Engineering Proteins by Domain Insertion

submitted to the
Department of Biology
of the
Technical University Darmstadt

for the degree of
Doctor rerum naturalium
(Dr. rer. nat.)

Dissertation by
M. Sc. Jan Mathony

First Referee: Professor Dr. Dominik Niopek
Second Referee: Professor Dr. Beatrix Süß

Darmstadt, 2023

Mathony, Jan: Engineering Proteins by Domain Insertion
Darmstadt, Technical University Darmstadt
Year of publication of the dissertation at TUprints: 2023
Date of the oral exam: 23.03.2023

Ehrenwörtliche Erklärung:

Ich erkläre hiermit, dass ich die vorliegende Arbeit entsprechend den Regeln guter wissenschaftlicher Praxis selbstständig und ohne unzulässige Hilfe Dritter angefertigt habe. Sämtliche aus fremden Quellen direkt oder indirekt übernommenen Gedanken sowie sämtliche von Anderen direkt oder indirekt übernommenen Daten, Techniken und Materialien sind als solche kenntlich gemacht. Die Arbeit wurde bisher bei keiner anderen Hochschule zu Prüfungszwecken eingereicht. Die eingereichte elektronische Version stimmt mit der schriftlichen Version überein.

Darmstadt, den 21.01.2023

Jan Mathony

Für Bastienne Zaremba
Für Tom, Inez und Axel Mathony

Acknowledgments

First and foremost, I would like to thank Professor Dr. Dominik Niopek! I am deeply grateful for his constant support and advice over the past years. Under his supervision, I had the unique chance to work independently and to explore diverse project directions, while I could always count on his full support at all stages of my project. His never-ending enthusiasm for science, his creativity and optimism during challenging project phases and the rare combination of perfectionism and pragmatism that he somehow balances with ease, make him an amazing supervisor and mentor.
I am very thankful to Professor Dr. Roland Eils for his continuous support. His trust and support over the years, first in the context of the iGEM competition and later, when I could start the Ph.D. journey in his department mean a lot to me.
I would also like to express my gratitude to Professor Dr. Beatrix Süß, Professor Dr. Robert Russell and Professor Dr. Gerhard Thiel for support during thesis advisory committee meetings and/or as referees and as members of my thesis examination board.

During my Ph.D., I had the privilege to work with a fantastic group of colleagues and student interns, who made this time an exciting and wonderful experience. I am very grateful to Sabine Aschenbrenner for the amazing support since my start in Dominik's group. As the "heart" of the lab she enabled a smooth and efficient transition of the laboratory from Heidelberg to Darmstadt. She also made important contributions to the AraC project and her enthusiasm for experiments of all kinds was always contagious. I would also like to thank Dr. Mareike Hoffmann for the efficient and fruitful work on the AcrIIC3 project.
Many thanks go out to our collaborators who made important contributions to the Acr projects. I had the pleasure to work together with the groups of Professor Dr. Bruno Correia, Professor Dr. Dirk Grimm, Professor Dr. Roland Eils and Professor Dr. Yanli Wang. In the context of this study, I am particularly thankful to the amazing teamwork with Dr. Carolin Schmelas, Dr. Zander Harteveld and Julius Upmeier zu Belzen.

Further, I would like to thank the German National Scholarship Foundation for funding me for the most part of this Ph.D. I am also very thankful towards Dr. Monika Langlotz at the ZMBH FACS core facility for her support during the project. I thank the workshop of the biology department at TU Darmstadt for the construction of custom illumination setups and the kind cooperation. Further I am grateful to the EMBL GeneCore facility for their fast and efficient NGS service. I thank Bastienne Zaremba and Benedict Wolf for proofreading the manuscript.

Finally, a very special thanks go to my friends and family. I would like to thank my parents Inez and Axel Mathony for their unconditional love and support, during all phases and situations in life. I am grateful to my brother Tom Mathony. With his kindness, honesty and determination, he is the best possible friend. I sincerely thank Bastienne Zaremba, for her constant support, believing in me and reminding me of what really matters.

Contributions

If not stated differently, all presented experiments were designed, executed and analyzed by myself under supervision of **Professor Dr. Dominik Niopek**.
Over the course of the presented work, I had the pleasure to collaborate with many exceptional scientists, who contributed to the success of all projects and I am very grateful to their phenomenal work and support. Throughout this thesis, I use the term "we" to describe experiments and project aspects to which other scientists contributed. I use the first person singular only if the data was obtained and analyzed by me alone. In addition, the contributions of colleagues are explicitly mentioned in the respective sections. Here, I will briefly summarize their contributions to this thesis:
The AcrIIC3 and AcrIIC1 projects were initiated at the time of my Master's thesis. Consequently, the data presented under "previous work" (section 1.4.2.4) was mostly obtained by me, prior to the start of my Ph.D. The dataset 1.8C, presented in section 1.4.2.4 was generated by **Dr. Carolin Schmelas** and **Sabine Aschenbrenner**. In addition, several DNA constructs used throughout these projects were cloned during that time (See appendix, supplementary table 1 for details). The AAVs used for transduction in the experiments resulting in Figure 2.3C-D were produced and purified by **Dr. Carolin Schmelas**. Finally, the structural models that were the basis of supplementary figures 2 and 3 were designed by **Julius Upmeier zu Belzen** and **Dr. Zander Harteveld**.
With respect to the domain insertion screening project, the cell sorting experiments were performed at the ZMBH FACS facility together with **Dr. Monika Langlotz**. Figure 2.19A in section 2.3.4 presents a comparison between my data and measurements published by **Weinberg et al.**, 2018. Finally, the characterization of the AraC hybrids, particularly the two datasets corresponding to Figure 2.26A and C, as well as the data for supplementary figure 6 were obtained jointly together with **Sabine Aschenbrenner**.

Publications

Mathony J*, Harteveld Z*, Schmelas C*, Upmeier zu Belzen J, Aschenbrenner S, Sun W, Hoffmann MD, Stengl C, Scheck A, Georgeon S, Rosset S, Wang Y, Grimm D, Eils R, Correia BE, Niopek D.
Computational design of anti-CRISPR proteins with improved inhibition potency.
Nature Chemical Biology. (2020). 16 (7), 725-730, DOI: 10.1038/s41589-020-0518-9.

Hoffmann MD*, Mathony J*, Upmeier zu Belzen J, Harteveld Z, Aschenbrenner S, Stengl C, Grimm D, Correia BE, Eils R, Niopek D.
Optogenetic control of Neisseria meningitidis Cas9 genome editing using an engineered, light-switchable anti-CRISPR protein.
Nucleic Acids Research. (2021). 49 (5), e29-e29, DOI: 10.1093/nar/gkaa1198.

Mathony J, Niopek D.
Enlightening Allostery: Designing Switchable Proteins by Photoreceptor Fusion.
Advanced Biology. (2021). 5 (5), 2000181, DOI: 10.1002/adbi.202000181.

Mathony J, Hoffmann MD, Niopek N.
Optogenetics and CRISPR: A New Relationship Built to Last.
Methods in Molecular Biology. (2020). 261-281, DOI: 10.1007/978-1-0716-0755-8_18.

Stadelmann T*, Heid D*, Jendrusch M*, Mathony J, Rosset S, Correia BE, Niopek D.
A deep mutational scanning platform to characterize the fitness landscape of anti-CRISPR proteins.
bioRxiv. (2021). DOI: 10.1101/2021.08.21.457204.

Adam L*, Stanifer M*, Springer F, Mathony J, Di Ponzio C, Eils R, Boulant S, Niopek D, Kallenberger SM.
Dynamics of SARS-CoV-2 host cell interactions inferred from transcriptome analyses.
bioRxiv. (2021). DOI: 10.1101/2021.07.04.450986.

The following publication relates to work performed prior to my PhD:

Upmeier zu Belzen J, Bürgel T, Holderbach S, Bubeck F, Adam L, Gandor C, Klein M, Mathony J, Pfuderer P, Platz L, Przybilla M, Schwendemann M, Heid D, Hoffmann MD, Jendrusch M, Schmelas C, Waldhauer M, Lehmann I, Niopek D, Eils R.
Leveraging implicit knowledge in neural networks for functional dissection and engineering of proteins.
Nature Machine Intelligence. (2019). (5), 225-235, DOI: 10.1038/s41589-020-0518-9.

*: Equal contribution

Poster and oral presentations (as presenting author)

"Computational design of anti-CRISPR proteins with improved inhibition potency" at the **Keystone Conference: Engineering the Genome**. Feb. 2020, Banff, Canada, (Talk and poster).

"Engineered anti-CRISPR proteins for precise control of gene editing" at the workshop **"CRISPR: Design, Strategy & Analysis" by CamBioScience**. Sep. 2020. Online, (Talk).

"Rational engineering of anti-CRISPR proteins for genome editing applications". At the **German Conference on Synthetic Biology**. Sep. 2021, online, (Talk and poster).

Abstract

Protein domains are structural and functional subunits of proteins. The recombination of existing domains is a source of evolutionary innovation, as it can result in new protein features and functions. Inspired by nature, protein engineering commonly uses domain recombination in order to create artificial proteins with tailor-made properties. Customized control over protein activity, for instance, can be achieved by harnessing switchable domains and functionally linking them to effector domains. Many natural protein domains exhibit conformational changes in response to exogenous triggers. The insertion of light-switchable receptor domains into an effector protein of choice, for instance, allows the control of effector activity with light. The resulting optogenetic proteins represent powerful tools for the investigation of dynamic cellular processes with high precision in time and space. On top, optogenetic proteins enable manifold biotechnological applications and they are even considered potential candidates for future therapeutics.

In this study, we first focused on CRISPR-Cas9 genome editing and applied a domain insertion strategy to genetically encoded inhibitors of the CRISPR nuclease from *Neisseria meningitidis* (*Nme*Cas9), which due to its small size and high DNA sequence-specificity is of great interest for CRISPR genome editing applications. Fusing stabilizing domains to the *Nme*Cas9 inhibitory protein AcrIIC1 allowed us to boost its inhibitory effect, thereby yielding a potent gene editing off-switch. Furthermore, the insertion of the light-responsive LOV2 domain from *Avena sativa* into AcrIIC3, the most potent inhibitor of *Nme*Cas9, enabled the optogenetic control of gene editing via light-dependent *Nme*Cas9 inhibition. Further investigation of the engineered inhibitors revealed the potential these proteins could have with respect to safe-guarding of the CRISPR technology by selectively reducing off-target editing.
The laborious optimization of the engineered CRISPR inhibitors necessary by the time motivated us to more systematically investigate possibilities and constraints of protein engineering by domain insertion using an unbiased insertion approach. Previously, single protein domains were usually introduced only at a few rationally selected sites into target proteins. Here, we inserted up to five structurally and functionally unrelated domains into several different candidate effector proteins at all possible positions. The resulting libraries of protein hybrids were screened for activity by fluorescence-activated cell sorting (FACS) and subsequent next-generation sequencing (Flow-seq). Training machine learning models on the resulting, comprehensive datasets allowed us to dissect parameters that affect domain insertion tolerance and revealed that sequence conservation statistics are the most powerful predictors for domain insertion success. Finally, extending our experimental Flow-seq pipeline towards the screening of engineered, switchable effector variants yielded two potent optogenetic derivatives of the *E. coli* transcription factor AraC. These novel hybrids will enable the co-regulation of bacterial gene expression by light and chemicals.
Taken together, our study showcases the design of functionally diverse protein switches for the control of gene editing and gene expression in mammalian cells and *E. coli*, respectively. In addition, the generation of a large domain insertion datasets enabled - for the first time - the unbiased investigation of domain insertion tolerance in several evolutionary unrelated proteins. Our study showcases the manifold opportunities and remaining challenges behind the engineering of proteins with new properties and functionalities by domain recombination.

Zusammenfassung

Domänen sind die strukturellen und funktionalen Untereinheiten von Proteinen. Die Rekombination bestehender Domänen dient als Quelle evolutionärer Innovationen, die neue Proteinfunktionen und -eigenschaften ermöglichen kann. Inspiriert von der Natur nutzt das Protein Engineering häufig die Domänenrekombination, um künstliche Proteine mit maßgeschneiderten Eigenschaften herzustellen. Die Kontrolle über Proteinaktivität kann beispielsweise erreicht werden, indem schaltbare Domänen genutzt und funktionell mit Effektordomänen verknüpft werden. Viele natürliche Proteindomänen zeigen Konformationsänderungen als Reaktion auf exogene Auslöser. Die Insertion licht-schaltbarer Rezeptordomänen in ein beliebiges Effektorprotein ermöglicht beispielsweise die Kontrolle der Effektoraktivität mit Licht. Die resultierenden optogenetischen Proteine sind leistungsstarke Werkzeuge zur Untersuchung dynamischer zellulärer Prozesse mit hoher zeitlicher und räumlicher Präzision. Darüber hinaus ermöglichen optogenetische Proteine vielfältige biotechnologische Anwendungen und gelten sogar als potenzielle Kandidaten für zukünftige Therapeutika.

In dieser Studie konzentrierten wir uns zunächst auf Geneditierung mittels CRISPR-Cas9 und wendeten eine Domäneninsertionsstrategie auf genetisch codierte Inhibitoren der CRISPR-Cas9-Nuklease aus *Neisseria meningitidis* (*Nme*Cas9) an. Diese Nuklease ist aufgrund ihrer geringen Größe und hohen DNA-Sequenzspezifität von großem Interesse für Anwendungen der CRISPR-Genomeditierung. Durch die Fusion stabilisierender Domänen mit dem *Nme*Cas9 Inhibitorprotein AcrIIC1 konnten wir dessen suppressive Wirkung verstärken und so einen effektiven Ausschalter für die Geneditierung herstellen. Darüber hinaus ermöglichte die Insertion der lichtempfindlichen LOV2-Domäne von *Avena sativa* in AcrIIC3, den wirksamsten *Nme*Cas9 Inhibitor, die optogenetische Kontrolle der Geneditierung durch lichtgesteuerte *Nme*Cas9-Hemmung. Weitere Untersuchungen der konstruierten Inhibitoren unterstrichen das Potenzial, das diese Proteine in Bezug auf die Absicherung der CRISPR-Technologie haben könnten, indem sie die Off-Target-Veränderung von DNA selektiv reduzieren.
Die mühsame Optimierung der konstruierten CRISPR-Inhibitoren, die zur damaligen Zeit notwendig war, motivierte uns, Potenzial und Limitationen des Protein-Engineerings durch Domäneninsertion unter Verwendung eines randomisierten Insertionsansatzes systematischer zu untersuchen. Früher wurden einzelne Proteindomänen normalerweise nur an wenigen bewusst ausgewählten Stellen in Zielproteine eingeführt. In dieser Studie dagegen, fügten wir bis zu fünf strukturell und funktionell nicht verwandte Domänen an allen möglichen Positionen in verschiedene Kandidatenproteine ein. Die resultierenden Bibliotheken von Proteinhybriden wurden durch fluoreszenzaktivierte Zellsortierung (FACS) und anschließende Next-Generation-Sequenzierung (Flow-seq) auf Aktivität gescreent. Das Trainieren von Machine Learning Modellen auf Grundlage der den resultierenden, umfassenden Datensätzen ermöglichte es uns, Parameter zu analysieren, die sich auf die Insertionstoleranz der Proteine auswirken. Es zeigte sich, dass Parameter der Sequenzkonservierung die besten Prädiktoren für den Erfolg von Domain-Insertionen sind. Schließlich führte die Erweiterung unserer experimentellen Flow-seq-Pipeline für das Screening von engineerten, schaltbaren Effektorvarianten zu zwei potenten optogenetischen Versionen des *E. coli*-Transkriptionsfaktors AraC. Diese neuartigen Hybride werden die Co-Regulierung der bakteriellen Genexpression durch Licht und Chemikalien ermöglichen.

Zusammenfassend zeigt unsere Studie das Design von funktionell unterschiedlichen Proteinschaltern für die Kontrolle der Geneditierung und Genexpression in Säugetierzellen bzw. *E. coli*. Darüber hinaus ermöglichte die Erstellung eines großen Domäneninsertionsdatensatzes erstmals die unvoreingenommene Untersuchung der Insertionstoleranz in mehreren evolutionär nicht verwandten Proteinen. Unsere Studie zeigt die vielfältigen Chancen und verbleibenden Herausforderungen hinter dem Engineering von Proteinen mit neuen Eigenschaften und Funktionalitäten durch Domänenrekombination.

Table of Contents

1 Introduction

1.1 Proteins – from sequence to function

Proteins represent one of the major macromolecule classes in living systems and are arguably the most versatile in nature. The variety of protein functions and properties is enormous. They serve as structure giving elements in the cytoskeleton, as enzymes in the metabolism, as regulators and actuators of practically all cellular processes ranging from cell cycle to cell differentiation, as signaling molecules and many more. In fact, proteins are involved in all key processes that make up the core foundations of life as we know it.

Proteins exhibit a modular architecture. They are composed of linear chains of covalently linked amino acids. 20 different canonical amino acids serve as their building blocks. Every protein is defined by a unique sequence of amino acids. The reason for the sheer endless diversity of proteins lies in the three-dimensional structure of these chains. All major properties, such as physical stability, catalytic activity or binding interfaces for the interaction with other molecules are the consequence of a protein's specific structure. In 1973, Christian Anfinsen postulated that each amino acid sequence possesses a single optimal fold defined by a minimum of free energy (Anfinsen, 1973). In other words: the three-dimensional structure of a protein is determined by its amino acid sequence (Fig. 1.1A). Over the last decades said dogma has been revised as it is well known today that proteins or parts of proteins can be disordered (Dyson & Wright, 2005). Moreover, even ordered structures often adopt various conformations (Motlagh *et al*, 2014). Nonetheless, the fact that proteins mostly exist in one or few three-dimensional conformations, which are determined by the primary amino acid sequence, still holds true. These profound assumptions are the very basis of many protein engineering strategies that I am going to describe later.

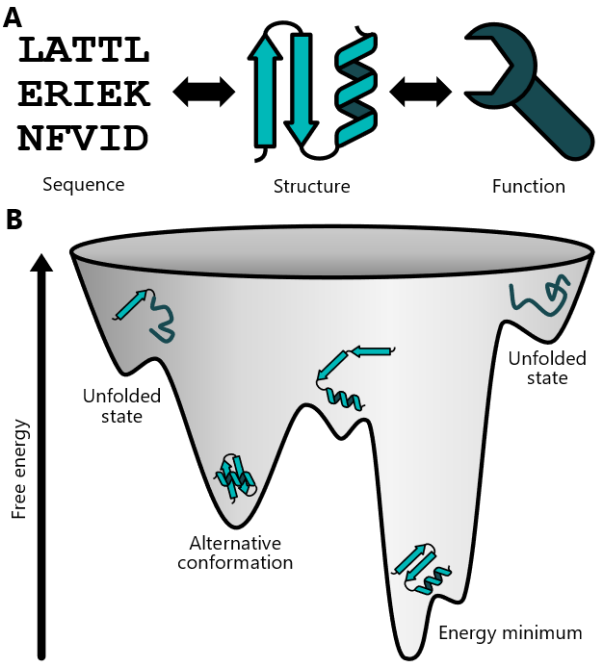


Figure 1.1: Proteins – from sequence over structure to function. (A) Sequence, structure and function of a protein are mutually dependent. (B) Schematic of the protein folding energy landscape.

The points discussed above result in an inter-dependence between the primary sequence of a protein, its structure and its function (Fig. 1.1A). It is thus no surprise, that the investigation of

protein function is always concerned with questions regarding sequence and structure. Basic research aiming to elucidate the role existing proteins play in nature and biotechnology, which envisions the design of new protein functions, are very similar in this regard.
The key challenge in protein science can be illustrated by two numbers games: Considering a small protein of 100 amino acids, in which the peptide bond connections between them had only two degrees of freedom, the number of possible protein conformations would be 2^{99}. Even if different three-dimensional folds could be sampled at a speed of 10^{13} conformations per second, the exploration of all possible conformations would take 100 million years. This problem, known as Levinthal's paradox (Levinthal, 1969), exemplifies that efficient protein folding must necessarily follow certain principles. Diverse theories on how proteins can adapt a stable structure within seconds *in vitro* have been discussed over the decades. These include the assumption that protein folding energy landscapes are funnel-shaped, i.e. they possess few relative energy minima corresponding to stable conformations with steep surroundings (Fig. 1.1B) (Ivankov & Finkelstein, 2020; Dill & MacCallum, 2012). From a practical perspective, the numbers impressively demonstrate the complexity of the folding problem and why the *in silico* prediction of protein structures is so challenging. Indeed, structure prediction algorithms only recently became powerful enough to be of practical use for the broader research community (AlQuraishi, 2019; Jumper *et al*, 2021) (see section 1.2.1.1.1).
The second example is not about the complexity of individual proteins, but considers protein sequence diversity as a whole. Assuming an average protein size of ~470 amino acids (AA) in eukaryotes (Tiessen *et al*, 2012) and 20 naturally occurring amino acid building blocks, the number of possible unique protein sequences, 20^{470}, is incredibly vast. This number gives an impression of the size of the protein sequence space, through which evolution navigates. At the same time, it leads to the main challenge behind every protein engineering approach, as it is impossible to systematically explore all sequences that could result in a desired structure and hence function. Thereby, the need for sophisticated strategies to generate proteins with new functionality becomes evident.
Taken together, the complexity of sequence-structure-function relationships in combination with the immense size of the protein sequence space are the reasons for our limited understanding of the protein world in general and the resulting challenges for protein engineering in particular.

1.1.1 Protein domains in natural proteins

The last section illustrated the enormous number of possible protein sequences. From a structural point of view, however, it is well-known that the number of unique folds found in natural proteins is surprisingly small (Ponting & Russell, 2002). Many proteins share similar structures or at least parts that are structurally related. In this context, the term "protein domain" becomes important (Baron *et al*, 1991). Protein domains describe families of similar amino acid sequences or structures that can be considered individual compact units. A protein can be composed of a single or several domains. The definition of such domains can be derived from sequence conservation, structural or functional similarity and even a combination thereof (Ponting & Russell, 2002). In the context of this thesis, I will look at domains mainly from a structural perspective, as protein structures are often more conserved than sequences (Glantz *et al*, 2016).

Although many single-domain proteins exist the majority of proteins consists of two or several domains (Apic *et al*, 2001). Interestingly, the same domains appear over and over again in nature (Ponting & Russell, 2002). Evolutionary processes can explain this phenomenon. It is for example well established that domain boundaries often correlate with the location of introns (Marsh & Teichmann, 2010). In the same line, the recombination or shuffling of domains was shown to be a driver of protein evolution that can give rise to proteins with new functionality (Peisajovich *et al*, 2010; Jin *et al*, 2009; Apic & Russell, 2010; Vogel *et al*, 2004). Contrary to the propagation from a common ancestor, convergent evolution towards specific folds was shown to contribute to the reoccurrence of domains in nature as well (Alva *et al*, 2010). Furthermore, not only protein domains are conserved, but also their combination, meaning that certain folds tend to appear in specific combinations. This feature has been referred to as "domain clubs" (Jin *et al*, 2009). On the contrary, very promiscuous domains, that are found in combination with a variety of other domains in natural proteins, are less common (Apic *et al*, 2001).

Despite these observations, the number of different domains that have been identified appears to be rather small, in relation the quantity known, as well as theoretically possible protein sequences. From an evolutionary perspective one could argue that the protein folding space has simply not been fully explored by nature. This theory is supported by the fact that proteins with new stable folds not found in nature can be artificially designed (Harbury *et al*, 1998; Kuhlman *et al*, 2003). The observation that many amino acid sequences are not able to fold into functional conformations and thus unlikely to exist in nature, adds another aspect (Finkelstein *et al*, 1993).

Taken together the abundance of natural proteins is, in fact, based on a limited pool of existing components. From a protein engineering perspective, the artificial recombination of natural protein domains represents an appealing strategy.

1.2 Protein engineering

Due to their broad application spectrum, engineered proteins play a pivotal role in the life sciences. Use case for engineered proteins range from the study of cellular processes in basic research, over biosensors for diverse molecule types in diagnostics towards medical applications such as chimeric antigen receptors in cancer therapy. The corresponding protein engineering goals and approaches are equally diverse. Proteins can be modified to tune their inherent properties, such as temperature stability or overall efficiency (Lovelock *et al*, 2022). Moreover, parameters, such as substrate specificity can be altered via protein engineering with the aim to create new protein functions (Kan *et al*, 2016; Lovelock *et al*, 2022). Fusion proteins link certain features of different proteins in time and space (Yu *et al*, 2015). Other approaches, in turn, aim at gaining control over the activity of a protein via exogenous stimuli (Mathony & Niopek, 2021; Herde *et al*, 2020). Even examples of entirely new proteins with defined functions designed from scratch have been reported (Hsia *et al*, 2016; Langan *et al*, 2019; Chen *et al*, 2020).

The following sections do not attempt to comprehensively cover this vast research area. Instead, I will give an overview of selected protein engineering methods, focusing on the main concepts and their strengths and weaknesses. I will further highlight aspects with particular relevance to this thesis.

1.2.1 Protein engineering approaches

In general, protein engineering is often based on heuristics informed by experimental results. More specifically, the functional and/or structural characterization of proteins provides information that can guide the engineering of proteins. Knowledge about active sites and other functionally important residues or structure elements give hints at which parts of a protein are suited for modifications and which ones should be left untouched. Engineering of the substrate specificity of enzymes, for instance, necessarily requires mutations at the substrate-binding pocket (Lovelock *et al*, 2022). If the stability of a protein should be optimized, in contrast, the active site will most likely not be the focus. Given the complexity of the protein engineering problem, prior experimental knowledge is often insufficient and trial and error approaches can be a very tedious procedure and may yield only suboptimal outcomes. Researchers have therefore developed a plethora of experimental and computational methods to more efficiently guide protein engineering efforts.

1.2.1.1 *Computational approaches*

Computationally-guided protein engineering aims at predicting promising protein variants *in silico,* which is then followed by testing and validation of lead candidates in the lab. To this end, mutations are computationally sampled and scored with respect to the desired protein property. The arguably most widely applied framework for this purpose is the Rosetta modeling suite (Rohl *et al*, 2004b). Originally developed by the group of David Baker, Rosetta has been improved and expanded over the years by many labs around the world (Das & Baker, 2008; Leman *et al*, 2020). Rosetta was initially developed for *in silico* protein structure prediction, but soon turned into a powerful tool for various kinds of protein engineering approaches. Its core feature is an energy function that scores putative protein conformations. This function captures a number of physical requirements of protein folding. Non-polar residues, for instance, must be buried within the protein core (Das & Baker, 2008). Van der Waals interactions and hydrogen bonds are further taken into account (Leman *et al*, 2020). Applying this function, Rosetta searches for conformations with local energy minima followed by optimization towards the absolute energy minimum (Das & Baker, 2008). Empirically informed strategies, in turn, are employed with respect to torsion angles and rotamers of amino acids. This procedure represents the basis for a plethora of protocols specific to different protein engineering goals, including the modeling of loops (Rohl *et al*, 2004a; Chivian & Baker, 2006), binding interface design (Barlow *et al*, 2018; Sevy *et al*, 2019; Shui *et al*, 2021) as well as the modeling of protein complexes (André *et al*, 2007; Courbet *et al*, 2022).

While Rosetta might well-represent the most comprehensive structure modeling environment, a large number of alternative stand-alone tools exist, including software packages for molecular dynamics (Brooks *et al*, 2009; Van Der Spoel *et al*, 2005) or tools for specific protein engineering tasks (Bienert *et al*, 2017; Tubert-Brohman *et al*, 2013; Trott & Olson, 2010). Moreover, the recent emergence of machine learning models trained on enormous amounts of protein data improves protein structure prediction and engineering at an astounding pace. The next section gives a brief overview of current structure prediction approaches relevant for this study.

1.2.1.1.1 Protein structure prediction with AlphaFold2 and related approaches

Protein structure prediction is often considered as one of the major unsolved questions in biology. The Critical Assessment of Techniques for Protein Structure Prediction (CASP), a community-driven recurring evaluation of the best structure prediction approaches has documented the persisting challenges in the research field over the last decades (Moult *et al*, 1995; Kryshtafovych *et al*, 2021). Recent advances in the area of machine learning (ML) gave rise to a rapid increase in prediction performance (Kryshtafovych *et al*, 2021). Importantly, ML models are not based on physics-inspired energy functions, but use neural networks, which are trained on structure and sequence information, obtained from publicly accessible protein repositories, such as UniProt (www.uniprot.org) (The UniProt Consortium, 2021) and the Protein Data Bank (PDB) (www.rcsb.org) (Berman *et al*, 2000). Various neural network architectures, such as graph neural networks (Ingraham *et al*, 2019), long short-term memory (LSTM) networks (AlQuraishi, 2019) and convolutional neural networks (Senior *et al*, 2020) have been explored and employed for structure prediction. Recently, a major leap in performance was achieved by DeepMind with the release of the AlphaFold2 (AF2) model (Jumper *et al*, 2021). Inspired by this work, the Baker lab constructed a similar network called RoseTTAfold with only slightly weaker performance, shortly thereafter (Baek *et al*, 2021). Both models build on the transformer architecture, which is based on a mechanism termed "self-attention", a concept well suited to represent long-range interactions within sequential data (Vaswani, 2017). This type of neural network was originally developed in the field of natural language processing (Vaswani, 2017; Devlin *et al*, 2019). Due to their general ability to efficiently learn sequence representations, transformers have since been widely adopted for the learning of biological information on the basis of DNA or Protein sequence data (Vig *et al*, 2020; Clauwaert & Waegeman, 2020).

AF2 and RoseTTAfold predict protein structures based on an input amino acid sequence. This input is used to construct multiple sequence alignments (MSAs) of homologues proteins, fetched from public databases. The MSA, in fact, is the key source of information used by AF2. In addition, structural templates are identified and harnessed by the network. As a consequence, the quality of the prediction depends to some degree on the availability of related sequences to generate meaningful MSAs, as well as structural templates. Of note, shortly after the release of AF2, alternative models with improved performance on single-sequence inputs (without MSA), reduced compute time and higher memory efficiency have been reported (Chowdhury *et al*, 2022; Ahdritz *et al*, 2022).

Given the large margin by which AF2 outperformed competing models in the CASP14 evaluation (Kryshtafovych *et al*, 2021), it became rapidly praised as one of the most transformative breakthroughs in biology in recent years (Callaway, 2022). As a note of caution, however, one must consider that the exceptional performance of AF2 is at least in parts limited to smaller more compact proteins, while the structures of large multi-domain proteins remain challenging to be precisely predicted (Jumper *et al*, 2021; Akdel *et al*, 2022). A vivid discussion has emerged with respect to the impact AF2 is going to have on the diverse research fields that rely on protein structure information (Diwan *et al*, 2021; Subramaniam & Kleywegt, 2022; Tong *et al*, 2021). It is highly debated, for instance, to which extend AF2 can be harnessed to explore the conformational diversity of proteins. It is further questionable, if AF2 is capable of

predicting the structural impact of small sequence changes that are induced by genetic variation or artificial mutagenesis (Diwan *et al*, 2021).
A number of recently published and ongoing investigations aim at answering these questions. Early studies came to the conclusion that AF2 would not be powerful enough to predict the structural effects of single point mutations (Diwan *et al*, 2021; Buel & Walters, 2022; Pak *et al*, 2021). The situation looks more promising though with respect to the modeling of conformational ensembles (del Alamo *et al*, 2022; Saldaño *et al*, 2022). Another aspect that gained increasing attention is the predicted local distance difference test (pLDDT) score, an AF2 internal measure of the position-wise quality of a structure model. This score was shown to be suited as an indicator for intrinsically disordered regions (Akdel *et al*, 2022; Wilson *et al*, 2022). Finally, first attempts to exploit AF2 for the design of new proteins have already been published (Jendrusch, 2021; Goverde *et al*, 2022). It remains to be seen, however, how widely these frameworks can be applied.
Taken together, the recent developments in protein structure prediction fuel new approaches within the protein engineering field. The exploration of AF2 and related models with respect to protein design has just begun.

1.2.1.2 *Directed evolution*

Computational modeling intends to predict promising mutations, so that only a limited number of candidates must be tested in the laboratory. Directed evolution follows a different, but complementary path. It aims at exploiting the principles of Darwinian evolution by selecting the best protein variants from a large pool of candidates. The two central steps of directed evolution are (i) the generation of a variant library, derived from a parent protein of choice and (ii) the subsequent selection of best performing variants. This procedure can be iteratively repeated. Various protocols describing different experimental procedures with diverse strengths and weaknesses have been developed over the decades. Overall, the concept has proven to be highly powerful and was recognized with a Nobel prize in 2018 awarded to the pioneers in the field, Frances Arnold, George P. Smith and Gregory Winter.
Traditional methods frequently used for the generation of sequence libraries are error-prone PCR (Chusacultanachai & Yuthavong, 2004) and DNA shuffling (Stemmer, 1994). An alternative that has become increasingly attractive due to more affordable DNA synthesis, is oligonucleotide-based saturation mutagenesis (Miyazaki & Arnold, 1999). It further allows the precise determination of regions or even individual codons that will be mutated. *In vivo* mutagenesis represents another well-established option, enabled by DNA propagation in mutator strains (Greener *et al*, 1997; Badran & Liu, 2015), although it comes at the cost of limited control over the mutated regions. The recent past has witnessed an increase in complexity with respect to the types of mutations that can be introduced into a coding DNA sequence. The deletion or insertion of several amino acids can nowadays easily be generated in bulk (Coyote-maestas *et al*, 2019; Guntas & Ostermeier, 2004; Macdonald *et al*, 2022). A global challenge with respect to directed evolution is the overall library complexity i.e., the number of different variants that can be created and screened. In order to effectively evolve or adapt a desired trait, introducing multiple mutations is often necessary, in particular when epistatic effects are considered (Bloom & Arnold, 2009; Voskarides, 2021). The size of comprehensive mutant libraries, however, increases exponentially when combinations of two or even more mutations should be covered. Although the cloning of libraries including millions

of candidates is nowadays feasible, the required complexity can still represent a crucial bottleneck for directed evolution experiments.

In this context, the screening method is of equal importance as the generation of complex libraries. While variants can certainly be individually characterized, this would drastically limit the throughput to a few dozens to several hundred candidates. In most cases, it is thus necessary to screen libraries in bulk, by establishing a robust genotype-phenotype linkage and then selecting lead candidates in high throughput. One of the first strategies to do so was phage display which enabled the evolution of protein interactions (McCafferty *et al*, 1990; Smith, 1985). Here, the fusion of a candidate library to a bacteriophage capsid protein is encoded within a phage genome. Recombinant phages that present the fusion protein on their surface and carry the corresponding coding sequence within their genome are exposed to a surface, coated with a binding partner for the protein of choice. Only proteins that can effectively bind to the partner stick to the surface, while non-functional variants can be washed off. Phages carrying the functional protein candidates are later eluted from the surface, so that the lead candidates can be recovered and sequenced.

Many alternative strategies use *in vivo* selection by expression of the library in a host organism, often *Escherichia coli* (*E. coli*). To this end, the activity of the candidates from the library must be coupled to the expression of a fluorescent reporter so that functional variants can be enriched via FACS (Zhao & Arnold, 1997). Alternatively, enrichments can be achieved by linking the phenotype of the candidate to cell survival. Important prerequisites for these procedures are monoclonality, i.e. the fact that every cell expresses only one variant, and a sufficient correlation between candidate protein activity and reporter expression.

A second, more recent development in the directed evolution field are continuous *in vivo* systems, comprehensively reviewed by Morrison et al. (Morrison *et al*, 2020). The idea behind continuous directed evolution is to combine the diversification via mutations and the selection of lead candidates within a single continuous process in living cells. Prominent examples are phage-assisted continuous evolution (PACE) (Esvelt *et al*, 2011) and OrthoRep (Ravikumar *et al*, 2018). Here, the propagation of phages (PACE) or yeast cells (OrthoRep) is coupled to the fitness of the protein variant they express. At the same time, the coding sequence of the protein to be evolved is continuously mutated inside the host itself, resulting in new variants with each generation of phages or cells. Such experiments can be performed in continuously growing cultures for days or even weeks, e.g. using chemostats (Esvelt *et al*, 2011). The big advantage of these methods is the enormous number of different variants that can be screened in parallel within a relatively small culture volume. The laborious generation of libraries, in contrast, is not required. In addition, due to the continuous character of the evolution, the currently best variants are permanently optimized, resulting in an automated iterative evolution process.

Finally, directed evolution has been interfaced with machine learning models in several instances over the past years (Alley *et al*, 2019; Romero *et al*, 2013; Yang *et al*, 2019; Bedbrook *et al*, 2019). The idea behind this concept is to use a limited number of experimentally screened candidates to train models that are then able to predict new, improved protein variants.

Coming back to the functional effects of mutations acquired during directed evolution experiments, it can be assumed that 30-50 % of random mutations impair protein function, while 50-70 % are neutral. Only a very small fraction, 0.01-0.5 %, are expected to be beneficial (Bloom & Arnold, 2009). The surprisingly large proportion of point mutations that do not result

in substantial effects on protein fitness, motivated the appreciation of genetic drift as an important contribution to evolution (Kimura, 1968). That means, depending on the sequence context, seemingly neutral mutations might become decisive for protein function due to epistatic coupling, for instance, if additional amino acid are changed later on (Davis *et al*, 2009). This observation underscores the need for large and complex libraries, as the effect of combinations of mutations can hardly be predicted from the individual amino acid changes. Practically, the situation is further complicated by the fact that experimental setups for directed evolution often select for only a certain feature. Assuming the optimization of enzyme activity as the goal of a directed evolution experiment, the performance of the candidate variants is usually measured by the amount of substrate they convert per time period. Unfortunately, such assays will most likely ignore deleterious effects on other protein properties, such as stability, as long as the defect is not substantial enough to be reflected by the respective protein activity. How exactly the factors described above for point mutations translate to other types of mutations, such as insertions or deletions, has never been systematically investigated. Recent studies probed the systematic insertion or deletion of few amino acids in membrane channels and Cas9. The results gave hints that proteins often tolerate short insertions as well as deletions at diverse sites (Shams *et al*, 2021; Macdonald *et al*, 2022). The datasets are, however, too small for a general interpretation regarding the functional effects insertions can evoke in directed evolution experiments.

1.3 Protein switches

1.3.1 Controlling protein activity

In the last section, I introduced protein engineering methods without going into the details of a certain application field. In the following, I will focus on the engineering on switchable proteins in particular. In the context of this study, I use the term "switchable protein" to describe proteins, the activity of which can be turned on/off or be modulated by an exogenous trigger. Exogenous triggers refer to signals such as light or small molecules. On the contrary, the many endogenous processes that involve changes in protein activity are reviewed elsewhere (Lee & Yaffe, 2016; Gebauer & Hentze, 2004) and not further considered here.

The application spectrum of switchable proteins is diverse. They can be employed, for instance, as biosensors the context diagnostics or to detect contaminants (Jayanthi *et al*, 2017; Mansouri *et al*, 2019). Protein switches are also frequently used in basic research to control and dissect the function of molecular pathways (Toettcher *et al*, 2013; Wilson *et al*, 2017). Similarly, their application in biotechnology enables the user-defined regulation of metabolic processes (Morgan *et al*, 2016; McCarty & Ledesma-Amaro, 2018). In a medical context, protein switches can in principal be harnessed to restrict the activity of protein drugs in time or space (Mathony *et al*, 2020b; Ye & Fussenegger, 2019). Mechanistically, the activity of proteins can be controlled opto- or chemogenetically in many ways. Here, I will mainly focus on optogenetic examples, since this category is most relevant in the context of my Ph.D. work.

The technically easiest way to regulate the activity of a protein species within a cell is to control its expression. A plethora of tools exist, to induce the expression of a protein of interest in a controlled manner (de Mena *et al*, 2018). The particular advantage of this strategy is the plug-and-play fashion by which the controlled DNA sequence encoding a protein of choice can be

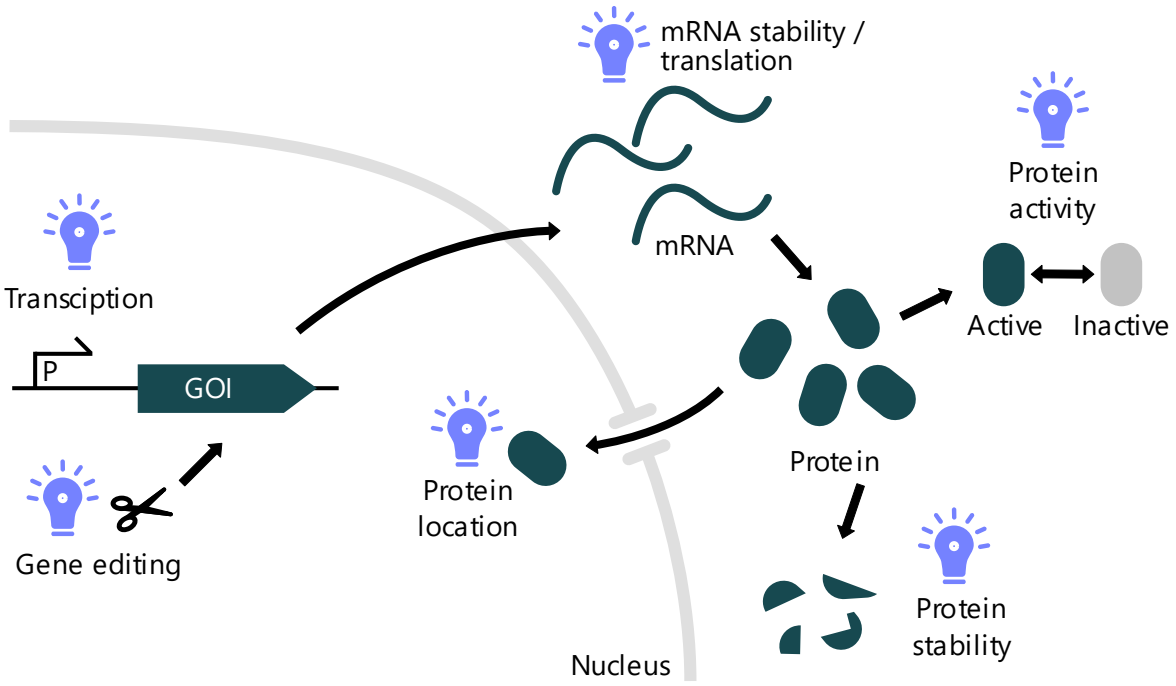


Figure 1.2: Possibilities to control genes and proteins optogenetically. Schematic of the various ways how genes and proteins can be optogenetically controlled. These include the light-dependent actuation of gene editing, transcription activation and repression, as well as the optogenetic regulation of protein activity, localization and stability. mRNA stability and translation, in turn are more often regulated via photo-cleavable RNA-modifications (Klöcker *et al*, 2022; Zhang *et al*, 2020). GOI, gene of interest.

exchanged. In section 1.4.1, I am going to describe the design of optogenetic transcription factors for bacteria in greater detail. Similar to its expression, the stability of a protein also affects its abundance and hence the resulting activity. To this end, optogenetically controlled degradation tags have been developed, which can be added as terminal fusions to a protein of choice (Bonger *et al*, 2014; Renicke *et al*, 2013). Collectively, these tools have in common that they regulate protein activity indirectly via control of its cellular abundance (Fig. 1.2). While this strategy might be well-suited for many use cases, it has also several drawbacks. The key issue is that only the abundance of a permanently active protein is regulated. The speed at which protein levels within a cell change are, however, determined by several factors, including transcription and translation rates, as well as protein stability. It is hardly possible to optogenetically control all these factors at once, resulting in limitations with respect to the speed at which the levels of a specific protein can be changed. A related technique enables the programmed import and export into or from the nucleus (Niopek *et al*, 2016, 2014). These and related methods were shown to respond rapidly to the light stimulus (Wang *et al*, 2016), rendering them an efficient option, if re-localization of the active protein is the desired mode of control.

The control over the actual activity of a pre-existing protein would, in contrast to all aforementioned strategies, enable immediate response to a trigger independent of expression levels or localization. As illustrated by the above points, this direct control of protein activity can be advantageous. It can be achieved by engineering of the target proteins themselves. The following sections describe methods for the direct optogenetic regulation of proteins.

1.3.2 Split proteins versus single-chain switches

The strategies to design switchable proteins can be roughly divided into two categories: split-proteins and single-chain approaches (Ostermeier, 2009). As the name already implies, split-protein switches are created by splitting the protein of choice into parts that are inactive on their own. Enabling the inducible reconstitution of the two parts allows to control the protein's activity (Shekhawat & Ghosh, 2011). Said reconstitution is usually mediated by fusion to additional protein domains that dimerize upon induction by light or a chemical trigger. Single-chain switches, in contrast, are engineered by the fusion of a protein of choice to a domain that affects its activity via structural rearrangements in an inducible fashion (Fig. 1.3) (Stratton & Loh, 2011).

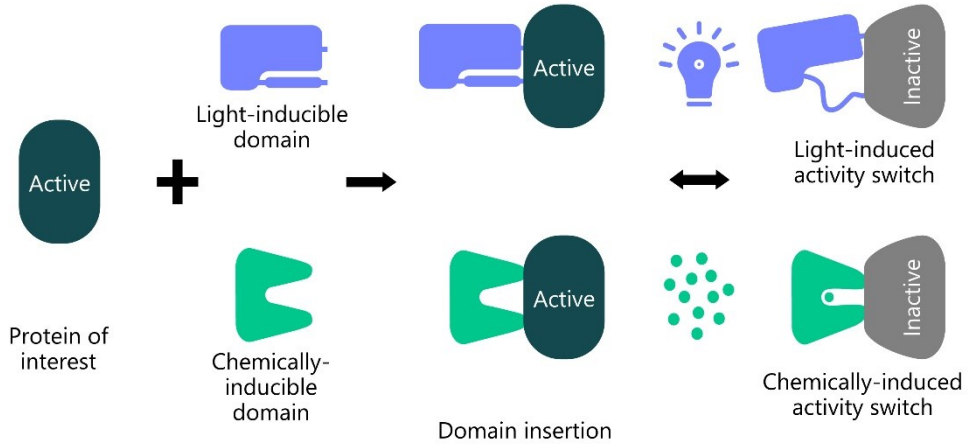


Figure 1.3: Controlling protein activity by domain insertion. Proteins can be controlled optogenetically or chemogenetically upon domain insertion. Light- or chemically inducible domains are inserted into a protein of choice in order to affect its conformation and activity upon induction. If the protein of choice is activated or deactivated by induction depends on specific structural adaptions of the insert domain in response to the trigger.

An efficient split-protein switch has to satisfy several requirements. Most importantly, suitable split-sites must be identified, meaning the division into parts that do not automatically reassemble into the original fold, when expressed individually. High rates of auto-reconstitution would lead to undesired background activity in the off-state. At the same time, the reassembly of a functional unit must still be possible when the two parts are brought into proximity by the induced dimerization. Moreover, a plethora dimerization domains exist, including dimerization mechanisms that are controlled by small molecules or light and can be employed to control the reassembly of the split-protein (Dagliyan *et al*, 2018; Guntas *et al*, 2015; Strickland *et al*, 2012; Jo *et al*, 2019; Rihtar *et al*, 2022; Shekhawat & Ghosh, 2011). Furthermore, examples of domain pairs enabling inducible dissociation have been reported (Wang *et al*, 2016). Generally, homo- and hetero-dimerizing domains are known. With respect to split proteins, hetero-dimers consisting of two different domains that bind to each other are often preferred (Shekhawat & Ghosh, 2011) as the dimerization of two copies of the same protein part can be prevented that way. Homo-dimerization systems are more frequently used if identical subunits are to be assembled (Romano *et al*, 2021; Wang *et al*, 2012).

In the context of this study, focus lies on single-chain protein switches. The core principle of these systems is a trigger-dependent conformational change, which brings the protein from an inactive conformation into an active state and *vice versa* (Fig. 1,3). This can be achieved with

help of protein domains that change their conformation in response to an input signal. Before the details of these switches are reviewed in the next section, it is important to consider the similarities and differences between single-chain and split-protein methods.
Split-protein systems are by definition more complex in several regards. Both protein parts have to be expressed individually, meaning that either two expression cassettes are necessary or both components have to be transcribed from a single promoter. The latter architecture requires the separation of the two parts via internal ribosomal entry sites (Pelletier & Sonenberg, 1988; Jang *et al*, 1988), 2A peptides (Luke *et al*, 2008) or bicistronic organization in bacteria. Here, it is important to ensure that the translation of a single full-length protein, e.g. via translation read-through, is prohibited. Furthermore, the dimerization of the split-parts partially depends on diffusion of both components within the cell, which could result in an activation delay, depending on the expression levels. Also, a complete activation of the expressed proteins cannot be reached, as it is unlikely that all parts present in a cell dimerize. These factors are irrelevant in the case of single-chain switches.
Considering the dynamic range of both approaches, the key difference lies in their mechanism of action. The activity of split proteins depends on the affinity of the dimerization domains. In addition, it relies on the reconstitution efficiency of the split parts. These factors define how "leaky" the system is, meaning to what degree the dimerization domains are able associate even in absence of the trigger and if the split-halves could re-assemble to some degree independent of the dimerization domains. The performance of single-chain protein switches, however, relies on biophysical restrictions and the forces driving the underlying conformational change. Here, the energy differences of the two conformational states (+/- stimulus) of the switchable domain, as well as the nature of this conformational change, are critical factors. Conformational changes usually underly equilibrium changes of structural ensembles (Yao *et al*, 2008; Motlagh *et al*, 2014). Consequently, not all molecules will exhibit the preferred conformation, resulting in some degree of natural leakiness in the system (Yao *et al*, 2008; Motlagh *et al*, 2014). In addition, the conformational regulation of protein activity can also be a question of the force a switchable domain can impose on the effector protein it is fused to (Dagliyan *et al*, 2016). In summary, the two strategies have different strengths and weaknesses that determine the performance of the resulting protein switches.

1.3.3 Engineering single-chain protein switches by domain insertion

The fact that the artificial insertion of additional amino acids into a natural protein is often possible without a loss of function is already known since the late 1980s (Starzyk *et al*, 1989; Freimuth *et al*, 1990; Ladant *et al*, 1992). Although the potential of insertional mutagenesis to alter protein activity was recognized early on (Shortle & Sondek, 1995), the effects reported in these studies mostly related to permanent changes in protein stability (Ladant *et al*, 1992). In 1997, the first successful insertion of larger random peptides (120-130 AA) into RNase H was described. In the same year the green fluorescent protein (GFP) was the first folded domain inserted into another protein with the aim to engineer a biosensor (Siegel & Isacoff, 1997). Siegel et al. inserted GFP into a potassium channel, achieving a voltage-dependent change in fluorescence of about ~5 %. Shortly thereafter, in 1998, a first domain insertion screen was performed, by random introduction of GFP into a cAMP-dependent protein kinase (Biondi *et al*, 1998). Several insertions with preserved kinase activity were thereby identified. Taking the

opposite route, Geoffrey et al. inserted ion-binding domains into GFP, which enabled the authors to create metal ion-specific biosensors (Baird *et al*, 1999). The same approach was used by Doi et al., who inserted a TEM1 β-lactamase into GFP (Doi & Yanagawa, 1999). To achieve measurable changes of fluorescence upon ligand binding, the scientists randomly mutated the fusion protein, followed by selection of variants with the desired properties. At this time, it was already assumed, that the structural changes of the ligand-binding domain (LBD) are allosterically linked to the GFP conformation, thus causing the observed change in fluorescence (Doi & Yanagawa, 1999). Functional examples of this kind were also taken as evidence that the proper folding of proteins does not require sequentially continuous domains (Collinet *et al*, 2000). Another study showed that different combinations of the same two domains can result in switches with different directionality, meaning an activation of fluorescence either in presence or absence of the ligand (Nagai *et al*, 2001). The dynamic range of this early generation of biosensors was often very limited. Nakai et al. were among the first to address this problem (Nakai *et al*, 2001). They created Ca^{2+}- sensors with improved signal-to-noise ratio by using enhanced GFP as effector domain, which exhibits brighter light emission as compared to related fluorescent proteins (Yang *et al*, 1996).

Over the years a plethora of allosteric protein switches have been developed (Skretas & Wood, 2005; Teasley Hamorsky *et al*, 2008; Ghanbarpour *et al*, 2019; Feil *et al*, 1997). The arguably most extensively studied and used class are biosensors based on luciferases (Fan *et al*, 2008; Taneoka *et al*, 2009) or fluorescent proteins (Siegel & Isacoff, 1997; Tallini *et al*, 2006; Akerboom *et al*, 2012). However, the application of protein switches is not limited to fluorescent and luminescent sensors. The group of Klaus Hahn has demonstrated the control of cell motility via optogenetically and chemogenetically triggered kinases (Dagliyan *et al*, 2016; Karginov *et al*, 2010; Chu *et al*, 2014; Dagliyan *et al*, 2013). Among others, they further showed that it is not only possible to insert single domains into proteins, but also tandems of linker-connected dimerization domains. This concept is exemplified by the rapamycin-dependent FKBP/FRB (FK506-binding protein; FKBP12-rapamycin binding protein) pair (Dagliyan *et al*, 2016) and the blue light-dependent homo-dimerizing vivid (VVD) domain (Shaaya *et al*, 2020) which were employed in potent single-chain protein switches. To this end, the domain pairs were connected via an amino acid linker and then inserted at the site of choice. The association/dissociation reactions of these domain pairs could then mediate the activity switch. Instead of the domain insertion at a single site, another architecture based on dimerization was used by Dueber et al. (Dueber *et al*, 2003). A PDZ domain and its cognate peptide-ligand were fused to either terminus of the neuronal Wiskott-Aldrich syndrome protein (N-WASP), respectively. The dimerization of these terminally linked domains sterically blocked the N-WASP induced increase of actin polymerization. In case a second separate ligand with higher affinity to the PDZ domain was provided, the interaction between N- and C-terminus of the fusion protein was revoked, resulting in the activation of N-WASP (Dueber *et al*, 2003). Similarly, the dimerizing fluorescent proteins pdDronpa were used to control the activity of kinases (Zhou *et al*, 2017a) or Cas9 nucleases (Zhou *et al*, 2017b).

In some cases, the simple terminal fusion of a switchable domain that sterically blocks important active sites can already be sufficient. The conformational change of the fused domain is then supposed to release the steric inhibition. This approach was nicely exemplified by Wu et al., who created a blue-light controlled Rac1 GTPase via terminal fusion to a light-

oxygen-voltage 2 (LOV2) domain from of *Avena sativa* phototropin-1 (*As*LOV2) (Wu *et al*, 2009). The *As*LOV2 domain was further employed as a lever arm, to control myosin action (Nakamura *et al*, 2014). To this end, the photoreceptor domain was fused as a molecular joint between myosin and α-actinin, so that it could control the angle between those proteins.

1.3.3.1 *Mechanistic investigation of switchable proteins*

Although allosteric protein switches are often engineered on the basis of structural considerations, the resulting fusion proteins have been rarely investigated, with respect to their actual structural adaptions. An exception is the calcium sensor GCaMP2, which consists of a calmodulin-M13 peptide insertion into eGFP. Crystal structures of the fusion proteins revealed that the binding of calcium, followed by conformational adaption of the insert, results structural changes at the interface between the domains (Akerboom *et al*, 2009). These effects stabilize the fluorophore's excited state and limit solvent access to the GFP core (Akerboom *et al*, 2009; Wang *et al*, 2008). In a different study, NMR-based investigation of the hybrid between a maltose binding domain and the TEM β-lactamase confirmed that the structures of the individual domains remain largely unaffected by the fusion. Still, the expected maltose-dependent conformational changes of key residues could also be observed (Wright *et al*, 2010). Beyond these exceptions, the structural consequences of domain insertion engineering attempts have barely been experimentally investigated.

1.3.4 Rational design of protein switches

Many of the examples discussed above were engineered based prior knowledge about the candidate proteins and domains. However, most of them were either constructed by trial and error (Biondi *et al*, 1998) or under consideration of aspects specific to certain proteins (Nakamura *et al*, 2014). Although the examination and utilization of specific structural features, such as the reversible formation of certain disulfide-bonds (Choi & Ostermeier, 2015) or the alternate folding of specific secondary structure elements (Yousef *et al*, 2004) can be successful, such approaches are hardly generalizable. This absence of broadly applicable strategies might also explain, why the same small set of effector proteins, such as GFP, β-lactamase or specific kinases have been repeatedly used in most previous studies.

Nonetheless, to overcome the limitation to specific use cases, diverse strategies for the rational design of switchable proteins have been proposed (Fig. 1.4).

1.3.4.1 *Mutually exclusive folding*

The group of Stewart Loh suggested a concept of mutual exclusive folding. To showcase their strategy, they inserted ubiquitin into a barnase (Radley *et al*, 2003). The termini of the ubiquitin domain are typically separated by 38.5 Å. Its insertion between P64 and T70 of the barnase was hence expected to result in drastic structural tension. Consequently, it must be impossible for both fusion partners to coexist in their native conformations, resulting in a structural "tug-of-war". The design was indeed successful, as the activity of ubiquitin could be switched by temperature changes or by the addition of a barnase binding partner (Radley *et al*, 2003). Similarly, a different combination between a barnase and a DNA-binding peptide (GCN4) resulted in switchable barnase activity, dependent on the presence of a cognate DNA motif (Ha *et al*, 2006). Regulation of the fusion protein's DNA-binding capabilities were enabled by changes of physical properties, such as temperature or buffer conditions.

The idea of mutual exclusive folding has since been adapted in several studies. Ha et al. for instance, inserted ubiquitin into a ribose binding protein (RBP) (Ha *et al*, 2015). The resulting fusion protein was inactive on its own, but could dimerize so that one RBP fragment could reconstitute a functional protein in trans, together with the complementary half from the identical partner molecule. Two other concepts showed that the competition for a structure element that is shared between two domains can also be used as a strategy for molecular switching. Sallee et al. designed proteins that react to the presence of the peptide Cdc42 (Sallee *et al*, 2007) and Strickland et al. merged the terminal helices of a LOV2 domain and the bacterial repressor TrpR, resulting in light-dependent DNA binding (Strickland *et al*, 2008). In both cases, the fusion proteins could only be active, when the shared structure element (the shared helix) adopted a specific conformation.
Similarly, the duplication of a C-terminal protein part and its subsequent fusion to the N-terminus of the same protein was exploited to establish conformational competition, as exemplified on the Ca^{2+}-binding protein calbindin D9k (Stratton *et al*, 2008). The duplicated parts of calbindin were now competing to associate with the unique middle domain of the protein to form a functional unit. By introducing a deactivating point mutation into one copy of the duplicated domain, a conformational switch was created. Proper folding of the wildtype domain resulted in a functional protein, while association of the central domain with the mutated copy led to an inactive state. The switching between the two conformations could, once again, be induced by changes in buffer composition.

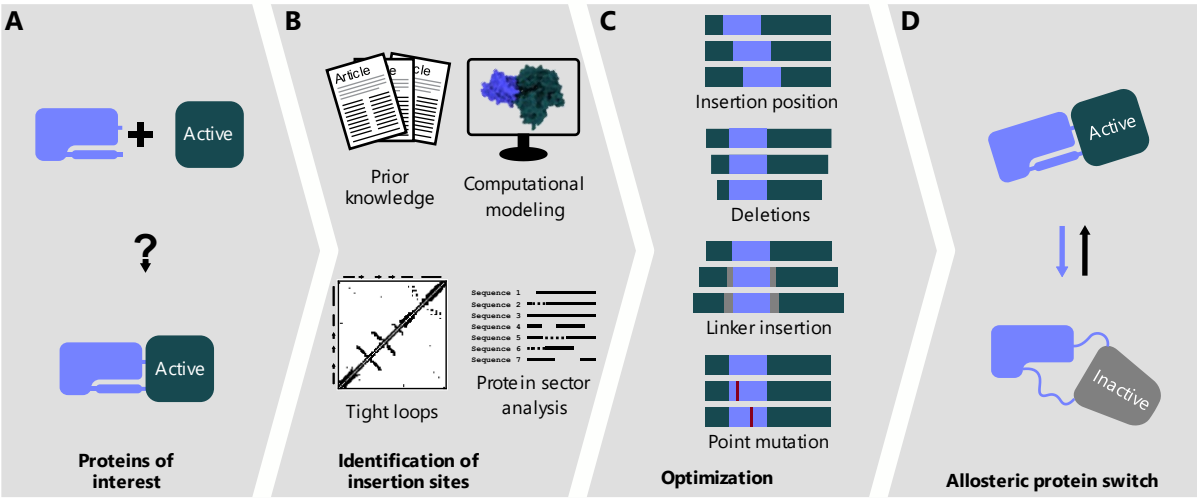


Figure 1.4: Workflow for the design of switchable proteins by domain insertions. (A, B) Surface sites that accept domain insertion (A) and ideally even control of protein conformation (D) are identified with help of prior experimental work, structural modeling, the analysis of surface exposed loops or harnessing information from MSAs. (C) Extensive optimization of the lead candidates is often required to result in a well-functioning conformational switch (D).

1.3.4.2 *Activity switching by induced disorder*

A different rational design strategy was proposed by the group of Klaus Hahn. Dagliyan et al. postulated the engineering of "extrinsic disorder" into proteins, by inserting switchable domains into so-called tight loops (Dagliyan *et al*, 2016, 2019). Tight loops refer to surface-exposed unstructured regions that bridge spatially aligned secondary structure elements. To achieve this, the authors employed the photo-switching mechanism of the *As*LOV2 domain, which results in a structural change of the domain's termini upon excitation by blue light. The

idea was that insertion of the domain does not interfere with the structure of the target protein in the dark state, i.e. when the terminal α-helices of *As*LOV2 are in close proximity. Upon light-induced unfolding of these helices, however, structural tension is imposed on the connected secondary structure elements, ultimately resulting in a conformational distortion of the fused effector protein and hence in an allostery-mediated loss of function. This concept was validated on several kinases, using *As*LOV2 and similarly uniRapR, a rapamycin responsive domain, as inserts (Dagliyan *et al*, 2016).

An elegant recent example made use of a peptide that undergoes a disorder-to-helix conversion, when a second helical peptide is present as a cognate binding partner (Plaper *et al*, 2022). In the bound state, both helices form a coiled-coil. When the unstructured peptide was inserted into the loop of a protein, it did not interfere with protein activity. Addition of the binding partner, however, caused the insert to adopt its helical conformation and to structurally disturb the activity of the parent protein. In contrast to the insertion of larger domains, such as *As*LOV2, this concept is compatible with many protein classes. The insert is relatively small and flexible in its unstructured state, so that more surface sites are suitable for insertion (Plaper *et al*, 2022). On the contrary, the system relies on helical peptides as input triggers, which need to be co-supplied. Non-invasive regulation with inputs such as light is not possible.

1.3.4.3 *Structural ensembles and protein switches*

The working principle of the examples above can be explained by the simplified mechanistic model of allosteric regulation via "induced fit". The addition of a ligand (and similarly light exposure) is believed to induce a conformational change of the protein from one state to another (Boehr *et al*, 2009). While this view on allosteric regulation of protein activity is still helpful and often sufficient as a conceptual basis, the more complex idea of conformational ensembles draws a more nuanced picture (Boehr *et al*, 2009; Motlagh *et al*, 2014; Weber, 1972). It assumes that several thermodynamically favorable conformations coexist in an equilibrium. The relative concentrations of the different folds in the equilibrium are defined by their energy, meaning that energetically favorable states are more prevalent. The coexisting conformations can also be perceived as local minima in an energy landscape (Boehr *et al*, 2009). With respect to conformational switches, it is assumed that a ligand selects for certain pre-existing conformation instead of actively inducing the conformational change (Choi *et al*, 2015; Motlagh *et al*, 2014). In other words, the ligand remodels the energy landscape in a way that the bound conformation becomes energetically more favorable than before (Motlagh *et al*, 2014). An example of how of important these considerations can actually be is given by the DNA-binding mechanism of certain transcription factors. Several variants of the catabolite activator protein (CAP) were shown to exhibit substantial differences in their ability to bind DNA, despite the fact that their respective DNA-binding interfaces are structurally identical (Tzeng & Kalodimos, 2012). It was shown, that the variants only differed with respect to their conformational entropy and that this structural flexibility was necessary for DNA binding (Tzeng & Kalodimos, 2012, 2013).

These observations raise the question: How can our knowledge of structural ensembles be used to guide and to improve the engineering of allosteric switches? Given that for most proteins no structural information or only the crystal structure of a single conformation is available, the nature of the structural ensemble is usually unknown. Nonetheless, Choi et al.

motivated an engineering strategy by conformational ensembles. First, they inserted the TEM1 β-lactamase into a maltose-binding protein (MBP), resulting in a hybrid, both components of which were still constitutively active (Choi *et al*, 2015). Based on this variant, they followed the hypothesis that the insertion of linkers would increase the conformational entropy, enabling inducible structure changes. Indeed, the addition of flexible G-rich linkers resulted in switchable variants, responsive to the addition of maltose, as well as temperature and pH changes (Choi *et al*, 2015). Thermodynamic predictions revealed a decrease in stability for one of the linker regions. Like many other strategies, this procedure has - to my best knowledge - not been validated for other effector proteins.

1.3.4.4 *Allostery prediction and switchable proteins*

The majority of examples discussed above are allosteric switches. A domain is inserted at a site distant to the catalytic center of a protein, but is still able to affect the proteins activity. This behavior must necessarily be caused by long-range structural effects and interdependencies. It is well known that even in proteins that are generally considered non-allosteric, single mutations can affect distant sites within the protein (Clarkson *et al*, 2006), which is in agreement with the assumption that all proteins could potentially be allosterically regulated (Gunasekaran *et al*, 2004). This hypothesis gives rise to the question whether certain surface sites are better suited for engineering of allosteric protein control and if these locations can be predicted? The investigation of allosteric proteins and the principles underlying allostery have been intensively studied (Dokholyan, 2016; Schueler-Furman & Wodak, 2016). However, most of the previous work focused on allosteric intervention via small molecule ligands and less on direct modification of a protein (Lu *et al*, 2014). Overall, the examination of allostery can be divided into analyses based on structural information, such as molecular dynamics or elastic network analysis (Ming & Wall, 2005; Su *et al*, 2014; Zheng *et al*, 2006), and the exploration of allosteric pathways informed by MSA-based coevolutionary insights (Dokholyan, 2016; Lee *et al*, 2008). Interestingly, only the latter approach has been adapted to engineer switches via domain fusions.

The idea to use coevolution in order to predict allosteric sites was pioneered by the group of Rama Ranganathan. It grounds on the observation of evolutionary couplings between functionally important residues and was mathematically formalized as statistically coupling analysis (SCA) (Lockless & Ranganathan, 1999; Halabi *et al*, 2009). The general assumption behind SCA is that the coevolution of residues has a functional meaning. The method uses MSAs to calculate the evolutionary coupling between amino acids (Rivoire *et al*, 2016) and predicts physically connected amino acid networks, comprising coevolving residues termed protein "sector" (Teşileanu *et al*, 2015; Halabi *et al*, 2009). The relevance of sector residues for protein function was demonstrated in several, previous studies (Lockless & Ranganathan, 1999; Süel *et al*, 2003; Halabi *et al*, 2009; Salinas & Ranganathan, 2018). Protein sectors often connect distant surface sites to the active center of a protein. On the basis of this observation, sector-connected surface sites were suggested as promising regions for the engineering of allosteric regulation by receptor fusion (Lee *et al*, 2008). In an initial study, one of two selected surface sites of the *E. coli* dihydrofolate reductase (DHFR) exhibited a modest blue light-induced change in activity upon insertion of the *As*LOV2 domain (Lee *et al*, 2008). In a follow-up study, a set 70 insertion sites was probed. 14 of the generated DHFR-LOV2 hybrids showed noticeable light-dependent activity. The insertion sites underlying these lead candidates were all located

close to predicted sector residues (distance below 4 Å) (Reynolds *et al*, 2011). Based on these results, Pincus et al. outlined an engineering strategy, based on protein sector analysis: the "rational engineering of allostery at conserved hotspots" (REACH) (Pincus *et al*, 2017). The authors suggested to predict surface-exposed sector sites, based on SCA. These sites could subsequently be screened by insertion of the domain of choice. To optimize the lead candidates, the additional incorporation of mutations was discussed (Pincus *et al*, 2017). A subsequent study experimentally analyzed mutational effects on one of the light-switchable DHFR-LOV2 hybrids, demonstrating that allostery-enhancing mutations tended to be enriched outside of the sector (McCormick *et al*, 2021). Interestingly, the combination of some beneficial mutations was shown to have an additive effect. It remains to be seen, whether these observations can be confirmed on different, structurally unrelated target proteins.

1.3.4.5 *Screening-based approaches to study switchable proteins*

Screening of randomized domain insertions in order to identify allosteric protein switches has long been considered a powerful method (Guntas *et al*, 2004; Guntas & Ostermeier, 2004). The concept behind this strategy is to randomly insert a domain of choice at many or even all positions of a protein followed by enrichment of variants with the desired properties. In essence, this method is a variation of the classic directed evolution approaches described in section 1.2.1.2. Fluorescence reporters or cell survival are typical readouts for insertion variant screening. The use of transposon libraries (Edwards *et al*, 2008) or DNA synthesis-based libraries (Coyote-maestas *et al*, 2019) enabled the comprehensive sampling of every possible insertion sites on practically any protein of interest. Although these procedures tend to be labor intensive, they have been successfully used in several cases. By sampling of large domain insertion libraries, allosteric fusions of the TEM1 β-lactamase with MBP (Guntas *et al*, 2004; Guntas & Ostermeier, 2004; Edwards *et al*, 2008), as well as GFP-MBP biosensors (Nadler *et al*, 2016), have been created in the past. Similarly, the screening of comprehensive insertion libraries led to the identification of Cas9 variants that can be activated by 4-hydroxytamoxifen (Oakes *et al*, 2016). Moreover, insertion of zinc-finger motifs into MBP resulted in maltose-dependent DNA-binding of the fusion protein (Younger *et al*, 2018).

While these examples mainly focused on the isolation of lead candidates, the relative fitness of every variant from such a comprehensive library can be assessed via next generation sequencing (NGS). This workflow, which was termed "domain insertion profiling with DNA sequencing" (DIP-seq) (Nadler *et al*, 2016) is based on the enrichment of variants from a library via FACS. In the sorted library, functional variants are enriched, while dysfunctional ones are depleted. Deep sequencing is then used to capture the distribution of variants. This strategy is of particular interest from a mechanistic standpoint, as it measures the insertion tolerance at every single site within a protein under conditions controlled by the experimenter. Studies on ion channels have recently highlighted the potential of such comprehensive domain insertion screens. Coyote-Maestas et al. randomly inserted a PDZ domain, Cib81, as well as short flexible peptides into the human potassium channel Kir2.1 (Coyote-Maestas *et al*, 2019). Using plasma membrane localization of Kir2.1 as readout, the authors first confirmed several trends that have previously been suggested in context of insertional tolerance. Overall, small flexible peptide insertions were better-tolerated than larger, structured inserts. Also, the unstructured C-terminal region of the channel showed a particularly high insertion tolerance. Functionally

important sites, in contrast, exhibited a more differential permissibility for the different inserts. Of particular note, unstructured loops did – unexpectedly – not show a particularly high domain insertion tolerance, albeit this is commonly assumed to be the case. Furthermore, the authors found that dynamic features derived from protein structures (normal modes) modestly correlated with insertion tolerance, while other factors, such as the predicted effects of point mutations at the respective site did not. It was further possible to train decision tree models based on normal modes that could to some degree identify sites that tolerate insertion. Finally, some variants carrying insertions of the Cib81 domain were light-switchable upon co-expression of its cognate binding partner Cry2.
In a large scale follow-up study, the same group randomly inserted 759 different motifs and domains at all possible positions into Kir2.1 (Coyote-Maestas *et al*, 2021). Most strikingly, different inserts tended to be tolerated at the same insertion sites, which could be viewed as domain insertion "hot-spots". In addition, the insertion of smaller motifs was, overall, better tolerated in the central parts of the channel, while successful insertions of larger domains were enriched in regions close to the protein's termini. Interestingly, correlations between the biophysical features of the different inserts and the tolerance of their insertion were site-dependent so that no general trends could be deduced for the dataset. Moreover, random forest models were trained on a combination of biophysical features of the insert motifs and features derived from the recipient protein. Overall, these models were able to explain about 40 % of the variance observed in the data. Hydrophobicity and sequence length were determined to be the most decisive insert properties, while the root mean square fluctuations of residues were most important for model performance on site of the recipient ion channel. It is also noteworthy, that the observed effects were transferrable to other evolutionary related ion channels. Overall, the authors emphasized flexible regions as ideal insertion sites. Although the two studies above may not be generalizable, since they employ protein localization as a readout (not function) and focus on ion channels only (Coyote-Maestas *et al*, 2019), their findings represent an important step towards a more systematic evaluation of domain insertion tolerance.

1.3.4.6 *The role of linkers and structural modeling*

In recent years, the role linker regions that connect insert and parent domains has gained attention in context of domain fusion approaches (Gräwe & Stein, 2020). The drastic effect linkers of varying length and flexibility can have on the performance of protein switches was demonstrated in several studies (Gräwe *et al*, 2022; Ranglack *et al*, 2020; Bubeck *et al*, 2018). However, a rational engineering strategy that predicts optimal linkers for a given fusion constructs has thus far not been developed and linker engineering still remains a trial an error procedure.
Finally, computational algorithms to model domain insertions have been created, building on the Rosetta framework (Berrondo *et al*, 2008; Blacklock *et al*, 2018). However, these strategies have to my best knowledge never been experimentally validated.

1.3.5 Protein domains as conformational switches

The last chapters highlighted the challenge to identify a proper insertion site, while selection of the insert domain was given less consideration. Although Coyote-Maestas et al. tested an impressive number of inserts in context of Kir2.1, the majority of them were small, often

unstructured motifs and the study was not focused on allosteric control of protein activity (Coyote-Maestas *et al*, 2021). In the literature, only a small set of insert domains is recurrently used. Among them is the estradiol-binding domain from human estrogen receptor-α (ERD), the uniRapR domain and the *As*LOV2 domain. All three domains have in common that they change their conformation upon stimulation with their respective trigger, which renders them good candidates for the engineering of allosteric protein switches by domain insertion.

1.3.5.1 *The Estradiol-binding domain*

The ERD has already been used for chemogenetic control of proteins for more than 20 years (Feil *et al*, 1997). In its apo state the termini of the ERD are widely separated with the C-terminal helix 12 sticking out from the protein core (Tanenbaum *et al*, 1998; Oakes *et al*, 2016). The distance between N- and C-terminus was determined to be 64 Å in the apo state (Oakes *et al*, 2016). Binding of the β-estradiol results in a conformational change leading to a decrease in the termini's distance to 37 Å (Wärnmark *et al*, 2002). Interestingly, binding of 4-hydroxytamoxifen (4-HT), an β-estradiol antagonist, results in a third, distinct conformation, which is even more compact (Shiau *et al*, 1998). Here, the termini are only separated by a distance of 21 Å. Feil et al. further identified a triple mutant G400V/M543A/L544A, that is only responsive to 4-HT, while not being able to bind β-estradiol anymore (Feil *et al*, 1997). The published ERD-based protein switches were all activated by addition of the ligand (Oakes *et al*, 2016; Feil *et al*, 1997). In the apo-state, the ERD presumably disturbs the conformation of the effector it is inserted into, due to the strain imposed at the insertion site by the ERD's termini. Ligand binding, however, is expected to release that strain by moving the ERD's termini closer together, hence allowing the fused effector protein to adopt its native, active structure.

1.3.5.2 *The uniRapR domain*

The uniRapR is a synthetic switch, based on FKBP12 and FRB (Dagliyan *et al*, 2013). Both domains are well-characterized and known to dimerize in response to rapamycin binding (Choi *et al*, 1996). More specifically, uniRapR is a synthetic construct created by fusing an engineered variant of FKBP12 (Karginov *et al*, 2010) to a truncated version of FRB. In response to rapamycin, the domain undergoes a stabilization through interaction of the two binding partners (Dagliyan *et al*, 2013). Upon insertion of the uniRapR domain into effector proteins, this stabilization was shown to affect the flexibility of certain protein regions, which is supposed to cause the activation of the target protein (Dagliyan *et al*, 2016, 2016).

1.3.5.3 *The AsLOV2 domain*

The *As*LOV2 domain, is the insert domain most extensively used in this study. It was identified as part of the plant photoreceptor phototropin-1 of the common oat, *Avena sativa*. The use of photoreceptors for targeted control of cells was first demonstrated by Edward Boyden and Karl Deisseroth, who employed the ion channel channelrhodopsin-2 from *Chlamydomonas reinhardtii* for the regulation of neuronal activity (Boyden *et al*, 2005; Nagel *et al*, 2003). In the two decades since this landmark study was published, numerous different photoreceptors have been employed for the optogenetic control of cells. In fact, *As*LOV2 is only one of several functionally similar LOV domains, used in optogenetics, although it is arguably the most intensively studied one (Hoffmann *et al*, 2018; Mathony & Niopek, 2021). LOV domains are part of Per-ARNT-Sim (PAS; period circadian protein–aryl hydro-carbon receptor nuclear

translocator protein–single-minded protein) domain class (Möglich *et al*, 2009, 2010). Their photo-activation is mediated by flavin nucleotide co-factors. In case of *As*LOV2, the chromophore is a flavine mononucleotide (FMN), which is bound by the LOV2 domain core (Fig.1.5). (Crosson & Moffat, 2002; Salomon *et al*, 2001). Excitation of the chromophore by absorption of blue light (up to ~490 nm) induces the formation of a covalent bond between the FMN and C450 (numbering follows the positions within the full-length phototropin-1) of the *As*LOV2 domain (Fig. 1.5) (Crosson & Moffat, 2002; Salomon *et al*, 2001). This results in a structural rearrangement driven by the release and formation of several hydrogen bonds in LOV2. Ultimately, these structural adaptions lead to the unfolding of the two terminal α-helices, A'α and Jα, which are initially attached to the domains core (Fig. 1.5) (Harper *et al*, 2003; Halavaty & Moffat, 2007; Zayner *et al*, 2012). Importantly, the excitation of *As*LOV2 is fully reversible, i.e. in the absence of light, the LOV2 domains falls back into the dark-adapted state within about ~1 min (Swartz *et al*, 2001).

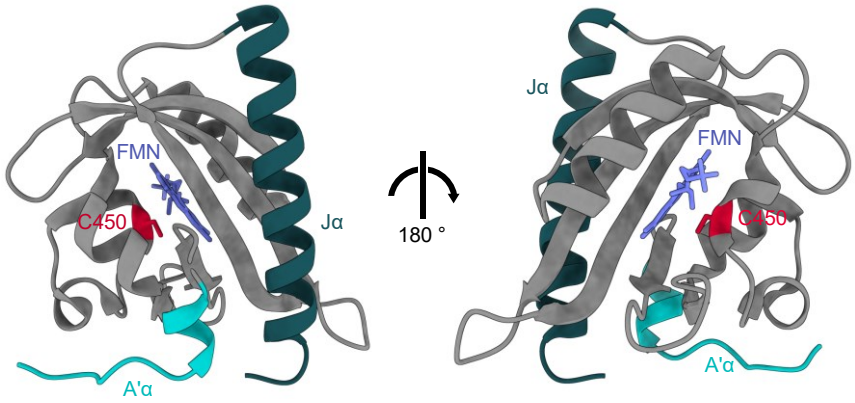


Figure 1.5: Structural features of the *As*LOV2 domain. A cryo-electron microscopy structure of the *As*LOV2 domain is shown. The FMN chromophore is indicated in blue and the functionally important residue C450 is shown in red. The terminal helices A'α and Jα unfold upon photo-excitation. They are marked in different tones of green. PDB-ID: 2V0U.

Over the years, a large number of mutants affecting the *As*LOV2 photo-cycle have been described. Among these are modifications that lock the domain in a pseudo dark-adapted (C450A and C450M) (Richter *et al*, 2005; Kothe *et al*, 2014; Wong *et al*, 2015) or pseudo light-adapted state (I510E, I532E, A536E and I539E) (Yao *et al*, 2008; Wang *et al*, 2016; Strickland *et al*, 2008). In addition, mutations that improve the docking of the Jα-helix to the protein core in the dark state (Strickland *et al*, 2010) or affect the duration of the photo-cycle have been identified (Zoltowski *et al*, 2009; Wang *et al*, 2016). For a more detailed description of *As*LOV2 mutants and their impact on LOV2 photoswitching and photocycle, I kindly refer the reader to reviews by Pudasaini et al. and Hoffmann et al. (Hoffmann *et al*, 2018; Pudasaini *et al*, 2015). As a note of caution, it has become apparent that the effect of the described mutations highly depends on the protein context in which the *As*LOV2 domain is used and a variability from case to case is to be expected (Hongdusit *et al*, 2020; Bubeck *et al*, 2018)

The adaption of the *As*LOV2 domain for the engineering of photo-switches is grounded on its compact size (16.5 kDa) and the reversible unfolding of the terminal α-helices, in particular the long C-terminal Jα helix (24 AA). Several studies, for instance, employed a photo-caging strategy based on the fusion of the C-terminal Jα helix to functional peptides. This approach enabled the light-controlled nuclear import or export (Niopek *et al*, 2014, 2016; Yumerefendi

et al, 2016, 2015), as well as the light-induced degradation (Bonger *et al*, 2014; Renicke *et al*, 2013) of proteins. To this end, hybrid sequences between localization- or degradation tags and the Jα-helix were created. In the dark-adapted state, the helical tag is folded and attached to the LOV2 domain, hence shielding these peptides from their cognate receptors. Upon illumination, in turn, unfolding of the Jα-helix exposes the peptide tag, which can then mediate the desired function.

The light-dependent relaxation of the *As*LOV2 terminal helices can also be employed to engineer allosteric effectors by LOV2 domain insertion. Importantly for this approach, the distance between the *As*LOV2 helical termini in the dark-state is only ~10 Å. Consequently, *As*LOV2 insertion tends to be well-tolerated in the dark (Dagliyan *et al*, 2016). Upon light absorption, however, the domain's terminal helices relax, which will create flexibility around the LOV2 insertion site. This "local disorder" can eventually disturb the structure and hence activity of the fused protein (Bubeck *et al*, 2018; Dagliyan *et al*, 2016). Functionally, this mechanism is practically the opposite of the chemically regulated domains described above (ERD, uniRapR), which cause a trigger-induced activation of the fused effector protein. Besides the common light-induced deactivation of protein activity via *As*LOV2, also light-activated *As*LOV2-based hybrids have been described (Reynolds *et al*, 2011; Gil *et al*, 2020), pointing towards a highly context-specific mediation of protein (de-)activation. Although the *As*LOV2 domain has been subject to extensive study and use by the optogenetics community, new applications in combination with different proteins still rely substantially on optimization by trial and error (Bubeck *et al*, 2018; Gil *et al*, 2020; Mathony & Niopek, 2021).

1.4 Applications of switchable proteins in transcription regulation and gene editing

In this study, the optogenetic control of proteins is exemplified and investigated in the context two different use cases, as well as protein classes: First, the optogenetic control of transcription in bacteria using engineered, light-dependent transcription factors; second, the light-mediated control of gene editing tools based on clustered regularly interspaced short palindromic repeats (CRISPR)-effectors.

1.4.1 Optogenetic control of transcription in bacteria

Transcription control in bacteria is a widely studied field of central importance for metabolic engineering, bioproduction and synthetic biology. In fact, the two seminal publications that laid the foundation for synthetic biology, described genetic circuits based on bacterial transcription factors (Gardner *et al*, 2000; Elowitz & Leibier, 2000). Since that time, optogenetics has long found its way into the area of bacterial gene expression control. Here, I will give a brief overview of the existing optogenetic transcription systems in *E. coli*.

1.4.1.1 *Diversity of optogenetic expression systems*

A variety of optogenetically regulated transcription systems for use in *E. coli* have been developed over the past 15 years. The strategies and functional mechanisms underlying these are diverse. Optogenetic systems can be categorized by several parameters, such as the used photoreceptor or the mechanism of action. Here, the existing tools are divided into direct

versus indirect transcription activators. Indirect means, that the optogenetic mechanism does not immediately affect transcription and at least one additional mediator protein is needed. Well-studied examples are bacterial adenylate cyclases (BAC), which convert adenosine triphosphate (ATP) into cyclic adenosine monophosphate (cAMP). Several natural light-dependent BACs are known (Raffelberg *et al*, 2013; Stierl *et al*, 2011). cAMP, in turn can activate gene expression by binding to the transcription factor CAP (Busby & Ebright, 1999). A number of optogenetic tools using light-dependent BACs have been described (Stüven *et al*, 2019; Blain-Hartung *et al*, 2018; Ryu *et al*, 2014). In addition, engineered BACs in which the native regulatory domains were exchanged to a red light responsive receptor were reported (Ryu *et al*, 2014, 2010; Blain-Hartung *et al*, 2018). The modular domain architecture of BACs enabled this domain swap. An advantage of the longer wave lengths of red light is an increased tissue penetration, as exemplified by BAC application in *Caenorhabditis elegans* (Ryu *et al*, 2014; Shu *et al*, 2009). On the other hand, red light-absorbing bacteriophytochromes require biliverdin as a co-factor, which is a natural metabolite in many eukaryotes, but not in *E. coli*. It thus needs to be co-supplied in the culture media or must be synthesized *in vivo* upon expression of additional enzymes (Ryu *et al*, 2014). Another drawback is the fact that cAMP is a rather universal second messenger in bacteria so that adverse side effects could arise when employing this molecule as mediator for gene expression control. An elegant solution for this problem is provided by the red light-inducible diguanylate cyclase BphS, which produces cyclic dimeric guanosine monophosphate (c-di-GMP) from GTP (Ryu & Gomelsky, 2014). Its reaction product, c-di-GMP, is not a natural metabolite of *E. coli*. The enzyme can be used in combination with the c-di-GMP-binding transcription factor MrkH. However, to keep c-di-GMP levels at a baseline a c-di-GMP phosphodiesterase had to be co-supplied.

Another mechanistically different strategy employs bacterial two-component systems, consisting of histidine kinases and their cognate transcription factors. Levskaya et al., for instance, harnessed a chimeric red light-responsive membrane-bound kinase for transcription activation (Levskaya *et al*, 2005) This engineered chimera was derived by fusing the phytochrome Cph1 from *Synechocystis* to a structurally similar histidine kinase EnvZ from *E. coli*. In the dark, auto-phosphorylation occurs, followed by phosphorylation of the transcription factor OmpR, thereby leading to the activation of gene expression. Upon stimulation with red light, however, the kinase becomes inactive and gene expression turns off. A number of similar systems were developed using natural and engineered kinases including photoreceptors that respond to blue and green light (Tabor *et al*, 2011; Ohlendorf *et al*, 2012; Schmidl *et al*, 2014; Ong & Tabor, 2018). It is important to note that the directionality of the switch depends on the used photoreceptor. The protein ccaS, for instance, initiates the phosphorylation cascade upon stimulation with green light (instead of inhibition by light) (Tabor *et al*, 2011). The activation/deactivation directionality could also be inverted by incorporation of an intermediate step in which the phosphorylated transcription activator induced the expression of a transcriptional repressor, such as LacI. This repressor, in turn, regulated the expression of the actual gene of interest (Multamäki *et al*, 2022; Lalwani *et al*, 2021). Despite increasing the complexity of the whole system, this approach was also successfully used to boost the dynamic range of a previously generated light switch following the above design principle (Lalwani *et al*, 2021).

In contrast to such mediator-dependent, indirect approaches, direct gene expression control can be achieved, e.g. using split-protein systems (refer to section 1.3.2). The widely applied RNA polymerase from the T7 phage mediates transcription only from its cognate T7 promoters. Several studies created Split-T7 polymerase variants fused to optogenetic homo- or heterodimerization systems, i.e. the so-called magnets and VVD domains, respectively, both of which react to blue light (Baumschlager *et al*, 2017; Han *et al*, 2017; Seifert *et al*, 2019). Also, the fusion of only one VVD domain to a split-part of the T7-polymerase turned out to be sufficient for optogenetic regulation, as the dark-adapted conformation of VVD already blocked the re-association of the full-length polymerase (Baumschlager *et al*, 2017). To enable the same photo-regulation by red light, Raghavan et al. linked the split-T7 parts to inteins, which were fused to the phytochrome B (phyB)/phytochrome-integrating factor 3 (PIF3) domain pair, a red light inducible dimerization system from *Arabidopsis thaliana* (Raghavan *et al*, 2020). Dimerization of the photo-receptors triggered trans-splicing of the inteins, resulting in reconstitution of the single chain T7-polymerase. This approach was, however, not reversible and, due to it being based on the PhyB/PIF3 system, required the chromophore phycocyanobilin to be exogenously supplied.

A particularly interesting example for gene expression control with split proteins are optogenetic recombinases re-constituted with help of optogenetic dimerizers, e.g. based on VVD (Sheets *et al*, 2020; Sheets & Dunlop, 2022). Similar to the two-component systems described above, recombinases do not directly control transcription, but can be used to initiate the reconstitution of a functional gene expression cassette, for instance by excising a terminator that is placed between a promoter and the coding sequence of a gene of interest (Sheets *et al*, 2020; Sheets & Dunlop, 2022). Interestingly, this strategy results in a binary behavior, since the removal of the terminator is a singular event, which irreversibly switches on gene expression.

Optogenetic transcription factors, in turn, represent an alternative to split-proteins. A natural example is the bacterial protein EL222. It consists of a helix-turn-helix (HTH) DNA-binding domain linked to a LOV2 domain and constitutes, like many transcription factors, a homodimer (Rivera-Cancel *et al*, 2012). Interestingly, EL222 can act as a repressor in the dark, as well as a transcription activator upon induction with blue light (Jayaraman *et al*, 2016; Camsund *et al*, 2021). Apart from light-regulated transcription by EL222 alone, hybrid promoters, bearing additional binding sites for the chemically inducible transcription factors AraC or LasR were constructed, which then respond to a combination of light and chemicals as input (Jayaraman *et al*, 2018). Similarly to EL222, the bacteriophytochrome photoreceptor 1 (BphP1) controls transcription via red light induction (Ong *et al*, 2018). However, this receptor depends, again, on the supply of biliverdin as a co-factor and can only repress transcription. A more unconventional mechanism of action was described for another repressor, CarH, which binds the coenzyme B12. Ligand binding leads to oligomerization and DNA-binding of the repressor (Ortiz-Guerrero *et al*, 2011). Light-exposure, in turn, triggers the dissociation of the co-factor and release of CarH from DNA (Ortiz-Guerrero *et al*, 2011). In contrast to other proteins, the photo-mechanism can be activated at diverse wavelengths, ranging from 360 nm to 540 nm.

Besides the adaption of natural transcription factors for the control of gene expression, also artificial transcription regulators have been engineered. To this end, the DNA-binding domain (DBD) of the widely used AraC and TetR proteins were fused to the VVD dimerization domains,

resulting in robust optogenetic switches (Komera *et al*, 2022; Romano *et al*, 2021). In these examples, the TetR hybrids act as repressors that are activated by light, while the AraC fusions inhibit transcription in the dark state. The reasons are different mechanisms of action: In case of TetR, dimerization is required for binding to its corresponding DNA recognition motif (Komera *et al*, 2022). AraC, in contrast, already binds to the pBAD promoter in the dark state where it acts as repressor, while the light-induced conformational change alters its properties toward an activator-state (Schleif, 2010). The details of AraC-DNA interactions are described in the next section (1.4.1.2). A similar approach was published by the group of Andreas Möglich, who fused a LOV domain from *Rhodobacter sphaeroides* (*Rs*LOV) to the DBD of TetR in order to render the transcription factor light-responsive (Dietler *et al*, 2021). Due to the intrinsic properties of *Rs*LOV, the resulting repressor was not only light controllable, but its activity was also temperature-dependent. While fusions with the wildtype *Rs*LOV exhibited limited switchability at higher temperatures around 37 °C, different mutants with optimized dynamic ranges were reported (Dietler *et al*, 2021).
The described strategies showcase the rich toolbox comprising diverse optogenetic transcription systems and their various mechanistic features. Importantly, all of these tools have specific properties and corresponding advantages and disadvantages due to differences in the activation wavelength, the specific dynamic range of activation, the number of genetically encoded components required, the degree of reversibility, requirements for exogenous co-factors and the leakiness of the systems in the off-state. Furthermore, the last years have witnessed the emergence of optogenetic gene expression systems based on more than one input (Dietler *et al*, 2021; Jayaraman *et al*, 2018). These allow to construct more complex genetic logic functions. Since the optogenetics field constantly innovates, the toolbox of light-controlled transcription regulation in bacteria is expected to further grow in the future.

1.4.1.2 *Structure, function and application of AraC*

AraC is a member of the AraC/XylS transcription factor family. In *E. coli*, AraC controls the expression of genes from the arabinose operon by regulating the activity of the pBAD promoter (Fig. 1.6) (Schleif, 2010). Importantly, AraC can act as both, a transcriptional repressor and an activator. It does so by binding to different operator sites within the pBAD promoter region. Two so-called araI1 and I2 half-sites are located close to the transcription start site, while an additional operator site (O2) is positioned several hundred base pairs upstream (Schleif, 2010). In its apo state, an AraC homodimer binds the O2 and I1 half-sites thus creating a DNA loop, which results in transcription repression (Lobell & Schleif, 1991). Upon binding of arabinose, the AraC dimer associates with both araI half-sites, now acting as a transcription activator by promoting RNA polymerase binding (Lobell & Schleif, 1991).
Structurally, each AraC monomer consists of two domains. An N-terminal arabinose-binding domain (Soisson *et al*, 1997) and a C-terminal α-helical DBD (Rodgers & Schleif, 2009). The DBD consists of two canonical HTH DNA-binding motifs, which are separated and flanked by additional α-helices (Rodgers & Schleif, 2009). The LBD constitutes a β-barrel, containing the arabinose-binding pocket (Soisson *et al*, 1997). An N-terminal "arm" comprising the first 20 amino acids remains unstructured in the apo-state and acts as a "lid" for the binding pocket once arabinose is bound (Soisson *et al*, 1997). The C-terminal end of the LBD is a coiled coil that mediates the dimerization of AraC in the active state. It is C-terminally connected to the

DBD via a longer interdomain linker. In absence of arabinose, dimerization is mediated by the β-barrel instead of the coiled-coil (Soisson *et al*, 1997).

The molecular mechanism underlying the arabinose-induced switching is still not fully understood. It is well established though that binding to both neighboring I1/2 sites in the uninduced state is prevented due to a lack of conformational flexibility. In consequence, AraC bridges the distant I1 and O2 sites (Harmer *et al*, 2001). Two structural elements have been of particular interest with respect to the AraC regulatory mechanism. First, the AraC N-terminal arm is known to attach more tightly to the LBD in presence of arabinose (Soisson *et al*, 1997). It is speculated, that this arm mediates contact to the DBD in absence of arabinose, supposedly resulting the restricted flexibility, which defines the DNA-binding preferences. This hypothesis was mainly studied by mutation experiments and molecular dynamics (MD) simulations (Lowe *et al*, 2014). It is supported by the fact that many mutations in the arm strengthened the activation of AraC (Tang & Cirino, 2010) or even led to constitutively active behavior (Dirla *et al*, 2009; Wu & Schleif, 2001a). Furthermore, constitutively negative mutations in the DBD could be rescued by additional mutations in the arm (Saviola *et al*, 1998; Wu & Schleif, 2001b). Nonetheless no direct evidence that the arm indeed affects the inter-domain mobility has so far been reported.

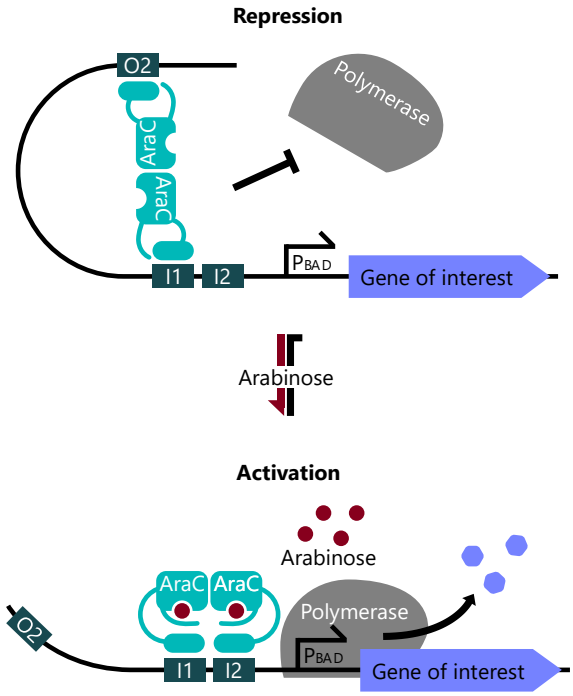


Figure 1.6: AraC – Mechanism of action. In absence of arabinose, the AraC dimer binds to the I1 and O2 operator sites, resulting in the repression of transcription (upper panel). Arabinose induction leads to a conformational change, accompanied by the "capping" of the arabinose-binding pocket via the N-terminal "arm" of the protein. The AraC dimer can now bind to the I1 and I2 half-sites, resulting in transcription activation (lower panel).

The second region of interest is the linker between DBD and LBD. It was suggested to play an important role by mediating the domain flexibility and in consequence DNA binding preferences (Seedorff & Schleif, 2011; Eustance *et al*, 1994). An arabinose-dependent conversion from a helical to an unstructured state has been proposed (Brown & Schleif, 2019; Malaga *et al*, 2016). At the same time, it is well established that this linker tolerates many mutations without measurable effects on AraC function, thus speaking against its critical role (Seedorff & Schleif, 2011; Malaga *et al*, 2016).

Despite our incomplete mechanistic understanding, AraC is widely employed as a tool for inducible protein expression. Its main advantage is the very low leakiness as compared to other gene expression control systems in *E. coli*, such as the IPTG-dependent lac promoter (Guzman *et al*, 1995; Balzer *et al*, 2013). However, expression from AraC inducible cassettes tends to exhibit an all or nothing behavior, limiting the gradual titration of expression levels (Siegele & Hu, 1997; Khlebnikov *et al*, 2002). Over the years, AraC was used for the design of genetic circuits in diverse application contexts (Otero-Muras & Banga, 2017; Daniel *et al*, 2013; Stricker

et al, 2008). Furthermore, engineered variants have been employed as biosensors for metabolites (Picard *et al*, 2022; Tang *et al*, 2008). Here, the vast number of naturally occurring homologues that respond to different ligands made the rational design of new variants with altered biding properties relatively easy (Cortés-Avalos *et al*, 2021). Recently, also an optogenetically switchable variant of AraC was published (see section 1.4.1.1 for details) (Romano *et al*, 2021). In summary, AraC is one of the best studied and most frequently used bacterial transcription factors.

1.4.2 CRISPR-Cas9

1.4.2.1 *The bacterial CRISPR-Cas immune system – a brief history*

CRISPR research dates back to the 1980s, when genomic sequence repeats of 29 base pairs (bp) were discovered in *E. coli* (Ishino *et al*, 1987). Over the next years, similar repats were described in diverse bacterial and archeal species, the function of which was entirely unclear at the time (Mojica *et al*, 1993, 2000). In 2002, the term CRISPR (for clustered regularly interspaced short palindromic repeats) was established and the presence of CRISPR associated (Cas) genes, located in vicinity to the CRISPR loci, was identified (Jansen *et al*, 2002). Later, it became clear that the spacers sequences separating the CRISPR repeats originated from phages and appeared to mediate resistance against the pathogens (Mojica *et al*, 2005; Bolotin *et al*). Phage infection was further proven to trigger the integration of new spacers into the CRISPR locus (Barrangou *et al*, 2007). The main missing link was made in 2008, when the groups of Eugene Koonin and John van der Oost showcased how RNAs transcribed from the CRISPR repeats are

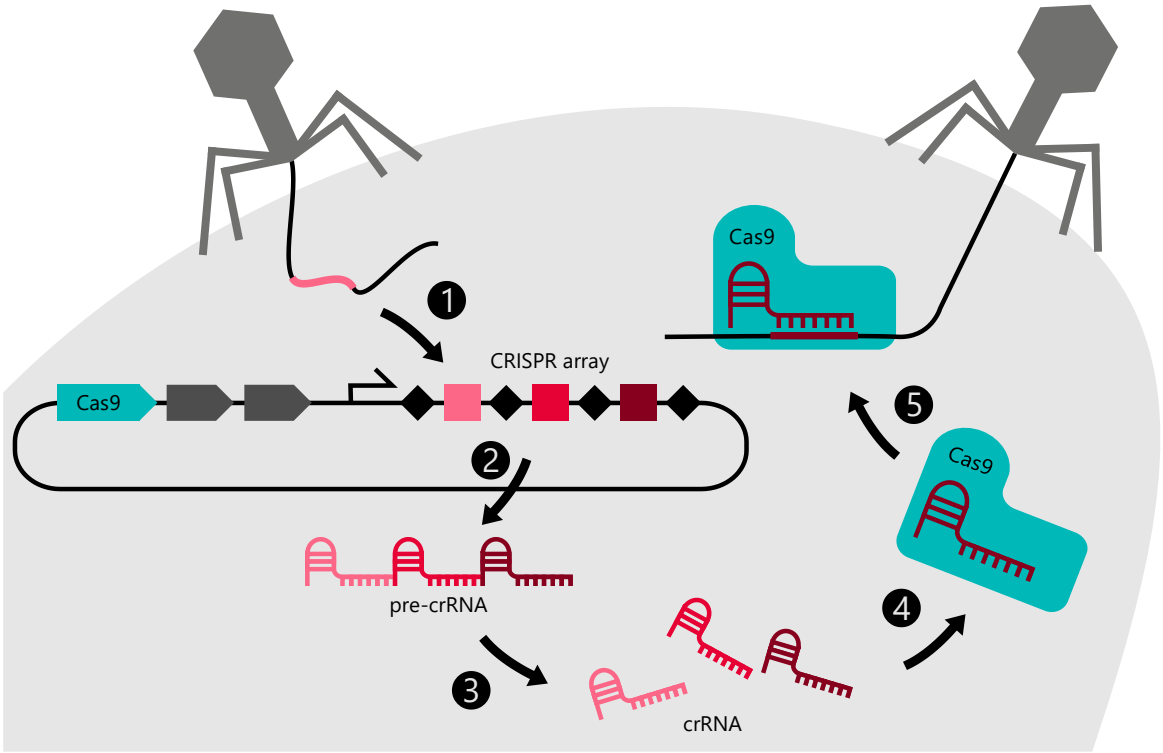


Figure 1.7: The adaptive CRISPR-Cas immune system of bacteria. Fragments from phage genomes are integrated into the CRISPR array upon phage infection (1). Expression of the array generates pre-crRNA (2) which is processed into mature crRNA (3). Cas proteins bind the crRNA with help of a second RNA (trans-activating RNA, not shown) (4), enabling them to selectively bind to and cleave invading DNA (5).

processed by complexes of Cas proteins, called Cascade (Brouns *et al*, 2008). The resulting small RNAs were shown to serve as guides for Cascade to target phage genomes, hence interfering with infections (Brouns *et al*, 2008). In 2012, seminal papers by Jinek et al. and Gasiunas et al. showed that a Cas protein (Cas9), guided by CRISPR-RNAs (crRNAs) was able to induce DNA double strand breaks (DSBs) *in vitro* (Gasiunas *et al*, 2012; Jinek *et al*, 2012).
Today, the process of CRISPR-mediated phage immunity is well understood, as nicely exemplified by the *Streptococcus pyogenes* (*S. pyogenes*) CRISPR-Cas9 system. CRISPR-Cas loci usually consist of the described CRISPR array and several Cas genes located in genomic proximity (Makarova *et al*, 2015). Upon phage infection, Cas1 and Cas2 mediate the acquisition of new protospacers, by inserting pieces of the invading DNA between the CRISP repeats into the bacterial genome (Fig. 1.7, (1)) (Arslan *et al*, 2014; Yosef *et al*, 2012; Nuñez *et al*, 2014). Cas1 and Cas2 are also the only two proteins present in all types of CRISPR sytems, pointing towards a mechanistically universal role (Makarova *et al*, 2012).
Expression of the CRISPR array containing the acquired spacers, first results in a pre-crRNA (Fig. 1.7, (2)), which is then cleaved into smaller mature crRNAs consisting of one protospacer and one of the repeats (Fig. 1.7, (3)) (Carte *et al*, 2008; Haurwitz *et al*, 2010; Wang *et al*, 2011; Gesner *et al*, 2011). This process is less conserved between different CRISPR classes. In *S. pyogenes* a trans-activating (tracr)RNA is required for crRNA to be processed by Csn1 and RNase III (Deltcheva *et al*, 2011). The mature crRNA in duplex with the tracrRNA is finally bound by Cas9 and the ribonucleoprotein complex cleaves DNA at sites which are complementary to the protospacer RNA sequence (Fig. 1.7, (4, 5)) (Jinek *et al*, 2012). In order to avoid targeting of the own CRISPR array, Cas9 requires a so-called protospacer adjacent motif (PAM) (Deveau *et al*, 2008; Mojica *et al*, 2009). This PAM sequence must be located directly 3' of the target site for Cas9 (Mojica *et al*). In case of Cas9 from *S. pyogenes* (*Spy*Cas9) the PAM consists of the nucleotides NGG (Jinek *et al*, 2012).

1.4.2.2 *Diversity of CRISPR-Cas systems*

The diversity of CRISPR-Cas systems in nature is stunning. They are found in approximately 50 % of all bacterial species (Makarova *et al*, 2015). The classification of CRISPR systems is defined by the organization of the CRISPR locus, as well as the sequence and function of Cas proteins (Makarova *et al*, 2015). CRISPR systems can be subdivided into two classes. Class I is characterized by a multi-protein nuclease complex and will not be discussed here as this architecture renders them less attractive for gene editing applications (Makarova *et al*, 2015, 2012). Class II systems in contrast, possess single-protein nucleases, such as Cas9, the simplicity of which explains the popularity they gained over the last decade. This class can be further grouped into the types II, V and VI. Again, genomic organization and their prevalence in different taxonomic clades are among the reasons for this classification. Interesting in this context are the properties of their respective nucleases. Type II systems, which were the first to be used for CRISPR gene editing (Cong *et al*, 2013; Mali *et al*, 2013b), are characterized by a Cas9 nuclease, which is guided by a complex of crRNA and tracrRNA (Makarova *et al*, 2015). Type V systems, in contrast, employ Cas12 nucleases, which are nowadays also frequently employed for genome engineering (Zhang, 2019, 12; Makarova *et al*, 2015). Apart from differences regarding the domain architecture of the proteins, Cas12 does not require a tracrRNA and is further able to process its own pre-crRNA without the need for helper proteins

(Dong *et al*, 2016, 1; Zetsche *et al*, 2015a, 1). Finally, type VI CRISPR systems target single-stranded RNA via Cas13 nucleases (Abudayyeh *et al*, 2016) and have been repurposed for RNA-editing and -interference (Abudayyeh *et al*, 2017; Cox *et al*, 2017).

1.4.2.3 *CRISPR-Cas as a gene editing tool*

The report of the targeted DNA cleavage by Cas9 (Gasiunas *et al*, 2012; Jinek *et al*, 2012) kicked off a revolution in the field of gene editing. To understand how disruptive the development of CRISPR-based technologies was, one should consider the problems of previous gene targeting techniques. Although genome engineering can principally be achieved without the use of nucleases, namely by harnessing homologous recombination, this method is extremely inefficient and highly dependent on the cell type (and species), as well as the cell cycle state (Capecchi, 1989; Thomas *et al*, 1986; Smithies *et al*, 1985).

It has long been known that DNA DSBs can lead to more efficient gene editing (Rudin *et al*, 1989; Rouet *et al*, 1994). In humans and many other eukaryotes, DSBs are mainly repaired by one of two competing pathways. During non-homologous end-joining (NHEJ) the open DNA ends are fused together without the need for repair templates (Weterings & Chen, 2008). This pathway is prone to errors though, often leading to the formation of small insertions or deletions (indels) comprising one to several base pairs (Rouet *et al*, 1994). When a DSB within the coding sequence of a gene is repaired via NHEJ, the occurring indels often result in frame shifts and premature stop codons. As a result, the correct gene product is no longer expressed. This is why the pathway can be employed to generate genetic knock-outs. The second repair option is the homology-directed repair, which enables perfect reconstitution of the original DNA sequence using a homologous template, usually the sister chromatid (Jasin & Rothstein, 2013). Due to the use of a repair template, the pathway is less prone to errors. In case an artificial template is provided, this process can be exploited to introduce knock-ins of custom DNA sequences (Storici *et al*, 2006, 2003). Importantly, the repair pathway choice is highly regulated and dependent on the cell cycle phase (Scully *et al*, 2019). DNA repair and its role in gene editing are reviewed in greater detail elsewhere (Jasin & Rothstein, 2013; Weterings & Chen, 2008; Scully *et al*, 2019).

For a long time, the targeted introduction of DSBs had been challenging. Initially programmable nucleases, that could be recruited to a DNA site of choice were designed as fusion proteins between DNA-binding zinc finger domains and DNA endonucleases, such as FokI (Kim *et al*, 1996; Bibikova *et al*, 2001; Porteus & Baltimore, 2003). Zinc fingers recognize specific combinations of three nucleotides. The consecutive fusion of several such modules enabled highly specific targeting of custom DNA sequences. Similarly, transcription activator-like effectors (TALEs) are protein domains that recognize single bases, which further increased the flexibility with respect to user-defined target sequences (Moscou & Bogdanove, 2009; Boch *et al*, 2009). Although being of great value, TALE nucleases (TALEN) and zinc finger nucleases (ZFN) suffer from the inherent disadvantage, that large parts of the respective protein need to be modified in order to target a specific DNA sequence.

Here the beauty of CRISPR-Cas systems becomes apparent, as they only require redesign of the 20-30 nt protospacer RNA sequence (Jinek *et al*, 2012; Gasiunas *et al*, 2012). To make things even simpler, Jinek et al. showed that the crRNA and tracrRNA could be fused together, resulting in a two-component gene editing system comprising the nuclease in complex with a

single guide RNA (sgRNA) (Jinek *et al*, 2012). It is no surprise, that this gene editing strategy was immediately adapted for the use in mammalian cells (Cong *et al*, 2013; Mali *et al*, 2013b).
Importantly, the application spectrum of Cas9 is not limited to gene editing. Catalytically dead variants (dCas9) can still be targeted to DNA and act as repressors (Qi *et al*, 2013). Fusion of additional repressor domains, such as the Kruppel-associated Box (KRAB) domain, to dCas9 can further improve the gene silencing efficiency (Gilbert *et al*, 2013). Vice versa, the linkage of dCas9 to transcription activation domains enable the induction of gene expression, when Cas9 is targeted to promoter regions (Gilbert *et al*, 2013; Perez-Pinera *et al*, 2013; Maeder *et al*, 2013; Cheng *et al*, 2013). CRISPR activation cannot only be achieved by protein fusions, but also by activator recruitment to the sgRNA scaffold (Zalatan *et al*, 2015; Konermann *et al*, 2015; Mali *et al*, 2013a). Moreover, dCas9 effector fusions are not limited to transcription regulation, but can be employed for the targeted regulation of epigenetic modifications, such as DNA methylation (Liu *et al*, 2016a; Amabile *et al*, 2016; Vojta *et al*, 2016; Xu *et al*, 2016) or histone acetylation (Kwon *et al*, 2017). dCas9, tagged with fluorescent proteins, was further used to image specific DNA loci (Chen *et al*, 2013; Ma *et al*, 2015b)
Another transformative set of innovations was developed by the lab of David Liu. The fusion of dCas9 and Cas9 nickases (nCas9) to Cytosine or Adenosine deaminases enabled the programmed conversion of Cytosines (C) into Thymines (T) or Adenosines (A) into Guanosines (G), respectively (Komor *et al*, 2016; Gaudelli *et al*, 2017). This method, called base editing, facilitated the precise re-writing of single nucleotides, without the need for introduction of DSBs. Moreover, the development of prime editing, even allowed the exchange of larger sequence stretches without DSBs through a fusion between nCas9 and a reverse transcriptase (Anzalone *et al*, 2019).
Finally the discovery of RNA targeting Cas orthologues (Abudayyeh *et al*, 2016, 1; East-Seletsky *et al*, 2016, 2) lead to the repurposing of the above-mentioned methods to the RNA level.
In this overview, only the core principles of the most important CRISPR tools were presented. The discussion of the whole diversity of CRISPR applications and especially the constant optimization of existing editors is far beyond the scope of this introduction and I refer the reader to respective reviews (Zhang, 2019; Adli, 2018).

1.4.2.4 *Challenges of CRISPR applications*

Despite the immense progress in the genome editing field, CRISPR-Cas tools still have their challenges and limitations. While a number of aspects, including the *in vivo* delivery, editing efficiencies and DNA repair pathway choice are worth discussion, I will focus on a critical and in the context of this study highly relevant topic, namely gene editing specificity. Early during the CRISPR tool development, it was shown that Cas9 frequently edits off-target sites, i.e. genomic sequences that exhibit high similarity to the actual target site (Fu *et al*, 2013; Hsu *et al*, 2013; Pattanayak *et al*, 2013). Although off-target activities tend to be much lower that the on-target editing rate (Jones *et al*, 2021), this undesired feature represents a major risk, especially with respect to gene therapy applications in patients. Furthermore, even larger genomic rearrangements have been detected upon Cas9 activity, albeit at very low frequencies (Kosicki *et al*, 2018; Frock *et al*, 2015).
To reduce the risk of unwanted Cas9 activity, diverse strategies have been proposed and tested. One solution is the use of Cas orthologues that naturally exhibit a higher target specificity

(Amrani *et al*, 2018, 9; Jones *et al*, 2021). In addition, a variety of engineered Cas9 variants with improved target specificity have been reported (Rees *et al*, 2017; Kleinstiver *et al*, 2016). Unfortunately, the increase in specificity was often accompanied by lower on-target editing rates (Schmid-Burgk *et al*, 2020; Jones *et al*, 2021). Also, the use of paired nickases with two sgRNAs, each enabling the cleavage of one DNA strand was shown to reduce off-targets (Ran *et al*, 2013). The reason is that two different ribonucleoproteins must be active at the same locus for a DSB to occur.

Looking at the problem from a mechanistic angle, one can state that off-target effects are also a result of Cas9 being active for too long. This is due to the fact that editing kinetics at off-target sites tend to be much slower, as compared to on-target sites (Shin *et al*, 2017; Jones *et al*, 2021). Methods that allow the timely inhibition or degradation of Cas9 and thus further contribute to increased editing specificity are highly desired. These approaches are described in the next section.

1.4.2.5 *Inducible CRISPR-Cas gene editors*

The realization that the tight control of Cas9 is crucial for the versatility and safety of the CRISPR method raises the question, how to effectively control the nuclease? As pointed out earlier, the direct control of protein conformation and hence function, rather than the regulation of protein expression, can provide a substantial advance with respect to response time and precision (refer to section 1.3.1). In case of Cas9, both of its components, the protein and the sgRNA are, in principle, amenable for engineering of switchable CRISPR effectors. On the sgRNA-level, the activity of Cas9 has been made inducible via photo-cleavable RNA modifications, that inhibited gene editing in the absence of light (Jain *et al*, 2016; Liu *et al*, 2020) or self-cleaving aptazymes, which react to chemical triggers (Tang *et al*, 2017; Ferry *et al*, 2017). Reversible activation in turn, could be reached by using aptamers, the conformation of which prevented gene targeting in absence of their cognate stimulus (Liu *et al*, 2016b; Kundert *et al*, 2019).

With respect to the control of the protein component, one has to differentiate between the regulation of DNA editing via Cas9-induced DSBs and Cas9 tools based on effector fusions. In the latter scenario, the use of dimerization domains represents an efficient method. Linking a dimerization domain to Cas9 and another one to the effector domain of choice can enable the inducible recruitment of the effector to the targeted locus. Many tools employing this principle have been developed over the years. Blue light-induced transcription control was, for example achieved by using the light-dependent hetero-dimerizer pairs CRY2 and CIB1 or CIBN for the recruitment of transcription activation domains to dCas9 (Polstein *et al*, 2015; Nihongaki *et al*, 2015b). Similarly, protein domains that dimerize upon chemical induction by abscisic acid or gibberellin have been exploited for the control of effector recruitment (Gao *et al*, 2016). The beauty of this approach lies in its simplicity as it can be viewed as a plug-and-play system, the effector domains of which can easily be exchanged. As a major disadvantage, Cas9 remains permanently bound to the DNA, which could cause unwanted side-effects such as the inhibition of transcription.

To regulate the gene editing activity of Cas9, it becomes necessary to engineer the nuclease itself. Although Cas9 is compatible with inducible degradation systems (Kleinjan *et al*, 2017; Senturk *et al*, 2017) and conditional nuclear import/export strategies (Zhao *et al*, 2018), these methods are irreversible and suffer from the dependence on the expression dynamics of new Cas9 protein. Similar to sgRNAs, also Cas9 can be photo-caged by chemically modified amino

acids, enabling its (irreversible) photo-activation (Hemphill *et al*, 2015). For the dynamic control of gene editing, Cas9 variants that can be reversibly activated are required. To this end, split-Cas9 systems were developed, the reconstitution of which is mediated by ligand dependent dimerizers (Nihongaki *et al*, 2019, 1; Zetsche *et al*, 2015b). Fusing the split halves of Cas9 to the rapamycin interacting domains FKBP and FRB enabled the rapamycin-dependent reconstitution of the full protein. Apart from this, also chemically activatable single-chain proteins were designed, for instance by insertion of the ERD as allosteric disrupter into the nuclease (Oakes *et al*, 2016; Davis *et al*, 2015). Also the steric occlusion of the DNA-binding groove of Cas9 via small molecule-induced dimerization domains that were covalently linked to Cas9 was demonstrated (Rose *et al*, 2018).

Optogenetic control over gene editing was achieved in similar ways. Nihongaki et al., for instance, described the design of light-dependent split Cas9 and Cas12 variants using the blue light-activated magnet system (Nihongaki *et al*, 2015a, 2019). A different dimerization approach was established by Richter et al., who inserted the *Rs*LOV domain into Cas9, resulting in an inactive dimer of Cas9-LOV2 hybrids that could be reactivated by exposure to blue light (Richter *et al*, 2016). Unfortunately, when tested in *E. coli*, the system was only active at temperatures around 30 °C. As an optogenetic single-protein approach, the fusion of Cas9 to a pair of pdDronpa domains was constructed (Zhou *et al*, 2017b). pdDronpa is a dimerizing GFP variant, which dissociates upon irradiation with ~500 nm light. In the dimerized state, the position of the pdDronpa domains was shown to inhibit DNA binding by Cas9. Since a single pdDronpa domain comprises 224 amino acids, the size of the resulting fusion protein was, however, rather large.

1.4.2.6 *Anti-CRISPR proteins*

Anti-CRISPR proteins (Acr) represent a diverse class of small, phage-derived proteins, able to inhibit CRISPR-Cas systems (Pawluk *et al*, 2017). They equip phages with the power to counteract the bacterial CRISPR defense system (Bondy-Denomy *et al*, 2013). Acrs were first discovered in 2013 by Bondy-Denomy et al. as inhibitors of type I-F CRISPR systems (Bondy-Denomy *et al*, 2013). Soon, it became clear that Acrs are a wide-spread class of functionally diverse proteins. Acrs that inhibit various types of CRISPR systems from both classes, have since been identified (Bondy-Denomy *et al*, 2013; Pawluk *et al*, 2016). In context of gene editing, inhibitors of several Cas9 orthologues (Rauch *et al*, 2017; Pawluk *et al*, 2016; Harrington *et al*, 2017), but also Cas12- (Marino *et al*, 2018) and Cas13-specific Acrs (Meeske *et al*, 2019) were described. The versatility of Acrs is not only striking from a phylogenetic, but also from a mechanistic perspective. Acrs can, for instance, block sgRNA loading onto Cas nucleases (Thavalingam *et al*, 2019), prevent their DNA binding (Bondy-Denomy *et al*, 2015; Chowdhury *et al*, 2017; Yang & Patel, 2017), inhibit target cleavage (Bondy-Denomy *et al*, 2015; Harrington *et al*, 2017), impair spacer acquisition (Philippe *et al*, 2022) or enzymatically modify Cas proteins (Athukoralage *et al*, 2020; Dong *et al*, 2019).

The application of Acrs in the gene editing field is currently focused on inhibitors of the type II Cas9 nucleases. A straightforward use case of Acrs is the reduction of off-target effects by temporal restriction of Cas9 activity. Here, blocking Cas9 after a period of activity is expected to mainly impair the editing at off-target sites, which are inefficiently targeted and hence take longer to be edited as compared to on-targets. This approach was successfully demonstrated

by temporally separating the delivery of Cas9 RNPs and Acrs (Shin *et al*, 2017), as well as the covalent fusion of attenuated Acrs to Cas9, which only slightly decreased the overall on-target activity (Aschenbrenner *et al*, 2020). Another way to improve DNA targeting specificity, especially with respect to medical applications, is blocking Cas9 activity in cell types or tissues, in which editing is undesired. Towards this goal, Hoffmann et al. placed Acr-encoding transgenes under regulation of cell-type specific microRNAs (miRNAs). Introducing miRNA binding sites into the 3'-UTR of Acr-encoding mRNAs enabled the miRNA-mediated degradation of the Acr transcript and hence released CRISPR-Cas9 activity selectively in the target cell type (expressing the microRNA) as compared to off-target cells (lacking the microRNA) (Hoffmann *et al*, 2019; Lee *et al*, 2019). The use of optogenetically controlled Acrs for the spatio-temporal control of Cas9 is another exciting use case and is described in detail under previous work in section 1.4.2.4.2.

Apart from the regulation of DNA cleavage, Acrs that inhibit DNA-binding are also compatible with many other CRISPR applications. These include the inhibition of Cas9-mediated transcription control (Nakamura *et al*, 2019), the regulation of gene drives (Basgall *et al*, 2018) and the confinement of base editing via Acrs (Liang *et al*, 2020). Besides their use for gene editing control, Acrs were also applied in biosensing gene circuits (Li *et al*, 2018) and as substitutes for antibodies with the aim to capture Cas9 (Johnston *et al*, 2019).

1.4.2.7 *Cas9 orthologues and Acrs used in this study*

1.4.2.7.1 Cas9 orthologues

A large number of Cas9 orthologues have been identified and characterized over the last decade. *Spy*Cas9 is still the most widely applied nuclease when it comes to gene editing, but several disadvantages promoted the use of alternative Cas variants (Table 1.1). First and foremost, the large size of *Spy*Cas9 (1,368 AA) can become a problem, in particular with respect to *in vivo* delivery. Many viral vectors have size restrictions with respect to the DNA that can be packaged into a capsid. The commonly used recombinant adeno-associated viruses (AAVs), for instance, exhibit a packaging capacity of ~4.7 kilobases (kb) (Wu *et al*, 2010). A *Spy*Cas9 expression cassette together with a sgRNA expression module would already exceed this limit. Moreover, many of the tools described above include the fusion of additional effector domains to Cas9, which further increases the construct's size. The aforementioned relatively high off-target rate of *Spy*Cas9 is an additional drawback (Fu *et al*, 2013; Pattanayak *et al*, 2013). Finally, as it is the case for all Cas9 nucleases, the PAM (NGG) represents a restriction with respect to the genomic sites that can be edited, i.e. AT-rich genome regions can hardly be targeted with *Spy*Cas9.

Table 1.1: List of Cas9 orthologues relevant for this study.

Protein	Source	AA seq. length	PAM	Structure
***Spy*Cas9**	*Streptococcus pyogenes*	1,368	NGG	4OO8
***Sau*Cas9**	*Staphylococcus aureus*	1,053	NNGRRT	5AXW
***Nme*Cas9**	*Neisseria meningitidis*	1,082	NNNNGATT	6J9N
***Nme*2Cas9**	*Neisseria meningitidis*	1,082	NNNNCC	6JFU

A different nuclease, relevant to this work is Cas9 from *Staphylococcus aureus* (*Sau*Cas9) (Ran *et al*, 2015; Kleinstiver *et al*, 2015). With a size of only 1053 amino acids, its encoding sequence easily fits into AAVs together with a sgRNA expression cassette, rendering it a good candidate for gene therapy applications (Friedland *et al*, 2015). Furthermore, split-*Sau*Cas9 architectures were successfully developed, enabling the distribution of the editor over two self-complementary AAVs (scAAV) (Schmelas & Grimm, 2018). These virus derivatives allow a faster expression of the cargo transcript, which is accompanied by reduced packaging capacities (only ~2.4 kb) (Schmelas & Grimm, 2018). The main disadvantage of *Sau*Cas9, however, is the longer four nucleotide PAM sequence, NNGRRT, which restricts the targetable DNA sequence space. It should be noted though, that the engineering of the PAM requirements of various Cas9 orthologues made substantial progress over the years, resulting in optimized *Sau*Cas9 mutants with altered PAM specificities (Ma *et al*, 2019).
Another orthologue with promising features is Cas9 from *Neisseria meningitidis* (*Nme*Cas9). In contrast to the previous two candidates, it originates from a type II-C instead of a type II-A

Table 1.2: List of anti-CRISPR proteins relevant for this study. *Cje*Cas9, Cas9 from *Campylobacter jejuni*; *Geo*Cas9, Cas9 from *Geobacillus stearothermophilus*; *Hpa*Cas9, Cas9 from *Haemophilus parainfluenza*; *Boe*Cas9, Cas9 from *Brackiella oedipodis*; *Kla*Cas9, Cas9 from *Kiloniella laminariae*.

Protein	AA seq. length	Targeted orthologues	Mechanism	Structure
AcrIIC1	86	*Nme*Cas9, *Nme*2Cas9, *Cje*Cas9, *Geo*Cas9, *Hpa*Cas9, *Boe*Cas9, *Kla*Cas9, *Sau*Cas9	Inhibition of DNA cleavage	5VGB
AcrIIC3	117	*Nme*Cas9, *Nme*2Cas9	Inhibition of DNA binding, dimerization	6JHW, 6JE9

CRISPR system (Esvelt *et al*, 2013; Hou *et al*, 2013). Alike *Sau*Cas9, *Nme*Cas9 is relatively compact nuclease (1,082 AA) and requires a four nucleotide PAM (NNNNGATT) (Esvelt *et al*, 2013; Hou *et al*, 2013). Interestingly, it also possesses the ability to bind and cleave single-stranded DNA and RNA (Rousseau *et al*, 2018; Ma *et al*, 2015a). Moreover, *Nme*Cas9 is a natural high-fidelity orthologue with almost no measurable off-target effects (Lee *et al*, 2016; Amrani *et al*, 2018). Finally, additional *Nme*Cas9 variants have been identified, which possess high sequence similarity (>86 %) to the "original" nuclease (Edraki *et al*, 2018). One of them, *Nme*2Cas9, is of identical size (1,082 AA) and exhibits an equally high target specificity, while requiring a more compact C-rich PAM sequence (NNNNCC). This discovery significantly advanced the genomic target range of the *N. meningitidis* nuclease family. Finally, the applicability and high specificity of both *Nme*Cas9 orthologues was also demonstrated *in vivo* (Ibraheim *et al*, 2018, 2021).

1.4.2.7.2 Acrs

This study focuses on the anti-CRISPR proteins AcrIIC1 and AcrIIC3. As the names already imply both inhibit type II-C Cas9 enzymes (Table 1.2). They are small proteins (86 and 117 amino acids, respectively) that share no sequence or structure similarity (Harrington *et al*, 2017; Pawluk *et al*, 2016). While several other AcrIIC inhibitors are known (Lee *et al*, 2018, 5), these two are the best characterized members of the family. As most striking difference, AcrIIC3 only inhibits

*Nme*Cas9 and the closely related orthologue from *Haemophilus parainfluenza*, while AcrIIC1 was the first described broad-spectrum inhibitor, being able to deactivate many Cas9 orthologues from type II-C systems (Table 1.2) (Harrington *et al*, 2017; Garcia *et al*, 2019, 5). It was even shown to have modest activity on the type II-A *Sau*Cas9 orthologue (Mathony *et al*, 2020a; Garcia *et al*, 2019, 5). Importantly, AcrIIC1 is a relatively weak inhibitor of Cas9, in stark contrast to the very strong activity of AcrIIC3 (Harrington *et al*, 2017). A reason for this notable difference may lie in the respective mechanisms of action. AcrIIC1 binds to the catalytic HNH domain, thus abolishing DNA cleavage, while still allowing Cas9 to bind to target DNA (Harrington *et al*, 2017). AcrIIC3 also binds the HNH domain, but at a different surface site, which might be a reason for the higher specificity towards *Nme*Cas9 (Kim *et al*, 2019; Zhu *et al*, 2019). In addition, AcrIIC3 further contacts the Rec lobe of Cas9 resulting in a dimerization of the nuclease (Sun *et al*, 2019) and inhibition of DNA binding (Harrington *et al*, 2017).

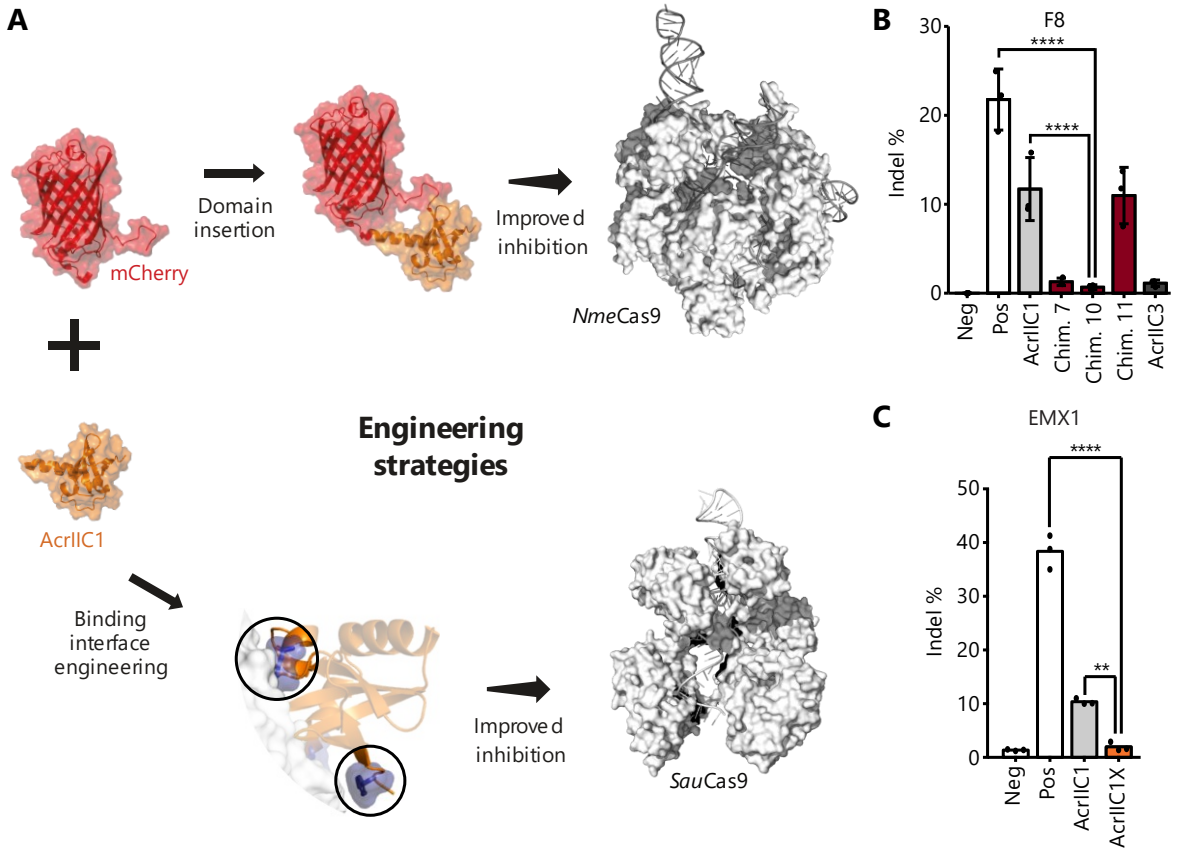


Figure 1.8: Previous work – improving the inhibition potency of AcrIIC1. (A) Insertion of mCherry into AcrIIC1 results in increased inhibition potency of *Nme*Cas9. The binding interface of AcrIIC1 was redesigned to bind to and inhibit the HNH-domain of *Sau*Cas9. (B) Analysis of gene editing efficiencies by T7 endonuclease assay. HEK293T cells were co-transfected with plasmids encoding *Nme*Cas9, a sgRNA targeting the endogenous F8 locus and the indicated Acr variants. The Cas9:Acr DNA ratio was 1:1. I acquired the data prior to the start of the Ph.D. Chimera 11 is a control construct carrying a PDZ domain insertion instead of mCherry. Chim., Chimera. (C) Analysis of the inhibition potency of AcrIIC1X on *Sau*Cas9. HEK293T cells were co-transfected with constructs encoding SauCas9, a sgRNA against the EMX1 locus and the indicated Acr variant. Gene editing efficiencies were assessed by TIDE sequencing. The experiments were performed by Sabine Aschenbrenner and Carolin Schmelas. (B, C) Individual data points from n=3 independent biological replicates are shown. Bars represent the mean and error bars represent the standard deviation (SD). Neg, negative control (Cas9 + non-targeting sgRNA). Pos, positive control (Cas9 + sgRNA). **P < 0.01, ****P < 0.0001 by one-way analysis of variance (ANOVA) with Bonferroni correction.

1.4.2.8 *Prior work on Anti-CRISPR proteins*

The work presented in this section stems from the time of my Master thesis, at which the Acr projects were initiated. It laid the foundation for the characterization and application of the Acr variants that I worked on during my subsequent Ph.D and which is described in chapter 2.1 of this thesis.

1.4.2.8.1 Improving the inhibition strength of Acrs

As stated in the last section, AcrIIC1 is of interest due to its high promiscuity, thus providing a possible "one fits many" solution for the control of gene editing tools. To be useful as molecular tools though, Acrs needs to show high levels of activity, which is unfortunately not the case for AcrIIC1. In order to increase the inhibition potency, we chose two different strategies.

First, by inserting the fluorescent protein mCherry into an unstructured loop around Y70 of AcrIIC1, we increased its stability and in consequence its inhibitory effect on *Nme*Cas9 (Fig. 1.8A, B). Testing different insertion variants, we identified two lead candidates, chimera 7 and 10 (Mathony *et al*, 2020a). The former variant carries mCherry between E68 and Y72 and chimera 10 bears the insert behind Y70, flanked by GSG linkers. While it was hardly possible to fully inhibit *Nme*Cas9 using the wildtype inhibitor, the engineered chimeric versions (chimera 7 and chimera 10) drastically reduced DNA cleavage (Fig. 1.8B).

Having shown that it is possible to increase the inhibition potency on AcrIIC1's natural target, *Nme*Cas9, we next wanted to effectively inhibit *Sau*Cas9 activity, an orthologue, only mildly affected by AcrIIC1 (Mathony *et al*, 2020a). At the time the project was initiated, no effective inhibitor of this orthologue was known. We reasoned that the lack of inhibition must be caused by a low binding affinity of AcrIIC1 to the HNH domain of *Sau*Cas9. In collaboration with the group of Bruno Correia, we hence redesigned the binding interface of AcrIIC1 to optimize it's affinity to *Sau*Cas9. The result was an AcrIIC1 triple mutant, N3F/D15Q/A48I, which was termed AcrIIC1X. Experiments performed by Carolin Schmelas and Sabine Aschenbrenner demonstrated that AcrIIC1X strongly outperformed wildtype AcrIIC1 with respect to *Sau*Cas9 inhibition in human cells (Fig. 1.8C).

1.4.2.8.2 Engineering of light-switchable Acrs

In the chapter 1.4.2.5, I pointed out, how important the control of Cas9 activity is for many applications. A general problem of most published tools is that they are specific to certain applications and are not necessarily compatible with the diverse Cas9 fusion proteins that are commonly used. An approach published by Bubeck et al. (Bubeck *et al*, 2018) from our group provided an elegant solution to this problem. Instead of engineering Cas9 itself, they inserted the *As*LOV2 domain into AcrIIA4, a protein inhibitor of *Spy*Cas9. The hybrid inhibitor was active in the dark, while the blue light-triggered conformational change of the LOV2 domain resulted in an inactive conformation of the fused Acr. Once the inhibitor was turned off, Cas9 became active and gene editing could progress (Fig. 1.9A). This approach was termed CASANOVA ("CRISPR–Cas9 activity switching via a novel optogenetic variant of AcrIIA4"). The advantage of the method lies in the fact that it can easily be combined with all sorts of CRISPR-derived tools, such as Cas9-based transcription activators or epigenetic modifiers without the need for re-engineering of the nuclease.

Similar to most other inducible CRISPR systems, CASANOVA is specific to a single Cas orthologue, namely *Spy*Cas9. The gene editing field, however, has moved far beyond the use

of only this one orthologue (refer to section 1.4.2.7.1). In order to evaluate to what extend the CASANOVA approach can be adapted to other Acrs, we set out to develop a light switchable variant of the *Nme*Cas9 inhibitor AcrIIC3. Following the screening of several insertion sites in this Acr (Hoffmann *et al*, 2021), we identified two variants that enabled the effective light control of gene editing (Fig. 1.9B). The respective variants carried *As*LOV2 domain insertions behind F59 of AcrIIC3. One derivative was the naïve insertion of the domain at this site, the other variant was flanked by single G's as linkers. The AcrIIC3-LOV2 hybrids were named CASANOVA-C3 (CN-C3) and CASANOVA-C3G (CN-C3G), respectively. When co-expressed with *Nme*Cas9 and a sgRNA into human cells, CN-C3 and CN-C3G of them efficiently inhibited gene editing in the dark, while allowing *Nme*Cas9 activity upon illumination, albeit at slightly lower efficiency (Fig. 1.9B). These results set the basis for the further characterization that is described in the results section.

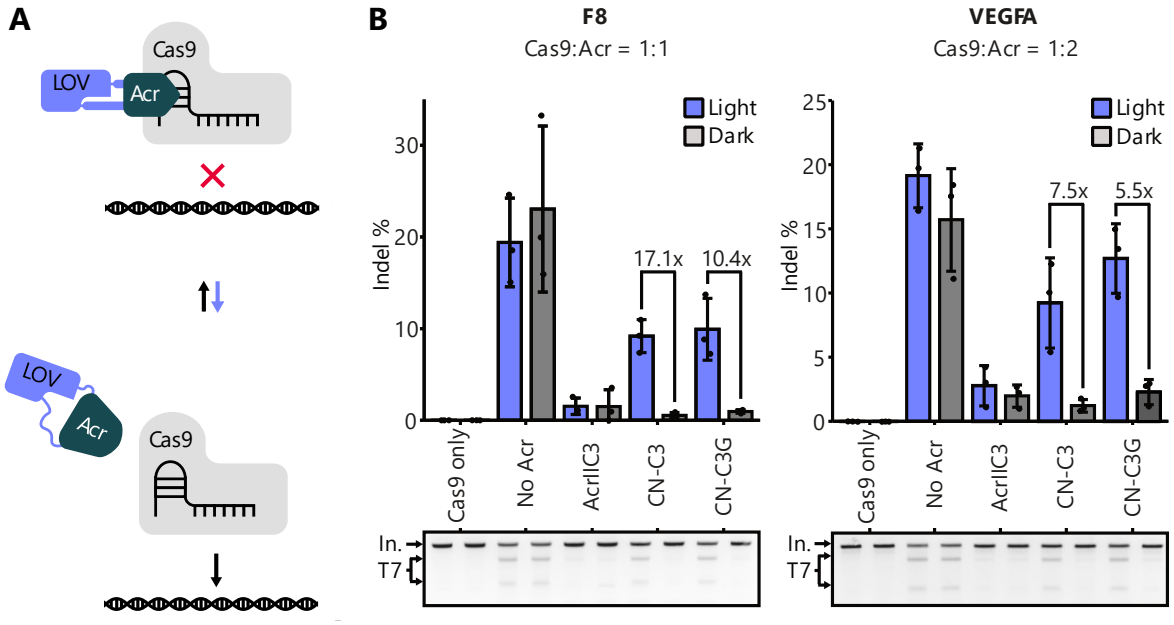


Figure 1.9: Previous work on CASANOVA-C3. (A) Working principle of CN-C3. (B) T7 endonuclease assay assessment of optogenetic control over gene editing by CN-C3. HEK293T cells were transfected with constructs expressing *Nme*Cas9, a sgRNA targeting the indicated loci and the respective Acr variants and incubated under blue-light irradiation or in the dark. Cas9:Acr vector mass ratios are shown above the panels. Individual data points from n=3 independent biological replicates are shown. Bars represent the mean and error bars represent the SD. A representative agarose gel image is shown. Fold changes are indicated for CN-C3 variants. I performed the experiments during my Master thesis. Fold changes in activity are indicated above the bars.

1.5 Aim of study

The control of protein activity via exogenous signals, such as light or chemicals, is widely used in basic research and biotechnology. It can be employed to restrict protein activity in time and space, hence enabling unprecedented direct control over cellular processes. The engineering of switchable proteins via the insertion of domains responsive to light or chemicals represents an elegant and frequently used strategy to create switchable proteins. At the same time, protein engineering by domain insertion is still hampered by an incomplete mechanistic understanding of proteins, so that extensive experimental optimization is often required.

In the first part of this study, we harnessed protein engineering to improve the control of compact and high-fidelity gene editing tools. We created Acrs with enhanced inhibition potency on the RNA-guided nucleases *Nme*Cas9 and *Sau*Cas9, as well as optogenetically switchable inhibitors for *Nme*Cas9. While strategies to control and safeguard CRISPR nucleases were previously focused on other Cas9 orthologues, our aim was to provide a versatile toolkit that enables the precise regulation of these two promising nucleases.
Focusing on the CRISPR inhibitor AcrIIC1, we first sought to improve its inhibition potency on *Nme*Cas9 and *Sau*Cas9, in order to guarantee a complete deactivation of nuclease activity. On top, we aimed to demonstrate the value of the engineered Acr by establishing a method to specifically activate *Sau*Cas9 only in selected cell types.
Next, using an exceptionally potent inhibitor of *Nme*Cas9, AcrIIC3, we planned to create an optogenetic CRISPR switch by insertion of the *As*LOV2 domain into this inhibitor. The photoreceptor insertion was supposed to enable the light-mediated deactivation of the Acr and in consequence activation of Cas9. The resulting tool would allow users to perform gene editing in a spatiotemporally defined manner.
Our experience with optogenetic Acrs, as well as previously published work showcased persisting challenges with respect to the design of switchable proteins. The identification of suitable insertion sites and the subsequent optimization of lead candidates are still cumbersome processes, depending on extensive trial and error.
In the second part of this study, we thus sought to dissect constraints underlying successful domain insertions in an unbiased fashion. To this end, we intended to perform comprehensive domain insertion screens with diverse insert and effector proteins. We selected four target protein candidates of great structural and functional diversity, including the transcription factor AraC, the Flp recombinase, the protease of the tobacco vein mottling virus (TVMV) and the Sigma factor F from *Bacillus subtilis (B. subtilis)*. In an unbiased fashion, we then inserted one or several different candidate domains into these proteins followed by FACS-mediated enrichment of active variants and next generation sequencing (Flow-seq). Thereby, we aimed at creating comprehensive datasets enabling us to dissect biophysical and evolutionary constraints of domain insertion tolerance.
Finally, we aimed at extending our Flow-seq approach towards the identification of switchable protein variants, thus providing a powerful strategy to engineer proteins to control selected cellular functions.

2 Results

The following chapter is subdivided into three parts. In the first part, I am going to describe the characterization and application of engineered Acrs that show enhanced or light-switchable Cas9 inhibition. An overview of experiments that were performed on these projects prior to my Ph.D. is provided in the introduction (section 1.4.2.8). All experiments described in the results chapter below, were, however, obtained during my Ph.D. and are reported in two peer-reviewed articles (Mathony et al., 2020 and Hoffmann, Mathony et al., 2021).
The second part describes an unbiased domain insertion screen as well as the corresponding bioinformatic analysis. Lastly, in the third part, I am going to present the characterization of two light-switchable variants of the transcription factor AraC that were identified during the screen.

2.1 Characterization of enhanced and light-switchable Cas9 inhibitors

2.1.1 Performance of improved Acrs on different *Nme*Cas9 orthologues

During my Master's thesis, I started the characterization of different domain fusions between the broad-spectrum Cas9 inhibitor AcrIIC1 and the fluorescent protein mCherry resulting in two lead candidates, chimeras 7 and 10 (refer to section 1.4.2.8.1). Since our previous analysis was limited to T7E assays, I first established a Tracing of indel by decomposition (TIDE)-based readout to measure Cas9 activity in the Acr application context (Brinkman *et al*, 2014). In case of high editing rates, TIDE sequencing is considered to give a qualitatively better representation of indel percentages than T7E assays do (Brinkman *et al*, 2014; Sentmanat *et al*, 2018). TIDE is based on Sanger sequencing of PCR products amplified from the edited locus. DNA repair at the target site via NHEJ results in the occurrence of errors, which in turn lead to a diverse set of indels in the targeted cell populations. In Sanger sequencing chromatograms, the varying length of the indels results in superposed sequence shifts within the chromatogram starting at the Cas9 cut site. Decomposition of the individual traces of the chromatogram allows the efficient assessment of indel frequencies. In order to evaluate the inhibition potency of the chimeras by TIDE, I targeted two endogenous loci by *Nme*Cas9 either in presence or absence of the respective engineered, chimeric inhibitors (Ch. 7, Ch. 10; see Fig. 1.8A, B) or wild-type AcrIIC1 and -C3 as controls (see below). Three days post transient transfection, the cells were lysed and indel frequencies were assessed (Fig. 2.1A) Indeed, the results exhibit higher editing efficiencies for the control with active Cas9 as compared to prior T7E-assays (Fig. 1.8B). At the same time, background signal of up to 5 % was detected, even in the non-targeting control. This phenomenon is caused by background signal in the sequencing read. Wildtype AcrIIC1 only mildly decreased the editing efficiency of *Nme*Cas9. The chimeras 7 and 10, instead drastically reduced editing to the background level (Fig. 2.1A). AcrIIC3 was included into the experiments as a positive control. In contrast to AcrIIC1, this inhibitor is known to be highlyefficient, while being also very specific, only inhibiting *Nme*Cas9 (Harrington *et al*, 2017;

Garcia *et al*, 2019, 5). The inhibition levels of the engineered chimeras and AcrIIC3 were comparable, indicating high inhibition potency of the AcrIIC1-mCherry chimeras as compared to wild-type AcrIIC1.

Shortly after the initial characterization of *Nme*Cas9 as a compact high-fidelity genome editor (Amrani *et al*, 2018; Lee *et al*, 2016), the closely-related *Nme*2Cas9 orthologue distinguished by a more compact PAM sequence (N_4CC instead of N_4GATT for *Nme*Cas9) was identified (Edraki *et al*, 2018). A detailed description of *Nme*2Cas9 and comparison with *Nme*Cas9 can be found in the introduction (section 1.4.2.7.1). To test, if our engineered inhibitors work comparably well on the *Nme*2Cas9 orthologue, I performed a similar assay as before. This time, I targeted the genomic loci VEGFA and FANCJ, since potent sgRNAs were reported for these loci (Edraki *et al*, 2018). As *Nme*2Cas9 exhibited a lower overall editing efficiency as compared to *Nme*Cas9, I assessed indel frequencies by T7E assay, which tends to be more sensitive in my experience. The results resembled very much the outcomes for *Nme*Cas9. While AcrIIC1

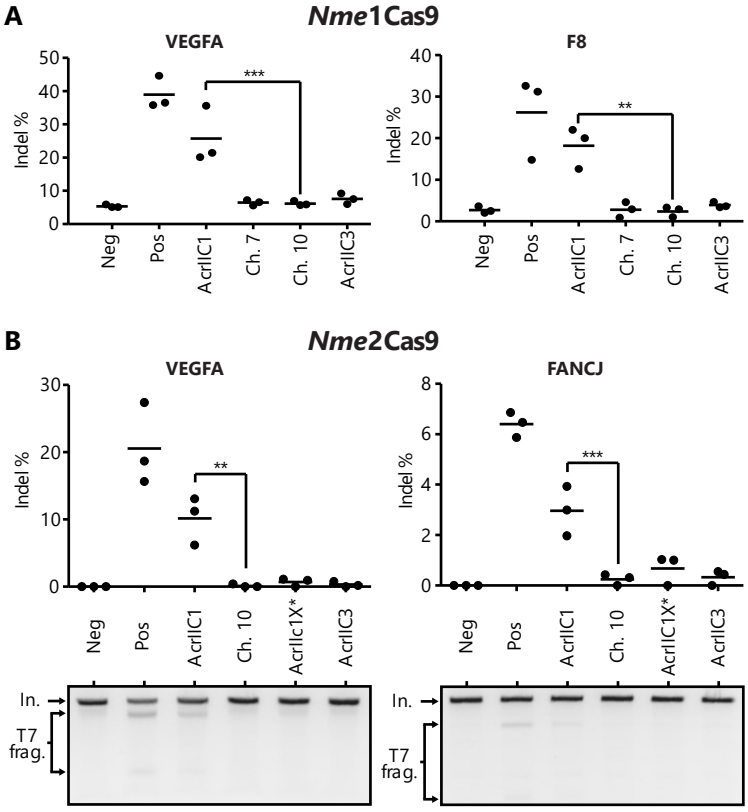


Figure 2.1: Performance of engineered AcrIIC1 variants on different *Nme*Cas9 orthologues. (A) Engineered Acr variants show improved inhibition potency. HEK293T cells were co-transfected with plasmids encoding *Nme*Cas9, a sgRNA targeting the VEGFA or F8 locus and the indicated Acr variants. The Cas9:Acr DNA mass ratio was 4:1. The performance of the different Acrs on *Nme*1Cas9 was assessed by TIDE sequencing. (B) engineered Acrs also exhibit enhanced inhibition of *Nme*2Cas9. HEK293T cells were co-transfected with plasmids expressing *Nme*2Cas9, a sgRNA targeting the indicated locus as well as the respective Acr variant with a Cas9:Acr plasmid mass ratio of 1:1. 72 h post-transfection, indel formation was assessed by T7E assay. Representative gel images are shown. (A,B) data points represent three independent biological replicates. Horizontal lines are the mean. Neg, negative control (Cas9 + non-targeting sgRNA). Pos, positive control (Cas9 + sgRNA). **P < 0.01, ***P < 0.001, calculated by one-way ANOVA with Bonferroni correction.

inhibited *Nme*2Cas9 only partially, the engineered AcrIIC1-mCherry chimeras, as well as wild-type AcrIIC3 completely blocked gene editing (Fig. 2.1B).
As a side note, this experiment included one additional Acr variant named AcrIIC1X*. This protein is a derivative of chimera 10 which carries three point mutations in the Cas9 binding surface of the Acr that increase its inhibition potency on *Sau*Cas9 (see chapter 2.1.2 and (Mathony *et al*, 2020a)). Interestingly, upon introduction of the three mutations, the capability of the chimeric Acr to efficiently inhibit *Nme*2Cas9 was still maintained.

2.1.2 Enabling cell type specific gene editing with AcrX

The chimeric Acrs described in the last section showed improved inhibition of *Nme*Cas9. AcrIIC1X, on the other hand, was engineered by us to very potently inhibit *Sau*Cas9 (see previous work in section 1.4.2.8). To demonstrate its applicability, I employed AcrIIC1X for cell type-specific activation of *Sau*Cas9. To this end, I built on a concept that had originally been developed by Mareike Hoffmann, a former Ph.D. student of our lab (Hoffmann *et al*, 2019). The strategy makes use of the fact that many miRNAs are expressed in humans in a highly tissue- and cell-type-specific manner (Landgraf *et al*, 2007; Patil *et al*, 2022). In order to restrict gene editing to the cell type or tissue of choice, Acr transgenes are co-delivered with Cas9/sgRNA. The Acr transgenes carry miRNA binding sites within the 3'-untranslated region (UTR) (Fig. 2.2). In cell types strongly expressing the corresponding miRNA, the Acr-encoding mRNA is hence degraded or its translation is inhibited by RNAi. This, in turn, prevents Acr expression and consequently enables gene editing by Cas9. In any off-target cell type not expressing the cognate miRNA, however, the Acr-encoding mRNA remains stable, therefore resulting Acr expression and hence in inhibition of genome editing. This strategy provides an additional safety layer for gene editing technologies. To prove the feasibility of this approach for the control of *Sau*Cas9 activity, I inserted binding sites for the hepatocyte-specific miRNA-122 (Lagos-Quintana *et al*, 2002) into the 3'-UTR of the AcrIIC1 as well as AcrIIC1X transgenes. As controls, versions of the same vectors with UTRs of the identical length, which carried a scaffold sequence instead of the miRNA binding site, were used. First, I assessed the system in HEK293T cells that do not express miRNA-122. To this end, cells were transfected with plasmids encoding *Sau*Cas9, a sgRNA targeting the endogenous EMX1 locus and the different Acr variants. On top, one batch of samples was co-

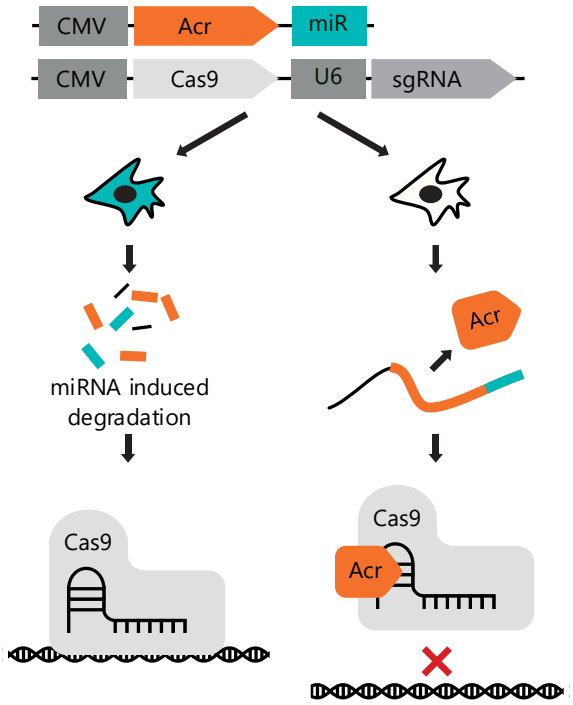


Figure 2.2: Enabling cell type specificity with help of miRNA-controlled Acrs. The Acr is co-expressed with Cas9 and a sgRNA. The mRNA encoding the inhibitor carries miRNA-binding sites in its 3'-UTR. When delivered to a cell type that expresses the cognate miRNA, the mRNA gets degraded and Cas9 becomes active. In off-target cell types, the mRNA remains stable. Consequently, the Acr is translated and inhibits CRISPR gene editing.

transfected with a vector expressing miRNA-122, while miRNA-155 was used as a control in a second set of samples. Finally, controls that did not overexpress either miRNA were included. The results were analyzed by T7E-assay (Fig. 2.3A), as well as by TIDE sequencing (Fig. 2.3B). As expected, the Cas9 control without Acr exhibited strong gene editing as indicated by high indel frequencies independent of the presence of a miRNA. Wildtype AcrIIC1 generally resulted in a noticeable, but incomplete Cas9 inhibition, as already observed in the previous experiments. TIDE sequencing revealed a significant release in genome editing efficiency when AcrIIC1-miRNA-122 was co-expressed with miRNA-122, but not upon co-expression with the control miRNA-155 (Fig. 2.3B). This effect was, however, not visible in the T7E-assay, possibly due to saturation of the measurable indel rate (Fig. 2.3A). Overall, Cas9 inhibition by AcrIIC1 tended to be slightly (albeit not significantly) weaker, when expressed from a transcript with miRNA-

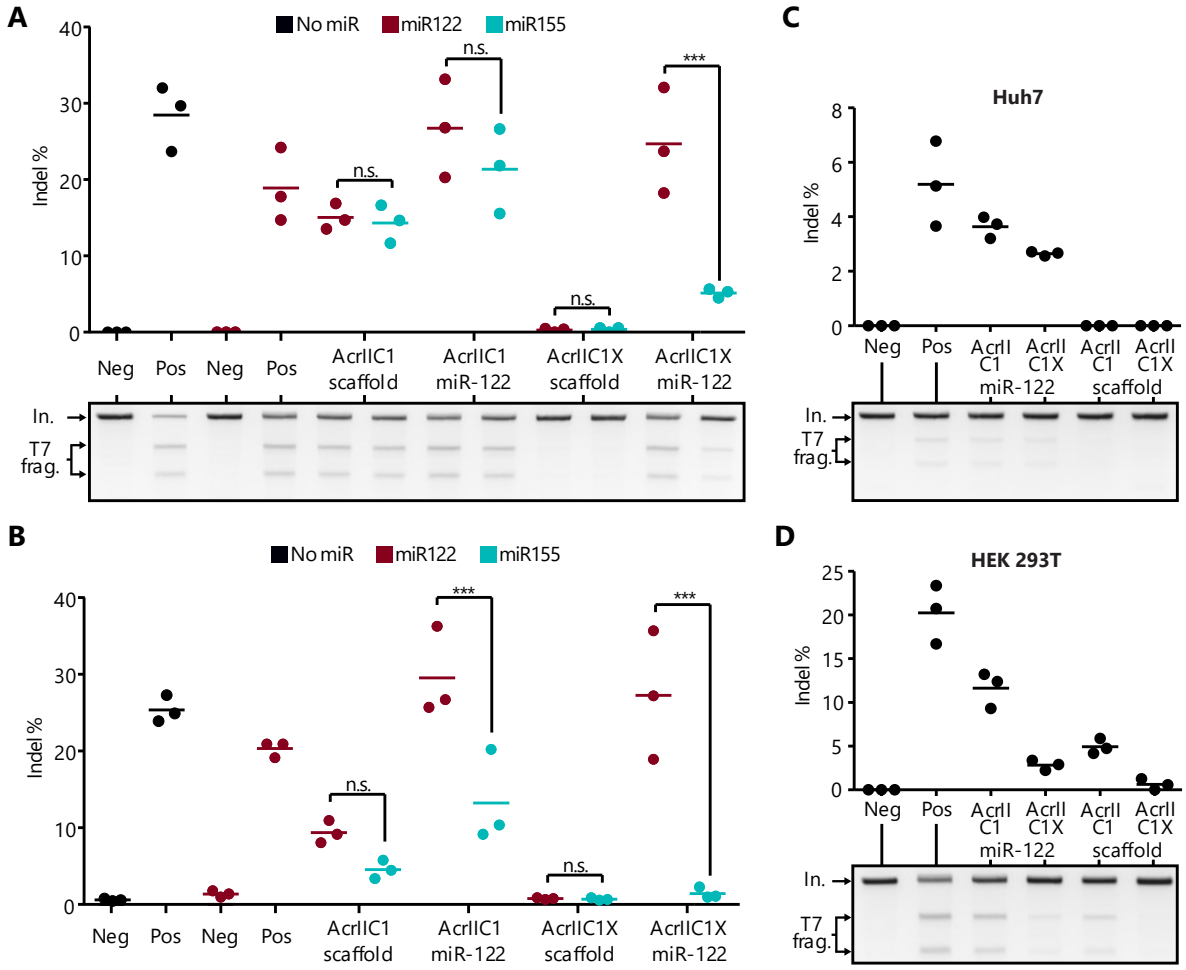


Figure 2.3: Hepatocyte-specific gene editing enabled by miRNA-controlled AcrIIC1X. (A, B) HEK293T cells were transfected with plasmids encoding *Sau*Cas9, a sgRNA targeting EMX1 and Acr derivatives carrying a miR-122-binding site, a miR-155-binding site or a scaffold of the same length in their 3'-UTR. In addition, a plasmid expressing the indicated miRNAs was co-supplied. Three days post transfection, indel formation was assessed by T7E assay (A) or TIDE sequencing (B). (C, D) HEK293T cells (C) and Huh7 (B) cells were transduced with AAVs encoding Cas9, a sgRNA targeting the EMX1 locus and the Acr transgenes with or without miR-122 binding sites. The multiplicities of infection (MOI) used during transfection were 10^5 and $5x10^4$ for *Sau*Cas9 and the Acrs, respectively. Gene editing was evaluated after three days by T7E assay. (A-D) Data points represent three independent biological replicates. Horizontal lines are the mean. Representative gel images of the T7 assays are shown below the graphs. Neg, negative control (Cas9 + non-targeting sgRNA). Pos, positive control (Cas9 + targeting sgRNA). n.s., not significant; ***P < 0.001 as calculated by one-way ANOVA with Bonferroni correction.

122-binding sites as compared to the scaffold controls, independent of the co-transfected miRNA. This might be due to some mild, basic expression of miRNA-122 in HEK293T cells or differences in the Acr's mRNA expression or stability due to the differences in the 3'UTRs. Irrespective of this mild difference, the inhibition of gene editing by AcrIIC1 was rather weak in all samples.
AcrIIC1X, in contrast, effectively blocked gene editing, when used in the scaffold configuration (Fig. 2.3A, B). The AcrIIC1X-miR-122 samples, in turn, showed potent Cas9 inhibition in absence of miRNA-122, but exhibited a drastic increase in the indel percentage upon miRNA-122 co-expression. Comparing the T7E assay and TIDE data, both readouts reveal the same qualitative trends and slightly differ with respect to the absolute indel values (Fig. 2.3A, B).
To take the concept further, Cas9, sgRNA and the aforementioned Acr transgenes (with and without miRNA-122 binding sites) were packaged into AAVs (serotype 2) and the efficiency of the system was tested in Huh7 cells, a hepatocyte cell line, which endogenously expresses miRNA-122. The respective AAVs were produced and purified by my colleague Dr. Carolin Schmelas and I performed the subsequent genome editing experiments. As the gene editing rates turned out to be rather low in Huh7 cells, likely due to inefficient viral transduction, the indel formation was assessed by T7E assay only. Nevertheless, co-transducing AAV encoding Cas9, a sgRNA and AcrIIC1 or AcrIIC1X resulted in enhanced genome editing for the miR-122-dependent Acr variants as compared to the control variants (scaffold; Fig. 2.3C). This indicates that miRNA-dependent genome editing can even be achieved with endogenous miRNAs.
Beyond the release of Cas9 activity within target cells (here: hepatocytes), the ability of the system to prevent editing in off-target cells is of equal importance. For this reason, the same experiment was performed in HEK293T cells, which served as an off-target cell line. As seen before upon plasmid transfection, AcrIIC1 showed a much weaker inhibition of gene editing as compared to AcrIIC1X when delivered by AAVs (Fig. 2.3D). The engineered inhibitor, in turn, efficiently blocked Cas9 activity. Again, the Acr variant carrying the miRNA-122 binding sites resulted in low, albeit detectable rates of editing in HEK293T cells. In sum, the presented data demonstrate the applicability of miRNA-regulated gene editing control concept in combination with *Sau*Cas9 and further shows that potent inhibitors, such as AcrIIC1X, are required to efficiently inhibit genome editing in off-target cell types. The results of this section together with additional experiments were published under the title "computational design of anti-CRISPR proteins with improved inhibition potency" (Mathony *et al*, 2020a).

2.1.3 Characterization of CASANOVA-C3

2.1.3.1 *Optogenetic control of gene editing at on- and off-target loci*

Apart from inhibition potency, described in the last section, reversibility is another key factor, when it comes to the effective control of gene editing. To this end, we designed CASANOVA-C3 (CN-C3), an optogenetically switchable *Nme*Cas9 inhibitor. As stated before, the initial identification and early characterization was performed by Mareike Hoffmann and me prior to this thesis and the datasets shown here are new experiments that I performed during my Ph.D. The previous work is outlined in section 1.4.2.8.2. Two variants of this inhibitor exist, both of which are based on AcrIIC3 carrying a *As*LOV2 domain inserted behind residue F59. While CN-C3 has no linker residues between the Acr and the LOV2 domain, the second variant, CN-C3G,

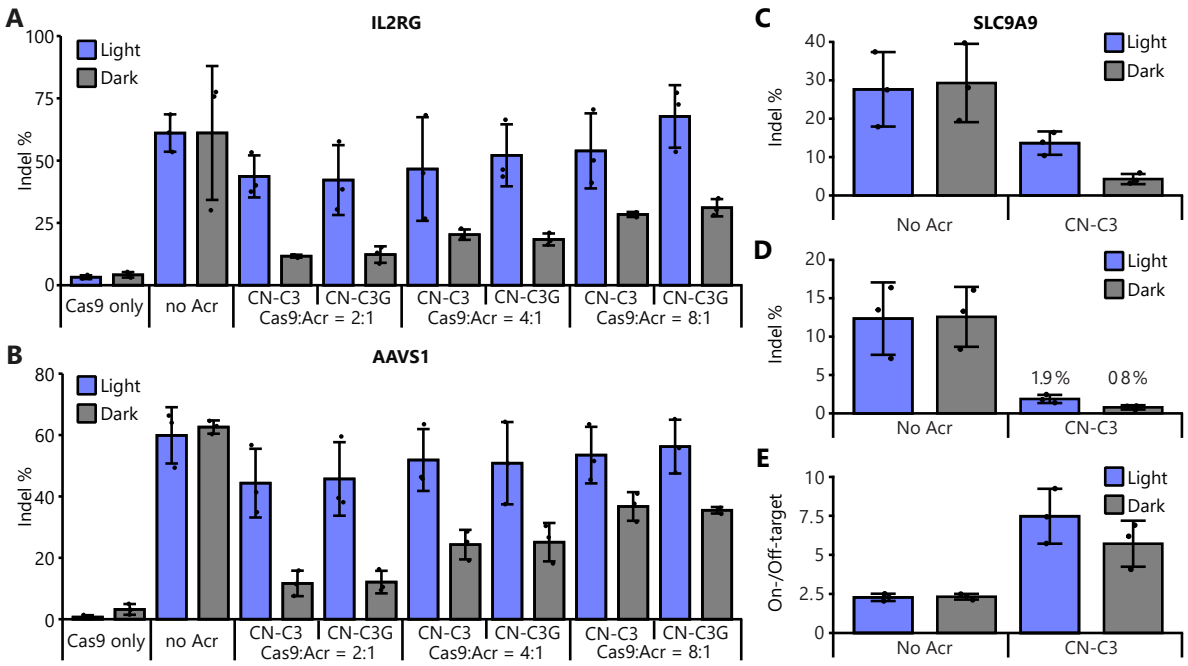


Figure 2.4: Optogenetic control of gene editing by CN-C3. (A-D) Cells were co-transfected with plasmids expressing *Nme*Cas9, a sgRNA targeting the annotated locus and CN-C3 or CN-C3G in the indicated vector mass ratios, followed by blue-light exposure or incubation in darkness. Indel formation was assessed after 72 h by TIDE sequencing (A, B) or targeted amplicon sequencing of the selected locus (C) or of a previously described off-target site (chr16:+:30756950:30756980) (D). (E) the ratio between on- and off-target gene editing from panels C and D is shown. (A-E) Data points represent n=3 individual biological replicates. Bars indicate the mean and error bars the SD.

carries single glycines as linkers. These hybrids were the lead candidates identified in a larger screen of diverse Acr-LOV2 combinations.

Following the initial characterization, I tested the properties of the AcrIIC3-LOV2 hybrids by co-transfecting different Cas9 to Acr vector mass ratios into HEK293T cells. The experiment was motivated by the fact that inhibition by Acrs is usually not 100 %. Consequently, it is important to assess, which ratio of nuclease and inhibitor results in minimal background (genome editing in the dark) as well as an efficient photo-switching of the system. On these grounds, I tested various vector mass ratios, ranging from 1:1 to an eight-fold excess of the *Nme*Cas9 plasmid on different genomic loci (Fig. 2.4A, B and Supp. Fig. 1). During this and the following experiments, two replicates of each sample were prepared, one of which was incubated in the dark, while the other one was illuminated by blue light at a wavelength of ~460 nm and an intensity of ~3 W/m^2. Editing rates of around 60 % were observed in the positive control, i.e. in absence of the inhibitor, irrespective of the light condition, as expected (Fig. 2.4A, B). The CN-C3(G) samples in turn, exhibited an up to 3-fold light switch in indel frequencies between the light and dark conditions. Comparing CN-C3 to CN-C3G, no significant differences could be found. Application of CN-C3(G) resulted in slightly decreased editing rates of about 45 % at the highest Acr concentration under light exposure, as compared to 55-60 % in the Cas9 control without inhibitor (Fig. 2.4A, B). At higher Cas9:Acr ratios, indel frequencies increased up to the positive control levels. With respect to gene editing in the darkness, leakiness of around 10 % was observed for both CN-C3 variants. As expected, background editing in the dark increased at lower Acr vector doses.

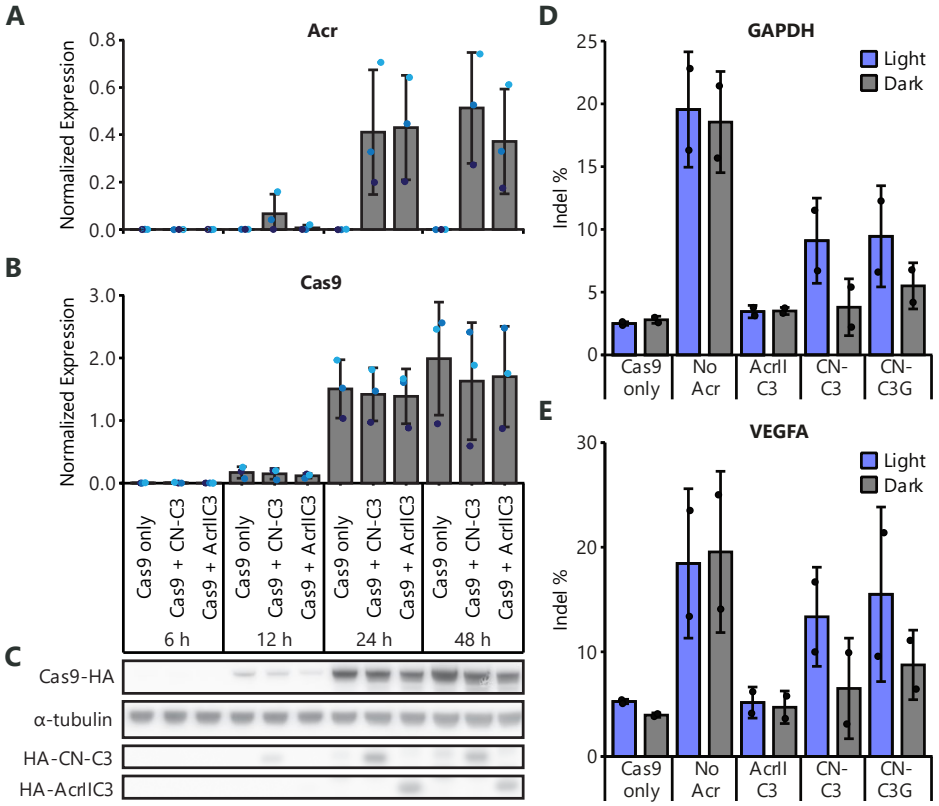


Figure 2.5: Optogenetic control is not mediated by changes in Acr stability and CN-C3 is generally compatible with *Nme*2Cas9. (A-C) HEK293T cells were co-transfected with vectors expressing *Nme*Cas9, a non-targeting sgRNA and the indicated Acr variant using a Cas9:Acr vector mass ratio of 1:1. Protein expression was assessed by Western blot and relative expression levels of the Acrs (A) and Cas9 (B) were calculated by normalizing the respective band intensities to the α-tubulin control. Representative gel images are shown (C). (D, E) HEK293T cells were transfected with constructs expressing *Nme*2Cas9, a sgRNA targeting the respective locus and the indicated Acr variant, followed by incubation under blue-light exposure or in darkness. The formation of indels was assessed by TIDE sequencing. (A, B, D, E) Points represent n=3 (A, B) or N=2 (D, E) biological replicates. Bars indicate the mean and error bars the SD. (A, B). The different colors of the data points represent the individual replicates.

Another key aspect of gene editing control by Acrs is the possibility to reduce off-target effects. It is known, that off-target editing is a matter of kinetics. A detailed description of this problem and its solutions is given in section 1.4.2.4. In short, it has previously been shown that the timed inhibition of Cas9 by Acrs can decrease the off-target editing rate, while only marginally affecting on-target activity of Cas9 (Shin *et al*, 2017). Practically, this has been achieved, by double transfection of cells, first with Cas9 and its sgRNA, then six hours later with the Acr. Unfortunately, this procedure is impractical especially with respect to potential applications outside the tissue culture. As CN-C3, even in its light-activated state, shows some degree of inhibition, we reasoned that this effect might be sufficient to decrease off-target gene editing, while preserving sufficiently high on-target activity. The advantage of this approach would be its applicability without the need for double transfections. One should note however, that *Nme*Cas9, the target of CN-C3 is already a Cas9 orthologue with high fidelity. In fact, only a single sgRNA (targeting SLC9A9) was available by the time, which caused significant off-target editing in my experimental setup (Fig. 2.4D). When I tested *Nme*Cas9 on this locus, indel rates around 25 % percent were observed for active Cas9 without the inhibitor (Fig. 2.4C). At the off-

target locus, substantial levels of undesired editing of ~12% were detected (Fig. 2.4D). Co-transfection of CN-C3 reduced the editing levels to 12 % and 1.9 % under light exposure at the on- and off-target site, respectively. In darkness, substantially lower editing rates of only ~5 % and 0.8 % were observed (Fig. 2.4C, D). In order to achieve the most accurate representation even at low residual editing rates, the reported indel percentages were measured by NGS instead of TIDE sequencing. In the absence of CN-C3, on-target editing exceeded the off-target effect by 2.5-fold (Fig. 2.4E). Application of CN-C3 increased this ratio to 6-7.5-fold. This improvement of specificity was slightly more pronounced in the light-activated state than the dark state, indicating that our optogenetic Acr can selectively decrease off-target editing.
Comparing the results shown Figure 2.4A-B with the on-target editing efficiency in Figure 2.4C, a locus dependency of the editing activity becomes apparent. Depending on the targeted site, differences in the inhibition strength of CN-C3, as well as its dynamic range of light-control were observed. An additional experiment, including another target locus and different Cas9:Acr vector mass ratios, further confirmed this effect (Supp. Fig. 1). This result is in line with previous reports (Bubeck *et al*, 2018; Mathony *et al*, 2020a; Hoffmann *et al*, 2019) and describes a phenomenon of locus-specific CRISPR inhibition, that we generally observe when working with Acrs and which is not specific to AcrIIC3 or CN-C3.

2.1.3.2 *Light-induced switching does not affect protein stability*

An important aspect, mechanistically and with respect to applicability of CN-C3, are the expression levels of the different components. We had previously shown that the introduction of an additional domain into AcrIIC1 increases its stability (Mathony *et al*, 2020a), a factor that likely contributed to its improved inhibition strength. In case of CN-C3, it was still unclear, whether LOV2 domain insertion affected only the inhibitor's activity or also its stability. To investigate this question, I performed a time-resolved Western blot experiment to determine protein expression at different time points in between 6 h to 48 h post transfection (Fig. 2.5A-C). Detection of the proteins was enabled by their fusion to HA-tags. Quantification of the results revealed substantial differences of the expression levels relative to the α-tubulin reference between replicates (Fig. 2.5A-C). The data points within a replicate, as well as the global trends, however, were very consistent: 6 h after transfection, neither Cas9 nor the Acrs were detectable. At the next time point, 12 h post-transfection, the relative CN-C3 expression started to rise (Fig. 2.5A), while wildtype AcrIIC3 was still undetectable. The expression of both Acrs plateaued after 24 h. Importantly, no effect of the LOV2 insertion on the final protein levels was observed. Relative expression levels of Cas9 behaved similarly, being rather low at the 12 h time point and reaching strong expression after 24 h. In one replicate, a slight increase was visible between the 24 h and the final 48 h time point.

2.1.3.3 *CN-C3 is principally compatible with Nme2Cas9*

Having investigated CN-C3 with respect to the inhibition of *Nme*Cas9, I further assessed, if the optogenetic control of *Nme*2Cas9 is also possible. Application of *Nme*2Cas9 resulted in editing levels around 20 % in HEK293T cells. (Fig. 2.5D, E). Background signals of 2.5 % to 5 % were, however, as visible in the TIDE sequencing results. As expected, wildtype AcrIIC3 completely inhibited *Nme*2Cas9 down to the background levels. CN-C3 and CN-C3G, exhibited only slightly increased background editing in the dark as compared to wildtype AcrIIC3, while upon light-induction indel rates increased to 10-15 %. Especially, on the GAPDH locus, this

represented a substantial decrease in gene editing efficiency as compared to the Cas9 only control. Different Cas9:Acr ratios, optimized for the locus of choice, might pose a solution to this problem. Taken together, only modest light-dependent editing was visible, which might be at least partially caused by the high background signal. A further, more detailed investigation of gene editing with *Nme*2Cas9 would be required in order to investigate the potential of CN-C3 on this Cas9 orthologue in detail.

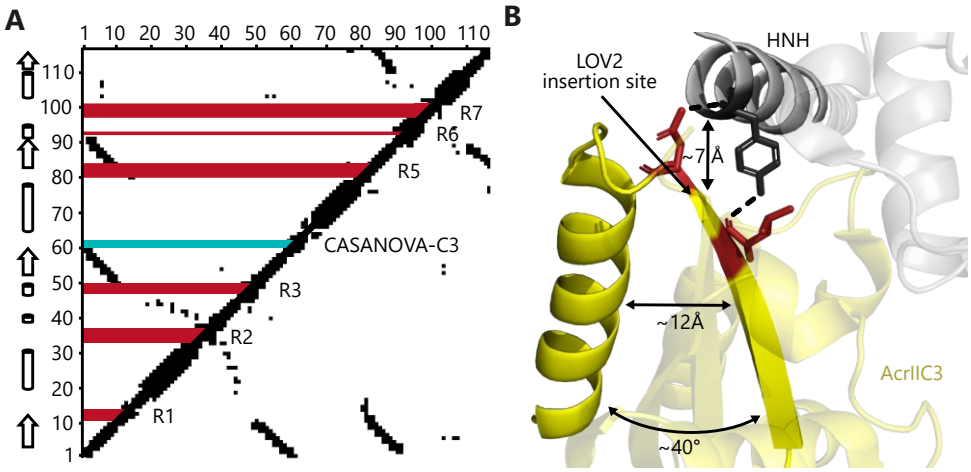


Figure 2.6: CN-C3 carries the LOV2 insertion at an unexpected site. (A) Residue contact map of AcrIIC3. Black squares mark residue pairs with relative distances <7 Å. Secondary structure elements are indicated next to the graph (left). Insertion sites that were experimentally tested and resulted in inactive AcrIIC3-LOV2 hybrids are marked in red. The successful insertion site for CN-C3 is indicated in green. R, region that was sampled. (B) Structural representation of the LOV2 insertion site for CN-C3. AcrIIC3 is shown in yellow and the *Nme*Cas9 HNH domain in grey. The residues that mediate the contact between the Acr and Cas9 are shown as sticks and highlighted in red. The distance and angle between the secondary structure elements flanking the insertion site are indicated. PDB-ID: 6J9N.

2.1.3.4 *CN-C3 carries the functional AsLOV2 insertion at an unexpected site*

From a protein engineering perspective, not only the performance of the engineered switches is of interest, but also possible mechanisms of action. Here, a better understanding of the interaction between CN-C3 and Cas9 is required. To this end, our collaboration partners, from the group of Professor Bruno Correia, designed computational models of the AcrIIC3-LOV2 hybrids (Supp. Fig. 2). Their analysis revealed that diverse LOV2 conformations relative to AcrIIC3 are possible for both variants, CN-C3 (Supp. Fig. 2A) and CN-C3G (Supp. Fig. 2B). Superimposing these models onto the experimentally resolved structure of AcrIIC3 bound to the *Nme*Cas9 HNH domain (Supp. Fig. 2C, D) showed that one predicted conformational cluster exists that would not sterically clash with the Acr's binding partner. This is in line with the experimental observation that the LOV2 insertion did not significantly reduce AcrIIC3 activity in the dark. More recently, complete structures of AcrIIC3, bound to the full-length *Nme*Cas9 have been published (Sun *et al*, 2019). The authors showed, that AcrIIC3 not only binds to the HNH domain, but also dimerizes the nuclease by making contacts with the REC2 lobe, resulting in a circular complex comprising two nucleases, as well as two inhibitors. This unique complex can be expected to drastically decrease the structural flexibility granted to the LOV2 domain positioning relative to AcrIIC3. Structural alignment of the hybrid inhibitor to the dimerized

structure (Supp. Fig. 3) confirmed, however, that the aforementioned conformational cluster would exactly fit into the limited space without clashing with either *Nme*Cas9 binding partner. The agreement between the modeled fusion protein and the structure of the *Nme*Cas9-AcrIIC3 complex further supported the validity of the structural model. At the same time, the structural restriction due to the dimerized conformation of the complex also exemplify that the selected insertion site must be one of very few positions that can tolerate a functional insertion of the LOV2 domain. This observation is backed by insertion screening experiments performed prior to this thesis, indicating that solely at the AcrIIC3 region around F59, LOV2 insertion results in functional AcrIIC3-LOV2 hybrids (Fig. 2.6A) (Hoffmann *et al*, 2021).

With respect to the general design of switchable proteins, previous publications suggested the selection of so called "tight loops", which bridge spatially aligned secondary structure elements often connected by hydrogen bonds (Dagliyan *et al*, 2016, 2019). These loops would enable the disturbance of the secondary structure integrity and position upon the light-induced conformational changes of the LOV2 domain (refer to section 1.3.4.2 for details). Such tight loops can be easily identified from contact maps, in which the aligned secondary structure elements are visible as lines orthogonal to the main diagonal (Fig. 2.6A). Although several of these loops exist in AcrIIC3, we know from the previously mentioned insertion screen experiments that none of these positions tolerated the insertion of the LOV2 domain (Fig. 2.6A) (Hoffmann *et al*, 2021). The structural constraints caused by the Cas9-dimerization mechanism, partially explain the failure of the "tight loop" concept in case of AcrIIC3.

The successful insertion site instead, although positioned within a surface exposed loop, does link an α-helix to a β-sheet, which are not closely connected by any hydrogen bonds (Fig. 2.6B). Instead, the insertion site is located between two key residues, L58 and N60, both of which directly contact the HNH domain of Cas9 (Kim *et al*, 2019, 3). It is at least surprising that this insertion site did not seem to interfere with AcrIIC3 activity. On the other side, it appears reasonable that only slight conformational changes of these important residues, caused by photo-switching, could affect Cas9 inhibition. Taken together, the structural analysis revealed unconventional mechanistic insights, that are distinct from previously studied *As*LOV2-based photo-switches.

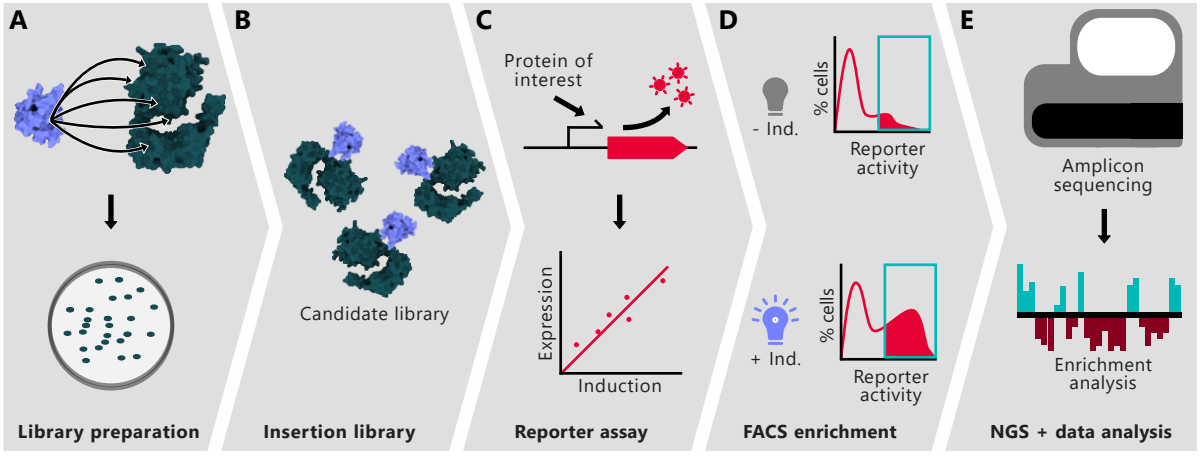


Figure 2.7: Overview of the domain insertion screen. (A, B) First, comprehensive insertion libraries are created. (C) A reporter assay to measure the activity of the parent protein is established, followed by FACS enrichment of functional variants from the domain insertion libraries (D). The input libraries, as well as the enriched subsets are finally sequenced and analyzed (E).

2.2 Unbiased insertion screens

Previous work by me and others in our lab has shown how cumbersome the identification of suitable insertion sites for the design of allosteric switches can be (Hoffmann *et al*, 2021; Bubeck *et al*, 2018). Currently the main hurdle is the fact that mostly positive examples are published and very few comprehensive domain insertion tolerance screens exist (Nadler *et al*, 2016; Oakes *et al*, 2016; Coyote-Maestas *et al*, 2019). To create a better basis for the biophysical and structural analysis of the domain insertion permissibility, I planned an unbiased screening of structurally and functionally diverse proteins. Towards this goal, the first prerequisite is an efficient cloning strategy for comprehensive domain insertion libraries (Fig. 2.17A). To this end, I used saturated programmable insertion engineering (SPINE) (Coyote-maestas *et al*, 2019). The details of the procedure are described in the methods (section 4.1.9.1). Once a candidate library is constructed (Fig. 2.17B), a reporter assay system is required to link the activity of the parent-insert hybrid proteins to a fluorescent readout (Fig. 2.17C). These assays are typically established and performed in *E. coli* due to the possibility to work in a monoclonal setting. Finally, functional hybrid proteins can be enriched, via FACS of the fluorescent fraction of cells (Fig. 2.17D). In case a switchable domain, such as *As*LOV2, is inserted, the FACS screen could further be performed after incubation of the culture in presence or absence of the cognate inducer. In the end, the enriched libraries are subjected to NGS, followed by bioinformatic analysis of the results (Fig. 2.17E). It is important, though, to also sequence the initial libraries, since the frequency of candidates in the input pool can already differ prior to the sorting.

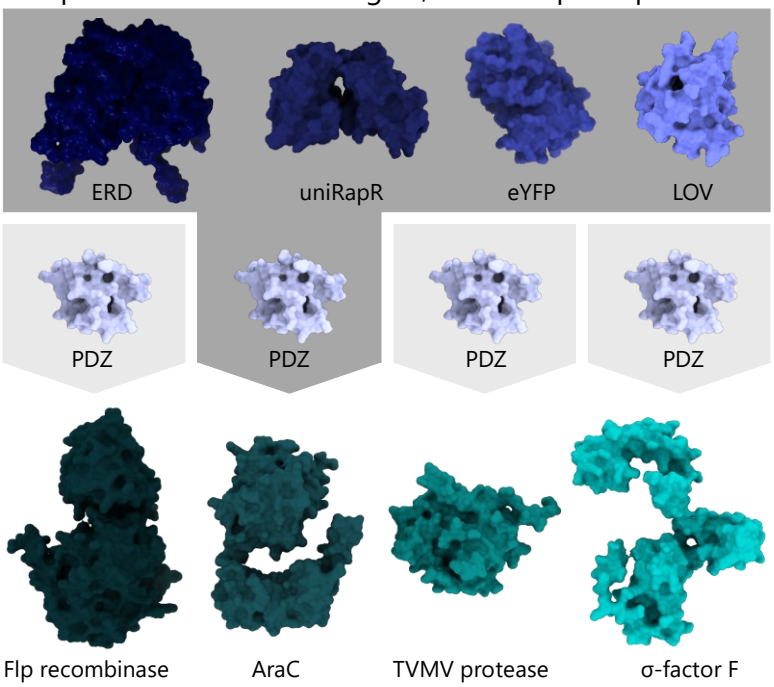


Figure 2.8: Candidate proteins and insert domains. A comprehensive domain insertion screen was performed for four candidate proteins: The Flp recombinase, the transcription factor AraC, the protease of TVMV and the SigF from *B. subtilis*. The PDZ domain from murine α1-syntrophin was used as insert. In case of AraC, additional libraries were generated and analyzed, comprising an ERD, uniRapR, eYFP and the *As*LOV2 domain as inserts. The depicted structures of the parent proteins are AF2 predictions. PDB-IDs of insert domains: 2V0U, 1FAP, 1Z86, 6ZQO, 1A52.

2.2.1 Insertion library screening of structurally and functionally diverse proteins

Four effector proteins were chosen for the domain insertion screen: The transcription factor AraC, the recombinase Flp, the TVMV protease and the σ-factor F (SigF) from *B. subtilis* (Fig. 2.8). This specific selection was made for several reasons. First, the proteins are structurally and

functionally diverse. Previous screens with different insert domains have primarily focused on closely related ion channels (Coyote-Maestas *et al*, 2019, 2020), which inherently limits the interpretability to this specific protein class. Second, the reporter assays for all four candidates could be adapted from published experiments, ensuring that the screening procedure could

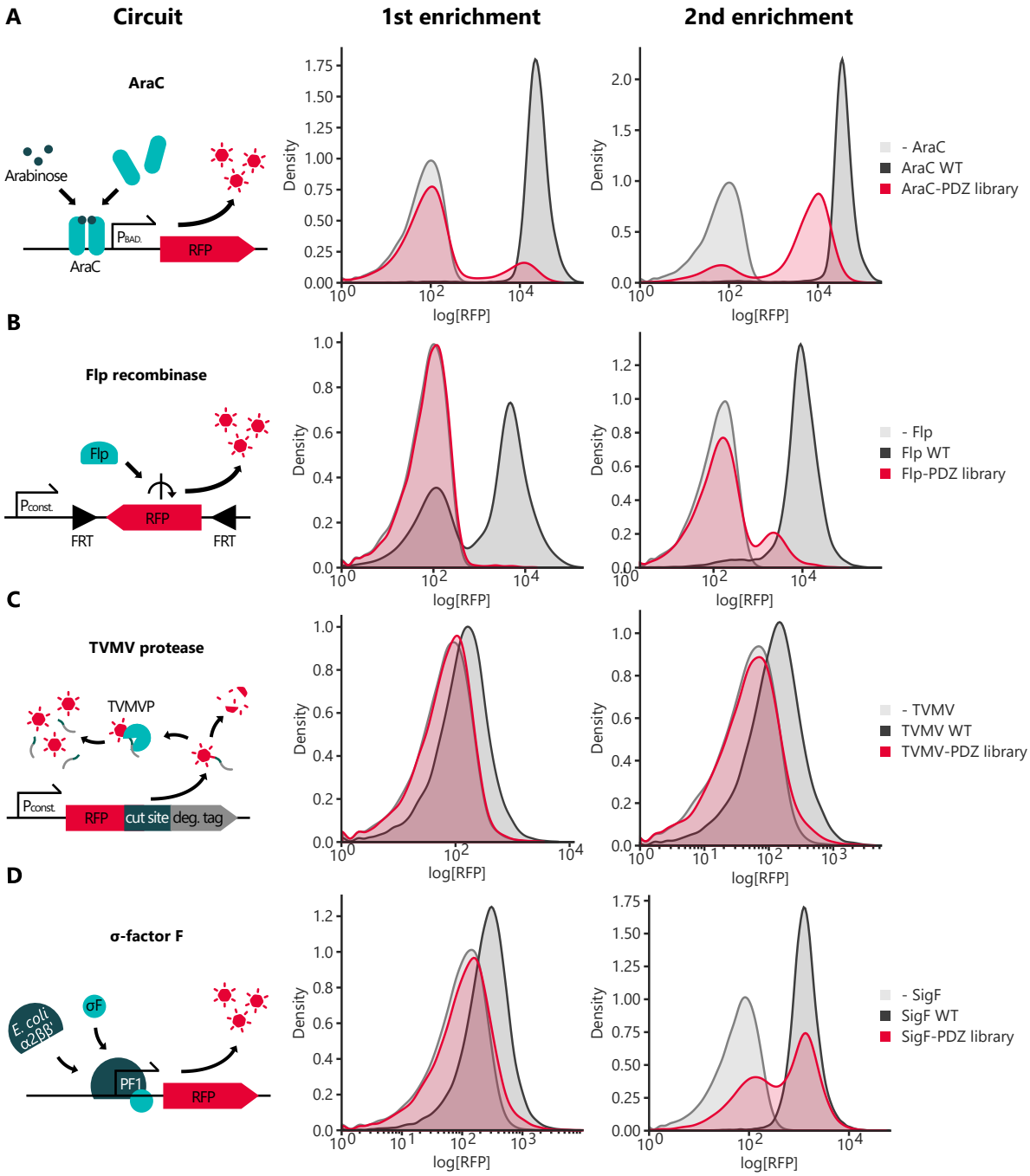


Figure 2.9: Reliable reporter assays enable the enrichment of the candidate libraries via FACS. Schematics of the reporter assays for AraC (A), the Flp recombinase (B), the TVMV protease (C) and SigF (D) are shown (left panel). *E. coli* cultures carrying the reporter construct and the domain insertion library were inoculated from precultures and grown under induction for 16 h, followed by FACS. Two rounds of sorting were performed. The panels in the middle and on the right show representative histograms generated from 25,000 gated events for the first and second round of enrichment, respectively. The (-) negative controls carried a plasmid expressing a different candidate protein.

rapidly be established in the lab. Third, all proteins have been thoroughly studied, thus providing substantial additional information with respect to structure and function as a rich resource for the subsequent analysis (refer to section 1.4.1.2). Finally, all proteins are widely used "workhorses" in synthetic biology, so that the information obtained in this study, as well as potential switchable variants derived from the screen would be of high relevance for the scientific community.

As insert, the PDZ domain from murine α1-syntrophin was selected. It is a relatively small domain (86 AA) with a compact globular fold, which has been used for similar purposes before (Oakes *et al*, 2016). Importantly, its termini are located in spatial proximity, so that the insertion into other proteins is possible without substantially affecting their overall structure. On top, I planned to test to what extend the domain identity affects the insertion tolerance. To this end, I selected four additional domains for the random insertion into AraC. They included the *As*LOV2 domain, the estradiol binding domain from the human estrogen receptor-α (ERD), an enhanced yellow fluorescent protein (eYFP) (Ormö *et al*, 1996) and the synthetic rapamycin-binding domain uniRapR (Dagliyan *et al*, 2013). These candidates extended the size range of inserts to up to 257 AA in case of the ERD and further included switchable candidates: The LOV2 domain reacts to light, while uniRapR and ERD change their conformation upon binding of rapamycin (Ormö *et al*, 1996) and β-estradiol or 4-hydroxytamoxifen (Shiau *et al*, 1998; Tanenbaum *et al*, 1998; Wärnmark *et al*, 2002), respectively.

The construction of the comprehensive libraries via SPINE was followed by NGS in order to ensure a good coverage of all possible insertion sites. The frequency of each variant was assessed, by selecting reads that span the boundaries between insert and parent protein. The processing of the sequencing data further included quality control steps, ensuring that only functional fusions without mutations at the ligation sites were considered valid. The complete processing procedure is outlined in the methods (section 4.2.3). Analysis of the initial libraries revealed near complete coverage of all possible insertion sites for most libraries (Supp. fig. 4). The only exception was the SigF-PDZ library, for which the overall sequencing depth was very low and the several insertions within the first 50 amino acids appeared to be missing.

To functionally screen the libraries, I constructed reporter plasmids, based on the monomeric red fluorescent protein 1 (RFP) (Campbell *et al*, 2002). The design of the reporter circuits is depicted in Figure 2.9 (left). In short, the AraC reporter was created simply by placing the RFP coding sequence under control of a pBAD promoter. In case of the Flp recombinase, RFP was expressed from a constitutive promoter (J23102, http://parts.igem.org/Promoters/Catalog/Anderson). However, the coding sequence was inverted and flanked by Flp recognition target (FRT) sites. In the ground state, a dysfunctional mRNA is transcribed and only upon inversion of the open reding frame (ORF) by the recombinase, RFP is expressed. It is important to note, that this reporter gives, in contrast to the other examples, a binary output, since its activity is permanently switched on by the single inversion event. A pseudo-analog behavior is achieved though, as several plasmid copies are present within the cells, which all must be activated individually by the recombinase. To report TVMV protease activity, a ssrA-like degradation tag (McGinness *et al*, 2006) was fused to a constitutively expressed RFP; a TVMV recognition site was placed in between RFP and the degradation tag. Active TVMV protease would thus cleave off the degron resulting in RFP

stabilization and an increase in fluorescence. Several similar designs of this reporter were initially tested, as described in supplementary note 1.

Many potyvirus proteases undergo a process called autolysis (Kapust *et al*, 2001), during which the protease cleaves off its own C-terminal region albeit at low efficiency. This results in a truncated protease with decreased activity. This behavior is, however, not expected for TVMV protease and a truncated TVMV version was shown to maintain wildtype activity (Sun *et al*, 2010). Nonetheless, to ensure that only one protein species could be present during all assays, the active, truncated version was selected for the experiments. Finally, a reporter for SigF was constructed, based on a SigF-specific promoter design that was previously reported by Bervoets et al. (Bervoets *et al*, 2018).

When I screened the libraries by FACS, the activity distribution differed greatly between the candidate proteins (Fig. 2.9A-D, middle). For AraC, a very clear separation between the controls was achieved (Fig. 2.9A, middle and right panel). The insertion library showed a large peak with supposedly inactive variants at low fluorescence levels and a much smaller population of active variants with high RFP fluorescence. Since the number of active variants seemed to be rather small, I decided to enrich all libraries a second time (Fig. 2.9A-D, right). During second enrichment, the active peak of the AraC library was already much more pronounced. Its maximum though, had a slightly weaker RFP signal than the wildtype AraC control. The Flp recombinase library, in contrast, appeared to be largely inactive (Fig. 2.9B). Only a small tail at the left end of the peak towards higher fluorescence levels indicated the presence of active variants. At the second enrichment step a more pronounced population with active insertion hybrids was clearly visible in the data. Still the active population was overall small, indicating the importance of the second sorting round.

The TEV protease reporter system initially showed a minor dynamic range, i.e. a small difference between the positive and negative controls (Fig. 2.9C). Nonetheless, a small but clearly visible

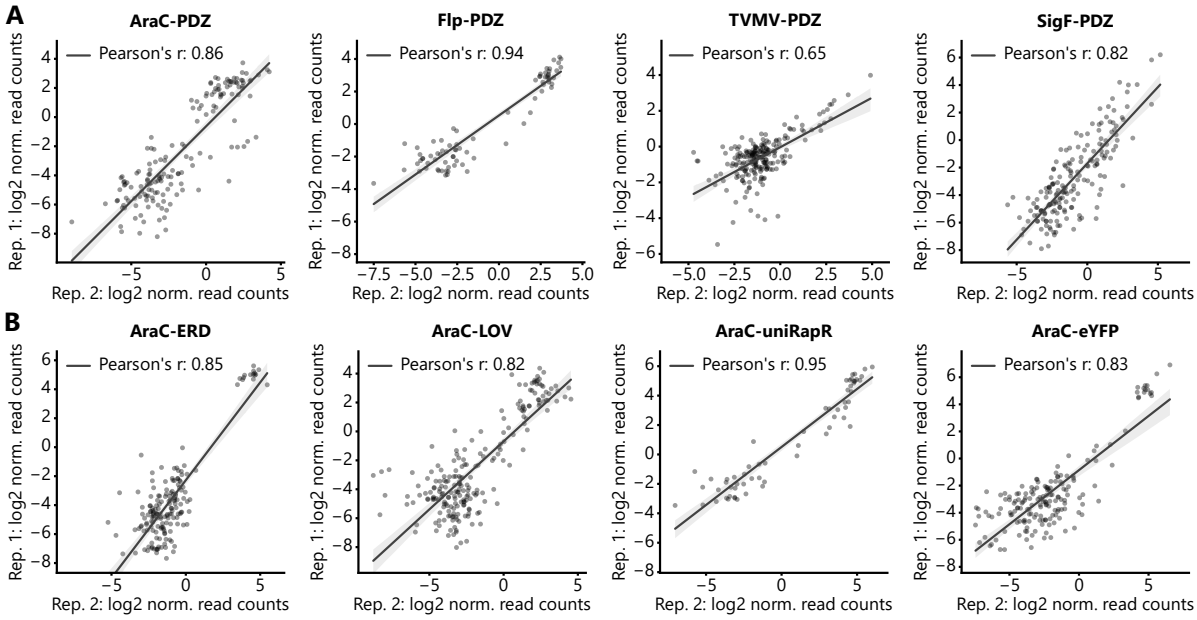


Figure 2.10: Domain insertion profiling outcomes are highly reproducible. The enrichment scores of biological replicate-1 are plotted against the respective scores from a second replicate-2 for all different parent proteins (A) and the additional AraC insertion libraries (B). (A, B) Only variants that were not fully depleted during enrichment are shown. A linear fit with 95 % confidence intervals is shown. Pearson correlations coefficients are indicated. Rep., replicate; norm., normalized.

shift of the library histogram towards higher RFP fluorescence was observed during the second round of enrichment (Fig. 2.9C, right). Finally, the library of SigF showed a growth defect at high induction rates. Interestingly, this effect disappeared after enrichment, hinting at a high proportion of toxic, probably misfolded proteins in the initial library (data not shown). To circumvent this problem, I sorted the initial library at low IPTG inducer concentration of 100 µM and doubled the amount to 200 µM for the second screening round. The difference in IPTG levels between the sorting rounds in case of SigF is also visible in the resulting histograms. For the positive control, a clear shift to higher fluorescence values during the second enrichment could be observed (Fig. 2.9D). A strong enrichment of functional SigF variants was achieved this way.

2.2.2 NGS of the enriched libraries reveals distinct patterns of successful insertions

Following the FACS screen, all enriched libraries were sequenced and subjected to the same processing pipeline as the input libraries (refer to methods, section 4.2.3 for details). The whole insertion screen and the sequencings were performed in biological duplicates. To calculate the enrichments of insert domains, the read counts from the sorted fractions were first normalized to the total number of reads from the respective sample. The resulting values were divided by the proportion of reads from the initial libraries at the respective position and finally Log2-scaled to equally weigh enrichment and depletion. The score is a measure for the fold enrichment per position and is referred to as "enrichment score" from now on. A score of -10 was automatically assigned to variants that went extinct during sorting. The value was rationally chosen to be just below the score of the variants with the strongest depletion still observable, i.e. with corresponding reads being present in the NGS dataset.

Following the data processing, I evaluated the reproducibility of the procedure. To this end, I analyzed the Pearson correlation between the replicates for all different proteins (Fig. 2.10A) as well as for AraC with its additional inserts (Fig. 2.10B and Supp. fig. 5). To provide a fair comparison, positions that were scored with -10 were excluded from that correlation as they would have artificially boosted the outcome (the same candidates were usually fully depleted in both replicates). The majority of the libraries showed strong correlations between biological replicates with a Pearson's $r > 0.8$ (Fig. 2.10A, B). An exception was the TVMV-PDZ library with a correlation coefficient of only 0.65 (Fig. 2.10A). I note that the strength of the enrichments was not always identical between replicates, as exemplified by regression lines with slopes different than 1.0 (Fig. 2.10A, B).

Next, we compared the NGS data with measurements of activity for individually selected variants using the previously described reporter assay. The obtained reporter activities were then compared between enriched and depleted variants (Supp. fig. 6). In most cases, drastic differences in activity between both groups were observed. One must also add that in two cases, only one enriched sample was measured. This was due to the fact that the experimentally validated variants were randomly selected and the majority of candidates within the libraries was depleted. Overall, the measurements were in good agreement with the sequencing-derived enrichments and depletions.

Mapping of the enrichment scores to the protein amino acid sequences for the PDZ libraries of all four proteins revealed different trends. Generally, roughly 80 % of the positions in each

protein were depleted (Fig. 2.11A-D). In case of AraC (Fig. 2.11A) and Flp (Fig. 2.11B), many variants disappeared completely after sorting in at least on replicate, while only very few insertion variants were completely abolished in case of the TVMV protease (Fig. 2.11C) and SigF (Fig. 2.11D). This has probably to do with the selection stringency that was applied during selection. In this regard, the relatively weak overall enrichment of the TVMV library is noteworthy, which corresponds well with the reduced resolution of the reporter assay, as stated in the last section.

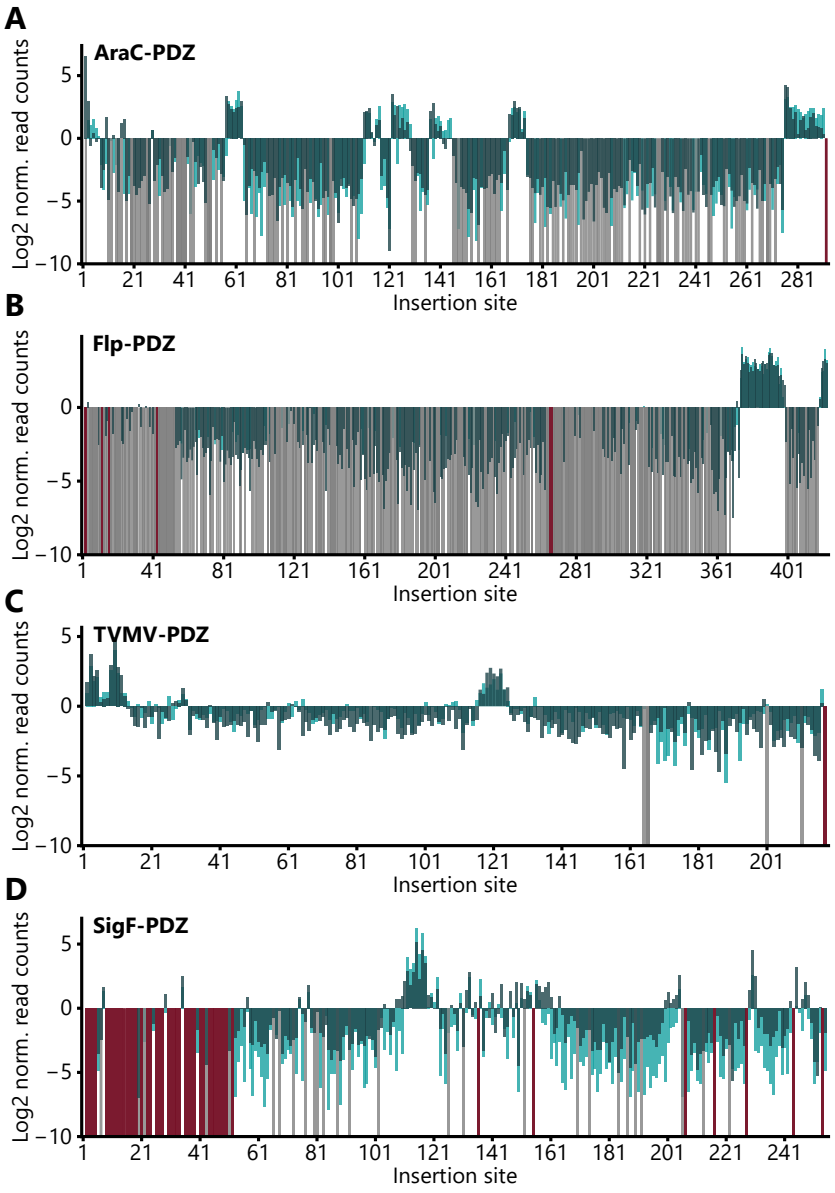


Figure 2.11: Domain insertion screens reveal clusters of surface sites that tolerate insertions. (A-D) The proportion of NGS read counts per insertion variant were normalized to the relative number of read counts derived from the input library, resulting in log2-normalized enrichment scores. Data from the candidate proteins AraC (A), Flp (B), TVMV protease (C) and SigF (D) with PDZ domain inserts are shown. Light green, dark green: individual replicates. Grey: variants with zero reads after enrichment. Red: variants missing in the initial library.

An interesting trend was the fact that positions tolerating insertions appeared in clusters spanning regions of different size. In this regard, my results differ from a previous study, which has shown a more scattered distribution of successful insertions (Oakes *et al*, 2016). Only very few single positions with enrichment, surrounded by depleted insertion sites, were detected for SigF. Also, the number of enrichment clusters differed between proteins. While the Flp recombinase showed only two regions with enrichments, both close to the C-terminus (Fig. 2.11B), AraC exhibited seven clusters, widely distributed over its primary sequence (Fig. 2.11A). After having gained a first impression with respect to domain insertion tolerance in the four effector proteins, I asked to which extend domain insertion depends on the nature of the inserted domain. This time, four different insert domains were included into the screen, i.e. LOV2, ERD, uniRapR, eYFP. To this end, I performed the library generation, FACS enrichment

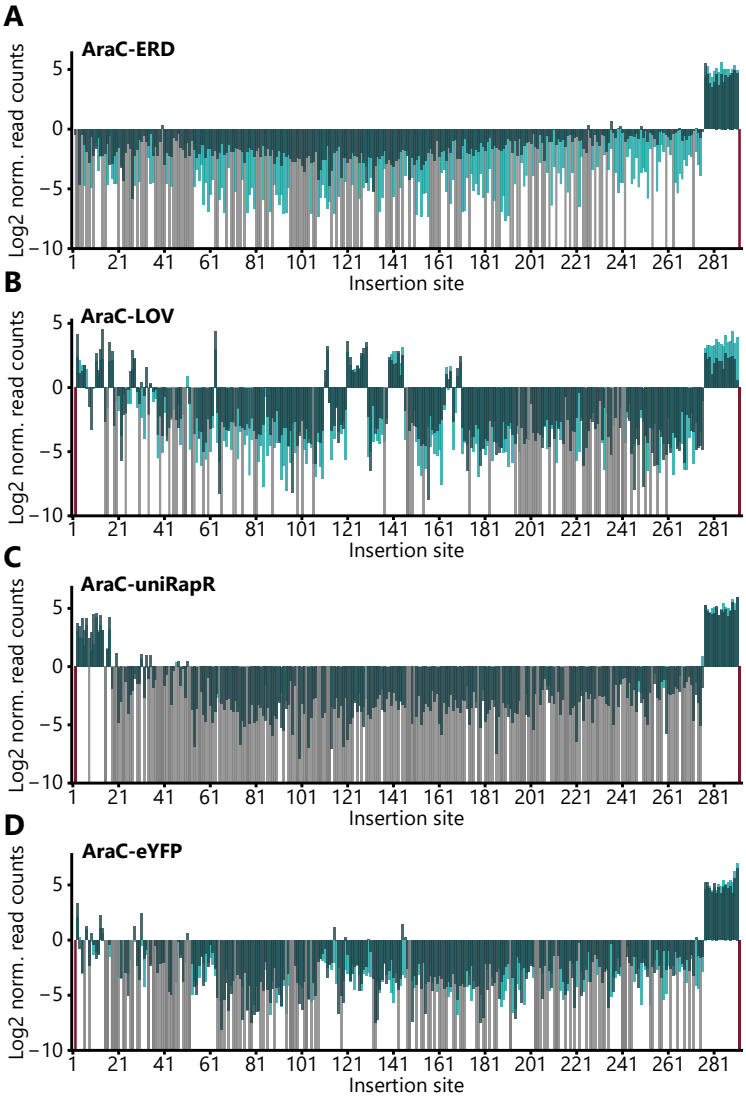


Figure 2.12: Insertion tolerance depends on the used insert domain. (A-D) After NGS of the enriched libraries, the proportion of read counts per insertion variant were normalized to the relative number of read counts derived from the input library, resulting in log2-normalized enrichment scores. Domain insertion data of AraC with the ERD (A), LOV2 (B), uniRapR (C) and eYFP (D) insert domains are shown. Light green, dark green: individual replicates. Grey: variants with zero reads after enrichment. Red: variants missing in the initial library.

and NGS for AraC, since it showed the broadest distribution of insertion-tolerating regions. Remarkably, the observed clusters of enriched insertion variants differed substantially between the insert domains (Fig. 2.12, supp. fig. 5). While the LOV2 domain (Fig. 2.12B) exhibited a pattern very similar, albeit not identical, to the PDZ domain (Fig. 2.11A), for the other three inserts substantially fewer enriched positions were detected (Fig. 2.12A, C, D). The uniRapR library was prominently enriched at the N- and C-terminal regions of AraC. The ERD and eYFP libraries, in contrast, resulted in peaks only at the C-terminal end of the protein. Taken together, it became apparent that rather promiscuous sites exist (especially at the C-terminus), which accepted the insertion of diverse domains, while other protein regions tolerated only the insertion of the more compact domains, such as LOV2 and PDZ. Interestingly, I hardly discovered insertion sites that were selective for just one specific domain.

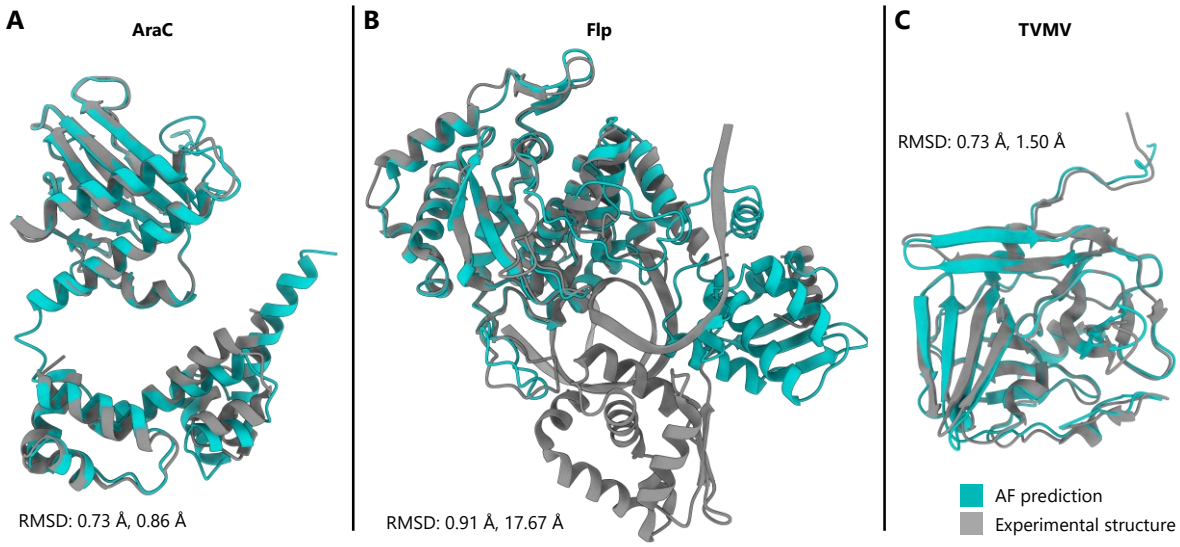


Figure 2.13: AlphaFold2 predictions accurately capture the structures of the candidate proteins. (A-C) Structural alignments between experimentally resolved structures (grey) and AlphaFold2 predictions (green) are shown for AraC (A), Flp (B) and the TVMV protease (C). The RMSD of the aligned residues as wells as the RMSD for all amino acids are shown. PDB-IDs: 2ARA, 2K9S, 1FLO, 3MMG.

Having described the observations from the screen on the sequence level, I next wanted to relate these findings to protein structural information. Unfortunately, experimentally resolved full-length structures were only available for the TVMV protease (Sun *et al*, 2010) and the Flp recombinase (Chen *et al*, 2000). In case of AraC, partial structures of the arabinose binding domain (Soisson *et al*, 1997), as well as the DNA-interacting domain (Rodgers & Schleif, 2009) exist. No experimentally resolved structure has so far been published for SigF. To circumvent this limitation, I predicted structures for all four proteins with AlphaFold2 (AF2) (Jumper *et al*, 2021) using the colabfold framework (Mirdita *et al*, 2022) (Fig. 2.13). The predictions of AraC and TVMV were in excellent agreement with the experimental structures. In case of the Flp recombinase, a large proportion of the structures aligned perfectly, while the position of the N-terminal domain was significantly shifted (Fig. 2.13B). The fold of this domain was, however, still highly similar to the experimental structure. Finally, although no comparison could be made for SigF, the conformation predicted by AF2 is in agreement with typical σ-factor folds (Paget, 2015). To be able analyze the structures of all proteins as well as for reasons of consistency, I show AF2 structures in the following section. To get a more detailed view of

possible Flp conformations, the experimentally resolved structure is additionally shown for comparison (Supp. Fig. 7).
Starting with AraC, the protein consists of an N-terminal arabinose binding domain and a C-terminal DNA-binding domain which are spatially clearly separated (Fig. 2.14A). Well-characterized functionally important residues were identified (marked in grey). These include R38 and W95 that make important contacts to bound arabinose (Soisson *et al*, 1997). Y103, E106 and Y146 instead form connections with a second AraC monomer in the activated, dimerized state (PDB-ID: 2ARC) (Soisson *et al*, 1997). The DNA-binding residues were identified from a structural alignment to the highly similar transcription factor Rob, of which a crystal structure in the DNA-bound form exists (PDB-ID: 1D5Y) (Kwon *et al*, 2000).
Mapping of the enrichment scores onto the structure revealed several trends (Figure 2.14A). Overall, insertions within the arabinose-binding β-barrel were mostly depleted, except for the two outward-pointing β-sheets (β4 and β5). At the ends of the two central α-helices that are important for AraC dimerization, several positions tolerated the insertion of the PDZ domain.

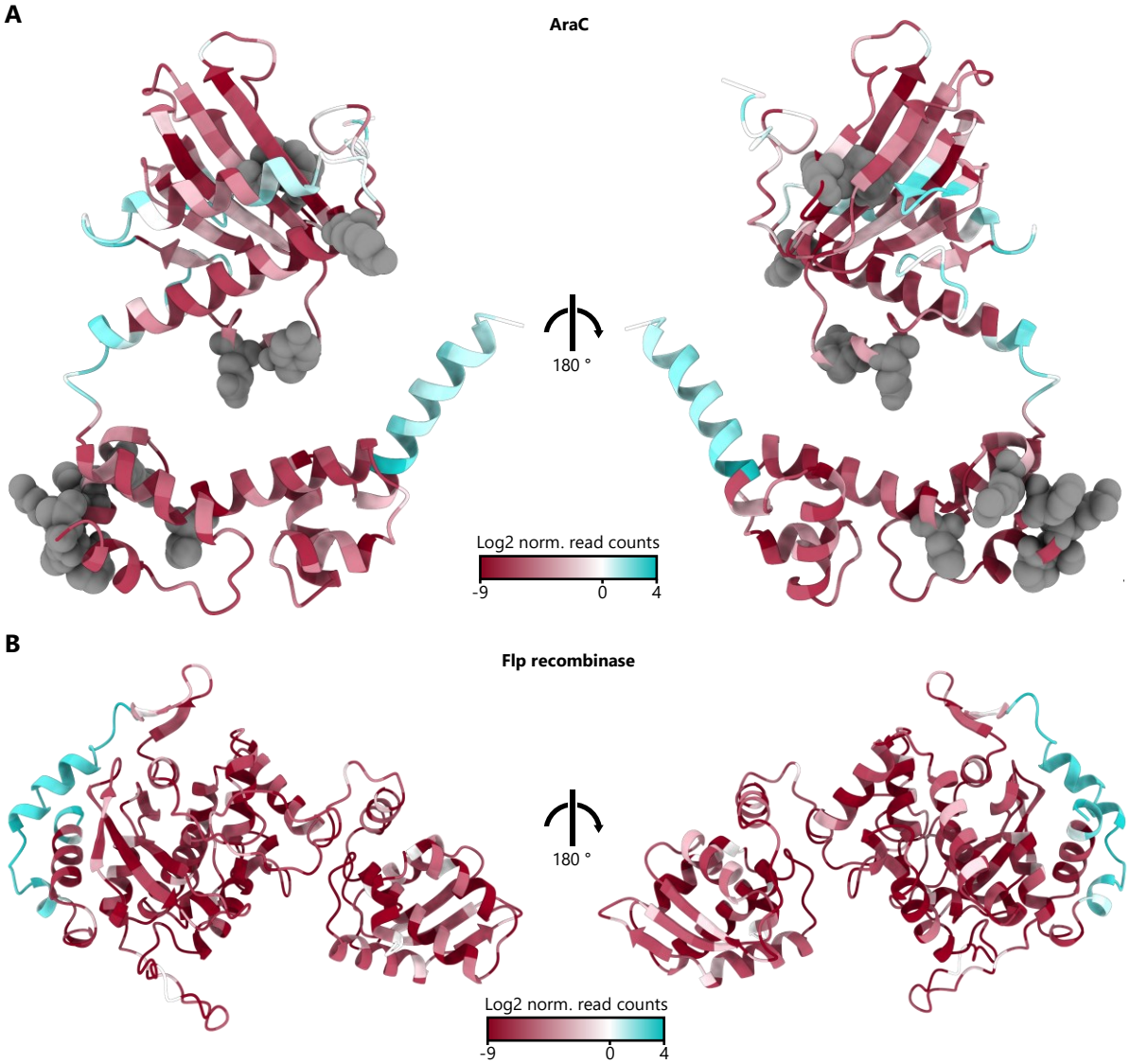


Figure 2.14: Positions with insertion tolerance are clustered at diverse, locally confined surface sites (i). (A, B) the insertion scores from the PDZ libraries are mapped onto the AF2 structure predictions of AraC (A) and Flp (B), respectively. The functional residues of AraC are indicated in grey.

The enrichment appeared to be the strongest within the domain-connecting linker. In line with physical constraints, the positions around the DNA-binding residues did not tolerate domain insertion. Within the DBD, only the most C-terminal α-helix of AraC exhibited high enrichment scores. Of note, this helix is oriented outwards, away from the supposed location of the DNA or the second AraC monomer, which explains the high scores. Overall, the depletion of insertions within the DBD appeared to be slightly stronger, as compared to the rest of the protein. Also, none of the loops within the AraC DBD exhibited domain insertion tolerance.
In case of the Flp recombinase, the situation differs significantly, as only two sequence stretches were enriched (Fig. 2.11 B). Structurally both clusters are in spatial proximity (Fig. 2.14B). The Flp recombinase consists of two domains, which bind the target DNA in their middle (Supp. Fig. 7). The C-terminal domain is responsive for DNA-cleavage (Lee *et al*, 1999). Interestingly, this domain carries the region that tolerated domain insertion. In this context, I note that in its DNA-bound state, the Flp recombinase exists as a tetramer, bound to two copies of the target DNA sequence, forming a Holliday-junction (Supp. Fig. 7) (Chen *et al*, 2000). Each pair of Flp

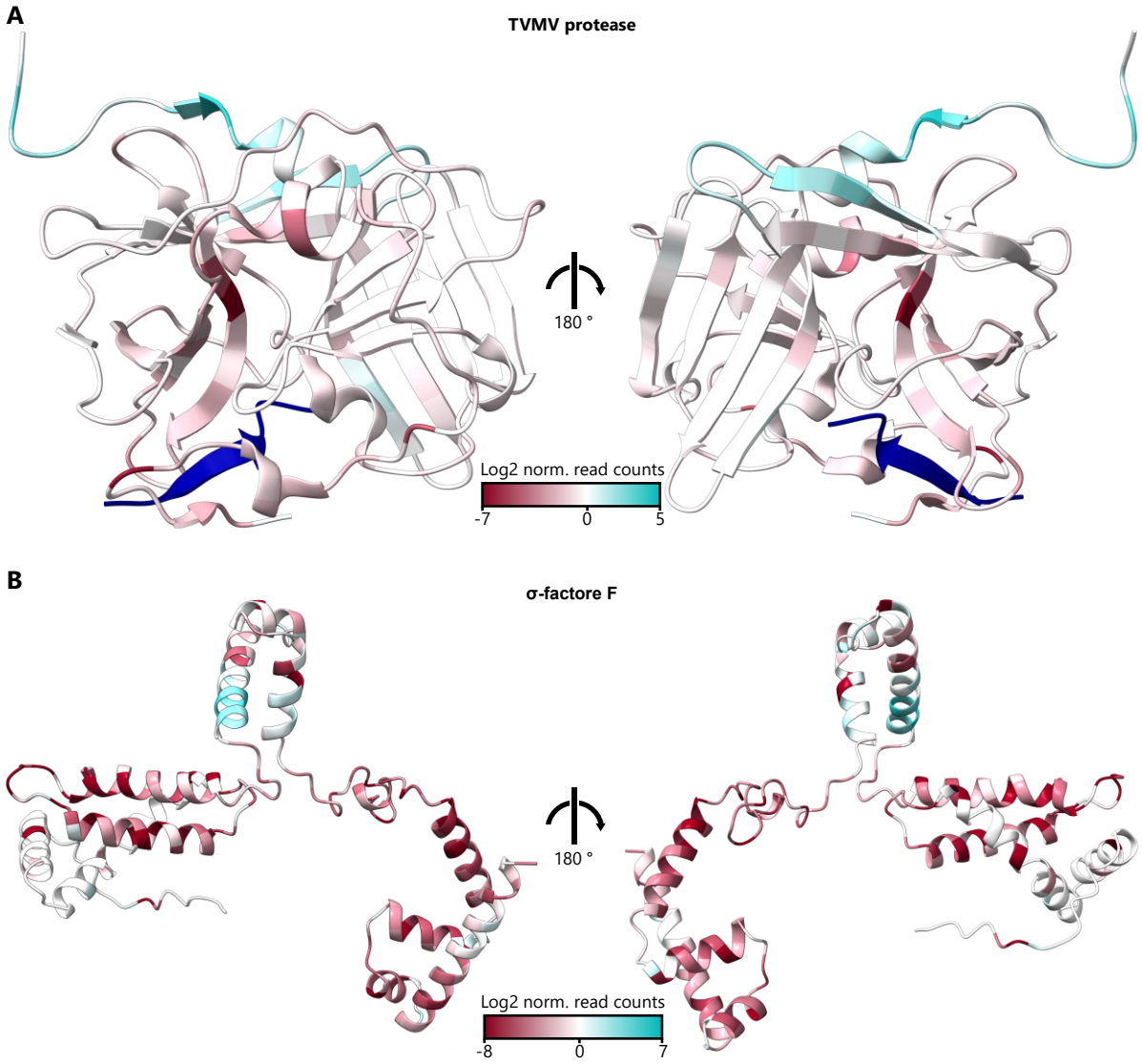


Figure 2.15: Positions with insertion tolerance are clustered at distanced locally confined surface sites (ii). (A, B) the insertion scores from the PDZ libraries are mapped onto the AF2 structure predictions of the TVMV protease (A) and SigF (B), respectively. The protease substrate is depicted in blue.

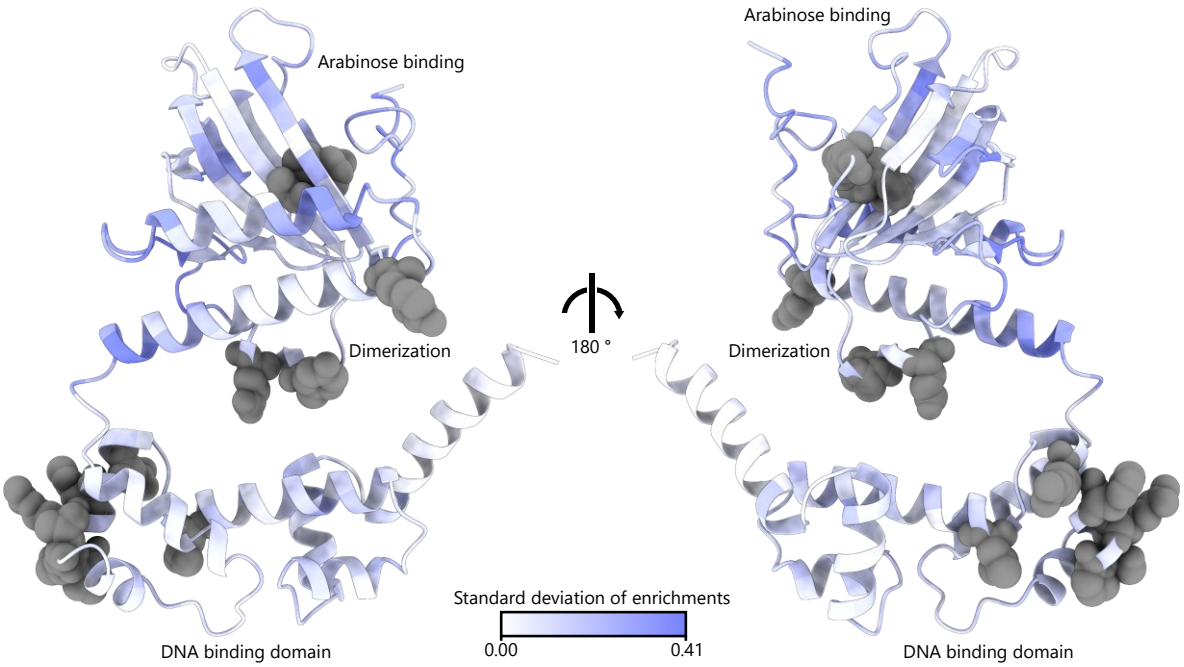


Figure 2.16: Insertion tolerant regions are domain-specific and are scattered across AraC. The enrichment scores for all insert libraries of AraC were min-max scaled and the SD was calculated for each position. The AF2 structure of AraC is colored by SD. Functionally critical residues are highlighted in grey.

proteins facing each other exhibit the same conformation, while the conformation of the other pair differs, respectively (Conway *et al*, 2003). Furthermore, Flp is known to cleave DNA in trans, meaning it does not catalyze the cleavage of the strand it is bound to, but of another DNA site within the complex (Lee *et al*, 1999). Altogether, the structure of the DNA-bound tetramer, as well as the cleavage mechanism are highly complex (Supp. Fig. 7). It is thus not surprising that only a very limited fraction of the protein's surface was amenable to domain insertion.

In case of the TVMV protease, the results overall trends share similarity with the Flp recombinase in the way that insertions were mainly tolerated within one surface patch (Fig. 2.15A). As mentioned before, enrichments and depletions were less pronounced as compared to the other candidate proteins. Only very few positions inside the protein's core and around the substrate-binding site were obviously depleted. The TVMV protease is the only protein from the set that acts on its own, i.e. it does not need any protein-protein interactions besides engaging with its substrate peptide sequence. Nonetheless, most surface sites appeared to only modestly tolerate domain insertion. In this context, one must consider that, in contrast to other protein effector candidates, the protease exhibits a rather compact fold without larger unstructured regions.

Finally, SigF is structurally constituted very differently from the TVMV protease. According to AF2, it consists of three rather small domains which are connected via long flexible linkers (Fig. 2.15B). With respect to possible domain insertion sites, I note that SigF is part of a large transcription initiation complex and responsible for recruitment of RNA-polymerase (Paget, 2015). Consequently, many of the sites in principle amenable to domain insertion are likely to be important protein-protein interaction interfaces. The SigF C- and N- terminal domains, which are supposed to bind DNA (Paget, 2015; Bervoets *et al*, 2018), were most prominently depleted in insertion variants (Fig. 2.15B). Here, only very few positions exhibited a modes enrichment. The middle domain, in contrast, is constituted by a coiled-coil motive that

tolerated insertions at several sites. Interestingly, the flexible inter-domain linkers, as well as some of the larger loop regions, were not enriched for domain insertions.
Finally, the differences in domain insertion tolerance of AraC with respect to the various inserts were investigated. To this end, data for each insert domain was min-max scaled and the standard deviation of the enrichment scores was calculated for each position. The results are shown in Figure 2.16. As it was already visible in the sequential representation, differential insertion tolerance seemed not to be associated with high or low enrichment scores (compare Figure 2.14A). For instance, parts of the β-barrel that generally did not tolerate insertions had a low SD, while the same applies to the C-terminal α-helix, which accepted the insertion of all five different domains. In sum, the domain insertion permissibility appeared to vary less in the DBD, as compared to the N-terminal fraction of the protein. In particular, the dimerizing α-helices and the neighboring β-sheet harbor positions with high variability. As shown above (Fig. 2.12), the differences in the standard deviation mainly arise from diverging scores between the more compact PDZ- and LOV2 domain, as compared to the remaining larger domains. Finally, it is striking that sites with high variability do not appear to be clustered around functionally important regions of the protein.

2.2.3 Single biophysical amino acid features at the insertion site do not explain insertion preferences

Having analyzed the data with focus on the individual proteins, I next aimed to investigate general trends in the datasets. In order to gain a complete picture of the obtained results, I started to explore the correlation of diverse features with domain insertion tolerance. A previous study already hinted at the absence of a clear link to general sequence or structure features (Coyote-Maestas *et al*, 2019). To validate these observations on my own, more diverse dataset, I analyzed possible preferences for certain amino acids or biophysical properties that could determine domain insertion permissibility. Such analyses are frequently performed following deep mutational scanning experiments, in which single amino acids are mutated (Willow Coyote-Maestas *et al*, 2022; Faure *et al*, 2022; Dunham & Beltrao, 2021). Point mutations, however, are positional changes by definition. Domains, instead, are inserted between two amino acids and can be considered much more drastic alterations, likely to structurally and/or functionally affect a larger proportion of the protein. To account for the issue of position assignment in the following analysis, the two neighboring amino acids were assigned to the insertion between them. Consequently, the mean value of their biophysical properties was considered to correspond to the enrichment of this insertion variant.
As a starting point, I looked at preferences for specific amino acids in the vicinity to insertion sites (Fig. 2.17). For none of the proteins, clear trends were visible. Between proteins, the means of the enrichment scores changed with the overall protein-specific insertion permissibility, meaning that the average insertion tolerance of an amino acid in the Flp recombinase necessarily had to be lower than the corresponding value for AraC. Overall, the distributions were rather similar for all amino acids with only weak tendencies into either direction. In AraC (Fig. 2.17A), for example, threonine seemed to be weakly associated with depleted variants, while for the Flp recombinase (Fig. 2.17B) tryptophane and glutamic acid were more frequently located near insertion sites with higher scores. However, these very weak trends were not conserved across proteins. In contrast to Flp, for instance, tryptophane was among the amino

acids with higher corresponding enrichment scores in case of the TVMV protease (Fig. 2.17C). The absence of clear trends is even more obvious, when focusing on different groups of side chains, such as aliphatic, or charged residues. Together, these observations indicate that amino acid identity at the insertion site cannot explain domain insertion tolerance. As stated earlier, this finding is in agreement with a previous report (Coyote-Maestas *et al*, 2019). The result further represents a major difference to DMS screens, where clear trends would be expected. This is not surprising, given the aforementioned, strong difference between the outcomes of point mutations and insertions. Finally, I note that due to the limited size of single proteins, only a relatively small number of data points were available for each amino acid. A larger dataset would be required to ensure that no trend was missed due to noise within the small dataset.

Since the identity of single residues was shown to be not informative with respect to domain insertion, I next considered structural aspects. Generally, it is to be expected that domain insertion would be tolerated more frequently at surface-exposed sites as compared to sites buried within the protein, since the latter would result in steric clashes and misfolding. To analyze the surface exposure requirement in context of our dataset, I calculated the average surface accessibility (ASA) for each position in our four effector proteins. As mentioned before, the mean of the values corresponding to the two neighboring amino acids was assigned to the insertion site effectively located between them. In principle, the expected trend was visible, since negative enrichment scores primarily had an ASA below 0.6 (Fig. 2.18A and Supp. fig. 8A). Also, variants with positive enrichment scores tended to correspond to higher ASA values.

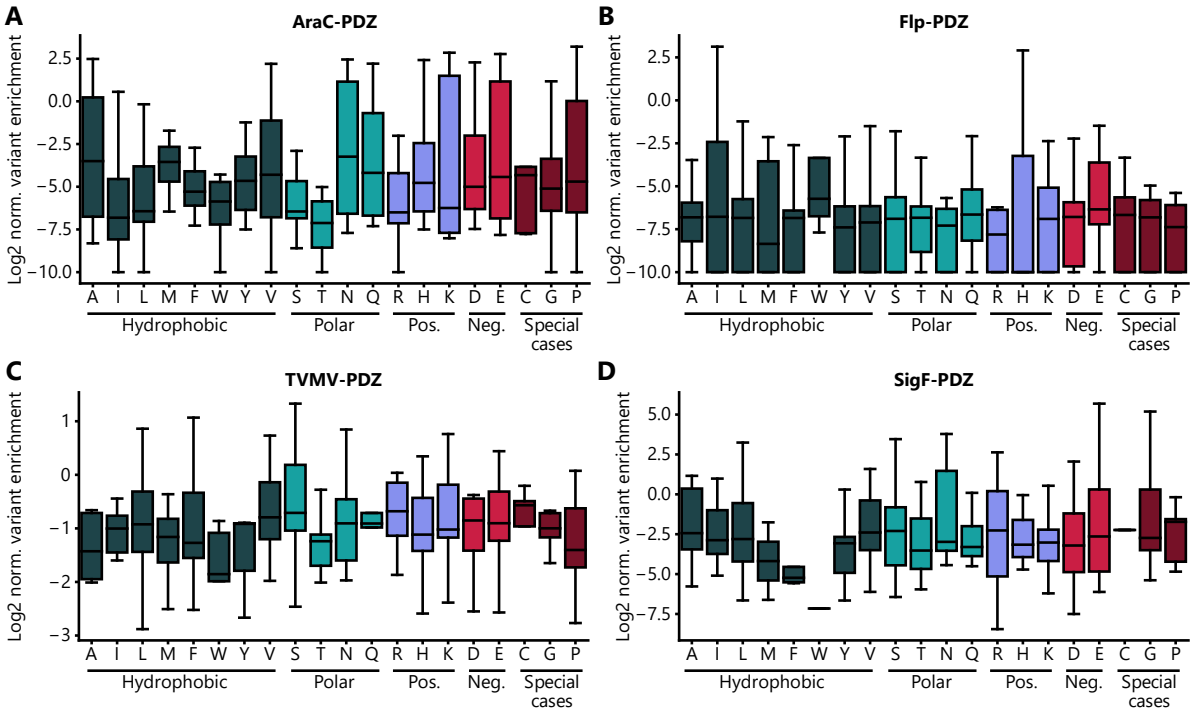


Figure 2.17: Successful domain insertion cannot be predicted from amino acid identity. (A-D) The enrichment score distribution for each amino acid is shown as boxplots for the PDZ libraries of AraC (A), Flp (B), TVMV protease (C) and SigF (D). Both residues neighboring an insertion site were taken into account for the calculations. The IQR is marked by the box and the median is represented by a line within the box. Whiskers extend to the 1.5-fold interquartile range (IQR) or to the value of the smallest or largest enrichment, respectively. Colors indicate the different amino acid categories as marked underneath the plots.

Surprisingly, these trends were, overall, very weak or completely absent in some libraries as indicated by spearman correlation coefficients around r=0.2.
Previous strategies for the design of hybrid proteins stressed the importance of unstructured loops as sites that frequently accept domain insertions. Thus, I continued the analysis by assessing the enrichment scores with respect their location within different secondary structure elements (Fig. 2.18B and Supp. fig. 8B). Surprisingly, the score distributions were very similar for all three major secondary structure elements. In case of most libraries, a bimodal distribution was visible, consisting of a large proportion of depleted variants and a smaller population of enriched candidates. Only for the AraC-PDZ library, a very weak trend towards higher enrichments in coils as compared to α-helices and β-sheets was observed (Fig. 2.18B). Interestingly, the different insert domains in combination with AraC did not result in major changes of the distributions. Although an enrichment in unstructured regions appeared to be absent, one has to take into account that there might be a tendency towards enrichment in structured elements at sites that are closer to loops, as it could be observed at some sites in AraC (Figure 2.14A). The definite confirmation of this trend, however, would require a larger dataset.
In order to obtain a more comprehensive overview over protein features that could affect domain insertion tolerance, I gathered a larger set of position-specific properties (Table 2.3.1).

Table 2.1: Position-specific properties included into the analysis of insertion tolerance.

Property	Description
ASA	Average surface accessibility
pLDDT	Position-wise pLDDT of AF2 models
Linker idx Suyama	Linker propensity index (Suyama & Ohara, 2003)
Linker idx George	Linker propensity index (George & Heringa, 2002)
Linker idx Bae	Linker index (Bae *et al*, 2005)
Hydrophobicity	Hydrophobicity (Prabhakaran, 1990)
Flexibility idx	Flexibility index (Bhaskaran & Ponnuswamy, 1988)
Molecular weight	Molecular amino acid weight
Average volume	Average amino acid volume
Positive charge	Positive charged amino acid
Negative charge	Negative charged amino acid
Net charge	Net charge of amino acid
Radius of gyration	Radius of gyration of the side chain
Side-chain stab idx	Side-chain contribution to protein stability (KJ/mol) (Takano & Yutani)
Buriability	Buriability (Zhou & Zhou, 2004)
KLD	KLD Kullback-Leibler divergence calculated from MSA
Insert frequency	Insert frequency at the respective position within the MSA
Deletion frequency	Deletion frequency at the respective position within the MSA
Mean ins len	Mean insertion length at the respective position within the MSA
Median ins len	Median insertion length at the respective position within the MSA

These features comprise a number of biophysical amino acid properties that were fetched from "AAindex", a database that curates a large set of diverse amino acid related statistics (Kawashima *et al*, 2008; Kawashima & Kanehisa, 2000). I also included three different linker propensity indices (Suyama & Ohara, 2003; George & Heringa, 2002; Bae *et al*, 2005). These describe to which extend amino acids tend to be present in inter-domain linkers. Regions with high linker propensities were supposed to be well suited for the insertion of domains, as will be discussed later (refer to section 3.4.4). Further, I included the pLDDT confidence score from AF2 models (Akdel *et al*, 2022; Tunyasuvunakool *et al*, 2021). It was previously shown that the

pLDDT correlates with intrinsically disordered sites (Akdel *et al*, 2022), which might tolerate insertions to larger extend, as compared to tightly structured regions. Finally, I also created MSAs for all four proteins and added features, such as the Kullback-Leibler divergence (KLD) as a measure of conservation. Additional scores with respect to the frequency of insertions and deletions were calculated from pairwise alignments between the protein of choice and all homologous sequences (see methods section 4.2.3 for details).

First, all pairwise Spearman correlations were determined for the AraC-PDZ dataset (Fig. 2.18C). Obviously, certain features, such as side-chain volume and molecular weight are by definition highly correlated, while others such as side-chain stability and flexibility must be negatively correlated. The pLDDT score and ASA were also anti-correlated because surface exposed elements can be more flexible and thus tend to have lower pLDDT scores. Also, the alignment-derived features corresponded to each other as expected. The appearance of insertions and deletions in the MSA correlated, but both features were negatively related to sequence conservation, i.e. KLD. Finally, it was striking, that no single feature showed a particularly strong correlation with the enrichment scores.

To investigate this phenomenon further, the same correlations were calculated for all datasets (Fig. 2.18D). Overall, no clear link between the enrichment scores and any of the features was found. Even the strongest correlation observed, i.e. relation between the AraC-LOV2 dataset and the position-wise insert frequency, was determined to have a Spearman's r of only 0.53. In most cases, the correlation coefficients were in the range between -0.2 and 0.2, indicating the absence of any clear interconnection. As the previous analysis already suggested, the results revealed, that no individual feature can explain the measured domain insertion tolerance.

When comparing the correlations for a certain feature from different datasets to one another, the overall tendencies were often in agreement (Fig. 2.18D). However, given the small values of Spearman's r, even a change in directionality of the correlation between two datasets should not be overinterpreted. In fact, only one slightly clearer global tendency was observed: The correlations of the sequence alignment derived features were slightly stronger, as compared to the amino acid specific, biophysical features, supporting the assumption that sequence context and conservation is more important than local information derived from a specific position alone, at least in context of domain insertion tolerance.

On a side note, I also considered a potential problem due to the fact that many biophysical properties were calculated on the basis of just the two residues, adjacent to the insertion site. As mentioned before, the introduction of a new domain must be seen in a larger context than just the neighboring residues. To this end, I also experimented with biophysical properties of larger patches, i.e. residues surrounding the respective insertion site, by averaging the features of residues within certain distances to the insertion. Also weighted averages, depending on spatial proximity were considered. This investigation did, however, not improve correlations or reveal stronger trends as compared to the simpler, residue pair-wise analysis.

Taken together, the analysis underlined the complexity of the problem and the lack of simple predictors for domain insertion tolerance.

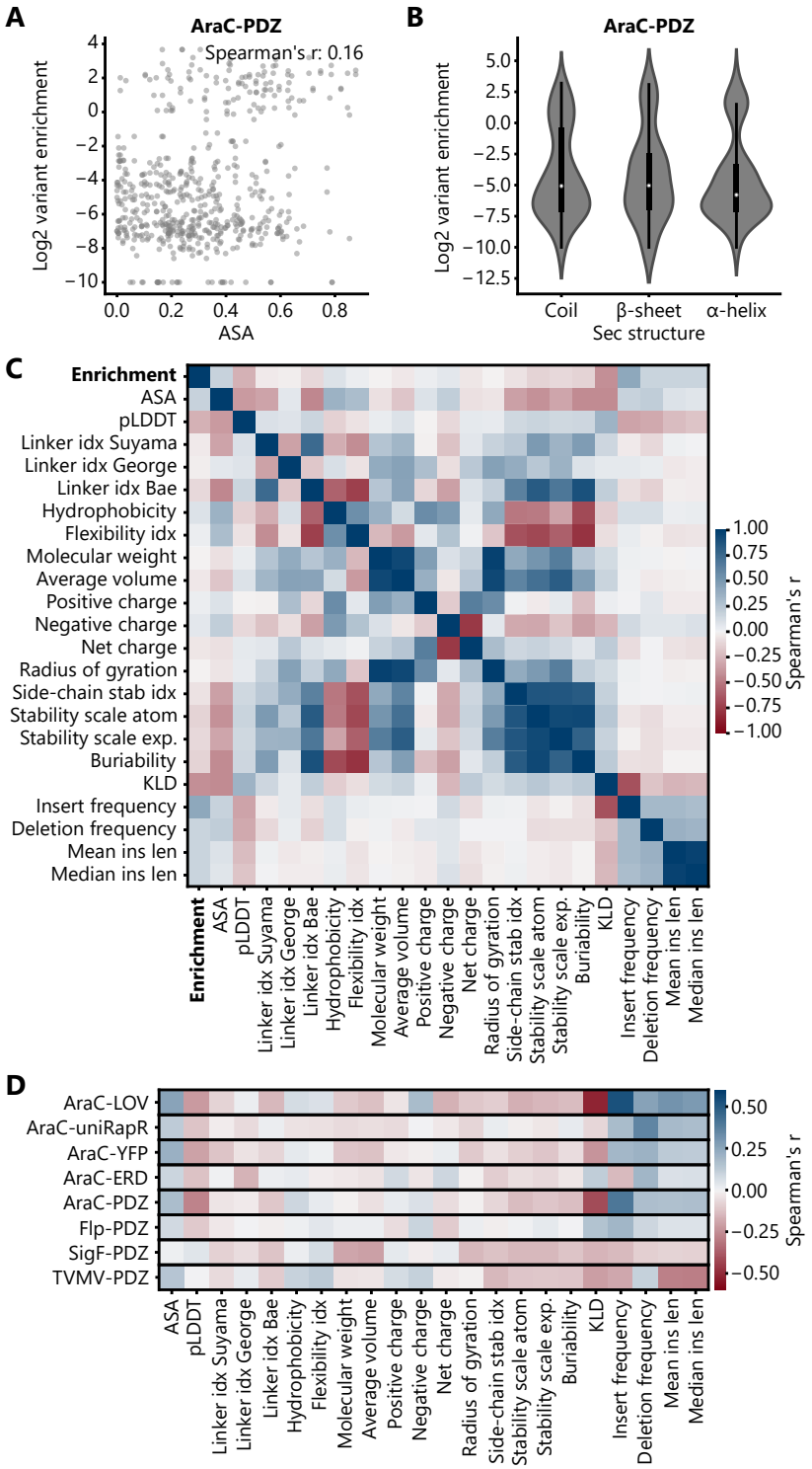


Figure 2.18: Secondary structure and amino acid features alone do not explain the observed preferences for domain insertions. (A) Correlation between variant enrichment and the average surface exposed area (ASA) of the residues neighboring an insertion site are plotted for AraC-PDZ. (B) The insertion score distribution with respect to different secondary structure elements is shown for the AraC-PDZ insertion library. For each insertion site, the secondary structure assignment of the amino acids prior and after the insertion were considered. The IQR is marked by the box and the median is represented by a white dot. Whiskers extend to the 1.5-fold IQR or to the value of the smallest or largest enrichment, respectively. (C) Pairwise correlations of all features are presented for the AraC-PDZ library as heatmap. (D) Spearman correlations between all datasets and diverse positional features are shown. (C, D) The last five features were calculated from sequence alignments. Linker idx: Different amino acid specific linker propensity indices that were reported by the indicated authors.

2.2.4 Comparing requirements for domain insertion tolerance to sites amenable to protein splitting

In the introduction, I pointed out that domain insertion is only one way to create switchable proteins. A different approach is the creation of split proteins, able to conditionally reassociate, as mediated by dimerization domains fused to its parts. At its core, this strategy is concerned with a similar question: How can sites be identified, at which a protein can be split into two parts, so that a re-constitution of the original fold is still possible, when both parts are fused to dimerizing domains? Despite the obvious differences, the selected sites must tolerate the presence of additional domains in both scenarios. A direct comparison that could reveal similarities between the approaches is therefore of interest. Conveniently, the Wilson Wong lab already reported a screen of several split candidates of the Flp recombinase (Weinberg *et al*, 2019). A comparison of their reported activities to the enrichment scores revealed that all enriched positions within the subset were also highly active split variants, while the majority of the depleted candidates corresponded to dysfunctional split-proteins (Fig. 2.19A). As a note of caution, only three of the reported variants were enriched in my screen, all of which carried the insertion in the same region close to the C-terminus. Furthermore, two of the depleted variants correspond to efficient split-protein designs.

Apart from experimental data, a computational model for the prediction of split sites has also been reported previously (Dagliyan *et al*, 2018). This model called "SPELL" calculates a "split energy", which is meant to indicate sites, which – when a protein is split at this position - would result in auto-reassembly of the corresponding N- and C-terminal protein fragments. However, when overlaying the SPELL predictions with the enrichment scores for Flp recombinase, no correlation was observed.

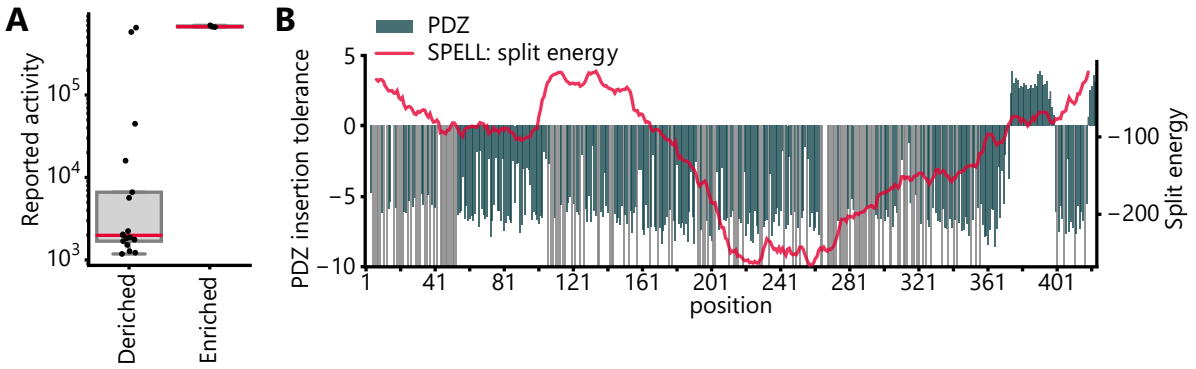


Figure 2.19: Domain insertion tolerance partially correlates with successful sites for split-proteins. (A) Boxplot showing the activity of inducible split-Flp recombinases reported by Weinberg et al. (Weinberg *et al*, 2019) for sites that were enriched or depleted in our domain insertion screening. Individual data points are shown. The IQR is marked by the box and the median is represented by a red line. Whiskers extend to the 1.5-fold IQR or to the value of the smallest or largest enrichment, respectively. (B) The Flp-PDZ enrichment histogram from Figure 2.11B is overlayed with the split-energy of the Flp recombinase for each position, as calculated by the SPELL algorithm (Dagliyan *et al*, 2018).

2.2.5 Assessment of AlphaFold2 structures in context of domain insertions

The field of protein structure prediction has recently seen major advances, most famously exemplified by Deepmind's AlphaFold2 (Jumper *et al*, 2021) and the Baker lab's RoseTTAfold (Baek *et al*, 2021). The applicability of these and similar models for protein engineering purposes is currently heavily explored (Jendrusch, 2021; Anishchenko *et al*, 2021; Pak *et al*,

2021; Akdel *et al*, 2022; Zhang *et al*, 2021; Dauparas *et al*, 2022). As discussed in section 1.2.1.1.1, the impact of the pLDDT score with respect to conformational flexibility and disorder is currently debated. Because structural flexibility is also of interest in the context of domain insertions, I analyzed the pLDDT scores of individual amino acids from an AF2 derived structure of wildtype AraC (Fig. 2.20). The observed values ranged from very weak prediction confidence of below 50 up to values close to the highest possible score of 100. While the sites with high insertion susceptibility were scattered across all pLDDT levels, positions with low insertion susceptibility were slightly enriched at higher pLDDT scores. The resulting correlation was, however, very weak with Spearman's r of -0.26.

Next, I aimed to investigate if the predictions of individual AraC-insert fusion exhibit stronger correlation to their respective enrichments. With this goal, AF2 structures of all possible PDZ insertions into AraC were predicted (Fig. 2.21A). To reduce the compute time to a minimum, only one model was predicted per candidate and a single "recycling step" was used. Note that conventionally, multiple models are predicted per input sequence and several "recycling steps" are employed to improve the accuracy of predicted models. First, the amino acid-wise pLDDT scores from each fusion variant were plotted as a heatmap for AraC (Fig. 2.21B). Of note, the heatmap does not include the PDZ domain insert sequences so that each column represents the same AraC amino acid in a different insertion context. The resulting representation allowed a direct investigation of the effect a certain insertion had on pLDDT values corresponding to the AraC amino acid sequence. Generally, the pLDDT scores tended to be high, >80 at most positions, indicating high quality predictions (Fig. 2.21B). Most prominent in the heatmap is a diagonal of decreased pLDDT values corresponding to the residues neighboring the respective position of the PDZ insertion. These lower values could implicate structural flexibility around the respective insertion site. The interpretation is backed by the fact that the unstructured loops of AraC are visible as vertical regions, also with decreased pLDDT scores. This observation is most pronounced for the inter-domain loop between the DBD and the LBD around amino acid position 170. The same applies to the protein's termini hinting towards their flexibility. Apart from that, the structure of the N-terminal β-barrel (AA 20-100) is implicitly visible in the heatmap by a symmetric pattern of locally decreased pLDDT scores in the upper left quarter (Fig. 2.21B). In summary, the pLDDT scores reflected structural features of AraC and potentially local conformational effects of insertions, albeit these findings remain speculative as this point.

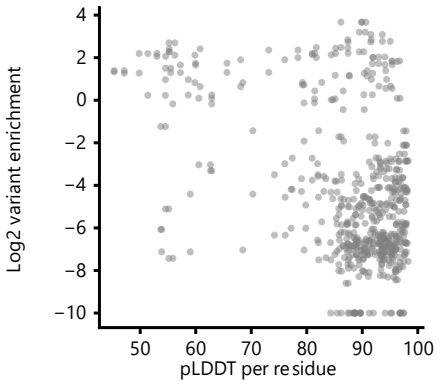


Figure 2.20: The position-specific pLDDT scores of wildtype AraC do not correlate with domain insertion susceptibility. Scatterplot of the relation between the enrichment scores of the AraC-PDZ library and the amino acid pLDDT scores from an AraC structure predicted by AF2. The corresponding Spearman's r is -0.26.

However, the pLDDT scores of specific AraC residues derived from the predictions of all different insertion variants did not correlate with the experimentally determined enrichment scores (Fig. 2.21C). In other words, changes of the confidence score of a residue with respect to the insertion site of PDZ did not correlate with the activity of AraC.

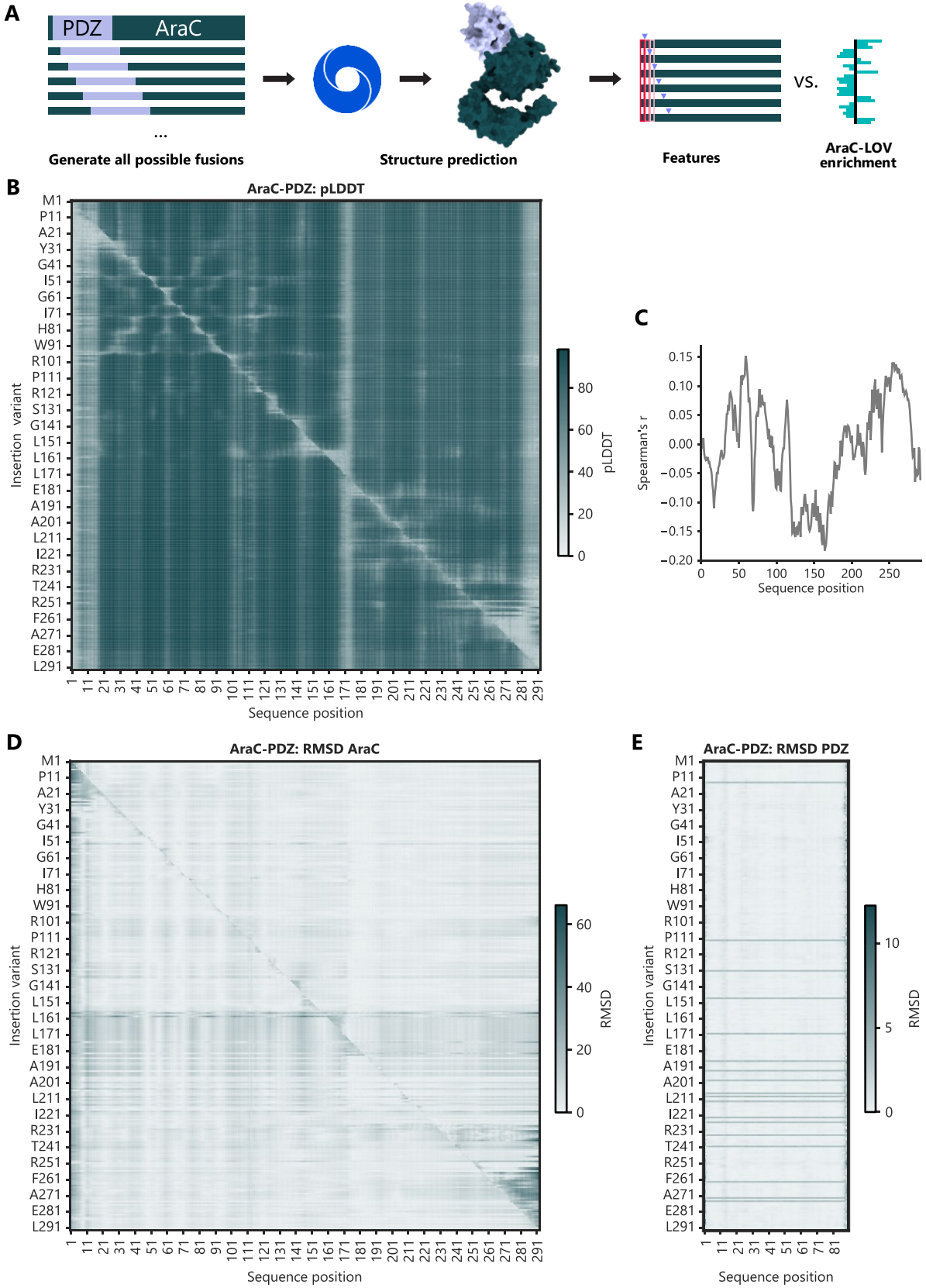


Figure 2.21: Correlations of structure predictions with domain insertion susceptibility.

Figure 2.21: Correlations of structure predictions with domain insertion susceptibility. (A) Depiction of the structure prediction workflow. Structures for all possible insertions of the PDZ domain into AraC were generated with AF2. Structural changes at single positions in response to different insertions were then compared and correlated to the experimental enrichments. (B) Structures of all possible PDZ insertions into AraC were predicted. The heatmap shows the pLDDT scores per position for each variant. Only AraC amino acids are depicted so that each column corresponds to pLDDT values from the same residue in different insertion variants. Rows, in turn, correspond to the different AraC-PDZ hybrids. (C) For each amino acid position, the pLDDT scores from all variants (columns in B) were correlated with the corresponding enrichment scores at these positions. The resulting Spearman correlation coefficients are shown. (D, E) The predicted AraC-PDZ structures were aligned to a predicted structure of wildtype AraC (D) or PDZ (E). The RMSDs between the wildtype and the respective part of the hybrid proteins are shown in the heatmap. Rows correspond to the different AraC-PDZ hybrids and columns to RMSD values of the same residue in different variants.

Apart from analyzing pLDDT scores, I also assessed the actual structure deviation between the hybrid proteins and the wildtype structures for all AraC-PDZ fusions. To achieve this, the structures of the individual AraC-PDZ hybrids were aligned to the structure of wildtype AraC and the distances between corresponding residues were measured. Comparison of these distances with respect to all hybrid proteins revealed a pattern that resembled the one of the pLDDT scores (Fig. 2.21D). A diagonal of local positional rearrangements around the respective insertion sites was observed, which was, however, less pronounced than the pLDDT score changes (Fig. 2.21B). Interestingly, no additional structural changes within the loop regions were visible. The predictions of the protein's termini, in turn, underwent considerable changes as compared to the wildtype conformation, when the domain was inserted nearby. Globally, the predicted structures shared high similarity, independent of the insertion site, with only very few outliers.

The situation was however different, when the predictions were compared to the structure of the PDZ domain, following the same procedure (Fig. 2.21E). In the majority of cases, the conformation of the domain was predicted to be identical to its wildtype structure independent of the site it was inserted at. In some AraC-PDZ hybrids, however, the PDZ conformation was predicted to be significantly distorted in comparison to the wildtype structure. These cases can be identified as dark green lines within the plot (Fig. 2.21E). Such conformational changes appeared more frequently, when the PDZ domain was inserted into the C-terminal DBD in contrast to insertions closer to the N-terminus. Interestingly, these sites were, again, not related to the experimentally observed patterns of domain insertion tolerance, suggesting no functional meaning.

Taken together, the exploration of state-of-the-art structure predictions suggested that AF2 is not able to capture the functional effects of domain insertions. The predicted structures of the hybrid proteins share high similarities with the individual structures of the insert domain or the parent protein. Nonetheless, AF2 predictions do reflect diverse structural features of AraC.

2.2.6 Machine learning models can guide the selection of sites susceptible to domain insertion

The results from the previous sections demonstrated the lack of clear indicators for domain insertion tolerance. Nonetheless, they also showed that very weak trends are indeed present. The logical follow-up question was, if a combination of the features analyzed above could enable the prediction of domain insertion tolerance. Towards this goal, machine learning models were trained on the entirety of the gathered insertion site properties in combination

with amino acid identity and secondary structure information as additional features. The main objective was to discriminate between sites that tolerated the insertion of a domain versus positions that resulted in inactive protein hybrids. Hence, the enrichment scores were binarized by assigning positions that were enriched during the screens to positive labels and vice versa the depleted sites to negative labels. Based on these datasets, I trained gradient boosting classifiers (Friedman, 2002) for each protein. This type of machine learning model is known to perform particularly well on tabular datasets (refer to methods section 4.2.4 for technical details) (Fig. 2.22). It was chosen after initial exploration of diverse classifier architectures. To evaluate the performance of a model, five-fold cross-validation was used.

The gradient boosting models reached surprisingly good performances on individual proteins ranging from a mean area under the receiving operator characteristic (AUROC) of 0.77 for SigF-PDZ (Fig. 2.22D) to 0.93 for Flp-PDZ (Fig. 2.22B). The corresponding average precisions ranged

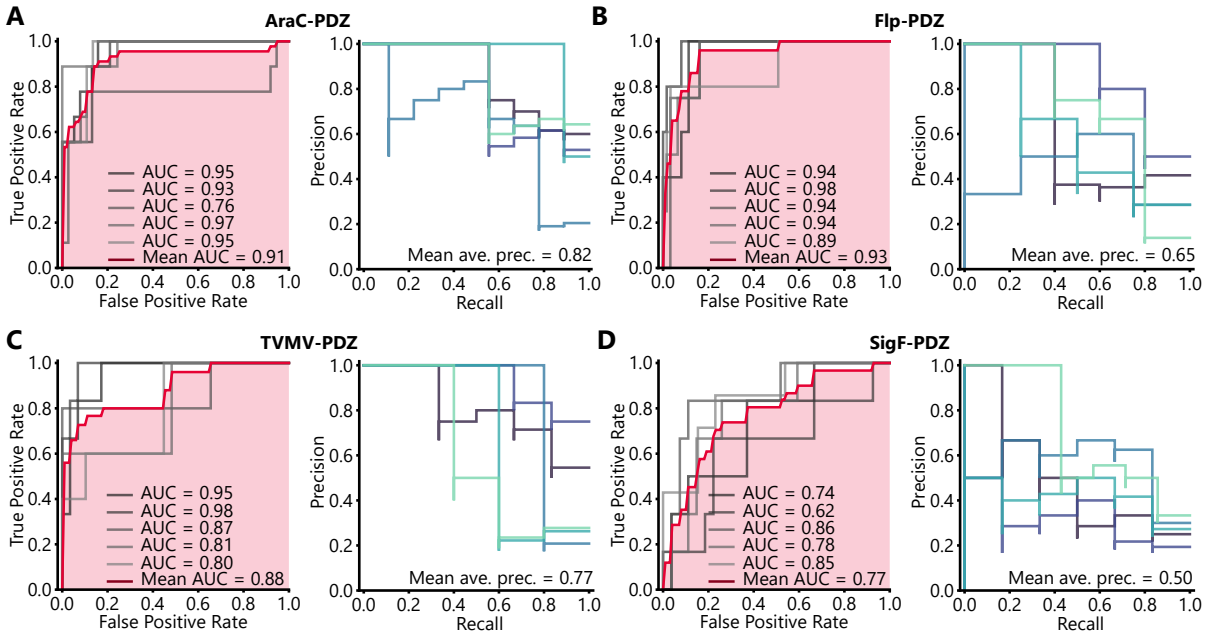


Figure 2.22: Gradient boosting models trained on positional features learn the insertion tolerance of individual proteins. (A-D) Models were trained on the PDZ datasets for AraC (A), Flp (B), TVMV protease (C) and SigF (D) with five-fold cross-validation. The ROC (left panel) is shown for individual folds in grey and the mean ROC in red. The mean AUC is marked in light red. Exact values are indicated. Precision-recall metrics for individual folds are shown in the right panels. The mean average precision is indicated.

from 0.5 for SigF-PDZ (Fig. 2.22D) to 0.82 in case of AraC-PDZ (Fig. 2.22A). With respect to weaker performance of the models for SigF-PDZ, it should be noted that the hyperparameters were not individually optimized for each protein (see section 4.2.4).

Encouraged by these results, I wondered whether a classifier could be trained on the combined dataset of hybrid proteins. Fortunately, the binarization during dataset preparation is expected to reduce batch effects between the enrichment scores of different proteins. The reason is that the main difference between the individual experiments was the varying enrichment stringency. While different stringencies can affect the absolute scores of all samples, only few candidates would be expected to switch from enriched to depleted or vice versa. The former effect, however, is negated by binarization, which, in turn, increases the comparability between the datasets of the different proteins. Optimizing the gradient boosting classifier on this complete training set resulted in a mean AUROC of 0.84 (Fig. 2.23A) and an average precision of 0.54

(Fig. 2.23B). These values were within the range of the previously trained models for the individual proteins.

To place these metrics into context, I compared them to several benchmarks. These included a dummy baseline, build on random choice of insertion sites, and the use of individual features as predictors (Fig. 2.23C, D, Supp. fig. 9). This analysis was performed on a separate test set,

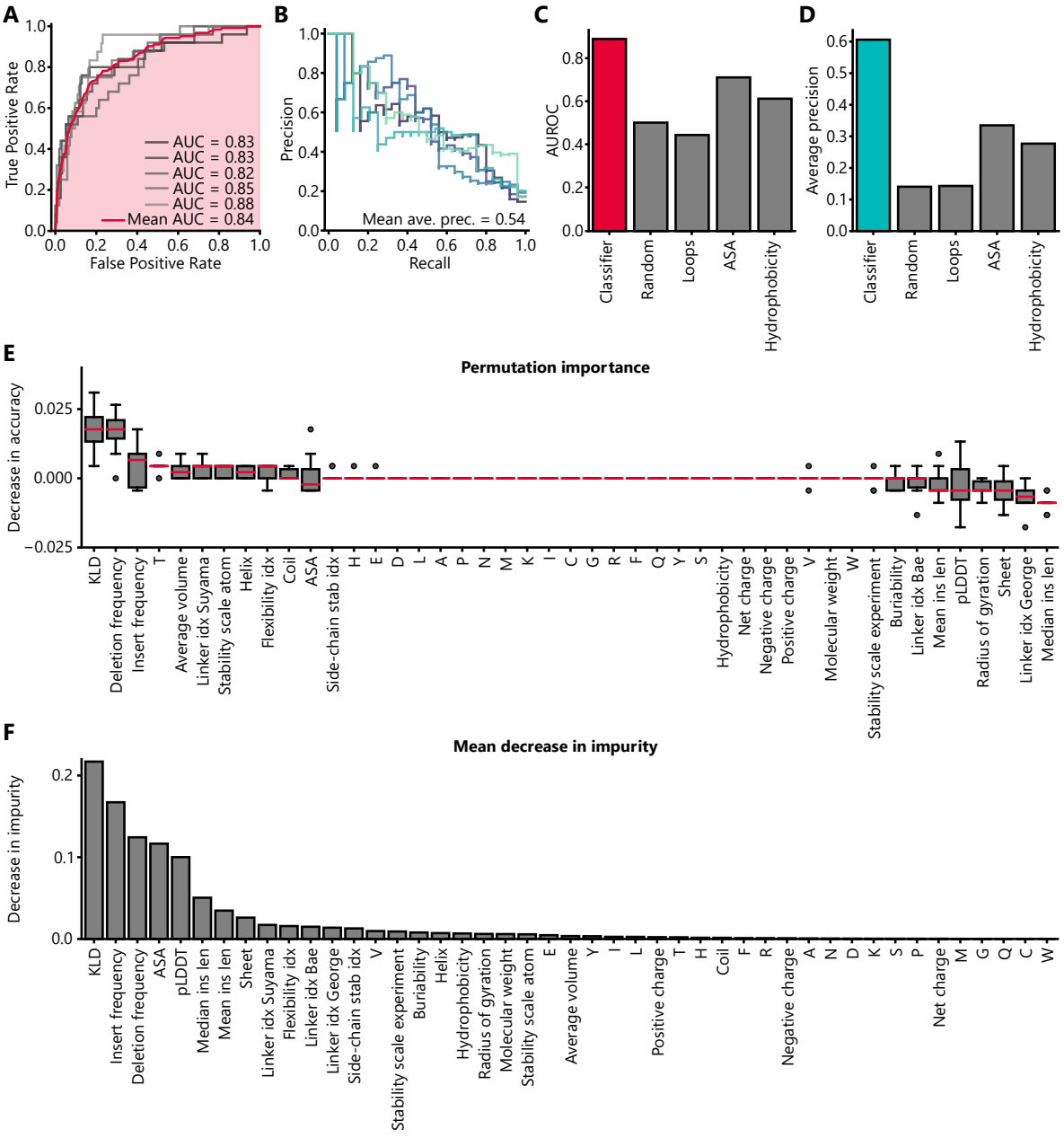


Figure 2.23: Gradient boosting models improve the prediction of domain insertion sites. (A, B) The model was trained on the combined PDZ datasets of all candidate proteins with five-fold cross-validation. (A) The ROC is shown for individual folds in grey and the mean ROC in red. The mean AUC is marked in light red. Precise values are indicated. (B) Precision-recall metrics for individual folds are shown. The mean average precision is indicated. (C, D) The AUROC (C) and average precision (D) of the trained model and different benchmarks are shown. The values were calculated from a previously withheld test set. The performance of the gradient boosting classifier is compared to a random baseline and several individual features. (E) The decrease in accuracy upon random permutation of the respective features is presented. The results were calculated individually for each fold in the cross-validation dataset. The IQR is marked by the box and the median is represented by a red line. Whiskers extend to the 1.5-fold IQR or to the value of the smallest or largest score, respectively. Outliers are shown as points. (F) Bar plot indicating the Gini importance of each feature.

withheld from the previously used cross-validation dataset the entire time. The random baseline had an AUROC and AP in the range of 0.50 and 0.16, respectively (Fig. 2.23C, D). Most of the individual features were also within this range or only slightly above, confirming again that they have no predictive power on their own. Only few parameters, including hydrophobicity and ASA, exhibited higher AUROCS and AP, with ASA being the best performing feature with values of 0,71 (AUROC) and 0.33 (AP), respectively. This result is in line with the previously discussed necessity of insertions not being tolerated at buried sites. My gradient boosting classifier showed a highly improved performance with an AUROC of 0.89 and an AP of 0.61, suggesting that the entirety of features implicitly provided information relevant for the successful prediction of domain insertion tolerance.

The remaining open question was, which features contributed the most to the predictions, i.e. carried the most essential information with respect to the prediction of domain insertion tolerance. To approach this question, the importance of individual properties for the model was assessed. Two complementing strategies were employed. First, the permutation importance of each feature was assessed, meaning that the loss in accuracy upon random permutation of the corresponding values was calculated (Fig. 2.23E). Second, the mean decrease in impurity (Gini importance) was measured, a method, which takes into account the rank of the decision nodes related to a feature (Louppe, 2015) (Fig. 2.23F). The results indicated, that the KLD and the frequency of insertions and deletions are the most important features, as they were top ranked by both methods. The only other two properties that were represented in the top ten most important features for both, the impurity and the permutation importance approaches, were ASA and the linker propensity index by Suyama et al. The majority of the other features did not seem to be of particular importance. The

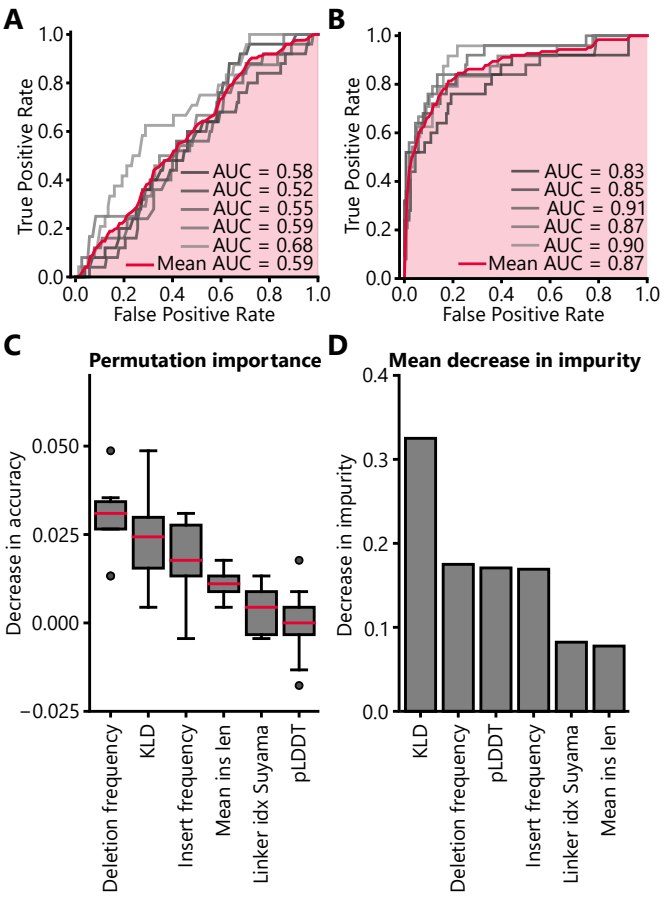


Figure 2.24: A set of six features determines the model's predictive power. (A) A gradient boosting model was trained exclusively on the amino acid identities. (B) The model was trained on a subset of features comprised of Deletion frequency, KLD, insert frequency, mean insertion length, the linker propensity index by Suyama (Suyama & Ohara, 2003) and the pLDDT score from AF2 structure predictions. (A, B) The ROC is shown for individual folds in grey and the mean ROC in red. The mean AUC is marked in red. Precise values are indicated. (C) The decrease in accuracy upon random permutation of the respective features are presented for the reduced model. The results were calculated individually for each fold in the cross-validation dataset. The IQR is marked by the box and the median is represented by a red line. Whiskers extend to the 1.5-fold IQR or to the value of the smallest or largest enrichment, respectively. Outliers are shown as points. (D) The Gini importance of each feature in the reduced model is shown.

permutation of some features even resulted in minimal gains of accuracy (Fig. 2.23E). In this context it should be considered, however, that the overall changes in accuracy were rather mild and in the range of only +/-0.03.
Nonetheless, the results raised the question if a subset of parameters might already contain all the information required for successful prediction of domain insertion sites. Given the accumulating evidence that amino acid identity is not an important feature, a model was trained on this information alone, which, as expected, did not "learn" much, indicated by a mean AUROC of 0,59 (Fig. 2.24A). As a consequence, amino acid information was removed from the training data. Next, additional features were depleted in a stepwise manner, always assuring that the performance of the model did not decrease upon feature removal. Following this procedure, I ended up with a reduced model, only trained on six features: KLD, deletion frequency, insertion frequency, mean insertion length, pLDDT and the linker index by Suyama et al. With an AUROC of 0.89 and an AP of 0.61, this model performed as good as the one trained on the dataset comprising all properties (Fig. 2.24B). Further reduction of features below the six above resulted in substantial decrease in model performance. Lastly, the feature importance analysis was repeated with the reduced model (Fig. 2.24C, D). Akin to the previous observations, KLD, insertion frequency and deletion frequency were detected as most important parameter explaining domain insertion tolerance.

2.3 Transcription control by optogenetic variants of AraC

2.3.1 Identification of light-switchable AraC variants

The last chapter dealt with the identification and prediction of sites that tolerate domain insertion. With respect to the design of switchable, allosteric proteins, however, this is only the first step. To gain a better understanding of which insertions actually mediate activity switching behavior and in order to identify optogenetic variants, I repeated the screen for the AraC-LOV2 library (refer to section 2.2.2), this time incubating the cultures under blue-light exposure instead of darkness prior to the sorting. Comparison of the previous enrichment in the dark (Fig. 2.25A) with the dataset enriched under illumination (Fig. 2.25B) revealed a high similarity between the resulting scores under both conditions. When only sites that tolerated the LOV2 insertion under at least on condition (light and/or dark) are included into a comparison, the differences between the datasets become more clearly visible (Fig. 2.25C). Here, only the dark-state replicates of these positions correlate well, while the light-induced replicate differs substantially. Thus, the observed similarities must mainly result from variants that were inactive under either condition. Indeed, when the enrichments from the light and dark experiments are plotted against each other (Fig. 2.25D), it becomes apparent that the majority of the hybrid proteins was inactive independent of light exposure. A smaller fraction appears to be constitutively active. Nonetheless, a remaining set of candidates exists that were enriched under one condition only and otherwise depleted. Interestingly, more variants could be found that are supposed to be active in the light and inactive in darkness than vice versa.
In order to obtain a better overview over the most promising, presumably switchable variants, the combined light and dark enrichment scores were plotted for the lead candidates (Fig. 2.25E). These variants were selected based on two requirements: One state had to exhibit an enrichment score >1, meaning that the corresponding AcaC-LOV2 fusion protein is supposed

to be highly active under one condition. Secondly, the difference between the light and dark scores had to be >2.5, hinting at a potent activity switch. The candidates fulfilling these prerequisites were located in four distinct clusters consisting of one to four individual variants. Interestingly, only one cluster represented hybrids, which seemed to be active in the dark, while all other variants were light-activated. Further, all members of each group exhibited the same switching directionality. The clusters all corresponded to positions within the N-terminal LBD or the linker region, while none was identified within the DBD.

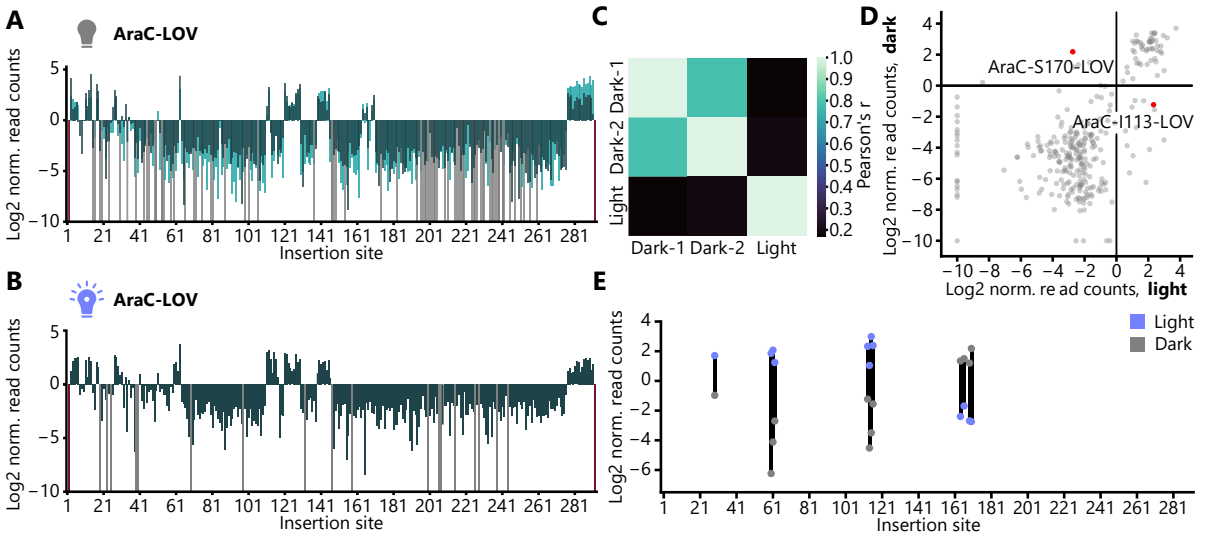


Figure 2.25: Domain insertion screening of an AraC-LOV2 hybrid library yields light-switchable variants. (A, B) The enrichment scores of AraC-LOV2 libraries that were sorted after incubation in darkness (A) or under blue-light exposure (B) are shown. The read counts per insertion variant obtained from NGS were normalized to read counts derived from the input library, resulting in log2-normalized enrichment scores. For the dark state, two biological replicates are shown. The scores for the light exposed sample resulted from a single experiment. Light green, dark green: individual replicates. Grey: variants with zero reads after enrichment. Red: variants missing in the initial library. (C) Pearson correlations between the different datasets are shown. Only positions of interest, that exhibited an enrichment in at least one replicate were included in the calculation. (D) Scatterplot showing the relation between the different enrichment conditions for each insertion variant. The two lead candidates are marked in red and their names are indicated. (E) Enrichment scores derived from experiments under light exposure or in darkness are marked by blue and grey points, respectively. Only datapoints from promising candidates with a log2 enrichment of at least one in the active state and a difference >2.5 between the light and dark states are shown.

2.3.2 Characterization of two potent light-switchable transcription factors

Having identified several promising candidate hybrids, I chose two of them for further characterization, one light-ON switch with the LOV2 insertion behind I113 (AraC-I113-LOV2) and a light-OFF variant carrying the insertion behind S170 (AraC-S170-LOV2). The performance of both proteins was assessed in joint work together with Sabine Aschenbrenner using the previously established RFP transcription reporter in *E. coli* (see section 2.2.1 for details). Together with wildtype AraC as a positive control and a strain expressing the TVMV protease as negative reference, the optogenetic hybrids were evaluated under a wide range of arabinose concentrations. Replicates under light exposure and in darkness were included for each concentration. The negative control underlined the exceptional tightness of the pBAD promoter with almost no leaky reporter expression detectable in absence of AraC (Fig. 2.26A). Wildtype AraC was completely inactive in absence of arabinose, while strongly activated gene expression was detected at increasing arabinose levels, reaching the full induction at 4 mM arabinose. Both AraC-LOV2 hybrids turned out to be still dependent on the presence of

arabinose. AraC-I113-LOV2 showed strong activity at high arabinose concentrations under illumination with 8 mM required for full activation. At this concentration, the RFP expression levels were only slightly below the corresponding values for wildtype AraC. In darkness, though, the expression levels were substantially lower at all tested concentrations. At very high arabinose levels though (16 mM and 25 mM), leaky reporter expression was detectable in the dark. The highest dynamic range of regulation was reached at a concentration of 4 mM arabinose, representing a 23-fold switch in activity. As expected, AraC-S170-LOV2 switched into the opposite direction, being completely inactive under light-exposure, while enabling RFP expression in the dark. The increase of RFP expression with rising arabinose concentrations turned out to be slower, as compared to the light-ON variant. The highest induction was reached at 16 mM arabinose. The slightly weaker activation came along with extremely low levels of leakiness under illumination. As a consequence, a 43-fold change in reporter activity was observed at 16 mM arabinose, indicating high performance of this optogenetic construct.

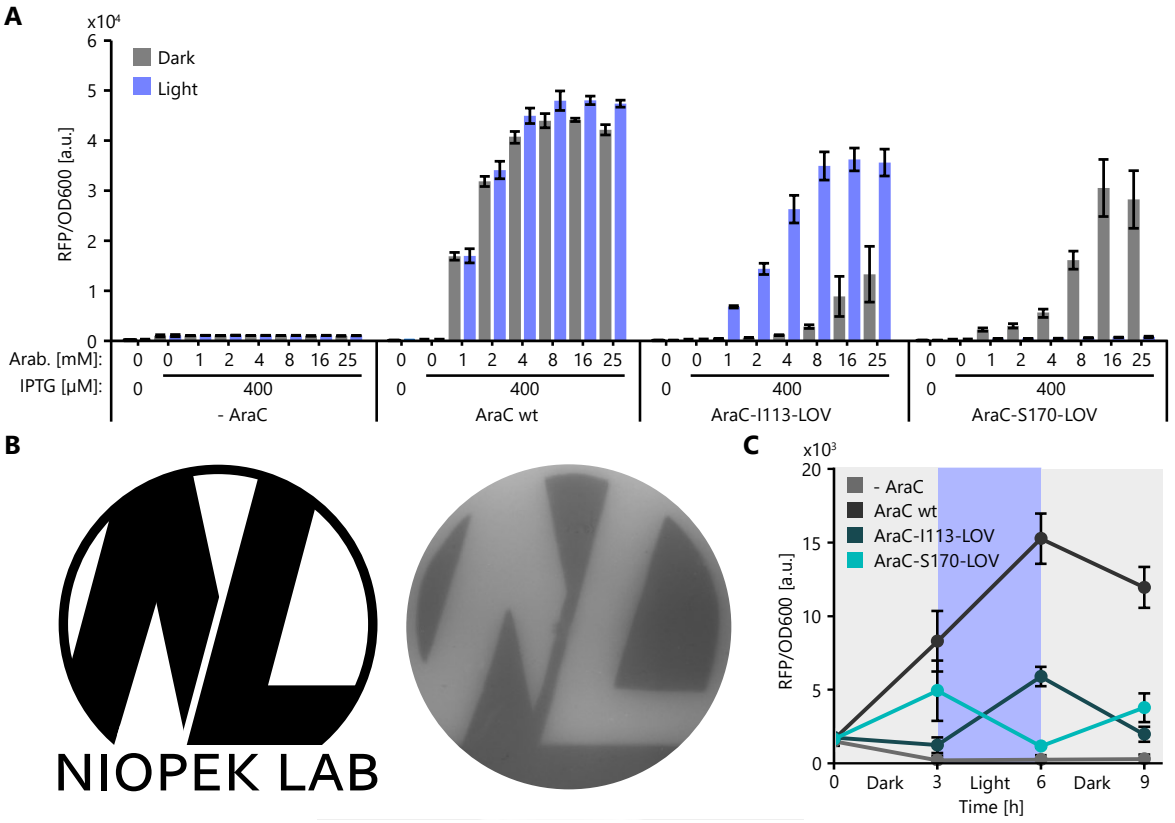


Figure 2.26: Optogenetic AraC variants mediate robust spatio-temporal gene expression control. (A) Cultures were inoculated from precultures carrying plasmids encoding an RFP reporter and the respective AraC variant. Inducers were added in the indicated concentrations. The samples were incubated for 16 h under light exposure or in darkness, followed by plate reader measurements of RFP fluorescence and the OD at 600 nm. The experiments were jointly performed with Sabine Aschenbrenner. Bars represent means from three independent biological replicates. Error bars show the SD. (B) Top agar mixed with inducers and bacteria carrying an RFP reporter plasmid and the AraC-S170-LOV2 variants was plated on an ager plate, which also contained arabinose and IPTG. The plate was incubated overnight, while being illuminated through a photo-mask of the logo on the left (without the text). A photo of the fluorescent signal from the RFP reporter was taken under UV-light exposure (right panel). (C) Cultures were inoculated with from precultures carrying plasmids encoding an RFP reporter and the respective AraC variant into media carrying 400 µM IPTG and 25 mM arabinose. The cultures were incubated either in darkness or under blue-light exposure. At the beginning of the experiment and every three hours from then, RFP fluorescence and OD600 were measured, followed by 1:30 dilution in fresh media. The experiment was performed together with Sabine Aschenbrenner. Points represent the mean of n=3 biological replicates. Error bars indicate the SD.

Having established the co-dependence of both systems on blue light, as well as arabinose, we next evaluated two of key advantages of optogenetic systems, their spatial and temporal resolution of control. Choosing AraC-S170-LOV2 for the test, I grew cells on an agar plate in presence of arabinose and IPTG (required for expression of AraC), while illuminating the plate through a photo-mask. When RFP expression was assessed under UV-light, the pattern of the mask was clearly visible with surprisingly sharp borders between fluorescent and dark regions (Fig. 2.26B). As a second experiment, we incubated liquid cultures of both variants in darkness or under illumination. Every three hours the light conditions were switched and the fluorescence was measured (Fig. 2.26C). The outcome shows the desired light-dependent increases and decreases of fluorescence. It is also visible that the wildtype AraC accumulated much higher RFP levels over time. Taken together, both results exemplify the versatility of this new optogenetic tool with respect to spatiotemporal control of gene expression in *E. coli*.

2.3.3 Structural analysis of the AraC-LOV2 hybrids

Apart from the experimental characterization, the structural aspects underlying the two optogenetic AraC-LOV2 fusion proteins were analyzed. Figure 2.27 shows the position of both insertion sites. Curiously, both sites are located in the region between the LBD and the DBD. This intermediate region consists of two parallel α-helices that are attached to the β-sheet core of the LBD. A longer linker connects them to the DBD. Both helices are oriented towards the dimerization interface. The light-ON insertion site (I113) is located within the loop prior to these two helices, relatively close to the LBD, while the light-OFF insertion site is occupying a position at the C-terminal end of the second helix. Both sites are distant, from functionally important residues so that the insertion would not be expected to permanently disturb protein activity. Finally, it is noted that both insertion sites are in proximity to larger loops and, thus, supposedly are flexible.

As mentioned before, the lead candidates, were members of clusters of switchable variants, each comprising four consecutive insertion sites. Figure 2.27 indicates that additional parts of

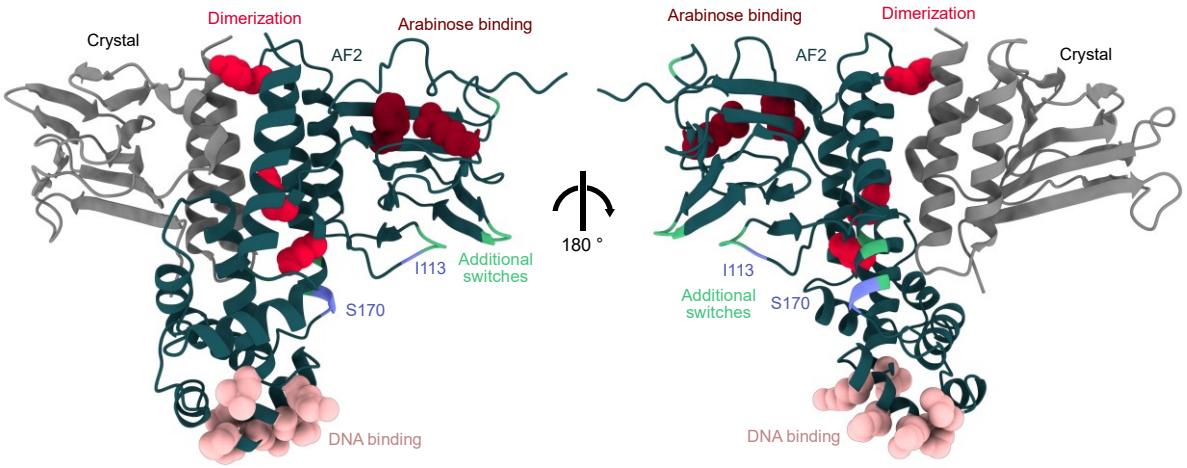


Figure 2.27: Optogenetic AraC variants carry the LOV2 domain in the linker region between the arabinose-binding domain and the DNA-binding domain. An AF2 prediction of the full-length AraC (green) is shown alongside the crystal structure (grey) of the arabinose binding domain. Relative positioning of the structures was obtained by superimposing the AF2 model onto crystalized dimer. Residues that bind to the operator are highlighted in pink, key residues for dimerization in red and the amino acids that are important for arabinose binding in vermilion. The insertion sites of the lead candidates are marked in blue. Additional insertion sites, corresponding to the switchable variants identified in Figure 2.25E are indicated in light green. PDB-ID: 2ARA.

the loop containing I113 were amenable to domain insertion resulting optogenetic switching. In case of the S170 region, the “switchable sites” extend into the neighboring α-helix, instead of the remaining part of the loop. The two additional clusters of switchable hybrids are located in the LBD, both within outward facing turns between two sheets of the arabinose-binding β-barrel. Based on these observations one could argue that in contrast to positions with general insertion tolerance, switchable candidates indeed carry insertions preferentially within surface-exposed loops. However, this observation is currently limited to the AraC protein.

3 Discussion and outlook

The unifying theme of this study was the engineering of proteins by the insertion of exogenous domains. The utility of this strategy was first exemplified for Acrs. By employing domain insertions to improve the stability of the *Nme*Cas9 inhibitor AcrIIC1, we created a highly potent off-switch for this RNA-guided nuclease. Using a different inhibitor, AcrIIC3, we further created a system for the light control of *Nme*Cas9 activity, by inserting the *As*LOV2 domain into the Acr. This design enabled the activation of gene editing by the illumination of cells with blue light.

Motivated by the experiences in the context of Acrs, we conducted a larger domain insertion screen, with the aim to further investigate the restrictions underlying insertions into diverse proteins. Based on a dataset derived from the screening of four different proteins, computational analysis revealed that simple biophysical features do not explain the observed distribution of successful insertions. Instead, machine learning models, based on a combination of features, most of which were derived from sequence alignments, were able to improve the prediction of domain insertion tolerance.

Finally, two potent optogenetically controlled variants of AraC were engineered. Both proteins harbored an insertion of the *As*LOV2 domain. Surprisingly, they differed with respect to the directionality of the light switch, meaning that one candidate, AraC-I113-LOV2, was activated by light, while the other variant, AraC-S170-LOV2 was light-deactivated. The further characterization of these constructs revealed their potential with respect to the spatiotemporally confined activation of transcription.

In this chapter, I am going to discuss the results obtained during these studies in context of related work. A particular focus will lie on the engineering strategies that have been developed over the years and how they compare to our examples and the results from the insertion screen. Finally, I will give an outlook with respect to future perspectives and challenges.

3.1 Improving the inhibition potency of AcrIIC1

3.1.1 Reasons for the increased inhibition potency

The work on AcrIIC1 was motivated by reports showing that the inhibitor has broad-spectrum activity, i.e. is able to inhibit several, biomedically relevant Cas9 orthologues, but with low efficiency (Harrington *et al*, 2017; Garcia *et al*, 2019). In previous work, we have already engineered different optimized AcrIIC1 variants: AcrIIC1X was a redesigned version of the inhibitor with improved potency on *Sau*Cas9; The chimera-7 and -10 were engineered by the insertion of mCherry, which increased the inhibition potency on *Nme*Cas9 and *Nme*2Cas9. Interestingly, the insertions had no effect on the inhibition of *Sau*Cas9 (Fig. 2.1 and Mathony *et al*, 2020a). Vice versa, AcrIIC1X exhibited increased performance only on *Sau*Cas9, but not on *Nme*Cas9 (Mathony *et al*, 2020a). The combination of both approaches in AcrIIC1X*, finally, resulted in an additive effect leading to the potent inhibition of all orthologues (Fig. 2.1 and Mathony *et al*, 2020a). The explanation for these outcomes lies in different mechanisms of actions. AcrIIC1X, for instance, was engineered with the aim to reach increased affinity to

*Sau*Cas9. Consequently, one would expect a decreased inhibition of *Nme*Cas9 due to mutations suboptimal for the interaction with this orthologue. Indeed, our collaborators measured a reduced affinity between AcrIIC1X and the HNH domain of *Nme*Cas9, as compared to the wildtype inhibitor (Mathony *et al*, 2020a). Interestingly however, my data obtained from gene editing experiments in human cells did not confirm these results, since no difference in inhibition potency could be measured between AcrIIC1 and AcrIIC1X (Mathony *et al*, 2020a). It is further unlikely that the differences were masked by saturated inhibition due to high inhibitor expression levels, considering the mild effect AcrIIC1 had on *Nme*Cas9 under the chosen experimental conditions. Thus, the mild reduction in affinity of AcrIIC1X for *Nme*Cas9 as compared to wild-type AcrIIC1 did not negatively impact its ability to inhibit *Nme*Cas9 *in vivo*.
More surprising were the results with respect to the AcrIIC1-mCherry chimeras. Together with Carolin Schmelas, I had already shown that the insertion of mCherry results in an increase of protein stability (Mathony *et al*, 2020a). Additional *in vitro* data from our collaboration partners Yanli Wang and Sun Wei indicated, that the domain insertion might lead to a slight increase of cleavage inhibition, independent from protein stability (Mathony *et al*, 2020a). Taken together, we expect the stabilizing effect of the additional mCherry domain to be the main driver of increased inhibition (Fig. 2.1A), while a mechanistic improvement could still play a minor role. The reason for this gain of function might lie in AcrIIC1's mechanism of action. By binding to the HNH nuclease domain, the inhibitor prevents the conformational change that brings the domain into proximity of the bound target DNA (Harrington *et al*, 2017). The presence of the fused mCherry domain might have increased this steric block and practically "locked" the Acr-Cas9 complex in a stable, inactive conformation.
While the inhibition of *Nme*Cas9 by the AcrIIC1-mCherry chimeras met our assumptions, the absence of inhibition improvements regarding *Sau*Cas9 was not expected (Mathony *et al*, 2020a). In principle, the increased stability of AcrIIC1 should have equally affected both target nucleases. Although the functionality of AcrIIC1 particularly with respect to *Sau*Cas9 has never been separately investigated, the nucleases high degree of structural and functional similarity to *Nme*Cas9 renders differences in the working mechanism rather unlikely. It is thus reasonable to assume, that the weaker affinity of AcrIIC1 (and its chimeras) to *Sau*Cas9 as compared to *Nme*Cas9 explains the lack of improvement. Perhaps, the increase of available inhibitor caused by the stabilization did not outweigh the low affinity AcrIIC1 has towards *Sau*Cas9. It can further not be ruled out that the small stability-independent increase in cleavage inhibition observed in *in vitro* experiments with *Nme*Cas9, is indeed specific to the precise structural configuration of this orthologue and does hence not translate to *Sau*Cas9. In summary, the observed effects appear to be predominantly caused by an increase in stability of the inhibitor and differences with respect to binding affinity. Certain aspects of the explanation, however, remain to be experimentally confirmed.
Finally, the fact that both approaches could be successfully combined in AcrIIC1X* (Fig. 2.1B) (Mathony *et al*, 2020a) provides additional evidence that the gains in inhibition potency did not come at the cost of inhibition efficiency on different orthologues. AcrIIC1X* seemed to be indistinguishable from the strongest inhibitor of *Nme*Cas9, AcrIIC3, with respect to inhibition potency, thus underlining the success of the engineering approach (Fig. 2.1B and (Mathony *et al*, 2020a).

3.1.2 The potential of engineered Acrs in comparison to natural inhibitors

The results discussed above demonstrated that protein engineering can be used to improve Cas9 inhibitors and that potent inhibition of several Cas9 orthologues by a single inhibitory protein is possible (Fig. 1.6 and Fig. 2.1). A remaining question is, to which extend such approaches will be required in the future to complement the increasingly rich repertoire of newly discovered and well-characterized, natural Acrs (Dong *et al*, 2018). In this regard, one has to take into account some of the inherent Acr characteristics. Anti-CRISPRs appear to be a common protein class in bacteriophages and the diversity of phages is still in large parts uncharacterized (Pawluk *et al*, 2017). In addition, the identification of new Acrs is non-trivial as they lack sequence conservation (Borges *et al*, 2017). Consequently, the discovery of new natural inhibitors is still cumbersome. With respect to Cas9 inhibition, AcrIIC1 was the first described broad-spectrum inhibitor (Harrington *et al*, 2017). Around the time we reported the engineered AcrIIC1 variants, a second promiscuous inhibitor, AcrIIA5, was characterized in two independent publications (Garcia *et al*, 2019; Song *et al*, 2019). Both reports showed that this inhibitor blocks the activity of a broad range of Cas9 orthologues, including the widely applied *Spy*Cas9, which is not targeted by AcrIIC1. Surprisingly, the publications differed in their mechanistic explanation of Cas9 inhibition by AcrIIA5. Song et al. reported inhibition of only DNA cleavage, but not DNA binding to be the driving mechanism, much alike AcrIIC1. However, they suspected AcrIIA5 to bind the second catalytic domain, RuvC, instead of the HNH domain (Song *et al*, 2019). Garcia et al. instead postulated a catalytic mechanism that involves sgRNA cleavage mediated by the inhibitor (Garcia *et al*, 2019). None of these postulated mechanisms has, however, been convincingly proven. From the application perspective, knowing the exact mechanism-of-action is highly relevant though. The catalytic mechanism, for instance, would result in a permanent loss of Cas9 RNP function, meaning that new ribonucleoprotein complexes have to be produced in order to reactivate gene editing. The steric inhibition, in turn, would only depend on the presence of the inhibitor. In summary, it remains unclear if AcrIIA5 is mechanistically similar to our engineered variants and could serve as a direct alternative to them or if it governs a second functionally complementary mechanism.

With respect to the full CRISPR application spectrum, AcrIIC1 and its derivatives, are restricted to the regulation of gene editing, as they only inhibit DNA cleavage (Harrington *et al*, 2017). The inhibition of CRISPRi or base editing would thus not be possible on the basis of AcrIIC1. Unpublished data from our group suggests that AcrIIA5, in contrast, can indeed inhibit CRISPRi and thus Cas9 DNA binding, hence favoring the postulated AcrIIA5 catalytic mechanism. However, the experiments revealed a limited inhibition potency that further depends on the targeted Cas9 orthologue. In this regard, our engineered AcrIIC1 variants would provide a more potent alternative, although a detailed side-by-side characterization is still needed.

Overall, the set of available inhibitors is limited and the current options differ greatly with respect to efficiency, mechanism of action and the range of targeted orthologues. Especially for the application of Acrs in genetic circuits (Nakamura *et al*, 2019), the discussed factors could have major impact on experimental outcomes. Thus, it appears plausible that a combination of the continued search for broad-spectrum inhibitors and the rational engineering focused on the precise tuning of desired properties will remain as complementary strategies for the foreseeable future.

3.1.3 Towards cell type-specific control of gene editing

The hepatocyte-specific gene editing technique enabled by global Cas9 expression in combination with AcrIIC1X, expressed in all, but the cell type of choice, demonstrated a promising use case for engineered CRISPR inhibitors (Figu. 2.2). It further exemplified how important the inhibition potency can be with respect to experimental outcomes and Acr applications. The miRNA-based control was assessed in a HEK293T system using miRNA overexpression (Fig. 2.3A, B), as well as in different cell types, where the observed effect depended on the endogenous levels of the respective miRNA trigger, miR-122 (Fig. 2.3A, B).

First, these experiments provided additional evidence that miRNA-regulated gene editing, which was initially developed by Hoffmann et al., is indeed a plug-and-play system that can be used in combination with different Cas9 and Acr variants (here *Sau*Cas9 and AcrIIC1X) (Hoffmann *et al*, 2019). Second, my data revealed facets of the system that are important to consider for future experiments. Very clearly, the data showed that the miRNA dose matters. In the overexpression experiment, in which the miRNA concentration was not the limiting factor, the Acr-encoding mRNA appeared to be fully degraded when the required miRNA binding site was present in the 3'-UTR (Fig. 2.3A, B). Consequently, the gene editing activity in AcrIIC1X-miR-122 samples did not differ from the control without an Acr expressing construct. In Huh7 cells, in which the system relied on endogenous miRNA-122, the observed effects were qualitatively similar, albeit a slightly decreased editing rate was measured in AcrIIC1X-miR-122 samples as compared to the control lacking the inhibitor (Fig. 2.3C). This observation confirmed that endogenous miRNA levels can be sufficient to activate gene editing via miRNA-regulated Acr transgene suppression. At the same time, it underlined the requirement for strongly expressed miRNAs, such as miR-122, to avoid incomplete Acr suppression and hence further reduction of gene editing activity in ON-target cells.

Another factor determining the success of the approach is the background expression level of the chosen miRNA. It was visible in HEK293T cells that AcrIIC1X represses gene editing more strongly when no miR-122 binding sites are present within its 3'-UTR (Fig. 2.3A, B). Residual levels of miR-122, naturally expressed by HEK293T cells, or effects of the altered 3'-UTR on the expression of AcrIIC1X (as compared to the control with a scaffold UTR) are possible explanations for the observation. Of note, a complete absence of the chosen miRNA in all off-target cell types cannot be expected. Even more important appears to be the choice of the inhibitor. Wildtype AcrIIC1 was much more affected by the described phenomena than the engineered AcrIIC1X (Fig. 2.3A-D), the latter of which could robustly suppress unintended editing in off-target cells.

In summary, the results demonstrated that gene editing in off-target cell types can be drastically reduced using our miRNA-dependent Acr transgene strategy. Further, our experiments revealed that the inhibition strength of the Acr is indeed an important factor to consider for prospective applications of this strategy.

3.1.3.1 *The role of methods for the detection of gene editing frequencies*

In the results section, I have already mentioned differences between T7E-assays and TIDE sequencing for the quantification of indels (section 2.1.1 and 2.1.2). Both measurement methods are still widely applied, although the NGS-based assessment of editing rates became increasingly popular over the last years. The advantages of the T7E-assay and TIDE are their

low costs per sample, speed and overall good reliability, especially if clear differences in editing rates should be distinguished, as is the case for many of the datasets presented in this thesis (Brinkman *et al*, 2014).
In the context of this study, the miRNA-related experiment provided a side-by-side comparison between both methods (Fig. 2.3A, B). Qualitatively, the results were similar, but the absolute numbers, i.e. indel %, differed between TIDE and T7E-assay. These differences were most prominent at intermediate editing rates. While the highest and the lowest values were nearly identical, editing efficiencies in the range of 10-20 % as measured by T7E-assay corresponded to lower values in TIDE sequencing (Fig. 2.3A, B). This effect was especially pronounced between the AcrIIC1-scaffold samples in Figure 2.3A and B. In addition, I note that many experiments in our lab over the years have shown that very low editing frequencies are often not detectable by TIDE sequencing, while still being visible in T7E-assays. The same trend became apparent in the miRNA-regulation experiments (Fig. 2.3A, B).
Importantly for the results of this project, most of the measured samples, exhibited a rather binary outcome, that is, efficient gene editing in controls versus complete inhibition of DNA cleavage by the improved Acrs. Consequently, differences in absolute values detected between T7E-assays and TIDE sequencing data do not affect the interpretation of the results. This is also an encouraging sign for the validity of experiments, in which only one method could be used.

3.2 Characterization of CASANOVA-C3

3.2.1 Performance of CASANOVA-C3

This project part transferred the concept of optogenetic Acrs from the *Spy*Cas9 inhibitor AcrIIA4 (Bubeck *et al*, 2018) to AcrIIC3, an anti-CRISPR protein specific to *Nme*Cas9. The successful implementation of CN-C3 enabled, for the first time, the direct optogenetic control of *Nme*Cas9 (Fig. 2.4) and *Nme*2Cas9 (Fig. 2.5D, E). This renders CN-C3 the first tool for the reversible light-mediated activation of type II-C CRISPR nucleases.
As previous work already indicated (Bubeck *et al*, 2018), the performance of CN-C3 is a question of dosage (Fig. 2.4A, B). Consequently, in order to identify the ideal dynamic range, the ratio between Cas9 and CN-C3 has to be chosen carefully. In addition, the results showed that a complete activation of gene editing in the light, as well as full inhibition in the dark, are difficult to achieve (Fig. 2.4A, B). The underlying reason is the nature of the *As*LOV2 photoswitch. The LOV2 domain exists in an equilibrium between its light- and dark-adapted states. Illumination substantially shifts this equilibrium towards the light-adapted conformation, but a completely binary behavior cannot be expected (Yao *et al*, 2008). The result is a natural limit to the dynamic range of engineered photo-switches. Moreover, the insertion into the Acr can also be suspected to result in physical strain on the domain towards a certain state, thus affecting the conformational ensemble (Dagliyan *et al*, 2019, 2016). This conformational stress might also reduce the efficiency of the light switch as compared to the natural photo-switching behavior of the *As*LOV2 domain.
With respect to the dynamic range, it is important to note that our group could show for several effector proteins an increase of switchability, when optogenetic Acrs were delivered by AAVs (Bubeck *et al*, 2018; Hoffmann *et al*, 2021). Higher transduction efficiencies in contrast to transient transfection likely explain this gain in performance. The efficient AAV delivery ensures

that almost every cell expresses similar amounts of Cas9 and the inhibitor. In consequence, the results in Fig. 2.4 might slightly underestimate the performance of CN-C3, as cells that were not co-transfected with both, the Acr and the Cas9/sgRNA constructs, might have been edited irrespective of the light condition due to the absence of the inhibitor. The opposite effect, namely no editing caused absence of Cas9 might have also played a role with respect to the upper limit of gene editing in the light condition.

Considering the specificity of genome editing, we showed that the use of CN-C3 can decrease off-target effects. Noteworthy, our group and others have reported various strategies to employ Acrs for enhancing target specificity by attenuation or timely control of Cas9 activity (Aschenbrenner *et al*, 2020; Shin *et al*, 2017; Liang *et al*, 2020). In case of *Nme*Cas9, the situation slightly differs from other use cases. Due to the fact that this nuclease already possesses a high target specificity (Amrani *et al*, 2018), I could demonstrate the off-target reduction for one sgRNA only (Fig. 2.4C-E). As the off-target editing was successfully reduced, *Nme*Cas9 in combination with CN-C3 could potentially be used as a gene editor with exceptionally high specificity. Nonetheless, convincingly demonstrating the superiority of the combination over comparable systems will require the examination of off-target reduction at a larger scale and ideally in an unbiased fashion using whole genome off-targeted detection methods. A future study, investigating these aspects could be of great value for the gene editing community, especially if performed under consideration of gene therapy applications. In this area of research, the minimization of off-targets is given particularly high priority and even incremental improvements could potentially have a major impact with respect to therapy outcomes and treatment acceptance.

When comparing CN-C3 to the AcrIIC1-mCherry chimeras, the differences observed on the corresponding protein levels are rather striking. We have previously shown that AcrIIC1, as well as the engineered chimeras, destabilize *Nme*Cas9 (Mathony *et al*, 2020a). Neither AcrIIC3, nor CN-C3 exhibited this feature under my experimental conditions (Fig. 2.5B, C). In both cases, Cas9 protein levels were identical when the wild-type AcrIIC3 inhibitor and its engineered variants were present (Fig. 2.5B, C) (Mathony *et al*, 2020a). The AcrIIC1 chimeras were significantly stabilized by mCherry insertion (Mathony *et al*, 2020a), while the LOV2 insertion into AcrIIC3 did not affect protein stability (Fig. 2.5A, C). It was principally not surprising, that domain insertions could have different effects on protein stability. Nonetheless, the results stress the importance to consider such effects, especially when the mechanism of action of a protein is not entirely understood.

3.2.2 Structural analysis of the *As*LOV2 insertions into AcrIIC3

In the introduction, I already highlighted the difficulty to identify suitable insertion sites (refer to section 1.3.4.5) and in the second part of this study, I focused on exactly this challenge. It makes thus sense to analyze CN-C3 from a structural perspective. To engineer its predecessor CASANOVA, a fusion of *As*LOV2 and AcrIIA4, Bubeck et al. successfully followed the principle of insertions into tight loops as it was proposed by Dagliyan et al. (Bubeck *et al*, 2018; Dagliyan *et al*, 2016, 2019). The rationale is described in detail in the introduction (refer to section 1.3.4.2). In short, the concept suggests the insertion of domains into loops, which connect aligned secondary structure elements. At these sites, the insertion can have the largest conditional impact on the structure of the target protein.

CN-C3 carries the LOV insertion at a surface site with unexpected properties (refer to section 2.1.3.4 and Fig. 2.6B). Based on structural considerations, one would probably avoid this site for the insertion of whole domains as a permanent loss of function appears to be likely. Indeed, the CN-C3 hybrid was originally created without this information, prior to the publication of AcrIIC3 structures. On the contrary, other insertion sites suggested by the engineering approach from Dagliyan et al., resulted in dysfunctional hybrids, presumably due to steric clashes with the *Nme*Cas9 binding partner (Fig. 2.6A) (Dagliyan *et al*, 2016, 2019). From the insertion strategy standpoint, the configuration of CN-C3 pronounces the limited understanding we still have with respect to the structural requirements of successful insertions. Existing strategies are not generalizable across protein classes and can be very successful for one candidate (AcrIIA4), while being not suited for the other (AcrIIC3).
Finally, the analysis of the AcrIIC3 structure in complex with the full-length *Nme*Cas9 (Supp. fig. 3 and Sun *et al*, 2019) point towards another aspect. Switchable proteins based on the insertion of LOV2 domains are typically supposed to be regulated allosterically by conformational changes the LOV2 domain imposes onto the fused protein. However, in the case of CN-C3, two additional mechanisms are possible. Due to the proximity of the insertion site to residues, which are in contact with Cas9, small local conformational changes of the residues directly connected to the LOV2 domain's terminal α-helices might already be sufficient to control AcrIIC3's activity (Fig. 2.6B). Long-range allosteric effects could then be neglected. However, considering the larger number of inter-residue contacts, AcrIIC3 makes with different parts of Cas9 (Supp. fig. 3), this explanation remains debatable (Sun *et al*, 2019). A second option is the steric inhibition of HNH domain-binding via the LOV2 domain. Rather than affecting the inhibitor's conformation, a light-induced change of the LOV2 domains position relative to AcrIIC3 could sterically block access of the binding interface to the HNH domain. Interestingly, the structural models of CN-C3 in supplementary figure 2 and 3 would also support this hypothesis.

3.2.3 Comparison of CN-C3 to other optogenetic CRISPR tools

In the introduction, I described the different approaches for the control of Cas9 activity (section 1.4.2.5). CN-C3 differs in several regards from alternative strategies for optogenetic CRISPR control. To begin with, CN-C3 is the first optogenetic tool for the control of a type II-C Cas9 orthologue. That said, also other approaches that were outlined in the introduction could prospectively be adapted for this class of Cas9 nucleases (section 1.4.2.5). The unique advantage of the CASANOVA concept in contrast to other methods, however, lies in the fact that only the inhibitor is engineered (altered) and not Cas9 itself. At first sight, this could be considered a disadvantage, as the number of individual components is increased. However, several benefits are linked to this feature. First, as explained above, the ratio of Cas9 to CN-C3 can easily be titrated, which is not possible for single-chain- or split-protein tools. The system can hence easily be fine-tuned, for instance with the goal to reach maximum efficiency under illumination or alternatively for minimal leakiness in the dark. Moreover, since the light-regulation module is separate from the actual gene editing component, researchers can easily use CN-C3 as an add-on system to be combined with CRISPR vectors, cell lines etc. already up and running in the lab. Finally, CN-C3 is expected to be fully compatible with all kinds of Cas9-derived tools. Base editing and prime editing, for example, rely on complex architectures

comprising Cas9 fusions to one or several additional domains. The combination of such a system with optogenetic concepts, that would require fusing even more domains to Cas9, could be problematic and lead to steric clashes. Moreover, adding optogenetic domains to the already large base- or prime editors would result in large proteins exceeding the size limit of most viral delivery systems. Light-regulated sgRNAs, in turn, would only provide a partial solution, here. While principally suitable for the control of base editing, prime editing requires a complex RNA architecture, called pegRNA (prime editing gRNA), which is likely incompatible with the majority of the published methods (Kundert *et al*, 2019; Jain *et al*, 2016). Employing the CASANOVA concept for the control of base- and prime editing is thus a logical and promising next step.

Despite the advantages of CASANOVA, it is always possible that alternative systems are better suited or more straightforward for certain applications. This leads to a last and arguably most important point: The choice of the optogenetic strategy is always a matter of the application and the details of the experimental setup it is used in. In this regard, the largest benefit of the optogenetic CRISPR-tools that we and others developed in recent years is their diversity and the manifold options and features that arise from it.

3.3 Outlook: applications of engineered Acrs

Acrs represent a fascinating class of proteins that has been discovered in the more recent past (Bondy-Denomy *et al*, 2013). They surprised with their diversity with respect to structure and inhibition mechanism (Chaudhary *et al*, 2018; Gussow *et al*, 2020). Among others, our group has substantially contributed to the application of Acrs for gene editing control (Mathony *et al*, 2020a; Hoffmann *et al*, 2021; Bubeck *et al*, 2018; Hoffmann *et al*, 2019; Aschenbrenner *et al*, 2020). Considering the body of work on Acr applications, published by now, one can state that the field is still in a proof-of-concept phase. The main lines of work that focus on actual application of Acrs have been the optogenetic control of gene editing (Hoffmann *et al*, 2021; Bubeck *et al*, 2018), cell-type specificity (Hoffmann *et al*, 2019; Lee *et al*, 2019; Mathony *et al*, 2020a) and the reduction of off-target effects (Aschenbrenner *et al*, 2020; Shin *et al*, 2017; Hoffmann *et al*, 2021; Liang *et al*, 2020). These different areas vary substantially though, with respect to their future perspectives.

CASANOVA, for instance is currently a tool for cell biology, allowing to dynamically regulate Cas9 activity, e.g. for the study of molecular pathways. Its expansion to *in vivo* applications is challenging due to its regulation by blue light. The tissue penetration of light at wavelengths in the range of 450 nm is highly restricted (Ash *et al*, 2017; Finlayson *et al*, 2022). Nonetheless, illumination of skin or the gastrointestinal tract is feasible and optogenetic experiments in the brain of mice are routinely performed using implanted fiber optics (Kim *et al*, 2017). A consequential application of CASANOVA in mouse models would hence be the investigation of skin phenotypes. In the context of biomedical research, wound healing or the study of melanoma are obvious candidates. In the future, CASANOVA could help to decipher the time-resolved dynamics of the processes underlying wound healing or cancer development. From a therapeutic standpoint, spatiotemporal restriction of Cas9 activity to increase the safety of gene therapy represents a long-term perspective. With respect to deeper layers of tissue, red light-induced systems could be beneficial, due to the increased tissue penetration of light at

longer wavelengths (Ash *et al*, 2017; Finlayson *et al*, 2022). Unfortunately, no compact red light-responsive domain, which undergoes considerable structural changes to be suited for insertion strategies, is known. Overall, the application of optogenetics in inner organs is challenging and alternative tools, e.g. based on chemical triggers or magnetism, might be more convenient in such use cases.

The development of cell-type specific gene editing, instead, is clearly targeted to applications in gene therapy. Initial experiments in mice demonstrated the huge potential of this technique (Lee *et al*, 2019). Prior to the use of Acrs, only the stability of the mRNA encoding for Cas9 could be directly regulated via miRNAs (Senís *et al*, 2014). The shortcoming of this previous strategy was, however, that gene editing could only be deactivated in a certain cell type or tissue, while being active in all other tissues. The opposite effect mediated by miRNA-regulated Acrs, namely the targeted activation of Cas9 only in a certain defined tissue, is much more relevant for practical use. This strategy in turn, shares similarity with other approaches, such as cell type-specific promoters (Senís *et al*, 2014) or tissue-specific AAVs (Schmidt & Grimm, 2015; Srivastava, 2016). Importantly, none of the individual strategies provides complete tissue specificity. Importantly, they are mechanistically compatible so that a combination of different approaches might provide the best solution in the future.

Finally, the application of Acrs for reduction in off-target editing, has been demonstrated in several previous studies. Firstly, the timed delivery of Acrs has been applied to limit the time-span in which Cas9 is active (Shin *et al*, 2017; Liang *et al*, 2020). Although effective, this approach is not suited for *in vivo* application and even in cell culture systems the required transfection at two separate time points is cumbersome. The fusion of attenuated Acrs and the use of the CASANOVA concept as demonstrated in this study provide promising alternatives (Hoffmann *et al*, 2021; Bubeck *et al*, 2018; Aschenbrenner *et al*, 2020). The method of choice will probably depend on the question if light-regulated control is desired or not.

Altogether, several promising concepts to apply Acrs in research and beyond have recently been developed. This study contributes to many aspects of the aforementioned applications. It will be interesting to see which lines of research will eventually take the leap towards *in vivo* use or clinical applications.

3.4 A comprehensive domain insertion screen of diverse protein classes

3.4.1 Creating randomized insertion libraries

In order to perform the domain insertion screen, I selected the proteins, AraC, Flp recombinase, SigF and the TVMV protease. The four candidates exhibited major differences with respect to their structure and function, which translated into different behaviors during the performed experiments (Fig. 2.9). For related future studies, several factors are important to consider.

First, the majority of proteins interacts with and binds to other macromolecules. Examples in this dataset are AraC, which acts as a dimer and binds to DNA (Schleif, 2010) or SigF which recruits a polymerase to DNA (Paget, 2015). These interactions are functionally essential and restrict the number of surface sites that are amenable to domain insertion.

Second, the assays I employed for FACS screening rely on the activation of a fluorescent reporter by the candidate protein. This is usually achieved via control of reporter transcription.

The type of assay naturally reflects a bias towards DNA-interacting proteins in my set. The protease is the only example of transcription-independent reporter regulation. Although many of such non-transcriptional reporter assays have been published, one should keep in mind that depending on the candidate protein, the setup of a functional assay could be cumbersome. Especially in case of enzymes that metabolize small molecules, this can be a problem. In this regard, two points are important to consider: Many molecules that serve as educts or products of enzymatic reactions cannot simply be linked to a (fluorescent) readout *in vivo*. On top, the enrichment of functional variants in *E. coli* relies on the measurement and sorting of individual cells. If the substrates or products of an enzyme can diffuse between cells, the read-out could be substantially blurred.

Referring back to the particular candidates from this study, the respective reporter assays differed substantially. Overall, the dynamic range was very large for AraC and the Flp recombinase, resulting in a clear separation between active and inactive variants upon FACS sorting (Fig. 2.9A, B). The assay for the TVMV protease exhibited much weaker separation between positive and negative controls (Fig. 2.9C). The reason lies in the inefficient stabilization of RFP. This can be explained either by incomplete cleavage of the degradation tag through the protease or an insufficient change in RFP stability after the tag was removed. It should be noted, that prior to this study, I tested additional reporter designs, including various degradation tags, promoters and also different positions of the cleavage site, none of which resulted in improvements regarding the dynamic ranges (refer to supplementary note 1, section 5.1). The truncated variant of the TVMV protease that was used for the screen might explain suboptimal cleavage efficiency, albeit this would contradict a previous study (Sun *et al*, 2010). Further investigation of the results with different TVMV protease variants will be necessary to elucidate this question.

With respect to the insert domains, PDZ was a convenient choice, as it has already been used for similar purposes (Oakes *et al*, 2016; Coyote-Maestas *et al*, 2019) and provides a suitable baseline as a small (86 AA) and compact protein domain. With reference to the body of work described in the introduction (see section 1.3.4), the additional insert domain, selected for AraC, are also self-explanatory, as they were frequently used in the context of domain insertions in the past. I chose AraC for the screening of additional domains, due to the very robust reporter assay (Fig. 2.9A) and because the optogenetic engineering of bacterial transcription factors had so far mainly been achieved by domain swapping (Romano *et al*, 2021; Komera *et al*, 2022; Dietler *et al*, 2021). It was thus interesting to investigate the combination of the original allosteric connection of arabinose-induced activation and the insertion of an additional switchable domain. Obviously, the used screening method is principally compatible with any type of insert. However, the success rate of insertions can vary drastically (Fig. 2.12).

Regarding library construction, the SPINE cloning strategy worked well and provided near complete coverage for all samples (Supp. fig. 4). SPINE requires the independent cloning of sub-libraries by splitting coding sequence of the effector protein into chunks of fifty codons. The different regions of the effector covered by individual sub-libraries, were still visible in the in the sequencing results as their coverage varied. This is a sign of differences in DNA concentration upon pooling of the sub-libraries. Overall, these effects were, however, in an unproblematic range, as no insertion variant was strongly depleted (Supp. fig. 4). A second observation that cannot be explained easily is the underrepresentation of the one or two

variants around the borders between two sub-libraries within the variant pool. Ligation errors during assembly of the sub-libraries could, in theory, be the reason, although the SPINE workflow is implemented in a way that such errors are unlikely to affect the representation of individual variants (refer to SPINE publication (Coyote-maestas *et al*, 2019)). As a note of caution, the sequencing coverage of the initial SigF-PDZ library was rather shallow and the sequencing will be repeated for future work on the protein (Supp. Fig. 4). The results for SigF were hence overall noisier (Fig. 2.3.5D).
When comparing this targeted insertion method to transposon-based libraries used in several publications (Coyote-maestas *et al*, 2019; Oakes *et al*, 2016; Nadler *et al*, 2016), the significantly increased coverage and more balanced distribution of candidates in the input libraries becomes apparent (Supp. fig. 4). Considering these benefits and the increasing affordability of the required DNA-oligonucleotide pools, I would argue that the SPINE technique provides superior efficiency as well as quality, as compared to conventional transposon libraries.

3.4.2 Data processing and quality control

The performed experiments can only be judged, based on the resulting sequencing data. The key parameter for data interpretation is the choice of the used scoring metric. The number of reads for a specific insertion site has to be normalized by the overall number of reads in the sample, as the sequencing depth varies between experiments. In addition, a previous study has shown that setting the read count of each candidate in proportion to their respective prevalence in the initial library considerably improved the quality of the results, as it corrects for artificial differences that arose during library preparation (Nadler *et al*, 2016). After this correction for biases, a Log2 transformation allowed to better distinguish enriched from depleted sites. This is especially important with respect to the later analysis (refer to section 4.2.3 for further methodological details).
An alternative metric that has previously been used is z-scored enrichment (Coyote-Maestas *et al*, 2019). I decided against it, since the expectation value of this standardization is zero by definition. Consequently, exactly half of the candidates would end up with a positive score although many of them might actually have decreased activity as compared to the wild-type effector protein. The counter-argument would be that in our screen, the enrichment was not benchmarked against the activity of the wildtype protein without insertion. Consequently, depending on the stringency of the enrichment, also the chosen scoring method could result in scores that either under- or overestimate the activity of variants, although probably to a lesser extent. Also, the z-score normalization could better correct for differing selection stringencies between experiments. What turned the balance upon consideration of the above factors, were the cytometry histograms prior to enrichments, indicating that a large majority of the candidates were inactive (Fig. 2.9). I thus concluded that applying the z-scoring would pose a strong bias on the data, which I intended to avoid.
Next, the quality of the experiments was judged, based on biological duplicates. Overall, experimental replicates were in good agreement (Fig. 2.11). The only outlier was the TVMV protease library with a Pearson's r of only 0.65 between replicates. This effect is likely caused by the overall weak enrichment. Nonetheless, several variants were enriched in both replicates. Apart from that, in most cases the best linear fit between the replicates revealed that the trends are in good agreement, although the absolute values differ (Fig. 2.11). Similarly, experimental

validation of the screen with individual variants proved the overall reliability of the measured categorization into enrichments versus depletions (Supp. fig. 6). The sparsity of robustly enriched variants resulted only relatively few enriched samples that were experimentally validated.

3.4.3 Analysis of the tolerated insertions

Most of the observed insertion preferences and enrichments can be explained by the experimental conditions. The less pronounced depletions within the TVMV library (Fig. 2.3.5C) resulted from the milder selection due to the reporter assay exhibiting a rather low dynamic measurement range (Fig. 2.9C). The complete extinction of many Flp variants upon enrichment (Fig. 2.3.5B) was caused by the stringent selection of the very small fraction of candidates, which were active in the starting library (Fig. 2.9B). Also, the global observation that libraries were dominated by inactive hybrids is in agreement with previous publications (Nadler *et al*, 2016; Oakes *et al*, 2016; Coyote-Maestas *et al*, 2019). In case of AraC, Flp and the TVMV protease, the enriched variants carried the insertion often close to the termini, which was recently reported by Coyote-Maestas et al. (Coyote-Maestas *et al*, 2021). In many regards though, the results differ significantly between the proteins. In the following, I will briefly discuss the implications of the observed enrichments with respect to both, protein sequence and structure.

3.4.3.1 *AraC*

AraC tolerated insertions at several different regions (Fig. 2.3.5A). Most striking was the C-terminal α-helix of the DBD, which appeared to be irrelevant to protein function (Fig. 2.14A and Fig. 2.16). As expected, insertions immediately neighboring the DNA-binding sites or the arabinose-binding β-barrel tended to be strongly depleted upon enrichment (Fig. 2.14A). Two findings, in turn, were rather unexpected. First, certain sites within the β-barrel tolerated insertions very well. Apparently, the overall stability of this domain is high enough not to be disturbed by an additional PDZ domain at these sites (Fig. 2.14A). The second observation refers to two regions that were previously described as being functionally crucial for AraC, the N-terminal arm and the linker region between the coiled-coil and the DBD (Harmer *et al*, 2001; Eustance *et al*, 1994; Seedorff & Schleif, 2011; Saviola *et al*, 1998; Wu & Schleif, 2001a). Both sites tolerated insertions relatively well, although the exact position appeared to matter a lot in case of the "arm" (Fig. 2.14A). PDZ-fusion were only accepted within the 6 N-terminal residues of the arm and just one of these variants was strongly enriched. Nonetheless, it is surprising that despite the supposed importance of the arm for AraC function the N-terminal fusion of a whole PDZ domain to this region seemed not to impair AraC activity (Fig. 2.14A).

The interpretation of these results is, unfortunately, non-trivial. Our incomplete mechanistic understanding of AraC is a central limitation here. Moreover, AraC plays a double role as repressor and activator, depending on its binding preferences for different operator sites. If the measured enrichments result from AraC adopting its activator state or if inhibition of the repressing conformation alone could have already led to these outcomes remains unclear.

The situation is further complicated by the results from experiments with the additional domains beyond PDZ. The insertions in the middle part of the protein, i.e. the inter-domain regions, were only tolerated in case of PDZ and LOV2 as inserts (Fig. 2.16). While both are rather small and compact domains, the other insert domains are, in turn, larger and their termini are separated by a longer distance. Intuitively, the observed size cutoff with respect to insertion

tolerance appears reasonable, since the middle part of AraC around the coiled-coils is involved in AraC dimerization (Soisson *et al*, 1997). Larger domains might hence interfere with the homo-dimer conformation. The C-terminal α-helix is the only stretch that tolerated the insertion of all five domains, again underscoring the rather minor role it must play in DNA-binding (Fig. 2.16). It is further surprising that, although this α-helix does not interact with DNA by itself (Fig. 2.16), also the DNA-binding of the neighboring parts within the DBD was apparently not affected by the presence of large additional domains at the C-terminus. Finally, uniRapR could be successfully inserted into the N-terminal arm (Fig. 2.12C). It is noteworthy that the insertion was tolerated throughout the "arm" until residue 18, in contrast to fewer sites with high insertion permissibility in case of the PDZ and LOV2 domain (Fig. 2.11A and Fig. 2.12B). This observation further supports our finding that the disordered N-terminus of AraC is not of functional relevance.

On a more global level, it is striking that with rare exceptions most AraC positions either tolerated the insertion of more than on domain or none (Fig. 2.11A and Fig. 2.12). This observation is in agreement with a previous study in ion channels (Coyote-Maestas *et al*, 2021) and suggests that tolerance towards insertions tends to be a general feature, which is not very selective for specific inserts. While in unstructured loops, this might not be too surprising, I identified several successful insertion sites in AraC, which are located in α-helices or β-sheets (Fig. 2.14A).

Coming to the deviation between the different insertion datasets, the largest differences can be spotted in the arabinose-binding domain and the linker region (Fig. 2.16). The DBD, in contrast exhibited similar enrichments for all domains. This means that the depletions at the DNA-interacting sites, as well as the enrichments at the most C-terminal α-helix are "conserved" over the different domains. An explanation might be derived from the evolutionary conservation of the HTH motifs in the DBD, which already indicate that the structural constraints might be high (Cortés-Avalos *et al*, 2021). More surprising are the differences observed within the β-barrel, which tolerates the insertion of some (PDZ and LOV2), but not all domains (Fig. 2.16). Consequently, this region must be more tolerant to structural strain, while still exhibiting certain constraints.

3.4.3.2 *Flp recombinase*

Like the other proteins, the Flp recombinase dataset differs substantially from the results for AraC. Only two sequence clusters tolerated the insertion of PDZ. Structurally they are located in proximity (Fig. 2.14B). In section 2.2.2, I described the complex mechanistic and structural features of Flp. Its two domains wrap around the target DNA and during recombination a holliday-junction complex is formed (Chen *et al*, 2000). This complex includes four DNA-bound Flp-monomers arranged in a square-like conformation. Moreover, this state can be conceived as a dimer of a dimer, since two pairs of Flp slightly different conformation are formed (Conway *et al*, 2003). This unique complex explains, why the overall insertion tolerance of Flp was low. The location of the identified insertion patch is in agreement with the observed surface accessibility within the complex (Supp. fig. 7). It remains unclear, however, why different sites with similar solvent accessibility did not tolerate insertions. Putatively, the conformational dynamics during recombination of the target DNA could be an underlying reason.

3.4.3.3 *TVMV protease*

The pattern from the TVMV protease library screening resembles the Flp recombinase in the sense that insertions were mostly tolerated within one region of the protein surface (Fig. 2.15A). Not too surprisingly, this site is located far away from the active center of the protease. The fact that successful insertions were relatively rare is stunning, however. The TVMV protease is the only protein from my set, which has no major binding partners and acts as a monomer. On the other hand, it adopts a rather compact globular fold, which might not accept much physical distortion (Sun *et al*, 2010). Interestingly, while few positions inside the protein core had strongly negative enrichment scores, the active site did not stand out as a highly depleted part (Fig. 2.14A).

3.4.3.4 *SigF*

This candidate differs from the others, in that no crystal structure was available. In contrast to AraC, the DNA-binding domains did not exhibit a homogenous depletion following screening (Fig. 2.15B). While many insertion sites were strongly depleted, a significant proportion had at least neutral or slightly positive scores. The only clearly enriched cluster was located in the coiled-coil part of the intermediate domain (Fig. 2.14A). Apparently, this region does not make close contacts to the polymerase during transcription initiation. Finally, SigF has several unstructured regions, located either terminally, in between domains, or occurring as loops between α-helices. These parts however, were not particularly enriched for insertions (Fig. 2.14A). Overall, a more detailed mechanistic understanding of SigF would be required to further interpret the observed enrichment pattern.

Taking the results from all four candidates together, the datasets provided a combination of easily interpretable features and more surprising enrichment trends. The characterization of the insertion patterns for AraC demonstrated that this method could even deliver hints with respect to mechanistic aspects. Overall, a general discussion regarding the function of the candidate proteins remains difficult and any trends should be interpreted carefully.

3.4.4 Decisive factors for domain insertion tolerance

Single, biophysical properties of amino acids located close to domain insertion sites did not correlate with domain insertion tolerance (Fig. 2.17). This observation was no surprise and showcases the main differences between a domain insertion scanning and the more common method of deep mutational scanning, which investigates point mutations. In case of DMS, a comparison between the mutated and the original amino acids inherently makes sense. This is because the exchange of amino acids with similar properties is more likely to be tolerated as compared to exchanges of biophysically very different ones. The effect of a domain insertion, however, cannot be assigned to a single, exact position at which the domain was introduced, but must be interpreted as a more global event with potentially larger implications on protein structure and function. Investigating which part or which feature of the protein determines if an insertion is tolerated at a specific site is thus a non-trivial task that will be discussed throughout a large part of the remaining chapter.

Coming back to the amino acid positions, the only effect one could have expected, was an enrichment close to prolins, as they tend to be located in loops (Suyama & Ohara, 2003; Bae *et al*, 2005). The absence of this trend is, however, in agreement with the observation that

insertions were not necessarily enriched within loops at all (Fig. 2.18B and Supp. fig. 8B). Another factor must be considered here and during the following analysis; Although our dataset is to date the largest of its kind to date, it is still relatively small. As a result, the statistical power, especially with respect to relatively rare amino acids or other sparse features, is limited. In contrast to the amino acid identity, surface accessibility and secondary structure elements are factors that were considered much more likely to affect the success of domain insertions. Interestingly and in contrast to conventional perception of domain insertion tolerance (Dagliyan *et al*, 2018, 2019; Lee *et al*, 2008), this was not the case (Fig. 2.18A, B and Supp. fig. 8). Starting with the surface accessibility, two aspects might mask a potential correlation with insertion tolerance. First, many surface sites present on our candidate proteins could be critical for interaction with other molecules/proteins or fulfill other functional roles (Fig. 2.18A and Supp. fig. 8A). In addition, domain insertion at some sites may permanently disturb the protein structure independent of how surface exposed this site is. Second, the candidate proteins used in this study are of small or medium size and especially AraC and SigF do not have one large protein core in which a significant proportion of residues is buried. As a result, less pronounced effects in relation to surface exposure must indeed be expected.

With respect to the secondary structure elements, I described the absence of any trend in section 2.2.3, again contradicting common assumptions (Dokholyan, 2016). It should be noted, however, that the categorization into three structure elements is rather superficial. For instance, not all loops can be considered identical. Linding et al. introduced the concept of "hot loops" (Linding *et al*, 2003). The idea is to distinguish between loops that are flexible and others that tend to keep a single conformation. Due to a lack of data, I could not make that or similar distinction in the study. One final point with respect to the statistic evaluation of insertion preferences for secondary structure elements has to be added. The shown distributions do not indicate if an insertion is tolerated at the middle of an α-helix of β-sheet, splitting this structure element in halves or if the insertion is located rather close to an unstructured loop.

Besides diverse biophysical features, I also included three linker-propensity indices into the analysis (Suyama & Ohara, 2003; Bae *et al*, 2005; George & Heringa, 2002). These indices are meant to identify inter-domain linkers. Several amino acids in a row with high linker propensity, have a higher probability of being unstructured and thus to link different domains. The insertion of a domain into a protein artificially creates new inter-domain regions. As a result, stretches with high linker propensity may be suited for insertion. The absence of any correlation with the observed enrichments was thus surprising and further underlines the difficulty of insertion site prediction (Fig. 2.18C, D). On a side note, this concept very much resembles the assumption of enrichments within unstructured loops.

That said, a further refinement of these correlations would still be possible. One could consider weighted contributions of the region around the insertion site which decreases with larger distance. Also, the assessment of surface spheres instead of volume, which also considers the protein core, could affect the outcome. Given the absence of any clear correlations with the observed enrichments and the overall simplicity of this approach, it seems reasonable that overall, more powerful methods are necessary.

3.4.5 Domain insertions versus split-proteins

The insertion of domains and the design of switchable split-proteins differ mechanistically. They are, however, similar in the regard that the insertion site as well as the split site carry additional domains, which must not interfere with protein function. It was thus interesting to see, whether these approaches would result in the identification of similar sites. The comparison of my enrichment scores to the 19 variants tested by Weinberg et al. showed some degree of similarity (Weinberg *et al*, 2019). As already stated in the results section, the outcome should not be overinterpreted. The total number of variants screened by Weinberg et al. is rather small and only three of them were enriched in my dataset (Fig. 2.19). Nonetheless only two successful split sites (out of the 19 tested sites), position 27 and 168, are in conflict with my data. This particular discrepancy could also be caused by the entirely different fusion domains underlying the respective datasets. Overall, the results provided a first indication that further investigation of a link between both methods could be promising.

The SPELL algorithm, in contrast, predicted high split energies close to the termini and between residues 100 and 180, which does not align with my data. That said, the SPELL algorithm has a different purpose (predicting split sites) based on different theoretical assumptions as well as engineering constraints that are likely distinct from the domain insertion problem.

3.4.6 Harnessing AlphaFold2 to analyze the dataset

AF2 has shaken up the field of structural biology over the last year. As described in the introduction (section 1.2.1.1.1), the debate about the impact AF2 is going to have on different research areas and with respect to different scientific challenges is currently ongoing. Concerning this study, the main question is: Can AF2 guide the insertion of domains into effector proteins? At this point, it should be stressed once again that AF2 performs best on single domain proteins (Jumper *et al*, 2021; Akdel *et al*, 2022). Hence, domain insertions are probably not the most ideal use case for AF2. That said, many impressive examples with respect to the prediction of complex protein structures have been shown (Jumper *et al*, 2021; Akdel *et al*, 2022; Tunyasuvunakool *et al*, 2021). Thus, it is still reasonable to analyze AF2 predictions of the hybrid proteins screened in this study.

As detailed in section 2.2.5, neither structure correlation between the fusions and the single proteins nor the pLDDT scores were correlated with insertion tolerance. Nonetheless, two clear conclusions can be drawn (Fig. 2.21). The role of the pLDDT score as a predictor of unstructured regions is currently discussed (Wilson *et al*, 2022; Akdel *et al*, 2022). My results also point into this direction, as a clear decrease of the score was visible in proximity to the loops of AraC (Fig. 2.21B). In addition, a pLDDT decrease in the region immediately neighboring the insertion site was observed. This trend demonstrates one of the very trivial aspects of domain insertion: local rearrangements. These local changes are expected, but have so far rarely been measured (Choi & Ostermeier, 2015; Wright *et al*, 2010). In this context, the AF2 predictions provide an additional indication.

When inserts were introduced close to local pLDDT minima, these regions of decreased pLDDT values were often extended towards the insertion site (Fig. 2.21B). A certain distance cutoff usually resulted in the reversion of the local minimum's size to its original proportion. This behavior is visible in the plot as triangular patterns of lower values observed around the "insertion diagonal". The same observation can also be made with inserts, which are located

close to the protein termini. Explained on the structural level, insertions that disrupt secondary structure elements close to a loop, cause a decrease in the pLDDT score of the faction between the insertion and the loop, likely due to a gain in flexibility. If the insertion site is located beyond a critical distance to the unstructured loop, the pLDDT scores within the region regain the higher default values. This might reflect that the secondary structure element between the insertion site and the loop can adapt its wildtype conformation, again.
On a global level, the predicted overall integrity of insert and parent protein irrespective of the insertion sites was striking, as none of the predictions resulted in larger structural rearrangements or general misfolding (Fig. 2.21D, E). One possible, technical explanation is a bias, caused by the MSAs that underly the predictions. Deep MSAs of the separate protein parts might bias the model towards the near-perfect prediction of their original conformation. Similar observations have been made in a different context (del Alamo *et al*, 2022). It is consequently unclear if or to which extend AF2 tends to predict reasonable structures for actually misfolded fusion hybrids.
Regarding the future of structure prediction models for similar tasks, different scenarios are possible. The latest results from CASP15 showed that the quality of structure prediction models is still improving (https://predictioncenter.org/casp15). The most important factor in recent models, which mostly built on AF2, seems to be quality of the MSAs. In anticipation of a further increase in model performance, accurate predictions of multi-domain proteins are now within reach. It could be possible, to build a framework on top of AF2 or future models with the aim to predict insertion sites. Such strategies have already been implemented for other tasks, including protein design or studying of conformational ensembles (Jendrusch, 2021; del Alamo *et al*, 2022; Goverde *et al*, 2022). This points towards another key aspect, namely the challenge of protein dynamics and conformational changes, which are highly relevant to domain insertion tolerance and allosteric protein regulation. It is well possible that current models are not capable to capture protein dynamics in a way that would allow the prediction of insertion sites. Finally, given the speed at which the field currently moves, it is also an option that neural network architectures that are completely unrelated to AF2 will provide better solutions in the near future.

3.4.7 Training gradient boosting classifiers on domain insertion datasets

3.4.7.1 *Model selection and training*

The initial analysis of individual amino acid features did not yield any significant correlation with the observed insertion data (Fig. 2.18), which motivated the subsequent machine learning approach. The models were employed to (i) elucidate if several combined features implicitly explain the observed insertion tolerance, (ii) investigate, which features are important and (iii) evaluate if useful predictions could be made based on the models.
The choice to train classifiers instead of regressors was motivated by trends within the data. First, most enrichment datasets showed a very clear separation between enriched and depleted variants (Fig. 2.3.5). A discrimination between those two states thus appeared the ideal learning goal. In addition, a classifier had the advantage that artifacts arising from the complete depletion of some variants would not affect the outcome.

With respect to the model choice, gradient boosting classifiers tend to perform very well on tabular data and performed best in preliminary tests. For hyperparameter optimization, grid search allowed a rather systematic evaluation of different settings.

3.4.7.2 *Performance and observations*

Models trained on individual datasets for AraC, TVMV protease and Flp, all showed a similar AUROC within a range of 0.88-0.93, while the average precision of the Flp model was with 0.65 considerably lower as compared to the other two candidates (0.82 and 0.77) (Fig. 2.22). The explanation lies in the effect that different label distributions have on both metrics. The imbalance between the number of negative and positive scores is greater in case of the recombinase, as compared to AraC or TVMV (see Fig. 2.11). The precision metric tends to be more sensitive towards such imbalances. It is also noteworthy that the differences with respect to the performance of the individual cross-validation folds was, in parts, relatively large (Fig. 2.22). The likely reason is, once again, the training data. First, the datasets were overall rather small as each of them contained only a few hundred data points. On top, the sparsity of positive labels resulted in a very small number of positive samples in each validation set, which further decreased the training stability. Finally with respect to the weaker performance of the classifier for SigF, the explanation could either be a reduced quality of the experimental data or that the protein was indeed a more difficult target for reasons that could not yet be elucidated.

Training of the model on the complete dataset revealed a performance in the range of the individual models, as expected (Fig. 2.23A, B). In addition, the fold-to-fold differences with respect to cross-validation were smaller, due to the larger sample size. This model also answered the first of the initial questions (i) that a combination of different features can indeed to some degree elucidate the sites at which insertions are tolerated (Fig. 2.23C, D).

Assessment of the feature importance, although being of great value for the further analysis, exhibited relatively weak trends (Fig. 2.23E, F). It appeared, that only a few top-ranked features, such as KLD and deletion frequency could be considered important, based on the analysis. The fact that the rank order differed significantly between both tested methods, however, showed that they were rather weak indicators of feature importance in the presented context (Fig. 2.23E, F). The median insert length, for instance, showed the lowest permutation importance, but was placed among the higher ranked features in the impurity analysis (Fig. 2.23E, F). A combination of aspects explains the absence of clearer trends. The overall differences upon permutation of a feature were relatively small, supposedly because the remaining data still provided considerable predictive power. In addition, several features were indeed meaningless to the model. As shown in Figure 2.24A, amino acid identities did not provide useful information for training of the classifier. Hence, it is not surprising that the permutation of these features was irrelevant (Fig. 2.23E). Another important factor is the issue of collinearity between certain properties. Residue volume and molecular weight, for instance, correlate by definition. If one feature is permutated, the model can compensate for the loss of information as the same data is still "stored" in another variable. As a result, the role of such features becomes only apparent if all correlating variables are permutated or deleted from the dataset. Of note, the final reduced model did exactly this, based on an iterative deletion of features (Fig. 2.24B). Here, the removal of all biophysical measures did not decrease the model performance,

demonstrating that these features are indeed dispensable (Fig. 2.24B-D). This also marks the second goal (ii), the successful identification of parameters relevant to insertion susceptibility. The final set of parameters was based on information derived from sequence alignments (Fig. 2.24C, D). This appears reasonable, since many modern machine learning approaches for protein data rely heavily on information on sequence conservation and coevolution.

3.4.7.3 *Limitations and future perspectives*

Although the described models performed very well, this study also shows their limitations and remaining challenges. First, the trained classifiers are still far from perfect with 20-50 % of the predictions still being false positives (Fig. 2.22 and Fig. 2.23B). As domain insertions are generally rather likely to be deleterious, it is no surprise that it remains challenging to achieve high precisions. Nonetheless, referring to the third goal (iii) from the last section, the trained models cannot be expected to completely replace the screening of several variants. In addition, the used validation sets necessarily included data from proteins that were already present in the training data. Cross-protein prediction, i.e. making predictions for proteins that were absent from the training data and thus completely unknown by the model was not possible to achieve. This lack of generalization is the biggest hurdle for the application of the discussed models as predictors for domain insertion sites. This issue is caused by the following factors: First, the dataset used in this study is still very small, as it comprises only four different proteins. It is reasonable to assume that a much larger dataset would be necessary to enable a model to generalize from specific example to more general aspects in order to make useful predictions for previously unseen proteins. Second, the described model learns only very simple, i.e. position-specific, representations of insertion sites. Given the complexity of protein structure and function, one could argue that even with a bigger dataset the selected model would reach its limits. The solution would be to train machine learning models on better representations of the insertion sites. The most simplistic solution is to integrate information of the insertion site surroundings, by using distance spheres. Preliminary tests on presented dataset explore this possibility (Supp. Fig. 10). Different levels "context" were provided for each insertion site. This was achieved by defining the features of a given insertion site as the mean of the values corresponding to residues within a specific distance radius around the insertion site. The stepwise increase of this radius from 2 Å to 10 Å was accompanied by an increase of the AUROC and average precision from 0.82 to 0.92 and from 0.54 to 0.69, respectively (Supp. Fig. 10). Despite the simplicity of the approach the result is encouraging with respect to the power that ML models trained on more detailed representations of an insertion site's surrounding could reach. I did not yet explore this direction further, as potential problems due to information that is shared between data points cannot be excluded at the given size of the dataset.

A more advanced approach to represent the region around the insertion sites is represented by graph-based models. Gainza et al. used such representations in order to predict protein-ligand interactions (Gainza *et al*, 2019). The ideal scenario, however, would be to use the whole protein as input data for the prediction of insertion sites. Here MSA- and structure-based representations of the data as they are used by many state-of-the-art models could be imagined (Jumper *et al*, 2021; AlQuraishi, 2019; Sverrisson *et al*, 2020). Harnessing the entire protein as information source to judge the insertion at a specific site, would enable complex models to consider long-distance interactions between residues and global effects an insertion could have on the overall stability or conformation of proteins. Obviously, this would require

datasets that are beyond the limits of current experimental methods and ML approaches beyond the scope of gradient boosting models.
This brings me back to the main problem at the current research state. My dataset, as well as the few comparable published ones (Coyote-Maestas *et al*, 2019, 2021), can only deliver information about very few individual proteins. Hence, the data availability lacks far behind the technical possibilities that modern machine learning approaches provide nowadays.
Nonetheless, while predicting insertion sites for proteins without prior random sampling might remain challenging, the trained classifiers clearly underlined the importance of evolutionary factors instead of biophysical properties. It will be interesting to see, how the evolutionary constrains that determine insertion tolerance can be most efficiently extracted.

3.4.8 Comparison of the gradient boosting models to related concepts

Predicting domain insertion tolerance is a challenging task and only few approaches with this aim have been proposed over the years (Dagliyan *et al*, 2016; Reynolds *et al*, 2011). In the last section, I pointed out that the main reason for the absence of reliable strategies might be a lack of data that can be exploited to develop such approaches. The same applies to the validation and comparison between existing approaches.
In case of the "extrinsic disorder" hypothesis for instance, a few kinases and CASANOVA are the only positive examples (Dagliyan *et al*, 2016; Bubeck *et al*, 2018; Gil *et al*, 2020). In this study, however, I could not confirm the expected trend of switchable variants close to tight loops in the presented data (refer to section 3.4.4). It must, however, be mentioned that I only screened for allosteric switching in two cases, AraC and AcrIIC3.
Most interesting in the given context is the protein sector analysis of the Ranganathan lab (Reynolds *et al*, 2011; Rivoire *et al*, 2016). The reason is that coevolution is computed from MSAs, meaning it is in agreement with my observation that conservation must be a determining factor for insertion tolerance (Fig. 2.24). As explained earlier, key to the analysis is the determination of a "sector" that is usually represented by a sparse but continuous network of residues that coevolve (Rivoire *et al*, 2016). To enable a comparison, I subjected AraC to SCA, which resulted in the prediction of a network of sector residues that was scattered across the protein (Supp. fig. 11). Interestingly, around most sector residues, insertions were not tolerated, supporting the expected functional relevance of these sites. Only very few sector residues were located next to enriched insertion sites. Importantly, only two of the four switchable clusters were in direct proximity to the sector (Supp. fig. 11), in contradiction to previous work (Reynolds *et al*, 2011).
On a conceptual level, several differences between the Ranganathan approach and our models exist. As the name already implies, statistical coupling analysis is a purely mathematic approach, following the premise that the information about the sector (and promising insertion sites) can be computed from an MSA (Rivoire *et al*, 2016). The modeling presented here, took a far more explorative route, considering diverse factors as potentially predictive (refer to section 2.2.6). Nonetheless, the features that were the basis for the final model, are also MSA-derived statistics (Fig. 2.24C, D). The beauty of SCA lies in the fact that the method follows a mathematically constructed rationale from front to back. The ML models trained in this study, instead process relatively simple input statistics in a harder to interpret fashion. The advantage of the ML

approach in turn, lies in the fact that it can potentially learn implicit factors that a researcher would not consider relevant.
Finally, Coyote-Maestas et al. trained decision trees and random forest models that exhibit high similarity to the gradient boosting classifiers presented here (Coyote-Maestas *et al*, 2019, 2021). While the decision trees showed only modest performance (Coyote-Maestas *et al*, 2019), their random forest models performed much better (Coyote-Maestas *et al*, 2021). The main difference to my models is that the authors combined information from the parent protein with features from the insert domain in their dataset. I used only one insert domain, while including different candidate proteins, whereas Coyote-Maestas et al. only focused on ion channels, albeit in combination with a much larger number of insert domains. Importantly, their study was centered solely on biophysical features and did not include sequence information. It would be interesting to see if the published models could be improved upon the inclusion of insertion statistics.
In summary, it is realistic to expect a combination of an increasing mechanistic understanding about the structural requirements and improvements in the ML-based prediction of domain insertion tolerance as the successful path for the future.

3.4.9 Outlook: How to improve the prediction of insertion sites?

The analysis above underlined the complexity of the domain insertion problem and the persisting need for better predictive models that could guide and improve protein engineering efforts. Like all previous methods, also the models described this study come along with their strengths and weaknesses. Overall, the problem can be pinpointed to the size of the available datasets as outlined in section 3.4.6.3. With four different target proteins, this study represents the largest collection of comprehensive domain insertions into candidates of different protein classes. The small number of proteins still sets the upper bar for all data analysis strategies.
With focus on future directions, experimental methods to create and screen domain insertion libraries in high throughput will be essential to increase the size of the available datasets. Unfortunately, a substantial expansion in assay throughput is currently unlikely. Here, the bottleneck are robust reporter assays for different classes of proteins. While the construction and screening of libraries could already be scaled up, establishing and validating functional assays as the basis for screening can still be a tedious and time-consuming task.
Alternatively, one could try to harness the immense resource of natural protein sequences and structures. Indeed, domain insertion represents a common feature in natural proteins. It is thus easily possible to gather large sets of over two million non-redundant protein sequences bearing domains, which are inserted into other domains. Together with a colleague, Benedict Wolf, I currently work on transformer models with the aim to create generalizable predictors for domain insertion tolerance. Due to the large dataset, it is possible to train such complex neural network architectures and to provide entire protein sequences as input for predictions. Our ongoing investigation suggests that domain architectures and the prediction of domain boundaries can efficiently be learned, although no final conclusions can yet be drawn at the current state of this new approach. A recently published study which aimed to predict domain annotations for proteins based on similar transformer networks is in line with our suggested approach (Bileschi *et al*, 2022). The main challenge will be the transfer of the learned information from natural proteins to the prediction of artificial domain insertions. First, domain

combinations in natural proteins were shaped over long periods of time, while we naively create new artificial combinations in the laboratory. This could be expected to result in inherent differences between natural and engineered examples, which transformers could unintendedly learn. Second and more importantly, nature provides only positive examples, i.e. successful insertions that are tolerated by the parent protein and are functionally relevant. For all sites of a protein at which no insertions are found in nature, it remains unclear if an insertion would indeed interfere with protein function or would still be tolerated, but did not evolve in the given biological context. To address this problem, we plan to build more heavily on MSAs instead of single sequences in the future. MSAs have the advantage that they can capture the information about insertions at different sites of a domain more comprehensively. This strategy is also in agreement with the results presented in this study.
The consideration of experimental screens and datasets derived from natural proteins exemplified the strengths and challenges of both approaches. Future progress in both areas, the experimental research as well as the further exploration of more powerful computational approaches based on the natural sequence repertoire, will likely be required for further advances in the field.

3.5 Identification and characterization of light-switchable AraC-LOV2 hybrids

3.5.1 Parallel screening at different conditions is a powerful method to identify allosteric switches

The comparison of the AraC-LOV2 libraries under light and dark conditions confirmed the expectation that many surface sites that accepted the introduction of domains did not result in LOV2-dependent switching of protein activity (Fig. 2.25). This finding is in line with previous reports by us and others (Bubeck *et al*, 2018; Reynolds *et al*, 2011). Of note, it is known, that the switchability can often be tuned by small changes, such as mutations or deletions around the insertion site (Bubeck *et al*, 2018; McCormick *et al*, 2021). In fact, Bubeck et. al started their optimization at a site that tolerated the insertion, but was hardly switchable and subsequently optimized the design, resulting in a powerful optogenetic tool (Bubeck *et al*, 2018). Consequently, the benefit of successful insertions, which are constitutively active is hard to judge without further experiments. It is however clear that the value of these constitutive sites should not be underestimated.
Coming back to the identification of switchable variants, our results further show that the simple parallel enrichment under different conditions is sufficient to identify switchable proteins (Fig. 2.25E). In a previous study several consecutive rounds of enrichments at alternating conditions were performed in order to reach a similar goal (Nadler *et al*, 2016). The procedure used here, however, allows a more straightforward and potentially faster screening in higher throughput.
The identified optogenetic AraC variants showed that LOV2 domain insertions clustered around four AraC sites (Fig. 2.25E). This fact supports an observation previously made with respect to anti-CRISPR proteins (Hoffmann *et al*, 2021; Bubeck *et al*, 2018): Allosteric regulation

is usually not restricted to one position, but to a stretch of 3-4 sequential insertion sites. Referring back to the previous results, the same tends to be true for insertion tolerant sites irrespective of switching (Fig. 2.11 and Fig. 2.25A, B). In this case, the clusters tended to be much larger, often in the range of 10-20 consecutive positions. In order to identify successful insertions sites or even switches, a comprehensive mapping of insertion tolerance might thus not always be necessary, since sampling of several selected sites could identify these hotspots.

3.5.2 The identified switches are clustered around functionally important sites

Zooming in on the location of the insertion sites that resulted in switchable transcription activation, some functional connections become apparent. The first three switchable clusters around position 29, 61 and 113 are all loop regions within the arabinose-binding β-barrel (Fig. 2.25 and Fig. 2.27). These variants differ from the distribution of tolerated insertions in that they are more strongly enriched within loops. In addition, all three clusters resulted in light-ON switches. In its structurally relaxed lit-state, the LOV2 domain apparently does not interfere with AraC activity when inserted at these sites, while it does so in the more compact dark-adapted conformation. In this regard it is also worth noting that the directionality of the switch did not change within a cluster. This fact is less trivial than it might appear, since several studies have shown that very small changes of only one or two amino acids around the fusion site can indeed turn an ON- into an OFF-switch in selected cases (Ryu *et al*, 2014; Etzl *et al*, 2018).

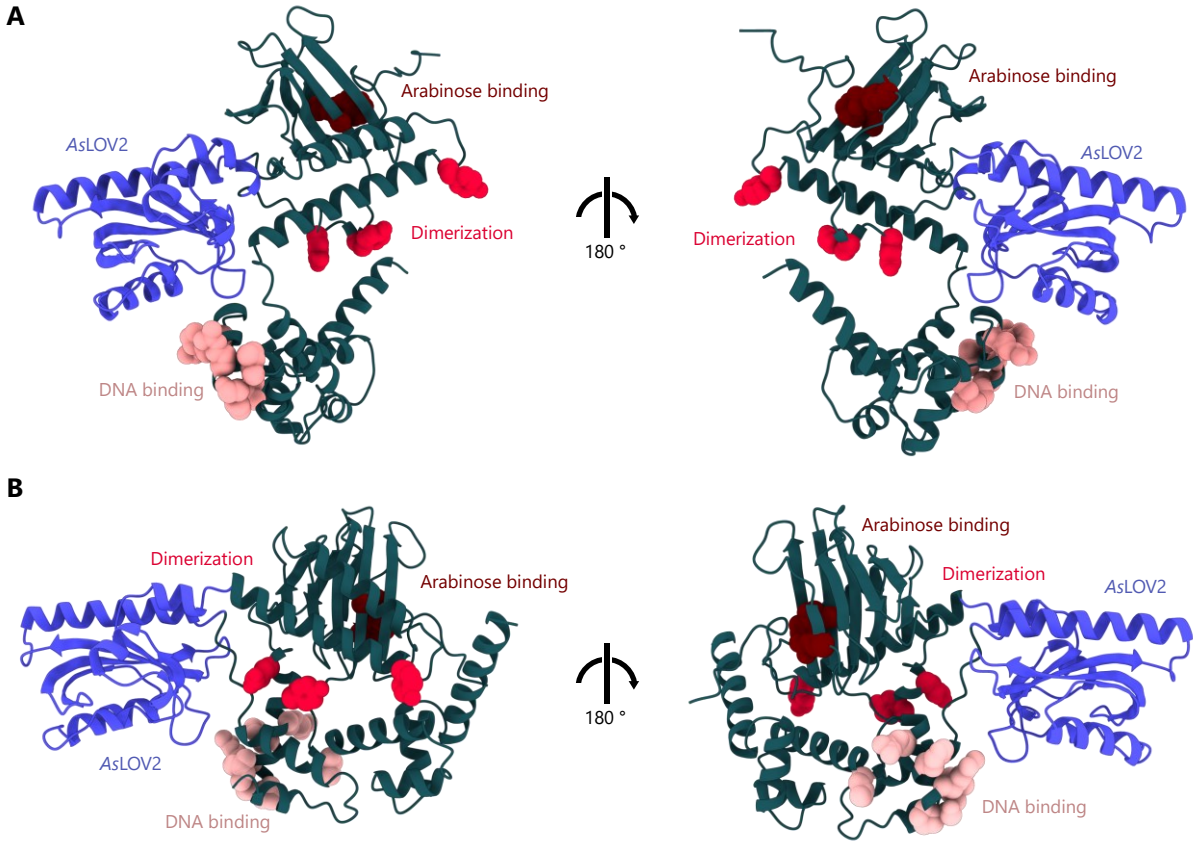


Figure 3.1: AlphaFold2 predicts different conformations for the lead AraC insertion variants. AF2 predictions of AraC-I113-LOV2 (A) and AraC-S170-LOV2 (B) are shown. AraC is depicted in green and the *As*LOV2 domain in blue. Residues that bind to the operator are highlighted in pink, key residues for dimerization in red and the amino acids that are important for arabinose binding in vermilion.

The fourth cluster around insertion position 170 differed in two ways. First, AraC activity was turned off by light and second the location was outside the arabinose-binding domain in the linker region close to the DBD (2.26E and Fig. 2.27). In this case, the best insertion sites were located within an α-helix. Due to the proximity to the dimerization interface of AraC, a steric mechanism of action instead allostery cannot be excluded. Still, a steric explanation remains the less likely option as it would be very surprising, if the LOV2 domain restricted dimerization just in its structurally relaxed light-adapted state. In addition, AF2 models of the two lead candidates, AraC-I113-LOV2 (Fig. 3.1A) and AraC-S170-LOV2 (Fig. 3.1B), predicted the LOV2 domain to point away from the dimerization interface, speaking in favor of an allosteric mechanism. Furthermore, the predictions suggest a more compact conformation of AraC in combination with the dark-adapted LOV2 domain state for the S170 variant (Fig. 3.1B), which is in agreement with the proposed active conformation of AraC (Schleif, 2010). Vice versa, the I113 variant is predicted to adopt a more stretched conformation (Fig. 3.1A), which would be compatible with DNA-looping and inhibition of switching (Schleif, 2010). Taken together, the structure predictions support the observed phenotypes. However, it is important to mention that AF2 always predicted the dark state conformation of the LOV2 domain. Although these considerations give a coherent image, I once again note the limitations of AF2 predictions and the open questions with respect to the AraC mechanism of action. In this light, the proposed explanation should rather be seen as a working model.

3.5.3 Characterization of optogenetic AraC variants

The characterization of the switches revealed a co-dependence on light and arabinose (Fig. 3.2A). Both variants exhibited high dynamic ranges and especially the light-OFF version performed exceptionally well (Fig. 2.26A). The co-dependence on two different inducers might contribute to the performance. Mechanistically, the combined response to two inputs represents a curious phenomenon. Apparently, two allosteric pathways must work in combination to control the activity of the protein. The fact that LOV2 insertions worked only in conjunction with arabinose induction, implies the explanation that the LOV2 domain hijacks the arabinose-induced allosteric mechanism. Mutational studies could help in the future to further dissect the mechanistic background.

The spatiotemporal precision further proved the versatility of these switches (Fig. 2.26B, C) and also the reached dynamic ranges are in the range or above comparable approaches (Romano *et al*, 2021; Dietler *et al*, 2021).

Although frequently used in combination, it is known that IPTG inducible promoters and pBAD can exhibit compatibility issues (Lee *et al*, 2007; Daniel *et al*, 2013; Stricker *et al*, 2008). Using the similarly well characterized TetR system as alternative, however, was not an option in context of our study, as the cognate inducer anhydrotetracycline is light-sensitive. In my experiments, the IPTG and arabinose-inducible promoters worked surprisingly well together and enabled strong as well as titratable expression levels (Fig. 2.26A and Fig. 2.9).

3.5.4 Outlook: Impact of optogenetic AraC-LOV2 switches and future directions

Transcription factors, the activity of which depends on two inputs enable to control the activation of transcription and the level of activity, separately. In the context of optogenetic control for instance, different parallel experiments could be performed under the same light regime, while transcription activity could be individually fine-tuned by arabinose induction. Previously, the combination of chemically inducible transcription factors and the light-responsive regulator EL222 has been used to this end (Jayaraman *et al*, 2018). The optogenetic variants presented here allow similar experimental setups, while at the same simplifying the underlying system to a single protein component.

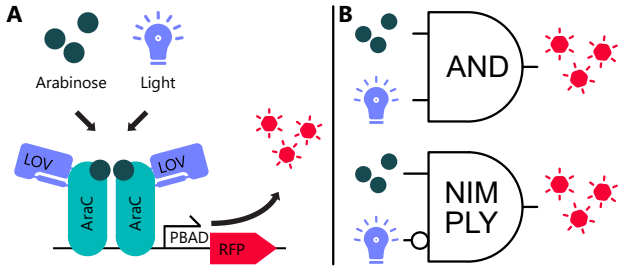


Figure 3.2: AraC-LOV2 hybrids represent single protein logic gates. (A) Schematic of the co-dependence of the AraC-LOV2 hybrids on arabinose and blue light. (B) AraC-I113-LOV2 acts as a single-protein AND gate and AraC-S170-LOV2 as a NIMPLY gate.

Recently a different optogenetically activatable AraC variant, BLADE, was published (Romano *et al*, 2021). The main difference to my proteins is that BLADE only represents a light-ON switch and is activatable by light alone. A direct side-by-side comparison was not performed.

Conceptually, the AraC-LOV2 fusions represent single protein boolean logic gates (Fig. 3.2B). The ON-switch AraC-I113-LOV2 acts as an AND-gate integrating blue light and arabinose as inputs, while AraC-S170-LOV2 represents a NIMPLY-gate. Such protein-based logic circuits recently gained attention as the group of Michael Elowitz created protease-based circuits and the lab of David Baker published work on protein computations with help of artificially designed α-helix bundles (Chen *et al*, 2020; Gao *et al*, 2018). Their work enabled DNA-independent computations with fast response times. In contrast to the tools presented here, these circuits all consisted of several components, for example split-proteases and their substrates (Gao *et al*, 2018). Engineered single-protein logic gates, in turn, provide the most compact and direct wiring from the input signals to the output computation and could potentially simplify the protein circuits in future work. To my knowledge, only one other example, an OR gate, constructed by fusing LOV2 and uniRapR domains to a kinase has so far been engineered (Vishweshwaraiah *et al*, 2021). My combination of the natural existing allostery with an artificially second input might be an approach that could be easily adapted to other proteins.

The theoretical background of single molecule computation agents has recently been discussed (Dokholyan, 2021). Apart from the aforementioned functional flexibility of proteins, which respond to different stimuli, I see the main future application in increasingly complex protein-based computations. For instance, a recent preprint demonstrated design of neural-network computations on the protein level (Chen *et al*, 2022), showcasing the increasing power of artificial protein and gene networks in living cells. The possibility to integrate complex information and wiring it to desired actuations is one of the main goals in synthetic biology. Such complex synthetic programs also require efficient processing on the level of the individual protein components. Integrating more functions into single amino acid chains has the potential to simplify molecular networks, release metabolic burden from the host cell (Ceroni *et al*, 2018)

and could reduce noise derived from stochastic fluctuations of the individual components (Eldar & Elowitz, 2010). This way, single-protein logic gates could contribute to future generations of synthetic biology strategies to program and re-wire cells.

4 Materials and Methods

This section describes the procedures applied for experiments and data analysis. The methods regarding the work on Acrs were also reported in two publications (Hoffmann *et al*, 2021; Mathony *et al*, 2020a). The descriptions of the AAV production and the CN-C3 structural modeling are based on reports by Carolin Schmelas and Julius Upmeier zu Belzen, respectively.

4.1 Experimental methods

4.1.1 Molecular cloning

All constructs used in this study are listed in supplementary table 1. The corresponding amino acid sequences of the encoded proteins are shown in supplementary table 2. Plasmids were generally constructed using Golden Gate assembly (Engler *et al*, 2008). In brief, DNA fragments were amplified by PCR (Q5 2x Master Mix, New England Biolabs (NEB)), with primers carrying type IIS restriction enzyme recognition sites in their overhangs, which enabled the scarless assembly of the constructs. PCRs were performed according to the NEB standard protocols. For Golden Gate assembly, the procedure described by Engler et al. was used (Engler *et al*, 2008). DNA-oligonucleotides were ordered from Merck and Integrated DNA Technologies (IDT). Double-stranded DNA fragments were purchased at IDT. PCR products were resolved on 0.5x Tris-acetate-EDTA (TAE) 1 % agarose gels and the corresponding bands were cut out and purified using the QIAquick Gel Extraction kit (Qiagen). Restriction enzymes and T4 DNA ligase were obtained from NEB and Thermo Fisher Scientific. Following DNA assembly, Top10 *E. coli* cells (Thermo Fisher Scientific) were transformed with the respective construct, followed by overnight incubation at 37 °C. Liquid cultures were inoculated from single colonies and grown overnight at 37 °C and 220 rounds per minute (rpm). DNA was purified using the QIAamp DNA Mini, Plasmid Plus Midi or Plasmid Maxi kit (all Qiagen). All constructs were verified using Sanger sequencing (Microsynth Seqlab and Genewiz).

The plasmids pEJS654 All-in-One AAV-sgRNA-h*Nme*Cas9 and *Nme*2Cas9_AAV encoding *Nme*Cas9 or *Nme*2Cas9 and a corresponding sgRNA expression cassette (Addgene #112139 and #119924) were a kind gift from Erik Sontheimer. The construct pX601-AAV-CMV::NLS-SaCas9-NLS-3xHA-bGHpA;U6::BsaI-sgRNA encoding *Sau*Cas9 together with a sgRNA expression cassette (Addgene #61591) was a kind gift from Feng Zhang.

4.1.2 Cell culture and transient transfection

4.1.2.1 *General cell culture procedures*

For most assays, HEK293T cells (human embryonic kidney) were used. Only in case of the miRNA-dependent Cas9 activity switch, Huh7 cells (hepatocyte cell line) were used as additional cell line. Cells were cultured at 37 °C in a humidified atmosphere of 5 % CO_2. Phenol red-free Dulbecco's Modified Eagle Medium (DMEM, Thermo Fisher Scientific) was used for cultivation. Media were supplemented with 10 % (v/v) fetal calf serum (Biochem AG), 2 mM L-glutamine (Thermo Fisher Scientific), 100 U per ml Penicillin (Thermo Fisher Scientific) and 100 and 100 µg/ml streptomycin (Thermo Fisher Scientific). To ensure integrity, cell lines were

authenticated prior to usage and regular tests for Mycoplasma contamination using the PCR Mycoplasma Detection Kit (ABM) were conducted.

4.1.2.2 *Transient transfection*

For all gene editing experiments, transient transfection was performed in 96-well plates, using a seeding density of 12,500 cell per well for HEK293T. For optogenetic experiments, black plates with clear bottom (Corning) were employed. All other experiments were performed in transparent multi-well plates (Corning). For the AcrIIC1 project, transfections were performed using Lipofectamine 3000, according to the manufacturer's protocol (Thermo Fisher Scientific). 200 ng of DNA were transfected per well, using 0.2 µl Lipofectamine and 0.4 µl p3000. The vector ratios between the respective Acr expressing construct and the corresponding all-in-one Cas9/sgRNA expressing constructs are indicated in the figure legends. For the controls, the Acr expressing plasmid was replaced by the vector pBluescript (Invitrogen), as stuffer DNA. In case of the negative controls, an all-in-one construct expressing a non-targeting sgRNA was used. The experiment involving miRNA overexpression was performed with 30 ng *Sau*Cas9 plasmid, 120 ng of the respective Acr construct and 80 ng of the miRNA plasmid per well, following the same transfections protocol.
The assays regarding CN-C3 were performed with 150 ng of total DNA per well, following the identical Lipofectamine 3000 protocol as described above. As sole exception, the CN-C3 titration experiment was performed using the JetPrime transfection reagent (Polyplus) following the manufacturer's recommendation, again, with 150 ng of DNA per well. The vector mass ratios are indicated in the respective figure panels/legends. To keep the amount of DNA constant, pBluescript was, used as stuffer. Transfections for Western Blots were performed in a six-well format, after seeding $2.5x10^5$ cells/well, the day before. 2 µg of DNA were used in total, comprising of the all-in-one Cas9 and sgRNA plasmid, as well as the Acr expressing plasmid in a vector mass ratio of 1:1. Transfection was performed using JetPrime following the manufacturer's protocol, as above.

4.1.3 AAV production and transduction

4.1.3.1 *AAV production*

The following section was adapted from a protocol kindly provided by Carolin Schmelas. In order to produce and purify AAVs, five 14 cm petri dishes were seeded with HEK293T cells at a density of $4x10^6$ cells per dish. Two days after seeding, cells were co-transfected with (i) a construct carrying the transgene, flanked by AAV-specific ITRs, (ii) a plasmid encoding the AAV *rep* and *cap* genes (serotype 2) and (iii) an adenoviral helper construct. A total amount of 14.6 µg of DNA was mixed with 6 ml H_2O, 7.9 ml of 300 mM NaCl (Sigma-Aldrich) and 1.75 ml polyethylenimine (PEI, Polyscience). Following 10 minutes of incubation at room temperature, 3.2 ml of transfection mix were slowly added dropwise per dish. After three days, the cells were harvested by incubation at 37 °C for 1 h in 5 ml of Benzonase buffer (50 mM Tris-HCl, 150 mM NaCl, 2 mM $MgCl_2$, pH 8.5). Addition of 1 µl of highly concentrated Benzonase (Merck Millipore) ensured the digestion of remaining DNA. Cell lysis was performed by five subsequent freeze and thaw cycle using liquid nitrogen and a 37 °C water bath. This step was followed by removal of cell debris via centrifugation at 4,000 g for 15 min. Finally, AAVs were purified using a iodixanol gradient, following a protocol described by Börner et al. (Börner *et al*, 2013). In short,

ultracentrifugation tubes (Seton Scientific), were filled with the AAV-containing supernatant, which was then underlaid with 1.5 ml of 15 %, 25 %, 40%, and 60 % iodixanol phases using a Pasteur pipet. Next, these gradients were centrifuged at 50,000 rpm at 4 °C for 2 h, so that the AAVs accumulated at the interface of the 40 % and 60 % phases. Finally, the virus was collected from the tube using a syringe and aliquots were stored at -80 °C until further use.
AAV yields were measured by quantitative PCR (qPCR) using the SensimixII Probe kit (Bioline) and the Rotor Gene 6000 qPCR cycler (Qiagen). The primers and the probe targeted the Cytomegalovirus (CMV) promoter from the transgenes (forward: 5`-AACGCCAATAGGGACTTTCC, reverse: 5`-GGGCGTACTTGGCATATGAT, probe: 5`-FAM-CGGTAAACTGCCCACTTGGCAGT-BHQ1). The following qPCR program was used: Initial denaturation for 10 min at 95 °C, followed by 40 cycles of denaturation at 95 °C for 10 s and elongation at 60 °C for 20 s. The results were analyzed using the RotorGene 6000 Series Software 1.7.

4.1.3.2 *AAV transduction*

AAV-transduction was performed at the 96-well format, seeding cells at a density of 12,500 or 3,000 cells per well for HEK293T and Huh7, respectively. Cells were transduced on two subsequent days, i.e. 24 h and 48 h after seeding. For the AAV encoding *Sau*Cas9 and the sgRNA, a MOI of 10^5 was used. The Acrs were supplied at an MOI of $5x10^4$. The cells were lysed two days after the second transduction and indel frequencies were measured by TIDE sequencing and T7E-assay.

4.1.4 Measurement of gene editing efficiencies

4.1.4.1 *T7-endonuclease assay*

Cells were lysed three days post transfection. Media was removed and 141 µl of 1x DirectPCR Lysis Reagent (Peqlab), supplemented with 200 µg/ml Proteinase K (Roche Diagnostics) was added. The lysis mix was incubated at 55 °C for at least 6 h, while shaking at 100 rpm. Lysis was followed by inactivation of proteinase K at 85 °C for 45 min. Next, the genomic loci targeted by Cas9 (Supplementary table 2) were amplified by the primers indicated in supplementary table 3, using the Q5 Hot Start High-Fidelity DNA Polymerase (New England Biolabs).
To perform the T7E assay, 5 µl of the amplicons were diluted to a volume of 20 µl in 1x NEB buffer 2. This mix was heated to 95 °C for 5 min in a PCR cycler, followed by re-annealing of the single-stranded DNA, using a cooling ramp rate of -2 °C/s at the temperature range between 95 °C and 85 °C. From thereon, a slower cooling rate of only -0.1 °C/s was applied until a temperature of 25 °C was reached. Samples were placed on ice, 0.5 µl of T7 endonuclease (New England Biolabs) was added, and samples were incubated at 37 °C for 15 min. Subsequently, the reaction was again placed on ice immediately and the samples were analyzed on 2 % Tris-borate-EDTA (TBE) agarose gels. Gel images were taken, while ensuring that the brightness levels of the DNA bands were not oversaturated. DNA band intensities were assessed using the ImageJ software (https://imagej.nih.gov/ij) (Rueden *et al*, 2017; Schneider *et al*, 2012). Indel frequencies were then calculated by the following formula: indel(%) = 100 x $(1 - (1- \text{fraction cleaved})^{1/2})$, where the fraction cleaved = Sum(cleavage product bands) / Sum(cleavage product bands + PCR input band).

4.1.4.2 *TIDE sequencing*

For TIDE sequencing, the same lysis and amplification procedure, as described for T7E-assays was applied. The amplicons were analyzed on a 0.5 x TAE 1 % agarose gel and the DNA bands were extracted using the QIAquick Gel Extraction Kit (Qiagen). DNA concentrations were measured using a nano-photometer (Nanodrop, Thermo Fisher Scientific). Next, samples with a concentration of approximately 75 ng/µl were sent for sanger sequencing (Genewiz or Eurofins) using the either the forward or the reverse PCR primer (Supplementary table 3) as sequencing primer. The resulting sequencing chromatogram was analyzed using the TIDE web tool (https://tide.nki.nl) (Brinkman *et al*, 2014).

4.1.4.3 *Targeted amplicon sequencing*

The cells were lysed and PCRs were performed as described in the T7E-assay section. In this case, however, modified PCR primers carrying the Illumina adapters, as well as custom barcodes for multiplexing were used. As before, the PCR products were resolved on an agarose gel, the desired bands were cut out and DNA was extracted. Finally, the DNA was diluted to 20 ng/µl and up to six different, individually barcoded samples were pooled. Targeted amplicon sequencing was performed using the commercial Genewiz Amplicon-EZ service. The samples were de-multiplexed using the Sabre package (https://github.com/najoshi/sabre). Indel frequencies were calculated using the CRISPresso 2.0 suite (https://github.com/pinellolab/CRISPResso2) (Clement *et al*, 2019).

4.1.5 Western blot

To lyse the cells at the time points indicated in the figures, the media was aspired, followed by washing with PBS. Next, 150 µl of protein lysis buffer (150 mM NaCl, 10 mM Tris, 1 mM Ethylenediaminetetraacetic acid (EDTA), 0.5 % NP-40 and 10 % cOmplete Protease Inhibitor (Roche Diagnostics), pH 8.0) was added per well. Cells were scraped off using a plastic spatula and the cell debris was removed by centrifugation at 10,000 g for 5 min at 4 °C. The supernatant was collected in new 1.5 ml reaction tubes. From now on, the samples were kept on ice. Protein concentrations were assessed by Bradford assay (Sigma-Aldrich) according to the manufacturer's protocol. Next, 30 µg of protein were diluted in Laemmli Sample Buffer (Bio-Rad) and the final volume was adjusted to 25 µl with lysis buffer. Finally, the samples were heated to 95 °C for 5 min, before being loaded on a 10 % Bis-Tris gel (Life Technologies). Electrophoresis was performed in 1x MOPS (3-(N-morpholino)-propanesulfonic acid) buffer (Life Technologies) at 130 V for 120 min. As ladder, the PageRuler Plus Prestained Protein Standard (Thermo Fisher Scientific) was used. The proteins were then blotted onto a nitrocellulose membrane (poresize: 0.2 µm) (Millipore) in 1x Towbin buffer at 120 V, again applied for 120 min. Membrane pieces covering the protein size ranges of <40 kDa, 40-80 kDa and >80 kDa were cut out and blocked for 1 h in 5 % (w/v) milk powder (Carl Roth) dissolved in tris-buffered saline (TBS) (ChemCruz) and supplemented with 1 % (v/v) Tween (TBS-T) (Carl Roth). Subsequently, the middle part of the membrane was incubated overnight with a primary antibody against α-tubulin (Santa Cruz, sc-32293, 1:1,000), while the other parts were incubated with a primary antibody against the HA-tag (Santa Cruz, sc-7392, 1:1,000) under constant rocking. Both antibodies were diluted with 5 % milk powder in TBS-T. The next day, membranes were washed with TBS-T three times for 10 min and then incubated with HRP-

(horse radish peroxidase)-conjugated secondary antibodies (anti-mouse antibody, 1:5,000 in 5% milk in TBS-T (Dianova)) for another hour on gyratory rocker. After washing the membranes again three times with TBS-T for 10 min under constant rocking, they were incubated with the SuperSignal West Pico PLUS Chemiluminescent Substrate (ThermoFisher) for 5 min. Finally, images were acquired with a ChemoStar detector (Intas). Band Quantification, was performed with ImageJ (https://imagej.nih.gov/ij) (Rueden *et al*, 2017; Schneider *et al*, 2012).

4.1.6 Illumination setup

4.1.6.1 *Illumination of mammalian cells*

To illuminate cells, a custom-made blue light setup was used, consisting of six high power LEDs (type CREE XP-E D5-15; emission peak ~460 nm; emission angle ~130°; LED-TECH.DE) individually mounted onto cooling elements. The LEDs were connected to a Switching Mode Power Supply (Manson; HCS-3102) and controlled via a Raspberry Pi, executing a custom Python script. The clear bottom 96-well plates were positioned on a table made of acrylic glass and illuminated by the LEDs from below. The whole setup was placed into a standard cell culture incubator. An illumination duty cycle of 5 s light-on and 10 s light-off was chosen at a light intensity of 3 W/m^2. The illumination intensity was regularly validated using a LI-COR LI-250A light meter. Control plates, with identical samples were kept within the same incubator, but were constantly protected from light.

4.1.6.2 *Illumination of E. coli*

For the illumination of liquid cultures, another custom-made LED setup was used. Eight blue light high-power LEDs (type CREE XP-E D5-15; emission peak ~460 nm; emission angle ~130°; LED-TECH.DE) were mounted onto an aluminum plate and connected to a Switching Mode Power Supply (Manson; HCS-3102). The LED-plate was installed upside down within a shaking incubator, so that the LEDs could illuminate the surface area of the shaking platform from a distance of approximately 30 cm. Liquid cultures were incubated in multi-well plates and illuminated at a constant intensity of 50 µmol/(m^2 s).

For the illumination of agar plates a different illumination device was applied. Here, a custom-made array of 96 LEDs (LB T64G-AACB-59-Z484-20-R33-Z, Osram, emission peak 469 nm, viewing angle 30 °, Mouser Electronics) mounted on circuit board were used at a light intensity of 15 µmol/(m^2 s), powered by a Switching Mode Power Supply (Manson; HCS-3102). A photo-mask made from black vinyl (Starlab) was cut out by hand and was directly attached to the bottom of the agar plate used in the experiment. This plate was placed above the LED array at a distance of ~5 cm. The whole setup was installed inside a standard bacteria incubator (Minitron, Infors). The electronic light setups were constructed by the workshop of the biology department at TU Darmstadt. I thank them very much for their great support.

4.1.7 TVMV reporter assay test

In order to test the different reporter constructs, plasmids were co-transformed with an IPTG inducible TVMV protease. Precultures of 1 ml, supplemented with 50 µg/ml chloramphenicol (Carl Roth) and 25 µg/ml of kanamycin (Carl Roth), were grown overnight at 37 °C and 220 rpm. The next day, 1 ml of media containing 0 µM, 200 µM or 400 µM of IPTG, again supplemented with identical concentrations of kanamycin and chloramphenicol, were

inoculated with 5 µl of each preculture. The main cultures were grown for 16 h at 37 °C and 220 rpm. After incubation, RFP fluorescence and OD_{600} were measured in a plate reader (Tecan Infinite 200 Pro). RFP levels were acquired at an excitation wavelength of 490 nm and an emission wavelength of 520 nm. The reported values were then calculated, by dividing measured fluorescence by the OD_{600} levels.

4.1.8 Optogenetic assays in *E. coli*

4.1.8.1 *Characterization of AraC-LOV2 hybrids*

Precultures of Oneshot Top10 cells (Thermo Fisher Scientific) carrying the RFP reporter plasmid for AraC and an IPTG inducible expression plasmid encoding the transcription factor or its derivatives, were inoculated from glycerol stocks into lysogeny broth (LB) (Carl Roth), supplemented with 50 µg/ml chloramphenicol (Carl Roth) and 25 µg/ml of kanamycin (Carl Roth). Cultures were prepared in 48-well plates (Corning), using a volume of 0.5 ml per well. The precultures were incubated for 16 h at 37 °C, while shaking at 220 rpm. Main cultures were similarly prepared in 48-well plates, using LB supplemented with 50 µg/ml chloramphenicol and 25 µg/ml of kanamycin, together with different amounts of IPTG (Carl Roth) and L-arabinose (Carl Roth). IPTG concentrations used in each sample are indicated in the corresponding figures/legends. The cultures were prepared in duplicates using with 5 µl from the respective precultures. Subsequently, one replicate was incubated under blue light illumination, while the other replicate was kept in the dark within the same incubator. The growth conditions were again 37 °C and 220 rpm for 16 h. After incubation, RFP fluorescence and OD_{600} were measured in a plate reader (Tecan Infinite 200 Pro). For RFP measurements, an excitation wavelength of 490 nm and an emission wavelength of 520 nm were used. The reported values were the calculated, by dividing measured fluorescence by the OD_{600} levels. Three independent biological replicates were generated by repeating experiments on different days.

4.1.8.2 *Agar plate photography*

Prior to the experiment, agar plates were poured using 1.5 % LB-agar, supplemented with 50 µg/ml chloramphenicol and 25 µg/ml of kanamycin, 400 µM IPTG and 25 mM L-arabinose (all Carl Roth). A preculture was prepared as described above (section 4.1.7.1) in a volume of 4 ml from Oneshot Top10 cells (Thermo Fischer Scientific), transformed with the pBAD-RFP reporter plasmid and a construct expressing AraC-S170-LOV2 from an IPTG-inducible trc (trp-lac) promoter. The preculture was incubated overnight at 37 °C and 220 rpm. The next day, 0.6 % LB-agar was freshly prepared and cooled to ~40 °C. Then, 3 ml of the liquid agar were supplemented with IPTG and L-arabinose to final concentrations of 400 µM and 25 mM, respectively. Finally, 300 µl of the preculture were quickly added to the agar, mixed by shaking and distributed on the previously prepared agar plates. After 30 minutes at room temperature, the top ager had solidified, and the photo-mask was glued to the bottom of the plate. Finally, the plate was incubated at 37 °C overnight, under constant blue light illumination. Images were acquired on the next day using a UV light source and camera.

4.1.8.3 *Reversibility experiment*

In a 48-well plate (Corning), 0.5 ml cultures were prepared, using LB media, supplemented with 50 µg/ml chloramphenicol and 25 µg/ml of kanamycin, 400 µM IPTG and 25 mM L-arabinose (all Carl Roth). The wells were inoculated with 5 µl of precultures that had been prepared as described in section 4.9.1, the day before. The cultures were then incubated at 37 °C and 220 rpm for three hours in darkness, followed by 3 h incubation under blue light exposure and a final step of 3 h in the dark. Prior to the first incubation step and after each following incubation period, the RFP fluorescence and the OD_{600} were measured via plate reader (Tecan Infinite 200 Pro). After every incubation period the samples were diluted 1:30 into new plates with pre-warmed fresh media, containing all supplements. The final relative fluorescence was obtained, by normalizing the RFP values to the measured OD_{600}. Three independent biological replicates were generated by repeating experiments on different days.

4.1.9 Comprehensive domain insertion screen

4.1.9.1 *Insertion library generation*

To construct comprehensive insertion libraries, I used saturated programmable insertion engineering (SPINE) (Coyote-maestas *et al*, 2019). The method builds on oligonucleotide pools (ordered at Agilent) that allow to order thousands of individual oligonucleotides with lengths up to 230 bases. The protein of interest was subdivided into chunks of ~50 amino acids. For each chunk, an oligonucleotide sub-pool was designed, comprising 50 individual sequences, each of which carried a Type IIS restriction enzyme recognition site behind a specific amino acid encoding triplet. A python pipeline for the automatic design of the required DNA sequences is provided Coyote-Maestas et al. (Coyote-maestas *et al*, 2019). The sub-pools were then individually introduced into an expression vector carrying the full coding sequence of the respective parent protein of interest. The sub-libraries were transformed into chemically competent Oneshot Top10 *E. coli* and grown overnight in liquid culture. To ensure an at least 40-fold coverage of the library, serial dilutions were plated on agar after transformation and the number of colony-forming units was calculated the next day. The sub-libraries were extracted from the bacteria using the QIAamp DNA Mini Preparation Kit (Qiagen) on a plate reader (Tecan Infinite 200 Pro). The DNA concentration was measured using the Quant-iT dsDNA (HS) assay kit (Thermo Fisher Scientific) and all sub-libraries from each protein were pooled at equal concentrations. To ensure that no wildtype protein contamination was carried on during cloning, the insertion handle was replaced by a kanamycin expression cassette via Golden Gate assembly. This insert was again flanked by a pair of different Type IIS restriction site, enabling its exchange to the actual domain of interest via Golden gate cloning. *E. coli* cells were transformed with the resulting kanamycin resistant library and plated on three 20 cm LB-agar plates, supplemented with 50 µg/ml chloramphenicol and 25 µg/ml of kanamycin (Carl-Roth). This procedure resulted in plates that were densely covered, but individual colonies were still visible. Again, a library coverage of at least 20× was observed by colony counting. The next day, each plate was rinsed with 3 ml of LB and the colonies were gently scraped off with a spatula. The resulting liquid cultures were collected from the plates and pooled for each protein. Plasmid DNA was then purified from the cultures and the kanamycin handle was replaced by the domain of choice in a final Golden Gate step. In the meantime, electrocompetent Oneshot Top10 *E. coli* cells carrying the respective RFP reporter plasmid had

been prepared. The cells were transformed with the assembled libraries by electroporation using the Gene Pulser Xcell electroporator (Bio-Rad). Following recovery in super optimal broth supplemented with 20 mM glucose (Carl Roth) (SOC) for one hour at 37 °C and 220 rpm, transformed cells were grown in LB (50 µg/ml chloramphenicol and 25 µg/ml of kanamycin) overnight. Serial dilutions plated on agar were performed. Plates were incubated overnight, and a library coverage was estimated from colony counts (coverage was >50-fold for all samples). Finally, glycerol stocks of the libraries were prepared, by mixing the culture with sterile 50 % (v/v) glycerol at a ratio of 1:1.

4.1.9.2 *Screening procedure*

Precultures of LB media (50 µg/ml of chloramphenicol and 25 µg/ml of kanamycin) were inoculated from glycerol stocks of *E. coli* strains carrying the insertion libraries. Positive controls expressing the wildtype parent protein without insert, as well as negative controls expressing a different protein of similar size from the same plasmid backbone, were included. The precultures were incubated for 16 h at 37 °C while shaking at 220 rpm. The next day, 1 ml LB cultures were inoculated with 10 µl from the precultures. These main cultures were supplemented with 16 mM L-arabinose and 400 µM IPTG for AraC, 400 µM IPTG for the TVMV protease, 200 µM IPTG for Flp, 100 µM IPTG for SigF during the first enrichment round and 200 µM for SigF during the second round of enrichment. These cultures were incubated for 16 h at 37 °C while shaking at 220 rpm. For the AraC-LOV2 libraries, two identical replicates were generated, one of which was incubated under blue light illumination and the other one in the dark. The next morning, the samples were diluted 1:100 in 1× PBS (Thermo Fisher Scientific) and kept on ice until sorting. FACS was performed on a FACSAria Fusion flow cytometer (BD Biosciences) at the ZMBH FACS facility (Heidelberg University). *E. coli* cells were identified and gated via their forward scatter (FSC) and side scatter (SSC). The red fluorescent peak was sorted from each library. If no clear peak was visible, the 5 % cells with the highest RFP levels were sorted. 25,000 cells were sorted, into LB media. After sorting, the collected cells were recovered for one hour without antibiotics at 37 °C and shaking at 220 rpm. Subsequently, 50 µg/ml chloramphenicol and 25 µg/ml of kanamycin were added and samples and incubation proceeded overnight. The next day, glycerol stocks were prepared from the libraries. As two rounds of sorting were necessary, the whole procedure was identically performed a second time, starting from the glycerol stocks of the first round of enrichment. The sorting data was analyzed using the python package (https://cytoflow.github.io/).

4.1.9.3 *Next generation sequencing*

The samples of the input libraries, as well as the enriched sorted fractions were objected to heat lysis. Cells were pelleted and resuspended in water. Aliquots were heated to 95 °C for 10 min, followed by centrifugation at 10,000 g for 10 min to remove cell debris. The supernatant was transferred to new tubes and stored at -20 °C for further usage. The coding sequence of the libraries was amplified using the Q5 Hot Start High- Fidelity DNA Polymerase (New England Biolabs) and the PCR amplicons separated from primer dimers on a 0.5x TAE 1 % agarose gel. The respective bands were excised and DNA was purified from them using the QIAquick Gel Extraction Kit (Qiagen). The DNA concentration was measured with the Quant-iT dsDNA (HS) assay kit (Thermo Fisher Scientific) on a plate reader (Tecan Infinite 200

Pro). Next, the DNA was fragmented and the sequencing libraries were prepared using the Illumina Nextera XT kit (Illumina). In general, the manufacturer's protocol was followed, with two modifications. First, to prevent under-tagmentation, only 0.2 ng of DNA was used as input and the tagmentation step was performed for 15 min, instead of 5 min. Second, during library preparation, samples to be pooled were barcoded using the Nextera XT Index Kit v2 (Illumina). The final sequencing libraries were then purified using AMPure XP magnetic beads (Beckman Coulter) according to the manufacturer's protocol. A two-sided size selection was performed using 25 µl beads together with 50 µl input reaction during the first size selection step and 100 µl of beads during the second. Following library clean-up, the DNA concentration was measured again using the Quant-iT dsDNA (HS) assay kit (Thermo Fisher Scientific) and the different libraries were pooled at equal concentrations. Next, library quality was assessed on a Bioanalyzer (Agilent) using the Agilent DNA 1000 Kit. NGS was performed as paired-end Illumina MiSeq and NextSeq runs at the EMBL Gene Core facility.

4.1.10 Experimental characterization of individual variants from the domain insertion screen

Switchable variants were isolated from the sorted fractions and stored as glycerol stocks in 25 % glycerol (Carl Roth). Additionally selected variants were cloned following the protocol in section 4.1.1. The variants tested are listed in supplementary table 1. Precultures of Oneshot Top10 cells (Thermo Fisher Scientific) carrying a RFP reporter plasmid specific for the respective candidate switch, as well as a plasmid encoding the respective switchable variant, were inoculated from glycerol stocks into lysogeny broth (LB) (Carl Roth), supplemented with 50 µg/ml chloramphenicol (Carl Roth) and 25 µg/ml of kanamycin (Carl Roth). Cultures were prepared in technical triplicates in 96-well plates (Corning), using a volume of 200 µl per well. The precultures were incubated for 16 h at 37 °C, while shaking at 220 rpm. Main cultures were similarly prepared in 96-well plates, using LB supplemented with 50 µg/ml chloramphenicol and 25 µg/ml of kanamycin, using the inducer concentration indicated in section 4.1.8.2 for each candidate protein, respectively. The cultures were inoculated with 3 µl from the respective precultures and grown at 37 °C and 220 rpm for 16 h. Following incubation, RFP fluorescence and OD_{600} were measured in a plate reader (Tecan Infinite 200 Pro). For RFP measurements, an excitation wavelength of 490 nm and an emission wavelength of 520 nm were used. The reported RFP/OD600 values were calculated, by dividing the measured fluorescence by the OD_{600} levels. Three independent biological replicates were generated at three different days.

4.2 Computational methods

4.2.1 Analysis of AcrIIC3 inter-residue contacts

The inter-residue contacts for AcrIIC3 were identified using the PyMol 2.4 contact map visualizer (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.) based on a published AcrIIC3 structure (PDB-ID: 6J9N), applying a cutoff of 7 Å.

4.2.2 Structural modeling

4.2.2.1 *Modeling of CN-C3*

The structural modeling of CN-C3 was performed by Julius Upmeier zu Belzen and Zander Harteveld. This section is adapted from their protocol. The AcrIIC3-LOV2 hybrid models are based on experimental structures from the individual proteins, AcrIIC3 (PDB-ID: 6J9N) and the *As*LOV2 domain (PDB-ID: 2V0W). The three N-terminal amino acids of the LOV2 domain were not present in the CN-C3 constructs and thus deleted from the structure. Structural modeling was performed using the Rosetta remodel application (Huang *et al*, 2011). To this end, the terminal parts of the LOV2 domain including the glycine linkers, in case of CN-C3G, were rebuild using the fragment insertion with cyclic coordinate descent (Canutescu & Dunbrack, 2003) and kinematic closure (Mandell *et al*, 2009; Harper *et al*, 2003) with default parameters. For both protein variants, 1,000 decoys were generated, 206 and 236 of which passed the chain-break filter for CN-C3 and CN-C3G, respectively. Finally, these models were clustered using a root-mean-square deviation of 5 Å as threshold, resulting in 8 clusters for CN-C3 and 17 clusters for CN-C3G, respectively.

4.2.2.2 *Structure prediction with AlphaFold2*

Sructures of AraC, SigF, the TVMV protease, Flp, as well as the AraC-LOV2 fusions were predicted by AlphaFold2 (Jumper *et al*, 2021) using the Colabfold implementation (Mirdita *et al*, 2022). This implementation makes use of the MMseqs2 algorithm for the generation of multiple sequence alignments (Steinegger & Söding, 2017). Structures were predicted using the "colabfold_batch" command with the "MMseqs2 (UniRef+Environmental)" MSA preferences. For the proteins without insertion, 5 models were run with three recycling iterations. To reduce compute time, only one model was predicted for the AraC-LOV2 hybrids, using a single recycling step. Images of the models were generated using UCSF ChimeraX (version 1.4) (Goddard *et al*, 2018; Pettersen *et al*, 2021). To compute the position-wise RMSDs for between the AraC-LOV2 hybrids and the respective wildtype structures, the AF2 structures of AraC and the LOV2 domain were separately superimposed onto the prediction of the fusion proteins and RMSDs were calculated amino acid-wise. Computations were performed on the KIT Horeka cluster.

4.2.3 NGS and data analysis

To analyze the sequencing data, the fastq files were de-multiplexed using the Sabre tool (https://github.com/najoshi/sabre). The domain insertion frequencies were then calculated using a slightly modified version of the DIP-seq library (Nadler *et al*, 2016). Next the enrichment scores were determined using the following equation:

$$Enrichment\ score_i = log_2\left[\frac{count\ enriched_i}{\sum_i^n count\ enriched_i} \Big/ \frac{count\ initial_i}{\sum_i^n count\ initial_i}\right]$$

where n are the insertion positions within a given protein, $count\ enriched$ represents the read counts after enrichment and $count\ initial$ indicates the read counts of the initial library that was used as input to the sorting experiments. Insertions that were missing from the initial libraries were not taken into account during analysis. Insertion variants that entirely disappeared during sorting were assigned a value of -10, which was slightly below the lowest obtained enrichment scores.

To gather position-wise protein features, diverse sources were used. Biophysical properties and linker propensity indices were fetched from the AAindex database (Kawashima & Kanehisa, 2000). Information about secondary structure, accessible surface area and pLDDT score were extracted from the AF2-predicted structures. To map these features to the enrichment scores, the mean of the respective feature corresponding to the two amino acids that neighbor the insertion site were assigned to the enrichment. For the machine learning applications described below, the categorical features, such as secondary structures were binarized similar to one-hot encodings, with the difference that every position could have two possible positive labels (for instance if the secondary structure assignments of the two neighboring residues differ). The KLD, as well as the insertion and deletion statistics were based on sequence alignments. To this end, similar sequences were gathered using position-specific iterated basic local alignment search (PSI-BLAST) (Altschul *et al*, 1997, 1990), with an expect threshold of 0.01 and a PSI-BLAST threshold of 0.005. The maximum number of sequences was limited to 5000. Based on these sequences, an MSA was calculated with MUSCLE (Edgar, 2004), using the Super5 algorithm with standard parameters. Finally, the KLD was calculated by the following equation:

$$Divergance_i = \sum_a f_i(a) \cdot log_{10} \frac{f_i(a)}{b(a)}$$

where the divergence is determined for the position i and $f(a)$ is the frequency of the amino acid a at the given position, while $b(a)$ represents the background frequency of the amino acid. The background frequencies were defined as the AA frequencies in SwissProt (Bairoch & Apweiler, 1997). Of note, the definition of the gap background frequencies is non-trivial, as discussed by Teşileanu et al. (Teşileanu *et al*, 2015). Here, gaps were not included and the KLD is only based on AA frequencies. The position-wise insertion and deletion frequencies as well as the scores for the mean and median insertion lengths were calculated from pairwise alignments between the sequence of the protein of interest and its related sequences gathered by PSI-BLAST.

4.2.4 Gradient boosting models

In order to train predictive models on the insertion data, the enrichment scores were first binarized. All sites exhibiting a positive enrichment were assigned the label 1 and all sites with negative insertions were labeled 0. All position-wise properties collected during data analysis were used as features. In addition, each amino acid and each secondary structure element represented individual additional features. Dataset construction and model training were performed using the Scikit-learn framework (Pedregosa *et al*, 2011). Individual datasets for every candidate protein, as well as a complete dataset using the combined data of all four proteins were constructed. A 80:20 train-test split was applied and the features were min-max scaled prior to training. Gradient boosting classifiers (Friedman, 2002) were trained using five-fold cross-validation. The hyperparameters were optimized on the complete dataset using grid search. For the final model, 100 estimators were trained using squared error and a learning rate of 0.1. The maximum depth of the trees was limited to four and the exponential loss was chosen. The maximum number of features parameter was kept at "auto". The receiving operator characteristic and the average precision were chosen as performance metrics. The permutation importance and loss of impurity were calculated using the respective Scikit-learn functions.

In order to create the datasets including the structural context, the insertion site was defined as the center between the Cα atoms of the neighboring amino acids. The values of the biophysical features were calculated as the mean of the values from the amino acids lying within the chosen radius of the insertion site. Alignment-derived features were not processed that way, since the occurrence of natural insertions and deletion already depend on the structural context. The linker propensity indices represent a special case because they were established with respect to linker sequences. Here, the mean of the sequence with a length of the defined context, symmetrically surrounding the insertion site, was used. The models based on these datasets were trained as described before. The same hyperparameters as for the original model were used.

4.2.5 Statistical coupling analysis of AraC

For SCA, the same MSA as described in section 4.2.3 was employed. The data was processed and analysis was performed as previously described by Rivoire et al. (Rivoire *et al*, 2016). The predicted sector residues were mapped onto the AF2 predicted structure of AraC (Supp. fig. 11).

4.2.6 Statistical analysis

Bars usually indicate the mean from three individual experiments, unless otherwise stated in the figure/legends. Error bars represent the standard deviation. In boxplots, the IQR is marked by the box and the median is represented by a line within the box. Whiskers extend to the 1.5-fold interquartile range (IQR) or to the value of the smallest or largest enrichment, respectively. Biological replicates were performed as independent experiments on different days. Differences in mean values were assessed for statistical significance by Bonferroni-corrected post-hoc one-way ANOVA. The p-values corresponding to the asterisks are reported in the figure legends.

4.2.7 Software

Structures were analyzed and images were rendered using UCSF ChimeraX (version 1.4) (Goddard *et al*, 2018; Pettersen *et al*, 2021) and PyMol (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.). Data analysis was performed using Python 3 (version 3.7.12) (Van Rossum *et al*, 2009), Biopython (version 1.79) (Cock *et al*, 2009), NumPy (version 1.21.6) (Harris *et al*, 2020) and Pandas (version 1.3.5) (McKinney, 2010). Plots were generated using the Matplotlib (version 3.2.2) (Hunter, 2007) and Seaborn (version 0.11) packages (Waskom, 2021). The machine learning models were trained using the Scikit-learn framework (version 1.0.2) (Pedregosa *et al*, 2011). Protein structures were predicted using the Colabfold implementation (Mirdita *et al*, 2022) of AF2 (Jumper *et al*, 2021). SCA was done using the pySCA package (Rivoire *et al*, 2016). ImageJ (version 1.46) (Rueden *et al*, 2017; Schneider *et al*, 2012) was used for image analysis. FACS data was analyzed using the Cytoflow software (version 2.1) (https://cytoflow.github.io/). Figures were created in Affinity Designer (version 1.10.5) and Zotero (version 1.6.19) was employed as reference manager. The thesis was written in Microsoft Word (Office 2021).

5 References

Abudayyeh OO, Gootenberg JS, Essletzbichler P, Han S, Joung J, Belanto JJ, Verdine V, Cox DBT, Kellner MJ, Regev A, *et al* (2017) RNA targeting with CRISPR-Cas13. *Nature* 550: 280–284

Abudayyeh OO, Gootenberg JS, Konermann S, Joung J, Slaymaker IM, Cox DBT, Shmakov S, Makarova KS, Semenova E, Minakhin L, *et al* (2016) C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* 353

Adli M (2018) The CRISPR tool kit for genome editing and beyond. *Nat Commun* 9, 1911

Ahdritz G, Bouatta N, Kadyan S, Xia Q, Gerecke W, O'Donnell TJ, Berenberg D, Fisk I, Zanichelli N, Zhang B, *et al* (2022) OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. 2022.11.20.517210 doi:10.1101/2022.11.20.517210 [PREPRINT]

Akdel M, Pires DEV, Pardo EP, Jänes J, Zalevsky AO, Mészáros B, Bryant P, Good LL, Laskowski RA, Pozzati G, *et al* (2022) A structural biology community assessment of AlphaFold2 applications. *Nat Struct Mol Biol* 29: 1056–1067

Akerboom J, Chen TW, Wardill TJ, Tian L, Marvin JS, Mutlu S, Calderón NC, Esposti F, Borghuis BG, Sun XR, *et al* (2012) Optimization of a GCaMP calcium indicator for neural activity imaging. *J Neurosci* 32: 13819–13840

Akerboom J, Rivera JDV, Rodríguez Guilbe MM, Malavé ECA, Hernandez HH, Tian L, Hires SA, Marvin JS, Looger LL & Schreiter ER (2009) Crystal structures of the GCaMP calcium sensor reveal the mechanism of fluorescence signal change and aid rational design. *J Biol Chem* 284: 6455–6464

del Alamo D, Sala D, Mchaourab HS & Meiler J (2022) Sampling alternative conformational states of transporters and receptors with AlphaFold2. *eLife* 11: e75751

Alley EC, Khimulya G, Biswas S, AlQuraishi M & Church GM (2019) Unified rational protein engineering with sequence-based deep representation learning. *Nat Methods* 16: 1315–1322

AlQuraishi M (2019) End-to-End Differentiable Learning of Protein Structure. *Cell Syst* 8: 292-301.e3

Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215: 403–410

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W & Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25: 3389–3402

Alva V, Remmert M, Biegert A, Lupas AN & Söding J (2010) A galaxy of folds. *Protein Sci* 19: 124–130

Amabile A, Migliara A, Capasso P, Biffi M, Cittaro D, Naldini L & Lombardo A (2016) Inheritable Silencing of Endogenous Genes by Hit-and-Run Targeted Epigenetic Editing. *Cell* 167: 219-232.e14

Amrani N, Gao XD, Liu P, Edraki A, Mir A, Ibraheim R, Gupta A, Sasaki KE, Wu T, Donohoue PD, *et al* (2018) NmeCas9 is an intrinsically high-fidelity genome-editing platform Jin-Soo Kim. *Genome Biol* 19: 1–25

André I, Bradley P, Wang C & Baker D (2007) Prediction of the structure of symmetrical protein assemblies. *Proc Natl Acad Sci* 104: 17656–17661

Anfinsen CB (1973) Principles that Govern the Folding of Protein Chains. *Science* 181: 223–230

Anishchenko I, Pellock SJ, Chidyausiku TM, Ramelot TA, Ovchinnikov S, Hao J, Bafna K, Norn C, Kang A, Bera AK, *et al* (2021) De novo protein design by deep network hallucination. *Nature* 600: 547–552

Anzalone AV, Randolph PB, Davis JR, Sousa AA, Koblan LW, Levy JM, Chen PJ, Wilson C, Newby GA, Raguram A, *et al* (2019) Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* 576: 149–157

Apic G, Gough J & Teichmann SA (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J Mol Biol* 310: 311–325

Apic G & Russell RB (2010) Domain Recombination: A Workhorse for Evolutionary Innovation. *Sci Signal* 3

Arslan Z, Hermanns V, Wurm R, Wagner R & Pul Ü (2014) Detection and characterization of spacer integration intermediates in type I-E CRISPR-Cas system. *Nucleic Acids Res* 42: 7884–7893

Aschenbrenner S, Kallenberger SM, Hoffmann MD, Huck A, Eils R & Niopek D (2020) Coupling Cas9 to artificial inhibitory domains enhances CRISPR-Cas9 target specificity. *Sci Adv* 6: eaay0187

Ash C, Dubec M, Donne K & Bashford T (2017) Effect of wavelength and beam width on penetration in light-tissue interaction using computational methods. *Lasers Med Sci* 32: 1909–1918

Athukoralage JS, McMahon SA, Zhang C, Grüschow S, Graham S, Krupovic M, Whitaker RJ, Gloster TM & White MF (2020) An anti-CRISPR viral ring nuclease subverts type III CRISPR immunity. *Nature* 577: 572–575

Badran AH & Liu DR (2015) Development of potent in vivo mutagenesis plasmids with broad mutational spectra. *Nat Commun* 6: 1–10

Bae K, Mallick BK & Elsik CG (2005) Prediction of protein interdomain linker regions by a hidden Markov model. *Bioinformatics* 21: 2264–2270

Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, *et al* (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373: 871–876

Baird GS, Zacharias DA & Tsien RY (1999) Circular permutation and receptor insertion within green fluorescent proteins. *Proc Natl Acad Sci U S A* 96: 11241–11246

Bairoch A & Apweiler R (1997) The SWISS-PROT protein sequence database: its relevance to human molecular medical research. *J Mol Med Berl Ger* 75: 312–316

Balzer S, Kucharova V, Megerle J, Lale R, Brautaset T & Valla S (2013) A comparative analysis of the properties of regulated promoter systems commonly used for recombinant gene expression in Escherichia coli. *Microb Cell Factories* 12: 26

Barlow KA, Ó Conchúir S, Thompson S, Suresh P, Lucas JE, Heinonen M & Kortemme T (2018) Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein-Protein Binding Affinity upon Mutation. *J Phys Chem B* 122: 5389–5399

Baron M, Norman DG & Campbell ID (1991) Protein modules. *Trends Biochem Sci* 16: 13–17

Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA & Horvath P (2007) CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science* 315: 1709–1712

Basgall EM, Goetting SC, Goeckel ME, Giersch RM, Roggenkamp E, Schrock MN, Halloran M & Finnigan GC (2018) Gene drive inhibition by the anti-CRISPR proteins AcrIIA2 and AcrIIA4 in saccharomyces cerevisiae. *Microbiol U K* 164: 464–474

Baumschlager A, Aoki SK & Khammash M (2017) Dynamic Blue Light-Inducible T7 RNA Polymerases (Opto-T7RNAPs) for Precise Spatiotemporal Gene Expression Control. *ACS Synth Biol* 6: 2157–2167

Bedbrook CN, Yang KK, Robinson JE, Mackey ED, Gradinaru V & Arnold FH (2019) Machine learning-guided channelrhodopsin engineering enables minimally invasive optogenetics. *Nat Methods* 16: 1176–1184

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN & Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* 28: 235–242

Berrondo M, Ostermeier M & Gray JJ (2008) Structure Prediction of Domain Insertion Proteins from Structures of Individual Domains. *Structure* 16: 513–527

Bervoets I, Van Brempt M, Van Nerom K, Van Hove B, Maertens J, De Mey M & Charlier D (2018) A sigma factor toolbox for orthogonal gene expression in Escherichia coli. *Nucleic Acids Res* 46: 2133–2144

Bhaskaran R & Ponnuswamy P k. (1988) Positional flexibilities of amino acid residues in globular proteins. *Int J Pept Protein Res* 32: 241–255

Bibikova M, Carroll D, Segal DJ, Trautman JK, Smith J, Kim YG & Chandrasegaran S (2001) Stimulation of homologous recombination through targeted cleavage by chimeric nucleases. *Mol Cell Biol* 21: 289–297

Bienert S, Waterhouse A, de Beer TAP, Tauriello G, Studer G, Bordoli L & Schwede T (2017) The SWISS-MODEL Repository-new features and functionality. *Nucleic Acids Res* 45: D313–D319

Bileschi ML, Belanger D, Bryant DH, Sanderson T, Carter B, Sculley D, Bateman A, DePristo MA & Colwell LJ (2022) Using deep learning to annotate the protein universe. *Nat Biotechnol*

Biondi RM, Baehler PJ, Reymond CD & Véron M (1998) Random insertion of GFP into the cAMP-dependent protein kinase regulatory subunit from Dictyostelium discoideum. *Nucleic Acids Res* 26: 4946–4952

Blacklock KM, Yang L, Mulligan VK & Khare SD (2018) A computational method for the design of nested proteins by loop-directed domain insertion. *Proteins Struct Funct Bioinforma* 86: 354–369

Blain-Hartung M, Rockwell NC, Moreno MV, Martin SS, Gan F, Bryant DA & Lagarias JC (2018) Cyanobacteriochrome-based photoswitchable adenylyl cyclases (cPACs) for broad spectrum light regulation of cAMP levels in cells. *J Biol Chem* 293: 8473–8483

Bloom JD & Arnold FH (2009) In the light of directed evolution: Pathways of adaptive protein evolution. *Proc Natl Acad Sci* 106: 9995–10000

Boch J, Scholze H, Schornack S, Landgraf A, Hahn S, Kay S, Lahaye T, Nickstadt A & Bonas U (2009) Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* 326: 1509–1512

Boehr DD, Nussinov R & Wright PE (2009) The role of dynamic conformational ensembles in biomolecular recognition. *Nat Chem Biol* 5: 789–796

Bolotin A, Quinquis B, Sorokin A & Ehrlich SD 2005 Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* 151: 2551–2561

Bondy-Denomy J, Garcia B, Strum S, Du M, Rollins MF, Hidalgo-Reyes Y, Wiedenheft B, Maxwell KL & Davidson AR (2015) Multiple mechanisms for CRISPR-Cas inhibition by anti-CRISPR proteins. *Nature* 526: 136–139

Bondy-Denomy J, Pawluk A, Maxwell KL & Davidson AR (2013) Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. *Nature* 493: 429–432

Bonger KM, Rakhit R, Payumo AY, Chen JK & Wandless TJ (2014) General Method for Regulating Protein Stability with Light. *ACS Chem Biol* 9: 111–115

Borges AL, Davidson AR & Bondy-Denomy J (2017) The Discovery, Mechanisms, and Evolutionary Impact of Anti-CRISPRs. *Annu Rev Virol* 4: annurev-virology-101416-041616

Börner K, Niopek D, Cotugno G, Kaldenbach M, Pankert T, Willemsen J, Zhang X, Schürmann N, Mockenhaupt S, Serva A, *et al* (2013) Robust RNAi enhancement via human Argonaute-2 overexpression from plasmids, viral vectors and cell lines. *Nucleic Acids Res* 41: e199

Boyden ES, Zhang F, Bamberg E, Nagel G & Deisseroth K (2005) Millisecond-timescale, genetically targeted optical control of neural activity. *Nat Neurosci* 8: 1263–1268

Brinkman EK, Chen T, Amendola M & Van Steensel B (2014) Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Res* 42: 1–8

Brooks BR, Brooks CL, MacKerell AD, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, *et al* (2009) CHARMM: The Biomolecular Simulation Program. *J Comput Chem* 30: 1545–1614

Brouns SJJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJH, Snijders APL, Dickman MJ, Makarova KS, Koonin EV & van der Oost J (2008) Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. *Science* 321: 960–964

Brown MJ & Schleif R (2019) Helical Behavior of the Interdomain Linker of the Escherichia coli AraC Protein. *Biochemistry* 58: 2867–2874

Bubeck F, Hoffmann MD, Harteveld Z, Aschenbrenner S, Bietz A, Waldhauer MC, Börner K, Fakhiri J, Schmelas C, Dietz L, *et al* (2018) Engineered anti-CRISPR proteins for optogenetic control of CRISPR–Cas9. *Nat Methods* 15: 924–927

Buel GR & Walters KJ (2022) Can AlphaFold2 predict the impact of missense mutations on structure? *Nat Struct Mol Biol* 29: 1–2

Busby S & Ebright RH (1999) Transcription activation by catabolite activator protein (CAP). *J Mol Biol* 293: 199–213

Callaway E (2022) What's next for AlphaFold and the AI protein-folding revolution. *Nature* 604: 234–238

Campbell RE, Tour O, Palmer AE, Steinbach PA, Baird GS, Zacharias DA & Tsien RY (2002) A monomeric red fluorescent protein. *Proc Natl Acad Sci* 99: 7877–7882

Camsund D, Jaramillo A & Lindblad P (2021) Engineering of a Promoter Repressed by a Light-Regulated Transcription Factor in Escherichia coli. *BioDesign Res* 2021

Canutescu AA & Dunbrack RL (2003) Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Sci* 12: 963–972

Capecchi MR (1989) Altering the genome by homologous recombination. *Science* 244: 1288–1292

Carte J, Wang R, Li H, Terns RM & Terns MP (2008) Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes Dev* 22: 3489–3496

Ceroni F, Boo A, Furini S, Gorochowski TE, Borkowski O, Ladak YN, Awan AR, Gilbert C, Stan G-B & Ellis T (2018) Burden-driven feedback control of gene expression. *Nat Methods* 15: 387–393

Chaudhary K, Chattopadhyay A & Pratap D (2018) Anti-CRISPR proteins: Counterattack of phages on bacterial defense (CRISPR/Cas) system. *J Cell Physiol* 233: 57–59

Chen B, Gilbert LA, Cimini BA, Schnitzbauer J, Zhang W, Li GW, Park J, Blackburn EH, Weissman JS, Qi LS, *et al* (2013) Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell* 155: 1479–1491

Chen Y, Narendra U, Iype LE, Cox MM & Rice PA (2000) Crystal structure of a Flp recombinase-Holliday junction complex: assembly of an active oligomer by helix swapping. *Mol Cell* 6: 885–897

Chen Z, Kibler RD, Hunt A, Busch F, Pearl J, Jia M, VanAernum ZL, Wicky BIM, Dods G, Liao H, *et al* (2020) De novo design of protein logic gates. *Science* 368: 78–84

Chen Z, Linton JM, Zhu R & Elowitz MB (2022) A synthetic protein-level neural network in mammalian cells. 2022.07.10.499405 doi:10.1101/2022.07.10.499405 [PREPRINT]

Cheng AW, Wang H, Yang H, Shi L, Katz Y, Theunissen TW, Rangarajan S, Shivalila CS, Dadon DB & Jaenisch R (2013) Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system. *Cell Res* 23: 1163–1171

Chivian D & Baker D (2006) Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. *Nucleic Acids Res* 34: e112

Choi J, Chen J, Schreiber SL & Clardy J (1996) Structure of the FKBP12-Rapamycin Complex Interacting with Binding Domain of Human FRAP. *Science* 273: 239–242

Choi JH, Laurent AH, Hilser VJ & Ostermeier M (2015) Design of protein switches based on an ensemble model of allostery. *Nat Commun* 6: 1–9

Choi JH & Ostermeier M (2015) Rational Design of a Fusion Protein to Exhibit Disulfide-Mediated Logic Gate Behavior. *ACS Synth Biol* 4: 400–406

Chowdhury R, Bouatta N, Biswas S, Floristean C, Kharkare A, Roye K, Rochereau C, Ahdritz G, Zhang J, Church GM, *et al* (2022) Single-sequence protein structure prediction using a language model and deep learning. *Nat Biotechnol*

Chowdhury S, Carter J, Rollins MF, Golden SM, Jackson RN, Hoffmann C, Nosaka L, Bondy-Denomy J, Maxwell KL, Davidson AR, *et al* (2017) Structure Reveals Mechanisms of Viral Suppressors that Intercept a CRISPR RNA-Guided Surveillance Complex. *Cell* 169: 47-57.e11

Chu PH, Tsygankov D, Berginski ME, Dagliyan O, Gomez SM, Elston TC, Karginov AV & Hahn KM (2014) Engineered kinase activation reveals unique morphodynamic phenotypes and associated trafficking for Src family isoforms. *Proc Natl Acad Sci U S A* 111: 12420–12425

Chusacultanachai S & Yuthavong Y (2004) Random Mutagenesis Strategies for Construction of Large and Diverse Clone Libraries of Mutated DNA Fragments. In *Parasite Genomics Protocols*, Melville SE (ed) pp 319–333. Totowa, NJ: Humana Press

Clarkson MW, Gilmore SA, Edgell MH & Lee AL (2006) Dynamic coupling and allosteric behavior in a nonallosteric protein. *Biochemistry* 45: 7693–7699

Clauwaert J & Waegeman W (2020) Novel transformer networks for improved sequence labeling in genomics.

Clement K, Rees H, Canver MC, Gehrke JM, Farouni R, Hsu JY, Cole MA, Liu DR, Joung JK, Bauer DE, *et al* (2019) CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nat Biotechnol* 37: 224–226

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, *et al* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25: 1422–1423

Collinet B, Hervé M, Pecorari F, Minard P, Eder O & Desmadril M (2000) Functionally Accepted Insertions of Proteins within Protein Domains. *J Biol Chem* 275: 17428–17433

Cong L, Ran FA, Cox D, Lin S, Barretto R, Hsu PD, Wu X, Jiang W & Marraffini L a (2013) Multiplex Genome Engineering Using CRISPR/Cas Systems. Science (New York, N.Y.). *Science* 339: 819–823

Conway AB, Chen Y & Rice PA (2003) Structural Plasticity of the Flp–Holliday Junction Complex. *J Mol Biol* 326: 425–434

Cortés-Avalos D, Martínez-Pérez N, Ortiz-Moncada MA, Juárez-González A, Baños-Vargas AA, Estrada-de los Santos P, Pérez-Rueda E & Ibarra JA (2021) An update of the unceasingly growing and diverse AraC/XylS family of transcriptional activators. *FEMS Microbiol Rev* 45: fuab020

Courbet A, Hansen J, Hsia Y, Bethel N, Boyken SE, Ueda G, Nattermann U, Nagarajan D, Silva D, Sheffler W, *et al* (2022) Computational design of mechanically coupled axle-rotor protein assemblies. 9

Cox DBT, Gootenberg JS, Abudayyeh OO, Franklin B, Kellner MJ, Joung J & Zhang F (2017) RNA Editing with CRISPR-Cas13. *Science* 358: 1019–1027

Coyote-Maestas W, He Y, Myers CL & Schmidt D (2019) Domain insertion permissibility-guided engineering of allostery in ion channels. *Nat Commun* 10: 1–14

Coyote-maestas W, Nedrud D, Okorafor S, He Y & Schmidt D (2019) Targeted insertional mutagenesis libraries for deep domain insertion profiling. *Nucleic Acids Res* 48: 1–14

Coyote-Maestas W, Nedrud D, Suma A, He Y, Matreyek KA, Fowler DM, Carnevale V, Myers CL & Schmidt D (2020) The biophysical basis of protein domain compatibility. 2020.12.09.418442 doi:10.1101/2020.12.09.418442 [PREPRINT]

Coyote-Maestas W, Nedrud D, Suma A, He Y, Matreyek KA, Fowler DM, Carnevale V, Myers CL & Schmidt D (2021) Probing ion channel functional architecture and domain recombination compatibility by massively parallel domain insertion profiling. *Nat Commun* 12: 7114

Crosson S & Moffat K (2002) Photoexcited structure of a plant photoreceptor domain reveals a light-driven molecular switch. *Plant Cell* 14: 1067–1075

Dagliyan O, Dokholyan NV & Hahn KM (2019) Engineering proteins for allosteric control by light or ligands. *Nat Protoc*: 1–21

Dagliyan O, Krokhotin A, Ozkan-Dagliyan I, Deiters A, Der CJ, Hahn KM & Dokholyan NV (2018) Computational design of chemogenetic and optogenetic split proteins. *Nat Commun* 9: 4042

Dagliyan O, Shirvanyants D, Karginov AV, Ding F, Fee L, Chandrasekaran SN, Freisinger CM, Smolen GA, Huttenlocher A, Hahn KM, *et al* (2013) Rational design of a ligand-controlled protein conformational switch. *Proc Natl Acad Sci* 110: 6800–6804

Dagliyan O, Tarnawski M, Chu PH, Shirvanyants D, Schlichting I, Dokholyan NV & Hahn KM (2016) Engineering extrinsic disorder to control protein activity in living cells. *Science* 354: 1441–1444

Daniel R, Rubens JR, Sarpeshkar R & Lu TK (2013) Synthetic analog computation in living cells. *Nature* 497: 619–623

Das R & Baker D (2008) Macromolecular modeling with rosetta. *Annu Rev Biochem* 77: 363–382

Dauparas J, Anishchenko I, Bennett N, Bai H, Ragotte RJ, Milles LF, Wicky BIM, Courbet A, de Haas RJ, Bethel N, *et al* (2022) Robust deep learning–based protein sequence design using ProteinMPNN. *Science* 378: 49–56

Davis BH, Poon AFY & Whitlock MC (2009) Compensatory mutations are repeatable and clustered within proteins. *Proc R Soc B Biol Sci* 276: 1823–1827

Davis KM, Pattanayak V, Thompson DB, Zuris JA & Liu DR (2015) Small molecule-triggered Cas9 protein with improved genome-editing specificity. *Nat Chem Biol* 11: 316–318

Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, Eckert MR, Vogel J & Charpentier E (2011) CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* 471: 602–607

Deveau H, Barrangou R, Garneau JE, Labonté J, Fremaux C, Boyaval P, Romero DA, Horvath P & Moineau S (2008) Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *J Bacteriol* 190: 1390–1400

Devlin J, Chang M-W, Lee K & Toutanova K (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (http://arxiv.org/abs/1810.04805) [PREPRINT]

Dietler J, Schubert R, Krafft TGA, Meiler S, Kainrath S, Richter F, Schweimer K, Weyand M, Janovjak H & Möglich A (2021) A Light-Oxygen-Voltage Receptor Integrates Light and Temperature. *J Mol Biol* 433: 167107

Dill KA & MacCallum JL (2012) The Protein-Folding Problem, 50 Years On. *Science* 338: 1042–1046

Dirla S, Chien JYH & Schleif R (2009) Constitutive mutations in the Escherichia coli AraC protein. *J Bacteriol* 191: 2668–2674

Diwan GD, Gonzalez-Sanchez JC, Apic G & Russell RB (2021) Next Generation Protein Structure Predictions and Genetic Variant Interpretation. *J Mol Biol* 433: 167180

Doi N & Yanagawa H (1999) Design of generic biosensors based on green fluorescent proteins with allosteric sites by directed evolution. *FEBS Lett* 453: 305–307

Dokholyan NV (2016) Controlling Allosteric Networks in Proteins. *Chem Rev* 116: 6463–6487

Dokholyan NV (2021) Nanoscale programming of cellular and physiological phenotypes: inorganic meets organic programming. *NPJ Syst Biol Appl* 7: 15

Dong C, Hao G-F, Hua H-L, Liu S, Labena AA, Chai G, Huang J, Rao N & Guo F-B (2018) Anti-CRISPRdb: a comprehensive online resource for anti-CRISPR proteins. *Nucleic Acids Res* 46: D393–D398

Dong D, Ren K, Qiu X, Zheng J, Guo M, Guan X, Liu H, Li N, Zhang B, Yang D, *et al* (2016) The crystal structure of Cpf1 in complex with CRISPR RNA. *Nature* 532: 522–526

Dong L, Guan X, Li N, Zhang F, Zhu Y, Ren K, Yu L, Zhou F, Han Z, Gao N, *et al* (2019) An anti-CRISPR protein disables type V Cas12a by acetylation. *Nat Struct Mol Biol* 26: 308–314

Dueber JE, Yeh BJ, Chak K & Lim WA (2003) Reprogramming Control of an Allosteric Signaling Switch Through Modular Recombination. *Science* 301: 1904–1908

Dunham AS & Beltrao P (2021) Exploring amino acid functions in a deep mutational landscape. *Mol Syst Biol* 17: e10305

Dyson HJ & Wright PE (2005) Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* 6: 197–208

East-Seletsky A, O'Connell MR, Knight SC, Burstein D, Cate JHD, Tjian R & Doudna JA (2016) Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection. *Nature* 538: 270–273

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32: 1792–1797

Edraki A, Mir A, Ibraheim R, Gainetdinov I, Yoon Y, Song C-Q, Cao Y, Gallant J, Xue W, Rivera-Pérez JA, *et al* (2018) A Compact, High-Accuracy Cas9 with a Dinucleotide PAM for In Vivo Genome Editing. *Mol Cell*: 1–13

Edwards WR, Busse K, Allemann RK & Jones DD (2008) Linking the functions of unrelated proteins using a novel directed evolution domain insertion method. *Nucleic Acids Res* 36

Eldar A & Elowitz MB (2010) Functional roles for noise in genetic circuits. *Nature* 467: 167–173

Elowitz MB & Leibier S (2000) A synthetic oscillatory network of transcriptional regulators. *Nature* 403: 335–338

Engler C, Kandzia R & Marillonnet S (2008) A One Pot, One Step, Precision Cloning Method with High Throughput Capability. *PLOS ONE* 3: e3647

Esvelt KM, Carlson JC & Liu DR (2011) A system for the continuous directed evolution of biomolecules. *Nature* 472: 499–503

Esvelt KM, Mali P, Braff JL, Moosburner M, Yaung SJ & Church GM (2013) Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nat Methods* 10: 1116–1123

Etzl S, Lindner R, Nelson MD & Winkler A (2018) Structure-guided design and functional characterization of an artificial red light–regulated guanylate/adenylate cyclase for optogenetic applications. *J Biol Chem* 293: 9078–9089

Eustance RJ, Bustos SA & Schleif RF (1994) Reaching Out: Locating and Lengthening the Interdomain Linker in AraC Protein. *J Mol Biol* 242: 330–338

Fan F, Binkowski BF, Butler BL, Stecha PF, Lewis MK & Wood KV (2008) Novel Genetically Encoded Biosensors Using Firefly Luciferase. *ACS Chem Biol* 3: 346–351

Faure AJ, Domingo J, Schmiedel JM, Hidalgo-Carcedo C, Diss G & Lehner B (2022) Mapping the energetic and allosteric landscapes of protein binding domains. *Nature* 604: 175–183

Feil R, Wagner J, Metzger D & Chambon P (1997) Regulation of Cre recombinase activity by mutated estrogen receptor ligand-binding domains. *Biochem Biophys Res Commun* 237: 752–757

Fernandez-Rodriguez J & Voigt CA (2016) Post-translational control of genetic circuits using Potyvirus proteases. *Nucleic Acids Res* 44: 6493–6502

Ferry QRV, Lyutova R & Fulga TA (2017) Rational design of inducible CRISPR guide RNAs for de novo assembly of transcriptional programs. *Nat Commun* 8: 14633

Finkelstein AV, Gutun AM & Badretdinov AYa (1993) Why are the same protein folds used to perform different functions? *FEBS Lett* 325: 23–28

Finlayson L, Barnard IRM, McMillan L, Ibbotson SH, Brown CTA, Eadie E & Wood K (2022) Depth Penetration of Light into Skin as a Function of Wavelength from 200 to 1000 nm. *Photochem Photobiol* 98: 974–981

Freimuth PI, Taylor JW & Kaiser ET (1990) Introduction of guest peptides into Escherichia coli alkaline phosphatase. Excision and purification of a dynorphin analogue from an active chimeric protein. *J Biol Chem* 265: 896–901

Friedland AE, Baral R, Singhal P, Loveluck K, Shen S, Sanchez M, Marco E, Gotta GM, Maeder ML, Kennedy EM, *et al* (2015) Characterization of Staphylococcus aureus Cas9: A smaller Cas9 for all-in-one adeno-associated virus delivery and paired nickase applications. *Genome Biol* 16: 1–10

Friedman JH (2002) Stochastic gradient boosting. *Comput Stat Data Anal* 38: 367–378

Frock RL, Hu J, Meyers RM, Ho Y-J, Kii E & Alt FW (2015) Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. *Nat Biotechnol* 33: 179–186

Fu Y, Foden JA, Khayter C, Maeder ML, Reyon D, Joung JK & Sander JD (2013) High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat Biotechnol* 31: 822–826

Gainza P, Sverrisson F, Monti F, Rodolà E, Boscaini D, Bronstein MM & Correia BE (2019) Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat Methods* 17

Gao XJ, Chong LS, Kim MS & Elowitz MB (2018) Programmable protein circuits in living cells. *Science* 361: 1252–1258

Gao Y, Xiong X, Wong S, Charles EJ, Lim WA & Qi LS (2016) Complex transcriptional modulation with orthogonal and inducible dCas9 regulators. *Nat Methods* 13: 1043–1049

Garcia B, Lee J, Edraki A, Hidalgo-Reyes Y, Erwood S, Mir A, Trost CN, Seroussi U, Stanley SY, Cohn RD, *et al* (2019) Anti-CRISPR AcrIIA5 Potently Inhibits All Cas9 Homologs Used for Genome Editing. *Cell Rep* 29: 1739-1746.e5

Gardner TS, Cantor CR & Collins JJ (2000) Construction of a genetic toggle switch in Escherichia coli. *Nature* 403: 339–342

Gasiunas G, Barrangou R, Horvath P & Siksnys V (2012) Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc Natl Acad Sci* 109: E2579–E2586

Gaudelli NM, Komor AC, Rees HA, Packer MS, Badran AH, Bryson DI & Liu DR (2017) Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* 551: 464–471

Gebauer F & Hentze MW (2004) Molecular mechanisms of translational control. *Nat Rev Mol Cell Biol* 5: 827–835

George RA & Heringa J (2002) An analysis of protein domain linkers: their classification and role in protein folding. *Protein Eng Des Sel* 15: 871–879

Gesner EM, Schellenberg MJ, Garside EL, George MM & Macmillan AM (2011) Recognition and maturation of effector RNAs in a CRISPR interference pathway. *Nat Struct Mol Biol* 18: 688–692

Ghanbarpour A, Pinger C, Esmatpour Salmani R, Assar Z, Santos EM, Nosrati M, Pawlowski K, Spence D, Vasileiou C, Jin X, *et al* (2019) Engineering the hCRBPII Domain-Swapped Dimer into a New Class of Protein Switches. *J Am Chem Soc* 141: 17125–17132

Gil AA, Carrasco-López C, Zhu L, Zhao EM, Ravindran PT, Wilson MZ, Goglia AG, Avalos JL & Toettcher JE (2020) Optogenetic control of protein binding using light-switchable nanobodies. *Nat Commun* 11: 1–12

Gilbert LA, Larson MH, Morsut L, Liu Z, Brar GA, Torres SE, Stern-Ginossar N, Brandman O, Whitehead EH, Doudna JA, *et al* (2013) CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. *Cell* 154: 442–451

Glantz ST, Carpenter EJ, Melkonian M, Gardner KH, Boyden ES, Wong GK-S & Chow BY (2016) Functional and topological diversity of LOV domain photoreceptors. *Proc Natl Acad Sci* 113

Goddard TD, Huang CC, Meng EC, Pettersen EF, Couch GS, Morris JH & Ferrin TE (2018) UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci Publ Protein Soc* 27: 14–25

Goverde C, Wolf B, Khakzad H, Rosset S & Correia BE (2022) De novo protein design by inversion of the AlphaFold structure prediction network. 2022.12.13.520346 doi:10.1101/2022.12.13.520346 [PREPRINT]

Gräwe A, Merkx M & Stein V (2022) iFLinkC-X: A Scalable Framework to Assemble Bespoke Genetically Encoded Co-polymeric Linkers of Variable Lengths and Amino Acid Composition. *Bioconjug Chem* 33: 1415–1421

Gräwe A & Stein V (2020) Linker Engineering in the Context of Synthetic Protein Switches and Sensors. *Trends Biotechnol*: 731–744

Greener A, Callahan M & Jerpseth B (1997) An efficient random mutagenesis technique using anE.coli mutator strain. *Mol Biotechnol* 7: 189–195

Gunasekaran K, Ma B & Nussinov R (2004) Is allostery an intrinsic property of all dynamic proteins? *Proteins Struct Funct Bioinforma* 57: 433–443

Guntas G, Hallett RA, Zimmerman SP, Williams T, Yumerefendi H, Bear JE & Kuhlman B (2015) Engineering an improved light-induced dimer (iLID) for controlling the localization and activity of signaling proteins. *Proc Natl Acad Sci* 112: 112–117

Guntas G, Mitchell SF & Ostermeier M (2004) A Molecular Switch Created by In Vitro Recombination of Nonhomologous Genes. *Chem Biol* 11: 1483–1487

Guntas G & Ostermeier M (2004) Creation of an Allosteric Enzyme by Domain Insertion. *J Mol Biol* 336: 263–273

Gussow AB, Park AE, Borges AL, Shmakov SA, Makarova KS, Wolf YI, Bondy-Denomy J & Koonin EV (2020) Machine-learning approach expands the repertoire of anti-CRISPR protein families. *Nat Commun* 11: 1–12

Guzman LM, Belin D, Carson MJ & Beckwith J (1995) Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *J Bacteriol* 177: 4121–4130

Ha JH, Butler JS, Mitrea DM & Loh SN (2006) Modular enzyme design: Regulation by mutually exclusive protein folding. *J Mol Biol* 357: 1058–1062

Ha JH, Karchin JM, Walker-Kopp N, Castañeda CA & Loh SN (2015) Engineered domain swapping as an on/off switch for protein function. *Chem Biol* 22: 1384–1393

Halabi N, Rivoire O, Leibler S & Ranganathan R (2009) Protein sectors: evolutionary units of three-dimensional structure. *Cell* 138: 774–86

Halavaty AS & Moffat K (2007) N- and C-terminal flanking regions modulate light-induced signal transduction in the LOV2 domain of the blue light sensor phototropin 1 from Avena sativa. *Biochemistry* 46: 14001–14009

Han T, Chen Q & Liu H (2017) Engineered Photoactivatable Genetic Switches Based on the Bacterium Phage T7 RNA Polymerase. *ACS Synth Biol* 6: 357–366

Harbury PB, Plecs JJ, Tidor B, Alber T & Kim PS (1998) High-Resolution Protein Design with Backbone Freedom. *Science* 282: 1462–1467

Harmer T, Wu M & Schleif R (2001) The role of rigidity in DNA looping-unlooping by AraC. *Proc Natl Acad Sci* 98: 427–431

Harper SM, Neil LC & Gardner KH (2003) Structural Basis of a Phototropin Light Switch. *Science* 301: 1541–1545

Harrington LB, Doxzen KW, Ma E, Liu JJ, Knott GJ, Edraki A, Garcia B, Amrani N, Chen JS, Cofsky JC, *et al* (2017) A Broad-Spectrum Inhibitor of CRISPR-Cas9. *Cell* 170: 1224-1233.e15

Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, *et al* (2020) Array programming with NumPy. *Nature* 585: 357–362

Haurwitz RE, Jinek M, Wiedenheft B, Zhou K & Doudna JA (2010) Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science* 329: 1355–1358

Hemphill J, Borchardt EK, Brown K, Asokan A & Deiters A (2015) Optical control of CRISPR/Cas9 gene editing. *J Am Chem Soc* 137: 5642–5645

Herde ZD, Short AE, Kay VE, Huang BD, Realff MJ & Wilson CJ (2020) Engineering allosteric communication. *Curr Opin Struct Biol* 63: 115–122

Hoffmann MD, Aschenbrenner S, Grosse S, Rapti K, Domenger C, Fakhiri J, Mastel M, Börner K, Eils R, Grimm D, *et al* (2019) Cell-specific CRISPR–Cas9 activation by microRNA-dependent expression of anti-CRISPR proteins. *Nucleic Acids Res* 47: e75–e75

Hoffmann MD, Bubeck F, Eils R & Niopek D (2018) Controlling Cells with Light and LOV. *Adv Biosyst* 2: 1800098

Hoffmann MD, Mathony J, Upmeier Zu Belzen J, Harteveld Z, Aschenbrenner S, Stengl C, Grimm D, Correia BE, Eils R & Niopek D (2021) Optogenetic control of Neisseria meningitidis Cas9 genome editing using an engineered, light-switchable anti-CRISPR protein. *Nucleic Acids Res* 49: 1–11

Hongdusit A, Zwart PH, Sankaran B & Fox JM (2020) Minimally disruptive optical control of protein tyrosine phosphatase 1B. *Nat Commun* 11: 1–44

Hou Z, Zhang Y, Propson NE, Howden SE, Chu L-F, Sontheimer EJ & Thomson JA (2013) Efficient genome engineering in human pluripotent stem cells using Cas9 from Neisseria meningitidis. *Proc Natl Acad Sci* 110: 15644–15649

Hsia Y, Bale JB, Gonen S, Shi D, Sheffler W, Fong KK, Nattermann U, Xu C, Huang PS, Ravichandran R, *et al* (2016) Design of a hyperstable 60-subunit protein icosahedron. *Nature* 535: 136–139

Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, Li Y, Fine EJ, Wu X, Shalem O, *et al* (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol* 31: 827–832

Huang PS, Ban YEA, Richter F, Andre I, Vernon R, Schief WR & Baker D (2011) Rosettaremodel: A generalized framework for flexible backbone protein design. *PLoS ONE* 6

Hunter JD (2007) Matplotlib: A 2D Graphics Environment. *Comput Sci Eng* 9: 90–95

Ibraheim R, Song CQ, Mir A, Amrani N, Xue W & Sontheimer EJ (2018) All-in-one adeno-associated virus delivery and genome editing by Neisseria meningitidis Cas9 in vivo. *Genome Biol* 19: 1–11

Ibraheim R, Tai PWL, Mir A, Javeed N, Wang J, Rodríguez TC, Namkung S, Nelson S, Khokhar ES, Mintzer E, *et al* (2021) Self-inactivating, all-in-one AAV vectors for precision Cas9 genome editing via homology-directed repair in vivo. *Nat Commun* 12: 6267

Ingraham J, Garg VK, Barzilay R & Jaakkola T (2019) Generative models for graph-based protein design. *Deep Gener Models Highly Struct Data DGSICLR 2019 Workshop*: 1–10

Ishino Y, Shinagawa H, Makino K, Amemura M & Nakata A (1987) Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. *J Bacteriol* 169: 5429–5433

Ivankov DN & Finkelstein AV (2020) Solution of Levinthal's Paradox and a Physical Theory of Protein Folding Times. *Biomolecules* 10: 250

Jain PK, Ramanan V, Schepers AG, Dalvie NS, Panda A, Fleming HE & Bhatia SN (2016) Development of Light-Activated CRISPR Using Guide RNAs with Photocleavable Protectors. *Angew Chem - Int Ed* 55: 12440–12444

Jang SK, Kräusslich HG, Nicklin MJ, Duke GM, Palmenberg AC & Wimmer E (1988) A segment of the 5′ nontranslated region of encephalomyocarditis virus RNA directs internal entry of ribosomes during in vitro translation. *J Virol* 62: 2636–2643

Jansen Ruud, Embden JanDA van, Gaastra Wim & Schouls LeoM (2002) Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* 43: 1565–1575

Jasin M & Rothstein R (2013) Repair of Strand Breaks by Homologous Recombination TL - 5. *Cold Spring Harb Perspect Biol* 5 VN-re

Jayanthi VSPKSA, Das AB & Saxena U (2017) Recent advances in biosensor development for the detection of cancer biomarkers. *Biosens Bioelectron* 91: 15–23

Jayaraman P, Devarajan K, Chua TK, Zhang H, Gunawan E & Poh CL (2016) Blue light-mediated transcriptional activation and repression of gene expression in bacteria. *Nucleic Acids Res* 44: 6994–7005

Jayaraman P, Yeoh JW, Zhang J & Poh CL (2018) Programming the Dynamic Control of Bacterial Gene Expression with a Chimeric Ligand- and Light-Based Promoter System. *ACS Synth Biol* 7: 2627–2639

Jendrusch M, Korbel JO, Sadiq SK (2021) AlphaDesign: A de novo protein design framework based on AlphaFold. Biorxiv, https://doi.org/10.1101/2021.10.11.463937

Jin J, Xie X, Chen C, Park JG, Stark C, James DA, Olhovsky M, Linding R, Mao Y & Pawson T (2009) Eukaryotic Protein Domains as Functional Units of Cellular Evolution. *Sci Signal* 2

Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA & Charpentier E (2012) A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* 337: 816–822

Jo R, Bauer MS, Schendel LC, Kluger C & Gaub HE (2019) Dronpa: A Light-Switchable Fluorescent Protein for Opto- Biomechanics.

Johnston RK, Seamon KJ, Saada EA, Podlevsky JD, Branda SS, Timlin JA & Harper JC (2019) Use of anti-CRISPR protein AcrIIA4 as a capture ligand for CRISPR/Cas9 detection. *Biosens Bioelectron* 141: 111361

Jones SK, Hawkins JA, Johnson NV, Jung C, Hu K, Rybarski JR, Chen JS, Doudna JA, Press WH & Finkelstein IJ (2021) Massively parallel kinetic profiling of natural and engineered CRISPR nucleases. *Nat Biotechnol* 39: 84–93

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, *et al* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* 596: 583–589

Kan SBJ, Lewis RD, Chen K & Arnold FH (2016) Directed evolution of cytochrome c for carbon-silicon bond formation: Bringing silicon to life. *Science* 354: 1048–1051

Kapust RB, Tözsér J, Fox JD, Anderson DE, Cherry S, Copeland TD & Waugh DS (2001) Tobacco etch virus protease: mechanism of autolysis and rational design of stable mutants with wild-type catalytic proficiency. *Protein Eng Des Sel* 14: 993–1000

Karginov AV, Ding F, Kota P, Dokholyan NV & Hahn KM (2010) Engineered allosteric activation of kinases in living cells. *Nat Biotechnol* 28: 743–747

Kawashima S & Kanehisa M (2000) AAindex: Amino Acid index database. *Nucleic Acids Res* 28: 374

Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T & Kanehisa M (2008) AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res* 36: D202–D205

Kennedy MB (1995) Origin of PDZ (DHR, GLGF) domains. *Trends Biochem Sci* 20: 350

Khlebnikov A, Skaug T & Keasling JD (2002) Modulation of gene expression from the arabinose-inducible araBAD promoter. *J Ind Microbiol Biotechnol* 29: 34–37

Kim CK, Adhikari A & Deisseroth K (2017) Integration of optogenetics with complementary methodologies in systems neuroscience. *Nat Rev Neurosci* 18: 222–235

Kim Y, Lee SJ, Yoon H, Kim N, Lee B & Suh J (2019) Anti-CRISPR AcrIIC3 discriminates between Cas9 orthologs via targeting the variable surface of the HNH nuclease domain. *FEBS J*: 1–14

Kim YG, Cha J & Chandrasegaran S (1996) Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. *Proc Natl Acad Sci U S A* 93: 1156–1160

Kimura M (1968) Evolutionary Rate at the Molecular Level. *Nature* 217: 624–626

Kleinjan DA, Wardrope C, Nga Sou S & Rosser SJ (2017) Drug-tunable multidimensional synthetic gene control using inducible degron-tagged dCas9 effectors. *Nat Commun* 8: 1–9

Kleinstiver BP, Pattanayak V, Prew MS, Tsai SQ, Nguyen NT, Zheng Z & Joung JK (2016) High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* 529: 490–495

Kleinstiver BP, Prew MS, Tsai SQ, Topkar VV, Nguyen NT, Zheng Z, Gonzales APW, Li Z, Peterson RT, Yeh JRJ, *et al* (2015) Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* 523: 481–485

Klöcker N, Weissenboeck FP, van Dülmen M, Špaček P, Hüwel S & Rentmeister A (2022) Photocaged 5′ cap analogues for optical control of mRNA translation in cells. *Nat Chem* 14: 905–913

Komera I, Gao C, Guo L, Hu G, Chen X & Liu L (2022) Bifunctional optogenetic switch for improving shikimic acid production in E. coli. *Biotechnol Biofuels Bioprod* 15: 13

Komor AC, Kim YB, Packer MS, Zuris JA & Liu DR (2016) Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* 533: 420–424

Konermann S, Brigham MD, Trevino AE, Joung J, Abudayyeh OO, Barcena C, Hsu PD, Habib N, Gootenberg JS, Nishimasu H, *et al* (2015) Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature* 517: 583–588

Kosicki M, Tomberg K & Bradley A (2018) Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements. *Nat Biotechnol* 36: 765–771

Kothe G, Lukaschek M, Link G, Kacprzak S, Illarionov B, Fischer M, Eisenreich W, Bacher A & Weber S (2014) Detecting a new source for photochemically induced dynamic nuclear polarization in the lov2 domain of phototropin by magnetic-field dependent 13C NMR spectroscopy. *J Phys Chem B* 118: 11622–11632

Kryshtafovych A, Schwede T, Topf M, Fidelis K & Moult J (2021) Critical assessment of methods of protein structure prediction (CASP)-Round XIV. *Proteins* 89: 1607–1617

Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL & Baker D (2003) Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Sci New Ser* 302: 1364–1368

Kundert K, Lucas JE, Watters KE, Fellmann C, Ng AH, Heineike BM, Fitzsimmons CM, Oakes BL, Qu J, Prasad N, *et al* (2019) Controlling CRISPR-Cas9 with ligand-activated and ligand-deactivated sgRNAs. *Nat Commun* 10: 2127

Kwon DY, Zhao Y-T, Lamonica JM & Zhou Z (2017) Locus-specific histone deacetylation using a synthetic CRISPR-Cas9-based HDAC. *Nat Commun* 8: 15315

Kwon HJ, Bennik MHJ, Demple B & Ellenberger T (2000) Crystal structure of the Escherichia coli Rob transcription factor in complex with DNA. *Nat Struct Biol* 7: 424–430

Ladant D, Glaser P & Ullmann A (1992) Insertional mutagenesis of Bordetella pertussis adenylate cyclase. *J Biol Chem* 267: 2244–2250

Lagos-Quintana M, Rauhut R, Yalcin A, Meyer J, Lendeckel W & Tuschl T (2002) Identification of tissue-specific microRNAs from mouse. *Curr Biol CB* 12: 735–739

Lalwani MA, Ip SS, Carrasco-López C, Day C, Zhao EM, Kawabe H & Avalos JL (2021) Optogenetic control of the lac operon for bacterial chemical and protein production. *Nat Chem Biol* 17: 71–79

Landgraf P, Rusu M, Sheridan R, Sewer A, Iovino N, Aravin A, Pfeffer S, Rice A, Kamphorst AO, Landthaler M, *et al* (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* 129: 1401–1414

Langan RA, Boyken SE, Ng AH, Samson JA, Dods G, Westbrook AM, Nguyen TH, Lajoie MJ, Chen Z, Berger S, *et al* (2019) De novo design of bioactive protein switches. *Nature*

Lee CM, Cradick TJ & Bao G (2016) The neisseria meningitidis CRISPR-Cas9 system enables specific genome editing in mammalian cells. *Mol Ther* 24: 645–654

Lee J, Jayaram M & Grainge I (1999) Wild-type Flp recombinase cleaves DNA in trans. *EMBO J* 18: 784–791

Lee J, Mir A, Edraki A, Garcia B, Amrani N, Lou HE, Gainetdinov I, Pawluk A, Ibraheim R, Gao XD, *et al* (2018) Potent Cas9 Inhibition in Bacterial and Human Cells by AcrIIC4 and AcrIIC5 Anti-CRISPR Proteins. *mBio* 9: 1–17

Lee J, Mou H, Ibraheim R, Liang SQ, Liu P, Xue W & Sontheimer EJ (2019) Tissue-restricted genome editing in vivo specified by microRNA-repressible anti-CRISPR proteins. *Rna* 25: 1421–1431

Lee J, Natarajan M, Nashine VC, Socolich M, Vo T, Russ WP, Benkovic SJ & Ranganathan R (2008) Surface Sites for Engineering Allosteric Control in Proteins. *Science* 322: 438–442

Lee MJ & Yaffe MB (2016) Protein Regulation in Signal Transduction. *Cold Spring Harb Perspect Biol* 8: a005918

Lee SK, Chou HH, Pfleger BF, Newman JD, Yoshikuni Y & Keasling JD (2007) Directed Evolution of AraC for Improved Compatibility of Arabinose- and Lactose-Inducible Promoters. *Appl Environ Microbiol* 73: 5711–5715

Leman JK, Weitzner BD, Lewis SM, Adolf-Bryfogle J, Alam N, Alford RF, Aprahamian M, Baker D, Barlow KA, Barth P, *et al* (2020) Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat Methods* 17: 665–680

Levinthal C (1969) How to Fold Graciously. *Mossbauer Spectrosc Biol Syst*: 22–24

Levskaya A, Chevalier AA, Tabor JJ, Simpson ZB, Lavery LA, Levy M, Davidson EA, Scouras A, Ellington AD, Marcotte EM, *et al* (2005) Engineering Escherichia coli to see light. *Nature* 438: 441–442

Li J, Xu Z, Chupalov A & Marchisio MA (2018) Anti-CRISPR-based biosensors in the yeast S. cerevisiae. *J Biol Eng* 12: 1–14

Liang M, Sui T, Liu Z, Chen M, Liu H, Shan H, Lai L & Li Z (2020) AcrIIA5 Suppresses Base Editors and Reduces Their Off-Target Effects. *Cells* 9: 1786

Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ & Russell RB (2003) Protein Disorder Prediction. *Structure* 11: 1453–1459

Liu XS, Wu H, Ji X, Stelzer Y, Wu X, Czauderna S, Shu J, Dadon D, Young RA & Jaenisch R (2016a) Editing DNA Methylation in the Mammalian Genome. *Cell* 167: 233-247.e17

Liu Y, Zhan Y, Chen Z, He A, Li J, Wu H, Liu L, Zhuang C, Lin J, Guo X, *et al* (2016b) Directing cellular information flow via CRISPR signal conductors. *Nat Methods* 13: 938–944

Liu Y, Zou RS, He S, Nihongaki Y, Li X, Razavi S, Wu B & Ha T (2020) Very fast CRISPR on demand. *Science* 368: 1265–1269

Lobell RB & Schleif RF (1991) AraC-DNA looping: Orientation and distance-dependent loop breaking by the cyclic AMP receptor protein. *J Mol Biol* 218: 45–54

Lockless SW & Ranganathan R (1999) Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. *Science* 286: 295–299

Louppe G (2015) Understanding Random Forests: From Theory to Practice. doi:10.48550/arXiv.1407.7502 [PREPRINT]

Lovelock SL, Crawshaw R, Basler S, Levy C, Baker D, Hilvert D & Green AP (2022) The road to fully programmable protein catalysis. *Nature* 606: 49–58

Lowe M, Gullotti D, Damjanovic A, Cheng A, Dirla S & Schleif R (2014) Computational and experimental investigation of constitutive behavior in AraC. *Proteins Struct Funct Bioinforma* 82: 3385–3396

Lu S, Huang W & Zhang J (2014) Recent computational advances in the identification of allosteric sites in proteins. *Drug Discov Today* 19: 1595–1600

Luke GA, de Felipe P, Lukashev A, Kallioinen SE, Bruno EA & Ryan MD (2008) Occurrence, function and evolutionary origins of '2A-like' sequences in virus genomes. *J Gen Virol* 89: 1036–1042

Ma D, Xu Z, Zhang Z, Chen X, Zeng X, Zhang Y, Deng T, Ren M, Sun Z, Jiang R, *et al* (2019) Engineer chimeric Cas9 to expand PAM recognition based on evolutionary information. *Nat Commun* 10: 560

Ma E, Harrington LB, O'Connell MR, Zhou K & Doudna JA (2015a) Single-Stranded DNA Cleavage by Divergent CRISPR-Cas9 Enzymes. *Mol Cell* 60: 398–407

Ma H, Naseri A, Reyes-Gutierrez P, Wolfe SA, Zhang S & Pederson T (2015b) Multicolor CRISPR labeling of chromosomal loci in human cells. *Proc Natl Acad Sci* 112: 3002–3007

Macdonald CB, Nedrud D, Grimes PR, Trinidad D, Fraser JS & Coyote-Maestas W (2022) Deep Insertion, Deletion, and Missense Mutation Libraries for Exploring Protein Variation in Evolution, Disease, and Biology Genomics

Maeder ML, Linder SJ, Cascio VM, Fu Y, Ho QH & Joung JK (2013) CRISPR RNA–guided activation of endogenous human genes. *Nat Methods* 10: 977–979

Makarova KS, Brouns SJJ, Horvath P, Sas DF & Wolf YI (2012) Evolution and classification of the CRISPR-Cas systems Kira. *Nat Rev …* 9: 467–477

Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, Barrangou R, Brouns SJJ, Charpentier E, Haft DH, *et al* (2015) An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Microbiol* 13: 722–736

Malaga F, Mayberry O, Park DJ, Rodgers ME, Toptygin D & Schleif RF (2016) A genetic and physical study of the interdomain linker of E. Coli AraC protein-a trans-subunit communication pathway. *Proteins Struct Funct Bioinforma* 84: 448–460

Mali P, Aach J, Stranges PB, Esvelt KM, Moosburner M, Kosuri S, Yang L & Church GM (2013a) CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat Biotechnol* 31: 833–838

Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE & Church GM (2013b) RNA-guided human genome engineering via Cas9. *Science* 339: 823–826

Mandell DJ, Coutsias EA & Kortemme T (2009) Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat Methods* 6: 551–552

Mansouri M, Strittmatter T & Fussenegger M (2019) Light-Controlled Mammalian Cells and Their Therapeutic Applications in Synthetic Biology. *Adv Sci* 6

Marino ND, Zhang JY, Borges AL, Sousa AA, Leon LM, Rauch BJ, Walton RT, Berry JD, Joung JK, Kleinstiver BP, *et al* (2018) Discovery of widespread type I and type V CRISPR-Cas inhibitors. *Science* 362: 240–242

Marsh JA & Teichmann SA (2010) How do proteins gain new domains? *Genome Biol* 11: 126

Mathony J, Harteveld Z, Schmelas C, Upmeier zu Belzen J, Aschenbrenner S, Sun W, Hoffmann MD, Stengl C, Scheck A, Georgeon S, *et al* (2020a) Computational design of anti-CRISPR proteins with improved inhibition potency. *Nat Chem Biol*

Mathony J, Hoffmann MD & Niopek D (2020b) Optogenetics and CRISPR: A New Relationship Built to Last. *Methods Mol Biol Clifton NJ* 2173: 261–281

Mathony J & Niopek D (2021) Enlightening Allostery: Designing Switchable Proteins by Photoreceptor Fusion. *Adv Biol* 5: 2000181

McCafferty J, Griffiths AD, Winter G & Chiswell DJ (1990) Phage antibodies: filamentous phage displaying antibody variable domains. *Nature* 348: 552–554

McCarty NS & Ledesma-Amaro R (2018) Synthetic Biology Tools to Engineer Microbial Communities for Biotechnology. *Trends Biotechnol* xx

McCormick JW, Russo MA, Thompson S, Blevins A & Reynolds KA (2021) Structurally distributed surface sites tune allosteric regulation. *eLife* 10: e68346

McGinness KE, Baker TA & Sauer RT (2006) Engineering Controllable Protein Degradation. *Mol Cell* 22: 701–707

McKinney W (2010) Data Structures for Statistical Computing in Python. *Proc 9th Python Sci Conf*: 56–61

Meeske AJ, Nakandakari-Higa S & Marraffini LA (2019) Cas13-induced cellular dormancy prevents the rise of CRISPR-resistant bacteriophage. *Nature* 570: 241–245

de Mena L, Rizk P & Rincon-Limas DE (2018) Bringing Light to Transcription: The Optogenetics Repertoire. *Front Genet* 9: 1–12

Ming D & Wall ME (2005) Quantifying allosteric effects in proteins. *Proteins Struct Funct Bioinforma* 59: 697–707

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S & Steinegger M (2022) ColabFold: Making Protein folding accessible to all. *Nat Methods*

Miyazaki K & Arnold FH (1999) Exploring Nonnatural Evolutionary Pathways by Saturation Mutagenesis: Rapid Improvement of Protein Function. *J Mol Evol* 49: 716–720

Möglich A, Ayers RA & Moffat K (2009) Design and Signaling Mechanism of Light-Regulated Histidine Kinases. *J Mol Biol* 385: 1433–1444

Möglich A, Ayers RA & Moffat K (2010) Addition at the Molecular Level : Signal Integration in Designed Per – ARNT – Sim Receptor Proteins. *J Mol Biol* 400: 477–486

Mojica FJM, Díez-Villaseñor C, García-Martínez J & Almendros C (2009) Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiol Read Engl* 155: 733–740

Mojica FJM, Díez-Villaseñor C, García-Martínez J & Almendros CY 2009 Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* 155: 733–740

Mojica FJM, Díez-Villaseñor C, García-Martínez J & Soria E (2005) Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements. *J Mol Evol* 60: 174–182

Mojica FJM, Díez-Villaseñor C, Soria E & Juez G (2000) Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Mol Microbiol* 36: 244–246

Mojica FJM, Juez G & Rodriguez-Valera F (1993) Transcription at different salinities of Haloferax mediterranei sequences adjacent to partially modified PstI sites. *Mol Microbiol* 9: 613–621

Morgan SA, Nadler DC, Yokoo R & Savage DF (2016) Biofuel metabolic engineering with biosensors. *Curr Opin Chem Biol* 35: 150–158

Morrison MS, Podracky CJ & Liu DR (2020) The developing toolkit of continuous directed evolution. *Nat Chem Biol* 16: 610–619

Moscou MJ & Bogdanove AJ (2009) A simple cipher governs DNA recognition by TAL effectors. *Science* 326: 1501

Motlagh HN, Wrabl JO, Li J & Hilser VJ (2014) The ensemble nature of allostery. *Nature* 508: 331–339

Moult J, Pedersen JT, Judson R & Fidelis K (1995) A large-scale experiment to assess protein structure prediction methods. *Proteins Struct Funct Bioinforma* 23: ii–iv

Multamäki E, García de Fuentes A, Sieryi O, Bykov A, Gerken U, Ranzani AT, Köhler J, Meglinski I, Möglich A & Takala H (2022) Optogenetic Control of Bacterial Expression by Red Light. *ACS Synth Biol* 11: 3354–3367

Nadler DC, Morgan SA, Flamholz A, Kortright KE & Savage DF (2016) Rapid construction of metabolite biosensors using domain-insertion profiling. *Nat Commun* 7: 1–11

Nagai T, Sawano A, Eun Sun Park & Miyawaki A (2001) Circularly permuted green fluorescent proteins engineered to sense Ca2+. *Proc Natl Acad Sci U S A* 98: 3197–3202

Nagel G, Szellas T, Huhn W, Kateriya S, Adeishvili N, Berthold P, Ollig D, Hegemann P & Bamberg E (2003) Channelrhodopsin-2, a directly light-gated cation-selective membrane channel. *Proc Natl Acad Sci U S A* 100: 13940–13945

Nakai J, Ohkura M & Imoto K (2001) A high signal-to-noise ca2+ probe composed of a single green fluorescent protein. *Nat Biotechnol* 19: 137–141

Nakamura M, Chen L, Howes SC, Schindler TD, Nogales E & Bryant Z (2014) Remote control of myosin and kinesin motors using light-activated gearshifting. *Nat Nanotechnol* 9: 693–697

Nakamura M, Srinivasan P, Chavez M, Carter MA, Dominguez AA, La Russa M, Lau MB, Abbott TR, Xu X, Zhao D, *et al* (2019) Anti-CRISPR-mediated control of gene editing and synthetic circuits in eukaryotic cells. *Nat Commun* 10: 1–11

Nihongaki Y, Kawano F, Nakajima T & Sato M (2015a) Photoactivatable CRISPR-Cas9 for optogenetic genome editing. *Nat Biotechnol* 33: 755–760

Nihongaki Y, Otabe T, Ueda Y & Sato M (2019) A split CRISPR–Cpf1 platform for inducible genome editing and gene activation. *Nat Chem Biol* 15: 882–888

Nihongaki Y, Yamamoto S, Kawano F, Suzuki H & Sato M (2015b) CRISPR-Cas9-based Photoactivatable Transcription System. *Chem Biol* 22: 169–174

Niopek D, Benzinger D, Roensch J, Draebing T, Wehler P, Eils R & Di Ventura B (2014) Engineering light-inducible nuclear localization signals for precise spatiotemporal control of protein dynamics in living cells. *Nat Commun* 5: 1–11

Niopek D, Wehler P, Roensch J, Eils R & Ventura BD (2016) Optogenetic control of nuclear protein export. *Nat Commun*: 1–9

Nuñez JK, Kranzusch PJ, Noeske J, Wright AV, Davies CW & Doudna JA (2014) Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nat Struct Mol Biol* 21: 528–534

Oakes BL, Nadler DC, Flamholz A, Fellmann C, Staahl BT, Doudna JA & Savage DF (2016) Profiling of engineering hotspots identifies an allosteric CRISPR-Cas9 switch. *Nat Biotechnol* 34: 646–651

Ohlendorf R, Vidavski RR, Eldar A, Moffat K & Möglich A (2012) From Dusk till Dawn: One-Plasmid Systems for Light-Regulated Gene Expression. *J Mol Biol* 416: 534–542

Ong NT, Olson EJ & Tabor JJ (2018) Engineering an E. coli Near-Infrared Light Sensor. *ACS Synth Biol* 7: 240–248

Ong NT & Tabor JJ (2018) A Miniaturized Escherichia coli Green Light Sensor with High Dynamic Range. *ChemBioChem* 19: 1255–1258

Ormö M, Cubitt AB, Kallio K, Gross LA, Tsien RY & Remington SJ (1996) Crystal Structure of the Aequorea victoria Green Fluorescent Protein. *Science* 273: 1392–1395

Ortiz-Guerrero JM, Polanco MC, Murillo FJ, Padmanabhan S & Elías-Arnanz M (2011) Light-dependent gene regulation by a coenzyme B12-based photoreceptor. *Proc Natl Acad Sci* 108: 7565–7570

Ostermeier M (2009) Designing switchable enzymes. *Curr Opin Struct Biol* 19: 442–448

Otero-Muras I & Banga JR (2017) Automated Design Framework for Synthetic Biology Exploiting Pareto Optimality. *ACS Synth Biol* 6: 1180–1193

Paget MS (2015) Bacterial Sigma Factors and Anti-Sigma Factors: Structure, Function and Distribution. *Biomolecules* 5: 1245–1265

Pak MA, Markhieva KA, Novikova MS, Petrov DS, Vorobyev IS, Maksimova ES, Kondrashov FA & Ivankov DN (2021) Using AlphaFold to predict the impact of single mutations on protein stability and function. *bioRxiv*: 2021.09.19.460937

Patil AH, Baran A, Brehm ZP, McCall MN & Halushka MK (2022) A curated human cellular microRNAome based on 196 primary cell types. *GigaScience* 11: giac083

Pattanayak V, Lin S, Guilinger JP, Ma E, Doudna JA & Liu DR (2013) High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat Biotechnol* 31: 839–843

Pawluk A, Amrani N, Zhang Y, Garcia B, Hidalgo-Reyes Y, Lee J, Edraki A, Shah M, Sontheimer EJ, Maxwell KL, *et al* (2016) Naturally Occurring Off-Switches for CRISPR-Cas9. *Cell* 167: 1829-1838.e9

Pawluk A, Davidson AR & Maxwell KL (2017) Anti-CRISPR: discovery, mechanism and function. *Nat Rev Microbiol* 16: 12–17

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, *et al* (2011) Scikit-learn: Machine Learning in Python. *J Mach Learn Res* 12: 2825–2830

Peisajovich SG, Garbarino JE, Wei P & Lim WA (2010) Rapid diversification of cell signaling phenotypes by modular domain recombination. *Science* 328: 368–372

Pelletier J & Sonenberg N (1988) Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. *Nature* 334: 320–325

Perez-Pinera P, Kocak DD, Vockley CM, Adler AF, Kabadi AM, Polstein LR, Thakore PI, Glass KA, Ousterout DG, Leong KW, *et al* (2013) RNA-guided gene activation by CRISPR-Cas9-based transcription factors. *Nat Methods* 10: 973–976

Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, Morris JH & Ferrin TE (2021) UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci Publ Protein Soc* 30: 70–82

Philippe C, Morency C, Plante P-L, Zufferey E, Achigar R, Tremblay DM, Rousseau GM, Goulet A & Moineau S (2022) A truncated anti-CRISPR protein prevents spacer acquisition but not interference. *Nat Commun* 13: 2802

Picard HR, Schwingen KS, Green LM, Shis DL, Egan SM, Bennett MR & Swint-Kruse L (2022) Allosteric regulation within the highly interconnected structural scaffold of AraC/XylS homologs tolerates a wide range of amino acid changes. *Proteins Struct Funct Bioinforma* 90: 186–199

Pincus D, Resnekov O & Reynolds KA (2017) An evolution-based strategy for engineering allosteric regulation. *Phys Biol* 14: 025002

Plaper T, Merljak E, Fink T, Lainšček D, Satler T, Jazbec V, Benčina M & Jerala R (2022) Designed allosteric protein logic Synthetic Biology

Pletneva NV, Maksimov EG, Protasova EA, Mamontova AV, Simonyan TR, Ziganshin RH, Lukyanov KA, Muslinkina L, Pletnev S, Bogdanov AM, *et al* (2021) Amino acid residue at the 165th position tunes EYFP chromophore maturation. A structure-based design. *Comput Struct Biotechnol J* 19: 2950–2959

Polstein LR, Gersbach C a, Carolina N, States U, Biology C, Carolina N & Carolina N (2015) A light-inducible CRISPR/Cas9 system for control of endogenous gene activation. *Nat Chem Biol* 11: 198–200

Ponting CP & Russell RR (2002) The Natural History of Protein Domains. *Annu Rev Biophys Biomol Struct* 31: 45–71

Porteus MH & Baltimore D (2003) Chimeric Nucleases Stimulate Gene Targeting in Human Cells. *Science* 300: 763–763

Prabhakaran M (1990) The distribution of physical, chemical and conformational properties in signal and nascent peptides. *Biochem J* 269: 691–696

Pudasaini A, El-Arab KK & Zoltowski BD (2015) LOV-based optogenetic devices: light-driven modules to impart photoregulated control of cellular signaling. *Front Mol Biosci* 2: 1–15

Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP & Lim WA (2013) Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* 152: 1173–1183

Radley TL, Markowska AI, Bettinger BT, Ha JH & Loh SN (2003) Allosteric switching by mutually exclusive folding of protein domains. *J Mol Biol* 332: 529–536

Raffelberg S, Wang L, Gao S, Losi A, Gärtner W & Nagel G (2013) A LOV-domain-mediated blue-light-activated adenylate (adenylyl) cyclase from the cyanobacterium Microcoleus chthonoplastes PCC 7420. *Biochem J* 455: 359–365

Raghavan AR, Salim K & Yadav VG (2020) Optogenetic Control of Heterologous Metabolism in E. coli. *ACS Synth Biol* 9: 2291–2300

Ran FA, Cong L, Yan WX, Scott DA, Gootenberg JS, Kriz AJ, Zetsche B, Shalem O, Wu X, Makarova KS, *et al* (2015) In vivo genome editing using Staphylococcus aureus Cas9. *Nature* 520: 186–91

Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA & Zhang F (2013) Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* 8: 2281–2308

Ranglack J, Weyrich A & Stein V (2020) iFLinkC : an iterative functional linker cloning strategy for the combinatorial assembly and recombination of linker peptides with functional domains Alexander Gr awe. *Nucleic Acids Res* 48: 1–11

Rauch BJ, Silvis MR, Hultquist JF, Waters CS, McGregor MJ, Krogan NJ & Bondy-Denomy J (2017) Inhibition of CRISPR-Cas9 with Bacteriophage Proteins. *Cell* 168: 150-158.e10

Ravikumar A, Arzumanyan GA, Obadi KA, Javanpour AA, Liu Correspondence CC, Obadi MKA & Liu CC (2018) Scalable, Continuous Evolution of Genes at Mutation Rates above Genomic Error Thresholds Resource Scalable, Continuous Evolution of Genes at Mutation Rates above Genomic Error Thresholds. *Cell* 175

Rees HA, Komor AC, Yeh WH, Caetano-Lopes J, Warman M, Edge ASB & Liu DR (2017) Improving the DNA specificity and applicability of base editing through protein engineering and protein delivery. *Nat Commun* 8: 1–10

Renicke C, Schuster D, Usherenko S, Essen LO & Taxis C (2013) A LOV2 domain-based optogenetic tool to control protein degradation and cellular function. *Chem Biol* 20: 619–626

Reynolds KA, McLaughlin RN & Ranganathan R (2011) Hot spots for allosteric regulation on protein surfaces. *Cell* 147: 1564–1575

Richter F, Fonfara I, Bouazza B, Schumacher CH, Bratovič M, Charpentier E & Möglich A (2016) Engineering of temperature- and light-switchable Cas9 variants. *Nucleic Acids Res* 44: 10003–10014

Richter G, Weber S, Römisch W, Bacher A, Fischer M & Eisenreich W (2005) Photochemically induced dynamic nuclear polarization in a C450A mutant of the LOV2 domain of the Avena sativa blue-light receptor phototropin. *J Am Chem Soc* 127: 17245–17252

Rihtar E, Lebar T, Lainšček D, Kores K, Lešnik S, Bren U & Jerala R (2022) Chemically inducible split protein regulators for mammalian cells. *Nat Chem Biol*

Rivera-Cancel G, Motta-Mena LB & Gardner KH (2012) Identification of Natural and Artificial DNA Substrates for Light-Activated LOV–HTH Transcription Factor EL222. *Biochemistry* 51: 10024–10034

Rivoire O, Reynolds KA & Ranganathan R (2016) Evolution-Based Functional Decomposition of Proteins. *PLoS Comput Biol* 12: 1–26

Rodgers ME & Schleif R (2009) Solution structure of the DNA binding domain of AraC protein. *Proteins Struct Funct Bioinforma* 77: 202–208

Rohl CA, Strauss CEM, Chivian D & Baker D (2004a) Modeling structurally variable regions in homologous proteins with rosetta. *Proteins* 55: 656–677

Rohl CA, Strauss CEM, Misura KMS & Baker D (2004b) Protein Structure Prediction Using Rosetta. In *Methods in Enzymology* pp 66–93. Academic Press

Romano E, Baumschlager A, Akmeriç EB, Palanisamy N, Houmani M, Schmidt G, Öztürk MA, Ernst L, Khammash M & Di Ventura B (2021) Engineering AraC to make it responsive to light instead of arabinose. *Nat Chem Biol* 17: 817–827

Romero PA, Krause A & Arnold FH (2013) Navigating the protein fitness landscape with Gaussian processes. *Proc Natl Acad Sci U S A* 110

Rose JC, Stephany JJ, Wei CT, Fowler DM & Maly DJ (2018) Rheostatic Control of Cas9-Mediated DNA Double Strand Break (DSB) Generation and Genome Editing. *ACS Chem Biol* 13: 438–442

Rouet P, Smih F & Jasin M (1994) Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. *Mol Cell Biol* 14: 8096–8106

Rousseau BA, Hou Z, Gramelspacher MJ & Zhang Y (2018) Programmable RNA Cleavage and Recognition by a Natural CRISPR-Cas9 System from Neisseria meningitidis. *Mol Cell* 69: 906-914.e4

Rudin N, Sugarman E & Haber JE (1989) Genetic and physical analysis of double-strand break repair and recombination in Saccharomyces cerevisiae. *Genetics* 122: 519–534

Rueden CT, Schindelin J, Hiner MC, DeZonia BE, Walter AE, Arena ET & Eliceiri KW (2017) ImageJ2: ImageJ for the next generation of scientific image data. *BMC Bioinformatics* 18: 529

Ryu M-H & Gomelsky M (2014) Near-infrared Light Responsive Synthetic c-di-GMP Module for Optogenetic Applications. *ACS Synth Biol* 3: 802–810

Ryu MH, Kang IH, Nelson MD, Jensen TM, Lyuksyutova AI, Siltberg-Liberles J, Raizen DM & Gomelsky M (2014) Engineering adenylate cyclases regulated by near-infrared window light. *Proc Natl Acad Sci U S A* 111: 10167–10172

Ryu M-H, Moskvin OV, Siltberg-Liberles J & Gomelsky M (2010) Natural and Engineered Photoactivated Nucleotidyl Cyclases for Optogenetic Applications*. *J Biol Chem* 285: 41501–41508

Saldaño T, Escobedo N, Marchetti J, Zea DJ, Mac Donagh J, Velez Rueda AJ, Gonik E, García Melani A, Novomisky Nechcoff J, Salas MN, *et al* (2022) Impact of protein conformational diversity on AlphaFold predictions. *Bioinformatics* 38: 2742–2748

Salinas VH & Ranganathan R (2018) Coevolution-based inference of amino acid interactions underlying protein function. *eLife* 7: 1–20

Sallee NA, Yeh BJ & Lim WA (2007) Engineering Modular Protein Interaction Switches by Sequence Overlap. *J Am Chem Soc* 129: 4606–4611

Salomon M, Eisenreich W, Dürr H, Schleicher E, Knieb E, Massey V, Rüdiger W, Müller F, Bacher A & Richter G (2001) An optomechanical transducer in the blue light receptor phototropin from Avena sativa. *Proc Natl Acad Sci U S A* 98: 12357–12361

Saviola B, Seabold R & Schleif RF (1998) Arm-domain interactions in AraC. *J Mol Biol* 278: 539–548

Schleif R (2010) AraC protein, regulation of the l-arabinose operon in Escherichia coli , and the light switch mechanism of AraC action. *FEMS Microbiol Rev* 34: 779–796

Schmelas C & Grimm D (2018) Split Cas9, Not Hairs – Advancing the Therapeutic Index of CRISPR Technology. *Biotechnol J* 13: 1–12

Schmid-Burgk JL, Gao L, Li D, Gardner Z, Strecker J, Lash B & Zhang F (2020) Highly Parallel Profiling of Cas9 Variant Specificity. *Mol Cell* 78: 794-800.e8

Schmidl SR, Sheth RU, Wu A & Tabor JJ (2014) Refactoring and Optimization of Light-Switchable Escherichia coli Two-Component Systems. *ACS Synth Biol* 3: 820–831

Schmidt F & Grimm D (2015) CRISPR genome engineering and viral gene delivery: A case of mutual attraction. *Biotechnol J* 10: 258–272

Schneider CA, Rasband WS & Eliceiri KW (2012) NIH Image to ImageJ: 25 years of image analysis. *Nat Methods* 9: 671–675

Schueler-Furman O & Wodak SJ (2016) Computational approaches to investigating allostery. *Curr Opin Struct Biol* 41: 159–171

Scully R, Panday A, Elango R & Willis NA (2019) DNA double-strand break repair-pathway choice in somatic mammalian cells. *Nat Rev Mol Cell Biol* 20: 698–714

Seedorff J & Schleif R (2011) Active Role of the Interdomain Linker of AraC. *J Bacteriol* 193: 5737–5746

Seifert S, Ehrt C, Lückfeldt L, Lubeck M, Schramm F & Brakmann S (2019) Optical Control of Transcription: Genetically Encoded Photoswitchable Variants of T7 RNA Polymerase. *ChemBioChem* 20: 2813–2817

Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AWR, Bridgland A, *et al* (2020) Improved protein structure prediction using potentials from deep learning. *Nature* 577: 706–710

Senís E, Fatouros C, Große S, Wiedtke E, Niopek D, Mueller AK, Börner K & Grimm D (2014) CRISPR/Cas9-mediated genome engineering: An adeno-associated viral (AAV) vector toolbox. *Biotechnol J* 9: 1402–1412

Sentmanat MF, Peters ST, Florian CP, Connelly JP & Pruett-Miller SM (2018) A Survey of Validation Strategies for CRISPR-Cas9 Editing. *Sci Rep* 8: 1–8

Senturk S, Shirole NH, Nowak DG, Corbo V, Pal D, Vaughan A, Tuveson DA, Trotman LC, Kinney JB & Sordella R (2017) Rapid and tunable method to temporally control gene editing based on conditional Cas9 stabilization. *Nat Commun* 8: 1–10

Sevy AM, Wu NC, Gilchuk IM, Parrish EH, Burger S, Yousif D, Nagel MBM, Schey KL, Wilson IA, Crowe JE, *et al* (2019) Multistate design of influenza antibodies improves affinity and breadth against seasonal viruses. *Proc Natl Acad Sci U S A* 116: 1597–1602

Shaaya M, Fauser J, Zhurikhina A, Conage-Pough JE, Huyot V, Brennan M, Flower CT, Matsche J, Khan S, Natarajan V, *et al* (2020) Light-regulated allosteric switch enables temporal and subcellular control of enzyme activity. *eLife* 9: e60647

Shams A, Higgins SA, Fellmann C, Laughlin TG, Oakes BL, Lew R, Kim S, Lukarska M, Arnold M, Staahl BT, *et al* (2021) Comprehensive deletion landscape of CRISPR-Cas9 identifies minimal RNA-guided DNA-binding modules. *Nat Commun* 12: 5664

Sheets MB & Dunlop MJ (2022) An Optogenetic Toolkit for Light-Inducible Antibiotic Resistance Synthetic Biology

Sheets MB, Wong WW & Dunlop MJ (2020) Light-Inducible Recombinases for Bacterial Optogenetics. *ACS Synth Biol* 9: 227–235

Shekhawat SS & Ghosh I (2011) Split-protein systems: Beyond binary protein-protein interactions. *Curr Opin Chem Biol* 15: 790–797

Shiau AK, Barstad D, Loria PM, Cheng L, Kushner PJ, Agard DA & Greene GL (1998) The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell* 95: 927–937

Shin J, Jiang F, Liu JJ, Bray NL, Rauch BJ, Baik SH, Nogales E, Bondy-Denomy J, Corn JE & Doudna JA (2017) Disabling Cas9 by an anti-CRISPR DNA mimic. *Sci Adv* 3: 1–10

Shortle D & Sondek J (1995) The emerging role of insertions and deletions in protein engineering. *Curr Opin Biotechnol* 6: 387–393

Shu X, Royant A, Lin MZ, Aguilera TA, Lev-Ram V, Steinbach PA & Tsien RY (2009) Mammalian expression of infrared fluorescent proteins engineered from a bacterial phytochrome. *Science* 324: 804–807

Shui S, Gainza P, Scheller L, Yang C, Kurumida Y, Rosset S, Georgeon S, Di Roberto RB, Castellanos-Rueda R, Reddy ST, *et al* (2021) A rational blueprint for the design of chemically-controlled protein switches. *Nat Commun* 12: 5754

Siegel MS & Isacoff EY (1997) A Genetically Encoded Optical Probe of Membrane Voltage. *Neuron* 19: 735–741

Siegele DA & Hu JC (1997) Gene expression from plasmids containing the araBAD promoter at subsaturating inducer concentrations represents mixed populations. *Proc Natl Acad Sci* 94: 8168–8172

Skretas G & Wood DW (2005) A Bacterial Biosensor of Endocrine Modulators. *J Mol Biol* 349: 464–474

Smith GP (1985) Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* 228: 1315–1317

Smithies O, Gregg RG, Boggs SS, Koralewski MA & Kucherlapati RS (1985) Insertion of DNA sequences into the human chromosomal β-globin locus by homologous recombination. *Nature* 317: 230–234

Soisson SM, MacDougall-Shackleton B, Schleif R & Wolberger C (1997) Structural Basis for Ligand-Regulated Oligomerization of AraC. *Science* 276: 421–425

Song G, Zhang F, Zhang X, Song G, Zhang F, Zhang X, Gao X, Zhu X, Fan D & Tian Y (2019) AcrIIA5 Inhibits a Broad Range of Cas9 Orthologs by Preventing DNA Target Cleavage Report AcrIIA5 Inhibits a Broad Range of Cas9 Orthologs by Preventing DNA Target Cleavage. *CellReports* 29: 2579-2589.e4

Srivastava A (2016) In vivo tissue-tropism of adeno-associated viral vectors. *Curr Opin Virol* 21: 75–80

Starzyk RM, Burbaum JJ & Schimmel P (1989) Insertion of new sequences into the catalytic domain of an enzyme. *Biochemistry* 28: 8479–8484

Steinegger M & Söding J (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 35: 1026–1028

Stemmer WP (1994) Rapid evolution of a protein in vitro by DNA shuffling. *Nature* 370: 389–391

Stierl M, Stumpf P, Udwari D, Gueta R, Hagedorn R, Losi A, Gärtner W, Petereit L, Efetova M, Schwarzel M, *et al* (2011) Light Modulation of Cellular cAMP by a Small Bacterial Photoactivated Adenylyl Cyclase, bPAC, of the Soil Bacterium Beggiatoa*♦. *J Biol Chem* 286: 1181–1188

Storici F, Durham CL, Gordenin DA & Resnick MA (2003) Chromosomal site-specific double-strand breaks are efficiently targeted for repair by oligonucleotides in yeast. *Proc Natl Acad Sci* 100: 14994–14999

Storici F, Snipe JR, Chan GK, Gordenin DA & Resnick MA (2006) Conservative Repair of a Chromosomal Double-Strand Break by Single-Strand DNA through Two Steps of Annealing. *Mol Cell Biol* 26: 7645–7657

Stratton MM & Loh SN (2011) Converting a protein into a switch for biosensing and functional regulation. 20: 19–29

Stratton MM, Mitrea DM & Loh SN (2008) A Ca2+-sensing molecular switch based on alternate frame protein folding. *ACS Chem Biol* 3: 723–732

Stricker J, Cookson S, Bennett MR, Mather WH, Tsimring LS & Hasty J (2008) A fast, robust and tunable synthetic gene oscillator. *Nature* 456: 516–519

Strickland D, Lin Y, Wagner E, Hope CM, Zayner J, Antoniou C, Sosnick TR, Weiss EL & Glotzer M (2012) TULIPs: Tunable, light-controlled interacting protein tags for cell biology. *Nat Methods* 9: 379–384

Strickland D, Moffat K & Sosnick TR (2008) Light-activated DNA binding in a designed allosteric protein. *Proc Natl Acad Sci* 105: 10709–10714

Strickland D, Yao X, Gawlak G, Rosen MK, Gardner KH & Sosnick TR (2010) Rationally improving LOV domain–based photoswitches. *Nat Methods* 7: 623–626

Stüven B, Stabel R, Ohlendorf R, Beck J, Schubert R & Möglich A (2019) Characterization and engineering of photoactivated adenylyl cyclases. *Biol Chem* 400: 429–441

Su JG, Qi LS, Li CH, Zhu YY, Du HJ, Hou YX, Hao R & Wang JH (2014) Prediction of allosteric sites on protein surfaces with an elastic-network-model-based thermodynamic method. *Phys Rev E Stat Nonlin Soft Matter Phys* 90: 022719

Subramaniam S & Kleywegt GJ (2022) A paradigm shift in structural biology. *Nat Methods* 19: 20–23

Süel GM, Lockless SW, Wall MA & Ranganathan R (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat Struct Biol* 10: 59–69

Sun P, Austin BP, Tözsér J & Waugh DS (2010) Structural determinants of tobacco vein mottling virus protease substrate specificity: Structure of TVMV Protease/Substrate Complex. *Protein Sci* 19: 2240–2251

Sun W, Yang J, Cheng Z, Lou J, Sontheimer EJ & States A (2019) Structures of Neisseria meningitidis Cas9 Complexes in Catalytically Poised and Anti-CRISPR- Inhibited States Article Structures of Neisseria meningitidis Cas9 Complexes in Catalytically Poised and. *Mol Cell*: 1–15

Suyama M & Ohara O (2003) DomCut: prediction of inter-domain linker regions in amino acid sequences. *Bioinformatics* 19: 673–674

Sverrisson F, Correia BE, Feydy J & Bronstein MM (2020) Fast end-to-end learning on protein surfaces. *bioRxiv*

Swartz TE, Corchnoy SB, Christie JM, Lewis JW, Szundi I, Briggs WR & Bogomolni RA (2001) The Photocycle of a Flavin-binding Domain of the Blue Light Photoreceptor Phototropin *. *J Biol Chem* 276: 36493–36500

Tabor JJ, Levskaya A & Voigt CA (2011) Multichromatic Control of Gene Expression in Escherichia coli. *J Mol Biol* 405: 315–324

Takano K & Yutani K A new scale for side-chain contribution to protein stability based on the empirical stability analysis of mutant proteins. 4

Tallini YM, Ohkura M, Choi BR, Ji G, Imoto K, Doran R, Lee J, Plan P, Wilson J, Xin HB, *et al* (2006) Imaging cellular signals in the heart in vivo: Cardiac expression of the high-signal Ca2+ indicator GCaMP2. *Proc Natl Acad Sci U S A* 103: 4753–4758

Tanenbaum DM, Wang Y, Williams SP & Sigler PB (1998) Crystallographic comparison of the estrogen and progesterone receptor's ligand binding domains. *Proc Natl Acad Sci U S A* 95: 5998–6003

Taneoka A, Sakaguchi-Mikami A, Yamazaki T, Tsugawa W & Sode K (2009) The construction of a glucose-sensing luciferase. *Biosens Bioelectron* 25: 76–81

Tang S-Y & Cirino PC (2010) Elucidating residue roles in engineered variants of AraC regulatory protein. *Protein Sci* 19: 291–298

Tang S-Y, Fazelinia H & Cirino PC (2008) AraC Regulatory Protein Mutants with Altered Effector Specificity. *J Am Chem Soc* 130: 5267–5271

Tang W, Hu JH & Liu DR (2017) Aptazyme-embedded guide RNAs enable ligand-responsive genome editing and transcriptional activation. *Nat Commun* 8: 15939

Teasley Hamorsky K, Ensor CM, Wei Y & Daunert S (2008) A Bioluminescent Molecular Switch For Glucose. *Angew Chem Int Ed* 47: 3718–3721

Teşileanu T, Colwell LJ & Leibler S (2015) Protein Sectors: Statistical Coupling Analysis versus Conservation. *PLoS Comput Biol* 11: 1–20

Thavalingam A, Cheng Z, Garcia B, Huang X, Shah M, Sun W, Wang M, Harrington L, Hwang S, Hidalgo-Reyes Y, *et al* (2019) Inhibition of CRISPR-Cas9 ribonucleoprotein complex assembly by anti-CRISPR AcrIIC2. *Nat Commun* 10: 2806

The UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 49: D480–D489

Thomas KR, Folger KR & Capecchi MR (1986) High frequency targeting of genes to specific sites in the mammalian genome. *Cell* 44: 419–428

Tiessen A, Pérez-Rodríguez P & Delaye-Arredondo LJ (2012) Mathematical modeling and comparison of protein size distribution in different plant, animal, fungal and microbial species reveals a negative correlation between protein size and protein number, thus providing insight into the evolution of proteomes. *BMC Res Notes* 5: 85

Toettcher JE, Weiner OD & Lim WA (2013) Using optogenetics to interrogate the dynamic control of signal transmission by the Ras/Erk module. *Cell* 155: 1422–1434

Tong AB, Burch JD, McKay D, Bustamante C, Crackower MA & Wu H (2021) Could AlphaFold revolutionize chemical therapeutics? *Nat Struct Mol Biol* 28: 771–772

Trott O & Olson AJ (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *J Comput Chem* 31: 455–461

Tubert-Brohman I, Sherman W, Repasky M & Beuming T (2013) Improved docking of polypeptides with Glide. *J Chem Inf Model* 53: 1689–1699

Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, Bridgland A, Cowie A, Meyer C, Laydon A, *et al* (2021) Highly accurate protein structure prediction for the human proteome. *Nature* 596: 590–596

Tzeng S & Kalodimos CG (2013) Allosteric inhibition through suppression of transient conformational states. *Nat Chem Biol* 9: 462–465

Tzeng SR & Kalodimos CG (2012) Protein activity regulation by conformational entropy. *Nature* 488: 236–240

Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE & Berendsen HJC (2005) GROMACS: fast, flexible, and free. *J Comput Chem* 26: 1701–1718

Van Rossum G, Van Rossum D & L. F (2009) Python 3 Reference Manual Scotts Valley, CA: CreateSpace

Vaswani A, Shazeer N, Parmar N, Uzkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention Is All You Need. Arxiv, arXiv:1706.03762.

Vig J, Madani A, Varshney LR, Xiong C, Socher R & Rajani NF (2020) BERTology Meets Biology: Interpreting Attention in Protein Language Models. Arxiv, arXiv:2006.15222

Vishweshwaraiah YL, Chen J, Chirasani VR, Tabdanov ED & Dokholyan NV (2021) Two-input protein logic gate for computation in living cells. *Nat Commun* 12: 6615

Vogel C, Bashton M, Kerrison ND, Chothia C & Teichmann SA (2004) Structure, function and evolution of multidomain proteins. *Curr Opin Struct Biol* 14: 208–216

Vojta A, Dobrinić P, Tadić V, Bočkor L, Korać P, Julg B, Klasić M & Zoldoš V (2016) Repurposing the CRISPR-Cas9 system for targeted DNA methylation. *Nucleic Acids Res* 44: 5615–5628

Voskarides K (2021) Directed Evolution. The Legacy of a Nobel Prize. *J Mol Evol* 89: 189–191

Wang H, Vilela M, Winkler A, Tarnawski M, Schlichting I, Yumerefendi H, Kuhlman B, Liu R, Danuser G & Hahn KM (2016) LOVTRAP: An optogenetic system for photoinduced protein dissociation. *Nat Methods* 13: 755–758

Wang Q, Shui B, Kotlikoff MI & Sondermann H (2008) Structural Basis for Calcium Sensing by GCaMP2. *Structure* 16: 1817–1827

Wang R, Preamplume G, Terns MP, Terns RM & Li H (2011) Interaction of the Cas6 riboendonuclease with CRISPR RNAs: recognition and cleavage. *Struct Lond Engl 1993* 19: 257–264

Wang X, Chen X & Yang Y (2012) Spatiotemporal control of gene expression by a light-switchable transgene system. *Nat Methods* 9: 266–269

Wärnmark A, Treuter E, Gustafsson J-A, Hubbard RE, Brzozowski AM & Pike ACW (2002) Interaction of transcriptional intermediary factor 2 nuclear receptor box peptides with the coactivator binding site of estrogen receptor alpha. *J Biol Chem* 277: 21862–21868

Waskom ML (2021) seaborn: statistical data visualization. *J Open Source Softw* 6: 3021

Weber G (1972) Ligand binding and internal equilibiums in proteins. *Biochemistry* 11: 864–878

Weinberg BH, Cho JH, Agarwal Y, Pham NTH, Caraballo LD, Walkosz M, Ortega C, Trexler M, Tague N, Law B, *et al* (2019) High-performance chemical- and light-inducible recombinases in mammalian cells and mice. *Nat Commun* 10: 4845

Weterings E & Chen DJ (2008) The endless tale of non-homologous end-joining. *Cell Res* 18: 114–124

Willow Coyote-Maestas, David Nedrud, Yungui He DS (2022) Determinants of trafficking, conduction, and disease within a K+ channel revealed through multiparametric deep mutational scanning.

Wilson CJ, Choy W-Y & Karttunen M (2022) AlphaFold2: A Role for Disordered Protein/Region Prediction? *Int J Mol Sci* 23: 4591

Wilson MZ, Ravindran PT, Lim WA & Toettcher JE (2017) Tracing Information Flow from Erk to Target Gene Induction Reveals Mechanisms of Dynamic and Combinatorial Control. *Mol Cell* 67: 757-769.e5

Wong S, Mosabbir AA & Truong K (2015) An engineered split intein for photoactivated protein trans-splicing. *PLoS ONE* 10: 1–16

Wright CM, Majumdar A, Tolman JR & Ostermeier M (2010) NMR characterization of an engineered domain fusion between maltose binding protein and TEM1 beta-lactamase provides insight into its structure and allosteric mechanism. *Proteins* 78: 1423–1430

Wu M & Schleif R (2001a) Strengthened Arm-Dimerization Domain Interactions in AraC*. *J Biol Chem* 276: 2562–2564

Wu M & Schleif R (2001b) Mapping arm-DNA-binding domain interactions in AraC. *J Mol Biol* 307: 1001–1009

Wu YI, Frey D, Lungu OI, Jaehrig A, Schlichting I, Kuhlman B & Hahn KM (2009) A genetically encoded photoactivatable Rac controls the motility of living cells. *Nature* 461: 104–108

Wu Z, Yang H & Colosi P (2010) Effect of Genome Size on AAV Vector Packaging. *Mol Ther* 18: 80–86

Xu X, Tao Y, Gao X, Zhang L, Li X, Zou W, Ruan K, Wang F, Xu G & Hu R (2016) A CRISPR-based approach for targeted DNA demethylation. *Cell Discov* 2: 1–12

Yan J, Wen W, Xu W, Long J, Adams ME, Froehner SC & Zhang M (2005) Structure of the split PH domain and distinct lipid-binding properties of the PH–PDZ supramodule of α-syntrophin. *EMBO J* 24: 3985–3995

Yang H & Patel DJ (2017) Inhibition Mechanism of an Anti-CRISPR Suppressor AcrIIA4 Targeting SpyCas9. *Mol Cell* 67: 117-127.e5

Yang KK, Wu Z & Arnold FH (2019) Machine-learning-guided directed evolution for protein engineering. *Nat Methods* 16: 687–694

Yang T-T, Cheng L & Kain SR (1996) Optimized Codon Usage and Chromophore Mutations Provide Enhanced Sensitivity with the Green Fluorescent Protein. *Nucleic Acids Res* 24: 4592–4593

Yao X, Rosen MK & Gardner KH (2008) Estimation of the available free energy in a LOV2-Jα photoswitch. *Nat Chem Biol* 4: 491–497

Ye H & Fussenegger M (2019) Optogenetic Medicine: Synthetic Therapeutic Solutions Precision-Guided by Light. *Cold Spring Harb Perspect Med* 9: a034371

Yosef I, Goren MG & Qimron U (2012) Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. *Nucleic Acids Res* 40: 5569–5576

Younger AKD, Su PY, Shepard AJ, Udani SV, Cybulski TR, Tyo KEJ & Leonard JN (2018) Development of novel metabolite-responsive transcription factors via transposon-mediated protein fusion. *Protein Eng Des Sel* 31: 55–63

Yousef MS, Baase WA & Matthews BW (2004) Use of sequence duplication to engineer a ligand-triggered, long-distance molecular switch in T4 lysozyme. *Proc Natl Acad Sci* 101: 11583–11586

Yu K, Liu C, Kim BG & Lee DY (2015) Synthetic fusion protein design and applications. *Biotechnol Adv* 33: 155–164

Yumerefendi H, Dickinson DJ, Wang H & Zimmerman SP (2015) Control of Protein Activity and Cell Fate Specification via Light-Mediated Nuclear Translocation. *PLOS ONE*: 1–19

Yumerefendi H, Lerner AM, Zimmerman SP, Hahn K, Bear JE, Strahl BD & Kuhlman B (2016) Light-induced nuclear export reveals rapid dynamics of epigenetic modifications. *Nat Chem Biol* 12: 399–401

Zalatan JG, Lee ME, Almeida R, Gilbert LA, Whitehead EH, La Russa M, Tsai JC, Weissman JS, Dueber JE, Qi LS, *et al* (2015) Engineering Complex Synthetic Transcriptional Programs with CRISPR RNA Scaffolds. *Cell* 160: 339–350

Zayner JP, Antoniou C & Sosnick TR (2012) The amino-terminal helix modulates light-activated conformational changes in AsLOV2. *J Mol Biol* 419: 61–74

Zetsche B, Gootenberg JS, Abudayyeh OO, Slaymaker IM, Makarova KS, Essletzbichler P, Volz SE, Joung J, Van Der Oost J, Regev A, *et al* (2015a) Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell* 163: 759–771

Zetsche B, Volz SE & Zhang F (2015b) A split-Cas9 architecture for inducible genome editing and transcription modulation. *Nat Biotechnol* 33: 139–142

Zhang D, Jin S, Piao X & Devaraj NK (2020) Multiplexed Photoactivation of mRNA with Single-Cell Resolution. *ACS Chem Biol* 15: 1773–1779

Zhang F (2019) Development of CRISPR-Cas systems for genome editing and beyond. *Q Rev Biophys* 52: e6

Zhang Y, Li P, Pan F, Liu H, Hong P, Liu X & Zhang J (2021) Applications of AlphaFold beyond Protein Structure Prediction Bioinformatics

Zhao C, Zhao Y, Zhang J, Lu J, Chen L, Zhang Y, Ying Y, Xu J, Wei S & Wang Y (2018) HIT-Cas9: A CRISPR/Cas9 Genome-Editing Device under Tight and Effective Drug Control. *Mol Ther - Nucleic Acids* 13: 208–219

Zhao H & Arnold FH (1997) Combinatorial protein design: strategies for screening protein libraries. *Curr Opin Struct Biol* 7: 480–485

Zheng W, Brooks BR & Thirumalai D (2006) Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. *Proc Natl Acad Sci* 103: 7664–7669

Zhou H & Zhou Y (2004) Quantifying the effect of burial of amino acid residues on protein stability. *Proteins Struct Funct Bioinforma* 54: 315–322

Zhou XX, Fan LZ, Li P & Lin MZ (2017a) Optical control of cell signaling by single-chain photoswitchable kinases. *Science* 2: 836–842

Zhou XX, Zou X, Chung HK, Gao Y, Liu Y, Qi LS & Lin MZ (2017b) A single-chain photoswitchable CRISPR-Cas9 architecture for light-inducible gene editing and transcription. *ACS Chem Biol*: acschembio.7b00603

Zhu Y, Gao A, Zhan Q, Wang Y, Feng H, Liu S, Gao G, Serganov A & Gao P (2019) Diverse Mechanisms of CRISPR-Cas9 Inhibition by Type IIC Anti-CRISPR Proteins. *Mol Cell* 74: 296-309.e7

Zoltowski BD, Vaccaro B & Crane BR (2009) Mechanism-based tuning of a LOV domain photoreceptor. *Nat Chem Biol* 5: 827–834

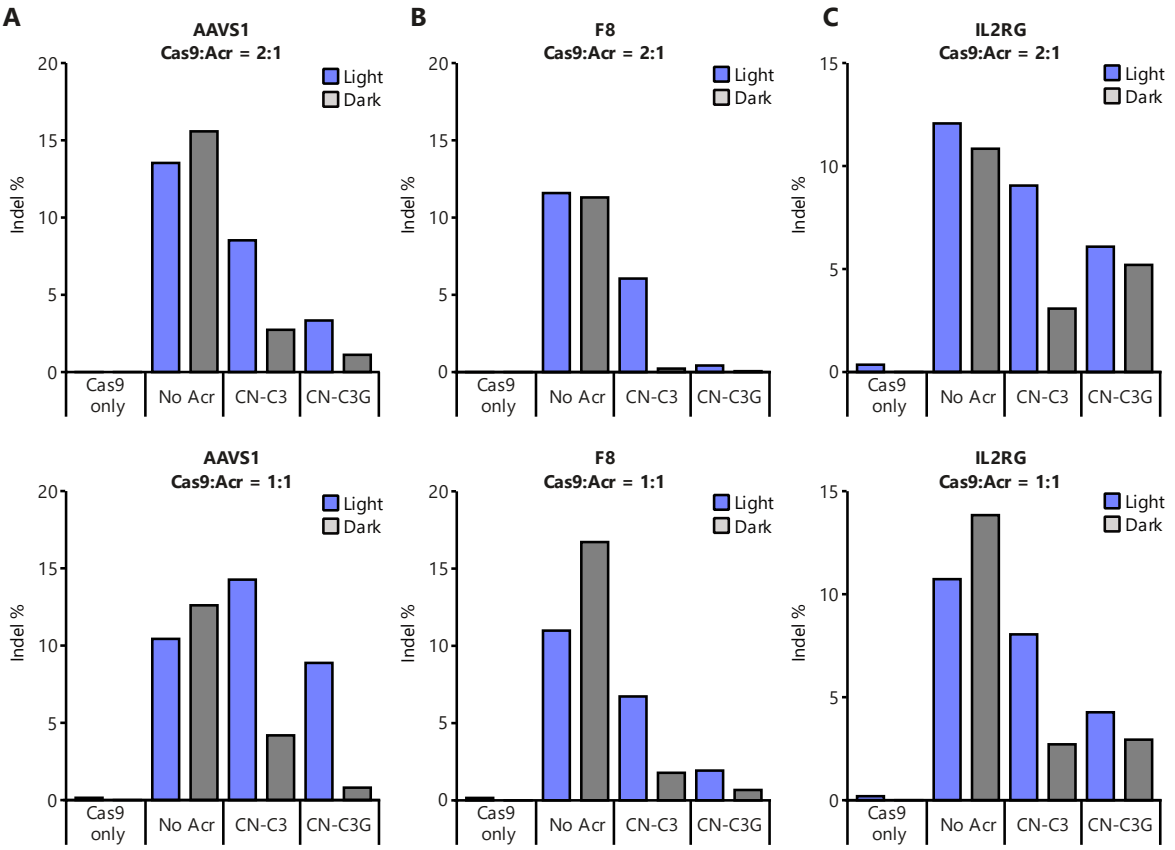
6 Appendix

6.1 Supplementary note 1 – Evaluation of different TVMV reporters

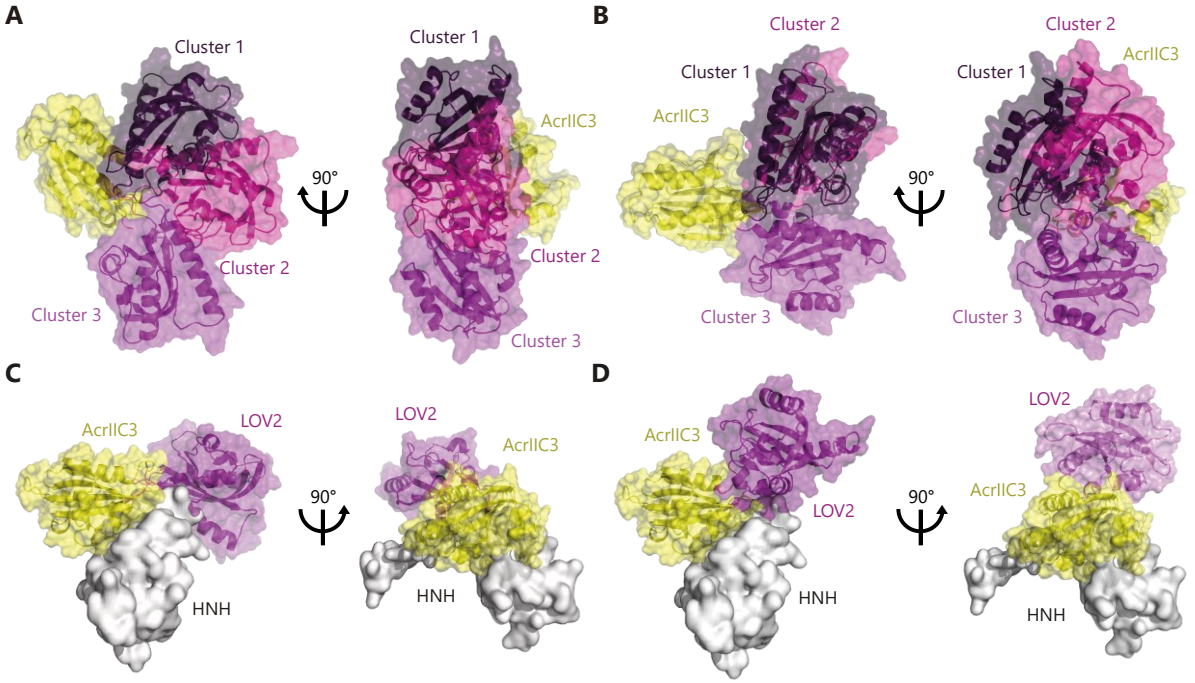
In contrast to other reporter assays, the dynamic range of the TVMV protease assay was relatively low. Prior to the FACS screening experiments, different reporter setups were tested. They comprised two promoter configurations driving the expression of the RFP degron reporter, the strong constitutive J23102 and the weaker J23105 promoter (http://parts.igem.org/Promoters/Catalog/Anderson). In addition, different degradation tags were tested. One strategy was to fuse the M0051 or M0052 (McGinness *et al*, 2006) tag C-terminally to RFP via the TVMV protease recognition site, which acted as linker. In this scenario, the active protease would cleave the degradation tag off, thus stabilizing the reporter. The second strategy was adapted from Fernandez-Rodriguez et al. (Fernandez-Rodriguez & Voigt, 2016). The protease recognition site was positioned at the N-terminus of the fusion protein, followed by either Y- or F-degron sequences and lastly the RFP reporter. In this case, the protease recognition site was expected to shield the tag so that it remained inactive. Upon proteolytic cleavage by TVMV, however, the degradation machinery recognizes the tag and hence degrades the reporter.

Evaluation of the reporters via measurements of the RFP fluorescence revealed modest changes of the reporter expression upon induction of protease expression (Supp. fig. 12). Here, only minimal differences were observed between the samples that were induced with 200 µM IPTG and those induced with 400 µM IPTG. In case of the stronger J23102 promoter, rather high fluorescence levels could be reached in combination with the M0052 tag (Supp. fig. 12A). However, the changes in fluorescence between the native and the induced state were rather small. The N-terminal degradation tags, in turn, resulted in much lower overall fluorescence levels (Supp. fig. 12A). Even lower fluorescence levels were measured when the reporter was under control of the J23105 promoter (Supp. fig. 12B). Here, the induction of degradation by the F- and Y- degron designs turned out to be extremely weak. A clearer effect could be measured with the C-terminal fusions though, which exhibited the expected increase in fluorescence upon induction of TVMV protease expression. Overall, the combination of the J23105 promoter with the M0051 degradation tag showed the highest dynamic range and was hence as TVMV reporter in the FACS screen (Supp. fig. 12A, B).

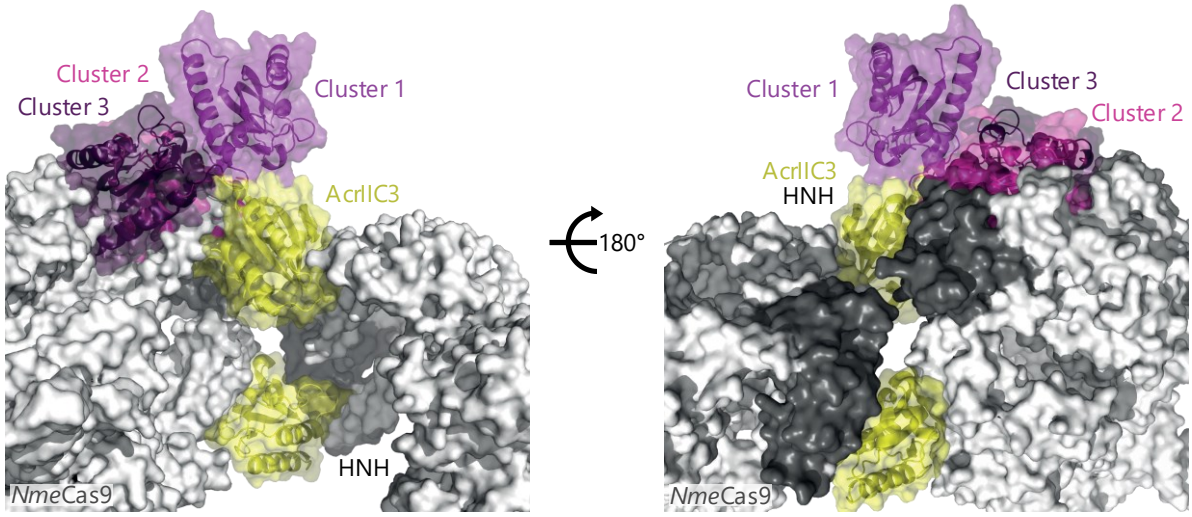
6.2 Supplementary figures



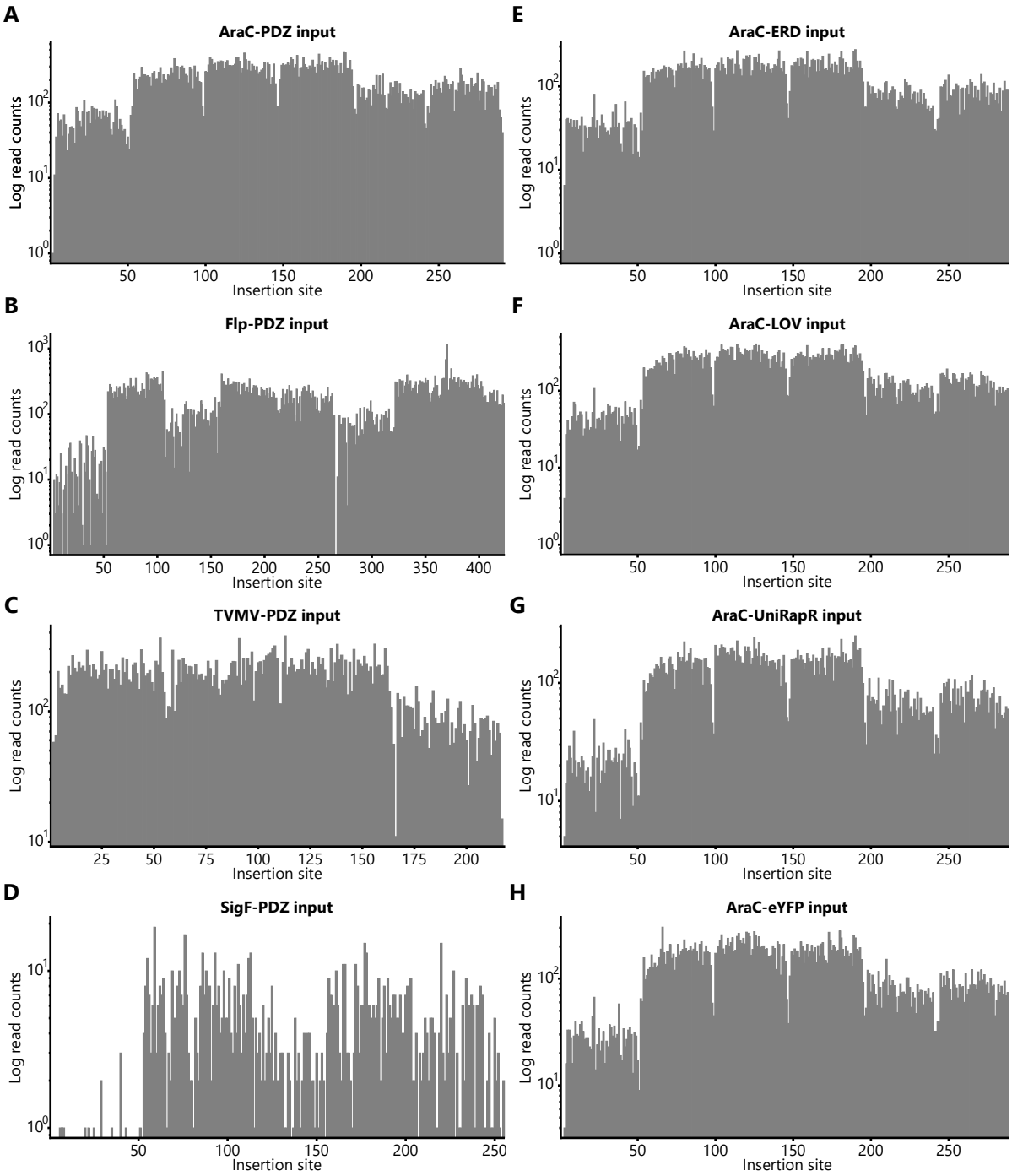
Supplementary figure 1: *Nme*Cas9 inhibition by CN-C3 can be tuned by adjusting the transfected vector dose. (A-C) HEK293T cells were transfected with plasmids encoding *Nme*Cas9, a sgRNA targeting the endogenous AAVS1 (A), F8 (B) or IL2RG (C) locus and the indicated Acr variant. Cas9:Acr vector mass ratios used for transfection are shown in the figure. The editing efficiencies were assessed by T7E assay 72 h post transfection. The bars represent indel frequencies of a single experiment.



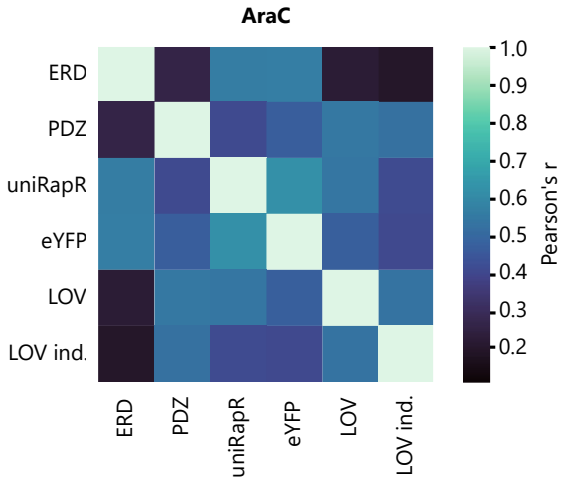
Supplementary figure 2: Structural modelling of CN-C3. (A-D) Structural models CN-C3 (A, C) and CN-C3G (B, D) are shown. The models were generated by domain assembly simulation, performed by Julius Upmeier zu Belzen. The most populated conformational clusters of the LOV2 domain are located in diverse positions relative to AcrIIC3 (A, B). Superimposition onto the structure of AcrIIC3 bound to the HNH domain of *Nme*Cas9, reveals a conformation without sterical clashes (C, D). PDB-IDs: 6J9N, 2V0W.



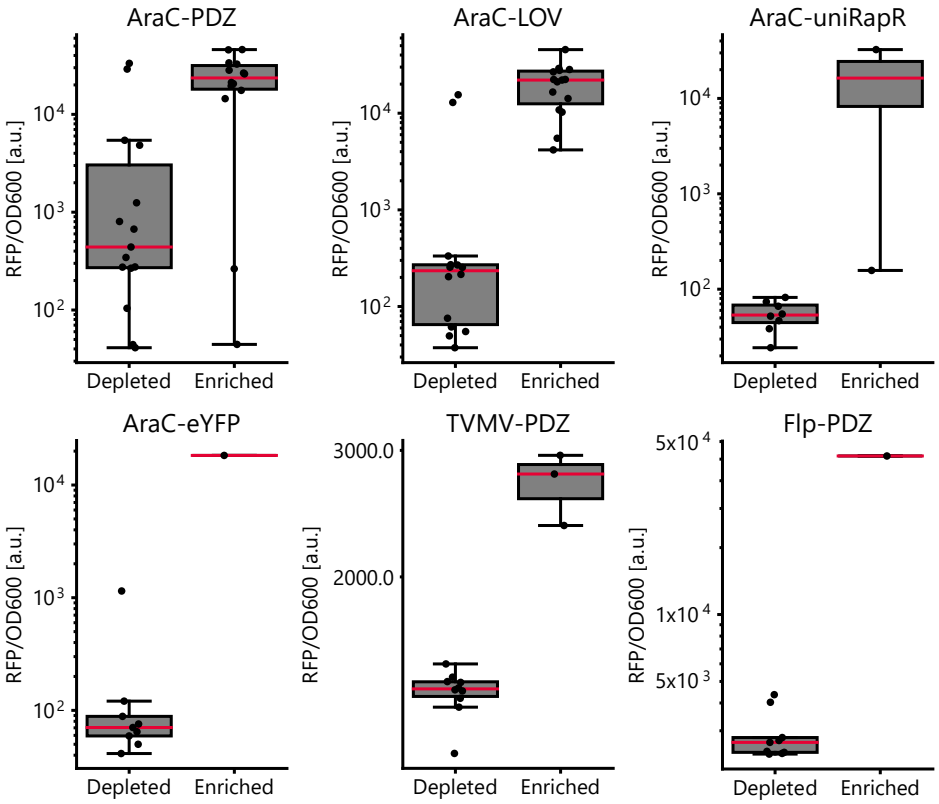
Supplementary figure 3: Two of the three most populated clusters of the LOV2 domain sterically clash with *Nme*Cas9. The structural model of CN-C3 was superimposed on the structure of the dimerized conformation of *Nme*Cas9 bound by AcrIIC3. The HNH domain is depicted in dark grey. PDB-IDs: 6J9N, 2V0W, 6JE9.



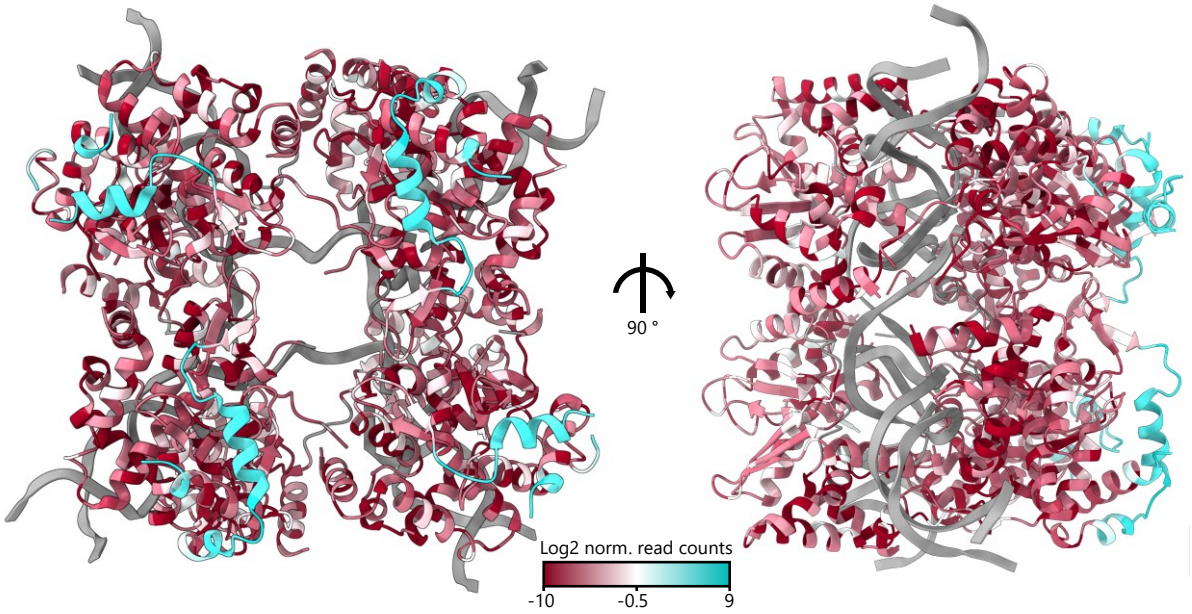
Supplementary figure 4: Cloning of domain insertion libraries via SPINE yields near-complete coverage of domain insertion positions. (A-H) The insertion library coverage was assessed via NGS. Histograms represent the log-normalized read counts for insertions at the respective position (amino acid/codon).



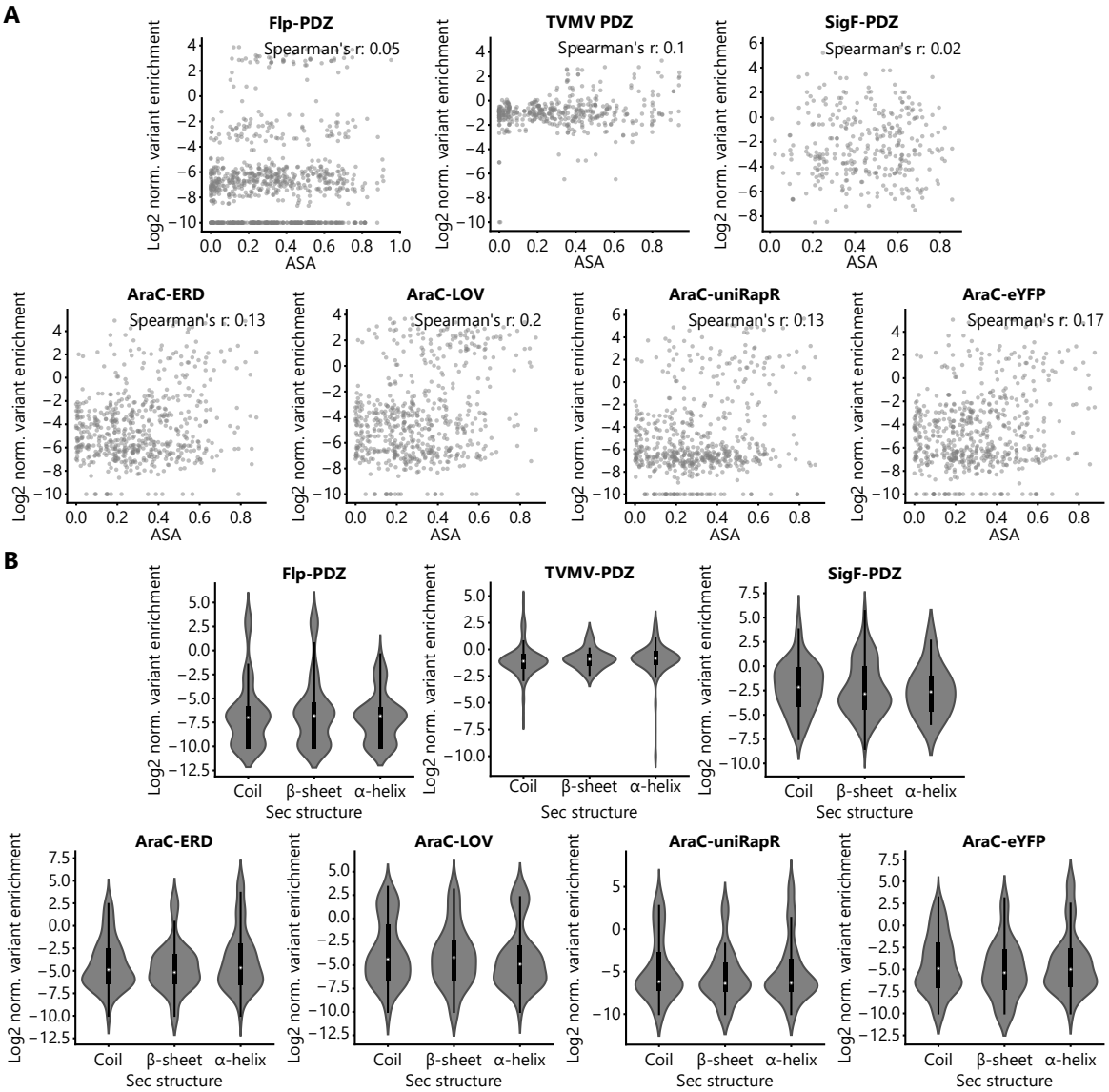
Supplementary figure 5: Pairwise correlations between enrichment scores of different domains inserted into AraC. The heatmap shows pairwise Pearson correlations between all domain inserted into AraC. Enrichments of the AraC-LOV2 library in darkness and under light induction (ind.) were assessed and are depicted separately.



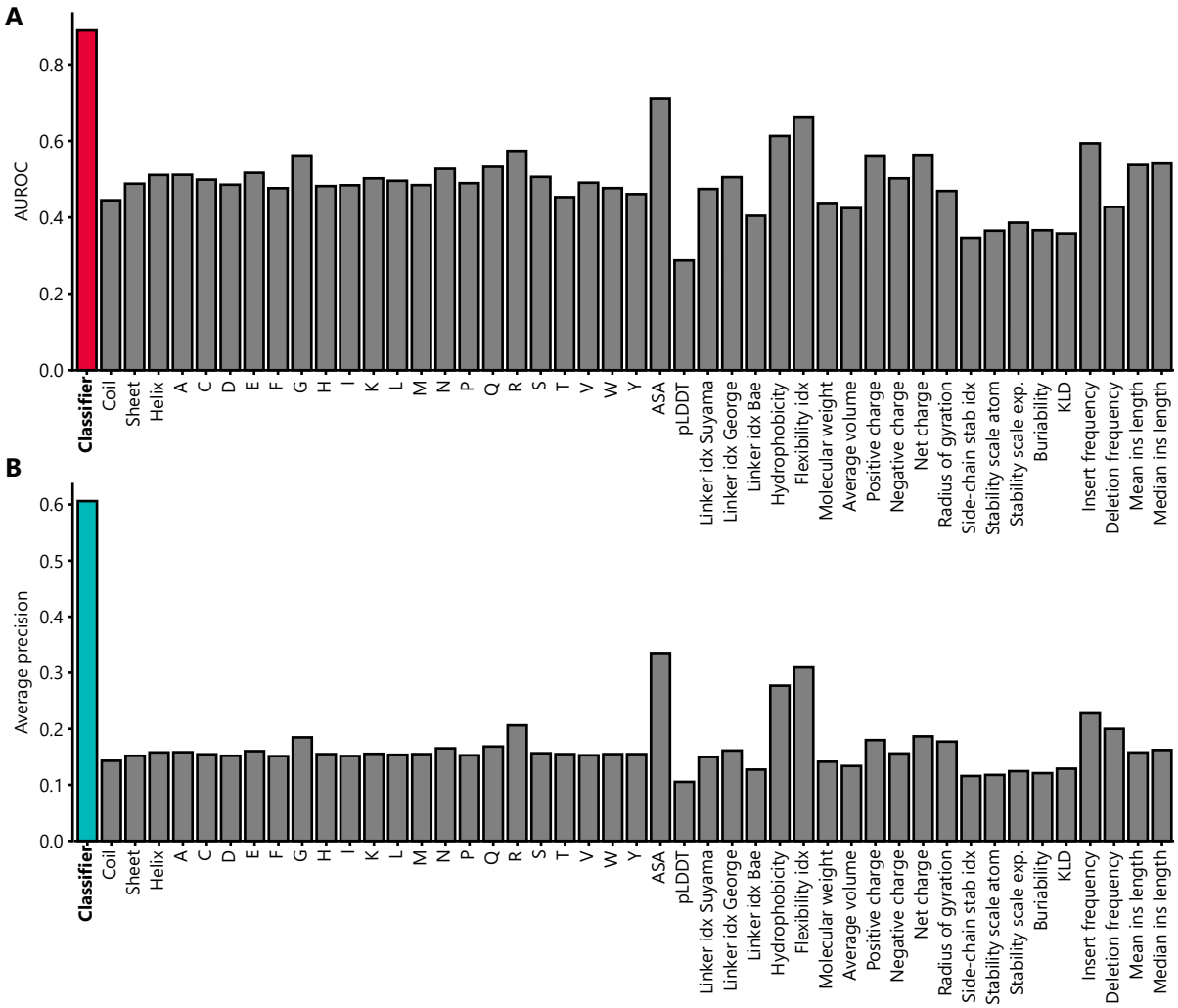
Supplementary figure 6: Validation of the domain insertion screen by characterization of individual insertion variants. Individual domain insertion variants were cloned and their activity was assessed using the respective RFP reporter assays. To this end, cells were grown 200 µl in 96-well plates in presence of inducers, overnight. RFP and OD600 values were assessed by plate reader measurements. Boxplots indicate the resulting normalized fluorescence by enriched and depleted candidates, respectively. Individual data points correspond to the mean of three biological replicates, each of which consisted of three technical replicates. The data was obtained in collaboration with Sabine Aschenbrenner. The IQR is marked by the box and the median is represented by a red line. Whiskers extend to the 1.5-fold IQR or to the value of the smallest or largest enrichment, respectively.



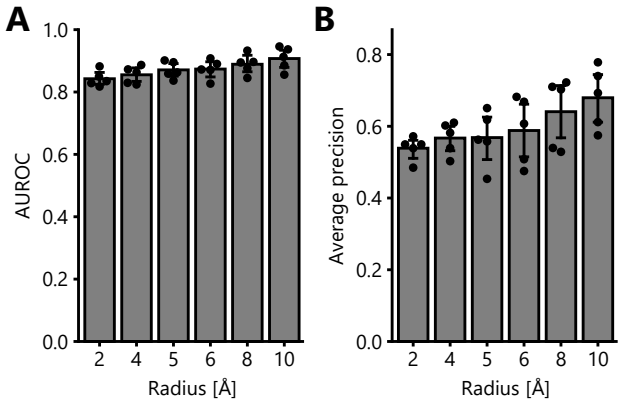
Supplementary figure 7: Enrichment scores mapped onto structures of the Flp-holliday junction complex.
PDB-ID: 1FLO.



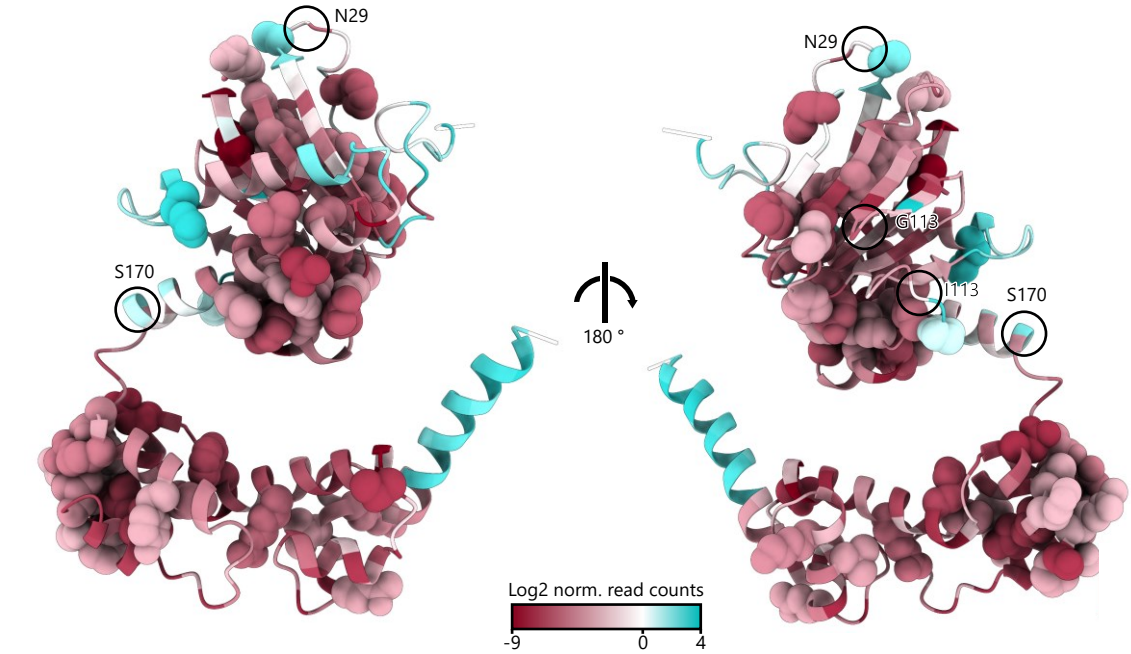
Supplementary figure 8: Correlations between the enrichment scores and surface accessibility or secondary structures. (A) Scatter plot showing the relation between variant enrichment and the average surface exposed area (ASA) of the residues neighboring an insertion site. (B) The insertion score in regions with the respective secondary structure element are shown. For each insertion site, the secondary structure assignment of the amino acid prior and after the insertion were considered. The IQR is marked by the box and the median is represented by the white dot. Whiskers extend to the 1.5-fold IQR or to the value of the smallest or largest enrichment, respectively.



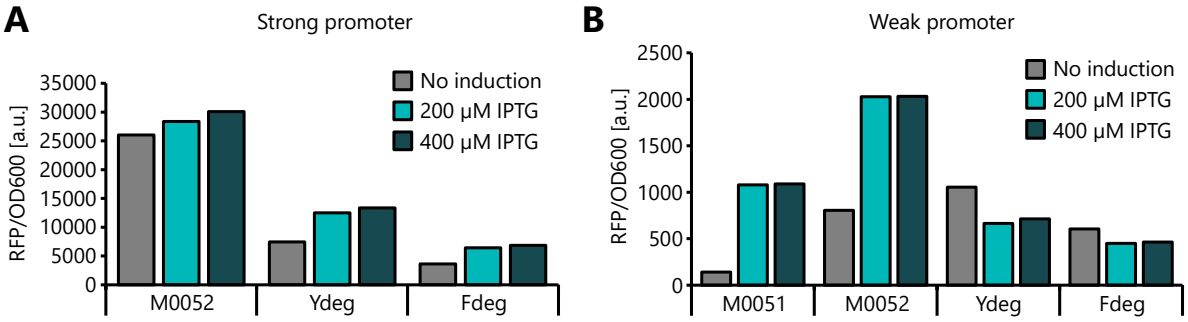
Supplementary figure 9: Full comparison of the trained classifier to baseline predictors. (A, B) The AUROC (A) and average precision (B) are shown. The values were calculated on a previously withheld test set. The performance of the gradient boosting classifier is compared to all individual features.



Supplementary figure 10: Influence of the insertion site context on classifier performance. The mean AUROC (A) and mean average precision (B) are indicated for models trained on datasets consisting of all four candidate proteins. Each data point in the training set represented the average of information derived from the residues within the indicated distance to the insertion site, resulting in different levels of surrounding context used for prediction. (A, B) Individual data points represent cross-validation folds. Bars represent the mean and error bars the SD.



Supplementary figure 11: Analysis of the AraC protein sector. An AF2 predicted full-length structure of AraC is shown. The enrichment scores are mapped onto the structure. Residues belonging to the protein sector are highlighted as spheres. Key residues within the allosterically switchable clusters are marked by circles.



Supplementary figure 12: Analysis of different TVMV protease reporters. (A, B) Samples were inoculated from precultures carrying plasmids encoding the TVMV protease and an RFP reporter under control of either (A) a strong constitutive promoter (J23102) or (B) a weaker one (J12305). Different reporter constructs that were tested are indicated below the bars. IPTG was added in the indicated concentrations. The samples were incubated for 16 h at 37 °C and 220 rpm before RFP fluorescence and the OD at 600 nm were assessed by plate reader measurements. Bars represent values from a single experiment.

6.3 Supplementary tables

Supplementary table 1: Constructs used in this study. NLS, nuclear localization signal; CMV, cytomegalovirus.

#	Name	**Description**. In sequential order	Source
1	pBluescript	empty vector	Invitrogen
2	h*Nme*Cas9 + sgRNA scaffold (pEJS654 All-in-One AAV-sgRNA-h*Nme*Cas9; Addgene plasmid: #112139)	U1a promoter, NLS h*Nme*Cas9 NLS 3xHA; U6 promoter, sgRNA scaffold	(Ibraheim *et al*, 2018)
3	h*Nme*Cas9 + VEGFA sgRNA (AAV)	U1a promoter, NLS h*Nme*Cas9 NLS 3xHA; U6 promoter, VEGFA sgRNA	(Hoffmann *et al*, 2019)
4	h*Nme*Cas9 + AAVS1 sgRNA (AAV)	U1a promoter, NLS h*Nme*Cas9 NLS 3xHA; U6 promoter, AAVS1 sgRNA	(Mathony *et al*, 2020a)
5	h*Nme*Cas9 + F8 sgRNA (AAV)	U1a promoter, NLS h*Nme*Cas9 NLS 3xHA; U6 promoter, F8 sgRNA	(Mathony *et al*, 2020a)
6	h*Nme*Cas9 + IL2RG sgRNA (AAV)	U1a promoter, NLS h*Nme*Cas9 NLS 3xHA; U6 promoter, IL2RG sgRNA	(Mathony *et al*, 2020a)
7	h*Nme*Cas9 + SLC9A9 sgRNA (AAV)	U1a promoter, NLS h*Nme*Cas9 NLS 3xHA; U6 promoter, SLC9A9 sgRNA	(Hoffmann *et al*, 2019)
8	h*Nme*2Cas9 + sgRNA scaffold	NLS h*Nme*2Cas9 NLS 3xHA; U6 promoter, sgRNA scaffold	(Edraki *et al*, 2018)
9	h*Nme*2Cas9 + VEGFA sgRNA	NLS h*Nme*2Cas9 NLS 3xHA; U6 promoter, VEGFA sgRNA	(Mathony *et al*, 2020a)
10	h*Nme*2Cas9 + FANCJ sgRNA	NLS h*Nme*2Cas9 NLS 3xHA; U6 promoter, FANCJ sgRNA	(Mathony *et al*, 2020a)
11	h*Nme*2Cas9 + GAPDH sgRNA	NLS h*Nme*2Cas9 NLS 3xHA; U6 promoter, GAPDH sgRNA	(Mathony *et al*, 2020a)
12	hSaCas9 + sgRNA scaffold	CMV promoter, NLS hSaCas9 NLS 3xHA; U6 promoter, sgRNA scaffold	(Ran *et al*, 2015)
13	hSaCas9 + EMX1 sgRNA	CMV promoter, NLS hSaCas9 NLS 3xHA; U6 promoter, EMX1 sgRNA	(Mathony *et al*, 2020a)
14	AcrIIC1	CMV promoter, AcrIIC1	(Hoffmann *et al*, 2019)
15	AcrIIC1 chimera-7	CMV promoter, AcrIIC1 with mCherry insertion between E68 and Y72	(Mathony *et al*, 2020a)
16	AcrIIC1 chimera-10	CMV promoter, AcrIIC1 with GSG-mCherry-GSG insertion behind Y70	(Mathony *et al*, 2020a)
17	AcrIIC3	CMV promoter, AcrIIC3	(Hoffmann *et al*, 2019)
18	AcrX (AcrIIC1_N3F/D15Q/A48I)	CMV promoter, AcrIIC1_N3F/D15Q/A48I	(Mathony *et al*, 2020a)
19	AcrX*	CMV promoter, AcrIIC1_N3F/D15Q/A48I with GSG-mCherry-GSG insertion behind Y70	(Mathony *et al*, 2020a)
20	AcrIIC1-scaffold	CMV promoter, AcrIIC1 with scaffold for miRNA binding site insertion in the 3'-UTR	(Hoffmann *et al*, 2019)
21	AcrIIC1-miR-122	CMV promoter, AcrIIC1 with 2x miR-122 binding sites in the 3'-UTR	(Mathony *et al*, 2020a)
22	AcrX-scaffold	CMV promoter, AcrX with scaffold for miRNA binding site insertion in the 3'-UTR	(Mathony *et al*, 2020a)
23	AcrX-miR-122	CMV promoter, AcrX with 2x miR-122 binding sites in the 3'-UTR	(Mathony *et al*, 2020a)
24	AcrIIC3 F59-LOV2-N60 (**CN-C3**) (Addgene: #137191)	CMV promoter, AcrIIC3 F59-LOV2-N60	(Hoffmann *et al*, 2021)
25	AcrIIC3 F59-G-LOV2-G-N60 (**CN-C3G**) (Addgene: #137192)	CMV promoter, AcrIIC3 F59-G-LOV2-G-N60	(Hoffmann *et al*, 2021)
26	HA-AcrIIC3	CMV promoter, HA-AcrIIC3	(Hoffmann *et al*, 2021), this work

27	HA-CN-C3	CMV promoter, HA-CN-C3	(Hoffmann *et al*, 2021), this work
28	RFP reporter for AraC	BAD promoter, mRFP1, LVA degradation tag	This work
29	RFP reporter for SigF	F1 promoter, mRFP1	This work
30	RFP reporter for Flp recombinase	J23102 promoter, inverted mRFP1, flanked by FRT sites	This work
31	RFP reporter for TVMV protease (J23105 + M0051)	J23105 promoter, mRFP1, TVMV recognition site, (M0051) DAS+2 degradation tag	This work
32	RFP reporter for TVMV protease (J23105 + M0052)	J23105 promoter, mRFP1, TVMV recognition site, M0052 degradation tag	This work
33	RFP reporter for TVMV protease (J23102 + M0051)	J23102 promoter, mRFP1, TVMV recognition site, (M0051) DAS+2 degradation tag	This work
34	RFP reporter for TVMV protease (J23102 + M0052)	J23102 promoter, mRFP1, TVMV recognition site, M0052 degradation tag	This work
35	RFP reporter for TVMV protease (J23105 + Ydeg)	J23105 promoter, TVMV recognition site, Ydeg tag, mRFP1	This work
36	RFP reporter for TVMV protease (J23102 + Ydeg)	J23102 promoter, TVMV recognition site, Ydeg tag, mRFP1	This work
37	RFP reporter for TVMV protease (J23102 + Fdeg)	J23102 promoter, TVMV recognition site, Fdeg tag, mRFP1	This work
38	AraC	TRC promoter, AraC	This work
39	Flp recombinase	TRC promoter, Flp recombinase	This work
40	TVMV protease	TRC promoter, TVMV protease	This work
41	SigF	TRC promoter, SigF	This work
42	AraC_S170_LOV2	TRC promoter, AraC with insertion of AsLOV2 behind S170	This work
43	AraC_I113_LOV2	TRC promoter, AraC with insertion of AsLOV2 behind I113	This work
44	AraC_E3_PDZ	AraC with insertion of PDZ behind E3	This work
45	AraC_S14_PDZ	AraC with insertion of PDZ behind S14	This work
46	AraC_N16_PDZ	AraC with insertion of PDZ behind N16	This work
47	AraC_A17_PDZ	AraC with insertion of PDZ behind A17	This work
48	AraC_L23_PDZ	AraC with insertion of PDZ behind L23	This work
49	AraC_E27_PDZ	AraC with insertion of PDZ behind E27	This work
50	AraC_T50_PDZ	AraC with insertion of PDZ behind T50	This work
51	AraC_Q60_PDZ	AraC with insertion of PDZ behind Q60	This work
52	AraC_E63_PDZ	AraC with insertion of PDZ behind E63	This work
53	AraC_S112_PDZ	AraC with insertion of PDZ behind S112	This work
54	AraC_I113_PDZ	AraC with insertion of PDZ behind I113	This work
55	AraC_N116_PDZ	AraC with insertion of PDZ behind N116	This work
56	AraC_R121_PDZ	AraC with insertion of PDZ behind R121	This work
57	AraC_H129_PDZ	AraC with insertion of PDZ behind H129	This work
58	AraC_G143_PDZ	AraC with insertion of PDZ behind G143	This work
59	AraC_E165_PDZ	AraC with insertion of PDZ behind E165	This work
60	AraC_S170_PDZ	AraC with insertion of PDZ behind S170	This work
61	AraC_T241_PDZ	AraC with insertion of PDZ behind T241	This work
62	AraC_D286_PDZ	AraC with insertion of PDZ behind D286	This work
63	AraC_E3_LOV2	AraC with insertion of LOV2 behind E3	This work
64	AraC_S14_LOV2	AraC with insertion of LOV2 behind S14	This work

65	AraC_N16_LOV2	AraC with insertion of LOV2 behind N16	This work
66	AraC_A17_LOV2	AraC with insertion of LOV2 behind A17	This work
67	AraC_L23_LOV2	AraC with insertion of LOV2 behind L23	This work
68	AraC_E27_LOV2	AraC with insertion of LOV2 behind E27	This work
69	AraC_T50_LOV2	AraC with insertion of LOV2 behind T50	This work
70	AraC_Q60_LOV2	AraC with insertion of LOV2 behind Q60	This work
71	AraC_E63_LOV2	AraC with insertion of LOV2 behind E63	This work
72	AraC_S112_LOV2	AraC with insertion of LOV2 behind S112	This work
73	AraC_I113_LOV2	AraC with insertion of LOV2 behind I113	This work
74	AraC_N116_LOV2	AraC with insertion of LOV2 behind N116	This work
75	AraC_R121_LOV2	AraC with insertion of LOV2 behind R121	This work
76	AraC_H129_LOV2	AraC with insertion of LOV2 behind H129	This work
77	AraC_G143_LOV2	AraC with insertion of LOV2 behind G143	This work
78	AraC_E165_LOV2	AraC with insertion of LOV2 behind E165	This work
79	AraC_S170_LOV2	AraC with insertion of LOV2 behind S170	This work
80	AraC_T241_LOV2	AraC with insertion of LOV2 behind T241	This work
81	AraC_D286_LOV2	AraC with insertion of LOV2 behind D286	This work
82	AraC_E3_eYFP	AraC with insertion of eYFP behind E3	This work
83	AraC_N16_eYFP	AraC with insertion of eYFP behind N16	This work
84	AraC_L23_eYFP	AraC with insertion of eYFP behind L23	This work
85	AraC_T50_eYFP	AraC with insertion of eYFP behind T50	This work
86	AraC_Q60_eYFP	AraC with insertion of eYFP behind Q60	This work
87	AraC_I113_eYFP	AraC with insertion of eYFP behind I113	This work
88	AraC_N116_eYFP	AraC with insertion of eYFP behind N116	This work
89	AraC_E165_eYFP	AraC with insertion of eYFP behind E165	This work
90	AraC_S170_eYFP	AraC with insertion of eYFP behind S170	This work
91	AraC_T241_eYFP	AraC with insertion of eYFP behind T241	This work
92	AraC_E3_ERD	AraC with insertion of ERD behind E3	This work
93	AraC_N16_ERD	AraC with insertion of ERD behind N16	This work
94	AraC_L23_ERD	AraC with insertion of ERD behind L23	This work
95	AraC_T50_ERD	AraC with insertion of ERD behind T50	This work
96	AraC_Q60_ERD	AraC with insertion of ERD behind Q60	This work
97	AraC_I113_ERD	AraC with insertion of ERD behind I113	This work
98	AraC_N116_ERD	AraC with insertion of ERD behind N116	This work
99	AraC_E165_ERD	AraC with insertion of ERD behind E165	This work
100	AraC_S170_ERD	AraC with insertion of ERD behind S170	This work
101	AraC_T241_ERD	AraC with insertion of ERD behind T241	This work
102	AraC_E3_uniRapR	AraC with insertion of uniRapR behind E3	This work
103	AraC_N16_uniRapR	AraC with insertion of uniRapR behind N16	This work
104	AraC_L23_uniRapR	AraC with insertion of uniRapR behind L23	This work
105	AraC_T50_uniRapR	AraC with insertion of uniRapR behind T50	This work
106	AraC_Q60_uniRapR	AraC with insertion of uniRapR behind Q60	This work
107	AraC_I113_uniRapR	AraC with insertion of uniRapR behind I113	This work
108	AraC_N116_uniRapR	AraC with insertion of uniRapR behind N116	This work
109	AraC_E165_uniRapR	AraC with insertion of uniRapR behind E165	This work

110	AraC_S170_uniRapR	AraC with insertion of uniRapR behind S170	This work
111	AraC_T241_uniRapR	AraC with insertion of uniRapR behind T241	This work
112	TVMV_L5_PDZ	TVMV with insertion of PDZ behind L5	This work
113	TVMV_D11_PDZ	TVMV with insertion of PDZ behind D11	This work
114	TVMV_G37_PDZ	TVMV with insertion of PDZ behind G37	This work
115	TVMV_I42_PDZ	TVMV with insertion of PDZ behind I42	This work
116	TVMV_L56_PDZ	TVMV with insertion of PDZ behind L56	This work
117	TVMV_T105_PDZ	TVMV with insertion of PDZ behind T105	This work
118	TVMV_S121_PDZ	TVMV with insertion of PDZ behind S121	This work
119	TVMV_H143_PDZ	TVMV with insertion of PDZ behind H143	This work
120	TVMV_F187_PDZ	TVMV with insertion of PDZ behind F187	This work
121	TVMV_D193_PDZ	TVMV with insertion of PDZ behind D193	This work
122	TVMV_W198_PDZ	TVMV with insertion of PDZ behind W198	This work
123	TVMV_F204_PDZ	TVMV with insertion of PDZ behind F204	This work
124	TVMV_I209_PDZ	TVMV with insertion of PDZ behind I209	This work
125	Flp_L15_PDZ	Flp with insertion of PDZ behind L15	This work
126	Flp_C42_PDZ	Flp with insertion of PDZ behind C42	This work
127	Flp_D115_PDZ	Flp with insertion of PDZ behind D115	This work
128	Flp_S129_PDZ	Flp with insertion of PDZ behind S129	This work
129	Flp_L151_PDZ	Flp with insertion of PDZ behind L151	This work
130	Flp_I239_PDZ	Flp with insertion of PDZ behind I239	This work
131	Flp_N290_PDZ	Flp with insertion of PDZ behind N290	This work
132	Flp_W330_PDZ	Flp with insertion of PDZ behind W330	This work
133	Flp_S397_PDZ	Flp with insertion of PDZ behind S397	This work
134	Flp_Y403_PDZ	Flp with insertion of PDZ behind Y403	This work

Supplementary table 2: Amino acid sequences of the used proteins and insert domains. Tags and NLS' linked to the proteins are marked in bold. The *As*LOV2 amino acids marked in red were only present in CN-C3, but not part of the *As*LOV2 used for insertion screens and AraC hybrids.

Protein	Sequence
AcrIIC1	MANKTYKIGKNAGYDGCGLCLAAISENEAIKVKYLRDICPDYDGDDKAEDWLRWGTDSRVKAA ALEMEQYAYTSVGMASCWEFVEL
AcrIIC3	MAFKRAIIFTSFNGFEKVSRTEKRRLAKIINARVSIIDEYLRAKDTNASLDGQYRAFLFNDES PAMTEFLAKLKAFAESCTGISIDAWEIEESEYVRLPVERRDFLAAANGKEIFKI
AraC	MSAEAQNDPLLPGYSFNAHLVAGLTPIEANGYLDFFIDRPLGMKGYILNLTIRGQGVVKNQGR EFVCRPGDILLFPPGEIHHYGRHPEAREWYHQWVYFRPRAYWHEWLNWPSIFANTGFFRPDEA HQPHFSDLFGQIINAGQGEGRYSELLAINLLEQLLLRRMEAINESLHPPMDNRVREACQYISD HLADSNFDIASVAQHVCLSPSRLSHLFRQQLGISVLSWREDQRISQAKLLLSTTRMPIATVGR NVGFDDQLYFSRVFKKCTGASPSEFRAGCEEKVNDVAVKLS**GHHHHHH**
*As*LOV2	LATTLERIEKNFVITDPRLPDNPIIFASDSFLQLTEYSREEILGRNCRFLQGPETDRATVRKI RDAIDNQTEVTVQLINYTKSGKKFWNLFHLQPMRDQKGDVQYFIGVQLDGTEHVRDAAEREGV MLIKKTAENIDEAAK**EL**
ER domain	GPLDNSLALSLTADQMVSALLDAEPPILYSEYDPTRPFSEASMMGLLTNLADRELVHMINWAK RVPGFVDLTLHDQVHLLECAWLEILMIGLVWRSMEHPGKLLFAPNLLLDRNQGKCVEGMVEIF DMLLATSSRFRMMNLQGEEFVCLKSIILLNSGVYTFLSSTLKSLEEKDHIHRVLDKITDTLIH LMAKAGLTLQQQHQRLAQLLLILSHIRHMSNKGMEHLYSMKCKNVVPLYDLLLEMLDAHRLHA PGSEL
eYFP	VSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTLVTT FGYGLQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLVNRIELK GIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQQNTPIG DGPVLLPDNHYLSYQSKLSKDPNEKRDHMVLLEFVTAAGITLGMDELYK
Flp recombinase	MSPQFGILCKTPPKVLVRQFVERFERPSGEKIALCAAELTYLCWMITHNGTAIKRATFMSYNT IISNSLSFDIVNKSLQFKYKTQKATILEASLKKLIPAWEFTIIPYYGQKHQSDITDIVSSLQL

	QFESSEEADKGNSHSKKMLKALLSEGESIWEITEKILNSFEYTSRFTKTKTLYQFLFLATFIN CGRFSDIKNVDPKSFKLVQNKYLGVIIQCLVTETKTSVSRHIYFFSARGRIDPLVYLDEFLRN SEPVLKRVNRTGNSSSNKQEYQLLKDNLVRSYNKALKKNAPYSIFAIKNGPKSHIGRHLMTSF LSMKGLTELTNVVGNWSDKRASAVARTTYTHQITAIPDHYFALVSRYYAYDPISKEMIALKDE TNPIEEWQHIEQLKGSAEGSIRYPAWNGIISQEVLDYLSSYINRRI**SGHHHHHH**
mRFP1	MASSEDVIKEFMRFKVRMEGSVNGHEFEIEGEGEGRPYEGTQTAKLKVTKGGPLPFAWDILSP QFQYGSKAYVKHPADIPDYLKLSFPEGFKWERVMNFEDGGVVTVTQDSSLQDGEFIYKVKLRG TNFPSDGPVMQKKTMGWEASTERMYPEDGALKGEIKMRLKLKDGGHYDAEVKTTYMAKKPVQL PGAYKTDIKLDITSHNEDYTIVEQYERAEGRHSTGA
*Nme*2Cas9	**MVPKKKRKVEDKRPAATKKAGQAKKKK**MAAFKPNPINYILGLDIGIASVGWAMVEIDEEENPI RLIDLGVRVFERAEVPKTGDSLAMARRLARSVRRLTRRRAHRLLRARRLLKREGVLQAADFDE NGLIKSLPNTPWQLRAAALDRKLTPLEWSAVLLHLIKHRGYLSQRKNEGETADKELGALLKGV ANNAHALQTGDFRTPAELALNKFEKESGHIRNQRGDYSHTFSRKDLQAELILLFEKQKEFGNP HVSGGLKEGIETLLMTQRPALSGDAVQKMLGHCTFEPAEPKAAKNTYTAERFIWLTKLNNLRI LEQGSERPLTDTERATLMDEPYRKSKLTYAQARKLLGLEDTAFFKGLRYGKDNAEASTLMEMK AYHAISRALEKEGLKDKKSPLNLSSELQDEIGTAFSLFKTDEDITGRLKDRVQPEILEALLKH ISFDKFVQISLKALRRIVPLMEQGKRYDEACAEIYGDHYGKKNTEEKIYLPPIPADEIRNPVV LRALSQARKVINGVVRRYGSPARIHIETAREVGKSFKDRKEIEKRQEENRKDREKAAAKFREY FPNFVGEPKSKDILKLRLYEQQHGKCLYSGKEINLVRLNEKGYVEIDHALPFSRTWDDSFNNK VLVLGSENQNKGNQTPYEYFNGKDNSREWQEFKARVETSRFPRSKKQRILLQKFDEDGFKECN LNDTRYVNRFLCQFVADHILLTGKGKRRVFASNGQITNLLRGFWGLRKVRAENDRHHALDAVV VACSTVAMQQKITRFVRYKEMNAFDGKTIDKETGKVLHQKTHFPQPWEFFAQEVMIRVFGKPD GKPEFEEADTPEKLRTLLAEKLSSRPEAVHEYVTPLFVSRAPNRKMSGAHKDTLRSAKRFVKH NEKISVKRVWLTEIKLADLENMVNYKNGREIELYEALKARLEAYGGNAKQAFDPKDNPFYKKG GQLVKAVRVEKTQESGVLLNKKNAYTIADNGDMVRVDVFCKVDKKGKNQYFIVPIYAWQVAEN ILPDIDCKGYRIDDSYTFCFSLHKYDLIAFQKDEKSKVEFAYYINCDSSNGRFYLAWHDKGSK EQQFRISTQNLVLIQKYQVNELGKEIRPCRLKKRPPVRED**KRPAATKKAGQAKKKKYPYDVPD YAGYPYDVPDYAGSYPYDVPDYAAAPAAKKKKLD**
*Nme*Cas9	**MVPKKKRKVED**AAFKPNSINYILGLDIGIASVGWAMVEIDEEENPIRLIDLGVRVFERAEVPK TGDSLAMARRLARSVRRLTRRRAHRLLRTRRLLKREGVLQAANFDENGLIKSLPNTPWQLRAA ALDRKLTPLEWSAVLLHLIKHRGYLSQRKNEGETADKELGALLKGVAGNAHALQTGDFRTPAE LALNKFEKESGHIRNQRSDYSHTFSRKDLQAELILLFEKQKEFGNPHVSGGLKEGIETLLMTQ RPALSGDAVQKMLGHCTFEPAEPKAAKNTYTAERFIWLTKLNNLRILEQGSERPLTDTERATL MDEPYRKSKLTYAQARKLLGLEDTAFFKGLRYGKDNAEASTLMEMKAYHAISRALEKEGLKDK KSPLNLSPELQDEIGTAFSLFKTDEDITGRLKDRIQPEILEALLKHISFDKFVQISLKALRRI VPLMEQGKRYDEACAEIYGDHYGKKNTEEKIYLPPIPADEIRNPVVLRALSQARKVINGVVRR YGSPARIHIETAREVGKSFKDRKEIEKRQEENRKDREKAAAKFREYFPNFVGEPKSKDILKLR LYEQQHGKCLYSGKEINLGRLNEKGYVEIDHALPFSRTWDDSFNNKVLVLGSENQNKGNQTPY EYFNGKDNSREWQEFKARVETSRFPRSKKQRILLQKFDEDGFKERNLNDTRYVNRFLCQFVAD RMRLTGKGKKRVFASNGQITNLLRGFWGLRKVRAENDRHHALDAVVVACSTVAMQQKITRFVR YKEMNAFDGKTIDKETGEVLHQKTHFPQPWEFFAQEVMIRVFGKPDGKPEFEEADTLEKLRTL LAEKLSSRPEAVHEYVTPLFVSRAPNRKMSGQGHMETVKSAKRLDEGVSVLRVPLTQLKLKDL EKMVNREREPKLYEALKARLEAHKDDPAKAFAEPFYKYDKAGNRTQQVKAVRVEQVQKTGVWV RNHNGIADNATMVRVDVFEKGDKYYLVPIYSWQVAKGILPDRAVVQGKDEEDWQLIDDSFNFK FSLHPNDLVEVITKKARMFGYFASCHRGTGNINIRIHDLDHKIGKNGILEGIGVKTALSFQKY QIDELGKEIRPCRLKKRPPVR**EDKRPAATKKAGQAKKKKYPYDVPDYAGYPYDVPDYAGSYPY DVPDYAAAPAAKKKKLD**
PDZ domain	RRRVTVRKADAGGLGISIKGGRENKMPILISKIFKGLAADQTEALFVGDAILSVNGEDLSSAT HDEAVQALKKTGKEVVLEVKYMK
*Sau*Cas9	**MAPKKKRKVGIHGVPAA**KRNYILGLDIGITSVGYGIIDYETRDVIDAGVRLFKEANVENNEGR RSKRGARRLKRRRRHRIQRVKKLLFDYNLLTDHSELSGINPYEARVKGLSQKLSEEEFSAALL HLAKRRGVHNVNEVEEDTGNELSTKEQISRNSKALEEKYVAELQLERLKKDGEVRGSINRFKT SDYVKEAKQLLKVQKAYHQLDQSFIDTYIDLLETRRTYYEGPGEGSPFGWKDIKEWYEMLMGH CTYFPEELRSVKYAYNADLYNALNDLNNLVITRDENEKLEYYEKFQIIENVFKQKKKPTLKQI AKEILVNEEDIKGYRVTSTGKPEFTNLKVYHDIKDITARKEIIENAELLDQIAKILTIYQSSE DIQEELTNLNSELTQEEIEQISNLKGYTGTHNLSLKAINLILDELWHTNDNQIAIFNRLKLVP KKVDLSQQKEIPTTLVDDFILSPVVKRSFIQSIKVINAIIKKYGLPNDIIIELAREKNSKDAQ KMINEMQKRNRQTNERIEEIIRTTGKENAKYLIEKIKLHDMQEGKCLYSLEAIPLEDLLNNPF NYEVDHIIPRSVSFDNSFNNKVLVKQEENSKKGNRTPFQYLSSSDSKISYETFKKHILNLAKG KGRISKTKKEYLLEERDINRFSVQKDFINRNLVDTRYATRGLMNLLRSYFRVNNLDVKVKSIN GGFTSFLRRKWKFKKERNKGYKHHAEDALIIANADFIFKEWKKLDKAKKVMENQMFEEKQAES MPEIETEQEYKEIFITPHQIKHIKDFKDYKYSHRVDKKPNRELINDTLYSTRKDDKGNTLIVN NLNGLYDKDNDKLKKLINKSPEKLLMYHHDPQTYQKLKLIMEQYGDEKNPLYKYYEETGNYLT KYSKKDNGPVIKKIKYYGNKLNAHLDITDDYPNSRNKVVKLSLKPYRFDVYLDNGVYKFVTVK NLDVIKKENYYEVNSKCYEEAKKLKKISNQAEFIASFYNNDLIKINGELYRVIGVNNDLLNRI EVNMIDITYREYLENMNDKRPPRIIKTIASKTQSIKKYSTDILGNLYEVKSKKHPQIIKKG**KR PAATKKAGQAKKKKGSYPYDVPDYAYPYDVPDYAYPYDVPDYA**
SigF	MSDVEVKKNGKNAQLKDHEVKELIKQSQNGDQQARDLLIEKNMRLVWSVVQRFLNRGYEPDDL FQIGCIGLLKSVDKFDLTYDVRFSTYAVPMIIGEIQRFIRDDGTVKVSRSLKELGNKIRRAKD

	ELSKTLGRVPTVQEIADHLEIEAEDVVLAQEAVRAPSSIHETVYENDGDPITLLDQIADNSEE KWFDKIALKEAISDLEEREKLIVYLRYYKDQTQSEVAERLGISQVQVSRLEKKILKQIKVQMD HTDG
TVMV protease	MSSKALLKGVRDFNPISACVCLLENSSDGHSERLFGIGFGPYIIANQHLFRRNNGELTIKTMH GEFKVKNSTQLQMKPVEGRDIIVIKMAKDFPPFPQKLKFRQPTIKDRVCMVSTNFQQKSVSSL VSESSHIVHKEDTSFWQHWITTKDGQCGSPLVSIIDGNILGIHSLTHTTNGSNYFVEFPEKFV ATYLDAADGWCKNWKFNADKISWGSFTLVE
uniRapR	TCVVHYTGMLEDGKKFDSSRDRNKPFKFMLGKQEVIRGWEEGVAQMSVGQRAKLTISPDYAYG ATGHGSGSGSGVKDLLQAWDLYYHVFRRISGPPGPGSGLWHEMWHEGLEEASRLYFGERNVKG MFEVLEPLHAMMERGPQTLKETSFNQAYGRDLMEAQEWCRKYMKSGSSGGSGSGIIPPHATLV FDVELLKLE

Supplementary table 3: Genomic target sites underlying the CRISPR experiments in mammalian cells. Protospacer sequences are underlined. The PAM sequence is in bold.

Locus	Target sequence including PAM (5′ to 3′)
AAVS1	ACCCCACAGTGGGGCCACTAGGGACAG**GATT**
EMX1	GGCCTCCCCAAAGCCTGGCCAGG**GAGT**
F8	GGTTTCTAGTTGTGACAAGAACACTGGT**GATT**
FANCJ	AAAATTGTGATTTCCAGATCCACAAGC**CC**
GAPDH	CAAGAGCACAAGAGGAAGAGAGAGAC**CC**
IL2RG	CTCTTTCTCCTCAAGGAACAATCAGTG**GATT**
SLC9A9	TGGTCTGGGGTACAGCCTTGGCATCAT**GATT**
VEGFA – *Nme*1Cas9	GCGGGGAGAAGGCCAGGGGTCACTCCAG**GATT**
VEGFA – *Nme*2Cas9	GTGTGTCCCTCTCCCCACCCGTCCCTGT**CC**

Supplementary table 4: Primers for T7E-assays and TIDE sequencing.

Locus	Direction	Primer sequence (5′ to 3′)
AAVS1	Forward	TGCTTTCTTTGCCTGGACAC
	Reverse	CCTCTCTGGCTCCATCGTAA
EMX1	Forward	GGAGCAGCTGGTCAGAGGGG
	Reverse	GGGAAGGGGGACACTGGGGA
F8	Forward	GGGAGAGAACCTCTAACAGAACG
	Reverse	GCTCCAGGTGATGGATCATCAG
FANCJ	Forward	GTTGGGGGCTCTAAGTTATGTAT
	Reverse	CTTCATCTGTATCTTCAGGATCA
GAPDH	Forward	TAAAAAGTGCAGGGTCTGGCG
	Reverse	CTAACCAGTCAGCGTCAGAGC
IL2RG	Forward	ATGACACTGGTGGGTGTTCAG
	Reverse	TCTTCACCTTGCAGGCTCTCT
SLC9A9	Forward	GCACTTATTCTGGCCCCTGACTGC
	Reverse	GAGAACCATGGTCTGGGGAAGAAGACC
SLC9A9, off-target	Forward	AGGCCTGGGCTTTATCCA
	Reverse	AGCAGTAGTTCTCAAACTATGT
VEGFA – *Nme*1Cas9	Forward	GTGTGCAGACGGCAGTCACTAG
	Reverse	CTCTGCGGACGCTCAGTGAAG
VEGFA – *Nme*2Cas9	Forward	ATCAAATTCCAGCACCGAGCGC
	Reverse	AGAACTCAGGACCAACTTATTCTG

Supplementary table 5: PDB IDs of protein structures shown and used in this study.

Structure	PDB-ID	Source
AcrIIC3, *Nme*Cas9	6JE9	(Sun *et al*, 2019)
AraC, apo-form	2ARA	(Soisson *et al*, 1997)
AraC, complexed with L-arabinose	2ARC	(Soisson *et al*, 1997)
AraC, DBD	2K9S	(Rodgers & Schleif, 2009)
*As*LOV2 domain	2V0W, 2V0U	(Halavaty & Moffat, 2007)

ERD	1A52	(Tanenbaum *et al*, 1998)
eYFP, F165G	6ZQO	(Pletneva *et al*, 2021)
Flp recombinase	1FLO	(Chen *et al*, 2000)
*Nme*HNH, AcrIIC3	6J9N	(Zhu *et al*, 2019)
PDZ	1Z86	(Yan *et al*, 2005)
Rob transcription factor	1D5Y	(Kwon *et al*, 2000)
TVMV protease	3MMG	(Sun *et al*, 2010)
uniRapR	1FAP	(Choi *et al*, 1996)

6.4 List of Figures

6.5 List of tables

6.6 Abbreviations

Units

Å	angstrom
g	gravity of earth
h	hour
kDa	kilodalton
l	liter
m	meter
M	molar
mg	milligramme
min	minute
ml	milliliter
mM	millimolar
mol	mole
nm	nenometer
s	second
V	volt
W	watt
λ	wavelength
µl	microliter
µm	micrometre

Amino acids

A	Alanine	L	Leucine
R	Arginine	K	Lysine
N	Asparagine	M	Methionine
D	Aspartic acid	F	Phenylalanine
C	Cysteine	P	Proline
E	Glutamic acid	S	Serine
Q	Glutamine	T	Threonine

G	Glycine	W	Tryptophan
H	Histidine	Y	Tyrosine
I	Isoleucine	V	Valine

Nucleobases

A	Adenine	N	Random base
C	Cytosine	U	Uracil
G	Guanine	R	Purine
T	Thymine		

Other Abbreviations

AA	Amino acid
AAV	Adeno-associated virus
Acr	Anti-CRISPR protein
AcrIIC1 chim.	AcrIIC1 chimera
AcrIIC1X	Engineered AcrIIC1 (N3F/D15Q/A48I)
AcrIIC1X*	Engineered AcrIIC1 (N3F/D15Q/A48I), GSG-mCherry-GSG insertion behind Y70
AF2	AlphaFold2
AMP	Adenosine monophosphate
ANOVA	Analysis of variance
AP	Average precision
Ara	Arabinose
ASA	Accessible surface area
*As*LOV2, LOV2	LOV2 domain of Phototropin 1 from *Avena sativa*
ATP	Adenosine triphosphate
AUC	Area under the curve
AUROC	Area under the receiving operator characteristic
B. subtilis	*Bacillus subtilis*
BAC	Bacterial adenylate cyclases
bp	Base pair
CAP	Catabolite activator protein
Cas	CRISPR associated
CASANOVA	CRISPR–Cas9 activity switching via a novel optogenetic variant of AcrIIA4
CASP	Critical Assessment of Techniques for Protein Structure Prediction
c-di-GMP	Cyclic dimeric guanosine monophosphate
CMV	Cytomegalovirus
CN-C3	CASANOVA-C3, AcrIIC3 with *As*LOV2 insertion behind F59
CN-C3G	CASANOVA-C3G, AcrIIC3 with *As*LOV2 insertion behind F59, flanked by glycine linkers
CRISPR	Clustered regularly interspaced short palindromic repeats
crRNA	CRISPR RNA
DBD	DNA-binding domain
dCas9	Catalytically dead Cas9
DHFR	Dihydrofolatereductase
DNA	Deoxyribonucleic acid

DSB	Double-strand break
E. coli	*Escherichia coli*
ERD	Estrogen receptor-α domain
eYFP	Enhanced yellow fluorescent protein
FACS	Fluorescnece activated cell sorting
FKBP	FK506 binding protein
FMN	Flavine mononucleotide
FRB	FKBP–rapamycin binding
FRT	Flp recognition target
HDR	Homology-directed repair
HEK293T	Human embryonic kidney 293 cells expressing a mutated SV40 large T antigen
HTH	Helix-turn-helix
indel	Insertion/deletion
IPTG	Isopropyl-β-D-thiogalactopyranosid
IQR	Interquartile range
kb	Kilo base
KRAB	Krüppel associated box
LBD	Ligand-binding domain
LOV2	Light, oxygen, voltage
ML	Machine learning
MBP	Maltose-binding protein
MD	Molecular dynamics
miRNA, miR	microRNA
MOI	Multiplicity of infection
MSA	Multiple sequence alignment
nCas9	Cas9 nickase
NGS	Next generation sequencing
NHEJ	Non-homologous end-joining
*Nme*Cas9	Cas9 orthologue from *Neisseria meningitidis*
NLS	Nuclear localization signal
NMR	Nuclear magnetic resonance
ORF	Open reading frame
PAM	Protospacer adjacent motif
PAS	Per-ARNT-Sim; period circadian protein–aryl hydro-carbon receptor nuclear translocator protein–single-minded protein
PCR	Polymerase Chain Reaction
PDB	Protein data bank
PDZ	post synaptic density protein (PSD95), Drosophila disc large tumor suppressor (Dlg1), and zonula occludens-1 protein (zo-1) (Kennedy, 1995)
POI	Protein of interest
PR	Precision recall
REACH	Rational engineering of allostery at conserved hotspots
RFP	Red fluorescent protein
RMSD	Root-mean-square deviation of atomic positions
RNA	Ribonucleic acid
ROC	Receiving operator characteristic

rpm	Rounds per minute
S. aureus	*Staphylococcus aureus*
S. pyogenes	*Streptococcus pyogenes*
*Sau*Cas9	Cas9 orthologue from *Staphylococcus aureus*
SCA	Statistical coupling analysis
scAAV	Self-complementary adeno-associated virus
SD	Standard deviation
sgRNA	Single guide RNA
SigF	Sigma factor F from *Bacillus subtilis*
SPELL	Split Protein Reassembly by Ligands or Light
SPINE	Saturated Programmable Insertion Engineering
SV40	Simian vacuolating virus 40
T7E assay	T7 endonuclease assay
TAE	Tris-acetate-EDTA
TALE	Transcription activator-like effector
TBE	Tris-borate-EDTA
TIDE	Tracking of indels by decomposition
tracrRNA	Trans-activating RNA
TVMV	Tobacco vein mottling virus
UTR	Untranslated region
VVD	vivid
WT	wildtype
ZFN	Zinc finger nuclease