AlphaDesign: A *de novo* protein design framework based on AlphaFold

Michael Jendrusch ,[*] Jan O. Korbel ,[†] and S. Kashif Sadiq [‡]
Genome Biology Unit, European Molecular Biology Laboratory (EMBL),
Meyerhofstrasse 1, Heidelberg 69117, Germany

De novo protein design is a longstanding fundamental goal of synthetic biology, but has been hindered by the difficulty in reliable prediction of accurate high-resolution protein structures from sequence. Recent advances in the accuracy of protein structure prediction methods, such as AlphaFold (AF), have facilitated proteome scale structural predictions of monomeric proteins. Here we develop AlphaDesign, a computational framework for *de novo* protein design that embeds AF as an oracle within an optimisable design process. Our framework enables rapid prediction of completely novel protein monomers starting from random sequences. These are shown to adopt a diverse array of folds within the known protein space. A recent and unexpected utility of AF to predict the structure of protein complexes, further allows our framework to design higher-order complexes. Subsequently a range of predictions are made for monomers, homodimers, heterodimers as well as higher-order homo-oligomers - trimers to hexamers. Our analyses also show potential for designing proteins that bind to a pre-specified target protein. Structural integrity of predicted structures is validated and confirmed by standard *ab initio* folding and structural analysis methods as well as more extensively by performing rigorous all-atom molecular dynamics simulations and analysing the corresponding structural flexibility, intramonomer and interfacial amino-acid contacts. These analyses demonstrate widespread maintenance of structural integrity and suggests that our framework allows for fairly accurate protein design. Strikingly, our approach also reveals the capacity of AF to predict proteins that switch conformation upon complex formation, such as involving switches from α-helices to β-sheets during amyloid filament formation. Correspondingly, when integrated into our design framework, our approach reveals *de novo* design of a subset of proteins that switch conformation between monomeric and oligomeric state.

Keywords: AlphaFold, computational protein design, self-assembly, biologics, network inversion, evolutionary algorithms, computational biophysics

I. INTRODUCTION

Proteins are the workhorses of the vast majority of life processes at the cellular scale. They carry out a myriad of functions ranging from the catalysis of biochemical reactions, mechanical functions involved in cell motility and the formation of sub-cellular architecture amongst many others. A central paradigm to understand their function has been based on the observation that proteins fold into complex, yet specific three-dimensional structures that vary depending on their amino-acid sequence, hypothesized to be their lowest energy state [1]. Thus experimental structure determination has been a primary pursuit of biology for the last 50 years [2] and given rise to over 10^5 distinctly solved structures [3]. These have revealed a diverse array of fold topologies and geometries of individual proteins, classified into 41 architectures with 1390 fold topologies [4] and that 51% of structures previously extracted [5] from the protein data bank (PDB) [3] form quaternary structure through oligomer and complex formation [5].

A central driving question has been to what extent amino-acid sequence encodes structure and whether it is then possible to predict structure from sequence. This has fuelled a plethora of diverse computational protein structure prediction approaches across decades of effort [6–14]. The quest for improved methods has been embodied in an independently assessed biennial community-wide competition (CASP) that ranks method accuracy of participants' approaches with respect to determined but unreleased structures resulting in steady progress towards predictive accuracy [15]. Recently, this has culminated in the development of AlphaFold [16] - hereafter, termed AF, a state-of-the-art neural network-based approach that has achieved comparable atomic accuracy with respect to crystal structure. AF has been applied at a proteome-wide scale across several species, resulting in a new database of predicted structures [17]. More recently, a community-wide assessment has revealed several important applications of AF ranging from amino acid variant effect prediction to cryo-EM model building [18]. Whilst intended for single-chain structure prediction, an unexpected consequence of AF's input protocol allows prediction of non-contiguous chains, thus enabling prediction of protein complexes using the existing trained network [18, 19]. This feature has also been applied at proteome-wide

[*] michael.jendrusch@embl.de
[†] jan.korbel@embl.org; corresponding author
[‡] kashif.sadiq@embl.de; corresponding author

scale to reveal novel core eukaryotic complexes [20]. These efforts culminated in the release of AlphaFold-Multimer [21], a version of AF explicitly trained for complex prediction.

Despite these advances, the protein folding problem remains far more complex than structure prediction alone. Proteins are not rigid structures, especially at physiological conditions. Firstly, proteins exhibit thermodynamic equilibrium between folded and unfolded forms [1]. Secondly, many proteins also undergo conformational changes with a well-described equilibrium between states, making use of these changes to enact function [22]. Moreover, there is an abundance of intrinsically disordered proteins (IDPs) - those that do not exhibit stable tertiary structures, or transiently fold depending on environmental context [22–25]. One notable biomedical example being the misfolding of amyloid-α helices into β-sheeted filaments associated with Alzheimer's disease [26]. Intriguingly, metamorphic proteins - those that form multiple, stable, yet different folds - have also been discovered [27].

From a biophysical perspective, protein folding is made comprehensible using the concept of a folding energy funnel within the atomic configuration space [28]. Multiple energy minima then correspond to alternate stable states that make transitions with characterisable rates. Based on this concept, computational physics methods such as molecular dynamics (MD) simulations have provided a route to characterise the dynamics, thermodynamics and kinetics of conformational transitions [29, 30], *ab initio* folding [31], disordered transitions [32, 33] as well as protein-protein [34, 35] and protein-ligand [36] binding.

Notwithstanding the already-mentioned complex structure and dynamics of naturally occurring proteins, evolution has still only explored an infinitesimal portion of the potential protein sequence landscape [37]. There is therefore enormous potential in unlocking the fundamental biophysical principles of protein folding to design and engineer novel proteins that can exploit this vast space. Recent examples include protein logic gates [38], self-assembling systems [39] and targeted therapeutics [40]. Established methods for protein engineering have until recently focused on tuning naturally occurring proteins through iterative experimental selection processes such as directed evolution [41–43]. More recently, computational design approaches have enabled *de novo* protein design, encompassing a full suite of functionalities ranging from rules for topology selection, protein backbone construction, sequence optimisation as well combinations of these approaches [37, 44].

The current *de facto* standard in computational protein design is embodied by the Rosetta suite of protein design tools [45, 46] and Rosetta Remodel [47] in particular. It combines tools for all steps from selecting a desired protein topology to designing and validating a folding protein sequence. Protein design within Rosetta Remodel is composed of a combination of four tasks: topology specification, backbone generation and fixed-backbone design [47]. After specifying a desired protein topology a backbone structure can be generated using matching fragments extracted from existing proteins [45, 47]. Fragments matching a desired secondary structure and set of contacts are selected from PDB structures. Selected fragments are then sampled using Markov-Chain Monte Carlo to arrive at a pool of candidate backbone geometries [47]. Candidate backbone structures can then be equipped with a sequence using fixed-backbone protein design, optionally optimising the backbone between design steps [48]. Given a desired backbone geometry, the goal is to find an amino acid sequence which will fold that structure. Here, the Rosetta Design protocol [48] starts by populating the desired backbone with an all-valine sequence and runs Markov-Chain Monte Carlo to arrive at a low-energy sequence. The general strategy of Rosetta-based protein design has been successfully applied to a variety of design problems. These applications range from the first *de novo* designed proteins [48] to synthetic vaccines [49, 50], complex assemblies [51, 52] and enzymes [53, 54]. However, Markov-Chain Monte Carlo in structure space can be time-consuming and computationally demanding.

To tackle this shortcoming of classical protein design, there has been an increase in approaches applying neural networks to various design problems. Several works train generative models to directly generate protein sequences with a desired function [55–57]. [55] have trained a language model on the UniProt sequence database [58] to generate sequences with a specified function. [59] have expanded on [55] by additionally fine-tuning the generative model on a specific protein family or function. Beyond language models, other types of generative models, such as variational autoencoders [56] and generative adversarial networks [57] have been explored for direct protein sequence generation. While these approaches have been successful in generating protein sequences associated with a given function, they do not explicitly take into account structural information and thus cannot be applied to protein design tasks involving constraints on tertiary structure.

On the other side of the spectrum of neural network-based protein design, generative models have been explored for protein structure generation [60–62]. [60, 63] have trained generative adversarial networks to generate realistic backbone distance maps and coordinates. [62] have achieved the same goal using variational autoencoders. Both approaches have demonstrated fine-grained control over designed backbone structures by latent variable manipulation. While these approaches are a viable alternative to classical backbone design, they offer no guarantees on the designability of their predicted backbones. In contrast, our approach produces backbones and sequences which are by construction designable and high-confidence under AF.

Bridging the gap between structure-only and sequence-only approaches, [64–67] have trained neural networks on the PDB database of protein structures to predict protein sequence given a fixed backbone structure. These approaches rely on network components taking into account the geometry of the protein backbone together with the protein sequence. [64, 65, 67] have used per-residue local coordinate systems to reason about protein geometry, while [66] use network architecture adapted to Euclidean coordinate data. [68] have framed fixed-backbone design as a constraint satisfaction problem and have trained their network as a constraint solver. These methods require a neural network trained for a specific protein design task – fixed-backbone protein design – and by themselves cannot easily be extended to other design applications without retraining.

A number of works explore re-using previously trained predictors of protein structure or function as parts of an *in silico* screening framework [69–76]. In these approaches, a neural network is treated as a score function to evaluate the quality of protein designs. Designs are then improved using gradient-based [71–73], gradient-free [69, 74, 76] or neural-network based [75] optimisation.

[70] were the first to use an optimisation loop incorporating a neural network for structure prediction for *de novo* protein design. Their approach has since been extended to fixed-backbone design [71, 73] and protein scaffold generation for protein motif stabilisation [72]. [71, 72] use both Markov-Chain Monte Carlo and gradient descent as their means for optimisation. All of these approaches have made use of trRosetta [77] as a structure predictor. With the new release of AF, [74] have used it as a structure predictor for fixed-backbone protein design using greedy optimisation, with sequences initialised from a trained model. We expand previous approaches by constructing a flexible family of target functions for optimisation using AF [16] and extending the range of possible design tasks.

In this work, we develop a flexible framework for protein design by sequence optimisation using evolutionary algorithms. We embed AF [16] into a design loop as a structure prediction oracle. Our framework extends previous work on protein design using structure predictors [70, 71] by defining a flexible family of target functions encoding various protein design tasks and integrate this design loop with extensive validation using both Rosetta [45] and molecular dynamics simulations. We apply our framework to design *de novo* protein monomers, dimers and oligomers. We further design binders for target proteins using only their sequence, as well as design proteins which change conformation upon complex formation.

II. *DE NOVO* DESIGN FRAMEWORK DEVELOPMENT

We frame protein design as a search problem to find the set of protein sequences for which a certain target function exceeds a fixed threshold. To take into account both sequence properties and all-atom protein structure, we integrate AF [16] into our target functions to provide high-quality structure prediction and measures of prediction confidence. We combine this with state-of-the art validation using Rosetta *ab initio* structure prediction [45] and molecular dynamics simulations. In the following we describe the components of our framework.

A. Optimisation loop

Our approach to *de novo* protein design follows prior work using trRosetta [70, 71, 77]. We set up an optimisation loop which continuously mutates a pool of sequences, scores them using a neural network-based cost function and updates the sequence pool with the mutated sequences and scores (Fig. 1 A). In contrast to previous approaches using trRosetta, we use AF as a structure prediction oracle [16] and define our optimisation target as a flexible function of protein sequence, all-atom structure and AF confidence.

AF is designed to process a sequence together with a corresponding multiple sequence alignment, it is possible to predict the structure for a single sequence by constructing an alignment containing that sequence alone [19]. While this decreases the capability of AF to predict correct structures with default parameters, previous work shows that increasing the number of AF iterations on a single input (referred to as recycling steps) can rescue prediction quality [19]. AF returns an all-atom structure for the input protein sequence together with predicted confidence measures (Fig. 1 A, AlphaFold outputs). AF confidence is expressed as a combination of the predicted local distance difference test (pLDDT) [78] and the predicted aligned error (pAE) [16]. pLDDT measures local model quality, while pAE provides a measure of confidence for each amino acid pair. We may then optimise sequences to maximise any target function $\mathcal{L}$ of these outputs. These can include minimising the RMSD to a target structure and maximising AF prediction confidence. As AF is suitable for complex structure prediction [18], target functions can also constrain oligomeric state or conformational change upon protein binding (Fig. 1 B). In general, any function of protein structure, sequence and prediction confidence may be used for optimisation.

For each sequence in the sequence pool we compute the value of the target function. Sequences are further recombined and mutated by an optimiser to explore sequence space. The sequence pool is updated with the mutated sequences. In this work, we use a simple evolutionary algorithm following [79]. However, other gradient-free or gradient-based optimisers can be substituted as desired. Throughout optimisation a protein changes its structure and both local and global confidence measures increase (Fig. 1 C). For a given protein design task, a threshold is selected above which a sequence is considered fully optimised. We return sequences above this threshold and submit a subset of them to validation using molecular dynamics simulations and Rosetta *ab initio* structure prediction [45].

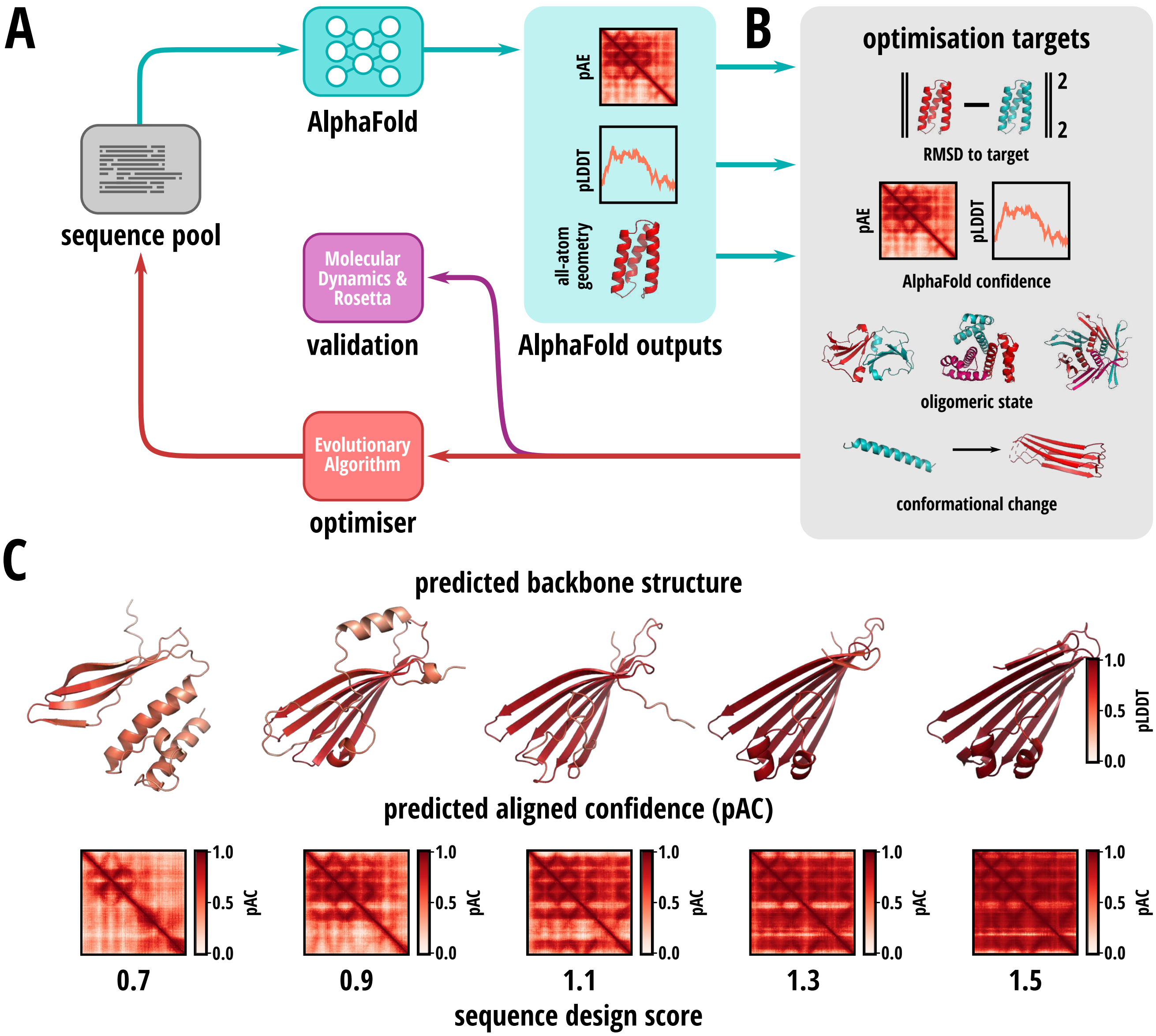

FIG. 1. **Method overview. (A)** Core optimisation loop of AlphaDesign. A pool of amino acid sequences (grey) is maintained throughout the design process. AlphaFold (blue) predicts the all-atom structure and confidence metrics (pLDDT, pAC) for each sequence (light blue). These are combined into a target function maximised in the optimisation loop. The sequences and their target values are fed to an optimiser (red), which updates the sequence pool. If the target value exceeds a desired threshold, the sequence and its structure are submitted for validation using Rosetta and molecular dynamics simulations (purple). **(B)** Target function components. The optimisation target may be any function of the amino acid sequence, predicted all-atom structure and confidence metrics (pLDDT, pAC). Examples include RMSD to a target structure, maximisation of confidence, complex formation and conformational change upon complex formation. **(C)** Snapshots of a sequence at increasing target values during optimisation. For increasing target values from 0.7 to 1.5 the predicted structure (top) changes. Simultaneously, local confidence for each amino acid (pLDDT, top) and predicted aligned confidence for each pair of amino acids (pAC, bottom) increase.

B. Target functions

When using gradient-free methods, we can use any function of the protein sequence, all-atom structure and AF confidence as the optimisation target. This opens up many possibilities for designing proteins and protein complexes with specific properties. In this work, we focus on cost functions making use of AF confidence and protein backbone geometry.

a. Confidence AF models return two main measures of confidence for a protein structure prediction. These are the predicted local distance difference test (pLDDT) [78] and the predicted aligned error (pAE) [16]. pLDDT provides a measure of local confidence for each amino acid, while pAE provides the predicted error of each amino acid position in the local coordinate frame of each other amino acid. For both confidence measures, AF predicts a binned distribution of values. Instead of working directly with pAE values, we instead convert to predicted aligned confidence (pAC), which translates to $pAC = 1 - \mu_{pAE}/pAE_{max}$ where μ_{pAE} is the mean pAE and pAE_{max} is the center of the highest pAE bin. Alternatively, we can work with the predicted template matching score (pTM) [16, 80].

All target functions in this work include a term maximising confidence in the form of:

$$\mathcal{L}_{conf}(X) = \frac{\mu_{pAE} + \mu_{pLDDT}}{2}$$

where $X = (x, pAE, pLDDT)$ is a tuple containing the all-atom structure x, as well as the pAE and pLDDT for each protein.

b. Geometry As AF returns protein all-atom coordinates [16], we may introduce target terms which depend on arbitrary geometrical features of a protein. The main building block for such terms are measures of difference between two sets of coordinates. We have implemented the following protein structure metrics to use within cost functions:

- Aligned error [16]:

$$AE_{ij}(x, y) = |T_{x,i}x_i - T_{y,j}y_i|$$

where $T_{x,i}$ is the coordinate transformation moving a vector to the local coordinate system at backbone atom x_i. This is the aligned error used in AF training [16].

- Frame-Aligned Point Error [16]:

$$FAPE(x, y) = \frac{1}{cN^2} \sum_{ij}^{N} clip(AE_{ij}, c)$$

where c is a cutoff for the maximum error. This is the FAPE loss used in AF training.

- distance RMSD:

$$dRMSD(x, y) = RMSD(d(x), d(y))$$

- Template Matching score [80]:

$$TM_{appx}(x, y) = \max_j \frac{1}{N} \sum_{i}^{N} f(AE_{ij}(x, y));$$

$$f(d) = \frac{1}{1 + (\frac{d}{d_0(N)})};\quad d_0(N) = 1.24 \sqrt[3]{\max(N, 19) - 15} - 1.8$$

This is an approximation of the template matching score which does not require structural alignment [16]. This is our preferred measure for structural similarity, as it is normalised to 1 and provides a smoother target for highly dissimilar structures.

Furthermore, we can enforce general shape constraints for proteins using a compactness target $\mathcal{L}_{comp}$, which penalises large protein radii:

$$\mathcal{L}_{comp}(X) = -\frac{1}{N} \sum_i |x_i - \mu_x| - \max_i |x_i - \mu_x|$$

where μ_x denotes the center of mass of the coordinates x.

c. Monomers To generate globular monomeric proteins, we combine the confidence target with a weighted compactness target as follows:

$$\mathcal{L}_{mon}(X) = 2\mathcal{L}_{conf}(X) - \frac{1}{15 \cdot N}\mathcal{L}_{comp}(X)$$

This trades off compactness for confidence in the case of elongated structures.

d. Complexes For protein complex design, we use the same target function $\mathcal{L}_{mon}$ for each constituent monomer. This ensures each monomer has a stable high-confidence structure. Previous work identified inter-monomer pAE as a predictor of complex formation [19]. Low inter-monomer pAE corresponds to a high confidence in complex prediction. Therefore, we also impose the same confidence and compactness target on the complex, resulting in:

$$\mathcal{L}_{cpx}(X, X^m) = \frac{1}{M+1}\left(2\mathcal{L}_{conf}(X) + \mathcal{L}_{comp}(X) + \sum_{i}^{M}\mathcal{L}_{mon}(X^m)\right)$$

where X denote the AF predictions for the complex and X^m denote the same for each monomer m.

e. Conformational change To steer towards conformational change upon complex formation, we introduce an additional term to our oligomer cost function:

$$\mathcal{L}_{cc}(X, X^m) = \frac{1}{M}\sum_{m}^{M}(1 - TM(monomer(x, m), x^m))$$

where $monomer(x, m)$ extracts the coordinates of monomer m from the structure of the oligomer x. Essentially, maximising this target function minimises the TM score [80] between a protein structure as part of a complex and its structure as a monomer.

C. Input representation

We represent amino acid sequences as one-hot encoded arrays with 20 classes, one for each standard amino acid. This allows us to use both gradient-free and gradient-based optimisers, as we can estimate the gradient through a one-hot representation using a straight-through estimator or similar [81]. As input to AF, we construct an additional multiple-sequence alignment with only the single input sequence, as well as blank template features.

a. Sequence templates To reduce search space size and implement exact sequence constraints without introducing additional cost functions, we represent optimised sequences as a pair (S, T) where $S : \mathbb{R}^V$ is a one-hot representation of all V variable amino-acids and $T : \mathbb{R}^V \to \mathbb{R}^N$ is a template mapping which assembles the V variable amino acids into the final evaluated protein sequence of length N. For example, to design a homodimer with monomer size 64, the actual number of variable amino acids is $V = 64$ and the template T simply concatenates the input sequence $S : \mathbb{R}^{64}$ with itself.

b. Complex representation Following [19], we represent chain breaks for complex prediction by introducing an increment greater than 32 to the residue indices of each additional chain beyond the first. We use this increment to separate all residues on different chains by more than 32, which is the maximum relative residue index difference separately embedded by AF [16]. In addition, following [19] we split the multiple-sequence alignment features, such that each monomer has a separate copy of its sequence alignment features, with gaps at the sequence positions of other monomers.

III. MATERIALS AND METHODS

A. Prediction studies

Complex and conformational change predictions were performed using Colabfold [19] using the advanced notebook. In all cases, all five AF parameter sets were used, and models were selected by highest predicted template matching score [80]. Complex queries were predicted without paired MSA. The resulting top model was structurally aligned using PyMol [82] for RMSD reporting. All prediction runs are summarised in Table S1.

B. Design studies

a. Optimisation For optimisation an evolution strategy optimiser following [79] was used. Population size was set as 10. During mutation population size was expanded by a factor of 2. Sequences with a suboptimality of at most 10% were considered for mutation and recombination. Recombination was applied by crossover with probability of 10% at each sequence position. Optimisation was considered complete for sequences with $\mathcal{L} > 1.5$ for $\mathcal{L}_{mon}$, $\mathcal{L}_{cpx}$ and $\mathcal{L} > 0.9$ for target functions including $\mathcal{L}_{cc}$. Tab. I and Tab. II contain further parameters of optimisation runs performed.

b. AlphaFold configuration For optimisation, AF was configured for single-sequence use by disabling ensembling, templates, extra MSA features and restricting the number of MSA features to the number of monomers modelled. The number of AF iterations (recycling steps) was kept as a parameter for each optimisation run (Tab. I, Tab. II). For larger protein complexes, the number of iterations was decreased to 2 to speed up computation. The parameter set `model_1_ptm` was used for all experiments.

name	sequence length	N monomers	homooligomer	N recycling	target function
monomers 32	32	1	–	4	$\mathcal{L}_{mon}$
monomers 64	64	1	–	4	$\mathcal{L}_{mon}$
monomers 128	128	1	–	4	$\mathcal{L}_{mon}$
monomers 128	128	1	–	4	$\mathcal{L}_{mon}$
homodimers 32	32	2	yes	4	$\mathcal{L}_{cpx}$
homodimers 64	64	2	yes	4	$\mathcal{L}_{cpx}$
homodimers 128	128	2	yes	2	$\mathcal{L}_{cpx}$
heterodimers 32	32	2	no	4	$\mathcal{L}_{cpx}$
heterodimers 64	64	2	no	4	$\mathcal{L}_{cpx}$
heterodimers 128	128	2	no	2	$\mathcal{L}_{cpx}$
trimers 32 nr4	32	3	yes	4	$\mathcal{L}_{cpx}$
trimers 64 nr4	64	3	yes	4	$\mathcal{L}_{cpx}$
trimers 32 nr2	32	3	yes	2	$\mathcal{L}_{cpx}$
trimers 64 nr2	64	3	yes	2	$\mathcal{L}_{cpx}$
tetramers 32 nr4	32	4	yes	4	$\mathcal{L}_{cpx}$
tetramers 64 nr4	64	4	yes	4	$\mathcal{L}_{cpx}$
tetramers 32 nr2	32	4	yes	2	$\mathcal{L}_{cpx}$
tetramers 64 nr2	64	4	yes	2	$\mathcal{L}_{cpx}$
pentamers 32 nr4	32	5	yes	4	$\mathcal{L}_{cpx}$
pentamers 64 nr4	64	5	yes	4	$\mathcal{L}_{cpx}$
pentamers 32 nr2	32	5	yes	2	$\mathcal{L}_{cpx}$
pentamers 64 nr2	64	5	yes	2	$\mathcal{L}_{cpx}$
conformation change dimers 32	32	2	no	4	$0.7 \cdot \mathcal{L}_{cpx} + 0.3 \cdot \mathcal{L}_{cc}$
conformation change dimers 64	64	2	no	4	$0.7 \cdot \mathcal{L}_{cpx} + 0.3 \cdot \mathcal{L}_{cc}$
conformation change trimers 32	32	3	yes	4	$0.7 \cdot \mathcal{L}_{cpx} + 0.3 \cdot \mathcal{L}_{cc}$

TABLE I. *De novo* protein design runs and parameters.

name	binder length	target	N recycling	target function
binders 64-64-0	64	monomers 64 result 0	4	$\mathcal{L}_{cpx}$
binders 64-64-2	128	monomers 64 result 2	4	$\mathcal{L}_{cpx}$

TABLE II. Binder design design runs and parameters.

C. Rosetta validation

For a subset of examples in all design studies, we validated designed proteins using the Rosetta suite of protein design and structure prediction tools [45]. We used fragment-assembly-based *ab initio* structure prediction [83] as an independent baseline for designed protein structures.

a. Ab initio structure prediction. Protein secondary structures were predicted using PSIPRED [84] and S4PRED [85] to provide secondary structures for fragment selection. 3-mer and 9-mer fragments were selected from the Rosetta fragment database based on secondary structure and sequence information. Alignment information was not used as designed sequences did not have sufficient homology to natural sequences. *Ab initio* structure prediction was performed by fragment-assembly using the AbinitioRelax protocol. Two starting conformations were evaluated: starting from an extended conformation and starting from the AF predicted structure. For both cases, 32000 decoys were generated. Decoys from both runs were pooled and scored using the Rosetta all-atom energy function [86]. The overall 10 lowest energy decoys were selected for relaxation. Decoys were relaxed using the Relax protocol. The AF predicted structure was also relaxed to generate 10 additional decoys in case *ab initio* failed to find a minimum energy structure. All decoys were rescored to compute directly comparable energies. The lowest energy relaxed decoy was selected as the final predicted structure and its Rosetta score and RMSD with respect to the AF predicted structure were evaluated.

b. Protein-protein interface analysis. As Rosetta does not allow for *ab initio* structure prediction of protein complexes other than in the case of symmetric homo-oligomers [87], interfaces of designed complexes were assessed using the InterfaceAnalyzer protocol [88, 89]. Structures were relaxed and Rosetta binding energy was computed by removing a single monomer from the complex, followed by repacking. As a further measure of interface quality, packing statistics were computed using the PackStat protocol [90], which detects solvent-inaccessible unfilled regions within an interface.

D. Molecular dynamics simulation-based validation

A subset of designed monomers and higher-order complexes were further rigorously validated by performing all-atom molecular dynamics (MD) simulations in explicit solvent and analysing the corresponding properties of structural flexibility and internal/intramonomeric and interfacial contacts (for multimeric proteins and protein complexes).

a. Initial system construction The standard AMBER force field (ff14sb) was used to describe all protein parameters [91]. Each protein or protein complex was solvated using atomistic TIP3P water [92] with a minimum of 10 Å of padding to form a cubic periodic box and then electrically neutralized with an ionic concentration of 0.15 M NaCl that used standard ionic parameters [93].

b. Simulation protocol A standardised minimisation, equilibration and simulation protocol consisting of 11 stages was developed for all systems. A set of restraints (RS) were applied to each system at specified stages of equilibration. These consisted of restraining all heavy (non-hydrogen) atoms of the proteins. Each system was subsequently minimised across four stages with 1500 steps (500 steepest-descent + 1000 conjugate-gradient) of minimization applying restraints RS with different force constants in each sequential stage: Stage 1: 10 kcal/molA^2, Stage 2: 5 kcal/molA^2, Stage 3: 1 kcal/molA^2, Stage 4: unrestrained. MD simulations were performed in all subsequent stages. The SHAKE algorithm was employed on all atoms covalently bonded to a hydrogen atom. A time-step of 2 fs was used. The long-range Coulomb interaction was handled using a GPU implementation of the particle mesh Ewald summation method (PME) [94, 95]. A nonbonded cutoff distance of 10 Å was used. In Stage 5, each system was heated from 10 K to 300 K in 1 ns and with RS ($k = 10$ kcal/molA^2). The temperature was subsequently maintained at 300 K using a Langevin thermostat with a damping constant of $\gamma = 5.0\,\mathrm{ps}^{-1}$ and in Stage 6 the systems equilibrated for 1 ns at constant volume, thus in the NVT ensemble. Subsequently the pressure was maintained at 1 atm using a Berendsen barostat with a pressure relaxation time of $\tau_p = 1.0$ ps and the systems simulated in the NPT ensemble for 100 ps for each of the subsequent stages with RS: Stage 7: $k = 10$ kcal/molA^2 , Stage 8: $k = 5$ kcal/molA^2, Stage 9: $k = 1$ kcal/molA^2, Stage 10: $k = 0.5$ kcal/molA^2. Finally in Stage 11, RS restraints were removed and the systems simulated in the NPT ensemble for a further 5 ns. Following this, a production simulation of 100 ns each was performed for each system in the NPT ensemble with the same conditions as Stage 11. Coordinate snapshots from production simulations were generated every 10 ps, resulting in a trajectory of 10,000 snapshots per system.

c. Structural flexibility and contact analysis Structural stability and flexibility were analysed by computing the root-mean squared deviation (RMSD) of the backbone atoms of the protein with respect to the initial predicted structure (after aligning the protein backbone) as well as the root-mean-squared fluctuation (RMSF) with respect to the average structure in the production MD. In the case of complexes, this was carried out both on the overall complex and for individual monomers separately. The number of sidechain-sidechain intramonomer contacts were computed for each snapshot of the production MD, based on a heavy atom distance threshold of 4 Å. Similarly, interfacial contacts were computed for each interface in simulated protein complexes based on the same criteria. An overall picture of protein contacts was provided by summing the total intramonomer and interfacial contacts. Finally, for all simulated systems, an approximate potential of mean force (G) was determined by computing the Boltzmann-weighted distribution ($G = -k_B T \ln(\rho)$) in the 2D collective variable space of the global RMSF and total contacts (total intramonomer + total interfacial) contacts from the last 50 ns of simulation, where ρ is the normalised frequency in the binned 2D landscape.

E. Structural clustering

Designed proteins of length 64 and larger were subdivided into a dictionary of fragments using Geometricus [96]. Fragments were collected using the k-mer method with $k = 16$ and the radius method with cutoff 10 Å. A bag of features representation was computed for all designed proteins and dimensionality was reduced using non-negative matrix factorisation (NMF) with 50 components [97]. Proteins were separated into 10 clusters using ward-linkage agglomerative clustering [98] on their NMF components.

IV. RESULTS

A. AlphaFold predicts the structure of protein complexes

As recently shown by [18, 19], AF as trained on single-chain protein structures can be used for protein complex structure prediction. To gain confidence in the suitability of AF to design protein complexes using AlphaDesign, we first predicted the structure of small protein dimers, trimers, tetramers and pentamers (Fig. 2, SI Tab. S1). For the complexes considered, AF predictions show low root-mean-squared deviation (RMSD) to the native structure < 3 Å. While this may be due to these structures being part of the PDB [3] and thus part of AF's training dataset [16], together with recent work on protein complex prediction with AF [18, 19], it provides some indication that AF could be suitable for the design of protein complexes using our framework.

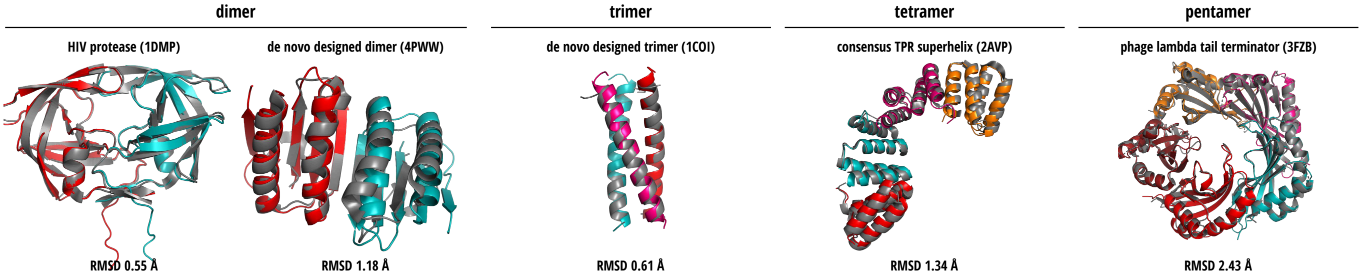
dimer
trimer
tetramer
pentamer
HIV protease (1DMP)
de novo designed dimer (4PWW)
de novo designed trimer (1COI)
consensus TPR superhelix (2AVP)
phage lambda tail terminator (3FZB)
RMSD 0.55 Å
RMSD 1.18 Å
RMSD 0.61 Å
RMSD 1.34 Å
RMSD 2.43 Å

FIG. 2. **AlphaFold prediction of protein complexes.** AlphaFold predictions of dimers (HIV protease 1DMP [99], de novo designed protein 4PWW [100]), trimers (de novo designed coiled coil 1COI [101]), tetramers (consensus TPR superhelix 2AVP), pentamers (phage lambda tail terminator 3FZB [102]). Predictions (colored) are aligned and overlaid onto the native structure (grey). All predictions show low RMSD to the the native structure.

B. Single-sequence optimisation designs diverse monomers

As an initial step towards AF-based *de novo* protein design, we optimised protein sequences of length 32, 64, 128 and 256 amino acids to design monomeric, globular proteins (Fig. 3). All designed monomers show a high predicted aligned confidence (Fig. 3, predicted error), all exceeding a mean predicted aligned confidence of 0.82 and mean pLDDT of 0.83 (normalised to the range $[0, 1]$). This indicates that AF assigns high confidence to the predicted structure, as pLDDT of 0.7 or higher corresponds to a confident prediction of backbone structure [16, 17]. To validate the AF-predicted structure, we performed fragment-based *ab initio* structure prediction using Rosetta [45] for a subset of designs, starting from an extended amino acid chain, as well as from the AF structure (SI Fig. S1). For each monomer, we then extracted the 10 lowest-energy Rosetta structures. We relaxed these structures as well as the AF structure using the Rosetta force field [86] and chose the lowest-energy structure. For most monomers, the RMSD between the AF structure (grey, Fig. 3 (A-D), structures) and the lowest-energy structure (red) is less than 3.0 Å. For structures with distances exceeding this threshold (Fig. 3, C: result 8) the larger deviation seems to be due to a flexible α-helix and the AF predicted structure has a comparably low Rosetta energy.

To visualise the energy landscape of folds for each designed monomer, we inspected the distribution of decoys from *ab initio* folding with their Rosetta energy [86] and RMSD to the AF structure (Fig. 3 (A-D), Rosetta score distribution). We note that for monomers of size 32 and 64 amino acids, relaxations of the predicted structure (blue) have comparable, yet higher energies compared to the best 10 decoys from *ab initio* prediction (red) (Fig. 3 (A, B), Rosetta score distribution, relaxed). Strikingly, for monomers of size 128 and 256, relaxed AF structures (blue) exhibit far lower energies compared to *ab initio* decoys (red) (Fig. 3 (C, D)). This indicates a failure of *ab initio* structure prediction to find a reasonable minimum energy structure for those proteins. Indeed, the distribution of decoys for monomers of size 32 and 64 shows funnel shaped distributions with a single minimum at both low energy and low RMSD characteristic for protein folding (Fig. 3 (A, B), Rosetta score distribution, abinitio). In contrast, for sizes 128 and 256, the distribution is much more diffuse with no clear minimum being visible, indicating failure of *ab initio* prediction to find a reasonable minimum (Fig. 3 (A, B), Rosetta score distribution, abinitio). However, the observation that relaxed AF structures exhibit very low Rosetta energy [86], increases confidence that designed structures are plausible.

As a further rigorous validation step, we assess the structural stability of designed monomers by performing all-atom molecular dynamics (MD) simulations. The Boltzmann-weighted frequency distribution in the 2D collective variable (CV) landscape consisting of backbone RMSF and intramonomer contacts shows that most systems exhibit a well-defined unimodal distribution centred on low RMSF and a significant number of contacts. This suggests the vast majority of conformers within the production ensemble sample a narrow range consistent with a maintenance of structural stability (Fig. 3 (A-D), MD metrics). Furthermore, the time evolution of the backbone RMSD with respect to the initial AF-predicted structure and backbone RMSF with respect to the MD-averaged structure are stably below 4 Å and 2 Å respectively in the majority of monomer systems (SI Fig. S1). Similarly, the time evolution of intramonomer contacts remains stable for all simulated monomers, suggesting a maintenance of folded structure. As expected the number of contacts grows with monomer size ranging from ∼50-100, ∼100-200, ∼300-400 and ∼600-800 for monomer lengths 32, 64, 128 and 256 amino-acids respectively (SI Fig. S1).

To assess the diversity of structures generated by AlphaDesign, we cluster all monomers of size 64 amino-acids or larger and extract representative structures for each cluster (Fig. 3 (E, F)). We embed structures using Geometricus [96], perform agglomerative clustering and visualise their distribution (Fig. 3 (E)). By visual inspection, cluster representatives encompass α-only (clusters 1, 2, 5, 6, 7), β-only (cluster 3) and mixed $\alpha\beta$ proteins (clusters 4, 8, 9, 10) (Fig. 3 (F)).

C. Searching for sequence pairs enables *de novo* dimer design

We next applied AlphaDesign to *de novo* design of protein homodimers and heterodimers. We optimised dimers with monomer size 32, 64 and 128 amino acids. As in our previous monomer designs, predicted dimers show high predicted aligned confidence for the AF structure (Fig. 4 (A-D), predicted error) with mean pAC > 0.75 and mean pLDDT > 0.83. By visual inspection, designed interfaces show a high level of shape complementarity. In lieu of *ab initio* structure prediction, which is in general harder to access for protein complexes using Rosetta [87], we opted for computing the Rosetta binding energy between monomers, as well as a score of interface packing statistics [90] as measures for structure quality. We applied this validation to a subset of designed dimers (SI Fig. S2). Dimers were relaxed using the Rosetta forcefield [86] noting that all relaxed structures have low RMSD with respect to the AF-predicted structure (Fig. 4, (A-F), complex structures). Furthermore, most dimers exhibit a packing statistic score greater than 0.6, which is comparable with the packing statistic of crystal structures with resolution 2.0 Å [90, 103]. Most structures show a Rosetta binding energy better than -40 REU hinting that predicted interfaces are stable under Rosetta [86].

MD simulations of dimer and higher order oligomer systems allow analysis of both intramonomer contacts as well as interfacial contacts between monomers. Again, the Boltzmann-weighted frequency distributions in the CV space of global RMSF and total contacts (intramonomer+interfacial) mostly show unimodal distributions with low mean RMSFs (≤ 2 Å) and significant numbers of total contacts that increase with complex size (Fig. 4, (A-F), MD metrics). Many dimer systems show stable time evolution of global RMSDs and RMSFs (SI Fig. S2), although significant deviation is observed in a few. Dissection of RMSDs, RMSFs, intramonomer and interfacial contacts, shows that both monomers in either the homodimer or heterodimer systems exhibit similar flexibility as well as number of intramonomer contacts. There are significant numbers of interfacial contacts in each dimer system, although, as expected these are fewer than the corresponding intramonomer contacts.

For synthetic biology applications, orthogonality is a desirable property for dimers [38]. That is, pairs of monomers designed to dimerise should only bind their designed partner, not monomers of other designed dimers. As a preliminary test for orthogonality of dimers designed using AlphaDesign, we predicted all combinations of monomers for a pair of designed dimers of monomer size 32 amino-acids (Fig. 4 (G)). On-target complexes were predicted by AF with high confidence (Fig. 4, (G), on-target predictions). In contrast, for off-target complexes – that is, complexes of monomers not designed for dimer formation – the mean predicted aligned confidence for inter-monomer amino acid pairs dropped below 0.5. This indicates that AF cannot confidently predict these off-target combinations as a dimer, providing preliminary evidence that dimers designed using AlphaDesign can indeed exhibit orthogonality.

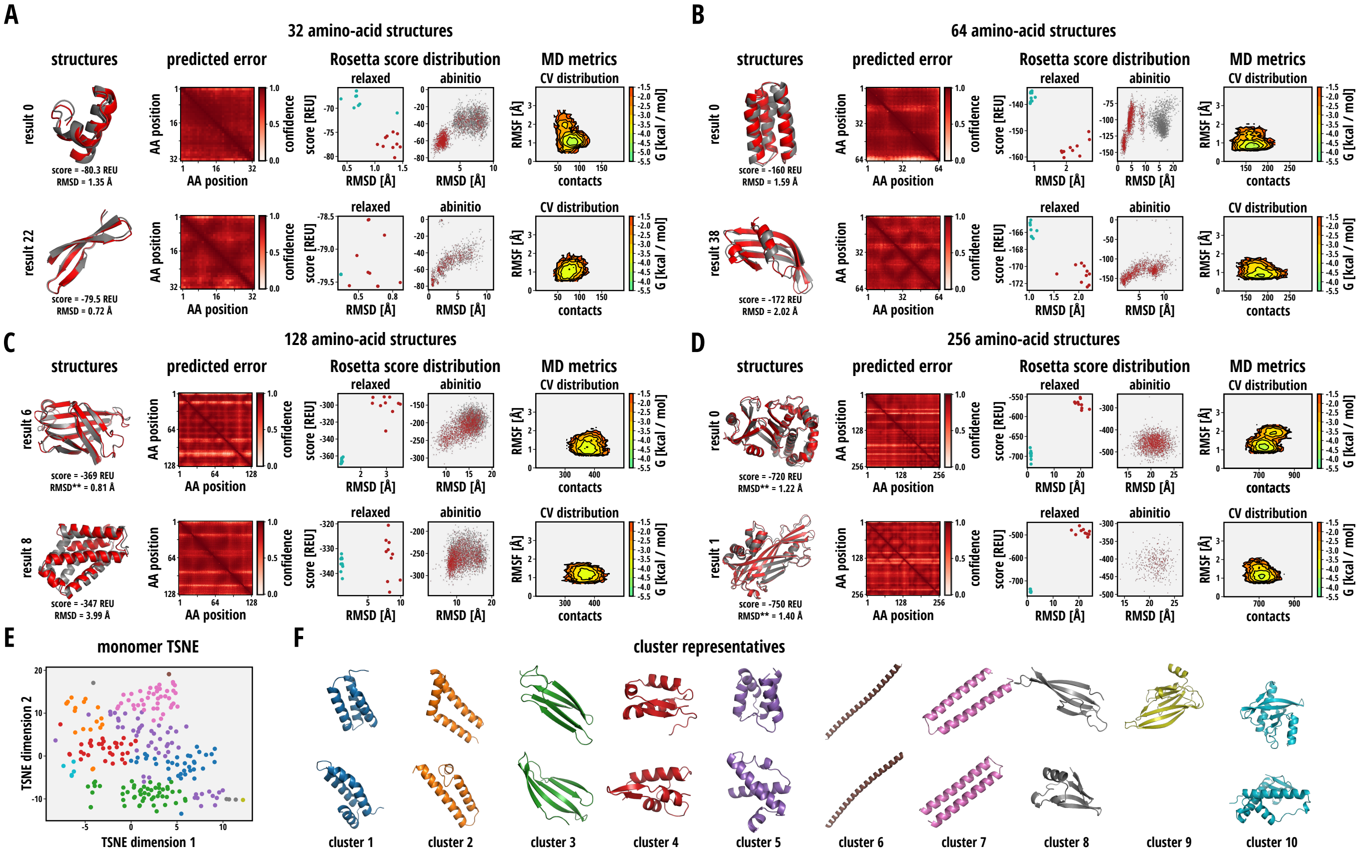

FIG. 3. **De novo monomer design. (A - D)** *ab initio* and molecular dynamics validation of designed monomers of length 32 - 256 amino acids. Designed monomers (grey) are overlaid with their lowest-energy structure (red) from Rosetta *ab initio* prediction or relaxation of their predicted all-atom structure (structures). Each monomer is reported with its RMSD with respect to the AlphaFold structure and Rosetta energy. Predicted aligned confidence (pAC) is shown for each designed monomer (predicted error). For validation using Rosetta, the distribution of Rosetta scores and RMSDs to the AlphaFold structure is shown (Rosetta score distribution): for the top 10 relaxed structures (relaxed) starting from *ab initio* prediction (red) and the AlphaFold structure (blue); for all decoys from Rosetta *ab initio* prediction (*ab initio*) starting from an extended conformation (grey) and the AlphaFold structure (red). For molecular dynamics validation (MD metrics), the Boltzmann-weighted collective variable (CV) distribution in the 2D landscape of the number of intramonomer amino acid contacts and backbone RMSF of the protein with respect to the averaged MD structure is shown (CV distribution). **(E)** TSNE of designed monomer structures of length 64 amino acids and larger. Structures are separated into 10 clusters using agglomerative clustering. **(F)** Representatives for each cluster, showing diverse structures designed using AlphaFold.

D. Multiple-sequence optimisation allows for homo-oligomer design

To demonstrate the feasibility of protein complex design beyond dimers, we proceeded to design homo-oligomers from trimers to hexamers. Monomer sizes of 32 and 64 amino acids were considered for trimers, tetramers, pentamers and 32 amino acids for hexamers. As before, we evaluated AF predicted aligned confidence, Rosetta binding energy, packing statistics and behaviour under 100 ns of molecular dynamics simulations for a subset of oligomers (SI Fig. S3). Designed structures show high predicted aligned confidence with most structures at pAC and pLDDT > 0.7 (Fig. 5, (A-G) predicted error), low RMSD to the Rosetta relaxed structure < 2.0 Å, good binding energy < -40 REU and packing statistics > 0.59 comparable to natural protein complexes [90, 103] (Fig. 5 (A-G) complex structures).

MD simulations of homo-oligomers exhibit similar properties to the dimers, with time evolution showing generally modest RMSDs (≤ 6 Å) and RMSFs (≤ 2 Å) and significant numbers of intramonomer and interfacial contacts (SI Fig. S3). Boltzmann-weighted distributions in the RMSF-contact space again show unimodal distributions with mean RMSFs remaining around 2 Å and total contacts scaling with complex size. Dissection of individual intramonomer and interfacial contacts shows notable contacts in each monomer and interface for almost all systems (Fig. 5 (A-G) MD metrics).

Interestingly, inspection of designed complexes reveals a subset of extended oligomers with exposed interfaces on both

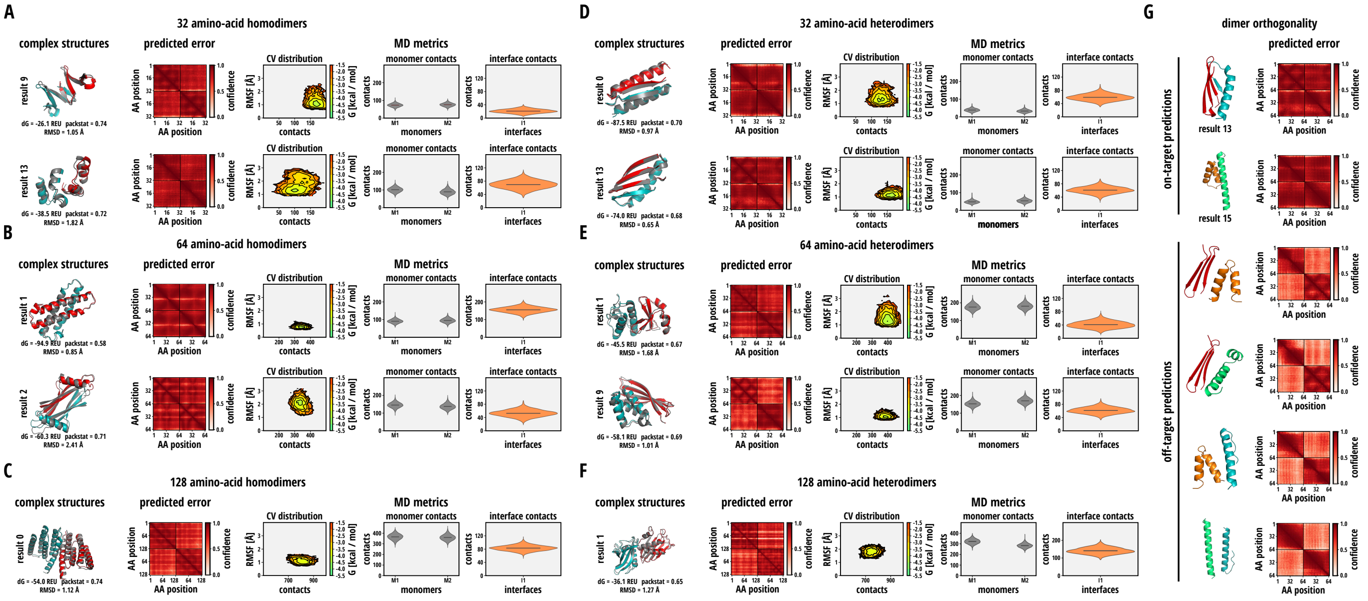

FIG. 4. **De novo dimer design. (A - C)** Rosetta and molecular dynamics validation of designed homodimers of length 32 - 128 amino acids. Designed homodimers (grey) are overlaid with their lowest-energy structure (coloured) from Rosetta relaxation of their predicted all-atom structure (complex structures). Each relaxed dimer is reported with its RMSD to the AlphaFold structure, Rosetta binding energy and packing statistics. Predicted aligned confidence (pAC) is shown for each designed dimer (predicted error). For molecular dynamics validation (MD complex metrics), the Boltzmann-weighted CV distribution in the 2D space of total amino acid contacts (intramonomer+interfacial) and RMSF of the complex is shown (CV distribution), together with the distribution of individual monomer contacts and interface contacts. **(D-F)** Rosetta and molecular dynamics validation of designed heterodimers of length 32 - 128 amino acids. Measures displayed are the same as in (A-C). **(G)** Orthogonality of designed dimers. For two distinct pairs of designed dimers, predicted structures for the on-target and off-target complexes are shown. Predicted confidence (predicted error) of the on-target dimers is consistently higher for inter-monomer pairs of amino acids than the same confidence for off-target dimers.

sides (Fig. 5 (G-H)). The designed sequences could potentially oligomerise into larger assemblies. Correspondingly, MD simulations of these complexes show maintenance of intramonomer contacts for all monomers and sequential interfacial contacts for all but one interface, as expected. This indicates the possibility of designing and validating large-scale assemblies in AF beyond simple oligomers.

E. Designing with a fixed monomer finds binders for target proteins

As a natural extension to dimer and oligomer design using AF with potential applications in biologics design and synthetic biology, we designed proteins binding to a fixed target protein (Fig. 6, SI Fig. S4). Here, we chose two proteins of length 64 amino-acids previously designed as target proteins and which exhibit distinctly different folds (Fig. 6 (A, B) target protein). We fixed the sequence for the target proteins during the design process while optimising the sequence of the designed binding proteins. Predicted structures of designed binding proteins exhibit high aligned confidence > 0.81, good Rosetta binding energy $< -58\,\mathrm{REU}$ and packing statistics > 0.67 comparable to natural protein interfaces [90, 103] (Fig. 6 (A, B)). This indicates that designed binders form dimers with the target proteins. By visual inspection we note that binders designed for each target protein seem to all bind at the same interface indicating preferences for binding sites during the design process.

MD simulations of the two designed binders for each of the two target proteins show unimodal distributions in the 2D CV space of global RMSF-total contacts, with mean RMSFs $< 2\,\text{Å}$ and mean number of total contacts ranging from 300-400, consisting of significant numbers of individual intramonomer (100-200) and interfacial (40-120) contacts.

F. Sequence design uncovers signatures of conformational change in AlphaFold

To ascertain whether AF sequence space contains information about protein conformational change when interacting with other proteins, we predict the structures of two proteins known to have different monomeric and oligomeric

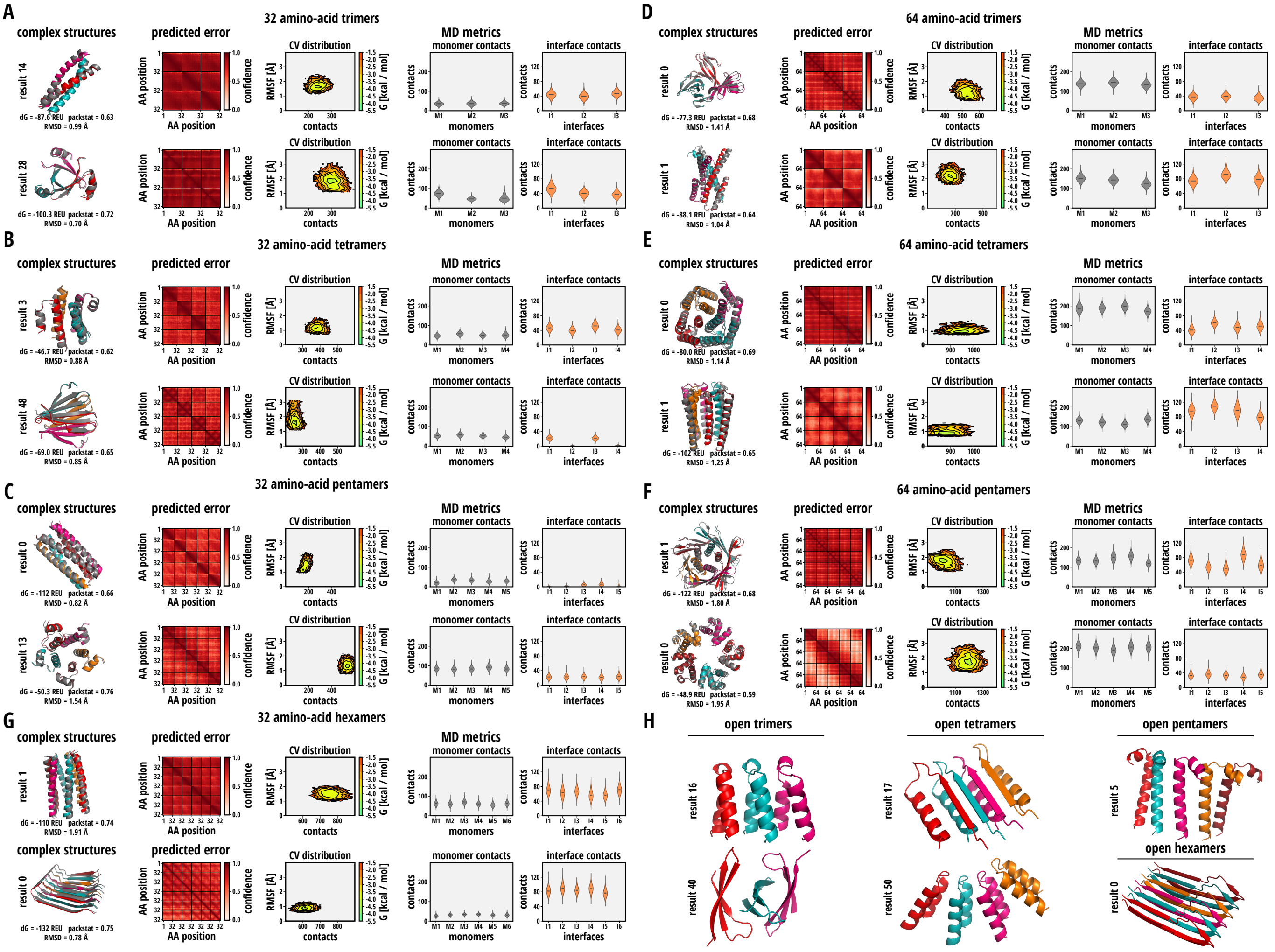

FIG. 5. **De novo oligomer design. (A - G)** Rosetta and molecular dynamics validation of designed trimers, tetramers and pentamers of monomer length 32 - 64 amino acids, as well as hexamers of monomer length 32 amino acids. Designed oligomers (grey) are overlaid with their lowest-energy structure (coloured) from Rosetta relaxation of their predicted all-atom structure (complex structures). Each relaxed oligomer is reported with its RMSD with respect to the AlphaFold structure, Rosetta binding energy and packing statistics. Predicted aligned confidence (pAC) is shown for each designed oligomer (predicted error). For molecular dynamics validation, the Boltzmann-weighted CV distribution in the 2D space of total amino acid contacts (intramonomer+interfacial) and RMSF of the complex is shown (CV distribution), together with the distribution of individual monomer contacts and interface contacts. **(H)** Additional structures of extended oligomers. These oligomers can be extended by additional monomers on each side.

states – KaiB, a circadian clock protein [104] and amyloid-β involved in Alzheimer's disease [105]. As a monomer, KaiB adopts its ground-state structure, KaiBgs [104]. In complex with KaiC, KaiB changes conformation to the fold-switch stabilised state KaiBfs [104]. The AF prediction of KaiB as a monomer shows good agreement (RMSD 0.69 Å) with the native KaiBgs structure, while the prediction of KaiB in complex with KaiC shows good agreement with the native structure of KaiBfs (RMSD 1.92 Å). However, structural alignment of the predicted complex of KaiB and KaiC shows high RMSD (6.11 Å) to the structure of KaiBgs (Fig. 7 (A)), indicating that the AF prediction has captured a part of the conformational change between KaiBgs and KaiBfs. In its monomeric state, native amyloid-β forms an α-helical structure and changes conformation to a parallel β-sheet upon aggregation (Fig. 7 (A) grey) [105]. While not as accurate in terms of RMSD (> 5 Å), AF captures the transition between the α-helical monomeric state and the parallel β-sheet oligomer well (Fig. (A) red). This indicates that indeed AF may have learned to predict conformational change upon protein binding in a limited way.

Furthermore, we identify a set of oligomer designs exhibiting conformational change upon complex formation (Fig. 7 (B), SI Fig. S5). Designed structures show a conformational change from an α-helix in the monomeric state (Fig. 7

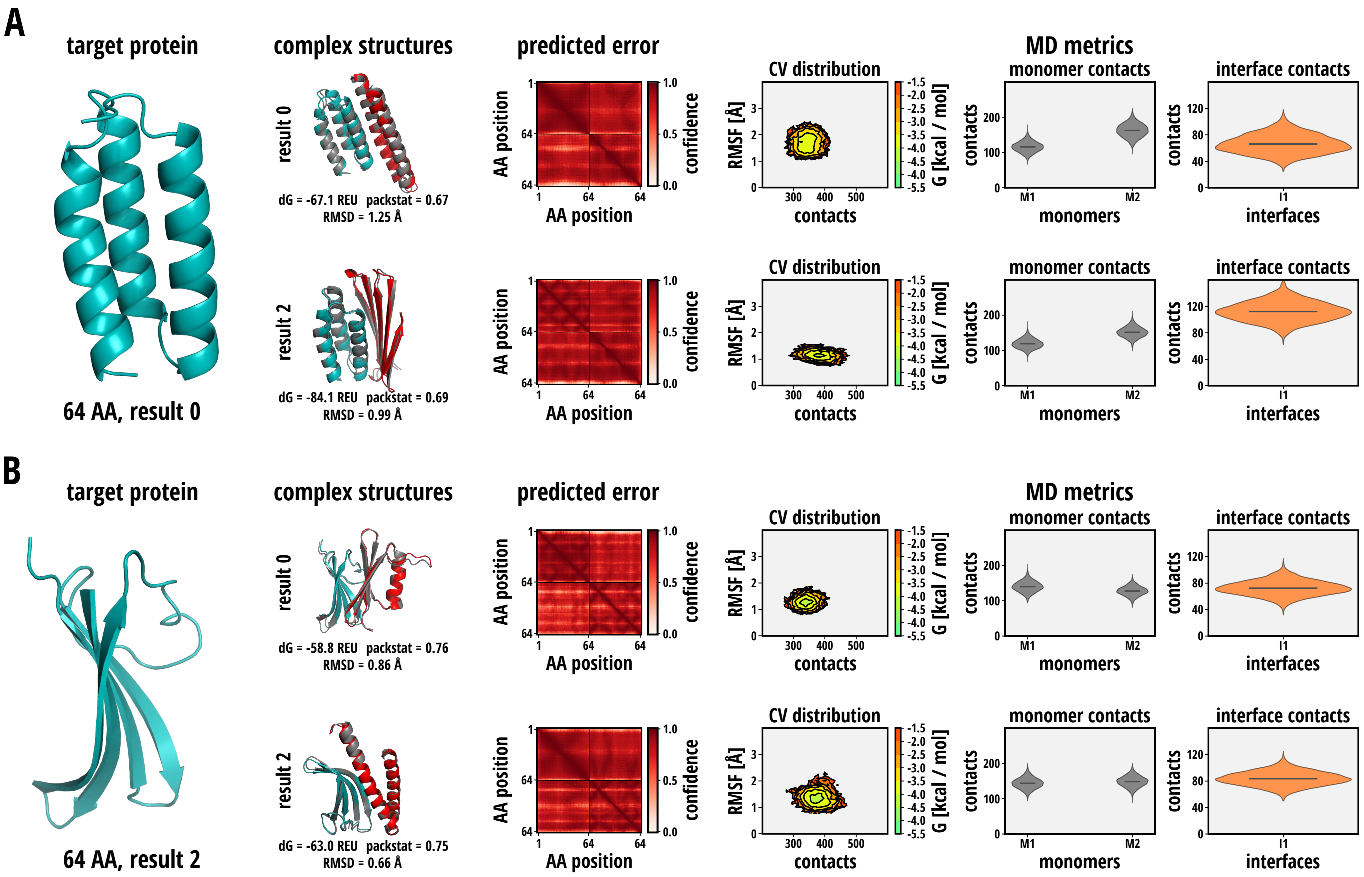

FIG. 6. **De novo binder design. (A, B)** Rosetta validation of designed binders for previously *de novo* designed proteins. Relaxed binders (red) bound to the target protein (blue) are overlaid with their predicted all-atom structure (grey) (complex structures). Each relaxed binder is reported with its RMSD with respect to the AlphaFold structure, Rosetta binding energy and packing statistics. Predicted aligned confidence (pAC) is shown for each designed binder (predicted error). For molecular dynamics validation, the Boltzmann-weighted CV distribution in the 2D space of total amino acid contacts (intramonomer+interfacial) and RMSF of the complex is shown (CV distribution), together with the distribution of individual monomer contacts and interface contacts.

(B), monomer structures) to a stack of parallel β-sheets characteristic for amyloids [105]. Both the monomeric and oligomeric state exhibit high predicted aligned confidence under AF (Fig. 7 (B) predicted error) and the oligomeric state shows binding energies $< -88\,\text{REU}$ and packing statistics > 0.65 for all oligomers. This indicates that the complexes are stable under the Rosetta forcefield [86].

Similarly, MD simulations show well-defined minima in the RMSF-contact space centering on low RMSFs ($< 2\,\text{Å}$) and a significant number of contacts (100-500). However, these conformation-switching open-ended oligomers do vary significantly compared to other oligomers including conformation-retaining open-ended oligomers, described previously. Whilst open-endedness is captured by a significant number of contacts in all-but-one sequential interfaces, the number of interfacial contacts (40-80) is either similar to or larger than the number of intramonomer contacts (10-60). This confirms the elongated conformational structure of monomers in the oligomeric state.

We next attempt to design heterodimers and homo-oligomers exhibiting conformational change upon complex formation. By maximising TM score between the monomeric and oligomeric state during optimisation, we find proteins exhibiting the desired conformation change (Fig. 7 (C)). Interestingly, we find the resulting proteins show conformational changes beyond the amyloid-like transition found in our previous design experiments, indicating a larger variety of conformation changing proteins present in AF sequence space.

V. DISCUSSION AND CONCLUSIONS

Here, we have developed a *de novo* protein design framework based on sequence optimisation using evolutionary algorithms. Extending on previous works that utilise structure predictors[70, 71], we have embedded AlphaFold (AF)

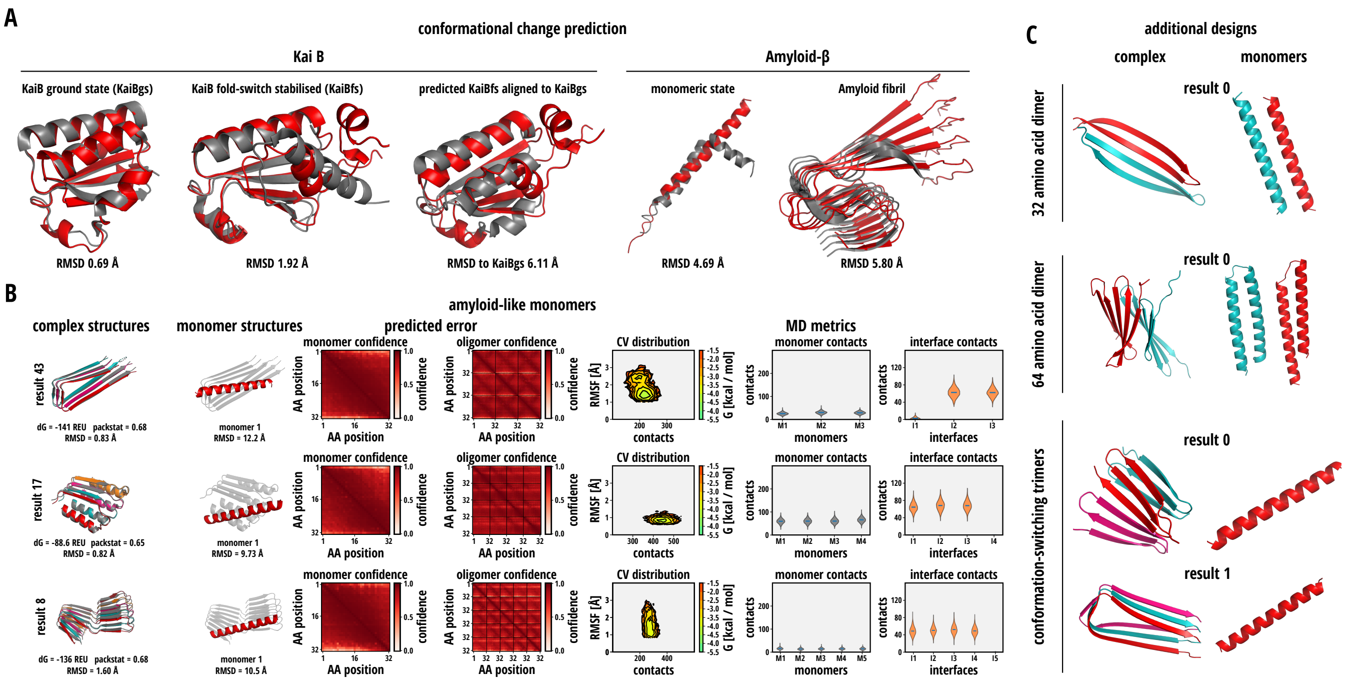

FIG. 7. **Conformational change in AlphaFold sequence space. (A)** AlphaFold prediction of conformational change upon complex formation. (left) Predicted structures for monomeric circadian clock protein KaiB (KaiB ground state, KaiBgs) and fold-switch stabilised KaiB (KaiBfs) predicted in complex with KaiC (red) overlaid with the native structure of KaiBgs (1R5P [106]) and KaiBfs (5JYT [107]). Both predictions show low RMSD with respect to the corresponding native structure. Predicted KaiBfs (red) overlaid with native KaiBgs (the incorrect conformation, grey) shows high RMSD. (right) Predicted structures for monomeric Amyloid-β and its pentamer (red) overlaid with the native monomer (1IYT [108]) and oligomer (2MXU [109]). AlphaFold predicts the conformational change from α-helix to parallel β-sheet characteristic for amyloids. **(B)** Rosetta and molecular dynamics validation of designed oligomers of monomer length 32 showing amyloid-like predicted conformational change. Designed oligomers (grey) are overlaid with their lowest-energy structure (coloured) from Rosetta relaxation of their predicted all-atom structure (complex structures). Each relaxed oligomer is reported with its RMSD to the AlphaFold structure, Rosetta binding energy and packing statistics. Monomers are shown in comparison to the oligomer structure (monomer structures) to illustrate conformational change upon oligomerisation. Predicted aligned confidence (pAC) is shown for each designed oligomer and its constituent monomers (predicted error). For molecular dynamics validation, the Boltzmann-weighted CV distribution in the 2D space of total amino acid contacts (intramonomer+interfacial) and RMSF of the complex is shown (CV distribution), together with the distribution of individual monomer contacts and interface contacts. **(C)** Additional proteins designed using a target function favouring conformational change upon complex formation.

[16] into the design loop as a prediction oracle. Importantly, optimisation is facilitated by AF's output predicted confidence measures, namely the predicted local distance difference test (pLDDT) [78] and the predicted aligned error (pAE) [16]. Furthermore, we have developed both a set of flexible target functions that encode various design tasks as well as an extendable platform for developing further target functions for solving bespoke design problems. This has enabled a range of applications including *de novo* design of protein monomers, dimers, oligomers, context dependent conformational switchers and binders to target proteins.

Our predicted structures are extensively validated using the Rosetta suite of protein design and structure prediction tools [45]. We also use fragment-assembly-based *ab initio* structure prediction [83] as an independent baseline for designed protein structures. In addition to this we develop a further rigorous validation protocol using all-atom molecular dynamics (MD) simulations that extends beyond conventionally used computational techniques for structure prediction evaluation. MD simulations enable extensive exploration of a putative native state, thus instabilities are picked up as increases in global structural flexibility, loss of internal contacts within protein monomers and/or loss of interfacial contacts within complexes.

We applied our framework to design *de novo* monomer proteins starting from completely random sequences ranging in size from 32 - 256 amino acids in length, based on a target function that combined pLDDT and pAE. This results in a range of structurally stable *de novo* designed monomer proteins with diverse folds. Using AF's recently noticed functionality to predict complexes we incorporated and tested complex prediction on a number of systems showing good structural agreement. By specifying a further target function based on globular compactness together with complex prediction, we further designed a range of stable *de novo* protein complexes including homodimers, heterodimers and homo-oligomers from trimers to hexamers. Interestingly, we demonstrate orthogonality between pairs of dimers, suggesting the approach could have applicability in designing mutually exclusive combinations, for example, in protein logic gates [38]. Moreover, we have observed a number of open-ended predicted complexes, providing a potential route to the design of self-assembling systems [39].

A particularly intriguing subset of our oligomer systems exhibit striking conformational changes between their monomeric and oligomeric state. Structural prediction of existing protein systems known to change conformation and/or fold between monomer and oligomer form, including amyloid $\alpha - \beta$ switching, show that AF inherently contains signatures of conformational change. We consequently observe conformation switching between monomer and oligomer forms in some *de novo* designed open-ended oligomer systems. By developing a target function that maximises the structural difference between monomer and oligomer forms, our framework is able to *de novo* design oligomers with this conformational switching property. Context dependent conformational switching is a desirable feature in synthetic biology applications. For example, designing proteins to self-assemble inside but not outside the cell may be achievable by design of membrane permeable α-helices that spontaneously switch into membrane impermeable β-sheeted filaments as accumulated intracellular concentration drives an equilibrium towards the oligomeric state.

Finally, our approach enabled us to select a protein to be unmodified during the design loop - thus by combining this with a *de novo* designed protein starting from a random sequence we could design monomeric binders for a pre-specified target protein. Application of this to a set of target proteins resulted in the design of stable binders that exhibit significant interfacial contacts across the same interface for a given target protein. This suggests our framework could be further optimised towards therapeutic applications in potent biologic design.

VI. ACKNOWLEDGEMENTS

J.O.K. acknowledges support by the BMBF (de.NBI project: 031A537B) and the Volkswagen Foundation (contract 95826). S.K.S. acknowledges support by the Volkswagen Foundation "Experiment! Funding Initiative" (grant no. 93874-1).

REFERENCES

[1] Christian B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181 4096:223–30, 1973.

[2] Christine Zardecki, Chenghua Shao, Maria Voigt, and Stephen K. Burley. Protein data bank: 50 years of macromolecular structures enabling research and education. *The FASEB Journal*, 35, 2021.

[3] Helen M. Berman, John D. Westbrook, Zukang Feng, Gary L Gilliland, Talapady N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Acta crystallographica. Section D, Biological crystallography*, 58 Pt 6 No 1:899–907, 2000.

[4] Ian P. W. Sillitoe, Nicola Bordin, Natalie L. Dawson, Vaishali P. Waman, Paul Ashford, Harry M. Scholes, Camilla S.M. Pang, Laurel Woodridge, Clemens Rauer, Neeladri Sen, Mahnaz Abbasian, Sean Le Cornu, Su Datt Lam, Karel Berka, Ivana Hutarová Vareková, Radka Svobodová Vareková, Jonathan G. Lees, and Christine A. Orengo. Cath: increased structural coverage of functional space. *Nucleic Acids Research*, 49:D266 – D273, 2021.

[5] K. Fujiwara and M. Ikeguchi. Oligami: Oligomer architecture and molecular interface. *Journal of Proteomics & Bioinformatics*, pages 248–248, 2008.

[6] Mohammed AlQuraishi. End-to-end differentiable learning of protein structure. *Cell systems*, 8 4:292–301.e3, 2019.

[7] Sheng Wang, S. Sun, Z. Li, Renyu Zhang, and Jinbo Xu. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Computational Biology*, 13, 2017.

[8] Ratul Chowdhury, Nazim Bouatta, Surojit Biswas, Charlotte Rochereau, George M Church, Peter Karl Sorger, and Mohammed N AlQuraishi. Single-sequence protein structure prediction using language models from deep learning. *bioRxiv*, 2021.

[9] John Ingraham, Adam J. Riesselman, Chris Sander, and Debora S. Marks. Learning protein structure with a differentiable simulator. In *ICLR*, 2019.

[10] Jiaxiang Wu, Tao Shen, Haidong Lan, Yatao Bian, and Junzhou Huang. Se(3)-equivariant energy-based models for end-to-end protein folding. *bioRxiv*, 2021.

[11] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

[12] Carlos Simmerling, Bentley Strockbine, and Adrian E. Roitberg. All-atom structure prediction and folding simulations of a stable protein. *Journal of the American Chemical Society*, 124 38:11258–9, 2002.

[13] Alexander Schug, Abhinav Verma, Kyu H. Lee, and Wolfgang Wenzel. Stochastic optimization methods for protein folding. 2005.

[14] Brian Kuhlman and Philip Bradley. Advances in protein structure prediction and design. *Nature Reviews Molecular Cell Biology*, 20:681–697, 2019.

[15] Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. Critical assessment of methods of protein structure prediction (casp) - round xiv. *Proteins*, 2021.

[16] John M Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Zídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andy Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David A. Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583 – 589, 2021.

[17] Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, et al. Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873):590–596, 2021.

[18] Mehmet Akdel, Douglas Eduardo Valente Pires, Eduard Porta Pardo, Jürgen Jänes, Arthur O. Zalevsky, Bálint Mészáros, Patrick Bryant, Lydia L. Good, Roman A. Laskowski, Gabriele Pozzati, Aditi Shenoy, Wensi Zhu, Petras J. Kundrotas, Victoria Ruiz Serra, Carlos H M Rodrigues, Alistair S Dunham, David Burke, Neera Borkakoti, Sameer Velankar, Adam Frost, Kresten Lindorff-Larsen, Alfonso Valencia, Sergey Ovchinnikov, Janani Durairaj, David B. Ascher, Janet M Thornton, Norman E. Davey, Amelie Stein, Arne Elofsson, Tristan I. Croll, and Pedro Beltrão. A structural biology community assessment of alphafold 2 applications. *bioRxiv*, 2021.

[19] Milot Mirdita, Sergey Ovchinnikov, and Martin Steinegger. Colabfold-making protein folding accessible to all. *bioRxiv*, 2021.

[20] Ian R. Humphreys, Jimin Pei, Minkyung Baek, Aditya Krishnakumar, Ivan Anishchenko, Sergey Ovchinnikov, Jing Zhang, Travis J. Ness, Sudeep Banjade, Saket Bagde, Viktoriya G. Stancheva, Xiao-Han Li, Kaixian Liu, Zhi Zheng, Daniel J. Barrero, Upasana Roy, Israel S. Fernández, Barnabas Szakal, Dana Branzei, Eric C. Greene, Sue Biggins, Scott Keeney, Elizabeth A. Miller, J. Christopher Fromme, Tamara L. Hendrickson, Qian Cong, and David Baker. Structures of core eukaryotic protein complexes. *bioRxiv*, 2021.

[21] Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Zídek, Russ Bates, Sam Blackwell, Jason Yim, Olaf Ronneberger, Sebastian Bodenstein, Michal Zielinski, Alex Bridgland, Anna Potapenko, Andrew Cowie, Kathryn Tunyasuvunakool, Rishub Jain, Ellen Clancy, Pushmeet Kohli, John M Jumper, and Demis Hassabis. Protein complex prediction with alphafold-multimer. *bioRxiv*, 2021.

[22] David D Boehr, Ruth Nussinov, and Peter E Wright. The role of dynamic conformational ensembles in biomolecular recognition. *Nature chemical biology*, 5(11):789–796, 2009.

[23] Peter E Wright and H Jane Dyson. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of molecular biology*, 293(2):321–331, 1999.

[24] Bálint Mészáros, István Simon, and Zsuzsanna Dosztányi. The expanding view of protein–protein interactions: complexes involving intrinsically disordered proteins. *Physical biology*, 8(3):035003, 2011.

[25] A Keith Dunker, Israel Silman, Vladimir N Uversky, and Joel L Sussman. Function and structure of inherently disordered proteins. *Current opinion in structural biology*, 18(6):756–764, 2008.

[26] Subramanian Vivekanandan, Jeffrey R Brender, Shirley Y Lee, and Ayyalusamy Ramamoorthy. A partially folded structure of amyloid-beta (1–40) in an aqueous environment. *Biochemical and biophysical research communications*, 411(2):312–316, 2011.

[27] Kulkarni Madhurima, Bodhisatwa Nandi, and Ashok Sekhar. Metamorphic proteins: the janus proteins of structural biology. *Open biology*, 11(4):210012, 2021.

[28] José Nelson Onuchic and Peter G Wolynes. Theory of protein folding. *Current opinion in structural biology*, 14(1):70–75, 2004.

[29] S Kashif Sadiq, Frank Noé, and Gianni De Fabritiis. Kinetic characterization of the critical step in hiv-1 protease maturation. *Proceedings of the National Academy of Sciences*, 109(50):20449–20454, 2012.

[30] S Kashif Sadiq, Abraham Muñiz Chicharro, Patrick Friedrich, and Rebecca C Wade. A multiscale approach for computing gated ligand binding from molecular dynamics and brownian dynamics simulations. *bioRxiv*, 2021.

[31] Kresten Lindorff-Larsen, Stefano Piana, Ron O Dror, and David E Shaw. How fast-folding proteins fold. *Science*, 334(6055):517–520, 2011.

[32] Tom Venken, Arnout Voet, Marc De Maeyer, Gianni De Fabritiis, and S Kashif Sadiq. Rapid conformational fluctuations of disordered hiv-1 fusion peptide in solution. *Journal of chemical theory and computation*, 9(7):2870–2874, 2013.

[33] Guillermo Pérez-Hernández, Fabian Paul, Toni Giorgino, Gianni De Fabritiis, and Frank Noé. Identification of slow molecular order parameters for markov model construction. *The Journal of chemical physics*, 139(1):07B604_1, 2013.

[34] Nuria Plattner, Stefan Doerr, Gianni De Fabritiis, and Frank Noé. Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and markov modelling. *Nature chemistry*, 9(10):1005–1011, 2017.

[35] Albert C Pan, Daniel Jacobson, Konstantin Yatsenko, Duluxan Sritharan, Thomas M Weinreich, and David E Shaw. Atomic-level characterization of protein–protein association. *Proceedings of the National Academy of Sciences*, 116(10):4244–4249, 2019.

[36] Neil J Bruce, Gaurav K Ganotra, Daria B Kokh, S Kashif Sadiq, and Rebecca C Wade. New approaches for computing ligand–receptor binding kinetics. *Current opinion in structural biology*, 49:1–10, 2018.

[37] Po-Ssu Huang, Scott E Boyken, and David Baker. The coming of age of de novo protein design. *Nature*, 537(7620):320–327, 2016.

[38] Zibo Chen, Ryan D Kibler, Andrew Hunt, Florian Busch, Jocelynn Pearl, Mengxuan Jia, Zachary L VanAernum, Basile IM Wicky, Galen Dods, Hanna Liao, et al. De novo design of protein logic gates. *Science*, 368(6486):78–84, 2020.

[39] Zibo Chen, Matthew C Johnson, Jiajun Chen, Matthew J Bick, Scott E Boyken, Baihan Lin, James J De Yoreo, Justin M Kollman, David Baker, and Frank DiMaio. Self-assembling 2d arrays with de novo protein building blocks. *Journal of the American Chemical Society*, 141(22):8891–8895, 2019.

[40] Aaron Chevalier, Daniel-Adriano Silva, Gabriel J Rocklin, Derrick R Hicks, Renan Vergara, Patience Murapa, Steffen M Bernard, Lu Zhang, Kwok-Ho Lam, Guorui Yao, et al. Massively parallel de novo protein design for targeted therapeutics. *Nature*, 550(7674):74–79, 2017.

[41] Michael J Dougherty and Frances H Arnold. Directed evolution: new parts and optimized function. *Current opinion in biotechnology*, 20(4):486–491, 2009.

[42] Moshe Goldsmith and Dan S Tawfik. Directed enzyme evolution: beyond the low-hanging fruit. *Current opinion in structural biology*, 22(4):406–412, 2012.

[43] Lynne Regan, Diego Caballero, Michael R Hinrichsen, Alejandro Virrueta, Danielle M Williams, and Corey S O'hern. Protein design: Past, present, and future. *Peptide Science*, 104(4):334–350, 2015.

[44] Xingjie Pan and Tanja Kortemme. Recent advances in de novo protein design: Principles, methods, and applications. *Journal of Biological Chemistry*, page 100558, 2021.

[45] Andrew Leaver-Fay, Michael D. Tyka, Steven M. Lewis, Oliver F. Lange, James M Thompson, Ron Jacak, Kristian Kaufman, Paul D. Renfrew, Colin A. Smith, William Sheffler, Ian W. Davis, Seth Cooper, Adrien Treuille, Daniel J. Mandell, Florian Richter, Yih-En Andrew Ban, Sarel Jacob Fleishman, Jacob E. Corn, David E. Kim, Sergey Lyskov, Monica Berrondo, Stuart G. Mentzer, Zoran Popovic, James J Havranek, John Karanicolas, Rhiju Das, Jens Meiler, Tanja Kortemme, Jeffrey J. Gray, Brian Kuhlman, David Baker, and Philip Bradley. Rosetta3: an object-oriented software suite for the simulation and design of macromolecules. *Methods in enzymology*, 487:545–74, 2011.

[46] Julia Koehler Leman, Brian D Weitzner, Steven M Lewis, Jared Adolf-Bryfogle, Nawsad Alam, Rebecca F Alford, Melanie Aprahamian, David Baker, Kyle A Barlow, Patrick Barth, et al. Macromolecular modeling and design in rosetta: recent methods and frameworks. *Nature methods*, 17(7):665–680, 2020.

[47] Po-Ssu Huang, Yih-En Andrew Ban, Florian Richter, Ingemar André, Robert M. Vernon, William R. Schief, and David Baker. Rosettaremodel: A generalized framework for flexible backbone protein design. *PLoS ONE*, 6, 2011.

[48] Brian Kuhlman, Gautam Dantas, Gregory C. Ireton, Gabriele Varani, Barry L. Stoddard, and David Baker. Design of a novel globular protein fold with atomic-level accuracy. *Science*, 302:1364 – 1368, 2003.

[49] Fabian Sesterhenn, Che Yang, Jaume Bonet, Johannes T. Cramer, Xiaolin Wen, Yimeng Wang, Chi-I Chiang, Luciano A. Abriata, Iga Kucharska, Giacomo Castoro, Sabrina S Vollers, Marie Galloux, Elie Dheilly, Stéphane Rosset, Patricia Corthésy, Sandrine Georgeon, Mélanie Villard, Charles-Adrien Richard, Delphyne Descamps, Teresa Delgado, Elisa Oricchio, Marie-Anne Rameix-Welti, Vicente Más, Sean Ervin, Jean-François Éléouët, Sabine Riffault, John T. Bates, Jean-Philippe Julien, Yuxing Li, Theodore S. Jardetzky, Thomas Krey, and Bruno E. Correia. De novo protein design enables precise induction of functional antibodies in vivo. *bioRxiv*, page 685867, 2020.

[50] Bruno E. Correia, John T. Bates, Rebecca J. Loomis, Gretchen Baneyx, Chris Carrico, Joseph G. Jardine, Peter Rupert, Colin E. Correnti, Oleksandr Kalyuzhniy, Vinayak Vittal, Mary J. Connell, Eric Stevens, Alexandria Schroeter, Man Chen, Skye MacPherson, Andreia M. Serra, Yumiko Adachi, Margaret A. Holmes, Yuxing Li, Rachel E. Klevit, Barney S. Graham, Richard T. Wyatt, David Baker, Roland K. Strong, James E. Crowe, Philip R. Johnson, and William R. Schief. Proof of principle for epitope-focused vaccine design. *Nature*, 507:201 – 206, 2014.

[51] Zibo Chen, Scott E. Boyken, Mengxuan Jia, Florian Busch, David Flores-Solis, Matthew J. Bick, Peilong Lu, Zachary L. VanAernum, Aniruddha Sahasrabuddhe, Robert A. Langan, Sherry Bermeo, T. J. Brunette, Vikram Khipple Mulligan, Lauren P. Carter, Frank Dimaio, Nikolaos G. Sgourakis, Vicki H. Wysocki, and David Baker. Programmable design of orthogonal protein heterodimers. *Nature*, 565:106–111, 2018.

[52] Yang Hsia, Jacob B. Bale, Shane Gonen, Dan Shi, William Sheffler, Kimberly K. Fong, Una Nattermann, Chunfu Xu, Po-Ssu Huang, Rashmi Ravichandran, Sue Yi, Trisha N. Davis, Tamir Gonen, Neil P. King, and David Baker. Design of a hyperstable 60-subunit protein icosahedron. *Nature*, 535:136 – 139, 2016.

[53] Florian Richter, Andrew Leaver-Fay, Sagar D. Khare, Sinisa Bjelic, and David Baker. De novo enzyme design using rosetta3. *PLoS ONE*, 6, 2011.

[54] Matthew D. Smith, Alexandre Zanghellini, and Daniela Grabs-Röthlisberger. Computational design of novel enzymes without cofactors. *Methods in molecular biology*, 1216:197–210, 2014.

[55] Ali Madani, Bryan McCann, Nikhil Naik, Nitish Shirish Keskar, Namrata Anand, Raphael R. Eguchi, Po-Ssu Huang, and Richard Socher. Progen: Language modeling for protein generation. *bioRxiv*, 2020.

[56] Alex Hawkins-Hooker, Florence Depardieu, Sebastien Baur, Guillaume Couairon, Arthur Chen, and David Bikard. Generating functional protein variants with variational autoencoders. *PLoS Computational Biology*, 17, 2021.

[57] Donatas Repecka, Vykintas Jauniskis, Laurynas Karpus, Elzbieta Rembeza, Jan Zrimec, Simona Povilonienė, Irmantas Rokaitis, Audrius Laurynėnas, Wissam Abuajwa, Otto Savolainen, Rolandas Meskys, Martin K. M. Engqvist, and Aleksej Zelezniak. Expanding functional protein sequence space using generative adversarial networks. *bioRxiv*, 2019.

[58] Rolf Apweiler, Amos Bairoch, Cathy H. Wu, Winona C. Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria Jesus Martin, Darren A. Natale, Claire O'Donovan, Nicole Redaschi, and Lai-Su Yeh. Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 32 Database issue:D115–9, 2004.

[59] Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Deep neural language modeling enables functional protein generation

across families. *bioRxiv*, 2021.

[60] Namrata Anand and Possu Huang. Generative modeling for protein structures. In *NeurIPS*, 2018.

[61] Sari Sabban and Mikhail G. Markovsky. Ramanet: Computational de novo helical protein backbone design using a long short-term memory generative neural network. *bioRxiv*, 2019.

[62] Hao Huang, Boulbaba Ben Amor, Xichan Lin, Fan Zhu, and Yi Fang. G-vae, a geometric convolutional vae for proteinstructure generation. *ArXiv*, abs/2106.11920, 2021.

[63] Namrata Anand, Raphael R. Eguchi, and Po-Ssu Huang. Fully differentiable full-atom protein backbone generation. In *DGS@ICLR*, 2019.

[64] John Ingraham, Vikas K. Garg, Regina Barzilay, and T. Jaakkola. Generative models for graph-based protein design. In *DGS@ICLR*, 2019.

[65] Namrata Anand-Achim, Raphael R. Eguchi, Irimpan I Mathews, Carla Patricia Perez, Alexander Derry, Russ B. Altman, and Po-Ssu Huang. Protein sequence design with a learned potential. *bioRxiv*, 2020.

[66] Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael J. L. Townshend, and Ron O. Dror. Learning from protein structure with geometric vector perceptrons. *ArXiv*, abs/2009.01411, 2021.

[67] Jingxue Wang, Huali Cao, John Zeng Hui Zhang, and Yifei Qi. Computational protein design with deep learning neural networks. *Scientific Reports*, 8, 2018.

[68] Alexey Strokach, David Becerra, Carles Corbi-Verge, Albert Perez-Riba, and Philip M. Kim. Fast and flexible protein design using deep graph neural networks. *Cell systems*, 2020.

[69] Surojit Biswas, Grigory Khimulya, Ethan C. Alley, Kevin M. Esvelt, and George M. Church. Low-n protein engineering with data-efficient deep learning. *bioRxiv*, 2020.

[70] Ivan Anishchenko, Tamuka Martin Chidyausiku, Sergey Ovchinnikov, Samuel J Pellock, and David Baker. De novo protein design by deep network hallucination. *bioRxiv*, 2020.

[71] Christoffer H Norn, Basile I. M. Wicky, David Juergens, Sirui Liu, David E. Kim, Doug K Tischer, Brian Koepnick, Ivan V. Anishchenko, David Baker, and Sergey Ovchinnikov. Protein sequence design by conformational landscape optimization. *Proceedings of the National Academy of Sciences of the United States of America*, 118, 2021.

[72] Doug K Tischer, Sidney Lisanza, Jue Wang, Runze Dong, Ivan V. Anishchenko, Lukas F. Milles, Sergey Ovchinnikov, and David Baker. Design of proteins presenting discontinuous functional sites using deep learning. *bioRxiv*, 2020.

[73] Johannes Linder and G. Seelig. Fast differentiable dna and protein sequence optimization for molecular design. *ArXiv*, abs/2005.11275, 2020.

[74] Lewis Moffat, Joe G Greener, and David T Jones. Using alphafold for rapid and accurate fixed backbone protein design. *bioRxiv*, 2021.

[75] Ethan C. Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M. Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, pages 1–8, 2019.

[76] Surojit Biswas, Gleb Kuznetsov, Pierce J Ogden, Nicholas Conway, Ryan P. Adams, and George M. Church. Toward machine-guided design of proteins. *bioRxiv*, 2018.

[77] Jianyi Yang, Ivan V. Anishchenko, Hahnbeom Park, Zhenling Peng, Sergey Ovchinnikov, and David Baker. Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences*, 117:1496 – 1503, 2020.

[78] Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. lddt: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29:2722 – 2728, 2013.

[79] S. Sinai, Richard Wang, Alexander Whatley, Stewart Slocum, Elina Locane, and Eric D. Kelsic. Adalead: A simple and robust adaptive greedy search algorithm for sequence design. *ArXiv*, abs/2010.02141, 2020.

[80] Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure*, 57, 2004.

[81] Eric Jang, Shixiang Shane Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *ArXiv*, abs/1611.01144, 2017.

[82] Warren L. Delano. The pymol molecular graphics system. 2002.

[83] Kim T. Simons, Richard Bonneau, Ingo Ruczinski, and David Baker. Ab initio protein structure prediction of casp iii targets using rosetta. *Proteins: Structure*, 37, 1999.

[84] D. T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology*, 292 2:195–202, 1999.

[85] Lewis Moffat and David T. Jones. A deep semi-supervised framework for accurate modelling of orphan sequences. *bioRxiv*, 2020.

[86] Rebecca F. Alford, Andrew Leaver-Fay, Jeliazko R. Jeliazkov, Matthew J. O'Meara, Frank Dimaio, Hahnbeom Park, Maxim V. Shapovalov, Paul D. Renfrew, Vikram Khipple Mulligan, Kalli Kappel, Jason W. Labonte, Michael S. Pacella, Richard Bonneau, Philip Bradley, Roland L. Dunbrack, Rhiju Das, David Baker, Brian Kuhlman, Tanja Kortemme, and Jeffrey J. Gray. The rosetta all-atom energy function for macromolecular modeling and design. *Journal of chemical theory and computation*, 13 6:3031–3048, 2017.

[87] Rhiju Das, Ingemar André, Yang Shen, Yibing Wu, A. S. Lemak, Sonal Bansal, Cheryl H. Arrowsmith, Thomas Szyperski, and David Baker. Simultaneous prediction of protein folding and docking at high resolution. *Proceedings of the National Academy of Sciences*, 106:18978 – 18983, 2009.

[88] Steven M. Lewis and Brian Kuhlman. Anchored design of protein-protein interfaces. *PLoS ONE*, 6, 2011.

[89] P. Benjamin Stranges and Brian Kuhlman. A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds. *Protein Science*, 22, 2013.

[90] William Sheffler and David Baker. Rosettaholes: Rapid assessment of protein core packing for structure prediction, refinement, design, and validation. *Protein Science*, 18, 2009.

[91] James Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin Hauser, and Carlos Simmerling. ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of chemical theory and computation*, 11 8:3696–713, 2015.

[92] William L. Jorgensen, Jayaraman Chandrasekhar, Jeffry D. Madura, Roger Impey, and Michael L. Klein. Comparison of simple potential functions for simulating liquid water. *Journal of Chemical Physics*, 79:926–935, 1983.

[93] InSuk Joung and Thomas E. Cheatham. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *The Journal of Physical Chemistry. B*, 112:9020 – 9041, 2008.

[94] Ulrich Essmann, Lalith E. Perera, Max L. Berkowitz, Thomas A. Darden, Hsing-Chou Lee, and Lee G. Pedersen. A smooth particle mesh ewald method. *Journal of Chemical Physics*, 103:8577–8593, 1995.

[95] P. P. Ewald. Die berechnung optischer und elektrostatischer gitterpotentiale. *Annalen der Physik*, 369:253–287, 1921.

[96] Janani Durairaj, Mehmet Akdel, Dick de Ridder, and Aalt D.J. van Dijk. Geometricus represents protein structures as shape-mers derived from moment invariants. *Bioinformatics*, 36 Supplement_2:i718–i725, 2020.

[97] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[98] Joe H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58:236–244, 1963.

[99] C. Nicholas Hodge, Paul Edward Aldrich, Lee Bacheler, C. H. Francis Chang, Charles Eyermann, Sena Garber, Mary F. Grubb, David Anthony Jackson, Prabhakar Kondaji Jadhav, Bruce D. Korant, Patrick Y. S. Lam, Michael B. Maurin, James L. Meek, Michael J. Otto, M. M. Rayner, Caroline Reid, Thomas Ray Sharpe, Linyee Shum, Dean L. Winslow, and Susan K. Erickson-Viitanen. Improved cyclic urea inhibitors of the hiv-1 protease: synthesis, potency, resistance profile, human pharmacokinetics and x-ray crystal structure of dmp 450. *Chemistry & biology*, 3 4:301–14, 1996.

[100] Sergey Vorobiev, Y.-R. Lin, Jayaraman Seetharaman, Rong Xiao, John K. Everett, Thomas B. Acton, David Baker, Gaetano T. Montelione, Liang Tong, and John F. Hunt. Crystal structure of engineered protein. northeast structural genomics consortium target or494. 2014.

[101] N. L. Ogihara, Manfred S. Weiss, David S. Eisenberg, and William F. DeGrado. The crystal structure of the designed trimeric coiled coil coil-vald: Implications for engineering crystals and supramolecular assemblies. *Protein Science*, 6, 1997.

[102] Lisa G. Pell, Amanda Liu, Lizbeth Edmonds, Logan W Donaldson, P. Lynne Howell, and Alan R. Davidson. The x-ray crystal structure of the phage lambda tail terminator protein reveals the biologically relevant hexameric ring structure and demonstrates a conserved mechanism of tail termination among diverse long-tailed phages. *Journal of molecular biology*, 389 5:938–51, 2009.

[103] Nicholas F. Polizzi, Yibing Wu, Thomas Lemmin, Alison M. Maxwell, Shao-Qing Zhang, Jeff Rawson, David N. Beratan, Michael J. Therien, and William F. DeGrado. De novo design of a hyperstable non-natural protein-ligand complex with sub-å accuracy. *Nature chemistry*, 9 12:1157–1164, 2017.

[104] Roger Tseng, Nicolette F Goularte, Archana G. Chavan, Jansen Luu, Susan E Cohen, Yong-Gang Chang, Joel Heisler, Sheng Li, Alicia K. Michael, Sarvind Tripathi, Susan S. Golden, Andy LiWang, and Carrie L. Partch. Structural basis of the day-night transition in a bacterial circadian clock. *Science*, 355:1174 – 1180, 2017.

[105] Rodrigo Gallardo, Neil A. Ranson, and Sheena E. Radford. Amyloid structures: much more than just a cross-β fold. *Current opinion in structural biology*, 60:7–16, 2019.

[106] Robert G. Garces, Ning Wu, Wanda Gillon, and Emil F. Pai. Anabaena circadian clock proteins kaia and kaib reveal a potential common binding site to their partner kaic. *The EMBO Journal*, 23, 2004.

[107] Roger Tseng, Nicolette F Goularte, Archana G. Chavan, Jansen Luu, Susan E Cohen, Yong-Gang Chang, Joel Heisler, Sheng Li, Alicia K. Michael, Sarvind Tripathi, Susan S. Golden, Andy LiWang, and Carrie L. Partch. Structural basis of the day-night transition in a bacterial circadian clock. *Science*, 355:1174 – 1180, 2017.

[108] Orlando Crescenzi, Simona Tomaselli, Remo Guerrini, Severo Salvadori, Anna Maria D'Ursi, Piero Andrea Temussi, and Delia Picone. Solution structure of the alzheimer amyloid beta-peptide (1-42) in an apolar microenvironment. similarity with a virus fusion domain. *European journal of biochemistry*, 269 22:5642–8, 2002.

[109] Yiling Xiao, Buyong Ma, Dan McElheny, Sudhakar Parthasarathy, Fei Long, Minako Hoshi, Ruth Nussinov, and Yoshitaka Ishii. Aβ(1–42) fibril structure illuminates self-recognition and replication of amyloid in alzheimer's. *Nature structural & molecular biology*, 22:499 – 505, 2015.