Thèse n° 9195

Towards automating de novo protein design for novel functionalities: controlling protein folds and protein-protein interactions

Présentée le 28 avril 2022

Faculté des sciences et techniques de l’ingénieur
Laboratoire de conception de protéines et d’immuno-ingénierie
Programme doctoral en biotechnologie et génie biologique

pour l’obtention du grade de Docteur ès Sciences

par

Zander HARTEVELD

Acceptée sur proposition du jury

Prof. A. L. A. Persat, président du jury
Prof. B. E. Ferreira De Sousa Correia, directeur de thèse
Prof. D. Woolfson, rapporteur
Dr S. Ovchinnikov, rapporteur
Prof. P. De Los Rios, rapporteur

■ École
polytechnique
fédérale
de Lausanne

2022

designs matter.

Acknowledgements

I am deeply thankful for the many people who have enriched my life and supported me during my time as a Ph.D. student.

First and foremost, to my advisor Bruno Correia, thank you for mentoring me and for making this possible. Thank you for your constant advice, strong support and positive energy to overcome the many roadblocks encountered during this journey. Your ability to think big, connect creative ideas from different fields and people in science is exceptional and shaped the last four years into an amazing counter-cultural scientific experience. Thank you for always having given me freedom to approach matters in my own way, and also to always have challenged me to move out of my comfort zone - searching for innovative solutions and grow personally and professionally. I am grateful for your trust in me and your help during these years. It has been an incredible pleasure.

I would like to deeply thank Dominik Niopek who has been both a collaborator and second mentor to me, sharing perspectives, support and advice. I would like to thank Micheal Bronstein and Pierre Vandergheynst for helping on the machine learning related projects and having given me the opportunity to work within their groups.

I would like to thank my good friend and collaborator Ahmed Sadek for his constant help during nearly half of my Ph.D. - thank you Ahmed for sharing your skills and scientific knowledge with me! I have learned so much from you and it's been an incredible ride that I feel lucky to have shared with you. Thank you Pablo Gainza - professionally, I owe you my deepest gratitude for your brilliant insights, the numerous and rich scientific discussions and your constant help. Thank you for mentoring me during all this time. Thank you to my friend and compatriot Freyr Sverrisson. Your brilliant feedback and stimulating discussions helped me to better understand the physics of proteins and to "machine-learn". I would like to thank Andreas Loukas and Michaël Defferrard. Thanks for your constant scientific input and technical assistance to merge the fields of computational protein design with deep learning! Thank you Jaume Bonet for having not only helped me towards becoming a bioinformatician and computational biologist, but also for having inspired me with incredible scientific graphics and figures throughout. I would like to thank former Ph.D. students Che Yang and Fabian Sesterhenn for their support, insightful questions, and feedback regarding my work. It was a privilege to have worked with you and learn many new skills. Andreas Scheck, Sarah Wehrle and Sailan Shui that have started their journey alongside with me - thank you for the all the joyful moments we shared. Much of this work would not have been possible without the generous support of Stéphane Rosset and Sandrine Georgeon. Thank you very much for all the time and effort you invested to help progress this work. To my close collaborators and members of the LPDI, thank you for your immense help, discussions, lunches breaks and beers; it's been incredible. To so many members of and passers-by of the LPDI — Karen Schriever, Jayne Marsden, Julius Upmeier zu Belzen, Tiana Schwab, Célina Chkroun, Charles Berger and Arthur Valentin. Thank you for all your work that has greatly contributed to my thesis. Specifically, I'd like to thank Leo Scheller, Ahmed Sadek, Freyr Sverrisson and Andreas Scheck for all effort the efforts spent proofreading my thesis.

To all my former and current flatmates - Félix, Chris, Béa, Bastien, Javier, Susi, Jasso, Sämi, Marie, Maggie - thank you for always making it a joy and for having made the restrictive COVID-times easier. I am so thankful to my good friends Ece, Michele, Yannic, Michael, Quentin, Joachim, Dominique, and many more for all their support, travel experiences, sport sessions, beers and wine, and parties during these years. It has been a blast. I would like to thank my cousin Alan for his constant support during those 4 years. Thanks Alan! A very special thank you to my siblings Killian, Daphne and Robin. Thank you for always helping me out!

Most importantly, I would like to deeply thank Claire and Ben Harteveld for their unconditioned support, help and love during all these years. Thank you for always being here for me.

Z. H.

Abstract

The sheer size of the protein sequence space is massive: a protein of 100 residues can have 20^{100} possible sequence combinations; and knowing that this exceeds the number of atoms in the universe, the chance of randomly discovering a stable new sequence with the desired characteristics is infinitesimally small. Therefore, computational methodologies that can search through the sequence space and expand beyond naturally occurring functional protein sequence variants hold enormous potential in biomedicine and nanotechnology.

My thesis work leverages machine learning, physics-based and data-driven techniques to design new protein molecules with distinct shapes (folds) so that they can precisely interact with other molecules to perform biological functions.

The first part of my thesis is dedicated to the design of functional proteins. The re-designed of an anti-CRISPR protein that can be controlled via blue light (optogenetic control) to regulate the genome editing activity of the enzyme CRISPR–Cas9 is presented. A surface-based design of a broad-spectrum inhibitory Acr towards another natural target (*Sau*Cas9) exemplifies the re-purposing of existing inhibitory molecules against other related targets. This ultimately led to the development of a general surface-centric design method for generating specific protein-protein interactions from scratch and exemplified by the successful design of novel PD-L1 inhibitors, an immune checkpoint that can halt the immune system from attacking the cancer cells.

The successful design of protein-protein interactions heavily relies on the underlying protein fold and structure stabilizing the functional motif in a protein. Because nature has only evolved a small set of protein folds, generated protein-interaction motifs can rarely be incorporated into existing protein structures. To address this problem, the second part of my thesis is dedicated to the development of computational *de novo* protein design methods for the crafting of proteins with customized folds. To this end, the TopoBuilder framework utilizes a large collection of native proteins to transform a literal description of a protein fold into a physically-realistic protein. Finally, Genesis, a deep neural networks approach for the tailored *de novo* protein design is presented. Employing both, the TopoBuilder and Genesis, proteins completely absent from the natural repertoire were designed and experimentally validated.

My thesis sets the path to explore possibilities of jointly optimizing the protein's shape and its surface geometry to master biological functions. We are now at entering a new era where newly designed protein-based drugs and materials with the potential to solve a vast array of technical challenges and open new avenues for next-generation precision drugs and advanced nanomaterials.

Key words: *De novo* protein design, protein-protein interactions, protein structure, deep learning

Résumé

La taille même de l'espace des séquences protéiques est énorme : une protéine de 100 résidus peut avoir 20^{100} combinaisons de séquences possibles ; sachant que cela dépasse le nombre d'atomes dans l'univers, la chance de découvrir au hasard une nouvelle séquence stable avec les caractéristiques désirées est infiniment petite. Par conséquent, les méthodologies informatiques qui peuvent rechercher dans l'espace des séquences et s'étendre au-delà des variantes de séquences de protéines fonctionnelles naturelles présentent un énorme potentiel en biomédecine et en nanotechnologie.

Mon travail de thèse s'appuie sur l'apprentissage automatique, les techniques basées sur la physique et les données pour concevoir de nouvelles molécules de protéines avec des formes distinctes (repliement) afin qu'elles puissent interagir précisément avec d'autres molécules pour remplir des fonctions biologiques.

La première partie de ma thèse est consacrée à la conception de protéines fonctionnelles. La re-conception d'une protéine anti-CRISPR qui peut être contrôlée via la lumière bleue (contrôle optogénétique) pour réguler l'activité d'édition du génome de l'enzyme CRISPR-Cas9 est présentée. Une conception basée sur la surface d'un Acr inhibiteur à large spectre vers une autre cible naturelle (*Sau*Cas9) illustre la réaffectation de molécules inhibitrices existantes contre d'autres cibles apparentées. Cela a finalement conduit au développement d'une méthode de conception centrée sur la surface générale pour générer des interactions protéine-protéine spécifiques à partir de zéro et illustrée par la conception réussie des nouveaux inhibiteurs PD-L1, un point de contrôle immunitaire qui peut empêcher le système immunitaire d'attaquer les cellules cancéreuses.

La conception réussie des interactions protéine-protéine repose fortement sur le repliement protéique sous-jacent et la structure stabilisant le motif fonctionnel dans une protéine. Parce que la nature n'a développé qu'un petit ensemble de plis protéiques, les motifs d'interaction protéique générés peuvent rarement être incorporés dans les structures protéiques existantes. Pour résoudre ce problème, la deuxième partie de ma thèse est consacrée au développement de méthodes computationnelles de conception de protéines *de novo* pour la fabrication de protéines avec des plis personnalisés. À cette fin, le cadre TopoBuilder utilise une grande collection de protéines natives pour transformer une description littérale d'un pli protéique en une protéine physiquement réaliste. Enfin, Genesis, une approche de réseaux de neurones profonds pour la conception sur mesure de protéines *de novo* est présentée. En utilisant à la fois le TopoBuilder et Genesis, des protéines totalement absentes du répertoire naturel ont été conçues et validées expérimentalement.

Ma thèse ouvre la voie pour explorer les possibilités d'optimiser conjointement la forme de la protéine et sa géométrie de surface pour maîtriser les fonctions biologiques. Nous entrons maintenant dans une nouvelle ère où des médicaments et des matériaux à base de protéines nouvellement conçus ont le potentiel de résoudre un vaste éventail de défis techniques et d'ouvrir de nouvelles voies pour les médicaments de précision de nouvelle génération et les nanomatériaux avancés.

Mots clefs : *De novo* conception de protéines, interactions protéine-protéine, structure des protéines, apprentissage profond

Contents

Acknowledgements **i**

Abstract **iii**

Résumé **v**

List of figures **xi**

List of supplementary figures **xiii**

List of abbreviations **xv**

1 Introduction **1**

1.1 Manual, minimal, and rational protein design 5
1.2 Computational protein design 6
1.3 *De novo* protein design – classical methods 8
1.4 Emerging *de novo* design methods 10
1.4.1 Structure prediction DNNs 12
1.4.2 Neural generation of protein backbones for fold discovery 14
1.4.3 *De novo* sequence generation conditioned on structure information . . . 15
1.4.4 Inverting protein structure prediction nets for *de novo* design 15
1.5 Protein design for biomolecular recognition and signaling 17
1.6 Objectives 24
1.6.1 Aim 1: Computational protein modeling frameworks for applied PPI design. 24
1.6.2 Aim 2: Hierarchical design of *de novo* proteins with native-like features. 24
1.6.3 Aim 3: Tailored *de novo* design using DNNs to probe the “darkmatter” of the protein universe. 25

2 Engineered anti-CRISPR proteins for optogenetic control of CRISPR–Cas9 **27**

2.1 Abstract 28
2.2 Main 28
2.3 Methods 33

2.3.1 Structure-based identification of the LOV2 insertion site 33
2.3.2 Domain assembly 33
2.3.3 Computational design of improved Acr–LOV mutants 33
2.3.4 General methods and cloning 34
2.3.5 Cell culture, transient transfection and AAV lysate production 35
2.3.6 Blue light setup 35
2.3.7 Luciferase reporter assays 36
2.3.8 T7 endonuclease assay and TIDE sequencing 36
2.3.9 Quantitative RT-PCR 37
2.3.10 Telomere labeling experiments 38
2.3.11 Automated image analysis for telomere labeling in KNIME 38
2.3.12 Membrane recruitment experiments 39
2.3.13 Statistical analysis 40
2.4 Supplementary information 41
2.4.1 Supplementary Note 1 53
2.4.2 Supplementary Note 2: Considerations for luciferase reporter cleavage assay data interpretation 53
2.4.3 Supplementary Note 3: Considerations for T7 assay data interpretation . 53
2.4.4 Supplementary Note 4 54
2.4.5 Supplementary Note 5 54
2.4.6 Supplementary Note 6: Comment on the used experimental timings . . 54
2.4.7 Supplementary Note 7 55
2.4.8 Supplementary Discussion 56

3 Computational design of anti-CRISPR proteins with improved inhibition potency 59
3.1 Abstract 60
3.2 Main 60
3.3 Results 61
3.3.1 Domain insertion into AcrIIC1 boosts *Nme*Cas9 inhibition 61
3.3.2 Mechanistic insights into effects of domain insertion 63
3.3.3 Computational design yields a potent *Sau*Cas9 inhibitor 65
3.3.4 A hepatocyte-specific *Sau*Cas9 ON-switch 67
3.4 Discussion 68
3.5 Methods 70
3.5.1 Modeling of AcrIIC1-mCherry fusions 70
3.5.2 AcrIIC1 interface design 70
3.5.3 Protein expression and purification 71
3.5.4 Circular dichroism 72
3.5.5 Size-exclusion chromatography combined with multi-angle light scattering 72
3.5.6 Affinity measurements 72
3.5.7 Mammalian expression construct design and cloning 73
3.5.8 Cell culture and AAV lysate production 73

3.5.9 Large-scale AAV production, purification and titration 74
3.5.10 Luciferase reporter assays 75
3.5.11 T7 endonuclease assay and TIDE sequencing 75
3.5.12 Western blot 77
3.5.13 *In vitro* DNA cleavage assay 78
3.5.14 Electrophoretic mobility shift assay 78
3.5.15 Statistical analysis 79
3.6 Supplementary information 80

4 *De novo* design of site-specific protein binders using learned surface fingerprints 97
4.1 Abstract 98
4.2 Main 98
4.3 Results 99
4.3.1 Design of *de novo* PPIs using learned surface fingerprints 99
4.3.2 Fingerprint-based *de novo* design of protein-binders against PD-L1 . . . 101
4.4 Discussion 105
4.5 Methods 107
4.5.1 Computational grafting and interface optimizations 107
4.6 Supplementary information 109

5 A generic framework for hierarchical *de novo* protein design 113
5.1 Abstract 113
5.2 Main 114
5.3 Results 115
5.3.1 Sketching native-like protein backbones from Forms 117
5.3.2 Backbone assembly and sequence design of native-like Sketches 118
5.3.3 *De novo* design of five protein folds 119
5.3.4 Corrections induce native features in idealized folds 120
5.3.5 Assessing designability by approximating the inherent sequence-to-structure compatibility 121
5.3.6 Experimental validation of novel sequences 123
5.4 Discussion 125
5.5 Methods 126
5.5.1 Correction module (InteractiveMaster) 126
5.5.2 Loop creation (LoopMaster) 126
5.5.3 The number of architectures and topologies from (multi-)Forms 127
5.5.4 Protein Expression and Purification 127
5.5.5 Circular dichroism spectroscopy 128
5.5.6 Size-exclusion chromatography combined with multi-angle light scattering128
5.6 Supplementary information 129
5.6.1 Supplementary methods 129
5.6.2 Supplementary figures 129

6 Tailored *de novo* protein design with deep neural networks **141**
6.1 Abstract 142
6.2 Main 142
6.3 Results 144
6.3.1 Sketching protein drafts from native backbones 144
6.3.2 *De novo* design through collaborative deep neural networks 146
6.3.3 Probing "dark matter" protein folds 150
6.4 Discussion 152
6.5 Methods 153
6.5.1 Dataset generation 153
6.5.2 Data splits 153
6.5.3 Data encoding 154
6.5.4 Genesis architecture 154
6.5.5 Loss function 154
6.6 Supplementary information 156

7 Conclusions & Perspectives **159**
7.1 Orchestrating PPIs and beyond 159
7.2 Probing the protein universe with designer proteins 162

A RosettaSilentToolbox **165**
A.1 Abstract 165
A.2 Background 166
A.3 Implementation 167
A.4 Results 170
A.4.1 Analysis of protein backbone features 170
A.4.2 Guiding iterative CPD workflows 170
A.4.3 Evaluation of designed proteins 172
A.4.4 Benchmarking design protocols 173
A.5 Discussion 174
A.6 Conclusions 176

B Selected sequences **179**

Bibliography **209**

Curriculum Vitae **211**

List of Figures

1.1 Orders of protein structure 3
1.2 Data growth 4
1.3 Computational protein design 9
1.4 DL-based *de novo* protein design methods 17
1.5 Protein interface re-design 19
1.6 Hotspot-centric *de novo* PPI design 20
1.7 MaSIF framework and applications 21
1.8 MaSIF-based *de novo* PPI design framework 22

2.1 Engineering and characterization of CASANOVA, an anti-CRISPR protein dependent on blue light 29
2.2 Optogenetic control of gene expression and telomere labeling. 31

3.1 Improving the efficacy of Acr proteins by protein engineering. 61
3.2 Domain insertion into AcrIIC1 yields a highly potent *Nme*Cas9 inhibitor. 62
3.3 Enhanced potency of the AcrIIC1-mCherry chimeric inhibitor arises from multiple factors. 64
3.4 Structure-guided design of AcrIIC1X, an Acr protein targeting *Sau*Cas9. 66
3.5 Characterization of AcrIIC1X, a designer *Sau*Cas9 inhibitor. 67

4.1 Surface-centric design of protein interactions 100
4.2 Seed-based design of PD-L1 binders 102
4.3 Design and optimization of the two PD-L1 binder designs 103
4.4 Structural validation of designs 106

5.1 Form parametrization for protein structures 116
5.2 Set-up for geometric parametric correction and test examples 117
5.3 Geometric corrections 120
5.4 Computational assessment of sequence quality 122
5.5 Experimental results 124

6.1 Genesis neural *de novo* protein design pipeline 144
6.2 Network architecture and data engineering 145
6.3 Pipeline performances 148
6.4 Computational design of native folds 150
6.5 Computational design of dark matter folds 151

A.1 rstoolbox library architecture. 168
A.2 Sample code for the evaluation of protein backbone dihedral angles and fragment quality. 169
A.3 Ramachandran plots and fragment quality profiles. 171
A.4 Guiding iterative design pipelines. 172
A.5 Sample code to guide iterative CPD workflows. 173
A.6 Multi-stage design. 174
A.7 Sample code for the evaluation of a multistep design pipeline. 175
A.8 Comparison and benchmarking of different design protocols. 176
A.9 Sample code for the comparison between 4 different decoy populations. 177

List of Supplementary Figures

2.1 Sampling of LOV2 insertion sites in AcrIIA4 loop 5. 41
2.2 Schematic and sequences of Acr–LOV hybrids. 41
2.3 Cas9 inhibition is dose dependent. 42
2.4 CASANOVA computational model. 42
2.5 Photoactivatable genome editing with CASANOVA. 43
2.6 TIDE sequencing analysis of light-mediated indel mutation. 44
2.7 Cas9 inhibition can be modulated via mutations that affect docking of the LOV2 terminal helices. 45
2.8 Acr–LOV hybrids carrying a C450A LOV2 pseudodark mutation are still light responsive. 46
2.9 *In silico* docking analysis reveals Acr mutations that improve CASANOVA performance. 47
2.10 Comparison of light-dependent indel mutation by CASANOVA and its corresponding S46D and T16F mutants. 48
2.11 Optogenetic control of xCas9. 48
2.12 CASANOVA-mediated IL1RN gene activation is reversible. 49
2.13 Fluorescence signal of single labeled telomeres increased after CASANOVA activation. 49
2.14 Reversible recruitment of Acr–LOV hybrid 5 to a plasma-membrane-targeted dCas9–gRNA complex. 50
2.15 CASANOVA mutants selected for experimental characterization. 50
2.16 List of constructs created and used in this study. 51
2.17 gRNA target sites. 52
2.18 Primers used for genomic PCRs and quantitative RT-PCR. 52
2.19 Full-length gel images. 58

3.1 Identification of a loop amenable to domain insertion. 80
3.2 Screening of AcrIIC1 domain insertion variants. 80
3.3 T7 endonuclease assay analysis of *N. meningitides* Cas9 inhibition by AcrIIC1 chimeras. 81

3.4 Acr-mCherry chimera conformations. 82
3.5 AcrIIC1 chimeras show improved inhibition of *Nme*2Cas9. 82
3.6 Alignment of the *Sau*- and *Nme*Cas9 HNH domains. 83
3.7 AcrIIC1 is able to partially inhibit genome editing by *S. aureus* Cas9. 83
3.8 Computational interface design. 84
3.9 Rosetta scores of single site mutants. 85
3.10 Screening and iterative optimization of computationally designed AcrIIC1 mutants. 86
3.11 Electrostatic potential of HNH domain and AcrIIC1 surfaces. 87
3.12 Biochemical analysis of AcrIIC1 and AcrIIC1X. 88
3.13 T7 endonuclease assays of *S. aureus* Cas9 inhibition by AcrIIC1X. 89
3.14 T7 endonuclease assays of *S. aureus* Cas9 inhibition by AcrIIC1X. 90
3.15 AcrIIC1X inhibits *Nme*Cas9 function, which can be further improved by domain insertion. 91
3.16 AcrIIC1X can be harnessed for cell type-specific, miRNA-dependent inhibition of *Sau*Cas9. 92
3.17 Polymorphism in the CCR5 gene. 93
3.18 Gel images and western blots. 93
3.19 Coomassie staining of PAGE loaded with purified proteins. 94
3.20 List of constructs. 94
3.21 Genomic target sites. 95
3.22 Primers for genomic PCRs. Fw, forward primer; rv, reverse primer. 95

4.1 Modeling buried surfaces as radial patches 109
4.2 Planarity of the PD-L1 interface 110
4.3 Seed clustering 111

5.1 Overview of the modules in the TopoBuilder *de novo* design framework 130
5.2 Ubiquitin-like fold geometric analysis. 131
5.3 Counts and examples for the correction search 132
5.4 Comparisons of naive and native-like Sketches 133
5.5 BLAST search against SwissProt 134
5.6 TopoBuilder *de novo* design pipelines 135
5.7 Computational assessment of sequence quality 136
5.8 ROC curves and ROC-AUC 137
5.9 SEC-MALS and thermal meltings for the designs 138
5.10 Sequence and structure similarities to the PDB 139
5.11 Sequence and structure similarities to the AF database 140

6.1 Sampling strategy 156
6.2 Pipeline performances with MSA design 157

List of abbreviations

AA Amino acid.

AF AlphaFold.

CASP Critical Assessment of Techniques for Protein Structure Prediction.

CATH Class, Architecture, Topology, Homology.

CNN Convolutional neural networks.

DEE Dead-end elimination.

DL Deep learning.

DNN Deep neural networks.

EM Electron microscopy.

GAN Generative adversarial network.

H-bond Hydrogen bond.

Ig-like Immunoglobulin-like.

MaSIF Molecular surface interaction fingerprinting.

MC Monte Carlo.

MSA Multiple sequence alignment.

NF Normalizing flows.

NMR Nuclear magnetic resonance.

PDB Protein Databank.

PPI protein-protein interaction.

RMSD Root-mean-square deviation.

RNA Ribonucleic acid.

SCOP Structural Classification of Proteins.

SSE Secondary structure element.

TB TopoBuilder.

trR Transformed-restrained Rosetta.

VAE Variational autoencoder.

vdW van der Waals.

1 Introduction

Around 13.8 billion years ago, the Big Bang induced a quick expansion and cooling of the universe allowing for basic atomic matter as we know it to emerge. With gravity and heat, atoms clustered together forming condensed clouds which further gave rise to stars and planets. Our planet Earth has formed around 4.5 billion years ago, at that time only composed of simple inorganic molecules. Lucky environmental conditions triggered chemical reactions that caused the formation of the Earth's prebiotic soup containing the four key families of organic molecules: lipids, carbohydrates, amino acids, and nucleic acids. The subtle interplay and fine balance between these four ingredients ultimately fostered fantastic complex molecular systems and the first living organisms.

—

Next to water, proteins are among the most abundant molecules in almost all of today's living organisms. They are of utmost importance to life — the human body, for example, produces ten thousand different proteins orchestrating nearly all vital functions. Proteins are encoded in the genetic material that is composed of nucleic acids (DNA) and stored in each and every cell. For a cell to make a protein, the genetic material is decoded and transcribed into messenger ribonucleic acid (messenger RNA, mRNA) by specialized machinery and subsequently translated by ribosomes into proteins.

Proteins are polypeptide chains composed of organic building blocks called amino acids (AAs) residues. The number and order of 20 chemically different natural AAs give rise to different proteins with unique sequences. All AAs share a common backbone constructed of four heavy atoms (Nitrogen: N, Carbon: C, Carbon: Cα, Oxygen: O) that are linked together through peptide bonds formed by the Cα and N. From the chiral Cα atoms, different side chains branch out which give each AA its unique physiochemical properties (hydrophobicity, polarity, charge), size and shape. Interestingly, the 20 AA choices at each position throughout the AA chain lead to an exponential number of sequence possibilities. For instance, a 100 AA sequence has 20^{100} ($\sim 10^{130}$) possible sequence combinations, exceeding the approximated count of atoms in the universe ($\sim 10^{89}$). Despite the colossal sequence possibilities, only a handful exist in Nature, effectively those that have functional importance to an organism [1, 2, 3].

Under standard physiological conditions, an AA sequence collapses or folds spontaneously into a well-defined and stable three-dimensional (3D) structure. The structure is frequently described through a set of translation- and rotation invariant (relative) coordinates. The protein backbone can be defined by three dihedral angles (φ, ψ, ω) describing the global shape and folding path of the sequence. The φ and ψ torsions describe the rotations around (C_{i-1}, N_i, $C\alpha_i$, C_i) and (N_i, $C\alpha_i$, C_i, N_{i+1}), respectively. The ω describes the torsion around ($C\alpha_i$, C_i, N_{i+1}, $C\alpha_{i+1}$) and is constrained to approximately 180 degrees (planar) due to the delocalized electrons between the O and N of the peptide bond, giving a partial double bond character. Therefore, the protein's overall structural flexibility is governed by the existence of energetically favored (φ, ψ)-pairs. Unfavorable (φ, ψ) combinations result from residues in very close 3D proximity inducing strong steric repulsions [4]. Aside from the backbone, the side chain conformations can be described by an additional set of torsional angles ($\chi_1, \dots, \chi_4$) for the rotatable bonds.

The rough shapes of proteins are commonly categorized into four hierarchical levels (Fig. 1.1). The protein's AA sequence represents the “primary structure”. During the folding process, the different portions of the sequence adopt one of the three local geometric sub-structures (i.e. “secondary structure" elements (SSEs)), namely α-helices, β-strands, and loops or coils. α-helices are stabilized by local hydrogen bonds (H-bonds), while β-strands arrange into β-sheets through non-local H-bonds. Contrary, loops are either loosely structured or unordered and generally connect SSEs within the protein. The 3D organization of SSEs composes the protein's “tertiary structure”, stabilized by non-local hydrophobic interactions. Despite often being stable on their own, globular tertiary structures can assemble into “quaternary structures” forming larger multi-domain complexes stabilized by hydrophobic interactions, H-bonds, salt bridges, and disulfide bonds. Each domain in a complex can either function independently of the others or the overall functioning is achieved through the cooperation of the domains. The protein “quinary structure” refers to features of protein surfaces ensuring that proteins interact with specific partners and thereby control and organize the cellular milieu.

The main thermodynamic driving force of the protein folding process is the hydrophobic effect. During folding, hydrophobic side chains are pushed into the core away from the aqueous environment reducing the hydrophobic-to-water contacts. Polar interactions are thought to be less involved, but often help to detail the correct geometries of protein structures [5]. The side chains in protein cores adopt low-energy conformations called rotamers [6] that induce a dense packing resembling a 3D jigsaw puzzle (compared with more loosely folded aggregates or molten globules) [7]. Furthermore, the SSE types have specific rotamer preferences [8, 9, 10].

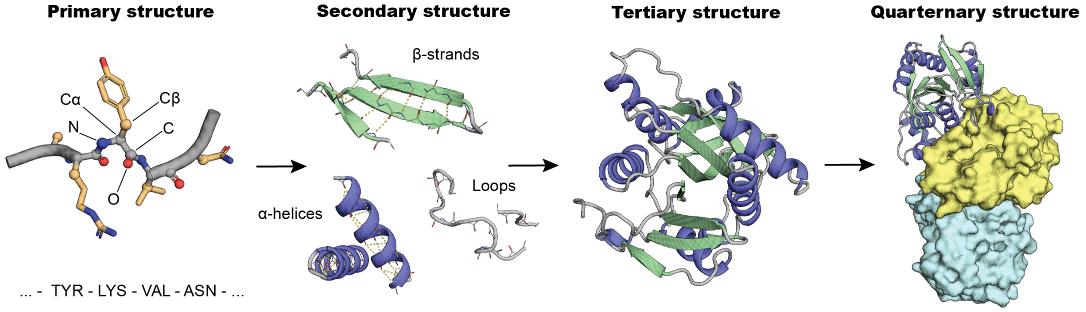


Figure 1.1: Orders of protein structure. The primary structure of a protein corresponds to its AA sequence. The backbone, shown in grey is composed of four atoms (N, $C\alpha$, C, O) and the first atom of the side chains (in wheat) is the $C\beta$. The secondary structure contains three different elements: helices, β-strands and loops. The tertiary structure referst to the spatial arrangement of the SSEs. A quaternary structure is formed upon assembling multiple protein domains.

In 1961 Anfinsen [11] discovered that the native state of a protein structure is fully defined by its underlying sequence (at least for most small and globular proteins) and represents the lowest Gibbs free energy state if the state is kinetically accessible. This dogma, seen from an entropic viewpoint may seem counter-intuitive: because of the very large number of degrees of freedom of an unfolded AA sequence, there exists an astronomically large number of possible conformations — $\sim 3^{198}$ for a protein with 100 AAs, assuming 198 φ and ψ torsion angles that can adopt 3 stable configurations [12, 13]. Folding of the sequence into a single conformation reduces the conformational freedom leading to a large entropic cost estimated to be around 70 kcal mol^{-1} for a 100-residue protein [14, 15, 16]. The entropy-balancing forces are of thermodynamic nature including the hydrophobic effect, H-bonds, van der Waals (vdW), and electrostatic interactions. They are individually weak, and the interplay of many coherent interactions is needed to counter the entropic penalty and stabilize a protein structure. For a protein structure, the change in free energy ΔG between the folded – unfolded state is relatively small of around $\Delta G = 5$ - 10 kcal mol^{-1}, consequently, only a few mutations are needed to disrupt protein stability [17, 18]. Additional aspects may further explain the marginal stability of natural proteins. First, some proteins sacrifice stability for function i.e., protein binding sites often contain exposed hydrophobic residues essential for binding [19]. Second, marginal stability may be important for cell homeostasis i.e., to regulate protein abundance [20, 21]. Lastly, the stability of proteins might be an equilibrium between maintaining essential activities and the accumulation of mutations through evolution i.e., natural selection only optimizes stability if there is a gain in fitness [22, 23].

By probing the protein conformational and energy landscape through folding experiments or simulations, one can show that natural proteins fold in a funnel-like fashion to reach the free energy minimum [25]. The funnel structure of natural proteins is smooth and enables rapid folding and escaping of misfolded states. The smoothness of the funnel is often explained by evolution and backbone optimality i.e., natural proteins have a minimally frustrated backbone structure [13].

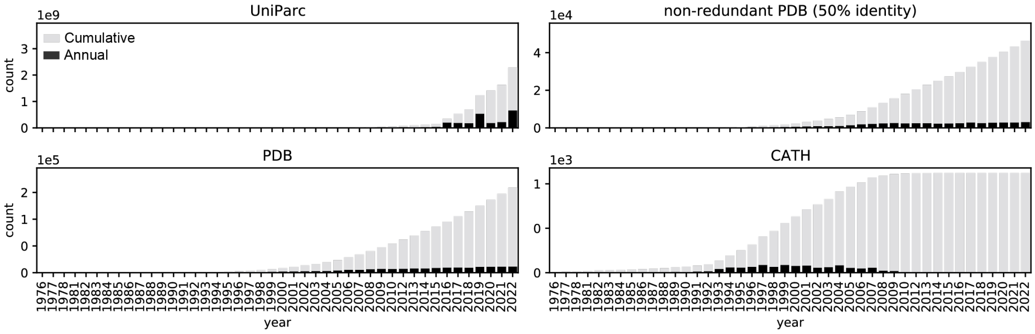


Figure 1.2: Data growth. The annual growth of available protein sequence and structure data. UniParc [24] is a container for non-identical sequences from various different sources. The PDB stores solved protein structure data. The CATH hierarchically classifies structures. Shown is the annual growth of folds, and strikingly only a few new folds have been discovered within the last 10 years.

Taken together, we can recapitulate three general rules for protein folding: (1) hydrophobic AAs are directed towards the core inducing a dense packing, (2) polar AAs are exposed to make favorable interactions with the hydrophilic solvent and fine-tune the geometry, and (3) the global conformation of the backbone is minimally frustrated exerting a smooth folding funnel.

A protein's function arises from its exact structure; for example, scissors-like folded proteins take part in nutrients' digestion while tunnel-shaped ones participate in cellular nutrient uptake and metabolic waste expulsion. Determining the structures of proteins is therefore crucial for understanding and revealing biological mechanisms. While experimental high-throughput sequencing techniques have enabled the discovery of full genomes of various organisms, protein structure determination remains a manual and arduous process. To date, protein structures are commonly solved experimentally through three main techniques: X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy (cryo-EM). Decades of rigorous efforts in solving the structures of proteins collectively yielded a database with over 150,000 structures, stored and publicly accessible in the Protein Data Bank (PDB) [26]. Considering the wealth of the available sequence data [27, 24], one would also expect a large variety of different protein shapes. Surprisingly, the classification of protein structures shows that nature has only evolved a limited set of 1,000 - 10,000 protein folds depending on the definition [28, 29, 30, 31]. Two prominent protein structure classification databases are the Structural Classification of Proteins (SCOP) [32, 33] and the Class, Architecture, Topology, Homology (CATH) [34, 35]. Both follow a similar systematic way of classifying protein structures. Commonly, at the top level, structures are split according to the SSE composition (e.g., mostly α, mixed-α/β, mostly β). Then, the SSE orientations and overall shapes are included to define protein architectures. Architectures can be classified into folds or topologies which respect the same SSE connectivity. Lower fine-level classifications are then mainly based on homology (common ancestor) and/or functional properties.

It is not clear why evolution has only generated a small set of protein folds (Fig. 1.2). An attractive hypothesis is that the evolved set of protein folds suffice any function needed for life [2, 3, 36]. Additionally, recycling existing protein folds for a battery of functions through gene duplication, divergence, and recombination could be a more efficient strategy than completely evolving new folds i.e., by ab initio invention [1]. An illustrative example is the Immunoglobulin-like (Ig-like) fold, found across numerous species as part of different protein complexes that harbor a variety of functions including antibodies for defending the organism against pathogens, cell surface receptors for signal recognition and transduction, and enzymes among others [37].

In view of the enormous space of unexplored proteins and their functional potential to solve vast arrays of bio- and nanotechnological issues, creating protein molecules with controlled geometries for tailored functions has been and remains a grand challenge for modern-day protein designers. Since the 1980s, scientists began to develop methods that allowed them to create proteins beyond the set of naturally occurring sequences and folds with improved or novel biochemical characteristics. Presently, several strategies exist, and the most important classical and emerging protein design concepts will be reviewed in the following sections.

1.1 Manual, minimal, and rational protein design

The field of protein design originated from protein engineering and the study of protein folding. For protein engineers, the goal is to enhance or build new functionalities such as solubility, enzymatic activity, or specific molecular recognition into existing proteins rather than creating new functional proteins from the ground up. Achieving these goals, two successful strategies employed are (1) directed evolution [38, 39], where experimentally mutations are induced to create a pool of variants that is screened for the desired property and (2) rational re-design [40, 41], where residues at specific positions along the sequence are mutated based on prior knowledge from biochemical or structural studies or physiochemical concepts. As both approaches use an initial (natural) protein structure, the stability and dynamics of the conformations remain limited. Importantly, protein engineering can lead to new variants that are in close sequence and structural proximity of natural proteins but is incapable (or, at least very inefficiently) of discovering completely novel proteins [42].

In contrast to protein engineering where the proteins' function was the primary objective, early protein folding studies aimed to understand the sequence-to-structure relationships and how this affects protein folding. To study how proteins fold, protein biochemists figured that it may be possible to understand the mechanisms by doing the inverse of protein folding [43] i.e., design a sequence which lowest free energy shape corresponds to a target backbone (protein design). Examples of the first manually designed proteins were the small RNA-binding peptide in 1979 by Bernd Gutte [44], or the β-sandwich mimicking proteins named “betadoublets” or “betabellins” by Jane and David Richardson towards the end of the 1980s [45, 46]. These designs were designed based on AA propensities i.e., the preferences of AAs to adopt specific SSEs. Nonetheless, the designed proteins came out short on solubility and tended to form aggregates [46]. This highlighted the complexity and the importance of a deep rational understanding of protein structures extending beyond their primary structure.

Subsequently, the first minimalistic approaches of protein design based on sequence pattering led to the deciphering of additional physiochemical features derived from biochemical- and empirical studies. Binary sequence patterns of polar (p) and hydrophobic (h) residues were found to generate small amphipathic α-helices ($[hhphhphhp]_n$) and β-strands ($[hphphphp]_n$) [47, 48].

Rational protein design embraces simple sequence-based rules together with the more complex sequence-to-structure relationships gathered from manual and early bioinformatic inspections of natural proteins. All together laid the foundation of rational design rules that could be implemented into protein design programs. At that dawn of designing proteins, designers specified the protein backbone trace using mathematical (parametric) equations [49, 50] and side chain repacking algorithms were applied to design the sequences [51, 52, 53]. The parametric design was extensively used for the design of helices and helical assemblies (coiled-coils) [54]. The helical system was preferred over other folds due to its symmetries and the design rules that could be efficiently leveraged and assisted in designing the first coiled-coils through designing individual helices first and then associating them [55].

The first successfully computationally from scratch (*"de novo"*) designed and verified protein was α3D [56], a globular three-helix bundle that was extensively analyzed [57, 58, 59]. Afterward, larger helical bundles were designed with geometries and association states previously unencountered in nature [60].

With the first *de novo* proteins, it also became apparent that a sequence not only needs to stabilize the target structure but should also destabilize closely related competing conformations. To achieve this, negative design concepts complemented the current design rules to favor the targeted structure and escape the closely related ones [49, 61, 62].

1.2 Computational protein design

With the increase of the available computational resources and sophisticated software combined with the development of high-throughput and cost-effective gene synthesis, computationally designing full-atomistic protein models at scale, selecting promising candidates and experimentally verifying them was becoming feasible. However, the core challenge of protein design persisted i.e., how to cope with the massive sequence and structure spaces. In fact, protein design is a very hard combinatorial problem (NP-hard i.e., there exist no known algorithm to solve the problem in polynomial time) [63] requiring efficient algorithms in order to explore the possibilities maximizing P(sequence|structure).

To efficiently search over sequences and conformations, modern protein design methods use an initial protein structure and iterate over two basic operations: (1) sampling possible side chain configurations and backbone conformations and (2) applying simplified force fields (energy or scoring functions) to evaluate the sequence-to-structure fitness and approximate the energy (stability) of each model. Through these two operations, protein design methods can traverse a vast space of sequences and conformations towards optimal or near-optimal low-energetic solutions [64].

The sampling procedures can be either stochastic or deterministic. Sampling deterministically guarantees that the solution found corresponds to the global minimum free energy through searching the space exhaustively. One of the most prominent examples of a deterministic sampling algorithm is the dead-end elimination (DEE) that identifies and prunes physically unrealistic and suboptimal combinations of side chain and backbone conformations ("dead ends") without losing the global minimum energy conformation [65]. The DEE algorithm is often combined with the A* search algorithm (DEE/A*) that starts from the pruned, pairwise decomposed energy matrix to find low-energy models such as implemented in OSPREY [66]. Nevertheless, deterministic algorithms are extremely computational and time demanding, especially for large and complex proteins. To overcome these drawbacks, stochastic sampling methods incorporate a random component and search through the space randomly. Because stochastic sampling methods do not consider the full space, they do not guarantee to return the global minimum solution, but rather return ensembles of low energetic solutions. Two prominent methods used are Monte Carlo (MC) sampling and genetic algorithms.

A frequently used framework for protein modeling and design is the Rosetta software [67]. The full-atomistic energy function implemented in Rosetta is based on a weighted linear combination of energy terms that consider geometric degrees of freedom and chemical identities to approximate the energy associated with a protein conformation [68]. The energy function has been parameterized using a collection of small-molecule and X-ray crystal structure data and describes (1) interactions between non-bonded atom pairs important for atomic packing, electrostatics, and solvation, (2) empirical potentials for the modeling of H-bonds, and disulfide bonds, (3) statistical potentials that evaluate the backbone and side chain conformations, and (4) additional scoring terms to capture native protein features. To sample the sequence and conformational space, Rosetta uses MC simulated annealing [69]. The MC algorithm randomly introduces mutations and small conformational movements which are then evaluated an energy function. Simple, a change is accepted if the energy of the model decreases and rejected if the energy increases based on the Metropolis criterion [70]. The drawback of this rigid procedure is the poor exploration of both the sequence and conformation spaces i.e., its "unfitness" to always sample the global minimum and remain stuck in a local minimum. To allow the algorithm to better explore the landscape, a Boltzmann probability is introduced including a temperature factor that enables the tuning of the criterion i.e., the algorithm accepts bad solutions with a certain probability. The temperature factor is high at the beginning (high acceptance and exploration) and is lowered ("cooled") over time so that only energetically favored changes are accepted (i.e., drilling down the found minimum). This technique allows the algorithm to efficiently escape from local minima traps. To date, stochastic sampling methods are often preferred over deterministic ones thanks to their speed and modularity that enable options to include additional design objectives such as specific rotameric libraries, backbone flexibilities, and multistate design [71].

1.3 *De novo* protein design – classical methods

A longstanding goal for protein designers is to create novel protein sequences and shapes *de novo* i.e., absent from the natural protein universe (Fig. 1.3). *De novo* design not only tests our understanding of the underlying physicochemical principles governing protein folding and structure, but also enables the generation of new proteins with targeted and unmatched functions to solve current biomedical and technological problems. Because the sequence space is very large and the main fraction is likely not viable, randomly searching for sequences that will adopt an intended structure or integrate a certain function is quasi-impossible [72, 73, 74]. Also, the experimental determination of a protein structure is difficult, expensive, and time-consuming, and it is therefore impossible to pursue structure determination at scale. Finally, predicting a protein structure from a single *de novo* sequence remains a difficult task [75].

To circumvent this challenge, classical *de novo* protein design methods start with specifying the overall protein's shape. This allows to drastically reduce the search to only sequence fitting the roughly drafted shape. Then, *de novo* design involves two iterative steps: first, backbones satisfying the initial shape constraints are sampled and second, low free energy sequences are fitted onto sampled backbones. Ultimately, the method yields an ensemble of low energetic sequences that are predicted to adopt the targeted shape. In combination with modern experimental techniques (e.g. parallel oligonucleotide synthesis, yeast display screening, and next-generation sequencing), *de novo* designs were successfully screened and tested leading to punctual successes for fully-α-helical-, mixed-α/β- and fully-β-folds, such as TIM-barrels [76], β-barrels [77], β-propellers [78], coiled-coils [79, 80, 60], and repeat proteins [81].

Different strategies exist to define the overall shape of a protein backbone. For symmetric or repetitive folds such as α-helical coiled-coils, parametric functions can be formulated describing the exact placements of the backbone atoms. A prominent example is ISAMBARD (Intelligent System for Analysis, Model Building, and Rational Design), a software that features the integration of less-common SSEs and generalizes parametric modeling and design [82]. Still, integrating structural irregularities found in natural SSEs into parametric frameworks remains difficult. Such local structural irregularities include bulges in β-strands, breakpoints in α-helices, or 3_{10} helices. Examples of two popular methods that can introduce structural irregularities are fragment assembly [83, 84] and SEWING (Structure Extension with Native-substructure Graphs) [85]. Fragment assembly uses small protein fragments of sizes 3 and 9

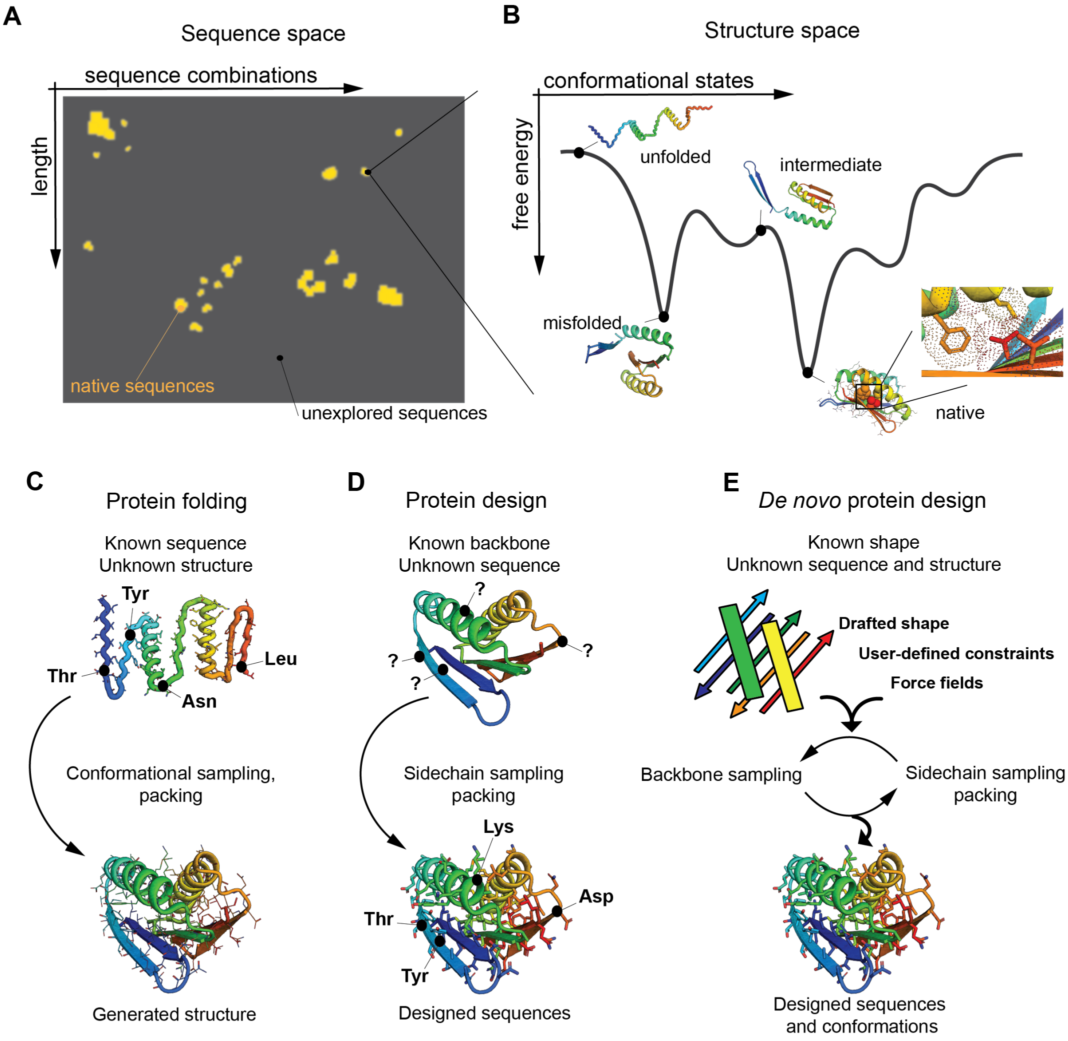


Figure 1.3: Computational protein design. **A:** Nature only sampled a tiny fraction of the possible sequence space. **B:** The conformational energy landscape of a sequence is funnel-shaped and contains multiple local minima and a global minimum corresponding to the well-folded, native structure. **C:** The task of protein folding is to predict the 3D structure from the AA sequence. **D:** The task of protein design is to predict sequences whose lowest energy structure corresponds to the target backbone, and can be seen as the inverse folding problem. **E:** For *de novo* protein design, both the exact sequence and structure are unknown. Only the overall shape of the target protein structure is specified. To search for potential sequences obeying the drafted shape, *de novo* design iteratively samples backbones and fits low-energetic sequences onto them, resulting in an conformational and sequence ensemble.

(3mers and 9mers) derived from the natural repertoire matching the target backbone locally and stitches them together to create a protein structure. The assembling is stochastic and guided by a scoring function to decide if the insertion is accepted or rejected. To scan the conformational and energetic landscape well and find low potential conformations, thousands of fragment and structure energy minimization moves are required which is computationally expensive and time-consuming. On the other hand, the SEWING method combines natural SSEs with their local irregularities through their regular regions and thereby avoids computationally expensive loop closures.

Despite the multiple triumphs, *de novo* design remains challenging [74, 77]. Many computationally designed proteins require multiple rounds of experimental and computer-guided opitmizations [16, 74]. Especially mixed-α/β- and fully-β-proteins are notoriously difficult to design [77, 86, 87, 88]. For complex protein folds, many specific and manual adjustments to the *de novo* design protocol are required, depending on the targeted protein fold and the design objective.

A major limitation is the crafting of "designable" protein backbones [89, 90, 91, 92, 93] i.e., backbones that are strain-less and physically realistic. For backbones to be designable, they need optimal SSE configurations with favored tertiary structure symmetries such that a well-packed core is realizable with the available SSEs [94, 95]. Bioinformatic studies showed that natural protein backbones differ in terms of their designability e.g., some backbones accommodate a much larger pool of different energetically favorable sequences than others. This has been linked with mutational robustness increasing thermodynamic and evolutionary stability [90]. Today, the sequence capacity of a backbone is often used to quantify its designability [89, 90, 95], however this metric remains elusive and difficult to interpret physically and chemically. Designability includes other factors that are difficult to measure such as fold specificity [89, 96], or native-like SSE arrangements [97]. Backbone designability necessitates the embedding of the local protein structure patterns and irregularities in the global structural context. While the local structural features have been well described [98], empirical design rules connecting these local patterns to tertiary structure elements (such as β/β-, β/α-, and α/β-units) and domains flourished during the last 15 years (especially for mixed-α/β and fully-β tertiary motifs). For example, Koga and colleagues [99] formulated a first set of rules to design "ideal" proteins. Here, ideal refers to globular proteins, without irregularities in their SSEs and small loops as connecting elements. The rules are based on loop lengths that induce the correct tertiary arrangements and packing. Since then, the rules have been steadily updated [100, 101] e.g., with structurally defined loops to bridge non-local motifs [100, 87], using β-strand register shifts and β-bulges to control the curvature and to carve cavities into proteins [76, 86], or strategically placed glycine residues that relieve backbone stress and allow the design of β-barrels [77].

1.4 Emerging *de novo* design methods

Technological advancements and the exponential growth of available protein sequence and structural data led to the development of methods capable of "learning" complex relationships of underlying physicochemical characteristics. A variety of deep learning (DL) methods borrowed from computer vision and natural language processing have substantially contributed towards improved understanding of the protein sequence-to-structure relationship and accelerated computational protein design (Fig. 1.4) [102, 103, 104].

Usually, a DL method is composed of multiple layers of artificial neural networks, often referred to as deep neural networks (DNNs). One can think of a single neural network layer as an interconnected stack of nodes that act as small functions and can communicate between themselves. When injected with data, each layer progressively integrates and propagates the processed signal further to the next layer. The connections between nodes within a layer are weighted and depending on the importance of the signals, the weights can be modulated. A successful adjustment of the weights gives DNNs the ability to internally learn meaningful representations called "embeddings", condensed numerical descriptions that are understood by DL methods and can be leveraged to reveal and better understand complex relationships within the data. The adjustment of the weights is achieved through the "training" procedure that requires a large set of diverse samples. A batch of samples is iteratively fed into the DNN and after each iteration, the DNN calculates the error (loss) of its own prediction with respect to a criterion. The error can be backpropagated through the DNN and used to derive the gradients with respect to the weights, and then readjust the weights slightly in the direction of the steepest descent. Following this scheme, the DNN iteratively minimizes the errors of the predictions. Ultimately, successfully end-to-end trained DNNs can be utilized to solve a particular task given new data such as predicting the protein structure from a sequence that was not included in the training samples [105, 106, 75, 107, 108].

An exciting field of DL is deep generative modeling where DNNs are used to approximate the underlying high-dimensional, complex distributions of the data. Specifically, generative frameworks can be used to infer the likelihood of data and then to generate new, artificial samples from the learned underlying distribution. Prominent examples are deep fakes that can generate realistic appearing fake portraits of celebrity images or text with particular emotions or humor [109].

Based on the impressive examples from computer vision and natural language processing, it sounds appealing to use deep generative modeling for *de novo* protein design i.e., creating artificial proteins based on the available protein sequence and structure data. Popular generative approaches include normalizing flows (NFs) [110], generative adversarial networks (GANs) [111], and variational autoencoders (VAEs) [112]. The latter is of particular interest owing to its flexibility to solve a wide range of different problems.

VAEs consists of an encoder, a decoder, and a specific loss function. The encoder is a DNN that compresses the input data as a distribution over the DNN internal - latent representation with the goal of losing the least information possible. The decoder then de-compresses the latent representation back into the input as best as possible. The degree of compression is controlled by the size of the latent space, while the compression quality is dependent on the depth of the DNNs. In practice, the samples are encoded as Gaussian distributions defined by a set of means (μ) and variances (σ^2) that can be then used to generate latent variables and reconstructed through the decoder to return the inputs. During training, a reconstruction loss is used to measure the error between the input and the prediction. Additionally, a regularization term is applied to the latent space to ensure an approximate standard Gaussian and well-behaved latent space structure. Hence, the latent space should be continuous and complete i.e., the μ for each of the latent variables should be “close” together with “overlapping” σ^2. Without any regularization, the encoder tends to predict tiny σ^2 and large μ leading to an overfitted latent space that is unable to generalize to new, unseen data. Usually, the regularization is formulated as the Kulback-Leibler divergence between the predicted distribution and a standard Gaussian, forcing the predicted distribution to have unit variance ($\sigma^2 = 1$) and centered means ($\mu = 0$). A well-trained VAE can interpolate across the continuous and complete latent space between variables and thereby sample novel data samples.

1.4.1 Structure prediction DNNs

In December 2018, a DL-based program called AlphaFold1 (AF1) released by DeepMind [105] won the 13th Critical Assessment of Techniques for Protein Structure Prediction (CASP) competition. AF1s’ rationale is that a set of evolutionarily related sequences (multiple sequence alignment (MSA) holds covariation statistics that can be leveraged to predict distance distributions between atoms and subsequently used to guide the computational structure predictions. Usually, for contact or distance predictions, the $C\beta$ atoms are used as they are the first atoms of the AA side chains and the $C\alpha$-$C\beta$ vectors determine whether a residue is pointing towards or away from the protein core [113, 114, 115, 116]. The AF1 model uses a series of two dimensional (2D) convolutional blocks transforming an encoded sequence and its MSA features into pairwise distance- and dihedral distributions. The distributions are constructing a protein-specific potential of mean force [117] i.e., a function that describes the free energy changes when the distances between two residues changes. The potential is then minimized via gradient descent to produce low-energy structural models.

Shortly after, the transformed-restrained Rosetta (trRosetta, or short trR) [106] was developed, an improved version of the AF1 model. Importantly, trR predicts inter-residue orientations in addition to the distance probabilities. The pairwise orientations between two residues i and j include three dihedral angles (ω, ψ_{ij}, ψ_{ji}) and two angles (φ_{ij}, φ_{ji}). The ω torsion is defined as the rotation around the virtual axis connecting $C\beta_i$ and $C\beta_j$ (through the four atoms ($C\alpha_i$, $C\beta_i$, $C\beta_j$, $C\alpha_j$)), and is symmetric. The ψ and φ orientations are asymmetric e.g., the orientation depends on the reference frame either at residue i or j. The ψ torsion is calculated using the atoms (N_i, $C\alpha_i$, $C\beta_i$, $C\beta_j$) and the φ angles using the atoms ($C\alpha_i$, $C\beta_i$, $C\beta_j$) in the case of setting the reference to residue i. The predicted inter-residue geometric potentials are converted into energy-term restraints guiding the Rosetta minimization protocol. trR was shown to be able to predict protein structures for a set of *de novo* designed sequences, elucidating its potential to evaluate novel protein sequences where no or only shallow MSAs can be constructed. The architectures of AF1 and trR are heavily inspired by DNNs from computer vision. Hence, they rely on a grid-like data representation and require additional programs to convert their 2D predictions into 3D structures.

The second version of AlphaFold (AF2) [108] was revealed in the CASP14 competition and achieved near-experimental accuracy for a majority of modeling challenges. The AF2 architecture was completely revised and is fully end-to-end i.e., the DNN predicts the 3D coordinates of all heavy atoms from an input sequence and its aligned homologs. Briefly, AF2 contains two main transformer-based modules that are employed sequentially. The first module (termed "Evoformer") is based on stacked equivariant transformer-type layers and uses the raw MSAs together with homologous structures or the self-distilled models (predicted models with high confidence), and then returns the processed MSA embeddings (m_{si}) and residue pair features (z_{ij}). Importantly, the Evoformer continuously mixes and synchronizes information from the MSA and pair representations enabling the discovery of spatial and evolutionary relationships between sequences. The MSA features are gated by a row- and column-wise self-attention (row acts on the sequences and columns on residues). The residue pair feature tensor is processed through a "triangular" attention-biasing operation restraining residue triplets to fulfill the triangular inequality. In general, the MSA features influence the pair features throughout the module whereas the pair representation softly directs the attention to the MSA features that distill evolutionary couplings. The second AF2 "structure" module uses the first row of the MSA feature (m_{i1}) which is the embedded original sequence, and the full pair embeddings from the Evoformer to generate 3D structures. This is achieved using 8 blocks of invariant point attention layer that predicts relative rotations- and translations for residues formatted as rigid-body frames. The side chain conformations (χ angles and atom positions) are predicted at the end of each block with a separate small module. Considerably, during the structure module, the chain structure is not imposed at any time to allow simultaneous local and global refinements. Feeding back the predicted outputs into the network ("recycling") improves the structure prediction accuracy. The network is supervised by a final frame-aligned point error loss (FAPE), and several auxiliary losses (structure violation loss and side chain loss, MSA BERT-like loss, distogram cross-entropy loss) are used at different stages during training to force the network to learn a meaningful and geometrically correct structure within the first few blocks of the DNN.

Taken together, the main improvements in AF2 are: (1) directly using the raw MSAs rather than starting from MSA-derived features such as covariance-inversion or pseudolikelihood models, (2) novel efficient attention layers throughout the model that mixes and captures important geometric information, (3) a heavy atom 3D structure generation module and (4) end-to-end differentiable learning with recycling iterations to refine predictions. Since AF2s initial communication multiple groups have either implemented related frameworks such as RoseTTAfold [107] or slightly modified AF2 to solve various biologically relevant tasks such as for predicting protein-protein interactions and peptide-protein interactions [40, 118] and improved X-ray and EM density model fitting [119, 120]. Furthermore, AF2 was used to predict models for protein sequences of almost the entire human proteome (98.5% of human proteins) [121].

The immense success of DL in protein structure prediction from sequence raises the question of whether these methods can also be used for the inverse task e.g., protein design, or even for *de novo* protein design. Many sequence-based DL methods for protein engineering have been developed [122, 123, 124, 125]. These methods focus on generating improved or novel functional variants without considering structural features to discover new protein topologies or folds.

1.4.2 Neural generation of protein backbones for fold discovery

An infinite number of representations of a protein backbone through cartesian coordinates exists by translations and/or rotations in 3D space. Due to the enormity of the space, it would be challenging for brute-forcing a DNN to learn all possibilities and thereof generate new ones. Hence, for DNNs to efficiently learn protein backbone representations, coordinates invariant to translations and rotations are needed. Two invariant representations of protein backbone are the (1) dihedral coordinates (φ, ψ) assuming ideal bond geometry, and (2) Z-matrix where an atom position a_i is relatively described by its three previous atoms in the chain (a_{i-1}, a_{i-2}, a_{i-3}) through a distance (a_{i-3}, a_{i-2}), an angle (a_{i-3}, a_{i-2}, a_{i-1}), and a torsion (a_{i-3}, a_{i-2}, a_{i-1}, a_i). While sequential invariant encodings have been used for protein structure generation from sequence [126, 127], they suffer from the integration error i.e., small errors along the sequence add up and lead to large deviations downstream the chain.

To cope with the integration error, the protein backbone geometry can be arranged into 2D representations. A frequent 2D encoding of molecules are distance matrices, and for proteins often either Cα or Cβ atoms are considered. Additional 2D feature maps exist such as the inter-residue orientation maps introduced in the trR framework. Not only do 2D representations break the integration of errors along the chain, but they also allow the use of 2D convolutional neural networks (CNNs). A 2D feature map can be thought of as an image, where its pairwise features are color channels. Deep 2D CNNs can efficiently identify secondary and tertiary structural patterns that are otherwise difficult to describe through conventional heuristics. Using 2D invariant encodings, for example, GANs have been trained to in-paint protein loops [128] or infer missing residues [129]. Similarly, VAE have been used to create protein structures [130].

Both, 1D and 2D representations typically need an additional step to recover the 3D cartesian coordinates of a structure. This has been difficult due to potential errors or invalidities within the representations. For example, cartesian coordinates can be recovered from a valid (perfect) Euclidean distance matrix through multidimensional scaling that uses the top three components of the Eigendecomposition of the Gram matrix [131]. However, for degenerate and low-resolution distance matrices with systematic noise such as generated by DNNs, multidimensional scaling is often prone to fail. To circumvent the 3D recovery issue, an additional 2D to 3D translation DNN was coupled to a GAN generating protein backbones to guide the generation process towards valid distance maps [132]. DNNs capable of directly generating 3D coordinates were also developed including a DNN trained on a data corpus of Ig-like domains with specific loss functions designed to preserve geometric correctness (IG-VAE) [133]. The IG-VAE can interpolate in the learned latent space and decode new Ig-like domains.

1.4.3 *De novo* sequence generation conditioned on structure information

Instead of following the standard *de novo* protein design framework where a backbone is first generated and subsequently designed, DL methods simultaneously incorporate the inputted structural information and generate sequences fulfilling the structural constraints. For example, a GAN has been developed to generate fold-specific sequences [134]. To achieve this, the sequence generation of the GAN is accompanied by two additional GANs, one classifies the generated sequences to fold families and the other predicts whether the sequences are alike sequences from the natural repertoire. The method enabled the design of diverse and novel fold-specific sequences with good overall predicted biophysical and biological properties. A second example is the "ProteinSolver" which uses a template fold to extract geometric constraints and then generates sequences that accurately obey those [135]. According to several computational metrics and experimental validations, the selected sequences adopt the intended fold [135]. Similarly, the "Structure Transformer" uses an encoder-decoder pair to autoregressively transform relative structure-derived node and edge features into meaningful embeddings to then generate structure compatible sequences [136]. The inputs to this framework can be relaxed to soft constraints such as a few contacts or H-bonds of the backbone which allows certain backbone flexibility and a more diverse sequence generation. Lastly, fitting sequences on a novel protein fold were generated through iteratively sampling over a VAE trained 4,000 structures and their sequence homologs conditioned through a rough topology description (SSE localization, direction, and length) [137]. The fold specific sequences threaded onto a models' backbone were shown to be stable for short molecular dynamics simulations (200 ns) [137].

1.4.4 Inverting protein structure prediction nets for *de novo* design

DNNs predicting structure from a sequence have become state-of-the-art. Thanks to modern graphical processing units (GPUs), these DNNs are extremely fast, enabling the prediction of a single structure from a sequence in seconds to minutes. Interestingly, the structure prediction engines can be "reversed" and used to design novel sequences. Several applications have shown that such structure prediction engines learn sequence-to-structure sufficient information to fabricate new sequences that adopt a well-folded confirmation as well as being unrelated to the naturally occurring pool of sequences and structures.

For instance, trR was used to hallucinate novel proteins [138]. Starting from randomly generated sequences, they were pushed through trR to predict initial diffuse and blurry distance and orientation maps. The predicted distributions were driven towards realistic distances and orientations by sampling AA substitutions that would maximize the contrast between the random (background) and the predicted maps. Similarly, trR was also utilized for fixed backbone design via backpropagating the errors between the predicted and the target maps to the sequence and specifically guide the optimization in the sequence space [139]. This procedure has the advantage of implicitly optimizing over the full sequence and structure landscape. This means that positive and negative design are linked, and thus the method searches for the lowest-energy sequence while maximizing the probability of the target structure relative to all other conformations.

By combining the hallucination and fixed backbone design strategies, partially guided hallucination for the stabilization of discontinues structural motifs was achieved with the trR, RoseTTAfold, and AF2 DNNs [140, 141]. To achieve this, a composite loss was minimized ensuring the recapitulation of the structural motif and its embedding into a hallucinated, well-structured scaffold. Particularly AF2, as it explicitly predicts all heavy atoms, a problem-specific coordinate loss was added to enhance the interactions with the target protein. This includes one term ensuring that the motif is surface exposed and another term constraining the rotamer configurations of the interacting residues of the motif.

Also AF2 can efficiently be used for fixed-backbone protein design. In a first approach, an autoregressive generative model was trained to recover masked AA on a large set of *de novo* designed sequences [142] and then AF2 was used to predict structural models for top thousand diverse sequences. They created sequence-structure pairs were exploited to bias the initial sequence for the AF2 based backbone design task i.e., the target backbone is compared to the structure models from the database, the best matches are retrieved, and a starting sequence is generated. The sequence is then optimized by iteratively mutating a couple of residues in the sequence, prediciting an AF2 model that is compared to the target backbone, and then either accepting or rejecting the mutations. Another similar approach uses an evolutionary algorithm that continuously optimizes and diversifies the input sequence pool scored by AF2 through generating structural models and leveraging confidence measures as well as other structural metrics till reaching the target structure [143]. This pipeline enabled designing monomers of various sizes, multimers, conformational switches, and protein binders that showed comparable quality with respect to several orthogonal computational metrics including molecular dynamics and Rosetta folding simulations [143]. Recently, RoseTTAfold was modified and retrained to jointly recover missing sequence and structure information [141]. This approach termed RoseTTAfold-joint (RF_joint) alleviates the need of iterating over the predictions and enables the completion of structures with a single forward pass. However, for RF_joint to perform properly, sufficient structural and sequence context are required *a priori*.

The introduced protein design methods have not only been for the generation of new sequences folding into a particular structure, but also to design for specific functions. Several functional protein design frameworks will be explained in the next section.

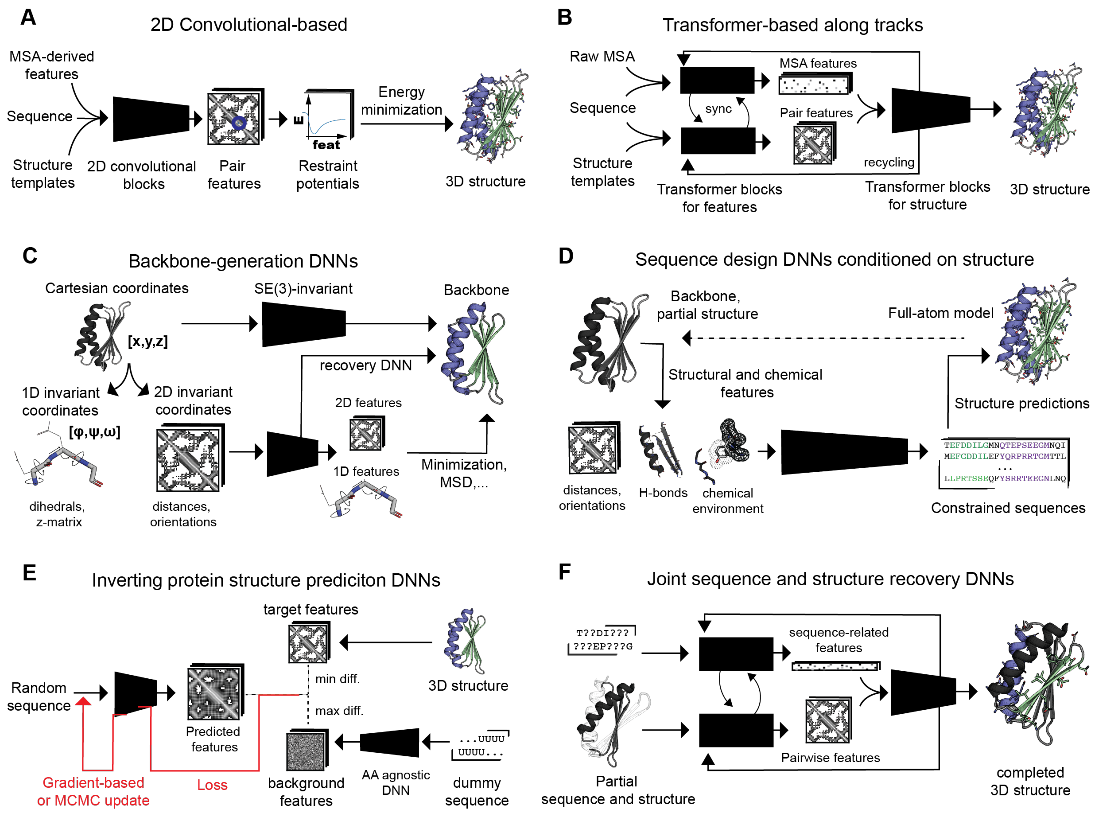


Figure 1.4: DL-based *de novo* protein design methods **A:** AF1 and trR are based on 2D convolutions to predict a structure from a sequence and MSA features. **B:** AF2 and RoseTTAfold are transformer-based frameworks with multiple tracks enabling storing and mixing information efficiently. Importantly, the methods contain invariant DNNs to predict the structure/backbone from the embeddings. **C:** DL-based methods that learned to generate designable protein backbones. If directly used on cartesian coordinates, a SE(3)-invariant DNNs is needed. However, often the cartesian coordinates are converted to translational -and rotational invariant representations and used with standard DNNs. **D:** DNNs can predict sequences that fit structural prerequisites. The sequences are validated by structure prediction tools to confirm that the encoded structure is correct. **E:** Sequences can be generated for a target backbone by iterating over a structure prediction DNN and using the gradient to maximize the probability of the sequence to fit the target backbone and simultaneously disfavor off-target conformations. Similarly, instead of using a target backbone, random background outputs can be used to drive the sequences to be realistic and have well-packed structures (hallucination). **F:** Protein structure prediction DNNs can be modified to recover masked sequence and structure regions, and thereby predict sequences and structures in forward pass.

1.5 Protein design for biomolecular recognition and signaling

Cells are densely packed with organic molecular matter. The packing is highly coordinated, with molecules undergoing permanent or transient interactions that induce short- or longlasting signaling cascades. These signaling networks underpin cellular activities from movement to division as well as cell-to-cell recognitions. Creating synthetic proteins to specifically perturb protein-protein interactions (PPIs) and thereby master biological activities is of significant interest to a wide range of applications in biotechnology and bioengineering.

Natural PPIs are often exquisitely specific i.e., the interactions occur at specific localized sites on both partners' surfaces. Inspections of native interfaces reveal a high-shape complementary and that approximately 1600 $Å^2$ of the solvent-accessible surface area is buried upon complexation [144, 145]. The strength (binding affinity) and length of interaction can vary between different interacting protein pairs [146]. Both, binding affinity and duration are dependent on the environment e.g., temperature, pH, ionic strength, or post-translational modifications. Biophysically, the binding affinity can be formulated as the difference between the free energy of the complex (bound state) relative to the free energies of the unbound states ($\Delta\Delta$G). Important thermodynamic contributors of binding are hydrophobic and electrostatic interactions, H-bonds and salt-bridges [147, 148]. Also, other factors such as conformational dynamics (induced-fit), coordinated waters along the interface, and overall entropy indirectly affect the proteins' binding [149].

The binding affinity can be quantified by the half-life of the complex (K_D) through factoring the association (k_{on}) and dissociation (k_{off}) rates. It has been seen that association rates are more strongly influenced by electrostatic interactions [150]. In fact, the specificity is often incorporated by electrostatic interactions which require precise geometric orientations (Coulomb and polarizartion can be strongly directional, while H-bond are weak directional [151]) of the involved side chains and charge complementarity. Hydrophobic interactions tend to be less specific and more forgiving regarding the geometry. Taking into account that many interactions occur in aqueous solution, hence the formation of H-bonds and/or salt-bridges require a dehydration/rehydration process that involves the transitioning over energetically unfavored states which can weaken or decelerate the binding [152].

With aforementioned computational protein design strategies at hand, multiple frameworks have attempted to tackle the design of PPIs at scale [153, 154]. However, even state-of-the-art computational tools lack the required precision resulting in frequent experimental failures [153, 154]. Current scoring functions excel at positive design capturing interactions such as vdW or H-bonds, but inherently lack negative design to avoid the myriad of off-target conformations [155, 156, 157, 62]. Most of specific binding proteins have been fully created or optimized through *in vitro* screening and selection of antibodies and specialized protein scaffolds such as ankyrins and fibronectin domains [158, 159, 160]. The successful computational design of protein binders requires tight feedback loops between experimental optimization and screening strategies (including site-directed/site-saturation mutagenesis and deep sequencing, yeast display, and fluorescence-activated cell sorting (FACS) technology). The successes and failures uncovering essential concepts of PPIs show that a robust computational protein design framework for generating PPIs remains elusive. In the following sections, important PPIs design strategies will be introduced.

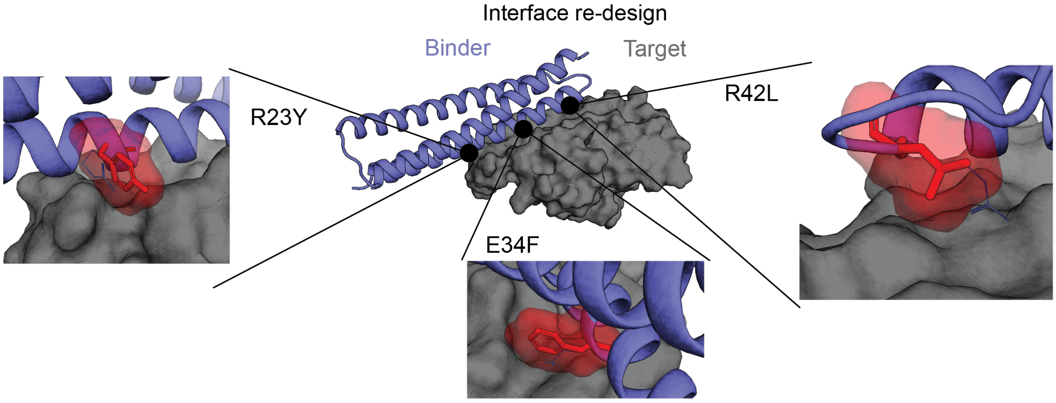


Figure 1.5: Protein interface re-design. The re-design of a binders' interface (purple) with predicted point mutations (red) to increase the affinity to the target structure (grey).

A simple approach to PPI design is to re-design interfaces of existing PPIs and to thereby enhance their binding affinity or modulate their specificity (Fig. 1.5) [161, 162]. Yet, this strategy is based on the existing two protein partners and therefore lacks control over the size and overall shape of the binder. Thus, it is more desirable to create binders *de novo* controlling their fold and function. A successfully used *de novo* PPI design method is grafting the interacting side chains of the natural binder with or without the underlying backbone segment onto another protein that acts as a "stabilizer" (scaffold) [163]. However, grafting is inherently limited to known interactions and cannot be used to target new protein sites from scratch. To address this issue, "hotspot-centric" design methods (Fig. 1.6) have been established such as inverse rotamers [164] or docking of disembodied side chains [165]. Interaction hotspots [166] are residues within the interface that provide a disproportionate amount of the binding energy and upon mutation of these hotspots to alanine, the binding affinity usually drops by several orders of magnitude [167]. Interaction hotspots have been observed within many PPIs and are energetically favorable and evolutionary more conserved than other residues of the interface [168, 147, 148, 169]. Essential is the formation of conformationally restricted hotspots through a dense interaction network involving surrounding side chains and thereby disfavoring alternative binding modes [170].

The main idea of hotspot-centric design is as follows: (1) a high-resolution steps attempts to create high-affinity interactions at the core of the interface by disembodied residue docking; (2) an independent low-resolution stage docks coarse-grained proteins onto the target sites to search for high-shape complementary conformations that can engage the site without major steric clashes; (3) the results of the two steps are combined through transferring as many hotspots as possible onto the docked conformations. The residues surrounding the hotspots are refined to collectively stabilize the defined hotspot configurations and contribute to the binding affinity.

The third step can be challenging as there is no guarantee that the backbones can well embed the hotspots. To optimize the embedding of the hotspots, several strategies exist. An example

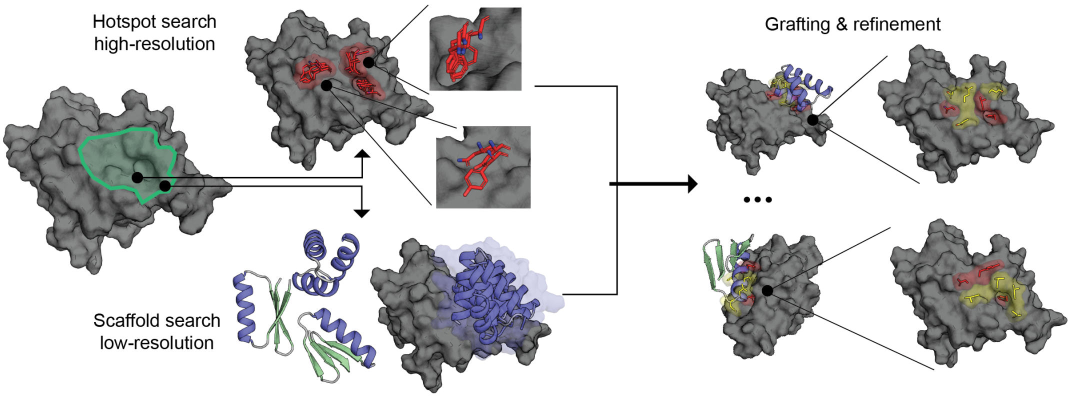


Figure 1.6: Hotspot-centric *de novo* PPI design. Hotspot-centric design for a particular surface site on the target (grey) starts with a high-resolution search for disembodied residues (red, hotspots) with high shape- and electrostatic complementarity. This is achieved through methods such as rigid-body docking or inverse rotamer searches. A low-resolution protocol docks scaffolds (green, purple) to retrieve potential binding conformations. Finally, the clustered hotspots are grafted onto the docked scaffolds and interface residues neighbouring the hotspots are refined and optimized (yellow). The figure is inspired by Gainza *et. al.* [171].

is the folding-and-design algorithm (FunFolDes (FFD)) that optimizes the embedding of the grafted hotspots into the scaffold and towards the target by computationally refolding and redesigning the scaffold while keeping the hotspots fixed and thereby relieving potential clashes or backbone strains [172]. Recently, the Rotamer Interaction Field (RIF) docking method was proposed [77]. The RIF algorithm seeds the interface with a large ensemble of discrete AAs that form H-bonds, hydrophobic interactions and favorable vdW contacts with the targeted interface. Subsequently, scaffolds are docked into the AAs cloud and hierarchically searching for favorable conformations capable of jointly harboring all necessary AA rotamers.

Despite impressive successes [154, 173], major limitations remained including: (1) the exact surface site of the desired target needs to be specified *a priori*, and identifying amenable ("targetable" or "druggable") binding sites remains elusive and often evolutionary or mechanistic insights are necessary; (2) the search for hotspot residues is challenging due to inaccuracies in energy functions and stochastic sampling; (3) the placement of the hotspot residues onto a scaffold is notoriously difficult, and oftentimes, there no natural scaffold that could stabilize the placed hotspots adequately exists [174, 175].

Recently, Gainza and colleagues developed a geometric DL framework called MaSIF (molecular surface interaction fingerprinting) to efficiently tackle the above-mentioned problematics of PPI design [171, 176]. Instead of using an atomistic representation of proteins, MaSIF operates on the high-level molecular surface. A proteins' surface is traced using the contacts between a "rolling" spherical probe over the molecule [177, 7]. Multiple discretized descriptions of surfaces exist e.g., point clouds, voxel, or graph representations [178, 179, 176]. Geometrical- and physiochemical features can be mapped on these representations [178, 180]. Using

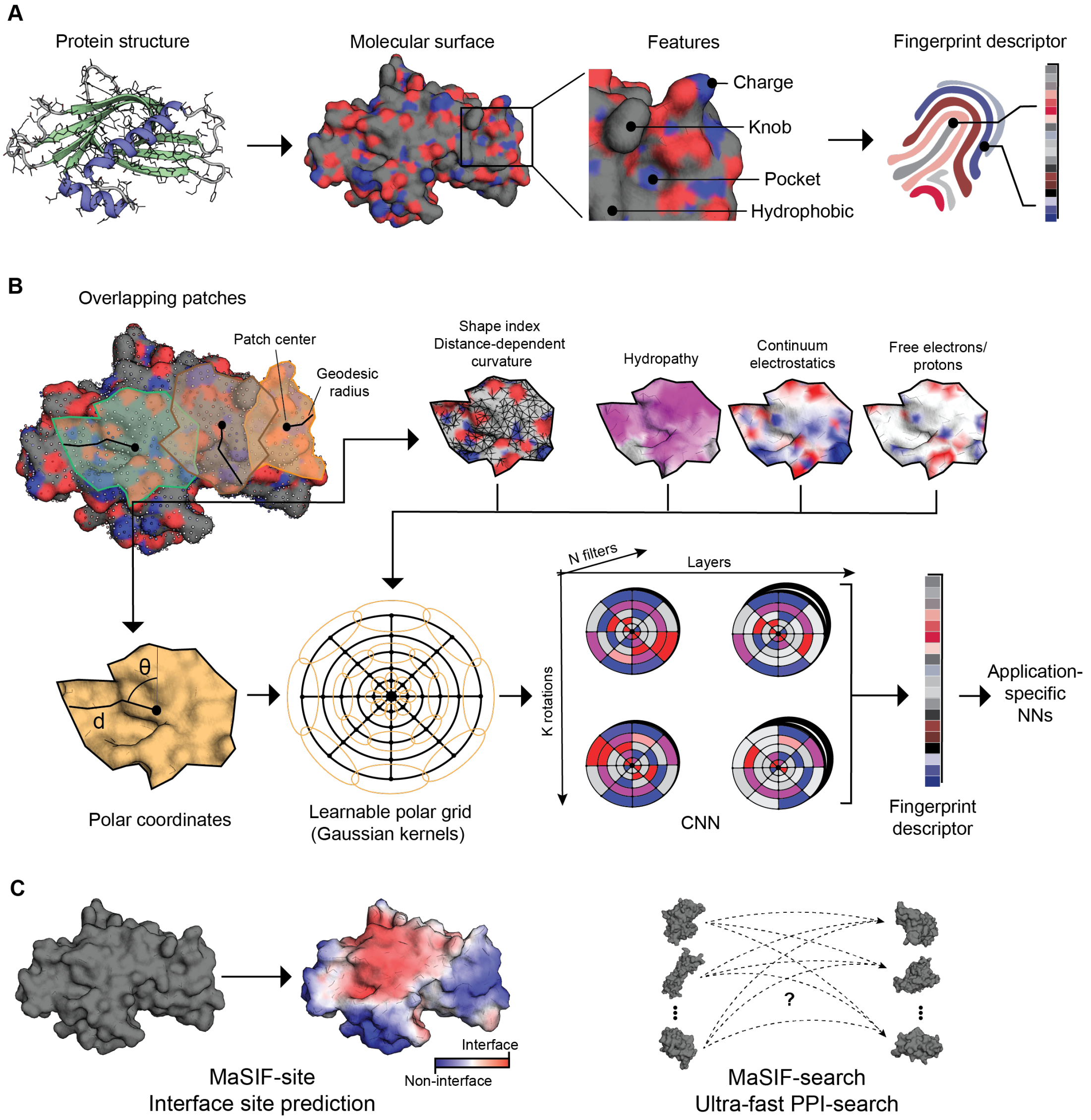


Figure 1.7: MaSIF framework and applications. **A:** MaSIF operates on the molecular surface representation of proteins that show distinct geometric and chemical features "fingerprinting" each surface regions (patches) of the surface. **B:** MaSIF geometric overview. First, the surface gets divided into multiple (usually > 36) overlapping patches of 9 Å or 12 Å geodesic radii. The patches are described geometrically through polar coordinates including the radial coordinate (geodesic radius) and an angular coordinate (angle theta) that are "sooften" using a mulivariate Gaussian kernel with learnable parameters (learnable soft polar grid). Chemical and shape-derived features are mapped onto the polar grid and a geometric CNN is used to distill the information into application-specific fingerprint descriptors (vectors). **C:** MaSIF-based applications include interface site prediction and ultra-fast PPI searches that is similar to rigid-body docking. The figure is inspired by Gainza *et. al.* [171].

surfaces, the solvation energies of proteins, catalytic rates of enzymes, or similarities between protein pockets shapes were studied [178, 181].

The central hypothesis of MaSIF is that surfaces displays important chemical and geometric features thought to prime and fingerprint interaction sites; and proteins participating in similar interactions share similar fingerprint descriptors that are physiochemically grounded and free from any evolutionary history (Fig. 1.7A). However, the fingerprint patterns can be difficult to recognize, especially by eye. To distill the geometric and physiochemical information stored in protein surfaces, MaSIF uses geometric DL (Fig. 1.7B) [182, 183]. First, MaSIF decomposes a protein surface into multiple overlapping patches with a geodesic radius of 9 Å or 12 Å, this is the distance "walking over the surface". The number of patches and geodesic radius varies depending on the needs of the specific application. For each patch a mesh is created and the nodes are described geometrically by a polar coordinates i.e., the geodesic distance and an angular coordinate (Θ). The Θ-angle is calculated by scaling the 3D surface into 2D through a multidimensional scaling and then choosing a random direction to calculate the angle. Each set of coordinates is fed through a multivariate Gaussian with learnable parameters (μ, σ^2) yielding a learnable soft polar grid. Multiple features such as the shape index or hydropathy are then mapped onto the soft polar grid. Because the angular coordinate has been chosen based on a random direction, the feature-loaded polar grids are rotated and passed through a CNN. The maximum activating rotation is chosen for each of the features and a final layer combines the information into a single fingerprint descriptor. Note that these fingerprint descriptors are not universal, but depend on the optimization objective towards a particular task.

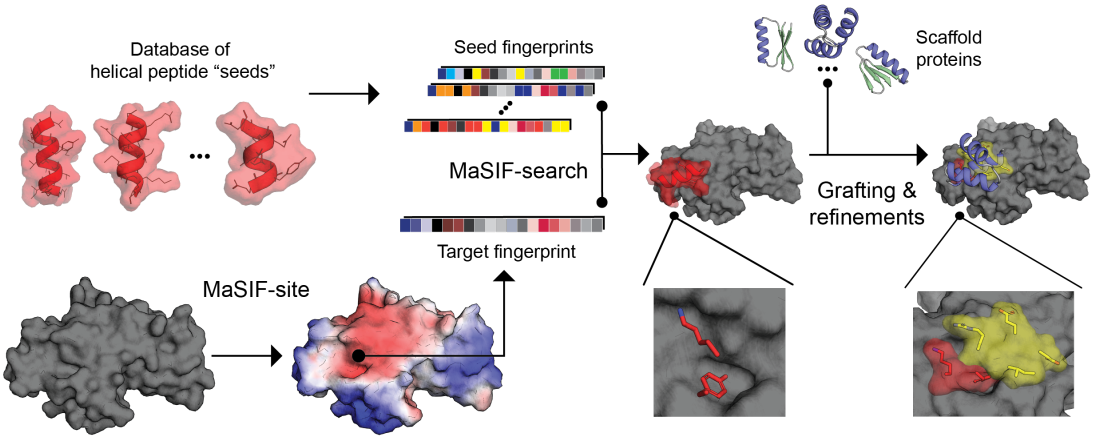


Figure 1.8: MaSIF-based *de novo* PPI design framework. **A:** The MaSIF-based *de novo* PPI design framework uses MaSIF-site to predicts interface sites on protein surfaces, and generates target fingerprint descriptors. Employing a preprocessed library of protein or peptide surface fingerprint descriptors acting as "seeds", MaSIF-search retrieves to the target complementary seed fingerprints and aligns them onto the target interface. Using standard grafting techniques, the seed hotspots (red) are optimized or transferred onto larger proteins to confer stability (yellow).

Following the conceptual MaSIF framework, two deep DNNs have been trained for the PPI design task (Fig. 1.7C): (1) A geometric DNN was trained to differentiate interface regions from non-interface regions (MaSIF-site) for efficient identification of potential interface sites, and (2) once interfaces are selected, a second DNN is trained to perform ultra-fast searches for complementary descriptors against the selected interface descriptor and return high-complementary fragments (MaSIF-search). Combining MaSIF with established grafting methods, the found fragments could be integrated into protein scaffolds and further optimized. MaSIF coupled with classic methods represents a comprehensive platform able to virtually design synthetic PPIs against any protein known to date (Fig. 1.8).

Multiple groups serendipitously discovered that RoseTTAfold and AF2 are able to predict structures of non-contiguous proteins by pseudo-multimer input (e.g. residue gap insertion or chains joined with a flexible linker), hence enabling the prediction of protein complexes [107, 184, 185, 186, 118, 187, 188]. Also, the AF2 system was revised to support multi-chain features and symmetry handling during training and inference (termed AlphaFold-Multimer), demonstrating better performance for predicting protein complexes than the previously mentioned pseudo-multimer inputs [189].

Altogether, computational protein design methods have rapidly evolved and extensively been further improved. The first two chapters present how established protein design methods can be employed to design functional proteins. In the third chapter a novel computational framework is introduced to design proteins binders from scratch interacting at specific locations on the target proteins. The last two chapter describe two computational methods for the *de novo* design of proteins with tailored shapes. Lastly, we discuss the impact of the scientific findings and how the computational design methods could be used jointly opening the door for a general computational protein design platform for the generation of functional proteins.

1.6 Objectives

My dissertation seeks the development of protein design methods incorporating DL, physics-based, and data-driven techniques for generating designer proteins with detailed geometrical shapes so that they can precisely interact with other molecules, ultimately leading to biological functions. The developed methods are relevant to the field of protein design as they (1) further automate the *de novo* protein design process, (2) illustrate innovative solutions to circumvent current *de novo* design challenges, and (3) can be joined with state-of-the-art interface design strategies for the design of biologically active proteins.

1.6.1 Aim 1: Computational protein modeling frameworks for applied PPI design.

The design of biologically functional proteins has the potential to solve major current bio- and nanotechnological challenges. In chapter 2, computational interface design is used to enhance the inhibition and light-switching behavior of an engineered anti-CRISPR protein (Acr) called CASANOVA. Subsequently, chapter 3 presents a surface-guided design strategy to create the first Acr (AcrIIC1X*) that can efficiently inhibit *Sau*Cas9, an important Cas9 orthologue for *in vivo* gene editing. Finally, chapter 4 introduces a MaSIF-based surface-centric *de novo* PPI design framework that is coupled with computational grafting methods for the design of novel PD-L1 inhibitors from scratch. Altogether, the first three chapters present the development and applications of computational interface design methods ranging from the established physics-based to recent DL strategies, accompanied by biochemical and cellular validations of the designs' behavior and functioning.

1.6.2 Aim 2: Hierarchical design of *de novo* proteins with native-like features.

De novo protein design is currently hindered by the necessity of sampling “designable” protein backbones. In chapter 5, we present the TopoBuilder *de novo* design method that generates proteins from an overall description of the targeted fold, "Sketches". An iterative, data-driven approach that extracts important structural characteristics from tertiary motifs of natural structures is introduced to supervise the backbone generation process and the subsequent sequence sampling stage. Importantly, the framework greatly reduces the dependency on hand-crafted rules and manual adjustments of the backbone. Ultimately, the *de novo* design of sequences for five protein folds and extensive computational and experimental characterizations demonstrates that several sequences adopted the correct shape. Thus, the TopoBuilder presents itself as a general-purpose *de novo* protein design framework and enables the custom-building of novel protein folds to comply with predefined structures and functions.

1.6.3 Aim 3: Tailored *de novo* design using DNNs to probe the “darkmatter” of the protein universe.

Most *de novo* protein design methods rely on extensive backbone sampling simulations to search for minimally frustrated conformations that are designable. In addition, recent energy functions suffer from inaccuracies and repeatedly fail in differentiating between stable, well-folded and unstable designs. With the advancements of DL for protein structure prediction and design, a DNN termed Genesis is trained to render protein SSE lattice models termed “Sketches” designable by denoising their 2D feature maps. Genesis is smoothly interfaced with trRosetta to efficiently design sequences and bypass arduous backbone sampling in 3D space. The Genesis framework enables extremely fast exploration of the sequence space independent of fold-specific restraints. Essentially, Genesis alleviates the backbone designability problem and could ultimately contribute to the *de novo* design of currently unexplored “darkmatter” proteins harboring new functions.

2 Engineered anti-CRISPR proteins for optogenetic control of CRISPR–Cas9

...

This chapter is a postprint version based on an article published in Nature Methods in 2018 (DOI: 10.1038/s41592-018-0178-9) in accordance with the publisher.

Authors

Felix Bubeck[1*], Mareike D. Hoffmann[1,2*], **Zander Harteveld**[3,4#], Sabine Aschenbrenner[1,2#], Andreas Bietz[1], Max C. Waldhauer[1], Kathleen Börner[5,6,7], Julia Fakhiri[5,6], Carolin Schmelas[5,6], Laura Dietz[1], Dirk Grimm[5,6,7], Bruno E. Correia[3,4], Roland Eils[1,2,8,9] and Dominik Niopek[1,2]

[*, #] These authors contributed equally.

Affiliations

[1] Synthetic Biology Group, Institute for Pharmacy and Biotechnology (IPMB) and Center for Quantitative Analysis of Molecular and Cellular Biosystems (BioQuant), University of Heidelberg, Heidelberg, DE. [2] Department of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, DE. [3] Institute of Bioengineering, École Polytechnique Fédérale de Lausanne, Lausanne, CH. [4] Swiss Institute of Bioinformatics (SIB), Lausanne, CH. [5] Department of Infectious Diseases, Virology, University Hospital Heidelberg, Heidelberg, DE. [6] BioQuant Center and Cluster of Excellence CellNetworks, Heidelberg University, Heidelberg, DE. [7] German Center for Infection Research (DZIF), Heidelberg, DE. [8] Digital Health Center, Berlin Institute of Health (BIH) and Charité, Berlin, DE. [9] Health Data Science Unit, University Hospital Heidelberg, Heidelberg, DE.

Author contributions

F.B. and D.N. conceived the study. F.B., M.D.H., A.B., M.C.W. and D.N. designed experiments. F.B., M.D.H., S.A. and A.B. conducted experiments. F.B., M.D.H. and D.N. analyzed and interpreted data. Z.H. and B.E.C. performed computational analysis and structural modeling. J.F. developed the luciferase cleavage reporter construct and the CFTR-targeting gRNA. C.S. and

L.D. made practical contributions. K.B. and D.G. provided expertise and practical support for AAV experiments. R.E. and D.N. jointly directed the work. D.N. wrote the manuscript with support from all other authors.

2.1 Abstract

Anti-CRISPR proteins are powerful tools for CRISPR–Cas9 regulation; the ability to precisely modulate their activity could facilitate spatiotemporally confined genome perturbations and uncover fundamental aspects of CRISPR biology. We engineered optogenetic anti-CRISPR variants comprising hybrids of AcrIIA4, a potent *Streptococcus pyogenes* Cas9 inhibitor, and the LOV2 photosensor from *Avena sativa.* Coexpression of these proteins with CRISPR–Cas9 effectors enabled light-mediated genome and epigenome editing, and revealed rapid Cas9 genome targeting in human cells.

2.2 Main

CRISPR–Cas9 technologies have transformed scientists' ability to manipulate genomes and study molecular networks, and concurrently opened novel avenues for the treatment of genetic diseases [190]. In order to improve the accuracy of Cas9-mediated genome perturbations, universal strategies enabling spatiotemporally confined Cas9 activation are highly desired [191]. In the recent past, several approaches facilitating conditional Cas9 activation via chemicals or light have been developed [191]. However, they typically limit users to specific, modified Cas9 or single guide RNA (sgRNA) variants.

Recently, phage-derived anti-CRISPR (Acr) proteins were discovered that naturally inhibit type II CRISPR systems, including the most widely used, *S. pyogenes* Cas9 [192, 193]. These CRISPR antagonists also function when heterologously expressed in yeast [194] and mammalian cells [195], which suggests that CRISPR inhibition could be a widely generalizable strategy for Cas9 regulation. Because of the lack of methods to easily confine Acr activity in time and space, however, the utility of this approach is currently limited.

We sought to overcome this limitation by engineering artificial Acr proteins that can be easily switched between a functional Cas9-inhibitory state and a nonfunctional state via a precise external light stimulus (Fig. 2.1a).

As a sensor, we chose the small (16.5 kDa) LOV2 domain from *A. sativa* phototropin-1, as its blue-light-induced Jα-helix photoswitching mechanism [196] has already been successfully harnessed for the engineering of inducible allostery on diverse effectors [197]. Recent data show that insertion of the LOV2 domain into selected, surface-exposed loops of mammalian motility enzymes enables optogenetic control of enzymatic function [198]. The N and C termini of LOV2 are in close proximity in the dark-adapted state, thereby preserving the native enzyme conformation after LOV2 insertion. Photoexcitation, however, results in unfolding of

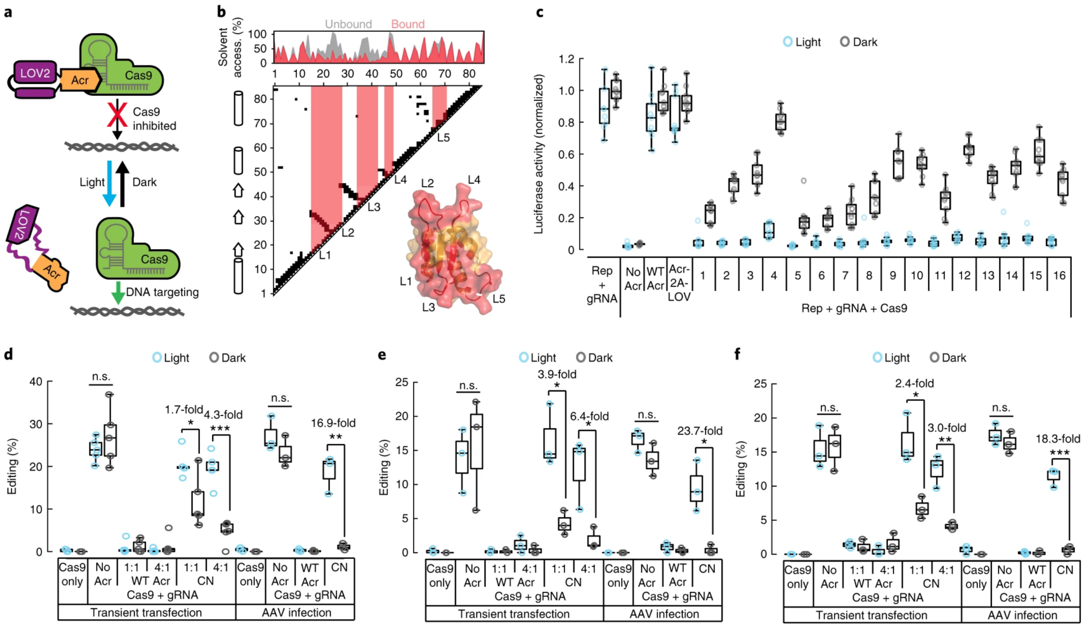


Figure 2.1: Engineering and characterization of CASANOVA, an anti-CRISPR protein dependent on blue light. **a:** Schematic of CASANOVA function. **b:** Structural analysis of AcrIIA4. Top: solvent accessibility (access.) in the Cas9-bound and unbound states. Center: contact map showing residues in close proximity (7 Å distance). Light red bars indicate Cas9-binding segments. Lower right: surface representation of AcrIIA4 (PDB 5VW1) with Cas9-interacting segments shown in light red. L, loop. **c:** Light control of luciferase reporter cleavage. HEK293T cells expressing Cas9, a luciferase reporter (Rep), a reporter-targeting gRNA and the indicated Acr–LOV variant (Sup. Fig. 2.2) were irradiated with blue light for 48 h or kept in the dark and then subjected to a luciferase assay. n=9 biologically independent samples (cell cultures). Acr-2A-LOV, control construct coexpressing AcrIIA4 and LOV2 via a P2A sequence. **d-f:** Light-mediated indel mutation of human CCR5 (d), CFTR (e) and EMX1 (f) loci. Cells expressing the indicated components were illuminated for 70 h or kept in the dark and then subjected to a T7 endonuclease assay. The vector mass ratio of the Acr:Cas9 construct is indicated. WT, wild-type; CN, CASANOVA. *P < 0.05, **P < 0.01, ***P < 0.001, two-sided Student's t-test (exact P values are stated in the Methods section). **d:** n = 5 independent transfection and 3 independent transduction experiments. **e,f:** n = 3 independent experiments. **c–f:** Box plots show the median (center line), first and third quartiles (box edges), 1.5× the interquartile range (whiskers) and individual data points (blue and gray symbols).

the LOV2 terminal helices. This locally induced disorder impairs enzyme function, provided the LOV2 insertion loop is conformationally coupled to the enzyme's active site [198]. Of note, the active site of an enzyme is often highly sensitive to minor structural perturbations. We therefore aimed to develop a similar approach for optogenetic control of nonenzymatic proteins such as Acrs (Supplementary Discussion).

As an actuator, we chose the 87 amino acids (AA) protein AcrIIA4 derived from *Listeria monocytogenes* prophage [192], which binds Cas9–sgRNA complexes with sub-nanomolar affinity [199], thereby efficiently blocking Cas9 DNA binding and nuclease activity [200, 201]. To assess the validity of the LOV2-insertion approach for the Acr target protein, we carried out detailed computational characterization of the reported AcrIIA4 structure in complex with Cas9 and an sgRNA [201]. Analysis of AcrIIA4 solvent accessibility, Cas9 binding segments and residue contacts within the AcrIIA4 structure suggested that loop L5 was a promising target site for LOV2 insertion (Fig. 2.1b and Supplementary Note 1).

We therefore inserted the LOV2 domain (phototropin-1 residues 404–546) at all possible positions into AcrIIA4 L5 and investigated Cas9 inhibition by the resulting Acr–LOV hybrids in HEK293T cells, using a luciferase-reporter cleavage assay (Supplementary Note 2). In the absence of light, the variant carrying the LOV2 domain between AcrIIA4 residues E66 and Y67 impaired Cas9 activity noticeably, although the inhibition was approximately fourfold weaker than that by wild-type AcrIIA4 (Sup. Fig. 2.1). Cas9 activity recovered fully in the presence of blue light, which suggests that the Acr–LOV light switch was functioning, even though room remained for optimization (Fig. 2.1c, variant 1). We hypothesized that the ~10 Å spacing between the LOV2 termini in the dark state might cause undesired strain on the AcrIIA4 structure, thus perturbing Cas9 binding. Therefore, we carried out stepwise deletion of AcrIIA4 residues that directly preceded the insertion site but did not mediate critical contacts with Cas9 [200], and optionally included linkers of variable length at the Acr–LOV boundaries (Sup. Fig. 2.2). Indeed, several of the so-obtained Acr–LOV hybrids showed potent Cas9 inhibition in the dark and almost full recovery of Cas9 activity after photoactivation (Fig. 2.1c, variants 2–16). As expected for a competitive inhibitor, the degree of Cas9 inhibition depended on the dose of transfected Acr–LOV hybrid (Sup. Fig. 2.3). The most potent hybrid inhibitor (Acr–LOV variant 4) carried a three AA deletion (ΔN64/Q65/E66) preceding the LOV domain insertion site, and no GS linkers (Fig. 2.1c and Sup. Fig. 2.2). The deletion is likely to facilitate smooth embedding of the LOV2 domain into the AcrIIA4 target loop without negatively affecting the overall conformation (Sup. Fig. 2.4). In the following, we refer to Acr–LOV variant 4 or further optimized mutants thereof as CASANOVA (for 'CRISPR–Cas9 activity switching via a novel optogenetic variant of AcrIIA4').

Using transient transfection or adeno-associated-virus-mediated transduction, we coexpressed CASANOVA with Cas9 and sgRNAs targeting various genomic loci in HEK293T cells. Insertion/deletion (indel) mutations were strongly light dependent (up to ~24-fold regulation) for all target loci as measured by T7 endonuclease assay (Fig. 2.1d–f, Sup. Fig. 2.5 and Supplementary Note 3). TIDE sequencing [202] further revealed a broad range of indels in the illuminated, but not the dark control, samples (Sup. Fig. 2.6). However, in the transfected samples, we observed significant background editing in the dark, which suggested that Cas9 inhibition was imperfect, at least under heterogeneous expression conditions (Fig. 2.1d–f, Supplementary Discussion). Therefore, we introduced mutations known to improve docking of the terminal helices against the LOV core in the dark [203, 204] into the LOV2 domain of CASANOVA. As expected, Cas9 background activity was reduced in several mutants, albeit at the cost of a lower dynamic range in most cases (Sup. Figs. 2.7 and 2.8, Supplementary Notes 4 and 5). We tuned the performance further by introducing mutations in the AcrIIA4 part of CASANOVA (Sup. Fig. 2.9). We screened these mutations computationally and carried out selection by manual inspection and application of several structural metrics, aiming to enhance Cas9 binding affinity (Methods). Two mutants obtained in this manner (CASANOVA

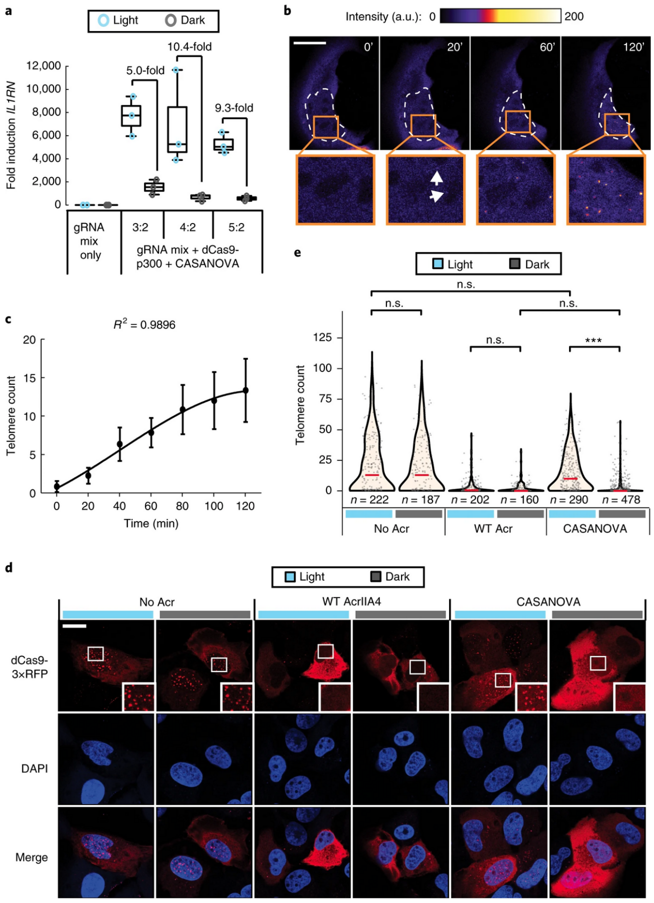


Figure 2.2: Optogenetic control of gene expression and telomere labeling. **a:** Light-dependent IL1RN activation. HEK293T cells expressing the indicated components were exposed to blue light for 44 h or kept in the dark and then subjected to quantitative RT-PCR. The vector mass ratio of the CASANOVA:dCas9–p300 constructs is indicated. Box plots show the median (center line), first and third quartiles (box edges), 1.5× the interquartile range (whiskers) and individual data points (blue and gray symbols). n = 3 independent experiments. **b,c:** Optogenetic recruitment of dCas9 to telomeres in living cells. U2OS cells expressing dCas9–3×RFP, a telomere-targeting gRNA and CASANOVA were exposed to blue light pulses every 30 s for 2 h. **b:** Representative RFP fluorescence images. Arrows point to the first visible fluorescent dots. Dashed lines indicate the nucleus boundary. Scale bar, 20 µm; zoomed-in views in the bottom row are magnified 3.2-fold relative to the images above. **c:** Quantification of labeled telomeres over time. The line indicates a third-order polynomial fit. Data are mean ± s.e.m. **b,c:** n = 3 biologically independent samples (cell cultures). **d,e:** Light-mediated telomere recruitment in samples fixed after 20 h of irradiation or incubation in the dark. n = 3 independent experiments for WT AcrIIA4 and CASANOVA samples; n = 4 independent experiments for no-Acr samples. **d:** Representative fluorescence images. Scale bar, 20 µm; insets are magnified 2.2-fold relative to primary images. **e:** Quantification of telomere labeling by automated image analysis (Methods). The violin plot shows the distribution; red bars and gray dots indicate the median and individual data points, respectively. ***P < 2.2 × 1016 by two-sided Wilcoxon rank-sum test. The total number of nuclei analyzed is indicated. n.s., not significant.

T16F and CASANOVA S46D) showed enhanced Cas9 inhibition in the dark without noticeably compromised light activation (Sup. Fig. 2.10). CASANOVA and several of its optimized mutants also conferred strong light regulation on xCas9, a recently developed protospacer-adjacent motif–relaxed, highly specific *Spy*Cas9 derivative [205] (Sup. Fig. 2.11).

Next, we investigated whether CASANOVA would enable light-mediated regulation of dCas9–effector fusions. To this end, we used a previously reported dCas9 variant fused to the p300 histone acetyltransferase core domain [206] and targeted the interleukin 1 receptor antagonist (IL1RN) promoter, known to be strongly activated upon induced H3K27 acetylation, in HEK293T cells via a combination of four guide RNAs (gRNAs). We titrated the transfected CASANOVA dose and incubated cells in the dark or light for 44 h before assessing IL1RN expression by quantitative RT-PCR. IL1RN transcript levels were increased up to tenfold in the illuminated samples compared with levels in the dark controls, indicating successful control of the dCas9–p300 epigenetic modifier (Fig. 2.2a). IL1RN levels continued to increase with prolonged illumination but decreased after the stimulus was withdrawn, indicating reversibility of the CASANOVA system (Sup. Fig. 2.12 and Supplementary Discussion).

Finally, we assessed CASANOVA's potential for studying the kinetics of Cas9 DNA targeting in living cells. To this end, we conducted a CRISPR labeling [207] experiment in which a dCas9–3×RFP fusion, a telomere-targeting gRNA and CASANOVA were coexpressed in U2OS cells [208]. Twenty-four hours after transfection, we irradiated the cells with a 488-nm laser beam, and we monitored the RFP signal over time by confocal microscopy. Strikingly, the first visible dots appeared in the cell nucleus only 20–40 min after irradiation, and their number and intensity increased rapidly over time, indicating dCas9–3×RFP recruitment to telomeres (Fig. 2.2b,c).

To rule out that the observed telomere recruitment was due simply to differential expression kinetics of the transfected components (dCas9–3×RFP, gRNA and CASANOVA) rather than to light-induced dCas9–3×RFP release from CASANOVA, we conducted a comprehensive control experiment. This time, we incubated transfected cells in either light or dark for 20 h, and we fixed the samples before microscopy analysis. We also included control samples expressing wild-type AcrIIA4 instead of CASANOVA or expressing no inhibitor at all, and analyzed telomere labeling in a fully unbiased manner using an automated image-analysis workflow implemented in KNIME (Methods and Supplementary Data). The CASANOVA samples showed strong telomere labeling similar to that in the positive control when incubated in light, whereas labeling was notably decreased and similar to that in the negative control after incubation in the dark (Fig. 2.2d,e and Sup. Fig. 2.13). The no-inhibitor and wild-type AcrIIA4 controls showed light-independent strong and weak telomere labeling, respectively (Fig. 2.2d,e and Sup. Fig. 2.13).

These observations suggested that dCas9 release from the CASANOVA trap occurred quickly after photoexcitation, which we verified by monitoring the reversible, light-dependent interaction of a fluorescently labeled CASANOVA variant and a plasma-membrane-targeted dCas9 (Sup. Fig. 2.14). Moreover, these data provide direct confirmation that Cas9 DNA targeting in mammalian cells is a rapid process that takes place on a scale of minutes (Fig. 2.2b,c, Supplementary Discussion).

CASANOVA not only is an important add-on to the CRISPR toolbox (Supplementary Protocol), but also conceptually advances the ability to confer light regulation on non-enzymatic proteins (Supplementary Discussion).

2.3 Methods

2.3.1 Structure-based identification of the LOV2 insertion site

We analyzed the X-ray structure of the *Spy*Cas9–sgRNA–AcrIIA4 [201] complex (PDB 5VW1) with CMView [209] (version 1.1) to generate the contact map. Contacts between residues were considered positive if their Cα atoms were less than 7 Å apart. The secondary structure of Acr was assigned with DSSP (version 2.0.4) [210], and the binding-interface segments were assigned by spatial proximity to the CRISPR molecule. Here, segments containing multiple residues of less than 4.5 Å were considered the binding interface. Finally, we analyzed the solvent-accessible surface area using the software NACCESS (version 2.1.1; http://www.bioinf.manchester.ac.uk/naccess/) with the default settings.

2.3.2 Domain assembly

To generate the domain fusions between the LOV and Acr domains, we used the Rosetta Remodel application (version 3.9) [211]. We used two input structures: the *Spy*Cas9–sgRNA–AcrIIA4 [201] (PDB 5VW1) complex and the LOV2 domain from *A. sativa* (PDB 2V0W). The C- and N-terminal helices of the LOV2 domain were rebuilt using loop fragments and then underwent cyclic coordinate descent [212] and kinematic closure [213, 214] refinement. A total of 331 decoys passed the chain break filter, out of 2,500 decoys attempted. In a second step, the 331 output decoys were clustered with an root-mean-square deviation (RMSD) threshold of 10 Å using Rosetta's clustering tool and further minimized. The 331 structures yielded a total of six clusters, and the best-scoring decoys of the top three populated clusters were selected to illustrate the potential structural diversity of the Acr–LOV fusion.

2.3.3 Computational design of improved Acr–LOV mutants

We carried out interface design for the interface residues in AcrIIA4 using the RosettaScripts application [165]. *In silico* saturation mutagenesis was carried out for residues in close spatial proximity (residues 16, 18 and 33 composed set 1, and residues 19, 28 and 45 constituted set 2). Designs with interaction energies within the same range (+2.5 Rosetta energy units) or lower than that of the wild-type complex were manually inspected, and the best mutations were selected for experimental characterization. Supplementary Table 2.15 presents the metrics of the experimentally characterized mutants.

2.3.4 General methods and cloning

A list of all constructs used and created in this study is shown in Supplementary Table 2.16. Annotated vector sequences are provided as Supplementary Data (GenBank files). Plasmids were created via classical restriction enzyme cloning, Golden Gate cloning [215] or Gibson assembly (New England Biolabs). Oligonucleotides were obtained from IDT or Sigma Aldrich. Synthetic double-stranded DNA fragments were obtained from IDT. The CMV-promoter-driven *Spy*Cas9 expression vector was obtained by PCR amplification of the *Spy*Cas9 gene from vector *pSp*Cas9(BB)-2A-GFP (kind gift from Feng Zhang (Addgene plasmid 48138)) followed by ligation into pcDNA3.1(–) (Thermo Fisher) via XhoI/HindIII. Adeno-associated virus (AAV) vectors encoding *Spy*Cas9 or a U6-promoter-driven, improved gRNA scaffold (F+E18) and RSV-promoter-driven GFP28 were used for gRNA expression. Annealed oligonucleotides corresponding to the target site sequence were cloned into the gRNA AAV vector via BbsI by Golden Gate cloning. All gRNA target sites relevant to this study are shown in Supplementary Table 2.17. The luciferase reporter for measuring *Spy*Cas9 activity (luciferase cleavage reporter) was developed by cloning of an H1-driven expression cassette encoding a firefly-luciferase-targeting gRNA into pAAVpsi229. The resulting vector coencoded an SV40-promoter-driven Renilla luciferase gene and a TK-promoter-driven firefly luciferase gene. The AcrIIA4 coding sequence was obtained as a human-codon-optimized synthetic DNA fragment from IDT and cloned into pcDNA3.1(–) via NheI/NotI. We created Acr–LOV hybrids by linearizing the Acr-encoding vector by PCR and then inserting a human-codon-optimized *A. sativa* LOV2-encoding fragment (IDT) via blunt-end ligation or Golden Gate cloning. GS linkers were optionally appended to the LOV-encoding DNA fragment via PCR before ligation. Mutations were introduced into the Acr part of the Acr–LOV hybrids by site-directed mutagenesis with 5-phosphorylated primers. Mutations were inserted into the LOV part of the Acr–LOV hybrids by PCR amplification of the LOV2 domain with primers introducing the mutations into the N- and C-terminal helix and cloning of the altered LOV fragment back into a PCR-linearized parent Acr–LOV hybrid vector by Golden Gate cloning. Note that wild-type Acr and all Acr–LOV hybrids bore an N-terminal SV40 nuclear localization signal, which we added to target the Cas9 inhibitor to the nucleus. The xCas9 cDNA was created by Gibson assembly on the basis of the reported *Spy*Cas9 mutations [205] using synthetic double-stranded DNA fragments cloned into pcDNA3.1(–). The dCas9–p300 construct was a kind gift from Charles Gersbach (Addgene plasmid # 61357). pEJS477 - pHAGE - TO - *Spy*dCas9_3XmCherry - sgRNA / Telomere - All-in-one was a gift from Erik Sontheimer (Addgene plasmid # 85717). Based on this vector, we created constructs coexpressing dCas9_3 x mCherry and CASANOVA or wild-type AcrIIA4 via a P2A peptide by cloning a P2A-CASANOVA or P2A-AcrIIA4 cDNA (IDT) behind the *Spy*Cas9-3 x mCherry coding sequence.

In all cloning procedures, PCR was done with Q5 Hot Start high-fidelity DNA polymerase (New England Biolabs) or Phusion Flash high-fidelity polymerase (Thermo Fisher). Agarose gel electrophoresis was used to analyze PCR products. Bands of the expected size were cut out and DNA was extracted with a QIAquick gel extraction kit (Qiagen). Ligations were performed with T4 DNA ligase (New England Biolabs) and optionally heat-inactivated at 70 °C for 5 min before transformation. Chemically competent Top10 cells (Thermo Fisher) were used for DNA vector amplification. Plasmid DNA was purified with the QIAamp DNA Mini, Plasmid Plus Midi or Plasmid Maxi kit (all from Qiagen).

2.3.5 Cell culture, transient transfection and AAV lysate production

Cells lines were cultured at 5% CO2 and 37 °C in a humidified incubator and passaged when they reached 70–90% confluency (every 2–4 d). HEK293T (human embryonic kidney) and U2OS (human osteosarcoma; kindly provided by Karsten Rippe, German Cancer Research Center (DKFZ), Heidelberg) cells were maintained in phenol-red-free DMEM (Thermo Fisher/Gibco) supplemented with 10% (v/v) FCS (Biochrom AG), 2 mM l-glutamine, and 100 U/ml penicillin + 100 µg/ml streptomycin (both Thermo Fisher/Gibco). The U2OS medium was additionally supplemented with 1 mM sodium pyruvate (Gibco). Cell lines were authenticated and tested for mycoplasma contamination before use via a commercial service (Multiplexion).

Transient transfections were performed with JetPrime (Polyplus Transfection) or Turbofect (Thermo Fisher) according to the manufacturer's protocols. Details are given in the corresponding experimental sections below. For the production of AAV-containing cell lysates, low-passage HEK293T cells were seeded into six-well plates (CytoOne) at a density of 350,000 cells per well. The next day, cells were triple-transfected with (i) the AAV vector plasmid, (ii) an AAV helper plasmid carrying AAV serotype 2 rep and cap genes and (iii) an adenoviral plasmid providing helper functions for AAV production, using 1.33 µg of each construct and 8 µl of Turbofect reagent per well. The AAV vector plasmid encoded (1) Cas9 driven from an engineered, short CMV promoter [216], (2) a U6-promoter-driven gRNA [216] (based on the improved F+E scaffold18) and an RSV-promoter-driven GFP marker or (3) a CMV-promoter-driven CASANOVA variant. Seventy-two hours after transfection, cells were collected in 300 µl of PBS and subsequently subjected to five freeze–thaw cycles alternating between snap-freezing in liquid nitrogen and 37 °C. Finally, the cell debris was removed by centrifugation at ~18,000g and the AAV-containing supernatant was stored at –20 °C until use.

2.3.6 Blue light setup

For blue light illumination of samples in the cell culture incubator, we used a custom-made blue light setup comprising six blue light high-power LEDs (type CREE XP-E D5-15; emission peak ~460 nm; emission angle ~130°; LED-TECH.DE) powered by a Switching Mode Power Supply (Manson; HCS-3102). We used a Raspberry Pi running a custom Python script to control the power supply. We irradiated samples from below, through the transparent bottom of the culture dishes or well plates, by positioning them on an acrylic glass table installed in the incubator with the LEDs located underneath the table. We used a pulsatile illumination regime (5 s on, 10 s off) for sample irradiation. The light intensity was ~3 W/m2 as measured with a LI-COR LI-250A light meter, unless indicated otherwise below.

2.3.7 Luciferase reporter assays

HEK293T cells were seeded into black, clear-bottom 96-well plates (Corning) at a density of ~12,500 cells per well. The next day, cells were cotransfected with 33 ng of a Cas9 or xCas9 expression vector, 33 ng of a construct coexpressing firefly and Renilla luciferase as well as an H1 promoter-driven gRNA targeting the firefly luciferase cDNA, and, in most cases, 33 ng of the CMV-promoter-driven Acr–LOV hybrid with 0.2 µl of JetPrime (amounts are per well). For the titration experiment in Sup. Fig. 2.3, 3, 10 or 33 ng of Acr–LOV hybrid was cotransfected with 30, 23 or 0 ng of an irrelevant DNA to vary the vector mass ratio of Cas9 and Acr–LOV construct between 10:1 and 1:1. For the experiment shown in Sup. Fig. 2.7a, 8.25 ng of Acr–LOV constructs and 24.75 ng of an irrelevant DNA were used. Six hours after transfection, the medium was exchanged and cells were exposed to blue light for 48 h or kept in the dark as a control (Supplementary Note 6). For the titration experiment shown in Sup. Fig. 2.3, the irradiation time was 30 h and the light intensity was ~2.5 W/m^2. Subsequently, a Dual-Glo luciferase assay system (Promega) was applied to quantify luciferase activity. In brief, cells were collected into the supplied lysis buffer, and firefly and Renilla luciferase activities were measured with a GLOMAX Discover or GLOMAX 96 microplate luminometer (both from Promega). The integration time was 10 s, and the delay time between automated substrate injection and measurement was 2 s. Firefly photon counts were normalized to Renilla photon counts, and the resulting values were further normalized to the positive control.

2.3.8 T7 endonuclease assay and TIDE sequencing

Cells were seeded into black, clear-bottom 96-well plates (Corning) at a density of 12,500 cells per well for transfection-based experiments or 3,500 cells per well for AAV transduction-based experiments. Transfections were performed with JetPrime using 0.3 µl of JetPrime reagent and 200 ng of total DNA per well, comprising one of the following mixes: 33 ng each of the gRNA, Cas9 and CASANOVA expression vectors, and 100 ng of an irrelevant DNA (1:1 ratio Cas9:CASANOVA); or 33 ng of gRNA, 33 ng of Cas9 and 133 ng of CASANOVA expression vector (1:4 ratio Cas9:CASANOVA). For AAV-based experiments, cells were cotransduced with 7 µl each of the Cas9, gRNA and CASANOVA AAV lysates on two subsequent days when the CCR5 locus was being targeted. For all other loci, 33 µl of the Cas9 and gRNA AAV lysate were used in combination with 20 µl (for CFTR gRNA2, CFTR gRNA3, mir-122 and VEGFA) or 33 µl (for CFTR and EMX1) of CASANOVA lysate. After transfection or transduction, cells were irradiated with blue light for 70 h or kept in the dark as a control (Supplementary Note 6). Cells were washed with PBS and collected in DirectPCR lysis reagent (Peqlab) supplemented with proteinase K (Sigma). The genomic CRISPR–Cas9 target locus was PCR-amplified with primers flanking the target site (Supplementary Table 4) using Q5 Hot Start high-fidelity DNA polymerase (New England Biolabs). For TIDE sequencing analysis [217], the amplicon was purified by gel electrophoresis followed by gel extraction with the QIAquick gel extraction kit (Qiagen) and by Sanger sequencing (GATC). Data analysis was carried out with the TIDE web tool (version 2.0.1; https://tide.deskgen.com/). To assess the indel frequency by T7 assay, we used a rapid T7 protocol [216]. We diluted 10 µl of the target locus amplicons 1:4 in 1× NEB buffer 2, heated the mixture to 95 °C, and then slowly cooled it to allow reannealing and formation of heteroduplexes using a nexus GSX1 Mastercycler (Eppendorf) and the following program: 95 °C/5 min, 95–85 °C at –2 °C per second, 85–25 °C at –0.1 °C per second. Subsequently, 0.5 µl of T7 endonuclease (New England Biolabs) was added, and samples were mixed and incubated at 37 °C for 15 min and then analyzed on a 2% Tris-borate-EDTA agarose gel. The Gel iX20 system equipped with a 2.8-megapixel/14-bit scientific-grade CCD (charge-coupled device) camera (Intas) was used for gel documentation. To calculate the indel percentages from the gel images, we subtracted the background from each lane and quantified T7 bands with the ImageJ (version 1.51n; http://imagej.nih.gov/ij/) gel analysis tool. Peak areas were measured and percentages of insertions and deletions (indel(%)) were calculated using the formula Indel(%) = 100 × (1–(1–Fraction cleaved) × 0.5), where

$$\text{Fraction cleaved} = \frac{\sum \text{Cleavage product bands}}{\sum \text{Cleavage product bands} + \text{PCR input band}}$$

We calculated the reported fold changes in editing efficiency by dividing the mean indel(%) of an illuminated sample by the mean indel(%) of its corresponding dark control sample. Full-length gel images are shown in Supplementary Note 7.

2.3.9 Quantitative RT-PCR

HEK293T cells were seeded into transparent six-well plates (CytoOne) at 250,000 cells per well. The next day, cells were cotransfected with (i) 750 ng of IL1RN gRNA construct mix17 (187.5 ng per vector); (ii) 500 ng of a construct encoding dCas9–p300–P2A–CASANOVA (or an irrelevant DNA as control); (iii) 250, 500 or 750 ng of CASANOVA-encoding vector (corresponding to vector mass ratios of 3:2, 4:2 and 5:2 as indicated in Fig. 2a); and (iv) 500, 250 or 0 ng of irrelevant stuffer DNA, using 6 µl of JetPrime reagent (all amounts are per well). The medium was replaced 4 h after transfection, and cells were irradiated with blue light pulses for 44 h or kept in the dark as a control (Supplementary Note 6). For the experiment shown in Sup. Fig. 2.12, we used the 4:2 vector mass ratio and no stuffer DNA. Furthermore, after the 44-h illumination period, samples were split into two separate six-well plates and illumination was continued for another 3 d, or cells were kept in the dark as controls.

Subsequently, cells were lysed with QIAzol lysis reagent (Qiagen) according to the manufacturer's instructions. Reverse transcription was performed with the RevertAid first strand cDNA synthesis kit (Thermo Fisher) and equal amounts of input RNA for each experiment. Real-time PCR reactions were set up with 2 µl of cDNA mix (25 ng/µl), 1.4 µl of each 10 µM primer (IL1RN forward or GAPDH forward and IL1RN reverse or GAPDH reverse; Supplementary Table 2.18), 10 µl of PowerSYBR Green PCR master mix (Thermo Fisher) and 5.2 µl of water. A StepOne Plus real-time PCR system (Applied Biosystems) was used with the following cycling conditions:

95 °C for 10 min for initial denaturation followed by 45 cycles of (95 °C for 15 s, 58 °C for 60 s). Fold changes in IL1RN transcript levels were then calculated via the $\Delta\Delta$Ct method [218].

2.3.10 Telomere labeling experiments

U2OS cells were seeded into four-compartment CELLview cell culture dishes (Greiner Bio-One) at a density of 30,000 cells per compartment. The next day, cells were transfected with vectors encoding (i) CMV-promoter-driven dCas9–3×RFP–P2A–CASANOVA and U6-promoter-driven telomere-targeting gRNA, (ii) a telomere-targeting gRNA and GFP transfection marker, and (iii) CMV-promoter-driven CASANOVA in a ratio of 20:6:3 using 362.5 ng of total DNA and 1.5 µl of JetPrime for transfection (per compartment). Four hours after transfection, the medium was changed.

For experiments shown in Fig. 2.2b,c, samples were kept in the dark for 24 h. Subsequently, imaging was done with a Leica SP8 confocal laser scanning microscope equipped with automated CO_2 and temperature control; UV, argon and solid-state lasers; and an HC PL APO 40×/1.3-NA (numerical aperture) oil objective. RFP fluorescence was recorded every 5 min for 2 h using the 552-nm laser line for excitation (1% laser power). The detection wavelength was set to 578–789 nm. In parallel, the field of view was scanned with a 488-nm laser beam (2% laser power) every 30 s. We analyzed images manually by counting RFP fluorescent spots (i.e., labeled telomeres) in the nucleus. Four people assessed the images independently, and their results were averaged for each nucleus at each time point.

For the experiments described in Fig. 2.2d,e, additional control samples were included. In the positive control samples, vector i was replaced by a vector encoding dCas9–3×RFP (without the P2A–CASANOVA) and a U6-promoter-driven telomere-targeting RNA, and vector iii was replaced by an irrelevant DNA. In the negative control samples, the CASANOVA in vectors i and iii was replaced by wild-type AcrIIA4. Four hours after transfection, cells were either irradiated with blue light pulses for 20 h or kept in the dark (Supplementary Note 6), after which samples were fixed with 4% PFA. SlowFade Diamond antifade mountant with DAPI (Invitrogen) was added and imaging was carried out with the aforementioned microscopy setup and the following excitation/detection settings: 405 nm (1% laser power)/410–490 nm for DAPI, 488 nm (2% laser power)/493–578 nm for GFP, or 552 nm (1% laser power)/578–789 nm for RFP. RFP fluorescent spots (i.e., labeled telomeres) were then detected and quantified via a fully automated image-analysis pipeline (described below).

2.3.11 Automated image analysis for telomere labeling in KNIME

We used the ImageJ2 (beta) Integration in KNIME Version 3.5.2 (KNIME AG) to create an automated image-processing and analysis pipeline for the quantification of labeled telomeres. The fully annotated workflow is provided as Supplementary Data. All images were analyzed with an identical workflow configuration, apart from the configuration of data input and output nodes. In brief, raw image stacks (.lif files) were imported into KNIME, and then the three fluorescence channels were split (DAPI, nuclear marker; GFP, a transfection marker coencoded on the gRNA vector; and RFP, corresponding to dCas9–3×mCherry). Nuclei were segmented on the basis of the DAPI signal. GFP– nuclear segments (i.e., negative for the telomere-targeting gRNA construct) were excluded from the analysis. Furthermore, nuclear segments with a mean RFP signal higher than 170 (as images were 8-bit, this corresponds to two-thirds of the maximum) were also excluded from the analysis, as the very high RFP background fluorescence impaired reliable spot detection. The Spot Detection node was used to identify and segment fluorescent spots in the RFP channel. All spots lying outside of the nuclear segments were excluded, and random fluorescence fluctuations were filtered out by selection for spots with an average fluorescence at least 1.7-fold higher than the RFP background fluorescence in the corresponding nuclear segment. The workflow output comprised a CSV table listing the nuclear segments and corresponding spots detected in each image. Subsequent data visualization and statistical analysis were done in R version 3.3.2.

2.3.12 Membrane recruitment experiments

HEK293T cells were seeded into four-compartment CELLview cell culture dishes (Greiner Bio-One) at a density of 60,000 cells per compartment. The next day, cells were cotransfected with (i) 900 ng of a vector encoding the Rosa26-1 gRNA31, which does not have a target site in human cells, (ii) 100 ng of a vector expressing dCas9–mVenus fused to the myristoylation palmitoylation domain from lyn kinase, and (iii) 5 ng of Acr–LOV hybrid 5 C-terminally fused to mCherry with 1 µl of JetPrime (amounts are per compartment). The medium was exchanged 4 h after transfection, and samples were kept in the dark for 24 h. For the experiments in Supplementary Fig. 14c, cells were seeded into one-compartment CELLview cell culture dishes (Greiner Bio-One) at a density of 360,000 cells per dish. Identical amounts of the aforementioned vectors and 2 µl of JetPrime reagent were used for transfection (amounts are per single dish). The aforementioned Leica SP8 microscopy setup was used for live-cell recruitment experiments. We irradiated cells with blue light pulses by scanning the field of view with a 488-nm laser beam (2% laser power) every 30 s. mVenus and mCherry signals were recorded in parallel every 5 min. Excitation/detection settings were as follows: 514 nm (0.5% laser power)/493–519 nm for mVenus, or 552 nm (1% laser power)/580–789 nm for mCherry. To quantify the ratio of membrane to cytoplasmic mCherry fluorescence, we segmented cells manually by drawing regions of interest (ROIs) around the plasma membrane of the cell (to calculate the total cell fluorescence) and inside the cell close to but not touching the plasma membrane (to calculate the cytoplasmic fluorescence) in ImageJ. The integrated fluorescence signal was measured for each ROI and at each time point, and the ratio of plasma membrane to cytoplasmic fluorescence was calculated from the obtained values via the following formula:

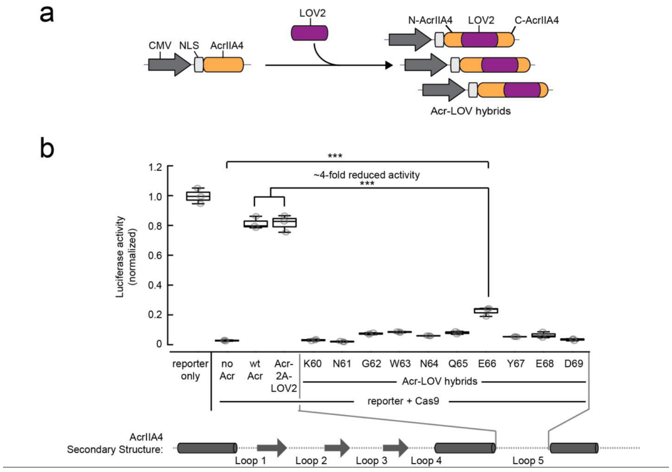
$$\text{membrane/cytoplasmic fluorescence} = \frac{\text{total cell fluorescence} - \text{cytoplasmatic fluorescence}}{\text{cytoplasmatic fluorescence}}$$

Finally, the resulting values were normalized to the corresponding value at time point 0 for each cell.

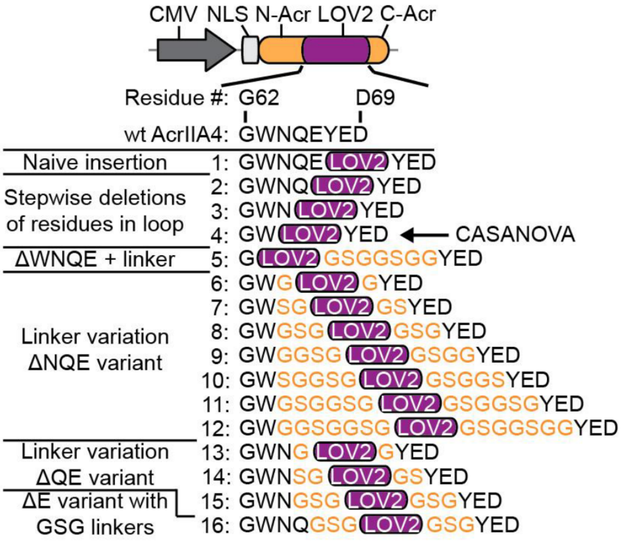
2.3.13 Statistical analysis

All box plots show the median (center line), first and third quartiles (box edges), and minimum/maximum values within 1.5× the interquartile range (whiskers). Uncertainties in the reported mean values are indicated as s.e.m. Statistical significance in reported differences was tested by two-sided Student's t-test. A two-sided Wilcoxon rank-sum test was applied for non-normally distributed data. P values < 0.05 were considered statistically significant. Statistical analysis was performed in R (version 3.1.0) and Microsoft Excel (version 14.0.7208.5000). Figure 1d: *P=0.0229, **P=0.0025, ***P=9.96×105. Figure 1e: *P=0.0128 (transfection, 1:1), *P=0.0299 (transfection, 1:4), *P=0.0142 (transduction). Figure 1f: *P=0.0137, **P=0.0045, ***P=0.0002.

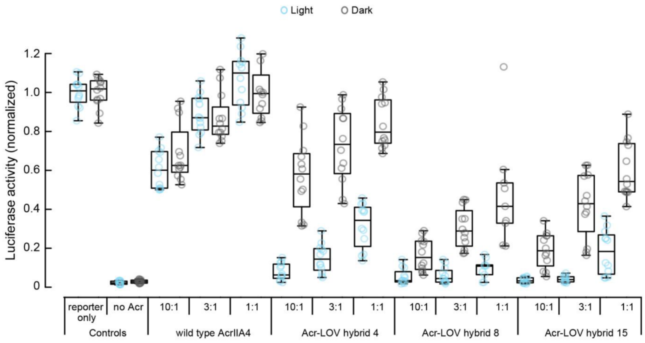
2.4 Supplementary information



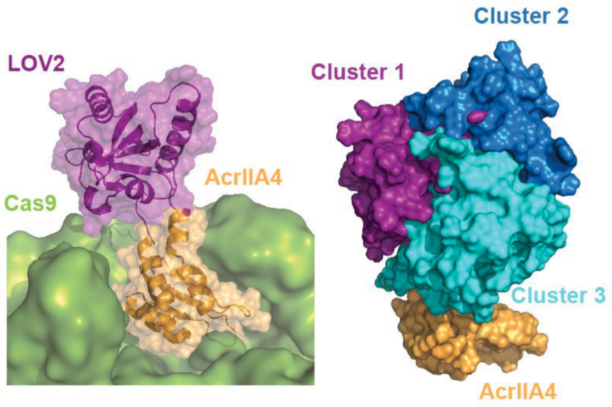
Supplementary Figure 2.1: Sampling of LOV2 insertion sites in AcrIIA4 loop 5. **a:** Acr–LOV hybrid generation. The Avena sativa LOV2 domain coding sequence from phototropin-1 was inserted at different positions in AcrIIA4 loop 5. NLS, SV40 nuclear localization signal. CMV, cytomegalovirus promoter. **b:** Luciferase reporter cleavage assay measuring Cas9 inhibition by different Acr–LOV hybrids. HEK293T cells were co-transfected with vectors encoding (i) the Acr–LOV hybrid, (ii) Cas9 and (iii) a luciferase reporter as well as a gRNA targeting the luciferase gene. Luciferase activity was assessed 48 h post-transfection. The AcrIIA4 residue behind which the LOV2 domain was inserted is indicated. Box plots show the median (center line), first and third quartiles (box edges), 1.5× the interquartile range (whiskers) and individual data points (circles). n=3 biologically independent samples (cell cultures). Wt, wild-type. Acr-2A-LOV2, control construct co-expressing wild-type AcrIIA4 and the LOV2 domain via a P2A sequence. ***P = 0.0004 (no Acr), ***P = 3.63 × 10–5 (wt Acr) and ***P = 8.87 × 10–5 (Acr-P2A-LOV2) by two-sided Student's t-test.



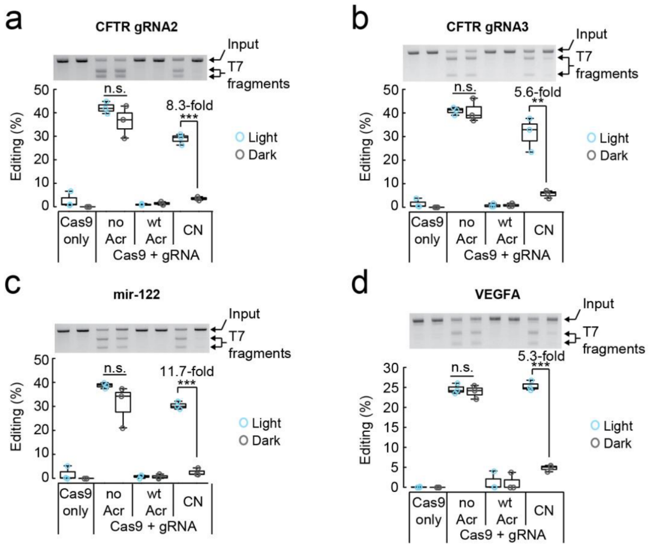
Supplementary Figure 2.2: Schematic and sequences of Acr–LOV hybrids. NLS, SV40 nuclear localization signal.



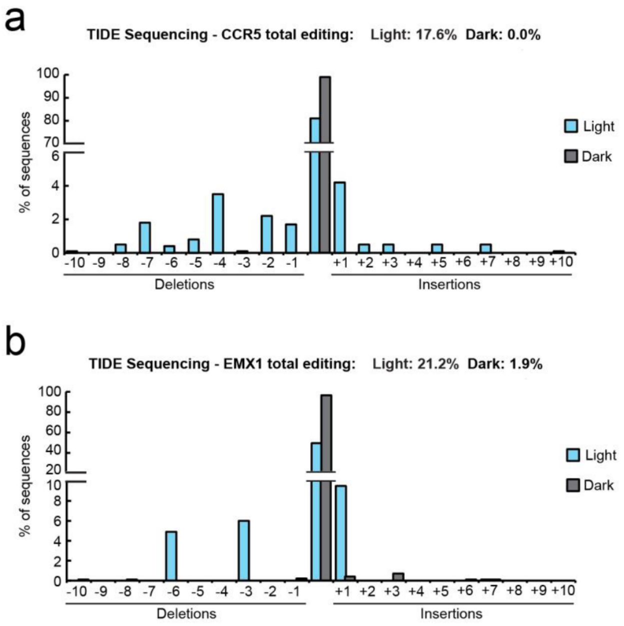
Supplementary Figure 2.3: Cas9 inhibition is dose dependent. HEK293T cells were co-transfected with plasmids encoding (i) Acr–LOV hybrid, (ii) Cas9 and (iii) a luciferase reporter as well as a gRNA targeting the luciferase gene. The vector mass ratio of the transfected Cas9 and Acr–LOV construct was varied between 10:1 and 1:1, as indicated. Six hours post-transfection, cells were irradiated with pulsatile blue light (5 s ON, 10 s OFF; 2.5 W per m2) for 30 h or kept in the dark as control before assessing luciferase activity. Box plots show the median (center line) and first and third quartiles (box edges), 1.5× the interquartile range (whiskers) and individual data points (circles). n = 9 biologically independent samples (cell cultures) for the Acr–LOV hybrid 8 (1:1) and n = 12 biologically independent samples (cell cultures) for all other conditions.



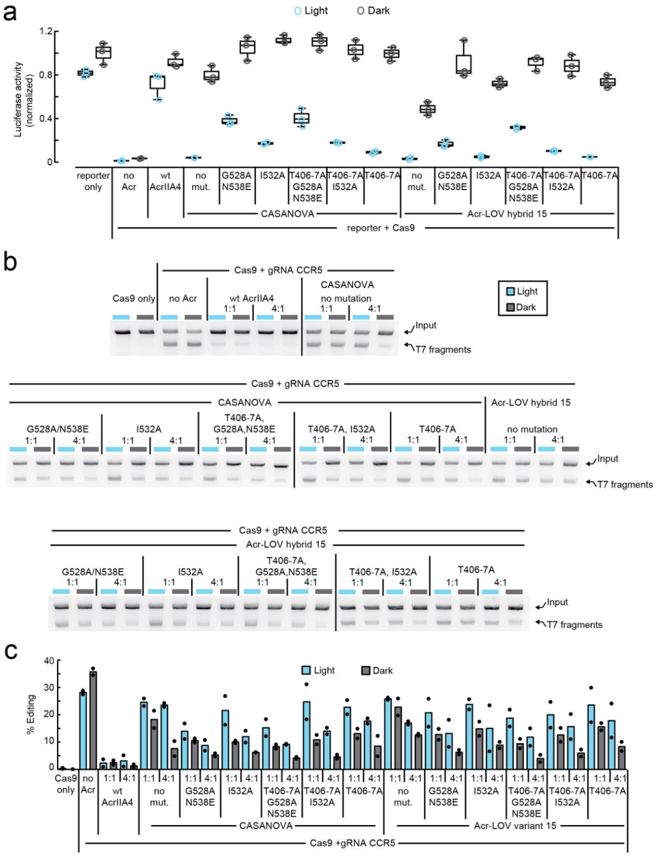
Supplementary Figure 2.4: CASANOVA computational model. Structural model of CASANOVA bound to a Cas9–gRNA complex (left). The three most populated clusters of CASANOVA conformations obtained through domain assembly simulations (Methods) are displayed on the right.



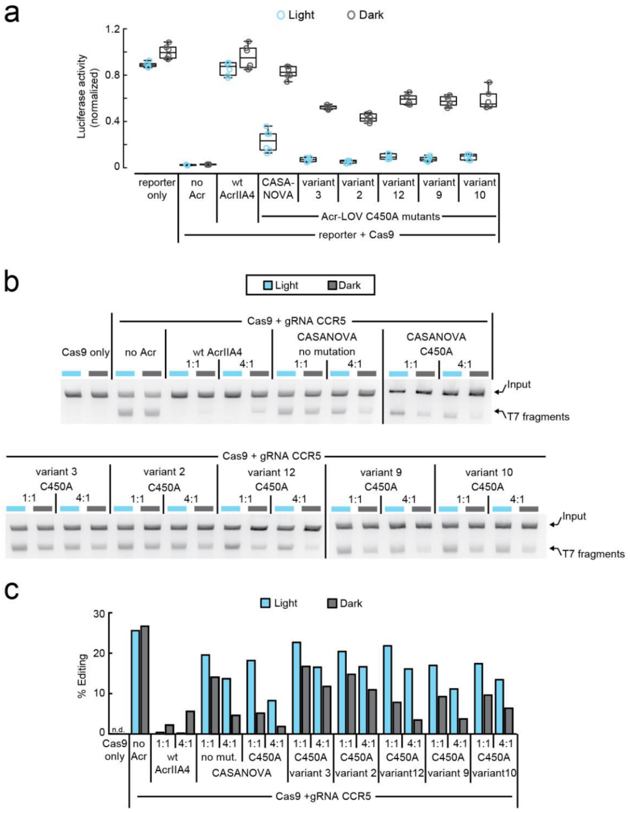
Supplementary Figure 2.5: Photoactivatable genome editing with CASANOVA. Light-mediated indel mutation of human CFTR locus ((a,b) with two different gRNAs), mir-122 locus (c) and VEGFA locus (d). HEK293T cells were co-transduced with AAV vectors encoding CASANOVA, Cas9 and the indicated gRNA and exposed to blue light for 70 h or kept in the dark as control. The target locus was then PCR amplified with primers flanking the estimated break point and the amplicon was denatured and re-annealed in a thermocycler to allow heteroduplex formation. Following digestion with T7 endonuclease, samples were analyzed on a 2% agarose gel (see Methods for details). Representative gel images and corresponding quantifications of editing frequencies are shown. Wt, wild-type. Box plots show the median (center line) and first and third quartiles (box edges), 1.5× the interquartile range (whiskers) and individual data points (circles). n = 3 independent experiments. All statistics by two-sided Student's t-test. (a) ***P = 5.83 × 10–5 (b) **P = 0.0039 (c) ***P = 2.85 × 10–5 (d) ***P = 2.01 × 10–5.



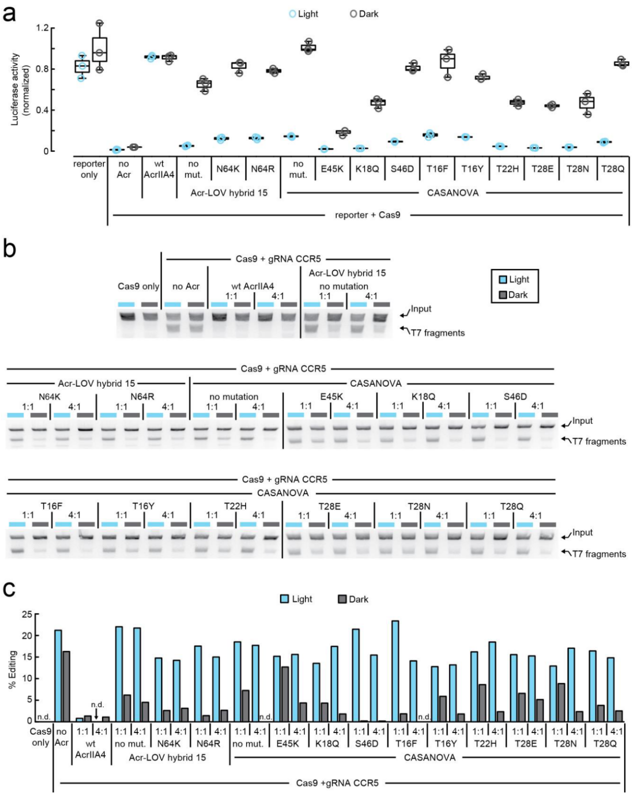
Supplementary Figure 2.6: TIDE sequencing analysis of light-mediated indel mutation. HEK293T cells were co-transduced with AAV vectors expressing the Cas9, CASANOVA and a gRNA targeting the CCR5 **a** or EMX1 **b** locus. Cells were exposed to blue light for 70 h or kept in the dark as control, followed by TIDE sequencing1. The target locus was PCR-amplified with primers flanking the expected breakpoint followed by Sanger sequencing of the amplicon. Total editing efficiencies and frequencies of individual insertions or deletions were then calculated by decomposition of the sequencing chromatogram using the TIDE web tool (https://www.deskgen.com/landing/tide.html). TIDE sequencing revealed a broad range of different insertions and deletions in the light, but not in the dark control samples. Data represent a single experiment.



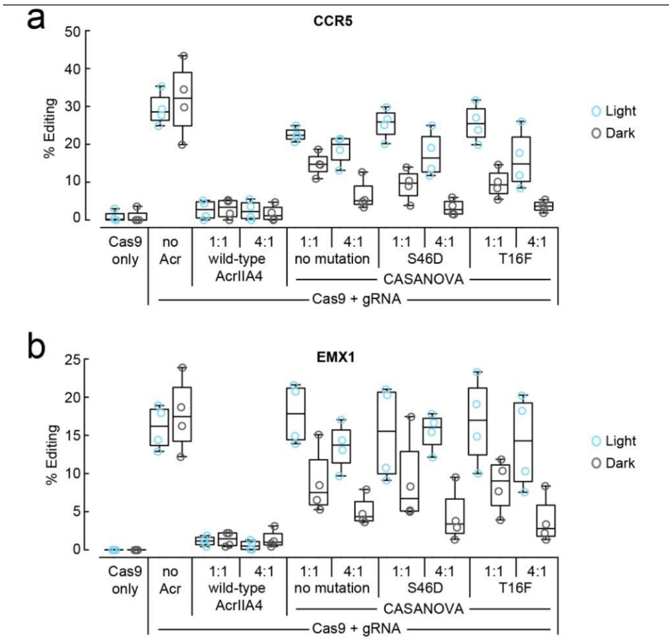
Supplementary Figure 2.7: Cas9 inhibition can be modulated via mutations that affect docking of the LOV2 terminal helices. **a:** Light-dependent luciferase reporter cleavage mediated by different Acr–LOV hybrid mutants. HEK293T cells were co-transfected with plasmids encoding (i) the indicated Acr—LOV hybrid variant, (ii) Cas9 and (iii) a luciferase reporter as well as a gRNA targeting the luciferase gene. Six hours post-transfection, cells were irradiated with pulsatile blue light for 48 h or kept in the dark as control before assessing luciferase activity. Box plots show the median (center line) and first and third quartiles (box edges), 1.5× the interquartile range (whiskers) and individual data points (circles). n = 3 biologically independent samples (cell cultures). **b:** T7 endonuclease assays and **c:** corresponding quantification of light-mediated indel mutation of the human CCR5 locus. HEK293T cells were co-transfected with constructs expressing the Cas9, the CCR5-locus-targeting gRNA and the indicated Acr–LOV hybrid variant. During transfection, the vector mass ratio of Acr–LOV:Cas9 construct was varied as indicated. Subsequently, cells were exposed to blue light for 70 h or kept in the dark as control. The target locus was then PCR-amplified with primers flanking the estimated break point and the amplicon was denatured and re-annealed in a thermocycler to allow heteroduplex formation. Following digestion with T7 endonuclease, samples were analyzed on a 2% agarose gel (see Methods for details). **c** Data are means and dots indicate individual data points. n = 2 independent experiments.



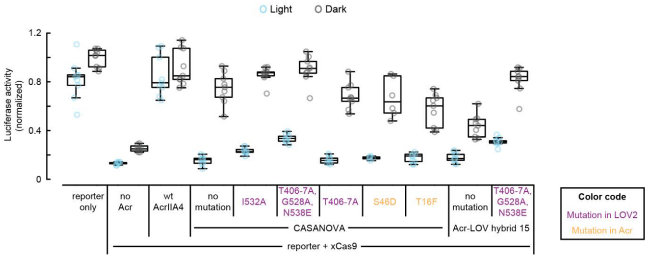
Supplementary Figure 2.8: Acr–LOV hybrids carrying a C450A LOV2 pseudodark mutation are still light responsive. **a:** Light-dependent luciferase reporter cleavage mediated by different Acr–LOV hybrid pseudodark mutants. HEK293T cells were co-transfected with plasmids encoding (i) the indicated Acr–LOV hybrid variant, (ii) Cas9 and (iii) a luciferase reporter as well as a gRNA targeting the luciferase gene. Six hours post-transfection, cells were irradiated with pulsatile blue light for 48 h or kept in the dark as control before assessing luciferase activity. Box plots show the median (center line) and first and third quartiles (box edges), 1.5× the interquartile range (whiskers) and individual data points (circles). n = 6 biologically independent samples (cell cultures). **b:** T7 endonuclease assays and **c:** corresponding quantification of light-mediated indel mutation of the human CCR5 locus. HEK293T cells were co-transfected with constructs expressing Cas9, CCR5-locus-targeting gRNA and the indicated Acr–LOV hybrid variant and exposed to blue light for 70 h or kept in the dark as control. During transfection, the vector mass ratio of Acr–LOV:Cas9 construct was varied as indicated. n.d., not determined. **b,c:** Data correspond to a single experiment.



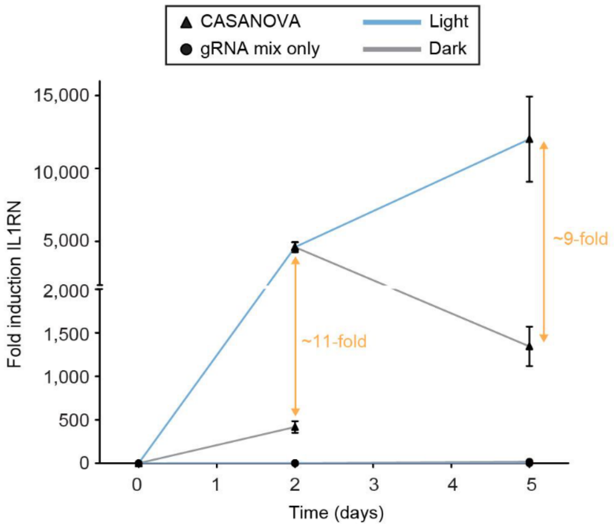
Supplementary Figure 2.9: *In silico* docking analysis reveals Acr mutations that improve CASANOVA performance. **a:** Light-dependent luciferase reporter cleavage mediated by different Acr–LOV hybrid mutants. HEK293T cells were co-transfected with vectors encoding (i) the indicated Acr–LOV hybrid variant, (ii) Cas9 and (iii) a luciferase reporter as well as a gRNA targeting the luciferase gene. Six hours post-transfection, cells were irradiated with pulsatile blue light for 48 h or kept in the dark as control before assessing the luciferase activity. Box plots show the median (center line) and first and third quartiles (box edges), 1.5× the interquartile range (whiskers) and individual data points (circles). n = 3 biologically independent samples (cell cultures). **b:** T7 endonuclease assays and **c:** corresponding quantification of light-mediated indel mutation of the human CCR5 locus. HEK293T cells were co-transfected with constructs expressing Cas9, the CCR5-locus-targeting gRNA and the indicated Acr–LOV hybrid variant and exposed to blue light for 70 h or kept in the dark as control. During transfection, the vector mass ratio of Acr–LOV:Cas9 construct was varied as indicated. Data represent a single experiment. n.d., not determined.



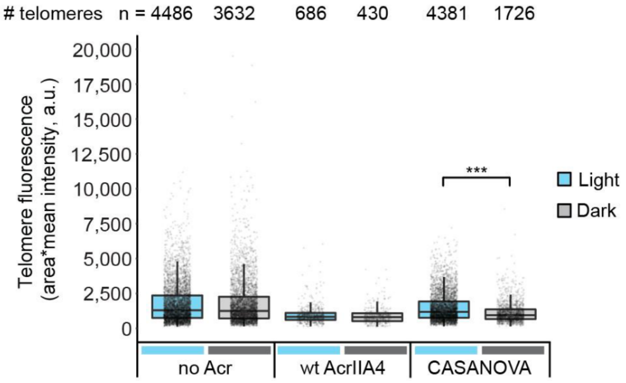
Supplementary Figure 2.10: Comparison of light-dependent indel mutation by CASANOVA and its corresponding S46D and T16F mutants. HEK293T cells were co-transfected with constructs expressing Cas9, a gRNA and the indicated CASANOVA variant and exposed to blue light for 70 h or kept in the dark as control. During transfection, the vector mass ratio of Acr–LOV:Cas9 construct was varied as indicated. Editing frequencies were evaluated by mismatch-sensitive T7 endonuclease assay. **a:** Indel mutation of CCR5 locus and **b:** indel mutation of EMX1 locus. **a,b:** Box plots show the median (center line) and first and third quartiles (box edges), 1.5× the interquartile range (whiskers) and individual data points (circles). n = 4 independent experiments.



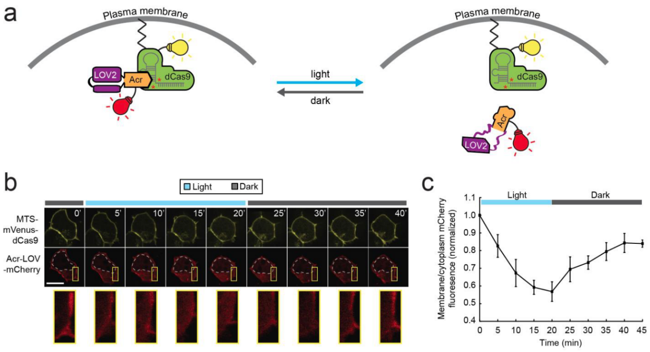
Supplementary Figure 2.11: Optogenetic control of xCas9. Light-dependent luciferase reporter cleavage mediated by different Acr–LOV hybrid mutants. HEK293T cells were co-transfected with vectors encoding (i) the indicated Acr–LOV hybrid variant, (ii) xCas9 and (iii) a luciferase reporter as well as a gRNA targeting the luciferase gene. Six hours post-transfection, cells were irradiated with pulsatile blue light for 48 h or kept in the dark as control before assessing luciferase activity. Box plots show the median (center line) and first and third quartiles (box edges), 1.5× the interquartile range (whiskers) and individual data points (circles). n = 6 biologically independent samples (cell cultures) for CASANOVA (S46D) and n = 9 biologically independent samples (cell cultures) for all other conditions.



Supplementary Figure 2.12: CASANOVA-mediated IL1RN gene activation is reversible. Light-dependent IL1RN gene activation in HEK293T cells expressing CASANOVA and a dCas9–p300 fusion targeted to the IL1RN promoter via a combination of four gRNAs (triangles) or the gRNA mix only (round dots; control). Cells were exposed to blue light or kept in the dark as indicated in the figure, and IL1RN expression was assessed by quantitative RT-PCR at the indicated time points. Gene activation in the light-induced CASANOVA sample (day 2) continues to rise when prolonging illumination, but decreases upon withdrawal of the light stimulus. Data are means ± s.e.m. n = 7 biologically independent samples (cell cultures).



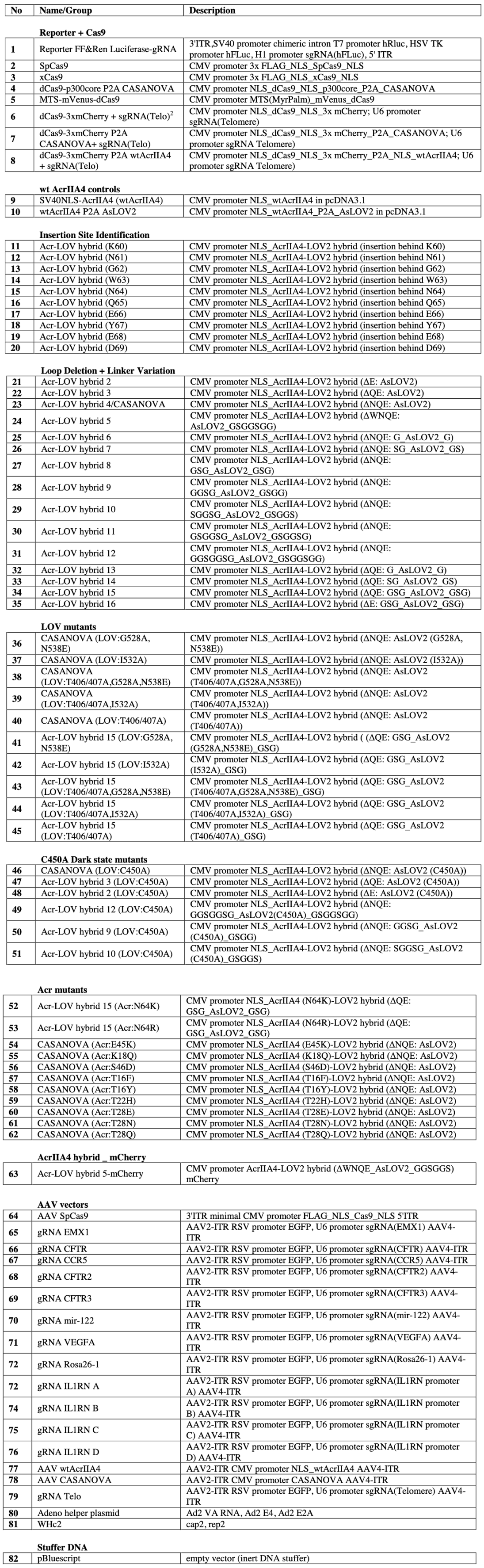
Supplementary Figure 2.13: Fluorescence signal of single labeled telomeres increased after CASANOVA activation. We calculated telomere fluorescence by multiplying the area of the fluorescent spot with its mean fluorescence intensity. Box plots show the median (center line) and first and third quartiles (box edges), 1.5× the interquartile range (whiskers) and individual data points (dots). Data correspond to the experiment described in Fig. 2 d,e. The total number of telomeres analyzed is indicated. ***P < 2.2 × 10–16 by a two-sided Wilcoxon rank-sum test.



Supplementary Figure 2.14: Reversible recruitment of Acr–LOV hybrid 5 to a plasma-membrane-targeted dCas9–gRNA complex. **a:** Light-inducible recruitment of a LOV–Acr hybrid fused to mCherry to a plasma-membrane-targeted dCas9–mVenus. **b:** Representative fluorescence images of HEK293T cells expressing LOV–Acr hybrid 5 fused to mCherry and an mVenus–dCas9 fusion targeted to the plasma membrane. Cells were irradiated with blue light pulses every 30 s using a 488-nm laser for 20 min, followed by 20 min dark recovery. mVenus and mCherry fluorescence images were recorded every 5 min. Dashed lines indicate the nucleus boundary. Yellow boxes in the Acr–LOV–mCherry images correspond to plasma membrane close-up views shown below the images. Scale bar, 20 µm. **c:** Quantification of the plasma membrane to cytoplasmic mCherry fluorescence ratio over time. Data are means ± s.e.m. n = 4 cells from biologically independent samples (cell cultures).

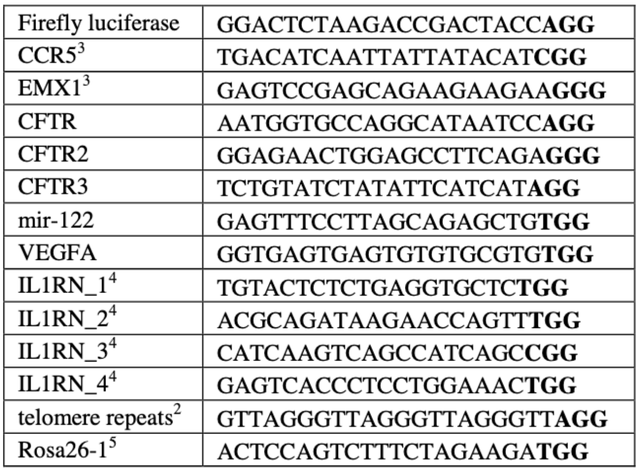
Construct	**Rosetta score**	**ddG**	**dHbond_gain_overall**
baseline	-1434.482	-124.294	0
T16Y	-1412.166	-124.173	1
T16F	-1432.37	-125.793	0
K18Q	-1435.027	-125.65	3
T22H	-1432.85	-125.192	0
T28E	-1436.272	-124.224	0
T28N	-1433.502	-125.348	1
T28Q	-1440.055	-125.657	0
E45K	-1438.617	-124.158	1
S46D	-1396.953	-122.129	1
N64K	-1435.808	-125.231	0
N64R	-1409.696	-124.297	1

Supplementary Figure 2.15: CASANOVA mutants selected for experimental characterization. ddG indicates the predicted change in free energy upon binding to the Cas9/gRNA complex. The dHbond_gain_overall shows the number of additionally formed buried hydrogen bonds of the designs compared to the wild-type (baseline).



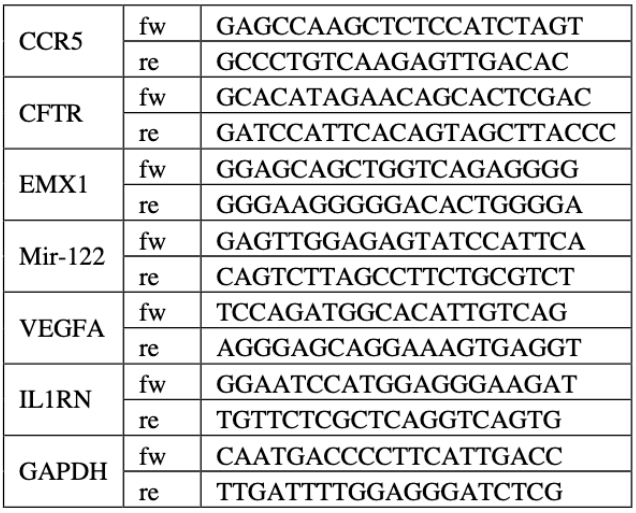
No	Name/Group	Description
	Reporter + Cas9	
1	Reporter FF&Ren Luciferase-gRNA	3'ITR,SV40 promoter chimeric intron T7 promoter hRluc, HSV TK promoter hFLuc, H1 promoter sgRNA(hFLuc), 5' ITR
2	SpCas9	CMV promoter 3x FLAG_NLS_SpCas9_NLS
3	xCas9	CMV promoter 3x FLAG_NLS_xCas9_NLS
4	dCas9-p300core P2A CASANOVA	CMV promoter NLS_dCas9_NLS_p300core_P2A_CASANOVA
5	MTS-mVenus-dCas9	CMV promoter MTS(MyrPalm)_mVenus_dCas9
6	dCas9-3xmCherry + sgRNA(Telo)[2]	CMV promoter NLS_dCas9_NLS_3x mCherry; U6 promoter sgRNA(Telomere)
7	dCas9-3xmCherry P2A CASANOVA+ sgRNA(Telo)	CMV promoter NLS_dCas9_NLS_3x mCherry_P2A_CASANOVA; U6 promoter sgRNA Telomere)
8	dCas9-3xmCherry P2A wtAcrIIA4 + sgRNA(Telo)	CMV promoter NLS_dCas9_NLS_3x mCherry_P2A_NLS_wtAcrIIA4; U6 promoter sgRNA Telomere)
	wt AcrIIA4 controls	
9	SV40NLS-AcrIIA4 (wtAcrIIA4)	CMV promoter NLS_wtAcrIIA4 in pcDNA3.1
10	wtAcrIIA4 P2A AsLOV2	CMV promoter NLS_wtAcrIIA4_P2A_AsLOV2 in pcDNA3.1
	Insertion Site Identification	
11	Acr-LOV hybrid (K60)	CMV promoter NLS_AcrIIA4-LOV2 hybrid (insertion behind K60)
12	Acr-LOV hybrid (N61)	CMV promoter NLS_AcrIIA4-LOV2 hybrid (insertion behind N61)
13	Acr-LOV hybrid (G62)	CMV promoter NLS_AcrIIA4-LOV2 hybrid (insertion behind G62)
14	Acr-LOV hybrid (W63)	CMV promoter NLS_AcrIIA4-LOV2 hybrid (insertion behind W63)
15	Acr-LOV hybrid (N64)	CMV promoter NLS_AcrIIA4-LOV2 hybrid (insertion behind N64)
16	Acr-LOV hybrid (Q65)	CMV promoter NLS_AcrIIA4-LOV2 hybrid (insertion behind Q65)
17	Acr-LOV hybrid (E66)	CMV promoter NLS_AcrIIA4-LOV2 hybrid (insertion behind E66)
18	Acr-LOV hybrid (Y67)	CMV promoter NLS_AcrIIA4-LOV2 hybrid (insertion behind Y67)
19	Acr-LOV hybrid (E68)	CMV promoter NLS_AcrIIA4-LOV2 hybrid (insertion behind E68)
20	Acr-LOV hybrid (D69)	CMV promoter NLS_AcrIIA4-LOV2 hybrid (insertion behind D69)
	Loop Deletion + Linker Variation	
21	Acr-LOV hybrid 2	CMV promoter NLS_AcrIIA4-LOV2 hybrid (ΔE: AsLOV2)
22	Acr-LOV hybrid 3	CMV promoter NLS_AcrIIA4-LOV2 hybrid (ΔQE: AsLOV2)
23	Acr-LOV hybrid 4/CASANOVA	CMV promoter NLS_AcrIIA4-LOV2 hybrid (ΔNQE: AsLOV2)
24	Acr-LOV hybrid 5	CMV promoter NLS_AcrIIA4-LOV2 hybrid (ΔWNQE: AsLOV2_GSGGSGG)
25	Acr-LOV hybrid 6	CMV promoter NLS_AcrIIA4-LOV2 hybrid (ΔNQE: G_AsLOV2_G)
26	Acr-LOV hybrid 7	CMV promoter NLS_AcrIIA4-LOV2 hybrid (ΔNQE: SG_AsLOV2_GS)
27	Acr-LOV hybrid 8	CMV promoter NLS_AcrIIA4-LOV2 hybrid (ΔNQE: GSG_AsLOV2_GSG)
28	Acr-LOV hybrid 9	CMV promoter NLS_AcrIIA4-LOV2 hybrid (ΔNQE: GGSG_AsLOV2_GSGG)
29	Acr-LOV hybrid 10	CMV promoter NLS_AcrIIA4-LOV2 hybrid (ΔNQE: SGGSG_AsLOV2_GSGGS)
30	Acr-LOV hybrid 11	CMV promoter NLS_AcrIIA4-LOV2 hybrid (ΔNQE: GSGGSG_AsLOV2_GSGGSG)
31	Acr-LOV hybrid 12	CMV promoter NLS_AcrIIA4-LOV2 hybrid (ΔNQE: GGSGGSG_AsLOV2_GSGGSGG)
32	Acr-LOV hybrid 13	CMV promoter NLS_AcrIIA4-LOV2 hybrid (ΔQE: G_AsLOV2_G)
33	Acr-LOV hybrid 14	CMV promoter NLS_AcrIIA4-LOV2 hybrid (ΔQE: SG_AsLOV2_GS)
34	Acr-LOV hybrid 15	CMV promoter NLS_AcrIIA4-LOV2 hybrid (ΔQE: GSG_AsLOV2_GSG)
35	Acr-LOV hybrid 16	CMV promoter NLS_AcrIIA4-LOV2 hybrid (ΔE: GSG_AsLOV2_GSG)
	LOV mutants	
36	CASANOVA (LOV:G528A, N538E)	CMV promoter NLS_AcrIIA4-LOV2 hybrid (ΔNQE: AsLOV2 (G528A, N538E))
37	CASANOVA (LOV:I532A)	CMV promoter NLS_AcrIIA4-LOV2 hybrid (ΔNQE: AsLOV2 (I532A))
38	CASANOVA (LOV:T406/407A,G528A,N538E)	CMV promoter NLS_AcrIIA4-LOV2 hybrid (ΔNQE: AsLOV2 (T406/407A,G528A,N538E))
39	CASANOVA (LOV:T406/407A,I532A)	CMV promoter NLS_AcrIIA4-LOV2 hybrid (ΔNQE: AsLOV2 (T406/407A,I532A))
40	CASANOVA (LOV:T406/407A)	CMV promoter NLS_AcrIIA4-LOV2 hybrid (ΔNQE: AsLOV2 (T406/407A))
41	Acr-LOV hybrid 15 (LOV:G528A, N538E)	CMV promoter NLS_AcrIIA4-LOV2 hybrid ((ΔQE: GSG_AsLOV2 (G528A,N538E)_GSG)
42	Acr-LOV hybrid 15 (LOV:I532A)	CMV promoter NLS_AcrIIA4-LOV2 hybrid (ΔQE: GSG_AsLOV2 (I532A)_GSG)
43	Acr-LOV hybrid 15 (LOV:T406/407A,G528A,N538E)	CMV promoter NLS_AcrIIA4-LOV2 hybrid (ΔQE: GSG_AsLOV2 (T406/407A,G528A,N538E)_GSG)
44	Acr-LOV hybrid 15 (LOV:T406/407A,I532A)	CMV promoter NLS_AcrIIA4-LOV2 hybrid (ΔQE: GSG_AsLOV2 (T406/407A,I532A)_GSG)
45	Acr-LOV hybrid 15 (LOV:T406/407A)	CMV promoter NLS_AcrIIA4-LOV2 hybrid (ΔQE: GSG_AsLOV2 (T406/407A)_GSG)
	C450A Dark state mutants	
46	CASANOVA (LOV:C450A)	CMV promoter NLS_AcrIIA4-LOV2 hybrid (ΔNQE: AsLOV2 (C450A))
47	Acr-LOV hybrid 3 (LOV:C450A)	CMV promoter NLS_AcrIIA4-LOV2 hybrid (ΔQE: AsLOV2 (C450A))
48	Acr-LOV hybrid 2 (LOV:C450A)	CMV promoter NLS_AcrIIA4-LOV2 hybrid (ΔE: AsLOV2 (C450A))
49	Acr-LOV hybrid 12 (LOV:C450A)	CMV promoter NLS_AcrIIA4-LOV2 hybrid (ΔNQE: GGSGGSG_AsLOV2(C450A)_GSGGSGG)
50	Acr-LOV hybrid 9 (LOV:C450A)	CMV promoter NLS_AcrIIA4-LOV2 hybrid (ΔNQE: GGSG_AsLOV2 (C450A)_GSGG)
51	Acr-LOV hybrid 10 (LOV:C450A)	CMV promoter NLS_AcrIIA4-LOV2 hybrid (ΔNQE: SGGSG_AsLOV2 (C450A)_GSGGS)
	Acr mutants	
52	Acr-LOV hybrid 15 (Acr:N64K)	CMV promoter NLS_AcrIIA4 (N64K)-LOV2 hybrid (ΔQE: GSG_AsLOV2_GSG)
53	Acr-LOV hybrid 15 (Acr:N64R)	CMV promoter NLS_AcrIIA4 (N64R)-LOV2 hybrid (ΔQE: GSG_AsLOV2_GSG)
54	CASANOVA (Acr:E45K)	CMV promoter NLS_AcrIIA4 (E45K)-LOV2 hybrid (ΔNQE: AsLOV2)
55	CASANOVA (Acr:K18Q)	CMV promoter NLS_AcrIIA4 (K18Q)-LOV2 hybrid (ΔNQE: AsLOV2)
56	CASANOVA (Acr:S46D)	CMV promoter NLS_AcrIIA4 (S46D)-LOV2 hybrid (ΔNQE: AsLOV2)
57	CASANOVA (Acr:T16F)	CMV promoter NLS_AcrIIA4 (T16F)-LOV2 hybrid (ΔNQE: AsLOV2)
58	CASANOVA (Acr:T16Y)	CMV promoter NLS_AcrIIA4 (T16Y)-LOV2 hybrid (ΔNQE: AsLOV2)
59	CASANOVA (Acr:T22H)	CMV promoter NLS_AcrIIA4 (T22H)-LOV2 hybrid (ΔNQE: AsLOV2)
60	CASANOVA (Acr:T28E)	CMV promoter NLS_AcrIIA4 (T28E)-LOV2 hybrid (ΔNQE: AsLOV2)
61	CASANOVA (Acr:T28N)	CMV promoter NLS_AcrIIA4 (T28N)-LOV2 hybrid (ΔNQE: AsLOV2)
62	CASANOVA (Acr:T28Q)	CMV promoter NLS_AcrIIA4 (T28Q)-LOV2 hybrid (ΔNQE: AsLOV2)
	AcrIIA4 hybrid _ mCherry	
63	Acr-LOV hybrid 5-mCherry	CMV promoter AcrIIA4-LOV2 hybrid (ΔWNQE_AsLOV2_GGSGGS) mCherry
	AAV vectors	
64	AAV SpCas9	3'ITR minimal CMV promoter FLAG_NLS_Cas9_NLS 5'ITR
65	gRNA EMX1	AAV2-ITR RSV promoter EGFP, U6 promoter sgRNA(EMX1) AAV4-ITR
66	gRNA CFTR	AAV2-ITR RSV promoter EGFP, U6 promoter sgRNA(CFTR) AAV4-ITR
67	gRNA CCR5	AAV2-ITR RSV promoter EGFP, U6 promoter sgRNA(CCR5) AAV4-ITR
68	gRNA CFTR2	AAV2-ITR RSV promoter EGFP, U6 promoter sgRNA(CFTR2) AAV4-ITR
69	gRNA CFTR3	AAV2-ITR RSV promoter EGFP, U6 promoter sgRNA(CFTR3) AAV4-ITR
70	gRNA mir-122	AAV2-ITR RSV promoter EGFP, U6 promoter sgRNA(mir-122) AAV4-ITR
71	gRNA VEGFA	AAV2-ITR RSV promoter EGFP, U6 promoter sgRNA(VEGFA) AAV4-ITR
72	gRNA Rosa26-1	AAV2-ITR RSV promoter EGFP, U6 promoter sgRNA(Rosa26-1) AAV4-ITR
72	gRNA IL1RN A	AAV2-ITR RSV promoter EGFP, U6 promoter sgRNA(IL1RN promoter A) AAV4-ITR
74	gRNA IL1RN B	AAV2-ITR RSV promoter EGFP, U6 promoter sgRNA(IL1RN promoter B) AAV4-ITR
75	gRNA IL1RN C	AAV2-ITR RSV promoter EGFP, U6 promoter sgRNA(IL1RN promoter C) AAV4-ITR
76	gRNA IL1RN D	AAV2-ITR RSV promoter EGFP, U6 promoter sgRNA(IL1RN promoter D) AAV4-ITR
77	AAV wtAcrIIA4	AAV2-ITR CMV promoter NLS_wtAcrIIA4 AAV4-ITR
78	AAV CASANOVA	AAV2-ITR CMV promoter CASANOVA AAV4-ITR
79	gRNA Telo	AAV2-ITR RSV promoter EGFP, U6 promoter sgRNA(Telomere) AAV4-ITR
80	Adeno helper plasmid	Ad2 VA RNA, Ad2 E4, Ad2 E2A
81	WHc2	cap2, rep2
	Stuffer DNA	
82	pBluescript	empty vector (inert DNA stuffer)

Supplementary Figure 2.16: List of constructs created and used in this study. ITR, inverted terminal repeat. TK, thymidine kinase. FF, Firefly. Ren, Renilla. MTS, membrane-targeting sequence.



Firefly luciferase	GGACTCTAAGACCGACTACC**AGG**
CCR5[3]	TGACATCAATTATTATACAT**CGG**
EMX1[3]	GAGTCCGAGCAGAAGAAGAA**GGG**
CFTR	AATGGTGCCAGGCATAATCC**AGG**
CFTR2	GGAGAACTGGAGCCTTCAGA**GGG**
CFTR3	TCTGTATCTATATTCATCAT**AGG**
mir-122	GAGTTTCCTTAGCAGAGCTG**TGG**
VEGFA	GGTGAGTGAGTGTGTGCGTG**TGG**
IL1RN_1[4]	TGTACTCTCTGAGGTGCTC**TGG**
IL1RN_2[4]	ACGCAGATAAGAACCAGTT**TGG**
IL1RN_3[4]	CATCAAGTCAGCCATCAGC**CGG**
IL1RN_4[4]	GAGTCACCCTCCTGGAAAC**TGG**
telomere repeats[2]	GTTAGGGTTAGGGTTAGGGTT**AGG**
Rosa26-1[5]	ACTCCAGTCTTTCTAGAAGA**TGG**

Supplementary Figure 2.17: gRNA target sites. Sequences are in 5’ to 3’ direction; the PAM sequence is indicated in bold. References indicate the publications originally reporting the corresponding gRNAs.



CCR5	fw	GAGCCAAGCTCTCCATCTAGT
	re	GCCCTGTCAAGAGTTGACAC
CFTR	fw	GCACATAGAACAGCACTCGAC
	re	GATCCATTCACAGTAGCTTACCC
EMX1	fw	GGAGCAGCTGGTCAGAGGGG
	re	GGGAAGGGGGACACTGGGGA
Mir-122	fw	GAGTTGGAGAGTATCCATTCA
	re	CAGTCTTAGCCTTCTGCGTCT
VEGFA	fw	TCCAGATGGCACATTGTCAG
	re	AGGGAGCAGGAAAGTGAGGT
IL1RN	fw	GGAATCCATGGAGGGAAGAT
	re	TGTTCTCGCTCAGGTCAGTG
GAPDH	fw	CAATGACCCCTTCATTGACC
	re	TTGATTTTGGAGGGATCTCG

Supplementary Figure 2.18: Primers used for genomic PCRs and quantitative RT-PCR. Sequences are in 5’ to 3’ direction.

2.4.1 Supplementary Note 1

Solvent accessibility analysis of the AcrIIA4-Cas9 complex suggested that the most C-terminal loop segment of AcrIIA4 (L5 in Fig. 2.1b) was the only insertion site amenable to LOV2 domain fusion, since all the other insertion points would cause steric clashes with Cas9. Detailed analysis of residue contacts (Online Methods) within the AcrIIA4 structure further consolidated this choice. The secondary structure elements bridged by L5 have a network of non-local contacts that, when destabilized, will impose structural distortions to the Cas9-binding surface of AcrIIA4. In turn, these would very likely result in a decrease in affinity to Cas9 (Fig. 2.1b). Finally, several residues in L5 (Y67, D69 and E70) mediate direct and critical contacts with Cas9 [202]. We thus speculated that LOV2 insertion in proximity to this functional site would offer an efficient means to control Cas9 binding.

2.4.2 Supplementary Note 2: Considerations for luciferase reporter cleavage assay data interpretation

For the luciferase reporter cleavage assay, we used a vector expressing (i) Firefly luciferase, (ii) a gRNA targeting the Firefly luciferase gene and, additionally, (iii) Renilla luciferase, which is employed for normalization purposes (Online methods). In this assay, the Cas9-targeted locus is supplied as a plasmid, many copies of which will enter a cell upon efficient transfection. Importantly, cells transfected not at all or only inefficiently are effectively excluded in this assay, as they will express low amounts of the luciferase reporters and will therefore contribute much less to the measurement outcome. In contrast, the very efficiently transfected cells (which will typically strongly express all required components (Cas9, gRNA, luciferase reporter, Acr-LOV2 hybrid)) contribute most to the measurement outcome. This renders the luciferase reporter cleavage assay extremely robust, even under the heterogeneous condition of a transient transfection.

2.4.3 Supplementary Note 3: Considerations for T7 assay data interpretation

In contrast to the luciferase reporter assay (see Supplementary Note 2 above), the T7 endonuclease assay measures the indel frequency for an endogenous locus targeted by Cas9. Unless some type of selection is applied to enrich transfected cells, which we did not do, all cells will equally contribute to the T7 assay result, regardless of whether they express the Cas9, gRNA and Acr-LOV transgenes strongly, weakly or not at all. Moderately transfected cells often show heterogeneity in transgene expression, i.e. the stoichiometry of Cas9-gRNA complexes and Acr-LOV hybrids present will highly vary between cells. This can lead to unintended Cas9 leakiness or – on the contrary - overly strong Cas9 inhibition in individual cells. Still, these cells will eventually contribute as much to the measurement outcome, as the most efficiently transfected ones.

2.4.4 Supplementary Note 4

Numerous mutations have been reported that improve docking of the AsLOV2 terminal helices against the LOV core in the dark [204, 203, 219, 220, 221]. It is important to note, however, that the effect they have on the performance of a particular optogenetic tool is difficult to predict a priori. To this end, we performed a systematic mutational study using sets of well-studied mutations: The T406-7A double mutation improving docking of the Aα helix [203] and two different mutations within the Jα helix, i.e., the I532A single mutation or G528A/N538E double mutation [204]. We tested the effects of each individual mutation (or double mutation) as well as the impact of their combinations on the performance of CASANOVA as well as LOV2-Acr hybrid 15. We observed that Cas9 background activity in the dark was reduced in several mutants, albeit this improvement came at the cost of a reduced dynamic range in most cases (Sup. Fig. 2.7). We and others have observed this trade-off between “leakiness” in the dark and dynamic range of activation before for several of the tested mutations in the context of unrelated optogenetic tools [203, 222]. Importantly, the effects of the tested mutations were similar for CASANOVA and Acr-LOV hybrid 15, i.e., mutations that caused a strong background activity in one Acr-LOV hybrid, for instance, did the same in the context of the other (Sup. Fig. 2.7).

2.4.5 Supplementary Note 5

The LOV2 cysteine 450 forms a covalent adduct with the flavin mononucleotide chromophore upon blue light excitation [217, 217], which is a key step in the LOV2 photocycle and triggers unfolding of the LOV2 terminal helices [196, 223, 224, 225]. Mutants of the C450 to alanine, methionine and serine are known to lock the LOV2 in a dark state-like conformation [226, 227] (pseudodark state) and are thus often employed as negative control when developing optogenetic constructs [228, 198, 229]. In the context of engineered, LOV2-dependent split inteins, it has recently been found that pseudodark mutants can still be excited upon irradiation for several hours and, strikingly, may even enhance the performance of an optogenetic tool [230]. Provided the experimental timing tolerates the slow activation kinetics (which is, however, often not the case for optogenetic experiments), pseudodark mutants are therefore an unconventional, but interesting resource for improving LOV2-based optogenetic tools. To test this concept in the context of our light-dependent Cas9 inhibitor, we introduced the C450A mutation into the LOV2 part of CASANOVA as well as five other Acr-LOV hybrids. Remarkably, all variants still showed strong, light-dependent Cas9 inhibition (Sup. Fig. 2.8). The CASANOVA C450A even outperformed the wild-type construct regarding its ability to inhibit Cas9 in the dark, albeit light-activation was markedly reduced (Sup. Fig. 2.8c).

2.4.6 Supplementary Note 6: Comment on the used experimental timings

For the telomere labeling experiment, we adapted the experimental timing from the study by Pawluk *et al.* [208], as we employed their optimized dCas9-3xRFP vector, the identical cell line (U2OS) and a similar experimental setup. Pawluk and co-workers performed microscopy analysis 24 h post-transfection. We started illumination 4 h post transfection and performed microcopy after 20 h of light induction (or 20 h of incubation in the dark), which sums up to the identical 24 h used by Pawluk and colleagues. This rather short time from transfection to measurement was further chosen as in this particular experiment, maximum expression (i.e. high dCas9-3xRFP levels) is not desired. This is because only a certain amount of dCas9-3xRFP molecules can bind to telomeres. Consequently, once the telomeres are efficiently occupied, any additional molecules will mainly result in nuclear background fluorescence and therefore lower the signal-to-background ratio. To enable robust, automated image analysis, however, a high signal-to-background ratio was critical.

For the T7 assays, the scenario is fundamentally different because here, efficient expression of Cas9 is crucial to enable efficient editing. Furthermore, only loci that have been repaired via non-homologous end-joining (NHEJ) following editing and thus carry mutations (mostly indels) will be detected in this assay. However, this repair requires additional time. Therefore, in transfection-based genome editing experiments, the incubation time used is typically 72 h or more. We used 70 h, which is largely congruent with the extensive CRISPR literature and also comparable to the timing (72 h) in the CRISPR landmark paper by Cong *et al.* [231].

For the luciferase reporter, which was newly developed by us, the timing is generally less critical as this assay is rather robust and thus tolerates fluctuations in transfection efficiencies or total expression levels (see Supplementary Note 2 above). Furthermore, this assay is independent of NHEJ. Therefore, since the experimental timing is overall less of a concern, we used 48 h of illumination (i.e. 48 h of incubation after transfection) in most cases, as expression of all components should have peaked at this time point. An exception is Sup. Fig. 2.3, where we used 30 h of illumination instead of 48 h. This, however, had no noticeable influence on the general performance of the assay, further evidencing its inherent robustness. Finally, for the gene activation experiments, we chose the 48 h incubation time (4 h incubation in the dark followed by 44 h in the light (or dark)) based on previous reports on Cas9-mediated gene activation [232, 233, 234] as well as a recent report [235] of a chemically inducible dCas9-p300 fusion. Notably, the fold induction in IL1RN expression we observed upon blue light induction (~7,000-fold; Fig. 2a) precisely corresponds to what Hilton *et al.* also reported as maximum activation of the same gene by dCas9-p300, namely, ~6,000 – 10,000-fold induction [206].

2.4.7 Supplementary Note 7

The full-length gel images corresponding to Sup. Figs. 2.5, 2.7, 2.8 and 2.9 are shown in Sup. Fig. 2.19 below.

2.4.8 Supplementary Discussion

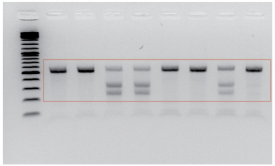
We developed CASANOVA, an engineered, light-inducible *Spy*Cas9 inhibitory protein and demonstrated its utility for controlling genome editing as well as recruitment of dCas9-effector fusions to selected genomic loci in human cells. Due to its small size of only 27.8 kDa (similar to GFP), CASANOVA can be easily expressed from plasmids or viral vectors to achieve robust and efficient Cas9 regulation. Interestingly, the tightness and dynamic range of genome editing control was generally higher when using AAV transduction for delivery as compared to transient transfection (Fig. 1d-f). The reason could be the stronger heterogeneity in expression of the different components upon transient transfection, likely resulting in unintended Cas9 activity in the dark in cells expressing high Cas9/gRNA, but rather low CASANOVA levels (see Supplementary Notes 2 and 3 above).

A particularly notable feature of CASANOVA is its versatility. Without any modification, CASANOVA allowed light-mediated control of *Spy*Cas9, xCas9 [205] as well as *Spy*dCas9-effector fusions, as exemplified for a Cas9-p300 epigenetic modifier. Remarkably, CASANOVA also enabled us to directly monitor the light-induced recruitment of dCas9-RFP fusions to telomeres, thereby confirming Cas9 initial binding to be a rapid process in human cells (Fig. 2.2b,c). Our data thereby complement previous live-cell studies, that investigated the replacement of dCas9 molecules bound to on and off-target loci via FRAP and spectroscopy [236, 237, 238], but not the binding of Cas9 to previously unbound loci. We note that in the telomere labeling experiments, we observed heterogeneity in dCas9-3xRFP nucleocytoplasmic localization in all samples. This is likely attributed to the large size of dCas9-3xRFP (~ 250 kDa), which renders this fusion protein a rather difficult target for nuclear import. Nevertheless, robust telomere labeling was observed in the positive control samples and in the CASANOVA sample induced with blue light. The fact that the CASANOVA dark control sample or wild-type AcrIIA4 controls showed hardly any telomere labeling, as expected, indicates that the heterogeneity in dCas9-3xRFP localization had no negative impact on assay performance (Fig. 2.2b-e).

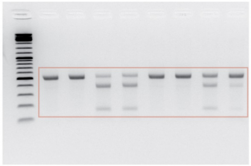
Another important feature of CASANOVA that should benefit numerous applications is its reversibility. We showed that CASANOVA rapidly releases Cas9 upon blue light stimulation and re-binds Cas9 when stopping light induction (Sup. Fig. 2.14), thereby regaining Cas9 inhibition (Sup. Fig. 2.12). To estimate the kinetics of CASANOVA reversibility, however, it is essential to realize that they do not solely depend on the kinetics of CASANOVA conformational change and consecutive binding of/release from Cas9-gRNA complexes. Akin to its parent anti-CRISPR protein AcrIIA4, CASANOVA acts as competitive inhibitor, i.e., it competes with the target DNA for free (d)Cas9-gRNA complexes [200, 199, 201, 239]. Very importantly, while AcrIIA4 binds free (d)Cas9-gRNA complexes with high affinity, it is unable to actively displace (d)Cas9-gRNA from bound DNA target loci [239]. Consequently, dark-state CASANOVA will most likely not be able to actively resolve (d)Cas9-gRNA:DNA ternary complexes that formed, e.g., during a preceding illumination phase. Notably, the half-life of the dCas9-gRNA:DNA complexes is in the range of several hours, at least for perfect target sites31. We thus expect the release of (d)Cas9-gRNA from the DNA target locus to bethe rate-limiting step during dark-state adaptation of the CASANOVA system upon withdrawal of the light trigger.

Apart from CASANOVA's manifold applications, this work also expands our ability to confer light regulation on selected proteins via LOV2-mediated, inducible disorder. When engineering CASANOVA, the choice of Acr surface sites amenable for LOV2 insertion was highly restricted by the small size, compact structure and limited solvent accessibility of AcrIIA4 when in complex with Cas9. The insertion site in AcrIIA4 loop 5 finally chosen based on our computational analysis directly precedes several, functional residues (Y67, D69, E70) mediating critical contacts with Cas9 [200, 201]. Importantly, in previous work by Dagliyan *et al.* on controlling mammalian enzymes [198], LOV2 domain insertion in close proximity to the enzyme's active site was purposely avoided, as even minor structural perturbations at such sites are likely to strongly impair enzymatic function. Concurrently, we observed that inserting the LOV2 domain into AcrIIA4 loop 5 resulted in a markedly impaired Cas9 inhibition even in the dark. However, we were able to restore close to wild-type inhibitor activity by carefully "embedding" the LOV2 domain into the target protein structure via systematic deletions of residues preceding the insertion site. Thereby, the proximity of the fused LOV2 domain to the AcrIIA4 functional site, initially imposed by the target protein structure, suddenly turned into an advantage, as the conformational change of the LOV2 was now directly coupled to AcrIIA4 functional residues. The robustness of CASANOVA performance under different experimental conditions suggests that this unconventional design could be an interesting blueprint for the optogenetic regulation of diverse Acrs and, potentially, many other proteins of interest.

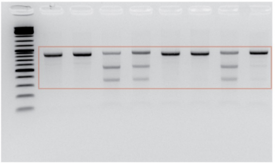
Supplementary Figure 5a



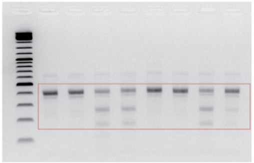
Supplementary Figure 5b

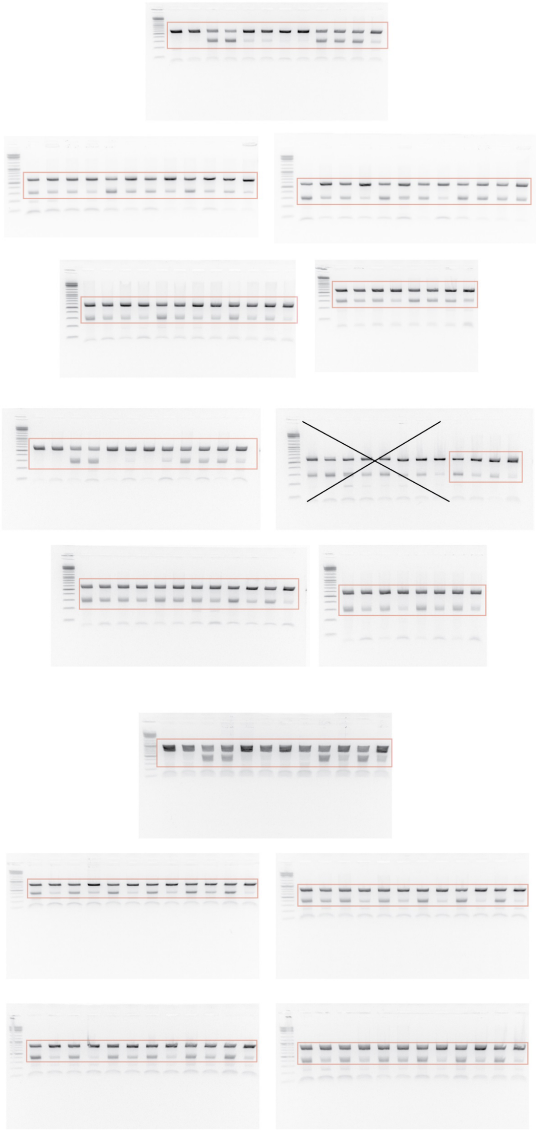


Supplementary Figure 5c



Supplementary Figure 5d





Supplementary Figure 2.19: Full-length gel images. The ladder is the Gene Ruler DNA Ladder Mix (Thermo Fisher).

3 Computational design of anti-CRISPR proteins with improved inhibition potency

...

This chapter is a postprint version based on an article published in Nature Chemical Biology in 2020 (DOI: 10.1038/s41589-020-0518-9) in accordance with the publisher.

Authors

Jan Mathony[1,2,*], **Zander Harteveld**[3,4,*], Carolin Schmelas[5,6,*], Julius Upmeier zu Belzen[1,2,7], Sabine Aschenbrenner[1,2,8], Wei Sun[9,10], Mareike D. Hoffmann[1,8], Christina Stengl[1], Andreas Scheck[3,4], Sandrine Georgeon[3,4], Stéphane Rosset[3,4], Yanli Wang[9,10], Dirk Grimm[5,6,11], Roland Eils[2,7], Bruno E. Correia[3,4] and Dominik Niopek[1,7]

[*] These authors contributed equally.

Affiliations

[1]Synthetic Biology Group, BioQuant Center, University of Heidelberg, Heidelberg, DE. [2]Digital Health Center, Berlin Institute of Health (BIH) and Charité, Berlin, DE. [3]Institute of Bioengineering, École Polytechnique Fédérale de Lausanne, Lausanne, CH. [4]Swiss Institute of Bioinformatics (SIB), Lausanne, CH. [5]Department of Infectious Diseases, Virology, University Hospital Heidelberg, Heidelberg, DE. [6]BioQuant Center and Cluster of Excellence CellNetworks at Heidelberg University, Heidelberg, DE. [7]Health Data Science Unit, University Hospital Heidelberg, Heidelberg, DE. [8]Department of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, DE. [9]National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing, CN. [10]Key Laboratory of RNA Biology, Institute of Biophysics, Chinese Academy of Sciences, Beijing, CN. [11]German Center for Infection Research (DZIF) and German Center for Cardiovascular Research (DZHK), Partner Site Heidelberg, Heidelberg, DE.

Author contributions

D.N. conceived the initial idea and refined it together with J.U.z.B. and B.E.C.. J.M., C. Schmelas, C. Stengl, S.A. and D.N. designed and performed experiments. S.A. and M.D.H. performed AAV production. S.R. and S.G. purified the Acrs and performed protein-biochemical characterization. Z.H., J.U.z.B., A.S. and B.E.C. performed *in silico* structural analysis and modeling. W.S. and Y.W. designed and performed in vitro DNA binding and cleavage assays. D.G. provided expertise on AAVs and *Sau*Cas9. R.E. supported the computational and experimental work. B.E.C. and D.N. jointly directed the work. B.E.C. and D.N. wrote the manuscript with support from all authors.

3.1 Abstract

Anti-CRISPR (Acr) proteins are powerful tools to control CRISPR–Cas technologies. However, the available Acr repertoire is limited to naturally occurring variants. Here, we applied structure-based design on AcrIIC1, a broad-spectrum CRISPR–Cas9 inhibitor, to improve its efficacy on different targets. We first show that inserting exogenous protein domains into a selected AcrIIC1 surface site dramatically enhances inhibition of *Neisseria meningitidis* (*Nme*)Cas9. Then, applying structure-guided design to the Cas9-binding surface, we converted AcrIIC1 into AcrIIC1X, a potent inhibitor of the *Staphylococcus aureus* (*Sau*)Cas9, an orthologue widely applied for *in vivo* genome editing. Finally, to demonstrate the utility of AcrIIC1X for genome engineering applications, we implemented a hepatocyte-specific *Sau*Cas9 ON-switch by placing AcrIIC1X expression under regulation of microRNA-122. Our work introduces designer Acrs as important biotechnological tools and provides an innovative strategy to safeguard CRISPR technologies.

3.2 Main

The detailed characterization of bacterial CRISPR–Cas systems [240] and their adaptation for precise genome engineering in mammalian cells [233, 231] has revolutionized the life sciences and enabled novel applications in biotechnology and medicine. The recent discovery of Acr proteins [241, 242, 243], that is, potent inhibitors of Cas effectors, provides a shut-off mechanism that can keep this powerful technology in check [194] and enhance the precision at which genome perturbations can be made [162, 244, 245, 239, 246]. Acrs originate from the coevolution of prokaryotes and phages. Bacteria employ the CRISPR adaptive immune system to destroy invading nucleic acids. Phages, on the other hand, circumvent the CRISPR defense by suppressing the activity of essential CRISPR components via Acrs. Mining of sequence databases and screening of phage libraries proved to be powerful strategies to discover Acrs that target CRISPR–Cas orthologues from various species [242, 243, 192, 247, 248, 249, 250, 251, 252]. However, these approaches are inherently limited to the naturally occurring protein repertoire. Moreover, for various Cas effectors of major biotechnological interest, nature might be lacking (efficient) Acr counterparts.

In this work, we applied protein engineering to design artificial Acrs that exhibit enhanced

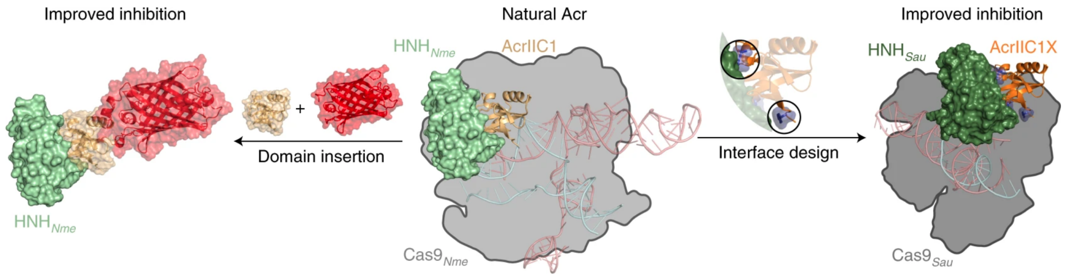


Figure 3.1: Improving the efficacy of Acr proteins by protein engineering. Domain insertion into AcrIIC1 loop 5 yields chimeric inhibitors showing improved inhibition of *Nme*Cas9, while structure-based engineering of its HNH-binding interface results in a potent *Sau*Cas9 inhibitor (PDB 4ZIN, 5VGB, 5F9R and 5CZZ).

inhibition potency on different Cas9 orthologues (Fig. 3.1). As starting point, we used AcrIIC1 (ref. [208]), a broad-spectrum inhibitor targeting various type II-C Cas9s (ref. [253]), including those from *Neisseria meningitidis*, *Geobacillus stearothermophilus* and *Campylobacter jejuni*. We show that domain insertion into a selected AcrIIC1 surface site dramatically enhances inhibition of *Neisseria meningitidis* (*Nme*)Cas9. On top, re-designing the Cas9-binding surface yielded AcrIIC1X, a potent *Staphylococcus aureus* (*Sau*)Cas9 inhibitor. Our work complements the discovery of natural Acrs by providing engineering strategies to improve Acr function for biotechnological applications.

3.3 Results

3.3.1 Domain insertion into AcrIIC1 boosts *Nme*Cas9 inhibition

AcrIIC1 binds the conserved catalytic HNH domain and locks Cas9 in a DNA binding-competent but catalytically inactive state. This unique inhibitory mechanism might explain why AcrIIC1 is a rather weak inhibitor [248] compared with its related proteins AcrIIC3, -C4 and -C5, which either interfere with Cas9 DNA binding [248, 253] or mediate cleavage of the Cas9-loaded single guide (sg)RNA20. Importantly, biochemical assays suggest tight binding of AcrIIC1 to the *Nme*Cas9 HNH domain [253]. We hypothesized that inserting an exogenous domain into AcrIIC1 could perturb the conformational freedom and/or stability of the components in the Cas9–sgRNA complex. This, in turn, could enhance Cas9 inhibition, provided the inhibitor would still bind tightly to the HNH domain (Fig. 3.1, left). To test this hypothesis, we first investigated the structure of the AcrIIC1–HNH domain complex for AcrIIC1 surface sites that could be amenable for domain insertion. AcrIIC1 loop 5 appeared to be an ideal candidate (Sup. Fig. 3.1), as this loop is surface exposed and distal from the HNH-interacting surface necessary for the activity that we wished to preserve. We then created 11 different AcrIIC1 domain fusions that carry an mCherry (~27 kDa), *Avena sativa* LOV2 (~17 kDa) or PDZ domain (~9 kDa) in loop 5 and optional, flanking GS-linkers or short deletions (Sup. Fig. 3.2a). These chimeric Acrs were screened for their ability to inhibit *Nme*Cas9 cleavage of the IL2RG locus in HEK 293T cells (Sup. Fig. 3.2b).

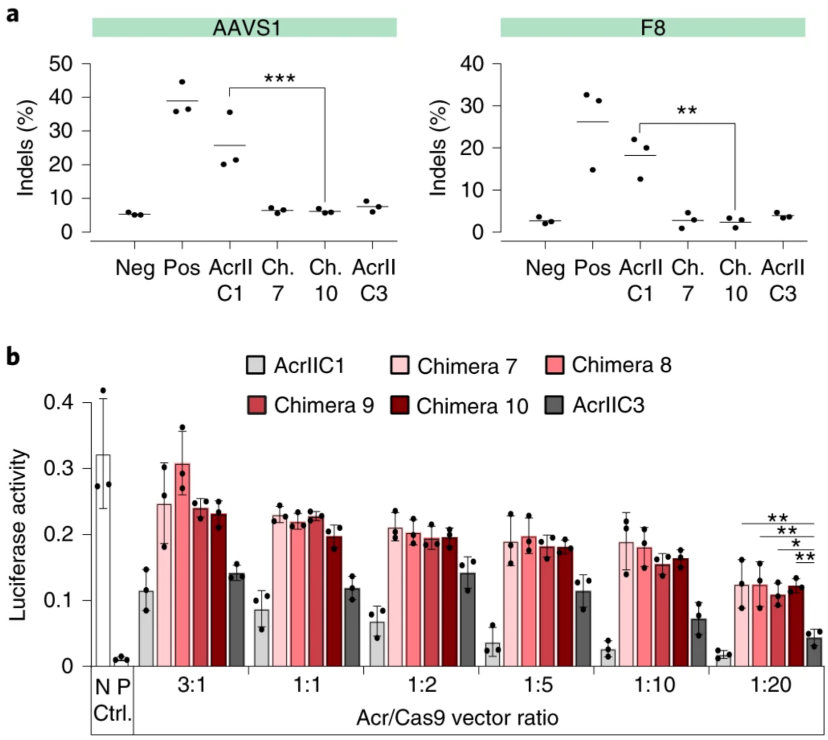


Figure 3.2: Domain insertion into AcrIIC1 yields a highly potent *Nme*Cas9 inhibitor. **a:** HEK 293T cells were cotransfected with plasmids expressing *Nme*Cas9, an sgRNA targeting the indicated locus as well as the indicated Acr. The vector mass ratio used during transfection was Acr/Cas9 = 1:4. At 72 h post-transfection, indel frequencies were analyzed by TIDE sequencing [202]. Lines in the plots indicate means, dots individual data points for n = 3 independent experiments. Neg, negative control (Cas9 only); Pos, positive control (Cas9 + sgRNA); Ch., AcrIIC1-mCherry chimera. **P < 0.01, ***P < 0.001 by one-way ANOVA with Bonferroni correction. **b:** AcrIIC1-mCherry chimeras outperform wild-type AcrIIC1 and AcrIIC3. Cells were cotransfected with vectors encoding *Nme*Cas9, a firefly luciferase reporter and corresponding reporter-targeting sgRNA as well as the indicated Acr followed by luciferase assay. Bars indicate means, error bars the s.d. for n = 3 independent experiments. **a,b:** N, negative (Cas9 only (a) or reporter only (b) control (Ctrl.)); P, positive control (Cas9 + sgRNA (a) or reporter + Cas9 (b)). *P < 0.05, **P < 0.01, by one-way ANOVA with Bonferroni correction. AAVS1, adeno-associated virus integration site 1; F8, Factor VIII gene.

To measure Cas9 inhibition, we employed tracking of indels by decomposition (TIDE) sequencing [202], which is a quantitative assay previously shown to correlate well with next-generation sequencing (NGS) data of indel measurements [254]. Complementarily, we also used T7 endonuclease I assay, which has been shown to have a rather limited correlation to NGS data [254] and was thus only used for qualitative assessment of Cas9 inhibition.

Remarkably, several AcrIIC1-LOV2 and AcrIIC1-mCherry chimeras mediated highly potent *Nme*Cas9 inhibition, largely exceeding that of the parent AcrIIC1, as indicated by TIDE (Fig. 3.2a) and qualitatively confirmed by T7 endonuclease assay (Sup. Figs. 3.2b–3.4).

Of note, the chimera containing the PDZ domain showed reduced inhibition compared with wild-type AcrIIC1 (Sup. Figs. 3.1b and 3.3a). The AcrIIC1-mCherry chimera no. 10 (Sup. Fig. 3.2a) showed a particularly strong improvement in *Nme*Cas9 inhibition in comparison with wild-type AcrIIC1 on the tested loci (Fig. 3.2a and Sup. Fig. 3.3). Moreover, this chimera was also superior to wild-type AcrIIC1 when tested on *Nme*2Cas9 (Sup. Fig. 3.5), which, in contrast to the four-nucleotide *Nme*Cas9 protospacer adjacent motif (PAM), only requires a dinucleotide PAM [255].

To further characterize the gain in inhibition, we employed a reporter assay in which *Nme*Cas9 cleaves a firefly luciferase transgene, thereby resulting in luciferase knockout. We cotransfected cells with the reporter, *Nme*Cas9 and either the parent AcrIIC1 or different, engineered AcrIIC1-mCherry chimeras. As a benchmark, we also included AcrIIC3 (Fig. 2b), so far the most potent *Nme*Cas9 inhibitor in mammalian cells [208]. During transfection, we varied the Acr/Cas9 vector ratios from 3:1 to 1:20. The chimeric inhibitors outperformed both wild-type AcrIIC1 as well as AcrIIC3 and showed potent Cas9 inhibition even at very low Acr/Cas9 vector ratios (Fig. 3.2b). Together, these experiments demonstrate that domain insertion can yield Acr proteins with superior inhibition potency than that of natural Acrs, thereby enabling extremely tight control of Cas9 activity.

3.3.2 Mechanistic insights into effects of domain insertion

To investigate the reasons behind this gain in inhibition potency, we first explored the inhibitory mechanism of AcrIIC1-mCherry. Electro mobility shift assays and complementary in vitro DNA cleavage with purified protein showed that, similar to wild-type AcrIIC1 (ref. [253]), AcrIIC1-mCherry chimera no. 10 blocks Cas9 catalytic function, but does not interfere with DNA binding (Fig. 3.3a,b). Of note, we observed that both AcrIIC1 and AcrIIC1-mCherry were able to impair DNA cleavage only upon pre-incubation of Acr with Cas9 before the addition of sgRNA, but not when the Acr was added to pre-assembled Cas9–sgRNA complexes (Fig. 3.3b and see Discussion). Importantly, AcrIIC1-mCherry was able to efficiently block Cas9-mediated DNA cleavage at lower Acr/Cas9 molar ratios as compared with wild-type AcrIIC1 (Fig. 3.3b), showing that the chimera is more potent in impairing cleavage activity. Next, we performed western blot experiments to assess the impact of domain insertion on Acr protein levels in cells. The chimeric inhibitor was expressed at higher levels compared with wild-type AcrIIC1, which likely contributes to the particularly potent inhibition observed *in vivo* (Fig. 3.3c).

These data indicate that AcrIIC1-mCherry retains the same inhibitory mechanism of wild-type AcrIIC1 but does so with increased inhibitory potency of the Cas9 catalytic activity, and abundance of the chimeric Acr in cells (Fig. 3.3c).

Finally, we also assessed the impact of the Acrs on *Nme*Cas9 protein levels. We found that *Nme*Cas9 levels were strongly reduced (>2-fold) in presence of both AcrIIC1 and AcrIIC1-mCherry, but not in presence of AcrIIA4 (an Acr specific to *Streptococcus pyogenes* (*Spy*)Cas9,

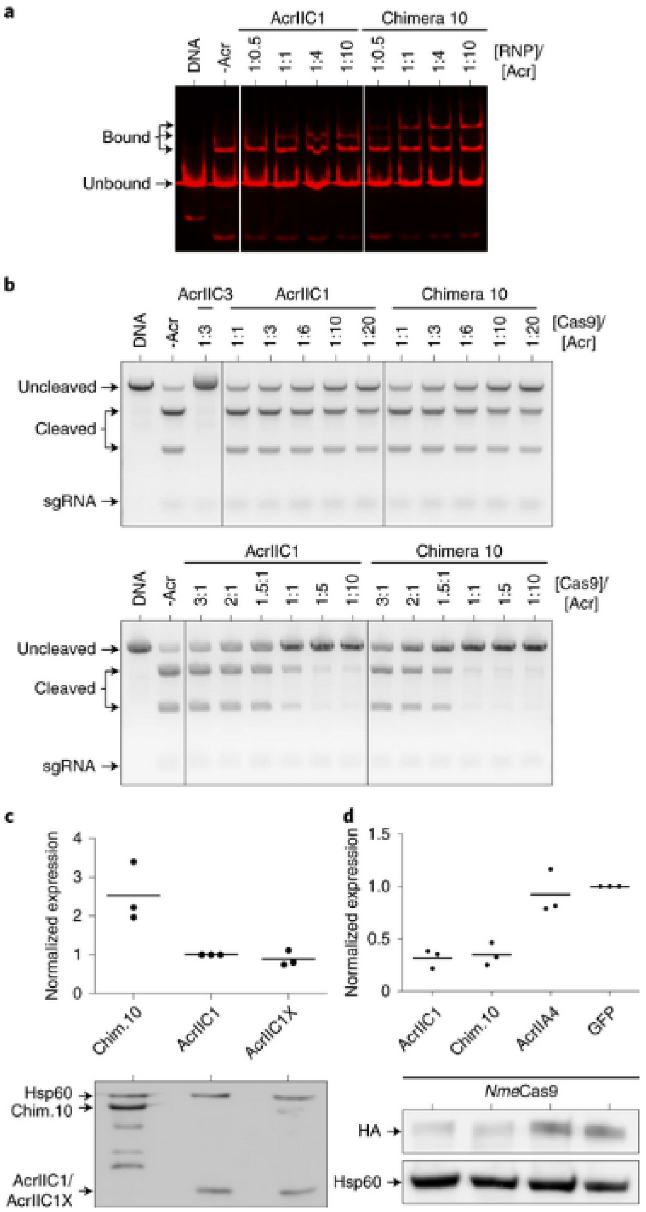


Figure 3.3: Enhanced potency of the AcrIIC1-mCherry chimeric inhibitor arises from multiple factors. **a:** Cy3-labeled targeting strand and 5-Cy5-labeled NTS were annealed and incubated with *Nme*Cas9, sgRNA and varying amounts of the indicated Acr. Cas9 DNA binding was analyzed on a native gel using Cy5 as readout. **b:** *In vitro* DNA cleavage assays conducted in presence of AcrIIC1 or AcrIIC1-mCherry chimera 10. The sgRNA was added to *Nme*Cas9 either before the Acr (top) or afterwards (bottom). **a,b:** Data correspond to a single experiment. **c:**, Analysis of Acr expression in HEK 293T cells by western blot 2 d post-transfection. Full-length gel image is shown in Sup. Fig. 3.18. **d:** Constructs encoding HA-tagged *Nme*Cas9 and (untagged) Acrs were cotransfected in a vector mass ratio of 1:1. *Nme*Cas9 protein levels were analyzed by western blot 2 d post-transfection. **c,d:** Top, lines in plots indicate means, dots individual data points, n = 3 independent experiments. Bottom, representative western blot image. Chim., chimera.

used as control) or GFP (control) (Fig. 3.3d). This suggests that on top of preventing Cas9 from taking on a cleavage-competent conformation, the presence of AcrIIC1-mCherry as well as wild-type AcrIIC1 reduces Cas9 protein levels, indicating that these Acrs might limit Cas9 expression and/or stability (see Discussion).

3.3.3 Computational design yields a potent *Sau*Cas9 inhibitor

The Cas9 from *S. aureus* is a type II-A CRISPR effector widely employed for *in vivo* genome editing [256]. Due to its favorable, small size (3.2 kb), *Sau*Cas9 can easily be packaged into Adeno-associated virus (AAV) particles [256], which are prime vector candidates for therapeutic CRISPR applications [216, 257]. We speculated that AcrIIC1 might represent an ideal starting point to engineer an artificial *Sau*Cas9 inhibitor (Fig. 3.1, right), as the overall structure of the *Sau*Cas9 HNH domain is similar to that of *Nme*Cas9 (ref. [253]), although substantial differences exist at the sequence level (sequence identity is only 33.7%; Sup. Fig. 3.6). Recent data indicate that AcrIIC1 can inhibit *Sau*Cas9 function, albeit incompletely [252]. To independently confirm *Sau*Cas9 inhibition by AcrIIC1 in human cells, we coexpressed AcrIIC1 in HEK 293T together with *Sau*Cas9 and sgRNAs targeting different loci, and then performed T7 endonuclease assays. Editing was still observed in presence of AcrIIC1, albeit lower as compared with the positive control (Sup. Fig. 3.7). This suggested that AcrIIC1 can also bind the *Sau*Cas9 HNH domain, though likely with a compromised affinity. This functional observation was further confirmed by surface plasmon resonance affinity measurements of recombinantly expressed and purified AcrIIC1 and the HNH domains of *Sau*Cas9 and *Nme*Cas9. We observed a striking difference when comparing the affinities of AcrIIC1 to *Nme*Cas9 HNH (K_D = 0.95 nM) and *Sau*Cas9 HNH (K_D = 370 nM) (Sup. Fig. 3.12i,k), explaining the low efficacy of AcrIIC1 on the *Sau*Cas9 target. To rationalize the affinity difference, we generated a structural model of the *Sau*Cas9 HNH domain in complex with AcrIIC1 and investigated the AcrIIC1 interacting surface as compared with the *Nme*Cas9 HNH domain (Fig. 3.4a). Two regions in *Sau*Cas9 HNH domain showed suboptimal contacts to corresponding AcrIIC1 residues (Fig. 3.4a). To screen for residue variants that could optimize the interfacial contacts in these two regions, we performed *in silico* mutagenesis using Rosetta design [258] followed by manual inspection, which suggested ten AcrIIC1 candidate mutations predicted to improve binding to the *Sau*Cas9 HNH domain (Sup. Figs. 3.8 and 3.9). We tested these mutants, first individually, in genome editing experiments targeting the EMX1 locus and then iteratively combined the most promising variants in subsequent screening rounds (Sup. Fig. 3.10). Of note, we decreased the Acr/Cas9 ratio with each screening round to better resolve the performance of improved candidates. After only three rounds, we arrived at a triple mutant referred to as AcrIIC1X (N3F, D15Q, A48I), which achieved a near-complete blockage of EMX1 editing (Sup. Fig. 3.10). According to our structural model, these point mutants improved the shape and chemical complementarity of the Cas9-interacting surface, providing a rational basis for the enhanced activity (Fig. 3.4b,c and Sup. Fig. 3.11). Importantly, similar to AcrIIC1, AcrIIC1X was monomeric (Sup. Fig. 3.12a–d) and well folded in solution according to circular dichroism spectroscopy (Sup. Fig. 3.12e–h). In line with our design hypothesis, AcrIIC1X exhibited improved affinity to the

*Sau*Cas9 HNH domain (K_D = 53 nM) as compared with wild-type AcrIIC1 (K_D = 370 nM) (Sup. Fig. 3.12k,l). While affinity of AcrIIC1X to the *Nme*Cas9 HNH domain was likewise reduced, it was still in the low nanomolar range (K_D = 6.9 nM; Sup. Fig. 3.12j).

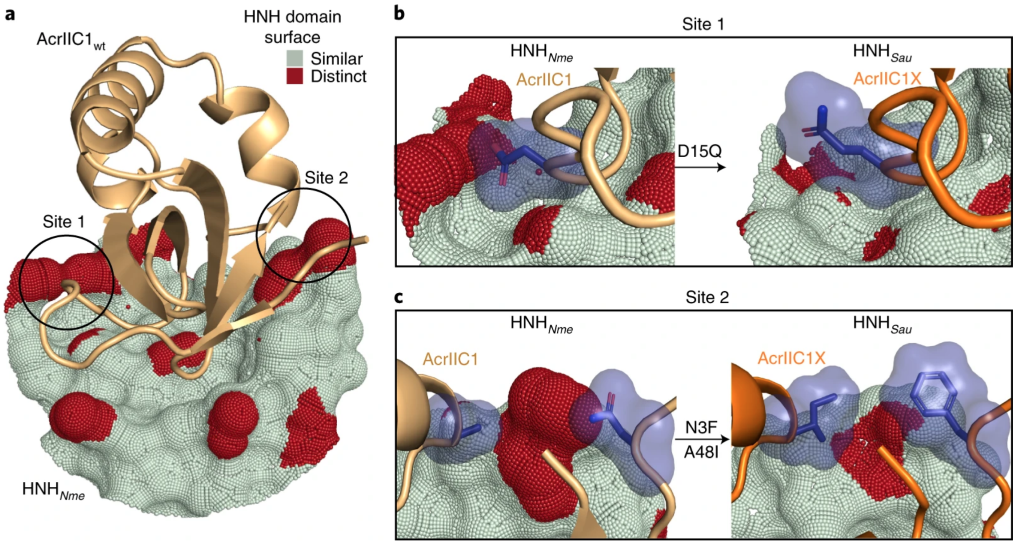


Figure 3.4: Structure-guided design of AcrIIC1X, an Acr protein targeting *Sau*Cas9. **a** Structure showing AcrIIC1 binding to the *Nme*Cas9 HNH domain surface. Red patches indicate regions at which the *Sau*Cas9 HNH surface displays deviations of at least 1 Å as compared with the *Nme*Cas9 HNH surface, highlighting the most important sites to target by mutagenesis. **b,c:** Comparison of wild-type AcrIIC1 residues (D15 (b), N3 and A48 (c)) binding to the *Nme*Cas9 HNH surface with the corresponding, engineered residues (15Q (b), 3F and 48I (c)) binding to the *Sau*Cas9 HNH surface. The designed mutants exhibited improved shape and chemical complementarity to *Sau*Cas9 HNH, providing a structural rationale for enhancements in binding affinity.

Next, we characterized AcrIIC1X performance in detail by targeting *Sau*Cas9 to different loci (Fig. 3.5a). AcrIIC1X efficiently suppressed *Sau*Cas9 genome editing at all tested loci, showing highly improved inhibition as compared with the parental AcrIIC1 according to TIDE measurements (Fig. 3.5b) and qualitatively confirmed by T7 endonuclease assays (Sup. Fig. 3.14a). Akin to wild-type AcrIIC1, AcrIIC1X was unable to fully inhibit *Nme*Cas9-mediated editing (Sup. Fig. 3.15). Fusion of an mCherry domain to AcrX loop 5 gave rise to a chimera (AcrIIC1X*), which was able to block *Nme*Cas9-mediated editing, while likewise maintaining *Sau*Cas9 inhibition (Sup. Fig. 3.15).

Subsequently, to test the performance of AcrIIC1X in different mammalian cell lines and upon viral delivery, we packaged (1) AcrIIC1X as well as (2) *Sau*Cas9 and sgRNAs targeting the EMX1, Grin2B or CXCR4 locus into AAV serotype 2. We then transduced HEK 293T (human embryonic kidney), U2OS (human osteosarcoma) or U87 (human primary glioblastoma) cells with these vectors and found that indel formation was reduced to undetectable levels in practically all samples that received AcrIIC1X (Fig. 3.5c and Sup. Fig. 3.14b).

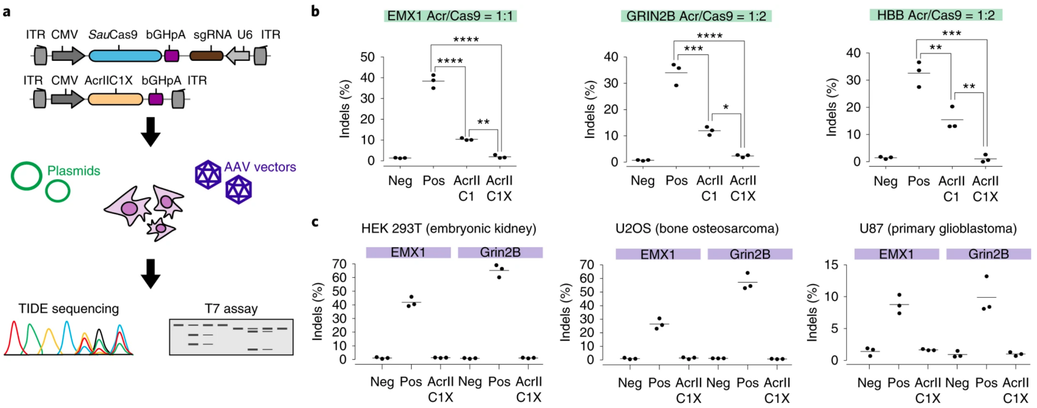


Figure 3.5: Characterization of AcrIIC1X, a designer *Sau*Cas9 inhibitor. **a:** Schematics of vectors and experimental setup. **b:** HEK 293T cells were cotransfected with plasmids expressing *Sau*Cas9; an sgRNA targeting the EMX1, GRIN2B or HBB locus; and either AcrIIC1 or AcrIIC1X, followed by TIDE sequencing [202]. *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001 by one-way ANOVA with Bonferroni correction. **c:** Cells were cotransduced with AAV vectors expressing (1) *Sau*Cas9 and an sgRNA targeting the indicated loci and (2) AcrIIC1X, followed by TIDE sequencing. **b,c:** Lines in the plots indicate means, dots individual data points for n = 3 independent experiments. ITR, inverted terminal repeat.

3.3.4 A hepatocyte-specific *Sau*Cas9 ON-switch

Finally, to demonstrate the great potential of AcrIIC1X for genome engineering applications, we aimed at employing AcrIIC1X to confine *Sau*Cas9 activity to selected cell types. To this end, we build on an approach we refer to as Cas-ON, which was previously developed by us [244] and others [259, 260]. Cas-ON harnesses cell-specific microRNAs (miRNAs), such as miR-122 which is solely expressed in hepatocytes, to de-target Acr expression from selected cell types [244]. To this end, binding sites for a given miRNA are inserted into the 3 UTR of an Acr transgene. Upon codelivery of the miRNA-dependent Acr transgene, Cas9 and an sgRNA, Acr expression is knocked down by RNA interference specifically in the target cell type, thereby permitting Cas9 to be active. In any off-target cell type lacking the miRNA trigger, the Acr is expressed and Cas9 is thus blocked. Importantly, the Cas-ON approach can be applied to cultured cells [244, 259, 260] and was recently demonstrated to function in mice as well [259]. Thus far, the Cas-ON strategy was only implemented for *Spy*Cas9 and *Nme*Cas9 using AcrIIA4, AcrIIC1 and AcrIIC3 (refs. [244, 259, 260]).

We hypothesized that AcrIIC1X should facilitate adapting the Cas-ON approach to *Sau*Cas9, which holds enormous potential for *in vivo* genome editing applications [256]. To this end, we inserted target sites for miR-122 into the 3 UTR of the AcrIIC1X transgene (or AcrIIC1). As control we employed Acr constructs with a 3 UTR of identical length, but lacking the miR-122 target sites (scaffold).

To assess whether the miR-dependent AcrIIC1X enables release of *Sau*Cas9 activity selectively in the presence of miR-122, we first cotransfected plasmids encoding *Sau*Cas9, an sgRNA targeting the EMX1 locus and the Acr transgenes into HEK 293T cells, which naturally do not express miR-122. We also artificially overexpressed miR-122 or a control miRNA (miR-155) and investigated indel formation at the target locus. In samples transfected with the miR-122-dependent AcrIIC1X construct, editing was efficiently suppressed in the absence of miR-122 (OFF state), but released almost to the levels of the positive control in the presence of miR-122 (ON state; Sup. Fig. 3.16a,b). In contrast, the miR-122-dependent AcrIIC1 showed substantial editing even in the absence of miR-122 (Sup. Fig. 3.16a,b), indicating that the OFF state of the Cas-ON switch is highly leaky when using AcrIIC1.

Finally, to investigate whether the miR-122-dependent AcrIIC1X facilitates cell type-specific editing, we packaged the different components of our Cas-ON system (AcrIIC1X or AcrIIC1 either with or without mir-122 target sites, *Sau*Cas9 and an sgRNA targeting EMX1) into AAV serotype 2. We then cotransduced Huh-7 cells, a hepatocyte-derived cell line naturally expressing high miR-122 levels [244], or HEK 293T cells (control cells) with these vectors and qualitatively measured genome editing efficiency in the presence of the different Acr variants by T7 endonuclease assay. In the Huh-7 samples, editing was observed in the presence of both the miR-122-dependent AcrIIC1 as well as AcrIIC1X (Sup. Fig. 3.16c). Very importantly, however, only the miR-122-dependent AcrIIC1X, but not AcrIIC1, potently inhibited editing in the off-target cell line (HEK 293T; Sup. Fig. 3.16d). These data demonstrate that AcrIIC1X, but not AcrIIC1, enables the implementation of the Cas-ON switch for *Sau*Cas9, showcasing the importance of highly potent Acrs in the context of biological applications.

3.4 Discussion

In this work, we applied protein engineering to improve the inhibition potency of AcrIIC1 for two different Cas9 orthologues, namely *Nme*Cas9 and *Sau*Cas9. By inserting exogenous protein domains into a computationally selected surface site on AcrIIC1, we first created chimeric Acrs most of which showed enhanced efficacy on the *Nme*Cas9 target (Fig. 3.2). The improvement in inhibition could be allocated to two synergistic effects. On the one hand, *in vitro* DNA cleavage assays showed that AcrIIC1-mCherry is more potent than wild-type AcrIIC1 in blocking Cas9 catalytic activity (Fig. 3.3b). This suggests that the fused mCherry domain improves the AcrIIC1 inhibitory mechanism, potentially by perturbing the Cas9 conformation. On the other hand, western blots indicated elevated protein levels for the chimeric inhibitor (Fig. 3.3c), which is likely to contribute to the observed improvement of inhibition in cells.

Apart from these particularly notable differences, we also made two interesting mechanistic observations that apply to both the chimeric Acrs and the parental AcrIIC1. First, we found that both efficiently block Cas9-mediated DNA cleavage *in vitro* only when pre-incubated with apo-Cas9 (that is, before adding the sgRNA). When applied to pre-assembled *Nme*Cas9–sgRNA complexes, however, AcrIIC1 can hardly block DNA cleavage (Fig. 3.3b). This suggests that the HNH domain is less accessible to the Acr in the sgRNA-bound Cas9 state, while upon addition of the Acr, the sgRNA can still bind. We speculate that this might be due to certain HNH domain conformations or orientations in the sgRNA-bound and -unbound states. Secondly, we found that the presence of AcrIIC1-mCherry as well as wild-type AcrIIC1 reduces Cas9 protein levels in cells, indicating that these Acrs might perturb Cas9 expression or stability. Of note, Cas9 degradation has very recently also been proposed as a mechanism underlying *Listeria monocytogenes* Cas9 inhibition by AcrIIA1 in bacteria [261]. This is particularly interesting as, similar to AcrIIC1, AcrIIA1 targets the Cas9 HNH domain.

We note that the Acr chimera bearing the smallest domain (PDZ) was weaker as compared with the parent AcrIIC1 (Sup. Fig. 3.2b). While we did not explore the reason behind the reduced efficacy in this single case, we speculate that the presence of the PDZ domain might slightly distort the AcrIIC1 structure or sterically clash with Cas9, resulting in decreased binding affinity.

When re-designing AcrIIC1 toward improved inhibition of *Sau*Cas9, it was interesting to see that although affinity to *Nme*Cas9 HNH domain dropped considerably (Sup. Fig. 3.12i,j), it remained sufficient as to facilitate *Nme*Cas9 inhibition in human cells (Sup. Fig. 3.15). Akin to its parent (AcrIIC1), *Nme*Cas9 inhibition by AcrIIC1X could also be further improved by domain insertion (AcrIIC1X*). Thus, apart from blocking *Sau*Cas9 function, AcrIIC1X and, in particular, AcrIIC1X* are also well suited for applications requiring simultaneous blockage of both *Nme*- and *Sau*Cas9. The promiscuity of AcrIIC1X/AcrIIC1X* can, however, also present a limitation; for example, when the aim is to independently control *Sau*Cas9 and *Nme*Cas9 function via multiple inhibitors. For such multiplexing applications, orthogonality would be desired.

Generally, we reason that the domain insertion approach presented here might be well suited for Acrs that already bind a given Cas orthologue with high affinity, but do not fully block Cas function. The engineering success will then mainly depend on the presence of a surface site on the Acr that (1) is amenable to domain insertion and (2) can interfere structurally with Cas regions whose accessibility or conformational freedom is critical for Cas function. Re-design of the binding surface, on the other hand, could aid in cases where the low binding affinity of a natural Acr would fail to potently inhibit the Cas orthologue of interest. We reason that surface re-design could be a promising approach for inhibitors that have considerable broad-spectrum activity [262, 251, 253] and already some residual inhibitory effect on the Cas orthologue of interest (as was the case for AcrIIC1 on *Sau*Cas9). Generally, this cross-reactivity is due to similarities at the structural level in the region targeted by the inhibitors. In the case of *Nme*Cas9 and *Sau*Cas9, the structure of the AcrIIC1-targeted HNH domain is indeed very similar (root-mean-square deviation (RMSD) of 1.06 Å (ref. [253])). In addition, we speculate that it should also be possible to re-design the Cas9 binding surface of AcrIIC1 towards structurally more distal orthologues such as *Spy*Cas9 (HNH domain structure RMSD of 3.0 Å (ref. [253])). However, this would likely require the computational and experimental sampling of a larger protein structural and sequence space.

Apart from re-designing natural Acrs, we note with excitement the rapid progress in the protein engineering field, in particular with respect to the *in silico* design of protein–protein interactions [174]. This suggests that, in the future, it might also be possible to create fully synthetic Acrs targeting selected Cas orthologues from scratch.

Recent reports show that *Sau*Cas9 can be inhibited by AcrIIA5 (refs. [252, 263, 264]) and AcrIIA13-15 (ref. [262]). Very importantly, our work is not to be seen as a potential replacement for the exploration of the natural Acr repertoires. On the contrary, we believe that structure-guided protein engineering will greatly complement these efforts by providing rational means to customize and improve CRISPR inhibitors beyond the limits of natural evolution. The resulting designer Acrs may find wide application in the context of biotechnology and CRISPR-based therapies [194, 162, 244, 245], and also provide an innovative strategy to safeguard CRISPR technology.

3.5 Methods

3.5.1 Modeling of AcrIIC1-mCherry fusions

We used the Rosetta remodel application [211] to generate the AcrIIC1-mCherry chimeras based on the structures for AcrIIC1 (PDB 5VGB) and mCherry (PDB 4ZIN). The N and C termini of mCherry were absent in the crystal structure and were rebuilt using fragment insertion together with cyclic coordinate descent [212] and kinematic closure [213, 214] with default values. For the designed chimera with a two-residue deletion, approximately 1,500 decoys were generated and subsequently clustered with an RMSD threshold of 5 Å into 27 clusters. For the chimera with additional GSG-linkers at the N and C termini, approximately 1,200 structures were clustered in 100 clusters with the same parameters. Representative examples of the three most populated clusters, which also have the lowest energies, are shown to illustrate the potential structural diversity of the AcrIIC1-mCherry chimeras (Sup. Fig. 3.4). Analyses of the Rosetta outputs, structural models and the biochemical data were performed using the rstoolbox [265].

3.5.2 AcrIIC1 interface design

To screen *in silico* for mutations that could enhance the affinity of AcrIIC1 to *Sau*Cas9, we modeled a complex of AcrIIC1 (originally crystalized with the HNH domain of *Nme*Cas9, PDB 5VGB) with the *Sau*Cas9 HNH domain (PDB 5CZZ). A first structural alignment was performed between the *Nme*- and *Sau*Cas9 domains using TM-align [266], revealing an RMSD of 2.32 Å and several structurally and sequence-conserved interface regions. These conserved interface regions were then used to refine the alignment, using the PyMOL (v.2.3.1) superposition function, to obtain the modeled complex used for the design simulations. We then analyzed the interfaces of both orthologues to pinpoint hotspots that could be designed in AcrIIC1 to enhance its interaction with *Sau*Cas9. These hotspots were visualized with surface point-wise distances between the two surfaces that were computed with a custom script. For each point on the reference surface (*Nme*), the distance to the closest point on the other surface (*Sau*) was calculated. In the visualizations, these distances were binarized by setting a cutoff of 1 Å (Fig. 3.4).

Next, we used RosettaScripts [267] to perform a single-site *in silico* mutagenesis, thereby allowing subsets of amino acids (AAs) for each of the selected residues on AcrIIC1. From our interface analysis we selected the following residues in AcrIIC1 for *in silico* mutagenesis: N3, D15, R36, D43, D45, D46, K47, A48 and M77. The design protocol consisted of two rounds of packing and minimization with fixed backbone. We generated a total of 51 designs and computed their change in binding free energy $\Delta\Delta$G (ddG), number of hydrogen bonds (H-bonds) across the interface, change in hydrophobic solvent-accessible surface area and interface shape complementarity with the target *Sau*Cas9 HNH domain. Designs with improved $\Delta\Delta$Gs compared with that of the AcrIIC1–*Sau*Cas9 complex (22 Rosetta energy units), increased hydrogen bonds across the interface, improvements in solvent-accessible surface area and shape complementarity were manually inspected. A total of ten substitutions on eight sites were selected for experimental validation. Mutations N3F, N3Y and A48I were designed to increase interface packing and π-stacking with the complementary hydrophobic patches on *Sau*Cas9. D43F and D45F were generated to fill voids within the interface boundaries and increase the hydrophobic packing. D15Q, R36D, D46E, K47Q and M77S were introduced to balance the underlying charge distribution on *Sau*Cas9 within the respective region. The overall design workflow is shown in Sup. Fig. 3.8.

After experimental validation and combination of the proposed mutations, the final AcrIIC1X (AcrIIC1 with N3F, D15Q, A48I) was modeled following the same protocol. Electrostatic properties for AcrIIC1, AcrIIC1X as well as the *Nme*Cas9 and the *Sau*Cas9 HNH domains were computed using the adaptive Poisson Boltzmann solver (APBS) plugin in PyMOL (Sup. Fig. 3.11). The mutation D15Q results in a less-negative potential, to optimize the interaction with a patch of the interface in *Sau*Cas9 that has a lower positive potential as compared with *Nme*Cas9.

3.5.3 Protein expression and purification

DNA sequences of the designs were purchased from Twist Bioscience. For bacterial expression, the DNA fragments were cloned via Gibson cloning [268] into a pET21b vector encoding a peptide sequence containing a tobacco etch virus (TEV) protease cleavage site followed by a terminal His-tag and transformed into *Escherichia coli* BL21(DE3). Expression was conducted in Terrific Broth supplemented with ampicillin (100 µg ml^{1}). Cultures were inoculated at an optical density (OD)600 of 0.1 from an overnight culture and incubated in a shaker at 37 °C and 220 r.p.m. After reaching an OD600 of 0.6, expression was induced by the addition of 0.4 mM IPTG and cells were further incubated overnight at 20 °C. Cells were harvested by centrifugation and pellets were resuspended in lysis buffer (50 mM TRIS, pH 7.5, 500 mM NaCl, 5% glycerol, 1 mg ml^{-1} lysozyme, 1 mM PMSF, 4 µg ml^{-1} DNase). Resuspended cells were sonicated and clarified by centrifugation. Ni-NTA purification of sterile-filtered (0.22

µm) supernatant was performed using a 5-ml His-Trap FF column on an ÄKTA pure system (GE Healthcare). Bound proteins were eluted using an imidazole concentration of 300 mM. Concentrated proteins were further purified by size exclusion chromatography on a Hiload 16/600 Superdex 75 pg column (GE Healthcare) using PBS buffer (pH 7.4) as mobile phase.

3.5.4 Circular dichroism

Far-UV circular dichroism spectra of AcrIIC1 and AcrIIC1X were collected between wavelengths of 190 and 250 nm on a Jasco J-815 circular dichroism spectrometer in a 1-mm pathlength quartz cuvette. Proteins were dissolved in 10 mM phosphate buffer at concentrations between 20 and 40 µM. Wavelength spectra were averaged from two scans with a scanning speed of 20 nm min^1 and a response time of 0.125 s. The thermal denaturation curves were collected by measuring the change in ellipticity at 220 nm from 20 to 90 °C with 2 or 5 °C increments.

3.5.5 Size-exclusion chromatography combined with multi-angle light scattering

Multi-angle light scattering was used to assess the monodispersity and molecular weight of the proteins. Samples containing 50–100 µg of protein in PBS buffer (pH 7.4) were injected into a Superdex 75 10/300 GL column (GE Healthcare) using an HPLC system (Ultimate 3000, Thermo Scientific) at a flow rate of 0.5 ml min^{-1} coupled in-line to a multi-angle light-scattering device (miniDAWN TREOS, Wyatt). Static light-scattering signal was recorded from three different scattering angles. The scatter data were analyzed by ASTRA software (version 6.1, Wyatt).

3.5.6 Affinity measurements

Surface plasmon resonance was used to determine the dissociation constants of the Acr designs to the *Sau*Cas9 and *Nme*Cas9 HNH domains. Experiments were performed on a Biacore 8K at room temperature with HBS-EP+ running buffer (10 mM HEPES pH 7.4, 150 mM NaCl, 3 mM EDTA, 0.005% v/v Surfactant P20) (GE Healthcare). Approximately 2,800 response units of *Sau*Cas9 HNH domain and 320 response units of *Nme*Cas9 HNH domain were immobilized via amine coupling on the methyl-carboxyl dextran surface of a CM5 chip (GE Healthcare). Varying protein concentrations were injected over the surface at a flow rate of 30 µl min^1 with a contact time of 120 s and a following-dissociation period of 600 s. Following each injection cycle, ligand regeneration was performed using 10 mM glycine pH 2.5 (GE Healthcare). Data analysis was performed using bivalent analyte kinetic fits within the Biacore evaluation software (GE Healthcare).

3.5.7 Mammalian expression construct design and cloning

Constructs used in this study are listed in Supplementary Table 2.15. Sequences for all plasmids created in this study are provided as GenBank files in the Supplementary Data 1. The following constructs were generated via classical restriction enzyme cloning or Golden Gate assembly [269]. Oligonucleotides and synthetic double-stranded DNAs were obtained from IDT. PCRs were performed either with Q5 Hot Start high-fidelity DNA polymerase (New England Biolabs) or Phusion Flash high-fidelity polymerase (Thermo Fisher Scientific). After separating PCR products or restriction digest products on agarose gels, bands of the desired size were cut out and the DNA was extracted using the QIAquick gel extraction kit (Qiagen). Restriction enzymes and T4 DNA ligase were obtained from Thermo Fisher Scientific. Constructs were transformed into chemical-competent Top10 cells (Thermo Fisher Scientific). DNA was purified using the QIAamp DNA Mini, Plasmid Plus Midi or Plasmid Maxi kit (all from Qiagen).

The plasmids pEJS654 All-in-One AAV-sgRNA-h*Nme*Cas9 and *Nme*2Cas9_AAV co-encoding *Nme*Cas9 or *Nme*2Cas9 and a corresponding sgRNA expression cassette were kind gifts from Erik Sontheimer (Addgene no. 112139 and no. 119924). The plasmid pX601-AAV-CMV::NLS-SaCas9-NLS-3xHA-bGHpA;U6::BsaI-sgRNA co-encoding *Sau*Cas9 and a corresponding sgRNA expression cassette were kind gifts from Feng Zhang (Addgene no. 61591). The luciferase reporter plasmid was previously reported by us [162] and modified as follows: an VEGFA target site (NTS33) [270] was inserted behind the firefly luciferase start codon and in frame with the luciferase gene; an NTS33-targeting sgRNA was subsequently inserted into the modified reporter. Vectors encoding AcrIIC1, AcrIIC3 and AcrIIA4 were previously reported by us [162, 244]. AsLOV2-, PDZ- and mCherry-encoding sequences were obtained from IDT as human-codon-optimized synthetic DNA fragments (gBlocks). AcrIIC1 chimeras were created by Golden Gate cloning as follows: the plasmid encoding AcrIIC1 was first linearized at a selected position in the AcrIIC1 coding sequence via around-the-horn PCR; LOV2-, PDZ- and mCherry-coding sequences were then amplified by matching primers introducing optional GS-linker-encoding sequences and ligated into the linearized vector backbone; point mutations and protein tags were introduced via around-the-horn PCR via the primer overhangs. An AcrIIC1-encoding vector bearing a cloning scaffold for the insertion of miRNA-binding sites within the 3 UTR was previously reported by us [244] (Addgene no. 120300). A corresponding construct for AcrIIC1X was generated by Golden Gate cloning. The miR-122-binding sites were introduced into these vectors as annealed oligonucleotides, also by Golden Gate cloning.

Annealed oligonucleotides corresponding to the target site sequence were cloned into the hybrid Cas9–sgRNA vectors via SapI (*Nme*Cas9 and *Nme*2Cas9) or BsaI (*Sau*Cas9) restriction sites as described previously [256, 270].

3.5.8 Cell culture and AAV lysate production

HEK 293T (human embryonic kidney), U87 (human primary glioblastoma; kindly provided by Kathleen Börner, Heidelberg University Clinics) and U2OS (human osteosarcoma; kindly provided by Karsten Rippe, German Cancer Research Center (DKFZ), Heidelberg) were cultured at 5% CO2 and 37 °C in a humidified incubator and maintained in phenol red-free DMEM (Thermo Fisher/GIBCO) supplemented with 10% (v/v) fetal calf serum (Biochrom AG), 2 mM l-glutamine, and 100 U ml^1 penicillin and 100 μg ml1 streptomycin (both Thermo Fisher/GIBCO). The U2OS medium was additionally supplemented with 1 mM sodium pyruvate (GIBCO). Cell lines were free of mycoplasma contamination and authenticated before usage (Multiplexion).

AAV-containing cell lysates were produced by seeding HEK 293T cells into six-well plates (Corning) at a density of 350,000 cells per well. On the next day, cells were cotransfected using 8 μl of Turbofect reagent (Thermo Fisher Scientific) per well and 1,333 ng of each of the following plasmids: (1) an AAV vector plasmid carrying the transgenes to be delivered flanked by inverted terminal repeats; (2) an AAV helper plasmid carrying rep and cap genes of AAV serotype 2; and (3) an adenoviral helper plasmid providing the required helper functions [271]. The AAV vector plasmid encoded (1) a dual transgene cassette expressing *Sau*Cas9 driven from a CMV promoter and sgRNA driven from a shortened U6 promoter, targeting the EMX1, Grin2B or CXCR4 locus, in a single-stranded AAV context; (2) the same *Sau*Cas9 cassette but together with an empty sgRNA expression cassette (negative control); or (3) a CMV promoter-driven AcrIIC1X transgene in a double-stranded AAV context. At 3 d after transfection, cells were collected in 300 μl of PBS and lysed by subjection to five alternating freeze-thaw cycles in liquid nitrogen and in a 37 °C water bath. Cell debris was separated by centrifugation and the supernatant containing the AAVs was stored at 4 °C (for a maximum of 2 weeks) before usage.

3.5.9 Large-scale AAV production, purification and titration

For each AAV construct, HEK 293T cells were seeded in five 14-cm petri dishes at a density of 4 × 106 cells per dish. After 2 d, cells were cotransfected with 14.7 μg per dish of each of the following plasmids: (1) a plasmid encoding the transgene flanked by AAV inverted terminal repeats; (2) a plasmid providing AAV rep and cap (from serotype 2); and (3) a plasmid providing the adenoviral helper functions. Therefore, the plasmid DNA (73.5 μg of each, plasmid 1, 2 and 3) was mixed with 6 ml of H2O, 7.9 ml of 300 mM NaCl (Sigma-Aldrich) and 1.75 ml of polyethylenimine (Polyscience). Mixes were incubated for 10 min and 3.2 ml was added dropwise to each plate. Then, 3 d later, cells were collected and resuspended in 5 ml of Benzonase buffer (50 mM Tris-HCl, 150 mM NaCl, 2 mM MgCl2, pH 8.5). Remaining plasmid DNA was digested by the addition of 1 μl of highly concentrated Benzonase (Merck Millipore) and 1 h incubation at 37 °C. Subsequently, cells were lysed by subjecting them to five freeze and thaw cycles and AAVs were collected with the supernatant after centrifugation at 4,000g and 4 °C for 15 min. The AAVs were purified using an iodixanol gradient as described by Börner *et al.* [272]. Therefore, the supernatant was placed in ultracentrifugation tubes (Seton Scientific) and underlaid with 1.5 ml of 15%, 25%, 40% and 60% iodixanol phases using a Pasteur pipet. The gradients were centrifuged at 50,000 r.p.m. at 4 °C for 2 h (Beckman Coulter) and subsequently the interface between the 40% and 60% iodixanol phase, containing the purified AAVs, was collected using a needle, aliquoted and stored at 80 °C.

To quantify AAV vector yields, quantitative PCR (qPCR) was performed using the RotorGene 6000 (QIAGEN), the SensimixII Probe kit (Bioline), primers (forward: 5-AACGCCAATAGGGACTTTCC; and reverse: 5-GGGCGTACTTGGCATATGAT) and probe (5-FAM-CGGTAAACTGCCCACTTGGCAGT-BHQ1) directed against the CMV promoter. RT–qPCR was performed using the following program: 10 min at 95 °C, followed by 40 cycles of heating at 95 °C for 10 s and elongation at 60 °C for 20 s. After the run, samples and standard curve (based on a dilution of a plasmid with known number of molecules) were analyzed with the accompanying RotorGene 6000 Series Software 1.7.

3.5.10 Luciferase reporter assays

HEK 293T cells were seeded into 96-well plates at a density of 12,500 cells per well. For titration experiments employing the chimeric Acrs (Fig. 3.2b), the cells were cotransfected on the following day with (1) 33 ng of a dual luciferase reporter plasmid encoding a firefly and Renilla luciferase gene and an sgRNA targeting the NTS33 site in the firefly reporter gene; (2) 33 ng of a vector coexpressing *Nme*Cas9 and an sgRNA targeting the NTS33 site; (3) 99, 33, 16.5, 6.6, 3.3 or 1.65 ng of Acr vector; and (4) 0, 66, 82.5, 92.4, 95.7 or 97.35 ng of an irrelevant stuffer plasmid (pBluescript), respectively. The stuffer plasmid was added to keep the total amount of DNA transfected constant in all samples. Transfections were performed using Lipofectamine 3000 reagent (Thermo Fisher Scientific) according to the manufacturer's protocol.

At 2 d post-transfection, the cells were washed with 1x PBS and lysed with 30 µl of passive lysis buffer (Promega) for 30 min, while being shaken on a thermomixer (Eppendorf) at 500 r.p.m. and at room temperature. Finally, luciferase activity was analyzed using the Dual-Glo luciferase assay system (Promega). In short, 10 µl of lysate was transferred to a white sample plate and photo counts were measured with a GLOMAX 96 microplate luminometer (Promega). Integration time was 10 s with a delay of 2 s between substrate injection and measurement. To calculate the reported luciferase activity values, firefly luciferase photon counts were normalized to those obtained for *Renilla* luciferase.

3.5.11 T7 endonuclease assay and TIDE sequencing

Genomic target sites relevant for T7 and TIDE experiments are listed in Supplementary Table 2.16.

For transfection-based experiments, HEK 293T cells were seeded in 96-well plates (Eppendorf) at a density of 12,500 cells per well. For AAV transduction-based experiments, HEK 293T, U2OS and U87 cells were seeded at a density of 3,500, 3,000 and 3,000 cells per well, respectively. For experiments with *Sau*Cas9, transfections were performed with JetPrime using 0.3 µl of JetPrime reagent per well, except for the experiment shown in Sup. Fig. 3.15, in which Lipofectamine

3000 was employed for transfection of all samples including those with *Sau*Cas9. Note that the CXCR4 target site in Sup. Fig. 3.14a is the CXCR4-1 site in Supplementary Table 2.16. For *Nme*Cas9 and *Nme*2Cas9 experiments (Fig. 3.2a and Sup. Figs. 3.3 and 3.5), transfections were conducted with Lipofectamine 3000 using 0.2 µl of Lipofectamine reagent, 0.4 µl of p3000 and 200 ng of total DNA per well. Cells were cotransfected with 100, 133 or 160 ng of Acr vector and 100, 67 or 40 ng of all-in-one Cas9/sgRNA vector, corresponding to Acr/Cas9 vector ratios of 1:1, 2:1 and 4:1, respectively, as indicated in Fig. 3.2a and Sup. Figs. 3.3, 3.5 and 3.15. Transfections for the initial screen of the chimeric AcrIIC1 variants (Sup. Fig. 3.2b) were performed with only 100 ng of total DNA per well, using a 1:1 ratio of Cas9/sgRNA and Acr vectors. For the miRNA overexpression experiment (Sup. Fig. 3.16a,b), 30 ng of SauCas9, 120 ng of AcrIIC1X (or AcrIIC1) and 80 ng of the miRNA-overexpressing plasmid were cotransfected using 0.2 µl of Lipofectamine 3000 and 0.4 µl of p3000 per well.

For AAV-based experiments (Fig. 3.5c), cells were cotransduced with 50 µl of AcrIIC1X and 50 µl of Cas9/sgRNA AAV lysates on 2 subsequent days. As negative and positive controls, cells were transduced with 50 µl of Cas9-only AAV lysate or Cas9/sgRNA AAV lysate, respectively, topped up to 100 µl with PBS (to keep the transduction volume identical in all samples). Note, the CXCR4 target site in Sup. Fig. 3.14b is the CXCR4-2 site in Supplementary Table 2.16.

For the miRNA-122-dependent editing experiments (Sup. Fig. 3.16c,d), HEK 293T cells or Huh-7 cells were cotransduced with titrated AAVs encoding *Sau*Cas9, an EMX1-targeting sgRNA and the respective Acrs on 2 subsequent days. The multiplicity of infection was 105 for *Sau*Cas9 and 5 × 104 for the inhibitors. Cells were lysed 2 d after the second transduction.

At 3 d post-transfection or post-(initial) transduction, cells were harvested in DirectPCR Lysis Reagent (Peqlab) supplemented with Proteinase K (Sigma) and incubated at 55 °C for at least 6 h, followed by Proteinase K inactivation at 85 °C for 45 min. The CRISPR–Cas9-targeted genomic loci were then amplified via PCR with appropriate primers (Supplementary Table 2.17) using Q5 Hot Start High-Fidelity DNA Polymerase (New England Biolabs). Indel frequencies were assessed by T7 endonuclease assay or TIDE sequencing [202].

For T7 assays, 5 µl of the target amplicons was diluted 1:4 in 1x NEB buffer 2 and subsequently denatured at 95 °C for 5 min and re-annealed by applying a ramp rate of 2 °C s1 at 95 to 85 °C and 0.1 °C s1 at 85 to 25 °C using a nexus GSX1 Mastercycler (Eppendorf). Subsequently, 0.5 µl of T7 endonuclease (New England Biolabs) was added, and samples were incubated at 37 °C for 15 min, followed by analysis on a 2% TBE agarose gel. The PCR input and T7 cleavage fragment bands were then quantified using the gel analysis tool in ImageJ [273, 274] (http://imagej.nih.gov/ij/). The frequency of insertions and deletions was calculated using the formula: Indel(%) = 100 × (1–(1–Fraction cleaved) × 0.5), whereas the fraction cleaved is calculated as

$$\text{Fraction cleaved} = \frac{\sum \text{Cleavage product bands}}{\sum \text{Cleavage product bands} + \text{PCR input band}}$$

Full-length gel images are shown in Sup. Fig. 3.18.

For TIDE sequencing analysis, the target locus PCR amplicon was purified from a 1% agarose gel using the QIAquick Gel Extraction Kit (Qiagen). The DNA concentration was determined using a nano-photometer (Nanodrop, Thermo Fisher Scientific) and DNA was diluted to a final concentration of 75 ng µl1 and sent for Sanger sequencing (Eurofins). Percentages of modified sequences were then quantified using the TIDE web tool (https://tide.deskgen.com/).

3.5.12 Western blot

HEK 293T cells were seeded in six-well plates with 4.5 × 105 cells per well. On the following day, 1,000 ng of total DNA was transfected using the Lipofectamine 3000 Transfection kit (Thermo Fisher Scientific). To investigate the effect of AcrIIC1, the chimera 10, AcrIIA4 or a GFP control construct on *Nme*Cas9 expression, 500 ng of the respective construct was cotransfected with 500 ng of a construct coexpressing HA-*Nme*Cas9 and a nontargeting sgRNA. To quantify expression levels of AcrIIC3, AcrIIC1, chimera 10 and AcrIIC1X, an N- and C-terminal flagged version of the respective inhibitor was created and 1,000 ng of corresponding DNA was transfected per well as follows. A DNA mix containing 125 µl of Optimem and 5 µl of P3000 and a Lipofectamine mix comprising 5 µl of Lipofectamine 3000 and 125 µl of Optimem were prepared according to the manufacturer's protocol. Subsequently, the Lipofectamine and DNA mixes were combined and incubated for 10 min and then added dropwise to the cells. Then, 2 d later, cells were collected in 150 µl of RIPA buffer (50 mM Tris (Roth) pH 8.0, 150 mM NaCl, 1 mM EDTA (GRÜSSING GmbH), 1% Triton (Merck), 0.1% SDS (SERVA Electrophoresis GmbH), 0.5% sodium deoxycholate (Merck), Protease Inhibitor (cOmplete, Roche)) per well and incubated on ice for 10 min. Subsequently, samples were centrifuged at 13,000 r.p.m. at 4 °C and the supernatant was collected. Protein concentration in the supernatant was measured using the BCA Protein Assay Kit (Thermo Scientific) and Laemmli Sample Buffer (BioRad) with 1:10 diluted β-mercaptoethanol (Sigma-Aldrich) was added to 400 µg of protein. For the *Nme*Cas9 experiment, 100 µg of protein was loaded on a 4–15%, ten-well precast gel (BioRad), as were the prestained protein ladder (Thermo Scientific) and the MagicMark ladder (Thermo Scientific) that is stained by the secondary antibody. For the inhibitor experiment, a 17.5% acrylamide gel was prepared using 2.65 ml of H2O, 2.2 ml of 40% acrylamide (Roth), 1.3 ml of 1.5 M Tris (pH 8.8), 50 µl of 10% SDS, 50 µl of APS and 6 µl of TEMED. The samples were transferred to a nitrocellulose membrane using a semi-dry transfer system (BioRad) and the membrane was blocked using 5% milk powder (Roth) in tris-buffered saline buffer (TBS) (ChemCruz) with 1% Tween (Roth) for 1 h. For the *Nme*Cas9 experiment, the membrane was cut into two pieces at around 80 kDa. The upper part was incubated with the primary antibody against the HA-tag (Santa Cruz, sc-7392, 1:1,000) and the lower part with the primary antibody against Hsp60 (Santa Cruz, sc-1052, 1:1,000) in 5% milk powder in TBS supplemented with Tween 20 (TBS-T) overnight. On the following day, the upper part of the membrane was treated with a secondary antibody against mouse IgG (Jackson Immuno Research, 115-035-068, 1:5,000) and the lower part of the membrane with a secondary antibody against goat IgG (Santa Cruz, sc-2768, 1:5,000) for 1 h. For the inhibitor experiment, the membrane was first incubated with a primary antibody against the Flag-tag (Sigma, F1804, 1:1,000) overnight and subsequently with the secondary antibody against mouse IgG for 1 h before it was imaged. In a second round, the membrane was first washed with TBS-T for 6 h and incubated with the primary antibody against Hsp60 overnight, followed by incubation with the secondary antibody against goat IgG. The image was obtained by using a two-component chemiluminescence substrate for the secondary antibody-coupled horseradish peroxidase (Biozym).

3.5.13 *In vitro* DNA cleavage assay

The target DNA for the *in vitro* cleavage assay was cloned into pUC19 and linearized by the ScaI restriction enzyme before use. A 10-µl reaction of 100 nM RNP-Acr protein mix and 300 ng of target DNA was set up in reaction buffer (20 mM Tris pH 7.5, 100 mM KCl, 5 mM MgCl2, 1 mM DTT, 5% glycerol) as follows (samples were kept on ice if not indicated otherwise). *Nme*Cas9 and the sgRNA were either pre-mixed in reaction buffer, then the mix incubated on ice for 15 min (to allow RNP formation) and then the Acr added followed by incubation for another 15 min; or the Acr and *Nme*Cas9 were mixed first, then incubated on ice for 15 min followed by addition of the sgRNA and incubation for another 15 min (as indicated in Fig. 3.3b). An *Nme*Cas9/sgRNA ratio of 1:1.1 was used. Ratios of RNP/Acr were as indicated in Fig. 3.3b. Final reaction mixes were incubated at 37 °C for 10 min. Subsequently, the reaction was stopped by adding 2 µl of 6× loading dye and target DNA cleavage was analyzed on a 1% agarose gel. Sup. Fig. 3.19 shows the Coomassie staining of the purified proteins used for the *in vitro* DNA cleavage assay.

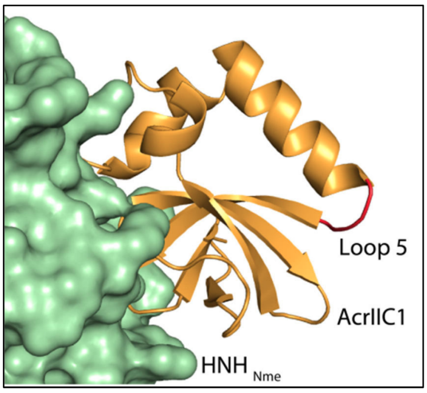
3.5.14 Electrophoretic mobility shift assay

Binding substrates were prepared by mixing equal volumes of 4 µM 5-Cy3-labeled 53-nucleotide targeting strand (TS) with 4 µM 5-Cy5-labeled 53-nucleotide nontargeting strand (NTS), followed by incubation at 95 °C for 10 min and strand annealing at room temperature. Forked double-strand DNA with the protospacer region not complementary was used to facilitate DNA binding to *Nme*Cas9. Then, 6 µl of reaction buffer (20 mM Tris pH 7.5, 150 mM NaCl) was incubated on ice with 1 µl of 2 µM *Nme*Cas9 and 1 µl of 1, 2, 8 or 20 µM Acr for 15 min. Next, 1 µl of 2 µM sgRNA was added and incubated on ice for 30 min. Finally, 1 µl of 2 µM dsDNA-53-32M was added, followed by incubation for another 30 min on ice. Cas9 DNA binding was analyzed by running a 5% native gel followed by detection of the NTS signal with a FluorChem system (the Cy3 signal was not used for analysis, as Cy3 exhibits a strong cross-talk with the mCherry part of AcrIIC1-mCherry).

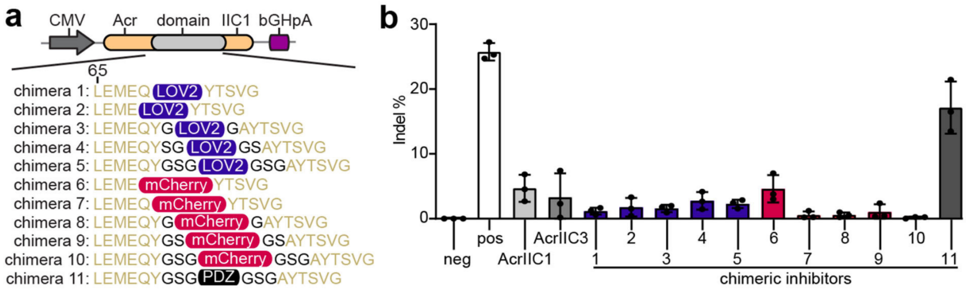
3.5.15 Statistical analysis

Individual data points correspond to independent experiments with cells that were seeded and transfected/transduced independently and on different days. Each data point shown for the luciferase experiments further represents the mean of three technical replicates, that is, cell cultures in different wells that were transfected and treated in parallel. Reported differences between groups were analyzed for statistical significance by one-way analysis of variance (ANOVA) and Bonferroni's corrected post-hoc test. $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$; $^{****}P < 0.0001$. $P < 0.05$ was considered statistically significant.

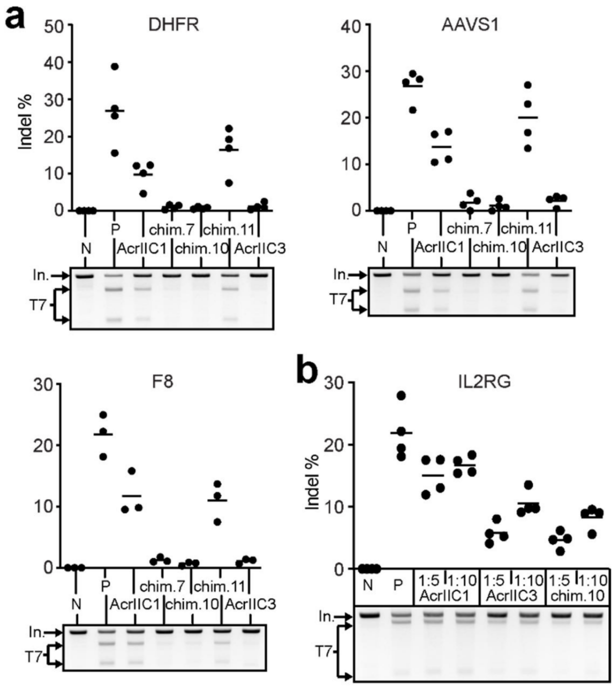
3.6 Supplementary information



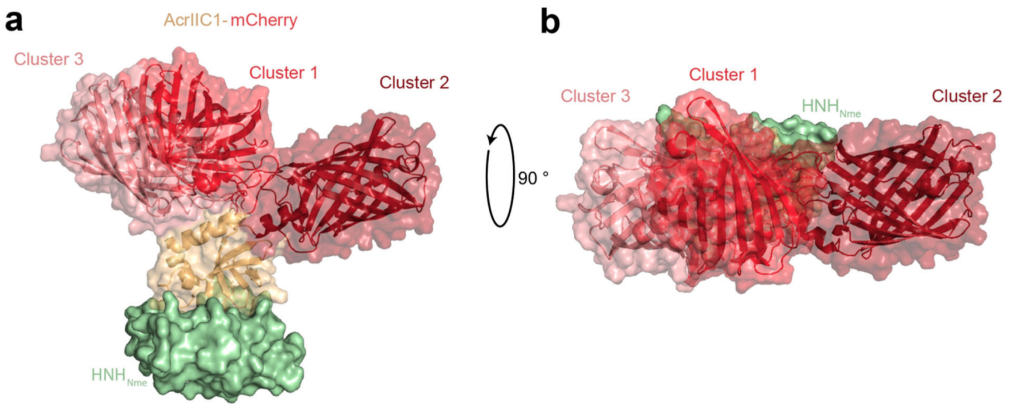
Supplementary Figure 3.1: Identification of a loop amenable to domain insertion. Loop 5 (in red) is located on the opposite site of the HNH-domain binding interface. It connects an α-helix and a β-sheet via two residues (Tyrosine 70, Alanine 71). Structures shown correspond to PDB 5VGB.



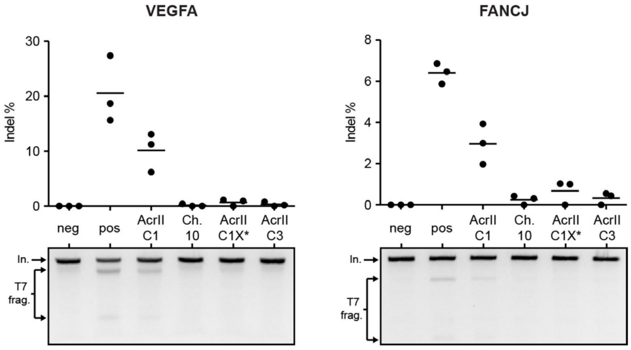
Supplementary Figure 3.2: Screening of AcrIIC1 domain insertion variants. **a:** Schematic of chimeric Acrs. The chimeras comprise AcrIIC1 bearing an *Avena sativa* LOV2, mCherry or PDZ domain inserted into loop 5. The constructs carry optional GS linkers flanking the inserted domain (black residues) or deletions in loop 5. AcrIIC1 residue L65 is indicated. **b:** Screen of chimeric Acrs in HEK 293T cells. Cells were co-transfected with vectors expressing *Nme*Cas9, the indicated Acr and a sgRNA targeting the IL2RG locus followed by T7 endonuclease assay. The Acr:Cas9 vector ratio during transfection was 1:1. Bars indicate means, error bars the SD and dots individual data points for n = 3 independent experiments. Numbers correspond to the constructs shown in a. Neg, negative control (Cas9 only). Pos, positive control (Cas9 + sgRNA).



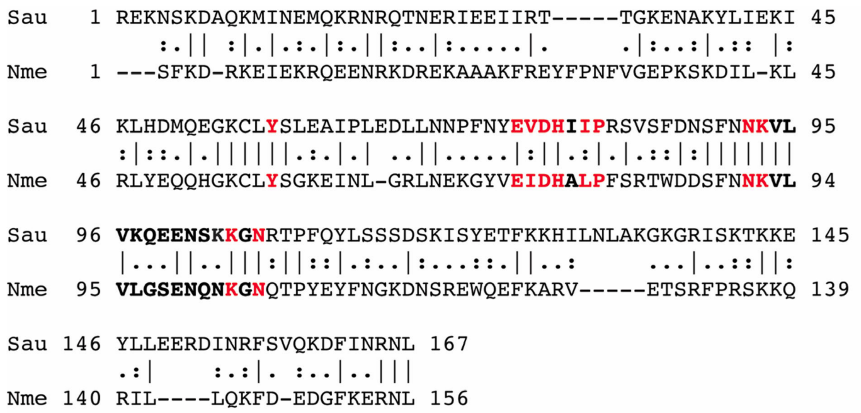
Supplementary Figure 3.3: T7 endonuclease assay analysis of *N. meningitides* Cas9 inhibition by AcrIIC1 chimeras. **a,b:** HEK 293T cells were co-transfected with vectors expressing *Nme*Cas9, the indicated Acr and sgRNAs targeting different genomic loci followed by T7 endonuclease assay. In **a**, Acr:*Nme*Cas9 vector ratio used during transfection was 1:1, while in **b**, the indicated, low Acr:*Nme*Cas9 vector ratios were used. Representative T7 gel images and corresponding quantification of indel frequencies are shown. Lines in plots show means, dots are individual data points for n = 4 (DHFR, AAVS1 and IL2RG locus) or n = 3 (F8 locus) independent experiments. Chim., AcrIIC1-mCherry chimeras in Sup. Fig. 3.2a. In., input band. T7, T7 cleavage fragments. N, negative Cas 9 only control. P, positive control Cas9 + sgRNA. Full-length gel images are shown in Sup. Fig. 3.18.



Supplementary Figure 3.4: Acr-mCherry chimera conformations. For the shown chimera, the mCherry was inserted after residue Q69 and before Y72. Residues Y70 and A71 were deleted. Representative examples of the three most populated clusters are shown, aligned to the structure of AcrIIC1 in complex with the *Nme*Cas9 HNH domain. **a** Side view. **b** Top view. The models are based on PDB 5VGB and 4ZIN.

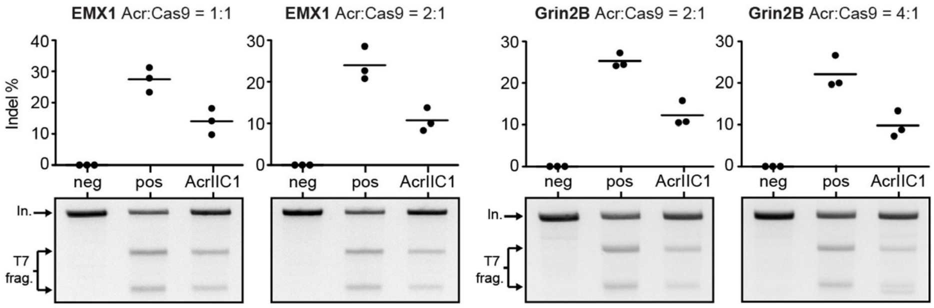


Supplementary Figure 3.5: AcrIIC1 chimeras show improved inhibition of *Nme*2Cas9. HEK 293T cells were co-transfected with vectors encoding *Nme*Cas9, the indicated Acr and sgRNAs targeting the VEGFA or FANCJ loci followed by T7 endonuclease assay. An Acr:*Nme*2Cas9 vector mass ratio of 1:1 was used during transfection. Representative T7 gel images and corresponding quantification of indel frequencies are shown. Lines in plots show means, dots represent individual data points for n = 3 independent experiments. Ch., AcrIIC1-mCherry chimeras. In., input band. T7 frag., T7 cleavage fragments. Neg, negative control. Pos, positive control. Full-length gel images are shown in Sup. Fig. 3.18.

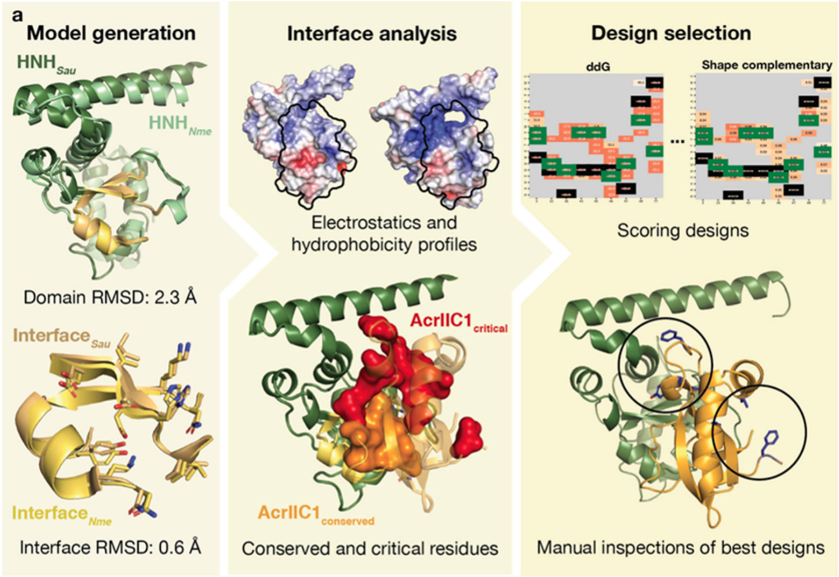


```
Sau   1 REKNSKDAQKMINEMQKRNRQTNERIEEIIRT-----TGKENAKYLIEKI 45
           :.|| :|.|.:.|:.||:..|:.....|.     .|:..:|.:: |:
Nme   1 ---SFKD-RKEIEKRQEENRKDREKAAAKFREYFPNFVGEPKSKDIL-KL 45

Sau  46 KLHDMQEGKCLYSLEAIPLEDLLNNPFNYEVDHIIPRSVSFDNSFNNKVL 95
        :|::.|.||||||.:.|.| ..||.....|:||.:|.|.::|:|||||||
Nme  46 RLYEQQHGKCLYSGKEINL-GRLNEKGYVEIDHALPFSRTWDDSFNNKVL 94

Sau  96 VKQEENSKKGNRTPFQYLSSSDSKISYETFKKHILNLAKGKGRISKTKKE 145
        |...||..|||:||::|.:..|:...::.||..:     ...|..::||:
Nme  95 VLGSENQNKGNQTPYEYFNGKDNSREWQEFKARV-----ETSRFPRSKKQ 139

Sau 146 YLLEERDINRFSVQKDFINRNL 167
        .:|    :.:|. :..|..|||
Nme 140 RIL----LQKFD-EDGFKERNL 156
```

Supplementary Figure 3.6: Alignment of the *Sau*- and *Nme*Cas9 HNH domains. Conserved residues within the AcrIIC1-binding interface are in red. Bold characters indicate residues that were used to align the interfaces.



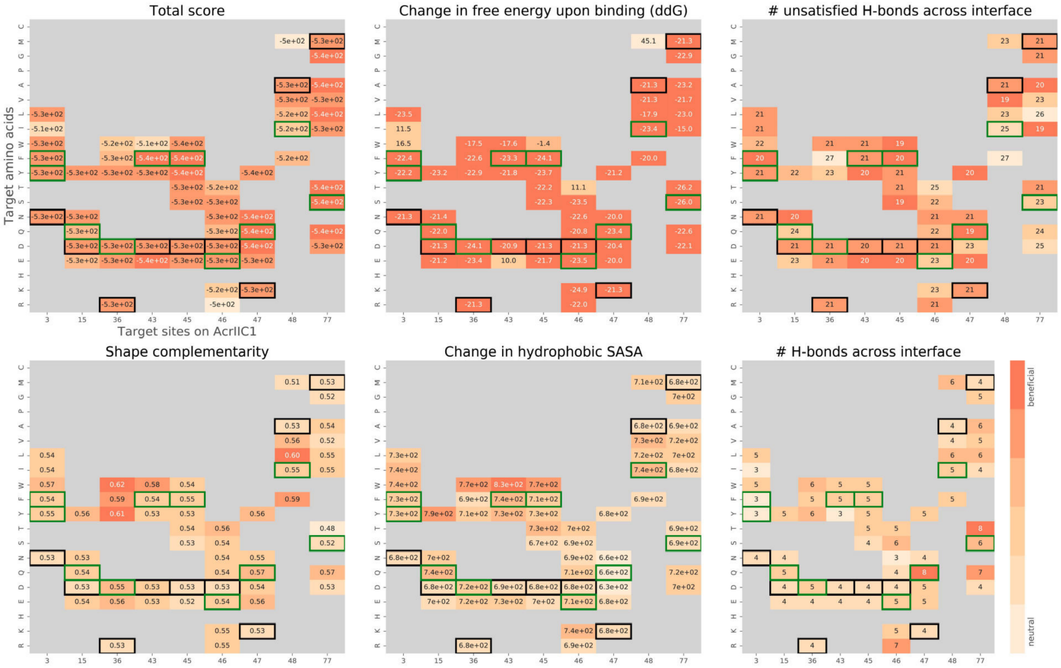
Supplementary Figure 3.7: AcrIIC1 is able to partially inhibit genome editing by *S. aureus* Cas9. HEK 293T cells were co-transfected with vectors expressing *Sau*Cas9, a sgRNA targeting the EMX1 or Grin2B locus and AcrIIC1 followed by T7 endonuclease assay. The Acr:Cas9 vector ratio used during transfection is indicated. Lines in the plots indicate means, dots individual data points for n = 3 independent experiments. In., input band. T7 frag., T7 cleavage fragments. Neg, negative control (Cas9 only). Pos, positive control (Cas9 + sgRNA). Full-length gel images are shown in Sup. Fig. 3.18.



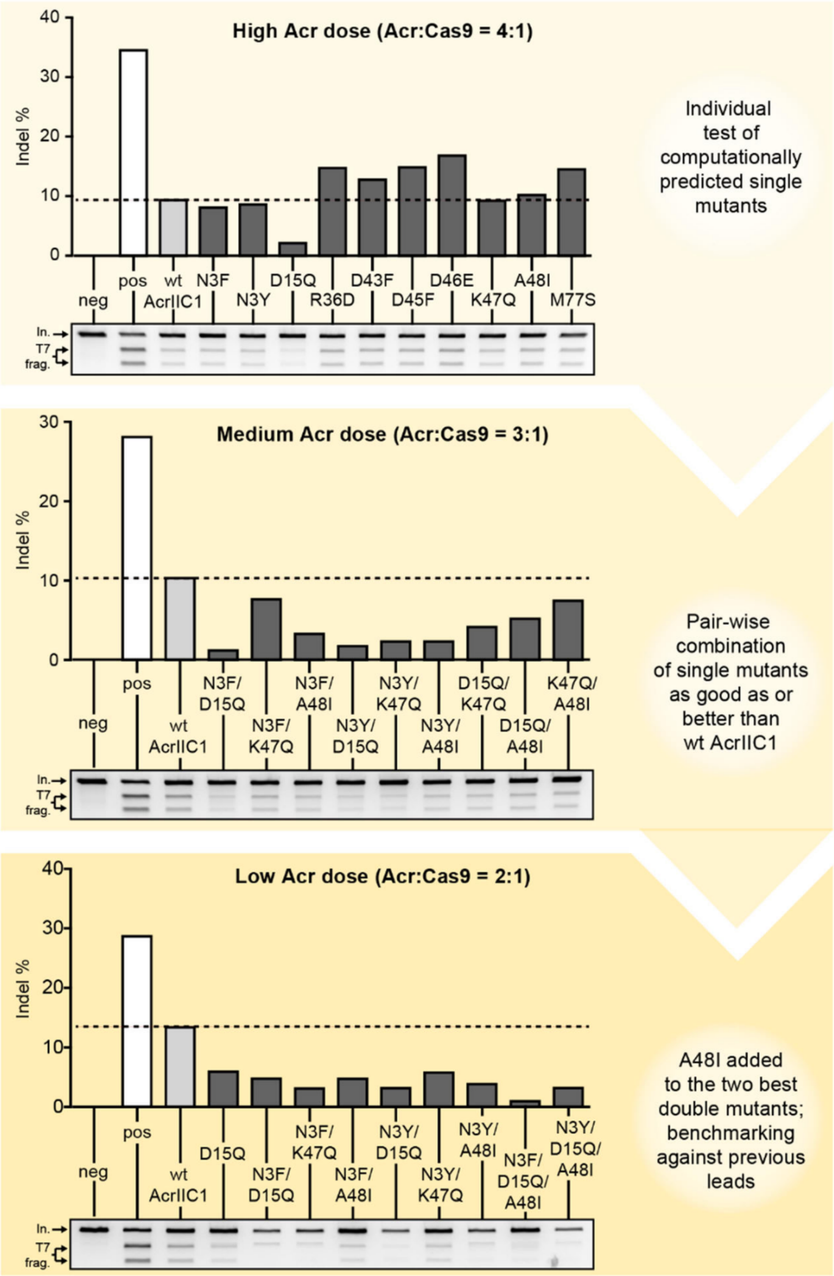
b

```
AcrIIC1    1 MANKTYKIGKNAGYDGCGLCLAAISENEAIKVKYLRDIC
AcrIIC1X   1 MAFKTYKIGKNAGYQGCGLCLAAISENEAIKVKYLRDIC
              3           15
AcrIIC1   40 PDYDGDDKAEDWLRWGTDSRVKAAALEMEQYAYTSVGMA
AcrIIC1X  40 PDYDGDDKIEDWLRWGTDSRVKAAALEMEQYAYTSVGMA
                    48
AcrIIC1   79 SCWEFVEL
AcrIIC1X  79 SCWEFVEL
```

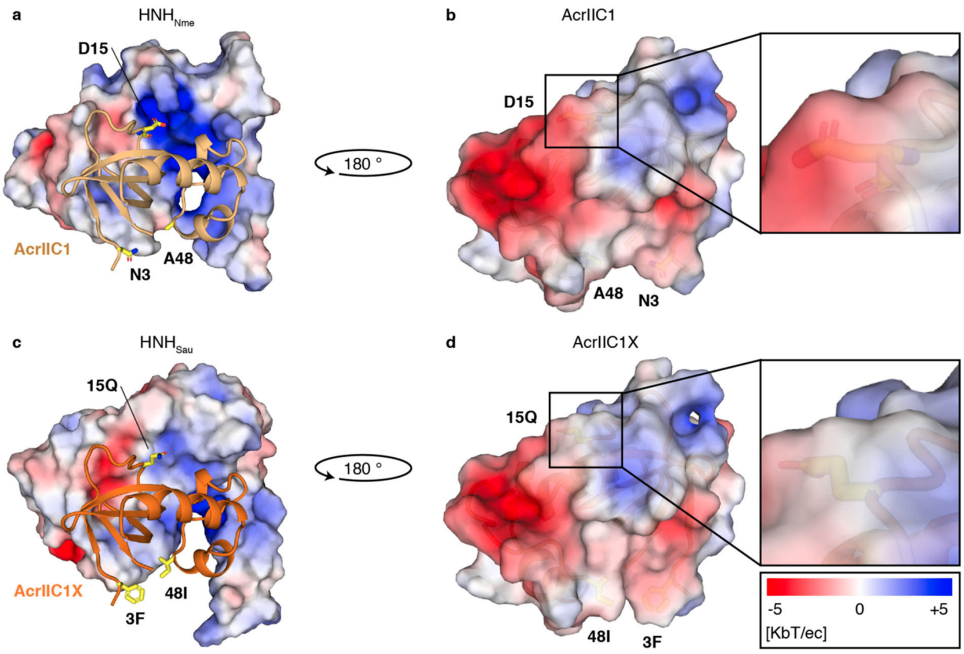
Supplementary Figure 3.8: Computational interface design. **a:** After generating the *Sau*Cas9 HNH – AcrIIC1 model by structural alignment of the *Nme*Cas9 and *Sau*Cas9 HNH domains and analysis of the interfaces, conserved and critical residues to be kept or mutated were determined. Single-site *in silico* mutation experiments were performed and top-scoring variants were experimentally validated after manual inspection. **b:** Sequence alignment of AcrIIC1 and AcrIIC1X. Mutations are marked in bold red characters.



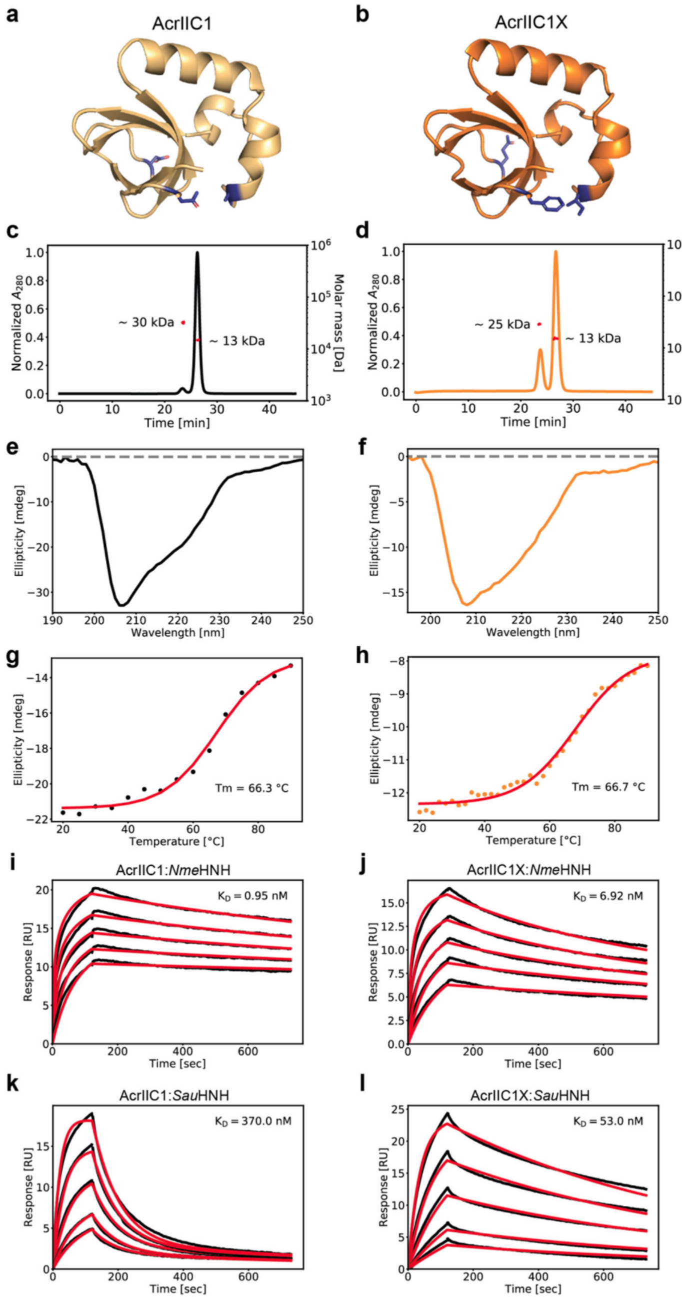
Supplementary Figure 3.9: Rosetta scores of single site mutants. Heatmaps of the computed structural metrics for the generated designs. Black boxes refer to the wild-type (AcrIIC1) scores per position, while green boxes indicate scores of selected designs for experimental validation.



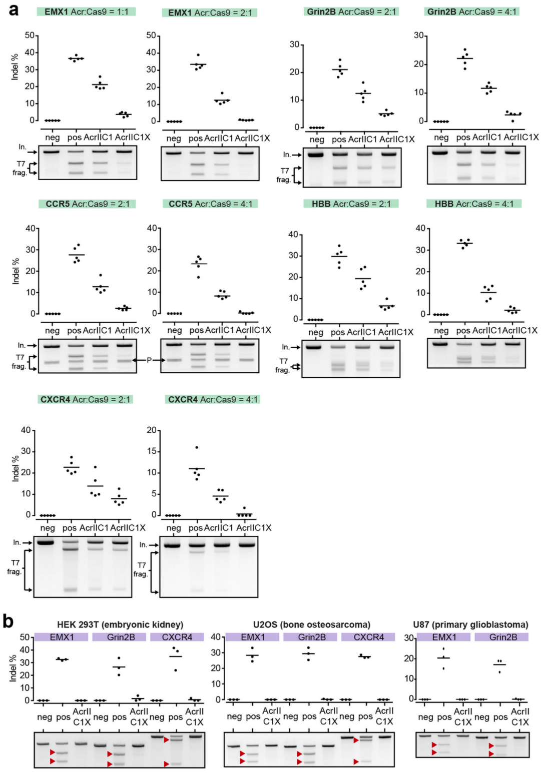
Supplementary Figure 3.10: Screening and iterative optimization of computationally designed AcrIIC1 mutants. HEK 293T cells were co-transfected with vectors expressing *Sau*Cas9, a sgRNA targeting the EMX1 locus and either wild-type (wt) AcrIIC1 or the indicated AcrIIC1 mutant followed by T7 endonuclease assay. The Acr:Cas9 vector ratio (indicated) used during transfection was decreased with every iteration. Representative T7 gel images and corresponding quantifications of indel frequencies are shown. Dotted lines indicate the editing frequency in the presence of wild-type AcrIIC1. In., input band. T7 frag., T7 cleavage fragments. Neg, negative control (Cas9 only). Pos, positive control (Cas9 + sgRNA). Full-length gel images are shown in Sup. Fig. 3.18.



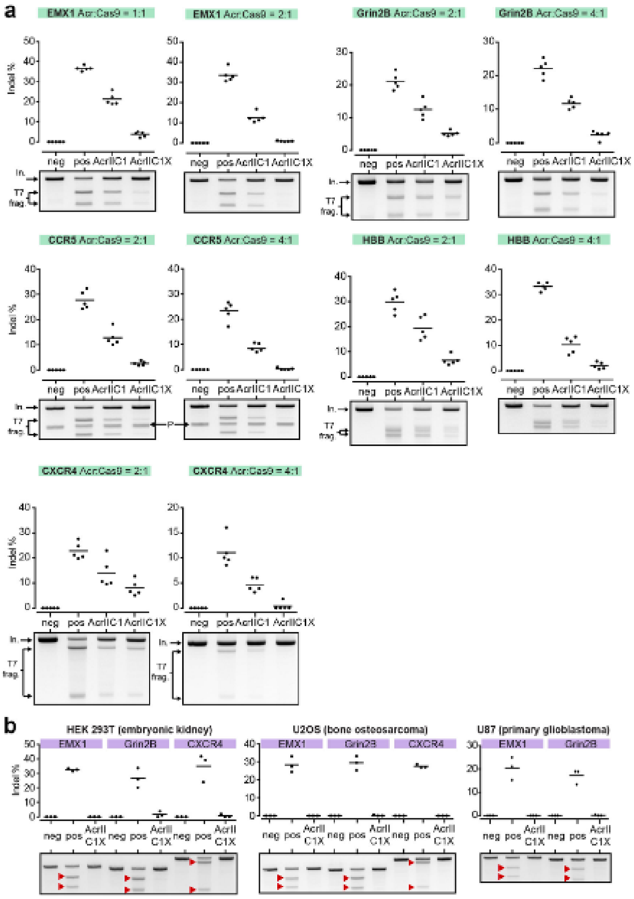
Supplementary Figure 3.11: Electrostatic potential of HNH domain and AcrIIC1 surfaces. Potentials were calculated using the Adaptive Poisson-Boltzmann Solver. **a:** The *Nme*Cas9 HNH domain has a strong positive potential close to residue 15 of AcrIIC1 in bound state. **b:** AcrIIC1 has a similarly strong negative potential in the Cas-binding site. Residue 15 is at the border of this negative patch. **c:** The *Sau*Cas9 HNH domain shows reduced positive potential around AcrIIC1 residue 15 as compared to *Nme*Cas9. **d:** Mutation of residue 15 from aspartic acid to glutamine reduces the negative potential of AcrIIC1 in this position (PDB 5VGB). **c,d:** AcrIIC1X is the AcrIIC1 N3F, D15Q, A48I triple mutant.



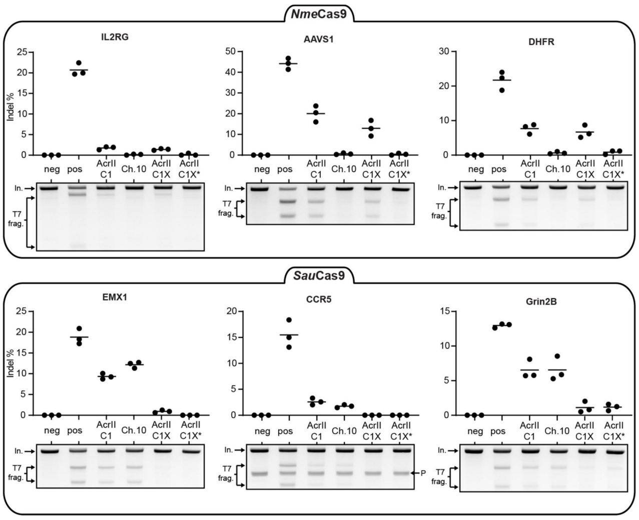
Supplementary Figure 3.12: Biochemical analysis of AcrIIC1 and AcrIIC1X. **a,b:** Structure of AcrIIC1 and model of AcrIIC1X. Shown in blue are the three points mutations differentiating AcrIIC1 from AcrIIC1X. **c,d:** SEC-MALS for AcrIIC1 and AcrIIC1X showing mainly monomeric forms and a small fraction of dimer species present in case of AcrIIC1X. **e,f:** CD spectra for AcrIIC1 and AcrIIC1X. Spectra for AcrIIC1 and AcrIIC1X show a minimum at around 207 nm, typical of mixed-αand βsecondary structures. **g,h:** Thermal melting CD spectra for AcrIIC1 and AcrIIC1X are shown. Both proteins have a melting point (Tm) of around 66 °C. **i,j:** Binding affinity determined by SPR for AcrIIC1 and AcrIIC1X to *Nme*Cas9 HNH domain. AcrIIC1 shows a K_D of 0.95 nM while AcrIIC1X shows a higher K_D of 6.92 nM. Experimental sensorgrams are shown in black and the fitted curves in red. **k,l:** Binding affinity determined by SPR for AcrIIC1 and AcrIIC1X to *Sau*Cas9 HNH domain. AcrIIC1 shows a K_D of 370 nM; AcrIIC1X shows a lower K_D of 53 nM. Experimental sensorgrams are shown in black and the fitted curves in red.



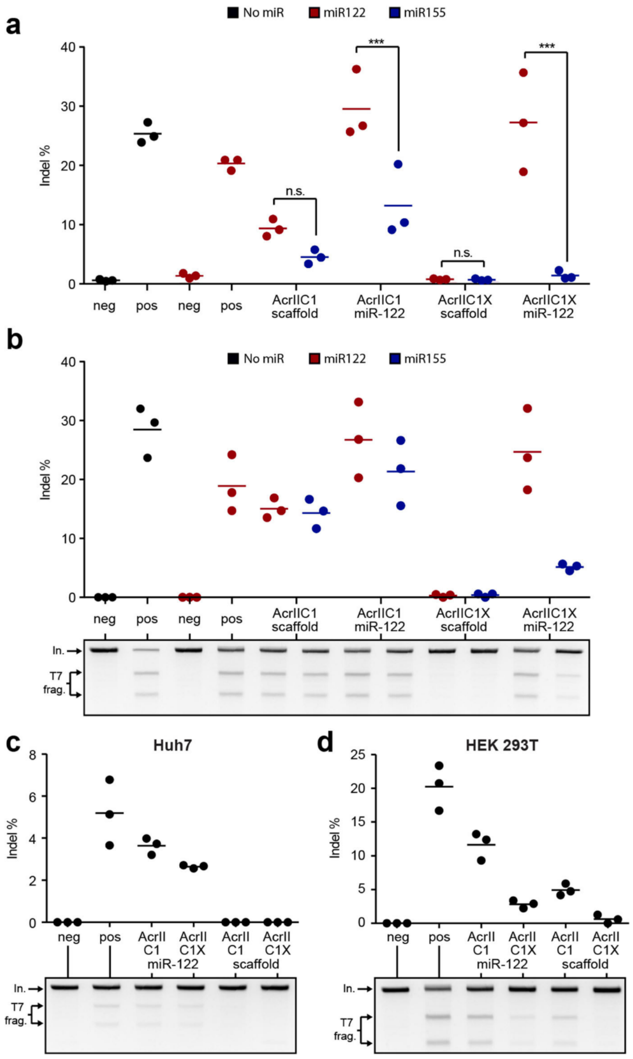
Supplementary Figure 3.13: T7 endonuclease assays of *S. aureus* Cas9 inhibition by AcrIIC1X. **a:** HEK 293T cells were co-transfected with vectors expressing (i) the indicated Acr and (ii) *Sau*Cas9 and a sgRNA targeting the indicated locus followed by T7 endonuclease assay. The Acr:Cas9 vector ratio used during transfection is indicated. P denotes a T7 cleavage band which is due to a polymorphism in the CCR5 gene (Sup. Fig. 3.17). In., input band. T7 frag., T7 cleavage fragments. **b:** AAV-mediated delivery of AcrIIC1X results in potent *Sau*Cas9 inhibition in different cell lines. Cells were co-transduced with AAV2 vectors expressing (i) *Sau*Cas9 and a sgRNA targeting the indicated loci and (ii) AcrIIC1X followed by T7 endonuclease assay. Red triangles point to T7 cleavage fragments. **a,b:** Representative T7 gel images and corresponding quantification of indel frequencies are shown. Lines in the plots indicate means, dots individual data points for n = 5 (**a**) or n = 3 (**b**) independent experiments. Neg, negative control (Cas9 only). Pos, positive control (Cas9 + sgRNA). Full-length gel images are shown in Sup. Fig. 3.18.



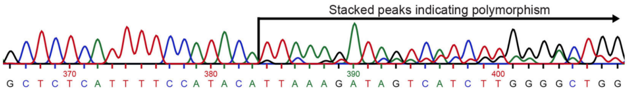
Supplementary Figure 3.14: T7 endonuclease assays of *S. aureus* Cas9 inhibition by AcrIIC1X. **a:** HEK 293T cells were co-transfected with vectors expressing (i) the indicated Acr and (ii) *Sau*Cas9 and a sgRNA targeting the indicated locus followed by T7 endonuclease assay. The Acr:Cas9 vector ratio used during transfection is indicated. P denotes a T7 cleavage band which is due to a polymorphism in the CCR5 gene (Sup. Fig. 3.17). In., input band. T7 frag., T7 cleavage fragments. **b:** AAV-mediated delivery of AcrIIC1X results in potent *Sau*Cas9 inhibition in different cell lines. Cells were co-transduced with AAV2 vectors expressing (i) *Sau*Cas9 and a sgRNA targeting the indicated loci and (ii) AcrIIC1X followed by T7 endonuclease assay. Red triangles point to T7 cleavage fragments. **a,b:** Representative T7 gel images and corresponding quantification of indel frequencies are shown. Lines in the plots indicate means, dots individual data points for n = 5 (**a**) or n = 3 (**b**) independent experiments. Neg, negative control (Cas9 only). Pos, positive control (Cas9 + sgRNA). Full-length gel images are shown in Sup. Fig. 3.18.



Supplementary Figure 3.15: AcrIIC1X inhibits *Nme*Cas9 function, which can be further improved by domain insertion. HEK 293T cells were co-transfected with vectors expressing (i) *Nme*Cas9 or *Sau*Cas9, (ii) a sgRNA targeting the indicated locus as well as the indicated Acr variant followed by T7 endonuclease assay. The Acr:Cas9 vector ratio used during transfection was 1:1 for the *Nme*Cas9 and 2:1 for the *Sau*Cas9 samples. Representative T7 gel images and corresponding quantification of indel frequencies are shown. Lines in the plots indicate means, dots individual data points for n = 3 independent experiments. In., input band. T7 frag., T7 cleavage fragments. Neg, negative control (Cas9 only). Pos, positive control (Cas9 + sgRNA). P denotes a T7 cleavage band which is due to a polymorphism in the CCR5 gene (Sup. Fig. 3.17). Ch. 10, Acr chimera no. 10 in Sup. Fig. 3.2. AcrIIC1X*, AcrIIC1X-mCherry chimera. Full-length gel images are shown in Sup. Fig. 3.1.

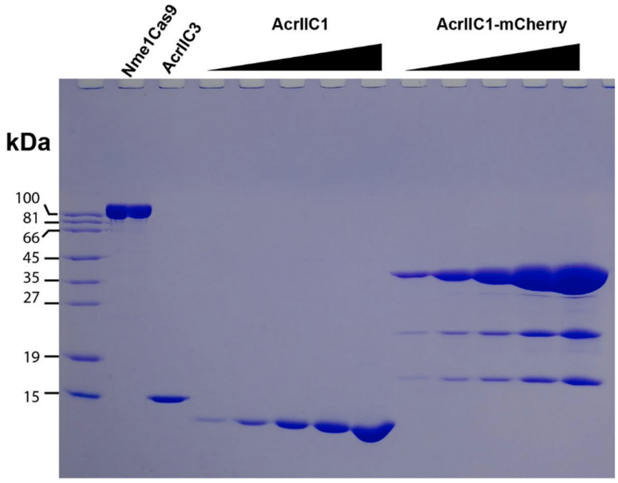


Supplementary Figure 3.16: AcrIIC1X can be harnessed for cell type-specific, miRNA-dependent inhibition of *Sau*Cas9. **a,b:** HEK 293T cells were co-transfected with vectors encoding *Sau*Cas9, an EMX1-targeting sgRNA, the indicated Acr variant alongside constructs overexpressing miR-122 (red) or miR-155 (blue), followed by TIDE sequencing (**a**) or T7 endonuclease assay (**b**). The Acrs were carrying either two miR-122 binding site in their 3'-UTR (miR-122) or contained a 3' UTR of identical length, but lacking the miRNA binding sites (scaffold). The *Sau*Cas9:Acr:miRNA vector mass ratio used during transfection was 1:4:2.7. Lines in the plots indicate means, dots individual data points for n = 3 independent experiments. Neg, negative control (Cas9 only). Pos, positive control (Cas9 + sgRNA). n.s., not significant; ***P < 0.001 by one-way ANOVA with Bonferroni correction. In **b**, representative T7 gel images are shown below the quantification of indel frequencies. In., input band. T7 frag., T7 cleavage fragments. **c,d:** Huh7 or HEK 293Tcells were co-transduced with AAV vectors encoding *Sau*Cas9, an EMX1 targeting sgRNA and the indicated Acrs carrying miRNA-122 binding sites in their 3'-UTR or not. MOIs of 105 for *Sau*Cas9 and 5x104 for the Acrs were used during transduction. Editing outcomes were analyzed by T7 endonuclease assay. Representative T7 gel images and corresponding quantification of indel frequencies are shown. Lines in the plots indicate means, dots individual data points for n = 3 independent experiments. In., input band. T7 frag., T7 cleavage fragments. Neg, negative control (Cas9 only). Pos, positive control (Cas9 + sgRNA). **b,d:** Full-length gel images are shown in Sup. Fig. 3.18.

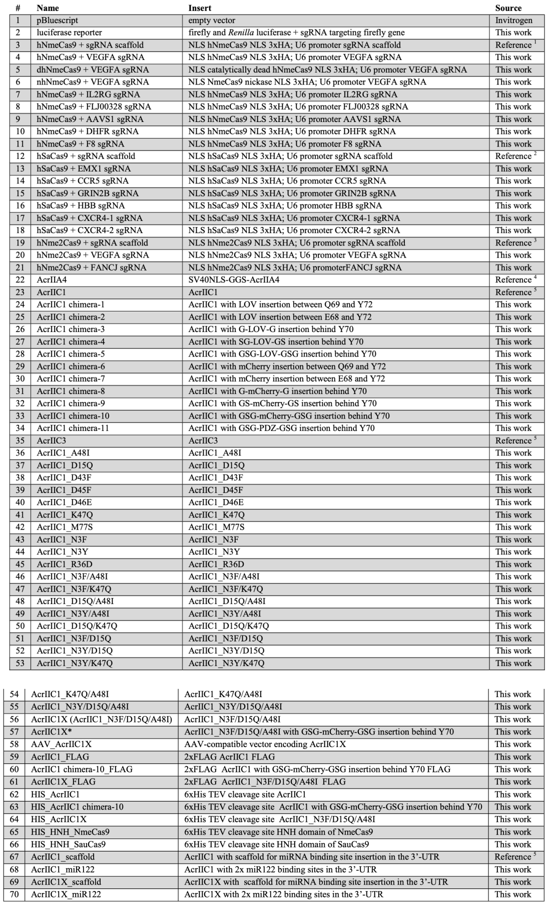


Supplementary Figure 3.17: Polymorphism in the CCR5 gene. Sanger sequencing confirms a polymorphism within the primer-flanked region in CCR5 (genomic DNA is derived from untreated HEK 293T cells).

Supplementary Figure 3.18: Gel images and western blots. The ladder is the Gene Ruler DNA Ladder Mix (Thermo Fisher).

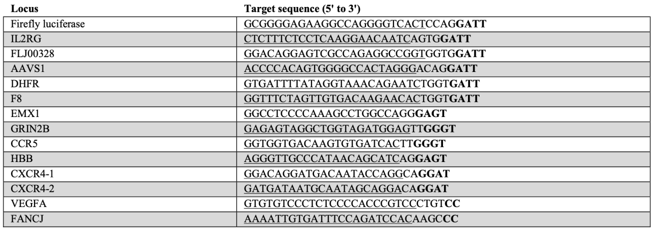


Supplementary Figure 3.19: Coomassie staining of PAGE loaded with purified proteins. Data corresponds to a single experiment.



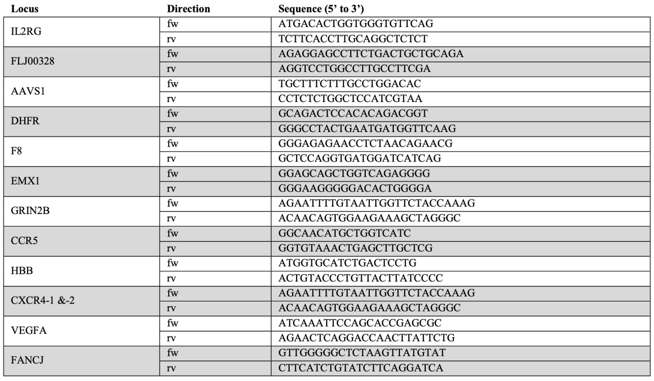
#	Name	Insert	Source
1	pBluescript	empty vector	Invitrogen
2	luciferase reporter	firefly and *Renilla* luciferase + sgRNA targeting firefly gene	This work
3	hNmeCas9 + sgRNA scaffold	NLS hNmeCas9 NLS 3xHA; U6 promoter sgRNA scaffold	Reference [1]
4	hNmeCas9 + VEGFA sgRNA	NLS hNmeCas9 NLS 3xHA; U6 promoter VEGFA sgRNA	This work
5	dhNmeCas9 + VEGFA sgRNA	NLS catalytically dead hNmeCas9 NLS 3xHA; U6 promoter VEGFA sgRNA	This work
6	nhNmeCas9 + VEGFA sgRNA	NLS NmeCas9 nickase NLS 3xHA; U6 promoter VEGFA sgRNA	This work
7	hNmeCas9 + IL2RG sgRNA	NLS hNmeCas9 NLS 3xHA; U6 promoter IL2RG sgRNA	This work
8	hNmeCas9 + FLJ00328 sgRNA	NLS hNmeCas9 NLS 3xHA; U6 promoter FLJ00328 sgRNA	This work
9	hNmeCas9 + AAVS1 sgRNA	NLS hNmeCas9 NLS 3xHA; U6 promoter AAVS1 sgRNA	This work
10	hNmeCas9 + DHFR sgRNA	NLS hNmeCas9 NLS 3xHA; U6 promoter DHFR sgRNA	This work
11	hNmeCas9 + F8 sgRNA	NLS hNmeCas9 NLS 3xHA; U6 promoter F8 sgRNA	This work
12	hSaCas9 + sgRNA scaffold	NLS hSaCas9 NLS 3xHA; U6 promoter sgRNA scaffold	Reference [2]
13	hSaCas9 + EMX1 sgRNA	NLS hSaCas9 NLS 3xHA; U6 promoter EMX1 sgRNA	This work
14	hSaCas9 + CCR5 sgRNA	NLS hSaCas9 NLS 3xHA; U6 promoter CCR5 sgRNA	This work
15	hSaCas9 + GRIN2B sgRNA	NLS hSaCas9 NLS 3xHA; U6 promoter GRIN2B sgRNA	This work
16	hSaCas9 + HBB sgRNA	NLS hSaCas9 NLS 3xHA; U6 promoter HBB sgRNA	This work
17	hSaCas9 + CXCR4-1 sgRNA	NLS hSaCas9 NLS 3xHA; U6 promoter CXCR4-1 sgRNA	This work
18	hSaCas9 + CXCR4-2 sgRNA	NLS hSaCas9 NLS 3xHA; U6 promoter CXCR4-2 sgRNA	This work
19	hNme2Cas9 + sgRNA scaffold	NLS hNme2Cas9 NLS 3xHA; U6 promoter sgRNA scaffold	Reference [3]
20	hNme2Cas9 + VEGFA sgRNA	NLS hNme2Cas9 NLS 3xHA; U6 promoter VEGFA sgRNA	This work
21	hNme2Cas9 + FANCJ sgRNA	NLS hNme2Cas9 NLS 3xHA; U6 promoterFANCJ sgRNA	This work
22	AcrIIA4	SV40NLS-GGS-AcrIIA4	Reference [4]
23	AcrIIC1	AcrIIC1	Reference [5]
24	AcrIIC1 chimera-1	AcrIIC1 with LOV insertion between Q69 and Y72	This work
25	AcrIIC1 chimera-2	AcrIIC1 with LOV insertion between E68 and Y72	This work
26	AcrIIC1 chimera-3	AcrIIC1 with G-LOV-G insertion behind Y70	This work
27	AcrIIC1 chimera-4	AcrIIC1 with SG-LOV-GS insertion behind Y70	This work
28	AcrIIC1 chimera-5	AcrIIC1 with GSG-LOV-GSG insertion behind Y70	This work
29	AcrIIC1 chimera-6	AcrIIC1 with mCherry insertion between Q69 and Y72	This work
30	AcrIIC1 chimera-7	AcrIIC1 with mCherry insertion between E68 and Y72	This work
31	AcrIIC1 chimera-8	AcrIIC1 with G-mCherry-G insertion behind Y70	This work
32	AcrIIC1 chimera-9	AcrIIC1 with GS-mCherry-GS insertion behind Y70	This work
33	AcrIIC1 chimera-10	AcrIIC1 with GSG-mCherry-GSG insertion behind Y70	This work
34	AcrIIC1 chimera-11	AcrIIC1 with GSG-PDZ-GSG insertion behind Y70	This work
35	AcrIIC3	AcrIIC3	Reference [5]
36	AcrIIC1_A48I	AcrIIC1_A48I	This work
37	AcrIIC1_D15Q	AcrIIC1_D15Q	This work
38	AcrIIC1_D43F	AcrIIC1_D43F	This work
39	AcrIIC1_D45F	AcrIIC1_D45F	This work
40	AcrIIC1_D46E	AcrIIC1_D46E	This work
41	AcrIIC1_K47Q	AcrIIC1_K47Q	This work
42	AcrIIC1_M77S	AcrIIC1_M77S	This work
43	AcrIIC1_N3F	AcrIIC1_N3F	This work
44	AcrIIC1_N3Y	AcrIIC1_N3Y	This work
45	AcrIIC1_R36D	AcrIIC1_R36D	This work
46	AcrIIC1_N3F/A48I	AcrIIC1_N3F/A48I	This work
47	AcrIIC1_N3F/K47Q	AcrIIC1_N3F/K47Q	This work
48	AcrIIC1_D15Q/A48I	AcrIIC1_D15Q/A48I	This work
49	AcrIIC1_N3Y/A48I	AcrIIC1_N3Y/A48I	This work
50	AcrIIC1_D15Q/K47Q	AcrIIC1_D15Q/K47Q	This work
51	AcrIIC1_N3F/D15Q	AcrIIC1_N3F/D15Q	This work
52	AcrIIC1_N3Y/D15Q	AcrIIC1_N3Y/D15Q	This work
53	AcrIIC1_N3Y/K47Q	AcrIIC1_N3Y/K47Q	This work
54	AcrIIC1_K47Q/A48I	AcrIIC1_K47Q/A48I	This work
55	AcrIIC1_N3Y/D15Q/A48I	AcrIIC1_N3Y/D15Q/A48I	This work
56	AcrIIC1X (AcrIIC1_N3F/D15Q/A48I)	AcrIIC1_N3F/D15Q/A48I	This work
57	AcrIIC1X*	AcrIIC1_N3F/D15Q/A48I with GSG-mCherry-GSG insertion behind Y70	This work
58	AAV_AcrIIC1X	AAV-compatible vector encoding AcrIIC1X	This work
59	AcrIIC1_FLAG	2xFLAG AcrIIC1 FLAG	This work
60	AcrIIC1 chimera-10_FLAG	2xFLAG AcrIIC1 with GSG-mCherry-GSG insertion behind Y70 FLAG	This work
61	AcrIIC1X_FLAG	2xFLAG AcrIIC1_N3F/D15Q/A48I FLAG	This work
62	HIS_AcrIIC1	6xHis TEV cleavage site AcrIIC1	This work
63	HIS_AcrIIC1 chimera-10	6xHis TEV cleavage site AcrIIC1 with GSG-mCherry-GSG insertion behind Y70	This work
64	HIS_AcrIIC1X	6xHis TEV cleavage site AcrIIC1_N3F/D15Q/A48I	This work
65	HIS_HNH_NmeCas9	6xHis TEV cleavage site HNH domain of NmeCas9	This work
66	HIS_HNH_SauCas9	6xHis TEV cleavage site HNH domain of SauCas9	This work
67	AcrIIC1_scaffold	AcrIIC1 with scaffold for miRNA binding site insertion in the 3’-UTR	Reference [5]
68	AcrIIC1_miR122	AcrIIC1 with 2x miR122 binding sites in the 3’-UTR	This work
69	AcrIIC1X_scaffold	AcrIIC1X with scaffold for miRNA binding site insertion in the 3’-UTR	This work
70	AcrIIC1X_miR122	AcrIIC1X with 2x miR122 binding sites in the 3’-UTR	This work

Supplementary Figure 3.20: List of constructs. NLS, nuclear localization signal; HA, Human influenza hemagglutinin.



Locus	Target sequence (5' to 3')
Firefly luciferase	GCGGGGAGAAGGCCAGGGGTCACTCCAG**GATT**
IL2RG	CTCTTTCTCCTCAAGGAACAATCAGTG**GATT**
FLJ00328	GGACAGGAGTCGCCAGAGGCCGGTGGTG**GATT**
AAVS1	ACCCCACAGTGGGGCCACTAGGGACAG**GATT**
DHFR	GTGATTTTATAGGTAAACAGAATCTGGT**GATT**
F8	GGTTTCTAGTTGTGACAAGAACACTGGT**GATT**
EMX1	GGCCTCCCCAAAGCCTGGCCAGG**GAGT**
GRIN2B	GAGAGTAGGCTGGTAGATGGAGTT**GGGT**
CCR5	GGTGGTGACAAGTGTGATCACTT**GGGT**
HBB	AGGGTTGCCCATAACAGCATCAG**GAGT**
CXCR4-1	GGACAGGATGACAATACCAGGCA**GGAT**
CXCR4-2	GATGATAATGCAATAGCAGGACA**GGAT**
VEGFA	GTGTGTCCCTCTCCCCACCCGTCCCTGT**CC**
FANCJ	AAAATTGTGATTTCCAGATCCACAAGC**CC**

Supplementary Figure 3.21: Genomic target sites. The sgRNA-complementary part is underlined. The PAM is shown in bold. Note, the HBB sgRNA contains a 5’ G which is not part of the genomic target site.



Locus	Direction	Sequence (5’ to 3’)
IL2RG	fw	ATGACACTGGTGGGTGTTCAG
	rv	TCTTCACCTTGCAGGCTCTCT
FLJ00328	fw	AGAGGAGCCTTCTGACTGCTGCAGA
	rv	AGGTCCTGGCCTTGCCTTCGA
AAVS1	fw	TGCTTTCTTTGCCTGGACAC
	rv	CCTCTCTGGCTCCATCGTAA
DHFR	fw	GCAGACTCCACACAGACGGT
	rv	GGGCCTACTGAATGATGGTTCAAG
F8	fw	GGGAGAGAACCTCTAACAGAACG
	rv	GCTCCAGGTGATGGATCATCAG
EMX1	fw	GGAGCAGCTGGTCAGAGGGG
	rv	GGGAAGGGGGACACTGGGGA
GRIN2B	fw	AGAATTTTGTAATTGGTTCTACCAAAG
	rv	ACAACAGTGGAAGAAAGCTAGGGC
CCR5	fw	GGCAACATGCTGGTCATC
	rv	GGTGTAAACTGAGCTTGCTCG
HBB	fw	ATGGTGCATCTGACTCCTG
	rv	ACTGTACCCTGTTACTTATCCCC
CXCR4-1 &-2	fw	AGAATTTTGTAATTGGTTCTACCAAAG
	rv	ACAACAGTGGAAGAAAGCTAGGGC
VEGFA	fw	ATCAAATTCCAGCACCGAGCGC
	rv	AGAACTCAGGACCAACTTATTCTG
FANCJ	fw	GTTGGGGGCTCTAAGTTATGTAT
	rv	CTTCATCTGTATCTTCAGGATCA

Supplementary Figure 3.22: Primers for genomic PCRs. Fw, forward primer; rv, reverse primer.

4 *De novo* design of site-specific protein binders using learned surface fingerprints

...

This chapter is based on ongoing work towards a more generic platform for the *de novo* design of site-specific PPIs exploiting learned surface fingerprint descriptors by MaSIF. An extensive protein design effort for the generation of novel protein binders is presented to target the PD-L1 at a specific location. Beyond the deepened understanding of the fundamental principles governing PPIs, this section also paves the way for an improved computer-driven PPI generation enabling the fast-track development of protein-based therapeutics and the targeting of sites previously thought to be “undruggable”. A manuscript based on this work is currently in preparation.

Preliminary authors

Sarah Wehrle[1,2,*], Pablo Gainza[1,2,*], Alexandra Van Hall-Beauvais[1,2,*], **Zander Harteveld**[1,2], Tan Shuguang[3], Jayne Marsden[1], Stéphane Rosset[1], Sandrine Georgeon[1], George F. Gao[3], Bruno E. Correia[1,2]

[*] These authors contributed equally.

Affiliations

[1] École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, CH. [2] Swiss Institute of Bioinformatics (SIB), Lausanne, CH. [3] Center for Molecular Immunology, Institute of Microbiology, Chinese Academy of Science, Beijing, CN.

Author contributions

B.E.C. conceived the initial idea and refined it together with P.G. and S.W.. Z.H. and P.G. developed the computational platform including MaSIF and Rosetta to generate protein binders. Z.H. designed the 3ONJ-scaffolded binder targeting the PD1 site of PD-L1. P.G. designed the 3S0D-scaffolded binder. S.W. and A.v.H.-B. performed biochemical experiments characterizing and optimizing the designs with the help of J.M., S.R., and S.G.. The solved x-ray structures of the PD-L1 - binder complexes were solved by our collaborators T.S. and G.F.G. B.E.C. directed the work. B.E.C., S.W. and P.G. wrote the manuscript with support from all authors.

4.1 Abstract

Protein-protein interactions (PPIs) play a crucial role in virtually all living processes. The *de novo* design of PPIs stands as a strict assessment to the understanding of the underlying principles driving molecular recognition and opens possibilities for the development of protein-based therapeutics to target specific sites on protein molecules previously thought of as “undruggable”. PPIs emerge between proteins that display highly complementary chemical and geometric molecular surfaces, forming buried interfaces with a large number of contacts. Here, we hypothesize that the key ingredients for the design novel, site-specific PPIs is to (1) identify surface regions on the target with a high propensity to become buried and then (2) identify complementary surfaces that can optimally complement the identified regions. We introduce a *de novo* PPI design platform based on learned surface fingerprints calculated through a geometric DNN. The fingerprint descriptors are thought to efficiently capture features that are important determinants for molecular recognition. Based on the surface fingerprints, we selected peptide-binding fragments to engage a defined site on PD-L1, and subsequently grafted them onto protein scaffolds to confer stability and refined them to add additional contacts. We designed helical binders that engage a region that overlaps with the PD1 binding site. The two experimentally improved binders showed a high affinity for PD-L1(K_Ds in the range of 100-50 nM). The solved crystallographic structure of the protein complexes and site saturation mutagenesis experiments on the binding interface revealed excellent agreements to the computational models. The binding motifs unveiled by our method display completely novel interaction motifs currently unobserved in nature. Ultimately, this work presents a surface-centric perspective to understand molecular recognition and presents a robust route for the *de novo* design of PPIs to generate targeted diagnostics and therapeutics.

4.2 Main

Designing novel protein-protein interactions (PPIs) remains a fundamental challenge in computational protein design, with broad basic and translational applications [161, 149, 275, 173]. *De novo* PPI design consists in generating protein amino acid (AA) sequences that can engage a site on a target protein and thereby form a *de novo* quaternary complex. Fundamentally, the problem tests our understanding of the forces that drive biomolecular interactions and our capacity for structure-based PPIs prediction. Also, *de novo* PPI design has a tremendous biomedical importance as it could be utilized to rapidly and affordably engineer protein-based therapeutics with tailored biophysical properties only difficult to achieve with conventional screening platforms [276].

Despite recent advances in rational PPI design [165, 277, 278] and PPI prediction driven by AlphaFold2 (AF2) [108], designing novel protein binders against specific targets remains a challenge, particularly when no structural elements from preexisting binders are used. Current state-of-the-art methods for *de novo* PPI design [279, 280, 154], in particular hotspot-centric approaches [281] and rotamer information fields (RIF) [77], rely on placing key interaction residues (hotspots) [166] on the interface followed by transplantation of these onto protein structures acting as scaffolds to optimally present them in an energetically-favored conformation. However, the main challenges current methods face include (1) the identification hotspot residues and (2) the placement of the hotspots onto a protein scaffold while designing for a well-packed protein interface around them. In fact, hotspot residues are often found in deep hydrophobic pockets [168], thus the challenge is amplified when targeting "flat" macromolecular interfaces with shallow pockets.

A long-standing model of molecular recognition proposes that PPIs form between proteins with geometric and chemical complementary molecular surfaces [282, 145]. Proteins in their apo state populate ensembles of low-energy conformations [283]. When two proteins with partially complementary conformations encounter each other [284], binding is induced through the formation an interface with a well-packed buried and hydrophobic core that is often surrounded by a polar rim region [166, 148]. Hence, a requirement for successfully creating novel protein complexes is the sculpting of high shape complementary and well-packed buried interfaces upon complexation.

Multiple success of rationally designed PPIs have led to protein-based therapeutics such as antibodies and inhibitors, vaccine design, and more [172, 285, 286]. However, more methods are needed for designing PPIs to various surface types and protein sites [287]. Here, we introduce a novel design approach based on learned surface fingerprints which we hypothesise to more efficiently capture features that are determinant for molecular recognition. We use this method on an impactful and therapeutically relevant target: PD-L1. The use of this target as test case for our method highlight the relevance of such methods and the need for continuous development of new approaches.

4.3 Results

4.3.1 Design of *de novo* PPIs using learned surface fingerprints

Previously, we have introduced a geometric deep learning (DL) framework termed MaSIF (molecular surface interaction fingerprinting) [171], to extract fingerprints from protein surfaces that can leveraged to learn deterministic patterns of interactions of proteins with other biomolecules. The fingerprints summarize the information present in the molecular surfaces and can concomitantly be used to identify patterns in the surfaces, such as the propensity of molecular surface regions to form buried interfaces [171].

Having shown that MaSIF had robust performances in prediction tasks, we sought to test whether we could leverage this framework for the design of novel PPIs by targeting defined sites in other proteins only using structurally-derived information from the target binder. We approached the *de novo* PPI design problem by devising an approach that relies on two objectives: (1) the prediction of surface sites with high binding propensity using the MaSIF-site predictor [171] and (2) employing surface fingerprints, a surface-centric search (MaSIF-seed-search) for short binding peptide motifs referred to as binding "seeds" that display the required features to engage the predicted site, and ultimately template for the design of a productive binding interaction.

For the seed search, we reasoned that a viable approach is to first identify seeds that display complementary surfaces with respect to the target surface, and use the structural fragment from the seed forming this complementary surface as a starting point to generate viable PPIs (Fig. 4.1a,b). We hypothesized that seeds can be found within the vast number of structural fragments available in known structures (Fig. 4.1a). Once identified, the atomic elements that form the seed are transferred to a protein scaffold to confer stability and form additional favorable contacts with the target protein (Fig. 4.1c) using established motif grafting techniques [163].

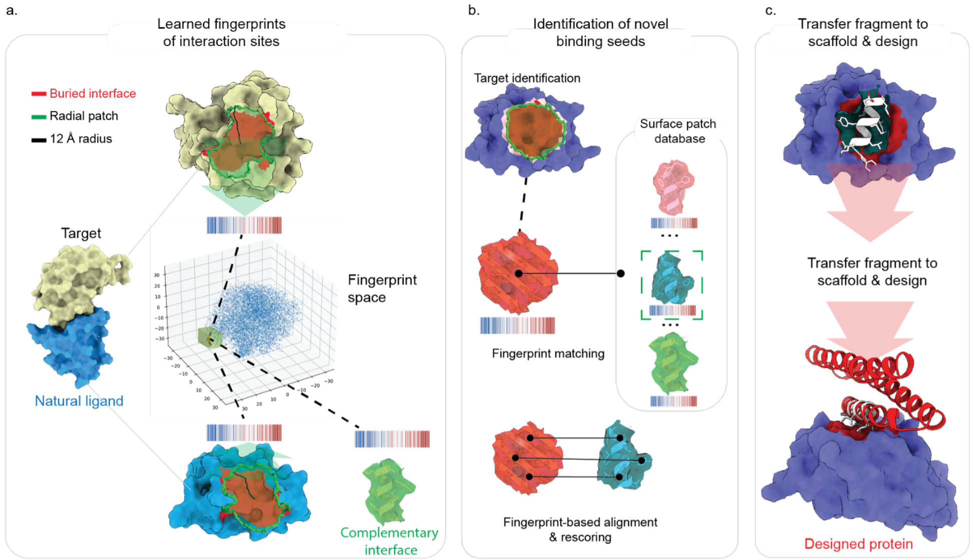

a.
Learned fingerprints of interaction sites
Buried interface
Radial patch
12 Å radius
Target
Natural ligand
Fingerprint space
Complementary interface
b.
Identification of novel binding seeds
Target identification
Surface patch database
Fingerprint matching
Fingerprint-based alignment & rescoring
c.
Transfer fragment to scaffold & design
Transfer fragment to scaffold & design
Designed protein

Figure 4.1: Surface-centric design of protein interactions. **a:** Protein binding sites are spatially embedded as fingerprints. Protein surfaces are decomposed into overlapping radial patches. A DNN learns to embed the complementary fingerprints within close regions of space. Here we used t-SNE [288] to visualize a sub-sample of the fingerprint space. A green box shows the location in space of complementary fingerprints. **b:** Protocol to identify new binding fragments. A target patch is automatically identified based on the propensity to form buried interfaces. A fingerprint is then computed on this patch and all complementary fingerprints in a large database (~140M patches) are compared. A short list of patches is selected, and the fingerprints are used to align and re-score patches. **c:** Transfer fragment and design. The selected fragment is transferred to a protein scaffold and the rest of the interface is designed. The designed protein is then tested experimentally.

The identification of binding motifs that can mediate high-affinity interactions remains difficult because the space of possible conformations is extremely diverse and sensitive to minor atomic-level changes. For example, misplaced methyl groups or incompatible charges can render the binding motifs incapable of sustaining a productive binding interaction. Therefore, a remaining challenge in computational protein design is to accurately identify viable binding motifs can effectively be recycled as seeds for the *de novo* design of PPIs.

MaSIF can search large libraries of potential binding seeds with remarkable speed relying on learned surface features instead of handcrafted descriptors, which may potentially increase the accuracy of seed identification. To perform binding seed searches, we developed a MaSIF-based *de novo* PPI generation platform called MaSIF-seed-search. The MaSIF-seed-search protocol can search a large database of seeds for potential interactions to a defined target surface. To achieve this, the protein molecular surfaces are first decomposed into overlapping radial patches with a radius of 12 Å (Sup. Fig. 4.1a) capturing enough buried surface area. For each point within the patch, we computed chemical and geometric features inputted into a deep neural network (DNN) that is trained to output fingerprints that are complementary between interacting pairs and dissimilar between non-interacting pairs [171]. This approach — when applied to patches centered at the core of buried interfaces— results in good discrimination between interacting and non-interacting pairs (Sup. Fig. 4.1e).

We sought to design *de novo* protein binders to engage an important and challenging protein target of biological relevance. For this, we computationally designed and experimentally validated protein binders against the PD-L1 protein, an important target in the field of immuno-oncology.

4.3.2 Fingerprint-based *de novo* design of protein-binders against PD-L1

PD-L1 is an all-βprotein adopting an Immunoglobulin (Ig)-like fold. Importantly, no known helical binders have been discovered for PD-L1. PD-L1 displays a rather flat interface (Sup. Fig. 4.2) and classifies as a notoriously “hard-to-drug” target using small molecules [289]. We selected the structure of PD-L1 co-crystallized with nanobody VHH [290] (PDB id: 5JDS) as our target. We compared the buried interface of PD-L1 in this crystal structure with the buried interfaces of 1,380 transient PPIs and found it to be among in the 99th percentile in terms of surface flatness (ranked #7 among 1,380 interfaces) (Sup. Fig. 4.2).

We computed the fingerprint for the patch in the center of the interface and compared it to our database of fingerprints. Fragments with similar fingerprints were selected for alignment and scoring (Fig. 4.1b, bottom), and re-ranked. We noticed that among the top results, helices clustered in orientations anti-parallel to the β-sheets of PD-L1 (Fig. 4.2a,b and Sup. Fig. 4.3a). Among the most populated cluster, we noticed a convergence of AA for a twelve-residue fragment (Fig. 4.2a,b).

Next, we used the MotifGraft program in Rosetta [163] to search for scaffolds that could display this fragment while having favorable steric interactions with the target, and then use

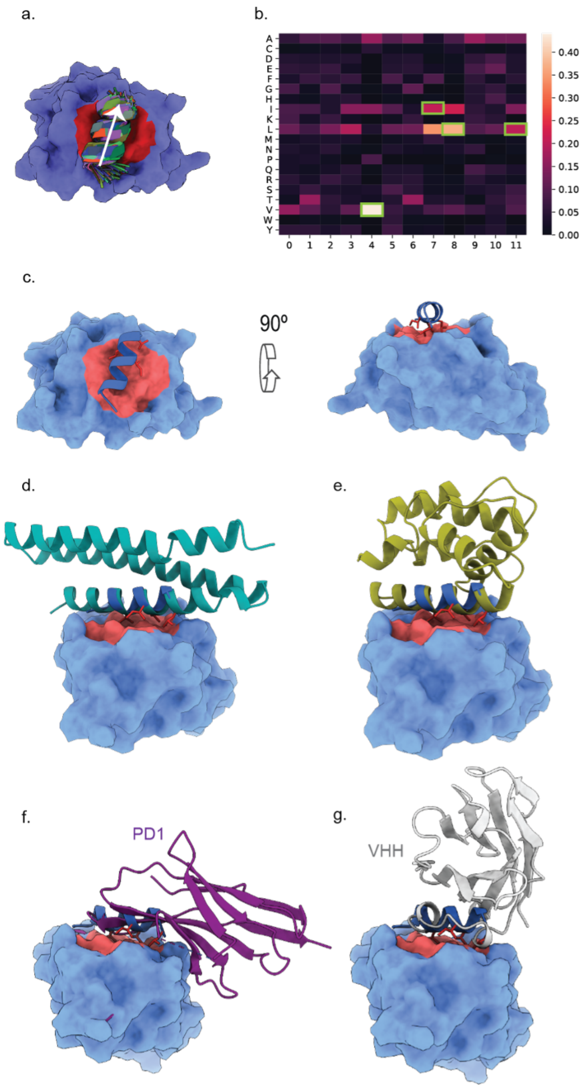


Figure 4.2: Seed-based design of PD-L1 binders. **a:** Clustering of top-ranking binding seeds (12 residues in length) from the most populated cluster. PD-L1 is shown in blue and the predicted buried interface is shown in red. **b:** Heat map (frequencies) of the amino acid identities of each position in the clustered binding seeds. The selected residue is outlined in green. **c:** Top and side view of the consensus helix. **d:** Model of the seed grafted onto the three-helix scaffold with PDB id: 3ONJ. **e:** Model of the seed grafted onto the helical scaffold with PDB id: 3S0D. **f-g:** Comparison of the binding seed with two known binders of PD-L1, the natural binder PD-1 (PDB id: 4ZQK) and a high affinity nanobody (PDB id: 5JDS).

RosettaDesign program [291] to optimize the contacts in the interface. Twenty-four designs were tested in yeast and two designs showed a binding signal above background noise. We selected these two designs for further optimization. One design was based on an odorant-binding protein from *Apis mellifera* (PDB 3S0D) and the other one on a SNARE protein found in yeast (PDB 3ONJ) (Fig. 4.2c,d,e). Both designs showed weak binding to PD-L1 when displayed on yeast. The binding signals were low but distinctly higher than binding signals of the wildtype protein controls. Soluble expression in *Escherichia coli (E. coli)* and mammalian cells failed for both proteins, hence we optimized binding affinity and solubility/stability of the proteins by designing computer-guided combinatorial libraries and sorting using yeast display.

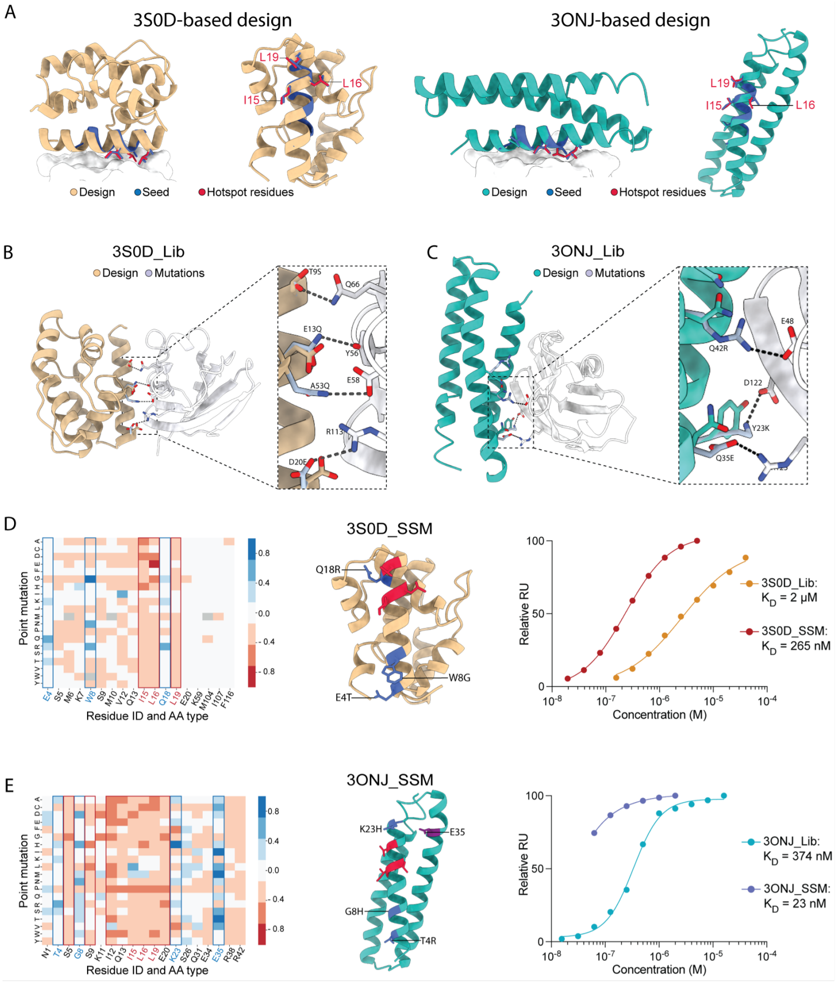


Figure 4.3: Design and optimization of the two PD-L1 binder designs. **a:** Both designs share the same seed (blue) with two leucines and one isoleucine as hotspot residues (red). **b:** Library generation and affinity sorting of the 3S0D-based design enabled the formation of an additional H-bonds between Q53 of the design and E58 of PD-L1. The other mutations maintain bond formations. The design binds to PD-L1 with an affinity of 2 µM. **c:** Affinity sorting of the 3ONJ-based design resulted in the formation of three additional salt bridges yielding an affinity of 374 nM. **d and e:** SSM library sorting and deep sequencing was conducted for both designs. The results confirm the high importance of the three hotspot residues and revealed potential positions to improve binding affinity. **d:** After introducing three mutations in the binding interface according to the SSM results (blue), the affinity of the 3SOD-based design improved to 265 nM. **e:** Mutation of three residues in the interface improved the affinity of the 3ONJ-based design to 23 nM. The SSM data also showed that the glutamate at position 35 was suboptimal. However, mutation of this position (purple) resulted in a loss of soluble expression.

The stability issues of the 3S0D protein mainly resulted from a large void in its core. To keep the library complexity in a feasible scope, we decided to optimize the affinity and the stability of the protein in two consecutive libraries. Mutations of the affinity-directed library were mainly conducted on the helix bearing the selected seed. A few positions potentially allowing additional contacts to PD-L1 without modifying the hotspot residues were selected. Yeasts were sorted with decreasing PD-L1 concentrations and sequencing performed after the third sort. The obtained mutations were mild but distinctly enriched. The mutations allowed for the new formation or improvement of hydrogen bonds (H-bonds) with PD-L1 increasing its binding affinity considerably on the surface of yeast. The most substantial change occurred at position 53, the mutation of alanine to glutamine allows for the formation of a new H-bond with the glutamate 58 in PD-L1. The mutation of aspartate to glutamate at position 20 maintains the formation of a salt bridge with arginine 113 in PD-L1 (Fig. 4.3a,b). However, the glutamate is beneficial for binding since no other amino acid mutation was found at this position after the third sort. Also, mutation to a small nonpolar residue, like alanine, or a bulky positively charged residue, like arginine, decreased the binding affinity significantly indicating the importance of this salt bridge for binding. As the protein was still not expressing in the soluble fraction in *E. coli* or mammalian cells, we designed a second library targeting the core residues of the protein. The allowed mutations aimed to increase the size of the core residues and thereby decrease the void in the core to stabilize the protein. The enriched mutations allowed for the soluble expression of the protein in mammalian cells. Binding affinity measurements of the soluble protein resulted in a binding affinity of 2 µM (Fig. 4.3d) which is comparable to binding of wildtype PD1 (K_D = 8.4 µM). Also, the binding kinetics of the protein with its fast on and off-rate is comparable to wildtype PD1.

The inability to express the 3ONJ protein was likely resulting from the large hydrophobic interface introduced during the design procedure. By targeting residues to increase the affinity as well as hydrophobic residues not directly being involved in binding would make the protein more hydrophilic. The library increased the affinity considerably on yeast. The enriched mutations not only increased the binding affinity but also improved the overall hydrophilicity of the protein and therefore enabled the soluble expression in *E. coli*. Most important for binding are the mutations at position 23, 35 and 42. All three mutations allow for the formation of additional salt bridges with PD-L1, increasing the binding affinity of the design (Fig. 4.3a,c). The protein was also monomeric, α-helical, and thermally stable. The affinity of soluble protein to PD-L1 yielded a K_D of 374 nM (Fig. 4.3d), more than ten-times higher than the affinity of wildtype PD1.

The importance of the hydrophobic seed residues for binding was shown with single point mutations. Exchanging valine 12 or leucine 16 with an arginine knocked out binding almost entirely. Mutating leucine 16 to an alanine had a weaker influence on affinity, due to its chemical and geometrical characteristics that are more similar to one and another. To further improve the affinity of the proteins and to learn more about their binding modes and potential shortcomings of our designs, we constructed a site saturation mutagenesis (SSM) library, targeting 24 residues in the binding interface of the 3S0D protein. The non-combinatorial character of the SSM library allows screening of many more positions and residues. A binding and a non-binding population of the library was sorted and analysed using next generation sequencing (NGS). The analysis showed that the three hotspot residues are crucial for binding, and that almost all mutations at these positions impair the binding to PD-L1. The data further revealed that the tryptophan at position 8 was sub-optimal and that a smaller residue at this position could improve binding. Also, the glutamate at position 4 and the glutamine at position 18 showed up as potential targets to improve the binding affinity. Different additional single mutants and combinations were tested for their binding affinity to PD-L1. Most of the mutations were able to improve the binding affinity. A combination of three mutations (E4T, W9N, Q19R) resulted in a ten-fold increase of the binding affinity, giving a K_D of 256 nM (Fig. 4.3d).

Since both designs shared the same seed we decided to apply the SSM mutations of the 3S0D design also to the 3ONJ design. The insertion of three mutations (E4T, W8G, V12I) improved the K_D to 120 nM. However, a more profound analysis of this protein binding mode was still crucial. Therefore, we again constructed an SSM library targeting 24 residues in the binding interface. The data confirmed the importance of the hotspot residues for PD-L1 binding, as mutations in these positions decreased the affinity of the protein. Some of the improvable positions that were already seen in the 3S0D design SSM library were also seen in the SSM library specific to 3ONJ design. Strikingly, the data revealed new positions that could potentially increase binding. The combination of three mutations (T4R, G8H, K23H) were able to improve the binding affinity by almost ten-fold to a K_D of 23 nM (Fig. 4.3d). Position 35 showed the most potential to improve binding affinity. However, mutating this position to threonine or isoleucine resulted in the inability to express the protein. Therefore, this position has a strong impact on solubility of the protein.

The SSM data of both designs revealed a high importance of the polar serine residues in the surrounding of the interface. To validate the binding mode, we co-crystallized the designs with PD-L1. The structures (Fig. 4.4) showed excellent agreement with our computational models. Our experimental data validates the accuracy of the selected seed as a PD-L1 binder and demonstrates the importance of the hotspot residues to confer for a specific binding interaction.

4.4 Discussion

Long-living PPIs are formed between proteins with complementary molecular surfaces. Here we build on this concept to propose that *de novo* PPI design benefits from a surface-centric approach, particularly when designing the buried surface areas of the interface. Molecular surfaces are highly diverse, but at the same time they are difficult to model, interpret, or even compare. In this work, we proposed a new method based on the geometric DL tool MaSIF, to overcome this limitation and both identify patches with a high propensity to form buried surfaces as well as to scan for binding seeds with complementary binding surfaces. The discovered binding seeds were then used as an interface core to design novel binding

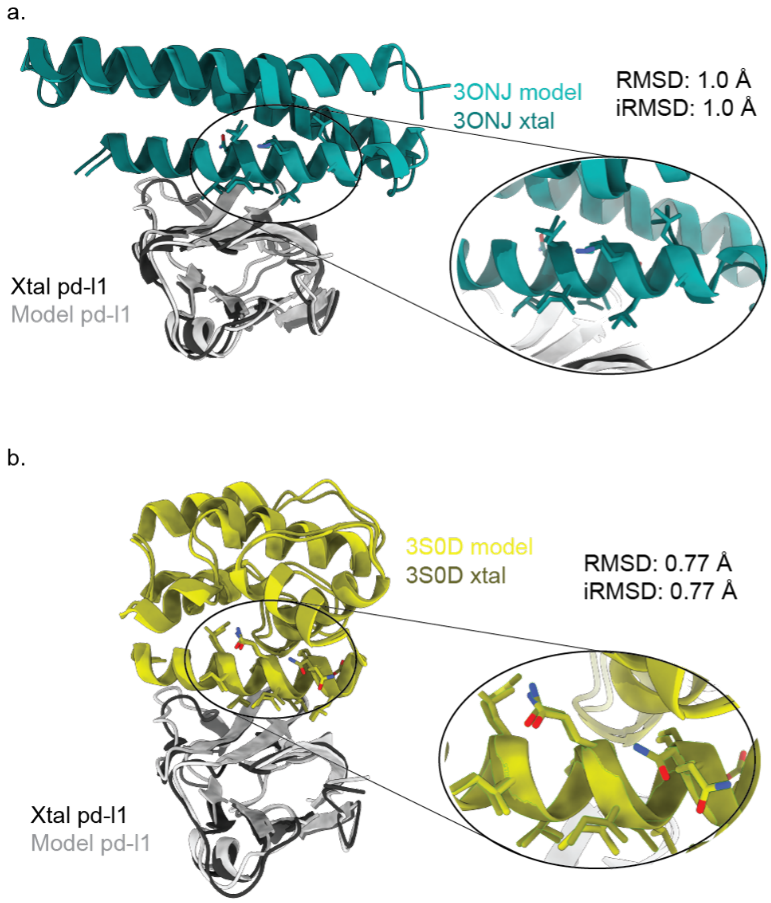


Figure 4.4: Structural validation of designs. **a and b:** Structural validation of designs, shown by aligning the computational model (lighter color) with the experimental structure (darker color). The inset shows the alignment of the residues in the binding seed. **a:** 3ONJ model. **b:** 3S0D model.

proteins against the PD-L1 protein, one of the flattest protein interfaces known.

MaSIF learns representations of patches in molecular surfaces that are involved in PPIs. By embedding our molecular surfaces as fingerprints, we overcome a technical challenge: how to rapidly identify surface fragments that can engage a specific target. We showed that the method is highly sensitive to small surface deviations, and thus able to scan for the correct patches in a set of benchmark helical-binding proteins among hundreds of millions of patches from helical fragments. In other words, in the absence of protein flexibility a single surface patch captures enough information to distinguish elements from highly similar motifs containing the same underlying secondary structure.

To our knowledge the PD-L1 protein does not form interactions with helical domains. We transferred the top found seed to a helical protein and redesigned the interface. Our designed binders showed exquisite agreements between experiment and structure, with an root-mean-square deviation (RMSD) at or below 1.0 Å. One limitation that arose from our work is that our initial computational designs showed binding in yeast display barely above background. We optimized the affinity with computationally-guided libraries, diversifying the polar residues surrounding the binding seed. After optimizing these designs and evaluating the top designs with SSM, we noticed that the main improvements resulted from the solubility of the designs (and thus, help with expression), and also from polar AAs (serine, glutamine) that are atypical hotspot residues. This highlights the need for improvements of current computational interface design for polar interactions, which has also been observed by other authors [148, 292]. We believe that the precision in terms of complex alignment that our method provides, along with the large amounts of data becoming available, and progress in the field of geometric DL will allow us to achieve this.

A second limitation arose from our reliance on natural proteins to serve as scaffolds for our identified fragments. Natural proteins are known to be marginally stable [293] and introducing mutations can result in proteins that do not express or express poorly. Other authors have also noticed this problem with natural proteins and moved altogether to highly stable *de novo* proteins which express at high level even in the presence of many surface mutations [294, 153].

A final limitation arises from our reliance on helical fragments to form the initial binding seed. We selected these because they are ubiquitous in nature, because they are relatively stable, and there is a long history of successful design efforts to transfer helical fragments to novel scaffolds. However, it is unlikely that all proteins can be targeted by a helical binding seed. So far, we have been unable to successfully apply our method differently structured seeds that are more diverse, unstable and difficult to transfer onto protein scaffolds. We believe this requires further methodological advances.

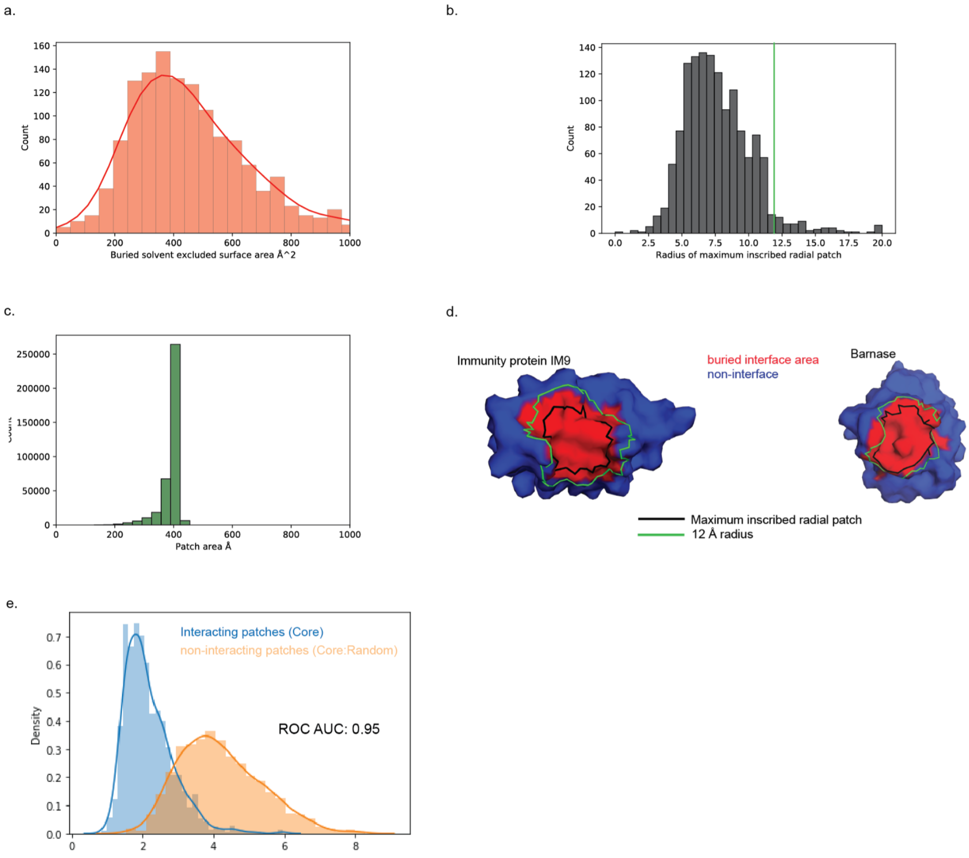
In this work we scanned large databases of protein fragments to identify complementary binding seeds. We envision that in the future, generative DNNs will be able to simultaneously process the fingerprint of the target and generate a complementary protein fragment. In parallel, current efforts to generate (or hallucinate) protein folds will also be applicable to generate scaffolds that will ideally support the generated binding seeds [138, 140, 141].

4.5 Methods

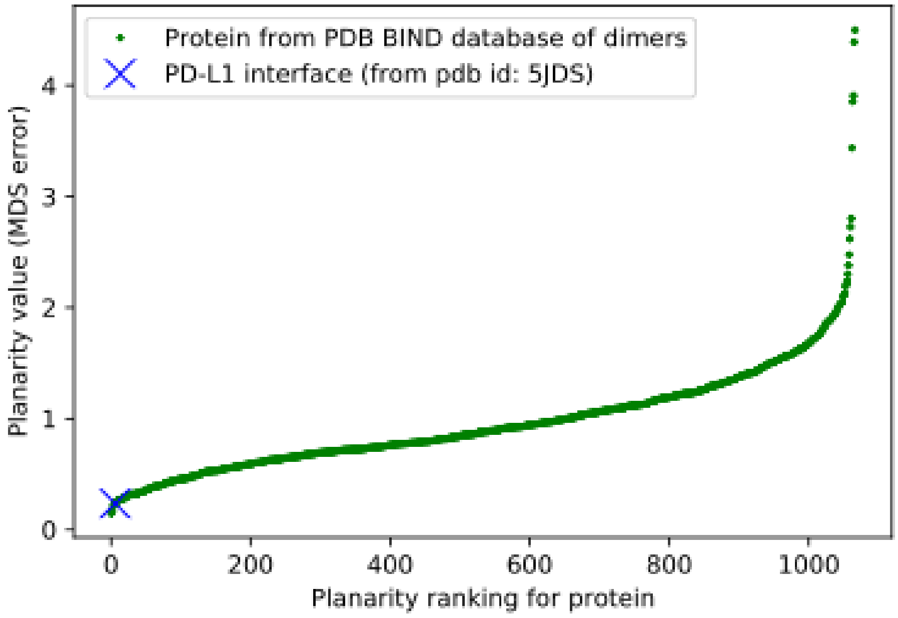
4.5.1 Computational grafting and interface optimizations

We used the Rosetta MotifGraft method [163] to search for potential scaffolds able to accommodate the hotpsot residues of the helical seed over a database of globular, medium sized (sizes 50 – 140 residues) proteins. We selected the top 20 grafts by RMSD between the hotspots on the seed and on the graft. Importantly, we verified that the grafts were reported monomeric and soluble. Next, we used fixed backbone interface design using Rosetta [291] on the non-hotspot residues to optimize interactions across the interface. We restricted the design to several AA types based on the underlying electrostatic potentials calculated with the Adaptive Poisson-Boltzmann Solver (APBS). We generated a set of 200 sequences and selected the top candidates by surface complementarity (> 0.6) and $\Delta\Delta G$ (ddG) score for experimental validation.

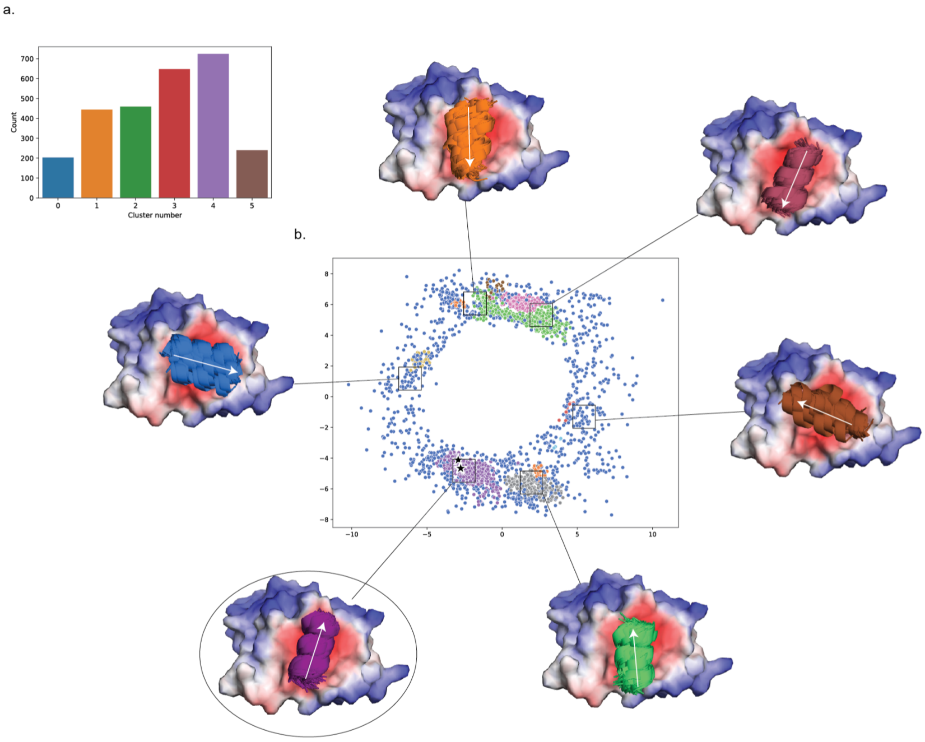
4.6 Supplementary information



Supplementary Figure 4.1: Modeling buried surfaces as radial patches **a:** Histogram of the area of the buried surface area on 1380 dimeric PPIs. We note that areas are computed for only one of the proteins (i.e. each subunit in a PPI is computed separately). **b:** Size of the maximum inscribed radial patch for the 1,380 proteins. Patch area for the patches used here (12 A), for a set of 30,000 randomly selected patches. **c:** Histogram of the patch areas of thousands of randomly selected protein patches with a fixed radius of 12 Å. **d:** Example of the buried interface area for two well known, high affinity binders, Immunity Protein IM9 (PDB id: 1EMV) and the protein Barnase (PDB id: 1BRS). The buried interface of each protein when bound to its partner is shown in red. The maximum inscribed radial patch's circumference is shown in black, and the circumference of a patch with radius 12 Å is shown in green. **e:** Histogram of similarities between MaSIF-seed-searches: (blue) pairs of patches that are co-crystallized from transient PPIs, with the fingerprint computed for the patch centered on the largest inscribed radial patch, and (orange) pairs of patches where one was taken from the center of the interface of a random PPI and the other was taken from a random patch surface.



Supplementary Figure 4.2: Planarity of the PD-L1 interface in structure with PDB id: 5JDS. y-axis: error in multidimensional scaling when flattening the patch from 3D to 2D. x-axis: ranking of each protein. The PD-L1 interface used here is marked with a blue X.



Supplementary Figure 4.3: Clusters of binding seeds docked on the PD-L1 surface (PDB id: 5JDS). 140 million patches from ~250,000 helices extracted from the PDB were compared and docked to the predicted interface in PD-L1 using MaSIF-seed-search. The top scoring seeds were selected for further processing. 12-residue fragments of these seeds that occupied the largest buried surface were then clustered using metric multi-dimensional scaling (MSD) of all pairwise RMSDs between all seeds. **a:** Histogram of clusters, showing the prevalence of each orientation. **b:** Plot of the clusters in the MDS. A box is drawn around the center of each cluster and the picture shows the selected helix orientation for all points inside the box. A star shows the location of the PD-L1 seed used for the designs.

5 A generic framework for hierarchical *de novo* protein design

...

This chapter describes the improvements and implementation of a hierarchical *de novo* protein design method termed "TopoBuilder". The code is available under https://github.com/LPDI-EPFL/topobuilder and the scripts and examples can be downloaded from https://github.com/LPDI-EPFL/tbpipeline. A manuscript is currently being finalized.

Authors

Zander Harteveld[1,2*], Jaume Bonet[1,2*], Stéphane Rosset[1], Che Yang[1,2], Fabian Sesterhenn[1,2], and Bruno E. Correia[1,2]

[*] These authors contributed equally.

Affiliations

[1] École Polytechnique Fédérale de Lausanne, Lausanne, CH. [2] Swiss Institute of Bioinformatics (SIB), Lausanne, CH.

Author contributions

B.E.C. and Z.H. conceived the initial idea and refined it together with J.B., Z.H., F.S., C.Y., and B.E.C.. Z.H. and J.B. designed and performed experiments. Z.H. and J.B. wrote the software. S.R. purified the designs and performed protein-biochemical characterization. Z.H. performed *in silico* structural analysis and modeling with the support of F.S. and C.Y.. B.E.C. directed the work. B.E.C. and Z.H. wrote the manuscript with support from all authors.

5.1 Abstract

De novo protein design enables the exploration of novel sequences and structures absent from the natural protein universe. *De novo* design also stands as a stringent test to our understanding of the underlying physical principles of protein folding and may lead to the development

of proteins with unmatched functional characteristics. The first fundamental challenge of *de novo* design is to devise “designable” structural templates leading to sequences that will adopt the predicted fold. Here, we built on the TopoBuilder *de novo* design method, to automatically assemble structural templates with native-like features starting from string descriptors that capture the overall topology of proteins. Our framework eliminates the dependency of hand-crafted and fold-specific rules through an iterative, data-driven approach that extracts geometrical parameters from structural tertiary motifs. We evaluated the TopoBuilder framework by designing sequences for a set of five protein folds and experimental characterization revealed that several sequences were folded and stable in solution. The TopoBuilder *de novo* design framework will be broadly useful to guide the generation of artificial proteins with customized geometries, enabling the exploration of the protein universe.

5.2 Main

Evolution has only explored a small subset of all possible amino acids (AA) sequences and structures [42]. The space of viable protein sequences, e.g. sequences that have a global free energy minimum representing a well-folded native state is small. Such a notion has been supported by several experimental studies showing that many random AA sequences have a rough energy landscape with multiple local minima representing aggregated or misfolded states [72, 73, 74, 153].

De novo design strategies stand as an essential tool to aid the exploration of the sequence space and thereby enabling the creation of new protein structures and functions. Classical *de novo* protein design generally entails two iterative steps: first, target folds are modeled (backbone generation); second, an AA sequence that stabilizes the lowest free energy state of the target backbone conformation is searched (sequence design). Despite multiple successes [295, 60, 296, 85], *de novo* design remains a challenging problem for protein designers given that it stands as a stringent test to our understanding of the principles that govern protein structures.

Successful structure-based *de novo* design largely relies on the crafting of “designable” protein backbones, meaning physically realistic and strainless backbones that are compatible with sequences that will yield a protein fold with a well-defined energy minimum [89, 92, 91, 297, 298, 93]. The designability of a protein backbone is generally proxied by the number of sequences that it can support [89, 90, 95]. For example, some natural protein structures can accommodate more sequences than the average and are thought to be more robust against random mutations, and therefore thermodynamically more stable which favors evolutionary stability [90]. Generating designable backbones is important as one would like to *a priori* limit the sampling space to engineer only reasonable shapes with inherent structure-to-sequence compatibility and discard presumptive non-viable structures. Many *de novo* design approaches are likely to fail due to the lack of designability of the starting structural templates, requiring multiple iterative rounds of human-guided and experimental optimizations [74, 153].

Quantifying the designability of protein backbones is difficult [94, 299]. Even recent energy functions fail to reliably capture global designability aspects of protein backbones but excel in assessing high-resolution details such as van der Waals (vdW) forces, steric repulsion, electrostatic interactions, and hydrogen bonds (H-bonds) [97, 281]. To facilitate the design process at early stages, it would be necessary to have low-resolution energy functions that could accurately capture the physicochemical determinants of realistic structures at the backbone level [300].

There has been considerable progress in developing parametric functions and general principles for describing ideal and less symmetric protein structures [95, 99, 301]. Often secondary structure elements (SSEs) are connected to create tertiary structural topologies by packing α-helices on paired β-strands through the control of the loop length and ABEGO residue structure. This approach made it possible to design a set of ideal protein structures, including TIM-barrels [76], β-barrels [77, 302], jelly-rolls [87], and Immunoglobulin-like domains [88]. Nonetheless, parametric definitions are often specifically framed for distinct protein classes or architecture types and cannot be generalized to other architectural configurations i.e. the Crick coiled-coil generating equations [80] or descriptive parametric models of β-barrels [303, 304].

In this work, we greatly enhanced the capabilities of the TopoBuilder framework by introducing a data-driven correction module to generate native-like backbones from a simplified description of a protein topology that we term “Sketch” (Fig. 5.1A and Sup. Fig. 5.1). This module is applicable to any protein topology that can be described by arranging ideal SSEs in layers [174, 175]. The correction module generates parametric refinements that geometrically optimize the SSEs of protein backbones towards native-like configurations, rendering them more designable. The set of corrections includes translational and rotational parameters jointly capturing key geometric features such as distances and angles of native tertiary motifs. To further aid the topology assembly step, we utilize structural fragments from naturally occurring loops to connect two subsequent SSEs. We evaluated the general framework by *de novo* designing five different folds and found that even a minimal set of corrections to the protein backbones are sufficient to improve sequence sampling and achieve a better sequence-to-structure compatibility and sequence quality overall according to a variety of computational metrics. Finally, we experimentally characterized 54 designs and obtained multiple sequences that were folded and stable in solution, including topologies that have been particularly hard for computational design such as all-β structures.

5.3 Results

Evolution proceeds incrementally through random and sparse sampling of the possible sequence space which in turn populates the protein structure space. However, nature seems to show a tendency to reuse the same protein structures repeatedly based on the observation that

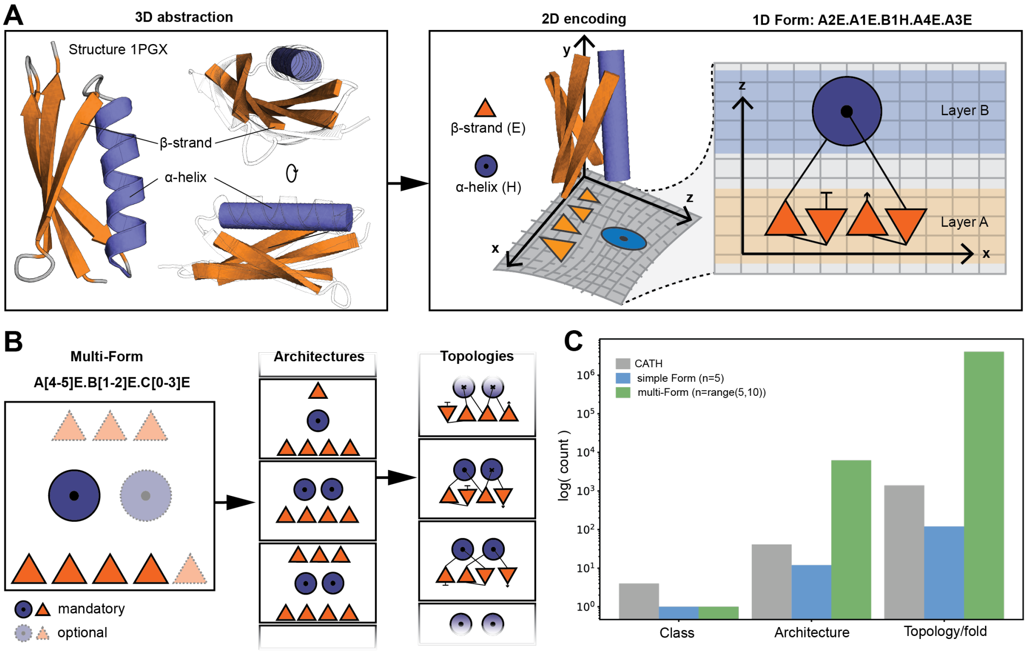


Figure 5.1: Form parametrization for protein structures. **A:** Form parametrization for a Ubiquitin-like fold (PDB 1PGX). A 3D abstraction of the structure is created that is encoded into a layered 2D lattice diagram where the sheet is assigned as layer A and the helix as layer B (layer assignment is arbitrary). The SSEs on a layer are dispersed on the x-axis, and layers are stacked onto each other following the z-axis. The lattice representation is summarized into a Form descriptor as shown on the top. The Form describes each SSE by the layer, relative position in the layer, and secondary structure type separated by a dot (N- to C-terminal sequence order is preserved). **B:** A multi-Form string created by assigning some SSE as mandatory and others as optional. The flexibility allows the sampling of a range of architectures and topologies. **C:** Comparing the exploratory capacity between a simple Form (five SSEs), a multi-Form (a minimum of five and a maximum of ten SSEs) and the known space of protein folds (as classified by CATH). The simple Form nearly samples as many existing topologies as known, while the multi-Form greatly generates more topologies than what can be found in nature.

the discovery of new protein folds has become rare [2, 3, 1]. In some regards, the mapping of structural space poses a number of challenges since it depends on the structural definition and coarseness of the structures. For a more systematical exploration of the structural space, Taylor and colleagues defined an idealized SSE lattice representation that can easily be captured through a simple string descriptor called Forms [305, 3] (Fig. 5.1A). The Form parametrization describes proteins as layered topologies, with each layer being composed of a defined number of either α-helices or hydrogen-bonded (H-bond) β-strands (Fig. 5.1A). Although constrained by its grid-like tabular system, a wide range of structural configurations can systematically be defined, potentially allowing to fully explore a protein topological space at orders of magnitude larger than the natural space currently characterized (Fig. 5.1B, C). Although Forms are well suited for protein topology comparison and classification [306], using them for *de novo* design is challenging due to the loss of crucial structural and sequence features, including native tertiary configurations of SSEs and side chain representations.

5.3.1 Sketching native-like protein backbones from Forms

Given the Form description, the TopoBuilder starts by placing ideal SSEs at their respective relative positions as specified by the Form description, creating a three-dimensional (3D) backbone object containing only SSEs, which we refer to as “Sketch”. We define the layer stacking along the z-axis, with inter-layer separations of 8 Å for β-sheets, 10 Å - 11 Å for α-helices, and mixed α/β structures [307]. The y-axis aligns along the direction of the SSEs i.e., following the sequence directions of the α-helices and β-strands. The intra-layer spacing between adjacent SSEs is therefore along the x-axis and typically of 10 Å for α-helices and 4.85 Å for β-strands (Fig. 5.1A) [308].

The Sketch representation does not contain loops connecting the SSEs and has no sequence information. Furthermore, the naive SSE assembly shows a rather non-native configuration of the protein structure strongly hinting towards structures that present non-designable configurations. Fine-grained structural details at the secondary structure and tertiary levels such as β-sheet pleatings and curvatures, and α-helical packing are absent. These features are difficult to sample automatically and correctly even with low resolution scoring functions.

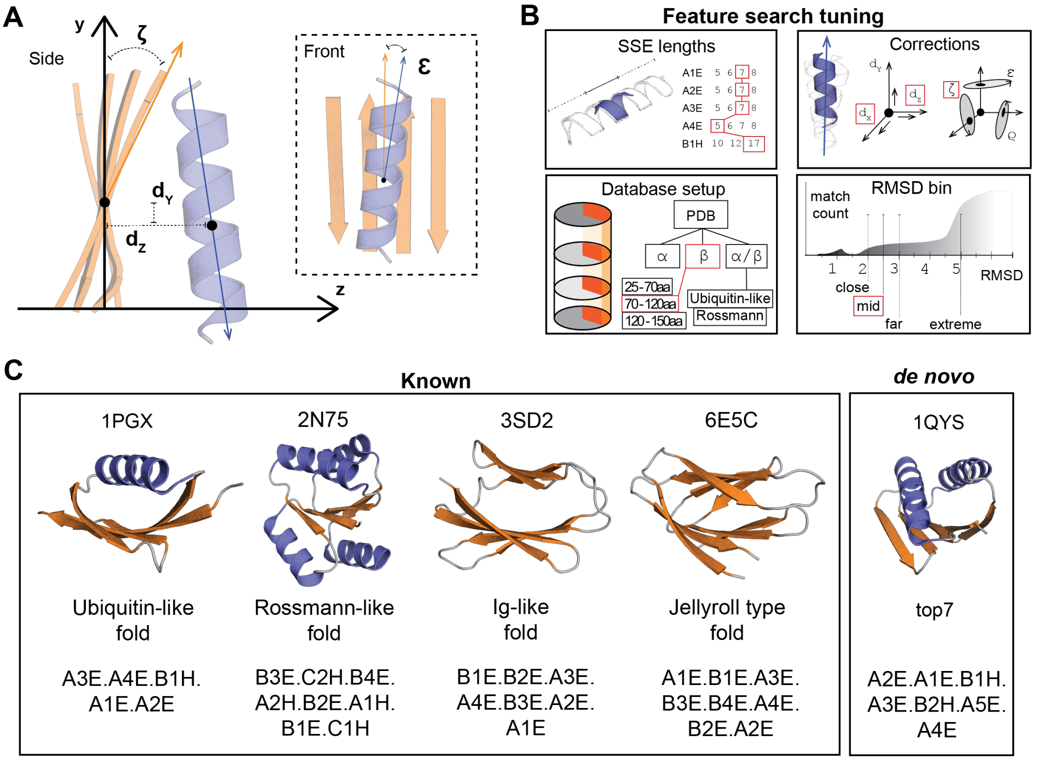


Figure 5.2: Set-up for geometric parametric correction and test examples. **A:** Example of four possible geometrical correction parameters calculated from the matches found by MASTER and exemplified on a Sketch. **B:** The correction module can be modulated, such as specifying the SSE lengths, specifically engineering databases for the MASTER searches, specifying corrections to be applied, and selecting the RMSD bin. **C:** Four known example protein structures and a *de novo* protein fold with their corresponding Form covering a variety of structural complexities.

We hypothesized that *grosso modo* parametric corrections per SSE could incorporate global native structural features and improve the Sketchs' designability (Fig. 5.2A). To do so, we implemented a module that calculates per SSE geometrical statistical corrections from native structures on-the-fly (see Methods). Briefly, the Sketch is first divided into two-layer components based on adjacent layers. These sub-structures are then iteratively matched against a database of natural protein structures using the software MASTER [309, 310] and structural geometry statistics are retrieved from the returned matches and used to correct the relative positioning of the SSEs in the Sketch. This results in a hierarchical refining procedure i.e. once a substructure is corrected, its improved geometry contextualizes the correction of the next sub-substructure.

A minimal set of parametric corrections include two rotational parameters and one translational parameter per SSE (Fig. 5.2A and Sup. Fig. 5.2). We compute the twist angle (ζ) which is the angle between the vector pointing along the length of the SSE to the plane spanned by the layer. Between two adjacent layers, we express the angle (ε) as the shear between the layer planes, and the inter-layer distance (d_Z) is the distance from one layer to the next one. The three parameters jointly shift the SSEs from a naive configuration towards a native arrangement. For example, a β-sheet is completely "flat" because the initial SSEs placed in the Sketch are fully ideal and aligned next to each other. The natural occurring twist within β-sheets can be approximated by twist angles ζ. Similarly, for small helical bundles, the twist angle $\zeta$and the shear angle εcan roughly approximate a coiled-coil [311]. Hence, the geometric corrections attempt to optimize the global arrangement of the SSEs and generate topologies with native-like features. Note that our framework allows choosing the set of parametric corrections together with other settings (Fig. 5.2B).

5.3.2 Backbone assembly and sequence design of native-like Sketches

Upon initial Sketch generation and the optimization stage that improves the native features of the SSE placements, several additional steps are necessary to obtain a well designed structure: (1) the building of loops connecting the SSEs; (2) structural diversification starting from the initial native-like Sketch; (3) sequence sampling and selection of best scoring designs. All these steps were performed using tools in the Rosetta software suite, and more details are given below and in the Methods section (Sup. Fig. 5.1A-G).

To fully compose the Sketchs' structural description, we query native loop segments that can bridge the gaps between the SSEs (see Methods). We avoid computational intensive and time-consuming loop closure algorithms by generating structural fragments (3-mers and 9-mers) using the structural information (ABEGO torsions) gathered from native loops and their anchored SSEs. We use Rosetta [67] to generate backbone conformation and perform sequence design under mid and long-range distance restraints (distances between Cα-atoms from different SSEs). Additionally, secondary structure assignments are extracted from the native-like Sketch input to the previously developed Rosetta FunFolDes (FFD) [172, 312] protocol to assemble an initial set of poly-valine backbone candidates. The structural fragments induce native backbone signatures at the local level, while the SSE arrangement is tightly controlled by the distance restraints. We modified the Rosetta energy function at every stage of the folding simulation to include H-bonds and SSE pairing terms to favor the correct pairing between β-strand [312]. Each assembled backbone is fitted with a set of optimal sequences via the Rosetta FastDesign protocol. During this stage, AA sampling restrictions per position were added, such as layer definitions (core, surface, or boundary, profiles from structural fragments) [211, 265], and secondary structure type (α-helix, loop, or β-strand) assignments. A bonus term enhancing SSE formation at defined positions was included in the energy function to enhance secondary structure formation at the desired positions [313].

5.3.3 *De novo* design of five protein folds

To showcase the TopoBuilder *de novo* design framework and assess its qualitative performance, we attempt to *de novo* design five different folds of variable structural complexities (Fig. 5.2C). We selected four native folds: a two-layered α/β Uiquitin-like fold, a two-layered β-sandwich Ig-like fold, a two-layered β-sandwich Jelly-roll, and a three-layered α/β/α Rossmann-like fold. In order to investigate the generalization capability of the framework to the space of novel folds, we include a two-layered α/β top7-like fold. Of note, the top7 structure (PDB 1QYS) was excluded from all databases used during the correction searches in order to avoid any biases coming from the solved structure. The Ubiquitin-like fold is built of a α-helix packed onto a four-stranded β-sheet. Both terminal β-strands pair in a parallel direction and are located in the center of the sheet making non-local H-bonds contacts, while the edge strands form a β-α-β-motif. The Jelly-roll and the Ig-like have both non-local β-β-motifs. Our drafted Jelly-roll has three β-arcade motifs and the Ig-like fold is made from a β-arcade on one and a long β-arch on the other side. The architecture of the intended Rossmann fold contains a four-stranded central β-sheet that is flanked by two helices on the top and two helices at the bottom and can be decomposed into three interlocked β-α-β-motifs. Lastly, the top7 fold is defined by two interlocked β-α-β-motifs with two additional terminal core strands. Each of the folds has its unique complexity with specific tertiary motifs that need to be arranged realistically with the correct geometries. Especially non-local interactions and connections have been difficult to build and design, and multiple detailed analyses were needed to discover effective design rules for single folds and domains [77, 87, 88, 314, 101].

For the generation of the backbones, we used default SSE lengths of 5 - 8 residues for β-strands, 17 residues for long α-helices, and 13 residues for short α-helices. Thus, we did not employ SSE lengths of specific native examples but rather used these as a rough guide for the overall topology. To probe the impact and contribution of the parametric corrections, we first performed baseline design simulations guided only by an uncorrected Sketch that we refer to as “naive Sketch”. We then corrected the Sketch using several combinations of parameters (termed “native-like Sketch”) in order to assess the interplay between them and find a minimal and optimal parameter mix. In the first scenario, we solely used the twist angle ζ-correction.

The second scenario consisted of the corrections ζ, d_Z, and the third scenario simulated with ζ, d_Z. ε. For each of the different scenarios, a total of 1,000 decoys was generated.

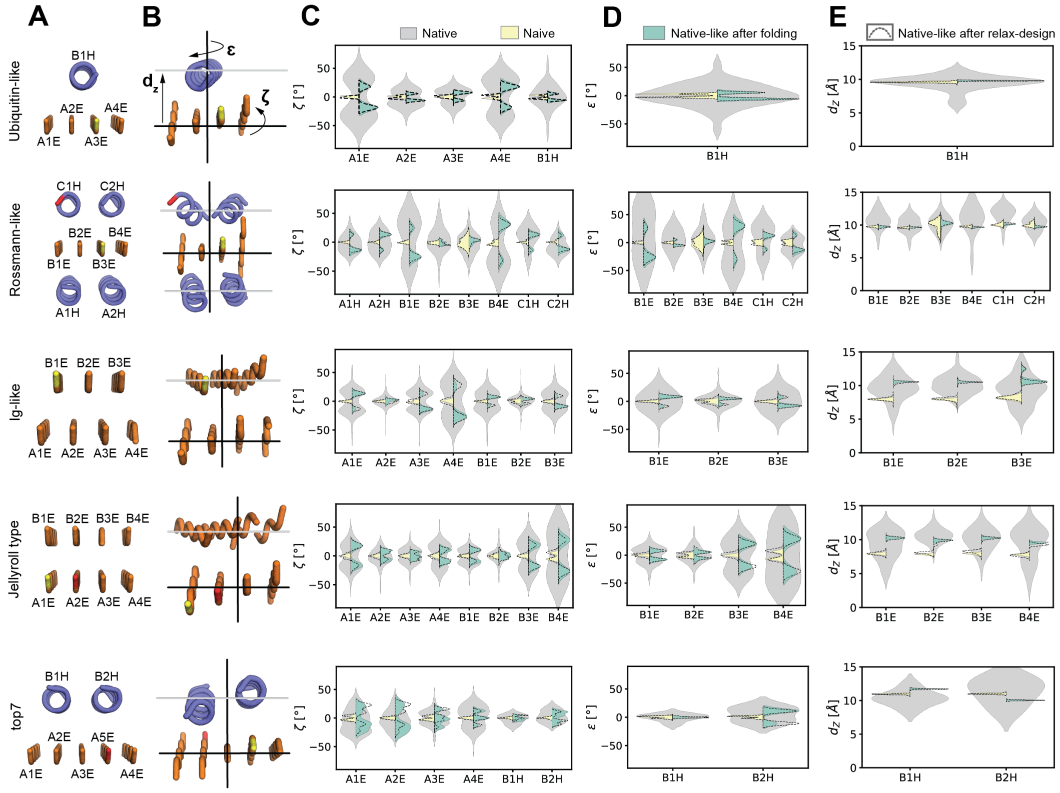


Figure 5.3: Geometric corrections. **A:** The native-like Sketches of the selected example folds with three of the geometric parameters indicated. **B, C, D:** The ζ-, ε-, and d_Z-correction parameter calculated from the matches found by MASTER. The native geometry distributions are shown in grey. In yellow, the output of a simulation (1,000 decoys) using a naive sketch without adaptation of the SSEs, which tends to result in flat β-sheets and helical packing. In green simulations (1,000 decoys) derived from a native-like Sketch where the outputs follow the native distributions.

5.3.4 Corrections induce native features in idealized folds

For the native folds, the corrections are derived from a set of ∼15 - 25 structurally distinct proteins (Sup. Fig. 5.3). The top7 fold-derived Sketch has a lower number of matches (9 distinct structures) that are extracted from larger protein domains with similar SSE dispositions and connectivities. We compared the native-like to the naive decoys for each of the five selected folds (Fig. 5.3A,B and Sup. Fig. 5.4). The ζ-angle (Fig. 5.3C) shows that the native-like decoys have distributions inducing a twist in β-strands and native side-to-side configurations for α-helices while the distributions of the naive decoys do not follow the ζ-angle geometry of natural protein structures. Rather, their ζ-angle remains around 0° indicating that no twisting has been induced during the folding and relaxation simulations. Similarly, the ε-angle (Fig. 5.3D) and the d_Z distance (Fig. 5.3E) that capture the layer packing geometry follow the natural distributions for the native-like decoys, while for the naive decoys the distributions remain similar to those of the naive sketches showing that fragment insertion protocols are insufficient to correct these overall topological features. Importantly, a relaxation without restraints after the folding and design simulations did not yield decoys with native geometries, suggesting that geometric corrections from the native-like sketches led to the generation of models and sequences that favored native-like backbone configurations.

5.3.5 Assessing designability by approximating the inherent sequence-to-structure compatibility

To assess the difference between sequences from the naively constructed and the native-like derived backbones, we first used BLASTp [315] to search for close sequence matches. However, the few hits (E-value < 0.01) found did not correspond to the intended fold based on the available AlphaFold2 (AF2) models, requiring methods with higher sequence-to-structure sensibilities to uncover the subtle differences (Sup. Fig. 5.5). Therefore, we used two orthogonal deep learning (DL) protein structure prediction engines trRosetta (trR) [106] and the recent AlphaFold2 AF2 [108] to predict structural models for all designed sequences (without MSA generation, i.e. in single-sequence input mode). The main advantages of trR and AF2 are their speed and accuracy in predicting structural models from their sequences (in the range of a couple of min./sequence), enabling the prediction of structural models for 1,000 sequences in a few hours using a computer cluster. We calculated the template modeling (TM)-scores and root-mean-square deviations (RMSDs) between our designed TopoBuilder (TB) decoy models and the trR and AF2 models (Fig 5.4A, B, Sup. Fig. 5.6). We hypothesized that, if our native-like backbones have improved designability, they could lead to sequences with stronger signatures for the respective fold and consequently lead to more accurate structure predictions with respect to the target folds in contrast to the naive Sketch-derived designs.

For the naive Sketch-derived designs, we observe low RMSDs at around ~2 Å for trR and ~1.8 Å for AF2 while simulations with the best combination of corrections achieve RMSDs of ~1.5 Å and ~1.2 Å for trR and AF2, respectively (Sup. Fig. 5.7C,D). We observe stronger pronounced tendencies for the corresponding fold-sensitive TM-scores (Sup. Fig. 5.7A,B). For the sequences and models generated with naive Sketches, the TM-scores peak around 0.7 for trR and 0.8 for AF2. The simulations with the best combination of corrections tend to TM-score one unit higher (TM-score ~0.8 for trR and ~0.9 for AF2).

Comparing the naive- and native-like derived sequences based on two metrics shows population differences (Fig. 5.4). We compare the AF2 predicted local Distance Difference Test (plDDT) versus the TM-score between the TB and the trR models (TM-score(TB,trR)). To gather the top double positive population, we set the plDDT threshold to a minimum of 60 and adjust the TM-score(TB,trR) gate. Analyzing the population difference of the double positives e.g., sequences with plDDT > 60 and high TM-score(TB,trR) shows an enrichment of the native-like derived populations of 9× for the Ubiquitin-like designs, 4× for the Rossmann

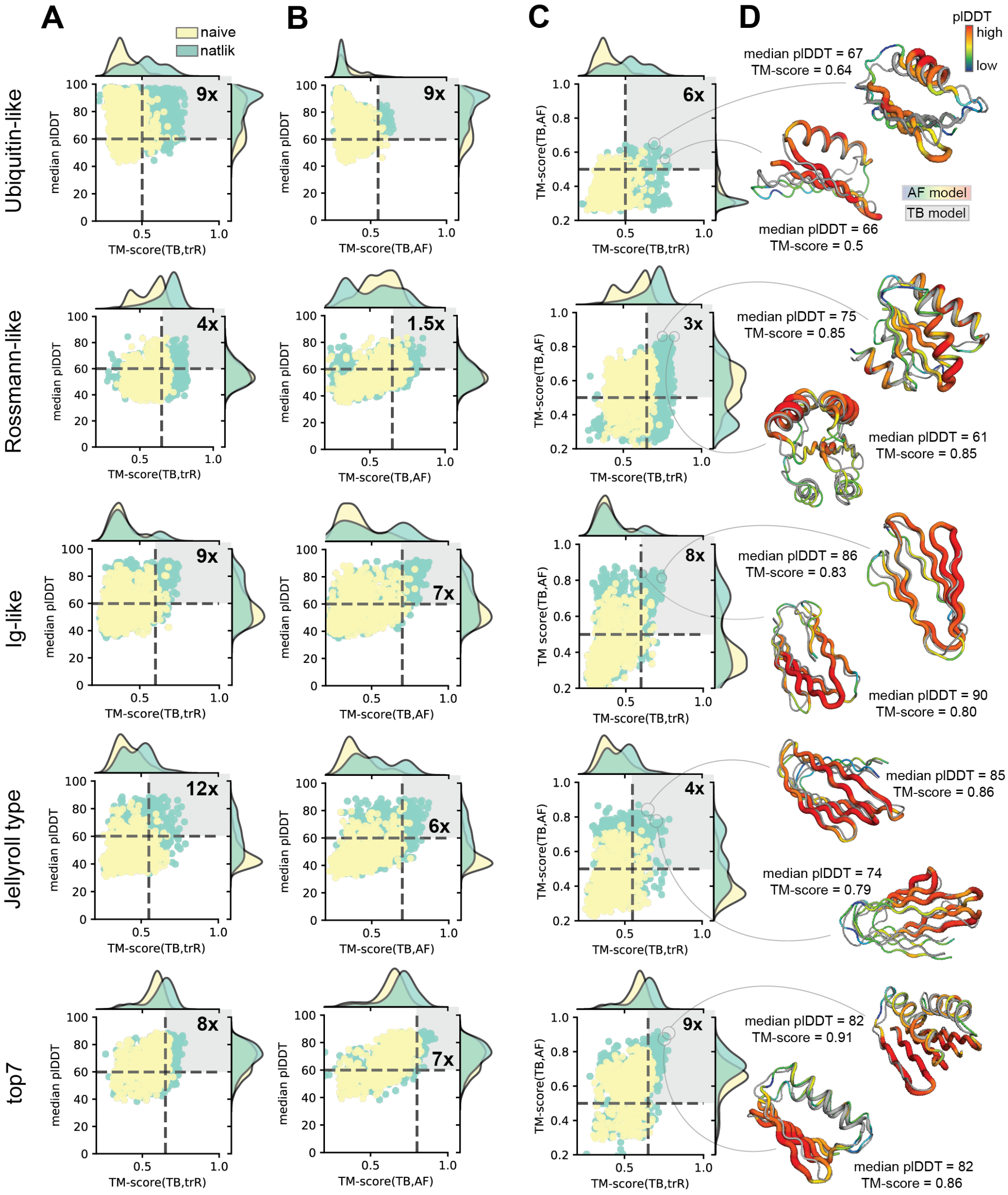


Figure 5.4: Computational assessment of sequence quality. Boost in quality for native-like derived sequences with N× more sequences in the upper right field (double positives). **A:** The plDDT scores versus the TM-scores calculated from the TB models aligned onto the predicted trR models. The plDDT threshold is fixed to 60, while the TM-score is gated in order to evaluate the double-positive populations. **B:** The plDDT scores versus the TM-scores calculated from the TB models aligned onto the predicted models from AF2. The plDDT is fixed again to 60 and the TM-score is used to select the respective populations. **C:** Comparing the TM-scores calculated between the TB models and the AF2 models or trR models. The AF2-based TM-score is fixed to 0.5 and the trR-based TM-score is changing depending on the fold to analyze the respective fractions of double-positive populations. **D:** Examples of designs of native-like designs. In color the AF2 model according to its confidence and in grey the TB model.

designs, 9× for the Ig-like designs, 12× for the Jelly-roll designs and 8× for the top7 designs (Fig. 5.4A). Similarly, we then compared the plDDT against the TM-score between the TB and the AF2 models (TM-score(TB,AF2)) and observed similar enrichments ranging from 1.5× for the Rossmann designs to 9× for the Ubiquitin-like designs (Fig. 5.4B). Lastly, we compare the TM-score(TB,AF2) versus the TM-score(TB,trR) fixing the TM-score(TB,AF2) threshold to 0.5 and adjusting the TM-score(TB,trR) gate. We see a clear boost of the native-derived double positives ranging from 3× for the Rossmann designs to 9× for the top7 designs (Fig. 5.4C). To assess the increased performance induced through the corrections, we projected the score-pairs onto their respective diagonal and calculated the receiver operating characteristic (ROC)-curve and the area under the curve (AUC) (Sup. Fig. 5.8). The ROC-AUC indicates the degree of separation between the naive- and native-derived projected distributions. Most ROC-AUC values are in the range of 0.63 - 0.70 across the three different score pairs additionally showing that our corrections improve the *de novo* design of proteins.

We argue that the better agreements between the native-like TB models and the predicted models together with high AF2 confidences result from an improved sequence-to-structure agreement that emerges from more designable backbones.

5.3.6 Experimental validation of novel sequences

We next sought to experimentally test whether the “topobuilt” proteins are folded and stable in solution. For each of the examples, we investigated the top 25 trR models by TM-score, obtained the synthetic genes for 54 designs, and expressed and purified them from *Escherichia coli* (*E. coli*) (see Methods). A total of three Ubiquitin-like, four Rossmann-like, three Ig-like, one Jelly-roll type, and two top7 like fold proteins expressed soluble (models of the designs show in Fig. 5.5A). All three Ubiquitin-like designs, two Ig-like fold designs, and the two top7 like folds had size exclusion chromatography (SEC-MALS) peaks (Sup. Fig. 5.9B) with the apparent molecular weights of the monomer or small oligomeric species. The peaks corresponding to the monomeric or small oligomeric species were examined by circular dichroism (CD) spectroscopy (Fig 5.5E). In all cases, the CD spectra were consistent with the respective target structures, with the characteristic profiles of α/β proteins. Most of the designs were thermostable, two of them with apparent melting temperatures above 50 °C and the remaining above 90 °C (Fig. 5.5E, Sup. Fig. 5.9C).

We compare the successfully expressed and characterized designs with their respective trR (Fig. 5.5A) and AF (Fig. 5.5B) models and perceive that two Rossmann, two Ig-like, and two top7 designs recapitulate the intended structures accurately, strongly indicating that our designs folded into the target shape. To evaluate the structural similarity to the natural repertoire, we searched within the PDB (Fig. 5.5C) and the AF2 (Fig. 5.5D) database for the closest protein structures. While for both, the two Rossmann-like and the two Ig-like designs we found first hits around ~4 Å RMSD, the two top7 designs are far away from any natural protein folds with the first hit ~5.5 Å RMSD.

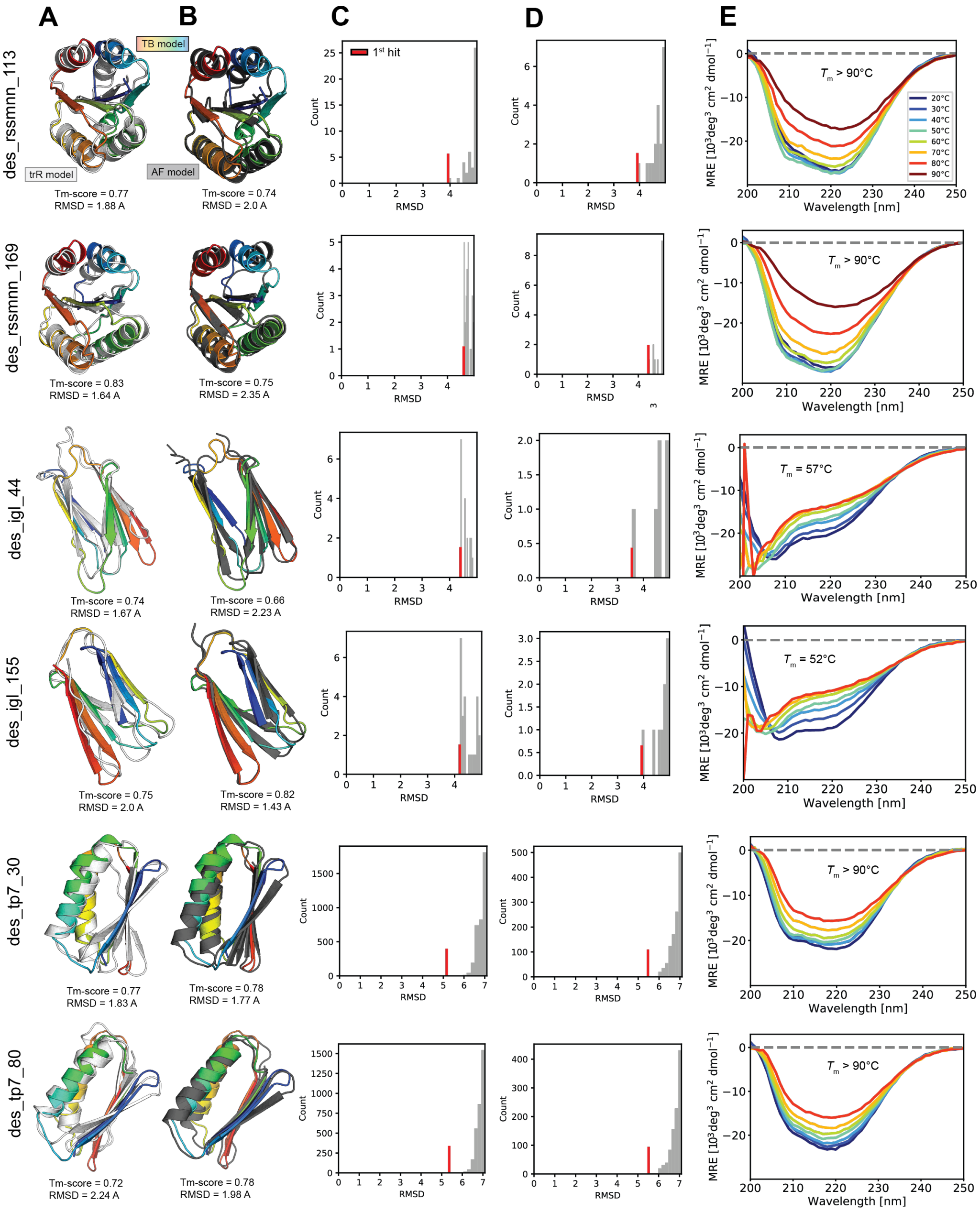


Figure 5.5: Experimental results. **A:** TB and trR models superimposed. **B:** TB and AF2 models superimposed. **C, D:** Fast RMSD-based structure search using MASTER on the PDB and AF2 databases. **E:** The far-UV CD spectra during thermal denaturation with the melting temperatures T_M obtained by fitting to the denaturation curves shown in 5.9

To reveal potential sequence families, we performed pairwise alignments, calculated BLOSUM62 distances for each alignment (i.e. the sum of each individual BLOSUM62 score), and clustered them hierarchically (Sup. Fig. 5.10A and Sup. Fig. 5.11A). Similarly, to categorize the conformations we calculated pairwise RMSDs followed by hierarchical clustering (Sup. Fig.

5.10B and Sup. Fig. 5.11B). Our designs are generally well integrated within the hierarchical cluster trees showing native compatibility. To search for close members in sequence and structure jointly, we gathered the sequence and structure features and projected the data into two dimensions through a principal component analysis (PCA) (Sup. Fig. 5.10C and Sup. Fig. 5.11C). We observe that our designs are close to native clusters, further indicating their sequence and structure nativeness. For the native folds, the matches are of the same fold family. Interestingly, des_rssmnn_113 has a *de novo* designed Rossmann fold (PDB 2LV8) and natural Rossmann domains (PDB 1MZP and 4IZ6) as cluster members, hence the design likely incorporated general native sequence and structure features. The des_tr7_30 based on the top7 *de novo* designed fold has native cluster members that structurally fit well, but have different connectivities (e.g. PDB 6NR1, 4QTP or 4QDJ).

Taken together, the data indicate that a total of six designs across three different folds adopt stable monomeric or dimeric structures with the predicted secondary structure content and correct AF2 predictions.

5.4 Discussion

The TopoBuilder *de novo* design method enables the generation of artificial proteins from a minimal Form description. The Form drafts the overall target topology and enables a fast and systematic fold-space exploration. Combined with the TopoBuilder *de novo* design framework, virtually any protein Form description can be constructed and designed.

Our computational and experimental assessments show that geometrical corrections inferred from existing layered native structural sub-motifs that compose the folds capture enough information to adapt and improve the designability of protein backbones. Minimal sets of geometric corrections lead to improved designs that have pre-defined folds and native-like characteristics [175]. When analyzing our designed sequences with state-of-the-art structure prediction tools, we identify a larger fraction of successfully recovered folds from sequences derived from corrected backbones than for sequences originating from naive backbones. The experimental characterization of multiple designs additionally shows that the TopoBuilder *de novo* design framework generates realistic designs that fold as modeled.

Ultimately, our analysis shows that the current scoring functions and fragment assembly are insufficient for the generation of designable backbones without the guidance of global correctly arranged SSEs. Recent work has focused on the discovery of fold or protein domain-specific rules through arduous analysis of natively available folds [76, 302, 312, 172], but these findings are difficult to translate to new folds. Additionally, most of the rules rely on the design of structured loops to guide the SSEs' placements. Here, we present an alternative and complementary solution that is fully automatic. Instead of focusing on structured loops, we optimize and correct the global placements of SSEs and thereby implicitly guide the loop geometries.

Our proposed strategy should enable *de novo* design to non-experts, improve and streamline future protein design efforts. The insights we gained from the parameters for a variety of complex fold examples can be harnessed and support the future discovery and understanding of protein architectural principles. Our work opens up a set of large avenues for computational protein architects and designers e.g., the scaffolding of functional proteins via incorporating known or predicted functional sites, or large complex protein machinery by assembling single *de novo* designed domains.

5.5 Methods

5.5.1 Correction module (InteractiveMaster)

We process each single MASTER match by first fitting a vector along each SSE (Sup. Fig. 5.2A,B,C). To do so, we perform a Principal Component Analysis (PCA, for more details, see [316]) overall Cα atoms within the SSE selection. Naturally, for SSE such as β-strands and α-helices, their first major eigenvector is returned by the PCA points along the SSE length, while the second is sideways and the third perpendicular.

For each full layer (including all SSE) we calculate the first three eigenvectors (Sup. Fig. 5.2D,E). This will result in a local coordinate system for the specific layer where the first eigenvector (ideally) is along the y-axis (along the lengths of the SSE), the second eigenvector the x-axis (towards the side), and the third eigenvector the z-axis. After, the first eigenplane can be interpreted as the layer (1-2) plane that is spanned by the first and second eigenvector. The 1-3 plane generated by the first and third eigenvector would splice the layer in the middle along the length of the SSE. Lastly, the 2-3 plane that can be calculated using the second and third eigenvector would half the layer roughly along the center of each SSE.

Having abstracted from atoms to simple geometric objects such as vectors and planes, multiple parameters can efficiently be calculated. Considering two adjacent layers, one can calculate various different geometric characteristics. Here, we use the ζ-angles, which are the angles between the first SSE eigenvectors and the layer (1-2) plane. The ε-angles, which are the shear angles between the two layers, can be calculated as the mean across all first SSE eigenvectors with the corresponding 1-3 eigenplane. The interlayer distance d_z can be computed as the mean across all SSE center distances to adjacent layers. Lastly, the sheer distances are calculated as the distances between the 2-3 planes to the SSE centers.

5.5.2 Loop creation (LoopMaster)

Using MASTER, we search for natural loops that can bridge the gap regions between consecutive SSEs in the sketch. For each gap, we perform independent searches with their corresponding SSE elements. The algorithm starts by iteratively matching two consecutive SSEs against a database of protein structures. For each gap, the matches are clustered based on their loop lengths, and the most populated cluster is selected (the maximum length allowed is 7 residues). Subsequently, the remaining loops are filtered with respect to their ABEGO torsion profile. Loops displaying the same classified ABEGO dihedral angles for each residue are dropped, leaving a single loop per ABEGO profile. We do not perform loop closure sampling to add the loops on the sketch as this would require computationally expensive structural sampling and is not guaranteed to effectively find a solution to close the gap effectively. Instead, directly protein structure fragments of sizes 3 and 9 (3mers and 9mers) are generated. By including their pre- and post-SSEs, we generate protein structure fragments of sizes 3 and 9 (3mers and 9mers) for the full native-like sketch, bypassing the need for sampling fragments for the SSE at a later stage.

5.5.3 The number of architectures and topologies from (multi-)Forms

To approximate the number of architectures and topologies from a Form, we assume that all SSEs elements are of the same type. Thus, if a topology consists of five SSE elements ($n = 5$), we do not differ between architectural and topological variations on the number of α-helical (H) and β-strand (E). The complete number of possible architectures A from a Form containing n SSEs can be calculated by enumerating all possible SSE placements on l layers ($l \leq n$). Given the largest lattice ($n \times n$), we would have a total of $\left(\frac{n^2}{n}\right.$ placement options. Unfortunately, this approach would first lead to many unrealistic configurations, such as low compactness and disembodied SSE elements, and second overcounts the number of architectures as it does not take into account rotational and translational variations. A more accurate prediction can be made by relating layered architectures to free polyominos (or square animals) without holes for with n cells (A000104, can be found under https://oeis.org/A000104). This enables us to simply and extremely fast look up the number of architectures for a Form with n SSE elements. The total number of topologies for one architecture can be calculated through $n!$. This will give all different ways of arranging n distinct SSEs into a sequence.

5.5.4 Protein Expression and Purification

DNA sequences of the designs were purchased from Twist Bioscience. For bacterial expression, the DNA fragments were cloned via Gibson cloning into a pET11b followed by a terminal His-tag and transformed into *E. coli* BL21(DE3). Expression was conducted in Terrific Broth supplemented with ampicillin (100 µg ml1). Cultures were inoculated at an optical density (OD) 600 of 0.1 from an overnight culture and incubated in a shaker at 37 °C and 220 r.p.m.. After reaching an OD600 of 0.6, expression was induced by the addition of 0.4 mM IPTG and cells were further incubated overnight at 20 °C. Cells were harvested by centrifugation and pellets were resuspended in lysis buffer (50 mM TRIS, pH 7.5, 500 mM NaCl, 5% glycerol, 1 mg ml^{-1} lysozyme, 1 mM PMSF, 4 µg ml1 DNase). Resuspended cells were sonicated and clarified by centrifugation. Ni-NTA purification of sterile-filtered (0.22 µm) supernatant was performed using a 5-ml His-Trap FF column on an ÄKTA pure system (GE Healthcare). Bound proteins were eluted using an imidazole concentration of 500 mM. Concentrated proteins were further purified by size exclusion chromatography on a Hiload 16/600 Superdex 75 pg column (GE Healthcare) using PBS buffer (pH 7.4) as mobile phase.

5.5.5 Circular dichroism spectroscopy

Far-UV circular dichroism spectra were collected between wavelengths of 190 and 250 nm on a Jasco J-815 circular dichroism spectrometer in a 1-mm path-length quartz cuvette. Proteins were diluted in 10 mM Phosphate-buffered saline (PBS) at concentrations between 20 and 40 µM. Wavelength spectra were averaged from two scans with a scanning speed of 20 nm min^{-1} and a response time of 0.125 s. The thermal denaturation curves were collected by measuring the change in ellipticity at 220 nm from 20 to 90 °C with 2 or 5 °C increments.

5.5.6 Size-exclusion chromatography combined with multi-angle light scattering

Multi-angle light scattering was used to assess the monodispersity and molecular weight of the proteins. Samples containing 80–100 µg of protein in PBS buffer (pH 7.4) were injected into a Superdex 75 10/300 GL column (GE Healthcare) using an HPLC system (Ultimate 3000, Thermo Scientific) at a flow rate of 0.5 ml min^{-1} coupled in-line to a multi-angle light-scattering device (miniDAWN TREOS, Wyatt). Static light-scattering signal was recorded from three different scattering angles. The scatter data were analyzed by ASTRA software (version 6.1, Wyatt).

5.6 Supplementary information

5.6.1 Supplementary methods

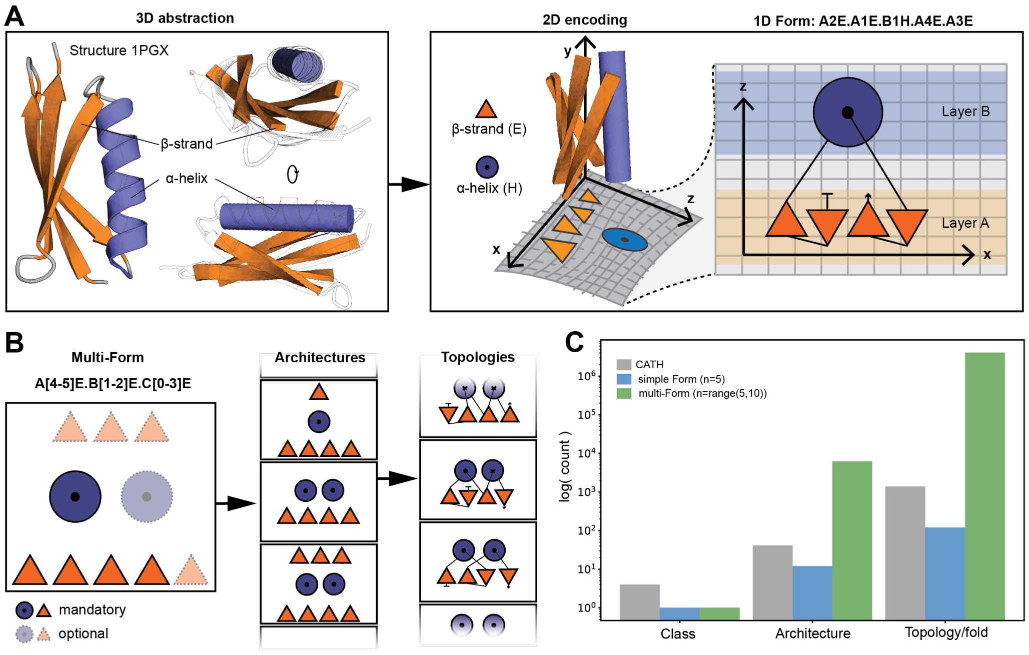
MASTER matching

We use the MASTER software to rapidly search for matches within a specified RMSD cutoff over the large protein structure databases. The method returns multiple alignments within the RMSD threshold where we only use the best alignment per protein structure for the calculations of the parametric corrections. We also use MASTER to search for potential loops between by using the segment feature of the MASTER software.

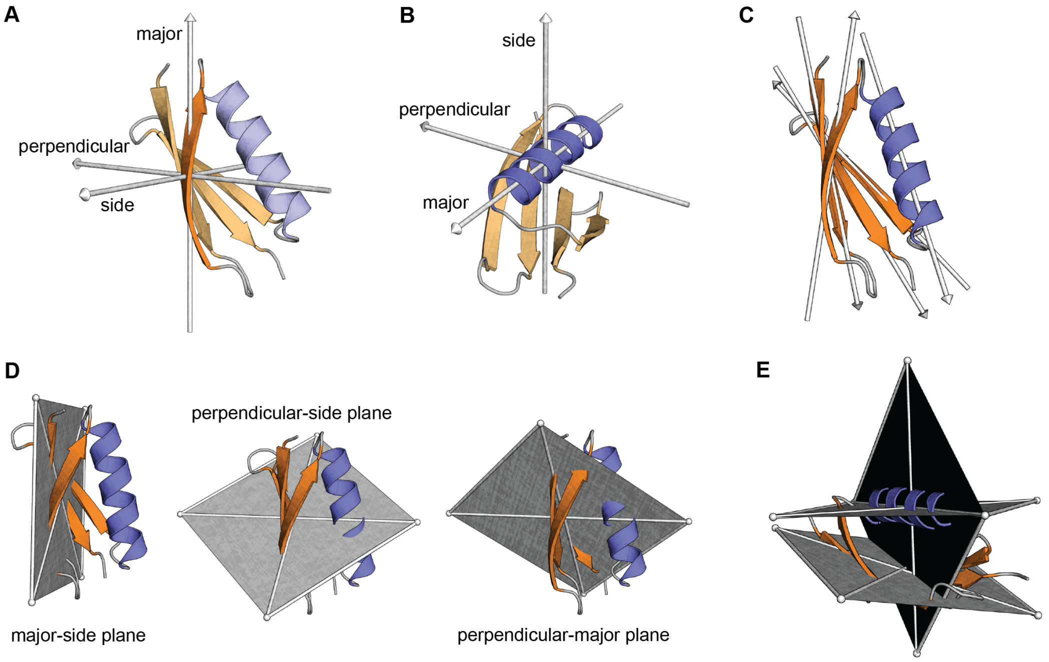
Protein-protein alignment by TM-align

For each alignment, TM-align optimizes and reports TM-score, a measure of the distance between Cα carbons of aligned residues in target and template, normalized by protein length. The optimization algorithm used by TM-align results in alignments where the superposition of segments with similar local structures is optimized over the superposition of segments with disparate local structures.

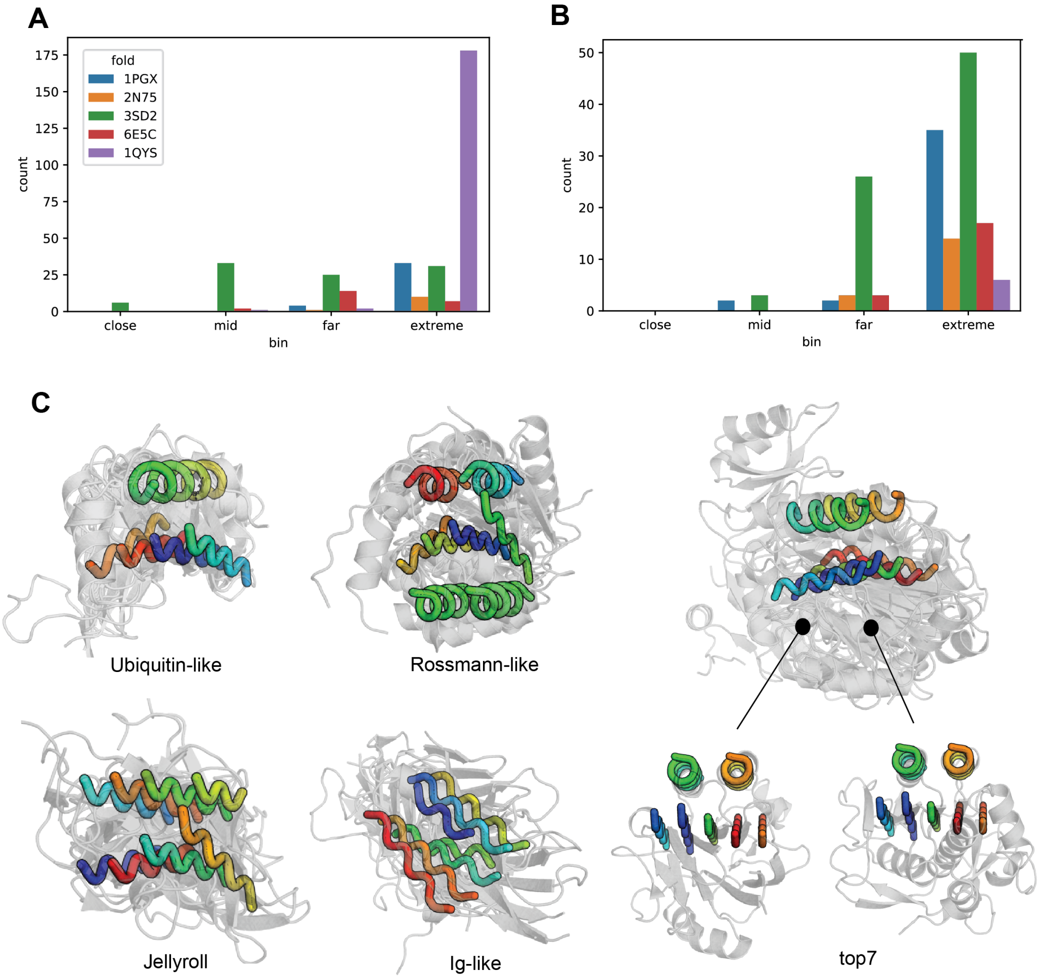
5.6.2 Supplementary figures



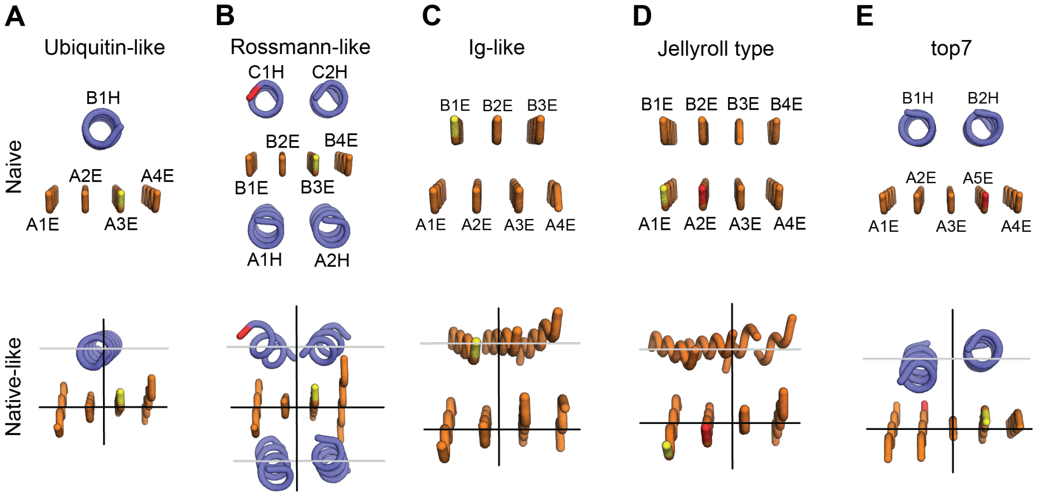
Supplementary Figure 5.1: Overview of the modules in the TopoBuilder *de novo* design framework. The TopoBuilder *de novo* design framework can easily be extended and modified by adding custom modules with the provided template coded in Python. **A:** The generic pipeline takes a Form string as input and generates a 3D expansion of the architecture or topology sketch. **B:** If the topology is not specified, all possible topologies are generated for the user to select a single topology from. **C:** At the heart of the pipeline, the correciton module (InteractiveMaster) protocol searches a database of natural protein structures for simple geometric features to recover a native-like configuration of the sketch. **D:** A provided module enables the incorporation of structural motifs. **E:** The loops are reconstructed implicitly via first searching for natural loops that are roughly capable of bridging the gap and secondly creating structure fragments from the found SSE and their loops. **F:** Finally, the structural model is built via Rosetta fragment assembly (FFD) and designed using the FastDesign method. **G:** The generic TopoBuilder pipeline starting from a Form that expands into a naive Sketch, which is then corrected into a native-like Sketch. Loops are searched able to connect the SSEs and fragments of size 3 and 9 are created (3mers and 9mers). Using the fragments and distance constraints form the native-like Sketch, N poly-valine backbones are assembled and NxM sequences designed.



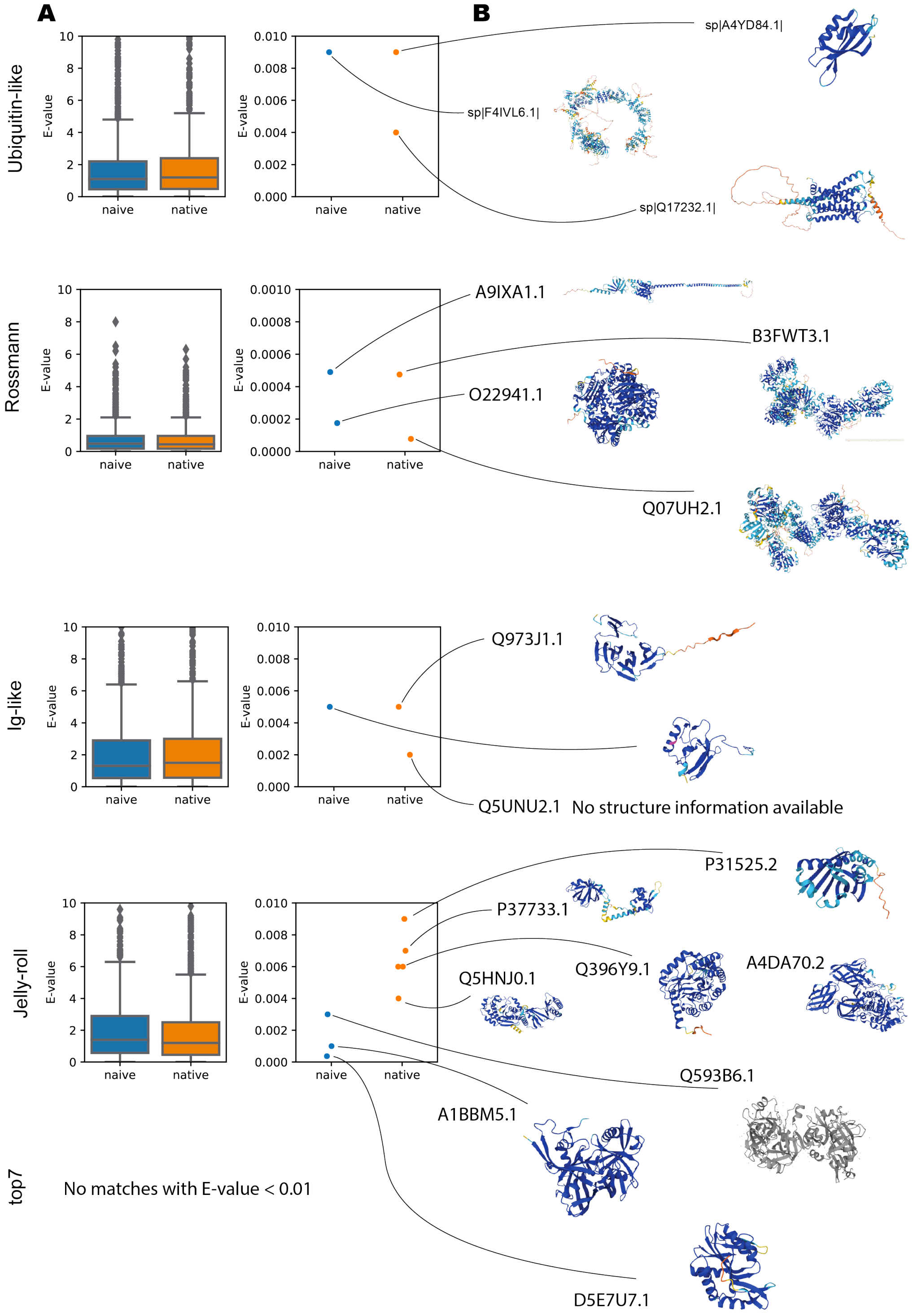
Supplementary Figure 5.2: Ubiquitin-like fold geometric analysis. **A:** The eigenvectors calculated for a single β-strand of the PDB 1PGX structure. The major eigenvector will always point along the length of the SSE. **B:** The eigenvectors calculated for a single α-helix of the 1PGX structure. The major eigenvector is along with the helical pitch. **C:** All major eigenvectors for each of the SSE of the 1PGX structure. The positions and directions of the SSE can be described by the major eigenvectors. **D:** The set of the eigenplanes calculated for the β-sheet (layer A) 1PGX structure. **E:** The major-side planes for the sheet and the helix of the 1PGX structure.



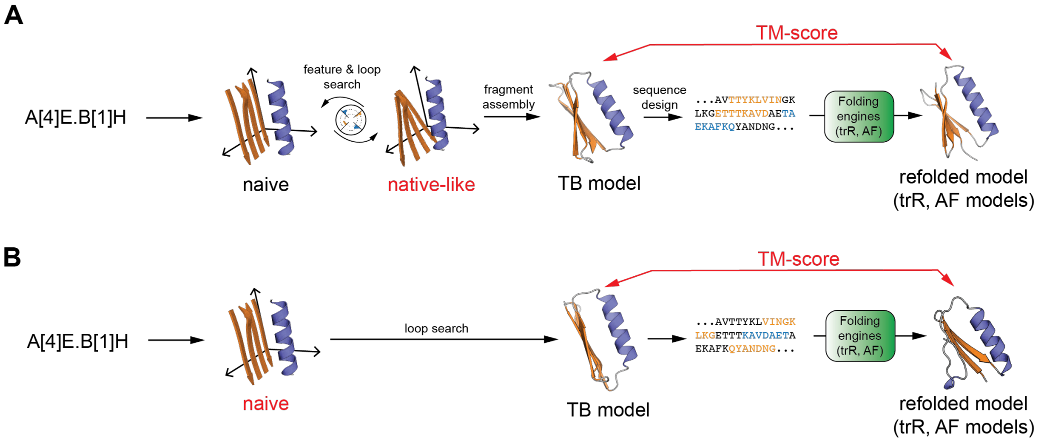
Supplementary Figure 5.3: Counts and examples for the correction search. **A:** Counts of the retrieved matches per RMSD-bin for the first layer only (first step corrections). As only a single layer is searched more matches are retrieved. **B:** Counts of the retrieved matches per RMSD-bin for the second layer (second step corrections). Here, two layers are included making the matching more stringent and less structures were retrieved. **C:** Examples of retrieved matches (second step corrections) to derive the geometrical corrections.



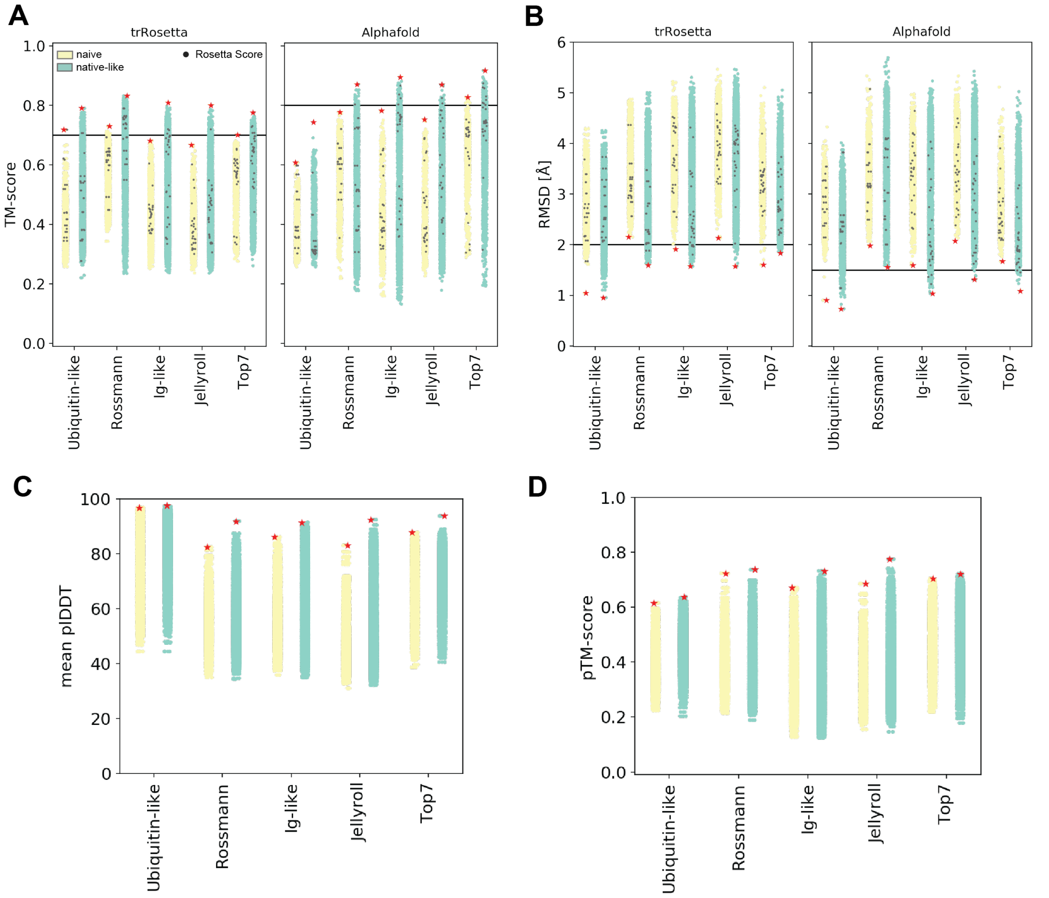
Supplementary Figure 5.4: Comparisons of naive and native-like Sketches. A front-view of naive and native-like Sketches. **A, B, C, D, E:** The naive Sketches (top row) with their respective Form element labeled for each of the five folds. The lower row shows the native-like Sketches with corrected SSE.



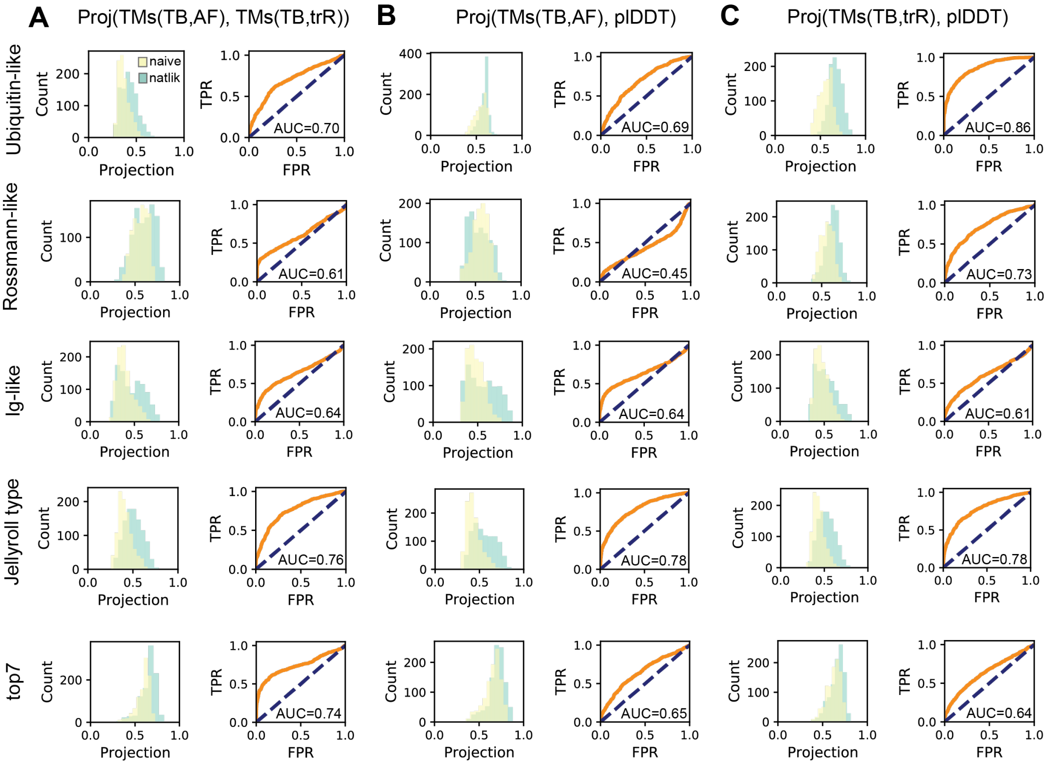
Supplementary Figure 5.5: BLAST search against SwissProt. **A:** Best E-values of the BLAST searches against the SwissProt database for each design from a target fold. **B:** Available AF2 models (colored) or solved structures (grey). None of the models or structures represents the target fold.



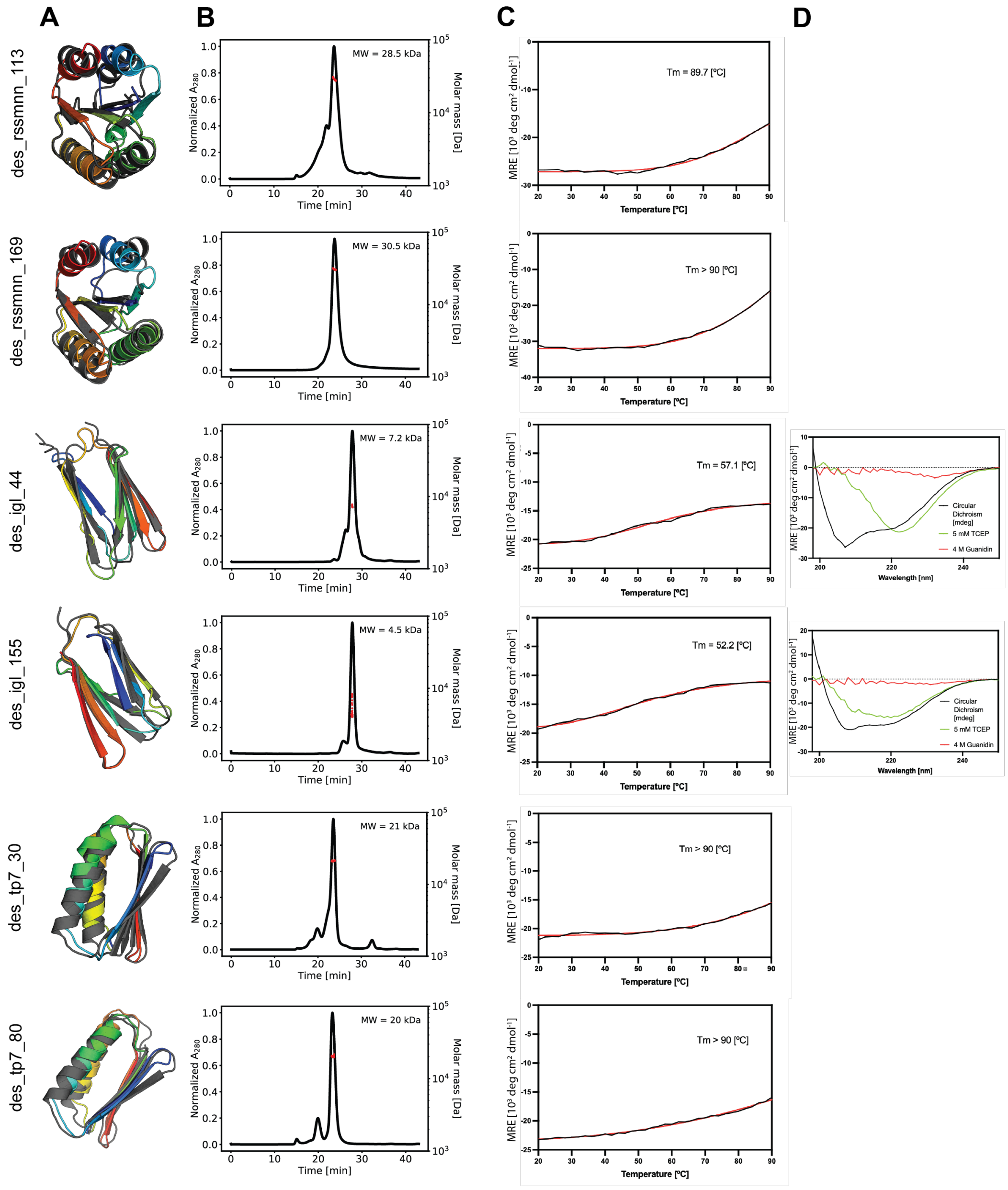
Supplementary Figure 5.6: TopoBuilder *de novo* design pipelines. **A:** The TopoBuilder *de novo* design pipeline including the feature search (corrections) rendering a naively created Sketch more native-like and designable. A full atomistic model is generated through the Rosetta fragment assembly protocol (FFD) and sequences are designed using the Rosetta FastDesign method. To evaluate if our sequences encode the necessary fold determining signatures strongly enough, we use state-of-the-art protein structure prediction engines (trR and AF2) to computationally predict a model structure that we then structurally compare to our TB model by calculating the TM-score. **B:** The TopoBuilder *de novo* design pipeline where the feature search module has been ablated.



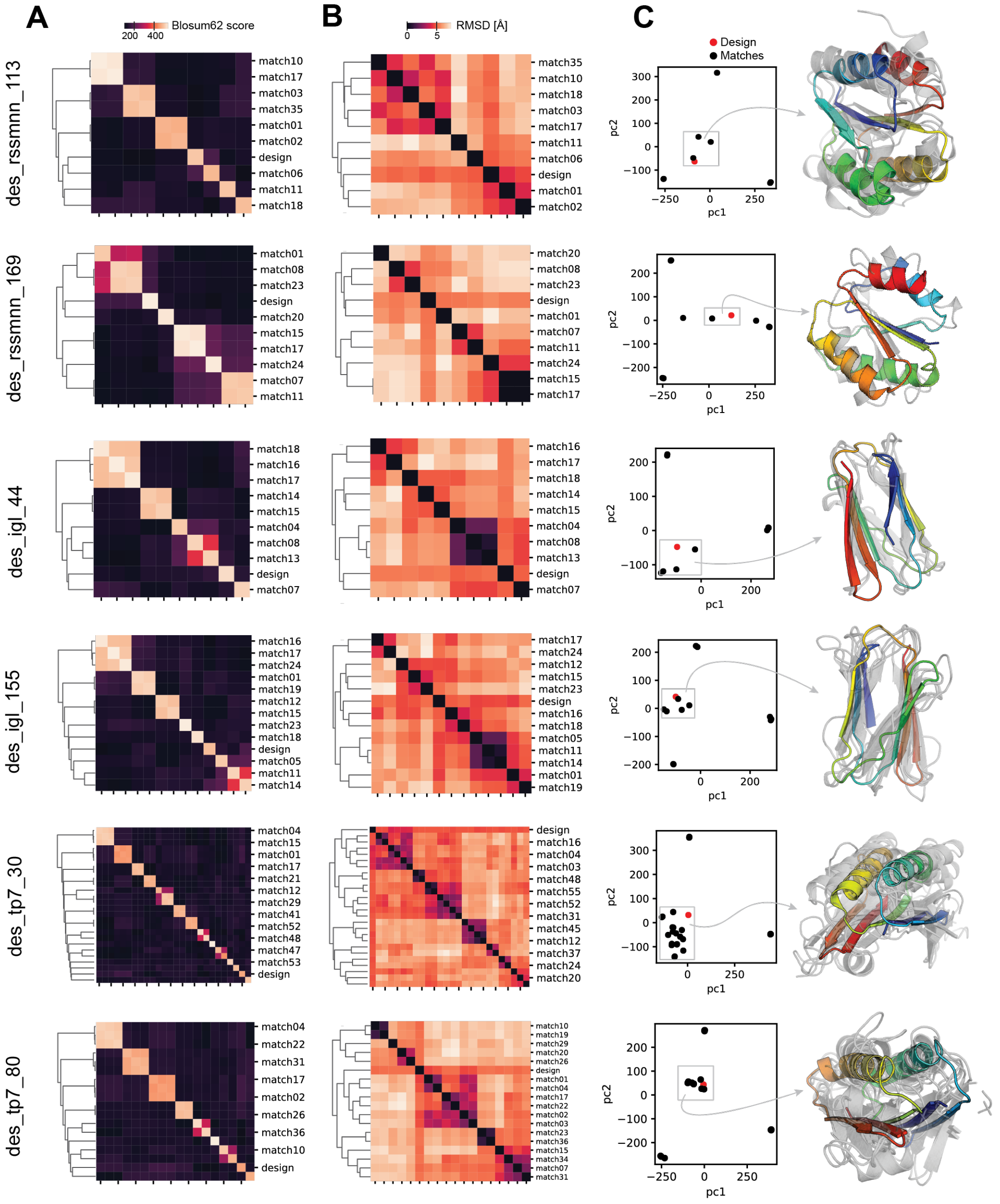
Supplementary Figure 5.7: Computational assessment of sequence quality. Th **A:** Pairwise TM-scores calculated between the TB and the trR or AF models (single sequence input without MSA) for each set of simulations. The black dots represent the 25 lowest scoring decoys by Rosetta energy. The red star indicates the best scoring decoy (either by TM-score or RMSD) for each of the simulations. **B:** RMSD using the best superposition and residue coverage between the TB and the trR or AF models. **C, D:** The mean plDDT and pTM-scores predicted by AF for each decoy.



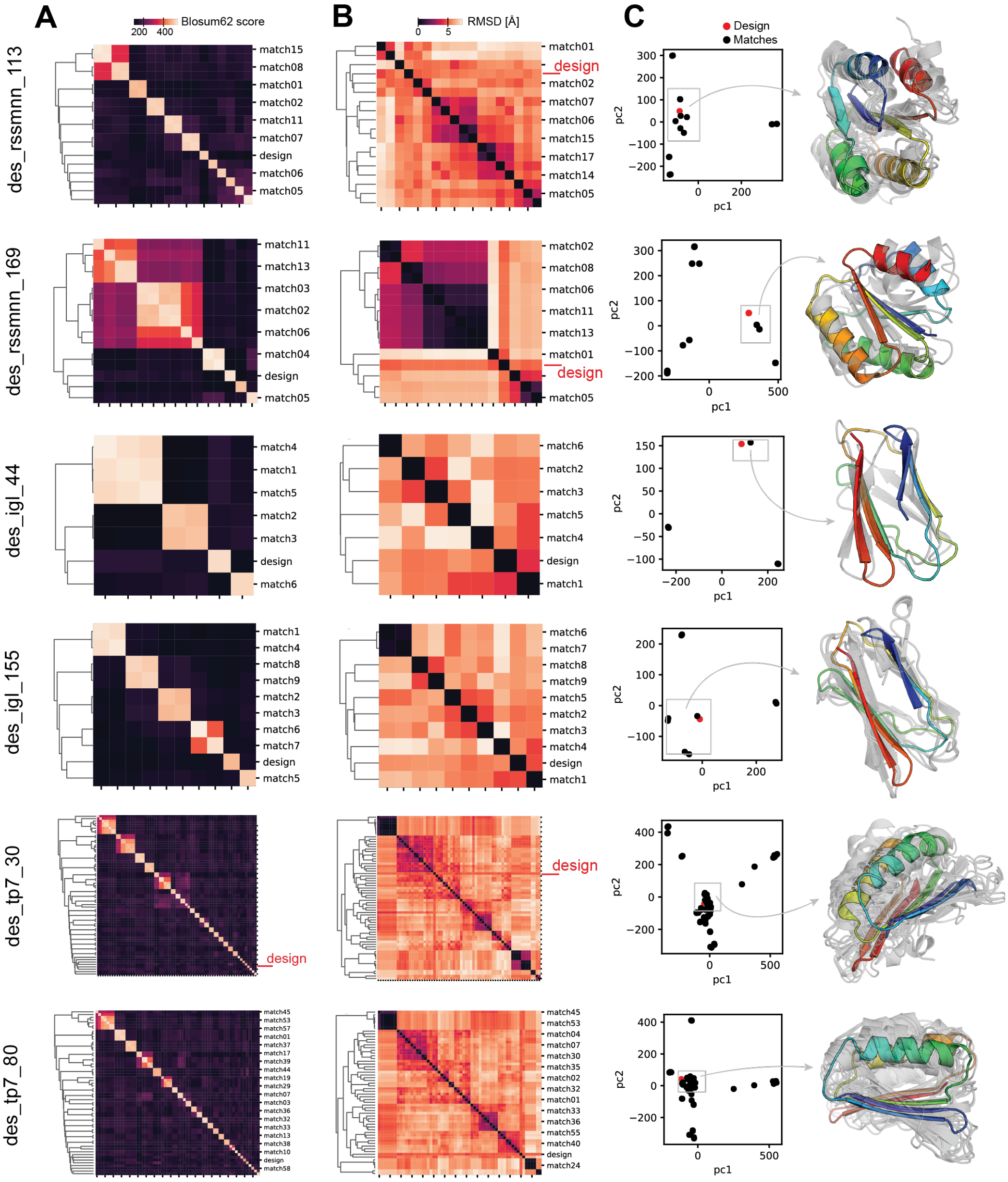
Supplementary Figure 5.8: ROC curves and ROC-AUCs. The double-positive scores are projected onto the diagonal. The distributions are shown on the left, and the ROC curve and the ROC-AUC are shown on the right indicating the strength of separation between the naive- and native-derived decoys. **A:** Projection onto diagonal of the TM-score(TB,AF), TM-score(TB,trR) scores. **B:** Projection onto diagonal of the TM-score(TB,AF), plDDT scores. **C:** Projection onto diagonal of the TM-score(TB,trR), plDDT scores.



Supplementary Figure 5.9: SEC-MALS and thermal meltings for the designs. **A:** Models of the designs (rainbow) with their corresponding AF2 predictions (black). **B:** SEC-MALS profiles for the designs show that both des_igl_44 and des_igl_155 are mostly monomeric, while the other designs show signals and molecular weights (MW) indicating oligomeric species (2mers, 3mers, and 4mers). **C:** Thermal denaturation curves at 220 nm for des_rssmnn_113, des_rssmnn_169, des_tp7_30 and des_tp7_80 and at 218 nm for des_igl_44 and des_igl_155 are shown. **D:** CD spectra collected with 5 mM of the reductant tris(2-carboxyethyl)phosphine (TCEP, green) and 4 M of the chemical denaturation guanidinium hydrochloride (GdnHCl, red) for des_igl$_4$4$anddes_igl_1$55.



Supplementary Figure 5.10: Sequence and structure similarities to the PDB. **A:** Global pairwise sequence alignments scored with a BLOSUM62 distance (sum of the individual BLOSUM62 scores). **B:** Pairwise structure RMSDs. **C:** Principal component analysis (PCA) of the combined sequence (BLOSUM62 distance) and structure (RMSDs) features. Models of the designs (rainbow) with their cluster members.



Supplementary Figure 5.11: Sequence and structure similarities to the AF database. **A:** Global pairwise sequence alignments scored with a BLOSUM62 distance (sum of the individual BLOSUM62 scores). **B:** Pairwise structure RMSDs. **C:** Principal component analysis (PCA) of the combined sequence (BLOSUM62 distance) and structure (RMSDs) features. Models of the designs (rainbow) with their cluster members.

6 Tailored *de novo* protein design with deep neural networks

...

This chapter describes ongoing work for a *de novo* protein design method employing deep neural network (DNN) modules. As described in chapter 5, we use idealized drafts of protein folds and train a DNN termed "Genesis" to generate "designable" structural constraints thereof. trRosetta is used to design sequences obeying the predicted constraints. Together, the Genesis-trRosetta framework circumvents the creation of backbones in 3D space, and compared to the TopoBuilder, Genesis-trRosetta can be used to virtually design any protein fold while being orders of magnitudes faster. A manuscript based on this work is currently in preparation.

Preliminary authors
Zander Harteveld[1,2*], Joshua Southern[3], Michëal Defferrard[1], Pierre Vandergheynst[1], Michael M. Bronstein[3], Andreas Loukas[1], and Bruno E. Correia[1,2]

[*] These authors contributed equally.

Affiliations
[1] École Polytechnique Fédérale de Lausanne, Lausanne, CH. [2] Swiss Institute of Bioinformatics (SIB), Lausanne, CH. [3] Imperial College London, London, UK.

Author contributions

B.E.C. and Z.H. conceived the initial idea and refined it together with A.L., M.M.B, M.D., J.S., Z.H. and B.E.C.. Z.H. generated the data and wrote the software. B.E.C. directed the work. Z.H. wrote the manuscript with support from all authors.

6.1 Abstract

De novo protein design aims to explore uncharted areas of the global protein structure and sequence spaces. Despite recent advances, the success of *de novo* design remains limited. One of the main challenges is defining the set of “designable” structural protein backbone templates which could in turn help to solve the sequence sampling stages. Many protein backbone design approaches suffer from inaccuracies in both energy functions and sampling algorithms, which often leads to a convergence in sequence space with a few similar sequence variants that frequently fail experimentally. To address these limitations, we build on recent advancements in protein structure prediction and design using deep neural networks. We train a convolutional variational autoencoder called Genesis that is trained to improve protein secondary structure lattice models termed Sketches by denoising their 2D feature maps. Genesis interfaces with the recent neural sequence design framework trRosetta to jointly optimize the protein sequence and structure in a higher dimensional feature space, thereby bypassing the arduous and time-consuming step of crafting designable backbones and fitting sequences in 3D space. We used Genesis-trRosetta to design large pools of diverse sequences for a set of protein folds and found that the framework is capable of sampling native-like features maps for known and novel protein topologies. The Genesis framework enables the exploration of the protein sequence and fold space within minutes and is not bound to specific protein topologies. Essentially, our method addresses the backbone designability problem and could ultimately contribute to the *de novo* design of proteins with new functions.

6.2 Main

Evolution is a slow and gradual process that has only sampled a tiny fraction of the possible protein amino acid (AA) sequence space [317, 318]. Natural sequences collapse into structures that can be clustered into a small set of protein shapes (folds). In order to explore new sequences that fold into well-defined 3-dimensional (3D) conformations outside the natural repertoire and are amenable to tailored functionalities, *de novo* protein design strategies have been developed [42, 319, 291, 320]. Currently, *de novo* protein design is an iterative process where (1) the protein shape is outlined and corresponding backbones are sampled, and (2) low energy AA sequences are fitted onto the generated backbones. Despite numerous successes [295, 60, 296, 85, 76], *de novo* design is hindered by inaccuracies in current energy functions and the heuristics within most sampling methods, often leading to experimental failures [321, 77, 74].

Many designed proteins fail due to having physically unrealistic backbones that are not “designable” in the first place. Designable backbones are strain-less and have optimal secondary structure configurations with favored tertiary structure symmetries such that they are realizable with the 20 natural AA and induce packed conformations [94, 89, 90, 95, 92]. Furthermore, it has been observed that highly designable backbones can accommodate a large variety of energetically favorable sequences [92, 94]. Large sequence capacity has been linked to mutational robustness which favors thermodynamic and evolutionary stability [322, 323]. Capturing designability quantitatively is challenging as it includes properties that are difficult to measure, such as fold specificity [89, 298], or native-like structural arrangements [97].

Multiple empirically derived principles have been formulated to encode strong sequence-to-structure relationships for protein structures and to alleviate the designability problem to a certain extent. For highly symmetric and repetitive folds such as α-helical bundles, parametric equations describe the shapes with a minimal number of variables [324, 303]. Koga and colleagues [99] formulated the first set of rules that, together with fragment assembly protocols [84], led to the design of "ideal" protein folds, this is with small loops and regular secondary structure elements (SSEs). The rules are based on loop lengths that embed the packing of local tertiary motifs such as β/β-, β/α-, and α/β-units to SSEs. These rules have been steadily updated. For instance, loops can be structurally defined to bridge non-local motifs [100, 87], cavities can be created by inducing strong curvatures into β-strands through controlling resisters shifts between- and β-bulges in -strands [76, 86], and strategically placed stress-relieving glycine residues allow the design of β-barrels [77]. Furthermore, methods that automatically identify fold-specific statistics from structural data have been shown to improve the design process [77, 64]. The structure extension with native-substructure graphs (SEWING) [85] method enables designing proteins with non-ideal SSE, where natural SSE with all their irregularities are pieced together.

Recent advances within deep neural networks (DNNs) combined with the availability of large-scale protein structure data in the protein data bank (PDB) have enabled highly accurate structure prediction from sequence [107, 108] (Fig. 6.1A). Interestingly, the trained structure prediction DNN can be "reversed" for the protein design task. A good example is the transform-restrained Rosetta (trRosetta) neural network [106] that was used to hallucinate novel proteins by using a specific loss that maximizes the contrast between random (background) and native distance predictions [138]. trRosetta can also be employed for fixed backbone design via backpropagating gradients from the target structure to the sequence, which has the effect of implicitly optimizing over the full sequence and structure landscape [139] (Fig. 6.1B). In the latter case, the method searches for the lowest-energy sequence while maximizing the probability of the target structure relative to all other conformations. Encouragingly, the trRosetta design framework is able to design new sequences for a target structure within minutes on modern graphical processing units (GPUs), enabling multi-state and high-throughput sampling of the design space.

Inspired by these recent advances, this work puts forth the hypothesis that trRosetta can also facilitate tailored *de novo* design, where the shape and SSE composition is controlled (Fig. 6.1C). To achieve this, we implemented a framework that uses a simple string description of a protein fold (termed "Form" [305]) and auto-generates realistic designed proteins. The framework first creates a 3D representation of the Form termed "Sketch". We trained a variational autoencoder (VAE) termed Genesis to encode the distances and orientations of a Sketch to a latent representation, sampled and then decoded close-native distance and

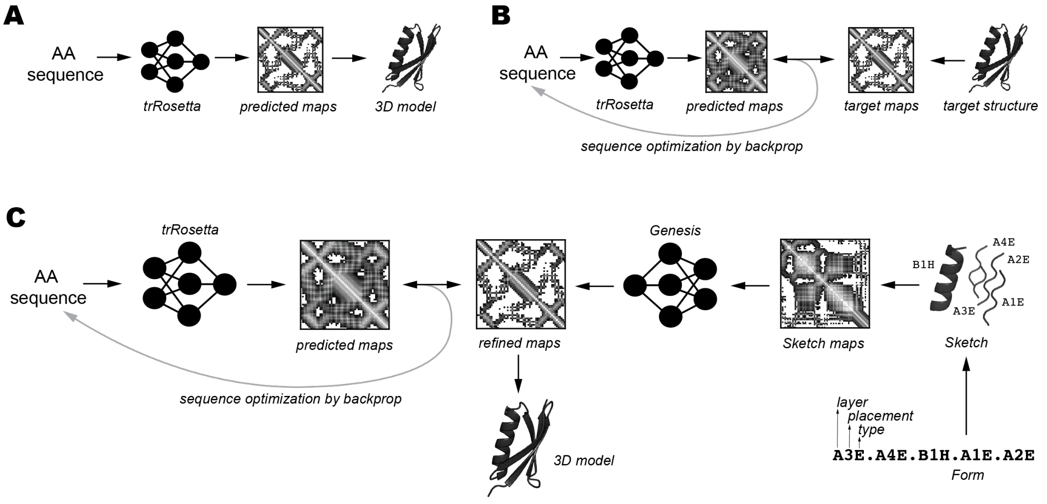


Figure 6.1: Genesis neural *de novo* protein design pipeline. **A:** The general trRosetta sequence to structure prediction pipeline. **B:** The trRosetta framework used for fixed backbone design maximizing the predicted probabilities towards the given target contacts. **C:** Our Genesis-trRosetta framework for *de novo* protein design. We use the trRosetta fixed backbone design strategy to design a sequence for the refined contacts from the roughly sketched contact of a protein fold through the Genesis DNN.

orientation probabilities from this representation ready for the trRosetta sequence design task. Our approach circumvents the need to create designable 3D backbones and the backbone generation task is, unlike conventional *de novo* design methods, not directly based on energy functions. This allows the *de novo* design process to be fast and efficient.

6.3 Results

6.3.1 Sketching protein drafts from native backbones

We delineate the shape of a protein through a string specifying the SSE types, lengths, and relative positions on a lattice, termed Form [305]. In a Form, each level or layer of the lattice can be populated by an arbitrary number of SSEs. The layers are equally spaced from each other by 8 Å for β-β-strand layers and 10 Å - 11 Å for α-α-helix or α-helix-β-layers [307]. A Form can be expanded into a 3D representation that we call a "Sketch". A Sketch is a rough 3D approximation of a native protein structure albeit lacking loops, native-like irregularities within SSEs, and side chains (or sequence).

We have previously described the TopoBuilder [174, 175], a method that can design *de novo* protein models and sequences given a Form or a Sketch using Rosetta FunFolDes (FFD) [312] and FastDesign with derived Cα-Cα distance restraints from the SSEs of the Sketch. This method yielded several successes [174, 175], although two major limitations remain: (1) the inherent non-designability of the Sketches and hence the imprecisely derived Cα-Cα restraints often guide the simulations towards incorrect solutions, and (2) the number of designs that are needed to be sampled (~2,000 - 10,000 sequences) in order to sufficiently probe the sequence landscape and be able to select potential low energy solutions is extremely high and therefore time and resource consuming.

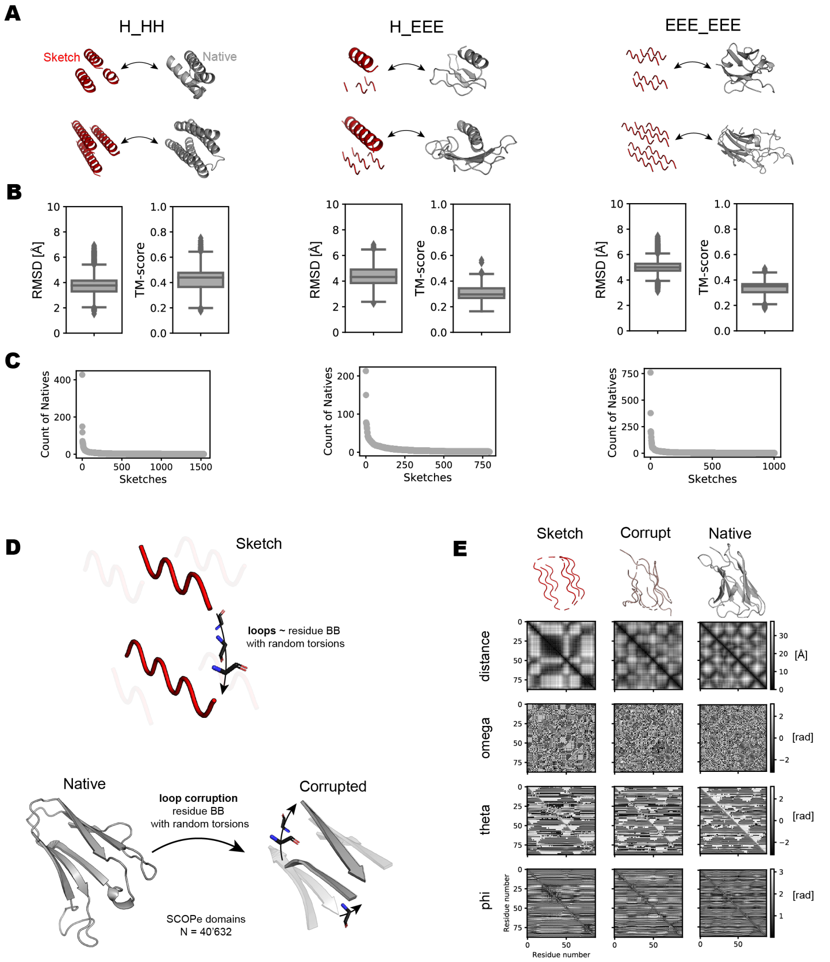


Figure 6.2: Network architecture and data engineering. **A:** Examples of Sketches (red) and a corresponding native structure (grey) for the major protein structure classes (HH_H: 3-helix bundle, H_EEE: mixed-α/β-sandwich, EEE_EEE: β-sandwich). **B:** Similarities between the Sketches and their corresponding native structures based on best-fit root-mean-squared deviation (RMSD) and the TM-score for major protein structure classes. **C:** The number of native structures that can be represented by an individual Sketch across the three major protein structure classes. **D:** Loops on the Sketch and corrupted structure are approximated by adding backbone residue atoms with random torsions along the shortest path between two consecutive SSEs. **E:** Comparison of the different feature maps (Sketch, corrupted Structures, and native Structure).

To alleviate these limitations, we employ a DNN to automatically learn to decipher important structural features and incorporate native-like patterns into the Sketches. The DNN takes the form of a VAE that is trained to transform a large dataset of Sketches into their respective native structures. The dataset was built by generating different sets of “mini-Sketches” and mapping them to their native counterparts (Fig. 6.2A). The sets encompass many 2- and 3-layer fully-β-,

α- and mixed-α/β topologies and capture a large scope of possible folds (see Methods). The mini-Sketches have small idealized SSEs (5 residues for β-strands and 9 residues for α-helices), no sequence information, and dummy backbone residues along the shortest path between end- and starting points of the SSEs representing the loops. The loops were modeled in this way because we do not have information about potential loop conformations (Fig. 6.2D). We also note that, though mini-Sketches can fit onto multiple native counterparts, the majority only map to 1 or 2 conformations (Fig. 6.2C).

Since not all protein domains can be formulated as a Form (e.g., β-barrels), we augment our data set by adding corrupted backbone structures, where the loops are replaced by dummy residues as done on the mini-Sketches (Fig. 6.2D). The corrupted backbones add architectural and structural diversity to our data set by retaining tertiary motif dispositions and secondary structure irregularities, respectively (Fig. 6.2E). We split our dataset into a series of training and test sets with different structural properties based on the Structural Classification of Proteins — extended (SCOPe) scheme. We then optimize Genesis for a training set consisting of a particular type of structure and test the validity of its predictions on proteins with increasingly different structural properties (see Methods). Our evaluation procedure quantitatively assesses the extent to which our framework generalizes beyond the distribution induced by a given training set. We argue that any method that facilitates *de novo* design should generalize to unknown subsets of the protein space.

6.3.2 *De novo* design through collaborative deep neural networks

We developed a convolutional VAE that operates on distance and orientation maps rather than atomistic coordinates (Fig. 6.3A) (See Methods for implementation details and data encoding). Importantly, distances and orientations are invariant with respect to translation and rotation which ensures stable and predictable performance in the presence of transformations of the data input under the special Euclidean group. Our VAE is conditioned on the real-valued distances and orientations of the mini-Sketches, and from the latent conditional distribution predicts distance- and orientation probabilities of native-like conformations.

We train the VAE in a supervised manner by minimizing the 1^{st} Wasserstein distance between the true feature maps and the distribution predicted by the VAE (see Methods). In contrast to the previously utilized cross-entropy loss [105, 106], the Wasserstein distance enables weighting individual errors between the distributions, i.e., penalizing large differences between the true and predicted distributions more than small differences. We follow a standard pre-train - fine-tune regimen. We pre-train the VAE on the corrupted structures with a learning rate set to 0.001 over 300 epochs and subsequently fine-tune the VAE for 500 epochs on the mini-Sketches. The pre-training slightly improves the performance on the test set when compared to directly training the VAE on the mini-Sketches (Fig. 6.2D).

We couple our fine-tuned VAE (called “Genesis”) with the trRosetta framework. We use the trRosetta fixed backbone design method to optimize sequences for our generated distance- and orientation probabilities [139]. We subsequently use the generated sequences and constraints from trRosetta to minimize the energy with gradient descent and generate 3D models using PyRosetta [325] (see Methods). In summary, the Genesis-trRosetta *de novo* design framework uses a Form to build a Sketch that is then refined by Genesis, designed through trRosetta, and finally assembled and minimized with a full atomistic energy function in Pyrosetta (Sup. Fig. 6.1).

Our ablation studies demonstrate the importance of both the Genesis and trRosetta modules. First, we remove the trRosetta design module by gathering structural restraints directly from the Genesis distance- and orientation probabilities and using a poly-valine AA sequence. We see a low performance with the template modeling (TM) score [266] median being below 0.5 and the root-mean-squared-deviations (RMSDs) median around 4 Å between the predicted 3D model and the native structure on the training and test set examples across all major classes (Fig. 6.3B). Second, we removed Genesis resulting in a framework where trRosetta is challenged to directly design sequences for a given Sketch. On the training examples, we measure a good TM-score median around 0.6 and an RMSD median around 2 Å for the three-helical architectures, while for the fully-β and mixed-α/β architectures the results are rather bad with a median TM-score around 0.4 and the median RMSD around 3.5 Å. The few selected test examples follow the same trend as the train examples: Genesis alone is not sufficient to build native-like poly-valine backbones, and simply using trRosetta to design sequences for Sketch results in a poor performance with sequences and constraints not recapitulating the intended shape of the Sketch. Thus, our experiments support the interpretation that, though Genesis cannot solve the backbone design problem by itself, its predicted features can guide trRosetta towards the sequences that correspond to specific folds. On the other hand, a Sketch alone lacks native-like features that could be identified by trRosetta to use and design fold-specific sequences.

We assess the performance of different variants of the Genesis pipeline. Using the basic framework "Sketch → Genesis → trRosetta → Rosetta", we achieve a TM-score of 0.8 and a median RMSD of 2 Å for fully-α-helical proteins, a median TM-score of ~0.55 and median RMSD ~3.5 Å for fully-β proteins, and a median TM-score of ~0.5 with a median RMSD ~4 Å for mixed-α/β proteins. We see an improvement when adding a simple relaxation with favoring secondary structure pairing and packing after the gradient descent minimization with median TM-scores of approximately 0.8, 0.6, 0.55 and RMSDs of 2 Å, 3 Å, 3.5 Å for alpha-, beta- and mixed-α/β proteins from the training set respectively. Importantly, adding a loss controlling the AA composition of the generating sequences within trRosetta resulted in a small performance drop.

We also tested the pipeline using the trRosetta hybrid-design protocol, where, instead of optimizing for a single sequence, the algorithm optimizes for multiple sequences from which a position-specific scoring matrix (PSSM) is generated and used to guide the sequence design task. The results were comparable to the standard pipeline in terms of TM-scores and RMSDs (Sup. Fig. 6.2).

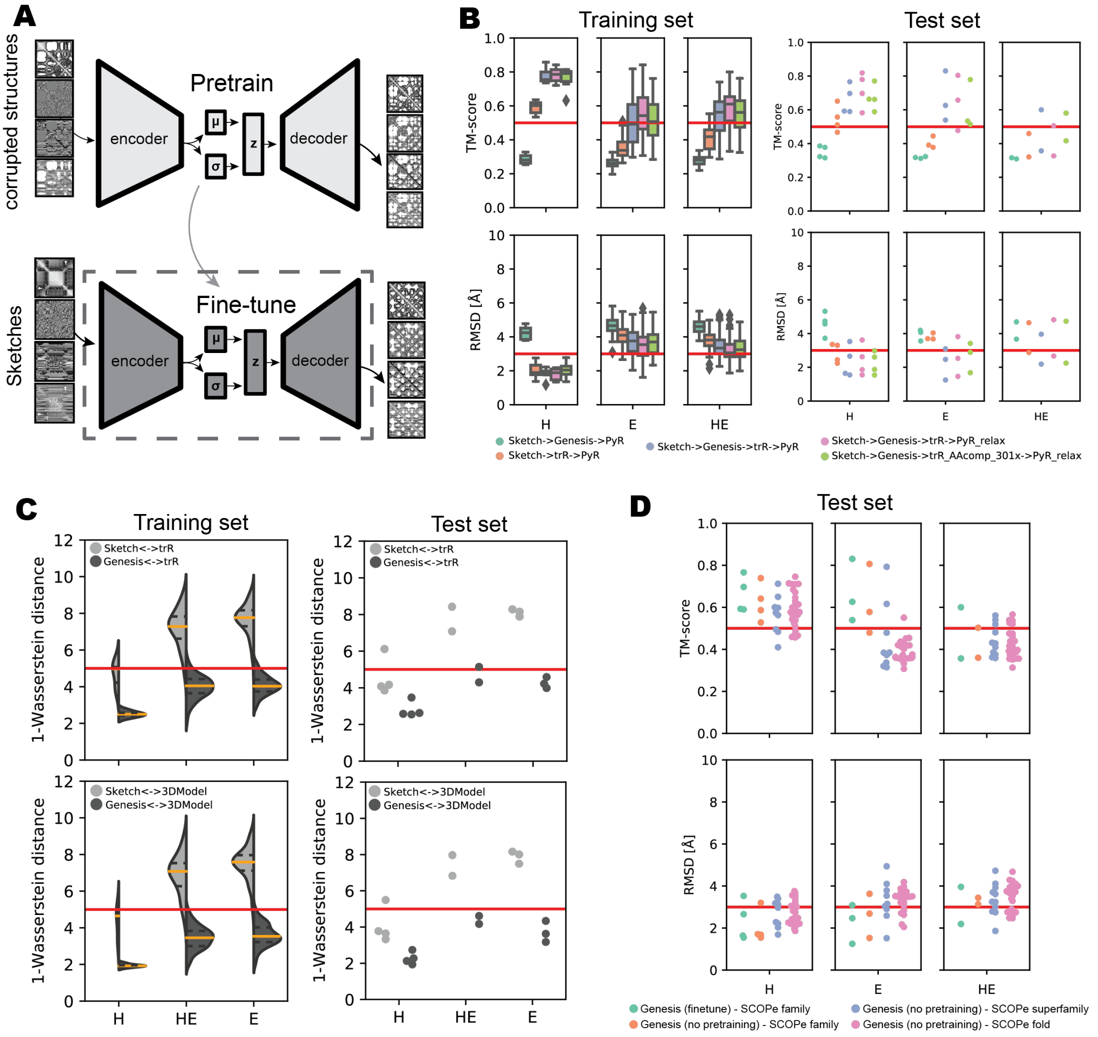


Figure 6.3: Pipeline performances. **A:** General architecture and training scheme of the Genesis module used. Genesis is first pre-trained with corrupted feature maps and subsequently fine-tuned with Sketch feature maps. **B:** Different pipelines and their performances for the different classes of proteins ("H": fully-α-helical, "E": fully-β, and "HE": mixed-α/β). The number of optimization steps is 101 if not differently indicated. "Sketch" represents the input feature maps from the Sketch, "Genesis" is the Genesis module to optimize the feature maps, "trR" is the trRosetta design module and "PyR" is the PyRosetta script to generate 3D models from the generated features and sequence. The first pipeline is an ablation of the trRosetta module, where restraints are derived directly from the Genesis generated feature maps using a poly-valine AA chain for the 3D model generation. The second pipeline is an ablation of the Genesis module where the trRosetta module is directly used to optimize the Sketch feature maps. The three subsequent pipelines are variations of the full pipeline, including additional relaxation steps (PyR_relax) and an adding AA composition loss with 301 optimization steps to the trRosetta module (trR_AAcomp_301x). **C:** Comparison between the Sketch maps - trRosetta / 3D model (3DModel) maps and the Genesis refined maps - trRosetta / 3D model (3DModel) using the 1^{st} Wasserstein distance. **D:** Performance of the standard Genesis pipeline across different difficulty levels according to the SCOPe structure classification.

In order to evaluate the generalization power of Genesis, we train and test Genesis on the series of training and test splits given by SCOPe (Fig. 6.3D) (see Methods). The SCOPe hierarchically classifies proteins based on structural similarities: The top level ("Class") divides proteins into major classes: fully-α, fully-β and mixed-α/β. The next lower level ("Fold") arranges the structures according to secondary structure disposition and connectivity. Two important additional lower levels exist ("Superfamily" and "Family") which take fine-grained structural and functional features into account.

For α-helical proteins, the TM-scores and RMSDs are around 0.6 and 3 Å, respectively, whereas for fully-β and mixed-α/β proteins, a degradation in performance is observed. At the superfamily level two out of 25 test proteins cross the critical 0.5 TM-score threshold for fully-β structures. One level higher (fold level) solely one out of the 75 test proteins crosses the 0.5 TM-threshold. For mixed-α/β proteins, around three test proteins are predicted with a TM-score above 0.5 for both superfamily and fold levels. These results indicate that Genesis is capable of generalizing across families quite well, while proteins with different connectivities and structural features are more difficult to generate successfully.

To showcase the Genesis-trRosetta *de novo* design framework, we conditionally sample five different folds. We sample a two-layer mixed-α/β Ubiquitin-like fold, where four strands are packed against a helix, and a three-layer mixed-α/β Rossmann fold with a central four stranded β-sheet and two exterior packing α-helices on both sides. We additionally challenge the framework by generating two different two-layer β-sandwiches, an Immunoglobulin (Ig) -like fold and a Jelly-roll fold. Finally, we design sequences that adopt the Top7 fold [326], a novel fold not observed in the natural repertoire and representing a generalization test to our method.

As we do not have prior knowledge about SSEs and loop lengths, we sample over 20 to 30 combinations (see Methods). For each combination, we refine three different feature maps using Genesis and for each design a set of 1,000 sequences through trRosetta. With the predicted sequence library we create a PSSM and use it together with the distance- and orientation restraints to design two low energy sequences and 3D models.

We realign each of the 3D models back to the input Sketch and collect all 3D models and sequences that have a TM-score above 0.5. A 3D model with the correct connectivity should have a TM-score around 0.5 to the Sketch (to be the same fold) (Fig. 6.4A). For the Rossmann-, Jelly-roll- and Top7 folds, over 50% of the designs passed the 0.5 TM-score threshold and were collected, while for the Ig-like and Ubiquitin-like folds, approximately 25% passed the TM-score threshold.

We use AlphaFold2 (AF2) to predict a model from the collected sequences and realign the AF2 model to the Genesis model (Fig. 6.4B). We observe that around 50% of the sequences have the expected fold (TM-score > 0.5) and the median RMSDs in the range of 3 Å to 4 Å.

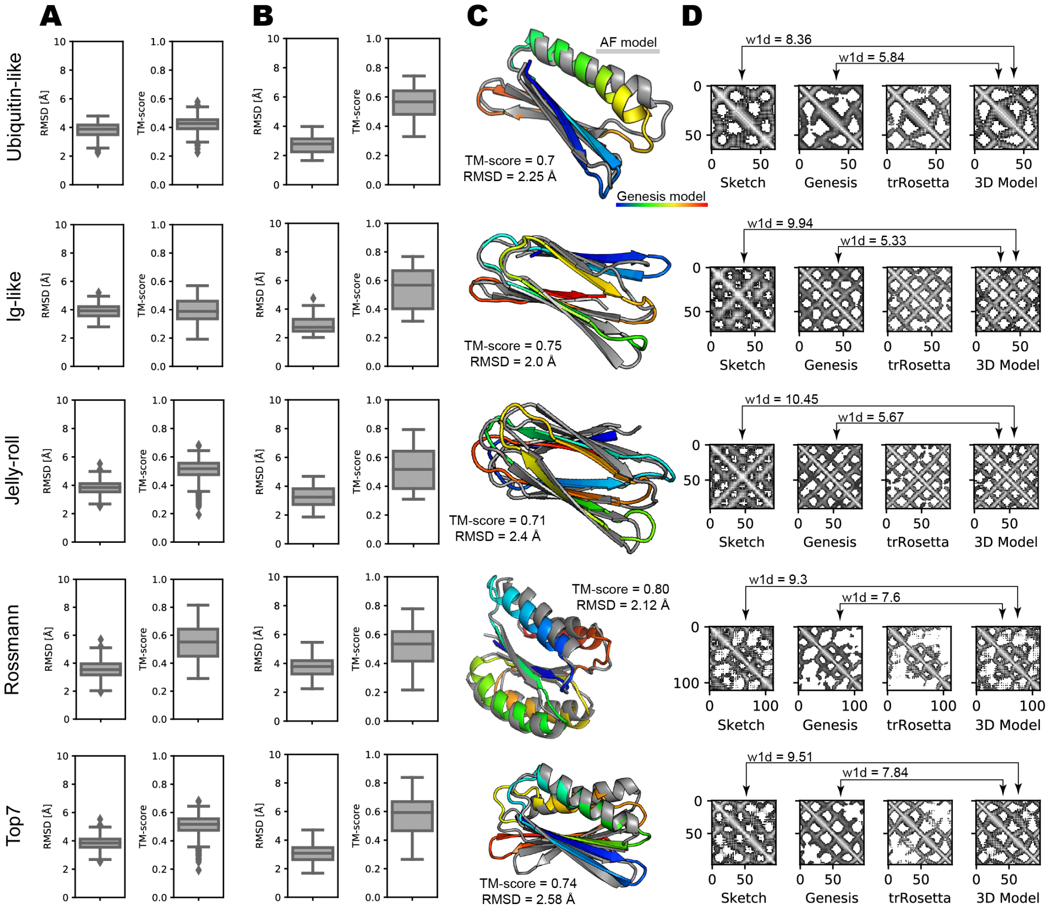


Figure 6.4: Computational results of the sampled native folds. **A:** The TM-scores and RMSDs between the sampled designs and the initial Sketch (input) for the native folds. **B:** The TM-scores and RMSDs between the select designs (TM-score > 0.5 design to Sketch) and their AlphaFold2 (AF2) predictions (without MSA input). **C:** Examples of sampled models (rainbow) and the corresponding AF2 models. D: The distance feature matrix of the Sketch, Genesis, trRosetta, and the 3D model generated using PyRosetta.

6.3.3 Probing "dark matter" protein folds

The ultimate goal of *de novo* protein design methods is to generate protein shapes not existent in nature. We asked the question if our framework based on DNNs and trained on natural derived structure data is capable of (besides the Top7 fold) generalizing outside the distribution of natural folds e.g., if our framework is able to generate out-of-distribution.

To this end, we sought to sample protein folds not included in the training set, nor observed in nature. Previously, Taylor and colleagues [306] computationally analyzed possibly unexplored regions of the three-layer mixed-α/β fold space through Cα-traces that obey constraints of natural protein structures, such as handedness of connection and loop-crossing. We further reduced the set by discarding Cα-traces that have mixed secondary structure types on the same layer, disembodied SSEs (unpacked) or nearly crossing loops. We selected three distinct folds to design with the Genesis-trRosetta method.

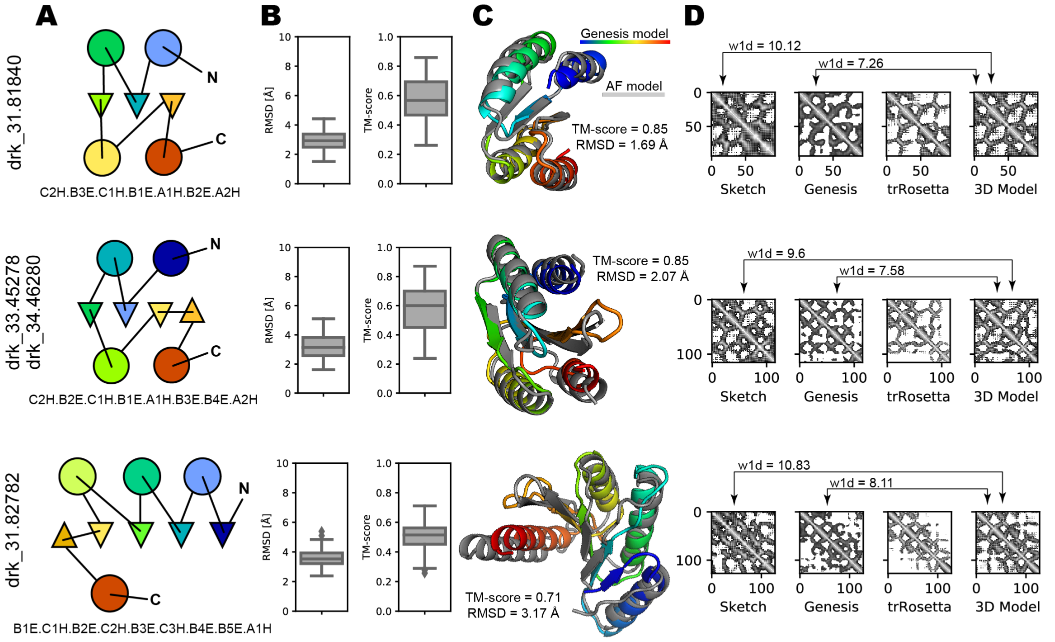


Figure 6.5: Computational results of dark matter folds. **A:** Form descriptions and 2D lattice diagram of the selected dark matter folds. **B:** The TM-scores and RMSDs between the select designs (TM-score > 0.5 design to Sketch) and their AF2 predictions (without MSA input). **C:** Examples of sampled models (rainbow) and the corresponding AF2 models. **D:** The distance feature matrix of the Sketch, Genesis, trRosetta, and the 3D model generated using PyRosetta.

The first novel fold (drk_31.81840) has a three-stranded central β-sheet with two helices on both sides. The top helices are connected through the middle and the side strand, and the bottom helices are connected through the other side strand (Fig. 6.5A top). The second novel fold (drk_33.45278 and drk_34.46280) is also a three-layer fold with a four stranded β-sheet in the middle and two helices on each side. This fold is similar to the first novel fold, but connects the top helices with two β-strands on one side and the two lower helices with the two β-strands on the other side (Fig. 6.5A middle). The third novel fold (drk_31.82782) consists of a five-stranded β-sheet sandwiched with three helices on top and a single helix on the bottom. This fold is “rolled” between the four consecutive strands and the three top helices and the last strand connects the lower helix that bases the full β-sheet (Fig. 6.5A bottom).

We used the Genesis-trRosetta framework to sample sequences using as input different Forms varying in the loop and SSE lengths. We collected the sequences with a TM-score above 0.5 between the 3D models and the Taylor Cα-trace to ensure the correct overall fold. We found that many of the collected 3D models often had single distorted helices or unpaired strands likely due to high resolution constraints and potentially suboptimal sequences sampled from the PSSMs. We therefore picked the two to five best generated 3D model backbones and their PSSM generated by trRosetta and additionally sampled 200 sequences and 3D models for each of the three novel folds.

All sampled sequences then were fed into AF2 (single-sequence prediction) and the AF2-predictions aligned to the initial 3D model (Fig. 6.5B). Interestingly, for the three novel folds the median TM-score was around 0.5 or higher and median RMSDs 3.7 Å or lower. Thus more than half of the sampled sequences have trRosetta models have folds that agree with their corresponding AF2 models.

6.4 Discussion

We show that a specialized VAE termed Genesis is able to encode representations of idealized protein folds and decode native-like conformations. By basing ourselves on distance- and orientation representations, we are able to alleviate the need of generating designable protein backbones in 3D space with fold-specific restraints and energy functions, and thereby also bypassing the need for designable backbones. We couple Genesis to the trRosetta design engine to generate multiple sequences for the sampled distance- and orientation representations for a set of known and novel folds.

Our results demonstrate that the Genesis-trRosetta framework is capable of designing new proteins adopting known folds and novel folds non-existent in nature. By changing secondary structure and loop lengths the overall size can be adjusted and different conformations sampled. The generalization capability of Genesis is significant and can very well drive the trRosetta-Rosetta hybrid design method to sequences with strong fold signatures. Using AF2 as an orthogonal test shows that many of these sequences adopt the intended target shape. Additionally, our framework is considerably fast, within minutes to generate a sequence and a 3D model for a given target protein shape even on a central processing unit (CPU). This demonstrates the usage of DNNs can leverage the automated generation of proteins normally only accessible through large-scale simulations [74, 153].

Our work opens exciting new horizons for *de novo* protein design where control over the shape is desired. For example, our method could be harvested to generate custom protein backbones such that they fit onto non-canonically structured protein interfaces. Often times, nanomaterials exhibit highly regular patterns, and could therefore be engaged by secondary structures that are placed respecting the regularity constraints. Another example where our method could be used is for the design of larger molecular assemblies that are constructed from smaller protein domains. Often, the overall shape of the assembly is controlled by the shape of the individual subunits. Hence, we expect that the versatility and speed of the Genesis-trRosetta method together with other potential DNN tools for protein design and engineering to explore the protein universe should be broadly useful.

6.5 Methods

6.5.1 Dataset generation

We created two distinct datasets from the SCOPe (v2.07 stable) [327] domains of medium sizes (40 - 128 residues). (1) The pre-training data set was created by corrupting existing protein structures by removing the loops based on the DSSP (hydrogen bond estimation algorithm) [210] assignments. We remodel the loops as done in a Sketch, e.g. we add dummy residues (N, C, Cα, O backbone atoms with randomized torsion angles) along the shortest path between the two endpoints Cα atoms of the consecutive SSEs. We add as many dummy residues as in the native structure, hence the corrupted structure has the same length as its native counterpart. This procedure leaves the native secondary structures dispositions that may incorporate important native structural features for the pre-training. In total, we created a total of 40,726 pairs. (2) We developed a program that creates small fold Sketches obeying simple topological rules such as non-crossing loops and loop distance restraints from the architecture types: EE_EEE, EEE_EEE, H_EEE, H_EEEE, H_EEE_H, HH_EE, HH_EEE, HH_EE_H, HHH, HHH_EE (where “_” represents a layer separation and E: β-strand and H: α-helix). We searched the SCOPe domains for partial structural matches within 3 Å RMSD using MASTER [309, 310] for each of the generated mini-Sketches. Importantly, a mini-Sketch can partially match onto a native domain. We crop the overlapping regions of the native domain at the first and the last residue of the matching Sketch. Secondary structures within the cropped domain that do not map to secondary structures in the Sketch are assigned as loops. Furthermore, we remove domains larger than 128 residues and identical matches for the mapping to the same Sketch. This resulted in a total of 35,435 Sketch - native domain pairs.

6.5.2 Data splits

Within SCOPe, protein structures are hierarchically classified into groups where the "Class" groups proteins based on secondary structure content and organization (fully-α, fully-β, mixed-α/β), "Fold" divides them based on SSE disposition and connectivity, "Superfamily" is based on structural features and "Family" contains the structures with similar sequences.

We pick protein families that represent compact structures with small loops for our family test set. The test set includes the SCOPe families b.1.22.1, b.11.1.6, b.69.2.3, b.70.2.1, b.82.1.22, b.114.1.1, a.7.2.0, a.7.2.1, a.7.8.2, a.7.12.1, a.8.11.1, a.24.10.3, a.24.13.1, a.60.9.0, a.160.1.2, c.2.1.7, c.25.1.2, c.118.1.0, c.93.1.0, c.56.5.6, d.110.4.3, e.51.1.1, c.97.1.5, d.17.1.5, d.58.3.2, d.58.10.0, d.58.23.1, d.92.1.13, d.230.1.1, d.240.1.0. We generate higher-level test sets (fold and superfamily) by removing all corresponding groups from the picked structures in the Family test set, e.g. for the family b.1.22.1 the superfamily is b.1.22 and the Fold is b.1.. Importantly, identical structures and mini-Sketches in the training set were removed in order to avoid any biases during testing.

6.5.3 Data encoding

The coordinates of the Sketches and their native counterparts are encoded into a total of four 2D distance- and orientation feature maps as done by trRosetta. Briefly, the first feature map is all-against-all Cβ distances. The second feature map is the dihedral "ω" that measures the rotation along the virtual axis of two connecting Cβ residues. The distances and ω-angles are symmetric, e.g. measuring from residue i to residue j will give the same result as measuring from residue j to residue i. The third and fourth feature map are the "ϑ" dihedrals and the "φ" angles specifying the direction of Cβ of residue i with respect to residue j. Both, ϑ and φ are asymmetric metrics. Together the four feature maps fully define a protein backbone in 3D space.

While we use real valued feature maps as input to Genesis, we bin the true feature maps according to the trRosetta scheme. The distances from 2 to 20 Å are binned into 36 equally spaced segments (0.5 Å each) and a 37$^{\text{th}}$ bin to indicate that pairs are not in contact. The dihedral (ω, ϑ) and angular (φ) features are binned into 15° segments yielding 24, 24, and 12 with an additional bin indicating no contact, respectively. Therefore, we have encoded the true feature maps into tensors of shape 128x128x1x37 for the distances, 128x128x1x25 for the dihedrals and 128x128x1x13 for the angles. Thus, at each "pixel" (each residue pair) we have an additional dimension that can be seen as a dirac distribution with a score of one for the bin with the distance and zero everywhere else.

6.5.4 Genesis architecture

The VAE includes an encoder, a decoder and a loss function. The input Sketch x feature maps (real-valued) of shapes 128x128x4 are processed by the encoder, a sequence of four convolutional blocks. A single block includes a 2D convolution, an instance norm and ELU activation followed by a 40% dropout. From the compressed data representation, we use two multilayer perceptrons (MLPs) to predict a normal distribution over the latent space $p(z|x)$ through predict means and covariances two vectors of size 128. Using the reparametrization trick, we sample a latent variable z from $p(z|x)$. The decoder $q(y|z)$ passes z through three blocks of 2D deconvolution, instance norm, ELU activation and 40% dropout to create a decompressed representation. The final layer of the decoder branches into four different heads. Each head is a convolutional block with a final softmax activation over each pixel yielding distance outputs of shape 128x128x1x37, two dihedral outputs of sizes 128x128x1x25 and an angular output of shape 128x128x1x13.

6.5.5 Loss function

Our loss function is composed of five individual losses (four reconstruction losses, and a loss on the latent space).

We use the Wasserstein distance (for details see [328]) as reconstruction loss. Let us define x P

and y Q and the corresponding densities as p and q, respectively. We assume that $(x, y) \in \mathbb{R}^d$. Additionally, let us denote $\mathscr{J}(P, Q)$ all joint distributions $\mathscr{J}$ for (x, y) that have marginals P and Q. Then the general Wasserstein distance can be written as

$$W_p(\mathscr{P}, \mathscr{Q}) = \left(\inf_{J \in \mathscr{J}(\mathscr{P}, \mathscr{Q})} \int \|x - y\|^p \, dJ(x, y) \right)^{1/p} \tag{6.1}$$

In the discrete case, when P and Q are distributions $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ the formulation becomes

$$W_p(\mathscr{P}, \mathscr{Q}) = \left(\sum_{i=1}^{n} \|x_i - y_i\|^p \right)^{1/p} \tag{6.2}$$

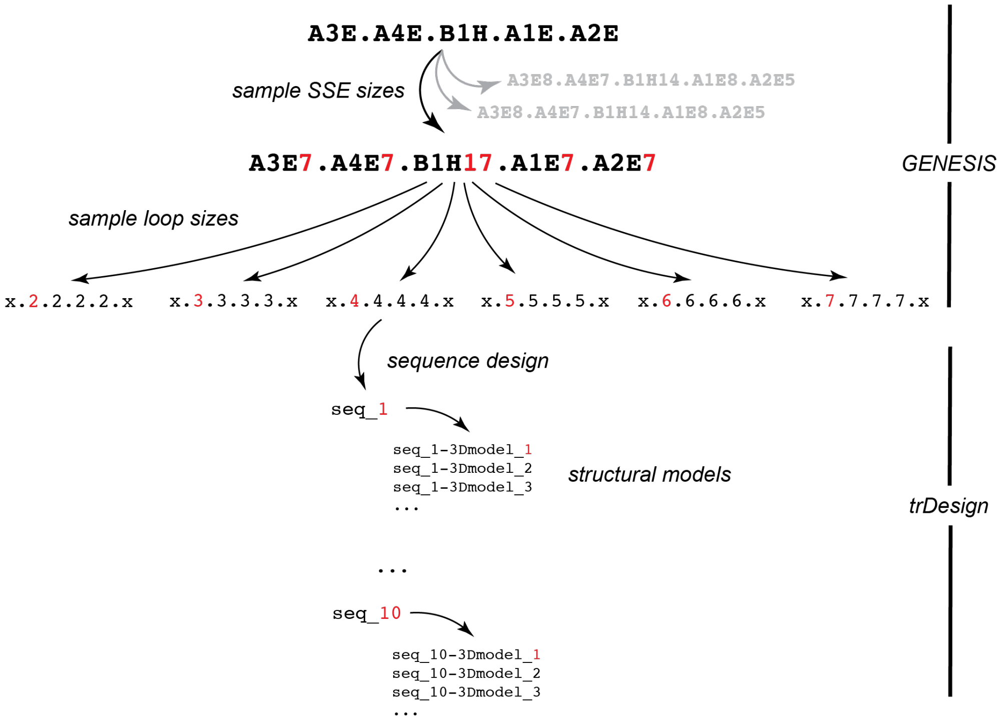
In the case of 1D discrete distributions $(p = 1)$, the 1-Wasserstein (W_1) distance is also called Earth mover's distance (EMD) and is efficiently computable. The main advantage of the 1-Wasserstein distance compared to other measures such as the binary cross-entropy and the Kullback-Leibler (KL) divergence is that it takes into account the metric space. This means that larger deviations from the predicted to the true distributions are more penalized while small errors are less penalized.

We define the reconstruction loss as the sum over the 1-Wasserstein distances between the predicted distributions $(\hat{D})$ and the true distributions (D) of each pixel normalized by the length of the protein (N_{AA}). Each pixel is defined as (i, j) where $i = 1, \ldots, n_w$ and $j = 1, \ldots, n_h$ with n_w being the width and n_h the height.

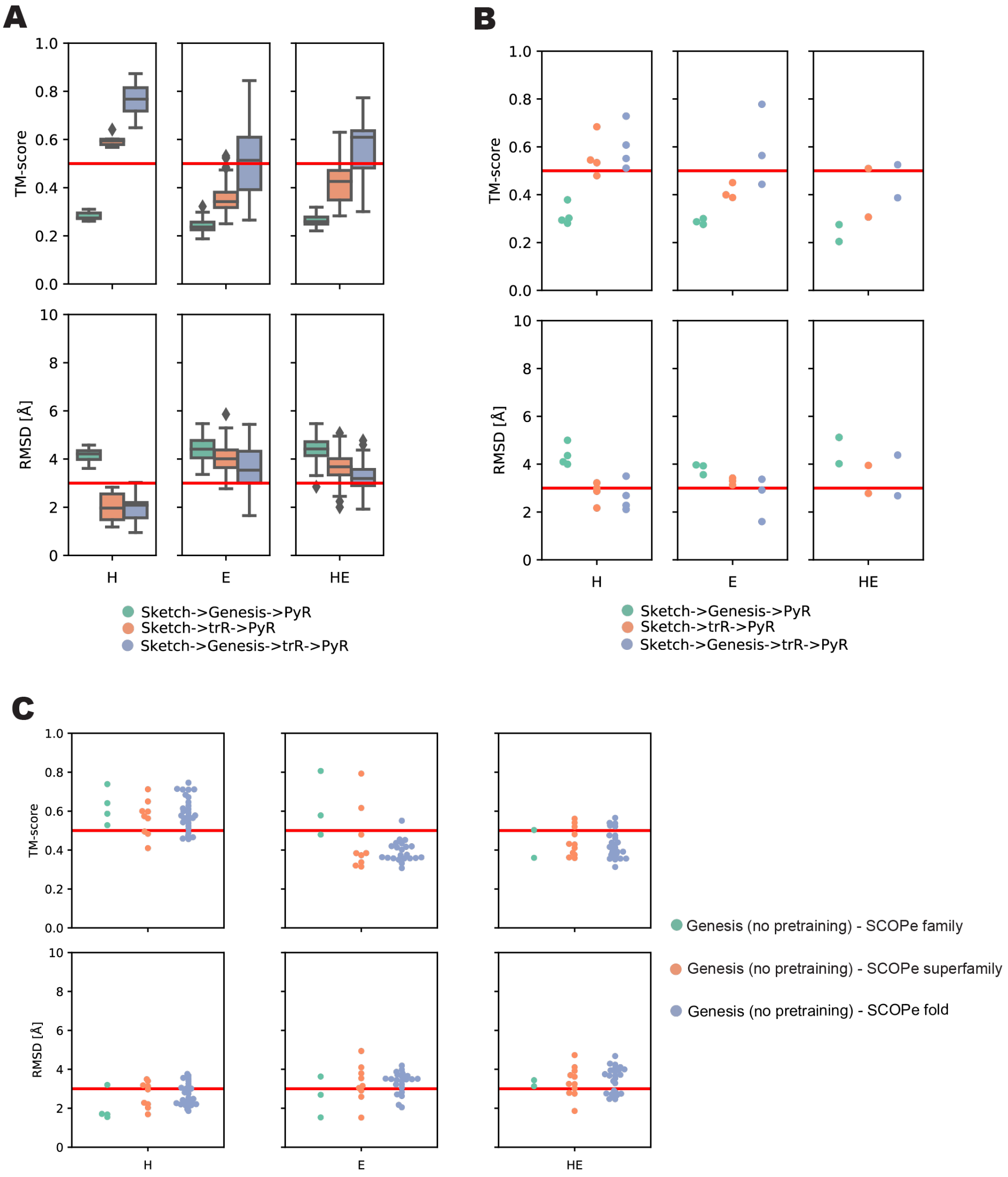
$$L_{\mathrm{rec}} = \frac{1}{N_{\mathrm{AA}}} \sum_{i=1}^{n_w} \sum_{j=1}^{n_h} W_1(D_{i,j}, \hat{D}_{i,j}) \tag{6.3}$$

Note that the true distribution is modeled as a Dirac distribution supported by the true values, whereas the predicted distribution $(\hat{D})$ is parametrized by the VAE decoder. We additionally use the Kullback-Leibler (KL) divergence on the latent space normalized by the length of the protein to penalize latent vectors not following a Normal distribution $\mathrm{KLD} = \frac{1}{N_{\mathrm{AA}}} \mathrm{KL}(p(z|x) \,\|\, p(z))$, with $p(z)$ $\mathrm{Normal}(0, 1)$. Thus the final loss is defined as

6.6 Supplementary information



Supplementary Figure 6.1: Sampling strategy We sample several combinations of SSE and loop sizes yielding different sized Sketches and different refined feature maps by Genesis. Using trRosetta, we design multiple sequences for each of the feature maps. Then, using PyRosetta, we generate multiple potential structural models per sequence including the distance- and orientation restraints from the feature maps.



Supplementary Figure 6.2: Pipeline performances with MSA design **A:** Training set: different pipelines and their performances for the different classes of proteins (“H”: fully-α-helical, “E”: fully-β, and “HE”: mixed-α/β) using the hybrid trRosetta design approach. The number of optimization steps is 101 if not differently indicated. **B:** Test set performances over different protein classes using the hybrid trRosetta design approach. **C:** Performance of the Genesis pipeline using the hybrid trRosetta design across different difficulty levels according to the SCOPe structure classification.

7 Conclusions & Perspectives

This dissertation presented the development of novel computational protein design platforms for automated *de novo* design of protein folds and PPIs. Since the start of my Ph.D. in October 2017, the field of *de novo* protein design has remarkably advanced; and each of the presented methods have (1) contributed to the field's scientific progress and (2) helped to answer or deepen our current understanding of of important problems related to *de novo* design of proteins and PPIs. In the following section, I will recapitulate our findings and then try to give future perspectives from what we have learned in the context of the overall scientific progress.

7.1 Orchestrating PPIs and beyond

The computational design of novel PPIs remains a challenging biological problem. In chapters 2 and 3, computational protein interface design strategies were employed to efficiently predict a set of potential point mutations that could increase the binding affinity to a specific target. Specifically, in chapter 2, a small protein that binds tightly to the *Spy*Cas9–sgRNA complex and inhibits DNA editing was engineered by fusing a light-sensitive domain LOV2 of *A. sativa* phototropin-1 to AcrIIA4. When blue light interacts with the LOV2 domain, a conformational change (i.e. partial unfolding) is prompted. In the case where the LOV2 domain is fused to another protein, the induced conformational change gets propagated and can disrupt the activity of the fused partner. The light-induced disruption strategy is generally applicable for engineering switches where the function of the partner protein depends on the subtle spatial organization of its atoms. For example, for enzymes to function, their catalytic site requires precisely coordinated residues; and for proteins to interact specifically, they require a geometric and electrostatic complementary interface with respect to their target site. Initial experimental testing of the switching behavior of the generated LOV2-AcrIIA4 hybrid was accompanied by significant inhibitory leakage i.e., the DNA-editing off-switching behaviour of the hybrid was incomplete. Although the hybrid was able to inhibit *Spy*Cas9, significant DNA editing was still observed. This could result from the fusion procedure disrupting the AcrIIA4 structure and its binding activity. This activity disruption after fusing two proteins together is frequently observed [329, 198]. To fully recover the inhibitory activity of the engineered hybrids, a computational single-sided interface design protocol was developed. This permitted the screening of a large set of potential mutations at the interface between the hybrid and the enzyme complex to increase the binding affinity and thereby reduce DNA-editing leaks. The calculations pointed towards two mutations that showed full recovery of the LOV2-AcrIIA4 hybrid inhibitory activity while retaining full switch-reversibility. The results highlight: (1) small conformational deformations can efficiently be corrected by few mutations restoring the initial activity, and (2) fast, simple modeling simulations enabled the screening of vast search space that is experimentally unattainable. The developed strategy's extends beyond switchable Cas9 inhibitors, for example to design improved protein binders or light-sensitive enzymes.

In chapter 3, improvements of the computer-guided interface design protocol led to the repurposing of AcrIIC1, a broad-spectrum inhibitor of various Cas9s towards a newly designed Acr with superior inhibition potency against *Staphylococcus aureus* (*Sau*)Cas9. This Cas9 orthologue is of substantial importance for *in vivo* genome editing due to its small size and efficient packaging into adenovirus-associated viral vectors (AAVs). No other rationally designed Acr inhibitor against *Sau*Cas9 exits, and only shortly after the project a natural inhibitor was discovered [262]. To create a *Sau*Cas9 specific Acr variant, a computational search for amino acid (AA) substitutions in the binding region of AcrIIC1 that must increase its shape and electrostatic complementarity to the *Sau*Cas9 structure was performed. From the simulations, ten mutations were proposed, three of which combined (N3F, D15Q, A48I) yielded AcrIIC1X: a potent (*Sau*)Cas9 inhibitor supported by experimental validation. Our methods provide the first rational basis for an engineering strategy to improve and redirect Acr functioning. It exemplifies that by minimal, but detailed optimization of the interface it is possible to (re)design for tight binding. Interestingly, no mutation within the center of the interface of AcrIIC1 was proposed to be beneficial by the computer simulations. This strengthens the hypothesis that the often-hydrophobic residues within the center of the interface aid non-specific binding, while the outer polar ring around the interface delineates the specificities. This concept also underlies the hotspot-centric design methods i.e., certain hotspot residues substantially power the interaction whereas the remaining interactions contribute little on their own but have a substantial additive effects [166, 281].

In chapters 2 and 3, we took advantage of known protein-protein binding sites and natural binders to extract hotspot residues and transplant them onto other protein structures. However, structures and interfaces are often times unknown or not readily available. In chapter 4, MaSIF applications were integrated into a *de novo* PPI design platform that allows for the design of new, site-specific interactions that can sustain a PPI. This platform is significant as it is the first method that is guided by learned surface fingerprint descriptors that can effectively be leveraged to create of novel PPIs. The surface-centric PPI design strategy allows for an ultra-fast search of complementary surfaces that can optimally complement the identified target region. This is difficult to achieve with current state-of-the-art. Hotspot-centric methods rely on computational docking of disembodied side chains to identify clusters of hotspot residues that could contribute to the mediation of high-affinity interactions. [281, 154]. However, they are limited due to the potential impossibilities to precisely integrate the gathered hotspots simultaneously and forming dense, target-complementary interaction surfaces thereof. Thanks to MaSIFs sensitivity to conformational changes, it makes the framework identify seeds containing hotspot residues with the exact side chain conformations needed for inducing realistic and specific PPIs.

The MaSIF-based *de novo* PPI design framework was employed to design completely new PD-L1 inhibitors, an immune checkpoint receptor. Besides being clinically revelant, the PD-L1 represents a very challenging target for designing specific PPIs for. It's surface lacks deep hydrophobic grooves that can serve as "anker" spot for large hydrophobic residues, and hence make it difficult to design specific and strong molecular interactions with hotspot-centric methods. Furthermore, only monoclonal antibodies (mAbs) targeting PD-L1 were successfully developed [330] and small molecule inhibitors are in early drug development stages [289]. Employing MaSIF's surface-based design strategy, the PD1 binding site of PD-L1 was targeted by searching for complementary surfaces. The fragments used were α-helical because (1) no known α-helical binder against PD-L1 exists to the best of our knowledge, and (2) α-helical fragments can readily be transferred to larger protein scaffolds to confer stability and further optimize the enlarge interface. Our lead designs showed nanomolar binding affinity to PD-L1, comparable to the natural antibody-antigen binding. The complex crystal structure solved by collaborators demonstrates that the optimized design - PD-L1 complex showing bound PD-L1 at the desired site with high atomic accuracy (RMSD = 1 Å).

Taken together, the MaSIF *de novo* design framework opens possibilities to target interfaces previously thought to be “undruggable”, as seen by the SAS-6 and PD-L1 designs. These results highlight the framework's potential to target any protein by crafting protein binders from the ground up without relying on native interactions. By operating on the molecular surfaces rather than the atomistic representation, this unlocks exciting possiblities to target modified proteins e.g., glycosylated, or post-translationally modified. This also extends to large protein assemblies such as amyloid fibrils and even beyond biological molecules to target structured nanoscale materials. Furthermore, MaSIFs' sensibility to small changes within the surface representations could effectively be used to decipher dynamic or neo-interfaces emerging from quaternary complexes. This could open possibilities to target PPIs during specific dynamic states that may not be observed with an atomistic representation.

7.2 Probing the protein universe with designer proteins

The precise structure of a protein dictates its function. Different protein structures can perform different functions, depending the specific atomic positions in 3D space. Thus, creating proteins with certain shapes and functions represents major aims for the field. A fundamental challenge is that not all AA sequences successfully fold into a protein and finding a viable sequence by randomly sampling through the sequence space is extremely unlikely. Instead of searching for new proteins at the sequence level, established *de novo* protein design methods start by defining the overall shape of the protein that one would like to create and thereby limit the sequence space to only sequences that could potential fit the predefined shape. Then, backbones of the shape are sampled, and for each of the sampled backbone conformations an AA sequence with a low free energy is fitted. Albeit remarkable successes in designing novel proteins [76, 77, 60], this method is far from optimal.

De novo protein design approaches face various hurdles. The sampling methods used are heuristic and the best conformation could be missed. Also, the current scoring functions approximating the stability of a protein are not accurate enough and can lead to many false positives i.e., low free energy sequences for a target backbone that do not fold experimentally. Furthermore, while excelling at quantifying low-resolution terms, energy functions lack a quantitative metric to evaluate whether a protein shape is realizable in 3D with the available set of AAs i.e., the designability. Lastly, classic protein design methods often converge in sequence space i.e., generating uniform or highly similar sequences rather than predicting a set of distinct variants. In a previous project [331], sequence optimization for a selection of protein structures starting from different sequence initializations were performed, however all simulations converged. The sequences were not identified by the standard bioinformatics tools (e.g., Hmmer [332] and BLAST [315]) to belong to the targeted structure. This highlighted the intrinsic inaccuracy of the scoring functions that drift their calculations towards sequences lacking identifiable natural sequence signatures. To address this problem, a genetic algorithm was developed that biases the sequence search towards the natural sequence patterns and allows for designing sequences with native-like features. A few of the designs were experimentally validated and found to fold into the targeted shape in addition to being thermodynamically stable. This highlights the inaccuracies of scoring functions and the sampling issues and proposes an alternative solution. To improve the performance of scoring functions an additional term estimating the "native-likeness" should be integrated i.e., with DL-based models such as transformers that have learned from massive amounts of sequences [103, 333]. The merge of DL-based terms would represent a first step towards a general, full-atomistic learned molecular scoring function with improved speed and accuracy.

In chapter 5, an enhanced version of the TopoBuilder to hierarchically construct protein architectures and folds was presented. A major hindrance for the *de novo* design of novel proteins is that many models are not designable due to the lack of native-like structural details insufficiently captured throughout the modeling calculations. Handling this issue, a new method that incorporates overall structural features from natural tertiary motifs rendering towards more native-like designs was developed. Concretely, the relative orientations of the SSE in the global context of the protein was automatically optimized. Our results show that with a minimal set of geometric corrections from native tertiary motifs that compose the folds capture enough information to adapt and improve the designability of protein backbones. Additionally, the fine-tuning interventions led to a designability boost of the backbones, indicating their importance and that current fragment insertion methods do not capture global fold determining features. Previous methods [99, 64] focused on empirical rules to design ideal structured loops that would force the SSE placements. The TopoBuilder represents an alternative and complementary solution that is largely automatic. Rather than focusing on structured loops, the global placements of SSEs is optimised and corrected and thereby implicitly guide the loop geometries. Eventually, the TopoBuilder lays the foundations for generating proteins with controlled folds that can be harvested to scaffold functional sites or create larger complex protein machinery by assembling single *de novo* designed domains.

Despite the qualitative improvements, the TopoBuilder is computationally demanding, requiring specialized cluster hardware and time-consuming simulations. In the last chapter, a novel and holistic approach Genesis to model and refine protein folds independent from scoring functions and computationally intensive calculations was proposed. Particularly, a variational autoencoder (VAE) was developed and trained on a large dataset of protein Sketches - protein structures. Genesis is able to convert protein descriptions into globally coherent structural models representations without the need of operating in 3D and the use of energy functions. Our results show that Genesis can encode representations of simple topological description of proteins and can readily decode native-like structure representations. Importantly, we avoid the need for sampling and scoring in 3D, hence bypassing the need for crafting designable backbones. By sampling novel topologies, we show that Genesis effectively learned to generate structure representations that, when coupled to trRosetta yielded sequences that are predicted to adopt the intended fold. This method's generalization capabilities and prediction speed contribute to a new tendency for *de novo* protein design, i.e., "neural *de novo* protein design".

A potential next step could be to interface Genesis with AlphaFold2 (AF2) and its full-atom predictions. Not only could this improve the sequence design quality of the framework, but also extend Genesis's ability to also denoise side chain conformations. Hence, this would enable the control over backbone and side chain configurations simultaneously, and could therefore be of great importance for multiple design tasks where highly accurate side chain geometries are required e.g., enzymatic sites, channels and cavities within proteins or PPIs.

In the long run, the presented work, the reported results and discovered design principles contribute to the shift towards a universal platform for generating *de novo* designed proteins and could be leveraged to tackle currently unsolved computational design problems. For example, MaSIFs sensitivity may enable the design of synthetic enzymes or epitope-stabilized vaccines, both requiring the exact display of the motif surfaces. Then, Genesis could be adapted for generating protein scaffolds that are able to stabilize very specific motif surfaces. Adding dynamics to the Genesis and MaSIF frameworks may help to better understand and control specific surface conformations and required fold configurations. Alltogether, the methods could jointly be used for larger biomolecular assemblies and machineries with specific incorporated dynamic behaviours to tightly control biological signals.

...

The coming years promise exciting new avenues for *de novo* design i.e., to study the fundamentals of biological systems or address unmet biomedical needs.

A rstoolbox - a Python library for large-scale analysis of computational protein design data and structural bioinformatics

...

This chapter is a postprint version based on an article pulished in BMC Bioinformatics in 2019 (DOI: 10.1186/s12859-019-2796-3) in accordance with the the publisher.

Authors
Jaume Bonet[1,2], **Zander Harteveld**[1,2], Fabian Sesterhenn[1,2], Andreas Scheck[1,2] and Bruno E. Correia[1,2*]

Authors and affiliations
[1]Institute of Bioengineering, École Polytechnique Fédérale de Lausanne, CH-1015, Lausanne, Switzerland. [2]Swiss Institute of Bioinformatics (SIB), CH-1015, Lausanne, Switzerland.

Author contributions

JB, ZH, FS and AS contributed to the code. JB and BEC devised the examples and wrote the manuscript. All authors read, contributed and approved the final version of the manuscript.

A.1 Abstract

Large-scale datasets of protein structures and sequences are becoming ubiquitous in many domains of biological research. Experimental approaches and computational modelling methods are generating biological data at an unprecedented rate. The detailed analysis of structure-sequence relationships is critical to unveil governing principles of protein folding, stability and function. Computational protein design (CPD) has emerged as an important structure-based approach to engineer proteins for novel functions. Generally, CPD workflows rely on the generation of large numbers of structural models to search for the optimal structure-sequence configurations. As such, an important step of the CPD process is the selection of a small subset of sequences to be experimentally characterized. Given the limitations of

current CPD scoring functions, multi-step design protocols and elaborated analysis of the decoy populations have become essential for the selection of sequences for experimental characterization and the success of CPD strategies.

Here, we present the rstoolbox, a Python library for the analysis of large-scale structural data tailored for CPD applications. rstoolbox is oriented towards both CPD software users and developers, being easily integrated in analysis workflows. For users, it offers the ability to profile and select decoy sets, which may guide multi-step design protocols or for follow-up experimental characterization. rstoolbox provides intuitive solutions for the visualization of large sequence/structure datasets (e.g. logo plots and heatmaps) and facilitates the analysis of experimental data obtained through traditional biochemical techniques (e.g. circular dichroism and surface plasmon resonance) and high-throughput sequencing. For CPD software developers, it provides a framework to easily benchmark and compare different CPD approaches. Here, we showcase the rstoolbox in both types of applications.

rstoolbox is a library for the evaluation of protein structures datasets tailored for CPD data. It provides interactive access through seamless integration with IPython, while still being suitable for high-performance computing. In addition to its functionalities for data analysis and graphical representation, the inclusion of rstoolbox in protein design pipelines will allow to easily standardize the selection of design candidates, as well as, to improve the overall reproducibility and robustness of CPD selection processes.

A.2 Background

The fast-increasing amounts of biomolecular structural data are enabling an unprecedented level of analysis to unveil the principles that govern structure-function relationships in biological macromolecules. This wealth of structural data has catalysed the development of computational protein design (CPD) methods, which has become a popular tool for the structure-based design of proteins with novel functions and optimized properties [334]. Due to the extremely large size of the sequence-structure space [306], CPD is an NP-hard problem [63]. Two different approaches have been tried to address this problem: deterministic and heuristic algorithms.

Deterministic algorithms are aimed towards the search of a single-best solution. The OSPREY design suite, which combines Dead-End Elimination theorems combined with A* search (DEE/A*) [335], is one of the most used software relying on this approach. By definition, deterministic algorithms provide a sorted, continuous list of results. This means that, according to their energy function, one will find the best possible solution for a design problem. Nevertheless, as energy functions are not perfect, the selection of multiple decoys for experimental validation is necessary [336, 337]. Despite notable successes [338, 53, 339], the time requirements for deterministic design algorithms when working with large proteins or *de novo* design approaches limits their applicability, prompting the need for alternative approaches for CPD.

Heuristic algorithms, such as those based on Monte Carlo (MC) sampling [340], use stochastic sampling methods together with scoring functions to guide the structure and sequence exploration towards an optimized score. These algorithms have the advantage of sampling the sequence-structure space within more reasonable time spans, however, they do not guarantee that the final solutions reached the global minimum [341]. Heuristic CPD workflows address this shortcoming in two ways: I) extensive sampling generating large decoy sets; II) sophisticated ranking and filtering schemes to discriminate and identify the best solutions. This general approach is used by the Rosetta modelling suite [68], one of the most widespread CPD tools.

For Rosetta, as with other similar approaches, the amount of sampling necessary scales with the degrees of freedom (conformational and sequence) of a particular CPD task. Structure prediction simulations such as *ab initio* or docking may require to generate up to 106 decoys to find acceptable solutions [97, 342]. Similarly, for different design problems the sampling scale has been estimated. Sequence design using static protein backbones (fixed backbone design) [297] may reach sufficient sampling within hundreds of decoys. Protocols that allow even limited backbone flexibility, dramatically increase the search space, requiring 104 to 106 decoys, depending on the number of residues for which sequence design will be performed. Due to the large decoy sets generated in the search for the best design solution, as well as the specificities of each design case, researchers tend to either generate one-time-use scripts or analysis scripts provided by third parties. In the first case, these solutions are not standardized and its logic can be difficult to follow. In the second case, these scripts can be updated over time without proper back-compatibility control. As such, generalized tools to facilitate the management and analysis of the generated data are essential to CPD pipelines.

Here, we present rstoolbox, a Python library to manage and analyse designed decoy sets. The library presents a variety of functions to produce multi-parameter scoring schemes and compare the performance of different CPD protocols. The library can be accessed by users within three levels of expertise: a collection of executables for designers with limited coding experience, interactive interfaces such as Ipython [343] for designers with basic experience in data analysis (i.e. pandas [344]), and a full-fledge API to be used by developers to benchmark and optimize new CPD protocols. This library was developed for direct processing of Rosetta output files, but its general architecture makes it easily adaptable to other CPD software. The applicability of the tools developed expands beyond the analysis of CPD data making it suitable for general structural bioinformatics problems (see extended_example notebook in the code's repository). Thus, we foresee that rstoolbox may provide a number of useful functionalities for the broad structural bioinformatics community.

A.3 Implementation

rstoolbox has been implemented extending from pandas [344], one of the most established Python libraries for high-performance data analysis. The rstoolbox library architecture is composed of 4 functional modules (Fig. A.1): I) rstoolbox.io - provides read/write functions for multiple data types, including computational design simulations and experimental data, in a variety of formats; II) rstoolbox.analysis - provides functions for sequence and structural analysis of designed decoys; III) rstoolbox.plot – plotting functionalities that include multiple graphical representations for protein sequence and structure features, such as logo plots [345], Ramachandran distributions [346], sequence heatmaps and other general plotting functions useful for the analysis of CPD data; IV) rstoolbox.utils – helper functions for data manipulation and conversion, comparison of designs with native proteins and the creation of amino acid profiles to inform further iterations of the design process.

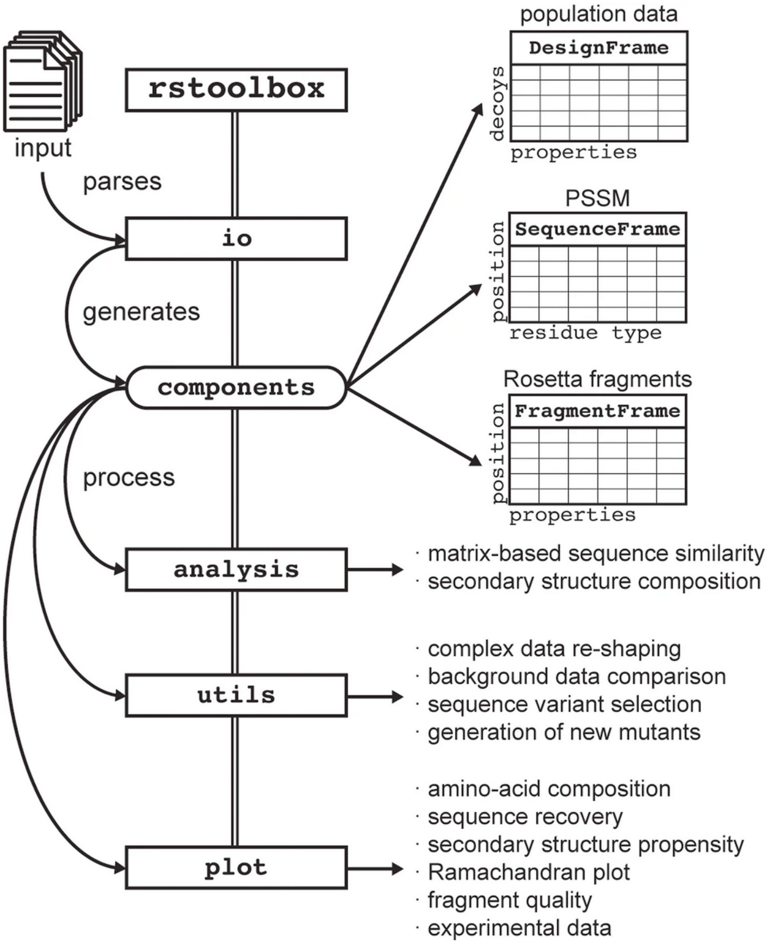
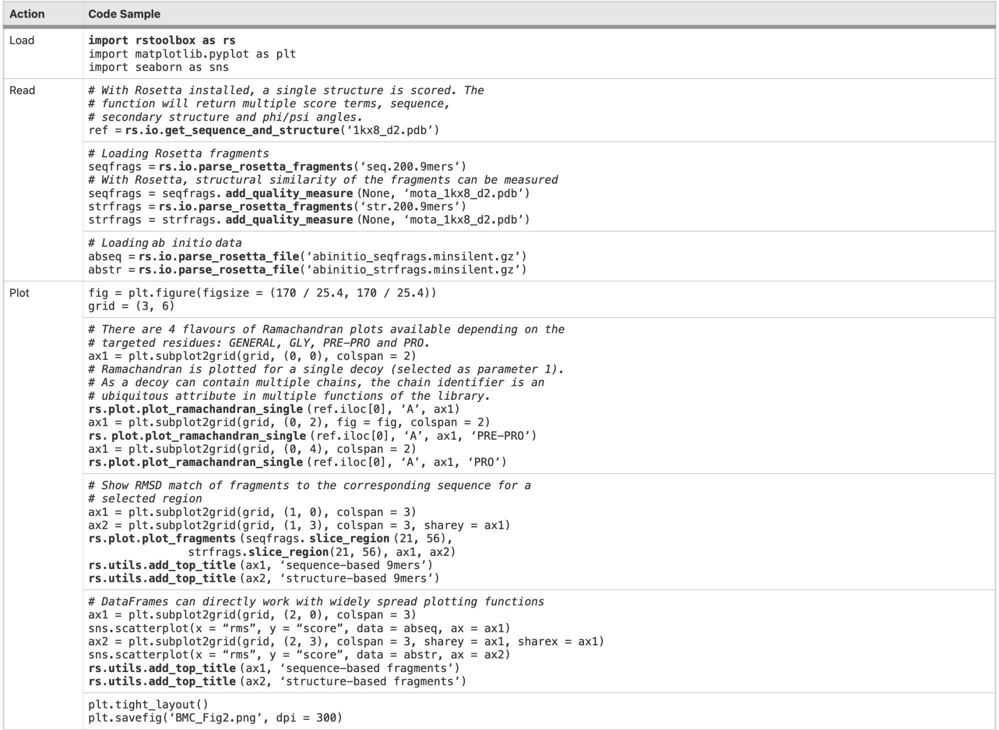


Figure A.1: rstoolbox library architecture. The io module contains functions for parsing the input data. The input functions in io generate one of the three data containers defined in the components module: DesignFrame for decoy populations, SequenceFrame for per-position amino acid frequencies and FragmentFrame for Rosetta's fragments. The other three modules analysis, utils and plot, provide all the functions to manipulate, process and visualize the data stored in the different components.

Additionally, rstoolbox contains 3 table-like data containers defined in the rstoolbox.components module (Fig. A.1): I) DesignFrame - each row is a designed decoy and the columns represent decoy properties, such as, structural and energetic scores, sequence, secondary structure,



Action	Code Sample

Load

```
import rstoolbox as rs
import matplotlib.pyplot as plt
import seaborn as sns
```

Read

```
# With Rosetta installed, a single structure is scored. The
# function will return multiple score terms, sequence,
# secondary structure and phi/psi angles.
ref = rs.io.get_sequence_and_structure('1kx8_d2.pdb')
```

```
# Loading Rosetta fragments
seqfrags = rs.io.parse_rosetta_fragments('seq.200.9mers')
# With Rosetta, structural similarity of the fragments can be measured
seqfrags = seqfrags. add_quality_measure (None, 'mota_1kx8_d2.pdb')
strfrags = rs.io.parse_rosetta_fragments('str.200.9mers')
strfrags = strfrags. add_quality_measure (None, 'mota_1kx8_d2.pdb')
```

```
# Loading ab initio data
abseq = rs.io.parse_rosetta_file('abinitio_seqfrags.minsilent.gz')
abstr = rs.io.parse_rosetta_file('abinitio_strfrags.minsilent.gz')
```

Plot

```
fig = plt.figure(figsize = (170 / 25.4, 170 / 25.4))
grid = (3, 6)
```

```
# There are 4 flavours of Ramachandran plots available depending on the
# targeted residues: GENERAL, GLY, PRE–PRO and PRO.
ax1 = plt.subplot2grid(grid, (0, 0), colspan = 2)
# Ramachandran is plotted for a single decoy (selected as parameter 1).
# As a decoy can contain multiple chains, the chain identifier is an
# ubiquitous attribute in multiple functions of the library.
rs.plot.plot_ramachandran_single (ref.iloc[0], 'A', ax1)
ax1 = plt.subplot2grid(grid, (0, 2), fig = fig, colspan = 2)
rs. plot.plot_ramachandran_single (ref.iloc[0], 'A', ax1, 'PRE–PRO')
ax1 = plt.subplot2grid(grid, (0, 4), colspan = 2)
rs.plot.plot_ramachandran_single (ref.iloc[0], 'A', ax1, 'PRO')
```

```
# Show RMSD match of fragments to the corresponding sequence for a
# selected region
ax1 = plt.subplot2grid(grid, (1, 0), colspan = 3)
ax2 = plt.subplot2grid(grid, (1, 3), colspan = 3, sharey = ax1)
rs.plot.plot_fragments (seqfrags. slice_region (21, 56),
               strfrags.slice_region(21, 56), ax1, ax2)
rs.utils.add_top_title (ax1, 'sequence–based 9mers')
rs.utils.add_top_title (ax2, 'structure–based 9mers')
```

```
# DataFrames can directly work with widely spread plotting functions
ax1 = plt.subplot2grid(grid, (2, 0), colspan = 3)
sns.scatterplot(x = "rms", y = "score", data = abseq, ax = ax1)
ax2 = plt.subplot2grid(grid, (2, 3), colspan = 3, sharey = ax1, sharex = ax1)
sns.scatterplot(x = "rms", y = "score", data = abstr, ax = ax2)
rs.utils.add_top_title (ax1, 'sequence–based fragments')
rs.utils.add_top_title (ax2, 'structure–based fragments')
```

```
plt.tight_layout()
plt.savefig('BMC_Fig2.png', dpi = 300)
```

Figure A.2: Sample code for the evaluation of protein backbone dihedral angles and fragment quality.

residues of interest among others; II) SequenceFrame - similar to a position-specific scoring matrix (PSSM), obtained from the DesignFrame can be used for sequence and secondary structure enrichment analysis; III) FragmentFrame - stores fragment sets, a key element in Rosetta's *ab initio* folding and loop closure protocols. Derived from pandas.DataFrame [344], all these objects can be casted from and to standard data frames, making them compatible with libraries built for data frame analysis and visualization.

The DesignFrame is the most general data structure of the library. It allows fast sorting and selection of decoys through different scores and evaluation of sequence and structural features. It can be filled with any tabulated, csv or table-like data file. Any table-formatted data can be readily input, as the generation of parsers and integration into the rstoolbox framework is effortless, providing easy compatibility with other CPD software packages, in addition to Rosetta. Currently, rstoolbox provides parsers for FASTA files, CLUSTALW [347] and HMMER [348] outputs, Rosetta's json and silent files (Fig. A.1).

The components of the library can directly interact with most of the commonly used Python plotting libraries such as matplotlib [349] or seaborn [350]. Additional plotting functions, such as logo and Ramachandran plots, are also present to facilitate specific analysis of CPD data. As mentioned, this library has been developed primarily to handle Rosetta outputs and thus, rstoolbox accesses Rosetta functions to extract structural features from designed decoys (e.g. backbone dihedral angles). Nevertheless, many of the rstoolbox's functionalities are independent of a local installation of Rosetta. rstoolbox is configured with a continuous integration system to guarantee a robust performance upon the addition of new input formats and functionalities. Testing covers more than 80% of the library's code, excluding functions that have external dependencies from programs like Rosetta [67], HMMER [348] or CLUSTALW [347]. To simplify its general usage, the library has a full API documentation with examples of common applications and can be directly installed with PyPI (pip install rstoolbox).

A.4 Results

A.4.1 Analysis of protein backbone features

A typical metric to assess the quality of protein backbone conformations is by comparison of the backbone dihedral angles with those of the Ramachandran distributions [346]. Such evaluation is more relevant in CPD strategies that utilize flexible backbone sampling, which have become increasingly used in the field (e.g. loop modelling [351], *de novo* design [326]). A culprit often observed in designs generated using flexible backbone sampling is that the modelled backbones present dihedral angles in disallowed regions of the Ramachandran distributions, meaning that such conformations are likely to be unrealistic. To identify these problematic structures, rstoolbox provides functions to analyse the dihedral angles of decoy sets and represent them in Ramachandran plots (Table A.2, Fig. A.3a).

Furthermore, structural prediction has also become an integral part of many CPD workflows [86]. Here, one evaluates if the designed sequences have energetic propensity to adopt the desired structural conformations. A typical example where prediction is recurrently used as a criterion to select the best designed sequences is on *de novo* design. To assess the ability of novel sequences to refold to the target structures, the Rosetta *ab initio* protocol is typically used [97]. Importantly, the quality of the predictions is critically dependent on the fragment sets provided as input as they are used as local building blocks to assemble the folded three-dimensional structures. The local structural similarity of the fragments to the target structure largely determines the quality of the sampling of the *ab initio* predictions. rstoolbox provides analysis and plotting tools to evaluate the similarity of fragment sets to a target structure (Fig. A.3b). In Fig. A.3c the impact of distinct fragment sets in *ab initio* predictions is shown where a clear folding funnel is visible for fragments with high structural similarity. This tool can also be useful for structural prediction applications to profile the quality of different fragment sets.

A.4.2 Guiding iterative CPD workflows

Many CPD workflows rely on iterative approaches in which multiple rounds of design are performed and each generation of designs is used to guide the next one.

The rstoolbox presents a diversity of functions that aid this process and perform tasks from se-

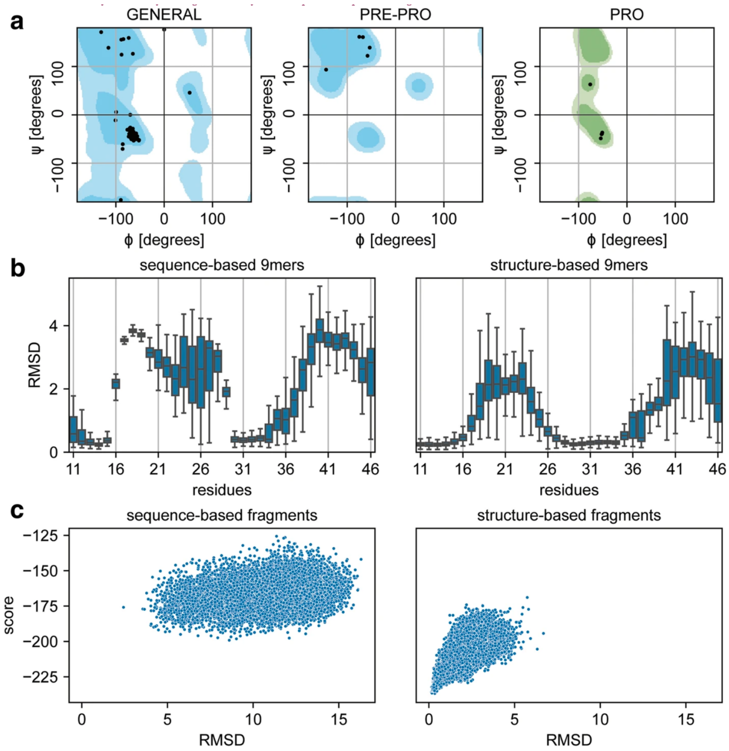


Figure A.3: Ramachandran plots and fragment quality profiles. Assessment of fragments generated using distinct input data and their effect on Rosetta *ab initio* simulations. With the exception of the panel identifiers, the image was created with the code presented in Table A.2. **a:** Ramachandran distribution of a query structure. **b:** Fragment quality comparison between sequence- and structure-based fragments. The plot shows a particular region of the protein for which sequence-based fragments present much larger structural deviations than structure-based fragments in comparison with the query protein. **c:** Rosetta *ab initio* simulations performed with sequence- (left) or structure-based (right) fragments. Fragments with a better structural mimicry relative to the query structure present an improved folding funnel.

lecting decoys with specific mutations of interest, to those that define residue sets for instance based in position weight matrices (generate_mutants_from_matrix()). When redesigning naturally occurring proteins, it also presents a function to generate reversions to wild-type residues (generate_wt_reversions()) to generate the best possible design with the minimal number of mutations. These functions will directly execute Rosetta, if installed in the system, but can also be used to create input files to run the simulations in different software suits. Code example for these functionalities is shown in Table A.5. The result of the code is depicted on Fig. A.4.

rstoolbox allows the user to exploit the data obtained from the analysis of designed populations in order to bias following design rounds. When using rstoolbox, this process is technically simple and clear to other users, which will improve the comprehension and reproducibility of iterative design pipelines.

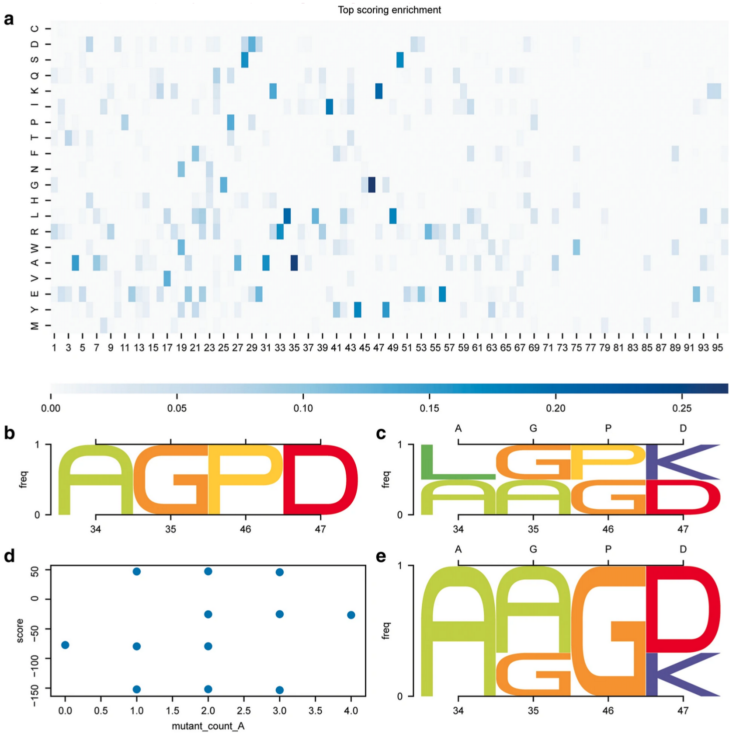
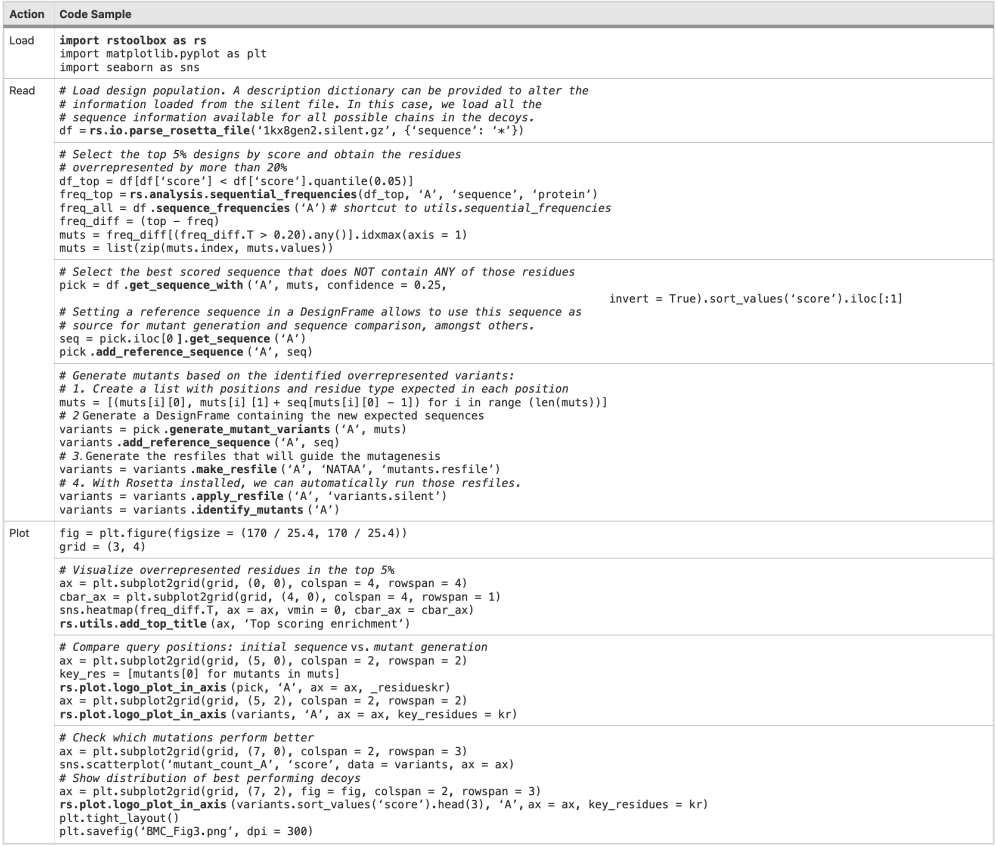


Figure A.4: Guiding iterative design pipelines. Information retrieved from decoy populations can be used to guide following generations of designs. With the exception of the panel identifiers, the image was directly created with the code presented in Table A.5. **a:** Mutant enrichment from comparison of the design on top 5% by score and the overall population. Positions 34, 35, 46 and 47 present a 20% enrichment of certain residue types over the whole population and are selected as positions of interest. **b:** Residue types for the positions of interest in the decoy selected as template of the second generation. **c:** Upon guided mutagenesis, we obtain a total of 16 decoys including the second-generation template. We can observe that the overrepresented residues shown in A are now present in the designed population. Upper x axis shows the original residue types of the template. **d:** Combinatorial targeted mutagenesis yields 16 new designs, three of which showed an improved total score relative to the second-generation template (mutant_count_A is 0). **e:** The three best scoring variants show mutations such as P46G which seem to be clearly favorable for the overall score of the designs. Upper x axis shows the original residue types of the template.

A.4.3 Evaluation of designed proteins

Recently, we developed the Rosetta FunFolDes protocol, which was devised to couple conformational folding and sequence design [312]. FunFolDes was developed to insert functional sites into protein scaffolds and allow for full-backbone flexibility to enhance sequence sampling. As a demonstration of its performance, we designed a new protein to serve as an epitope-scaffold for the Respiratory Syncytial Virus site II (PDB ID: 3IXT [352]), using as scaffold the A6 protein of the Antennal Chemosensory system from Mamestra brassicae (PDB ID: 1KX8 [353]). The designs were obtained in a two-stage protocol, with the second generation being based on the optimization of a small subset of first-generation decoys. The code presented in Table A.7 shows how to process and compare the data of both generations. Extra plotting



Action **Code Sample**

Load

```
import rstoolbox as rs
import matplotlib.pyplot as plt
import seaborn as sns
```

Read

```
# Load design population. A description dictionary can be provided to alter the
# information loaded from the silent file. In this case, we load all the
# sequence information available for all possible chains in the decoys.
df = rs.io.parse_rosetta_file('1kx8gen2.silent.gz', {'sequence': '*'})
```

```
# Select the top 5% designs by score and obtain the residues
# overrepresented by more than 20%
df_top = df[df['score'] < df['score'].quantile(0.05)]
freq_top = rs.analysis.sequential_frequencies(df_top, 'A', 'sequence', 'protein')
freq_all = df .sequence_frequencies ('A') # shortcut to utils.sequential_frequencies
freq_diff = (top − freq)
muts = freq_diff[(freq_diff.T > 0.20).any()].idxmax(axis = 1)
muts = list(zip(muts.index, muts.values))
```

```
# Select the best scored sequence that does NOT contain ANY of those residues
pick = df .get_sequence_with ('A', muts, confidence = 0.25,
                                                                                  invert = True).sort_values('score').iloc[:1]
# Setting a reference sequence in a DesignFrame allows to use this sequence as
# source for mutant generation and sequence comparison, amongst others.
seq = pick.iloc[0 ].get_sequence ('A')
pick .add_reference_sequence ('A', seq)
```

```
# Generate mutants based on the identified overrepresented variants:
# 1. Create a list with positions and residue type expected in each position
muts = [(muts[i][0], muts[i] [1] + seq[muts[i][0] − 1]) for i in range (len(muts))]
# 2 Generate a DesignFrame containing the new expected sequences
variants = pick .generate_mutant_variants ('A', muts)
variants .add_reference_sequence ('A', seq)
# 3. Generate the resfiles that will guide the mutagenesis
variants = variants .make_resfile ('A', 'NATAA', 'mutants.resfile')
# 4. With Rosetta installed, we can automatically run those resfiles.
variants = variants .apply_resfile ('A', 'variants.silent')
variants = variants .identify_mutants ('A')
```

Plot

```
fig = plt.figure(figsize = (170 / 25.4, 170 / 25.4))
grid = (3, 4)
```

```
# Visualize overrepresented residues in the top 5%
ax = plt.subplot2grid(grid, (0, 0), colspan = 4, rowspan = 4)
cbar_ax = plt.subplot2grid(grid, (4, 0), colspan = 4, rowspan = 1)
sns.heatmap(freq_diff.T, ax = ax, vmin = 0, cbar_ax = cbar_ax)
rs.utils.add_top_title (ax, 'Top scoring enrichment')
```

```
# Compare query positions: initial sequence vs. mutant generation
ax = plt.subplot2grid(grid, (5, 0), colspan = 2, rowspan = 2)
key_res = [mutants[0] for mutants in muts]
rs.plot.logo_plot_in_axis (pick, 'A', ax = ax, _residueskr)
ax = plt.subplot2grid(grid, (5, 2), colspan = 2, rowspan = 2)
rs.plot.logo_plot_in_axis (variants, 'A', ax = ax, key_residues = kr)
```

```
# Check which mutations perform better
ax = plt.subplot2grid(grid, (7, 0), colspan = 2, rowspan = 3)
sns.scatterplot('mutant_count_A', 'score', data = variants, ax = ax)
# Show distribution of best performing decoys
ax = plt.subplot2grid(grid, (7, 2), fig = fig, colspan = 2, rowspan = 3)
rs.plot.logo_plot_in_axis (variants.sort_values('score').head(3), 'A', ax = ax, key_residues = kr)
plt.tight_layout()
plt.savefig('BMC_Fig3.png', dpi = 300)
```

Figure A.5: Sample code to guide iterative CPD workflows.

functions to represent experimental data obtained from the biochemical characterization of the designed proteins is also shown. The result of this code is represented in Fig. A.6.

A.4.4 Benchmarking design protocols

One of the main novelties of FunFolDes was the ability to include a binding partner during the folding-design simulations. This feature allows to bias the design simulations towards productive configurations capable of properly displaying the functional motif transplanted to the scaffold. To assess this new feature, we used as a benchmark test the previously computationally designed protein BINDI, a 3-helix bundle that binds to BHRF1 [285]. We performed simulations under four different conditions: no-target (binding-target absent), static (binding-target without conformational freedom), pack (binding-target with side-chain repacking) and packmin (binding-target with side chain repacking and backbone minimization) and evaluated the performance of each simulation. Specifically, we analysed how the design populations performed regarding energetic sampling (Fig. A.8a) and the mimicry of BINDI's conformational shift from the original scaffold (Fig. A.8a). In addition, we quantified the sequence recovery relative to the experimentally characterized BINDI sequence (Fig. A.8b

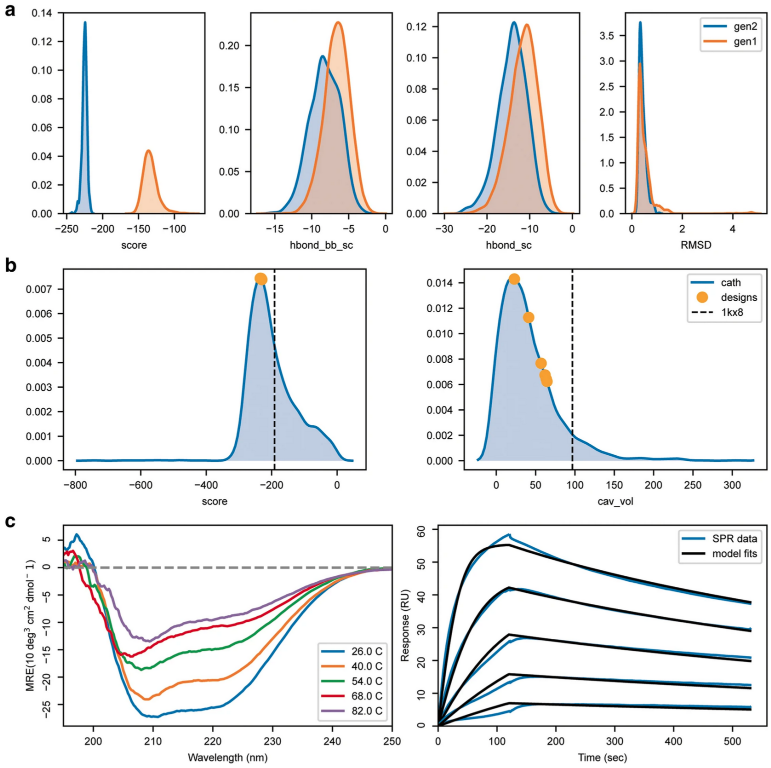
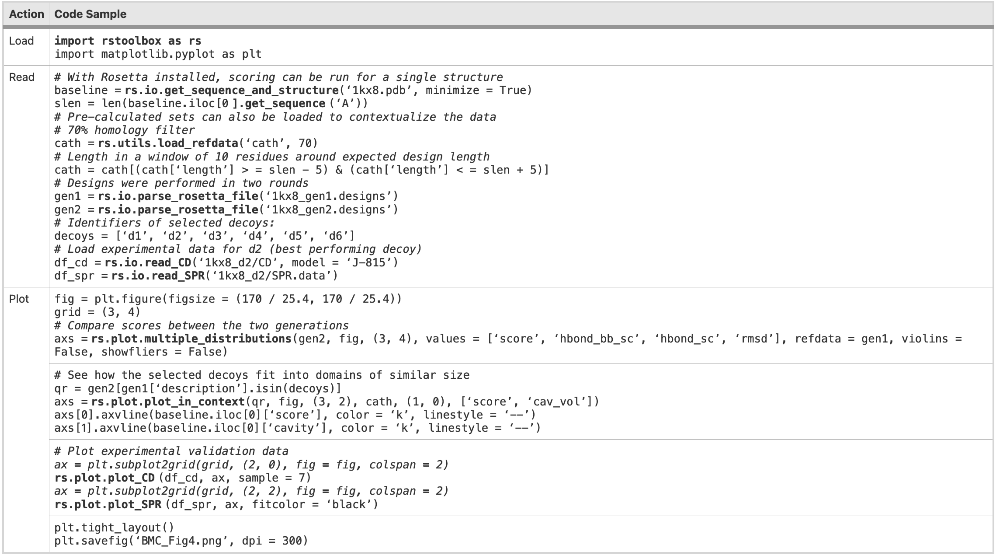


Figure A.6: Multi-stage design. Comparison with native proteins and representation of experimental data for 1kx8-based epitope-scaffold. Analysis of the two-step design pipeline, followed by a comparison of the distributions obtained for native proteins and the designs and plotting of biochemical experimental data. With the exception of the panel identifiers, the image was directly created with the code presented in Table A.7. **a:** Comparison between the first (orange) and the second (blue) generation of designs. score – shows the Rosetta energy score; hbond_bb_sc – quantifies the hydrogen bonds between backbone and side chain atoms; hbond_sc - quantifies the hydrogen bonds occurring between side chain atoms; RMSD – root mean square deviation relative to the original template. Second-generation designs showed minor improvements on backbone hydrogen bonding and a substantial improvement in overall Rosetta Energy. **b:** Score and cavity volume for the selected decoys in comparison with structures of CATH [34] domains of similar size. The vertical dashed black line represents the score and cavity volume of the original 1kx8 after minimization, highlighting the improvements relative to the original scaffold. **c:** Circular Dichroism and Surface Plasmon Resonance data for the best design shows a well folded helical protein that binds with high affinity to the expected target.

and c). Table A.9 exemplifies how to easily load and combine the generated data and create a publication-ready comparative profile between the four different approaches (Fig. A.8).

A.5 Discussion

The analysis of protein structures is an important approach to enable the understanding of fundamental biological processes, as well as, to guide design endeavours where one can alter and improve the activity and stability of newly engineered proteins for a number of



Action | Code Sample

Load

```
import rstoolbox as rs
import matplotlib.pyplot as plt
```

Read

```
# With Rosetta installed, scoring can be run for a single structure
baseline = rs.io.get_sequence_and_structure('1kx8.pdb', minimize = True)
slen = len(baseline.iloc[0 ].get_sequence ('A'))
# Pre-calculated sets can also be loaded to contextualize the data
# 70% homology filter
cath = rs.utils.load_refdata('cath', 70)
# Length in a window of 10 residues around expected design length
cath = cath[(cath['length'] > = slen - 5) & (cath['length'] < = slen + 5)]
# Designs were performed in two rounds
gen1 = rs.io.parse_rosetta_file('1kx8_gen1.designs')
gen2 = rs.io.parse_rosetta_file('1kx8_gen2.designs')
# Identifiers of selected decoys:
decoys = ['d1', 'd2', 'd3', 'd4', 'd5', 'd6']
# Load experimental data for d2 (best performing decoy)
df_cd = rs.io.read_CD('1kx8_d2/CD', model = 'J-815')
df_spr = rs.io.read_SPR('1kx8_d2/SPR.data')
```

Plot

```
fig = plt.figure(figsize = (170 / 25.4, 170 / 25.4))
grid = (3, 4)
# Compare scores between the two generations
axs = rs.plot.multiple_distributions(gen2, fig, (3, 4), values = ['score', 'hbond_bb_sc', 'hbond_sc', 'rmsd'], refdata = gen1, violins =
False, showfliers = False)
```

```
# See how the selected decoys fit into domains of similar size
qr = gen2[gen1['description'].isin(decoys)]
axs = rs.plot.plot_in_context(qr, fig, (3, 2), cath, (1, 0), ['score', 'cav_vol'])
axs[0].axvline(baseline.iloc[0]['score'], color = 'k', linestyle = '--')
axs[1].axvline(baseline.iloc[0]['cavity'], color = 'k', linestyle = '--')
```

```
# Plot experimental validation data
ax = plt.subplot2grid(grid, (2, 0), fig = fig, colspan = 2)
rs.plot.plot_CD (df_cd, ax, sample = 7)
ax = plt.subplot2grid(grid, (2, 2), fig = fig, colspan = 2)
rs.plot.plot_SPR (df_spr, ax, fitcolor = 'black')
```

```
plt.tight_layout()
plt.savefig('BMC_Fig4.png', dpi = 300)
```

Figure A.7: Sample code for the evaluation of a multistep design pipeline.

important applications. In the age of massive datasets, structural data is also quickly growing both through innovative experimental approaches and more powerful computational tools. To deal with fast-growing amounts of structural data, new analysis tools accessible to users with beginner-level coding experience are urgently needed. Such tools are also enabling for applications in CPD, where large amounts of structural and sequence data are routinely generated. Here, we describe and exemplify the usage of rstoolbox to analyse CPD data illustrating how these tools can be used to distil large structural datasets and produce intuitive graphical representations.

CPD approaches are becoming more popular and achieving important milestones in generating proteins with novel functions [334]. However, CPD pipelines remain technically challenging with multiple design and selection stages which are different for every design problem and thus often require user intervention. Within the applications of rstoolbox, several functionalities can aid in this process, by providing an easy programmatic interface to perform selections, comparisons with native proteins, graphical representations and informing follow-up rounds of design in iterative, multi-step protocols. The tools presented here were devised for Rosetta CPD calculations, nevertheless the table-like data structure used allows for the easy creation of parsers for other protein modelling and design tools. This is especially relevant in other modelling protocols that require large sampling such as protein docking [354]. Importantly, rstoolbox can also be useful for structural bioinformatics and the analysis of structural features which have become more enlightening with the growth of different structural databases (e.g. PDB [26], SCOP [355], CATH [34]).

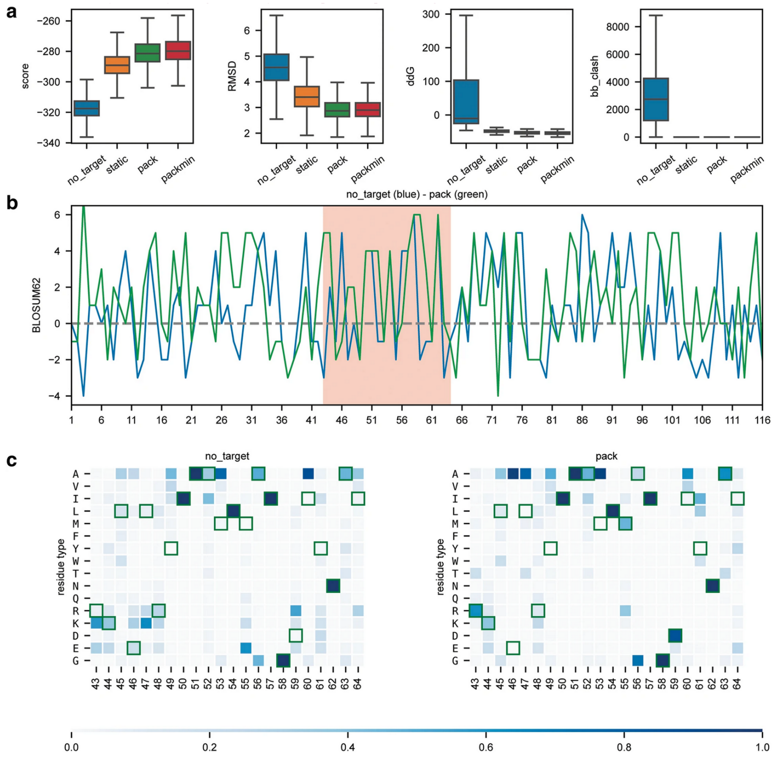
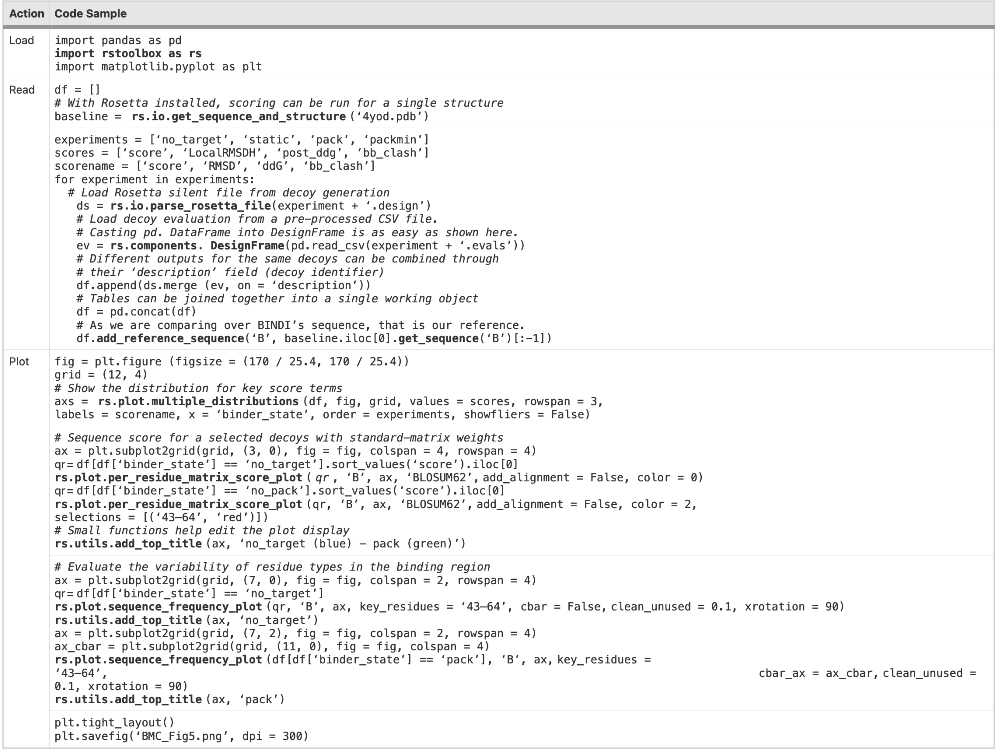


Figure A.8: Comparison and benchmarking of different design protocols. Representation of the results obtained using four different design protocols. With the exception of the panel identifiers, the image was directly created with the code presented in Table A.9. **a:** Representation of four scoring metrics in the design of a new protein binder. score – shows the overall Rosetta score; RMSD – root mean square deviation relative to BINDI; ddG –Rosetta energy for the interaction between two proteins; bb_clash - quantifies the backbone clashes between the binder and the target protein; **b:** BLOSUM62 positional sequence score for the top design of the no_target (blue) and pack (green) design populations showcases how to analyse and compare individual decoys. The higher the value, the more likely two residue types (design vs. BINDI) are to interchange within evolutionary related proteins. Special regions of interest can be easily highlighted, as for instance the binding region (highlighted in salmon). **c:** Population-wide analysis of the sequence recovery of the binding motif region for no_target and pack simulations. Darker shades of blue indicate a higher frequency and green frames indicate the reference residue type (BINDI sequence). This representation shows that the pack population explores more frequently residue types found in the BINDI design in the region of the binding motif.

A.6 Conclusions

Here, we present the rstoolbox, a Python library for the analysis of large-scale structural data tailored for CPD applications and adapted to a wide variety of user expertise. We endowed rstoolbox with an extensive documentation and a continuous integration setup to ensure code stability. Thus, rstoolbox can be accessed and expanded by users with beginner's level programming experience guaranteeing backward compatibility. The inclusion of rstoolbox in design, protocol development and structural bioinformatics pipelines will aid in the comprehension of the human-guided decisions and actions taken during the processing of large structural datasets, helping to ensure their reproducibility.



Action **Code Sample**

Load

```
import pandas as pd
import rstoolbox as rs
import matplotlib.pyplot as plt
```

Read

```
df = []
# With Rosetta installed, scoring can be run for a single structure
baseline =  rs.io.get_sequence_and_structure ('4yod.pdb')
```

```
experiments = ['no_target', 'static', 'pack', 'packmin']
scores = ['score', 'LocalRMSDH', 'post_ddg', 'bb_clash']
scorename = ['score', 'RMSD', 'ddG', 'bb_clash']
for experiment in experiments:
  # Load Rosetta silent file from decoy generation
   ds = rs.io.parse_rosetta_file(experiment + '.design')
   # Load decoy evaluation from a pre-processed CSV file.
   # Casting pd. DataFrame into DesignFrame is as easy as shown here.
   ev = rs.components. DesignFrame(pd.read_csv(experiment + '.evals'))
   # Different outputs for the same decoys can be combined through
   # their 'description' field (decoy identifier)
   df.append(ds.merge (ev, on = 'description'))
   # Tables can be joined together into a single working object
   df = pd.concat(df)
   # As we are comparing over BINDI's sequence, that is our reference.
   df.add_reference_sequence('B', baseline.iloc[0].get_sequence('B')[:-1])
```

Plot

```
fig = plt.figure (figsize = (170 / 25.4, 170 / 25.4))
grid = (12, 4)
# Show the distribution for key score terms
axs =  rs.plot.multiple_distributions (df, fig, grid, values = scores, rowspan = 3,
labels = scorename, x = 'binder_state', order = experiments, showfliers = False)
```

```
# Sequence score for a selected decoys with standard-matrix weights
ax = plt.subplot2grid(grid, (3, 0), fig = fig, colspan = 4, rowspan = 4)
qr= df[df['binder_state'] == 'no_target'].sort_values('score').iloc[0]
rs.plot.per_residue_matrix_score_plot ( qr , 'B', ax, 'BLOSUM62', add_alignment = False, color = 0)
qr= df[df['binder_state'] == 'no_pack'].sort_values('score').iloc[0]
rs.plot.per_residue_matrix_score_plot (qr, 'B', ax, 'BLOSUM62', add_alignment = False, color = 2,
selections = [('43–64', 'red')])
# Small functions help edit the plot display
rs.utils.add_top_title (ax, 'no_target (blue) – pack (green)')
```

```
# Evaluate the variability of residue types in the binding region
ax = plt.subplot2grid(grid, (7, 0), fig = fig, colspan = 2, rowspan = 4)
qr= df[df['binder_state'] == 'no_target']
rs.plot.sequence_frequency_plot (qr, 'B', ax, key_residues = '43–64', cbar = False, clean_unused = 0.1, xrotation = 90)
rs.utils.add_top_title (ax, 'no_target')
ax = plt.subplot2grid(grid, (7, 2), fig = fig, colspan = 2, rowspan = 4)
ax_cbar = plt.subplot2grid(grid, (11, 0), fig = fig, colspan = 4)
rs.plot.sequence_frequency_plot (df[df['binder_state'] == 'pack'], 'B', ax, key_residues =
'43–64',                                                                     cbar_ax = ax_cbar, clean_unused =
0.1, xrotation = 90)
rs.utils.add_top_title (ax, 'pack')
```

```
plt.tight_layout()
plt.savefig('BMC_Fig5.png', dpi = 300)
```

Figure A.9: Sample code for the comparison between 4 different decoy populations.

B Selected sequences

...

Main sequences of chapter 4.

Name	Sequence
3onja_9.1_computational	SLLESYEWSFIVQLILAKLELAYAPSQPLSQRNEQLKRVEQQQDQLFDLLDQMDVEVNNSIGDASERATYKAKLREWKKTIQSDIKRPLQSLVDSG
3onja_9.1_SSM	NLLTSYEGSFKIQLILAKLELAKAPSQPLSQRNEELKRVEQRQDRLFDLLDQMDVEVNNSIGDASERATYKAKLREWKKTIQSDIKRPLQSLVDSG

Selected sequences for experimental characterization of chapter 5.

Name	Sequence
srch1PGX_117	TKIQIHHEQRNQTINISEDDEEKAKREAHELIKKLQVKVEEEQNESREEVHIKNKLEGSGSGSGSGSGSGSWGSGSGSGSGSGSGS
srch1PGX_125	YSMRIQNNSRNEEIRIEDDDKEKLKKLAEEYLRRIKLEYEEHEEEKHDRIEIRIKLEGSGSGSGSGSGSGSWGSGSGSGSGSGSGS
srch1PGX_156	VTLEIKTEQENQETEWRDTDEEKLKRKAKEYVERKQMETEQHENESENRYELRLRLEGSGSGSGSGSGSGSWGSGSGSGSGSGSGS
srch1PGX_66	VEVHLENERHNEKQTYHTTDAERLKRKFEEIYQKKKFDRKEEEENKDEEKVKVRFRLEGSGSGSGSGSGSGSWGSGSGSGSGSGSGS
srch1PGX_8	VQVQIRNEKHNEEINITFGPNQLEEAKKMAKEILKKLQVKQEHEENEDHEELRIRIQLEGSGSGSGSGSGSGSWGSGSGSGSGSGSGS
srch2N75_113	SFVLVLSNNDDEIKEYMKAFSTLVLHFTLMKDNDYKRIVEEAMKHLSELLIIILVTEDEDLLKTWQKAAKKYHNNVEMVQTSTLEEAKKITKK
srch2N75_169	KIIILFVMNKTDEELKELMEKYTKLFFQFTYDPKKDEAKKAIKKAMEIAKKYADNLFIVILTDDETYIRWIEEWMKQMQVNTQLYVTKDWKLVKEVIEK
srch2N75_176	VLIWILLQGYEDEELKKVMRKEEKEVVHFKFSDDEDEVRKVMEKALKEAKKVQSEFLLFIYRLDETAKRIAEELAKRAWDNIQIYTTEDWKEWQKVMEK
srch2N75_18	TYIIVFSTSSDKIHEAWKKAASNLFFFEEKDKSKLEQMIKEAMKLYTEFVFILITEDDDMKRLVREAVKKMQPELRLVETEDPKLAEEYIRK
srch2N75_59	TVVLILMHPTDHWKEIYKKLGEILLMISMTDKEKYIKKAMELLQKYNDELIIIILTEDEDLKKKFEKWVRIIKGEQKLIKISTPEQAEQHLRK
srch3SD2_116	WHLTKDGLTMKVQLSKGDTIHMETTDEQLEYRAEGDSMNVEVRIPKPLTFKVEIKQNGQQLSTEIRLEGSGSGSGSGSGSGSWGSGSGSGSGSGSGS
srch3SD2_155	QVSIEDKTKIRIQLSPDTRIEITINGQNHTFQADKTSRVELQMERPVEIRIKIEEGDKEEHLRVELEGSGSGSGSGSGSGSWGSGSGSGSGSGSGS
srch3SD2_2	QIQFEDKKRLRIEVTQGVEVHIELNGQQLHFKANTNYQIEIQLTNFEEIHVTLRTEENEYHYTLKLEGSGSGSGSGSGSGSWGSGSGSGSGSGSGS
srch3SD2_44	SIEVKDRTHMELEMRKNWEVRVEINGDQREHKGTENDKLEIHIDDPRFIKMKVNEDGKEIEVRVHLEGSGSGSGSGSGSGSWGSGSGSGSGSGSGS
srch3SD2_73	SVEIRDKTNLHITVSKGITVQIEISGTQLRYEAKDDNFNVHVHIEPGLEMRVRIEEGDKEIRIKVKLEGSGSGSGSGSGSGSWGSGSGSGSGSGSGS
srch3SD2_75	EITYRNRSEIEIQLETGMTIEITLNGKELRFQGTDHKDKVEIHDPTMRNIELRVKLPDKHRKYRIELEGSGSGSGSGSGSGSWGSGSGSGSGSGSGS
srch3SD2_8	EIKMEDKTKIHLTFSKGVKLEIQINGRKFEYRATDNINIHMQLSHPAELRVKIEEGSKEIRVKLELEGSGSGSGSGSGSGSWGSGSGSGSGSGSGS
srch6E5C_136	ETVKFELTEPREVEFRIKLTKKIQINVEPGMKMHLRVNDGQVHLEPDESKRLEYTIEDGNEVKIKMETRNIKVTIEE
srch6E5C_150	TTQEIRVTGDNEYTVTVNLTKRIEIEVSKDLTMHVTVNGVEIKFQLGSNVNMEIQLNTEININIQVSRGEVTIKIHD
srch6E5C_26	REYTIEMNEPTRIEVHIKVDDEAPLEIEVRVHYEKVQIKINISGQKFELHAKTENVELRIEFDVLGEITIEVEVPYKVEVHVKM
srch6E5C_62	ETWHMNMSEPTTITLKITMSSEDSITIHLEFTEDVNVEVRVNNINYEFKITRNLKMQVNIETPGEVEIHINYKFSKVTVKI
srch6E5C_77	MSKEMNMKDGSNLEIRYEVKKDGPVEIHIEVEQDLEVHVHISGDQREVHPGPNKKITVKVNVGSNIRIELKYNYITIKVQE
srch6E5C_94	KTWKWNYNDSSTVEFESRITEPGEIEIRITINADNVNVQVSHSSDSVSVSGGTKNMNYEIHKGDSTNFKITVKVQGKVHVELHL
srch1QYS02_2_0006	SLTIRLEIDGFTVEINIEGSDELIEELLKKLLEKLLKLTTVKLTLEVEDTEKIKELFEELAKALKKEGIVTSLESRTKDGKFEFHLHS
srch1QYS02_30_0008	QITVSLEVKGKTETISIERDDEIMEEIIKKLKEELKKDKDVKLTINVEGSDKVAELLKELARALVAEKKGITISFKKKDGKVELHLHF
srch1QYS02_66_0005	TLHLEIKVGDKRVEINIEEDKRSARELVEKLKEELKQGKIKNITINIDGDDEIQELVKELAEEIKKSLDDLTLEIRKKDGKVEIRLHM
srch1QYS02_77_0007	TYTLSLELKKETLEINLEDDDKIAEQLYEELKERLDTSVEIRLTISVELDDRIEELVKELAEELKKEKDEMSLEIRKTKDSLEIRLRF
srch1QYS02_80_0010	SYTISITSNGTTLEINLSESDEIQEEIIERLKELLKQGKIKEFTFESEDNDKLAELIKELAKALAKSGSGSSLSIQKNKDSVRIELHI
srch1QYS02_82_0010	TIQIRLESDDETIEINIEDSPELLEELLERLKELLKKGKLKKITLNVSDSDKVEEIARELYKAARDLVNAKTIESRTNGDRVEVHIHF
srch1QYS02_83_0008	SLTIEIRSKDETVEINVEESPELAEEYLRRAEELLRKFDNVEVRINTNGDDEIEELIKELIKALEKELNVSEQRTEKHGDKSSFEIHM
srch1QYS02_98_0003	TLRIEIQTKDERIEINLEDDDEKLEELMERLKELIKKGKVKELRFNFEDTDKIIELLRELAEALKKIGDLKTVSEETSDDKTHVEIRL
srch1QYS02_9_0007	EITIRLETSDKTLEINISSSDEILEELVERLKELLKEVTTLRLSITINSDDKAERLFRELLKAILKRLNGSNVRIETHGDEVHFQLHS
srch3SD202_10_0005	KLTVTYNKETRRLEITVSPGVKITVELNGKKLTYTIDKDARVRLEITTPGDDLKFRLEFTLDGKTYTYSWE
srch3SD202_126_0005	NTNLKINKDTNEIELSITKGVTITIELEGGKFTVSADPREEVTVKLSSELASAKIRLEVEVPDQTITLEVE
srch3SD202_142_0005	DITVTYNKEEKKLEIKVEPGVKVTLENNGRKTTVEFDPGEEIRIEITDSTGDLKIRIEITIGDKTFTIRIE
srch3SD202_148_0004	KVTIKYNKEENKLEIRLEPGVTITVEINGKKLTYSASSSSEVTIEITDKDPSAKIRLEITVGDKTITIEWE
srch3SD202_163_0006	NVTVKYNKETKTLEIRLEPNVTVTLTFKGKKLEFSVNSSSEFRLTVTLEDDELKLRVEITIGDKTFTWRFE
srch3SD202_177_0004	NITITYNKETNRLEITISPGVKVTLTFNGKKFEFTGTKGDEIHITVSSELASLKIRIEITINGKTITIEME
srch3SD202_17_0004	NVTVTYNKERNKLEINISPGVKVQIEINDKKLTFSGDDSSEIRIEIQLDDPSAKIKLRIEKGDKTITIRIE
srch3SD202_197_0002	NVTLKLNKEENRVEVSLSKDVTLRIELNGVKFEYSGSSGTEVTVEVSSEAAKDKIKVEVTVPDKTITVTLE
srch3SD202_197_0007	EVTLTINKEENRIEITVSKDVTITIELSGKKFTYSGTSGSEVRVEVSSEAAKDTIKLEVTLPDKTFRYEVQ
srch3SD202_1_0005	NVTVTYNEETKTLEIRLKPDSKLTLEFSNGKFEFEFPPGTEVRIEISSSLASAKINIKVTEGDKTITIRME
srch3SD202_1_0006	DVTVTYNEETNELTIKVEPKSSVTIEFENGKFTFTVTPGTELRLEFSKDLASAKITIKVTEGDKTITIRMK
srch3SD202_2_0002	HVTITYNEETNRLELTFEPGMTIKLELKGGKFTVKVDTDQEIRIEVSSELVKLKISLRVEEEDSTKTFRIE
srch3SD202_3_0008	NVTITYNKETNELEISISPGVKITLTFNGKKFEYTVPPGQEFHLKISSDLVKLKIEIEITIPDKTFTWRFE
srch3SD202_44_0010	NVTVTYNKETNTLEIRVEPDTKITIELNGKKITVSGDKEREVRIKIQNEIPDPKIRIEFTTKDKTITIRIE
srch3SD202_53_0010	KYSLTVNKDTKTLELTLEPGFKVTVEISGKKLTYTGSKDEEVRVRVEDEGGKGTIKLEVTVGDETITVRWE
srch3SD202_88_0008	NITVTYNEETNTLEISLSPGVKITIEINGKKLEFSKDGSENLRLRLEESPPDIKLRLEFTVNGKTITIRFE
srch6E5C02_104_0008	TNITVTISEDEEVTVRIEVQKPSKIEIHITKDATVNSRDSSNVTLSTNTTITLSTNLGPSLTITIKKGKLKVTIHL
srch6E5C02_14_0005	TNLTFTLEPNSEITLEISPGKDVSIRLEVTKDATVELRDSDNTTTSSKSTIQLHKPANDSITQTLKEGELKLTLHM
srch6E5C02_156_0009	INVNVSLSKDQTLHMRVSPSTEVSITVTVSKGATMRVTTPDNLTVSGNESFTLTYKKDVEVTISITDGKLKVTITL
srch6E5C02_163_0001	KNLTFNLKESSKVTVKISPFDEIRVTLTLTEDGTVELRTSDNITFSAKESLTLHFHGPVEITVTLKEGSSVLTVSV
srch6E5C02_166_0002	SNFTFELKPGSEYTITLSPGKPISITIEIKSDATVESRLSNNTTVSDKSTITITSTLNDSVQITVKVGEVKITVTI
srch6E5C02_173_0005	TNLEFEIKKGEEITVQLEGGKPISVTLHITGSGTLEFRTEDNITVTSNSTVTFSSNPNATVTLTVKKGEVKITLTH

Main sequences of of chapter 6.

Name	Sequence
native_1	IVHVFFLRPNDEDIEKRLREELERHQKGGGDKFEIWWLTVNGSEFEEKAKKLLKKLKQYGLVIFIILGDSEIQKAAKKVEELAKEVQVFILQINVDGDPSIAEKVEKEAKKKIH
native_2	IRIELHEHGSHVTVKVELEPGLQLKLEVTSDPGNKVSSSSLDKGTSSYEVTVSGTTLEIEIHNPRHQETIRQEY
native_3	EVEVHLHKHGDKIEVEVEIQLKTDNQEIRFEFQSSSDGASTSYYSIQLGKGGKLTIRIEFGDHVPITIKLSETGPAGEQLSTTQIQ
native_4	RVQLYKNEDKLKLTVELDGTEVTIHIKIEGLSSQSVTDSTGGSVTLEISSGLKGSVTVTIHGKNETIHLK
native_5	ILFIFYSDETEKLFEEIKKKSQKEKQFEFHKFEFTSSEEAKKIIKEIIKRWGEEEIIVIVITHDPRQEEAAKEAAKEVPAQKVIVVKVDGNDEKWKEKIKELIEK
native_6	TWYLILTEVNKKYLEEIIKKTQKHKEIHIIEIRVSSNEELEKVIKELIKRYLGDEIILIILTNTEELKQIAERAAREAGLEKFLVVEQDPSDEKAKEQIEELVKR
native_7	QRVELQNTEERITVTVQLTGNVKIEIEVHVEGGHTKSKSDTDSKQLTVTWIPDSYEITVRVHGENGSEVQKLQ
native_8	QSITIEVGRRKESLELHIELGPNYSIEIRFEIDGESLEIHIEVHIGNKTKHYTYQIKNGKQTIHITLDTHGPITFRVRLTKTELRLTVHRY
native_9	LIKENIDGSSIKITIPGGGGTITIRVSKDATLHIYNGRQSSTLTGKDSLQITITINDDLTIEVHIQGNVTLELR
native_10	STLQFEPGKTQTLSDQVDSKSTVVTITISSGSITIQVHFGGSARLSYEFSDGDGGVTVTLTVHEGVEVEYRWTGGDGSVEIHS
native_11	IRFEFKEGTNKFTIEVEVSGGEQVTIELHTEGDETIRVTLTGDTKIKLEFSSKPLEFHIEHHTKDRTEHTTLY
native_12	DEVHIQLNIGGDDFNWTWHYRTDDVSEALEKAKRIAEKYANGPATSFKVREKQAPDDKIRQVEITV
native_13	NIKFKLENNSQEFEITSGTGDKIREVFEKIVRKLLSLPTKLTIEIHLSKDESEEIARALEEIAKRINPNQEQETETESDNRITLKLI
native_14	LSVELQITTEGPTEIRFSTSIPDDGGGDELLKKIVQKWIEELQRRLSPKGSSYQLEKKQEGDEHKIQVKL
native_15	WTITITVEKRDGESETRTENVGPNDTEVQKWVKKIVKEFVKDGPLTIQVTVTLDGNGDKVLQLVLEAAKEAIDDLGDSLEQETRQDENGQRTSTIRIK
native_16	VTIKITIHKSDDTTEQYQYTFSFEGTPFAEVAAEVVKRYKNVPQVEIHIEVELDSEENKLIELWEKAARKVADGFATSKSTKEHQDHNGTKHIQFHFL
native_17	FEIEVQQHGDTYEVRLKTEPNATIRITITSENGQTFTENKEPTEKITVHVSSGKVEIELRITTKDGTYTNKYK
native_18	LSVTVHIQHGDHPSETYQYSSSAEGTKVVEWALEVIERLEDYGQIRIEVTIKVDGPGDEIIEILQKVLHKVKGKVGGSYHYERHSDEEGNSTITLKFY
native_19	IYQHDSKSDGNITVQVTVNLGGGGTVRLRWKVKIDHHAEVHLEFNVNRVHKYSQSQTVHGTGELELELKVRLDNGVSITLTVHGPHGTIEVEVHIE
native_20	FFIIVISDGTTKWLHELIEIWRRRYKGPIELITDKIDPRDEKKIRELAHEFAKRVSGKIVFLIYIGDETHRIAEVVEEALRKVLPVPVILLKFSDPKDAIEIALKLIEKYLK
denovo_1	IEELARKYIEQFRGGNKVLLVFIEEKDARRAAEIAKKELKKLLGEVLVIIGPSDEILEIAKELVKKEQATYLIFFIDKDPRIEDKVKKAKKEIE
denovo_2	FEKAQKWIREILEKGTEVLVVIFIEDEVQKEVEELLKKVEGSGQNFLVFPGNDKEIAERAAEEAVKWSGIIIIFLINDVTVIEVGGDEAKKKIRELFKKLI
denovo_3	IEELAHIVLELAKQGIKVVILIFYPTVYQKLQKILKELKIEALLQIIVVPKTDKEAIREWIERLAENAQLILFLTEGRVIQIENTNTKARQEYEELIRKLQ
denovo_4	EEYLERLVREWIKKSDGEIVIIFIVVGSEDAEKEAKKAVEIIRRLIGWTVYLYRVSEGATDQVKKIIKELLKKLQNYIYIIIIIISDGPGNQIKIFVFTGPDEEREKEWEEAWKQD
denovo_5	YVSINGRPEDAKKQLQEILKKGGEVEVELSYEGGGDNDKRIKWILKVIEELVRAGGEFIFLVEVFTKDKVKDLVEELRKILLILIIIHTNNKGNFTTSYILTKGKDTKEVKEKAKKATKKVIKKAQKD
denovo_6	DEEEARRIAKLVKDGSLVLLVFFNGSDAEKQAEKIKQIIEKVIGSVIVISGPTEELAKIVQKIVKTYNVHYLFIWVDSDPSLQDYAKKVKERAE
denovo_7	EEEAARHIAELYRRGIDIFVVITLTTVAKKVHEIVRKLKVEKVLEIQEVPTTDKKLIEEILREAAKKWEVVVVLFKDQIITITNKNTEAKKQAEEAIKKTL
denovo_8	VKAEEVVEEVWHRYKHHKVLFILFVTHTEDAKKWAKIAKKRLHELGVEEVRIIELEDEESWKKAIEYVQKQIKKTKDGYIVIFFIIKQENSSFKIFILVLTTDHEKQLKELEEKLE
denovo_9	SINVHGSAKDAAKLIQELLKKGGHVEIQVHFEIGGDTEKAYRKVAELIKLLLELGPKLRFTFEVTTEDLARTIAELWARYVALVIVKFTSSKGSLLTYVIGSEGENTKEAQEYVKKAQKKLEEELKKK
denovo_10	AEEWAERWRKIFKDGKKFVLFFFDKENLEREARKAVKIAQEKVGSIEVLIDHTEELAKKIKEIVKKKQVEVLFFFWSSDPRIETKIRKWQKEIQ
denovo_11	IEEVVEIAVQLLRKGITVWIVIFYQTIAEKIEKLIRKKKAKDITEIQTWQETDKELIRKILEEAAEKADLVVFVKEGEVEVIQHGNDKVKKEIKKLAEKWE
denovo_12	EELLHKWVQEIVKKSDGKFLFIFIVLGDKDVEEIIEKIVEYIRKVFGHTVILFKISKGTTEQIKKIIKEILKKIQEYIVIFVFILTDGNGKQIKIILITGQDEEIEKYIEELLKQL
denovo_13	VEEAKKTIREWLEKGVRVIIVVFVDSEVQEEIRKLIEKVEGSGYSLIVLPTNDKELIKKYWKKLARDPGWLIFIHKDTIITVKLSGDEAQKKLEELAERKA
denovo_14	DEELAKRAEKLAKDGDLIIFVFFDEEDARKIAEKIKKYVEQTLGSVYVIVGPTEEALKIAKEIFKKHNFKLIFLWFTSDPRQETKIKQAKKHIQ
denovo_15	IREVAEELRKKIDEGSIILIVIIEQETAEKELREVVKLLQKKFGSIWVISGSTDEVLKKAKEYFHKWNVQYIFFFWSSDPRQETKVKRVWKEIQ
denovo_16	QIEFQGKVDDVLKILKEIKKHGGSVKITVKVSSGGSSEQQIKLLIKLVKKLFEHGPELILEVEVTTTEDAKTLSKLLEKYVFVIITNVSQSTGNVEVNFIGSLGRNTKKVEEIIKRAQKEIQEKVRKE
denovo_17	KEYVKKILKEIYKKSDGEFIILFLVTGQEDVKKWAREAVKLARELAGVTVIFIEITKGETELIVKYIQKILKEFKHAILLYIFLFFEGPGNQVKVIVFDGRTEELEKIVREYLKQL
denovo_18	ELLQEVLKRLAEEYSSTKIRIIIIITKDESIKKLAKQAIKILKKIGIEEIELIEVNTDENIKKILEELEERLKKTDDGIWIIVIIVAKDNSSYSLVIILVGSKEKEQIEELKKKAK
denovo_19	KEEIERVLREIAEKQDVTLVVFLSDTLVEEAKEAAKRVWHPKHNIVFVTGPDDERVVKEFVKAWKKYPGWVWIFIDPSAKELKEKIEEAVKK
denovo_20	KEELKRIFEETWRRSPAVIIIYLTSTLAEEAREILHQVLREEHNIFVFTEPNHEEVIKKFVEAVKRYPYDFYIFLDPDAQRLKKIIEERVKK
denovo_21	YEVKEIAEEVARRYKDTKLTFIFFVTNDETAKRIAKEAAKLLHKLGVERVEIYELNTEESLKKILKKFKEWLQKSEDGLWIVFFIIIHENSSYSIWWLISGTDEKELAEKIYKHLK
denovo_22	HEELKRAWEEIVKKEDTTILLFLASTLAEEAREIIERLLRDKNNIWLFTDPNDERIWHEIAEQWKKIPYKVWIFVDPDAQELAKKVEEWVKK
denovo_23	LEEIIKKLQEQAHKQEDLILIIILVGRTAEELVQEALEKIKEVVKLIFITVLQTTEELEKLVEIAKKLTGGKVILFIVVGNDSTFVIHLDKDDEEKAKKIIEKLVK
denovo_24	IEEVKKYIKEALKKGQPLLIIIFHDSKIQKEVEEALREVEGDGKKYKTFTANDKENVQEIWRRVARDPGWLLFIFENEVYIFKVTGSDAEKYLHELAKKYA
denovo_25	DEEAAKRAQKAIKDGNKIVLFFVVTETAHKVAEKILKIVQKVAGEVILVTGDTDRALKKFKELVKKWNVQVVVFWFDHDPSIRSKVEEWLKQAR
denovo_26	IKLQEVIEEYVRKYKDEQLIFFFLITRDETAEKYAQEARKTAHKLGVEEVRIIKLNDDRSIEEILKKIEEVARKVPHGKVVIFILKLHENSSYKLVVITITSDREKQLEEALKKLE
denovo_27	VEEVLKKLEERLRKQKDIVVIVLVGRTAKEKVKEVLQRVKEKVRIEVYELITTSEQLEELVKIAQKLLGGEVWIFFQVGNDSFYVIEIEKDRKEEAERQAKKYIK
denovo_28	AEEIAKEIKELLEKQEDIKIIILVADTAEEILRRALEAVKEKVQFEIYKILTTTDQIEKVKEVAQKLLGGIVYVVIIFGNSSYYELKLEKDREEEFKELLKQYKK
denovo_29	KKFEEYIEELARKYKETDILFIILVSKTEDLRELAQQAARIAHEIGIKEVIIIEIKTEENLQRATKIAEEIIKKTNSGIIFLFIISKTDNSSFKVYYLTLPSDREKEIEEYIKKAR
denovo_30	KEWLKKLIEEIVKKSDGKIIFVFIVVGDEDAHEIVKEIVEYARKAGDVEVTVFHISKGYTDIIAEIVEKLLKKFKDYIVLYLFIIIDGNGEQIKIVLFDGRTEKLEEIVKKYIKRT
denovo_31	FEEVAKRIVKKFKEGSLILIVFVDGDDARKEAEKAKEILKKYLGSVLFLIDSTDEALKKAKEVIKKYQFEYLFWIFSQDPSLQDKIEKVKREAH
denovo_32	KEEIERIIKRTVKEKDTSIVFVVASTAEEEVREIAHEALREKNNIFIYTDPNDERVFKAFAQWAKKLPGEVIFFLDDDAKELWKKVEEWIKK

Bibliography

[1] C. Chothia, J. Gough, C. Vogel, and S. A. Teichmann. “Evolution of the protein repertoire”. *Science* 300.5626 (2003), pp. 1701–1703.

[2] M. Levitt. “Growth of novel protein structural data”. *Proc Natl Acad Sci U S A* 104.9 (2007), pp. 3183–3188.

[3] C. Chothia and J. Gough. “Genomic and structural aspects of protein evolution”. *Biochem J* 419.1 (2009), pp. 15–28.

[4] V. Sasisekharan and G. N. Ramachandran. “Studies on collagen”. *Proc Indian Acad Sci* 45.6 (1957), pp. 363–376.

[5] A. R. Fersht and L. Serrano. “Principles of protein stability derived from protein engineering experiments”. *Curr Opin Struct Biol* 3.1 (1993), pp. 75–83.

[6] J. Janin and S. Wodak. “Conformation of amino acid side-chains in proteins”. *J Mol Biol* 125.3 (1978), pp. 357–386.

[7] F. M. Richards. “Areas, volumes, packing and protein structure”. *Annu Rev Biophys Bioeng* 6 (1977), pp. 151–176.

[8] M. J. McGregor, S. A. Islam, and M. J. E. Sternberg. “Analysis of the relationship between side-chain conformation and secondary structure in globular proteins”. *J Mol Biol* 198.2 (1987), pp. 295–310.

[9] R. L. Dunbrack and M. Karplus. “Backbone-dependent rotamer library for proteins. Application to side-chain prediction”. *J Mol Biol* 230.2 (1993), pp. 543–574.

[10] R. L. Dunbrack and F. E. Cohen. “Bayesian statistical analysis of protein side-chain rotamer preferences”. *Protein Sci* 6.8 (1997), pp. 1661–1681.

[11] C. B. Anfinsen, E. Haber, M. Sela, and F. H. White. “The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain”. *Proc Natl Acad Sci U S A* 47 (1961), pp. 1309–1314.

[12] C. Levinthal. “Levinthal’s paradox”. *Proceedings of a Meeting held at Allerton House, Monticello, IL*. 1969, pp. 22–24.

[13] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes. “Funnels, pathways, and the energy landscape of protein folding: A synthesis”. *Proteins: Struct Funct Genet* 21.3 (1995), pp. 167–195.

[14] G. P. Brady and K. A. Sharp. “Entropy in protein folding and in protein-protein interactions”. *Curr Opin Struct Biol* 7.2 (1997), pp. 215–221.

[15] C. Levinthal. “Are there pathways for protein folding?” *J Chim Phys* 65 (1968), pp. 44–45.

[16] S. J. Fleishman and D. Baker. “Role of the Biomolecular Energy Gap in Protein Design, Structure, and Evolution”. *Cell* 149.2 (2012), pp. 262–273.

[17] K. A. Dill. “Dominant forces in protein folding”. *Biochemistry* 29.31 (1990), pp. 7133–7155.

[18] A. Fersht. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*. Vol. 9. Series in Structural Biology. 2017.

[19] I. Halperin, H. Wolfson, and R. Nussinov. “Protein-Protein Interactions: Coupling of Structurally Conserved Residues and of Hot Spots across Interfaces. Implications for Docking”. *Structure* 12.6 (2004), pp. 1027–1038.

[20] G. McLendon and E. Radany. “Is protein turnover thermodynamically controlled?” *J Biol Chem* 253.18 (1978), pp. 6335–6337.

[21] D. A. Parsell and R. T. Sauer. “The structural stability of a protein is an important determinant of its proteolytic susceptibility in Escherichia coli”. *J Biol Chem* 264.13 (1989), pp. 7590–7595.

[22] S. Warszawski, R. Netzer, D. S. Tawfik, and S. J. Fleishman. “A “Fuzzy”-Logic Language for Encoding Multiple Physical Traits in Biomolecules”. *J Mol Biol* 426.24 (2014), pp. 4125–4138.

[23] R. A. Goldstein. “The evolution and evolutionary consequences of marginal thermostability in proteins”. *Proteins* 79.5 (2011), pp. 1396–1407.

[24] The UniProt Consortium. “UniProt: the universal protein knowledgebase in 2021”. *Nucleic Acids Res* 49.D1 (2021), pp. D480–D489.

[25] P. G. Wolynes. “Evolution, Energy Landscapes and the Paradoxes of Protein Folding”. *Biochimie* 119 (2015), pp. 218–230.

[26] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. “The Protein Data Bank”. *Nucleic Acids Res* 28.1 (2000), pp. 235–242.

[27] NCBI. “Database resources of the National Center for Biotechnology Information”. *Nucleic Acids Res* 44.Database issue (2016), pp. D7–D19.

[28] C. Chothia. “One thousand families for the molecular biologist”. *Nature* 357.6379 (1992), pp. 543–544.

[29] C. A. Orengo, D. T. Jones, and J. M. Thornton. “Protein superfamilles and domain superfolds”. *Nature* 372.6507 (1994), pp. 631–634.

[30] A. F. W. Coulson and J. Moult. “A unifold, mesofold, and superfold model of protein fold use”. *Proteins* 46.1 (2002), pp. 61–71.

[31] S. Govindarajan, R. Recabarren, and R. A. Goldstein. “Estimating the total number of protein folds”. *Proteins* 35.4 (1999), pp. 408–414.

[32] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. “SCOP: a structural classification of proteins database for the investigation of sequences and structures”. *J Mol Biol* 247.4 (1995), pp. 536–540.

[33] A. Andreeva, E. Kulesha, J. Gough, and A. G. Murzin. “The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures”. *Nucleic Acids Research* 48.D1 (2020), pp. D376–D382.

[34] I. Sillitoe, N. Dawson, T. E. Lewis, S. Das, J. G. Lees, P. Ashford, A. Tolulope, H. M. Scholes, I. Senatorov, A. Bujan, F. Ceballos Rodriguez-Conde, B. Dowling, J. Thornton, and C. A. Orengo. “CATH: expanding the horizons of structure-based functional annotations for genome sequences”. *Nucleic Acids Res* 47.D1 (2019), pp. D280–D284.

[35] I. Sillitoe, N. Bordin, N. Dawson, V. P. Waman, P. Ashford, H. M. Scholes, C. S. M. Pang, L. Woodridge, C. Rauer, N. Sen, M. Abbasian, S. Le Cornu, S. D. Lam, K. Berka, I. H. Varekova, R. Svobodova, J. Lees, and C. A. Orengo. “CATH: increased structural coverage of functional space”. *Nucleic Acids Res* 49.D1 (2021), pp. D266–D273.

[36] R. Kolodny, L. Pereyaslavets, A. O. Samson, and M. Levitt. “On the universe of protein folds”. *Annu Rev Biophys* 42 (2013), pp. 559–582.

[37] A. N. Barclay. “Ig-like domains: Evolution from simple interaction molecules to sophisticated antigen recognition”. *Proc Natl Acad Sci U S A* 96.26 (1999), pp. 14672–14674.

[38] P. A. Romero and F. H. Arnold. “Exploring protein fitness landscapes by directed evolution”. *Nat Rev Mol Cell Biol* 10.12 (2009), pp. 866–876.

[39] F. H. Arnold. “Directed Evolution: Bringing New Chemistry to Life”. *Angew Chem* 57.16 (2018), pp. 4143–4148.

[40] P. N. Bryan. “Protein engineering”. *Biotechnol Adv* 5.2 (1987), pp. 221–224.

[41] B. A. van den Berg, M. J. T. Reinders, J.-M. van der Laan, J. A. Roubos, and D. de Ridder. “Protein redesign by learning from data”. *Protein Eng Des Sel* 27.9 (2014), pp. 281–288.

[42] P.-S. Huang, S. E. Boyken, and D. Baker. “The coming of age of de novo protein design”. *Nature* 537.7620 (2016), pp. 320–327.

[43] K. E. Drexler. “Molecular engineering: An approach to the development of general capabilities for molecular manipulation”. *Proc Natl Acad Sci U S A* 78.9 (1981), pp. 5275–5278.

[44] B. Gutte, M. Däumigen, and E. Wittschieber. “Design, synthesis and characterisation of a 34-residue polypeptide that interacts with nucleic acids”. *Nature* 281.5733 (1979), pp. 650–655.

[45] J. S. Richardson and D. C. Richardson. “Some design principles: Betabellin”. *Protein Eng* (1987), pp. 149–163.

[46] T. P. Quinn, N. B. Tweedy, R. W. Williams, J. S. Richardson, and D. C. Richardson. “Betadoublet: de novo design, synthesis, and characterization of a beta-sandwich protein”. *Proc Natl Acad Sci U S A* 91.19 (1994), pp. 8747–8751.

[47] S. Kamtekar, J. M. Schiffer, H. Xiong, J. M. Babik, and M. H. Hecht. “Protein design by binary patterning of polar and nonpolar amino acids”. *Science* 262.5140 (1993), pp. 1680–1685.

[48] W. F. DeGrado and J. D. Lear. “Induction of peptide conformation at apolar water interfaces. 1. A study with model peptides of defined hydrophobic periodicity”. *J Am Chem Soc* 107.25 (1985), pp. 7684–7689.

[49] S. P. Ho and W. F. DeGrado. “Design of a 4-helix bundle protein: synthesis of peptides which self-associate into a helical protein”. *J Am Chem Soc* 109.22 (1987), pp. 6751–6758.

[50] P. B. Harbury, B. Tidor, and P. S. Kim. “Repacking protein cores with backbone freedom: structure prediction for coiled coils”. *Proc Natl Acad Sci U S A* 92.18 (1995), pp. 8408–8412.

[51] J. W. Ponder and F. M. Richards. “Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes”. *J Mol Biol* 193.4 (1987), pp. 775–791.

[52] J. R. Desjarlais and T. M. Handel. “New strategies in protein design”. *Curr Opin Biotechnol* 6.4 (1995), pp. 460–466.

[53] B. I. Dahiyat and S. L. Mayo. “De Novo Protein Design: Fully Automated Sequence Selection”. *Science* 278.5335 (1997), pp. 82–87.

[54] D. N. Woolfson. “The design of coiled-coil structures and assemblies”. *Adv Protein Chem* 70 (2005), pp. 79–112.

[55] D. Eisenberg, W. Wilcox, S. M. Eshita, P. M. Pryciak, S. P. Ho, and W. F. DeGrado. “The design, synthesis, and crystallization of an alpha-helical peptide”. *Proteins* 1.1 (1986), pp. 16–22.

[56] S. T. R. Walsh, H. Cheng, J. W. Bryson, H. Roder, and W. F. DeGrado. “Solution structure and dynamics of a de novo designed three-helix bundle protein”. *Proc Natl Acad Sci U S A* 96.10 (1999), pp. 5486–5491.

[57] Y. Zhu, D. O. V. Alonso, K. Maki, C.-Y. Huang, S. J. Lahr, V. Daggett, H. Roder, W. F. DeGrado, and F. Gai. “Ultrafast folding of alpha3D: a de novo designed three-helix bundle protein”. *Proc Natl Acad Sci U S A* 100.26 (2003), pp. 15486–15491.

[58] S. Park, Y. Xu, X. F. Stowell, F. Gai, J. G. Saven, and E. T. Boder. “Limitations of yeast surface display in engineering proteins of high thermostability”. *Protein Eng Des Sel* 19.5 (2006), pp. 211–217.

[59] Y. Maruyama and A. Mitsutake. “Stability of Unfolded and Folded Protein Structures Using a 3D-RISM with the RMDFT”. *J Phys Chem B* 121.42 (2017), pp. 9881–9885.

[60] A. R. Thomson, C. W. Wood, A. J. Burton, G. J. Bartlett, R. B. Sessions, R. L. Brady, and D. N. Woolfson. “Computational design of water-soluble α-helical barrels”. *Science* 346.6208 (2014), pp. 485–488.

[61] S. Nautiyal, D. N. Woolfson, D. S. King, and T. Alber. “A designed heterotrimeric coiled coil”. *Biochemistry* 34.37 (1995), pp. 11645–11651.

[62] G. Grigoryan, A. W. Reinke, and A. E. Keating. “Design of protein-interaction specificity affords selective bZIP-binding peptides”. *Nature* 458.7240 (2009), pp. 859–864.

[63] N. A. Pierce and E. Winfree. “Protein Design is NP-hard”. *Protein Eng Des Sel* 15.10 (2002), pp. 779–782.

[64] E. Marcos and D.-A. Silva. “Essentials of de novo protein design: Methods and applications”. *Wiley Interdiscip Rev Comput Mol Sci* 8.6 (2018), e1374.

[65] J. Desmet, M. D. Maeyer, B. Hazes, and I. Lasters. “The dead-end elimination theorem and its use in protein side-chain positioning”. *Nature* 356.6369 (1992), pp. 539–542.

[66] P. Gainza, K. E. Roberts, I. Georgiev, R. H. Lilien, D. A. Keedy, C.-Y. Chen, F. Reza, A. C. Anderson, D. C. Richardson, J. S. Richardson, and B. R. Donald. “OSPREY: Protein Design with Ensembles, Flexibility, and Provable Algorithms”. *Methods Enzymol* 523 (2013), pp. 87–107.

[67] J. K. Leman et al. “Macromolecular modeling and design in Rosetta: recent methods and frameworks”. *Nat Methods* 17.7 (2020), pp. 665–680.

[68] R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O’Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel, J. W. Labonte, M. S. Pacella, R. Bonneau, P. Bradley, R. L. Dunbrack, R. Das, D. Baker, B. Kuhlman, T. Kortemme, and J. J. Gray. “The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design”. *J Chem Theory Comput* 13.6 (2017), pp. 3031–3048.

[69] H. Kawai, T. Kikuchi, and Y. Okamoto. “A prediction of tertiary structures of peptide by the Monte Carlo simulated annealing method”. *Protein Eng Des Sel* 3.2 (1989), pp. 85–94.

[70] W. K. Hastings. “Monte Carlo sampling methods using Markov chains and their applications”. *Biometrika* 57.1 (1970), pp. 97–109.

[71] J. A. Davey and R. A. Chica. “Multistate approaches in computational protein design”. *Protein Sci* 21.9 (2012), pp. 1241–1252.

[72] T. H. LaBean, S. A. Kauffman, and T. R. Butt. “Libraries of random-sequence polypeptides produced with high yield as carboxy-terminal fusions with ubiquitin”. *Mol Divers* 1.1 (1995), pp. 29–38.

[73] A. R. Davidson and R. T. Sauer. “Folded proteins occur frequently in libraries of random amino acid sequences”. *Proc Natl Acad Sci U S A* 91.6 (1994), pp. 2146–2150.

[74] G. J. Rocklin, T. M. Chidyausiku, I. Goreshnik, A. Ford, S. Houliston, A. Lemak, L. Carter, R. Ravichandran, V. K. Mulligan, A. Chevalier, C. H. Arrowsmith, and D. Baker. “Global analysis of protein folding using massively parallel design, synthesis, and testing”. *Science* 357.6347 (2017), pp. 168–175.

[75] R. Chowdhury, N. Bouatta, S. Biswas, C. Rochereau, G. M. Church, P. K. Sorger, and M. N. AlQuraishi. “Single-sequence protein structure prediction using language models from deep learning”. *bioRxiv* (2021), p. 2021.08.02.454840.

[76] P.-S. Huang, K. Feldmeier, F. Parmeggiani, D. A. Fernandez Velasco, B. Höcker, and D. Baker. “De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy”. *Nat Chem Biol* 12.1 (2016), pp. 29–34.

[77] J. Dou, A. A. Vorobieva, W. Sheffler, L. A. Doyle, H. Park, M. J. Bick, B. Mao, G. W. Foight, M. Y. Lee, L. A. Gagnon, L. Carter, B. Sankaran, S. Ovchinnikov, E. Marcos, P.-S. Huang, J. C. Vaughan, B. L. Stoddard, and D. Baker. “De novo design of a fluorescence-activating betaDe novo design of a non-local β-sheet protein with high stability and accuracy-barrel”. *Nature* 561.7724 (2018), pp. 485–491.

[78] A. R. D. Voet, H. Noguchi, C. Addy, D. Simoncini, D. Terada, S. Unzai, S.-Y. Park, K. Y. J. Zhang, and J. R. H. Tame. “Computational design of a self-assembling symmetrical beta-propeller protein”. *Proc Natl Acad Sci U S A* 111.42 (2014), pp. 15102–15107.

[79] C. W. Wood, M. Bruning, A. Á. Ibarra, G. J. Bartlett, A. R. Thomson, R. B. Sessions, R. L. Brady, and D. N. Woolfson. “CCBuilder: an interactive web-based tool for building, designing and assessing coiled-coil protein assemblies”. *Bioinformatics* 30.21 (2014), pp. 3029–3035.

[80] P.-S. Huang, G. Oberdorfer, C. Xu, X. Y. Pei, B. L. Nannenga, J. M. Rogers, F. DiMaio, T. Gonen, B. Luisi, and D. Baker. “High thermodynamic stability of parametrically designed helical bundles”. *Science* 346.6208 (2014), pp. 481–485.

[81] T. J. Brunette, F. Parmeggiani, P.-S. Huang, G. Bhabha, D. C. Ekiert, S. E. Tsutakawa, G. L. Hura, J. A. Tainer, and D. Baker. “Exploring the repeat protein universe through computational protein design”. *Nature* 528.7583 (2015), pp. 580–584.

[82] C. W. Wood, J. W. Heal, A. R. Thomson, G. J. Bartlett, A. Á. Ibarra, R. L. Brady, R. B. Sessions, and D. N. Woolfson. “ISAMBARD: an open-source computational environment for biomolecular analysis, modelling and design”. *Bioinformatics* 33.19 (2017), pp. 3043–3050.

[83] J. U. Bowie and D. Eisenberg. “An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function”. *Proc Natl Acad Sci U S A* 91.10 (1994), pp. 4436–4440.

[84] C. A. Rohl, C. E. M. Strauss, K. M. S. Misura, and D. Baker. “Protein Structure Prediction Using Rosetta”. Ed. by B. .-. M. i. Enzymology. Vol. 383. Numerical Computer Methods, Part D. Academic Press, 2004, pp. 66–93.

[85] T. M. Jacobs, B. Williams, T. Williams, X. Xu, A. Eletsky, J. F. Federizon, T. Szyperski, and B. Kuhlman. “Design of structurally distinct proteins using strategies inspired by evolution”. *Science* 352.6286 (2016), pp. 687–690.

[86] E. Marcos, B. Basanta, T. M. Chidyausiku, Y. Tang, G. Oberdorfer, G. Liu, G. V. T. Swapna, R. Guan, D.-A. Silva, J. Dou, J. H. Pereira, R. Xiao, B. Sankaran, P. H. Zwart, G. T. Montelione, and D. Baker. “Principles for designing proteins with cavities formed by curved beta-sheets”. *Science* 355.6321 (2017), pp. 201–206.

[87] E. Marcos, T. M. Chidyausiku, A. C. McShan, T. Evangelidis, S. Nerli, L. Carter, L. G. Nivón, A. Davis, G. Oberdorfer, K. Tripsianes, N. G. Sgourakis, and D. Baker. “De novo design of a non-local beta-sheet protein with high stability and accuracy”. *Nat Struct Mol Biol* 25.11 (2018), pp. 1028–1034.

[88] T. M. Chidyausiku, S. R. Mendes, J. C. Klima, U. Eckhard, S. Houliston, M. Nadal, J. Roel-Touris, T. Guevara, H. K. Haddox, A. Moyer, C. H. Arrowsmith, F. X. Gomis-Rüth, D. Baker, and E. Marcos. “De Novo Design of Immunoglobulin-like Domains”. *bioRxiv* (2021), p. 2021.12.20.472081.

[89] S. Govindarajan and R. A. Goldstein. “Why are some proteins structures so common?” *Proc Natl Acad Sci U S A* 93.8 (1996), pp. 3341–3345.

[90] R. Helling, H. Li, R. Mélin, J. Miller, N. Wingreen, C. Zeng, and C. Tang. “The designability of protein structures”. *J Mol Graph Model* 19.1 (2001), pp. 157–167.

[91] P. Koehl and M. Levitt. “De novo protein design. I. in search of stability and specificity”. *J Mol Biol* 293.5 (1999), pp. 1161–1181.

[92] H. Li, R. Helling, C. Tang, and N. Wingreen. “Emergence of Preferred Structures in a Simple Model of Protein Folding”. *Science* 273.5275 (1996), pp. 666–669.

[93] J. Zhang, F. Zheng, and G. Grigoryan. “Design and designability of protein-based assemblies”. *Current Opinion in Structural Biology*. Membranes / Engineering and design 27 (2014), pp. 79–86.

[94] J. L. England and E. I. Shakhnovich. “Structural determinant of protein designability”. *Phys Rev Lett* 90.21 (2003), p. 218101.

[95] G. Grigoryan and W. F. DeGrado. “Probing Designability via a Generalized Model of Helical Bundle Geometry”. *J Mol Biol* 405.4 (2011), pp. 1079–1100.

[96] N. S. Wingreen, H. Li, and C. Tang. “Designability and thermal stability of protein structures”. *Polymer*. Conformational Protein Conformations 45.2 (2004), pp. 699–705.

[97] K. T. Simons, R. Bonneau, I. Ruczinski, and D. Baker. “Ab initio protein structure prediction of CASP III targets using ROSETTA”. *Proteins: Struct Funct Genet* 37.S3 (1999), pp. 171–176.

[98] J. S. Richardson. “The anatomy and taxonomy of protein structure”. *Adv Protein Chem* 34 (1981), pp. 167–339.

[99] N. Koga, R. Tatsumi-Koga, G. Liu, R. Xiao, T. B. Acton, G. T. Montelione, and D. Baker. “Principles for designing ideal protein structures”. *Nature* 491.7423 (2012), pp. 222–227.

[100] Y.-R. Lin, N. Koga, R. Tatsumi-Koga, G. Liu, A. F. Clouser, G. T. Montelione, and D. Baker. "Control over overall shape and size in de novo designed proteins". *Proc Natl Acad Sci U S A* 112.40 (2015), E5478–E5485.

[101] N. Koga, R. Koga, G. Liu, J. Castellanos, G. T. Montelione, and D. Baker. "Role of backbone strain in de novo design of complex α/β protein structures". *Nat Commun* 12.1 (2021), p. 3921.

[102] Y. Du, J. Meier, J. Ma, R. Fergus, and A. Rives. "Energy-based models for atomic-resolution protein conformations". *arXiv* (2020).

[103] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. Fergus. "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences". *Proc Natl Acad Sci U S A* 118.15 (2021).

[104] J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, and A. Rives. "Language models enable zero-shot prediction of the effects of mutations on protein function". *bioRxiv* (2021), p. 2021.07.09.450648.

[105] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu, and D. Hassabis. "Improved protein structure prediction using potentials from deep learning". *Nature* 577.7792 (2020), pp. 706–710.

[106] J. Yang, I. Anishchenko, H. Park, Z. Peng, S. Ovchinnikov, and D. Baker. "Improved protein structure prediction using predicted interresidue orientations". *Proc Natl Acad Sci U S A* 117.3 (2020), pp. 1496–1503.

[107] M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. v. Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read, and D. Baker. "Accurate prediction of protein structures and interactions using a three-track neural network". *Science* (2021).

[108] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, 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, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. "Highly accurate protein structure prediction with AlphaFold". *Nature* 596.7873 (2021), pp. 583–589.

[109] D. Jin, Z. Jin, Z. Hu, O. Vechtomova, and R. Mihalcea. "Deep Learning for Text Style Transfer: A Survey". *arXiv* (2021).

[110] E. G. Tabak and E. Vanden-Eijnden. "Density estimation by dual ascent of the log-likelihood". *Commun Math Sci* 8.1 (2010), pp. 217–233.

[111] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. “Generative Adversarial Networks”. *arXiv* (2014).

[112] D. P. Kingma and M. Welling. “Auto-Encoding Variational Bayes”. *arXiv* (2014).

[113] D. T. Jones and S. M. Kandathil. “High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features”. *Bioinformatics* 34.19 (2018), pp. 3308–3315.

[114] S. Wang, S. Sun, Z. Li, R. Zhang, and J. Xu. “Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model”. *PLoS Comput Biol* 13.1 (2017), e1005324.

[115] S. Ovchinnikov, H. Kamisetty, and D. Baker. “Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information”. *eLife* 3 (2014). Ed. by B. Roux, e02030.

[116] M. Fantini, S. Lisi, P. De Los Rios, A. Cattaneo, and A. Pastore. “Protein Structural Information and Evolutionary Landscape by In Vitro Evolution”. *Mol Biol Evol* 37.4 (2020), pp. 1179–1192.

[117] J. G. Kirkwood. “Statistical Mechanics of Fluid Mixtures”. *J Chem Phys* 3.5 (1935), pp. 300–313.

[118] J. Ko and J. Lee. “Can AlphaFold2 predict protein-peptide complex structures accurately?” *bioRxiv* (2021), p. 2021.07.27.453972.

[119] A. J. McCoy, M. D. Sammito, and R. J. Read. “Possible Implications of AlphaFold2 for Crystallographic Phasing by Molecular Replacement”. *bioRxiv* (2021), p. 2021.05.18.444614.

[120] T. C. Terwilliger, B. K. Poon, P. V. Afonine, C. J. Schlicksup, T. I. Croll, C. Millán, J. S. Richardson, R. J. Read, and P. D. Adams. “Improved AlphaFold modeling with implicit experimental information”. *bioRxiv* (2022), p. 2022.01.07.475350.

[121] K. Tunyasuvunakool, J. Adler, Z. Wu, T. Green, M. Zielinski, A. Žídek, A. Bridgland, A. Cowie, C. Meyer, A. Laydon, S. Velankar, G. J. Kleywegt, A. Bateman, R. Evans, A. Pritzel, M. Figurnov, O. Ronneberger, R. Bates, S. A. A. Kohl, A. Potapenko, A. J. Ballard, B. Romera-Paredes, S. Nikolov, R. Jain, E. Clancy, D. Reiman, S. Petersen, A. W. Senior, K. Kavukcuoglu, E. Birney, P. Kohli, J. Jumper, and D. Hassabis. “Highly accurate protein structure prediction for the human proteome”. *Nature* 596.7873 (2021), pp. 590–596.

[122] K. K. Yang, Z. Wu, and F. H. Arnold. “Machine-learning-guided directed evolution for protein engineering”. *Nat Methods* 16.8 (2019), pp. 687–694.

[123] B. L. Hie and K. K. Yang. “Adaptive machine learning for protein engineering”. *Curr Opin Struct Biol* 72 (2022), pp. 145–152.

[124] Z. Wu, K. E. Johnston, F. H. Arnold, and K. K. Yang. “Protein sequence design with deep generative models”. *Curr Opin Chem Biol.* Mechanistic Biology * Machine Learning in Chemical Biology 65 (2021), pp. 18–27.

[125] T. Bepler and B. Berger. “Learning the protein language: Evolution, structure, and function”. *Cell Syst* 12.6 (2021), 654–669.e3.

[126] M. AlQuraishi. “End-to-End Differentiable Learning of Protein Structure”. *Cell Syst* 8.4 (2019), 292–301.e3.

[127] J. Ingraham, A. Riesselman, C. Sander, and D. Marks. “Learning Protein Structure with a Differentiable Simulator”. 2018.

[128] Z. Li, S. P. Nguyen, D. Xu, and Y. Shang. “Protein Loop Modeling Using Deep Generative Adversarial Network”. *IEEE 29th International Conference on Tools with Artificial Intelligence.* 2017, pp. 1085–1091.

[129] N. Anand and P. Huang. “Generative Modeling for Protein Structures”. *International Conference on Learning Representations (workshop)* (2018).

[130] X. Guo, Y. Du, S. Tadepalli, L. Zhao, and A. Shehu. “Generating Tertiary Protein Structures via an Interpretative Variational Autoencoder”. *arXiv* (2021).

[131] I. Dokmanic, R. Parhizkar, J. Ranieri, and M. Vetterli. “Euclidean Distance Matrices: Essential theory, algorithms, and applications”. *IEEE Signal Process Mag* 32.6 (2015), pp. 12–30.

[132] N. Anand, R. Eguchi, and P.-S. Huang. “Fully differentiable full-atom protein backbone generation”. *International Conference on Learning Representations (ICLR)* (2019).

[133] R. R. Eguchi, N. Anand, C. A. Choe, and P.-S. Huang. “IG-VAE: Generative Modeling of Immunoglobulin Proteins by Direct 3D Coordinate Generation”. *bioRxiv* (2020), p. 2020.08.07.242347.

[134] M. Karimi, S. Zhu, Y. Cao, and Y. Shen. “De Novo Protein Design for Novel Folds Using Guided Conditional Wasserstein Generative Adversarial Networks”. *J Chem Inf Model* 60.12 (2020), pp. 5667–5681.

[135] A. Strokach, D. Becerra, C. Corbi-Verge, A. Perez-Riba, and P. M. Kim. “Fast and Flexible Protein Design Using Deep Graph Neural Networks”. *Cell Syst* 11.4 (2020), 402–411.e4.

[136] J. Ingraham, V. K. Garg, R. Barzilay, and T. Jaakkola. “Generative Models for Graph-Based Protein Design”. *Conference and Workshop on Neural Information Processing Systems (NeurIPS)* (2019).

[137] J. G. Greener, L. Moffat, and D. T. Jones. “Design of metalloproteins and novel protein folds using variational autoencoders”. *Sci Rep* 8.1 (2018), p. 16189.

[138] I. Anishchenko, S. J. Pellock, T. M. Chidyausiku, T. A. Ramelot, S. Ovchinnikov, J. Hao, K. Bafna, C. Norn, A. Kang, A. K. Bera, F. DiMaio, L. Carter, C. M. Chow, G. T. Montelione, and D. Baker. “De novo protein design by deep network hallucination”. *Nature* 600.7889 (2021), pp. 547–552.

[139] C. Norn, B. I. M. Wicky, D. Juergens, S. Liu, D. Kim, D. Tischer, B. Koepnick, I. Anishchenko, F. Players, D. Baker, and S. Ovchinnikov. “Protein sequence design by conformational landscape optimization”. *Proc Natl Acad Sci U S A* 118.11 (2021).

[140] D. Tischer, S. Lisanza, J. Wang, R. Dong, I. Anishchenko, L. F. Milles, S. Ovchinnikov, and D. Baker. “Design of proteins presenting discontinuous functional sites using deep learning”. *bioRxiv* (2020), p. 2020.11.29.402743.

[141] J. Wang, S. Lisanza, D. Juergens, D. Tischer, I. Anishchenko, M. Baek, J. L. Watson, J. H. Chun, L. F. Milles, J. Dauparas, M. Expòsit, W. Yang, A. Saragovi, S. Ovchinnikov, and D. Baker. “Deep learning methods for designing proteins scaffolding functional sites”. *bioRxiv* (2021), p. 2021.11.10.468128.

[142] L. Moffat, J. G. Greener, and D. T. Jones. “Using AlphaFold for Rapid and Accurate Fixed Backbone Protein Design”. *bioRxiv* (2021), p. 2021.08.24.457549.

[143] M. Jendrusch, J. O. Korbel, and S. K. Sadiq. “AlphaDesign: A de novo protein design framework based on AlphaFold”. *bioRxiv* (2021), p. 2021.10.11.463937.

[144] L. Lo Conte, C. Chothia, and J. Janin. “The atomic structure of protein-protein recognition sites”. *J Mol Biol* 285.5 (1999), pp. 2177–2198.

[145] M. C. Lawrence and P. M. Colman. “Shape complementarity at protein/protein interfaces”. *J Mol Biol* 234.4 (1993), pp. 946–950.

[146] G. Schreiber and A. E. Keating. “Protein binding specificity versus promiscuity”. *Curr Opin Struct Biol* 21.1 (2011), pp. 50–61.

[147] Y. Ofran and B. Rost. “Protein–Protein Interaction Hotspots Carved into Sequences”. *PLoS Comput Biol* 3.7 (2007), e119.

[148] Z. Hu, B. Ma, H. Wolfson, and R. Nussinov. “Conservation of polar residues as hot spots at protein interfaces”. *Proteins* 39.4 (2000), pp. 331–342.

[149] G. Schreiber and S. J. Fleishman. “Computational design of protein-protein interactions”. *Curr Opin Struct Biol* 23.6 (2013), pp. 903–910.

[150] T. Selzer, S. Albeck, and G. Schreiber. “Rational design of faster associating and tighter binding protein complexes”. *Nat Struct Biol* 7.7 (2000), pp. 537–541.

[151] T. Clark. “Directional Electrostatic Bonding”. *The Chemical Bond.* John Wiley & Sons, Ltd, 2014, pp. 523–536.

[152] P. Schmidtke, F. J. Luque, J. B. Murray, and X. Barril. “Shielded hydrogen bonds as structural determinants of binding kinetics: application in drug design”. *J Am Chem Soc* 133.46 (2011), pp. 18903–18910.

[153] A. Chevalier, D.-A. Silva, G. J. Rocklin, D. R. Hicks, R. Vergara, P. Murapa, S. M. Bernard, L. Zhang, K.-H. Lam, G. Yao, C. D. Bahl, S.-I. Miyashita, I. Goreshnik, J. T. Fuller, M. T. Koday, C. M. Jenkins, T. Colvin, L. Carter, A. Bohn, C. M. Bryan, D. A. Fernández-Velasco, L. Stewart, M. Dong, X. Huang, R. Jin, I. A. Wilson, D. H. Fuller, and D. Baker. “Massively parallel de novo protein design for targeted therapeutics”. *Nature* 550.7674 (2017), pp. 74–79.

[154] L. Cao, B. Coventry, I. Goreshnik, B. Huang, J. S. Park, K. M. Jude, I. Marković, R. U. Kadam, K. H. G. Verschueren, K. Verstraete, S. T. R. Walsh, N. Bennett, A. Phal, A. Yang, L. Kozodoy, M. DeWitt, L. Picton, L. Miller, E.-M. Strauch, S. Halabiya, B. Hammerson, W. Yang, S. Benard, L. Stewart, I. A. Wilson, H. Ruohola-Baker, J. Schlessinger, S. Lee, S. N. Savvides, K. C. Garcia, and D. Baker. “Robust de novo design of protein binding proteins from target structural information alone”. *bioRxiv* (2021), p. 2021.09.04.459002.

[155] J. S. Richardson, D. C. Richardson, N. B. Tweedy, K. M. Gernert, T. P. Quinn, M. H. Hecht, B. W. Erickson, Y. Yan, R. D. McClain, and M. E. Donlan. "Looking at proteins: representations, folding, packing, and design". *Biophys J* 63.5 (1992), pp. 1185–1209.

[156] J. Karanicolas, J. E. Corn, I. Chen, L. A. Joachimiak, O. Dym, S. H. Peck, S. Albeck, T. Unger, W. Hu, G. Liu, S. Delbecq, G. T. Montelione, C. P. Spiegel, D. R. Liu, and D. Baker. "A De Novo Protein Binding Pair By Computational Design and Directed Evolution". *Molecular Cell* 42.2 (2011), pp. 250–260.

[157] J. J. Havranek and P. B. Harbury. "Automated design of specificity in molecular recognition". *Nat Struct Mol Biol* 10.1 (2003), pp. 45–52.

[158] A. Koide and S. Koide. "Monobodies: antibody mimics based on the scaffold of the fibronectin type III domain". *Methods Mol Biol* 352 (2007), pp. 95–109.

[159] H. K. Binz, P. Amstutz, A. Kohl, M. T. Stumpp, C. Briand, P. Forrer, M. G. Grütter, and A. Plückthun. "High-affinity binders selected from designed ankyrin repeat protein libraries". *Nat Biotechnol* 22.5 (2004), pp. 575–582.

[160] H. K. Binz, M. T. Stumpp, P. Forrer, P. Amstutz, and A. Plückthun. "Designing repeat proteins: well-expressed, soluble and stable proteins from combinatorial libraries of consensus ankyrin repeat proteins". *J Mol Biol* 332.2 (2003), pp. 489–503.

[161] T. Kortemme and D. Baker. "Computational design of protein–protein interactions". *Curr Opin Chem Biol* 8.1 (2004), pp. 91–97.

[162] F. Bubeck, M. D. Hoffmann, Z. Harteveld, S. Aschenbrenner, A. Bietz, M. C. Waldhauer, K. Börner, J. Fakhiri, C. Schmelas, L. Dietz, D. Grimm, B. E. Correia, R. Eils, and D. Niopek. "Engineered anti-CRISPR proteins for optogenetic control of CRISPR–Cas9". *Nat Methods* 15.11 (2018), pp. 924–927.

[163] D.-A. Silva, B. E. Correia, and E. Procko. "Motif-Driven Design of Protein-Protein Interfaces". *Methods Mol. Biol.* 1414 (2016), pp. 285–304.

[164] J. J. Havranek and D. Baker. "Motif-directed flexible backbone design of functional interactions". *Protein Sci* 18.6 (2009), pp. 1293–1305.

[165] S. J. Fleishman, T. A. Whitehead, D. C. Ekiert, C. Dreyfus, J. E. Corn, E.-M. Strauch, I. A. Wilson, and D. Baker. "Computational Design of Proteins Targeting the Conserved Stem Region of Influenza Hemagglutinin". *Science* 332.6031 (2011), pp. 816–821.

[166] T. Clackson and J. A. Wells. "A hot spot of binding energy in a hormone-receptor interface". *Science* 267.5196 (1995), pp. 383–386.

[167] T. Clackson, M. H. Ultsch, J. A. Wells, and A. M. de Vos. "Structural and functional analysis of the 1:1 growth hormone:receptor complex reveals the molecular basis for receptor affinity". *J Mol Bio* 277.5 (1998), pp. 1111–1128.

[168] A. A. Bogan and K. S. Thorn. "Anatomy of hot spots in protein interfaces". *J Mol Bio* 280.1 (1998), pp. 1–9.

[169] B. Ma, T. Elkayam, H. Wolfson, and R. Nussinov. "Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces". *Proc Natl Acad Sci U S A* 100.10 (2003), pp. 5772–5777.

[170] S. J. Fleishman, S. D. Khare, N. Koga, and D. Baker. "Restricted sidechain plasticity in the structures of native proteins and complexes". *Protein Sci* 20.4 (2011), pp. 753–757.

[171] P. Gainza, F. Sverrisson, F. Monti, E. Rodolà, D. Boscaini, M. M. Bronstein, and B. E. Correia. "Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning". *Nat Methods* 17.2 (2020), pp. 184–192.

[172] B. E. Correia, J. T. Bates, R. J. Loomis, G. Baneyx, C. Carrico, J. G. Jardine, P. Rupert, C. Correnti, O. Kalyuzhniy, V. Vittal, M. J. Connell, E. Stevens, A. Schroeter, M. Chen, S. Macpherson, A. M. Serra, Y. Adachi, M. A. Holmes, Y. Li, R. E. Klevit, B. S. Graham, R. T. Wyatt, D. Baker, R. K. Strong, J. E. Crowe, P. R. Johnson, and W. R. Schief. "Proof of principle for epitope-focused vaccine design". *Nature* 507.7491 (2014), pp. 201–206.

[173] G. Giordano-Attianese, P. Gainza, E. Gray-Gaillard, E. Cribioli, S. Shui, S. Kim, M.-J. Kwak, S. Vollers, A. D. J. Corria Osorio, P. Reichenbach, J. Bonet, B.-H. Oh, M. Irving, G. Coukos, and B. E. Correia. "A computationally designed chimeric antigen receptor provides a small-molecule safety switch for T-cell therapy". *Nat Biotechnol* 38.4 (2020), pp. 426–432.

[174] F. Sesterhenn, C. Yang, J. Bonet, J. T. Cramer, X. Wen, Y. Wang, C.-I. Chiang, L. A. Abriata, I. Kucharska, G. Castoro, S. S. Vollers, M. Galloux, E. Dheilly, S. Rosset, P. Corthésy, S. Georgeon, M. Villard, C.-A. Richard, D. Descamps, T. Delgado, E. Oricchio, M.-A. Rameix-Welti, V. Más, S. Ervin, J.-F. Eléouët, S. Riffault, J. T. Bates, J.-P. Julien, Y. Li, T. Jardetzky, T. Krey, and B. E. Correia. "De novo protein design enables the precise induction of RSV-neutralizing antibodies". *Science* 368.6492 (2020), eaay5051.

[175] C. Yang, F. Sesterhenn, J. Bonet, E. A. van Aalen, L. Scheller, L. A. Abriata, J. T. Cramer, X. Wen, S. Rosset, S. Georgeon, T. Jardetzky, T. Krey, M. Fussenegger, M. Merkx, and B. E. Correia. "Bottom-up de novo design of functional proteins with complex structural features". *Nat Chem Biol* (2021), pp. 1–9.

[176] F. Sverrisson, J. Feydy, B. E. Correia, and M. M. Bronstein. "Fast End-to-End Learning on Protein Surfaces". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 15272–15281.

[177] M. L. Connolly. "Analytical molecular surface calculation". *J Appl Crystallogr* 16.5 (1983), pp. 548–558.

[178] D. Kihara, L. Sael, R. Chikhi, and J. Esquivel-Rodriguez. "Molecular surface representation using 3D Zernike descriptors for protein shape comparison and docking". *Curr Protein Pept Sci* 12.6 (2011), pp. 520–530.

[179] L. Sael, D. La, B. Li, R. Rustamov, and D. Kihara. "Rapid comparison of properties on protein surface". *Proteins* 73.1 (2008), pp. 1–10.

[180] J. Ryu, R. Park, and D.-S. Kim. “Molecular surfaces on proteins via beta shapes”. *Computer-Aided Design* 39.12 (2007), pp. 1042–1057.

[181] S. Yin, E. A. Proctor, A. A. Lugovskoy, and N. V. Dokholyan. “Fast screening of protein surfaces using geometric invariant fingerprints”. *Proc Natl Acad Sci U S A* 106.39 (2009), pp. 16622–16626.

[182] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. “Geometric deep learning: going beyond Euclidean data”. *IEEE Signal Process Mag* 34.4 (2017), pp. 18–42.

[183] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković. “Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges”. *arXiv:2104.13478* (2021).

[184] Yoshitaka Moriwaki. *AlphaFold2 can also predict heterocomplexes. All you have to do is input the two sequences you want to predict and connect them with a long linker (https://twitter.com/Ag_smith/status/1417063635000598528)*. Tweet. 2021.

[185] M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, and M. Steinegger. “ColabFold - Making protein folding accessible to all”. *bioRxiv* (2021), p. 2021.08.15.456425.

[186] Minkyung Baek. *Adding a big enough number for "residue_index" feature is enough to model hetero-complex using AlphaFold (https://twitter.com/minkbaek/status/1417538291709071362)*. Tweet. 2021.

[187] T. Tsaban, J. K. Varga, O. Avraham, Z. Ben-Aharon, A. Khramushin, and O. Schueler-Furman. “Harnessing protein folding neural networks for peptide–protein docking”. *Nat Commun* 13.1 (2022), p. 176.

[188] P. Bryant, G. Pozzati, and A. Elofsson. “Improved prediction of protein-protein interactions using AlphaFold2”. *bioRxiv* (2021), p. 2021.09.15.460468.

[189] R. Evans, M. O’Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper, and D. Hassabis. “Protein complex prediction with AlphaFold-Multimer”. *bioRxiv* (2021), p. 2021.10.04.463034.

[190] R. Barrangou and J. A. Doudna. “Applications of CRISPR technologies in research and beyond”. *Nat Biotechnol* 34.9 (2016), pp. 933–941.

[191] F. Richter, I. Fonfara, R. Gelfert, J. Nack, E. Charpentier, and A. Möglich. “Switchable Cas9”. *Curr Opin Biotechnol.* Chemical biotechnology • Pharmaceutical biotechnology 48 (2017), pp. 119–126.

[192] B. J. Rauch, M. R. Silvis, J. F. Hultquist, C. S. Waters, M. J. McGregor, N. J. Krogan, and J. Bondy-Denomy. “Inhibition of CRISPR-Cas9 with Bacteriophage Proteins”. *Cell* 168.1 (2017), 150–158.e10.

[193] A. P. Hynes, G. M. Rousseau, M.-L. Lemay, P. Horvath, D. A. Romero, C. Fremaux, and S. Moineau. “An anti-CRISPR from a virulent streptococcal phage inhibits Streptococcus pyogenes Cas9”. *Nat Microbiol* 2.10 (2017), pp. 1374–1380.

[194] E. M. Basgall, S. C. Goetting, M. E. Goeckel, R. M. Giersch, E. Roggenkamp, M. N. Schrock, M. Halloran, and G. C. Finnigan. “Gene drive inhibition by the anti-CRISPR proteins AcrIIA2 and AcrIIA4 in Saccharomyces cerevisiae”. *Microbiology* 164.4 (2018), pp. 464–474.

[195] J. Shin, F. Jiang, J.-J. Liu, N. L. Bray, B. J. Rauch, S. H. Baik, E. Nogales, J. Bondy-Denomy, J. E. Corn, and J. A. Doudna. “Disabling Cas9 by an anti-CRISPR DNA mimic”. *Sci Adv* 3.7 (2018), e1701620.

[196] S. M. Harper, L. C. Neil, and K. H. Gardner. “Structural Basis of a Phototropin Light Switch”. *Science* 301.5639 (2003), pp. 1541–1544.

[197] M. D. Hoffmann, F. Bubeck, R. Eils, and D. Niopek. “Controlling Cells with Light and LOV”. *Adv Biosyst* 2.9 (2018), p. 1800098.

[198] O. Dagliyan, M. Tarnawski, P.-H. Chu, D. Shirvanyants, I. Schlichting, N. V. Dokholyan, and K. M. Hahn. “Engineering extrinsic disorder to control protein activity in living cells”. *Science* 354.6318 (2016), pp. 1441–1444.

[199] I. Kim, M. Jeong, D. Ka, M. Han, N.-K. Kim, E. Bae, and J.-Y. Suh. “Solution structure and dynamics of anti-CRISPR AcrIIA4, the Cas9 inhibitor”. *Sci Rep* 8.1 (2018), p. 3883.

[200] D. Dong, M. Guo, S. Wang, Y. Zhu, S. Wang, Z. Xiong, J. Yang, Z. Xu, and Z. Huang. “Structural basis of CRISPR–SpyCas9 inhibition by an anti-CRISPR protein”. *Nature* 546.7658 (2017), pp. 436–439.

[201] H. Yang and D. J. Patel. “Inhibition Mechanism of an Anti-CRISPR Suppressor AcrIIA4 Targeting SpyCas9”. *Mol Cell* 67.1 (2017), 117–127.e5.

[202] E. K. Brinkman, T. Chen, M. Amendola, and B. van Steensel. “Easy quantitative assessment of genome editing by sequence trace decomposition”. *Nucleic Acids Res* 42.22 (2014), e168.

[203] D. Strickland, Y. Lin, E. Wagner, C. M. Hope, J. Zayner, C. Antoniou, T. R. Sosnick, E. L. Weiss, and M. Glotzer. “TULIPs: tunable, light-controlled interacting protein tags for cell biology”. *Nat Methods* 9.4 (2012), pp. 379–384.

[204] D. Strickland, X. Yao, G. Gawlak, M. K. Rosen, K. H. Gardner, and T. R. Sosnick. “Rationally improving LOV domain–based photoswitches”. *Nat Methods* 7.8 (2010), pp. 623–626.

[205] J. H. Hu, S. M. Miller, M. H. Geurts, W. Tang, L. Chen, N. Sun, C. M. Zeina, X. Gao, H. A. Rees, Z. Lin, and D. R. Liu. “Evolved Cas9 variants with broad PAM compatibility and high DNA specificity”. *Nature* 556.7699 (2018), pp. 57–63.

[206] I. B. Hilton, A. M. D’Ippolito, C. M. Vockley, P. I. Thakore, G. E. Crawford, T. E. Reddy, and C. A. Gersbach. “Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers”. *Nat Biotechnol* 33.5 (2015), pp. 510–517.

[207] B. Chen, L. A. Gilbert, B. A. Cimini, J. Schnitzbauer, W. Zhang, G.-W. Li, J. Park, E. H. Blackburn, J. S. Weissman, L. S. Qi, and B. Huang. "Dynamic Imaging of Genomic Loci in Living Human Cells by an Optimized CRISPR/Cas System". *Cell* 155.7 (2013), pp. 1479–1491.

[208] A. Pawluk, N. Amrani, Y. Zhang, B. Garcia, Y. Hidalgo-Reyes, J. Lee, A. Edraki, M. Shah, E. J. Sontheimer, K. L. Maxwell, and A. R. Davidson. "Naturally Occurring Off-Switches for CRISPR-Cas9". *Cell* 167.7 (2016), 1829–1838.e9.

[209] C. Vehlow, H. Stehr, M. Winkelmann, J. M. Duarte, L. Petzold, J. Dinse, and M. Lappe. "CMView: Interactive contact map visualization and analysis". *Bioinformatics* 27.11 (2011), pp. 1573–1574.

[210] W. Kabsch and C. Sander. "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features". *Biopolymers* 22.12 (1983), pp. 2577–2637.

[211] P.-S. Huang, Y.-E. A. Ban, F. Richter, I. Andre, R. Vernon, W. R. Schief, and D. Baker. "RosettaRemodel: A Generalized Framework for Flexible Backbone Protein Design". *PLoS One* 6.8 (2011), e24109.

[212] A. A. Canutescu and R. L. Dunbrack. "Cyclic coordinate descent: A robotics algorithm for protein loop closure". *Protein Sci* 12.5 (2003), pp. 963–972.

[213] E. A. Coutsias, C. Seok, M. P. Jacobson, and K. A. Dill. "A kinematic view of loop closure". *J Comput Chem* 25.4 (2004), pp. 510–528.

[214] D. J. Mandell, E. A. Coutsias, and T. Kortemme. "Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling". *Nat Methods* 6.8 (2009), pp. 551–552.

[215] C. Engler, R. Kandzia, and S. Marillonnet. "A One Pot, One Step, Precision Cloning Method with High Throughput Capability". *PLoS One* 3.11 (2008), e3647.

[216] E. Senís, C. Fatouros, S. Große, E. Wiedtke, D. Niopek, A.-K. Mueller, K. Börner, and D. Grimm. "CRISPR/Cas9-mediated genome engineering: An adeno-associated viral (AAV) vector toolbox". *Biotechnol J* 9.11 (2014), pp. 1402–1412.

[217] M. Salomon, W. Eisenreich, H. Dürr, E. Schleicher, E. Knieb, V. Massey, W. Rüdiger, F. Müller, A. Bacher, and G. Richter. "An optomechanical transducer in the blue light receptor phototropin from Avena sativa". *Proc Natl Acad Sci U S A* 98.22 (2001), pp. 12357–12361.

[218] K. J. Livak and T. D. Schmittgen. "Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2$\Delta\Delta$CT Method". *Methods* 25.4 (2001), pp. 402–408.

[219] G. Guntas, R. A. Hallett, S. P. Zimmerman, T. Williams, H. Yumerefendi, J. E. Bear, and B. Kuhlman. "Engineering an improved light-induced dimer (iLID) for controlling the localization and activity of signaling proteins". *Proc Natl Acad Sci U S A* 112.1 (2015), pp. 112–117.

[220] D. Niopek, P. Wehler, J. Roensch, R. Eils, and B. Di Ventura. “Optogenetic control of nuclear protein export”. *Nat Commun* 7.1 (2016), p. 10624.

[221] R. M. Melero-Fernandez de Mera, L.-L. Li, A. Popinigis, K. Cisek, M. Tuittila, L. Yadav, A. Serva, and M. J. Courtney. “A simple optogenetic MAPK inhibitor design reveals resonance between transcription-regulating circuitry and temporally-encoded inputs”. *Nat Commun* 8.1 (2017), p. 15017.

[222] D. Niopek, D. Benzinger, J. Roensch, T. Draebing, P. Wehler, R. Eils, and B. Di Ventura. “Engineering light-inducible nuclear localization signals for precise spatiotemporal control of protein dynamics in living cells”. *Nat Commun* 5.1 (2014), p. 4404.

[223] X. Yao, M. K. Rosen, and K. H. Gardner. “Estimation of the available free energy in a LOV2-J alpha photoswitch”. *Nat Chem Biol* 4.8 (2008), pp. 491–497.

[224] A. S. Halavaty and K. Moffat. “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* 46.49 (2007), pp. 14001–14009.

[225] J. P. Zayner, C. Antoniou, and T. R. Sosnick. “The amino-terminal helix modulates light-activated conformational changes in AsLOV2”. *J Mol Biol* 419.1-2 (2012), pp. 61–74.

[226] G. Kothe, M. Lukaschek, G. Link, S. Kacprzak, B. Illarionov, M. Fischer, W. Eisenreich, A. Bacher, and S. Weber. “Detecting a new source for photochemically induced dynamic nuclear polarization in the LOV2 domain of phototropin by magnetic-field dependent (13)C NMR spectroscopy”. *J Phys Chem B* 118.40 (2014), pp. 11622–11632.

[227] G. Richter, S. Weber, W. Römisch, A. Bacher, M. Fischer, and W. Eisenreich. “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* 127.49 (2005), pp. 17245–17252.

[228] Y. I. Wu, D. Frey, O. I. Lungu, A. Jaehrig, I. Schlichting, B. Kuhlman, and K. M. Hahn. “A genetically encoded photoactivatable Rac controls the motility of living cells”. *Nature* 461.7260 (2009), pp. 104–108.

[229] H. Wang, M. Vilela, A. Winkler, M. Tarnawski, I. Schlichting, H. Yumerefendi, B. Kuhlman, R. Liu, G. Danuser, and K. M. Hahn. “LOVTRAP: an optogenetic system for photoinduced protein dissociation”. *Nat Methods* 13.9 (2016), pp. 755–758.

[230] S. Wong, A. A. Mosabbir, and K. Truong. “An Engineered Split Intein for Photoactivated Protein Trans-Splicing”. *PLoS One* 10.8 (2015), e0135965.

[231] L. Cong, F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, X. Wu, W. Jiang, L. A. Marraffini, and F. Zhang. “Multiplex genome engineering using CRISPR/Cas systems”. *Science* 339.6121 (2013), pp. 819–823.

[232] S. Konermann, M. D. Brigham, A. E. Trevino, J. Joung, O. O. Abudayyeh, C. Barcena, P. D. Hsu, N. Habib, J. S. Gootenberg, H. Nishimasu, O. Nureki, and F. Zhang. “Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex”. *Nature* 517.7536 (2015), pp. 583–588.

[233] P. Mali, J. Aach, P. B. Stranges, K. M. Esvelt, M. Moosburner, S. Kosuri, L. Yang, and G. M. Church. “CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering”. *Nat Biotechnol* 31.9 (2013), pp. 833–838.

[234] M. L. Maeder, S. J. Linder, V. M. Cascio, Y. Fu, Q. H. Ho, and J. K. Joung. “CRISPR RNA-guided activation of endogenous human genes”. *Nat Methods* 10.10 (2013), pp. 977–979.

[235] T. Chen, D. Gao, R. Zhang, G. Zeng, H. Yan, E. Lim, and F.-S. Liang. “Chemically Controlled Epigenome Editing through an Inducible dCas9 System”. *J Am Chem Soc* 139.33 (2017), pp. 11337–11340.

[236] H. Ma, L.-C. Tu, A. Naseri, M. Huisman, S. Zhang, D. Grunwald, and T. Pederson. “CRISPR-Cas9 nuclear dynamics and target recognition in living cells”. *J Cell Biol* 214.5 (2016), pp. 529–537.

[237] S. C. Knight, L. Xie, W. Deng, B. Guglielmi, L. B. Witkowsky, L. Bosanac, E. T. Zhang, M. El Beheiry, J.-B. Masson, M. Dahan, Z. Liu, J. A. Doudna, and R. Tjian. “Dynamics of CRISPR-Cas9 genome interrogation in living cells”. *Science* 350.6262 (2015), pp. 823–826.

[238] P. Qin, M. Parlak, C. Kuscu, J. Bandaria, M. Mir, K. Szlachta, R. Singh, X. Darzacq, A. Yildiz, and M. Adli. “Live cell imaging of low- and non-repetitive chromosome loci using CRISPR-Cas9”. *Nat Commun* 8.1 (2017), p. 14725.

[239] J. Shin, F. Jiang, J.-J. Liu, N. L. Bray, B. J. Rauch, S. H. Baik, E. Nogales, J. Bondy-Denomy, J. E. Corn, and J. A. Doudna. “Disabling Cas9 by an anti-CRISPR DNA mimic”. *Sci Adv* 3.7 (2017), e1701620.

[240] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier. “A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity”. *Science* 337.6096 (2012), pp. 816–821.

[241] J. Bondy-Denomy, A. Pawluk, K. L. Maxwell, and A. R. Davidson. “Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system”. *Nature* 493.7432 (2013), pp. 429–432.

[242] J. Bondy-Denomy, B. Garcia, S. Strum, M. Du, M. F. Rollins, Y. Hidalgo-Reyes, B. Wiedenheft, K. L. Maxwell, and A. R. Davidson. “Multiple mechanisms for CRISPR-Cas inhibition by anti-CRISPR proteins”. *Nature* 526.7571 (2015), pp. 136–139.

[243] A. Pawluk, J. Bondy-Denomy, V. H. W. Cheung, K. L. Maxwell, and A. R. Davidson. “A new group of phage anti-CRISPR genes inhibits the type I-E CRISPR-Cas system of Pseudomonas aeruginosa”. *mBio* 5.2 (2014), e00896.

[244] M. D. Hoffmann, S. Aschenbrenner, S. Grosse, K. Rapti, C. Domenger, J. Fakhiri, M. Mastel, K. Börner, R. Eils, D. Grimm, and D. Niopek. "Cell-specific CRISPR-Cas9 activation by microRNA-dependent expression of anti-CRISPR proteins". *Nucleic Acids Res* 47.13 (2019), e75.

[245] M. Nakamura, P. Srinivasan, M. Chavez, M. A. Carter, A. A. Dominguez, M. La Russa, M. B. Lau, T. R. Abbott, X. Xu, D. Zhao, Y. Gao, N. H. Kipniss, C. D. Smolke, J. Bondy-Denomy, and L. S. Qi. "Anti-CRISPR-mediated control of gene editing and synthetic circuits in eukaryotic cells". *Nat Commun* 10.1 (2019), p. 194.

[246] S. Aschenbrenner, S. M. Kallenberger, M. D. Hoffmann, A. Huck, R. Eils, and D. Niopek. "Coupling Cas9 to artificial inhibitory domains enhances CRISPR-Cas9 target specificity". *Sci Adv* 6.6 (2020), eaay0187.

[247] A. Pawluk, R. H. J. Staals, C. Taylor, B. N. J. Watson, S. Saha, P. C. Fineran, K. L. Maxwell, and A. R. Davidson. "Inactivation of CRISPR-Cas systems by anti-CRISPR proteins in diverse bacterial species". *Nat Microbiol* 1.8 (2016), p. 16085.

[248] J. Lee, A. Mir, A. Edraki, B. Garcia, N. Amrani, H. E. Lou, I. Gainetdinov, A. Pawluk, R. Ibraheim, X. D. Gao, P. Liu, A. R. Davidson, K. L. Maxwell, and E. J. Sontheimer. "Potent Cas9 Inhibition in Bacterial and Human Cells by AcrIIC4 and AcrIIC5 Anti-CRISPR Proteins". *mBio* 9.6 (2018), e02321–18.

[249] A. P. Hynes, G. M. Rousseau, D. Agudelo, A. Goulet, B. Amigues, J. Loehr, D. A. Romero, C. Fremaux, P. Horvath, Y. Doyon, C. Cambillau, and S. Moineau. "Widespread anti-CRISPR proteins in virulent bacteriophages inhibit a range of Cas9 proteins". *Nat Commun* 9.1 (2018), p. 2919.

[250] K. E. Watters, C. Fellmann, H. B. Bai, S. M. Ren, and J. A. Doudna. "Systematic discovery of natural CRISPR-Cas12a inhibitors". *Science* 362.6411 (2018), pp. 236–239.

[251] N. D. Marino, J. Y. Zhang, A. L. Borges, A. A. Sousa, L. M. Leon, B. J. Rauch, R. T. Walton, J. D. Berry, J. K. Joung, B. P. Kleinstiver, and J. Bondy-Denomy. "Discovery of widespread type I and type V CRISPR-Cas inhibitors". *Science* 362.6411 (2018), pp. 240–242.

[252] B. Garcia, J. Lee, A. Edraki, Y. Hidalgo-Reyes, S. Erwood, A. Mir, C. N. Trost, U. Seroussi, S. Y. Stanley, R. D. Cohn, J. M. Claycomb, E. J. Sontheimer, K. L. Maxwell, and A. R. Davidson. "Anti-CRISPR AcrIIA5 Potently Inhibits All Cas9 Homologs Used for Genome Editing". *Cell Rep* 29.7 (2019), 1739–1746.e5.

[253] L. B. Harrington, K. W. Doxzen, E. Ma, J.-J. Liu, G. J. Knott, A. Edraki, B. Garcia, N. Amrani, J. S. Chen, J. C. Cofsky, P. J. Kranzusch, E. J. Sontheimer, A. R. Davidson, K. L. Maxwell, and J. A. Doudna. "A Broad-Spectrum Inhibitor of CRISPR-Cas9". *Cell* 170.6 (2017), 1224–1233.e15.

[254] M. F. Sentmanat, S. T. Peters, C. P. Florian, J. P. Connelly, and S. M. Pruett-Miller. "A Survey of Validation Strategies for CRISPR-Cas9 Editing". *Sci Rep* 8.1 (2018), p. 888.

[255] A. Edraki, A. Mir, R. Ibraheim, I. Gainetdinov, Y. Yoon, C.-Q. Song, Y. Cao, J. Gallant, W. Xue, J. A. Rivera-Pérez, and E. J. Sontheimer. “A Compact, High-Accuracy Cas9 with a Dinucleotide PAM for In Vivo Genome Editing”. *Mol Cell* 73.4 (2019), 714–726.e4.

[256] F. A. Ran, L. Cong, W. X. Yan, D. A. Scott, J. S. Gootenberg, A. J. Kriz, B. Zetsche, O. Shalem, X. Wu, K. S. Makarova, E. V. Koonin, P. A. Sharp, and F. Zhang. “In vivo genome editing using Staphylococcus aureus Cas9”. *Nature* 520.7546 (2015), pp. 186–191.

[257] F. Schmidt and D. Grimm. “CRISPR genome engineering and viral gene delivery: a case of mutual attraction”. *Biotechnol J* 10.2 (2015), pp. 258–272.

[258] A. Leaver-Fay, M. Tyka, S. M. Lewis, O. F. Lange, J. Thompson, R. Jacak, K. Kaufman, P. D. Renfrew, C. A. Smith, W. Sheffler, I. W. Davis, S. Cooper, A. Treuille, D. J. Mandell, F. Richter, Y.-E. A. Ban, S. J. Fleishman, J. E. Corn, D. E. Kim, S. Lyskov, M. Berrondo, S. Mentzer, Z. Popović, J. J. Havranek, J. Karanicolas, R. Das, J. Meiler, T. Kortemme, J. J. Gray, B. Kuhlman, D. Baker, and P. Bradley. “ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules”. *Methods Enzymol* 487 (2011), pp. 545–574.

[259] J. Lee, H. Mou, R. Ibraheim, S.-Q. Liang, P. Liu, W. Xue, and E. J. Sontheimer. “Tissue-restricted genome editing in vivo specified by microRNA-repressible anti-CRISPR proteins”. *RNA* 25.11 (2019), pp. 1421–1431.

[260] M. Hirosawa, Y. Fujita, and H. Saito. “Cell-Type-Specific CRISPR Activation with MicroRNA-Responsive AcrllA4 Switch”. *ACS Synth Biol* 8.7 (2019), pp. 1575–1582.

[261] B. A. Osuna, S. Karambelkar, C. Mahendra, K. A. Christie, B. Garcia, A. R. Davidson, B. P. Kleinstiver, S. Kilcher, and J. Bondy-Denomy. “Listeria Phages Induce Cas9 Degradation to Protect Lysogenic Genomes”. *Cell Host Microbe* 28.1 (2020), 31–40.e9.

[262] K. E. Watters, H. Shivram, C. Fellmann, R. J. Lew, B. McMahon, and J. A. Doudna. “Potent CRISPR-Cas9 inhibitors from Staphylococcus genomes”. *Proc Natl Acad Sci U S A* 117.12 (2020), pp. 6531–6539.

[263] G. Song, F. Zhang, X. Zhang, X. Gao, X. Zhu, D. Fan, and Y. Tian. “AcrIIA5 Inhibits a Broad Range of Cas9 Orthologs by Preventing DNA Target Cleavage”. *Cell Rep* 29.9 (2019), 2579–2589.e4.

[264] D. Agudelo, S. Carter, M. Velimirovic, A. Duringer, J.-F. Rivest, S. Levesque, J. Loehr, M. Mouchiroud, D. Cyr, P. J. Waters, M. Laplante, S. Moineau, A. Goulet, and Y. Doyon. “Versatile and robust genome editing with Streptococcus thermophilus CRISPR1-Cas9”. *Genome Res* 30.1 (2020), pp. 107–117.

[265] J. Bonet, Z. Harteveld, F. Sesterhenn, A. Scheck, and B. E. Correia. “rstoolbox - a Python library for large-scale analysis of computational protein design data and structural bioinformatics”. *BMC Bioinformatics* 20.1 (2019), p. 240.

[266] Y. Zhang and J. Skolnick. “TM-align: a protein structure alignment algorithm based on the TM-score”. *Nucleic Acids Res* 33.7 (2005), pp. 2302–2309.

[267] S. J. Fleishman, A. Leaver-Fay, J. E. Corn, E.-M. Strauch, S. D. Khare, N. Koga, J. Ashworth, P. Murphy, F. Richter, G. Lemmon, J. Meiler, and D. Baker. “RosettaScripts: A Scripting Language Interface to the Rosetta Macromolecular Modeling Suite”. *PLoS One* 6.6 (2011), e20161.

[268] D. G. Gibson, L. Young, R.-Y. Chuang, J. C. Venter, C. A. Hutchison, and H. O. Smith. “Enzymatic assembly of DNA molecules up to several hundred kilobases”. *Nat Methods* 6.5 (2009), pp. 343–345.

[269] C. Engler and S. Marillonnet. “Combinatorial DNA assembly using Golden Gate cloning”. *Methods Mol Biol* 1073 (2013), pp. 141–156.

[270] N. Amrani, X. D. Gao, P. Liu, A. Edraki, A. Mir, R. Ibraheim, A. Gupta, K. E. Sasaki, T. Wu, P. D. Donohoue, A. H. Settle, A. M. Lied, K. McGovern, C. K. Fuller, P. Cameron, T. G. Fazzio, L. J. Zhu, S. A. Wolfe, and E. J. Sontheimer. “NmeCas9 is an intrinsically high-fidelity genome-editing platform”. *Genome Biol* 19.1 (2018), p. 214.

[271] A.-K. Herrmann, C. Bender, E. Kienle, S. Grosse, J. El Andari, J. Botta, N. Schürmann, E. Wiedtke, D. Niopek, and D. Grimm. “A Robust and All-Inclusive Pipeline for Shuffling of Adeno-Associated Viruses”. *ACS Synth Biol* 8.1 (2019), pp. 194–206.

[272] K. Börner, D. Niopek, G. Cotugno, M. Kaldenbach, T. Pankert, J. Willemsen, X. Zhang, N. Schürmann, S. Mockenhaupt, A. Serva, M.-S. Hiet, E. Wiedtke, M. Castoldi, V. Starkuviene, H. Erfle, D. F. Gilbert, R. Bartenschlager, M. Boutros, M. Binder, K. Streetz, H.-G. Kräusslich, and D. Grimm. “Robust RNAi enhancement via human Argonaute-2 overexpression from plasmids, viral vectors and cell lines”. *Nucleic Acids Res* 41.21 (2013), e199.

[273] C. T. Rueden, J. Schindelin, M. C. Hiner, B. E. DeZonia, A. E. Walter, E. T. Arena, and K. W. Eliceiri. “ImageJ2: ImageJ for the next generation of scientific image data”. *BMC Bioinformatics* 18.1 (2017), p. 529.

[274] C. A. Schneider, W. S. Rasband, and K. W. Eliceiri. “NIH Image to ImageJ: 25 years of image analysis”. *Nat Methods* 9.7 (2012), pp. 671–675.

[275] S. J. Fleishman, T. A. Whitehead, E.-M. Strauch, J. E. Corn, S. Qin, H.-X. Zhou, J. C. Mitchell, O. N. A. Demerdash, M. Takeda-Shitaka, G. Terashi, I. H. Moal, X. Li, P. A. Bates, M. Zacharias, H. Park, J.-s. Ko, H. Lee, C. Seok, T. Bourquard, J. Bernauer, A. Poupon, J. Azé, S. Soner, Ş. K. Ovalı, P. Ozbek, N. B. Tal, T. Haliloglu, H. Hwang, T. Vreven, B. G. Pierce, Z. Weng, L. Pérez-Cano, C. Pons, J. Fernández-Recio, F. Jiang, F. Yang, X. Gong, L. Cao, X. Xu, B. Liu, P. Wang, C. Li, C. Wang, C. H. Robert, M. Guharoy, S. Liu, Y. Huang, L. Li, D. Guo, Y. Chen, Y. Xiao, N. London, Z. Itzhaki, O. Schueler-Furman, Y. Inbar, V. Potapov, M. Cohen, G. Schreiber, Y. Tsuchiya, E. Kanamori, D. M. Standley, H. Nakamura, K. Kinoshita, C. M. Driggers, R. G. Hall, J. L. Morgan, V. L. Hsu, J. Zhan, Y. Yang, Y. Zhou, P. L. Kastritis, A. M. J. J. Bonvin, W. Zhang, C. J. Camacho, K. P. Kilambi, A. Sircar, J. J. Gray, M. Ohue, N. Uchikoga, Y. Matsuzaki, T. Ishida, Y. Akiyama, R. Khashan, S. Bush, D. Fouches, A. Tropsha, J. Esquivel-Rodríguez, D. Kihara, P. B. Stranges, R.

Jacak, B. Kuhlman, S.-Y. Huang, X. Zou, S. J. Wodak, J. Janin, and D. Baker. “Community-Wide Assessment of Protein-Interface Modeling Suggests Improvements to Design Methodology”. *J Mol Biol* 414.2 (2011), pp. 289–302.

[276] H. Lu, Q. Zhou, J. He, Z. Jiang, C. Peng, R. Tong, and J. Shi. “Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials”. *Sig Transduct Target Ther* 5.1 (2020), pp. 1–23.

[277] N. P. King, W. Sheffler, M. R. Sawaya, B. S. Vollmar, J. P. Sumida, I. André, T. Gonen, T. O. Yeates, and D. Baker. “Computational Design of Self-Assembling Protein Nanomaterials with Atomic Level Accuracy”. *Science* 336.6085 (2012), pp. 1171–1174.

[278] S. E. Boyken, Z. Chen, B. Groves, R. A. Langan, G. Oberdorfer, A. Ford, J. M. Gilmore, C. Xu, F. DiMaio, J. H. Pereira, B. Sankaran, G. Seelig, P. H. Zwart, and D. Baker. “De novo design of protein homo-oligomers with modular hydrogen-bond network–mediated specificity”. *Science* 352.6286 (2016), pp. 680–687.

[279] G. W. Foight, Z. Wang, C. T. Wei, P. Jr Greisen, K. M. Warner, D. Cunningham-Bryant, K. Park, T. J. Brunette, W. Sheffler, D. Baker, and D. J. Maly. “Multi-input chemical control of protein dimerization for programming graded cellular responses”. *Nat Biotechnol* 37.10 (2019), pp. 1209–1216.

[280] L. T. Dang, Y. Miao, A. Ha, K. Yuki, K. Park, C. Y. Janda, K. M. Jude, K. Mohan, N. Ha, M. Vallon, J. Yuan, J. G. Vilches-Moure, C. J. Kuo, K. C. Garcia, and D. Baker. “Receptor subtype discrimination using extensive shape complementary designed interfaces”. *Nat Struct Mol Biol* 26.6 (2019), pp. 407–414.

[281] S. J. Fleishman, J. E. Corn, E.-M. Strauch, T. A. Whitehead, J. Karanicolas, and D. Baker. “Hotspot-Centric De Novo Design of Protein Binders”. *J Mol Biol* 413.5 (2011), pp. 1047–1062.

[282] C. Chothia and J. Janin. “Principles of protein–protein recognition”. *Nature* 256.5520 (1975), pp. 705–708.

[283] O. Keskin, A. Gursoy, B. Ma, and R. Nussinov. “Principles of ProteinProtein Interactions: What are the Preferred Ways For Proteins To Interact?” *Chem Rev* 108.4 (2008), pp. 1225–1244.

[284] D. Rajamani, S. Thiel, S. Vajda, and C. J. Camacho. “Anchor residues in protein–protein interactions”. *Proc Natl Acad Sci U S A* 101.31 (2004), pp. 11287–11292.

[285] E. Procko, G. Y. Berguig, B. W. Shen, Y. Song, S. Frayo, A. J. Convertine, D. Margineantu, G. Booth, B. E. Correia, Y. Cheng, W. R. Schief, D. M. Hockenbery, O. W. Press, B. L. Stoddard, P. S. Stayton, and D. Baker. “A Computationally Designed Inhibitor of an Epstein-Barr Viral Bcl-2 Protein Induces Apoptosis in Infected Cells”. *Cell* 157.7 (2014), pp. 1644–1656.

[286] D. Baran, M. G. Pszolla, G. D. Lapidoth, C. Norn, O. Dym, T. Unger, S. Albeck, M. D. Tyka, and S. J. Fleishman. “Principles for computational design of binding antibodies”. *Proc Natl Acad Sci U S A* 114.41 (2017), pp. 10900–10905.

[287] A. Scheck, S. Rosset, M. Defferrard, A. Loukas, J. Bonet, P. Vandergheynst, and B. E. Correia. “RosettaSurf - a surface-centric computational design approach”. *bioRxiv* (2021), p. 2021.06.16.448645.

[288] L. v. d. Maaten and G. Hinton. “Visualizing Data using t-SNE”. *J Mach Learn Res* 9.86 (2008), pp. 2579–2605.

[289] C. Liu, N. P. Seeram, and H. Ma. “Small molecule inhibitors against PD-1/PD-L1 immune checkpoints and current methodologies for their development: a review”. *Cancer Cell Int* 21.1 (2021), p. 239.

[290] F. Zhang, H. Wei, X. Wang, Y. Bai, P. Wang, J. Wu, X. Jiang, Y. Wang, H. Cai, T. Xu, and A. Zhou. “Structural basis of a novel PD-L1 nanobody for immune checkpoint blockade”. *Cell Discov* 3 (2017), p. 17004.

[291] X. Pan and T. Kortemme. “Recent advances in de novo protein design: Principles, methods, and applications”. *J Biol Chem* 296 (2021), p. 100558.

[292] P. B. Stranges and B. Kuhlman. “A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds”. *Protein Sci* 22.1 (2013), pp. 74–82.

[293] A. Goldenzweig and S. J. Fleishman. “Principles of Protein Stability and Their Application in Computational Design”. *Annu Rev Biochem* 87 (2018), pp. 105–129.

[294] L. Cao, I. Goreshnik, B. Coventry, J. B. Case, L. Miller, L. Kozodoy, R. E. Chen, L. Carter, A. C. Walls, Y.-J. Park, E.-M. Strauch, L. Stewart, M. S. Diamond, D. Veesler, and D. Baker. “De novo design of picomolar SARS-CoV-2 miniprotein inhibitors”. *Science* 370.6515 (2020), pp. 426–431.

[295] N. H. Joh, T. Wang, M. P. Bhate, R. Acharya, Y. Wu, M. Grabe, M. Hong, G. Grigoryan, and W. F. DeGrado. “De novo design of a transmembrane Zn2+-transporting four-helix bundle”. *Science* 346.6216 (2014), pp. 1520–1524.

[296] J. B. Bale, S. Gonen, Y. Liu, W. Sheffler, D. Ellis, C. Thomas, D. Cascio, T. O. Yeates, T. Gonen, N. P. King, and D. Baker. “Accurate design of megadalton-scale two-component icosahedral protein complexes”. *Science* 353.6297 (2016), pp. 389–394.

[297] B. Kuhlman and D. Baker. “Native protein sequences are close to optimal for their structures”. *Proc Natl Acad Sci U S A* 97.19 (2000), pp. 10383–10388.

[298] J. Miller, C. Zeng, N. S. Wingreen, and C. Tang. “Emergence of highly designable protein-backbone conformations in an off-lattice model”. *Proteins: Struct Funct Genet* 47.4 (2002), pp. 506–512.

[299] F. Pan, Y. Zhang, X. Liu, and J. Zhang. *Estimating the Designability of Protein Structures*. Tech. rep. bioRxiv, 2021, p. 2021.11.03.467111.

[300] S. P. Leelananda, R. L. Jernigan, and A. Kloczkowski. “Predicting Designability of Small Proteins from Graph Features of Contact Maps”. *J Comput Biol* 23.5 (2016), pp. 400–411.

[301] S. Minami, N. Kobayashi, T. Sugiki, T. Nagashima, T. Fujiwara, R. Koga, G. Chikenji, and N. Koga. “Exploration of novel αβ-protein folds through de novo design”. *bioRxiv* (2021), p. 2021.08.06.455475.

[302] A. A. Vorobieva, P. White, B. Liang, J. E. Horne, A. K. Bera, C. M. Chow, S. Gerben, S. Marx, A. Kang, A. Q. Stiving, S. R. Harvey, D. C. Marx, G. N. Khan, K. G. Fleming, V. H. Wysocki, D. J. Brockwell, L. K. Tamm, S. E. Radford, and D. Baker. “De novo design of transmembrane beta-barrels”. *Science* 371.6531 (2021), eabc8182.

[303] G. Offer, M. R. Hicks, and D. N. Woolfson. “Generalized Crick Equations for Modeling Noncanonical Coiled Coils”. *J Struct Biol* 137.1 (2002), pp. 41–53.

[304] J. Novotný, R. E. Bruccoleri, and J. Newell. “Twisted hyperboloid (strophoid) as a model of β-barrels in proteins”. *J Mol Biol* 177.3 (1984), pp. 567–573.

[305] W. R. Taylor. “A ‘periodic table’ for protein structures”. *Nature* 416.6881 (2002), pp. 657–660.

[306] W. R. Taylor, V. Chelliah, S. M. Hollup, J. T. MacDonald, and I. Jonassen. “Probing the "dark matter" of protein fold space”. *Structure* 17.9 (2009), pp. 1244–1252.

[307] C. Chothia and A. V. Finkelstein. “The Classification and Origins of Protein Folding Patterns”. *Annu Rev Biochem* 59.1 (1990), pp. 1007–1035.

[308] F. E. Cohen, M. J. E. Sternberg, and W. R. Taylor. “Analysis and prediction of protein β -sheet structures by a combinatorial approach”. *Nature* 285.5764 (1980), pp. 378–382.

[309] J. Zhou and G. Grigoryan. “Rapid search for tertiary fragments reveals protein sequence–structure relationships”. *Protein Sci* 24.4 (2015), pp. 508–524.

[310] J. Zhou and G. Grigoryan. “A C++ library for protein sub-structure search”. *bioRxiv* (2020), p. 2020.04.26.062612.

[311] W. R. Taylor, G. J. Bartlett, V. Chelliah, D. Klose, K. Lin, T. Sheldon, and I. Jonassen. “Prediction of protein structure from ideal forms”. *Proteins: Structure, Function, and Bioinformatics* 70.4 (2008), pp. 1610–1619.

[312] J. Bonet, S. Wehrle, K. Schriever, C. Yang, A. Billet, F. Sesterhenn, A. Scheck, F. Sverrisson, S. Vollers, R. Lourman, M. Villard, S. Rosset, and B. E. Correia. “Rosetta FunFolDes - a general framework for the computational design of functional proteins”. *bioRxiv* (2018), p. 378976.

[313] G. Bhardwaj, V. K. Mulligan, C. D. Bahl, J. M. Gilmore, P. J. Harvey, O. Cheneval, G. W. Buchko, S. V. Pulavarti, Q. Kaas, A. Eletsky, P.-S. Huang, W. A. Johnsen, P. Greisen, G. J. Rocklin, Y. Song, T. W. Linsky, A. Watkins, S. A. Rettie, X. Xu, L. P. Carter, R. Bonneau, J. M. Olson, E. Coutsias, C. E. Correnti, T. Szyperski, D. J. Craik, and D. Baker. “Accurate de novo design of hyperstable constrained peptides”. *Nature* 538.7625 (2016), pp. 329–335.

[314] R. Koga, M. Yamamoto, T. Kosugi, N. Kobayashi, T. Sugiki, T. Fujiwara, and N. Koga. “Robust folding of a de novo designed ideal protein even with most of the core mutated to valine”. *Proc Natl Acad Sci U S A* 117.49 (2020), pp. 31149–31156.

[315] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. “Basic local alignment search tool”. *J Mol Biol* 215.3 (1990), pp. 403–410.

[316] J. Lever, M. Krzywinski, and N. Altman. “Principal component analysis”. *Nat Methods* 14.7 (2017), pp. 641–642.

[317] F. B. Salisbury. “Natural selection and the complexity of the gene”. *Nature* 224.5217 (1969), pp. 342–343.

[318] J. M. Smith. “Natural selection and the concept of a protein space”. *Nature* 225.5232 (1970), pp. 563–564.

[319] I. V. Korendovych and W. F. DeGrado. “De novo protein design, a retrospective”. *Q Rev Biophys* 53 (2020).

[320] D. N. Woolfson. “A Brief History of De Novo Protein Design: Minimal, Rational, and Computational”. *J Mol Biol* (2021), p. 167160.

[321] D. Baker. “What has de novo protein design taught us about protein folding and biophysics?” *Protein Sci* 28.4 (2019), pp. 678–683.

[322] A. Wagner. “Robustness and evolvability: a paradox resolved”. *Proc R Soc B: Biol Sci* 275.1630 (2008), pp. 91–100.

[323] J. D. Bloom, D. A. Drummond, F. H. Arnold, and C. O. Wilke. “Structural determinants of the rate of protein evolution in yeast”. *Mol Biol Evol* 23.9 (2006), pp. 1751–1761.

[324] F. H. C. Crick. “The Fourier transform of a coiled-coil”. *Acta Cryst* 6.8-9 (1953), pp. 685–689.

[325] S. Chaudhury, S. Lyskov, and J. J. Gray. “PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta”. *Bioinformatics* 26.5 (2010), pp. 689–691.

[326] B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker. “Design of a novel globular protein fold with atomic-level accuracy”. *Science* 302.5649 (2003), pp. 1364–1368.

[327] N. K. Fox, S. E. Brenner, and J.-M. Chandonia. “SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures”. *Nucleic Acids Res* 42.D1 (2014), pp. D304–D309.

[328] V. M. Panaretos and Y. Zemel. “Statistical Aspects of Wasserstein Distances”. *Annu Rev Stat Appl* 6.1 (2019), pp. 405–431.

[329] R. N. McLaughlin Jr, F. J. Poelwijk, A. Raman, W. S. Gosal, and R. Ranganathan. “The spatial architecture of protein function and adaptation”. *Nature* 491.7422 (2012), pp. 138–142.

[330] F. Martins, L. Sofiya, G. P. Sykiotis, F. Lamine, M. Maillard, M. Fraga, K. Shabafrouz, C. Ribi, A. Cairoli, Y. Guex-Crosier, T. Kuntzer, O. Michielin, S. Peters, G. Coukos, F. Spertini, J. A. Thompson, and M. Obeid. “Adverse effects of immune-checkpoint inhibitors: epidemiology, management and surveillance”. *Nat Rev Clin Oncol* 16.9 (2019), pp. 563–580.

[331] G. Sormani, Z. Harteveld, S. Rosset, B. Correia, and A. Laio. “A Rosetta-based protein design protocol converging to natural sequences”. *J Chem Phys* 154.7 (2021), p. 074114.

[332] S. R. Eddy. “Multiple alignment using hidden Markov models”. *Proc Int Conf Intell Syst Mol Biol* 3 (1995), pp. 114–120.

[333] R. Rao, J. Meier, T. Sercu, S. Ovchinnikov, and A. Rives. “Transformer protein language models are unsupervised structure learners”. *International Conference on Learning Representations (ICLR)*. 2020, p. 2020.12.15.422761.

[334] P. Gainza-Cirauqui and B. E. Correia. “Computational protein design—the next generation tool to expand synthetic biology applications”. *Curr Opin Biotechnol.* Tissue, Cell and Pathway Engineering 52 (2018), pp. 145–152.

[335] P. Gainza, K. E. Roberts, I. Georgiev, R. H. Lilien, D. A. Keedy, C.-Y. Chen, F. Reza, A. C. Anderson, D. C. Richardson, J. S. Richardson, and B. R. Donald. “OSPREY: Protein Design with Ensembles, Flexibility, and Provable Algorithms”. *Methods Enzymol* 523 (2013), pp. 87–107.

[336] C.-Y. Chen, I. Georgiev, A. C. Anderson, and B. R. Donald. “Computational structure-based redesign of enzyme activity”. *Proc Natl Acad Sci U S A* 106.10 (2009), pp. 3764–3769.

[337] K. M. Frey, I. Georgiev, B. R. Donald, and A. C. Anderson. “Predicting resistance mutations using protein design algorithms”. *Proc Natl Acad Sci U S A* 107.31 (2010), pp. 13707–13712.

[338] D. N. Bolon and S. L. Mayo. “Enzyme-like proteins by computational design”. *Proc Natl Acad Sci U S A* 98.25 (2001), pp. 14274–14279.

[339] M. Shimaoka, J. M. Shifman, H. Jing, J. Takagi, S. L. Mayo, and T. A. Springer. “Computational design of an integrin I domain stabilized in the open high affinity conformation”. *Nat Struct Mol Biol* 7.8 (2000), pp. 674–678.

[340] Z. Li and H. A. Scheraga. “Monte Carlo-minimization approach to the multiple-minima problem in protein folding”. *Proc Natl Acad Sci U S A* 84.19 (1987), pp. 6611–6615.

[341] P. Gainza, H. M. Nisonoff, and B. R. Donald. “Algorithms for protein design”. *Curr Opin Struct Biol.* Engineering and design • Membranes 39 (2016), pp. 16–26.

[342] D. E. Kim, B. Blum, P. Bradley, and D. Baker. “Sampling Bottlenecks in De novo Protein Structure Prediction”. *J Mol Biol* 393.1 (2009), pp. 249–260.

[343] F. Perez and B. E. Granger. “IPython: A System for Interactive Scientific Computing”. *Comput Sci Eng* 9.3 (2007), pp. 21–29.

[344] W. McKinney. “Data Structures for Statistical Computing in Python”. *Proc 9th Python Sci Conf* (2010), pp. 56–61.

[345] T. D. Schneider and R. Stephens. “Sequence logos: a new way to display consensus sequences”. *Nucleic Acids Res* 18.20 (1990), pp. 6097–6100.

[346] G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. “Stereochemistry of polypeptide chain configurations”. *J Mol Biol* 7.1 (1963), pp. 95–99.

[347] J. D. Thompson, T. J. Gibson, and D. G. Higgins. “Multiple sequence alignment using ClustalW and ClustalX”. *Curr Protoc Bioinformatics* Chapter 2 (2002), Unit 2.3.

[348] S. C. Potter, A. Luciani, S. R. Eddy, Y. Park, R. Lopez, and R. D. Finn. “HMMER web server: 2018 update”. *Nucleic Acids Res* 46.W1 (2018), W200–W204.

[349] J. D. Hunter. “Matplotlib: A 2D Graphics Environment”. *Comput Sci Eng* 9.3 (2007), pp. 90–95.

[350] M. Waskom, O. Botvinnik, D. O’Kane, P. Hobson, J. Ostblom, S. Lukauskas, D. C. Gemperline, T. Augspurger, Y. Halchenko, J. B. Cole, J. Warmenhoven, J. d. Ruiter, C. Pye, S. Hoyer, J. Vanderplas, S. Villalba, G. Kunter, E. Quintero, P. Bachant, M. Martin, K. Meyer, A. Miles, Y. Ram, T. Brunner, T. Yarkoni, M. L. Williams, C. Evans, C. Fitzgerald, Brian, and A. Qalieh. *mwaskom/seaborn: v0.9.0 (July 2018)*. 2018.

[351] A. Stein and T. Kortemme. “Improvements to Robotics-Inspired Conformational Sampling in Rosetta”. *PLoS One* 8.5 (2013), e63090.

[352] J. S. McLellan, M. Chen, A. Kim, Y. Yang, B. S. Graham, and P. D. Kwong. “Structural basis of respiratory syncytial virus neutralization by motavizumab”. *Nat Struct Mol Biol* 17.2 (2010), pp. 248–250.

[353] A. Lartigue, V. Campanacci, A. Roussel, A. M. Larsson, T. A. Jones, M. Tegoni, and C. Cambillau. “X-ray Structure and Ligand Binding Study of a Moth Chemosensory Protein*”. *J Biol Chem* 277.35 (2002), pp. 32094–32098.

[354] R. G. Coleman, M. Carchia, T. Sterling, J. J. Irwin, and B. K. Shoichet. “Ligand Pose and Orientational Sampling in Molecular Docking”. *PLoS One* 8.10 (2013), e75992.

[355] A. Andreeva, D. Howorth, J.-M. Chandonia, S. E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin. “Data growth and its impact on the SCOP database: new developments”. *Nucleic Acids Res* 36.Database issue (2008), pp. D419–425.

Zander HARTEVELD

Avenue de Cour 11, 1007 Lausanne, CH
(+41) 76 466 50 81
zandermilanh@gmail.com - github.com/zanderharteveld

PERSONAL INFORMATION

Citizenship: Swiss, Dutch

SUMMARY

Advance, accelerate and re-imaging next generation diagnostics and therapeutics through deep learning and structural biology. Strengths in communication, coordination, and a drive to implement efficient solutions. Very strong team player with a preference for working in small targeted groups where innovation and rationality are highly valued.

EDUCATION

Ph.D., Biotechnology and Bioengineering, EPFL, Switzerland, October 2017

- Dissertation title, supervisor*: Towards automated functional de novo protein design methodologies*, Prof. Bruno Correia

MS, Life Sciences and Technology, EPFL, Switzerland, September 2015

- Thesis title, supervisors: *SOLUXION2.0 - A deep neural network approach to predict protein solubility*. Prof. Bruno Correia, Jens Kringelum (Ph.D., Evaxion Biotech), Christian Skjødt Hansen (Ph.D., Evaxion Biotech)

BS, Health Sciences and Technology, ETH , Switzerland, September 2012

WORK EXPERIENCE

Laboratory of Protein Design and Immunoengineering (LPDI), EPFL, Switzerland, October 2017

- Functional de novo design methodologies:
    - Dissertation encompassing the (1) implementation of functional de novo protein design methodologies, (2) usage therefrom to solve and answer complex biological problem sets and (3) statistical analysis and interpretation of biochemical and biophysical experimental data.
        - Interdisciplinary laboratory (computational and experimental arms); closely work with biochemists, cell biologists and neuroscientists for designing experiments and analyzing experimental data.
        - Collaborations with theoretical machine learning laboratories (e.g. Prof. Pierre Vandergheynst, LTS2 EPFL); closely work on the development and applying of novel theoretical frameworks to biological relevant problems.
    - *De novo* protein fold design.
        - I led the development of *de novo* protein design tools Genesis and Topobuilder and applied both to design synthetic proteins. Topobuilder automatically builds proteins with customized geometries from minimal topological descriptions. Genesis uses a neural network to ultra-fast build physically realistic protein backbones from topological structure descriptions.
    - *De novo* PPI design.
        - Developed pipelines for *de novo* PPI design and applied to design experimentally tested proteins for Parkinson's disease, cancer immunotherapy and antivirals. Helical peptides against specific sites on alpha-synuclein fibrils (Parkinson's) were first placed using MaSIF-site, and subsequently designed with a specific protocol in Rosetta adhering to the structural repetitiveness of the fibril. A *de novo* designed PD-L1 inhibitor for cancer immunotherapy and an orthologue specific anti-CRISPR protein for in cellulo Cas-9 control were designed with adapted protocols.

EMD Serono, Lausanne, Switzerland, December 2017 - December 2018

- Computational design of optimized immunogens:
    - Combination of several computational techniques ranging from motif grafting and oligomer stabilization to protein surface design using Rosetta.
    - Data analysis and candidate selections for in vitro characterizations.
    - Evaluate, report and communicate results to collaborators.

Evaxion Biotech, Copenhagen, Denmark, January 2017 - October 2017

- MHC II epitope prediction:
    - Implementation of a neural network for MHC II epitope prediction from mass spectrometry MHC eluted ligand data.
    - Method leverages mixed data types and ligands with multiple potential allele annotations for pan-specific predictions.
    - Software in use to rank and select potential antigens for in vitro and in vivo characterization.
- Protein solubility prediction:
    - Design and implementation of a recurrent neural network to predict protein solubility in E. coli solely from sequence information.
    - Software guides the companies' decision on the antigen selection process for in vitro and in vivo characterization.
    - Pathed the way for several new generations neural networks to predict protein solubility within the company.

SKILLS

Computer Languages:	Python, C++ (basics)
Software & Tools:	PyTorch, PyTorch geometric, Numpy, Pandas, Rosetta, MaSIF
Languages:	German, French, English
Interests:	Structural and systems biology, machine learning, computer vision, organic chemistry

SERVICE

Weekly IBI-EDBB Grad Student Mini-Symposium, Committee Member and chairing sessions, January 2018
NCCR Chemical Biology Member, Interdisciplinary Research in Switzerland, January 2018

MENTORING

Master Students:
Tiana Schwab, now Quality Assurance at CSL Behring, Switzerland
Jayne Marsden, now Quality Assurance Specialist at Vifor Pharma, Switzerland
Julius Upmeier zu Belzen, now Ph.D. at Berlin Institute of Health, Germany
Karen Schriever, now Ph.D. at KTH Stockholm, Sweden

Bachelor students:
Rachael White (Rosetta intern)
Célina Chkroun, now MS in Life Sciences (Internet of Things), EPFL
Charles Berger, now MS in Life Sciences (Biophotonic and Bioimaging), EPFL
Arthur Valentin, now MS in Life Sciences (Biocomputing), EPFL

TEACHING ASSISTANT

Biological Chemistry II (+100 Students), 2018/2019 and 2019/2020
Biological Chemistry I (+100 Students), 2018/2019 and 2019/2020
Mathematics 2 MAN (+50 students, linear algebra I, analytical geometry I) for 2nd semester bachelor students (any field), 2017/2018

ORAL COMMUNICATIONS

Coming:
Keystone Symposium, Computational Design and Modeling of Biomolecules; joint with: Antibody as drugs, January 2022
Poster: Genesis: Tailored *de novo* protein design with deep neural networks.

Previous:
The Protein Society Symposium, *remote*, July 2021
Poster: A generic framework for layered de novo protein design.

Winter RosettaCon, *remote*, March 2021
Talk: TopoBuilder: A generic framework for layered de novo protein design.

EPFL Bioengineering Day, Lausanne, Switzerland, October 2017
Poster: Z Harteveld, J Bonet and BE Correia. Denovo design of functional immunoglobulin - like folds.

RosettaCon, Leavenworth, Washington, USA, August 2017
Poster: Z Harteveld, C Garde, BE Correia, AH Mattsson, C S Hansen. SOLUXION2.0 - A deep neural network approach to predict protein solubility.

EPFL Bioengineering Day, Lausanne, Switzerland, October 2016
Poster: Z Harteveld, M Garg, C Yang, J Bonet, BE Correia. Computational Interface Design - Motavizumab Re-purposing from RSV to MPV.

PUBLICATIONS

In progress:

Z Harteveld, J Bonet, S Rosset, C Yang, F Sesterhenn, BE Correia. A generic framework for layered de novo protein design.

Z Harteveld*, A Van Hall-Beauvais*, J Southern*, M Defferrard, P Vandergheynst, MM Bronstein, A Loukas, and BE. Correia. Tailored *de novo* protein design with deep neural networks.

P Gainza, S Wehrle, A Van Hall-Beauvais, **Z Harteveld**, T Shuguang, A Petruzzella, J Marsden, S Rosset, S Georgeon, E Oricchio, G Gao, BE Correia. De novo design of site-specific protein inhibitors using interaction fingerprints.

Published:

MD Hoffmann, J Mathony, J Upmeier Zu Belzen, **Z Harteveld**, C Stengl, BE Correia, R Eils, D Niopek (2021). Optogenetic control of Neisseria meningitidis Cas9 genome editing using an engineered, light-switchable anti-CRISPR protein. Nucleic Acids Research. https://doi.org/10.1093/nar/gkaa1198

G Sormani, **Z Harteveld**, S Rosset, B Correia, A Laio (2021). A Rosetta-based protein design protocol converging to natural sequences. The Journal of Chemical Physics. https://doi.org/10.1063/5.0039240

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

J Bonet, **Z Harteveld**, F Sesterhenn, A Scheck, BE Correia (2019). rstoolbox - A Python libraryfor large-scale analysis of computational protein design data and structural bioinformatics. BMC Bioinformatics. https://doi.org/10.1186/s12859-019-2796-3

F Bubeck, MD Hoffmann, **Z Harteveld**, S Aschenbrenner, A Bietz, MC Waldhauer, K Börner, J Fakhiri, C Schmelas, L Dietz, D Grimm, BE Correia, R Eils, D Niopek (2018). Engineered anti-CRISPR proteins for optogenetic control of CRISPR–Cas9. Nature Methods. https://doi.org/10.1038/s41592-018-0178-9

* co-first authorship

REFERENCES

Prof. Bruno E. Correia, EPFL (bruno.correia@epfl.ch)
Prof. Dominik Niopek, TU Darmstadt (dominik.niopek@tu-darmstadt.de)
Jens Kringelum, Ph.D., Evaxion Biotech (jkgm@evaxion-biotech.com)
Christian Skjødt Hansen, Ph.D., Evaxion Biotech (csh@evaxion-biotech.com)
Pablo Gainza Cirauqui, Ph.D., Monte Rosa therapeutics (pgainza@monterosatx.com)
Andreas Loukas, Ph.D., Scientist EPFL (andreas.loukas@epfl.ch)