Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Rapid Assessment of Chemical Complementarity of Ligands for Protein Design

Derek Woolfson

D.N.Woolfson@bristol.ac.uk

University of Bristol https://orcid.org/0000-0002-0394-3202

Rokas Petrenas

University of Bristol

Katarzyna Ożga

University of Bristol

Joel Chubb

University of Bristol

Andrey Romanyuk

University of Bristol

Jennifer McManus

University of Bristol https://orcid.org/0000-0003-3630-5108

Graham Leggett

University of Sheffield

Nigel Scrutton

University of Manchester https://orcid.org/0000-0002-4182-3500

Tom Oliver

University of Bristol

Article

Keywords:

Posted Date: December 10th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7592160/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Rapid Assessment of Chemical Complementarity of Ligands for Protein Design

Rokas Petrenas[1,†], Katarzyna Ożga[1,†], Joel J. Chubb[1,†], Andrey V. Romanyuk[1,2], Jennifer J. McManus[3], Graham J. Leggett[4], Nigel S. Scrutton[5,6], Thomas A. A. Oliver[1], and Derek N. Woolfson[1,2,7,*]

[1]School of Chemistry, University of Bristol, Bristol, UK.

[2]Max Planck-Bristol Centre for Minimal Biology, University of Bristol, Bristol, UK.

[3]HH Wills Physics Laboratory, University of Bristol, Bristol, UK.

[4]Department of Chemistry, University of Sheffield, Sheffield, UK.

[5]Manchester Institute of Biotechnology, University of Manchester, UK

[6]Department of Chemistry, University of Manchester, UK

[7]School of Biochemistry, University of Bristol, Medical Sciences Building, Bristol, UK.

[†]These authors contributed equally: Rokas Petrenas, Katarzyna Ożga, Joel J. Chubb.

*To whom correspondence should be addressed: d.n.woolfson@bristol.ac.uk

Abstract (152 words)

Computational protein design is delivering *de novo* structures rapidly and reliably. This opens new frontiers for functional design, for instance, for binding small molecules tightly and specifically. However, achieving predictable and tuneable binding is challenging. Here we introduce a rapid physics-based method to generate isosteric and chemically complementary binding pockets for small-molecule targets in *de novo* proteins. We test this by constructing and characterizing binding proteins for several synthetic and natural chromophores. By evaluating only single-digit numbers of designs, the pipeline delivers stable proteins with pre-organised binding sites confirmed by X-ray crystallography, which bind the targets with micromolar affinities or better. To illustrate the scope and applications of this approach, we incorporate selective and coupled chromophore-binding sites in a two-domain *de novo* protein enabling controlled energy transfer between the two sites, and we develop a small *de novo* binding protein that can be used in live mammalian cells to visualise sub-cellular structures.

Main

Driven by deep-learning methods, computational protein design is advancing rapidly, and it is now possible to generate many *de novo* protein structures quickly and robustly[1-4]. This sets new frontiers for the field. One such area is to design proteins that bind small molecules tightly and specifically. Although this has been a target for computational protein design for some time[5-12], understanding the non-covalent interactions that underpin such designs to make binding predictable and tuneable remains challenging. Solutions to these problems would enable functions like small-molecule sensing, sequestration, delivery, and catalysis[5,6,13].

To date, design studies in this area have highlighted the difficulty of achieving high-affinity and highly specific interactions. The routes taken have encompassed all approaches in *de novo* design, including minimal[14], rational[5,15,16], and computational strategies[7]. Many of these focus on specific ligands rather than small-molecule binding in general and the need to validate selectivity. For example, using computational design, Baker et al. have used Rosetta to computationally design binders for the steroid digoxigenin into a natural scaffold, initially achieving low-to-mid micromolar affinity, and then improving this to picomolar with library-based mutagenesis and selection[7]. Kortemme et al. have expanded the scope of computational binding-site design by fragmenting ligands and using the fragments in PDB searches to identify optimal contacts[8]. Polizzi and DeGrado combine fragment-based binding-site design with parametric scaffold generation to develop an apixaban binder with sub-micromolar affinity, and low-nanomolar binders for rucaparib[9-11]. Also, Höcker and Jürgens have engineered the *E. coli* tryptophan repressor to bind auxin as a component of a biosensor in plants[12]. Despite these and other successes, these largely physics- and knowledge-based methods are computationally slow, can have limited ligand scope, and usually require careful input and filtering from an expert user[3].

Currently, protein design is undergoing a step-change with the application of deep-learning methods[3,4]. While AI has accelerated and simplified the sampling of designable *de novo* binders, it is not yet clear how well these tools predict which designs are most likely to succeed experimentally or how they advance mechanistic understanding of protein function. Specifically, early applications of deep-learning methods to deliver small molecule-binding proteins, whilst achieving low-micromolar to nanomolar affinities for cholic acid and methotrexate for instance, still require experimental screening of large numbers of designs per target[13]. New pipelines based on scaffold generation using RFDiffusionAA[17] do show improvement for targets that are well represented in structural databases such as haem and bilin. Whilst impressive advances, these studies highlight the considerable challenges that remain in the computational design of small molecule-binding proteins, with even the most advanced methods requiring extensive optimization and screening to achieve experimental successes[3,18].

To address these issues, we introduce RASSCCoL, a computational pipeline for the Rapid ASSessment of Chemical Complementarity of Ligands in protein design. This is a robust and interpretable physics-based method for incorporating small molecule-binding sites into protein scaffolds. RASSCCoL's streamlined pipeline uses the volume and functional groups of the ligand to etch complementary pockets into the proteins. Two key aspects of this approach are: (i) the use of fast volume calculations to limit the number of side-chain combinations that need to be assessed, and (ii) the leveraging of a robust, highly stable, and large family of *de novo* designed α-helical scaffolds. Effectively, these allow near-exhaustive sampling of sequences that are most likely to generate pre-organised, isosteric binding sites for the target ligand while bypassing the sampling of backbone flexibility. The resulting sequences are repacked onto the

input backbone and are evaluated by ligand docking using AutoDock Vina[19], and then by regenerating the protein folds and binding pockets using AlphaFold2 predictions[20]. To address the challenge of exponential scaling for larger ligands and concomitant sequence spaces, we employ machine learning to score designs at the sequence level. This restricts the computationally expensive explicit modelling of structures to the most-promising sequences.

To validate this approach, we create binding sites in single-chain coiled coil-based scaffolds generated through rationally seeded computational protein design[21,22]. These scaffolds have advantages of small size, stability, and well-understood sequence-to-structure relationships. Specifically, they have layered hydrophobic cores that facilitate ligand alignment and side-chain replacement in the computational design. Coiled-coil proteins and protein-protein interfaces are ubiquitous in nature[23], and are amongst the most successful *de novo* designed proteins targeted to date[2,24]. In principle, however, RASSCCoL can be applied to any protein scaffold. Combined, this pipeline simplifies the design process, requires minimal user input, produces robust *in silico* models that are chemically derived and interpretable, limits the number of constructs for experimental testing, and yields high success rates of confirmed binders. Importantly, this successful rational and physical approach provides a baseline for increasingly complex design pipelines. Finally, we show that RASSCCoL can be used to introduce orthogonal but coupled binding sites to produce more-complex, functional *de novo* proteins; and to deliver small *de novo* chromophore-binding proteins for use as fluorescent reporters in cells.

Rapid physics-based design pipeline

Initially to develop and illustrate RASSCCoL, we chose a compact and stable *de novo* designed 4-helix bundle, sc-apCC-4 (PDB id 8A3K), as the protein scaffold[21]; and two synthetic, environment-sensitive dyes—1,6-diphenyl-1,3,5-hexatriene (DPH) and Nile red—as the target ligands[25]. Our computational pipeline for generating small molecule-binding sites had three generalisable steps (Fig. 1): (i) binding-site selection; (ii) sequence generation and culling; (iii) model building, ligand docking, and evaluation.

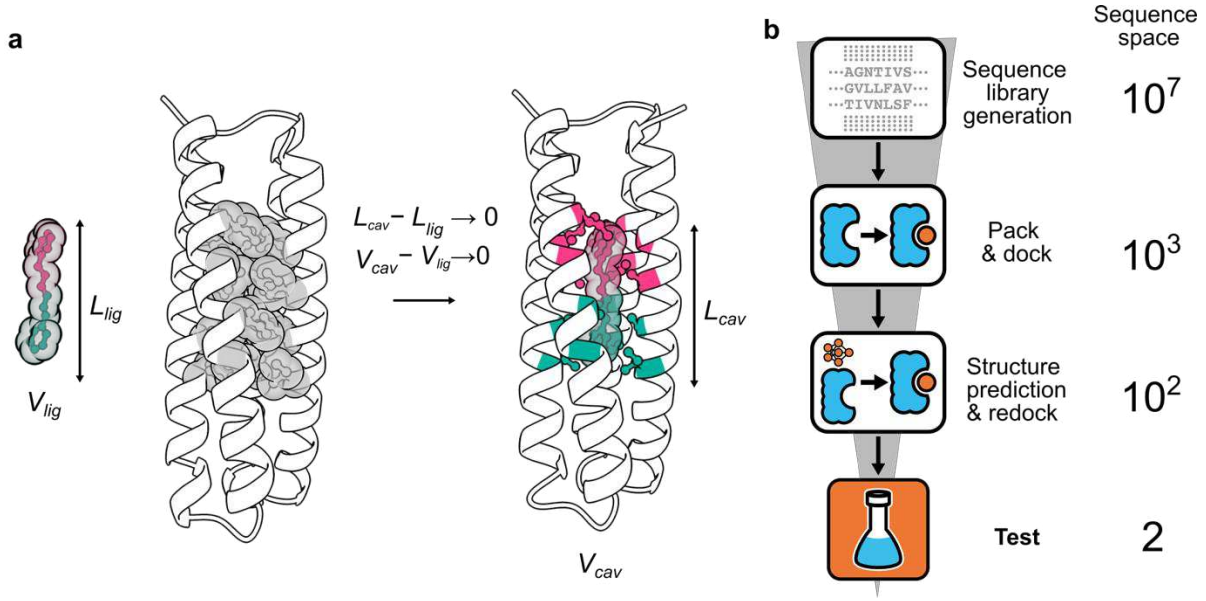


Fig. 1: The RASSCCoL computational protein design pipeline illustrated by DPH binding. a, A potential site for DPH binding (L_{cav}) is identified based on the maximum ligand length (L_{lig}). To match the ligand's shape, the protein scaffold is segmented into corresponding layers (magenta and green). A volume constraint is used to minimise the difference between the binding cavity (V_{cav}) and the ligand (V_{lig}) in each layer; effectively, replacing atoms of the side chains with those of the ligand. **b,** Design and evaluation strategy showing the numbers of sequences at each step of DPH-binder design. In this case, the sequence library allows only aliphatic amino acids in the binding pocket. Sequences passing the volume constraint are packed onto the scaffold for ligand docking. Structures with the most-negative predicted binding energies in AutoDock Vina are modelled using AlphaFold2 (AF2) and the ligands redocked. Two high-confidence designs with the best redocking scores are selected for experimental testing.

First, the positions of binding-site residues were identified based on ligand length. The core packing in sc-apCC-4 is layered with four side chains per layer and a spacing of ≈5 Å between layers. Thus, the number of layers that need to be redesigned to accommodate a ligand can be

estimated as maximum ligand length in Å divided by 5 (Fig. 1a). Second, a volume calculation based on Bondi radii[26] was used to identify side-chain permutations to generate voids that match the ligand's volume most closely. The 2-fold symmetric and entirely hydrophobic ligand DPH (length, L ≈ 15 Å; and volume, V ≈ 274 Å^3; Fig. 1a) spans three 4-residue layers of sc-apCC-4. Within these layers, we retained some of the Ile residues important for fold stability[21]. This left ten designable positions in total, which we restricted to combinations of the aliphatic residues, Gly, Ala, Val, Leu and Ile (Supplementary Table 1). Thus, potentially, ≈10 million sequence combinations (5^{10}) could have been searched. However, this reduced to ≈3,200 sequences by applying the ligand-volume and ligand-symmetry constraints (Fig. 1b, Supplementary Table 1). Third, using FASPR[27] each sequence was repacked into the scaffold structure, and DPH was docked into the resulting models using AutoDock Vina (Fig. 1b)[19]. AutoDock Vina has a physics-based scoring function to predict ligand-binding poses and affinities, which we used to score the models. Combined, these computational steps took tenths of a second for each sequence (Supplementary Fig. 1), enabling the entire design protocol to finish within 30 minutes on a single CPU core (Supplementary Table 1). Finally, AlphaFold2 (AF2) was used to predict structures for a small number of the top-scoring sequences, followed by docking to evaluate ligand-pose recovery (Fig. 1b, Supplementary Table 2). These provided assessment metrics to choose only a small number of designs for experimental production and characterization (Supplementary Fig. 3, Supplementary Table 3). To evaluate how RASSCCoL extended to asymmetric ligands, we targeted Nile red (L ≈ 13 Å; V ≈ 324 Å^3; Supplementary Fig. 2). In this case, the sequence space was significantly larger at ≈400,000 possible sequences after applying the volume constraint (Supplementary Table 1, Supplementary Fig. 4). By parallelizing this approach to utilise multiple CPU threads, the computations for Nile red were completed in less than 2 hours.

Thus, RASSCCoL enables *in silico* screening of hundreds of thousands of designed sequences within hours even with modest computational resources (Supplementary Table 1, Supplementary Fig. 1). This allows users to embellish the RASSCCoL pipeline as needed, for instance to include molecular dynamics (MD) simulations to explore wider backbone and side-chain conformational space (Supplementary Figs. 8 – 9).

Robust design of small-molecule binders

For experimental testing, we targeted five ligands of varying volume, polarity, and heteroatom content (Fig. 2a, Supplementary Fig. 2): the aforementioned hydrophobic fluorophores DPH ($L \approx 15$ Å; $V \approx 274$ Å^3) and Nile red ($L \approx 13$ Å; $V \approx 324$ Å^3); more-hydrophilic and biologically relevant lumiflavin ($L \approx 11$ Å; $V \approx 257$ Å^3); and biologically active arcyriaflavin A ($L \approx 12$ Å; $V \approx 310$ Å^3) and SN38 ($L \approx 14$ Å; $V \approx 396$ Å^3). For each ligand, we introduced up to ten mutations into three layers of the sc-apCC-4 scaffold (Fig. 2b and Supplementary Table 1). For lumiflavin, arcyriaflavin A and SN38, the side-chain libraries were diversified to include small polar residues, Ser, Thr, Asn, and Asp (Supplementary Table 1). In this way, we aimed to accommodate the polar functional groups of these ligands, hypothesizing that these would be captured by the AutoDock scores in the RASSCCoL pipeline (Supplementary Figs. 5 – 8). Indeed, binding-site sequences with polar amino acids were returned for these ligands (Fig. 2b, Supplementary Fig. 10). RASSCCoL calculations and full-length sequences of designs made as synthetic genes are summarised in Supplementary Tables 1 and 3.

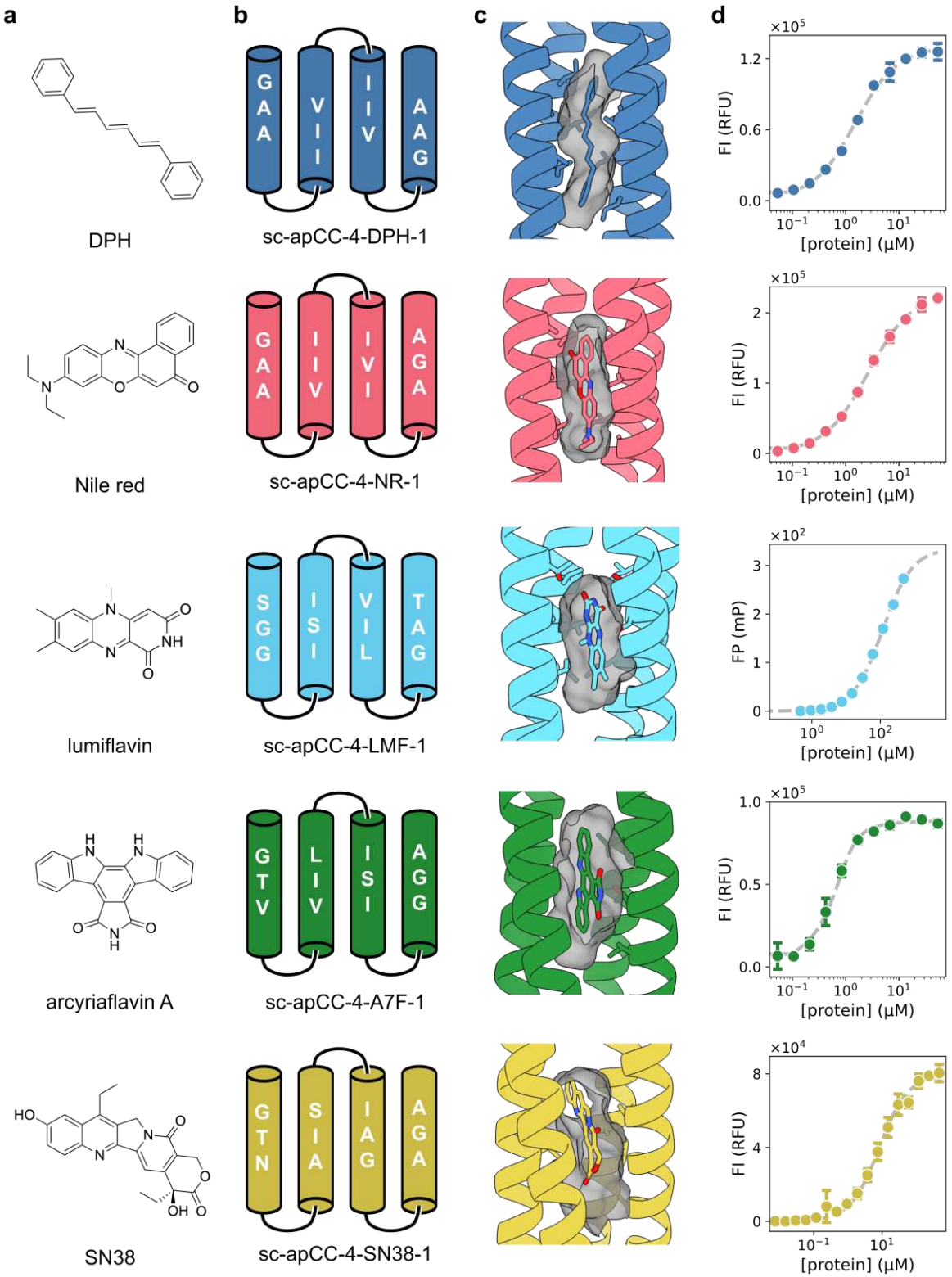


Fig. 2: RASSCCoL generates low-μM binders to a range of ligands in a starting scaffold devoid of cavities. a, Ligands tested for binding, top to bottom: DPH (blue-coloured panels), Nile red (red), lumiflavin (cyan), arcyriaflavin A (green), and SN38 (yellow). **b,** Binding-site residues within the hydrophobic core of the 4-helix bundle, sc-apCC-4 (PDB id 8A3K). **c,** Structure predictions of the binders from AF2 with ligands docked using Autodock Vina. **d,** Saturation-binding data measured by fluorescence intensity (FI) or fluorescence polarization (FP). Fitted dissociation constants, top to bottom: $K_{D(DPH)}$ = 1.2 ± 0.1 μM; $K_{D(Nile\ red)}$ = 2.4 ± 0.1 μM; $K_{D(lumiflavin)}$ > 50 μM; $K_{D(arcyriaflavin\ A)}$ = 0.11 ± 0.02 μM; $K_{D(SN38)}$ = 7.8 ± 0.6 μM.

Conditions: 0.5 μM DPH/Nile red or 1 μM lumiflavin/SN38/arcyriaflavin A, with either 10% v/v MeCN and 0 – 60 μM protein (DPH/Nile red/arcyriaflavin A) or 10% v/v DMSO and 0 – 480 μM protein (lumiflavin/SN38). DPH/Nile red/arcyriaflavin A/lumiflavin were measured in PBS, pH 7.4. SN38 was measured in ABS and at pH 4 to preserve the bioactive lactone ring. The data shown are the mean of three independent repeats, with error bars representing the standard deviations from the means. For simplicity, only the highest-affinity protein for each ligand is shown, see Supplementary Figs. 15 – 19 for the other designs.

To test the robustness of both RASSCCoL and the sc-apCC-4 scaffold, we selected just thirteen sequences for characterization based on the docking scores: two for DPH, with clockwise and anticlockwise arrangements of the helices predicted by AF2 (Supplementary Fig. 11) but with the same binding-site solution; two each for Nile red and arcyriaflavin A; three for lumiflavin; and four for SN38 (Supplementary Table 3 and Supplementary Figs. 3 – 8). All proteins expressed well, were monomeric, α helical, and thermally stable (Supplementary Figs. 12 and 16). Binding assays using fluorescence intensity (FI) or fluorescence polarization (FP) measurements returned sub-micromolar affinities for DPH and arcyriaflavin A, low-micromolar affinities for Nile red and SN38, and a weaker ≈50 uM affinity for lumiflavin (Figure 2d, Supplementary Figs. 18 – 22). Interestingly, and although ligand selectivity was not explicitly included at this stage in the RASSCCoL pipeline, we found that several of the designed proteins discriminated between the different small molecules (Supplementary Table 5, Supplementary Fig. 23). Notably, the sc-apCC-4-DPH-1 and sc-apCC-4-NR-1 proteins, with designed hydrophobic binding pockets, did not bind the more hydrophilic lumiflavin—despite its smaller molecular volume compared to DPH and Nile red—consistent with the lack of chemical complementarity (Supplementary Table 5, Supplementary Fig. 23).

As controls to test the robustness of the scaffold and any ligand binding through non-specific interactions, we mutated the targeted binding site residues to either all Gly or all Ala to give sc-apCC-4-Gly and sc-apCC-4-Ala, respectively. Despite high AF2 confidence (Supplementary Table 2), the sc-apCC-4-Gly was completely unfolded (Supplementary Fig.

12, top). However, the sc-apCC-4-Ala mutant was monomeric, α helical, and thermally stable (Supplementary Figs. 12 and 16). It bound target ligands at least one order of magnitude weaker than the best designed binders (Supplementary Table 5, Supplementary Figs. 18 – 22). Interestingly, arcyriaflavin A was the most promiscuous ligand in our panel, and it was the only ligand to bind the Ala-based control with μM affinity.

We determined high-resolution crystal structures of the ligand-free forms for all five *de novo* binders. (Fig. 3, Supplementary Tables 6 – 7). The DPH binder crystalised as a domain-swapped dimer despite being monomeric in solution (Supplementary Figs. 16 and Supplementary Figures 24 – 25). Nonetheless, both domains aligned well with the AF2 model, with the designed binding site geometry maintained (Fig. 3a). The Nile red, lumiflavin, arcyriaflavin A and SN38 binders matched the respective AF2 models closely with overall root-mean square differences between the C_a atoms (C_α-RMSDs) of < 0.5 Å (Fig. 3b – e). All five apo-protein structures retained the overall backbone conformations close to the models predicted by AF2 (Fig. 3, and Supplementary Fig. 26), and revealed solvent accessible, pre-organised binding pockets with side-chain geometries and binding-site heavy atom RMSDs of < 0.5 Å. Thus, the observed cavity shapes and volumes matched the computational designs closely. Despite multiple attempts, we could not obtain any crystals for the ligand:protein complexes from mixtures or by soaking crystals in ligand solutions. However, redocking the ligands into the crystal structures returned almost identical poses to the computational design models (Fig. 3a – d). The only exception was the SN38 binder, where docking positioned the ligand within the binding pocket of the crystal structure, but not in the AF2 model (Fig. 3e). Moreover, for lumiflavin, arcyriaflavin A and SN38, inspection of the structures and docking predictions revealed hydrogen bonds between the ligands and introduced polar side chains (Fig. 3c – e).

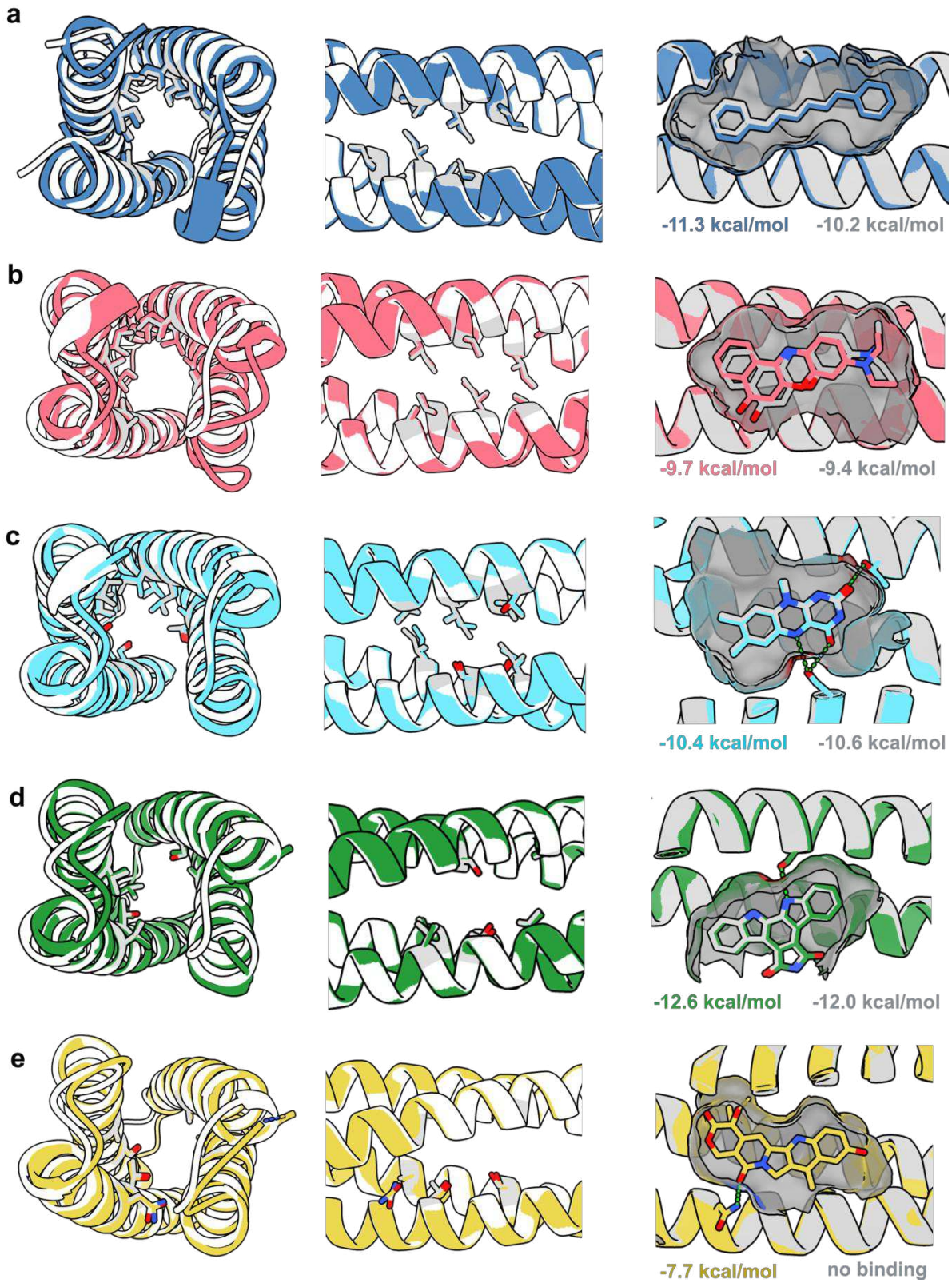


Fig. 3. Binding-site models generated by RASSCCoL closely match ligand-free crystal structures. Binding-site heavy atom alignments between AlphaFold2 models (grey) and the experimental X-ray crystal structures (coloured as in Fig. 2). **a,** The DPH binder, sc-apCC-4-DPH-2, at 2.4 Å resolution (with a root-mean square difference between the side-chain heavy atoms (sc-RMSD) = 0.45 Å, PDB id 9R1N). Only one domain of the domain-swapped dimer is shown, see Supplementary Fig. 25. **b,** The Nile red binder, sc-apCC-4-NR-1, at 1.7 Å resolution (sc-RMSD = 0.33 Å, PDB id 9R1J). **c,** The lumiflavin binder, sc-apCC-4-LMF-1, at

2.4 Å resolution (sc-RMSD = 0.30 Å, PDB id 9R1M). **d,** The arcyriaflavin A binder, sc-apCC-4-A7F-1, at 2.3 Å resolution (sc-RMSD = 0.35 Å, PDB id 9R1K). **e,** The SN38 binder, sc-apCC-4-SN38-1, at 1.85 Å resolution (sc-RMSD = 0.38 Å, PDB id 9R1L). The right-hand column shows docked poses and AutoDock Vina scores for the target ligands within the crystal structures, overlaid with the docked poses within the AF2 models (grey). Binding-site residues are highlighted and shown as sticks, with oxygen atoms in red and nitrogen atoms in blue. Hydrogen bonds are shown as dashed green lines. The X-ray crystal structures were energy-minimised using Rosetta prior to docking. This was done to relax some of the peripheral side chains that were not fully resolved in the experimental maps.

RASSCCoL scales to larger ligands and scaffolds

Traditional protein design methods that explore combinations of side chains do not scale well with increasing number of designable positions and large amino-acid palettes[28-30]. For our *de novo* scaffolds, significant reduction of the sequence space is possible by exploiting the symmetry, layered structures, and well-understood sequence-to-structure relationships of α-helical bundles. However, in more general cases using RASSCCoL, even with its optimised and relatively fast model-building and docking steps, the process would become computationally expensive once sequence space exceeded a few million. To mitigate this, we implemented an active sequence sampling strategy using Gradient Boosted Trees (GBT)[31]. This prioritises the most-informative data points rather than exhaustively docking all possible sequences, significantly reducing computational cost and time (Supplementary Figs. 27 – 28).

To test this approach, we used a recently designed antiparallel 6-α-helix barrel, sc-apCC-6 (PDB id 8QAD) [22]. Unlike sc-apCC-4, this has a central hydrophobic cavity that binds DPH with low-micromolar affinity but has a weaker affinity for Nile red (Supplementary Fig. 29). We aimed to reverse this selectively and design a variant that bound Nile red but not DPH by exploiting only shape complementarity using RASSCCoL. The smaller 4-helix bundle required redesigning both ***a*** and ***d*** core positions of the ***abcdefg*** coiled-coil sequence repeats to generate binding sites (Supplementary Fig. 30). However, due to the increased radius and

changes in core packing of the 6-helix barrel[22], we reasoned that redesigning only ***a*** positions should accommodate the ligands (Supplementary Fig. 30). In addition, the larger scaffold allowed the amino-acid pallet to be expanded to include Phe, which the 4-helix bundle cannot accommodate. Even by restricting designs to just the ***a*** sites of three layers and a subset of hydrophobic residues (Gly, Ala, Val, Leu, Ile, and Phe), this generated ≈1.3 million sequences after the cavity volume filter. Using active sampling with GBT, only 100,000 sequences were evaluated using Auto Dock Vina, thus reducing the computation by over tenfold.

In more detail, GBT predictors were trained initially on a small, randomly selected subset of RASSCCoL designs of Nile red binders in the sc-apCC-6 background, and these were used to predict AutoDock Vina scores for the remaining sequences. The training set was then iteratively expanded by explicitly docking sequences with the highest prediction uncertainty (i.e., knowledge gaps), the lowest predicted docking scores (i.e., most-promising binders), and randomly chosen sequences to ensure smooth sampling of the sequence space. This process of docking and retraining continued until no further improvement in predictor's mean-squared error was observed on a randomly selected validation set. At this stage, any remaining unsampled sequences with predicted docking scores below a user-defined cutoff were explicitly docked and passed to the final stages of the RASSCCoL pipeline — filtering and refolding.

Again, we only tested two of the resulting designs experimentally: one intended to bind both ligands equally well (sc-apCC-6-DPH-NR, labelled P for promiscuous), and another specific for Nile red (sc-apCC-6-NR, labelled S for specific) (Fig. 4a). Both proteins expressed well, were monomeric, α helical, and thermally stable (Supplementary Figs. 14 and 17). As predicted, we observed low-μM affinities for DPH and Nile red in the promiscuous binder (Fig.

4b, panel P). By contrast, the selective binder exhibited 170-fold tighter binding for Nile red compared to DPH, with a sub-μM dissociation constant for the former (Fig. 4b, panel S). Inspection of the docked AF2 models suggests that selectivity against DPH—which is ≈2 Å longer than Nile red—arises from subtle shortening of the binding pocket, causing steric clashes between DPH and residues at the extremes of the binding pocket (Fig. 4c).

Nile red is an environment-sensitive probe[32]. This is apparent in changes in its emission spectrum upon binding to the P and S designs (Fig. 4d): *i.e.*, the resolved vibronic bands similar to those seen in n-hexane; and the blue-shifted fluorescence peak relative to that observed in dioxane and the more-solvent-exposed binding site of sc-apCC-4-NR-1[25,33]. This infers binding to tight hydrophobic pockets of the P and S designs. Finally, we solved a structure of P without ligand to 3.2 Å. This matched the AF2 model with Cα-RMSD of 0.3 Å (Fig. 4e, PDB id 9R1O). Despite the medium resolution, electron density for the designed residues—particularly the Phe side chains—was well-resolved (Fig. 4e). Redocking the ligands into this structure produced poses that closely aligned with the designed models (Fig. 4f).

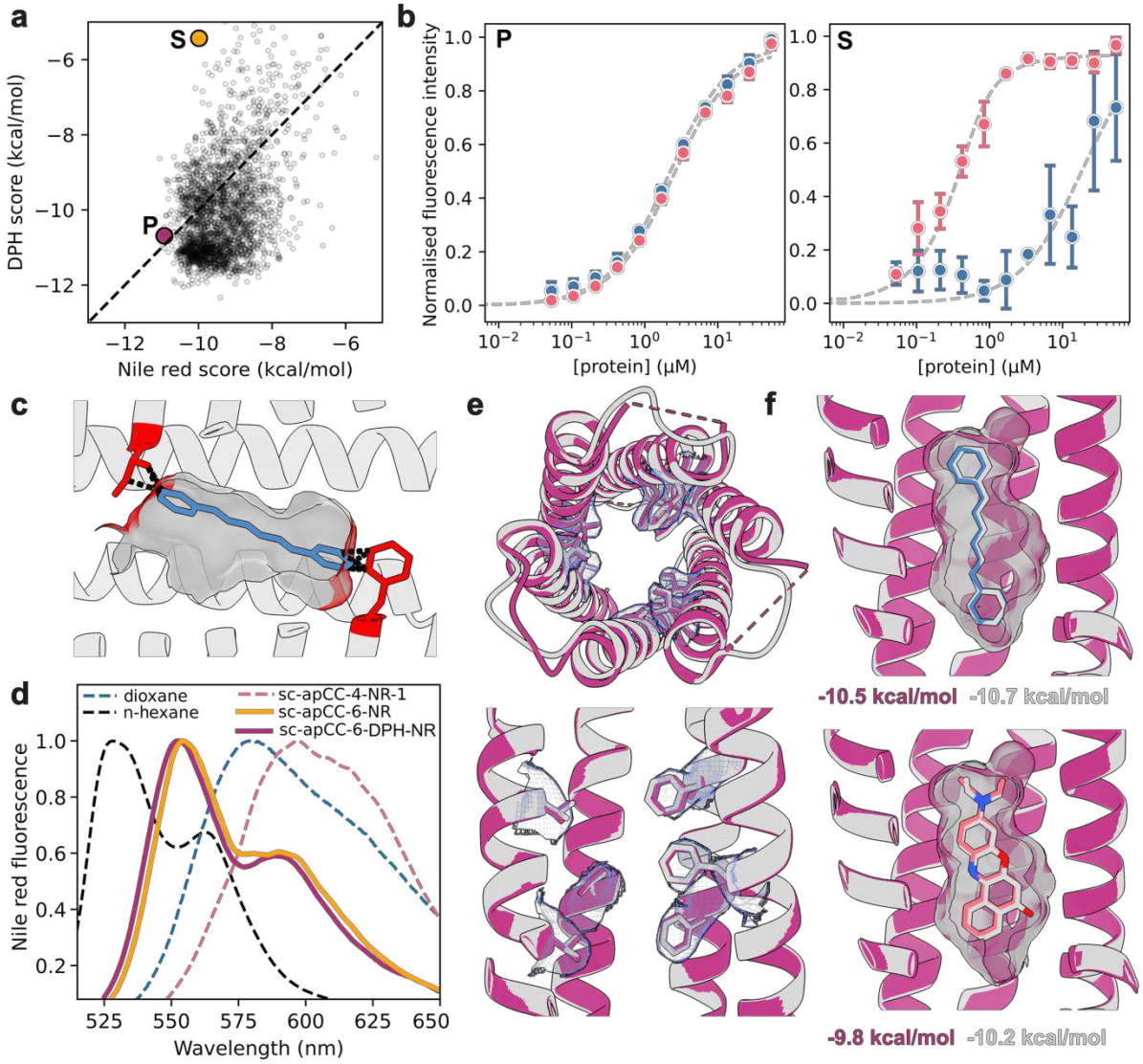


Fig. 4. RASSCCoL-generated promiscuous and selective binders for DPH and Nile red. **a,** AutoDock Vina scores for DPH and Nile red docked into AF2 models for the top 1,000 generated designs. The dashed line is the 1:1 diagonal, and the Nile red binders are labelled P (promiscuous) and S (selective). **b,** Saturation binding curves and fits for Nile red and DPH to the proteins. **Left**, promiscuous binder, $K_{D(DPH)}$ = 1.9 ± 0.2 μM, $K_{D(Nile\ red)}$ = 2.2 ± 0.2 μM. **Right**, the specific binder, $K_{D(DPH)}$ > 20 μM, $K_{D(Nile\ red)}$ = 0.11 ± 0.04 μM. **c,** AF2 model for the specific binder with DPH docked. Selectivity against DPH is achieved by Leu and Phe residues (red) that shorten the binding site. Severe clashes between DPH and the side chains (overlap > 1 Å) are shown in black. **d,** Normalised emission spectrum for Nile red after excitation at 500 nm in aqueous solution with the specific and promiscuous 6-helix binders, compared to the 4-helix Nile red binder and organic solvents. **e,** Structural alignment between the AlphaFold2 model (grey) and the experimental X-ray crystal structure of the promiscuous binder P (sc-apCC-6-DPH-NR) at 3.2 Å resolution (Cα RMSD = 0.294 Å, PDB id 9R1O). NCS-averaged electron density for the designed residues is shown at 1σ level in blue. **f,** Docked poses of DPH and Nile red in the crystal structure of the promiscuous binder P, overlaid with the design target (docked ligands within the AF2 model; grey). Oxygen atoms shown in red and nitrogen atoms in blue. The crystal structure was energy-minimised using Rosetta prior to docking. Binding conditions: 0.5 μM ligand, 0 – 60 μM protein, PBS, 10% v/v MeCN,

pH 7.4. Data are the mean of three independent repeats, error bars represent the standard deviation from the mean.

Binders function *in vitro* and in cells

Building beyond single-domain, small molecule-binding proteins will require generating binding sites (or binding modules) that are orthogonal and can be co-assembled in modular manner into more-complex *de novo* frameworks. To explore this, we used two of the designed 4-helix bundles that bind DPH and Nile red selectively (sc-apCC-4-DPH-1 and sc-apCC4-NR-1, Supplementary Table 5, Supplementary Fig. 23) to construct a two-domain *de novo* protein. For this, the domains were fused *via* a rigid, read-through helical linker (Fig. 5a). Since DPH and Nile red are a known Förster resonance energy transfer (FRET) pair[25,34] (Supplementary Fig. 32), we designed this construct to position the two dyes strategically at the Förster distance of ≈50 Å. The resulting protein was expressed in *E. coli*, and was monomeric, α helical and thermally stable (Supplementary Figs. 12 and 16). Small-angle X-ray scattering (SAXS) indicated an extended overall conformation consistent with the model (Fig. 5a,b). The protein retained its affinities for both ligands (Fig. 5c) and the presence of one ligand did not substantially alter the binding of the other (Supplementary Fig. 31). Moreover, with both ligands present, selective excitation of DPH resulted in both DPH and Nile red fluorescence (Fig. 5d and Supplementary Figs. 34 – 35). The latter must arise from FRET and thus indicates a population of the protein with both DPH and Nile red bound (Supplementary Table 9). The FRET lifetime of ≈2 ns gave a predicted distance of ≈51 Å, consistent with the computational design model and MD simulations of it (Supplementary Fig. 36). This experimentally confirmed the design of the dual binder (Fig. 5e).

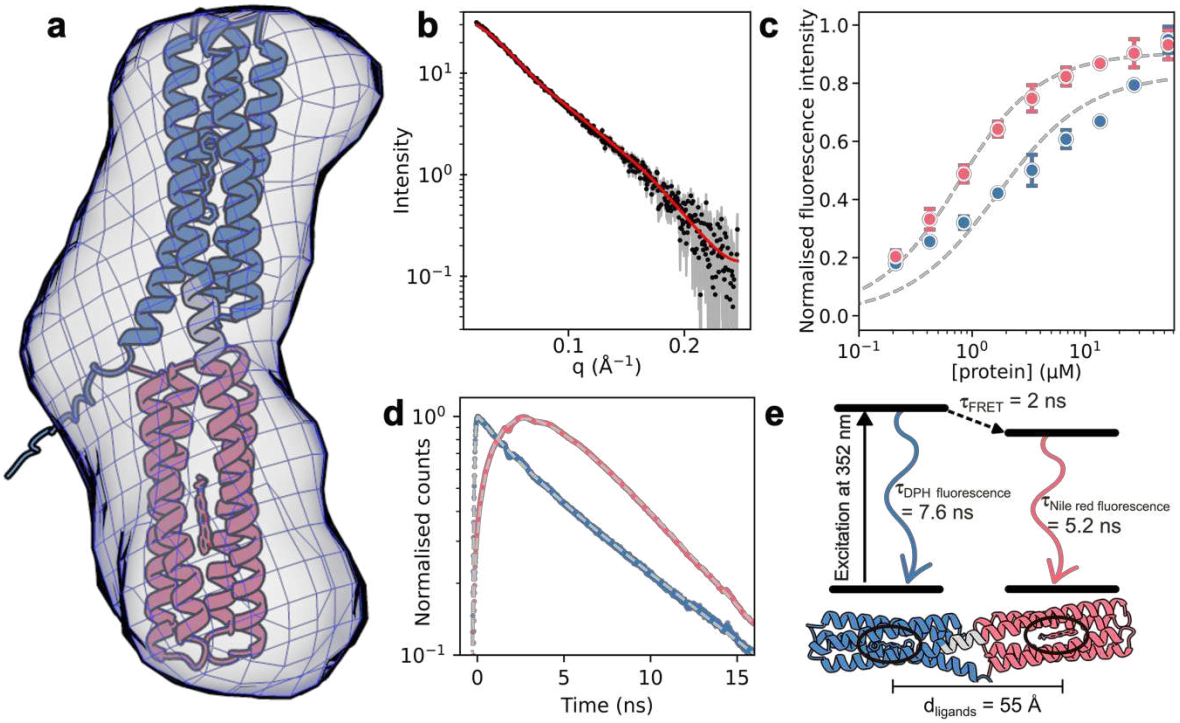


Fig. 5: RASSCCoL-derived double-binder that transfers energy from DPH (blue) to Nile red (red). a, AF2 model of the double-binder that combines sc-apCC-4-DPH-1 and sc-apCC-4-NR-1, superposed to the averaged *ab initio* model from small angle X-ray scattering data. DPH and Nile red are docked with AutoDock Vina. **b,** Small angle X-ray scattering data and FoXS [35] fit to the AF2 model (red, $\chi 2 = 1.44$). Standard error shown in grey. **c,** Saturation binding curves and fits: blue, $K_{D(DPH)} = 1.3 \pm 0.4$ μM; red, $K_{D(Nile\ red)} = 0.53 \pm 0.05$ μM. **d,** Fluorescence lifetime decay measurements for DPH (blue, < 450 nm) and Nile red (red, >600 nm) after selective excitation of DPH at 352 nm. The delayed rise in Nile red emission over the first few nanoseconds originates from Förster resonance energy transfer between DPH and Nile red. In contrast, control measurements that excited only Nile red (Supplementary Fig. 35) showed no delayed emission. Bi-exponential fit of the data returned $t_{FRET} = 2.0 \pm 0.2$ ns, yielding an inter-ligand distance of 51 ± 1 Å. The expected distance from MD simulations of the model from panel (**a**) was 55 ± 1 Å, see Supplementary Fig. 36 and Supplementary Table 9. **e,** Kinetic model for energy transfer between DPH and Nile red, derived from panel **(d)** and direct excitation of DPH and Nile red in isolation (Supplementary Figs. 34 and 35)**.** Binding conditions: 0.5 μM ligand, 0 – 60 μM protein in PBS, 10% v/v MeCN, pH 7.4. Binding data are the mean of three independent repeats, error bars represent the standard deviation from the mean. SAXS conditions**:** 5 mg/ml protein in PBS, pH 7.4.

Finally, we explored the potential for using the *de novo* synthetic chromophore-binding proteins to target and visualise subcellular structures in mammalian cells. There are advantages of using *de novo* fluorescent proteins for such applications. For instance, at ≈130 residues and ≈15 kDa, sc-apCC-4 variants are approximately half the size of commonly used GFP variants[36]

and on par with some of the smallest FPs, such as: iLOV (13 kDa)[37]; SNAP-tag FP mimics (20 kDa)[38]; *de novo* β barrels (14 kDa)[39]; and transmembrane 4-helix bundles (21 kDa)[40]. Moreover, bright fluorophore encapsulation within proteins generally leads to improved and tuneable spectral properties compared to conventional FPs. To examine this, we found that Nile red bound to sc-apCC-4-SN38-1, had a high fluorescence quantum yield and thus brightness (e_{552} = 52 400 cm^{-1} M^{-1}; Φ = 0.9), and a long fluorescence lifetime (τ = 4.5 ns) (Supplementary Table 8 and Supplementary Fig. 37). This is brighter and longer lived than conventional fluorescent proteins in the same spectral region, such as mCherry (e_{587} = 72 000 cm^{-1} M^{-1}; Φ = 0.2; τ = 1.4 ns) [41]. Furthermore, emission wavelengths can be tuned toward the near-infrared (NIR) by substituting Nile red for derivatives with similarly high quantum yields, brightness and long fluorescence lifetimes; e.g., Nile blue (e_{627} = 34 400 cm^{-1} M^{-1}; Φ = 0.74; τ = 4.2 ns) (Supplementary Table 8 and Supplementary Figs. 38 – 40)[42]. On this basis, we tested several of our *de novo* protein modules directly in HeLa cells as fusions to GFP, without and with peptide tags for directing the constructs to different subcellular regions (Supplementary Table 4, Supplementary Fig. 41). This is illustrated below for sc-apCC-4-SN38-1 (Fig. 6), which binds both Nile red and Nile blue; $K_{D,\text{Nile Red}} \approx 0.2$ μM and $K_{D,\text{Nile blue}} \approx 0.4$ μM (Supplementary Table 5, Supplementary Figs. 23 and 38).

As a control, we fused the parent sc-apCC-4 to the *N* terminus of mEmerald variant of the green fluorescent protein and transformed the construct into HeLa cells (Supplementary Table 4, Supplementary Fig. 41). The transiently expressed fusion protein was dispersed throughout the cell. Next, we added short, *N*-terminal peptide sequences to the fusion to target: the nucleus with a nuclear-localization signal (NLS)[43]; the plasma membrane with a Fyn-tag[44]; and the actin network with LifeAct[45] (Supplementary Table 4). In all cases, the mEmerald signal was redistributed and located as expected (Supplementary Fig. 41). These experiments were

404 performed with 0.1 μM Nile red added to the cell media 30 minutes before imaging. However,
405 with these non-Nile-red-binding controls, none of the images showed any coincidence of the
406 mEmerald and Nile red signals. By contrast, when sc-apCC-4 was substituted with sc-apCC-
407 4-NR-1 or sc-apCC-4-SN38-1, the mEmerald and Nile red signals were coincident as evident
408 in the merged images (Fig. 6 and Supplementary Fig. 41). These indicate that the dye had been
409 taken up by cells and sequestered by the designed Nile red-binding components of the fusion
410 proteins. Interestingly, the images were sharper and clearer with the sc-apCC-4-SN38-1
411 design. We attribute this to its sub-μM affinity for Nile red, which is 10-fold tighter than for
412 sc-apCC-4-NR-1 (Supplementary Table 5 and Supplementary Fig. 23). Finally, to demonstrate
413 that the sc-apCC-4 variant alone is sufficient as a *de novo* fluorescent protein chassis, we
414 repeated the experiments without the mEmerald reporter, and these worked similarly to the
415 fusion constructs showing robust staining of targeted subcellular structures (Supplementary
416 Fig. 42).
417

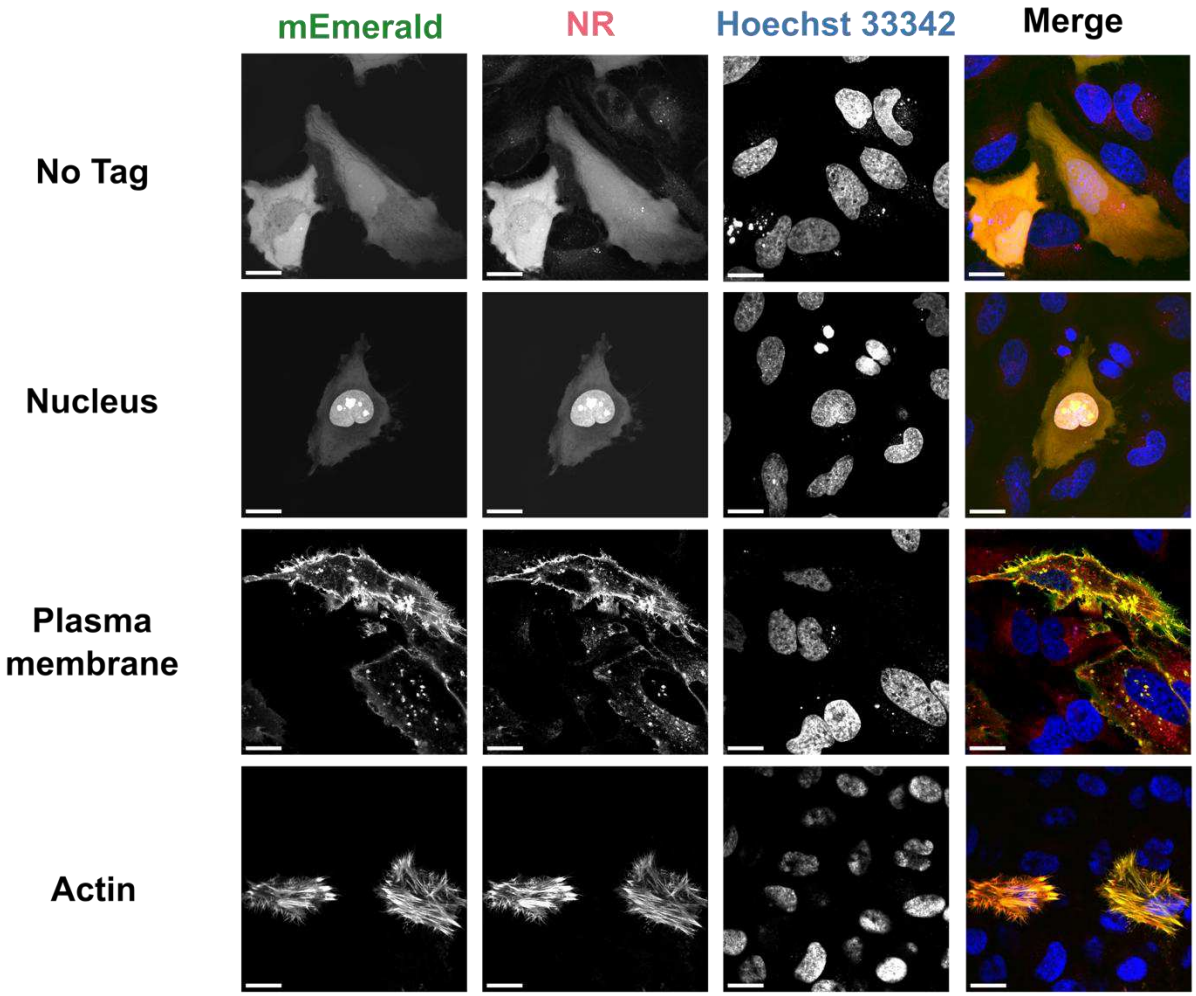


Fig 6: *De novo* designed fluorescent protein for subcellular imaging. (**Top row**) Fluorescence confocal microphotographs of live HeLa cells transiently expressing a fusion of sc-apCC-4-SN38-1 and mEmerald, and stained by adding 0.1 μM Nile red and Hoechst 33342 to the cell media 30 minutes before imaging. (**Rows 2 – 4**) Similar images for additional fusions with *N*-terminal tags for locating the proteins to the nucleus, plasma membrane, and actin filaments. Merge (right-hand column): mEmerald (green), NR (red), and Hoechst 33342 (blue). Maximum z-projections are shown for constructs with cytosolic and nuclear localization, single-z slices are shown for actin- and membrane-targeting fusions; scalebars = 20 μm.

Conclusion

Designing proteins that bind small molecules tightly and specifically is a key challenge for protein design and synthetic biology, with potential applications including bioimaging, *in vivo* sensing, catalysis and therapeutics[3]. Here we present RASSCCoL, a fast, efficient, and intuitive computational strategy for introducing small molecule-binding sites into *de novo* designed

proteins. This works by replacing side chains of the *de novo* scaffold to create cavities that are isosteric and chemically complementary with the target ligand. The calculations are vastly accelerated by: i) using a volume filter to enrich for sequences most likely to accommodate the ligand; ii) rapid modelling with FASPR[27] and AutoDock[19]; (iii) evaluation of the models using statistical and physical metrics; and iv), where needed for large targets, an active-learning procedure to reduce the number of explicit modelling steps. To put this approach into practice, we leverage well-understood, mutable, and thermostable *de novo* proteins as the scaffolds; namely, a growing family of single-chain α-helical bundles and barrels[21,22]. This streamlines the design process even further by bypassing backbone optimization and focusing on sequence design and refinement.

Specifically, we have designed low-to-sub-micromolar protein binders for two hydrophobic fluorophores, DPH and Nile red; for lumiflavin, a proxy for flavin-based cofactors; and for two cytotoxic compounds, arcyriaflavin A and SN38. The last three feature polar functional groups that RASSCCoL accommodates with complementary polar side chains. In each case, successful designs are achieved by testing just 1 – 4 expressed genes in *E. coli*. The approach scales well to larger scaffolds, or with additional design constraints such as ligand selectivity, which we also demonstrate experimentally. To test the modularity of some of the resulting designs, we combine two different binding modules for DPH and Nile red (a known FRET pair) in a two-domain protein that holds the chromophores in a specified orientation and at a set distance to enable energy transfer between the two sites. Finally, we demonstrate the potential of our designs as small (≈15 kDa), stable, mutable, and adaptable fluorescent protein tags for visualizing subcellular structures and processes in mammalian cells.

Together with advances by others[10,11,40], our findings reinforce how the use of straightforward, well-understood, and fully characterised *de novo* scaffolds can accelerate functional protein design. In this way, RASSCCoL offers a robust and interpretable platform for creating next-generation protein-based tools and modules for broader applications in cell and synthetic biology, diagnostics, and therapeutics.

Methods

General

All solvents, chemicals, and reagents were purchased from commercial sources and used without further purification. Luria-Broth (LB), antibiotics, IPTG, and L-rhamnose were purchased from Thermo Fisher. All other chemicals were reagent grade and purchased from Sigma-Aldrich. Protein biophysical characterization were recorded in a PBS buffer (50 mM sodium phosphate, 150 mM NaCl, pH 7.4).

Computational design

Van der Waals volumes for ligands, side chains, and cavities were calculated from Bondi radii[26]. Binding-site sequences were generated and filtered for cavity volumes matching ligand size within a user-defined cutoff. For SN38, the closed-lactone form representing the pharmacologically active tautomer was used during the computational design process, as only this form engages in canonical binding interactions *in vivo*[46]. Passing sequences were packed with FASPR[27]. Ligands were docked with AutoDock Vina[19] using parameters optimised for speed (exhaustiveness of 1, box size of 1.5 × ligand length, spacing of 0.5 Å, and 5,000 evaluations). Sequences below a user-defined energy cutoff were refolded with AF2, see Supplementary Figs. 3 – 8. For Nile red, lumiflavin, arcyriaflavin A and SN38, only AF2 model

4 was used. Earlier DPH designs used all five AF2 models. Candidates for experimental validation were chosen based on AutoDock Vina[19] scores after redocking with default parameters (exhaustiveness of 16, spacing of 0.375 Å, and 10,000 evaluations). Throughout the pipeline, no backbone or side-chain minimisation or relaxation was performed.

For the active sampling protocol, binding site cavity volumes and integer encodings for amino acids were used as features to train Gradient Boosted Trees (GBT). For the first sampling iteration, 20,000 sequences were picked randomly, repacked, scored with AutoDock Vina and split 80/20 into train and validation sets. These were used to train GBTs to predict Vina scores for the remaining sequences. 1,000 sequences with the best scores, 1,000 with the highest prediction uncertainty and 1,000 random sequences were scored with Vina and added to the training set. This was iterated until the Mean Squared Error on the validation set stopped decreasing; this took 10 iterations. The trained model was then used to predict Vina scores for the remaining sequences; 50,000 best predictions were then taken further. This pipeline is implemented using TensorFlow Decision Forests[31].

Additional details of RASSCCoL pipeline are available on Woolfson Lab GitHub (https://github.com/woolfson-group/RASSCCoL).

Molecular dynamics simulations

OpenMM 7.7 was used for running MD simulations[47]. Simulations were run with the Amber ff14SB force field, Langevin Middle integrator with an integration step of 4 fs, 1.5 au H-mass repartitioning and collision rate of 1 ps^{-1}.

For sc-apCC-4-DPH-1-sc-apCC-4-NR-1, the system was solvated with tip3p waters and 100 mM NaCl in a rhombic dodecahedron with a padding of 1 nm, particle mesh Ewald with a cut-off of 1 nm was used for long-range electrostatics, and the pressure was kept constant at 1.01325 bar with a MonteCarlo barostat. The AlphaFold2 model with the ligands docked was initialised at 10 K with 5 kcal/mol restraints on the protein backbone and ligand heavy atoms and minimised. The system was then heated to 300 K over 600 ps. Heavy atom restraints were relaxed over 3.5 ns and the unrestrained system was further equilibrated for 10 ns. A production run was then performed for 100 ns. Periodic boundary conditions were removed using GROMACS trajconv[48], trajectories were analysed using MDAnalysis[49].

Short MD simulations to evaluate SN38 binding were run with implicit water (obc2) using a repacked 4-helix bundle (PDB id 8A3K) with SN38 docked within the binding pocket. The structure was initialised at 0 K, minimised, heated to 400 K over 400 ps, and then equilibrated for 1.6 ns. A production run was performed for 5 ns. Trajectories were analysed using MDAnalysis[49], and ligand binding affinities were analysed using AutoDock Vina. The simulations were run with and without 20 cal mol^{-1} $Å^{-2}$ backbone heavy atom restraints to minimize protein unfolding at elevated temperatures.

Protein expression and purification

All genes were directly cloned into pET28a vectors, transformed and then expressed in *E. coli* Lemo21-DE3 (New England Biolabs). Flasks containing 1 L of Miller's Luria Broth–kanamycin–chloramphenicol and 0.5 mM L-rhamnose were inoculated with 5 ml of overnight cultures and incubated to an optical density at 600 nm of ~0.6 at 37 °C with 200 r.p.m. shaking. Expression was induced with 0.5 mM isopropyl-β-D-thiogalactoside, and cultures were incubated at 37 °C overnight with 200 r.p.m. shaking. Following expression, cultures were

pelleted and resuspended in 20 ml lysis buffer (50 mM Tris, pH 7.4, 500 mM NaCl, 30 mM imidazole, 1 mg ml^{-1} lysozyme) for 30 min at 37 °C. Resuspended pellets were sonicated using a Biologics Model 3000 Ultrasonic homogenizer with settings at 50% power and 90% pulser (1 pulse per second) for 5 min and then clarified at 25,500 *g* for 30 min. The clarified lysate was heat shocked at 75 °C for 10 min and then cooled on ice for 10 min before reclarifying at 25,500 *g* for 10 min. The expressed proteins were first purified with Ni-based immobilised metal affinity chromatography at room temperature. Filtered lysate was loaded onto an ÄKTAprime plus (GE, PrimeView 5.31) equipped with a HisTrap HP 5-ml column (Cytiva). His-tagged proteins were eluted using a single step gradient from 0 to 55% buffer B (buffer A consisted of 50 mM Tris, 500 mM NaCl and 30 mM imidazole at pH 7.4; buffer B consisted of 50 mM Tris, 500 mM NaCl and 300 mM imidazole at pH 7.4). Fractions were combined and further purified by SEC using a HiLoad 16/600 Superdex 200-pg or 75-pg size exclusion column (Cytiva) equilibrated in buffer containing 50 mM sodium phosphate and 150 mM NaCl (pH 7.4) at room temperature. Eluted fractions were pooled, concentrated and separated using SDS–PAGE to confirm protein identities.

Circular dichroism spectroscopy

Circular dichroism data were collected on a JASCO J-810 or J-815 spectropolarimeter fitted with a Peltier temperature controller in the far UV region. Spectra Manager (1.55) was used for data collection. Protein samples were prepared as 5-15 μM solutions in 50 mM sodium phosphate and 150 mM NaCl (pH 7.4) at 5 °C. Data were collected in a 1-mm quartz cuvette between wavelengths of 190 nm and 260 nm with the instrument set as follows: band width, 1 nm; data pitch, 1 nm; scanning speed, 100 nm min^{-1}; response time, 1 s. Each circular dichroism spectrum was obtained by averaging eight scans and subtracting the background signal of the buffer and cuvette. For thermal response experiments, the circular dichroism

signal at a 222 nm wavelength was monitored over the temperature range 5 – 95 °C at a ramping rate of 60 °C per hour with the same settings and protein concentrations given above. The spectra were converted from ellipticities (mdeg) to mean residue ellipticities (deg·cm^2·dmol^{-1}·res^{-1}) by normalizing for concentration of peptide bonds and the cell path length using the equation:

$$MRE = \frac{\theta \times 10^6}{c \times l \times n} \quad (1)$$

where the variable θ is the measured difference in absorbed circularly polarized light in millidegrees, c is the micromolar concentration of the compound, l is the path length of the cuvette in millimeters, and n is the number of amide bonds in the polypeptide.

Analytical ultracentrifugation

AUC was performed on a Beckman Optima X-LA or X-LI analytical ultracentrifuge with an An-50-Ti or An-60-Ti rotor (Beckman-Coulter) equipped with ProteomeLab XL-A (5.5) software. Buffer densities, viscosities, and peptide and protein partial specific volumes ($\bar{v}$) were calculated using SEDNTERP (http://rasmb.org/sednterp). For sedimentation velocity, protein samples were prepared in PBS at a 30 μM protein concentration and placed in a sedimentation velocity cell with a two-channel centerpiece and quartz windows. The samples were centrifuged at 50 k.r.p.m. at 20 °C, with a total of 120 absorbance scans taken over a radial range of 5.8 – 7.3 cm at 5 min intervals. Data from a single run were fitted to a continuous $c(s)$ distribution model using SEDFIT (v15.2b) at a 95% confidence level [50]. Residuals for sedimentation velocity experiments are shown as a bitmap in which the grayscale shade indicates the difference between the fit and raw data (residuals, <−0.05 black and >0.05 white). Good fits are uniformly gray without major dark or light streaks.

Ligand binding

Ligand-binding experiments were performed in triplicate. The total concentration of ligand was kept constant: 0.5 μM DPH/Nile red or 1 μM arcyriaflavin A in 10% v/v MeCN; 1 μM lumiflavin/SN38 in 10% v/v DMSO. The concentration of the *de novo* protein design was varied from 0 – 54 μM. Fluorescence intensity and polarisation data were collected on a Clariostar plate reader (BMG Labtech) using the following excitation wavelength: λ_{DPH} = 352 nm, $\lambda_{Nile\ red}$ = 520 nm, $\lambda_{arcyriaflavin\ A}$ = 320 nm, $\lambda_{lumiflavin}$ = 450 nm, λ_{SN38} = 350 nm. Binding constants were extracted by fitting the data to the following equation:

$$(2) \quad y = B_{max}\ \frac{(c + x + K_D) + \sqrt{(c + x + K_D)^2 - 4cx}}{2c}$$

where c is the total concentration of the constant component (ligand), x is the concentration of variable component (protein), Bmax is the fluorescence intensity/polarisation signal when all the constant component is bound, and y is the fraction of bound component being monitored via fluorescence signal.

Small-angle X-ray scattering

Data for sc-apCC-4-DPH-1-sc-apCC-4-NR were collected at The European Synchrotron Radiation Facility beamline BM29. Batch mode samples were prepared at 0.5 - 5 mg/mL in 50 mM sodium phosphate pH 7.4 and 150 mM NaCl. The His-tag was not removed as we observed non-specific cleavage at the domain linker after incubation with TEV protease.

Data for sc-apCC-4-DPH-2 were obtained at the Diamond Light Source beamline B21. The His-tag was removed using thrombin protease. Size-exclusion chromatography mode samples were prepared at 20 mg/mL in 20 mM Tris pH 8.0, 50 mM NaCl, 1% v/v glycerol. Shodex KW403 column was equilibrated in the same buffer at 4 °C.

For both designs, buffer subtraction and data merging was performed with ScÅtter[51]. q_{min} was taken as the first point of the linear Guinier region, q_{max} was calculated using ShaNum through ATSAS interface[52]. MultiFoxS software (Sali Lab) was used to compare experimental scattering profiles to design models and assess quality of fit by calculating χ^2 [35,53]. Sc-apCC-4-DPH-1-sc-apCC-4-NR data were fit to a His-tagged AlphaFold2 model. Sc-apCC-4-DPH-2 data were fit to a dimer observed in the crystal structure (PDB id 9R1N) and a monomeric AlphaFold2 model.

X-ray crystallography

Crystals were grown using a sitting-drop vapour-diffusion method. Proteins in 20 mM Tris pH 8.0, 50 mM NaCl were concentrated to 20 mg/mL. Commercially available sparse matrix screens were used (Morpheus®, JCSG-plus™, Structure Screen 1 and 2, Pact Premier™, ProPlex™; Molecular Dimensions), and the drops were dispensed using a robot (Oryx8; Douglas Instruments). For each well of an MRC 2 drop plate, 0.3 µL of peptide or protein solution and 0.3 µL of reservoir solution in parallel with 0.4 µL of the peptide or protein solution and 0.2 µL of reservoir solution were mixed and the plate was incubated at 20 °C. Diffraction-quality crystals were grown through multiple rounds of seeding. Crystals generally formed within a month, and after looping were soaked in reservoir solution containing 25% glycerol as a cryoprotectant. Final crystallization conditions for are provided in Supplementary Table 6.

Diffraction data for the crystals were obtained at the Diamond Light Source (Didcot, UK) on beamlines I04 or I24. Sc-apCC-4-DPH-2, sc-apCC-4-NR-1, sc-apCC-4-SN38-1 and sc-apCC-4-A7F-1 data were processed using the automated Xia2 pipeline[54], which ports data through DIALS[55] to POINTLESS and AIMLESS[56] as implemented in the CCP4 suite[57]. Sc-apCC-6-

DPH-NR and sc-apCC-4-LMF-1 were processed using the AUTOPROC pipeline, which use the same integrating and data reduction software in addition to STARANISO[58]. All structures were phased by molecular replacement using the AlphaFold2 model for PHASER[59]. Final structures were obtained after iterative rounds of model building with COOT[60] and refinement with Phenix Refine[61]. Solvent-exposed atoms lacking map density were deleted. Data merging and refinement statistics for are provided in Supplementary Table 7.

Fluorescence lifetime measurements

Time-correlated single photon counting (TCSPC) data were acquired using a home-built TCSPC apparatus previously reported.[62] The output of a tunable high-power ultrafast oscillator (Chameleon Ultra II, 3.7 W, 80 MHz, Coherent) was frequency doubled to generate the required 352 nm pulses for DPH excitation. For Nile red and Nile blue, the light was first frequently shifted to 1120 nm with an optical parametric oscillator (Chameleon Compact, APE) and then frequency doubled to generate 560 nm excitation pulses. To avoid re-excitation of samples, the repetition rate was reduced to 6.67 MHz using a pulse picker (cavity dumper, APE). The diffracted output was spatially filtered with a pinhole to remove residual zeroth order diffraction, and the resulting light was focused into a sample (1 cm path length cell). Fluorescence was collected at 90° relative to excitation using an infinity-corrected microscope objective (4×/0.2 NA Plan Apochromat, Nikon) and through a series of achromatic lenses collimated and focused onto an avalanche photodiode detector (ID100-50-ULN, IDQ). Nile red and Nile blue fluorescence was collected with a >600 nm filter, DPH with a 450 nm bandpass filter. Photon count arrival times were recorded with a time-to-digital converter (Time Tagger 20, Swabian Instruments) and accumulated in 10 ps bins.

The instrument response function (IRF) was 188 ps, as determined by recording laser scatter from solvent in a 1 cm path length cuvette. All data were collected at room temperature (20 °C) using the magic angle condition.

Calculation of FRET distance between DPH and Nile red in sc-apCC-4-DPH-1-sc-apCC-4-NR-1

Lifetime decays after the direct excitation of DPH and Nile red were modelled as a mono- or bi-exponential decay convolved with a Gaussian IRF function.

The energy transfer from the donor (DPH) was modelled as a bi-exponential decay accounting for the population of DPH (A_2) that are not involved in FRET:

$$(3) \quad I_{DPH} = A_1 e^{-(1/\tau_{FRET}+1/\tau_{DPH})t} + A_2 e^{-t/\tau_{DPH}}$$

The energy transfer to the acceptor (Nile red) was modelled as a bi-exponential function where B_1 is a negative amplitude reflecting the rise of acceptor fluorescence due to FRET:

$$(4) \quad I_{NR} = B_1 e^{-t/\tau_{FRET}} + B_2 e^{-t/\tau_{NR}}$$

These were fitted to DPH and Nile red channels simultaneously while floating the exponential pre-factors (A_1, A_2, B_1, B_2) and k_{FRET}. Average k_{DPH} and k_{NR} were kept fixed based on control measurements of the dyes alone within the protein. For readability, the equations shown above exclude the convolution terms required to model the instrument response function.

Assuming $1/\tau_{FRET} = k_{FRET}$, the Förster distance was calculated as:

$$(5) \quad R^6 = 8.785\ x\ 10^{-5} \frac{\phi_{DPH}}{\tau_{DPH}} \frac{1}{k_{FRET}} \frac{\kappa^2 I}{n^4}$$

where Φ_{DPH} and τ_{DPH} are the fluorescence quantum yield and the lifetime of DPH inside the protein, κ is the orientation factor between the dipoles, I is the Förster spectral overlap integral

of DPH fluorescence and Nile red absorption within the αHB, and n is the refractive index of the buffer (1.33). κ^2 was derived from the molecular dynamics simulations. DPH fluorescence quantum yields within the protein were estimated by reference quantum yield measurements of DPH in acetonitrile (Φ_{MeCN} = 0.17) as previously described[25]. The Förster spectral overlap integral was obtained using software freely available at FluorTools.com. The data were collected using 20 μM of each ligand, 30 μM protein, PBS, 10% v/v MeCN, pH 7.4.

Nile red and Nile blue fluorescence quantum yield measurements in sc-apCC-4-SN38-1

Absolute fluorescence quantum yields were measured using an integrating sphere on a FS5 spectrofluorometer (Edinburgh Instruments) with λ_{ex} = 510 nm, λ_{em} = 485 – 870 nm for Nile red and λ_{ex} = 530 nm, λ_{em} = 510 – 870 nm for Nile blue. The data were collected at 2 μM of each ligand, 50 μM protein, PBS, 10% v/v MeCN, pH 7.4. The quantum yield for Nile blue was corrected for self-absorption. For Nile red, the self-absorption was less than the error.

Transfection and imaging of HeLa cells

The linear DNA fragments for the designer constructs were synthesized by GeneArt or IDT and then subcloned in pTwist CMV vector (Twist Bioscience) for expression in HeLa cells. When required, PCR reactions were carried out using Q5 High Fidelity Hot Start DNA Polymerase (New England Biolabs (NEB)), following manufacturers' instructions. PCR products were purified using Monarch® PCR & DNA Cleanup Kit (NEB). Restriction digest reactions were carried out using NEB restriction enzymes at 37 °C for 1 h. Ligation reactions were carried out using T4 DNA Ligase and Rapid ligation buffer (Promega) for 1-2 h at 16 °C. Transformation was carried out in competent (subcloning efficiency) *Escherichia coli* cells (NEB® 5-alpha from NEB or DH5α from Invitrogen) for 30 min on ice, followed by heat

shocking (30 s, 42 °C), and a further 5 min on ice. Cells were recovered by incubating in LB media at 37 °C for 1 h and then plated on LB-agar plates containing the ampicillin.

Plasmid DNA samples for transfection were purified using Plasmid Midi Kit according to the manufacturer's instructions (QIAGEN) and verified by DNA sequencing using Rapid Sequencing Service provided by Source Bioscience.

HeLa cells (ECACC, UK Health Security Agency) were maintained in high glucose Dulbecco's Modified Eagle's Medium with 10% (v/v) foetal calf serum (Sigma-Aldrich) and 5% penicillin/streptomycin (PAA) (herein referred to as DMEM) without phenol red at 37 °C and 5% CO_2. For transfection, cells were seeded on tissue culture treated 35 mm CELLview cell culture dishes with 10 mm glass bottom (Greiner) at a density of $1x10^5$ cells per well and incubated at 37 °C and 5% CO2 for 16 h prior to transfection. Cells were transfected with 0.8 µg of the indicated plasmid DNA using Effectene transfection reagent according to the manufacturer's instructions (QIAGEN). After transfection cells were incubated at 37 °C, 5% CO_2 for 6 hours, washed with fresh DMEM and incubated further for 18 hours. Before imaging, cells were washed with PBS, stained with Hoechst 33342 (1 µg/mL) and Nile Red (0.1 µM) at 37 °C for 30 minutes, washed again, and then the medium was replaced for fresh DMEM.

Confocal images were collected using an Olympus IXplore SpinSR system with a 60x objective lens at 37 °C using the following imaging channels (λ_{ex}-max λ_{em}/width λ_{em}): 405-447/60, 488-525/50, and 561-617/73 nm. Figures were assembled using the Fiji distribution of ImageJ2[63].

Data availability

The coordinate and structure factor files for sc-apCC-4-DPH-2, sc-apCC-4-NR-1, sc-apCC-4-LMF-1, sc-apCC-4-A7F-1, sc-apCC-4-SN38-1 and sc-apCC-6-DPH-NR have been deposited in the PDB with accession codes 9R1N, 9R1J, 9R1M, 9R1K, 9R1L, 9R1K and 9R1O, respectively. Data from this study are openly available in Zenodo at DOI: 10.5281/zenodo.15333152.

Code availability

The code to run the RASSCCoL protocol can be found in the Woolfson group GitHub (https://github.com/woolfson-group/RASSCCoL).

References

1 Korendovych, I. V. & DeGrado, W. F. De novo protein design, a retrospective. *Q. Rev. Biophys.* **53**, e3 (2020). https://doi.org/10.1017/s0033583519000131

2 Woolfson, D. N. A Brief History of De Novo Protein Design: Minimal, Rational, and Computational. *J. Mol. Biol.* **433**, 167160 (2021). https://doi.org/10.1016/j.jmb.2021.167160

3 Kortemme, T. De novo protein design-From new structures to programmable functions. *Cell* **187**, 526-544 (2024). https://doi.org/10.1016/j.cell.2023.12.028

4 Albanese, K. I., Barbe, S., Tagami, S., Woolfson, D. N. & Schiex, T. Computational protein design. *Nat. Rev. Methods Primers* **5**, 13 (2025). https://doi.org/10.1038/s43586-025-00383-1

5 Robertson, D. E. *et al.* Design and synthesis of multi-haem proteins. *Nature* **368**, 425-432 (1994). https://doi.org/10.1038/368425a0

6 Feldmeier, K. & Höcker, B. Computational protein design of ligand binding and catalysis. *Curr. Opin. Chem. Biol.* **17**, 929-933 (2013). https://doi.org/10.1016/j.cbpa.2013.10.002

7 Tinberg, C. E. *et al.* Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* **501**, 212-216 (2013). https://doi.org/10.1038/nature12443

8 Lucas, J. E. & Kortemme, T. New computational protein design methods for de novo small molecule binding sites. *PLoS Comput. Biol.* **16**, e1008178 (2020). https://doi.org/10.1371/journal.pcbi.1008178

9 Polizzi, N. F. *et al.* De novo design of a hyperstable non-natural protein–ligand complex with sub-Å accuracy. *Nat. Chem.* **9**, 1157-1164 (2017). https://doi.org/10.1038/nchem.2846

10 Lu, L. *et al.* De novo design of drug-binding proteins with predictable binding energy and specificity. *Science* **384**, 106-112 (2024). https://doi.org/doi:10.1126/science.adl5364

11 Polizzi, N. F. & DeGrado, W. F. A defined structural unit enables de novo design of small-molecule–binding proteins. *Science* **369**, 1227-1233 (2020). https://doi.org/doi:10.1126/science.abb8330

12 Herud-Sikimic, O. *et al.* A biosensor for the direct visualization of auxin. *Nature* **592**, 768-772 (2021). https://doi.org/10.1038/s41586-021-03425-2

13 An, L. *et al.* Binding and sensing diverse small molecules using shape-complementary pseudocycles. *Science* **385**, 276-282 (2024). https://doi.org/doi:10.1126/science.adn3780

14 Hecht, M. H. *et al.* De novo heme proteins from designed combinatorial libraries. *Protein Sci.* **6**, 2512-2524 (1997). https://doi.org/https://doi.org/10.1002/pro.5560061204

15 Bender, G. M. *et al.* De Novo Design of a Single-Chain Diphenylporphyrin Metalloprotein. *J. Am. Chem. Soc.* **129**, 10732-10740 (2007). https://doi.org/10.1021/ja071199j

16 Koder, R. L. *et al.* Design and engineering of an O2 transport protein. *Nature* **458**, 305-309 (2009). https://doi.org/10.1038/nature07841

17 Krishna, R. *et al.* Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* **384**, eadl2528 (2024). https://doi.org/doi:10.1126/science.adl2528

18 Ertelt, M., Moretti, R., Meiler, J. & Schoeder, C. T. Self-supervised machine learning methods for protein design improve sampling but not the identification of high-fitness variants. *Sci. Adv.* **11**, eadr7338 (2025). https://doi.org/10.1126/sciadv.adr7338

19 Eberhardt, J., Santos-Martins, D., Tillack, A. F. & Forli, S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J. Chem. Inf. Model.* **61**, 3891-3898 (2021). https://doi.org/10.1021/acs.jcim.1c00203

20 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021). https://doi.org/10.1038/s41586-021-03819-2

21 Naudin, E. A. *et al.* From peptides to proteins: coiled-coil tetramers to single-chain 4-helix bundles. *Chem. Sci.* **13**, 11330-11340 (2022). https://doi.org/10.1039/D2SC04479J

22 Albanese, K. I. *et al.* Rationally seeded computational protein design of α-helical barrels. *Nat. Chem. Biol.* **20**, 991-999 (2024). https://doi.org/10.1038/s41589-024-01642-0

23 Schweke, H. *et al.* An atlas of protein homo-oligomerization across domains of life. *Cell* **187**, 999-1010 e1015 (2024). https://doi.org/10.1016/j.cell.2024.01.022

24 Woolfson, D. N. Understanding a protein fold: The physics, chemistry, and biology of alpha-helical coiled coils. *J. Biol. Chem.* **299**, 104579 (2023). https://doi.org/10.1016/j.jbc.2023.104579

25 Petrenas, R. *et al.* Confinement and Catalysis within De Novo Designed Peptide Barrels. *J. Am. Chem. Soc.* **147**, 3796-3803 (2025). https://doi.org/10.1021/jacs.4c16633

26 Zhao, Y. H., Abraham, M. H. & Zissimos, A. M. Fast calculation of van der Waals volume as a sum of atomic and bond contributions and its application to drug compounds. *J. Org. Chem.* **68**, 7368-7373 (2003). https://doi.org/10.1021/jo034808o

27 Huang, X., Pearce, R. & Zhang, Y. FASPR: an open-source tool for fast and accurate protein side-chain packing. *Bioinform.* **36**, 3758-3765 (2020). https://doi.org/10.1093/bioinformatics/btaa234

28 Dauparas, J. *et al.* Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49-56 (2022). https://doi.org/10.1126/science.add2187

29 Dauparas, J. *et al.* Atomic context-conditioned protein sequence design using LigandMPNN. *Nat. Methods*, 717-723 (2025).

30 Desmet, J., De Maeyer, M., Hazes, B. & Lasters, I. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* **356**, 539-542 (1992). https://doi.org/10.1038/356539a0

31 Guillame-Bert, M., Bruch, S., Stotz, R. & Pfeifer, J. in *KDD '23: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* 4068-4077 (Association for Computing Machinery, 2023).

32 Sackett, D. L. & Wolff, J. Nile red as a polarity-sensitive fluorescent probe of hydrophobic protein surfaces. *Anal. Biochem.* **167**, 228-234 (1987). https://doi.org/https://doi.org/10.1016/0003-2697(87)90157-6

33 Gajo, C. *et al.* Nile Red Fluorescence: Where's the Twist? *J. Phys. Chem. B* **128**, 11768-11775 (2024). https://doi.org/10.1021/acs.jpcb.4c06048

34 Jain, B. & Das, K. Fluorescence resonance energy transfer between DPH and Nile Red in a lipid bilayer. *Chem. Phys. Lett.* **433**, 170-174 (2006). https://doi.org/https://doi.org/10.1016/j.cplett.2006.11.050

35 Schneidman-Duhovny, D., Hammel, M., Tainer, J. A. & Sali, A. FoXS, FoXSDock and MultiFoXS: Single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles. *Nucleic Acids Res.* **44**, W424-429 (2016). https://doi.org/10.1093/nar/gkw389

36 Cranfill, P. J. *et al.* Quantitative assessment of fluorescent proteins. *Nat. Methods* **13**, 557-562 (2016). https://doi.org/10.1038/Nmeth.3891

37 Chapman, S. *et al.* The photoreversible fluorescent protein iLOV outperforms GFP as a reporter of plant virus infection. *Proc. Natl. Acad. Sci. USA* **105**, 20038-20043 (2008). https://doi.org/10.1073/pnas.0807551105

38 Zhang, D. S. *et al.* Design of a palette of SNAP-tag mimics of fluorescent proteins and their use as cell reporters. *Cell Discov.* **9**, 56 (2023). https://doi.org/10.1038/s41421-023-00546-y

39 Dou, J. Y. *et al.* De novo design of a fluorescence-activating β-barrel. *Nature* **561**, 485-491 (2018). https://doi.org/10.1038/s41586-018-0509-0

40 Zhu, J. *et al.* De novo design of transmembrane fluorescence-activating proteins. *Nature* **640**, 249-257 (2025). https://doi.org/10.1038/s41586-025-08598-8

41 Shaner, N. C. *et al.* Improved monomeric red, orange and yellow fluorescent proteins derived from *Discocoma* sp. red fluorescent protein. *Nat. Biotechnol.* **22**, 1567-1572 (2004). https://doi.org/10.1038/nbt1037

42 Wang, H. *et al.* Near-infrared pH-sensitive probes based on aza-Nile Blue for detecting interactions between mitochondria and lysosomes. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **338**, 126169 (2025). https://doi.org/10.1016/j.saa.2025.126169

43 Beetz, C. *et al.* Identification of nuclear localisation sequences in spastin (SPG4) using a novel Tetra-GFP reporter system. *Biochem. Biophys. Res. Commun.* **318**, 1079-1084 (2004). https://doi.org/10.1016/j.bbrc.2004.03.195

44 Liang, X. Q., Lu, Y., Wilkes, M., Neubert, T. A. & Resh, M. D. The N-terminal SH4 region of the Src family kinase Fyn is modified by methylation and heterogeneous fatty acylation - Role in membrane targeting, cell adhesion, and spreading. *J. Biol. Chem.* **279**, 8133-8139 (2004). https://doi.org/10.1074/jbc.M311180200

45 Riedl, J. *et al.* Lifeact: a versatile marker to visualize F-actin. *Nat. Methods* **5**, 605-607 (2008). https://doi.org/10.1038/nmeth.1220

46 Qi, Q. R., Tian, H., Yue, B. S., Zhai, B. T. & Zhao, F. Research Progress of SN38 Drug Delivery System in Cancer Treatment. *Int. J. Nanomed.* **19**, 945-964 (2024). https://doi.org/10.2147/Ijn.S435407

47 Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **13**, e1005659 (2017). https://doi.org/10.1371/journal.pcbi.1005659

48 Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1-2**, 19-25 (2015). https://doi.org/https://doi.org/10.1016/j.softx.2015.06.001

49 Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.* **32**, 2319-2327 (2011). https://doi.org/10.1002/jcc.21787

50 Schuck, P. Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and lamm equation modeling. *Biophys. J.* **78**, 1606-1619 (2000). https://doi.org/10.1016/s0006-3495(00)76713-0

51 Förster, S., Apostol, L. & Bras, W. Scatter: software for the analysis of nano- and mesoscale small-angle scattering. *J. Appl. Crystallogr.* **43**, 639-646 (2010). https://doi.org/https://doi.org/10.1107/S0021889810008289

52 Manalastas-Cantos, K. *et al.* ATSAS 3.0: expanded functionality and new tools for small-angle scattering data analysis. *J. Appl. Crystallogr.* **54**, 343-355 (2021). https://doi.org/10.1107/S1600576720013412

53 Schneidman-Duhovny, D., Hammel, M., Tainer, J. A. & Sali, A. Accurate SAXS profile computation and its assessment by contrast variation experiments. *Biophys. J.* **105**, 962-974 (2013). https://doi.org/10.1016/j.bpj.2013.07.020

54 Winter, G. xia2: an expert system for macromolecular crystallography data reduction. *J. Appl. Crystallogr.* **43**, 186-190 (2010). https://doi.org/https://doi.org/10.1107/S0021889809045701

55 Winter, G. *et al.* DIALS: implementation and evaluation of a new integration package. *Acta Crystallogr. D Biol. Crystallogr.* **74**, 85-97 (2018). https://doi.org/10.1107/S2059798317017235

56 Evans, P. R. & Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallogr. D Biol. Crystallogr.* **69**, 1204-1214 (2013). https://doi.org/10.1107/S0907444913000061

57 Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 235-242 (2011). https://doi.org/10.1107/S0907444910045749

58 Vonrhein, C. *et al.* Data processing and analysis with the autoPROC toolbox. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 293-302 (2011). https://doi.org/10.1107/S0907444911007773

59 McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658-674 (2007). https://doi.org/10.1107/S0021889807021206

60 Casanal, A., Lohkamp, B. & Emsley, P. Current developments in Coot for macromolecular model building of Electron Cryo-microscopy and Crystallographic Data. *Protein Sci.* **29**, 1069-1078 (2020). https://doi.org/10.1002/pro.3791

61 Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213-221 (2010). https://doi.org/10.1107/S0907444909052925

62 Polak, D. W., Hannon, A. D. P., Marczak Giorio, G. A., Hawkins, O. A. & Oliver, T. A. A. The Solvent-Dependent Photophysics of Diphenyloctatetraene. *J. Phys. Chem. B* **127**, 8199-8207 (2023). https://doi.org/10.1021/acs.jpcb.3c03737

63 Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676-682 (2012). https://doi.org/10.1038/Nmeth.2019

Acknowledgments

R.P. is supported by a BBSRC-funded PhD studentship and by Rosa Biotech through the South West Biosciences Doctoral Training Partnership (BB/T008741/1). K.O. is supported by a BBSRC grant to N.S.S. and D.N.W. (BB/X003027/1). J.J.C. was supported by an Engineering and Physical Sciences Research Council (EPSRC) program grant to G.J.L. and D.N.W. (EP/T012455/1). A.V.R. is supported by a Leverhulme Trust grant to J.J.M. and D.N.W. (RGP-2021-049). T.A.A.O. thanks the Royal Society for the award of a URF (URF\R\201007) that enabled TCSPC measurements.

T.A.A.O., D.N.W. and R.P. thank Olivia Hawkins and Camilla Gajo for assisting with the fluorescence lifetime measurements. We thank the Diamond Light Source for access to beamlines B21, I03 and I24 (proposals mx37593 and mx31440) and The European Synchrotron Radiation Facility for access to beamline BM29 (proposal mx2551). Finally, we thank the members of the Woolfson group and collaborators from the Manchester Institute of Biotechnology, particularly Dr Derren Heyes, and the University of Sheffield for many valuable discussions.

Authors contributions

D.N.W., J.J.C., K.O., and R.P. conceived the project. J.J.C., K.O., and R.P. wrote and tested the RASSCCoL code. J.J.C. and K.O. designed and characterised the lumiflavin and SN38 binders. R.P. designed and characterised the DPH and Nile red binders. R.P. and K.O. designed the arcyriaflavin A binders. R.P. collected the X-ray diffraction data and solved the crystal

structures. A.V.R. devised and performed cell experiments. J.J.C, K.O., R.P., A.V.R. and D.N.W. wrote the paper, and all authors contributed to its revision. J.J.M., G.J.L., N.S.S., and D.N.W. raised the funding.

Competing interests

Authors have no competing interests.

Supplementary information

Supplementary Tables 1–8, Supplementary Figs. 1–43 and references.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- RASSCCoLNatureChemBiolDec2025SI.pdf