunoglobulin Superfamily--Domains for Cell Surface Recognition." *Annual Review of Immunology* 6: 381–405. https://doi.org/10.1146/annurev.iy.06.040188.002121.

Williams, Alan F., and A. Neil Barclay. 1988. "The Immunoglobulin Superfamily—domains for Cell Surface Recognition." *Annual Review of Immunology* 6 (1): 381–405. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.822.6038&rep=rep1&type=pdf.

Wojtowicz, Woj M., Wei Wu, Ingemar Andre, Bin Qian, David Baker, and S. Lawrence Zipursky. 2007. "A Vast Repertoire of Dscam Binding Specificities Arises from Modular Interactions of Variable Ig Domains." *Cell* 130 (6): 1134–45. https://doi.org/10.1016/j.cell.2007.08.026.

Wu, T. T., and E. A. Kabat. 1970. "An Analysis of the Sequences of the Variable Regions of Bence Jones Proteins and Myeloma Light Chains and Their Implications for Antibody Complementarity." *The Journal of Experimental Medicine* 132 (2): 211–50. https://doi.org/10.1084/jem.132.2.211.

Wu, Wei, Goran Ahlsen, David Baker, Lawrence Shapiro, and S. Lawrence Zipursky. 2012. "Complementary Chimeric Isoforms Reveal Dscam1 Binding Specificity in Vivo." *Neuron* 74 (2): 261–68. https://doi.org/10.1016/j.neuron.2012.02.029.

Xue, Xiaoguang, Qingyu Lu, Hui Wei, Dongli Wang, Dongwei Chen, Guangjun He, Li Huang, Hanzhong Wang, and Xinquan Wang. 2011. "Structural Basis of Chemokine Sequestration by CrmD, a Poxvirus-Encoded Tumor Necrosis Factor Receptor." *PLoS Pathogens* 7 (7): e1002162. https://doi.org/10.1371/journal.ppat.1002162.

Xu, Shuwa, Qi Xiao, Filip Cosmanescu, Alina P. Sergeeva, Juyoun Yoo, Ying Lin, Phinikoula S. Katsamba, et al. 2018. "Interactions between the Ig-Superfamily Proteins DIP-α and Dpr6/10 Regulate Assembly of Neural Circuits." *Neuron* 100 (6): 1369–84.e6. https://doi.org/10.1016/j.neuron.2018.11.001.

Yang, Shangyu, Yong Wang, Feiyang Yu, Rao Cheng, Yiwei Zhang, Dan Zhou, Xuanxiu Ren, Zengqin Deng, and Haiyan Zhao. 2023. "Structural and Functional Insights into the Modulation of T Cell Costimulation by Monkeypox Virus Protein M2." *Nature Communications* 14 (1): 5186. https://doi.org/10.1038/s41467-023-40748-2.

Yang, Yang, Lu Guo, Lin Chen, Bo Gong, Da Jia, and Qingxiang Sun. 2023. “Nuclear Transport Proteins: Structure, Function, and Disease Relevance.” *Signal Transduction and Targeted Therapy* 8 (1): 425. https://doi.org/10.1038/s41392-023-01649-4.

Yin, Zheng, Jishou Zhang, Shuwan Xu, Jianfang Liu, Yao Xu, Junping Yu, Mengmeng Zhao, Wei Pan, Menglong Wang, and Jun Wan. 2022. “The Role of Semaphorins in Cardiovascular Diseases: Potential Therapeutic Targets and Novel Biomarkers.” *FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology* 36 (10): e22509. https://doi.org/10.1096/fj.202200844R.

Youkharibache, Philippe. 2019. “Protodomains: Symmetry-Related Supersecondary Structures in Proteins and Self-Complementarity.” *Methods in Molecular Biology* 1958: 187–219. https://doi.org/10.1007/978-1-4939-9161-7_10.

———. 2021. “Topological and Structural Plasticity of the Single Ig Fold and the Double Ig Fold Present in CD19.” *Biomolecules* 11 (9). https://doi.org/10.3390/biom11091290.

Zhang, Yang, and Jeffrey Skolnick. 2005. “TM-Align: A Protein Structure Alignment Algorithm Based on the TM-Score.” *Nucleic Acids Research* 33 (7): 2302–9. https://doi.org/10.1093/nar/gki524.

Zinn, Kai, and Engin Özkan. 2017. “Neural Immunoglobulin Superfamily Interaction Networks.” *Current Opinion in Neurobiology* 45 (August): 99–105. https://doi.org/10.1016/j.conb.2017.05.010.

Zipursky, S. Lawrence, Woj M. Wojtowicz, and Daisuke Hattori. 2006. “Got Diversity? Wiring the Fly Brain with Dscam.” *Trends in Biochemical Sciences* 31 (10): 581–88. https://doi.org/10.1016/j.tibs.2006.08.003.

A universal residue numbering scheme for the Immunoglobulin-fold (Ig-fold) to study Ig-Proteomes and Ig-Interactomes

Caesar Tawfeeq[1], Jiyao Wang[2], Umesh Khaniya[3], Thomas Madej[2], James Song[2], Ravinder Abrol[1*], and Philippe Youkharibache[3*].

[1]Department of Chemistry and Biochemistry, California State University Northridge, CA 91330, USA
[2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA
[3]Cancer Data Science Laboratory, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA.

[*]Corresponding authors. e-mail address: abrol@csun.edu; philippe.youkharibache@nih.gov

SUPPLEMENTARY INFORMATION

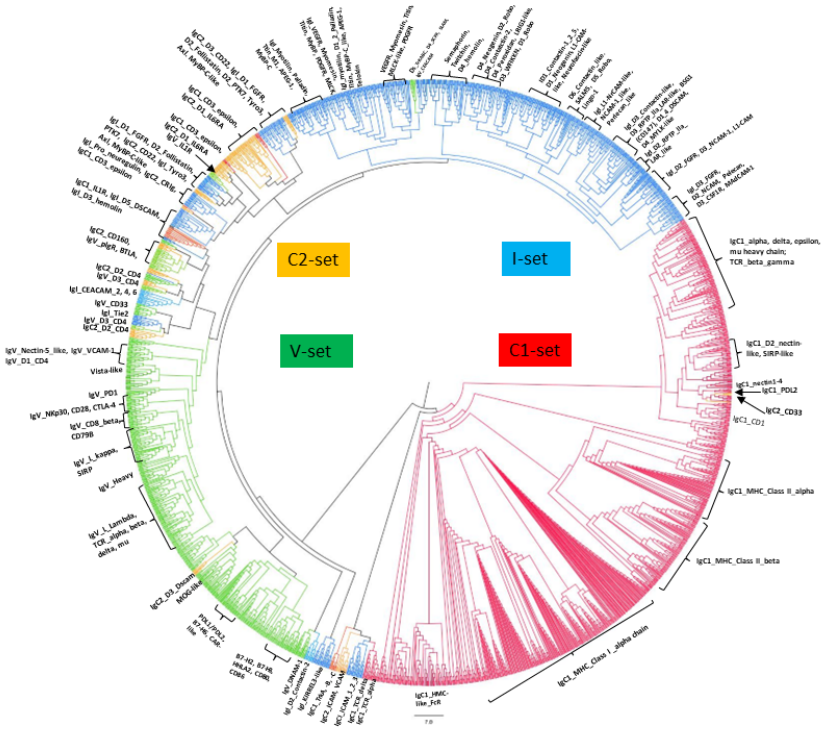


Figure S1: CDD IgSF Tree. Constructed with BLOSUM 62 with the Score of Optimally-Extended Block method

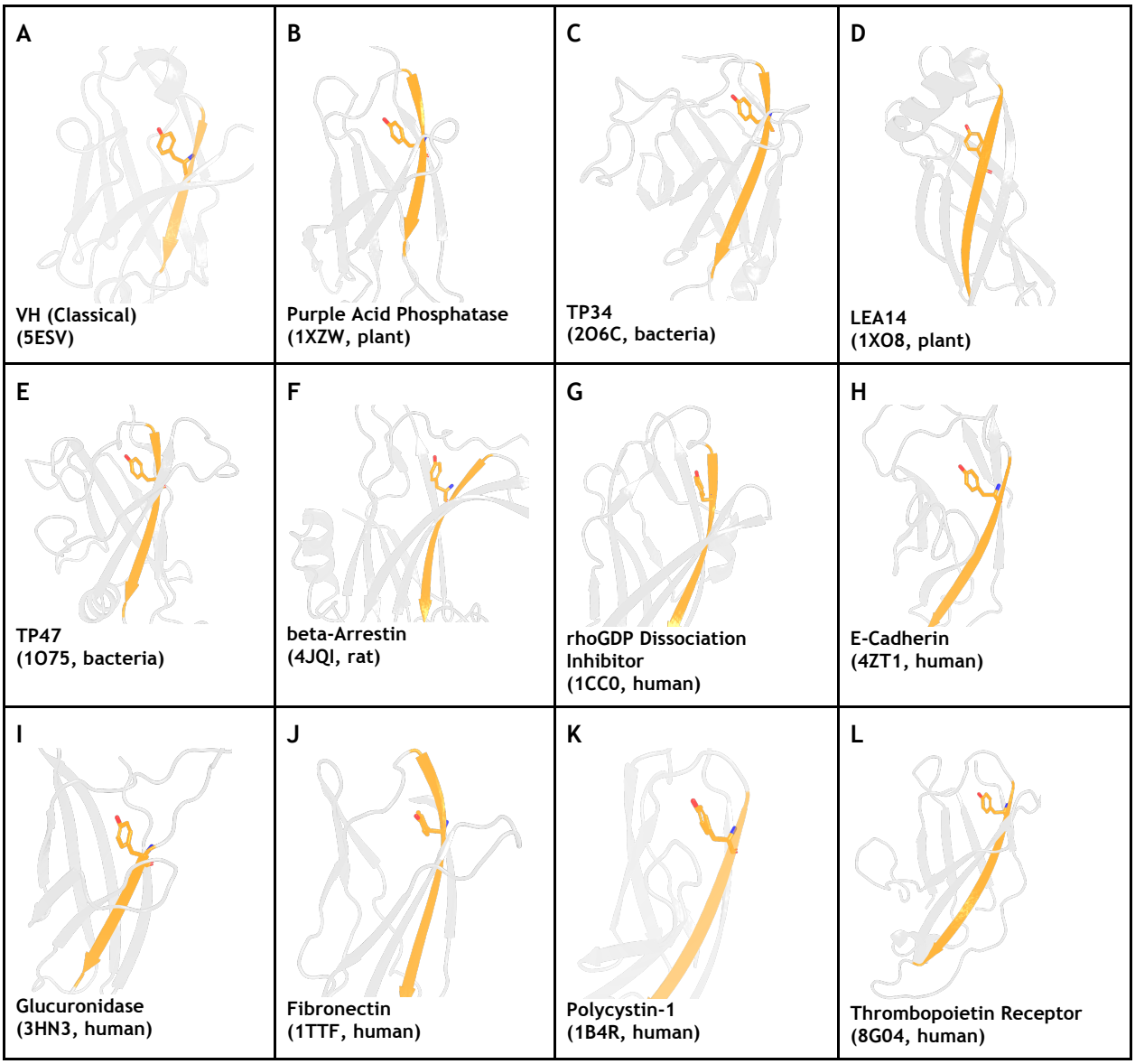


Figure S2: F strand Tyr (igs# 8548 or 8546) is highly conserved across many eukaryotic and bacterial Ig-like domains despite topostructural variations in the fold. This Tyr appears to play similar roles by interacting with the EF loop, and may provide insights into the evolution of a number of proteins sharing the Ig-fold. It certainly points toward a key structural residue.

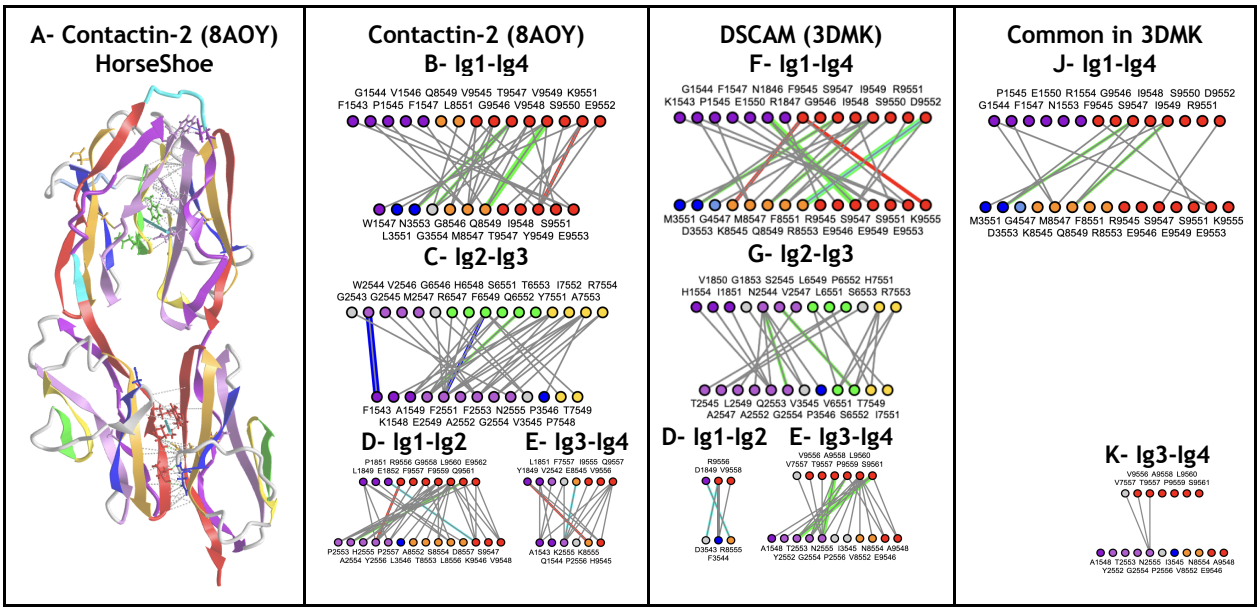


Figure S3: Tertiary Intrachain interfaces in the Horseshoe superdomain formed by the four N-terminal residues in Ig-chains. This superdomain allows a certain level of plasticity observed in comparing contactin-2 and DSCAM. The RMSD is 4.9Å. A) Contactin-2 Horseshoe (Pdbid 8A0Y) shows an Ig1-Ig4 and a Ig2-Ig3 antiparallel interfaces. B) Ig1-Ig4 interactome in contactin-2. C) Ig2-Ig3 interactome in contactin-2. D) Ig1-Ig2 interactome in contactin-2. E) Ig3-Ig4 interactome in contactin-2. F) Ig1-Ig4 interactome in DSCAM. G) Ig2-Ig3 interactome in DSCAM. H) Ig1-Ig2 interactome in DSCAM. I) Ig3-Ig4 interactome in DSCAM. J) Common interactions in Ig1-Ig4. This pairwise interactome is more conserved than Ig2-Ig3 that is more plastic. K) Common interactions in Ig3-Ig4.

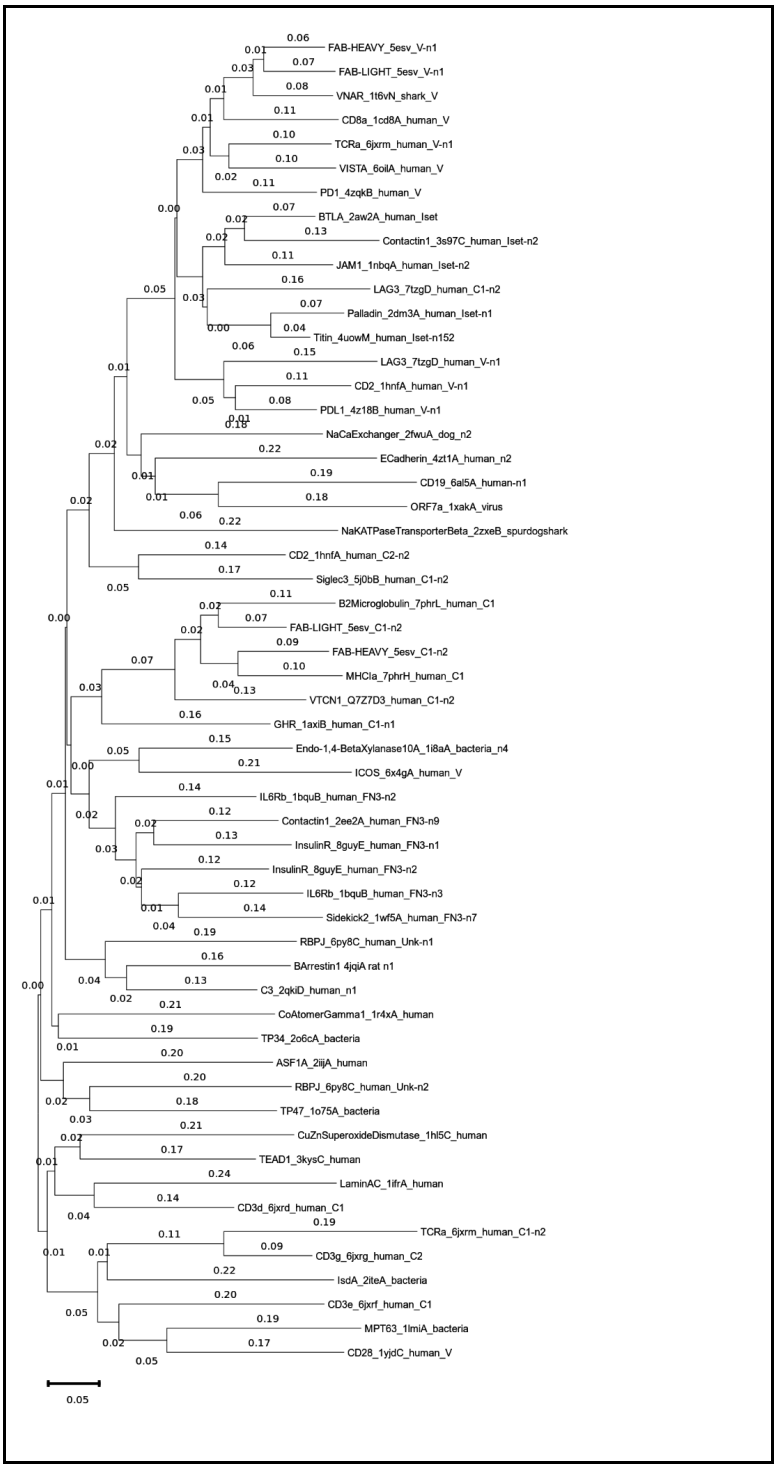


Figure S4: Clustering Ig templates using MEGA11. The graph shows the distances between templates. The corresponding TM-score is equal to 1 - distance.

Supplementary Files:

SF1	Multiple Sequence Alignment – IgV domains – fasta format
SF2	Multiple Sequence Alignment – IgI domains – fasta format
SF3	Multiple Sequence Alignment – IgC1 domains – fasta format
SF4	Multiple Sequence Alignment – IgC2 domains – fasta format
SF5	Multiple Sequence Alignment – IgFN3 domains – fasta format
SF6	Multiple Sequence Alignment – IgCad (Cadherin) domains – fasta format
SF7	Ig-domains templates - Table 2 in Excel format
SF8	IgStrand Alignment of 258 SARS-CoV2 VH, VL,CH1,CL domains