

Targeting protein–ligand neosurfaces with a generalizable deep learning tool

<https://doi.org/10.1038/s41586-024-08435-4>

Received: 7 June 2024

Accepted: 20 November 2024

Published online: 15 January 2025

Open access

 Check for updates

Anthony Marchand^{1,8}, Stephen Buckley^{1,8}, Arne Schneuing^{1,8}, Martin Pacesa¹, Maddalena Elia¹, Pablo Gainza^{1,7}, Evgenia Elizarova¹, Rebecca M. Neeser^{2–5}, Pao-Wan Lee³, Luc Reymond⁴, Yangyang Miao¹, Leo Scheller¹, Sandrine Georgeon¹, Joseph Schmidt¹, Philippe Schwaller², Sebastian J. Maerkl³, Michael Bronstein^{5,6} & Bruno E. Correia¹✉

Molecular recognition events between proteins drive biological processes in living systems¹. However, higher levels of mechanistic regulation have emerged, in which protein–protein interactions are conditioned to small molecules^{2–5}. Despite recent advances, computational tools for the design of new chemically induced protein interactions have remained a challenging task for the field^{6,7}. Here we present a computational strategy for the design of proteins that target neosurfaces, that is, surfaces arising from protein–ligand complexes. To develop this strategy, we leveraged a geometric deep learning approach based on learned molecular surface representations^{8,9} and experimentally validated binders against three drug-bound protein complexes: Bcl2–venetoclax, DB3–progesterone and PDF1–actinonin. All binders demonstrated high affinities and accurate specificities, as assessed by mutational and structural characterization. Remarkably, surface fingerprints previously trained only on proteins could be applied to neosurfaces induced by interactions with small molecules, providing a powerful demonstration of generalizability that is uncommon in other deep learning approaches. We anticipate that such designed chemically induced protein interactions will have the potential to expand the sensing repertoire and the assembly of new synthetic pathways in engineered cells for innovative drug-controlled cell-based therapies¹⁰.

Protein–protein interactions (PPIs) have essential roles in healthy cell homeostasis but are also involved in numerous diseases¹¹. For this reason, several therapies targeting PPIs have been developed over the past decades, and several computational tools for the design of new protein interactions have recently been proposed¹². The governing principles determining the propensity of proteins to form interactions are intricate owing to the interplay of several contributions, such as geometric and chemical complementarity, dynamics and solvent interactions. Therefore, it remains challenging to predict and design new PPIs, especially in the absence of evolutionary constraints. Native PPIs can also be controlled by regulatory layers such as allostery², posttranslational modifications³ or direct ligand binding^{4,5}. Compound-bound surfaces, which we refer to as neosurfaces, are among the most fascinating and challenging molecular recognition instances, as relatively minor changes at the protein binding site can have a large impact on binding affinities. Interest in such interactions has been fuelled by the development of new drug modalities, specifically, molecular glues that form neosurfaces to trigger protein interactions for degradation and other applications^{13,14} and thus represent a promising route for the development of innovative therapeutics.

In synthetic biology, molecular components that rely on small-molecule-induced neosurfaces have been used to engineer chemically responsive systems with precise spatiotemporal control of cellular activities¹⁵. Small-molecule triggers have been used to both induce and disrupt PPIs, thereby functioning as ON or OFF switches for engineered cellular functions^{10,15,16}. There are several practical advantages to using small molecules as triggers, including their simple administration, biodistribution, cell permeability, safety, and high affinity and specificity to their target proteins. Protein-based switches controlled by small molecules have already been used to regulate transcription¹⁷, protein degradation^{18,19} and protein localization^{20,21}, among many other applications. In addition to their use in basic research, engineering molecular switches are increasingly used to control protein-based and cellular therapeutics, the activity of which may need to be regulated to mitigate potentially dangerous side effects^{10,22,23}. Although several chemically disruptable heterodimer (OFF-switch) systems have been proposed^{10,15,22}, computationally designed chemically induced dimerization (CID, ON-switch) systems remain challenging owing to the complexity of modelling neosurfaces. Previous attempts to design CID systems primarily relied on experimental methods^{15,17,24–26}, and,

¹Laboratory of Protein Design and Immunoengineering, Institute of Bioengineering, Ecole polytechnique fédérale de Lausanne, Lausanne, Switzerland. ²Laboratory of Chemical Artificial Intelligence, Institute of Chemical Sciences and Engineering, Ecole polytechnique fédérale de Lausanne, Lausanne, Switzerland. ³Laboratory of Biological Network Characterization, Institute of Bioengineering, Ecole polytechnique fédérale de Lausanne, Lausanne, Switzerland. ⁴Biomolecular Screening Core Facility, School of Life Sciences, Ecole polytechnique fédérale de Lausanne, Lausanne, Switzerland. ⁵Department of Computer Science, University of Oxford, Oxford, UK. ⁶Aithyra Research Institute for Biomedical Artificial Intelligence, Austrian Academy of Sciences, Vienna, Austria. ⁷Present address: Monte Rosa Therapeutics, Boston, MA, USA. ⁸These authors contributed equally: Anthony Marchand, Stephen Buckley, Arne Schneuing.

✉e-mail: bruno.correia@epfl.ch

despite the emergence of artificial intelligence and numerous computational tools, only a few tools can generalize to both proteins and small molecules as targets for protein design; this has resulted in a lack of suitable approaches for design of new chemically induced PPIs. Computational methods to design new CIDs have mostly relied on transplanting an existing drug-binding site to a known heterodimer interface⁷ or using docking of putative pre-existing proteins (that is, scaffolds) followed by interface optimization⁶. However, these approaches can have limitations such as the risk of drug-independent dimerization, a lack of suitable scaffold proteins for design or extensive need for in vitro maturation techniques.

We recently reported a geometric deep-learning-based framework called MaSIF (molecular surface interaction fingerprinting)⁸ for the study of protein surface features and design of new PPIs⁹. In this study, we aimed to test whether our surface-centric approach could generalize to non-protein ligands without further training data using a higher-level representation, namely the geometric and chemical features found on the molecular surface. To do so, we designed site-specific binders that target neosurfaces composed of a small-molecule ligand and protein surface moieties, resulting in de novo ligand-dependent protein interactions. Although state-of-the-art tools showed good performance in the prediction and design of ligand–protein interactions²⁷, they were not suitable for the design of de novo ternary complexes, which are particularly challenging owing to the scarcity of data. Here we successfully designed and characterized new drug-inducible protein binders recognizing the B-cell lymphoma 2 (Bcl2) protein in complex with clinically approved inhibitor venetoclax²⁸, progesterone-binding antibody DB3 in complex with its ligand²⁹ and, finally, peptide deformylase 1 (PDF1) protein from *Pseudomonas aeruginosa* in complex with antibiotic actinonin³⁰. Last, we show that such ligand-controlled systems can be used in both in vitro and cellular contexts for a range of synthetic biology applications, unlocking possibilities for the development and regulation of new therapeutic approaches.

Neosurface features captured by MaSIF

Within our geometric deep learning framework MaSIF⁸, we previously developed two applications: (1) MaSIF-site, to accurately predict regions of a protein surface with a high propensity to form an interface with another protein; and (2) MaSIF-search, to rapidly find and dock protein partners on the basis of complementary surface patches. In MaSIF-search, we extract surface patch descriptors (fingerprints), so that patches with complementary geometry and chemistry have similar fingerprints, whereas non-interacting patches have low fingerprint similarity. Surface fingerprints enable an initial ultrafast search in an alignment-free manner using the Euclidean distances between them. Patches with fingerprint distances below a threshold are then further aligned in three dimensions and scored with an interface postalignment (IPA) score to refine the selection.

In its initial conception, MaSIF only considered canonical amino acids as part of the protein molecular surface and was not compatible with small molecules, glycans or other ligands. Thus, we present here MaSIF-neosurf, which incorporates small molecules as part of the molecular surface representation of the target protein to predict interfaces and partners on the basis of neosurface fingerprints (Fig. 1a and Methods). MaSIF was initially trained to operate on general chemical and geometric surface properties of biomolecules, while abstracting the underlying structure. Thus, it is not restricted to protein surfaces but should in principle also capture the surface patterns arising from non-protein surfaces. Upon generation of the molecular surface of the protein–ligand complex, MaSIF-neosurf computes two geometric features: shape index³¹ and distance-dependent curvature³². In addition, three chemical features are used: Poisson–Boltzmann electrostatics, which can be computed directly from the small molecule; and hydrogen bond donor/acceptor propensity³³ and hydrophobicity^{34–36}, for

which we developed new featurizers tailored to capture the chemical properties of the small molecules (Methods and Supplementary Fig. 1).

To assess the capabilities of MaSIF-neosurf, we benchmarked its performance on several ternary complexes whose interfaces are composed of protein and ligand surfaces. We aimed to recover known binding partners for proteins with small molecules at the binding interface. We assembled a list of 14 ligand-induced protein complexes, then split each of the complexes into two subunits, resulting in 28 independent benchmarking cases, and processed them with and without the small molecule bound. The ligand-free protein surfaces, together with 8,879 decoy proteins involved in PPIs, constituted our database, which we queried with surface patches from all 28 protein–ligand complexes. As each of the 8,907 protein candidates was decomposed into almost 4,000 patches on average, the database represented a large search space with more than 35 million potential binding sites. We then evaluated whether the correct binding partner was retrieved and docked in the correct rigid-body orientation. When considering the protein–ligand complex as a docking partner, MaSIF-neosurf recovered more than 70% (20) of the correct binding partners and their binding poses (Fig. 1b), whereas the state-of-art RoseTTAFold All-Atom²⁷ recovered only 14% (4) of correct binding poses (Supplementary Fig. 2). Only a small subset of test cases could be recovered in the absence of the ligand; the general trend was that in such cases the protein surface was a large contributor to the overall protein interaction (Supplementary Fig. 3). The ability to capture neosurface properties was further supported by an increased descriptor distance score between interacting partners (that is, increased complementarity between interacting fingerprints) and an increased IPA score (Methods) in the presence of the small molecule compared with the case without (Fig. 1c,d). Although both geometric and chemical input features have been shown to contribute to MaSIF's performance⁸, ablating individual inputs did not seem to have a substantial effect (Supplementary Fig. 4), probably owing to a certain degree of redundancy within features. Overall, MaSIF-neosurf captured, in many instances, factors that were determinant of ligand-mediated protein interactions. To further test its capabilities, we sought to de novo design interactions of this type.

Designing new ligand-induced PPIs

Recently, we proposed the MaSIF-seed pipeline for the design of de novo site-specific protein binders⁹. Given the performance of MaSIF-seed against several therapeutically relevant targets, we sought to determine whether such an approach could generalize to design site-specific binders to neosurfaces composed of ligand and protein atoms. By doing so, we would tackle the challenge of designing chemically controlled protein interactions and test our understanding of molecular recognition events mediated by neosurfaces. We therefore adapted our MaSIF-seed pipeline to the newly developed MaSIF-neosurf framework (Fig. 2a). When neosurfaces had been computed for a given protein–ligand complex, we first used MaSIF-site to identify the regions most likely to become buried in an interface. Then, an extensive fingerprint search was used to identify complementary structural motifs (binding seeds) from a database of approximately 640,000 structural fragments (402 million surface patches/fingerprints). Therefore, by focusing on the predicted buried regions of the interface and searching for highly complementary motifs, we could quickly reduce the vast space of patches and binding motifs to the most promising candidates. Finally, the top seeds were refined by sequence optimization and grafted with Rosetta³⁷ on recipient proteins (scaffolds) to stabilize the binding motif. Last, a final round of sequence design was performed to improve atomic contacts at the interface.

We designed ligand-dependent protein binders targeting ligand-bound proteins from different families: Bcl2 in complex with clinically approved drug venetoclax, antiprogestosterone antibody DB3 in complex with its ligand and PDF1 from *P. aeruginosa* in complex

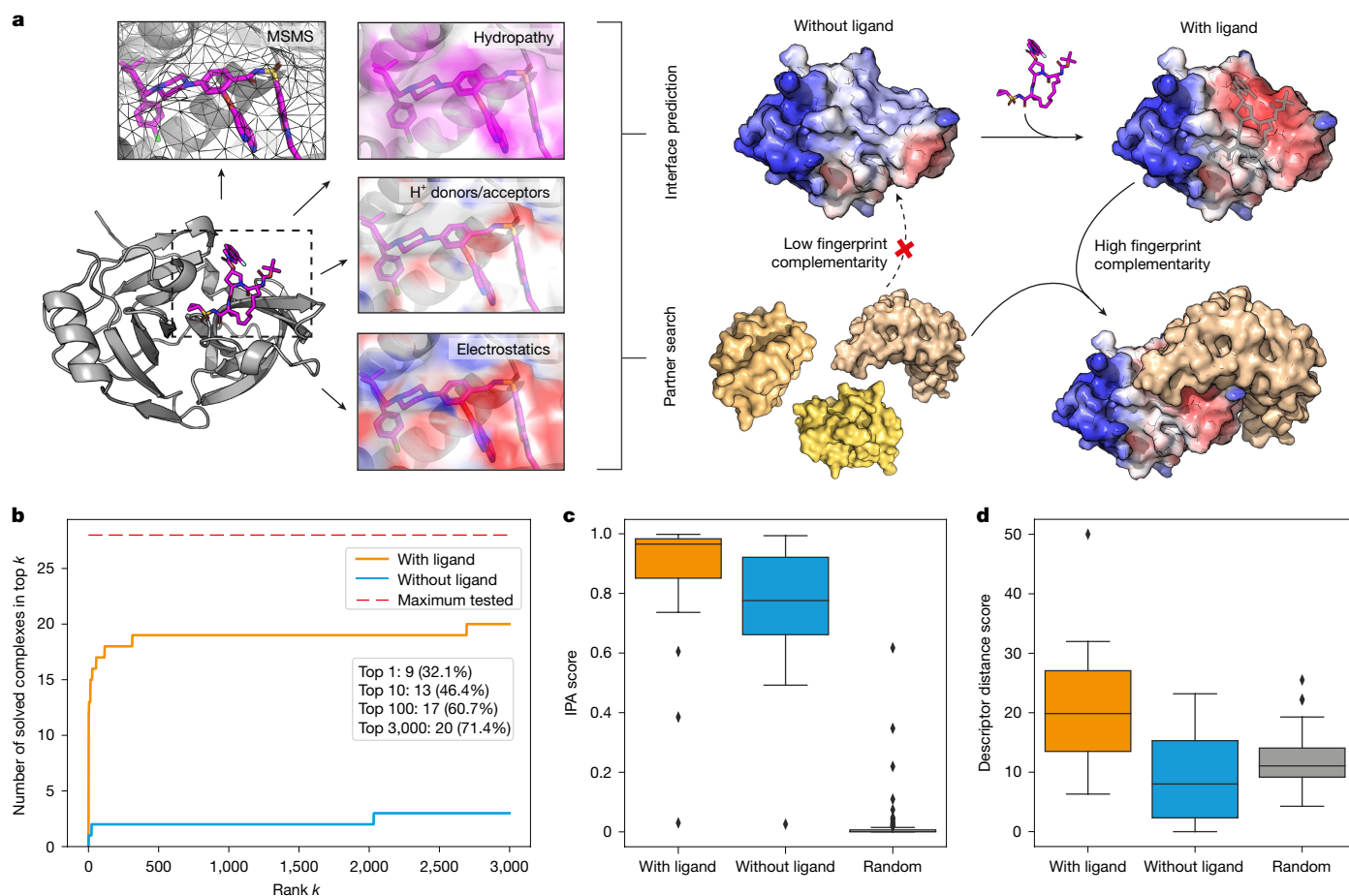


Fig. 1 | Neosurface properties are captured to identify interface sites and binding partners. **a**, Geometric and chemical features of the ligand–protein complexes are computed, including the molecular surface representation (MSMS), hydropathy score, proton donors/acceptors and Poisson–Boltzmann electrostatics. Surface features are vectorized in a descriptor (also referred to as a fingerprint) and used by MaSIF-neosurf for interface propensity prediction or protein partner search. The ligand-containing fingerprint is then used to find complementary fingerprints in a patch database. **b**, Ranking predictions using MaSIF-neosurf on a benchmark dataset of known ternary complexes and

a set of 8,879 decoys. Complementary partner search was performed in the presence (orange) and absence (blue) of the respective small-molecule ligand. **c,d**, IPA scores (**c**) and descriptor distance scores (**d**; Methods) of the interacting complexes in the presence (orange) and absence (blue) of the small molecule compared with a set of random patch alignments (grey). Boxes represent quartiles, and whiskers show data points within 1.5× the interquartile range. Data outside this range are shown as flier points. Twenty-eight complexes are plotted with ligand and 26 without ligand, and there are 74 (**c**) and 104 (**d**) random alignments.

with antibiotic actinonin (Fig. 2b). We first identified a moderate to high interface propensity of these neosurfaces with MaSIF-neosurf, selected one to three relevant interface patches depending on the solvent-accessible surface area exposed by the ligand (Fig. 2b) and searched for complementary fingerprints in our seed database. Top-ranking seeds were selected (around 100–120 for each target), refined and grafted on to multiple recipient scaffolds, and approximately 2,000 final designs per target complex were selected with computational filters (Fig. 2c, Supplementary Table 1 and Methods). Our pipeline generated designs with diverse helical and β -sheet-based binding motifs, as well as various protein folds; it thus sampled a wide space of sequences and topologies (Fig. 2c). All selected designs were predicted to favourably engage the neosurface by showing increased interface structural metrics (such as the predicted binding energy, buried surface area and number of atomic contacts) in the presence of the ligand (Supplementary Fig. 5).

Experimental validation of designed CIDs

The computational designs were screened by yeast display³⁸, and, after two rounds of fluorescence-activated cell sorting, enriched clones were deep sequenced (Supplementary Fig. 6 and Supplementary Table 2).

We show one binder targeting each of the selected test cases in Fig. 3a. The best designs showed no binding in the absence of the corresponding small molecules, whereas modest to high binding signals were observed for the ligands in yeast display experiments (Fig. 3b). These changes in binding signal upon addition of small molecules are consistent with the expected behaviour of a chemically induced PPI. Notably, small molecules contributed about 10–12% of the predicted target buried surface area, but they improved the predicted binding energy ($\Delta\Delta G$) of the interface compared with the ligand-unbound form by 17.0–27.7%. This result demonstrates the small yet critical contribution of each ligand to the binding event and highlights the difficulty of the design problem (Supplementary Table 3).

Moreover, point mutants at the interface hotspot residues abrogated binding to the target complex, further supporting the designed binding mode (Fig. 3c). No binding was observed with the native scaffolds used for seed grafting and interface design, supporting the critical role of the interface design pipeline (Fig. 3c). Finally, specificity towards the desired ligand was confirmed using control compounds: S55746 for Bcl2, 19-*O*-benzoyl-progesterone (OBz-Pro) for DB3 IgG and tertbutyldimethylsilyl-actinonin (TBDMS-Act) for PDF1 (Fig. 3d and Supplementary Fig. 7). These analogues retained binding to the protein target (Supplementary Fig. 7). However, no binding to the designs was

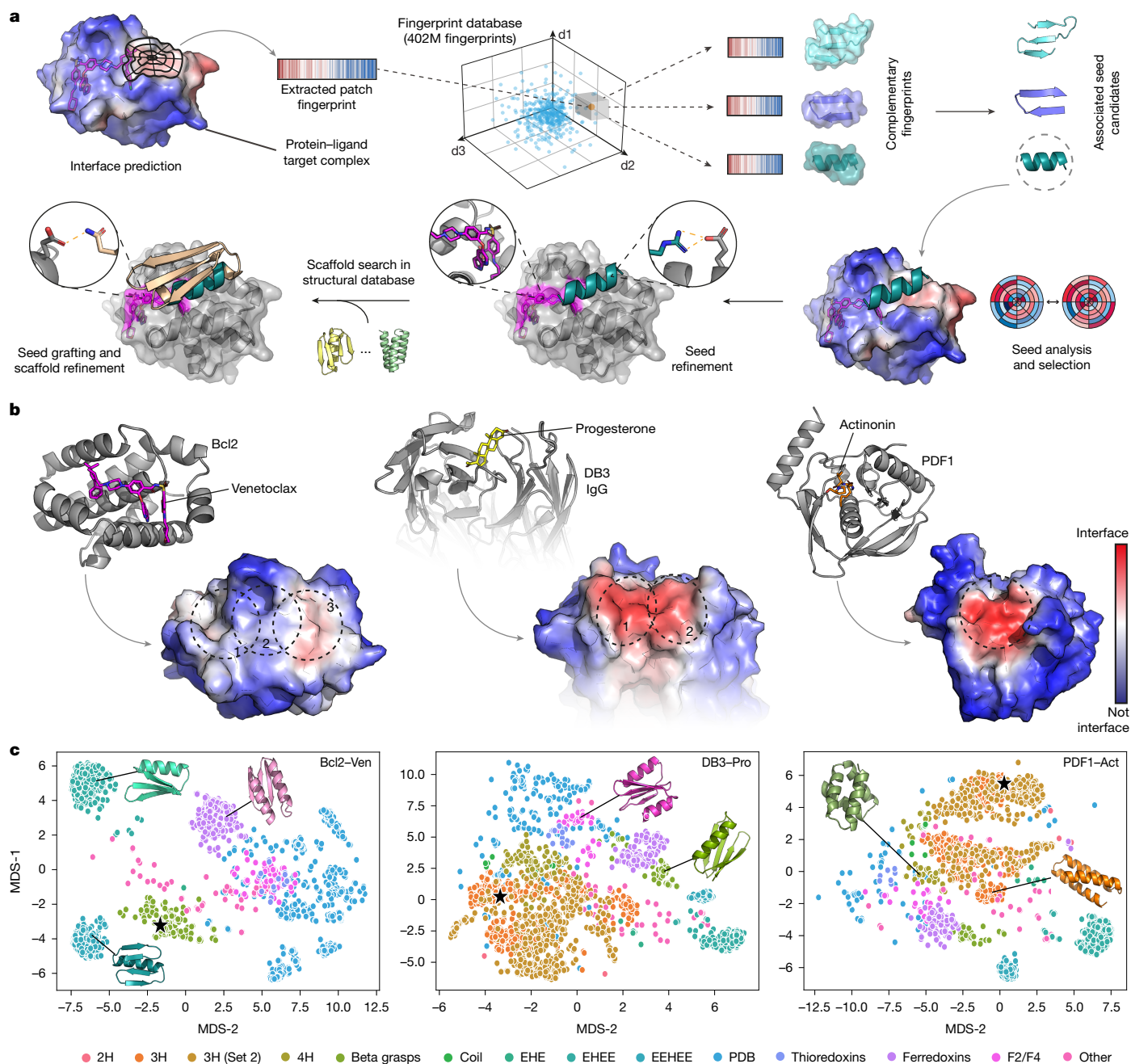


Fig. 2 | Design of ligand-induced protein interactions with MaSIF-neosurf.
a, To design new ligand-induced protein interactions, potential interface sites are first identified on the target protein-ligand complex. The corresponding patches are then used to find complementary fingerprints in a patch database. The top patches are aligned and scored to refine the selection. Associated binding motifs (seeds) undergo sequence optimization with an emphasis on designing new hydrogen bond networks with the target protein and small molecule. Seeds are then grafted on suitable scaffolds from a structural database, and the rest of the scaffold interface is redesigned using Rosetta. Finally, the top (approximately) 2,000 designs, according to different structural

metrics, are selected and screened experimentally. **b**, Target candidates in complex with their respective small molecules (top row). Neosurfaces with their protein binding propensities (bottom row). Sites selected for binder design are highlighted with dashed circles. **c**, Diversity of the computational designs mapped using multidimensional scaling (MDS) of pairwise r.m.s.d. values between all designs. Experimentally confirmed binders are highlighted with a star. In total, 1,995 computational designs were plotted for Bcl2-venetoclax (Bcl2-Ven), 1,998 for DB3-progesterone (DB3-Pro) and 1,997 for PDF1-actinonin (PDF1-Act).

observed, confirming that the correct interface on the target complex was engaged with high ligand specificity (Fig. 3d).

Biochemical and structural validation

To map the binding site with high confidence and identify potential beneficial mutations, we performed a site-saturation mutagenesis

(SSM) study (Supplementary Fig. 8). To assess the effects of the different mutations over the designed ligand-dependent interaction, we computed the average enrichment score of each mutation when comparing binding versus non-binding populations on yeast display experiments, similar to other deep saturation mutagenesis studies^{39,40}. Globally, we observed that such interactions had exquisite sensitivity to single-point mutants and that residues with high sensitivity mapped

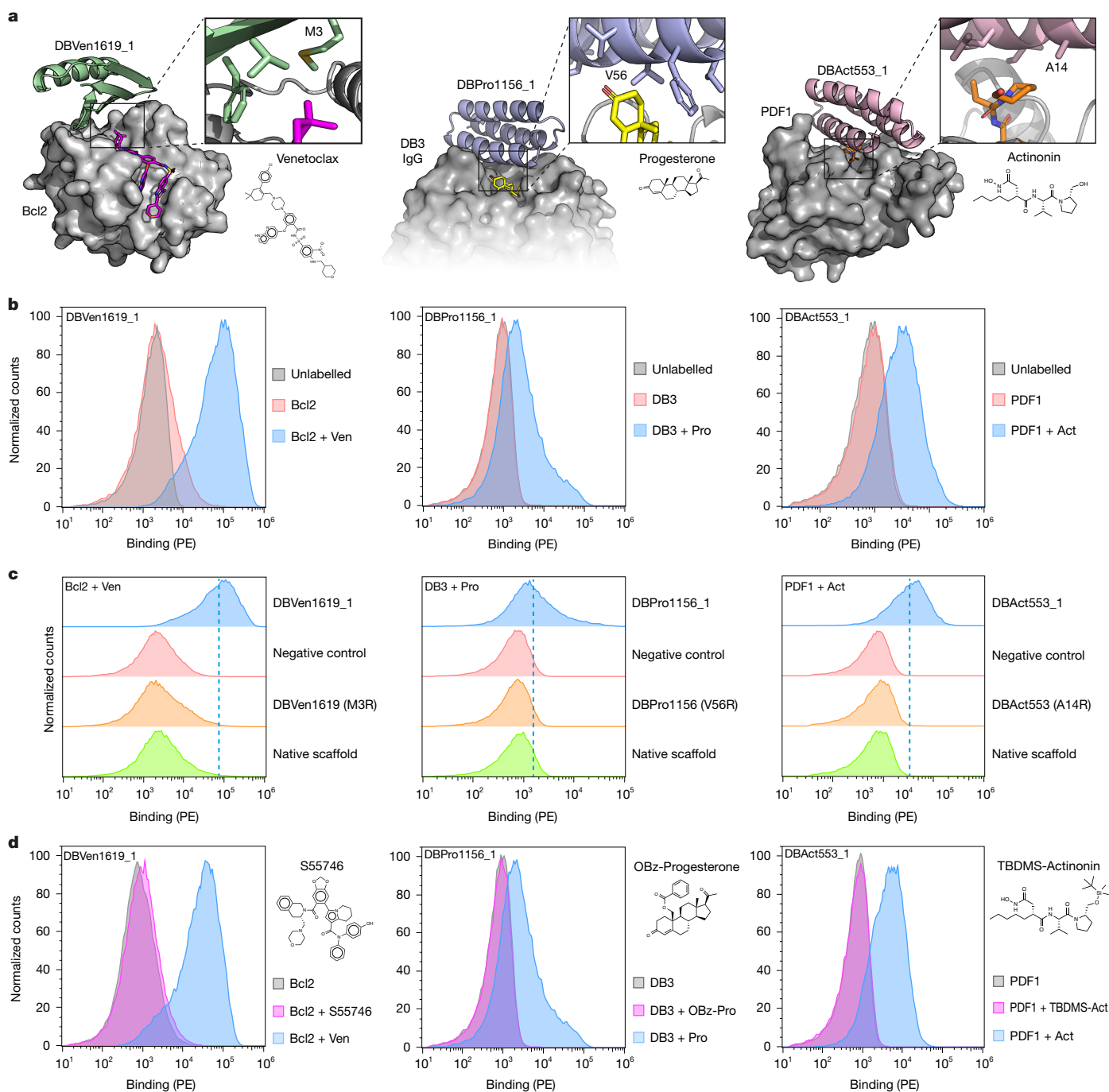


Fig. 3 | De novo design and screening of small-molecule-dependent binders. **a**, Models of the designed binders in complex with their respective target complexes: Bcl2–venetoclax, DB3–progesterone and PDF1–actinonin. **b**, Histograms of the binding signal (phycoerythrin; PE) measured by flow cytometry on yeast displaying the designed binders. Yeast were either unlabelled or labelled with 500 nM of the respective target protein preincubated with the ligand or with the target protein alone. **c**, Histograms of the binding signal (PE) measured by flow cytometry on yeast displaying designed binders, a mutated version with a single-point mutant at the predicted interface and the

starting scaffold used for the design process. Yeast cells were labelled with 500 nM of their respective ligand–protein complex. Dashed lines represent the geometric mean of the designed binder signal. **d**, Binding measured on yeast displaying DBVen1619_1, DBPro1156_1 or DBAct553_1 labelled with the target protein alone (grey), the target protein in complex with the original small molecule (blue) or the target protein in complex with the small-molecule analogue (magenta). Control analogues were S55746, OBz-Pro and TBDMS-Act. Detailed structures of the small molecules and their analogues are shown in Supplementary Fig. 7.

very closely to the designed interfaces, supporting the accuracy of our computational models (Fig. 4a).

The initial successful designs were expressed and purified for further biophysical characterization. All designs were monomeric, folded and highly stable in solution (Extended Data Fig. 1). All three designs showed binding affinities in the range of native transient DBPs⁴¹, from

mid-nanomolar to low micromolar, after pure in silico generation (Extended Data Fig. 2). Specifically, DBAct553_1 showed a binding affinity (K_D) of 542 nM, and DBVen1619_1 and DBPro1156_1 showed affinities of 4 μ M and greater than 10 μ M, respectively.

In the SSM scan, some mutations were associated with potential improvements in affinity (Supplementary Fig. 8 and Extended Data

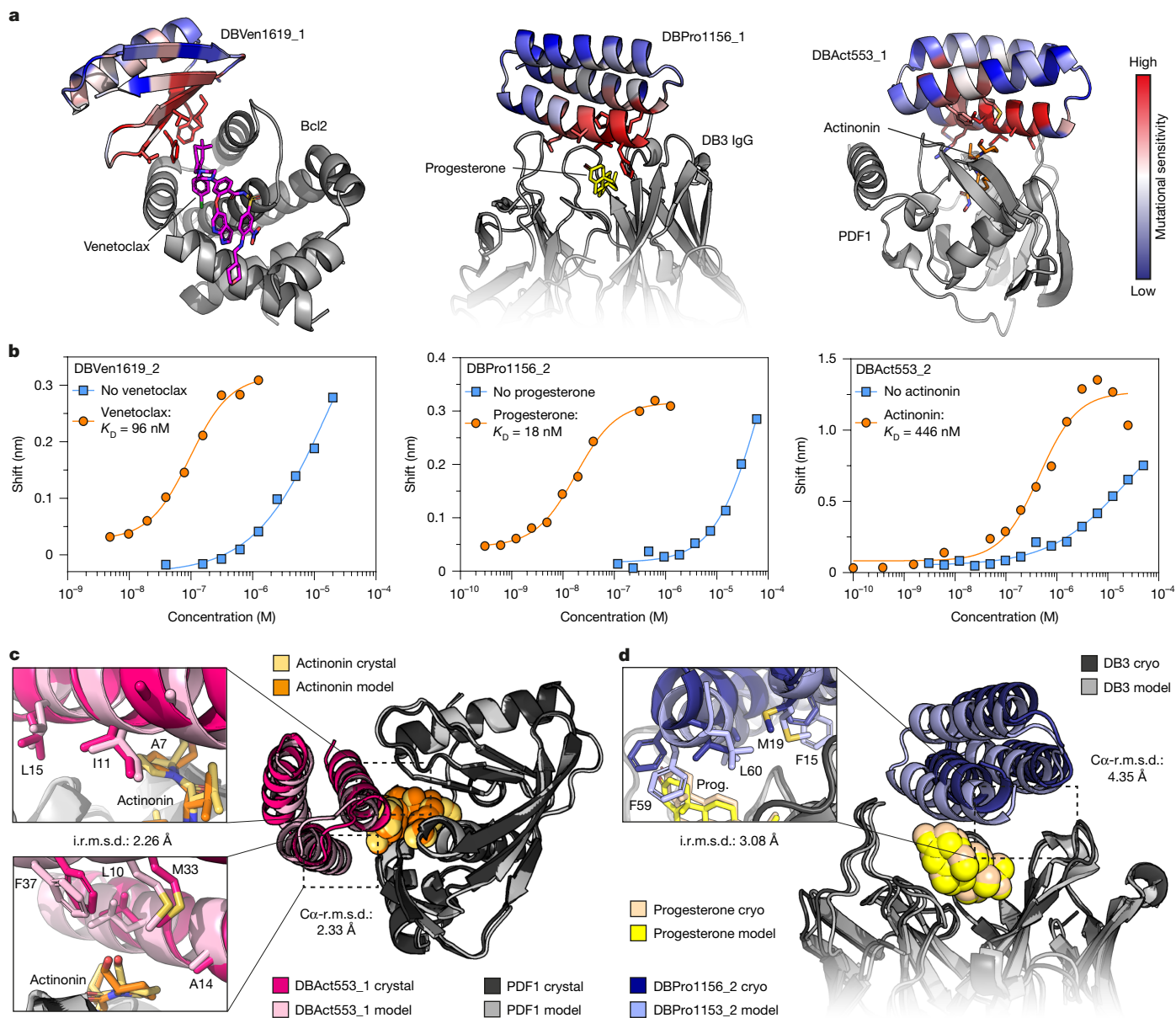


Fig. 4 | Binding mode, affinity and structure determination of the designed binders. **a**, Computational model coloured with the average enrichment score in the SSM for each amino acid position of the designed binder. Red indicates that an amino acid position is sensitive to mutations, whereas blue indicates a more tolerant amino acid position. Target proteins are shown in grey. **b**, Affinity measurements for DBVen1619_2, DBPro1156_2 and DBAct553_2 by biolayer interferometry. Each measurement was obtained in the presence (orange) or absence (blue) of the respective small molecule. The fits were calculated using

a nonlinear four-parameter curve-fitting analysis. **c**, Crystal structure of DBAct553_1 in complex with actinonin-bound PDF1 (PDB 8S1X). The computational model (light pink) is aligned with the crystal structure (magenta). Inset shows the alignment of the residues at the interface. **d**, Cryo-electron microscopy (cryo-EM) structure obtained for DBPro1156_2 in complex with progesterone (prog.)-bound DB3. The computational model (light blue) is aligned with the cryo-EM structure (dark blue). Inset shows the alignment of the residues at the interface.

Fig. 2). Owing to the large number of beneficial mutation candidates for DBVen1619_1, we created a combinatorial library covering six residues, sampling a set of favourable amino acids identified by SSM (Supplementary Fig. 9). Three of the six positions converged into single mutations (K1Q, M3L, I13K), whereas the remaining three residues did not converge. We engineered a variant, DBVen1619_2, with the three beneficial mutations and confirmed the binding improvement on yeast display (Supplementary Fig. 9). Among the favourable mutations, M3L in the core of the interface between Bcl2–venetoclax and DBVen1619_2 had a crucial role (Extended Data Fig. 2). The conformational rigidity of a leucine is likely to be preferred to the rotameric flexibility of a methionine⁴², reducing the entropic cost of the binding interaction⁴³.

On the other hand, the second beneficial mutation (I13K) is likely to provide a favourable electrostatic interaction with a glutamate nearby. Overall, the incorporation of the three mutations resulted in a 42-fold improvement of the affinity ($K_D = 96$ nM, Fig. 4b).

For the progesterone-dependent binder DBPro1156_1, four favourable mutations were identified by SSM and showed increased binding on yeast display (Supplementary Fig. 10). Two mutations (Y12W and S16G) significantly improved the binding signal and showed an additive effect in the resulting design, DBPro1156_2. Modelling of the two mutations suggested increased interface packing (Y12W) and removal of a steric clash (S16G) (Extended Data Fig. 2). DBPro1156_2 showed a binding affinity of 18 nM, representing an improvement of three orders

of magnitude relative to the parent design, solely with two mutations (Fig. 4b).

Several mutations were found to slightly improve binding of DBAct553_1 to actinonin-bound PDF1 (Supplementary Fig. 11). Most of these mutations (for instance, R7N and A8R) were predicted to result in a more elaborate hydrogen bond network across the interface (Extended Data Fig. 2). Of note, the combination of I3E with R7N was found to be deleterious for binding (Supplementary Fig. 11), probably because the spatial proximity of these mutations might trigger unwanted side chain rearrangement. A combination of the beneficial mutations (R7N and A8R) gave rise to DBAct553_2, which bound with an affinity of 446 nM to actinonin-bound PDF1 (Fig. 4b).

To evaluate the structural accuracy of our computational design approach, we cocrystallized the ternary complex of actinonin-bound PDF1 with DBAct553_1 (PDB 8S1X, Fig. 4c). The crystal structure closely resembled the computational model, with a C α root mean square deviation (r.m.s.d.) of 2.33 Å and a full-atom interface r.m.s.d. (i.r.m.s.d.) of 2.26 Å, demonstrating the accuracy of our design pipeline. The deviation from our initial model could to a large extent be attributed to a misplaced residue (Y2) in the model of the design scaffold, which induced a slight shift of the N-terminal helix (Extended Data Fig. 3). Consequently, the C α r.m.s.d. of our model deviated by 0.93 Å from that of the experimental structure (Extended Data Fig. 3). Of note, the AlphaFold2 (ref. 44) prediction of the monomeric designed binder aligned perfectly with our structure, with a C α r.m.s.d. of 0.49 Å, placing residue Y2 with the correct orientation. Overall, this observation, together with previous findings, suggests that increased use of deep learning tools such as AlphaFold could significantly increase model accuracy and therefore the success rate⁴⁵. Finally, we solved a cryo-electron microscopy structure (3.23-Å local resolution) of DBPro1156_2 in complex with DB3 Fab and progesterone that confirmed the designed binding mode and interface engagement with the small molecule (Fig. 4d and Supplementary Fig. 12). Despite the absence of structural data for the remaining design, the mutational sensitivity as assessed by SSM (Fig. 4a) and the lack of binding with the small-molecule analogue (Fig. 3d) suggest that the binder engages the target interface with a binding mode in agreement with our computational model.

Functionalization in cell-based systems

Chemically controllable components have important applications in synthetic biology and have been shown to be useful in modulating the activity of emerging cell-based therapies^{10,15,46}. To test whether our computationally designed CIDs would assemble in a more complex cellular context, we engineered reporter proximity-based systems that were expressed in a cell-free system or mammalian cells and could activate a signalling pathway or lead to the reconstitution of a reporter protein in the presence of the small molecule. The most natural functional logic for chemically induced protein interactions is to function as ON-switch systems (Fig. 5, Extended Data Fig. 4 and Supplementary Fig. 13).

We first repurposed a previously described heterodimerization-based reporter system⁴⁷ to test the DB3 antibody as a single-chain variable fragment (scFv) binding to DBPro1156_2. Here DB3 was fused to a zinc-finger 438 transcription factor and DBPro1156_2 to a T7 RNA polymerase (Fig. 5a), and testing was performed in a cell-free reporter system. Heterodimerization in the presence of progesterone induced proximity between the T7 RNA polymerase and the transcription factor, leading to the transcription of a reporter linear DNA template and its translation into a red fluorescent protein (mCherry). Whereas only baseline fluorescence was observed in the absence of progesterone, a 15.8-fold increase was observed after addition of progesterone (Fig. 5b). Similarly, progesterone titration demonstrated a dose–response relationship, suggesting its possible use as a new cell-free biosensor (Fig. 5c).

To test the chemically induced activity of the designed modules in mammalian cells, we used a previously described system called

generalized extracellular molecule sensor (GEMS)¹⁷. Briefly, the target protein and the designed binder are both fused to an erythropoietin receptor (EpoR) linked to an intracellular domain of human interleukin-6 receptor subunit B (IL-6RB) (Extended Data Fig. 4). Transcription of a reporter gene (NanoLuciferase)⁴⁸ is triggered upon a conformational change induced by heterodimerization in the presence of the drug. After incorporating Bcl2 and DBVen1619_2 in the GEMS system, we observed a 26.8-fold change in luminescence in the presence of venetoclax, whereas minimal background was observed in the absence of the drug (Extended Data Fig. 4). These results show the desired behaviour of an ON-switch system. In addition, our modified GEMS system showed heightened sensitivity to the drug, with a half-maximal effective concentration (EC₅₀) of 0.31 nM, probably owing to colocalization of the sensing modules in the cell membrane (Extended Data Fig. 4).

Next, we designed a cytoplasmic system to respond to actinonin and fused PDF1 and DBAct553_1 to two moieties of a split NanoLuc (Fig. 5d). In this system, we also observed a significant increase in signal (19.1-fold) upon dosing of the cells with actinonin (Fig. 5e). This ON-switch system was also highly sensitive to the presence of the drug, as shown by the titration reporting an EC₅₀ of 27 nM (Fig. 5f).

Finally, we demonstrated that our CID system could be used to control tumour-killing activity in primary murine T cells engineered to express a chimeric antigen receptor (CAR). Whereas the classical second-generation CAR (2G-CAR) has an extracellular recognition domain and an intracellular activation domain⁴⁹, we decoupled the two CAR components into two chains, using Bcl2 and DBVen1619_2, that dimerize in the presence of venetoclax (split CID-CAR, Fig. 5g). With this split CID-CAR, we observed inducible CAR-T cell killing of HER2-expressing tumour cells (MC38) upon addition of venetoclax, whereas no difference was observed with the classical 2G-CAR (Fig. 5h). Despite the desired effect, residual tumour killing in the absence of the drug and a slightly lower potency compared with that of the 2G-CAR were observed. Nevertheless, tumour killing was stable over time (up to 48 h), and significant efficacy was achieved at a concentration of 10 nM venetoclax (Fig. 5i and Supplementary Fig. 14). Overall, we showed that our computationally designed CIDs could be used to functionalize molecular components in cellular systems, suggesting a promising route for the development of new modules for synthetic biology, including a wide range of biosensors and cell-based applications.

Optimization for further binders

Considering the difficulty of the design task, the experimental success rate remained modest (one binder out of 2,000 designs). However, incorporating AlphaFold2 (refs. 44,50) as a filtering step represents a promising approach to improve design success rates, as a substantial proportion of the computational designs were predicted to not fold in silico (Extended Data Fig. 5). Recently, deep learning tools such as LigandMPNN have been proposed for sequence design tasks that include small molecules^{27,51}. Using LigandMPNN to optimize the sequences of the 2,000 designs tested for each protein–ligand complex (Methods), we observed an increased folding rate in silico. The top 500 designs for each target complex were selected (excluding known binders) on the basis of the computational metrics previously described. Using yeast display, we screened and isolated one new binder for Bcl2–venetoclax and 12 for PDF1–actinonin, representing improvements of 4-fold and 52-fold in the success rate, respectively, compared with the original approach (Extended Data Figs. 6 and 7). Most newly identified designs were reported as misfolded in the original pool, before optimization with LigandMPNN. Of note, each newly identified seed grafted on these new designs demonstrated specificity upon a point mutation at the interface (Extended Data Fig. 8). Thus, we foresee a promising synergy between emerging novel sequence design tools and our MaSIF-neosurf approach for challenging design tasks.

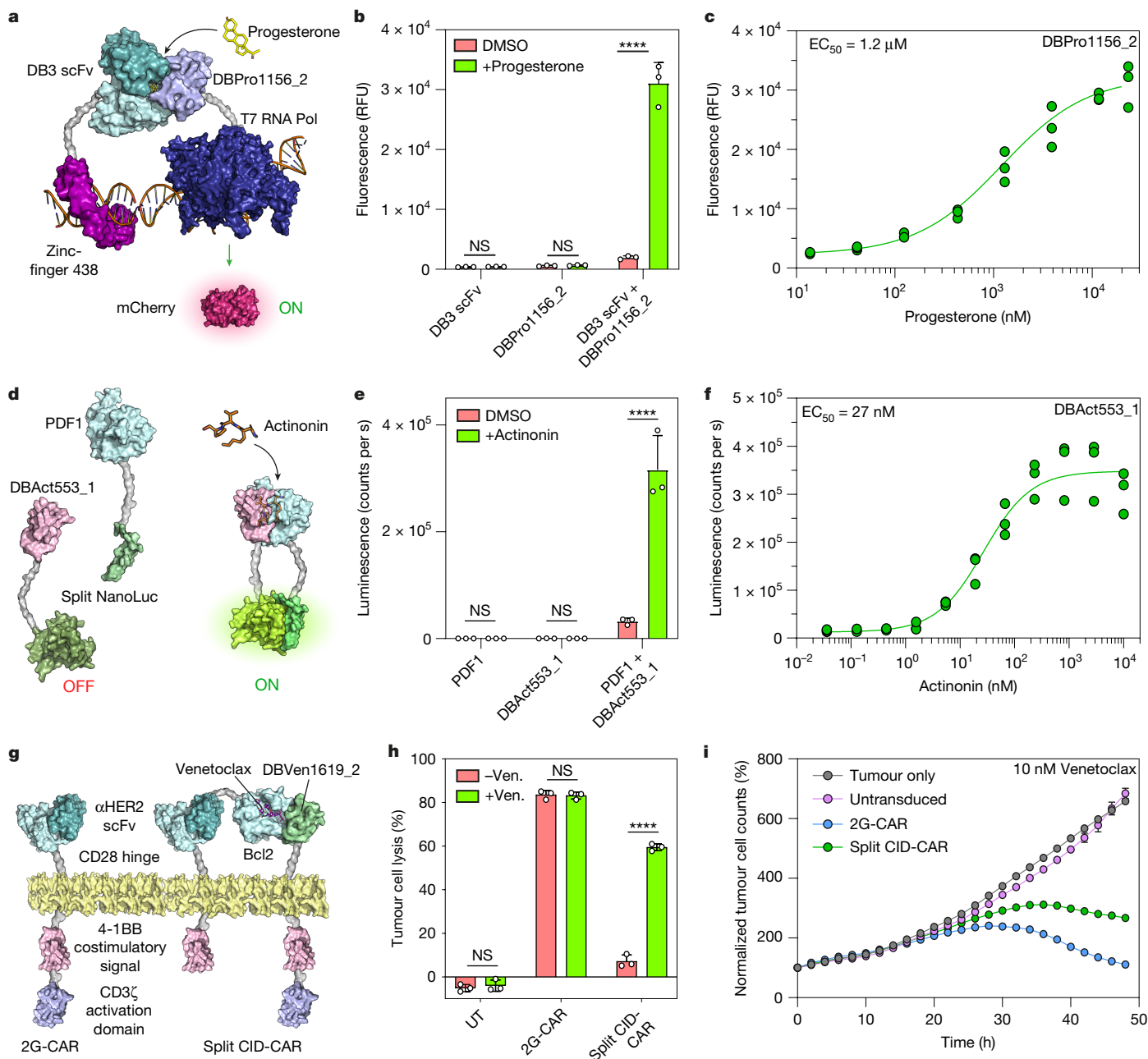


Fig. 5 | Computationally designed CIDs are functional in cell-based systems. **a**, Schematic of the cell-free expression system with scFv DB3 fused to a zinc-finger transcription factor and DBPro1156_2 fused to T7 RNA polymerase (Pol). **b**, Fluorescence (relative fluorescence units; RFU) measured with each monomeric component or mixed, with or without 20 μ M progesterone. **c**, Progesterone-dose-dependent responses performed in a cell-free system containing both components. **d**, Schematic of the split NanoLuc system functionalizing DBAct553_1 and PDF1. **e**, Intracellular NanoLuc luminescence of HEK293T transfected with C-terminal split NanoLuc-fused PDF1 only, N-terminal split NanoLuc-fused DBAct553_1 only or both together, with or without 10 μ M actinonin. **f**, Actinonin-dose-dependent responses performed on HEK293T transfected with both components. **g**, Schematic representation of α HER2-specific 2G-CAR and the drug-inducible α HER2-CAR split system

(split CID-CAR). The two domains assembled upon addition of venetoclax. **h**, Killing efficiency of CAR-T cells with and without venetoclax using untransduced (UT) murine primary T cells or cells transduced with 2G-CAR or the split CID-CAR. Tumour cell lysis was measured after 48 h of cocultivation with target cells. The percentage of live target cells was normalized to the number of live cells in each well at $t = 0$ h and further normalized to the growth of target cells cultured alone. **i**, Killing efficiency of CAR-T cells measured over time. Tumour cell counts at different time points were normalized to the number of live cells in each well at $t = 0$ h. A concentration of 10 nM venetoclax was used. Two-way ANOVA with Tukey's multiple comparison test; NS, not significant. **** $P < 0.0001$. Data are presented as mean \pm s.d. Data points are derived from three technical replicates (a–f) or three biological replicates (g–i).

Discussion

Most deep-learning-based protein design pipelines are primarily conditioned on the natural amino acid repertoire^{52–54} and therefore lack generalization to the design of interactions involving small molecules.

This gap is mainly due to the scarcity of protein–ligand structural data, especially for ternary complexes, in training sets based on the Protein Data Bank (PDB), in which such structures are rare^{55–57}. Geometric deep learning approaches principled in the physical and chemical features of the molecular surface can overcome these limitations and provide

joint representations for protein and small-molecule complexes. The resulting neosurfaces capture and present generalizable molecular features that enable the challenging task of designing protein binders targeting these hybrid interfaces. Using the MaSIF-neosurf framework, we successfully designed three specific binders against Bcl2–venetoclax, DB3–progesterone and PDF1–actinonin complexes. All designed binders showed high stability, specificity and native-like affinity for their target complexes by pure in silico generation. The affinities were experimentally optimized to nanomolar range, and the binding mode was confirmed through mutational and structural characterization, demonstrating the accuracy of our design pipeline. Notably, our pipeline captured subtle yet crucial contributions of each ligand (10–12% of the buried solvent-accessible surface area only; Supplementary Table 3) to induce protein interactions. This sensitivity represents a further layer of complexity in the task of designing highly sensitive CIDs, compared with previous attempts targeting large ligand interfaces⁶.

To demonstrate the functionality of our designed CID systems, we probed their efficiency and specificity in the context of a complex cellular environment. They exhibited robust ON-switch behaviour in both cytoplasmic and membrane-bound circuits, indicating their potentially wide applicability in mammalian systems as logic gates, synthetic circuits or new biosensors for detecting specific metabolites^{15,17}. This relevance is further underscored by our use of the FDA-approved drug venetoclax for treating leukaemia²⁸, the natural product actinonin with potentially chemotherapeutic effects³⁰ and the endogenous hormone progesterone⁵⁸. These can be used for combined anticancer therapies with CAR-T cells, which are often hindered by off-target toxicities^{10,59}. Addition of synthetic small-molecule activators could allow finer control of their activity and elevate their safety profile.

Although the design of specific protein–ligand interactions remains challenging, the results presented here lay a strong foundation for further innovations. Experimental methods such as antibody screening platforms²⁶ are agnostic to where and how proteins engage their respective targets. Deep learning tools, such as the one presented here, can be used to control these parameters and offer more modalities in terms of protein shapes, folds, sizes and thermal stability. However, some challenges remain, as state-of-the-art deep-learning-based structure validation methods including RoseTTAFold²⁷ failed to predict our validated complexes (Supplementary Fig. 2). New tools such as AlphaFold3 demonstrated good performance for ligand–protein complex prediction; however, their limited scope of use poses non-negligible hurdles to further advances in the field of drug and/or protein design⁶⁰. We foresee that approaches such as surface fingerprinting could represent a suitable alternative for targeting neosurfaces. Despite the achievements of our pipeline, we were unsuccessful in the case of the BRD4–JQ1 complex, probably owing to the flexibility of the ligand and the inferior computational metrics of the designs compared with those of other test cases (Supplementary Fig. 15). Most deep learning methods, including ours, exhibit superior performance on hydrophobic patches, whereas significant challenges persist in accurately modelling polar interfaces^{9,40}. Overall, we expect that surface-based representation will contribute to solving molecular design problems in low-data regimens, such as the design of protein-based molecules with non-natural amino acids. The ability to extract expressive fingerprints from protein–ligand complexes opens up the possibility of rationally designing innovative drug modalities, such as on-demand cell-based therapies^{10,19}, controllable biologics^{22,24} and molecular glues, which has remained an outstanding challenge in drug development^{13,14}.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions

and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-024-08435-4>.

- Janin, J., Bahadur, R. P. & Chakrabarti, P. Protein–protein interaction and quaternary structure. *Q. Rev. Biophys.* **41**, 133–180 (2008).
- Monod, J., Changeux, J.-P. & Jacob, F. Allosteric proteins and cellular control systems. *J. Mol. Biol.* **6**, 306–329 (1963).
- Seet, B. T., Dikic, I., Zhou, M.-M. & Pawson, T. Reading protein modifications with interaction domains. *Nat. Rev. Mol. Cell Biol.* **7**, 473–483 (2006).
- Patel, D., Kopec, J., Fitzpatrick, F., McCorvie, T. J. & Yue, W. W. Structural basis for ligand-dependent dimerization of phenylalanine hydroxylase regulatory domain. *Sci. Rep.* **6**, 23748 (2016).
- Schlessinger, J. Ligand-induced, receptor-mediated dimerization and activation of EGF receptor. *Cell* **110**, 669–672 (2002).
- Foight, G. W. et al. Multi-input chemical control of protein dimerization for programming graded cellular responses. *Nat. Biotechnol.* **37**, 1209–1216 (2019).
- Glasgow, A. A. et al. Computational design of a modular protein sense-response system. *Science* **366**, 1024–1028 (2019).
- Gainza, P. et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods* **17**, 184–192 (2020).
- Gainza, P. et al. De novo design of protein interactions with learned surface fingerprints. *Nature* **617**, 176–184 (2023).
- Giordano-Attianese, G. et al. A computationally designed chimeric antigen receptor provides a small-molecule safety switch for T-cell therapy. *Nat. Biotechnol.* **38**, 426–432 (2020).
- Gonzalez, M. W. & Kann, M. G. Chapter 4: protein interactions and disease. *PLoS Comput. Biol.* **8**, e1002819 (2012).
- Marchand, A., Van Hall-Beauvais, A. K. & Correia, B. E. Computational design of novel protein–protein interactions – an overview on methodological approaches and applications. *Curr. Opin. Struct. Biol.* **74**, 102370 (2022).
- Schreiber, S. L. The rise of molecular glues. *Cell* **184**, 3–9 (2021).
- Oleinikovas, V., Gainza, P., Ryckmans, T., Fasching, B. & Thomä, N. H. From thalidomide to rational molecular glue design for targeted protein degradation. *Annu. Rev. Pharmacol. Toxicol.* **64**, 291–312 (2024).
- Shui, S., Buckley, S., Scheller, L. & Correia, B. E. Rational design of small-molecule responsive protein switches. *Protein Sci.* **32**, e4774 (2023).
- Shui, S. et al. A rational blueprint for the design of chemically-controlled protein switches. *Nat. Commun.* **12**, 5754 (2021).
- Scheller, L., Strittmatter, T., Fuchs, D., Bojar, D. & Fussenegger, M. Generalized extracellular molecule sensor platform for programming cellular behavior. *Nat. Chem. Biol.* **14**, 723–729 (2018).
- Wells, J. A. & Kumru, K. Extracellular targeted protein degradation: an emerging modality for drug discovery. *Nat. Rev. Drug Discov.* <https://doi.org/10.1038/s41573-023-00833-z> (2023).
- Jan, M. et al. Reversible ON- and OFF-switch chimeric antigen receptors controlled by lenalidomide. *Sci. Transl. Med.* **13**, eabb6295 (2021).
- Ishida, M. et al. Synthetic self-localizing ligands that control the spatial location of proteins in living cells. *J. Am. Chem. Soc.* **135**, 12684–12689 (2013).
- Gibson, W. J. et al. Bifunctional small molecules that induce nuclear localization and targeted transcriptional regulation. *J. Am. Chem. Soc.* **145**, 26028–26037 (2023).
- Marchand, A. et al. Rational design of chemically controlled antibodies and protein therapeutics. *ACS Chem. Biol.* **18**, 1259–1265 (2023).
- Mata, M. et al. Inducible activation of MyD88 and CD40 in CAR T cells results in controllable and potent antitumor activity in preclinical solid tumor models. *Cancer Discov.* **7**, 1306–1319 (2017).
- Martinko, A. J. et al. Switchable assembly and function of antibody complexes in vivo using a small molecule. *Proc. Natl Acad. Sci. USA* **119**, e2117402119 (2022).
- Spencer, D. M., Wandless, T. J., Schreiber, S. L. & Crabtree, G. R. Controlling signal transduction with synthetic ligands. *Science* **262**, 1019–1024 (1993).
- Hill, Z. B., Martinko, A. J., Nguyen, D. P. & Wells, J. A. Human antibody-based chemically induced dimerizers for cell therapeutic applications. *Nat. Chem. Biol.* **14**, 112–117 (2018).
- Krishna, R. et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* <https://doi.org/10.1126/science.adl2528> (2024).
- Roberts, A. W. et al. Targeting BCL2 with venetoclax in relapsed chronic lymphocytic leukemia. *N. Engl. J. Med.* **374**, 311–322 (2016).
- Arevalo, J. H., Stura, E. A., Taussig, M. J. & Wilson, I. A. Three-dimensional structure of an anti-steroid Fab' and progesterone-Fab' complex. *J. Mol. Biol.* **231**, 103–118 (1993).
- Chen, D. Z. et al. Actinonin, a naturally occurring antibacterial agent, is a potent deformylase inhibitor. *Biochemistry* **39**, 1256–1262 (2000).
- Koenderink, J. J. & van Doorn, A. J. Surface shape and curvature scales. *Image Vis. Comput.* **10**, 557–564 (1992).
- Yin, S., Proctor, E. A., Lugovskoy, A. A. & Dokholyan, N. V. Fast screening of protein surfaces using geometric invariant fingerprints. *Proc. Natl Acad. Sci. USA* **106**, 16622–16626 (2009).
- Morozov, A. V. & Kortemme, T. Potential functions for hydrogen bonds in protein structure prediction and design. *Adv. Protein Chem.* **72**, 1–38 (2005).
- Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).
- Wildman, S. A. & Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **39**, 868–873 (1999).
- Eisenberg, D., Schwarz, E., Komaromy, M. & Wall, R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.* **179**, 125–142 (1984).
- Fleishman, S. J. et al. RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite. *PLoS ONE* **6**, e20161 (2011).

38. Chao, G. et al. Isolating and engineering human antibodies using yeast surface display. *Nat. Protoc.* **1**, 755–768 (2006).
39. Cao, L. et al. De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. *Science* **370**, 426–431 (2020).
40. Cao, L. et al. Design of protein-binding proteins from the target structure alone. *Nature* **605**, 551–560 (2022).
41. Perkins, J. R., Diboun, I., Dessailly, B. H., Lees, J. G. & Orengo, C. Transient protein-protein interactions: structural, functional, and network properties. *Structure* **18**, 1233–1243 (2010).
42. Najmanovich, R., Kuttner, J., Sobolev, V. & Edelman, M. Side-chain flexibility in proteins upon ligand binding. *Proteins* **39**, 261–268 (2000).
43. Bissantz, C., Kuhn, B. & Stahl, M. A medicinal chemist's guide to molecular interactions. *J. Med. Chem.* **53**, 5061–5084 (2010).
44. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
45. Bennett, N. R. et al. Improving de novo protein binder design with deep learning. *Nat. Commun.* **14**, 2625 (2023).
46. Shui, S., Scheller, L. & Correia, B. E. Protein-based bandpass filters for controlling cellular signaling with chemical inputs. *Nat. Chem. Biol.* <https://doi.org/10.1038/s41589-023-01463-7> (2023).
47. Hussey, B. J. & McMillen, D. R. Programmable T7-based synthetic transcription factors. *Nucleic Acids Res.* **46**, 9842–9854 (2018).
48. England, C. G., Ehlerding, E. B. & Cai, W. NanoLuc: a small luciferase is brightening up the field of bioluminescence. *Bioconjug. Chem.* **27**, 1175–1187 (2016).
49. Zhao, Y. et al. IL-10-expressing CAR T cells resist dysfunction and mediate durable clearance of solid tumors and metastases. *Nat. Biotechnol.* <https://doi.org/10.1038/s41587-023-02060-8> (2024).
50. Mirdita, M. et al. ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
51. Dauparas, J. et al. Atomic context-conditioned protein sequence design using LigandMPNN. Preprint at *bioRxiv* <https://doi.org/10.1101/2023.12.22.573103> (2023).
52. Khakzad, H. et al. A new age in protein design empowered by deep learning. *Cell Syst.* **14**, 925–939 (2023).
53. Watson, J. L. et al. De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).
54. Dauparas, J. et al. Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
55. Rui, H., Ashton, K. S., Min, J., Wang, C. & Potts, P. R. Protein-protein interfaces in molecular glue-induced ternary complexes: classification, characterization, and prediction. *RSC Chem. Biol.* **4**, 192–215 (2023).
56. Ferreira De Freitas, R. & Schapira, M. A systematic analysis of atomic protein-ligand interactions in the PDB. *MedChemComm* **8**, 1970–1981 (2017).
57. Orasch, O. et al. Protein-protein interaction prediction for targeted protein degradation. *Int. J. Mol. Sci.* **23**, 7033 (2022).
58. Nagy, B. et al. Key to life: physiological role and clinical implications of progesterone. *Int. J. Mol. Sci.* **22**, 11039 (2021).
59. Morgan, R. A. et al. Case report of a serious adverse event following the administration of T cells transduced with a chimeric antigen receptor recognizing ERBB2. *Mol. Ther.* **18**, 843–851 (2010).
60. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025

Methods

Incorporation of small molecules in MaSIF

Molecular surface meshes were triangulated using the MSMS program⁶¹, and radial patches (geodesic radius 12 Å) were computed following the original MaSIF preprocessing scripts⁸. Before MaSIF's geodesic convolutional layers are applied, five input features are computed for each patch: shape index³¹, distance-dependent curvature³², Poisson–Boltzmann continuum electrostatics, hydrogen bond donor and acceptor potential³³, and hydrophobicity^{34–36}. The first two features are purely geometric and are calculated analogously to protein surfaces alone. Moreover, the APBS program⁶² used to compute the Poisson–Boltzmann electrostatics on the surface supports small molecules in the MOL2 file format and hence does not require us to treat them in a conceptually different way. The remaining two chemical input features are computed as described below.

Hydrogen bond donors and acceptors. The hydrogen bond propensity feature assigns a positive value to points on the molecular surface near the optimal direction in which a hydrogen bond could be formed with an acceptor atom. It is determined by the direction of the covalent bond between a donor atom and its hydrogen (Supplementary Fig. 1b,c). Likewise, a negative value is assigned to points corresponding to hydrogen bond acceptors. For different acceptor types, the theoretically optimal position for forming a hydrogen bond can either lie on a cone (Supplementary Fig. 1d–f) or in a small number of specific directions that can be derived from the molecular geometry. We assign different magnitudes of the donor or acceptor feature on the basis of the angular deviation from the ideal hydrogen bond geometry according to a quadratic function.

The optimal direction of the hydrogen bond was determined using the RDKit software package⁶³, and surface points were assigned positive (donor) or negative (acceptor) values between –1 and +1 on the basis of their angular deviation from the ideal direction. For potential acceptors, RDKit was also used to determine whether the idealized location of the hydrogen bond lay on a cone or in one or more discrete directions.

Hydrophobicity. MaSIF's hydrophobicity feature makes use of the Kyte–Doolittle scale³⁴, which is exclusively defined for amino acids. Equivalent values for small molecules thus need to be approximated on the basis of a more general hydrophobicity measure that can be estimated computationally, such as the logarithm of the octanol–water partition coefficient (logP)³⁵. To this end, we developed a nonlinear function that maps logP values to the Kyte–Doolittle scale. We fit the parameters of this function to find an optimal match for the Kyte–Doolittle and logP values of all 20 amino acids. As the best functional form of this mapping was not immediately obvious from the raw values (Supplementary Fig. 1l), we experimented with different hydrophobicity scales as intermediates and found that the Eisenberg scale³⁶ had approximately linear and exponential relationships with logP and Kyte–Doolittle values of amino acids, respectively. We first computed the optimal parameters of the mappings from logP to Eisenberg scale (Supplementary Fig. 1g) and Eisenberg scale to Kyte–Doolittle scale (Supplementary Fig. 1h) and then composed these two functions to establish the desired relationship between logP and Kyte–Doolittle values (Supplementary Fig. 1i). Finally, we also restricted the outputs to the valid interval of Kyte–Doolittle values [–4.5, 4.5] to ensure that the feature did not leave the domain on which MaSIF was trained.

Furthermore, as some ligands can cover large surface patches, we aimed to capture local variations of the hydrophobicity by fragmenting the molecules before calculating their hydrophobicity scores. We used the BRICS algorithm⁶⁴ to decompose molecules and compute estimates of the logP value of each fragment with RDKit. The resulting fragments were more similar in size to amino acids and tended to have

less extreme hydrophobicity scores than whole ligands, moving the distribution of this feature closer to that expected on protein surfaces (Supplementary Fig. 1k–l). To translate from logP to the Kyte–Doolittle scale, we parameterized a function so that it approximated the relationship between these hydrophobicity values for the 20 amino acids. Kyte–Doolittle and Eisenberg values of all amino acids are available in tabular form, whereas we computed their logP with RDKit to fit the curves. The final function was:

$$\text{Kyte–Doolittle} = \text{clip}(-6.2786 + \exp(0.4772 \times \log P + 1.8491), \\ \text{min} = -4.5, \text{max} = 4.5).$$

After computing equivalent Kyte–Doolittle values for all small-molecule fragments, we assigned the resulting hydrophobicity score of the closest fragment to each surface vertex.

To create the histograms in Supplementary Fig. 1k–l, we extracted 20,363 unique small-molecule ligands from the Binding MOAD⁶⁵ dataset, fragmented each and removed duplicates. This resulted in 9,362 unique fragments that were compared with the set of ligands and the 20 standard amino acids.

Target protein selection

The target proteins were selected on the basis of several factors including the reported protein–ligand affinity⁶⁶, the resolution of the structural data, the interface propensity, and the solvent-accessible surface area of the small molecule when bound to the receptor to ensure a measurable interface with the designed binders. More practical considerations such as small molecule purchase availability or feasibility of target protein expression were also considered.

Binding site identification

MaSIF-site⁸ was trained on a dataset of known PPIs to predict regions on protein surfaces with high propensity to form a buried interface. The neural network takes a protein–ligand complex decomposed into 12-Å (geodesic radius) overlapping patches as input and generates a per-vertex regression score, indicating the propensity of each point to become a buried surface area within a protein interaction. In this study, we used MaSIF-site to predict interfaces and guide the selection of target patches both in our computational benchmark and for all three target complexes for design (Bcl2–venetoclax, DB3–progesterone and PDF1–actinonin). In the computational benchmark, we conducted the search only for the three patches with the highest interface propensity near the centre of the binding site. For design, the number of targeted sites overlapping with the protein–ligand neosurface depended on the solvent-accessible surface area of each ligand to ensure that all the ligand-exposed surface was covered during the complementary motif search. This number was 1 for PDF1–actinonin, 2 for DB3–progesterone and 3 for Bcl2–venetoclax.

Binding seed identification

The fingerprints of the predicted 12-Å (geodesic radius) patches comprising both protein target and bound small molecule were used to find a complementary fingerprint in the MaSIF-seed database⁹, which contains approximately 640,000 continuous structural fragments (seeds) amounting to 402 million surface patches (also known as fingerprints). The seed database covers distinct secondary structures with approximately 390,000 sheet-based and 250,000 helical motifs. The MaSIF-search algorithm was trained to make patch fingerprints similar for interacting patches and dissimilar for non-interacting patches. Seeds with interface propensity scores above the defined threshold and with fingerprint distances (Euclidean distance between target and seed fingerprint) below the defined thresholds were selected. In second-stage alignment and scoring using the RANSAC algorithm, seeds were selected on the basis of IPA score. Cutoffs used for the seed selection are summarized in Supplementary Table 1.

Scoring aligned structures

We consider two descriptor-based postalignment scores. The descriptor distance score is a simple heuristic that aggregates descriptor distances across the predicted binding interface and is based on the squared Euclidean distances between interacting patches on each side of the interface. Two patches are considered to interact with each other if their centre points are less than 1.5 Å apart. The descriptor distance score is computed according to the following formula:

$$\text{DDS} = \sum_i \frac{1}{\|\text{binder_desc}(i) - \text{target_desc}(\text{NN}(i))\|^2}$$

where DDS is the descriptor distance score, i indexes interacting patches of the first protein and $\text{NN}(i)$ returns the index of the spatially nearest neighbour on the other protein. Higher scores mean higher complementarity.

The IPA score is computed by a neural network that was trained to discriminate between near-native and high-r.m.s.d. poses of docked proteins⁸. The inputs of this predictor are three-dimensional Euclidean distances, descriptor distances and dot products between surface normals of up to 200 pairs of corresponding patches at the predicted interface. The predictor outputs values between 0 and 1, where larger values indicate higher confidence in the presented interface.

Computational binder recovery benchmark

The binder recovery experiment was performed for 14 known ligand-induced protein complexes, in which both proteins involved in the interaction are considered as separate items, resulting in 28 search queries. In addition, we included 8,907 decoys based on 2,852 PPIs from the PDBbind (v.2020)⁶⁶ database. We split the provided structures into separate chains and only applied light filtering to remove nuclear magnetic resonance (NMR) structures, duplicate sequences within the same structure, and structures that could not be processed. All benchmark complexes and decoys are listed in Supplementary Table 4 and in the GitHub repository respectively ('Code availability'). After triangulating and featurizing all protein surfaces with and without ligands, we screened the database and docked candidates, analogous to the binding seed search. Here we assumed the location of the binding site on the target protein was known and selected the three surface vertices with the largest predicted surface propensity within 10 Å of the centre of this site as input patches. The centre of the binding site was approximated with a simple heuristic. We first identified interface atoms as those within 4 Å of any atom from the binding protein in the original complex structure. This could and typically did include atoms belonging to the small molecule. Then, we defined the average of the coordinates of all interface atoms of the target protein as the centre of the binding site. Furthermore, we declared a binder to be correctly recovered if its i.r.m.s.d. compared with the ground truth structure of the same protein was less than 5 Å, where i.r.m.s.d. considered only heavy atoms in the immediate vicinity of the target protein (less than 5 Å).

Seed and interface refinement

To optimize binding energy of the seed for the target complex, seeds were refined using a FastDesign protocol on Rosetta³⁷ with a penalty for buried unsatisfied polar atoms in the scoring function⁶⁷. Refined seeds were then selected on the basis of the computed binding energy (ddG), shape complementarity, number of interface hydrogen bonds, number of buried unsatisfied polar atoms and number of atoms in contact with the small molecule. β -sheet-based motifs making more than 33% contact with the target complex using loop regions were discarded. Moreover, the uniqueness of each seed was assessed by a pairwise alignment of the hotspot residues. For seeds showing more than 70% hotspot identity with another seed, only the one with the best surface-normalized ddG was kept.

Seed grafting and computational design

For each target, approximately 100–120 selected seeds were subsequently grafted with a Rosetta MotifGraft⁶⁸ protocol to stabilize the binding motif and bring further contacts with the target complex. Each seed was matched with a database of around 6,500 small protein scaffolds (less than 90 amino acids) originating from small globular monomeric proteins from the PDB⁶⁹ and four computationally designed miniprotein databases that had been experimentally validated^{70–73}. Before grafting on multiple scaffolds, seeds were cropped to the minimum number of residues making contact with the target, and loop motifs were removed from β -sheet-based seeds to optimize the grafting success rate. Once grafting had been performed, scaffolds underwent sequence optimization using a FastDesign protocol on Rosetta with a penalty for buried unsatisfied polar atoms in the scoring function. Final designs were selected based on the ddG, shape complementarity, number of interface hydrogen bonds and count of buried unsatisfied polar atoms. A similar number of designs per seed was ensured by setting dynamic cutoffs of these metrics adjusted for each seed.

Design optimization with LigandMPNN

Designs that did not show any binding in the first round of experimental screening underwent sequence optimization with LigandMPNN⁵¹. Ten sequences per design were generated and folded with AlphaFold2 in the ColabFold software⁵⁰ (single sequence mode). α -r.m.s.d. values between AlphaFold2 predictions and the original model were measured, and only one sequence per design with the lowest r.m.s.d. was selected. Designs in complex with their respective target were relaxed with Rosetta and filtered based on the ddG, shape complementarity, number of interface hydrogen bonds and number of buried unsatisfied polar atoms. Five-hundred designs per target complex were selected and rescreened by yeast display.

Library screening

For each target complex, around 2,000 protein designs were reverse-translated into DNA and purchased from Twist Bioscience as oligo pools with 18-bp homology overhangs. Oligo pools underwent two rounds of PCR: (1) for amplification of the library using the 18-bp overhangs; and (2) for addition of 45-bp homology with the yeast display vector (57.5 °C annealing for 30 s, 72 °C extension time for 1 min, 15 cycles). EB100 yeast was transformed by electroporation using the amplified inserts and linearized HA-tagged pCTcon2 vector as described previously³⁸. A similar approach was used for the SSM library of single designs. Transformed yeast cells were grown in minimal glucose medium (SDCAA) at 30 °C and induced with minimal galactose medium (SGCAA) overnight before sorting.

Yeast surface display of single designs

Genes encoding single designs were purchased from Twist Bioscience with an approximately 25-bp homology overhang for cloning. Each design was cloned into an HA-tagged pCTcon2 plasmid using Gibson assembly and transformed into XL10-Gold or HB101 bacteria for DNA production. The purified and sequence-approved DNA was then used to transform competent EB100 yeast using a Frozen-EZ Yeast Transformation II Kit (Zymo Research). For libraries, transformed yeast cells were grown in minimal glucose medium (SDCAA) at 30 °C and induced with minimal galactose medium (SGCAA) overnight before flow cytometry analysis.

Flow cytometry analysis and sorting

Induced yeast cells were washed with phosphate-buffered saline (PBS) supplemented with 0.1% bovine serum albumin and then labelled with the respective binding target for 2 h at 4 °C. Before labelling, protein–ligand complexes were preincubated at room temperature for 5 min with a 1:5–10 ratio. Cells were then washed and labelled with

an FITC-conjugated goat anti-HA tag antibody (Bethyl, A190-138F; display tag; 1:100 dilution) and a PE-conjugated goat anti-human Fc antibody (Invitrogen, 12-4317-87; binding tag; 1:100 dilution) for 30 min at 4 °C. Cells were washed, resuspended in an appropriate volume of buffer and analysed on a Gallios flow cytometer (Beckman Coulter) or sorted with a Sony SH800 cell sorter. Kaluza software (Beckman Coulter, v.1.1.20388.18228) and LE-SH800SZFCPL Cell Sorter (Sony, v.2.1.5) were respectively used for the data acquisition. In the case of cell sorting, each designed library was sorted for binding and non-binding populations separately. Flow cytometry data were then analysed using FlowJo (BD Biosciences, v.10.8.1).

Library sequencing

Sorted yeasts were cultured and plasmids encoding protein designs were extracted using a Zymo Prep Yeast Plasmid Miniprep II (Zymo Research) following the manufacturer's protocol. The sequence of interest was then amplified by PCR with vector-specific primers flanking the protein design gene. A second PCR was performed to add Illumina adaptors and Nextera barcodes, and the PCR product was desalted and purified using a Qiaquick PCR purification kit (Qiagen). An Illumina MiSeq system with 500 cycles was used for next-generation sequencing. Around 0.8–1.2 millions reads per sample were obtained; these were translated into the appropriate reading frame and matched with expected input sequences from the libraries. The enrichment of each design was calculated by normalizing the counts in the binding population with the counts in the non-binding populations. Hits were identified if the enrichment was more than ten-fold and the number of counts in the binding population was greater than 10,000.

Protein expression and purification

A list of protein sequences can be found in Supplementary Table 5. Genes encoding the 6xHis-tagged and/or human Fc-tagged protein of interest were purchased from Twist Bioscience, cloned into pET11 (bacterial vector) or pHLSec (mammalian vector) by Gibson assembly and transformed into XL10-Gold or HB101 bacteria. Plasmids were extracted using a GeneJET plasmid Miniprep kit (Thermo Fisher, for bacterial vector) or a PureLink Fast Low-Endotoxin Midi plasmid purification kit (Invitrogen, for mammalian vector) and checked by Sanger sequencing. Proteins were purified using bacterial or mammalian expression systems. Mammalian expression was performed using an Expi293 expression system (Thermo Fisher, A14635). Cells were authenticated (short tandem repeat (STR) genotyping) and tested negative for mycoplasma contamination (quantitative PCR) by the provider. Supernatants were collected after 6 days and filtered and purified as described below. For bacterial expression, BL21(DE3) or T7 Express Competent *Escherichia coli* were transformed with the plasmid of interest and grown as a pre-culture overnight. Precultures were inoculated 1:50 in Terrific Broth medium and incubated at 37 °C until they reached an optical density at 600 nm (OD_{600}) of approximately 0.7. Then, bacteria were induced with 1 mM isopropyl β -D-1-thiogalactopyranoside (IPTG) and incubated overnight at 18–20 °C. Cells were collected by centrifugation at 4,000g for 10 min, resuspended in lysis buffer (50 mM Tris, pH 7.5, 500 mM NaCl, 5% glycerol, 1 mg ml⁻¹ lysozyme, 1 mM phenylmethylsulfonyl fluoride (PMSF) and 1 μ g ml⁻¹ DNase) and lysed by sonication. Lysates were then clarified by centrifugation at 30,000g for 30 min and filtered.

All 6xHis-tagged proteins were purified using an ÄKTA Pure system (GE Healthcare) Ni-NTA HisTrap affinity column, followed by size-exclusion chromatography on a Superdex HiLoad 16/600 75 pg or 200 pg depending on the size of the protein. All proteins were concentrated in PBS as a final buffer.

Surface plasmon resonance

Affinity measurements were performed on a Biacore 8K (GE Healthcare, software v.4.0.8.19879) using HBS-EP+ as a running buffer (10 mM HEPES at pH 7.4, 150 mM NaCl, 3 mM EDTA, 0.005% v/v surfactant

P20; GE Healthcare). All proteins were immobilized on a CMS5 chip (GE Healthcare, catalogue no. 29104988) by means of amine coupling to reach 500–1,000 response units. Analytes were then injected in serial dilutions using the running buffer. The flow rate was 30 μ l min⁻¹ for a contact time of 120 s, followed by 400 s of dissociation time. Surface plasmon resonance data were fitted in steady-state affinity mode by reporting the relative response units for each concentration.

Biolayer interferometry

Biolayer interferometry measurements were performed on a Gator system using GatorOne software (Gator Bio, v.2.7.3.0728). The running buffer consisted of 500 mM NaCl and 50 mM Tris pH 7.5 or HPS-P+ buffer (10 mM HEPES pH 7.4, 150 mM NaCl, 1 μ M NiSO₄, 0.005% v/v surfactant P20; GE Healthcare) supplemented with 100 nM venetoclax or 5 μ M actinonin if needed. Fc-tagged proteins were immobilized at a concentration of 7 μ g ml⁻¹ on protein A probes (1.5 to 2.5 nm immobilized) and dipped into serial dilutions of the ligand. Steady-state responses were normalized with the maximum value and plotted using a nonlinear four-parameter curve-fitting analysis.

Grating-coupled interferometry

Grating-coupled interferometry measurements were performed on a Creoptix WAVE system (Malvern Panalytical) using Creoptix WAVE control software (Malvern Panalytical, v.4.5.18). The running buffer consisted of HPS-P+ buffer (10 mM HEPES pH 7.4, 150 mM NaCl, 0.005% v/v surfactant P20; GE Healthcare). All protein targets were immobilized on a 4PCH chip (Malvern Panalytical) by means of amine coupling to reach 7,000–10,000 pg mm⁻². An intermediate injection with 1 μ M NiSO₄ was used for PDF1 protein. S55746, OBz-Pro and TBDMS-Act were then injected sequentially as analytes at concentrations of 2, 2.5 and 5 μ M, respectively, using the waveRAPID (repeated analyte pulses of increasing duration) kinetic assay⁷⁴. The flow rate was 100 μ l min⁻¹ for an injection duration of 25 s followed by 300 s of dissociation time for TBDMS-Act, whereas an injection duration of 50 s followed by 600 s of dissociation time was used for S55746 and OBz-Pro. Measurements were fitted with either a 1:1 model (for Bcl2–S55746 and PDF1–TBDMS-Act) or with a mass transport model (for BD3–OBz-Pro).

Size-exclusion chromatography combined with multiangle light scattering

Size-exclusion chromatography combined with multiangle light scattering (miniDAWN TREOS, Wyatt) was performed to determine the molecular weights of the purified designs. The final concentration was approximately 1 mg ml⁻¹ in PBS (pH 7.4), and 100 μ l of the sample was injected into a Superdex 75 10/300 GL column (GE Healthcare) with a flow rate of 0.5 ml min⁻¹. Ultraviolet absorbance at 280 nm, differential refractive index and light scattering signals were recorded. Molecular weight was determined using ASTRA software (v.6.1, Wyatt).

Circular dichroism

Far-ultraviolet circular dichroism spectra were obtained with a Chirascan spectrometer (AppliedPhotophysics). Protein samples were diluted in PBS at a protein concentration of 300 μ g ml⁻¹ and placed in 1-m path-length cuvettes. Wavelengths between 200 nm and 250 nm were recorded with a scanning speed of 20 nm min⁻¹ and a response time of 0.125 s. All spectra were corrected for buffer absorption. Temperature ramping melts were performed from 20 to 90 °C with an increment of 2 °C min⁻¹. Thermal denaturation curves were plotted by the change in ellipticity at the global curve minimum. If possible, melting temperatures were determined after fitting the data with a sigmoid curve equation in GraphPad Prism.

Cell transfection and induction

Human embryonic kidney cells (HEK293T; Invitrogen, R70007) were cultured in Dulbecco's modified Eagle medium (DMEM; 41966-029,

Article

Gibco) supplemented with 10% (v/v) fetal bovine serum (FBS; A5256701, Gibco) and 1% (v/v) penicillin–streptomycin (15140-122, Gibco). Cells were authenticated by the provider (STR genotyping) and tested negative for mycoplasma contamination (quantitative PCR). Cells were maintained at 37 °C with 5% CO₂ and passaged every 2–3 days at around 80% confluence. Cells were seeded into the inner 60 wells of a 96-well plate at 10,000 cells per well 24 h before transfection. Cells were transfected by layering 50 µl from a mixture of 330 µl DMEM, 825–850 ng total DNA and 4.125 µg polyethylenimine (24765-1, Polysciences) on top of the medium in each well, enough for each six-well column with a 10% extra margin, as described previously⁷⁵. Cells were left to incubate overnight, for a minimum of 12 h. The next morning, medium was replaced with fresh medium including the respective dilutions of the inducing agent.

Cellular detection assay

In the secreted NanoLuc assays, cells were seeded into clear 96-well cell culture plates (655-180, Greiner Bio-One) and transfected the next day. In the venetoclax-induced GEMS assay, cells were transfected with STAT3 (100 ng), STAT3-NanoLuc reporter (150 ng), and either a single GEMS receptor chain containing Bcl2 or DBVen1619_2 (600 ng) or both chains together (300 ng each). In the secreted split NanoLuc progesterone-induced assay (Extended Data Fig. 4), cells were transfected with 412.5 ng of scFv-DB3(V_H/V_H)-N-term-NanoLuc and 412.5 ng of DBPro1156_2-C-term-NLuc or 825 ng of a single plasmid. Cells were induced with their respective agent the following day. After 24 h of induction, 5 µl medium was transferred to a black 384-well plate (3820, Corning) and mixed with 5 µl diluted substrate from the Nano-Glo Luciferase Assay kit (N1120, Promega). After gentle shaking, plates were measured on a Tecan Spark plate reader with an integration time of 1,000 ms.

For intracellular NanoLuc assays, cells were plated in black 96-well cell culture plates (655086, Greiner). The next day, cells were transfected with either a single chain of PDF1-C-term-NanoLuc or DBAct553_1-N-term-NanoLuc (825 ng) or both chains together (412.5 ng each). The following day, cells were induced with different dilutions of the inducing agent actinonin. After 24 h of induction, intracellular nanoluciferase activity was measured using a Nano-Glo Live Cell Assay kit (N2012, Promega). Medium was aspirated and replaced with 24 µl RPMI medium (52400-025, Gibco) containing 10% v/v FBS, and 6 µl diluted substrate was added to each well. After gentle shaking, plates were measured on a Tecan Spark plate reader with an integration time of 1,000 ms. All cell-based fits presented in Fig. 5 and Extended Data Fig. 4 were calculated from technical replicates ($n = 3$) using a nonlinear four-parameter curve-fitting analysis. All statistical analyses were based on two-way analysis of variance (ANOVA) with multiple comparisons.

Cell-free reporter system

The gene encoding the 6xHis-DBPro1156_2 protein fused to T7 RNA polymerase (*T7RNAP*) was cloned into a pQE30 plasmid using Gibson assembly. The plasmid was then transformed into NEBExpress Iq competent *E. coli* (NEB, C30371) for protein expression. Bacteria were precultured overnight and inoculated to a 500 ml Luria-Bertani (LB)-medium culture, grown until the OD₆₀₀ was approximately 0.7 and then induced with 0.1 mM IPTG for 3 h. The cells were collected by centrifugation at 4,000g and lysed by sonication. Proteins were purified using Ni-NTA IMAC Sepharose gravity columns.

The ZF438-DB3 scFv (V_H/V_L) fusion protein was expressed using a PURExpress kit from NEB (E6800S) with the addition of a disulfide bond enhancer (E6820S). The reaction volume was 10 µl, containing 4 µl of solution A, 3 µl of solution B, 0.4 µl of NEB disulfide bond enhancer 1, 0.4 µl of NEB disulfide bond enhancer 2, 2 µl of DNA template (10 ng µl⁻¹) and 0.2 µl of water. The reaction was incubated at 34 °C for 3 h and used for the following reporter reaction.

A PURExpress kit from NEB (E6800S) with disulfide bond enhancer (E6820S) was used to set up the mCherry reporter expression as well.

The reporter-expressing reaction also included 100 nM purified DBPro1156_2-T7RNAP and ZF438-DB3 scFv pre-expressed with PURExpress. The DNA template for the mCherry gene was set to 4 nM, and the mCherry gene was transcribed under the regulation of a truncated T7 promoter downstream of the zinc-finger 438 protein binding site, which requires a zinc-finger protein for activation of transcription. Progesterone was dissolved in 2% dimethyl sulfoxide. Then, 10-µl reactions with different conditions were loaded into a 384-well plate. The mCherry fluorescence intensity was measured on a BioTek Synergy HI Multimode Reader (Agilent) with an excitation wavelength of 565 nm and an emission wavelength of 615 nm at 34 °C for 8 h with 2-min intervals. All fits presented in Fig. 5 and Extended Data Fig. 4 were calculated from technical replicates ($n = 3$) using a nonlinear four-parameter curve-fitting analysis. All statistical analyses were based on two-way ANOVA with multiple comparisons.

Retrovirus production and primary murine T cell transduction

Retrovirus production and transduction of activated primary murine T cells were carried out as previously described⁷⁶. Briefly, Phoenix-ECO cells (ATCC, CRL-3214) were seeded in a T125 flask and, after 48 h, transfected with polyethylenimine and plasmid mix. Cells were authenticated by the provider (STR genotyping) and tested negative for mycoplasma contamination (MycoAlert Mycoplasma Detection Kit, LT07-318). At 48 and 72 h after transfection, the supernatant containing the virus was collected, mixed, filtered through a 0.45-µm filter, concentrated using ultracentrifugation (24,000g, 2 h, 4 °C) and then stored at -80 °C.

Primary murine T cells were isolated from C57BL/6 mouse spleens using a specific isolation kit (Miltenyi Biotec, 130-095-130) and cultured in T cell medium (RPMI 1640 medium supplemented with GlutaMAX, 10% (v/v) FBS, 100 U ml⁻¹ penicillin, 100 µg ml⁻¹ streptomycin sulfate, 1 mM sodium pyruvate, 50 µM 2-mercaptoethanol). Primary murine T cells were tested negative for mycoplasma (MycoAlert Mycoplasma Detection Kit, LT07-318). Cells were activated using αCD3/CD28 activation beads (11452D, Gibco) at a cell concentration of 0.5 × 10⁶ cells ml⁻¹ in T cell medium supplemented with 50 IU ml⁻¹ of human IL-2 (200-02, PeproTech). Retroviruses were added to plates precoated with protamine and spun at 2,000g for 1.5 h at 32 °C. Activated T cells (0.5 × 10⁶ cells per well) were transferred to each well. T cells were passaged 48 h posttransduction and maintained at 0.5 × 10⁶ cells ml⁻¹ in T cell medium supplemented with 10 ng ml⁻¹ of human IL-7/IL-15 (200-7/200-15, PeproTech). Transduction efficiency was assessed by flow cytometry by measuring binding of a biotinylated HER2 protein (AcroBiosystems, HE2-H822R; 1:100 dilution) labelled with PE-conjugated streptavidin (Invitrogen, 12-4317-87; 1:100 dilution). For the transduction efficiency of the double chain, the chain containing FLAG-tagged Bcl2 and αHER2 was labelled with an A647-conjugated anti-FLAG antibody (Thermo Fisher, MA1-142-A647; 1:100 dilution), and the chain containing V5-tagged DBVen1619 was labelled with a fluorescein isothiocyanate (FITC)-conjugated anti-V5 antibody (GeneTex, GTX21209; 1:100 dilution).

Cytotoxicity assay of murine CAR-T cells

On day 10 after transduction, untransduced T cells, 2G-CAR-T cells and split CID-CAR-T cells (10 × 10⁴) were cocultured with HER2-transduced MC38 mouse colon cancer cells (MC38-HER2; provided by L. Tang at EPFL) with an effector to target cell ratio of 1:1 in 96-well flat-bottomed plates. The number of CAR-positive cells was normalized to match the lowest transduction efficiency of the CID-CAR-T cells (Supplementary Fig. 16) by adding untransduced cells to achieve the same number of CAR-positive cells and maintain a consistent total cell count per well. Cytotoxicity activity of CAR-T cells was monitored for 48 h at different inducer concentrations. Target cells were labelled using Incucyte NuLight Red to enable real-time counting of viable tumour cells with the IncuCyte live cell imaging system. All cell-based data presented

in Fig. 5 were calculated from biological replicates ($n = 3$) and fitted using a nonlinear four-parameter curve-fitting analysis. MC38-HER2 cells were tested negative for mycoplasma (MycoAlert Mycoplasma Detection Kit, LT07-318).

Protein purification for crystallography

The 6xHis-tagged PDF1 from *P. aeruginosa* and DBAct553_1 were expressed in *E. coli* (BL21 T7 Express). Amino acid sequences of both proteins are shown in Supplementary Table 5. For PDF1, cells were grown in LB medium supplemented with 100 mM NiSO₄ up to an OD₆₀₀ of 0.7 at 37 °C, then induced with 1 mM IPTG and allowed to continue to grow overnight at 18 °C. For DBAct553_2, cells were grown in auto-induction medium up to an OD₆₀₀ of 0.7 at 37 °C and then overnight at 18 °C. Cells were collected by centrifugation at 4,000g for 10 min, resuspended in lysis buffer (50 mM Tris, pH 7.5, 500 mM NaCl, 5% glycerol, 1 mg ml⁻¹ lysozyme, 1 mM PMSF and 1 µg ml⁻¹ DNase) and lysed by sonication. Lysates were then clarified by centrifugation at 30,000g for 30 min and filtered. Proteins were purified using an ÄKTA Pure system (GE Healthcare) Ni-NTA HisTrap affinity column, followed by size-exclusion chromatography on a Superdex HiLoad 16/600, 75 µg, with Tris-buffered saline (50 mM Tris pH 7.5, 250 mM NaCl, 10 µM NiSO₄) as a final buffer. PDF1, DBAct553_2 and actinonin were mixed at final concentrations of 35 µM, 105 µM and 300 µM, respectively, and incubated on ice for 1 h. Proteins were then concentrated by centrifugation before crystallization.

Crystallographic data collection and structure determination

The actinonin-bound PDF1-DBAct553_1 complex (5 mg ml⁻¹) was crystallized using a sitting-drop vapour diffusion setup at 18 °C with 200 nl of protein and 200 nl crystallization solution consisting of 0.2 M sodium formate, 0.1 M sodium phosphate pH 6.2, 20% (v/v) PEG and 10% (v/v) glycerol. Crystals were cryoprotected with 25% glycerol and flash-cooled in liquid nitrogen. Diffraction data were collected at a temperature of 100 K at the European Synchrotron Radiation Facility (ESRF Grenoble, France). Raw data were processed and scaled with XDS (10 Jan. 2022, BUILT = 20220220) and then processed using the autoPROC package⁷⁷ (GlobalPhasing, v.20230222). Phases were obtained by molecular replacement using the Phaser module of the Phenix package (v.1.20.1-4487) and a model from PDB 1LRY in complex with our designed binder DBAct553_1 (ref. 78). Atomic model adjustment and refinement were completed using COOT (v.0.9.5) and Phenix.refine^{79,80} (v.1.20.1-4487). Finally, MolProbity⁸¹ (v.4.5.1) was used to assess the quality of the refined model. Details of data collection and refinement statistics are shown in Extended Data Table 1.

Cryo-EM preparation and data acquisition

A chimeric DB3 Fab (Supplementary Table 5) was produced using the Expi293 expression system from Thermo Fisher Scientific (A14635). An anti-kappa light chain Fab⁸² (Supplementary Table 5) was produced using ExpiCHO-S cells (Thermo Fisher Scientific, A29127) growing in a ProCHO-5 medium (Lonza) supplemented with 2% dimethyl sulfoxide. Supernatants were collected 6 and 7 days, respectively, after transfection and filtered and purified by Ni-NTA affinity chromatography, followed by size-exclusion chromatography on a Superdex HiLoad 16/600, 75 µg. All proteins were concentrated in PBS as a final buffer. DBPro1156_2 was purified as described previously ('Protein expression and purification').

DB3 Fab, anti-kappa light chain Fab, DBPro1156_2 and progesterone were mixed with a molar ratio of 1:0.9:3:2, supplemented with 0.1% *n*-dodecyl-β-D-maltoside and concentrated to 3.87 mg ml⁻¹. Proteins were applied to a glow-discharged 300-mesh holey carbon grid (Au 1.2/1.3, Quantifoil Micro Tools), blotted for 4 s at 95% humidity, 10 °C, plunge-frozen in liquid ethane (Vitrobot, Thermo Fisher Scientific) and stored in liquid nitrogen. Data collection was performed with automation program EPU (Thermo Fisher Scientific, v.2.12.1) on a 300 kV FEI

Titan Krios G4 microscope equipped with a FEI Falcon IV detector. Micrographs were recorded at a calibrated magnification of ×120,000 with a pixel size of 0.658 Å and a nominal defocus ranging from -1.0 µm to -1.7 µm.

Cryo-EM image processing

Acquired cryo-EM data were processed (Supplementary Fig. 12) using cryoSPARC (v.4.4.1). Gain-corrected micrographs were imported, and micrographs with a resolution estimation worse than 5.5 Å were discarded after patch contrast transfer function estimation. A total of 16,038 micrographs were used for this complex. Initial particles were picked using a blob picker with 90–150-Å particle size. Particles were extracted with a box size of 360 × 360 pixels, downsampled to 140 × 140. After two-dimensional classification, clean particles were used for ab initio three-dimensional reconstruction. After several rounds of three-dimensional classification, the class with most detailed features was reextracted using full box size and subjected to non-uniform and local refinement to generate high-resolution reconstructions. The local resolution was calculated and visualized using ChimeraX⁸³ (v.1.3, UCSF).

For structure building, we used ColabFold⁵⁰ repredictions of the anti-kappa and DB3 Fabs, as well as the designed binder. Subsequent manual model adjustment and refinement were completed using Coot⁷⁹ (v.0.9.5). Atomic model refinement was performed using Phenix.real_space_refine⁸⁰ (v.1.20.1-4487). The quality of the refined model was assessed using MolProbity⁸¹ (v.4.5.1). Structural figures were generated using PyMOL (v.2.4, Schrödinger). The refined atomic models and corresponding cryo-EM maps were deposited under PDB accession code 9FKD and EMDB accession code EMD-50522. Details of data collection and refinement statistics are shown in Extended Data Table 2.

Chemical synthesis

All chemical reagents and solvents for synthesis were purchased from commercial suppliers (Sigma-Aldrich, Fluka, Acros) and were used without further purification or distillation. The composition of mixed solvents is given as a volume ratio (v/v). The ¹H NMR spectra were recorded on a Bruker DPX 400 (400 MHz for ¹H) with chemical shifts (δ) reported in ppm relative to the solvent residual signals (7.26 ppm for CDCl₃; 3.31 ppm for MeOD) (Supplementary Fig. 17). Coupling constants are reported in Hz. Liquid chromatography coupled with mass spectrometry (LC-MS) was performed on a Shimadzu MS2020 connected to a Nexera UHPLC system equipped with a Waters ACQUITY UPLC BEH Phenyl 1.7 µm 2.1 × 50 mm column. Buffer A consisted of 0.05% HCOOH in H₂O; buffer B was 0.05% HCOOH in acetonitrile. The liquid chromatography gradient was as follows: 10% to 90% B within 6.0 min with 0.5 ml min⁻¹ flow. Preparative high-performance liquid chromatography (HPLC) was performed on a Dionex system equipped with an UltiMate 3000 diode array detector for product visualization on a Waters SymmetryPrep C18 column (7 µm, 7.8 × 300 mm). Buffer A consisted of 0.1% v/v trifluoroacetic acid in H₂O; buffer B was acetonitrile. The gradient was from 25% to 90% B within 30 min with 3 ml min⁻¹ flow.

19-O-benzoyl-progesterone. First, 19-hydroxyprogesterone (2.0 mg, 6.1 µmol, 1 eq.) was dissolved in pyridine (0.5 ml); then, benzoyl chloride (0.9 µl, 7.9 µmol, 1.3 eq.) was added. The reaction mixture was stirred for 3 h. LC-MS analysis showed reaction completion, and 10 µl methanol was added. After 30 min, the solvents were evaporated under reduced pressure. The residue was dissolved in a minimum of acetonitrile and subjected to preparative HPLC. The fractions containing the product were pooled and lyophilized. The yield was 1.1 mg (41%). ¹H NMR (400 MHz, CDCl₃) δ 7.89 (d, *J* = 8.4 Hz, 2H), 7.56 (t, *J* = 7.4 Hz, 1H), 7.42 (t, *J* = 7.8 Hz, 2H), 5.98 (s, 1H), 4.81 (d, *J* = 11.3 Hz, 1H), 4.46 (d, *J* = 11.3 Hz, 1H), 2.68 (ddd, *J* = 17.0, 13.8, 5.9 Hz, 1H), 2.57–2.32 (m, 4H), 2.26–2.06 (m, 5H), 2.03–1.63 (m, 6H), 1.55–1.37 (m, 2H), 1.36–1.06 (m, 5H), 0.69 (s, 3H). HRMS (ESI/QTOF) *m/z*: [M + H]⁺ calcd for C₂₈H₃₅O₄⁺ 435.2530; found 435.2528.

Article

TBDMS-Act. Actinonin (2.0 mg, 5.2 μmol , 1 eq.) and 4-dimethylaminopyridine (3.8 mg, 31.2 μmol , 6 eq.) were suspended in dichloromethane (0.5 ml). TBDMS-Cl (2.5 mg, 16.6 μmol , 3.2 eq.) was added, and the reaction was stirred for 5 h at room temperature. The solvent was evaporated under reduced pressure, the residue was dissolved in MeOH (0.5 ml), water (50 μl) was added, and the reaction was heated to 60 °C for 5 h. The solvents were evaporated again, and the residue was dissolved in a minimum of dichloromethane and subjected to preparative thin-layer chromatography using dichloromethane/MeOH 9:1 as the eluent. The yield was 2.0 mg (77%). $^1\text{H NMR}$ (400 MHz, MeOD) δ 4.38 (d, J = 8.5 Hz, 1H), 4.13 (s, 1H), 3.89 (dt, J = 10.0, 6.8 Hz, 1H), 3.79 (dd, J = 9.9, 5.3 Hz, 1H), 3.68 (dd, J = 9.9, 2.8 Hz, 1H), 3.63–3.42 (m, 1H), 2.83–2.75 (m, 1H), 2.34 (dd, J = 14.5, 8.0 Hz, 1H), 2.24–1.84 (m, 6H), 1.67–1.48 (m, 1H), 1.46–1.18 (m, 6H), 1.02–0.94 (m, 7H), 0.93–0.86 (m, 12H), 0.07 (s, 3H), 0.05 (s, 3H). HRMS (ESI/QTOF) m/z : $[\text{M}+\text{Na}]^+$ calcd for $\text{C}_{25}\text{H}_{49}\text{N}_3\text{NaO}_5\text{Si}^+$ 522.3334; found 522.3342.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The crystal structure of DBAct553_2 in complex with actinonin-bound PDF1 has been deposited at the PDB under accession code 8S1X. The refined atomic models and corresponding cryo-EM maps of DBPro1156_2 in complex with progesterone-bound DB3 Fab were deposited under PDB accession code 9FKD and EMDB accession code EMD-50522. The scaffold database generated for grafting the seed provided by MaSIF-neosurf is available in part at Zenodo (<https://zenodo.org/records/7643697>)⁸⁴ and in part at GitHub (<https://github.com/strauchlab/DBP> and https://github.com/strauchlab/scaffold_design/). Data used to generate Figs. 1–5, Extended Data Figs. 1–8 and Supplementary Figs. 1–17, as well as the relevant plasmid maps, are available at Zenodo (<https://zenodo.org/records/13737922>)⁸⁵.

Code availability

MaSIF-neosurf and the Rosetta design scripts are available at GitHub (<https://github.com/LPDI-EPFL/masif-neosurf>).

- Sanner, M. F., Olson, A. J. & Spohner, J. C. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers* **38**, 305–320 (1996).
- Jurru, E. et al. Improvements to the APBS biomolecular solvation software suite. *Protein Sci.* **27**, 112–128 (2018).
- Landrum, G. *RDKit: open-source cheminformatics* <https://www.rdkit.org/> (2021).
- Degen, J., Wegscheid-Gerlach, C., Zaliani, A. & Rarey, M. On the art of compiling and using ‘drug-like’ chemical fragment spaces. *ChemMedChem* **3**, 1503–1507 (2008).
- Hu, L., Benson, M. L., Smith, R. D., Lerner, M. G. & Carlson, H. A. Binding MOAD (Mother Of All Databases). *Proteins* **60**, 333–340 (2005).
- Liu, Z. et al. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* **31**, 405–412 (2015).
- Coventry, B. & Baker, D. Protein sequence optimization with a pairwise decomposable penalty for buried unsatisfied hydrogen bonds. *PLoS Comput. Biol.* **17**, e1008061 (2021).
- Silva, D.-A., Correia, B. E. & Procko, E. in *Computational Design of Ligand Binding Proteins* Vol. 1414 (ed. Stoddard, B. L.) 285–304 (Springer New York, 2016).
- Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
- Rocklin Gabriel, J. et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **357**, 168–175 (2017).

- Bhardwaj, G. et al. Accurate de novo design of hyperstable constrained peptides. *Nature* **538**, 329–335 (2016).
- Tobin, A. R. et al. Inhibition of a malaria host–pathogen interaction by a computationally designed inhibitor. *Protein Sci.* **32**, e4507 (2023).
- Linsky, T. W. et al. Sampling of structure and sequence space of small protein folds. *Nat. Commun.* **13**, 7151 (2022).
- Kartal, Ö., Andres, F., Lai, M. P., Nehme, R. & Cottier, K. waveRAPID—a robust assay for high-throughput kinetic screens with the Creoptix WAVESystem. *SLAS Discov.* **26**, 995–1003 (2021).
- Scheller, L. in *Mammalian Cell Engineering* Vol. 2312 (ed. Kojima, R.) 15–33 (Springer US, 2021).
- Lanitis, E. et al. Optimized gene engineering of murine CAR-T cells reveals the beneficial effects of IL-15 co-expression. *J. Exp. Med.* **218**, e20192203 (2021).
- Kabsch, W. XDS. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125–132 (2010).
- Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzym.* **276**, 307–326 (1997).
- Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).
- Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).
- Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 12–21 (2010).
- Macdonald, L. E. et al. Kappa-on-Heavy (KoH) bodies are a distinct class of fully-human antibody-like therapeutic agents with antigen-binding properties. *Proc. Natl Acad. Sci. USA* **117**, 292–299 (2020).
- Goddard, T. D. et al. UCSF ChimeraX: meeting modern challenges in visualization and analysis. *Protein Sci.* **27**, 14–25 (2018).
- Gainza, P. et al. De novo design of site-specific protein interactions with learned surface fingerprints. *Zenodo* <https://doi.org/10.1101/2022.06.16.496402> (2023).
- Marchand, A. et al. Targeting protein-ligand neosurfaces with a generalizable deep learning tool. *Zenodo* <https://doi.org/10.5281/zenodo.13737921> (2024).

Acknowledgements We thank the staff at PTPSP at EPFL, F. Pojer, K. Lau, A. Larabi, L. Durrer and S. Quinche for advice on the biophysical characterization of proteins and work on structural validations; the staff at the Dubochet Center for Imaging (DCI) in Lausanne for cryo-EM data collection and processing; SCITAS at EPFL for support with the computational simulations; the staff at GECF for assistance with deep sequencing and members of FCCS for assistance with fluorescence-activated cell sorting. We also thank A. Moretti from Malvern Panalytical for his support in running grating-coupled interferometry on Creoptix Wave and for providing access to the instrument; E.-M. Strauss for providing de novo hyperstable proteins for our scaffold database; L. Tang for his help with isolation and enrichment of T cells from mouse spleen; and N. Thomä and K. Lau for their feedback on the manuscript. This work was supported by Swiss National Science Foundation grants 310030_197724 (B.E.C., A.M., M.E.), TMGC-3_213750 (B.E.C., S.B.) and 200020_214843 (P.-W.L., S.J.M.); National Center of Competence in Research in Molecular Systems Engineering grant 182895 (B.E.C. and A.M.); National Center of Competence in Research in Catalysis grant 180544 (P.S.); EPSRC Turing AI World-Leading Research Fellowship No. EP/X040062/1 (M.B.); Microsoft Research AI4Science (B.E.C. and A.S.); VantAI (R.M.N.); Huawei Technologies Düsseldorf (B.E.C., L.S.); Reprodivac grant SEFRI 22.00135 (B.E.C., E.E.); H2020 Marie Skłodowska-Curie EPFL-Fellows grant (P.G.) and the “Peter und Traudl Engelhorn Stiftung” (M.P.).

Author contributions A.M., S.B. and A.S. contributed equally to this work. A.M. and B.E.C. led the project. A.M., S.B., P.-W.L. and M.E. performed the experimental work. A.M., S.B., M.P. and L.S. designed the experimental methodology. A.M., A.S. and Y.M. performed the computational work and protein design. A.M., A.S., P.G., E.E. and R.M.N. contributed to the design of the computational pipeline. M.P. solved the crystal structure. L.R. synthesized the small-molecule analogues. S.G. and J.S. participated in the expression and purification of proteins. P.S., S.J.M., M.B. and B.E.C. provided supervision and acquired the necessary funding. A.M., S.B., A.S., M.P. and B.E.C. wrote the manuscript with input from all authors.

Competing interests Ecole Polytechnique Fédérale de Lausanne (EPFL) has filed a patent application that incorporates findings presented previously in MaSIF-seed. P.G., A.M., M.B. and B.E.C. are named as coinventors on this patent (US Patent Office, US20230395187A1).

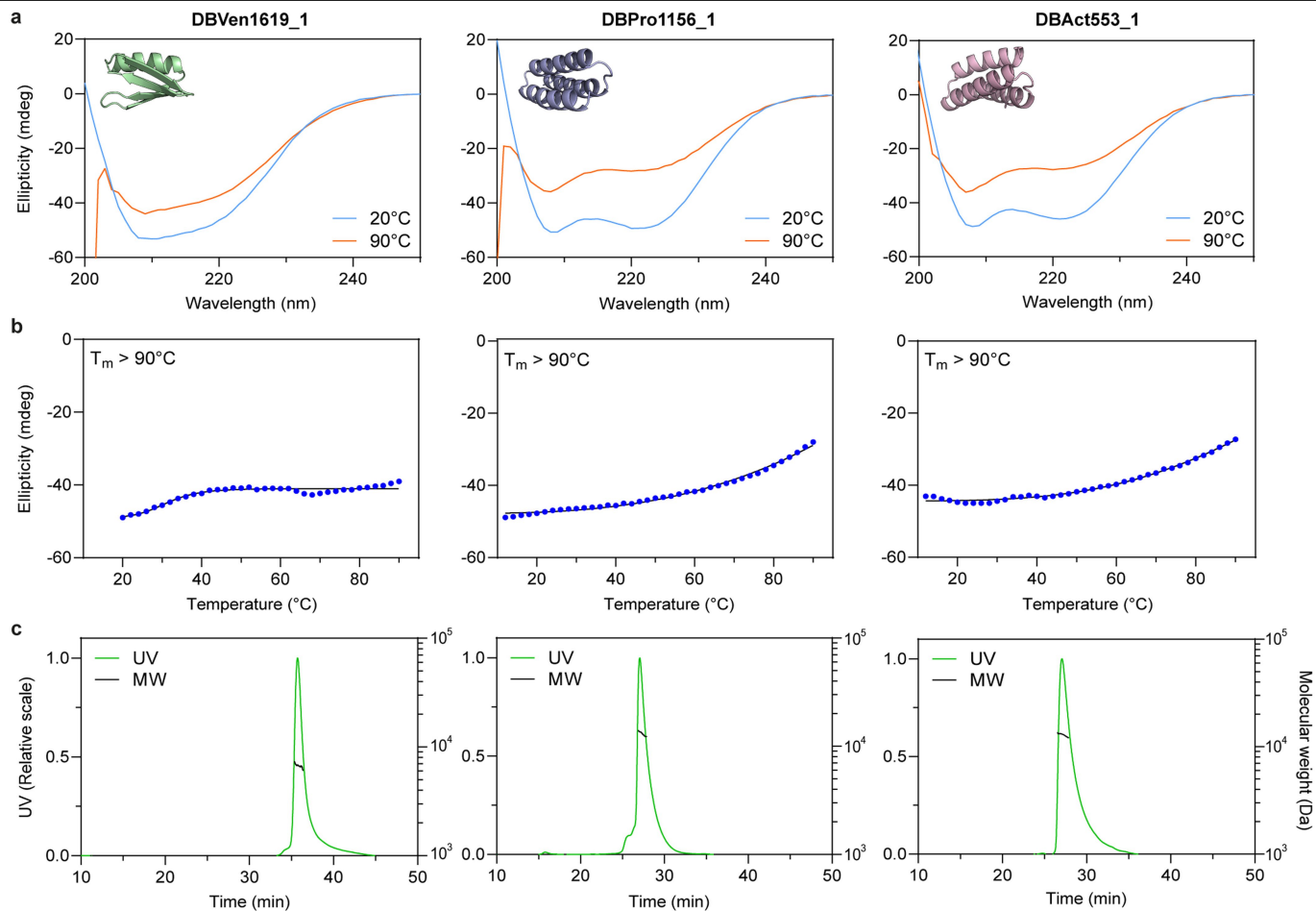
Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41586-024-08435-4>.

Correspondence and requests for materials should be addressed to Bruno E. Correia.

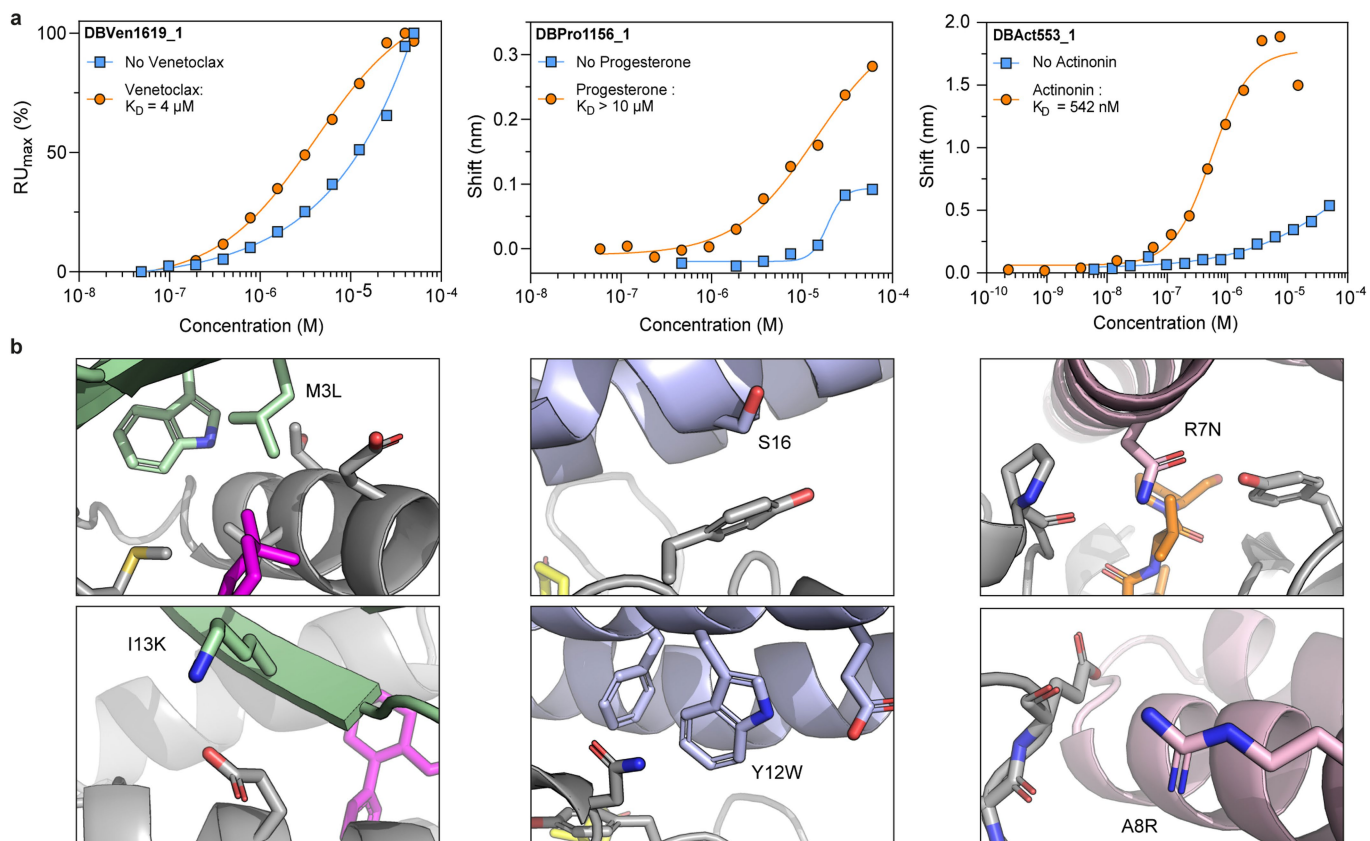
Peer review information Nature thanks Jiankun Lyu, Stuart Schreiber and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



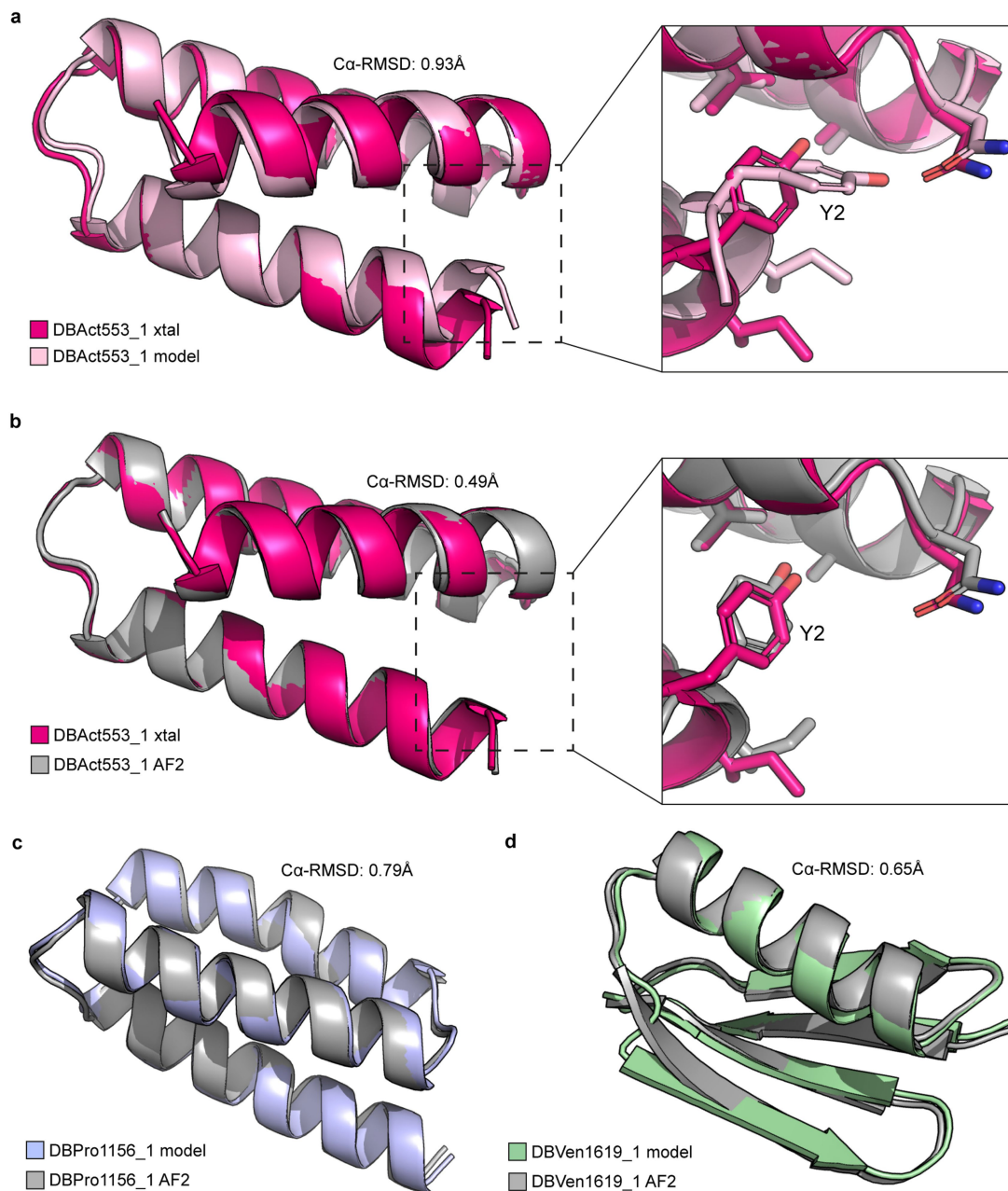
Extended Data Fig. 1 | Biophysical characterization of purified binders.
a. Protein folding of the purified binder measured by circular dichroism at 20 $^\circ\text{C}$ (blue) or 90 $^\circ\text{C}$ (orange). **b.** Thermal stability determined by measuring

the ellipticity at 218 nm at increasing temperature. **c.** Oligomeric state determined by size-exclusion multi-angle light scattering (SEC-MALS).



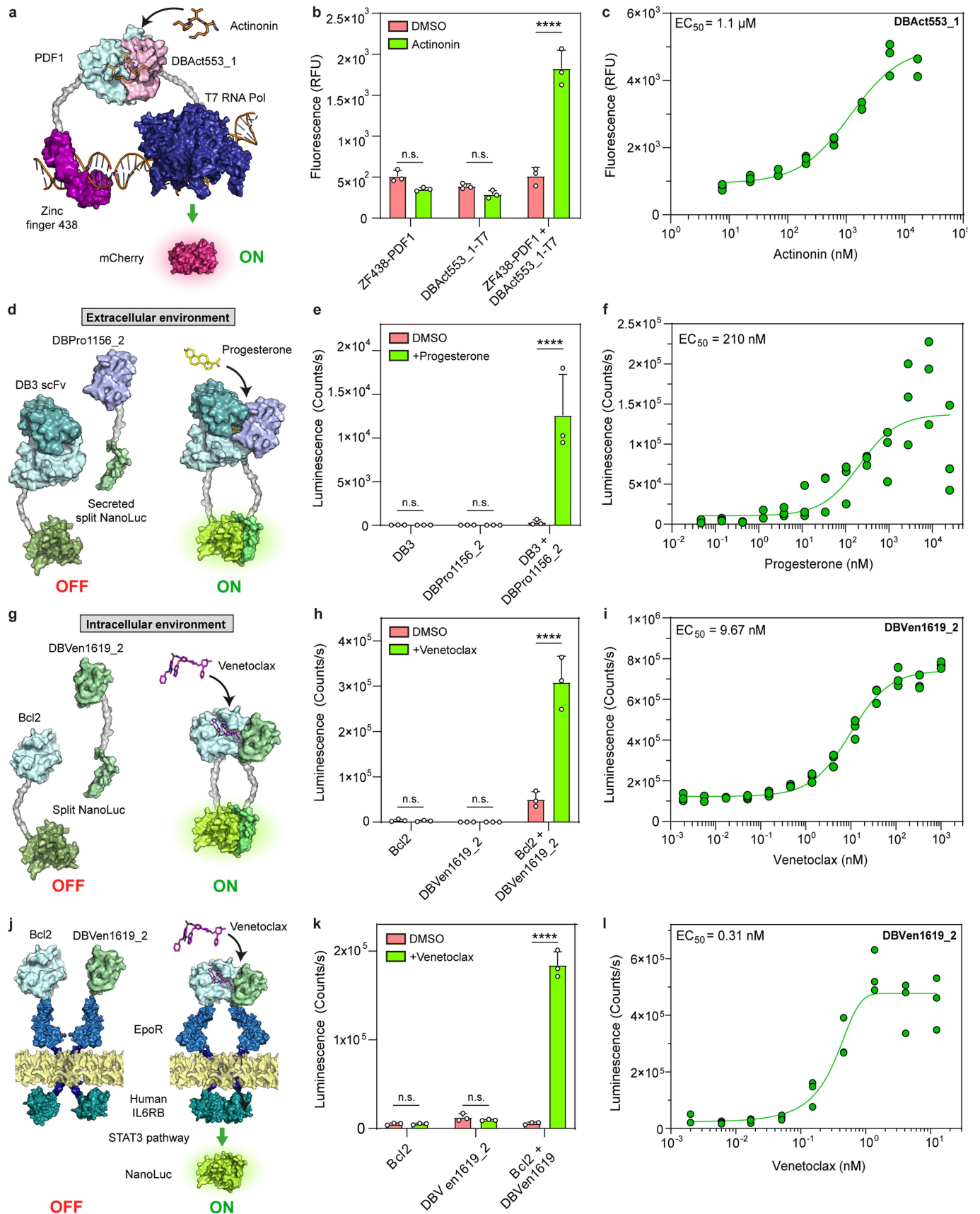
Extended Data Fig. 2 | Affinity measurements of first-generation binders and identified beneficial mutations. a. Affinity measurements for DBVen1619_1, DBPro1156_1 and DBAct553_1 performed by surface plasmon resonance (DBVen1619_1) or biolayer interferometry (DBPro1156_1 and DBAct1156_1). Each measurement was performed in presence (orange) or

absence (blue) of the respective small molecule. The fits were calculated using a nonlinear four-parameter curve fitting analysis. **b.** Computational models incorporating the beneficial mutations that improved the affinity of the designed binders. Target proteins are shown in gray and designed binders in their respective color.



Extended Data Fig. 3 | Comparison between crystallographic data and AlphaFold2 predictions. **a.** Computational model of DBAct553_1 (light pink) aligned with its crystal structure (magenta) with a close-up on tyrosine-2. **b.** AlphaFold2 (AF2) prediction of DBAct553_1 (gray) aligned with its crystal

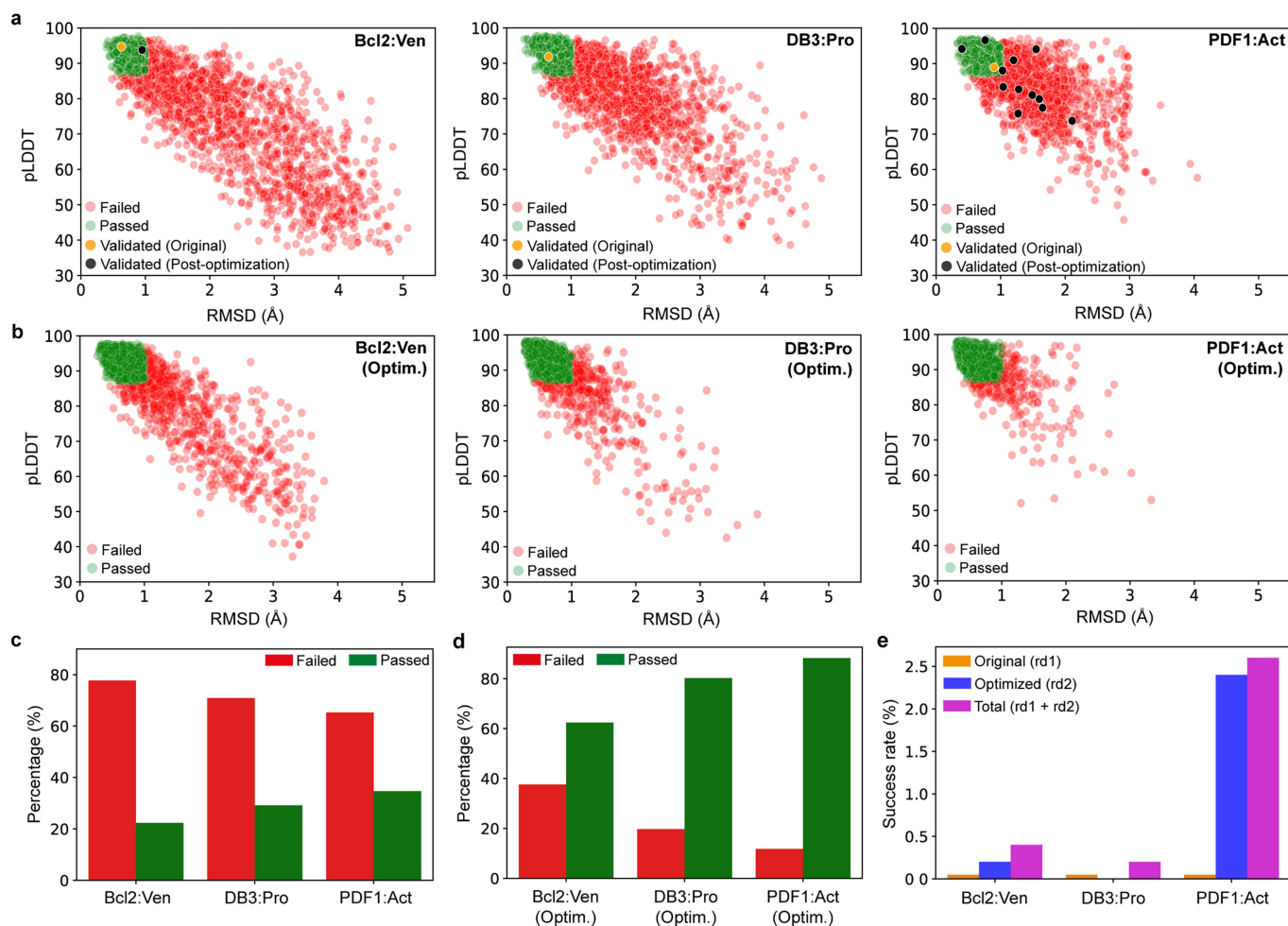
structure (magenta) with a close-up on tyrosine-2. **c-d.** Comparison between computational models of DBPro1156_1 (c) and DBVen1619_1 (d) and their respective AlphaFold2 prediction as monomers.



Extended Data Fig. 4 | See next page for caption.

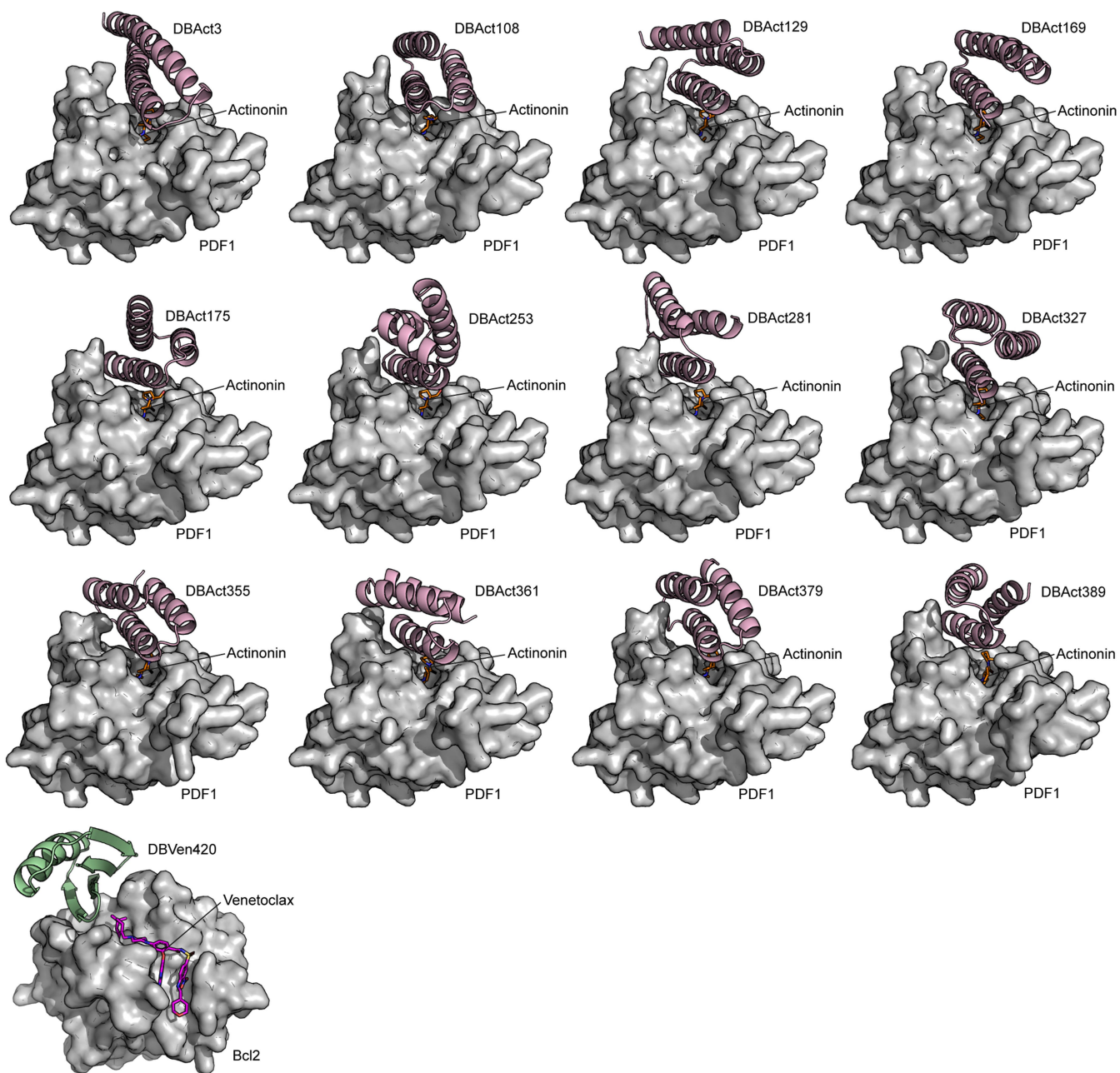
Extended Data Fig. 4 | Designed CIDs can be utilized in different cell-free and cell-based sensing schemes. **a.** Schematic of the cell free-expression system with PDF1 protein fused to a zinc finger transcription factor and DBAct553_1 fused to T7 RNA polymerase. **b.** Fluorescence (Relative fluorescence unit; RFU) measured in wells containing each monomeric component or mixed, without or with 16 μ M Actinonin. $p < 0.0001$ (****). **c.** Actinonin dose-dependent responses performed in a cell free system containing both components. **d.** Schematic of the extracellular split NanoLuc system functionalizing DB3 scFv and DBPro1156_2. **e.** Extracellular NanoLuc luminescence of HEK293T transfected with C-term split NanoLuc-fused DB3, N-term split NanoLuc-fused DBPro1156_2 or both together without or with 25 μ M Progesterone. **f.** Progesterone dose-dependent responses performed on HEK293T transfected with split-NanoLuc DB3 scFv and DBPro1156_2. $p < 0.0001$ (****). **g.** Schematic of the intracellular split NanoLuc system functionalizing Bcl2 and DBVen1619_2. **h.** Intracellular NanoLuc luminescence of HEK293T

transfected with C-term split NanoLuc-fused Bcl2 only, N-term split NanoLuc-fused DBVen1619_2 only or both together without or with 1 μ M Venetoclax. **i.** Venetoclax dose-dependent responses performed on HEK293T transfected with split-NanoLuc Bcl2 and DBVen1619_2. **j.** Schematic of the GEMS reporter system functionalizing Bcl2-based CID. Both protein components of the CID are individually fused to erythropoietin receptor (EpoR) chains linked to an intracellular human IL6RB domain, which induces the expression of a reporter gene (secreted NanoLuc luciferase) when activated. **k.** NanoLuc luminescence of HEK293T cells transfected with Bcl2-GEMS only, DBVen1619_2 only or both together without or with 12 nM Venetoclax. $p < 0.0001$ (****). **l.** Venetoclax dose-dependent responses performed on HEK293T transfected with Bcl2 and DBV1619 GEMS receptors. $p < 0.0001$ (****). Two-way ANOVA with Tukey's multiple comparison test, non-significant (ns). Barplots are presented as mean \pm standard deviations. Data points are derived from three technical replicates (a-i) or three biological replicates (j-l).

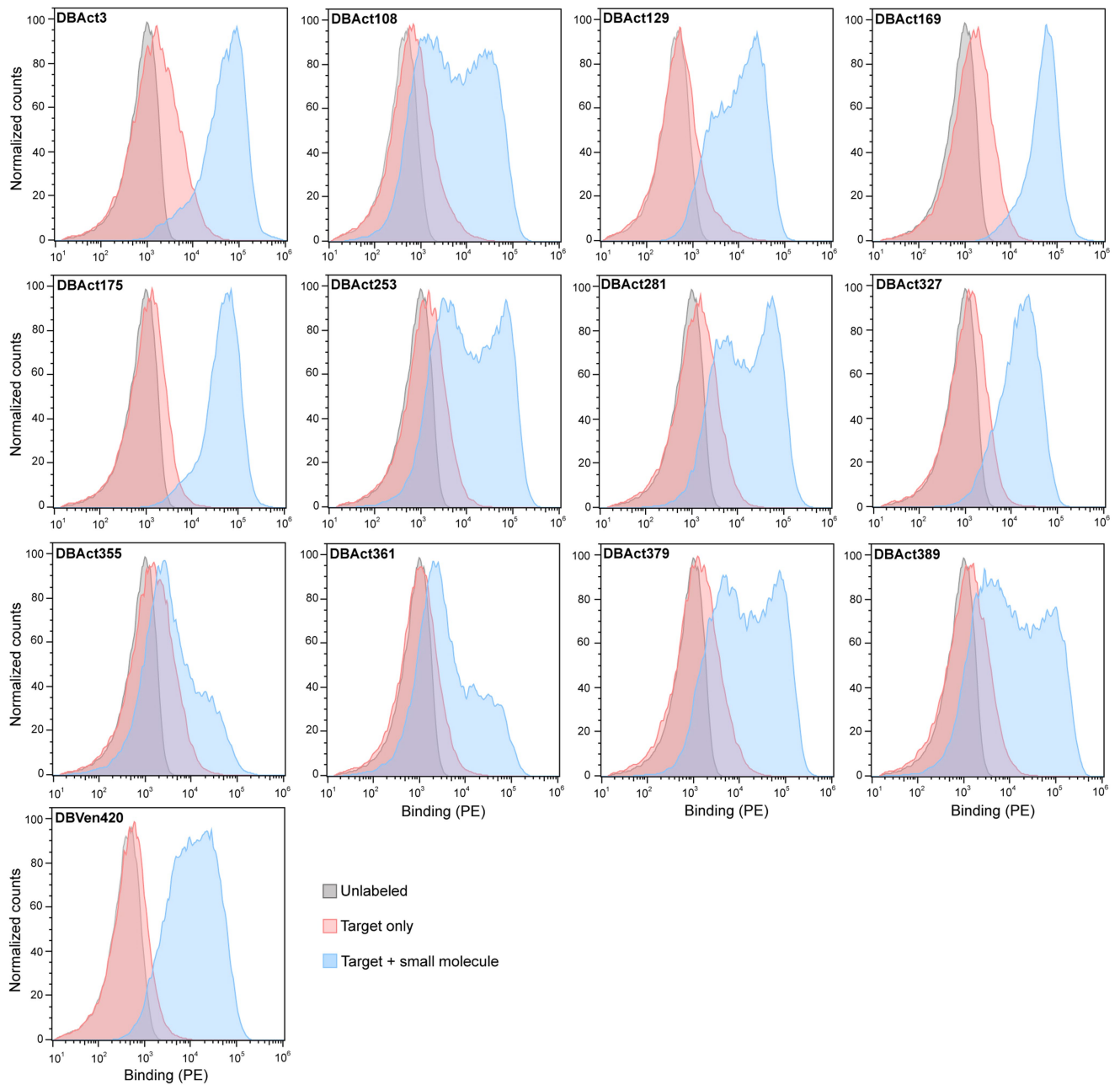


Extended Data Fig. 5 | AlphaFold filtering of original and LigandMPNN-optimized designs. a-b. AlphaFold monomer prediction (single sequence mode) of the ~2000 designs generated against each protein-ligand complex. Predictions were made with the original pool of designs (a) or after optimization with LigandMPNN (b). Prediction confidence (pLDDT) and root mean square deviation (RMSD) from the computational models are plotted. Designs that would pass a strict filtering ($\text{RMSD} \leq 1 \text{ \AA}$ and $\text{pLDDT} \geq 87$) are colored in green, while ones that failed filtering are colored in red. Validated binders are colored

in orange. Binders that were validated after optimization with LigandMPNN are colored in black **c-d.** Percentage of generated designs that failed (red) or passed (green) the strict AlphaFold2 filtering prior (c) and after LigandMPNN optimization (d). **e.** Experimental success rate obtained with the original pool of designs (round 1, ~2000 designs each, orange), with the LigandMPNN-optimized pool of designs (round 2, 500 designs each, blue) and the sum of all validated binders with the reduced selection (round 1 + 2, 500 designs each, pink).

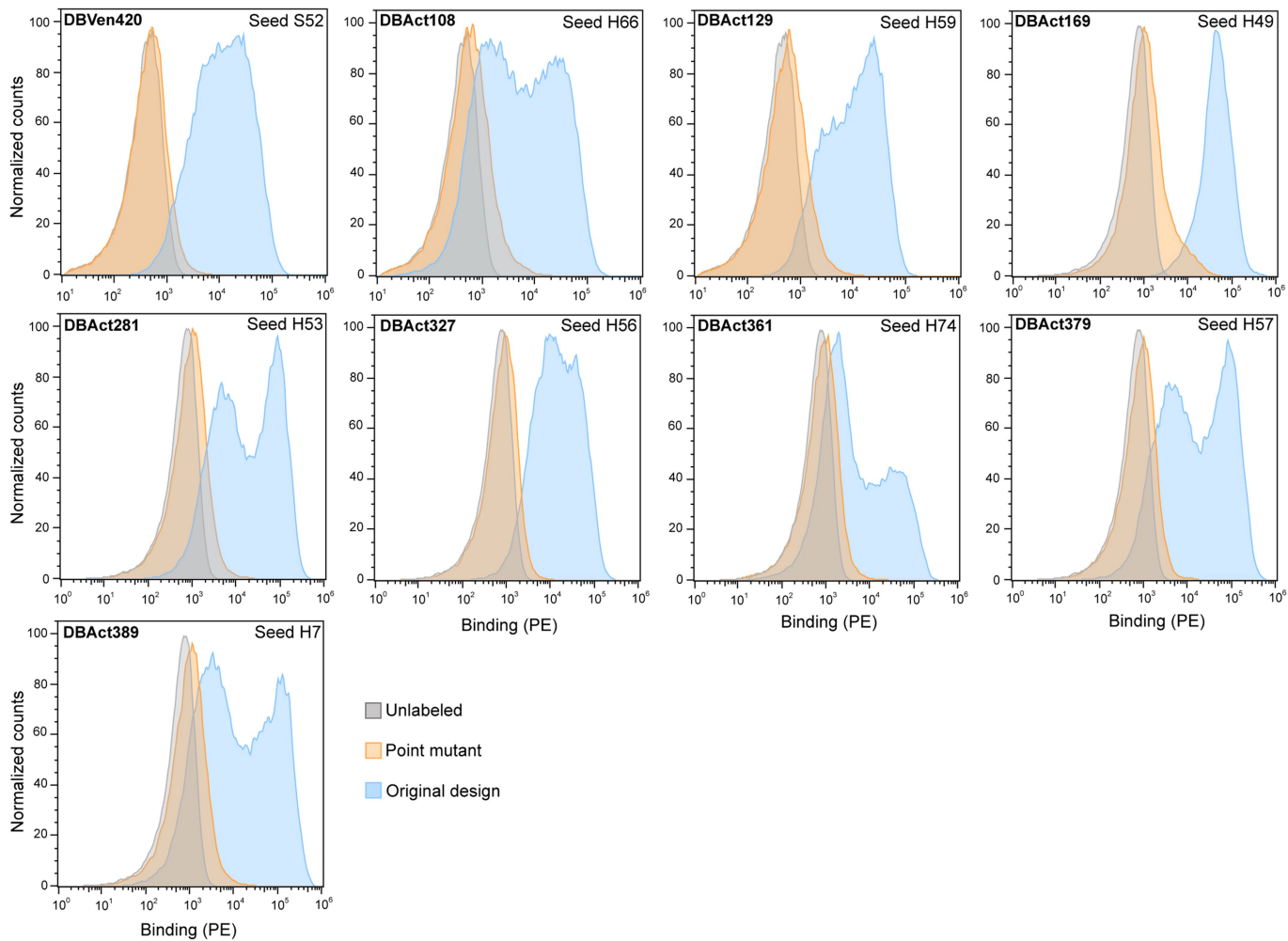


Extended Data Fig. 6 | Computational model of new binders screened with an optimized library. 12 new designs targeting PDF1:Actinonin complex (pink) and 1 targeting Bcl2:Venetoclax complex (green) were identified. Target proteins are colored in gray.



Extended Data Fig. 7 | Ligand-inducible binding of new binders screened with an optimized library. Histograms of the binding signal (PE, phycoerythrin) measured by flow cytometry on yeast displaying the designed binders. Yeast

were either unlabeled (gray) or labeled with 500 nM of their respective target protein preincubated with the ligand (blue), or with the target protein alone (red).



Extended Data Fig. 8 | Binding abrogation with point mutations at the designed interface. Histograms of the binding signal (PE, phycoerythrin) measured by flow cytometry on yeast displaying the designed binders (blue) or its point mutant (orange). All yeasts were labeled with 500 nM of the target

protein with the respective small molecule. Unlabeled controls are colored in gray. One design per identified seed was selected for a point mutation. Point mutants were respectively: L3R, A36R, A14R, A15R, A49R, A15R, A51R, A31R and L19R.

Article

Extended Data Table 1 | Crystallographic data collection and refinement statistics

<i>DBAct553_1:Actinonin:PDF1</i> (PDB: 8S1X)	
Data collection	
Space group	P 2 ₁ 2 ₁ 2 ₁
Cell dimensions	
<i>a</i> , <i>b</i> , <i>c</i> (Å)	49.44, 75.01, 83.16
<i>a</i> , <i>b</i> , <i>g</i> (°)	90.00, 90.00, 90.00
Resolution (Å)	55.70 - 1.88 (1.96 - 1.88)
R _{merge}	0.044 (1.125)
<i>I</i> / σ <i>I</i>	15.0 (1.3)
Completeness (%)	96.9 (99.6)
Redundancy	4.3 (4.4)
Refinement	
Resolution (Å)	55.70 - 1.88
No. reflections	24982 (2801)
R _{work} /R _{free}	0.1838/0.2030
No. atoms	
Protein	1850
Ligand/ion	73
Water	76
<i>B</i> -factors	
Protein	56.9
Ligand/ion	66.7
Water	55.3
R.m.s. deviations	
Bond lengths (Å)	0.007
Bond angles (°)	0.730

Extended Data Table 2 | Cryo-EM data collection and model validation statistics

*DBPro1156_2:Progesterone:DB3_Fab:
anti-Kappa:Fab (PDB: 9FKD, EMDB:
50522)*

Data collection and processing

Magnification	120,000
Voltage (kV)	300
Electron exposure (e ⁻ /Å ²)	60.0
Defocus range (μm)	-1.2-2.4
Pixel size (Å)	0.658
Symmetry imposed	C1
Initial particle images (no.)	4,048,193
Final particle images (no.)	192,055
Map resolution (Å)	3.23
FSC threshold	0.143
Map resolution range (Å)	3.7 – 7.0

Refinement

Initial model used (PDB code)	1DDB
Model resolution (Å)	3.4
FSC threshold	0.5
Model resolution range (Å)	3.1 – 3.4
Map sharpening <i>B</i> factor (Å ²)	-121.8
Model composition	
Non-hydrogen atoms	7193
Protein residues	933
Ligands	1
<i>B</i> factors (Å ²)	
Protein	117.1
Ligand	61.4
R.m.s. deviations	
Bond lengths (Å)	0.004
Bond angles (°)	0.594
Validation	
MolProbity score	1.11
Clashscore	3.17
Poor rotamers (%)	0.74
Ramachandran plot	
Favored (%)	98.48
Allowed (%)	1.52
Disallowed (%)	0.00

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

XDS v Jan 10, 2022 (BUILT=20220220) and autoPROC (GlobalPhasing, version 20230222) package for processing of crystallographic data; Automation program EPU (ThermoFischer Sci., v2.12.1) for cryoEM data collection; Biacore 8K control software (Cytiva, v4.0.8.19879) for measuring surface plasmon resonance; LE-SH800SZFCPL Cell Sorter software (Sony, v2.1.5) for sorting and collecting FACS data; Kaluza for Galios (Beckman Coulter, v1.1.20388.18228) for collecting flowcytometry data; Chromeleon (ThermoFischer Sci, v7.2.10) and Astra (Wyatt Tech., v8.0.2.5) for performing and collecting data of SEC-MALS; GatorOne (Gator Bio, v.2.7.3.0728) for biolayer interferometry; Python (v3.6.5) as well as APBS (v1.5), MSMS (v2.6.1), RDKit (v2020.09.1.0) to run our MaSIF-neosurf pipeline.

Data analysis

Coot (v0.9.5) for structure building; Prism (GraphPad, v10) for graphs generation; FlowJo (BD Bioscience, v10.8.1) for flowcytometry analysis; PyMol (Schrödinger, v2.4) and ChimeraX (UCSF, v1.3) for protein visualization and structural graphic generation; MolProbity (v4.5.1) for structure evaluation; Phenix (v1.20.1-4487) for molecular replacement phasing and structure refinement; cryoSPARC (v4.4.1) for cryoEM structure determination, Biacore Insight Evaluation Software (Cytiva, v4.0.8.19879) for evaluating surface plasmon resonance measurements, Rosetta modelling suite (v3.13) for protein design and analysis; Custom scripts to analyze protein modeling data are available on Github (https://github.com/LPDI-EPFL/masif_seed)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Crystal structure of DBAct553_2 in complex with Actinonin-bound PDF1 has been deposited at the PDB under the accession code 8S1X (DOI: <https://doi.org/10.2210/pdb8S1X/pdb>). The refined atomic models and corresponding cryoEM maps of DBPro1156_2 in complex with progesterone-bound DB3 Fab were deposited under PDB: 9FKD (<https://doi.org/10.2210/pdb9FKD/pdb>) and EMDB: 50522. The scaffold database generated for grafting the seed provided by MaSIF-neosurf is partly available at Zenodo (<https://zenodo.org/records/7643697#.Y-z533ZKhaQ>) and partly on Github (<https://github.com/strauchlab/DBP> and https://github.com/strauchlab/scaffold_design/). Data used to generate the figures and supplementary materials, as well as the relevant plasmid maps, were deposited on Zenodo (<http://doi.org/10.5281/zenodo.13737922>).

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender	<input type="text" value="N/A"/>
Reporting on race, ethnicity, or other socially relevant groupings	<input type="text" value="N/A"/>
Population characteristics	<input type="text" value="N/A"/>
Recruitment	<input type="text" value="N/A"/>
Ethics oversight	<input type="text" value="N/A"/>

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For the protein designs, ~2000 protein sequences for each target (Bcl2:venetoclax, DB3:Progesterone & PDF1:Actinonin) were filtered computationally and selected by yeast display. Based on experimental success rate observed in previous findings (doi: 10.1038/s41586-023-05993-x), 2000 designs were considered as appropriate sample size.
Data exclusions	There is no data exclusion in this study.
Replication	Each binding candidate detected by yeast display and/or deep-sequencing showed reproducible results and were also tested with negative controls (unrelated protein ligand or unlabeled yeast) and point mutants. Result reproducibility was also confirmed with alternative methods (e.g. SPR)
Randomization	Randomization does not apply in our study as oligos encoding the one-shot designs were pooled together for a high-throughput screening (one oligopool per target protein) and then sorted by yeast display using flow cytometry. The goal was to test binding capacity of each individual design.
Blinding	Researchers were not blinded as this does not apply to our study (no bias expected from cell sorting)

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

Methods

- n/a Involved in the study
- Antibodies
- Eukaryotic cell lines
- Palaeontology and archaeology
- Animals and other organisms
- Clinical data
- Dual use research of concern
- Plants

- n/a Involved in the study
- ChIP-seq
- Flow cytometry
- MRI-based neuroimaging

Antibodies

Antibodies used

- 1) Anti-HA tag, FITC - Ref : A190-138F - Manufacturer : Bethyl - Clone : Unknown - Dilution: 1:100
- 2) Anti-human IgG, PE - Ref : 12-4998-82 - Manufacturer : Invitrogen - Clone : Polyclonal - Dilution: 1:100
- 3) Biotinylated Human Her-2 - Ref: HE2-H822R - AcroBiosystems - Clone : N/A - Dilution: 1:100
- 4) Streptavidin, PE - Ref: 12-4317-87 - Invitrogen - Clone: N/A - Dilution: 1:100
- 5) Anti-V5 tag, FITC - Ref: GTX21209 - GeneTex - Clone: Polyclonal - Dilution: 1:100
- 6) Anti-FLAG tag, AlexaFluor647 - Ref: MA1-142-A647 - ThermoFisher - Clone: L5 - Dilution: 1:100

Validation

- For the commercially available antibodies, no validation reports were provided other than publications citing the products or examples on manufacturer's webpage :
- 1) 10.21037/atm.2020.03.74
 - 2) <https://www.thermoFisher.com/antibody/product/Goat-anti-Human-IgG-Fc-Secondary-Antibody-Polyclonal/12-4998-82>
 - 3) 10.2967/jnumed.118.223636
 - 4) 10.1016/j.xpro.2020.100129
 - 5) <https://www.genetex.com/Product/Detail/V5-tag-antibody-FITC/GTX21209>
 - 6) 10.1038/s41589-021-00832-4

Eukaryotic cell lines

Policy information about [cell lines and Sex and Gender in Research](#)

Cell line source(s)

Expi293 were purchased from ThermoFischer Sci. (Ref: A14635)
 HEK293T were purchased from Invitrogen (Ref: R70007)
 Phoenix-ECO were purchased from ATCC (Ref: CRL-3214)
 MC38-HER2 were generated as described previously (doi: 10.1038/s41590-021-00940-2) and provided by Prof. Li Tang (EPFL)
 Primary murine T cell were isolated from C57BL/6 mice

Authentication

Expi293 were authenticated by the provider using short tandem repeat (STR) genotyping
 HEK293T were authenticated by the provider using short tandem repeat (STR) genotyping
 Phoenix-ECO were authenticated by the provider using short tandem repeat (STR) genotyping
 MC38-HER2 was not authenticated in this study.
 Primary murine T cells were specifically isolated using a pan T cell kit (Miltenyi Biotec; Ref: 130-095-130)

Mycoplasma contamination

Expi293: Cells were tested negative for mycoplasma contamination by the manufacturer (qPCR) and no additional test was performed as the cells were used for protein expression.
 HEK293T : Cells were tested negative for mycoplasma (qPCR)
 Phoenix-ECO : Cells were tested negative for mycoplasma (MycoAlert® Mycoplasma Detection Kit - Ref: LT07-318)
 MC38-HER2 : Cells were tested negative for mycoplasma (MycoAlert® Mycoplasma Detection Kit - Ref: LT07-318)
 Primary murine T cell : Cells were tested negative for mycoplasma (MycoAlert® Mycoplasma Detection Kit - Ref: LT07-318)

Commonly misidentified lines (See [ICLAC](#) register)

No commonly misidentified lines were used in this study

Plants

Seed stocks

N/A

Novel plant genotypes

N/A

Authentication

N/A

Flow Cytometry

Plots

Confirm that:

- The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- All plots are contour plots with outliers or pseudocolor plots.
- A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Yeasts cells (EBY-100) were labeled with the protein target and then with an anti-HA (FITC) for the display signal and an anti-Fc (PE) for binding to the protein target. Cells were washed with PBS supplemented with 0.1% BSA. See methods for more details

Instrument

Sony SH800 (Sorting) and Beckman Coulter Gallios (Analysis)

Software

Sony LE-SH800SZFCPL Cell Sorter software (v2.1.5), Kaluza for Gallios (v1.1.20388.18228) and FlowJo (v10.8.1)

Cell population abundance

Transformed yeasts underwent two rounds of sorting for target binding and amplifications in culture before being sequenced for the isolation of single binding clones. Single clone candidates (re-transformation of naive yeasts with pure DNA) were then individually tested with controls.

Gating strategy

Yeast cells without target labeling served as a negative control and yeasts showing binding signal above this negative threshold were collected (See supplementary fig. 6).

- Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.