Efficient Design of Protein-based Binders

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Kateryna Maksymenko

aus Kyjiw, Ukraine

Tübingen

2023

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	15.02.2024
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Andrei N. Lupas
2. Berichterstatter:	Prof. Dr. Ralf-Peter Jansen
3. Berichterstatterin:	Prof. Dr. Birte Höcker

Acknowledgements

I would like to start by expressing my deepest appreciation to my supervisors, who guided me through the different stages of my doctoral journey. Patrick Müller gave me the opportunity to start my PhD, mentored me during my time at the Friedrich Miescher Laboratory and taught me the importance of rigorous and systematic scientific attitude. Andrei Lupas advised me after my transition to Department 1 at the Max Planck Institute. He provided me with his exceptional expertise in protein biology, granted me research freedom, always offered his time when I needed it, and was an example of genuine excitement for the world of sequences and structures. Julia Skokowa was a tremendous help in the functional evaluation of my designs. She generously offered her excellent ideas during our weekly meetings and greatly expanded my view of the therapeutic relevance of proteins. I would also like to thank her and Karl Welte for involving me in the joyful scientific and social events with their research group.

All of the work presented in this dissertation would not have happened without Mohammad ElGamacy, who spearheaded both HECTOR and Damietta projects. He introduced me to the main concepts of protein design and was extremely supportive of my computational and experimental efforts throughout.

I sincerely thank my thesis advisory committee members, Ralf-Peter Jansen and Karl Forchhammer, for their detailed scientific input and useful recommendations on navigating my doctoral studies.

I am also very grateful to my friends, colleagues and collaborators for their fruitful discussions, research contributions, and an amazing working environment. Particularly, I would like to acknowledge Birte Hernandez Alvarez, Murray Coles, Marcus Hartmann, Reinhard Albrecht, Vincent Truffault, John Weir, Narges Aghaallaei, Natalia Pashkovskaia, Carolin Schnedermann, Yimin Hu, and Silvie Deiss. Special thanks go to the Design crew - Alberto Piovesan, Timo Ullrich, Matej Milijas-Jotic, and Valeriia Hatskovska. I want to acknowledge the IMPRS program for funding my work and the Researcher Support Team for making all of the organizational steps happen.

Finally, many thanks to my family, especially my parents, my sister, and my husband, for all their love and faith in me.

Zusammenfassung

Protein-Protein-Interaktionen bilden die Grundlage verschiedenster Prozesse der Homöostase und der Entstehung von Krankheiten. Folglich können Proteinbinder, die diese Interaktionen fördern oder ihnen entgegenwirken, als potente Werkzeuge sowohl in der Forschung als auch zu therapeutischen Zwecken dienen. Obwohl sich die Methoden des Proteindesigns rasch weiterentwickeln, bleibt das Design von Epitop-gerichteten Bindern, die allein auf der Zielstruktur basieren, eine enorme Herausforderung. Bestehende Methoden des *De-novo*-Designs liefern hochaffine Binder bislang nur durch das experimentelle Screening großer Bibliotheken der designten Kandidaten. Die niedrigen Erfolgsquoten lassen sich auf die Komplexität der simultanen Suche nach dem optimalen Bindungsgerüst, der Positionierung und der Sequenz zurückführen. Zudem ist es eine Herausforderung, die zahlreichen Faktoren, die zu einem Bindungsereignis beitragen, präzise einzuschätzen, wodurch die Bewertung der Designs zusätzlich erschwert wird. Ziel dieser Arbeit ist die Erforschung neuer Designstrategien zur ressourceneffizienten Entwicklung von bedarfsorientierten Proteinbindern.

Zuerst bewerte ich den Nutzen eines mehrstufigen Ansatzes, der das Docking vom Interfacedesign trennt, um die Komplexität des Prozesses zu reduzieren. Beim Docking wird eine neuartige Fingerprinting-Methode verwendet, die eine rapide Abschätzung der Komplementarität von Moleküloberflächen ermöglicht und geeignete Binderstrukturen aus einer Proteinstrukturdatenbank abruft. Zum Nachweis des Konzepts wende ich diese Strategie an, um Binder zu entwerfen, die den vaskulären endothelialen Wachstumsfaktor (VEGF) binden. VEGF ist ein wichtiges angiogenetisches Molekül, das in der Pathogenese verschiedener Krebsarten eine bedeutende Rolle spielt. Ich charakterisiere experimentell eine kleine Auswahl von Designkandidaten und demonstriere, dass zwei von ihnen eine nanomolare Affinität zu VEGF haben. Darüber hinaus hemmen sie die Proliferation und das Überleben von VEGF-abhängigen Zellen und weisen *in vivo* eine VEGF-supprimierende Wirkung auf.

Des Weiteren untersuche ich die Umsetzbarkeit der Tensorisierung von Energieberechnungen für das Proteindesign. Durch die direkte Projektion atomarer Interaktionsfelder in dreidimensionale Tensoren werden Energieauswertungen zu einer einzigen Matrixoperation zusammengefasst, was den Rechenaufwand erheblich

vermindert. Durch retrospektive Validierung zeige ich, dass das tensorisierte Verfahren andere Designmethoden in Bezug auf Geschwindigkeit und Genauigkeit übertrifft. Zur prospektiven Validierung setze ich dieses Verfahren ein, um multispezifische Bindeproteine für Liganden des epidermalen Wachstumsfaktor-Rezeptors (EGFR) zu entwickeln. Die getesteten Designs zeigen eine starke Bindung an ihre Zielmoleküle und hemmen die EGFR-Aktivität *in vitro* und *in vivo*.

Diese Arbeit bietet innovative Lösungen für Problemstellungen im Bereich des Proteindockings und -designs. Innerhalb eines einzigen *In-silico*-Durchgangs können diese Lösungsansätze genutzt werden, um in kürzester Zeit Binder gegen verschiedenste Zielproteine zu entwickeln.

Abstract

Protein-protein interactions form the basis of diverse processes in homeostasis and disease. Consequently, protein binders that promote or antagonize these interactions can serve as potent tools for both research and therapeutic purposes. While protein design methods are rapidly advancing, the design of epitope-directed binders relying on the target structure alone remains a formidable challenge. Existing *de novo* design approaches yield high-affinity binders only through the experimental screening of large libraries of designed candidates. The low success rates can be attributed to the dimensionality of the simultaneous search for an optimal binder scaffold, pose, and sequence. Moreover, the limitations in accurately estimating numerous factors contributing to a binding event further complicate the scoring process. This work aims to explore new design strategies to create on-demand protein binders in a resource-efficient manner.

First, I evaluate the utility of a tiered approach that separates the docking task from interface design to reduce complexity of the problem. The docking step uses a novel surface fingerprinting method, which enables ultra-fast estimation of surface complementarity and retrieves viable binder scaffolds from a protein structure database. As proof-of-concept, I adopt this strategy to design binders targeting the vascular endothelial growth factor (VEGF), a key angiogenic molecule implicated in pathogenesis of various cancers. I experimentally characterize a small number of design candidates and show that two of them have nanomolar affinity to VEGF, inhibit proliferation and survival of VEGF-dependent cells, and finally have a VEGF-suppressing effect *in vivo*.

Second, I investigate the feasibility of tensorizing energy calculations for protein design. The direct projection of atomic interaction fields in three-dimensional tensors condenses energy evaluations into a single matrix operation, greatly simplifying the computational load. Through retrospective validation, I demonstrate that the tensorized framework outperforms other design engines in terms of speed and accuracy. For prospective validation, I deploy this framework to design multi-specific binders against ligands of the epidermal growth factor receptor (EGFR). The tested designs bind strongly to their targets and inhibit EGFR activity *in vitro* and *in vivo*.

This work offers innovative solutions to protein docking and design problems. Integrated into the design framework, these solutions can be used to rapidly create protein binders against diverse targets through a single *in silico* round.

List of publications

Publications included in this dissertation:

Maksymenko K., Maurer A., Aghaallaei N., Barry C., Borbaran-Bravo N., Ullrich T., Dijkstra T.M.H., Hernandez Alvarez B., Müller P., Lupas A.N., Skokowa J., ElGamacy M. (2023). The design of functional proteins using tensorized energy calculations. Cell Rep Methods. 3(8): https://doi.org/10.1016/j.crmeth.2023.100560

Maksymenko K., Coles M., Aghaallaei N., Pashkovskaia N., Volz M., Pereira J., Hartmann M.D., Tabatabai G., Liebau S., Müller P., Lupas A.N., Skokowa J., ElGamacy M. (2023). A complementarity-based approach to *de novo* binder design generates VEGF inhibitors (in revision).

Publications not included in this dissertation:

Hernandez Alvarez B., Skokowa J., Coles M., Mir P., Nasri M., **Maksymenko K.**, Weidmann L., Rogers K.W., Welte K., Lupas A.N., Müller P., ElGamacy M. (2020). Design of novel granulopoietic proteins by topological rescaffolding. PLoS Biol. 18(12): https://doi.org/10.1371/journal.pbio.3000919

Skokowa J., Hernandez Alvarez B., Coles M., Ritter M., Nasri M., Haaf J., Aghaallaei N., Xu Y., Mir P., Krahl A.-C., Rogers K.W., **Maksymenko K.**, Bajoghli B., Welte K., Lupas A.N., Müller P., ElGamacy M. (2022). A topological refactoring design strategy yields highly stable granulopoietic proteins. Nat Commun. 13(2948): https://doi.org/10.1038/s41467-022-30157-2

Ullrich T., Pollmann C., Ritter M., Haaf J., Aghaallaei N., Hatskovska V., Tesakov I., **Maksymenko K.**, El-Riz M., Kandabarau S., Klimiankou M., Lengerke C., Welte K., Hernandez-Alvarez B., Müller P., Lupas A.N., Piehler J., Skokowa J., ElGamacy M. (2023). Tuning of granulopoietic signaling by *de novo* designed agonists (in preparation).

List of abbreviations

3D	3-dimensional
AF2	AF2
CTLA-4	Cytotoxic T-lymphocyte-associated protein 4
DARPin	Designed ankyrin repeat proteins
DEE	Dead-end elimination
DL	Deep learning
EGF	Epidermal growth factor
EGFR	Epidermal growth factor receptor
EpCAM	Epithelial cell adhesion molecule
ERK	Extracellular signal-regulated kinase
Fab	Fragment antigen binding
GPU	Graphics processing unit
HER2	Human epidermal growth factor receptor 2
LJ	Lennard-Jones
MD	Molecular dynamics
PD-1	Programmed cell death protein 1
PDB	Protein Data Bank
PD-L1	Programmed death-ligand 1
pLDDT	Predicted value of the local distance difference test
POI	Protein of interest
PPI	Protein-protein interaction
Prdm14	PR-domain containing protein 14
RIF	Rotamer interaction field
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
SHP2	Src homology 2 domain-containing tyrosine phosphatase
TPU	Tensor processing units
VEGF	Vascular endothelial growth factor

Table of contents

Chapter 1

Introduction

Over the years, our understanding of protein folding and functioning has significantly deepened. This theoretical progress, along with technological advancements, has enabled the design of proteins from first principles. While computational methods show considerable success in generating well-folded *de novo* structures, the *in silico* design of new protein functions remains challenging. Particularly, design of protein binders capable of targeting pre-defined molecular sites with high affinity and specificity stands as a tempting yet unrealized endeavor.

Currently, obtaining protein-based binders for medical and biotechnological needs involves laborious processes of animal immunization or large-scale screening of random protein libraries. Computational protein design could potentially provide a more rational and efficient means to create affinity reagents with tailored features. However, recent *de novo* design attempts revealed that only a tiny fraction of the designed candidates (<1%) exhibit the intended properties in terms of expressibility, stability, and binding [1, 2].

To better comprehend the reasons behind this low success rate, the problem of binder design can be formulated as a combined high-throughput docking and sequence design problem. The goal of the docking step is to identify a protein backbone that, in its specific orientation, shows high shape complementarity to the target epitope and can serve as a viable template for future binders. This task is computationally demanding, since it requires sampling an immense space of available scaffolds, considering their potential rotations and translations. The sequence design problem, on the other hand, relates to sampling and scoring all possible combinations of amino acid identities and side chain conformations to find a sequence that would stabilize a desired structure and guarantee its binding to the target protein. The primary difficulties here are insufficient sampling and the lack of accurate scoring functions to jointly minimize the free energies of bound and solvated states of the binder. Thus, throughout my research, I aim to explore alternative solutions for the docking and sequence design problems.

To address the first issue, I test a theoretical framework for performing massive-scale molecular docking in a synthetic geometry that allows to estimate complementarity between unaligned surface patches from different proteins. This can be advantageous for the rapid selection of promising binder scaffolds from large structure datasets. I apply this framework to design *de novo* binders quenching the receptor-binding site of the vascular endothelial growth factor (VEGF), an important therapeutic target in oncology, cardiology, and ophthalmology.

On the design problem front, my objective is to validate a new scoring framework that tensorizes all of the non-bonded energy terms and promises enhanced design throughput and accuracy. This approach avoids iterative loops for calculating atom-atom interactions, but rather evaluates the interaction energy between two bodies (i.e. an inbound rotamer and its chemical environment) through a single matrix operation. I integrate this innovative way of scoring into a design workflow and use it to create inhibitors of the epidermal growth factor receptor (EGFR) with anti-cancer function.

Overall, the research presented here focuses on establishing efficient, high-accuracy methods for protein binder design, aiming to eliminate the need for resource-intensive rounds of experimental binder optimization.

This thesis is organized as follows. Chapter 2 emphasizes the importance of protein-based binders by giving an overview of their diverse applications. Chapter 3 covers the basic concepts of protein-protein association and then dives into existing computational methods for designing novel binders. Chapter 4 highlights the persisting challenges in the field. Chapter 5 describes a new solution to tackle high-throughput docking problem in context of *de novo* binder design and presents experimental results on designing VEGF binders. Chapter 6 introduces a tensorized framework for accelerated design calculations and shows its utility for designing functional proteins. Chapter 7 discusses the strengths and limitations of the presented approaches and outlines future directions based on previous findings. Finally, Chapter 8 summarizes the main achievements of this work.

Chapter 2

Protein binders and their applications

Proteins exhibit exceptional functional versatility in living cells. They act as transporters, signaling molecules, catalysts, anti-infective agents, structural scaffolds, and much more. Therefore, the tools to control protein activities would greatly benefit both applied and fundamental biological research. Protein-based binders - reagents that bind a specific target protein - represent one of such tools to modulate or disrupt protein functions. The two following sections describe the use of protein binders in different fields.

2.1 Protein binders in the biomedical field

Over the past few decades, protein-based therapeutics have revolutionized medicine. Notably, half of the top ten selling drugs in 2023 are projected to be protein biologics [3]. Proteins represent a rivaling alternative to historically dominant small molecule drugs primarily for two reasons. First, owing to their complex three dimensional shapes, proteins can target diverse molecular surfaces that are otherwise undruggable by simple chemical compounds. And second, protein therapeutics are less likely to cause off-target adverse effects, since protein-protein interactions are highly specific [4, 5].

Protein-based binders serve a wide range of biomedical applications and can be classified functionally into the following groups: 1) protein agonists; 2) protein inhibitors; 3) proteins for targeted delivery; 4) protein diagnostics.

The first group includes therapeutics that replace natural protein activators in cases of endogenous protein deficiencies. A classical example is the treatment of diabetes with exogenous recombinant insulin, which binds to the insulin receptor and initiates signaling essential for glucose metabolism [6]. Other protein binders in this category are administered to enhance existing cellular activities, especially in haematological

pathways and immune responses. For instance, in patients with chemotherapy-induced neutropenia or anemia, recombinant cytokines such as granulocyte-colony stimulating factor and erythropoietin boost the production of neutrophils and erythrocytes in the bone marrow, respectively [7, 8].

The inhibitors group comprises binders that block functions of the target proteins. The major representatives of this group are monoclonal antibodies, which has been successfully used in oncology, inflammatory and infectious diseases. The prominent examples are pembrolizumab, an antibody suppressing the PD-1 receptor on lymphocytes to allow immune system to destroy cancer cells [9], and adalimumab, an antibody inactivating the tumor necrosis factor-alpha to treat rheumatoid arthritis and other autoimmune conditions [10]. Lately, non-immunoglobulin protein inhibitors with simpler architecture are also being developed extensively. Binders based on anticalin, affibody, monobody, and DARPin scaffolds are advancing through early stages of clinical study [11].

The third group refers to protein binders that facilitate the delivery of other proteins or small molecules to the intended therapeutic site. This targeted drug delivery enhances treatment efficacy while minimizing unwanted side effects. Several antibody-drug conjugates are currently in clinical use, such as brentuximab vedotin, which combines a synthetic chemotherapeutic agent with an anti-CD30 monoclonal antibody. The medication selectively targets tumor cells expressing the CD30 antigen, a defining marker of Hodgkin lymphoma and systemic anaplastic large cell lymphoma [12]. Additionally, a relatively young field of research explores the potential of protein binders in receptor-mediated intracellular and transcellular delivery. Sahtoe et al. designed small proteins that bind to the human transferrin receptor, which shuttles interacting proteins across the blood–brain barrier [13]. This holds promise for drug delivery into the brain and treatment of central nervous system diseases.

Besides therapeutic applications, protein binders are also utilized for *in vivo* and *in vitro* medical diagnostics. Radiolabeled binders serve as imaging reagents in positron emission tomography and single-photon emission computed tomography [14]. They allow to detect presence, localization, or dynamic changes of different molecular targets. For example, anti-granulocyte antibody besilesomab labeled with radioactive technetium-99m (^{99m}Tc) locates areas of bone infection [15], while ^{99m}Tc-labeled Fab fragment nofetumomab binds to the pancarcinoma antigen EpCAM and enables staging of breast, lung, kidney and other cancers [16]. Among non-immunoglobulin protein binders, affibody ABY-025 labeled with indium-111 or gallium-68 demonstrated ability to accurately quantify HER2 expression in breast cancer metastases [17, 18].

In the future, as disease etiologies are more clearly identified at the molecular level, biomedical applications of protein binders will definitely continue to expand.

2.2 Protein binders in basic research

Protein-based affinity reagents increasingly complement classical genetic approaches in elucidating biological processes. Life science research relies on protein binders to visualize a specific protein of interest (POI) or to manipulate its function in a predictable manner.

Visualization of proteins enables the investigation of protein localization and dynamics in living cells. Previously, *in vivo* visualization was achieved through direct tagging of an endogenous POI with a fluorescent protein. However, modification of an endogenous locus can alter expression and cellular behavior of a POI. Protein binders fused to fluorescent proteins, known as chromobodies, have emerged as an alternative approach. These protein binders target POIs and trace them without functional interference [19, 20]. Prominent examples include actin-binding proteins, which are widely used for live imaging of the cytoskeleton with high spatiotemporal resolution [21, 22]. Notably, protein binders can be designed to specifically target different isoforms or post-translational modifications of a POI. For instance, a DARPin-based binder derivatized with a merocyanine dye serves as a biosensor specific for phosphorylated ERK. It could selectively visualize activated endogenous ERK in mouse fibroblasts, revealing ERK translocation patterns [23].

The ability of binders to either agonize or antagonize protein functions is instrumental in dissecting intracellular signaling pathways. Numerous binding reagents have been developed to interfere with disease-related POIs, yielding valuable insights into how changes in indivividual interactions influence cellular physiology [24]. For example, monobodies targeting separate domains of the oncogenic phosphatase SHP2 helped to reveal its role in the context of leukemia signaling network [25]. Similarly, synthetic proteins binding to the transcriptional regulator Prdm14 contributed to uncovering the molecular mechanism by which Prdm14 preserves stem cell pluripotency [26].

Binding proteins can also work as customized ligands for affinity chromatography. Binders derived from repebody [27] and affitin [28] scaffolds have been successfully immobilized on a resin and used for affinity purification of various target proteins with high degree of purity and recovery in a single step. Additionally, some protein binders serve as crystallization chaperones for structure determination of poorly crystallizable POIs [29, 30, 31]. The reason for this is that a tight binding partner can reduce the conformational flexibility of a target and facilitate the formation of well-ordered crystals.

To conclude, the possible applications of protein-based binders are highly diverse and extend beyond the above-mentioned examples.

Chapter 3

Computational design of protein binders

As discussed earlier, the capacity to create on-demand protein binders has vast potential for advancing therapeutics and synthetic biology tools. To design such binders, it is necessary to understand the core principles of protein-protein interactions (PPIs). Thus, this chapter first covers forces governing biomolecular recognition and then outlines how computational approaches can transform our knowledge about protein association into designing novel PPIs.

3.1 Driving forces for protein-protein association

The association of protein complex in aqueous solution arises from multiple processes, such as initial formation of intermolecular contacts, rearrangement of interfacial water, and establishment of short-range interactions between binding partners (Fig. 1).

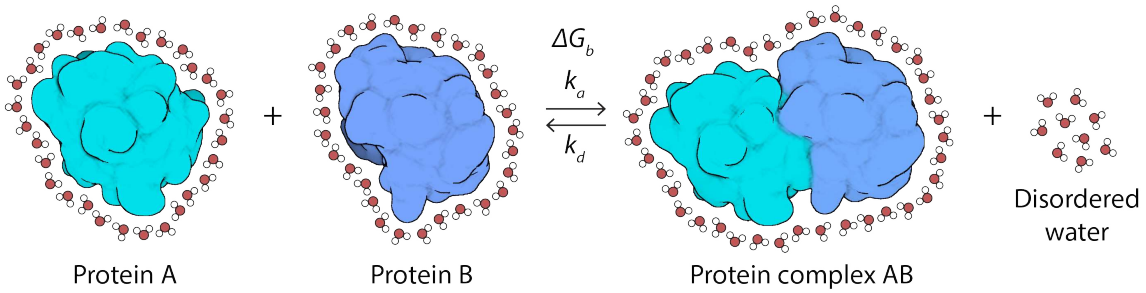


Figure 1: Association of two proteins (A and B) in water solution.

In its simplest form, reversible complex formation can be expressed as:

$$A + B \underset{k_d}{\overset{k_a}{\rightleftharpoons}} AB \tag{1}$$

where k_d is the first-order dissociation rate constant and k_a is the second-order association rate constant for the interaction of A with B. According to the law of mass action, the ratio between the rate constants define the equilibrium constants for dissociation (K_d) and association (K_a):

$$\frac{[A][B]}{[AB]} = K_d = \frac{1}{K_a} = \frac{k_d}{k_a} \tag{2}$$

Dissociation constant, in turn, represents the strength of biomolecular interaction - the smaller K_d, the higher affinity between a protein and its ligand. This parameter can also be used to calculate the binding energy, i.e. Gibbs free energy difference between the bound and unbound states (ΔG_b):

$$\Delta G_b = RT\ln(K_d) \tag{3}$$

where R is the gas constant and T is the temperature in Kelvin.

Similar to protein folding, protein association is guided by energy minimization. With biologically relevant K_d values ranging from 10^{-4} to 10^{-14} M, the binding energy falls within the negative range of approximately -6 to -19 kcal/mol [32]. To further comprehend the events driving PPIs, it is essential to break down the free energy into its enthalpic (ΔH) and entropic (ΔS) contributions. Eq. 4 illustrates the relationship between the aforementioned thermodynamic parameters:

$$\Delta G_b = \Delta H - T\Delta S \tag{4}$$

The binding enthalpy generally refers to the energy changes resulting from the formation of hydrogen bonds, van der Waals and electrostatic interactions at the protein-protein interface. Additionally, changes in other electromagnetic interactions of the system, for example loss of hydrogen bonds between water molecules and a protein, should also be considered. In the study summarized by Stites [33], binding enthalpies for 69 protein-protein and protein-peptide complexes were calorimetrically determined. ΔH values spanned from positive to negative. However, in most cases (74%) protein association was enthalpically controlled, with $\Delta H < 0$. In contrast, entropic contribution was favorable ($\Delta S > 0$) only in 55% of complexes.

The entropic component reflects changes in the number of possible arrangements that the system can adopt. It mainly accounts for accessible macromolecular ($\Delta S_{protein}$) and solvent ($\Delta S_{solvent}$) configurations (Eq. 5). Macromolecular entropy represents protein states with respect to its conformational (i.e. backbone and side chain conformations), rotational and translational freedom. While typically upon protein association backbone reorganizations are minor, flexibility of the side chains at the interface decreases, leading to an entropic loss [34]. Additionally, complex formation causes reduction in

rotational-translation entropy of a protein compared to its unbound state [35].

$$\Delta S = \Delta S_{protein} + \Delta S_{solvent} \tag{5}$$

Shifting to solvent entropy, PPIs lead to rearrangements of water solvating each protein (Fig. 1). Water molecules that were initially engaged in strong hydrogen bonds at nonpolar surface patches form weaker hydrogen bonds in the bulk, resulting in decreased order and increased entropy. This tendency to minimize the number of ice-like ordered water molecules by burying nonpolar patches in the protein interior or at the protein-protein interface is known as the hydrophobic effect. Mostly entropic in nature, the hydrophobic effect is one of the major driving forces of protein folding and association.

While high-resolution structures of protein complexes, in conjunction with calorimetric measurements, largely elucidate the origin of the binding enthalpy, contributions to the total binding entropy, especially its conformational component, remain challenging to experimentally dissect and, as a result, to model. Therefore, the design field traditionally simplifies the total system entropy to the solvent entropy only and correlates it with changes in the accessible surface area of the proteins upon complexation [36, 37].

As introduced above (Eq. 4), both enthalpic and entropic components are crucial in determining the sign and magnitude of the binding free energy. Empirical data on thermodynamics of PPIs indicate that the association enthalpy ΔH alone is not necessarily predictive of ΔG_b. Likewise, there is no obvious correlation between the entropic component $-T\Delta S$ and ΔG_b (Fig. 2A,B).

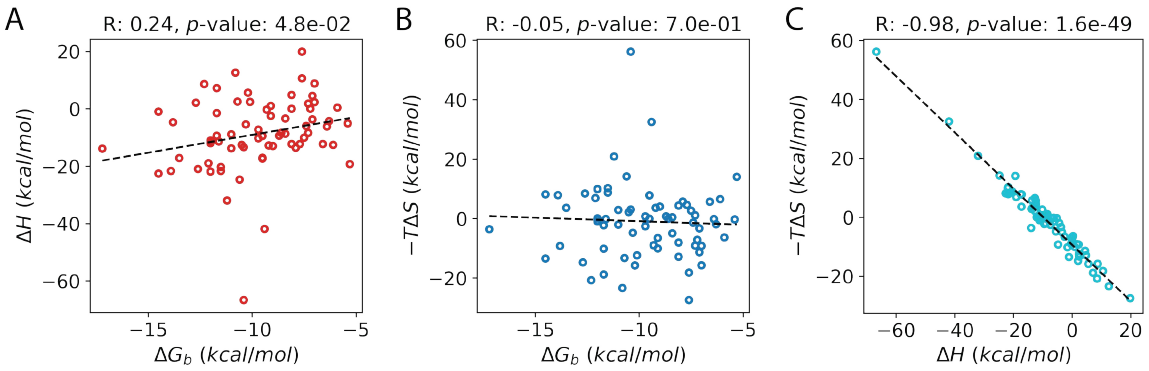


Figure 2: Correlations among thermodynamic properties in protein-protein association. Experimental data reported in [33] show a weak correlation between enthalpy **(A)** or entropy **(B)** and the free energy of binding. **(C)** Correlation between the enthalpic and entropic components of binding energy exhibits apparent compensation behavior. The Pearson correlation coefficients (R) and the correlation p-values are shown.

Notably, the contributions of ΔH and $-T\Delta S$ are closely related to each other (Fig. 2C). Formation of multiple noncovalent bonds at the protein-protein interface causes a large, favorable, negative enthalpy change. At the same time, this is typically

coupled with a decrease in entropy as the mobility of the interacting partners is constrained. Such enthalpy-entropy compensation results in the medium-magnitude free energy change and complicates the prediction of binding energetics. For binder design, this compensation phenomenon means that improving the enthalpy or entropy of interaction can be completely counterbalanced, leading to no net gain in binding affinity [38]. For example, a favorable hydrogen bond with an energy of around 1.5 kcal/mol should theoretically reduce K_d by a factor of ten at 298 K. In practice, however, additional hydrogen bond between a protein and its ligand often leaves affinity unaltered [39]. Similarly, an ionic bond between a pair of oppositely charged aspartate and arginine residues was found to have a negative enthalpy of -3.3 kcal/mol, while the corresponding interaction free energy approached zero [40]. Chapter 4 will describe how existing computational approaches address the challenges related to accurate prediction of binding energies.

The final protein association factor I would like to introduce here is the shape complementarity of interacting surfaces. The key role of shape matching in PPIs became evident from the earliest reports on the tightly packed natural protein interfaces [41, 42]. Although geometric complementarity does not directly represent a physical interaction, it is linked to nonpolar desolvation and van der Waals forces. A recent study by Desantis et al. [43] demonstrates that both local shape complementarity and van der Waals interactions exhibit a linear correlation with binding affinity of protein-protein complexes. Analogously to colloid systems, shape complementarity can be also considered as an attractive depletion force in a solution, where smaller solutes are preferentially excluded from the proximity of large proteins [44, 45]. Simulating generic depletion interactions between protein pairs shows that in some cases shape alone (without electrostatic and hydrophobic complementarity) is sufficient for protein assembly.

It is also important to note that binding interfaces often contain cavities, most of which are filled with water molecules to maintain close atom packing. These water molecules can participate in hydrogen bonds with polar protein groups and bridge interactions between two proteins. However, a currently prevailing concept suggests that only a small fraction of buried waters are truly bridging, whereas the majority forms favorable interactions with only one protein and has no significant effect on protein binding [46, 47].

Taken together, hydrophobic effect, van der Waals and electrostatic interactions are the major driving forces of protein association and each should be considered when designing novel binders.

3.2 Advantages of computational binder design over empirical methods

During the previous decades, various approaches were developed to create custom protein binders with desired affinity and specificity. Directed evolution has proven to be the most productive strategy thus far [48]. This approach deploys iterative rounds of *in vitro* display to screen protein libraries based on immunoglobulin or non-immunoglobulin (e.g. DARPins, anticalins, affibodies) scaffolds and select candidates with a desired phenotype. While still the method of choice, *in vitro* display requires considerable experimental effort and often yields binding to irrelevant epitope rather than the functionally important target sites. Moreover, the maximum library size that can be practically displayed strongly limits the number of mutable positions that can be diversified per screening round [49]. Even techniques like ribosome [50], mRNA [51], and CIS [52] display handle no more than 10^{13} unique sequences per cycle, which corresponds to a maximum of 10 mutable positions.

Computational protein design, in contrast, stands as an innovative and accelerated approach that provides control over the biophysical properties of resulting binders and allows to target predefined epitopes [53]. In broad terms, protein design is an *in silico* search for a sequence that guides folding and stabilization of a query protein structure or function. Being built on the hypothesis that proteins fold into the lowest energy state, protein design process can be viewed as sampling of sequence and conformational spaces followed by free energy estimation of the sampled states. Unlike empirical selection techniques, protein design rationalizes the discovery process and hence can broaden our understanding of fundamental concepts that govern biomolecular interactions [54].

3.3 Strategies to design binding proteins

Previous studies have demonstrated successful protein binder design utilizing different strategies. I will give a brief overview of some key achievements in the field, categorizing them into two groups: 1) template-based design and 2) *de novo* design (Fig. 3).

Template-based design. In the majority of cases, information from the structures of existing protein complexes serves as input for the design process. Natural binders or separate binding domains can be reengineered by introducing individual or combinatorial mutations with the aim of improving the initial thermostability, solubility, or affinity. A wild-type angiotensin-converting enzyme 2 was computationally redesigned with the Rosetta software [55] yielding mutants that bind to the SARS-CoV-2 spike receptor binding domain from 3- to 11-fold tighter compared to the wild-type [56]. Likewise,

using the OSPREY design framework [57] Lowegard et al. discovered a novel point mutation that enhances affinity of the Ras-binding domain of the c-Raf kinase to K-Ras, a protein implicated in difficult-to-treat cancers [58].

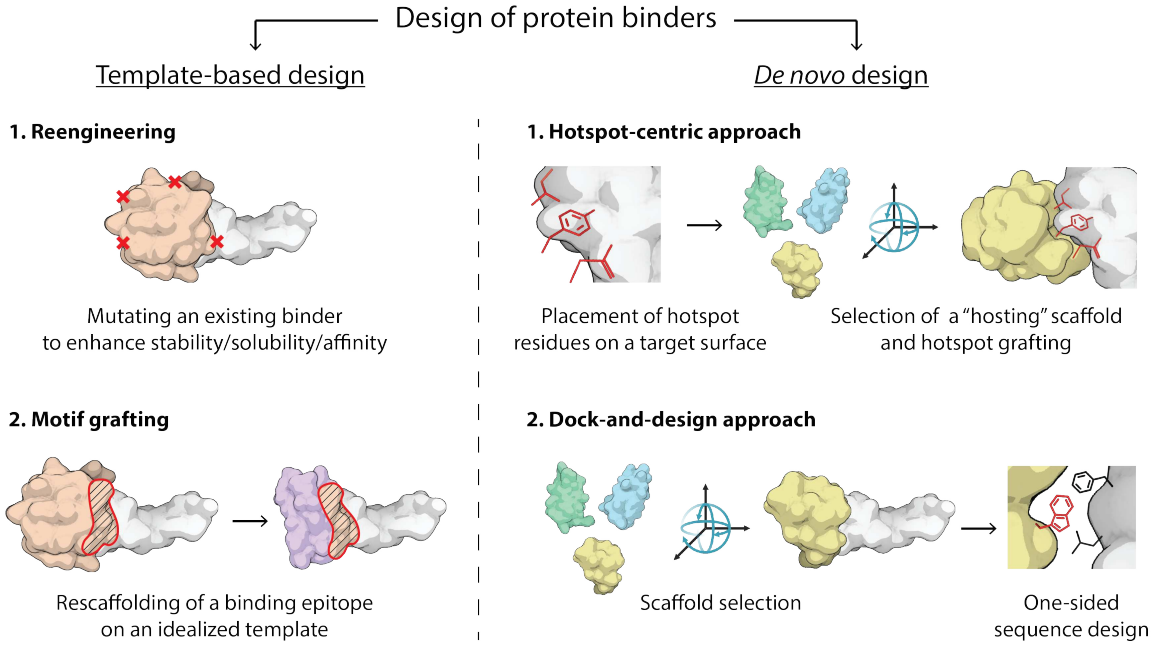


Figure 3: Methods to design protein binders. While template-based approaches rely on existing binding interfaces for a given protein target, *de novo* design allows to generate site-specific binders from the target structure alone.

Another approach that also relies on formerly described binding interfaces is incorporating key interaction residues into new geometrically-accommodating idealized scaffolds, i.e. motif grafting. For example, the inhibitor of the Epstein-Barr virus anti-apoptotic protein (BHRF1, a Bcl-2 homolog) was designed by transplanting a functional continuous motif from its natural partner on several fixed backbones [59]. This work was further expanded by rescaffolding additional amino acids on the previously designed inhibitor in order to tune specificity for multiple Bcl-2 homologs [60]. Recently, Bryan et al. also employed "natural hotspot" idea to create binders targeting PD-1 protein, which is essential for proper functioning of immune system [61]. In this case, instead of being mounted onto selected scaffolds, a natural binding motif was used as a seed around which *de novo* backbones were generated via Monte Carlo-based fragment assembly. Correia's lab showed utility of similar function-centric strategy (i.e. "bottom-up" design) to target not only contiguous sequence motifs, but also complex epitopes consisting of non-adjacent elements [62].

Over the past few years, deep learning (DL) approaches have gained substantial power and widespread use in protein modeling and design. Dauparas et al. lately described ProteinMPNN, a DL method which generates amino acid sequences given a

protein backbone [63]. The authors argued that sequences suggested by ProteinMPNN are predicted to fold to native backbones more confidently than the original native sequences. Therefore, ProteinMPNN holds great potential for enhancing expression and stability of existing proteins while preserving their functionality.

Motif grafting can also benefit from DL techniques, such as hallucination and inpainting. Unlike classical protein design frameworks, these methods simultaneously generate both sequence and structure starting from a given functional site. Hallucination coupled with further inpainting allowed to transplant two identical helical segments on a single chain, yielding a bivalent binder of the nerve growth factor receptor [64]. Hallucinating diverse protein binders is feasible even from minimalist functional sites. However, this approach is quite slow, especially for large systems. Meanwhile, inpainting is much more computationally efficient, but fails in case of small input motifs. Probably the most promising DL method for epitope scaffolding hitherto is RFdiffusion. By utilizing the 3D coordinates of a defined motif as input, RFdiffusion outputs a backbone capable of holding the motif in precise geometry. Generation of hosting templates with RFdiffusion followed by sequence optimization enabled site grafting with a success rate of around 50%. Designed binders carry a helix from the tumor-suppressing protein p53 and potentially can disrupt its interaction with the negative regulator MDM2 [65].

Overall, template-based strategies are rapidly advancing, delivering successful design outcomes. Yet, a remaining limitation of these methods is their dependence on previously observed PPIs. To go beyond existing interfaces and to extend the range of targetable protein sites, *de novo* design approaches are needed.

De novo design. Template-free binder design, solely based on the structural information of the target, remains a formidable challenge for the field. It requires a comprehensive understanding of biomolecular recognition alongside technological advancements. Nevertheless, some initial milestones in that direction have been already achieved.

Current strategies for *de novo* binder design can be divided into sequence-first (i.e. hotspot-centric) or scaffold-first (i.e dock-and-design) approaches, depending on whether sequence optimization or scaffold docking problem is prioritized (Fig. 3). The hotspot-centric approach was initially introduced by Fleishman et al. based on the observation that the free energy of binding is not evenly distributed across protein-protein binding region. Instead, interfaces typically contain residue clusters, known as hotspots, that are involved in highly optimized hydrogen bonding, tight van der Waals packing, and favorable electrostatics; and contribute a large fraction of binding energy [66, 67]. The design protocol starts with constructing a hotspot region by docking the disembodied amino-acid residues against the target surface and proceeds to mount these hotspots onto selected shape-complementary scaffolds with

final interface refinement. These steps were implemented to design binders targeting the stem of the influenza hemagglutinin [68]. Among the 73 experimentally tested designs, two exhibited interaction with their target. However, due to relatively low affinity, they had to undergo two rounds of *in vitro* maturation to improve binding. Notably, the described method relies on a small number of hotspots (i.e. 2-3 residues) with high energetic signatures, ideally placed in conspicuous pockets on the target surface. This significantly restricts possible interaction modes and excludes targeting of flat interfaces that lack deep cavities. Also, at the time of this study, selecting hotspot identities required human expertise and was not automated due to poor modeling of energetics for disembodied residues.

Recently, Baker's group computerized and generalized the hotspot-centric approach and applied it to make binders against SARS-CoV-2 spike protein [69] and twelve other diverse targets [1]. The new idea is to identify as many individual interactions with the target as possible, even if they are weak, rather than concentrating on a few strong interactions. For this, the framework takes advantage of the rotamer interaction field (RIF) method, originally developed for design of the small-molecule binding sites [70]. Briefly, discrete amino acid rotamers are sowed over the target epitope with the aim of introducing favorable hydrogen bonds and hydrophobic packing interactions. All RIF rotamers are stored and can be rapidly looked up for scaffold matching using a docking grid-based search algorithm. This allows to consider billions of potential interfaces. Although this work is one of the first to demonstrate purely *de novo* binder design, it has significant drawbacks. The presented protocol is complicated and time-consuming to run, since it entails numerous extra steps of fast scoring, interface design, motif extraction, re-grafting, and others. More importantly, this method suffers from a low experimental success rate (below 1%), requiring many thousands of designs to be screened for each design campaign.

One of the main reasons for the low success of RIF method, and sequence-first strategies in general, can be attributed to the extreme combinatorial load of discrete rotamer sampling in the absence of a backbone. The complexity of hotspot grafting grows even higher, given for instance tens of thousands of possible protein scaffolds and nearly 10^{16} interface sequences per scaffold placement [1]. To reduce this complexity, an alternative geometric DL tool, MaSIF-seed, has been developed by Gainza et al. [71, 2]. Based on geometrical and chemical complementarity, it identifies structural motifs, such as helical segments and 2- or 3-strand β-sheets, able to engage specific epitopes. Subsequently, these structural motifs are transplanted onto protein scaffolds using established grafting techniques. MaSIF-seed was applied to design protein binders for four different therapeutically-relevant targets - SARS-CoV-2 spike, PD-1, PD-L1, and CTLA-4. However, among 2000 tested sequences, only one PD-1 binder showed

specific interaction with its target with K_d of 4 μM, while binders to other targets were subjected to experimental optimization by *in vitro* evolution to achieve reasonable affinities. This highlights that although contacts at the buried interface region drives molecular recognition, they are insufficient for tight binding.

Undoubtedly, the above-mentioned studies represent a major step forward in design of novel PPIs. Nonetheless, given the current state of sampling and scoring methods, optimizing a binding sequence before selecting a structural template apparently complicates the design pipeline rather than brings benefits. An orthogonal strategy, therefore, is to avoid hotspot or seed construction, begin straight from rigid-body docking of a scaffold against the target epitope, and afterwards proceed with one-sided sequence design to form favorable interactions with the target. Rosetta-based DDMI (dock, design, and minimize interface) protocol was one of the earliest attempts in this direction [72]. Proteins targeting p21-activated kinase 1 were designed starting from a small helical bundle scaffold. 3 out of 6 tested designs exhibited weak micromolar binding. This emphasized the necessity to screen larger pools of scaffolds and to develop more efficient methods for protein-protein docking [73]. Despite the dock-and-design idea originating about a decade ago, not much progress has been reported on this front. Hence, during my PhD I aimed to establish and validate a new scaffold-first framework to improve the success of a *de novo* design campaign. Chapters 4 and 5 will provide further elaboration.

From another perspective, innovative DL methods, particularly diffusion models, can potentially transform dock-and-design problem into generate-and-design, where binder template is directly generated in the context of the target. First promising results were obtained using RFdiffusion method [65]. They demonstrated generation of binding scaffolds with further sequence optimization for five targets. The overall experimental success rate was 19%, the highest for *de novo* binder design thus far.

In conclusion, computational methods to design novel protein binders are continuously expanding and addressing tasks of different complexity. For interfaces with previously known binding partners, template-based approaches provide reliable solutions. Meanwhile, template-free design remains an outstanding challenge, motivating the development of additional *de novo* strategies.

Chapter 4

Challenges in *de novo* binder design and ways to address them

4.1 Single binder, multiple expectations

To effectively serve its function, an ideal binder must possess maximum affinity and specificity to the pre-defined epitope. Apart from this, it must exhibit high conformational and colloidal stability in the absence and presence of its target. It is commonly observed, however, that improving one of these properties often compromises the others [74, 75]. For instance, hydrophobic mutations at the binding site are expected to be beneficial for affinity, but might be detrimental for specificity and monomer solubility. Likewise, interfacial hydrogen bonds and salt bridges ensure specificity of association, but entail a risk of complex destabilization due to the desolvation effect [76, 77]. For design process, it means that all determinants of successful binder should be co-optimized simultaneously. This poses the scoring problem, which relates to accurate estimation of the multitude of factors such as shape complementarity, hydrophobic and polar interactions at the interface, and the overall solvation energy of both bound and free forms. Detailed insights into the scoring problem and potential solutions to mitigate it will be presented in Sections 4.3 and 4.4.

Another challenge is the extreme dimensionality of the search space, given that binder design process implies selection of: 1) a protein backbone from a pool of scaffolds N_{scaf}; 2) a suitable pose to the target epitope from possible translations and rotations $N_{trans}^3 \cdot N_{rot}^3$; 3) amino acid identities at the binding interface from $20^{N_{mut}}$ combinations for the mutable positions N_{mut}; and 4) side chain rotamers $N_{rotam}^{N_{mut}}$ for the mutable positions N_{mut}. This huge solution space of $N_{scaf} \cdot N_{trans}^3 \cdot N_{rot}^3 \cdot 20^{N_{mut}} \cdot N_{rotam}^{N_{mut}}$ cannot be covered systematically (Fig. 4).

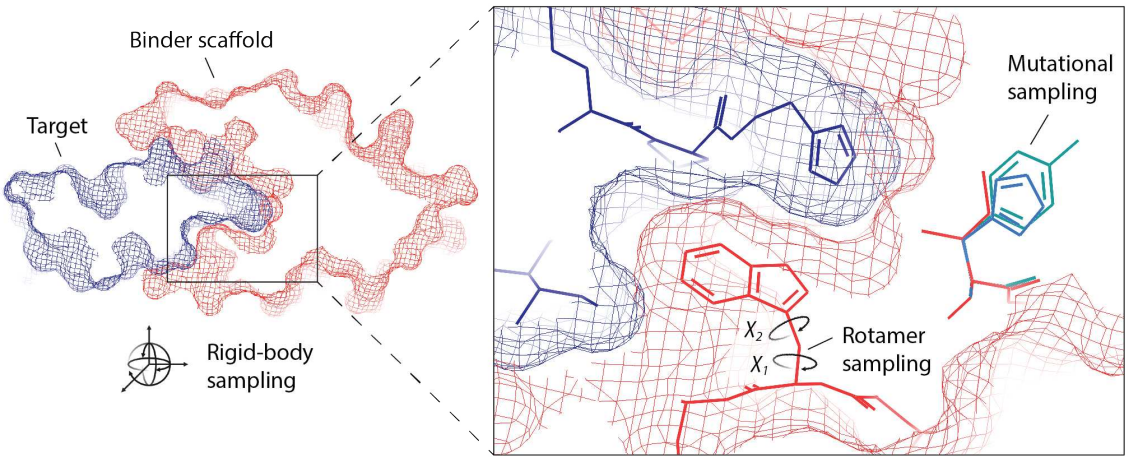


Figure 4: The complexity of binder design arises from combined high-throughput docking and design problems, where scaffolds, rigid-body transformations, side chain identities and rotamers are combinatorially sampled.

A workaround to this challenge is to separate docking stage from sequence design (i.e. dock-and-design strategy). This approach prioritizes straightforward geometric features before delving into more complex physicochemical calculations. The pre-selection of a binder scaffold and subsequent design stage within a fixed-backbone context enable a more comprehensive exploration of the design space. Following this decomposition idea, the next sections will discuss high-throughput docking and sequence design as standalone aspects.

4.2 Protein-protein docking problem

In contrast to empirical methods that rely on a limited set of binder scaffolds, computational design can leverage a pool of all structurally characterized or modeled proteins as potential backbones. This significantly enhances the likelihood of finding suitable molecular topographies to quench a specific epitope. However, the difficulty lies in efficiently processing the vast number of protein structures and selecting those matching the target shape.

To enable protein-protein docking within a feasible timeframe, proteins are typically treated as rigid bodies and the docking is considered as a steric problem. Yet, the computational load of the process remains huge, given extensive rotational and translational transformations that two 3D objects must undergo to find their mutual orientation with maximized complementarity. For years, different search algorithms have emerged, such as fast Fourier transform [78, 79], spherical Fourier transform [80], local shape feature matching [81, 82], and randomized search [83, 84]. While more productive than the standard exhaustive global search, mentioned approaches are still resource-intensive. The average running time for docking a pair of proteins using a

single CPU ranges from minutes to hours [85], which is incompatible with screening large up-to-date structural databases.

An alternative solution for docking without explicit 3D alignment involves employing geometric invariant descriptors, or fingerprints. A geometric fingerprint captures surface patterns and represents them as sets of scalar values that remain constant upon translation and rotation [86]. Initially used in computer vision for object recognition, fingerprinting idea found application in protein biology for fast structure similarity searches [86, 87, 88]. In case of docking, however, it is essential to evaluate protein shape complementarity rather than similarity. To this end, Venkatraman et al. presented the LZerD algorithm that utilizes rotation-invariant 3D Zernike descriptors to score local complementarity of docking models [89]. Although LZerD showed promising accuracy in predicting protein complexes, the assessment of complementarity proved to be much slower than similarity search. It can be attributed to the time-consuming pre- and post-processing steps of the method, such as generation of docking decoys by geometric hashing and final clashes scoring [90].

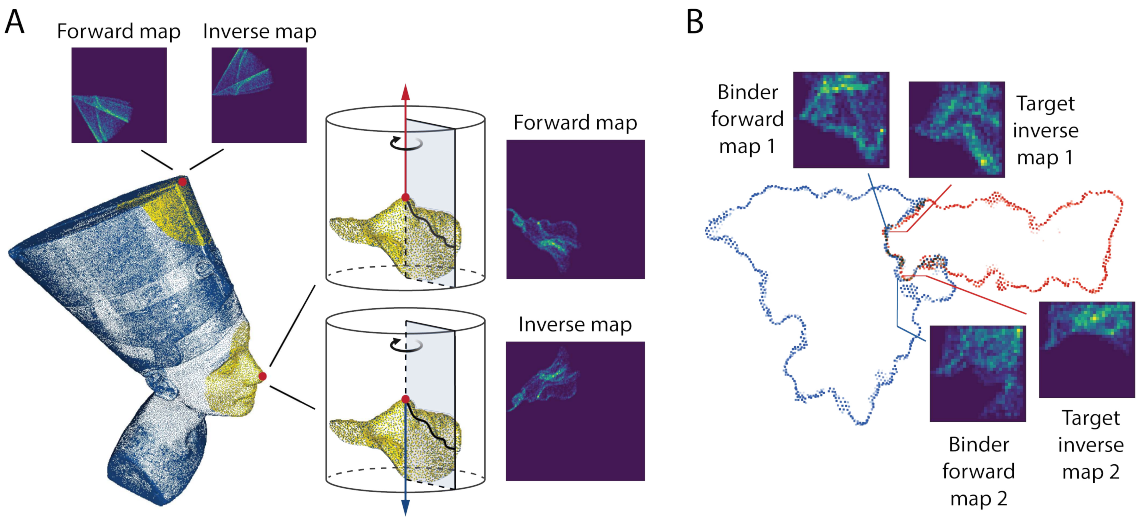


Figure 5: The HECTOR fingerprint captures local surface complementarity. **(A)** The HECTOR mapper takes a dot-surface of an object as input. Surface patches are extracted around a selected dot (e.g. red dot at the center of the golden face and crown patches from Nefertiti Bust). A map is compiled by rotational accumulation of dot density on a matrix. Depending if the surface normal pointing outwards or inwards of an object, the map is defined as "forward" or "inverse". **(B)** Forward and inverse maps at highly complementary interfaces are highly similar. Shown is an example of a tight interface between VEGF (red surface) and its antibody (bevacizumab; blue surface).

In light of the persistent challenges in protein-protein docking, my goal was to establish a new pipeline for high-throughput screening of protein structural databases and identifying potential binder scaffolds based on their complementarity to the target epitope. The starting point for this work was the innovative fingerprinting method -

Highly Efficient Complementarity Testing by Obverse Residuals, HECTOR - developed in our lab. In contrast to previously reported geometric descriptors, the HECTOR fingerprint represents a 3D molecular surface patch as a 2D matrix, which is not only rotation- and translation-invariant, but most importantly can be "flipped" to describe the ideally complementary patch (Fig. 5). Such fingerprint invertibility transforms complementarity evaluation task into simple matrix-matrix comparison, where surface patches of a target are inverse-mapped, and surface patches of putative scaffolds are forward-mapped. A single comparison between query and subject maps takes less than a microsecond, owing to efficient implementation on graphics processing units (GPUs). Consequently, blind docking of two proteins, represented on average by several thousands of surface patches, would take fraction of a second. Fig. 6 illustrates the main steps of selecting a template for binder design using HECTOR fingerprinting.

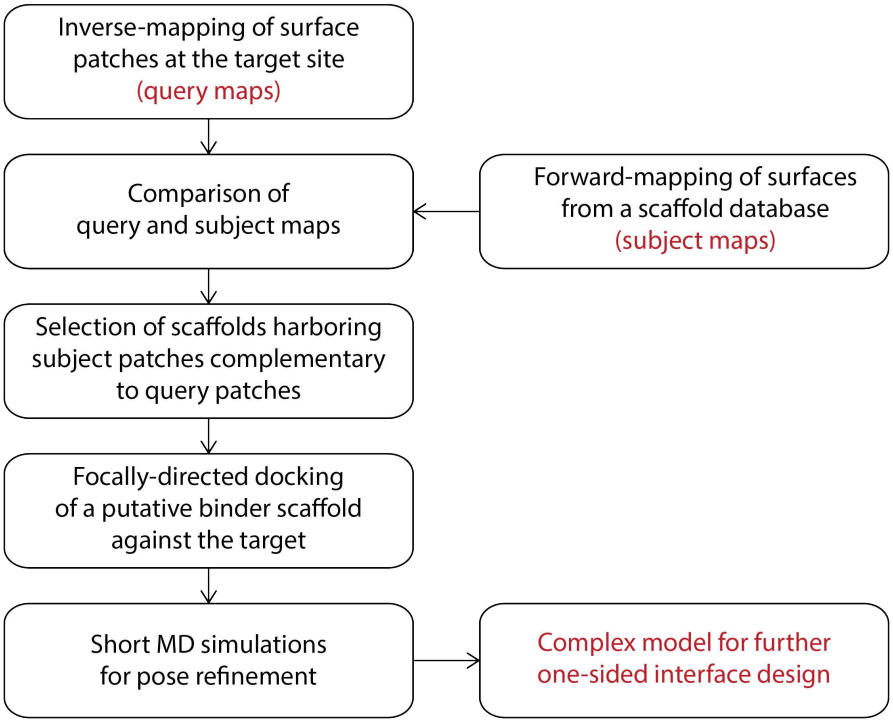


Figure 6: The workflow for identifying binder scaffolds with high shape complementarity to the target site. The input includes a target structure and a database of potential scaffolds. The output is a predicted model of a complex between the selected scaffold and the target, which can be used at the following sequence design stage.

To validate the utility of this docking approach on a real-world problem, I sought to use it for designing *de novo* binders against VEGF, a key angiogenic molecule, involved in the pathogenesis of diverse cancers, cardiovascular, and ophthalmic diseases [91]. Starting from two surface patches at the VEGF binding site, HECTOR software successfully identified the complementary backbones among crystal structures from

the Protein Data Bank (PDB). Two best candidates were selected for further design. Both scaffolds showed strong shape complementarity to the target site and much simpler topology compared to available antibody-based VEGF inhibitors. In Chapter 5, I describe the whole design pipeline in detail and present results of experimental characterization of the designed binders.

4.3 Trade-off between design speed and accuracy

After identifying shape-matching scaffolds, the second step in *de novo* binder design protocol involves selection of amino acid identities and side chain conformations to stabilize the backbone and to optimize interactions at the binding site. Since protein folding and protein association processes are driven by entropic effects and electromagnetic forces, the points addressed here are relevant for both general sequence optimization, and design of binding interfaces.

Sampling. Sequence design is typically formulated as a discrete optimization problem where the goal is to find a combination of rotamers minimizing the energy of a given structure [92]. The rationale behind discrete rotamer sampling has two aspects. First, statistical analysis of experimental protein structures together with MD simulations revealed that side chain torsion angles (χ_i) are not evenly distributed in dihedral space, but they rather cluster around certain values [93, 94]. This allows to construct rotamer libraries by space binning and determining a representative conformation in each bin. Second, although less realistic, discrete rotamer sampling facilitates a broader exploration of design solutions and enables energy barrier hopping, which is essential to overcome entrapment in local minima [54].

Small-scale design problems with a few mutable positions can be solved using deterministic algorithms, such as dead-end elimination (DEE) [95, 96]. DEE yields reproducible solutions and reduces the size of conformational search space by pruning branches of a decision tree if they inherit rotamers that are provably not part of the global minimum energy conformation. However, as the number of designable residues increases, the side-chain optimization problem becomes intractable for exhaustive deterministic sampling. Instead, full-fledged *de novo* design problems are usually handled with stochastic methods, such as simulated annealing or genetic algorithms [97]. For instance, Rosetta, the most widely used protein design software [55], takes advantage of Monte Carlo sampling with simulated annealing. Rosetta's design algorithm walks through the search space randomly rather than exhaustively. It accepts or rejects single amino acid mutations or rotamer substitutions on the basis of Metropolis criterion. Mutations that lower energy of the system are accepted, and mutations that increase energy of the system are accepted only at some probability dictated by energy change

and simulation temperature [98, 99]. Unlike DEE algorithms, Monte Carlo sampling does not guarantee to find an exact global minimum; nevertheless, it provides a population of decoys, among which some may be sufficiently accurate for various tasks.

Scoring. Amino acid mutations and their associated side chain conformations are evaluated using energy scoring functions, which aim to approximate atomic contacts in a physical environment. The most critical energy terms for protein folding and binding describe non-bonded interactions. While earliest scoring metrics considered only van der Waals forces [100], advanced physics-based functions now also account for electrostatics and desolvation [55]. Estimating these interaction energies requires calculating all pairwise atom-atom relations, which in turn makes scoring process extremely costly. Accelerated approaches, meanwhile, often come at the expense of scoring accuracy. For example, instead of a full-atom residue modeling, side chains can be represented as a single "centroid" pseudo-atom [101]. The centroid's location is defined using known structures from PDB as the average relative location of side-chain atoms of the same residue type. This simplified representation, coupled with a reduced number of energy terms, enables a low-resolution scoring function to rapidly scan conformational space, which can be valuable for the initial filtering of design candidates.

Another way to address scoring complexity is to use statistical terms rather than physical ones. Predicting hydrogen bond formation can serve here as an illustrative example. Previous studies highlighted extreme importance of hydrogen bonds at the protein-protein interfaces for binding specificity. However, current physics-based energy functions fail to 1) account for complex nature of hydrogen bonding, and 2) capture the counterbalancing effect of desolvation penalties [102]. Instead, Kortemme et al. have developed a simple orientation-dependent hydrogen bonding function based on PDB statistics and electronic structure calculations [103, 104]. This knowledge-based potential is incorporated into the current version of Rosetta score function. The limitations of PDB-based energy terms can be attributed to inaccuracies in the deposited structures, frequently associated with cryogenic measurement conditions and ensemble averaging. Additionally, knowledge-based terms are generally less interpretable and transferable when compared to physics-based terms.

With the emergence of machine-learning approaches, it is now also possible to build models that learn residue-level patterns for local backbone structure and chemical environment. These models predict the probability distribution of amino acid identities at each residue position and can be used for sequence design. Recently, several DL-based methods were developed and successfully validated using crystallography and cryogenic electron microscopy [63, 105]. Again, the drawback of these approaches lies in the lack of physical transparency, as residue choice is solely guided by biases in sequence distributions among database structures.

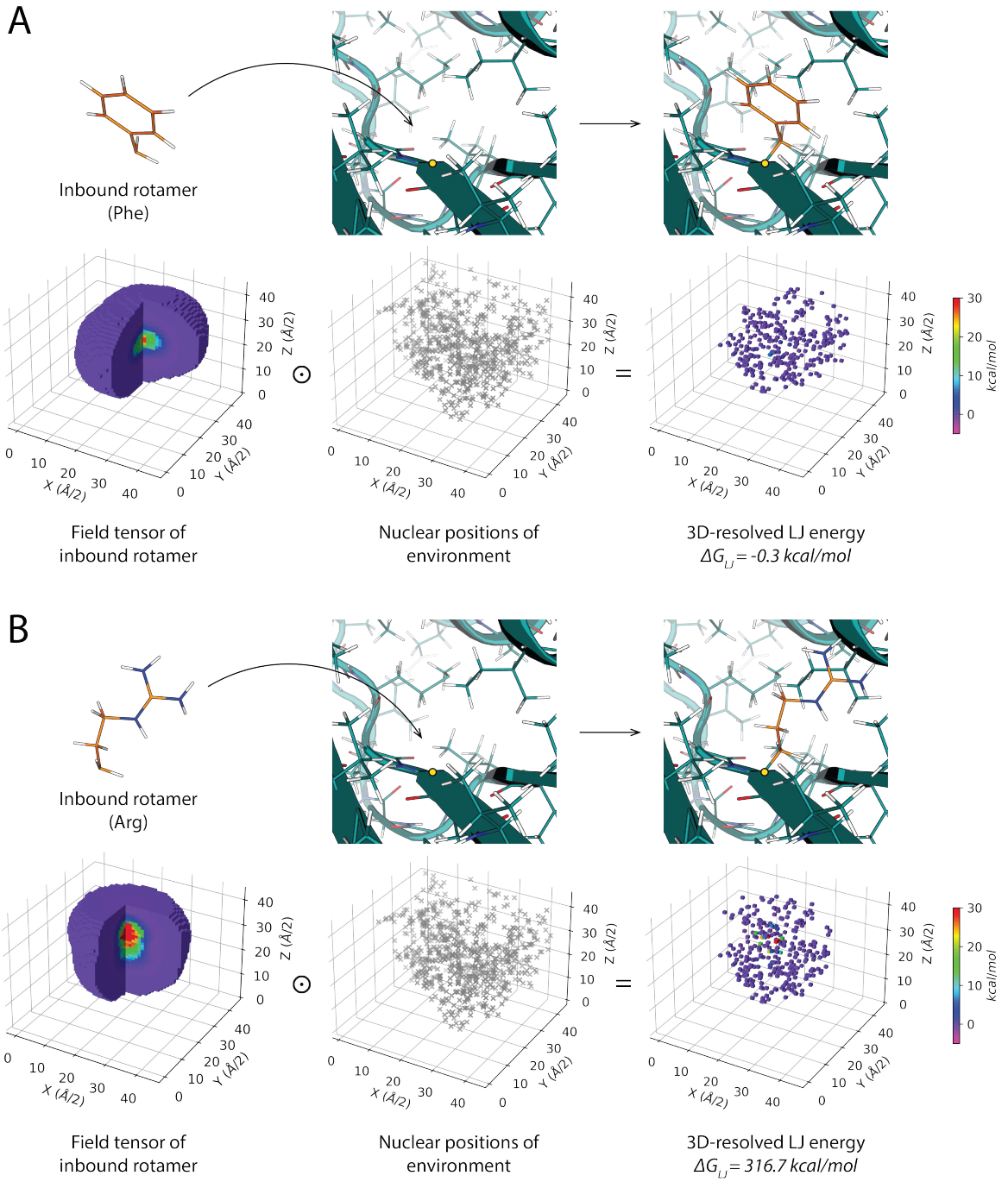


Figure 7: The concept of tensorized energy calculations. **(A)** and **(B)** show examples of scoring the Lennard-Jones (LJ) potential ΔG_{LJ} following the mutation of a single position to phenylalanine and arginine, respectively. While the phenylalanine rotamer fits well into the molecular environment, the arginine rotamer clashes with surrounding atoms, resulting in much higher LJ interaction energy. The workflow starts with removing the side chain of the designable residue from the input structure. Remaining atoms form the "environment". The environment tensor is populated using values of the atoms' positions. The rotamer library provides the pre-computed, smooth interaction fields (i.e. field tensors), which through a single element-wise multiplication with the environment tensor yields the spatially resolved energy. In case of calculating electrostatic or solvation terms, environment tensor contains not only information on the atoms' positions, but also on their partial charges or surface solvation energies.

To summarize, the shortcomings of protein design process are associated with either incomplete conformational sampling or inaccurate scoring functions. Therefore, our group started a project aiming to introduce an alternative framework for design calculations that enables simultaneous improvements in accuracy and speed on the software and hardware levels. The main concept is to substitute exhaustive calculations of the pairwise atom-atom interactions with a single matrix operation that evaluates interaction energy between two bodies - an inbound rotamer and a chemical environment. It can be done by representing the dense atomic interaction fields as 3D projections (i.e. tensors) (Fig. 7). Uniform tensor dimensions allow to achieve ideal process parallelization, compatible with high-performance computers. Further acceleration comes from pre-computing most information of an interaction field in a discrete rotamer library. On the accuracy front, this framework uses a self-consistent energy function that can be derived from any established force fields. Therefore, it avoids overfitting and training bias typical for knowledge-based terms.

My goal in this project was to investigate the utility of the tensorized scoring function (referred to as Damietta potential) for combinatorial protein design problem. Particularly, I applied Damietta potential to design multispecific binders against the epidermal growth factor (EGF)-like ligands. These binders are needed for broad inhibition of EGFR signaling and can serve as a new therapeutic modality for several types of cancer [106]. In Chapter 6, I present experimental results for the designed EGFR inhibitors and demonstrate how tensorized calculations can be integrated into design pipeline to efficiently identify low-energy amino acid sequences for a given structure.

4.4 Filtering of successful designs

Several challenges persist in the estimation of folding and binding energies. For instance, accurate capturing of long-range interactions, modeling of hydrogen bonds or direct evaluation of entropy remains unresolved. Therefore, additional ranking of the designed sequences with orthogonal scoring schemes could considerably increase success rates.

To this end, MD-based routines greatly complement the static scoring and can serve as a final filtering step of the design workflow [107, 108]. While conventional MD simulations of *ab initio* folding from an extended peptide are computationally intense even for very small proteins, faster techniques, such as temperature accelerated MD, have been developed to assess structural deviations from the designed coordinates [7, 109]. To precisely measure the binding energy across a protein-protein interface, ElGamacy et al. established protocols for steered MD, where external pulling forces

facilitate disintegration of a complex structure [110, 111]. In my projects, described in Chapter 5 and 6, I take advantage of both temperature-accelerated and steered MD simulations to prioritize a small number of designed candidates for experimental testing.

Unprecedented accuracy of AlphaFold2 (AF2) [112] in predicting protein structure makes DL methods a potent alternative for *in silico* design validation. A recent study showed that complementing energy-based protein binder design with AF2 filtering increases success rates nearly 10-fold [113]. It should be cautioned, though, that AF2 produces correct monomeric structures more consistently compared to correct heteromeric complexes [114, 115]. Therefore, in its current state, AF2 can be utilized to assess the likelihood of the designed sequence folding into the intended monomer structure rather than its actual binding to the target.

Chapter 5

A complementarity-based approach to *de novo* binder design generates VEGF inhibitors

Abstract

De novo design of binders capable of targeting arbitrarily selected epitopes remains a substantial challenge. Here, we present a generalizable computational strategy to design site-specific protein binders, obviating steps of extensive empirical optimization or *in vitro* screening. Our dock-and-design pipeline retrieves complementary scaffolds from a protein structure database to a given query epitope, where the scaffold is mutated to carve a binding site *de novo*. The docking step utilizes a novel fingerprint that greatly simplifies and accelerates the surface complementarity evaluation. As proof-of-concept, we designed protein inhibitors blocking the receptor-binding site of the vascular endothelial growth factor (VEGF), a key angiogenic molecule involved in the pathogenesis of diverse cancers, cardiovascular, and ophthalmic diseases. We experimentally characterized 16 designs based on scaffolds that belong to two different folds. Several designs bound VEGF with nanomolar affinity and showed VEGF-inhibiting activity *in vitro* and *in vivo*.

The full text of the manuscript is available in Appendix I.

Author contributions

Author	**Author position**	**Scientific ideas %**	**Data generation %**	**Analysis and interpretation %**	**Paper writing %**
Kateryna Maksymenko	1	25	40	35	35
Murray Coles	2		6	6	
Narges Aghaallaei	3		13	13	10
Natalia Pashkovskaia	4	5	7	15	13
Mareike Volz	5		7		
Joana Pereira	6		3	3	
Marcus D. Hartmann	7		3	6	
Ghazaleh Tabatabai	8				
Stefan Liebau	9				
Patrick Müller	10	7			
Andrei N. Lupas	11				
Julia Skokowa[1]	12	13		3	7
Mohammad ElGamacy[1]	13	50	21	19	35
Titel of paper:		A complementarity-based approach to *de novo* binder design generates VEGF inhibitors			
Status in publication process:		In revision			

[1]Corresponding author

Chapter 6

The design of functional proteins using tensorized energy calculations

Abstract

In protein design, the energy associated with a huge number of sequence-conformer perturbations has to be routinely estimated. Hence, enhancing the throughput and accuracy of these energy calculations can profoundly improve design success rates and enable tackling more complex design problems. In this work, we explore the possibility of tensorizing the energy calculations and apply them in a protein design framework. We use this framework to design enhanced proteins with anti-cancer and radio-tracing functions. Particularly, we designed multispecific binders against ligands of the epidermal growth factor receptor (EGFR), where the tested design could inhibit EGFR activity *in vitro* and *in vivo*. We also used this method to design high-affinity Cu^{2+} binders that were stable in serum and could be readily loaded with copper-64 radionuclide. The resulting molecules show superior functional properties for their respective applications and demonstrate the generalizable potential of the described protein design approach.

The publication is available in Appendix II.

Author contributions

Author	**Author position**	**Scientific ideas %**	**Data generation %**	**Analysis and interpretation %**	**Paper writing %**
Kateryna Maksymenko	1	15	60	57	45
Andreas Maurer	2		8	8	
Narges Aghaallaei	3		3.5	2.5	10
Caroline Barry	4		2	2	
Natalia Borbarán- Bravo	5		0.5	0.5	
Timo Ullrich	6		0.5	0.5	
Tjeerd M.H. Dijkstra	7	5			
Birte Hernandez Alvarez	8		0.5	0.5	
Patrick Müller	9	5			
Andrei N. Lupas	10	5			
Julia Skokowa	11	5		4	
Mohammad ElGamacy[1]	13	65	25	25	45
Titel of paper:		The design of functional proteins using tensorized energy calculations			
Status in publication process:		Published [116]			

[1]Corresponding author

Chapter 7

Discussion

Throughout my research, I tested the applicability of HECTOR docking software and Damietta engine for complex design tasks. Specifically, these frameworks were shown to facilitate successful design of protein-based binders. In the following sections, I will discuss advantages and limitations of the presented methods, as well as future directions that can build upon the current results.

7.1 Massive-scale docking with HECTOR algorithm

Advantages. Shape complementarity is a central component of the scoring functions for protein-protein docking. Conventional docking approaches extensively sample spatial arrangements of the receptor and ligand molecules to generate putative binding poses. These poses are subsequently ranked based on numerical analysis of steric matching, for example by measuring the distance between two points on opposing surfaces [42, 117]. In contrast to trial-and-error methods, HECTOR evaluates shape complementarity between unaligned surface patches from different proteins, through an analytical solution. It is achieved by using invertible fingerprints with reduced dimensionality, wherein a 3D surface patch is represented by a 2D matrix. Quantifying complementarity between the query and subject patches is thus simplified to a trivial comparison of the two corresponding matrices. This setup enables high-performance docking, which is essential for selecting binder scaffolds compatible with the target epitope.

This year, two studies reported the development of DL methods, also aiming to address the challenge of *de novo* binder design. Therefore, I would like to contrast them with HECTOR, particularly in how they approach the scaffold selection task.

The first study by Gainza et al. [2] uses surface fingerprinting method MaSIF to search for complementary structural motifs capable of engaging the target site. There are several key differences between HECTOR and MaSIF that relate to geometric mapping,

information compression, and complementarity matching aspects. For instance, while HECTOR maps are rotation- and translation-invariant, MaSIF protocol results in rotation-variant radar maps, which later should be resolved by convolution neural networks. Moreover, the geodesic polar coordinate system and curvature accounting adopted by MaSIF are much more computationally intensive when compared to the cylindrical basis that HECTOR employs. Additionally, MaSIF encodes not only the steric, but also electrostatic and hydropathy surface information that is eventually compressed onto a 1D vector, in a more lossy fashion than HECTOR. Altogether, this makes MaSIF reliant on the processing of surface descriptors with neural networks, and complicates the retrospective analysis of design successes and failures.

Another recent study by Watson et al. [65] presents a completely different solution to binder scaffold selection that circumvents the docking problem. It employs the conditional RFdiffusion model, which can directly generate a protein backbone around the hotspot residues on the target structure. This allows to create binders without resorting to the scaffold database. Although extremely promising, RFdiffusion approach has its shortcomings. This model outputs poly-glycine structures, which necessitates further full-fledged sequence design, for example using ProteinMPNN neural network [63]. Unfortunately, due to the nascency of the field, there are not enough experimentally validated structures to indicate the accuracy of generated backbones after a sequence design step. Since the quality of a scaffold significantly influences the ultimate success of a design campaign [1], it may be more advantageous to rely on the datasets of the high-resolution experimental structures, as HECTOR does. Additionally, it has been observed that the default RFdiffusion model generates mostly helical binders. However, some cases, such as quenching convex epitopes, require a greater diversity of topologies.

Limitations. Despite the numerous benefits that HECTOR docking brings to the design pipeline, there are also several limitations to consider. First, although the complementarity evaluation process itself is ultra-fast, the large-scale tessellation and fingerprinting of the entire structural database (e.g. PDB) can be computationally expensive. Nevertheless, this is justified, as the library of subject maps should be computed only once and afterwards can be used for multiple design projects. Second, representing 3D surfaces with 2D maps definitely results in a certain loss of information, which can possibly lead to false positive hit selection. Lastly, HECTOR approach perfectly suits the binder design problem, where optimizing shape complementarity between the ligand and receptor is generally rewarding. However, natural protein interfaces, even those with high affinity, often exhibit irregularities and contain immobilized solvent molecules [32]. Therefore, HECTOR cannot be confidently used to predict native protein binding sites.

Outlook. In the work presented here, binder scaffolds were retrieved from PDB database through a constrained two-vs-two map search. To enhance the specificity of the search, other matching schemes should also be tested. I anticipate that comparison between cliques of three or more maps would serve as a more stringent complementarity filter, with minimal computational overhead.

In its current implementation, HECTOR map covers a radial distance of 12 Å and consists of 30 × 30 bins, with 0.4 Å bin resolution. While increasing the map size or improving its resolution could theoretically enhance the quality of HECTOR search, it would simultaneously slow down the matrix-matrix comparison process. Thus, in future I plan to explore a tiered HECTOR scoring, where only surface patches that pass complementarity criteria at a low resolution are mapped and re-ranked again at a finer resolution.

Regarding natural protein-protein complexes, chemical complementarity significantly contributes to molecular recognition, along with steric matching. Thus, to tackle docking problem outside of the design context, it is essential to consider additional scoring terms, for example electrostatics and desolvation [85]. To this end, HECTOR mapping algorithm should be modified to separately encode information on surface shape, atomic partial charges and desolvation energies. This modification would facilitate a graded evaluation of patch compatibility based on different features and broaden HECTOR applicability to modeling natural PPIs.

Invertibility of the HECTOR fingerprints allows for estimating both local similarity, and local complementarity between the query and subject surfaces. Owing to the high-throughput nature of mapping, the software can be adopted to construct large-scale surface complementarity and similarity networks using databases of experimentally determined or predicted protein structures. This can be a stepping stone for either design of binders or migrating functional epitopes across a huge pool of structural scaffolds.

7.2 Damietta: framework for sequence design and beyond

Advantages. Damietta design engine, introduced in this work, not only accelerates the interaction energy scoring by orders-of-magnitude, but also opens room for profound improvements in accuracy. I will begin by outlining the key features of the framework that enable these advancements.

Typically, scoring functions consist of multiple energy terms. The imbalance between these components, caused for example by mixing physics-based and knowledge-based terms, leads to accuracy errors [54]. To avoid this, Damietta uses an energy function

that is readily derived from any established, self-consistent force field. The same force field is used to create the rotamer library through long MD simulations of isolated amino acids. This setup excludes overfitting and training biases, and attributes inaccuracies directly to the force field parameters, allowing improvements to be more systematic. Additionally, since Damietta's rotamer library is not knowledge-based, it can include on-demand generated rotamers for non-standard amino acids or small-molecule ligands, which are underrepresented in structure databases.

In contrast to traditional evaluation of interatomic interaction energies through excessive looping, Damietta adopts tensorized calculations that allow single instruction, multiple data processing. This framework can maximally exploit massively parallel computing architectures, such as GPUs and TPUs (i.e. tensor processing units). Moreover, most information on interaction fields is pre-computed in a discrete rotamer library, considerably reducing runtime load.

Limitations. Some of the current Damietta limitations can be attributed to approximations in the scoring function. For instance, the way of calculating the Lennard-Jones potential assumes symmetric LJ parameters for the interacting atom pairs, where a certain atom of the inbound rotamer "senses" surrounding atoms as the self-same type. Also, the framework applies isotropic charges for electrostatic energy scoring.

Damietta evaluates non-bonded interactions between the atoms of an inbound rotamer and all the "environment" atoms within a box of 22 Å×22 Å×22 Å dimensions. Thus, interactions further than the defined distance cutoff are ignored. Increasing the box size may enhance accuracy; however, this would come at the expense of calculation speed.

At present, Damietta framework is restricted to the fixed-backbone design, rendering it sensitive to input structures. Even a minor adjustment to the backbone conformation can lead to a substantial energy change, especially for the LJ potential that scales as the 12^{th} power of distance when two atoms are in close proximity [92]. A workaround to this limitation is performing short MD simulations before and after design steps. Alternatively, an ensemble of multiple conformations can be used as an input, rather than a single protein structure.

Outlook. To simplify representation of the electrostatic properties, most design software treat charges as fixed, atom-centered, and isotropic. Considering the intrinsic non-uniform distribution of charges, along with the redistribution of electron density in response to the local environment, can significantly enhance scoring accuracy [118]. Within Damietta framework, addressing this aspect can be achieved by employing polarizable force fields at the rotamer library level, at no additional run-time cost.

For calculating the LJ potential, Damietta represents information on the interaction between a multi-atom environment (1st body) and an inbound rotamer (2nd body) in a way, where 1st body encodes only 3D occupancy of its atomic positions, while the 2nd body encodes the dense, real-valued energy field (Fig. 7). To overcome the limitation of symmetric LJ parameters mentioned above, 1st body can be populated not only using values for positions of atoms, but also coefficients for rescaling LJ parameters depending on the atom type.

Further accuracy gains can be achieved by implementing a new version of rotamer library based on the isolated amino acid MD simulations at higher temperatures. Our preliminary tests show that increasing temperature provides broader rotamer distributions, which can be used in designing conformational motifs rarely sampled in nature.

Concerning current Damietta calculations in a fixed-backbone context, I assume that further developments of the framework will allow it to support motion perturbations. Potentially, the gradients of the LJ and electrostatic fields can be used to derive pre-computed vector fields that encode the force vectors associated with each rotamer. Through multiplication of the vector field with the environment tensor mapping nuclear positions and/or their partial charges, forces exerted by the rotamer on the environment can be derived. The resulting transformation matrix can be applied to the atomic coordinates in order to introduce motions into the overall structure during every design or repacking step. These developments may give rise to a completely new class of conformational samplers with a wide range of applications.

Ongoing work in our group is also dedicated to applying Damietta for multi-conditional design. As an example, the protocol can favor binding of the designed candidates to a first given target while penalizing interactions with a second one. Alternatively, it can be used to design binders with multi-specificity to different targets.

Lastly, I would like to mention the development of Damietta server. Inspired by existing web resources that offer instant access to various tools for protein analysis and design [119, 120, 121], we are currently building a design toolkit allowing everyone to run Damietta through an intuitive interface, without the need to set up the tool on a personal server.

7.3 Synergy between physics-based and machine-learning methods

Latest attempts of different research groups to design protein binders *de novo* revealed benefits and shortcomings of both physics-based and DL methods. Therefore, combination of these approaches might be an ideal solution. I would like to describe here

a potential workflow (Fig. 8) that takes advantage of the design strategies introduced in this thesis, along with recently published DL methods.

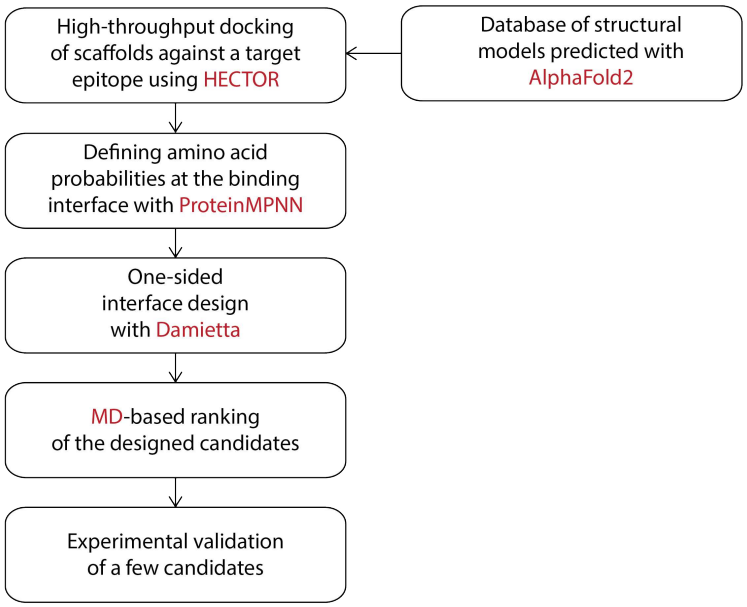


Figure 8: The design workflow integrating both physics-based and machine-learning methods.

The unprecedented computational performance of HECTOR docking algorithm allows for screening large databases. In turn, databases of structural models predicted by AF2 [112], which are orders of magnitude larger than PDB, can serve as a source of potential scaffolds [122, 123]. Since the quality of a template structure is a prerequisite for successful interface design, only a subset of high-confidence predictions should be considered. Based on a previous analysis of binder success rates, it was suggested that templates with a C_α RMSD between experimental and modeled structures of less than 0.7 Å are acceptable for further design [1]. This backbone quality can be correlated with a predicted value of the local distance difference test (pLDDT) higher than 90 for AF2 structures [124].

The efficiency of a sequence design step considerably improves by constraining the amino acid types allowed at each residue position. For this task, ProteinMPNN, a graph-based neural network [63], can be of great utility. For each mutable position, amino acids with ProteinMPNN-predicted probability higher than a certain cutoff should be included into a specification file for a combinatorial Damietta run. Given that current DL methods lack high precision at the atomic level [125], complementing ProteinMPNN choices with rigorous physics-based Damietta scoring would provide reliable solutions. To select few candidates for experimental validation, MD-based ranking described in Chapter 5 and Chapter 6 can be applied.

Chapter 8

Conclusions

Current shortcomings of *de novo* binder design relates to 1) the lack of methods for large-scale docking of potential scaffolds against the target structure, and 2) the low speed and accuracy of the design calculations. In this work, I explored novel strategies that could overcome the limitations on both fronts.

On the docking side, I showed that assessing shape complementarity of molecular surfaces in a reduced mathematical representation allows for efficient sampling of structure space. It enables fast docking of a target epitope against the large databases of protein structures. Using the HECTOR docking approach followed by interface design and MD-based filtering, I designed binders targeting receptor-binding site of VEGF. In contrast to previously reported *de novo* binder design campaigns that relied on screening thousands of candidates, I identified sufficiently tight binders from just a few tested sequences.

On the design side, I demonstrated that tensorizing energy calculations significantly accelerates the design process without compromising scoring accuracy. Retrospective validation of the Damietta software guided significant improvements in the scoring function. Meanwhile, prospective validation highlighted the ability of this framework to tackle complex design tasks. Particularly, Damietta was successfully applied for creating multi-specific binders against EGF-like ligands.

Finally, the molecules designed in this work hold therapeutic potential for anti-VEGF and anti-EGFR treatment. Moreover, I anticipate that combining HECTOR and Damietta approaches would offer a generalizable and straightforward pipeline for site-specific targeting of other application-relevant proteins with high success rates.

Bibliography

[1] L. Cao, B. Coventry, I. Goreshnik, B. Huang, W. Sheffler, J. S. Park, K. M. Jude, I. Marković, R. U. Kadam, K. H. G. Verschueren, K. Verstraete, S. T. R. Walsh, N. Bennett, A. Phal, A. Yang, L. Kozodoy, M. DeWitt, L. Picton, L. Miller, E.-M. Strauch, N. D. DeBouver, A. Pires, A. K. Bera, S. Halabiya, B. Hammerson, W. Yang, S. Bernard, L. Stewart, I. A. Wilson, H. Ruohola-Baker, J. Schlessinger, S. Lee, S. N. Savvides, K. C. Garcia, and D. Baker, "Design of protein-binding proteins from the target structure alone," *Nature*, vol. 605, pp. 551–560, Mar. 2022.

[2] P. Gainza, S. Wehrle, A. V. Hall-Beauvais, A. Marchand, A. Scheck, Z. Harteveld, S. Buckley, D. Ni, S. Tan, F. Sverrisson, C. Goverde, P. Turelli, C. Raclot, A. Teslenko, M. Pacesa, S. Rosset, S. Georgeon, J. Marsden, A. Petruzzella, K. Liu, Z. Xu, Y. Chai, P. Han, G. F. Gao, E. Oricchio, B. Fierz, D. Trono, H. Stahlberg, M. Bronstein, and B. E. Correia, "De novo design of protein interactions with learned surface fingerprints," *Nature*, vol. 617, pp. 176–184, Apr. 2023.

[3] A. Brown, "Top product forecasts for 2023," *Nature Reviews Drug Discovery*, vol. 22, p. 8, Jan. 2023.

[4] S. B. Ebrahimi and D. Samanta, "Engineering protein-based therapeutics through structural and chemical design," *Nature Communications*, vol. 14, p. 2411, Apr. 2023.

[5] B. Leader, Q. J. Baca, and D. E. Golan, "Protein therapeutics: a summary and pharmacological classification," *Nature Reviews Drug Discovery*, vol. 7, pp. 21–39, Jan. 2008.

[6] E. K. Sims, A. L. J. Carr, R. A. Oram, L. A. DiMeglio, and C. Evans-Molina, "100 years of insulin: celebrating the past, present and future of diabetes therapy," *Nature Medicine*, vol. 27, pp. 1154–1164, July 2021.

[7] J. Skokowa, B. Hernandez Alvarez, M. Coles, M. Ritter, M. Nasri, J. Haaf, N. Aghaallaei, Y. Xu, P. Mir, A.-C. Krahl, K. W. Rogers, K. Maksymenko,

B. Bajoghli, K. Welte, A. N. Lupas, P. Müller, and M. ElGamacy, "A topological refactoring design strategy yields highly stable granulopoietic proteins," *Nature Communications*, vol. 13, p. 2948, May 2022.

[8] T. Susantad, M. Fuangthong, K. Tharakaraman, P. Tit-oon, M. Ruchirawat, and R. Sasisekharan, "Modified recombinant human erythropoietin with potentially reduced immunogenicity," *Scientific Reports*, vol. 11, Jan. 2021.

[9] J. McDermott and A. Jimeno, "Pembrolizumab: PD-1 inhibition as a therapeutic strategy in cancer," *Drugs of Today*, vol. 51, no. 1, p. 7, 2015.

[10] A. Reimold, "The role of adalimumab in rheumatic and autoimmune disorders: comparison with other biologic agents," *Open Access Rheumatology: Research and Reviews*, p. 33, May 2012.

[11] M. Gebauer and A. Skerra, "Engineered protein scaffolds as next-generation therapeutics," *Annual Review of Pharmacology and Toxicology*, vol. 60, pp. 391–415, Jan. 2020.

[12] P. D. Senter and E. L. Sievers, "The discovery and development of brentuximab vedotin for use in relapsed hodgkin lymphoma and systemic anaplastic large cell lymphoma," *Nature Biotechnology*, vol. 30, pp. 631–637, July 2012.

[13] D. D. Sahtoe, A. Coscia, N. Mustafaoglu, L. M. Miller, D. Olal, I. Vulovic, T.-Y. Yu, I. Goreshnik, Y.-R. Lin, L. Clark, F. Busch, L. Stewart, V. H. Wysocki, D. E. Ingber, J. Abraham, and D. Baker, "Transferrin receptor targeting by de novo sheet extension," *Proceedings of the National Academy of Sciences*, vol. 118, Apr. 2021.

[14] J. Garousi, A. Orlova, F. Y. Frejd, and V. Tolmachev, "Imaging using radiolabelled targeted proteins: radioimmunodetection and beyond," *EJNMMI Radiopharmacy and Chemistry*, vol. 5, June 2020.

[15] W. S. Richter, V. Ivancevic, J. Meller, O. Lang, D. L. Guludec, I. Szilvazi, H. Amthauer, F. Chossat, A. Dahmane, C. Schwenke, and A. Signore, "99mTc-besilesomab (scintimun®) in peripheral osteomyelitis: comparison with 99mTc-labelled white blood cells," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 38, pp. 899–910, Feb. 2011.

[16] H. B. Breitz, A. Tyler, M. J. Bjorn, T. Lesley, and P. L. Weiden, "Clinical experience with tc-99m nofetumomab merpentan (verluma) radioimmunoscintigraphy," *Clinical Nuclear Medicine*, vol. 22, pp. 615–620, Sept. 1997.

[17] M. Sandström, K. Lindskog, I. Velikyan, A. Wennborg, J. Feldwisch, D. Sandberg, V. Tolmachev, A. Orlova, J. Sörensen, J. Carlsson, H. Lindman, and M. Lubberink, "Biodistribution and radiation dosimetry of the anti-HER2 affibody molecule 68Ga-ABY-025 in breast cancer patients," *Journal of Nuclear Medicine*, vol. 57, pp. 867–871, Feb. 2016.

[18] D. Sandberg, V. Tolmachev, I. Velikyan, H. Olofsson, A. Wennborg, J. Feldwisch, J. Carlsson, H. Lindman, and J. Sörensen, "Intra-image referencing for simplified assessment of HER2-expression in breast cancer metastases using the affibody molecule ABY-025 with PET and SPECT," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 44, pp. 1337–1346, Mar. 2017.

[19] U. Rothbauer, K. Zolghadr, S. Tillib, D. Nowak, L. Schermelleh, A. Gahl, N. Backmann, K. Conrath, S. Muyldermans, M. C. Cardoso, and H. Leonhardt, "Targeting and tracing antigens in live cells with fluorescent nanobodies," *Nature Methods*, vol. 3, pp. 887–889, Oct. 2006.

[20] S. Harmansa and M. Affolter, "Protein binders and their applications in developmental biology," *Development*, vol. 145, Jan. 2018.

[21] J. Riedl, A. H. Crevenna, K. Kessenbrock, J. H. Yu, D. Neukirchen, M. Bista, F. Bradke, D. Jenne, T. A. Holak, Z. Werb, M. Sixt, and R. Wedlich-Soldner, "Lifeact: a versatile marker to visualize f-actin," *Nature Methods*, vol. 5, pp. 605–607, June 2008.

[22] C. R. Schiavon, T. Zhang, B. Zhao, A. S. Moore, P. Wales, L. R. Andrade, M. Wu, T.-C. Sung, Y. Dayn, J. W. Feng, O. A. Quintero, G. S. Shadel, R. Grosse, and U. Manor, "Actin chromobody imaging reveals sub-organellar actin dynamics," *Nature Methods*, vol. 17, pp. 917–921, Aug. 2020.

[23] L. Kummer, C.-W. Hsu, O. Dagliyan, C. MacNevin, M. Kaufholz, B. Zimmermann, N. V. Dokholyan, K. M. Hahn, and A. Plückthun, "Knowledge-based design of a biosensor to quantify localized ERK activation in living cells," *Chemistry & Biology*, vol. 20, pp. 847–856, June 2013.

[24] A. L. Marschall, S. Dübel, and T. Böldicke, "Specific in vivo knockdown of protein function by intrabodies," *mAbs*, vol. 7, pp. 1010–1035, Aug. 2015.

[25] F. Sha, E. B. Gencer, S. Georgeon, A. Koide, N. Yasui, S. Koide, and O. Hantschel, "Dissection of the BCR-ABL signaling network using highly specific monobody inhibitors to the SHP2 SH2 domains," *Proceedings of the National Academy of Sciences*, vol. 110, pp. 14924–14929, Aug. 2013.

[26] N. Nady, A. Gupta, Z. Ma, T. Swigut, A. Koide, S. Koide, and J. Wysocka, "ETO family protein mtgr1 mediates prdm14 functions in stem cell maintenance and primordial germ cell formation," *eLife*, vol. 4, Nov. 2015.

[27] W. Heu, J.-M. Choi, J.-J. Lee, S. Jeong, and H.-S. Kim, "Protein binder for affinity purification of human immunoglobulin antibodies," *Analytical Chemistry*, vol. 86, pp. 6019–6025, May 2014.

[28] B. Mouratou and F. Pecorari, "Application of affitins for affinity purification of proteins," in *Methods in Molecular Biology*, pp. 37–48, Springer US, 2022.

[29] G. Sennhauser and M. G. Grütter, "Chaperone-assisted crystallography with DARPins," *Structure*, vol. 16, pp. 1443–1453, Oct. 2008.

[30] M. A. Bukowska and M. G. Grütter, "New concepts and aids to facilitate crystallization," *Current Opinion in Structural Biology*, vol. 23, pp. 409–416, June 2013.

[31] A. Chevrel, A. Mesneau, D. Sanchez, L. Celma, S. Quevillon-Cheruel, A. Cavagnino, S. Nessler, I. L. de la Sierra-Gallay, H. van Tilbeurgh, P. Minard, M. Valerio-Lepiniec, and A. Urvoas, "Alpha repeat proteins as expression and crystallization helpers," *Journal of Structural Biology*, vol. 201, pp. 88–99, Feb. 2018.

[32] A. V. Veselovsky, Y. D. Ivanov, A. S. Ivanov, A. I. Archakov, P. Lewi, and P. Janssen, "Protein-protein interactions: mechanisms and modification by drugs," *Journal of Molecular Recognition*, vol. 15, no. 6, pp. 405–422, 2002.

[33] W. E. Stites, "Protein-protein interactions: Interface structure, binding thermodynamics, and mutational analysis," *Chemical Reviews*, vol. 97, pp. 1233–1250, Aug. 1997.

[34] J. Wang, Z. Szewczuk, S.-Y. Yue, Y. Tsuda, Y. Konishi, and E. O. Purisima, "Calculation of relative binding free energies and configurational entropies: A structural and thermodynamic analysis of the nature of non-polar binding of thrombin inhibitors based on hirudin55–65," *Journal of Molecular Biology*, vol. 253, pp. 473–492, Oct. 1995.

[35] G. P. Brady and K. A. Sharp, "Entropy in protein folding and in protein—protein interactions," *Current Opinion in Structural Biology*, vol. 7, pp. 215–221, Apr. 1997.

[36] V. J. Hilser, B. G.-M. E., T. G. Oas, G. Kapp, and S. T. Whitten, "A statistical thermodynamic model of the protein ensemble," *Chemical Reviews*, vol. 106, pp. 1545–1558, Mar. 2006.

[37] J. A. Caro, K. W. Harpole, V. Kasinath, J. Lim, J. Granja, K. G. Valentine, K. A. Sharp, and A. J. Wand, "Entropy in molecular recognition by proteins," *Proceedings of the National Academy of Sciences*, vol. 114, pp. 6563–6568, June 2017.

[38] J. D. Chodera and D. L. Mobley, "Entropy-enthalpy compensation: Role and ramifications in biomolecular ligand recognition and design," *Annual Review of Biophysics*, vol. 42, pp. 121–142, May 2013.

[39] J. M. Fox, M. Zhao, M. J. Fink, K. Kang, and G. M. Whitesides, "The molecular origin of enthalpy/entropy compensation in biomolecular recognition," *Annual Review of Biophysics*, vol. 47, pp. 223–250, May 2018.

[40] C. A. Blasie and J. M. Berg, "Entropy-enthalpy compensation in ionic interactions probed in a zinc finger peptide," *Biochemistry*, vol. 43, pp. 10600–10604, July 2004.

[41] C. Chothia and J. Janin, "Principles of protein-protein recognition," *Nature*, vol. 256, pp. 705–708, Aug. 1975.

[42] M. C. Lawrence and P. M. Colman, "Shape complementarity at protein/protein interfaces," *Journal of Molecular Biology*, vol. 234, pp. 946–950, Dec. 1993.

[43] F. Desantis, M. Miotto, L. D. Rienzo, E. Milanetti, and G. Ruocco, "Spatial organization of hydrophobic and charged residues affects protein thermal stability and binding affinity," *Scientific Reports*, vol. 12, July 2022.

[44] Y. Li, X. Zhang, and D. Cao, "The role of shape complementarity in the protein-protein interactions," *Scientific Reports*, vol. 3, Nov. 2013.

[45] F. Gao, J. Glaser, and S. C. Glotzer, "The role of complementary shape in protein dimerization," *Soft Matter*, vol. 17, no. 31, pp. 7376–7383, 2021.

[46] D. Reichmann, Y. Phillip, A. Carmi, and G. Schreiber, "On the contribution of water-mediated interactions to protein-complex stability," *Biochemistry*, vol. 47, pp. 1051–1060, Dec. 2007.

[47] M. H. Ahmed, F. Spyrakis, P. Cozzini, P. K. Tripathi, A. Mozzarelli, J. N. Scarsdale, M. A. Safo, and G. E. Kellogg, "Bound water at protein-protein

interfaces: Partners, roles and hydrophobic bubbles as a conserved motif," *PLoS ONE*, vol. 6, p. e24712, Sept. 2011.

[48] I. V. Korendovych and W. F. DeGrado, "De novo protein design, a retrospective," *Quarterly Reviews of Biophysics*, vol. 53, 2020.

[49] L. Sellés Vidal, M. Isalan, J. T. Heap, and R. Ledesma-Amaro, "A primer to directed evolution: current methodologies and future directions," *RSC Chemical Biology*, vol. 4, pp. 271–291, Apr. 2023.

[50] J. Hanes and A. Plückthun, "In vitro selection and evolution of functional proteins by using ribosome display," *Proceedings of the National Academy of Sciences*, vol. 94, pp. 4937–4942, May 1997.

[51] D. S. Wilson, A. D. Keefe, and J. W. Szostak, "The use of mRNA display to select high-affinity protein-binding peptides," *Proceedings of the National Academy of Sciences*, vol. 98, pp. 3750–3755, Mar. 2001.

[52] R. Odegrip, D. Coomber, B. Eldridge, R. Hederer, P. A. Kuhlman, C. Ullman, K. FitzGerald, and D. McGregor, "CIS display: In vitro selection of peptides from libraries of protein–DNA complexes," *Proceedings of the National Academy of Sciences*, vol. 101, pp. 2806–2810, Feb. 2004.

[53] D. N. Woolfson, "A brief history of de novo protein design: Minimal, rational, and computational," *Journal of Molecular Biology*, vol. 433, p. 167160, Oct. 2021.

[54] M. ElGamacy, "Accelerating therapeutic protein design," in *Protein Design and Structure*, pp. 85–118, Elsevier, 2022.

[55] J. K. Leman and et al., "Macromolecular modeling and design in rosetta: recent methods and frameworks," *Nature Methods*, vol. 17, pp. 665–680, June 2020.

[56] A. Glasgow, J. Glasgow, D. Limonta, P. Solomon, I. Lui, Y. Zhang, M. A. Nix, N. J. Rettko, S. Zha, R. Yamin, K. Kao, O. S. Rosenberg, J. V. Ravetch, A. P. Wiita, K. K. Leung, S. A. Lim, X. X. Zhou, T. C. Hobman, T. Kortemme, and J. A. Wells, "Engineered ACE2 receptor traps potently neutralize SARS-CoV-2," *Proceedings of the National Academy of Sciences*, vol. 117, pp. 28046–28055, Oct. 2020.

[57] M. A. Hallen, J. W. Martin, A. Ojewole, J. D. Jou, A. U. Lowegard, M. S. Frenkel, P. Gainza, H. M. Nisonoff, A. Mukund, S. Wang, G. T. Holt, D. Zhou, E. Dowd, and B. R. Donald, "OSPREY 3.0: Open-source protein redesign for you, with powerful new features," *Journal of Computational Chemistry*, vol. 39, pp. 2494–2507, Oct. 2018.

[58] A. U. Lowegard, M. S. Frenkel, G. T. Holt, J. D. Jou, A. A. Ojewole, and B. R. Donald, "Novel, provable algorithms for efficient ensemble-based computational protein design and their application to the redesign of the c-raf-RBD:KRas protein-protein interface," *PLOS Computational Biology*, vol. 16, p. e1007447, June 2020.

[59] E. Procko, G. Y. Berguig, B. W. Shen, Y. Song, S. Frayo, A. J. Convertine, D. Margineantu, G. Booth, B. E. Correia, Y. Cheng, W. R. Schief, D. M. Hockenbery, O. W. Press, B. L. Stoddard, P. S. Stayton, and D. Baker, "A computationally designed inhibitor of an epstein-barr viral bcl-2 protein induces apoptosis in infected cells," *Cell*, vol. 157, pp. 1644–1656, June 2014.

[60] S. Berger, E. Procko, D. Margineantu, E. F. Lee, B. W. Shen, A. Zelter, D.-A. Silva, K. Chawla, M. J. Herold, J.-M. Garnier, R. Johnson, M. J. MacCoss, G. Lessene, T. N. Davis, P. S. Stayton, B. L. Stoddard, W. D. Fairlie, D. M. Hockenbery, and D. Baker, "Computationally designed high specificity inhibitors delineate the roles of BCL2 family proteins in cancer," *eLife*, vol. 5, Nov. 2016.

[61] C. M. Bryan, G. J. Rocklin, M. J. Bick, A. Ford, S. Majri-Morrison, A. V. Kroll, C. J. Miller, L. Carter, I. Goreshnik, A. Kang, F. DiMaio, K. V. Tarbell, and D. Baker, "Computational design of a synthetic PD-1 agonist," *Proceedings of the National Academy of Sciences*, vol. 118, July 2021.

[62] C. Yang, F. Sesterhenn, J. Bonet, E. A. van Aalen, L. Scheller, L. A. Abriata, J. T. Cramer, X. Wen, S. Rosset, S. Georgeon, T. Jardetzky, T. Krey, M. Fussenegger, M. Merkx, and B. E. Correia, "Bottom-up de novo design of functional proteins with complex structural features," *Nature Chemical Biology*, vol. 17, pp. 492–500, Jan. 2021.

[63] J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker, "Robust deep learning–based protein sequence design using ProteinMPNN," *Science*, vol. 378, pp. 49–56, Oct. 2022.

[64] J. Wang, S. Lisanza, D. Juergens, D. Tischer, J. L. Watson, K. M. Castro, R. Ragotte, A. Saragovi, L. F. Milles, M. Baek, I. Anishchenko, W. Yang, D. R. Hicks, M. Expòsit, T. Schlichthaerle, J.-H. Chun, J. Dauparas, N. Bennett, B. I. M. Wicky, A. Muenks, F. DiMaio, B. Correia, S. Ovchinnikov, and D. Baker, "Scaffolding protein functional sites using deep learning," *Science*, vol. 377, pp. 387–394, July 2022.

[65] J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. V. Torres, A. Lauko, V. D. Bortoli, E. Mathieu, S. Ovchinnikov, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek, and D. Baker, "De novo design of protein structure and function with RFdiffusion," *Nature*, vol. 620, pp. 1089–1100, July 2023.

[66] A. A. Bogan and K. S. Thorn, "Anatomy of hot spots in protein interfaces," *Journal of Molecular Biology*, vol. 280, pp. 1–9, July 1998.

[67] S. J. Fleishman, J. E. Corn, E.-M. Strauch, T. A. Whitehead, J. Karanicolas, and D. Baker, "Hotspot-centric de novo design of protein binders," *Journal of Molecular Biology*, vol. 413, pp. 1047–1062, Nov. 2011.

[68] S. J. Fleishman, T. A. Whitehead, D. C. Ekiert, C. Dreyfus, J. E. Corn, E.-M. Strauch, I. A. Wilson, and D. Baker, "Computational design of proteins targeting the conserved stem region of influenza hemagglutinin," *Science*, vol. 332, pp. 816–821, May 2011.

[69] L. Cao, I. Goreshnik, B. Coventry, J. B. Case, L. Miller, L. Kozodoy, R. E. Chen, L. Carter, A. C. Walls, Y.-J. Park, E.-M. Strauch, L. Stewart, M. S. Diamond, D. Veesler, and D. Baker, "De novo design of picomolar SARS-CoV-2 miniprotein inhibitors," *Science*, vol. 370, pp. 426–431, Oct. 2020.

[70] J. Dou, A. A. Vorobieva, W. Sheffler, L. A. Doyle, H. Park, M. J. Bick, B. Mao, G. W. Foight, M. Y. Lee, L. A. Gagnon, L. Carter, B. Sankaran, S. Ovchinnikov, E. Marcos, P.-S. Huang, J. C. Vaughan, B. L. Stoddard, and D. Baker, "De novo design of a fluorescence-activating beta-barrel," *Nature*, vol. 561, pp. 485–491, Sept. 2018.

[71] P. Gainza, F. Sverrisson, F. Monti, E. Rodola, D. Boscaini, M. M. Bronstein, and B. E. Correia, "Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning," *Nature Methods*, vol. 17, pp. 184–192, Dec. 2019.

[72] R. K. Jha, A. Leaver-Fay, S. Yin, Y. Wu, G. L. Butterfoss, T. Szyperski, N. V. Dokholyan, and B. Kuhlman, "Computational design of a PAK1 binding protein," *Journal of Molecular Biology*, vol. 400, pp. 257–270, July 2010.

[73] A. Marchand, A. K. Van Hall-Beauvais, and B. E. Correia, "Computational design of novel protein–protein interactions – an overview on methodological

approaches and applications," *Current Opinion in Structural Biology*, vol. 74, p. 102370, June 2022.

[74] L. A. Rabia, A. A. Desai, H. S. Jhajj, and P. M. Tessier, "Understanding and overcoming trade-offs between antibody affinity, specificity, stability and solubility," *Biochemical Engineering Journal*, vol. 137, pp. 365–374, Sept. 2018.

[75] E. K. Makowski, P. C. Kinnunen, J. Huang, L. Wu, M. D. Smith, T. Wang, A. A. Desai, C. N. Streu, Y. Zhang, J. M. Zupancic, J. S. Schardt, J. J. Linderman, and P. M. Tessier, "Co-optimization of therapeutic antibody affinity and specificity using machine learning models that generalize to novel mutational space," *Nature Communications*, vol. 13, July 2022.

[76] C. D. Waldburger, J. F. Schildbach, and R. T. Sauer, "Are buried salt bridges important for protein stability and conformational specificity?," *Nature Structural Molecular Biology*, vol. 2, pp. 122–128, Feb. 1995.

[77] D. Kuroda and J. J. Gray, "Shape complementarity and hydrogen bond preferences in protein-protein interfaces: implications for antibody modeling and protein–protein docking," *Bioinformatics*, vol. 32, pp. 2451–2456, Apr. 2016.

[78] E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A. A. Friesem, C. Aflalo, and I. A. Vakser, "Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques.," *Proceedings of the National Academy of Sciences*, vol. 89, p. 2195–2199, Mar. 1992.

[79] R. Chen and Z. Weng, "Docking unbound proteins using shape complementarity, desolvation, and electrostatics," *Proteins: Structure, Function, and Bioinformatics*, vol. 47, p. 281–294, Mar. 2002.

[80] D. W. Ritchie and G. J. Kemp, "Protein docking using spherical polar fourier correlations," *Proteins: Structure, Function, and Genetics*, vol. 39, p. 178–194, May 2000.

[81] D. Duhovny, R. Nussinov, and H. J. Wolfson, *Efficient Unbound Docking of Rigid Molecules*, p. 185–200. Springer Berlin Heidelberg, 2002.

[82] D. Schneidman-Duhovny, Y. Inbar, R. Nussinov, and H. J. Wolfson, "Patchdock and symmdock: servers for rigid and symmetric docking," *Nucleic Acids Research*, vol. 33, p. W363–W367, July 2005.

[83] M. Zacharias, "Protein–protein docking with a reduced protein model accounting for side-chain flexibility," *Protein Science*, vol. 12, p. 1271–1282, June 2003.

[84] S. Chaudhury, M. Berrondo, B. D. Weitzner, P. Muthu, H. Bergman, and J. J. Gray, "Benchmarking and analysis of protein docking performance in rosetta v3.2," *PLoS ONE*, vol. 6, p. e22477, Aug. 2011.

[85] S.-Y. Huang, "Exploring the potential of global protein–protein docking: an overview and critical assessment of current programs for automatic ab initio docking," *Drug Discovery Today*, vol. 20, p. 969–977, Aug. 2015.

[86] S. Yin, E. A. Proctor, A. A. Lugovskoy, and N. V. Dokholyan, "Fast screening of protein surfaces using geometric invariant fingerprints," *Proceedings of the National Academy of Sciences*, vol. 106, pp. 16622–16626, Sept. 2009.

[87] L. Sael, B. Li, D. La, Y. Fang, K. Ramani, R. Rustamov, and D. Kihara, "Fast protein tertiary structure retrieval based on global surface shape similarity," *Proteins: Structure, Function, and Bioinformatics*, vol. 72, p. 1259–1273, Mar. 2008.

[88] L. Sael and D. Kihara, "Improved protein surface comparison and application to low-resolution protein structure data," *BMC Bioinformatics*, vol. 11, Dec. 2010.

[89] V. Venkatraman, Y. D. Yang, L. Sael, and D. Kihara, "Protein-protein docking using region-based 3d zernike descriptors," *BMC Bioinformatics*, vol. 10, Dec. 2009.

[90] D. Kihara, L. Sael, R. Chikhi, and J. Esquivel-Rodriguez, "Molecular surface representation using 3d zernike descriptors for protein shape comparison and docking," *Current Protein and Peptide Science*, vol. 12, p. 520–530, Sept. 2011.

[91] R. S. Apte, D. S. Chen, and N. Ferrara, "Vegf in signaling and disease: Beyond discovery and development," *Cell*, vol. 176, p. 1248–1264, Mar. 2019.

[92] X. Pan and T. Kortemme, "Recent advances in de novo protein design: Principles, methods, and applications," *Journal of Biological Chemistry*, vol. 296, p. 100558, Jan. 2021.

[93] R. Chanhrasekaran and G. N. Ramachandran, "Studies on the conformation of amino acids: Xi. analysis of the observed side group conformations in proteins1," *International Journal of Protein Research*, vol. 2, p. 223–233, Dec. 1970.

[94] R. L. Dunbrack, "Rotamer libraries in the 21st century," *Current Opinion in Structural Biology*, vol. 12, p. 431–440, Aug. 2002.

[95] J. Desmet, M. D. Maeyer, B. Hazes, and I. Lasters, "The dead-end elimination theorem and its use in protein side-chain positioning," *Nature*, vol. 356, p. 539–542, Apr. 1992.

[96] D. B. Gordon and S. L. Mayo, "Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem," *Journal of Computational Chemistry*, vol. 19, p. 1505–1514, Oct. 1998.

[97] B. Kuhlman and P. Bradley, "Advances in protein structure prediction and design," *Nature Reviews Molecular Cell Biology*, vol. 20, p. 681–697, Aug. 2019.

[98] B. Kuhlman and D. Baker, "Native protein sequences are close to optimal for their structures," *Proceedings of the National Academy of Sciences*, vol. 97, p. 10383–10388, Sept. 2000.

[99] A. Leaver-Fay, M. Tyka, S. M. Lewis, O. F. Lange, J. Thompson, R. Jacak, K. W. Kaufman, P. D. Renfrew, C. A. Smith, W. Sheffler, I. W. Davis, S. Cooper, A. Treuille, D. J. Mandell, F. Richter, Y.-E. A. Ban, S. J. Fleishman, J. E. Corn, D. E. Kim, S. Lyskov, M. Berrondo, S. Mentzer, Z. Popović, J. J. Havranek, J. Karanicolas, R. Das, J. Meiler, T. Kortemme, J. J. Gray, B. Kuhlman, D. Baker, and P. Bradley, *Rosetta3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules*, p. 545–574. Elsevier, 2011.

[100] M. Levitt and S. Lifson, "Refinement of protein conformations using a macromolecular energy minimization procedure," *Journal of Molecular Biology*, vol. 46, p. 269–279, Dec. 1969.

[101] J. J. Gray, S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C. A. Rohl, and D. Baker, "Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations," *Journal of Molecular Biology*, vol. 331, p. 281–299, Aug. 2003.

[102] P. B. Stranges and B. Kuhlman, "A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds," *Protein Science*, vol. 22, p. 74–82, Nov. 2012.

[103] T. Kortemme, A. V. Morozov, and D. Baker, "An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein–protein complexes," *Journal of Molecular Biology*, vol. 326, p. 1239–1259, Feb. 2003.

[104] A. V. Morozov, T. Kortemme, K. Tsemekhman, and D. Baker, "Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations," *Proceedings of the National Academy of Sciences*, vol. 101, p. 6946–6951, Apr. 2004.

[105] N. Anand, R. Eguchi, I. I. Mathews, C. P. Perez, A. Derry, R. B. Altman, and P.-S. Huang, "Protein sequence design with a learned potential," *Nature Communications*, vol. 13, Feb. 2022.

[106] S. Guardiola, M. Varese, M. Sánchez-Navarro, and E. Giralt, "A third shot at egfr: New opportunities in cancer therapy," *Trends in Pharmacological Sciences*, vol. 40, p. 941–955, Dec. 2019.

[107] A. Chevalier, D.-A. Silva, G. J. Rocklin, D. R. Hicks, R. Vergara, P. Murapa, S. M. Bernard, L. Zhang, K.-H. Lam, G. Yao, C. D. Bahl, S.-I. Miyashita, I. Goreshnik, J. T. Fuller, M. T. Koday, C. M. Jenkins, T. Colvin, L. Carter, A. Bohn, C. M. Bryan, D. A. Fernández-Velasco, L. Stewart, M. Dong, X. Huang, R. Jin, I. A. Wilson, D. H. Fuller, and D. Baker, "Massively parallel de novo protein design for targeted therapeutics," *Nature*, vol. 550, p. 74–79, Sept. 2017.

[108] M. C. Childers and V. Daggett, "Insights from molecular dynamics simulations for computational protein design," *Molecular Systems Design and Engineering*, vol. 2, no. 1, p. 9–33, 2017.

[109] B. Hernandez Alvarez, J. Skokowa, M. Coles, P. Mir, M. Nasri, K. Maksymenko, L. Weidmann, K. W. Rogers, K. Welte, A. N. Lupas, P. Müller, and M. ElGamacy, "Design of novel granulopoietic proteins by topological rescaffolding," *PLOS Biology*, vol. 18, p. e3000919, Dec. 2020.

[110] M. ElGamacy, M. Coles, and A. Lupas, "Asymmetric protein design from conserved supersecondary structures," *Journal of Structural Biology*, vol. 204, p. 380–387, Dec. 2018.

[111] M. ElGamacy, M. Coles, P. Ernst, H. Zhu, M. D. Hartmann, A. Plückthun, and A. N. Lupas, "An interface-driven design strategy yields a novel, corrugated protein architecture," *ACS Synthetic Biology*, vol. 7, p. 2226–2235, Aug. 2018.

[112] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, p. 583–589, July 2021.

[113] N. R. Bennett, B. Coventry, I. Goreshnik, B. Huang, A. Allen, D. Vafeados, Y. P. Peng, J. Dauparas, M. Baek, L. Stewart, F. DiMaio, S. De Munck, S. N. Savvides,

and D. Baker, "Improving de novo protein binder design with deep learning," *Nature Communications*, vol. 14, May 2023.

[114] R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper, and D. Hassabis, "Protein complex prediction with alphafold-multimer," Oct. 2021.

[115] W. Zhu, A. Shenoy, P. Kundrotas, and A. Elofsson, "Evaluation of alphafold-multimer prediction on multi-chain protein complexes," *Bioinformatics*, vol. 39, July 2023.

[116] K. Maksymenko, A. Maurer, N. Aghaallaei, C. Barry, N. Borbaran-Bravo, T. Ullrich, T. M. Dijkstra, B. H. Alvarez, P. Müller, A. N. Lupas, J. Skokowa, and M. ElGamacy, "The design of functional proteins using tensorized energy calculations," *Cell Reports Methods*, vol. 3, p. 100560, Aug. 2023.

[117] Y. Yan and S.-Y. Huang, "Pushing the accuracy limit of shape complementarity for protein-protein docking," *BMC Bioinformatics*, vol. 20, Dec. 2019.

[118] G. A. Cisneros, M. Karttunen, P. Ren, and C. Sagui, "Classical electrostatics for biomolecular simulations," *Chemical Reviews*, vol. 114, p. 779–814, Aug. 2013.

[119] L. Zimmermann, A. Stephens, S.-Z. Nam, D. Rau, J. Kübler, M. Lozajic, F. Gabler, J. Söding, A. N. Lupas, and V. Alva, "A completely reimplemented mpi bioinformatics toolkit with a new hhpred server at its core," *Journal of Molecular Biology*, vol. 430, p. 2237–2243, July 2018.

[120] J. P. Kynast and B. Höcker, "Atligator web: A graphical user interface for analysis and design of protein–peptide interactions," *BioDesign Research*, vol. 5, Jan. 2023.

[121] N. Ferruz, S. Schmidt, and B. Höcker, "Proteintools: a toolkit to analyze protein structures," *Nucleic Acids Research*, vol. 49, p. W559–W566, May 2021.

[122] K. Tunyasuvunakool, J. Adler, Z. Wu, T. Green, M. Zielinski, A. Žídek, A. Bridgland, A. Cowie, C. Meyer, A. Laydon, S. Velankar, G. J. Kleywegt, A. Bateman, R. Evans, A. Pritzel, M. Figurnov, O. Ronneberger, R. Bates, S. A. A. Kohl, A. Potapenko, A. J. Ballard, B. Romera-Paredes, S. Nikolov, R. Jain, E. Clancy, D. Reiman, S. Petersen, A. W. Senior, K. Kavukcuoglu, E. Birney, P. Kohli, J. Jumper, and D. Hassabis, "Highly accurate protein

structure prediction for the human proteome," *Nature*, vol. 596, p. 590–596, July 2021.

[123] I. R. Humphreys, J. Pei, M. Baek, A. Krishnakumar, I. Anishchenko, S. Ovchinnikov, J. Zhang, T. J. Ness, S. Banjade, S. R. Bagde, V. G. Stancheva, X.-H. Li, K. Liu, Z. Zheng, D. J. Barrero, U. Roy, J. Kuper, I. S. Fernández, B. Szakal, D. Branzei, J. Rizo, C. Kisker, E. C. Greene, S. Biggins, S. Keeney, E. A. Miller, J. C. Fromme, T. L. Hendrickson, Q. Cong, and D. Baker, "Computed structures of core eukaryotic protein complexes," *Science*, vol. 374, Dec. 2021.

[124] T. C. Terwilliger, D. Liebschner, T. I. Croll, C. J. Williams, A. J. McCoy, B. K. Poon, P. V. Afonine, R. D. Oeffner, J. S. Richardson, R. J. Read, and P. D. Adams, "Alphafold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination," *Nature Methods*, Nov. 2023.

[125] H. Khakzad, I. Igashov, A. Schneuing, C. Goverde, M. Bronstein, and B. Correia, "A new age in protein design empowered by deep learning," *Cell Systems*, vol. 14, p. 925–939, Nov. 2023.

Appendix

I A complementarity-based approach to *de novo* binder design generates VEGF inhibitors

II The design of functional proteins using tensorized energy calculations

A complementarity-based approach to *de novo* binder design generates VEGF inhibitors

Kateryna Maksymenko[1,2], Murray Coles[1], Narges Aghaallaei[3,4], Natalia Pashkovskaia[5], Mareike Volz[5], Joana Pereira[1,6], Marcus D. Hartmann[1,7], Ghazaleh Tabatabai[8], Stefan Liebau[5], Patrick Müller[2,9], Andrei N. Lupas[1], Julia Skokowa[3,†], Mohammad ElGamacy[1,2,3,†]

[1] Max Planck Institute for Biology, Department of Protein Evolution, 72076, Tübingen (Germany)
[2] Friedrich Miescher Laboratory of the Max Planck Society, 72076, Tübingen (Germany)
[3] University Hospital Tübingen, Division of Translational Oncology, 72076, Tübingen (Germany)
[4] Present address: Medical University of Vienna, Ludwig Boltzmann Institute for Hematology and Oncology, 1090, Vienna (Austria)
[5] Eberhard Karls University Tübingen, Institute of Neuroanatomy and Developmental Biology, 72074, Tübingen (Germany)
[6] Present address: University of Basel, Biozentrum and Swiss Institute of Bioinformatics, 4056, Basel (Switzerland)
[7] Eberhard Karls University Tübingen, Interfaculty Institute of Biochemistry, 72076, Tübingen (Germany)
[8] University Hospital Tübingen, Department of Neurology and Interdisciplinary Neuro-Oncology and Hertie Institute for Clinical Brain Research, 72076, Tübingen
[9] Present address: University of Konstanz, Department of Biology, 78464 Konstanz (Germany)
[†]Corresponding author

Abstract

De novo design of binders capable of targeting arbitrarily selected epitopes remains a substantial challenge. Here, we present a generalizable computational strategy to design site-specific protein binders, obviating steps of extensive empirical optimization or *in vitro* screening. Our dock-and-design pipeline retrieves complementary scaffolds from a protein structure database to a given query epitope, where the scaffold is mutated to carve a binding site *de novo*. The docking step utilizes a novel fingerprint that greatly simplifies and accelerates the surface complementarity evaluation. As proof-of-concept, we designed protein inhibitors blocking the receptor-binding site of the vascular endothelial growth factor (VEGF), a key angiogenic molecule involved in the pathogenesis of diverse cancers, cardiovascular, and ophthalmic diseases. We experimentally characterized 16 designs based on scaffolds that belong to two different folds. Several designs bound VEGF with nanomolar affinity and showed VEGF-inhibiting activity *in vitro* and *in vivo*.

Introduction

Protein-protein interactions play a central role in all biological processes, thus the ability to design and manipulate such interactions is of immense value for basic and applied research. Particularly, the design of on-demand binders against a predefined target epitope greatly broadens the range of accessible therapeutic applications [1]. Such epitope targeting, in addition to the control over the binder's shape and biophysical properties, can be achieved through computational protein design [2]. Previous studies have demonstrated successful protein binder design based on information derived from natural complex structures, where key interaction residues were incorporated into new, geometrically-accommodating scaffolds [3-6]. Alternatively, *de novo* backbones have been built around natural binding motifs [7, 8]. Although these template-based approaches are rapidly advancing, they cover only a narrow range of targetable surfaces and are not applicable for novel interface design. Meanwhile,

template-free binder design – i.e., without using structural information from known binders – remains an outstanding challenge [9]. Recent studies contributed towards *de novo* binder design. Cao *et al.* applied the rotamer interaction field (RIF) docking method to generate binders against SARS-CoV-2 spike protein [10], and twelve other diverse target proteins [11]. An alternative geometric deep learning tool, MaSIF-seed, has been proposed recently to identify structural motifs able to engage specific epitopes based on geometrical and chemical complementarity. This tool enabled design of protein binders for three different therapeutically relevant targets - SARS-CoV-2 spike, PD-1, and PD-L1 [12]. However, all these approaches suffer from low success rates (below 1 %), and thus have to rely on the screening of thousands of designs. This motivates the development and evaluation of further strategies for *de novo* design of protein binders.

An ideal binder must possess maximum binding affinity and specificity, while being stable in the presence and absence of its target. This is a formidable challenge since improving one of these properties often comes at the cost of the other two [13]. Moreover, binder design requires the simultaneous optimization of shape complementarity, polar and hydrophobic interactions at the interface, and the overall solvation energy of the bound and free forms. In this work, we explore a tiered approach that first prioritizes the maximization of shape complementarity, then follows to optimize and refine the other designable parameters. The first step of identifying complementary binding scaffolds is achieved through high-throughput docking of a database of protein structures against the target epitope. To this end, we devised an ultra-fast algorithm (Highly Efficient Complementarity Testing by Obverse Residuals, HECTOR), which enables the evaluation of surface complementarity through a vectorized, lower dimensional surface representation. The second step involves the sequence design and *in silico*-filtering of the binding candidates (Fig. 1).

To test the utility of our pipeline, we sought to design binders targeting the receptor-binding site of vascular endothelial growth factor (VEGF), with the aim of inhibiting its activity. VEGF is a key regulator of normal and pathological angiogenesis, and is a validated target for anti-cancer, cardiovascular, and ophthalmic therapies [14, 15]. VEGF forms a 2:2 complex with the ectodomain of its receptor (VEGFR), which triggers activation of the intracellular receptor tyrosine kinase domain and the subsequent signaling cascades [16] (Fig. 2A). Current anti-VEGF therapies include both small molecules and protein-based inhibitors. Small molecules act by inhibiting the VEGFR intracellular kinase domain, however, they are less specific and more toxic than the available protein-based therapies [17]. On the other hand, protein-based inhibitors, which block the VEGF:VEGFR interaction, are large, complex molecules that are post-translationally modified [18, 19]. Therefore, we aimed to design inhibitors based on simple, single-domain scaffolds. Having tested only 16 candidates, we identified proteins that showed strong VEGF binding, anti-VEGF activity in cell-based and *in vivo* assays, highlighting their therapeutic potential.

Results

Strategy for de novo design of epitope-targeted binders

The first step of our pipeline aims at identifying binder scaffolds with highly complementary surface patches to the query epitope. Docking a query epitope against a large database of template structures is, however, a computationally demanding task, particularly given the rotational and translational degrees of freedom associated with the docking problem. In order to achieve ultra-fast steric docking evaluation for two (query and subject) surface patches, we derived a surface fingerprint. The HECTOR fingerprint combines the two advantageous

properties of *invertibility* and *compression*. Through the first property, a fingerprint describing a query surface patch can be *inverted* through a single transformation to describe the ideally complementary surface patch (Fig. S1A). Consequently, maximizing the similarity of the *inverted* query fingerprint to a subject fingerprint becomes a simple and straightforward optimization, that effectively maximizes the complementarity between the underlying query and subject surfaces (Fig. S1B). The second property relates to the lower dimensional projection of a three-dimensional surface patch by a two-dimensional matrix. While this compression is lossy, it is necessary to achieve rotational invariance of a fingerprint (Materials and methods).

In this layout, a protein structure database is used as a source of scaffolds. Specifically, here we used a snapshot of high-resolution X-ray structures from the Protein Data Bank (PDB). The dot-surfaces of all-atom representations of these structures are derived (adding the coordinates of missing atoms) and stored. Overlapping surface patches are then extracted from these dot-surfaces, which are then forward-mapped to yield a database of fingerprints. While this large-scale mapping tessellation and fingerprinting can be computationally expensive, it is performed only once. A query epitope (one or more) undergoes the same procedure, however, the z-coordinate is obversely scaled (i.e. $k = -1$) to obtain the inverse fingerprint (Materials and methods). The complementarity between the query and subject surface patches is evaluated as an R-factor that quantifies the dissimilarity between the two corresponding matrices (i.e. fingerprints). Therefore, lower R-factor values indicate higher complementarity (Fig. S2). Particularly, identifying clusters of query and subject surface patches that are spatially proximal can be used to identify putative binding sites (Fig. S3, S4). The R-factor calculation could be efficiently implemented in a vectorized form on graphics processing units (GPUs), where a single evaluation fell within sub-microsecond timeframe. This allowed us to carry out large-scale, shape-based docking at a high sampling granularity, where the overlapping surface patch centers were on average 1.5 Å apart.

Upon identifying scaffolds that harbor highly complementary surface patches, complementary patches would provide distance restraints for docked pose generation and refinement. These highly complementary complex models would serve as input for one-sided interface design, where the scaffold residues at the binding interface are designed to optimize the interactions with the query epitope. To more rigorously rank the designed complexes, two types of molecular dynamics (MD) simulations were used to predict the configurational stability and binding energy of the designed complexes.

Computational design of the VEGF binders

Using described strategy, we sought to design proteins able to prevent activation of the VEGFR by quenching the receptor-binding site of VEGF (Fig. 2A). In order to find scaffolds of complementary shape, we applied the HECTOR software to two surface patches on the VEGF binding site, which is structurally well characterised [20, 21]. These surface patches were inverse-mapped and docked against a library of fingerprints pre-computed from a high-resolution subset of PDB crystal structures. We retrieved six HECTOR hits (Fig. S3) and locally docked them against VEGF using PatchDock [22]. We selected two of these for further design: a bacterial ketosteroid isomerase (PDB: 1OH0) and a nitrophorin, hemoprotein of blood-feeding insects (PDB: 1PM1). Both templates showed a strong shape complementarity signal to the VEGF β-hairpin loop at their respective catalytic pockets. Both also possess much simpler architecture than published anti-VEGF antibodies and have molecular weights of 15 kDa and 20 kDa, respectively. We named the designs utilizing the first scaffold Sam and those using the second one Sima (Fig. 2B).

We then performed interface design by designing positions on the binder structure chosen automatically based on their proximity from the target surface (C_α distance ≤ 8 Å). Sequence and conformers sampling was done using the RosettaScripts framework [23]. The interface design was followed by two MD scoring steps, i.e. temperature-accelerated MD and potential of mean force calculations from non-equilibrium steered MD. We selected eight of the best-scoring Sam designs (Sam0.1 – Sam0.8) for experimental validation. The final sequences (Table S1) have 19 to 24 mutations compared to the starting template. For Sima proteins, we chose four candidates for testing where the native disulfide bonds of the scaffold were retained (Sima1.1 – Sima4.1) and four candidates where the disulfide bonds were eliminated (Sima1.2 – Sima4.2). The final Sima sequences (Table S1) contain 25 to 28 mutations compared to the scaffold.

Biophysical characterization of the designed VEGF inhibitors

As a first experimental step, we expressed the chosen designs in *E. coli* and double-purified them, in most cases, with good yields (Table S2). We then tested their oligomeric state with analytical size exclusion chromatography, revealing all the Sam designs to be monomeric and the Sima designs to exist as monomeric, dimeric, and tetrameric forms (Fig. S5). Based on expression yield and preliminarily estimated binding to VEGF (Fig. S6, Fig. S7), we chose the best-performing design from each group for full characterization; Sam0.7 and Sima3.2. We evaluated the thermostability and aggregation propensity of the proteins adopting nano differential scanning fluorimetry and backreflection measurements. Sam0.7 and Sima3.2 have similar melting temperatures (63 °C and 65 °C, respectively, Fig. 3A). However, light scattering thermograms revealed that Sima3.2 has a higher propensity for aggregation with an onset temperature of around 40 °C, compared to 60 °C for Sam0.7 (Fig. 3B). To characterize the kinetics and affinity of interactions between the designed proteins and VEGF, we performed surface plasmon resonance-based binding assays. Analysis of the kinetics across the injection dilution series, assuming 1:1 binding, resulted in dissociation constants (K_d) of 190 nM and 14 nM for Sam0.7 and Sima3.2, respectively. The higher affinity of Sima3.2 could be ascribed to its slower dissociation rate, compared to Sam0.7 (Fig. 3C, Fig. S8).

Structural characterization of the designs

To evaluate the accuracy of the computational design, we sought to determine experimental structures of the designed binders. We obtained crystal structures for the unbound Sam0.2 and Sam0.7 proteins. Alignment of the experimental structures to the corresponding design models showed that there is an atomic-level agreement with overall backbone Cα root mean square deviation (RMSD) of 1.8 Å for Sam0.2 and 1.7 Å for Sam0.7 (Fig. 4A, Fig. S9, Table S3). Further analysis of the per residue deviations revealed that regions with the highest mismatch (Cα RMSD per residue > 3 Å) correspond to the outer rim of the Sam binding site (e.g., R56 – M58, A64, F95, A96, D119 – V121 for Sam0.7, Fig. 4B). We ascribe this to the opened and closed conformations of Sam in its bound and unbound states, respectively. While designed coordinates of Sam were derived from its complex with VEGF (MD-relaxed structure, Materials and methods), the crystal structures represent isolated Sam monomers.

To further characterize the Sam binding mode, we expressed ^{13}C,^{15}N-labelled protein for NMR spectroscopy. The ^{15}N-HSQC fingerprint of the protein is well dispersed, indicative of a folded protein. However, experiments acquired toward sequential assignment of the protein showed multiple peaks for several residues (e.g., G60 and A61, Fig. S10). This is indicative of conformational exchange that is slow on NMR chemical shift timescales; i.e. milliseconds. For other residues ^{15}N-HSQC were not assigned due to low intensities or were missing altogether. These residues broadly coincide with loops on the parent structure. Despite this, sufficient residues were assigned to allow mapping of the VEGF binding epitope. We titrated labelled

Sam with unlabelled VEGF and recorded ^{15}N-HSQC at two VEGF concentrations. Given the measured affinity of Sam for VEGF, we should expect slow exchange between free and bound species at the concentrations used for the NMR samples; i.e. new peaks corresponding to bound Sam should appear as VEGF concentrations increase, while peaks for free Sam disappear. However, these observations are hampered by the large size of the Sam/VEGF complex, together with the low intensity of many free VEGF peaks in the region of the binding epitope. Nevertheless, peaks for some residues show the expected behaviour (Fig. S10) and map broadly to the designed epitope. Other peaks show gradual shifts on titration of VEGF, which could be the result of weak binding to a secondary epitope.

In vitro and in vivo evaluation of anti-VEGF activity

VEGF secreted by cancer cells provides mitogenic and survival stimuli for endothelial cells, leading to the formation of new blood vessels, and in turn, tumour expansion [15]. In order to assess the inhibitory capability of the designed proteins, we tested their anti-angiogenic effect on VEGF-responsive primary human endothelial cells, HUVEC. Indeed, the addition of either Sam0.7 or Sima3.2 into a culture medium in the low micromolar range significantly decreased VEGF-induced cell survival (Fig. 5A).

In addition to the regulation of angiogenesis, autocrine and paracrine VEGF/VEGFR signaling in tumour cells contributes to cancer cell proliferation, survival, induction of the epithelial-mesenchymal transition, and metastasis [24, 25]. For example, autocrine expression of VEGF has been reported to play a role in hematopoietic malignancies by stimulating growth and migration of leukemic cells [26]. We, therefore, decided to evaluate the efficacy of our designs on the proliferation of the acute myeloid leukemia cell line, U937, that is sensitive to VEGF. Treatment of U937 cells with increasing concentrations of the designed binders reduced cell proliferation down to 60 % in case of Sam0.7 and down to 50 % in case of Sima3.2 (Fig. 5B). Further analysis revealed half-maximal inhibitory concentrations (IC_{50}) to be 8.0 µg/ml (533 nM) and 3.6 µg/ml (180 nM) for Sam0.7 and Sima3.2, respectively (Fig. S11). To test whether detected anti-VEGF activity of our designs is mediated by residues introduced during interface design step, we performed U937 proliferation assay for the initial design templates of Sam and Sima (Sam_cntrl and Sima_cntrl, respectively). No inhibitory effect was observed (Fig. 5B). Additionaly, we evaluated the effect of the designed binders on the human embryonic kidney cell line, HEK293T, that does not express endogenous VEGFR. We observe a mild reduction in cell growth only after treatment of HEK293T cells with the highest concentration of Sam0.7 (100 µg/ml), demonstrating that our designs specifically target VEGF-dependent cells (Fig. S12).

After confirming the inhibitory effect of Sam0.7 on VEGF-dependent cell types, we proceeded to evaluate its anti-angiogenetic activity, using an *in vitro* microvascular formation model. In this experiment, a self-assembled 3D microvasculature was created by co-culturing hiPSC-derived endothelial cells and pericytes on a 3D fibrin gel. The microvasculature was treated with either Sam0.7 or Sam_cntrl (Fig. 6A). Treatment with Sam0.7 at both tested concentrations (50 µg/ml and 100 µg/ml) significantly reduced the angiogenesis *in vitro* in a dose-dependent manner (Fig. 6B). Specifically, the microvasculature treated with Sam0.7 exhibited lower area coverage and thinner blood vessels compared to the microvasculature treated with Sam_cntrl. Other microvascular parameters, such as connectivity, total vessel length, amount of branches, and others remained unchanged (Fig. S13). Moreover, Sam0.7 showed stronger effect on the microvasculature formation then 100 µg/ml bevacizumab - a highly potent antibody-based VEGF inhibitor that is approved for the treatment of several cancer types [27] (Fig. S14).

To further evaluate the pharmacological potential of the designs, we studied their effects on the VEGF-dependent solid tumor and blood cancer cells *in vivo*. Injection of 3.5 mg/ml (4 nl) of Sam0.7 into the brain of zebrafish embryos at 33 hours post fertilization (hpf) led to a decrease in the number of transplanted LNZ308 glioma cells after 2 days treatment (Fig. 7) at a comparable level to injection of 25 mg/ml of bevacizumab. A similar effect was observed when the design was tested on a leukemic cell line. Injection of Sam0.7 in the bloodstream of zebrafish embryos, where the transplanted leukemia cells, U937, were engrafted (Fig. S15), led to their reduction. Treatment of zebrafish embryos with 2.4 mg/ml (4 nl) of Sima3.2 over 2 days had a significant inhibiting effect on LNZ308 cells when it was injected directly into the brain (Fig. 7). Injection of Sima3.2 into the bloodstream of 33 hpf embryos did not reduce U937 cells and showed high lethality rate as opposed to PBS control and Sam0.7 injections (Fig. S15). This we attribute to the higher aggregation propensity of Sima3.2 in solution compared to Sam0.7.

Discussion

Protein design has demonstrated tremendous potential in creating new molecules for therapeutic applications, however, sculpting a binding pocket *de novo* remains a major scientific and technological challenge [28, 29]. This is due to the fact that a protein-protein binding event occurs under the influence of several interaction forces and major dynamic rearrangements of both the solute and solvent molecules [30, 31]. The accuracy of estimating energy change associated with these different factors is not uniform using the existing methods. Therefore, we argued that a tiered approach to binder design, where the simpler binding factors are optimized first, can overcome current limitations. Shape complementarity between the interacting molecular surfaces is one of the major drivers of binding [32], and can be geometrically quantified with a very high accuracy [33-36]. The relative importance of shape complementarity is particularly magnified in cases of maximum solvent-exclusion and substantial asymmetry at the binding interface, which can enhance binding affinity and specificity, respectively [37]. Here, we explored a complementarity-first approach, which simplifies the shape-docking problem through fast and efficient comparisons of surface fingerprints, with the aim of identifying binding scaffolds from a structural database. The most complementary scaffolds undergo one-sided interface design, followed by querying the dynamics of designed interfaces for more rigorous scoring.

The docking algorithm introduced in this work can identify design templates with locally-optimal surface topographies from among a large diversity of folds. This overcomes the conventional reliance of design on a small set of binder scaffolds (e.g. immunoglobulins [38], DARPins [39], anticalins [40], affibodies [41]). Although in this study we used a subset of the X-ray structures from the PDB, the computational performance of the ultra-fast docking algorithm allows the screening of much bigger databases. Particularly, databases of structural models predicted by AlphaFold2 [42], which are orders of magnitude larger than the PDB, can provide a vast source of potential scaffolds for our design approach [43, 44]. The scaffold selection stage enables selecting small, topologically simple, and PTM-free scaffolds. These properties are highly desirable for therapeutic proteins, as they simplify their production, processing, and quality control [45, 46].

To demonstrate the utility of our computational approach we set out to design binders targeting receptor-binding epitope on the VEGF surface to inhibit VEGF/VEGFR signaling. We tested only a small number of designs (16 candidates) belonging to two different folds. These designs had low molecular weight, were mostly well-expressed in *E. coli*, soluble, and stable. Affinity measurements showed a few designs to bind VEGF with nanomolar dissociation constants.

Notably, the two templates deployed here mount highly concave epitopes on mostly β-sheet structures. This has shown particularly advantageous for quenching the convex VEGF epitope, which is likely inaccessible by helical-scaffolds, hitherto the most-used scaffolds for *de novo* binders [47]. Our designs could inhibit the VEGF-dependent proliferation of leukemia cells, as well as the survival of primary human endothelial cells. The anti-VEGF activity of the designed Sam0.7 molecule was further confirmed using a human 3D microvasculature *in vitro* model: treatment with Sam0.7 resulted in the dose-dependent reduction in the area covered by blood vessels and thinning of the capillaries. Finally, our designs were effective in reducing the size of tumor in glioma zebrafish xenograft, highlighting their *in vivo* activity and therapeutic potential. Given the generality of this framework, it can be used to rapidly create sufficiently tight *de novo* binders against diverse targets through a single design round.

Due to invertibility of the HECTOR fingerprints, our algorithm can be adopted to evaluate both local similarity, and local complementarity between the query and subject surfaces. While in this work we focused only on complementarity to solve the shape-docking problem, we anticipate that the same approach could find wide utility in the field of protein structure alignment. Particularly, this can be used to identify local surface similarities and thus map remote functional relationships across large protein structural databases. Additionally, this approach is most powerful in comparing surface patches that are mounted on different tertiary motifs, which is unachievable through existing alignment methods [48-50]. The identification of such local similarities can provide an alternative approach to function inpainting [6, 51, 52].

Materials and methods

Computational design of de novo binders

The surface mapping and steric docking step begins by surface tessellation of the entire protein structure, to yield a dot-surface S_c of the molecular surface in the Cartesian coordinate system [53]. The associated vector field describing the outward-pointing surface normal vector at each surface dot S_n was derived. The position vector and surface normal at each dot would thus define the local origin and z-axis at each local surface patch. The protein's dot-surface was split into a number of overlapping surface patches with an average origin-to-origin spacing of 1.5 Å, where each surface patch $F_i = \{\mathbf{s_c} | \mathbf{s_c} \in S_c, \frac{\|(\mathbf{s_c}-\mathbf{s_{c_i}})\times\mathbf{s_{n_i}}\|}{\|\mathbf{s_{n_i}}\|} < R_{max}, \left|(\mathbf{s_{c_i}}+\mathbf{s_{n_i}})\cdot(\mathbf{s_c})\right| < \frac{H_{max}}{2}\}$ is a set of the Cartesian coordinates (x, y, z) of surface dot $\mathbf{s_c}$ that lie within a maximum distance R_{max} from the surface normal $\mathbf{s_{n_i}}$ of the i^{th} dot $\mathbf{s_{c_i}}$ (i.e. the centre of the patch), and are within half the maximum axial distance (along the surface normal vector $\mathbf{s_n}$). These individual surface patches contain their frame of reference where $\mathbf{s_{c_i}}$ is the origin and $\mathbf{s_{n_i}}$ is the local z-axis to be projected into a cylindrical coordinate system to yield a forward or inverse 3D map G_i according to:

$$T_p: f(x,y,z) \rightarrow g(r,\theta,l) \coloneqq \begin{bmatrix} \sqrt{x^2+y^2} \\ 1/\tan\left(\frac{y}{x}\right) \\ k \cdot z \end{bmatrix} \qquad (1)$$

A projected G_i can be mapped in either a forward (G_{+i}) or an obverse (G_{-i}) direction, depending on the value of $k = \begin{cases} +1 & forward \\ -1 & obverse \end{cases}$ to describe a forward- or obverse-mapped surface patch. At this stage, projected $G_{\pm i}$ patches are translation-invariant, and rotation-invariant along two principal planes; i.e. the 3D projections are only sensitive to alignments

along the θ dimension. Therefore, the indefinite integral across the θ dimension (eq. 2): i) reduces the dimensionality of the 3D projections into compressed 2D density projects; albeit at the loss of some information, ii) renders the 2D maps rotationally-invariant, and iii) possess constant dimensions.

$$T_r: g_{\pm}(r,\theta,l) \to h_{\pm}(r,l) \coloneqq \int g_{\pm}(r,\theta,l)d\theta \qquad (2)$$

Lastly, to decrease the sensitivity to minor resolution errors, a bilinear interpolation step (eq. 3) is applied to $H_{\pm i}$ to blur slight deviations in surface-dot densities, and allow for a degree of conformational insensitivity:

$$T_r: h_{\pm}(r,l) \to j_{\pm}(r,l) \coloneqq \sum_{i=0}^{1}\sum_{j=0}^{1} a_{ij}x^i y^j \qquad (3)$$

This yields the final form of the fingerprint $j_{\pm}(r,l)$ that can be directly used for complementarity matching by calculating an R-factor $R_{s,q}$ that best describes the normalized differences between two matrices [54] describing the forward- and obverse-mapped query and subject maps, respectively:

$$R_{s,q} = \frac{\Sigma\sqrt{\|j_{s,\pm}(r,l) - j_{q,\mp}(r,l)\|}}{\Sigma\sqrt{\|j_{s,\pm}(r,l)\|}} \qquad (4)$$

A 2018 Protein Data Bank snapshot was filtered down to the X-ray structures with resolution better than 2.0 Å, their hydrogen atoms coordinates were guessed, and the structures dot-surfaces were derived. These were forward-mapped as subject patches $j_{s,+}(r,l)$, where r covers radial distance range of $[0,12]$ Å from the surface normal axis of the patch, and l covers the axial distance range of $[-6,6]$ Å. The final maps were composed of 30×30 bins, where the resolution is 0.4 Å.

Two receptor binding surface patches were used as query in a two-against-two search, which was bounded by a distance restraint with a tolerance of ±3 Å. The retrieved hits were docked against VEGF based on their surface complementarity using PatchDock [22]. For that, the residues corresponding to patches with low R-factor in both the hit and VEGF were set as binding residues and used as constraints for rigid body docking using default parameters, as set by the `buildParamsToOutfile.pl` PatchDock utility (Supplementary methods). For each job, the top few poses were selected and further refined by using short, low-temperature molecular dynamics simulations. These simulations were run with generalized Born implicit solvent, where the temperature was set to 250 K using a Langevin thermostat through the NAMD2 software [55].

For the one-sided interface design stage, a design protocol was implemented in RosettaScripts [23] that deploys multiple design movers interlaced with backbone movers. The movers were bundled within two main iterative generic Monte Carlo sampling stages, which implemented with interface ΔΔG and packstat [56] filters for scoring. This design protocol was run for 100 instances per input decoy, and the entire procedure was performed in 4-6 successive design rounds (Supplementary methods). The top few thousand decoys were clustered to unique sequences and were further filtered using two-stage molecular dynamics ranking. The first stage filtered the designs using serial tempering simulations in implicit solvent between cool and hot basins of 250 K and 370 K for 6 and 8 ps, respectively. This was conducted for a total of 60 cycles separated by 100 steps of conjugate gradient minimization per simulation, for 3 replicas per decoy. The conformational homogeneity across the cycles and replicas were quantified for each decoy as a function of the all-against-all RMSD [57]. This reduced the decoys to a few hundred candidates that were further filtered by a second stage of steered molecular dynamics in explicit solvent under constant pressure and temperature. In the latter

stage, the binder's backbone atoms were fixed using harmonic restraints along the z-dimension, and VEGF was pulled along the z-dimension at a fixed velocity of 2 Å/ns using a force constant of 50 kcal/mol·Å. The estimated work of dissociation was derived from the potential of mean force as previously described [58, 59].

Bacterial protein expression and purification

Synthetic genes encoding the human VEGF165 (Gly34-Asp135) and the designs were cloned into the pET28a(+) expression vector between the *NdeI* and *XhoI* cloning sites in-frame with a thrombin cleavage site and an N-terminal poly-histidine purification tag (General Biosystems, Inc.; Synbio Technologies, Inc.). Plasmids were transformed into chemically competent *E. coli* BL21(DE3) using the heat shock method. Transformed cells were grown in LB medium supplemented with 40 μg/ml kanamycin at 37 °C. At OD600 of 0.6-1.0, cells were induced with 1mM IPTG and incubated overnight at 25°C for protein expression. For purification of all Sam designs and Sima3.2, cells were harvested by centrifugation at 5000 g at 4°C for 20 min and lysed in 25 ml of lysis buffer (1M guanidinium chloride, 100 mM NaCl, 50mM Tris-HCl pH 8.0) supplemented with a tablet of the cOmplete, EDTA-free Protease Inhibitor Cocktail (Roche, 5056489001) and 3 mg of lyophilized DNase I (PanReac AppliChem, A3778) using a Branson Sonifier S-250 (Fisher Scientific). The lysate was cleared by centrifugation at 28000 g for 50 min and the supernatant was passed through a 0.45 μm filter (Millipore, SLHV033RS). The sample was applied to a 5 ml HisTrap HP column (Cytiva, GE17-5248-01). The running buffer was 150 mM NaCl, 30 mM Tris-HCl pH 8.0. After sequential washing the column with 30 ml of the running buffer supplemented with 0 or 50 mM imidazole, fractions were collected by linear gradient elution using 150 mM NaCl, 30 mM Tris-HCl pH 8.0, 500 mM imidazole buffer. In case of other Sima designs and human VEGF, proteins were extracted from the insoluble fraction of the cell pellet by stirring in a phosphate-buffered saline (PBS) with 5 M guanidinium chloride, and 25 mM DTT for 2 h at room temperature. The mixture was gradually diluted 5 times with PBS and cleared by centrifugation at 28000 g for 50 min. The proteins were purified from the filtered supernatants by immobilized metal-affinity chromatography (IMAC) as described above. The only difference is that the composition of the running buffer was 150 mM NaCl, 30 mM Tris-HCl pH 8.0, 1 M urea, 1.25 mM reduced glutathione, 0.25 mM oxidized glutathione. The eluted fractions containing the protein of interest were pooled, concentrated using 10 kDa MWCO centrifugal filters (Millipore, UFC901024), and further purified on a Superdex Increase 75 10/300 gel filtration column (Cytiva, 29148721) using PBS. Gel filtration fractions containing pure protein in the desired oligomeric state (monomeric fraction of Sam0.7, dimeric fraction of Sima3.2) were pooled, concentrated, and stored at -20 °C for subsequent analyses. Both IMAC and gel filtration steps were performed on an Äkta Pure chromatography system (Cytiva).

Thermostability analysis

Nanoscale differential scanning fluorimetry (nanoDSF) using Prometheus NT.48 (Nanotemper) was applied to evaluate thermostability of the designs. Capillaries (Nanotemper, PR-C002) were filled with 1 mg/ml protein samples in three replicates. Melting scan was performed across the temperature range from 20 °C to 90 °C with a temperature ramp of 1 °C/min. In addition to measuring the intrinsic fluorescence intensity ratio (350/330 nm), light intensity loss due to scattering (backreflection) was measured to detect protein aggregation.

Isothermal Titration Calorimetry (ITC)

To determine whether Sam designs bind to VEGF, ITC experiments were performed using MicroCal PEAQ-ITC (Malvern Panalytical). All samples were prepared in PBS. The cell was loaded with VEGF solution at concentrations from 80 to 500 μM for different measurements.

The syringe was loaded with around 400 µM of Sam designs (375 µM of Sam0.2, 416 µM of Sam0.7). All measurements were carried out with the following parameters: 18 injections of 2 µl with 150 s between injections, reference power 41.9 µW, stirring speed 750 rpm, temperature 25 °C. Experimental data were fitted using MicroCal PEAQ-ITC Analysis Software. K_d values were corrected taking into account N-value that describes active fraction of protein in the cell and in the syringe.

Microscale thermophoresis (MST)

To test ability of Sima designs to bind VEGF, MST measurements were performed on Monolith NT.115 (Nanotemper). Sima1.1, Sima3.1, and Sima3.2 were labeled with fluorescent dye using Protein Labeling Kit RED-NHS 2nd Generation (Nanotemper, MO-L011). For each experiment 16 samples were prepared with constant concentration of labeled Sima protein (100 nM of Sima1.1, 40 nM of Sima3.1, and 80 nM of Sima3.2) and decreasing concentrations of VEGF (2-fold dilutions from 12.5 µM to 0.4 nM for Sima1.1 measurements, from 64 µM to 2 nM for Sima3.1 measurements, and from 2 µM to 61 pM for Sima3.2 measurements). All dilutions were made in PBS supplemented with 0.1 % Pluronic®F-127 (Nanotemper). Measurements were done in Monolith NT.115 Premium capillaries (Nanotemper, MO-K025) at 25 °C. Dissociation constants (K_d) were derived by fitting normalized fluorescence values (cold region: -2 - -1 s, hot region: 4-6 s) against ligand concentrations.

Surface plasmon resonance (SPR) binding assay

Multi-cycle kinetics experiments were performed on a Biacore X100 system (GE Healthcare Life Sciences). Recombinant Human VEGF165 (rhVEGF) (R&D Systems, 293-VE-010/CF) was diluted to 50 µg/mL in 10 mM acetate buffer pH 5.0 and immobilized on the surface of a CM5 sensor chip (Cytiva, 29149604) using standard amine coupling chemistry. The designs were diluted in the running buffer (PBS with 0.05% v/v Tween-20). Analyses were conducted at 25°C at a flow rate of 10 µL/min. Five increasing concentrations of the sample solution (111 nM, 134 nM, 161 nM, 193 nM, 231 nM for Sam0.7; 44 nM, 66 nM, 100 nM, 148 nM, 222 nM for Sima3.2) were injected over the functionalized sensor chip surface for 60 s, followed by a 60 s dissociation with running buffer. At the end of each run, the sensor surface was regenerated with a 30 s injection of 50 mM NaOH at a flow rate of 10 µL/min. The reference responses and zero-concentration sensograms were subtracted from each dataset (double-referencing). Association rate (k_a), dissociation rate (k_d), and equilibrium dissociation (K_d) constants were obtained using the linearization method described in [6].

X-ray structure determination

Crystallization screens were set up with a Mosquito robot (TTP Labtech) in 96-well plates at 21 °C, using 75 µl of reservoir solution and sitting drops containing 400 nl of reservoir and 400 nl of protein solution of either Sam0.2 or Sam0.7 at a concentration of 8 mg/ml. Within 7-14 days, crystals of Sam0.2 and Sam0.7 grew in a condition containing 20 % (w/v) PEG 3350 and 0.2 M KSCN. Crystals were cryoprotected through the addition of 15% PEG 400, flash cooled and stored in liquid nitrogen until data collection. Diffraction data were collected at 100K on an EIGER2 16M detector at beamline X10SA at the Swiss Light Source (SLS). Data were processed and scaled using XDS [60] and the structures solved using molecular replacement with MOLREP [61] and the computational models of Sam0.2 and Sam0.7 as search models. Structures were completed by cyclic refinement with REFMAC5 [62] and modelling using Coot [63]. Data collection and refinement statistics are summarized together with PDB accession codes in Table S3.

Nuclear magnetic resonance (NMR) spectroscopy

Spectra were recorded at 310 K on Bruker AVIII-600 and AVIII-800 spectrometers. Backbone sequential assignments were made using standard triple-resonance experiments, including 3D-(H)CC(CA)NH-TOCSY and 3D-(H)CC(CO)NH-TOCSY spectra. Sequential assignment was achieved for the majority of the protein, however a larger stretch between V36 and A64, which corresponds to the outer rim of the binding site, could not be assigned due to multiple signals for some residues and weak or missing signals for others. This is consistent with this region undergoing slow conformational exchange processes (i.e. over millisecond timescales). To test VEGF binding, ~50 µM ^{15}N,^{13}C-labelled Sam0.2 was titrated with unlabelled VEGF at a molar ratio of 1:2 and 1:3. ^{15}N-HSQC spectra were recorded at each titration point and compared with that of the free protein.

HUVEC survival assay

HUVECs (Sigma, 200-05N) were cultured in Endothelial Cell Growth Medium (EGM) (Cell Applications, Inc., 211-500) and used at passages 4 to 6. 300 µl of cell suspension were seeded in a 48-well plate (Nunc, 150687) at a density of 10^5 cells/ml. When the culture reached about 80 % confluency, cells were washed once with DPBS and medium was changed to Endothelial Cell Basal Medium (EBM) (Cell Applications, Inc., 210-500) or EBM supplemented with 30 ng/ml rhVEGF, or with 30 ng/ml rhVEGF and different concentrations of VEGF inhibitors varying from 1 µg/ml to 100 µg/ml. Before the experiment, inhibitors were preincubated with rhVEGF for 30 min at room temperature. After incubation for 48 hours at 37°C, 5% CO_2, 60 µl of CellTiter-Blue Reagent (Promega, G8080) were added to the wells and the plate was incubated for an additional 1 h under the same conditions. Cell survival was monitored by measuring fluorescence (560/590 nm) using a Synergy HTX Microplate Reader (BioTek). The results were normalized to the average of fluorescence values from the wells with EBM only.

U937/HEK293T cell proliferation endpoint analysis

U937 cells (DSMZ, ACC 5) were cultured in RPMI 1640 medium (Gibco, 22400071) supplemented with 10 % FBS (Gibco, 10082147). HEK293T cells (DSMZ, ACC 635) were cultured in DMEM medium (Gibco, 41966029) supplemented with 10 % FBS (Gibco, 10082147). Cells were pelleted by centrifugation at 300 g for 5 min, washed once with DPBS (Gibco, 14190144) and once with non-supplemented medium. After the last washing step, cells were resuspended in RPMI 1640 (for U937) or DMEM (for HEK293T) medium supplemented with 1 % FBS. 100 µl of U937 cell suspension were seeded in a 96-well plate (Sarstedt, 83.3925.500) at a density of 2×10^5 cells/ml. For HEK293T, 100 µl of cell suspension were seeded in a 96-well plate (Corning, 3596) at a density of 5×10^4 cells/ml. Different concentrations of Sam0.7 or Sima3.2 varying from 1 µg/ml to 100 µg/ml were added to the wells in triplicates. DPBS was added to the wells serving as an untreated control. After incubation for 72 h at 37°C, 5% CO_2, 20 µl of CellTiter-Blue Reagent (Promega, G8080) were added to the wells and the plate was incubated for an additional 1 h under the same conditions to allow cells to convert resazurin to resorufin. Cell viability was monitored by measuring fluorescence (560/590 nm) using a Synergy HTX Microplate Reader (BioTek). The average of fluorescence values of the culture medium background was subtracted from all fluorescence values of the experimental wells. The data were presented as a percentage of untreated control fluorescence values.

In vitro 3D microvasculature analysis

Human episomal iPSC line from a healthy female donor, E1 (ThermoFisher Scientific, A18945) was cultured on Geltrex-coated (Gibco, A1413302) plates in Essential 8™ Medium (Gibco, A1517001). Endothelial cells (ECs) and pericytes (PCs) were differentiated following

established protocols. ECs were labeled with the pJG-IRBP-mCherry viral vector to enable live cell imaging. Differentiated ECs were cultured on 0.2% gelatin-coated plates (Sigma-Aldrich, G2500) in Endothelial Cell Growth Medium (PromoCell, C-22020) supplemented with 30 ng/mL vascular endothelial growth factor A (PeproTech, 100-20) and 20 ng/ml fibroblast growth factor 2 (PeproTech, 100-18B); PCs were cultured in Essential 6™ Medium (Gibco, A1516501) supplemented with 10% FBS. Confluent ECs and PCs were detached with Accumax™ (Sigma-Aldrich, A7098) pelleted by centrifugation at 300 g for 3 min and resuspended in a fibrin gel composed of 5 mg/mL fibrinogen (Sigma-Aldrich, F6755) and 4 U/mL thrombin (Sigma-Aldrich, T9549), with a final cell concentration 3×10^6 cell/mL of each cell type. 3D fibrin drops were formed using 5 µL of cell suspension in 96-well plates (Falcon, 351172) and cultured in Endothelial Cell Growth Medium supplemented with 20 ng/ml fibroblast growth factor 2. Two concentrations of Sam0.7 (50 and 100 µg/ml) were added to the wells, with 5 repeats for each condition; Sam_ctrl protein was added to the wells as control. The microvasculature was grown for 7 days and then imaged with an Axio Observer Z1/7 (Zeiss) microscope. Fiji/ImageJ software was utilized to analyze the microvasculature parameters. The statistical analysis was done using GraphPad Prism 10 software. All procedures were in accordance with the Helsinki Convention and approved by the Ethical Committee of the Eberhard Karls University Tübingen (no. 396/2021BO2).

Xenotransplantation of U937-GFP or LNZ308-GFP cells in zebrafish embryos

Zebrafish lines were maintained according to standard protocols and handled in accordance with European Union animal protection directive 2010/63/EU and approved by the local government (Tierschutzgesetz §11, Abs. 1, Nr. 1, husbandry permit 35/9185.46/Uni TÜ). The role of the designed inhibitors in cell survival and proliferation was evaluated by xenotransplantation of U937-GFP and LNZ308-GFP cell lines in zebrafish embryo at 1.5 dpf. U937-GFP cells were generated by lentiviral transduction with lentiviral construct expressing GFP, pRRL.PPT.SF..i2GFPpre, kindly provided by A. Schambach, Hannover Medical School. LNZ308-GFP cells were generated by lentiviral transduction with a third generation GFP-expressing lentivirus based on the lentiviral vector pLJM1-EGFP (Plasmid #19319, Addgene). 1 nl of U937-GFP cell suspension was injected at a density of 2×10^5 cells/µL (around 200 cells) into the perivitelline space, whereas LNZ308-GFP glioma cells (1 nl, around 200 cells) were orthopically implanted into the brain of anesthesized embryos at 29 hpf. The embryos were incubated at 35°C after transplantation. After 4 to 5 hours alive embryos with good engraftment were injected with same volume (4 nl) of either control (PBS or 3.0 mg/ml of inactive protein mvn_cntrl [6] as a negative control, or 25 mg/ml of Bevacizumab as a positive control) or inhibitors (3.5 mg/ml of Sam0.7, and 2.4 mg/ml of Sima3.2). Injections were done either in the brain or bloodstream of zebrafish embryos transplanted with LNZ308-GFP cells or U937-GFP cells, respectively. The embryos were positioned and orientated laterally within cavities formed in 1 % agarose on a 96-well plate for imaging. To quantify the level of transplanted human cells, zebrafish embryos were imaged by a Nikon fluorescent stereomicroscope (SMZ18). All images taken under same condition were analyzed with Imaris software. The fluorescently labeled cells were quantified by surface measurement option with background subtraction. These values were calculated in each embryo of each group and plotted using Grafpad. The statistical analysis was performed using Prism 7 software.

Data accessibility

Coordinates and structure factors of Sam0.2 and Sam0.7 were deposited in the PDB under the accession codes 8BL5 and 8BL9, respectively.

Acknowledgments

This project has received funding from the IMPRS (KM), the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement No 863952 (ACE-OF-SPACE)) (PM), the M. Schickedanz Kinderkrebsstiftung (ME, JS, PM, NA), "German Universities Excellence Initiative" of the Tübingen University (JS), DFG (No 500215849) (ME, JS), Fortüne program of the University of Tübingen (2667-0-0) (NP).

Contributions

Conceptualization: ME, AL, PM, JS, JP; HECTOR software and construct design: ME; biophysical characterization: KM, ME; cell-based experiments: KM, GT; crystallography: KM, MDH; NMR: MC, ME, KM; in vitro microvaculature experiments: NP, MV, SL; zebrafish experiments: NA; funding acquisition: AL, PM, JS, ME; resources: AL, PM, JS; supervision: AL, ME, JS, PM; writing – original draft: KM, ME, MC, NA; writing – review and editing: all authors.

Figure 1

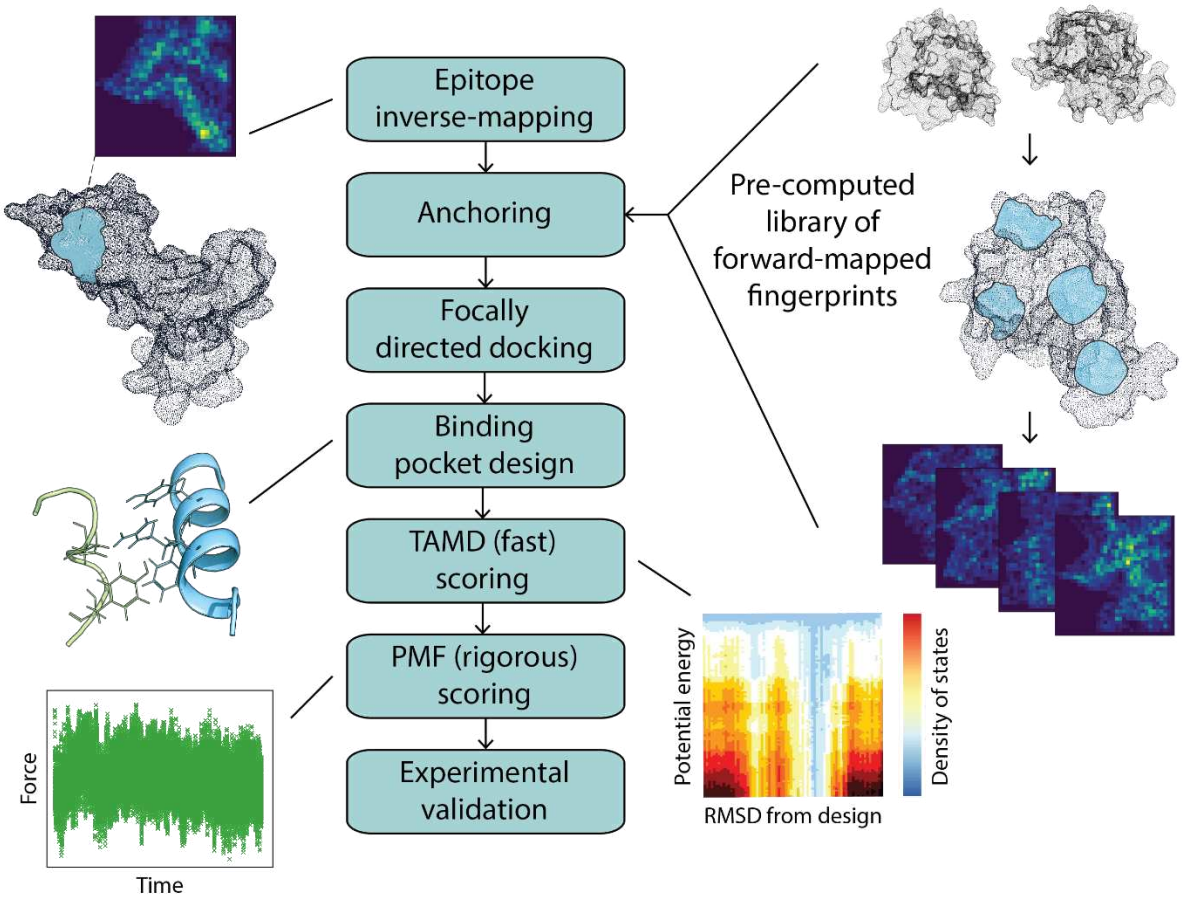


Figure 1. Design strategy of epitope-directed binders. The design workflow primarily relies on simplifying the complexity of the docking process and decoupling it from the design stage. The docking step particularly is based on a novel surface mesh fingerprinting protocol that analytically defines a complementary surface fingerprint and trivially matches it to a query surface. This is followed by steps of high-resolution docking, sequence design, and molecular dynamics-based testing of binding interface stability. TAMD - temperature-accelerated molecular dynamics. PMF - potential of mean force. RMSD - root-mean-square deviation.

Figure 2

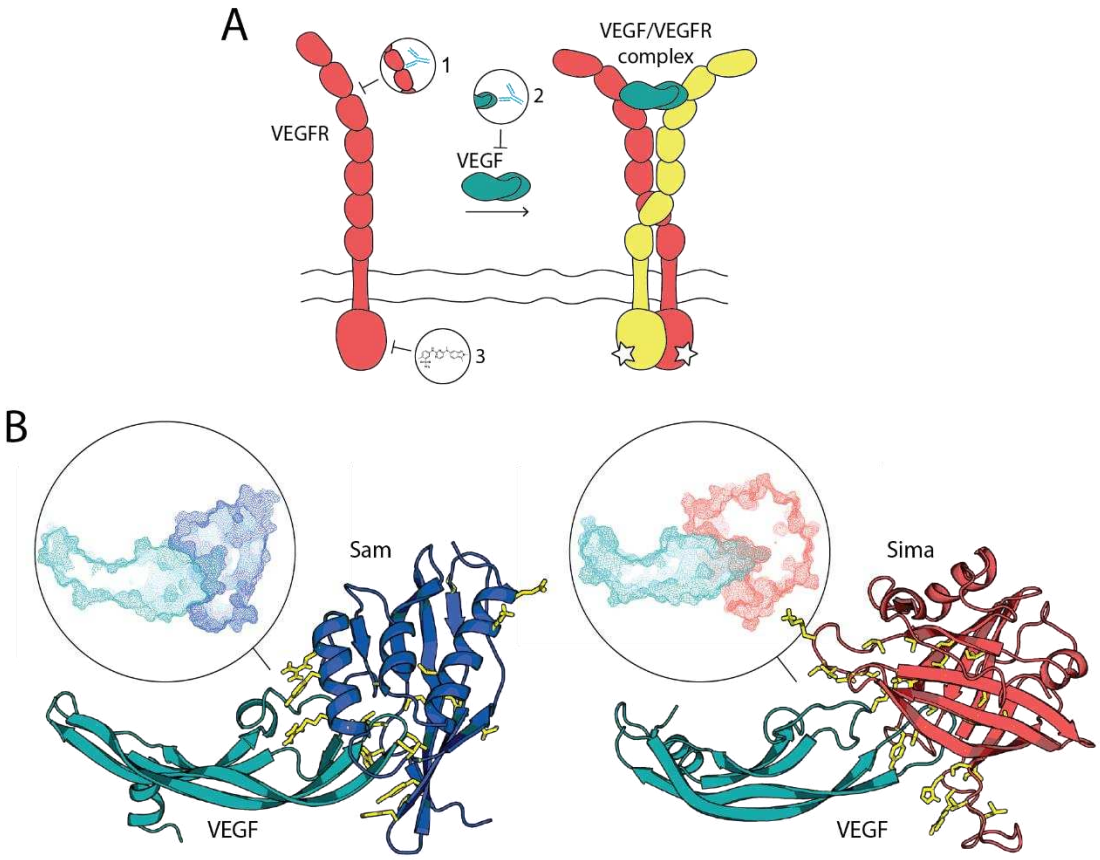


Figure 2. Mechanism of VEGFR activation and strategies for its inhibition. **(A)** VEGFR subunit is composed of 7 extracellular domains, a transmembrane segment, and an intracellular kinase domain (coral or yellow). VEGFR is activated when dimeric VEGF (teal) binds, and in turn, dimerizes two receptor subunits. This dimeric receptor configuration triggers the intracellular domain kinase activity. VEGFR signaling can be inhibited via: 1) protein-based quenching of the ligand binding site on the receptor surface, 2) protein-based sequestration of VEGF itself at its receptor binding site, or 3) inhibition of the kinase activity through the use of small molecules. **(B)** Anti-VEGF binders were designed based on two different scaffolds, Sam (blue) and Sima (coral), that showed a high shape complementarity to the receptor-binding site of VEGF (teal). Amino acids mutated during binder pocket design are colored in yellow.

Figure 3

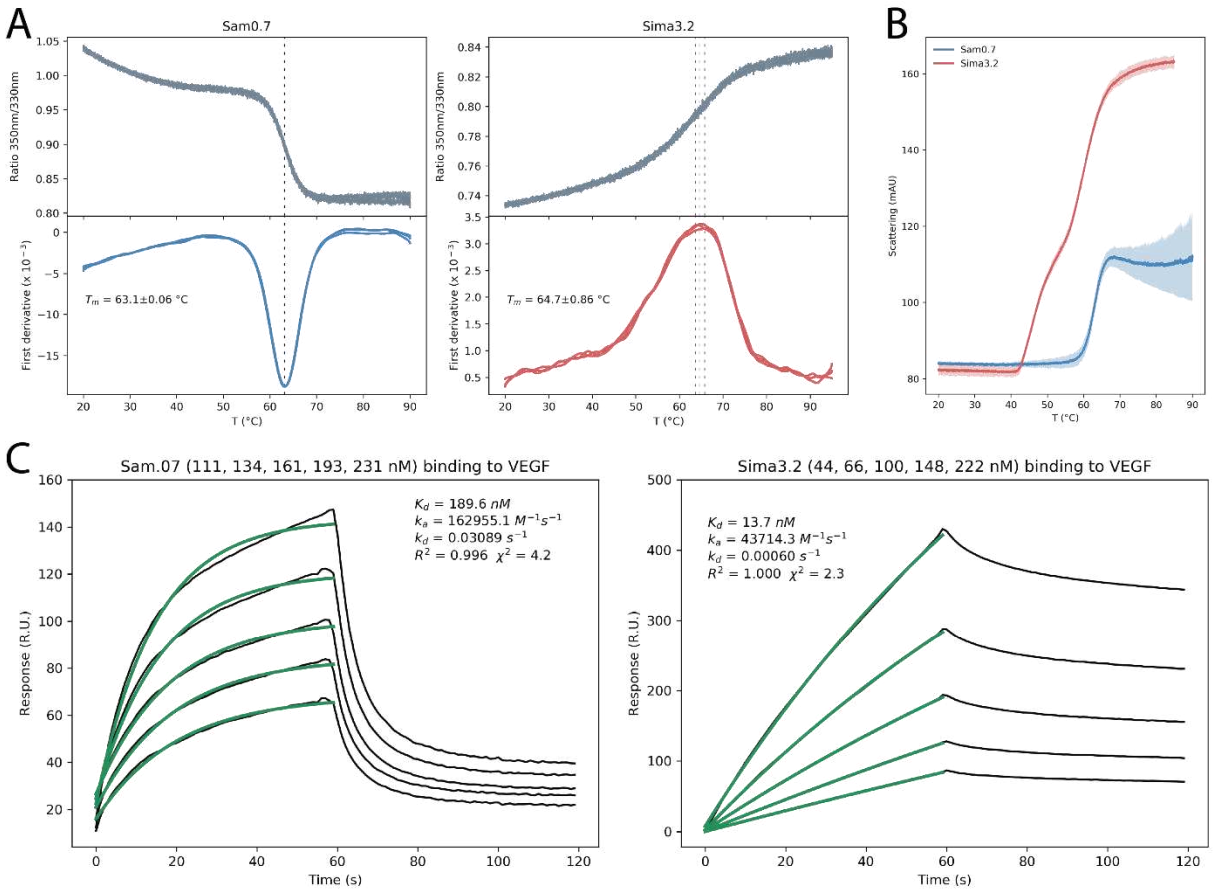


Figure 3. Biophysical characterization of the designed proteins. **(A)** Thermal unfolding curves show melting temperatures (T_m) of both Sam0.7 and Sima3.2 designs to be higher than 60 °C. **(B)** Light scattering thermograms indicate that Sima3.2 has higher propensity for aggregation with an onset temperature of around 40 °C, compared to an onset temperature of 60 °C for Sam0.7. Shades represent the standard deviation across three replicates. **(C)** The designed proteins bind VEGF with nanomolar affinity as measured by SPR.

Figure 4

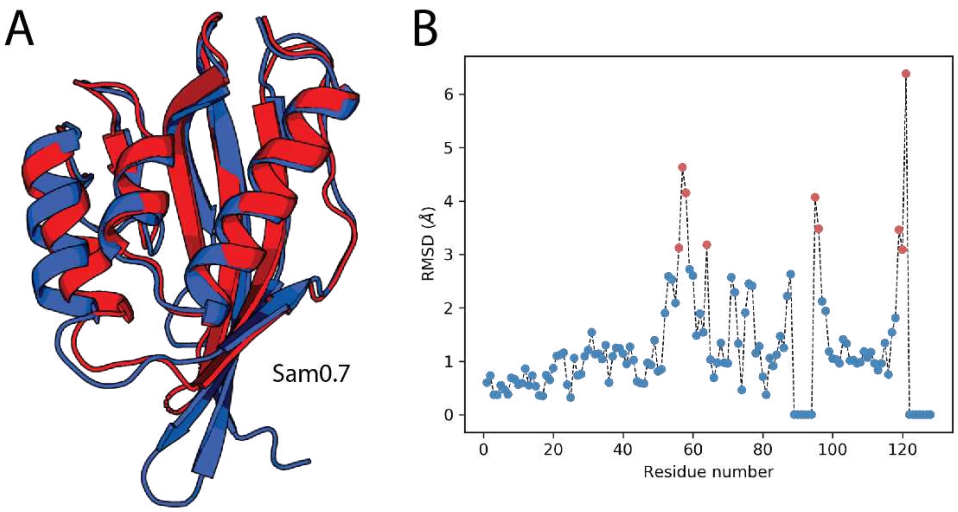


Figure 4. Crystal structure of Sam0.7 matches the computational design model with atomic-level accuracy. (A) Superimposition of the Sam0.7 design model (blue) and the experimentally determined crystal structure (red). **(B)** RMSD between coordinates of the C_α atoms in the Sam0.7 design model and corresponding C_α atoms in the crystal structure. RMSD values higher than 3 Å are colored in red. Residues in gaps are assigned an RMSD value of 0.

Figure 5

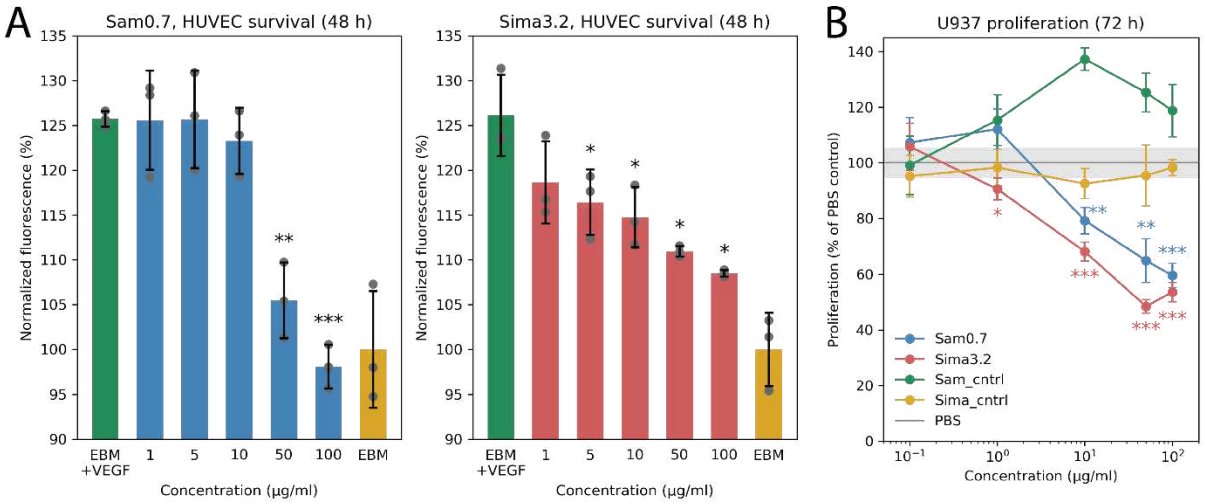


Figure 5. The designs inhibit proliferation and survival of VEGF-dependent cells. **(A)** VEGF-dependent survival of HUVEC primary cells was significantly reduced in a dose-dependent manner by the designed binders. Yellow bars indicate the survival of the cells in an endothelial cell growth basal medium without VEGF, whereas green bars correspond to survival of the cells in a basal medium with 30 nM of VEGF. Blue and red bars show results on cell survival in a basal medium with 30 nM of VEGF and increasing concentrations of Sam0.7 and Sima3.2, respectively. **(B)** Treatment of U937 acute myeloid leukemia cell line with Sam0.7 and Sima3.2 proteins at low micromolar concentrations decreased the cell growth. In contrast, unmutated scaffolds, Sam_cntrl and Sima_cntrl, did not show any inhibitory activity. Error bars represent the standard deviations across three replicates from one experiment. Statistical significance was calculated using Fisher's one-sided t-test (*, $p \leq 0.05$, **, $p \leq 0.01$, ***, $p \leq 0.001$ vs. the PBS group).

Figure 6

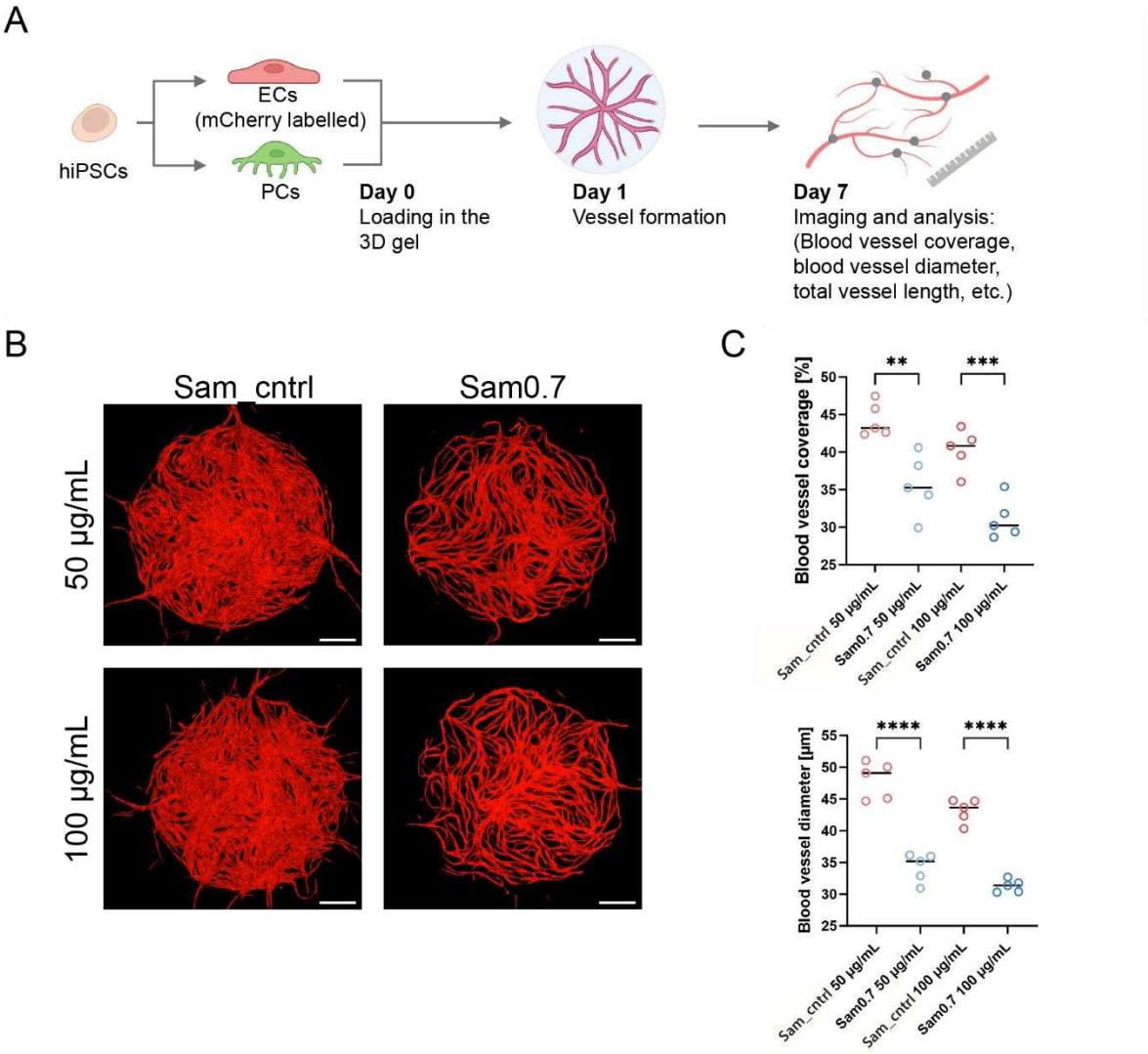


Figure 6. The designed binder Sam0.7 reduces angiogenesis *in vitro*. (A) Schematic representation of *in vitro* microvasculature formation and analysis: human induced pluripotent stem cell (hiPSC)-derived endothelial cells (ECs) and pericytes (PCs) were co-cultured in the fibrin gel with Sam0.7 or Sam_cntrl as a negative control. The formed microvasculature was imaged at day 7, and images were analyzed to calculate microvasculature parameters. The figure was created with BioRender. **(B)** Representative images showing *in vitro* microvasculature formation in the presence of Sam0.7 or Sam_cntrl at two working concentrations (50 and 100 µg/mL). The scale bar is 500 µm. **(C)** Quantitative analysis of microvasculature formation. The upper plot shows the percentage of the area covered with blood vessels, and the bottom plot shows the blood vessel diameter. Statistical significance was calculated using the one-way ANOVA test (** $p \leq 0.01$, *** $p \leq 0.001$, **** $p \leq 0.0001$ treated Sam0.7 vs. Sam_cntrl group).

Figure 7

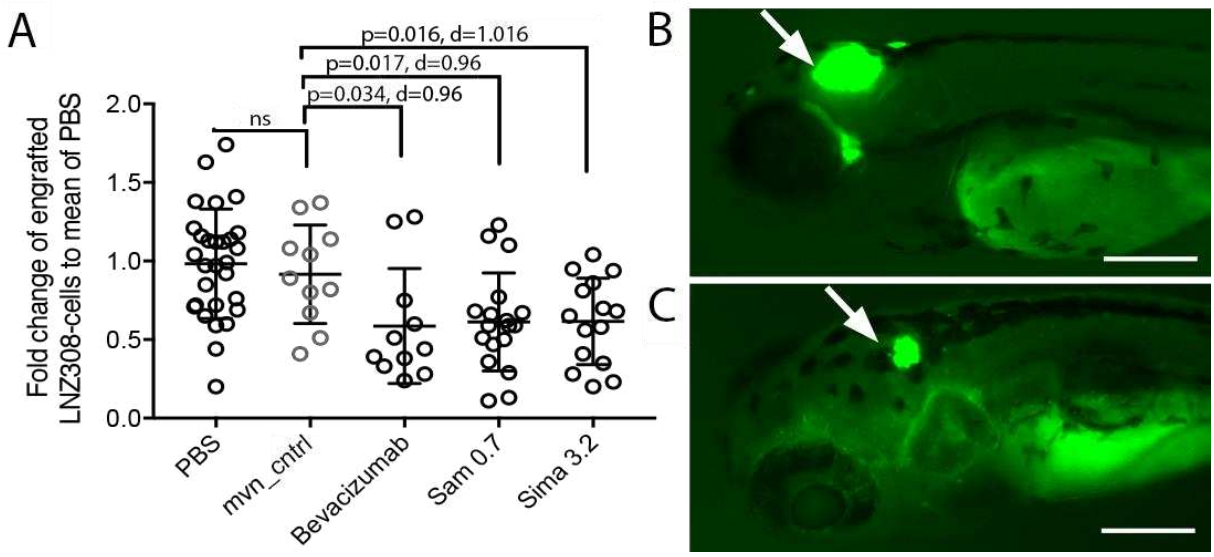


Figure 7. The designed binders show *in vivo* activity in zebrafish xenograft. **(A)** Quantification of the engrafted LNZ308-GFP glioma cells in zebrafish embryos that were injected with PBS, inactive protein (mvn_cntrl) as negative control, Bevacizumab as positive control, Sam 0.7, or Sima 3.2. Each dot indicates one embryo. p-value was calculated by Mann Whitney two tailed test. d – Cohen's d value. **(B)** Representative zebrafish xenograft treated with PBS. **(C)** Representative zebrafish xenograft treated with Sam 0.7. Arrowheads indicate transplanted LNZ308-GFP cells in the brain. The scale bar is 200µm.

References

1. Gebauer, M. and A. Skerra, *Engineered Protein Scaffolds as Next-Generation Therapeutics.* Annual Review of Pharmacology and Toxicology, 2020. **60**(1): p. 391-415.
2. Korendovych, I.V. and W.F. DeGrado, *De novo protein design, a retrospective.* Q Rev Biophys, 2020. **53**: p. e3.
3. Correia, B.E., et al., *Computational design of epitope-scaffolds allows induction of antibodies specific for a poorly immunogenic HIV vaccine epitope.* Structure, 2010. **18**(9): p. 1116-26.
4. Procko, E., et al., *A computationally designed inhibitor of an Epstein-Barr viral Bcl-2 protein induces apoptosis in infected cells.* Cell, 2014. **157**(7): p. 1644-1656.
5. Chevalier, A., et al., *Massively parallel de novo protein design for targeted therapeutics.* Nature, 2017. **550**(7674): p. 74-79.
6. Hernandez Alvarez, B., et al., *Design of novel granulopoietic proteins by topological rescaffolding.* PLoS Biol, 2020. **18**(12): p. e3000919.
7. Bryan, C.M., et al., *Computational design of a synthetic PD-1 agonist.* Proc Natl Acad Sci U S A, 2021. **118**(29).
8. Yang, C., et al., *Bottom-up de novo design of functional proteins with complex structural features.* Nat Chem Biol, 2021. **17**(4): p. 492-500.
9. Pan, X. and T. Kortemme, *Recent advances in de novo protein design: Principles, methods, and applications.* J Biol Chem, 2021. **296**: p. 100558.
10. Cao, L., et al., *De novo design of picomolar SARS-CoV-2 miniprotein inhibitors.* Science, 2020. **370**(6515): p. 426-431.
11. Cao, L., et al., *Design of protein-binding proteins from the target structure alone.* Nature, 2022. **605**(7910): p. 551-560.
12. Gainza, P., et al., *De novo design of site-specific protein interactions with learned surface fingerprints.* bioRxiv, 2022: p. 2022.06.16.496402.
13. Rabia, L.A., et al., *Understanding and overcoming trade-offs between antibody affinity, specificity, stability and solubility.* Biochem Eng J, 2018. **137**: p. 365-374.
14. Apte, R.S., D.S. Chen, and N. Ferrara, *VEGF in Signaling and Disease: Beyond Discovery and Development.* Cell, 2019. **176**(6): p. 1248-1264.
15. Karaman, S., V.M. Leppänen, and K. Alitalo, *Vascular endothelial growth factor signaling in development and disease.* Development, 2018. **145**(14).
16. Shaik, F., et al., *Structural Basis for Vascular Endothelial Growth Factor Receptor Activation and Implications for Disease Therapy.* Biomolecules, 2020. **10**(12).
17. Schneider, B.P. and G.W. Sledge, Jr., *Drug insight: VEGF as a therapeutic target for breast cancer.* Nat Clin Pract Oncol, 2007. **4**(3): p. 181-9.
18. Thurber, G.M., M.M. Schmidt, and K.D. Wittrup, *Antibody tumor penetration: transport opposed by systemic and antigen-mediated clearance.* Adv Drug Deliv Rev, 2008. **60**(12): p. 1421-34.
19. Kintzing, J.R., M.V. Filsinger Interrante, and J.R. Cochran, *Emerging Strategies for Developing Next-Generation Protein Therapeutics for Cancer Treatment.* Trends Pharmacol Sci, 2016. **37**(12): p. 993-1008.
20. Muller, Y.A., et al., *Vascular endothelial growth factor: crystal structure and functional mapping of the kinase domain receptor binding site.* Proc Natl Acad Sci U S A, 1997. **94**(14): p. 7192-7.
21. Brozzo, M.S., et al., *Thermodynamic and structural description of allosterically regulated VEGFR-2 dimerization.* Blood, 2012. **119**(7): p. 1781-8.

22. Schneidman-Duhovny, D., et al., *PatchDock and SymmDock: servers for rigid and symmetric docking.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W363-7.
23. Fleishman, S.J., et al., *RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite.* PLoS One, 2011. **6**(6): p. e20161.
24. Perrot-Applanat, M. and M. Di Benedetto, *Autocrine functions of VEGF in breast tumor cells: adhesion, survival, migration and invasion.* Cell Adh Migr, 2012. **6**(6): p. 547-53.
25. Goel, H.L. and A.M. Mercurio, *VEGF targets the tumour cell.* Nat Rev Cancer, 2013. **13**(12): p. 871-82.
26. Wegiel, B., et al., *The role of VEGF and a functional link between VEGF and p27Kip1 in acute myeloid leukemia.* Leukemia, 2009. **23**(2): p. 251-61.
27. Ferrara, N. and A.P. Adamis, *Ten years of anti-vascular endothelial growth factor therapy.* Nat Rev Drug Discov, 2016. **15**(6): p. 385-403.
28. Kuhlman, B. and P. Bradley, *Advances in protein structure prediction and design.* Nat Rev Mol Cell Biol, 2019. **20**(11): p. 681-697.
29. ElGamacy, M., *Accelerating therapeutic protein design.* Adv Protein Chem Struct Biol, 2022. **130**: p. 85-118.
30. Siebenmorgen, T. and M. Zacharias, *Computational prediction of protein–protein binding affinities.* WIREs Computational Molecular Science, 2020. **10**(3): p. e1448.
31. Chen, J., N. Sawyer, and L. Regan, *Protein-protein interactions: general trends in the relationship between binding affinity and interfacial buried surface area.* Protein Sci, 2013. **22**(4): p. 510-5.
32. Gao, F., J. Glaser, and S.C. Glotzer, *The role of complementary shape in protein dimerization.* Soft Matter, 2021. **17**(31): p. 7376-7383.
33. Milanetti, E., et al., *2D Zernike polynomial expansion: Finding the protein-protein binding regions.* Computational and Structural Biotechnology Journal, 2021. **19**: p. 29-36.
34. Gainza, P., et al., *Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning.* Nature Methods, 2020. **17**(2): p. 184-192.
35. Daberdaku, S. and C. Ferrari, *Antibody interface prediction with 3D Zernike descriptors and SVM.* Bioinformatics, 2019. **35**(11): p. 1870-1876.
36. Yin, S., et al., *Fast screening of protein surfaces using geometric invariant fingerprints.* Proc Natl Acad Sci U S A, 2009. **106**(39): p. 16622-6.
37. Li, Y., X. Zhang, and D. Cao, *The Role of Shape Complementarity in the Protein-Protein Interactions.* Scientific Reports, 2013. **3**(1): p. 3271.
38. Lu, R.M., et al., *Development of therapeutic antibodies for the treatment of diseases.* J Biomed Sci, 2020. **27**(1): p. 1.
39. Plückthun, A., *Designed ankyrin repeat proteins (DARPins): binding proteins for research, diagnostics, and therapy.* Annu Rev Pharmacol Toxicol, 2015. **55**: p. 489-511.
40. Deuschle, F.C., E. Ilyukhina, and A. Skerra, *Anticalin® proteins: from bench to bedside.* Expert Opin Biol Ther, 2021. **21**(4): p. 509-518.
41. Frejd, F.Y. and K.T. Kim, *Affibody molecules as engineered protein drugs.* Exp Mol Med, 2017. **49**(3): p. e306.
42. Jumper, J., et al., *Highly accurate protein structure prediction with AlphaFold.* Nature, 2021. **596**(7873): p. 583-589.
43. Tunyasuvunakool, K., et al., *Highly accurate protein structure prediction for the human proteome.* Nature, 2021. **596**(7873): p. 590-596.
44. Humphreys, I.R., et al., *Computed structures of core eukaryotic protein complexes.* Science. **374**(6573): p. eabm4805.

45. Tripathi, N.K. and A. Shrivastava, *Recent Developments in Bioprocessing of Recombinant Proteins: Expression Hosts and Process Development.* Front Bioeng Biotechnol, 2019. **7**: p. 420.
46. Gebauer, M. and A. Skerra, *Engineered Protein Scaffolds as Next-Generation Therapeutics.* Annu Rev Pharmacol Toxicol, 2020. **60**: p. 391-415.
47. Marchand, A., A.K. Van Hall-Beauvais, and B.E. Correia, *Computational design of novel protein-protein interactions - An overview on methodological approaches and applications.* Curr Opin Struct Biol, 2022. **74**: p. 102370.
48. Zhang, Y. and J. Skolnick, *TM-align: a protein structure alignment algorithm based on the TM-score.* Nucleic Acids Res, 2005. **33**(7): p. 2302-9.
49. Holm, L., *Using Dali for Protein Structure Comparison.* Methods Mol Biol, 2020. **2112**: p. 29-42.
50. van Kempen, M., et al., *Foldseek: fast and accurate protein structure search.* bioRxiv, 2022: p. 2022.02.07.479398.
51. Liu, S., et al., *Nonnatural protein-protein interaction-pair design by key residues grafting.* Proc Natl Acad Sci U S A, 2007. **104**(13): p. 5330-5.
52. Wang, J., et al., *Scaffolding protein functional sites using deep learning.* Science, 2022. **377**(6604): p. 387-394.
53. Xu, D. and Y. Zhang, *Generating Triangulated Macromolecular Surfaces by Euclidean Distance Transform.* PLOS ONE, 2009. **4**(12): p. e8140.
54. ElGamacy, M., et al., *Mapping Local Conformational Landscapes of Proteins in Solution.* Structure, 2019. **27**(5): p. 853-865.e5.
55. Kalé, L., et al., *NAMD2: Greater Scalability for Parallel Molecular Dynamics.* Journal of Computational Physics, 1999. **151**(1): p. 283-312.
56. Sheffler, W. and D. Baker, *RosettaHoles: rapid assessment of protein core packing for structure prediction, refinement, design, and validation.* Protein Sci, 2009. **18**(1): p. 229-39.
57. Skokowa, J., et al., *A topological refactoring design strategy yields highly stable granulopoietic proteins.* Nature Communications, 2022. **13**(1): p. 2948.
58. ElGamacy, M., et al., *An Interface-Driven Design Strategy Yields a Novel, Corrugated Protein Architecture.* ACS Synth Biol, 2018. **7**(9): p. 2226-2235.
59. ElGamacy, M., M. Coles, and A. Lupas, *Asymmetric protein design from conserved supersecondary structures.* J Struct Biol, 2018. **204**(3): p. 380-387.
60. Kabsch, W., *XDS.* Acta Crystallogr D Biol Crystallogr, 2010. **66**(Pt 2): p. 125-32.
61. Vagin, A. and A. Teplyakov, *Molecular replacement with MOLREP.* Acta Crystallogr D Biol Crystallogr, 2010. **66**(Pt 1): p. 22-5.
62. Murshudov, G.N., A.A. Vagin, and E.J. Dodson, *Refinement of macromolecular structures by the maximum-likelihood method.* Acta Crystallogr D Biol Crystallogr, 1997. **53**(Pt 3): p. 240-55.
63. Emsley, P., et al., *Features and development of Coot.* Acta Crystallogr D Biol Crystallogr, 2010. **66**(Pt 4): p. 486-501.

Supplementary material

Figure S1

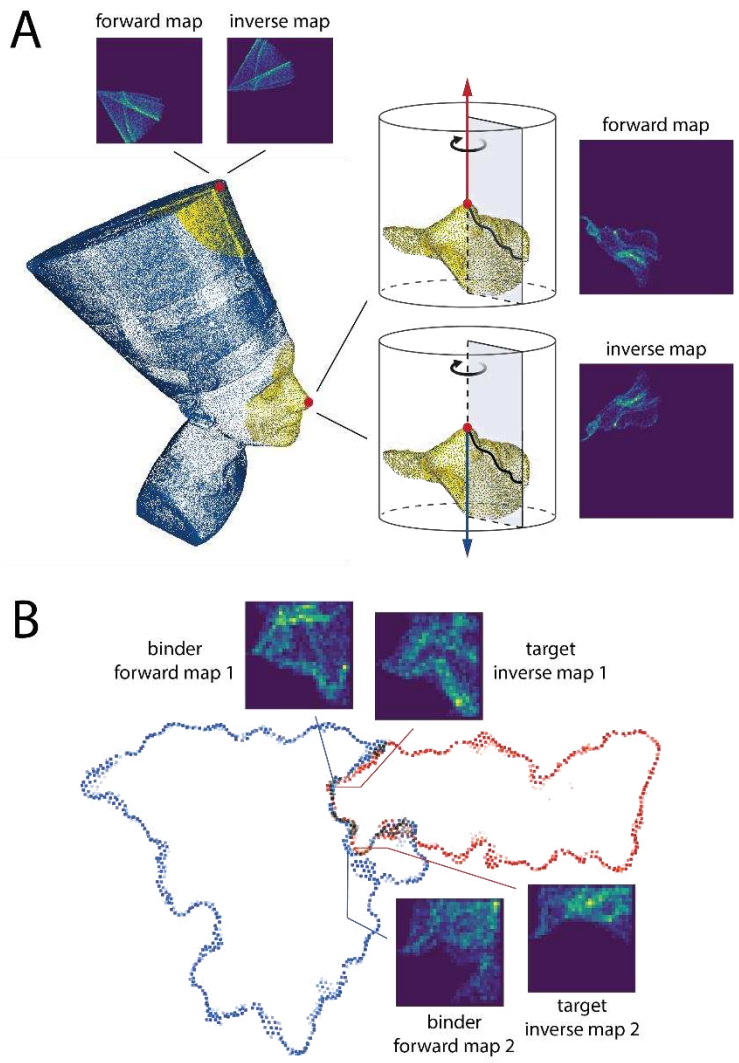
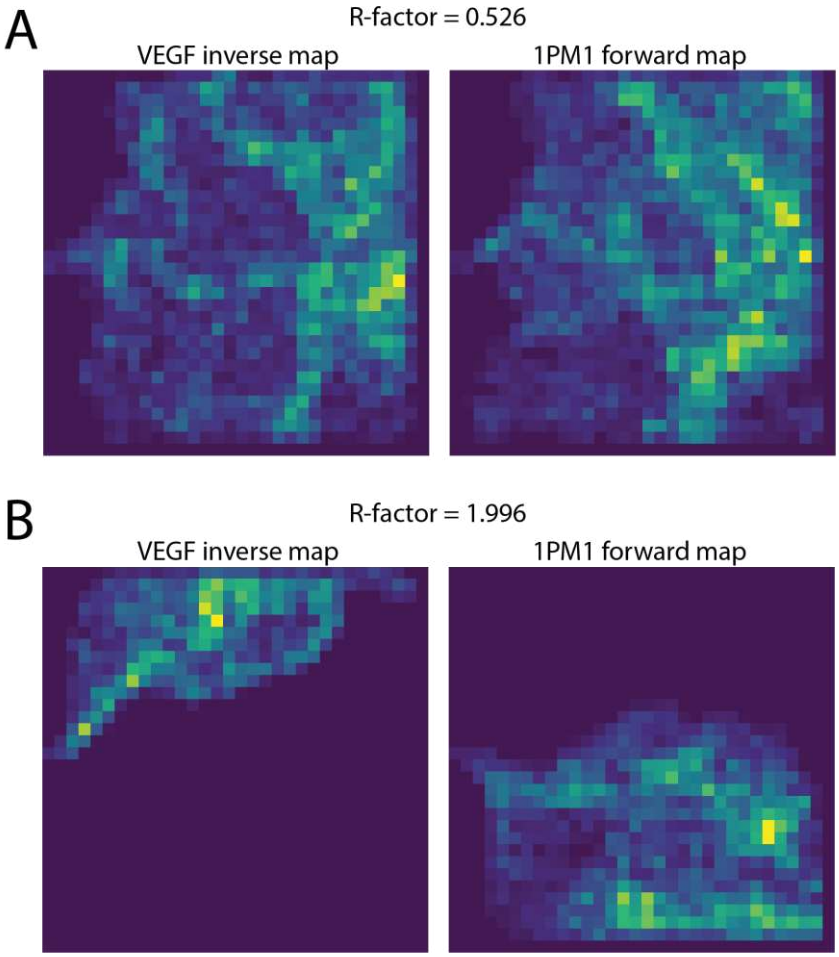


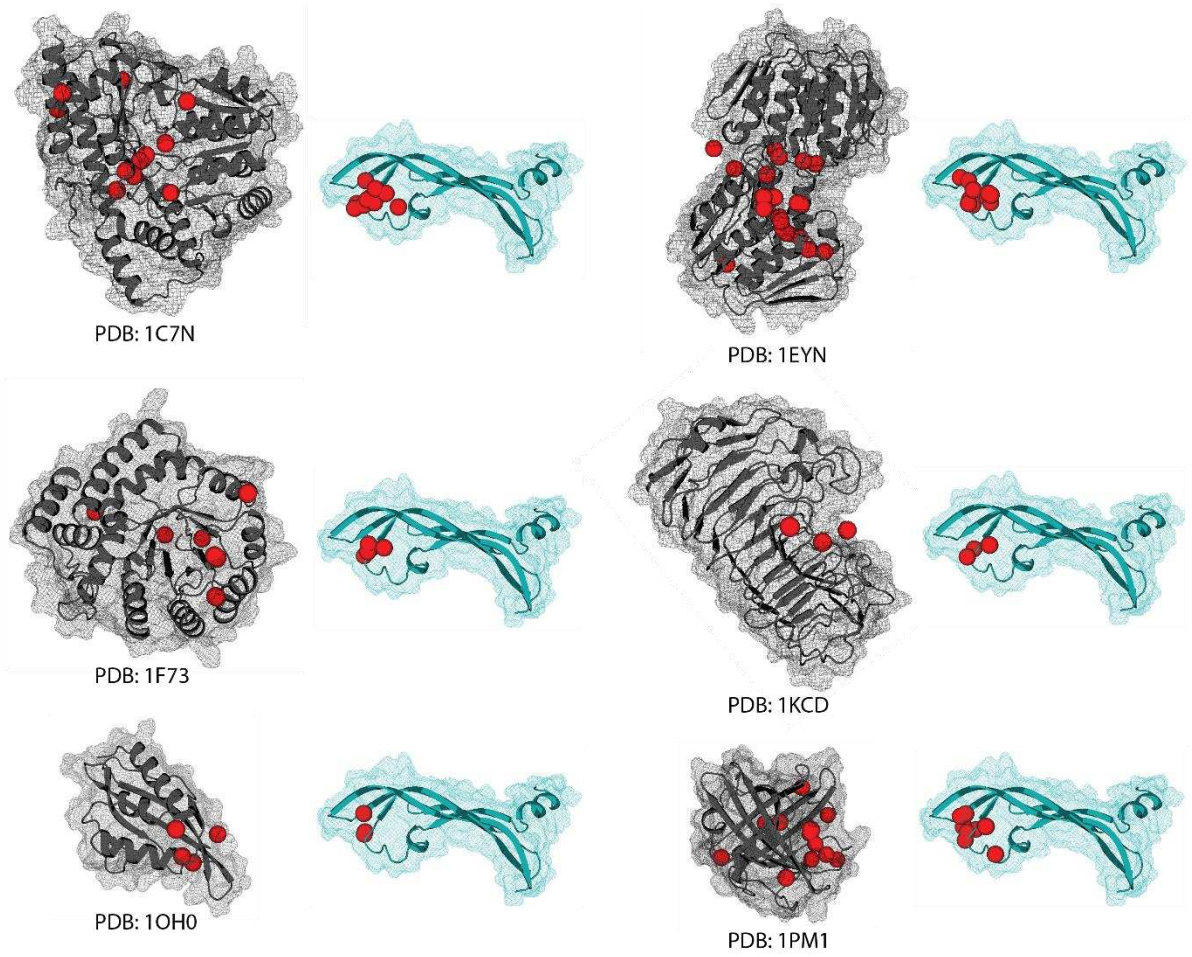
Figure S1. The HECTOR fingerprint captures local surface complementarity. (A) The HECTOR mapper takes a dot-surface of an object (e.g. molecular surface of a protein) as input. Surface patches are extracted around a selected dot (e.g. red dot at the centre of the golden face and crown patches from Nefertiti Bust). A map is compiled by rotational accumulation of dot density on a matrix. Depending if the surface normal pointing outwards or inwards of an object, the map is defined as "forward" or "inverse". **(B)** Forward and inverse maps at highly complementary interfaces are highly similar. Shown is an example of a tight interface between VEGF (red surface) and its antibody (bevacizumab; blue surface).

Figure S2



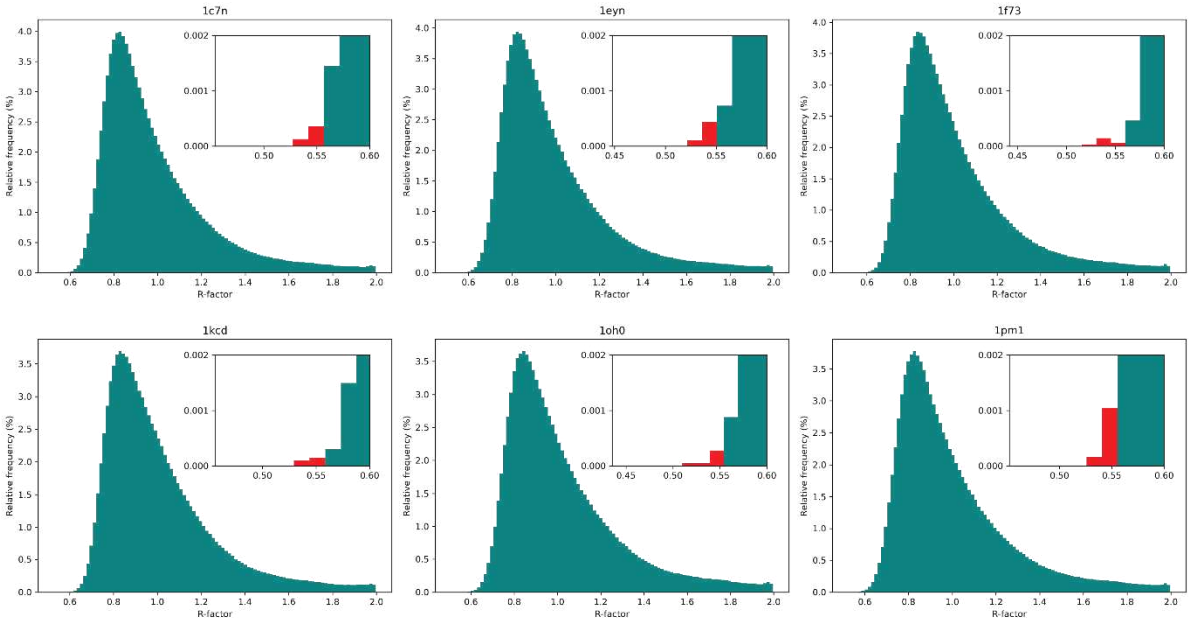
Supplementary figure 2. Representative fingerprints for surface patches with low and high R-factor. Inverse fingerprints (maps) for the surface patches at the receptor-binding site of VEGF were compared with the forward fingerprints for surface patches of a potential scaffold (PDB: 1PM1). **(A)** An example of two fingerprints corresponding to two surface patches with high shape complementarity (low *R*-factor). **(B)** An example of two fingerprints corresponding to two surface patches with low shape complementarity (high *R*-factor).

Figure S3



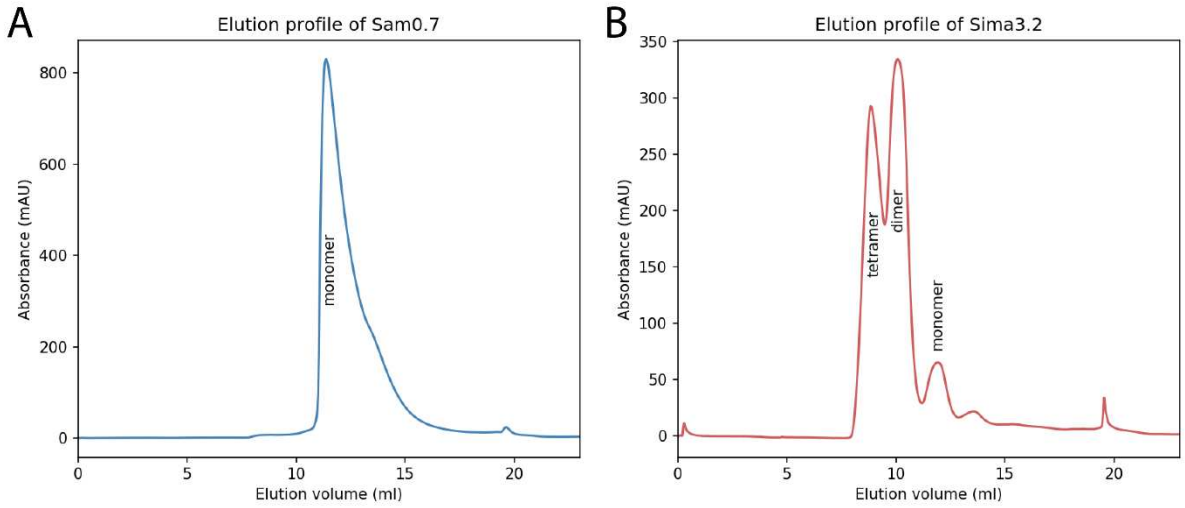
Supplementary figure S3. PDB hits with high shape complementarity to the receptor-binding site of VEGF. The inverse fingerprints corresponding to the surface patches within a binding site of VEGF (patches with y-coordinate > -5) were compared with the fingerprints of six HECTOR hits. Patches with R-factor lower than 0.55 are shown as red spheres. Coordinates of a sphere center correspond to coordinates of a patch center.

Figure S4



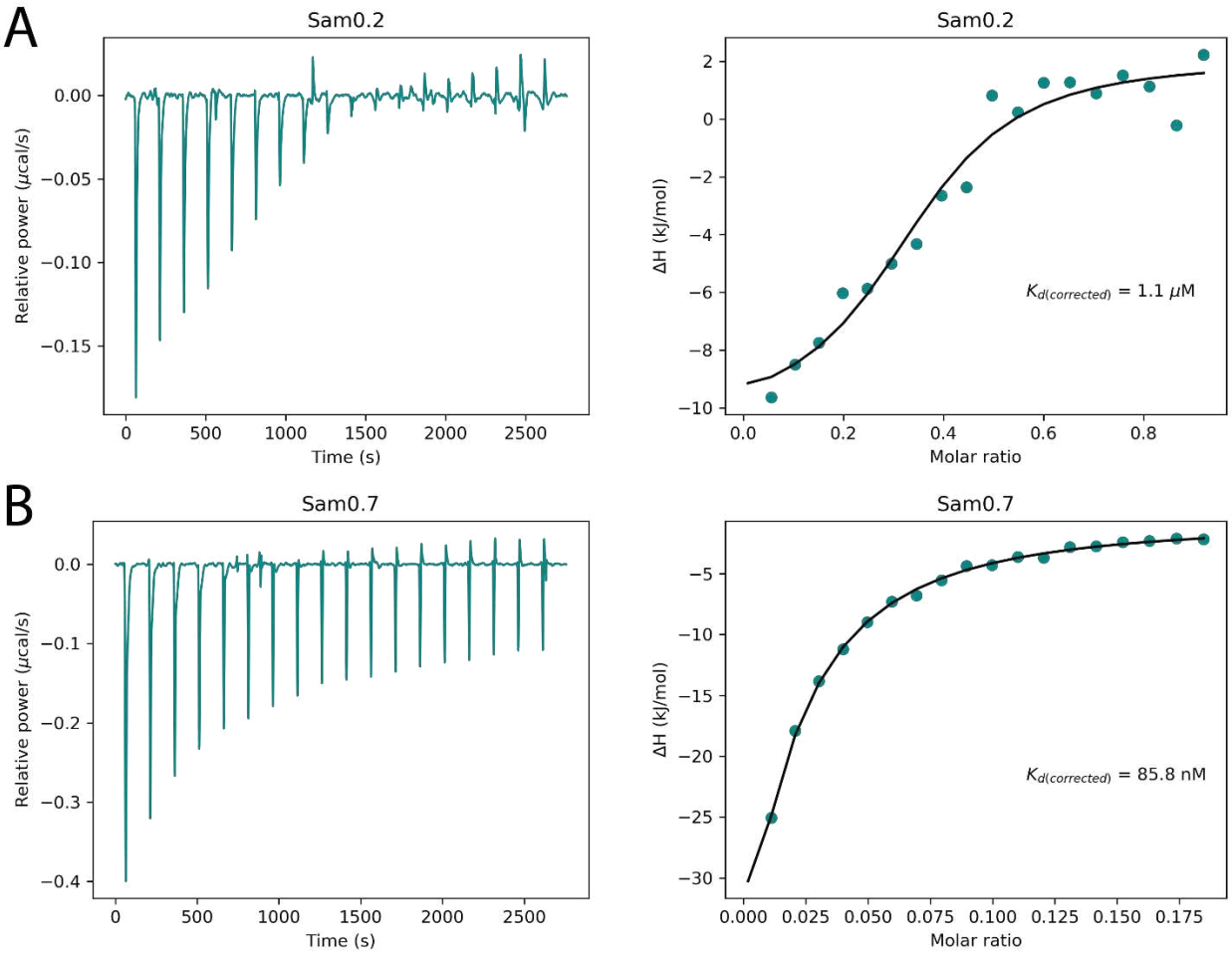
Supplementary figure 4. R-factor distribution for surface patches of PDB hits with high shape complementarity to the receptor-binding site of VEGF. Data are presented as histograms of R-factor values derived from comparison between the fingerprints of six HECTOR hits and the inverse fingerprints corresponding to the surface patches within a binding site of VEGF. Data points with R-factor values lower than 0.55 are colored in red.

Figure S5



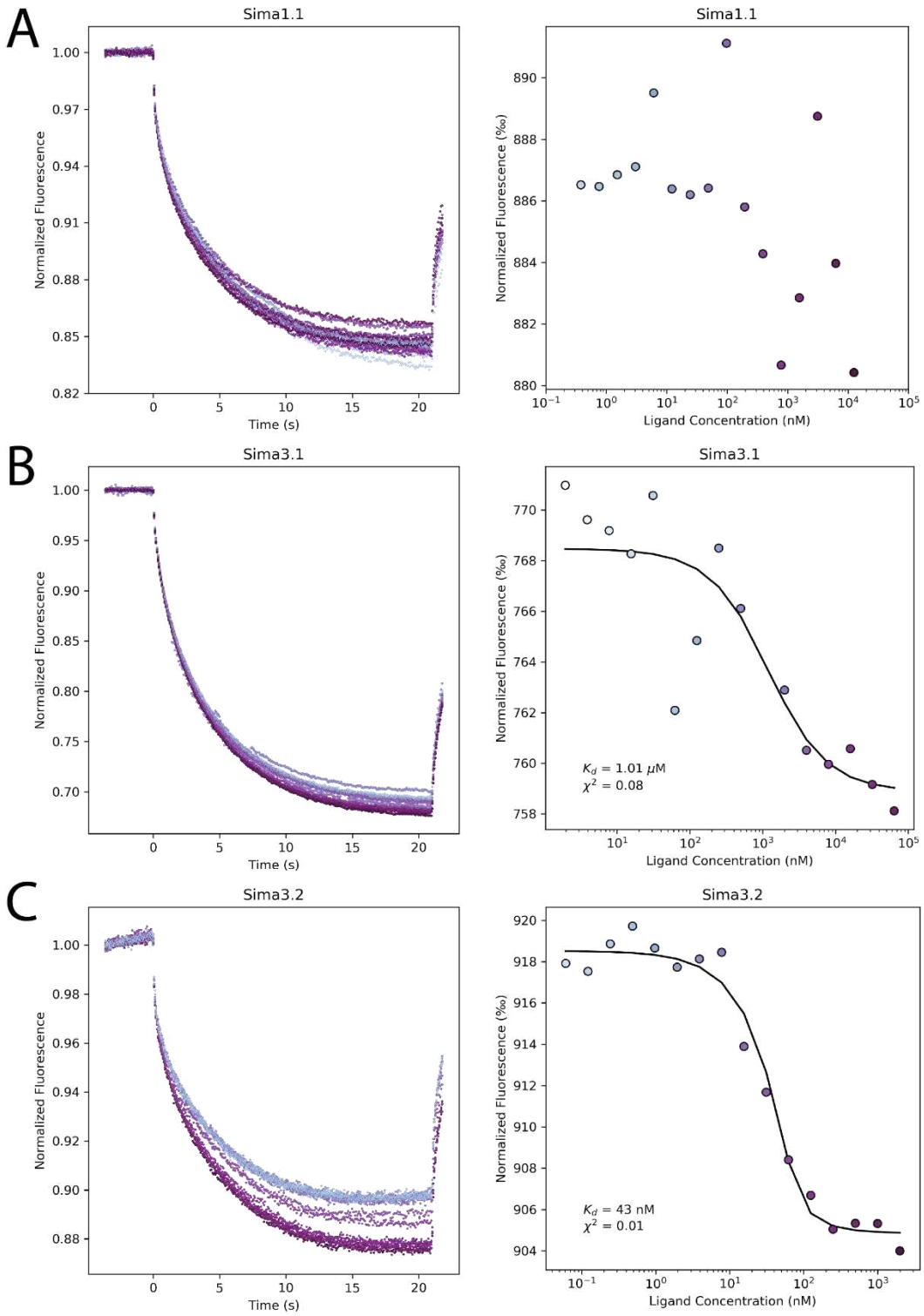
Supplementary figure 5. Size-exclusion elution profile of the designs. (A) The chromatogram shows Sam0.7 to elute as a monomeric protein. **(B)** Elution profile of Sima3.2 shows monomeric, dimeric, and tetrameric species. For experimental characterization, dimeric fraction of Sima3.2 was used.

Figure S6



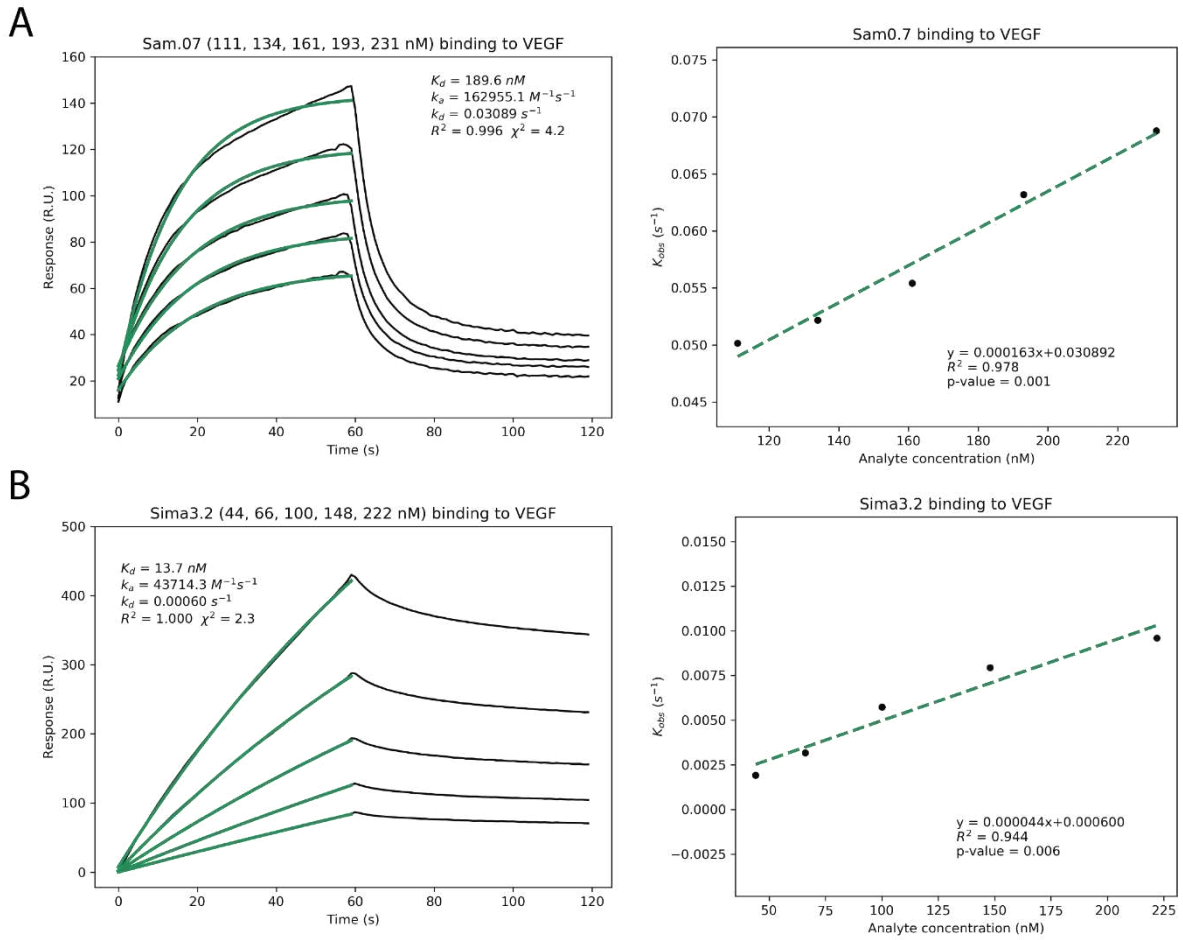
Supplementary figure 6. Calorigrams (left) and corresponding fitted curves (right) obtained from isothermal titration of VEGF with Sam0.2 (A) and Sam0.7 (B).

Figure S7



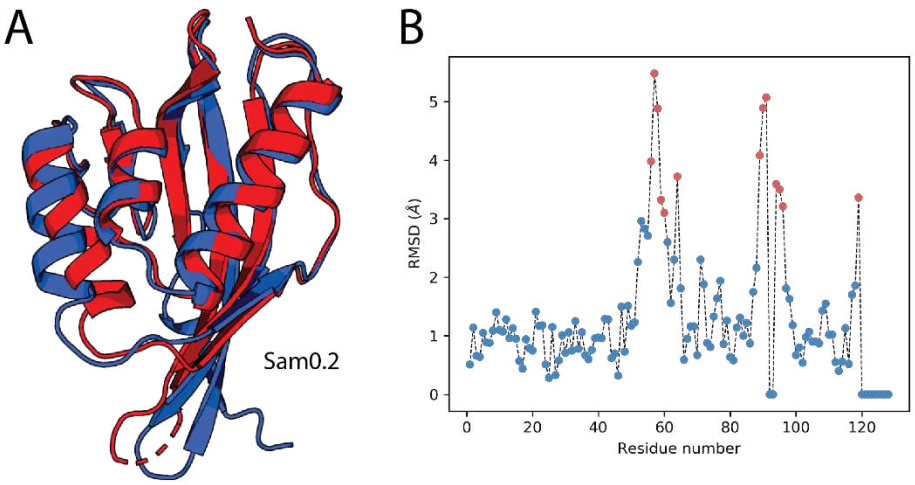
Supplementary figure 7. Microscale thermophoresis traces (left) and dose-response curves (right) for the binding interaction between VEGF and Sima designs.

Figure S8



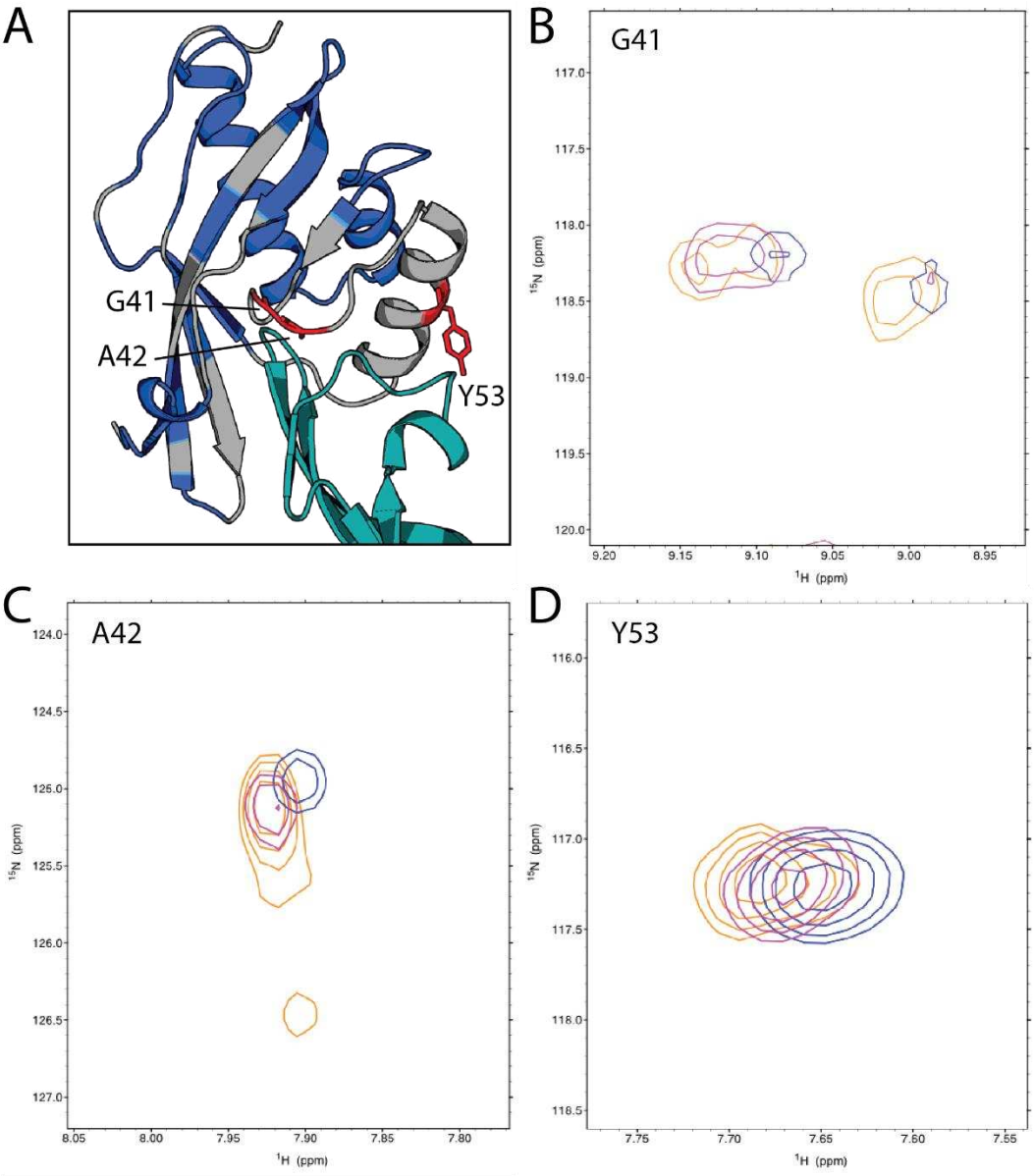
Supplementary figure 8. SPR sensorgrams of (A) Sam0.7 and (B) Sima3.2 and their binding kinetics fits. Sensograms (shown also in Fig. 3C) and association phase fits are shown (left-side panes; data points: black dots, fits: green curves) against their respective k_{obs} fits (right-side panes).

Figure S9



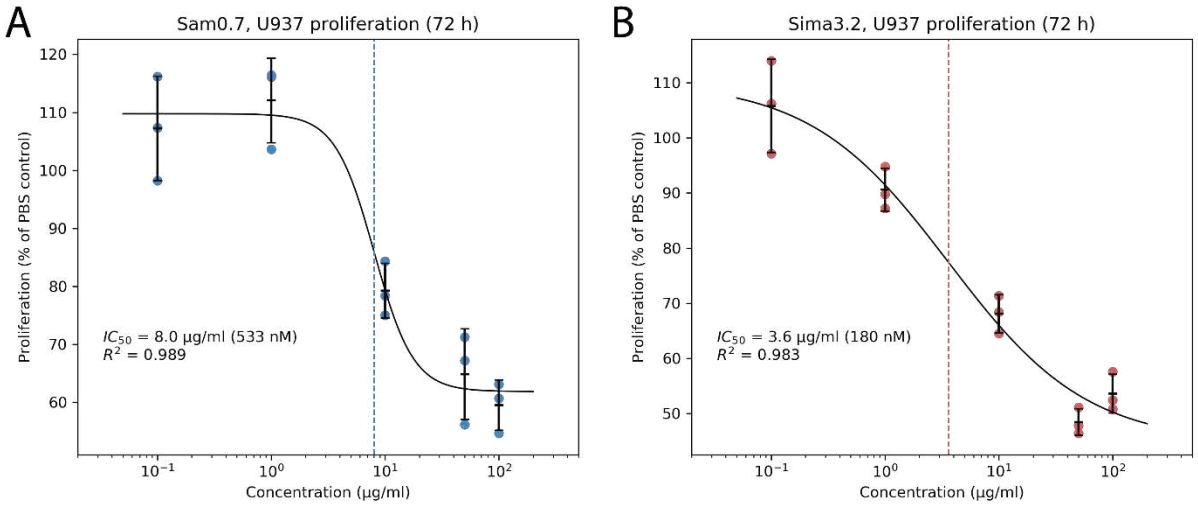
Supplementary figure 9. Crystal structure of Sam0.2 matches the computational design model with atomic-level accuracy. (A) Superimposition of the Sam0.2 design model (blue) and the experimentally determined crystal structure (red). **(B)** RMSD between coordinates of the C_α atoms in the Sam0.2 design model and corresponding C_α atoms in the crystal structure. RMSD values higher than 3Å are colored in red. Residues in gaps are assigned an RMSD value of 0.

Figure S10



Supplementary figure 10. VEGF titration causes NMR chemical shift perturbations of Sam0.2. (A) Sequential assignment of ^{15}N-HSQC peaks was achieved for majority of Sam0.2 residues (blue). Although many residues at the outer rim of the binding site could not be assigned (grey), peaks for some residues (red) showed significant chemical shift changes upon VEGF titration. **(B-D).** The overlaid 2D ^{1}H-^{15}N HSQC spectra for ^{15}N,^{13}C-labelled Sam0.2, residues G41, A42, Y53, in the absence (blue) or presence of VEGF at a molar ratio of 1:2 (magenta) and 1:3 (orange).

Figure S11



Supplementary figure 11. Proliferation assays using the VEGF-dependent U937 acute myeloid leukemia cell line. Sam0.7 **(A)** and Sima3.2 **(B)** suppressed proliferation of U937 cells with IC_{50} values of 533 nM and 180 nM, respectively. Error bars represent the standard deviation across three replicates.

Figure S12

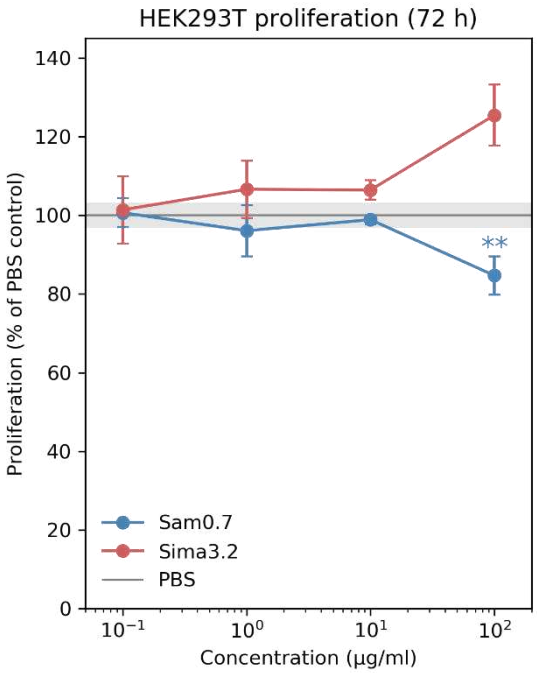
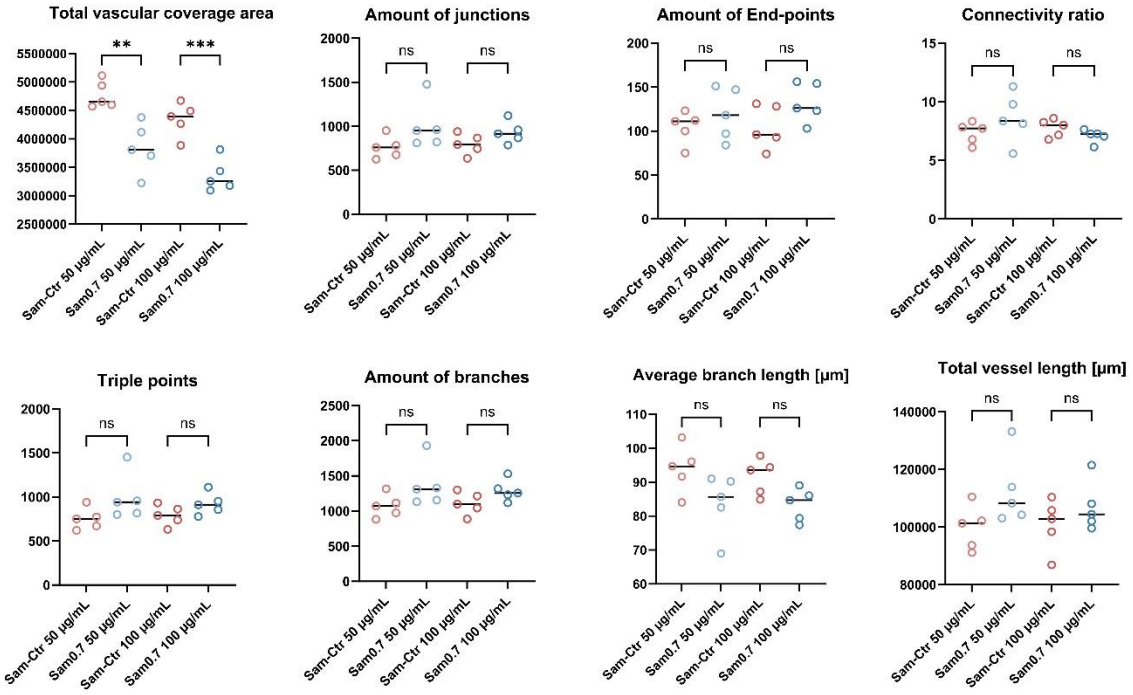


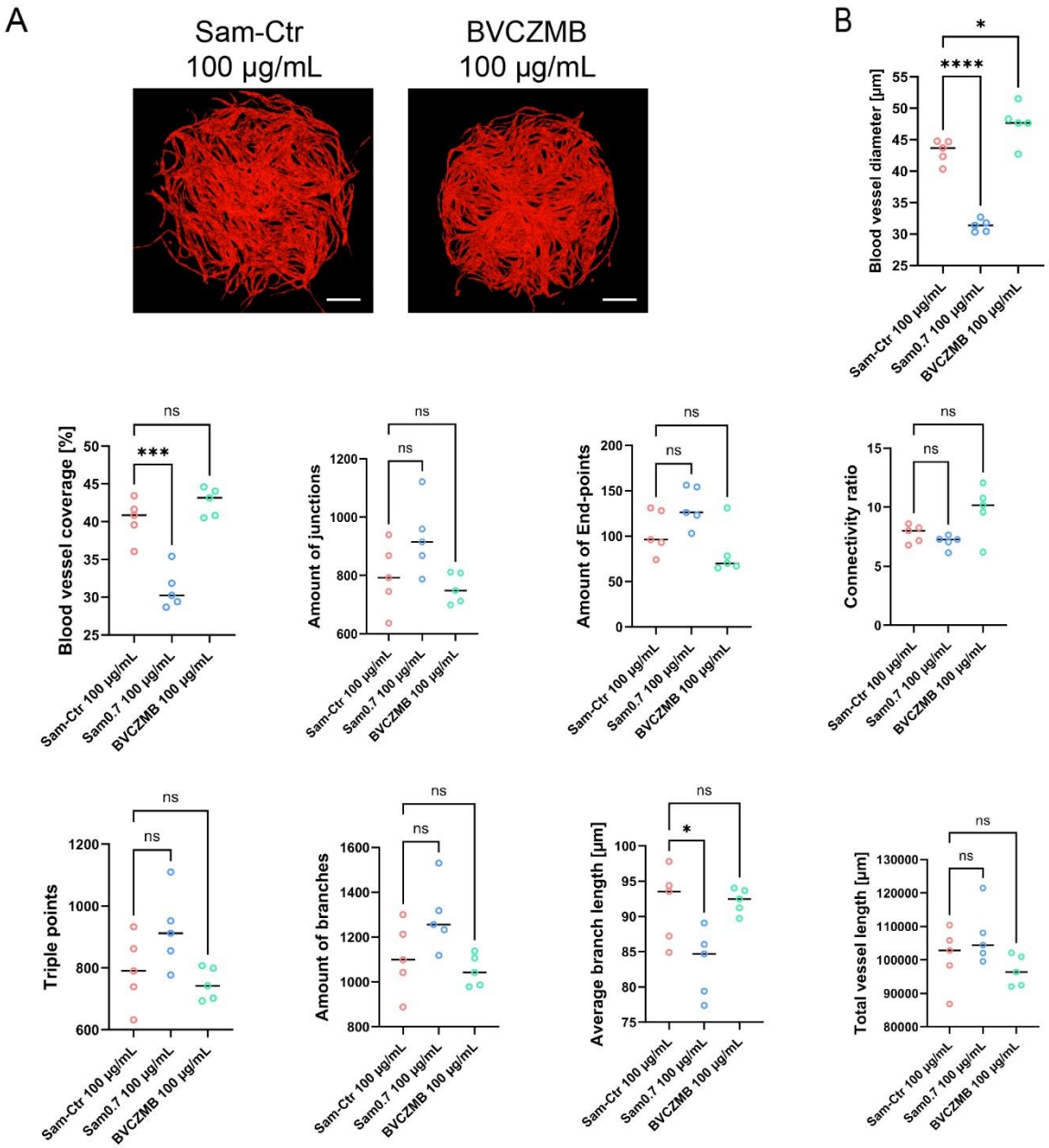
Figure S12. The designs did not show a strong inhibitory effect on proliferation of the human embryonic kidney cell line. Proliferation of HEK293T cells was mostly not affected by the designed binders. Only treatment with the highest concentration of Sam0.7 (100 µg/ml) could decrease the cell growth. Error bars represent the standard deviations across three replicates from one experiment. Statistical significance was calculated using Fisher's one-sided t-test (**, $p \leq 0.01$ vs. the PBS group).

Figure S13



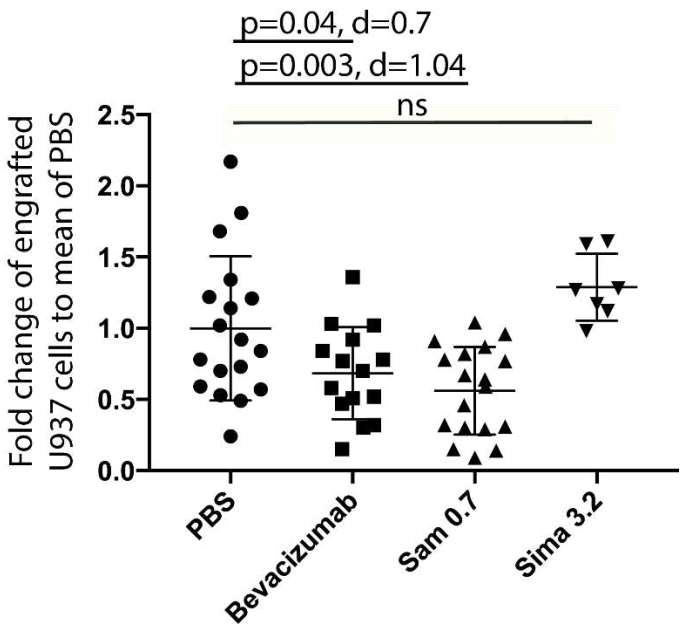
Supplementary figure S13. The effect of the designed binder Sam0.7 on microvasculature formation *in vitro*. Quantification of microvasculature parameters of capillaries treated with Sam0.7 or Sam_cntrl (as a negative control). The connectivity ratio is defined as the ratio of junctions to end-points. Statistical significance was calculated using the one-way ANOVA test (** $p \leq 0.01$, *** $p \leq 0.001$ treated Sam0.7 vs. Sam_cntrl group).

Figure S14



Supplementary figure S14. The effect of Bevacizumab on the microvasculature formation *in vitro*. **(A)** Representative images showing *in vitro* microvasculature formation in the presence of Bevacizumab or Sam_cntrl as a negative control at the same working concentrations (100 µg/mL). The scale bar is 500 µm. **(B)** Quantitative analysis of microvasculature formation. Statistical significance was calculated using the one-way ANOVA test (* $p \leq 0.05$, *** $p \leq 0.001$, **** $p \leq 0.0001$ treated Sam0.7 or Bevacizumab vs. Sam_cntrl group).

Figure S15



Supplementary figure 15. Evaluation of inhibitory activity of the designs in leukemia zebrafish xenograft. Quantification of the engrafted U937-GFP leukemia cells in zebrafish embryos that were injected with PBS, Bevacizumab as positive control, Sam 0.7, or Sima 3.2. Each dot indicates one embryo. p-value was calculated by Mann Whitney two tailed test. d – Cohen's d value.

Table S1. Protein sequences of Sam/Sima templates and the designed proteins

Name	Sequence	Global sequence identity with the initial template	Molecular weight (Da)
PDB 1OH0	LPTAQEVQGLMARYIELVDVGDIEAIVQMYADD ATVEDPFGQPPIHGREQIAAFYRQGLGGGKVRA CLTGPVRASHNGCGAMPFRVEMVWNGQPCALDV IDVMRFDEHGRIQTMQAYWSEVNLSVREP	-	14162
Sam0.1	LPTAQEVQGLMARYIELMDVGDIEAIVQMYADD ATVEAPFGAPPIHGRERIAYFYRRMLGGGIARA TLTGPVRASHNGTGAMPFRVEFVFNGQPYAMDV RVEMRFDEHGRIQTMQAYWSWVNLSVREP	109/128 (85 %)	14437
Sam0.2	LPTAEEVQKLMARYIELMDVGDIEAIVQMYADD ATVEAPFGAPPIHGRERIAYFYRRMLGEGIARA TLTGPVRASHNGTGAMPFRVEYVFNGQPYAMDV RVEMRFDEHGRIQTMQAYWSEVNDSVREP	106/128 (83 %)	14543
Sam0.3	LPTAEEVQKLMARYIELMDKGDIEAIVQMYADD ATVEAPFGAPPIHGRERIAYFYRRMLGEGIARA TLTGPVRASHNGTGAMPFRVEFVFNGQPYAMDV RVEMRFDEHGRIQTMQAYWSWVNLSVREP	105/128 (82 %)	14611
Sam0.4	LPTAEEVQKLMARYIELLDVGDIEAIVQMYADD ATVEAPFGSPPIHGRERIAYFYRRMLGEGIARA TLTGPVRASHNGTGAMPFRVEYVLNGQPVAMDI RVEMRFDEHGRIQTMQAYWSWVNVSVREP	104/128 (81 %)	14498
Sam0.5	LPTAEEVQKLMARYIELMDVGDIEAIVQMYADD ATVEAPFGAPPIHGRERIAYFYRRMLGGGIARA TLTGPVRASHNGTGAMPFRVEYVFNGQPYAMDV RVEMRFDEHGRIQTMQAYWSWVNLSVREP	107/128 (84 %)	14526
Sam0.6	LPTAEEVQKLMARYIELMDVGDIEAIVQMYADD ATVEAPFGAPPIHGRERIAYFYRRMLGGGIARA TLTGPVRASHNGTGAMPFRVEFVFNGQPFAMDV RVEMRFDEHGRIQTMQAYWSWVNLSVREP	107/128 (84 %)	14494
Sam0.7	LPTAEEVQKLMARYIELMDVGDIEAIVQMYADD ATVEAPFGAPPIHGRERIAYFYRRMLGGGIARA TLTGPVRASHNGTGAMPFRVEYVFNGQPFAMDV RVEMRFDEHGRIQTMQAYWDWVNLSVREP	106/128 (83 %)	14538
Sam0.8	LPTAEEVQKLMARYIELMDVGDIEAIVQMYADD ATVEAPFGAPPIHGRERIAYFYRRMLGGGIARA TLTGPVRASHNGTGAMPFRVEYVFNGQPYAMDV RVEMRFDEHGRIQTMQAYWSWVNISVREP	106/128 (83 %)	14526
PDB 1PM1	DCSTNISPKQGLDKAKYFSGKWYVTHFLDKDPQ VTDQYCSSFTPRESDGTVKEALYHYNANKKTSF YNIGEGKLESSGLQYTAKYKTVDKKKAVLKEAD EKNSYTLTVLEADDSSALVHICVREGSKDLGDV	-	19894

	YTVLTHQKDAEPSAKVKSAVTQAGLQLSQFVGT KDLGCQYDDQFTSL		
Sima1.1	DCSTNISPKQGLDKAKYFSGKWYLTHVLIKDPV AVTQFCSSFTPRESDGTVKEAIYVYLAIKKTSE YAIGEGKLESSGLQYTATFKTVDKKKAVLKEWD ERYSYTITVLEADDSSALTHVCTREGSKDYGDY YHVLTHQKDAEPSAKVKSAVTQAGLQLSQFVGT KDLGCQYDDQFTSL	153/179 (85 %)	19957
Sima2.1	DCSTNISPKQGLDKAKYFSGKWYLTHVLIKDPV AVTQFCSSFTPRESDGTVKEAIYVYLAIKKTSE YAIGEGKLESSGLQYTATFKTVDKKKAVLKEWD ERYSYTITVLEADDSSALTHVCTREGSKDYGDY YMVLTHQKDAEPSAKVKSAVTQAGLQLSQFVGT KDLGCQYDDQFTSL	153/179 (85 %)	19951
Sima3.1	DCSTNISPKQGLDKAKYFSGKWYLTHVLIKDPK IVSQFCASFTPRESDGTVKIAVYLYLAIKKTSD YAIGEGKLESSGLQYTATSKVVDKKKAVLKELD ERHSYTVTVLEADDSSALTHICVREGSKDYGDY YLVLTHQKDAEPSAKVKSAVTQAGLQLSQFVGT KDLGCQYDDQFTSL	152/179 (85 %)	19781
Sima4.1	DCSTNISPKQGLDKAKYFSGKWYLTHVLIKDPV AVTQFCSSFTPRESDGTVKVAIYVYLAIKKTSE YAIGEGKLESSGLQYTATFKTVDKKKAVLKEFD ERHSYTITVLEADDSSALVHVCVREGSKDYGDY YNVLTHQKDAEPSAKVKSAVTQAGLQLSQFVGT KDLGCQYDDQFTSL	154/179 (86 %)	19835
Sima1.2	STNISPKQGLDKAKYFSGKWYLTHVLIKDPVAV TQFCSSFTPRESDGTVKEAIYVYLAIKKTSEYA IGEGKLESSGLQYTATFKTVDKKKAVLKEWDER YSYTITVLEADDSSALTHVTTREGSKDYGDYYH VLTHQKDAEPSAKVKSAVTQAGLQLSQFVGTKD LGCQYDDQFTSL	150/177 (85 %)	19737
Sima2.2	STNISPKQGLDKAKYFSGKWYLTHVLIKDPVAV TQFCSSFTPRESDGTVKEAIYVYLAIKKTSEYA IGEGKLESSGLQYTATFKTVDKKKAVLKEWDER YSYTITVLEADDSSALTHVTTREGSKDYGDYYM VLTHQKDAEPSAKVKSAVTQAGLQLSQFVGTKD LGCQYDDQFTSL	150/177 (85 %)	19731
Sima3.2	STNISPKQGLDKAKYFSGKWYLTHVLIKDPKIV SQFCASFTPRESDGTVKIAVYLYLAIKKTSDYA IGEGKLESSGLQYTATSKVVDKKKAVLKELDER HSYTVTVLEADDSSALTHITVREGSKDYGDYYL VLTHQKDAEPSAKVKSAVTQAGLQLSQFVGTKD LGCQYDDQFTSL	149/177 (84 %)	19561
Sima4.2	STNISPKQGLDKAKYFSGKWYLTHVLIKDPVAV TQFCSSFTPRESDGTVKVAIYVYLAIKKTSEYA IGEGKLESSGLQYTATFKTVDKKKAVLKEFDER	151/177 (85 %)	19615

	HSYTITVLEADDSSALVHVTVREGSKDYGDYYN VLTHQKDAEPSAKVKSAVTQAGLQLSQFVGTKD LGCQYDDQFTSL		

Table S2. Final protein yield after two-step purification

Name	Yield per litre of culture (mg)
Sam0.1	4.0
Sam0.2	5.1
Sam0.3	5.2
Sam0.4	6.0
Sam0.5	13.2
Sam0.6	protein could not be neither purified from a soluble fraction, nor refolded
Sam0.7	7.8
Sam0.8	2.8
Sima1.1	1.1
Sima1.2	0.7
Sima2.1	0.7
Sima2.2	0.3
Sima3.1	2.6
Sima3.2	1.2
Sima4.1	0.7
Sima4.2	0.4

Table S3. Crystallographic Data Collection and Refinement Statistics

Structure	**Sam0.2**	**Sam0.7**
Data collection		
Space group	$P6_322$	$P6_522$
Cell parameters a, b, c (*Å)*	*77.81, 77.81, 78.16*	*40.55, 40.55, 253.12*
Wavelength (*Å)*	1.000	1.000
Resolution limits (*Å)*[a]	39.08-2.65 (2.81-2.65)	35.12-1.80 (1.91-1.80)
Unique reflections	4410 (686)	12500 (1934)
Completeness (%)	100 (100)	100 (99.9)
Redundancy	36.70 (40.09)	32.77 (34.94)
I/σI	31.5 (1.36)	19.0 (1.65)
R_{merge} (%)	8.4 (294.7)	10.3 (162.3)
CC(1/2)	100 (72.7)	99.9 (96.5)
Refinement		
Resolution limits (*Å)*	39.08-2.65	35.12-1.80
R_{cryst} (%)	25.7	25.7
R_{free} (%)	29.8	29.6
Protein molecules / asymmetric unit	1	1
Mean B value (*Å²)*	110.5	50.0
PDB code	**8BL5**	**8BL9**

[a] Values in parenthesis refer to the highest-resolution shell.

Supplementary methods

PatchDock protocol:

```
import os
import sys

import subprocess as sp

#define software localisations
patchdock_location = 'PatchDock/'

# patchdock routines

def make_patchdock_parameter_file(receptor_name, ligand_name,
cluster_radius, outID, out_dict = out_docked_dict):

    outfile = '{}{}.params'.format(out_dict,outID)

    make_parm =
sp.Popen(["{}buildParamsToOutfile.pl".format(patchdock_location),
target_name, fragment_name, outfile, str(cluster_radius)],
stdout=sp.PIPE, stderr=sp.PIPE, stdin=sp.PIPE)
    stdout, stderr = make_parm.communicate()

    return outfile

def create_active_site_file(molecule_name, chain, active_site):

    active_site_file = '{}_active_site.txt'.format(molecule_name)
    ofile = open(active_site_file, "w")

    for residue in active_site:
        ofile.write("{} {}\n".format(residue, receptor_chain))
    ofile.close()

    return active_site_file

def add_active_site_to_patchdock_param_file(parameter_file,
active_site_file, mode = 'receptor'):

    with open(parameter_file, 'r') as pfile:
        data = pfile.readlines()
    pfile.close()

    for i in range(len(data)):
        if '#{}ActiveSite'.format(mode) in data[i]:
            data[i] = '{}ActiveSite {}\n'.format(mode,
active_site_file)

    with open(parameter_file, 'w') as pfile:
        pfile.writelines(data)

    pfile.close()
    return
```

```
def generate_empty_patchdock_outfile(receptor_name, ligand_name,
outID):

    # Shall generate a provisional patchdock outfile with the same
parameters as
    # those default from PatchDock and where no transformation is
applied to the ligand

    patchdock_outfile = '{}{}.dock_out'.format(out_docked_dict,outID)

    with open(patchdock_outfile, 'w') as outfile:


outfile.write('************************************************************
*\n')
        outfile.write('Program parameters\n\n')
        outfile.write('baseParams    (Str)   4.0 13.0 2\n')
        outfile.write('clusterParams (Str)   0.1 4 2.0 4.0\n')
        outfile.write('desolvationParams     (Str)   500.0 1.0\n')
        outfile.write('ligandGrid    (Str)   0.5 6.0 6.0\n')
        outfile.write('ligandMs      (Str)   10.0 1.8\n')
        outfile.write('ligandPdb     (Str)   {}\n'.format(ligand_name))
        outfile.write('ligandSeg     (Str)   10.0 20.0 1.5 1 0 1 0\n')
        outfile.write('log-file      (Str)   patch_dock.log\n')
        outfile.write('log-level     (Str)   2\n')
        outfile.write('matchAlgorithm        (Str)   1\n')
        outfile.write('matchingParams        (Str)   1.5 1.5 0.4 0.5
0.9\n')
        outfile.write('protLib       (Str)   PatchDock/chem.lib\n')
        outfile.write('receptorGrid  (Str)   0.5 6.0 6.0\n')
        outfile.write('receptorMs    (Str)   10.0 1.8\n')
        outfile.write('receptorPdb   (Str)   {}\n'.format(receptor_name))
        outfile.write('receptorSeg   (Str)   10.0 20.0 1.5 1 0 1 0\n')
        outfile.write('scoreParams   (Str)   0.3 -5.0 0.5 0.0 0.0 1500 -8
-4 0 1 0\n\n')

outfile.write('************************************************************
*\n\n')
        outfile.write('   # | score | pen.  | Area    | as1   | as2   |
as12  | ACE     | hydroph | Energy  |cluster| dist. || Ligand
Transformation\n')
        outfile.write('   1 | 00000 | 00000 | 0000000 | 00000 | 00000 |
00000 | 0000000 | 0000000 | 0000000 | 00000 | 00000 || 0.0 0.0 0.0 0.0
0.0 0.0\n')

    return patchdock_outfile


def dock(receptor_name, ligand_name, outID, cluster_radius = 4.0,
add_active_site = 'Yes', receptor_binding_site = None,
ligand_binding_site = None):

    # prepare parameter file
    parameter_file =  make_patchdock_parameter_file(receptor_name,
ligand_name, cluster_radius = cluster_radius, outID = outID, out_dict =
out_docked_dict)

    if add_active_site == 'Yes':
```

```
        receptor_active_site_file =
create_active_site_file(receptor_name.split('.')[0], receptor_chain =
'A', active_site = receptor_binding_site)
        add_active_site_to_patchdock_param_file(parameter_file,
receptor_active_site_file)

        ligand_active_site_file =
create_active_site_file(ligand_name.split('.')[0], receptor_chain =
'A', active_site = ligand_binding_site)
        add_active_site_to_patchdock_param_file(parameter_file,
ligand_active_site_file)

    # dock fragment to target
    patchdock_outfile = '{}{}.dock_out'.format(out_docked_dict,outID)
    run_patchdock =
sp.Popen(["{}patch_dock.Linux".format(patchdock_location),
parameter_file, patchdock_outfile], stdout=sp.PIPE, stderr=sp.PIPE,
stdin=sp.PIPE)
    stdout, stderr = run_patchdock.communicate()

    # delete parameters file
    delete_params = sp.Popen(["rm", parameter_file], stdout=sp.PIPE,
stderr=sp.PIPE, stdin=sp.PIPE)
    stdout, stderr = delete_params.communicate()

    return patchdock_outfile



# ----------------------
# MAIN

receptor = sys.argv[1]
ligand   = sys.argv[2]
receptor_binding_site = sys.argv[3].split(',')
ligand_binding_site = sys.argv[4].split(',')

patchdock_outfile = dock(receptor, ligand, 'HECTOR', cluster_radius =
4.0, add_active_site = 'Yes', receptor_binding_site =
receptor_binding_site, ligand_binding_site = ligand_binding_site)
```

One-sided interface design protocol:

```
<ROSETTASCRIPTS>
    <TASKOPERATIONS>
        <ReadResfile name="rrf" filename="%%resfile%%"/>
        <ProteinInterfaceDesign name="pid" repack_chain1="1"
repack_chain2="1" design_chain1="1" design_chain2="0"
interface_distance_cutoff="10"/>
        <IncludeCurrent name="currentTask"/>
        <RestrictToRepacking name="repackonly"/>
        <RestrictToInterfaceVector name="vectorTask" chain1_num="1"
chain2_num="2" CB_dist_cutoff="10.0" nearby_atom_cutoff="6.0"
vector_angle_cutoff="65.0" vector_dist_cutoff="8.0"/>
        <InitializeFromCommandline name="cmdTask"/>
    </TASKOPERATIONS>

    <FILTERS>
        <PackStat name="holes_1" threshold="%%pck_scr1%%" chain="0"
repeats="5"/>
        <PackStat name="holes_2" threshold="%%pck_scr2%%" chain="0"
repeats="5"/>
        <Ddg name="ddG" scorefxn="talaris2013" repack="true"
threshold="-20" repack_bound="true" repeats="2"/>
        <EnergyPerResidue name="nrgy_per_res" scorefxn="talaris2013"
score_type="total_score" whole_interface="1" jump_number="1"
interface_distance_cutoff="8.0" bb_bb="1"/>
        <ScoreType name="ttl_scr" scorefxn="talaris2013"
score_type="total_score" threshold="%%ttl_scr_thrshld%%"/>
    </FILTERS>

    <MOVERS>
        <DockingProtocol name="dock" docking_local_refine="1"
docking_score_high="soft_rep" ignore_default_docking_task="1"
dock_min="1"
task_operations="vectorTask,cmdTask,repackonly,currentTask"/>
        <TaskAwareMinMover name="minmover_rpck" scorefxn="soft_rep"
chi="1" bb="1" jump="1"
task_operations="vectorTask,cmdTask,currentTask,repackonly"/>
        <GreedyOptMutationMover name="grdy_opt_mut" filter="ddG"
filter_delta="0.5" scorefxn="soft_rep" relax_mover="minmover_rpck"
sample_type="low" repack_shell="7.5" task_operations="rrf"/>
        <Backrub name="backrub"/>
        <BackrubDD name="backrubdd" partner1="1" partner2="0"
interface_distance_cutoff="8.0" moves="1000" sc_move_probability="0.2"
scorefxn="talaris2013" small_move_probability="0.3"
bbg_move_probability="0.4" task_operations="rrf"/>
        <RepackMinimize name="des1" scorefxn_repack="soft_rep"
scorefxn_minimize="soft_rep" minimize_bb="0" task_operations="rrf"
design_partner1="1" design_partner2="0"/>
        <RepackMinimize name="des2" scorefxn_repack="talaris2013"
scorefxn_minimize="talaris2013" minimize_bb="0" design_partner1="1"
design_partner2="0" task_operations="rrf"/>
        <RepackMinimize name="des3" design_partner1="1"
design_partner2="1" minimize_bb="0" task_operations="rrf"/>
        <FastRelax name="relax" scorefxn="talaris2013" repeats="2"
task_operations="currentTask,repackonly,cmdTask"/>
        <ParsedProtocol name="design">
            <Add mover_name="des1"/>
```

```
            <Add mover_name="backrubdd"/>
            <Add mover_name="des1"/>
            <Add mover_name="des2"/>
            <Add mover_name="des2"/>
            <Add mover_name="backrubdd"/>
            <Add mover_name="dock"/>
            <Add mover_name="des3"/>
            <Add filter="holes_1"/>
            <Add mover_name="backrub"/>
            <Add mover_name="des3"/>
            <Add filter="holes_2"/>
        </ParsedProtocol>
        <GenericMonteCarlo name="iterate" filter_name="ddG"
sample_type="low" scorefxn_name="talaris2013" mover_name="design"
trials="3"/>
        <GenericMonteCarlo name="iterate_h" filter_name="holes_1"
scorefxn_name="talaris2013" mover_name="design" trials="2"/>
    </MOVERS>

    <PROTOCOLS>
        <Add mover="iterate"/>
        <Add filter="holes_1"/>
        <Add mover="iterate_h"/>
        <Add filter="holes_2"/>
        <Add filter="ttl_scr"/>
        <Add filter="ddG"/>
    </PROTOCOLS>
</ROSETTASCRIPTS>


<!— This script was invoked as:

for fn in *.pdb; do for e in `seq 001 100`; do qsub -l h_vmem=2G -l
h_rt=15:0:0 -cwd -N des_fat -o ./ -b y
~/rstta_bin/rosetta_scripts.static.linuxgccrelease -database ~/rstta_db
-s gq01_0001.pdb -docking:dock_pert 1 2 -nstruct 1 -out:prefix $e -
out:file:silent silent_$e -out:file:scorefile score_$e -parser:protocol
../intrfc_fixbb_10.xml -parser:script_vars resfile='rf_gq01'
pck_scr1='0.35' pck_scr2='0.45' ttl_scr_thrshld='-150.0' -
ignore_zero_occupancy false -mute all -ex1 -restore_talaris_behavior;
done; done -->
```

This protocol was run for 4-6 rounds in series with successively higher thresholds, where the output of each round was filtered by rosettaholes score and ddG and forwarded as input of the following round.

The resfiles of the two tested designs were as follows:

- Sam resfile (input template: 1OH0):

```
NATAA
start

210 A PIKAA EQ EX 1 LEVEL 3 EX 2 LEVEL 3 EX 3 LEVEL 1
211 A POLAR EX 1 LEVEL 3 EX 2 LEVEL 3 EX 3 LEVEL 1
214 A PIKAA RKQ EX 1 LEVEL 3 EX 2 LEVEL 3 EX 3 LEVEL 1
220 A APOLAR EX 1 LEVEL 3 EX 2 LEVEL 3
222 A POLAR EX 1 LEVEL 3 EX 2 LEVEL 3
238 A APOLAR EX 1 LEVEL 3 EX 2 LEVEL 3 EX 3 LEVEL 1
```

```
240 A APOLAR EX 1 LEVEL 3 EX 2 LEVEL 3 EX 3 LEVEL 1
242 A PIKAA YR EX 1 LEVEL 3 EX 2 LEVEL 3 EX 3 LEVEL 1
243 A POLAR EX 1 LEVEL 3
244 A ALLAAxc EX 1 LEVEL 3
245 A ALLAAxc EX 1 LEVEL 3
246 A ALLAAxc EX 1 LEVEL 3
257 A ALLAAxc EX 1 LEVEL 3
259 A ALLAAxc EX 1 LEVEL 3
261 A PIKAA LM EX 1 LEVEL 3 EX 2 LEVEL 3 EX 3 LEVEL 1
263 A PIKAA QENDY EX 1 LEVEL 3 EX 2 LEVEL 3 EX 3 LEVEL 1
264 A POLAR EX 1 LEVEL 3 EX 2 LEVEL 3
266 A ALLAAxc EX 1 LEVEL 3
268 A ALLAAxc EX 1 LEVEL 3
285 A POLAR EX 1 LEVEL 3 EX 2 LEVEL 3
288 A APOLAR EX 1 LEVEL 3 EX 2 LEVEL 3
290 A ALLAAxc EX 1 LEVEL 3
291 A ALLAAxc EX 1 LEVEL 3 EX 2 LEVEL 3
292 A ALLAAxc EX 1 LEVEL 3
293 A ALLAAxc EX 1 LEVEL 3
294 A ALLAAxc EX 1 LEVEL 3
297 A ALLAAxc EX 1 LEVEL 3 EX 2 LEVEL 3 EX 3 LEVEL 3
298 A ALLAAxc EX 1 LEVEL 3
301 A APOLAR EX 1 LEVEL 3 EX 2 LEVEL 3
303 A APOLAR EX 1 LEVEL 3 EX 2 LEVEL 3
318 A ALLAAxc EX 1 LEVEL 3
321 A ALLAAxc EX 1 LEVEL 3
322 A PIKAA EYLIW EX 1 LEVEL 3
325 A ALLAAxc EX 1 LEVEL 3 EX 2 LEVEL 3 EX 3 LEVEL 1
326 A ALLAAxc EX 1 LEVEL 3
```

- Sima resfile (intput template: 1PM1):

```
NATAA
start

 24 A ALLAAxc
 27 A ALLAAxc
 29 A ALLAAxc
 31 A ALLAAxc
 32 A ALLAAxc
 33 A ALLAAxc
 34 A ALLAAxc
 36 A ALLAAxc
 38 A ALLAAxc
 40 A ALLAAxc
 53 A ALLAAxc
 55 A ALLAAxc
 57 A ALLAAxc
 59 A ALLAAxc
 62 A ALLAAxc
 66 A ALLAAxc
 67 A ALLAAxc
 68 A ALLAAxc
 70 A ALLAAxc
 84 A ALLAAxc
 85 A ALLAAxc
 87 A ALLAAxc
 96 A ALLAAxc
 98 A ALLAAxc
```

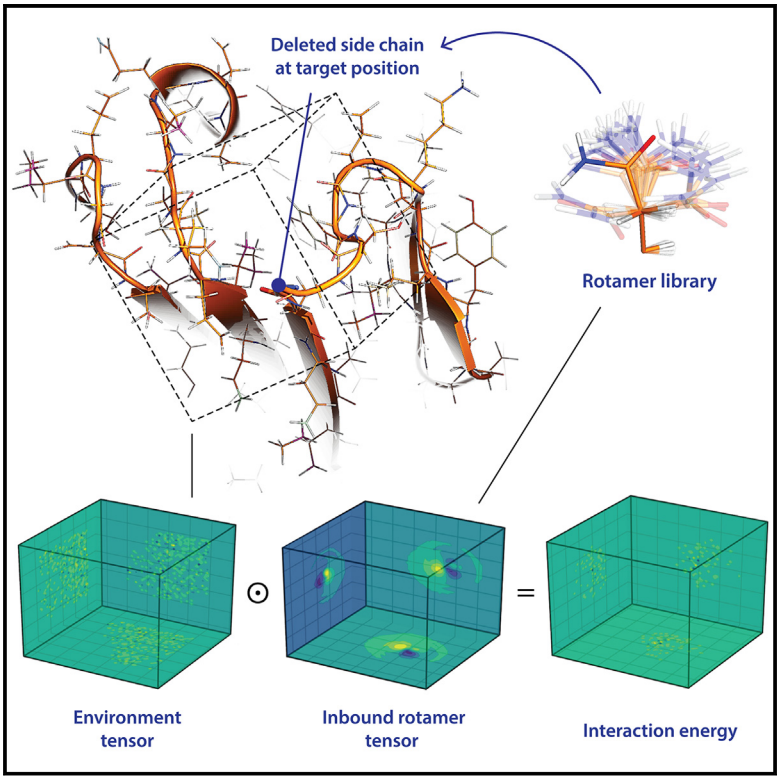
```
100 A ALLAAxc
102 A ALLAAxc
104 A ALLAAxc
106 A ALLAAxc
118 A ALLAAxc
119 A ALLAAxc
120 A ALLAAxc
122 A ALLAAxc
128 A ALLAAxc
129 A ALLAAxc
130 A ALLAAxc
132 A ALLAAxc
134 A ALLAAxc
136 A ALLAAxc
```

Cell Reports
Methods

Article

The design of functional proteins using tensorized energy calculations

Graphical abstract



Highlights

- Tensorized energy calculations accelerate protein design
- A rotamer library can cover residues and ligands not present in structure databases
- The design framework does not rely on any training data
- This framework is applicable to diverse protein design problems

Authors

Kateryna Maksymenko, Andreas Maurer, Narges Aghaallaei, ..., Andrei N. Lupas, Julia Skokowa, Mohammad ElGamacy

Correspondence

mohammad.elgamacy@med.uni-tuebingen.de

In brief

Maksymenko et al. present an approach for protein design that accelerates interaction energy calculations, generates a rotamer library relying only on force-field parameters, and offers a training-free scoring function. The authors apply this framework to design proteins with EGFR-inhibiting and radio-tracing functions.

Maksymenko et al., 2023, Cell Reports Methods *3*, 100560
August 28, 2023
https://doi.org/10.1016/j.crmeth.2023.100560

CellPress

Cell Reports Methods

CellPress
OPEN ACCESS

Article

The design of functional proteins using tensorized energy calculations

Kateryna Maksymenko,[1,2] Andreas Maurer,[3,4] Narges Aghaallaei,[5,9] Caroline Barry,[1,6,10] Natalia Borbarán-Bravo,[5] Timo Ullrich,[1,2] Tjeerd M.H. Dijkstra,[1,7,8] Birte Hernandez Alvarez,[1] Patrick Müller,[2,11] Andrei N. Lupas,[1] Julia Skokowa,[5] and Mohammad ElGamacy[1,2,5,12,*]

[1]Department of Protein Evolution, Max Planck Institute for Biology, 72076 Tübingen, Germany
[2]Friedrich Miescher Laboratory of the Max Planck Society, 72076 Tübingen, Germany
[3]Werner Siemens Imaging Center, Department of Preclinical Imaging and Radiopharmacy, Eberhard Karls University, 72076 Tübingen, Germany
[4]Cluster of Excellence iFIT (EXC 2180) "Image Guided and Functionally Instructed Tumor Therapies," Eberhard Karls University, 72076 Tübingen, Germany
[5]Division of Translational Oncology, University Hospital Tübingen, 72076 Tübingen, Germany
[6]Krieger School of Arts and Sciences, Johns Hopkins University, Washington, DC 20036, USA
[7]Department for Women's Health, University Hospital Tübingen, 72076 Tübingen, Germany
[8]Translational Bioinformatics, University Hospital Tübingen, 72072 Tübingen, Germany
[9]Present address: Ludwig Boltzmann Institute for Hematology and Oncology, Medical University of Vienna, 1090 Vienna, Austria
[10]Present address: Institute of Molecular Biology, 55128 Mainz, Germany
[11]Present address: Department of Biology, University of Konstanz, 78464 Konstanz, Germany
[12]Lead contact
*Correspondence: mohammad.elgamacy@med.uni-tuebingen.de
https://doi.org/10.1016/j.crmeth.2023.100560

MOTIVATION Calculating the interaction energies between atoms is a central process in biomolecular simulations. Traditionally, these calculations are performed exhaustively for each atom pair, which constitutes the computational bottleneck. In this study, we introduce a framework that instead represents the dense atomic interaction fields as three-dimensional projections. These projections can condense energy evaluations into a single matrix operation, greatly simplifying the computational load. We apply this framework to the complex protein design problem in order to identify a low-energy amino acid sequence for a target structure.

SUMMARY

In protein design, the energy associated with a huge number of sequence-conformer perturbations has to be routinely estimated. Hence, enhancing the throughput and accuracy of these energy calculations can profoundly improve design success rates and enable tackling more complex design problems. In this work, we explore the possibility of tensorizing the energy calculations and apply them in a protein design framework. We use this framework to design enhanced proteins with anti-cancer and radio-tracing functions. Particularly, we designed multispecific binders against ligands of the epidermal growth factor receptor (EGFR), where the tested design could inhibit EGFR activity *in vitro* and *in vivo*. We also used this method to design high-affinity Cu^{2+} binders that were stable in serum and could be readily loaded with copper-64 radionuclide. The resulting molecules show superior functional properties for their respective applications and demonstrate the generalizable potential of the described protein design approach.

INTRODUCTION

Protein design processes search for sequences to fill up a given target structure while minimizing the free energy of this defined configuration. Under the layout of fixed-backbone design, amino acid side chains and conformations are sampled at the designable positions and scored for their energy within their local environment. Thus, protein design simulations typically sample a large number of sequence-conformer combinations even for a small number of designable positions. Moreover, the computational load increases steeply with the difficulty of the specific design problem.[1] This demands scoring functions to be sufficiently fast to cover large sequence sub-spaces that contain viable solutions.[2] The inherent trade-off between the scoring

speed and the accuracy has led to the broad utility of fast energy functions and trained or knowledge-based models. Previous efforts include the use of knowledge-based scoring terms,[3] coarse-grained representation,[4] geometric filters,[5] and directly[6] or indirectly[7] learned sequence-structure relationships.

In this work, we explore the feasibility of tensorizing energy calculations for molecular mechanics applications and, particularly, evaluate its usefulness for protein design simulations. In protein design, the evaluation of the energy associated with non-bonded interactions represents the computational bottleneck. We seek to demonstrate the accessible performance gains from reformulating the non-bonded energy terms (i.e., the Lennard-Jones [LJ], electrostatic, and solvation energy terms) to best suit the computing hardware architecture. Specifically, conducting energy calculations as large-matrix (or tensor) operations enables substantial efficiency gains on both conventional central processing units (CPUs) and massively parallel processing hardware (Figure 1A). Here we use an energy function that is readily derived from established, self-consistent force fields. This obviates the need for empirically optimizing a scoring function against one or more training datasets and thus avoids overfitting and training bias. Our approach also attributes inaccuracies directly to the force-field parameters, allowing improvements to be more systematic. Finally, as tensorization increases the throughput of evaluating sequence:conformer combinations, it raises the probability of finding lower-energy solutions, which can improve the experimental success rate.

As an overview, our design workflow starts by pre-computing a discrete rotamer library from molecular dynamics (MD) simulations of isolated amino acids. These simulations are run under the same force field from which the scoring function terms are derived. The resulting conformational pool of each amino acid is then partitioned into clusters with a constant number of rotamers. Each rotamer in the database is dually represented by its atomic coordinates and by its interaction fields projections (field tensors). To evaluate the interaction energy between an inbound rotamer and a host structure, the existing side chain at the designable position is removed (Figure 1B), and the surrounding environment is projected as a three-dimensional histogram of the constituting atoms (environment tensor). The element-wise multiplication of a rotamer's field tensor and a structure's environment tensor yields the interaction energy between them (Figure 1C).

We use this framework for designing two different classes of proteins. The first class aims at inhibiting soluble growth factors, particularly, the epidermal growth factor (EGF)-like ligands. Multispecific quenchers of the several soluble ligands of the EGF receptor (EGFR) are needed for broad inhibition of EGFR signaling and can serve as a new therapeutic modality for several types of cancer.[8] In this work, we redesigned a minimal receptor domain to maximally stabilize it in its ligand-bound form. This yielded small, single-domain binders (18 kDa) that are biophysically and functionally superior to the natural template. The designs showed multispecific binding to their target growth factors and potently inhibited EGFR signaling *in vitro* and *in vivo*. Our second objective is to develop a new class of protein-based radiotracers for use as genetically encodable molecules for positron-emission tomography (PET) imaging. Using our framework, we could redesign a natural copper-storage protein into a developable form as a diagnostic protein tag. Two designs were monomeric and showed superior stability and high production yield, in contrast to the starting template. These designs were capable of binding Cu^{2+} with high affinity and low off rates, and, importantly, were sufficiently stable for hours in serum *in vitro*, highlighting their translational potential.

RESULTS AND DISCUSSION

Rotamer library construction

Typically, discrete rotamer libraries are constructed from amino acid conformations pooled from known structures. As the structural databases grew in size, more stringent inclusion criteria were imposed, greatly improving the quality of the available libraries.[9] Nonetheless, PDB-based rotamer libraries still provide a sparse coverage of the rotameric space and entrain undesirable factors pertinent to protein structure determination, such as cryogenic measurement condition or ensemble averaging. Thus, we use MD to achieve more extensive sampling of the conformational tendencies of amino acids in folded proteins.[10] Furthermore, MD simulations of isolated amino acids can cover a broader conformational distribution that is unbiased by the choice of input protein structures. Such conformational distributions more faithfully reproduce the tendencies of the random coil state and represent a reference energy distribution prior to any folding event.[11] We have therefore chosen to build the rotamer library using MD simulations of capped amino acids (i.e., Ac-X-NHMe).[12] The internal energies related to backbone conformational preferences were derived from the occupancy of backbone bins of the dihedral space (i.e., (φ,ψ) angles). The side-chain conformational preferences were nonetheless mapped in the Cartesian space by means of root-mean-square deviation (RMSD) clustering after alignment to a (C^{β}-C^{α}-N) frame of reference. This was sought to obtain a constant number of rotamer clusters for every (φ,ψ) bin, regardless of the number of atoms or conformational tendencies of the amino acid. Each cluster would be represented by a single rotamer in the final library, where the relative energy of the rotamer relates to its respective cluster size (STAR Methods).

This approach brings several advantages. First, the internal energies derived from the conformational preferences of amino acids are consistent with the other non-bonded energy terms used in design calculations, as both rely on the same force field. Second, the generated rotamers can implicitly encode the dynamic influences on bond angles, bond stretching, and improper dihedrals. These subtle deformations are generally dismissed in other rotamer libraries, but have been shown to significantly impact the energy gap between native and non-native states of a protein.[13] Third, this approach offers great versatility regarding the energy function of choice. Whereas here we used the CHARMM force field[14] for both the rotamer library creation and the design energy function, more complex potentials (e.g., polarizable force fields) can be deployed. The library can be readily extended to cover rotameric distributions in different pH or sequence (e.g., tri- or pentapeptide) contexts and can generate on-demand rotamers for non-standard amino acids or ligands in a consistent way.[15] Fourth, the long MD trajectories

CellPress
OPEN ACCESS

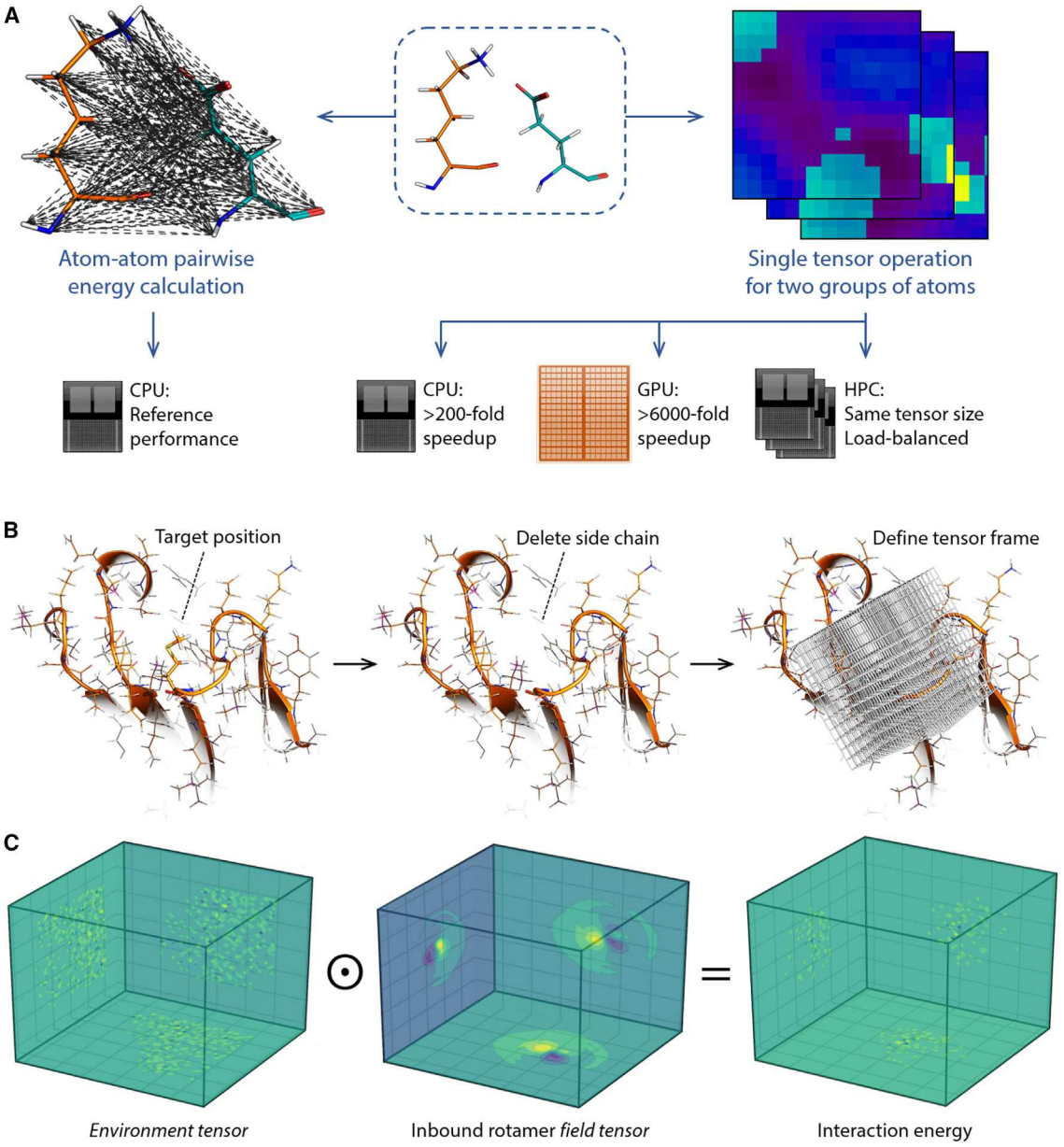


Figure 1. The concept of the tensorized design framework

(A) The non-bonded interactions between an amino acid and its molecular environment (e.g., another proximal amino acid) entail the calculation of all atom-atom pairwise potential energies within a distance cutoff. The intensive execution of a large number of distance- and energy-evaluation instructions, as well as memory handling processes, slows the overall performance. Thus, formulating the energy evaluation problem between two groups of atoms as a single tensor operation not only speeds up the scoring on conventional processors, but also renders the calculation highly compatible with stream processors. Moreover, the constant dimensions of the used tensors enable ideal load balancing on high-performance computers. The performance enhancement figures were calculated for the Lennard-Jones potential function, a cubic tensor representation of 22 Å side and 0.5 × 0.5 × 0.5 Å voxels. The computing operations are performed as in (B) and (C).

(B) Once a mutable or repackable target residue is defined, the input structure is transformed to the frame of reference with respect to the residue's backbone coordinates. The side-chain atoms of the target residue are then deleted, leaving behind the "environment" atoms.

(C) Proxy values for the atoms' positions, partial charges (illustrated here as plane projections; yellow, negative; blue, positive) or their surface solvation energies are projected onto the voxels of a constructed tensor. The rotamer library comprises the more expensive, pre-computed smooth interaction fields (i.e., field tensors; example shown for an asparagine side chain), which through a single element-wise multiplication with the environment tensor yields the spatially resolved energy in a "two-body" format.

effectively sample a broader space compared with what is traditionally obtained from the PDB-derived rotamer libraries[16] (Figure S1A). Furthermore, by raising the temperature of the isolated amino acid MD simulations, broader coverage of otherwise poorly sampled (φ, ψ) regions can be covered more effectively than in PDB-based rotamer libraries (Figure S1B). The former can be particularly successful in better accessing rare "linchpin" rotamers reported to constitute sampling bottlenecks.[17] Last, the extraction of a constant number of rotamers plays an important role downstream of the algorithm, as it dictates a defined rotameric sampling granularity. It also guarantees a uniform load balancing during the design calculations, particularly in parallelized implementations.

Tensorized molecular interaction fields, energies, and mechanics

Our framework combines two principles to maximize the computational efficiency of energy calculations. The first principle is to pre-compute and store most of the information needed for energy calculation. The second principle is to deploy a tensorized form of the energy functions to better fit a single-instruction, multiple data processing paradigm, a hallmark of modern computing technology (Figure 1A).

In this framework, the scoring problem is simplified into a multiatom, two-body problem, where the 1st body represents the multiatom chemical environment surrounding the side chain at the designable position, wherein the atoms of this side chain are absent. The 2nd body represents the inbound side-chain rotamer, aligned to the same frame of reference. The information on this two-body interaction is encoded in an asymmetric fashion in which the 1st body only encodes the three-dimensional occupancy of its atomic positions and charges, while the 2nd body encodes the net, real-valued energy field around all of its respective atoms (Figure 1C). The computationally expensive step is the projection of energy fields, which in this case is restricted to the 2nd body (i.e., the rotamer) and is hence pre-computed once and stored in a look-up table. This restricts the run-time computing load to simply mapping the 1st-body three-dimensional occupancy, substantially reducing the run time needed at every designable position. Such a representation benefits from further speed-up when implemented in a tensorized fashion. In this manner, the scalar-valued interaction energy between the two bodies is obtained by the sum of the element-wise product of the two tensors (representing environment atoms and the rotamer energy field). This format is ideally suited for evaluating both the LJ and the electrostatic potentials, albeit at the cost of assuming symmetric LJ parameters for the interacting atom pairs, according to the atom's respective encoding in the 2nd body tensor. Nonetheless, an atom-type rescaling of atoms in the 1st body can be used to correct for this in the future. Applying the tensorization framework would differ, however, in the case of encoding a surface area-based solvation potential. In the latter situation, the 1st body tensor has to fully describe the environment's surface solvation energy field. Hence, the solvation energy per unit surface area will be normalized by the number of voxels representing the atomic surface (STAR Methods). Such a tensor is pre-computed for the 2nd body (i.e., the inbound side-chain rotamer) and is computed on the fly for the 1st body as well as for the combined two bodies (STAR Methods). This renders the solvation term the relatively more expensive energy term to compute.

To assess the accuracy of this energy function retrospectively, we predict the energy change associated with single-point mutations. To avoid the confounding effects of combinatorial repacking and iterative energy minimization, the energy values were evaluated without any combinatorial side-chain optimization, but by finding the lowest-energy rotamer at the designated position only (STAR Methods). We used the dataset of mutants of the β1 domain of streptococcal protein G, constituting the largest thermodynamic stability dataset collected in a single experimental setup to date.[18] This dataset covers most of the single-point mutagenesis landscape of the Gβ1 protein and represents a broad range of burial and secondary structure contexts. In this setup, $\Delta\Delta G$ values obtained by the tensorized potential (herein referred to as the Damietta potential) showed a better correlation ($R = 0.46$, $p = 7.3 \times 10^{-34}$) compared with the reported Rosetta score[18] ($R = 0.36$, $p = 2.6 \times 10^{-21}$) (Figure S2), where both methods were used without minimization. By testing our potential against other large datasets generated from proteolytic stability assays,[19] we obtained $\Delta\Delta G$ correlation coefficients ranging between 0.26 and 0.41. These datasets comprise diverse folds of the N-terminal domain of the phage 434 repressor (1,046 mutants; $R = 0.41$, $p = 1.1 \times 10^{-43}$), the SH3 domain of human obscurin (1,097 mutants, $R = 0.26$, $p = 2.1 \times 10^{-18}$), the N-terminal domain of ribosomal protein 493 L9 (725 mutants, $R = 0.35$, $p = 4.9 \times 10^{-22}$), and r11_829_TrROS protein designed by trRosetta (833 mutants, $R = 0.31$, $p = 1.7 \times 10^{-20}$) (Figure S2). We also sought to evaluate the native side-chain conformer recovery for a dataset of proteins with available X-ray and NMR structures.[6] The overall recovery rates obtained by the Damietta potential were around 70% for χ_1 and 50% for $\chi_{1\&2}$ (Figure S3A). As expected, buried residues had a higher prediction accuracy (~90% for χ_1 and 65%–85% for $\chi_{1\&2}$; Figure S3B), given the constraining chemical environment around the amino acids at the protein core. We further used the same dataset of X-ray structures to evaluate native sequence recovery, where the results showed very low recovery rates compared with other design methods (Figure S3C). This can be attributed to the fact that our potential was not trained to maximize sequence nativeness, which is not necessarily a proxy of sequence optimality.

Combinatorial design by decision tree swarm

The highly dimensional nature of combinatorial design of more than a few amino acid positions severely limits the usefulness of exact sampling algorithms in finding global minimum solutions within reasonable computing time frames. Nonetheless, despite the non-additive effects of correlated mutations, favorable mutations generally tend to cluster closely in the sequence space.[20] Thus, a swarm of greedy samplers traversing several sequence optimization paths simultaneously can generally reduce entrapment within local minima and lead to near-optimal solutions. This is clearly demonstrated by the success of stochastic design algorithms.[21] In order to enable the exploration of a sizable number of mutations, we developed a combinatorial sampling strategy that searches for successive minima

CellPress
OPEN ACCESS

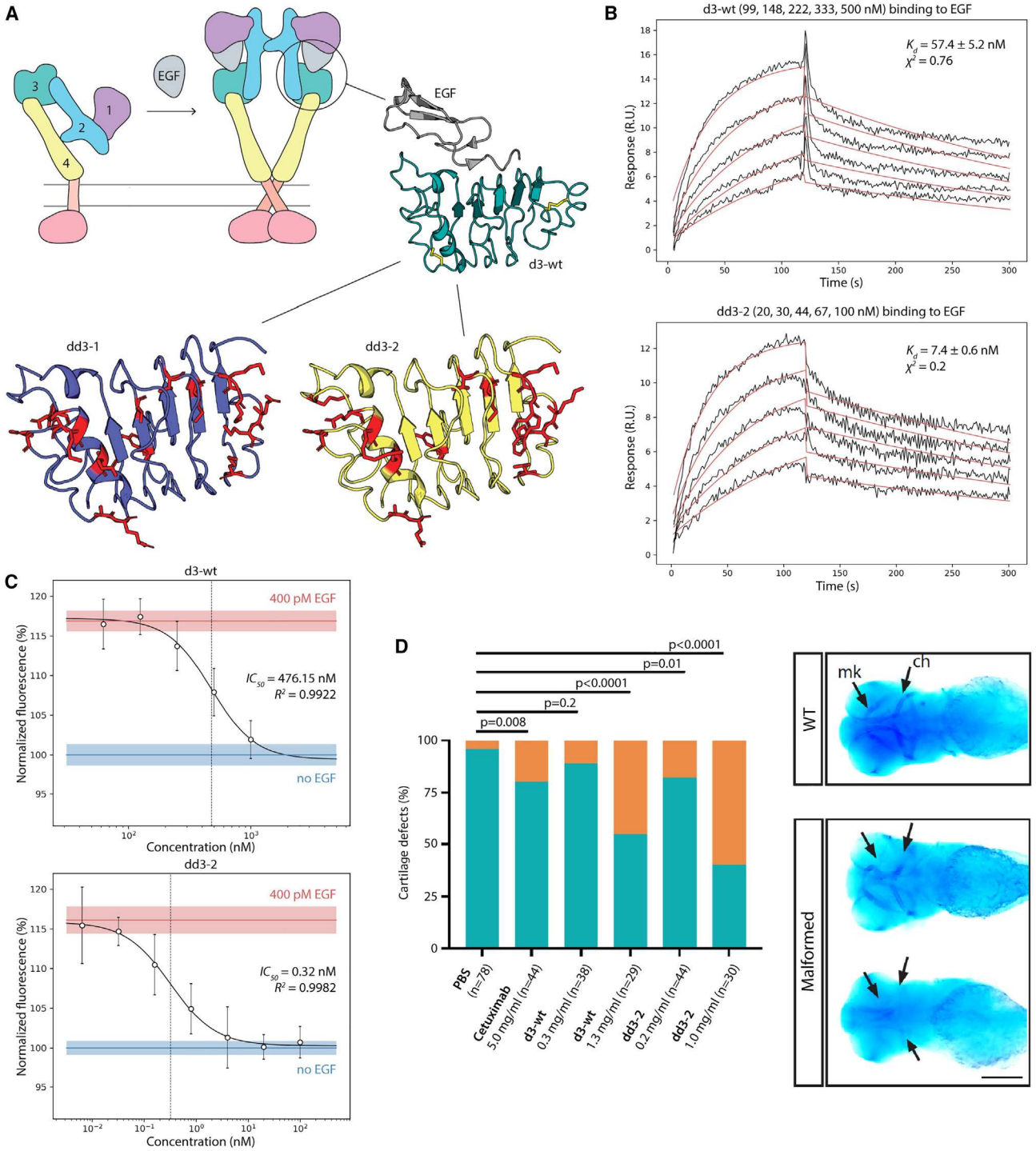


Figure 2. The design and characterization of EGFR inhibitors

(A) The EGFR extracellular segment consists of four domains (d1, violet; d2, cyan; d3, teal; d4, yellow). In the absence of a ligand, it lies in a closed monomeric configuration. Upon ligand binding (here EGF, gray) the receptor adopts an open, dimeric configuration triggering intracellular signaling. As the third domain of the EGFR is reported to hold most of the binding affinity to the EGF ligand, it was used as a template to design soluble EGF binders. A close-up view of the EGF in complex with the wild-type d3 domain (d3-WT) is shown in gray and teal, respectively (PDB: 1IVO). Disulfide bridges in the d3-WT structure are shown as yellow sticks. Using the described energy function, the highest energy residues were identified. These residues were defined as mutable (shown in red) and designed using the combinatorial sampler. Two design models (dd3-1, purple; dd3-2, yellow) were finally chosen for experimental characterization.

(legend continued on next page)

by spanning a few semi-independent search paths. This strategy builds on the power of parallel, loosely communicating conformational samplers that assume a globally smooth, but locally rugged, landscape as was demonstrated with the SARS[22] and FLAPS[23] algorithms. One way of implementing this approach in a design context is through a ''few-to-many, many-to-few'' scheme, whereby designable amino acid positions are arranged as depth levels within a decision tree, and nodes at each level represent the mutational decisions. As branching represents the expansion to many mutant combinations, a ranking-and-trimming step that keeps only a few lowest-energy designs is imposed between the layers of the tree (Figure S4). This alternation between the many combinations and the few best intermediary decoys guarantees the traversal of a pre-set number of branches (n_paths) down the tree depth at any given level. This scheme restrains the combinational load complexity and enables several parallelization schemes. Measuring the overall performance of the algorithm against varying design loads shows the algorithm to greatly simplify the sampling complexity, while enabling an arbitrary level of parallel sampling through the number of paths (n_paths). These paths not only generate diverse output from a single run, but also keep track of several local minima, which improves the overall minimization outcome when the sampling complexity grows. Particularly, under the same sampling complexity, minimizing across a larger number of paths leads to lower energy decoys (Table S1). These results also highlight a rotamer sampling performance in the microsecond range, which, compared with other methods, indicates substantial performance gains (Table S1). We expect further performance optimization to greatly improve these figures in the future.

In most stochastic design algorithms, higher energy mutations can still be accepted at a lower probability in favor of basin hopping and diversity generation. This, however, adds random noise to the already heterogeneous uncertainty of the scoring function (i.e., scoring error). Instead, here we introduce diversity by randomizing the ordering of the designable positions along the decision tree (i.e., across independent design simulation replicas). In this setup, the parallel, deterministic sampling trials are more dispersed across the solution space and their optimization paths can be easily retraced, in comparison with the use of a Metropolis criterion. This ''few-to-many-to-few'' combinatorial sampler is thus best run through several independent replicas (while randomizing the order of designable positions) with iterated traversals of the same decision tree (n_{iter}) in order to improve the search convergence within each simulation replica (Figure S4).

Design and characterization of EGFR inhibitors

As a proof-of-principle, we applied our framework to create inhibitors of EGFR signaling, a key pathway involved in the survival, proliferation, and dissemination of tumor cells.[24] EGFR (HER1) is a receptor tyrosine kinase that represents an important target for modulating signal transduction cascades, as it dimerizes upon ligand-induced conformational change.[25] Approved inhibitors of EGFR signaling are either small-molecule inhibitors of the receptor's intracellular kinase domain or monoclonal antibodies blocking its ectodomain dimerization.[26] These are indicated for treating different EGFR-dependent cancers, e.g., colon cancer and epidermoid carcinomas.[27] Nonetheless, these two inhibition modalities (i.e., tyrosine kinase inhibitors and dimerization-inhibiting monoclonal antibodies) have been shown to be subject to evasion by cancer cells through numerous evolution and resistance mechanisms.[26] Binders targeting the ligand itself, i.e., EGF,[28,29] can provide a new class of inhibitors with potential synergistic effects when combined with existing drugs. However, the cross-activity of more than one EGF-family ligand against the EGFR (and its related receptors, particularly the HER4 receptor) complicates this endeavor. For instance, the heparin-binding EGF-like growth factor (HB-EGF), transforming growth factor α (TGF-α), and amphiregulin (AR) also play roles in activating these receptors and promote cancer progression.[30]

The multiligand nature of EGFR signaling is thus better tackled through the development of polyspecific binders capable of quenching more than one growth factor, ideally, with high affinity. Previous attempts to engineer a recombinant form of the entire extracellular segment of the EGF and HER3 receptors could achieve broad ligand-binding specificity.[30,31] These binders were constructed as dimeric IgG1 Fc-fragment fusions with the four extracellular domains of the receptor. While achieving broad inhibition of EGFR and HER3 ligands, these constructs require recombinant expression in mammalian cells and possess a large molecular weight of approximately 190 kDa, which can hamper their bioavailability at the relevant tumor tissue.[32]

In this study, we aimed to create a miniature binder, using only one of the EGFR ligand-binding domains as a starting template and stabilizing it in its ligand-bound conformation by sequence redesign. Previous work has shown the human EGFR domain 3 (herein referred to as d3-WT) to be the ectodomain encoding most of the binding information to EGF.[33] The d3-WT template sequence (Table S2) was restricted to a stretch of 168 amino acids, which contains disulfide bridges at the beginning and the end of the domain. The designs were instead based on a truncated domain boundary to include only 160 amino acids. All cysteine residues were excluded to improve downstream

(B) SPR sensograms show the dd3-2 design to bind EGF tighter than d3-WT. A similar pattern with improved affinity of the design was observed toward other EGFR ligands (Figures S6 and S7). K_d is represented as the mean ± standard deviation (SD). The χ^2 value represents the difference between the experimental data and the fitted curve averaged over the whole sensogram. Experimental data, black; fit, red.

(C) Proliferation inhibition assays were done using the EGFR signaling-dependent A431 cells. The inhibition of cell proliferation was observed to be much stronger for dd3-2 (IC_{50} = 0.32 nM) than for d3-WT (IC_{50} = 476 nM). The positive and negative control values of cell proliferation with (400 pM) and without EGF treatment are indicated by red and blue lines, respectively. Shades and error bars represent SD across three replicates.

(D) Pharyngeal skeleton of zebrafish embryos was stained with Alcian blue. First-arch Meckel's cartilage (mk) and second-arch derivative ceratobranchials (ch) are observable (arrows). Upon EGFR inhibition, embryos with partial absence of Meckel's cartilage and ceratobranchials or without any cartilage formation are observed and categorized in the malformed class. Cartilage defection upon injection of PBS, cetuximab, dd3-2, and d3-WT is shown in percentages for each group. Two-tailed p values were analyzed by a 2 × 2 contingency table in GraphPad. n, the number of evaluated embryos. Scale bar, 250 μm.

Table 1. SPR-derived binding parameters of d3-WT, dd3-1, and dd3-2 to different EGFR ligands

	k_a (1/Ms)	k_d (1/s)	K_D (nM)
d3-WT			
EGF	$49.7 \times 10^3 \pm 0.7 \times 10^3$	$2.8 \times 10^{-3} \pm 0.2 \times 10^{-3}$	56 ± 2.1
HB-EGF	$47.2 \times 10^3 \pm 4.4 \times 10^3$	$10.7 \times 10^{-3} \pm 3.7 \times 10^{-3}$	370.2 ± 256.3
TGF-α	$38.4 \times 10^3 \pm 26.9 \times 10^3$	$57.7 \times 10^{-3} \pm 20.9 \times 10^{-3}$	$2.55 \times 10^3 \pm 2.52 \times 10^3$
dd3-2			
EGF	$248.7 \times 10^3 \pm 53.7 \times 10^3$	$2.2 \times 10^{-3} \pm 0.1 \times 10^{-3}$	9.2 ± 2.6
HB-EGF	$162.7 \times 10^3 \pm 68.5 \times 10^3$	$4.3 \times 10^{-3} \pm 0.8 \times 10^{-3}$	34.5 ± 8.1
TGF-α	$181.6 \times 10^3 \pm 22.6 \times 10^3$	$37.8 \times 10^{-3} \pm 12.2 \times 10^{-3}$	231.9 ± 54.6
dd3-1			
EGF	$260.2 \times 10^3 \pm 43.6 \times 10^3$	$2.7 \times 10^{-3} \pm 0.08 \times 10^{-3}$	10.4 ± 1.5

Sensograms are shown in Figures 2B, S5D, S6, and S7.

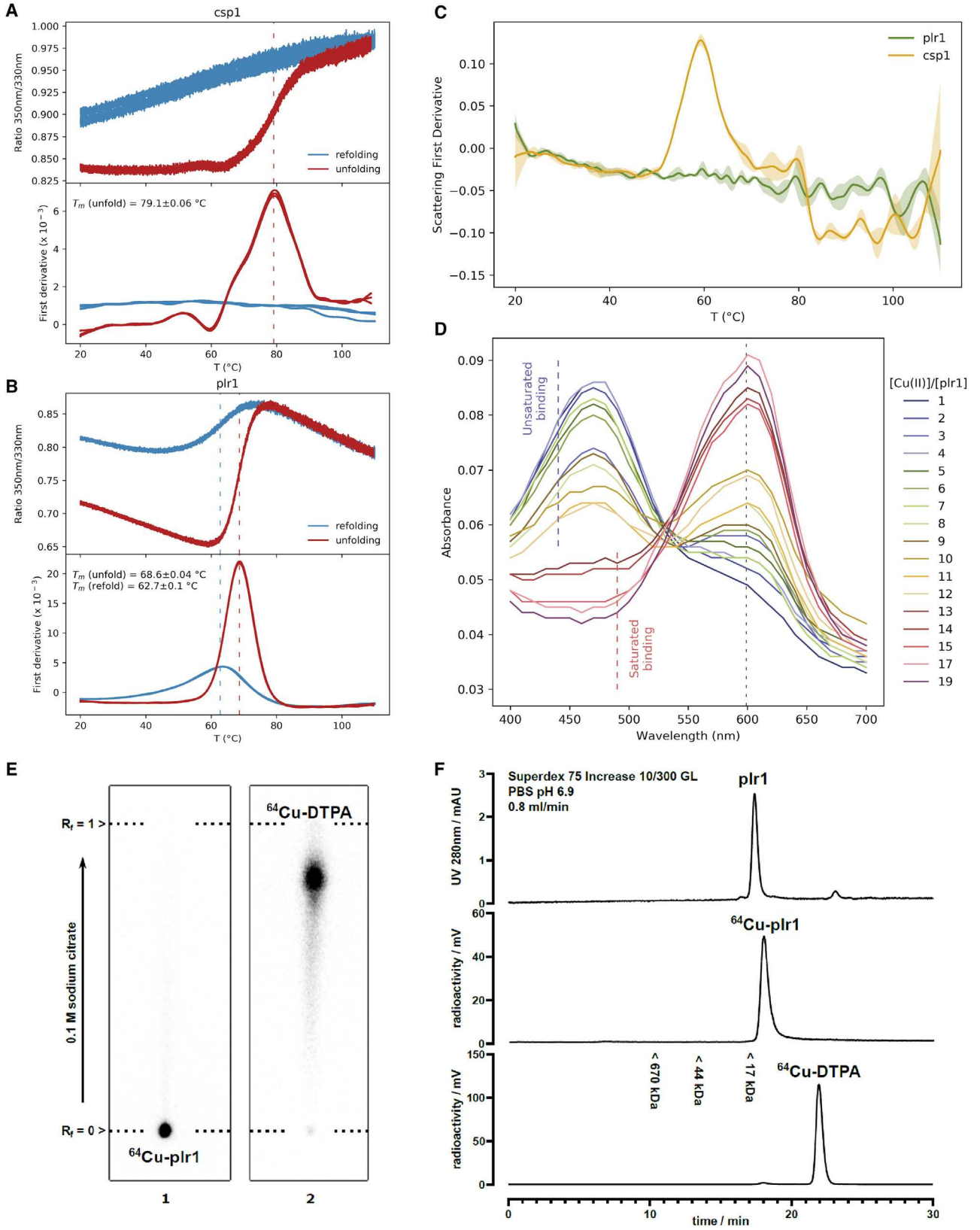
properties of the designs. The designable positions were set to comprise all residues with energy higher than a set threshold, which were identified using the "repack all" (ra) protocol. Running 100 instances of the tree swarm combinatorial sampler (cs_f2m2f) with randomized order of the designable positions yielded about 200 decoys with unique sequences. These were subject to accelerated MD (aMD) simulations and were ranked according to their conformational stability. The conformational stability scores as well as the RMSF plots derived from the latter simulations indicated a better stability of the designs compared with the d3-WT model, where the cysteines were reduced (Figure S5A). The two most mutually distant sequences in the top 10 designs were eventually selected for experimental evaluation, named dd3-1 and dd3-2 (designed domain 3; Figure 2A).

The starting template (d3-WT) and two designs (dd3-1 and dd3-2) were expressed in *E. coli.* Double purification from the soluble fraction showed the three proteins to have similar yields of approximately 0.2 mg per liter of culture. The designs also showed thermostability similar to that of d3-WT, as evaluated by nanoscale differential scanning fluorimetry (nanoDSF) (Figure S5B). However, the designs exhibited much stronger EGF inhibition activity in proliferation assays using the EGF-dependent epidermoid carcinoma cell line A431. Particularly, dd3-2 was the most active design, with a proliferation inhibition IC_{50} more than 1,000-fold lower compared with d3-WT ($IC_{50,dd3\text{-}2}$ = 0.32 nM vs. $IC_{50,d3\text{-}WT}$ = 476 nM) (Figure 2C) and only 3-fold higher than that of cetuximab, a therapeutic anti-EGFR antibody[34] (Figure S5C). Next, to evaluate the difference in binding affinities toward EGF, we carried out surface plasmon resonance (SPR) titrations of our binders against immobilized EGF. The results showed that dd3-1 and dd3-2 bind EGF around 6-fold tighter than d3-WT, where dissociation constants (K_d) were 10, 9, and 56 nM for dd3-1, dd3-2, and d3-WT, respectively (Table 1; Figures 2B, S5D, S6, and S7). This enhanced binding can be the result of stabilizing the ligand-bound conformation. Compared with previous results that weaponized an Fc chimera of the entire EGFR extracellular segment (~95 kDa per subunit),[31] our dd3-2 design is far smaller (18 kDa), leading to a better protein efficiency (i.e., $\Delta G_{bind}/MW$[35]) of the latter (−2.6 kJ/kDa) vs. the former (−0.5 kJ/kDa). To further evaluate the ability of our designs, particularly dd3-2, to bind other related EGFR ligands, we performed SPR binding experiments against HB-EGF and TGF-α, which are also important therapeutic targets for treatment of EGFR-dependent cancers.[30] The results showed dd3-2 to bind HB-EGF 10-fold tighter than d3-WT, where K_d values of 35 and 370 nM were observed for dd3-2 and d3-WT, respectively, and demonstrated dd3-2 to bind TGF-α 11-fold tighter than d3-WT, with K_d values of 232 and 2,550 nM, respectively (Table 1; Figures S6 and S7). This polyspecificity of d3 proteins could be an explanation of their high inhibitory activity against the A431 cell line, given the complex signaling interplay among autocrine and paracrine EGFR ligands.[36] In addition, the stronger inhibition by the designs can be attributed to their improved stability in the monomeric form in solution, as observed during the proteins' purification and analytical size exclusion (Figure S8).

To investigate potential effects of the designed inhibitors *in vivo*, we injected equal volumes of PBS solution containing cetuximab (positive control), d3-WT, or dd3-2 into zebrafish embryos. As a negative control, pure PBS was injected. Inhibitors were administered starting at 4–6 h post-fertilization during 4 days. As a first step, the survival of the embryos was determined every day from 1 to 4 days post-fertilization (dpf). While almost no effect on survival was observed at any time point following injection of PBS, injections of cetuximab (5 mg/mL), d3-WT (0.3 mg/mL, 1.3 mg/mL), or dd3-2 (0.2 mg/mL, 1.0 mg/mL) were found to be lethal to different extents (Table S3). Next, we evaluated the morphological defects present in the surviving embryos at 4 dpf. Since it has been previously shown that EGFR inhibitors cause developmental defects in head cartilage,[37] we analyzed head cartilage formation by Alcian blue staining (Figure 2D). In comparison to wild-type embryos with completely formed cartilaginous elements of the pharyngeal skeleton, the embryos with cartilaginous defects or even without any cartilage formation were classified as the malformed group. The results indicate cetuximab, d3-WT, and dd3-2 to affect skeletal development in a manner typical of EGFR signaling impairment. In line with the above-described biophysical and cell-based experiments, dd3-2 caused stronger effects in zebrafish embryos compared with d3-WT. Interestingly, both d3-WT and dd3-2 were more active than the anti-EGFR antibody cetuximab (Figure 2D).

Design and characterization of copper binders

To further test the performance of the Damietta potential, we sought to design metallic radionuclide-binding proteins. Metal-binding proteins serve essential functions, including catalysis,



(legend on next page)

sensing, transport, and storage.[38] Designed metalloproteins can be tailored to encode one or more of such functions and be useful for a range of biomedical applications.[39,40] Particularly, metalloproteins capable of high-affinity metal binding, efficient storage, and transport can serve as electron microscopy contrast agents,[41] probes for magnetic resonance imaging,[42] or targeted radioactive tracers for radiotherapy and diagnostic imaging purposes.[43] We specifically aimed to design proteins to bind the radioactive $^{64}Cu^{2+}$ ions. Such genetically encodable radiotracers can be fused with targeting proteins for high-resolution PET imaging or radioligand therapy with the therapeutic radioisotope ^{67}Cu, forming an ideal therapeutic/diagnostic pair.[44]

Given this intended function, we based our designs on helical bundles to create modules with robust and stable folding and thus minimal interference with any fused homing protein (e.g., tumor cell-targeting antibody fusions).[45] We chose a cysteine-rich helical bundle protein (Csp1) as a starting template, which was shown to bind 13 Cu^{+} ions along its core.[46] Although Csp1 has a low molecular weight (13 kDa) and possesses a simple up-down four-helix structure, it suffers several drawbacks. Specifically, Csp1 is unstable, is tetrameric, has low bacterial production yield, and has complex purification requirements.[46] We therefore redesigned 22 amino acid positions, which were mostly surface exposed, to disrupt the oligomerization interface of the tetrameric Csp1 template and improve solubility and stability of the helical bundle (STAR Methods). The two most conformationally stable designs as assessed by MD simulations (named plr1 and plr2) were selected for experimental characterization (Table S2). The synthetic genes encoding plr1, plr2, and Csp1 were cloned without purification tags in a vector for expression in *E. coli*. The soluble expression levels were highest for plr1, followed by plr2, both being higher than Csp1. In contrast to the template, which has a net negative charge of −2, the designs possessed high net-positive charges (plr1, +14; plr2, +17). This particularly facilitated the purification of the supercharged designs using ion-exchange chromatography. We restricted our further characterization to the plr1 design given its high purification yield of >50 mg per liter of culture (>20-fold higher than Csp1). Analytical size exclusion showed plr1 to be monomeric, in contrast to the template Csp1, which was tetrameric and showed significant aggregation (Figure S8). Thermostability analysis indicated Csp1 and plr1 to have melting temperatures (T_m) of 79°C and 69°C, respectively (Figures 3A and 3B). However, Csp1 displayed a lower aggregation temperature (T_{agg}), which was 60°C and >110°C for Csp1 and plr1, respectively (Figure 3C). Similarly, irreversible thermal denaturation was observed for Csp1, in contrast to the reversible folding of plr1 (Figures 3A and 3B). The colloidal stability of the plr1 design is important for its clinical usefulness, given that aggregation tendency can greatly reduce the efficacy and raise the immunogenicity risk of biopharmaceuticals.[47] The difficulty of handling Csp1 protein restricted our further functional analysis to the designed forms only.

By titrating copper and using a chromophoric probe, we observed a mid-point indicating approximately 13 Cu^{2+} binding sites on plr1 (Figure 3D). This high metal-binding capacity of almost 1 metal/kDa of protein is consistent with that originally reported by Csp1.[46] This binding ratio could be beneficial for high imaging sensitivity and efficacy in radiotracer imaging and radioimmunotherapy applications, respectively. To test suitable labeling conditions with $^{64}Cu^{2+}$, we incubated plr1 with the buffered radioisotope at a specific radioactivity of 2 GBq/mg. After incubation for 60 min at 35°C, the radioactivity was efficiently (>90%) sequestered into plr1, as judged by radio-thin-layer chromatography (radio-TLC) (Figure 3E), and eluted from size-exclusion chromatography with the same profile as unlabeled plr1 (Figure 3F). These results strongly support the capacity of the design to readily and stably chelate copper radioisotopes through a simple incubation procedure.

Exploring the determinants of structural stability of the copper-binding proteins can guide the generation of enhanced variants for clinical applications. Through a second design round, we aimed to create two classes of variants to evaluate their metal-binding stability. The first class involved repacking core residues to eliminate three or six core cysteine residues, cr3, and cr61 or cr62 designs, respectively (Figure 4A; Table S2). These variants have their cysteine-lined lumen plugged at the solvent-accessible end, which could restrict the outward diffusion of coordinated metal ions. The second class was negatively supercharged designs (neg1 and neg2), where the positively charged residues of plr1 were forcibly redesigned into neutral or negatively charged residues (Figure 4A; Table S2). These variants would provide a favorable electrostatic environment for the coordinated metal ions, especially given the +2 oxidation state of the target copper ions. While the five new designs were all well expressed, the three repacked core variants (cr3, cr61, and cr62) were majorly dimeric in solution and therefore were excluded from further analysis. Conversely, the neg1 variant could be purified in a monomeric state. In radio-TLC experiments, neg1 bound $^{64}Cu^{2+}$, which was also observed for plr1 (Figure S9A). Competitive binding assays showed neg1 to bind Cu^{2+} 3-fold tighter compared with the plr1 design (Figures 4B

Figure 3. Design of stabilized $^{64}Cu^{2+}$ binding proteins

(A) NanoDSF melting curves show Csp1 to unfold at 79°C (red curves), without a refolding transition upon cooling (blue curves).

(B) The plr1 design has a lower melting temperature of 69°C (red curves) but refolds upon cooling (blue curves). Melting temperatures (T_m) are represented as the mean ± SD.

(C) Csp1 scattering signal, however, shows an aggregation mid-point at 60°C, highlighting its colloidal instability, while plr1 does not show a change in scattering and does not precipitate in solution. Shading represents SD across three replicates.

(D) Titrations performed using the chromophoric change of zincon indicate plr1 to bind between 12 and 13 Cu^{2+} ions per molecule.

(E) Radiographic images of silica TLC plates with $^{64}Cu^{2+}$-loaded plr1 (1) and the same sample stripped with DTPA for 3 h (2) show plr1 to bind $^{64}Cu^{2+}$. TLC plates were developed with 0.1 M sodium citrate (pH 5). Proteins stay at the starting spot and DTPA migrates near the front line under these conditions. Similar results were observed for neg1 (Figure S9A).

(F) Size-exclusion chromatogram of plr1 at 280 nm (top), radioactive signal of runs with $^{64}Cu^{2+}$-loaded plr1 (middle), and a sample stripped with DTPA (bottom). plr1 binds $^{64}Cu^{2+}$ and elutes corresponding to the same size as non-loaded plr1.

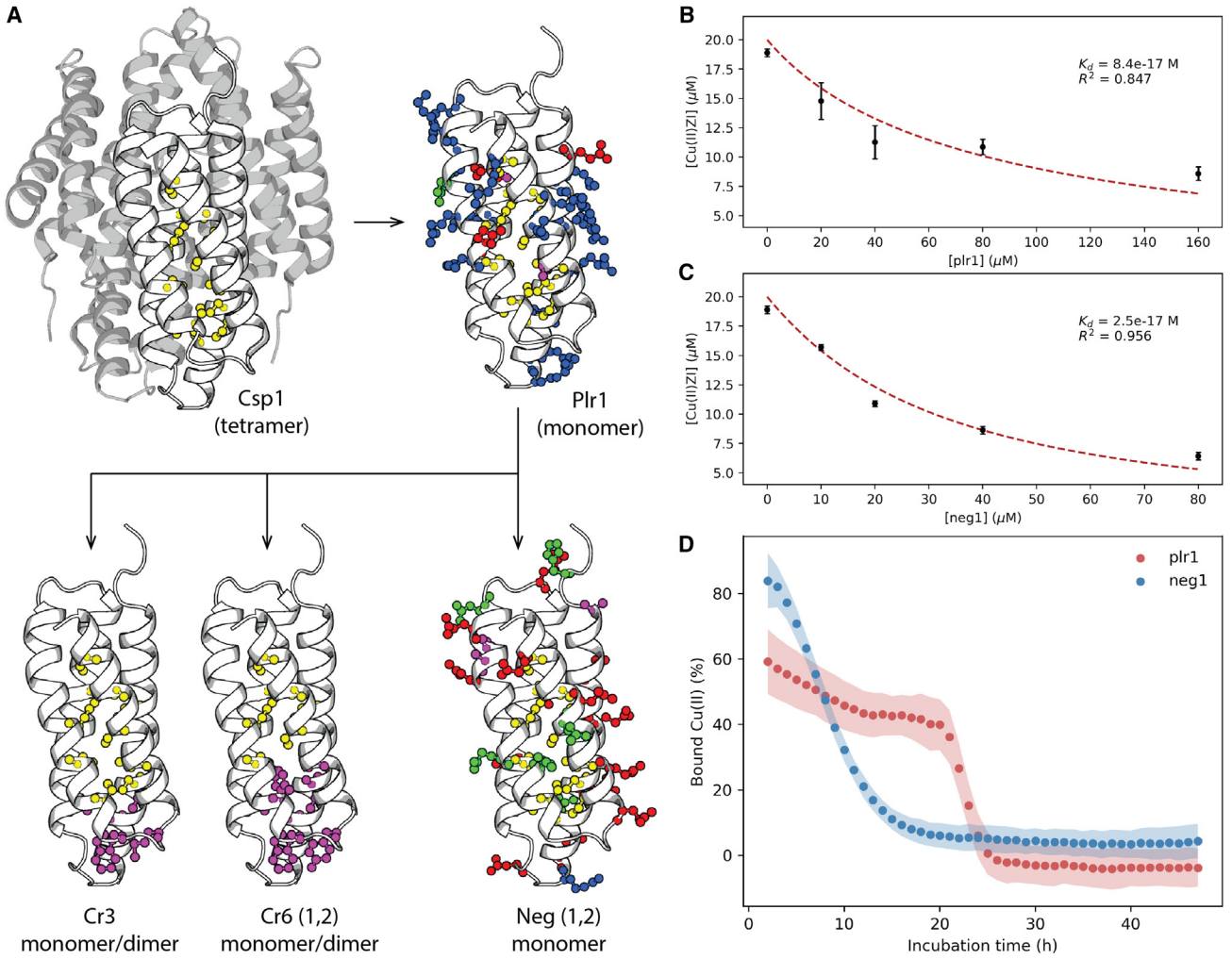


Figure 4. Exploring the sequence-stability relationship of copper-binding proteins

(A) A phylogeny of the redesigned copper-binding proteins. The Csp1 template is shown in white in its tetrameric configuration (gray protomers) and the cysteine side chains at the core are depicted in a ball-and-stick representation (PDB: 5FJE). The polar side chains introduced in the first-generation design (plr1 model) are shown, yielding a monomeric, positively supercharged protein. Starting from the plr1 sequence, second-generation designs belong to three classes: core-repacked designs where either three (cr3) or six (cr61 and cr62) core cysteine residues were eliminated or negatively supercharged designs (neg1, neg2). Side chains colored yellow are cysteines, blue are positively charged, red are negatively charged, green are polar, and purple are non-polar.

(B and C) Competitive Cu^{2+} binding assays using zincon for plr1 and neg1 designs show sub-femtomolar dissociation constants, whereby neg1 binds Cu^{2+} more than 3-fold tighter than plr1. Error bars represent SD across two replicates.

(D) The Cu^{2+} release upon incubating protein:Cu^{2+} complexes in 4-fold diluted serum. Plr1 remains associated with Cu^{2+} despite the lower initial affinity, followed by cooperative dissociation. neg1, on the other hand, displays higher initial affinity, but faster Cu^{2+} dissociation. This highlights the higher proteolytic resistance of plr1 compared with neg1, which corresponds to their expected thermostabilities (Figures 3B and S9B). Shading represents SD across three replicates.

and 4C), pointing to the possible stabilization of the metal:protein complex by negative charges. However, the thermal stability of the neg1 apoprotein decreased in comparison to plr1, where an earlier melting transition was observed for neg1, despite its reversible unfolding (Figure S9B). This affinity/stability trade-off was also evident when the copper-binding stability was assessed in untreated fetal bovine serum *in vitro* (Figure 4D). Whereas neg1 initially bound more copper ions than plr1, it released copper faster. Fitting to a first-order decay model yields dissociation rate constants in 4-fold diluted serum of 0.094 h^{-1} and 0.054 h^{-1} for neg1 and plr1, respectively. Notably, plr1 displayed a more complex copper dissociation behavior with a possible cooperative dissociation step, which might be due to accelerated chemical degradation upon copper ion release. These results guide further, more detailed, investigation of protein charge tuning and *trans*-chelation to maximize the binding stability of the radioisotope.

Radiolabeling of proteins (e.g., tumor-targeting antibodies) for PET imaging is not only important for routine tumor-imaging applications, but also for tracking the biodistribution of protein- and cell-based therapies *in vivo*. Currently, associating a metallic radionuclide (such as $^{64}Cu^{2+}$) to a protein is mostly performed

through chemical coupling of chelating agents (e.g., DOTA) to the protein of interest (e.g., NHS coupling).[48] This undirected chemical coupling requires additional processing steps and introduces positional and stoichiometric heterogeneity of the labeled proteins, lowering their fidelity and usefulness for clinical applications. Conversely, our designed copper binders can be used as genetically encodable PET labeling tags that can be expressed on a target cell surface or as a single-chain fusion with the protein of interest. Given the high affinity of these proteins to Cu^{2+} ions, they can be readily loaded with copper radionuclides under mild conditions, greatly simplifying the radiolabeling procedure.

Limitations of the study

Currently, the described framework is restricted to fixed-backbone sequence design. However, we foresee that further developments of the framework will allow it to support motion. Also, while in its current state the framework applies isotropic charges, implementing more advanced electrostatic potential to describe the rotamer library can greatly enhance the scoring accuracy, at no added run-time cost.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL DETAILS
  - Bacterial protein expression system
  - Cell line
  - Zebrafish
- METHOD DETAILS
  - The rotamer library
  - The energy function
  - Evaluation of scoring accuracy against stability benchmarks
  - Evaluation of rotamer and sequence recovery
  - Design of EGFR inhibitors
  - Purification of EGFR inhibitors
  - Thermostability analysis of EGFR inhibitors
  - Surface plasmon resonance binding assays
  - A431 cell proliferation assay
  - Effect of EGFR inhibitors on zebrafish embryos
  - Design of copper-binding proteins
  - Purification of copper-binding proteins
  - Thermostability analysis of copper binders
  - Cu^{2+}-binding affinity, capacity, and stability
  - Loading of copper binders with ^{64}Cu
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2023.100560.

ACKNOWLEDGMENTS

This project has received funding from the IMPRS (K.M.), the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement 863952 [ACE-OF-SPACE]) (P.M.), the M. Schickedanz Kinderkrebsstiftung (M.E., J.S., and N.A.), the "German Universities Excellence Initiative" of Tübingen University (J.S.), and the Deutsche Forschungsgemeinschaft (DFG; no. 500215849) (M.E., B.H.A., and J.S.). Some computations described in this work were performed on the HPC system Raven at the Max Planck Computing and Data Facility. We acknowledge support from the Open Access Publishing Fund of the University of Tübingen. We also thank Dominik Seyfried and Johannes Kinzler for excellent technical assistance and Prof. Gerald Reischl for the supply of ^{64}Cu.

AUTHOR CONTRIBUTIONS

M.E. conceptualized the computational design framework and developed the software. C.B., K.M., M.E., and T.M.H.D. improved the energy function and performed and analyzed MD simulations. J.S. and P.M. selected design targets. K.M. and M.E. performed the computational design. A.M., B.H.A., K.M., M.E., N.B.-B., and T.U. experimentally characterized the designed proteins. N.A. designed, performed, and analyzed the results of zebrafish experiments. A.M., J.S., K.M., M.E., and N.A. wrote, reviewed, and edited the manuscript. A.N.L., J.S., M.E., P.M., and T.M.H.D. contributed to supervision and project administration. A.N.L., A.M., J.S., M.E., and P.M. provided resources. A.N.L., J.S., M.E., and P.M. acquired funding.

DECLARATION OF INTERESTS

The designed EGFR inhibitors and the copper binders described in this study are part of patent applications nos. EP22190708 (inventors: K.M., M.E., and J.S.) and EP22206059 (inventors: M.E. and J.S.), respectively. These applications were filed by Eberhard Karls Universität Tübingen and Max-Planck-Gesellschaft zur Förderung der Wissenschaften e.V. M.E. is a co-founder of Heliopolis Biotechnology LLC.

Received: January 22, 2023
Revised: May 25, 2023
Accepted: July 21, 2023
Published: August 15, 2023

REFERENCES

1. Fleishman, S.J., and Baker, D. (2012). Role of the Biomolecular Energy Gap in Protein Design, Structure, and Evolution. Cell *149*, 262–273. https://doi.org/10.1016/j.cell.2012.03.016.
2. Kuhlman, B., and Bradley, P. (2019). Advances in protein structure prediction and design. Nat. Rev. Mol. Cell Biol. *20*, 681–697. https://doi.org/10.1038/s41580-019-0163-x.
3. Alford, R.F., Leaver-Fay, A., Jeliazkov, J.R., O'Meara, M.J., DiMaio, F.P., Park, H., Shapovalov, M.V., Renfrew, P.D., Mulligan, V.K., Kappel, K., et al. (2017). The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. J. Chem. Theor. Comput. *13*, 3031–3048. https://doi.org/10.1021/acs.jctc.7b00125.
4. Rohl, C.A., Strauss, C.E.M., Misura, K.M.S., and Baker, D. (2004). Protein Structure Prediction Using Rosetta. In Methods in Enzymology (Academic Press), pp. 66–93. https://doi.org/10.1016/S0076-6879(04)83004-0.
5. Sheffler, W., and Baker, D. (2009). RosettaHoles: rapid assessment of protein core packing for structure prediction, refinement, design, and validation. Protein Sci. *18*, 229–239. https://doi.org/10.1002/pro.8.
6. Zhou, J., Panaitiu, A.E., and Grigoryan, G. (2020). A general-purpose protein design framework based on mining sequence-structure relationships in known protein structures. Proc. Natl. Acad. Sci. USA *117*, 1059–1068. https://doi.org/10.1073/pnas.1908723117.

7. Norn, C., Wicky, B.I.M., Juergens, D., Liu, S., Kim, D., Tischer, D., Koepnick, B., Anishchenko, I., Foldit Players; Baker, D., and Ovchinnikov, S. (2021). Protein sequence design by conformational landscape optimization. Proc. Natl. Acad. Sci. USA *118*, e2017228118. https://doi.org/10.1073/pnas.2017228118.

8. Vokes, E.E., and Chu, E. (2006). Anti-EGFR therapies: clinical experience in colorectal, lung, and head and neck cancers. Oncology *20*, 15–25.

9. Dunbrack, R.L., Jr. (2002). Rotamer libraries in the 21st century. Curr. Opin. Struct. Biol. *12*, 431–440. https://doi.org/10.1016/s0959-440x(02)00344-5.

10. Towse, C.-L., Rysavy, S.J., Vulovic, I.M., and Daggett, V. (2016). New Dynamic Rotamer Libraries: Data-Driven Analysis of Side-Chain Conformational Propensities. Structure *24*, 187–199. https://doi.org/10.1016/j.str.2015.10.017.

11. Childers, M.C., Towse, C.L., and Daggett, V. (2016). The effect of chirality and steric hindrance on intrinsic backbone conformational propensities: tools for protein design. Protein Eng. Des. Sel. *29*, 271–280. https://doi.org/10.1093/protein/gzw023.

12. Vitalini, F., Noé, F., and Keller, B.G. (2016). Molecular dynamics simulations data of the twenty encoded amino acids in different force fields. Data Brief *7*, 582–590. https://doi.org/10.1016/j.dib.2016.02.086.

13. Conway, P., Tyka, M.D., DiMaio, F., Konerding, D.E., and Baker, D. (2014). Relaxation of backbone bond geometry improves protein energy landscape modeling. Protein Sci. *23*, 47–55. https://doi.org/10.1002/pro.2389.

14. MacKerell, A.D., Jr., Banavali, N., and Foloppe, N. (2000). Development and current status of the CHARMM force field for nucleic acids. Biopolymers *56*, 257–265. https://doi.org/10.1002/1097-0282(2000)56:4<257::Aid-bip10029>3.0.Co;2-w.

15. Childers, M.C., Towse, C.-L., and Daggett, V. (2018). Molecular dynamics-derived rotamer libraries for d-amino acids within homochiral and heterochiral polypeptides. Protein Eng. Des. Sel. *31*, 191–204. https://doi.org/10.1093/protein/gzy016.

16. Shapovalov, M.V., and Dunbrack, R.L., Jr. (2011). A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure *19*, 844–858. https://doi.org/10.1016/j.str.2011.03.019.

17. Kim, D.E., Blum, B., Bradley, P., and Baker, D. (2009). Sampling Bottlenecks in De novo Protein Structure Prediction. J. Mol. Biol. *393*, 249–260. https://doi.org/10.1016/j.jmb.2009.07.063.

18. Nisthal, A., Wang, C.Y., Ary, M.L., and Mayo, S.L. (2019). Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis. Proc. Natl. Acad. Sci. USA *116*, 16367–16377. https://doi.org/10.1073/pnas.1903888116.

19. Tsuboyama, K., Dauparas J., Chen J., Laine E., Mohseni Behbahani Y., Weinstein J.J., Mangan N.M., Ovchinnikov S., and Rocklin G.J. (2023). Mega-scale experimental analysis of protein folding stability in biology and design. Nature. https://doi.org/10.1038/s41586-023-06328-6.

20. Romero, P.A., and Arnold, F.H. (2009). Exploring protein fitness landscapes by directed evolution. Nat. Rev. Mol. Cell Biol. *10*, 866–876. https://doi.org/10.1038/nrm2805.

21. Leaver-Fay, A., Tyka, M., Lewis, S.M., Lange, O.F., Thompson, J., Jacak, R., Kaufman, K.W., Renfrew, P.D., Smith, C.A., Sheffler, W., et al. (2011). Chapter nineteen - Rosetta3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. In Computer Methods, Part C, M.L. Johnson and L. Brand, eds. (Academic Press), pp. 545–574. https://doi.org/10.1016/B978-0-12-381270-4.00019-6.

22. ElGamacy, M., Riss, M., Zhu, H., Truffault, V., and Coles, M. (2019). Mapping Local Conformational Landscapes of Proteins in Solution. Structure *27*, 853–865.e5. https://doi.org/10.1016/j.str.2019.03.005.

23. Weiel, M., Götz, M., Klein, A., Coquelin, D., Floca, R., and Schug, A. (2021). Dynamic particle swarm optimization of biomolecular simulation parameters with flexible objective functions. Nat. Mach. Intell. *3*, 727–734. https://doi.org/10.1038/s42256-021-00366-3.

24. Woodburn, J.R. (1999). The epidermal growth factor receptor and its inhibition in cancer therapy. Pharmacol. Ther. *82*, 241–250. https://doi.org/10.1016/s0163-7258(98)00045-x.

25. Ogiso, H., Ishitani, R., Nureki, O., Fukai, S., Yamanaka, M., Kim, J.-H., Saito, K., Sakamoto, A., Inoue, M., Shirouzu, M., and Yokoyama, S. (2002). Crystal Structure of the Complex of Human Epidermal Growth Factor and Receptor Extracellular Domains. Cell *110*, 775–787. https://doi.org/10.1016/S0092-8674(02)00963-7.

26. Chong, C.R., and Jänne, P.A. (2013). The quest to overcome resistance to EGFR-targeted therapies in cancer. Nat. Med. *19*, 1389–1400. https://doi.org/10.1038/nm.3388.

27. Chan, D.L.H., Segelov, E., Wong, R.S., Smith, A., Herbertson, R.A., Li, B.T., Tebbutt, N., Price, T., and Pavlakis, N. (2017). Epidermal growth factor receptor (EGFR) inhibitors for metastatic colorectal cancer. Cochrane Database Syst. Rev. *6*, CD007047. https://doi.org/10.1002/14651858.CD007047.pub2.

28. Schrank, Z., Chhabra, G., Lin, L., Iderzorig, T., Osude, C., Khan, N., Kuckovic, A., Singh, S., Miller, R.J., and Puri, N. (2018). Current Molecular-Targeted Therapies in NSCLC and Their Mechanism of Resistance. Cancers *10*, 224.

29. Guardiola, S., Varese, M., Sánchez-Navarro, M., and Giralt, E. (2019). A Third Shot at EGFR: New Opportunities in Cancer Therapy. Trends Pharmacol. Sci. *40*, 941–955. https://doi.org/10.1016/j.tips.2019.10.004.

30. Yotsumoto, F., Sanui, A., Fukami, T., Shirota, K., Horiuchi, S., Tsujioka, H., Yoshizato, T., Kuroki, M., and Miyamoto, S. (2009). Efficacy of ligand-based targeting for the EGF system in cancer. Anticancer Res. *29*, 4879–4885.

31. Sarup, J., Jin, P., Turin, L., Bai, X., Beryt, M., Brdlik, C., Higaki, J.N., Jorgensen, B., Lau, F.W., Lindley, P., et al. (2008). Human epidermal growth factor receptor (HER-1:HER-3) Fc-mediated heterodimer has broad antiproliferative activity in vitro and in human tumor xenografts. Mol. Cancer Therapeut. *7*, 3223–3236. https://doi.org/10.1158/1535-7163.MCT-07-2151.

32. Li, Z., Li, Y., Chang, H.-P., Chang, H.-Y., Guo, L., and Shah, D.K. (2019). Effect of Size on Solid Tumor Disposition of Protein Therapeutics. Drug Metab. Dispos. *47*, 1136–1145. https://doi.org/10.1124/dmd.119.087809.

33. Lax, I., Fischer, R., Ng, C., Segre, J., Ullrich, A., Givol, D., and Schlessinger, J. (1991). Noncontiguous regions in the extracellular domain of EGF receptor define ligand-binding specificity. Cell Regul. *2*, 337–345. https://doi.org/10.1091/mbc.2.5.337.

34. Baselga, J. (2001). The EGFR as a target for anticancer therapy—focus on cetuximab. Eur. J. Cancer *37*, 16–22. https://doi.org/10.1016/S0959-8049(01)00233-7.

35. ElGamacy, M. (2022). Accelerating therapeutic protein design. In Advances in Protein Chemistry and Structural Biology (Academic Press). https://doi.org/10.1016/bs.apcsb.2022.01.004.

36. Hoesl, C., Röhrl, J.M., Schneider, M.R., and Dahlhoff, M. (2018). The receptor tyrosine kinase ERBB4 is expressed in skin keratinocytes and influences epidermal proliferation. Biochim. Biophys. Acta Gen. Subj. *1862*, 958–966. https://doi.org/10.1016/j.bbagen.2018.01.017.

37. Pruvot, B., Curé, Y., Djiotsa, J., Voncken, A., and Muller, M. (2014). Developmental defects in zebrafish for classification of EGF pathway inhibitors. Toxicol. Appl. Pharmacol. *274*, 339–349. https://doi.org/10.1016/j.taap.2013.11.006.

38. Malmstrom, B.G., and Neilands, J.B. (1964). Metalloproteins. Annu. Rev. Biochem. *33*, 331–354. https://doi.org/10.1146/annurev.bi.33.070164.001555.

39. Lu, Y., Yeung, N., Sieracki, N., and Marshall, N.M. (2009). Design of functional metalloproteins. Nature *460*, 855–862. https://doi.org/10.1038/nature08304.

40. Chalkley, M.J., Mann, S.I., and DeGrado, W.F. (2022). De novo metalloprotein design. Nat. Rev. Chem *6*, 31–50. https://doi.org/10.1038/s41570-021-00339-5.

41. Ellisman, M.H., Deerinck, T.J., Shu, X., and Sosinsky, G.E. (2012). Picking faces out of a crowd: genetic labels for identification of proteins in correlated light and electron microscopy imaging. Methods Cell Biol. *111*, 139–155. https://doi.org/10.1016/b978-0-12-416026-2.00008-x.

42. Matsumoto, Y., and Jasanoff, A. (2013). Metalloprotein-based MRI probes. FEBS Lett. *587*, 1021–1029. https://doi.org/10.1016/j.febslet.2013.01.044.

43. Sawyer, J.R., Tucker, P.W., and Blattner, F.R. (1992). Metal-binding chimeric antibodies expressed in Escherichia coli. Proc. Natl. Acad. Sci. USA *89*, 9754–9758. https://doi.org/10.1073/pnas.89.20.9754.

44. Keinänen, O., Fung, K., Brennan, J.M., Zia, N., Harris, M., van Dam, E., Biggin, C., Hedt, A., Stoner, J., Donnelly, P.S., et al. (2020). Harnessing 64Cu/67Cu for a theranostic approach to pretargeted radioimmunotherapy. Proc. Natl. Acad. Sci. USA *117*, 28316–28327. https://doi.org/10.1073/pnas.2009960117.

45. ElGamacy, M., and Hernandez Alvarez, B. (2021). Expanding the versatility of natural and de novo designed coiled coils and helical bundles. Curr. Opin. Struct. Biol. *68*, 224–234. https://doi.org/10.1016/j.sbi.2021.03.011.

46. Vita, N., Platsaki, S., Baslé, A., Allen, S.J., Paterson, N.G., Crombie, A.T., Murrell, J.C., Waldron, K.J., and Dennison, C. (2015). A four-helix bundle stores copper for methane oxidation. Nature *525*, 140–143. https://doi.org/10.1038/nature14854.

47. Pham, N.B., and Meng, W.S. (2020). Protein aggregation and immunogenicity of biotherapeutics. Int. J. Pharm. *585*, 119523. https://doi.org/10.1016/j.ijpharm.2020.119523.

48. Rolle, A.-M., Hasenberg, M., Thornton, C.R., Solouk-Saran, D., Männ, L., Weski, J., Maurer, A., Fischer, E., Spycher, P.R., Schibli, R., et al. (2016). ImmunoPET/MR imaging allows specific detection of Aspergillus fumigatus lung infection in vivo. Proc. Natl. Acad. Sci. USA *113*, E1026–E1033. https://doi.org/10.1073/pnas.1518836113.

49. Phillips, J.C., Hardy, D.J., Maia, J.D.C., Stone, J.E., Ribeiro, J.V., Bernardi, R.C., Buch, R., Fiorin, G., Hénin, J., Jiang, W., et al. (2020). Scalable molecular dynamics on CPU and GPU architectures with NAMD. J. Chem. Phys. *153*, 044130. https://doi.org/10.1063/5.0014475.

50. Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: visual molecular dynamics. J. Mol. Graph. *14*, 33–38.27.28. https://doi.org/10.1016/0263-7855(96)00018-5.

51. Bottaro, S., Lindorff-Larsen, K., and Best, R.B. (2013). Variational Optimization of an All-Atom Implicit Solvent Force Field to Match Explicit Solvent Simulation Data. J. Chem. Theor. Comput. *9*, 5641–5652. https://doi.org/10.1021/ct400730n.

52. Stone, J.E., Phillips, J.C., Freddolino, P.L., Hardy, D.J., Trabuco, L.G., and Schulten, K. (2007). Accelerating molecular modeling applications with graphics processors. J. Comput. Chem. *28*, 2618–2640. https://doi.org/10.1002/jcc.20829.

53. Pritchard, R.B., and Hansen, D.F. (2019). Characterising side chains in large proteins by protonless 13 C-detected NMR spectroscopy. Nat. Commun. *10*, 1747. https://doi.org/10.1038/s41467-019-09743-4.

54. Peterson, L.X., Kang, X., and Kihara, D. (2014). Assessment of protein side-chain conformation prediction methods in different residue environments. Proteins *82*, 1971–1984. https://doi.org/10.1002/prot.24552.

55. Hernandez Alvarez, B., Skokowa, J., Coles, M., Mir, P., Nasri, M., Maksymenko, K., Weidmann, L., Rogers, K.W., Welte, K., Lupas, A.N., et al. (2020). Design of novel granulopoietic proteins by topological rescaffolding. PLoS Biol. *18*, e3000919. https://doi.org/10.1371/journal.pbio.3000919.

56. Skokowa, J., Hernandez Alvarez, B., Coles, M., Ritter, M., Nasri, M., Haaf, J., Aghaallaei, N., Xu, Y., Mir, P., Krahl, A.-C., et al. (2022). A topological refactoring design strategy yields highly stable granulopoietic proteins. Nat. Commun. *13*, 2948. https://doi.org/10.1038/s41467-022-30157-2.

57. Kocyła, A., Pomorski, A., and Krężel, A. (2017). Molar absorption coefficients and stability constants of Zincon metal complexes for determination of metal ions and bioinorganic applications. J. Inorg. Biochem. *176*, 53–65. https://doi.org/10.1016/j.jinorgbio.2017.08.006.

58. Griessinger, C.M., Maurer, A., Kesenheimer, C., Kehlbach, R., Reischl, G., Ehrlichmann, W., Bukala, D., Harant, M., Cay, F., Brück, J., et al. (2015). 64Cu antibody-targeting of the T-cell receptor and subsequent internalization enables in vivo tracking of lymphocytes by PET. Proc. Natl. Acad. Sci. USA *112*, 1161–1166. https://doi.org/10.1073/pnas.1418391112.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Cetuximab	MedChemExpress	Cat#HY-P9905
Bacterial and virus strains		
BL21(DE3) Competent Cells - Novagen	Sigma-Aldrich	Cat#69450
Chemicals, peptides, and recombinant proteins		
cOmplete, EDTA-free Protease Inhibitor Cocktail	Roche	Cat#5056489001
DNase I	PanReac AppliChem	Cat#A3778
DMEM, high glucose, pyruvate	Gibco	Cat#41966029
Fetal Bovine Serum, certified, heat inactivated	Gibco	Cat#10082147
DPBS, no calcium, no magnesium	Gibco	Cat#14190144
Alcian Blue 8 GX	Sigma-Aldrich	Cat#A5268
Zincon monosodium salt	Supelco	Cat#96440
Recombinant Human EGF	PeproTech	Cat#AF-100-15
Recombinant Human HB-EGF	R&D Systems	Cat#259-HE-050/CF
Recombinant Human TGF-alpha	R&D Systems	Cat#239-A-100
Critical commercial assays		
CellTiter-Blue® Cell Viability Assay	Promega	Cat#G8080
Deposited data		
Molecular dynamics simulation data for amino acids in explicit water performed under the CHARMM27 force field	Vitalini et al.[12]	ftp://bdg.chemie.fu-berlin.de/Ac-X-NHMe/
Thermodynamic stability data for protein G mutants	Nisthal et al.[18]	ProtaBank ID: gwoS2haU3
Protein folding stability data for protein mutants measured by cDNA display proteolysis	Tsuboyama et al.[19]	https://doi.org/10.5281/zenodo.7844779
Experimental models: Cell lines		
A431 Cell Line human	ECACC	Cat#85090402; RRID:CVCL_0037
Experimental models: Organisms/strains		
Zebrafish: wild-type: AB	N/A	RRID:ZIRC_ZL1
Software and algorithms		
Damietta	This paper	https://doi.org/10.5281/zenodo.8152656
NAMD	Phillips et al.[49]	RRID:SCR_014894
VMD	Humphrey et al.[50]	RRID:SCR_001820
Prism	GraphPad Software, Inc.	RRID:SCR_002798
PyMOL	Schrödinger, Inc.	RRID:SCR_000305
Biacore X100 Evaluation Software	Cytiva	RRID:SCR_015936
Other		
Gene synthesis	Synbio Technologies, Inc., BioCat GmbH	N/A
Millex-HV Filter, 0.45 μm, PVDF, 33 mm	Millipore	Cat# SLHV033RS
Amicon Ultra-15 centrifugal filter unit, 10 kDa	Millipore	Cat# UFC901024
HisTrap Excel column, 5 ml	Cytiva	Cat#GE17-3712-06
Superdex 75 Increase 10/300 GL column	Cytiva	Cat#29148721
HiTrap Capto Q column, 5 ml	Cytiva	Cat#11001303
HiTrap Capto S column, 5 ml	Cytiva	Cat#17544123

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Prometheus Standard Capillaries	Nanotemper	Cat#PR-C002
Sensor Chip CM5	Cytiva	Cat#BR100012
96 Well TC-Treated Microplates	Corning	Cat#3596
UV-Star plate, 384 well, F-bottom	Greiner	Cat#781801

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Mohammad ElGamacy (mohammad.elgamacy@med.uni-tuebingen.de).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.
- Damietta software is available at https://bio.mpg.de/damietta. An archival DOI is listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL DETAILS

Bacterial protein expression system

BL21(DE3) competent *E. coli* cells were used for transformation and high-level protein expression using a T7 RNA polymerase-IPTG induction system.

Cell line

A431 cells (ECACC 85090402; RRID: CVCL_0037) were cultured in DMEM medium supplemented with 10% fetal bovine serum (FBS), at 37 °C, 5% CO_2. Sub-confluent cultures were split 1:10 twice a week.

Zebrafish

Wild-type zebrafish line (AB, RRID:ZIRC_ZL1) was used for the experiments. Treatment of zebrafish embryos with tested inhibitors started at 4-6 hpf and continued until 4 dpf. The sex of the embryos was not taken into consideration. Zebrafish were maintained according to standard protocols and handled in accordance with European Union animal protection directive 2010/63/EU and approved by the local government (Tierschutzgesetz §11, Abs. 1, Nr. 1, husbandry permit 35/9185.46/Uni TÜ).

METHOD DETAILS

The rotamer library

Large-scale molecular dynamics simulations of capped amino acids (i.e. Ac-X-NHMe, where X represents the amino acid symbol) were used as conformational pools from which representative rotamers were sampled. These trajectories can be performed under a predefined force field in explicit water, yielding a pool of freely inter-changing conformations. In the current implementation, we have used a set of trajectories performed under the CHARMM27 force field in explicit solvent by Vitalini et al.[12] The trajectory of every Ac-X-NHMe amino acid is 1-μs-long and was uniformly partitioned into 36 (φ, ψ) bins (within (±60, ±60) intervals). The propensity of each backbone conformational state (i.e. a (φ, ψ) bin) was used to represent its relative energy with respect to all of the other conformational bins as:

$$\Delta G_{pp} = -k_b T \ln \frac{O_{(\varphi,\psi)_m}}{N} \quad \text{(Equation 1)}$$

Where k_b is the Boltzmann constant in kcal·mol^{-1}K^{-1}, T is the temperature in Kelvins (here given as constant values of 0.001985875 and 298, respectively), $O_{(\varphi,\psi)_m}$ is the number of observations in the m^{th} (φ, ψ) conformation bin, and N is the total number of conformations. A similar explicit internal energy term is used to describe the side chain conformational preference, where all conformations within each $(\varphi, \psi)_m$ bin are aligned to a single frame-of-references using the three atoms C^{β} (or Gly $H^{\alpha 1}$), C^{α}, N. In this frame of

reference C^{α} is positioned at the origin, $C^{\alpha} \rightarrow C^{\beta}$ (or $C^{\alpha} \rightarrow H^{\alpha 1}$) vector is aligned along the z-axis, and the $C^{\alpha} \rightarrow N$ vector lies along the xy-plain. The aligned conformers undergo a 3D *k*-means clustering including all of their atomic coordinates, resulting in *k* representative side chain conformers at the center of each cluster. Here the values of *k* was set to 50, and 100 conformational clusters to build the 50- and 100-rotamer libraries, respectively. The representative conformer of each cluster was taken to be the one with the lowest RMSD to the average structure of the entire cluster, where the energy of every cluster is defined as:

$$\Delta G_k = -k_b T \ln \frac{O_{(k_n|(\varphi,\psi)_m)}}{O_{(\varphi,\psi)_m}} \quad \text{(Equation 2)}$$

Where $O_{(k_n|(\varphi,\psi)_m)}$ is the number of observations in conformation bin $(\varphi,\psi)_m$ of the n^{th} cluster out of *k* clusters. In cases where the entire molecular dynamics simulation results in a conformational bin that is underpopulated, i.e. $O_{(\varphi,\psi)_m} < k$, the entire bin is not represented in the library.

The energy function

The energy function is composed of 5 terms representing the energy difference to a ground state of a solvated, capped amino acid, as follows: backbone internal energy (ΔG_{pp}, Equation 1), side chain conformational energy (ΔG_k, Equation 2), Lennard-Jones interaction energy (ΔG_{LJ}), solvation energy (ΔG_{solv}), and electrostatic interaction energy (ΔG_{elec}). The total energy is a weighted sum of these terms, as:

$$\Delta G_{total} = w_{pp}\Delta G_{pp} + w_k \Delta G_k + w_{LJ}(\Delta G_{LJ} - \Delta G_{LJ,ref}) + w_{solv}(\Delta G_{solv} - \Delta G_{solv,ref}) + w_{solv}(\Delta G_{elec} - \Delta G_{elec,ref}) \quad \text{(Equation 3)}$$

The energy calculation scheme follows a two-body formulation whereby the interactions are only calculated between two sets of atoms belonging to the *1st-body* and the *2nd-body*. The *1st-body* atoms are all the atoms within a bounding box of dimensions $d_{box} \times d_{box} \times d_{box}$ excluding the side chain atoms of the mutable residue, where the bounding box is centered at the Cα atom of the mutable residue. The *2nd-body* atoms represent the side chains of the rotamer to be placed in the mutable residue position, as sampled from the rotamer library. This scheme applies to the interaction energy terms (ΔG_{LJ} and ΔG_{elec}) as well as the solvation free energy term (ΔG_{solv}). The non-bonded energy terms between the inbound sidechain atoms and the environment are corrected by subtracting a reference energy ($\Delta G_{LJ,ref}$, $\Delta G_{solv,ref}$, and $\Delta G_{elec,ref}$; Equation 3). These reference values describe the interaction energies between the side chain of a rotamer and its backbone atoms in the respective conformation pooled from the MD. These reference energy values are precomputed with the rotamer library and are subtracted from the final interaction energy before it is weighted (Equation 3). To preserve the compatibility between different energy terms, the partial charges and Lennard-Jones parameters are also obtained from the CHARMM27 force field parameters,[14] and the surface-area based solvation energy term relied on the CHARMM-based parameters for the EEF1-SB model.[51] Moreover, the ΔG_{pp} (Equation 1) and ΔG_k (Equation 2) terms, are derived from conformational distributions extracted from simulations that used the CHARMM27 force field.[12]

For evaluating electrostatic interaction energies, an approximation of the Generalised Born model was used. The interactions are calculated between the *1st-body* atoms $i \in I$ of the protein where the mutable residue side chain atoms are deleted, and *I* protein atoms exist within a bounding simulation box, and the *2nd-body* atoms $j \in J$ that constitute the inbound side chain atoms looked up from the rotamer library. The electrostatics function was composed of three terms as follows:

$$\Delta G_{elec} = \frac{332 q_i q_j}{\varepsilon(r) r_{ij}} \left(1 - \frac{\varepsilon_p \bar{r}}{\left(\bar{r}_{ij}^2 + \bar{b}^2 e^{-\frac{\bar{r}_{ij}^2}{4\bar{b}^2}} \right)^{\frac{1}{2}}} \left(\frac{1}{\varepsilon_p} - \frac{1}{\varepsilon_s} \right) \right) - 166 \frac{q_i^2}{\bar{b}} \left(\frac{1}{\varepsilon_p} - \frac{1}{\varepsilon_s} \right) \quad \text{(Equation 4)}$$

Where q_i and q_j represent the partial charges of atoms *i* and *j*, separated by the distance r_{ij}. A distance-dependent dielectric function of $\varepsilon(r) = r$ was assumed. As the electrostatic interactions cutoff was set to 7.0 Å, the value of $\varepsilon(r)$ ranged as $\sigma_{LJ} < \varepsilon(r) < 7.0$, where σ_{LJ} of a carbon-carbon interaction, for example, is approximately 4 Å. ε_p and ε_s represent the dielectric constant of protein core ($\varepsilon_p = 8$) and water ($\varepsilon_s = 75$), respectively.

Equation 4 was derived in this form to reduce the computing cost, whereby the first term is precomputed for 2nd body as $\frac{322 q_j}{\varepsilon(r) r}$, that is multiplied by the 1st body partial charges tensor (i.e. q_i), which is computed rapidly on the fly. The second and third terms represent the charge solvation corrections according to the Generalised Born model and are computed as a tensor scalar and an additive term, respectively. In the second term, the average interatomic distance in the simulation cube $\bar{r}_{ij}^2$ was used instead of the individual distances, and born radius was taken as the average born radius in the simulation cube; $b_i = b_j = \bar{b}$. The average interatomic distance within the simulation cube $\bar{r}_{ij}$ was set to $0.6617 d_{box}$, while the average born radius $\bar{b}$ was approximated according to the fraction of filled volume of the simulation cube (i.e. $d_{box} \frac{V_{filled}}{V_{total}}$).

The Lennard-Jones function was implemented as a piece-wise function to avoid sensitivity to interatomic clashes, which can be relaxed upon MD-based minimization. The piece-wise function consists of three components; a standard LJ term in the attractive range of inter-atomic distances, a slow-growing repulsive term across a band of the atomic crust, and a flat maximum at a defined atomic core, as follows:

$$\Delta G_{LJ} = \begin{cases} \varepsilon_{LJ,j}\left(\frac{\sigma_{LJ,j}^{12}}{r_{ij}^{12}} - \frac{\sigma_{LJ,j}^{6}}{r_{ij}^{6}}\right) & (\sigma_{LJ,j} - c_{soft}) < r_{ij} \leq c_{lr} \\ \varepsilon_{LJ,j}\frac{\sigma_{LJ,j}^{2}}{r_{ij}^{2}} & (\sigma_{LJ,j} - c_{hard}) < r_{ij} \leq (\sigma_{LJ,j} - c_{soft}) \\ \varepsilon_{LJ,j}\frac{\sigma_{LJ,j}^{4}}{(\sigma_{LJ,j} - c_{hard})^{4}} & 0 < r_{ij} \leq (\sigma_{LJ,j} - c_{hard}) \end{cases} \quad \text{(Equation 5)}$$

Where $\varepsilon_{LJ,j}$ and $\sigma_{LJ,j}$ are the minimum LJ energy value (in kcal·mol^{-1}) and the LJ radius (in Å) of atom j as obtained from the CHARMM27 parameters.[14] The cutoffs c_{soft} and c_{hard} were set to 0.25 Å and $\sigma_{LJ,j}/2$, respectively. The use of $\varepsilon_{LJ,j}$ and $\sigma_{LJ,j}$ parameters of atom j instead of the averaged parameters for atoms i and j was aimed at lowering the computing cost, since the entire LJ interaction fields are pre-calculated for inbound rotamer atoms (i.e. $j \in J$ atoms). Both the LJ and electrostatic interaction fields are calculated for all values of $0 < r_{ij} \leq c_{lr}$ at a resolution of 0.5 Å, where c_{lr} is the long-range cutoff that is set to 7.0 Å. Such fields are stored in the chemical library provided with the software, and are looked up during the design process depending on the mutable position (φ, ψ) bin.

These interactions are calculated within a cube where a single side of the cube (d_{box}) has the size of 22 Å, containing 44× 44× 44 voxels, where the voxel resolution is 0.5 Å. To evaluate the impact of varying the voxel resolution on the calculation accuracy, we emulated the interaction of carbon atoms at varying resolutions. The results showed most of the energy error to stem from the repulsive part of the function (i.e. $r_{ij} < \sigma_{LJ,ij}$), which was largely positive in value as r_{ij} is floored to the nearest discrete bin (Figure S10A). The results also showed the energy error to substantially decrease above a resolution of 0.4 Å (Figure S10A). It worth highlighting that this approach is dissimilar to other parallel energy computations that are primarily thread-based, where a constant set of instructions can access a large number of shared variables (here; atom attributes) across the main (or a GPU) memory.[52] Instead, the presented framework simplifies these calculations by encoding most of the energy function information into smooth fields associated with each discrete rotamer (Figure S10B). This leaves only populating the positions of the environment atoms as the quick on-the-fly step that is performed once for each designable position (Figure S10C). This renders the total energy much faster to compute through only 2 instructions and 2 variables; as $E_{LJ} = \sum(v_{i,full} \odot v_{j \in J})$ (Figure S10D). Our method further stacks all rotamer fields belonging to one side chain into a single variable (i.e. 4D tensors encoding 100 *rotamers* × dim_x × dim_y × dim_z) in order to reduce the number of memory calls. Unlike the standard way of computing the LJ function, this operation avoids any exponentiation or square roots, and offers substantially higher arithmetic intensity.

A generic solvation free energy term based on a surface area method was also put in place to account for the hydrophobic effect. This term was adapted from the EEF1-SB energy model parameters,[51] and follows the form:

$$\Delta G_{solv} = \sum_{I} \sigma_{solv,I} A_I(\boldsymbol{r_I}) \quad \text{(Equation 6)}$$

Where $\sigma_{solv,I}$ is the solvation energy per unit surface area (in kcal·mol^{-1}Å^{-2}) of atom I, with solvent-exposed surface area A_I when located at position vector $\boldsymbol{r_I}$. An approximation of the $A_I(\boldsymbol{r_I})$ function is derived from the non-occluded vdW surface area of slightly inflated vdW radii. This radial inflation was performed here by an added 0.5 Å to the atomic vdW radii, while correcting for the $\sigma_{solv,I}$ for the larger atomic surface area to keep the solvation energy per atom type constant. The implementation relied on a voxelized representation of atomic crusts encoding the $\sigma_{solv,I}$ values in tensorial forms as well as core-masking tensors which are used to exclude the occluded atomic surfaces. This energy term is calculated separately for the *1st-body* atoms ($\Delta G_{solv,1}$), the *2nd-body* ($\Delta G_{solv,2}$), and the *1st-body* and *2nd-body* combined ($\Delta G_{solv,1,2}$). These values would represent the solvated free energies of the protein environment with the removed side chain atoms at the mutable residue, the inbound side chain atoms of a rotamer from the rotamer library, and the combined protein environment and rotamer side chain atoms after rotamer placement, respectively. Given these three values, the final solvation free energy term is calculated as follows:

$$\Delta G_{solv} = \Delta G_{solv,1,2} - (\Delta G_{solv,1} + \Delta G_{solv,2}) \quad \text{(Equation 7)}$$

Finally, while the different terms of the energy function are compatible as they were derived from the same force field, the softening of the repulsive component of the LJ term necessitates the downscaling of the electrostatic term ($w_{elec} = 0.25$), in order to avoid highly clashing configurations with optimal electrostatic interactions. Additionally, the w_k is recommended to be set to 0 for mutagenesis tasks, and to 1 for repacking tasks. This is as ΔG_k is not directly comparable across different amino acid types, given that different amino acid types have drastically different chemical exchange timeframes in solution, even when their backbone is fixed.[53] Otherwise, the other weighting factors were all set to 1.0; $w_{pp} = w_k = w_{LJ} = w_{solv} = 1.0$. Additionally, to deploy some tiered scoring to avoid calculating all energy terms for highly clashing rotamers, a maximum LJ energy value was set to 5.0 kcal·mol^{-1}.

Evaluation of scoring accuracy against stability benchmarks

Ability of the framework to evaluate stability of protein mutants was benchmarked using five independent previously reported experimental dataset of: 1) mutants of the β1 domain of streptococcal protein G (PDB: 1PGA);[18] 2) mutants of the N-terminal domain of phage 434 repressor (PDB: 1R69); 3) mutants of the SH3 domain in human obscurin (PDB: 1V1C), 4) mutants of the N-terminal domain of ribosomal protein L9 (PDB: 2HBB), and 5) mutants of the r11_829_TrROS protein designed by trRosetta hallucination (an AlphaFold model of the design was used).[19] Generating mutants and estimating their energies was done using a single-point (sp) routine with the parameters (-max_lj 10.0, -w_pp 1.0, -w_k 1.0, -w_lj 1.0, -w_solv 1.0, -w_elec 0.25). $\Delta\Delta G$ was calculated as the difference in free energy between the mutant and the wild type reference. Predictive potential of a software was assessed using a Pearson correlation coefficient (R) for computed energy values against experimental data. For the dataset of Gβ1 mutants, performance of Damietta was analyzed in comparison with performance of Rosetta framework (ddg_monomer application, NoMin protocol) described before by Nisthal et al.[18]

Evaluation of rotamer and sequence recovery

For testing the ability of Damietta single-point mutagenesis routine to recover native side-chain conformations (rotamer recovery) the dataset described by Zhou et al.[6] was used. The dataset consists of 9 pairs of structures, where each pair represents one X-ray and one NMR structure of the same monomeric protein (PDB: 1TVG, 1XPW; 3C4S, 2JZ2; 2O0Q, 2JQN; 2Q00, 2JPU; 3IDU, 2KL6; 3K63, 2KRT; 3FIF, 2JN0; 3H9X, 2KFP; 1TTZ, 1XPV). For NMR entries, first model was used (out of twenty models in each ensemble). For each structure in the dataset, the repack all (ra) routine was run for 10 successive rounds, where side chains of all amino acids were repacked from N- to C-terminus in each round. Rotamer recovery was evaluated in terms of predicted χ_1 and χ_2 side-chain torsion angles. An angle prediction was considered correct if the torsion error was in the range of ±40° from the native angle.[54] χ_1 accuracy was defined as a percentage of residues within a protein with correctly predicted χ_1 angle. $\chi_{1\&2}$ accuracy was defined as a percentage of residues within a protein with correctly predicted both χ_1 and χ_2 angles. Results were presented as a boxplot, where lower and upper hinges represent first and third quartile, respectively, and the whiskers extend from the box by 1.5 times the inter-quartile range (Figure S3). Core residues were identified as residues with solvent exposed surface area less than 5 $Å^2$.

To evaluate the recovery of native amino acid identities (sequence recovery) a dataset of 9 above-mentioned X-ray structures was used (PDB: 1TVG, 3C4S, 2O0Q, 2Q00, 3IDU, 3K63, 3FIF, 3H9X, 1TTZ). Each amino acid position (except of N-terminal and C-terminal residues) was mutated individually using the single-point (sp) routine with the following parameters: 20 target amino acids, -max_lj 10.0, -w_pp 1.0, -w_k 1.0, -w_lj 1.0, -w_solv 1.0, -w_elec 0.25). Sequence recovery was calculated as a percentage of amino acid positions within a protein at which the lowest-energy residue selected by sp sampler is identical to the native amino acid.

Design of EGFR inhibitors

The design was performed using the EGF:d3 structure (PDB: 1IVO) as a template (residue range: 313-480 for d3-wt, and 313-472 for designed proteins). The input structure for Damietta applications has to be a single-chain structure that is CHARMM-typed, with all hydrogens included and no missing atoms. The input coordinates of the d3-wt domain structure was CHARMM-typed using the automatic PSF generation plugin (autopsf, version 1.8) as implemented in VMD (version 1.9.3).[50] Using the repack-all application (damietta_ra), we had identified residues with energy higher than 20 kcal/mol, as well as all cysteine residues. These residues were subject to combinatorial design using the few-to-many-to-few sampler (damietta_cs_f2m2f). The combinatorial sampler mutates and moves the designable residues, and moves the repackable residues as specified in the spec file (Methods S1). The average energy per residue is calculated for both the mutable and repackable residues and are deterministically minimized. The order by which residues are optimized is randomizable and multiple traversals of the mutagenesis decision tree can be conducted. The context of the spec file contained the following parameters: the mutable residues (mut_res) and their target mutations; the repackable residues (rpk_res); a scrambled order of the mutable residues in every instance (scramble_order); the top mutations considered for combinatorial optimization (m_mutations); the number of parallel paths traversed down the decision tree (n_paths); the number of repeat iterations by which the tree is traversed (n_iters). The target mutations in the spec file were specified according to a sequence profile of d3 homologues obtained from closest 500 homologous sequences in the nr protein database. The spec file was run 100 instances, for 100 CPU hours/instance. The resulting decoys were further filtered according to their stability in accelerated molecular dynamics (aMD) simulations that follow a serial tempering routine previously described.[55,56] These simulations were conducted using NAMD[49] with a generalized Born implicit solvent model and a timestep of 1 fs. The tempering scheme starts by 500 steps of conjugate gradient minimization followed by an annealing cycle that was repeated for 160 rounds, with one configuration dump at the end of each cycle. The annealing cycle follows the sequence of: 100 minimization steps, 3000 timesteps (i.e. 3 ps) in a 370 K Langevin bath, 4000 timesteps (i.e. 4 ps) in a 250 K Langevin bath, and 100 minimization steps. The two most conformationally homogeneous designs (dd3-1 and dd3-2) were accordingly selected for experimental characterization. Conformational homogeneity was quantified as the average all-*vs.*-all RMSD averaged across all frames (i.e. 160 frames) output from the aMD simulations using VMD.[50]

Purification of EGFR inhibitors

Sequences of d3-wt and the designed proteins (dd3-1, dd3-2, Table S2) were ordered as synthetic genes in pET-28a(+) expression vector (Synbio Technologies, Inc.). Plasmids were transformed into chemically competent *E. coli* BL21(DE3) using the heat shock method. Transformed cells were grown in LB medium supplemented with 40 μg/ml kanamycin at 37 °C. At OD600 of around

0.6-0.8, cells were induced with 1mM IPTG and incubated overnight at 25°C for protein expression. Cells were harvested by centrifugation at 5000 g at 4°C for 15 min and lysed in 30 ml of lysis buffer (1M guanidinium chloride, 100 mM NaCl, 50mM Tris-HCl pH 8.0) supplemented with a tablet of the cOmplete, EDTA-free Protease Inhibitor Cocktail (Roche) and 3 mg of lyophilized DNase I (PanReac AppliChem) using a Branson Sonifier 250 (Fisher Scientific). The lysate was cleared by centrifugation at 28000 g at 4°C for 50 min and the supernatant was filtered through a 0.45 μm filter (Millipore). The sample was applied to a 5 ml HisTrap Excel column (Cytiva). The running buffer was 200 mM NaCl, 30 mM Tris-HCl pH 8.0. After sequential washing the column with 20 ml of the running buffer and 20 ml of the running buffer supplemented with 50 mM imidazole, fractions were collected by linear gradient elution using 150 mM NaCl, 30 mM Tris-HCl pH 8.0, 500 mM imidazole buffer. The eluted fractions containing the protein of interest were pooled, concentrated using 10 kDa MWCO centrifugal filters (Millipore), and further purified on a Superdex 75 Increase 10/300 GL gel filtration column (Cytiva) using PBS. Gel filtration fractions containing pure protein in the desired oligomeric state were pooled, concentrated, and stored at -20 °C for subsequent analyses. Both IMAC and gel filtration steps were performed on an Äkta Pure chromatography system (Cytiva).

Thermostability analysis of EGFR inhibitors

NanoDSF measurements using Prometheus NT.48 (Nanotemper) were performed to evaluate thermostability of d3-wt and the designs (dd3-1, dd3-2). Capillaries (Nanotemper) were filled with 0.5 mg/ml protein samples in three replicates. Melting scan was performed across the temperature range from 20 °C to 90 °C with a temperature ramp of 1 °C/min.

Surface plasmon resonance binding assays

Multi-cycle kinetics experiments were performed on a Biacore X100 system (Cytiva). For measuring binding to EGF, EGF (Peprotech) was diluted to 100 μg/mL in 10 mM acetate buffer pH 4.0 and immobilized on the surface of a CM5 sensor chip (Cytiva) using standard amine coupling chemistry. Five different concentrations of the sample solution (nanomolar range) were injected over the functionalized sensor chip surface for 120 s, followed by a 180 s dissociation with the running buffer. At the end of each run, the sensor surface was regenerated with a 30 s injection of 50 mM HCl at a flow rate of 10 μL/min. For measuring binding to HB-EGF, HB-EGF (R&D Systems) was diluted to 20 μg/mL in 10 mM acetate buffer pH 5.0 and immobilized on the surface of a CM5 sensor chip using standard amine coupling chemistry. Five different concentrations of the sample solution (nanomolar range) were injected over the functionalized sensor chip surface for 180 s, followed by a 180 s dissociation with the running buffer. At the end of each run, the sensor surface was regenerated with a 30 s injection of 50 mM NaOH at a flow rate of 10 μL/min. For measuring binding to TGF-α, TGF-α (R&D Systems) was diluted to 100 μg/mL in 10 mM acetate buffer pH 4.5 and immobilized on the surface of a CM5 sensor chip using standard amine coupling chemistry. Five different concentrations of the sample solution (nanomolar – low micromolar range) were injected over the functionalized sensor chip surface for 60 s, followed by a 60 s dissociation with the running buffer. At the end of each run, the sensor surface was regenerated with a 60 s injection of 10 mM glycine-HCl pH 1.5 at a flow rate of 10 μL/min. In all experiments, reference surfaces were treated in the same manner (surface activation and deactivation with amine coupling reagents), except that no ligand was added. Test proteins were diluted in the running buffer (PBS supplemented with 0.05% v/v Tween-20). Analyses were conducted at 25°C at a flow rate of 10 μL/min. The reference responses and zero-concentration sensograms were subtracted from each dataset (double-referencing). Association rate (k_a), dissociation rate (k_d), and equilibrium dissociation (K_d) constants were obtained using the Biacore X100 Evaluation Software. Fitting was performed with global parameter settings where single parameter value applies to the whole titration series. To estimate the reliability of the fit, the fitting procedure was repeated using 4 out of 5 analyte concentrations (excluding one concentration at a time), yielding average values and standard deviations for a single titration series. For binding assays of d3-wt and dd3-2 vs EGF, HB-EGF, and TGF-alpha, two independent titration series were performed.

A431 cell proliferation assay

A431 cells were cultured in DMEM medium (Gibco) supplemented with 10 % FBS (Gibco). Cells were pelleted by centrifugation at 300 g for 5 min, washed once with DPBS (Gibco) and once with non-supplemented DMEM medium. After the last washing step, cells were resuspended in DMEM medium supplemented with 1 % FBS and 400 pM EGF (Peprotech). Cell suspension was seeded in a 96-well plate (Corning), 100 μl/well, at a density of 8000 cells/well. Different concentrations of d3-wt (0.032 nM – 500 nM), dd3-2 (0.0064 nM – 100 nM), or cetuximab (0.0064 nM – 100 nM) were added to the wells in triplicates. PBS was added to the wells serving as an untreated control. Several wells contained cells in DMEM medium supplemented with 1 % FBS, but without EGF, as an unstimulated control. After incubation for 72 h at 37°C, 5% CO_2, 20 μL of CellTiter-Blue® Reagent (Promega) were added to the wells and the plate was incubated for additional 2 h under the same conditions to allow cells to convert resazurin to resorufin. Cell viability was monitored by measuring fluorescence (560/590 nm) using a Synergy HTX Microplate Reader (BioTek). The data were presented as percentage of unstimulated (i.e., without EGF) control fluorescence values.

Effect of EGFR inhibitors on zebrafish embryos

Eggs were collected and placed at 28 °C in E3 medium (5 mM NaCl, 0.17 mM KCl, 0.4 mM $CaCl_2$, and 0.16 mM $MgSO_4$). The age of the embryos and larvae is indicated as hours postfertilization (hpf) or days post fertilization (dpf). All experiments described in the present study were conducted on embryos younger than 5 dpf. To test the effect of the designed inhibitors on zebrafish, we injected

around 4 nl of Cetuximab (5.0 mg/ml), dd3-2 (0.2 and 1.0 mg/ml), d3-wt (0.3 and 1.3 mg/ml) into the yolk of embryos at 4-6 hpf and then continued the treatment by adding the inhibitors into medium until 4 dpf. Embryos were distributed in pools of 10-15 into 24-well plates in E3 medium. Survival ratio was assessed every day from 1 dpf to 4 dpf and morphological or developmental defects were analyzed on fixed and stained embryos with Alcian blue at 4 dpf using a Nikon SMZ18 stereomicroscope with a DS-Fi3 camera (5,9 MP). GraphPad Prism software (version 7) was used for graphing and statistical analysis. Cartilage was stained with Alcian Blue 8 GX (Sigma). Zebrafish larvae were fixed in PFA 4 % for 2 h at room temperature, rinsed with PBST and stained overnight with 10 mM $MgCl_2$/80 % ethanol/0.04 % Alcian Blue solution. Embryos were rinsed with 80 % ethanol/10 mM $MgCl_2$ and washed stepwise with 70 %, 50 %, 30 % ethanol and PBST. Pigments were bleached in H_2O_2 3 %/formamide 5 %/20X SSC 2,5 % up to 30 minutes. Embryos were stored in 80 % glycerol for imaging.

Design of copper-binding proteins

Combinatorial design simulations were run to redesign the template structure of apo-protein form (PDB: 5FJD), given the limitations of the described method and the used classical mechanics force field to adequately describe coordinated metal ions. Most surface positions were set as designable to break the oligomerization interfaces and stabilize the helical structures. The spec file (Methods S1) was run 100 instances, for 125 CPU hours/instance. The final sequences were chosen according to the same aMD filtering described above. Through a second round of computational design, we sought to create constructs with a more sealed core. This was done by mutating 3 or 6 cysteine residues (and their surrounding positions) into well-packed hydrophobic residues at the end of the cysteine-lined lumen of the plr1 model. Three such design candidates were also synthesized and tested: cr3, cr61, and cr62. Additionally, the surface positions of plr1 were also redesigned to bias them towards negative supercharged variants, where two candidates were selected and tested: neg1 and neg2.

Purification of copper-binding proteins

The synthetic genes for all tested designs, and the design template (Table S2) were cloned without purification tags in a pET28b(+) vector (BioCat GmbH). The proteins were transformed and expressed in *E. coli* BL21 (DE3). Expression was induced in 2-litre LB medium supplemented with 40 μg/ml kanamycin at an optical density (OD600) of 0.8, and was done overnight at 25 °C. Cells were harvested by centrifugation at 5000 g at 4 °C for 15 min and lysed in 30 ml of lysis buffer (for positively charged variants: 100 mM NaCl, 50 mM Tris-HCl pH 8.0; for negatively charged variants: 2 mM EDTA, 20 mM Tris-HCl pH 8.0) supplemented with a tablet of the cOmplete, EDTA-free Protease Inhibitor Cocktail (Roche) and 3 mg of lyophilized DNase I (PanReac AppliChem) using a Branson Sonifier 250 (Fisher Scientific). The lysate was cleared by centrifugation at 28000 g at 4°C for 50 min and the supernatant was filtered through a 0.45 μm filter (Millipore). The sample was diluted 5-fold and applied to a 5 ml HiTrap Capto Q or S columns depending on their isoelectric point (Cytiva). Positively charged proteins were eluted in 20 mM HEPES, 1 mM DTT buffer pH 7.4, using a gradient of 0 to 1.5 M KCl. Negatively charged variants where eluted in 20 mM HEPES, 1 mM DTT buffer pH 8.0, using a gradient of 0 to 1.5 M NaCl. The relevant fractions were identified by SDS-PAGE analysis, and further purified on a Superdex 75 Increase 10/300 GL gel filtration column (Cytiva) using 20 mM HEPES, 150 mM NaCl buffer pH 7.4. Gel filtration fractions containing pure protein in the desired oligomeric state were pooled, concentrated, and stored at -20°C for subsequent analyses.

Thermostability analysis of copper binders

NanoDSF measurements using Prometheus NT.48 (Nanotemper) were performed to evaluate thermostability of selected designs, as well as thermostability of Csp1. Capillaries (Nanotemper) were filled with 0.1-1 mg/ml protein samples in 3 replicates. Melting scan was performed across the temperature range from 20 °C to 110 °C with a temperature ramp of 1 °C/min. In addition to measuring the intrinsic fluorescence intensity ratio (350/330 nm), light intensity loss due to scattering (backreflection) was measured to detect protein aggregation.

Cu^{2+}-binding affinity, capacity, and stability

To evaluate how many Cu^{2+} ions can be bound within the core of plr1, absorption spectra (from 400 nM to 700 nM) were recorded for samples containing 20 μM of Zincon (Supelco), 20 μM of $CuSO_4$ and varying concentrations of plr1 (to provide ratio Cu^{2+}/plr1 from 1 to 19). In case, when plr1 is saturated with Cu^{2+} ions, complex between Cu^{2+} and Zincon forms and characteristic Cu^{2+}ZI peak at 599 nM can be observed on the spectrum.

The binding affinities of the designs to Cu^{2+} were determined using a modified Zincon assay described by Kocyla et al.[57] Zincon competition tests with plr1 and neg1 were performed in 20 mM HEPES buffer, pH 7.4, containing 150 mM NaCl. 50 μM of Zincon was mixed with 20 μM of $CuSO_4$ and different concentrations of a protein solution (0 – 160 μM for plr1, 0 – 80 μM for neg1). Samples were incubated for 8 h at 25 °C. The exact concentrations of Cu^{2+}ZI complex present in each sample were calculated based on the absorbances at 599 nM using the molar absorption coefficient of Cu^{2+}ZI at pH 7.4, 26100 $M^{-1}cm^{-1}$. Absorbances of the samples were measured on a Synergy HTX Microplate Reader (BioTek) in a 384-well plate (Greiner). The dissociation constant of the designed proteins (K_d^{Des}) was calculated as follows:

$$K_d^{Des} = \frac{K_d^{Cu^{2+}ZI}}{K_{ex}} \quad \text{(Equation 8)}$$

where $K_d^{Cu^{2+}ZI}$ is a dissociation constant of Cu^{2+}ZI at pH 7.4, 4.68×10^{-17} M, and K_{ex} is a constant describing the reaction of Cu^{2+} ion transfer from Cu^{2+}ZI to the tested copper-binding design. K_{ex} was determined by fitting the experimental data to the following equation:

$$K_{ex} = \frac{[ZI]\left[Cu^{2+}Des\right]}{\left[Cu^{2+}ZI\right][Des]} \quad \text{(Equation 9)}$$

To measure off rates for Cu^{2+} dissociating from plr1 or neg1, we performed the dissociation experiments in 20 mM HEPES buffer, pH 7.4, containing 150 mM NaCl and 25 % v/v of untreated fetal bovine serum. Samples containing 50 μM of Zincon, 50 μM of $CuSO_4$ and 150 μM of either plr1 or neg1 were incubated in 384-well plate (Greiner) for 48 h at 37 °C. Every hour the absorbances at 599 nM were recorded. Samples containing no Cu^{2+} and no protein were used as a control, and average absorbance of these samples was subtracted from all tested values. Results are presented as a decrease in amount of Cu^{2+} bound (Cu^{bound}) to the protein over time (t). Absorbance values from the wells containing Zincon and Cu^{2+}, but no protein were used for normalization and were referred to as 100 % of unbound Cu^{2+}. Dissociation rates (K_{off}) were determined by fitting the experimental data to the equation:

$$Cu^{bound} = Cu_0^{bound} \times e^{-K_{off}\times t} \quad \text{(Equation 10)}$$

where Cu_0^{bound} is a percent of Cu^{2+} bound to the protein at time zero.

Loading of copper binders with ^{64}Cu

^{64}Ni (98 % enrichment) was electroplated on a Pt/Ir plate (90/10) and irradiated with 12.5 MeV protons on the Tübingen PETtrace cyclotron (GE Healthcare) to produce ^{64}Cu via the ^{64}Ni(p,n)^{64}Cu route. The target was dissolved using concentrated HCl and $^{64}Cu^{2+}$ was purified using ion chromatography as described before.[58] The obtained radioisotope solution in 0.1 M HCl was buffered with 1.5 volumes of 0.5 M ammonium acetate pH 4.1 before addition of the protein (2 μg per MBq). After 30 min of incubation at 35 °C incorporation of the radioactivity was analyzed by thin layer chromatography (stationary phase: Polygram SIL G UV254, Macherey-Nagel; mobile phase: 0.1 M sodium citrate pH 5) with autoradiographic detection using a phosphor imager (Cyclone Plus Storage Phosphor System, Perkin Elmer). Size exclusion chromatography with radioactivity detector (1260 Infinity II, Agilent; Superdex 75 Increase 10/300, Cytiva) was used to analyze the elution profile of the protein and bound radioactivity. For stripping of the bound radioactivity DTPA (final concentration 0.14 mg/ml) was added to the labeled protein and re-analyzed after incubation at room temperature.

QUANTIFICATION AND STATISTICAL ANALYSIS

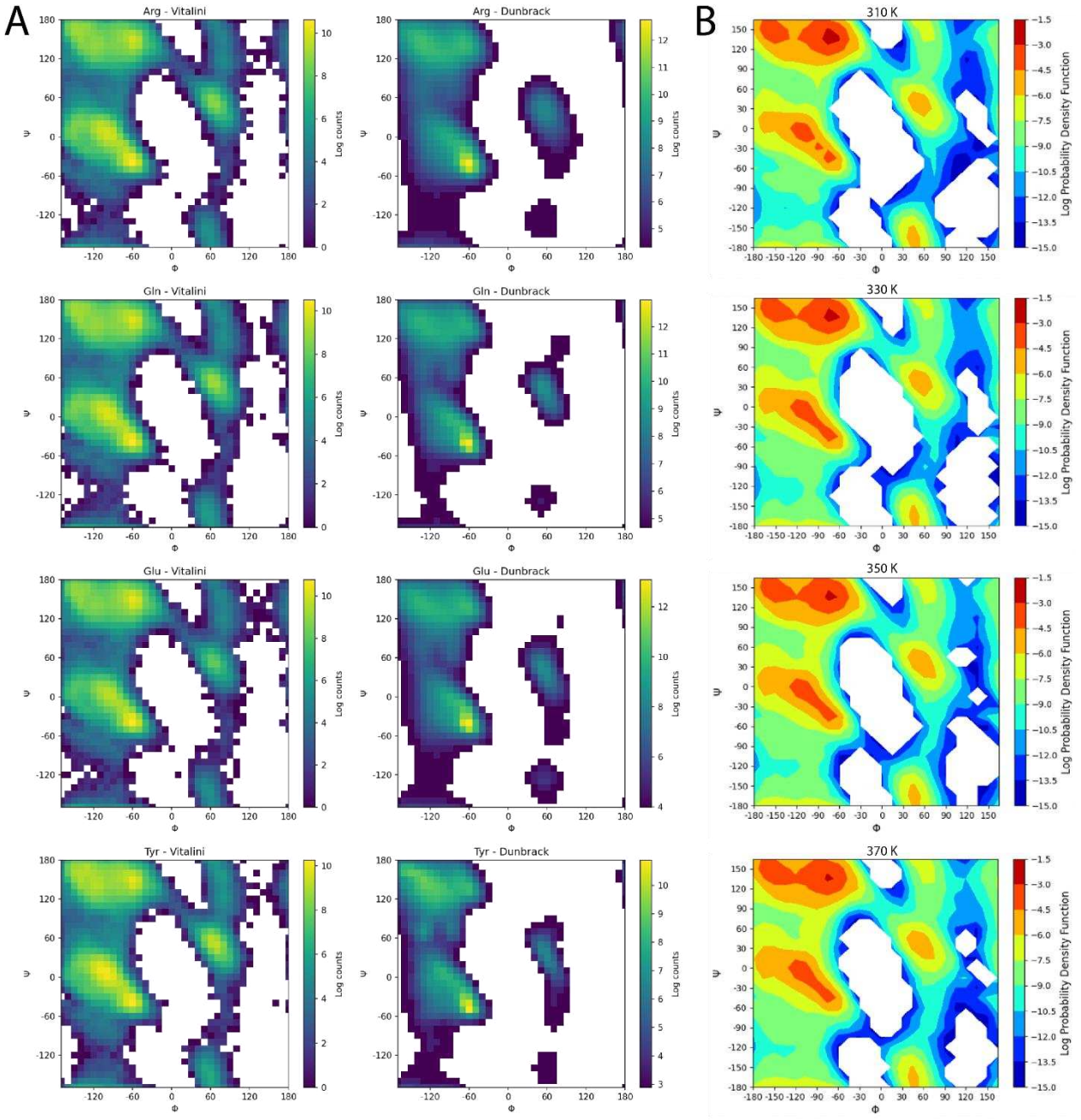
To evaluate scoring accuracy of the Damietta potential against different stability benchmarks the Pearson correlation coefficients and the correlation p-values were calculated using the scipy.stats sub-package in the SciPy (version 1.8.0). For SPR measurements, A431 cell proliferation, as well as copper-binding experiments, mean values and standard deviations were calculated and described in the Figure legends. Statistical analysis for zebrafish experiment was performed using the GraphPad Prism software (version 7).

Cell Reports Methods, Volume *3*

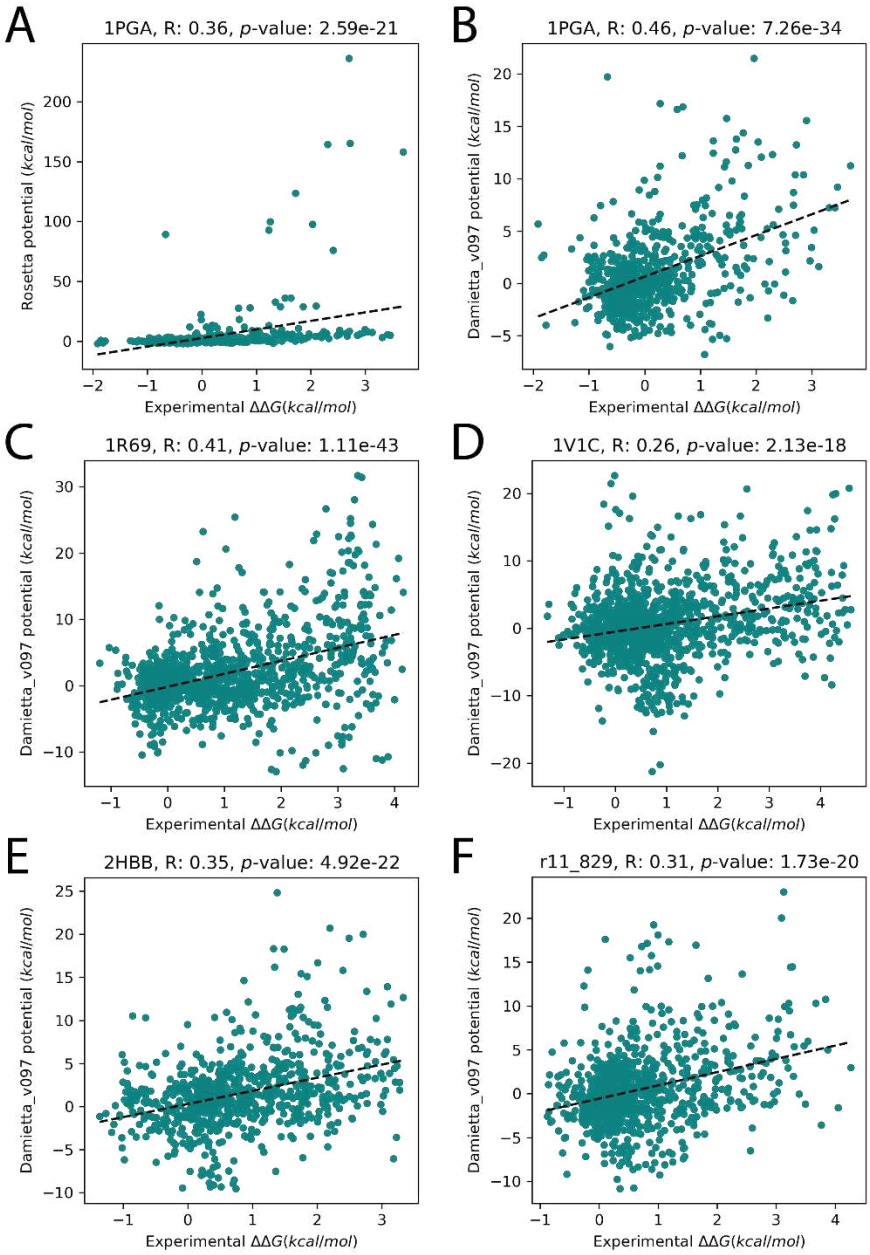
Supplemental information

The design of functional proteins using tensorized energy calculations

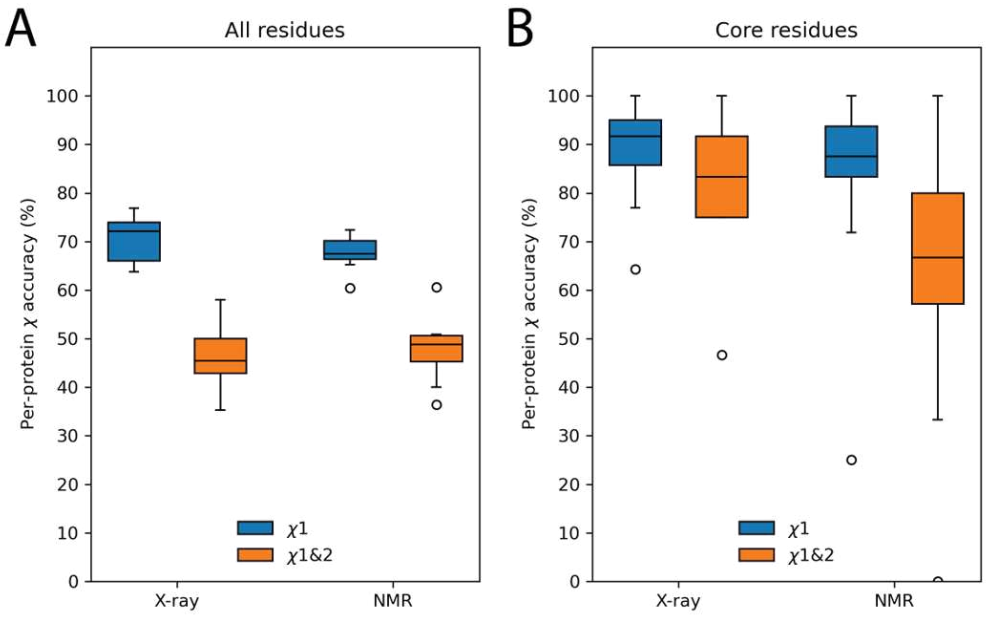
Kateryna Maksymenko, Andreas Maurer, Narges Aghaallaei, Caroline Barry, Natalia Borbarán-Bravo, Timo Ullrich, Tjeerd M.H. Dijkstra, Birte Hernandez Alvarez, Patrick Müller, Andrei N. Lupas, Julia Skokowa, and Mohammad ElGamacy



Supplemental figure 1: MD-derived conformational distributions cover broad conformational space, related to STAR Methods. (A) The figure shows conformational distributions for four amino acid examples with diverse physical properties (Arg, Gln, Glu, and Tyr; top to bottom). Left column shows distributions used to drive the described rotamer library (originally described by Vitalini et al. [S1]). Right column shows PDB-derived distributions described by Shapovalov and Dunbrack [S2]. **(B)** MD performed at elevated temperatures can be used to obtain broadened conformational distributions to better cover rare conformations. Long MD simulations (1 μs) of capped GAG tripeptide yield broader coverage of the alanine (φ, ψ)-space at successively higher temperatures of 310 K, 330 K, 350 K, 370 K, without perturbation of the overall distribution.



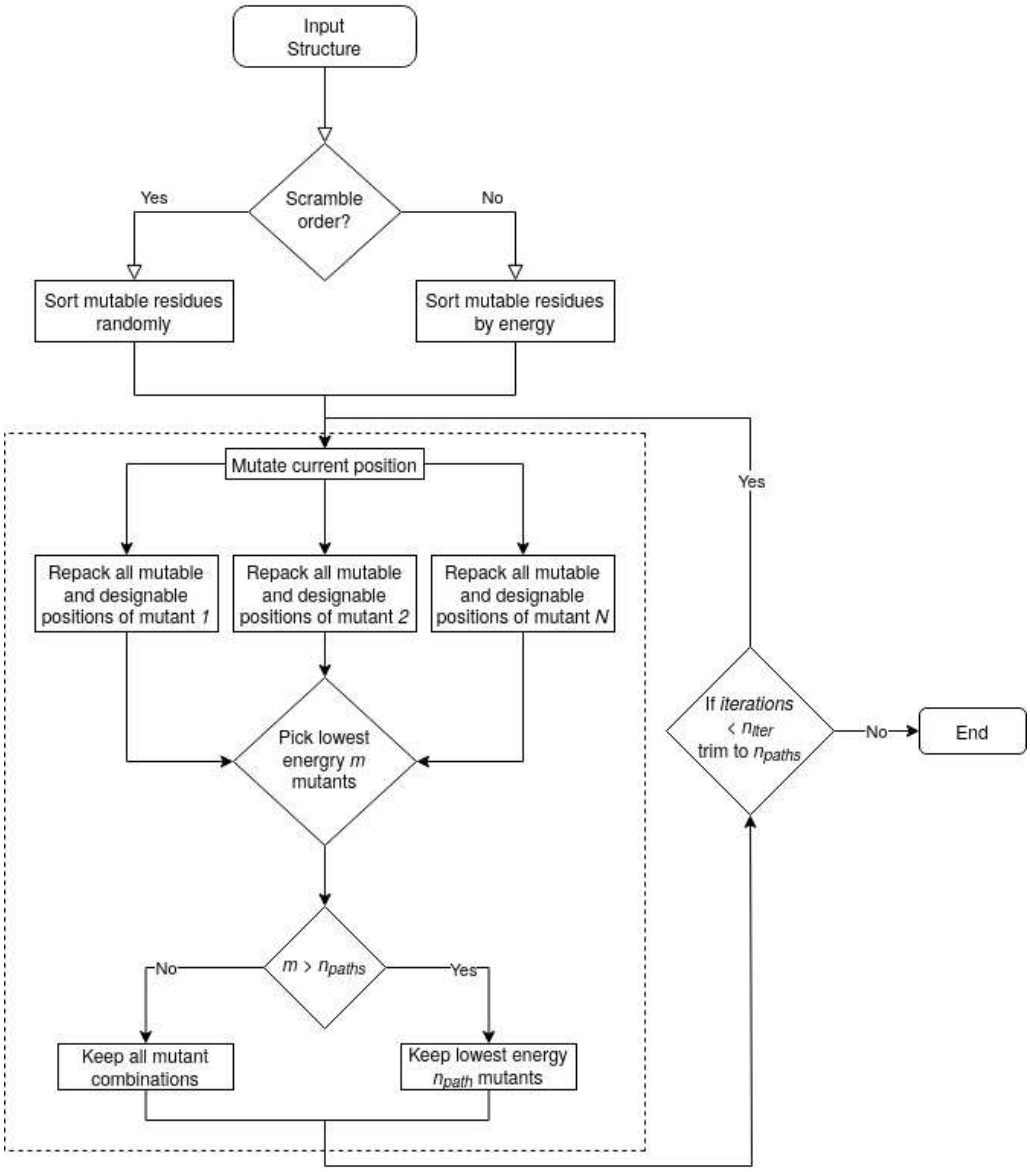
Supplemental figure 2: Retrospective validation of the Damietta potential against five different benchmarks, related to Fig 1. A comparison between the Rosetta **(A)** and Damietta **(B)** energies correlation with the change in folding free energy (ΔΔG) values for mutants of Gβ1 protein. Evaluation of the Damietta energy function performance against a benchmark of mutants of the N-terminal domain of phage 434 repressor (PDB 1R69) **(C)**, SH3 domain in human obscurin (PDB 1V1C) **(D)**, N-terminal domain of ribosomal protein L9 (PDB 2HBB) **(E)**, and the hallucination design r11_829_TrROS **(F)**. The Pearson correlation coefficients (R) and the correlation p-values are shown.



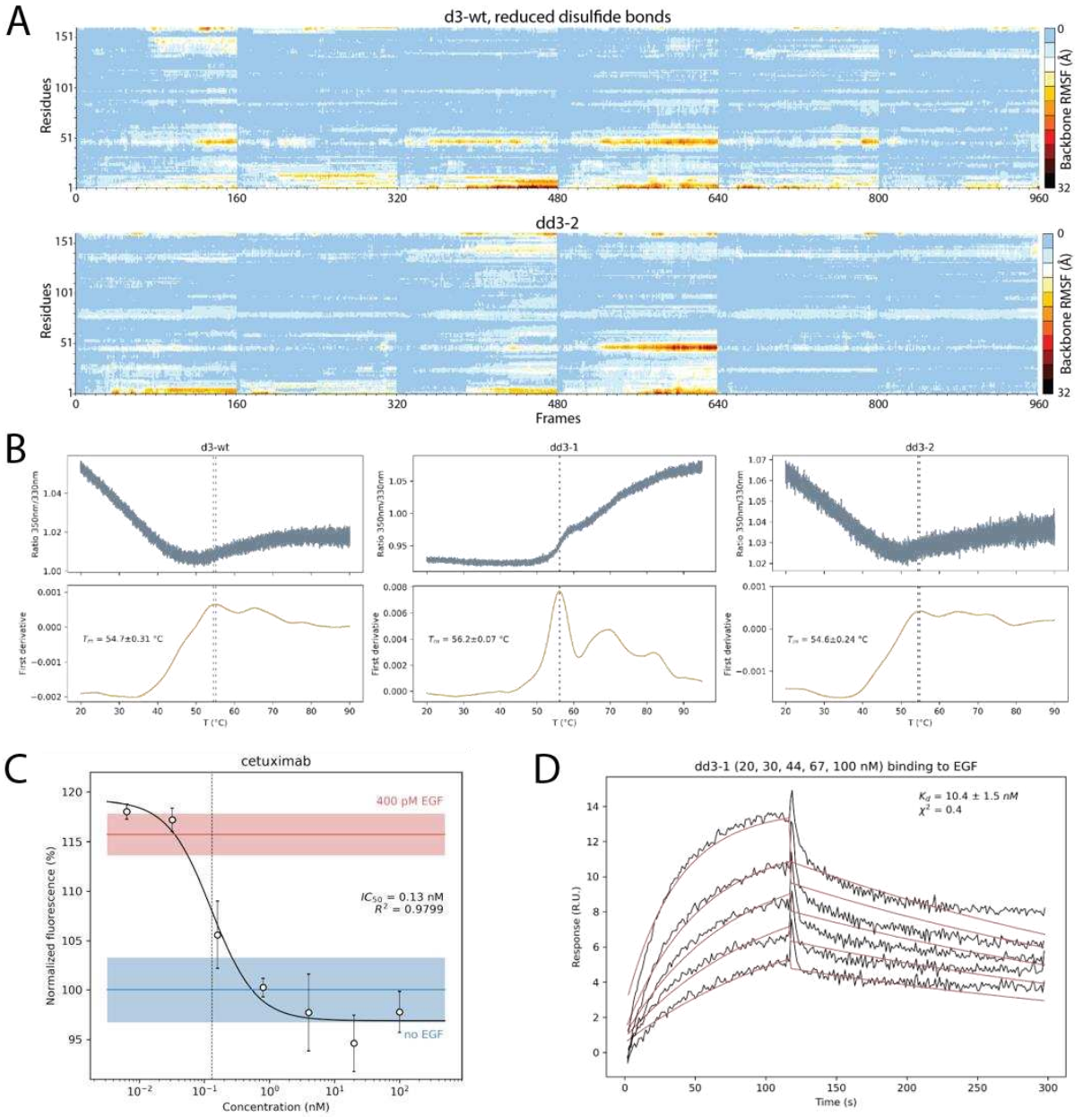
C

PDB	1TTZ	1TVG	2O0Q	2Q00	3C4S	3FIF	3H9X	3IDU	3K63
Sequence recovery (%)	12.33	11.94	14.41	10.92	12.73	18.75	11.30	11.43	9.52

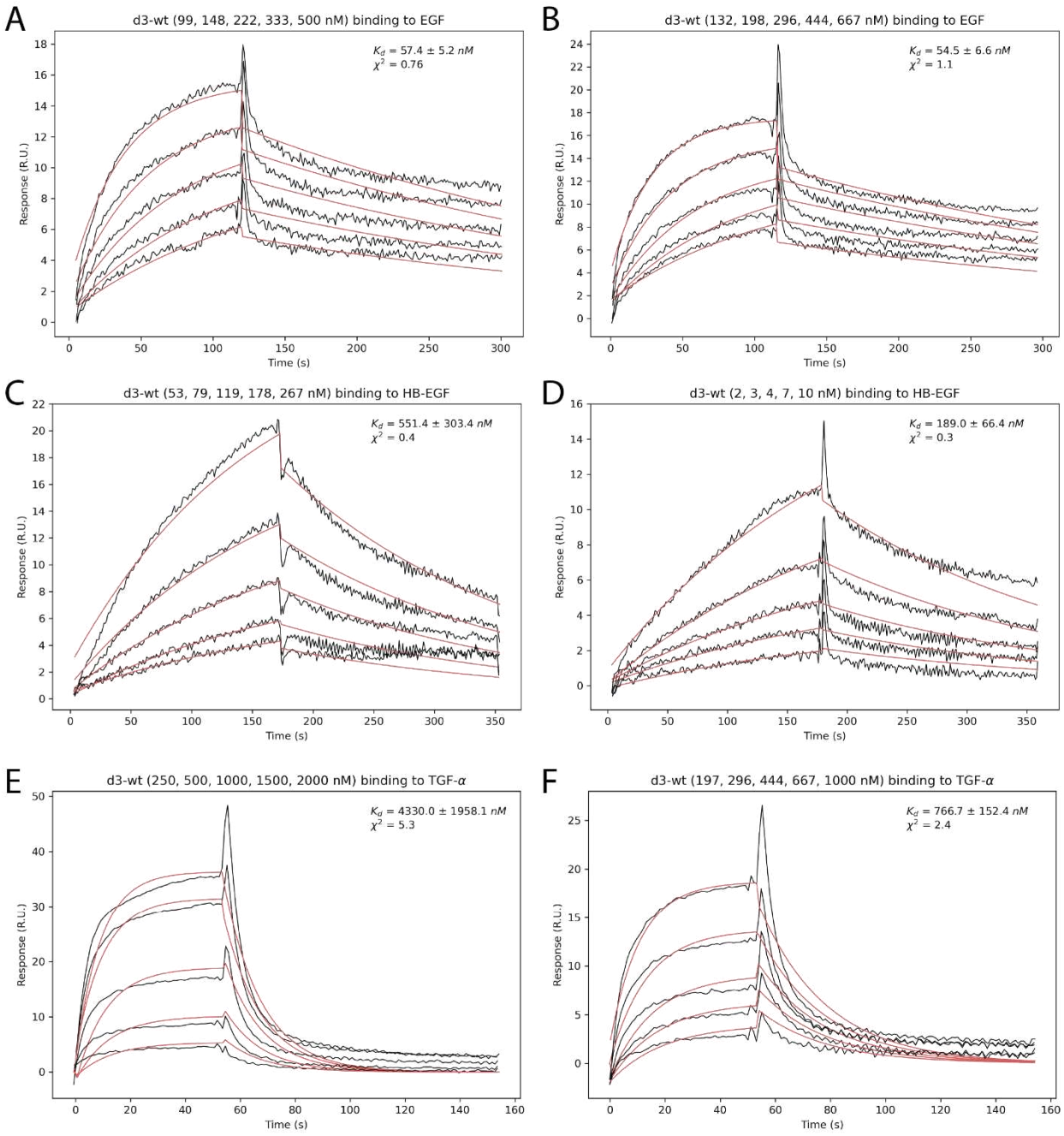
Supplemental figure 3: Results of native side-chain conformation recovery and single-position sequence recovery, related to Fig. 1. (A) Boxplot shows the median χ_1 prediction accuracy to be around 70 % for both X-ray and NMR structures. $\chi_{1\&2}$ accuracies are roughly 20 % lower than χ_1 accuracies. **(B)** Rotamer recovery for core positions was found to be high with the median χ_1 accuracy of around 90 % and the median $\chi_{1\&2}$ accuracy of around 65-85 %. **(C)** Sequence recovery rates for 9 different crystal structures are presented as a percentage of amino acid positions within a protein at which the lowest-energy residue selected by `sp` sampler is identical to the native amino acid.



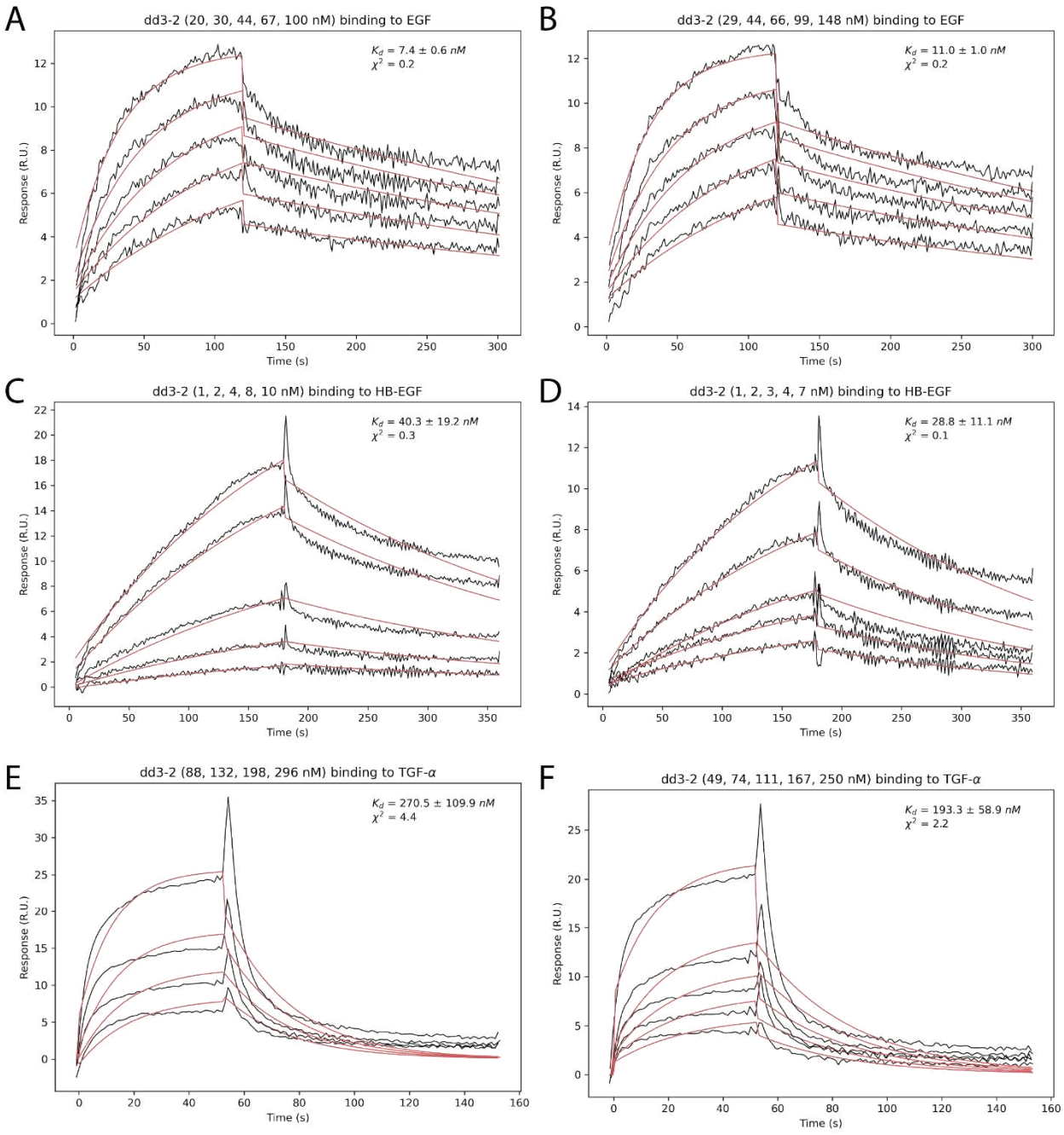
Supplemental figure 4: A flow chart of the greedy branch-and-rank combinatorial sampling algorithm used for design, related to Fig. 1. Scheme of the few-to-many-to-few combinatorial sampler. The algorithm assumes all mutable and repackable residues are inter-dependent and follows a maximum number of paths n_{paths} down the decision tree. At the start of the design simulation all of the combinatorially generated sequences are further designs so long as their number is $\leq n_{paths}$. But as the sequence combinations grow $> n_{paths}$, the algorithm ranks all of the designs by the average energy per residue (evaluated at the mutable and repackable positions). Only a small number ($= n_{paths}$) of lowest energy mutants is kept for the next position mutagenesis round. The order of mutable positions can be preset or randomized as the defined by the user. The main flow of sampling therefore follows: i) mutate position, ii) calculate average energy of every mutant at that position, iii) combine the top m mutations with all of the previously kept mutants from the previous cycle to evaluate the average energy per residue, iv) if the generated combinations are larger in number than n_{paths}, only keep the lowest energy n_{paths} sequences, v) move to the next mutable position and repeat step, vi) repeat these iterations over all mutable residues. This cycle is repeated n_{iters} times for better convergence.



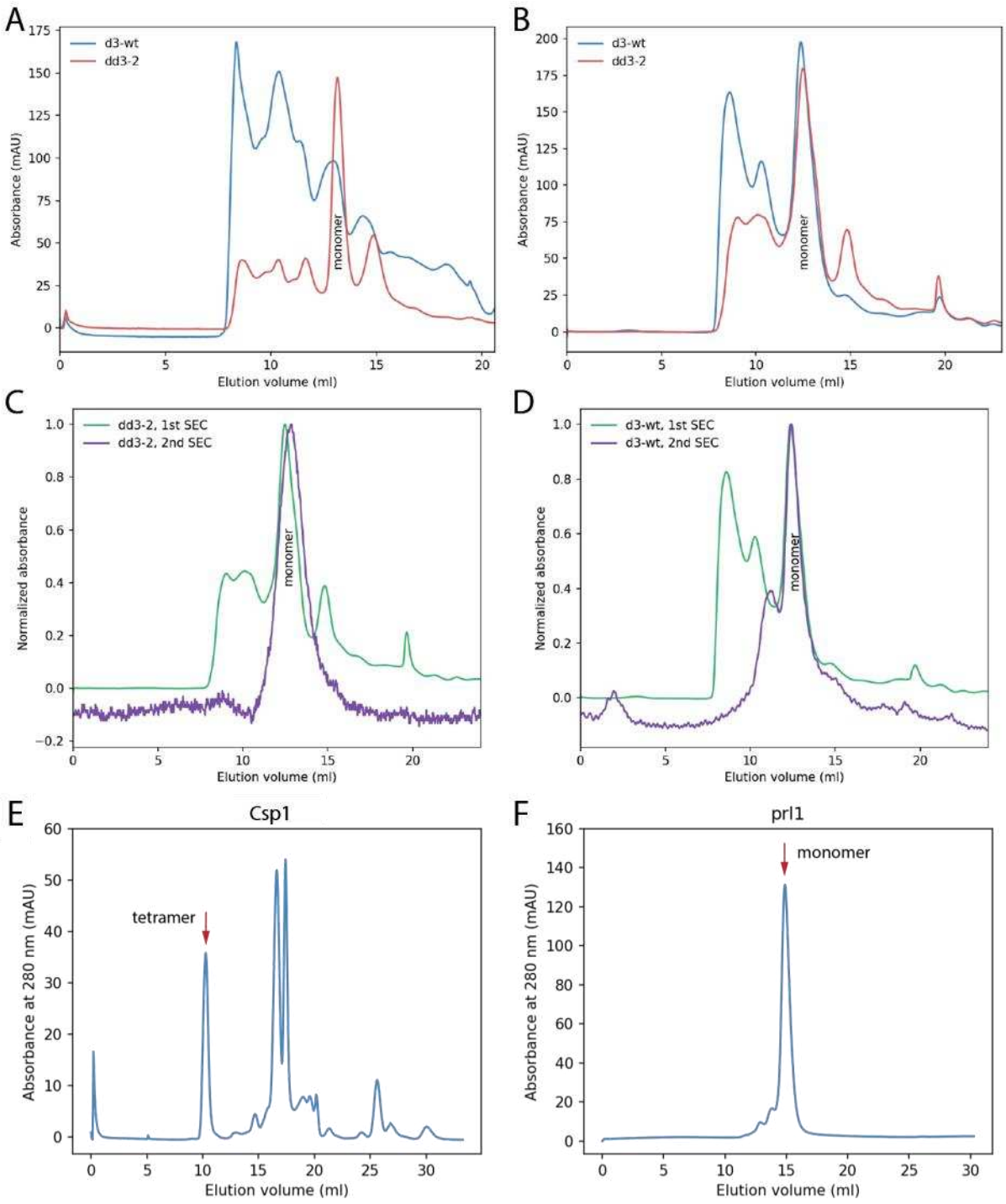
Supplemental figure 5: Computational and experimental characterization of EGFR inhibitors, related to Fig. 2. (A) RMSF heatmaps indicate improved conformational stability of dd3-2 (bottom) compared to the reduced cysteines form of d3-wt (top). A single data point in the heatmap represents RMSF for each residue's backbone atoms and its reference residue's backbones atoms (frame 0) through tempering MD trajectories (160 frames, 6 replicas). **(B)** NanoDSF measurements showed no significant difference in melting temperatures between designed proteins (dd3-1, dd3-2) and the wild type (d3-wt). Melting temperatures (T_m) are represented as mean ± SD. **(C)** EGFR-blocking antibody Cetuximab inhibited proliferation of A431 cells with IC_{50} of 0.13 nM, which is only 3-fold lower compared to IC_{50} of dd3-2 (Fig. 2C). The positive and negative control values of cell proliferation with and without EGF-treatment are indicated by red and blue lines, respectively. Shades and error bars represent the standard deviation across three replicates. **(D)** SPR sensograms showed dd3-1 design to bind EGF with K_d value of 10 nM, which is approximately 6-fold tighter compared to d3-wt (Fig. 2B). K_d is represented as mean ± standard deviation (SD). Experimental data, black; fit, red.



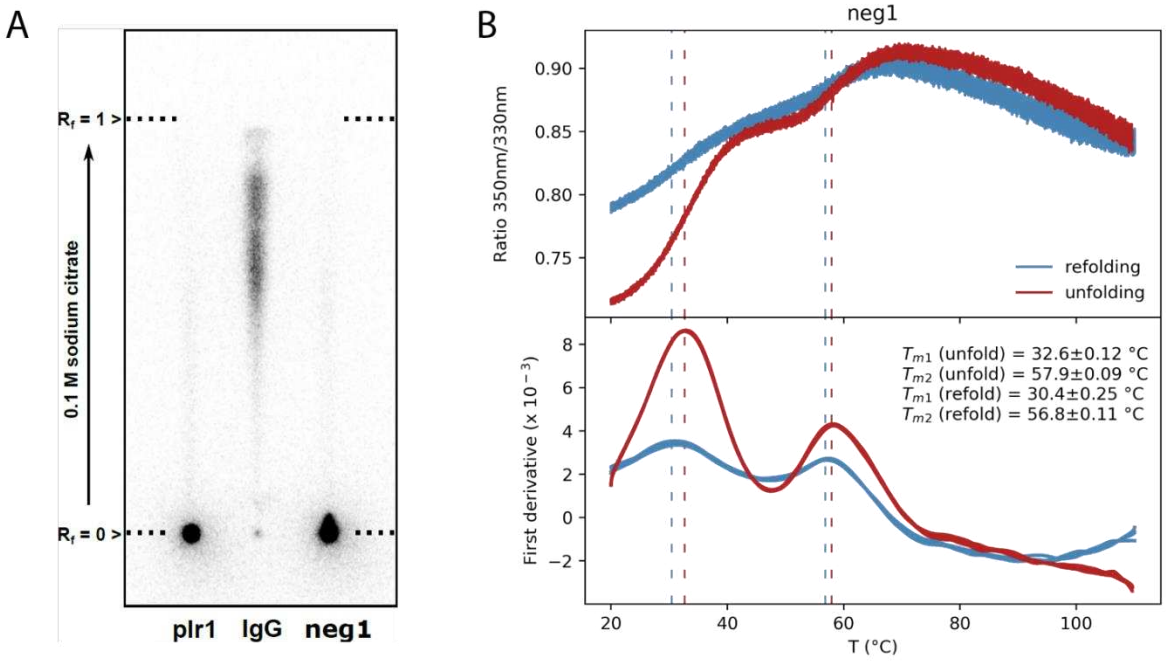
Supplemental figure 6: SPR titrations of d3-wt binding to immobilized EGFR ligands, related to Table 1. Sensograms of d3-wt binding to **(A, B)** EGF, **(C, D)** HB-EGF, and **(E, F)** TGF-alpha. Results were obtained from two independent experiments. K_d is represented as mean ± standard deviation (SD). Experimental data, black; fit, red.



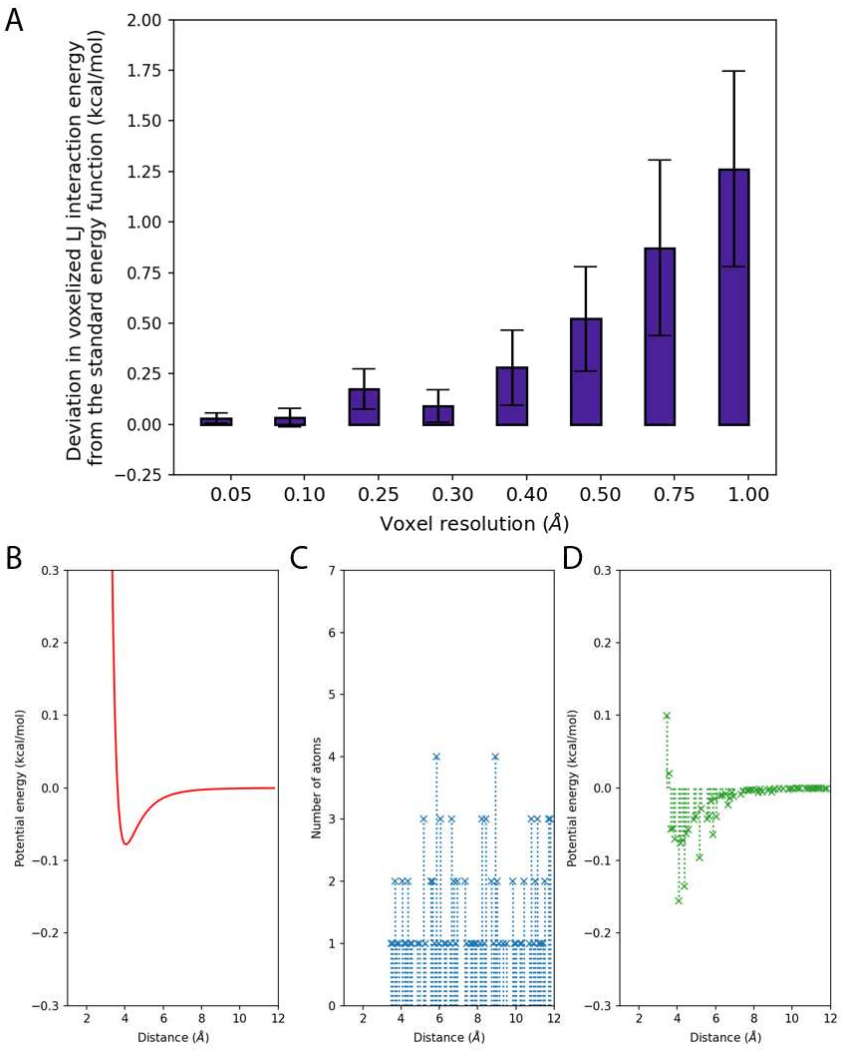
Supplemental figure 7: SPR titrations of dd3-2 binding to immobilized EGFR ligands, related to Table 1. Sensograms of dd3-2 binding to **(A, B)** EGF, **(C, D)** HB-EGF, and **(E, F)** TGF-alpha. Results were obtained from two independent experiments. K_d is represented as mean ± standard deviation (SD). Experimental data, black; fit, red.



Supplemental figure 8: Analytical size-exclusion profiles of EGFR-inhibiting and copper-binding proteins, related to Fig. 2 and Fig. 4A. (A, B) The chromatograms for d3-wt and dd3-2 proteins, obtained from two independent rounds of expression and purification, show d3-wt to have higher aggregation propensity than dd3-2. For experimental evaluation, monomeric fractions of dd3-2 and d3-wt were collected. **(C)** Analytical size-exclusion elution profile of the collected dd3-2 monomeric fraction. **(D)** Analytical size-exclusion elution profile of the collected d3-wt monomeric fraction. **(E)** Csp1 protein elutes as a tetramer and forms aggregates in solution. **(F)** plr1 elutes as a homogenous monomer with no signs of aggregation.



Supplemental figure 9: Experimental characterization of copper-binding designs, related to Fig. 3. **(A)** Radio-TLC shows the specific binding of plr1 or neg1, but not IgG (negative control) to $^{64}Cu^{2+}$ ions. **(B)** NanoDSF measurement shows substantial destabilization of the negatively supercharged variant neg1 when compared to plr1 (Fig. 3B). Specifically, two melting transitions T_{m1} and T_{m2} occur at around 33 °C and 58 °C, respectively (red lines show differential fluorescence data points and first derivatives of three replicas of the heating ramp). These two unfolding events however appear to be reversible (blue lines show the behavior down the cooling ramp). Melting temperatures (T_m) are represented as mean ± SD.



Supplemental figure 10: Array-oriented calculation of interactions and the effect of voxel resolution on energy error, related to STAR Methods. (A) The Y-axis shows the residual energy values by subtracting energy values calculated by the standard LJ function from the discretized LJ function; i.e. $E_{LJ,discretized} - E_{LJ,smooth}$. The discretization was performed at voxel resolutions of 0.05, 0.10, 0.25, 0.30, 0.40, 0.50, 0.75 and 1.00 Å. In this demonstration, the interactions between one sp^3 methyl carbon ($\varepsilon_{LJ} = -0.078\ kcal/mol$, $\sigma_{LJ} = 3.63$ Å) and 100 inbound sp^3 methyl carbons, that are randomly pooled from a uniform distribution, bounded at the distance range of 3.5 to 12.0 Å. Error bars represent standard deviation values for 10 replicas, with randomized atomic positions. **(B-D)** The presented framework simplifies interection calculations to an expensive pre-computed field of a discrete rotamer (**B**; here shown as a 1D LJ potential of a single atom for simplicity) and a quickly populated histogram of the environment atomic positions (**C**; here depicted as 1D array for simplicity). A single multiplication process yields the "energies array" **(D)**.

Table S1: The time performance of the combinatorial sampler under different simulation parameters, related to Fig. 1. These design simulations were run on the ubiquitin crystal structure (PDB: 1UBQ) as template.

	Simulation 1	Simulation 2	Simulation 3	Simulation 4	Simulation 5
Mutable residues	10	20	30	40	40
Repackable residues	10				
Target mutations per position	20				
Conformers per repacking task	100				
Number of iterations (`n_iters`)	1				
Theoretical sequence complexity	$1{\cdot}10^{13}$	$1{\cdot}10^{26}$	$1.1{\cdot}10^{39}$	$1.1{\cdot}10^{52}$	$1.1{\cdot}10^{52}$
Theoretical rotameric complexity	$1{\cdot}10^{53}$	$1{\cdot}10^{86}$	$1.1{\cdot}10^{119}$	$1.1{\cdot}10^{152}$	$1.1{\cdot}10^{152}$
Intermediate mutants (`m_muts`)	3				
swarm paths (`n_paths`)	6				1
Actual rotamer evaluations	$1.4{\cdot}10^{7}$	$9.2{\cdot}10^{7}$	$3.3{\cdot}10^{8}$	$6.4{\cdot}10^{8}$	$1.9{\cdot}10^{8}$
Wall-clock time (s)	$1.3{\cdot}10^{3}$	$3.5{\cdot}10^{3}$	$7.3{\cdot}10^{3}$	$1.1{\cdot}10^{4}$	$3.7{\cdot}10^{3}$
CPU threads	8 (2.3 GHz)				
Time/residue* (s)	$9.3{\cdot}10^{-3}$	$3.8{\cdot}10^{-3}$	$2.2{\cdot}10^{-3}$	$1.7{\cdot}10^{-3}$	$1.9{\cdot}10^{-3}$
Time/rotamer (s)	$9.3{\cdot}10^{-5}$	$3.8{\cdot}10^{-5}$	$2.2{\cdot}10^{-5}$	$1.7{\cdot}10^{-5}$	$1.9{\cdot}10^{-5}$
Time/atom-step**	$8.5{\cdot}10^{-8}$	$3.5{\cdot}10^{-8}$	$2.0{\cdot}10^{-8}$	$1.6{\cdot}10^{-8}$	$1.7{\cdot}10^{-8}$
Lowest-energy decoy (kcal/mol)***	-11.0	-11.9	-12.3	-12.2	-11.9

* Evaluating 100 rotamers per residue ranged between 0.002 ~ 0.009 s. This is in comparison with ~ 0.2 s for RaSP, 100 ~ 200 s for Rosetta, 200 ~ 400 s for FoldX, 0.7 ~ 0.8 s for ACDC-NN, or 4 ~ 6 s for ThermoNet [S3].

** Assuming 9.7 $Å^3$/atom as an average atomic volume occupancy per simulation cube (1098 atoms / 10,648 $Å^3$). The tabulated results describe the overall performance across the entire control flow of the `cs_f2m2f` application. For instance, the isolated speed of the LJ energy function, on a single CPU core gives 130 ns/atom-step (including environment mapping and rotamer·environment tensor multiplication), which is substantially faster than recently described approaches [S4].

*** As the average energy per sampled residues (i.e. all repackable or mutable residues).

Table S2: Protein sequences of template and designed EGFR inhibitors and copper-binding proteins, related to STAR Methods. Mutated residues are highlighted in yellow.

Name	Sequence
d3-wt	CNGIGIGEFKDSLSINATNIKHFKNCTSISGDLHILPVAFRGDSFTHTPPLDPQELDILKTVKEITGFLLI QAWPENRTDLHAFENLEIIRGRTKQHGQFSLAVVSLNITSLGLRSLKEISDGDVIISGNKNLCYANTINWK KLFGTSGQKTKIISNRGENSCKATGQ
dd3-1	ANGIGIGEFKDSLSIWAWNIKHFKNARSISGDLHILPVAFRGDSFTHTPPLEPKELEILKTVKEITGYLLV QAWPENRTDLHAFENLEIIRGRTKQHGQFSLAVVSLNVTSLGLRSLKEISDGDVIMSGNKNLKMANEMNWK KMFGTSGQKTKIISNRGE--------
dd3-2	ANGIGIGEFKDSLSINAWNIKHFKNAQSISGDLHILPVAFRGDSFTHMPPLEPKELEILKTVKEITGYLLI QAWPENRTDLHAFENLEIIRGRTKQHGQFSLAVVSLNVTSLGLRSLKEISDGDVIMSGNKNLKWANKWNWK KVFGTSGQKTKIISNRGE--------
Csp1	M--GAKYKALLDSSSHCVAVGEDCLRHCFEMLAMNDASMGACTKATYDLVAACGALAKLAGTNSAFTP AFAKVVADVCAACKKECDKFPSIAECKACGEACQACAEECHKVA
plr1	M--GAKYKALLESSRRCVRVGERCLRHCREMLRRNDASMGACTKATYDLVKACAELAKLAGTNSARTP KKAKQVARVCEKCKKECDKFPSIAECKACAEACKKCAEECRKVA
plr2	M--GAKYKALLRSSRRCVKVGEECLRHCREMLKRNDASMGACTKATYDLVKACARLAKLAGTNSARTP RRAKRVARVCERCKKECDKFPSIAECKACAEACQRCAEECKKVA
cr3	MGHGAKYKALLESSRRCVRVGERCLRHAREMLRRNDASMGALTKATYDLVKACAELAKLAGTNSARTP KKAKQVARVCEKCKKECDKWPSMAEAKACAEACKKCAEECRKVA
cr61	MGHGAKYKALLESSRRCVRVGERALRHAREMLRRNDASMGAATKAFYDLVKACAELAKLAGTNSARTP KKAKQVARVCEKCKKEADKWPSYAEAKAAAEACKKCAEECRKVA
cr62	MGHGAKYKALLESSRRCVRVGERWLRHAREMLRRNDASMGAATKAAYDLVKACAELAKLAGTNSARTP KKAKQVARVCEKCKKEADKWPSMAEAKAAAEACKKCAEECRKVA
neg1	MGHGAHYAALLESSERCVEVGERCLEHCQEMLEKNDESMGACTKATEDLVKACEELAKLAGTESAQTP ELAAEVARVCEQCQKECDKFPSIEECKECAEACQECAEECEKVA
neg2	MGHGAHYEALLESSERCVEVGERCLEHCQEMLEKNDESMGACTKATEDLVKACEELAKLAGTESAQTP ELAAEVARVCRQCAKECDKFPSIEECKECAEACEECAEECRKVA

Table S3: Survival of zebrafish embryos exposed to different concentrations of inhibitors, related to Fig. 2D.

Inhibitor	Concentration	Lethality (%)
Cetuximab	5.0 mg/ml (34 µM)	30
d3-wt	0.3 mg/ml (16 µM)	10
	1.3 mg/ml (70 µM)	23
dd3-2	0.2 mg/ml (11 µM)	20
	1.0 mg/ml (55 µM)	48

Supplemental Methods S1: Spec files for d3-wt and Csp1 redesign, related to STAR Methods

d3-wt redesign

```
library                         /damietta_v022/libv021_100
input                           d3wt_autopsf.pdb

# designable residues
mut_res                 1                       AFILMVWY
mut_res                 16                      DNHILMWYF
mut_res                 18                      ANSTQNHILMVWYF
mut_res                 26                      AFILMVWY
mut_res                 27                      STEDNQR
mut_res                 48                      TFILMVWY
mut_res                 52                      DE
mut_res                 54                      QREK
mut_res                 57                      DE
mut_res                 68                      HIFWY
mut_res                 71                      ILV
mut_res                 109                     ILVM
mut_res                 127                     ILMV
mut_res                 134                     DKRQNSTEH
mut_res                 135                     ADEFHIKLMNQRSTVWY
mut_res                 138                     DKRQNSTEH
mut_res                 139                     FILMVWY
mut_res                 144                     FILMVWY

# repacking residues
rpk_res                 23
rpk_res                 29
rpk_res                 46
rpk_res                 55
rpk_res                 60
rpk_res                 69
rpk_res                 78
rpk_res                 81
rpk_res                 96
rpk_res                 104
rpk_res                 107
rpk_res                 112
rpk_res                 117
rpk_res                 125
rpk_res                 130
rpk_res                 133
rpk_res                 154
rpk_res                 157

# sampling parameters (optional)
scramble_order                  1
m_mutations                     3
n_paths                         7
n_iters             5 # default:= 1

# mutagenesis scoring weights (optional)
mut_max_lj                      25.0
mut_w_pp                        1.0
mut_w_k                         0.0
mut_w_lj                        1.0
mut_w_solv                      1.0
mut_w_elec                      0.125

# repacking scoring weights (optional)
rpk_max_lj              25.0
rpk_w_pp                1.0
rpk_w_k                 1.0
rpk_w_lj                1.0
rpk_w_solv              1.0
rpk_w_elec              0.125
```

Csp1 redesign

```
library                         /damietta_v022/libv021_100
input                           cu3_wt_autopsf.pdb

# designable residues
mut_res                 21                      KREQ
mut_res                 24                      KREQ
mut_res                 25                      KREQ
mut_res                 28                      KREQ
mut_res                 32                      KREQ
mut_res                 38                      KREQ
mut_res                 42                      DKRQNSTEH
mut_res                 43                      DKRQNSTEH
mut_res                 49                      KREQ
mut_res                 56                      KREQ
mut_res                 60                      KREQ
mut_res                 64                      KREQ
mut_res                 75                      DKRQNSTEH
mut_res                 78                      DKRQNSTEH
mut_res                 79                      KREQ
mut_res                 82                      KREQ
mut_res                 85                      KREQ
mut_res                 88                      KREQ
mut_res                 89                      KREQ
mut_res                 111                     KREQ
mut_res                 112                     KREQ
mut_res                 118                     KREQ

# repacking residues
rpk_res                 17
rpk_res                 31
rpk_res                 35
rpk_res                 39
rpk_res                 53
rpk_res                 57
rpk_res                 67
rpk_res                 81
rpk_res                 92
rpk_res                 93
rpk_res                 96
rpk_res                 115
rpk_res                 119

# sampling parameters (optional)
scramble_order                  1 # default:= 0
m_mutations                     3 # default:= 3
n_paths                         7 # default:= 1
n_iters                         5 # default:= 1

# mutagenesis scoring weights (optional)
mut_max_lj                      25.0
mut_w_pp                        1.0
mut_w_k                         0.0
mut_w_lj                        1.0
mut_w_solv                      1.0
mut_w_elec                      0.125

# repacking scoring weights (optional)
rpk_max_lj              25.0
rpk_w_pp                1.0
rpk_w_k                 1.0
rpk_w_lj                1.0
rpk_w_solv              1.0
rpk_w_elec              0.125
```

Supplemental references

1. Vitalini, F., Noé, F., and Keller, B.G. (2016). Molecular dynamics simulations data of the twenty encoded amino acids in different force fields. Data Brief *7*, 582-590. 10.1016/j.dib.2016.02.086.
2. Shapovalov, M.V., and Dunbrack, R.L., Jr. (2011). A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure *19*, 844-858. 10.1016/j.str.2011.03.019.
3. Blaabjerg, L.M., Kassem, M.M., Good, L.L., Jonsson, N., Cagiada, M., Johansson, K.E., Boomsma, W., Stein, A., and Lindorff-Larsen, K. (2023). Rapid protein stability prediction using deep learning representations. Elife *12*. 10.7554/eLife.82593.
4. Rapaport, D.C. (2022). GPU molecular dynamics: Algorithms and performance. Journal of Physics: Conference Series *2241*, 012007. 10.1088/1742-6596/2241/1/012007.