

# Dissecting the Determinants of Domain Insertion Tolerance and Allostery in Proteins

Jan Mathony,\* Sabine Aschenbrenner, Philipp Becker, and Dominik Niopek\*

**Domain insertion engineering is a promising approach to recombine the functions of evolutionarily unrelated proteins. Insertion of light-switchable receptor domains into a selected effector protein, for instance, can yield allosteric effectors with light-dependent activity. However, the parameters that determine domain insertion tolerance and allostery are poorly understood. Here, an unbiased screen is used to systematically assess the domain insertion permissibility of several evolutionary unrelated proteins. Training machine learning models on the resulting data allow to dissect features informative for domain insertion tolerance and revealed sequence conservation statistics as the strongest indicators of suitable insertion sites. Finally, extending the experimental pipeline toward the identification of switchable hybrids results in opto-chemogenetic derivatives of the transcription factor AraC that function as single-protein Boolean logic gates. The study reveals determinants of domain insertion tolerance and yielded multimodally switchable proteins with unique functional properties.**

stable protein folds and their corresponding functions, while enabling evolutionary innovation by exploring novel combinations and interdependencies thereof.<sup>[1,2]</sup> This observation has inspired protein engineering approaches that combine evolutionary unrelated protein domains into single polypeptide chains, thereby creating hybrid proteins with new-to-nature properties.<sup>[3–7]</sup> From a synthetic biology perspective, a particularly interesting strategy is the insertion of receptor domains into effector proteins with the aim to allosterically couple the effector conformation to the receptor state.<sup>[3,5,8]</sup> Receptor activation, e.g. via chemicals or light, will induce an allosteric signal relaying to the effector's active site (e.g., a catalytic surface or binding site), thereby enabling highly targeted control of the effector-mediated cellular function.

Although a number of hybrid proteins have been created by domain insertion

## 1. Introduction

The recombination of protein domains is an important driver of evolution. It allows nature to repeatedly build on the same set of

engineering over the past years, their rational design remains challenging and screening of larger libraries and iterative optimization is commonly required to obtain functional hybrids.<sup>[9–13]</sup> Importantly, the identification of an insertion site at which the fusion of two protein domains results in their functional coupling and does not irreversibly interfere with the activity of either protein part represents a largely unsolved problem. These persisting challenges can be explained by our limited understanding of the structural and biophysical requirements and constraints that generally determine suitable domain insertion sites.


Advances in the generation of comprehensive domain insertion libraries via transposon<sup>[12,14]</sup> or oligonucleotide pool-based cloning,<sup>[15]</sup> as well as the coupling of fluorescence-activated cell sorting (FACS) to deep sequencing, facilitate the efficient generation and subsequent investigation of larger domain insertion datasets.<sup>[11,12,16]</sup> Employing such experimental approaches, recent studies investigated the impact of domain insertion on the membrane localization of potassium ion channels.<sup>[16,17]</sup> Using the resulting data to train random forest models, the authors analyzed biophysical properties that contribute to domain insertion permissibility in ion channels.<sup>[17]</sup> This previous research was centered around a single type of membrane protein as well as the impact of domain insertion on subcellular protein localization. To render domain insertion engineering a broadly-applicable strategy, however, studying the domain insertion tolerance at the functional level as well as deciphering the determinants of functional coupling between re-combined protein domains will be essential.

J. Mathony, P. Becker  
Center for Synthetic Biology  
Technical University of Darmstadt  
64287 Darmstadt, Germany  
E-mail: jan.mathony@uni-heidelberg.de

J. Mathony, P. Becker  
Department of Biology  
Technical University of Darmstadt  
64287 Darmstadt, Germany

P. Becker  
Department of Biotechnology and Biomedicine  
Technical University of Denmark  
Kongens Lyngby 2800, Denmark

J. Mathony, S. Aschenbrenner, D. Niopek  
Institute of Pharmacy and Molecular Biotechnology (IPMB)  
Faculty of Engineering Sciences  
Heidelberg University  
69120 Heidelberg, Germany  
E-mail: dominik.niopek@uni-heidelberg.de

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/advs.202303496>

© 2023 The Authors. Advanced Science published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

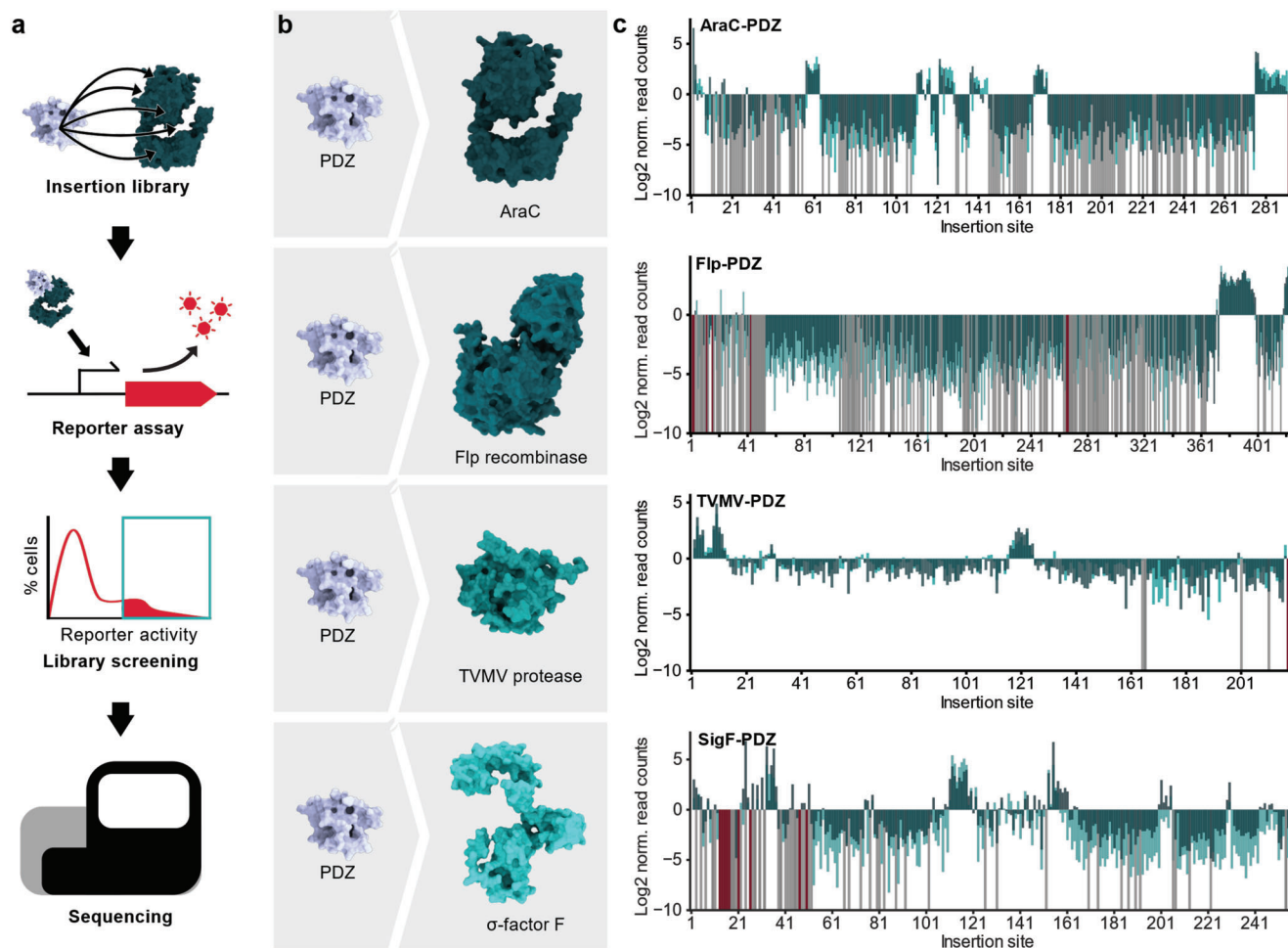
DOI: 10.1002/advs.202303496

Here, we set out to broaden the understanding of domain insertion requirements in diverse protein classes. Toward this goal, we inserted up to five structurally and functionally unrelated domains into several different, unrelated candidate effector proteins covering nearly all possible sequence positions. Using gene circuits that relay effector activity to a fluorescent readout, the resulting, comprehensive libraries of protein hybrids were screened for active variants by FACS and subsequent next-generation sequencing (NGS). Training of machine learning models on the resulting datasets allowed us to dissect parameters that affect domain insertion tolerance and revealed sequence conservation statistics as the most powerful predictors for domain insertion success. Finally, extending our experimental pipeline toward the screening of engineered, switchable effector variants yielded two potent optogenetic derivatives of the *E. coli* transcription factor AraC that function as single-protein chemo-optogenetic Boolean logic gates.

## 2. Results

### 2.1. A functional FACS-NGS Screen of Domain Insertion Tolerance

To elucidate the domain insertion tolerance within an evolutionarily and functionally diverse set of effector proteins, we first constructed comprehensive insertion libraries. The libraries comprised of effector proteins carrying insert domains at all possible sequence positions (Figure 1A). Four structurally unrelated proteins that are widely applied in synthetic and cell biology were chosen as effector protein scaffolds: the transcription factor AraC, the recombinase Flp, a previously described variant of the TMV protease,<sup>[18]</sup> and  $\sigma$ -factor F (SigF) from *Bacillus subtilis* (Figure 1B). Protein hybrid libraries were generated via saturated programmable insertion engineering (SPINE) for all four candidates using the PDZ domain from murine  $\alpha$ 1-syntrophin



**Figure 1.** Domain insertion profiling of functionally and structurally diverse proteins. A) Flow chart of the domain insertion screening workflow. B) Overview of the screened PDZ-domain insertion libraries. The depicted structures of the parent proteins are AF2 predictions. PDB-ID of PDZ: 1Z86. C) Enrichment score histograms for the different candidate proteins are shown. The Log<sub>2</sub> norm. read counts correspond to the fraction of reads after enrichment normalized to the fraction of read counts within the initial library. Data from the four candidate proteins AraC, Flp, TMV protease, and SigF with PDZ domain inserts are shown. Enrichments are mapped to the respective insertion site as indicated by the position of the acceptor proteins preceding the insertion. Light green, dark green: individual replicates. Grey: variants with zero reads after enrichment. Red: variants missing in the initial library. Insertion sites correspond to residues preceding the inserted domain.

as insert (Figure 1B).<sup>[15]</sup> With its small size of 86 amino acids, its globular fold and the N- and C-terminus located in close proximity ( $\sim 10$  Å), the PDZ domain is ideally suited for domain insertion screening (Table S1, Supporting Information).<sup>[11]</sup> Further, to elucidate how the domain identity would affect the functionality of the resulting protein hybrids, four additional insert domains of varying size and structure (see Table S1, Supporting Information for details) were selected and fused at all possible sequence positions into one of the candidate proteins, AraC. These included the AsLOV2 (*Avena sativa*) domain, the estradiol binding domain from human estrogen receptor- $\alpha$  (ERD), an enhanced yellow fluorescent protein (eYFP)<sup>[19]</sup> and the synthetic rapamycin receptor uniRapR.<sup>[20]</sup> Following the construction of all eight libraries, a nearly complete coverage of all possible insertion sites was observed by deep sequencing (Figure S1, Supporting Information).

To enable functional screening of these libraries in *Escherichia coli*, we next created reporter gene circuits that robustly couple the activity of the effector protein to the expression or stability of a red fluorescent protein (RFP) (Figure S2A, Supporting Information, Methods). We then co-transformed *E. coli* Top10 cells with the reporters and their corresponding effector-insert hybrid libraries, followed by an analysis of the reporter activity via FACS. Fluorescence histograms of the initial libraries showed a large fraction of non-functional hybrid protein candidates as indicated by a large proportion of non- or low fluorescent cells (Figure S2B, Supporting Information). Still, a small but considerable fraction corresponding to medium to high fluorescent cells and hence active protein hybrids was observed. Sorting this fraction resulted in a clear enrichment of cells expressing high RFP levels in the case of AraC and SigF and less pronounced, but still visible enrichments in fluorescent cells for Flp and the TVMV protease (Figure S2C, Supporting Information). Quantitative differences between the four effector library pools were caused by varying proportions of active versus inactive hybrid protein candidates in the initial libraries as well as differences in the dynamic range of the reporter assays (Figure S2, Supporting Information, controls). To ensure a significant enrichment of active variants, we sorted each library in two consecutive rounds. Next, we assessed enrichment or depletion of each individual domain insertion variant in the sorted libraries by adapting the previously published DIP-seq pipeline.<sup>[12]</sup> In short, the fraction of read counts corresponding to a variant after enrichment was normalized by the fraction of read counts from the initial library and the resulting scores were log<sub>2</sub>-scaled. Variants that went extinct during sorting and thus had a read count of zero were assigned a log<sub>2</sub> value of  $-10$ , since this represents the assay's detection limit. To ensure the reproducibility of the workflow, the whole screening and sequencing process was performed in two independent replicates.

Results from different replicates correlated well, with a Pearson correlation coefficient (Pearson's  $r$ )  $> 0.8$  in all cases except one (Pearson's  $r$  for TVMV-PDZ = 0.65), while the level of enrichment/depletion differed between replicates for individual variants (Figure S3, Supporting Information). As cross-validation of our enrichment and analysis pipeline, we experimentally measured the activity (RFP expression) for a set of hybrids individually and compared it to the variant enrichment scores obtained by NGS. As expected, a drastic difference in activity between the enriched and the depleted variants was measured in most cases

(Figure S4, Supporting Information). For the following analysis, the mean of the two biological replicates was used.

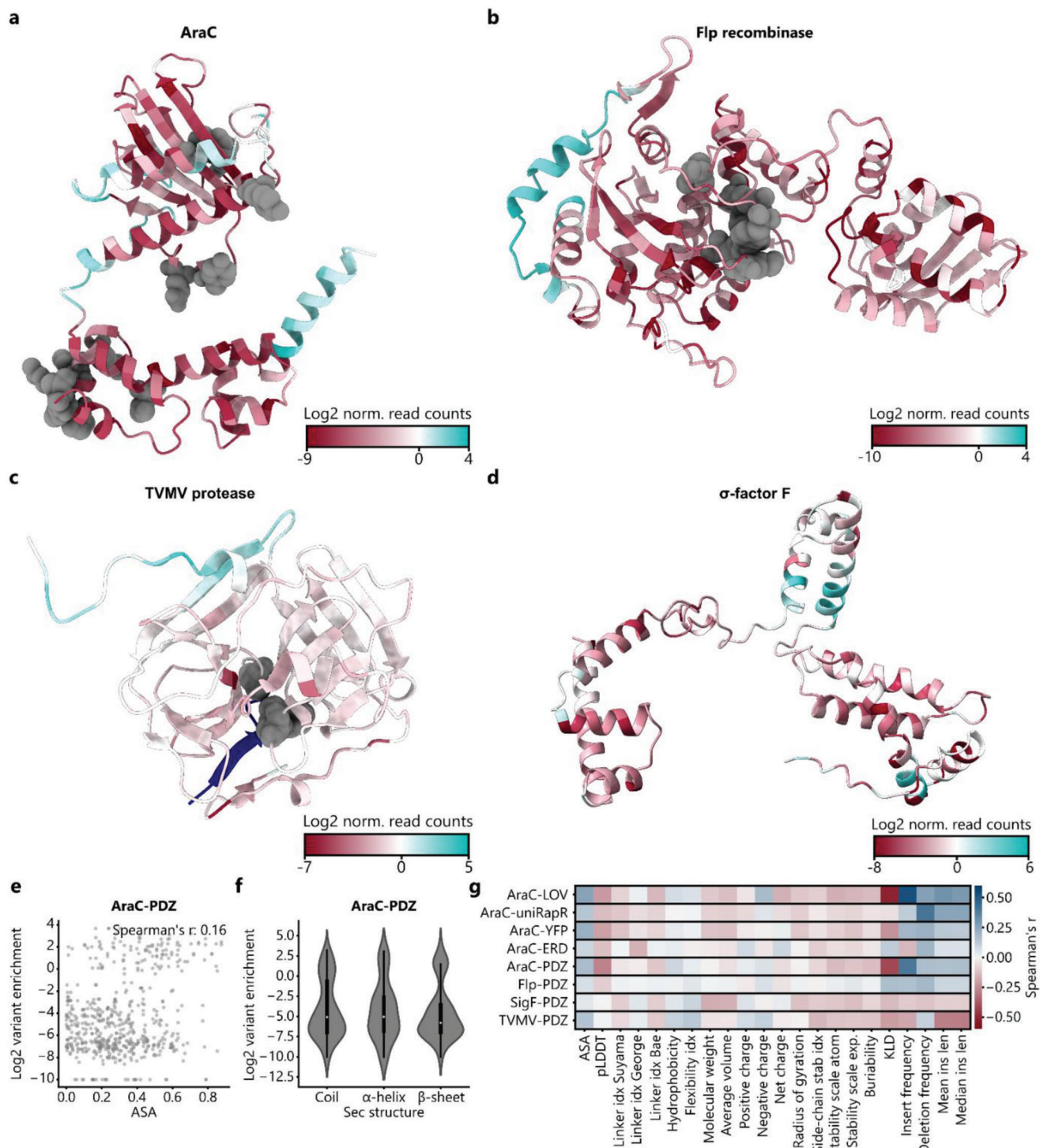
## 2.2. Domain Insertion Permissibility is Sequentially and Structurally Clustered

Mapping the enrichment scores of the PDZ insertion libraries to the amino acid sequences of the respective, four effector proteins revealed that positions tolerating insertions occurred in clusters spanning regions of  $\approx 10$ – $30$  consecutive amino acids (Figure 1C). Insertion tolerance thus appears to be regionally confined, rather than being determined by features of individual residues or positions. Roughly 80 % of the insertions within each protein were depleted, i.e., they do not tolerate domain fusion (Figure 1C).

Moreover, the number of clusters with enrichments differed substantially between the insert domains tested in combination with the AraC effector (Figure S5, Supporting Information). For the LOV2 insert domain, we observed several insertion-permissive regions throughout the sequence of AraC comparable to those for the PDZ insert. In contrast, the other three insert domains were enriched at substantially fewer positions, mainly at the C-terminus of AraC. As LOV2 and PDZ are considerably smaller ( $< 150$  AA) than the other tested domains, insert size appears to be a determining factor for insertion tolerance. In addition, the relative distance of the PDZ- and LOV2 domain's termini (14.1 Å and 20.7 Å, respectively; note this is the distance as measured from the terminal residues in the structures from Table S4, Supporting Information) are smaller as compared to the other insert domains, although uniRapR exhibits an only marginally larger distance between N- and C-termini (24.4 Å) (Table S1, Supporting Information). Interestingly, we hardly observed insertion sites selective for just one specific insert domain. This indicates that domain insertion permissibility is a general property of protein regions rather than a lock-key relation between an insertion site and an individual insert domain.

Next, we mapped the enrichment scores onto structures of the respective effector proteins. To this end, we used AlphaFold2 (AF2)-predicted protein structures, as well as experimentally resolved full length structures if available<sup>[21,22]</sup> (Figure 2A–D; Figures S6–S8, Supporting Information). Importantly, the predicted structures were generally in excellent agreement with the available experimentally validated (partial) folds (Figure S9, Supporting Information). Structural analysis revealed strong depletions around functionally critical regions, such as the DNA- and arabinose-binding sites of AraC, the catalytic center of the Flp recombinase, or the DNA-binding region of SigF (Figure 2A–D; Figures S6 and S7, Supporting Information). For TVMV protease, depletions within the hydrophobic core and around the active site were observed, albeit trends were overall less pronounced for this candidate protein (Figure 2C; Figure S6C, Supporting Information). Interestingly and in contrast to common assumptions underlying domain insertion engineering strategies, no clear enrichment at surface-exposed unstructured loops could be identified for any of the candidates. Rather, insert sites were observed at similar frequency in helices, sheets, and loops (Figure 2A–D).





**Figure 2.** Secondary structure and amino acid features alone do not explain the experimentally observed domain insertion patterns. A) Domain insertion permissive positions are clustered at diverse, locally confined surface sites. The insertion scores from the PDZ libraries are mapped onto the AF2 structure predictions of the candidate proteins namely AraC A) and Flp recombinase B) the crystal structure of the TVMV protease (PDB-ID: 3MMG) C) and an AF2 structure prediction of SigF D). Functionally critical residues of AraC, Flp, and the TVMV protease are indicated in grey. E) Correlation between variant enrichment and the average surface exposed area (ASA) of the residues neighboring an insertion site are plotted for AraC-PDZ. Spearman's  $r$  is indicated. F) Violin plot of 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 to and after the insertion was 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. G) Spearman correlations between all datasets and diverse positional features are shown (Table S2, Supporting Information). Linker idx: Different amino acid specific linker propensity indices that were reported by the indicated authors.

Next, to quantitatively analyze these qualitative observations, we correlated the measured enrichments with a set of basic positional properties such as the average solvent accessible area (ASA), secondary structure, and amino acid identity of the residues neighboring a respective insertion site (Figure 2E,F; Figures S10 and S11, Supporting Information). Of note, none of these basic properties explained the observed enrichments. In order to obtain a more comprehensive overview of protein features that could affect domain insertion success, a larger set of position-specific features was gathered (Table S2, Supporting Information, Methods). Further, these comprised a number of biophysical amino acid properties, fetched from the "AAindex" database,<sup>[23,24]</sup> as well as several previously published linker propensity indices.<sup>[25–27]</sup> These indices describe to which extent amino acids tend to be present in inter-domain linkers. Regions with high linker propensities are commonly expected to be well suited for the insertion of domains. Further, we included the pLDDT confidence score from AF2 models, which was previously shown to correlate with intrinsically disordered sites.<sup>[28]</sup> Moreover, the Kullback-Leibler divergence (KLD), a measure for sequence conservation, was extracted from multiple sequence alignments of the candidate protein with natural homologs. Finally, additional scores, such as the frequency of insertions and deletions at every position in evolutionary related sequences, were included (refer to Methods). Spearman correlations between all enrichment scores for the screened libraries and each feature revealed overall weak trends, with the majority of the correlation coefficients lying in the range between  $-0.2$  and  $0.2$  (Figure 2G; Figure S12, Supporting Information). This observation is in agreement with previous results in the context of ion channels.<sup>[16,17]</sup> Additionally, we confirmed that AF2-based structure predictions of insertion variants could not explain the observed enrichment trends (Note S1 and Figures S13 and S14, Supporting Information).

### 2.3. Machine Learning Reveals Statistical Features Predicting Domain Insertion Tolerance

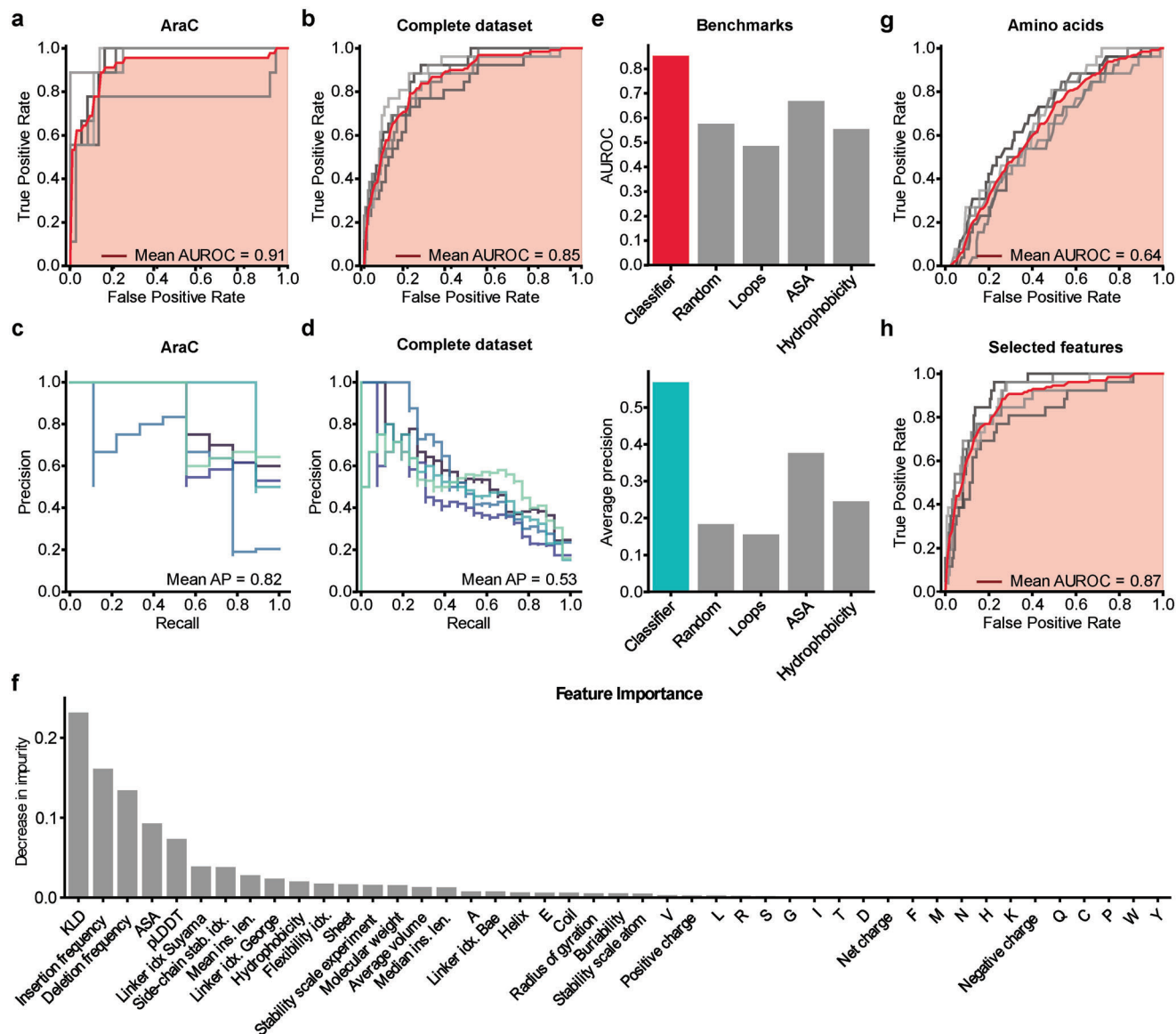
The absence of any clear correlation between the experimental data and positional protein properties raised the question if a combination of the above features would enable the prediction of domain insertion tolerance. To address this question, machine learning models were trained on the entirety of the gathered insertion site properties in combination with amino acid identity and secondary structure information as additional features. The learning objective was to discriminate between enriched sites that tolerated the insertion of a domain versus depleted positions, as these states appeared to be well separated in the data (Figure 1C). As model architecture, we chose a gradient boosting classifier,<sup>[29]</sup> i.e., an algorithm that additively combines multiple simpler machine learning models (in this case basic regression trees) by minimizing a loss functions. Such algorithms are known to perform particularly well on tabular datasets. The model was trained for each protein using five-fold cross-validation. We assessed the classifier's performance on each cross-validation test set, using standard metrics including the area under the receiving operator characteristic (AUROC) and average precisions (AP) (refer to the experimental section

for details). The models reached surprisingly good performances on datasets derived from individual candidate proteins ranging from a mean AUROC of 0.72 for SigF-PDZ to 0.92 for Flp-PDZ (Figure 3A,B; Figure S15, Supporting Information). The corresponding AP ranged from 0.41 (SigF-PDZ) to 0.82 (AraC-PDZ) (Figure 3A,B; Figure S15, Supporting Information). The lower AP values are caused by the high proportion of negative labels in the respective datasets. Encouraged by these results, we optimized the model on a complete training set including all four proteins, which resulted in a mean AUROC of 0.84 and a mean AP of 0.54 (Figure 3C,D). To place the classifiers performance into context, we compared it to several benchmarks on a previously withheld test set. These included a random choice baseline, and the use of individual features as predictors. Our classifier exhibited highly improved predictive power as compared to all individual features, reaching an AUROC of 0.85 and an AP of 0.56, suggesting that the entirety of input features implicitly provided the information necessary for successful prediction of domain insertion tolerance (Figure 3E; Figure S16, Supporting Information).

Finally, we aimed at identifying the key features most informative for the prediction of domain insertion tolerance. To this end, the influence of individual features on the model's performance was assessed by measuring the permutation importance of each feature as well as its Gini importance (Figure 3F; Figure S17A, Supporting Information).<sup>[30]</sup> The permutation importance measures the decrease of a model's accuracy upon random shuffling of the values for an individual feature. The Gini importance, in contrast, measures the average importance of regression tree nodes corresponding to a certain feature by calculating the respective gain in impurity. Both measures indicated that most parameters were dispensable, while the alignment-derived properties were most critical for successful prediction. In that line, a model trained solely on information about the identity of insertion-adjacent amino acids did reach an AUROC of 0.64 (Figure 3G). As a consequence, we depleted features from the input data in a stepwise manner, while ensuring the performance of the model did not decrease upon feature removal. Following this procedure, we were able to train a reduced model, only based on six features: KLD, deletion frequency, insertion frequency, mean insertion length, pLDDT, and the linker index by Suyama et al.<sup>[25]</sup> With an AUROC of 0.87 and an AP of 0.55, the reduced model performed as good as the original one trained on all features (Figure 3H). Lastly, the feature importance analysis was repeated with the reduced model. Akin to the previous observations, KLD, insertion frequency, and deletion frequency, i.e., evolutionary and statistical features derived from MSAs, were detected as most important parameters explaining domain insertion tolerance (Figure S17B,C, Supporting Information).

### 2.4. Identification of Potent Light-Switchable AraC Variants

Up to this point, we focused on features determining the preservation of function upon domain fusion into an effector protein. Taking our experimental screening approach one step further, we next investigated to which extent insertions can mediate allosteric behavior, i.e., a functional link between an insert and the

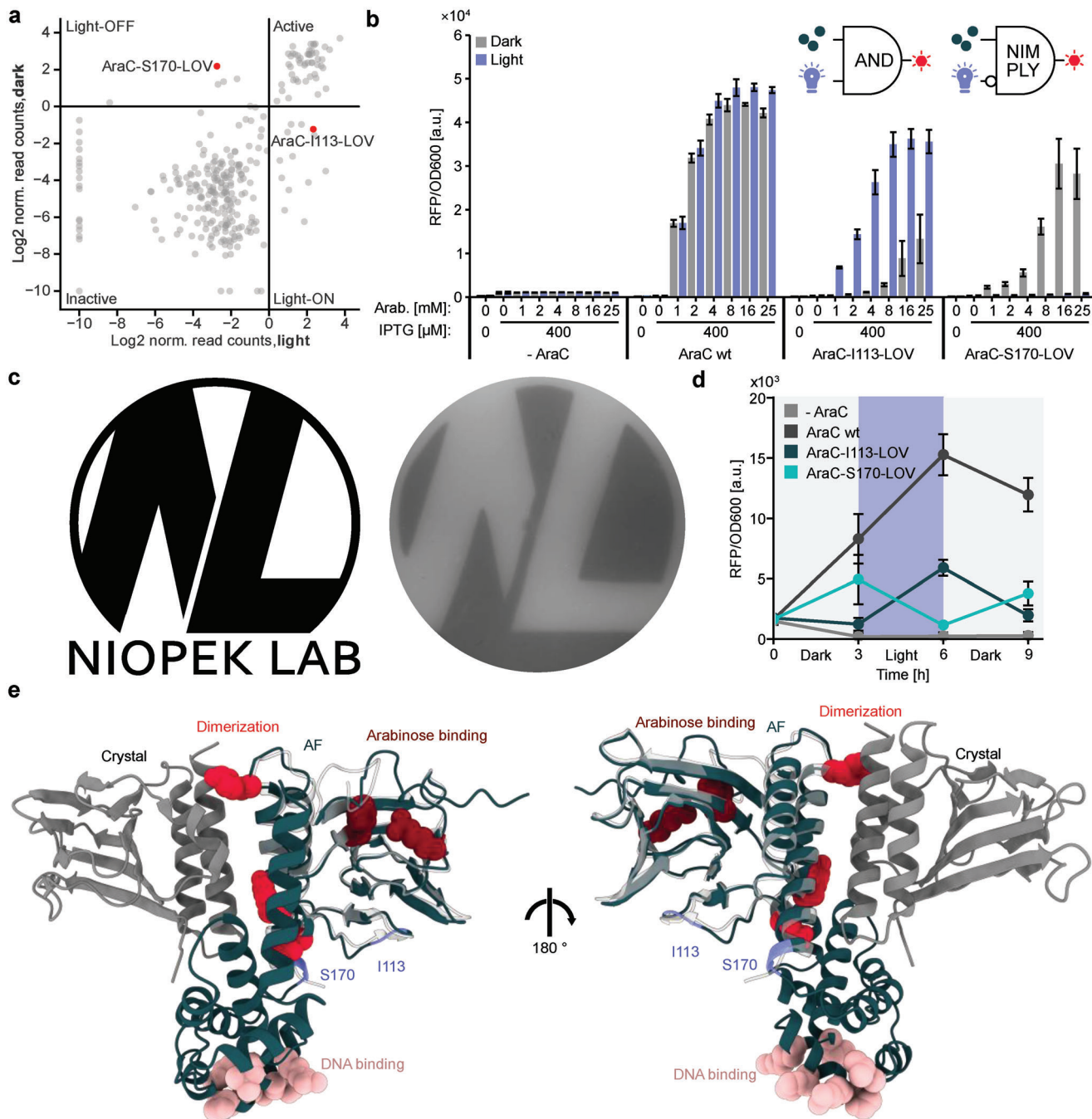


**Figure 3.** Gradient boosting classifier models reveal parameters informative of domain insertion tolerance. A,B) ROC curves of the model trained with fivefold cross-validation on the AraC-PDZ dataset A) or the combined PDZ datasets of all candidate proteins B). Results from individual cross-validations are shown in grey and the mean ROC is depicted in red (see Experimental Section 4.13 for details on the used metrics). C,D) Precision-recall metrics for individual cross-validation folds are shown. The mean average precision (Mean AP) is indicated. E), The AUROC and average precision of the trained classifier and different benchmarks are shown. The metrics were assessed on a previously withheld test set. F) Bar plot indicating the Gini importance (i.e., mean decrease in impurity) of each feature for the model trained on the full dataset. G) The ROC metric of a gradient boosting model that was trained exclusively on the amino acid identities is shown. H) ROC of a model that was trained on a subset of features comprised of Deletion frequency, KLD, insert frequency, mean insertion length, the linker propensity index by Suyama<sup>[25]</sup> and the pLDDT score from AF2 structure predictions. A,B,G,H) The ROC is depicted for individual folds in grey and the mean ROC in red. The mean AUC is marked in light red. Precise values are indicated.

effector. Such switchable hybrids are of great interest for various applications in biology and bioengineering. Toward this goal, we re-visited our initial AraC-LOV hybrid library. The AsLOV2 domain is known to reversibly unfold its two terminal helices in response to blue light ( $\approx 450$  nm), a property that has been harnessed for the development of light-switchable effector proteins in optogenetics.<sup>[3,31]</sup> It was hence interesting to explore, whether screening our comprehensive AraC-LOV library could readily reveal potent, optogenetic AraC variants.

We, therefore, repeated the screen for the AraC-LOV library, this time incubating the cultures under blue-light exposure prior to FACS sorting. The resulting variant enrichment was then compared to that of the same library sorted upon incubation of cultures in the absence of light (Figure S5, Supporting Information). Globally, we observed a high similarity between the resulting enrichment scores for each position under both conditions (Figure 4A; Figure S18A, Supporting Information). However, a subset of regions showed significant differences between





**Figure 4.** LOV2 domain insertion screening yields chemo-optogenetic AND and NIMPLY gates. A) Scatterplot showing the relation between the enrichment scores of individual variants for the libraries incubated in the light and dark. B) Characterization of light-responsive AraC variants. Inducers were supplied in the indicated concentrations. The samples were incubated under light exposure or in darkness, followed by measurements of reporter fluorescence (RFP) and OD600. Bars represent means from three independent replicates. Error bars show the SD. The corresponding logic gates are indicated. C) Agar photograph generated via an AraC-S170-LOV2 controlled RFP reporter. Top agar mixed with inducers and bacteria carrying an RFP reporter plasmid and the AraC-S170-LOV2 variants were plated on an agar 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). D) Cultures were inoculated into media carrying 400 μM IPTG and 25 mM arabinose. The samples 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. Points represent the mean of  $n = 3$  biological replicates. Error bars indicate the SD. E) An AF2 prediction of the full length AraC (green) is shown alongside the crystal structure (grey and white) of the arabinose binding domain. The relative positioning of the structures was obtained by superimposing the AF2 model onto a dimer crystal structure. Insertion sites and key residues are highlighted and their function is indicated. PDB-ID: 2ARA.

the enrichment scores obtained for the libraries cultured in the dark and light (Figure S18B,C, Supporting Information). Strikingly, further analyzing the insertion variants in these regions revealed a plethora of presumably light-activatable as well as light-inhibited AraC-LOV hybrids corresponding to multiple different AraC insertion sites (Figure S18B,C, Supporting Information).

From this set of optogenetic variants, we chose two AraC-LOV hybrids for further characterization, one light-ON switch carrying the LOV2 insertion behind I113 (AraC-I113-LOV) and a light-OFF switch with the LOV2 insertion behind S170 (AraC-S170-LOV). We then assessed the performance of these AraC-LOV hybrids using the previously established RFP transcription reporter in *E. coli* under varying arabinose concentrations, as well as light conditions. Interestingly, the activity of both AraC-LOV hybrids was co-dependent on the arabinose concentration and the light stimulus (Figure 4B; Figure S18D, Supporting Information). The AraC-I113-LOV samples showed a 23-fold increase in reporter expression upon illumination at an arabinose concentration of 4 mM. At higher arabinose concentrations, increasing fluorescence levels were also observed for samples incubated in the dark, indicating that the chemical inducer could, to some extent, override the light-mediated regulation. Vice versa, the AraC-S170-LOV samples showed efficient, light-dependent repression of reporter activity practically to baseline with a 43-fold switch in reporter activity at 16 mM arabinose. Moreover, the light-regulation was in this case not affected by high arabinose concentrations. Comparing the overall activation of the AraC variants in response to arabinose, the activity of the wildtype saturates already at an inducer concentration of 4 mM, while the LOV2-hybrids, in particular AraC-S170-LOV, require higher arabinose concentrations (up to 16 mM) to trigger maximum reporter activity. This suggests that LOV2 insertion weakened the sensitivity of AraC to arabinose. The observed behavior establishes the AraC-I113-LOV and AraC-S170-LOV hybrids as single-protein Boolean logic devices capable of integrating light and arabinose as inputs and functioning as AND and NIMPLY gates, respectively (Figure 4B; Note S2, Supporting Information).

Next, we investigated if these new optogenetic AraC variants facilitate spatiotemporal control of gene expression. Growing the AraC-S170-LOV reporter strain on agar while illuminating it through a photo-mask confined reporter RFP expression to light-shielded regions and hence resulted in display of the photomask shape on the fluorescent cell layer (Figure 4C). Moreover, incubating AraC-S170-LOV and AraC-I113-LOV reporter strain cultures while alternating between light and dark conditions resulted in reporter expression oscillation, the phase of which depended on the AraC-LOV variant used (Figure 4D). Taken together, the results showcase the versatility of this new chemoptogenetic toolkit with respect to spatiotemporal control of gene expression in *E. coli*.

On a structural level it is striking that most insertion sites resulting in switchable AraC behavior are located within the region between the ligand-binding domain (LBD) and the DNA-binding domain (LBD) of AraC (Figure 4E). This trend can be explained by the functional role this region has, serving as a dimerization interface upon AraC activation and by mediating the relative flexibility of both domains.<sup>[32,33]</sup> It is thus no surprise that LOV2 domain insertions in this area can influence AraC function. Of note, AF2 structure predictions of AraC-I113-

LOV and AraC-S170-LOV capture the former variant in a more compact conformation, which is in agreement with the less flexible repressor state of wildtype AraC<sup>[32,33]</sup> (note: AF2 predicts the LOV2 structure in its dark-adapted state) (Figure S19, Supporting Information). AraC-S170-LOV, in turn, was predicted to have a more relaxed conformation, as would be expected for an active AraC (Figure S19, Supporting Information). To further investigate the robustness of allosteric coupling in both hybrid proteins, we screened a set of point mutants for their effects on wildtype AraC and its engineered derivatives (Figure S20, Supporting Information). The majority of mutations did not improve the AraC-I113-LOV switch, but rather reduced reporter activity, in the active (light) state or increased leakiness, i.e., reporter activity in the dark. Excitingly, several AraC-S170-LOV point mutants (e.g., T50S, G141D, and V284F; mutations correspond to residues in wildtype AraC) showed an increased level of activity in the dark as compared to the initial variant, while likewise retaining potent reporter repression upon illumination. The mutations E3I and T241C, in turn, permanently impaired the function of the AraC-S170-LOV variant, while having no significant effect on AraC-I113-LOV. Finally, none of the tested mutations had major effects on the activity of wildtype AraC. Collectively, our data highlight (i) the variant-specificity of mutational effects in the engineered allosteric AraC-LOV hybrids and (ii) their increased functional and likely structural sensitivity toward minor sequence alterations. Moreover, the mutational data in conjunction with the arabinose-dependency data (Figure 4B) indicate the interconnection of the natural arabinose-mediated allosteric regulation with a LOV2-induced artificial allosteric pathway.

### 3. Conclusion

In this study, we investigated the constraints of domain insertion engineering at the functional and structural level. Thereby, we considerably extended the existing body of work toward new protein families and, for the first time, compared the insertion tolerance of several evolutionary unrelated proteins side-by-side directly using effector protein function as readout. In agreement with previous studies,<sup>[16,17]</sup> our data showcases the absence of any simplistic explanations for domain insertion permissibility. In contrast, we demonstrated that gradient boosting classifiers can help to decipher the importance of factors underlying domain insertion tolerance. Our models identified MSA-derived conservation statistics as main determinants of domain insertion tolerance, thus suggesting an evolutionarily informed approach to be particularly promising for domain insertion engineering (Figure 3). In this context, parallels can be drawn with statistical coupling analysis (SCA), a method for identifying co-evolving residues based on the statistical evaluation of MSAs.<sup>[5,13,34]</sup> The SCA-derived residue patterns termed “protein sectors” have been proposed to be functionally critical and well suited for identifying allosteric sites to engineer protein switches.<sup>[13]</sup> In contrast, our work underscores the indicative value of evolutionary insertion/deletion events.

We note that in context of domain insertions, the predictive power of machine learning models is still constrained by the amount of available training data, which is, in turn, restricted by the current experimental capacity limits. The use of experimental data, such as the presented insertion library screens, in



combination with larger datasets extracted from public protein sequence databases might provide an elegant solution to address this limitation in data size.

In addition, it will be interesting to see to what extent the observed trends are replicated across entirely unrelated protein classes, such as enzymes, which are particularly vulnerable even to minor structural changes in the active site or proteins the activity of which depends on complex domain motions. In these cases, the preservation of activity might rely on factors that cannot easily be inferred from conservation statistics.

With respect to allosteric proteins, the screening pipeline developed here was efficient in identifying allosteric switches (Figure 4; Figure S18, Supporting Information). In previous work, a GFP-maltose-binding protein insertion library was enriched alternatingly in the presence and absence of the input trigger in three consecutive rounds.<sup>[12]</sup> Our adaption of the method using parallel enrichment of the same library under different conditions (here culturing samples in the presence or absence of light) turned out to be sufficient to reliably identify light-switchable proteins. A more stringent selection regime during FACS could potentially render even a single round of enrichment sufficient, which would further simplify and streamline the workflow for the engineering of switchable effector proteins.

We note that several optogenetic bacterial expression systems exist.<sup>[35–37]</sup> These include the light-responsive AraC variant BLADE, which is based on the Vivid LOV domain from *Neurospora crassa* functioning via light-induced AraC dimerization.<sup>[35]</sup> In contrast to these previous examples, the transcription factors developed here are co-dependent on two stimuli, namely light and arabinose. This has interesting implications for synthetic biology applications and gene circuit control. Transcription factors co-dependent on two inputs enable the independent control of the state (on/off) and amplitude of activation for genetic programs. Previously, the combination of chemically inducible transcription factors and light-responsive regulators had to be combined within far more complex circuits to achieve the same goal.<sup>[37–39]</sup> The optogenetic variants presented here highly simplify such experimental setups by reducing the underlying system to a single protein component (see Note S2, Supporting Information). Such single-protein Boolean logic gates could considerably streamline the design and increase the robustness of complex genetic circuits and biocomputing programs by reducing the number of required components and through the direct integration of signals within a single molecule.

In summary, our study pinpoints determinants of domain insertion tolerance and showcases the power of unbiased domain insertion screens for the engineering allosteric effector proteins with applications in synthetic biology and beyond.

## 4. Experimental Section

**Molecular Cloning:** All constructs used in this study are listed in Table S3 (Supporting Information). The corresponding amino acid sequences of the encoded proteins are shown in Table S4 (Supporting Information). Plasmids were constructed using Golden Gate assembly.<sup>[40]</sup> 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 5'-overhangs, which enabled the scarless assembly of constructs. PCRs were performed according to the NEB stan-

dard protocols. For Golden Gate assembly, the procedure described by Engler et al. was followed.<sup>[40]</sup> DNA-oligonucleotides were ordered from Merck and Integrated DNA Technologies (IDT). Double-stranded DNA fragments were purchased at IDT. Point mutants were cloned by introducing the changes via mismatching primers upon amplification of the full plasmid and subsequent phosphorylation and ligation. 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, plated on agar, and incubated overnight at 37 °C. Liquid cultures were inoculated from single colonies and grown overnight at 37 °C while shaking at 220 rounds per minute (rpm). DNA was purified using the QIAamp DNA Mini kit (Qiagen). All constructs were sequence-verified using Sanger sequencing (Microsynth Seqlab and Genewiz). The plasmid pTKEI-Dest, which served as a backbone for the insertion libraries, was a gift from David Savage (Addgene plasmid # 79784).<sup>[12]</sup>

**Reporter Assays:** All reporter circuits used the monomeric red fluorescent protein 1 (RFP) as readout.<sup>[41]</sup> The design of the genetic circuits is depicted in Figure S2A (Supporting Information). In short, the AraC reporter was created by placing the RFP coding sequence under the 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 RFP open reading frame by the recombinase, RFP is expressed. To measure TVMV protease activity, a ssrA-like degradation tag<sup>[42]</sup> 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 degenon resulting in RFP stabilization and an increase in fluorescence. Many related potyvirus proteases undergo a process called autolysis,<sup>[43]</sup> during which the protease cleaves off its own C-terminal region albeit at low efficiency. This results in a truncated protease with decreased activity. To ensure that only one TVMV protein species would be present during all assays, a previously reported, truncated TVMV version<sup>[44]</sup> was used for insertion library generation. Finally, a reporter for SigF was constructed, based on a SigF-specific promoter design previously reported by Bervoets et al.<sup>[45]</sup>

**Domain Insertion Library Generation:** To generate insertion libraries covering all possible effector protein positions, saturated programmable insertion engineering (SPINE) was used.<sup>[15]</sup> In short, the protein of interest was subdivided into chunks of  $\approx 50$  amino acids. For each chunk, an oligonucleotide sub-pool (Agilent) was designed, comprising 50 individual DNA sequences, each of which carried a Type IIS restriction enzyme recognition site handles behind a specific amino acid encoding triplet. A python pipeline for the automatic design of the required DNA sequences provided by Coyote-Maestas et al.<sup>[15]</sup> was employed for oligo pool design. The sub-pools were then individually cloned into an expression vector carrying the full-length coding sequence of the respective effector protein of interest and transformed into chemically competent Oneshot Top10 *E. coli*. To ensure at least 40-fold coverage of the library, serial dilutions were plated on agar plates following transformation and the number of colony-forming units was calculated. The plasmid sub-libraries were purified from the bacteria using the QIAamp DNA Mini Preparation Kit (Qiagen). The DNA concentration was measured using the Quant-iT dsDNA (HS) assay kit (Thermo Fisher Scientific) and all sub-libraries for each individual effector protein were pooled using equal DNA 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. *E. coli* cells were transformed and plated on three 20 cm LB-agar plates, supplemented with  $50 \mu\text{g mL}^{-1}$  chloramphenicol and  $25 \mu\text{g mL}^{-1}$  of kanamycin (Carl-Roth). Again, a library coverage of at least 20x was ensured by serial dilutions and 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 insert

domain of choice, again using Golden Gate cloning. Finally, Oneshot Top10 *E. coli* carrying the respective reporter plasmid were transformed with the assembled libraries by electroporation. Following a recovery in super optimal broth supplemented with 20 mM glucose (Carl Roth) (SOC) for 1 h at 37 °C and 220 rpm, transformed cells were grown in LB (50 µg mL<sup>-1</sup> chloramphenicol and 25 µg mL<sup>-1</sup> 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 cultures with sterile 50% (v/v) glycerol at a ratio of 1:1, and stocks were stored at -80 °C until usage.

**FACS-Based Library Enrichment:** Precultures of LB media (50 µg mL<sup>-1</sup> of chloramphenicol and 25 µg mL<sup>-1</sup> of kanamycin) were inoculated from glycerol stocks of *E. coli* strains carrying the insertion libraries. Positive control samples expressing the wildtype effector protein without insert, as well as negative controls expressing a different protein of similar size (not activating the reporter) 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 for 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 using the forward scatter (FSC) and side scatter (SSC) values (Figure S21, Supporting Information). 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 for each library into LB media. Next, the collected cells were recovered for one hour in LB media without antibiotics at 37 °C and shaking at 220 rpm. Subsequently, 50 µg mL<sup>-1</sup> chloramphenicol and 25 µg mL<sup>-1</sup> of kanamycin were added, followed by incubation of cultures overnight. The next day, glycerol stocks were prepared from the cultures representing sorted libraries. A second round of FACS-sorting and enrichment was performed by repeating the procedure starting from the glycerol stocks after the first round of enrichment. FACS data were analyzed using the cytoflow python package (<https://cytoflow.github.io/>).

**Next Generation Sequencing:** The input libraries, as well as the enriched sorted fractions were subjected 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 until further use. The coding sequence of the libraries was amplified using the Q5 Hot Start High-Fidelity DNA Polymerase (NEB) and the PCR amplicons were separated from primer dimers on a 0.5× TAE 1% agarose gel. The bands representing the protein hybrid libraries were excised and DNA was purified using the QIAquick Gel Extraction Kit (Qiagen). The DNA concentration was then measured with the Quant-iT dsDNA (HS) assay kit (Thermo Fisher Scientific) using 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). 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, the 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 step. 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) us-

ing the Agilent DNA 1000 Kit. Finally, samples were sequenced using the paired-end Illumina MiSeq and NextSeq sequencing services at the EMBL Gene Core facility (Heidelberg).

**Experimental Characterization of Individual Variants from the Domain Insertion Screen:** Individual protein hybrids were isolated from the sorted fractions or cloned individually and stored as glycerol stocks in 25% glycerol (Carl Roth). The variants tested are listed in Table S3 (Supporting Information). Precultures of Oneshot Top10 cells carrying a RFP reporter plasmid specific to the respective protein hybrid, 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<sup>-1</sup> chloramphenicol (Carl Roth) and 25 µg mL<sup>-1</sup> 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<sup>-1</sup> chloramphenicol and 25 µg mL<sup>-1</sup> of kanamycin, using the same induction scheme as for the FACS screen. 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<sub>600</sub> were measured on 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/OD<sub>600</sub> values were calculated by dividing the measured fluorescence by the OD<sub>600</sub> levels. Three independent biological replicates prepared and measured on different days were generated for each variant.

**Illumination Setup:** For the illumination of liquid cultures, a 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 ≈30 cm. Liquid cultures were incubated in multi-well plates and illuminated at a constant intensity of 50 µmol m<sup>-2</sup> s<sup>-1</sup> (≈5 W m<sup>-2</sup>).

For the illumination of agar plates (see "agar plate photography" below), 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 onto a circuit board was used, applying a light intensity of 15 µmol m<sup>-2</sup> s<sup>-1</sup> (≈1.5 W m<sup>-2</sup>). This device was again 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. The plate was then placed above the LED array at a distance of ≈5 cm. The whole setup was installed inside a standard bacteria incubator (Minitron, Infors). The LED devices were custom-made by the workshop of the biology department at TU Darmstadt.

**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 LB (Carl Roth), supplemented with 50 µg mL<sup>-1</sup> chloramphenicol (Carl Roth) and 25 µg mL<sup>-1</sup> 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<sup>-1</sup> chloramphenicol and 25 µg mL<sup>-1</sup> 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 and inoculated with 5 µL from the respective precultures. Subsequently, one replicate was incubated under blue light exposure, while the other replicate was kept in the dark within the same incubator. The growth conditions were again at 37 °C and 220 rpm for 16 h. Following incubation, RFP fluorescence and OD<sub>600</sub> were measured in a plate reader. As before, an excitation wavelength of 490 nm and an emission wavelength of 520 nm were used and the fluorescence was normalized to the OD<sub>600</sub>. Experiments were performed in three independent replicates.

Activity measurements of the AraC derivatives carrying point mutations were performed identically using an arabinose concentration of 8 mM.

**Agar Plate Photography:** Prior to the experiment, agar plates were prepared using 1.5% LB-agar, supplemented with 50  $\mu\text{g mL}^{-1}$  chloramphenicol and 25  $\mu\text{g mL}^{-1}$  of kanamycin, 400  $\mu\text{M}$  IPTG and 25 mM L-arabinose (all Carl Roth). A preculture of the AraC-S170-LOV reporter strain was incubated overnight at 37 °C and 220 rpm. The next day, 0.6% LB-agar was freshly prepared and cooled to  $\approx 40$  °C. Next, 3 mL of the liquid agar were supplemented with IPTG and L-arabinose to final concentrations of 400  $\mu\text{M}$  and 25 mM, respectively. Finally, 300  $\mu\text{L}$  of the preculture was quickly added to the agar, mixed by shaking and distributed on the previously prepared agar plates. After 30 min at room temperature, the top agar 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, high-pass filter, and camera.

**Reversible Optogenetic Gene Expression Control:** In a 48-well plate (Corning), 0.5 mL cultures were prepared, using LB media, supplemented with 50  $\mu\text{g mL}^{-1}$  chloramphenicol and 25  $\mu\text{g mL}^{-1}$  of kanamycin, 400  $\mu\text{M}$  IPTG and 25 mM L-arabinose (all Carl Roth). The wells were inoculated with 5  $\mu\text{L}$  of precultures that had been prepared as described above. The samples were then incubated at 37 °C and 220 rpm for 3 h 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<sub>600</sub> were measured in a plate reader. Following 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<sub>600</sub>. Three independent replicates were generated by repeating experiments on different days.

**Structure Prediction with AlphaFold2:** Full-length structures of AraC, SigF, the TVMV protease, Flp, as well as the AraC-LOV2 fusions were obtained by AlphaFold2<sup>[21]</sup> using the Colabfold implementation.<sup>[22]</sup> Structures were predicted using the “colabfold\_batch” command with the “MMseqs2 (UniRef+Environmental)” MSA preferences. For the proteins without insertion, five 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).<sup>[46,47]</sup> To compute the position-wise RMSDs for between the AraC-LOV2 hybrids and the respective wild-type 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.

Besides AF2-predicted structures, several previously published experimental structures are shown in several figures (Table S5, Supporting Information).

**NGS and Data Analysis:** To analyze the sequencing data, 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.<sup>[12]</sup> Briefly, the sequencing data were subjected to quality control, i.e., corrupted or mutant reads were filtered out. Next, reads that contained the insert sequence were selected and the insertion site was determined. Then, the enrichment scores were calculated using the following Equation 1:

$$\text{Enrichment score}_i = \log_2 \left[ \frac{\text{count enriched}_i / \text{count initial}_i}{\sum_i^n \text{count enriched}_i / \sum_i^n \text{count initial}_i} \right] \quad (1)$$

where  $i$  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 the analysis. Insertion variants that entirely disappeared during sorting and could thus not be log<sub>2</sub>-scaled, were assigned a value of −10, which was in the range of the lowest experimentally obtained enrichment scores.

To gather position-wise protein features, diverse feature sources were used. Biophysical properties and linker propensity indices were fetched

from the AAindex database.<sup>[23,24]</sup> 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 (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),<sup>[48,49]</sup> 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,<sup>[50]</sup> using the Super5 algorithm with standard parameters. Finally, the KLD was calculated as indicated in Equation 2:

$$\text{Divergence}_i = \sum_a f_i(a) \cdot \log_{10} \frac{f_i(a)}{b(a)} \quad (2)$$

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. Background frequencies were defined as the AA frequencies in SwissProt.<sup>[51]</sup> Of note, the definition of the gap background frequencies is non-trivial, as discussed by Teşileanu et al.<sup>[52]</sup> 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.

**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.<sup>[53]</sup> 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. As model architecture, Gradient boosted regression trees were used.<sup>[29]</sup> Gradient boosting models are ensemble models that iteratively use simple models to optimize a loss function. Here, the gradient boosting classifier implementation from “Scikit learn” was used, which employs regression trees as base models. The model was optimized on the training data set using five-fold cross-validation. 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 (ROC) and precision-recall plots were chosen as performance metrics. ROC curves illustrate the classification performance setting the true positive classification rate in relation to the false positive classification rate for different classification thresholds. The area under the ROC thus summarizes the relation between true positives and false positives in a single value. Precision recall plots instead, show the precision that a model reaches in relation to its recall or sensitivity. The average precision refers to the weighted mean of the calculated precisions. The permutation importance and loss of impurity were calculated using the respective Scikit-learn functions.

**Statistical Analysis:** The domain insertion screen was performed in two independent replicates. Pearson correlations were calculated, to assess the similarity between replicates. For the analysis of domain insertion tolerance, the mean of the two replicates was used. In order to analyze the influence of positional protein features on domain insertion



permissibility, Spearman correlations between the measured enrichment scores and the respective features were calculated and the Spearman  $r$  values are reported. The experimental validation of individual variants and the characterization of the AraC-LOV hybrids were performed in  $n = 3$  independent replicates. The mean of the measurements, as well as the standard deviation are indicated in the respective figures.

## Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

## Acknowledgements

The authors thank the members of the Niopek lab for helpful discussions. Further, the authors are grateful to the ZMBH flow cytometry core facility (Heidelberg University) for support with cell sorting and the EMBL Genomics Core Facility (EMBL, Heidelberg) for performing deep sequencing. Finally, the authors sincerely thank the workshop at the Biology Department of the Technical University Darmstadt for the construction of customized illumination setups. Funded by the European Union (ERC, DaVinci-Switches, project number 101041570). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. D.N. is also grateful for funding from the German Research Foundation (DFG) [project no. 453202693], the Schwiete Stiftung, and the Aventis foundation. J.M. was partially funded by the German Academic Scholarship Foundation.

## Conflict of Interest

The authors declare no conflict of interest.

## Author Contributions

D.N. and J.M. conceived the study. J.M., S.A., and P.B. designed and performed the experiments. J.M. implemented the computational analysis. D.N. directed the work and secured funding. J.M. and D.N. wrote the manuscript with support from all authors.

## Data Availability Statement

The data that support the findings of this study are available in the supplementary material of this article. The computational analysis, as well as experimental raw data are available at Github under: [https://github.com/Niopek-Lab/DI\\_screen](https://github.com/Niopek-Lab/DI_screen). The structures shown in the figures are including all color codes are provided on the Github repository as ChimeraX session files. The AF2-predicted Structures of all possible AraC-PDZ hybrids (corresponding to figure S14) are available as PDB files on Github. Plasmids encoding the AraC-I113-LOV, AraC-S170-LOV, AraC-S170-LOV\_G141D and AraC-S170-LOV\_T50S are available on Addgene (Addgene-IDs: #206804; #206805).

## Keywords

allostery, domain insertion, optogenetics, protein engineering

Received: May 30, 2023

Revised: July 21, 2023

Published online: August 10, 2023

- [1] C. P. Ponting, R. R. Russell, *Annu. Rev. Biophys. Biomol. Struct.* **2002**, 31, 45.
- [2] J. Jin, X. Xie, C. Chen, J. G. Park, C. Stark, D. A. James, M. Olhovsky, R. Linding, Y. Mao, T. Pawson, *Sci Signal* **2009**, 2, p ra76.
- [3] O. Dagliyan, M. Tarnawski, P. H. Chu, D. Shirvanyants, I. Schlichting, N. V. Dokholyan, K. M. Hahn, *Science* **2016**, 354, 1441.
- [4] M. S. Siegel, E. Y. Isacoff, *Neuron* **1997**, 19, 735.
- [5] J. Lee, M. Natarajan, V. C. Nashine, M. Socolich, T. Vo, W. P. Russ, S. J. Benkovic, R. Ranganathan, *Science* **2008**, 322, 438.
- [6] M. Ostermeier, *Protein Eng Des Sel* **2005**, 18, 359.
- [7] G. Guntas, T. J. Mansell, J. R. Kim, M. Ostermeier, *Proc. Natl. Acad. Sci. USA* **2005**, 102, 11224.
- [8] O. Dagliyan, N. V. Dokholyan, K. M. Hahn, *Nat. Protoc.* **2019**, 14, 1863.
- [9] M. D. Hoffmann, J. Mathony, J. Upmeier Zu Belzen, Z. Hartevelde, S. Aschenbrenner, C. Stengl, D. Grimm, B. E. Correia, R. Eils, D. Niopek, *Nucleic Acids Res.* **2021**, 49, e29.
- [10] F. Bubeck, M. D. Hoffmann, Z. Hartevelde, S. Aschenbrenner, A. Bietz, M. C. Waldhauer, K. Börner, J. Fakhiri, C. Schmelas, L. Dietz, D. Grimm, B. E. Correia, R. Eils, D. Niopek, *Nat. Methods* **2018**, 15, 924.
- [11] B. L. Oakes, D. C. Nadler, A. Flamholz, C. Fellmann, B. T. Staahl, J. A. Doudna, D. F. Savage, *Nat. Biotechnol.* **2016**, 34, 646.
- [12] D. C. Nadler, S. A. Morgan, A. Flamholz, K. E. Kortright, D. F. Savage, *Nat. Commun.* **2016**, 7, 12266.
- [13] K. A. Reynolds, R. N. McLaughlin, R. Ranganathan, *Cell* **2011**, 147, 1564.
- [14] W. R. Edwards, K. Busse, R. K. Allemann, D. D. Jones, *Nucleic Acids Res.* **2008**, 36, e78.
- [15] W. Coyote-maestas, D. Nedrud, S. Okorafor, Y. He, D. Schmidt, *Nucleic Acids Res.* **2019**, 48, e11.
- [16] W. Coyote-Maestas, Y. He, C. L. Myers, D. Schmidt, *Nat. Commun.* **2019**, 10, 290.
- [17] W. Coyote-Maestas, D. Nedrud, A. Suma, Y. He, K. A. Matreyek, D. M. Fowler, V. Carnevale, C. L. Myers, D. Schmidt, *Nat. Commun.* **2021**, 12, 7114.
- [18] J. Fernandez-Rodriguez, C. A. Voigt, *Nucleic Acids Res.* **2016**, 44, 6493.
- [19] M. Ormö, A. B. Cubitt, K. Kallio, L. A. Gross, R. Y. Tsien, S. J. Remington, *Science* **1996**, 273, 1392.
- [20] O. Dagliyan, D. Shirvanyants, A. V. Karginov, F. Ding, L. Fee, S. N. Chandrasekaran, C. M. Freisinger, G. A. Smolen, A. Huttenlocher, K. M. Hahn, N. V. Dokholyan, *Proc. Natl. Acad. Sci. USA* **2013**, 110, 6800.
- [21] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Židek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, et al., *Nature* **2021**, 596, 583.
- [22] M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, M. Steinegger, *Nat. Methods* **2022**, 19, 679.
- [23] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, M. Kanehisa, *Nucleic Acids Res.* **2008**, 36, D202.
- [24] S. Kawashima, M. Kanehisa, *Nucleic Acids Res.* **2000**, 28, 374.
- [25] M. Suyama, O. Ohara, *Bioinform.* **2003**, 19, 673.
- [26] R. A. George, J. Heringa, *Protein Eng Des Sel* **2002**, 15, 871.
- [27] K. Bae, B. K. Mallick, C. G. Elsik, *Bioinform.* **2005**, 21, 2264.
- [28] M. Akdel, D. E. V. Pires, E. P. Pardo, J. Jänes, A. O. Zalevsky, B. Mészáros, P. Bryant, L. L. Good, R. A. Laskowski, G. Pozzati, A. Shenoy, W. Zhu, P. Kundrotas, V. R. Serra, C. H. M. Rodrigues, A. S. Dunham, D. Burke, N. Borkakoti, S. Velankar, A. Frost, J. Basquin, K. Lindorff-Larsen, A. Bateman, A. V. Kajava, A. Valencia, S. Ovchinnikov, J. Durairaj, D. B. Ascher, J. M. Thornton, N. E. Davey, et al., *Nat. Struct. Mol. Biol.* **2022**, 29, 1056.
- [29] J. H. Friedman, *Comput Stat Dat Anal.* **2002**, 38, 367.
- [30] G. Louppe, 10.48550/arXiv.1407.7502, **2015**.

- [31] J. Mathony, D. Niopek, *Adv. Biol.* **2021**, *5*, 2000181.
- [32] S. M. Soisson, B. MacDougall-Shackleton, R. Schleif, C. Wolberger, *Science* **1997**, *276*, 421.
- [33] R. Schleif, *FEMS Microbiol. Rev.* **2010**, *34*, 779.
- [34] N. Halabi, O. Rivoire, S. Leibler, R. Ranganathan, *Cell* **2009**, *138*, 774.
- [35] E. Romano, A. Baumschlager, E. B. Akmeriç, N. Palanisamy, M. Houmani, G. Schmidt, M. A. Öztürk, L. Ernst, M. Khammash, B. Di Ventura, *Nat. Chem. Biol.* **2021**, *17*, 817.
- [36] J. Dietler, R. Schubert, T. G. A. Krafft, S. Meiler, S. Kainrath, F. Richter, K. Schweimer, M. Weyand, H. Janovjak, A. Möglich, *J. Mol. Biol.* **2021**, *433*, 167107.
- [37] X. Li, C. Zhang, X. Xu, J. Miao, J. Yao, R. Liu, Y. Zhao, X. Chen, Y. Yang, *Nucleic Acids Res.* **2020**, *48*, e33.
- [38] P. Jayaraman, K. Devarajan, T. K. Chua, H. Zhang, E. Gunawan, C. L. Poh, *Nucleic Acids Res.* **2016**, *44*, 6994.
- [39] P. Jayaraman, J. W. Yeoh, J. Zhang, C. L. Poh, *ACS Synth. Biol.* **2018**, *7*, 2627.
- [40] C. Engler, R. Kandzia, S. Marillonnet, *PLoS One* **2008**, *3*, e3647.
- [41] R. E. Campbell, O. Tour, A. E. Palmer, P. A. Steinbach, G. S. Baird, D. A. Zacharias, R. Y. Tsien, *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 7877.
- [42] K. E. McGinness, T. A. Baker, R. T. Sauer, *Mol. Cell* **2006**, *22*, 701.
- [43] R. B. Kapust, J. Tózsér, J. D. Fox, D. E. Anderson, S. Cherry, T. D. Copeland, D. S. Waugh, *Protein Eng. Des. Sel.* **2001**, *14*, 993.
- [44] P. Sun, B. P. Austin, J. Tózsér, D. S. Waugh, *Protein Sci.* **2010**, *19*, 2240.
- [45] I. Bervoets, M. Van Brempt, K. Van Nerom, B. Van Hove, J. Maertens, M. De Mey, D. Charlier, *Nucleic Acids Res.* **2018**, *46*, 2133.
- [46] T. D. Goddard, C. C. Huang, E. C. Meng, E. F. Pettersen, G. S. Couch, J. H. Morris, T. E. Ferrin, *Protein Sci.* **2018**, *27*, 14.
- [47] E. F. Pettersen, T. D. Goddard, C. C. Huang, E. C. Meng, G. S. Couch, T. I. Croll, J. H. Morris, T. E. Ferrin, *Protein Sci.* **2021**, *30*, 70.
- [48] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, *Nucleic Acids Res.* **1997**, *25*, 3389.
- [49] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **1990**, *215*, 403.
- [50] R. C. Edgar, *Nucleic Acids Res.* **2004**, *32*, 1792.
- [51] A. Bairoch, R. Apweiler, *J Mol Med (Berl)* **1997**, *75*, 312.
- [52] T. Teşileanu, L. J. Colwell, S. Leibler, *PLoS Comput. Biol.* **2015**, *11*, e1004091.
- [53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, *J Mach Learn Res* **2011**, *12*, 2825.