nature

https://doi.org/10.1038/s41586-022-04654-9

Accelerated Article Preview

Design of protein binding proteins from target structure alone

Received: 28 September 2021

Accepted: 15 March 2022

Accelerated Article Preview
Published online: 24 March 2022

Cite this article as: Cao, L. et al. Design of protein binding proteins from target structure alone. *Nature* https://doi.org/10.1038/s41586-022-04654-9 (2022).

Longxing Cao, Brian Coventry, Inna Goreshnik, Buwei Huang, Joon Sung Park, Kevin M. Jude, Iva Marković, Rameshwar U. Kadam, Koen H. G. Verschueren, Kenneth Verstraete, Scott Thomas Russell Walsh, Nathaniel Bennett, Ashish Phal, Aerin Yang, Lisa Kozodoy, Michelle DeWitt, Lora Picton, Lauren Miller, Eva-Maria Strauch, Nicholas D. DeBouver, Allison Pires, Asim K. Bera, Samer Halabiya, Bradley Hammerson, Wei Yang, Steffen Bernard, Lance Stewart, Ian A. Wilson, Hannele Ruohola-Baker, Joseph Schlessinger, Sangwon Lee, Savvas N. Savvides, K. Christopher Garcia & David Baker

Article

Design of protein binding proteins from target structure alone

https://doi.org/10.1038/s41586-022-04654-9

Received: 28 September 2021

Accepted: 15 March 2022

Published online: 24 March 2022

Longxing Cao[1,2,22], Brian Coventry[1,2,3,22], Inna Goreshnik[1,2], Buwei Huang[1,2,4], Joon Sung Park[5], Kevin M. Jude[6,7,8], Iva Marković[9,10], Rameshwar U. Kadam[11], Koen H. G. Verschueren[9,10], Kenneth Verstraete[9,10], Scott Thomas Russell Walsh[12,13], Nathaniel Bennett[1,2,3], Ashish Phal[1,4,14], Aerin Yang[6,7,8], Lisa Kozodoy[1,2], Michelle DeWitt[1,2], Lora Picton[6,7,8], Lauren Miller[1,2], Eva-Maria Strauch[15], Nicholas D. DeBouver[16,17], Allison Pires[17,18], Asim K. Bera[1,2], Samer Halabiya[19], Bradley Hammerson[17], Wei Yang[1,2], Steffen Bernard[11], Lance Stewart[1,2], Ian A. Wilson[11,20], Hannele Ruohola-Baker[1,14], Joseph Schlessinger[5], Sangwon Lee[5], Savvas N. Savvides[9,10], K. Christopher Garcia[6,7,8] & David Baker[1,2,21]

The design of proteins that bind to a specific site on the surface of a target protein using no information other than the three-dimensional structure of the target remains an outstanding challenge[1–5]. We describe a general solution to this problem which starts with a broad exploration of the very large space of possible binding modes to a selected region of a protein surface, and then intensifies the search in the vicinity of the most promising binding modes. We demonstrate its very broad applicability by de novo design of binding proteins to 12 diverse protein targets with very different shapes and surface properties. Biophysical characterization shows that the binders, which are all smaller than 65 amino acids, are hyperstable and following experimental optimization bind their targets with nanomolar to picomolar affinities. We succeeded in solving crystal structures of five of the binder-target complexes, and all five are very close to the corresponding computational design models. Experimental data on nearly half a million computational designs and hundreds of thousands of point mutants provide detailed feedback on the strengths and limitations of the method and of our current understanding of protein-protein interactions, and should guide improvement of both. Our approach now enables targeted design of binders to sites of interest on a wide variety of proteins for therapeutic and diagnostic applications.

Protein interactions play critical roles in biology, and general approaches to disrupt or modulate these with designed proteins would have huge impact. While empirical laboratory selection approaches starting from very large antibody, DARPIN or other protein scaffold libraries can generate binders to protein targets, it is difficult at the outset to target a specific region on a target protein surface, and to sample the full space of possible binding modes. Computational methods can target specific target surface locations and provide a more principled and potentially much faster approach to binder generation than random library selection methods, as well as insight into the fundamental properties of protein interfaces (which must be understood for design to be successful). Most current methods for computationally designing proteins to bind to a target surface utilize information derived from native complex structures on specific sidechain interactions or protein backbone placements optimal for binding[1–3]. Computational docking of antibody scaffolds with varied loop geometries has yielded binders, but the designed binding modes have rarely been validated with high-resolution structures[4]. Binders have been generated starting from several computationally identified hot-spot residues, which were then used to guide the positioning of naturally occurring protein scaffolds[5]. However, for many target proteins, there are no obvious pockets or clefts on the protein surface into which a small number of privileged sidechains can be placed, and guidance by a small number of hotspot residues limits the approach to a small fraction of possible interaction modes.

[1]Department of Biochemistry, University of Washington, Seattle, WA, USA. [2]Institute for Protein Design, University of Washington, Seattle, WA, USA. [3]Molecular Engineering Graduate Program, University of Washington, Seattle, WA, USA. [4]Department of Bioengineering, University of Washington, Seattle, WA, USA. [5]Department of Pharmacology, Yale University School of Medicine, New Haven, CT, USA. [6]Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA, USA. [7]Department of Structural Biology, Stanford University School of Medicine, Stanford, CA, USA. [8]Department of Molecular and Cellular Physiology, Stanford University School of Medicine, Stanford, CA, USA. [9]VIB-UGent Center for Inflammation Research, Ghent, Belgium. [10]Unit for Structural Biology, Department of Biochemistry and Microbiology, Ghent University, Ghent, Belgium. [11]Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA, USA. [12]National Cancer Institute, National Institutes of Health, Chemical Biology Laboratory, 1050 Boyles Street, Building 376, Frederick, MD, USA. [13]J.A.M.E.S. Farm, 13615 Highland Road, Clarksville, MD, USA. [14]Institute for Stem Cell and Regenerative Medicine, University of Washington, Seattle, WA, USA. [15]Dept. of Pharmaceutical and Biomedical Sciences, University of Georgia, Athens, GA, USA. [16]UCB Pharma., 7869 NE Day Road West, Bainbridge Island, WA, USA. [17]Seattle Structural Genomics Center for Infectious Disease (SSGCID), Seattle, WA, USA. [18]Seattle Children's Center for Global Infectious Disease Research, Seattle, WA, USA. [19]Department of Electrical and Computer Engineering, University of Washington, Seattle, WA, United States. [20]The Skaggs Institute for Chemical Biology, The Scripps Research Institute, La Jolla, CA, USA. [21]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. [22]These authors contributed equally: Longxing Cao, Brian Coventry. e-mail: dabaker@uw.edu

Design Method

We sought to develop a general approach to design of high affinity binders to arbitrary protein targets that addresses two major challenges. First, in the general case, there are no clear sidechain interactions or secondary structure packing arrangements that can mediate strong interactions with the target; instead there are a very large number of individually very weak possible interactions. Second, the number of ways of choosing which of these numerous weak interactions to incorporate into a single binding protein is combinatorially large, and any given protein backbone is unlikely to be able to simultaneously present sidechains that can encompass any preselected subset of these interactions. To motivate our approach, consider the simple analogy of a very difficult climbing wall with only a few good footholds or handholds distant from each other. Previous "hotspot" based approaches correspond to focusing on routes involving these footholds/handholds, but this greatly limits the possibilities and there may be no way to connect them into a successful route. An alternative is to: first, identify all possible handholds and footholds, no matter how poor; second, have thousands of climbers select subsets of these, and try to climb the wall; third, identify those routes that were most promising; and fourth, have a second group of climbers explore them in detail. Following this analogy, we devised a multi-step approach to overcome the above two challenges by: 1) enumerating a large and comprehensive set of disembodied sidechain interactions with the target surface; 2) identifying from large in silico libraries of protein backbones those that can host many of these sidechains without clashing with the target; 3) identifying recurrent backbone motifs in these structures; and 4) generating and placing against the target a second round of scaffolds containing these interacting motifs (Fig. 1a). Steps 1 and 2 search the space very widely, while steps 3 and 4 intensify search in the most promising regions. We describe and motivate each step below.

We began by docking disembodied amino acids against the target protein, and storing the backbone coordinates and target binding energies of the typically billions of amino acids making favorable hydrogen bonding or non-polar interactions in a 6-dimensional spatial hash table for rapid lookup (Fig. 1a; see Methods). This "rotamer interaction field" (RIF) enables rapid approximation of the target interaction energy achievable by a protein scaffold docked against a target based on its backbone coordinates alone (with no need for time consuming sidechain sampling)--for each dock, the target interaction energies of each of the matching amino acids in the hash table are summed. A related approach was used for small molecule binder design;[6] since protein targets are so much bigger, and non-polar interactions are the primary driving force for protein-protein interaction, we focused the RIF generation process on non-polar sites in specific surface regions of interest: for example in the case of inhibitor design, interaction sites with biological partners. The RIF approach improves upon previous discrete interaction-sampling approaches[5] by reducing algorithmic complexity from O(N) or O(N^2) to O(1) with respect to the number of sidechain-target interactions considered, allowing for billions, rather than thousands, of potential interfaces to be considered.

For docking against the rotamer interaction field, it is desirable to have a very large set of protein scaffold options, as the chance that any one scaffold can house many interactions is small. The structure models of these scaffolds must be quite accurate so that the positioning is correct. Using fragment assembly[7], piecewise fragment assembly[8], and helical extension[9], we designed a large set of miniproteins ranging in length from 50 to 65 amino acids containing larger hydrophobic cores than previous miniprotein scaffold libraries[1], which makes the protein more stable and more tolerant to introduction of the designed binding surfaces. 84,690 scaffolds spanning five different topologies with structural metrics predictive of folding were encoded in large oligonucleotide arrays and 34,507 were found to be stable using a high-throughput proteolysis based protein stability assay[10].

We experimented with several approaches for docking these stable scaffolds against the target structure rotamer interaction field, balancing overall shape complementarity with maximizing specific rotamer interactions. The most robust results were obtained using direct low resolution shape matching[11] followed by grid-based refinement of the rigid body orientation in the RIF (RIFDock). This approach resulted in better Rosetta binding energies (ddGs) and packing (contact molecular surface, see below) after sequence design than shape matching alone with PatchDock (Fig. 1b red and green), and more extensive non polar interaction with the target than hierarchical search without PatchDock shape matching[6] (Extended Data Fig. 2a).

Because of the loss in resolution in the hashing used to build the RIF, and the necessarily approximate accounting for interactions between sidechains (see Methods), we found that evaluation of the RIF solutions is considerably enhanced by full combinatorial optimization using the Rosetta forcefield, allowing the target sidechains to repack and the scaffold backbone to relax. Full combinatorial sequence optimization is quite CPU intensive, however, and to enable rapid screening through millions of alternative backbone placements, we developed a rapid pre-screening method using Rosetta to identify promising RIF docks. We found that including only hydrophobic amino acids, using a reduced set of rotamers than in standard Rosetta design calculations, and a more rapidly computable energy function sped design more than 10-fold while retaining a strong correlation with results after full sequence design (next paragraph); this pre-screen (referred to as the "Predictor" below) substantially improved the binding energies and shape complementarity of the final designs as far more RIF solutions could be processed (Extended Data Fig. 2b).

We observed that application of standard Rosetta design to the set of filtered docks in some cases resulted in models with buried unsatisfied polar groups and other suboptimal properties. To overcome these limitations, we developed a combinatorial sequence design protocol that maximizes shape and chemical complementarity with the target while avoiding buried polar atoms. Sequence compatibility with the scaffold monomer structure was increased using a structure based sequence profile[12], the cross-interface interactions were upweighted during the Monte Carlo-based sequence design stage to maximize the contacts between the binder and the target (ProteinProteinInterfaceUpweighter; see Methods), and rotamers containing buried unsatisfiable polar atoms were eliminated prior to packing and buried unsatisfied polar atoms penalized by a pair-wise decomposable pseudo-energy term[13]. This protocol yielded amino acid sequences more strongly predicted to fold to the designed structure (Extended Data Fig. 2c) and to bind the target (Extended Data Fig. 2d) than standard Rosetta interface design.

In the course of developing the overall binder design pipeline, we noticed upon inspection that even designs with favorable Rosetta binding free energies, large changes in Solvent Accessible Surface Area (SASA) upon binding, and high shape complementarity (SC) often lacked dense packing and interactions involving several secondary structural elements. We developed a quantitative measure of packing quality in closer accord with visual assessment -- the contact molecular surface (see Methods) -- which balances interface complementarity and size in a manner that explicitly penalizes poor packing. We used this metric to help select designs at both the rapid Predictor stage and after full sequence optimization (see Methods).

The space sampled by the search over structure and sequence space is enormous: tens of thousands of possible protein backbones × nearly one billion possible disembodied sidechain interactions per target × 10^{16} interface sequences per scaffold placement. Sampling of spaces of this size is necessarily incomplete, and many of the designs at this stage contained buried unsatisfied polar atoms (only rotamers that cannot make hydrogen bonds in any context are excluded at the packing stage) and cavities. To generate improved designs, we intensified the search around the best of the designed interfaces. We developed a resampling protocol which extracts all the secondary structural motifs making

good contacts with the target protein from the first "broad search" designs, clusters these motifs based on their backbone coordinates and rigid body placements, and then selects the binding motif in each cluster with the best per-position weighted Rosetta binding energy; around 2,000 motifs were selected for each target. These motifs, which in many cases resemble TERMs[14], are privileged because they contain a much greater density of favorable side chain interactions with the target than the rest of the designs. The motifs were used to guide another round of docking and design: scaffolds from the library were superimposed on the motifs, the favorable-interacting motif residues transferred to the scaffold, and the remainder of the scaffold sequence optimized to make further interactions with the target, allowing backbone flexibility through backbone torsion-angle minimization to increase shape complementarity with the target (Fig. 1a). Interface metrics for the designs based on the resampling protocol were considerably improved relative to those of the designs from the broad searching stage (Fig. 1b). The designs with the most favorable protein folding and protein interface metrics from both the broad searching and resampling stages were selected for experimental validation.

Experimental testing

Previous protein binder design approaches have been tested on only one or two targets, which limits assessment of their generality. To robustly test our new binder design pipeline, we selected thirteen native proteins of considerable current interest spanning a wide range of shapes and biological functions. These proteins fall into two classes: first, human cell surface or extracellular proteins involved in signaling, for which binders could have utility as probes of biological mechanism and potentially as therapeutics (Tropomyosin receptor kinase A (TrkA)[15], Fibroblast growth factor receptor 2 (FGFR2)[16], Epidermal growth factor receptor (EGFR)[17], Platelet-derived growth factor receptor (PDGFR)[18], Insulin receptor (InsulinR)[19], Insulin-like growth factor 1 receptor (IGF1R)[20], Angiopoietin-1 receptor (Tie2)[21], Interleukin-7 receptor alpha (IL-7Rα)[22], CD3 delta chain (CD3δ)[23], Transforming growth factor beta (TGF-β)[24]); and second, pathogen surface proteins for which binding proteins could have therapeutic utility (Influenza A H3 hemagglutinin (H3)[25], VirB8-like protein from Rickettsia typhi (VirB8)[26], and the SARS-CoV-2 coronavirus spike protein) (Fig. 2 and Fig. 3). For each target, we selected one or two regions to direct binders against for maximal biological utility and for potential downstream therapeutic potential. These regions span a wide range of surface properties, with diverse shape and chemical characteristics (Fig. 2, Fig. 3 and Extended Data Fig. 3). Some of the selected targeting regions overlap with the native interfaces, but no native interface information or native hotspots were used during the binder design process. For some targets (e.g. CD3δ and VirB8), no native complex structures were available and there were no proteins known to bind at the targeted region.

Using the above protocol, we designed 15,000-100,000 binders for each of 13 target sites on the 12 native proteins (see Methods; we chose two sites on the EGF receptor). Synthetic oligonucleotides (230 bp) encoding the 50-65 residue designs were cloned into a yeast surface expression vector, the designs were displayed on the surface of yeast, and those which bind their target enriched by several rounds of fluorescence-activated cell sorting using fluorescently labeled target proteins. The starting and enriched populations were deep sequenced, and the fraction of each design after each sort was determined by comparing the frequency of the design in the parent and child pools. From multiple sorts at different target protein concentrations, we determined, as a proxy for binding K_d's, the midpoint concentration (SC_{50}) in the binding transitions for each design in the library (Extended Data Table 1 and Methods).

To assess whether the top enriched designs for each target fold and bind as in the corresponding computational design models, and to investigate the sequence dependence of folding and binding, we generated high-resolution footprints of the binding surface by sorting site saturation mutagenesis libraries (SSMs) in which every residue was substituted with each of the 20 amino acids one at a time. For the majority, but not all, of the enriched designs, substitutions at the binding interface and in the protein core were less tolerated than substitutions at non-interface surface positions (Fig. 2, Fig. 3 and Extended Data Fig. 5), and all of the cysteines were highly conserved in designs containing disulfides. The effects of each mutation on both binding energy and monomer stability were estimated using Rosetta design calculations, and a reasonable correlation was found between the predicted and experimentally determined effect of mutations (Extended Data Fig. 6a). In almost all cases, a small number of substitutions were found to increase apparent binding affinity, and we generated libraries combining 5-15 of these and sorted for binding under increasingly stringent (lower target concentrations) conditions. Many of these affinity-enhancing substitutions were mutations to tyrosine (Extended Data Fig. 6b), consistent with the high relative frequency of tyrosine in natural protein interfaces[27]. The set of affinity increasing substitutions provide valuable information for improving the approach as these substitutions ideally would have been identified in the computational sequence design calculations (see discussion below).

We expressed the highest affinity combinatorially-optimized binders for each target in *E. coli* for more detailed structural and functional characterization. All of the designs were in the soluble fraction, and could be readily purified by Ni^{2+}-NTA chromatography. All had circular dichroism spectra consistent with the design model, and most (9 out of 13) were stable at 95 °C (Fig. 2, 3 and Table 1). The binding affinities for the targets were assessed by biolayer interferometry, and found to range from 300 pM to 900 nM (Fig. 3, Table 1 nd Extended Data Fig. 4). The sequence mapping data report on the residues on the design critical for binding, but only weakly on the region of the target bound. We investigated this using a combination of binding competition experiments, biological assays, and structural characterization of the complexes. For the nine targets for which these were available, this characterization suggested binding modes consistent with the design models, as described in the following paragraphs.

Cell Receptors involved in signaling

The receptor tyrosine kinases TrkA, FGFR2, PDGFR, EGFR, InsulinR, IGF1R and Tie2 are key regulators of cellular processes and are involved in the development and progression of many types of cancer[28]. We designed binders targeting the native ligand binding sites for PDGFR, EGFR (on both domain I and domain III, the binders are referred to as EGFRn_mb and EGFRc_mb respectively), InsulinR, IGF1R and Tie2, and targeting surface regions proximal to the native ligand binding sites for TrkA and FGFR2 (Fig. 2 and Fig. 3 and see methods for criteria). We obtained binders to all eight target sites; the binding affinities of the optimized designs ranged from ~1 nM or better for TrkA and FGFR2, to 860 nM for IGF1R (Table 1). Competition experiments with nerve growth factor (NGF), Platelet Derived Growth Factor-BB (PDGF-BB), insulin, insulin growth factor-1 (IGF-1) and Angiopoietin 1 (Ang1) on yeast suggest that the binders for TrkA, PDGFR, InsulinR, IGF1R and Tie2 bind to the targeted sites (Extended Data Fig. 7), consistent with the computational design models. The receptor tyrosine kinase binders as monomers are all expected to be antagonists, and we tested the effect on signaling through TrkA, FGFR2 and EGFR of the cognate binders on cells in culture. Strong inhibition of signaling by the native agonists was observed in all three cases (Fig. 3c, Extended Data Fig. 8 and Extended Data Fig. 9).

Binding of IL-7 to the IL-7α receptor subunit leads to recruitment of the γ_c receptor, forming a tripartite cytokine-receptor complex crucial to several signaling cascades leading to the development and homeostasis of T and B cells[29]. We obtained a picomolar affinity binder for IL-7Rα targeting the IL-7 binding site, and found that it blocks STAT5

signaling induced by IL-7 (Fig. 3c and Table 1). We also obtained binders to CD3δ, one of the subunits of the T-cell receptor, and the signaling molecule TGF-β, which play critical roles in immune cell development and activation (Fig. 2 and Table 1).

Pathogen target proteins

Hemagglutinin (HA) is the main target for influenza A virus vaccine and drug development, and can be genetically classified into two main subgroups, group 1 and group 2[30,31]. The HA stem region is an attractive therapeutic epitope, as it is highly conserved across all influenza A subtypes and targeting this region can block the low pH-induced conformational rearrangements associated with membrane fusion, which is essential for virus infection[32,33]. Neutralizing antibodies targeting the stem region of group 2 HA have been identified through screening of large B-cell libraries after vaccination or infection that neutralize both group 1 and group 2 influenza A viruses[34,35]. Protein[1,5], peptide[36] and small molecule inhibitors[37] have been designed to bind to the stem region of group 1 HA and neutralize the influenza A viruses, but none recognize the group 2 HA. The design of small proteins or peptides that can bind and neutralize both group 1 HA and group 2 HA has been challenging due to three main differences between group 1 HA and group 2 HA: first, the group 2 HA stem region is more hydrophilic, containing more polar residues; second, in group 2 HA, Trp21 adopts a configuration roughly perpendicular to the surface of the targeting groove, which makes the targeted groove much shallower and less hydrophobic; and third, the group 2 HA is glycosylated at Asn38 with the carbohydrate side chains covering the hydrophobic groove (Extended Data Fig. 10a). We used our new method to design binders to H3 HA (A/Hong Kong/1/1968), the main pandemic subtype of group 2 influenza virus, and obtained a binder with an affinity of 320 nM to wild-type H3 (Fig 2 and Table 1) and 28 nM to the deglycosylated H3 variant (N38D) (Extended Data Fig. 10b); the reduction in affinity is likely due to the entropy loss of the glycan upon binding and/or the steric clash with the glycan. The binder also binds to H1 HA (A/Puerto Rico/8/1934) which belongs to the main pandemic subtype of group 1 influenza virus (Extended Data Fig. 10b); the binding to both H1 and H3 HA is competed by the stem region binding neutralizing antibody FI6v3[34] on the yeast surface (Extended Data Fig. 10c), suggesting that the binder binds the hemagglutinin at the targeted site. We also designed binders to the prokaryotic pathogen protein VirB8 which belongs to the type IV secretion system of *Rickettsia typhi*, which is the causative agent of murine typhus[26]. We selected the surface region composed of the second and the third helices of VirB8, and obtained binders with 510 pM affinity (Fig. 2 and Table 1).

With the outbreak of the SARS-CoV-2 coronavirus pandemic we applied our method to design miniproteins targeting the receptor binding domain of the SARS-CoV-2 spike protein near the ACE2 binding site to block receptor engagement. Due to the pressing need for coronavirus therapeutics, we recently described the results of these efforts[38] ahead of those described in this manuscript; As in the case of FGFR2, IL-7Rα and VirB8, the method yielded picomolar binders, which are among the most potent compounds known to inhibit the virus in cell culture (IC_{50} 0.15 ng/ml) and subsequent animal experiments have shown that they provide potent protection against the virus in vivo[39]. The modular nature of the miniprotein binders enables their rapid integration into designed diagnostic biosensors for both influenza and SARS-CoV-2 binders[40].

The designed binding proteins are all very small proteins (<65 amino acids), and many are 3-helix bundles. To evaluate their target specificity, we tested the highest affinity binder to each target for binding to all other targets. There was very little cross reactivity (Fig. 4a), likely due to their quite diverse surface shapes and electrostatic properties (Fig. 4b). Consistent with previous observations with affibodies[41], this suggests that a wide variety of binding specificities can be encoded in simple helical bundles; in our approach, scaffolds are customized for each target, so the specificity arises both from the set of sidechains at the binding interface, and the overall shape of the interface itself.

High-resolution structural validation

High-resolution structures are critical for evaluating the accuracy of computational protein designs. We succeeded in obtaining crystal structures of the unbound miniprotein binders for FGFR2 and IL-7Rα, as well as co-crystal structures of the miniprotein binders of H3, TrkA, FGFR2, IL-7Rα and VirB8 in complex with their targets (Extended Data Table 2). The H3 binder binds to the shallow groove of the stem region of HK68/H3 HA in the crystal structure as designed; the Cα root-mean-square deviation (RMSD) over the entire miniprotein binder is 1.91 Å using the HA as the alignment reference (Fig. 5a). The binder makes extensive hydrophobic interactions with HA and almost all of the designed interface side chain configurations are recapitulated in the crystal structure (Fig. 5a). There is a clear reorientation of the oligosaccharide at Asn38 compared with the unbound HK68/H3 structure (Fig. 5a and Extended Data Fig. 10a; this has also been observed in HK68/H3 HA structures bound to stem region neutralizing antibodies[34,35]), consistent with the higher binding affinity for the N38D variant than for wild-type H3 HA (A/Hong Kong/1/1968) in BLI assays (Table 1 and Extended Data Fig. 10b). The crystal structure of the TrkA binder in complex with TrkA was very close to the design model (Fig. 5b). After aligning the crystal structure and design model on TrkA, the Cα RMSD over the entire miniprotein binder is 2.41 Å, and over the two interfacial binding helices, 1.20 Å. The crystal structures of the FGFR2 binder by itself (Extended Data Fig. 11a) and in complex with the third Ig-like domain of FGFR4 (Fig. 5c) match the design models with near atomic accuracy, with Cα rmds of 0.58 Å for the binder alone and 1.33 Å over the entire complex. The TrkA binder and the FGFR2 binder bind to the curved sheet side of the ligand binding domain of TrkA and FGFR4 with extensive hydrophobic and polar interactions, and most of the key hydrophobic interactions as well as the primarily polar interactions in the computational design models are largely recapitulated in the crystal structures (Fig 5b, c). The binding interfaces partially overlap with the native ligand binding sites of nerve growth factor (NGF) and fibroblast growth factor (FGF); however, the detailed sidechain interactions are entirely different in the designed and native complexes (Extended Data Fig. 3). For IL-7Rα, the crystal structure of the monomer is close to that of the design, with a Cα RMSD of 0.63 Å (Extended Data Fig. 11b) and the co-crystal structure with IL-7Rα also matches closely with the design model, with a Cα RMSD of 2.2 Å using IL-7Rα as the reference (Fig 5d). Both the de novo IL-7Rα binder and the native IL-7 use two helices to bind with IL-7Rα, but the binding orientations are totally different (Extended Data Fig. 3). The VirB8 binder makes extensive interactions with the helical regions of VirB8 as designed; no native proteins have been identified to bind to this region. The Cα RMSD over the entire miniprotein binder is 2.54 Å using the VirB8 as the alignment reference, and the sidechain configurations of key interface residues are largely recapitulated (Fig. 5e). The heavy-atom RMSDs over the buried sidechains at the interface (within 8 Å of the target in the design models) are 0.71 Å (H3), 1.10 Å (TrkA), 1.29 Å (FGFR2), 1.63 Å (IL-7Rα) and 1.52 Å (VirB8), close to the core sidechain RMSDs (mean 0.90 Å). Further highlighting the accuracy of the protein interface design method, cryoEM structures of the SARS-CoV-2 binders LCB1 and LCB3 in complex with the virus are also nearly identical to the design models, with Cα RMSD of 1.27 Å and 1.9 Å respectively[38] (Fig. 5f). While we were not able yet to solve structures for the remainder of the designs, the high-resolution sequence footprinting (Fig. 2 and Fig. 3) and competition results (Extended Data Fig. 7) suggest that the interfaces involve both the designed residues and the intended regions on the target. The very close agreement between the experimentally determined structures and the original design models suggests that the substitutions

required to achieve high affinity play relatively subtle roles in tuning interface energetics; the overall structure of the complex, including the structure of the monomer binders and the detailed target binding mode, are determined by the computational design procedure.

Determinants of design success

For our de novo design strategy to be successful, we must encode in the ~60-residue designed sequences both information on the folded monomer structures, and on the target binding interfaces: designs which do not fold to the correct structure, or which fold to the intended structures but do not bind to the target will fail. To assess the accuracy with which the monomer structure must be designed, we carried out an additional calculation and experiment for the IL-7Rα target. Large numbers of scaffolds were superimposed onto 11 interface helical binding motifs identified in the first broad design search, and sequence design was carried out as described above. A strong correlation was found between the extent of binding and the RMSD to the binding motif (Extended Data Fig. 12a), suggesting that designed backbones must be quite accurate to achieve binding.

To assess the determinants of binding of the designed interfaces, assuming that the designs fold to the intended monomer structures, we took advantage of the large data set (810,000 binder designs and 240,000 single mutants) generated in this study. Design success rates varied considerably between the different targets: for some (FGFR2 and PDGFR), hundreds of binders were generated, while for others (Tie2 and CD3δ), fewer than 10 binders were obtained from libraries of 100,000 designs (Extended Data Table 1). Across all targets, there was a strong correlation between success rate and the hydrophobicity of the targeted region (Extended Data Fig. 12b), and designs observed experimentally to bind their targets tended to have stronger predicted binding energy, and larger contact molecular surfaces (Extended Data Fig. 13). As found previously for design of protein stability[10], iterative design-build-test cycles in which the design method is updated at each iteration to incorporate feedback from the previous design round should lead to systematic improvement in the design methodology and success rate.

Conclusions

Our success in designing nM affinity binders for 14 target sites demonstrates that binding proteins can be designed de novo using only information on the structure of the target protein, without need for prior information on binding hotspots or fragments from structures of complexes with binding partners. The success also suggests that our design pipeline provides a quite general solution to the de novo protein interface design problem that goes far beyond previously described methods. However, there is still considerable room for improvement. Only a small fraction of designs bind, and in almost all cases, the best of these require additional (5-14) substitutions to achieve high affinity binding. Furthermore, the design of binders to highly polar target sites remains a considerable challenge-the sites targeted here all contain at least four hydrophobic residues. The datasets generated in this work -- both the information on binders versus non binders, and the feedback on the effects of individual point mutants on binding -- should help guide the development of methods for designing high affinity binders directly from the computer with no need for iterative experimental optimization. More generally, the de novo binder design method and the large data set generated here provide a starting point for investigating the fundamental physical chemistry of protein-protein interactions, and for developing and assessing computational models of protein-protein interactions.

This work represents a major step forward towards the longer range goal of direct computational design of high affinity binders starting from structural information alone. We expect that the binders created here, and new ones created with the method moving forward, will find wide utility as signaling pathway antagonists as monomeric proteins and as tunable agonists when rigidly scaffolded in multimeric formats, and in diagnostics and therapeutics for pathogenic disease. Unlike antibodies, the designed proteins are soluble when expressed in *E. coli* at high levels and are thermostable, and hence could form the basis for a next generation of lower cost protein therapeutics. More generally, the ability to rapidly and robustly design high affinity binders to arbitrary protein targets could transform the many areas of biotechnology and medicine that rely on affinity reagents.

Online content

Any methods, additional references, Nature Research 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/s41586-022-04654-9.

1. Chevalier, A. et al. Massively parallel de novo protein design for targeted therapeutics. *Nature* **550**, 74–79 (2017).
2. Strauch, E. M. et al. Computational design of trimeric influenza-neutralizing proteins targeting the hemagglutinin receptor binding site. *Nat. Biotech.* **35**, 667–671 (2017).
3. Silva, D. A. et al. De novo design of potent and selective mimics of IL-2 and IL-15. *Nature* **565**, 186–191 (2019).
4. Baran, D. et al. Principles for computational design of binding antibodies. *Proc. Natl Acad. Sci. U.S.A.* **114**, 10900–10905 (2017).
5. Fleishman, S. J. et al. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science* **332**, 816–821 (2011).
6. Dou, J. et al. De novo design of a fluorescence-activating β-barrel. *Nature* **561**, 485–491 (2018).
7. Koga, N. et al. Principles for designing ideal protein structures. *Nature* **491**, 222–227 (2012).
8. Linsky, T. et al. Sampling of Structure and Sequence Space of Small Protein Folds. Preprint at *bioRxiv* https://doi.org/10.1038/d41586-022-00463-2 (2021).
9. Maguire, J. B. et al. Perturbing the energy landscape for improved packing during computational protein design. *Proteins* **89**, 436–449 (2021).
10. Rocklin, G. J. et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **357**, 168–175 (2017).
11. Schneidman-Duhovny, D., Inbar, Y., Nussinov, R. & Wolfson, H. J. PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucl. Acids Res.* **33**, W363–367 (2005).
12. Brunette, T. J. et al. Modular repeat protein sculpting using rigid helical junctions. *Proc. Natl Acad. Sci. U.S.A.* **117**, 8870–8875 (2020).
13. Coventry, B. & Baker, D. Protein sequence optimization with a pairwise decomposable penalty for buried unsatisfied hydrogen bonds. *PLoS Comp. Biol.* **17**, e1008061 (2021).
14. Mackenzie, C. O., Zhou, J. & Grigoryan, G. Tertiary alphabet for the observable protein structural universe. *Proc. Natl Acad. Sci. U.S.A.* **113**, E7438–E7447 (2016).
15. Wiesmann, C., Ultsch, M. H., Bass, S. H. & de Vos, A. M. Crystal structure of nerve growth factor in complex with the ligand-binding domain of the TrkA receptor. *Nature* **401**, 184–188 (1999).
16. Plotnikov, A. N., Hubbard, S. R., Schlessinger, J. & Mohammadi, M. Crystal structures of two FGF-FGFR complexes reveal the determinants of ligand-receptor specificity. *Cell* **101**, 413–424 (2000).
17. Garrett, T. P. et al. Crystal structure of a truncated epidermal growth factor receptor extracellular domain bound to transforming growth factor α. *Cell* **110**, 763–773 (2002).
18. Shim, A. H. et al. Structures of a platelet-derived growth factor/propeptide complex and a platelet-derived growth factor/receptor complex. *Proc. Natl Acad. Sci. U.S.A.* **107**, 11307–11312 (2010).
19. Croll, T. I. et al. Higher-Resolution Structure of the Human Insulin Receptor Ectodomain: Multi-Modal Inclusion of the Insert Domain. *Structure* **24**, 469–476 (2016).
20. Xu, Y. et al. How ligand binds to the type 1 insulin-like growth factor receptor. *Nat. Commun.* **9**, 821 (2018).
21. Barton, W. A. et al. Crystal structures of the Tie2 receptor ectodomain and the angiopoietin-2-Tie2 complex. *Nat. Struct. Mol. Biol.* **13**, 524–532 (2006).
22. McElroy, C. A., Dohm, J. A. & Walsh, S. T. Structural and biophysical studies of the human IL-7/IL-7Rα complex. *Structure* **17**, 54–65 (2009).
23. Arnett, K. L., Harrison, S. C. & Wiley, D. C. Crystal structure of a human CD3-ε/δ dimer in complex with a UCHT1 single-chain antibody fragment. *Proc. Natl Acad. Sci. U.S.A.* **101**, 16268–16273 (2004).
24. Radaev, S. et al. Ternary complex of transforming growth factor-β1 reveals isoform-specific ligand recognition and receptor recruitment in the superfamily. *J. Biol. Chem.* **285**, 14806–14814 (2010).
25. Ekiert, D. C. et al. Cross-neutralization of influenza A viruses mediated by a single antibody loop. *Nature* **489**, 526–532 (2012).
26. Gillespie, J. J. et al. Structural Insight into How Bacteria Prevent Interference between Multiple Divergent Type IV Secretion Systems. *mBio* **6**, e01867–01815 (2015).
27. Birtalan, S. et al. The intrinsic contributions of tyrosine, serine, glycine and arginine to the affinity and specificity of antibodies. *J. Mol. Biol.* **377**, 1518–1528 (2008).
28. Lemmon, M. A. & Schlessinger, J. Cell signaling by receptor tyrosine kinases. *Cell* **141**, 1117–1134 (2010).

29. Markovic, I. & Savvides, S. N. Modulation of Signaling Mediated by TSLP and IL-7 in Inflammation, Autoimmune Diseases, and Cancer. *Frontiers Immunol.* **11**, 1557 (2020).
30. Webster, R. G., Bean, W. J., Gorman, O. T., Chambers, T. M. & Kawaoka, Y. Evolution and ecology of influenza A viruses. *Microbiol. Rev.* **56**, 152–179 (1992).
31. Nobusawa, E. et al. Comparison of complete amino acid sequences and receptor-binding properties among 13 serotypes of hemagglutinins of influenza A viruses. *Virology* **182**, 475–4853 (1991).
32. Bullough, P. A., Hughson, F. M., Skehel, J. J. & Wiley, D. C. Structure of influenza haemagglutinin at the pH of membrane fusion. *Nature* **371**, 37–43 (1994).
33. Ekiert, D. C. et al. Antibody recognition of a highly conserved influenza virus epitope. *Science* **324**, 246–251 (2009).
34. Corti, D. et al. A neutralizing antibody selected from plasma cells that binds to group 1 and group 2 influenza A hemagglutinins. *Science* **333**, 850–856 (2011).
35. Joyce, M. G. et al. Vaccine-Induced Antibodies that Neutralize Group 1 and Group 2 Influenza A Viruses. *Cell* **166**, 609–623 (2016).
36. Kadam, R. U. et al. Potent peptidic fusion inhibitors of influenza virus. *Science* **358**, 496–502 (2017).
37. van Dongen, M. J. P. et al. A small-molecule fusion inhibitor of influenza virus is orally active in mice. *Science* **363**, eaar6221 (2019).
38. Cao, L. et al. De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. *Science* **370**, 426–431 (2020).
39. Case, J. B. et al. Ultrapotent miniproteins targeting the SARS-CoV-2 receptor-binding domain protect against infection and disease. *Cell Host & Microbe* **29**, 1151–1161 (2021).
40. Quijano-Rubio, A. et al. De novo design of modular and tunable protein biosensors. *Nature* **591**, 482–487 (2021).
41. Frejd, F. Y. & Kim, K. T. Affibody molecules as engineered protein drugs. *Experimental Molec. Med.* **49**, e306 (2017).

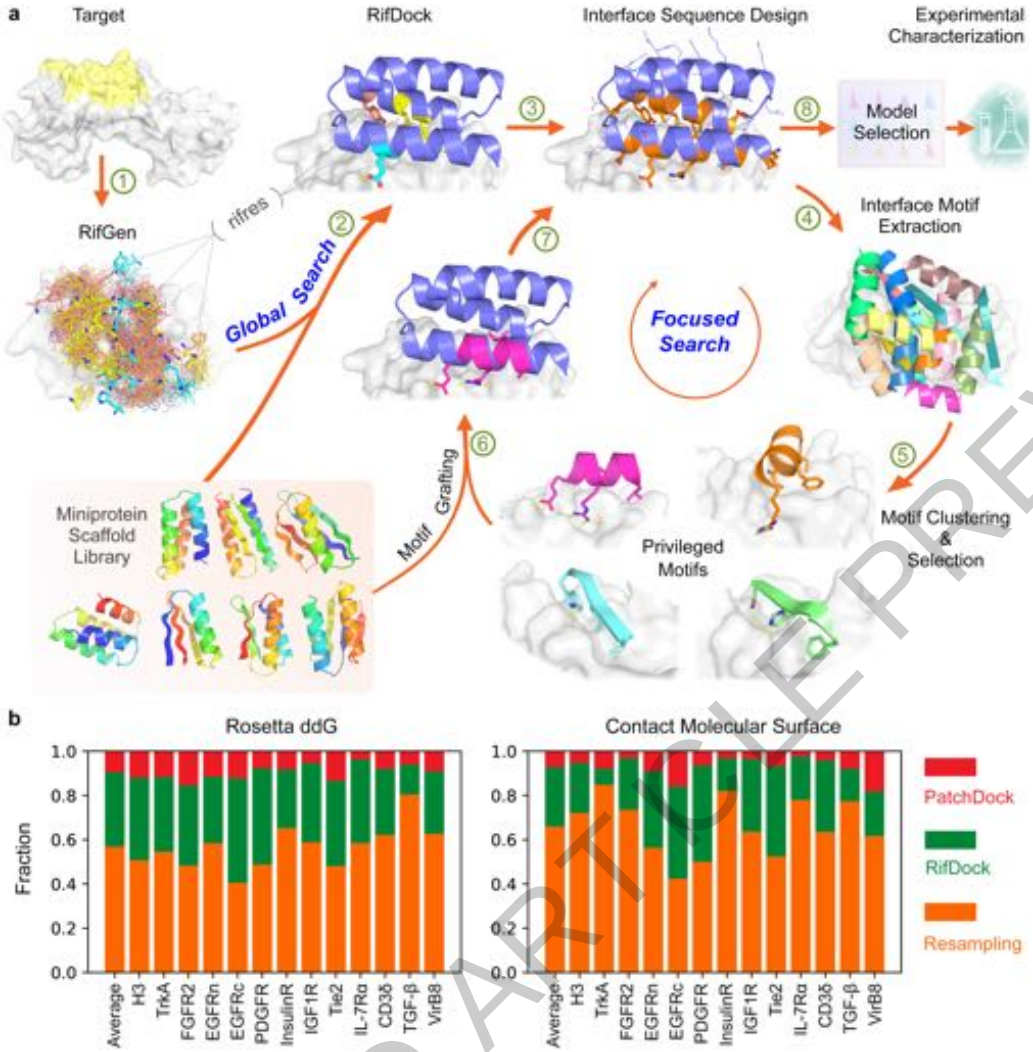

Fig. 1 | Overview of the de novo protein binder design pipeline. a, Schematic of our two stage binder design approach. In the global search stage, billions of disembodied amino acids are docked onto the selected targeting region and the positioning of the scaffolds is guided by the favorable sidechain interactions. The interface sequences are then designed to maximize interaction with the target. In the focused search stage, the interface motifs are extracted, clustered. The privileged motifs are then selected to guide another round of docking and design. Designs are then selected for experimental characterization based on computational metrics. See Extended Data Fig. 1 for a more detailed flow chart of the de novo binder design pipeline. **b**, Comparison of sampling efficiency of PatchDock, RifDock, and resampling protocols. Bar graph shows the distribution over the three approaches of the top 1% of binders based on Rosetta ddg and contact molecular surface after pooling equal-CPU-time dock-and-design trajectories for each of the 13 target sites and averaging per-target distributions (see Methods).

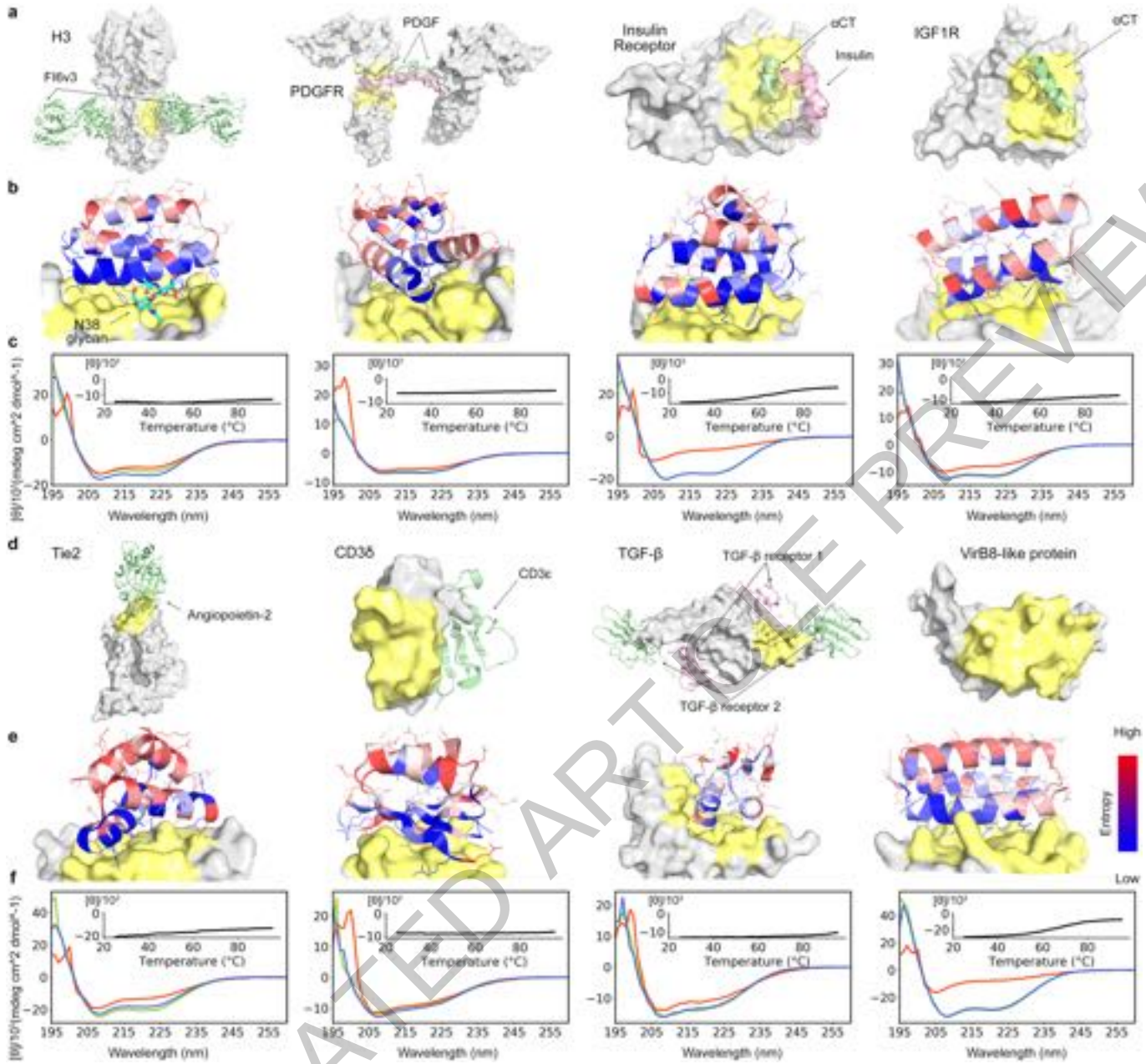

Fig. 2 | De novo design and characterization of miniprotein binders. a and **d**, Naturally-occurring target protein structures shown in surface representation, with known interacting partners shown in cartoons where available. Regions targeted for binder design are colored in pale yellow or green; the remainder of the target surface is in gray. See Extended Data Fig. 3 for side-by-side comparisons of the native binding partners and the computation design models. The PDB ID codes are 3ZTJ (H3), 3MJG (PDGFR), 4OGA (InsulinR), 5U8R (IGF1R), 2GY7 (Tie2), 1XIW (CD3δ), 3KFD (TGF-β) and 4O3V (VirB8). **b** and **d**, Computational models of designed complexes colored by site saturation mutagenesis results. Designed binding proteins (cartoons) are colored by positional Shannon entropy, with blue indicating positions of low entropy (conserved) and red those of high entropy (not conserved); target surface is in gray and yellow. The core residues and binding interface residues are more conserved than the non-interface surface positions, consistent with the computational models. Full SSM maps over all positions of all the de novo designs are provided in the Supplementary Information. **c** and **f**, Circular dichroism spectra at different temperatures (green: 25 °C, red: 95 °C, blue: 95 °C followed by 25 °C) and (insert) CD signal at 222-nm wavelength as a function of temperature for the optimized designs. See Extended Data Fig. 4 for the biolayer interferometry characterization results of the optimized designs.

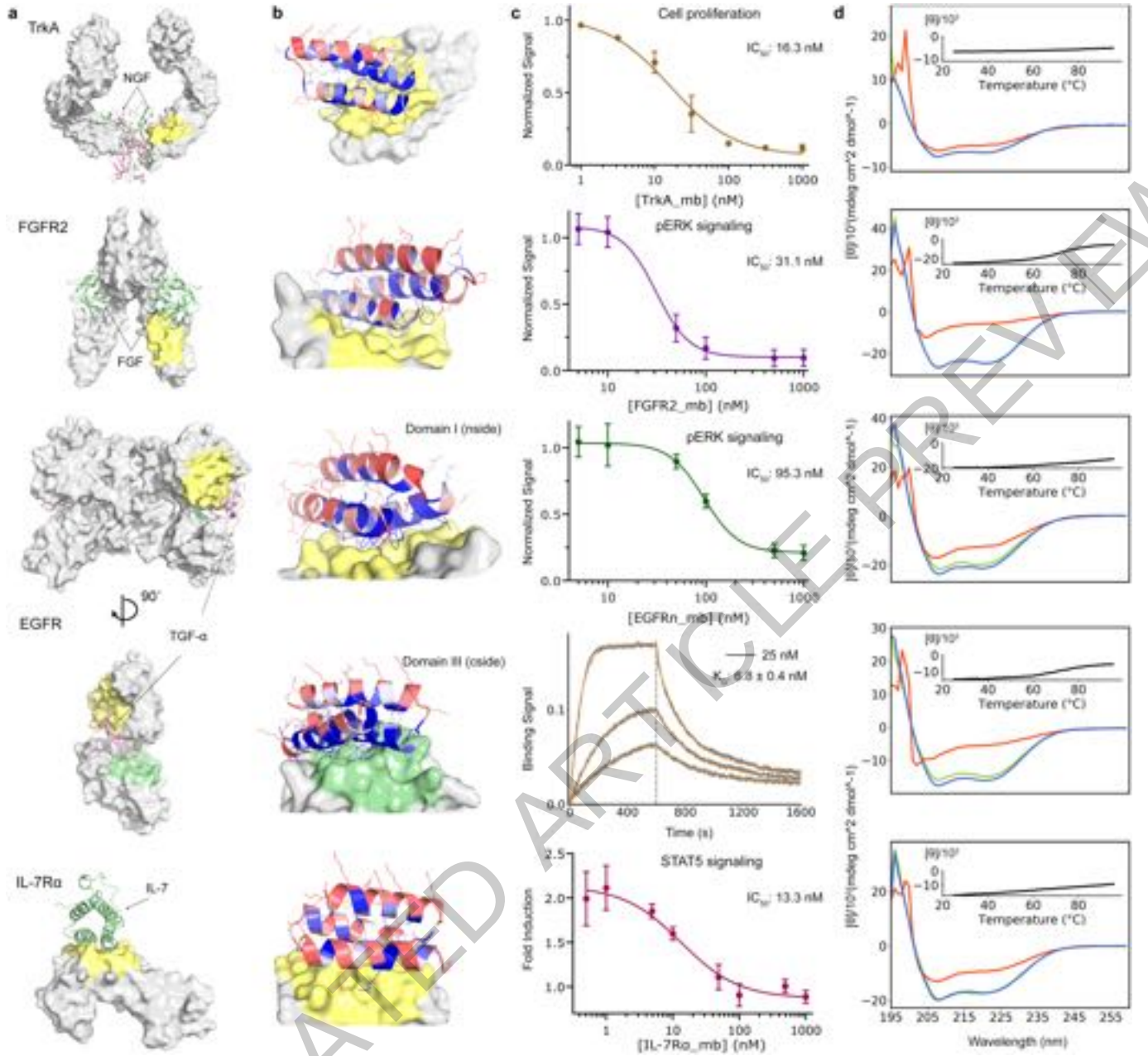

Fig. 3 | De novo design and inhibition of native signaling pathways by designed miniproteins. See the panel descriptions in Fig. 2 legend for (**a**), (**b**) and (**d**), and the PDB ID codes are 2IFG (TrkA), 1DJS (FGFR2), 1MOX (EGFR), 3DI3 (IL-7Rα) for (**a**). **c**, For TrkA, the dose-dependent reduction in cell proliferation after 48 hr of TF-1 cells with increase in TrkA minibinder concentration is shown. 8.0 ng/ml human β-NGF was used for competition. Titration curves at different concentrations of NGF and the effects of the miniprotein binders on cell viability are in Extended Data Fig. 8. For FGFR2, the dose-dependent reduction pERK signaling elicited by 0.75 nM βFGF in HUVECs with increasing FGFR2 minibinder concentration is shown. For EGFR N-side binder, the dose-dependent reduction in pERK signaling elicited by 1 nM EGF in HUVECs with increase in EGFR n-side minibinder concentration is shown. See Extended Data Fig. 9 and methods for the experimental details. For the EGFR C-side binder, the biolayer interferometry results are shown. See Extended Data Fig. 4 for the biolayer interferometry characterization results of the other optimized designs. For IL-7R, the reduction in STAT5 activity induced by 50 pM of hIL-7 in HEK293T cells in the presence of increased hIL-7Rα minibinder concentrations is shown. The mean values were calculated from triplicates for the cell signaling inhibition assays measured in parallel, and error bars represent standard deviations. IC_{50} was calculated using a four-parameter-logistic equation in GraphPad Prism 9 software.

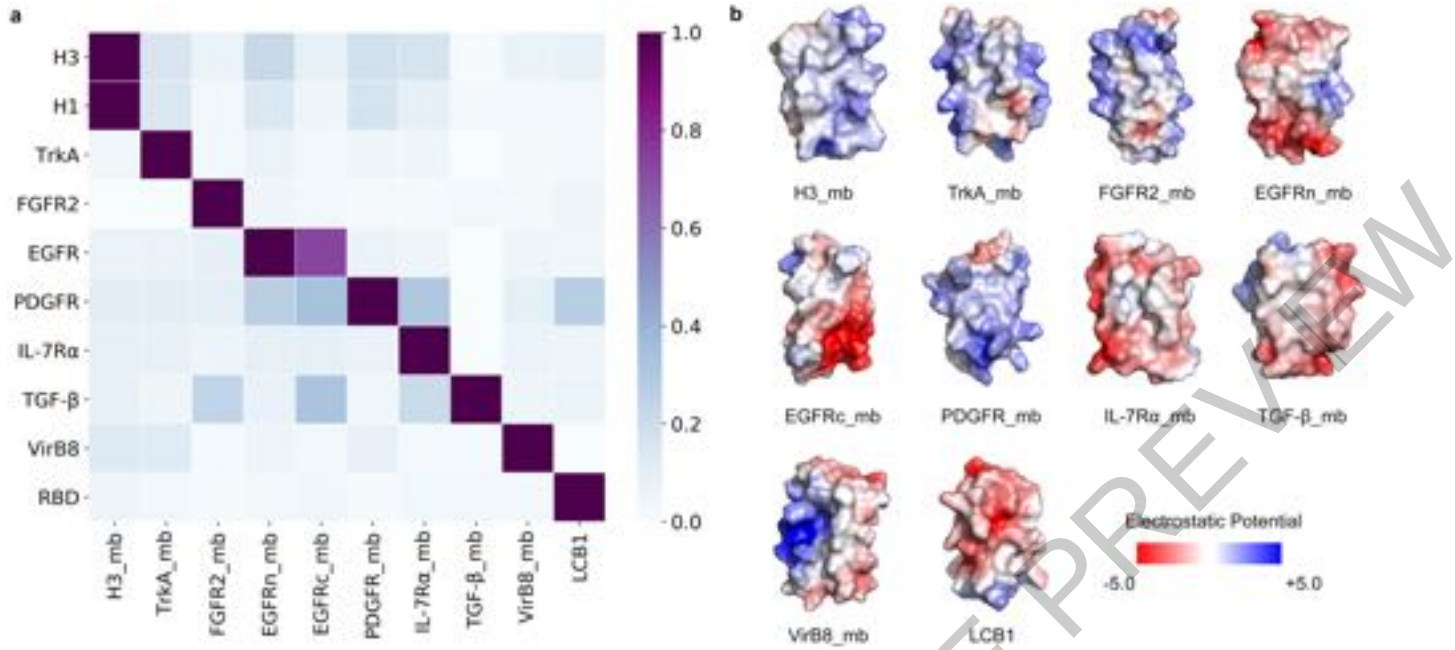

Fig. 4 | Designed binders have high target specificity. To assess the cross reactivity of each miniprotein binder with each target protein, the biotinylated target proteins were loaded onto biolayer interferometry SA sensors, allowed to equilibrate, and baseline signal set to zero. The BLI tips were then placed into 100 nM binder solution for 5 min, washed with buffer, and dissociation was monitored for an additional 10 min. Heatmap shows the maximum response signal for each binder-target pair normalized by the maximum response signal of the cognate designed binder-target pair. The raw BLI traces are shown in the supplementary data. **b**, Surface shape and electrostatic potential (generated with the APBS Electrostatics plugin in Pymol; red positive, blue, negative) of the designed binding interfaces.

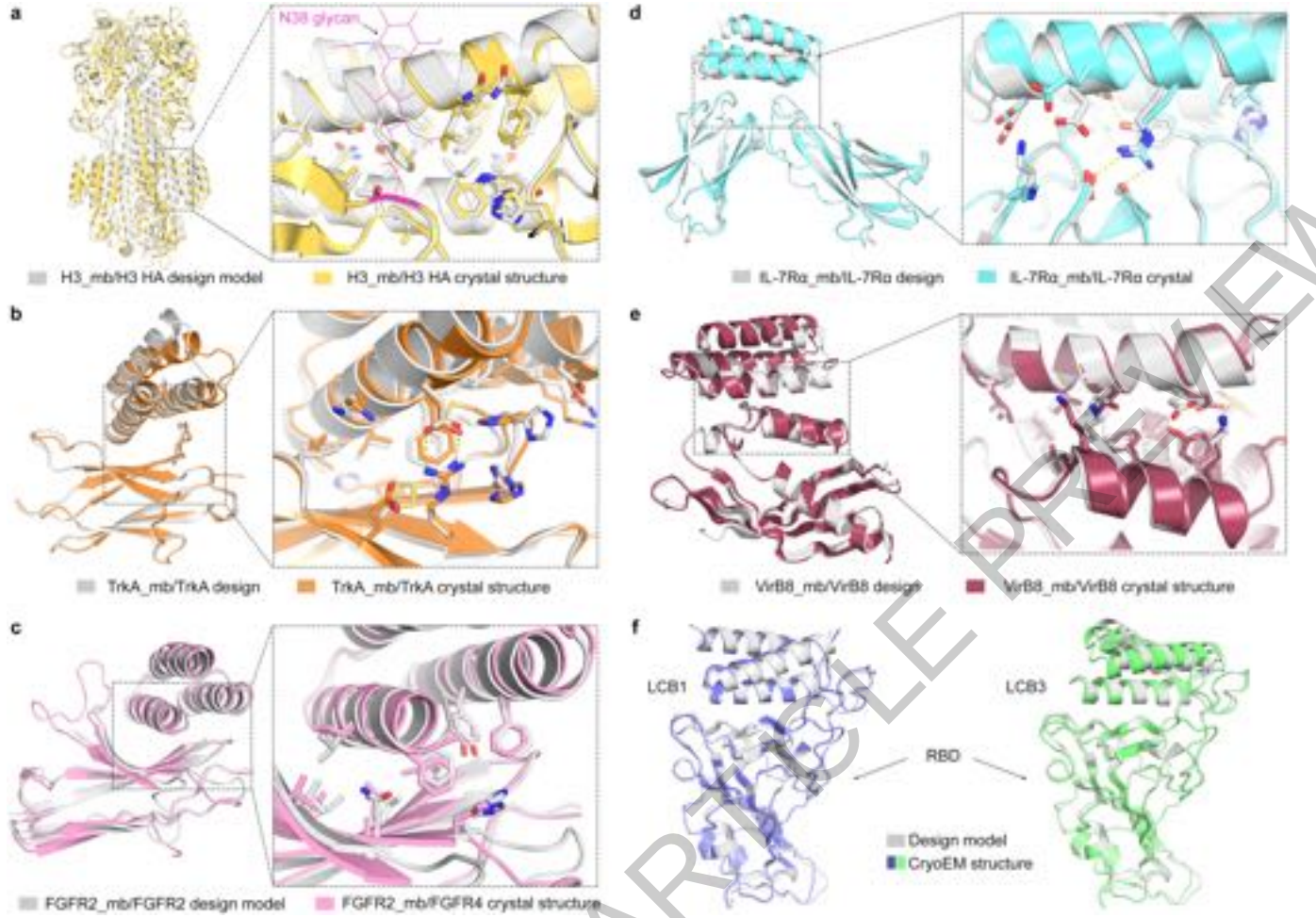

Fig. 5 | High-resolution structures of miniprotein binders in complex with target proteins are very close to the computational design models. (a-e). (left) Superimposition of computational design model (silver) on experimentally determined crystal structure. (right) Zoom-in view of designed interface, with interacting side chains as sticks. **a**. H3 HA, **b**. TrkA, **c**. FGFR2, **d**. IL-7Rα, **e**. VirB8. **f**, Superimposition of the computational design model and refined cryo-EM structures of LCB1 (left) and LCB3 (right) bound to receptor binding domain of SARS-CoV-2 spike protein (design models are in gray and cryoEM structures are in pale blue and green).

Table 1 | Physicochemical properties of the optimized de novo miniprotein binders

	H3	TrkA	FGFR2	EGFRn	EGFRc	PDGFR	InsulinR	IGF1R	Tie2	IL-7Rα	CD3δ	TGF-β	VirB8
K_D (nM)	320 ± 24.0	1.4 ± 0.02	243 ± 59.0	1.2 ± 0.01	6.8 ± 0.3	82 ± 25	210 ± 39	860 ± 270	584 ± 35	0.31 ± 0.004	612 ± 30	113 ± 4.4	0.51 ± 0.005
TM (°C)	> 95.0	> 95.0	71.1	> 95.0	71.2	> 95.0	65.0	> 95.0	> 95.0	> 95.0	> 95.0	> 95.0	66.2

The binding affinity and melting temperature of the optimized de novo miniprotein binders. See Fig. 2, Fig. 3 for the circular dichroism spectra and the raw biolayer interferometry traces in Fig. 3 and Extended Data Fig. 4. The experimental details can be found in the corresponding figure legends and the method section.

Methods

Broad search stage

The crystal structures of HA (PDB: 4FNK)[25], EGFR (PDB: 1MOX, 4UV7)[17,42], PDGFR (PDB: 3MJG)[18], IR (PDB: 4ZXB)[19], IGF1R (PDB: 5U8R)[20], Tie2 (PDB: 2GY7)[21], IL-7Rα (PDB: 3DI3)[22], CD3 (PDB: 1XIW)[23], TGF-β (PDB: 3KFD)[24] and VirB8 (PDB: 4O3V)[26] were refined in the Rosetta energy field constrained by experimental diffraction data. The crystal structures of TrkA (PDB: 1WWW)[15] and FGFR2 (PDB: 1EV2)[16] were refined with the Rosetta FastRelax protocol with coordinate constraints. The targeting chain or the selected targeting region were extracted and used as the starting point for docking and design. To run PatchDock[11], the scaffolds were mutated to poly-valine first and default parameters were used to generate the raw docks. Rifdock was used to generate the rotamer interacting field by docking billions of individual disembodied amino acids to the selected targeting regions[6]. In detail, hydrophobic sidechain R-groups are docked against the target using a branch-and-bound search to quickly identify favorable interactions with the target, and polar sidechain R-groups are enumeratively sampled around every target hbond donor or acceptor. To identify backbone placements from which these interactions can be made, side chain rotamer conformations are grown backwards for all R-group placements, and their backbone coordinates stored in a 6-dimensional spatial hash table for rapid lookup. For the hierarchical searching protocol, the miniprotein scaffold library (50-65 residues in length) was docked into the field of the inverse rotamers using a branch-and-bound searching algorithm from low resolution spatial grids to high resolution spatial grids. For the PatchDock+Rifdock protocols, the PatchDock outputs were used as seeds for the initial positioning of the scaffolds and the docks were further refined in the finest resolution rotamer interaction field. These docked conformations were further optimized to generate shape and chemically complementary interfaces using the Rosetta FastDesign protocol, activating between side-chain rotamer optimization and gradient-descent-based energy minimization. Serval improvements were added to the sequence design protocol to generate better sequences for both folding and binding. These include a better repulsive energy ramping strategy[9], upweighting cross-interface energies, a pseudo-energy term penalizing buried unsatisfied polar atoms[13] and a sequence profile constraint based on native protein fragments[12]. Computational metrics of the final design models were calculated using Rosetta, which includes ddg, shape complementary and interface buried solvent accessible surface area, contact molecular surface, etc, for design selection. All the script and flag files to run the programs are in the Supplementary Information.

Focused search stage

The binding energy and interface metrics for all the continuous secondary structure motifs (helix, strand and loop) were calculated for the designs generated in the broad search stage. The motifs with good interaction (based on binding energy and other interface metrics, like SASA, contact molecular surface) with the target were extracted and aligned using the target structure as the reference. All the motifs were then clustered based on an energy based-TMalign like clustering algorithm. Briefly, all the motifs were sorted based on the interaction energy with the target, and the lowest energy motif in the unclustered pool was selected as the center of the first cluster. A similar score between this motif and every motif remaining in the unclustered pool was calculated based on the TMalign algorithm[43] without any further superimposition. Those motifs within a threshold similar score (default 0.7) from the current cluster center were removed from the unclustered pool and added to the new cluster. The lowest-energy motif remaining in the unclustered pool was selected as the center of the next cluster, and the second step was repeated. This process continued for subsequent clusters until no motifs remained in the unclustered pool. The best motif from each cluster was then selected based on the per-position weighted Rosetta binding energy, using the average energy across all the aligned motifs at each position as the weight. Around 2,000 best motifs were selected and the scaffold library was superimposed onto these motifs using the MotifGraft mover[44]. Interface sequences were future optimized and computational metrics were computed for the final optimized designs as described in the broad search stage. CPU-time requirements to produce 100,000 designed binders to be tested experimentally were typically around 100,000 CPU-hours (usually at least 10x as many binders were computationally designed than were ordered).

Rapid Rosetta packing evaluation (The Predictor)

A severe speed mismatch exists between the docking methods (RifDock and Focused search) and the subsequent full sequence design step. While the docking methods can typically produce outputs every 1 to 3 sec, the full sequence design can take upwards of 4 min. To remedy this situation, a step was designed to take about 20 sec that would be more predictive than metrics evaluated on raw docks, but faster than the full sequence design.

A stripped down version of the Rosetta beta_nov16 score function was used to design only with hydrophobic amino acids. Specifically, fa_elec, lk_ball[iso,bridge,bridge_unclp], and the _intra_ terms were disabled as these proved to be the slowest energy methods by profiling. All that remained were Lennard-Jones, implicit solvation, and backbone-dependent one-body energies (fa_dun, p_aa_pp, rama_prepro). Additionally, flags were used to limit the number of rotamers built at each position (See Supplementary Information).

After the rapid design step, the designs are minimized twice: once with a low-repulsive score function and again with a normal-repulsive score function. Metrics of interest were then evaluated including like Rosetta ddG, Contact Molecular Surface, and Contact Molecular Surface to critical hydrophobic residues.

Using the observation that these predicted metrics correlate with the values after full sequence design, a Maximum Likelihood Estimator (functional form similar to logistic regression) was used to give each predicted design a likelihood that it should be selected to move forward. A subset of the docks to be evaluated are subjected to the full sequence design, and their final metric values calculated. With a "goal threshold" for each filter, each fully-designed output can be marked as "pass" or "fail" for each metric independently. Then, by binning the fully-designed outputs by their values from the rapid trajectory and plotting the fraction of designs that pass the "goal threshold", the probability that each predicted design passes each filter can be calculated (sigmoids are fitted to smooth the distribution). From here, the probability of passing each filter may be multiplied together to arrive at the final probability of passing all filters. This final probability can then be used to rank the designs and pick the best designs to move forward to full sequence optimization.

Note: the rapid design protocol here is used merely to rank the designs, not to optimize them; the raw, non-rapid-designed docks are the structures carried forward.

Contact molecular surface

Solvent-accessible surface area (SASA) is a measure of the exposure of amino acids to the solvent and it is typically calculated by methods involving in-silico rolling of a spherical probe, which approximates a water molecule (radius 1.4Å), around a full-atom protein model. Delta-SASA upon protein-protein binding has been widely used to analyze native protein interactions. Unlike the crystal structures of the native protein complexes, design models for the de novo interactions are usually imperfectly packed, and contain many holes or cavities. If the sizes of the holes or cavities in the interface are smaller than the rolling probe, the SASA cannot capture those holes and cavities and the real contacts are usually overestimated by the delta-SASA metric. The contact molecular surface was developed to capture the flaws

of the de novo designed interactions. Firstly, the molecule surfaces of the binder and the target were calculated by the triangularization algorithm in the Rosetta shape complementary filter. For each triangle, the distance to the closest triangle on the other side was calculated and used to down-weight the area of the triangle by the equation: A' = A * exp(-0.5*distance^2). Then all the down-weighted areas were summed up to get the contact molecular surface. In this way, the real contacts between the target and the binder are penalized by the cavities and holes in the interface. The contact molecular surface was implemented as the ContactMolecularSurface filter in the Rosetta macromolecular modelling suite.

Upweighted protein interface interactions

Rosetta sequence design starts from generating an interaction graph by calculating the energies between all designable rotamer pairs[45]. The best rotamer combinations are searched using a Monte Carlo Simulated Annealing protocol by optimizing the total energy of the protein (monomer/complex). To obtain more contacts between the binder and the target protein, we can upweight the energies of all the cross interface rotamer pairs by a defined factor. In this way, the Monte Carlo protocol will be biased to find solutions with better cross interface interactions. The upweighted protein interface interaction protocol was implemented as the ProteinProteinInterfaceUpweighter task operation in the Rosetta macromolecular modelling suite.

Comparison of sampling efficiency of PatchDock, RifDock, and resampling protocols

The top 30 PatchDock outputs for the 1,000 helical scaffolds tested were designed using the RosettaScripts protocol (red). RifDock seeded with PatchDock outputs generated 300 outputs per scaffold which were trimmed to a total of 19,500 docks with the Predictor (see Methods) and subsequently designed (green). The top 150 RifDock outputs per scaffold were trimmed to 9,750, designed, and 300 motifs were extracted. The motifs were grafted into the scaffold set to produce 150,000 docks, which were trimmed to 9,750, designed, and combined with the earlier 9,750 (orange).

DNA library preparation

All protein sequences were padded to 65 aa by adding a (GGGS)n linker at the C terminus of the designs, to avoid the biased amplification of short DNA fragments during PCR reactions. The protein sequences were reversed translated and optimized using DNAworks2.0[46] with the *S. cerevisiae* codon frequency table. Oligo pool encoding the *de novo* designs and the point mutant library were ordered from Agilent Technologies. Combinatorial libraries were ordered as IDT (Integrated DNA Technologies) ultramers with the final DNA diversity ranging from 1e6 to 1e7.

All libraries were amplified using Kapa HiFi Polymerase (Kapa Biosystems) with a qPCR machine (BioRAD CFX96). In detail, the libraries were firstly amplified in a 25 µl reaction, and PCR reaction was terminated when the reaction reached half maximum yield to avoid over amplification. The PCR product was loaded to a DNA agarose gel. The band with the expected size was cut out and DNA fragments were extracted using QIAquick kits (Qiagen, Inc.). Then, the DNA product was re-amplified as before to generate enough DNA for yeast transformation. The final PCR product was cleaned up with a QIAquick Clean up kit (Qiagen, Inc.). For the yeast transformation, 2-3 µg of linearized modified pETcon vector (pETcon3) and 6 µg of insert were transformed into EBY100 yeast strain using the protocol as described[47].

DNA libraries for deep sequencing were prepared using the same PCR protocol, except the first step started from yeast plasmid prepared from 5×10^7 to 1×10^8 cells by Zymoprep (Zymo Research). Illumina adapters and 6-bp pool-specific barcodes were added in the second qPCR step. Gel extraction was used to get the final DNA product for sequencing. All different sorting pools were sequenced using Illumina NextSeq sequencing.

Target protein preparation

Influenza A hemagglutinin (HA) ectodomain was expressed using a baculovirus expression system as described previously[25,48]. Briefly, each HA was fused with gp67 signal peptide at the N terminus and to a BirA biotinylation site, thrombin cleavage site, trimerization domain and His-tag at the C terminus. Expressed HAs were purified using metal affinity chromatography using Ni^{2+}-NTA resin. For binding studies, each HA was biotinylated with BirA and purified by gel filtration using S200 16/90 column on ÄKTA protein purification system (GE Healthcare). The biotinylation reactions contained 100 mM Tris (pH 8.5), 10 mM magnesium acetate,10 mM ATP, 50 µM biotin and <50 mM NaCl, and were incubated at 37 °C for 1 hr.

For TrkA, the DNA encoding human TrkA ECD (residues 36-382) was cloned into pAcBAP, a derivative of pAcGP67-A modified to include a C-terminal biotin acceptor peptide (BAP) tag (GLNDIFEAQKIEWHE) followed by a 6xHIS tag for affinity purification. It was then transfected into *Trichoplusia ni* (High Five) cells (InVitrogen) using the BaculoGold baculovirus expression system (BD Biosciences) for secretion and purified from the clarified supernatant via Ni-NTA followed by size exclusion chromatography with a Superdex-200 column in sterile phosphate-buffered saline (PBS) (Cat. 20012-027; Gibco). The ectodomains of FGFR2 (residues 147-366, Uniprot ID P21802), EGFR (residues ID 25-525, Uniprot ID P00552), PDGFR (residues 33-314, Uniprot ID P09619), InsulinR (residues ID 28-953, Uniprot ID P06213), IGF1R (residues 31-930, Uniprot ID P08069), Tie2 (residues 23-445, Uniprot ID Q02763), IL-7Rα (residues 37-231, Uniprot ID P16871) were expressed in mammalian cells with a IgK Signal peptide (METDTLLLWVLLLWVPG STG) at the N terminus and a C-terminal tag (GSENLYFQGSHHHHH HGSGLNDIFEAQKIEWHE) that contains a TEV cleavage site, a 6-His-tag and an AviTag. VirB8 was expressed in *E. coli* with a C-terminal AviTag as previously described[26]. The proteins were purified by Ni^{2+}-NTA, and polished with size exclusion chromatography. Then, the AviTag-proteins were biotinylated with the BirA biotin-protein ligase bulk reaction kit (Avidity) following the manufacturer's protocol and the excessive biotin was removed through size exclusion chromatography. Biotinylated CD3 protein was bought from Abcam (Cat# ab205994). TGF-β was bought from Acro Biosystems (Cat# TG1-H8217). IGF1 was bought from Sigma (Cat# 407251-100ug). Insulin was bought from Abcam (Cat# ab123768). The caged Ang1-Fc protein was prepared as described previously[49], and was kindly provided by George Ueda. The FI6v3 antibody was kindly provided by D.H. Fuller at University of Washington.

Yeast surface display

S. cerevisiae EBY100 strain cultures were grown in C-Trp-Ura media supplemented with 2% (w/v) glucose. For induction of expression, yeast cells were centrifuged at 6,000x g for 1 min and resuspended in SGCAA media supplemented with 0.2% (w/v) glucose at the cell density of $1x10^7$ cells per ml and induced at 30 °C for 16–24 hr. Cells were washed with PBSF (PBS with 1% (w/v) BSA) and labelled with biotinylated targets using two labeling methods, with-avidity and without-avidity labeling. For the with-avidity method, the cells were incubated with biotinylated target, together with anti-c-Myc fluorescein isothiocyanate (FITC, Miltenyi Biotech) and streptavidin–phycoerythrin (SAPE, ThermoFisher). The concentration of SAPE in the with-avidity method was used at ¼ concentration of the biotinylated targets. For the without-avidity method, the cells were firstly incubated with biotinylated targets, washed, secondarily labelled with SAPE and FITC. All the original libraries of de novo designs were sorted using the with-avidity method for the first few rounds of screening to fish out weak binder candidates, followed by several without-avidity sorts with different concentrations of targets. For SSM libraries, two rounds of without-avidity sorts were applied and in the third round of screening, the libraries were titrated with a series of decreasing concentrations of targets to enrich mutants with beneficial mutations. The combinatorial

libraries were sorted to convergence by decreasing the target concentration with each subsequent sort and collecting only the top 0.1% of the binding population. The final sorting pools of the combinatorial libraries were plated on C-trp-ura plates and the sequences of individual clones were determined by Sanger sequencing. The competition sort was done following the without-avidity protocols with a very minor modification. Briefly, the biotinylated target proteins (H1, H3, TrkA, InsulinR, IGF1R, PDGFR and Tie2) were first incubated with an excessive amount of competitors (FI6v3, FI6v3, NGF, insulin, IGF1, PDGF and caged Ang1-Fc) respectively for 10 mins, and the mixture was used for labeling the cells. The non-specificity reagent was prepared using the protocol as described[50]. For non-specificity sort, the cells were firstly washed with PBSF and incubated with the non-specificity reagent at the concentration of 100 ug/ml for 30 min. The cells were then washed and secondarily labelled with SAPE and FITC for cell sorting. The cells were then labeled with RBD using the above mentioned protocol.

Miniprotein expression

Genes encoding the designed protein sequences were synthesized and cloned into modified pET-29b(+) *E. coli* plasmid expression vectors (GenScript, N-terminal 8 His-tag followed by a TEV cleavage site). For all designed proteins, the sequence of the N-terminal tag is MSHHHHHH HHHSENLYFQSGGG (unless otherwise noted), which is followed immediately by the sequence of the designed protein. For proteins expressed with the maltose binding protein (MBP) tag, the corresponding genes were subcloned into a modified pET-29b(+) *E. coli* plasmid, which has a N-terminal 6 His-tag and a MBP tag. Plasmids were transformed into chemically competent *E. coli* Lemo21 cells (NEB). For the designs for TrkA, FGFR2, EGFR, IR, IGF1R, Tie2, IL-7Rα, TGF-β and the MBP tagged miniproteins, protein expression was performed using the Studier autoinduction media supplemented with antibiotic, and cultures were grown overnight. For designs for HA, PDGFR and CD3δ, the *E. coli* cells were grown in LB media at 37 °C until the cell density reached 0.6 OD_{600}. Then, IPTG was added to the final concentration of 500 mM and the cells were grown overnight at 22 °C for expression. The cells were harvested by spinning at 4,000x g for 10 min and then resuspended in lysis buffer (300 mM NaCl, 30 mM Tris-HCL (pH 8.0), with 0.25% CHAPS for cell assay samples) with DNase and protease inhibitor tablets. The cells were lysed with a QSONICA SONICATORS sonicator for 4 minutes total (2 min on time, 10 sec on-10 sec off) with an amplitude of 80%. Then the soluble fraction was clarified by centrifugation at 20,000x g for 30 min. The soluble fraction was purified by Immobilized Metal Affinity Chromatography (Qiagen) followed by FPLC size-exclusion chromatography (Superdex 75 10/300 GL, GE Healthcare). All protein samples were characterized with SDS-PAGE with the purity higher than 95%. Protein concentrations were determined by absorbance at 280 nm measured using a NanoDrop spectrophotometer (Thermo Scientific) using predicted extinction coefficients.

Circular dichroism

Far-ultraviolet CD measurements were carried out with an JASCO-1500 equipped with a temperature-controlled multi-cell holder. Wavelength scans were measured from 260 to 190 nm at 25, 95 °C and again at 25 °C after fast refolding (~5 min). Temperature melts monitored dichroism signal at 222 nm in steps of 2 °C/minute with 30s of equilibration time. Wavelength scans and temperature melts were performed using 0.3 mg/ml protein in PBS buffer (20mM $NaPO_4$, 150mM NaCl, pH 7.4) with a 1 mm path-length cuvette. Melting temperatures were determined fitting the data with a sigmoid curve equation. 9 out of the 13 designs retained more than half of the mean residue ellipticity values, which indicated the Tm values are greater than 95 °C. Tm values of the other designs were determined as the inflection point of the fitted function.

Biolayer interferometry

Biolayer interferometry binding data were collected on an Octet RED96 (ForteBio) and processed using the instrument's integrated software. For minibinder binding assays, biotinylated targets were loaded onto streptavidin-coated biosensors (SA ForteBio) at 50 nM in binding buffer (10 mM HEPES (pH 7.4), 150 mM NaCl, 3 mM EDTA, 0.05% surfactant P20, 1% BSA) for 6 min. Analyte proteins were diluted from concentrated stocks into the binding buffer. After baseline measurement in the binding buffer alone, the binding kinetics were monitored by dipping the biosensors in wells containing the target protein at the indicated concentration (association step) and then dipping the sensors back into baseline/buffer (dissociation). The binding affinities of Tie2- and IGF1R- mini binders were low, and MBP tagged proteins were used for the binding assay to amplify the binding signal. The binding assay for the Insulin receptor (IR) designs were conducted with Amine Reactive Second-Generation (AR2G ForteBio) Biosensors with the recommended protocol. In brief, the miniproteins were immobilized onto the AR2G tips and the InsulinR were used as the analyte with the indicated concentrations. Data was analysed and processed by ForteBio Data Analysis Software Version 9.0.0.14.

For the cross-reactivity assay, each target protein was loaded onto SA tips at the concentration of 50nM for 325s. The tips were dipped into the miniprotein wells for 300s (association) and then dipped into the blank buffer wells for 600s (dissociation). The maximum raw bio-layer Interferometry signal binding was used as the indicator of binding strength. The maximum signal among all the miniprotein binders for a specific target was used to normalize the data for heatmap plotting.

Crystallization and structure determination of the H3_mb in complex with HK68/H3

To prepare the H3_mb-HK68/H3 HA complex for crystallization, a five-fold molar excess of H3_mb was mixed with ~2 mg/mL of HK68/H3 HA in 20 mM Tris (pH 8.0), 150 mM NaCl. The mixture was incubated overnight at 4 °C to allow complex formation. Saturated complexes were then purified from unbound HB_mb by gel filtration. Gel filtration fractions containing the H3_mb-HK68/H3 HA complex were concentrated to ~7 mg/mL in 20 mM Tris (pH 8.0), 150 mM NaCl. Crystallization screens were set up using the sitting drop vapor diffusion method with our automated CrystalMation robotic system (Rigaku) at The Scripps Research Institute. Within 3-7 days, diffraction quality crystals had grown in 0.2 M sodium thiocyanate and 20% (w/v) polyethylene glycol 3350 as a precipitant. The resulting crystals were cryoprotected by addition of 5-15% ethylene glycol, flash cooled, and stored in liquid nitrogen until data collection. Diffraction data were collected at 100 K at the Stanford Synchrotron Radiation Lightsource (SSRL) beamline 12-1 and processed with HKL-2000[51]. Initial phases were determined by molecular replacement using Phaser[52,53] with an HA model from PDB code 4FNK (apo HK68/H3 HA). Refinement was carried out in Phenix[54], alternating with manual rebuilding and adjustment in COOT[55]. Electron density maps were calculated using Phenix Data collection and refinement statistics are summarized in Extended Data Table 2. The final coordinates were validated using MolProbity[56].

Crystal structure of TrkA in complex with the miniprotein binder

The human TrkA receptor extracellular domain was produced in insect cells using baculovirus prepared as described[57]. Hi5 cells were coinfected in shaking Fernbach flasks with baculoviruses encoding TrkA ECD and endoglycosidase H in the presence of kifunensin. Cultures were allowed to progress for 65 hr before the supernatant was recovered by centrifugation. Media components were precipitated by the addition of 50 mM Tris (pH 8.0), 1 mM $NiCl_2$, and 5 mM $CaCl_2$, and the supernatant was filtered over diatomaceous earth. The filtrate was batch-bound to Ni^{2+}-NTA resin, eluted with 200 mM imidazole in HBS (HEPES-buffered saline: 10 mM HEPES (pH 7.3), 150 mM NaCl), and purified by size exclusion chromatography (SEC) on a Superdex-75 column (Cytiva Life Sciences). To prepare the TrkA/miniprotein complex, an

excess amount of miniprotein was mixed with TrkA, digested overnight at 4 °C with 1:100 (w/w) carboxypeptidases A and B, and purified by SEC.

For crystallization, the TrkA/ligand complex was concentrated to 38 mg/ml in HBS and screened in sitting drop format using a Mosquito crystallization robot (SPT Labtech). Initial "sea urchin"-like crystals were obtained from the MCSG1 screen (Anatrace-Microlytic) in 0.17 M ammonium acetate, 0.085 M sodium citrate (pH 5.6), 25.5% PEG 4000, and 15% glycerol. These crystals were crushed and used to microseed the MCSG1 screen again at a ratio of 3:2:1 protein:precipitant:seed stock, resulting in single plate-like crystals grown from 0.2 M ammonium sulfate, 0.1 M bis-Tris (pH 6.5), and 25% PEG 3350. After further optimization to 0.4 M ammonium sulfate, 0.1 M bis-Tris (pH 6.2), and 20% PEG 3350, new seeds were prepared for final seeding into 0.4 M ammonium sulfate, 0.1 M bis-Tris (pH 6.2), and 16% PEG 3350.

Crystals were cryoprotected by addition of ethylene glycol to 30% (v/v) and flash cooled in liquid nitrogen. Diffraction data to 1.84 Å resolution were collected at 100 K using an X-ray wavelength of 1.033 Å at the Stanford Synchrotron Radiation Laboratory (SSRL) beamline 12-2. Crystals were assigned to space group P21 with unit cell dimensions a = 42.20 Å, b = 205.70 Å, c = 72.57 Å, β = 106.42°. Data were indexed, integrated, and scaled using XDS[58,59] and merged using Pointless and Aimless from the CCP4 suite[60–62].

The structure was solved by molecular replacement in Phaser[52] using separated domains of TrkA ECD (PDB accession 2IFG) and the predicted model of the ligand as search models to place two copies of the complex in the asymmetric unit. Initial rebuilding was completed with phenix.autobuild[63] followed by iterative rounds of manual rebuilding in Coot[64] and refinement in Phenix[65–67]. TLS parameters were chosen using TLSMD[68] and NCS restraints were used throughout refinement[69]. The final resolution of the data was selected as 1.84 Å by comparing the results of paired refinements at 1.84, 1.90, 1.95, 2.00, and 2.05 Å resolution[70]. The final refined model included 97.26% of residues in the favored region of the Ramachandran plot with 0.25% outliers as calculated by MolProbity[56].

Crystallographic software used in this study was configured and installed by SBGrid[71]. Diffraction images have been deposited in the SBGrid Data Bank with ID 839 and the final model and reflections have been deposited in the RCSB Protein Data Bank with ID 7N3T.

Crystal structures of FGFR2_mb in complex with FGFR4 domain 3 and FGFR2_mb alone

cDNA of human Fibroblast Growth Factor receptor 4 (FGFR4) domain 3 ($FGFR4_{D3}$, amino acids S245–D355) was amplified with PCR and cloned into pET-28a(+) plasmid (Novagen). The plasmid containing $FGFR4_{D3}$ with N-terminal hexa-histidine tag was transformed into BL21(DE3) cells. The transformed cells were grown in LB at 37 °C until OD_{600} reached 0.5, induced with 1.0 mM IPTG, grown for additional 4 hr at 37 °C, and harvested. The bacterial cells were resuspended and lysed by sonication. $FGFR4_{D3}$ was refolded from insoluble fractions using previously reported procedure[16,72,73], and purified to homogeneity using nickel affinity chromatography (Ni^{2+}-NTA agarose, Qiagen) followed by size exclusion chromatography (Superdex 200 Increase 10/300 GL, Cytiva) equilibrated with a buffer containing 200 mM NaCl, 25 mM HEPES (pH 8.0), and 5% glycerol. The purified $FGFR4_{D3}$ was mixed with a 1.2-fold molar excess of FGFR2_mb and subjected to another round of size exclusion chromatography to isolate the $FGFR4_{D3}$:FGFR2_mb complex. Fractions containing $FGFR4_{D3}$ bound to FGFR2_mb were pooled and concentrated to 12 mg/mL and screened for crystallization using commercially available crystallization screening kits using Mosquito Crystal liquid handler (SPT Labtech). Crystals of $FGFR4_{D3}$:FGFR2_mb complex were obtained with the ProPlex screening solution (Molecular Dimensions) containing 0.2 M sodium chloride, 0.1 M MES pH 6.0, and 20% PEG 3,350 at 4 °C. The crystals were cryoprotected using the mother liquor supplemented with 25% glycerol before being flash-cooled in liquid nitrogen.

Crystals of FGFR2_mb were obtained using the solution containing alcohols (0.02 M 1,6-hexanediol, 0.02 M 1-butanol, 0.02 M 1,2-propanediol, 0.02 M 2-propanol, 0.02 M 1,4-butanediol, 0.02 M 1,3-propanediol), buffer mixture (0.1 M Tris and BICINE adjusted to pH 8.5), and precipitants (12.5% v/v MPD, 12.5% PEG 1000, 12.5% w/v PEG 3,350) by hanging-drop vapor diffusion method at 20 °C, which were directly flash-cooled in liquid nitrogen for X-ray diffraction data collection.

X-ray diffraction data were collected at the NE-CAT 24ID-E beam line of Advanced Photon Source (Argonne National Laboratory) and processed with XDS[74]. The initial structure of FGFR2_mb was obtained by molecular replacement with PHASER[52,75] using the designed model, which was iteratively refined using PHENIX[67,75] followed by manual building with COOT[64]. The structure of $FGFR4_{D3}$:FGFR2_mb complex was obtained by molecular replacement with Phaser[52,75] using the coordinates corresponding to the domain 3 region of FGFR1c[72] (PDB ID: 1CVS) and the coordinates of FGFR2_mb as the search model, followed by iterative refinements using PHENIX[67,75] and COOT[64]. The final structures were validated with MolProbity[75,76]. Data collection and refinement statistics are provided in Extended Data Table 2.

Crystal structure of unbound IL-7Rα minibinder

To facilitate crystallization, the N-terminal His-Tag was removed using TEV protease and the protein was concentrated to 40 mg/mL in 30 mM Tris-HCl (pH 8.0) and 150 mM NaCl. Sparse-matrix crystal screening was performed using kits from Hampton Research (Index-HT, PEGRx-HT, and PEG/Ion-HT) at room temperature. A Mosquito nanoliter crystallization robot was used to set up sitting drops consisting of 200 nL of protein and 200 nL of each reservoir solution with 80 µL of reservoir solution in MRC-2 plates. Promising prism-shaped crystals grew from IndexHT C3 condition and optimal conditions ranged from 2.4-3.0 M sodium malonate (pH 7.0). Protein crystals were cryo-cooled directly into liquid nitrogen. Initial X-ray diffraction experiments were carried out on a home source system equipped with MicroMax-007 HF rotating anode with a Dectris Eiger R 4M single-photon counting device. X-ray diffraction data on optimized protein crystals were collected at the Advanced Photon Source synchrotron beamline 23ID-D of GM/CA with a Dectris Pilatus3-6M detector. All X-ray data was processed with XDS. Molecular replacement using the *de novo* designed model was used to solve the crystal structure using Phaser within the Phenix package. Two molecules were located in the asymmetric unit. Structural refinement used Phenix using no NCS restraints. Data collection and refinement statistics are given in Extended Data Table 2.

Crystal structure of IL-7Rα in complex with the minibinder

The ectodomain of human IL-7Rα was produced and purified as previously described[77]. The anti-IL-7Rα minibinder was prepared as described above. The IL-7Rα:minibinder complex was formed by adding a molar excess of purified minibinder to recombinant IL-7Rα. The IL-7Rα:minibinder complex was purified using size-exclusion chromatography using a Superdex-75 column (Cytiva Life Sciences) with HBS buffer (pH 7.4) as the running buffer. Fractions corresponding to the IL-7Rα-minibinder complex were pooled and concentrated by centrifugal ultra-filtration to a concentration of 3.9 mg/ml. Sparse-matrix crystallization screens were carried out using the BCS-Screen (Molecular Dimensions) at 293K via the sitting-drop. The vapour-diffusion geometry was used to set up sitting drops consisting of 200 nL of protein and 100 nL of each reservoir solution. using a Mosquito nanoliter crystallization robot (TTP Labtech). The IL-7Rα-mininbinder complex crystallized in condition A5 (0.1 M phosphate / citrate (pH 5.5), 25.0% PEG Smear Medium). Crystals were cryo-protected with mother liquor supplemented with 25% v/v polyethylene glycol 400 and cryo-cooled by direct plunging into liquid nitrogen. X-ray diffraction data of protein crystals were collected at beamline ID23-2 of the ESRF (Grenoble, France) with a Dectris PILATUS3 X 2M detector and were processed

with XDS[58]. The structure was determined by maximum-likelihood molecular replacement in Phaser using the crystal structure of IL-7Rα (PDB entry 3DI2) as a search model[52]. Three copies of the complex were located in the asymmetric unit. Model (re)building was performed in Coot[64], and coordinate and ADP refinement was performed in PHENIX[65] and autoBuster[78]. Model and map validation tools in Coot, the PHENIX suite and the PDB_REDO server[79] were used to validate the quality of crystallographic models. The final model and reflections have been deposited in the RCSB Protein Data Bank with access code 7OPB. Data collection and refinement statistics are provided in Extended Data Table 2.

Crystal structure of VirB8 like protein in complex with the minibinder

VirB8-like protein of type IV secretion system from Rickettsia typhi (Uniprot ID: Q68X84) in complex with 0.75 mM VirB8 miniprotein binder was suspended in a buffer containing 20 mM HEPES pH 7.0, 300 mM NaCl, and 5% glycerol. The complex was crystallized using the sitting drop vapor diffusion method at 14 °C with drops composed of 0.4 mL of the complex at 9.9 mg/mL mixed with 0.4 mL crystallant (sparse matrix screen JCSG Top96 (Rigaku Reagents) condition G9: 100 mM sodium acetate/hydrochloric acid (pH 4.6), 25% (w/v) PEG 4000, 200mM Ammonium sulfate) equilibrated against 80 mL crystallant in the reservoir. Crystals were cryoprotected in the crystallant supplemented with 15% (v/v) ethylene glycol. X-ray diffraction data of the VirB8 protein/miniprotein binder complex was collected at the LS-CAT beamline 21-ID-F at the Advanced Photon Source. Data were integrated in XDS and reduced using XSCALE[58]. Data quality assessed using POINTLESS[80]. Molecular replacement was performed using Phaser[52] using search models comprised of previously solved crystal structure of Rickettsia typhi VirB8-like of type IV secretion system (PDBID: 403V) and an Alphafold2[81] predicted model of the VirB8 miniprotein binder. Iterative manual model building and refinement were carried out using Coot[64] and Phenix[65]. Structure quality was assessed using Molprobity[56] prior to deposition in the Protein Data Bank[82,83] (See Extended Data Table 2). Diffraction images are available on Integrated Resource for Reproducibility in Macromolecular Crystallography[84,85].

Comparison between the crystal structures and design models

For the structures of the miniprotein binders in complex with the targets, the whole structures were aligned using the target as the references first. The root-mean-square deviation (RMSD) over the Cα atoms of the entire miniprotein binder was calculated. For the unbound crystal structures of the FGFR2 miniprotein binder and the IL-7Rα miniprotein binder, the RMSD were calculated over all the Cα atoms after superimposition. For the analysis of the heavy atoms of the interface core residues, the structures were aligned using the target as references first. Interface residues of the binders were selected as long as there is one residue on the target that has Cβ-Cβ distance less than 8 Å using the NeighborhoodResidueSelector, and core residues were selected using the LayerSelector in Rosetta with the default burial cutoff value. Then heavy atoms of the interface core residues were used to calculate the RMSD values. Four, eight, six and six residues were considered as interface core residues for the H3, FGFR2, IL-7Rα and VirB8 complex structures respectively.

TrkA minibinder antagonist assay

The Phospho-flow signaling assay was used to characterize the antagonistic properties of the TrkA minibinder. TF-1 cells (ATCC CRL-2003) were starved for 4 hr in base media without NGF or other cytokines prior to signaling assays. Cells were plated in 96-well plates with different concentrations of TrkA binder and stimulated with human beta-NGF (R&D) for 10 min at 37 °C, followed by fixation with 1.6% paraformaldehyde for 10 min at room temperature. Then, cells were permeabilized by resuspension in ice-cold methanol and stored at -20 °C until flow cytometry analysis. For intracellular staining, the permeabilized cells were washed and incubated with Alexa Fluor-488 conjugated anti-ERK1/2 pT202/pY204 antibody (BD) and Alexa Fluor-647 conjugated anti-Akt pS473 antibody (Cell Signaling Technology) for 1 hr at room temperature. After washing with autoMACS running buffer (Miltenyi), fluorescence intensity of each antibody staining level was acquired using CytoFlex flow cytometer (Beckman Coulter). Mean fluorescence intensity (MFI) values were background subtracted and normalized to the maximal MFI value in the absence of TrkA binder and plotted in Prism 9 (GraphPad). The dose-response curves were generated using the "sigmoidal dose-response" analysis.

For the cell proliferation assay, TF-1 cells were plated in 96-well plate and cultured in RPMI-1640 media containing 2% FBS and different concentrations of TrkA binder and NGF for 48 hr at 37 °C. Cell proliferation rate was assessed by measuring cellular ATP level using CellTiter-Glo 2.0 Cell Viability Assay reagent (Promega) according to the manufacturer's protocol. Luminescent signal was measured using SpectraMax Paradigm plate reader, and the data were plotted and analyzed by Prism 9 (GraphPad). The dose-response curves were generated using the "sigmoidal dose-response" analysis.

FGFR2 and EGFR minibinder antagonist assay

Cell Culture: Human Umbilical Vein Endothelial Cells, or HUVECs (Lonza, Germany, Catalog #C2519AS) were grown in EGM2 media on 35-mm cell culture dishes coated with 0.1% gelatin. Briefly, EGM2 is composed of 20% Fetal Bovine Serum, 1% penicillin-streptomycin, 1% GlutaMAX (Gibco, Cat. #35050061), 1% ECGS (Endothelial cell growth factor), 1 mM sodium pyruvate, 7.5 mM HEPES, 0.08 mg/mL heparin, and 0.01% amphotericin B in a mixture of 1x RPMI-1640 with and without glucose (final glucose concentration = 5.6 mM). Media was filtered through a 0.2-μm filter. HUVECs were serially passaged and expanded before cryopreservation.

FGFR/EGFR Antagonist Assay

Frozen HUVECs were thawed and cultured in a 35-mm dish in EGM2 media until confluency was reached. After that, EGM2 media was aspirated and rinsed twice with 1x PBS. Cells were then serum-starved by adding 2 mL of DMEM serum-free media (1 g/L glucose, Gibco) for 16 hr, after which the starvation media was aspirated. The cells were then treated with the FGFR2 mini binder or the EGFR mini binder for 1 hr at 37 °C and at concentrations varying between 5 nM and 1 μM of minibinder. This was followed by stimulation with βFGF (0.75 nM, Fisher Scientific) or EGF (1 nM, Peprotech) respectively, for 15 min at 37 °C. After treatment, the media was aspirated, and cells were washed once with 1x PBS before harvesting total protein for analysis.

Total Protein Isolation

After mini binder treatment, the cells were gently rinsed in 1X PBS before lysis with 130 μL of lysis buffer containing 20 mM Tris-HCL (pH 7.5), 150 mM NaCL, 15% glycerol, 1% triton, 3% SDS, 25 mM β-glycerophosphate, 50 mM NaF, 10 mM sodium pyrophosphate, 0.5% orthovanadate, 1% PMSF (all obtained from Sigma-Aldrich, St. Louis, MO), benzonase nuclease (EMD Chemicals, Gibbstown, NJ), protease inhibitor cocktail (PierceTM Protease Inhibitor Mini Tablets, Thermo Scientific, USA), and Phosphatase Inhibitor Cocktail 2 (Cat. #P5726). Cell lysate was collected in a fresh Eppendorf tube. 43.33 μL of 4X Laemmli sample buffer (Bio-Rad, USA) (containing 10% β-mercaptoethanol) was added to the cell lysate and then heated at 95 °C for 10 min. The boiled samples were either used for western blot analysis or stored at -80 °C.

Western Blotting

30 μL of protein lysate was loaded per well and separated on a 4-20% SDS-PAGE gel for 30 min at 250 V. Proteins were then transferred onto a nitrocellulose membrane for 12 min using the semi-dry turbo transfer apparatus (Bio-Rad, USA). The membranes were blocked in 5% BSA for 1 hr, after which they were probed overnight with respective primary

antibodies on a rocker at 4 °C. The primary antibodies used in this assay were β-actin (1:10,000, Cell Signaling Technologies), p-ERK1/2 p44/42 (1:10,000, Cell Signaling Technologies) and p-AKT S473 (1:2,000, Cell Signaling Technologies). The next day, membranes were washed thrice with 1x TBS-T and then incubated with anti-rabbit HRP conjugated secondary antibody (1:10,000, Bio-Rad Laboratories) for 1 hr. For p-AKT S473, following washes, the membrane was blocked in 5% milk at room temperature for 1 hr and then incubated in the respective HRP-conjugated secondary antibody (1:2,000) prepared in 5% milk, for 1 hr. They were developed using Immobilon Western chemiluminescent substrate (EMD Millipore), followed by quantification using the NIH ImageJ analysis software. The raw scans of the Western results are shown as Supplementary Data Figure 5. Quantifications were done by calculating the peak area for each band. Inhibition curve fit and corresponding IC_{50}'s were determined using GraphPad Prism 9 software.

IL-7Rα minibinder antagonist assay

HEK293T cells were cultured in DMEM medium with 10% FBS at 37 °C and 5% CO_2. Cells were co-transfected with 1000 ng pcDNA3-γ common, 300 ng pMET7-HA-IL-7Rα, 200 ng pMX-IRES-GFP-hJak3, 300 ng empty pMET7 vector and 200 ng pGL3-b-casein-luci STAT5 reporter plasmid per well of a 6-well plate. One day after transfection, cells were detached with cell dissociation buffer (Life Technologies), re-suspended in DMEM + 10% FCS and 2% of cells were seeded in 96-well plate as previously described[77] and stimulated overnight with 50 pM human IL-7 (Immunotools GmbH) and increasing concentrations of IL-7Rα minibinder. STAT5-dependent luciferase activity was measured on the next day using a GloMax 96 Microplate Luminometer. The -fold induction of luciferase activity was calculated by the ratio of the luminescence signal from cells stimulated with IL-7 to the signal from the unstimulated cells. The data were plotted and fitted to a log inhibitor versus response curve in GraphPad Prism. The pcDNA3-gamma common was a kind gift from Dr J.C. Renauld (Faculty of Medicine and Dentistry, UC Louvain, Belgium) and the pMX-IRES-GFP-hJak3 vector[86] was kindly provided by Dr S.N. Constantinescu (Ludwig Institute for Cancer Research, Belgium). The pMET7-HA-IL-7Rα, empty pMET7 and pGL3-β-casein-luci vectors were kindly provided by Dr. F. Peelman (UGent, Belgium).

Apparent SC_{50} estimation from FACS and NGS

The Pear program[87] was used to assemble the fastq files from the Next Generation Sequencing runs. Translated, assembled reads were matched against the ordered designs to determine the number of counts for each design in each pool.

The critical assumption to the fitting here is to pretend that the yeast cells displaying a particular design will follow a modified version of the standard KD equation relating fraction bound to concentration:

$$fraction\_collected_i = \frac{concentration}{(concentration + SC_{50,i})} \text{EQ-KD}$$

where $fraction\_collected_i$ is the fraction of the yeast cells displaying design i that were collected, concentration is the target concentration for sorting, and $SC_{50,i}$ is the apparent SC_{50} of the design (the concentration where 50% of the cells would be collected).

The next assumption is that all designs have the same expression level on yeast surface and that 100% of yeast cells express well enough to be collected in the "expression" gate (i.e. the right population in Supplementary Figure 7).

These two assumptions, while probably false, allow fitting the data with only one free parameter per design and no global free parameters. The correct version of EQ-KD for this experiment likely has a different shape and slope from a perfect sigmoid, the net effect of correcting this would be that all SC_{50} are scaled by a constant factor (which would not affect the relative comparisons made here). It can be shown by analyzing the data that different designs result in different expression levels on yeast (one can examine the $fraction\_collected_i$ for strong binders at concentrations where binding should be saturated). The net result is that experimentally, EQ-KD is multiplied by a constant between 0 and 1 for each design. This constant seems to range from 0.2 to 0.7. As such, when fitting the data, $fraction\_collected_i$ values above 0.2 are considered saturating. However, because the 0.2 mark may represent 90% collection for poorly-expressing designs and 30% collection for strongly-expressing designs, the resulting SC_{50} fits may vary by up to five-fold. The alternative is to try to estimate an expression level; however, this becomes increasingly difficult with weaker binders that never saturate the experiment.

Apparent SC_{50} estimation from FACS and NGS: Point estimates

The following equation may be used to determine the $fraction\_collected_i$ for a single design in a single sort:

$$fraction\_collected_i = \frac{proportion\_child\_pool_i}{proportion\_parent\_pool_i} \times facs\_collection\_fraction \text{EQ-FRAC}$$

where $fraction\_collected_i$ is the proportion of cells carrying design i that were collected during the sort, $proportion\_child\_pool_i$ is the proportion of the total NGS counts for design i from the pool that was collected, $proportion\_parent\_pool_i$ is the proportion of the total NGS counts for design i from the pool that was the input for the sorter, and facs_collection_fraction was the fraction of the yeast cells collected during the specific sort (a number extracted from the FACS machine itself).

This point-estimate method is best suited for asking the question: which designs have $SC_{50} < SC_{50,0}$? by determining the expected $fraction\_collected_i$ for a given sorting concentration and $SC_{50,0}$. The sorting concentration and $SC_{50,0}$ should be selected such that EQ-KD results in an expected $fraction\_collected_i$ less than 0.2 in order to circumvent the expression issues mentioned above. Then, any designs with $fraction\_collected_i$ greater than the cutoff may say their SC_{50} is less than $SC_{50,0}$. Designs with low numbers of counts are suspect, see the Doubly-Tranformed Yeast Cells section. For this analysis, any designs with fewer than max_possible_passenger_cells cells were eliminated.

This method may be applied to avidity sorts, however, the resulting SC_{50} would be the SC_{50} during avidity experiments. It is unclear to the authors what the precise mathematical effect of avidity is and as such we do not compare avidity SC_{50}s with non-avidity SC_{50}s.

Apparent SC_{50} estimation from FACS and NGS: Doubly-transformed yeast cells

Doubly-transformed yeast cells represent a major source of error in these experiments. While rare, a yeast cell that contains two plasmids, one of a strong binder and one of a non-binder, will carry the non-binder plasmid through the sorting process. The net result is that the non-binder will end up with counts that track the strong binder, however, at a greatly reduced absolute number. (Rare is a relative term here. While the odds of any two specific plasmids being in one cell is low, in the entire pool of yeast, doubly transformed cells seem to be quite common.)

We chose to address this issue by making the following assumption: non-binders that take advantage of a doubly-transformed yeast cell do so from precisely one double-transformation event. In other words, we assume that the same non-binding plasmid did not get doubly transformed into two separate strong-binding yeast. This assumption allows us to estimate the largest number of cells we would expect to see from a doubly-transformed plasmid:

$$max\_possible\_passenger\_cells = \frac{cells\_collected_{i\_max}}{cells\_sorted\_R1_{i\_max}} \times cell\_copies\_before\_first\_sort \text{EQ-MAX}$$

where max_possible_passenger_cells is the highest number of cells that we would expect a non-binding plasmid to occupy, cells_collected$_{i_max}$ is the number of cells collected in this round for the design with the most number of cells collected, cells_sorted_R1$_{i_max}$ is the number of cells sorted for design i_max (the same design from cells_collected$_{i_max}$), and cell_copies_before_first_sort is the number of copies of each cell that occurred before the first sort ($2^{\#cell\_divisions}$). The number of cells_collected$_i$ may be approximated by multiplying the number of cells the FACS machine collected by the proportion of the pool that design i represents. The number of cells_sorted$_i$ may be estimated by either dividing the cells_collected$_i$ by the facs_collection_fraction or by multiplying the number of cells fed to the FACS machine by the proportion of design i in that pool.

With this number in hand, one can set a floor for the number of cells that one would expect to see. Any design with fewer than this number of cells cannot be considered for calculations because it is unclear whether or not that cell is part of a doubly-transformed yeast cell. On the whole, this method reduces false-positive binders, but also removes true-positive binders that did not transform well. It is wise to simply drop designs from the downstream calculations that did not transform well.

Apparent SC_{50} estimation from FACS and NGS: Full estimate

Estimation of an upper and lower bound on the SC_{50} from the data may be performed by looking at an arbitrary number of sorting experiments. Taking a $P(SC_{50} == SC_{50,0} \mid data)$ and performing Bayesian analysis, one arrives at a confidence interval for the actual SC_{50} value. This analysis may be performed at every sort and the resulting distributions combined to produce a robust estimate.

Each sort may be modeled as a binomial distribution where: p=fraction_collected from EQ-KD using concentration=sorting_concentration and SC_{50}=$SC_{50,0}$; n=cells_sorted$_i$; and x=cells_collected$_i$. By performing this analysis at a range of $SC_{50,0}$ values and examining the probability this could happen by the binomial distribution, one arrives at $P(SC_{50} == SC_{50,0} \mid data)$. Specifically for this analysis, the cdf of the binomial was used with the null hypothesis that $SC_{50} == SC_{50,0}$.

Care should be taken for the valid range of p. As stated previously, it is wise to cap the expected value of p to 0.2 to account for expression levels and to floor the value such that n*p does not fall below max_possible_passenger_cells. In our implementation, if x falls into a range that has been clipped, a probability of 1 is returned.

Code to perform this entire analysis is available in the Supplementary Information.

SSM validation: Relax protocol

In order to remove artifacts from design and to discover the best orientation for each SSM mutation, all binders were relaxed using the Rosetta beta_nov16 score function before calculations began (30 replicates using 5 repeats of cartesian FastRelax taking the best scoring model). Relaxation of point mutants then used the standard cartesian FastRelax procedure and allowed all residues within 10 Å of the mutation to relax. The backbone coordinates of those residues on the binder were allowed to relax while the target was held constant. The best of three (as evaluated by Rosetta energy) was chosen as the representative model. An xml is provided in the Supporting Information to perform this relaxation.

SSM validation: Entropy score

In order to validate that the designed binder was folded into the correct shape and was using its designed interface to bind to the target, the entropy of the interface, monomer core, and monomer surface were examined. For each position on the binder, the sequence entropy (Shannon entropy) of each position was calculated using the observed frequencies of each amino acid in the Next Generation Sequencing data. The specific pool that was chosen for this analysis was the pool with concentration closest to 10-fold lower than the calculated SC_{50} of the parent.

After the per-position sequence entropy was calculated, the average per-position entropy of the SASA-hidden positions contacting the target (interface core), the SASA-hidden positions not contacting the target (monomer core), and the fully exposed positions not contacting the target (monomer surface) were calculated. A simple subtraction was performed according to EQ-ENTROPY:

$$intermediate\_entropy\_score = S_{monomer\_core} + S_{interface\_core} - S_{monomer\_surface} \text{EQ-ENTROPY}$$

where S_{region} is the average entropy of that region.

Finally, the probability that the score could have come from totally random data was computed by performing the above calculation on the actual data, and then performing the same calculation 100 times, but randomly mismatching the observed counts among all SSM point mutations. In this way, the experimental noise is kept constant among the 100 decoy datasets. The final step to arrive at a p-value was to calculate the mean and standard deviation of the 100 decoy intermediate_entropy_scores and to find the p-value with the Normal CDF function of the binder's intermediate_entropy_score.

SSM validation: Rosetta accuracy score

In order to further assess the accuracy of the design model, the correlation between the predicted effect on binding by Rosetta was compared with the experimental data. The effect from Rosetta can be broken into two components: monomer stabilization/destabilization and interface stabilization/destabilization. The effect on the monomer energy will affect the fraction of the proteins that are folded in solution. This fraction of folded proteins will then worsen the affinity because only the folded proteins are able to bind. The effect on the monomer stability was estimated by taking the difference in Rosetta energy between the native relaxed dock and the mutant relaxed dock and looking only at the change in Rosetta score of the docked protein (excluding energies arising from cross-interface edges). The effect on the target energy was calculated the same was and was considered to directly affect the binding energy. The binding energy was calculated by taking the difference in Rosetta score between the docked and undocked conformations (but with no repacking or minimization in the unbound form). An xml exists in the Supplementary Information to perform this calculation.

The effect on the P(fold_monomer) was estimated by first determining the predicted ΔG_{fold} of the native protein.

$$P(fold\_monomer) = exp(\frac{\Delta G_{fold} + \Delta G_{mutant\_effect}}{kT}) \text{EQ-PFOLD}$$

$$\Delta ddG_{monomer\_effect} = kTln(\frac{P(fold\_monomer)_{native}}{P(fold\_monomer)_{mutant}}) \text{EQ-ddg\_monomer}$$

Where k is the Boltzmann constant and T is temperature which was set to 300 K for this calculation.

Using EQ-ddg_monomer and EQ-PFOLD, the predicted ΔG_{fold} for the native design was estimated by performing a least-squares fit of all mutations that did not occur in residues at the interface. A rudimentary confidence interval was created by allowing all ΔG_{fold} values that resulted in a root mean squared error of within 0.25kcal/mol of the best ΔG_{fold} value. Typical confidence intervals spanned 3 kcal/mol.

$$\Delta ddG_{Rosetta} = \Delta ddG_{monomer\_effect} + \Delta ddG_{interface\_effect} + \Delta ddG_{target\_effect} \text{ EQ-DDG\_SUM}$$

With the ΔG_{fold} in hand, the predicted effect on the binding energy could be computed according to EQ-DDG_SUM. The values of ΔG_{fold} inside the confidence range for ΔG_{fold} that produced the largest and smallest $\Delta ddG_{Rosetta}$ were used to produce a confidence interval for $\Delta ddG_{Rosetta}$.

The per-position accuracy was assessed by determining whether the confidence interval for $\Delta ddG_{Rosetta}$ was compatible with the confidence interval for the SC_{50} from the experimental data. A buffer of 1kcal/mol was allowed.

With the per-position accuracies in hand, the overall percentage of mutations that Rosetta was able to explain in the monomer_core and interface_core was assessed. This produced an overall Rosetta accuracy score.

In the same way as the Entropy score, 100 decoys with randomly shuffled SC_{50} values were subjected to the same procedure. The mean and standard deviation of the decoys was determined and the p-value for the Rosetta score was determined using the Normal CDF function.

Reporting summary

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

Data availability

The atomic coordinates and experimental data of H3_mb in complex with H3 HA, TrkA_mb in complex with TrkA, unbound FGFR2_mb, FGFR2_mb in complex with FGFR4, unbound IL-7Rα_mb, IL-7Rα_mb in complex with IL-7Rα and VirB8_mb in complex with VirB8 have been deposited in the RCSB Protein Database with the accession numbers of 7RDH, 7N3T, 7N1K, 7N1J, 7S5B, 7OPB and 7SH3 respectively. Diffraction images for the TrkA minibinder complex have been deposited in the SBGrid Data Bank with ID 838. The Rosetta macromolecular modeling suite (https://www.rosettacommons.org) is freely available to academic and non-commercial users. Commercial licenses for the suite are available via the University of Washington Technology Transfer Office.

Code availability

The Rosetta macromolecular modelling suite (https://www.rosettacommons.org) is freely available to academic and non-commercial users. Commercial licenses for the suite are available via the University of Washington Technology Transfer Office. The design scripts and main pdb models, computational protocol for data analysis, experimental data and analysis scripts, the whole miniprotein scaffold library, all the design models and next-sequencing results used in this paper can be downloaded from file servers hosted by Institute for Protein Design: http://files.ipd.uw.edu/pub/robust_de_novo_design_minibinders_2021/supplemental_files/scripts_and_main_pdbs.tar.gz, http://files.ipd.uw.edu/pub/robust_de_novo_design_minibinders_2021/supplemental_files/computational_protocol_analysis.tar.gz, http://files.ipd.uw.edu/pub/robust_de_novo_design_minibinders_2021/supplemental_files/experimental_data_and_analysis.tar.gz, http://files.ipd.uw.edu/pub/robust_de_novo_design_minibinders_2021/supplemental_file s/scaffolds.tar.gz, http://files.ipd.uw.edu/pub/robust_de_novo_design_minibinders_2021/supplemental_files/design_models_pdb.tar.gz and http://files.ipd.uw.edu/pub/robust_de_novo_design_minibinders_2021/supplemental_files/design_models_silent.tar.gz. All the files are stored in compressed gzip format. Download and decompress the files, there is a detailed description of the binder design pipeline and the whole process can be reproduced based on those files. The source code for RIF docking implementation is freely available at https://github.com/rifdock/rifdock.

42. Lim, Y. et al. GC1118, an Anti-EGFR Antibody with a Distinct Binding Epitope and Superior Inhibitory Activity against High-Affinity EGFR Ligands. *Mol. Cancer Therapeutics* **15**, 251–263 (2016).
43. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucl. Acids Res.* **33**, 2302–2309 (2005).
44. Silva, D. A., Correia, B. E. & Procko, E. Motif-Driven Design of Protein-Protein Interfaces. *Methods Mol. Biol.* **1414**, 285–304 (2016).
45. Leaver-Fay, A. et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* **487**, 545–574 (2011).
46. Hoover, D. M. & Lubkowski, J. DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucl. Acids Res.* **30**, e43 (2002).
47. Benatuil, L., Perez, J. M., Belk, J. & Hsieh, C. M. An improved yeast transformation method for the generation of very large human antibody libraries. *Protein engineering, design & selection: PEDS* **23**, 155–159 (2010).
48. Stevens, J. et al. Structure of the uncleaved human H1 hemagglutinin from the extinct 1918 influenza virus. *Science* **303**, 1866–1870 (2004).
49. Divine, R. et al. Designed proteins assemble antibodies into modular nanocages. *Science* **372**, eabd9994 (2021).
50. Xu, Y. et al. Addressing polyspecificity of antibodies selected from an in vitro yeast presentation system: a FACS-based, high-throughput selection and analytical tool. *Prot. Eng. Des. Sel.* **26**, 663–670 (2013).
51. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
52. McCoy, A. J. et al. Phaser crystallographic software. *J. Applied Crystallog.* **40**, 658–674 (2007).
53. McCoy, A. J., Grosse-Kunstleve, R. W., Storoni, L. C. & Read, R. J. Likelihood-enhanced fast translation functions. *Acta Crystallographica. Section D, Biological crystallography* **61**, 458–464 (2005).
54. Adams, P. D. et al. PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallographica. Section D, Biological crystallography* **58**, 1948–1954 (2002).
55. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallographica. Section D, Biological crystallography* **60**, 2126–2132 (2004).
56. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica. Section D, Biological crystallography* **66**, 12–21 (2010).
57. Wehrman, T. et al. Structural and mechanistic insights into nerve growth factor interactions with the TrkA and p75 receptors. *Neuron* **53**, 25–38 (2007).
58. Kabsch, W. Xds. *Acta Crystallographica. Section D, Biological crystallography* **66**, 125–132 (2010).
59. P.,L. XDSME: XDS Made Easier. GitHub repository, https://github.com/legrandp/xdsme. https://doi.org/10.5281/zenodo.837885 (2017)
60. Evans, P. R. An introduction to data reduction: space-group determination, scaling and intensity statistics. *Acta Crystallographica. Section D, Biological crystallography* **67**, 282–292 (2011).
61. Evans, P. R. & Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallographica. Section D, Biological crystallography* **69**, 1204–1214 (2013).
62. Winn, M. D. et al. Overview of the CCP4 suite and current developments. *Acta Crystallographica. Section D, Biological crystallography* **67**, 235–242 (2011).
63. Terwilliger, T. C. et al. Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *Acta Crystallographica. Section D, Biological crystallography* **64**, 61–69 (2008).
64. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallographica. Section D, Biological crystallography* **66**, 486–501 (2010).
65. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica. Section D, Biological crystallography* **66**, 213–221 (2010).
66. Echols, N. et al. Graphical tools for macromolecular crystallography in PHENIX. *J. Applied Crystallog.* **45**, 581–586 (2012).
67. Afonine, P. V. et al. Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallographica. Section D, Biological crystallography* **68**, 352–367 (2012).
68. Painter, J. & Merritt, E. A. Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. *Acta Crystallographica. Section D, Biological crystallography* **62**, 439–450 (2006).
69. Headd, J. J. et al. Flexible torsion-angle noncrystallographic symmetry restraints for improved macromolecular structure refinement. *Acta Crystallographica. Section D, Biological crystallography* **70**, 1346–1356 (2014).
70. Karplus, P. A. & Diederichs, K. Linking crystallographic model and data quality. *Science* **336**, 1030–1033 (2012).
71. Morin, A. et al. Collaboration gets the most out of software. *eLife* **2**, e01456 (2013).
72. Plotnikov, A. N., Schlessinger, J., Hubbard, S. R. & Mohammadi, M. Structural basis for FGF receptor dimerization and activation. *Cell* **98**, 641–650 (1999).
73. Schlessinger, J. et al. Crystal structure of a ternary FGF-FGFR-heparin complex reveals a dual role for heparin in FGFR binding and dimerization. *Mol. Cell* **6**, 743–750 (2000).
74. Kabsch, W. Integration, scaling, space-group assignment and post-refinement. *Acta Crystallographica. Section D, Biological crystallography* **66**, 133–144 (2010).
75. Liebschner, D. et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallographica. Section D, Structural biology* **75**, 861–877 (2019).
76. Williams, C. J. et al. MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci.* **27**, 293–315 (2018).
77. Verstraete, K. et al. Structure and antagonism of the receptor complex mediated by human TSLP in allergy and asthma. *Nat. Commun.* **8**, 14937 (2017).
78. BUSTER version 2.10.2 Cambridge, United Kingdom:Global Phasing Ltd. (Cambridge, United Kingdom:Global Phasing Ltd., Cambridge, United Kingdom:Global Phasing Ltd., 2016).

79. Joosten, R. P., Long, F., Murshudov, G. N. & Perrakis, A. The PDB_REDO server for macromolecular structure model optimization. *IUCrJ* **1**, 213–220 (2014).
80. Evans, P. Scaling and assessment of data quality. *Acta Crystallographica. Section D, Biological crystallography* **62**, 72–82 (2006).
81. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
82. Berman, H. M. et al. The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242 (2000).
83. Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.* **10**, 980 (2003).
84. Grabowski, M. et al. A public database of macromolecular diffraction experiments. *Acta Crystallographica. Section D, Structural biology* **72**, 1181–1193 (2016).
85. Grabowski, M. et al. The Integrated Resource for Reproducibility in Macromolecular Crystallography: Experiences of the first four years. *Struct. Dyn.* **6**, 064301 (2019).
86. Hornakova, T. et al. Acute lymphoblastic leukemia-associated JAK1 mutants activate the Janus kinase/STAT pathway via interleukin-9 receptor α homodimers. *J. Biol. Chem.* **284**, 6773–6781 (2009).
87. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2014).
88. Buchan, D. W. A. & Jones, D. T. The PSIPRED Protein Analysis Workbench: 20 years on. *Nucl. Acids Res.* **47**, W402–W407 (2019).
89. Lauer, T. M. et al. Developability index: a rapid in silico tool for the screening of antibody aggregation propensity. *J. Pharma. Sci.* **101**, 102–115 (2012).

Acknowledgements This work was supported by DARPA Synergistic Discovery and Design (SD2) HR0011835403 contract FA8750-17-C-0219 (L.C., B.C., S.H., D.B.), The Audacious Project at the Institute for Protein Design (L.K.), the Open Philanthropy Project Improving Protein Design Fund (B.C., D.B.), funding from Eric and Wendy Schmidt by recommendation of the Schmidt Futures program (I.G., L.M.), an Azure computing resource gift for COVID-19 research provided by Microsoft (L.C., B.C.), the National Institute of Allergy and Infectious Diseases (HHSN272201700059C, D.B., B.H., L.S.; NIH R01 AI140245 to E.M.S.; NIH R01 AI150855 to I.A.W.), the National Institute on Aging (R01AG063845, B.H., D.B.), the Defense Threat Reduction Agency (HDTRA1-16-C-0029, D.B., E-M.S.), The Donald and Jo Anne Petersen Endowment for Accelerating Advancements in Alzheimer's Disease Research (N.B.), a gift from Gates Ventures (M.D.), The Human Frontier Science Program (A.Y.) and The Howard Hughes Medical Research Institute (K.M.J, K.C.G., D.B.). Use of SSRL at Stanford Linear Accelerator Center (SLAC) National Accelerator Laboratory is supported by the US Department of Energy Office of Science, Office of Basic Energy Sciences under contract DE-AC02-76SF00515. The SSRL Structural Molecular Biology Program is supported by the Department of Energy, Office of Biological and Environmental Research and the National Institutes of Health, National Institute of General Medical Sciences (including P41GM103393). A part of this work is based upon research conducted at the Northeastern Collaborative Access Team beamlines, which are funded by the National Institute of General Medical Sciences from the National Institutes of Health (P30 GM124165). The Eiger 16M detector on the 24-ID-E beam line is funded by a NIH-ORIP HEI grant (S10OD021527). STRW was supported by the CCR intramural research program of NCI-NIH. GM/CA at the Advanced Photon Source at Argonne National Laboratory has been funded by the National Cancer Institute (ACB-12002) and the National Institute of General Medical Sciences (AGM-12006, P30GM138396). This research used resources of the Advanced Photon Source, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under Contract No. DE-AC02-06CH11357. The Eiger 16M detector at GM/CA-XSD was funded by NIH grant S10 OD012289. We thank the staff of beamline ID23-2 (ESRF) for technical support and beamtime allocation. S.N.S. acknowledges research support from Research Foundation Flanders (grants G0C2214N and G0E1516N), and the Hercules Foundation (no. AUGE- 11-029). S.N.S. is a principal investigator of the VIB (Belgium). SSGCID is funded by federal funds from the National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health (NIH), Department of Health and Human Services, under contract no. HHSN272201700059C from September 1, 2017. APS/LSCAT research used resources of the Advanced Photon Source, a U.S. Department of Energy (DOE), Office of Science user facility operated for the DOE Office of Science by Argonne National Laboratory under contract no. DE-AC02- 06CH11357. Use of LS-CAT Sector 21 was supported by the Michigan Economic Development Corporation and the Michigan Technology Tri-Corridor (grant 085P1000817). We thank G. Ueda for kindly providing the Ang1 protein for the TrkA competition assay and D.H. Fuller for kindly providing the FI6v3 antibody for the HA competition assay. We thank Y.-J. Park, A. Walls, and D. Veesler for their collaborative research and cryoEM structure determination for minibinders targeting SARS-CoV-2 spike. We would also like to thank K. Van Wormer and A. Curtis Smith for their tremendous laboratory support during COVID-19.

Author contributions L.C., B.C. and D.B. designed the research; L.C. and B.C. contributed equally; L.C. and B.C. developed the method; L.C., B.C. and E-M.S. designed the scaffold library; L.C., B.C., B.H. and N.B. designed the binders; L.C., B.C., I.G., B.H., N.B., L.K., M.D. L.M. S.H. and W.Y. performed the yeast screening, expression and binding experiments; R.U.K., S.B. and I.A.W. prepared the H3 protein and solved the H3_mb complex structure; L.P., K.M.J. and A.Y. prepared the target protein, solved the complex structure and performed the competition assay for TrkA; J.S.P., J.S. and S.L. solved the FGFR2_mb structures; A.P. performed the competition assay for FGFR2 and EGFR; I.M., K.H.G.V., K.V. and S.N.S. performed the IL7Rα competition assay and solved the complex structure of IL7Rα_mb; S.T.R.W. solved the structure of the unbound IL7Rα_mb; B.H., N.D.D., A.P. and A.B. prepared the VirB8 target protein and solved the complex structure. All authors analysed data. L.S., I.A.W., H.R-B., J.S., S.L., S.N.S., K.C.G. and D.B. supervised research. L.C., B.C. and D.B. wrote the manuscript with the input from the other authors. All authors revised the manuscript.

Competing interests L. C., B.C., I.G., B.H., N.B., E-M.S., L.S. and D.B. are co-inventors on a provisional patent application (21-0753-US-PRO) that incorporates discoveries described in this manuscript.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-022-04654-9.

Correspondence and requests for materials should be addressed to David Baker.

Peer review information *Nature* thanks the anonymous reviewers for their contribution to the peer review of this work.