Fold fragments in protein evolution and design

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Francisco Antonio Lobos González
aus Concepción, Chile

Tübingen
2022

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation: 15.11.2022

Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatterin:	Prof. Dr. Birte Höcker
2. Berichterstatter:	Prof. Dr. Oliver Kohlbacher

Contents

List of Tables

List of Figures

Abbreviations & Acronyms

Å Ångström
$Abs_{280\,nm}$ Absorbance at 280 nm
CSB Computational Structural Biology Toolbox
Cα Alpha carbon
DNA Deoxyribonucleic acid
FAD Flavin adenine nucleotide
Fld Flavodoxin
FMN Flavin mononucleotide
fTHF 10-formyltetrahydrofolate
GAR Glycinamide ribonucleotide
HMM Hidden Markov Model
M Mean
Mdn Median
MRE Mean residue ellipticity
MSA Multiple sequence alignment
NAD Nicotinamide adenine nucleotide
NADP NADP phosphate
PAGE Polyacrylamide gel electrophoresis
PBP Periplasmic binding protein
PDB Protein Data Bank
PLIP Protein-Ligand Interaction Profiler
PP_i Pyrophosphate
ProFAR *N*′-[(5′-phosphoribosyl)formimino]-5-aminoimidazole-4-carboxamide ribonucleotide
PRPP Phosphoribosyl pyrophosphate
PROSS Protein Repair One-Stop Shop
PSA Pairwise sequence alignment
PSI-BLAST Position-Specific Iterative Basic Local Alignment Search Tool
REU Rosetta Energy Unit
RMSD Root mean square deviation
SAM S-adenosyl-L-methionine
sccs SCOP(e) concise classification string
SCOPe Structural Classification of Proteins — extended

SD Standard deviation
SDS Sodium dodecyl sulfate
$\theta_{222\,nm}$ Ellipticity at 222 nm
TIM Triose phosphate isomerase
TM-score Template modeling score
TPR Tetratricopeptide repeat

Secondary structure representation

Throughout this text secondary structure elements will be represented as follows:

Zusammenfassung

Domänen, die strukturellen, funktionellen und evolutionären Komponenten von Proteinen, fungieren als Bausteine, die zur Diversifizierung des Proteinuniversums, wie wir es kennen, beitragen. Es wird angenommen, dass Domänen aus ursprünglichen Fragmenten in Subdomänengröße entstanden sind und sich weiterentwickelt haben. Diese Fragmente wurden in bestehenden Proteinfaltungen in unterschiedlichen strukturellen und/oder funktionellen Zusammenhängen wiederverwendet. Darüber hinaus kann die Rekombination dieser Elemente zur Erzeugung von Proteinen mit neuartigen Strukturen und Funktionen führen, was einen innovativen Ansatz für ein rationales Proteindesign darstellt. Das Auffinden und Beschreiben dieser Fragmente mit den derzeitigen Domänenklassifikationssystemen ist jedoch keine triviale Angelegenheit. Hier zeigen wir durch die Analyse von Fuzzle, einer von unserer Gruppe aufgebauten Datenbank, neue konservierte Fragmente zwischen Proteinfaltungen, von denen wir bisher annahmen, dass sie evolutionär nicht verwandt sind. Mehrere dieser Fragmente, bei denen es sich meist um die Rossmann-Faltung oder ihre Derivate handelt, wurden genauer untersucht, wobei der Schwerpunkt auf ligandenbindenden Eigenschaften lag, die über die evolutionäre Zeit hinweg erhalten geblieben sind. Weiterhin wurden einige dieser Fragmente zu rekombinant exprimierten Faltungschimären rekombiniert, wobei zwei von ihnen, *Ec*PurT/*Bh*PurN und *Ec*PurT/*Gk*PurN, Segmente aus den preATP-grasp- und Formyltransferase-Faltungen aufwiesen und eine stabile Struktur in Übereinstimmung mit dem computerberechneten Modell aufwiesen. Insgesamt geben diese Ergebnisse einen allgemeinen Einblick in die Entwicklung von Faltungen in Proteindomänen und werfen ein Licht auf deren Evolution. Ferner zeigen sie eine Möglichkeit auf,

den bekannten Proteinsequenzraum zu erweitern, indem entfernt konservierte Faltungsfragmente zu realisierbaren Chimären kombiniert werden. Für die Zukunft planen wir weitere Charakterisierungen der Fuzzle-Fragmente und der evolutionären Zusammenhänge zwischen ihnen. Zudem erwarten wir, dass wir weitere Fragmente finden werden, die die Erzeugung von Faltungschimären mit Funktionalität in ihnen ermöglichen.

Abstract

Domains, the structural, functional, and evolutionary units of proteins, act as building blocks that contribute to the diversification of the protein universe as we know it. It is thought that domains emerged and evolved employing a set of ancestral, sub-domain sized fragments that have been reused in extant protein folds under different structural and/or functional contexts. Moreover, recombination of these elements may result in the generation of proteins with novel structures and functions, offering an innovative approach for rational protein design. Finding and describing these fragments with current domain classification systems, however, is not a straightforward task. Here we show, through the analysis of Fuzzle, a database built by our group, novel conserved fragments between protein folds assumed to be evolutionary unrelated until now. Several of these fragments, involving mostly the Rossmann fold or its derivatives, were examined in more detail, with a focus on ligand-binding features that have remained over evolutionary time. Furthermore, some of these fragments were recombined into fold chimeras that were expressed recombinantly, with two of them, *Ec*PurT/*BhPurN* and *Ec*PurT/*GkPurN*, incorporating segments from the preATP-grasp and formyltransferase folds and adopting a stable structure in accordance with the computationally designed model. Overall, these results give a general perspective of fold development in protein domains, shedding light on their evolution. In addition, they demonstrate a way of expanding the known protein sequence space by combining remotely conserved fold fragments into feasible chimeras. In the future we envision further characterization of Fuzzle fragments and the evolutionary connections between them. Additionally, we expect to find more fragments that will allow the generation of fold chimeras with proper functionality within them.

Introduction

LIFE ON EARTH EXISTS IN MANY DIFFERENT FORMS. In most of the planet's environments, no matter how extreme their physical conditions may be, there are organisms able to survive and thrive in them. This is reflected, for instance, in extremophiles that withstand challenging temperatures, pH values, salt concentrations, among others. Other scenarios can be seen in rainforests housing a comparatively high number of species with respect to their land area. The ability to conquer and tolerate an ecosystem is brought in part by adaptation, i.e., the process by which living beings change over generations to enhance their survival or reproductive success.[1] Over geological time, adaptation to different environments has resulted in a wide array of approaches that living things, especially microorganisms, have employed and honed.

[1] Keller & Lloyd, 1992

Nature as a tinkerer

THIS VARIETY IN LIFE STRATEGIES is also reflected in the broad range of metabolic strategies adopted by different lifeforms as a result of the evolutionary process. Yet all living organisms share the same set of basic biochemical building blocks, functions, and mechanisms. This conservation in chemical components shows how most aspects present in biological systems did not arise out of nothing. Instead, evolutionary innovation has usually built from existing, already available objects, reshaping them into novel, complex traits. As François Jacob described in his essay "Evolution and Tinkering":[2]

[2] Jacob, 1977

> Natural selection has no analogy with any aspect of human behavior. However, if one wanted to play with a comparison, one would have to say that natural selection does not work as an engineer works. It works like a tinkerer—a tinkerer who does not know exactly what he is going to produce but uses

> whatever he finds around him whether it be pieces of string, fragments of wood, or old cardboards; in short it works like a tinkerer who uses everything at his disposal to produce some kind of workable object.

Examples of evolutionary novelty by tinkering abound and are present at all levels of organization, from molecules to organisms, creating and developing uniqueness through random combination of preexisting characters, yielding imperfect products with no foresight and no specific long-term goal in mind. Despite this lack of purpose, they are able to endure and respond to environmental constraints, and to persist and proliferate in living beings. One of the most prominent cases of tinkering throughout the history of life on Earth is the emergence and posterior evolution of proteins.

The protein space is wide and diverse

POLYPEPTIDES ARE THE MAIN WORKHORSES in cellular systems, carrying out the vast majority of tasks encoded genetically. The proteome provides catalysis, signal transduction, ligand binding, and transport, it supplies structure and motion, among a myriad of other processes (Figure 1). Proteins are a

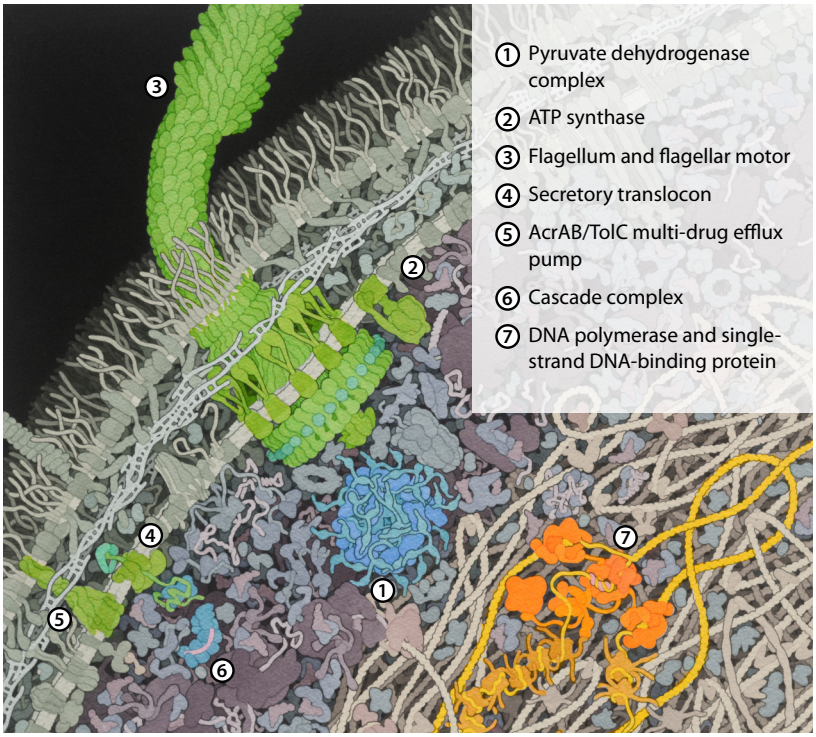


Figure 1: Molecular landscape of an *Escherichia coli* cell. Highlighted are proteins involved in ①② catalysis, ③ structure and motion, ④⑤ transport, ⑥ immunity, and ⑦ information storage and replication. Adapted from Goodsell (2021).

highly versatile and optimized kind of macromolecule, able to adopt a wide range of sequences, structures, and functions within the cell. The process by which this degree of variability was achieved is nevertheless an unanswered question. Polypeptides present in current living beings originated from prebiotic precursors, but there is no consensus about the mechanisms that produced the first amino acid chains. Amino acid synthesis and polymerization reactions can be conducted abiotically, and several hypotheses have been proposed. There is, however, no definite scenario explaining how the first functionally active peptides originated from these precursors. Likewise, little is known about the processes that occurred during early life conditions, allowing for further structural and functional divergence and giving rise to the diverse and sophisticated protein universe we know and see today.

New proteins require new genes

TO UNDERSTAND how the set of extant proteins appeared and expanded, the link between gene and protein needs to be illustrated. As Francis Crick proposed in the *sequence hypothesis*,[3] the three-dimensional structure and therefore the function of a polypeptide is fully encoded in the nucleic acid sequence of the corresponding gene. Therefore, to generate new proteins means to add new protein-coding genes into a genome. This can be achieved by means of two mechanisms: duplication or *de novo* birth. The gene, after being created, and depending on its evolved functionality, can undergo different fates.

[3] Crick, 1958

Gene duplication

Duplication is considered the most common process to generate new genetic material. There are several ways a gene can be duplicated in a genome:[4]

[4] Futuyma, 2018

- *Unequal crossing-over*, where recombination occurs between misaligned homologous chromosomes during meiosis, resulting in two copies of the gene in adjacent chromosomal loci.
- *Replication slippage*: during replication, dissociation of DNA polymerase and reattachment at the incorrect site can cause the propagation of tandem repeats of genetic segments.

- *Retrotransposition*, where mRNA is reverse-transcribed to DNA and then inserted into a random location in the genome.
- *Exon shuffling*, where exons from preexisting genes assemble and form genes with new exon combinations. This is usually mediated by intronic recombination.
- *Whole genome duplication*: if sister chromatids fail to separate properly during cell division an individual can become polyploid, receiving an additional copy of the parental genome.
- *Horizontal gene transfer*, the nonsexual transmission of genetic material between unrelated species, is a common mechanism among prokaryotes. It is important, among others, in antibiotic resistance acquisition and the spreading of virulence factors.

De novo *gene birth*

IN CONTRAST TO DUPLICATION, the emergence of new genes from non-coding DNA is very unlikely but not unheard of. In the last decades studies have been conducted in several model organisms searching for *orphan genes*, i.e. genes with no known ortholog in other species, including closely-related ones. Since functionality and phenotypical effects of this kind of genes are difficult to evaluate, not much is known about the vast majority of them. However, one well-characterized example is the emergence of antifreeze protein genes in Arctic codfishes, which evolved from a non-genic precursor region.[5]

See Zhao et al., 2014 for an example in *Drosophila*, arguably the best-studied case.

[5] Baalsrud et al., 2017

Possible fates for a new gene

No matter how it was conceived, a new gene will most likely be removed by either random drift or selection. However, a small fraction of them will spread over the population, fixating themselves in the species' genome. In the case of duplications, after fixation in the genome, the newly duplicated gene has a number of potential outcomes (Figure 2). For instance, the paralog can diverge by the acquisition of a completely different function (*neofunctionalization*). Alternatively, both the parental gene and its copy can evolve independently, specializing in some of the functions of the ancestral gene (*subfunctionalization*). Subfunctionalization can also be incomplete, giving rise

to genes coding proteins with *moonlighting* functions.[6] The replica can suffer less favorable consequences as well. Over time the duplicate may lose functional or regulatory elements, becoming a non-working *pseudogene*.

[6] Espinosa-Cantú et al., 2015

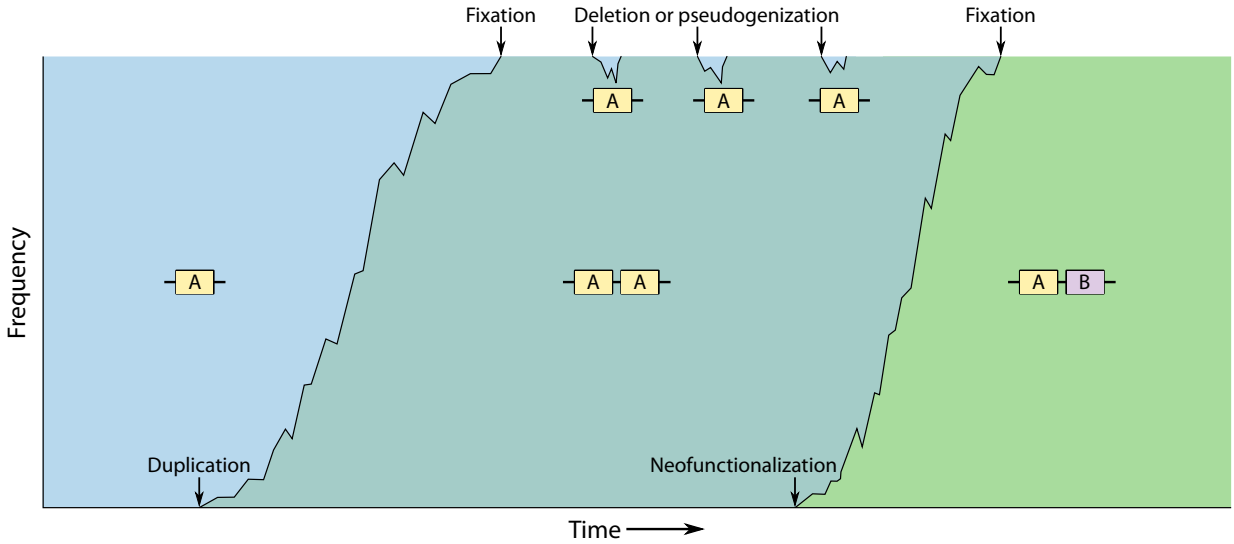


Figure 2: Example evolutionary trajectory of gene A and its neofunctionalized duplicate, B. Adapted from Innan & Kondrashov (2010).

In addition to the previously mentioned whole-gene strategies, proteins have developed a modular approach for evolutionary innovation: the combination and reuse of domains within them. Protein domains have been instrumental in the expansion of the protein universe, enabling an increase in structural and functional complexity from a finite set of basic components.

What is a protein domain?

THE DEFINITION OF "PROTEIN DOMAIN" is rather flexible and depends to a great extent upon the conceptual context where that term is being used. The reasons for this are historical and are in part a result of the ambiguity and subjectiveness about what a domain actually entails and how these criteria were applied in the different domain assignment and classification schemes developed over time. Broadly speaking, domains are regarded as self-contained, conserved, independently structural, functional, and evolutionary units of a protein. In the following sections I will expand on the different aspects that can define a domain and how they are reconciled in different

classification systems.

Domains in protein structure

In 1973, after surveying by visual inspection several solved globular protein structures available at the time, Donald Wetlaufer postulated the existence of compact, independently-folding structural regions contained within a single polypeptide chain.[7] This premise can be considered an extension or generalization of the *domain hypothesis* in immunoglobulins proposed by Edelman in 1969,[8] where immunoglobulins acquired structurally separated regions, labeled *domains*, that can be grouped according to homology (Figure 3). In this particular example, each domain belongs to the constant or variable region of the immunoglobulin molecule. Within each group, domains are expected to be conserved structurally and functionally.

[7] Wetlaufer, 1973

[8] Edelman et al., 1969

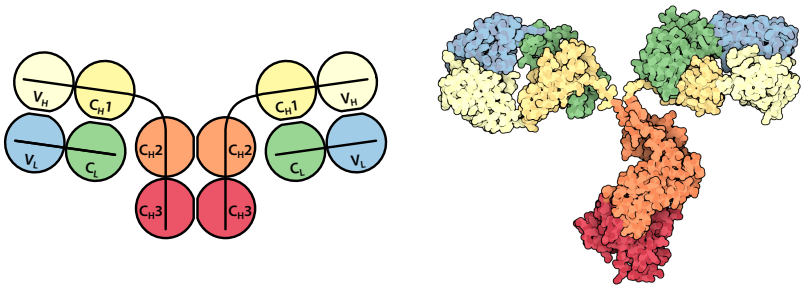


Figure 3: The domain hypothesis in immunoglobulins. Left: Domain arrangement of immunoglobulin G, adapted from Edelman (1973). Each circle corresponds to a single immunoglobulin domain. Right: Three-dimensional structure of immunoglobulin G (PDB ID 1IGT).

Initial definitions for structural domains were subjective, but algorithms for objective, quantitative descriptions of *domain* were developed shortly afterwards, based mostly on C_α-C_α contact maps and solvent accessible surface areas.[9] At that point, some general properties of domains as structural units of proteins were established. They have compact, isolated hydrophobic cores, with no continuation of beta-sheet hydrogen bonding patterns and no secondary structure elements shared with other domains. Short-distance, long-range (i.e. distant in sequence) interactions are maximized, while interactions with non-domain regions are minimized. Overall, these traits imply that a domain folds autonomously and independently from the rest of the protein chain. This aspect led to the proposal that individual domains may have a role in the folding of a protein chain, acting as nucleation spots for the

[9] Janin & Wodak, 1983

process to take place. Independent folding of domains in multidomain proteins might speed up the overall folding process by avoiding misfolding interactions with their neighbors as the whole protein chain adopts its native structure.[10]

[10] Han et al., 2007

Domains in protein evolution

With the advent of protein sequencing in the 1950s, early efforts were directed at grouping and classifying protein sequences based on their presumed common ancestry or homology. The underlying assumption was that functionally essential residues would be under strong selective pressure and thus remain identical across different species.[11] An early example of such an endeavor is the Atlas of Protein Sequence and Structure,[12] the first computerized database of protein sequences.

[11] Strasser, 2010

[12] Dayhoff, 1965

In the 1970s, with collections of sequences already established, algorithms for global and local comparison between biological sequences were developed.[13] Structural comparisons demonstrated that structurally and functionally similar domains can have very low levels of sequence identity, within a range known as the *twilight zone* of sequence homology, leading to the development of sensitive methods to detect homology that did not rely uniquely on sequence identity. Furthermore, the exponential growth in the amount of deposited biological sequences asked for efficient ways to curate, annotate, and classify them. These may involve pairwise sequence similarity searches such as FASTA and BLAST[14], but later on multiple sequence alignment (MSA) based approaches using position-specific scoring matrices (PSSMs) or hidden Markov models (HMMs) could identify large amounts of homologous sequences without a substantial increase in computational time. These approaches were initially applied on whole protein sequences, but once domains started to be defined based on their structure, the same principles began to be applied to them.

[13] Needleman & Wunsch, 1970; Smith & Waterman, 1981

[14] Pearson & Lipman, 1988; Altschul et al., 1990

Over the course of evolution, domains can change their sequences, diverging and generating a *family*, i.e., a set of different protein domains descended from a single, common ancestor. It is assumed that proteins in a family are closely related, sharing significant sequence similarity and having conserved

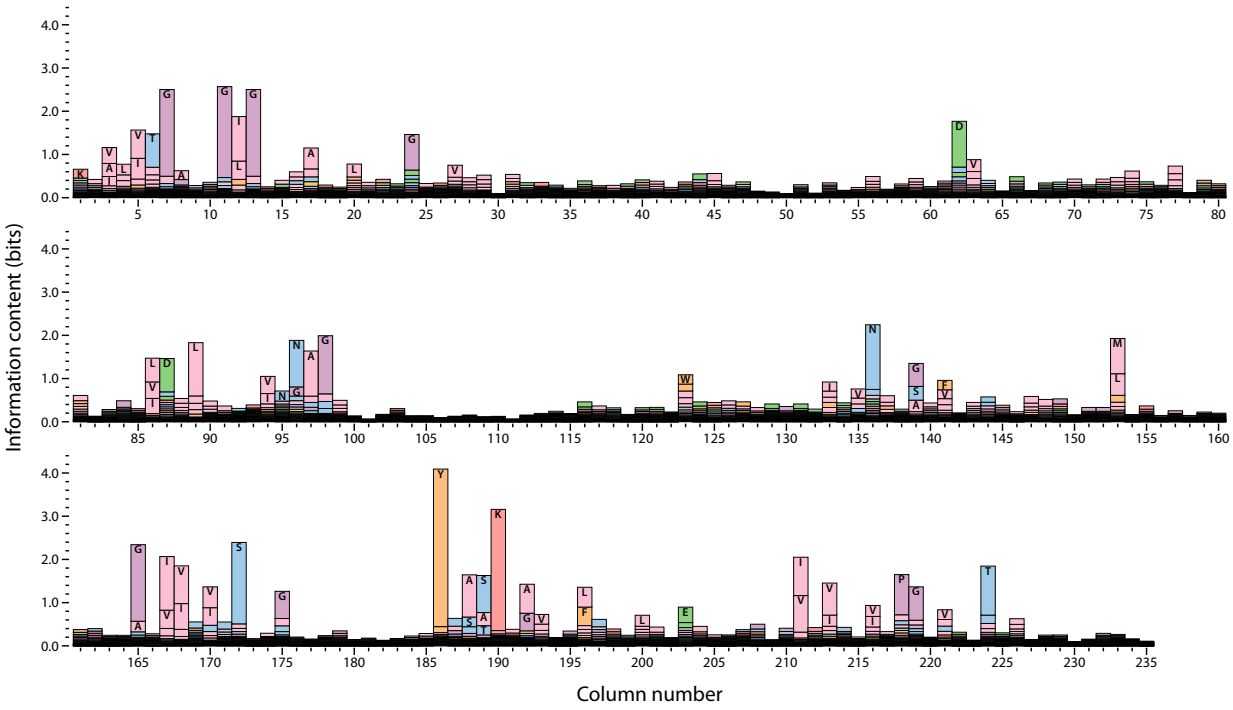


Figure 4: Example sequence logo representing the short-chain dehydrogenase/reductase family (SDR, Pfam ID PF00106). Conserved regions (i.e. with high information content) are distributed throughout the family's MSA.

sequence blocks or motifs derived from functional constraints (see Figure 4 for an example). However, proteins can be considered homologous even with no detectable sequence similarity, based on shared functional or structural features. To address this problem, Margaret Dayhoff introduced in 1974 the concept of *superfamily*: a group of distant but nevertheless evolutionarily related families.[15]

[15] Dayhoff, 1974

At the sequence level, protein domains can be defined as recurring, homologous sequence portions, present in different genetic contexts, that have remained intact through evolution.[16] In other words, domains can be considered discrete evolutionary units of proteins. Consequently, genomic analysis in different organisms have showed that a large fraction of genome-encoded proteins (up to 80% in eukaryotic species) contains more than one domain,[17] and that most domain families have sustained extensive duplication, recombination and shuffling, generating as a result conserved multidomain architectures. Despite the randomness and high frequency of recombination events, only a small fraction of all possible domain arrangements can be detected in proteins, suggesting that they have been subject to strong evolutionary pressure (Figure 5).

[16] Dawson et al., 2017

[17] Apic et al., 2001; Liu & Rost, 2004

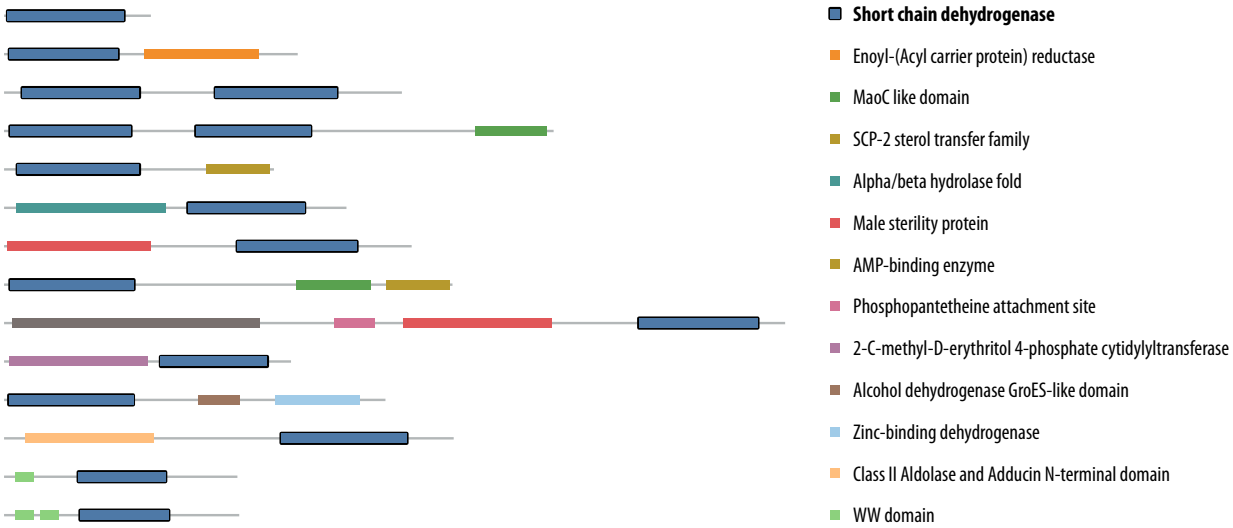


Figure 5: Most frequent domain combinations within the SDR family depicted in Figure 4.

[18] Vogel et al., 2004

Furthermore, some of these combinations, termed *supra-domains*, recur in different contexts, with different partner domains.[18] Supra-domains have usually a single, specific N-to-C-terminal order believed to be consequence of the initial fixation event, i.e. a result of historical rather than spatial or functional constraints, as the latter two should not depend on the domain order within a supra-domain.

Domains in protein function

Single domains are considered functional modules of proteins and are generally associated to a specific activity conserved, at least partially, throughout the whole domain family. This modularity allows for the reuse of domains in different circumstances, where the combination of domains will determine the functions of the protein as a whole. The combination can represent, for example, contiguous steps in a metabolic pathway, since having such domains in close proximity may aid in the transfer of intermediates that could be unstable or in low concentration during these cascade reactions. An example is shown in Figure 6A. Other advantages of pathway-related multidomain proteins are the possibility of allosteric regulation and the guarantee of a fixed stoichiometric ratio of enzymatic domains involved, in a manner analogous to prokaryotic polycystronic mRNAs. This segregation of activities allows for different domain combinations and their reuse, as similar domains can be found in proteins with different overall functions.

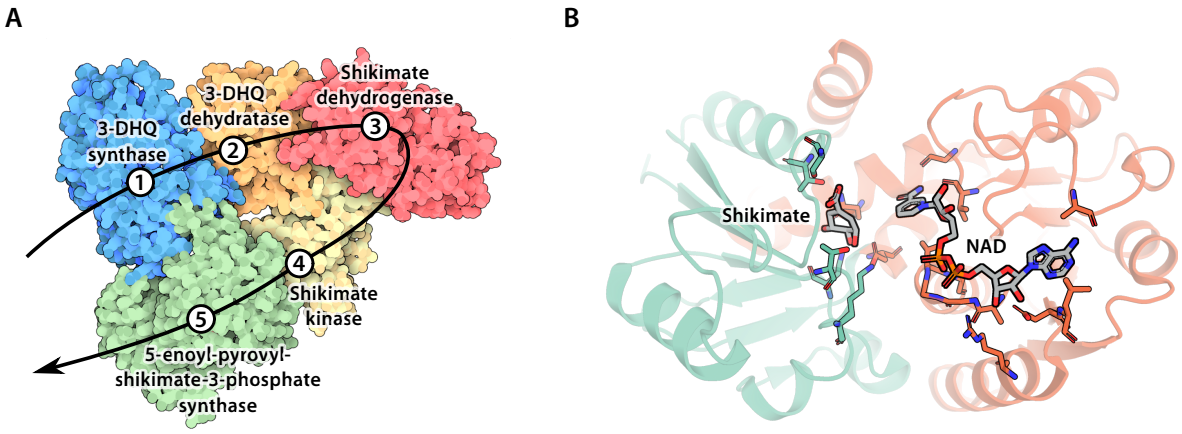


Figure 6: Function in multidomain proteins. A: The multidomain AROM protein (PDB ID 6HQV) catalyzes five consecutive steps in the shikimate pathway. B: The active site in quinate/shikimate dehydrogenase (PDB ID 3JYQ) consists of residues from both of its domains.

However, it is also possible that the protein function lies at the interface between domains, with two or more domains contributing altogether to the global activity of the protein. For example, the active site of an enzyme can be provided by residues from different domains, allocating substrates in the interdomain cleft (Figure 6B).

A domain can perform the same function in different molecular contexts, combined with different kinds of partner domains. Nevertheless, in some particular cases the domain can diverge and obtain a new or modified function, different from the ones commonly attributed to the domain's superfamily. Borrowing from linguistics, these two cases have been described as syntactic and semantic shifts, respectively.[19]

Assigning function to a domain is usually based on homology and done by sequence and structural comparisons to existing domains, looking at conserved residues or regions known to contribute to activity. In some cases, a newly discovered domain cannot be described in terms of homologous domains and is therefore classified as a *domain of unknown function* or DUF.

[19] Vogel et al., 2004

Classifying protein domains

DOMAINS HAVE BEEN COMMONLY ORGANIZED according to their sequence, their structure, or a combination thereof. In the case of sequence-based databases, features defining each family are usually encoded in patterns or profiles, represented as PSSMs or HMMs derived from seed MSAs. Family assign-

ment of newly released domain sequences is done routinely in an automated fashion. Examples of these kind of databases are PROSITE and Pfam.[20]

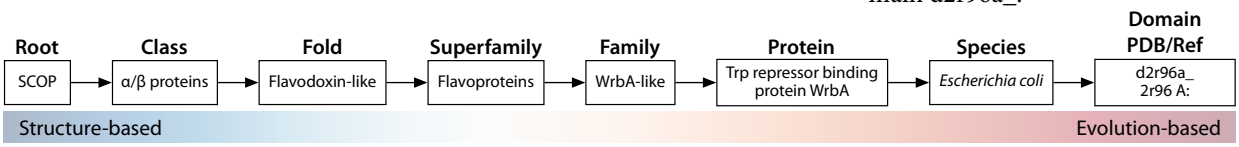
[20] Sigrist et al., 2012; Mistry et al., 2020

Very remote homolog domains might stay undetected in sequence-based comparisons, due to their low sequence identity. Since structure is more conserved than sequence, structural comparisons might be able to recognize these connections. Moreover, structural analysis of domain families can reveal non-conserved embellishments contributing to variability. Like sequence-based systems, most structural classifications group domains into families, but others, like DALI,[21] adopt an all-against-all approach. Depending on the database, the assignment can be manual or automated. In hierarchical systems, such as CATH and SCOP,[22] higher levels are exclusively based on structure and do not imply homology. Lower levels, on the other hand, integrate evolutionary information incrementally (Figure 7).

[21] Holm, 2020

[22] Sillitoe et al., 2020; Fox et al., 2013

Figure 7: SCOP hierarchy of domain d2r96a_.



Subdomain-sized fragments in protein evolution

AS STATED BEFORE, THE PROTEIN UNIVERSE has used and reused domains taken from a limited repertoire of families and folds (10^4 to 10^5 and 10^3 to 10^4, respectively). Nevertheless, the mechanism by which this domain pool originated and diversified remains elusive. One hypothesis proposes that domains derive from an ancestral set of subdomain-sized peptides that acted as cofactors for self-replicating ribozymes on the early Earth (the RNA world). This collection of peptides gained complexity in a gradual manner, ending up as the full-sized domains seen in actual proteins. Drawing from the above-mentioned tactics used for whole-gene generation, three proposed processes could reach this goal: repetition, accretion, and recombination.

Repetition

Repetition is, without doubt, the most well studied strategy in domain recapitulation. Some folds are composed of distinguishable repeats, which provide hints about their origin and subsequent evolution. It is assumed that repeat-containing folds emerged through the tandem duplication and fusion of supersecondary structural motifs. Most examples in literature explore solenoid folds, with a varying number of repeating helical motifs forming a superhelical structure. To date, several idealized scaffolds of them have been built based on single subunits (Figure 8A and B), including armadillo, HEAT, ankyrin (all of them α-solenoid), and leucine-rich (αβ-solenoid) repeats.[23] The case of the α-helical tetratricopeptide (TPR) repeat is of particular interest. In addition to TPR fold domains based on consensus repeats,[24] a TPR-like protein was built from a non-repetitive helical hairpin, found through a profile-based approach, in the ribosomal protein RPS20.[25] While the protein with the wild-type motifs was soluble but unstructured, it needed only a small number of mutations to adopt a superhelical fold (Figure 8C).

Toroidal folds have also been subject of fold reconstruction to shed light on their evolutionary history. The internal symmetry present in their structures allowed the discovery of subdomain-sized repeats, with the triose phosphate isomerase (TIM) barrel, β-propeller, outer membrane protein (OMP) bar-

[23] Reichen et al., 2014; Urvoas et al., 2010; Kohl et al., 2003; Stumpp et al., 2003

[24] Main et al., 2003

[25] Zhu et al., 2016

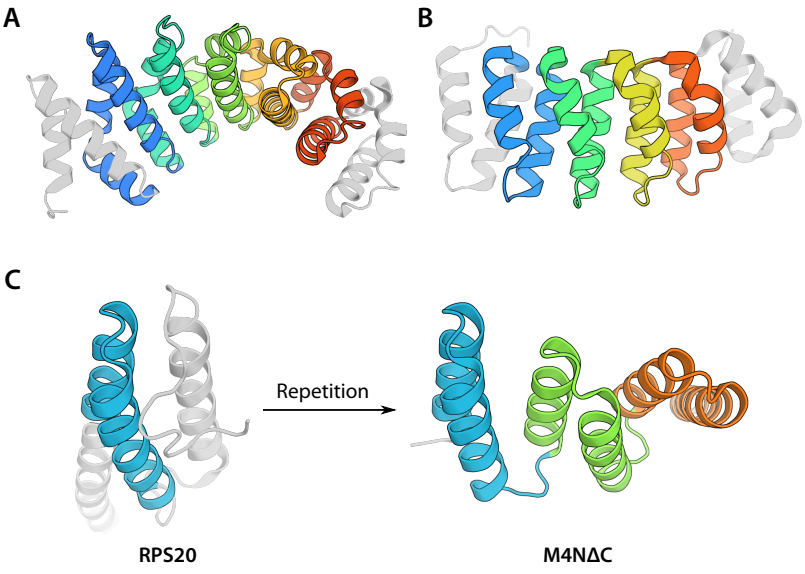


Figure 8: Repetition in solenoid folds A: Designed armadillo repeat protein (PDB ID 4PLQ). B: Artificial HEAT-like repeat protein (PDB ID 3LTJ). C: Helical hairpin of *Thermus thermophilus* ribosomal protein RPS20 (PDB ID 4GKJ) and its amplification into a TPR-like protein (PDB ID 5ZFR). Each repeat is highlighted with a different color. Non-repeat regions are shown in gray.

rels, and β-trefoil folds as remarkable examples.

TIM barrels (Figure 9A) are made of 8 βα units, which suggested that said fold arose through a series of duplications: first as $(\beta\alpha)_2$ quarters, then $(\beta\alpha)_4$ halves, and finally the present $(\beta\alpha)_8$ topology. Indeed, profile HMM-based studies have found remote homology between these elements.[26] Experimental evidence has been obtained in the form of half barrels forming either homo- or heterodimers, showing that a $(\beta\alpha)_4$ duplication approach is feasible.[27] Nevertheless, there are some cases where truncated $(\beta\alpha)_6$ barrels have been found, suggesting, additionally, a quarter-based incremental or decremental strategy.[28]

β-propellers (Figure 9B) have β-meander repeats (or blades) arranged around a central axis. The number of blades range from four to ten, depending on the protein family and function.[29] This fold has been reconstructed from NHL blades, forming a family of fully symmetrical six-bladed proteins denominated Pizza.[30] Moreover, an ancestrally reconstructed blade starting from a tachylectin repeat gave a functional, five-bladed, lectin-like oligomer.[31]

The β-trefoil (Figure 9C), a threefold pseudosymmetric globular fold, can be recovered from single β_3-loop-β units.

[26] Söding et al., 2006

[27] Höcker et al., 2001

[28] Setiyaputra et al., 2011

[29] Chen et al., 2011

[30] Voet et al., 2014

[31] Smock et al., 2016

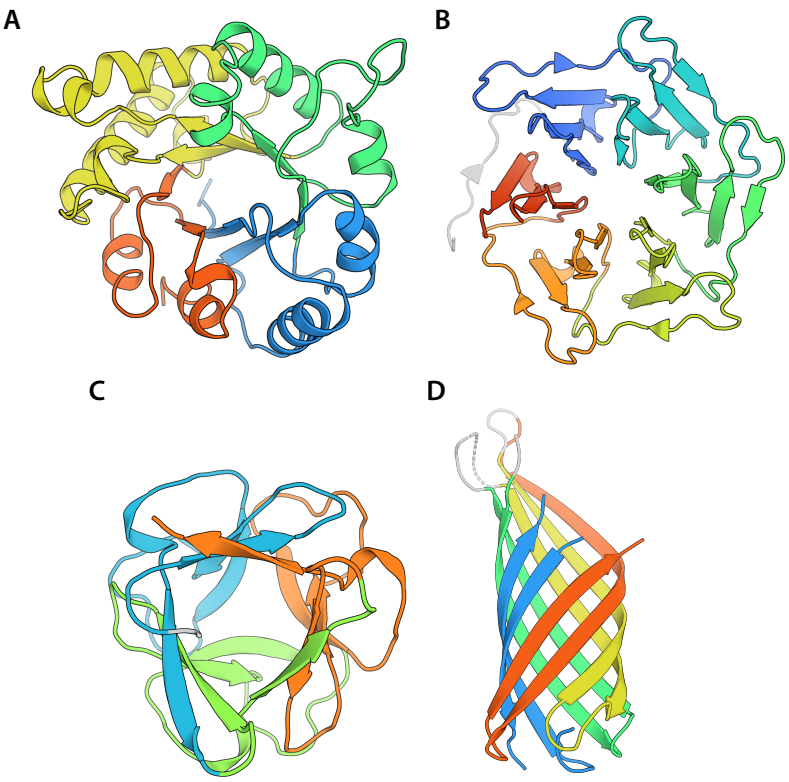


Figure 9: Repetition in toroidal folds. A: *Gallus gallus* triose phosphate isomerase (PDB ID 8TIM). B: Designed β-propeller protein Pizza6 (PDB ID 3WW9). C: Designed β-trefoil Mitsuba-1 (PDB ID 5XG5). D: *Escherichia coli* OmpA membrane domain (PDB ID 1QJP). Each repeat is highlighted with a different color. Non-repeat regions are shown in gray.

Furthermore, trefoil constructs with perfect sequence symmetry between its subdomains have been developed, with one case, called Mitsuba-1, retaining its parent's lectin and cancer cell binding activity.[32]

[32] Terada et al., 2017

OMP barrels (Figure 9D), with an all-β topology, resulted from the amplification of a β-hairpin motif culminating in barrels with different number of strands.[33] An analysis of the sequence and structure of over 50 000 OMP barrel homologs revealed evolutionary pathways that are different from the expected linear increase of repeats, including loop-to-hairpin transitions and large, non-duplicative rearrangements.[34]

[33] Remmert et al., 2010

[34] Franklin et al., 2018

Accretion and recombination

Accretion and recombination of domain fragments have not been studied as deeply as duplication. However, even in the absence of direct experimental evidence, some of the previously mentioned evolutionary pathways of OMP barrels can be considered a case of accretion.

Not to be confused with *domain accretion*, a similar concept applied to multidomain architecture formation as opposed to single domains.

Empirical proofs of domain fragment extension by accretion and recombination have been scarce, but there are a number of works delving into that sort of events. Some of the studies have involved the recombination with other fragments from random extant domains. In these, a fragment of an extant domain has been recombined randomly with DNA fragments and selected by phage display.

One of these efforts recombined a segment from the cold shock protein CspA of *Escherichia coli* with different fragments of natural proteins at random.[35] One of the resulting combinatorial proteins, 1b11, incorporated a segment from the S1 domain of the 30S ribosomal subunit from *E. coli*. Its structure could be solved, showing that it forms a swapped, six-stranded β-barrel (Figure 10A).[36] It is worth mentioning that the CspA fragment retains its original three-stranded conformation (Figure 10B). Another study employed as bait a DNA segment from a human immunoglobulin Vκ domain encoding a β-sheet with strands adjacent in structure but not in sequence.[37] Surprisingly, after recombination with random human cDNA fragments, the most stable chimera had integrated an antisense region of the YIPF3 gene. These unexpected products exhibit what could have happened during very early protein evolu-

[35] Riechmann & Winter, 2000

[36] Bono et al., 2005

[37] Fischer et al., 2004

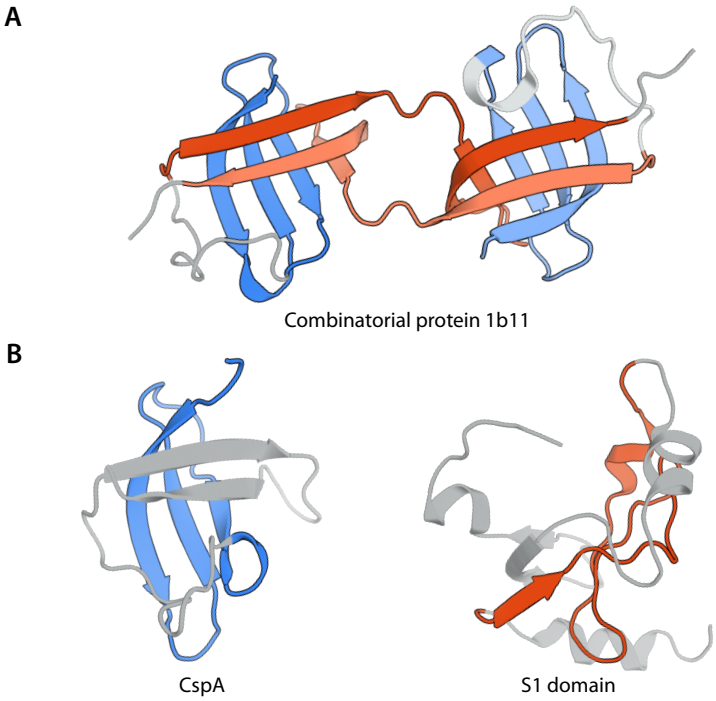


Figure 10: Structure of (A) 1b11 forming a swapped dimer (PDB ID 2BH8) and (B) its parental proteins, CspA (PDB ID 1MJC) and S1 (PDB ID 5XQ5). CspA and S1 fragments are colored blue and red, respectively.

tion, where indiscriminate recombination of small fragments by pure chance seemed to be crucial for the genesis of domain precursors.

Conserved domain fragments between different folds

IN A MANNER ANALOGOUS TO WHAT IS SEEN in multidomain proteins, subdomain-sized fragments can be shared between domains adopting different, seemingly unrelated folds. Alva et al. conducted a survey of modern proteins and found 40 of such fragments, with lengths ranging from 9 to 38 residues.[38] Some of them are present in what are considered the most ancient folds and are involved in nucleotide or metal binding, which led to the suggestion that this set of ancestral fragments represent the remains of a primordial yet catalytically active RNA-peptide world.

[38] Alva et al., 2015

In addition to the fragments found in the previous study, larger domain segments might have played a role in the posterior fold diversification. Our group previously identified one of such fragments between $(\beta\alpha)_4$ halves of TIM barrel and $(\beta\alpha)_5$ flavodoxin-like (Fld-like) domains (Figure 11).[39] Initially this fragment, encompassing a $(\beta\alpha)_3\beta$ motif, was found solely on grounds of structural similarity between HisF (imidazole glyc-

[39] Bharat et al., 2008

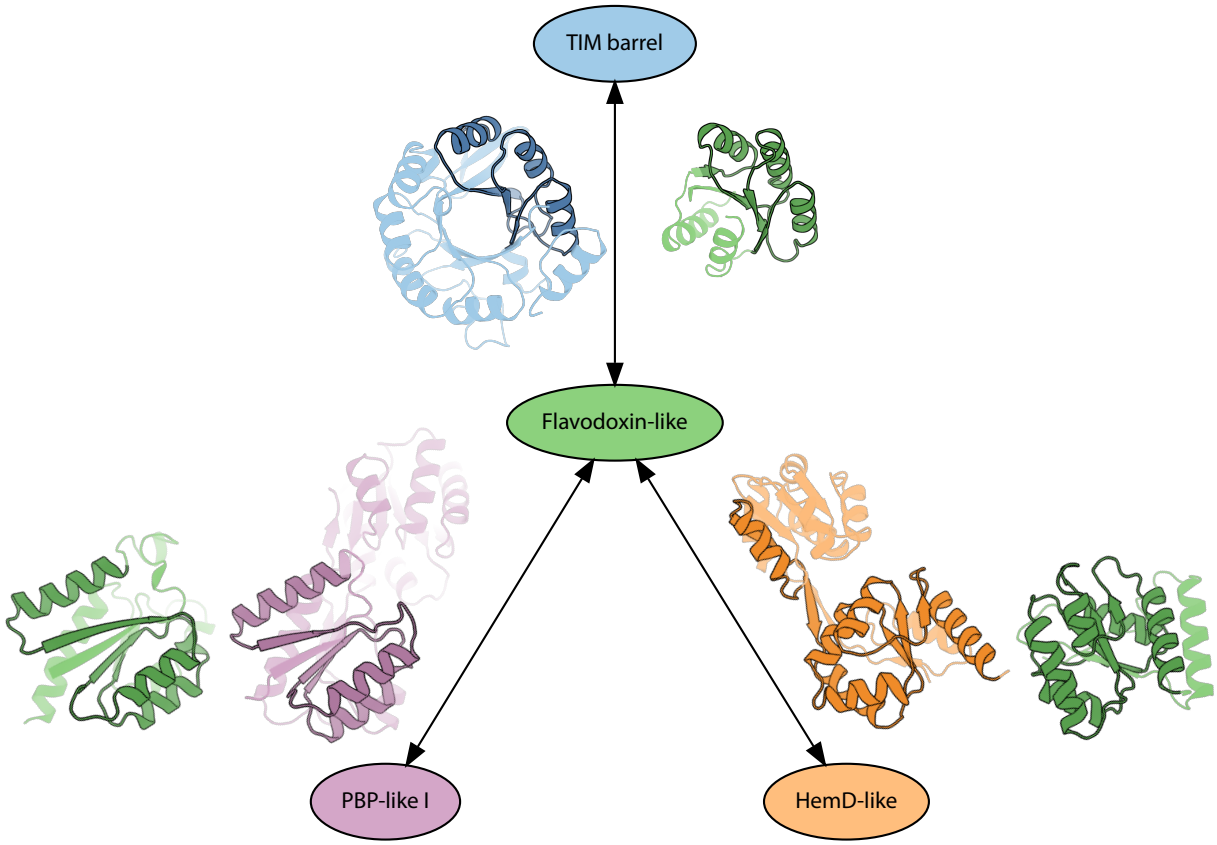


Figure 11: Conserved fragments between different folds already found by our group.

erol phosphate synthase) and the chemotaxis protein CheY. However, in a follow-up work, sequence-based, profile HMM comparisons confirmed these fragments as remotely homologous and not a result of convergent evolution.[40] Further surveys of the protein sequence space identified an intermediate fragment from a then uncharacterized protein named N-TM0182 (N-terminal, B_{12} binding domain of *Thermotoga maritima* TM0182). This segment was expressed and its structure solved by X-ray crystallography. In it, N-TM0182 forms a strand-swapped dimer in the solved structure, but an $(\alpha\beta)_2$ fragment showed the highest structural similarity with both TIM barrels and Fld-like domains.

[40] Farías-Rico et al., 2014

Encouraged by this outcome, other fold combinations were explored and as a result additional conserved fragments were identified, namely the $(\beta\alpha)_4$ one between the PBP-like I and Fld-like domains[41] and another between HemD-like and Fld-like folds.[42] Both PBP-like I and HemD-like folds present common properties, such as a bilobular architecture connected by a hinge region and interlobe ligand-binding interfaces, but the

[41] Farías Rico, 2014

[42] Toledo-Patiño et al., 2019

linker region between both lobes and their overall connectivity is different (Figure 11).

The HemD-like example is of relevance since it has been proposed, based on sequence and structural information, that the HemD-like fold arose from a flavodoxin-like precursor. According to this hypothesis, the Fld-like ancestral fragment sustained an insert-mediated segment swap and swiveling followed by duplication and fusion, achieving the two-lobed HemD-like architecture. Remarkably, these events can be undone experimentally in a stepwise manner, yielding a Fld-like structure starting from a HemD-like half.

All the previously mentioned instances of fragments are conserved in sequence and structure, meaning that, in a manner analogous to full-sized domains, recombination using regions from domain fragments as crossover points should be possible. This concept would enable the rational design of proteins representing domain chimeras, even with fragments originating from different folds, as long as an evolutionary link is established.

Protein design

The discipline of protein design refers to the rational design of novel sequences that produce an *a priori* specified structure and/or activity. This field was set in motion in the late 1970s by Bernd Gutte, who designed in a rational fashion protein binders for the anticodon of yeast $tRNA^{Phe}$ (the GAA trinucleotide) and the insecticide DDT.[43] While successful in assessing their function, the structural characterization of the designs was incomplete. Nevertheless, these early results showed the feasibility of artificial enzyme design from first principles.

[43] Gutte et al., 1979; Moser et al., 1983

The first fully successful designs involved helical bundles. There were two concurrent strategies pursuing this goal. The first approach, from the deGrado group, was a modular one using designed amphiphilic identical helices.[44] This four-helix bundle was, at the time, the best characterized artificial protein (Figure 12A). In contrast, the Richardson group sought a "protein-like" sequence as "natural" as possible. Their design was non-repetitive, with no sequence homology to known proteins.[45]

[44] Ho & DeGrado, 1987

[45] Hecht et al., 1990

Eventually, the introduction of computer-assisted design

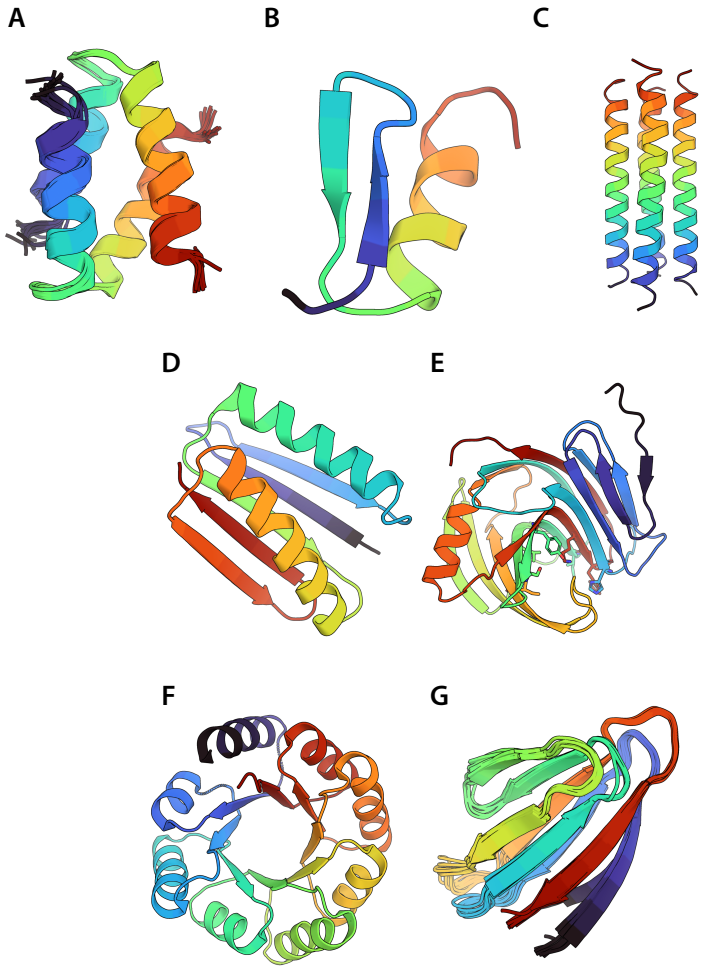


Figure 12: Hallmarks of protein design. A: NMR ensemble of the four-helix bundle α_2D (PDB ID 1QP6). B: FSD-1 (PDB ID 1FSD). C: Right-handed coiled-coil RH4 (PDB ID 1RH4). D: Top7 (PDB ID 1QYS). E: Retroaldolase RA61 variant M48K (PDB ID 3B5L). F: Four-fold symmetric TIM barrel (PDB ID 5BVL). G: NMR ensemble of designed double-stranded β-helix BH_10 (PDB ID 6E5C).

protocols allowed for the automated generation of precise protein conformations, as well as improved fine-tuning of the intended designs. In 1997, the first successful fully automated *de novo* designed protein was published by the group of Stephen Mayo.[46] The target and experimental structures resembled a ββα motif based on a zinc finger backbone (Figure 12B). Structural motifs rarely seen in naturally occurring proteins, such as right-handed coiled-coils (Figure 12C), have also been designed.[47] In 2003, the group of David Baker solved the structure of the first *de novo* protein design with a fold not found in natural proteins (Figure 12D).[48] The year 2008 saw a series of synthetic enzymes for the retro-aldol reaction, catalyzing a carbon-carbon bond breakage (Figure 12E).[49] Finally, in the past few years several folds have been designed *de novo*, such as TIM barrels and all-β proteins with non-local sheet geometry (Figure 12F and G).[50]

[46] Dahiyat & Mayo, 1997

[47] Harbury et al., 1998

[48] Kuhlman et al., 2003

[49] Jiang et al., 2008

[50] Huang et al., 2015; Marcos et al., 2018

Despite all these advances in the area, the design of folds from scratch, even naturally occurring ones, remains a challenge. As an example, it was not until 2015, despite decades of efforts,[51] that the first *de novo* idealized four-fold symmetric TIM barrel design was successfully achieved and characterized.[52] Likewise, the computational design of functional biocatalysts has slowed down due to the low efficiency that is generally achieved with this strategy.[53] Furthermore, designs usually have to be complemented by additional experimental procedures such as directed evolution to increase solubility and/or catalytic efficiency. As a consequence the field has focused mostly on the design of folds and protein binders in the last years.

[51] Goraj et al., 1990; Figueroa et al., 2016

[52] Huang et al., 2015

[53] Vaissier Welborn & Head-Gordon, 2019

One of the underlying problems in protein design is the fact that, for a given conformation, the search space for the most likely sequence is immense and grows exponentially with sequence length. It has been demonstrated that computational protein design is an NP-hard optimization problem, meaning that no known algorithm can solve it in polynomial time.[54] This issue can be solved with different approaches and methods, each with their own advantages and trade-offs. These can be classified broadly into two categories: provable and heuristic. Provable algorithms give an exact solution (in this case, the sequence of the global minimum energy conformation or GMEC) but are computationally expensive. Heuristic algorithms, on the other hand, can be considerably faster but a GMEC solution is not mathematically guaranteed given the stochastic strategies used to sample conformational space.

A protein with m residues, 20 possible amino acids per residue, and a mean of n rotamers per amino acid has $(20 \times n)^m$ potential conformations.

[54] Pierce & Winfree, 2002

An alternative way to tackle this concern and simplify the search space for sequences that are suitable for design is to recombine natural backbone fragments and then optimize the sequence to get a well-behaved protein. With this combinatorial strategy the complexity of the search process can be reduced by using protein fragments that have already been showed to fold properly, while still being able to sample non-natural, untested structures.[55] The problem lies, then, in the choice of suitable fragments that would allow us to follow this strategy.

[55] Khersonsky & Fleishman, 2016

Subdomain-sized fragments as building blocks for protein design and engineering

Since conserved domain fragments are structurally similar, recombination on these regions should keep the overall conformation present in the parental fragments. Directed evolution protocols that mimic non-homologous recombination have been developed in the previous decades. For instance, ITCHY (incremental truncation for the creation of hybrid enzymes)[56] is a method that allows for the random recombination of two genes. Its advantage lies in the fact that, unlike other recombination methods, no sequence homology is required at the crossover region. ITCHY has been used to generate interspecies hybrids that retain the parental family activity. More recently, ITCHY was performed on a genome-wide scale in *Escherichia coli*, resulting in a number of functional ORF hybrids that could rescue auxotrophic *E. coli* strains.[57] Unfortunately, only one of those proteins, the SgbE147/MioC82 hybrid, could be expressed and purified in soluble form and, even though biophysical characterization showed the presence of stable secondary structural elements, further structural and functional studies were not carried out.

[56] Ostermeier, Shim, et al., 1999

Other techniques such as DNA shuffling require genes with at least 70% identity.

[57] Rawcliffe, 2019

As shown above, domain fragments can be merged in a manner akin to domain recombination and generate stably-folded chimeric proteins. On a functional level, and as seen with multidomain proteins, fragments or regions adjacent to them could provide ligand-binding pockets, catalytic residues, allosteric sites, and the like. To put it another way, a chimera made from different domain fragments would integrate functional properties from the parental domains.

Our group has explored this approach to design and evaluate chimeras from different folds. Expanding on the fragment shared between TIM barrel and Fld-like folds one chimera, CheYHisF, was built by recombining CheY with the C-terminal half of HisF.[58] The first version of CheYHisF, while already soluble, had a ninth strand consisting of the histdine tag residues. Removal of the tag sequence coupled with design optimization with Rosetta resulted in a soluble, eight-stranded version of the chimera (Figure 13).[59] Additionally, the phosphate binding site provided by the HisF fragment is conserved in CheYHisF, where it is interacting with a sulfate ion in the

[58] Bharat et al., 2008

[59] Eisenbeis et al., 2012

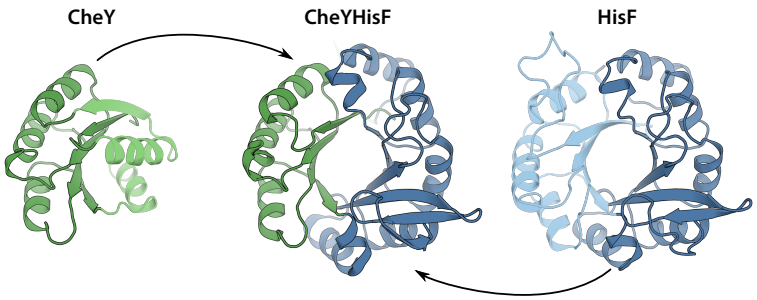


Figure 13: Structure of the eight-stranded variant of CheYHisF (PDB ID 2LLE) and its parental domains (CheY: PDB ID 1TMY; HisF: PDB ID 1THF).

crystal structure.

Building upon this result, additional chimeras have been designed and tested, taking advantage of the conserved fold fragments identified earlier on. A leucine-binding protein domain (LBP) was fused to CheY, where one lobe of the PBP-like I (LBP) was replaced by a CheY fragment, resulting in a chimera adopting a PBP-like structure.[60] A structure of the LBP/CheY chimera was later solved, having the canonical PBP-like topology.[61] The HemD-like/Fld-like fragment was also subject to chimeragenesis studies. A fragment of the cobalamin-binding domain of methylmalonyl-CoA mutase (MCM, a Fld-like enzyme) replaced the N- and C- terminal lobe of the HemD-like uroporphyrinogen III synthase,[62] with the aim of transferring and maintaining the cobalamin-binding capabilities of MCM into a HemD-like scaffold. The resulting chimera was soluble and was able to bind cobalamin based on spectroscopic evidence. However, most of the solved crystallographic structures were of the apoprotein alone, lacking any ligands. Only one structure in complex with the derivative 5,6-dimethylbenzimidazole was obtained.

[60] Farías Rico, 2014

[61] PDB ID 4QWV

[62] Toledo Patiño, 2019

On a related note, computational strategies that mimic the recombination process have been developed, such as SCHEMA and SEWING. The SCHEMA algorithm identifies protein fragments that can be recombined without disrupting the structure of the hybrid protein.[63] It has been used mostly in recombination studies to increase thermostability or functional properties within single protein families. SEWING, on the other hand, is a method that builds *de novo* structural models from pieces of naturally occurring domains.[64] It extracts supersecondary structure motifs and combines them in a continuous or discontinuous way. Since there is no specific target fold, multiple conformations can be freely explored.

[63] Voigt et al., 2002

[64] Jacobs et al., 2016

The combinatorial approach to protein design, albeit promising, is still in its infancy. Domain fragments have been studied mostly in the context of their evolution, focusing primarily on their application to fold recapitulation. However, as another application, recombination of fragments may include, in an additive fashion, functional properties from each individual domain segment. In a best case scenario, this mix-and-match approach would contribute to the expansion of the protein sequence and functional space through the generation of bespoke fold chimeras with novel activities.

Fuzzle, a database of conserved domain fragments

Previous studies performed in our group have already demonstrated the usefulness of subdomain-sized motifs in both protein design and fold reconstruction scenarios. Initially, these relationships were established based solely on structural similarities, as shown in the TIM barrel and Fld-like examples.[65] Later on, sequence similarities could be detected by the adoption of high-sensitivity profile HMM-based algorithms. As a consequence, evolutionary links between Fld-like and other folds were found.[66] In light of these promising results, the next logical step was to extend this analysis over a larger set of domains, families, superfamilies, and folds. Said set should be as comprehensive and illustrative of the diversity of folds as possible, while minimizing redundancy for performance reasons.

[65] Shanmugaratnam et al., 2012

[66] Farías Rico, 2014; Toledo-Patiño et al., 2019

The Fuzzle (*Fold Puzzle*) database[67] is the culmination of this approach. It is the result of an all-against-all pairwise search over a representative set of SCOP domains encompassing the vast majority of protein families present in their classification. Detailed examination of Fuzzle should result in novel insights for our fragment-based view of the protein domain universe, showing how the reuse of these fragments was key for the structural and functional diversification of proteins.

[67] Ferruz et al., 2020

Problem definition

THERE IS A GROWING AMOUNT of theoretical and experimental evidence for the existence of remotely homologous domain

fragments, along with their potential for their fusion into stable, functional domain chimeras. Despite the extent of these studies, and given the different criteria under which domains can be classified, conserved, subdomain-sized segments between folds normally considered non-homologous cannot be easily found. Higher SCOP and CATH hierarchy levels rely uniquely on structural features, disregarding any evolutionary relationship that domains within structural classes may have. That is, since there is no hierarchy of fragments, different folds with different topologies can share fragments, but this evolutionary link will not be reflected in the database. It is therefore necessary to devise a system that evaluates exhaustively these elusive connections between potentially related protein folds.

Aim and goals

THIS DISSERTATION covers two interrelated projects, each trying to answer a specific question from a protein evolution or design point of view. First, from the evolutionary side, I intend to shed light on the diversification process that proteins and protein domains have sustained over the course of their evolution. To achieve this I had to describe conserved fragments representing ancestral states that were present during early stages of life on Earth as well as recapitulate the events that could have happened back then, resulting in evolutionary innovation. Additionally, as a side effect of this examination, evolutionary relationships between remotely homologous folds would be unraveled and defined for the first time, extending our understanding of the nature of the known protein space. Careful descriptions of said connections may help to chronicle the similarities that have remained in related folds as well as their differences acquired over time.

Second, I propose to apply these strategies in a protein design context. This involved the search for fragments suitable for functional design of hybrid fold chimeras. Once found, the conserved regions of said fragments would guide the recombination of partial sections of the parental domains and computational models of the chimeric structures were generated. After evaluating them, I tested the designed chimeras *in vitro* for proper folding and, if possible, for binding and function.

Both lines of work use the analysis of a preliminary version

of the Fuzzle database as a starting point, using the information available therein to survey the extant protein universe and find evolutionarily relevant and potentially engineerable conserved domain segments.

Materials & Methods

Materials

Bacterial strains

- *E. coli* DH5α Competent Cells (Invitrogen)
- *E. coli* TOP10 Competent Cells (Invitrogen)
- *E. coli* BL21 (DE3) Competent Cells (Novagen)
- T7 Express Competent *E. coli* (NEB)

Cloning vectors

- pET21a(+) DNA (Novagen)

Enzymes

- NdeI restriction endonuclease (NEB)
- XhoI restriction endonuclease (NEB)
- T4 DNA Ligase (NEB)
- Taq DNA Polymerase (Thermo Scientific)

DNA purification kits

- DNA Clean & Concentrator (Zymo Research)
- NucleoSpin Plasmid purification kit (MACHEREY-NAGEL)

DNA and protein standards

- 1 kb DNA Ladder (NEB)
- PageRuler Protein Ladder (Thermo Scientific)
- Gel Filtration LMW Calibration Kit (GE Healthcare)

Culture media

- LB medium: 1% w/v peptone, 0.5% w/v yeast extract, 0.5%

w/v NaCl.
- LB-Amp medium: LB medium enriched with 100 µg/mL ampicillin.

Buffers and solutions

- CutSmart Buffer (10×) (NEB)
- 10× Taq Buffer (Thermo Scientific)
- Gel Loading Dye, Purple (6×) (NEB)
- TAE buffer: 40 mM Tris, 20 mM acetic acid, 1 mM ethylenediaminetetraacetic acid
- DNA Stain G (SERVA Electrophoresis GmbH)
- Ni-IMAC loading buffer: 20 mM Tris pH 7.6, 10 mM β-mercaptoethanol, 200 mM sodium chloride, 20 mM imidazole
- Ni-IMAC loading buffer: 20 mM Tris pH 7.6, 10 mM β-mercaptoethanol, 200 mM sodium chloride, 500 mM imidazole
- Size exclusion chromatography (SEC) buffer: 20 mM Tris pH 7.6, 10 mM β-mercaptoethanol, 200 mM sodium chloride
- Refolding buffer 1: 20 mM Tris pH 7.6, 200 mM sodium chloride, 10 mM β-mercaptoethanol, 6M guanidine hydrochloride
- Refolding buffer 2: 20 mM Tris pH 7.6, 200 mM sodium chloride, 10 mM β-mercaptoethanol, 1M guanidine hydrochloride
- Refolding buffer 3: 20 mM Tris pH 7.6, 200 mM sodium chloride, 10 mM β-mercaptoethanol, 2M guanidine hydrochloride
- SDS-PAGE stacking gel: 0.33 M Bis-Tris pH 6.5, 5% w/v 29:1 acrylamide:bisacrylamide, 0.01% v/v ammonium persulfate, 0.001% v/v tetramethylethylenediamine
- SDS-PAGE separating gel: 0.33 M Bis-Tris pH 6.5, 13.3% w/v 29:1 acrylamide:bisacrylamide, 0.01% v/v ammonium persulfate, 0.001% v/v tetramethylethylenediamine
- 2× SDS sample buffer: 100 mM Tris pH 6.8, 20% glycerol, 4% SDS, 200 mM DTT, 0.01% bromophenol blue.
- SDS-PAGE running buffer (VWR): 1% w/v 4-morpholineethanesulfonic acid, 0.6% w/v Tris, 0.1% w/v SDS, 0.035% w/v ethylenediaminetetraacetic acid

- SDS-polyacrylamide gel staining solution: 110 mg/L Coomassie Blue G-250, 5% w/v ethanol, 8% w/v phosphoric acid, 1% w/v cyclodextrin.

Equipment

- ÄKTA FPLC chromatography system (GE Healthcare)
- ÄKTA Prime Plus chromatography system (GE Healthcare)
- ÄKTA pure chromatography system (GE Healthcare)
- ÄKTA purifier chromatography system (GE Healthcare)
- HiLoad 16/60 Superdex 75 prep grade size exclusion column (Amersham Biosciences)
- HiLoad 16/60 Superdex 200 prep grade size exclusion column (Amersham Biosciences)
- Superdex 75 100/300 GL size exclusion column (Amersham Biosciences)
- Superdex 200 increase 100/300 GL size exclusion column (Amersham Biosciences)
- Avanti J-26S XPI High Performance Centrifuge (Beckman Coulter)
- JLA-8.1000 Fixed-Angle Aluminum Rotor (Beckman Coulter)
- JA-25.50 Fixed-Angle Aluminum Rotor (Beckman Coulter)
- BD 53 incubator (Binder)
- BLstar 16 blue light LED illuminator (Biometra GmbH)
- T3000 thermocycler (Biometra GmbH)
- Mini-PROTEAN Tetra vertical electrophoresis cell (Bio-Rad Laboratories)
- PowerPac Basic Power Supply (Bio-Rad Laboratories)
- 102c ultrasonic sonifier (Branson)
- BioPhotometer (Eppendorf AG)
- ThermoMixer C dry block (Eppendorf AG)
- Centrifuge 5920 R (Eppendorf AG)
- Centrifuge 5810 R (Eppendorf AG)
- Centrifuge 5427 R (Eppendorf AG)
- Centrifuge 5424 (Eppendorf AG)
- S-4x750 swing-bucket rotor (Eppendorf AG)
- A-4-44 swing-bucket rotor (Eppendorf AG)
- FA-45-48-11 fixed-angle rotor (Eppendorf AG)
- EPS 301 power supply (GE Healthcare)
- C24 incubator shaker (New Brunswick Scientific)
- Innova 44 incubator shaker (New Brunswick Scientific)
- VX-150 vertical floor-standing autoclave (Systec)
- Degassing Station (TA Instruments)
- E-Box VX2/20M gel documentation imaging system (Vilber)
- SRT6 analogue tube roller (Stuart Equipment)
- Multi Reax shaker and mixer (Heidolph Instruments)
- MR 3000 magnetic stirrer (Heidolph Instruments)
- Hei-Mix L magnetic stirrer (Heidolph Instruments)
- MR Hei-Standard magnetic stirrer (Heidolph Instruments)
- TW-26 white light transilluminator (UVP)

- Duomax 1030 platform shaker (Heidolph Instruments)
- Thermoblock V5
- pH 211 pH meter (Hanna Instruments)
- 572 precision balance (Kern & Sohn GmbH)
- ALS 120-4 analytical balance (Kern & Sohn GmbH)
- Heraeus Fresco 17 microcentrifuge (Thermo Scientific)
- Vortex mixer (VWR)
- HERAfreeze ultra-low temperature freezer (Thermo Scientific)
- Premium refrigerator (Liebherr)
- Pipetus battery powered pipette filler (Hirschmann-Laborgeräte GmbH)
- Pipetman pipettes (Gilson)
- LAB 500 CL Laboratory Washer (Steelco)
- Unichromat 700 chromatography cabinet for ÄKTA systems (UniEquip GmbH)
- Cary 50 UV-Vis spectrophotometer (Varian)
- FP-6500 spectrofluorimeter (JASCO)
- J-15 spectropolarimeter (JASCO)
- MicroPulser electroporator (BioRad)
- miniVE vertical electrophoresis system (Amersham Biosciences)
- Phoenix pipetting robot (Art Robbins Instruments)
- CryoLoops (Hampton Research)
- 3-well Intelli-Plates (Art Robbins Instruments)
- MRC Maxi 48 well crystallization plates (SWISSCI)

DNA sequences

The genes coding all designed chimeras used in this work are shown below in FASTA format.

```
>HiTyrA/DgFld
ATGGGTTTTAAAACTATCAATTCTGATATTCACAAAATTGTTATTGTGGGCGGTTATGGTAAATTAGGCGGCTTATTTGCCCGTTATTTACGTGCATCTG
GCTATCCAATTTCTATTTTAGATCGCGAAGATTGGGCGGTGGCTGAAAGTATTTTAGCGAATGCTGATGTCGTGTTGTTGGGTTGTAGCACCTGGGGGGA
CGACGAAATCGAGCTGCAAGAAGATTTTGTGCCGCTGTACGAGGACCTGGATCGCGCGGGTCTGAAGGATAAAAAAGTCGGCGTGTTCGGCTGCGGTGAT
TCCTCTTATACCTATTTCTGCGGTGCGGTCGACGTTATTGAAAAAAAAGCCGAGGAACTGGGCGCGACGTTGGTCGCAAGCAGCTTGAAGATCGATGGCG
AGCCGGATAGCGCGGAAGTTCTTGATTGGGCACGTGAAGTCCTGGCACGCGTTCTCGAGCACCACCACCACCACCACTGA

>HiTyrA/DgFld2
ATGGGTTTCAAAACCATTAATTCTGACATCCATAAAATCGTGGTTGTTGGCGGTTACGGTAAACTGGGCGGTTTGATTGCGCGTTATCTGCGTGCGAGCG
GCTACCCGATCAGCATTTTGGACCGCGAAGACTGGGCCGTCGCAGAGTCTATCCTGGCGAACGCAGACGTCGTTATCTTGGGTTGCTCCACTTGGGGCGA
CGACGAAATTGAGCTGCAAGAAGACTTCCTGCCACTGTACGAAGACCTGGATCGCGCGGGCCTGAAGGACAAGAAAGTAGGTGTGGCCGGCTGTGGTGAT
TCCAGCTATACCTATGAGTGCGGTGCAGTGGATGTCATTGAGAAGAAGGCTGAGGAACTGGGCGCCACGCTGGTGGCGTCCAGCCTGAAGATTGACGGTG
AGCCGGACTCGGCGGAGGTCCTGGACTGGGCGCGTGAAGTCTTGGCCCGCGTGGAAAACTTGTATTTCCAGGGCCACCACCACCATCACCACCACTAA

>HiTyrA/DgFld3
ATGGGTTTCAAAACGATCAATTCTGACATCCACAAAATTGTCGTCGTTGGTGGCTACGGCAAACTGGGCGGTCTGTTCGCGCGTTACCTGCGTGCGGCGG
GCTATCCGATTATTATTCTGGATCGTGAAGATTGGGCAGTGGCTGAAAGCATTCTGGCGAACGCGGATCTGGTGCTGCTGGGTTGCAGCACGTGGGGTGA
CGACGAGATCGAACTGCAGGAGGATTTCCTGCCTCTGTACGAGGATCTGGATCGCGCAGGTCTGAAAGATAAGAAGGTCGGTGTGTTTGGTTGTGGCGAT
AGCAGCTATACGTACGAATGTGGCGCGGTGGACGTCATCGAGAAAAAGGCGGAAGAGCTGGGCGCCACGCTGGTTGCGTCCAGCTTGAAGATTGACGGCG
AGCCGGATAGCGCTGAAGTGCTCGATACCGCCCGTGAGGTTCTGGCTCGCGTTGAGAACCTGTATTTCCAGGGTCATCATCACCACCATCACCACCACTA
A
```

>BsKtrA/DgFld
ATGAACAAACAATTCGCAGTGATCGGTTTGGGTCGTTTCGGTGGTTCGATTGTGAAGGAGCTGCACCGTATGGGCCATGAGGTGCTGGCGGTTGATATTA
ACGAGGAAAAGGTTAATGCGTACGCGAGCTACGCCACCCACGCAGTTATTGCTAATGCAACCGAGGAAAACGAACTGCTGTCCCTGGGCATCCGTAATTT
CGAGTACGTTTTGCTGGGCTGCTCTACCTGGGGTGACGACGAGATCGAGCTGCAGGAGGACTTTGTGCCGCTGTATGAAGACCTGGATCGTGCTGGCCTG
AAGGATAAGAAGGTGGGCGTTTTTGGTTGCGGTGACAGCTCCTACACGTACTTCTGCGGTGCAGTAGACGTTATTGAGAAAAAAGCTGAAGAGTTAGGCG
CAACCCTGGTGGCGTCTTCCCTGAAGATTGACGGCGAGCCGGATAGCGCGGAGGTTCTGGATTGGGCGCGCGAGGTTTTAGCGCGTGTTGAAAATCTGTA
CTTTCAGGGTCATCATCACCACCACCACCACCACTGA

>BsKtrA/DgFld2
ATGAACAAGCAATTCGCGGTCATCGGTTTGGGTCGCTTTGGTGAATCTGTCGTTAAGGAACTGCACCGTATGGGTCATGAGGTCCTGGCGGTTGACATCA
ACGAGGAGAAAGTTAACGCGGTGGCAAGCTACGCGACCCACGCGGTGATCGCGAACGCTACGAATGAAAATGAGCTGCTGAGCCTGGGTATCCGTAACTT
TGAATATGTGCTGTTGGGCGCTAGCACGTGGGGTGACGATGAGATCGAGCTGCAGGAGGACTTCCTGCCGCTGTACGAGGACCTGGATCGCGCGGGTCTG
AAAGATAAGAAAGTAGGTGTTTTCGGTAGCGGTGATAGCAGCTACACGTACTTCTGCGGTGCAGTTGATGTTATTGAGAAAAAAGCCGAAGAATTGGGCG
CGACCCTGGTGGCCAGCAGCCTGAAGATTGACGGCGAGCCGGACAGCGCGGAAGTGCTGGATTGGGCGCGCGAAGTGTTGGCACGCGTCGAGAATCTGTA
CTTTCAAGGTCATCATCATCACCACCACCATCACTGA

>BmMDH/MtFprA
ATGCGCAGCCATCATCATCATCACCATCATCATGGCTCTGAAAACCTGTACTTCCAAGGTGCACGTAATAAAATCGCGCTGATTGGTAGCGGCATGATCG
GTGGCACCCTGGCGCACCTGGCGGGTCTGAAGGAACTGGGCGATGTTGTGCTGTTCGACATTGCCGAGGGTACTCCGCAAGGCAAAGGCCTGGACATTGC
GGAGTCGAGCCCGGTGGATGGTTTTGACGCGAAATTCACCGGTGCCAATGACTACGCTGCGATCGAGGGTGCAGACGTGGTCTTAGTCGGTTCCCCGACC
ATTAACAATGACATCCTGCCGGTCGTGAGCCCATTGCTGGATGACCTGGTCGGCCTGCGTCCAAAGAATAAGGTAGGCCTGGCGTTTGGTGCCTACGGTT
GGGGTGGTGGCGCGCAAAAAATTTTAGAGGAACGTCTGAAGGCGGCGAAAATTGAGCTGATCGCAGAACCGGGCCCGACCGTACAGTGGGTCCCACGTGG
TGAAGATCTGCAACGTTGCTACGAACTGGGCCGTAAAATCGCGGCACGCATCGCAGATTAA

>BsKtrA/BmP450
ATGCGCTCTCATCACCACCATCACCACCACCATGGTAGCGAGAACTTGTATTTTCAGGGCAACAAGCAATTTGCCGTGATCGGTCTGGGTCGTTTCGGCG
GCAGCATTGTGAAGGAACTGCACCGCATGGGTCACGAGGTCCTCGCGGTTGACATTAATGAAGAGAAAGTCAACGCGTACGCTTCTTACGCGACCCATGC
AGTCATCGCTAATGCGACCGAGGAAAATGAATTGCTGTCTCTGGGTATTCGCAACTTCGAATACGTTCTGATCGTCACCGCGTCTTACAACGGCCACCCA
CCGGACAATGCGAAACAATTCGTCGACTGGTTGGATCAAGCTAGCGCCGATGAAGTGAAGGGCGTTAGATACTCCGTCTTCGGTTGCGGCGATAAGAATT
GGGCGACGACCTATCAAAAAGTCCCGGCATTTATCGATGAGACCCTGGCCGCGAAAGGTGCGGAGAACATCGCAGACCGTGGTGAGGCTGACGCAAGCGA
TGACTTCGAAGGCACGTACGAAGAATGGCGCGAACACATGTGGAGCGATGTGGCCGCATACTTCAACCTGTAA

>TmGDH/MtFprA
ATGCACCATCACCATCATCACGAGAATTTGTATTTCCAGTCTCGTTTCTTTGTGCTGGGTGCAGGCAGCTGGGGTACCGTGTTCGCGCAAATGCTGCATG
AGAATGGTGAGGAGGTCATTCTGTGGGCACGCCGTAAAGAGATCGTTGACCTGATCAATGTCAGCCACACGAGCCCGTACGTTGAAGAGTCCAAAATTAC
CGTGCGCGCGACGAATGATCTGGAAGAAATTAAAAAAGAGGATATTCTGGTGGTCGGTAGCCCGACGATTAATAACGACATTCTGCCAGTTGTTAGCCCG
CTGCTGGATGATCTGGTGGGTCTGCGCCCGAAGAATAAGGTGGGTTTAGCCTTCGGTGCATACGGCTGGGGCGGTGGCGCCCAGAAAATCCTGGAAGAGC
GTCTGAAAGCGGCAAAAATTGAACTCATTGCTGAACCGGGCCCTACTGTCCAATGGGTCCCGCGTGGTGAGGACCTGCAACGTTGTTATGAACTGGGCCG
CAAGATTGCAGCCCGTATTGCAGATTAA

>SsTyrA/ScLOT6
ATGCACCATCACCATCATCACGAGAATTTGTATTTCCAGTCTAAGATTGGCGTCGTTGGCCTGGGCCTGATCGGCGCATCTCTGGCCGGTGATCTACGTC
GTCGTGGTCATTACCTGATCGGCGTTTCCCGTCAACAAAGCACCTGCGAGAAGGCGGTTGAGCGCCAGCTGGTCGACGAGGCAGGTCAAGATTTGAGCCT
GCTGCAGACTGCGAAAATCATTGTCTTTGTTACCCCGCAGTACAACTGGGGCTATCCGGCAGCGCTGAAAAACGCTATCGATCGCCTGTATCACGAATGG
CACGGTAAGCCGGCGCTGGTGGTTAGCTACGGTGGCCACGGTGGTTCTAAATGCAACGATCAGCTGCAGGAAGTTCTGCATGGCCTCAAAATGAACGTTA
TCGGTGGTGTGGCGGTGAAAATTCCGGTCGGTACTATCCCGCTGCCGGAGGATATCGTGCCTCAGTTGTCCGTCCATAATGAGGAGATTTTGCAACTGCT
GGCCAGCTGTATCTGA

>TmGDH/DrWrbA
ATGCACCATCACCATCATCACGAGAATTTGTATTTCCAGTCTCGTTTTTTTGTTCTGGGTGCCGGCAGCTGGGGTACCGTTTTTGCTCAGATGCTGCATG
AAAACGGCGAGGAAGTTATCCTGTGGGCACGTCGTAAGGAGATTGTCGATCTGATCAATGTTAGCCACACGAGCCCGTATGTTGAAGAATCCAAGATTAC
GGTTCGCGCAACCAACGACCTGGAGGAGATTAAGAAGGAAGACATCCTGGTCTTCAGCTCCCCAACCCGTTTCGGTGGCGCAACCTCCCAAATGCGCGCA
TTTATCGACACCTTGGGTGGCCTGTGGAGCAGCGGTAAGCTGGCAAACAAAACGTTCAGCGCCATGACCAGCGCACAGAACGTGAACGGTGGCCAGGAGA
CCACGCTGCAAACCCTGTACATGACTGCAATGCATTGGGGTGCGGTTTTGACCCCGCCGGGTTATACGGACGAGGTTATCTTCAAGTCCGGTGGCAATCC
GTACGGTGCGAGCGTTACCGCGAACGGTCAACCGTTGTTGGAAAACGATCGTGCGAGCATCCGTCATCAGGTGCGTCGTCAGGTGGAGCTGACGGCGAAG
CTGCTGGAAGGCTAA

>TmGDH/DgROO
ATGCACCATCACCATCATCACGAGAATTTGTATTTCCAGTCTCGTTTTTTTGTTCTGGGCGCGGGTAGCTGGGGTACCGTTTTTGCCCAGATGCTGCACG
AGAATGGTGAGGAGGTTATTCTGTGGGCGCGCCGCAAAGAAATTGTTGACCTGATCAACGTAAGCCACACCAGCCCGTATGTGGAGGAAAGCAAGATTAC
CGTGCGTGCGACCAATGACCTGGAGGAGATCAAAAAAGAAGATATTCTGGTCGTTGGTAGCCCGACCCATAACAACGGTATTCTGCCGTACGTTGCAGGC
ACGTTGCAATACATTAAGGGTCTGCGTCCGCAAAATAAGATCGGTGGTGCGTTTGGTAGCTTCGGTTGGAGCGGCGAGAGCACGAAAGTGTTAGCGGAGT
GGTTGACCGGTATGGGTTTTGACATGCCGGCGACCCCGGTGAAAGTGAAGAACGTGCCGACGCATGCGGATTATGAACAGCTGAAGACCATGGCGCAAAC
GATTGCACGTGCTCTGAAGGCAAAGCTGGCAGCGTGA

>AfFNO/DdFld
ATGCACCATCACCATCATCACGAGAATTTGTATTTCCAGTCTCGTGTGGCGCTGCTGGGTGGCACCGGCAACCTGGGCAAAGGCCTGGCGCTGCGCTTAG
CAACGCTGGGCCATGAAATTGTTGTTGGTTCCCGCCGTGAGGAGAAGGCCGAAGCGAAGGCTGCCGAGTACCGTCGTATCGCAGGTGATGCCTCTATTAC
CGGCATGAAGAATGAGGACGCGGCGGAGGCCTGTGACATCGCGCTGTTCGGCTGTAGCGCCTGGGGCATGGAGGACCTGGAGATGCAGGACGATTTTTTG
TCTCTGTTCGAGGAATTTGACCGTATTGGCTTGGCAGGTCGTAAAGTTGCCGCATTTGCGAGCGGCGACCAAGAGTATGAACATTTCTGCGGCGCAGTCC
CGGCGATCGAGGAGCGTGCGAAAGAATTGGGTGCGACCATCATCGCGGAGGGTCTCAAGATGGAGGGCGATGCCAGCAACGATCCGGAGGCAGTGGCGAG
CTTCGCTGAGGATGTGCTGAAGCAACTGTAA

>EcPurT/EcPurN
ATGACGTTATTAGGCACTGCGCTGCGTCCGGCAGCAACTCGCGTGATGTTATTAGGCTCCGGTGAACTGGGTAAAGAAGTGGCAATCGAGTGTCAGCGTC
TCGGCGTAGAGGTGATTGCCGTCGATCGCTATGCCGACGCACCAGCCATGCATGTCGCGCATCGCTCCCATGTCATTAATATGCTTGATGGTGATGCATT
ACGCCGTGTGGTTGAACTGGAAAAACCACATTATATCGTGCTGGCTGGTTTTATGCGCATTCTCAGCCCGGCGTTTGTCTCCCACTATGCCGGGCGTTTG
CTGAACATTCACCCTTCTCTGCTGCCGAAATATCCCGGATTACACACCCATCGTCAGGCGCTGGAAAATGGCGATGAAGAGCACGGTACATCGGTGCATT
TCGTCACCGATGAACTGGACGGTGGCCCGGTTATTTTACAGGCGAAAGTCCCGGTATTTGCTGGTGATTCGGAAGATGACATCACCGCCCGCGTGCAAAC
CCAGGAACACGCCATTTATCCACTGGTGATTAGCTGGTTTGCCGATGGTCGTCTGAAAATGCACGAAAACGCCGCGTGGCTGGATGGTCAACGTCTGCCG
CCGCAGGGCTACGCTGCCGACGAGGGCAGCGAGAACTTGTACTTTCAAAGCGGTAGCCACCATCATCACCACCACTAA

>EcPurT/BhPurN
ATGACGTTATTAGGCACTGCGCTGCGTCCGGCAGCAACTCGCGTGATGTTATTAGGCTCCGGTGAACTGGGTAAAGAAGTGGCAATCGAGTGTCAGCGTC
TCGGCGTAGAGGTGATTGCCGTCGATCGCTATGCCGACGCACCAGCCATGCATGTCGCGCATCGCTCCCATGTCATTAATATGCTTGATGGTGATGCATT
ACGCCGTGTGGTTGAACTGGAAAAACCACATTATGTTGTGCTGGCAGGCTACATGCGTCTGGTTGGTCCGACGCTGTTAGGCGCATATGAAGGCCGCATT
GTTAACATTCACCCGTCCCTCCTGCCGGCGTTCCCGGGTCTGCATGCGATTGAACAGGCCATCCGTGCGAACGTGAAGGTGACCGGTGTCACTATTCATT
ATGTCGATGAGGGTATGGATACCGGTCCGATTATCGCGCAAGAGGCGGTTAGCATCGAGGAGGAGGATACTTTGGAGACCCTGACCACCAAAATTCAAGC
TGTAGAACACCGTCTGTACCCGGCGACGCTGCACAAACTGCTGAGCAAGGGCAGCGAGAACTTGTACTTTCAAAGCGGTAGCCACCATCATCACCACCAC
TAA

>EcPurT_GkPurN
ATGACGTTATTAGGCACTGCGCTGCGTCCGGCAGCAACTCGCGTGATGTTATTAGGCTCCGGTGAACTGGGTAAAGAAGTGGCAATCGAGTGTCAGCGTC
TCGGCGTAGAGGTGATTGCCGTCGATCGCTATGCCGACGCACCAGCCATGCATGTCGCGCATCGCTCCCATGTCATTAATATGCTTGATGGTGATGCATT
ACGCCGTGTGGTTGAACTGGAAAAACCACATTATATCGCGCTCGCAGGTTATATGCGCCTGATTGGTCCGACCCTGTTGTCTGCGTACGAAGGTAAAATC
GTTAATATTCACCCGTCGCTGCTGCCGGCCTTTCCGGGCAAAGACGCAATTGGCCAAGCGTACCGTGCTGGCGTTTCTGAGACGGGTGTAACCGTGCATT
ACGTGGATGAGGGCATGGACACGGGTCCGGTTATTGCGCAGCGTGTGGTTCCGATCGTCCCGGGCGAGCCTATTGAAGCTCTGGAAGAGCGTATTCACCA
AGTGGAGCACGAGCTGTATCCGACCGTGCTGCGTATGCTGCTGGGCGAGAAAGAGCAGCAAGAAGAACGTATCGAAAATGACGGCTCGGAGACGAGCATC
GACCAGCGCGTTCAGGGCAGCGAGAACTTGTACTTTCAAAGCGGTAGCCACCATCATCACCACCACTAA

>EcPurT_BaFmt
ATGACGTTATTAGGCACTGCGCTGCGTCCGGCAGCAACTCGCGTGATGTTATTAGGCTCCGGTGAACTGGGTAAAGAAGTGGCAATCGAGTGTCAGCGTC
TCGGCGTAGAGGTGATTGCCGTCGATCGCTATGCCGACGCACCAGCCATGCATGTCGCGCATCGCTCCCATGTCATTAATATGCTTGATGGTGATGCATT
ACGCCGTGTGGTTGAACTGGAAAAACCACATTATATCGTTACCGCAGCATTCGGCCAAATTGTGCCGAACGAGATTTTAGAGGCCCCGAAATATGGTTGT
ATCAACGTTCACGCGAGCCTGCTGCCGGAACTGCGTGGCGGTGCACCGATCCACTACGCGATTATGGAGGGTAAAGAAAAAACTGGCATCACGATTATGT
ACATGGTCGAAAAACTGGACGCCGGCGACATTCTGACCCAGGTGGAGGTGGAAATTGAGGAGCGCGAAACCACCGGCAGCCTGTTCGATAAGTTGTCGGA
GGCGGGCGCGCACCTGCTGTCGAAGACCGTTCCGCTGCTGATTCAGGGCAAACTGGAACCTATCAAGCAGAACGAGGAGGAGGTCACCTTTGCGTACAAT
GGCAGCGAGAACTTGTACTTTCAAAGCGGTAGCCACCATCATCACCACCACTAA

>EcPurN_EcPurT
ATGAATATTGTGGTGCTTATTTCCGGCAACGGAAGTAATTTACAGGCAATTATTGACGCCTGTAAAACCAACAAAATTAAAGGCACCGTACGGGCAGTTT
TCAGCAATAAGGCCGACGCGTTCGGCCTTGAACGCGCCCGCCAGGCGGGTATTGCAACGCATACGCTCATCGCCAGCGCGTTTGACAGTCGTGAAGCCTA
TGACCGGGAGTTGATTCATGAAATCGACATGTACGCACCCGATGTGGTCGTGCTGGAGATCGAAGCTATTGCCACCGATATGCTGATCCAACTTGAAGAG
GAAGGACTGAATGTTGTCCCCTGCGCTCGCGCAACGAAATTAACGATGAATCGCGAGGGTATCCGTCGCCTGGCGGCAGAAGAGCTGCAGCTGCCCACTT
CCACTTATCGTTTTGCCGATAGCGAAAGCCTTTTCCGCGAGGCGGTTGCTGACATTGGCTATCCCTGCATTGTAAAACCGGTGATGAGCTCTTCCGGCAA
GGGGCAGACGTTTATTCGTTCTGCAGAGCAACTTGCTCAGGCATGGAAGTACGCTCAGCAAGGCGGTCGCGCCGGAGCGGGCCGCGTAATTGTTGAAGGC
GTCGTTAAGTTTGACTTCGAAATTACCCTGCTAACCGTCAGCGCGGTGGATGGCGTCCATTTCTGTGCACCAGTAGGTCATCGCCAGGAAGATGGCGACT
ACCGTGAATCCTGGCAACCACAGCAAATGAGCCCGCTTGCCCTTGAACGTGCGCAGGAGATTGCCCGTAAAGTGGTGCTGGCACTGGGCGGTTATGGGTT
GTTTGGTGTCGAGCTATTTGTCTGTGGTGATGAGGTGATTTTCAGTGAGGTCTCCCCTCGTCCACATGATACCGGGATGGTGACGTTAATTTCTCAAGAT
CTCTCAGAGTTTGCCCTGCATGTACGTGCCTTCCTCGGACTTCCGGTTGGCGGGATCCGTCAGTATGGTCCTGCAGCTTCTGCCGTTATTCTGCCACAAC
TGACCAGTCAGAATGTCACGTTTGATAATGTGCAGAATGCCGTAGGCGCAGATTTGCAGATTCGTTTATTTGGTAAGCCGGAAATTGATGGCAGCCGTCG
TCTGGGGGTGGCACTGGCTACTGCAGAGAGTGTTGTTGACGCCATTGAACGCGCGAAGCACGCCGCCGGACAGGTAAAAGTACAGGGTGGCAGCGAGAAC
TTGTACTTTCAAAGCGGTAGCCACCATCATCACCACCACTAA

>BhPurN_EcPurT
ATGAAACGTGTCGCTATCTTCGCTTCCGGCTCTGGTACCAATGCAGAGGCAATCATCCAGTCTCAAAAGGCAGGCCAGCTGCCATGTGAAGTAGCACTGC
TGATCACGGACAAACCGGGTGCGAAAGTTGTTGAACGTGTCAAAGTGCACGAGATTCCGGTCTGCGCATTGGACCCGAAGACGTACCCAAGCAAAGAGGC
ATATGAGATTGAAGTGGTCCAACAACTGAAGGAGAAGCAAATCGACTTCATCGTGCCGGAGATCGAAGCTATTGCCACCGATATGCTGATCCAACTTGAA
GAGGAAGGACTGAATGTTGTCCCCTGCGCTCGCGCAACGAAATTAACGATGAATCGCGAGGGTATCCGTCGCCTGGCGGCAGAAGAGCTGCAGCTGCCCA
CTTCCACTTATCGTTTTGCCGATAGCGAAAGCCTTTTCCGCGAGGCGGTTGCTGACATTGGCTATCCCTGCATTGTAAAACCGGTGATGAGCTCTTCCGG
CAAGGGGCAGACGTTTATTCGTTCTGCAGAGCAACTTGCTCAGGCATGGAAGTACGCTCAGCAAGGCGGTCGCGCCGGAGCGGGCCGCGTAATTGTTGAA
GGCGTCGTTAAGTTTGACTTCGAAATTACCCTGCTAACCGTCAGCGCGGTGGATGGCGTCCATTTCTGTGCACCAGTAGGTCATCGCCAGGAAGATGGCG
ACTACCGTGAATCCTGGCAACCACAGCAAATGAGCCCGCTTGCCCTTGAACGTGCGCAGGAGATTGCCCGTAAAGTGGTGCTGGCACTGGGCGGTTATGG
GTTGTTTGGTGTCGAGCTATTTGTCTGTGGTGATGAGGTGATTTTCAGTGAGGTCTCCCCTCGTCCACATGATACCGGGATGGTGACGTTAATTTCTCAA
GATCTCTCAGAGTTTGCCCTGCATGTACGTGCCTTCCTCGGACTTCCGGTTGGCGGGATCCGTCAGTATGGTCCTGCAGCTTCTGCCGTTATTCTGCCAC
AACTGACCAGTCAGAATGTCACGTTTGATAATGTGCAGAATGCCGTAGGCGCAGATTTGCAGATTCGTTTATTTGGTAAGCCGGAAATTGATGGCAGCCG
TCGTCTGGGGGTGGCACTGGCTACTGCAGAGAGTGTTGTTGACGCCATTGAACGCGCGAAGCACGCCGCCGGACAGGTAAAAGTACAGGGTGGCAGCGAG
AACTTGTACTTTCAAAGCGGTAGCCACCATCATCACCACCACTAA

>GkPurN_EcPurT
ATGAAACGTTTGGCCGTTTTCGCATCCGGTTCTGGTACCAATTTTCAAGCTATCGTTGACGCAGCTAAGCGCGGTGACCTGCCGGCACGTGTAGCGCTGC
TGGTTTGCGATCGTCCGGGTGCAAAAGTTATCGAGCGTGCAGCGCGCGAAAATGTCCCTGCGTTTGTGTTTAGCCCAAAGGATTACCCGAGCAAAGCCGC
TTTTGAGAGCGAGATTTTGCGCGAGCTGAAGGGTCGTCAGATCGACTGGATCGTGCCGGAGATCGAAGCTATTGCCACCGATATGCTGATCCAACTTGAA
GAGGAAGGACTGAATGTTGTCCCCTGCGCTCGCGCAACGAAATTAACGATGAATCGCGAGGGTATCCGTCGCCTGGCGGCAGAAGAGCTGCAGCTGCCCA
CTTCCACTTATCGTTTTGCCGATAGCGAAAGCCTTTTCCGCGAGGCGGTTGCTGACATTGGCTATCCCTGCATTGTAAAACCGGTGATGAGCTCTTCCGG
CAAGGGGCAGACGTTTATTCGTTCTGCAGAGCAACTTGCTCAGGCATGGAAGTACGCTCAGCAAGGCGGTCGCGCCGGAGCGGGCCGCGTAATTGTTGAA
GGCGTCGTTAAGTTTGACTTCGAAATTACCCTGCTAACCGTCAGCGCGGTGGATGGCGTCCATTTCTGTGCACCAGTAGGTCATCGCCAGGAAGATGGCG
ACTACCGTGAATCCTGGCAACCACAGCAAATGAGCCCGCTTGCCCTTGAACGTGCGCAGGAGATTGCCCGTAAAGTGGTGCTGGCACTGGGCGGTTATGG
GTTGTTTGGTGTCGAGCTATTTGTCTGTGGTGATGAGGTGATTTTCAGTGAGGTCTCCCCTCGTCCACATGATACCGGGATGGTGACGTTAATTTCTCAA
GATCTCTCAGAGTTTGCCCTGCATGTACGTGCCTTCCTCGGACTTCCGGTTGGCGGGATCCGTCAGTATGGTCCTGCAGCTTCTGCCGTTATTCTGCCAC
AACTGACCAGTCAGAATGTCACGTTTGATAATGTGCAGAATGCCGTAGGCGCAGATTTGCAGATTCGTTTATTTGGTAAGCCGGAAATTGATGGCAGCCG
TCGTCTGGGGGTGGCACTGGCTACTGCAGAGAGTGTTGTTGACGCCATTGAACGCGCGAAGCACGCCGCCGGACAGGTAAAAGTACAGGGTGGCAGCGAG
AACTTGTACTTTCAAAGCGGTAGCCACCATCATCACCACCACTAA

>BaFmt_EcPurT
ATGATTAAAGTTGTGTTCATGGGTACGCCTGATTTCTCCGTCCCTGTTCTCAGACGTCTGATCGAGGACGGCTATGACGTCATCGGTGTGGTTACCCAGC
CGGACCGTCCGGTGGGTCGCAAGAAAGTGCTGACCCCGACTCCAGTGAAGGTGGAGGCGGAGAAGCATGGTATTCCGGTCCTGCAACCGTTGCGTATCCG
TGAAAAAGACGAATATGAGAAGGTCCTGGCGTTGGAGCCGGACCTGATTGTGCCGGAGATCGAAGCTATTGCCACCGATATGCTGATCCAACTTGAAGAG
GAAGGACTGAATGTTGTCCCCTGCGCTCGCGCAACGAAATTAACGATGAATCGCGAGGGTATCCGTCGCCTGGCGGCAGAAGAGCTGCAGCTGCCCACTT
CCACTTATCGTTTTGCCGATAGCGAAAGCCTTTTCCGCGAGGCGGTTGCTGACATTGGCTATCCCTGCATTGTAAAACCGGTGATGAGCTCTTCCGGCAA
GGGGCAGACGTTTATTCGTTCTGCAGAGCAACTTGCTCAGGCATGGAAGTACGCTCAGCAAGGCGGTCGCGCCGGAGCGGGCCGCGTAATTGTTGAAGGC
GTCGTTAAGTTTGACTTCGAAATTACCCTGCTAACCGTCAGCGCGGTGGATGGCGTCCATTTCTGTGCACCAGTAGGTCATCGCCAGGAAGATGGCGACT
ACCGTGAATCCTGGCAACCACAGCAAATGAGCCCGCTTGCCCTTGAACGTGCGCAGGAGATTGCCCGTAAAGTGGTGCTGGCACTGGGCGGTTATGGGTT
GTTTGGTGTCGAGCTATTTGTCTGTGGTGATGAGGTGATTTTCAGTGAGGTCTCCCCTCGTCCACATGATACCGGGATGGTGACGTTAATTTCTCAAGAT
CTCTCAGAGTTTGCCCTGCATGTACGTGCCTTCCTCGGACTTCCGGTTGGCGGGATCCGTCAGTATGGTCCTGCAGCTTCTGCCGTTATTCTGCCACAAC
TGACCAGTCAGAATGTCACGTTTGATAATGTGCAGAATGCCGTAGGCGCAGATTTGCAGATTCGTTTATTTGGTAAGCCGGAAATTGATGGCAGCCGTCG
TCTGGGGGTGGCACTGGCTACTGCAGAGAGTGTTGTTGACGCCATTGAACGCGCGAAGCACGCCGCCGGACAGGTAAAAGTACAGGGTGGCAGCGAGAAC
TTGTACTTTCAAAGCGGTAGCCACCATCATCACCACCACTAA

Software

- Rosetta 3.6-3.12
- R 3.3-3.6
- Python 2.7 and 3.8
- PostgreSQL 9.5
- CLANS

Methods

Fuzzle deposition pipeline

The starting point of this doctoral project was the analysis of the Fuzzle database. It is however essential to explain the database design and build process to understand all variables and parameters used in the examination procedure.

The initial input to be processed are sequences (in FASTA format) and structures (in PDB format) from the SCOPe release 2.06 clustered at a 95% identity cutoff (SCOPe95).[1] This set contains 28 010 representative domains covering 4 775 protein families in total.

[1] Fox et al., 2013

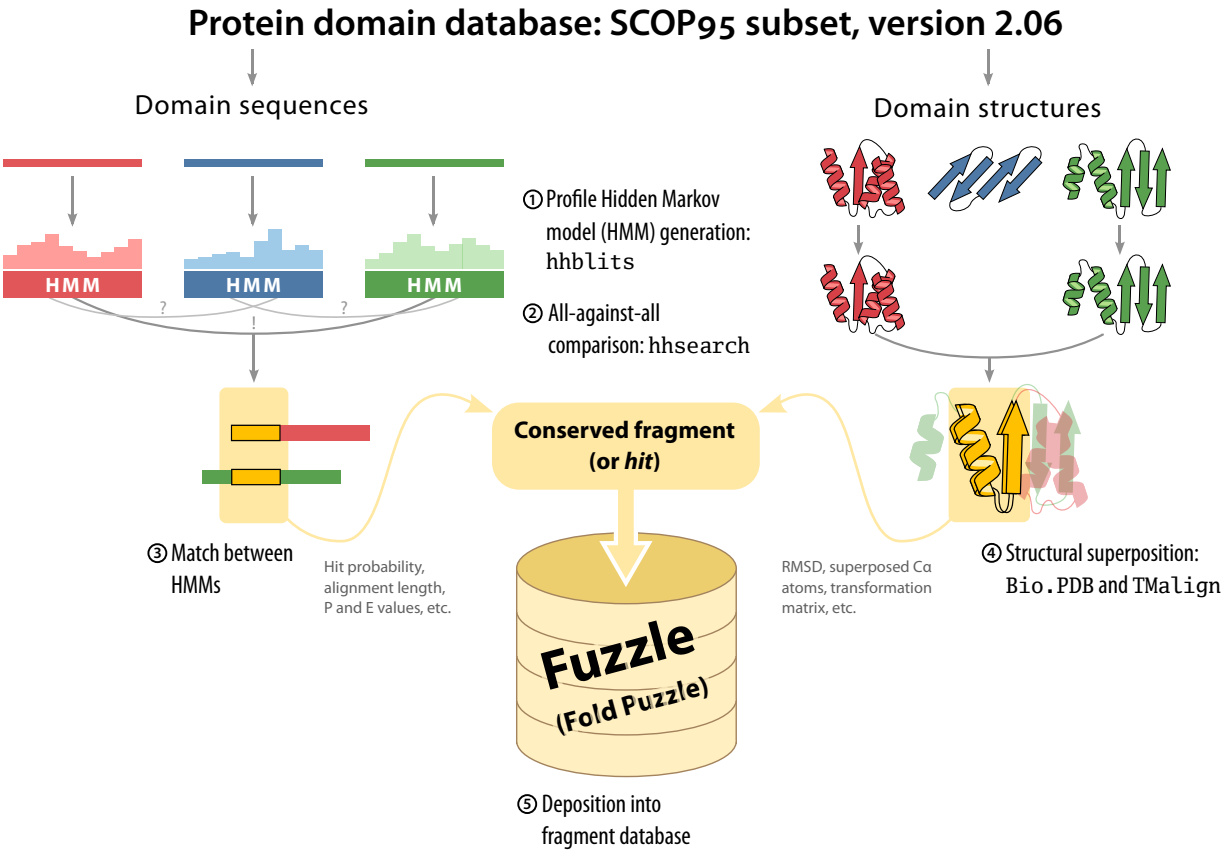


Figure 14: Simplified Fuzzle pipeline showing the sequence-based and structure-based steps and the storage of hits in the database.

The data deposition pipeline is shown in Figure 14. For each query domain sequence, an alignment in A3M format was generated by running the `buildali.pl` script from HH-suite 1.5 and PSI-BLAST with the nr20 sequence database. This A3M alignment is then used as an input for hhmake to generate an HHsearch format (`.hhm`) profile HMM file. After generat-

ing HMM files for all domain sequences in the SCOP domain subset, `hhsearch` was run for every one of them against all other HMM files as subject database with secondary structure scoring disabled (`-ssm 0`) to prevent possible biases originating from the PSIPRED prediction implemented in HHsearch. This operation gave as a result HHsearch output files for every single domain in the dataset. Each one of these consist in a list of hits that was then parsed with CSB (Computational Structural Biology Toolbox)[2] and relevant columns were extracted (see Table 1). Afterwards, for each single HHsearch hit in those lists, query and subject domain structures were loaded and trimmed according to the start and end residues present in the hit. These substructures were superimposed with the `Superimposer` module from Biopython[3] and the query-subject residue pairings defined by the HHsearch pairwise alignment ($RMSD_{Seq}$). From this superposition both RMSD and number of aligned Cα atoms were extracted.

[2] Kalev et al., 2012

[3] Hamelryck & Manderick, 2003

Then both structure fragments were superimposed with TM-align[4] and the HHsearch alignment as the initial state ($RMSD_{TM}$). Relevant parameters such as RMSD, number of aligned Cα atoms, and TM-score were extracted. This same procedure was repeated with the full domain structures. Finally, all hits and their extracted values were deposited in Fuzzle as a PostgreSQL database after adding a unique integer identifier to each of them. Additionally, the TM-align transformation matrix of every hit between domains from different folds was stored. Analysis of Fuzzle hits was performed in either PostgreSQL or R, the latter being used mostly for statistical assessments.

[4] Zhang & Skolnick, 2005

Table 1: Variables stored for each Fuzzle hit

Column	Description
`id`	Hit index
`query`	Query domain stable domain identifier (sid)
`q_scop_id`	Query domain SCOP(e) concise classification string identifier (sccs)
`no`	Hit number in query hitlist
`sbjct`	Subject domain sid
`s_scop_id`	Subject sccs
`s_desc`	Subject description according to scop
`prob`	HHsearch hit probability

(Continued on next page.)

Table 1: (cont.)

Column	Description
`eval`	HHsearch hit *E*-value
`pval`	HHsearch hit *P*-value
`score`	HHsearch alignment raw score
`ss`	Secondary structure score (always set to 0)
`cols`	Number of aligned match columns in the HMM-HMM alignment
`q_start`	Start residue for query domain
`q_end`	End residue for query domain
`s_start`	Start residue for subject domain
`s_end`	End residue for subject domain
`hmm`	Number of match states in subject HMM
`ident`	Pairwise sequence identity of the local alignment
`q_sufam_id`	Query sccs up to the superfamily level
`s_sufam_id`	Subject sccs up to the superfamily level
`q_fold_id`	Query sccs up to the fold level
`s_fold_id`	Subject sccs up to the fold level
`rmsd_pair`	RMSD using RMSD_{Seq} approach
`ca_pair`	Number of aligned Cα pairs using RMSD_{Seq} approach
`rmsd_tm_pair`	RMSD using RMSD_{TM} approach
`score_tm_pair`	TM-score using RMSD_{TM} approach
`ca_tm_pair`	Number of aligned Cα pairs using RMSD_{TM} approach
`rmsd_tm`	Full-domain RMSD using RMSD_{TM} approach
`score_tm`	TM-score of full-domain superposition
`ca_tm`	Number of aligned Cα pairs using RMSD_{TM} approach
`q_tm_start`	Start residue for query domain of superposition using RMSD_{TM} approach
`q_tm_end`	End residue for query domain of superposition using RMSD_{TM} approach
`s_tm_start`	Start residue for subject domain of superposition using RMSD_{TM} approach
`s_tm_end`	End residue for subject domain using RMSD_{TM} approach

Clustering analysis of Fuzzle hits

To generate and analyze hit networks, Fuzzle hits of interest were loaded into CLANS,[5] which implements a version of the Fruchterman-Reingold force-directed graph layout algorithm.[6] In these graphs, nodes represent domains while edges represent hits with their corresponding HHsearch *P*-value. Hits were clustered according to a previously defined *P*-value cutoff. The clustering procedure was run in 2D until equilibrium, i.e., when the overall node movement was negligible

[5] Frickey & Lupas, 2004

[6] Fruchterman & Reingold, 1991

Construction of a domain-ligand interaction dataset

Domain fragments from different folds may have conserved ligand-interacting regions which transcend fold boundaries. These motifs can provide additional information when evolutionary connections are described or at the moment of designing functional chimeric proteins. It is worthwhile, then, to search these kind of patterns in Fuzzle hits and analyze them in depth. Unfortunately, the PDB-style files available in SCOP — the same ones used in the Fuzzle pipeline — lack heteroatoms, which makes them inadequate to assess domain-ligand interactions in the first place. To identify them among all SCOP domains, domain atom coordinates were extracted from the parental PDB files according to the corresponding domain descriptor from SCOP, while keeping all heteroatoms present in the original coordinate file. PLIP (Protein-Ligand Interaction Profiler)[7] was then run on the domain PDB files to generate a list of ligand-domain relationships.

[7] Salentin et al., 2015

For selected hits in Fuzzle, query and subject domains were superimposed with the `Superimposer` module of Biopython and the transformation matrix previously calculated with TM-align. SCOP domain structures homologous to either the query or subject domain (as defined by ASTRAL[8]) were subsequently superimposed and then the ligand-domain interactions previously validated with PLIP were filtered. Atom-atom distances between query and subject ligand clusters were then measured with Biopython and stored for posterior analysis (Figure 15).

[8] Chandonia et al., 2004

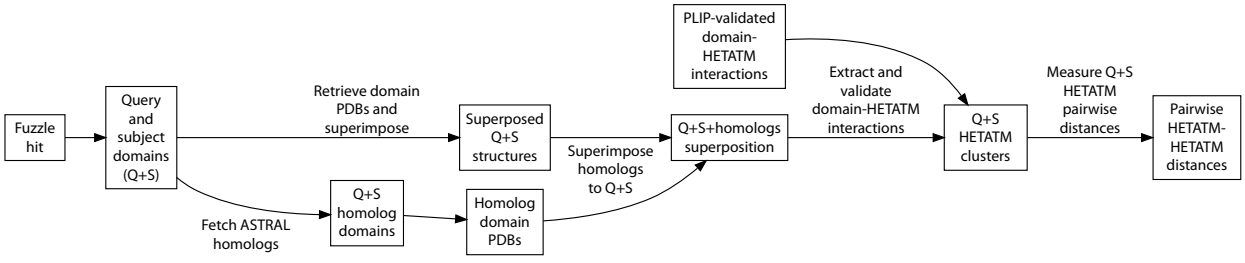


Figure 15: Ligand analysis pipeline for Fuzzle hits.

Ligand interaction in Fuzzle fragments was analyzed by mapping the ligand-interacting residues found with PLIP to the HHsearch alignment of the Fuzzle hit (Figure 16). Ligand-interacting residues were considered conserved if both sequences have ligand-binding interactions in the same region of the alignment.

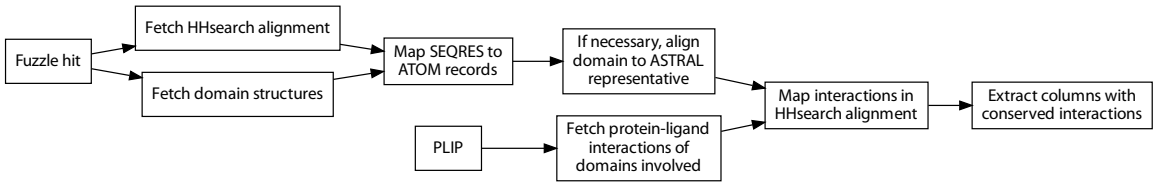


Figure 16: Pipeline for the discovery of conserved ligand-interacting residues in Fuzzle hits.

Creation of chimera structural models

Once a hit was chosen to generate a fold chimera, query and subject domains from the target hit were superimposed according to their TM-align transformation matrix. Steric clashes between both fragments were avoided with the help of contact maps calculated with Biopython. The crossover point was chosen based on Cα-Cα distances between adjacent residues in both domains. In other words, the distance between the last residue of the N-terminal parental fragment and the first residue of the C-terminal one should be as close to 3.8 Å (the average Cα-Cα distance in a polypeptide backbone) as possible. In some cases shape complementarity of the fragments was evaluated by rigid-body protein-protein docking using the Cluspro server.[9] After merging the atomic coordinates of N- and C-terminal halves based on the transformation matrix of the target Fuzzle hit, the chimera models were refined with the `FastRelax` mover in RosettaScripts.[10] Relax is an all-atom refinement algorithm that performs several rounds of side-chain packing and backbone minimization, with a gradual increase on the repulsive weight in the scoring function. The `ref2015` scoring function and either the `legacy` or `MonomerRelax2019` relax script were used and a total of 10 000 relax runs (resulting in the same amount of poses) were performed for each chimera model.

[9] Kozakov et al., 2017

[10] Tyka et al., 2011; Fleishman et al., 2011

For mutational studies fixed-backbone sequence design was performed on the lowest-energy pose obtained in the previous relax stage using RosettaScripts and the `PackRotamers` side chain optimization mover with sequence constraints. To encourage the generation of trajectories with few mutations containing large score differences, as opposed to a high amount of mutated residues with comparatively low scoring changes,

a `FavorNativeResidue` mover was applied to include a scoring bonus if the residue went unchanged during the design procedure. In other words, a mutation in the structure will be accepted only if said scoring bonus is surpassed.

To evaluate the models an approximate energy landscape (RMSD to the lowest-energy pose vs. Rosetta score of the pose) was built. This landscape was used subsequently to assess alternate conformations. Amino acid residues in the chimera, especially the ones at the fragments' interface, might acquire beneficial or detrimental effects due to the new environment they are brought into. This was estimated by relaxing parental and chimera structures as described before and then filtering the 500 lowest-scoring poses in each case. Then the average score for each residue was derived and the difference between the residue average in every chimera residue and its corresponding parental one was calculated.

Clusters of hydrophobic residues in the parental fragments should be combined in the final chimera in order to avoid large breaks or rearrangements that could disrupt the desired structure. Therefore, hydrophobic residue (ILV) clusters of the lowest-energy pose as well as the parental domains were determined using the BASiC server.[11]

[11] Kathuria et al., 2016

Cloning

Once the chimera models were built and evaluated, genes for the selected ones were designed, codon-optimized with PySplicer[12] and ordered as either gene fragments containing NdeI and XhoI flanking restriction sites or readily cloned pET21a(+) vectors. Gene fragments were resuspended in nuclease-free water to a concentration of 10 ng/μL. 100 ng of this solution were digested with NdeI and XhoI and the mix shown in Table 2.

[12] Available at `https://gitlab.com/cathalgarvey/pysplicer`

Table 2: Reaction mix for DNA digestion.

Reagent	Volume
Substrate DNA	Variable
CutSmart Buffer (10×)	4 µL
NdeI	30 units
XhoI	10 units
H_2O	to 30 µL

The digestion mix was incubated at 37°C for 1 hour. The digested gene fragment was then purified with the DNA Clean & Concentrator kit according to the manufacturer's indications. After purification the DNA fragment was ligated into 50 ng of predigested pET21a(+) plasmid using a 1:5 molar ratio of vector to insert and the ligation mix (Table 3).

Table 3: Reaction mix for gene fragment ligation.

Reagent	Volume
Digested gene fragment	Variable
Digested vector	Variable
T4 buffer (10×)	1 µL
T4 ligase	400 units
H_2O	to 10 µL

The ligation mix was then incubated at 4°C overnight. The ligated vector was used to transform chemically competent DH5α or Top10 cells. Frozen cell aliquots were thawed on ice before adding ligated DNA and incubating on ice for 15 min. Cells were then heat shocked at 42°C for 45 s. After further incubating on ice for 10 min, the cells were resuspended in 1 mL of LB medium without antibiotic and incubated under shaking at 37°C for 45 min. 200 µL of the resuspended cells were plated on agar plates with LB medium and 50 µg/mL ampicillin. The plates were then incubated for 16 hours at 37°C.

Colony PCR

To screen for properly ligated constructs individual colonies were picked and resuspended in the PCR mix (Table 4).

Table 4: Reaction mix for colony PCR.

Reagent	Volume
10× Taq Buffer	2.5 µL
dNTPs (10 µM)	0.5 µL
Gene-specific Forward primer (10 µM)	0.5 µL
Gene-specific Reverse primer (10 µM)	0.5 µL
Taq polymerase	0.125 µL
Nuclease free H_2O	20.875 µL
Total	25 µL

The PCR reactions were run with the following program (Table 5):

Table 5: Thermocycler program for colony PCR.

Temperature (°C)	Time	
95	5 min	
95	20 s	× 30 cycles
55	20 s	
72	1 min/kb	
72	5 min	
4	∞	

After amplification the PCR products were mixed with DNA loading buffer in a 5:1 ratio, loaded in a 1% agarose gel submerged in TAE buffer and run at 80 V for 30 minutes.

Transformation of expression strains

Once a construct sequence was verified by DNA sequencing, A 5 mL overnight culture was set up and the plasmid containing the chimera gene was isolated using the NucleoSpin Plasmid purification kit. Then the purified plasmid was used to transform either chemically competent *E. coli* cells such as BL21, as described before, or electrocompetent cells. In the case of T7 Express electrocompetent *E. coli* cells, 50 ng of plasmid were added to 100 µl of competent cells. After mixing, the cell-DNA mixture was transferred to a chilled electroporation cuvette and electroporated with the MicroPulser electroporator and its Ec2 settings (2.5 kV, 1 pulse). Afterwards, 900 µl LB medium

were added to the cell mixture, mixed, transferred to a microcentrifuge tube and incubated for 60 min at 37°C. Finally, 10 µl of the transformed cells were plated on agar plates with LB medium and 50 µg/mL ampicillin and incubated overnight at 37°C.

Test expression of recombinant proteins

One colony of the transformed cells on the plates was picked and used to inoculate 5 mL of LB-Amp medium. This culture was grown overnight at 37°C. Then, a 250 mL flask containing 25 mL of LB-Amp medium were inoculated with 200 µL of the overnight culture and incubated with agitation at 37°C until it reached an OD_{600} of approximately 0.7. The culture was then split into two flasks, where one of them had 1 mM of IPTG added into it. Both flasks were further incubated for 4 hours. Cells were then harvested by centrifugation at 4°C and 4 000 rpm for 15 minutes. The resulting cell pellets were stored at -20°C if they were not employed immediately. Cell pellets were resuspended in 1 mL of SEC buffer and sonicated on ice with a small-sized tip and the following program: Output control 3, 30% duty cycle, 2 × 3 minutes with a 1 minute pause between them. This lysate was centrifuged at 4°C and 18 000 rpm for 1 hour. The composition of the supernatant and insoluble fraction was then assessed by SDS-PAGE.

SDS-PAGE

To analyze the purity and size range of protein samples, they were mixed with an equal volume of 2x SDS sample buffer. Then they were denatured by heating at 95°C for five minutes. After centrifuging at 18 000 rpm for a minute, 10 µL of the sample was loaded in a separating SDS-PAGE gel topped with a stacking one and ran at 100 V until the dye front was nearly at the bottom of the gel. Gels were then rinsed with distilled water and incubated with the gel staining solution.

Large scale protein expression

After test expression the culture volume was upscaled, usually to 1 or 6 L. Per liter of culture, 10 mL of LB-Amp medium were inoculated with a transformed colony and grown overnight at

37°C. This preculture was used to inoculate a 5 L flask containing 1 L of LB-Amp medium that was incubated at 37°C under shaking until the culture reached an OD_{600} of approximately 0.7. Then expression was induced with the addition of 1 mM IPTG and further incubated for 4 hours. After four hours have elapsed, cells were harvested by centrifugation at 4 000 rpm and 4°C. The resulting cell pellets were either immediately used for purification or stored at -20°C for later use.

Protein purification

Cell pellets were resuspended in 10 mL of Ni-IMAC loading buffer per liter of original culture and sonicated on ice with a medium-sized tip and the following program: Output control 4, 50% duty cycle, 2 times 2 min with a 1 min pause. The lysate was then centrifuged at 18 000 rpm and 4°C for 60 min. In case of refolding, the insoluble pellet was taken and further processed. The supernatant, on the other hand, was passed through a 0.22 µm syringe filter and then loaded into a HisTrap HP/FF 5 mL column previously equilibrated with 5 column volumes of loading buffer. After loading the sample, the column was washed with 10 column volumes of loading buffer and then the protein was eluted using a 0-100% linear gradient of elution buffer. Elution fractions containing protein (i.e, absorbing at 280 nm) were collected and loaded into polyacrylamide gels. Fractions with the protein of interest were pooled, concentrated and loaded into the preparative size exclusion column previously equilibrated with SEC buffer. Collected fractions with absorption at 280 nm were analyzed by SDS-PAGE and, when the protein of interest was present with high purity, pooled and concentrated for biophysical characterization.

Refolding

In the case a protein expressed as inclusion bodies, refolding was attempted. For each liter of culture, the insoluble pellet obtained in the sonication step was washed in 10 mL of SEC buffer and centrifuged at 18 000 rpm and 4°C for 60 min. After repeating the wash step, the resulting pellet was resuspended in 5 mL of refolding buffer 1 and stirred slowly at 4°C for 60 min. Then 5 mL of refolding buffer 2 were added to the refolding solution, stirred one more time at 4°C for 60 min,

and centrifuged at 18 000 rpm and 4°C for 60 min. When the centrifugation was finished the supernatant was taken and refolding buffer 3 was added to a final volume of 25 mL, which was then dialyzed at 4°C against 5 L of SEC buffer with three buffer changes. The dialyzed sample was then centrifuged at 18 000 rpm and 4°C for 60 min, and purified as described above.

Biophysical characterization

Purified protein samples were subjected to analytical size exclusion chromatography to estimate their molecular weight in solution. An analytical size exclusion column was equilibrated with 5 column volumes of SEC buffer. The samples were passed through a 0.22 µm syringe filter and then injected into the ÄKTA system. Then the sample was eluted with two column volumes of SEC buffer.

Circular dichroism (CD) of the protein samples was measured in a J-15 spectropolarimeter using a 1mm quartz cuvette. The instrument was set as follows: 195-250 nm wavelength range, bandwidth 1 nm, response 2 s, data pitch 0.1 nm, scanning speed 100 nm/min, accumulation 5. The melting temperature of the protein samples was determined by thermal denaturation and CD monitoring at 222 nm. The temperature range analyzed was from 20 to 95°C in 1°C increments. The same settings as the CD spectrum measurements were used, with a temperature slope of 1°C/min. After denaturation the sample was cooled down slowly and a CD spectrum was acquired to check for refoldability.

Protein crystallography

Crystallization trials were set up using a sparse matrix screen (JCSG Core suite) in 3-well Intelli-Plates. Drops were pipetted using a Phoenix pipetting robot. Crystal hits were obtained by sitting drop vapour diffusion with a protein concentration of 26~mg/mL, drop volume of 0.8 µL, and drop ratio of 1:1 at 293~K. The initial hit condition (Core III B11, 0.2 M di-Ammonium hydrogen phosphate, 20% PEG 3350) was further optimized by varying the mean PEG molecular weight and its concentration using MRC Maxi 48 well plates and a drop volume of 2 µL. The crystal used for data collection was ob-

tained in 0.2 M di-ammonium hydrogen phosphate and 18% PEG 3000. Crystals were mounted in CryoLoops and flash-cooled in liquid nitrogen after addition of 30% PEG 400 as cryoprotectant. Diffraction data was collected at 100 K at the BESSY II synchrotron source located in Berlin, Germany. Data processing was done using XDSAPP2,[13] where the data cutoff was chosen based on assessing limits for data completeness (75%), I/σ (0.5), and $CC_{½}$ (30%) in the highest resolution shell. Phases were solved by molecular replacement using *phenix.phaser*[14] and a computed model of the chimera as the search model. Iterative model refinements were done automatically by *phenix.refine*[15] and manually with Coot.[16]

[13] Sparta et al., 2016

[14] McCoy et al., 2007

[15] Afonine et al., 2012

[16] Emsley et al., 2010

Results

Analysis of the Fuzzle database

EARLIER WORK IN OUR GROUP focused on the evolutionary relationships between a small number of folds that hinted at the presence of a diverse set of conserved, sub-domain sized fragments among the known protein universe. These remotely homologous motifs have found their use in protein evolution and design. Thus, a generalization of the search methods used earlier, combining pairwise sequence and structure comparisons, was implemented over the entirety of domain families classified in SCOP.

The resulting all-against-all database, Fuzzle, consists of *conserved fragments* (or *hits*) between query and subject SCOP domains, revolving around sequence (based on HMM-HMM alignments) and structural (relying upon the TM-align algorithm) similarities. It was built using the 2.06 version of SCOP95 and comprises 8 109 195 hits. Out of these, 1 852 531 hits (22.84%) occur between domains belonging to different folds. The most common kind of hits in this *interfold* subset are between domains from the α/β class (806 127 hits, Figure 17). All-α/all-α and all-β/all-β hits are the second and third most frequent class pairing.

The interfold subset contains 27 888 unique fold combinations. It is therefore required to filter hits within this group to find fold pairs with fragments that are appealing for evolutionary characterization and/or combinatorial design. The information included in each hit allows us to determine trends that will narrow our search and guide our decision-making process. As an example, given the different algorithms used by HHsearch and TM-align for their resulting alignments, subtle discrepancies among them are expected in each hit. Indeed, the number of equivalent residues between the sequence align-

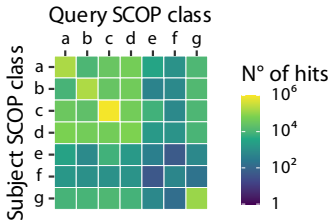


Figure 17: SCOP class distribution in interfold Fuzzle hits. a: all α, b: all β, c: α/β, d: α+β, e: multi-domain proteins, f: membrane and cell surface proteins, g: small proteins.

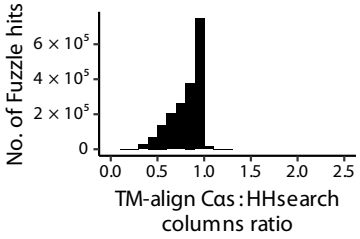


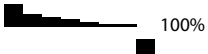
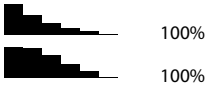
Figure 18: Distribution of Cα:columns ratios in Fuzzle hits. Intervals are left-closed, right-open. Hits with `ca_tm_pair` = 0 were omitted.

ments is usually higher than the one on the structural alignments, with the vast majority of hits having a `ca_tm_pair:cols` ratio close to 1 or below that value (Figure 18, $M = 0.818$, $SD = 0.173$). Likewise, the start and end residues in both kinds of alignments generally vary within a hit. In the interfold subset, 443 313 hits (23.93%) have the same start and end points for their sequence and structural alignments, and 183 453 (9.90%) have the same sequence and structural alignment or, in other words, identical columns in both cases. Regarding fragment length, hits within the interfold subset have a mean of 40.54 residues ($SD = 29.21$). Sequence identities are for the most part low ($M = 19.6\%$, $SD = 9.75\%$), as expected from remotely homologous hits detected with HHsearch.

Unless otherwise stated, "fragment length" will refer exclusively to `cols`.

The 25 most frequent fold pairs in the interfold subset are represented on Table 6. Sixteen pairs occur between α/β folds, with the Rossmann fold (c.2) being present in 10 of them. It is followed by the flavodoxin-like fold (c.23, 4 pairs), TIM barrels, and beta-beta-alpha zinc fingers (c.1 and g.37 respectively, 3 pairs each). Only one pair contains folds between different structural classes: b.1 and c.56 (immunoglobulin-like beta-sandwich and phosphorylase/hydrolase-like folds). Despite the high frequency of hits in the last case, the distribution of probabilities shows that practically all of them have values below 30% ($M = 24.28\%$, $SD = 2.89$).

Table 6: Top 25 fold pairs in the interfold Fuzzle set.

Fold pair	Number of hits	Probab. distrib.[a]	Fold pair	Number of hits	Probab. distrib.[a]
c.2-c.66	112 157 (6.05%)		c.2-c.79	15 024 (0.81%)	
c.2-c.3	110 665 (5.97%)		c.37-c.91	14 751 (0.80%)	
b.1-b.6	55 428 (2.99%)		c.1-c.93	12 988 (0.70%)	
c.2-c.37	39 279 (2.12%)		b.1-c.56	12 700 (0.69%)	
c.1-c.23	37 524 (2.03%)		c.1-c.67	12 580 (0.68%)	
a.4-a.35	36 717 (1.98%)		c.2-c.4	11 084 (0.60%)	
g.37-g.41	22 993 (1.24%)		b.6-b.34	10 262 (0.55%)	
c.2-c.23	22 508 (1.21%)		c.2-c.87	8 659 (0.47%)	
g.37-g.39	21 079 (1.14%)		c.2-c.65	8 547 (0.46%)	
c.2-c.78	19 933 (1.08%)		c.23-c.66	8 482 (0.46%)	
c.2-c.30	18 554 (1.00%)		a.3-a.138	8 110 (0.44%)	
a.4-a.39	16 570 (0.89%)		g.37-g.50	8 010 (0.43%)	
c.23-c.93	15 626 (0.84%)				

[a] Probability distribution ranging from 20 to 100%, in 10% increments.

When looking at hits from the interfold subset with TM-score ≥ 0.3, HHsearch probabilities adopt a bimodal distribution, clustering at values ≤ 30 and ≥ 80 (Figure 19). 206 740 hits (11.1% of the interfold subset) are located in the region of TM-score ≥ 0.3 and probability ≥ 80%. Hits between different folds with high probabilities and TM-scores can be considered remotely homologous and are therefore of interest for further evolutionary studies as well as design purposes. It is essential, then, to set proper cutoffs and refine the search to fetch relevant fragments within this interfold subset of hits.

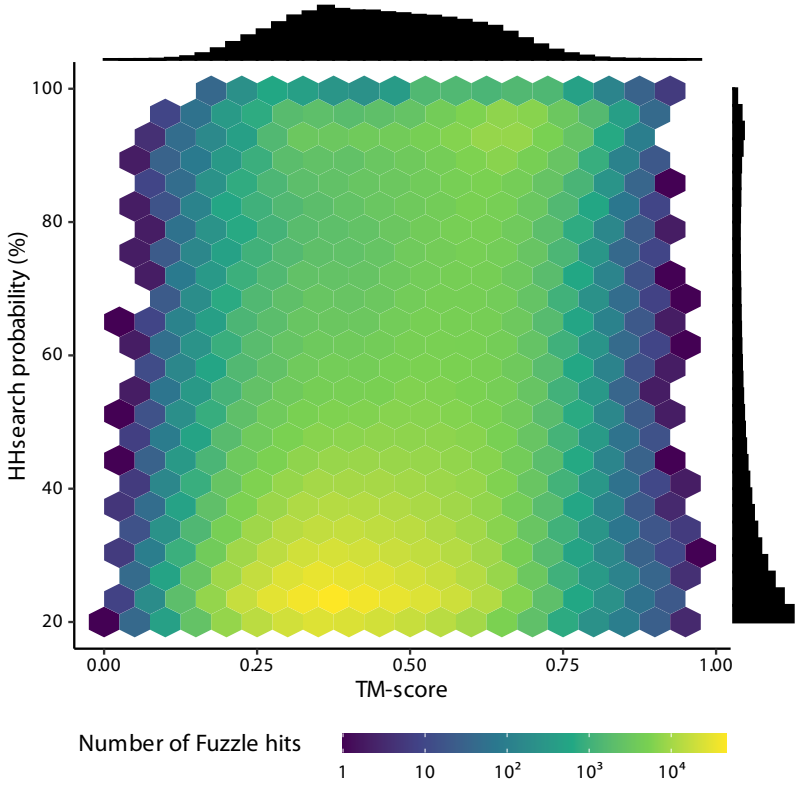


Figure 19: Hexagonal heatmap of TM-score and HHsearch probabilities in the interfold subset of Fuzzle. Hits with TM-score = 0 and/or probability ≤ 20% were filtered out. Marginal distributions for both variables are included as well.

Based on the previous observations, the aforementioned interfold dataset was further filtered with the following parameters: `prob` >= 50.0, `score_tm_pair` >= 0.3, `cols` >= 30, and `ca_tm_pair` >= 30.

A 30-residue cutoff will increase the likelihood of finding more than one secondary structure element within the hit.

This *filtered* dataset consists of 348 893 hits, out of which 183 322 are bidirectional, i.e., when a query-subject domain pair has a corresponding hit with the roles reversed. After filtering, the resulting set has a mean fragment length of 66.21 residues (*SD* = 36.75) and most fragments concentrate at either

35 or 70 residues approximately (Figure 20). This implies that a considerable amount of these hits have several secondary structure elements, covering a significant portion of the domains involved.

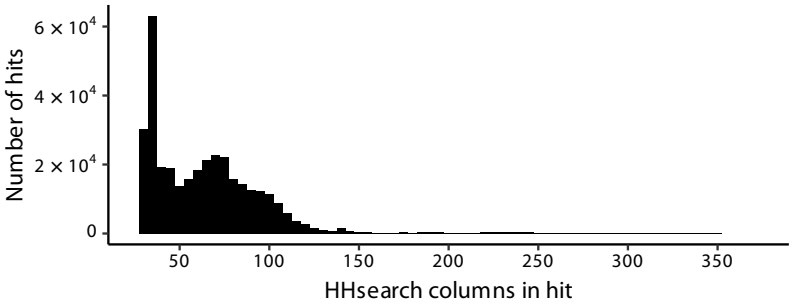


Figure 20: Fragment size distribution in the filtered interfold subset of Fuzzle.

The number of unique fold pairs in the filtered subset decreased to 2 604, a 90% reduction. There is further enrichment of α/β-α/β hits in the filtered dataset (Figure 21), suggesting that a large proportion of conserved fragments with high probability, high TM-score, and length ≥ 30 residues have a combination of alternating βα elements, as opposed to folds with purely α-helical or β-stranded motifs.

The previous statement is confirmed by looking at the top 25 fold pairs in the filtered interfold subset (Table 7). In this subset, the Rossmann fold is still the most prevalent one (Figure 22), appearing on 14 pairs, followed by the flavodoxin-like fold and the periplasmic binding protein-like I (PBP-like I, c.93), with 3 pairs each. Fragments outside α/β ones in this list are between differently-sized β-propellers (b.68, b.69, b.70), fragments between DNA/RNA-binding 3-helical bundles, and lambda repressor-like DNA-binding domains (a.4–a.35), a.4 with the EF hand-like (a.39) and a.39 with the type I dockerin domain folds (a.139).

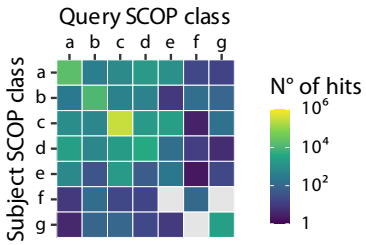
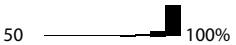
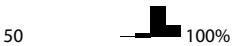
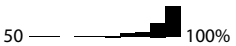
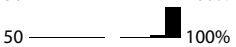
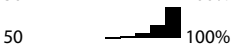


Figure 21: SCOP class distribution in the filtered interfold subset of Fuzzle. Gray cells represent zero hits.

Table 7: Top 25 fold pairs in the filtered interfold Fuzzle set.

Fold pair	Number of hits	Probab. distrib.[a]	Fold pair	Number of hits	Probab. distrib.[a]
c.2-c.3	71 461 (20.48%)		a.4-a.35	3 677 (1.05%)	
c.2-c.66	62 278 (17.85%)		b.68-b.69	3 582 (1.03%)	
c.1-c.23	18 335 (5.26%)		a.4-a.39	3 319 (0.95%)	
c.2-c.30	12 540 (3.59%)		a.39-a.139	3 158 (0.91%)	
c.2-c.78	11 928 (3.42%)		c.2-c.65	3 063 (0.88%)	
c.2-c.37	8 165 (2.34%)		c.2-c.72	2 686 (0.77%)	
c.2-c.4	7 949 (2.28%)		b.69-b.70	2 592 (0.74%)	
c.23-c.93	7 212 (2.07%)		c.16-c.93	2 260 (0.65%)	
c.2-c.5	4 883 (1.40%)		c.2-c.32	2 030 (0.58%)	
c.2-c.23	4 622 (1.32%)		c.3-c.4	1 950 (0.56%)	
c.2-c.111	4 387 (1.26%)		c.2-c.87	1 678 (0.48%)	
c.2-c.79	4 348 (1.25%)		c.55-c.95	1 493 (0.43%)	
c.1-c.93	3 755 (1.08%)				

[a]Probability distribution ranging from 50 to 100%, in 5% increments.

More that two thirds of the fragments found in the filtered subset belong to one of the following seven folds (Figure 22): NAD(P)-binding Rossmann-fold domains (c.2), FAD/NAD(P)-binding domain (c.3), S-adenosyl-L-methionine-dependent methyltransferases (c.66), flavodoxin-like (c.23), TIM beta/alpha-barrel (c.1), periplasmic binding protein-like I (c.93), and preATP-grasp domain (c.30).

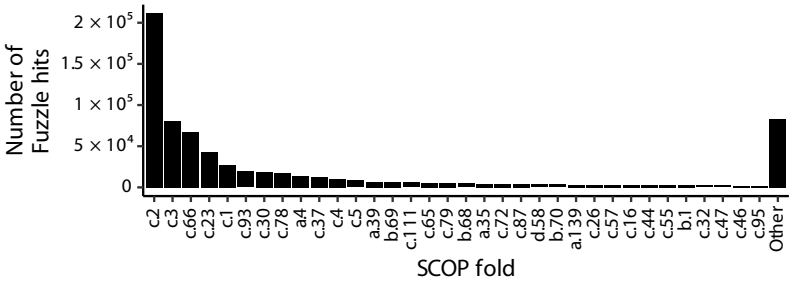


Figure 22: Fold frequency of Fuzzle hits in the filtered interfold dataset. All folds mentioned in Table 7 are present.

Most domains in the filtered interfold subset have few hits (M = 38.6, SD = 93.6, Mo = 1, Mdn = 5) and hit few folds as well (M = 3.25, SD = 4.01, Mo = 1, Mdn = 1). However, excluding the TIM barrel fold, all folds mentioned above have a high number of hits per domain (Figure 23A) and hit several different folds (Figure 23B).

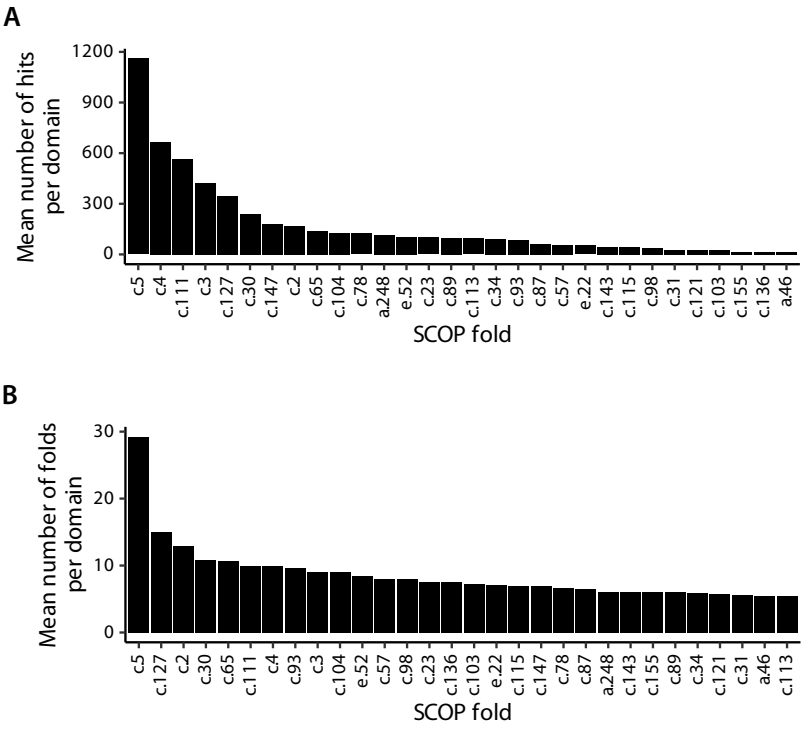


Figure 23: Folds with high average number of hits per domain (A) and high average number of hit folds per domain (B).

Fragments involving c.1, c.23 and c.93 domains have been already characterized by our group. Thus I revisited and characterized them carefully in order to gain new insights regarding their evolutionary connection.

Hits involving TIM barrel domains

IT HAS BEEN PROPOSED that domains arose through duplication of smaller, ancestral protein fragments. One of the best studied examples of that phenomenon is the TIM barrel (SCOP fold c.1). The TIM barrel fold is considered one of the most ancient ones, catalyzing a wide diversity of chemical reactions. Given the two-fold symmetry in some TIM barrel domain structures, it has been proposed that its $(\beta\alpha)_8$ topology was a result of the duplication and subsequent fusion of two ancestral $(\beta\alpha)_4$ halves or four $(\beta\alpha)_2$ quarters. These fragments, in turn, present common ancestry with at least one other fold, the flavodoxin-like one. For this reason hits within the TIM barrel fold as well as its relationships with other different folds will we analyzed next.

Intrafold TIM barrel fragments

In Fuzzle, $(\beta\alpha)_8$ barrel domains can show intrafold hits despite having very low sequence identity (M = 16.9%, SD = 9.57). Excluding cases where the starting residue lies in the same βα element for both query and subject domains, the fragment starting positions are mostly offset by one or two $(\beta\alpha)_2$ elements or, in other words, by a quarter or half barrel (Figure 24A). Contrary to expectations, and with the exception of the first β-strand, most intrafold barrel fragments start with an α-helix (Figure 24B). Within these sort of fragments, the most

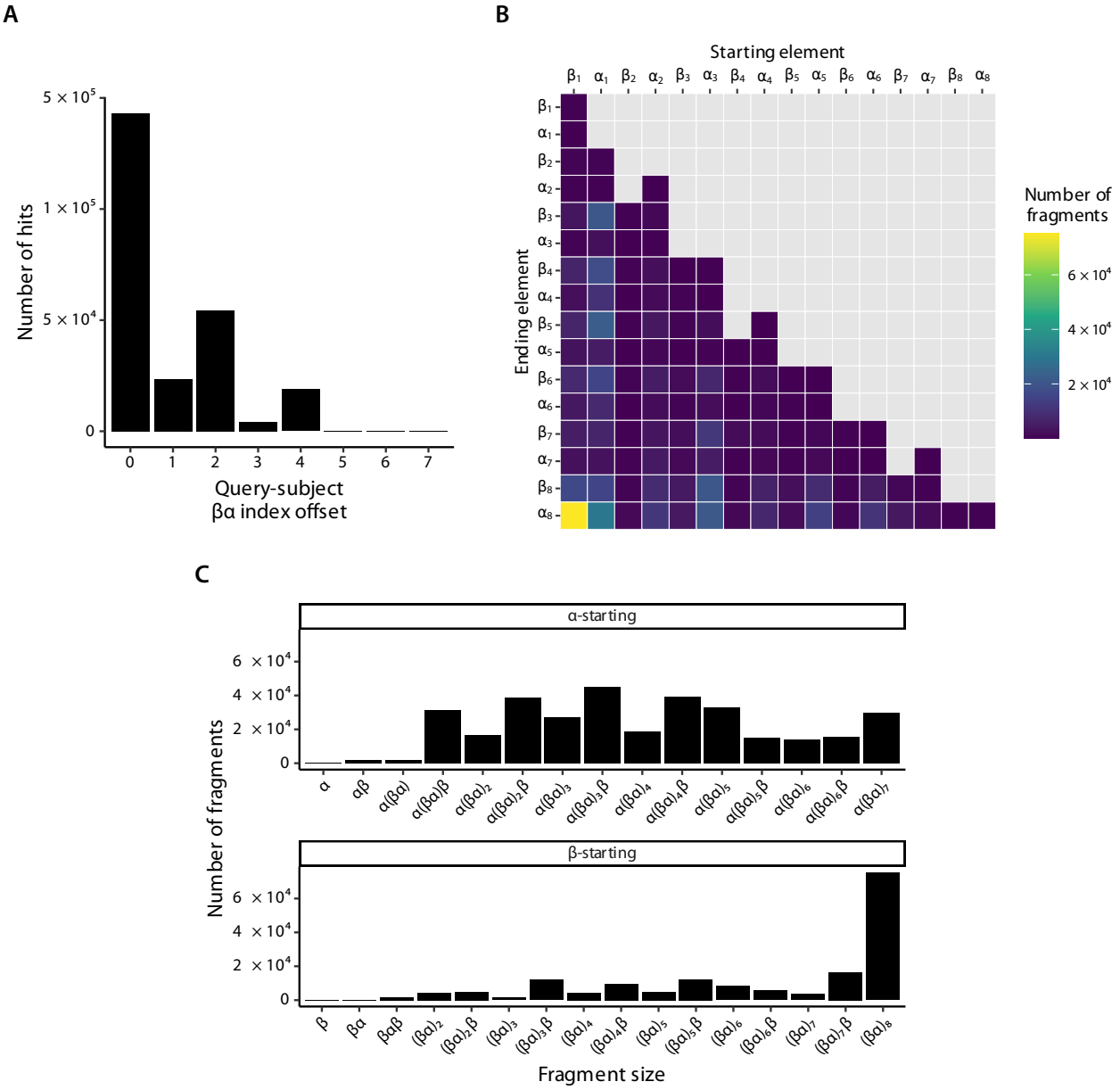


Figure 24: Starting and ending α/β elements in intrafold TIM barrel hits. Hits were filtered for probability ≥ 50%. A: Absolute differences in the location of fragment-starting βα elements. B: Starting and ending α/β elements in intrafold TIM barrel hits. Counts equal to zero are shown in gray. C: Fragment size distribution.

frequent ones are: $\alpha(\beta\alpha)\beta$, $\alpha(\beta\alpha)_2\beta$, $\alpha(\beta\alpha)_3\beta$, $\alpha(\beta\alpha)_4\beta$, $\alpha(\beta\alpha)_5$, and $\alpha(\beta\alpha)_7$. That is, fragments starting with an α-helix have a beta-alpha element flanked by either a helix or strand on each terminus. In the case of strand-starting fragments, the size trends are different, with $(\beta\alpha)_3\beta$, $(\beta\alpha)_4\beta$, $(\beta\alpha)_5\beta$, $(\beta\alpha)_7\beta$, and $(\beta\alpha)_8$ being the most common lengths.

A particular case of fragments are the ones where one domain hits multiple times with itself. A hit covering the full length of the domain is obviously expected, but in some domains one additional, smaller fragment is found. This is consistent with the view of a duplication-based evolutionary development of the fold. 176 folds in Fuzzle have at least one domain hitting more than once with itself (Figure 25).

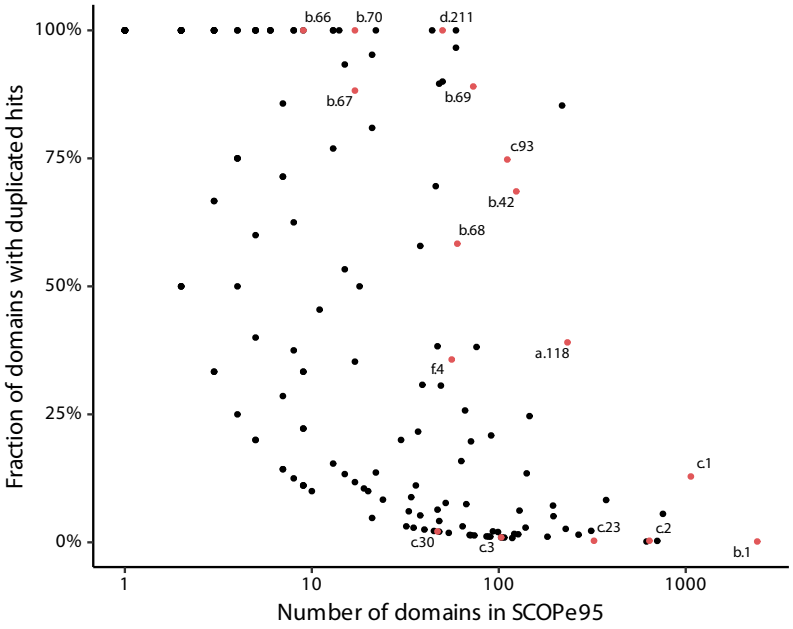


Figure 25: Folds with duplicated hits in Fuzzle. Some previously mentioned folds are highlighted in red. a.118: α-α superhelix (armadillo, HEAT, TPR); b.1: Immunoglobulin-like β-sandwich; b.42: β-trefoil; b.66-b.70: β-propellers; d.211: β-hairpin-α-hairpin (ankyrin) repeat; f.4: Transmembrane β-barrels.

Interestingly, roughly 13% of TIM barrel domains contain duplicated hits with themselves, a figure higher than other similarly populated folds. The size of these interbarrel fragments depends on the starting element (Figure 26A). Helix-starting fragments are smaller, with $\alpha(\beta\alpha)_2\beta$, $\alpha(\beta\alpha)_3$ and $\alpha(\beta\alpha)_3\beta$ being the most common ones. In the strand-starting group there is a shift towards larger sizes, particularly $(\beta\alpha)_7$, $(\beta\alpha)_7\beta$, and, as expected, full $(\beta\alpha)_8$ barrels. Overall, intradomain fragments concentrate around half- or full barrels. Isolated quarter fragments, i.e $(\beta\alpha)_2$ ones, are absent from this set of hits. The smallest fragment found here is an $\alpha(\beta\alpha)\beta$ one, that is, a single βα element

In other words, $(\alpha\beta)_3$, $(\alpha\beta)_3\alpha$, and $(\alpha\beta)_4$

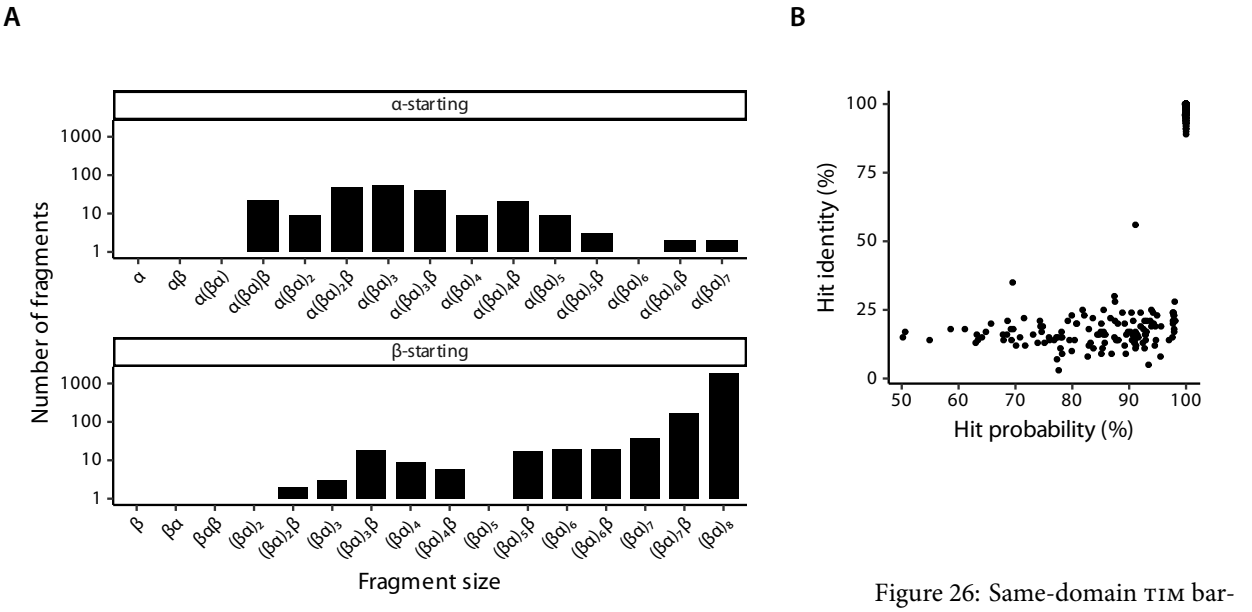
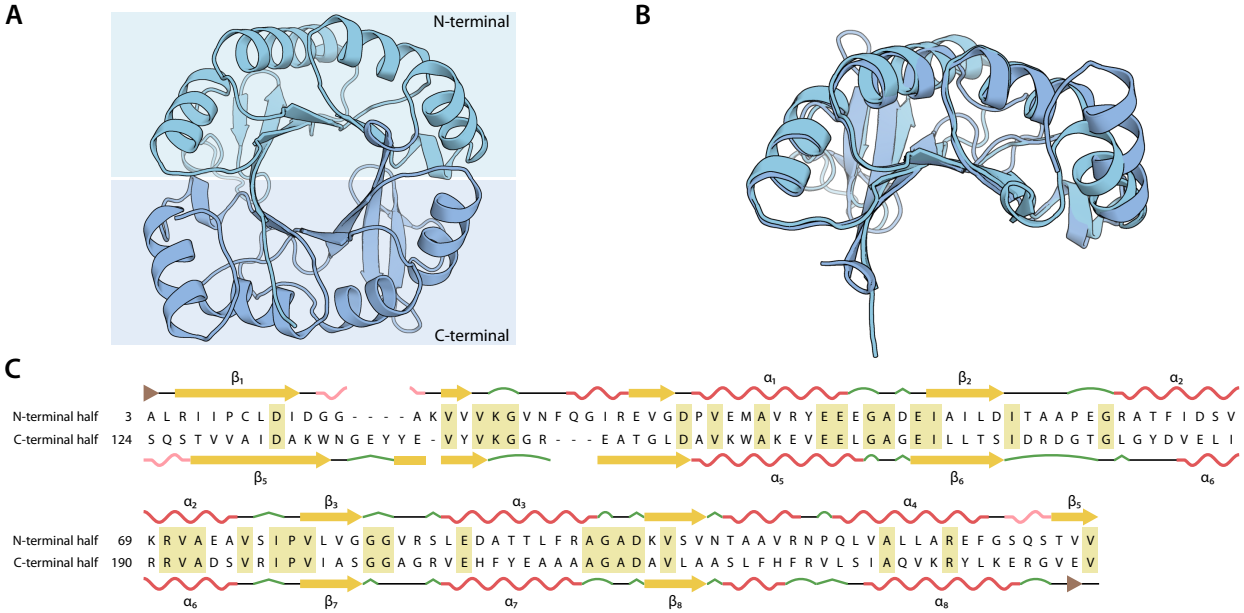


Figure 26: Same-domain TIM barrel Fuzzle hits. A: Size distribution. B: Relationship between HHsearch probability and sequence identity.

flanked by a helix and a strand from different adjacent quarters. Besides, the smallest β-starting fragment detected was $(\beta\alpha)_2\beta$, two quarters extended by an additional strand.

Identity does not correlate with HHsearch probability in these intrabarrel fragments. In fact, if hits with probability 99.9% or above are excluded, hits within the same TIM barrel domain have a mean identity of 18.4% (Figure 26B). Figure 27 shows an example half barrel hit. Despite both halves aligning perfectly with each other, the sequence identity between them is only 21%.

Figure 27: Half-barrel sized hits in TIM barrel domains. A: Structure of *Pyrobaculum aerophilum* HisF (SCOP ID d1h5ya_) and its two halves. B: Structural superposition of both HisF halves as found in the Fuzzle hit. C: HHsearch sequence alignment of the hit.



Interfold TIM barrel hits

Within Fuzzle, TIM barrel domains have hits mostly with the flavodoxin-like (c.23), the PBP-like I (c.93), and the composite domain of metallo-dependent hydrolases (b.92) folds (Figure 28). This last relationship is of interest because the TIM barrel domains involved in it are from the metallo-dependent hydrolase (c.1.9) superfamily. Metallo-dependent hydrolases have two domains, a catalytic TIM barrel domain and a composite β-sandwich one, with the former interrupting the latter. Even though the hydrolase and composite domains have Fuzzle hits with high probability (M = 72.6%, SD = 23.8), their TM-score is low (M = 0.230, SD = 0.078) due to their different structural classes. Some noteworthy hits within this group are between c.1 and b.92 domains situated in the same PDB chain. Figure 29 shows one of such examples from human guanine deaminase. Despite having a low identity (13%) and 4.5 Å RMSD over 46 residues (out of 63 aligned HHsearch columns), this hit has a 92.3% HHsearch probability, suggesting a common ancestral fragment spanning regions of both folds.

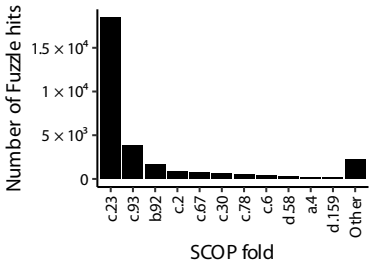


Figure 28: Fold distribution in interfold TIM barrel hits. Hits with probability < 50% were filtered out.

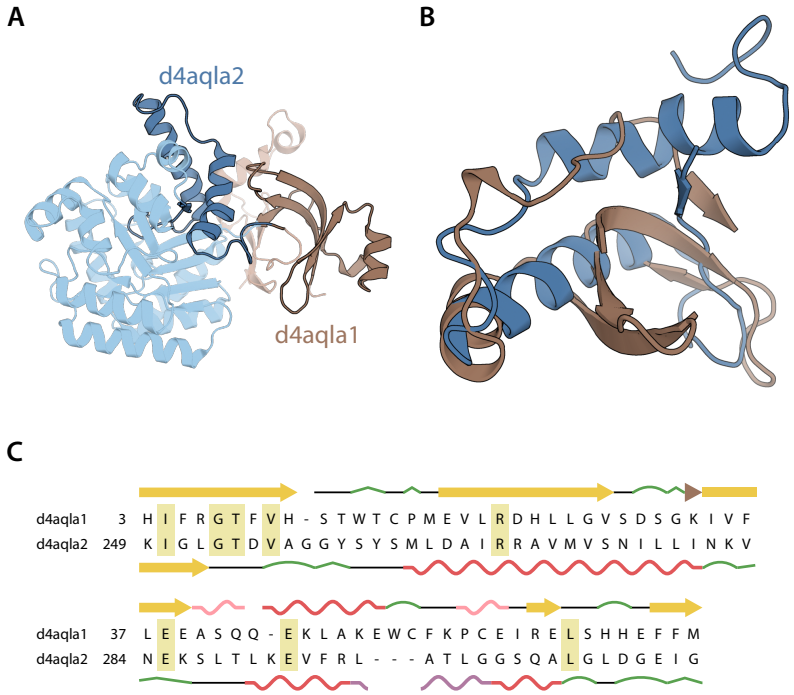


Figure 29: Conserved fragment between the c.1 and b.92 folds. A: Structure of guanine deaminase from *Homo sapiens* (PDB ID 4AQL) with its TIM barrel (SCOP ID d4aqla2, blue) and composite β-sandwich (SCOP ID d4aqla1, brown) domains. The shared fragment is highlighted in darker shades. B: TM-align superposition of the conserved fragment. C: HHsearch alignment of the hit.

TIM barrel and Fld-like domain fragments

DISTANT EVOLUTIONARY RELATIONSHIPS have been previously assessed between TIM barrel (c.1) and flavodoxin-like (c.23) domains, where a $(\beta\alpha)_3$ fragment is shared between both folds. Fuzzle has hits that confirm this link. The distribution of them among different superfamilies is uneven, with the CheY-like and cobalamin (vitamin B_{12})-binding domain ones (c.23.1 and c.23.6 respectively) linking to a wide range of different c.1 superfamilies (Figure 30). Hits from the FMN-linked oxidoreductases (c.1.4), inosine monophosphate dehydrogenase (c.1.5), (trans)glycosidases (c.1.8), aldolase (c.1.10), enolase C-terminal domain-like (c.1.11), phosphoenolpyruvate/pyruvate domain (c.1.12), and nicotinate/quinolinate PRTase C-terminal domain-like superfamilies are likewise spread over three or more Fld-like ones.

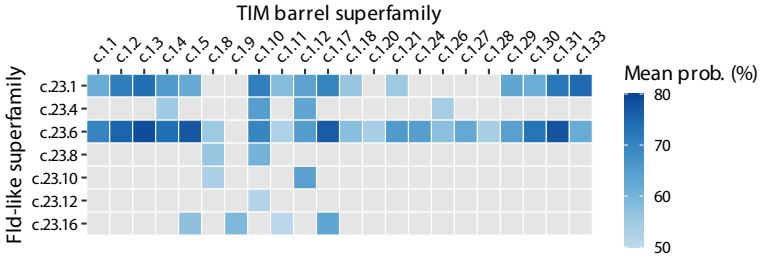


Figure 30: TIM barrel/Fld-like superfamilies in filtered subset hits. Superfamiliy pairs without hits are shown in gray.

In these hits the TIM barrel fragments cover mostly from the third to the fifth, or alternatively, from the fourth to the seventh βα element. On the other hand, the majority of Fld-like fragments start at the first or second βα element and end at the fourth one (Figure 31).

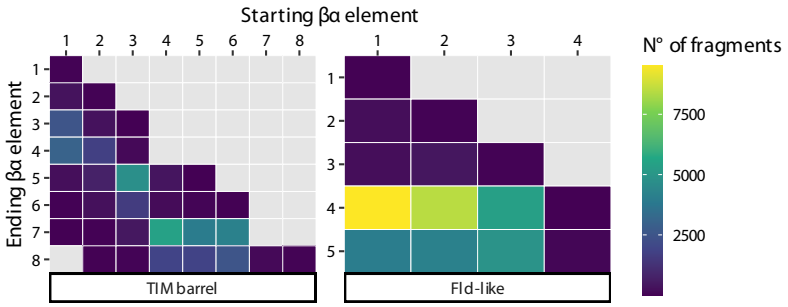


Figure 31: Starting and ending βα elements in TIM barrel/Fld-like hits.)

In the filtered dataset, the range of aligned HHsearch columns in hits between c.1 and c.23 is wide, with three max-

ima at around 40, 70, and 90 residues. However, in the same set of hits, the number of alpha carbons aligned by TM-align shows only the first two maxima, omitting the largest one at 90 columns (Figure 32A).

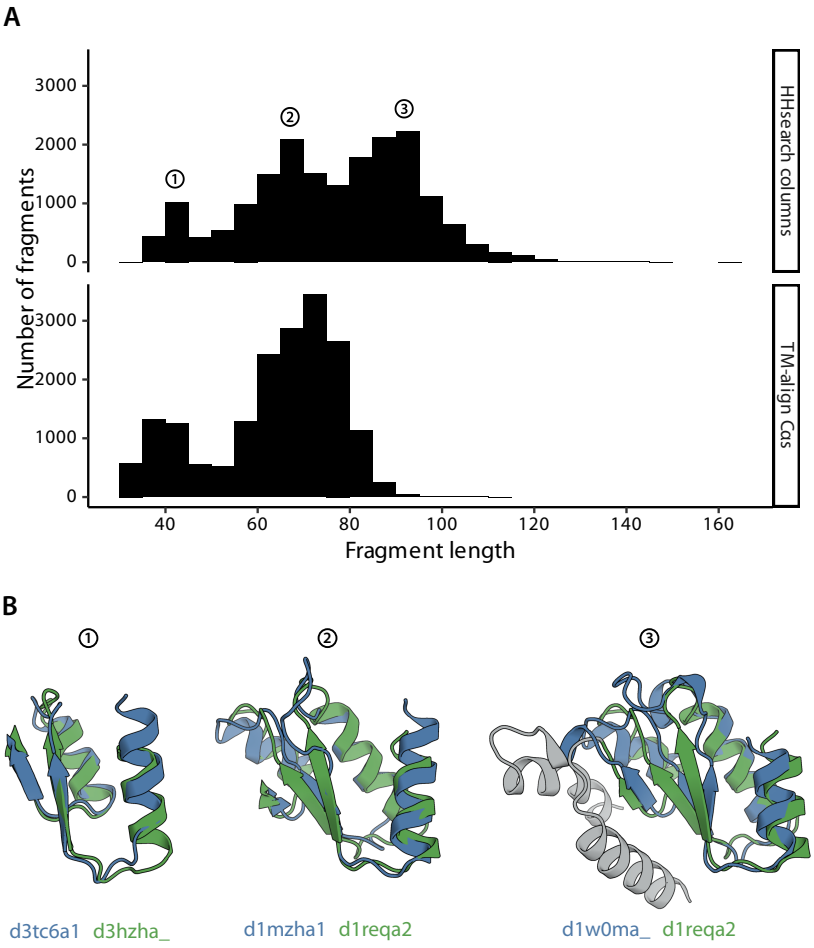


Figure 32: HHsearch columns and TM-align Cαs distributions in c.1-c.23 hits. A: Differences between both distributions. B: Representative c.1-c.23 hits from each maximum labeled in A.

A closer look at the structural superpositions of larger fragments shows mismatches on the secondary structure elements at the N-terminal end of both domains (Figure 32B). This is caused by the different β-strand order in both folds: TIM barrels have a sequential N- to C-terminal strand order (12345678) whereas in flavodoxin-like domains it is discontinuous (21345). Terminal helices can be misaligned as well, due to the different orientations that they adopt in both folds to provide proper core shielding. Helices in TIM barrels are always on the outer face of the β-strand, while in flavodoxin-like proteins the terminal and middle helices are positioned on opposite sides, forming their well-known sandwich-like architecture.

Interestingly, the database has a number of duplicated domain hits between c.1 and c.23 folds, wherein a flavodoxin-like fold domain shares a fragment with two different regions of a TIM barrel domain. Said hit pairs can have overlapping regions, but in some particular cases they do not, aligning with the N- and C-terminal halves of the TIM barrel domain (Figure 33).

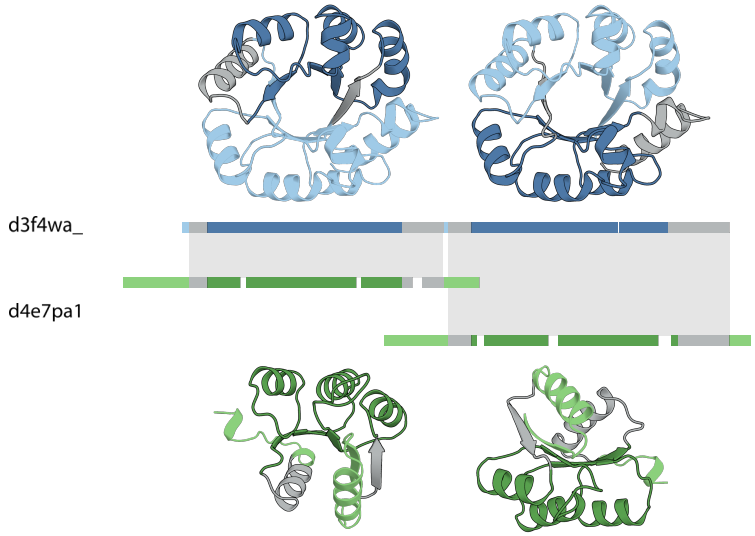


Figure 33: Duplicated half barrel sized hits between TIM barrel and Fld-like domains. d3f4wa_: 3-hexulose-6-phosphate synthase from *Salmonella typhimurium*. d4e7pa1: *Streptococcus pneumoniae* response regulator spr1814. Regions not present in the Fuzzle hit are colored in lighter shades. Structurally unaligned regions are colored gray. Residues aligning in both sequence and structure are highlighted in darker shades.

Ligand interactions in TIM barrel/Fld-like hits

The TIM barrel fold is an extreme example of functional diversity, with five out of seven EC (Enzyme Commission) classes having TIM barrel proteins.[1] They share, however, some conserved functional properties. Many TIM barrel domains bind substrates or cofactors with a phosphate group, with residues in loops 7 and 8 as well as an α-helix within loop 8 involved in this interaction. The specific type of cofactor or substrate associated to the domain is dependent on the latter's superfamily and includes alkyl phosphates, nucleotides such as IMP, NAD, or FMN, DHAP, and thiamin phosphate.

[1] Nagano et al., 1999

Likewise, different Fld-like superfamilies bind different ligands. Most Fuzzle hits between TIM barrel and Fld-like domains include a member of the CheY-like (c.23.1) cobalamin-binding domain (c.23.6) superfamilies. Cobalamin-binding domains use this coenzyme to catalyze mainly, depending on the enzyme family, methyl transfer or radical-mediated isomerization reactions. In contrast, CheY-like domains act as response regulators in bacterial two-component signaling

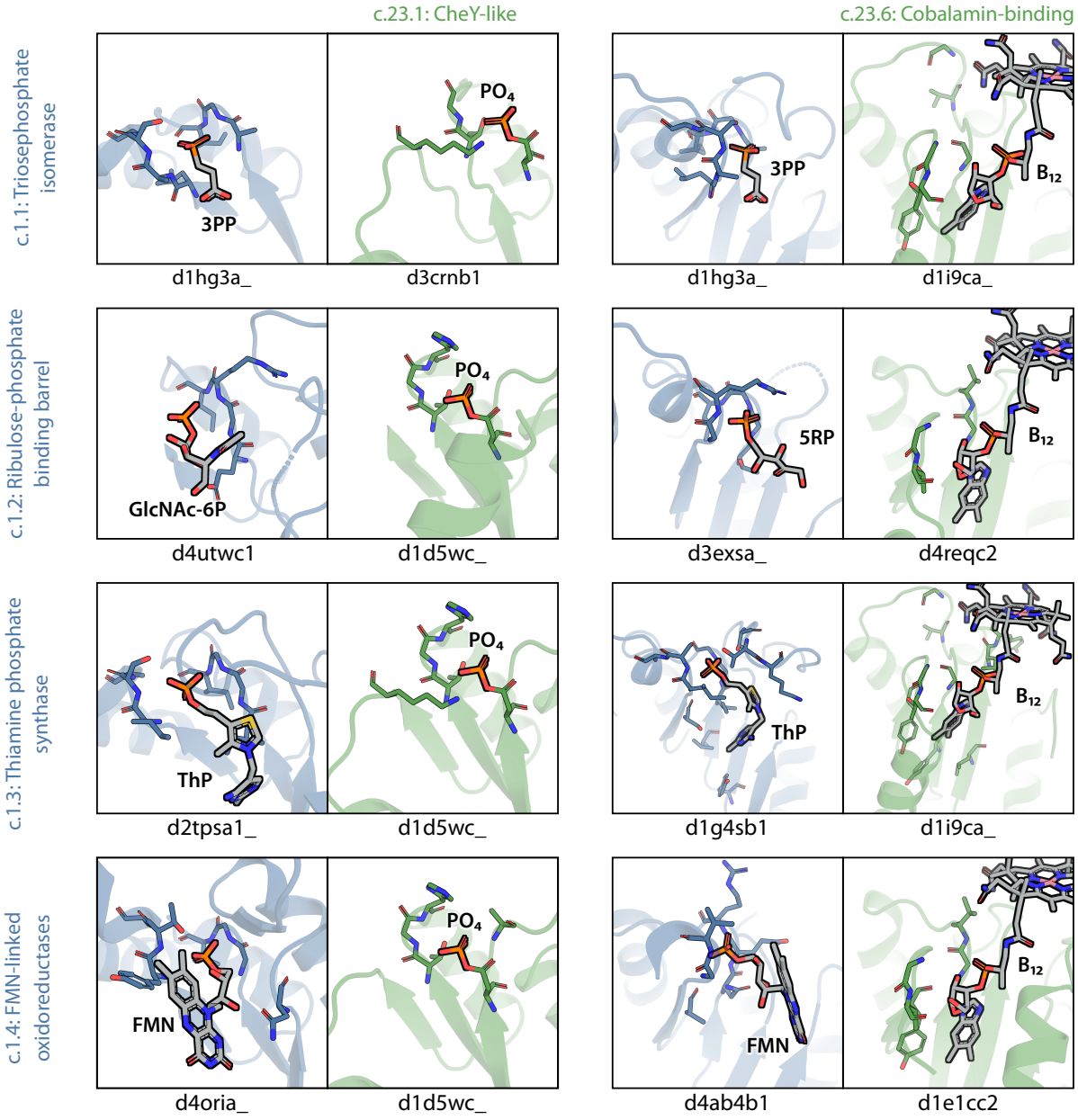


Figure 34: Conserved ligand-binding sites in TIM barrel and Fld-like domains. Ligands are gray, while conserved TIM barrel and Fld-like residues are shown as blue and green sticks respectively. 3PP: 3-phosphonopropanoic acid. GlnNAc-6P: N-acetyl-D-glucosamine-6-phosphate. 5RP: ribulose-5-phosphate. THP: thiamin phosphate.
(Continued on next page.)

systems, where they receive the signal from a sensor partner, usually a histidine protein kinase. CheY-like domains do not bind any ligand molecule. Instead, they share a conserved aspartic acid residue at the end of the third β-strand, whose phosphorylation is crucial for response activation and modulation. Despite these functional differences between CheY-like and cobalamin-binding superfamilies, several TIM barrel domain superfamilies hit both Fld-like ones in Fuzzle and present conserved ligand-interacting residues in both contexts (Figure 34).

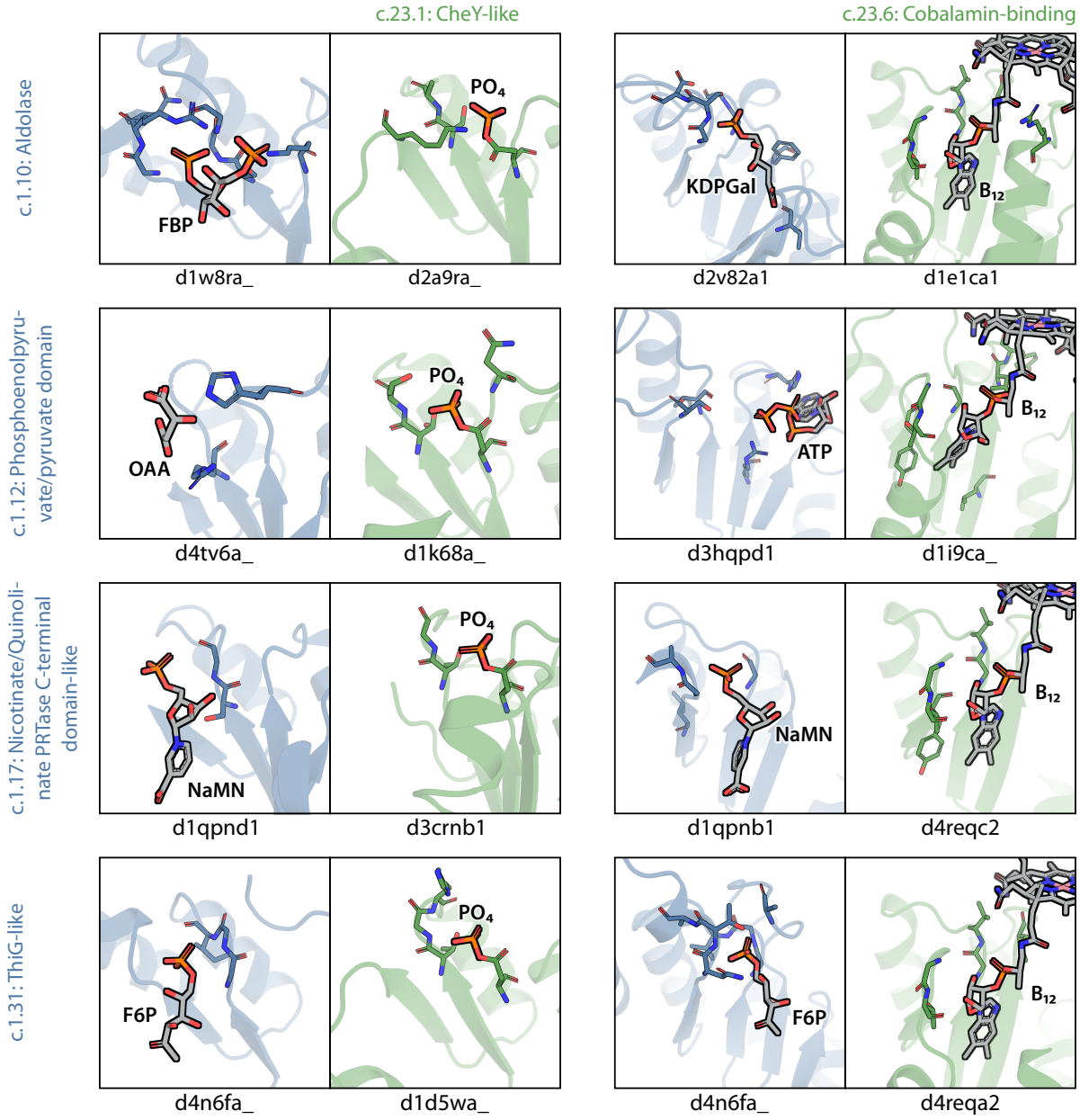


Figure 34: (cont.)
FBP: fructose 1,6-bisphosphate. KDPGal: 2-keto-3-deoxy-6-phosphogalactonate. OAA: oxaloacetate. NaMN: nicotinamide mononucleotide. F6P: fructose-6-phosphate.

Hits from the cobalamin-binding superfamily (SCOP ID c.23.6), have vitamin B_{12}-interacting residues in their loops, overlapping with the previously mentioned phosphate-binding motif found in TIM barrels. Figure 35 shows a Fuzzle hit comprised by a uridine monophosphate (UMP) binding TIM barrel and a vitamin B_{12}-binding Fld-like domain. Despite being similarly located, cobalamin-binding residues interact mostly with the ribose group of vitamin B_{12} and, to a lesser extent, the phosphate, benzimidazolyl, and corrin side groups. In contrast, the TIM barrel residues interact mainly with phosphate groups,

with other accessory moieties having less interactions.

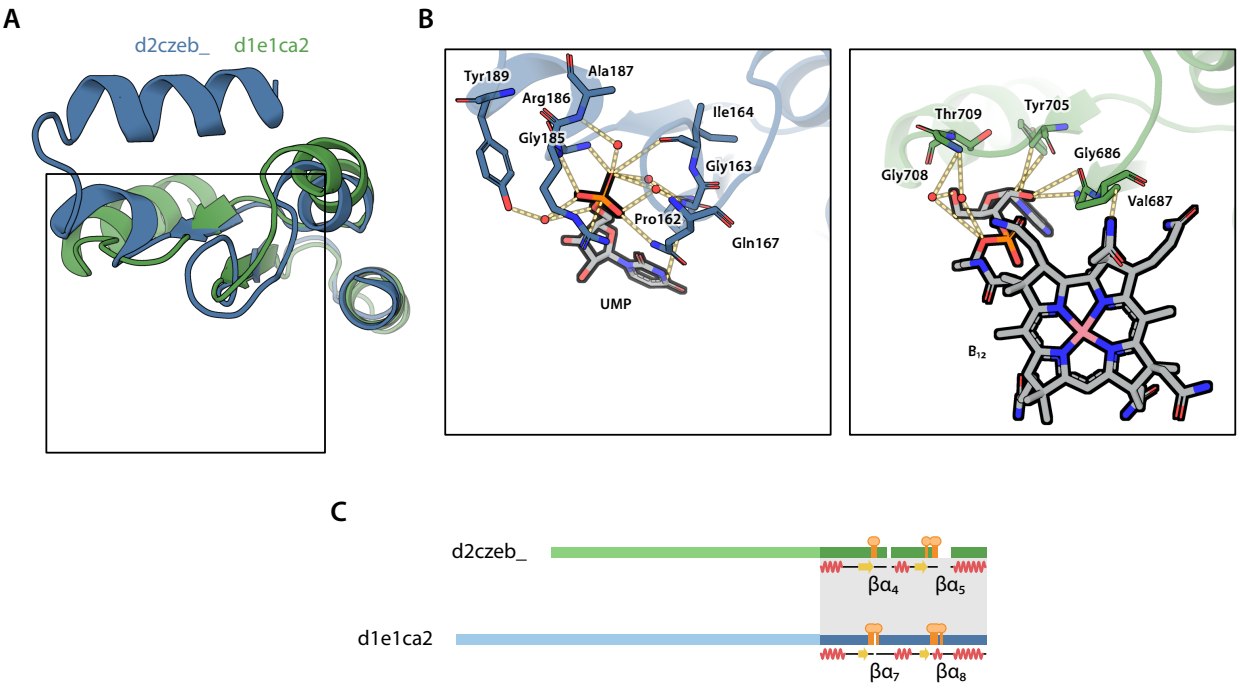


Figure 35: Conservation of ligand-interacting residues in TIM barrel/cobalamin-binding fragments. A: Fuzzle hit between *Pyrococcus horikoshii* orotidine 5′-phosphate decarboxylase (SCOP ID d2czeb_, blue) and *Propionibacterium freudenreichii* methylmalonil-CoA mutase (SCOP ID d1e1ca2, green). B: Detail of conserved protein-ligand interactions present in the fragment. C: Allocation of conserved protein-ligand interactions at the sequence level. Ligand-interacting residues are highlighted in orange. For the sake of simplicity, only helices and strands are shown in the secondary structure representation.

A more direct phosphate-phosphate relationship ocurrs between TIM barrel domains and domains from the CheY-like superfamily (SCOP id c.23.1). Some Fuzzle hits have CheY-like domains associated with FMN-binding TIM barrel domains. In their structural superposition the phosphate moieties of phosphoaspartate (pAsp) and FMN are located close to each other. Furthermore, some phosphate-interacting residues in the Fld-like domain overlap with FMN-binding ones. The latter are located mostly around the phosphate group of FMN, either forming hydrogen bonds or water bridges with it (Figure 36). The regions where the overlapping residues correspond to loops 3-5 in CheY-like and loops 6-8 in TIM barrels. The last loop of each group is coincidentally where the structural superposition of both folds diverge. In the sequence alignment, however, the location of the interacting residues is close in both groups. These conserved phosphate-binding residues in both folds hint at functional conservation derived from an ancient fragment shared between them and serve as additional evidence for the evolutionary relationship between these ancestral folds.

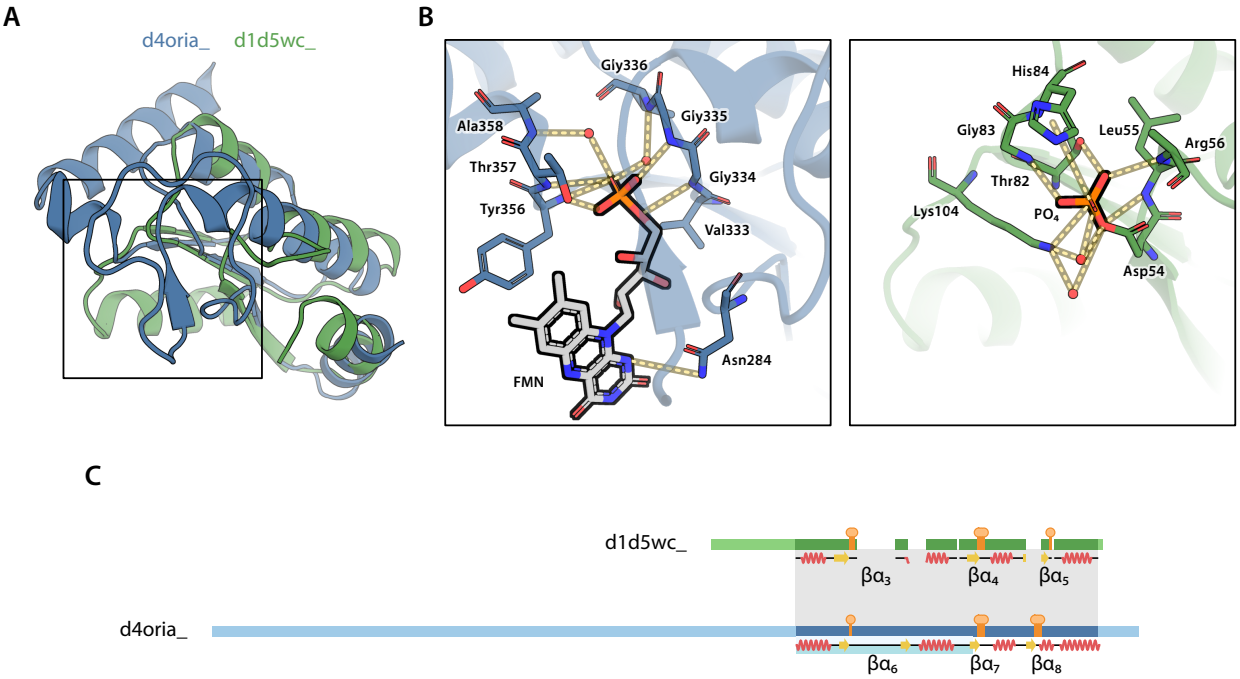


Figure 36: Conservation of phosphate-binding sites in TIM barrel/CheY-like fragments. A: Fuzzle hit between *Rattus norvegicus* dihydroorotate dehydrogenase (SCOP ID d4oria_, blue) and *Rhizobium meliloti* transcriptional regulatory protein FixJ (SCOP ID d1d5wc_, green). B: Detail of conserved protein-ligand interactions present in the fragment. C: Allocation of conserved protein-ligand interactions at the sequence level. The same coloring as in Figure 35 is used.

Apart from c.23.1 and c.23.6, some additional relationships with other Fld-like superfamilies are present. For instance, inosine monophosphate dehydrogenase (IMPDH, SCOP superfamily c.1.5) contains additional conserved fragments with another Fld-like superfamily, class I glutamine amidotransferase-like (GATase-like, c.23.16). Specifically, hits involve the small subunit of carbamoyl phosphate synthetase (CPSase), which catalyzes the hydrolysis of glutamine to ammonia, In this case conserved interactions with glutamine are present (Figure 37), including a catalytic cysteine essential for the amidotransferase activity. The corresponding residues in IMPDH bind the ribose phosphate group of the substrate.

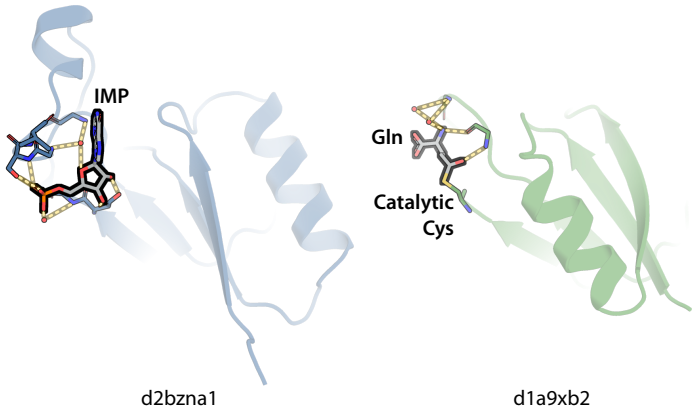


Figure 37: Conserved ligand-interacting residues between IMPDH and GATase domains. Fuzzle hit between *Homo sapiens* GMP reductase 2 (SCOP ID d2bzna1) and *Escherichia coli* CPSase (SCOP ID d1a9xb2).

c.1.12 (Phosphoenolpyruvate/pyruvate domain) and c.23.10 (SGNH hydrolase) both contain domains with a conserved residue located in the conserved fragment (Figure 38). In c.1.12 an aspartate located in a loop helix after the sixth beta-strand, part of a well known motif in PEP-utilizing enzymes, coordinates a Mg^{2+} ion. The Fld-like hydrolase superfamily, on the other hand, has an asparagine residue in an equivalent helix located between strand and helix 3 serving as a proton donor in the oxyanion hole of the enzyme. This asparagine is part of the SGNH motif and thus fully conserved within the members of this superfamily.

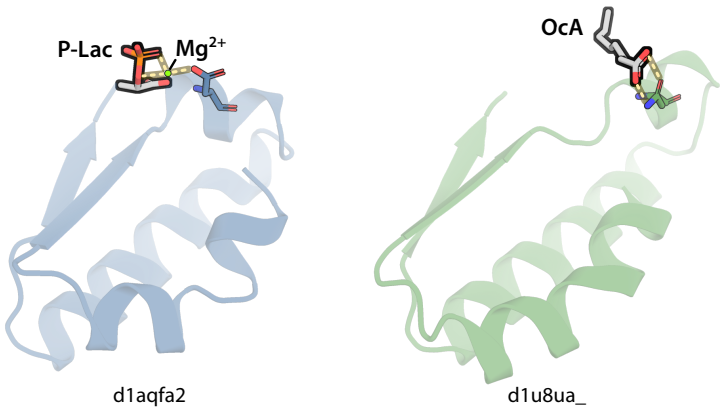


Figure 38: Conserved ligand-interacting residues between c.1.12 and SGNH hydrolases. Fuzzle hit between *Oryctolagus cuniculus* pyruvate kinase (SCOP ID d1aqfa2) and *Escherichia coli* thioesterase 1/protease 1/lysophospholipase L1 (SCOP ID d1u8ua_) P-Lac: L-phospholactate, OcA: octanoic acid.

Among the superfamily combinations found to have conserved interactions, one is unexpectedly missing: FMN-linked oxidoreductases (c.1.4) and flavoproteins (c.23.5). Since both superfamilies bind and have FMN-dependent activity, an evolutionary link is expected, yet no hits between them were found in the whole Fuzzle dataset. A closer inspection of both folds show different FMN-binding pockets in them and different orientations of the ligand (Figure 39), suggesting separate emergence events.

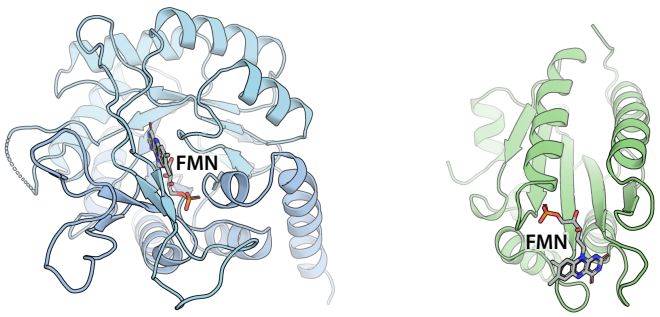


Figure 39: FMN-binding in TIM barrel (SCOP ID d1gtea2, left) and Fld-like (SCOP ID d3qe2a1, right) domains.

Duplication of ligand interaction motifs in TIM barrel/Fld-like fragments

IF THE TIM BARREL FOLD was a result of fragment duplication, then functional features already present in the ancestral state should have been propagated as well and lay symmetrically in both halves. An example of this are the HisA/PriA enzymes from the ribulose-phosphate binding barrel superfamily (c.1.2), which adopt a TIM barrel fold and catalyze the isomerization of *N*′-[(5′-phosphoribosyl)formimino]-5-aminoimidazole-4-carboxamide ribonucleotide (ProFAR). ProFAR has two phosphate groups, one at each end, which interact at identical places in each half barrel, namely loops 3-4 and 7-8 respectively. Members from this family have hits with CheY-like as well as cobalamin-binding members. Moreover, some domain pairs have duplicated hits, where the phosphate-interacting residues in the CheY-like domain overlap with each phosphate-binding site in both TIM barrel domain halves (Figure 40).

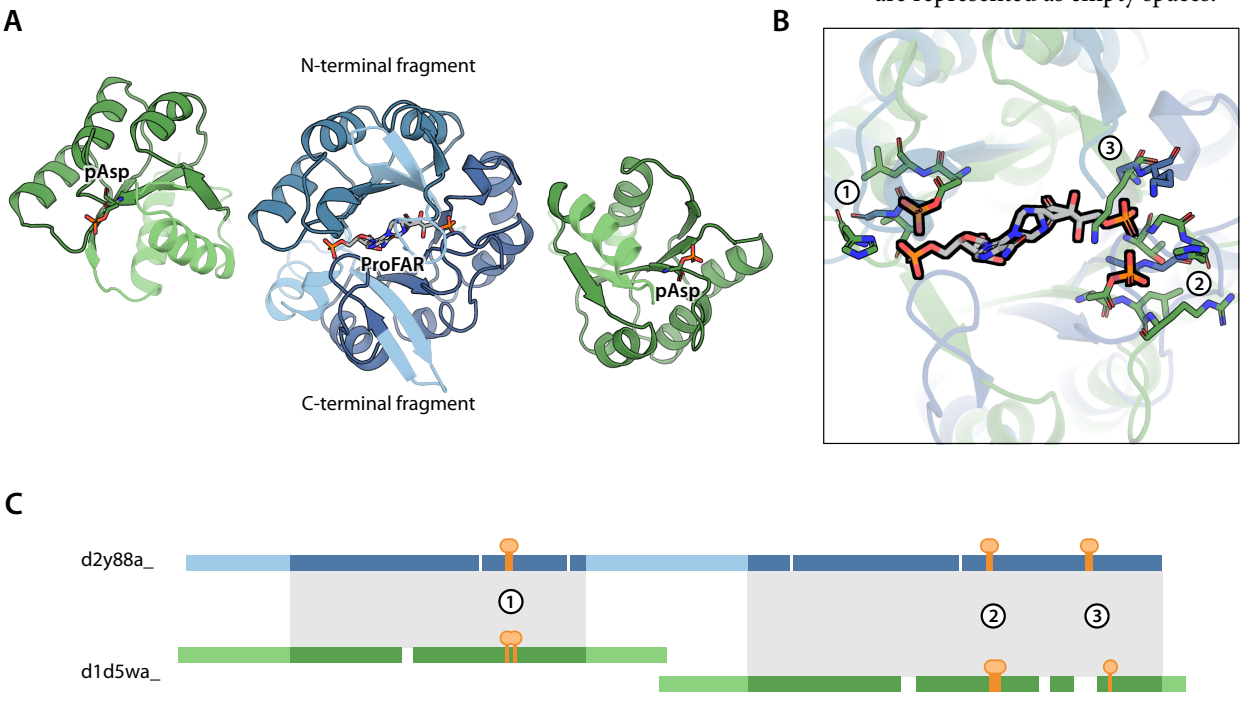


Figure 40: Fragment duplication in phosphorylated ligand binding. A: Phosphoribosyl isomerase A (PriA, SCOP ID d2y88a_, shown in blue). receiver domain from FixJ (FixJN, SCOP ID d1d5wa_, shown in green). B: Detail from A showing overlapping interacting residues. C: Sequence alignment of both fragments. Conserved ligand-interacting residues are highlighted in orange, while gaps are represented as empty spaces.

Hits between TIM barrel and PBP-like I folds

After flavodoxin-like domains, the PBP-like I fold is the most frequent one in TIM barrel hits. PBP-like I fold domains comprise two compact, six-stranded α/β/α lobes connected by a flexible linker region that acts as a hinge. The strand order in both β-sheets is discontinuous, where the first lobe contains strands 1-5 and 11, and the second one strands 6-10 and 12 (Figure 41). The function of PBP-like I domains, usually sensing and transport of small molecules, happens at the inter-lobe cavity, where ligand binding induces a large conformational change, closing and twisting both lobes.

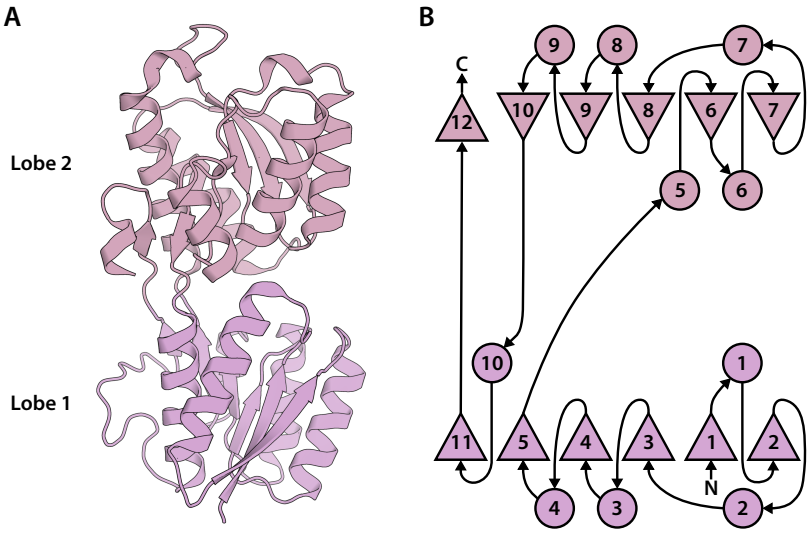


Figure 41: Topology of an archetypical PBP-like I domain (PDB ID 2FVY). Ribbon (A) and topology (B) diagrams of the domain structure are shown. α-helices and β-strands are depicted as circles and triangles, respectively. Adapted from Fukami-Kobayashi, Tateno, & Nishikawa (1999).

Similar to what is seen in TIM barrel domains, the PBP-like I fold contains repetitions in its intrafold hits. In fact, 75% of c.93 domains contain repeat hits with themselves (Figure 25). Hits between TIM barrel and PBP-like I domains have a wide range of sizes, with most fragments having approximately 80 or 140 residues (Figure 42).

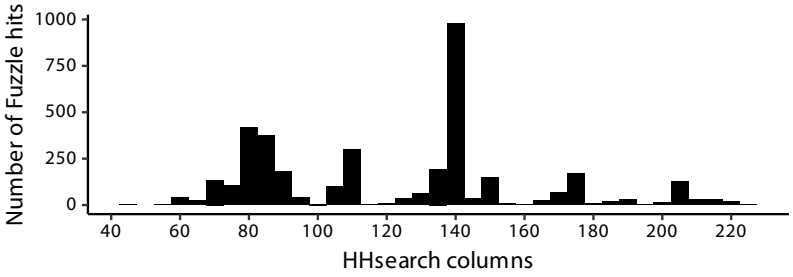


Figure 42: Fragment size distribution in c.1-c.93 hits.

There are large regions in the fragments without structural matches between both folds. Upon closer inspection, the only segment where structure and sequence alignments agree is a $(\alpha\beta)_2$ or $(\alpha\beta)_3$ element, depending on the number of aligned HHsearch columns. There are also discrepancies between sequence and structural comparisons in c.1-c.93 hits. As the number of aligned columns increases, the resemblance between the HHsearch and TM-align alignments decreases sharply (Figure 43), up to a point where, at around 140 columns, both alignments in most hits are completely different.

Not all TIM barrel superfamilies hit PBP-like I domains. In fact, 95% of the c.1-c.93 hits contain domains from the aldolase superfamily (c.1.10, Figure 44).

A look into conserved domain-ligand interactions reveals some c.1-c.93 hits with conserved ligand-interacting residues distributed throughout the fragments. Some of them are present in the structurally aligned loops, but conserved residues involved in ligand binding are located in the structurally divergent section as well. The example in Figure 45

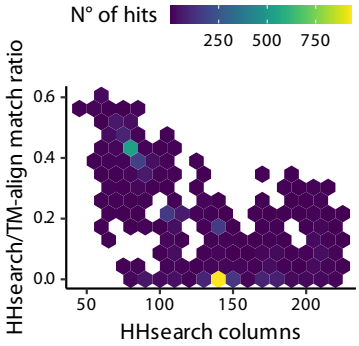


Figure 43: Equivalent matching residues in c.1-c.93 hits. A ratio of 1.0 means that the HHsearch and TM-align alignments are identical. At ratio zero no column between them is alike.

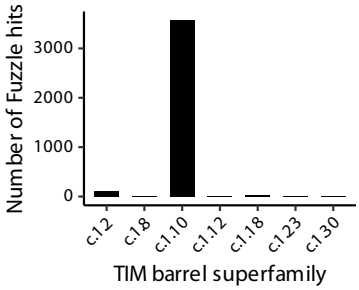


Figure 44: TIM barrel superfamily distribution in c.1-c.93 hits.

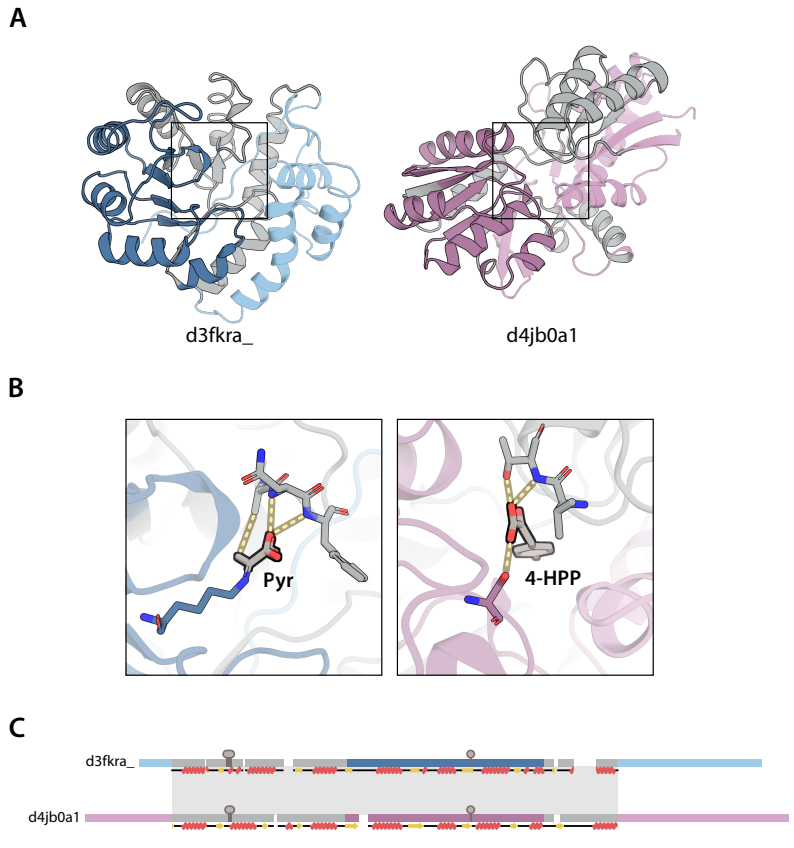
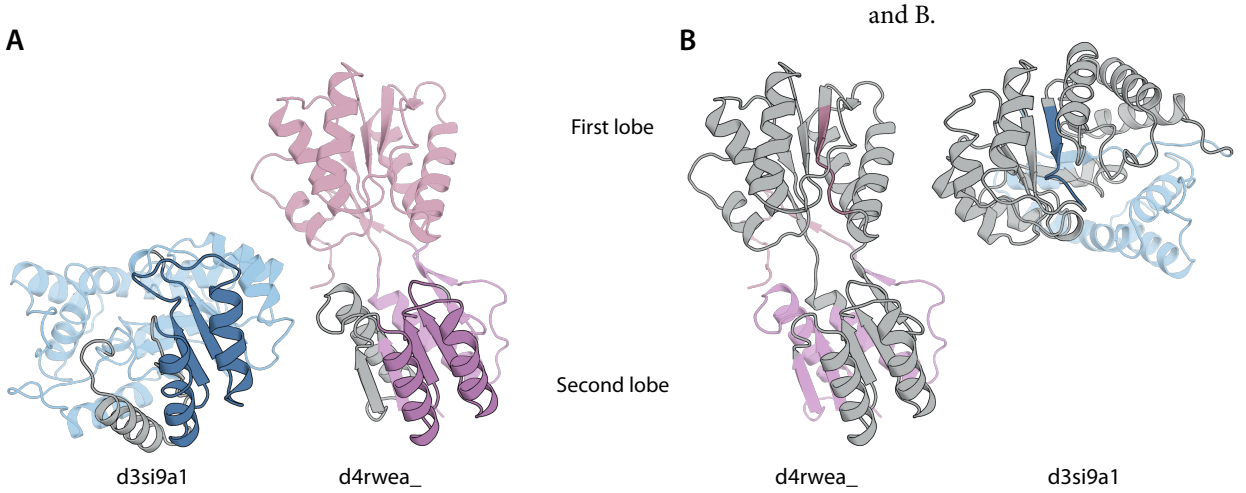


Figure 45: Conserved ligand-interacting residues in c.1-c.93 hits. A: Fuzzle hit between *Azospirillum brasilense* L-2-keto-3-deoxyarabonate dehydratase (left, SCOP ID d3fkra_) and *Rhodopseudomonas palustris* CouP (right, SCOP ID d4jb0a1). B: Detail from A showing conserved ligand-binding residues. Pyr: pyruvate. 4-HPP: 4-hydroxyphenylpyruvate. C: Sequence alignment of the hit. Conserved ligand-binding residues are marked with gray circles.

depicts a TIM barrel and a PBP-like I domain binding pyruvate and a pyruvate derivative, respectively. Even though the orientation of the pyruvate group in both folds is not the same, residues involved in binding are located in a similar manner both in sequence and structure, allowing the substrate to be kept in the binding pocket.

As is the case with c.1-c.23 fragments, some c.1-c.93 domain pairs contain duplicated hits, with a TIM barrel domain hitting both lobes of a single PBP-like I domain (Figure 46).

Figure 46: Duplicated hits between c.1 and c.93 domains. A and B: Fuzzle hits between *Bartonella henselae* 4-hydroxy-tetrahydrodipicolinate synthase (SCOP ID d3si9a1) and *Yersinia pestis* sugar-binding transport protein (SCOP ID d4rwea_). C and D: Corresponding HHsearch alignments for A and B.



C

```
d4rwea_   9  V G D L A N P F F V Q I T K G V E L E A R K L A G D K V K V T L V S S G Y D L G Q Q V A Q I D N F I A A K V D M I I L N
d3si9a1  48  S P T L T H E E H K R I I E L C V E Q V A K R - - - - V P V V A G A G S N S T S E A V E L A K H A E K A G A D A V L V V

d4rwea_  69  A A - - - - - D S K G I G P A V K R A K D A - G I V V V A V D V A
d3si9a1 104  T P Y Y N R P N Q R G L Y T H F S S I A K A I S I P I I I Y N I P
```

D

```
d4rwea_  42  S S G Y D L G Q Q V A Q I D N F I A A K V D M I I L N A A D S K G - - - - - - - - - - I G P A V K R A K D A G I V V V A
d3si9a1  16  N G A I D E K A F C N F V E W Q I T Q G I N G V S P V G T T G E S P T L T H E E H K R I I E L C V E Q V A K R V P V V -

d4rwea_  92  V D V A A E G A D A T I T S D N T Q A G A M A C K Y I S D R L K E K G N V V I I N G P P V S A I Q N R V E G C E S E F K
d3si9a1  75  - - - - - - - - - A G A G S N S T S E A V E L A K H A E K A G A D A - - V L V V T P Y Y N R P N Q R G L Y T H F S S I A

d4rwea_ 152  K Y P D I K V L S S N Q N A K G S R E G G L E V M T S L L A V N P K I D G V F A I N D P T A I G A D L A A K Q A Q R N E
d3si9a1 124  K A I S I P I I I Y N I P S R S V I D M A V E T M R D L C R D F K N I I G V K D A T G K I - - E R A S E Q R E K C G K D

d4rwea_ 212  F F I V G V D G S P D A E E A L K R G G N T L F V A T P A Q D P Q V M
d3si9a1 182  F V Q L S G D D C T - A L G F N A H G G V G C I S V S S N V A P K L C
```

There are varying degrees of structural similarity within these duplicated hits. In the example displayed here, the smaller fragment only has mismatches at the N-terminal region, an αβ segment. The larger fragment, on the other hand, despite aligning almost 200 residues, merely shows TM-align-derived structural matches in a single β-strand between the TIM barrel domain and the PBP-like I lobe. Duplication of ligand-binding residues in these fragments could not be found.

Fragments between Fld-like and PBP-like I folds

AS MENTIONED IN THE INTRODUCTION, our group has previously shown that c.93 domains are related to the Fld-like fold. Specifically speaking, a remotely homologous $(\beta\alpha)_4\beta$ fragment was found between the Fld-like CheY and the PBP-like I leucine-binding protein. It was proposed that both folds share an ancestral region, which in the case of the PBP-like I fold was duplicated, leading to its bilobed structure. This assumption prompted us to analyze this evolutionary link more in depth. 7 212 hits between these two folds exist in the filtered subset of Fuzzle (Table 7), and most fragments lie within two

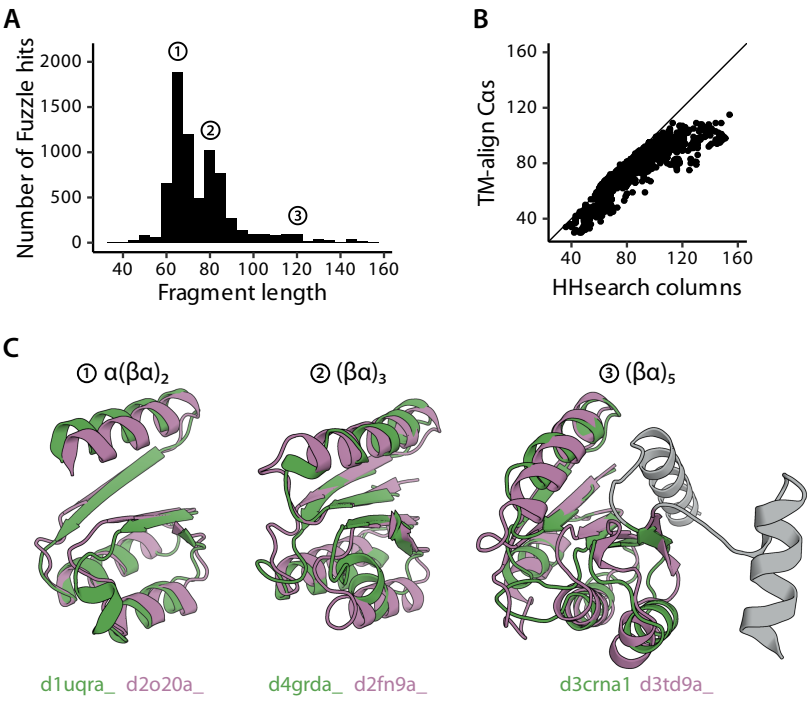


Figure 47: Size distribution in c.23-c.93 hits from Fuzzle's filtered subset. A: Distribution of HHsearch columns. B: Relationship between HHsearch columns and TM-align superimposed Cαs. The identity line is shown in black. C: Size differences in c.23-c.93 fragments. Fld-like and PBP-like I domain structures are colored green and purple, respectively. Structurally divergent residues are shown in gray.

ranges of length: 65-70 and 80-90 columns (Figure 47A). Both groups of fragments differ by one β-strand. If the number of aligned columns in the hit is below 90, the `ca_tm_pair`:`cols` ratio remains fairly constant, averaging 0.95 (*SD* = 0.05, Figure 47B). But at larger fragment sizes (`cols` ≥ 90) this proportion decreases (*M* = 0.83, *SD* = 0.09). Similar to the case of TIM barrel/Fld-like hits, this is explained by the structural divergence at the C-terminal end of the fragments, where the PBP-like I domain crosses over to the other lobe. As with the c.1-c.23 case, the secondary structure of the corresponding fragment residues is preserved despite the differences in spatial arrangement (Figure 47C).

There is a marked tendency for these hits to start and end at the N-terminal lobe of the PBP-like I domains, followed by fragments that start at the N-terminal lobe and cross over to the C-terminal one (Table 8). These two categories account for 87% of the total hits.

Table 8: Fragment positioning in c.23-c.93 hits.

Start lobe	End lobe	Number of hits
1	1	5 310
1	2	958
2	1	244
2	2	700

The distribution of Fld-like superfamilies in c.23-c.93 reveals that four superfamilies are mainly represented (Figure 48): CheY-like (c.23.1), N^5-CAIR mutase (c.23.8), type II 3-dehydroquinate dehydratase (c.23.13), and class I glutamine amidotransferase-like (c.23.16). On the other hand, all PBP-like I domains belong to a single superfamily.

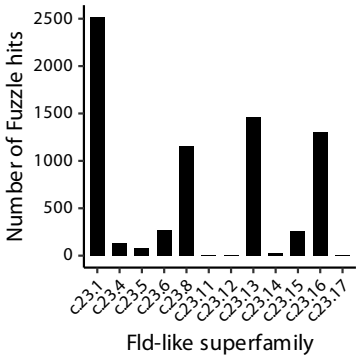


Figure 48: Fld-like superfamily distribution in c.23-c.93 hits.

Similar to what is observed in the TIM barrel/Fld-like case described previously, a group of hits exist where a PBP-like I domain is hit twice by a Fld-like domain. In the vast majority domain pairs with this property, each hit targets a different PBP I-like lobe with no overlaps. An example hit is shown in Figure 49, where both fragments are long enough to span the two lobes. Instances of duplication were found in four c.23 superfamilies: c.23.1 (CheY-like), c.23.8 (N^5-CAIR mutase), c.23.15 (ribosomal protein S2), and c.23.16 (class I glutamine amidotransferase-like), with the latter being the most frequent

one. The presence of these kind of repeated fragments gives support to the theory that an ancestral fragment duplication took place during the evolution of the PBP-like I fold.

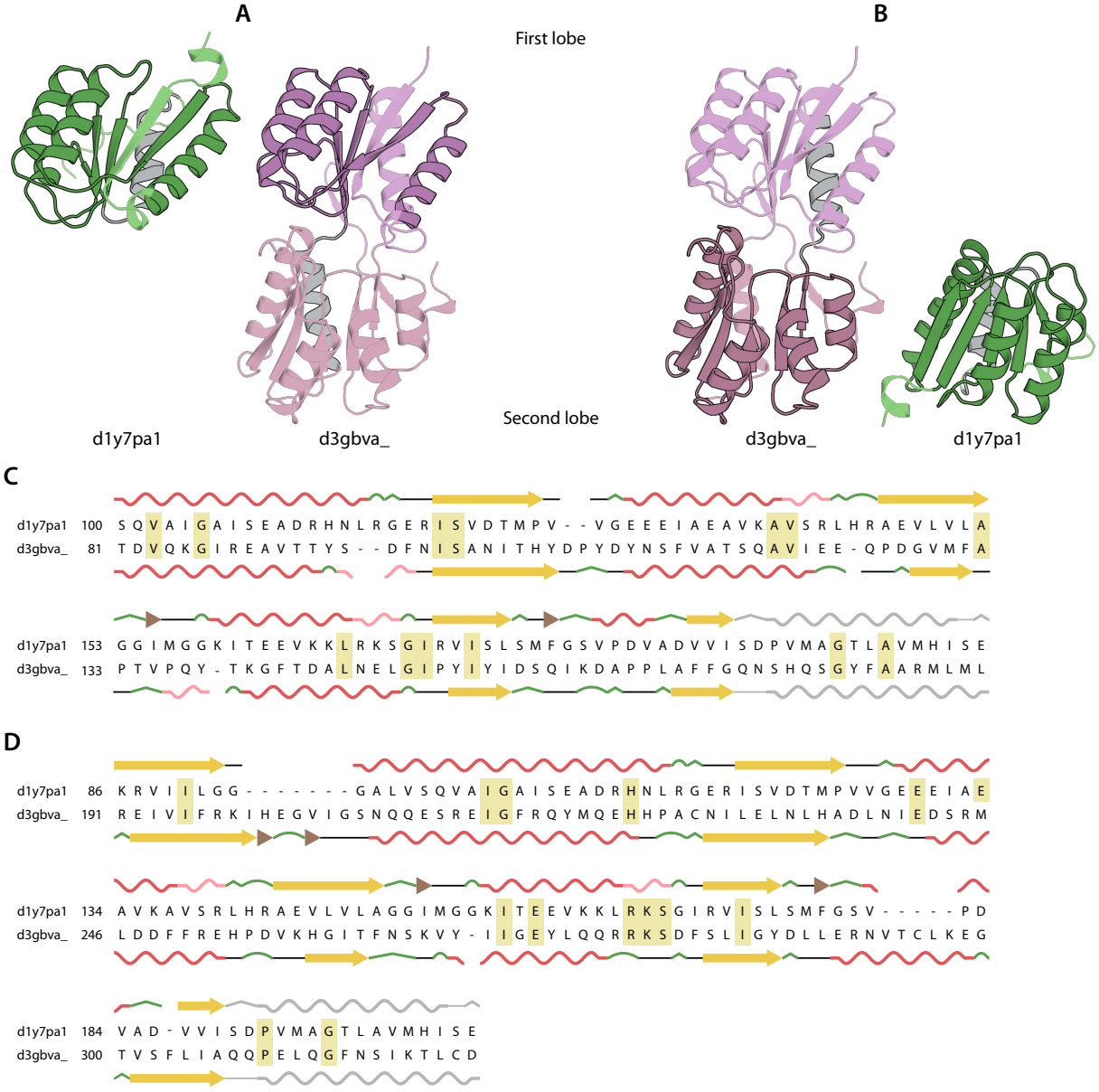


Figure 49: Fragment duplication in c.23-c.93 hits. A and B: Representative pair of hits between the C-terminal domain of hypothetical protein AF1403 from *Archaeoglobus fulgidus* (SCOP ID d1y7pa1) and a putative LacI-family transcriptional regulator from *Bacteroides fragilis* (SCOP ID d3gbva_). Domains are colored as described in Figure 47. C and D: Corresponding HHsearch alignments for A and B.

Six Fld-like superfamilies contain hits with conserved residues involved in ligand binding (Figure 50). The nature of the ligands in both folds vary in each case, but some common patterns can be found and described. In the case of ligands containing phosphate (pAsp, FMN, AIR, T6P), most conserved residues are interacting with this group. This is clearly shown in hits between PBP-like I domains and the c.23.8 superfamily, where both ligands have similar phosphate-binding sites.

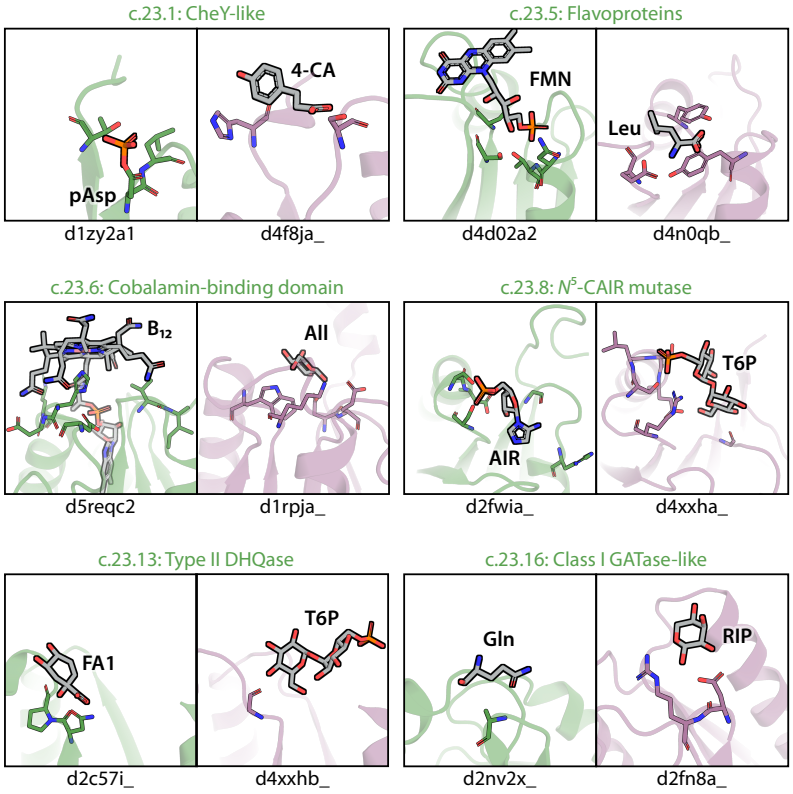


Figure 50: Ligand interactions in c.23-c.93 hits. One representative hit is shown for every Fld-like superfamily of interest. Folds are colored as described in Figure 47, while ligands are shown in gray. pAsp: phosphoaspartate, 4-CA: *p*-coumaric acid, FMN: flavin mononucleotide, Leu: leucine, B_{12}: cobalamin, All: D-allose, AIR: 5-aminoimidazole ribonucleotide. T6P: trehalose-6-phosphate, FA1: 2,3-anhydro-quinic acid, Gln: glutamine, RIP: β-D-ribopyranose.

The cobalamin-binding superfamily of Fld-like shows conserved ligand-binding patterns similar to what was described in c.1-c.23 fragments, with one exception. The histidine residue present in the DXHXXG motif, responsible for the coordination of the cobalt ion found in cobalamin, aligns with loop residues in PBP-like I where interactions with sugar molecules take place (see the c.23.6 subpanel in Figure 50). The structural alignment in these kind of hits shows different loop conformations, suggesting the presence of evolutionary pressure for the accomodation of different sorts of substrates.

Duplicated ligand-binding hits were found only within the N^5-CAIR mutase superfamily (PurE, c.23.8). In these, both hits have overlapping residues of interest. The example pictured in Figure 51A shows the Fld-like domain of PurE (N^5-carboxyaminoimidazole ribonucleotide mutase) hitting both lobes of a PBP-like I D-galactose-binding periplasmic protein. In both instances, residues involved in binding their corresponding ligands (N^5-CAIR and galactose, respectively) are positioned similarly in their sequences and structures (Figure 51B and C).

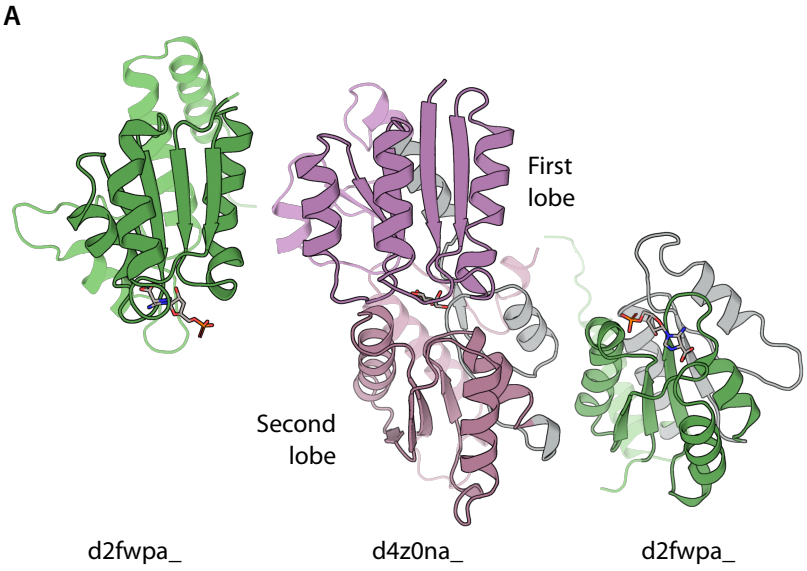


Figure 51: Conserved ligand-interacting residues in duplicated c.23-c.93 hits. A: Duplicated Fuzzle hits between *Acetobacter aceti* PurE (SCOP ID d2fwpa_) and *Streptobacillus moniliformis* D-galactose-binding PBP (SCOP ID d4z0na_) bound to isoCAIR (4(*R*)-carboxy-5-iminoimidazoline ribonucleotide) and D-galactose, respectively. B: Conserved ligand-interacting residues found in said hits. C: Location of conserved interacting residues in the domain sequences.

Combined evolutionary relationships between c.1, c.23 and c.93

Since c.1, c.23 and c.93 seem to be closely related folds, Fuzzle hits between them were clustered (Figure 52) to analyze their joint evolutionary relationships. Out of all 33 TIM barrel superfamilies only one of them, aldolase (c.1.10), connects directly both Fld-like and PBP-like I domains. Likewise, two Fld-like superfamilies, the CheY-like and cobalamin-binding domain ones (c.23.1 and c.23.6 respectively) connect TIM barrel and PBP-like I domains.

It is worth mentioning that, as previously reported and unlike the TIM barrel fold, Fld-like superfamilies do not form a cohesive cluster, with some of them even being connected exclusively to non-Fld-like domains. This can be explained as an instance of convergent evolution among them, where the Fld-like architecture has appeared independently several times in different evolutionary contexts and fulfilling distinct functions.

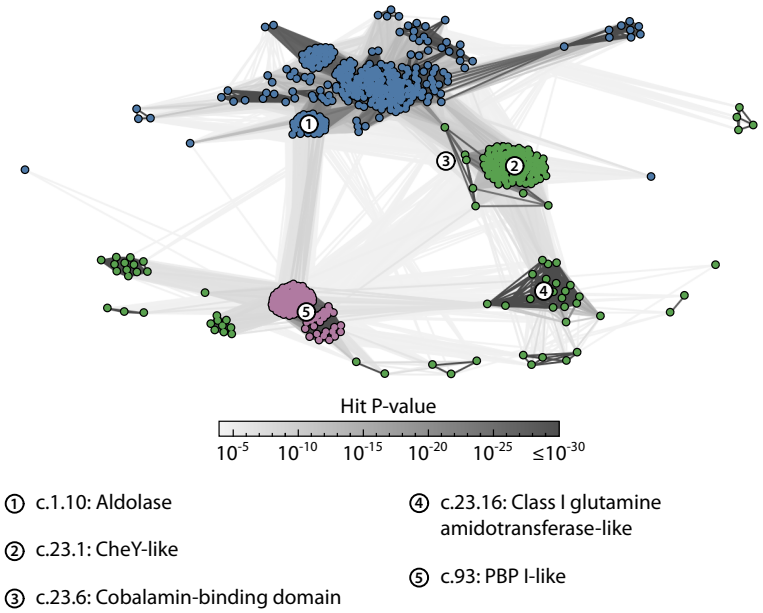


Figure 52: Cluster map of c.1, c.23 and c.93 domains.

Conserved fragments in Rossmann fold domains

THE ROSSMANN FOLD IS CONSIDERED one of the most ancient ones, having been present already in early metabolic networks.[2] Furthermore, Rossmann fold domains are among the most widely distributed ones in present proteins and are likewise the most frequent fold in interfold Fuzzle hits: looking at the filtered subset, 211 401 hits include a Rossmann fold domain (Table 7, Figure 22). Most of them (about 60%) involve either FAD/NAD(P)-binding domains (SCOP fold c.3, Figure 53) or S-adenosyl-L-methionine (SAM) dependent methyltransferases (c.66). Despite having a similar number of hits, the allocation of fragment sizes within them differs greatly. In fact, by looking at the size distributions in a number of folds sharing fragments with the Rossmann fold, some of them are markedly right-skewed, with most hits having less than 40 residues (Figure 54, see c.3 and c.4 for example). Other folds have mostly fragments larger than 50 residues (e.g. c.78 and c.79 in Figure 54), while others include conserved fragments in a broad range of lengths.

Given the prevalence of Rossmann fold hits in Fuzzle, we characterized the relationships with some of its most frequently connected folds to find common features that could explain the presence of these conserved fragments and their permanence along the evolutionary history of these protein folds.

[2] Caetano-Anollés et al., 2007

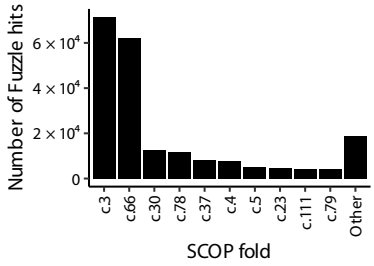


Figure 53: Fold distribution of Rossmann fold-hitting fragments within the interfold subset.

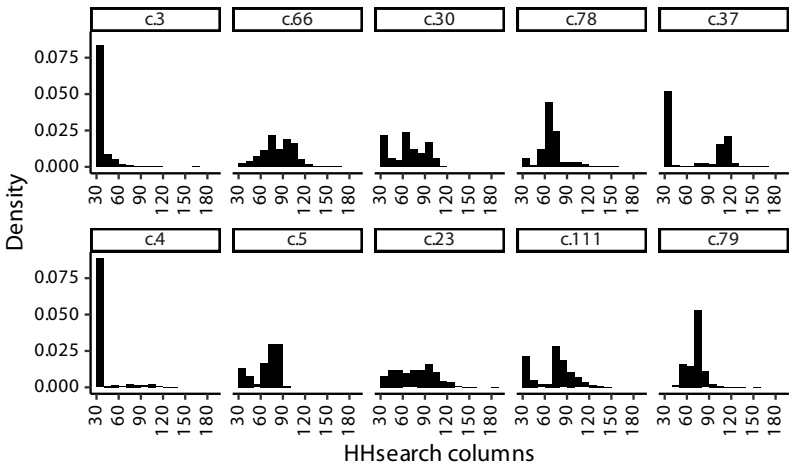


Figure 54: Length distribution of Rossmann fold hits in the 10 most frequent corresponding folds from the filtered interfold subset. Densities are scaled to integrate to 1.

Hits between Rossmann and FAD/NAD(P)-binding folds

Most Rossmann fold hits are associated with FAD/NAD(P)-binding domains (SCOP ID c.3). While the former has a typical α/β doubly-wound architecture with a 321456 sheet order, the latter forms a β/β/α sandwich with a 32145 topology in its central β-sheet. Despite the similarity in secondary structure placement, most fragments between both folds are rather short, having an average of 36.9 residues (*SD* = 9.5 in the filtered subset) and comprising a βαβ motif at the N-terminus of the domains involved (Figure 55).

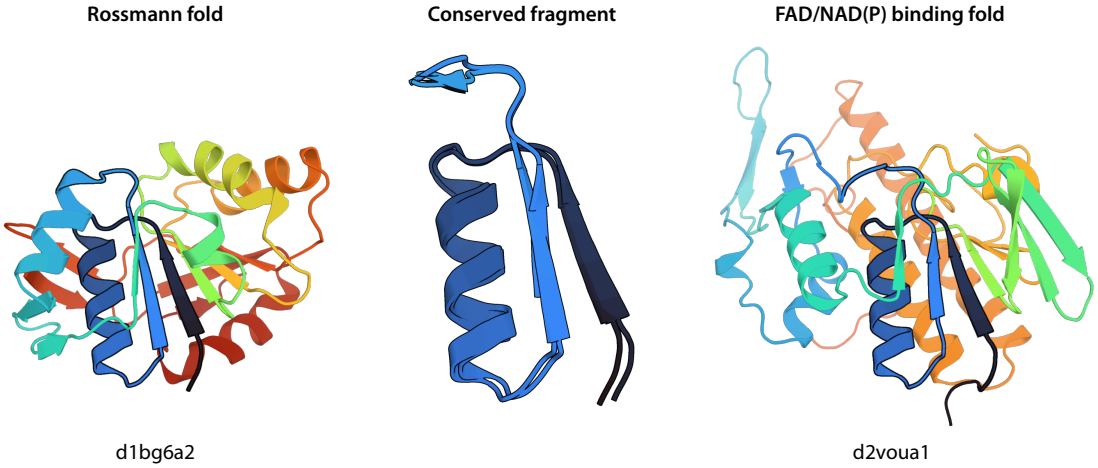


Figure 55: Representative Rossmann/FAD/NAD(P)-binding hit. N- and C-termini are colored blue and red respectively.

As their name indicates, FAD/NAD(P)-binding domains use said cofactors mainly in reductase proteins. It should be noted that, despite the relatively small fragment size, both folds contain several conserved residues in them that bind the adenylate moiety of their corresponding dinucleotides: NAD(P) in the case of c.2 and FAD or NAD(P) in c.3 domains (Figure 56A). In these fragments, cofactor interacting residues are located at the N-terminal end of the helix and the loops following every β-strand. Ligand-binding interactions are fully conserved between both folds. Loop 1 and the helix interact mostly with the ribose and phosphate groups from the adenine side. Loop 2, on the other hand, interacts with the adenylate group as well as the 2′-phosphate of NADP. In NAD-binding Rossmann fold domains a conserved, ribose-interacting aspartate residue can be found instead of the phosphate-binding one. Additionally, the adenylate moieties are positioned identically in c.2 and c.3. In contrast, the orientation of the non-adenosyl moiety

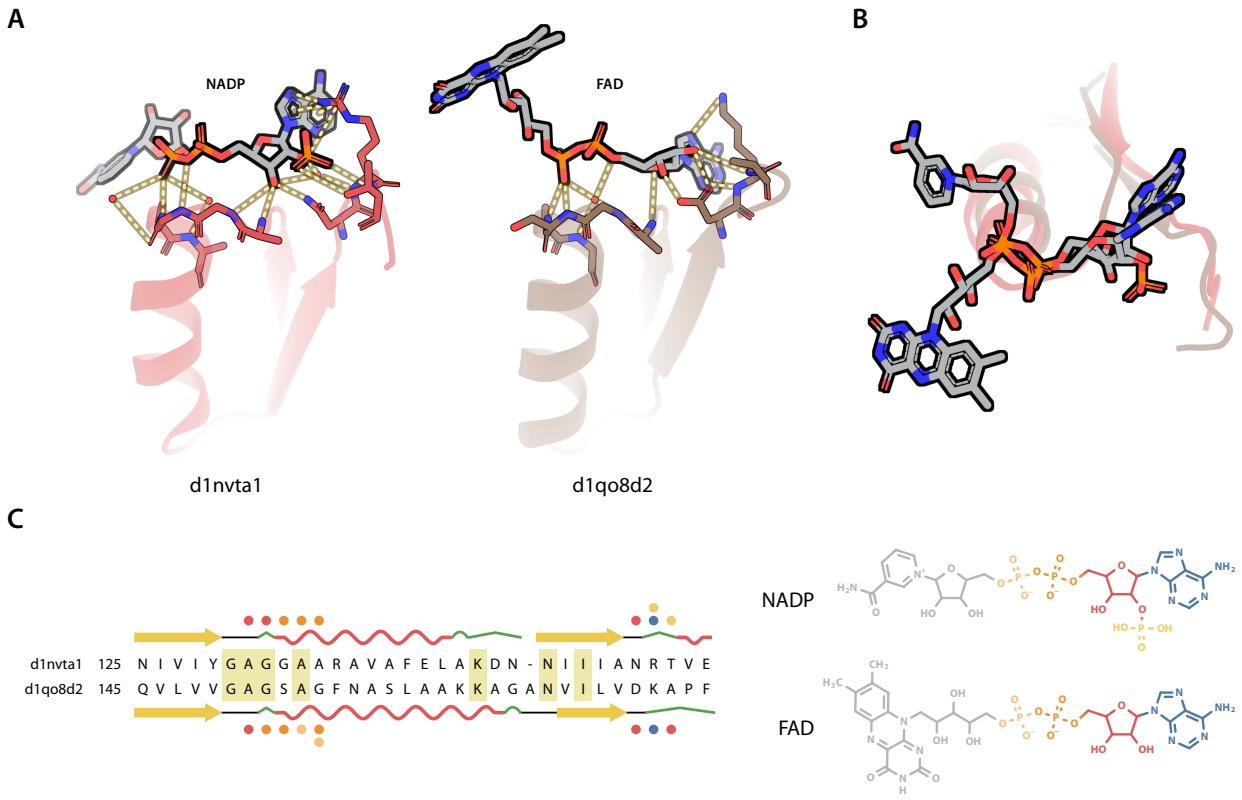


Figure 56: NAD(P)/FAD interactions in c.2-c.3 fragments. A: Fuzzle hit between *Methanococcus jannaschii* shikimate 5-dehydrogenase (SCOP ID d1nvta1) and *Shewanella frigidimarina* fumarate reductase flavoprotein subunit (d1qo8d2). Conserved ligand-binding residues within the fragments are shown. B: Superposition of the fragments and ligands in A. C: HHsearch alignment of the hit. Dots depict ligand-binding residues. Their color legend is shown on the right hand side.

is different in both folds and so is the location of the interacting residues. Therefore, no conserved residues could be found between them (Figure 56B).

Hits with SAM-dependent methyltransferases

The second most frequent fold hitting Rossmann fold domains is the SAM-dependent methyltransferase (SAM MTase, c.66) fold. SAM MTase domains have a β/α/β sandwich-like fold with a mixed β-sheet: its order is 3214576, with the seventh β-sheet positioned antiparallel to the rest. Most conserved fragments have alignments in the 70-75 and 95-100 residue range, but some trends around 40-45 and 55-60 can also be seen, albeit with lower frequencies (Figure 57A). These local maxima represent fragments with an incrementing number of secondary structure elements: $(\beta\alpha)_2$, $(\beta\alpha)_2\beta$, $(\beta\alpha)_3\beta$, and $(\beta\alpha)_4\beta$, respectively (Figure 57B). Despite the wide range of sizes, and unlike the case of c.1-c.23 hits, differences between `cols` and `ca_tm_pair` remain small throughout the whole extent of it: the `cols`:`ca_tm_pair` ratio is, on average, 0.946, (*SD* = 0.055, Figure 57C).

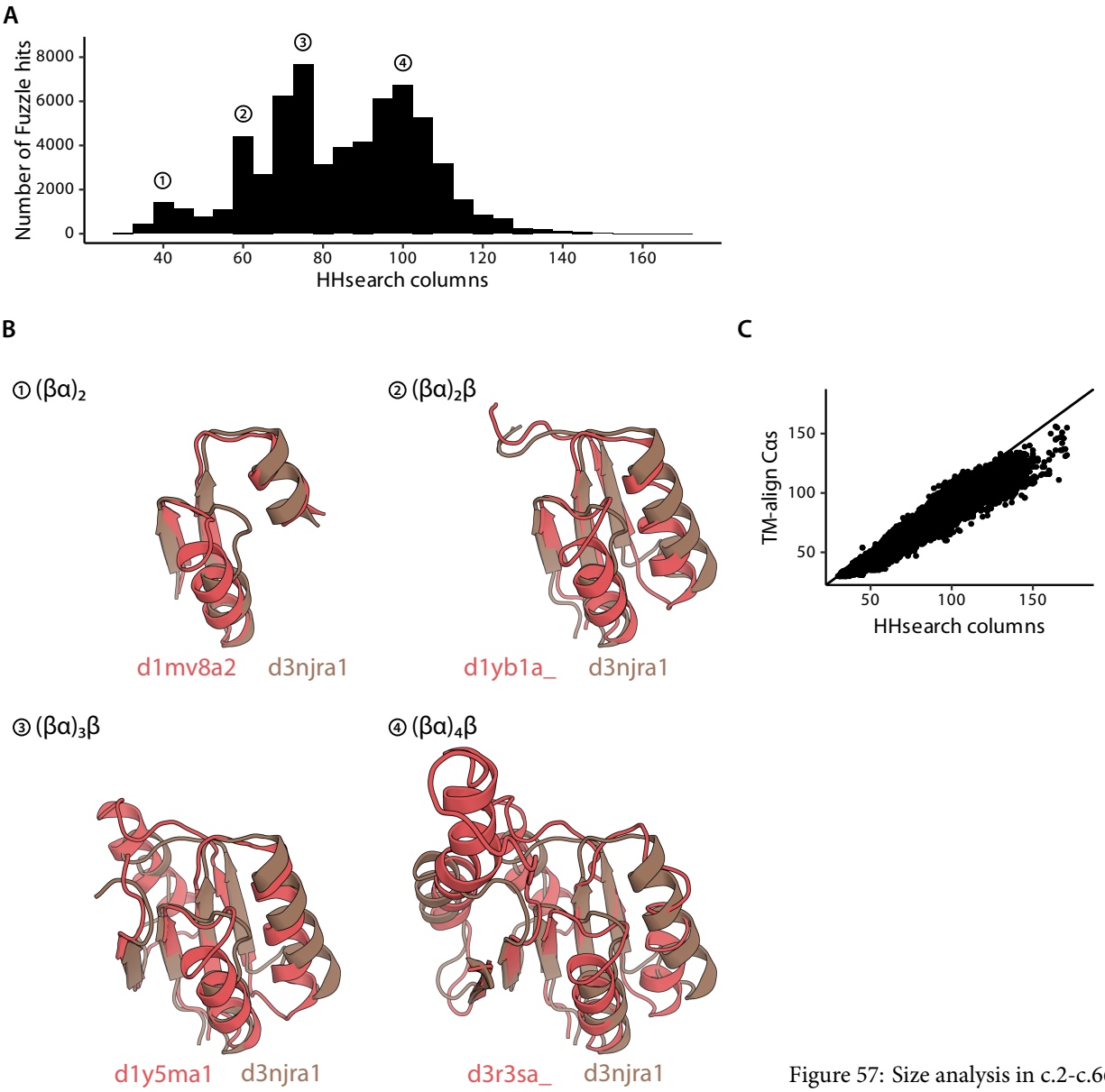


Figure 57: Size analysis in c.2-c.66 hits. A: Fragment length distribution. B: Example fragments with different lengths. C: Relationship between HHsearch and TM-align fragment lengths.

Both folds have well-defined nucleotide cofactors involved in their function. SAM MTases use S-adenosyl-L-methionine (SAM) as a methyl group donor. In Fuzzle hits, conserved residues involved in cofactor binding are present in loops 1-4 from both folds (Figure 58). Most identical interactions in c.2 and c.66 are related to the adenosyl group present in NAD(P) and SAM, where aligned residues, located mostly in loops 2 and 3 of the domains, are associated to the adenine ring as well as the ribose groups. As mentioned in the previous section, loop 1, the glycine-rich one present in the Rossmann fold, interacts with the phosphate groups present in NAD(P). On the other hand, its spatial equivalent in methyltransferases, since SAM lacks a phosphate group, binds mainly the amino, α-carbon,

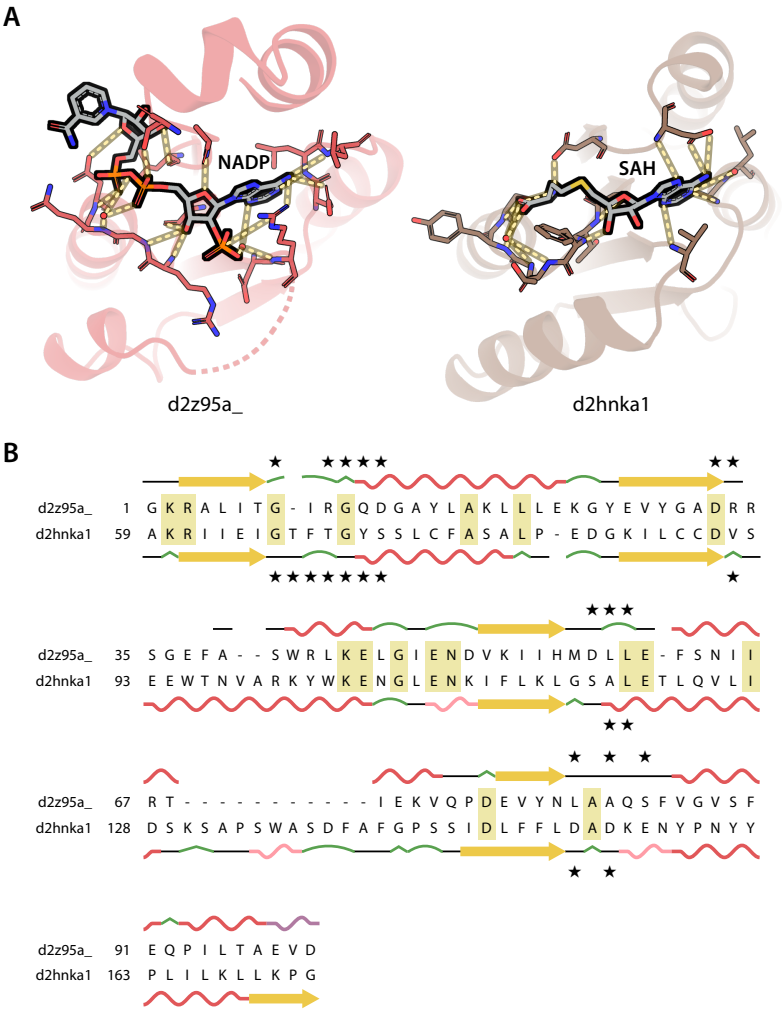


Figure 58: Conserved ligand interactions in c.2-c.66 hits. A: Fuzzle hit between *Aquifex aeolicus* GDP-D-mannose dehydratase (c.2, left) and *Leptospira interrogans* SAM-dependent O-methyltransferase (c.66 right) in complex with their respective cofactors. Note that SAH (S-adenosyl-L-homocysteine) is the product of SAM demethylation. B: HHsearch alignment of the hit. Ligand-interacting residues are marked with a star.

and carboxyl groups of the methionyl moiety of the cofactor. Lastly, loop 4 in Rossmann fold domains interacts with the ribose group of both nucleotides in NAD(P). In SAM MTase domains this loop harbors a metal-binding site necessary for catalysis in metal-dependent methyltransferases.

Relationships between Rossmann fold and preATP-grasp domains

The preATP-grasp fold (SCOP ID c.30) accommodates a number of different enzyme families, including biotin carboxylase-like ones, D-alanine—D-alanine ligase, and glutathione synthetase. The fold is comprised of α/β/α layers arranged in a parallel or mixed β-sheet of 4 to 6 strands.

Fuzzle shows that, in the filtered interfold subset, both c.2 and c.30 folds hit a higher-than-average number of distinct do-

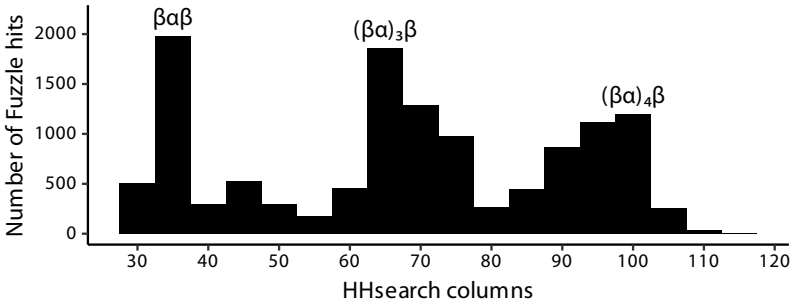


Figure 59: Fragment size distribution in c.2-c.30 hits.

mains and folds (Figure 23). Hits in the filtered subset between them lean towards high probabilities, with 36.4% of them having probability ≥ 95% (Table 7). The distribution of fragment sizes contains three maxima representing βαβ, $(\beta\alpha)_3\beta$, and $(\beta\alpha)_4\beta$ motifs (Figure 59).

The c.30 fold contains nine protein families, out of which one, the biotin carboxylase N-terminal domain-like (SCOP ID c.30.1.1), is predominantly hit by Rossmann fold domains (Figure 60). Not all c.30 families have conserved substrate-interacting residues with the Rossmann fold. Conserved interactions with only two c.30 ligands, glutathione (GSH) and

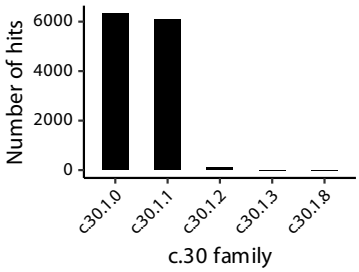


Figure 60: PreATP-grasp families in c.2-c.30 hits. Note that c.30.1.0 correspond to automated matches that have not been classified into families yet. In this particular case the majority of c.30.1.0 domains have high sequence similarity with c.30.1.1.

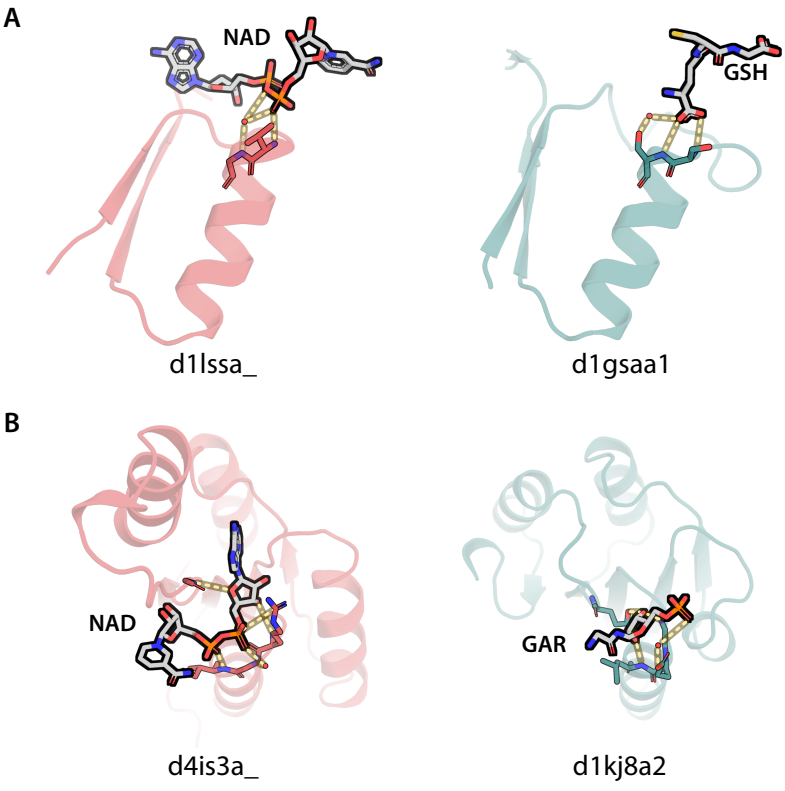


Figure 61: Conserved ligand-interacting residues between NAD-binding c.2 domains, a c.30 GSH-binding (A) and a GAR-binding (B) domain.

glycinamide ribonucleotide (GAR), could be found (Figure 61). Residues interacting with the adenylate moiety's 5′-phosphate group in NAD overlap with the ones making interactions with glutamate's α-carboxyl group in GSH. On the other hand, residues interacting with GAR associate mostly with the ribose group, with one residue forming a water bridge with the phosphate group. In this case, the Rossmann fold fragment interacts with both 5′-phosphate groups and the adenosine-associated ribose group of NAD.

Relationships between the Rossmann and P-loop containing folds

It has been proposed that the Rossmann and P-loop containing NTPase (c.37) folds arose from a common ancestral motif, which diverged and evolved distinctive ligand interaction sequence motifs.[3] The canonical P-loop binding motif, also known as Walker-A, is a glycine-rich, highly conserved sequence motif related to phosphate binding. Rossmann fold proteins, on the other hand, developed a different signature, involving a glycine-rich loop between the first α-helix and β-strands and a ribose-interacting carboxylate side chain at the end of β-strand 2.

[3] Longo et al., 2020

Fuzzle contains hits between the Rossmann fold and P-loop containing NTPases, the latter being the fifth most frequent fold in Rossmann fold-containing hits. Hits of these kind contain the N-terminal section of the domains, including the phosphate-binding motifs (Figure 62). Although they align in sequence, there can be mismatches in the structural alignment. This is expected, as different c.37 superfamilies can have differ-

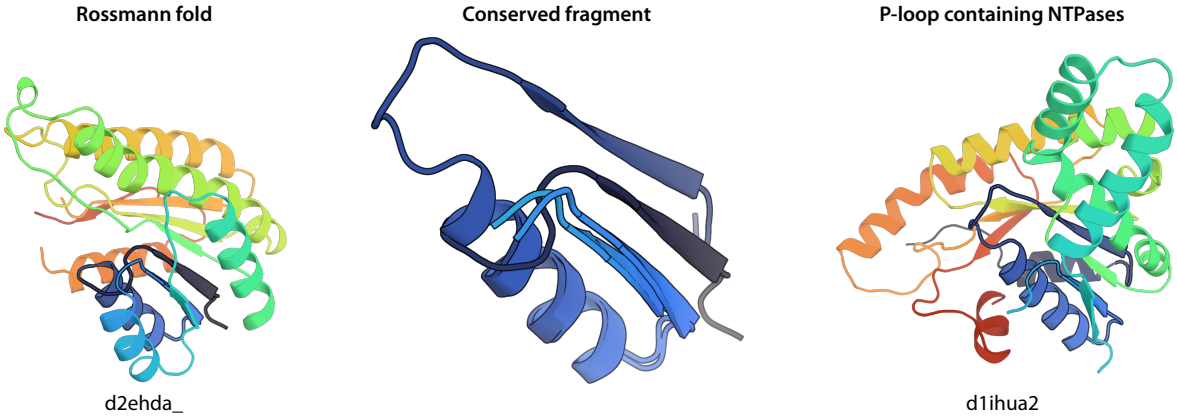


Figure 62: Representative c.2-c.37 hit.

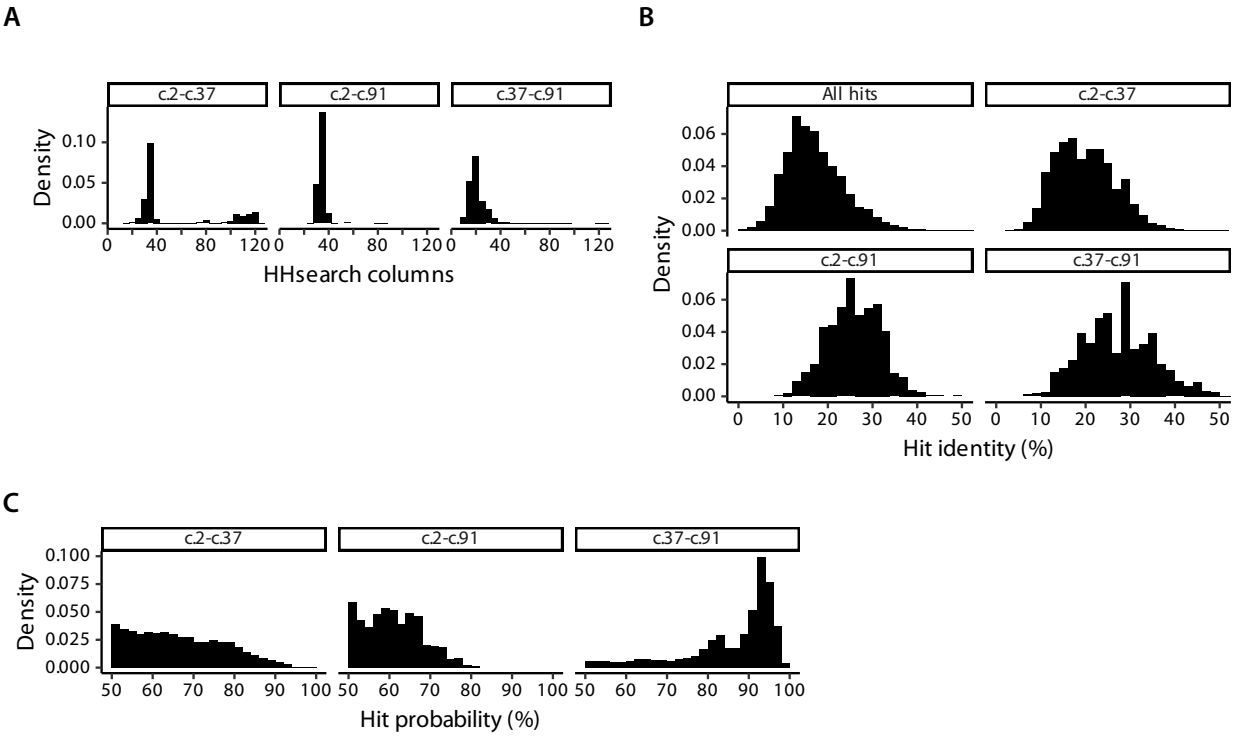


Figure 63: Characterization of c.2-c.37, c.2-c.91, and c.37-c.91 hits in the unfiltered interfold subset of Fuzzle. A: Hit length distribution. B: Hit identity distribution. C: Hit probability distribution.

ent sheet order. Additionally, both folds hit PEP carboxykinase-like domains (SCOP ID c.91), another fold having P-loop motifs. Within the interfold subset of Fuzzle these fold pairings have distinct fragment size patterns (Figure 63A). c.2-c.37 fragments have a bimodal HHsearch columns distribution, with two maxima at around 35 and 120 residues. c.2-c.91 fragments, on the other hand, trend only at a length of approximately 35 residues. c.37-c.91 fragments are even smaller, having their average near 20 residues. For this reason all hits described further in this section were selected using the same parameters as with the filtered intrafold set, but lowering the minimum length cutoff to 20 residues instead of the usual value of 30.

An unusual property of c.2-c.91 and c.37-c.91 fragments is their high sequence identity compared to most fold pairs, including c.2-c.37 (Figure 63B). Specifically, their identities averaged 26.3 and 27.7% (*SD* = 5.96 and 8.02 respectively). Moreover, the fragments differ in their HHsearch probabilities. c.37-c.91 tend towards high-probability hits (≥ 90%), unlike the other two fold pairings that have mostly lower probability values (Figure 63C).

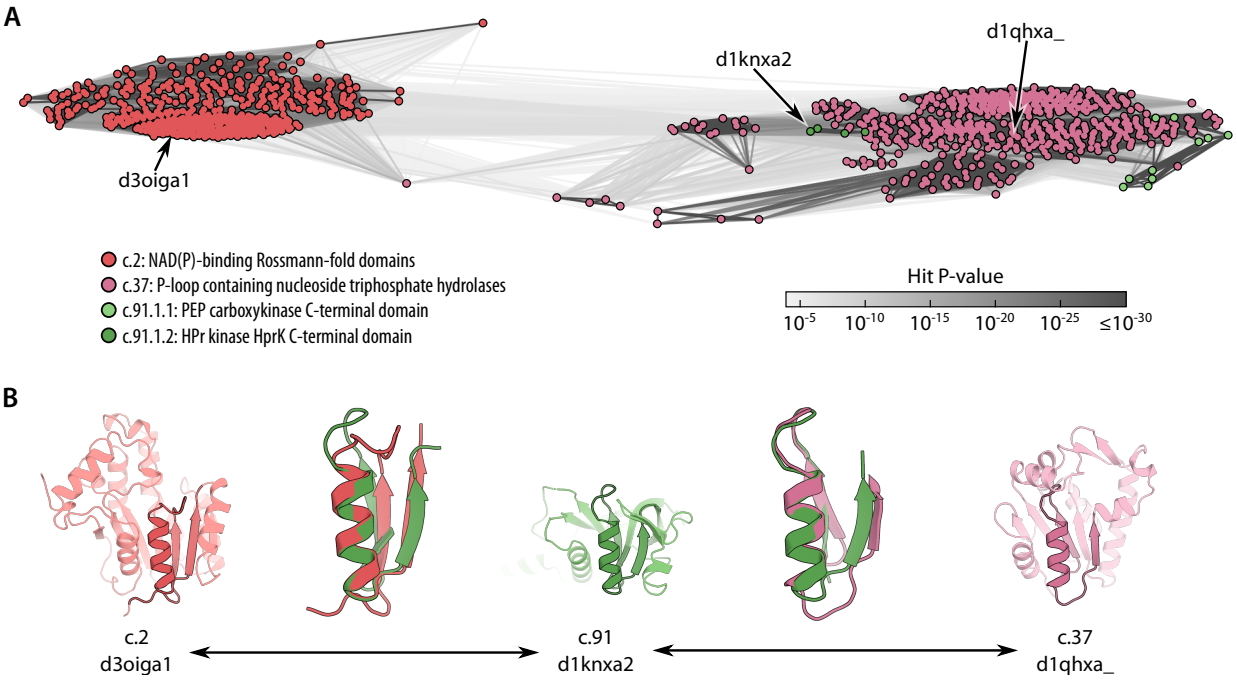


Figure 64: Clustering analysis of c.2, c.37, and c.91 domains. A: Cluster map of Fuzzle hits involving these domains. B: Representative hits connecting all three folds.

Domains from all three folds were clustered according to their Fuzzle hits present in the filtered subset (Figure 64). As expected, and owing to their higher probability values, c.91 domains cluster much closer to the c.37 fold. However, the clustering shows that one of the two classified c.91 families in SCOP, HPr kinase HprK C-terminal domain, (HPr-C, c.91.1.2) has hits connecting both c.2 and c.37 folds at significant *P*-values. The remaining family, PEP carboxykinase C-terminal domain (PEPCK, c.91.1.1), hits exclusively P-loop containing NTPase domains (Figure 64A). Both c.2-c.91 and c.37-c.91 fragments consist of an N-terminal βαβ motif (Figure 64B). Similar to the c.2-c.37 case, c.2-c.91 fragments exhibit differences in their sheet topology and phosphate-binding loop geometry.

Given the connection among c.2, c.37 and c.91, we searched Fuzzle for additional intermediate folds that might lie between c.2 and c.37. Including c.91, fifteen folds were found to contain hits connecting both c.2 and c.37 domains (Figure 65). Most of these intermediates are situated closer to the Rossmann fold, but certain folds and superfamilies lie halfway through both c.2 and c.37 domains. Specifically, two Fld-like superfamilies are present there: CheY-like (c.23.1) and β-D-glucan exohydrolase, C-terminal domain-like (c.23.11). Twelve of the intermediate folds bind nucleotide-based and/or phosphorylated ligands or

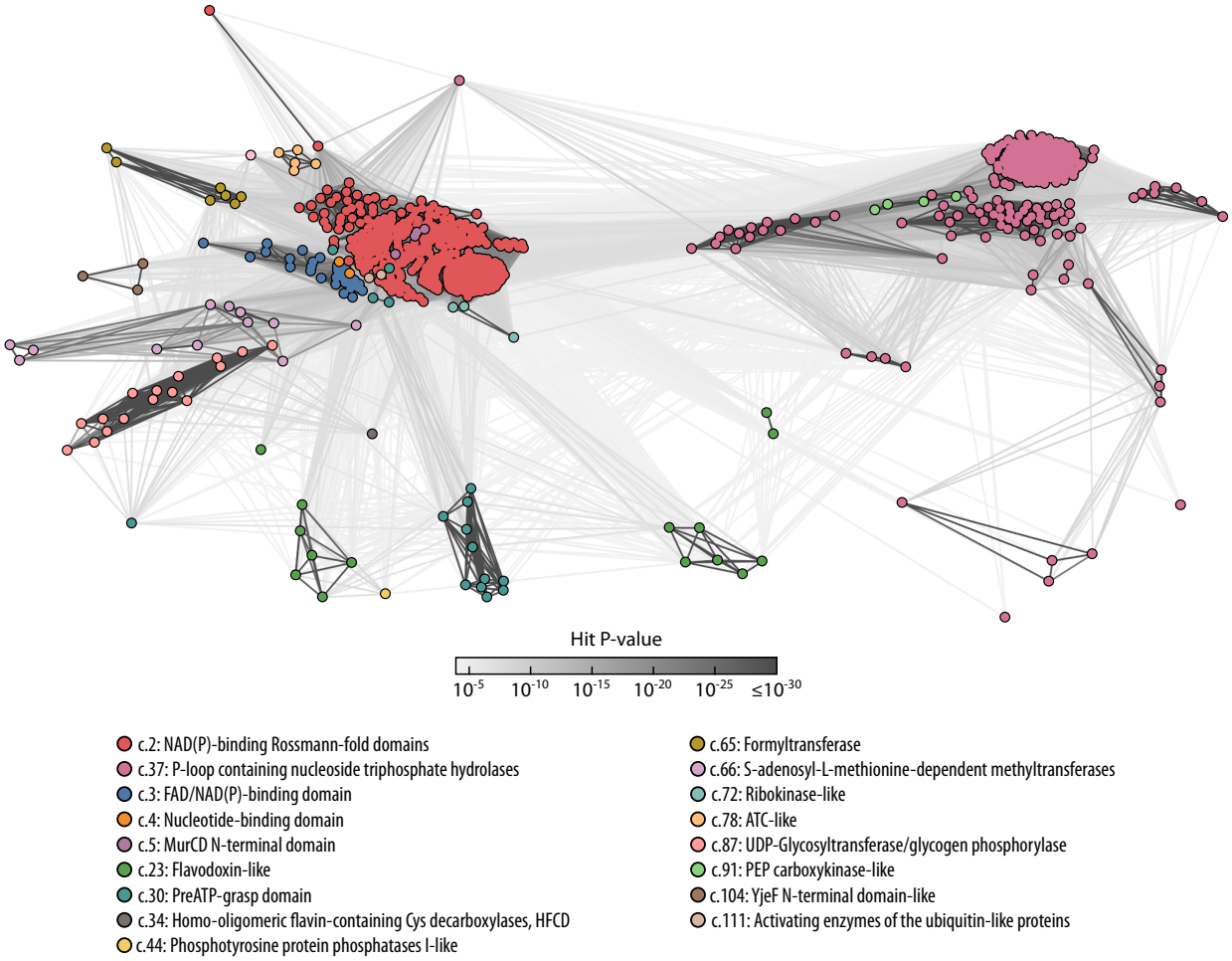


Figure 65: Intermediate hits between c.2 and c.37 domains.

cofactors and were shown to carry residues involved in ligand interaction that are conserved with both c.2 and c.37 folds (Figure 66). One fold, the PEP carboxykinase-like (c.91), is not binding a ligand but acts as a protein kinase instead. Conserved residues from c.91 domains hold a phosphate group and can interact with phosphorylable residues from the kinase targets as well.

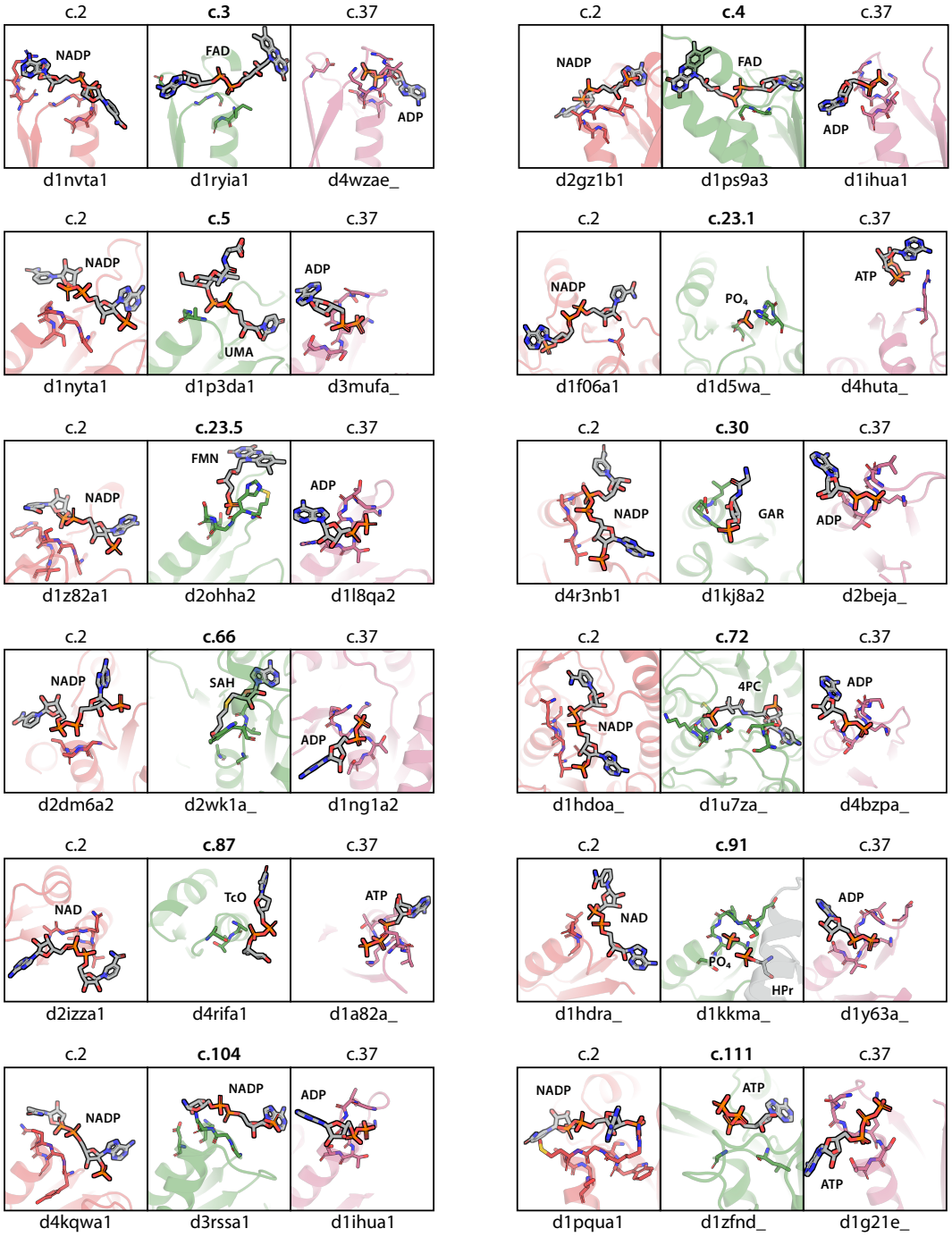


Figure 66: Conserved ligand-interacting residues between c.2, c.37 and intermediate fold hits. The largest combined intermediate fold fragment between c.2 and c.37 is shown in each middle panel. 4PC: 4'-phosphopantothenate. TcO: TDP-carba-D-olivose. HPr: Phosphocarrier protein HPr.

Conserved fragments with MurCD N-terminal domains

The MurCD N-terminal domain fold (SCOP ID c.5) is present in enzymes involved in peptidoglycan biosynthesis, assisting in the addition of a pentapeptide to UDP-N-acetylmuramic acid. The fold is composed of an α/β/α sandwich with a 5-stranded parallel β-sheet. Despite having only five domains in the SCOPe95 set, this fold is the seventh most frequent one in Rossmann fold hits and, like c.2, hits many different folds and domains (Figure 23). Furthermore, most probabilities in c.2-c.5 hits are very high, with 58.7% of the hits displaying a probability higher than 95%.

Regarding fragment size, there are two trends: A smaller βαβ one (~35 residues), and a broad peak (~75-85) containing $(\beta\alpha)_3\beta$ and $(\beta\alpha)_4\beta$ fragments.

MurCD proteins bind UDP-N-acetylmuramic acid (UDP-MurNAc) and its derivatives. An assessment of ligand-binding residues present in the fragments shows that most of them are conserved in both folds. In MurCD domains, the fragment interacts with the UDP and the N-acetylmuramoyl moieties. In Rossmann fold domains, the equivalent residues interact with the whole nucleotide cofactor, making contacts with every group within it. UDP-interacting residues overlap with the ones associated with the adenosyl group, while MurNAc-binding residues correspond to nicotinamide-interacting ones (Figure 67). Additionally, the nucleotide moiety in UDP-MurNAc and the adenosyl one in NAD(P) have similar geometries.

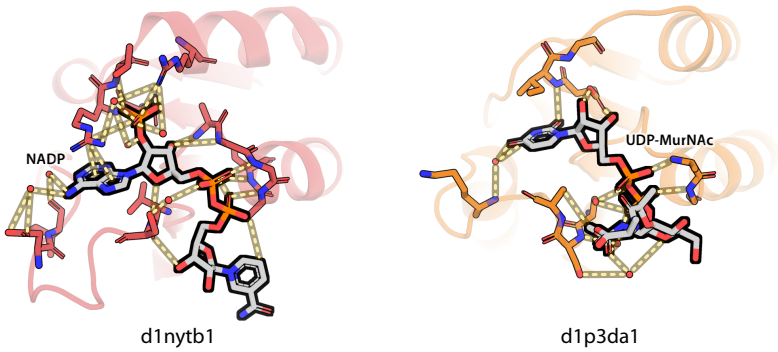


Figure 67: Conserved ligand interactions in c.2-c.5 hits. UMA: UDP-N-acetylmuramoyl-L-alanine.

Conserved fragments between Rossmann and Fld-like folds

In addition to the fold pairs mentioned earlier, flavodoxin-like domains also hit Rossmann fold proteins. The former is the eighth most frequent fold hitting Rossmann fold domains, despite both of them having different sheet orders (21345 versus 321456). This difference in topologies can be surmounted in different ways. Some hits omit in the alignment the third βα element from the Rossmann fold domain, being shown as an insertion (Figure 68A). Other domains, however, contain deletions or modified secondary structure motifs in the same region, allowing a more permissive alignment (Figure 68B).

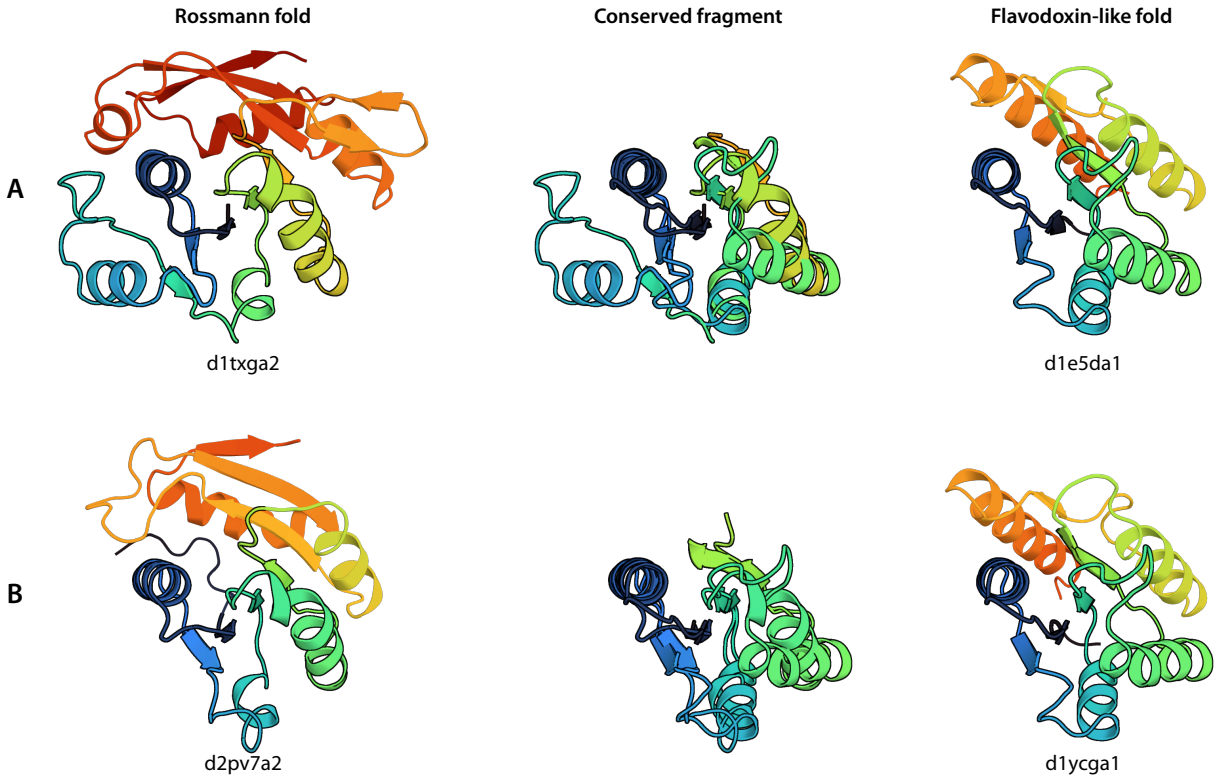


Figure 68: Representative Rossmann/Fld-like hits. N- and C-termini are colored blue and red respectively.

The extent of fragment sizes reveals that most conserved fragments carry around 55, 80 and 95 residues (Figure 69A). The former two lengths represent $(\beta\alpha)_3$ and $(\beta\alpha)_4$ fragments. The latter, on the other hand, and depending on the starting residues of the hit, can be described as $(\beta\alpha)_5$ or $\alpha(\beta\alpha)_4\beta$. Eleven Fld-like superfamilies contain hits with Rossmann fold domains, with flavoproteins (c.23.5) and CheY-like (c.23.1) being the most frequent ones (Figure 69B).

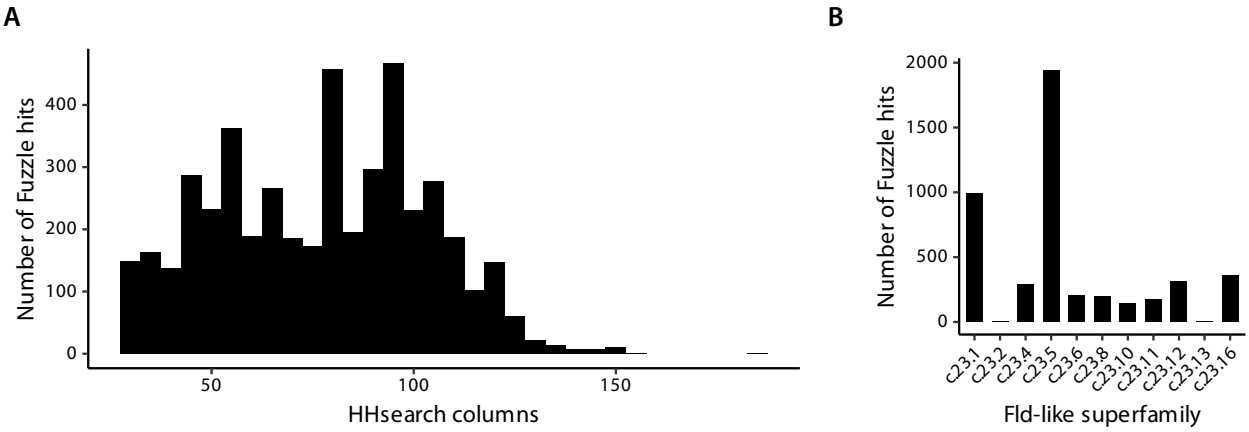


Figure 69: Fragment size (A) and superfamily (B) distribution in c.2-c.23 hits.

Regarding ligand binding, conserved interacting residues are distributed mostly among the loops following the fragments' strands and are dependent upon the Fld-like superfamily (Figure 70). The phosphate binding, Gly-rich loop of the Rossmann fold overlaps with phosphate-interacting residues in loop 1 of CheY-like (c.23.1), succinyl-CoA synthetase domains (c.23.4), flavoproteins (c.23.5), N^5-CAIR mutase superfamily (c.23.8), and formate/glycerate dehydrogenase catalytic domain-like (c.23.12). In the case of c.23.1 and c.23.4, the phosphate group is bound to specific conserved residues involved in function (aspartate and histidine, respectively). Most phosphate groups are located similarly in all these fragments. One particular exception is N^5-CAIR mutase. In this case the phosphate group of the substrate, AIR, is positioned in a different orientation compared to NAD. Phosphate-interacting residues remain, however, comparable. Conserved binding patterns are also present in cobalamin-binding domains (c.23.6), in a manner similar to what is seen in previously mentioned folds. Additional interactions between c.2 and c.23 are unveiled as well. For instance, in c.23.10 (SGNH hydrolase), the serine, glycine, and asparagine residues forming part of their namesake overlap with phosphate-interacting ones in Rossmann fold domains that utilize coenzyme-A. In class I glutamine amidotransferase-like domains (c.23.16) there are no conserved phosphate-binding interactions, but a catalytic cysteine residue overlaps with a Rossmann one that is binding the nicotinamide mononucleotidyl moiety.

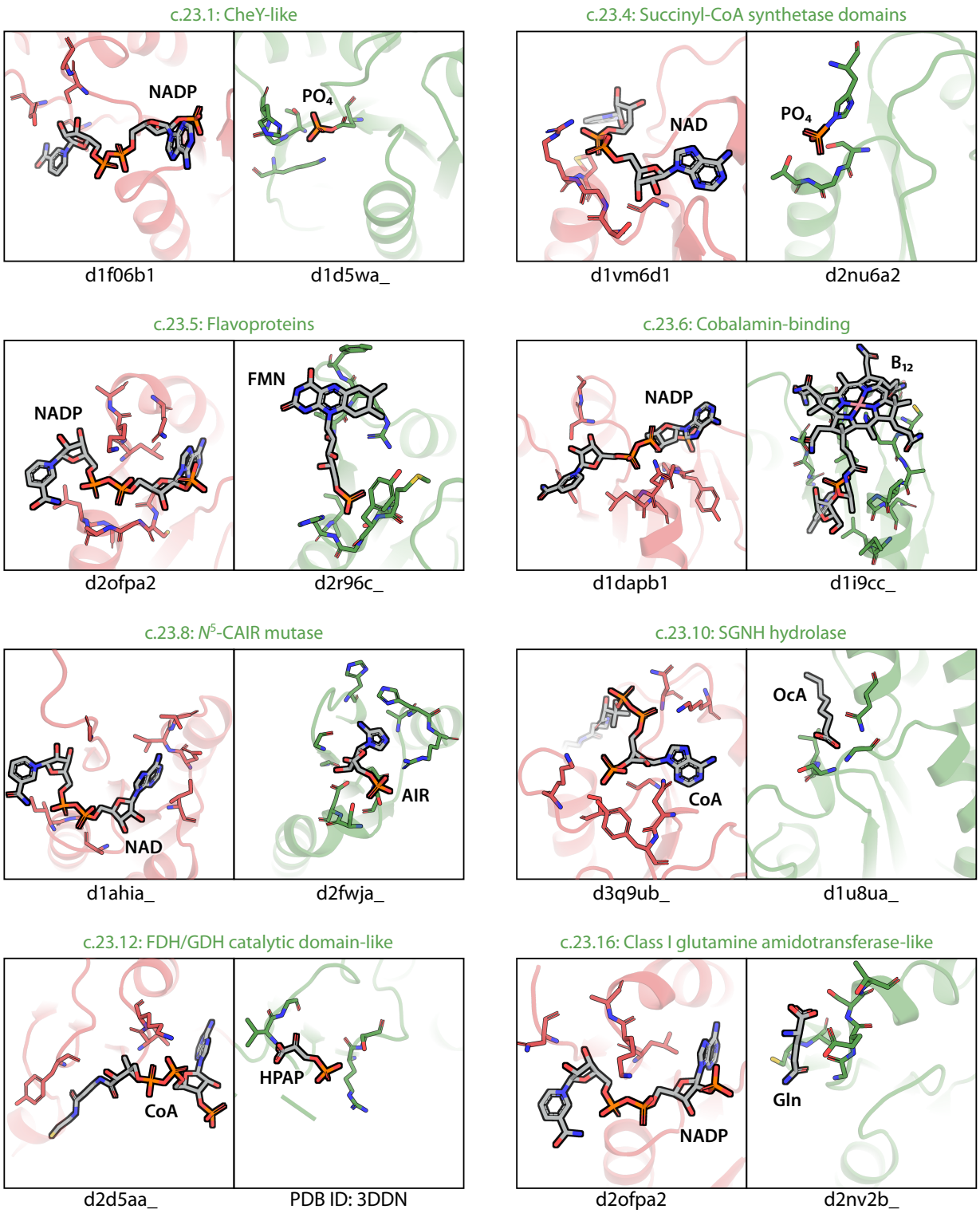


Figure 70: Conserved ligand-interacting residues in c.2-c.23 hits. OcA: Octanoic acid. HPAP: Hydroxypyruvic acid phosphate. Note that the 3DDN structure is a member of the c.23.12 superfamily but is not classified as a domain in SCOP.

Building a Rossmann fold/Fld-like chimera

SO FAR, WE HAVE FOCUSED on the characterization of remotely homologous fragments between domains belonging to different folds. However, one of the main goals of this PhD project was to develop functional chimeras through the recombination of ligand-binding domains from different folds, using their conserved fragments to guide their fusion (Figure 71). In order to add a layer of functionality, the orientation of the parental binding pockets in the resulting protein must be such that the ligand geometry allows the catalytic mechanism of a potential reaction between them to take place. The previous statement implies that the interaction sites from both parental domains will be merged, giving rise to a chimeric active site.

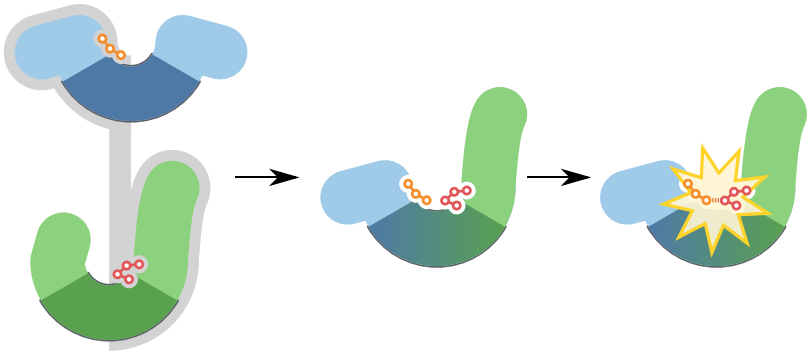


Figure 71: Concept for building functional chimeras by merging binding pockets. The conserved region between the green and blue domain as well as the recombination point lie within the darker-shaded regions.

Finding a conserved fragment and a suitable ligand pair is only the initial step. Several aspects must be addressed during the fragment picking and design process. First, ligands must be in close proximity but cannot overlap, as this will most likely result in steric clashes and unwanted binding modes for at least one of the substrates and/or a suboptimal geometry of the ligand-binding interface. Next, the structural complementarity between both polypeptide fragments must be as optimal as possible to maximize beneficial interactions and prevent unfavorable ones at their interface, which may result in a misfolded chimeric protein. Finally, the recombination site should lie in a segment with high structural similarity within the conserved fragment to avoid spatial rearrangements or unexpected conformations in the chimera. For this same reason steric clashes between regions outside of the fragment must be minimized.

With all these aspects in consideration, we performed an

exhaustive search within Fuzzle hits that could fulfill our criteria. One candidate set of hits is the one comprising NAD-binding Rossmann fold domains and the Fld-like flavoproteins (SCOP superfamily c.23.5, see page 103). Figure 72 depicts the hit between the Rossmann fold domain from *Haemophilus influenzae* TyrA (*Hi*TyrA) and *Desulfovibrio gigas* flavodoxin (*Dg*Fld).

*Hi*TyrA interacts with NAD, while *Dg*Fld binds FMN. Residues in both domains associated with phosphate groups are conserved (Figure 72B and C). There can be, however,

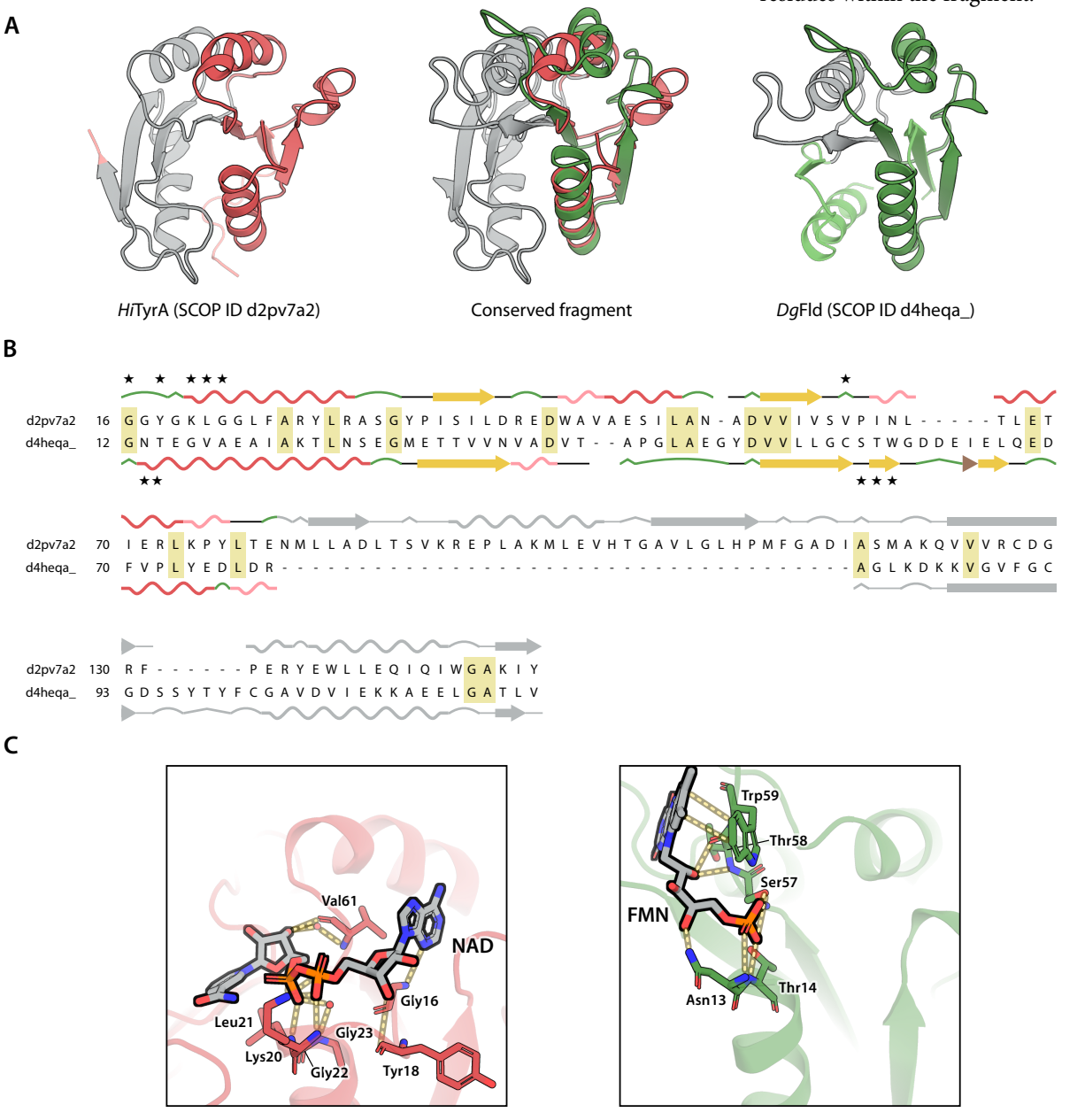


Figure 72: Fuzzle hit between *Hi*TyrA and *Dg*Fld. A: Structural superposition. B: HHsearch alignment. Residues involved in ligand binding are marked with a star. C: Conserved ligand-interacting residues within the fragment.

slight differences in the phosphate interaction patterns and geometry. For instance, the phosphate group in FMN lies between the two NAD ones, namely where the phosphoanhydride bond is positioned in the latter, and is further apart from the β-strand ends. In *Dg*Fld the phosphate-interacting loop and the C-terminal end of the adjacent helix are extended to reach the group in FMN. Residues further downstream within the fragment are also conserved and involved in binding. In *Hi*TyrA they associate mostly with the nicotinamide mononucleotidyl moiety of NAD. *Dg*Fld, on the other hand, bind the nucleoside subunit of FMN, that is, the ribityl and isoalloxazine group, in this region.

Looking at the relative position of the adenosine phosphate moiety of NAD and the FMN molecule (Figure 73A), their geometry and the distance between their phosphate groups suggest two reactions that might take place if the binding pockets present in both conserved fragments are combined. The first

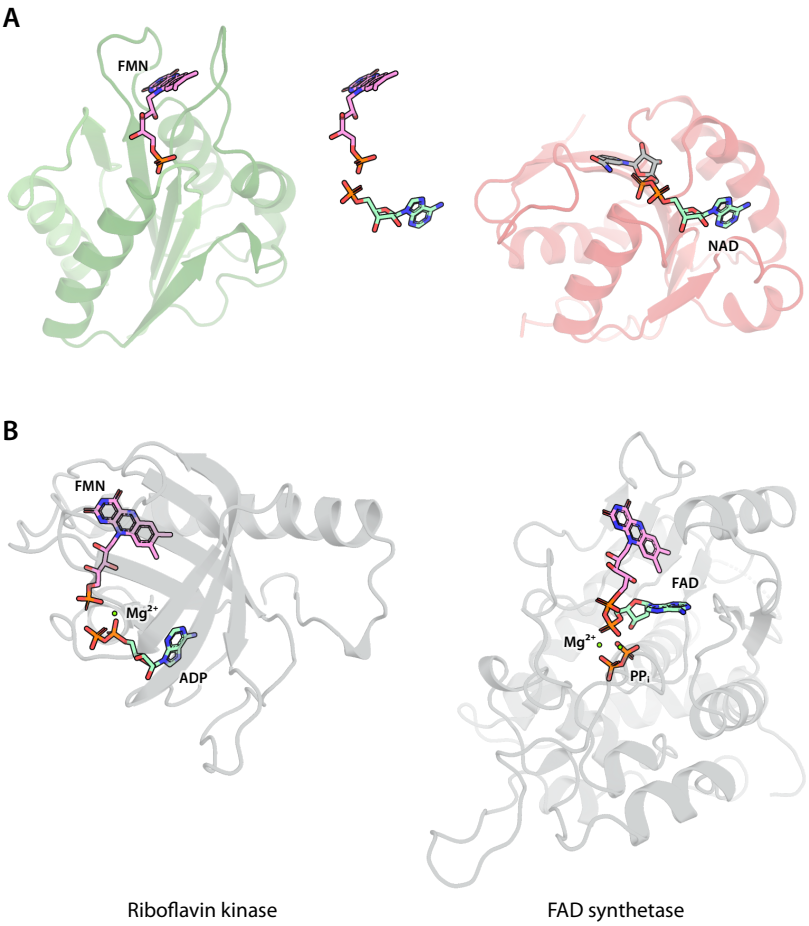


Figure 73: Ligand geometry in the *Hi*TyrA/*Dg*Fld hit. A: Combined ligand orientation of FMN (pink) and the adenosyl phosphate moiety of NAD (cyan) in the Fuzzle hit (see Figure 72). B: Orientation of the products in riboflavin kinase (PDB ID 5A8A, left) and FAD synthetase (PDB ID 3FWK, right).

one is the phosphorylation of riboflavin:

$$\text{Riboflavin} + \text{ATP} \longrightarrow \text{FMN} + \text{ADP}$$

The mechanism would occur via activation of the O5′ hydroxyl group of riboflavin and posterior phosphoryl transfer through a nucleophilic attack against the γ-phosphate group of ATP. An alternative, feasible reaction between the ligand pair is the adenylation of FMN:

$$\text{FMN} + \text{ATP} \longrightarrow \text{FAD} + \text{PP}_\text{i}$$

In this case the reaction mechanism consists of a nucleophilic attack by the 5′-phosphate of FMN on the α-phosphate of ATP, releasing FAD and pyrophosphate (PP_i) as products. The U-shaped substrate geometry and the aforementioned mechanisms are reminiscent of existing riboflavin kinases and FAD synthetases,[4] which additionally use magnesium ions as cofactor (Figure 73B). Taking these hits as a starting point, several Rossmann/Fld-like hits from Fuzzle were analyzed to generate computational models of their theoretical chimeras. A preliminary rigid-body docking test confirmed the complementarity of the chosen fragments, which assemble in a mode similar to a theoretical *Hi*TyrA/*Dg*Fld hybrid and interface at the purposed recombination point (Figure 74). A model of the *Hi*TyrA/*Dg*Fld chimera was built and optimized with Rosetta. After relaxing, it preserves the binding pockets from both parental domains and is able to dock a transition state analog for the proposed reaction.

[4] The latter is also known as FMN adenylyltransferase.

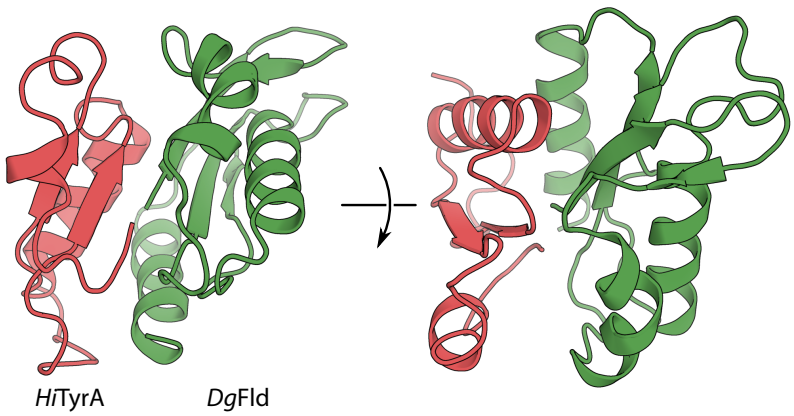


Figure 74: Rigid-body docking model of *Hi*TyrA and *Dg*Fld fragments.

Recombinant *Hi*TyrA/*Dg*Fld was expressed in *E. coli*, but soluble protein could only be recovered from inclusion bodies after denaturation and subsequent refolding. Analytical size

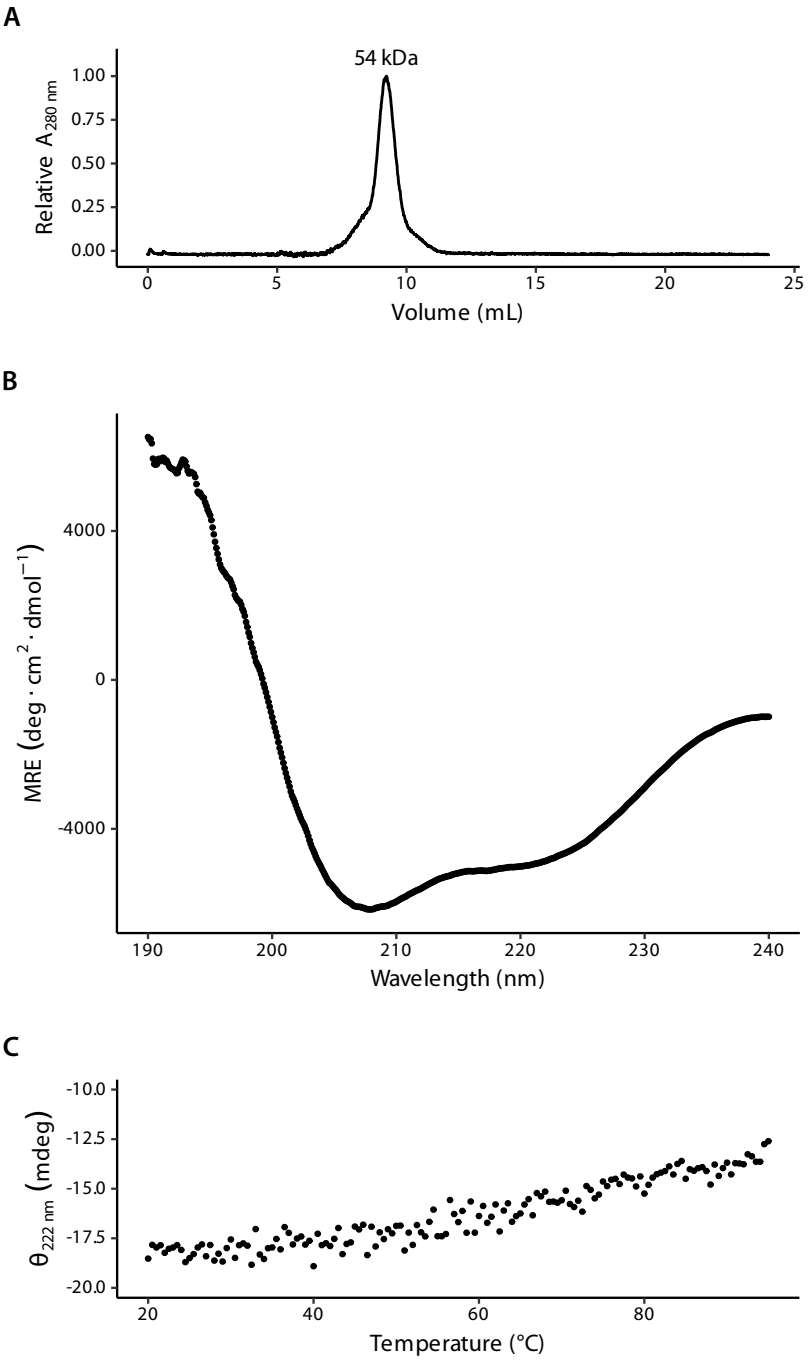


Figure 75: Biophysical characterization of *Hi*TyrA/*Dg*Fld. A: Analytical size exclusion chromatogram. B: Circular dichroism spectrum. C: Thermal melting curve.

exclusion reveals that most of the protein is present in a molecular weight corresponding to a trimer, but larger sizes can also be found (Figure 75A), hinting at the presence of protein aggregates. Secondary structure estimation by circular dichroism spectra showed that the chimera adopted a molten globule-like state (Figure 75B) and did not denature at high temperatures (Figure 75C).

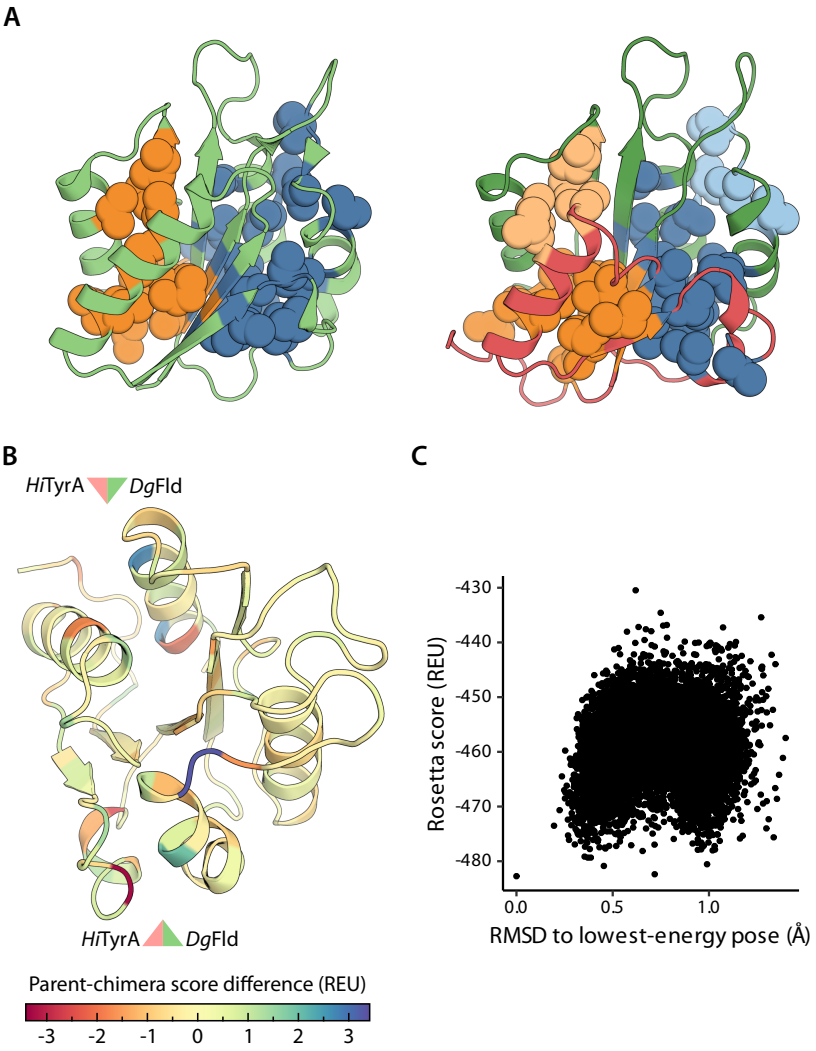


Figure 76: Issues in the *Hi*TyrA/*Dg*Fld model. A: Hydrophobic residue clusters in the parental *Dg*Fld structure (left) and the lowest-scoring chimera model (right). Differently colored spheres represent different clusters. B: Rosetta residue score differences between parent and chimera model structure. C: RMSD-score relationship in the *Hi*TyrA/*Dg*Fld Rosetta models.

Given this outcome, the computational model of the chimera was further analyzed. The fragments in it have isolated hydrophobic residue clusters separated at the interface, breaking the continuous clusters present in the Fld-like fold (Figure 76A). In addition, a residue energy breakdown and comparison between the parental domains and the *in silico* model showed issues in the fragment interface (Figure 76B). It is in this region where the largest score differences between native and chimera structures were found, hinting at detrimental interactions between both sections of the chimera. A look at the sampled structural space shows a bimodal distribution of conformational states, suggesting that the protein is unable to adopt a single, low-energy conformation *in vitro* (Figure 76C).

To solve this problem fixed-backbone sequence design was performed on TyrA/Fld. One of the resulting variants, TyrA/Fld3, expressed solubly without any additional aid but formed large complexes or aggregates. Attempts to concentrate the purified protein resulted in precipitation starting at values around 1 mg/mL. Several other designs with different parental domains were tested, and the most successful ones expressed solubly without any mutations (Table 9). They suffered, however, from the same aggregation issue as TyrA/Fld. This roadblock prevented further biophysical and structural characterization as well as the assessment of the cofactor binding properties of these chimeras. In light of these discouraging results, we expanded our search for additional fold pairs with suitable ligand-binding properties that may be combined into fold chimeras. They will be described in the next section.

Table 9: Tested Rossmann/Fld-like constructs.

Chimera name	Rossmann fold domain	Flavodoxin-like domain	Comments
*Hi*TyrA/*Dg*Fld	*Haemophilus influenzae* prephenate dehydrogenase (SCOP ID d2pv7a_)	*Desulfovibrio gigas* flavodoxin (d4heqa_)	Insoluble.
*Hi*TyrA/*Dg*Fld2	*H. influenzae* prephenate dehydrogenase	*D. gigas* flavodoxin	Variant with 6 mutations. Insoluble.
*Hi*TyrA/*Dg*Fld3	*H. influenzae* prephenate dehydrogenase	*D. gigas* flavodoxin	Variant with 7 mutations. Soluble, forms aggregates.
*Bs*KtrA/*Dg*Fld	*Bacillus subtilis* K^+-uptake protein (d2hmua_)	*D. gigas* flavodoxin	Insoluble.
*Bs*KtrA/*Dg*Fld2	*B. subtilis* K^+-uptake protein	*D. gigas* flavodoxin	Variant with 7 mutations. Insoluble.
*Tm*GDH/*Mt*FprA	*Thermotoga maritima* glycerol-3-phosphate dehydrogenase (d1z82a1)	*Moorella thermoacetica* nitric oxide reductase (d1ycga1)	Insoluble.
*Bm*MDH/*Mt*FprA	*Brucella melitensis* lactate/malate dehydrogenase (d3gvha1)	*M. thermoacetica* nitric oxide reductase	Insoluble.
*Bs*KtrA/*Pm*P450	*B. subtilis* K^+-uptake protein	*Priestia megaterium* cytochrome P450 BM-3 (d1bvya_)	Insoluble.
*Ss*TyrA/*Sc*LOT6	*Synechocystis sp.* prephenate dehydrogenase (d2f1ka2)	*Saccharomyces cerevisiae* NADP(H)-dependent FMN reductase (d1t0ia_)	Insoluble.
*Tm*GDH/*Dr*WrbA	*T. maritima* glycerol-3-phosphate dehydrogenase	*Deinococcus radiodurans* NAD(P)H dehydrogenase (d1ydga1)	Insoluble.
*Tm*GDH/*Dg*ROO	*T. maritima* glycerol-3-phosphate dehydrogenase	*D. gigas* rubredoxin-oxygen oxidoreductase (d1e5da_)	Insoluble.
*Af*FNO/*Dd*Fld	*Archaeoglobus fulgidus* F420-dependent NADP reductase (d1jaya_)	*Desulfovibrio desulfuricans* flavodoxin (d3f6ra_)	Soluble, forms aggregates.

Relationships between PurN and PurT in evolution and design

THE *DE NOVO* PURINE BIOSYNTHESIS PATHWAY generates inosine monophosphate (IMP) using phosphoribosyl pyrophosphate (PRPP) as the initial precursor.[5] IMP is in turn an intermediate for the synthesis of purine nucleotides, namely AMP and GMP. The third step of the *de novo* pathway, shown in Figure 77, involves the transfer of a formyl group to glycinamide ribonucleotide (GAR) to generate N-formylglycinamide ribonucleotide (FGAR). This reaction can be catalyzed by either of two enzymes: PurN (phosphoribosylglycinamide formyltransferase) and PurT (formate-dependent phosphoribosylglycinamide formyltransferase). The eukaryotic PurN ortholog, GART, is part of a trifunctional enzyme having PurD (GAR synthetase) and PurM (aminoimidazole ribonucleotide synthetase) activities as well.

[5] Zhang et al., 2008

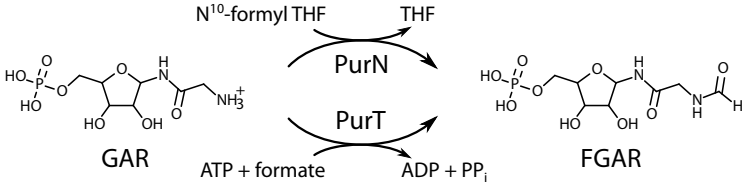


Figure 77: Reaction catalyzed by PurN and PurT.

It has been assumed that PurN and PurT are redundant, analogous enzymes and thus with no recognizable common ancestor. Several lines of evidence seem to support this hypothesis.

First, their distribution is heavily dependent on taxonomy (Figure 78), with eukaryotic organisms relying solely on GART, archaeal species having either PurN or PurT in their genome, and different bacterial classes having only PurN or a combination of both genes.

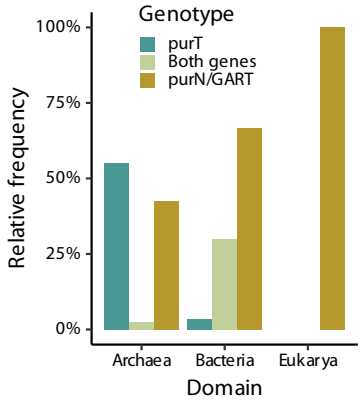


Figure 78: purN and purT genotypes in different species, according to entries in UniProt.

Second, regarding their structure, PurN and PurT present different domain arrangements (Figure 79). The PurN enzyme is composed of a single domain adopting a three-layered formyltransferase fold (SCOP ID c.65). This fold has a mixed β-sheet in a 3214567 order, with the sixth strand being antiparallel to the others. On the other hand, PurT has a multidomain architecture involving an α/β preATP-grasp (SCOP fold c.30), an α+β ATP-grasp (d.142) participating in catalysis, and an all-

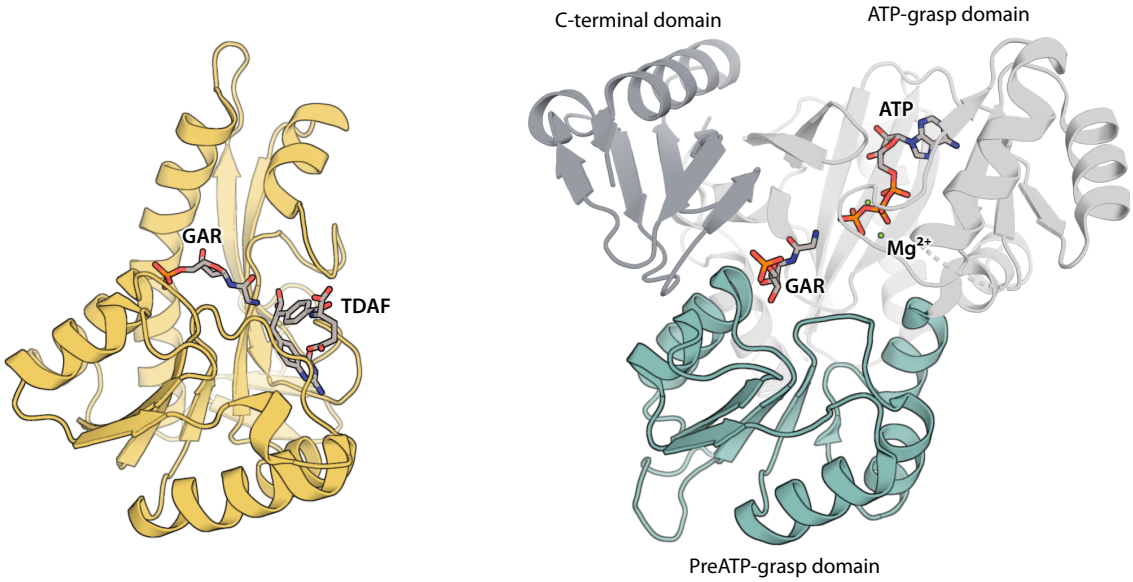


Figure 79: Structures of *Escherichia coli* PurN (left, PDB ID 1C2T) and PurT (right, PDB ID 1KJ8) in their holo forms. The formyltransferase and the preATP-grasp domains are depicted in gold and teal, respectively. TDAF: 5,8,10-trideazafolic acid, a cofactor analog.

β C-terminal (b.84) domain. The preATP-grasp fold is three-layered (α/β/α) and can have parallel or mixed β-sheets of 4 to 6 strands, depending on the enzyme family.

Third, the formylation reaction in PurN and PurT takes place with different participating cofactors and under different catalytic mechanisms.[6] The former uses 10-formyl-tetrahydrofolate (10-fTHF) as a formyl group donor, while the latter carries out the transfer through an ATP-activated acyl-phosphate intermediate,[7] with a formate group provided by PurU (formyltetrahydrofolate deformylase). The formyltransferase domain of PurN binds the GAR substrate and the folate-based cofactor in close proximity, allowing the former to perform a nucleophilic attack in the latter.

Despite these precedents, Fuzzle has several hits between preATP-grasp and formyltransferase-like domains, including the ones present in PurN and PurT. The conserved fragment found between *Escherichia coli* PurN and PurT has a $(\beta\alpha)_4\beta$ topology (Figure 80A) and comprises the N-terminal half of PurN and the majority of PurT's preATP-grasp domain. Both domain fragments have GAR in their proximity, suggesting the presence of residues involved in substrate binding. Additionally, the PurN fragment interacts with N^{10}-fTHF. An analysis of structure-ligand interactions with PLIP showed that the position of GAR-interacting residues is conserved in the sequence alignment (Figure 80B).

[6] Warren et al., 1996

[7] Marolewski et al., 1997

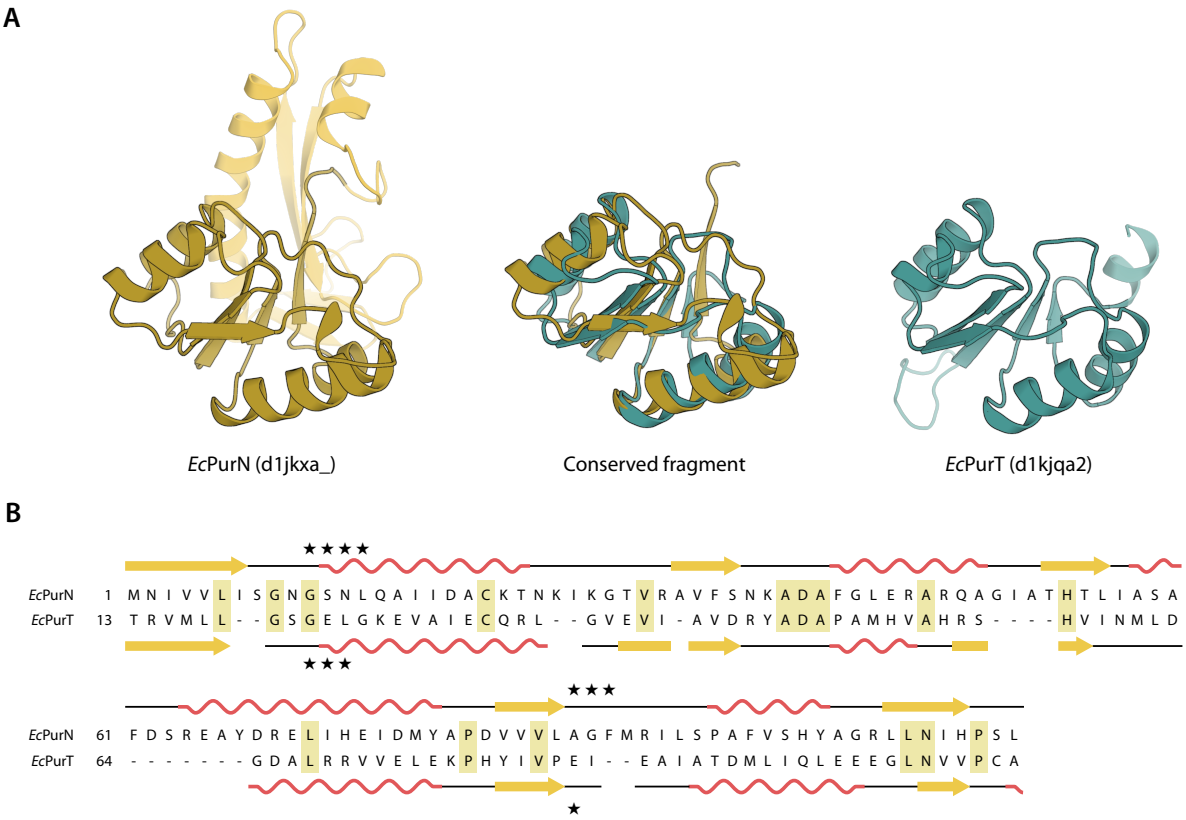


Figure 80: Fuzzle hit between *Escherichia coli* PurN and PurT. A: Structural superposition of the fragments. B: HHsearch alignment of the hit. GAR-interacting residues are marked with a star.

While both PurN and PurT fragments interact with GAR using conserved residues, the binding mode and the nature of the interactions are different (Figure 81). The PurN motif resembles the phosphate-interacting, glycine-rich one found in dinucleotide-binding Rossmann fold domains. Residues present in the fragment interact with the phosphate and glycinamide moieties of GAR, while Glu173, located outside the fragment, establishes hydrogen bonds with the hydroxyl groups of the ribose.

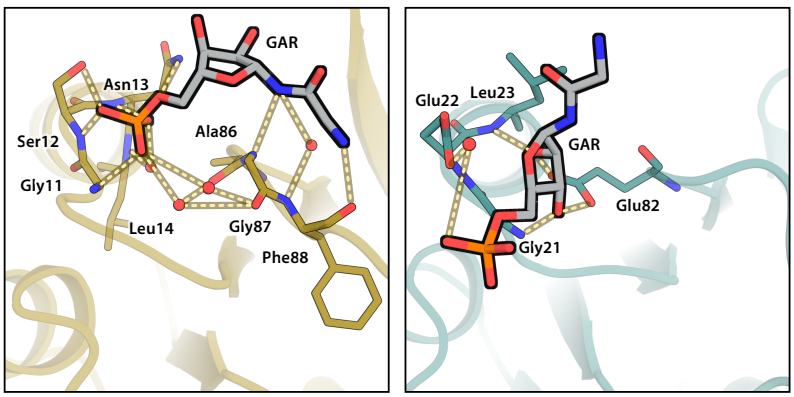


Figure 81: GAR-protein interactions within the conserved *E. coli* PurN (left, SCOP ID d1kj8a1) and PurT (right, SCOP ID d1c2ta_) fragments.

In contrast, the preATP-grasp domain fragment in PurT contains only three residues (Gly21, Glu22, Leu23, and Glu81) that interact directly with GAR, forming hydrogen bonds exclusively with the ribose hydroxyl groups. The corresponding residues in PurN interact with the phosphate group, but in PurT the Leu23 side chain takes the place where phosphate is supposed to be located, occluding said space and preventing any PurN-like phosphate-protein interaction. Despite this loss of key interactions, the GAR-binding capabilities of PurT are kept through additional interactions between the full enzyme and the glycinamide and phosphate groups. The residues responsible for the compensating interactions are located outside the conserved fragment, mainly within the all-β C-terminal domain that comes after the ATP-grasp. Overall, these residues allow the substrate to stay bound in a similar region of the fragment relative to PurN, albeit in a different orientation.

These insights lead to the design of chimeras between PurN and PurT domains, where fragments from both domains would be grafted in a way that exchanges the GAR binding modes and reaction mechanisms (Figure 82). Despite the different sub-

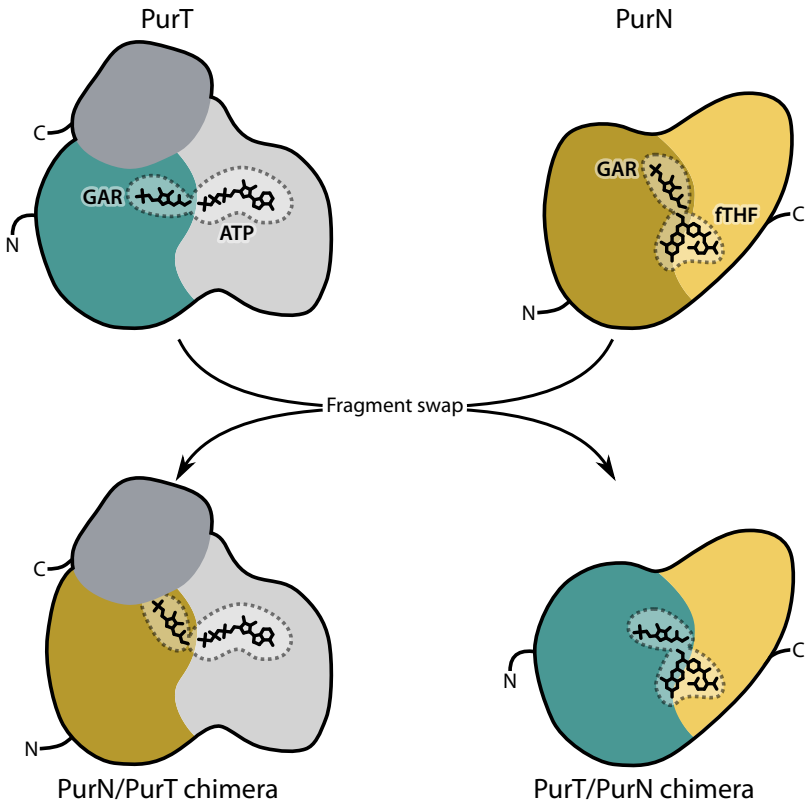


Figure 82: PurN/PurT chimera design strategy.

strate orientation, a formylation reaction may still be possible if the chimeras adopt the fold of one of their parental structures. Moreover, if the relative geometry between substrate and cofactor is favorable, the reaction should be feasible as well. Based on several c.30-c.65 hits present in Fuzzle, a set of PurT/PurN (N-terminal PurT, C-terminal PurN) and PurN/PurT (N-terminal PurN, C-terminal PurT) chimeras were designed *in silico* (Table 10). One PurT preATP-grasp domain (from *Escherichia coli*) and three PurN homologs (from *Escherichia coli*, *Bacillus halodurans*, and *Geobacillus kaustophilus*) were used as fragment donors, in addition to *Bacillus anthracis* methionyl-tRNA formyltransferase which adopts the c.65 fold as well but formylates a different substrate, the methionyl group present in methionyl-tRNA.

Table 10: Designed PurT ↔ PurN chimeras.

Chimera set	**PreATP-grasp domain**	**Formyltransferase domain**
*Ec*PurT ↔ *Ec*PurN	*Escherichia coli* PurT	*Escherichia coli* PurN
*Ec*PurT ↔ *Bh*PurN	*Escherichia coli* PurT	*Bacillus halodurans* PurN
*Ec*PurT ↔ *Gk*PurN	*Escherichia coli* PurT	*Geobacillus kaustophilus* PurN
*Ec*PurT ↔ *Ba*Fmt	*Escherichia coli* PurT	*Bacillus anthracis* Met-tRNA formyltransferase

Learning from the Rossmann/Fld cases, special care was taken in building and evaluating the resulting computational chimera models. All models adopted the target parental fold and were able to recover the hydrophobic residue clusters present in the target parental fold. In addition, energy issues in the fragment interface were minimal, with no large detrimental score differences, and the sampled conformational landscape showed no local minima, i.e., no alternate conformations. Figure 83 illustrates the evaluation of *Ec*PurT/*Bh*PurN as an example.

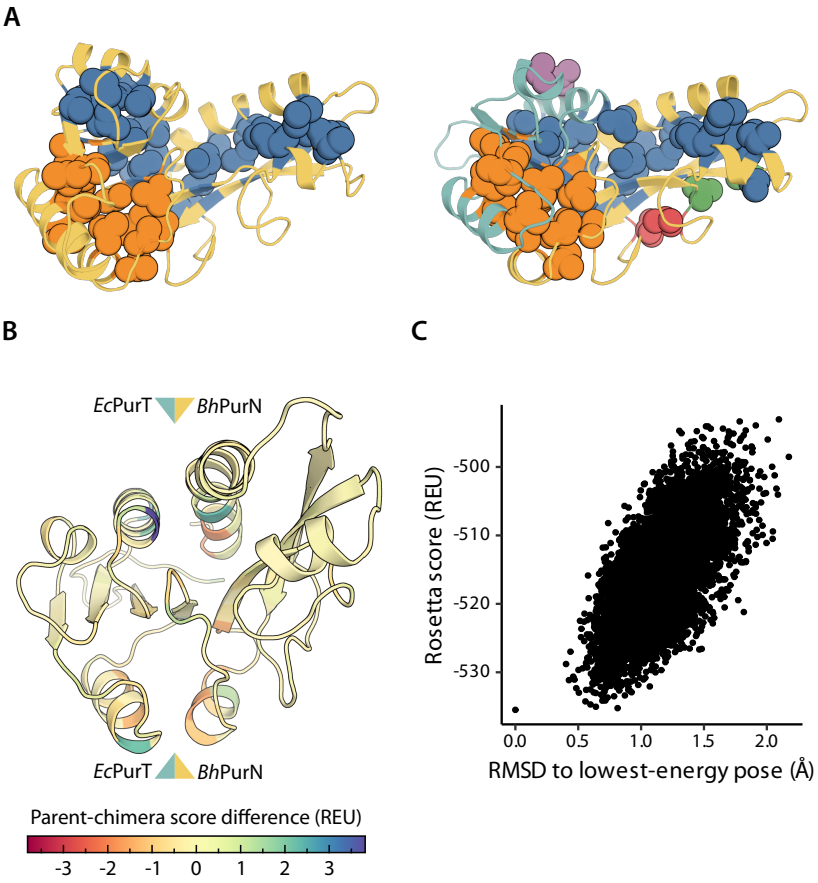


Figure 83: *Ec*PurT/*Bh*PurN model evaluation. A: Hydrophobic clusters in the parental *Bh*PurN structure (left) and the lowest-scoring chimera model (right). B: Rosetta residue score differences between parent and chimera model structure. C: RMSD-score relationship in the *Ec*PurT/*Bh*PurN Rosetta models.

After the computational evaluation of the designs, test expressions of all eight constructs were carried out. Two of them, *Ec*PurT/*Gk*PurN and *Ec*PurT/*Bh*PurN, could be expressed solubly in *Escherichia coli* (Table 11). Unfortunately, none of the PurN/PurT chimeras were soluble. Of the two soluble chimeras, *Ec*PurT/*Bh*PurN was the best-behaved one and therefore most of the characterization was focused on it.

Table 11: Tested PurT ↔ PurN constructs.

Chimera name	N-terminal fragment	C-terminal fragment	Solubility
*Ec*PurT/*Ec*PurN	*E. coli* PurT (SCOP ID d1kjqa1)	*E. coli* PurN (d1jkxa_)	Insoluble
*Ec*PurT/*Bh*PurN	*E. coli* PurT	*B. halodurans* PurN (d3p9xa1)	Soluble
*Ec*PurT/*Gk*PurN	*E. coli* PurT	*G. kaustophilus* PurN (d3av3a1)	Soluble
*Ec*PurT/*Ba*Fmt	*E. coli* PurT	*B. anthracis* Met-tRNA formyltransferase (d3rfoa1)	Insoluble
*Ec*PurN/*Ec*PurT	*E. coli* PurN	*E. coli* PurT	Insoluble
*Bh*PurN/*Ec*PurT	*B. halodurans* PurN	*E. coli* PurT	Insoluble
*Gk*PurN/*Ec*PurT	*G. kaustophilus* PurN	*E. coli* PurT	Insoluble
*Ba*Fmt/*Ec*PurT	*B. anthracis* Met-tRNA formyltransferase	*E. coli* PurT	Insoluble

Purified *Ec*PurT/*Bh*PurN was characterized experimentally. Analytical size exclusion chromatography showed a single peak representing a dimer in solution. As shown by the far-UV CD spectrum (Figure 84A), *Ec*PurT/*Bh*PurN is a folded protein, containing both α-helices and β-strands in its secondary structure profile. The chimera's T_m, as determined by a thermal melt assay, corresponds to 53.6°C, which is close to the measured T_m of the parental *Bh*PurN, 54.4°C. The same assay showed that the protein suffers irreversible unfolding upon thermal denaturation.

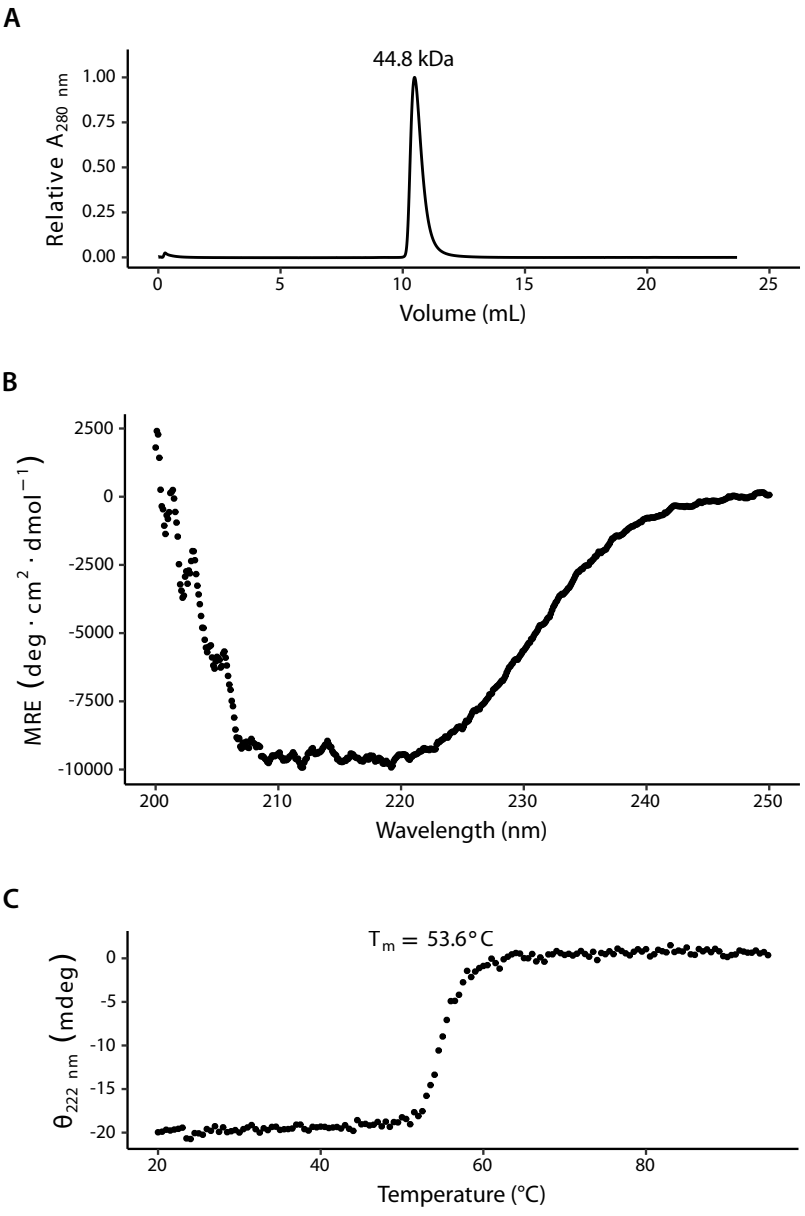


Figure 84: Biophysical characterization of *Bh*PurT/*Ec*PurN. A: Analytical size exclusion chromatogram. B: Circular dichroism spectrum. C: Thermal melting curve.

Besides biophysical characterization, crystallization of *Ec*PurT/*Bh*PurN was achieved and a preliminary crystallographic structure of PurN/PurT was solved for residues 12 to 188 (Figure 85A, Table 12). *Ec*PurT/*Bh*PurN exhibits a PurN-like conformation, reconstructing successfully the formyltransferase fold. It resembles closely the theoretical Rosetta model as well (Figure 85B), with a full-atom RMSD value of 2.7 Å (Cα RMSD: 2.1 Å). RMSD values between the isolated parental fragments and their corresponding region in the crystallographic structure were lower (Figure 85C), hinting at slight differences in the relative orientations of the recombined fragments. The regions with the highest structural divergence between model and experimental structure are located in helices 3 and 4, along with their preceding loops, and loop 6 (Figure 85D). Helices 3 and 4 are located at one side of the fragment interface. The remaining pair adopt, however, the conformation present in the Rosetta model.

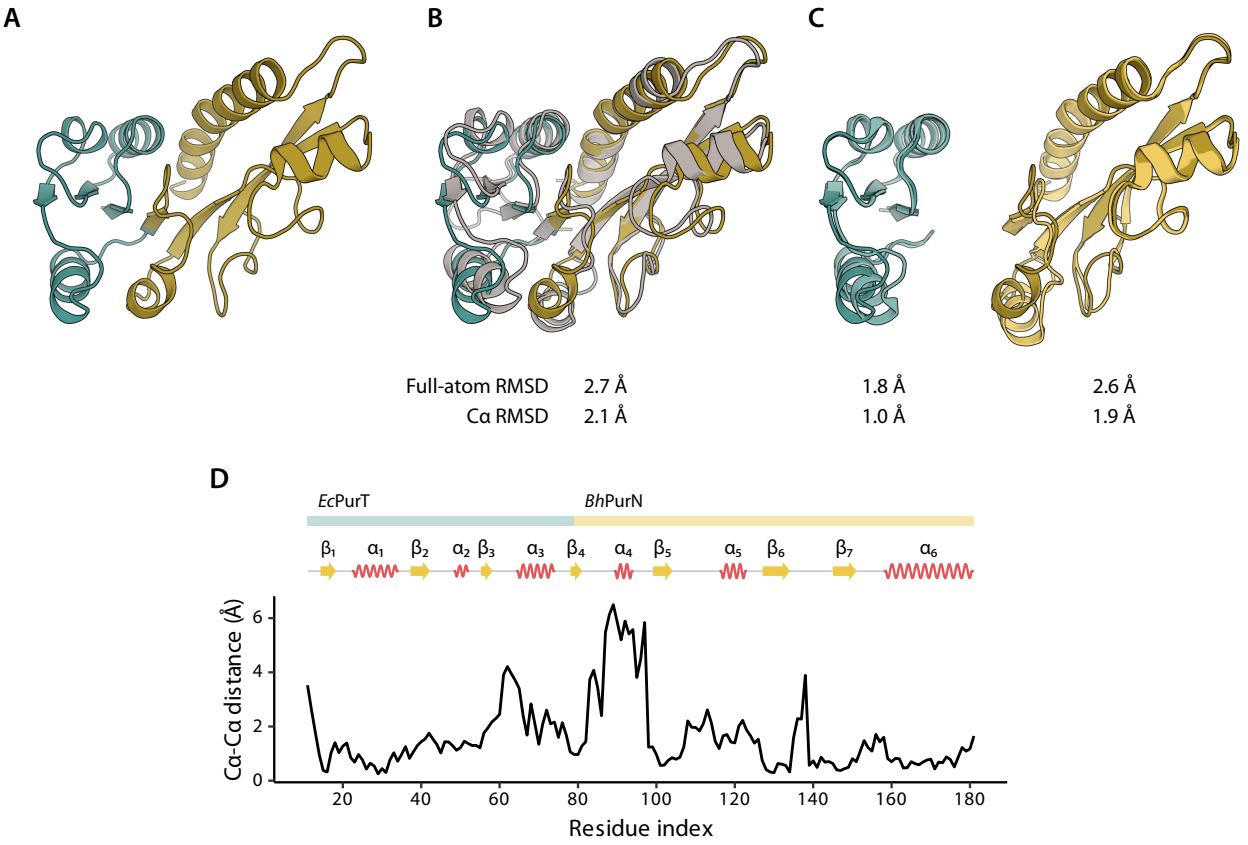


Figure 85: Crystallographic structure of EcPurT/BhPurT. A: Structure B: Superposition with Rosetta model (in gray) C: Separate superpositions with wild-type fragments (in lighter shades). D: Cα-Cα distances between the solved structure and the Rosetta model, according to the superposition in B.

Table 12: Data collection and refinement statistics for the preliminary *Ec*PurT/*Bh*PurN model.

Wavelength (nm)	0.9184
Resolution range (Å)	22.66 - 2.12 (2.196 - 2.12)
Space group	P 3_1 2 1
Unit cell	81.631 81.631 59.062 90 90 120
Total reflections	143552 (14514)
Unique reflections	13192 (1300)
Multiplicity	10.9 (11.2)
Completeness (%)	99.77 (99.69)
Mean I/σ(I)	13.22 (0.80)
Wilson B-factor	49.92
R_{merge}	0.1245 (2.936)
R_{meas}	0.1307 (3.077)
R_{pim}	0.03958 (0.9145)
$CC_{½}$	0.999 (0.229)
CC^{*}	1 (0.611)
Reflections used in refinement	13186 (1300)
Reflections used for R-free	660 (65)
R-work	0.2116 (0.3268)
R-free	0.2580 (0.3666)
CC_{work}	0.966 (0.535)
CC_{free}	0.957 (0.420)
Number of non-hydrogen atoms	1375
macromolecules	1312
ligands	16
solvent	47
Protein residues	171
$RMS_{lengths}$ (Å)	0.006
RMS_{angles} (°)	0.93
Ramachandran favored (%)	91.12
Ramachandran allowed (%)	5.92
Ramachandran outliers (%)	2.96
Rotamer outliers (%)	0.00
Clashscore	11.15
Average B-factor	67.51
macromolecules	67.13
ligands	96.53
solvent	68.17
Number of TLS groups	1

Statistics for the highest-resolution shell are shown in parentheses.

Evolutionary relationships between the c.30 and c.65 folds

To further characterize the evolutionary relationships within the preATP-grasp and the formyltransferase folds, a clustering analysis involving all their inter- and intrafold hits was carried out (Figure 86). While c.65 comprises a single family of proteins, c.30 contains nine different ones. Even at a hit P-value threshold of 10^{-5}, most preATP-grasp fold families do not cluster together, with few hits between them. The family containing PurT's preATP-grasp domain, biotin carboxylase N-terminal domain-like (SCOP ID c.30.1.1), is connected to three other c.30 families: D-alanine ligase N-terminal domain (c.30.1.2), prokaryotic glutathione synthetase, N-terminal domain (c.30.1.3), and lysine biosynthesis enzyme LysX, N-terminal domain (c.30.1.6). Within c.30.1.1 domains, PurT lies closer to GAR synthetase (GAR-syn, PurD). Formyltransferase domains, on the other hand, hit exclusively the c.30.1.1 family and their conserved fragments bear higher similarity to PurD and PurT.

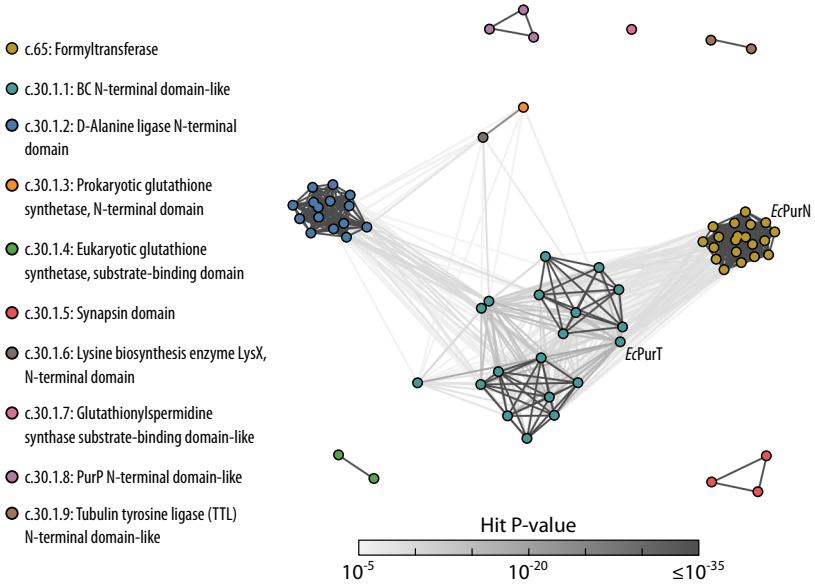


Figure 86: Cluster map of c.30-c.65 hits. BC: biotin carboxylase.

Sequence intermediates between c.30 and c.65

The identification of intermediate fold fragments that share features with both preATP-grasp and formyltransferase domains can be helpful to clarify their evolutionary history. Therefore,

hits involving domains that connect both c.30 and c.65 domains were searched in Fuzzle. A clustering analysis of them (Figure 87) found intermediate domains stemming from five folds: Rossmann (c.2), ATC-like (aspartate/ornithine carbamoyl transferase, c.78), UDP-glycosyltransferase/glycogen phosphorylase (c.87), chelatase-like (c.92), and CAC2185-like (c.147). Similar to the c.30-c.65 clustering, preATP-grasp domain families are not grouped together, with different ones forming distinct clusters. The PurT domain lies very close to Rossmann fold ones. Unlike the previous clustering effort, the PurN family is separated from other c.65 families, the latter ones lying closer to the Rossmann fold. The distance between both c.65 groups is, though, not large.

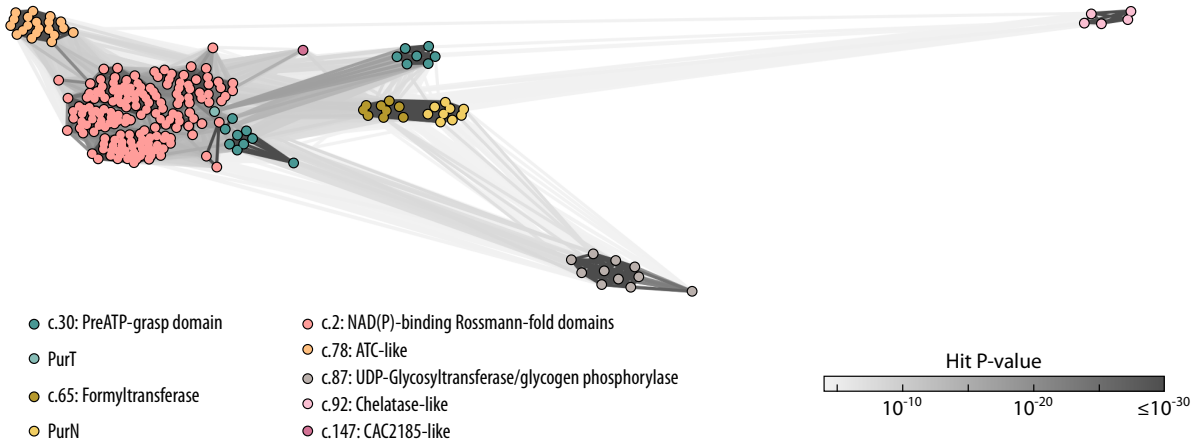


Figure 87: Cluster map of c.30-c.65 and intermediate hits. BC: biotin carboxylase; CPS: carbamoyl phosphate synthase

As shown before, the phosphate-interacting loop from Rossmann folds has conserved residues involved in interactions with the ribose from the adenosyl group, all phosphate groups, including the 2′-phosphate, and the nicotinamide group (Figure 88). The ligand-interacting regions in c.2-c.30 fragments are structurally more similar than the ones found in c.2-c.65 hits. In particular, there is a loop extension in the formyltransferase fold that is absent in both Rossmann and preATP-grasp domains. The nature of the ligand-protein interactions and the geometry of the nucleotides being bound in c.2-c.30 hits, however, seem to be less alike than the c.2-c.65 case, where the interactions with phosphate groups are emphasized. In ATC-like domains, conserved residues interact with the carboxyl group of the amino acid ligand. In the c.87 fold, which binds nucleotide-linked sugars, the fragment interacts

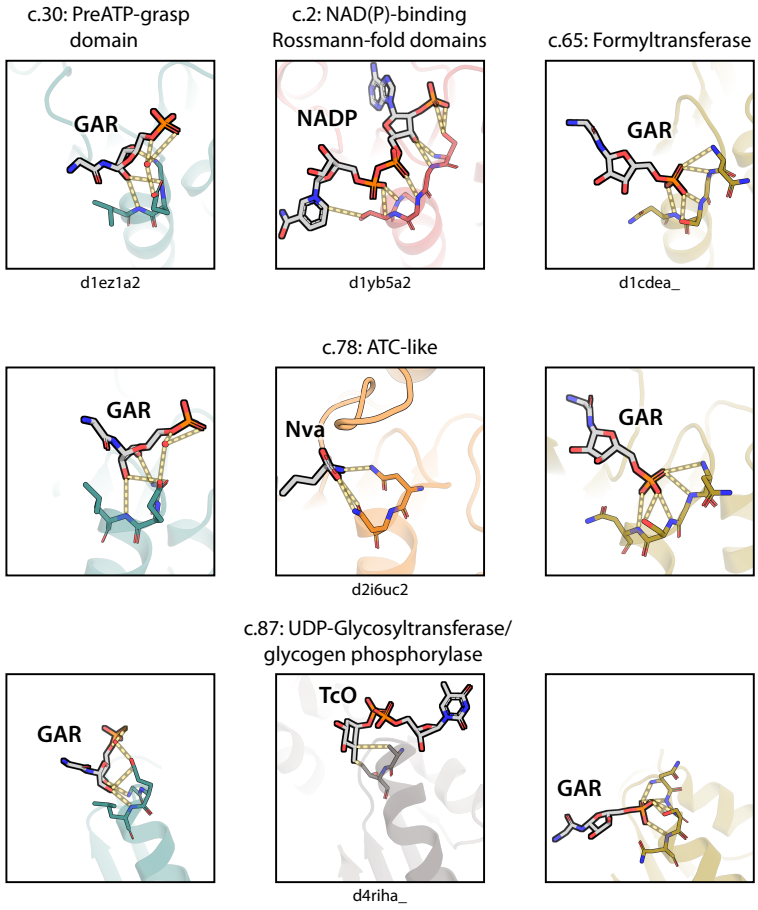


Figure 88: Conserved ligand-interacting residues between c.30, c.65 and intermediate fold hits. The largest combined intermediate fragment between c.30 and c.65 is shown in each middle panel. Nva: L-norvaline.

exclusively with the sugar moiety.

The c.147 fragment, comprising the N-terminal region of the hypothetical protein CAC2185, has no known ligand. Despite this absence, the fragment lies between c.2, c.30, and c.65 in the cluster map. Structurally speaking, however, the fragment bears higher similarity to c.2 and PurT-like c.30 fragments (Figure 89), suggesting a nucleotide binding role.

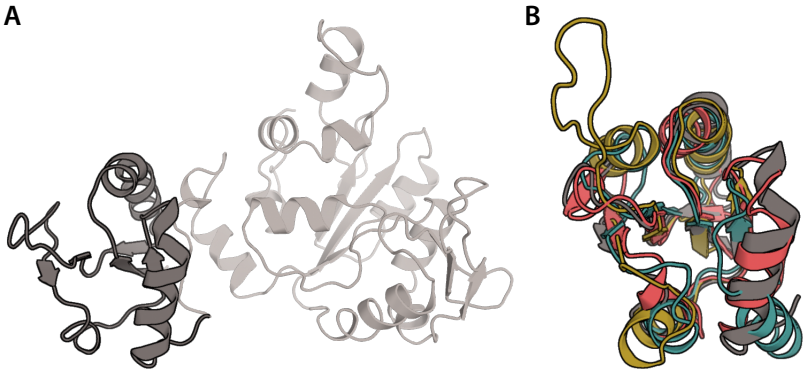


Figure 89: Fuzzle fragment shared between c.30, c.65, and c.147 folds. A: Fragment in the context of the whole c.147 domain (hypothetical protein CAC2185 from *Clostridium acetobutylicum*, SCOP ID d2g6ta1) B: Structural superposition of the CAC2185-like fragment with preATP-grasp (SCOP ID d1kjqa2), formyltransferase (d3rfoa1), and Rossmann fold (d2dt5a2) domains.

Discussion

THE WORK PRESENTED IN THIS DISSERTATION aims to increase our understanding of two aspects involving protein innovation: First, the way protein innovation occurred over geological timescales, and second, how Nature's ability for innovation might be harnessed and applied to protein design. One element situated at the intersection of both subjects is the presence of conserved, ancestral subdomain fragments. Moreover, analysis of the ligands associated to the fragments found regions involved in binding that are conserved in sequence and structure.

Fragments in Fuzzle

Fuzzle is a large database containing roughly 8.1×10^6 hits between different domains. This amount is, however, two orders of magnitude lower than the theoretical maximum, suggesting that the matching between different domains did not occur in an indiscriminate way. Most of these hits correspond to domain pairs within the same fold, but hits between domains from different folds and low, yet statistically significant similarity are also present within the database.

The 2.06 version of SCOPe 95 could hold theoretically $28\,010^2 \approx 7.84 \times 10^8$ hits.

Despite the common notion that structure is more conserved than sequence, most hits in Fuzzle have a higher count of conserved residues in their HHsearch alignments compared to the TM-align ones (Figure 18). This disparity can be seen in TIM barrel/Fld-like hits, where some terminal helices face in different directions (Figure 32). Other cases are more extreme, such as the one seen in some TIM barrel/PBP-like I hits, where the structural superposition is largely unrelated to the sequence alignment (Figure 43). In such instances, however, the secondary structure assigned in the sequence can remain aligned, hinting at the existence of conformational flexibility

in the conserved fragments. During the evolution of each fold, this variability could have been enforced by the different physical contexts that may constrain the folding space, resulting, for instance, in structural dislocations such as the one present in PBP-like I domains.

Fragment duplication in fold development

Fuzzle has instances of duplicated hits within and between certain folds. Some of them, such as α-helical repeat and β-propeller folds (Figure 25, have been already extensively characterized, but other, unexpected fold pairs were also found. Included in this group are TIM barrel/Fld-like, TIM barrel/PBP-like I, Fld-like/PBP-like I fragments. Said hits give support to a vision of domain development centered on duplication, where replication of short fragments gave rise to the precursor of present domain folds.

Other less evident evolutionary relationships can be found in Fuzzle as well. The third most frequent fold hitting TIM barrel domains is from a different structural class, the all-β composite domain of metallo-dependent hydrolases (b.92). Domains from these folds within a PDB chain can hit with each other with a high HHsearch probability, despite having little structural similarity (Figure 29). This finding is suggestive of an early duplication event in the evolution of these hydrolases, ending up with the two-domain architecture we see nowadays. Protein repeat identification methods such as HHrepID[1] did not find any kind of repeats within any whole two-domain metallo-dependent hydrolase chain.

[1] Biegert & Söding, 2008

Reexamination of previously described fragments

As expected, fold fragments analyzed earlier by our group are present in Fuzzle. We focused and elaborated in two of such cases: TIM barrel/Fld-like, and Fld-like/PBP-like I. Additionally, we found fragments between aldolase-like TIM barrels and PBP-like I, and therefore all three folds can be clustered together. In all these examples duplicated hits within a domain pair could be found. As shown in previous works,[2] most Fld-like fold superfamilies are not very close with each other.

[2] Farías-Rico et al., 2014

Furthermore, including the c.93 fold in the clustering does not seem to change these dissimilarities, further hinting at multiple emergence events of the Fld-like fold in its evolutionary history. This outcome implies also that the grouping of the fold performed by SCOP is based mostly on shared structural, i.e. topological, features. Regarding ligand binding, conserved features of the fragments in the group are varied and superfamily specific (Figure 34 and 48). In the examples mentioned above, phosphate-binding motifs are preserved in TIM barrel halves as well as Fld-like fragments. The duplication allows the former to bind ligands with two phosphate groups such as ProFAR. In the TIM barrel-PBP-like I case, on the other hand, ligand-binding residues are conserved only in sequence due to the loss of structural similarity (Figure 45). Conserved fragments between Fld-like and PBP-like I domains, by contrast, can be conserved in both sequence and structure as well, depending on the lobe the Fld-like domain is hitting the PBP-like I one. Overall, these results expand the scope of this ancestral fragment and contribute the understanding behind the evolutionary relationships that these three folds have.

The Rossmann fold as a main contributor of fragments

If we look away from the c.1, c.23, and c.93 folds and get a more global view of the protein universe we are able to detect additional dominant fragments within it. Rossmann fold domains are present in the large majority of Fuzzle hits, namely more than 60% of them. It also hits, on average, a large number of different folds, being only surpassed by c.5 and c.127 (Figure 23). Among the folds hit by the Rossmann fold are several ones, like c.2 itself, considered ancestral, including the Fld-like and P-loop-containing folds such as the c.37 NTPases. The Rossmann fold fragment shared with the other folds are usually N-terminal $(\beta\alpha)_n$ motifs containing the conserved phosphate-binding loop. The N-terminal region of Rossmann fold proteins have been studied in depth recently[3] and has been described as containing an ancient fingerprint at the end of the second β-strand. In it, a highly conserved aspartic acid residue interacts with the ribose moiety in the adenosyl group

[3] Laurino et al., 2016

of Rossmann-related cofactors. Furthermore, the geometry of the interaction is distinct and absent in other folds with similar topology and ligands, such as P-loop NTPases. A follow-up study proposed that both folds could have emerged from a common ancestral βαβ segment.[4] Most Rossmann fold hits are with folds considered Rossmann-like or Rossmannoids, such as c.3 and c.66. The Fld-like fold is sometimes considered Rossmannoid as well, but its topology is different than the Rossmann fold, having a 21345 strand order as opposed to a 321456 one. Nevertheless, a study suggested an evolutionary relationship between Rossmann fold domains and the CheY-like family of Fld-like ones, based on core packing conserved motifs.[5]

[4] Longo et al., 2020

[5] Cammer & Carter, 2010

Role of subdomain fragments in overall fold development

Most studies on domain fragments have focused on finding and classifying them, but little has been discussed about the role and properties of said fragments. For instance, what is actually encoded in them? Given their pervasiveness, it should be something that has been, at least marginally, positively selected over evolution. Structural stability, for instance, is one reasonable possibility. This work focused on the ligand-binding capabilities of such fragments, showing that they usually contained conserved residues associated to ligands, in particular known cofactors.

Besides ligand binding, another properties may also be explored. As an example, one could attempt to elucidate the role of these conserved regions in the origin and development of protein-protein interfaces. It has been shown that hydrophobic mutational ratchets allow the accumulation of multimer-neutral, monomer-deleterious substitutions, entrenching those multimers over evolutionary time, even if the interface itself has no functional advantage.[6] In an similar way, if we consider ancestral fragments in a domain as a facultative heterodimer, entrenchment could have been a mechanism involved in the recruitment and consolidation of fragment-fragment interfaces before or after recombination. The fragments found in Fuzzle should contain this kind of information. Additionally, protein-

[6] Hochberg et al., 2020

nucleic acid interactions could be studied as well, since they must have been very relevant in the ancestral protein universe, where protein-nucleic acid complexes are presumed to have been prevalent in the prebiotic world. Whether the nucleotide-binding conserved motifs found in Fuzzle are related to them or arose independently remains to be determined.

Conserved fragments between domains from different folds show the limitations of structure-based fold classification systems. It is clear that evolutionary criteria should be taken into account in domain classification and there are many efforts trying to integrate them. Said efforts involve either the actual domain databases (SCOP2 and ECOD,[7]) or third-party ones derived from them, including Fuzzle.[8] All of them reach the same conclusion: most relevant folds, including Rossmann, are linked by small ancestral βα fragments. These fragments, when represented as a network, form one large, single connected component.

[7] Andreeva et al., 2014; Cheng et al., 2014

[8] Alva et al., 2015; Nepomnyachiy et al., 2017; Ferruz et al., 2020

Additional methods, such as evolutionary coupling analysis, could be used to further assess common ancestry between different folds. In interfold relationships with large structural disagreements, such as the cases involving the PBP-like I fold, evolution-enforced contacts should be lost in the regions without structural similarity. In theory it could be possible to analyze this kind of connections using Fuzzle hits. The main obstacle would be, given the low sequence identity present in interfold hits, the generation of plausible MSAs that preserve the information contained in the PSAs.

Fold fragments as material for protein chimeras

Apart from evolutionary purposes, the fragments in Fuzzle may serve for chimeragenesis as well. The structural similarity they have makes them suitable for recombination. There are, however, some caveats.

The Rossmann fold/Fld-like chimeras showed several issues that hindered their stability in solution. While some mutants could be expressed in soluble form, the aggregation issues could not be completely eliminated. The instability of the chimeras seems to stem from the nonexhaustive sampling used

in the modeling process. One way to check for inaccuracies and topological preferences of the model would be to perform *ab initio* forward folding, i.e., the computational refolding of the chimera sequence. If the sequence can recovered the modeled structure, then the likelihood of a stable *in vitro* chimera should increase. Another possibility is to employ large-scale combinatorial approaches using fragment libraries to find the best-fitting pairs. Additional techniques aimed to increase stability such as directed evolution or circular permutation might be used, but in both cases the chimeric sequence can be altered considerably. Another possible strategy would be to add a linker segment connecting the fragments, to increase the conformational freedom between them and possibly increase the complementarity at the interface.

The PurT/PurN chimeras, on the other hand, owing to their more rigorous computational design pipeline, proved more successful, with two of them being solubly expressed without any aggregation observed. Additionally, the structure of *Ec*PurT/*Bh*PurN could be solved by X-ray crystallography. It resembled greatly the theoretical model and the parental formyltransferase, with a slight rearrangement of the domain fragments relative to each other. This change in orientation, however, was not enough to destabilize the chimera. Despite folding correctly, binding of neither the GAR substrate nor the folate cofactor could not be assessed at this time.

Along the same line, some projects from the group of Stephen Benkovic involved hybrid proteins using enzymes from the formyltransferase fold. In one work, they generated *E. coli* PurN/PurU (PurU: 10-formyltetrahydrofolate hydrolase) hybrids that had the catalytic machinery of PurN, with PurU providing the fTHF binding pocket. The hybrid enzyme expressed insolubly and thus required purification under denaturing conditions followed by refolding. Moreover, they had a formyltransferase activity 100- to 1000-fold lower that wild-type PurN. In a follow-up study,[9] the most active hybrid underwent random mutagenesis assays by DNA shuffling and selection by β-galactosidase complementation. The variant with the highest activity had acquired five mutations, four of which mapped to the vicinity of the fragments' interface, although none of them was directly involved in inter-fragment contacts. Based on these results the authors propose that the

[9] Nixon & Benkovic, 2000

initial hybrid exists in two states: a closed one, where both substrate-binding domain fragments are close enough for the formyltransferase reaction to take place, and an open one, where only the PurU fragment is able to function properly, hydrolyzing the cofactor without any possible transfer.

A related project[10] used incremental truncation libraries (see ITCHY, page 38) of *E. coli* PurN fragments to generate heterodimers formed by N- and C-terminal segments of varying lengths. Several functional variants were found, with the most active ones having between 5 and 10 percent of the monomeric, wild-type PurN activity. Unfortunately, no structural studies were done on said complexes and no comments about conformational similarities or differences between the PurN monomer and the engineered heterodimers were made. Along the same line, it is not known whether the relative orientation of both halves in the functional PurN heterodimers remained similar to the wild-type enzyme. Like the previous PurN/PurU case, this study suggests the presence of a "division of labor" between both halves, with the N-terminal fragment in charge of binding the GAR substrate, and the C-terminal one associating with the formyl-containing cofactor. They expanded on this topic by fusing an N-terminal fragment of *E. coli* PurN with a C-terminal fragment of human glycinamide ribonucleotide formyltransferase (hGART).[11] Both enzymes have 45.8% identity at the DNA level but could nevertheless be recombined into a chimeric formyltransferase. Like all the previously mentioned studies, no structural characterization was reported. PurN/hGART was used as an example of SCHEMA recombination.[12]

[10] Ostermeier, Nixon, et al., 1999

[11] Ostermeier, Shim, et al., 1999

[12] Voigt et al., 2002

In our work, the chimeras that folded solubly had PurN as the enzyme supplying the fTHF binding site, while PurT brought the GAR-interacting fragment. The way these fragments associate with each other resembles what is seen in Benkovic's constructs. However, since substrate and cofactor binding has not been assessed yet, we can only speculate about the effect of the recombination on it. The PurT fragment interacts mostly with the ribose moiety of GAR and as a consequence a decrease or abolishment of its binding can be expected. The fTHF binding pocket is mostly unaffected in the chimera and therefore the levels of cofactor binding should be similar to the parental PurN.

Evolutionary relationship between PurN and PurT

PurN and PurT have long been considered analogous enzymes in the *de novo* purine biosynthesis pathway, based on their low sequence identity and different domain architecture as well as catalytic mechanisms. In this work we present sequence- and structure-based evidence that suggests a common origin based on evolutionarily conserved domain fragments, which in this particular case appears to have been selected for substrate binding. Furthermore, additional connections between both PurN and PurT to other folds were found. The link found between PurN, PurT, and Rossmann fold is of particular interest as PurN and PurT are involved in nucleotide synthesis, while the conserved fragment in Rossmann fold domains interact with the adenosine phosphate segment of nucleotide coenzymes. These properties hint at the essential role during early life conditions that these remotely conserved fragments might have fulfilled and consequently kept, at least partially, in extant proteins. It is worth noticing, though, that despite PurT lying closer to Rossmann fold domains than PurN, the GAR binding mode in PurT does not resemble the one seen in PurN or Rossmann fold proteins. In contrast, PurN binds GAR in a position similar to the nicotinamide mononucleotidyl moiety of NAD in Rossmann fold enzymes (Figure 81 and 88).

Functional differences between PurN and PurT

Despite having a remote evolutionary link, an explanation for the simultaneous presence of PurN and PurT in proteobacteria remains to be found. There have been remarkably few attempts to explain differences between both enzymes, but recent studies in several bacterial species offer some clues.

PurN and PurT seem identical *in vitro* but appear to have different roles under certain *in vivo* circumstances. Both enzymes were the target of one work involving the *Riptortus-Burkholderia* symbiotic system, aimed at unveiling the effects of *Burkholderia purN* and *purT* deletions on its biofilm-forming abilities and symbiotic properties.[13] Δ*purN* and Δ*purT* exhibited normal growth but, unexpectedly, only the latter strain presented impaired biofilm formation levels and defects

[13] Kim et al., 2014

in their symbiotic association, caused by decreased levels of the second messenger cyclic-di-GMP. Additionally, host insects with Δ*purT* symbionts exhibited lower bacterial density, slower growth, and lighter body weight. Another analysis revealed that both enzymes are redundant in pathogenic *Salmonella enterica* grown *in vitro*.[14] However, deletion of either gene affected infection in mice, and propagation of the Δ*purN* mutant but not Δ*purT* was attenuated in cultured macrophages. This outcome suggests that, during invasion, DNA damage can be repaired in the absence of PurT but, if PurN is not present, PurT cannot support bacterial propagation by itself. Andersen-Civil et al. found that single and double knockouts of *purN* and *purT* in an uropathogenic *E. coli* strain (UPEC) affected its invasiveness but had no effect on the survival rate after *in vivo* infection.[15] The explanation for these results was that during invasion, UPEC grows in nutrient-deficient conditions (urine, for example) and thus the need for *de novo* purine synthesis arises. Interestingly, no differential effects between Δ*purN* and Δ*purT* mutants were found, suggesting functional redundancy under the conditions of the study. Summing up, these studies suggest that differences between PurN and PurT are subtle and only noticeable at an *in vivo* metabolic level.

[14] Jelsbak et al., 2016

[15] Andersen-Civil et al., 2018

Characterization of the preATP-grasp fold

To date, little information about the preATP-grasp fold can be found in the literature. Reviews have focused on the contiguous ATP-grasp fold,[16] with only passing mentions of c.30. SCOP describes the preATP-grasp fold as a "possible rudiment form of Rossmann-fold domain", but no citation is brought up and no publication mentioning this hypothesis could be found. Similar to what is seen in the Fld-like fold, most families within the c.30 fold do not seem to be evolutionarily related, as the clustering of c.30 domains produces connections only between some of them (Figure 86). It may be possible that families under this fold arose independently in an analogous manner. Nevertheless, this only depicts a fraction of hits involving preATP-grasp domains and therefore characterization of the fold in the context of the whole Fuzzle database is still needed. Besides GAR, the preATP-grasp fold holds a variety of ligands, including D-alanine, biotin, and glutathione. The binding mode

[16] Galperin & Koonin, 1997; Fawaz et al., 2011

of GAR is different from the other substrates, being oriented towards the N-terminal half of the preATP-grasp domain in PurT. Most c.30 substrates, on the other hand, lie usually closer to the ATP-grasp domain and therefore closer to the ATP co-factor. This could be explained in two ways: convergence or repurposing of the substrate binding site within preATP-grasp domain families.

References

Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev, A., Zwart, P. H., & Adams, P. D. (2012). Towards automated crystallographic structure refinement with *phenix.refine*. *Acta Crystallographica Section D*, *68*(4), 352–367. https://doi.org/10.1107/S0907444912001308

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Alva, V., Söding, J., & Lupas, A. N. (2015). A vocabulary of ancient peptides at the origin of folded proteins. *eLife*, *4*, e09410. https://doi.org/10.7554/eLife.09410

Andersen-Civil, A. I. S., Ahmed, S., Guerra, P. R., Andersen, T. E., Hounmanou, Y. M. G., Olsen, J. E., & Herrero-Fresno, A. (2018). The impact of inactivation of the purine biosynthesis genes, *purN* and *purT*, on growth and virulence in uropathogenic *E. coli*. *Molecular Biology Reports*, *45*(6), 2707–2716. https://doi.org/10.1007/s11033-018-4441-z

Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., & Murzin, A. G. (2014). SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Research*, *42*(D1), D310–D314. https://doi.org/10.1093/nar/gkt1242

Apic, G., Gough, J., & Teichmann, S. A. (2001). Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *Journal of Molecular Biology*, *310*(2), 311–325. https://doi.org/10.1006/jmbi.2001.4776

Baalsrud, H. T., Tørresen, O. K., Solbakken, M. H., Salzburger, W., Hanel, R., Jakobsen, K. S., & Jentoft, S. (2017). *De Novo* Gene Evolution of Antifreeze Glycoproteins in Codfishes Revealed by Whole Genome Sequence Data. *Molecular Biology and Evolution*, *35*(3), 593–606. https://doi.org/10.1093/molbev/msx311

Bharat, T. A. M., Eisenbeis, S., Zeth, K., & Höcker, B. (2008). A βα-barrel built by the combination of fragments from different folds. *Proceedings of the National Academy of Sciences*, *105*(29), 9942–9947. https://doi.org/10.1073/pnas.0802202105

Biegert, A., & Söding, J. (2008). De novo identification of highly diverged protein repeats by probabilistic consistency. *Bioinformatics*, *24*(6), 807–814. https://doi.org/10.1093/bioinformatics/btn039

Bono, S. de, Riechmann, L., Girard, E., Williams, R. L., & Winter, G. (2005). A segment of cold shock protein directs the folding of a combinatorial protein. *Proceedings of the National Academy of Sciences*, *102*(5), 1396–1401. https://doi.org/10.1073/pnas.0407298102

Caetano-Anollés, G., Kim, H. S., & Mittenthal, J. E. (2007). The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. *Proceedings of the National Academy of Sciences*, *104*(22), 9358–9363. https://doi.org/10.1073/pnas.0701214104

Cammer, S., & Carter, J., Charles W. (2010). Six Rossmannoid folds, including the Class I aminoacyl-tRNA synthetases, share a partial core with the anti-codon-binding domain of a Class II aminoacyl-tRNA synthetase. *Bioinformatics*, *26*(6), 709–714. https://doi.org/10.1093/bioinformatics/btq039

Chandonia, J.-M., Hon, G., Walker, N. S., Lo Conte, L., Koehl, P., Levitt, M., & Brenner, S. E. (2004). The ASTRAL Compendium in 2004. *Nucleic Acids Research*, *32*, D189–D192. https://doi.org/10.1093/nar/gkh034

Chen, C. K., Chan, N. L., & Wang, A. H. (2011). The many blades of the β-propeller proteins: conserved but versatile. *Trends Biochem. Sci.*, *36*(10), 553–561.

Cheng, H., Schaeffer, R. D., Liao, Y., Kinch, L. N., Pei, J., Shi, S., Kim, B.-H., & Grishin, N. V. (2014). ECOD: An evolutionary classification of protein domains. *PLOS Computational Biology*, *10*(12), 1–18. https://doi.org/10.1371/journal.pcbi.1003926

Crick, F. H. (1958). On protein synthesis. *Symp. Soc. Exp. Biol.*, *12*, 138–163.

Dahiyat, B. I., & Mayo, S. L. (1997). De novo protein design: Fully automated sequence selection. *Science*, *278*(5335), 82–87. https://doi.org/10.1126/science.278.5335.82

Dawson, N., Sillitoe, I., Marsden, R. L., & Orengo, C. A. (2017). The classification of protein domains. In *Methods in Molecular Biology* (pp. 137–164). Springer New York. https://doi.org/10.1007/978-1-4939-6622-6_7

Dayhoff, M. O. (Ed.). (1965). *Atlas of protein sequence and structure*. National Biomedical Research Foundation.

Dayhoff, M. O. (1974). Computer analysis of protein sequences. In *Computers in Life Science Research* (pp. 9–14). Springer US. https://doi.org/10.1007/978-1-4757-0546-1_3

Edelman, G. M. (1973). Antibody structure and molecular immunology. *Science*, *180*(4088), 830–840. https://doi.org/10.1126/science.180.4088.830

Edelman, G. M., Cunningham, B. A., Gall, W. E., Gottlieb, P. D., Rutishauser, U., & Waxdal, M. J. (1969). The covalent structure of an entire γG immunoglobulin molecule. *Proceedings of the National Academy of Sciences*, *63*(1), 78–85. https://doi.org/10.1073/pnas.63.1.78

Eisenbeis, S., Proffitt, W., Coles, M., Truffault, V., Shanmugaratnam, S., Meiler, J., & Höcker, B. (2012). Potential of fragment recombination for rational design of proteins. *Journal of the American Chemical Society*, *134*(9), 4019–4022. https://doi.org/10.1021/ja211657k

Emsley, P., Lohkamp, B., Scott, W. G., & Cowtan, K. (2010). Features and development of *Coot*. *Acta Crystallographica Section D*, *66*(4), 486–501. https://doi.org/10.1107/S0907444910007493

Espinosa-Cantú, A., Ascencio, D., Barona-Gómez, F., & DeLuna, A. (2015). Gene duplication and the evolution of moonlighting proteins. *Frontiers in Genetics*, *6*, 227. https://doi.org/10.3389/fgene.2015.00227

Farías Rico, J. A. (2014). *Evolutionary relationships beyond fold boundaries* [PhD thesis, Universität Tübingen]. https://publikationen.uni-tuebingen.de/xmlui/handle/10900/55864

Farías-Rico, J. A., Schmidt, S., & Höcker, B. (2014). Evolutionary relationship of two ancient protein superfolds. *Nature Chemical Biology*, *10*(9), 710–715. https://doi.org/10.1038/nchembio.1579

Fawaz, M. V., Topper, M. E., & Firestine, S. M. (2011). The ATP-grasp enzymes. *Bioorganic Chemistry*, *39*(5), 185–191. https://doi.org/10.1016/j.bioorg.2011.08.004

Ferruz, N., Lobos, F., Lemm, D., Toledo-Patino, S., Farías-Rico, J. A., Schmidt, S., & Höcker, B. (2020). Identification and analysis of natural building blocks for evolution-guided fragment-based protein design. *Journal of Molecular Biology*. https://doi.org/10.1016/j.jmb.2020.04.013

Figueroa, M., Sleutel, M., Vandevenne, M., Parvizi, G., Attout, S., Jacquin, O., Vandenameele, J., Fischer, A. W., Damblon, C., Goormaghtigh, E., Valerio-Lepiniec, M., Urvoas, A., Durand, D., Pardon, E., Steyaert, J., Minard, P., Maes, D., Meiler, J., Matagne, A., … Van de Weerdt, C. (2016). The unexpected structure of the designed protein Octarellin V.1 forms a challenge for protein structure prediction tools. *Journal of Structural Biology*, *195*(1), 19–30. https://doi.org/10.1016/j.jsb.2016.05.004

Fischer, N., Riechmann, L., & Winter, G. (2004). A native-like artificial protein from antisense DNA. *Protein Engineering, Design and Selection*, *17*(1), 13–20. https://doi.org/10.1093/protein/gzh002

Fleishman, S. J., Leaver-Fay, A., Corn, J. E., Strauch, E.-M., Khare, S. D., Koga, N., Ashworth, J., Murphy, P., Richter, F., Lemmon, G., Meiler, J., & Baker, D. (2011). RosettaScripts: A scripting language interface to the rosetta macromolecular modeling suite. *PLOS ONE*, *6*(6), 1–10. https://doi.org/10.1371/journal.pone.0020161

Fox, N. K., Brenner, S. E., & Chandonia, J.-M. (2013). SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research*, *42*(D1), D304–D309. https://doi.org/10.1093/nar/gkt1240

Franklin, M. W., Nepomnyachyi, S., Feehan, R., Ben-Tal, N., Kolodny, R., & Slusky, J. S. (2018). Evolutionary pathways of repeat protein topology in bacterial outer membrane proteins. *eLife*, *7*, e40308. https://doi.org/10.7554/eLife.40308

Frickey, T., & Lupas, A. (2004). CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics*, *20*(18), 3702–3704. https://doi.org/10.1093/bioinformatics/bth444

Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, *21*(11), 1129–1164. https://doi.org/https://doi.org/10.1002/spe.4380211102

Fukami-Kobayashi, K., Tateno, Y., & Nishikawa, K. (1999). Domain dislocation: A change of core structure in periplasmic binding proteins in their evolutionary history. *Journal of Molecular Biology*, *286*(1), 279–290. https://doi.org/10.1006/jmbi.1998.2454

Futuyma, D. (2018). *Evolution*. Sinauer Associates.

Galperin, M. Y., & Koonin, E. V. (1997). A diverse superfamily of enzymes with ATP-dependent carboxylate—amine/thiol ligase activity. *Protein Science*, *6*(12), 2639–2643. https://doi.org/10.1002/pro.5560061218

Goodsell, D. (2021). Escherichia coli *bacterium*. RCSB Protein Data Bank. https://doi.org/10.2210/rcsb_pdb/goodsell-gallery-028

Goraj, K., Renard, A., & Martial, J. A. (1990). Synthesis, purification and initial structural characterization of octarellin, a *de novo* polypeptide modelled on the α/β-barrel Proteins. *Protein Engineering, Design and Selection*, *3*(4), 259–266. https://doi.org/10.1093/protein/3.4.259

Gutte, B., Däumigen, M., & Wittschieber, E. (1979). Design, synthesis and characterisation of a 34-residue polypeptide that interacts with nucleic acids. *Nature*, *281*(5733), 650–655. https://doi.org/10.1038/281650a0

Hamelryck, T., & Manderick, B. (2003). PDB file parser and structure class implemented in Python. *Bioinformatics*, *19*(17), 2308–2310. https://doi.org/10.1093/bioinformatics/btg299

Han, J.-H., Batey, S., Nickson, A. A., Teichmann, S. A., & Clarke, J. (2007). The folding and evolution of multidomain proteins. *Nature Reviews Molecular Cell Biology*, *8*(4), 319–330. https://doi.org/10.1038/nrm2144

Harbury, P. B., Plecs, J. J., Tidor, B., Alber, T., & Kim, P. S. (1998). High-resolution protein design with backbone freedom. *Science*, *282*(5393), 1462–1467. https://doi.org/10.1126/science.282.5393.1462

Hecht, M., Richardson, J., Richardson, D., & Ogden, R. (1990). De novo design, expression, and characterization of Felix: A four-helix bundle protein of native-like sequence. *Science*, *249*(4971), 884–891. https://doi.org/10.1126/science.2392678

Ho, S. P., & DeGrado, W. F. (1987). Design of a 4-helix bundle protein: Synthesis of peptides which self-associate into a helical protein. *Journal of the American Chemical Society*, *109*(22), 6751–6758. https://doi.org/10.1021/ja00256a032

Hochberg, G. K. A., Liu, Y., Marklund, E. G., Metzger, B. P. H., Laganowsky, A., & Thornton, J. W. (2020). A hydrophobic ratchet entrenches molecular complexes. *Nature*, *588*(7838), 503–508. https://doi.org/10.1038/s41586-020-3021-2

Höcker, B., Beismann-Driemeyer, S., Hettwer, S., Lustig, A., & Sterner, R. (2001). Dissection of a $(\beta\alpha)_8$-barrel enzyme into two folded halves. *Nature Structural Biology*, *8*(1), 32–36.

https://doi.org/10.1038/83021

Holm, L. (2020). DALI and the persistence of protein shape. *Protein Science*, *29*(1), 128–140. https://doi.org/https://doi.org/10.1002/pro.3749

Huang, P.-S., Feldmeier, K., Parmeggiani, F., Velasco, D. A. F., Höcker, B., & Baker, D. (2015). *De novo* design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nature Chemical Biology*, *12*(1), 29–34. https://doi.org/10.1038/nchembio.1966

Innan, H., & Kondrashov, F. (2010). The evolution of gene duplications: Classifying and distinguishing between models. *Nature Reviews Genetics*, *11*(2), 97–108. https://doi.org/10.1038/nrg2689

Jacob, F. (1977). Evolution and tinkering. *Science*, *196*(4295), 1161–1166.

Jacobs, T. M., Williams, B., Williams, T., Xu, X., Eletsky, A., Federizon, J. F., Szyperski, T., & Kuhlman, B. (2016). Design of structurally distinct proteins using strategies inspired by evolution. *Science*, *352*(6286), 687–690. https://doi.org/10.1126/science.aad8036

Janin, J., & Wodak, S. J. (1983). Structural domains in proteins and their role in the dynamics of protein function. *Prog. Biophys. Mol. Biol.*, *42*(1), 21–78.

Jelsbak, L., Mortensen, M. I. B., Kilstrup, M., & Olsen, J. E. (2016). The *in vitro* redundant enzymes PurN and PurT are both essential for systemic infection of mice in *Salmonella enterica* serovar Typhimurium. *Infection and Immunity*, *84*(7), 2076–2085. https://doi.org/10.1128/IAI.00182-16

Jiang, L., Althoff, E. A., Clemente, F. R., Doyle, L., Röthlisberger, D., Zanghellini, A., Gallaher, J. L., Betker, J. L., Tanaka, F., Barbas, C. F., Hilvert, D., Houk, K. N., Stoddard, B. L., & Baker, D. (2008). De novo computational design of retro-aldol enzymes. *Science*, *319*(5868), 1387–1391. https://doi.org/10.1126/science.1152692

Kalev, I., Mechelke, M., Kopec, K. O., Holder, T., Carstens, S., & Habeck, M. (2012). CSB: a Python framework for structural bioinformatics. *Bioinformatics*, *28*(22), 2996–2997. https://doi.org/10.1093/bioinformatics/bts538

Kathuria, S. V., Chan, Y. H., Nobrega, R. P., Özen, A., & Matthews, C. R. (2016). Clusters of isoleucine, leucine, and valine side chains define cores of stability in high-energy states of globular proteins: Sequence determinants of structure and stability. *Protein Science*, *25*(3), 662–675. https://doi.org/10.1002/pro.2860

Keller, E. F., & Lloyd, E. A. (Eds.). (1992). *Keywords in evolutionary biology*. Harvard University Press.

Khersonsky, O., & Fleishman, S. J. (2016). Why reinvent the wheel? Building new proteins based on ready-made parts. *Protein Science*, *25*(7), 1179–1187. https://doi.org/https://doi.org/10.1002/pro.2892

Kim, J. K., Jang, H. A., Won, Y. J., Kikuchi, Y., Heum Han, S., Kim, C.-H., Nikoh, N., Fukatsu, T., & Lee, B. L. (2014). Purine biosynthesis-deficient *burkholderia* mutants are incapable of symbiotic accommodation in the stinkbug. *The ISME Journal*, *8*(3), 552–563.

https://doi.org/10.1038/ismej.2013.168

Kohl, A., Binz, H. K., Forrer, P., Stumpp, M. T., Plückthun, A., & Grütter, M. G. (2003). Designed to be stable: Crystal structure of a consensus ankyrin repeat protein. *Proceedings of the National Academy of Sciences*, *100*(4), 1700–1705. https://doi.org/10.1073/pnas.0337680100

Kozakov, D., Hall, D. R., Xia, B., Porter, K. A., Padhorny, D., Yueh, C., Beglov, D., & Vajda, S. (2017). The ClusPro web server for protein–protein docking. *Nature Protocols*, *12*(2), 255–278. https://doi.org/10.1038/nprot.2016.169

Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., & Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science*, *302*(5649), 1364–1368. https://doi.org/10.1126/science.1089427

Laurino, P., Tóth-Petróczy, Á., Meana-Pañeda, R., Lin, W., Truhlar, D. G., & Tawfik, D. S. (2016). An ancient fingerprint indicates the common ancestry of Rossmann-fold enzymes utilizing different ribose-based cofactors. *PLOS Biology*, *14*(3), 1–23. https://doi.org/10.1371/journal.pbio.1002396

Liu, J., & Rost, B. (2004). CHOP: parsing proteins into structural domains. *Nucleic Acids Research*, *32*, W569–W571. https://doi.org/10.1093/nar/gkh481

Longo, L. M., Jabłońska, J., Vyas, P., Kanade, M., Kolodny, R., Ben-Tal, N., & Tawfik, D. S. (2020). On the emergence of P-Loop NTPase and Rossmann enzymes from a Beta-Alpha-Beta ancestral fragment. *eLife*, *9*, e64415. https://doi.org/10.7554/eLife.64415

Main, E. R. G., Xiong, Y., Cocco, M. J., D'Andrea, L., & Regan, L. (2003). Design of stable α-helical arrays from an idealized TPR motif. *Structure*, *11*(5), 497–508. https://doi.org/10.1016/S0969-2126(03)00076-5

Marcos, E., Chidyausiku, T. M., McShan, A. C., Evangelidis, T., Nerli, S., Carter, L., Nivón, L. G., Davis, A., Oberdorfer, G., Tripsianes, K., Sgourakis, N. G., & Baker, D. (2018). De novo design of a non-local β-sheet protein with high stability and accuracy. *Nat Struct Mol Biol*, *25*(11), 1028–1034. https://doi.org/10.1038/s41594-018-0141-6

Marolewski, A. E., Mattia, K. M., Warren, M. S., & Benkovic, S. J. (1997). Formyl phosphate: a proposed intermediate in the reaction catalyzed by *Escherichia coli* PurT GAR transformylase. *Biochemistry*, *36*(22), 6709–6716. https://doi.org/10.1021/bi962961p

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., & Read, R. J. (2007). *Phaser* crystallographic software. *Journal of Applied Crystallography*, *40*(4), 658–674. https://doi.org/10.1107/S0021889807021206

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2020). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1), D412–D419. https://doi.org/10.1093/nar/gkaa913

Moser, R., Thomas, R. M., & Gutte, B. (1983). An artificial crystalline DDT-binding polypeptide. *FEBS Letters*, *157*(2), 247–251. https://doi.org/10.1016/0014-5793(83)80555-9

Nagano, N., Gail Hutchinson, E., & Thornton, J. M. (1999). Barrel structures in proteins: Automatic identification and classification including a sequence analysis of TIM barrels. *Protein Science*, *8*(10), 2072–2084. https://doi.org/10.1110/ps.8.10.2072

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453. https://doi.org/10.1016/0022-2836(70)90057-4

Nepomnyachiy, S., Ben-Tal, N., & Kolodny, R. (2017). Complex evolutionary footprints revealed in an analysis of reused protein segments of diverse lengths. *Proceedings of the National Academy of Sciences*, *114*(44), 11703–11708. https://doi.org/10.1073/pnas.1707642114

Nixon, A. E., & Benkovic, S. J. (2000). Improvement in the efficiency of formyl transfer of a GAR transformylase hybrid enzyme. *Protein Engineering, Design and Selection*, *13*(5), 323–327. https://doi.org/10.1093/protein/13.5.323

Ostermeier, M., Nixon, A. E., Shim, J. H., & Benkovic, S. J. (1999). Combinatorial protein engineering by incremental truncation. *Proceedings of the National Academy of Sciences*, *96*(7), 3562–3567. https://doi.org/10.1073/pnas.96.7.3562

Ostermeier, M., Shim, J. H., & Benkovic, S. J. (1999). A combinatorial approach to hybrid enzymes independent of DNA homology. *Nature Biotechnology*, *17*(12), 1205–1209. https://doi.org/10.1038/70754

Pearson, W. R., & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, *85*(8), 2444–2448. https://doi.org/10.1073/pnas.85.8.2444

Pierce, N. A., & Winfree, E. (2002). Protein Design is NP-hard. *Protein Engineering, Design and Selection*, *15*(10), 779–782. https://doi.org/10.1093/protein/15.10.779

Rawcliffe, G. (2019). *Exploring the protein universe: A study of subdomain driven evolution* [PhD thesis, University of Otago]. http://hdl.handle.net/10523/9772

Reichen, C., Hansen, S., & Plückthun, A. (2014). Modular peptide binding: From a comparison of natural binders to designed armadillo repeat proteins. *Journal of Structural Biology*, *185*(2), 147–162. https://doi.org/10.1016/j.jsb.2013.07.012

Remmert, M., Biegert, A., Linke, D., Lupas, A. N., & Söding, J. (2010). Evolution of Outer Membrane β-Barrels from an Ancestral ββ Hairpin. *Molecular Biology and Evolution*, *27*(6), 1348–1358. https://doi.org/10.1093/molbev/msq017

Riechmann, L., & Winter, G. (2000). Novel folded protein domains generated by combinatorial shuffling of polypeptide segments. *Proceedings of the National Academy of Sciences*, *97*(18), 10068—10073. https://doi.org/10.1073/pnas.170145497

Salentin, S., Schreiber, S., Haupt, V. J., Adasme, M. F., & Schroeder, M. (2015). PLIP: fully automated protein–ligand interaction profiler. *Nucleic Acids Research*, *43*(W1), W443–W447. https://doi.org/10.1093/nar/gkv315

Setiyaputra, S., Mackay, J. P., & Patrick, W. M. (2011). The structure of a truncated phosphoribosylanthranilate isomerase suggests a unified model for evolution of the $(\beta\alpha)_8$ barrel fold. *Journal of Molecular Biology*, *408*(2), 291–303. https://doi.org/10.1016/j.jmb.2011.02.048

Shanmugaratnam, S., Eisenbeis, S., & Höcker, B. (2012). A highly stable protein chimera built from fragments of different folds. *Protein Engineering, Design and Selection*, *25*(11), 699–703. https://doi.org/10.1093/protein/gzs074

Sigrist, C. J. A., Castro, E. de, Cerutti, L., Cuche, B. A., Hulo, N., Bridge, A., Bougueleret, L., & Xenarios, I. (2012). New and continuing developments at PROSITE. *Nucleic Acids Research*, *41*(D1), D344–D347. https://doi.org/10.1093/nar/gks1067

Sillitoe, I., Bordin, N., Dawson, N., Waman, V. P., Ashford, P., Scholes, H. M., Pang, C. S. M., Woodridge, L., Rauer, C., Sen, N., Abbasian, M., Le Cornu, S., Lam, S. D., Berka, K., Varekova, I. H., Svobodova, R., Lees, J., & Orengo, C. A. (2020). CATH: increased structural coverage of functional space. *Nucleic Acids Research*, *49*(D1), D266–D273. https://doi.org/10.1093/nar/gkaa1079

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, *147*(1), 195–197. https://doi.org/10.1016/0022-2836(81)90087-5

Smock, R. G., Yadid, I., Dym, O., Clarke, J., & Tawfik, D. S. (2016). De novo evolutionary emergence of a symmetrical protein is shaped by folding constraints. *Cell*, *164*(3), 476–486. https://doi.org/10.1016/j.cell.2015.12.024

Söding, J., Remmert, M., & Biegert, A. (2006). HHrep: de novo protein repeat detection and the origin of TIM barrels. *Nucleic Acids Research*, *34*, W137–W142. https://doi.org/10.1093/nar/gkl130

Sparta, K. M., Krug, M., Heinemann, U., Mueller, U., & Weiss, M. S. (2016). *XDSAPP2.0*. *Journal of Applied Crystallography*, *49*(3), 1085–1092. https://doi.org/10.1107/S1600576716004416

Strasser, B. J. (2010). Collecting, Comparing, and Computing Sequences: The Making of Margaret O. Dayhoff's Atlas of Protein Sequence and Structure, 19541965. *Journal of the History of Biology*, *43*(4), 623–660. https://doi.org/10.1007/s10739-009-9221-0

Stumpp, M. T., Forrer, P., Binz, H., & Plückthun, A. (2003). Designing repeat proteins: Modular leucine-rich repeat protein libraries based on the mammalian ribonuclease inhibitor family. *Journal of Molecular Biology*, *332*(2), 471–487. https://doi.org/10.1016/S0022-2836(03)00897-0

Terada, D., Voet, A. R. D., Noguchi, H., Kamata, K., Ohki, M., Addy, C., Fujii, Y., Yamamoto, D., Ozeki, Y., Tame, J. R. H., & Zhang, K. Y. J. (2017). Computational design of a symmetrical β-trefoil lectin with cancer cell binding activity. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-06332-7

Toledo Patiño, S. (2019). *On the emergence of the hemD-like fold and its use for fold-chimeragenesis* [PhD thesis, Universität Tübingen]. https://doi.org/10.15496/PUBLIKATION-34903

Toledo-Patiño, S., Chaubey, M., Coles, M., & Höcker, B. (2019). Reconstructing the remote origins of a fold singleton from a flavodoxin-like ancestor. *Biochemistry*, *0*(0), null. https://doi.org/10.1021/acs.biochem.9b00900

Tyka, M. D., Keedy, D. A., André, I., DiMaio, F., Song, Y., Richardson, D. C., Richardson, J. S., & Baker, D. (2011). Alternate states of proteins revealed by detailed energy landscape mapping. *Journal of Molecular Biology*, *405*(2), 607–618. https://doi.org/10.1016/j.jmb.2010.11.008

Urvoas, A., Guellouz, A., Valerio-Lepiniec, M., Graille, M., Durand, D., Desravines, D. C., Tilbeurgh, H. van, Desmadril, M., & Minard, P. (2010). Design, production and molecular structure of a new family of artificial alpha-helicoidal repeat proteins (αRep) based on thermostable HEAT-like repeats. *Journal of Molecular Biology*, *404*(2), 307–327. https://doi.org/10.1016/j.jmb.2010.09.048

Vaissier Welborn, V., & Head-Gordon, T. (2019). Computational design of synthetic enzymes. *Chemical Reviews*, *119*(11), 6613–6630. https://doi.org/10.1021/acs.chemrev.8b00399

Voet, A. R. D., Noguchi, H., Addy, C., Simoncini, D., Terada, D., Unzai, S., Park, S.-Y., Zhang, K. Y. J., & Tame, J. R. H. (2014). Computational design of a self-assembling symmetrical β-propeller protein. *Proceedings of the National Academy of Sciences*, *111*(42), 15102–15107. https://doi.org/10.1073/pnas.1412768111

Vogel, C., Bashton, M., Kerrison, N. D., Chothia, C., & Teichmann, S. A. (2004). Structure, function and evolution of multidomain proteins. *Current Opinion in Structural Biology*, *14*(2), 208–216. https://doi.org/10.1016/j.sbi.2004.03.011

Voigt, C. A., Martinez, C., Wang, Z.-G., Mayo, S. L., & Arnold, F. H. (2002). Protein building blocks preserved by recombination. *Nature Structural Biology*. https://doi.org/10.1038/nsb805

Warren, M. S., Mattia, K. M., Marolewski, A. E., & Benkovic, S. J. (1996). The transformylase enzymes of *de novo* purine biosynthesis. *Pure and Applied Chemistry*, *68*(11), 2029–2036. https://doi.org/10.1351/pac199668112029

Wetlaufer, D. B. (1973). Nucleation, rapid folding, and globular intrachain regions in proteins. *Proceedings of the National Academy of Sciences*, *70*(3), 697–701. https://doi.org/10.1073/pnas.70.3.697

Zhang, Y., Morar, M., & Ealick, S. E. (2008). Structural biology of the purine biosynthetic pathway. *Cellular and Molecular Life Sciences*, *65*(23), 3699–3724. https://doi.org/10.1007/s00018-008-8295-8

Zhang, Y., & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, *33*(7), 2302–2309. https://doi.org/10.1093/nar/gki524

Zhao, L., Saelao, P., Jones, C. D., & Begun, D. J. (2014). Origin and spread of de novo genes in *Drosophila melanogaster* populations. *Science*, *343*(6172), 769–772. https://doi.org/10.1126/science.1248286

Zhu, H., Sepulveda, E., Hartmann, M. D., Kogenaru, M., Ursinus, A., Sulz, E., Albrecht, R., Coles, M., Martin, J., & Lupas, A. N. (2016). Origin of a folded repeat protein from an intrinsically disordered ancestor. *eLife*, *5*, e16761. https://doi.org/10.7554/eLife.16761

Contributions

Steffen Schmidt and Dominik Lemm wrote and executed the Fuzzle generation pipeline. Noelia Ferruz and Saacnicteh Toledo helped with the analysis of the Fuzzle database. Lucas Krauss and Eduardo Merino assisted in the expression, purification, and characterization of the Rossmann/Fld-like chimeras. Sooruban Shanmugaratnam and Katja Mindler took part in the expression, purification, characterization and crystallization of the PurT ↔ PurN chimeras.

Acknowledgements

First, I would like to thank Prof. Dr. Birte Höcker for the opportunity to work at her research group. Likewise, I want to extend my gratitude to all current and former members of the AG Höcker for the nice work environment, the fruitful conversations and fun activities. I would also like to acknowledge all the people I met at the Max Planck Institute for Developmental Biology in Tübingen and the Department of Biochemistry at the University of [redacted] and thank them for the good time I had at those locations.

On a personal level, I'm grateful to all the friends I left in Chile and who tried their best (and succeeded!) to stay in touch with me. I hope to have brought you enough Bavarian beer when I went there on vacation. And, to the new friends I made in Germany, I hope to have brought enough Chilean wine and pisco, customs fines notwithstanding. I would also like to give a big shout-out to the Göttingen and München crews, in particular to Verena, Facundo, Susanne, Jürgen, Laura, Fabián, Ksenia, and Drago. Thank you all for your endless hospitality and for making me feel at home, especially during those times when [redacted] felt distant and alienating, if not downright hostile. I owe you a substantial part of my sanity during my stay there.

(Which happened more often than I'm willing to admit.)

Another considerable amount thereof was contributed by my family. I will be forever indebted to my parents, Patricia and Alfredo, as well as my siblings—Benjamín, Valeria, Loreto, and Vanessa—for their unrelenting, unwavering, and unconditional support when it mattered the most, those moments where things seemed too difficult or didn't turn out as expected. This effort wouldn't have been possible without you.

The work shown here was funded by a Consolidator Grant from the European Research Commission.

Concepción, November 2021

Colophon

This thesis was typeset in `pandoc` with a mishmash of Markdown, XeLaTeX, and the `tufte-latex` template. Robert Slimbach's Minion Pro was chosen as text typeface, while Slimbach and Carol Twombly's Myriad Pro is used as display and sans-serif typeface. Monospaced text is typeset in Kris Holmes and Charles Bigelow's Go Mono. The bibliography was processed with BibTeX and `pandoc-citeproc`.