Edit Preview PDF

Biophysical Chemistry xxx (xxxx) 106971

Contents lists available at ScienceDirect

Biophysical Chemistry

journal homepage: www.elsevier.com/locate/biophyschem

Comprehensive classification of proteins based on structures that engage lipids by COMPOSEL

Michael Overduin [a,*], Troy A. Kervin [a], Zachary Klarenbach [b], Trixie Rae C. Adra [a], Rakesh K. Bhat [a]

[a] Department of Biochemistry, University of Alberta, Edmonton, AB, Canada
[b] Department of Physiology, University of Alberta, Edmonton, AB, Canada

CORRECTED PROOF

ARTICLE INFO

Keywords:
Bilayer insertion
Intrinsically disordered region
Ligand complex
Lipid interaction
Monotopic
Myristoylation
Palmitoylation
Polytopic
Protein classification
Peripheral membrane domain
Transmembrane

ABSTRACT

Structures can now be predicted for any protein using programs like AlphaFold and Rosetta, which rely on a foundation of experimentally determined structures of architecturally diverse proteins. The accuracy of such artificial intelligence and machine learning (AI/ML) approaches benefits from the specification of restraints which assist in navigating the universe of folds to converge on models most representative of a given protein's physiological structure. This is especially pertinent for membrane proteins, with structures and functions that depend on their presence in lipid bilayers. Structures of proteins in their membrane environments could conceivably be predicted from AI/ML approaches with user-specificized parameters that describe each element of the architecture of a membrane protein accompanied by its lipid environment. We propose the Classification Of Membrane Proteins based On Structures Engaging Lipids (COMPOSEL), which builds on existing nomenclature types for monotopic, bitopic, polytopic and peripheral membrane proteins as well as lipids. Functional and regulatory elements are also defined in the scripts, as shown with membrane fusing synaptotagmins, multidomain PDZD8 and Protrudin proteins that recognize phosphoinositide (PI) lipids, the intrinsically disordered MARCKS protein, caveolins, the β barrel assembly machine (BAM), an adhesion G-protein coupled receptor (aGPCR) and two lipid modifying enzymes - diacylglycerol kinase DGKε and fatty aldehyde dehydrogenase FALDH. This demonstrates how COMPOSEL communicates lipid interactivity as well as signaling mechanisms and binding of metabolites, drug molecules, polypeptides or nucleic acids to describe the operations of any protein.

1. Introduction

The universe of proteins can be divided roughly in half between those that are membrane-associated and those that are not. The many ways in which human proteins directly interact with lipid bilayers are quantified in Fig. 1. Those with membrane-spanning α helices represent 26% of all human proteins [1]. Another nine employ β-barrels to span plasma and mitochondrial outer membranes, with β-barrels also predominating in chloroplast, plastid and bacterial outer membranes [2,3]. Membrane anchoring is reversibly mediated by palmitoyl groups, which are attached to 10% of human proteins [4]. Another 2%, 1% and 0.9% of human proteins are permanently modified by myristoyl [5], isoprenyl [6,7] and glycosylphosphatidylinositol (GPI) groups [8], respectively. Others such as Hedgehog are altered by the attachment of cholesterol groups. All these lipid anchors insert into membranes [9]. Peripheral membrane domains transiently recognize lipids on membrane surfaces, and include 70 different phosphoinositide (PI) code readers such as FYVE, PH and PX domains that are present in 9% of human proteins [10]. Over 0.4% of human proteins are monotopic, typically being tethered to one bilayer leaflet through helical structures [11]. Approximately 4.9% of yeast proteins contain disordered lipid binding regions (DLBRs) [12], with a similar proportion likely in humans [13]. Hence it is more common for proteins to partially insert into only one leaflet rather than being transmembrane. The true number of membrane proteins remains unclear as many remain uncharacterized and most exhibit several modes of lipid engagement. In fact, it remains possible that more proteins interact with membranes than do not.

Our understanding of membrane interactions is informed by the 5% of 200,000 structures in the protein data bank (PDB) that constitute membrane proteins [14]. However, the folds of the majority of membrane proteins remain structurally uncharacterized [15], and the lipid bilayers that stabilize their conformations are rarely retained or visible. These problems could conceivably be resolved with AI/ML approaches that predict protein structures and could expand to include membrane complexes if provided with the relevant restraints. mEven more challenging is modeling transmembrane membrane reader (TMMR) proteins that fuse two juxtaposed membranes [16]. Protein fold prediction methods including AlphaFold2 have modelled hundreds of millions of soluble and membrane proteins, yet while the former may rival experimental structures in their accuracy [17], the latter clearly do not, as shown below. We think this is because current AI/ML methods lack information on the membrane environment that plays a fundamental role in the structure of membrane proteins but varies from protein to protein. The COMPOSEL formalism presented here could improve membrane protein models generated by such algorithms. It can be used to concisely describe the diverse ligands and mechanisms of any protein, whether soluble or membrane, and may be used for teaching and reporting purposes.

---

* Corresponding author.
E-mail address: overduin@ualberta.ca (M. Overduin).

https://doi.org/10.1016/j.bpc.2023.106971
Received 24 December 2022; Accepted 5 February 2023
0301-4622/© 20XX

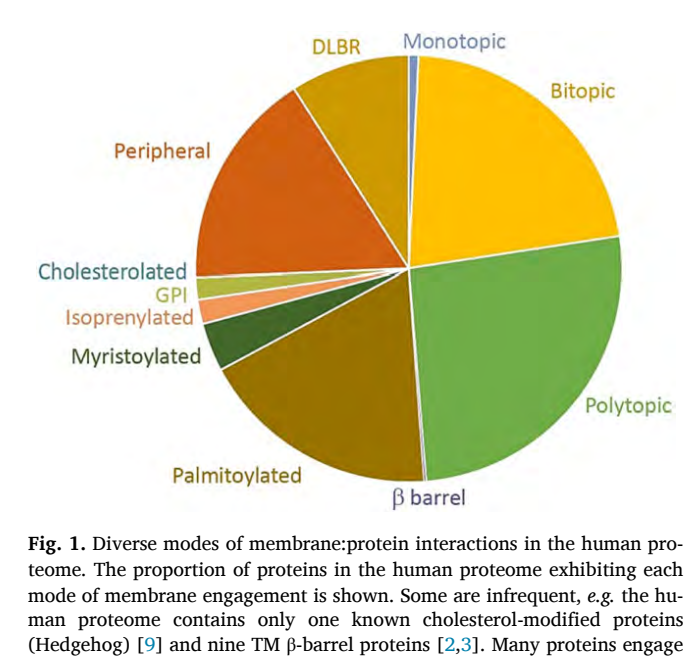

Fig. 1. Diverse modes of membrane:protein interactions in the human proteome. The proportion of proteins in the human proteome exhibiting each mode of membrane engagement is shown. Some are infrequent, *e.g.* the human proteome contains only one known cholesterol-modified proteins (Hedgehog) [9] and nine TM β-barrel proteins [2,3]. Many proteins engage membranes through multiple modes. The proportions are indicative as the precise numbers of DLBR and monotopic membrane-associated proteins remain unclear, and the peripheral fraction refers to those proteins containing membrane reader domains that can recognize PIs [10], with many other proteins interacting with other lipid bilayer surfaces.

Two major obstacles to the accurate prediction of membrane protein structures are the dearth of experimental data and the multiplicity of forces that determine the folds of membrane proteins. Experimental data on biological bilayer complexes can be acquired with synthetic polymers that spontaneously convert membranes containing protein assemblies into native nanodiscs [18]. This is revealing otherwise inaccessible *ex vivo* structures including of lipid-stabilized folding assemblies [19]. In contrast, more widely used detergents can dislocate bound lipids, alter protein conformations and compromise stabilizing and regulatory interactions [20], limiting our insights. Moreover, the forces that drive protein folding in membranes differ from those experienced in solution [21]. Hydrophobic, entropic and hydrogen bonding forces are uniquely exerted upon membrane proteins during ribosome exit, membrane insertion, protein compaction, folding and complex assembly. Different folding determinants including matching helix and strand lengths to bilayer thickness come into play with membrane proteins. The dynamics of various lipid tails and headgroups differ, as do pressure profiles across the lipid bilayer. A variety of chaperones and folding machines reside in organelles and employ enzymes and templates to usher polypeptides stepwise into their membranes. The AI/ML algorithms used for protein structure prediction generally do not utilize such information. We feel that accurate membrane protein structure prediction requires invoking additional restraints that are appropriate for lipid bilayers.

Lipid interactions are increasingly understood to be fundamentally important. The genetic code that articulates the structure of proteins is extended by the membrane code whereby folded proteins read, write, erase and move lipid signals in membranes. This concept is best exemplified by PI code readers that recruit proteins to organelle surfaces [10]. The membrane code is broadly applicable to any membrane:protein complex ("memtein"), which is defined as any membrane protein together with its directly associated layer of bound lipids [18,20]. The set of bound lipids is organelle-specific and vital for the physiological stability and dynamics of any membrane protein, preserving its structural, functional and regulatory states. Sets of lipid molecules can be specifically recognized by membrane readers to drive protein subcellular targeting and functions such as signaling or catalytic activity. These form "lipidons" that are akin to a DNA codon and are specifically recognized in a lipid bilayer surface rather than in helical duplex of nucleic acids . The operation of lipidons throughout the subcellular compartments determines the dynamic organization of cells, and is a culmination of the expression of the genome [10]. Being held together noncovalently, lipidons are difficult to purify, but their docking surfaces in membrane readers are identifiable, being typically larger and more exposed than the pockets that bind single lipid molecules. Lipidon recognition is regulated by PIP-stops [22] and MET-stops [23], which include phosphorylations and acetylations of membrane docking surfaces, respectively. Such signature features of the membrane code could be elaborated by AI/ML methods with the guidance of COMPOSEL.

Although structures of membrane proteins bound to biologically-relevant lipid bilayers are rarely available, their interaction surfaces can be identified by automated tools. The residues that bind membrane surfaces can be predicted using programs including DREAMM [24,25], Ez-3D [26], Positioning of Proteins in Membranes (PPM) [27,28] and Membrane Optimal Docking Area (MODA) [29]. The latter has been used to identify lipid binding surfaces in PI recognition domains [10,30], bacterial and viral trafficking proteins [31,32], prions [33], and coronavirus spikes [34,35]. Due to its accessibility, sensitivity and coverage MODA is used here to identify residues that engage lipids within simple scripts. This allows identification of previously undiscovered peripheral lipid binding residues in ordered and disordered protein modules, which have been under-represented in databases but are found to be pervasive here. In fact, COMPOSEL is applicable to any protein and can communicate the totality of their structures and interactions including all known and novel membrane engagements.

2. Materials and methods

2.1. Sequences and modifications

Protein sequences were obtained from UniProt [36] and domains were identified in the CATH, SCOP [37] and SMART resources [38]. Signal sequences were predicted using Philius [39], Phobius [40], PrediSi [41] and SignalP [42]. The Group-based Prediction System (GPS) Lipid [43], SwissPalm [44] and pCysMod [45] tools were used to identify protein palmitoylation, myristoylation, and prenylation sites. PIP-stops [22,46] and MET-stops [23] were identified in protein sequences from post-translational modifications reported in Phosphosite [47] close to membrane binding sites.

2.2. Protein structures

Membrane protein structures were obtained by surveying the RCSB Protein DataBank (PDB) [48], Membranome [49], OPM [28], PDBTM [50], MemProtMD [51], mpstruc [52] and PDBTM [50]. Models of structures were generated using ColabFold [53], AlphaFold2 [54], I-TASSER [55] and Rosetta [56]. Coiled coils, which often increase binding avidity of proteins for membranes, were predicted using Marcoil [57]. Protein structures were visualized with PyMol [58] and ICM Browser programs.

2.3. Transmembrane helix prediction

Membrane topologies of TM proteins were predicted using DeepTMHMM [59], HTP [60], BOCTUPUS [61], Philius [39], Phobius [40], Prodiv [62] and TMFoldWeb [63]. The results were compared to the output of the AlphaFold2, MODA, Rosetta, SignalP and PrediSi programs to identify consensus signal and membrane interacting sequences.

2.4. Membrane docking site prediction

The MODA program (Molsoft) was used to predict membrane-docking surfaces. This algorithm was trained on a family of well-characterized peripheral membrane proteins and assigns a membrane binding propensity score to each residue based on its likelihood of interacting with phospholipid bilayers [29]. Missing coordinates in structure files were not modelled; instead, other similar structures were considered where gaps existed. Sets of at least two residues that have MODA scores of 30 were ranked as having significant membrane binding propensities. The output from MODA was analyzed in Excel (Microsoft) to assess consistency across homologous structures and to generate heatmaps showing consensus membrane binding profiles. Novel membrane interacting surfaces in peripheral membrane domains were confirmed using PPM 3.0 [27,28], which predicts positions of protein structures on fluid anisotropic solvent slabs that represent membranes. Orientations of full-length membrane proteins were compared to those modelled in Membranome [49].

3. Results

3.1. Membrane protein classification schemes

The proposed COMPOSEL system incorporates the hallmarks of current membrane protein classifications. Membrane proteins are traditionally divided into monotopic, bitopic and polytopic classes based on whether they contain partial, single or multiple membrane spanning elements, respectively, with signal and stop transfer sequences dictating their biogenesis pathways [64]. Such proteins were later categorized into eight topological types [65]. Of these, type 1 proteins contain cleavable N-terminal signal peptide (SP) sequences, which are followed by stop-transfer sequences that forms TM helices and leave the C-terminus in the cytosol (Fig. 2). Type 2 proteins possess a single N-terminal TM helix spanning the membrane with the N-terminus left in the cytosol, while type 3 proteins have a single TM helix that crosses in the opposite direction with a cytosolic C-terminus. In contrast type 4 proteins contain a TM helix at their C-terminus, while type 5 proteins contain multiple TM helices. Type 6 and 7 proteins contain lipid chains or GPI anchors that insert into a leaflet of the membrane, respectively, while type 8 proteins are peripheral, with residues that reversibly engage membrane surfaces.

Several organelle-specific machines guide protein folding into membranes. The assembly of β barrels in mitochondria, chloroplast and bacterial outer membranes is mediated by SAM, Oep80 and BAM complexes, respectively [66]. Type 1, 2, 3 and 5 proteins fold through the signal recognition particle, while type 4 proteins use the guided entry of tail anchored (GET) insertase [67]. The signature sequences include cleavable signal peptide, stop-transfer and re-integration motifs that are differentially processed at the ER [68,69] to anchor hydrophobic N- or C-terminal regions by the Sec complex [70,71.] F olding of type 5 proteins also involves a recently resolved multipass translocon and is guided by their glycosylation [72,73]. Basic elements are preferentially displayed on the cytoplasmic leaflet according to the "positive inside rule", although mitochondrial and nuclear bitopic proteins do not obey this [74]. A variety of computational programs such as MemType-2 L [65] have been developed to assign protein sequences to these types, but they have inherent limitations as they are not 3D structure-based. Moreover, the various classification schemes are not integrated, universally applicable or inherently obvious to the nonexpert, prompting development of the COMPOSEL system (Fig. 2) to describe any protein simply and intuitively.

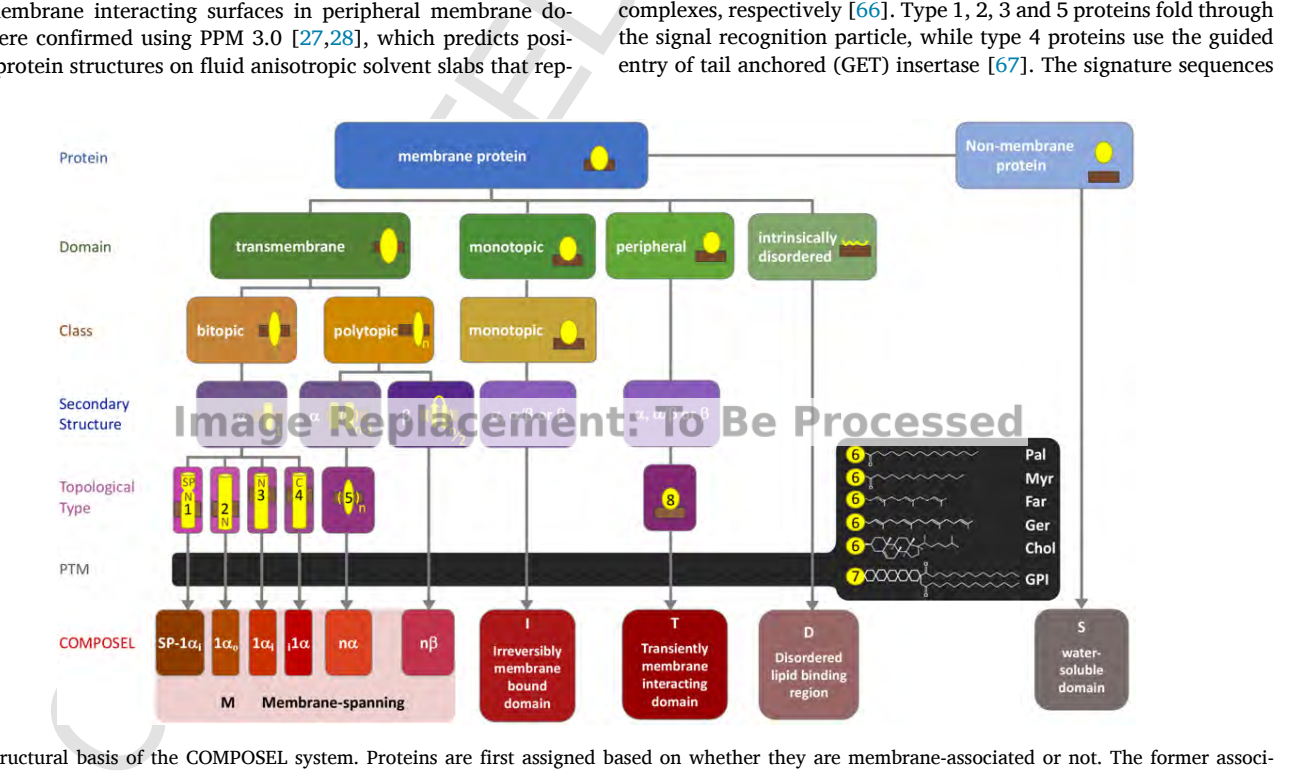

Fig. 2. Structural basis of the COMPOSEL system. Proteins are first assigned based on whether they are membrane-associated or not. The former associate with lipid bilayers *via* structural domains that are either transmembrane, monotopic, or peripheral depending on whether they cross the membrane, insert stably, or bind reversibly to one leaflet. Some intrinsically disordered regions lacking stable structure also interact transiently with lipid bilayer surfaces *via* basic and hydrophobic motifs. Proteins containing transmembrane domains are either bitopic if they have one TM element or polytopic if they have two or more TM elements, in which case they are referred to as type 5. The membrane-spanning domains of bitopic proteins are composed of a number "n" of TM helices or β strands. Monotopic proteins can be further divided into four types depending on whether they have a signal peptide (SP) at their N-terminus (type 1), a short N-terminal region preceding their TM helix, which is inside or outside the cytosol for types 2 and 3, respectively, or a C-terminal TM helix for type 4 proteins. Peripheral membrane domains are termed type 8 and are distinguished from monotopic proteins as their lipid bilayer interactions are transient rather than permanent. Type 6 membrane proteins are post-translationally modified by covalent attachment of palmitoyl (Pal), myristoyl (Myr), farnesyl (Far), geranylgeranyl (Ger) or cholesterol (Chol) groups while type 7 proteins contain GPI anchors. These features are all captured by the COMPOSEL scripts.

The historical membrane protein classification schemes do not reflect our current understanding of the membrane proteome for reasons outlined below:

- No membrane-embedded β-barrels are designated as a class or type, although parallel β-barrels of caveolins are integrated into the plasma membrane [3] and many antiparallel β-barrels span chloroplast, mitochondrial and bacterial outer membranes [2,66].
- Many intrinsically disordered proteins mediate important interactions with membranes [12,75] but are also missing in current membrane protein classification systems.
- PTMs are crucial to membrane interactions and membrane protein folding and stability, yethere is no differentiation between reversible and irreversible cysteine palmitoylation or glycine myristoylation, which are produced by different pathways with distinct effects, being permanent and reversible, respectively [5,44], while attachments of lipids such as farnesyl, geranylgeranyl or cholesterol groups are known [76] but are also not specifically incorporated into classification schemes, nor are glycosylations that guide multipass membrane protein folding or PIP-stops or MET-stops that regulate membrane recognition.
- There is no indication of which of the hundreds of polytopic folds [15] are present or the number or order of their TM elements. Even the common folds of tetraspanins or G-protein coupled receptors (GPCRs) which could easily be defined on the basis of their distinctive 4 or 7 helical bundles are not specified in the current schemes.
- Bitopic proteins are a highly populated class and typically form stable dimers [74], however oligomerization is not addressed by any membrane protein classifier.
- Monotopic proteins that stably sit inside a membrane leaflet form a unique class of membrane proteins but are not included as any specific topological type [65].
- Despite being widespread [10] peripheral membrane domains are often unrecognized and under-reported in classification schemes, including in the case of TMMR proteins that engage two distinct membranes simultaneously [16].
- Membrane proteins often belong to multiple topological types. For example, palmitoylated TMMRs such as synaptotagmins typically contain an N-terminal TM helix and two membrane binding C2 domains and thus belongs to types 3, 6 and 8. However analysis using MemType-2 L [65] predicts that they are not membrane proteins of any type.
- Functionally useful information such as a subcellular location, signaling pathway or biochemical reaction is not designated in current classifiers, nor are effects or identities of lipid ligands, conformational changes, dynamic processes or regulatory events.
- Some proteins assume multiple structural folds with very different physiological and pathological activities depending on its membrane environment context. For example, lipids are involved in the conversion of normal helical structures of membrane proteins into neuropathological β-sheet fibril forms as in the cases of α synuclein, β amyloid, prions and tau [77].

Such limitations and ambiguities contrast with the thousands of soluble protein fold families that are accurately assigned in proteins by bioinformatics tools such as InterPro [78]. Hence, despite their utility in the past, the old membrane protein classification schemes are insufficiently descriptive, inclusive or predictive to be broadly useful, particularly at proteome scales. Here COMPOSEL is proposed as a new formalism that extends beyond the historical classifiers, and is accessible, descriptive and scalable to all soluble and membrane proteins and their complexes by any researcher or computer language.

3.2. Components of the COMPOSEL system

We propose the COMPOSEL system in which protein-membrane complexes are denoted by the following formula:

$$\mathrm{SP} - \left[n\alpha[m - aa - s]^{x} : \mathrm{O}\left[\mathrm{L}[l - s]\right]\right] - \left((\mathrm{T}(m - aa - s)) : \mathrm{L}(l - s)\right)^{y}_{e}$$

where the components in order from left to right are the signal peptide, membrane spanning and bilayer leaflet engaging regions, respectively.

Proteins are composed of sets of domains that are classified as either membrane spanning and placed within square brackets, *i.e.* [M], or non-membrane spanning. The latter can either be categorized as an Irreversibly membrane-inserted monotopic domain (I), a Disordered lipid-binding region (D), a water-Soluble domain not contacting membranes or electrostatically repulsed away from them (S), or a Transiently membrane-interacting peripheral domain (T). The single letter names of protein module types form the word "MIDST" to aid in remembering the five modes of membrane engagement. Each letter type can be replaced with the standard names of domains, such as those from CATH or SCOP, that are found in a given protein [37]. Multidomain proteins have a number "y" of such domains, these are listed consecutively according to their order in the polypeptide sequence. The parameters describing protein components are as follows.

Domain descriptors provide information about components of a protein domain or region:

SP: cleavable signal peptide found at the N-terminus of type 1 proteins

n: number of TM elements

α: membrane-spanning secondary structure, either "α" or "3/10" helix or "β" strand

aa: amino acid residue number

m: post-translational modification indicated as a prefix before an amino acid number. These include lipidations such as Pal, Myr, Far, Ger, GPI or Chol groups, as well as "Nglyc" and "Oglyc for Asn- and Ser-linked glycosylations, "PO-" (phosphate) such as in a regulatory PIP-stop [22] and "Ac-" or "Me-" for acetylated or methylated residues that represent the MET-stops blocking membrane interactions [23].

Interactors are listed after a colon, and refer to molecular partner, which can be described as:

L: multicomponent ligand that is specifically recognized by a domain such as a polypeptide or nucleic acid sequence or lipid motif, which are underlined, italicized or unmodified, respectively, for clarity, which is particularly useful in cases of where ligand binding sites overlap. A domain and its ligand that engage in one bilayer leaflet only are placed within (parentheses), while those that bind nucleic acid or protein motifs are placed within {curly} or <angle> brackets, respectively, to separate types of ligand recognition events for legibility by computer programs.

l: component of a ligand such as a peptide residue, *nucleic acid base residue*, bilayer lipid, or **small molecule** such as a metabolite or drug, can be differentiated by underlined, *italicized*, unmodified, or **bolded** text, respectively, for both the ligand and interacting residues, to provide visual clarity for people,

-s: substituent group such as a hydroxyl (-OH) or 3-phosphate (3P), which is added as a suffix after an amino acid residue or ligand

x: oligomeric state, for example, "2" or "5" indicate a dimer or pentamer, respectively.

Subcellular environment of a domain, as is indicated by the subscripted suffix "e" in the generic COMPOSEL formula above, could be any of a variety of lower case letters such as:

e: extracellular
i: inside the cytosol
u: luminal
o: outside the cytosol
p: periplasmic

The subcellular organelle (O) destination follows the colon after a membrane-associated protein domain or region, and can include abbreviations such as:

EE: early endosome
ER: endoplasmic reticulum
LE: late endosome
MOM: mitochondrial outer membrane
OM: outer membrane
PM: plasma membrane
PO: peroxisome
SE: sorting endosome
SV: synaptic vesicle
TGN: trans-Golgi network
VM: viral membrane

In some cases, although the organelle destination of a protein may be unclear, the specific lipid ligands may be known. These can be defined with the bilayer lipid entity being recognized as follows:

$$\mathrm{B}\left[z_i\_\mathrm{L}_i\left[c_{ij} : d_{ij}/c_{ik} : d_{ik}\right]\right]$$

where
B indicates that bilayer lipids are being specified
z_i indicates the relative proportion of lipid species i
L_i refers to lipid species i, such as PA, PC, PE, PI or PS
c_{ij} indicates the average acyl chain length for chain j of lipid i
d_{ij} indicates the average number of degrees of unsaturation in acyl chain j of lipid i

With two lipids this would be:

$$\mathrm{B}\left[z_1 - \mathrm{L}_1\left[c_{1j} : d_{1j}/c_{1k} - d_{1k}\right], z_2 - \mathrm{L}_2\left[c_{2j} : d_{2j}/c_{2k} : d_{2k}\right]\right]$$

and a proposed lipidon recognized by the EEA1 FYVE domain with a 1:2:3 ratio of PI3P:PS:PC lipid molecules containing 1-palmitoyl and 2-oleoyl chains [79–81] would be expressed as:

$$\mathrm{B}\left[1 - \mathrm{PI3P}[16:0/18:1], 2 - \mathrm{PS}[16:0/18:1], 3 - \mathrm{PC}[16:0/18:1]\right]$$

As such, the lipid nomenclature used in COMPOSEL is based on systematic names and abbreviations as defined by the International Union of Pure and Applied Chemists and the International Union of Biochemistry and Molecular Biology (IUPAC-IUBMB) and recommended by the Lipid Maps consortium [82].

This COMPOSEL system is designed to be intuitive and compatible as it is based on commonly accepted terms. It is also intended to be as simple as possible while capturing the essential complexity, with the five types of protein modules being treated equally by the parameters. That is, thee MIDST protein types are distinguished principally by the nature of their membrane interactions, *i.e.* membrane-spanning, irreversibly inserted, disordered binder, water soluble non-binder, or transiently lipid-bound, respectively, but otherwise can be defined by the same COMPOSEL descriptors.

The environment of any domain outside the membrane is designated by a superscripted letter as in "e", "i", "o", "u", or "p" for extracellular, inside the cytosol, outside the cytosol, luminal (within an organelle), or periplasmic, respectively.

-Subservient elements can optionally be included in bracketed, subscripted or superscripted formats. For example, atoms are subservient to amino acid residues, which are subservient to domains, which are subservient to proteins within the cells of an organism. These be written in the form of a nested set of brackets species(cell(protein(domain(residue(atom))))) to be more machinereadable. To provide visual clarity for people the constituent elements can be subscripted to indicate boundaries or environments such as SP_{1-20} or Kinase_i to indicate a 20-residue signal peptide or cytosolic kinase, respectively, or superscripted to indicate ligand interactions as in $\mathrm{FYVE}^{356}\mathrm{:LE}^{\mathrm{PIP2}}$ to indicate that Protrudin protein's FYVE domain residue 356 recognizes a PIP2 lipid in the late endosome membrane.

Where there are a series of items such as several residues that recognize a shared ligand they are separated by commas, whereas those that recognize distinct ligands are separated by semicolons. All such items are listed in order by sequence for linear polymers such as proteins or as a set of membrane ligands within a lipidon based on standard lipid nomenclatures [82].

To illustrate how this COMPOSEL system works, we work through examples of different types of membrane proteins using predicted models and experimentally determined structures (Table 1). The process includes running related structures and models through MODA to identify consensus lipid interacting residues. The sequence is analyzed for the positions of TM elements (*e.g.* using DeepTHMM, Phobius [40], or HTP [60]), for modifications using GPS-Lipid [43] or Swiss-Palm [44], and for coiled coils and signal peptides using online tools such as Marcoil [57] and PrediSi [41], respectively. To improve confidence, the conservation of structural and membrane interacting elements within homologous families can also be analyzed using programs such as ClustalOmega [83]. The consensus predictions are then integrated into the COMPOSEL nomenclature to define how the protein interacts with membranes.

The COMPOSEL scripts designate a protein's fold, lipid binding elements, ligands and localization, and can be used by people to communicate biochemical functions or by programs to predict memtein structures. The intermolecular restraints are already used by programs like HADDOCK to calculate bilayer-bound structures of eighteen transmembrane proteins [84], three bilayer-docked enzymes [85–87], a spike fusion polypeptide [88], oncogenic Ras [89] as well as FYVE, PX and PH domains [46,90,91]. The application to AI/ML algorithms could incorporate PTM-based and sequence conservation restraints to further enhance accuracy. The script information is intended to be concise and intuitive while encompassing the range of cellular and molecular details of interest to computational and structural biologists in a scalable format, as illustrated by the examples below.

3.3. Protrudin

Protrudin is expressed in neurons where it controls neurite growth by directing membrane trafficking ("protusions") between the ER and endosomes. This trafficking protein is also known as Zfy27 engages partners including PDZD8 [92], VAP [93] and KIF5 [94] to assist in trafficking at endosomes to ensure neuronal polarity and integrity. Its coiled coil mediates dimerization and tetramerization [95], and it interacts with RAB GTPases [96,97]. The N-terminal half of protrudin forms four predicted TM helices that anchor into the ER membrane based on HTP analysis, although neither AlphaFold2 nor Rosetta arrange them as a tightly packed bundle (Fig. 3). The splayed helices and their lack of interdigitation in Protrudin models from across evolutionary space is surprising given the critical role of strong anchoring at membrane contact sites where lipids are transferred. Hence this splaying could be an artifact, as stably packed parallel oligomeric structures are likely form in biological membranes.

Table 1
Selected membrane proteins and their classifications. The lipid ligands, cellular locations (ER, endoplasmic reticulum; MOM, mitochondrial outer membrane; PM, plasma membrane). The proposed new membrane protein classification scheme indicates the type, class, number and type of TM elements and positions of peripheral membrane and monotopic domains. Multimeric states of domains are indicated with a superscripted number.

Protein	Uniprot	Major Class	Type	COMPOSEL
ADGRF1	Q5T601	polytopic	5, 8	$(\mathrm{NTD\text{-}SEA\text{-}GAIN})_e\text{-}[7\alpha\mathrm{:PM}]$
BamA	P0A940	β-barrel	1, 8	$\mathrm{SP}_{1-21}\text{-}(\mathrm{Potra\text{-}Potra\text{-}Potra}^{205,\ 206}\text{-}\mathrm{Potra\text{-}Potra})_p\text{-}[16\beta\mathrm{:OM}]$
BamB	P77774	peripheral	1, 6, 8	$\mathrm{SP}_{1-19}\text{-}(\mathrm{PQQ}^{\mathrm{Pal}\text{-}20,\ 100-104,\ 106,\ 204,\ 255-256,\ 273}\mathrm{:OM})_p$
Cav1	Q03135	monotopic	1, 6, 8	$\mathrm{D}\text{-}(\mathrm{Cav}^{81-83,\ 86,\ 90,\ 94,\ \mathrm{Pal}\text{-}133,\ \mathrm{Pal}\text{-}143,\ \mathrm{Pal}\text{-}156,\ 172,\ 174,\ 176})_i{}^{11}\mathrm{:PM}^{\mathrm{PI(4,5)P2,\ PS,\ Chol}}$
DGKε	P52429	bitopic	3, 8	$[1\alpha\mathrm{:PM}]\text{-}(\mathrm{C1\text{-}C1\text{-}DAGKc})_i$
EEA1	Q15075	peripheral	8	$\mathrm{C_2H_2\text{-}CC^2\text{-}(FYVE}^{1370,\ 1371,\ +1372,\ +1373,\ 1375}\mathrm{:EE}^{\mathrm{PI3P,PS,PC}})_i$
FALDH	P51648	bitopic	4, 6, 8	$(\mathrm{AldD\text{-}BD^2\text{-}GK})_i\text{-}[1\alpha\mathrm{:ER}]$
Gag	P12497	peripheral	6, 8	$(\mathrm{MA}^{\mathrm{Pal}\text{-}2;\ 30,\ 32}\mathrm{:PM}^{\mathrm{PI(4,5)P2}})\text{-}(\mathrm{CA}^{\underline{316},\ \underline{317}})^2\text{-}\{\mathit{NC:RNA}\}^2$
GDAP1	Q8TB36	bitopic	4, 8	$(\mathrm{GST})_i\text{-}[1\alpha\mathrm{:MOM}]$
MARCKS	P29966	DLBR	6, 8	$(\mathrm{D}^{\mathrm{Myr}\text{-}1}\text{-}\mathrm{ED:PM}^{\mathrm{PI(4,5)P2}})_i$
PDZD8	Q8NEN9	bitopic	3, 8	$[1\alpha\mathrm{:ER}]\text{-}(\mathrm{SMP:ER}^{\mathrm{PI(4,5)P2}}\text{-}\mathrm{PDZ\text{-}C1:LE}^{\mathrm{PI4P}})_i\text{-}\mathrm{CC}$
PRNP	P04156	peripheral	1, 7, 8	$\mathrm{SP}\text{-}(\mathrm{D}^{99,\ 122}\text{-}\mathrm{PrP}^{169,\ 222-228;\ \mathrm{GPI}\text{-}230}\mathrm{:PM}^{\mathrm{PE}})_i\text{-}\mathrm{SP}$
Protrudin	Q5T4F4	polytopic	5, 8	$[4\alpha\mathrm{:ER}]\text{-}(\mathrm{CC^2\text{-}FYVE:LE}^{\mathrm{PIP2}})_i$
Spike	P0DTC2	bitopic	3, 6, 8	$(\mathrm{NTD\text{-}RBD})^3{}_e\text{-}[1\alpha\mathrm{:VM}]\text{-}(\mathrm{Cte}^{10\text{-}\mathrm{Pal}})_i$
Syt1	P21579	bitopic	3, 6, 8	$[1\alpha^{\mathrm{Pal}\text{-}\ 75,\ -76,\ -78,\ -80,\ -83}\mathrm{:SV}]\text{-}(\mathrm{C2:PM}^{\mathrm{PS,PC}}\text{-}\mathrm{C2:PM}^{\mathrm{PI(4,5)P2,\ PS}})_i$

Phospholipid interacting motifs are predicted by MODA in the disordered juxtamembrane segments in the AlphaFold2 and Rosetta-based models. These include GVRYLLR-66, QEVCRAR-116, AYRVLHW-176 and YR-227 and sit peripherally next to each TM helix on the cytosolic side of the ER. The various TM predictors generally fail to correctly predict the Protrudin topology found by AlphaFold2 and Rosetta, with Philius, Phobius and Prodiv only identifying 3 helices, or 2 long TM helices in the case of HMMTOP. Only the 4-helix topology complies with the positive inside rule as well as the function of this protein, which mediates membrane contact sites with late endosomes. To carry this out its FYVE domain must recognize PI and extract lipids from these compartments *via* cytoplasmic access, which is also needed for partner protein interactions [97,98].

Zinc-stabilized FYVE domain structures including PI complexes have been solved by NMR [99,100] and X-ray crystallography [101], allowing us to predict the phospholipid bilayer binding motifs (TFSVL-360 and KRRS-365) of Protrudin's FYVE domain using MODA and sequence alignments. Comparison with other FYVE domain structures and binding studies is needed in this case to weed out inconsistent inferences. In particular, the AlphaFold2 and Rosetta models lack the two zinc ions that stabilize these folds and incorrectly predict zinc coordinating motifs as membrane binding, and GPS-Lipid incorrectly predicts that the cysteines involved may be lipidated. Nonetheless such errors are relatively obvious and easily excluded out by structural analysis.

The short COMPOSEL script of protrudin indicating the four N-terminal TM helices and the membrane-dissociated FYVE domain located in the cytosol is:

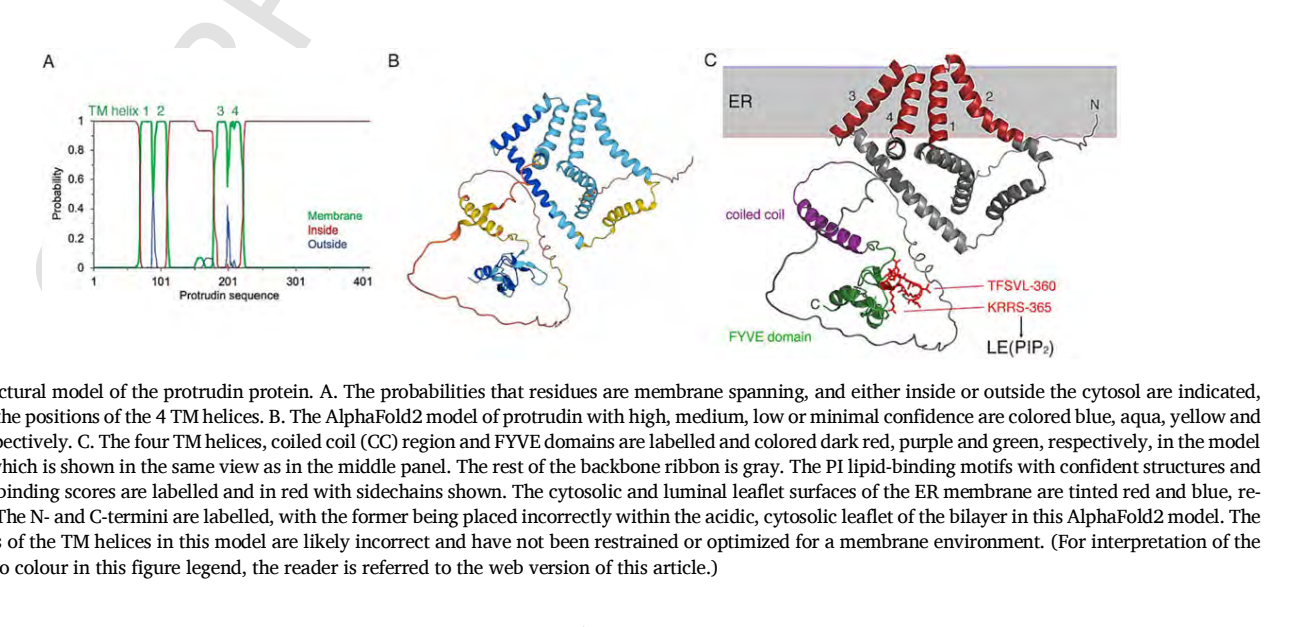

Fig. 3. Structural model of the protrudin protein. A. The probabilities that residues are membrane spanning, and either inside or outside the cytosol are indicated, along with the positions of the 4 TM helices. B. The AlphaFold2 model of protrudin with high, medium, low or minimal confidence are colored blue, aqua, yellow and orange, respectively. C. The four TM helices, coiled coil (CC) region and FYVE domains are labelled and colored dark red, purple and green, respectively, in the model structure, which is shown in the same view as in the middle panel. The rest of the backbone ribbon is gray. The PI lipid-binding motifs with confident structures and membrane binding scores are labelled and in red with sidechains shown. The cytosolic and luminal leaflet surfaces of the ER membrane are tinted red and blue, respectively. The N- and C-termini are labelled, with the former being placed incorrectly within the acidic, cytosolic leaflet of the bilayer in this AlphaFold2 model. The orientations of the TM helices in this model are likely incorrect and have not been restrained or optimized for a membrane environment. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

$$[4\alpha] - (\mathrm{CC} - \mathrm{FYVE})_i$$

Upon late endosome binding, protrudin's cellular compartments and primary lipid ligand type would be specified as:

$$[4\alpha : \mathrm{ER}] - (\mathrm{CC} - \mathrm{FYVE} : \mathrm{LE}(\mathrm{PIP}_2))_i$$

where ER and LE stand for endoplasmic reticulum and late endosome, respectively, and PIP$_2$ refers to phosphatidylinositol bisphosphate.

The fold of the TM helices in the protrudin model can also be indicated by bracketing the packed sets according to the density of interhelical contacts. In the case of AlphaFold2 model of protrudin (Fig. 3) only the first and last protrudin TM helices interdigitate, while the other two TM helices are structurally isolated. Thus the AlphaFold2 model infers (potentially incorrectly) that there are 3 separately folded TM structural domains, which can be defined as the following:

$$[2\alpha[\alpha1 : \alpha4]] - [1\alpha[\alpha2]] - [1\alpha[\alpha3]] : \mathrm{ER} - (\mathrm{CC} - \mathrm{FYVE} : \mathrm{LE}(\mathrm{PIP}_2))_i$$

In contrast, it is possible that the four TM helices form a single tertiary structural domain composed of 4 helices, with the amino acid residues inserting into membranes and binding lipids being stated as :

$$\left[4\alpha : \left[[\alpha]^{70-87}, \alpha4^{201-223}\right], \alpha2^{89-109}, \alpha3^{179-199}\right] : \mathrm{ER}] - \left(\mathrm{CC} - \mathrm{FYVE}^{356-360,362-365} : \mathrm{LE}^{\mathrm{PIP2}}\right)_i$$

Reactions and soluble domain interactions can also be defined. Protrudin dimerizes *via* its coiled coil element, while its FYVE domain is in equilibrium between being either free of ligand and bound to PIP$_2$ molecules in late endosomal membranes. This reaction can be written as:

$$[4\alpha : \mathrm{ER}] - (\mathrm{CC}^2 - \mathrm{FYVE} + \mathrm{LE}(\mathrm{PIP}_2))_i \Leftrightarrow [4\alpha : \mathrm{ER}] - (\mathrm{CC}^2 - \mathrm{FYVE} : \mathrm{LE}(\mathrm{PIP}_2))_i$$

To calculate membrane-bound structural models, the residues known to interact with specific lipid groups can also be included. Identifying lipid recognition determinants requires experimental structures of complexes that are lacking for protrudin but can be extrapolated from analysis of the EEA1 FYVE domain's lipid interactions [80,81,99–101], which show its RRHHCR-1375 sequence interacting specifically with the phosphatidylinositol 3-phosphate group, in addition to nonspecific interactions with PC molecules, as superscripted below. The angle at which the EEA1 FYVE domain orients on the early endosome (EE) membrane also depends on co-ligation of phosphatidylserine (PS), which also forms part of its lipidon [80,81]. Membrane binding is an order of magnitude tighter upon protonation of the two His residues, and hence is favoured low pH endocytic compartments, as indicated by the "+" signs in the script below. In fact, subcellular localization is increasingly recognized to be driven by complementarity of pH-dependent protein and membrane charges [102].

$$\left(\mathrm{FYVE}^{1370,1371,+1372,+1373,1375} : \mathrm{EE}^{\mathrm{PI3P,PS,PC}}\right)_i$$

EEA1 interacts with activated Rab5 switch residues I53 and Y82 through its N-terminal C$_2$H$_2$ zinc finger residues F41 and I42 [103] and forms stable homodimers *via* its central coiled coil (CC) region [104]. The Rab5 GTPase is stably anchored into sorting endosome (SE) membranes and is geranylgeranylated on cysteines 212 and 213 [105]. Hence the COMPOSEL script of the full-length dimeric membrane complex of EEA1 with protein-interacting residues underlined is:

$$\mathrm{C_2H_2}^{\underline{41,42}} : \mathrm{Rab5}^{\underline{53,82};\mathrm{Ger}-212,213:\mathrm{SE}} - \mathrm{CC}^2 - \left(\mathrm{FYVE}^{1370-1375} : \mathrm{EE}^{\mathrm{PI3P,PS,PC}}\right)_i$$

These scripts are designed as logical expressions. Elements of descending order such as domains, residues and atoms (or organelles, lipids and their substituents) are listed from left to right, with the smaller elements superscripted for clarity. Series of domains in a protein are linked in order from N- to C-termini with hyphens. Each protein domain and their known ligand are on the left and right of a colon, respectively. Sets of lipids recognized specifically are included together, with the primary ligand indicated first where known. Sets of TM elements are listed in topological order and bracketed according to packing density to define TM folds. COMPOSEL is intended to allow description of both soluble and membrane components at levels of resolution from atomic and up. To show general applicability we work through additional cases below.

3.4. Membrane-fusing synaptotagmins

The synaptotagmin (Syt) family facilitates calcium-dependent membrane fusion events. The seventeen human Syt proteins shared a conserved architecture characterized by an N-terminal membrane anchor with two C2 domains connected *via* a variable linker region [106]. Within this family, Syt1 functions as a calcium-sensor and fuses synaptic vesicles (SV) to the plasma membrane to enable neurotransmitter release [107]. Its structure (Fig. 4) is embedded in presynaptic vesicles by an N-terminal TM helix [108] along with five palmitoylated cyteines [109] that are necessary for presynaptic trafficking to intracellular vesicle [110]. Intriguingly, these modified residues may form an extended conformation in lipid bilayers [111] rather than an extension of the TM helix. The C2 domains mediate membrane fusion in conjunction with soluble N-ethylmaleimide sensitive factor attachment protein receptor (SNARE) proteins, which are located on the opposing plasma membrane [112]. While resting in the absence of calcium, the second C2 domain associates with the plasma membrane lipids, especially PI(4,5)P$_2$ as well as PS, *via* its lysine-rich patch [107,111]. Clusters of PI(4,5)P$_2$ on the plasma membrane can signal synaptotagmin association and subsequent membrane fusion [113,114]. This electrostatic interaction with the anionic phospholipids' drives MODA positive residues 367 and 368 in Ca^{2+} binding loop 3 of the second C2 domain to insert within the membrane plasma membrane [115,116]. The first C2 domain does not insert into the membrane without conformational changes caused by calcium ion binding [117]. Hence, the COMPOSEL script for Syt1 in the absence of calcium signaling is:

$$[\alpha1]^{\mathrm{Pal}-75,76,78,80,83} : \mathrm{SV} - \left(\mathrm{C2} - \mathrm{C2}^{367-368} : \mathrm{PM}^{\mathrm{PI(4,5)P2,PS}}\right)_i$$

The changes induced by calcium alter Syt1's mode of membrane association. The first C2 domain associates with calcium ions sequentially in three interrelated binding sites [118]. Calcium complexes with both the membrane and the first C2 domain, coordinating with anionic phospholipids and aspartates [118–120], thus inducing conformational changes and promoting membrane insertion. On the calcium binding, quenching experiments show F232 and F235 become embedded in the membrane [120], while NMR implicates R234 and F235 as membrane binding [121]. MODA analysis indicates that residues 232–236 have high propensities for membrane docking, as confirmed by PPM3, while structurally proximal residues 174 and 200 are also predicted to bind membranes. Experimental evidence points to PS as being a key ligand [122] alongside phosphatidylcholine (PC) [123]. Typically, membrane tethering of the first C2 occurs with *trans*-membranes, in this case the plasma membrane [107,124], as the membranes fuse to release neurotransmitters stored in the synaptic vesicle. In response to calcium, both of the C2 domains simultaneously embed within phospholipid bilayers [123], with the second C2 domain showing a similar membrane interactivity profile. Therefore, a COMPOSEL script of the form found mediating Syt1 membrane-membrane contact sites can be represented as:

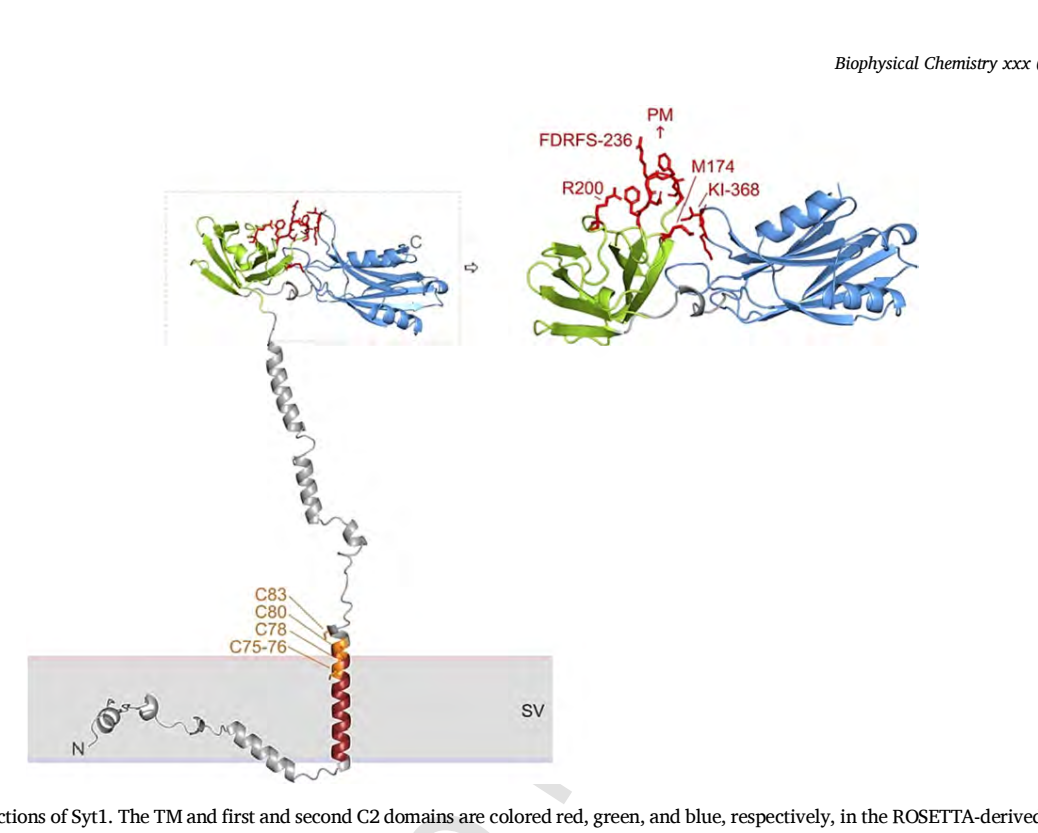

Fig. 4. Membrane interactions of Syt1. The TM and first and second C2 domains are colored red, green, and blue, respectively, in the ROSETTA-derived model of the structure of the entire Syt1 protein. The five palmitoylated cysteine residues are colored orange and labelled, and the rest of the backbone is gray. The C2 domain residues which display membrane binding propensities of at least 30 are drawn in red with sidechains sticks and labelled in the expanded view on the right. The membrane is shown with a slab with the cytoplasmic and luminal leaflets edged in red and blue, respectively. The positions of the membrane binding elements and soluble helices have not been restrained and are not correctly positioned in this model. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

$$[\alpha1]^{\mathrm{Pal}-75,-76,-78,-80,-83} : \mathrm{SV} - \left(\mathrm{C2}^{174,200,232-236} : \mathrm{PM}^{\mathrm{PS,PC}} - \mathrm{C2}^{367-368} : \mathrm{P}\right.$$

The syntaptotagmin membrane fusion machines assemble into multimeric rings that pin membranes together [125]. Their assembly is dynamic and regulated, not only by calcium. In particular, Syt1's membrane binding interactions are also blocked by PIP-stops [22,46] formed by phosphorylation of membrane-proximal T202, Y230, S236 and Y365 residues, as well as a MET-stop through acetylation of K237 [23]. These post-translational modifications (PTMs) would be predicted to detach the C2 domains from membrane as in the following state:

$$[\alpha1]^{\mathrm{Pal}-75,-76,-78,-80,-83} : \mathrm{SV} - \mathrm{C2}_i{}^{\mathrm{PO}-202,230,236;\mathrm{Ac}-237} - \mathrm{C2}^{\mathrm{PO}-365}$$

3.5. Multidomain PDZD8 protein at membrane contact sites

PDZD8 is an ER-resident protein found at membrane contact sites with the late endosome and mitochondria. It acts as a tether by complexing with other proteins, thus contributing to cytoskeletal organization and regulation of calcium dynamics and lipid transfer [98,126]. Its structure is anchored to the ER membrane through a single TM domain, followed by SMP and PDZ domains, which mediate lipid extraction (Fig. 5). The SMP domain transports lipids including PI(4,5)P$_2$ in *C. elegans* membranes, which regulate PIP distribution at ER-LE contact sites [127]. The tethering role of PDZD8 is supported by the C1 domain, which interacts with PS and PI(4)P in endosome membranes, as well as the coiled-coil domain (CC), which interacts with Rab7, which is geranylated and binds GTP and membranes [92]. The short COMPOSEL script for PDZD8 is:

$$[1\alpha : \mathrm{ER}] - (\mathrm{SMP} - \mathrm{PDZ} - \mathrm{C1})_i - \mathrm{CC}$$

Its interaction with the membranes *via* lipid ligands is as follows:

$$[1\alpha : \mathrm{ER}] - \left(\mathrm{SMP} : \mathrm{ER}^{\mathrm{PI(4,5)P2}} - \mathrm{PDZ} - \mathrm{C1} : \mathrm{LE}^{\mathrm{PI4P,PS}}\right)_i - \mathrm{CC}$$

The binding of PDZD8 CC residues L1001 and L1090 by a pair of Rab7 molecules in a GTP-dependent manner occurs through residue Y78 in the "A" molecule and F45, W62 and F77 in the "B" molecule, respectively [92]. The two states in the interaction can be described as:

$$[1\alpha : \mathrm{ER}] - \left(\mathrm{SMP}^{\mathrm{PI(4,5)P2}} - \mathrm{PDZ} - \mathrm{C1} : \mathrm{LE}^{\mathrm{PI\,4P,PS}}\right)_i - \mathrm{CC} + \mathrm{Rab7} + \mathbf{GTP} \rightarrow [1\alpha : \mathrm{ER}] - \left(\mathrm{SMP}^{\mathrm{PI(4,5)P2}} - \mathrm{PDZ} - \mathrm{C1} : \mathrm{LE}^{\mathrm{PI\,4P,PS}} - < \mathrm{CC}^{\underline{1001;1090}} : \mathbf{GTP}\right.$$

Only the structure of the CC domain of PDZD8 has been experimentally resolved (PDB: 7f6j) [92]. Consequently membrane-binding propensities of the full length protein were obtained from structural models generated by AlphaFold2 [54], Membranome [49] and Rosetta [56] as well as by ColabFold [53]. Residues consistently scoring over 20 in MODA were considered to be membrane-interacting. The scores of the free protein were also compared with those of the available crystal structure to determine which CC residues interact with the protein partner Rab7. The PDZD8 residues that insert into the membrane *via* the lipid-binding PDZ domain, as well as the pair of isoprenylated Rab7 subunits in the complex in COMPOSEL are expressed as:

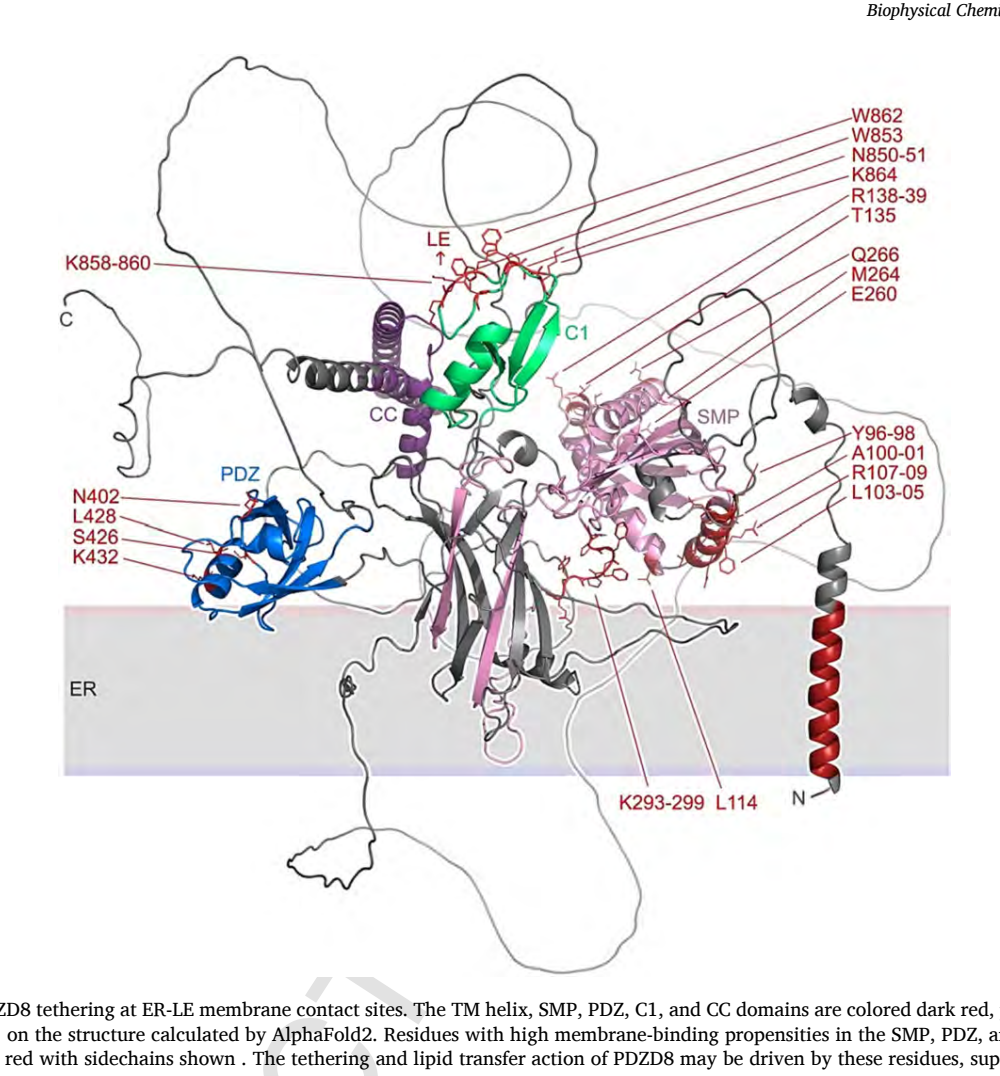

Fig. 5. Model of PDZD8 tethering at ER-LE membrane contact sites. The TM helix, SMP, PDZ, C1, and CC domains are colored dark red, pink, blue, green and purple, respectively, on the structure calculated by AlphaFold2. Residues with high membrane-binding propensities in the SMP, PDZ, and C1 domain are labelled and drawn in red with sidechains shown . The tethering and lipid transfer action of PDZD8 may be driven by these residues, supported by its interaction with Rab7-GTP. The positions of the membrane binding elements in this model have not been restrained with regards to lipid interactions and are not correctly positioned with respect to the bilayer. The ER membrane in which the membrane-spanning helix is shown but not the *trans*-membrane recognized by the peripheral membrane domains. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

$$[1\alpha : \mathrm{ER}] - \left(\mathrm{SMP}^{\mathrm{PI(4,5)P2}} - \mathrm{PDZ} - \mathrm{C1}^{850,851,853,858-860,862,864} : \mathrm{LE}^{\mathrm{PI\,4P,PS}} -\right.$$

This script describes the membrane tethering action of PDZD8, which binds directly to lipids in the LE, as well as indirectly through Rab7, two copies of which bind its CC region.

3.6. Intrinsically disordered MARCKS protein

The myristoylated alanine-rich C-kinase substrate (MARCKS) protein is one of several intrinsically disordered proteins that cluster negatively charged lipids on the plasma membrane *via* electrostatic interactions [128]. Related proteins GAP43 and CAP23 have been identified to carry out similar functions, and all three are substrates for protein kinase C, suggesting a specialized signaling role for these proteins [129]. MARCKS can also release PI(4,5)P$_2$ in response to calcium/calmodulin (CaM), which wraps around its basic effector domain (ED) [130], as seen in a crystalized complex where its F157 and L159 residues pack against CaM residues F92 and F12, respectively [131]. Thus, MARCKS acts as a PI(4,5)P$_2$ capacitor capable of releasing free lipid molecules in response to calcium. MARCKS uses its ED to engage the plasma membrane and cluster three PI(4,5)P$_2$ molecules [132,133], with MODA predicting additional residues that interact with the membrane near its myristoylated N-terminus (Fig. 6). Thus, the COMPOSEL scripts of these two states can be written as:

$$\left(\mathrm{D}^{\mathrm{Myr}-1} : \mathrm{PM} - \mathrm{ED}^{147-166:\mathrm{PI(4,5)P2}}\right)_i + \mathrm{CaM} \rightarrow \left(\mathrm{D}^{\mathrm{Myr}-1} : \mathrm{PM}\right) - < \mathrm{ED}^{\underline{157;159}} : \mathrm{CaM}^{\underline{93;12}} >$$

These states are shown in Fig. 6 along with an AlphaFold2 model that predicts that full-length MARCKS is unstructured. The effector sequence that interacts with calmodulin or membrane surfaces can be phosphorylated by PKC at serines 159, 163 and 170, which disrupts these interactions [134] and serve as PIP-stops [22,46], shifting the protein to a largely membrane-free state stabilized only weakly by the N-terminal myristoylated motif:

$$\left(\mathrm{D}^{\mathrm{Myr}-1} : \mathrm{PM} - \mathrm{ED}^{147-166:\mathrm{PI(4,5)P2}}\right)_i + \mathrm{PKC} \rightarrow \mathrm{D}^{\mathrm{Myr}-1} - \mathrm{ED}^{\mathrm{PO}-159,163,170}$$

3.7. Fatty aldehyde dehydrogenase

Proteins and lipids in cells are protected from damage by reactive aldehydes by the action of a family of 24 aldehyde dehydrogenases [135]. Of these, FALDH converts long-chain fatty aldehydes (FAL) into their corresponding fatty acid (FA) molecules. The dimeric protein is normally anchored into the ER, or in the case of the longer variant isoform, into peroxisomes, by a single C-terminal TM helix [136]. The palmitoylation of Cys425 of human FALDH has been identified by mass spectrometry in HEK293 cells [137] and would also insert into the membrane.

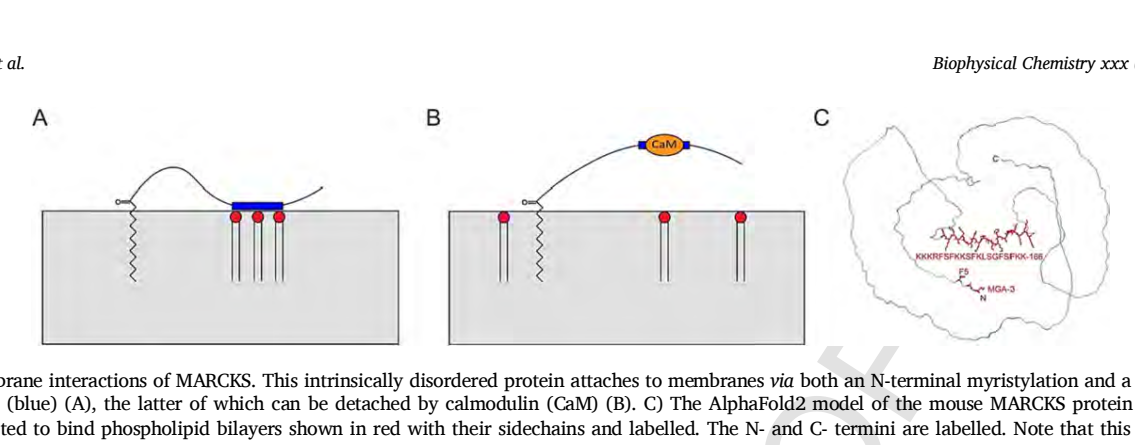

Fig. 6. Membrane interactions of MARCKS. This intrinsically disordered protein attaches to membranes *via* both an N-terminal myristylation and a basic effector domain (ED) (blue) (A), the latter of which can be detached by calmodulin (CaM) (B). C) The AlphaFold2 model of the mouse MARCKS protein with the elements predicted to bind phospholipid bilayers shown in red with their sidechains and labelled. The N- and C- termini are labelled. Note that this model of this disordered protein is of low confidence, as are the resulting membrane binding residue predictions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

MODA analysis of the crystal structure of human FALDH [138] as well as full length models from AlphaFold2 and Rosetta reveals membrane binding sites in the bridging domain that also forms a dimer interface, and a pair of short helices preceding the C-terminal TM helix (Fig. 7). The C-terminal helical extension (αCt) found in the long variant is positively charged, as is the cytosolic juxtamembrane region of both isoforms. The membrane interactions of residues M91, F453 and F454 are also predicted by PPM3, affirming of their key roles. Together these sites coordinate access to lipid substrates and passage into the substrate funnel, which harbours a lipid molecule in the crystal structure. Mutations that compromise FALDH activity lead to abnormal metabolism of lipids including sphingosine 1-phosphate, and cause the metabolic disorder Sjögren–Larsson Syndrome. These include pathogenic mutations such as R423H [139] and K447E [140], which lead to this autosomal recessively inherited neurocutaneous disorder and would compromise the membrane interactions of this enzyme.

The short and long FALDH isoforms found in different subcellular compartments are:

$$\left(\mathrm{AldD} - \left(\mathrm{BD}^{\mathrm{Pal}-425}\right)^2 - \mathrm{GH}\right)_i - [1\alpha : \mathrm{ER}]$$

$$\left(\mathrm{AldD} - \left(\mathrm{BD}^{\mathrm{Pal}-425}\right)^2 - \mathrm{GH}\right)_i - [1\alpha : \mathrm{PO}] - \alpha\mathrm{Ct}$$

where homodimerization occurs *via* the BD, and PO stands for the peroxisome membrane, which is engaged by the C-terminal α helix of the long isoform, and Cys425 is modified with a palmitoyl group (Pal).

Including the peroxisome membrane interacting residues of the long FALDH isoform generates the COMPOSEL script:

$$\left(\mathrm{AldD} - \left(\mathrm{BD}^{88,91,92,\mathrm{Pal}-425,432,434,435,438}\right)^2 - \mathrm{GH}^{i,445,446,449,450,453,454}\right)_i - [1\alpha$$

The specific lipid ligands are not known, but acidic phospholipids in both leaflets of the peroxisome membrane are likely given the basic character of the interacting sites.

The reaction catalyzed by FALDH involves a substrate molecule from the ER being bound by the catalytic domain, where it is cleaved and the product released into the local environment. This can be expressed as:

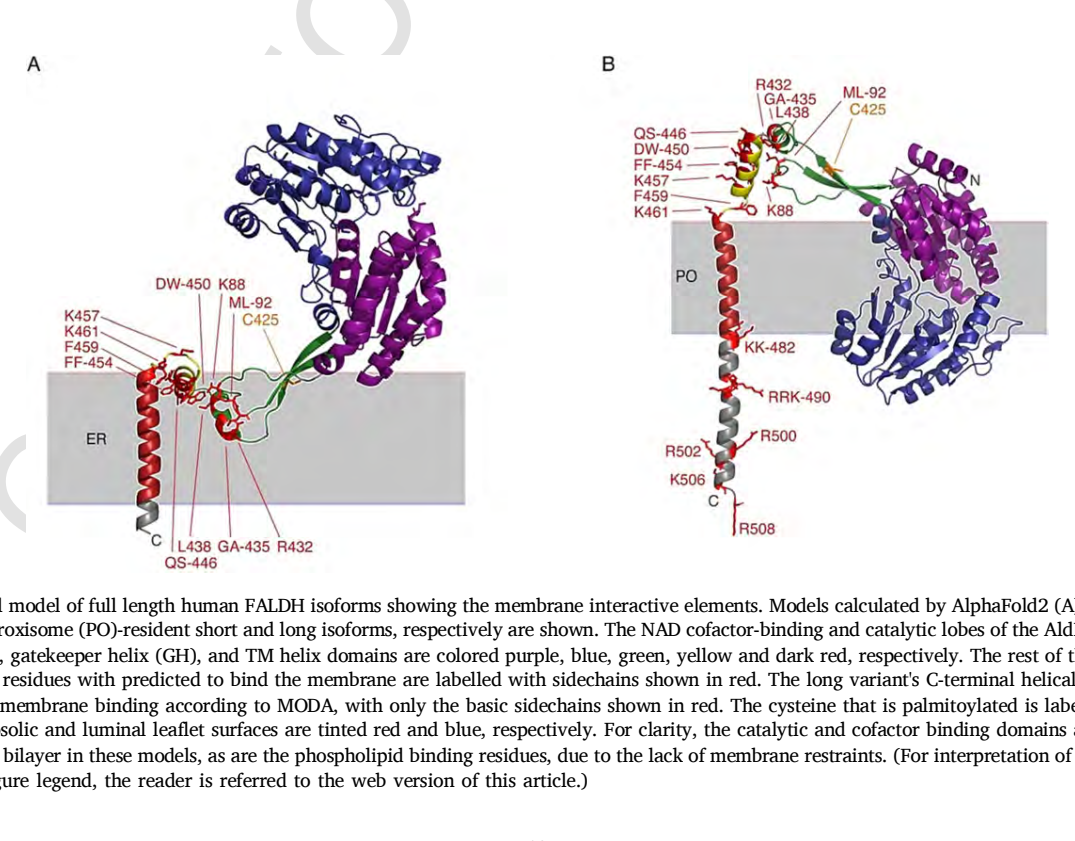

Fig. 7. Structural model of full length human FALDH isoforms showing the membrane interactive elements. Models calculated by AlphaFold2 (A) and Rosetta (B) of the ER and peroxisome (PO)-resident short and long isoforms, respectively are shown. The NAD cofactor-binding and catalytic lobes of the AldH domain, bridging domain (BD), gatekeeper helix (GH), and TM helix domains are colored purple, blue, green, yellow and dark red, respectively. The rest of the backbone ribbon is gray. The residues with predicted to bind the membrane are labelled with sidechains shown in red. The long variant's C-terminal helical end from K481-R508 is entirely membrane binding according to MODA, with only the basic sidechains shown in red. The cysteine that is palmitoylated is labelled and colored orange. The cytosolic and luminal leaflet surfaces are tinted red and blue, respectively. For clarity, the catalytic and cofactor binding domains are placed incorrectly within the bilayer in these models, as are the phospholipid binding residues, due to the lack of membrane restraints. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

$$\left(\mathrm{AldD}^{\mathrm{FAL}} - \mathrm{BD}^2 - \mathrm{GH}\right)_i - [1\alpha : \mathrm{ER}] \rightarrow \left(\mathrm{AldD} - \mathrm{BD}^2 - \mathrm{GH}\right)_i - [1\alpha : \mathrm{ER}] + \mathrm{FA}$$

3.8. Diacylglycerol kinase

Of the ten human diacylglycerol (DG) kinases, the epsilon isofom is unique in containing a TM helix. DGKε is found in the ER and plasma membranes [141,142], and carries out ATP-dependent phosphorylation of DG species to produce phosphatidic acid (PA). Its catalytic domain interfaces with a pair of C1 domains, these project into the cytosol [143] and influence DAG binding and fatty acyl specificity [144]. The AlphaFold2 model predicts that the TM helix orients away from the first C1 domain, while the second C1 domain packs against the conjoined catalytic and accessory domains (Fig. 8). A signal peptide is predicted by PrediSi although the resulting cleavage at Thr37 has not been verified experimentally. This would cut the predicted transmembrane helix in half, leaving it a monotopic protein that could explain its partial liberation from membranes with high salt or high pH [141]. However, a transmembrane orientation is supported by *in vitro* protein synthesizing system [143] and seems more likely given the position of the putative cut site inside the bilayer. Hence we infer that DGKε is a type 3 rather than type 1 membrane protein.

Based on MODA analysis of the AlphaFold2 and Rosetta models the first C1 domain interacts with membranes *via* the DLF-68, SQP-71 and LQ-83 motifs. These orient towards the TM helix and the major predicted membrane binding site of the catalytic domain, and could conceivably swival to sit on the inner leaflet. Together they form a broad surface that could influence catalytic rate and specificity for lipid substrates and highly curved membranes [145], while also engaging the PA product [146]. Curvature sensing could be further influenced by DGKε dimerization, which is maintained in gentle detergent and does not require the TM domain [147].

The DGKε protein name in a short COMPOSEL script including is:

$$[1\alpha] - (\mathrm{C1} - \mathrm{C1} - \mathrm{DAGKc})_i$$

Including the predicted lipid interacting residues gives the longer script:

$$\alpha1 - [1\alpha] - \left(\mathrm{C1}^{66-67,69-71,82-83} - \mathrm{C1} - \mathrm{DAGKc}^{358-364,394-396,398-399,401-40}\right.$$

The reaction carried out by DGKε on its preferred DG substrate bearing 18:0/20:4 acyl chains in the plasma membrane is:

$$[1\alpha : \mathrm{PM}] - (\mathrm{C1} - \mathrm{C1} - \mathrm{DAGKc} : \mathrm{DG}(18:0/20:4))_i + \mathbf{ATP} \rightarrow [1\alpha : \mathrm{PM}] - (\mathrm{C1} - \mathrm{C1} - \mathrm{DAGKc} : \mathrm{PA})_i + \mathbf{ADP}$$

3.9. Glutathione S-transferase protein

The glutathione S-transferase (GST) superfamily contains 23 human members including two with C-terminal TM helices that anchor into the mitochondrial outer membrane (MOM). Of these, the ganglioside-induced differentiation-associated protein 1 (GDAP1) is well-characterized but lacks canonical GST activity and glutathione binding capacity [148]. Although its natural ligand remains unclear, membrane association is critical for its ability to influence mitochondrial morphology and ER contact sites [149].

The crystal structure of GDAP1's core domain bound to hexadecanoic acid suggests that FA molecules are its ligands [150] (Fig. 9A). Analysis of the AlphaFold2 model of full length GDAP1 indicates a proximal membrane binding elements within the GST domain (N64, P66, Y155, and R159) and an α helical linker that precedes a single C-terminal TM helix (Fig. 9). Additionally, nearby residues P59 and L60 display significant membrane binding propensities in the apo-structures and could interact with lipids. A long loop is not visible in the crystal structures but is bordered by an exposed helical extension, and may be dynamic and could bind membranes, although this compacts and is not predicted to bind membranes in the full-length models. Altogether the predicted phospholipid docking residues circumscribe a surface that could sit on the bilayer surface to provide mitochondrial targeting [151]. Such interactions could contribute to the observed autoinhibition of GDAP1's GST activity and tubulation of liposomes that mimic mitochondrial outer membranes [152].

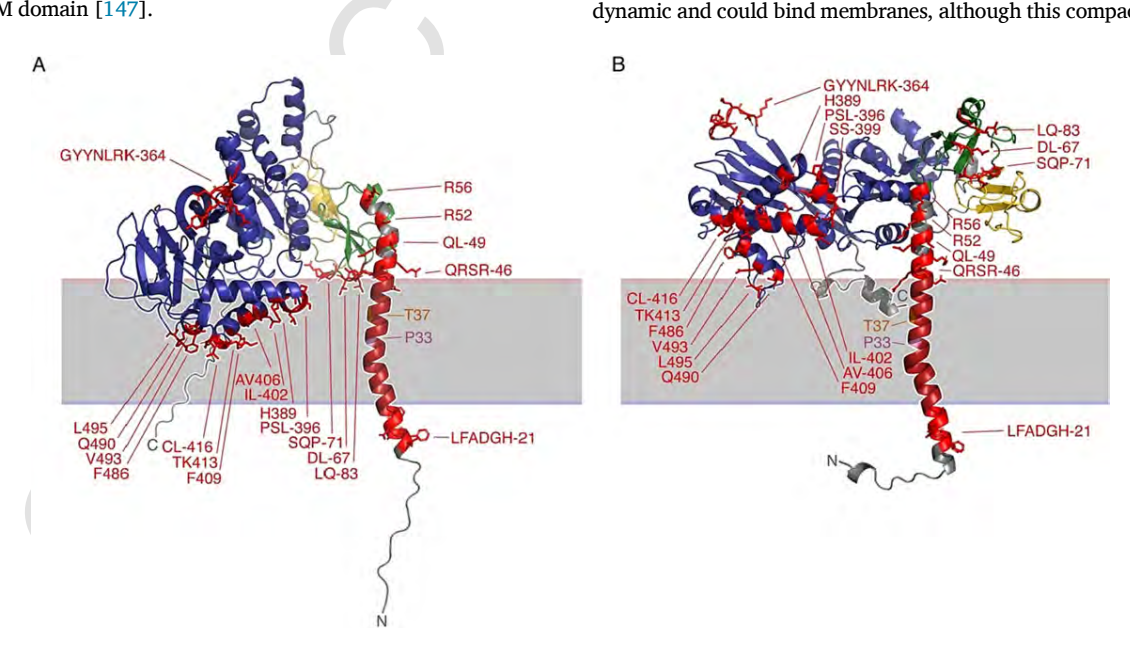

Fig. 8. Plasma membrane interactions in the DGKε structure. The TM helix, first and second C1 domains and DAG kinase catalytic (DAGKc) domain are colored dark red, green, yellow and blue, respectively, in these images of the AlphaFold2 (A) and Rosetta (B) models. The remainder of the backbone is gray, and the N- and C-termini are labelled. The residues in the peripheral domains with significant membrane binding propensity are labelled and colored red with their sidechains shown, and the proline and threonine residues in the TM helix that represent bend and putative cleavage sites are colored pink and orange, respectively. The inner and outer leaflet surfaces are tinted red and blue, respectively. For clarity, the catalytic and C-terminus are placed incorrectly within the bilayer, as are phospholipid binding residues, due to the lack of membrane restraints used to calculate these models. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

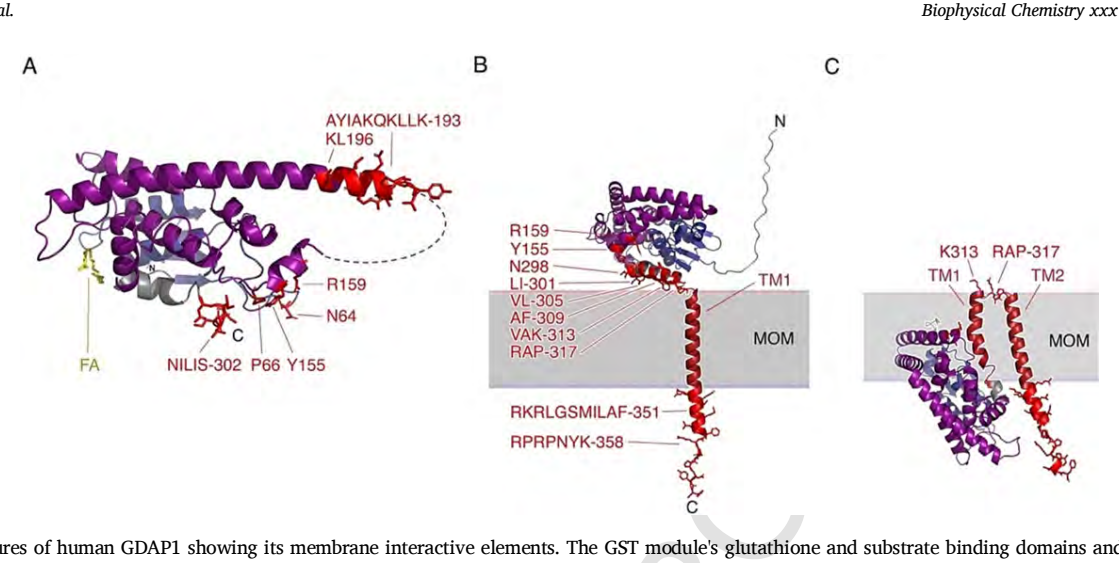

Fig. 9. Structures of human GDAP1 showing its membrane interactive elements. The GST module's glutathione and substrate binding domains and TM helix are colored blue, purple, and dark red, respectively, on (A) the crystal structure of the core GST module bound to hexadecenoic acid, which is colored yellow and labelled "FA" (PDB: 7aia) [150], (B) the AlphaFold2-based and (C) Rosetta models of the full-length protein in the mitochondrial outer membrane. The rest of the backbone ribbon is gray. The peripheral residues displaying membrane binding propensities of at least 30 are drawn in red as sticks and labelled. The N- and C-termini are labelled. For clarity, the soluble and membrane binding elements of these bilayer models are not placed correctly as no membrane restraints were used to predict the structural models. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Additional membrane binding poses can be considered. The Rosetta-derived model (Fig. 9C) indicates that an additional TM helix from residues 292 to 312 packs against the C-terminal helix, in agreement with sequence-based predictors such as HMMTOP, HTP and Phobius but not Memsat or Octopus. However, this pose is impossible for a tail anchored (TA)-protein with a cytosolic GST domain. Hence, we infer that the AlphaFold2 model (Fig. 9B) is more likely.

The isolated and crystallized GST domain forms monomers [148] as well as dimers [150]. The dimer interface involves C88 on the surface away from the membrane, with an Y29F substitution affects the dimerization of GDAP1 [150]. Dimerization is not required for activity although it could affect membrane interactions.

Mutations in GDAP1 implicated in Charcot–Marie–Tooth disease, a common inherited peripheral polyneuropathy, could alter the membrane interaction. For example, mutations linked to neuropathies positioned to affect membrane interactions include P153L, A156G, T157P, R161H, N297K, T307A, R310E, T322M, G327D, A336S, L344R, M347V and N356D in the Inherited Neuropathy Variant Browser database [153].

The models of GDAP1 predicted by AlphaFold2 and Rosetta differ, containing one or two TM helices near the C-terminus that position the GST domain either inside or outside the cytosol, respectively. The corresponding COMPOSEL scripts are:

$$\mathrm{GST}_i - [1\alpha : \mathrm{MOM}] \ \text{or} \ \mathrm{GST}_o - [2\alpha : \mathrm{MOM}]$$

Including the predicted membrane binding residues of the former state along with those of intervening helical linker (L) that interfaces with the membrane gives the following COMPOSEL script for ligand-bound GDAP1:

$$\mathrm{GST}_i^{64,66,155,159} : \mathrm{FA} - \left(\alpha\mathrm{L}^{298,300-301,304-305,308-309,311-313,315-317}\right) - [1\alpha : \mathrm{MOM}] - \mathrm{Ct}\alpha$$

Including the interactions described in these scripts could be used to predict more realistic models of full length GDAP1. In particular this would exclude the models that position the core domain incorrectly either inside or on the wrong side of the membrane (Fig. 9B, C).

3.10. BAM complexes that assemble β barrels

Outer membranes surround Gram-negative bacteria, mitochondria and chloroplast, and are packed with β barrel proteins. In bacteria the β barrel assembly machine (BAM) inserts and folds β barrels into this asymmetric membrane, which contains lipopolysaccharides in the external leaflet and phospholipids in the internal leaflet. The core subunit in *E. coli* is the β barrel protein BamA, which contains a N-terminal series of five polypeptide transport associated (Potra) domains. The resting state of BamA protein and the four lipoprotein subunits BamB, BamC, BamD and BamE in detergent has been resolved by X-ray crystallography [154]. The structure of their complex with a folding substrate in native outer membrane extracted using styrene maleic acid lipid particle (SMALP) has been determined by cryo-electron microscopy (cEM) [19]. Analysis of these assemblies predicts potential membrane interaction sites in all subunits of the free (Fig. 10A) and complexed states (Fig. 10B), with the MODA data shown in Fig. 10C. The N-terminal cysteines are acylated and insert into the bilayer but are not visible in the structures, nor are some loops and dynamic elements. We propose that the multiple membrane binding surfaces are differentially engaged during the assembly and operation of the machine, as indicated by changes in their MODA profiles.

The BamA protein structure includes a signal peptide, which is cleaved at residue 21. Its membrane interacting W205 and W206 residues [155] are within the third Potra domain and are facing the periplasmic leaflet, while the β barrel fold sits inside the bacterial outer membrane (OM). This yields the COMPOSEL script:

$$\mathrm{SP}_{1-21} - \left(\mathrm{Potra} - \mathrm{Potra} - \mathrm{Potra}^{205,206} - \mathrm{Potra} - \mathrm{Potra}\right)_p - [16\beta : \mathrm{OM}]$$

The BAM assembly in its resting state contains with peripherally inserted residues consistently predicted in the models and structures (Fig. 10) enable the expected structural poses [154,156,157] and insertion of palmitoyl groups. For example, the bound BamB lipoprotein subunit, which forms a β-propeller repeat fold known as PQQ, can be described in COMPOSEL as:

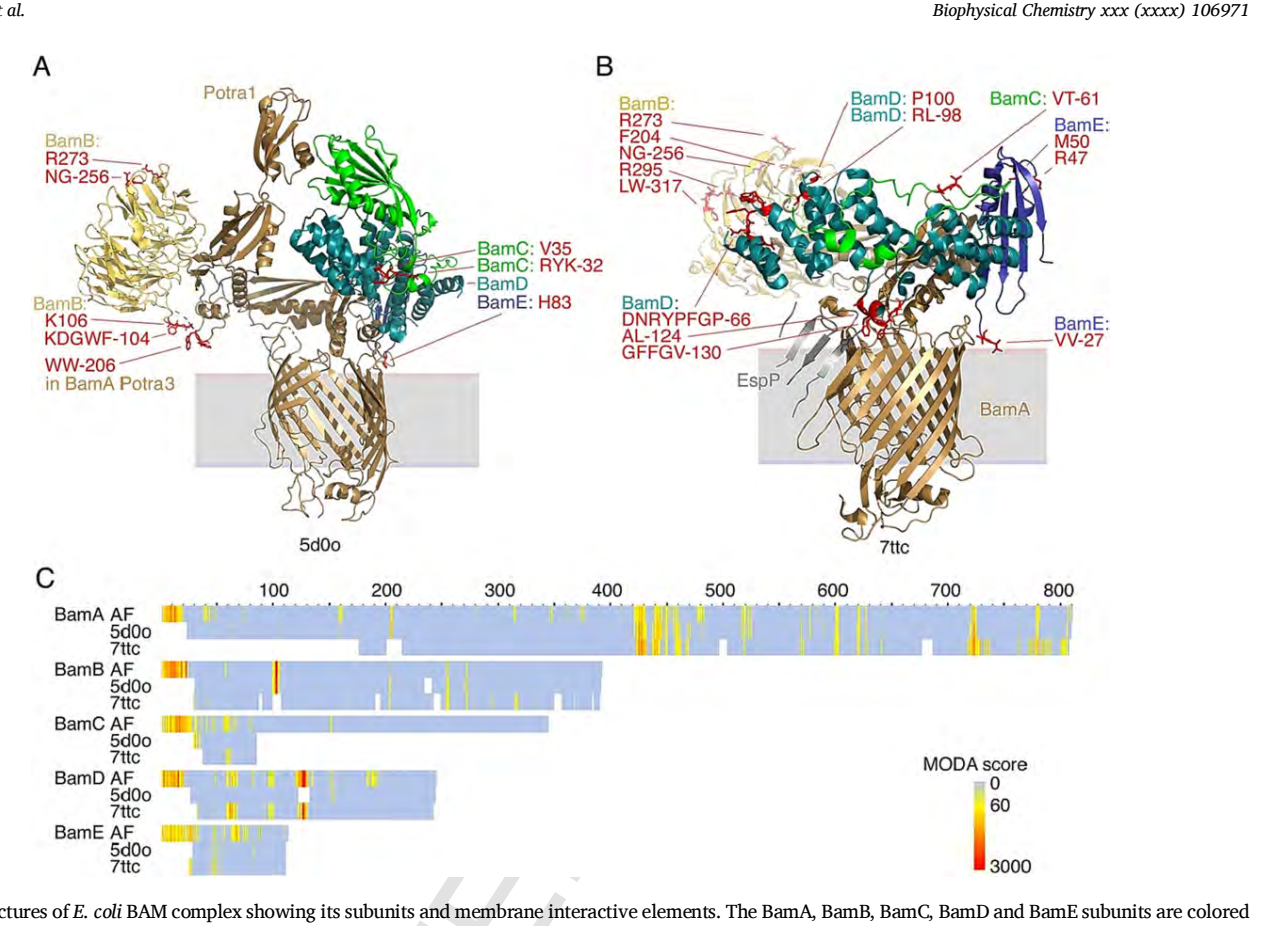

Fig. 10. Structures of *E. coli* BAM complex showing its subunits and membrane interactive elements. The BamA, BamB, BamC, BamD and BamE subunits are colored beige, yellow, green, teal, and blue, respectively, in (A) the crystal structure of the entire BAM assembly [154] and (B) its complex with the EspP folding substrate [19]. The resolved β strands of the EspP barrel substrate are in gray. The peripheral BAM residues (except for those in the BamA β barrel) which display membrane binding propensities of at least 30 are drawn in red as sticks and labelled. The membrane is shown with a slab with the extracellular and periplasmic leaflets edged in blue and red, respectively, with the PDB entry number shown below. C. The MODA scores of all the residues of each BAM subunit in the two structures and AlphaFold2 (AF) monomer models are shown in the heatmap. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

$$\mathrm{SP}_{1-19} - \left(\mathrm{PQQ}^{\mathrm{Pal}-20,100-104,106,204,255-256,273} : \mathrm{OM}\right)_p$$

The formation of the intermediate complex with a folding substrate yields conformational changes in the BAM structure [19]. This results in significantly higher membrane binding propensities in the folding machine intermediate with BamB described by a COMPOSEL script along with the β barrel and three other lipoprotein subunits included, as is stated below:

$$[\mathrm{BamA} : \mathrm{OM}] : \left(\mathrm{BamB}^{\mathrm{Pal}-20,100-104,106,203-205,255-257,271-273} : \mathrm{BamC}^{\mathrm{Pal}-25}\right.$$

Here the superscripted numbers indicate some of the residues in the folding intermediate that are predicted to be membrane interactive during the assembly process. It is also possible to use COMPOSEL scripts to indicate the protein:protein interactions. For example, the antiparallel pairing between G424 and F426 in the first strand of the BamA barrel and F1300 and Y1298 in the β signal strand of EspP, respectively, [19] are shown below with interacting peptide elements underlined and separated in order by semicolons:

$$\left[< \mathrm{BamA}^{\underline{\beta1C(424(CO);426(NH))}} : \mathrm{EspP}^{\underline{\beta S:1300(NH);1298(CO)}} >: \mathrm{OM}\right] - (\mathrm{BamB} : \mathrm{BamC} : \mathrm{BamD} : \mathrm{BamE})_p$$

Hence, COMPOSEL scripts are useful for communicating protein-protein and protein-membrane interactions, with the peptide or lipid ligand positioned after the colon of the contacting residue and in underlined or plain text to distinguish them. As such, such scripts represent simplified ways to convey the complex information within structures of large assemblies, allowing the reader to focus on the key element of any particular interaction.

3.11. Adhesion G-protein-coupled receptors

The largest superfamily of membrane proteins are the G-protein-coupled receptors (GPCRs). They are defined by the presence of seven TM helices and signal inside the cell through C-terminal interactions with G-proteins that bind and hydrolyze GTP. Their N-terminal extracellular domains are diverse. Those of adhesion G protein-coupled receptors (aGPCRs) contain a variable number of modules including the GPCR autoproteolysis-inducing (GAIN) domain with an internal GPCR Proteolysis Site (GPS) [158]. When cleaved the new N-terminal stalk can act as an intramolecular agonist. The autoproteolytic mechanism of the ~320 residue GAIN domain has been revealed by X-ray crystallography studies of its bi-lobed structure [158]. Nonetheless the detailed mechanisms of this subfamily of 33 GPCRs remain poorly understood, although their interactions with lipids including PS [159] and proteins [160] are thought to promote cell-to-cell and cell-to-matrix interactions.

Structures have been elucidated of ADGRF1 protein (Fig. 11A), a receptor for the brain-enriched lipid metabolite synaptamide [161]. This endogenous ligand is likely to bind between the two GAIN subdomains near residues Q511, N512 and Y513, whereupon it induces conformational changes and signals downstream through β-arrestin [162]. The structure of cleaved ADGRF1 bound to G-proteins elucidated by cEM reveals a lysophosphatidylcholine molecule bound in an interhelical cleft [163]. A series of five G-protein complexes reveals consistent interactions for signaling and activation by the tethered stalk peptide [164] (Fig. 11B). This aGPCR contains an amphipathic C-terminal helix (αC) and is also palmitoylated on cysteines 704 and 836, which also mediate interactions with the plasma membrane.

Once activated by autoproteolysis ADGRF1 forms a signaling complex with three G-proteins. In the case of the G$_{\alpha12}$ complex (Fig. 11B) this involves contacts between its residues L367, L371 and L376 and the receptor residues I686, L769, and I795, as underlined in the script below. The G$_{\alpha12}$ subunit contains myristoylated G3 and palmitoylated C4 residues that also mediate membrane interactions [165]. The G$_{\gamma2}$ subunit is isoprenylated at C68, which along with R62 engages plasma membrane inner leaflet lipids including phosphatidylethanolamine (PE) [166]. Hence a COMPOSEL script for auto-activated, cleaved ADGRF1 focussing on its complex with G-proteins is:

$$(\mathrm{NTD} - \mathrm{SEA} - \mathrm{GAIN})_e - [7\alpha : \mathrm{PM}]^{\underline{<686,769,795}} : \left(\mathrm{G}_{\alpha12}{}^{\underline{367,371,374}>;\mathrm{Myr}-3,\mathrm{Pal}-}\right.$$

3.12. Caveolins

The COMPOSEL script for ADGRF1 interacting with plasma membrane is as follows:

$$\begin{aligned}&(\mathrm{NTD}-\mathrm{SEA}-\mathrm{GAIN})_e\\&-\left[7\alpha^{\mathrm{Pal}-704,\mathrm{Pal}-836}:\mathrm{PM}\right]\\&-\left(\alpha\mathrm{C}^{843,846,847,848,850,852-854,856,857}\right)_i\end{aligned}$$

The COMPOSEL script for ADGRF1 indicating our discovery here of putative membrane binding residues on the novel N-terminal structural domain is:

$$\begin{aligned}&\left(\mathrm{NTD}^{3,6-7,10,14,19,68,71,73-74,76}-\mathrm{SEA}-\mathrm{GAIN}\right)_e\\&-\left[7\alpha^{\mathrm{Pal}-704,\mathrm{Pal}-836}:\mathrm{PM}\right]-\alpha\mathrm{C}_i\end{aligned}$$

Caveolae are small invaginations found in plasma membranes of many cell types, and play roles in signaling, intracellular trafficking and lipid homeostasis. Humans express three caveolins, which are monotopic proteins that oligomerize to form ~15 nm diameter disc-shaped pores, as resolved by cEM [3]. Their central helices align into a pinwheel-like ring about an inner β barrel, with significant membrane binding propensities on the outer rim and inner hub of all 11 subunits, which assemble into caveolae with polyhedral symmetry (Fig. 12). The N-terminus is predicted to be disordered and is not evident in the cEM structure, while cysteines 133, 143, and 156 in the membrane-facing surface are palmitoylated. Caveolins displace hundreds of lipids and cluster with PI(4,5)P_2, PS, and cholesterol [167] as their disc shapes insert deep into the cytoplasmic leaflet and curve membranes. The following COMPOSEL script identifies the phospholipid bilayer docking residues predicted by MODA in the perimeter

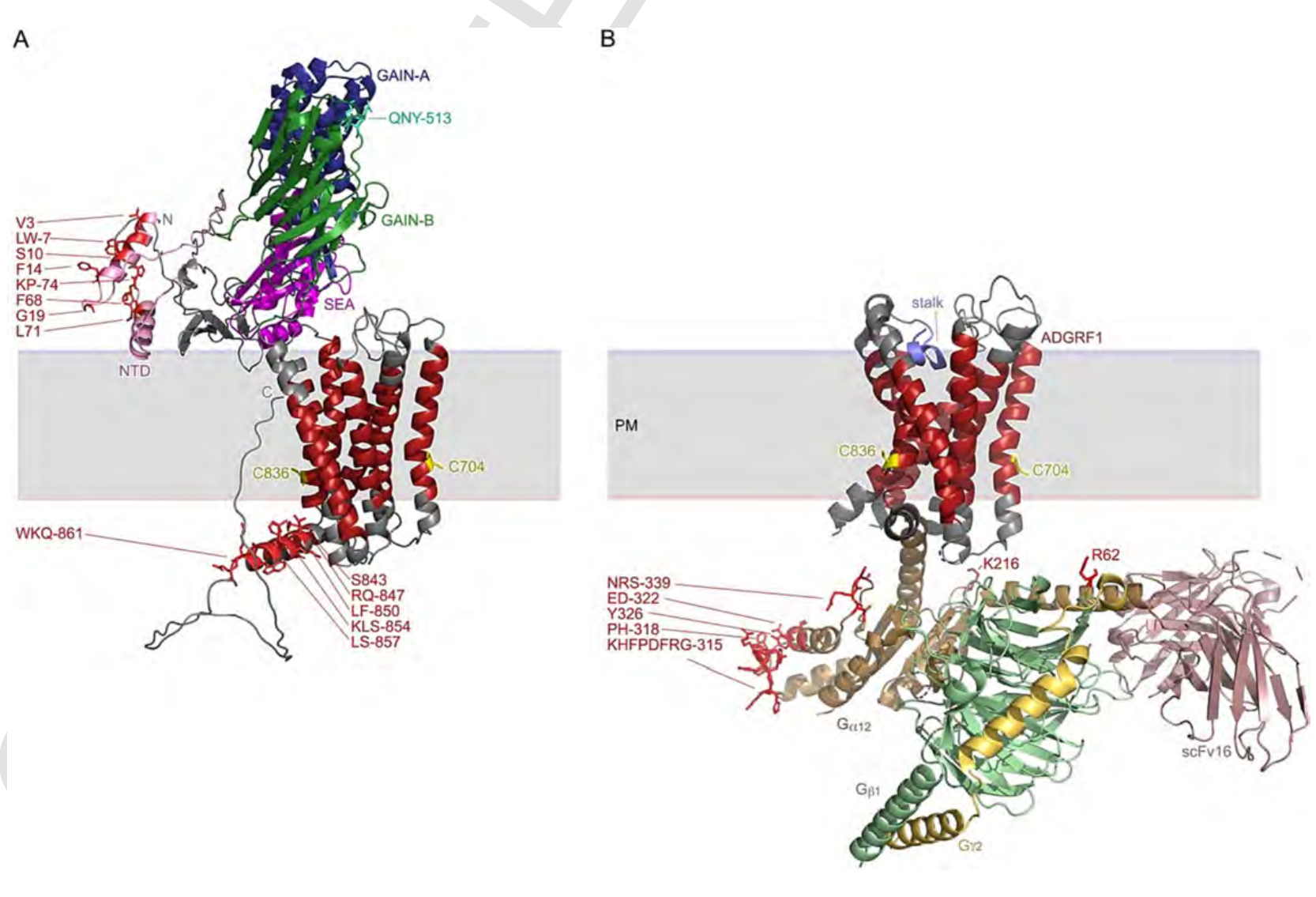

Fig. 11. Structures of the ADGRF1 indicating structural elements that engage lipid. A) An AlphaFold2 model is shown with a novel N-terminal domain (NTD), SEA domain, GAIN subdomains A and B, and 7 TM helix bundle are colored pink, magenta, blue, green and dark red, respectively. Residues with membrane binding propensities in peripheral domains are colored red with sidechains shown. The positions of the membrane binding elements have not been restrained and are not necessarily positioned accurately with respect to the membranes. B) The structure of the cleaved ADGRF1 receptor and bound to $G_{\alpha12}$, $G_{\beta1}$ and $G_{\gamma2}$ subunits and a stabilizing antibody, which are colored sand, light green, yellow and light violet, respectively, are shown. Predicted membrane binding residues in the $G_{\alpha12}$ and $G_{\gamma2}$ subunits shown in red with their sidechains and labelled, and the stalk peptide is depicted in light purple (PDB: 7wy7). For clarity the membrane binding elements are not correctly positioned with regards to the bilayer surfaces. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

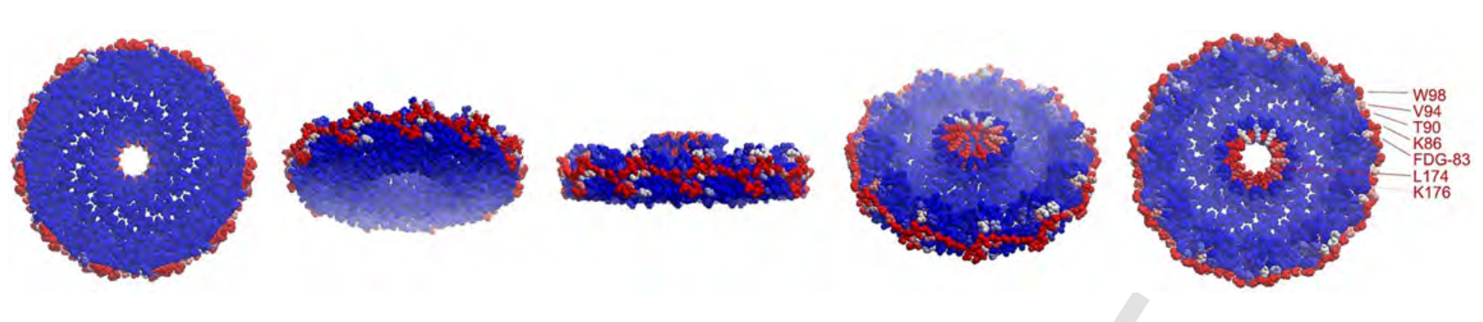

Fig. 12. Structure of the monotopic caveolin multimer (PDB: 7SC0) showing the membrane binding surfaces. Five perspectives of the human Cav1 structure are shown from the extracellular to cytoplasmic views from left to right. Residues in the eleven subunits that display significant membrane binding propensities based on MODA analysis are shown in red and are labelled, with the remainder of the surface is blue in these images generated in ICM Browser. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and center of the pinwheel fold. Additional contacts with cholesterol and the buried surface of the disc as well as disordered N-terminal residues are likely during the membrane insertion, bending and budding processes that caveolins mediate.

$$\mathrm{D}-\left(\mathrm{Cav}^{81-83,86,90,94,\mathrm{Pal}-133,\mathrm{Pal}-143,\mathrm{Pal}-156,172,174,176}:\mathrm{PM}^{\mathrm{PI(4,5)P2,PS,Chol}}\right)_i^{11}$$

4. Discussion

Collective progress in science depends on having useful classification systems, whether for a species, gene or protein. The form and function of membrane proteins are key to communication, interrogation and exploitation of the membrane code, for which a coherent classification system is proposed here. Current categorizations of membrane proteins by fold [168,169], topological class [64] or type [65] are not widely used by people or databases such as SMART [38] or Uniprot [36] and recent advances remain underappreciated. This gap is addressed here by establishment of COMPOSEL as a new classification system along with examples to show how it works with diverse protein architectures and memteins. Moreover, this facilitates identification of novel lipid binding domains and highlights the interdependency of complex reaction pathways, domain-domain connections and multimerizations. While a standardized system for annotating folded protein domains is already available [37], there is no such system for membrane proteins or memteins. Features incorporated here for full-length proteins and domains re intended to fully communicate their interactions and mechanisms and could inform the design of membrane readers, fusion machines and entirely novel assemblies.

The language we currently use for membrane proteins can be unnecessarily ambiguous and biased. "Membrane proteins" are usually assumed to refer to "transmembrane proteins", although we suggest here that this may only be true about half of the time. This study reveals that most membrane proteins contain peripheral membrane elements, monotopic domains, DLBRs or modifications that engage lipid bilayers. The latter "non-transmembrane" protein elements are thought to be more difficult to predict, and hence remain underappreciated, although automated tools like MODA exist for their rapid and convenient identification. Some transmembrane proteins also mediate peripheral interactions with a "*trans*-membrane" to form contact sites between organelles, as in the cases of Syt1, Protrudin and PDZD8. In such cases the term "transmembrane protein" can be confusing. The term "integral membrane protein" refers to transmembrane and monotopic membrane proteins while excluding peripheral or other membrane proteins that are transiently associated. However, proteins like Syt17 lack TM and monotopic domains but are tightly affixed to membranes by multiple palmitoylated cysteines and two C2 domains [106], and may actually be integral membrane proteins. Hence the overarching classifier of "membrane protein" is more accurate, particularly when combined with a description of the component interactions. This can be framed in the COMPOSEL language, which is the first structure/function-based system for unambiguous and universal classification for all proteins that concisely specifies the number and type of membrane interactive units including peripheral membrane and TM domains, disordered regions and lipidation sites.

Further advances in AI/ML could allow the complete universe of protein structures to be more accurately modelled [54,56] with inclusion of membrane interaction and regulatory modification data. The new restraints could originate from the MODA program, which is freely available and generates membrane binding propensity data in seconds, as well as public PTM databases, allowing communication of physiological functions of proteins and memteins. Obtaining validated lipidon specificities remains laborious although advances such as nanodiscs could be used to screen and cross-validate interactions with organelle membrane sections [18]. By refining structural models with restraints for multiple specific lipid ligands and protein docking surfaces in realistic bilayers could yield more accurate, complete and dynamic models of memteins. Indeed, more realistic computational approaches based on lipid compositions and dimensions that mirror those in biology are already becoming available [170]. Programs such as HADDOCK already incorporate such restraints and have been used to calculate structures of dozens of bilayer-bound proteins [46,84–91], providing a potential pathway forward.

COMPOSEL is designed to be an adaptable and scalable system that can be extended to describe complex disease mechanisms. The functions of most proteins, particularly those that serve as hubs, bridges and assembly points, involve multiple interactions. Viral proteins tend to especially polyfunctional with several binding sites that cooperate. This complexity can be challenging to describe and follow. For example, the Gag protein binds nanoclusters of PI(4,5)P_2 in the plasma membrane through residues including K30 and K32 of its myristoylated matrix (MA) domain [171–174]. It also oligomerizes in part through its capsid (CA) domain residues W316 and M317, and packages RNA through its nucleocapsid (NC) domains to facilitate budding of HIV particles. The state of the membrane bound Gag dimer complex with RNA can be simply stated as:

$$\begin{aligned}&\left(\mathrm{MA}^{\mathrm{Myr}-2;30,32}:\mathrm{PM}^{\mathrm{PI(45)P2}}\right)\\&-\left\{\mathrm{CA}^{\underline{316,317}}\right\}^2-\{NC:RNA\}^2\surd\end{aligned}$$

The membrane interactions and pathological rearrangements of multidomain proteins and intrinsically disordered regions can be covered. For example, the prion protein Prnp contains cleavable N- and C-terminal signal peptides and engages membrane lipids through DLBRs and its Prp domain, with conformational rearrangement leading to progression of a variety of prion diseases. Due to mutations or spontaneous β fibril formation the cellular form (C) converts into the scrapie (Sc) form with more extensive interactions with lipids including phosphatidylethanolamine (PE) [33,175]. This conversion by the cleaved Prnp protein can be described as:

$$\mathrm{D}^{99,122}-\mathrm{PrP\,(C)}^{169,222-228;\mathrm{GPI}-230}:\mathrm{PM}\rightarrow\mathrm{PrP\,(Sc)}^{117-124,154-157,160-161,173-17}$$

Multicomponent interactions of proteins and lipids are the norm in membranes. Unlike soluble proteins that float in three dimensional

spaces, membrane proteins are restricted to and fold in the two dimensional spaces of bilayers, hence they are more likely to form homooligomeric structures, with lipids often stabilizing their interfaces [176]. Although predicting structures of multimeric assemblies remains difficult, progress is being made. For example deep learning predictions are being developed for membrane protein complexes [177] and MemSTATS allows symmetries in membrane proteins and their quaternary complexes to be characterized [178]. Formation of such oligomeric complexes and their many lipid interactions can be challenging to describe but is key for understanding function, with COMPOSEL allowing their features to be more easily articulated.

The extended biological code that governs high-fidelity formation of DNA, RNA, protein and membrane structures of the cell can be succinctly described in COMPOSEL. For example, the following script captures how the human *EEA1* gene is expressed from chromosome 12q22, with codons including the five below being transcribed and translated into the RRHHCR-1375 motif that contacts the PI3P headgroup. When folded this element is key to recognition of the PI3P, PS and PC-containing lipidon that encodes the membrane identity of the early endosome. These four levels of the genetic and membrane code can be simply conveyed as:

$$\begin{aligned}&12\mathrm{q}22\left\{EEA1^{cgg\ \ cat\ \ cac\ \ tgc\ \ cga}\right\}\\&\rightarrow\mathrm{mRNA}\left\{EEA1^{CGG\ \ CAU\ \ CAC\ \ UGC\ \ CGA}\right\}\\&\rightarrow\mathrm{EEA1}^{1370-1375}\rightarrow\mathrm{EE}^{\mathrm{PI\,3P,PS}}\end{aligned}$$

An organism, tissue and cell in which biomolecular interactions takes place can also be specified in COMPOSEL. For example, infection of a *homo sapiens* (Hs) lung epithelial cell (LEPC) with SARS-CoV-2 virus (SC2) involves binding of the peptidase domain residues M82 and Y83 in the dimeric human ACE2 receptor and plasma membrane lipids of a host cell by F486 in the receptor binding motif (RBM) and residues K444 and V445 in the receptor binding domain (RBD) of the viral spike protein, respectively. This extracellular encounter is mediated by the trimeric spike which also contains an N-terminal domain (NTD) and is anchored in the viral membrane by a TM helix and ten palmitoylated cysteines within its C-terminal extension (Cte) [34,35,179]. The complex formed by the ACE2 dimer and plasma membrane of a human host cell and the trimerized spike RBD of the virus can be summarized as:

$$\begin{aligned}&\mathrm{Hs}\left(\mathrm{LEPC}\left(\mathrm{ACE2}\left(\mathrm{PD}^{\underline{82,83}}\right)_e^{2}-[1\alpha:\mathrm{PM}]\right)\right):\\&\mathrm{SC2}\left(\mathrm{Spike}\left(\left(\mathrm{NTD}-\mathrm{RBD}^{444,445;\underline{486}}\right)_e^{3}\right.\right.\\&\left.\left.-[1\alpha:\mathrm{VM}]-\left(\mathrm{Cte}^{10-\mathrm{Pal}}\right)_i\right)\right)\end{aligned}$$

In this way the complexity of biological codes, architectures or interactions within any organism, cell, virus or biomacromolecule can be described in a universally accessible language. As an intuitive nomenclature intended for people to read and write, we hope that this system will be considered, used and improved by the research community. It is geared to be useable as an extendable and compressible programming language in order to develop more realistic models of multicomponent proteins in membrane environments. Ultimately, this could provide a deeper understanding of how biological information is encoded and works at all levels of resolution to aid in teaching, experimental design and therapeutic intervention.

Funding

This work was supported by NSERC Discovery Grant (RGPIN- 2018-04994) and Campus Alberta Innovates Program (#RCP-12-002C) grants to MO.

Data availability statement

Not applicable.

CRediT authorship contribution statement

Michael Overduin : Investigation, Writing – review & editing, Conceptualization, Methodology, Funding acquisition, Resources, Supervision. **Troy A. Kervin :** Investigation, Writing – review & editing. **Zachary Klarenbach :** Investigation, Writing – review & editing. **Trixie Rae C. Adra :** Investigation, Writing – review & editing. **Rakesh K. Bhat :** Investigation, Writing – review & editing.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Michael Overduin reports financial support was provided by Natural Sciences and Engineering Research Council of Canada. Michael Overduin reports financial support was provided by Campus Alberta Innovates Program.

Acknowledgments

We thank Molsoft LLC for making the ICM program and MODA available, and those who have offered suggestions on the COMPOSEL system and syntax, with further input from the research community welcomed.

References

[1] L. Fagerberg, et al., Prediction of the human membrane proteome, Proteomics 10 (6) (2010) 1141–1149.
[2] L.S. Wenz, et al., Biogenesis and folding of β-barrel membrane proteins, Cell Cycle 13 (2) (2014) 169–170.
[3] J.C. Porta, et al., Molecular architecture of the human caveolin-1 complex, Sci. Adv. 8 (19) (2022) p. eabn7232.
[4] S.S. Sanders, et al., Curation of the mammalian Palmitoylome indicates a pivotal role for Palmitoylation in diseases and disorders of the nervous system and cancers, PLoS Comput. Biol. 11 (8) (2015) e1004405.
[5] T. Meinnel, C. Dian, C. Giglione, Myristoylation, an ancient protein modification mirroring Eukaryogenesis and evolution, Trends Biochem. Sci. 45 (7) (2020) 619–632.
[6] W.W. Epstein, et al., Quantitation of prenylcysteines by a selective cleavage reaction, Proc. Natl. Acad. Sci. U. S. A. 88 (21) (1991) 9668–9670.
[7] M.J. Blanden, et al., Efficient farnesylation of an extended C-terminal C, J. Biol. Chem. 293 (8) (2018) 2770–2785.
[8] M.H. Gíslason, et al., Prediction of GPI-anchored proteins with pointer neural networks, Curr. Res. Biotechnol. 3 (2021) 6–13.
[9] H. Jiang, et al., Protein Lipidation: occurrence, mechanisms, biological functions, and enabling technologies, Chem. Rev. 118 (3) (2018) 919–988.
[10] M. Overduin, T.A. Kervin, The phosphoinositide code is read by a plethora of protein domains, Expert Rev. Proteomics 18 (7) (2021) 483–502.
[11] C.I. Pataki, et al., Proteomic analysis of monolayer-integrated proteins on lipid droplets identifies amphipathic interfacial α-helical membrane anchors, Proc. Natl. Acad. Sci. U. S. A. 115 (35) (2018) E8172–E8180.
[12] A. Katuwawala, B. Zhao, L. Kurgan, DisoLipPred: accurate prediction of disordered lipid binding residues in protein sequences with deep recurrent networks and transfer learning, Bioinformatics 38 (1) (2021) 115–124.
[13] J. Liu, J.R. Faeder, C.J. Camacho, Toward a quantitative theory of intrinsically disordered proteins and their function, Proc. Natl. Acad. Sci. U. S. A. 106 (47) (2009) 19819–19823.
[14] S.K. Burley, et al., Protein data Bank: a comprehensive review of 3D structure holdings and worldwide utilization by researchers, educators, and students, Biomolecules 12 (10) (2022) 1425.
[15] A. Oberai, et al., A limited universe of membrane protein families and folds, Protein Sci. 15 (7) (2006) 1723–1734.
[16] M. Overduin, et al., Transmembrane membrane readers form a novel class of proteins that include peripheral phosphoinositide recognition domains and viral spikes, Membranes (Basel) 12 (11) (2022) 1161.
[17] K. Tunyasuvunakool, et al., Highly accurate protein structure prediction for the human proteome, Nature 596 (7873) (2021) 590–596.
[18] M. Overduin, et al., Structures and dynamics of native-state transmembrane protein targets and bound lipids, Membranes (Basel) 11 (6) (2021) 451.

[19] M.T. Doyle, et al., Cryo-EM structures reveal multiple stages of bacterial outer membrane protein folding, Cell 185 (7) (2022) 1143–1156.e13.
[20] M. Overduin, M. Esmaili, Memtein: the fundamental unit of membrane-protein structure and function, Chem. Phys. Lipids 218 (2019) 73–84.
[21] K. Corin, J.U. Bowie, How physical forces drive the process of helical membrane protein folding, EMBO Rep. 23 (3) (2022) e53025.
[22] T.A. Kervin, M. Overduin, Regulation of the phosphoinositide code by phosphorylation of membrane readers, Cells 10 (5) (2021) 1205.
[23] T.A. Kervin, B.C. Wiseman, M. Overduin, Phosphoinositide recognition sites are blocked by metabolite attachment, Front. Cell Dev. Biol. 9 (2021) 690461.
[24] A. Chatzigoulas, Z. Cournia, DREAMM: a web-based server for drugging protein-membrane interfaces as a novel workflow for targeted drug design, Bioinformatics 38 (24) (2022) 5449–5451.
[25] A. Chatzigoulas, Z. Cournia, Predicting protein-membrane interfaces of peripheral membrane proteins using ensemble machine learning, Brief. Bioinform. 23 (2) (2022) bbab518.
[26] C.A. Schramm, et al., Knowledge-based potential for positioning membrane-associated structures and assessing residue-specific energetic contributions, Structure 20 (5) (2012) 924–935.
[27] M.A. Lomize, et al., OPM database and PPM web server: resources for positioning of proteins in membranes, Nucleic Acids Res. 40 (Database issue) (2012) D370–D376.
[28] A.L. Lomize, S.C. Todd, I.D. Pogozheva, Spatial arrangement of proteins in planar and curved membranes by PPM 3.0, Protein Sci. 31 (1) (2022) 209–220.
[29] I. Kufareva, et al., Discovery of novel membrane binding structures and functions, Biochem. Cell Biol. 92 (6) (2014) 555–563.
[30] M. Lenoir, et al., Membrane and protein interactions of the Pleckstrin homology domain superfamily, Membranes (Basel) 5 (4) (2015) 646–663.
[31] C. Bissig, et al., Viral infection controlled by a calcium-dependent lipid-binding module in ALIX, Dev. Cell 25 (4) (2013) 364–373.
[32] J.A. Bryant, et al., Structure of dual-BON domain protein DolP identifies phospholipid binding as a new mechanism for protein localization, Elife 9 (2020) e62614.
[33] M. Overduin, H. Wille, D. Westaway, Multisite interactions of prions with membranes and native nanodiscs, Chem. Phys. Lipids 236 (2021) 105063.
[34] M. Overduin, T.A. Kervin, A. Tran, Progressive membrane-binding mechanism of SARS-CoV-2 variant spike proteins, iScience 25 (8) (2022) 104722.
[35] A. Tran, T.A. Kervin, M. Overduin, Multifaceted membrane binding head of the SARS-CoV-2 spike protein, Curr. Res. Struct. Biol. 4 (2022) 146–157.
[36] Consortium, U., UniProt: the universal protein knowledgebase in 2021, Nucleic Acids Res. 49 (D1) (2021) D480–D489.
[37] V.P. Waman, et al., The Genome3D consortium for structural annotations of selected model organisms, Methods Mol. Biol. 2165 (2020) 27–67.
[38] I. Letunic, S. Khedkar, P. Bork, SMART: recent updates, new developments and status in 2020, Nucleic Acids Res. 49 (D1) (2021) D458–D460.
[39] S.M. Reynolds, et al., Transmembrane topology and signal peptide prediction using dynamic bayesian networks, PLoS Comput. Biol. 4 (11) (2008) e1000213.
[40] L. Käll, A. Krogh, E.L. Sonnhammer, Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server, Nucleic Acids Res. 35 (Web Server issue) (2007) W429–W432.
[41] K. Hiller, et al., PrediSi: prediction of signal peptides and their cleavage positions, Nucleic Acids Res. 32 (Web Server issue) (2004) W375–W379.
[42] F. Teufel, et al., SignalP 6.0 predicts all five types of signal peptides using protein language models, Nat. Biotechnol. 40 (7) (2022) 1023–1025.
[43] Y. Xie, et al., GPS-lipid: a robust tool for the prediction of multiple lipid modification sites, Sci. Rep. 6 (2016) 28249.
[44] M. Blanc, F.P.A. David, F.G. van der Goot, SwissPalm 2: protein S-Palmitoylation database, Methods Mol. Biol. 2009 (2019) 203–214.
[45] S. Li, et al., pCysMod: prediction of multiple cysteine modifications based on deep learning framework, Front. Cell Dev. Biol. 9 (2021) 617366.
[46] M. Lenoir, et al., Phosphorylation of conserved phosphoinositide binding pocket regulates sorting nexin membrane targeting, Nat. Commun. 9 (1) (2018) 993.
[47] P.V. Hornbeck, et al., PhosphoSitePlus, 2014: mutations, PTMs and recalibrations, Nucleic Acids Res. 43 (Database issue) (2015) D512–D520.
[48] H.M. Berman, et al., The Protein Data Bank, Nucleic Acids Res. 28 (1) (2000) 235–242.
[49] A.L. Lomize, et al., Membranome: a database for proteome-wide analysis of single-pass membrane proteins, Nucleic Acids Res. 45 (D1) (2017) D250–D255.
[50] D. Kozma, I. Simon, G.E. Tusnády, PDBTM: protein data Bank of transmembrane proteins after 8 years, Nucleic Acids Res. 41 (Database issue) (2013) D524–D529.
[51] T.D. Newport, M.S.P. Sansom, P.J. Stansfeld, The MemProtMD database: a resource for membrane-embedded protein structures and their lipid interactions, Nucleic Acids Res. 47 (D1) (2019) D390–D397.
[52] S.H. White, Biophysical dissection of membrane proteins, Nature 459 (7245) (2009) 344–346.
[53] M. Mirdita, et al., ColabFold: making protein folding accessible to all, Nat. Methods 19 (6) (2022) 679–682.
[54] M. Varadi, et al., AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models, Nucleic Acids Res. 50 (D1) (2022) D439–D444.
[55] J. Yang, et al., The I-TASSER suite: protein structure and function prediction, Nat. Methods 12 (1) (2015) 7–8.
[56] M. Baek, et al., Accurate prediction of protein structures and interactions using a three-track neural network, Science 373 (6557) (2021) 871–876.
[57] L. Zimmermann, et al., A completely Reimplemented MPI bioinformatics toolkit with a new HHpred server at its Core, J. Mol. Biol. 430 (15) (2018) 2237–2243.
[58] W.L. DeLano, The PyMOL Molecular Graphics System, Version 1.8, Schrödinger LLC, 2014. http://www.pymol.org.
[59] Hallgren , J., et al., DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. bioRxiv.
[60] L. Dobson, I. Reményi, G.E. Tusnády, The human transmembrane proteome, Biol. Direct 10 (2015) 31.
[61] S. Hayat, et al., Inclusion of dyad-repeat pattern improves topology prediction of transmembrane β-barrel proteins, Bioinformatics 32 (10) (2016) 1571–1573.
[62] H. Viklund, A. Elofsson, Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information, Protein Sci. 13 (7) (2004) 1908–1917.
[63] D. Kozma, G.E. Tusnády, TMFoldWeb: a web server for predicting transmembrane protein fold class, Biol. Direct 10 (2015) 54.
[64] G. Blobel, Intracellular protein topogenesis, Proc. Natl. Acad. Sci. U. S. A. 77 (3) (1980) 1496–1500.
[65] K.C. Chou, H.B. Shen, MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM, Biochem. Biophys. Res. Commun. 360 (2) (2007) 339–345.
[66] K.A. Diederichs, S.K. Buchanan, I. Botos, Building better barrels - β-barrel biogenesis and insertion in Bacteria and mitochondria, J. Mol. Biol. 433 (16) (2021) 166894.
[67] I. Sinning, M.A. McDowell, Cryo-EM insights into tail-anchored membrane protein biogenesis in eukaryotes, Curr. Opin. Struct. Biol. 75 (2022) 102428.
[68] M. Spiess, T. Junne, M. Janoschke, Membrane protein integration and Topogenesis at the ER, Protein J. 38 (3) (2019) 306–316.
[69] A.M. Liaci, F. Förster, Take me home, protein roads: structural insights into signal peptide interactions during ER translocation, Int. J. Mol. Sci. 22 (21) (2021) 11871.
[70] X. Wu, C. Cabanos, T.A. Rapoport, Structure of the post-translational protein translocation machinery of the ER membrane, Nature 566 (7742) (2019) 136–139.
[71] S. Itskanov, E. Park, Structure of the posttranslational sec protein-translocation channel complex from yeast, Science 363 (6422) (2019) 84–87.
[72] L. Smalinskaitė, et al., Mechanism of an intramembrane chaperone for multipass membrane proteins, Nature611(7934) (2022)161-166.
[73] A. Sundaram, et al., Substrate-driven assembly of a translocon for multipass membrane proteins, Nature611(7934) (2022) 167–172.
[74] I.D. Pogozheva, A.L. Lomize, Evolution and adaptation of single-pass transmembrane proteins, Biochim. Biophys. Acta Biomembr. 1860 (2) (2018) 364–377.
[75] M.K. Araya, Y. Zhou, A.A. Gorfe, Remodeling of the plasma membrane by surface-bound protein monomers and oligomers: the critical role of intrinsically disordered regions, J. Membr. Biol 255 (6) (2022) 651–663.
[76] A. Hentschel, R.P. Zahedi, R. Ahrends, Protein lipid modifications--More than just a greasy ballast, Proteomics 16 (5) (2016) 759–782.
[77] J.M. Sanderson, The association of lipids with amyloid fibrils, J. Biol. Chem. 298 (8) (2022) 102108.
[78] M. Blum, et al., The InterPro protein families and domains database: 20 years on, Nucleic Acids Res. 49 (D1) (2021) D344–D354.
[79] A.H. Larsen, et al., Membrane-binding mechanism of the EEA1 FYVE domain revealed by multi-scale molecular dynamics simulations, PLoS Comput. Biol. 17 (9) (2021) e1008807.
[80] R. Brunecky, et al., Investigation of the binding geometry of a peripheral membrane protein, Biochemistry 44 (49) (2005) 16064–16071.
[81] S.A. Lee, et al., Targeting of the FYVE domain to endosomal membranes is regulated by a histidine switch, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 13052–13057.
[82] E. Fahy, et al., Update of the LIPID MAPS comprehensive classification system for lipids, J. Lipid Res. (2009) S9–14 50 Suppl S9-14.
[83] F. Sievers, D.G. Higgins, Clustal omega for making accurate alignments of many protein sequences, Protein Sci. 27 (1) (2018) 135–145.
[84] J. Roel-Touris, B. Jiménez-García, A.M.J.J. Bonvin, Integrative modeling of membrane-associated protein assemblies, Nat. Commun. 11 (1) (2020) 6210.
[85] R.K. Koppisetti, et al., Ambidextrous binding of cell and membrane bilayers by soluble matrix metalloproteinase-12, Nat. Commun. 5 (2014) 5552.
[86] T.C. Marcink, et al., MT1-MMP binds membranes by opposite tips of its β propeller to position it for Pericellular proteolysis, Structure 27 (2) (2019) 281–292.e6.
[87] S.H. Prior, et al., Charge-triggered membrane insertion of matrix Metalloproteinase-7, supporter of innate immunity and tumors, Structure 23 (11) (2015) 2099–2110.
[88] R.K. Koppisetti, Y.G. Fulcher, S.R. Van Doren, Fusion peptide of SARS-CoV-2 spike rearranges into a wedge inserted in Bilayered micelles, J. Am. Chem. Soc. 143 (33) (2021) 13205–13211.
[89] Z. Fang, et al., Inhibition of K-RAS4B by a unique mechanism of action: stabilizing membrane-dependent occlusion of the effector-binding site, Cell Chem. Biol. 25 (11) (2018) 1327–1336.e4.
[90] F. Dancea, K. Kami, M. Overduin, Lipid interaction networks of peripheral membrane proteins revealed by data-driven micelle docking, Biophys. J. 94 (2) (2008) 515–524.
[91] M. Lenoir, et al., Structural basis of wedging the Golgi membrane by FAPP pleckstrin homology domains, EMBO Rep. 11 (4) (2010) 279–284.
[92] H. Khan, et al., Structural basis of human PDZD8-Rab7 interaction for the ER-late endosome tethering, Sci. Rep. 11 (1) (2021) 18859.

[93] S. Saita, et al., Promotion of neurite extension by protrudin requires its interaction with vesicle-associated membrane protein-associated protein, J. Biol. Chem. 284 (20) (2009) 13766–13777.
[94] F. Matsuzaki, et al., Protrudin serves as an adaptor molecule that connects KIF5 and its cargoes in vesicular transport during process formation, Mol. Biol. Cell 22 (23) (2011) 4602–4620.
[95] D.V. Pantakani, et al., Oligomerization of ZFYVE27 (Protrudin) is necessary to promote neurite extension, PLoS One 6 (12) (2011) e29584.
[96] C. Raiborg, et al., Repeated ER-endosome contacts promote endosome translocation and neurite outgrowth, Nature 520 (7546) (2015) 234–238.
[97] J.E. Gil, et al., Phosphoinositides differentially regulate protrudin localization through the FYVE domain, J. Biol. Chem. 287 (2012) 41268–41276.
[98] M. Shirane, et al., Protrudin and PDZD8 contribute to neuronal integrity by promoting lipid extraction required for endosome maturation, Nat. Commun. 11 (1) (2020) 4576.
[99] T.G. Kutateladze, et al., Phosphatidylinositol 3-phosphate recognition by the FYVE domain, Mol. Cell 3 (6) (1999) 805–811.
[100] T. Kutateladze, M. Overduin, Structural mechanism of endosome docking by the FYVE domain, Science 291 (5509) (2001) 1793–1796.
[101] J.J. Dumas, et al., Multivalent endosome targeting by homodimeric EEA1, Mol. Cell 8 (2001) 947–958.
[102] A.A. Tokmakov, A. Kurotani, K.I. Sato, Protein pI and intracellular localization, Front. Mol. Biosci. 8 (2021) 775736.
[103] A. Mishra, et al., Structural basis for Rab GTPase recognition and endosome tethering by the C2H2 zinc finger of early endosomal autoantigen 1 (EEA1), Proc. Natl. Acad. Sci. U. S. A. 107 (24) (2010) 10866–10871.
[104] J. Callaghan, et al., The endosome fusion regulator early-endosomal autoantigen 1 (EEA1) is a dimer, Biochem. J. 338 (Pt 2) (1999) 539–543.
[105] E. Münzberg, M. Stein, Structure and dynamics of mono- vs. doubly Lipidated Rab5 in membranes, Int. J. Mol. Sci. 20 (19) (2019) 4773.
[106] A.C. Wolfes, C. Dean, The diversity of synaptotagmin isoforms, Curr. Opin. Neurobiol. 63 (2020) 198–209.
[107] W. Vennekate, et al., Cis- and trans-membrane interactions of synaptotagmin-1, Proc. Natl. Acad. Sci. U. S. A. 109 (27) (2012) 11037–11042.
[108] E.R. Chapman, How does synaptotagmin trigger neurotransmitter release? Annu. Rev. Biochem. 77 (2008) 615–641.
[109] U. Heindel, M.F. Schmidt, M. Veit, Palmitoylation sites and processing of synaptotagmin I, the putative calcium sensor for neurosecretion, FEBS Lett. 544 (1–3) (2003) 57–62.
[110] R. Kang, et al., Presynaptic trafficking of synaptotagmin I is regulated by protein palmitoylation, J. Biol. Chem. 279 (48) (2004) 50524–50536.
[111] P. Ruchala, et al., Insights into the structure and molecular topography of the fatty acylated domain of synaptotagmin-1, Biochim. Biophys. Acta Biomembr. 1861 (3) (2019) 677–684.
[112] Q. Zhou, et al., Architecture of the synaptotagmin-SNARE machinery for neuronal exocytosis, Nature 525 (7567) (2015) 62–67.
[113] A. Honigmann, et al., Phosphatidylinositol 4,5-bisphosphate clusters act as molecular beacons for vesicle recruitment, Nat. Struct. Mol. Biol. 20 (6) (2013) 679–686.
[114] J. Bai, W.C. Tucker, E.R. Chapman, PIP2 increases the speed of response of synaptotagmin and steers its membrane-penetration activity toward the plasma membrane, Nat. Struct. Mol. Biol. 11 (1) (2004) 36–44.
[115] D.Z. Herrick, et al., Position of synaptotagmin I at the membrane interface: cooperative interactions of tandem C2 domains, Biochemistry 45 (32) (2006) 9668–9674.
[116] E. Rufener, et al., Membrane-bound orientation and position of the synaptotagmin C2B domain determined by site-directed spin labeling, Biochemistry 44 (1) (2005) 18–28.
[117] S.B. Nyenhuis, A. Thapa, D.S. Cafiso, Phosphatidylinositol 4,5 bisphosphate controls the cis and trans interactions of Synaptotagmin 1, Biophys. J. 117 (2) (2019) 247–257.
[118] A. Radhakrishnan, et al., The Ca2 + affinity of synaptotagmin 1 is markedly increased by a specific interaction of its C2B domain with phosphatidylinositol 4,5-bisphosphate, J. Biol. Chem. 284 (38) (2009) 25749–25760.
[119] X. Zhang, J. Rizo, T.C. Südhof, Mechanism of phospholipid binding by the C2A-domain of synaptotagmin I, Biochemistry 37 (36) (1998) 12395–12403.
[120] A.F. Davis, et al., Kinetics of synaptotagmin responses to Ca2 + and assembly with the core SNARE complex onto membranes, Neuron 24 (2) (1999) 363–376.
[121] Y.K. Chae, et al., Lipid binding ridge on loops 2 and 3 of the C2A domain of synaptotagmin I as revealed by NMR spectroscopy, J. Biol. Chem. 273 (40) (1998) 25659–25663.
[122] R. Li, et al., Targeting phosphatidylserine with calcium-dependent protein-drug conjugates for the treatment of Cancer, Mol. Cancer Ther. 17 (1) (2018) 169–182.
[123] E. Hui, J. Bai, E.R. Chapman, Ca2 + −triggered simultaneous membrane penetration of the tandem C2-domains of synaptotagmin I, Biophys. J. 91 (5) (2006) 1767–1777.
[124] J. Lee, J.T. Littleton, Transmembrane tethering of synaptotagmin to synaptic vesicles controls multiple modes of neurotransmitter release, Proc. Natl. Acad. Sci. U. S. A. 112 (12) (2015) 3793–3798.
[125] O.D. Bello, et al., Synaptotagmin oligomerization is essential for calcium control of regulated exocytosis, Proc. Natl. Acad. Sci. U. S. A. 115 (32) (2018) E7624–E7631.
[126] Y. Hirabayashi, et al., ER-mitochondria tethering by PDZD8 regulates ca, Science 358 (6363) (2017) 623–630.
[127] D. Jeyasimman, et al., PDZD-8 and TEX-2 regulate endosomal PI(4,5)P, Nat. Commun. 12 (1) (2021) 6065.
[128] S. Manenti, et al., Affinity purification and characterization of myristoylated alanine-rich protein kinase C substrate (MARCKS) from bovine brain. Comparison of the cytoplasmic and the membrane-bound forms, J. Biol. Chem. 267 (31) (1992) 22310–22315.
[129] T. Laux, et al., GAP43, MARCKS, and CAP23 modulate PI(4,5)P(2) at plasmalemmal rafts, and regulate cell cortex actin dynamics through a common mechanism, J. Cell Biol. 149 (7) (2000) 1455–1472.
[130] T. Porumb, et al., Calcium binding and conformational properties of calmodulin complexed with peptides derived from myristoylated alanine-rich C kinase substrate (MARCKS) and MARCKS-related protein (MRP), Eur. Biophys. J. 25 (4) (1997) 239–247.
[131] E. Yamauchi, et al., Crystal structure of a MARCKS peptide containing the calmodulin-binding domain in complex with Ca2 + −calmodulin, Nat. Struct. Biol. 10 (3) (2003) 226–231.
[132] A. Gambhir, et al., Electrostatic sequestration of PIP2 on phospholipid membranes by basic/aromatic regions of proteins, Biophys. J. 86 (4) (2004) 2188–2207.
[133] S. McLaughlin, A. Aderem, The myristoyl-electrostatic switch: a modulator of reversible protein-membrane interactions, Trends Biochem. Sci. 20 (7) (1995) 272–276.
[134] S. Tzlil, D. Murray, A. Ben-Shaul, The "electrostatic-switch" mechanism: Monte Carlo study of MARCKS-membrane interaction, Biophys. J. 95 (4) (2008) 1745–1757.
[135] V. Vasiliou, et al., Aldehyde dehydrogenases: from eye crystallins to metabolic disease and cancer stem cells, Chem. Biol. Interact. 202 (1–3) (2013) 2–10.
[136] B. Ashibe, et al., Dual subcellular localization in the endoplasmic reticulum and peroxisomes and a vital role in protecting against oxidative stress of fatty aldehyde dehydrogenase are achieved by alternative splicing, J. Biol. Chem. 282 (28) (2007) 20763–20773.
[137] M.T. Forrester, et al., Site-specific analysis of protein S-acylation by resin-assisted capture, J. Lipid Res. 52 (2) (2011) 393–398.
[138] M.A. Keller, et al., A gatekeeper helix determines the substrate specificity of Sjögren-Larsson syndrome enzyme fatty aldehyde dehydrogenase, Nat. Commun. 5 (2014) 4439.
[139] G. Carney, S. Wei, W.B. Rizzo, Sjögren-Larsson syndrome: seven novel mutations in the fatty aldehyde dehydrogenase gene ALDH3A2, Hum. Mutat. 24 (2) (2004) 186.
[140] Y. Tachibana, et al., A case of Sjögren-Larsson syndrome with minimal MR imaging findings facilitated by proton spectroscopy, Pediatr. Radiol. 42 (3) (2012) 380–382.
[141] M. Decaffmeyer, et al., Determination of the topology of the hydrophobic segment of mammalian diacylglycerol kinase epsilon in a cell membrane and its relationship to predictions from modeling, J. Mol. Biol. 383 (4) (2008) 797–809.
[142] N. Kobayashi, et al., Differential subcellular targeting and activity-dependent subcellular localization of diacylglycerol kinase isozymes in transfected cells, Eur. J. Cell Biol. 86 (8) (2007) 433–444.
[143] M.H. Nørholm, et al., Flanking residues help determine whether a hydrophobic segment adopts a monotopic or bitopic topology in the endoplasmic reticulum membrane, J. Biol. Chem. 286 (28) (2011) 25284–25290.
[144] T.B. Ware, et al., Reprogramming fatty acyl specificity of lipid kinases via C1 domain engineering, Nat. Chem. Biol. 16 (2) (2020) 170–178.
[145] J.C. Bozelli, et al., Human diacylglycerol kinase ε N-terminal segment regulates the phosphatidylinositol cycle, controlling the rate but not the acyl chain composition of its lipid intermediates, ACS Chem. Biol. 17 (9) (2022) 2495–2506.
[146] J.C. Bozelli, et al., Regulation of DGKε activity and substrate acyl chain specificity by negatively charged phospholipids, Biophys. J. 118 (4) (2020) 957–966.
[147] A.O. Dicu, et al., Role of the hydrophobic segment of diacylglycerol kinase epsilon, Biochemistry 46 (20) (2007) 6109–6117.
[148] M.R. Googins, et al., Structural and functional divergence of GDAP1 from the glutathione S-transferase superfamily, FASEB J. 34 (5) (2020) 7192–7207.
[149] D. Pla-Martín, et al., Silencing of the Charcot-Marie-tooth disease-associated gene GDAP1 induces abnormal mitochondrial distribution and affects Ca2 + homeostasis by reducing store-operated Ca2 + entry, Neurobiol. Dis. 55 (2013) 140–151.
[150] G.T.T. Nguyen, et al., Structure of the complete dimeric human GDAP1 Core domain provides insights into ligand binding and clustering of disease mutations, Front. Mol. Biosci. 7 (2020) 631232.
[151] K.M. Wagner, et al., Targeting and function of the mitochondrial fission factor GDAP1 are dependent on its tail-anchor, PLoS One 4 (4) (2009) e5160.
[152] N. Huber, et al., Glutathione-conjugating and membrane-remodeling activity of GDAP1 relies on amphipathic C-terminal domain, Sci. Rep. 6 (2016) 36930.
[153] L. Cantarero, et al., Mitochondria-lysosome membrane contacts are defective in GDAP1-related Charcot-Marie-tooth disease, Hum. Mol. Genet. 29 (22) (2021) 3589–3605.
[154] Y. Gu, et al., Structural basis of outer membrane protein insertion by the BAM complex, Nature 531 (7592) (2016) 64–69.
[155] P.J. Fleming, et al., BamA POTRA domain interacts with a native lipid membrane surface, Biophys. J. 110 (12) (2016) 2698–2709.
[156] M.G. Iadanza, et al., Lateral opening in the intact β-barrel assembly machinery captured by cryo-EM, Nat. Commun. 7 (2016) 12865.
[157] M.G. Iadanza, et al., Distortion of the bilayer and dynamics of the BAM complex in lipid nanodiscs, Commun. Biol. 3 (1) (2020) 766.
[158] D. Araç, et al., A novel evolutionarily conserved domain of cell-adhesion GPCRs mediates autoproteolysis, EMBO J. 31 (6) (2012) 1364–1378.
[159] D. Park, et al., BAI1 is an engulfment receptor for apoptotic cells upstream of the ELMO/Dock180/Rac module, Nature 450 (7168) (2007) 430–434.

[160] V.A. Jackson, et al., Super-complexes of adhesion GPCRs and neural guidance receptors, Nat. Commun. 7 (2016) 11184.
[161] J.W. Lee, et al., Orphan GPR110 (ADGRF1) targeted by N-docosahexaenoylethanolamine in development of neurons and cognitive function, Nat. Commun. 7 (2016) 13123.
[162] B.X. Huang, et al., Synaptamide activates the adhesion GPCR GPR110 (ADGRF1) through GAIN domain binding, Commun. Biol. 3 (1) (2020) 109.
[163] X. Qu, et al., Structural basis of tethered agonism of the adhesion GPCRs ADGRD1 and ADGRF1, Nature 604 (7907) (2022) 779–785.
[164] X. Zhu, et al., Structural basis of adhesion GPCR GPR110 activation by stalk peptide and G-proteins coupling, Nat. Commun. 13 (1) (2022) 5513.
[165] T.L. Jones, J.S. Gutkind, Galpha12 requires acylation for its transforming activity, Biochemistry 37 (9) (1998) 3196–3202.
[166] M.A. Noguera-Salvà, et al., Role of the C-terminal basic amino acids and the lipid anchor of the Gγ, Biochim. Biophys. Acta Biomembr. 1859 (9 Pt B) (2017) 1536–1547.
[167] Y. Zhou, et al., Caveolin-1 and cavin1 act synergistically to generate a unique lipid environment in caveolae, J. Cell Biol. 220 (3) (2021) e202005138.
[168] S. Neumann, et al., Classification of α-helical membrane proteins using predicted helix architectures, PLoS One 8 (10) (2013) e77491.
[169] S. Hermansen, D. Linke, J.C. Leo, Transmembrane β-barrel proteins of bacteria: from structure to function, Adv. Protein Chem. Struct. Biol. 128 (2022) 113–161.
[170] R.F. Alford, et al., Protein structure prediction and Design in a Biologically Realistic Implicit Membrane, Biophys. J. 118 (8) (2020) 2042–2055.
[171] C. Sumner, A. Ono, Relationship between HIV-1 gag Multimerization and membrane binding, Viruses 14 (3) (2022) 622.
[172] J. Chen, et al., HIV-1 RNA genome dimerizes on the plasma membrane in the presence of gag protein, Proc. Natl. Acad. Sci. U. S. A. 113 (2) (2016) E201–E208.
[173] Y. Wen, et al., Mechanisms of PI(4,5)P2 enrichment in HIV-1 viral membranes, J. Mol. Biol. 432 (19) (2020) 5343–5364.
[174] H. Zhao, et al., Nucleic acid-induced dimerization of HIV-1 gag protein, J. Biol. Chem. 294 (45) (2019) 16480–16493.
[175] M. Esmaili, et al., Infectious lipid-bound prion multimers in custom native nanodiscs, J. Biol. Chem. 295 (25) (2020) 8460–8469.
[176] K. Gupta, et al., The role of interfacial lipids in stabilizing membrane protein oligomers, Nature 541 (7637) (2017) 421–424.
[177] I.R. Humphreys, et al., Computed structures of core eukaryotic protein complexes, Science 374 (6573) (2021) p. eabm4805.
[178] A.A. Aleksandrova, E. Sarti, L.R. Forrest, MemSTATS: a benchmark set of membrane protein symmetries and Pseudosymmetries, J. Mol. Biol. 432 (2) (2020) 597–604.
[179] F.S. Mesquita, et al., S-acylation controls SARS-CoV-2 membrane lipid organization and enhances infectivity, Dev. Cell 56 (20) (2021) 2790–2807.e8.

CORRECTED PROOF