frontiers in Bioinformatics

METHODS
published: 13 April 2021
doi: 10.3389/fbinf.2021.657529

# Deriving and Using Descriptors of Elementary Functions in Rational Protein Design

*Melvin Yin*[1†], *Alexander Goncearenco*[2†] *and Igor N. Berezovsky*[1,3*]

[1] Bioinformatics Institute, Agency for Science, Technology, and Research (A*STAR), Singapore, Singapore, [2] National Center for Biotechnology Information, National Institute of Health (NIH), Bethesda, MD, United States, [3] Department of Biological Sciences (DBS), National University of Singapore (NUS), Singapore, Singapore

The rational design of proteins with desired functions requires a comprehensive description of the functional building blocks. The evolutionary conserved functional units constitute nature's toolbox; however, they are not readily available to protein designers. This study focuses on protein units of subdomain size that possess structural properties and amino acid residues sufficient to carry out elementary reactions in the catalytic mechanisms. The interactions within such elementary functional loops (ELFs) and the interactions with the surrounding protein scaffolds constitute the descriptor of elementary function. The computational approach to deriving descriptors directly from protein sequences and structures and applying them in rational design was implemented in a proof-of-concept DEFINED-PROTEINS software package. Once the descriptor is obtained, the ELF can be fitted into existing or novel scaffolds to obtain the desired function. For instance, the descriptor may be used to determine the necessary spatial restraints in a fragment-based grafting protocol. We illustrated the approach by applying it to well-known cases of ELFs, including phosphate-binding P-loop, diphosphate-binding glycine-rich motif, and calcium-binding EF-hand motif, which could be used to jumpstart templates for user applications. The DEFINED-PROTEINS package is available for free at https://github.com/MelvinYin/Defined_Proteins.

**Keywords: protein function, protein design, elementary functional loops, elementary function, descriptor of the elementary function, DEFINED-PROTEINS software package**

OPEN ACCESS

***Edited by:***
*Michael Gromiha, Indian Institute of Technology Madras, India*

***Reviewed by:***
*Selvaraj Samuel, Bharathidasan University, India*
*Kumar Yugandhar, Cornell University, United States*

***Correspondence:**
*Igor N. Berezovsky*
*igorb@bii.a-star.edu.sg*

[†] *These authors have contributed equally to this work*

***Specialty section:***
*This article was submitted to Protein Bioinformatics, a section of the journal Frontiers in Bioinformatics*

***Received:*** *23 January 2021*
***Accepted:*** *15 March 2021*
***Published:*** *13 April 2021*

***Citation:***
*Yin M, Goncearenco A and Berezovsky IN (2021) Deriving and Using Descriptors of Elementary Functions in Rational Protein Design. Front. Bioinform. 1:657529. doi: 10.3389/fbinf.2021.657529*

## INTRODUCTION

Contemporary views of the enzymatic functions are dominated by either consideration of functional domains or the catalytic active sites as their minimal structural/functional units (Marchler-Bauer et al., 2015; Finn et al., 2016; Trudeau and Tawfik, 2019). The relationships in protein function evolution, however, are far more complex than the sequence-based models can describe (Nath et al., 2014; Goncearenco and Berezovsky, 2015; Aziz et al., 2016; Romero Romero et al., 2016; Berezovsky et al., 2017a). In many cases, the closely related structurally similar folds can carry completely different biochemical functions, while, on the other hand, the same function can be performed by many different protein folds (Goncearenco and Berezovsky, 2015; Berezovsky, 2019). According to the Enzyme Commission number (EC) nomenclature (Bairoch, 2000), the current number of enzymatic functions reaches up to 5,000. The number of underlying biochemical mechanisms (Holliday et al., 2012), however, does not exceed 500; moreover, the

number of elementary chemical reactions (Holliday et al., 2005) is less than 50. The differences in the order of magnitude between the numbers of enzymatic functions, biochemical mechanisms, and elementary chemical reactions prompt one to consider the biochemical function as a combination of elementary ones. Also, the domains themselves had to evolve from some primitive forms (Berezovsky, 2003, 2019; Nath et al., 2014; Goncearenco and Berezovsky, 2015; Aziz et al., 2016; Romero Romero et al., 2016, 2018; Berezovsky et al., 2017a,b). It has been hypothesized that the first group of enzymatic domains emerged as combinations of prebiotic (Romero Romero et al., 2016) ring-like peptides with simple chemical transformation (Trifonov et al., 2001; Goncearenco and Berezovsky, 2015; Berezovsky et al., 2017a; Berezovsky, 2019). Closed loops of preferential 25 to 35-amino acid residue size, a universal basic element of soluble proteins, are the descendants of the prebiotic peptides (Berezovsky et al., 2000, 2017a,b; Berezovsky, 2003) determined by the polymer nature of polypeptide chains (Yamakawa and Stockmayer, 1972; Shimada and Yamakawa, 1984; Berezovsky et al., 2000, 2017a; Orevi et al., 2013; Jacob et al., 2018).

Previous studies showed that biochemical functions can be represented as a combination of the elementary ones, provided by the elementary functional loops (EFLs), which are closed loops with specific signatures that perform elementary steps of biochemical transformations (Goncearenco and Berezovsky, 2010, 2011, 2012, 2015). We suggest that EFLs can be considered as potential elementary units in the design of biochemical functions, which requires an exhaustive description of their characteristics that are important for building the required catalytic site in the environment of the particular protein fold. Even though EFLs perform only elementary steps of the biochemical reactions, the relationship between their sequences, structures, and functions is complex. For example, CxxC motifs (Goncearenco and Berezovsky, 2011) are known to be involved in a variety of functions, such as metal or metal-containing cofactor binding or redox reactions. Consequently, structures of the EFLs containing this signature in many folds differ significantly (Goncearenco and Berezovsky, 2012; Zheng et al., 2016), depending on both the structural environment in the protein and its overall biochemical function. The interactions between the EFL and the substrate and between the EFL and the rest of the structure will also depend on the fold and its function (Berezovsky, 2019). Therefore, the descriptor of EFL has to capture structure- and function-dependent interaction propensities.

The protein design adventure (Das and Baker, 2008) started more than 50 years ago from the general protein folding problem (Dill and MacCallum, 2012) formulated in terms of polymer and statistical physics (Shakhnovich and Gutin, 1993) of biomolecules (Sali et al., 1994; Shakhnovich, 2006), in terms of statistical predictions of structures from the sequence (Sippl, 1990; Crippen, 1996), and as an inverse protein-folding problem (Rooman et al., 1990; Bowie et al., 1991; Rooman and Wodak, 1995) of finding the sequence that can be threaded into the certain fold structure. Current progress in the evolutionary-inspired fragment-based (Hocker, 2014) and *de novo* (Huang et al., 2016a; Silva et al., 2019) design is described in several original works (Brunette et al., 2015) and reviews (Lechner et al., 2018; Baker, 2019; Berezovsky, 2019). It is further facilitated by advances in machine learning and artificial intelligence, as well as by the quality and quantity of high-throughput sequence and structural data available, leading to significant improvements in the performance of computational approaches (Senior et al., 2020). Despite significant progress in the computational design of protein structures, the journey toward solving the great challenge of the *de novo* design of protein functions is, as of yet, at its very beginning (Huang et al., 2016a; Lechner et al., 2018; Baker, 2019; Berezovsky, 2019). Although the repertoire of conserved continuous functional units is available on the sequence level, a more comprehensive characterization is required to define spatial and interaction restraints (Berezovsky, 2019). The computational framework presented in this study facilitates the derivation of the descriptor of elementary function and conceptualizes the objective function for protein engineering and design applications using the descriptor. It merges structure, sequence, and interaction features important for defining elementary functions on a residue level. In protein design, descriptors of elementary functional units serve as off-the-shelf building blocks, for instance, in protein grafting, while the objective function optimizes the choice of such building blocks from a library, considering their geometry and interactions with the protein scaffold, particularly in the key catalytic or binding residues.

We illustrate this approach by calculating the descriptors for three ubiquitous ELFs: the calcium-binding EF-hand, the phosphate-binding in mononucleotide-containing ligands, and the phosphate-binding in dinucleotide-containing ligands, such as ATP, nicotinamide–adenine–dinucleotide (NAD), and NAD phosphate (NADP), in a variety of structural scaffolds. We also model a hypothetical grafting experiment by swapping the EFLs and EFL-derived chimeras among the scaffolds.

## MATERIALS AND METHODS

The proof-of-concept computational framework is aimed at, first, derivation of the descriptor of elementary function and, second, application of descriptors to design proteins with desired structures and functions. A descriptor represents a set of characteristics of the EFL (Goncearenco and Berezovsky, 2010, 2011, 2012; Berezovsky et al., 2017a), including the position-specific information on the sequence, and several structural features encoded as probabilistic distributions (Berezovsky, 2019). An elementary function is defined as the smallest structural unit sufficient to carry out an elementary reaction in a biochemical transformation. Depending on the protein engineering task, it may be needed to introduce or replace an elementary function in an existing protein of interest or build and design a protein with the required structure and function *de novo* (Berezovsky, 2019). The flowchart in **Supplementary Figure 1** illustrates the sequence of steps described below.

### Deriving the Descriptor

Motivated by the biophysical constraints of the polypeptide chain (Goncearenco and Berezovsky, 2010, 2011; Berezovsky et al.,

2017a; Berezovsky, 2019), the procedure starts from 30-residue long seed sequence fragments of the functional loops represented by a gapless multiple sequence alignment and a position-specific scoring matrix (PSSM) profile. The sequence profile is then iteratively scanned against the non-redundant UniRef database (Hunter et al., 2009) with an expectation–maximization (EM)-like algorithm (Goncearenco and Berezovsky, 2010, 2011) converging to an expanded sequence profile of the descriptor with a functional signature (Berezovsky et al., 2003a,b). The corresponding structures characterizing the functional loop are then obtained by looking for profile matches against the sequences in the Protein Data Bank (Berman et al., 2000; wwPDB consortium, 2019), followed by extraction and encoding of the structural feature in the form of parametrized probability distribution functions. Structural and functional annotations are extracted from the corresponding databases, such as Uniprot, MaCiE, and Conserved Domains Database (CDD) (Bairoch, 2000; Andreini et al., 2009; Fischer et al., 2010; Holliday et al., 2012; Marchler-Bauer et al., 2013; Akiva et al., 2014; Furnham et al., 2014). The structural features include dihedral angles, Van der Waals (VdW) interactions, and hydrogen bonds (H-bonds). Intra-EFL interactions and interactions between the functional loop and the rest of the structure are encoded separately. Thus, a descriptor contains information about the immediate environment of the functional loop in all protein scaffolds and enzymatic functions where it was encountered.

## Objective Function for Protein Engineering and Design

In protein engineering and *de novo* design, the descriptors of elementary function need to be integrated into a given structural scaffold. Once the elementary function that is required to be incorporated into the protein is selected, the sequence, structure, and interactions that would fit best into the scaffold have to be determined. The objective function scores how well a given structural loop fits in and should be maximized to obtain the best matching implementation of the descriptor with an assumption that the native structure has the best fit. Essentially, the score represents the joint likelihood of all amino acid positions in the grafted loop with respect to distributions parametrized in the descriptor with N-residue positions and M features: $F = \sum_{\mathrm{i}}^{\mathrm{N}} \sum_{\mathrm{j}}^{\mathrm{M}} W_{\mathrm{i}}^{P} W_{\mathrm{ij}}^{\mathrm{F}} S_{\mathrm{ij}}$. The weight that is given to a residue position $\mathrm{W_i^P}$ reflects the relative degree of conservation of features in each position: $\mathrm{W_i^P} = \frac{\sum_{\mathrm{j=0}}^{\mathrm{M}} \mathrm{W_{ij}^F}}{\sum_{\mathrm{i=0}}^{\mathrm{N}} \sum_{\mathrm{j=0}}^{\mathrm{M}} \mathrm{W_{ij}^F}}$ . Descriptor features $j$ enumerate along a sequence signature ($j = a$), dihedral angles ($j = d$), H-bonds ($j = h$), and vdW interactions ($j = v$), with the corresponding scores $S_{\mathrm{ij}}$ and weights $w_{\mathrm{ij}}$. Score $S_{i,a}$ is the log-odds score for amino acid substitution according to BLOSUM62 (Henikoff and Henikoff, 1993). The weight of the residue feature is given by the relative frequency of the two most frequent residues in the sequence profile $\mathrm{W_{i,a}^F} = \frac{\sum_{\mathrm{k}}^{2} (\mathrm{argmax_k w_k})}{\sum^{20} \mathrm{w_k}}$, where $\mathrm{w_k}$ is amino acid frequency. Dihedral angles are first clustered as two-dimensional vector quantities using the EM algorithm as implemented in scikit-learn, and their weights and fitting scores are derived from parameters in the trained model with the following equation:

$\mathrm{S_{i,d}} = 1 - \mathrm{erf}\left(\sum \mathrm{D^T C D}\right)$, where *erf* refers to the error function, D = X - Y is the displacement of the compared dihedral angle points, normalized by the median of all points in the descriptor, expressed as phi-psi angles, and *C* is a matrix proportional to the degree of confidence that the point belongs to the distribution. Given the precision matrix $\Lambda = \sigma^{-2}$ and posterior probability matrix *P* obtained from the clustered model, $\mathrm{C} = \Lambda\ \mathrm{P}$.

The isotropic VdW distribution score, $\mathrm{S_{i,v}}$, is based on counting the number of VdW contacts with non-hydrogen atoms within the 5-Å radius. Hydrogen bonds are additionally split into acceptors and donors: $\mathrm{S_{i,hA}}$ and $\mathrm{S_{i,hD}}$, respectively. They are scalar quantities and are measured as the number of donor H-bonds present at the residue position. For H-bonds, the effective radius is 3.5 Å. The score is the ratio between the absolute difference in the number of bonds ($\delta$) to the higher number of contacts for either structure (VdW interactions and H-bonds): $\mathrm{S_{i,\ [d,v,hA,hD]}} = \frac{\mathrm{abs}(\delta_{\mathrm{x}} - \delta_{\mathrm{y}})}{\mathrm{max}(\delta_{\mathrm{x}},\ \delta_{\mathrm{y}})}$. The weight of the feature is the SD of the spread of a half-normal distribution fitted onto the data $\mathrm{W^F_{i\ [b,c,d,e]}} = \sigma$. Each feature weight is further tuned by an empirically derived scalar factor to account for differences in the spread of absolute values returned by the scoring functions.

# RESULTS

## Deriving the Descriptors of EFs

**Figure 1** contains examples of some of the structural features for three descriptors: the calcium-binding $Ca^{2+}$-binding helix–loop–helix EF-hand motif (Gifford et al., 2007) (with the characteristic signature DxDxD, **Figure 1A**), the glycine-rich motif of the phosphate-binding in dinucleotide-containing ligands (with the characteristic signature GxGxxG, **Figure 1B**), and the phosphate-binding P-loop in nucleotide-containing ligands (with the characteristic signature GxxGxG, **Figure 1C**). For EF-hand (Gifford et al., 2007), dihedral angles on amino acid positions before and after the calcium-binding sites form tight clusters, i.e., the backbone is structurally conserved. Donor-acceptor hydrogen bond pairs spaced three to four residue positions apart are also present in the first and last 10 residues of the EF-section (**Figure 1A**), representing, together with conserved dihedral angles, two α-helices in the structural motif. Distributions of VdW contacts in the EF-hand motif show a large proportion of the conserved intrinsic contacts flanking the α-helices, whereas the central flexible link between them forms vital external contacts with the rest of the fold. The same patterns of the internal contacts are observed in the second halves of the GxGxxG and GxxGxG motif structures also containing the α-helices. The β-strands in these loops, as expected, interact more strongly with the rest of the structure via VdW interactions (**Supplementary Figure 2**) and H-bonds (**Supplementary Figures 3**, **4**). In addition to the interactions with ligands, the substrate-binding residues in the descriptor are characterized by multiple interactions with the fold.

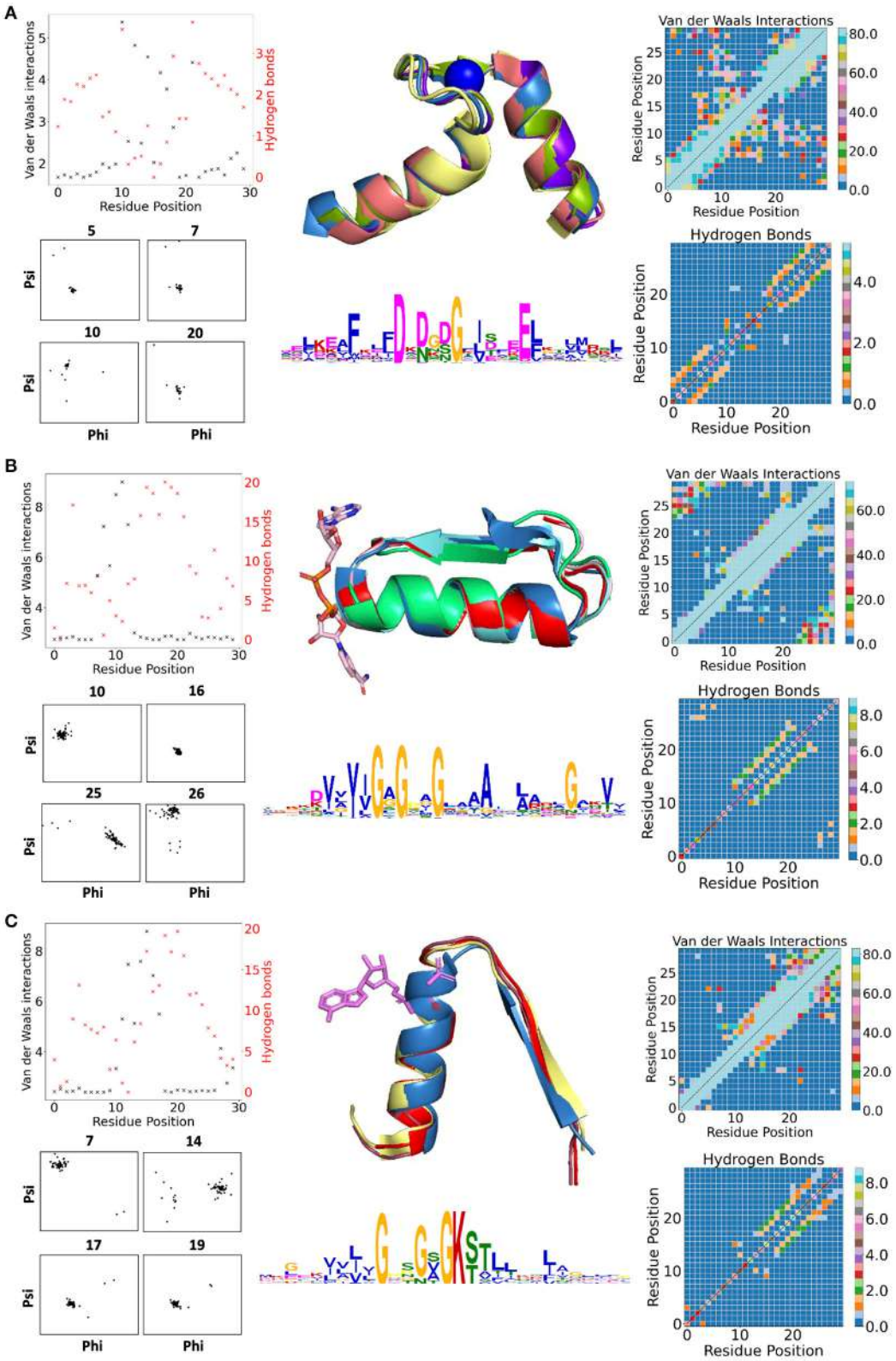


**FIGURE 1 | (A)** Descriptors for $Ca^{2+}$-binding helix-loop-helix EF-hand motif (Gifford et al., 2007), **(B)** phosphate-binding loop in dinucleotide-containing ligands, and **(C)** phosphate-binding P-loop. Charts in the left column show per-residues numbers of Van der Waals (VdW) interactions and hydrogen bonds (H-bonds); contact maps in the right column show the total number of VdW contacts in H-bonds in all structures used for the derivation of corresponding descriptors. The central column contains examples of structures used for derivation of corresponding descriptors (represented in the form of structurally aligned segments) along with the logos representing the position-specific matrices of derived descriptors. The numbering of residues is sequential and follows positions in the logo.

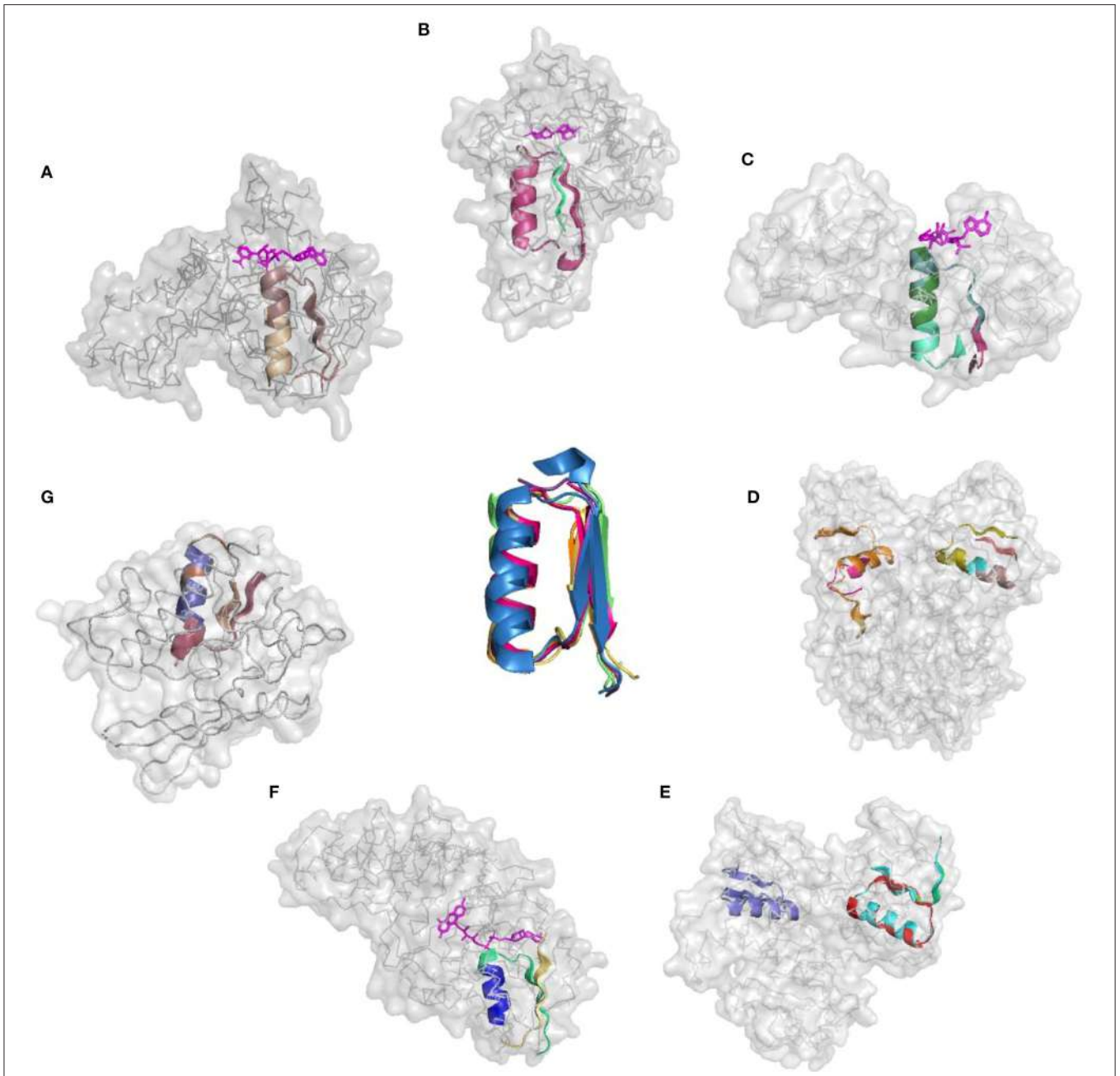


**FIGURE 2 |** Realizations of the descriptor of Gly-rich signature of the phosphate binding in dinucleotide-containing ligands. **(A–G)** PDB IDs of corresponding structures are: 1A5Z, 1JWB, 1LUA, 1PS9, 1ZMC, 2BI7, and 2CUL, respectively.

## Swapping the EFLs Within the Same Structural/Functional Context

The objective function was calibrated and exemplified on three motifs: (i) the EF-hand from *Bos taurus* calcium-binding protein structure (PDBID 1A29) with the $Ca^{2+}$ ligand; (ii) the phosphate-binding motif (GxGxxG) from *Equus caballus* oxidoreductase (PDBID 1A71) with dinucleotide-containing NAD ligand; and (iii) the phosphate-binding motif (GxxGxG) from *Methanocaldococcus jannaschii* ABC transporter (PDBID 1G6H) with nucleotide-containing an ADP ligand. These motifs were replaced with the realization of the corresponding descriptors (see **Supplementary Figure 5** and explanations in the figure caption).

**Supplementary Figure 6** illustrates how objective function can be used to obtain the best descriptor realization in the corresponding engineering or design tasks. First, a segment

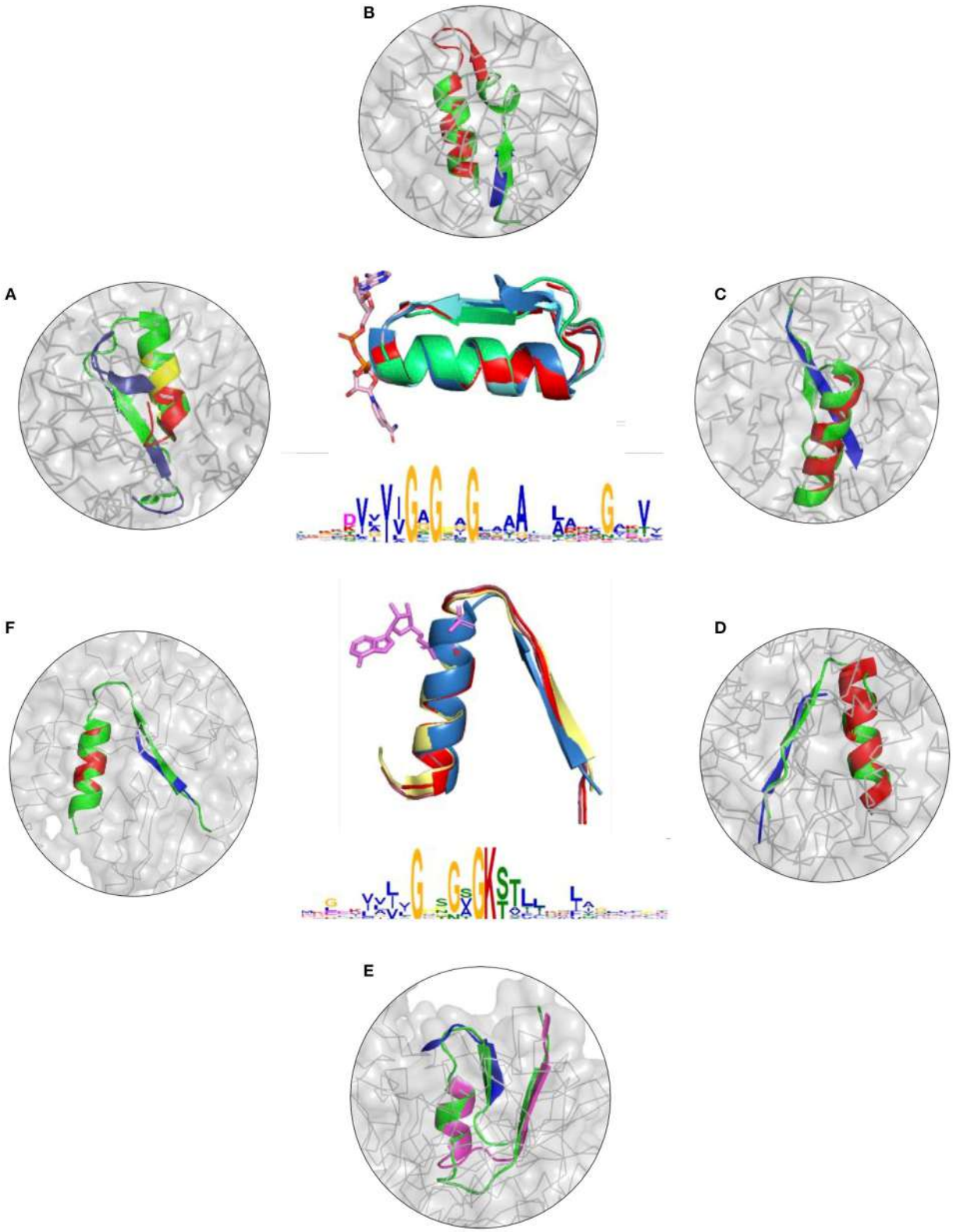


**FIGURE 3 | (A–F)** Cross-grafting of functional loops of the phosphate binding in dinucleotide (GxGxxG) and nucleotide-containing (GxxGxG) ligands. **(A–C)** Grafting of the phosphate-binding signature in dinucleotide-containing ligands (GxGxxG) in proteins with P-loop (GxxGxG) elementary function; recombinant realizations of descriptors are shown in proteins with PDB IDs: 1H5Y, 1BWV, and 1O5K. **(D–F)** Recombinant realizations of the descriptor of P-loop (GxxGxG) elementary function in proteins (1SKY, 1II2, and 1NI3, respectively) with elementary functional loops of the phosphate-binding in dinucleotide-containing ligands (GxGxxG). The original structures are shown in green.

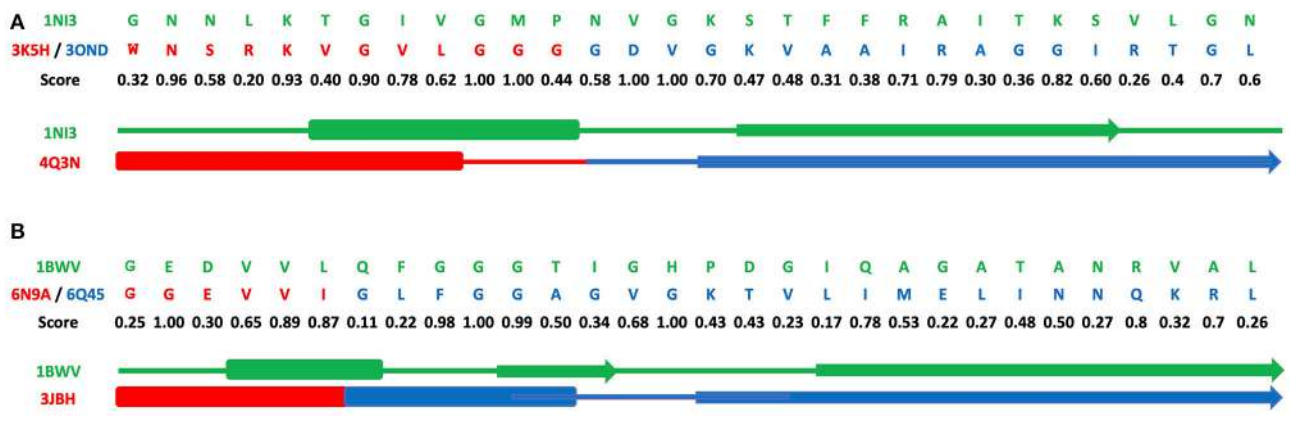


**FIGURE 4 |** Alignments for recombinant cross-grafting of the phosphate-binding loop in dinucleotide (GxGxxG) and nucleotide-containing (GxxGxG) ligands. **(A)** Replacement of the P-loop elementary function in 1NI3 and **(B)** the phosphate-binding signature in dinucleotide-containing ligands in 1BWV. The thick, thin, and arrow sections represent the beta sheet, turn section, and alpha helix secondary structures, respectively.

of seven consecutive residues (half of the typical functional signature; Berezovsky et al., 2017a) with the highest cumulative fitting score is determined. Then, this segment is extended with residues that contribute the highest scores to the fitting function. There might be insertions/deletions in the descriptor realization, or the reference could be undefined as in the case of missing data, with disordered or structurally unresolved regions in the query structure. In such a case, the algorithm first scores all valid residue positions and finds the appropriate segments to merge if the best match does not come from a singular structure. Next, adjacent segments are extended, possibly from both ends in the case of a gap, to fill in the uncharacterized positions. This returns the best-effort re-engineered structure that partially relies on the input structure and fills in the rest, where data are missing. **Supplementary Figure 6** exemplifies the case of descriptor realization that builds the functional loops out of two segments, providing the most optimal score.

## Swapping EFLs Between Different Functions and Structures in the Cross-Validation Experiment

To assess the robustness of the derived descriptors and versatility of the fitting function, we selected seven structures with the phosphate-binding functionality, but of different folds, origins, and ligands: NADP-binding *Thermotoga maritima* lactate dehydrogenase (1A5Z) and *Methylobacterium extorquens AM1* methylene-tetrahydromethanopterin dehydrogenase (1LUA), AMP-binding *Escherichia coli* MoeB-MoaD protein complex (1JWB), flavin–adenine dinucleotide (FAD)-binding *Klebsiella pneumoniae* udp-galactopyranose mutase (2BI7) and *Thermus thermophilus GidA-related protein*, and both binding sites for FAD and NADP in *E. coli* CoA reductase (1PS9) and human dihydrolipoamide dehydrogenase (1ZMC).

The donor functional loop that had to be transplanted is an elementary function of phosphate binding in dinucleotide-containing ligands (GxGxxG). We derived its descriptor from the dataset that excluded the aforementioned structures. Seven proteins representing variations in the structure and sequences of the phosphate-binding functional loops (both in nucleotide- and dinucleotide-containing ligands) were used as the targets or acceptors for the EFL-replacement procedure, using the above descriptor. **Figure 2** illustrates the fit between the re-engineered functional loop into the original structure, which binds various ligands, including AMP, FAD, and NAD(P). Additionally, recombinant functional loops can be built from the best matching segments of multiple structures. Taking 1PS9 structure as an example, its NADP binding site is replaced with the loops from structures 1KF6 from *E. coli* quinol-fumarate reductase; 1VRP from *C. sarcosine* oxidase; 6GNC from *Clostridium acetobutylicum* thioredoxin reductase; and 5ER0 from *Lactobacillus* oxidase. All of these structures contain FAD-binding sites replacing the original NADP-binding site.

## Proof-of-Principle Grafting Experiment Using Descriptors of the Phosphate-Binding EFLs With GxGxxG and GxxGxG Signatures

To further illustrate the utility of descriptors of elementary functions and the potential of the DEFINED-PROTEINS package in protein design applications, we set up initial conditions for the grafting procedure where we replaced the original functional loop in a given protein with the non-native one representing a different elementary function. We recruited two distinct, but not opposite, elementary functions, for which we have already derived descriptors: binding of the phosphate in dinucleotide-(GxGxxG) and nucleotide-containing P-loop (GxxGxG) ligands (Zheng et al., 2016). Thereby, we cross-grafted nucleotide-binding function into protein folds with native dinucleotide-ligand binding (**Figures 3A–C** and **Supplementary Figures 7A–C**), and cross-grafted dinucleotide-binding function into folds with the native mononucleotide binding ability (**Figures 3D–F** and **Supplementary Figures 7D–F**), respectively. Although both are

elementary functions of the phosphate binding, the latter belong to different dinucleotide- and nucleotide-containing ligands that determine the corresponding diversity of protein functions that these elementary functions evolved into.

The signatures of EFLs and the most frequent interactions with distinct parts of ligands were discussed elsewhere (Berezovsky, 2019), showing both conservatism and diversity depending on the position in the functional loop. From the structural perspective, both EFLs have similar secondary structures of β-turn-α-helix composition and architecture. The latter results in a high number of intrinsic H-bonds and VdW contacts in its second α-helical part, while β-strand elements of both loops form more contacts with the surrounding structure. At the same time, the dinucleotide-binding-GxGxxG-loop is a more compact structure by itself, with more intrinsic contacts between its α and β elements. We replaced the original functional segments with recombinant (**Figure 3**) and deconvoluted single-loop (**Supplementary Figure 7**) functional loops sampled from the corresponding descriptors. In all of the cases presented in **Figure 3** and **Supplementary Figure 7** replacements were done with the highest-scoring matches.

**Figures 3A–C** and **Supplementary Figures 7A–C** show results of grafting of the P-loop (GxxGxG) descriptor in places of Gly-rich signature (GxGxxG) of the phosphate-binding in dinucleotide-containing ligands. While obtaining a good match between the original and replacement loops in both unblended single-loop and recombinant cases, the latter was allowed to obtain a better per-residue fit between the original and replacement loops. Recombinant loop replacements consist of several fragments (typically, two to three), covering most of the loop (see examples in **Figure 4** and **Supplementary Figure 8**) with higher scores in the functional positions at the expense of non-function-bearing positions in the secondary structure elements. It agrees with less conserved dihedral angles in the turn segments of the loops that contain functional signatures, which also result in different angles between the secondary structure elements of the loop and difference in the intra-loop contacts in GxxGxG and GxGxxG loops.

**Figure 4** and **Supplementary Figure 8** show a comparison of secondary structure alignments in several examples of the single-loop and recombinant replacements. The secondary structure is affected by the environment (Minor and Kim, 1996), pointing to the need for adjustments and optimization in length and location after the original descriptor realization is placed instead of the natural ELF. Overall, examples of structural replacements (**Figures 3** and **Supplementary Figure 7**) and alignments (**Figure 4** and **Supplementary Figure 8**) show that realizations of descriptors can be used as a starting point for further optimization of positions and interactions involving functional residues and structures of elementary functional loops to obtain required modifications/design in the context of new protein structures and functions.

It is important to note that, in general, scores obtained with the objective function depend on the sequence/structure characteristics of the elementary function and the type of procedure in which the descriptor is used. The major contributors to the objective function score are sequence conservation and dihedral angles, with weights 0.62/0.52 and 0.3/0.38 for the GxxGxG/GxGxxG signatures, respectively. Weights for VdW interactions and H-bonds are smaller (see **Supplementary Table 1** legend for details): the VdW interactions are weak though omnipresent, whereas H-bonds between the loops and rest of the folds are rare, making the weights of both smaller. Nevertheless, contributions of all characteristics to the weight for each individual position are calculated as a sum of their weights normalized by the total feature weights across all residue positions. The score of the reference state for the descriptor, which can be indicative of the latter and can be used as a guideline in design, should be obtained for each descriptor. The straightforward way is to perform the cross-validation experiment, which assesses the robustness of the descriptor and provides the score that can be used as a ground state score. **Supplementary Table 1B** contains averaged scores obtained in the cross-validation experiment (**Figure 2**), which can be used as a reference for the engineering and design of corresponding descriptors in other folds. As an example, **Supplementary Table 2B** shows scores in case of cross-grafting of descriptors of the phosphate-binding signatures in the nucleotide- (GxxGxG) and dinucleotide-containing (GxGxxG) ligands, revealing the deviation from scores in cross-validation that can further increase in case of *de novo* design of functions based on the descriptors. The utility of averaged (**Supplementary Tables 1A**, **2A**) and per-residue scores is in the information on the relative conservatism (importance) of positions in the descriptor and of the overall match between its realization and the rest of the fold and its capacity to contribute to engineered/designed function. For example, depending on the requirements on the interactions within the ELF and between the loop and the rest of the fold, H-bonds can be introduced in positions with a conservatism level allowing to do so, but the same holds for changing other characteristics.

## DISCUSSION

There are two major tasks, engineering/modification and *de novo* design, in which descriptors of elementary functions can be used (Berezovsky, 2019). The former is a modification of the natural protein function by replacing one or several natural functional loops (elementary functions) in the protein with other elementary function(s) encoded in the corresponding descriptor(s). The goal of this engineering effort can be to change a substrate specificity to modify a biochemical function and/or mechanism of the enzyme (Babbitt et al., 1996; Pegg et al., 2006; Berezovsky, 2019). The quest on *de novo* design of the protein with required function can be set by providing the sequence, structure, or the sequence–structure combination (Leaver-Fay et al., 2011; Berezovsky, 2019). The specific task that DEFINED-PROTEINS presented in this study addresses the finding of a structural segment according to the functional descriptor, which will fit best into the original fold, providing required functional and stabilizing interactions with the rest of the fold. Ultimately, DEFINED-PROTEINS is aiming to build catalytic sites with

desired activity and interactions while maintaining the overall fold structure and stability.

Steady progress in the computational design of new topologies and functions (Huang et al., 2014, 2016b; King et al., 2015) and even a stronger drive toward *de novo* protein design (Huang et al., 2016a; Lechner et al., 2018; Baker, 2019; Silva et al., 2019) prompt the use of the basic units that would possess all traits determined by the polymer nature of proteins (Yamakawa and Stockmayer, 1972; Shimada and Yamakawa, 1984; Berezovsky et al., 2000, 2017a; Orevi et al., 2013; Jacob et al., 2018), their evolutionary history, and requirements on the structural stability and dynamics, as well as show the required functional activity (Berezovsky, 2019; Romero-Romero et al., 2021). The concept of the elementary function standard/descriptor allows one to consider individual steps of biochemical functions provided by physics-based and evolutionary selected ELFs (Berezovsky, 2019). An exhaustive description of elementary functions including all sequence, structure, and functional information that can be used in the design of new biochemical functions consisting of different combinations of elementary ones. Ultimately, the set of descriptors should, as exhaustively as possible, represent the diversity of sequence signatures that perform this elementary function, as well as the diversity of their structural implementations in different protein folds. In the combinations of descriptors into the desired biochemical function, the observables of the parameters of descriptors will be the result of the interference between their distributions and the type of the final structure that carries the function, the type of the overall transformation, interactions with other descriptors involved in the construction, and interactions with the substrate. Ideally, it should be possible to use descriptors of elementary functional units to build the geometry, the set of interactions, and the environment necessary for specified biochemical function. Then, the library of these units is supposed to be used in design efforts such as, the Rosetta enzyme design protocol (Leaver-Fay et al., 2011), coupled with which the placement and refinement of a DEFINED-PROTEINS functional unit into a designable scaffold, which would include modeling of intra-loop and loop-fold interactions and energy minimization toward stable structure with required dynamics, would be enabled.

## CONCLUSIONS

The DEFINED-PROTEINS is a proof-of-concept implementation that provides the tools to: (i) derive descriptor of elementary functions of interest directly from protein structures and (ii) apply descriptors in computational engineering and design. The software is available as a Python package that allows for integration in custom protein design workflows. We exemplified the derivation of the descriptor on three elementary functions: the calcium-binding EF-hand (Gifford et al., 2007), the glycine-rich motif of the phosphate-binding in dinucleotide-containing ligands, and the mononucleotide phosphate-binding P-loop. We also assessed the robustness of derived descriptors and the objective function by replacing phosphate-binding EFLs in seven proteins with different functions derived on the set of proteins excluding the above seven proteins. Finally, we demonstrated a proof-of-principle grafting experiment by cross-replacing functional loops between P-loop containing proteins and those with the elementary function of the phosphate-binding in dinucleotide-containing ligands.

## DATA AVAILABILITY STATEMENT

The software was built primarily using Python3 (http://www.python.org), C with an optional MPI requirement, and C++. Django (https://djangoproject.com) was used as the web framework with embedded interactive Bokeh plots (https://bokeh.org/). Docker container is available for OS-independent deployment. The software is BSD-licensed.

## AUTHOR CONTRIBUTIONS

IB: conceptualization, supervision, project administration, and funding acquisition. AG and IB: methodology. AG, MY, and IB: investigation, formal analysis, and writing—review and editing. MY: software and visualization. MY and IB: writing—original draft. All authors contributed to the article and approved the submitted version.

## FUNDING

This financial support provided by the Biomedical Research Council, via Agency for Science, Technology, and Research (A*STAR), is greatly appreciated. AG was supported in part by the Intramural Research Programs of the National Library of Medicine, National Institutes of Health.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbinf.2021.657529/full#supplementary-material

## REFERENCES

Akiva, E., Brown, S., Almonacid, D. E., Barber, A. E. 2nd, Custer, A. F., Hicks, M. A., Huang, C. C., et al. (2014). The structure-function linkage database. *Nucleic Acids Res.* 42, D521–D530. doi: 10.1093/nar/gkt1130

Andreini, C., Bertini, I., Cavallaro, G., Holliday, G. L., and Thornton, J. M. (2009). Metal-MACiE: a database of metals involved in biological catalysis. *Bioinformatics* 25, 2088–2089. doi: 10.1093/bioinformatics/btp256

Aziz, M. F., Caetano-Anolles, K., and Caetano-Anolles, G. (2016). The early history and emergence of molecular functions and modular scale-free network behavior. *Sci. Rep.* 6:25058. doi: 10.1038/srep25058

Babbitt, P. C., Hasson, M. S., Wedekind, J. E., Palmer, D. R., Barrett, W. C., Reed, G. H., et al. (1996). The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids. *Biochemistry* 35, 16489–16501. doi: 10.1021/bi9616413

Bairoch, A. (2000). The ENZYME database in *Nucleic Acids Res.* 28, 304–305. doi: 10.1093/nar/28.1.304

Baker, D. (2019). What has *de novo* protein design taught us about protein folding and biophysics? *Protein Sci.* 28, 678–683. doi: 10.1002/pro.3588

Berezovsky, I. N. (2003). Discrete structure of van der Waals domains in globular proteins. *Protein engineering* 16, 161–167. doi: 10.1093/proeng/gzg026

Berezovsky, I. N. (2019). Towards descriptor of elementary functions for protein design. *Curr. Opin. Struct. Biol.* 58, 159–165. doi: 10.1016/j.sbi.2019.06.010

Berezovsky, I. N., Grosberg, A. Y., and Trifonov, E. N. (2000). Closed loops of nearly standard size: common basic element of protein structure. *FEBS Lett.* 466, 283–286. doi: 10.1016/S0014-5793(00)01091-7

Berezovsky, I. N., Guarnera, E., and Zheng, Z. (2017a). Basic units of protein structure, folding, and function. *Progr. Biophys. Mol. Biol.* 128, 85–99. doi: 10.1016/j.pbiomolbio.2016.09.009

Berezovsky, I. N., Guarnera, E., Zheng, Z., Eisenhaber, B., and Eisenhaber, F. (2017b). Protein function machinery: from basic structural units to modulation of activity. *Curr. Opin. Struct. Biol.* 42, 67–74. doi: 10.1016/j.sbi.2016.10.021

Berezovsky, I. N., Kirzhner, A., Kirzhner, V. M., Rosenfeld, V. R., and Trifonov, E. N. (2003a). Protein sequences yield a proteomic code. *J. Biomol. Struct. Dyn.* 21, 317–325. doi: 10.1080/07391102.2003.10506928

Berezovsky, I. N., Kirzhner, A., Kirzhner, V. M., and Trifonov, E. N. (2003b). Spelling protein structure. *J. Biomol. Struct. Dyn.* 21, 327–339. doi: 10.1080/07391102.2003.10506929

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic Acids Res.* 28, 235–242. doi: 10.1093/nar/28.1.235

Bowie, J. U., Luthy, R., and Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253, 164–170. doi: 10.1126/science.1853201

Brunette, T. J., Parmeggiani, F., Huang, P. S., Bhabha, G., Ekiert, D. C., Tsutakawa, S. E., et al. (2015). Exploring the repeat protein universe through computational protein design. *Nature* 528, 580–584. doi: 10.1038/nature16162

Crippen, G. M. (1996). Failures of inverse folding and threading with gapped alignment. *Proteins* 26, 167–171. doi: 10.1002/(SICI)1097-0134(199610)26:2<167::AID-PROT6>3.0.CO;2-D

Das, R., and Baker, D. (2008). Macromolecular modeling with rosetta. *Annu. Rev. Biochem.* 77, 363–382. doi: 10.1146/annurev.biochem.77.062906.171838

Dill, K. A., and MacCallum, J. L. (2012). The protein-folding problem, 50 years on. *Science* 338, 1042–1046. doi: 10.1126/science.1219021

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., et al. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44, D279–D285. doi: 10.1093/nar/gkv1344

Fischer, J. D., Holliday, G. L., and Thornton, J. M. (2010). The CoFactor database: organic cofactors in enzyme catalysis. *Bioinformatics* 26, 2496–2497. doi: 10.1093/bioinformatics/btq442

Furnham, N., Holliday, G. L., de Beer, T. A., Jacobsen, J. O., Pearson, W. R., and Thornton, J. M. (2014). The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res.* 42, D485–D489. doi: 10.1093/nar/gkt1243

Gifford, J. L., Walsh, M. P., and Vogel, H. G. (2007). Structures and metal-ion-binding properties of the $Ca^{2+}$-binding helix–loop–helix EF-hand motifs. *Biochem. J.* 405, 199–221. doi: 10.1042/BJ20070255

Goncearenco, A., and Berezovsky, I. N. (2010). Prototypes of elementary functional loops unravel evolutionary connections between protein functions. *Bioinformatics* 26, i497–i503. doi: 10.1093/bioinformatics/btq374

Goncearenco, A., and Berezovsky, I. N. (2011). Computational reconstruction of primordial prototypes of elementary functional loops in modern proteins. *Bioinformatics* 27, 2368–2375. doi: 10.1093/bioinformatics/btr396

Goncearenco, A., and Berezovsky, I. N. (2012). Exploring the evolution of protein function in Archaea. *BMC Evol. Biol.* 12:75. doi: 10.1186/1471-2148-12-75

Goncearenco, A., and Berezovsky, I. N. (2015). Protein function from its emergence to diversity in contemporary proteins. *Phys. Biol.* 12:045002. doi: 10.1088/1478-3975/12/4/045002

Henikoff, S., and Henikoff, J. G. (1993). Performance evaluation of amino acid substitution matrices. *Proteins* 17, 49–61. doi: 10.1002/prot.340170108

Hocker, B. (2014). Design of proteins from smaller fragments-learning from evolution. *Curr. Opin. Struct. Biol.* 27, 56–62. doi: 10.1016/j.sbi.2014.04.007

Holliday, G. L., Andreini, C., Fischer, J. D., Rahman, S. A., Almonacid, D. E., Williams, S. T., et al. (2012). MACiE: exploring the diversity of biochemical reactions. *Nucleic Acids Res.* 40, D783–D789. doi: 10.1093/nar/gkr799

Holliday, G. L., Bartlett, G. J., Almonacid, D. E., O'Boyle, N. M., Murray-Rust, P., Thornton, J. M., et al. (2005). MACiE: a database of enzyme reaction mechanisms. *Bioinformatics* 21, 4315–4316. doi: 10.1093/bioinformatics/bti693

Huang, P. S., Boyken, S. E., and Baker, D. (2016a). The coming of age of *de novo* protein design. *Nature* 537, 320–327. doi: 10.1038/nature19946

Huang, P. S., Feldmeier, K., Parmeggiani, F., Velasco, D. A. F., Hocker, B., and Baker, D. (2016b). *De novo* design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nat. Chem. Biol.* 12, 29–34. doi: 10.1038/nchembio.1966

Huang, P. S., Oberdorfer, G., Xu, C., Pei, X. Y., Nannenga, B. L., Rogers, J. M., et al. (2014). High thermodynamic stability of parametrically designed helical bundles. *Science* 346, 481–485. doi: 10.1126/science.1257481

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., et al. (2009). InterPro: the integrative protein signature database. *Nucleic Acids Res.* 37, D211–D215. doi: 10.1093/nar/gkn785

Jacob, M. H., D'Souza, R. N., Schwarzlose, T., Wang, X., Huang, F., Haas, E., et al. (2018). Method-unifying view of loop-formation kinetics in peptide and protein folding. *J. Phys. Chem. B* 122, 4445–4456. doi: 10.1021/acs.jpcb.8b00879

King, I. C., Gleixner, J., Doyle, L., Kuzin, A., Hunt, J. F., Xiao, R., et al. (2015). Precise assembly of complex beta sheet topologies from *de novo* designed building blocks. *Elife* 4:e53865. doi: 10.7554/eLife.11012.020

Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Thompson, J., et al. (2011). ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* 487, 545–574. doi: 10.1016/B978-0-12-381270-4.00019-6

Lechner, H., Ferruz, N., and Hocker, B. (2018). Strategies for designing non-natural enzymes and binders. *Curr. Opin. Chem. Biol.* 47, 67–76. doi: 10.1016/j.cbpa.2018.07.022

Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., et al. (2015). CDD: NCBI's conserved domain database. *Nucleic Acids Res.* 43, D222–D226. doi: 10.1093/nar/gku1221

Marchler-Bauer, A., Zheng, C., Chitsaz, F., Derbyshire, M. K., Geer, L. Y., Geer, R. C., et al. (2013). CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res.* 41, D348–352. doi: 10.1093/nar/gks1243

Minor, D. L. Jr., and Kim, P. S. (1996). Context-dependent secondary structure formation of a designed protein sequence. *Nature* 380, 730–734. doi: 10.1038/380730a0

Nath, N., Mitchell, J. B., and Caetano-Anolles, G. (2014). The natural history of biocatalytic mechanisms. *PLoS Comput. Biol.* 10:e1003642. doi: 10.1371/journal.pcbi.1003642

Orevi, T., Rahamim, G., Hazan, G., Amir, D., and Haas, E. (2013). The loop hypothesis: contribution of early formed specific non-local interactions to the determination of protein folding pathways. *Biophys. Rev.* 5, 85–98. doi: 10.1007/s12551-013-0113-3

Pegg, S. C., Brown, S. D., Ojha, S., Seffernick, J., Meng, E. C., Morris, J. H., et al. (2006). Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database. *Biochemistry* 45, 2545–2555. doi: 10.1021/bi052101l

Romero Romero, M. L., Rabin, A., and Tawfik, D. S. (2016). Functional proteins from short peptides: dayhoff's hypothesis Turns 50. *Angew. Chem. Int. Ed. Engl.* 55, 15966–15971. doi: 10.1002/anie.201609977

Romero Romero, M. L., Yang, F., Lin, Y. R., Toth-Petroczy, A., Berezovsky, I. N., Goncearenco, A., et al. (2018). Simple yet functional phosphate-loop proteins. *Proc. Natl. Acad. Sci. U.S.A.* 115, E11943–E11950. doi: 10.1073/pnas.1812400115

Romero-Romero, S., Kordes, S., Michel, F., and Hocker, B. (2021). Evolution, folding, and design of TIM barrels and related proteins. *Curr. Opin. Struct. Biol.* 68, 94–104. doi: 10.1016/j.sbi.2020.12.007

Rooman, M. J., Rodriguez, J., and Wodak, S. J. (1990). Relations between protein sequence and structure and their significance. *J. Mol. Biol.* 213, 337–350. doi: 10.1016/S0022-2836(05)80195-0

Rooman, M. J., and Wodak, S. J. (1995). Are database-derived potentials valid for scoring both forward and inverted protein folding? *Protein Eng.* 8, 849–858. doi: 10.1093/protein/8.9.849

Sali, A., Shakhnovich, E., and Karplus, M. (1994). How does a protein fold? *Nature* 369, 248–251. doi: 10.1038/369248a0

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature* 577, 706–710. doi: 10.1038/s41586-019-1923-7

Shakhnovich, E. (2006). Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet. *Chem. Rev.* 106, 1559–1588. doi: 10.1021/cr040425u

Shakhnovich, E. I., and Gutin, A. M. (1993). Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl. Acad. Sci. U.S. A.* 90, 7195–7199. doi: 10.1073/pnas.90.15.7195

Shimada, J., and Yamakawa, H. (1984). Ring-closure probabilities for twisted wormlike chains. Application to DNA. *Macromolecules* 17, 689–698. doi: 10.1021/ma00134a028

Silva, D. A., Yu, S., Ulge, U. Y., Spangler, J. B., Jude, K. M., Labao-Almeida, C., et al. (2019). *De novo* design of potent and selective mimics of IL-2 and IL-15. *Nature* 565, 186–191. doi: 10.1038/s41586-018-0830-7

Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* 213, 859–883. doi: 10.1016/S0022-2836(05)80269-4

Trifonov, E. N., Kirzhner, A., Kirzhner, V. M., and Berezovsky, I. N. (2001). Distinct stages of protein evolution as suggested by protein sequence analysis. *J. Mol. Evol.* 53, 394–401. doi: 10.1007/s002390010229

Trudeau, D. L., and Tawfik, D. S. (2019). Protein engineers turned evolutionists-the quest for the optimal starting point. *Curr. Opin. Biotechnol.* 60, 46–52. doi: 10.1016/j.copbio.2018.12.002

wwPDB consortium (2019). Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* 47, D520–D528. doi: 10.1093/nar/gky949

Yamakawa, H., and Stockmayer, W. H. (1972). Statistical mechanics of wormlike chains. II. Excluded volume effects. *J. Chem. Phys.* 57, 2843–2854. doi: 10.1063/1.1678675

Zheng, Z., Goncearenco, A., and Berezovsky, I. N. (2016). Nucleotide binding database NBDB–a collection of sequence motifs with specific protein-ligand interactions. *Nucleic Acids Res.* 44, D301–D307. doi: 10.1093/nar/gkv1124

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Copyright © 2021 Yin, Goncearenco and Berezovsky. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*