Nanobody-aided structural study of the activity-regulated cytoskeleton-associated protein (Arc) using synchrotron radiation and cryo-EM

Lasse Neset

This thesis is submitted in partial fulfilment of the requirements for the degree of Master of Science in Biomedical Sciences

University of Bergen

Department of Biomedicine

Faculty of Medicine

2023

Acknowledgments

I would like to start by thanking my supervisors Professor Petri Kursula and Professor Clive Bramham for allowing me to take on this project and guiding me along the way. Andrea Johana Moreno Lopez for teaching me most of the methods I have learned this year. José Miguel Godoy Muñoz for being a great partner and officemate for this study.

Thank you everyone at the Neurotargeting group, the Bramham group, and the Petri Kursula group, both in Bergen and in Oulu, for providing a great environment both socially and academically. Selina Cannon Homaei and Oda Caspara Krokengen for keeping me company this year and also giving me good advice or someone to talk to. I wish you the best of luck in the last year of your PhDs. Thank you, Anne Baumann, Ju Xu, and Sudarshan Patil for teaching me many of the fundamental methods I have used this year.

Thank you to the staff at the Umeå Center for Electron Microscopy for allowing me to stay and learn cryo-EM for five days. It was a fun and enlightening experience. The staff at the CoSAXS beamline at the MAX IV synchrotron in Lund for providing such excellent facilities and guidance. Thank you again to Petri Kursula for allowing me to travel and have these first-hand scientific learning experiences.

Finally, I would like to thank my friends and family for supporting me and motivating me throughout this year, even though my calls and visits have become less and less frequent the closer I get to the deadline. Thank you Torild and Ivar for taking such good care of me during my time in Bergen. It has made it a lot easier to be on the other side of the country as my family. I want to especially thank my mom Birgitte and dad Leif Arne for always believing in me and encouraging me to be curious all my life. Even though I might not always have appreciated it, I now understand how hard you have worked your entire lives to achieve your goals. I have struggled with motivation and not putting in maximum effort at times, and only now in the last few years do I understand how rewarding it is to work hard for something and see the results. Thank you for encouraging my dream of working in science, which I certainly would have given up by now if not for your support.

Table of contents

Abbreviations

AMPA	α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid
Arc	Activity-regulated cytoskeleton-associated protein
C11	Anti-Arc nanobody C11 binding to the C-lobe of Arc
CaMKII	Calcium/calmodulin-dependent protein kinase II
CDR	Complementarity-determining region
CNS	Central nervous system
Cryo-EM	Cryogenic electron microscopy
CTD	C-terminal domain
dArc	*Drosophila Melanogaster* Arc
DLS	Dynamic light scattering
DTT	Dithiotheitol
E5	Anti-Arc nanobody E5 binding to the peptide-binding pocket of Arc's N-lobe
Fab	Antibody fragment
Gag	Group-specific antigen protein homologous to Arc
GKAP	Guanylate kinase-associated protein
hArc	Human Arc
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
IDP	Intrinsically disordered protein
IDR	Intrinsically disordered region
IEG	Immediate early gene
LTD	Long-term depression
LTP	Long-term potentiation
MALS	Multi-angle light scattering
mArc	Mammalian Arc
MSA	Multiple sequence alignment
NabFab	Nanobody-binding antigen-binding fragment
Nb	Nanobody
NiNTA	Ni^{2+}-nitrilotriacetic acid
NMDA	N-methyl-D-aspartate
NMR	Nuclear magnetic resonance
NT	Neurotransmitter
NTD	N-terminal domain
PKC	Protein kinase C
PNS	Peripheral nervous system
PSD	Postsynaptic density
PSD95	Post-synaptic density protein 95
rArcFL7A	Rat Arc mutant with a 7-alanine substitution at the oligomerization motif
SAXS	Small-angle X-ray scattering
SDS-PAGE	Sodium dodecyl sulfate–polyacrylamide gel electrophoresis
SEC	Size-exclusion chromatography
Stg	Stargazin. A peptide subunit of AMPARs and a natural ligand of Arc
TCEP	Tris(2-carboxyethyl)phosphine
TEM	Transmission electron microscopy
Tris	Tris(hydroxymethyl)aminomethane

Summary

Arc is an immediate early gene that regulates synaptic plasticity in glutamatergic neurons. The formation of new long-term memories requires functioning Arc protein. Arc is both a protein interaction hub at the dendritic spines and is able to encapsulate its own mRNA in virus-like capsids that transfect nearby cells. Relatively little is known about the structure of the mammalian Arc protein. It consists of mainly α-helical structures that make up the N- and C-terminal domain, which are connected by a flexible linker and flanked by flexible N- and C-terminal tails. Arc is found in many oligomeric states ranging from dimers to the predicted 140-subunit capsids. This study aimed to solve the full-length structure of dimeric mammalian Arc using X-ray crystallography and single-particle cryo-EM. Nanobodies that bind to Arc with high affinity were used to stabilize and enlarge the dimeric complex. Structural information about the Arc-nanobody complex was gathered using SAXS and compared with structures predicted using AlphaFold. The results show that AlphaFold struggles to predict structures that match the low-resolution structures of Arc in solution obtained from SAXS, likely due to Arc's structural flexibility and tendency to oligomerize. The study has also provided insight into the binding dynamics of these nanobodies to Arc and highlighted their many uses in addition to structural chaperones. Optimization of the sample preparation and data collection were performed for the use in single-particle cryo-EM to solve the dimeric structure of full-length Arc, although the data collection and processing have not been finished as of this moment. This data could provide valuable new structural information about Arc, which will help better the understanding of Arc's functions and roles in disease.

1. Introduction

1.1. The central nervous system

The brain is mainly made up of two types of cells, neurons which are responsible for signal transduction, and glial cells which have many different functions such as structural support, supplying nutrients to the neurons, maintaining homeostasis, and protection from viruses and pathogens[1]. Thanks to the works of Ramon y Cajal and Golgi in the late 19th century, we know that the nervous system is composed of distinct cells. Up until then, there was an ongoing debate on whether the neurons fused to form a continuous network, or if the neurons made up a network of independent cells. This debate was finally laid to rest with the development of the electron microscope (EM) in the 1950s. Palay & Palade[2] and De Robertis & Bennett[3] were able to image the synaptic cleft between two neurons to show they were, in fact, separate cells.

The exact number of neurons and glial cells in the human brain is still under debate, but a rough estimate can be set to 100 billion neurons and at least as many glial cells[4], although some scientists estimate the number of glial cells to be 10-50 times higher[4]. Glial cells do not transmit electrical signals like neurons, but instead, they perform many roles to support the nervous system. Oligodendrocytes and Schwann cells insulate the axons of neurons in the central nervous system (CNS) and peripheral nervous system (PNS) respectively, in order to increase the transduction velocity of electrical signals[5]. Microglial cells are responsible for the immune response in the CNS. Astroglial cells or astrocytes have many roles in the CNS. They provide biochemical control of the endothelial cells that make up the blood-brain barrier, provide nutrients to neurons, maintain the extracellular ion balance, and more[6].

Neurons are generally dipolar cells consisting of dendrites, a soma, and axons. There is a huge amount of diversity in the size, shape, and function amongst neurons. For example, Purkinje cells and motor neurons look very distinctly different, because of the different roles they have in the nervous system[7]. The soma or cell body is where the nucleus of the cell is located and where most of the synthesis and metabolic activity occurs. Dendrites are where the neuron synapses with the axons of other neurons. They vary wildly in size and shape depending on the role of the neuron. Neurons generally have one axon, although some have more than one, and are responsible for transmitting the electrical signal to other neurons or muscles. Electrical signals in the neuron usually travel from the synapse to the cell body and then along the axon.

1.2. The synapse

The location where the axon of one neuron meets the dendrite of another neuron is called a synapse. Since neurons are discrete cells, the electrical signal cannot travel from one neuron to the next, at least in most animals this is true. A recent study found that in certain ctenophores, the neurons fuse to form a continuous plasma membrane[8]. Whether the nervous system in all animals evolved from a single origin, or if it originated more than once is still uncertain. Aside from that tangent, most neurons need to make use of chemical signaling with neurotransmitters (NT) to send a signal across the synapse[1].

Chemical signaling makes use of neurotransmitters to convey the signal across the synaptic cleft[9]. Synapses can be divided into the pre- (axon) and post-synaptic cell (dendrite). The pre-synaptic cell is the one that sends the signal with NTs, and the post-synaptic cell receives the signal. There is also signaling occurring in the other direction, but the majority of the signaling goes from the pre-synaptic cell to the post-synaptic cell. Following a synaptic potential to the soma, and along the axon to the axon terminal, vesicles containing a NT (for example glutamate or serotonin) are stimulated and will fuse with the pre-synaptic membrane to release its contents into the synaptic cleft. There they will activate the corresponding receptors at the membrane of the post-synaptic cell. When the NTs bind to the receptors, ion channels will open and let ions pass into the cell, reducing the resting membrane potential in the cell. If the membrane potential reaches its threshold for activation, the membrane will depolarize, and an electrical signal (synaptic potential) will move along the dendrite toward the soma[1].

1.3. Synaptic plasticity and memory

In most parts of the CNS (apart from the dentate gyrus and subventricular zone), most neurons stop being generated after birth. Some neurons might be generated up until we reach early adulthood or even later in life[10-13]. Yet we are still able to form new memories and learn new skills. The plasticity of the brain can be described as its ability to modify neuronal circuits based on the activity of the neurons. The activation or lack of activation of neurons can modify their morphology and activity, both in the short and long-term[14]. This phenomenon is complex and not entirely understood. Therefore, there is not one single mechanism we know that regulates this. Donald Hebb has been credited with first hypothesizing whether this synaptic plasticity forms the basis for long-term memory storage[15], and subsequent studies have backed this theory[16,17]. This theory called the synaptic plasticity and memory hypothesis, is considered

to be the leading theory[18,19], although there is much more to learn about the mechanisms behind memory. One of the more studied pieces of this puzzle is long-term potentiation (LTP) and long-term depression (LTD). Patterns of high-frequency stimulation can initiate LTP in certain neurons, which will increase synaptic transmission. Other patterns of more moderate stimulus can initiate LTD, in which the synapse decreases the transmission strength across the synapse.

These forms of synaptic plasticity have been most extensively studied in excitatory neurons in the hippocampus region of the brain. This region of the brain is important for the retrieval of some types of memories such as spatial memory, and the formation of new long-term memories. LTP was discovered by Terje Lømo and Timothy Bliss in the laboratory of Per Andersen in Oslo 1966[20,21]. One of the more well-described forms of LTP can be found in excitatory glutamatergic neurons and is relying on the functions of the α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) receptor and the N-methyl-D-aspartate (NMDA) receptor. In these synapses, upon activation of the presynaptic cell, glutamate-containing vesicles will fuse with the membrane and glutamate will be released into the synaptic cleft. Both AMPA receptors and NMDA receptors are glutamate-gated cation channels found at the membrane of the post-synaptic cell. AMPARs are the most rapidly acting receptors at the post-synaptic membrane. Once glutamate is bound to AMPARs, a structural change of the AMPAR is induced and opens a channel through which Na^+ ions can pass through. This causes a depolarization of the membrane potential. NMDARs are also activated by glutamate binding, and most notably allow Ca^{2+}-ions to pass the membrane into the postsynaptic cell. There is another prerequisite for opening the ion channel in addition to glutamate binding, which is the depolarization of the membrane. This causes the Mg^{2+} blocking the pore of the NMDAR ion channel to be removed. Therefore, NMDARs can be described as coincidence detectors, requiring both the activation of the pre and post-synaptic cells to induce signal transduction[1,22]. Ca^{2+} in the cell acts as a second messenger, activating multiple complexes such as Ca^{2+}/calmodulin-dependent protein kinase II (CaMKII) and protein kinase C (PKC)[23,24]. Both of these kinases phosphorylate AMPARs, which increases their conductance, thus further increasing glutamate signal transduction. The insertion of more AMPARs into the post-synaptic membrane is also induced during LTP. This all leads to the post-synaptic cell being more responsive to glutamate following LTP induction by high-frequency stimulation. This sets in motion other signaling pathways and mediators, which for example act on protein synthesis and gene transcription, leading to a permanent increase in synaptic transmission across the synapse[25]. On the other side of the coin, LTD is induced when a smaller amount of glutamate is released by the pre-synaptic cell. This modest stimulation leads to an influx of

Ca^{2+} as in LTP, but at a lower level. This leads to the endocytosis of AMPARs from the membrane and the activation of protein kinases that reduce the activity of AMPARs. Thus, the effect of glutamate on signal transduction is permanently reduced across the synapse[26]. Both LTP and LTD can be induced by NMDAR activation and the subsequent influx of Ca^{2+} in the post-synaptic cell, even though these are effects that pull in the opposite direction. What determines whether LTP or LTD should be induced is the levels of Ca^{2+}. High Ca^{2+} levels (often induced by high-frequency stimulation) cause LTP induction and lower Ca^{2+} levels (often induced by low-frequency stimulation) cause LTD induction[27,28].

This is where the main focus of the project enters the picture. Activity-regulated cytoskeleton-associated (Arc) protein is known to be involved in both of these processes and is known to be crucial for the formation of new long-term memories.

1.4. Arc function
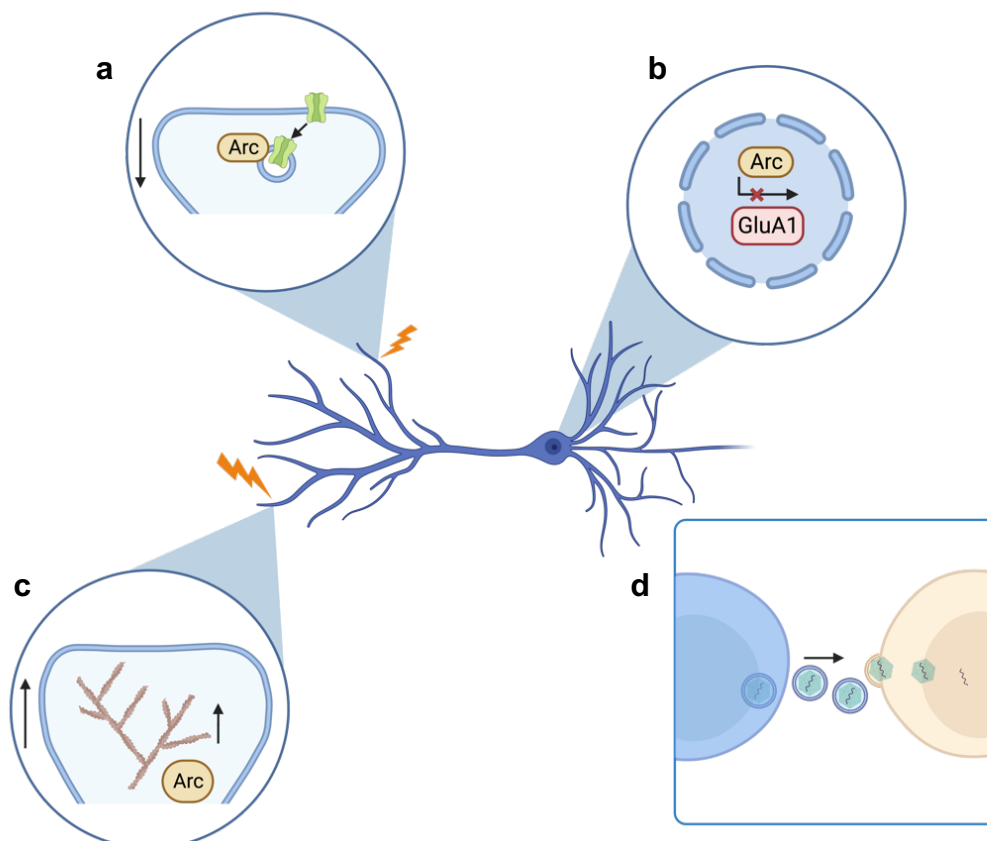
Figure 1.1: A schematic of the different roles of Arc in the neuron adapted from the 2022 review by Eriksen and Bramham[29]. (**a**) Arc's role in stimulating LTD by promoting the clathrin-mediated endocytosis of AMPARs from the post-synaptic membranes of dendritic spines. (**b**) Arc in the nucleus where it downregulates the transcription of the AMPAR subunit GluA1, resulting in a cell-wide downscaling of AMPARs. (**c**) Arc indirectly interacts with the actin cytoskeleton of active dendritic spines to stabilize newly polymerized actin filaments. (**d**) Arc is able to encapsulate its own mRNA and form large virus-like capsids, which are released by vesicles and taken up by nearby cells where they can be translated. This role in intercellular signaling and the structure and exact contents of these capsids is still a mystery.

Arc is a highly conserved protein among vertebrates. It is an immediate early gene (IEG) found mainly in glutamatergic neurons, that is rapidly expressed once the synapse is activated. IEGs are often transcription factors and are transcribed and expressed very quickly following novel stimuli. Upon activation of synapses, the mRNA of Arc is transcribed and translocated to active dendritic spines, where it is translated into the protein[30]. Arc mRNA is then subsequently translocated from the synapses to the cytosol where it is degraded[31]. Unlike a lot of IEGs, Arc is an effector protein and not a DNA-binding transcription factor. Arc is known as a master regulator of synaptic plasticity and is essential to the formation of new long-term memories. This was shown by Plath et al. when they tested the ability of Arc knockout mice to form new long-term memories. Learning and long-term memory formation were shown to be significantly reduced in Arc knockout mice, compared to wild-type mice[32]. The evidence is clear that Arc is essential for the regulation of synaptic plasticity and homeostatic scaling[33], and that dysregulation of Arc can have large consequences.

Another fascinating aspect of Arc is its ability to oligomerize and form large virus-like capsids[34]. These capsids resemble that of a retrovirus (such as HIV), and Arc has likely been repurposed from a retrotransposon. Over half of the mammalian genome is derived from viral or retrotransposon origin[35]. Most of these elements are non-coding and are not translated into proteins, although evidence suggests that these transposable elements can provide materials for the emergence of new regulatory elements and functions[36]. This includes dozens of protein-coding genes in vertebrates. Genes, many of which are expressed in the brain, originate from sequences previously encoded by retroviruses[37] and transposons[38]. The exact contents of these Arc capsids are not fully known yet, although it is known that Arc encapsulates its own mRNA, and that mRNA is necessary for the formation of the capsids[34]. The capsids are released by neurons in extracellular vesicles, containing Arc mRNA and most likely some other cargo. One theory is that this cargo would then alter the state of nearby cells, allowing Arc to not only regulate active or inactive cells but also nearby cells. Due to the effects observed when knocking out Arc, it is thought that this is a fundamental aspect of the consolidation of information[34].

Arc can also be found as two isoforms in *Drosophila Melanogaster*, however, there are significant structural differences between *Drosophila* Arc (dArc) and mammalian Arc (mArc). dArc lacks the N-terminal domain (NTD) of mArc, yet it is able to oligomerize and form large capsids of up to 240 Arc subunits which have been structurally solved using cryo-EM[39]. The N-terminal is predicted to have a helical structure which might play a role in oligomerization[40],

and the formation of the spikes on the outside of the capsid[39]. The C-terminal domain (CTD) on the other hand is more conserved across the different species, although the CTD of dArc has a tendency to oligomerize[41], whereas the CTD of mArc is thought to be monomeric in solution[42,43]. With all this taken into consideration, dArc and mArc likely originated in an evolutionary convergent manner, from two different domestication events[34].

1.5. Arc interactions

Arc regulates synaptic plasticity in several different ways[44]. One way is through its binding to Stargazin (Stg), which is a subunit of the AMPAR. Arc is localized to the Stg of AMPARs in active synapses, where it recruits other binding partners[45]. These binding partners include dynamin and endophilin which both play important roles in clathrin-mediated synaptic endocytosis[46]. Another binding partner of Arc is clathrin-adapter protein 2 (AP2) which is essential for the formation of the endocytic clathrin-coated vesicles in which the AMPARs are internalized[47]. Thus, high expression levels of Arc induce the endocytosis of AMPARs from the post-synaptic membrane, terminating LTP and inducing LTD. Homeostatic plasticity in this case can be described as the process of scaling neuronal output while maintaining the relative strength of individual synapses. Arc regulates homeostatic synaptic plasticity by scaling the surface levels of AMPARs depending on the activity of the synapses[33].

Calcium/Calmodulin-dependent protein kinase II (CaMKII) can be found in the active state (CaMKIIα) and the inactive state (CaMKIIβ). Arc was shown to be an interaction partner of CaMKIIβ, which is located at silent synapses or synapses entering LTD[48]. In the same study, Arc was proven to favor the inactive CaMKIIβ rather than the active CaMKIIα, found in active synapses or synapses in the later stage of LTP. CaMKIIβ acts as a scaffold for Arc to bind to and remain in the inactive synapse for longer. In the synapse, Arc promotes LTD by aiding the endocytosis of AMPARs as described above. Without CaMKIIβ as a binding partner, Arc will more freely diffuse from the synaptic bouton before it has a chance to interact with AMPARs. This way Arc can be targeted to inactive synapses rather than active synapses, maintaining the synaptic homeostasis.

NMDARs are an essential part of the activation of LTP, and Arc has been shown to interact with NMDARs in different ways. One way is through direct interactions with the Glu2NA and Glu2NB subunits of NMDARs[45]. Not much is known about the function of this interaction. This binding competes with the binding of Arc to AMPARs, which could be a way of

downregulating the effect of Arc in active synapses[43,49]. Another way Arc interacts with NMDARs is through its interaction with post-synaptic density protein 95 (PSD95). The post-synaptic density (PSD) is located under the postsynaptic membrane and is composed of over 200 multiprotein complexes. The proteome of the PSD consists of a set of 1000 highly conserved proteins among vertebrates. These multiprotein complexes can be divided into a hierarchy, in which PSD95 forms the prototype complex[50,51]. PSD95 is a scaffolding protein that mediates interactions with numerous synaptic proteins through its different domains (SH3 domain and guanylate kinase domain)[51]. Arc needs PSD95 to localize to the postsynaptic membrane, and Arc colocalizes with NMDARs in these multiprotein complexes[50]. The binding to PSD95 allows Arc to interact with multiple different proteins which are also recruited to the PSD by PSD95; among these are AP2, dynamin, and endophilin mentioned above. The synthesis of Arc is necessary for both the expression and consolidation of LTP[52,53]. Arc interacts with drebrin A and combined they regulate the activity of cofilin, which in turn regulates the stabilization of actin filaments in the dendritic spines[54]. This means that without the synthesis of new Arc, the stability of actin in dendritic spines will be negatively affected. Arc also plays a role in the nucleus where it associates with TIP60[55] and CREB binding protein[56], which are histone acetylases regulating transcription. Arc in the nucleus was shown to downregulate the transcription of GluA1, a subunit of AMPARs, thus supporting a synaptic downscaling of the dendrites[56].

Post-translational modifications (PTMs) of Arc such as ERK-dependent phosphorylation, palmitoylation, and SUMOylation have been observed. These PTMs likely mediate the localization of Arc to the nucleus or PSD, and thus the different functions of Arc[57,58]. The conclude, Arc has many different roles such as a protein interaction hub at the PSD, a regulator of transcription in the nucleus, and an mRNA-carrying capsid that is able to transfect nearby cells to name a few. These separate roles of Arc have been studied in some detail, but less is understood about how Arc is regulated in order to perform the different functions. One hypothesis is that the different oligomeric states of Arc perform these different functions[29]. Most of the information on Arc oligomerization has been gathered from in-vitro biochemical assays using purified recombinant protein. Therefore, there is a need to perform more studies in-vivo to evaluate the results from the in-vitro studies.

Arc is involved in multiple plasticity-related neuropathies, such as Alzheimer's disease[59,60] and fragile X syndrome[61]. It was shown that over 100 Arc-dependent genes were involved in Alzheimer's disease[60], though the direct link between Arc and Alzheimer's disease has yet to

be discovered. Arc knockout mice are unable to form new long-term memories[32] and additionally show an increased susceptibility to seizures, which may be related to Arc's role in homeostatic scaling[33,62]. Autism spectrum disorders have many contributing causes, but one similarity between many of these causes is the alteration of synaptic homeostasis[63]. Since Arc plays a large role in the regulation of homeostatic plasticity, it is not unlikely that Arc plays a role in autism spectrum disorders. There are currently no known diseases caused directly by mutations in the expression of Arc. Rather, mutations in other proteins that regulate Arc have been observed[64]. The fact that there are no known disease variants of Arc, as well as the high conservation of Arc across vertebrates, indicates the importance of Arc. In order to understand the disease mechanisms of Arc, it is first necessary to understand its structure and molecular function. There is a need for basal knowledge of Arc's role in the synapse before one can fully understand Arc's role in diseases and start searching for potential treatments.

1.6. Arc structure

The full-length structure of mArc has not yet been solved at high resolution. The structure of the individual domains of Arc has been solved through X-ray diffraction and NMR[42,43,65]. Arc consists of two folded domains connected by a flexible linker region. Both domains are connected to a disordered N- and C-terminal region.

The CTD, also known as the capsid domain (CA) of Arc, is negatively charged and composed of two lobes connected by a dynamic helical structure [42]. This helix has been shown to have different conformations linked to the oligomeric state; the helix can be straight or angled at the Gly277 (human Arc numbering). This suggests it might play a role in regulating the different functions of Arc[66]. The CTD of Arc shares 3D homology with the capsid domain of retroviral Group-specific antigen (Gag) protein of the Ty3/Gypsy family of retrotransposons[34,43,45]. This bi-lobar capsid structure of Arc is observed in most of the species expressing Arc protein, including insects, fish, and reptiles in addition to mammals[34].
Both the N-terminal lobe (N-lobe) and C-terminal lobe (C-lobe) consist of bundles of 4-5 shorter α-helical structures.
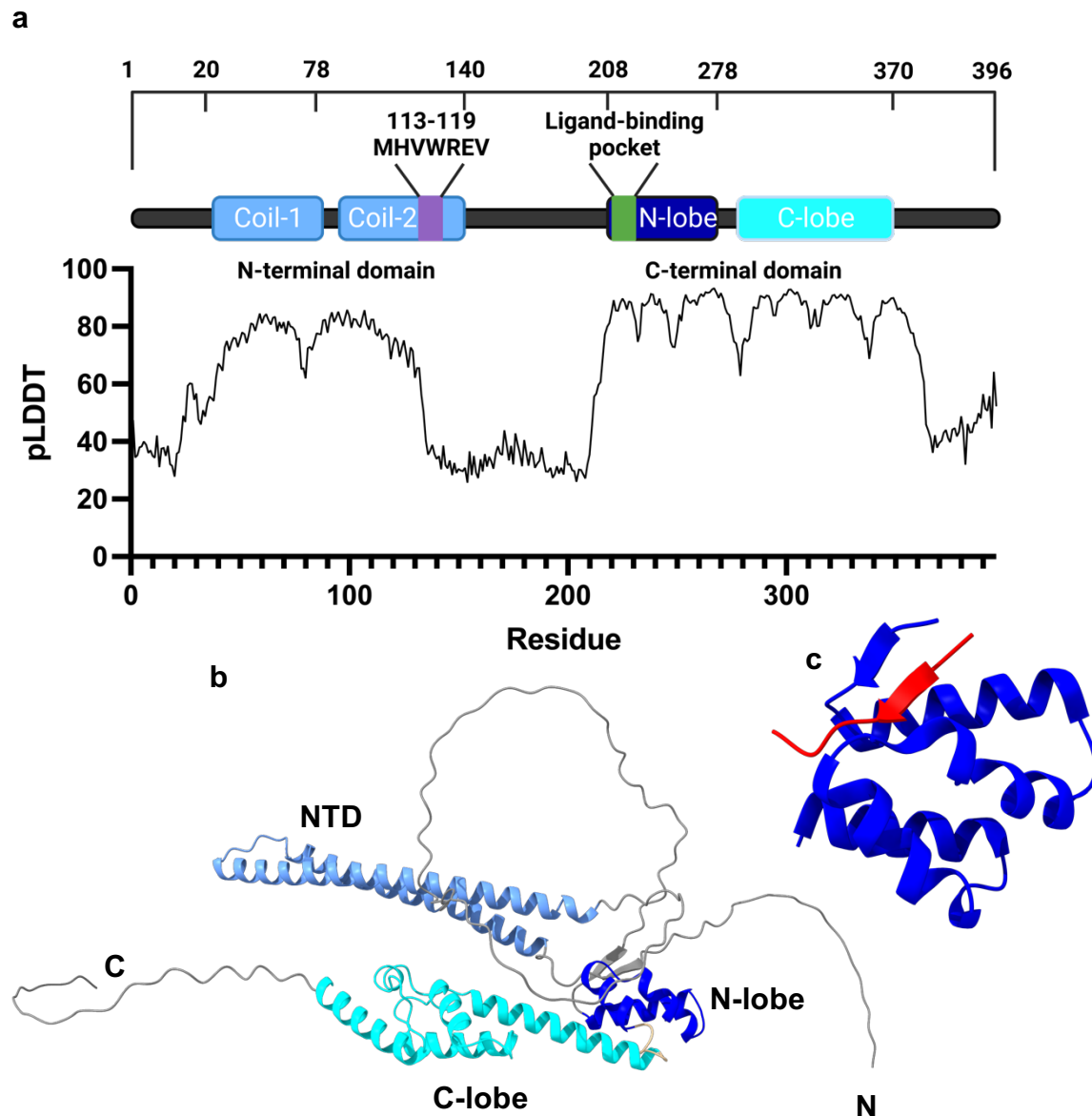
Figure 1.2: AlphaFold predicted structure of human Arc. **(a)** Here is shown the sequence of Arc with its domains colored. Low pLDDT values represent unstructured regions like the central linker and tails. **(b)** The predicted structure of human Arc retrieved from the AlphaFold DB[67], with the three regions colored. **(c)** The structure of the N-lobe of mArc in blue, bound to its peptide ligand Stargazin (Stg) in red (PDB ID: 6TNO[42]).

The N-lobe contains a peptide-binding pocket, which is responsible for many of the protein-protein interactions of Arc[42,45]. Some of the different peptide binding partners of Arc include Stargazin, Glu2NA which is a subunit of the NMDAR, and guanylate kinase-associated protein (GKAP) which also binds to different scaffolding proteins in the PSD. Molecular dynamics studies of the N-lobe crystal structure show that the flexibility of the N-lobe is changing when there is a ligand bound. When the ligand Stg is bound to the peptide binding pocket, the N-terminal strand folds to form a β-sheet over the ligand, stabilizing the structure[42]. In the 2022

study by Markússon et al.[66], it was hypothesized that the binding of a ligand induces a structural change in the linker region, which will change the relative orientations of the N and CTDs. This might partially explain how Arc can be found in different oligomeric states.

The crystal structure of the C-lobe has been solved through X-ray diffraction[42]. The C-lobe is a structural homolog to the N-lobe, and it was hypothesized that the longer peptides could bind to both the N-lobe and the C-lobe at the same time. This was however disproven in the same study when the N-lobe alone and the CTD as a whole construct were proven to have the same binding affinity towards ligands[42]. dArc C-lobe will form dimers, both in crystal structures and capsids[39,40]. There is no evidence for the role of the mArc C-lobe in oligomerization. Instead, the NTD harbors the oligomerization function in mArc.

The NTD, also known as the oligomerization domain, of Arc is positively charged and made up of an antiparallel coiled-coil structure[65]. Interestingly, this domain is only found in the mammalian version of Arc. It is thought to have evolved from the matrix domain of the Gag polyprotein[68]. The function of this domain is not as well studied as that of the CTD. The NTD mediates the binding of Arc to membranes[69], much like the matrix domain of the Gag polyprotein[39,70]. PTMs such as palmitoylation is thought to enhance this[65,70]. Eriksen Et al. were able to identify a 28-amino-acid stretch of the NTD in rat Arc, termed the oligomerization region, crucial for the formation of Arc oligomers. Within this region is a 7-amino-acid critical motif, [113]MHVWREV[119], which is the main contributor to the oligomerization. Through mutagenesis, they were able to create a mutant containing a seven-alanine substitution in this region, [113]AAAAAAA[119] (this mutant will be referred to as rArcFL7A), which binds as a dimer with no higher-order oligomers. Whether Arc forms dimers through an NTD-NTD interaction or through domain swapping where the NTD of one monomer binds to the CTD of the other[65], is still a question that needs solving.

1.7. Structural biology

The world-renowned physicist and Nobel laureate, Richard Feynman, said this about biology: "It is very easy to answer many of these fundamental biological questions; you just *look at the thing*!"[71, p. 67]. Of course, it is not always as simple as that. However, a lot can be learned about the function of a protein by looking at its structure. One might discern a catalytic site, a transmembrane region, a DNA or RNA binding site, how different monomers are arranged in a complex, or the shape of the protein which will affect its membrane permeability, just to name a few examples. Structure and function go hand in hand[72], so to understand the function

of a protein and apply that to something, it is often a good idea to first try to solve the protein 3D structure.

All proteins are made up of the same 20 amino acids, yet the size, structure, and function of proteins vary drastically. Long chains of amino acids make up peptides, which in turn make up the 2-dimensional structure of proteins. The main structural components of proteins are α-helices and β-sheets. Knowing the amino acid sequence of a protein and the secondary structural components they make up, will in most cases not be enough to deduce its tertiary structure[73]. The way a protein is folded in three-dimensional (3D) space, is what enables the function of the protein. This process is reversible in many cases, so if a protein is unfolded, it can be refolded again to regain its function[74]. The protein will naturally fold itself into the most thermodynamically stable conformation in native conditions[73]. Being able to translate the amino acid sequence into the 3D structure of the protein is still very difficult. This point becomes quite obvious when looking at the number of protein structures (over 177 000) in the Protein Data Bank (PDB)[75] compared to the 245 million entries in the UniProt TrEMBL database[76].

There are many methods that can be used to solve protein structures, and each method has its strengths and weaknesses. Out of the roughly 177 000 protein structures submitted to the PDB[75], 154 000 of those structures were solved using X-ray diffraction, 10 000 were solved using electron microscopy (EM), and 12 000 were solved using Nuclear Magnetic Resonance (NMR), to name the most common methods. X-ray diffraction has been an established method for many years now (perhaps the most famous example is the discovery of the DNA structure by Rosalind Franklin, James Watson, and Francis Crick), and is responsible for 87% of the protein structures submitted to the PDB. It can provide atomic-level resolution up to 1.0 Å and beyond for proteins of different sizes. The process involves crystallizing the protein of interest and then exposing the crystal to an X-ray beam. The diffraction of the X-ray beam is recorded by a detector, and the 3D structure of the protein can be deduced from these diffraction patterns[77]. The resolution mainly depends on the quality of the protein crystal and the detector. The main drawback of this method is the need for proteins to be crystallized[78]. This is not an easy process and can take years to achieve for a single protein, and some proteins might not even be possible to crystallize at all. Especially for intrinsically disordered proteins and membrane proteins, this problem is a bottleneck for protein structure determination.

EM has been an established method for decades. It has been used with great success to study larger structures, such as cells or tissue slices. However, there are limitations that have made it unsuitable for studying small, 3D molecules like protein structures. The addition of vitrifying the samples and keeping them at liquid N_2 temperatures has made cryo-EM one of the most promising methods in structural biology[79], even leading to the 2017 Nobel Prize in Chemistry being awarded to Jacques Dubochet[80], Joachim Frank[81], and Richard Henderson[82] for the development of the method. In single-particle cryo-EM, the sample is applied to a grid, containing usually a carbon film with holes in it. Then, the grid is plunge frozen into liquid ethane, whereupon the protein solution will be frozen into vitreous ice in the holes of the carbon film. This ice layer will hopefully be of even thickness and the protein should be evenly distributed throughout the ice. The aim is for the protein to be distributed in all 3D orientations and conformations, as *in vivo*. The grid is then imaged with the electron microscope, and 2D micrographs of the protein are collected. Particles in these micrographs will be categorized into the different orientations of the protein in 3D space. Once this process has been fine-tuned, it is possible to put together a 3D map of the protein from these 2D classes[83]. One of the main drawbacks of this method is the size limitation. The smaller the protein, the lesser the signal is and the harder it is to distinguish the different 2D micrographs from one another [84]. This lower size threshold of roughly 100 kDa is shrinking every year with the improvement of the detectors and data processing methods. Cryo-EM is also applicable to membrane proteins, unlike X-ray because of the difficulty in crystallizing these proteins in their most natural state. Membrane proteins can be imaged in detergent micelles, nanodiscs, or lipid bilayers using single-particle cryo-EM[85]. This has contributed to the growing amount of structures solved using cryo-EM, however, there is no reason to believe the increasing influence of cryo-EM will slow down any time soon.

Another useful method that provides structural information about proteins is small-angle X-ray scattering (SAXS). SAXS does not provide as high-resolution data as the methods described above; however, it can be done with proteins in solution, or membrane proteins in detergent micelles, bicelles, or nanodiscs, and the data analysis is relatively quick and intuitive[86]. This makes it suitable especially for flexible proteins[87,88], intrinsically disordered proteins (IDPs, IDPs will refer to both intrinsically disordered proteins and proteins containing intrinsically disordered regions in this thesis)[89] and small membrane proteins[90,91] which are all hard to crystallize or image using cryo-EM. There is quite a lot that can be learned about the protein from the low-resolution structures obtained from SAXS. For example, the overall size and

shape of the protein, oligomeric state, flexibility, and position of domains relative to each other can be studied, to name a few examples[86]. Large specialized facilities with a powerful source of X-rays, such as a synchrotron, are needed to perform SAXS with the highest resolution. SAXS can be paired with other methods like size exclusion chromatography (SEC) to further optimize the data collection. In SEC-SAXS, the protein is exposed to the X-ray beam as it is being extruded from the gel-filtration column. A detector will record the scattering of the X-ray beam hitting the particles in the sample[92]. From this scattering data, different protein parameters such as molecular weight, size, and low-resolution structure can be calculated. This data is often used to complement higher-resolution structures of proteins or domains thereof[93-95]. In this study, SEC-SAXS is used to study the binding dynamics of nanobodies to the Arc protein.

1.8. Nanobodies
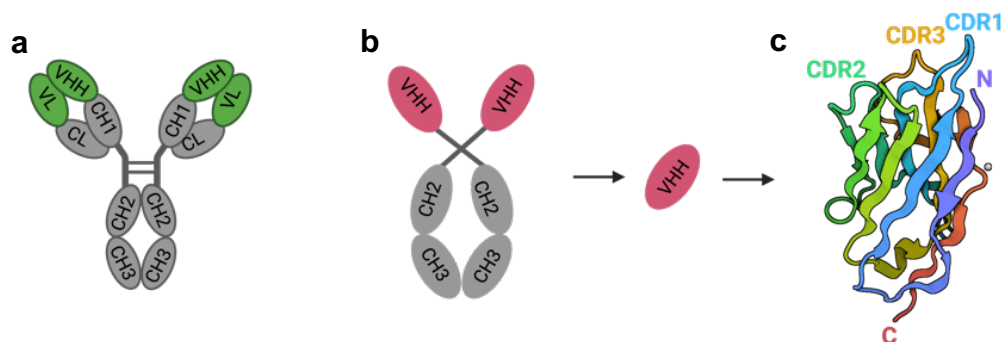
Figure 1.3: Representation of a normal IgG antibody, a camelid antibody and a nanobody. **(a)** A standard heterotetrameric IgG antibody consisting of two heavy chains (CH1, CH2, CH3 and VHH) and two light chains (VL and CL). **(b)** A camelid homodimeric antibody consisting of two heavy chains (CH2, CH3 and VHH), where the heavy chain domain (VHH) can be expressed recombinantly to produce a nanobody. **(c)** The crystal structure of a nanobody (PDB ID: 6XYF) showing the N-terminus (N), the three complementarity-determining regions (CDR1, CDR2, CDR3) and the C-terminus (C). Created with BioRender.com

Nanobodies (Nbs) are single domain-antibodies that are derived from single-chain antibodies of camelids. They consist of only the heavy chain variable domain of these antibodies and no light chains[96]. Nanobodies are between 12-15 kDa in size, compared to normal IgG antibodies which are 150 kDa, yet nanobodies still possess the specific binding affinities that make antibodies so useful. Nanobodies have three complementarity-determining regions (CDRs) connecting the β-sheet structures and providing binding specificity for the nanobodies. These CDRs can be grafted onto other nanobodies to generate a nanobody with a new binding partner[97]. The small size of nanobodies allows them to bind to epitopes where a larger antibody

might not fit and enables their use in high-resolution microscopy. They can be produced by immunization in alpacas in the same way as antibodies. Additionally, they can be expressed and purified from bacterial cells, which makes them cheaper and easier to produce in large amounts [98-102]. The idea is that these nanobodies can be used in a wide variety of applications. Nanobodies can be used to label and track endogenous proteins *in vivo*[103], as chaperones for the crystallization of proteins[66], to increase the size of protein complexes and break symmetry for cryo-EM analysis[97], and to inhibit protein functions for example by antagonistic binding to a binding site[99,103]. In addition, there is a lot of ongoing research on the use of nanobodies in cancer therapeutics[104].

1.9. AlphaFold

A milestone was reached in 2018 when the company DeepMind used artificial intelligence (AI) to predict protein structures with the AlphaFold program. AlphaFold was able to predict protein structures with an accuracy never before seen[105]. Two years later, they were able to once again push the boundaries of protein structure prediction with the development of AlphaFold2[67]. AlphaFold is a neural network that uses deep learning to predict the 3D structure using the amino acid sequence of the protein and the known structure of other homologous proteins. They are able to do this by training their neural network with the solved protein structures deposited in the PDB. Homology modeling has been used quite successfully for many years[106], however, the addition of AI to the modeling has drastically improved the accuracy. Does this newfound success of AI-based structure prediction mean that experimental structural biology will no longer be needed? Not yet. There is a need to verify these structure predictions experimentally. AlphaFold uses the amino acid sequence for its predictions. It does not take into account known ligands, cofactors, or metal ions that interact with the protein, and also struggles to predict protein complexes and oligomers, albeit this is something that is being worked on[107]. These factors play a part in how proteins are folded[108,109]. Proteins are also often found in different states. AlphaFold predicts the most thermodynamically stable conformation of the protein. This could be problematic, especially for intrinsically disordered proteins, which can have many different orientations[110]. AI-based prediction models have also not been well enough established yet that their output can be taken as the solution. There is still a need to experimentally prove these structures and compare them to the prediction. That being said, AlphaFold can be a very helpful tool both in designing experiments to solve protein structures and to give insight into the function of a protein.

1.10. Aims of the study

The structures of some individual domains of Arc have been solved, yet the high-resolution structure of full-length Arc remains elusive. The main aim of this project was to learn more about the 3D structure of mArc, with the help of nanobodies. The idea was to use the nanobodies as crystallization chaperones and to increase the size of Arc in order for it to be more easily visible during cryo-EM analysis. Arc has many functions, none of which are fully understood. How does Arc oligomerize? What are the functions of the different oligomeric states? What are the contents of the Arc capsids? How are new long-term memories formed? What is the role of Arc role in diseases such as Alzheimer's disease, autism spectrum disorders, and fragile-X syndrome, and can they be treated? A high-resolution structure of full-length Arc in different oligomeric states could help solve some of these questions in the future.

Another goal of the study was to further characterize the binding of the anti-Arc nanobodies to Arc using SAXS. These nanobodies can be useful tools to study both the function and structure of Arc. They could be used as crystallization chaperones, to increase the size of Arc and break symmetry for cryo-EM studies, as antagonists blocking the functions of Arc *in vivo*, or as probes to track the activity and localization of Arc in live cells and for immunohistochemical staining of fixed tissue. Learning more about the nanobodies could enable all of these studies, and new knowledge might open up even more possible applications. Thus, learning more about these anti-Arc nanobodies will aid the search for the high-resolution structure of Arc and further understanding of Arc's function in the cell and diseases.

2. Materials and methods

2.1. Materials

Unless otherwise stated, all materials were purchased from Thermo Fisher Scientific (MA, USA), and all chemicals from Sigma Aldrich-Merck (MA, USA).

2.2. Sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE)

To assess the purity of the recombinant proteins, analytical SDS-PAGE was performed according to Laemmli[111]. 10 µL of sample was mixed with 3 µL of 5x loading dye (300mM Tris-HCl pH 6.8, 50% (v/v) glycerol, 10% (w/v) SDS, (w/v) 0.05% bromophenol blue and 10% (v/v) β-mercaptoethanol. Samples were then heated at 95 °C for 5 min and quickly centrifuged to collect the condensates, before loading 10µL onto a Mini-PROTEAN TGX Precast 4-20% 15-well Gel (BioRad Laboratories, CA, USA). 2 µL of PageRuler Prestained Protein Ladder 10-180 kDa (Thermo scientific, MA, USA: 26616) was also loaded onto the gel. Gels were run for 35 min at 200 V until the bromophenol blue reached the end of the gel. 25 mM Tris-HCl, 0.1% (w/v) SDS, 192 mM glycine, pH 8.3 was used as the running buffer. Gels were washed with water and stained for at least 30 min up to overnight using InstantBlue Coomassie Protein Stain (Abcam, Cambridge, UK). The stained gels were then washed with water and imaged using the ChemiDoc XRS+ Gel Imaging system (BioRad Laboratories).

2.3. Transformation, protein expression, and purification
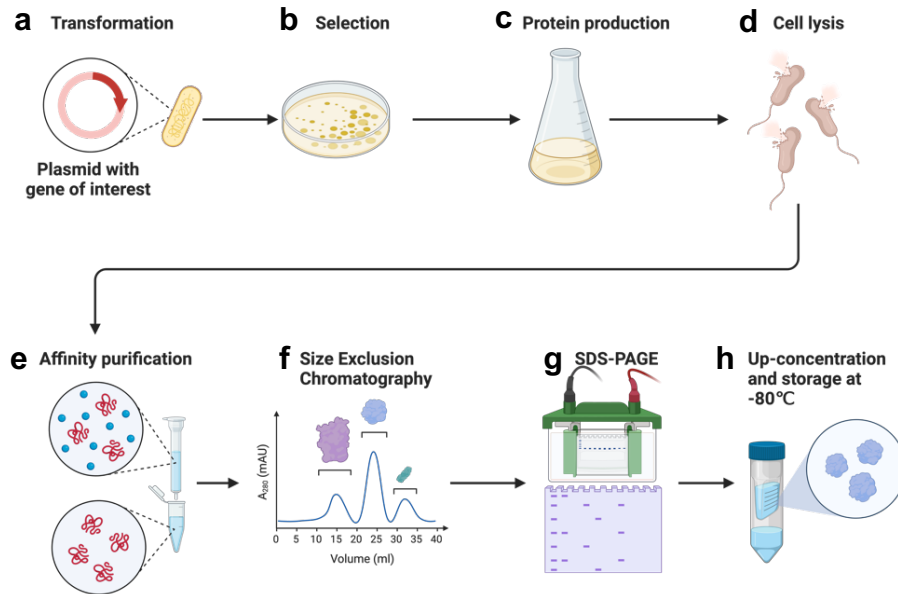
Figure 2.1: An overview of the protein purification process. **(a)** The plasmid vector containing the gene of interest and antibiotic resistance is transfected into E. coli cells using heat shock. **(b)** The E. coli are plated onto an LB agar plate containing antibiotics and allowed to grow overnight. **(c)** The following day, a bacterial colony is picked and added to LB media containing antibiotics to grow until the desired OD_{600} was achieved, when protein expression is induced via IPTG. **(d)** The cells are then harvested by ultra-speed centrifugation and lysed using lysozyme and liquid N_2. **(e)** The protein undergoes the first affinity purification, overnight proteolysis, and another affinity purification. **(f)** SEC is used to further purify the proteins. **(g)** The fractions are sampled and analyzed using SDS-PAGE to determine the purity. **(h)** Finally, the protein is up-concentrated and stored at -80 °C. Created with BioRender.com

2.3.1. Transformation

All plasmids were transformed into BL21-competent E. coli cells using the following protocol. 1 μL of plasmids were added to 100 μL of cells before being gently mixed and left on ice for 30 min. The cells were then heat shocked at 42 °C for 45 s and kept on ice for 2 min. Afterward, 500 μL of sterile LB media was added, and the cells were kept at 30 °C for 2 h while gently shaking. The cells were then centrifuged for 2 min until a pellet formed. Then half of the supernatant was discarded, and the pellet was resuspended in the remaining supernatant. 50 μL of these cells were plated on 1% (w/v) agar LB plates containing 50 mg/ml kanamycin. The plates were incubated at 37 °C overnight.

2.3.2. Expression

The starter culture was prepared by picking a colony from the agar plate and adding it to 10 mL of sterile LB media containing 50 μg/mL kanamycin. The starter culture was then incubated overnight at 37 °C 170 rpm. The following day, 10 mL of starter culture was added to 1 L of

sterile LB media containing 50 μg/mL kanamycin and incubated for 3-4 h at 37 ℃ 200 rpm until the measured cell concentration (OD_{600}) registered 0.4-0.6. Once the desired OD_{600} was reached, the temperature was decreased to 30 ℃, and 1 mM IPTG was added to induce protein expression. After 4 h, the cells were harvested by centrifuging for 20 min at 4 ℃ and 5 000 rpm (5 400 xg) using the Sorval Lynx 6000 Superspeed Centrifuge (Thermo Scientific, (Rotor: Fiberlite F9-6x1000 LEX)). The supernatant was discarded, and the pellet was resuspended in 35 mL of lysis buffer (40 mM HEPES pH 7.5, 100 mM NaCl, 20 mM imidazole), and 0.1 mg/ml hen egg-white lysozyme was added to further lyse the cells. The cells were then snap-frozen in liquid N_2 and stored at -20 ℃.

2.3.3. Individual protein purification

All the nanobodies were purified in the same way, and the Arc constructs followed the same protocol with some slight changes. Cells were thawed and sonicated on ice for 7 cycles of 3x 25 W for 10 s and 0 W for 30 s. After being lysed, the cell contents were harvested by centrifuging for 30 min at 4 ℃ and 17 000 g using the Sorvall Lynx 6000 Superspeed Centrifuge (Thermo Scientific, (Rotor: Fiberlite F21-8x50y)). The supernatant was filtered using a 0.45 μm syringe filter and the pellet was discarded. All the next steps were done at 4 ℃ to preserve the proteins. A Ni^{2+}-nitriloacetic acid (NiNTA) matrix equilibrated with washing buffer (40 mM HEPES pH 7.5, 100 mM NaCl, and 20 mM imidazole) was prepared, and the filtered supernatant was passed through the NiNTA column three times to bind the His-tagged protein. The column was then washed with 20 mL of the same washing buffer. Afterward, the column was then capped and 6 mL of elution buffer (40 mM HEPES pH 7.5, 100 mM NaCl, and 300 mM imidazole) and left for 10 min before uncapping. After the elution buffer had all passed through the column, another 4 mL of elution buffer was added. The appropriate (200 uL at 30 μM) amount of Tobacco Etch Virus (TEV) protease was then added to the eluted fraction to cleave the His-tag from the protein before being added to a Spectra/Por 1 Dialysis Membrane MWCO: 6-8 kDa (Spectrum Laboratories, CA, USA). It was then dialyzed against 1 L of 40 mM HEPES pH 7.5, 100 mM NaCl and 1 mM DTT overnight at 4 ℃ gently stirring to allow the cleaved His-tags to be removed. The following day, the contents of the dialysis bag were added to a NiNTA matrix equilibrated with 40 mM HEPES pH 7.5 and 100 mM NaCl and allowed to pass through 3 times and the flowthrough was collected. Bound proteins were eluted with the same buffer as above.

The Arc constructs contain an MBP fusion tag which will bind to amylose and increase the solubility of the protein[112,113]. Therefore, a second matrix was prepared with amylose resin equilibrated with the same buffer. The sample was allowed to flow through three times before elution with 10 mL of 40 mM HEPES pH 7.5, 100 mM NaCl and 10 mM maltose. Again, the flow through was collected and concentrated up to 2 mL using an Amicon Ultra 15 10 kDa MWCO (Merck Millipore, Ireland) centrifugal filter for the nanobodies, and an Amicon Ultra 15 30 kDa MWCO centrifugal filter for the Arc construct.

The sample was then filtered through an Ultrafree 0.22 μm centrifugal filter (Merck Millipore), before being loaded onto a HiLoad 16/60 Superdex 75 pg size exclusion chromatography (SEC) column (GE Healthcare, IL, US) for the nanobodies, or a HiLoad 16/60 Superdex 200 pg size exclusion chromatography (SEC) column (GE Healthcare) for the Arc construct. The column had been equilibrated with 20 mM Tris pH 7.5 and 150 mM NaCl, and the SEC run was performed at 1mL per min using the Äkta Pure chromatography system (GE Healthcare). Some of the earlier purifications were performed on a BioLogic DuoFlow chromatography system (BioRad) with a BioLogic BioFrac fraction collector (BioRad) until it finally broke down after decades of service. 1 mL fractions were collected and analyzed by SDS-PAGE. The fractions containing the pure protein peak were pooled together, up-concentrated, aliquoted, and snap-frozen in liquid N_2 before being stored at -80 °C until used.

2.3.4. Arc-nanobody complex purification

Equimolar amounts of rArcFL7A, E5, and C11 were removed from storage at -80 °C and thawed on ice. The proteins and nanobodies were gently mixed and incubated on ice for 1 h to ensure binding. After incubation, the proteins were filtered using an Ultrafree 0.22 μm centrifugal filter (Merck Millipore) to remove any aggregation. The sample was then gel filtrated on a HiLoad 10/300 Superdex 200 pg SEC column (GE Healthcare), equilibrated with 20 mM Tris pH 7.5, 150 mM NaCl and 0.5 mM TCEP. Once the size and purity of the peaks were assessed using SDS-PAGE, the protein concentration was measured using a NanoDrop 2000 Spectrophotometer (Thermo Scientific), and the protein complex was kept on ice at 4 °C until used. Freezing the protein complex is not recommended, so this purification had to be done every time an experiment was run.

2.4. Size exclusion multi-angle light scattering (SEC-MALS)

SEC-MALS was performed to estimate the oligomeric state and stoichiometry of the rArcFL7A dimer and its nanobody complexes. SEC-MALS was performed with help from Anne Baumann at the Biophysics, Structural Biology, and Screening (BiSS) core facility at the Department of Biomedicine. SEC was performed on a HiLoad 16/60 Superdex 200 pg SEC column (GE Healthcare) equilibrated with 20 mM Tris pH 7.5, 150 mM NaCl and 0.5 mM TCEP using a Shimadzu Prominence-i LC-2030C 3D HPLC unit (GMI, MN, USA) with an LC-2030/2040 PDA UV detector (Shimadzu, Kyoto, Japan). The system was calibrated using bovine serum albumin (BSA) shortly before running the samples. 50 µg of each protein complex in the aforementioned buffer were injected at a flow rate of 0.5 mL/min and the light scattering was recorded using a miniDAWN TREOS detector (Wyatt Technologies, CA, USA) and a RefractoMax 520 refractometer (Wyatt Technologies). Data collection and SEC-MALS analysis were carried out using the ASTRA 6.1 software (Wyatt Technologies).

2.5. Crystallization

Protein crystallization was attempted using commercial crystallization screens JCSG+ and PACT-Premier (Molecular Dimensions, Sheffield, UK) in a Triple Sitting Drop 96-well plate (SPT Labtech, Melbourn, UK) using a Mosquito LCP crystallization robot (SPT Labtech).
The protein-nanobody complex was purified as described above. The screening plates were set up with different protein concentrations (2-10 mg/mL), temperatures (8 °C and 20 °C), and protein to buffer volume ratio (1:2, 1:1, and 2:1). The 20 °C plates were kept in a Formulatrix Rock Imager 182 (Formulatrix, MA, USA) and the 8 °C plates were stored in a Molecular Dimensions MD5-603 Bench Top Incubator (Molecular Dimensions). As no large crystals were produced, X-ray diffraction was not performed.

2.6. Negative staining transmission electron microscopy (TEM)

Negative staining TEM was performed in order to assess the sample purity and observe how the protein complex behaves in solution. Some optimization was needed to find the optimal protein concentration and sample preparation. The protein complex was purified as described above and diluted to 0.02-0.05 mg/ml concentration. The samples were then filtered through

an Ultrafree 0.22μm centrifugal filter (Merck Millipore) to remove any aggregation. Different grids were used to fixate the protein solution Formvar/carbon-coated nickel grids 200 mesh (Polysciences, PA, USA), Formvar/carbon-coated carbon grids 300 mesh (Electron Microscopy Sciences, PA, USA), and in-house prepared carbon-coated copper grids 300 mesh. All grids were glow discharged before use. 3 μL of the sample was added to the grid and left for 60 s before carefully being dried off with paper. Then, 3 μL of buffer (20 mM Tris pH 7.5 and 150 mM NaCl) was added and dried off with paper after 15 s. Then another 3 μL of the same buffer was applied for 15 s before being dried off again. 3 μL of uranyl acetate at 2% was applied and dried off with paper after 2 s. Then, another 3 μL of uranyl acetate was added and left for 30 s before being dried with paper. The grids were prepared and kept in a grid box at room temperature, away from light, until being used. For all grids, a Hitachi HT7800 TEM (Hitachi High-Technologies Corporation, Tokyo, Japan) was used with an EMSIS XAROSA B20T camera (EMSIS, Münster, Germany) attached. Collected images were processed in Fiji[114].

2.7. SEC-SAXS

In SEC-SAXS, the sample is extruded from the gel filtration column directly into an X-ray beam, and the scattering of the X-rays is measured. The SEC part of this method was performed as described above, with some minor changes. This time, instead of equimolar amounts of protein and nanobodies, they were combined in a 1:1.3 protein-to-nanobody ratio to ensure all the Arc protein is bound to the nanobodies. Since the volume needed for SEC-SAXS is modest, a smaller column was used, Agilent Bio SEC-3 Column (Agilent Technologies, CA, USA) or HiLoad 16/60 Superdex 200 pg SEC column (GE Healthcare).

SEC-SAXS data collection was performed twice. First at the CoSAXS beamline[115] at MAX IV (Lund, Sweden) and again at the SWING beamline[93] at the SOLEIL synchrotron (Gif-sur-Yvette, France). Buffer subtraction and frame selection were performed in CHROMIXS [116], primary analysis in PRIMUS[117] and distance distribution function analysis using GNOM[118]. *Ab initio* model building was done with DAMMIN[119] and GASBOR[120]. CRYSOL[121] was used to calculate the theoretical scattering curves of the models created using AlphaFold[67] (described in more detail below).

2.8. Cryogenic electron microscopy (cryo-EM)

Cryo-EM and sample preparation were performed at the University of Umeå, at the Umeå Centre for Electron Microscopy (UCEM). The proteins were transported on dry ice to the facility, where the complexes were purified as described above. Two types of grids, Quantifoil 2/1 300 mesh copper grids (Quantifoil Micro Tools GmbH, Jena, Germany) and Quantifoil 2/2 300 mesh copper grids with 2 nm carbon layer (Quantifoil Micro Tools GmbH) were glow discharged for 20 s and 4 µL of the protein was applied. The grids were then blotted for 5 s at 4 °C and 100% humidity and plunge-frozen into liquid ethane using a Vitrobot Mark IV (Thermo Scientific). Afterward, the grids were stored in liquid N_2. The samples were screened, and data was collected on a 200 keV Titan Glacios Cryo-TEM with a Falcon4 detector (Thermo Scientific) at 190 000x nominal magnification in counting mode, with a pixel size of 0.746 $Å^2$ and objective aperture set to 100 µm. Images were acquired at a defocus range of -1.5 to -3.0 µm, an exposure time of 3.49 s, resulting in a total dose of 60 electrons per $Å^2$. A total of 7 000 images were collected with one image collected per hole, using the automated software EPU v 3.2.0 (Thermo Scientific). Data collection on the Titan Krios Cryo-TEM (Thermo Scientific) was scheduled to be performed in June 2023, as promising results from the initial screening were obtained. Preliminary data processing was performed using CryoSPARC[122] at UCEM, and unfortunately due to hardware malfunction in the home laboratory, further data analysis was not completed in time for the submission of this work.

2.9. AlphaFold modeling

AlphaFold2[67] and AlphaFold-Multimer[107] were run using the COSMIC[2] science gateway[123] and the ColabFold server[124], using the amino acid sequences for the different complexes (rArcFL7A, rArcFL7A+E5, rArcFL7A+C11, and rArcFL7A+E5+C11) to build multimers. The models were visualized using PyMOL[125] and UCSF ChimeraX[126].

3. Results

3.1. Protein Purification

3.1.1. rArcFL7A

Full-length mArc will oligomerize and form large virus-like capsids after being expressed[34]. This makes protein purification problematic when studying monomeric Arc. Thankfully, Eriksen et al. were able to identify the rArcFL7A mutant, [113]MHVWREV[119] to [113]AAAAAAA[119], which forms a dimer[65]. This construct was chosen for this study because it is easier to produce and perform structural studies on than the wild-type full-length Arc. There has been done a lot of characterization of this construct already in the lab, and an optimized purification protocol exists[66]. rArcFL7A was expressed from the pHMGWA vector[127] with an N-terminal His_6-tag and with MBP as a fusion partner. The vector was inserted into BL21 cells, and protein was expressed in LB medium using IPTG to induce the expression. Then followed affinity purification and protease treatment to cleave the fusion partners. Lastly, the protein was purified using SEC to ensure protein purity. SEC separates proteins based on their hydrodynamic radius, and the different collected fractions can be sampled and analyzed using SDS-PAGE to see what size proteins they contain. The addition of β-mercaptoethanol with SDS and the heating step dissociates the dimeric complex, and Arc runs as a monomer in the gel. The yield of the purification is roughly 10 mg of pure protein per liter of bacterial culture.
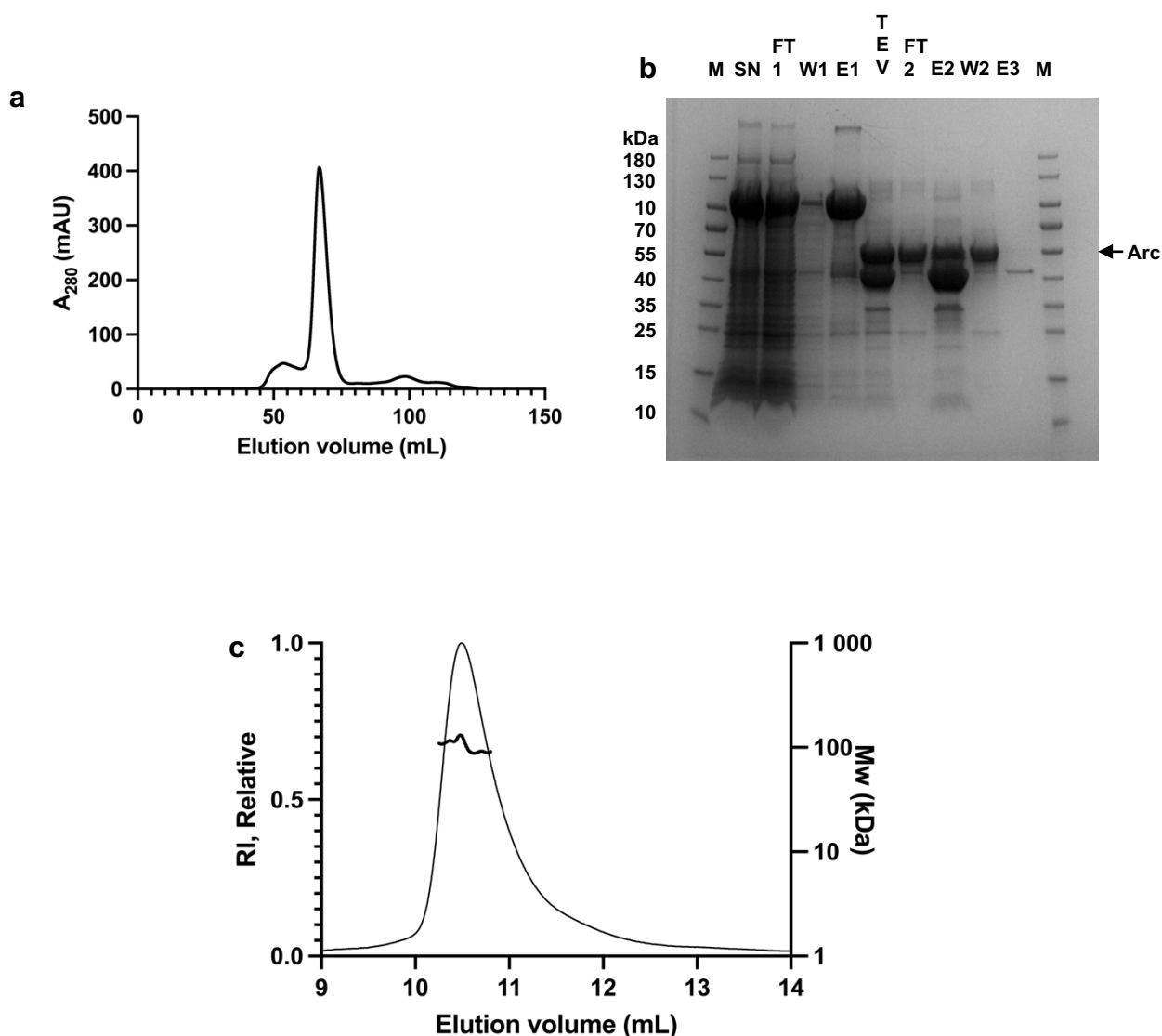
Figure 3.1: rArcFL7A purification and SEC-MALS. **(a)** SEC elution profile for rArcFL7A following affinity purification performed on a HiLoad Superdex 75 16/60 column. **(b)** The wells are labeled as follows: Marker (M), Supernatant after harvesting (SN), First NiNTA flow through (FT 1), First NiNTA wash (W1), First NiNTA elution (E1), NiNTA elution post TEV treatment (TEV), Second NiNTA flow through (FT 2), Second NiNTA elution (E2), Second wash with an amylose column this time (W2), Third elution this time with an amylose column (E3). SDS-PAGE analysis from expression and the different steps of the affinity purification leading up to the SEC. The band corresponding to the size of Arc is highlighted with an arrow. Arc runs closer to 55 kDa in gels rather than the actual size of 45 kDa. **(c)** Elution profile of the SEC-MALS run of rArcFL7A with relative refractive index (RI) plotted on the left Y-axis and MW (kDa) on the right Y-axis. The MW estimation extracted from Table 1 is 104.9 kDa with an estimated error of 26.8%. The signal was very noisy, as can be seen in Figure 3.5, explaining why the estimated error is so high. Even though the MW estimation is so uncertain, it is certain that rArcFL7A is in a dimeric state (90 kDa) as opposed to a monomeric state (45 kDa).

3.1.2. Nanobodies
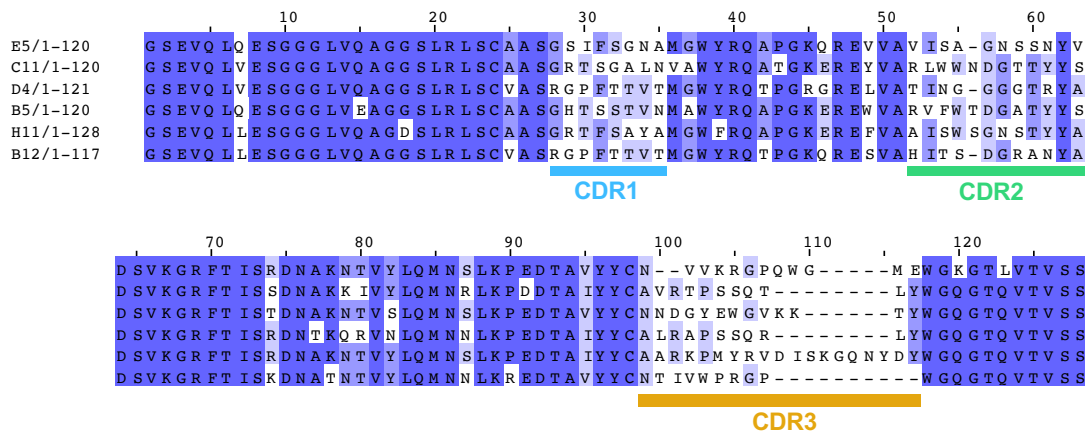
Figure 3.2: Aligned sequences of all six nanobodies raised against Arc. The sequences are conserved across the structured regions, whereas the three CDR loops are unique to each nanobody. Sequences were aligned with Clustal [128] and visualized with Jalview [129]. Nanobodies E5 and C11 were selected for this study.

Six nanobodies, E5, C11, D4, B5, H11, and B12, were previously obtained[66] from alpacas immunized against Arc through NanoTag (Göttingen, Germany). Two of these Nbs, H11, and E5, were shown to bind to the N-lobe of Arc and the rest to the C-lobe[103]. Since the nanobodies do not all bind to the same site, either H11 or E5 can be bound to rArcFL7A in combination with one of the other four nanobodies. Differential scanning fluorimetry (DSF) and isothermal titration calorimetry (ITC) had been carried out in order to find the thermal stability and binding affinity of rArcFL7A bound to these nanobodies. It was previously found that H11, C11 and E5 had the highest binding affinity and temperature stabilizing effect[66]. Since the combination of H11 and C11 with rArcFL7A had already been attempted to crystallize, it was decided that this study would focus on characterization and attempted crystallization of rArcFL7A with E5 and C11 simultaneously bound. Purification of E5 and C11 followed the same protocol described in the methods and materials section. The constructs were inserted in the pTN1433 expression vector with a His-tag fusion partner and expressed in *E. coli* BL21(DE3) cells. After harvesting the cells, the proteins were purified using affinity purification and subjected to protease treatment to cleave the His-tag. After the SEC run the fractions would be sampled and analyzed using SDS-PAGE to determine the size and purity. The total protein yield would vary between roughly 1-10 mg of protein per liter of bacterial culture.
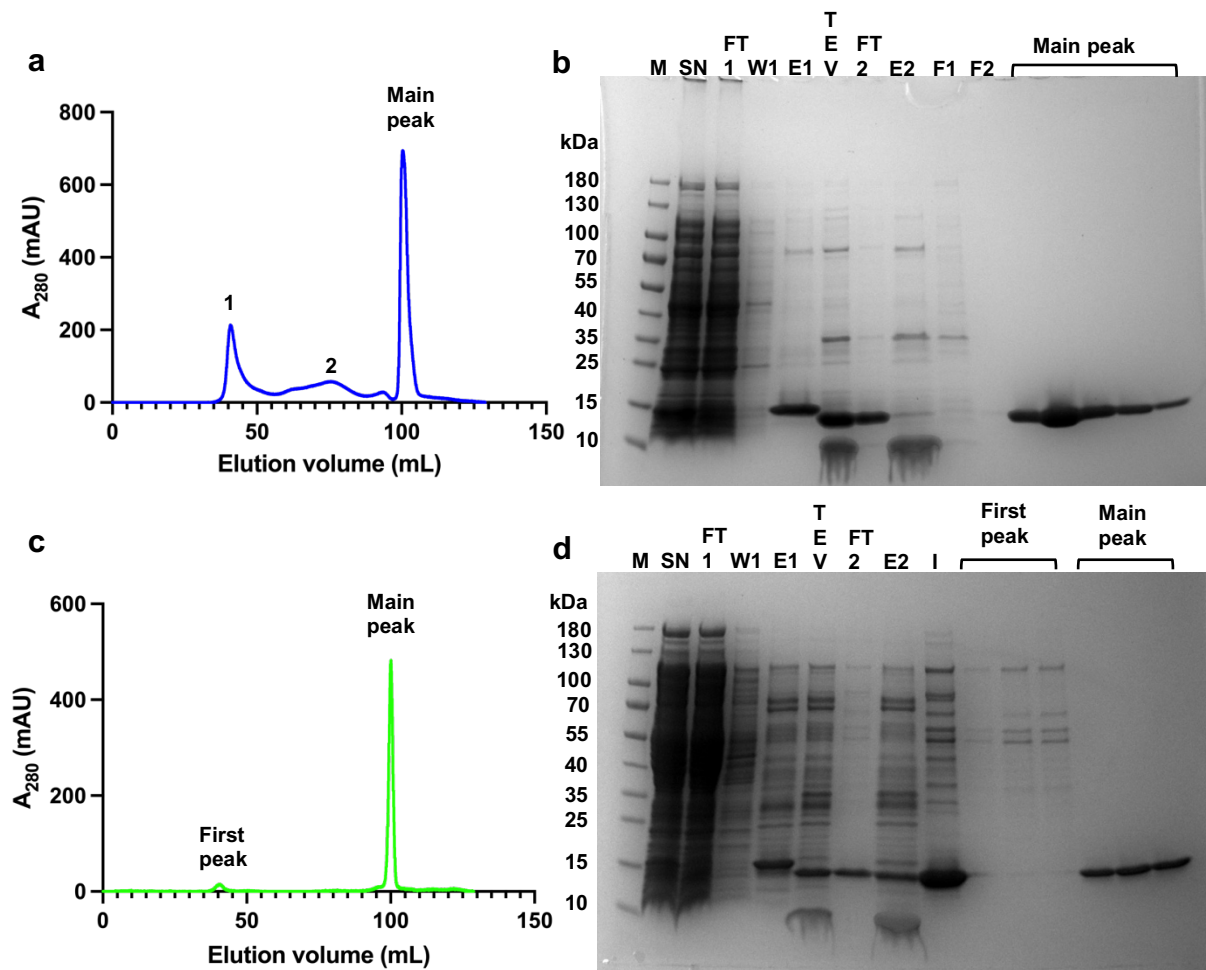
Figure 3.3: E5 and C11 purification (a) SEC elution profile for E5 following affinity purification performed on a HiLoad Superdex 75 10/300 column. **(b)** The wells are labeled as follows: Marker (M), Supernatant after harvesting (SN), First NiNTA flow through (FT 1), First NiNTA wash (W1), First NiNTA elution (E1), NiNTA elution post TEV treatment (TEV), Second NiNTA flow through (FT 2), Second NiNTA elution (E2), Up-concentrated sample injected into the gel filtration column (I), The numbered fractions from the elution profile (F1 and F2). SDS-PAGE analysis from expression and the different steps of the affinity purification and SEC. **(c)** SEC elution profile for C11 following affinity purification performed on a HiLoad Superdex 200 10/300 column. **(d)** SDS-PAGE analysis from expression and the different steps of the affinity purification and SEC.

3.1.3. Arc+E5+C11

As described above, the main focus of this study was to characterize the rArcFL7A+E5+C11 complex through various methods. The purification process started with mixing the pure protein and nanobodies in a 1:1.3:1.3 molar ratio of rArcFL7A:E5:C11. This complex was gently mixed and incubated on ice for 30-60 min to ensure complete binding of the nanobodies to Arc. The complex was then filtered to remove any aggregates before the SEC run. After the SEC run, an SDS-PAGE analysis was performed to see the content and purity of the fractions. The samples were run under reducing conditions after being heated. Thus, the complex has

been dissociated, and the bands correspond to each of the monomers (Arc 45 kDa, E5 12.76 kDa, and C11 13.11 kDa).

Table 1: Molecular weight calculated from the refractive index and light scattering.

Sample	MW (kDa)	Est. error (%)
rArcFL7A	104.9	26.8
rArcFL7A+E5	122.0	11.2
rArcFL7A+C11	123.4	10.6
rArcFL7A+E5+C11	163.0	16.6
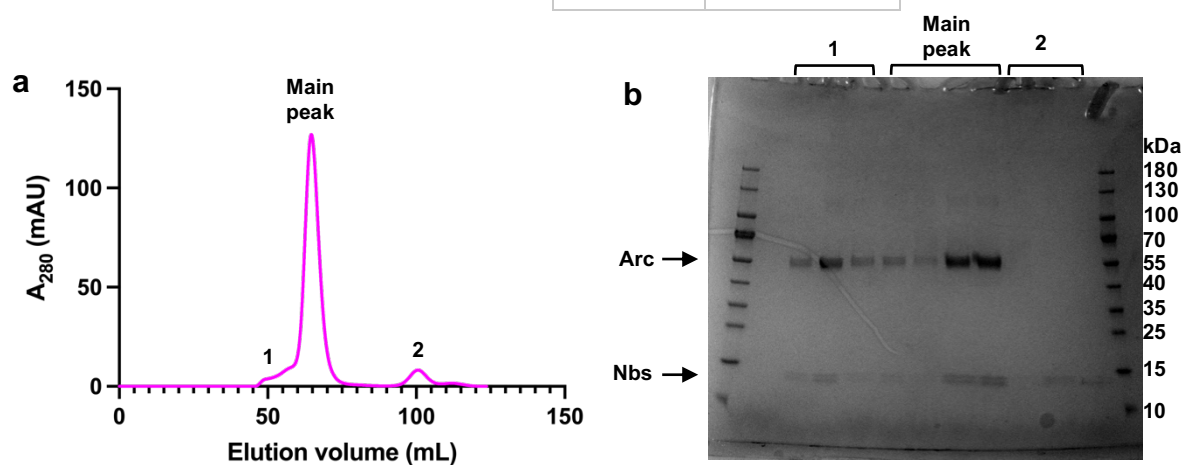
Figure 3.4: rArcFL7A+E5+C11 purification. (a) SEC elution profile for the rARcFL7A+E5+C11 complex performed on a HiLoad Superdex 200 10/300 column. **(b)** SDS-PAGE analysis of the samples taken from the SEC fractions. The shoulder of the peak contains aggregated protein and was discarded. The main peak contains pure protein complex and the final peak contains the excess nanobodies not bound to Arc. Two bands around 13 kDa can be seen, indicating the slight size difference between E5 and C11. Both Arc and the two nanobodies are highlighted with arrows.

After purification of all the proteins, it was decided to run SEC-MALS with rArcFL7A, rArcFL7A with E5 bound, rArcFL7A with C11 bound, and rArcFL7A with E5 and C11 bound. This was done in addition to the analytical SDS-PAGE to ensure that both nanobodies are able to bind at the same time. The system was calibrated using BSA right before running the samples, but after each sample had been run, the scattering signal was getting noisier and noisier, resulting in the suboptimal molecular weight estimations shown in Figure 3.5 and Table 1. Ideally, the samples would have been run again, this time with a larger protein amount and washing the column between each sample. Unfortunately, time did not allow for this. Still, there is a clear trend to be seen. The estimated size of the complex increases with each additional nanobody added, proving that E5 and C11 do not compete for binding to rArcFL7A.

a

b
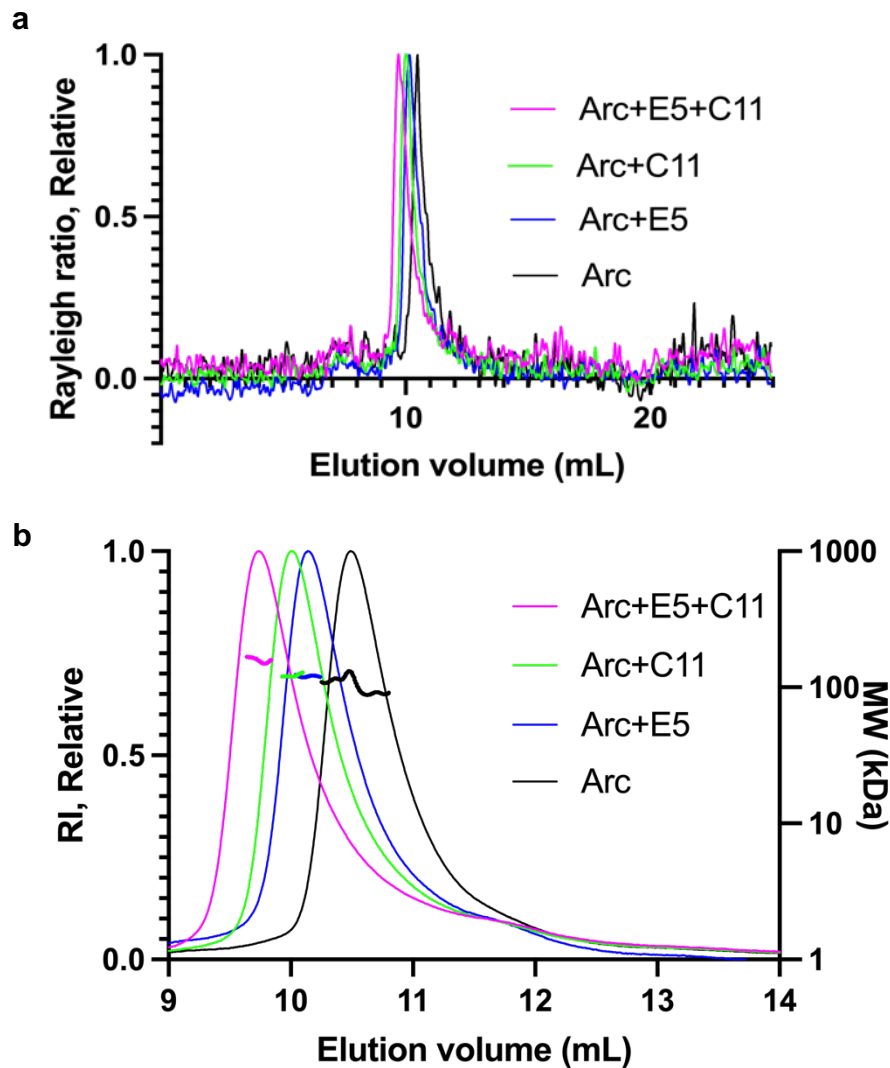
Figure 3.5 : SEC-MALS analysis of the four rArcFL7A and nanobody complexes. **(a)** Elution volume plotted against the scattering, showing the large noise to signal ratio of the samples. **(b)** Elution volume plotted against the relative refractive index (RI) on the left Y-axis and estimated MW (kDa) on the right Y-axis. The low signal to noise ratio accounts for the imprecise MW estimation. Still, both the UV peak of the protein and the MW estimation changes as each nanobody is added.

3.1.4. Anti-Arc anti-Fab nanobody C11

One other nanobody was acquired that binds to Arc and to an antibody fragment (Fab). This was done in order to further enlarge and rigidify the protein complex, which might be beneficial for both crystallization and cryo-EM. The sequence coding for the CDR loops of C11 are grafted onto an existing anti-Fab nanobody, creating a new construct that will bind to both Arc and the Fab. This Fab-binding region is situated at the distal end in relation to the CDR loops, so as to not interfere with the binding or the protein of interest[97]. The structure is shown in Figure 4.5. This new construct also included MBP as a fusion partner. It was attempted to

express and purify this new anti-Arc anti-Fab nanobody following a non-optimized purification protocol. This unsuccessful attempt is shown in Figure 3.6.
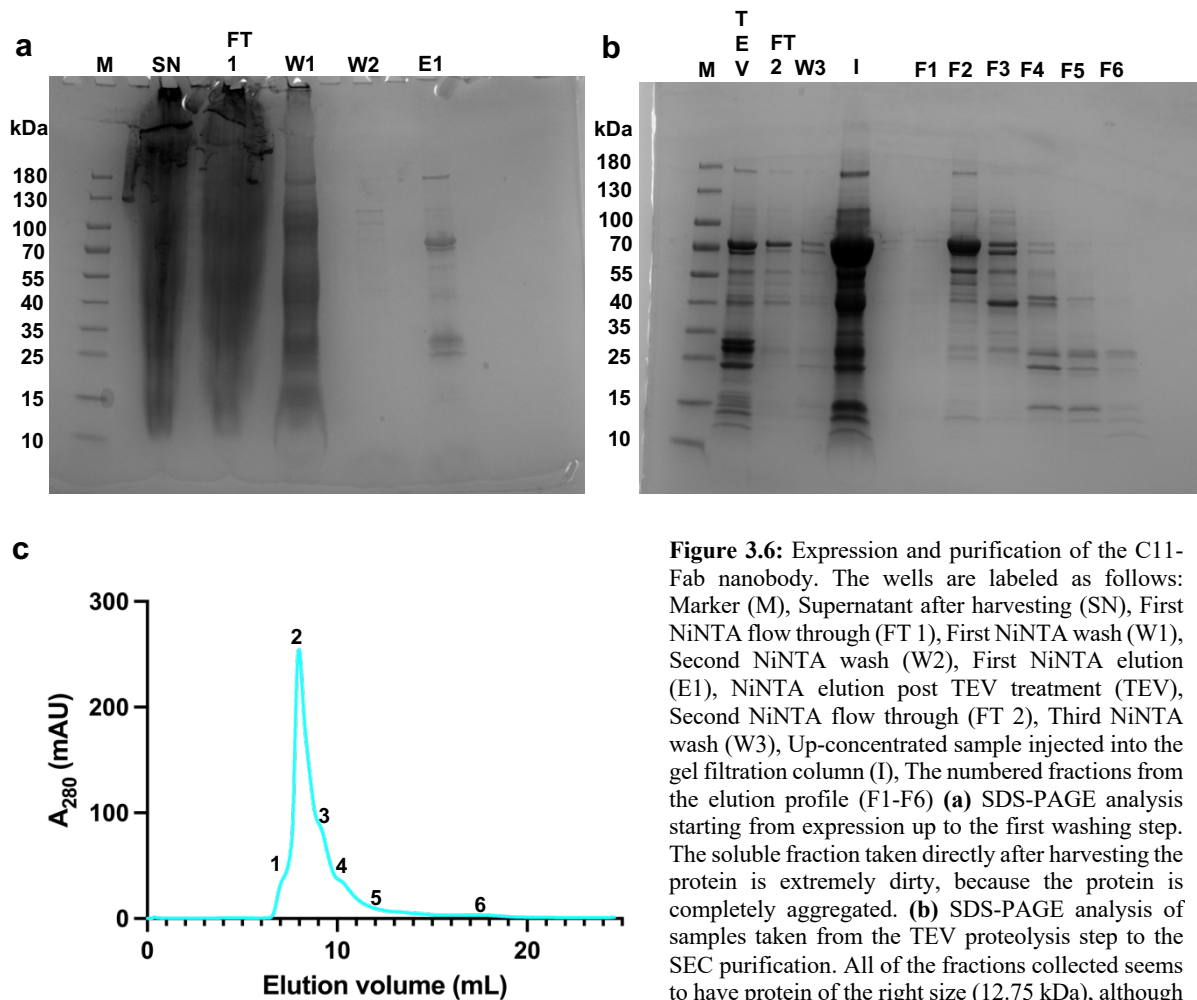
Figure 3.6: Expression and purification of the C11-Fab nanobody. The wells are labeled as follows: Marker (M), Supernatant after harvesting (SN), First NiNTA flow through (FT 1), First NiNTA wash (W1), Second NiNTA wash (W2), First NiNTA elution (E1), NiNTA elution post TEV treatment (TEV), Second NiNTA flow through (FT 2), Third NiNTA wash (W3), Up-concentrated sample injected into the gel filtration column (I), The numbered fractions from the elution profile (F1-F6) **(a)** SDS-PAGE analysis starting from expression up to the first washing step. The soluble fraction taken directly after harvesting the protein is extremely dirty, because the protein is completely aggregated. **(b)** SDS-PAGE analysis of samples taken from the TEV proteolysis step to the SEC purification. All of the fractions collected seems to have protein of the right size (12.75 kDa), although neither of the fractions contain pure protein. **(c)** SEC elution profile of C11-Fab. Most of the protein is eluted in or right after the void volume of the column, signifying that most of the protein is in an aggregated state.

Due to time constraints, this was the first and only attempt to purify this nanobody. The expression and purification would need optimization in order to be successful. If there was more time available, this would have been prioritized. It was observed after harvesting and lysing the cells that the cell lysate was extremely sticky. This is most likely due to the cells growing too fast, and the protein aggregating as a result. The first change would be to try expressing the nanobody in a different cell line more suited for nanobodies, such as the SHuffle *E. coli* cell line[130]. Another change would be to try a different vector which might also be more suited for these types of small proteins. This combined with a slower expression (for example

autoinduction instead of IPTG) at a lower temperature, might result in a more stable expression of the protein.

3.2. SAXS

SAXS data was collected from the four samples (rArcFL7A, rArcFL7A+E5, rArcFL7A+C11, and rArcFL7A+E5+C11) on two separate occasions. The first time was at the CoSAXS beamline on the MAX IV synchrotron and the second time was at the SWING beamline on the SOLEIL synchrotron. The strong X-ray beam at MAX IV led to problems with capillary fouling, which is when the protein aggregates and sometimes even burn to the capillary due to continuous exposure to an intense X-ray beam[131]. This has been observed previously in the group when measuring this specific protein at a similarly strong X-ray beam. This can be seen in the shifting baseline before and after the main protein peak in Figure 3.7. Therefore, it was decided to measure the samples again at the SWING beamline which has a weaker X-ray beam, hopefully more suited for these protein complexes.
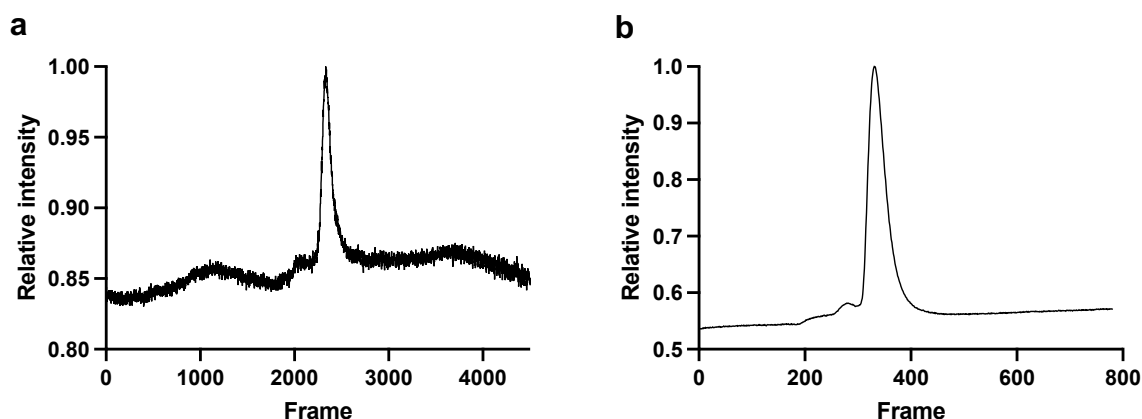
Figure 3.7 : Size exclusion chromatograms during the SAXS experiments. **(a)** SEC elution profile from the first measurements of rArcFL7A on the CoSAXS beamline at the MAX IV synchrotron in Lund. There is a fluctuation in the baseline signal and a shift before and after the main protein has been eluded, due to the protein being sensitive to X-rays. This will affect the data quality, which is why the samples were measured again. **(b)** SEC elution profile from the second measurements of rArcFL7A on the SWING beamline at the SOLEIL beamline near Paris. The shifting baseline is much better than in Lund due to the X-ray beam being weaker, although there still seems to be some radiation damage. To prevent this and ensure the best data quality, one could add more additives (such as DTT in addition to the TCEP present) to protect the protein from X-ray damage. It could also be possible to run the SEC at a higher flowrate, decreasing the X-ray exposure time of the protein.

SAXS was used to assess the binding sites of the nanobodies and to observe potential structural changes upon nanobody binding. The resulting data is shown in Table 1 and Figure 3.8. The quality of the obtained data was not optimal as there seems to be some aggregation in the samples, which can be seen in the first data points in the scattering curve. The buffer baseline subtraction is also not perfect, due to the slightly shifted baseline before and after the protein

has been eluted. The parameters derived from the scattering curves, however, all indicate that rArcFL7A is a dimer and that E5 and H11 are able to bind simultaneously in a 2:2:2 stoichiometry. The inherent low resolution of SAXS, the intrinsically disordered regions of Arc, and the relatively small size of the nanobodies compared to Arc, meant that accurate mapping of the binding epitopes of the nanobodies and the relative position of Arc's domains was not possible. The experiment did, however, confirm the simultaneous binding of two nanobodies to Arc, and a rough estimate of the binding sites. The binding of E5 to rArcFL7A resulted in a larger change in D_{max} and a smaller change in R_g (Table 2) than the binding of C11 to rArcFL7A. These observations indicate that E5 binds closer than C11 to the end of the longest axis. Since it is known that E5 binds the N-lobe and C11 binds the C-lobe, then those regions would be similarly located.
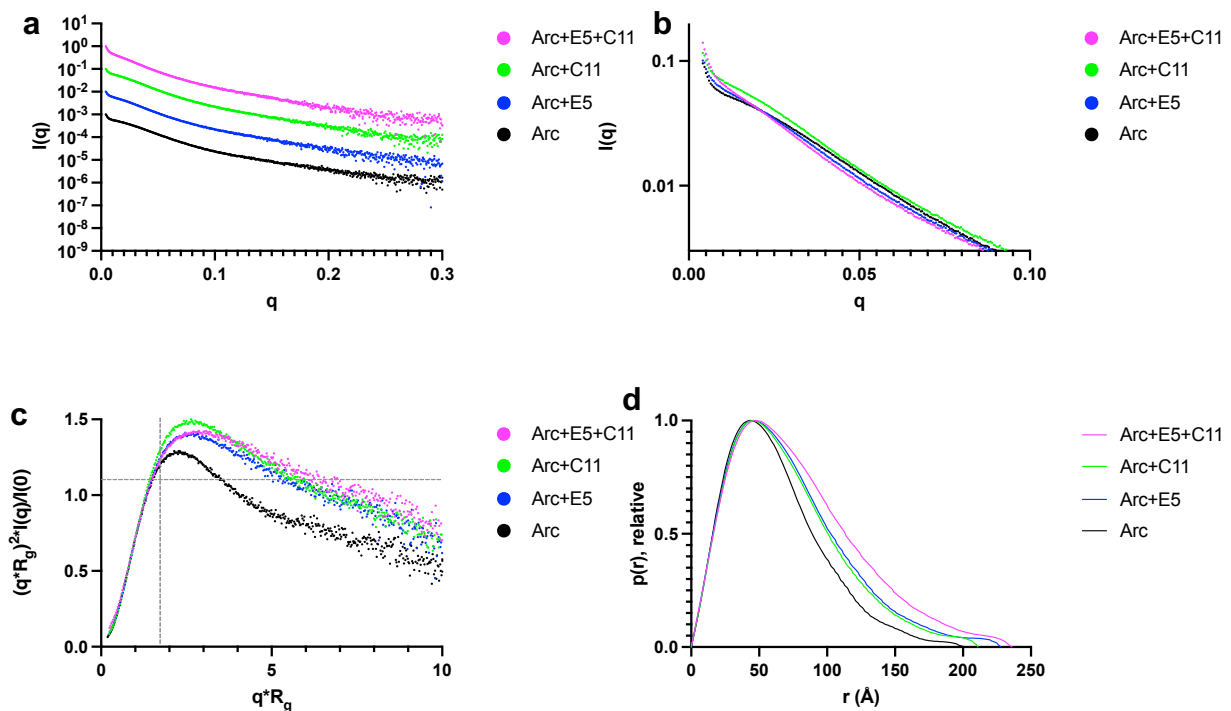
Figure 3.8: Plots obtained from the SAXS datasets. **(a)** Logarithmic representation of the raw scattering curves from the 4 samples, offset by one logarithmic unit along the Y-axis to visualize them better. There is some aggregation in the sample as can be seen by the steep slope of the first data points. **(b)** Highlighted the first region of the 4 scattering curves. The addition of nanobodies slightly changes the shape and size of the object, although not to a great extent. Here the aggregation in the samples can be seen more clearly. **(c)** Dimensionless Kratky plot of the 4 samples. The grey lines mark the maximum of an ideal rigid spherical particle ($\sqrt{3}$, $3/e$). The 4 complexes seem to be similarly rigid. Since Arc has both rigid and flexible domains, this means the Kratky plot will look somewhere in between that of a rigid object and an unfolded object. The addition of nanobodies does not seem to affect the folding and flexibility of Arc, only the size and overall shape. **d)** Pairwise Distance Distribution plot for the 4 samples. The highest X-value is known as the maximum distance within the object (D_{max}). The D_{max} can be seen to increase with the addition of nanobodies, indicating the particle becomes more elongated. E5 seems to have a larger impact on D_{max} than C11, indicating it binds further away

Table 2: Parameters derived from the SAXS measurements of the four complexes. In parenthesis after each value is the relative value compared to rArcFL7A.

Complex	rArcFL7A	+E5	+C11	+E5 and C11
Oligomeric state/ stoichiometry	Dimer	2:2	2:2	2:2:2
Mass (kDa)	90.0	115.6 (+25.6)	116.2 (+26.2)	141.7 (+51.7)
Bayesian Inference Mw est. (kDa)	109.1	130.1 (+21.0)	138.2 (+29.1)	157.1 (+48.0)
R_g *(Å)*	47.38	55.79 (+8.41)	57.17 (+6.66)	59.75 (+12.37)
P(r) R_g *(Å)*	50.51	58.49 (+7.98)	56.56 (+6.05)	63.71 (+13.2)
D_{max} *(Å)*	200	228 (+28)	211 (+11)	236 (+36)
Porod volume ($Å^3$)	230 220	297 891 (+67 671)	293 508 (+63 288)	350 343 (+120 123)
I_0	0.056	0.064 (+0.008)	0.075 (+0.019)	0.067 (+0.011)

This data was used to generate *ab initio* dummy models shown in Figure 3.9. The arrows indicate the possible binding sites of the nanobodies. As can be deduced from the R_g and D_{max} values in Table 2, C11 binds closer to the center of the object than E5 which binds close distal ends. The different domains of Arc are not able to be distinguished in this low-resolution model. This is due to the flexibility of the disordered regions of Arc, resulting in a conformational heterogeneity of the sample. The model-building software tries to create one model when, in reality, there are many different conformations in the sample that all contribute to the scattering signal. It does seem like the N-lobe which binds E5 is located closer to the ends of the object than the C-lobe where C11 is bound. The binding of nanobodies to Arc might also induce fluctuations in the intrinsically disordered regions, further complicating the process. The data and models can be used in comparison with the predicted AlphaFold structures, described below.
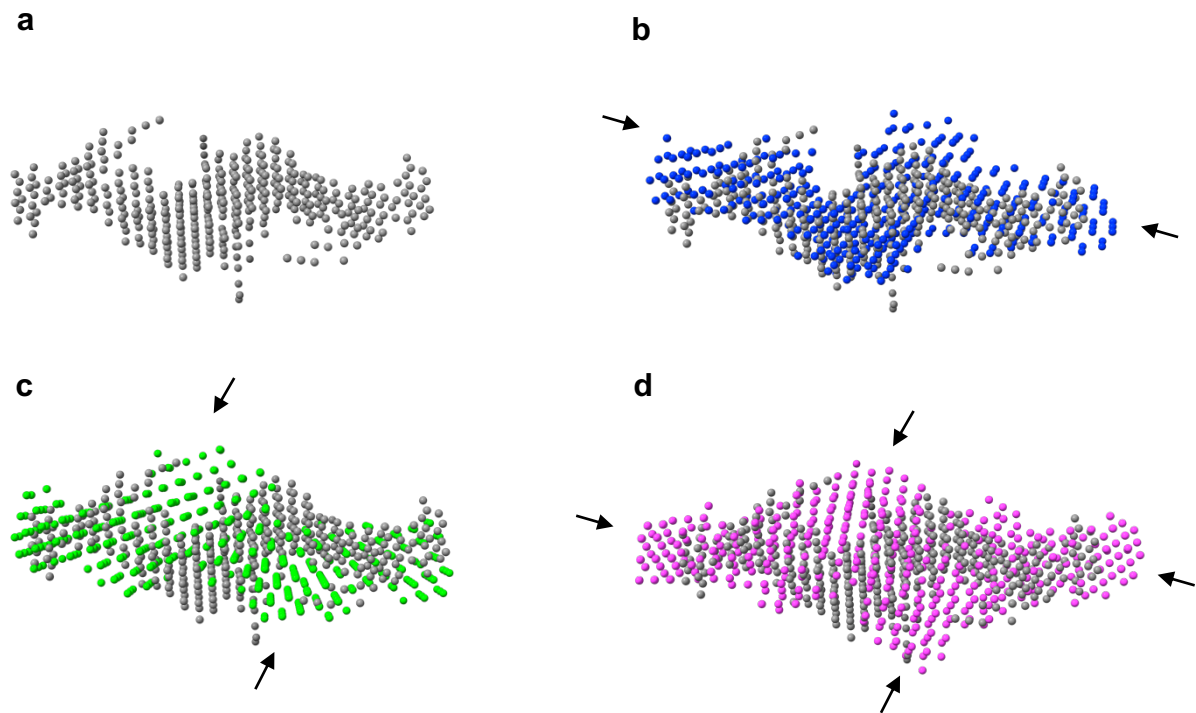
Figure 3.9: *Ab initio* low-resolution DAMMIN models of the rArcFL7A in grey, with E5 in blue, with C11 in green, and with E5 and C11 in magenta. **(a)** Dimeric rArcFL7A, X^2 value of 0.9702 **(b)** Dimeric rArcFL7A with E5 bound (blue), X^2 value of 1.299. The arrows indicate where the nanobodies might be bound. **(c)** rArcFL7A with C11 bound, X^2 value of 1.685. Arrows indicating the nanobody **(d)** rArcFL7A with E5 and C11 bound, X^2 value of 1.169. The two arrows pointing to the distal side of the model indicate E5 binding, and the two central arrows indicate C11 binding. This corresponds nicely with the R_g and D_{max} values presented in Table 2. All models were produced in DAMMIN with forced symmetry.

3.3. Crystallization

The crystallization of proteins is a process that is not very well understood, as described by McPherson and Gavira in their 2013 review. "There is no comprehensive theory, or even a very good base of fundamental data, to guide our efforts, although they are being accumulated at this time. As a consequence, macromolecular crystal growth is largely empirical in nature, and demands patience, perseverance and intuition."[78]. Hence, the approach is largely based on systematic trial and error. A good starting point is often commercially available screens; JCSG+, and PACT-Premier from Molecular Dimensions were used in this study. Crystallization of full-length mArc has been attempted by many and achieved by none. WT Arc oligomerizes and forms capsids[34]. The structural heterogeneity of the protein prevents the expected lattice structure of the crystals. The rArcFL7A mutant does not form higher-order oligomers above the dimer state and is, therefore, considered a better candidate. It does have

multiple disordered regions however, which are problematic for crystallization. The idea here was to use nanobodies as chaperones to hopefully induce crystal formation[132]. This has been done successfully in the lab to crystallize the individual domains of Arc[40,66], and attempted with rArcFL7A in combination with nanobodies H11 and C11. Therefore, it was decided to attempt the crystallization of rArcFL7A with E5 and C11. Different protein concentrations and temperatures were set up, in order to screen as many conditions as possible. No large crystals were observed during this project. There were some promising-looking conditions that might have been able to form protein crystals if optimized as can be seen in Figure 3.10. It is not certain that these UV-positive particles are of the full construct. Previous work to crystallize rArcFL7A has resulted in crystals that only contained the CTD of Arc[66]. Flexible regions are more susceptible to proteolytic degradation than folded regions[133]. It was therefore hypothesized that Arc was cleaved at the linker region by a minor protease contamination, resulting in *in situ* limited proteolysis[134]. The insolubility of the NTD makes it hard to crystallize in isolation, which is why the crystals only contained the CTD. Protein crystallization also requires quite large amounts of protein compared to EM studies. Because of the time constraints in this project, it was decided to focus on EM methods rather than optimizing crystal conditions.

a

b
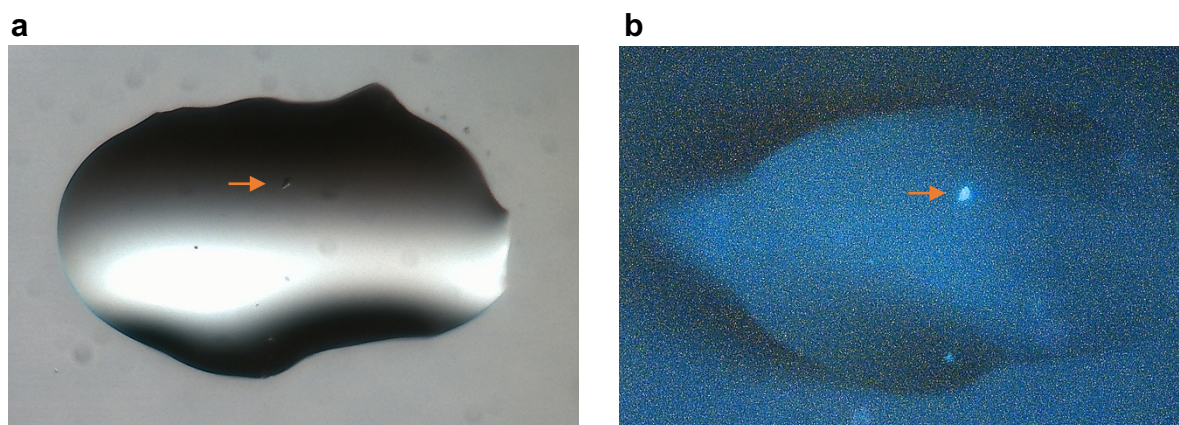
Figure 3.10: Microscopy images of a crystallization drop. **(a)** Image of a drop with a particle indicated by the arrow under visible light. **(b)** The same drop under UV light, indicating that the particle is protein and not a salt crystal. The crystal is not large enough to pick and collect data from. There is also no way of knowing what protein formed this crystal unless data was collected through X-ray diffraction. The crystal could be the full complex, rArcFL7A alone, the nanobodies alone or the CTD. This condition and other similar drops would ideally have been optimized further had there not been time constraints.

3.4. Negative staining TEM

Negative staining TEM works by creating contrast around particles by staining them with uranyl acetate which is a heavy metal salt. Heavier molecules are less transparent to electrons and will appear darker in contrast to the more transparent protein particles. Thus, the background will be stained dark and the proteins lighter, making them possible to observe[135,136]. In this study, negative staining TEM was used to observe sample homogeneity and optimize the conditions for cryo-EM sample preparation[137]. Conditions, such as protein concentration, aggregation, particle size and shape, and the homogeneity of the sample in the grid holes, are all important conditions to consider during cryo-EM. Performing this initial optimization using negative staining TEM is both cheaper and faster than using cryo-EM[138], and there are currently no available cryo-EM facilities in Norway. To apply for time at a high-resolution cryo-EM facility, you are required to perform this preliminary screening in order to reduce the time spent using the high-resolution microscopes, which is a major bottleneck.

Different conditions were tested in order to find the optimal conditions for cryo-EM grid preparation of the rArcFL7A+E5+C11 complex. Protein concentrations ranging from 0.02-1 mg/mL were tested, different staining times with the uranyl acetate, the addition of reducing agents, such as TCEP, to the SEC running buffer in order to reduce aggregation, and whether to wash the grid with buffer or not after applying the protein. In the end, the optimal conditions found were found and described in the methods section. The most meaningful discoveries were the addition of add TCEP to the running buffer and the optimal protein concentration of 0.02-0.05 mg/mL for TEM. In TEM, the grids are dried so that the sample is applied in a thin layer, whereas in cryo-EM the sample is frozen in a much thicker layer inside the grid holes. The heavy metal staining also provides higher contrast than the native conditions in cryo-EM. Therefore, cryo-EM requires a higher protein concentration than TEM to observe the same particle density. A good starting point is to aim for ten times higher protein concentration in cryo-EM than in TEM. In order to know if the particles were of the right size and shape, the images were compared with the SAXS data collected from the same complex. The shapes of such small proteins are not easy to observe at the magnification available to TEM, but they can be determined with enough detail to distinguish certain features. Based on the SAXS data, slightly elongated shapes roughly 20 nm were expected, depending on the orientation of the protein in 3D space. The particles observed in the TEM images are around 20 nm in size and their shape fits nicely with the SAXS data. The particles seem to be distributed evenly throughout the grid with minimal aggregation.
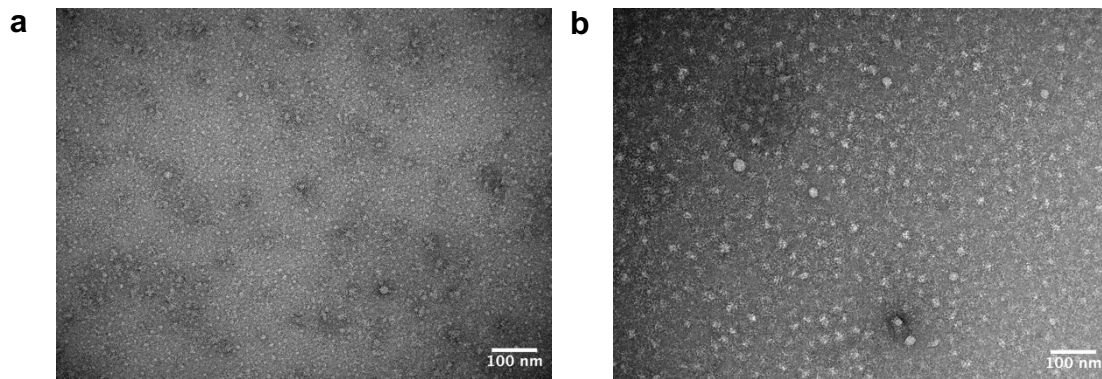
Figure 3.11: Negative staining TEM images of the rArcFL7A+E5+C11 complex at 70 000x magnification. The protein concentration in image **(a)** is 0.02 mg/mL and 0.05 mg/mL in image **(b)**. The particles, which are around 20 nm in size and of the expected shape, seem to be quite monodispersed apart from some aggregation and sticking together seen in some parts of the grid.

3.5. Cryo-EM

Sample preparation, grid preparation, screening and data collection were all performed at UCEM with the help of the staff. The individual proteins had been purified and aliquoted in Bergen and were brought to Umeå on dry ice. The protein complex purification was done on site the same day as the grid preparation, as described in the methods section. The protein concentrations of the resulting fractions were measured and found to range from 0.09-1.89 mg/mL. Aiming for a concentration of 0.2-0.5 mg/mL (10x the concentration used in TEM) as a starting point, three protein concentrations were selected at 0.19, 0.54, and 1.08 mg/mL. Two different grid types were used (Quantifoil 2/1 300 mesh copper grids (Quantifoil Micro Tools GmbH, Jena, Germany) and Quantifoil 2/2 300 mesh copper grids with 2nm carbon layer (Quantifoil Micro Tools GmbH)) and duplicates were prepared for each sample, resulting in a total of 12 grids being prepared. While screening, the highest concentration of protein (1.08 mg/mL) on the grids without a carbon layer looked optimal for data collection. Data collection was set up and resulted in the micrograph and preliminary 2D classes seen in Figure 3.12.

a
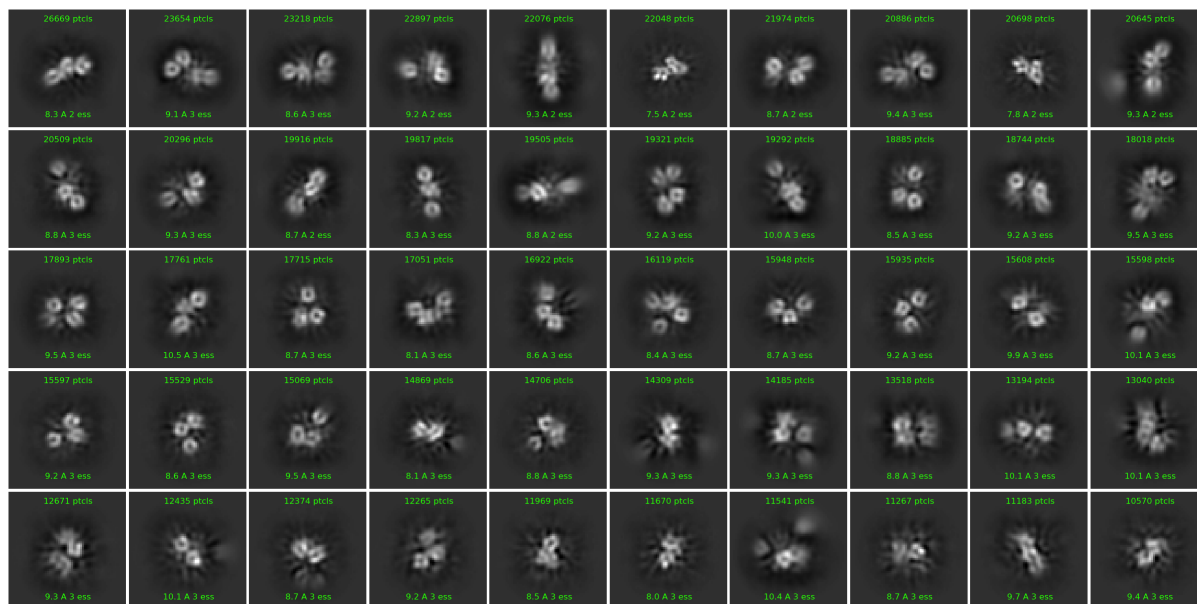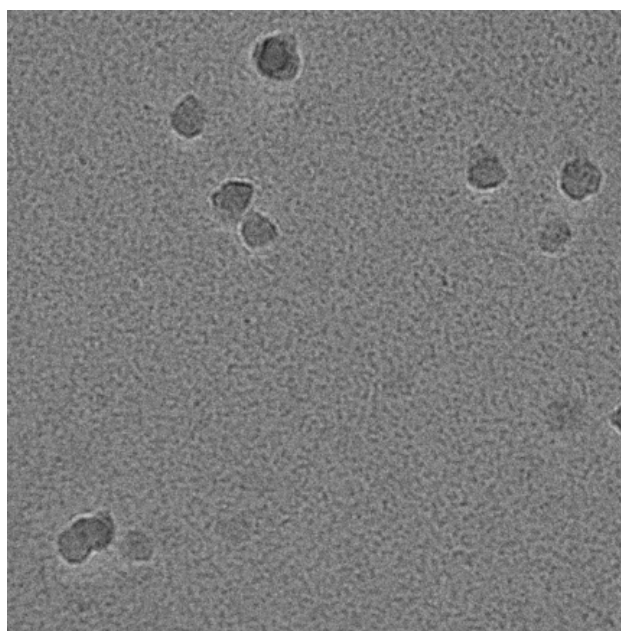
Figure 3.12: Preliminary 2D classes and TEM images. **(a)** These 2D classes were generated from 7000 micrographs on a Titan Glacios cryo-TEM (Thermo Scientific). Even though this is very preliminary data that has not been processed much, it is possible to see distinct particles. **(b)** A representative TEM micrograph from one of the grid holes is shown. The particles can be seen distributed equally across the background, although they are too small to see clearly at this magnification. The larger objects seem to be some ice formation present in all of the images, meaning that the sample preparation could be optimized further for better results. Perhaps changing the Vitrobot settings or type of grid could improve the data quality.

b

3.6. AlphaFold

AlphaFold is a neural network that predicts protein structure from a given amino acid sequence. This AI is trained on all the known protein structures deposited into the PDB. The first step is to perform a multiple sequence alignment (MSA) of the given amino acid sequence. AlphaFold uses the findings from the MSA to gradually build a model and then repeats this step until it has generated a series of residues. Then for each residue, the rotation and translation are

calculated, and the side chains are predicted using a relatively simple network. Once a model has been generated, it is run through a relaxation step to remove violations such as clashing side chains. The final model is then given as output along with two confidence metrics.

The predicted Local Distance Difference Test (pLDDT) measures the percentage of correctly predicted interatomic distances. It essentially acts as a confidence metric of individual domains of the structure and gives values between 0-100, whereby 100 is the most confident. Regions with a low score on the pLDDT should not be interpreted as a predicted structure, but rather as a region that is predicted to be unstructured in isolation. pLDDT is also not a measurement of the confidence of the relative positions of each domain. For that another metric is needed.

To assess the relative positions of each domain with respect to each other, one needs to look at the Predicted Aligned Error (PAE) plot. This metric predicts the position error at residue x, if the predicted and the true structures were aligned on residue y. The expected position error between two residues is depicted by color. Since the confidence tends to be higher within a domain than between domains, the individual domains of a protein can easily be distinguished in this plot[67,107].

AlphaFold was ran with the sequences for human Arc, rat Arc and rArcFL7A (both with and without nanobodies). Five models are given for each sequence, and they are ranked based on the highest total pLDDT score. The models with the highest pLDDT are the ones are shown. Below is shown the predicted structure of rArcFL7A given the amino acid sequence and dimeric state as input. It was also attempted to see if AlphaFold could predict the binding of nanobody E5 which is known to bind to the peptide-binding pocket of the N-lobe.

As can be seen in the AlphaFold models, Arc is predicted to dimerize through an NTD-NTD interaction in a parallel manner with the CTD domains flanking this structure on either side. The pLDDT values show high confidence in the CTD and NTD domains, and low confidence in the disordered regions indicates structural flexibility. This NTD-NTD conformation is present in all of the predicted models, whereas the exact position of the CTD slightly varies in each model. These observations are consistent with the PAE plot, which shows the low certainty of the relative positions of each domain, except for the predicted dimerization of the two NTDs.
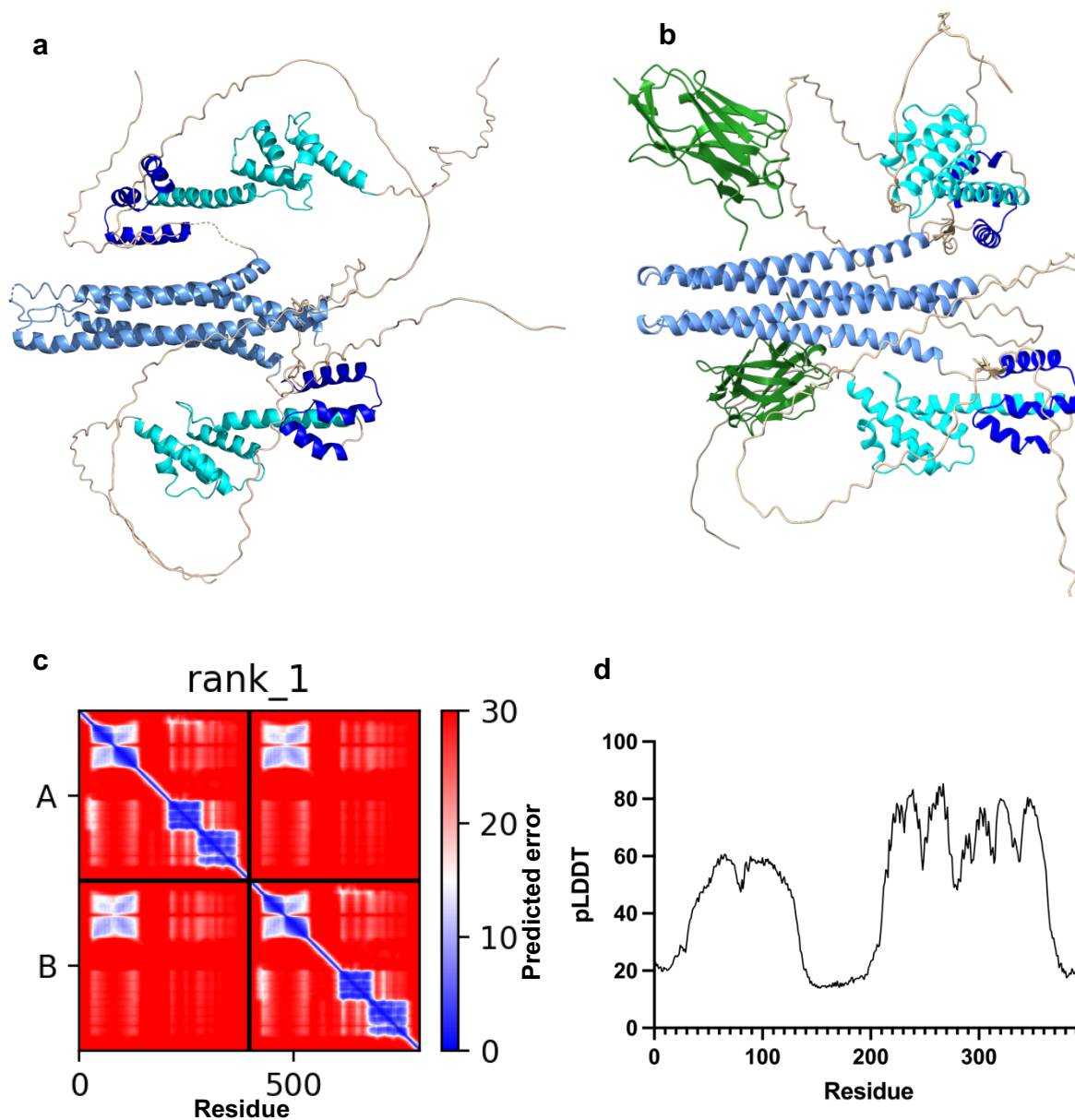
Figure 3.13: AlphaFold models of rArcFL7A with and without nanobody E5, PAE, and pLDDT plots. **(a)** AlphaFold predicted dimeric structure given the sequence of rArcFL7A. The NTDs are colored light blue, N-lobes colored dark blue, and C-lobes cyan. The dimer is predicted to align with the two NTDs in an antiparallel orientation. **(b)** AlphaFold predicted the dimeric structure of rArcFL7A with nanobody E5 bound (colored green). AlphaFold was not able to predict the binding of E5 to Arc with this input. It has been shown from crystal structures that the nanobodies bind either to the peptide binding pocket of the N-lobe or to the C-lobe[66]. **(c)** The PAE plot for the predicted dimeric rArcFL7A structure. The positions of the residues within each domain have very high confidence, whereas the confidence in the relative positions of the different domains is low. Except the position of the NTDs which it has relatively high confidence in. Seeing as there are no structures of how the NTDs dimerize, this is an interesting observation. **(d)** The pLDDT plot shows the confidence in each residue. This plot shows accurately which regions are predicted to be structured (NTD and CTD) and which are disordered (The central linker and both terminal tails).

4. Discussion

4.1. Crystallization

The crystallization process can be quite time-consuming because of the systematic approach. Some proteins crystallize on the first attempt, while others might be impossible to crystallize. It is not possible to prove that a protein cannot be crystallized, whereas one protein crystal will prove it can be. Therefore, it is necessary to try as many different conditions as possible until the protein is crystallized, or time runs out. The somewhat random, trial-and-error nature of crystallization also means that a protein crystal might not be reproducible, even using the same conditions. Because of the limited time available, it was here decided to not continue with the crystallization of the Arc-nanobody complex, and instead focus on EM and SAXS studies. Both of these methods require smaller amounts of protein than crystallization. SAXS has been done successfully before with the same Arc construct and single antibodies[66], whereas crystallization has still not been achieved after many attempts. Single-particle cryo-EM has not been attempted with this dimeric construct and cryo-EM was deemed to be a potential candidate method to provide the first high-resolution structure of full-length mArc.

For crystallization attempts in the future, there are a few things that should be attempted. The first is to try other commercial screens with different concentrations of protein. Another approach is to carefully observe, which conditions look better or worse than the others, and change certain parameters of the most promising drops to screen as many conditions as possible. Such screening includes temperature, pH, buffer composition, precipitant type, and composition. Another solution is to try the seeding of microcrystals. If there are wells with small crystals that are too small or fragile to diffract, they can be crushed up and used as seeds in new drops. Using a hair, these microcrystals are streaked across a new drop in order to act as nuclei for crystal growth[139]. The addition of nanobodies as crystallization chaperones has been used successfully for other proteins[132]. The idea is that the nanobodies change the surface of the protein complex, making it more likely to form the crystal lattice structure[66]. This methodology was applied successfully in order to crystallize the mammalian Arc CTD with C11 and H11 and the N-lobe in combination with E5 and C11 separately[66]. There are even more antigen-binding fragments that can be added to nanobodies to make the complex even larger[97]. These new nanobody constructs were obtained, but due to time limitations, they were not able to be expressed and purified. The time it takes for a crystal to grow is also highly variable. Some crystallize within a few days, whereas others can take months. If no crystals are

obtained, then there is not much data to be analyzed and discussed. Because the MSc project only lasts a year, and given the previous unsuccessful attempts to crystallize full-length mArc, crystallization was deemed more of a side project rather than the main focus of this project.

4.2. SAXS and AlphaFold
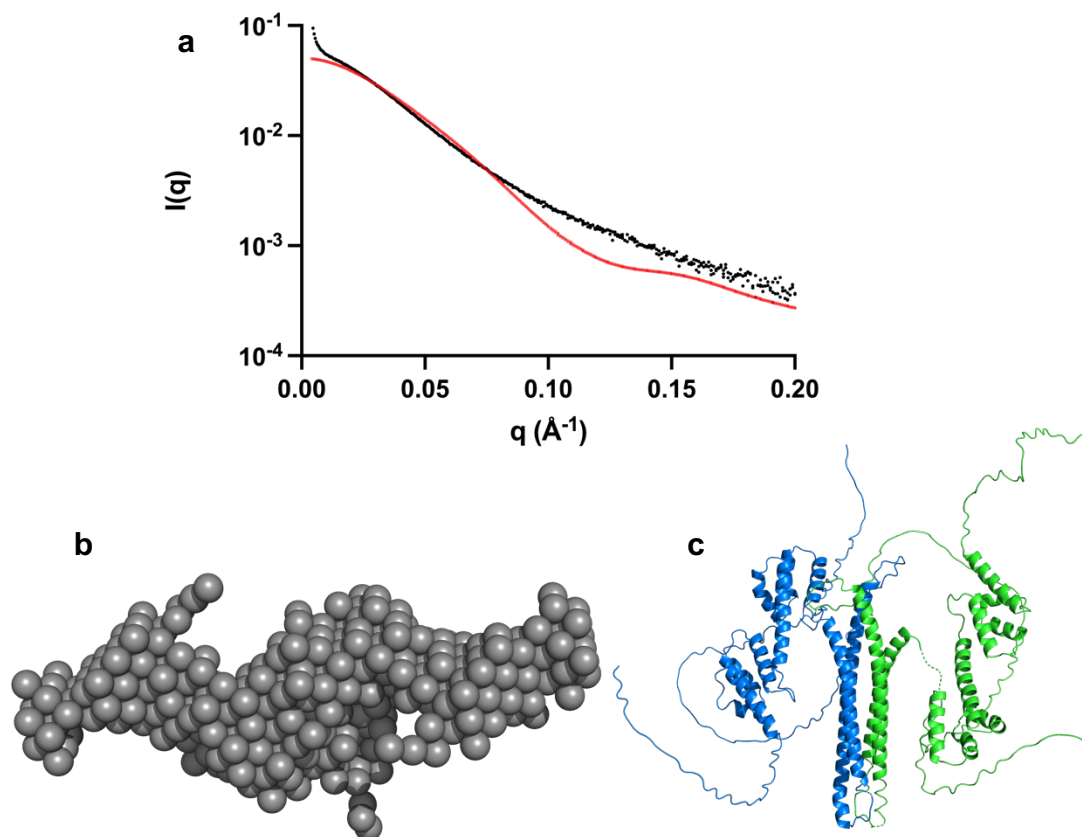
Figure 4.1: Comparing the AlphaFold model with the SAXS data. **(a)** The predicted AlphaFold model of rArcFL7A (red), fitted to the SAXS data for the same complex (black) using CRYSOL. The fit has a X^2 value of 65.82. AlphaFold predicts a more globular structure than what can be seen from the SAXS data. This is likely due to AlphaFold being trained on crystal structures which tend to be more compacted than structures in solution. **(b)** The DAMMIN model of rArcFL7A constructed from the SAXS dataset. **(c)** The AlphaFold predicted structure of rArcFL7A with each monomer colored. This is one possible way the structure could fit inside the DAMMIN model. As the two models are different, it is hard to compare these models. It does show how different the structure of rArcFL7A in solution and the predicted structure are. This is likely due to AlphaFold struggling to predict flexible proteins and oligomers.

As far as we know, mArc consists of two structured domains (NTD and CTD) with a flexible linker between, and a disordered tail on both termini. When comparing the SAXS data obtained from the rArcFL7A dimer and the predicted model (Figure 4.1), it becomes obvious the fit is not good. The predicted structure looks more globular than the more extended structure of Arc in solution, which can be seen in the shape of the scattering curves. The same can be seen when looking at the DAMMIN model next to the AlphaFold model. AlphaFold is trained mainly on crystal structures deposited to the PDB[67], and will therefore be biased towards predicting a

structure that looks closer to what the protein would look like had it been crystallized. However, the SAXS data is measured from the protein in solution, not in a crystal lattice. Proteins in a crystal lattice are more compact than in solution, and these packing interactions between the proteins will inhibit the conformational flexibility of the proteins[140]. Protein structures derived from crystals will, therefore, have a lower conformational flexibility than the protein in solution[141]. Thus, AlphaFold might be inclined to predict structures that would fit in a crystal lattice, rather than a less compacted structure found in solution[142,143]. In addition, since IDPs are notoriously hard to crystallize, the majority of the crystal structures deposited to the PDB are of highly structured proteins. The number of experimentally determined structures of conformationally structured IDRs is tiny compared to the structures of folded proteins, and these structures are not deposited to the PDB[144,145]. For these proteins, one often uses NMR for small proteins, or low-resolution methods such as SAXS. This again means that AlphaFold will have fewer examples of IDPs to learn from, contributing to the bias in predicting more compacted and structured regions. IDPs do not exist in a single 3D folded structure but rather fluctuate between a variety of conformations[146]. Since Arc contains three of these disordered regions, there is a lot of conformational heterogeneity in the sample, making it impossible to extract one correct model from the SAXS data. If AlphaFold predicts the disordered regions to be in one of these conformations, and the SAXS models indicate another conformation, the models will not match. This could also be the reason why the models of the four complexes derived from the SAXS data look so different (Figure 3.9).

The pLDDT value given by AlphaFold are a confidence score given to each residue of the protein. pLDDT values >90 can be considered highly accurate, values >70 are considered to be mostly accurate, scores >50 are considered to have low accuracy, and values <50 are considered very low accuracy. Regions within the lower threshold of <50 can be considered unstructured [147]. Over 90% of the residues in the solved structures deposited in the PDB have a pLDDT score >70. That number goes down to below 40% when looking at the unresolved structures in the PDB[147]. Some of those unsolved disordered proteins might be part of the "dark proteome", proteins that have no detectable similarity to any structures in the PDB[148]. A large part of this dark proteome was illuminated by the emergence of AlphaFold, going from 26% to 10% of the human proteome when including the AlphaFold database[149]. The proteins in this dark proteome might have a structure, but no discovered structural homologs. Therefore, AlphaFold is unable to predict the structure, as it uses solved homologous structures to generate its prediction. This does not mean that all regions given a low pLDDT score have a structure

that has not been solved yet. Instead, a region with a low pLDDT value should be considered intrinsically disordered[147,150], although a portion of these regions might have a structure that has not yet been solved[151]. Another portion of these regions with low pLDDT values belongs to the category of conditionally disordered proteins: domains that are disordered in isolation, and fold when interacting with their biological targets[152,153]. A recent study found that AlphaFold is able to predict with high accuracy the folded state of 60% of these conditionally disordered proteins[154], which backs up the assumption that regions with low pLDDT values should be considered fully disordered. Following this rule of thumb, Arc's central linker region should be considered fully disordered, although it might still be folded when in the correct environment, for example interacting with a binding partner. Hallin et al. showed that ligand binding did not alter the conformation of the CTD, by collecting SAXS data from the CTD of Arc with and without ligands bound[69]. However, in a subsequent study using molecular dynamics simulations, it was shown that the N-terminal strand of the N-lobe adjoining the linker region folds into a β-sheet, stabilizing the structure[42]. This observation entertains the possibility that ligand binding might impact the flexibility of the linker region, and thus the relative orientation of the two domains[29]. The two oppositely charged domains are predicted to interact with each other, and the presence of a flexible linker region between the two domains suggests conformational flexibility[155]. Perhaps this linker region opens or closes when interacting with some of its partners such as mRNA, other Arc subunits in the capsid, or another unknown partner[29]. The linker region could be conditionally folded in the right environment, for example in the capsid structure. Understanding this aspect would better the understanding of how Arc performs its many functions. Having high-resolution structures of Arc in its different oligomeric states would provide much-needed information. Using AlphaFold to guide the discovery of new protein structures will provide more data that AlphaFold can be trained on, in turn leading to increased accuracy of AlphaFold. This positive feedback loop combined with the continuous improvement of cryo-EM and other structural biology methods, will help the community get closer to reaching the end goal of mapping the structure of the entire proteome.

4.3. The role of disordered proteins and regions

IDPs are thought to make up about 30% of the human proteome[146,156]. Intrinsic disorder is common among all eukaryotes and is also present in viral proteins to a lesser extent[157]. Eukaryotes have evolved to be more complex than prokaryotes, and so has the eukaryotic proteome in multiple ways. Eukaryotic proteins are more disordered[158], more likely to include multiple domains[159], contain more PTMs, and are longer in general[160]. The more complex eukaryotes require more precise regulation and signaling than prokaryotes. This could in part be explained by the increased rate of IDPs, which play a key role in the regulatory processes of the cell[161,162]. IDPs lack a fixed fold, making them unable to perform the role of an enzyme. Instead, they often act as interaction hubs[163] and bind to other partners, such as DNA, RNA, or other proteins[110]. The functional flexibility of IDRs is further increased by their high likelihood to be modified post-translationally[164,165]. Disordered linkers are not only more abundant in eukaryotes than prokaryotes but are also more disordered and have an altered amino acid composition[166]. Whether this difference in amino acid composition is the reason behind or the cause of higher disorder is not known.

mArc has two disordered terminal tails and a flexible linker region fitting the description of an IDR. Not much is known about this linker connecting the two domains. Arc shares many of the same characteristics as common IDPs described above. mArc binds to PSD95 as well as other proteins and acts as a protein interaction hub[50,51], it is targeted by PTMs such as palmitoylation (although this occurs at the structured NTD and not a flexible region)[70], and binding to its ligand is hypothesized to induce a conformational change in its flexible regions[29]. Arc is even able to form large virus-like capsids, which internalize its own mRNA and can transfect nearby cells[34]. Whether the Arc disordered linker region is directly involved in any of these functions is unclear. Considering the flexible functionality of Arc, it is hard to look past its flexible structure as an important factor. Arc can also be found in many different oligomeric states, monomeric, dimeric, tetrameric, large oligomeric, and capsid. Eriksen and Bramham hypothesize in their recent review that the different oligomeric states of Arc are responsible for each of these functions and that the oligomeric state is determined by PTMs, protein interactions, and the local concentration of Arc[29]. Perhaps these determining factors induce structural changes in the linker region, resulting in Arc preferring one oligomeric state over the other.

4.4. Arc capsids
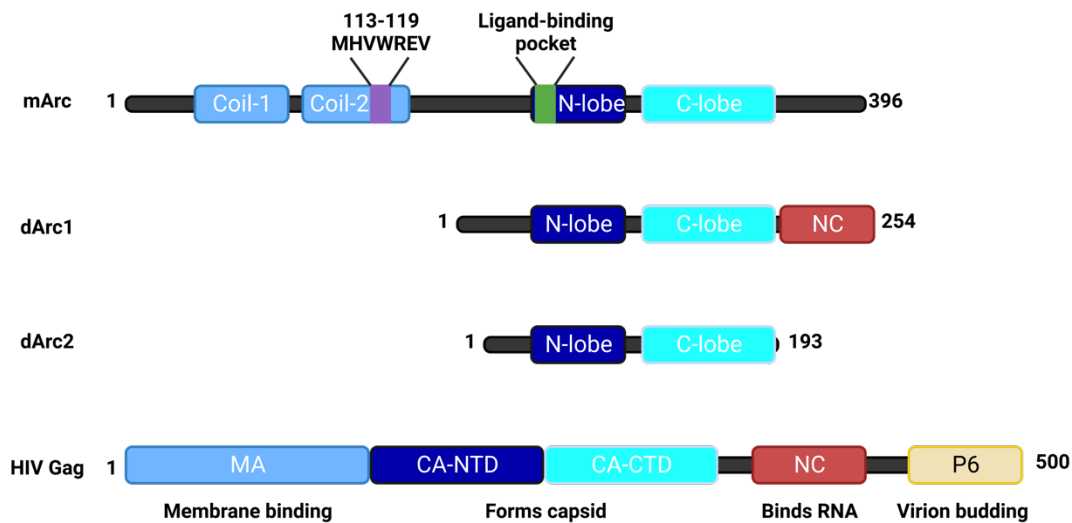
Figure 4.2: Structural homology of mammalian Arc, *Drosophila* Arc 1 and 2, and HIV Gag protein. All proteins belong to the Ty3/Gypsy family of retrotransposons and are able to form capsids. dArc1 and dArc2 are lacking a structured N-terminal domain. A capsid forming, bi-lobar C-terminal domain is conserved in all 4 proteins. dArc1 and HIV Gag protein contain a C-terminal RNA binding domain. The critical oligomerization motif (purple) and ligand-binding pocket (green) can be found only in mArc. Figure adapted from[29] and created with BioRender.com

Human Arc was discovered to have conserved homology to the retroviral polyprotein Gag[68], suggesting that Arc has evolved from the retrotransposon element Ty3/Gypsy. Arc can be found as two genes in *Drosophila* (dArc1 and dArc2), where the CA domain is structurally similar to the mArc CTD and Gag protein[41]. The main structural difference between mArc and dArc is the presence of a structured N-terminal domain in mArc. These observations indicate that mArc and dArc have been repurposed from different retrotransposable elements in the Ty3/Gypsy family in separate evolutionary events[34,167]. Both mArc and dArc are able to self-assemble into virus-like capsids, encapsulating their own mRNA. These capsids are then released from the neuron in extracellular vesicles, which are taken up by nearby cells, allowing the RNA to be released and translated in these cells[34,167]. The two versions of Arc found in *Drosophila* are similar, apart from the C-terminal structured zinc-finger pair homologous to the nucleocapsid (NC) domain of retroviruses found in dArc1, which is predicted to bind dArc1 mRNA[39]. The CA domain of dArc is oligomeric in solution, whereas the CTD of mArc is monomeric. Hallin et al. used truncated dArc2 N-lobe constructs lacking the N-terminal tail to obtain a dimeric crystal structure[40]. This domain-swapped dimeric construct does not match the penta/hexameric structure found in the dArc2 capsids[39], suggesting the N-terminal tail plays a role in the oligomerization and subsequent capsid formation.
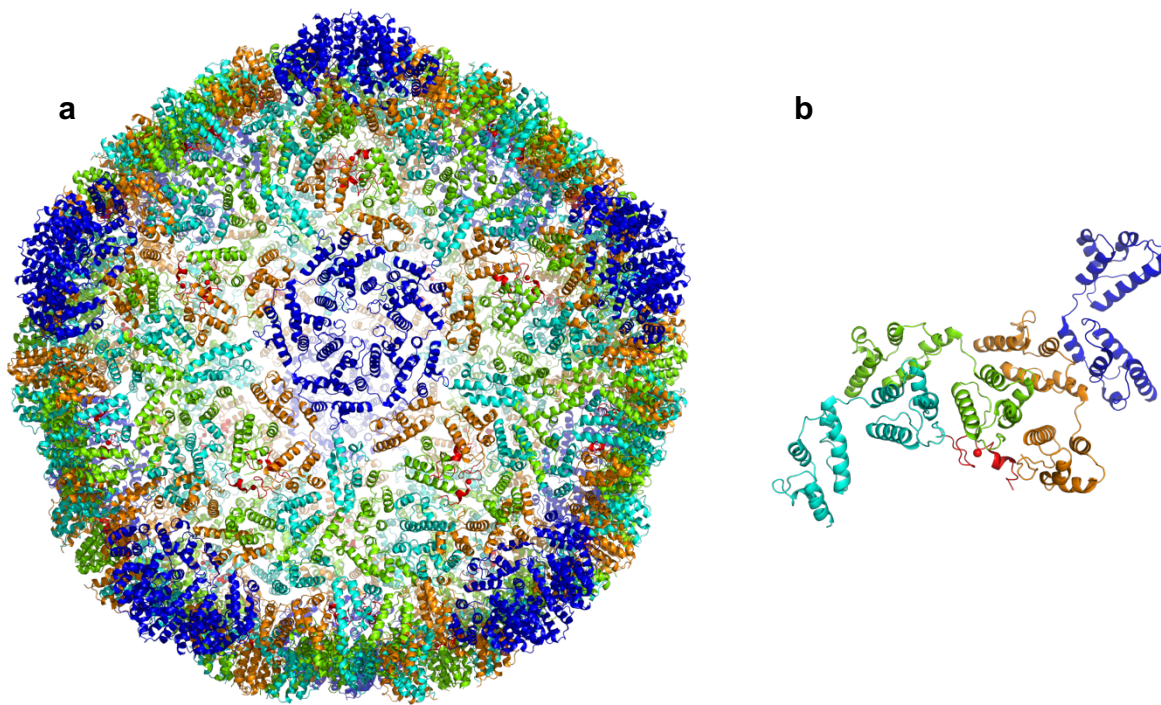
Figure 4.3: dArc1 capsid and asymmetrical unit (PDB ID: 6TAP). **(a)** Structure of dArc1 capsid solved using cryo-EM[39]. The capsids consist of 240 dArc1 subunits, making up 60 asymmetrical units, each containing 4 dArc1 proteins and one zinc finger domain. The capsid consists of 12 pentagonal five-fold capsomeres and 30 hexagonal two-fold capsomers. The N-lobes of 5 dArc1 subunits come together at the five-fold capsomers to form this pentagonal structure seen in blue. The N-terminal tails (not shown in this structure) are predicted to protrude outwards from these structures to form the spikes. There are also unresolved electron densities on the inside of the capsid, theorized to be the N-terminal tails forming spikes that stretch inwards to bind RNA non-specifically. The zinc finger domains can be seen in red on the inside of the capsids, likely to bind dArc1 RNA as it was found to be enriched in these capsids. **(b)** The five dArc1 copies in each asymmetrical unit are colored blue, orange, green and cyan, and the zinc finger is colored red.

The structures of dArc1 and dArc2 capsids have been solved using cryo-EM, revealing a 240-unit icosahedral capsid[39]. dArc lacks a structured N-terminal domain and instead has an N-terminal tail, which is folded as an amphipathic α-helix in the capsid structure, forming spikes that protrude both inward and outward from the capsid layer. These spikes are hypothesized to mediate both membrane binding on the outside of the capsid, and mRNA binding on the inside. The inward-facing spikes might be responsible for non-specific mRNA binding, whereas the NC domain specifically recognizes dArc1 RNA. Apart from dArc1 mRNA the contents of these capsids are not known, but there is likely other RNA bound by these spikes. It is possible that the N-terminal tail of dArc has evolved to perform the same role as the NTD in mArc. Less is known about the mArc NTD, as there are no crystal structures of the full NTD. Part of the coil 2 of the NTD of hArc has been crystallized, residues 99-132, containing the critical oligomerization motif [113]MHVWREV[119]. This structure proved that this region in isolation will dimerize in an antiparallel manner, with the residues M113 and W116 being crucial to this

interaction[65]. Mutations in this 7-amino acid stretch inhibit oligomerization and limit Arc to a dimeric structure, as in the case with the rArcFL7A mutant. Whether full-length mArc dimers bind in this antiparallel manner to align the M113 of one monomer with the W116 of the other monomer is not known. The NTD in isolation and the full-length protein have not been successfully crystallized, and the low resolution of methods like SAXS does not give enough insight to observe details of binding. AlphaFold predicts the rArcFL7A mutant to dock the two NTDs in a parallel rather than an antiparallel way (Figure 3.13), and the confidence is relatively high as can be seen in the PAE plot. Thus, the oligomerization motif could mediate binding between two Arc dimers instead of two monomers to form tetramers. This could be the reason why the rArcFL7A mutant is able to dimerize, but not able to form higher-order oligomers beyond the dimer.

The NTD is positively charged and has been proven to promote binding to membranes[65,70]. By its positive charge, it is also possible that the NTD mediates the non-specific binding of mRNA. This extended helical structure formed by the NTDs observed in the AlphaFold models could in theory function the same way as the spikes of the dArc capsids. The outwards-facing spikes could mediate membrane binding, and the inwards-facing spikes could perform mRNA binding. This is of course hard to prove without any high-resolution structures of any of the oligomeric states of mArc. Given the flexible structure of Arc, it seems like capsid could be the most promising candidate out of the different states of Arc. Cryo-EM was used successfully to solve the structure of both dArc1 and dArc2 capsids[39], proving it can be done.
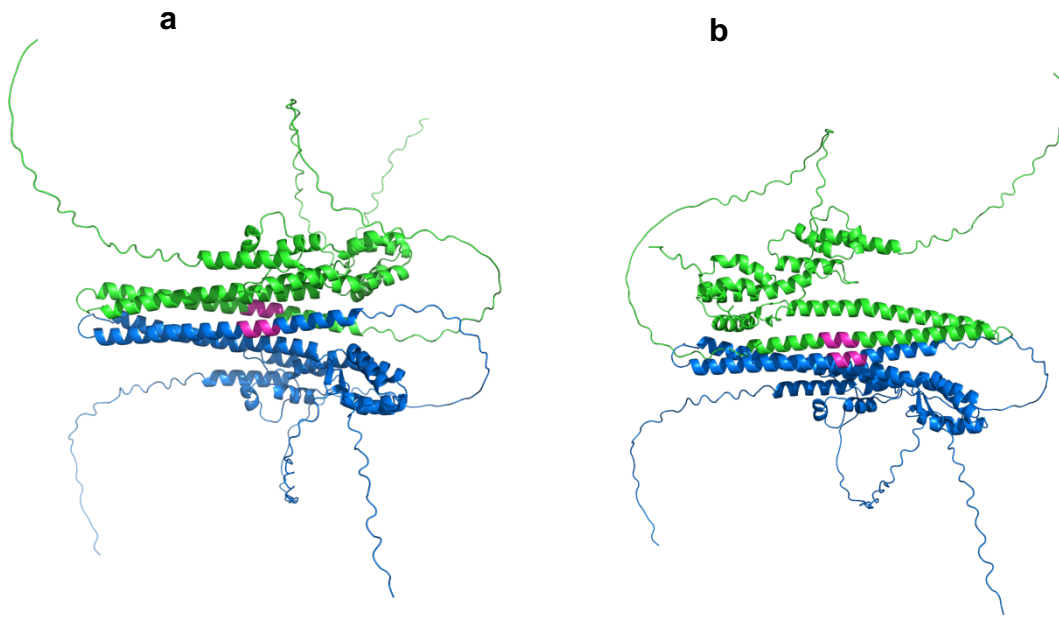
Figure. 4.4: Two different ways mArc can dimerize. **(a)** The AlphaFold predicted structure of dimeric hArc. The two subunits are colored green and blue, and the oligomerization motif magenta. Arc can be seen to dimerize along the NTDs in a parallel manner. According to the PAE plot (Figure 3.13 c) the confidence in the relative position of the two NTDs is relatively high compared to the rest of the structure. This way the two oligomerization motifs do not align to dock in the way seen in the crystal structure of the second coil of the NTD of hArc[65]. Instead, the oligomerization motifs could mediate binding between two of these dimers. Another interesting observation is that these parallel α-helical structures might resemble the spikes found on the outside of retrovirus capsids. The NTD of mArc are positively charged and have been shown to bind membranes[65,70]. **(b)** A different possible way Arc can dimerize, with the NTDs binding in an antiparallel manner. This means the oligomerization motifs might align as was seen in the aforementioned crystal structure of the second coil of the NTD of hArc.

4.5. TEM and cryo-EM

As with most methods in structural biology, flexible proteins are not optimal for single-particle cryo-EM. Flexible regions connecting the structured domains can be orientated in many different ways. This means that the same protein can have different conformations, which all look different in the 2D micrograph. When trying to piece together all the 2D images into a 3D structure, it becomes much harder when the protein is flexible. The same protein observed from the same angle can look different, depending on the conformation of the flexible regions. This structural heterogeneity makes it hard to create the different 2D classifications when the orientation of the protein is unknown. Therefore, it is unlikely to get a near-atomic resolution structure of flexible proteins of this small size. Still, it is possible to acquire a lower-resolution map. Unfortunately, it was not possible to process the datasets which were collected, because the cryo-EM workstation in the lab stopped working and had to be sent for repair. Luckily, preliminary 2D classes could be obtained as the data was being collected using CryoSPARC Live[122], and the micrographs can be looked at individually.

Distinct sphere-like objects can clearly be seen in the 2D classes. The number and relative position of these spheres vary, although there never appear to be more than four in each 2D class. Each of these spheres might be one of the four domains of the Arc dimer. Since some of the images show less than four spheres might be because some of them could be positioned behind the other domains in their shadow so to speak. Because of the flexibility of the linker region, the relative position of the individual domains could be changed. This makes it hard to tell if the 3D orientation of the protein is different, or if the domains of the protein are in different conformations. This problem can be helped by adding nanobody complexes that break the symmetry of the protein. Adding an asymmetrical antibody-fragment (Fab) to the protein will further make 3D orientation much quicker and more accurate. These Fabs will also increase the size of the protein complex, making it possible to study proteins smaller than the general lower threshold of 100 kDa[168,169]. New nanobodies that bind to both Arc and the synthetic Fab were generated and purchased from GeneArt (Thermo Scientific) by hybridizing nanobody C11 with the nanobody that binds to these Fab fragments[97]. Unfortunately, they could not be expressed and purified in time for the cryo-EM data collection shown here. They would most likely not have made the complex less flexible, but it would have made it easier to identify the different structural elements and their relative orientation, and the collected particles would have been larger in size.
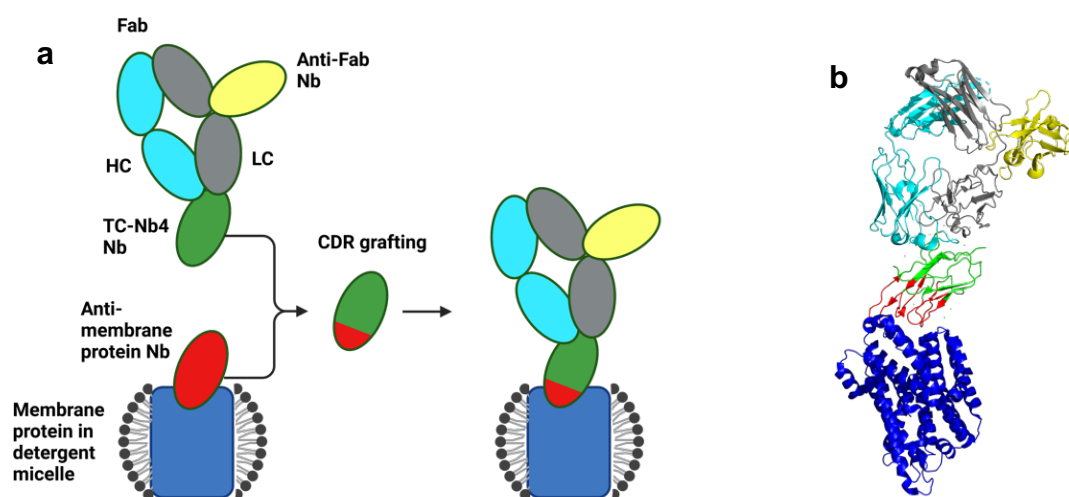
Figure 4.5: Nanobody-binding antigen-binding fragment (NabFab) complex. **(a)** In blue can be seen the target protein, in this case a membrane protein bound in a detergent micelle. This protein could be any protein as long as a nanobody has been immunized against it. The CDR loops of this nanobody (red) are grafted with the TC-Nb4 nanobody (green) which binds to the synthetic antigen-binding fragment (Fab). This Fab consists of a heavy chain (cyan) and a light chain (grey) which provide the rigid scaffold. A second antibody (yellow) binds to the distal end of the light chain to break symmetry, making it easier to determine the orientation of the particle. **(b)** The same complex visualized as a 3D protein structure solved using cryo-EM (PDB ID: 7PIJ). The CDR loops (red) on one end of the nanobody (green) can be seen binding to the membrane protein, while the Fab (cyan and grey) can be seen binding to the distal end of the CDR loops. The Fab scaffold was designed to not interfere with the protein of interest. Figure adapted from[97] and created with BioRender.com

4.6. Future prospects

The propensity of mArc to oligomerize makes it difficult to achieve a homogenous sample. The different states of Arc are somehow regulated in the cell, whether it is through PTMs, interactions with other proteins or macromolecules, localization in the cell, pH, or other unknown factors. If these determining factors were better understood they could be replicated *in vitro,* enabling the purification of stable Arc in its different states. The crystallization of Arc has proven to be difficult with the multiple flexible regions. Perhaps Arc is less flexible and willing to crystallize in another state than the dimer. Nanobodies have been shown to compete with Arc's natural ligands. It might also be possible to generate nanobodies that bind to the oligomerization motif of the NTD to prevent Arc from oligomerizing. Adding even more Fabs onto these nanobodies could change the surface of the protein enough for it to become stable in the crystal lattice, making it possible to solve the crystal structure. These same nanobodies and Fabs could be used to optimize the single-particle cryo-EM study of dimeric Arc. Perhaps new nanobodies could stabilize the flexible regions of Arc, making it more suitable for single-particle cryo-EM. Another interesting application for single-particle cryo-EM would be to study the mammalian Arc capsids. Imaging large symmetrical structures such as virus capsids is one of the strongest features of cryo-EM. The main issue here would be to purify the Arc capsids so that the sample would be as homogenous as possible. The anti-Arc nanobodies have been shown to co-purify with the capsids without disrupting their shape[66], and Fabs could be used to make the capsids more homogenous and rigid. This is proven to be a hurdle but if overcome, achieving a high-resolution structure of the mammalian Arc capsids could be possible. As more structures are solved, the accuracy of AlphaFold and other AI-based prediction models increases, especially concerning disordered or oligomeric proteins which is currently a big limitation. These predictions have already proven their usefulness as a tool to use alongside experimental structural biology methods, and their influence will only increase as the knowledge within the field expands. Using these rapidly advancing tools, it should be possible to solve the structure of full-length mArc in its different oligomeric states. Not only will these structures provide direct information about the function of Arc, but the knowledge will enable functional studies as well. This will lead to a deeper understanding of synaptic plasticity and memory, and hopefully treatment of diseases where Arc is involved.

5. Conclusion

The structure and function of proteins go hand in hand. The 3D folding of the polypeptide confers the protein its function. Thus, to fully understand the role of a protein, it is important to know its 3D structure. Arc is a flexible multidomain protein that is a master regulator of synaptic plasticity. Arc exerts its effect on synaptic plasticity both as a hub protein at the post-synaptic density of glutamatergic neurons and through the transfer of mRNA to nearby cells in virus-like capsids formed by Arc. The structural information on Arc is quite limited due to its flexibility and willingness to oligomerize, which makes crystallization challenging. The use of AI-based structure prediction models was compared to structural information of dimeric Arc in solution, indicating the limitations of AlphaFold when predicting the structure of flexible oligomeric proteins. The data presented in this study introduces new ideas about the oligomerization of Arc, and the use of nanobodies as tools both for structural and functional studies of proteins. Single-particle cryo-EM datasets were collected but not yet processed, which might contain the first high-resolution structure of full-length mammalian Arc. With the emergence of cryo-EM as the up-and-coming method for solving protein structures, AlphaFold as a way of predicting 3D structures and nanobodies as a highly versatile set of tools, many unsolved protein structures will be solved in the near future.

6. References

1. Dale Purves GJA, David Fitzpatrick, William C. Hall, Anthony-Samuel LaMantia, Richard D. Mooney, Michael L. Platt, and Leonard E. White. *Neuroscience*. 6th ed. Oxford University Press; 2017.

2. Palay SL, Palade GE. The fine structure of neurons. *J Biophys Biochem Cytol*. Jan 1955;1(1):69-88. doi:10.1083/jcb.1.1.69

3. De Robertis ED, Bennett HS. Some features of the submicroscopic morphology of synapses in frog and earthworm. *J Biophys Biochem Cytol*. Jan 1955;1(1):47-58. doi:10.1083/jcb.1.1.47

4. von Bartheld CS, Bahney J, Herculano-Houzel S. The search for true numbers of neurons and glial cells in the human brain: A review of 150 years of cell counting. *J Comp Neurol*. Dec 15 2016;524(18):3865-3895. doi:10.1002/cne.24040

5. Hartline DK. What is myelin? *Neuron Glia Biol*. May 2008;4(2):153-63. doi:10.1017/s1740925x09990263

6. Bélanger M, Magistretti PJ. The role of astroglia in neuroprotection. *Dialogues Clin Neurosci*. 2009;11(3):281-95. doi:10.31887/DCNS.2009.11.3/mbelanger

7. Zeng H, Sanes JR. Neuronal cell-type classification: challenges, opportunities and the path forward. *Nature Reviews Neuroscience*. 2017/09/01 2017;18(9):530-546. doi:10.1038/nrn.2017.85

8. Burkhardt P, Colgren J, Medhus A, et al. Syncytial nerve net in a ctenophore adds insights on the evolution of nervous systems. *Science*. 2023;380(6642):293-297. doi:doi:10.1126/science.ade5645

9. Pereda AE. Electrical synapses and their functional interactions with chemical synapses. *Nature Reviews Neuroscience*. 2014/04/01 2014;15(4):250-263. doi:10.1038/nrn3708

10.   Altman J, Das GD. Autoradiographic and histological evidence of postnatal hippocampal neurogenesis in rats. *Journal of Comparative Neurology*. 1965;124(3):319-335. doi:https://doi.org/10.1002/cne.901240303

11.   Altman J, Das GD. Postnatal Neurogenesis in the Guinea-pig. *Nature*. 1967/06/01 1967;214(5093):1098-1101. doi:10.1038/2141098a0

12.   Knoth R, Singec I, Ditter M, et al. Murine Features of Neurogenesis in the Human Hippocampus across the Lifespan from 0 to 100 Years. *PLOS ONE*. 2010;5(1):e8809. doi:10.1371/journal.pone.0008809

13.   Sorrells SF, Paredes MF, Cebrian-Silla A, et al. Human hippocampal neurogenesis drops sharply in children to undetectable levels in adults. *Nature*. 2018/03/01 2018;555(7696):377-381. doi:10.1038/nature25975

14.   Citri A, Malenka RC. Synaptic Plasticity: Multiple Forms, Functions, and Mechanisms. *Neuropsychopharmacology*. 2008/01/01 2008;33(1):18-41. doi:10.1038/sj.npp.1301559

15.   Hebb DO. *The organisation of behaviour*. Wiley and Sons; 1949.

16.   Milner PM. The cell assembly: Mark II. *Psychological Review*. 1957;64:242-252. doi:10.1037/h0042287

17.   Stent GS. A physiological mechanism for Hebb's postulate of learning. *Proc Natl Acad Sci U S A*. Apr 1973;70(4):997-1001. doi:10.1073/pnas.70.4.997

18.   Martin SJ, Grimwood PD, Morris RG. Synaptic plasticity and memory: an evaluation of the hypothesis. *Annu Rev Neurosci*. 2000;23:649-711. doi:10.1146/annurev.neuro.23.1.649

19.   Abraham WC, Jones OD, Glanzman DL. Is plasticity of synapses the mechanism of long-term memory storage? *npj Science of Learning*. 2019/07/02 2019;4(1):9. doi:10.1038/s41539-019-0048-y

20. Bliss TV, Lomo T. Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *J Physiol*. Jul 1973;232(2):331-56. doi:10.1113/jphysiol.1973.sp010273

21. Nicoll RA. A Brief History of Long-Term Potentiation. *Neuron*. 2017/01/18/ 2017;93(2):281-290. doi:https://doi.org/10.1016/j.neuron.2016.12.015

22. Bliss TVP, Collingridge GL. A synaptic model of memory: long-term potentiation in the hippocampus. *Nature*. 1993/01/01 1993;361(6407):31-39. doi:10.1038/361031a0

23. Malenka RC, Kauer JA, Perkel DJ, et al. An essential role for postsynaptic calmodulin and protein kinase activity in long-term potentiation. *Nature*. 1989;340(6234):554-557.

24. Malenka RC, Kauer JA, Zucker RS, Nicoll RA. Postsynaptic Calcium Is Sufficient for Potentiation of Hippocampal Synaptic Transmission. *Science*. 1988;242(4875):81-84. doi:doi:10.1126/science.2845577

25. Hayashi Y. Molecular mechanism of hippocampal long-term potentiation – Towards multiscale understanding of learning and memory. *Neuroscience Research*. 2022/02/01/ 2022;175:3-15. doi:https://doi.org/10.1016/j.neures.2021.08.001

26. Collingridge GL, Peineau S, Howland JG, Wang YT. Long-term depression in the CNS. *Nature Reviews Neuroscience*. 2010/07/01 2010;11(7):459-473. doi:10.1038/nrn2867

27. Yashiro K, Philpot BD. Regulation of NMDA receptor subunit expression and its implications for LTD, LTP, and metaplasticity. *Neuropharmacology*. Dec 2008;55(7):1081-94. doi:10.1016/j.neuropharm.2008.07.046

28. Malenka RC, Bear MF. LTP and LTD: an embarrassment of riches. *Neuron*. Sep 30 2004;44(1):5-21. doi:10.1016/j.neuron.2004.09.012

29. Eriksen MS, Bramham CR. Molecular physiology of Arc/Arg3.1: The oligomeric state hypothesis of synaptic plasticity. *Acta Physiologica*. 2022;236(3):e13886. doi:https://doi.org/10.1111/apha.13886

30.   Steward O, Wallace CS, Lyford GL, Worley PF. Synaptic activation causes the mRNA for the IEG Arc to localize selectively near activated postsynaptic sites on dendrites. *Neuron*. Oct 1998;21(4):741-51. doi:10.1016/s0896-6273(00)80591-7

31.   Guzowski JF, McNaughton BL, Barnes CA, Worley PF. Environment-specific expression of the immediate-early gene Arc in hippocampal neuronal ensembles. *Nature Neuroscience*. 1999/12/01 1999;2(12):1120-1124. doi:10.1038/16046

32.   Plath N, Ohana O, Dammermann B, et al. Arc/Arg3.1 is essential for the consolidation of synaptic plasticity and memories. *Neuron*. Nov 9 2006;52(3):437-44. doi:10.1016/j.neuron.2006.08.024

33.   Shepherd JD, Rumbaugh G, Wu J, et al. Arc/Arg3.1 Mediates Homeostatic Synaptic Scaling of AMPA Receptors. *Neuron*. 2006/11/09/ 2006;52(3):475-484. doi:https://doi.org/10.1016/j.neuron.2006.08.034

34.   Pastuzyn ED, Day CE, Kearns RB, et al. The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer. *Cell*. Jan 11 2018;172(1-2):275-288.e18. doi:10.1016/j.cell.2017.12.024

35.   Smit AFA. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Current Opinion in Genetics & Development*. 1999/12/01/ 1999;9(6):657-663. doi:https://doi.org/10.1016/S0959-437X(99)00031-3

36.   Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. *Nature Reviews Genetics*. 2017/02/01 2017;18(2):71-86. doi:10.1038/nrg.2016.139

37.   Kaneko-Ishino T, Ishino F. The role of genes domesticated from LTR retrotransposons and retroviruses in mammals. *Front Microbiol*. 2012;3:262. doi:10.3389/fmicb.2012.00262

38.   Feschotte C, Pritham EJ. DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet*. 2007;41:331-68. doi:10.1146/annurev.genet.40.110405.090448

39.   Erlendsson S, Morado DR, Cullen HB, Feschotte C, Shepherd JD, Briggs JAG. Structures of virus-like capsids formed by the Drosophila neuronal Arc proteins. *Nat Neurosci*. Feb 2020;23(2):172-175. doi:10.1038/s41593-019-0569-y

40.   Hallin EI, Markússon S, Böttger L, Torda AE, Bramham CR, Kursula P. Crystal and solution structures reveal oligomerization of individual capsid homology domains of Drosophila Arc. *PLoS One*. 2021;16(5):e0251459. doi:10.1371/journal.pone.0251459

41.   Cottee MA, Letham SC, Young GR, Stoye JP, Taylor IA. Structure of Drosophila melanogaster ARC1 reveals a repurposed molecule with characteristics of retroviral Gag. *Sci Adv*. Jan 2020;6(1):eaay6354. doi:10.1126/sciadv.aay6354

42.   Hallin EI, Bramham CR, Kursula P. Structural properties and peptide ligand binding of the capsid homology domains of human Arc. *Biochem Biophys Rep*. Jul 2021;26:100975. doi:10.1016/j.bbrep.2021.100975

43.   Nielsen LD, Pedersen CP, Erlendsson S, Teilum K. The Capsid Domain of Arc Changes Its Oligomerization Propensity through Direct Interaction with the NMDA Receptor. *Structure*. 2019/07/02/ 2019;27(7):1071-1081.e5. doi:https://doi.org/10.1016/j.str.2019.04.001

44.   Nikolaienko O, Patil S, Eriksen MS, Bramham CR. Arc protein: a flexible hub for synaptic plasticity and cognition. *Seminars in Cell & Developmental Biology*. 2018/05/01/ 2018;77:33-42. doi:https://doi.org/10.1016/j.semcdb.2017.09.006

45.   Zhang W, Wu J, Ward Matthew D, et al. Structural Basis of Arc Binding to Synaptic Proteins: Implications for Cognitive Disease. *Neuron*. 2015/04/22/ 2015;86(2):490-500. doi:https://doi.org/10.1016/j.neuron.2015.03.030

46.   Chowdhury S, Shepherd JD, Okuno H, et al. Arc/Arg3.1 Interacts with the Endocytic Machinery to Regulate AMPA Receptor Trafficking. *Neuron*. 2006/11/09/ 2006;52(3):445-459. doi:https://doi.org/10.1016/j.neuron.2006.08.033

47.   DaSilva LL, Wall MJ, L PdA, et al. Activity-Regulated Cytoskeleton-Associated Protein Controls AMPAR Endocytosis through a Direct Interaction with Clathrin-Adaptor Protein 2. *eNeuro*. May-Jun 2016;3(3)doi:10.1523/eneuro.0144-15.2016

48.   Okuno H, Akashi K, Ishii Y, et al. Inverse Synaptic Tagging of Inactive Synapses via Dynamic Interaction of Arc/Arg3.1 with CaMKIIβ. *Cell*. 2012/05/11/ 2012;149(4):886-898. doi:https://doi.org/10.1016/j.cell.2012.02.062

49.   Volianskis A, Bannister N, Collett VJ, et al. Different NMDA receptor subtypes mediate induction of long-term potentiation and two forms of short-term potentiation at CA1 synapses in rat hippocampus in vitro. *J Physiol*. Feb 15 2013;591(4):955-72. doi:10.1113/jphysiol.2012.247296

50.   Fernández E, Collins MO, Frank RAW, et al. Arc Requires PSD95 for Assembly into Postsynaptic Complexes Involved with Neural Dysfunction and Intelligence. *Cell Rep*. Oct 17 2017;21(3):679-691. doi:10.1016/j.celrep.2017.09.045

51.   Frank RAW, Zhu F, Komiyama NH, Grant SGN. Hierarchical organization and genetically separable subfamilies of PSD95 postsynaptic supercomplexes. *J Neurochem*. Aug 2017;142(4):504-511. doi:10.1111/jnc.14056

52.   Messaoudi E, Kanhema T, Soulé J, et al. Sustained Arc/Arg3.1 synthesis controls long-term potentiation consolidation through regulation of local actin polymerization in the dentate gyrus in vivo. *J Neurosci*. Sep 26 2007;27(39):10445-55. doi:10.1523/jneurosci.2883-07.2007

53.   Fukazawa Y, Saitoh Y, Ozawa F, Ohta Y, Mizuno K, Inokuchi K. Hippocampal LTP is accompanied by enhanced F-actin content within the dendritic spine that is essential for late LTP maintenance in vivo. *Neuron*. May 8 2003;38(3):447-60. doi:10.1016/s0896-6273(03)00206-x

54.   Nair RR, Patil S, Tiron A, et al. Dynamic Arc SUMOylation and Selective Interaction with F-Actin-Binding Protein Drebrin A in LTP Consolidation In Vivo. *Front Synaptic Neurosci*. 2017;9:8. doi:10.3389/fnsyn.2017.00008

55.   Wee CL, Teo S, Oey NE, Wright GD, VanDongen HM, VanDongen AM. Nuclear Arc Interacts with the Histone Acetyltransferase Tip60 to Modify H4K12 Acetylation(1,2,3). *eNeuro*. Nov-Dec 2014;1(1)doi:10.1523/eneuro.0019-14.2014

56.   Korb E, Wilkinson CL, Delgado RN, Lovero KL, Finkbeiner S. Arc in the nucleus regulates PML-dependent GluA1 transcription and homeostatic plasticity. *Nat Neurosci*. Jul 2013;16(7):874-83. doi:10.1038/nn.3429

57.   Nikolaienko O, Eriksen MS, Patil S, Bito H, Bramham CR. Stimulus-evoked ERK-dependent phosphorylation of activity-regulated cytoskeleton-associated protein (Arc) regulates its neuronal subcellular localization. *Neuroscience*. Sep 30 2017;360:68-80. doi:10.1016/j.neuroscience.2017.07.026

58.   Walczyk-Mooradally A, Holborn J, Singh K, et al. Phosphorylation-dependent control of Activity-regulated cytoskeleton-associated protein (Arc) protein by TNIK. *J Neurochem*. Sep 2021;158(5):1058-1073. doi:10.1111/jnc.15440

59.   Palop JJ, Chin J, Bien-Ly N, et al. Vulnerability of dentate granule cells to disruption of arc expression in human amyloid precursor protein transgenic mice. *J Neurosci*. Oct 19 2005;25(42):9686-93. doi:10.1523/jneurosci.2829-05.2005

60.   Leung HW, Foo G, VanDongen A. Arc Regulates Transcription of Genes for Plasticity, Excitability and Alzheimer's Disease. *Biomedicines*. Aug 11 2022;10(8)doi:10.3390/biomedicines10081946

61.   Park S, Park JM, Kim S, et al. Elongation factor 2 and fragile X mental retardation protein control the dynamic translation of Arc/Arg3.1 essential for mGluR-LTD. *Neuron*. Jul 10 2008;59(1):70-83. doi:10.1016/j.neuron.2008.05.023

62.   Peebles CL, Yoo J, Thwin MT, Palop JJ, Noebels JL, Finkbeiner S. Arc regulates spine morphology and maintains network stability in vivo. *Proc Natl Acad Sci U S A*. Oct 19 2010;107(42):18173-8. doi:10.1073/pnas.1006546107

63.   Toro R, Konyukh M, Delorme R, et al. Key role for gene dosage and synaptic homeostasis in autism spectrum disorders. *Trends Genet*. Aug 2010;26(8):363-72. doi:10.1016/j.tig.2010.05.007

64.   Greer PL, Hanayama R, Bloodgood BL, et al. The Angelman Syndrome protein Ube3A regulates synapse development by ubiquitinating arc. *Cell*. Mar 5 2010;140(5):704-16. doi:10.1016/j.cell.2010.01.026

65.   Eriksen MS, Nikolaienko O, Hallin EI, et al. Arc self-association and formation of virus-like capsids are mediated by an N-terminal helical coil motif. *The FEBS Journal*. 2021;288(9):2930-2955. doi:https://doi.org/10.1111/febs.15618

66.   Markússon S, Hallin EI, Bustad HJ, et al. High-affinity anti-Arc nanobodies provide tools for structural and functional studies. *PLoS One*. 2022;17(6):e0269281. doi:10.1371/journal.pone.0269281

67.   Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021/08/01 2021;596(7873):583-589. doi:10.1038/s41586-021-03819-2

68.   Campillos M, Doerks T, Shah PK, Bork P. Computational characterization of multiple Gag-like human proteins. *Trends in Genetics*. 2006/11/01/ 2006;22(11):585-589. doi:https://doi.org/10.1016/j.tig.2006.09.006

69.   Hallin EI, Eriksen MS, Baryshnikov S, et al. Structure of monomeric full-length ARC sheds light on molecular flexibility, protein interactions, and functional modalities. *Journal of Neurochemistry*. 2018;147(3):323-343. doi:https://doi.org/10.1111/jnc.14556

70.   Barylko B, Wilkerson JR, Cavalier SH, et al. Palmitoylation and Membrane Binding of Arc/Arg3.1: A Potential Role in Synaptic Depression. *Biochemistry*. Feb 6 2018;57(5):520-524. doi:10.1021/acs.biochem.7b00959

71.   Hey A. *Feynman And Computation*. 1st ed. CRC Press; 2002.

72.   Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of Molecular Biology*. 1999/10/22/ 1999;293(2):321-331. doi:https://doi.org/10.1006/jmbi.1999.3110

73.   Anfinsen CB. Principles that Govern the Folding of Protein Chains. *Science*. 1973;181(4096):223-230. doi:doi:10.1126/science.181.4096.223

74.   Goldberger RF, Epstein CJ, Anfinsen CB. Acceleration of Reactivation of Reduced Bovine Pancreatic Ribonuclease by a Microsomal System from Rat Liver. *Journal of Biological Chemistry*. 1963/02/01/ 1963;238(2):628-635. doi:https://doi.org/10.1016/S0021-9258(18)81309-6

75.   Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. *Nucleic Acids Research*. 2000;28(1):235-242. doi:10.1093/nar/28.1.235

76.   Consortium TU. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*. 2022;51(D1):D523-D531. doi:10.1093/nar/gkac1052

77.   Smyth MS, Martin JH. x ray crystallography. *Mol Pathol*. Feb 2000;53(1):8-14. doi:10.1136/mp.53.1.8

78.   McPherson A, Gavira JA. Introduction to protein crystallization. *Acta Crystallogr F Struct Biol Commun*. Jan 2014;70(Pt 1):2-20. doi:10.1107/s2053230x13033141

79.   Nogales E. The development of cryo-EM into a mainstream structural biology technique. *Nat Methods*. Jan 2016;13(1):24-7. doi:10.1038/nmeth.3694

80.   Dubochet J. On the Development of Electron Cryo-Microscopy (Nobel Lecture). *Angew Chem Int Ed Engl*. Aug 20 2018;57(34):10842-10846. doi:10.1002/anie.201804280

81.   Frank J. Single-Particle Reconstruction of Biological Molecules-Story in a Sample (Nobel Lecture). *Angew Chem Int Ed Engl*. Aug 20 2018;57(34):10826-10841. doi:10.1002/anie.201802770

82.   Henderson R. From Electron Crystallography to Single Particle CryoEM (Nobel Lecture). *Angew Chem Int Ed Engl*. Aug 20 2018;57(34):10804-10825. doi:10.1002/anie.201802731

83.   Cianfrocco MA, Kellogg EH. What Could Go Wrong? A Practical Guide to Single-Particle Cryo-EM: From Biochemistry to Atomic Models. *J Chem Inf Model*. May 26 2020;60(5):2458-2469. doi:10.1021/acs.jcim.9b01178

84.   Wu S, Avila-Sakar A, Kim J, et al. Fabs enable single particle cryoEM studies of small proteins. *Structure*. Apr 4 2012;20(4):582-92. doi:10.1016/j.str.2012.02.017

85.   Kalienkova V, Alvadia C, Clerico Mosina V, Paulino C. Single-Particle Cryo-EM of Membrane Proteins in Lipid Nanodiscs. *Methods Mol Biol*. 2020;2127:245-273. doi:10.1007/978-1-0716-0373-4_17

86.   Kursula P. Small-angle X-ray scattering for the proteomics community: current overview and future potential. *Expert Rev Proteomics*. Jun 2021;18(6):415-422. doi:10.1080/14789450.2021.1951242

87.   Bernadó P, Mylonas E, Petoukhov MV, Blackledge M, Svergun DI. Structural characterization of flexible proteins using small-angle X-ray scattering. *J Am Chem Soc*. May 2 2007;129(17):5656-64. doi:10.1021/ja069124n

88.   Cordeiro TN, Herranz-Trillo F, Urbanek A, et al. Structural Characterization of Highly Flexible Proteins by Small-Angle Scattering. *Adv Exp Med Biol*. 2017;1009:107-129. doi:10.1007/978-981-10-6038-0_7

89.   Kachala M, Valentini E, Svergun DI. Application of SAXS for the Structural Characterization of IDPs. *Adv Exp Med Biol*. 2015;870:261-89. doi:10.1007/978-3-319-20164-1_8

90.   Molodenskiy DS, Mertens HDT, Svergun DI. An automated data processing and analysis pipeline for transmembrane proteins in detergent solutions. *Sci Rep*. May 15 2020;10(1):8081. doi:10.1038/s41598-020-64933-1

91.   Koutsioubas A. Low-Resolution Structure of Detergent-Solubilized Membrane Proteins from Small-Angle Scattering Data. *Biophys J*. Dec 5 2017;113(11):2373-2382. doi:10.1016/j.bpj.2017.10.003

92.   Gräwert TW, Svergun DI. Structural Modeling Using Solution Small-Angle X-ray Scattering (SAXS). *J Mol Biol*. Apr 17 2020;432(9):3078-3092. doi:10.1016/j.jmb.2020.01.030

93.   Bizien T, Durand D, Roblina P, Thureau A, Vachette P, Pérez J. A Brief Survey of State-of-the-Art BioSAXS. *Protein Pept Lett*. 2016;23(3):217-31. doi:10.2174/0929866523666160106153655

94.   Aznauryan M, Delgado L, Soranno A, et al. Comprehensive structural and dynamical view of an unfolded protein from the combination of single-molecule FRET, NMR, and SAXS. *Proc Natl Acad Sci U S A*. Sep 13 2016;113(37):E5389-98. doi:10.1073/pnas.1607193113

95.   Kikhney AG, Svergun DI. A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Letters*. 2015/09/14/ 2015;589(19, Part A):2570-2577. doi:https://doi.org/10.1016/j.febslet.2015.08.027

96.   Hamers-Casterman C, Atarhouch T, Muyldermans S, et al. Naturally occurring antibodies devoid of light chains. *Nature*. 1993/06/01 1993;363(6428):446-448. doi:10.1038/363446a0

97.   Bloch JS, Mukherjee S, Kowal J, et al. Development of a universal nanobody-binding Fab module for fiducial-assisted cryo-EM studies of membrane proteins. *Proceedings of the National Academy of Sciences*. 2021;118(47):e2115435118. doi:doi:10.1073/pnas.2115435118

98.   Dong J-X, Lee Y, Kirmiz M, et al. A toolbox of nanobodies developed and validated for use as intrabodies and nanoscale immunolabels in mammalian brain neurons. *eLife*. 2019/09/30 2019;8:e48750. doi:10.7554/eLife.48750

99.   Beghein E, Gettemans J. Nanobody Technology: A Versatile Toolkit for Microscopic Imaging, Protein–Protein Interaction Analysis, and Protein Function Exploration. Review. *Frontiers in Immunology*. 2017-July-04 2017;8doi:10.3389/fimmu.2017.00771

100.   Li T, Vandesquille M, Koukouli F, et al. Camelid single-domain antibodies: A versatile tool for in vivo imaging of extracellular and intracellular brain targets. *Journal of Controlled Release*. 2016/12/10/ 2016;243:1-10. doi:https://doi.org/10.1016/j.jconrel.2016.09.019

101.   Gerdes C, Waal N, Offner T, et al. A nanobody-based fluorescent reporter reveals human α-synuclein in the cell cytosol. *Nature Communications*. 2020/06/01 2020;11(1):2729. doi:10.1038/s41467-020-16575-0

102.   Muyldermans S. A guide to: generation and design of nanobodies. *The FEBS Journal*. 2021;288(7):2084-2102. doi:https://doi.org/10.1111/febs.15515

103.   Ishizuka Y, Mergiya TF, Baldinotti R, et al. Development and Validation of Arc Nanobodies: New Tools for Probing Arc Dynamics and Function. *Neurochem Res*. Sep 2022;47(9):2656-2666. doi:10.1007/s11064-022-03573-5

104.   Steeland S, Vandenbroucke RE, Libert C. Nanobodies as therapeutics: big opportunities for small antibodies. *Drug Discovery Today*. 2016/07/01/ 2016;21(7):1076-1113. doi:https://doi.org/10.1016/j.drudis.2016.04.003

105.   Senior AW, Evans R, Jumper J, et al. Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins: Structure, Function, and Bioinformatics*. 2019;87(12):1141-1148. doi:https://doi.org/10.1002/prot.25834

106.   Waterhouse A, Bertoni M, Bienert S, et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res*. Jul 2 2018;46(W1):W296-w303. doi:10.1093/nar/gky427

107.   Evans R, O'Neill M, Pritzel A, et al. Protein complex prediction with AlphaFold-Multimer. *bioRxiv*. 2021:2021.10.04.463034. doi:10.1101/2021.10.04.463034

108.   Choi UB, Sanabria H, Smirnova T, Bowen ME, Weninger KR. Spontaneous Switching among Conformational Ensembles in Intrinsically Disordered Proteins. *Biomolecules*. 2019;9(3):114.

109.   Strodel B. Energy Landscapes of Protein Aggregation and Conformation Switching in Intrinsically Disordered Proteins. *Journal of Molecular Biology*. 2021/10/01/ 2021;433(20):167182. doi:https://doi.org/10.1016/j.jmb.2021.167182

110.   van der Lee R, Buljan M, Lang B, et al. Classification of Intrinsically Disordered Regions and Proteins. *Chemical Reviews*. 2014/07/09 2014;114(13):6589-6631. doi:10.1021/cr400525m

111.   Laemmli UK. Cleavage of Structural Proteins during the Assembly of the Head of Bacteriophage T4. *Nature*. 1970/08/01 1970;227(5259):680-685. doi:10.1038/227680a0

112.   Fox JD, Waugh DS. Maltose-binding protein as a solubility enhancer. *Methods Mol Biol*. 2003;205:99-117. doi:10.1385/1-59259-301-1:99

113.   Nallamsetty S, Austin BP, Penrose KJ, Waugh DS. Gateway vectors for the production of combinatorially-tagged His6-MBP fusion proteins in the cytoplasm and periplasm of Escherichia coli. *Protein Sci*. Dec 2005;14(12):2964-71. doi:10.1110/ps.051718605

114.   Schindelin J, Arganda-Carreras I, Frise E, et al. Fiji: an open-source platform for biological-image analysis. *Nature Methods*. 2012/07/01 2012;9(7):676-682. doi:10.1038/nmeth.2019

115.   Kahnt M, Klementiev K, Haghighat V, et al. Measurement of the coherent beam properties at the CoSAXS beamline. *J Synchrotron Radiat*. Nov 1 2021;28(Pt 6):1948-1953. doi:10.1107/s1600577521009140

116.   Panjkovich A, Svergun DI. CHROMIXS: automatic and interactive analysis of chromatography-coupled small-angle X-ray scattering data. *Bioinformatics*. Jun 1 2018;34(11):1944-1946. doi:10.1093/bioinformatics/btx846

117. Manalastas-Cantos K, Konarev PV, Hajizadeh NR, et al. ATSAS 3.0: expanded functionality and new tools for small-angle scattering data analysis. *J Appl Crystallogr*. Feb 1 2021;54(Pt 1):343-355. doi:10.1107/s1600576720013412

118. Svergun DI. Determination of the regularization parameter in indirect-transform methods using perceptual criteria. *Journal of Applied Crystallography*. 1992;25(4):495-503. doi:https://doi.org/10.1107/S0021889892001663

119. Svergun DI. Restoring Low Resolution Structure of Biological Macromolecules from Solution Scattering Using Simulated Annealing. *Biophysical Journal*. 1999/06/01/ 1999;76(6):2879-2886. doi:https://doi.org/10.1016/S0006-3495(99)77443-6

120. Svergun DI, Petoukhov MV, Koch MH. Determination of domain structure of proteins from X-ray solution scattering. *Biophys J*. Jun 2001;80(6):2946-53. doi:10.1016/s0006-3495(01)76260-1

121. Franke D, Petoukhov MV, Konarev PV, et al. ATSAS 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *J Appl Crystallogr*. Aug 1 2017;50(Pt 4):1212-1225. doi:10.1107/s1600576717007786

122. Punjani A, Rubinstein JL, Fleet DJ, Brubaker MA. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nature Methods*. 2017/03/01 2017;14(3):290-296. doi:10.1038/nmeth.4169

123. https://cosmic-cryoem.org/

124. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. *Nature Methods*. 2022/06/01 2022;19(6):679-682. doi:10.1038/s41592-022-01488-1

125. *The PyMOL Molecular Graphics System*. Version 2.5.4 Schrödinger. LLC;

126. Pettersen EF, Goddard TD, Huang CC, et al. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci*. Jan 2021;30(1):70-82. doi:10.1002/pro.3943

127.   Busso D, Delagoutte-Busso B, Moras D. Construction of a set Gateway-based destination vectors for high-throughput cloning and expression screening in Escherichia coli. *Analytical Biochemistry*. 2005/08/15/ 2005;343(2):313-321. doi:https://doi.org/10.1016/j.ab.2005.05.015

128.   Larkin MA, Blackshields G, Brown NP, et al. Clustal W and Clustal X version 2.0. *Bioinformatics*. Nov 1 2007;23(21):2947-8. doi:10.1093/bioinformatics/btm404

129.   Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics*. May 1 2009;25(9):1189-91. doi:10.1093/bioinformatics/btp033

130.   Lobstein J, Emrich CA, Jeans C, Faulkner M, Riggs P, Berkmen M. SHuffle, a novel Escherichia coli protein expression strain capable of correctly folding disulfide bonded proteins in its cytoplasm. *Microbial Cell Factories*. 2012/05/08 2012;11(1):753. doi:10.1186/1475-2859-11-56

131.   Brookes E, Vachette P, Rocco M, Pérez J. US-SOMO HPLC-SAXS module: dealing with capillary fouling and extraction of pure component patterns from poorly resolved SEC-SAXS data. *J Appl Crystallogr*. Oct 1 2016;49(Pt 5):1827-1841. doi:10.1107/s1600576716011201

132.   Duhoo Y, Roche J, Trinh TTN, et al. Camelid nanobodies used as crystallization chaperones for different constructs of PorM, a component of the type IX secretion system from Porphyromonas gingivalis. *Acta Crystallogr F Struct Biol Commun*. May 1 2017;73(Pt 5):286-293. doi:10.1107/s2053230x17005969

133.   Fontana A, Polverino de Laureto P, De Filippis V, Scaramella E, Zambonin M. Probing the partly folded states of proteins by limited proteolysis. *Fold Des*. 1997;2(2):R17-26. doi:10.1016/s1359-0278(97)00010-2

134.   Dong A, Xu X, Edwards AM, et al. In situ proteolysis for protein crystallization and structure determination. *Nature Methods*. 2007/12/01 2007;4(12):1019-1021. doi:10.1038/nmeth1118

135. Ohi M, Li Y, Cheng Y, Walz T. Negative staining and image classification—powerful tools in modern electron microscopy. *Biological procedures online*. 2004;6:23-34.

136. Cheng Y, Grigorieff N, Penczek PA, Walz T. A primer to single-particle cryo-electron microscopy. *Cell*. Apr 23 2015;161(3):438-449. doi:10.1016/j.cell.2015.03.050

137. Scarff CA, Fuller MJG, Thompson RF, Iadanza MG. Variations on Negative Stain Electron Microscopy Methods: Tools for Tackling Challenging Systems. *J Vis Exp*. Feb 6 2018;(132)doi:10.3791/57199

138. Gonen S. Progress Towards CryoEM: Negative-Stain Procedures for Biological Samples. *Methods Mol Biol*. 2021;2215:115-123. doi:10.1007/978-1-0716-0966-8_5

139. Stura EA, Wilson IA. Applications of the streak seeding technique in protein crystallization. *Journal of Crystal Growth*. 1991/03/01/ 1991;110(1):270-282. doi:https://doi.org/10.1016/0022-0248(91)90896-D

140. Zhu L, Sage JT, Rigos AA, Morikis D, Champion PM. Conformational interconversion in protein crystals. *Journal of Molecular Biology*. 1992/03/05/ 1992;224(1):207-215. doi:https://doi.org/10.1016/0022-2836(92)90584-7

141. Sousa R. Use of glycerol, polyols and other protein structure stabilizing agents in protein crystallization. *Acta Crystallogr D Biol Crystallogr*. May 1 1995;51(Pt 3):271-7. doi:10.1107/s0907444994014009

142. Fowler NJ, Williamson MP. The accuracy of protein structures in solution determined by AlphaFold and NMR. *Structure*. 2022/07/07/ 2022;30(7):925-933.e2. doi:https://doi.org/10.1016/j.str.2022.04.005

143. Huang YJ, Zhang N, Bersch B, et al. Assessment of prediction methods for protein structures determined by NMR in CASP14: Impact of AlphaFold2. *Proteins: Structure, Function, and Bioinformatics*. 2021;89(12):1959-1976. doi:https://doi.org/10.1002/prot.26246

144. Lazar T, Martínez-Pérez E, Quaglia F, et al. PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. *Nucleic Acids Res*. Jan 8 2021;49(D1):D404-d411. doi:10.1093/nar/gkaa1021

145. Varadi M, Kosol S, Lebrun P, et al. pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucleic Acids Res*. Jan 2014;42(Database issue):D326-35. doi:10.1093/nar/gkt960

146. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. *Nature Reviews Molecular Cell Biology*. 2005/03/01 2005;6(3):197-208. doi:10.1038/nrm1589

147. Tunyasuvunakool K, Adler J, Wu Z, et al. Highly accurate protein structure prediction for the human proteome. *Nature*. 2021/08/01 2021;596(7873):590-596. doi:10.1038/s41586-021-03828-1

148. Perdigão N, Heinrich J, Stolte C, et al. Unexpected features of the dark proteome. *Proc Natl Acad Sci U S A*. Dec 29 2015;112(52):15898-903. doi:10.1073/pnas.1508380112

149. Porta-Pardo E, Ruiz-Serra V, Valentini S, Valencia A. The structural coverage of the human proteome before and after AlphaFold. *PLOS Computational Biology*. 2022;18(1):e1009818. doi:10.1371/journal.pcbi.1009818

150. Akdel M, Pires DE, Pardo EP, et al. A structural biology community assessment of AlphaFold2 applications. *Nature Structural & Molecular Biology*. 2022:1-12.

151. Bruley A, Mornon J-P, Duprat E, Callebaut I. Digging into the 3D Structure Predictions of AlphaFold2 with Low Confidence: Disorder and Beyond. *Biomolecules*. 2022;12(10):1467.

152. Dyson HJ, Wright PE. Coupling of folding and binding for unstructured proteins. *Current Opinion in Structural Biology*. 2002/02/01/ 2002;12(1):54-60. doi:https://doi.org/10.1016/S0959-440X(02)00289-0

153. Hausrath AC, Kingston RL. Conditionally disordered proteins: bringing the environment back into the fold. *Cell Mol Life Sci*. Sep 2017;74(17):3149-3162. doi:10.1007/s00018-017-2558-1

154. Alderson TR, Pritišanac I, Kolarić Đ, Moses AM, Forman-Kay JD. Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2. *bioRxiv*. 2022:2022.02. 18.481080.

155. Myrum C, Baumann A, Bustad HJ, et al. Arc is a flexible modular protein capable of reversible self-oligomerization. *Biochem J*. May 15 2015;468(1):145-58. doi:10.1042/bj20141446

156. Cumberworth A, Lamour G, Babu MM, Gsponer J. Promiscuity as a functional trait: intrinsically disordered regions as central players of interactomes. *Biochemical Journal*. 2013;454(3):361-369. doi:10.1042/bj20130545

157. Uversky VN. Intrinsic disorder here, there, and everywhere, and nowhere to escape from it. *Cell Mol Life Sci*. Sep 2017;74(17):3065-3067. doi:10.1007/s00018-017-2554-5

158. Xue B, Dunker AK, Uversky VN. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *J Biomol Struct Dyn*. 2012;30(2):137-49. doi:10.1080/07391102.2012.675145

159. Ekman D, Björklund AK, Elofsson A. Quantification of the elevated rate of domain rearrangements in metazoa. *J Mol Biol*. Oct 5 2007;372(5):1337-48. doi:10.1016/j.jmb.2007.06.022

160. Ekman D, Björklund AK, Frey-Skött J, Elofsson A. Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *J Mol Biol*. Apr 22 2005;348(1):231-43. doi:10.1016/j.jmb.2005.02.007

161. Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Biol*. Jan 2015;16(1):18-29. doi:10.1038/nrm3920

162.   Tompa P, Schad E, Tantos A, Kalmar L. Intrinsically disordered proteins: emerging interaction specialists. *Current Opinion in Structural Biology*. 2015/12/01/ 2015;35:49-59. doi:https://doi.org/10.1016/j.sbi.2015.08.009

163.   Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews Molecular Cell Biology*. 2015/01/01 2015;16(1):18-29. doi:10.1038/nrm3920

164.   Pejaver V, Hsu WL, Xin F, Dunker AK, Uversky VN, Radivojac P. The structural and functional signatures of proteins that undergo multiple events of post-translational modification. *Protein Sci*. Aug 2014;23(8):1077-93. doi:10.1002/pro.2494

165.   Iakoucheva LM, Radivojac P, Brown CJ, et al. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res*. 2004;32(3):1037-49. doi:10.1093/nar/gkh253

166.   Basile W, Salvatore M, Bassot C, Elofsson A. Why do eukaryotic proteins contain more intrinsically disordered regions? *PLoS Comput Biol*. Jul 2019;15(7):e1007186. doi:10.1371/journal.pcbi.1007186

167.   Ashley J, Cordy B, Lucia D, Fradkin LG, Budnik V, Thomson T. Retrovirus-like Gag Protein Arc1 Binds RNA and Traffics across Synaptic Boutons. *Cell*. Jan 11 2018;172(1-2):262-274 e11. doi:10.1016/j.cell.2017.12.022

168.   Uchański T, Masiulis S, Fischer B, et al. Megabodies expand the nanobody toolkit for protein structure determination by single-particle cryo-EM. *Nature Methods*. 2021/01/01 2021;18(1):60-68. doi:10.1038/s41592-020-01001-6

169.   Wu X, Rapoport TA. Cryo-EM structure determination of small proteins by nanobody-binding scaffolds (Legobodies). *Proceedings of the National Academy of Sciences*. 2021;118(41):e2115001118. doi:doi:10.1073/pnas.2115001118