Principles of *de novo* protein antiviral development for pandemic preparedness

Jeremiah Nelson Sims

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2024

Reading Committee:

David Baker, Chair

Deborah Fuller

Michael MacCoss

Program Authorized to Offer Degree:

Molecular and Cellular Biology

University of Washington

Abstract

Principles of *de novo* protein antiviral development for pandemic preparedness

Jeremiah Nelson Sims

Chair of the Supervisory Committee:

David Baker

Biochemistry

As we navigate the aftermath of the COVID-19 pandemic, the call for advanced medical countermeasures to preempt the next pandemic has echoed across the realm of drug development. Within the pandemic toolbox, drugs and vaccines have been demonstrated as invaluable agents to curb the impact and disease burden of widely disseminated infection, though discovery of novel therapeutic agents remains a major bottleneck. The advent of deep learning-guided protein design offers promise in rationally designing novel, broadly neutralizing therapeutic agents that can be useful for pandemic scenarios. Herein, I apply these deep learning tools to develop miniproteins that cross-react and neutralize viruses with pandemic potential. As a proof of concept, I focus on the henipaviruses (namely Nipah and Hendra virus), a family of zoonotic viruses that top biosecurity watchlists because of their propensity to infect via a variety of routes, their difficulty in diagnosis, and their lack of approved treatments – drugs or vaccines. I also use this platform to identify strategies to improve drug-like properties of proteins in low/medium-throughput assays. Further, I develop and optimize a high-throughput screening modality that tags protein library members with short peptide barcodes designed for readout on high resolution mass spectrometry. I then use this method to screen *de novo* protein

libraries for soluble expression, designed assembly state, and reactivity with their designed target. Looking forward, this work provides a framework for developing drug-like miniproteins that cross-react with multiple homologs using known structural & functional information. Moreover, this peptide barcoding screen offers a streamlined approach for multi-parameter evaluation of entire protein libraries. Combined with *in silico* design and prediction, such an assay could be useful for informing computational tools to improve future rounds of computational design.

Table of contents

Chapter 1 – Introduction 7
1.1. – Just in case: antiviral protein design for pandemic viruses 7
1.2 – Just in time: accelerated development of antiviral proteins 9
1.3 – A survey of protein high throughput screening modalities 10
1.4 – Dissertation overview 14

Chapter 2 – Computational design of potent, broadly neutralizing anti-henipavirals 16
2.0 – HNV preface, authors, and abstract 16
2.1 – HNV introduction 17
2.2 – G protein receptor decoy design 19
2.3 – Characterization of oligomerized minibinders 23
2.4 – Discussion & ongoing work 26
2.5 – HNV materials & methods 27

Chapter 3 – Massively parallel assessment of designed protein solution properties using mass spectrometry and peptide barcoding 30
3.0. – MS barcoding preface, authors, and abstract 30
3.1. – MS-barcoding introduction 31
3.2. – Screening beta barrel monomers 38
3.3. – Screening hallucinated oligomers 42
3.4 – Screening large helical oligomers 46
3.5. – Evaluation of v2 barcode library on tetrahedral nanoparticles 50
3.6. – MS barcoding discussion 53
3.7. – MS barcoding materials & methods 54
3.8. – MS barcoding acknowledgements 64
3.9. – MS barcoding supplementary information 66

Conclusion 83

References 84

Acknowledgements

In chronological-ish order:

Firstly, I thank the village of Black folks – family, and friends, and otherwise kin – who have supported, nourished, encouraged, and trained me to prioritize kindness and humanity. Specifically, I thank Hannah, Stephanie, Rick, Memaw & Papa, Grandma & Grandpa, for bearing with grace our family's traditions. Secondly, I thank those who have shaped my spirit – from Richmond to Yale to DC to Seattle – you all have given me a reason to work and think and love with vigor every day. Special thanks to Deven Shakya, Joey Lew, the XI (each of you, and collectively), Beah Jacobson, Adam Echelman, Meera Garriga, Naveen Jasti, Miguel Paredes, Sanjay Srivatsan, Katya Cherukumilli, Sidney Lisanza, Barbara Reynolds, Jordan Drew, and all the many others who've enriched me. Thirdly, I thank my scientific mentors who have enabled my mind and spirit to flourish. Thank you to Drs. Elsa Yan, Yingying Cai, Kelly Culhane, Sanjay Desai, David Feldman, Basile Wicky, Lukas Milles, and Robert Ragotte who've trained me in the mental and physical discipline of science. Finally, I thank the two who have graciously incorporated my work and dreams into their lives – my brother, Devante Shands, and my partner, Bear Aragon.

None of this work would be possible or worthwhile without your guiding light.

The pursuit of a PhD requires an extraordinary investment of energy, diligence, and time; I have been proud to dedicate my nights, weekends, and holidays to this work for the past three and a half years. I'd like to acknowledge the University of Washington for compensating my PhD colleagues and me for 20 hours of work per week.

Dedication

I dedicate this work to all those Black people who have tried to forge a path forward for others, and to the many more that would, if they had the means.

Chapter 1 – Introduction

The myriad of functions of (and indeed the wonder of) biology can be briefly summarized by its central dogma – where DNA (or RNA) is the script of biology that stores and preserves the identity of an organism, and the resulting protein structure, encoded by its DNA, are the actors that carry out the functions of a host and allow it to interact with its surrounding environment. In reality, there are many more biomolecular cast members that are omitted from this overly simplified narrative, but this simplified view of biology helps us understand the underpinnings of disease – infection, auto-immune, cancer, etc.

1.1 – Just in case: antiviral protein design for pandemic viruses

Viral infection, and our bodies' processes in combating them, are one example of how biology propagates. Viruses are self-replicating, patterned protein capsules, sometimes surrounded by a lipid membrane, that have glycoproteins on their surface that control which target cells they can infect. Once inside of the cell, their protein capsules are programmed to disassemble in response to an environmental trigger, releasing their genetic cargo (DNA or RNA) into the host cell. This genetic cargo undergoes two important events: 1) replication[1] to amplify the number of copies of genetic cargo for new viral particles (virions) and 2)

[1] occasionally with mistakes that change the identity and fitness of the virus

transcription/translation to create the viral proteins necessary for nascent virions to assemble intracellularly and mediate host cell escape, thereby restarting the cycle of infection.

Though millenia of evolutionary competition has prepared the human immune system to be well adapted to combating certain viral families, it is apparent in epidemic/pandemic scenarios that new viral variants arise that are able to evade our immune system enough to cause infect, replicate, spread, and, at worst, cause severe infection and mortality. As such, biosecurity organizations such as CEPI, GAVI, WHO, the Rockefeller Institute, etc. emphasize building a pandemic response that incorporates both therapeutics – exogenous agents that combat the virus – and vaccines – virus component mimics to prime the immune system against that viral threat. At present, there is a bottleneck in lead candidate identification and downstream testing, though evidence from previous epidemic/pandemics suggests stockpiling is a cost-effective strategy (von Delft et al. 2023; Siddiqui and Edmunds 2008; Plans-Rubió 2020; Balicer et al. 2005). While small molecule-based drugs have been an invaluable part of the pharmacopeia, protein-based therapeutics, spearheaded by antibodies – a major class of our bodies' set of immune proteins – have gained traction due to their safety, proven clinical path, and potency. Indeed, antibody engineering has emerged as a field to explore the breadth of this protein class. However, antibodies, owing to their large size and varying stability, are difficult to manufacture and require administration by a healthcare expert, which limit use cases in field scenarios, like at the site of an outbreak (Sifniotis et al. 2019).

Protein design, the pursuit of generating custom biomolecules for a particular task, has demonstrated promise in leveraging the genetic code to make proteins that have the potency and safety of antibodies, while leveraging the scalability and ease of administration that many small molecule drugs offer. Proteins, antibodies and otherwise, are encoded by the same 20 amino acid building blocks that are chained together like beads on a string. This sequence of amino

acids dictates how the protein chain will fold, and the fold will dictate its function. As such, antibodies largely have similar structures, and their overall similar functions in the immune system are defined by their fold.

Protein design affords the possibility of sampling outside of the antibody structural space to identify new solutions, and new drugs. Prior work in the antiviral space leveraged parametric protein design (using idealized sequence and structural elements) to generate a potent antiviral to bind and neutralize SARS-CoV-2 (Cao et al. 2020). Of note, one particularly broadly neutralizing miniprotein grafted the binding helix of human receptor and SARS-CoV-2 entry receptor hACE-2 onto a soluble 3-helix bundle scaffold. These soluble receptor decoys were then incorporated into a trimeric structure (displaying 3 copies of decoys per molecule) that could simultaneously engage all three copies of the receptor binding domain of SARS-CoV-2 (Hunt et al. 2021). The final candidate oligomer receptor decoy was demonstrated to react with the gamut of variants of concern, highlighting the fitness cost of viral escape – mutation away from the receptor decoy results in decreased interaction with the host receptor that bears the same sequence. Thus, receptor decoy strategies and targeting otherwise highly conserved sites on the virus (epitopes) are a promising and underexplored field in antiviral development, one that has become increasingly available with the advancement of protein design.

1.2 – Just in time: accelerated development of antiviral proteins

During an outbreak, in which case number is growing rapidly, it is paramount to be able to quickly characterize antiviral strategies and deploy them to the site of concern. The Baker Lab's previous *de novo* design characterization pipelines utilized chip-based screening and three rounds of iterative yeast surface display to identify binders, identify favorable binding mutations, and explore the epistatic effects of combining these mutations to improve binding

affinity (Cao et al. 2022). Because optimization only occurs with respect to binding (specifically off-rate, k_{off}), the designs that are optimized for affinity at this early stage exhibit a variety of soluble expression and oligomerization profiles that preclude or complicate downstream characterization. In total, an optimistic time estimate is 3 months for binding data collection prior to interrogating the ability to produce the lead candidate with *E. coli* soluble expression. To circumvent this issue, I explored other high throughput assay modalities that could support the modular addition of multiple solution-phase screens, permitting simultaneous evaluation of multiple properties of library members.

1.3 – A survey of protein high throughput screening modalities

High throughput screening modalities can be broken down into two broad categories: 1) surface display techniques and 2) tagged screening techniques. Both are pooled approaches with many members represented in the same sample. Importantly, both are reliant upon linking the protein of interest's performance in the screen (phenotype) to a readout by some unique molecular identifier (genotype) in order to assess each member of the library.

Surface display technologies (phage, yeast, mammalian, bacterial, etc.) are those technologies which utilize the host cell's native surface presentation machinery to display the protein of interest on the cell (Pande, Szewczyk, and Grover 2010; Smith 1985; Boder and Wittrup 1997; Freudl et al. 1986; Charbit et al. 1986; Ho, Nagata, and Pastan 2006; Ho and Pastan 2009). To do so, genetic material encoding the protein of interest is introduced to and maintained by the cell, and the encoding protein is transcribed/translated and presented on the surface. As such, they have inherent phenotype-genotype linkage. When a screen is conducted to separate cells in the overall pool based on their performance (fractionation), the cells in each fraction can be lysed to release the DNA that encodes the protein of interest. With the advent of cost-effective next-generation sequencing tools (Zhang et al. 2020; van der Reis et al. 2022;

Vanderpoel et al. 2022) , surface display techniques have become an invaluable tool to study interactions in high throughput. The ability to achieve tens of millions of sequencing reads in a few days has enabled screening of libraries of tens of millions of proteins. In fact, the practical throughput limitations for these techniques (for non-mutational libraries) lie in the cost and length restrictions (≤180 amino acids for 2-oligo assemblies) of chip-based oligo pool synthesis, not in the complexity of the pools that can be screened. Additionally, surface display on host cells typically involves presentation of multiple copies of the same protein on the surface (multivalent presentation), which permit robust detection of weak or transient binding interactions. As a result, surface display has become a widely utilized technology for interrogation of binding affinities, antibody auto-reactivity, etc. (Cao et al. 2022; Wang et al. 2022).

The major disadvantage of surface display screening approaches is that the protein of interest is immobilized on the surface of a living cell, which precludes assessment of crucial solution-phase properties that are important to consider for biomanufacturing – soluble expression, oligomerization state, thermo-, acid-/base- stability, etc. Furthermore, *in cellulo* artifacts are prevalent due to the inherent biological constraints on the system. For example, it is impossible (without phase separation techniques, which are outside of the scope of this work) to ensure that each cell contains one copy of a DNA molecule, and therefore there are cells that display different numbers of the same protein on a cell, as well as cells that display multiple library members. Further, the relative prevalence of library members in a pool is partially a function of host growth rate, which adds noise to each measurement. These inherent biological noise properties complicate extrapolation of display performance to true values of the parameter of interest.

In contrast, tagged screening technologies (cDNA display, mRNA display, ribosome display, peptide barcoding) are typically performed in solution and incorporate a minimally perturbative tag that serves as a unique molecular identifier (Yamaguchi et al. 2009; Wilson,

Keefe, and Szostak 2001; Zahnd, Amstutz, and Plückthun 2007; Egloff et al. 2019). Since the tag is directly conjugated to the protein of interest, these technologies are well suited for screening either via multivalent display on a variety of surfaces and substrates or via direct, in-solution fractionation. In fact, tagged screening modalities have been gaining popularity due to the possibility of screening a library for multiple properties in series or in parallel, allowing for the identification of library members that satisfy multiple criteria of biological interest (Howell et al. 2014). Because the attached tag is the genotype surrogate, the complexity and throughput of these assays are largely determined by either cost limitations of library synthesis or the upper bounds of the readout technology.

Nucleic acid display technologies (cDNA, mRNA, ribosome) utilize next-generation sequencing formats that permit screening of samples as complex as display technologies ($>10^7$ library members) (Newton et al. 2020). With DNA oligos in hand, these techniques utilize *in vitro* transcription/translation (IVT/IVTT) to make assay-ready library material in a single day, albeit in amounts much less than cell-based production systems. However, the single-molecule sensitivity of next-generation sequencing has transformed this sample production limitation into an advantage, where very few reagents are needed to screen and fractionate these libraries, and very few copies of each library member are needed for a robust, highly reproducible readout.

The major disadvantage to nucleic acid display strategies is that the appended tag is large, often dwarfing the size of the cargo of interest, and is susceptible to natural enzymatic degradation of these tags, presenting a barrier for *in vivo* library screening. While these tags do not present problems in most pooled binding or stability experiments, there are functional barriers to access screens that fractionate on size (e.g. oligomerization state evaluation), screens that require high sample input (e.g. *in vivo* experiments), or screens that evaluate propensity for *in cellulo* production. Nonetheless, nucleic acid display is an attractive option for rapid lead identification and downselecting from otherwise prohibitively complex pools.

Another approach within the realm of tagged screening is peptide barcoding, introduced by Egloff and coworkers (Egloff et al. 2019). This approach features short (7-13 residues), genetically encoded peptides to the N or C terminus of a protein of interest, flanked by cleavage sites for site-specific proteases for facile isolation from the protein of interest and other sequence tags. In the original work, these peptides were designed to be doubly charged precursors that span the range of hydrophobicity and optimal Orbitrap m/z detection (450 - 1000 m/z units), providing a two-dimensional space of 5.3×10^8 peptides from which barcode libraries can be selected based on optimal separation of hydrophobicity and m/z. Upon proteolytic liberation from the fully expressed sequence, these peptides can be reliably detected on high resolution Orbitrap mass spectrometers. One advantage of using peptides is that they are themselves encoded by nucleic acid material, so they can be both sequenced (to identify genotype-phenotype linkages) as well as expressed in *in vitro* and *in cellulo* production systems. An additional advantage that peptide tags offer over nucleic acid tags is that they are small, oftentimes a fraction of the length and size of the protein of interest. As such, they can be used to probe similar solution phase properties as nucleic acid tags, while also being well suited for screens for size, stability, and *in cellulo* production. Thus, peptide barcoding is another approach within tagged screening modalities that is poised to address fundamental and translational questions at the intersection of protein sequence-structure and biomanufacturing and behavior in complex systems. Much like other tagged screening approaches, peptide barcoding has its throughput limited by either the cost of working with complex libraries or the upper complexity bound of a mass spec readout.

The limitations of the peptide barcoding pooled library format are similarly due to the functional behavior of a peptide tag. Because this tag is made of the same building blocks of the protein of interest, there is a non-negligible potential for any given protein tag to interact with and/or influence the protein of interest, resulting in tag-based artifacts. To address this issue, Egloff and coworkers assigned multiple barcodes per protein such that the true behavior of the

protein of interest could be inferred by the average behavior of all the barcoded variants. However, this demonstrated need for library member degeneracy presents a limitation at the level of the mass spectrometer, which, under standard protocols, can differentiate up to 120,000 peptides in the desired space of hydrophobicity and m/z. In brief, the more barcodes needed more design, the more the maximum diversity of the library is restricted. In theory, the maximal diversity of peptide barcoding libraries (using LC-MS/MS) is 2+ orders of magnitude lower than nucleic acid tagged screening. In practice, given the current cost of DNA oligo synthesis, this limitation only presents with libraries of extreme complexity (e.g. deep mutational scanning libraries with 2+ mutations per member). Another practical limitation to this approach is the need for both next generation sequencing and LC-MS/MS, both of which are costly instruments and the latter of which requires technical proficiency. In their work, Egloff and coworkers shotgun cloned a library of ~1000 nanobody sequences into a pool of ~10^8 barcoded vectors, resulting in a stochastic assignment of barcodes to nanobodies that required library downsampling to maximize the number of unique nanobody-barcode pairings (otherwise all barcodes would be assigned to all designs). To uncover genotype – i.e. nanobody-barcode pairings – the authors used next generation sequencing. These libraries were then screened using size exclusion chromatography to fractionate samples based on oligomerization state and target molecule reactivity. However, upon reading and detecting isolated barcodes on MS, they observed 30% of barcodes detected, further imposing a limit on the total throughput of this assay.

1.4 – Dissertation overview

As I began my dissertation work, I set out with the goal of rapid *in vitro* multiple parameter optimization for antiviral proteins, which, for aforementioned reasons, was best to address with a tagged screening approach. In this work, I use protein design to explore the antiviral development and preclinical pipeline for designed protein therapeutics for

henipaviruses, and I create proteins that neutralize henipaviruses *in vitro* as potently as the leading antibody (Chapter 2). Understanding that main failure modes for antiviral miniproteins (and minibinder characterization writ large) were lack of soluble expression and the propensity for off-target oligomerization states, I worked closely with Dr. David Feldman to optimize the peptide barcoding assay for faster time to results and higher fidelity data, enabling its use for screening *de novo* protein libraries. I demonstrate that the barcode identification rate on LC-MS/MS can be increased from ~30% to >90%, and that fewer barcodes can be assigned per library member with no compromise to data quality. (Chapter 3). I hope you enjoy this work as much as I've enjoyed doing it.

Chapter 2 – Computational design of potent, broadly neutralizing anti-henipavirals

2.0 – HNV preface, authors, and abstract

Preface:

Note: this chapter is borrowed directly from the manuscript in preparation of the same name, which is the original copy. I, Jeremiah Sims, am co-first and corresponding author for this work.

Authors:

Jeremiah N. Sims*[1,2,3], Zhaoqian Wang*[4, 5, 6], Kaitlin Sprouse[6], Moushimi Amaya[7], Brendan Larsen[8], Robert Ragotte[1,4], Xinting Li[1,4], Dionne Vafeados[1,2], Jesse Bloom[8,9], Christopher Broder[7], David Veesler[4,9], David Baker[1,4,9]

1. Institute for Protein Design, University of Washington, Seattle, WA 98105, USA
2. Department of Molecular & Cellular Biology, University of Washington, Seattle, WA 98105, USA
3. Medical Scientist Training Program, University of Washington, Seattle, WA 98105, USA
4. Department of Biochemistry, University of Washington, Seattle, WA 98195, USA.
5. School of Medicine, Westlake University, Hangzhou, Zhejiang, 10024, China
6. School of Life Sciences, Westlake University, Hangzhou, Zhejiang, 10024, China
7. Department of Microbiology and Immunology, Uniformed Services University of the Health Sciences, Bethesda, Maryland 20814, USA
8. Basic Sciences Division and Computational Biology Program, Fred Hutch Cancer Center, Seattle, WA 98109, USA
9. Howard Hughes Medical Institute, University of Washington, Seattle, WA 98105, USA

* These authors contributed equally to this work

Abstract:

The henipaviruses (HNV), particularly Nipah virus (NiV) and Hendra virus (HeV), lack approved vaccines or therapeutics despite having cyclical outbreaks with variable case fatality rates ranging from 40-90% in humans. To address this lack of field-ready treatment, we used RFDiffusion to design human EphrinB2 mini-protein receptor decoys that potently cross-react with both NiV and HeV G attachment glycoproteins with subnanomolar affinities. Further, we demonstrate that tetravalent minibinders inhibit viral entry *in vitro* with subnanomolar IC_{50}s comparable to m102.4, the leading anti-HNV antibody used for compassionate care. Unlike antibody therapeutics, we envision these stable oligomers to be amenable to a variety of field-ready routes of administration (respiratory, subcutaneous, intramuscular, etc.).

2.1 – HNV introduction

The henipaviruses (HNV) are zoonotic paramyxoviruses that are responsible for cyclical annual outbreaks in East & South Asia, Australia, Ghana, and Uganda. Using *Pteropus* bats as a reservoir, these viruses are known to infect a variety of mammalian hosts, most notably livestock (horses, pigs, etc.) that permit their transmission to humans via fluid contact or respiratory routes (Luby and Gurley 2012; Weatherman, Feldmann, and de Wit 2018). Two of the notable clade members, Nipah and Hendra viruses, have gained attention from pandemic surveillance agencies for their propensity to cause severe respiratory and neurological disease, with case fatality rates ranging from 40-90% in humans (Kenmoe et al. 2019; Vasudevan et al. 2024; Satter et al. 2023). At present, there are no approved vaccines or specific therapeutics capable of mitigating transmission and disease severity, though antibody-based therapies are currently in

clinical trials for two antibodies, m102.4 & hu1F5, that neutralize by inhibiting the function of the surface glycoproteins (Playford et al. 2020; Zeitlin et al. 2024).

The henipaviruses use two surface glycoproteins for attachment (HNV G protein) and fusion (HNV F protein) to gain entry to cells. The HNV G protein is a tetrameric type-II integral membrane protein that features a beta-propeller fold at the receptor binding domain that binds with exceptionally high affinity to its cognate human receptor, EphrinB2 (EFNB2), as well as EphrinB3 albeit with lower affinity (Bonaparte et al. 2005; Narayanan et al. 2023b; Larsen et al. 2024). Upon receptor engagement, there is evidence to suggest that HNV G receptor engagement triggers a conformational change in the trimeric HNV F protein at neutral pH to initiate membrane fusion with the host cell (Liu et al. 2013). Owing to the cooperation of these two proteins at the cell surface, antibodies inhibiting attachment and fusion have been a primary focus for neutralizing therapeutics.

Antibody m102.4 is a bivalent competitive inhibitor of NiV and HeV G proteins, and it has been demonstrated to protect Syrian hamsters and African green monkeys from lethal disease outcomes in disease challenge models (Dong et al. 2020; Zeitlin et al. 2024). However, m102.4 proved less effective than hu1F5 *in vivo* against $NiV_{Bangladesh}$, a highly virulent strain, despite comparable efficacies *in vitro*. This lack of efficacy is thought to stem from the NiV G tetramer maintaining open binding sites when an antibody binds, an effect that may be mitigated by engineering a tetravalent competitive inhibitor (Zeitlin et al. 2024). In the same study, anti-HNV F monoclonal antibody hu1F5, a bivalent fusion inhibitor that binds at a highly conserved quaternary epitope of the HNV F protein, protected animal models against all tested strains of HNV *in vivo*. Notably, the F protein trimer relies on intact function for all three of its intertwined leaflets, resulting in a lower barrier for effective neutralization.

These antibody-based therapeutics have potential to be potent additions to the pandemic toolbox, owing to their synergistic mechanisms of neutralization. However, antibody-based therapeutics suffer from practical manufacturing and distribution concerns that reduce their

utility in resource-limited field scenarios (Sifniotis et al. 2019). To address the concern of scalable and stable formulations for non-infusion routes of administration, we employed computational protein design to generate potent and stable receptor decoys HNV G to minimize escape potential and off-target binding. Further, we demonstrate that designed tetravalent HNV G decoys neutralize with *in vitro* efficacies comparable to m102.4. We envision that these inhibitors, screened for high soluble expression and stability, will be useful for field-ready routes of administration – intranasal, subcutaneous, intramuscular, etc.

2.2 – G protein receptor decoy design

Using a crystal structure of the complex of hEphrinB2 and HeV G (PDB code: 2VSM), the key G-H loop and beta strand of the EFNB2 binding interface (residues 109-128, QDIKFTIKFQEFSPNLWGLE) were chosen as a minimal interface to scaffold (Figure 1a). Previous research supports the importance of the G-H loop for binding in the HNV pocket (Narayanan et al. 2023a; Bowden et al. 2008), and we hypothesized that the adjacent beta strand itself would serve as an important secondary structural feature to guide RFdiffusion to create a soluble scaffold. As input to the RFDiffusion scaffold backbone generation, we provided the G-H loop and preceding beta strand (EFNB2's residues 109-128) and a truncated version of NiV G from 2VSM that maintains all interface contacts for the included EFNB2 domain (Figure 1b). To ensure that an appropriate diversity scaffold could be generated without restricting the overall position of the fixed EFNB2 motif, RFdiffusion was permitted to sample from a range of 15-60 residues on either side of the motif, allowing a maximum backbone length range of 55-100 residues (including the EFNB2 motif). A total of 18,933 backbones were generated, which then served as input for sequence design.

To assign amino sequences to each backbone, ProteinMPNN was alternated with FastRelax as previously described ((Watson et al. 2023), Methods), assigning two sequences per backbone, and each relaxed structure with side chains was output. During sequence design, we

fixed the residue identity of the scaffolded EFNB2 motif to design a sequence around the motif that would scaffold the EFNB2 interface without compromising its structure and function. A total of 37,873 design models with unique sequences were used as input to AlphaFold2 (AF2, (Jumper et al. 2021) for single sequence folding prediction without multiple sequence alignment.

We first attempted to use AF2 with initial guess (Watson et al. 2023), which initializes complex prediction with the designed interaction to assist with prediction of interchain contacts, as previously described. Briefly, we aimed to use the predicted local distance difference test (pLDDT, unitless) score as a metric for AF2's confidence in the predicting the structure of the receptor decoys, while we set out to use predicted aligned error (pAE, Å) to evaluate AF2 confidence in the per-residue pairwise interactions between the receptor decoy and the truncated G protein.

While the pLDDT values for these designs ranged from low confidence (<30) to high confidence (>85), AF2 failed to recapitulate the interface for any of the designs, resulting in pAE values ≥ 20 Å for all designs (Supplementary Figure 1). In output structures, receptor decoys were placed distantly from NiV G, suggesting that AF2 was unable to predict the known interface of the grafted EFNB2 motif with NiV G. To overcome this complex prediction artifact, we templated a minimal motif of 4 key residues of the EFNB2 loop (residues 120-123, FSPN) that make deep pocket interactions with NiV. When these residues were fixed in the pocket, we recovered 17.5% of designs with pAE < 20Å (Supplementary Figure 1). We downselected to 22 designs with favorable *in silico* metrics by **(1)** filtering for high AF2 confidence for non-EFNB2 motif residues in the *de novo* scaffold (scaffold pLDDT > 80) **(2)** identifying designs predicted to complex with NiV G (pAE < 10 Å), and **(3)** electing design models that agree with AF2 predictions (C_α-RMSD < 1 Å between AF2 prediction and relaxed MPNN design model).

The top 22 designs (all ~10 kDa) were cloned into a pET-28b(+)-based expression plasmid from e-blocks, expressed in BL21 *E. coli,* and purified at 50mL scale with a C-terminal

tandem SNAC tag and 6x-His tag. After Ni-NTA purification, candidate designs were sized in Tris-buffered saline (50mM Tris, 150 mM NaCl, pH 8) on an Superdex 75 Increase 10/300 GL to screen for oligomerization state. Out of the screened designs, 18/22 expressed, and 6/22 had prominent peaks in the expected monomer fraction (2/22 monodisperse) between 14-16 mL. We used Octet biolayer interferometry to measure binding affinity of the monomer fraction (as the immobilized ligand) to NiV G titrated 2-fold from 100 nM to 1.5625 nM, and one design, hnb.v1.10, had a measured affinity of ~ 5nM. Further, we validated competitive binding by immobilizing EFNB2 and measuring binding to NiV G in the presence and absence of 100x molar excess of hnb.v1.10. Following up with a sequence-verified clone of hnb.v1.10, we used SPR to confirm its binding to immobilized NiV G (3.6 nM) and HeV G (2.4 nM); no detectable binding signal was observed to Ghanaian bat virus G, a distant henipaviral clade member with 30% sequence similarity to either NiV G or HeV, despite its use of EFNB2 as an entry receptor (Figure 1c).

Acknowledging the high affinity of the cognate EFNB2-NiV G complex (~340 pM), we used partial diffusion and ProteinMPNN (as previously described in (Watson et al. 2023)) to improve the affinity of 2vsm10 by redesigning its sequence over two rounds of design. In the first round, we generated 95,988 sequences, and we selected 64 sequences to screen as e-blocks at 4-mL culture scale (described previously in (Watson et al. 2023)). After expression and purification by Ni-NTA, proteins were screened by SEC and SPR for minimally aggregating designs with subnanomolar affinity to NiV G. Three hits, called hnb.v2.E2, hnb.v2.F1, and hnb.v2.G9, had prominent peaks in the expected elution fraction by SEC and improved affinities (0.133 nM, 0.546 nM, 0.176nM, respectively). Only hnb.v2.F1 and hnb.v2.G9 replicated soluble expression and binding at the 50mL scale with sequence-verified clones, though hnb.v2.G9 did have a minor aggregate peak. We solved a cryo-EM structure of hnb.v2.G9 to 3Å resolution, validating that the cognate EFNB2 loop-strand interface is recapitulated when binding to NiV G (Figure 1d).

We next used ProteinMPNN's soluble weights to address hnb.v2.G9's aggregation by redesigning its core and solvent-exposed residues. We designed 101 sequences and selected 35 designs passing *in silico* filters for expression and purification at 4-mL scale. Of the 35 proteins, 8/35 were monomeric by SEC and 3/8 maintained binding < 200 pM to NiV G. From the screen and clonal follow-up, we identified hnb.v3.1 and hnb.v3.23 as highly soluble, well expressing, high affinity candidate monomers to both NiV G (69 pM & 105 pM, resp.) and HeV G (72 pM & 81 pM, resp., Figure 1e); these two sequences have 86% sequence identity.

In summary, we used our computational design pipeline to identify a lead candidate with initial 5 nM affinity, and we were able to computationally improve its affinity 50x and maintain high soluble expression. Understanding that NiV G is a tetramer, we next sought to potentiate neutralization by exploring the impact of oligomerization on neutralization potential.

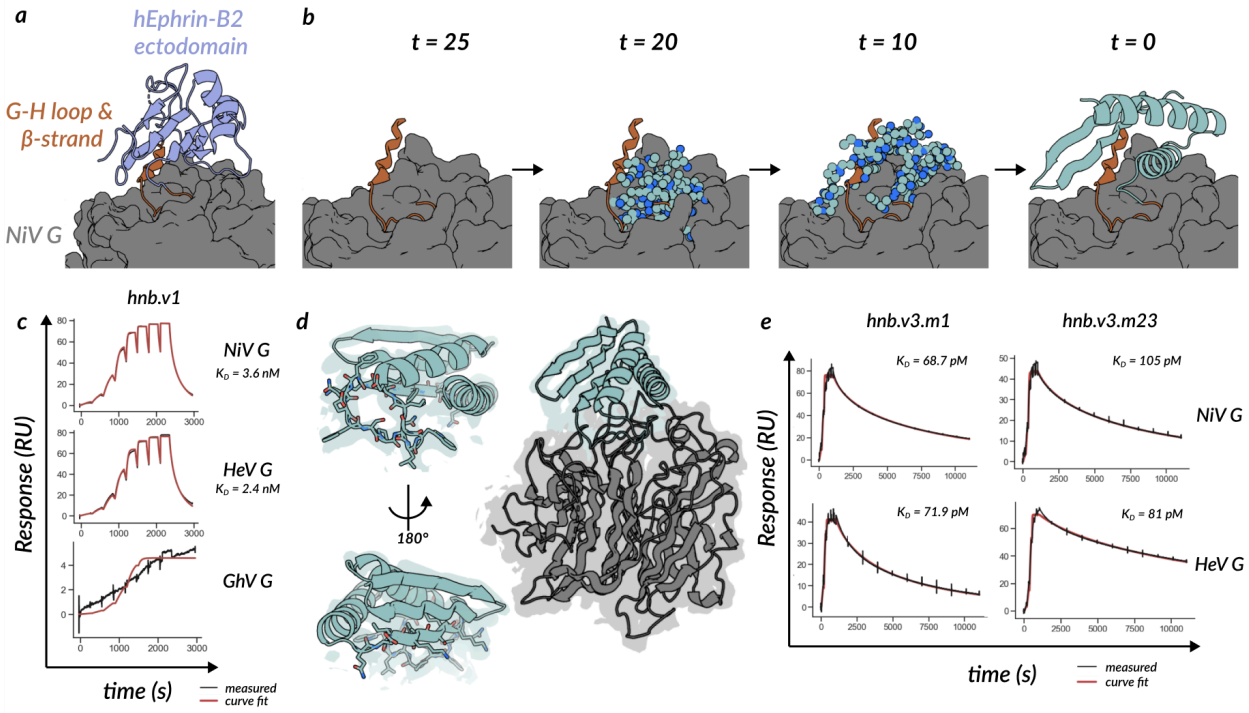


Figure 1. Design and characterization of anti-HNV minibinders. (a) Crystal structure (PDB: 2VSM) of hEphrinB2 (hEFNB2, purple) ectodomain bound to the head of Nipah virus (NiV) G protein (grey). The G-H loop of EFNB2 (red) was chosen for scaffolding because it constitutes the majority of the EFNB2-NiV G interface and is essential for HNV G attachment. **(b)** RFDiffusion timesteps to generate novel scaffolds (teal) for the G-H loop (red). **(c)** Surface plasmon resonance (SPR) single-cycle kinetics of first generation HNV G miniprotein hnb.v1 binding to immobilized NiV G and HeV G with single-digit nanomolar affinities. Curve fit (red) to measured signal (black) was

used to determine kinetic and equilibrium binding constants. **(d)** Cryo-electron microscopy of hnb.v2.G9 (teal) bound to NiV G (grey) solved to 3.7 Å. Design models (cartoon) were docked into electron densities (cloud) for second-generation hnb.v2.G9 bound to NiV G. **(e)** SPR single-cycle kinetics of third-generation soluble, monomeric minibiners hnb.v3.m1 and hnb.v3.m1 binding to immobilized NiV and HeV G.

2.3 – Characterization of oligomerized minibinders

The tetrameric state of EFNB2 and its high affinity interaction with HNV G proteins present a high barrier to neutralization by competitive inhibition. Zeitlin and coworkers recently investigated the efficacy of anti-G antibodies *in vivo* and determined that m102.4 failed to protect African Green Monkeys from fatality when challenged with $NiV_{Bangladesh}$, the most virulent known strain of any HNV to date, despite demonstrating comparable potency across all strains of interest *in vitro (Zeitlin et al. 2024)*. They posited that the $NiV_{Bangladesh}$ tetrameric G protein, bound by bivalent m102.4, maintained unoccupied EFNB2 binding sites that permitted host receptor attachment and subsequent viral invasion. In line with this hypothesis, we constructed a library of previously characterized cyclic oligomers (ranging from C1-C8) that would be suitable for flexible fusion of the minibinder to the N- or C-terminus via glycine-serine repeat linker. We hypothesized that highly avid minibinder-oligomer fusions would potentiate viral neutralization over the monomer by 1) increasing complex half-life with the receptor decoy and 2) reducing the fraction of unbound HNV G protein receptor binding sites.

To assess if multivalent presentation would prolong complex half-life, we constructed a small library of 48 vectors, each containing the sequence of one oligomer and a terminal tandem

SNAC tag and 6x-His tag (Figure 2a). We first cloned hnb.v1.10, which has a $t_{1/2}$ of 83s to NiV G, in these vectors, and we expressed and purified these cultures at 4-mL culture scale. We tested 8 child oligomers with monodisperse SEC for prolonged off-rate on SPR with immobilized G protein (Figure 2b). All 8 of these oligomers bound immobilized NiV G without complete dissociation after 10,000s, confirming the hypothesis that multivalent display of anti-henipaviral minibinders prolong complex half-life.

We then used a similar oligomer vector library of 48 vectors to clone sequences for hnb.v3.m1 and hnb.v3.23 into these vectors to create a library of 96 unique minibinder-oligomer combinations. Expressed and purified at 4-mL culture scale, we evaluated oligomerization state by size exclusion chromatography. For parent monomers hnb.v3.m1 and hnb.v3.m23, 18/48 hnb.v3.m1 oligomers and 10/48 hnb.v3.m23 oligomers had monodisperse elution profiles with peaks at the expected elution volume. We scaled 12 total oligomers (7 for hnb.v3.m1 and 5 for hnb.v3.m23) spanning C2-C8 geometries for 50 mL expression. After sizing, we rebound protein to Ni-NTA resin for overnight SNAC cleavage (previously described) and subsequently sized via SEC to buffer exchange into TBS and remove $Ni2^{2+}$-induced aggregates. Candidate oligomers with their parent monomers were assessed for tag cleavage (mass spectrometry) and dispersity of oligomerization state (mass photometry). Sizing by mass photometry is only sensitive enough for oligomers >50 kDa in mass, restricting analysis of monomers and C2s (6 designs). Nonetheless, 6/6 of C3+ oligomers were revealed to have peaks at the expected molecular weight, and 2/6 were monodisperse, indicating that these designs are robust oligomeric assemblies. All 12 designs, 2 parent monomers and 12 child oligomers were then subjected to an *in vitro* neutralization assay to interrogate the impact of oligomerization on neutralization.

To assay neutralization, we generated a pseudotyped HIV expressing NiV G protein on the surface, and we assessed their ability to infect cells expressing hEphrinB2 in the presence of increasing concentrations of inhibitor. We benchmarked monomer and oligomer minibinders against m102.4 to measure inhibition (as IC_{50}, concentration at which 50% viral entry is

achieved) relative to the leading clinical candidate. In neutralization assays (n=6 for all data points), all child oligomers neutralized more potently than parent monomers (Figure 2c), confirming that multivalent display and prolonged complex half-life translates to improved neutralization. Further, we observed that 1 dimer, 1 trimer, and 3 tetramers inhibited viral entry with comparable IC_{50} values to m102.4. One leading tetramer, hnb.v3.m1C12 yielded reproducible subnanomolar IC50s 2-4x lower than m102.4 ($p = 4.78 \times 10^{-5}$, $\alpha = 0.05/12 = 0.004$), while IC_{50} values for potent tetramers hnb.v3.m1C12 ($p = 0.00732$) and hnb.v3.m23C11 ($p = 0.197$) were not significantly different than m102.4.

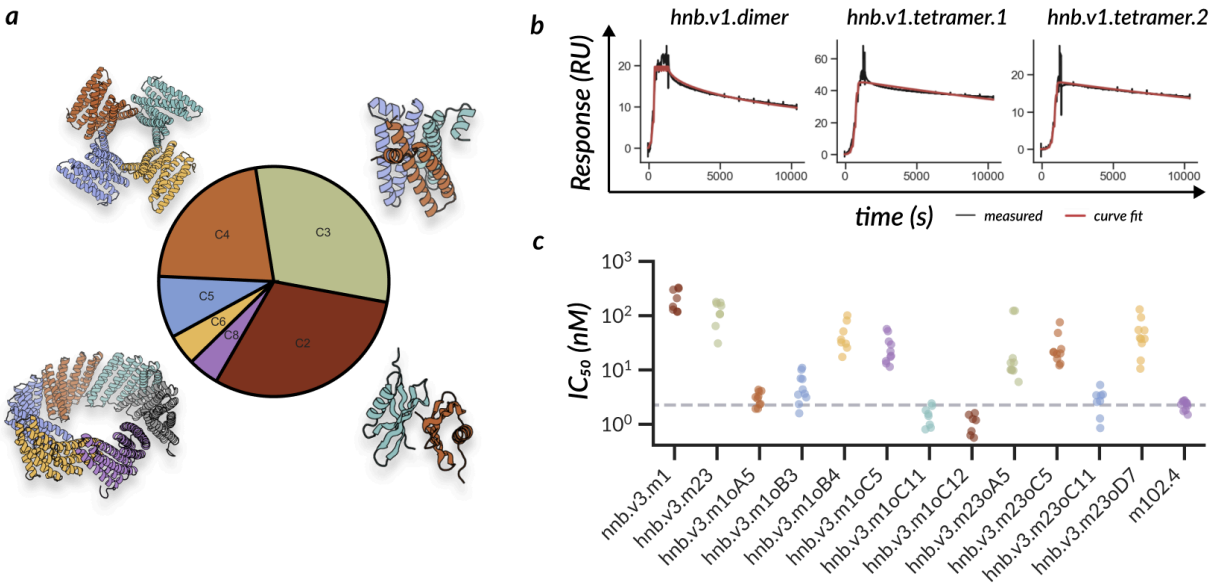


Figure 2. Constitution and characterization of anti-HNV G oligomers. (a) Distribution of validated *de novo* cyclic oligomers used for multivalent minibinder display based on accessibility of N- and/or C- termini. (b) SPR dissociation over 10,000s of oligomerized first generation minibinder hnb.v1 as a dimer or two different tetramers binding to immovilized NiV G. Multivalent minibinders fail to fully dissociate over the course of the experiment, whereas monomer hnb.v1 dissociates with a $t_{1/2}$ of ~90s (Figure 1c). (c) *in vitro* neutralization of third generation minibinder monomers (hnb.v3.m1 and hnb.v3.m23) and their child oligomers benchmarked against m102.4 (grey dotted line = mean IC_{50}). Two tetramers (hnb.v3.m1oC11 and hnb.v3.m1oC12) inhibit 2-4x more potently than m102.4.

2.4 – Discussion & ongoing work

This work explores the process of solubilizing and affinity maturation of a EFNB2 receptor decoy tailored to HNV G glycoproteins. This work, while preliminary, highlights the power using deep learning-guided protein design to generate receptor decoys to viruses with pandemic potential. Further, we highlight how cutting-edge protein backbone and sequence design tools can be leveraged to optimize manufacturability of protein products to maximize their potential for downstream success. Over three rounds of designs, we improve both affinity and solubility of the monomer, which permits scaffolding on stable oligomers to achieve potencies *in vitro* greater than that of a leading antibody.

Looking into the next few months, we are exploring a few key routes to map out a playbook for antiviral protein design:

First, we aim to explore viral escape potential from both monomers and the most potent child oligomers. To do so, we are simultaneously performing deep mutational scanning of 1) the monomers for mutations that improve affinity to HNV G (and rescue cross-reactivity to GhV G), and 2) a bat variant of EFNB2 (so as not to conduct gain of function research) to anticipate which mutations permit escape of our receptor decoys without loss of fitness to the receptor of interest. In this work, we hypothesize that non-native scaffold contacts to the receptor surface abrogate binding to GhV G – as such, there are likely mutations in NiV G or HeV G that could selectively reduce affinity to our receptor decoy without compromising fitness to hEFNB2.

Second, we seek to test our receptor decoys *in vivo* to explore whether tetravalent receptor decoys are effective in protecting against severe pathology for the range of available NiV and HeV variants. The standard animal models for HNV challenge studies, Syrian golden hamsters and African green monkeys (AGMs), have been useful in benchmarking the neutralization potential of leading antibodies (Zeitlin et al. 2024). Since m102.4's bivalency was hypothesized to be the cause of only partial protection from $NiV_{Bangladesh}$ in AGMs, we hypothesize

that tetravalency might afford enhanced protection for this highly virulent strain. Failure to protect might indicate the need for dual therapy targeting the F protein. Further, we want to explore a variety of administration regimes (intranasal, subQ, IM, etc.) to explore the options for field-ready formulations for protein antivirals.

Third, we are working on using computational design to bind highly conserved neutralizing epitopes of the F protein. At the time of this writing, we have identified a number of candidate miniproteins that cross react with NiV and HeV F protein; however this preliminary work is outside of the scope of this dissertation. We envision that a strategy of neutralization via both prominent surface glycoproteins will create a potent and escape resistant formulation to address the threat of a henipaviral epidemic.

2.5 – HNV binder materials & methods

Golden gate assembly and chemical transformation

As previously described (Watson et al. 2023), DNA synthesized as E-blocks from IDT encoding the protein of interest were cloned into pET-29b(+) vectors with C-terminal tandem SNAC and 6x-His tags modified with BsaI cut sites, leaving compatible sticky ends (5' –AGGA–gene of interest–TTCC – SNAC – 6x-His – 3') following vendor instructions for BsaI-HF (New England Biolabs). Plasmid was transformed without cleanup into chemically competent BL21-DE3 (New England Biolabs), and cells were recovered for 1 hour at 37 °C in SOC Outgrowth Medium (New England Biolabs). In the 96-well format, transformants were used to inoculate 4 x 1 mL plates of auto-induction media (Watson et al. 2023), and a 1 mL plates with LB + 50 μg/mL kanamycin sulfate was prepared for glycerol stocks after overnight growth at 37 °C with shaking at 1000 rpm. Otherwise, transformants were spread on plates with LB + 100 μg/mL kanamycin sulfate, and colonies were sequenced. Colonies were then used to inoculate 50 mL cultures in 250-mL baffled flasks for 20h growth at 37 °C with shaking at 250 rpm.

50-mL expression, purification, and SEC

Proteins were purified as described previously (Watson et al. 2023). Cell pellets were isolated with centrifugation, resuspended in lysis buffer (25 mM Tris HCl pH 8, 300 mM NaCl, 40 mM Imidazole, 1 mM DNase I, 10 μg /mL lysozyme), and and cells were lysed via sonication (5 min, 85% amplitude, 15s on/off cycles). Centrifugation was used to clarify lysate (14,000 x *g*, 20 min.), prior to its addition to 1-mL Ni-NTA columns pre-equilibrated with wash buffer (25 mM Tris HCl pH 8, 300 mM NaCl, pH 8). Resin was washed with 5 column volumes of wash buffer, and proteins were eluted in 1.6 mL elution buffer (50 mM Tris HCl pH 8, 300 mM NaCl). Samples of 1 mL were injected onto a Superdex-75 Increase 10/300 GL (Cytiva) or a Superdex-200 Increase 10/300 GL (Cytiva) for monomers and oligomers, respectively, and absorbance at 280 nm was used to track protein elution.

Octet biolayer interferometry

NiV G (prepared as described in (Cao et al. 2022; Wang et al. 2022)) was chemically conjugated as ligand to tips from an Amine coupling kit (Sartorius) according to manufacturer instructions in acetate buffer pH 5. Binders were diluted serially in octet buffer (HBS-EP (+), pH 7.4, Cytiva) as analyte 2x from 1μM in 6 wells, with two wells reserved as no-analyte and no-ligand controls. After establishing a baseline in octet buffer, ligand was exposed to binder for association and then placed into octet buffer for dissociation. The same setup was repeated for successful binders with binder immobilized as ligand, and NiV G titrated as analyte.

96-well, 4-mL protein purification and size-exclusion chromatography (SEC)

Proteins were purified similarly (described in (Watson et al. 2023)) to the 50-mL protocol with a few exceptions. After pellet isolation, pellets were lysed with B-PER™ Bacterial Protein Extraction Reagent (Thermo Scientific) at 37 °C for 15 min at 250 rpm. Plate-based Ni-NTA

pulldowns used 50 µL Ni-NTA resin, and proteins were eluted in 100 µL elution buffer, spun through a 0.22 µm filter plate (Pall – AcroPrep Advance 96-well Filter Plates - 2 mL, 0.2 µm wwPTFE membrane). Sizing was performed on a Superdex-75 Increase 5/150 GL (Cytiva) or a Superdex-200 Increase 5/150 GL (Cytiva) for monomers and oligomers, respectively.

Surface plasmon resonance (SPR) kinetic measurements

HNV G (prepared as described in (Cao et al. 2022; Wang et al. 2022)) was immobilized on a CM5 SPR chip (Cytiva) in acetate buffer, pH 5, and binder plates were prepared with 6 x 5-fold serial dilutions in HBS-EP (+), pH 7.4 (Cytiva) for a single-cycle kinetics experiment, as described previously ((Watson et al. 2023)). To assay reactivity to EphB4, we immobilized Fc-tagged EphB4 (AcroBiosystems) to a protein A chip (Cytiva) for single-cycle kinetics experiments.

Chapter 3 – Massively parallel assessment of designed protein solution properties using mass spectrometry and peptide barcoding

3.0 – MS barcoding preface, authors, and abstract

Preface:

Note: this chapter is borrowed directly from the manuscript in preparation of the same name, which is the original copy. I, Jeremiah Sims, am co-first and corresponding author for this work.

Authors:

David Feldman*[1,2], Jeremiah N. Sims*[1,3,4], Xinting Li[1,2], Richard Johnson[5], Stacey Gerben[1,2], David Kim[1,2], Christian Richardson[1,6], Brian Koepnick[1,2], Helen Eisenach[1,2], Derrick Hicks[1,2], Erin Yang[1,2], Basile Wicky[1,2], Lukas Milles[1,2], Asim K. Bera[1,2], Alex Kang[1,2], Evans Brackenbrough[1,2], Emily Joyce[1,2], Banumathi Sankaran[7], Josh Lubner[1,2], Inna Goreshnik[1,2], Dionne Vafeados[1,2], Aza Allen[1,2], Lance Stewart[1,2], Michael J. MacCoss[5], David Baker[1,2,8]

1. Institute for Protein Design, University of Washington, Seattle, WA 98105, USA
2. Department of Biochemistry, University of Washington, Seattle, WA 98105, USA
3. Department of Molecular & Cellular Biology, University of Washington, Seattle, WA 98105, USA
4. Medical Scientist Training Program, University of Washington, Seattle, WA 98105, USA
5. Department of Genome Sciences, University of Washington, Seattle, WA 98105, USA
6. Department of Bioengineering, University of Washington, Seattle, Washington 98105, United States

7. Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA.
8. Howard Hughes Medical Institute, University of Washington, Seattle, WA 98105, USA

* These authors contributed equally to this work

Abstract

Library screening and selection methods can determine the binding activities of individual members of large protein libraries given a physical link between protein and nucleotide sequence, which enables identification of functional molecules by DNA sequencing. However, the solution properties of individual protein molecules cannot be probed using such approaches because they are completely altered by DNA attachment. Mass spectrometry enables parallel evaluation of protein properties amenable to physical fractionation such as solubility and oligomeric state, but current approaches are limited to libraries of 1,000 or fewer proteins. Here, we improved mass spectrometry barcoding by co-synthesizing proteins with barcodes optimized to be highly multiplexable and minimally perturbative, scaling to libraries of >5,000 proteins. We use these barcodes together with mass spectrometry to assay the solution behavior of libraries of *de novo*-designed monomeric scaffolds, oligomers, binding proteins and nanocages, rapidly identifying design failure modes and successes.

3.1 – MS-barcoding introduction

Rapid improvements in computational methods have enabled the design of proteins with increasingly sophisticated structure and function (Watson et al. 2023; Dauparas et al. 2022; Baek et al. 2021). Nevertheless, for many tasks the success rate remains low, and the deficiencies in computational models are unclear. Library screens address this gap via efficient pooled testing of thousands of designs, identifying candidates for further development and providing

feedback to improve the design process. A wide variety of enrichment methods have been applied to protein libraries, including display methods to measure binding affinity (Boder and Wittrup 1997) and protease stability (Tsuboyama et al. 2023), fluorescent substrate turnover for enzymatic activity (Harris et al. 2000), and split GFP or enzymatic reporters as proxies for solubility (Cabantous and Waldo 2006; Zutz et al. 2021). However, these methods rely on physical linkage between the protein and the DNA sequence which encodes it, and this requirement has precluded directly screening proteins in solution for key biophysical properties such as oligomeric state, solubility, and expression yield (Egloff et al. 2019). Mass spectrometry, by contrast, enables parallel evaluation of protein properties such as aggregation propensity and resistance to chemical or thermal denaturation in solution. Physical fractionation methods such as size exclusion chromatography or bead-based pulldown of substrate-bound protein can be performed on a sample containing many proteins of interest, and the identities of the proteins in each fraction determined by mass spectrometry. However, direct application of shotgun proteomics to large libraries of designed proteins is frequently limited by high sequence similarity among designs. To circumvent sequence similarity, mass spectrometry multiplexing via peptide barcodes was recently developed to evaluate nanobody solution binding, monomericity, and expression (Egloff et al. 2019). However, this method used stochastic linking of peptide barcodes with randomized sequences to designs and was thus limited to a pre-enriched pool of 1000 nanobodies attached to ~12,000 total barcodes.

We reasoned that an approach combining mass spectrometry with peptide barcodes could provide a powerful way of assessing the properties of thousands of designed proteins in solution. We set out to optimize mass spectrometry barcoding for measuring the properties of diverse *de novo*-designed proteins, including monomeric and oligomeric scaffolds, minibinders, and nanocages. Since the previous approach used a shotgun cloning strategy to stochastically assign barcodes to a small nanobody library, there was an inherent need to use NGS to identify unique

barcode assignments to designs. While the previous approach proved effective, we reasoned that pre-assignment of barcodes at the DNA oligo level would afford greater overall throughput than a shotgun cloning approach, because pre-assignment would mitigate non-unique pairings of barcode and designed protein. Additionally, the previous approach suffered from barcode dropout at the level of mass spectrometry, likely due to barcode-specific differences in ionization efficiency. Thus, we aimed to identify a set of barcodes that would ionize reliably, thereby further increasing the fidelity of barcode identification, and subsequently, improving throughput. We began by seeking to design peptide barcodes *in silico* that are (1) co-synthesized with designs on an oligonucleotide array; (2) easily purified for mass spectrometry (Figure 3a); (3) minimally perturbative to the attached designs; and (4) efficiently separated and quantified by high resolution orbitrap liquid chromatography-coupled tandem mass spectrometry (LC-MS/MS) (Figure 3b).

To meet criteria (1) and (2), we adapted pET-28, a T7 expression vector, for library cloning of 300-nt oligos containing barcodes fused to protein designs of up to 74 amino acids, or 154 amino acids if using oligo assembly. Proteins expressed from this vector contain a barcode flanked by a designed protein and either an N- or C-terminal His-tag. Arginine and lysine residues were restricted to the barcode boundaries to enable facile isolation of barcodes from the protein of interest and tags by sequential protease digest by LysC digest, His-tag purification, and then trypsin digest. Barcodes were limited to 8 to 13 amino acids in length and predicted to generate doubly-charged precursors by electrospray ionization, in order to optimally cover the mass-to-charge range of high resolution orbitrap mass spectrometry. To meet criterion (4), the amino acid content of the barcodes was limited to avoid bulky hydrophobic residues likely to disrupt folding, as well as residues that affect net charge in electrospray ionization, residues likely to undergo chemical modification, and residues that interfere with tryptic cleavage (Methods). Finally, LC indexed retention time (iRT, a standardized measurement for predicting

elution, (Escher et al. 2012)) and MS/MS fragmentation spectra were predicted for candidate barcodes using Prosit, and a first-generation set of up to 100,000 barcodes was defined based on separability at the expected m/z and LC resolution.

As an initial test, a subset of 5,000 first-generation barcodes was synthesized and appended to Foldit1, a highly stable and soluble protein characterized previously (Koepnick et al. 2019). The barcodes were expressed in subpools and mixed to generate a standard curve for ratiometric quantification. After expression, barcodes were purified directly from *E. coli* lysate and analyzed by data-dependent acquisition (DDA) on an Orbitrap mass spectrometer (Methods), and barcode intensities were quantified by integrated MS1 area (Methods). There was little contamination from E coli proteins, while LC elution times for 88.08% of barcodes were accurately predicted (Figure 3c), confirming that Prosit predictions generalize with high accuracy to synthetic peptide sequences (Figure 3c).

Application of the first-generation barcodes to the larger design libraries in this study revealed that fewer barcoded designs were identified via mass spectrometry than by DNA sequencing (~30% detection rate across libraries with >10,000 barcodes), similar to the barcode design detection rate reported in Egloff, 2019. We suspected that this detection failure might be due to both unstable designs and suboptimal barcode sequence content for our mass spectrometry protocol; to isolate the latter scenario, we barcoded muGFP, a hyperstable monomeric variant of green fluorescent protein (Figure 3d). We sought to improve the barcode detection rate by optimizing the sensitivity of our mass spectrometry protocol and devising a second-generation barcode set based on the experimental data. Stochastic detection and quantification is a known limitation of data-dependent acquisition (DDA) (Barkovits et al. 2020), which relies on ion abundance in MS1 scans to trigger MS2 scans for peptide identification. This focus on abundant ions is advantageous when parent ions of the highest abundance are of primary interest

(Bateman et al. 2014). In contrast, data-independent acquisition (DIA) protocols have less bias because the MS2 isolation windows are systematically stepped throughout a precursor m/z range such that all precursors in that m/z range are subjected to MS2 analysis, regardless of the precursor abundance (Egertson et al. 2015; Kawashima et al. 2019). While these data require more intensive deconvolution than DDA spectra, available software can deconvolute DIA spectra with high quantitative accuracy (Lou et al. 2023; Demichev et al. 2020).

We compared DDA and DIA for barcode identification by linking ~25,000 first-generation barcodes to muGFP for single-sample readouts that lack the peptide identification benefit of matching peptide IDs between runs. Whereas DIA used the MS2 signals for quantification, our DDA protocol relied upon the MS1 signal for this purpose. The unoptimized DIA protocol (at 500,000 resolution, 5µm packing silica) yielded more peptide identifications than our DDA protocol, supporting the hypothesis that DIA is more reliable in detection of these barcodes at higher pool complexities. We further improved sensitivity and quantitative accuracy using columns with finer particle sizes (from 5µm to 1.9µm) that provided higher chromatographic resolution and reducing the orbitrap MS1 resolution (from 500,000 to 30,000) to permit a more even survey of the sample. Together these resulted in a doubling of the barcode identification rate (30% to 60%, Figure 3d).

Despite the increase in peptide identification rate from our optimized DIA protocol, we still observed 40% barcode dropout. Barcodes with a high hydrophilicity score, and particularly barcodes rich in acidic residues Asp and Glu, were detected at lower rates (Supplementary Figure 1). Using this information, we generated a second-generation barcode set with fewer acidic residues that achieved an overall 86.6% barcode recovery rate with the optimized DIA protocol (Figure 3d).

We investigated whether barcode pre-assignment could reduce the high rates of barcode swapping observed by Egloff and colleagues during PCR amplification steps and explained as mega-primer formation due to high sequence homology. We sought to reduce barcode swapping by maximizing coding sequence distance over the set of proteins being examined, maintaining the same coding DNA sequence for each barcoded variant of the same protein. For our single oligo libraries (those not requiring two-oligo assembly), 99% of barcodes exclusively mapped to the design of interest, and for the two-oligo assembly, requiring an additional qPCR amplification to stitch together 5' and 3' oligos from separately amplified subpools using homologous junctions unique to each design, 75% of barcodes exclusively mapped to the intended design (Figure 3e). The low rate of barcode mismapping considerably increases library coverage, enabling screening of larger libraries.

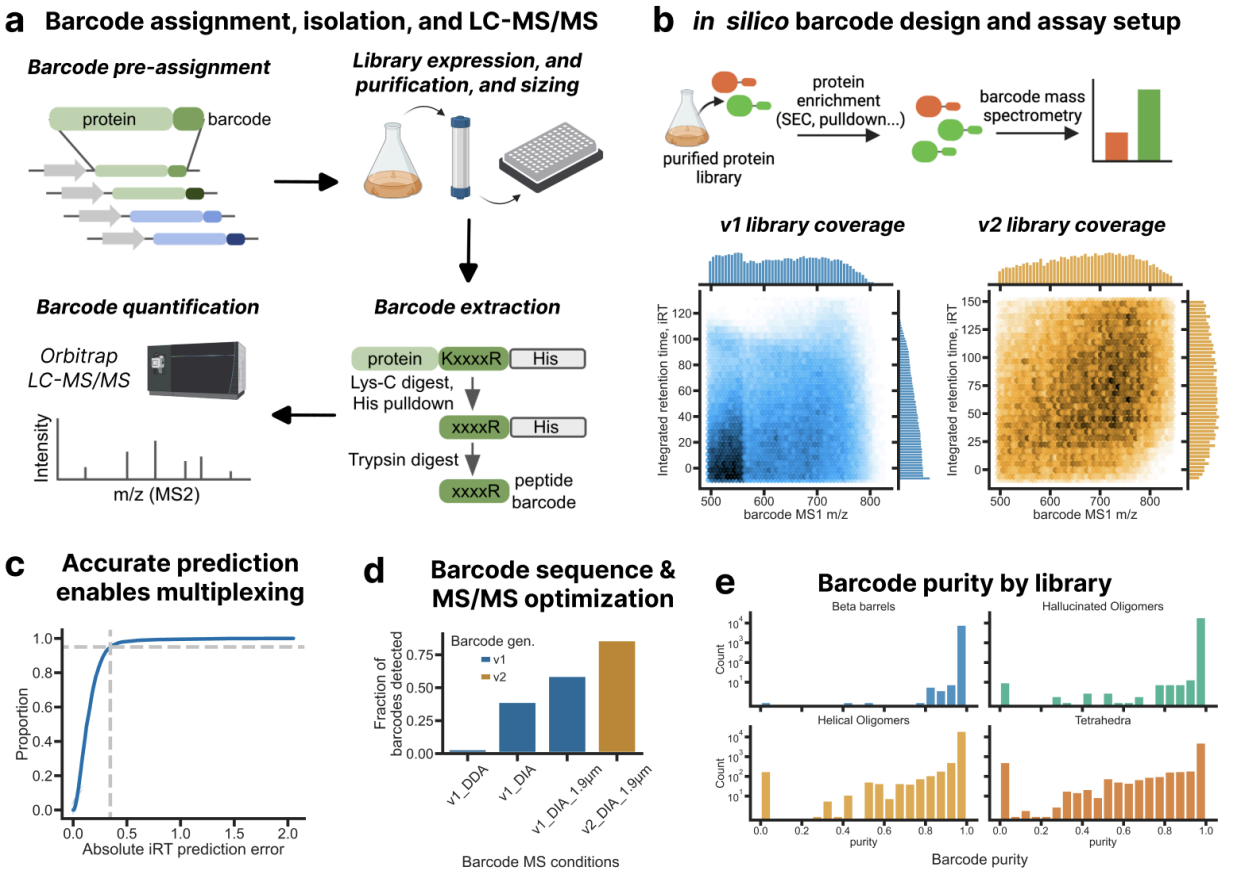


Figure 3: Barcode design and characterization. **(a)** Schematic of mass spectrometry barcoding workflow. Orthogonal barcode sets are designed *in silico* based on predicted LC-MS/MS performance and biophysical properties. Designed protein-barcode pairs are synthesized as an oligo pool for proteins up to 70 aa (154 aa with oligo assembly). The protein library is then expressed and fractionated as a single sample. Barcodes are purified from each fraction and measured by orbitrap LC-MS/MS. Schematics created with BioRender. **(b)** Protein enrichments are estimated from normalized barcode intensity, which requires efficient discrimination of barcodes. LC-MS/MS permits separation of barcodes by hydrophobicity (which influences iRT) and by parent ion mass-to-charge ratios (m/z). Barcode libraries were designed to traverse the optimal m/z range for DDA/DIA Orbitrap methods (450-1000 m/z) while spanning the space of iRT. **(c)** Accuracy of barcode elution from the LC column was measured empirically by comparing the predicted iRT (via Prosit) and the experimentally measured iRT (via DIA-NN) for the v2 barcoding set tagged to muGFP. A difference of < 0.5 iRT between actual and predicted iRT was observed for 95% (grey dotted lines), demonstrating high accuracy for *in silico* prediction of barcode LC retention. **(d)** An improved barcode detection rate was achieved across benchmarking muGFP barcoding experiments by 1) switching from DDA to DIA, 2) employing columns packed with finer silica (1.9 µm > 5 µM), and 3) restricting the amino acid composition of barcodes. **(e)** The frequency that a barcode is correctly mapped to a design (barcode purity) for different libraries. For the single oligo libraries, 99% of barcodes exclusively mapped to the intended design, and ~80% for 2-oligo designs (requiring pooled PCR homologous assembly).

3.2 – Screening beta barrel monomers

We applied our improved v1 mass spectrometry barcoding approach to a set of *de novo*-designed small beta barrels with six different barrel topologies, four of which were previously identified as protease-resistant via yeast surface display (Kim et al. 2023). Motivated by the observation that many of the designs were either insoluble or not monomeric when expressed individually (likely due to off-target strand pairing interactions), we set out to use size exclusion chromatography (SEC) to identify well-behaved designs from a set of individually pre-characterized controls (104 designs, not screened for protease resistance) and protease-resistant designs (416 designs) (Figure 4a).

An oligonucleotide library was synthesized with a median of 18 randomly assigned v1 barcodes per design (total of 520 designs; 9,080 barcodes, ranging from 9-18 barcodes per design) (Figure 4a, Supplementary Figure 2a). Deep sequencing the *E. coli* library prior to protein expression confirmed high sequence fidelity; 98.9% of reads with exact barcode sequences also contained the exact intended design sequence (Figure 3e, Supplementary Figure 2b). The NGS results indicate a low frequency of barcode-design mismapping, which validates barcode pre-assignment as a strategy to streamline the library preparation process.

Pooled expression, Ni-NTA purification, SEC fractionation, and barcode purification resulted in 21 samples which were analyzed by LC-MS/MS (Figure 4b). A total of 2804 barcodes were detected accounting for 504 designs, with all 2804 barcodes detected in at least 5 samples (median 5 barcodes per design) (Figure 4c, Supplementary Figure 2c & 2d). The distribution of barcodes detected per design was highly skewed and correlated with design quality measured by AlphaFold pLDDT, with 29% of designs accounting for 50% of detected barcodes (Figure 4c, Supplementary Figure 2e), in part due to insoluble designs not being represented in SEC.

We reconstructed a per-barcode elution profile using the relative abundance of barcodes across SEC fractions. After removing barcode elution profiles with fewer than 5 data points, we defined a consensus elution profile for each design, normalized to the column volume (Figure 4d, Methods). In most cases, individual barcodes for the same design showed highly similar elution profiles, with elution peaks of individual barcodes having <10% coefficient of variation for 81% of designs, suggesting that most barcodes are non-perturbative to design behavior. (Supplementary Figure 2f). We took the fraction of the elution profile within the expected range for a protein aggregate as an aggregate score; this led to a well-defined separation between control designs previously reported to be monomer, dimer, or aggregate based on individual SEC characterization (Figure 4d; all designs were intended to be monomeric, so those that formed dimers are also more likely to have aggregate species). A subset of designs exhibited high variation among barcodes (Supplementary Figure 3); these may be marginally stable designs and hence more perturbed by barcode attachment than stably-folded designs.

To determine if pooled MS barcoding can identify successful designs, we selected 19 designs for individual characterization that showed consensus elution peaks in the expected range for monomers or dimers and low variation among barcode elution profiles (Figure 4e). For 16 designs we observed high soluble expression, with SEC elution peaks correlated strongly between pooled and individual expression (Figure 4d, Supplementary Figure 4). Notably, 16/19 (84%) screen hits designed using the SH3, barrel6, or barrel5, and OB beta barrel topologies eluted within 1 mL of the expected monomer elution volumes, whereas these designs showed a success rate of only 13/24 when selected solely based on protease resistance (Kim, et al. 2023). Consistent with these results, we found minimal correlation ($r^2 = 0.08$) between protease stability score and either pooled SEC elution peak or number of barcodes detected (Figure 4e). Further structural characterization is described in Kim et al 2023.

We observed that success rates for monomeric design varied across six fold families, four native (1) oligosaccharide-binding – OB, 2) Src homology 3 – SH3, and two of SH3 subclasses: 3) small barrel – sm (Youkharibache et al. 2019), and 4) Tudor domains (Lasko 2010), and two *de novo* scaffolds, 5) b5 and 6) b6, previously described (Lasko 2010; Kim et al. 2023). Of these six families, barrels with sm folds had a higher monomer success rate than all other families (Figure 4f). On the other hand, non-Tudor/non-sm SH3, OB, b5, and b6 barrels tended to adopt higher order oligomers. Many of the Tudor barrels eluted as dimers, consistent with previous observations that Tudor domains can form homodimers (Tong et al. 2015; Cui et al. 2012; Zhao et al. 2007).

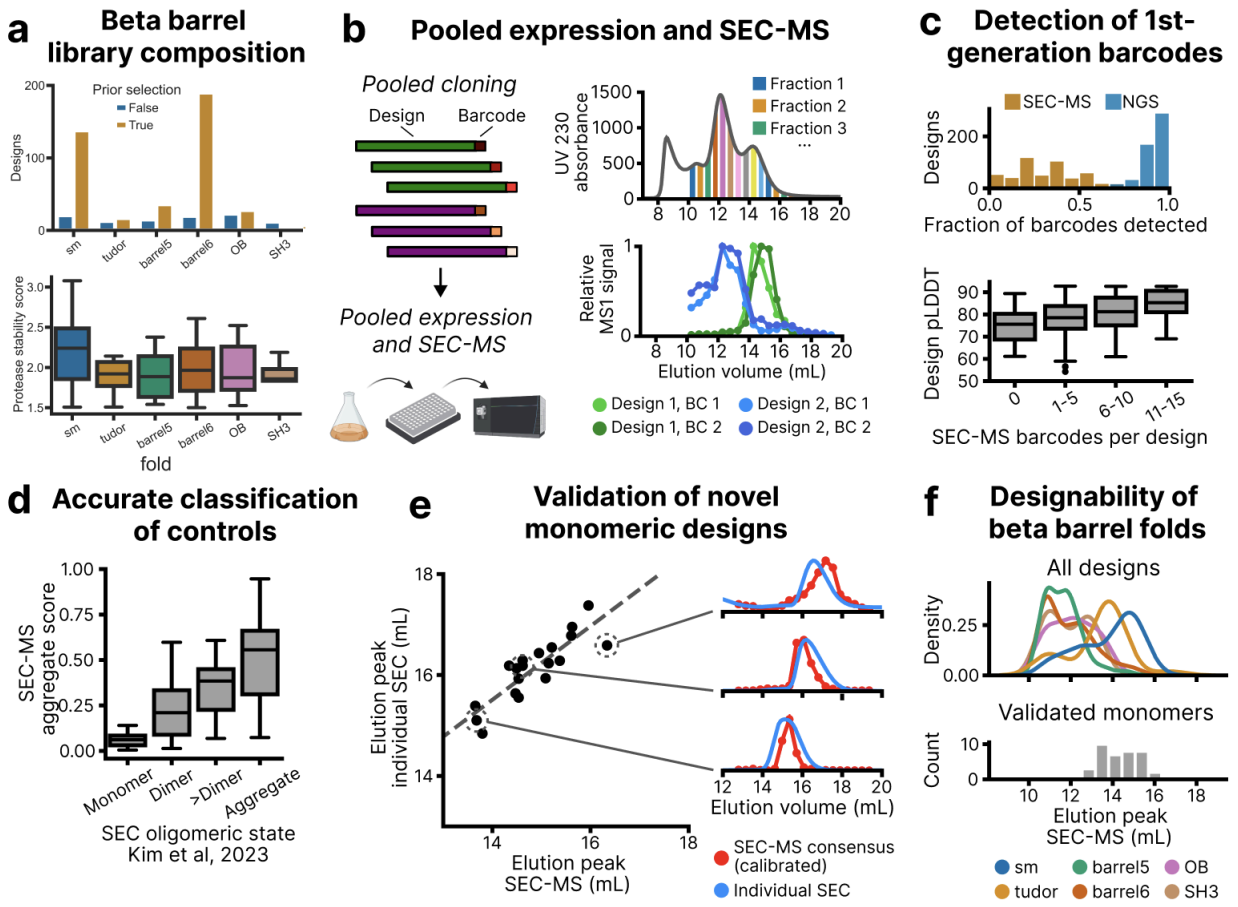


Figure 4: **Evaluation of barcoding approach on library of designed monomer beta barrels. (a)** Library distribution. Most designs were pre-selected for protease stability via yeast display (true and false in panel legend). An additional 104 controls were chosen solely by *in silico* metrics and individually characterized separately. **(b)** Designs and cognate barcodes were co-synthesized on oligonucleotide arrays. The library was cloned, expressed, and purified as a single pooled sample. After fractionation by SEC (top), barcodes were purified and quantified from each fraction by LC-MS/MS. Relative barcode abundance across fractions was used to reconstruct SEC traces (bottom). **(c)** Most barcodes were detected via next-generation sequencing (NGS), whereas a substantial fraction of barcodes were not detected in SEC-MS analysis (top). Barcode detection rate in SEC-MS was correlated with design quality measured by AlphaFold pLDDT (bottom). **(d)** For control designs individually characterized in a prior publication, aggregate score (fraction of consensus eluting between 10 and 12 mL) was predictive of off-target oligomeric state (all beta barrels were designed to be monomeric). **(e)** Novel monomeric designs were individually expressed without barcodes and characterized by SEC. Elution peaks agreed well between the pooled SEC-MS and individual SEC, up to a fixed offset (dashed line; added as a fixed global offset to SEC-MS consensus in examples on the right). **(f)** Comparison of designs grouped by fold (top) to validated monomers used as internal library controls (bottom) identified highly designable beta barrel folds.

3.3 – Screening hallucinated oligomers

Encouraged by the ability to identify successful monomeric designs, we next applied pooled MS barcoding to a ~10x larger library of 4,495 oligomers. The assembly of oligomers in solution is not well recapitulated in yeast display, so library-scale methods have not been applied to designed oligomers to date. For this design library, we expected successful designs to form highly stable complexes due to their large oligomeric interfaces. These complexes first assemble in the clonal environment of individual cells, and if stable should be maintained through cell lysis, pooled purification, and pooled SEC (Figure 5a). We selected oligomers with a protomer size comparable to the beta barrel library (65 amino acids) designed via a recently described hallucination method (Wicky et al. 2022). Of note, most (but not all) of the designs exhibit cyclic symmetry, so we opt for using C-notation herein to describe the number of subunits.

A library of 4,495 designs was synthesized, including 82 controls from the beta barrel monomer library, 127 hallucinated controls published with detailed experimental characterization, and 4,286 new hallucinated designs (Figure 5a, Supplementary Figure 5a), ranging from C1 to C7 (22,475 barcodes, 5 barcodes per design). Sequencing quality control, pooled expression, purification, SEC, and MS barcode analysis were performed as for the beta barrel monomer library (Supplementary Figure 5), except that two separate SEC-MS runs were performed with S75 and S200 columns to capture the full range of elution volumes. After filtering for barcodes detected in at least 5 SEC samples, we obtained 5,200 barcodes representing 2,909 designs (23.14% barcodes representing 63.11% of ordered designs, Supplementary Figure 5). Of these 2,909 designs, 1,130 (39.81% of the library) had peak elution volume coefficients of variation <10% (Supplementary Figure 5e & 5j).

A calibration curve measuring elution volumes of proteins previously characterized by SEC, SEC-MALS, and/or crystallography was constructed to predict the elution volume of library members as a function of molecular weight and hydrodynamic radius (Figure 5b). The filtered dataset was then categorized based on whether the observed SEC-MS elution volume was within a bootstrapped 95% confidence interval, corresponding to the fraction with the highest MS1 area. Out of the 2,909 designs, 43% had elution volumes that were predicted by the calibration curve (Figure 5b, Supplementary Figure 6a). We defined success at the level of designs based on agreement between observed and predicted elution volumes, and at the level of individual backbones based on whether the backbone gave rise to at least one successful design. Compared to monomer design, which might fail due to 1) lack of soluble expression or 2) aggregation, the design of higher symmetries is further complicated by two additional failure modes: 3) lack of assembly or 4) formation of off-target assemblies. At the SEC stage, we observed aggregation failure modes for all cyclic oligomers: for C1 designs (designed monomers), an asymmetric bimodal distribution indicated that most C1 designs eluted at either the expected elution volume or in the void as an aggregate (Supplementary Figure 6a). For C2+ designs, we observed trimodal distributions, with densities observed a) where C1 designs elute, b) at the expected elution volume, and c) in the void where aggregates elute (Supplementary Figure 6). Higher order symmetries C4+ with MWs approaching 40 kDa had major peaks that were poorly resolved from the void peak on the S75 column, so we analyzed the elution profiles for these symmetries on an S200, which has greater resolution in the 30-100 kDa range (Figure 5c). Here, C4+ symmetries were better differentiated from the void volume, and we observed broad overlapping peaks for designed C4-6 oligomers. Given the oligomers have similar molecular weights and diameters, this result suggests a multiplicity of design states for these oligomers. The difference in closing angle between C3, C4, and C5 is greater than the difference in closing angle between C6, C7, and C8, and hence obtaining specificity for the intended oligomerization state is more difficult for higher cyclic symmetries (Edman et al. 2023). Indeed, within each

oligomer class, the success rate decreased as symmetry number increased, from 64.62% for C1s (n = 506) to 34.00% for C5 (n = 100) (Figure 5d, Figure 5e, Supplementary Figure 6c; C6 was an exception). The AF2 prediction metrics pAE and pTM (which were not used to curate the input design set) were not significantly different between successful and unsuccessful designs (Figure 5e), but the unsuccessful designs were overall somewhat less compact than backbones with high success rates (Figure 5e).

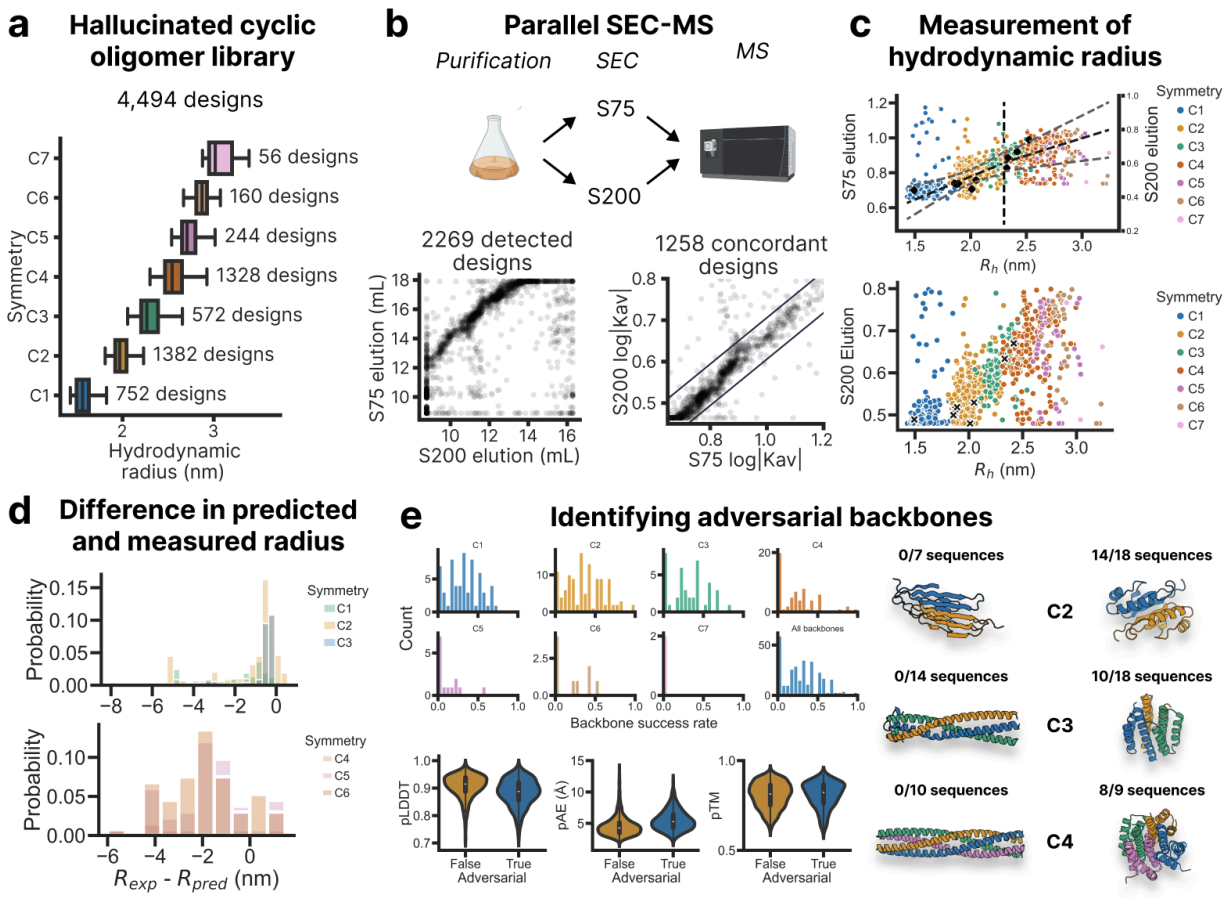


Figure 5: **Characterization of small cyclic oligomers.** **(a)** Cyclic homo-oligomers were designed by AlphaFold hallucination of backbone geometry with a symmetric loss function followed by sequence design with ProteinMPNN (top). Up to 18 sequences were designed per backbone in order to investigate the contribution of backbone quality to design success (bottom). **(b)** The library was fractionated using S75 and S200 SEC columns in order to capture a wide range of sizes. Sizing across the two runs was consistent for a majority of detected designs (bottom left). Designs were declared concordant if their normalized retention time (Kav) followed the expected linear relationship (bottom right). **(c)** For many concordant designs, rescaled Kav was proportional to hydrodynamic radius predicted from the design model. Shown in black are controls individually validated by SEC in published work. **(d)** Difference between the experimental and predicted radius of concordant designs for different symmetries. **(e)** A per-backbone failure rate was defined as the ratio of concordant designs to designs detected by NGS for each parent backbone (top). Adversarial backbones, defined as backbones with failure rate >0.85, had AlphaFold metrics indistinguishable from non-adversarial backbones (bottom) despite dramatically different success rates (examples on right).

3.4 – Screening large helical oligomers

We sought to apply pooled MS barcoding to larger proteins by using oligo assembly to circumvent the length limits of pooled oligonucleotide synthesis. Commercially available oligo pools have a maximum length of approximately 300 nt, which constrains barcoded designs to ≤74 aa to permit padding with subpools specific adapter sequences. Pairwise oligo assembly enables barcoding of much larger designs (≤154 aa), at the cost of decreased sequence representation due to differences in assembly efficiency among designs, and decreased sequence fidelity due to chimeric off-target assemblies. Chimeric assemblies are particularly concerning for peptide-based mass spectrometry, which relies on short sub-sequences as proxies for full-length proteins. To mitigate these issues, we developed a sequence design pipeline that minimized DNA homology among designs and used the same reverse translation for all barcoded variants of the same design, so assembly complexity scales with the number of designs but not the number of barcodes (Figure 6a).

To test pooled MS barcoding of larger proteins, we barcoded a set of 5,068 homo-oligomers designed from curved helical repeats, ranging in symmetry from C2 to C6 (26,805 barcodes, minimum 5 barcodes per design, design length 119-156 aa). Deep sequencing after *E. coli* transformation showed 77.8% of barcodes covering 92.8% of all designs, with a median of 5 barcodes detected per design (Supplementary Figure 7). Of all barcoded designs, 99.4% of barcodes correctly and uniquely mapped to the design of interest, determining that the intended barcode pre-assignment was maintained throughout pooled DNA assembly. Pooled expression, purification, SEC, and MS barcode analysis were performed as for the beta barrel monomer library, except that SEC conditions were adjusted for an Superdex 200 10/300 GL column to optimize separation of oligomers. After filtering for barcodes detected in at least 5 SEC samples,

we obtained 8311 barcodes (3060 designs, 60.38% of all designs), 52.1% of which had peak coefficients of variation < 10%.

We selected 28 designs with ≥2 barcodes detected, low peak coefficient of variation (<10%) among barcodes, and elution peaks within the expected range for their given symmetry and radius for individual characterization (Figure 6b). 26/28 designs showed high soluble expression, and SEC elution peaks correlated strongly between pooled and individual expression (Supplementary Figure 7). Overall, 27 (96.1%) of these designs eluted within the 95% confidence interval of the standard curve calibrated on hydrodynamic radius, and 15 of these designs (57.7%) eluted within 1 mL of the expected elution volume (Supplementary Figure 8). We selected one of these successful oligomers, sg266, to carry forward into crystallographic studies. We determined sg266 to adopt the modeled trimeric state by x-ray crystallography with an RMSD of 1.36Å between design model and the solved structure (Figure 6c). Additional structural characterization for these designs are published in (Gerben et al. 2023).

We further increased the size range to ~215 kDa in a SEC-MS screen of one-component I3 nanoparticles interfaces with constant scaffolds (n=1173 cages, with 4 barcodes per design). The library was designed to interrogate interface metrics for assembly of icosahedral particles with identical trimer frameworks and fully redesigned interfaces. This library was subjected to pooled expression, purification, and SEC, and subsequent MS barcode isolation and analysis from SEC fractions were performed to analyze barcode traces. Filtering for barcodes detected in at least 5 SEC samples, we obtained 2434 barcodes mapping to 1090 designs with 643 (56.4%) having barcode peak elution volume coefficients of variation <10% (Figure 6d). For validation, 12 designs demonstrating assembly in the library format were individually expressed alongside aggregated and unassembled designs (Supplementary Figure 10). One of these hits, I3-08, was taken forward for structural characterization (Figure 6e), which revealed it to be an icosahedron

of the on-target state. Upon expression of individual clones at the 50mL scale, cages appeared to adopt multiple oligomerization states on SEC, with peak elution volumes agreeing with peak values observed on SEC-MS, suggesting the desired assembly state is most abundant. We reasoned that these nanoparticles have a propensity to aggregate at the high concentrations expected at clonal 50mL scale, but not observed at the SEC-MS 50mL expression scale, which contains a heterogeneous pool of mixed expression. Nonetheless, the predominance of a peak at the expected elution volume corroborates the utility of this assay in large nanocage assemblies en masse.

Collectively, these data demonstrate that the oligomerization states of thousands of protein assemblies ranging in size from 10 kDa to 200+ kDa can be resolved in a single SEC run.

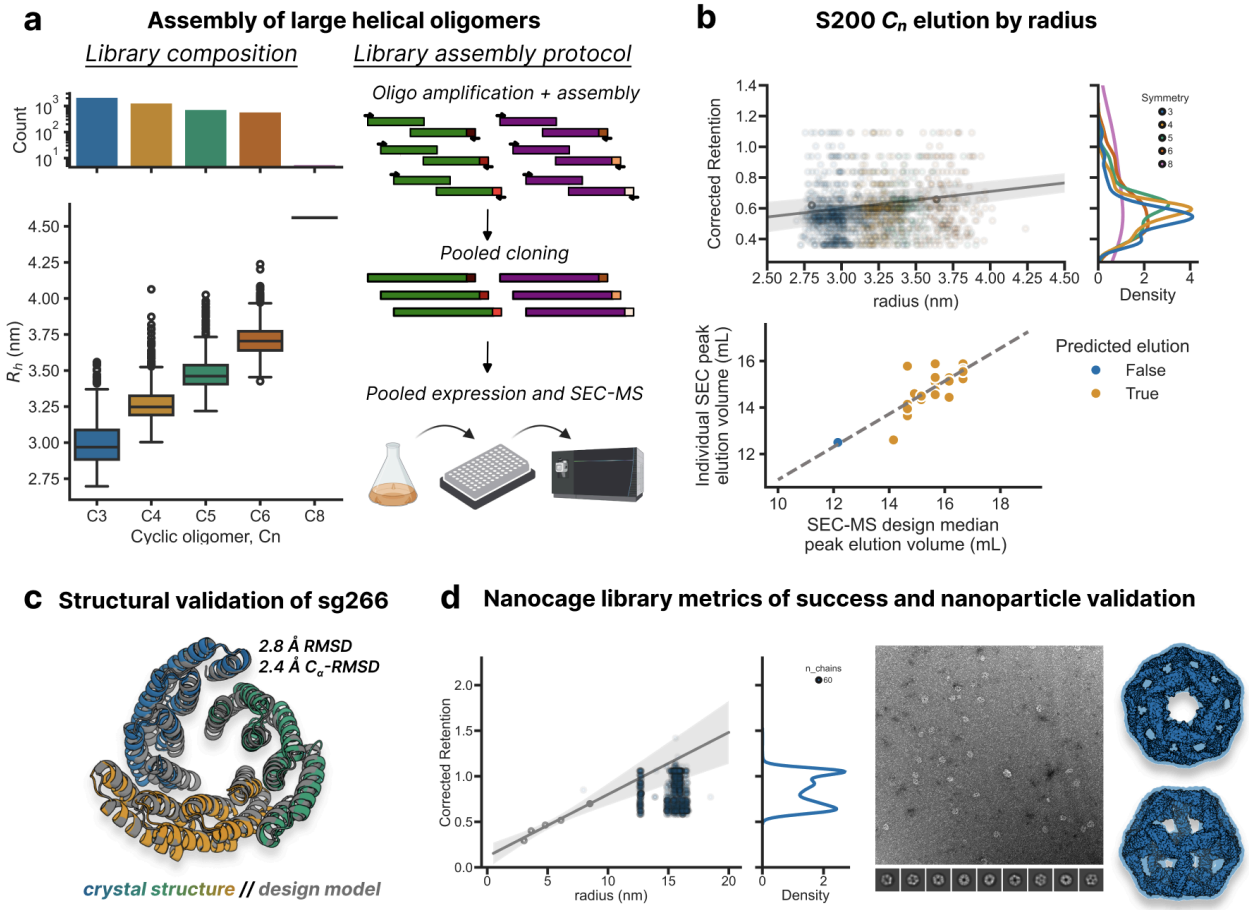


Figure 6: **Characterization of libraries of large oligomer designs. (a)** A library of designed large cyclic helical oligomers (>74 aa) of varying hydrodynamic radius with symmetries ranging from C3-C8 was assigned barcodes prior to chip-based DNA oligo synthesis (left). Owing to their large size, DNA oligos were subjected to assembly in pooled format, requiring a unique homologous sequence between part A (5' fragment) and part B (3' fragment) oligos to minimize formation of chimera sequences with mis-assigned barcodes. **(b)** Cyclic symmetries were fractionated on an S200, which permitted separation across the range of assembly states (top). Designs elected for follow-up were clonally expressed and purified at 50mL scale without barcodes, and their peak elution volumes on an S200 (bottom, y-axis) were compared to peak volumes from pooled expression (bottom, x-axis). **(c)** Crystal structure of sg266 (blue, green, and orange), a helical trimer hit from the screen, overlain with its design model. (d) Pooled elution profile of the I3 nanocage library (left). Negative stain electron micrograph (middle) and cryo-EM structure (right) of a hit from the screen, I3-08.

3.5 – Evaluation of v2 barcode library on tetrahedral nanoparticles

To evaluate the utility of the v2 barcode library, we set out to screen a library of putative *de novo* tetrahedra (n = 1,187 designs of 80 aa each, 7 barcodes per design, 8309 total barcoded designs) for homogenous nanoparticle assembly (Figure 7a). RFDiffusion has been successful in generating novel folds and topologies (Dauparas et al. 2022; Watson et al. 2023), and deep learning based sequence design has enabled the design of highly soluble proteins (Dauparas et al. 2022). While a diverse array of structures can be generated with these new design methods, robust nanoparticles are challenging to identify because these multi-component assemblies require reversible assembly of multiple weak interfaces to ensure particle homogeneity (Wargacki et al. 2021). As such, we aimed to use our higher fidelity barcoding set to identify novel, highly soluble tetrahedral nanoparticles.

From two ssDNA oligos per barcoded design, we used pairwise oligo assembly to construct full-length dsDNA encoding each design and a unique C-terminal barcode prior to cloning into a pET29b(+) plasmid encoding a C-terminal 6x-His tag. The library was then transformed, expressed, and purified prior to injection onto an S200 Increase 10/300 GL, as previously described for the large helical oligomers. NGS at 169x coverage identified 7429 correctly mapped barcodes (89.3% of all ordered barcodes) representing 1141 designs (96.1% of all ordered designs, Supplementary Figure 11). Of all detected barcodes on NGS, 80.9% exclusively mapped to the intended design (Figure 3E, Supplementary Figure 11). Barcodes were purified from 250μL SEC fractions and spectra were acquired with the optimized DIA LC-MS/MS acquisition protocol (Methods). Corroborating the results from the muGFP optimization study, 7222 / 7421 (97.3%) of NGS-verified barcodes were detected pre-filtering, and 98.3% of designs had ≥5

barcodes detected per design, with 39.4% of designs accounting for 50% of all detected barcodes (Supplementary Figure 12).

Peak elution volumes for all barcodes for a given design displayed low overall variation, with 81.3% of designs demonstrating CV < 10%, validating that the v2 barcode set is mostly well tolerated. Using a standard curve constructed from peak elution volumes of globular standards (Methods), 14 hits with peak elution volume CV < 10% that fell within the 95% CI of the standard curve for radius and/or MW (Methods, Supplementary Figure 13) were selected for follow-up expression and purification at 50mL scale with non-barcoded clones. Of these 14 designs, 13/14 expressed and 9/14 were monodisperse by SEC. Fractions at the expected nanoparticle elution volume were collected and subjected to complex sizing by mass photometry and hydrodynamic diameter measurement by dynamic light scattering (Figure 7c). Of the 9 hits, 8/9 had peaks corresponding to the expected molecular weight of a 12-mer assembly, and 8/9 had predominant peaks on DLS at the expected hydrodynamic diameter. In total, 3 of the 9 had monodisperse peaks by SEC, mass photometry (which exhibits bias towards incomplete assemblies), and DLS (which exhibits towards aggregates, respectively), suggesting robust assemblies with no significant off-target or fractional assemblies (Figure 7c, Supplementary Figure 13, 14). Of the 9 hits, 4/9 diffracting structures were solved to ≤3 Å resolution (Figure 7c). The structures were very similar to the design models for each of these designs with ≤ 3 Å all-atom root-mean-square deviation to the design model across all 12 chains, corroborating the design hypotheses for these cages.

Overall, the v2 barcoding set increases the barcode detection rate from 30% to 90% which greatly enables identification of rare complex assemblies in a diverse structural pool.

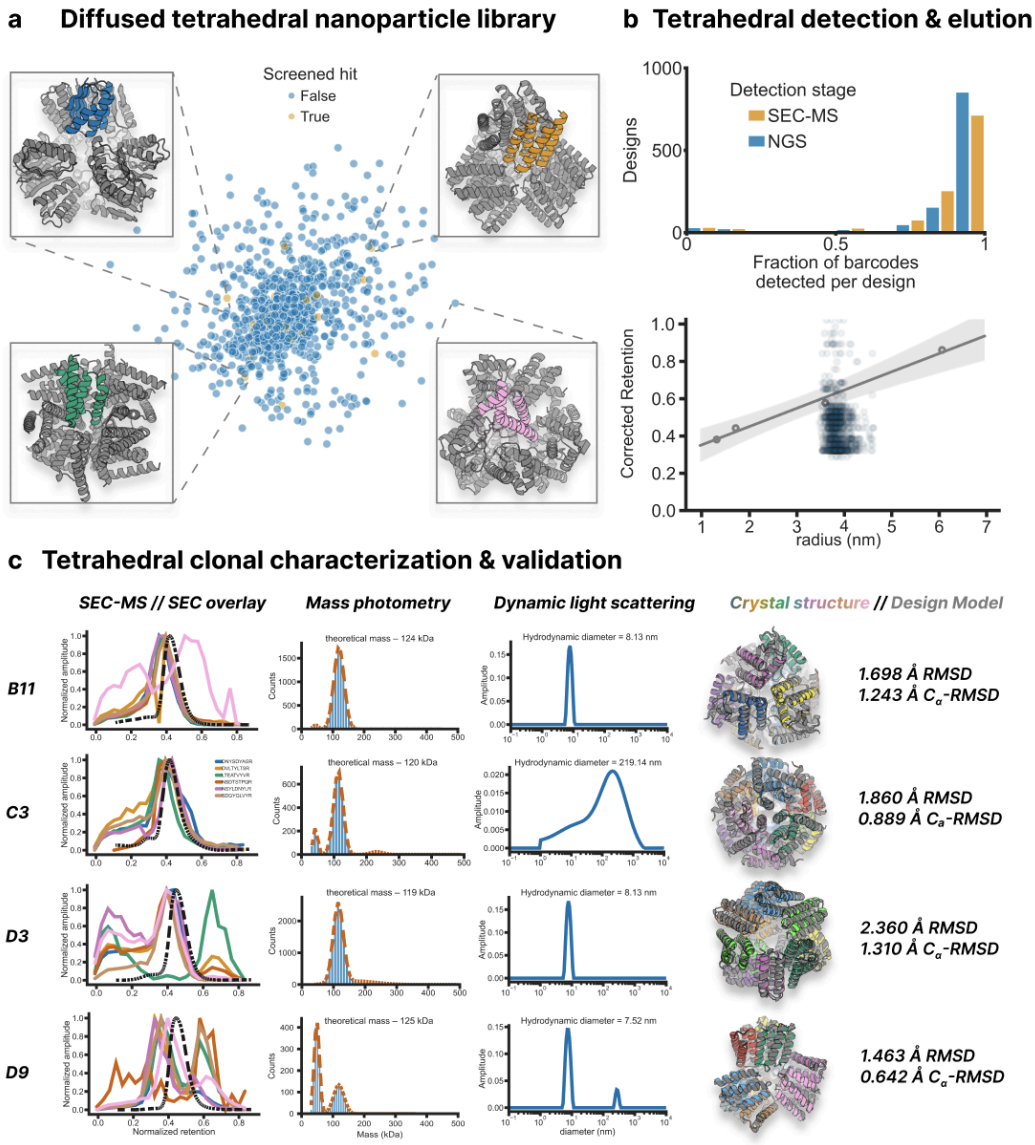


Figure 7: **Tetrahedral nanoparticle characterization using 2nd generation barcodes. (a)** A library of one-component tetrahedral backbones were generated using RFDiffusion and ProteinMPNN was used to design backbones. Sequences were then filtered by AlphaFold2 (pLDDT > 90) to predict the structure of the protomer (in color). Protomers were assessed for similarity by TM-score against all other library members. Multidimensional scaling of the all-by-all TM-score was plotted in 2D space. **(b)** Each library member was assigned 7 barcodes from the v2 barcoding set and ordered as a C-terminal genetic fusion as a single DNA oligo pool (n=8309 oligos). The DNA library was amplified, assembled, and transformed before being subjected to NGS and SEC-MS. Barcodes representing all 1,187 designs were detected at a comparable rate to NGS (top, median of 7, mean of 6.27 barcodes per design). The library was fractionated with an S200 SEC column in order to capture the expected range of assembly states. Designs were declared concordant if their corrected retention time ($\sqrt{(-\log_{10}(Kav))}$ followed the expected linear relationship (bottom). **(c)** Of the 13 hits identified by SEC-MS, four resulted in crystal structures. These four designs demonstrate concordance between the SEC-MS (color, solid), peak elution volumes and the clonally expressed (black, dashed) SEC trace at 50 mL scale (far left). Mass photometry (middle left) and dynamic light scattering (middle right) provide finer grained insight into the multiplicity of assembly states for C3 and D9 that is not evident by SEC. Crystal structures (far right) were solved to ≤3 Å resolution that corroborate the design model hypothesis.

3.6 – MS barcoding discussion

We were able to substantially increase the power of the peptide barcoding coupled to mass spectrometry approach introduced by Egloff, Seeger, and coworkers. Egloff, et al. used random peptide barcodes stochastically linked to a nanobody library derived from llama immunization to measure solution-phase properties. We extend this work in two key ways. First, we leveraged pooled oligo synthesis and assembly to generate libraries directly from *in silico* sequences up to 154 aa in length. We streamlined library construction by a) omitting cloning of the binder library into *E. coli* for pre-enrichment and b) attaching barcodes at the oligo level to circumvent barcode mapping via next-generation sequencing. Second, we explicitly designed barcode peptides to optimize mass spectrometry based detection and to minimize perturbation to attached proteins. Our approach should enable scaling the number of barcodes to up to 120,000 to greatly increase the number of proteins that can be analyzed simultaneously. Moving forward, recent advances in deep learning-based sequence design now enable elimination of possible interference from appended barcodes by embedding the barcode directly in the coding sequence.

Our mass spectrometry-barcoding approach enables large-scale testing of designed protein properties inaccessible to previous screening methods based on surface display and proxy reporters. Using this approach, we assessed the solubility and oligomerization state of thousands of computational designed proteins in parallel. Our results across diverse folds – beta barrels, hallucinated cyclic oligomers, large helical cyclic oligomers, and protein nanoparticles – suggest that despite potential liabilities of pooling– cross-design interaction and low concentration of each design–designs with the desired properties (intended oligomerization state, for example) can be robustly identified. Data generated by large-scale testing can provide feedback on the design process, exemplified by our analysis of systematic failure modes in cyclic

oligomers generated by hallucination and deep learning-based sequence design. While we focused here on solubility and size-separation, pooled MS barcoding generalizes to virtually any protein property amenable to solution-phase fractionation, such as binding, folding stability, and permeability across biological barriers.

3.7 – MS barcoding materials & methods

N.B. – all buffers prepared in Milli-Q-grade distilled deionized water (ddH_2O) unless otherwise specified

Single oligo library cloning

Subpools were amplified from an oligonucleotide pool (Twist) with subpool-specific primers using qPCR with KAPA HiFi HotStart ReadyMix (Roche) and EvaGreen Dye (Biotium). To reduce template switching that could scramble the design-barcode linkage, a minimal PCR cycle count was determined for each sample via a test reaction. Amplified oligos were size-selected with a 2% agarose E-gel EX (ThermoFisher), quantified using a Qubit dsDNA Quantification Assay (Thermo), and cloned into a modified pET28b plasmid using NEBridge Golden Gate Assembly Kit (BsaI-HF v2) (New England Biolabs). Approximately 100 ng of assembled plasmid was transformed into *E. cloni* Express BL21(DE3) Electrocompetent Cells (Lucigen). After recovery for 1h at 37°C, cells were transferred to a 50 mL LB overnight culture with kanamycin selection, from which glycerol stocks were prepared. Transformation efficiency was assessed by counting CFUs from a 1:50,000 dilution of the recovery culture on LB-Kan100 plates (Teknova).

Primer sequences:

>DF_Nterm_pT05_fwd_1

CTACTGGTCTCaCAGTCGA

>DF_Cterm_pT05_rvr_1

ACTGGAGACGGTCTCaGTTA

>DF_Cterm_pT05_fwd_2

GACACGGTCTCtCAGTCGA

>DF_Cterm_pT05_rvr2

ACGATTCTGGGTCTCtGTTA

>DF_Cterm_pT09_fwd_1

GCTCTCGGTCTCgTACCATG

>DF_Cterm_pT09_rvr_1

AGATGGACTGGTCTCgTGCG

>DF_Cterm_pT09_fwd_2

CTATCATTCGGTCTCcTACC

>DF_Cterm_pT09_rvr_2

AGTCTAATTGGTCTCcTGCG

>oDF-287_foldit_fwd_0_NGS_tP5

TCCCTACACGACGCTCTTCCGATCTctactggtctcacagtcga

>oDF-288_foldit_rev_0_NGS_tP7

GTTCAGACGTGTGCTCTTCCGATCTactggagacggtctcagtta

>oDF-289_foldit_fwd_1_NGS_tP5

TCCCTACACGACGCTCTTCCGATCTgacacggtctctcagtcga

>oDF-290_foldit_rev_1_NGS_tP7

GTTCAGACGTGTGCTCTTCCGATCTacgattctgggtctctgtta

>DF_pT15_1_short_fwd

GGGTCTAGggtctcaAGGA

>DF_pT15_1_short_rvr

AGTACTCGggtctcaCGCT

Adapter sequences:

Adapter sequences are terminal sequences that flank the DNA of the protein of interest. They are used to selectively amplify subpools from pooled single stranded oligo DNA. Their order on a sequence is:

5' end adapter – coding sequence of protein of interest – 3' end adapter.

Adapters for beta barrels (pair with DF_pT09 primers):

>DF_Cterm_pT09_1

5' end – GCTCTCGGTCTCgTACCATG

3' end – CGCAcGAGACCAGTCCATCT

Adapters for small cyclic oligomers (pair with DF_pT15_1_short primers:

>DF_pT15_1

5' end – TGCTTTGGGTCTAGggtctcaAGGA

3' end – AGCGtgagaccCGAGTACTTCTGGT

>DF_pT15_2

5'end – GACATGATCTAGAGggtctcaAGGA

3’ end – AGCGtgagaccTCCGACCATTCTTT

Adapters for large helical oligomers (pair with):

Adapters for nanocages (pair with):

>jason_01_A

5’ end – GCGACTAGGGGTATGCTG

3’ end – AAgcttacgggcaacATGG

>jason_01_B

5’end –GCGAcccactggcataaTT

3’ end – gcctgtgcgaaacATGG

>jason_02_A

5’ end – GCGAcgtgaccaccaagg

3’ end – AAcccagtaagggtcATGG

>jason_02_B

5’ end – GCGAgcgaccttagagtTT

3’ end – cagagaggtcagcATGG

>jason_03_A

5’ end – GCGAagtcccttaccctt

3’ end –AAgctcccgtatcagATGG

>jason_03_B

5’ end – GCGAgagggtctacagaTT

3’ end – ctcggtccacgatATGG

Two-oligo assembly library cloning

The previously described protocol for two-oligo assembly (Klein et al. 2016) mirrored the single oligo library construction protocol with two exceptions. First, chip oligo ssDNA for 5’ and 3’ fragments was amplified with KAPA HiFi HotStart Uracil+ReadyMix (Roche) along with uracil-containing adapters on the inner termini (3’ end of the 5’ fragment and 5’ end of the 3’ fragment). After gel extraction, fragments were resuspended in 20µL ddH_2O and incubated with 2 µL USER Enzyme (New England BioLabs) in a thermocycler at 37°C for 15 minutes followed by 22°C for 15 minutes. After cooling, the entire USER digest mixture was mixed with 10 µL 10x NEBNext End Repair Enzyme Mix (New England BioLabs), 5 µL NEBNext End Repair Reaction Buffer (New England BioLabs), and 63 µL ddH_2O for incubation on a thermocycler at 20°C for 30 minutes. DNA was then cleaned up and prepared for qPCR amplification with the primers to the outer adapters (5’ end of the 5’ fragment and 3’ end of the 3’ fragment) and KAPA HiFi Hotstart *without uracil*. The rest of the protocol is reflected in the single oligo library construction protocol, where the amplicon running at the combined length of the 5’ + 3’ fragments is extracted, cleaned up, and cloned into pET28b(+).

Next Generation Sequencing

Plasmid DNA was extracted from overnight cultures using a ZymoPURE II Plasmid Midiprep Kit (Zymogen). The cloned insert containing design and barcode was PCR-amplified and size-selected using the same protocol as for library cloning. Sample barcodes and Illumina

sequencing adapters were added in a second qPCR, and paired-end 2x300 nt reads were acquired on MiSeq (lllumina).

Protein purification and size-exclusion chromatography (SEC)

Glycerol stocks (100µL) were used to inoculate 50mL Studier's Autoinduction media (Studier 2014) in 250 mL baffled flasks. Cultures were grown for 20h at 37 °C, and cell pellets were harvested via centrifugation (10 min, 4000 x *g*). Pellets were resuspended in 25 mL lysis buffer (25 mM Tris HCl pH 8, 300 mM NaCl, 40 mM Imidazole, 1 mM DNase I, 10 µg /mL lysozyme) and lysed by sonication (5 min, 85% amplitude, 15s on/off cycles). Lysate was clarified by centrifugation (14,000 x *g*, 20 min.). The pellet was resuspended in 1 mL of SEC running buffer (50 mM Tris HCl pH 8, 150 mM NaCl) to create the "insoluble fraction" sample, while the supernatant was defined as the "soluble fraction" sample. Ni-NTA resin (Qiagen) was used to isolate His-tagged proteins from the soluble fraction. Specifically, wash buffer (25 mM Tris HCl pH 8, 300 mM NaCl, 40 mM imidazole) and then elution buffer (25 mM Tris HCl pH 8, 150 mM NaCl, 400 mM imidazole) were applied to the resin. Libraries were eluted in 2 mL, of which 100 µL was defined as the "injection" sample, and 1 mL was used for SEC. SEC was performed on an equilibrated S75 Increase 10/300 GL or S200 Increase 10/300 GL at a flow rate of 0.8 mL/min in SEC running buffer. Fractions were collected in 0.25 mL intervals from 8 mL to 20 mL (Wicky et al. 2022).

Crystal Structure Determination

Crystals were produced using the sitting drop vapor diffusion method. Drops with volumes of 200 nL in ratios of 1:1, 2:1, and 1:2 (protein:crystallization) were set up in 96-well plates at 20 °C, using the Mosquito from SPT Labtech. Drops were monitored using the JANSi UVEX imaging system.

For sg266, diffraction-quality crystals appeared in a mixture of 0.2 M Ammonium sulfate, 0.1 M HEPES pH 7.5, 25% w/v Polyethylene glycol 3,350.

Diffraction quality crystals appeared in 0.1 M Sodium chloride, 1.6 M ammonium sulfate and 0.1 M Sodium HEPES pH 7.5 for B11; 10 % v/v PEG 400, 0.05 M MES pH 6, 0.1 M Potassium chloride/ 2 mM Magnesium chloride hexahydrate for C3; 1.4 M Sodium malonate dibasic monohydrate pH 7.0 for D3 and 21 % w/v PEG 3350, 0.1 M MES pH 6.0 and 0.15 M Sodium chloride for D9.

Crystals were cryoprotected before being flash frozen in liquid nitrogen before being shipped for data collection at synchrotron. Data collection was performed with synchrotron radiation at the Advanced Light Source (ALS) on beamline 8.2.2/8.2.1 or NSLS2 on beam line FMX/AMX.

X-ray intensities and data reduction were evaluated and integrated using either XDS (Kabsch 2010) and merged and scaled using Pointless and Aimless in the CCP4 program suite (Winn et al. 2011). Structure determination and refinement starting phases were obtained by molecular replacement using Phaser (McCoy et al. 2007) using the design model for the structures. Following molecular replacement, the models were improved using Phenix autobuild (McCoy et al. 2007; Adams et al. 2010); efforts were made to reduce model bias by setting rebuild-in-place to false and using simulated annealing. Structures were refined in Phenix (Adams et al. 2010). Model building was performed using COOT (Emsley and Cowtan 2004). The final model was evaluated using MolProbity (Emsley and Cowtan 2004; Williams et al. 2018). Data collection and refinement statistics are available in Supplementary Table 1**.** Data deposition, atomic coordinates, and structure factors reported in this paper have been deposited in the PDB 8VEA.

Barcode design

Barcodes consisting of 8-12 variable amino acids were appended to the N or C-terminus of designed amino acid sequences prior to reverse translation and pooled, single stranded oligo DNA synthesis. Each sequence was tagged with multiple barcodes to average out barcode-specific perturbations. Barcodes were designed as KxxxxR, to enable tryptic digest would yield a doubly-charged precursor of the form xxxxR. Notably the xxxx regions omitted certain amino acids: F & W to minimize hydrophobicity, C & M to minimize barcode cross-reactivity, H, K, & R to avoid alternate charge states greater than +2 during positive mode ionization. In the v2 barcoding set, the number of D & E was restricted to be ≤ 2 to minimize the number of alternate charge states. Barcode sequences were selected to be orthogonal by high resolution LC-MS/MS, with a minimum chromatographic separation of 8 iRT units (predicted using Prosit (Gessulat et al. 2019), and a minimum MS1 mass-to-charge separation of 10 parts per million. Sequences containing designs and appended barcodes were reverse translated as in the yeast surface protease assay, except for specifying E. coli codon usage, and ordered as a 300 nt oligo pool (Twist Bioscience).

Barcode Isolation

For each sample, 100µL of sample was added to 100 µL of Lys-C buffer (8M urea, 100mM Tris HCl, pH 8) plus 1 µg of Endoproteinase LysC (New England Biolabs). Samples were incubated in a shaker at 37°C for 4-6 hours. To isolate His-tagged barcodes, 15 µL of Dynabeads™ His-Tag Isolation and Pulldown (Thermo) was added to each sample, incubated at 25°C for 5 minutes, and a magnetic rack was used to separate beads from supernatant. Beads for each sample were washed twice with 200 µL of barcode wash buffer (50 mM Tris HCl, pH 8, 150 mM NaCl, 0.1% Tween-20), followed by two 300µL washes with trypsin buffer (50 mM Tris HCl, pH 8). Beads were then resuspended in 50µL trypsin buffer plus 0.25 µg trypsin Trypsin-ultra (New England Biolabs) and incubated at 37°C overnight for at least 6 hours shaking at >1000 rpm.

Mass Spectrometry

NanoLC analytical columns were prepared with capillary with 75um inner diameter, cut to 40 cm with laser puller, and packed with ReproSil-Pur 120 C18Aq media 5µm or 1.9µm silica (Dr. Maisch) at 1000 psi to a final column length of 15-20 cm. Trap columns were prepared with capillary with 75um inner diameter and cured overnight for polymer frit formation. Trap columns were packed at 500 psi with ReproSil-Pur 120 C18Aq media 5µm silica (Dr. Maisch) to a final column length of 1-3 cm. Columns were equilibrated on an EASY-nLC 1200 (Thermo). LC-MS/MS was run using an EASY-nLC 1200 and either an Orbitrap Fusion Lumos Tribrid (Thermo) or an Orbitrap Exploris 480 (Thermo) Mass Spectrometer at the University of Washington Proteomics Resource. Samples were prepared 50% in 0.1% trifluoroacetic acid, and 8µL of 1:2 diluted sample was loaded onto the trap columns at 2.5 µL/min before separation at 300 nL/min on the analytical column with an 89 min gradient (Solvent A = ddH_2O with 0.1% formic acid, Solvent B = 80% acetonitrile with 0.1% formic acid, Gradient = 6-40% Solvent B). Peptides were ionized with nanospray ionization with a positive ion voltage of 2100 V at 300°C. Under the DDA protocol, MS data were acquired over 120 min with a cycle time of 3s utilizing the Thermo "Advanced peak determination" setting. MS1 scans were collected in profile with an Orbitrap resolution of 500,000 (480000 on the Exploris) and a precursor scan range of 450-900 m/z with the following parameters: RF lens – 30%, AGC target – Custom, Normalized AGC target – 175%, Polarity – positive. MS1 data went through MIPS (monoisotopic precursor selection) filtering for peptide peak determination, a charge state filter for 2-5 charge states, a dynamic exclusion filter (excludes precursor after n=1 times for 10 seconds, with a 10ppm low and high tolerance, and excludes isotopes) prior to MS2. Centroid MS2 data were acquired with isolation windows of 1.6 m/z and a normalized HCD collision energy of 27% at 15,000 resolution. MS1 scans collected under our optimized DIA protocol reduced MS1 resolution to 30,000.

Under the DIA protocol, samples are loaded at 2.5µL/min and run on a gradient at 400nL/min (Solvent A = ddH_2O with 0.1% formic acid, Solvent B = 80% acetonitrile with 0.1% formic acid, Gradient = 6-40% Solvent B) over 89m. Nanospray ionization with positive spray voltage of 2000V, internal mass calibration of Easy-IC, expected peak width of 30s and charge state of 2 for MS1 settings. MS1 resolution w/ resolution of 30000 in a scan range 450-1000, RF lens% = 40, data collected in centroid. Standard AGC target, "automatic" maximum injection time mode, 1 microscan per scan. MS2 has an isolation window of 6m/z without isolation offset with a resolution of 15000 acquired over 150-2000m/z. Collision energy uses normalized HCD collision energy of 27%. RF lens% = 50. AGC target set for 1000% normalized AGC target with "auto" maximum injection time. 1 microscan is performed, and data is collected in centroid. The DIA window starts from 493.4742 and collects 61 spectra with a 6 m/z window to an end mass of 850.6366 and is under N loop control.

In-silico design of tetrahedral nanoparticles

Tetrahedral nanoparticle backbones of 80AA per subunit were generated with RoseTTAFold Diffusion as previously described ((Watson et al. 2023)). Following diffusion, backbones were minimally downsampled, only filtering on external C-termini (for experimental purification) and inter-subunit contacts to ensure sufficient interface size. Backbones were then designed as homooligomers with ProteinMPNN at a sampling temperature of 0.1 and with 8 sequences per RFdiffusion-generated backbone (Dauparas, et al.). Candidate sequences were then predicted (asymmetric unit only) with AlphaFold2 (Jumper, et al.) and designs were filtered on pLDDT ≥ 85 and RMSD ≤ 1 Å, with 1,187 designs passing.

Mass photometry

Mass photometry measurements were collected as previously described (Pillai et al. 2023) with a TwoMP (Refeyn) mass photometer. Pooled fractions corresponding to cage assembly were

diluted 1:10,000 in SEC running buffer (50mM Tris, 300 mM NaCl, pH 8), and 10 µL dilute solution was applied to a gasket well for focusing and collection (1 minute) under a large field of view. Ratiometric values were converted to masses in kDa using known sizes of a 20 nM β-amylase standard in the same buffer (tetramer 224 kDa tetramer, 112 kDa dimer, and 56 kDa monomer). Expected masses were computed based on expressed protein sequence multiplied by 12, the number of subunits in a 1-component tetrahedron.

Dynamic light scattering

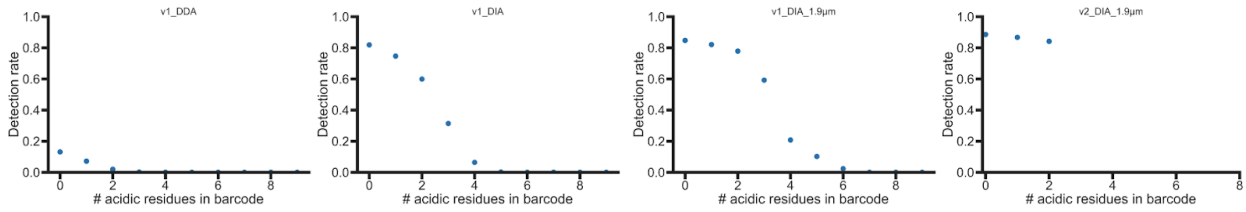
Measurements were performed with standard settings for polydispersity and sizing with an UNcle (Unchained Labs), as previously described (Yang et al. 2024). To a glass cuvette, 8.8 µL of sample ~1 mg/mL was applied, and data were collected at 25 °C with a 1s incubation time.

3.8 – MS barcoding acknowledgements

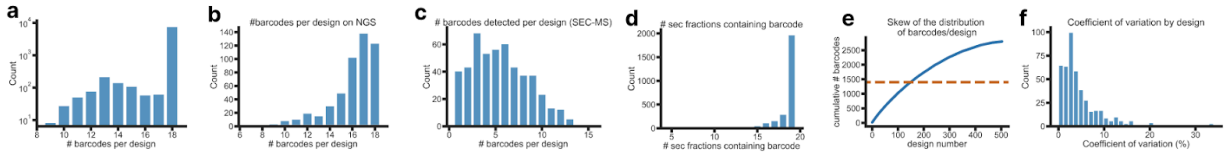
Many thanks to the community of colleagues, friends, and scientists who've assisted with this work. Specifically, we thank Robert Ragotte, Matthias Glögl, and Samuel J. Pellock for thoughtful discussions and critical feedback. We thank Kandise VanWormer, Rafael Ticzon, Hernan Nuñez-Ortega, Ratika Krishnamurty, Kristina Herrera, Lance Stewart, and the myriad support personnel at the Institute for Protein Design who have provided an enabling environment to carry out this work. Additionally, we want to thank the Advanced Light Source (ALS) beamline 8.2.2/8.2.2 at Lawrence Berkeley National Laboratory and National Synchrotron Light Source II for X-ray crystallography data collection. The Berkeley Center for Structural Biology is supported in part by the National Institutes of Health (NIH), National Institute of General Medical Sciences, and the Howard Hughes Medical Institute. The ALS is supported by the Director, Office of Science, Office of Basic Energy Sciences and US Department of Energy (DOE) (DE-AC02-05CH11231). At NSLSII the Center for Bio-Molecular Structure

(CBMS) is primarily supported by the NIH-NIGMS through a Center Core P30 Grant (P30GM133893), and by the DOE Office of Biological and Environmental Research (KP1607011). NSLS2 is a U.S.DOE Office of Science User Facility operated under Contract No. DE-SC0012704. This publication resulted from the data collected using the beamtime obtained through NECAT BAG proposal # 311950 and # 313951.

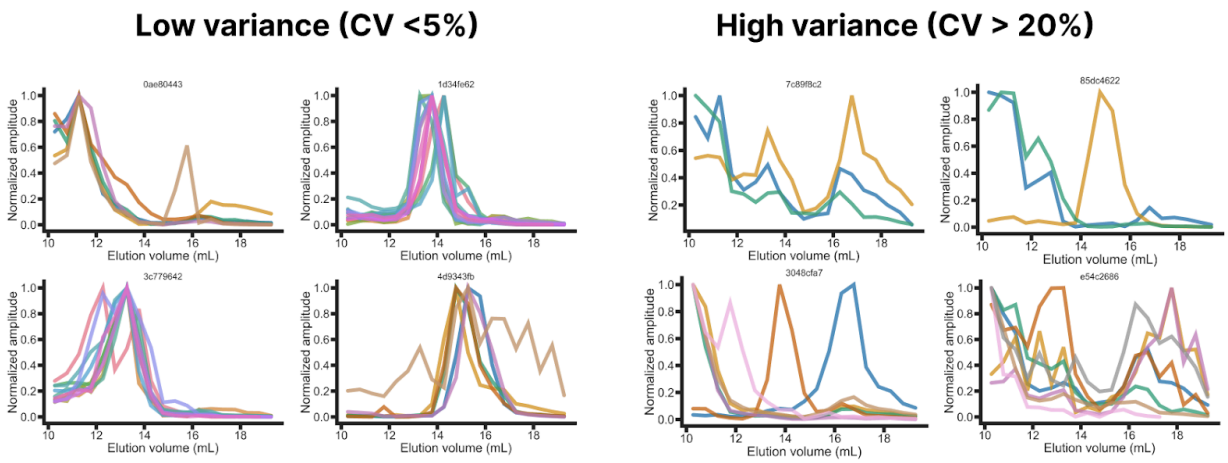
3.9 – MS barcoding supplementary information



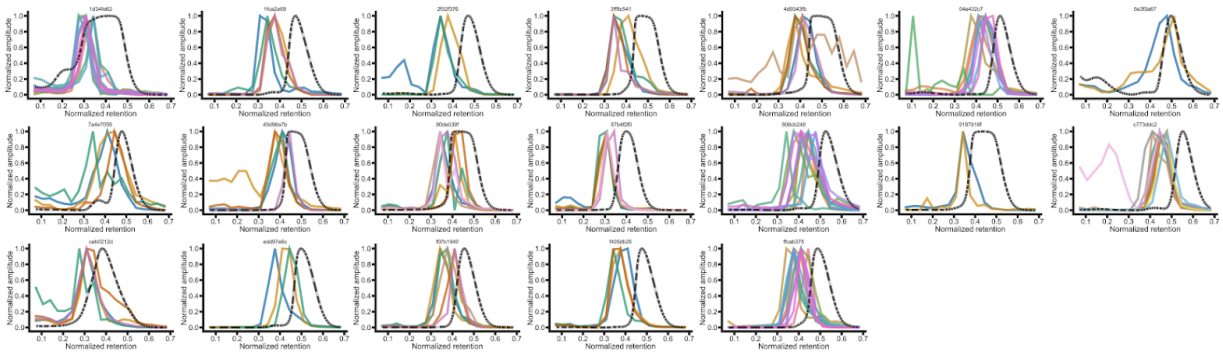
Supplementary Figure 1. Barcode detection rate by acidic residues across protocols using muGFP tagged with v1 (C-terminal) or v2 (N-terminal) barcodes. Barcode detection rate was quantified using the Skyline analysis pipeline for DDA or the DIA-NN analysis pipeline for DIA. The v2 library was restricted to < 3 acidic residues (Asp & Glu) to ensure higher detection rates.



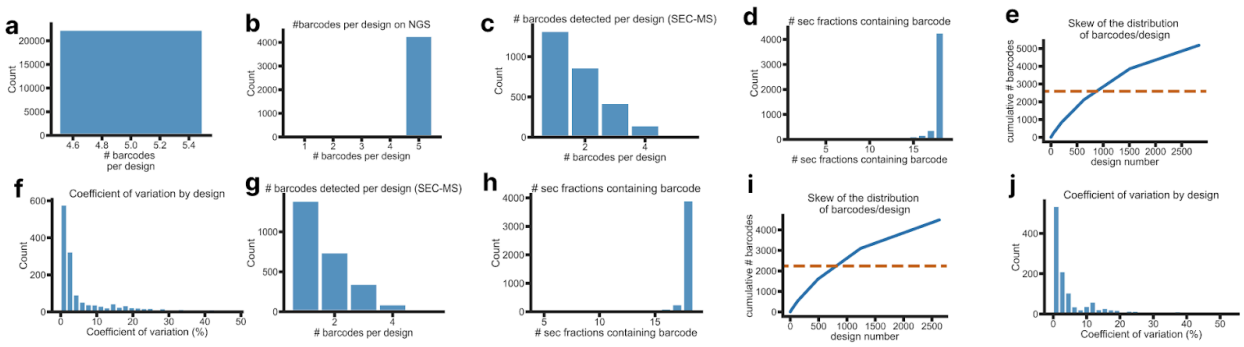
Supplementary Figure 2. Beta barrel library design and detection statistics. **(a)** distribution of barcodes ordered per design. **(b)** distribution of barcodes per design as detected by NGS. **(c)** distribution of barcodes per design as detected by SEC-MS. **(d)** distribution of barcode detection rate by SEC fraction. **(e)** skew of design representation based on SEC-MS barcode detection. 50% of the detected barcodes account for 30% of all designs. **(f)** coefficient of variation (CV) of barcode elution volume for SEC-MS. 81% of designs have CV < 10%.



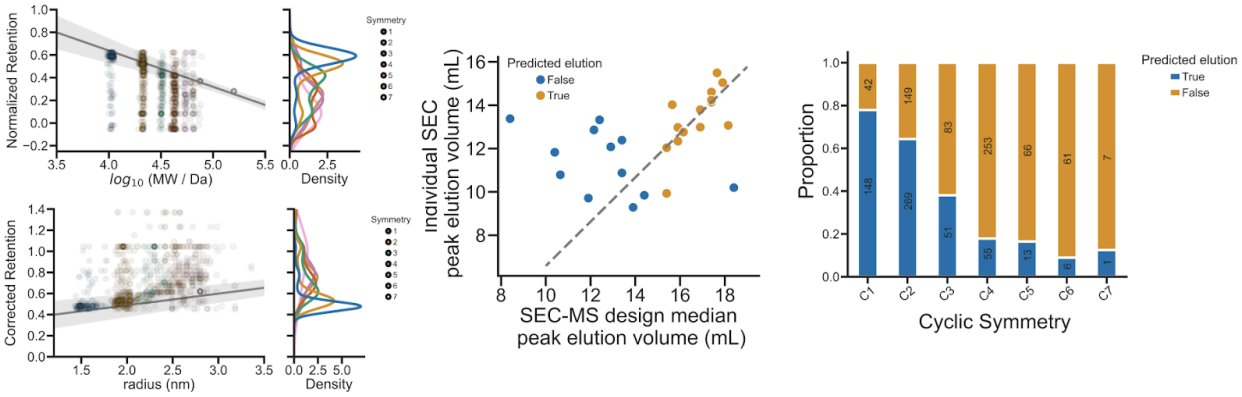
Supplementary Figure 3. Beta barrel library SEC-MS traces with low variability (< 5% coefficient of variation, CV) in peak elution volume (left) and high variability (>20%, right).



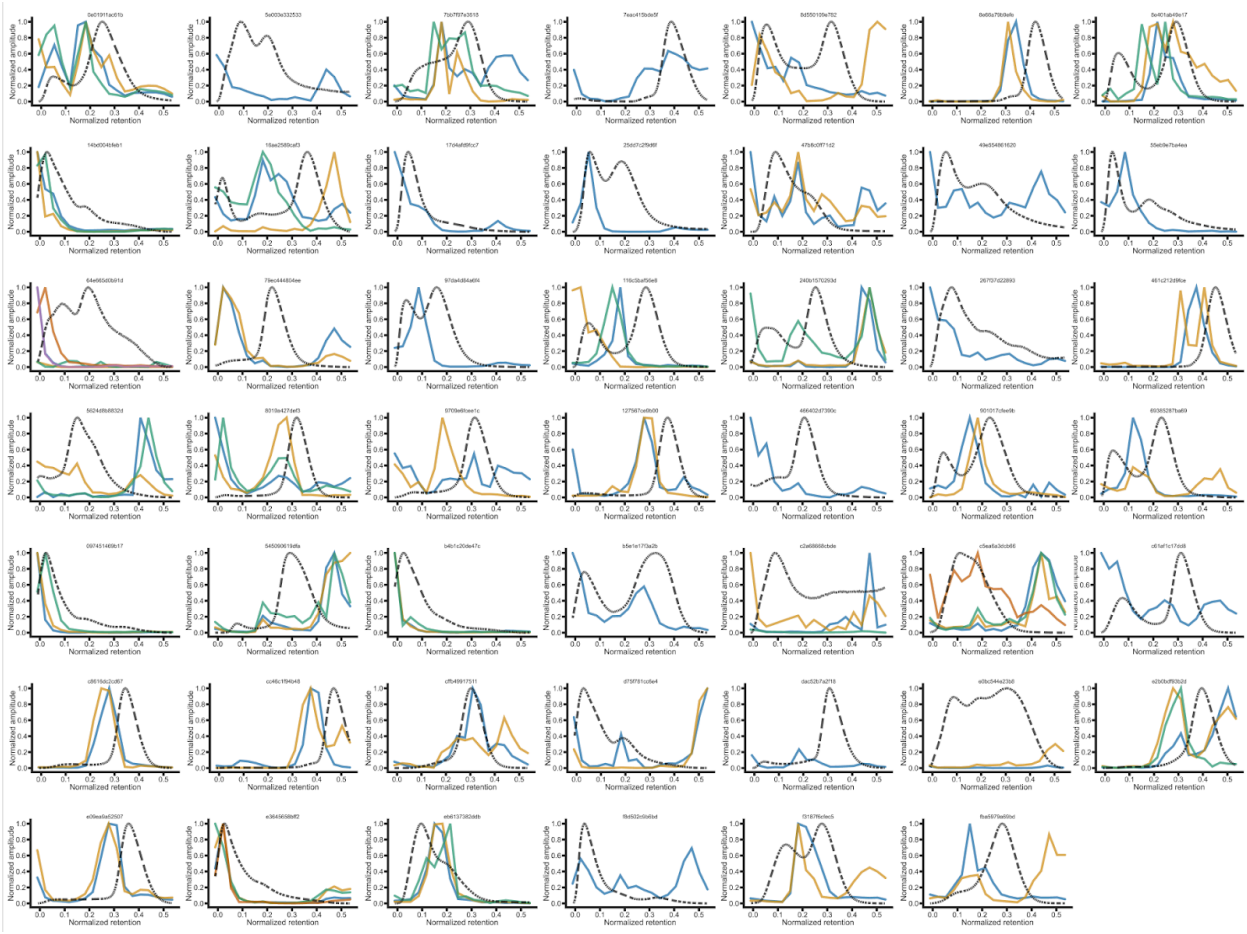
Supplementary Figure 4. Individual beta barrel SEC traces of validated monomers without barcodes (absorbance at 230nm, black dotted) overlaid with their corresponding barcoded SEC-MS traces (solid color).



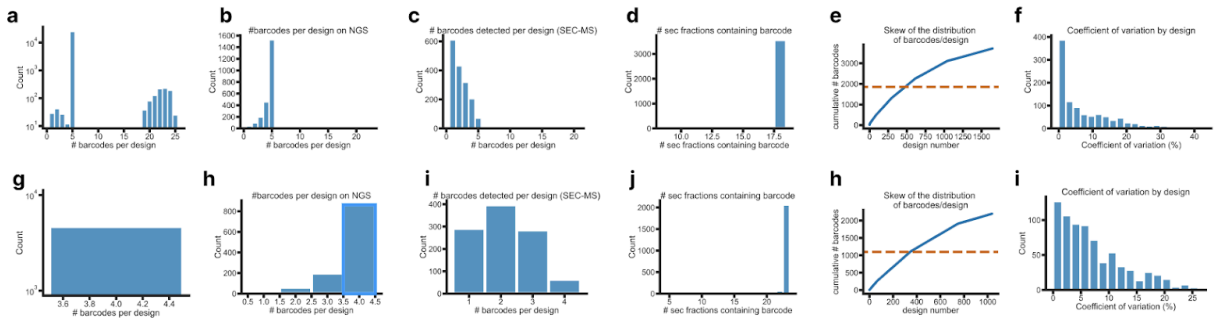
Supplementary Figure 5. Hallucinated cyclic oligomer design and detection statistics for S75 and S200 runs. **(a)** distribution of barcodes ordered per design. **(b)** distribution of barcodes per design as detected by NGS. **(c)** distribution of barcodes per design as detected by S75 SEC-MS. **(d)** distribution of barcode detection rate by S75 SEC fraction. **(e)** skew of design representation based on S75 SEC-MS barcode detection. 50% of the detected barcodes (orange, dotted) account for 29% of all designs. **(f)** coefficient of variation of barcode elution volume for S75 SEC-MS. **(g)** distribution of barcodes per design as detected by S200 SEC-MS. **(h)** distribution of barcode detection rate by S200 SEC fraction. **(i)** skew of design representation based on S200 SEC-MS barcode detection. 50% of the detected barcodes (orange, dotted) account for 29% of all designs. **(j)** coefficient of variation of barcode elution volume for S75 SEC-MS.



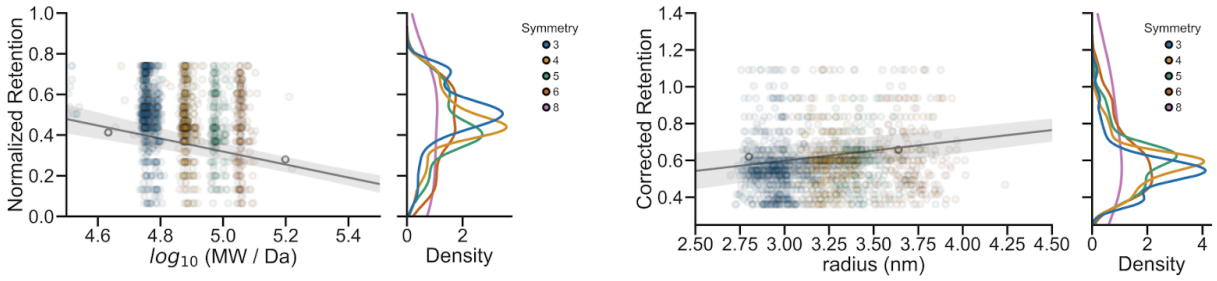
Supplementary Figure 6. a) Hallucinated cyclic oligomer library SEC elution profiles for S75 (top) and S200 (bottom). b) SEC peak elution volumes of hits expressed clonally at 50-mL culture scale (y-axis) vs barcode peak elution volume (x-axis) on an S200 column. c) Design success rate based on concordance of predicted and measured elution volumes on SEC.



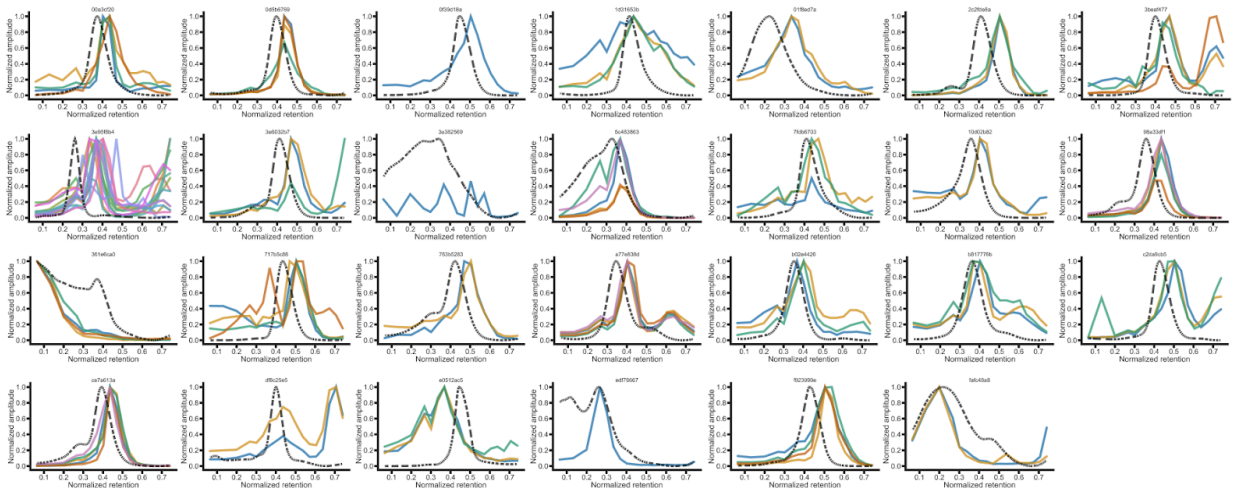
Supplementary Figure 7. Hallucinated oligomer hits S75 SEC-MS (solid color) overlaid with clonal SEC S75 data (Absorbance @ 280 nm, black dotted). Sequence-verified clones were expressed with a C-terminal 6x-His tag (without barcodes), purified via Ni-NTA, and sized on an S75 Increase 10/300 GL in 50 mM Tris, 150 mM NaCl, pH 8.



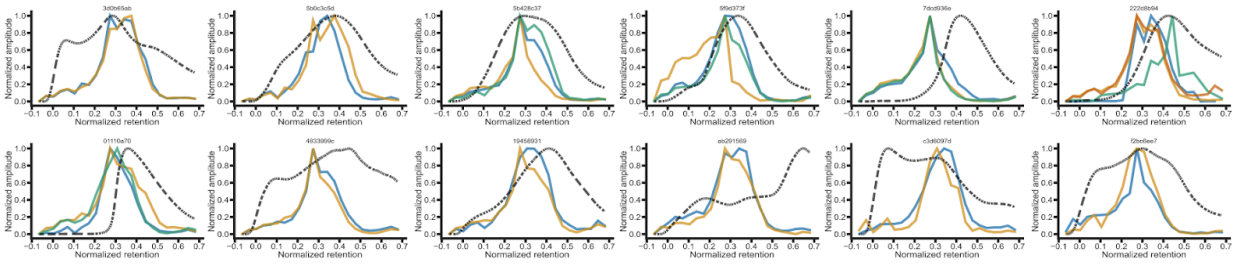
Supplementary Figure 8. Large helical oligomer **(top)** and I3 icosahedral nanocage **(bottom)** design and detection statistics. **(a & g)** distribution of barcodes ordered per design. **(b & h)** distribution of barcodes per design as detected by NGS. **(c and i)** distribution of barcodes per design as detected by SEC-MS **(d & j)** distribution of barcode detection rate by SEC fraction. **(e & k)** skew of design representation based on SEC-MS barcode detection. 50% of the detected barcodes account for 29% (top) and 33% (bottom) of all designs. **(f & i)** coefficient of variation (CV) of barcode elution volume for large helical oligomer (top) and I3 icosahedra (bottom) SEC-MS. 81% of large helical oligomer designs have CV < 10%, and 53% of I3 icosahedral designs have CV < 10%.



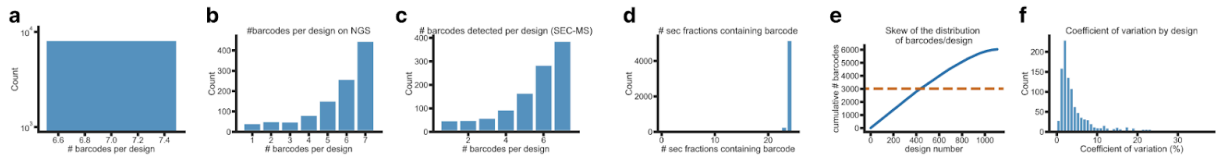
Supplementary Figure 9. S200 elution profiles of large helical oligomers along a standard curve. For each design, the median barcode elution profile was plotted vs either $\log_{10}$(molecular weight) in Da, or via computed hydrodynamic radius (see Supplementary Figure 11). Normalized retention (also known as K_{av}) is based on the void volume of the S200 column (by elution of blue dextran, where $K_{av} = 0$) and the known column volume (24 mL – $K_{av} = 1$).



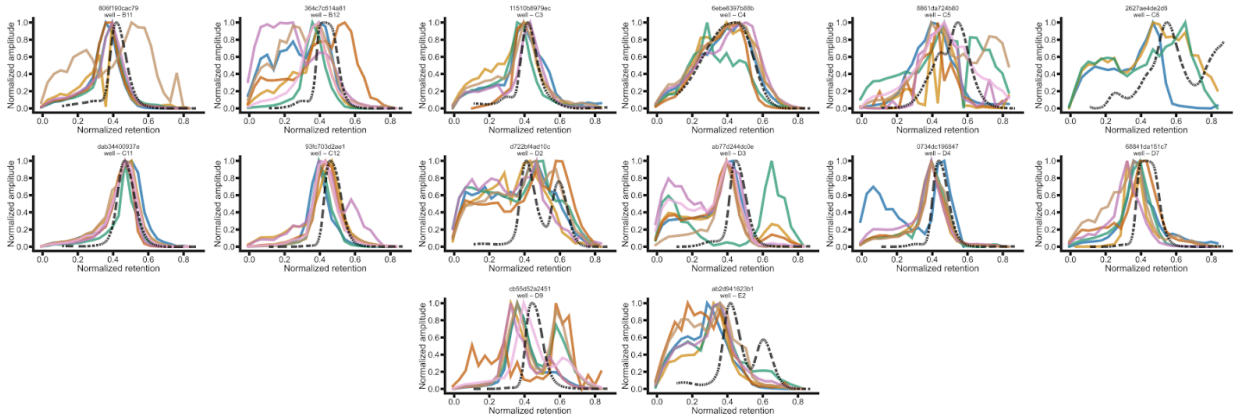
Supplementary Figure 10. Large helical oligomer hits SEC-MS overlaid (solid color) with clonal SEC data (Absorbance @ 280 nm, black dotted). Sequence-verified clones were expressed with a C-terminal 6x-His tag (without barcodes), purified via Ni-NTA, and sized on an S200 Increase 10/300 GL in 50 mM Tris, 300 mM NaCl, pH 8.



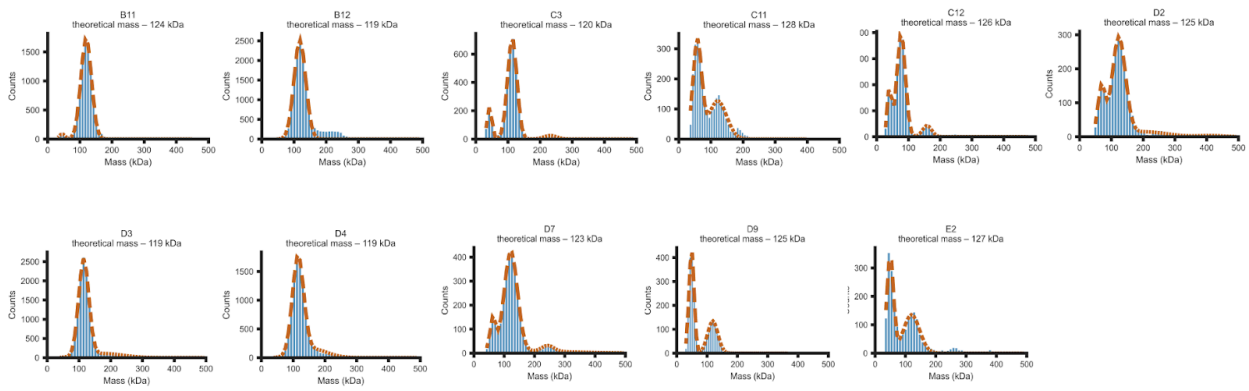
Supplementary Figure 11. I3 icosahedral hits SEC-MS overlaid (solid color) with clonal SEC data (Absorbance @ 280 nm, black dotted). Sequence-verified clones were expressed with a C-terminal 6x-His tag (without barcodes), purified via Ni-NTA, and sized on an S200 Increase 10/300 GL in 50 mM Tris, 300 mM NaCl, pH 8.



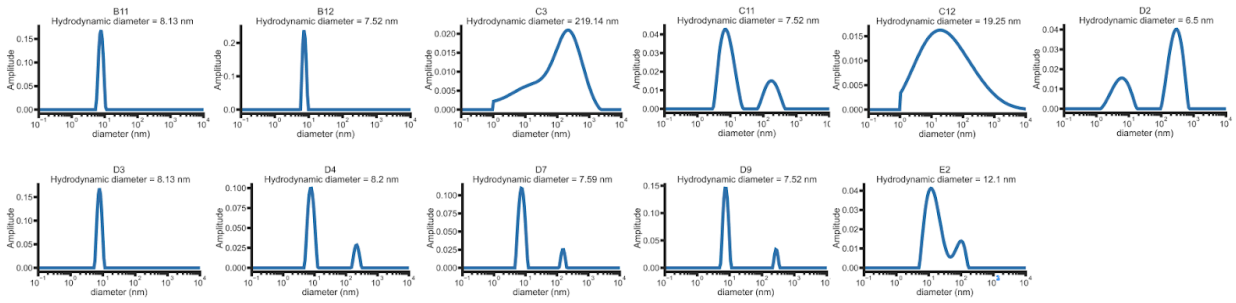
Supplementary Figure 12. Tetrahedra library design and detection statistics. **(a)** distribution of barcodes ordered per design. **(b)** distribution of barcodes per design as detected by NGS. **(c)** distribution of barcodes per design as detected by SEC-MS. **(d)** distribution of barcode detection rate by SEC fraction. **(e)** skew of design representation based on SEC-MS barcode detection. 50% of the detected barcodes account for 39% of all designs. **(f)** coefficient of variation (CV) of barcode elution volume for SEC-MS. 81% of designs have CV < 10%.



Supplementary Figure 13. Tetrahedral hits SEC-MS overlaid (solid color) with clonal SEC data (black dotted). Sequence-verified clones were expressed with a C-terminal 6x-His tag (without barcodes), purified via Ni-NTA, and sized on an S200 Increase 10/300 GL in 50 mM Tris, 300 mM NaCl, pH 8. Absorbance @ 230 nm was recorded as many of these proteins lack aromatics that absorb at 280 nm.



Supplementary Figure 14. Mass photometry data of hits from tetrahedral library. Sequence-verified clones were expressed, purified, and sized on an S200 in 50mM Tris, 300mM NaCl, pH 8. Peaks corresponding to expected elution volume were collected and 10 µL of 1:1000 diluted solution (in aforementioned SEC running buffer) was subjected to mass photometry analysis.



Supplementary Figure 15. Dynamic light scattering data of hits from tetrahedral library. Sequence-verified clones were expressed, purified, and sized on an S200 in 50mM Tris, 300mM NaCl, pH 8. Peaks corresponding to expected elution volume were collected and 8.8 µL of 1mg/mL solution to dynamic light scattering analysis.

Supplementary Table 1. X-ray diffraction data collection and refinement statistics

	SG266 (8VEA)	B11	C3	D3	D9
Resolution range	41.58 - 3.30 (3.55 - 3.30)	45.71 - 2.52 (2.62 - 2.52)	36.32 - 2.87 (3.10 - 2,87)	46.54 - 2.54 (2.6 - 2.54)	31.86 - 2.04 (2.15 - 2.04)
Space group	*P 63*	*P 21 21 21*	*I 2 3*	*P 21*	*I 2 2 2*
Unit cell	115.33, 115.33, 59.99; 90, 90, 120	45.71, 49.55, 76.02; 90, 90, 9	72.64, 72.64, 72.64; 90, 90, 90	75.03, 119.44, 116.12; 90, 91.62, 90	67.97, 69.27, 84.46; 90, 90, 90
Unique reflections	6977 (1381)	6161 (654)	1542 (305)	306782 (5350)	13040 (1879)
Multiplicity	15.9 (14.7)	12.8 (13)	38 (38)	3.4 (2.0)	10.9 (11.2)
Completeness (%)	99.69 (99.78)	99.3 (97.8)	100.00 (100.00)	99.19 (99.05)	100 (100)
Mean I/sigma (I)	28.45 (10.97)	14.8 (2.4)	19.8 (2.8)	9.0 (1.0)	8.3 (1.2)
Wilson B-factor	101.26	46.36	87.47	49.72	38.36
R-merge	0.074 (0.292)	0.123 (1.065)	0.172 (1.826)	0.167 (2.056)	0.159 (2.387)
R-pim	0.018 (0.078)	0.037 (0.315)	0.029 (0.302)	0.074 (0.900)	0.053 (0.782)
$CC_{1/2}$	1.00 (0.987)	0.999 (0.866)	0.999 (0.810)	0.998 (0.363)	0.996 (0.416)
Reflections used in refinement	6977 (1381)	5752 (1148)	1539 (1539)	67006 (4788)	11792 (1220)
R-work	0.1850 (0.2742)	0.2650 (0.3202)	0.2376 (0.2376)	0.2131 (0.3106)	0.2220 (0.3138)
R-free	0.2423 (0.3231)	0.3051 (0.3978)	0.2722 (0.2722)	0.2624 (0.3693)	0.2781 (0.3671)
Number of non-hydrogen atoms	3492	1283	609	14856	1921
macromolecules	3492	1261	609	14848	1906
Solent	n/a	22	n/a	8	15
Protein residues	429	159	79	1979	239
RMS (bonds)	0.004	0.003	0.004	0.002	0.002
RMS (angles)	0.64	0.549	0.70	0.49	0.46
Ramachandran favored (%)	96.93	96.77	93.51	98.29	99.14
Ramachandran allowed (%)	2.84	3.23	6.49	1.71	0.86
Ramachandran outliers (%)	0.24	0.00	0.00	0.00	0.00

Average B-factor	98	51	81	57	52
macromolecules	98	51	81	57	52
Solvent	n/a	42	n/a	49	48

Statistics for the highest-resolution shell are shown in parentheses.

Conclusion

In this work, I demonstrated an approach to develop protein-based receptor decoys that leverages deep-learning protein backbone and sequence design tools, and further, how these tools can be used to optimize a hit for manufacturability. Additionally, I improved the analysis pipeline and detection rate of an existing protein barcoding method and applied this approach to screen novel nanoparticles for assembly. While this is a brief summary of the work I've been able to carry out, I highlight these two points because they emphasize a major focus of my PhD – the synergy of *in silico* design and robust wet lab characterization.

The foundation of deep learning is a large dataset. Unsurprisingly, the massive advances in protein design have benefited from the existence of the Protein Data Bank, a publicly available dataset of high quality protein structures. As we usher in the next era of protein design, there will be increasing attention on generation of datasets to explore how protein sequence and structure interact with other physical properties and systems. Using the resources at my disposal, I was able to optimize an existing technique to make viable the screening of larger libraries with fewer barcodes – hopefully, such a tool may lower the barrier for protein designers who wish to generate their own protein datasets in house.

Looking forward, one of the goals of employing larger datasets is to design proteins to carry out highly specific functions – all on the computer. Since my PhD began, the field of protein design has progressed leaps and bounds. In less than a year, I was able to develop a protein that can neutralize a virus with potencies rivaling those of antibodies, all while testing fewer than 150 unique sequences. While we are certainly far off from the realm of zero-shot design, it has been a wonderful experience to explore protein design on the computer and in the lab in an age where the tools on both sides are rapidly evolving.

References

Adams, Paul D., Pavel V. Afonine, Gábor Bunkóczi, Vincent B. Chen, Ian W. Davis, Nathaniel Echols, Jeffrey J. Headd, et al. 2010. "PHENIX: A Comprehensive Python-Based System for Macromolecular Structure Solution." *Acta Crystallographica. Section D, Biological Crystallography* 66 (Pt 2): 213–21.

Baek, Minkyung, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, et al. 2021. "Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network." *Science* 373 (6557): 871–76.

Balicer, Ran D., Michael Huerta, Nadav Davidovitch, and Itamar Grotto. 2005. "Cost-Benefit of Stockpiling Drugs for Influenza Pandemic." *Emerging Infectious Diseases* 11 (8): 1280–82.

Barkovits, Katalin, Sandra Pacharra, Kathy Pfeiffer, Simone Steinbach, Martin Eisenacher, Katrin Marcus, and Julian Uszkoreit. 2020. "Reproducibility, Specificity and Accuracy of Relative Quantification Using Spectral Library-Based Data-Independent Acquisition." *Molecular & Cellular Proteomics: MCP* 19 (1): 181–97.

Bateman, Nicholas W., Scott P. Goulding, Nicholas J. Shulman, Avinash K. Gadok, Karen K. Szumlinski, Michael J. MacCoss, and Christine C. Wu. 2014. "Maximizing Peptide Identification Events in Proteomic Workflows Using Data-Dependent Acquisition (DDA)." *Molecular & Cellular Proteomics: MCP* 13 (1): 329–38.

Boder, E. T., and K. D. Wittrup. 1997. "Yeast Surface Display for Screening Combinatorial Polypeptide Libraries." *Nature Biotechnology* 15 (6): 553–57.

Bonaparte, Matthew I., Antony S. Dimitrov, Katharine N. Bossart, Gary Crameri, Bruce A. Mungall, Kimberly A. Bishop, Vidita Choudhry, et al. 2005. "Ephrin-B2 Ligand Is a Functional Receptor for Hendra Virus and Nipah Virus." *Proceedings of the National Academy of Sciences of the United States of America* 102 (30): 10652–57.

Bowden, Thomas A., A. Radu Aricescu, Robert J. C. Gilbert, Jonathan M. Grimes, E. Yvonne Jones, and David I. Stuart. 2008. "Structural Basis of Nipah and Hendra Virus Attachment to Their Cell-Surface Receptor Ephrin-B2." *Nature Structural & Molecular Biology* 15 (6): 567–72.

Cabantous, Stéphanie, and Geoffrey S. Waldo. 2006. "In Vivo and in Vitro Protein Solubility Assays Using Split GFP." *Nature Methods* 3 (10): 845–54.

Cao, Longxing, Brian Coventry, Inna Goreshnik, Buwei Huang, William Sheffler, Joon Sung Park, Kevin M. Jude, et al. 2022. "Design of Protein-Binding Proteins from the Target Structure Alone." *Nature* 605 (7910): 551–60.

Cao, Longxing, Inna Goreshnik, Brian Coventry, James Brett Case, Lauren Miller, Lisa Kozodoy, Rita E. Chen, et al. 2020. "De Novo Design of Picomolar SARS-CoV-2 Miniprotein Inhibitors." *Science* 370 (6515): 426–31.

Charbit, A., J. C. Boulain, A. Ryter, and M. Hofnung. 1986. "Probing the Topology of a Bacterial Membrane Protein by Genetic Insertion of a Foreign Epitope; Expression at the Cell Surface." *The EMBO Journal* 5 (11): 3029–37.

Cui, Gaofeng, Sungman Park, Aimee I. Badeaux, Donghwa Kim, Joseph Lee, James R. Thompson, Fei Yan, et al. 2012. "PHF20 Is an Effector Protein of p53 Double Lysine Methylation That Stabilizes and Activates p53." *Nature Structural & Molecular Biology* 19 (9): 916–24.

Dauparas, J., I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, et al. 2022. "Robust Deep Learning-Based Protein Sequence Design Using ProteinMPNN." *Science* 378 (6615): 49–56.

Delft, Annette von, Matthew D. Hall, Ann D. Kwong, Lisa A. Purcell, Kumar Singh Saikatendu, Uli Schmitz, John A. Tallarico, and Alpha A. Lee. 2023. “Accelerating Antiviral Drug Discovery: Lessons from COVID-19.” *Nature Reviews. Drug Discovery* 22 (7): 585–603.
Demichev, Vadim, Christoph B. Messner, Spyros I. Vernardis, Kathryn S. Lilley, and Markus Ralser. 2020. “DIA-NN: Neural Networks and Interference Correction Enable Deep Proteome Coverage in High Throughput.” *Nature Methods* 17 (1): 41–44.
Dong, Jinhui, Robert W. Cross, Michael P. Doyle, Nurgun Kose, Jarrod J. Mousa, Edward J. Annand, Viktoriya Borisevich, et al. 2020. “Potent Henipavirus Neutralization by Antibodies Recognizing Diverse Sites on Hendra and Nipah Virus Receptor Binding Protein.” *Cell* 183 (6): 1536–50.e17.
Egertson, Jarrett D., Brendan MacLean, Richard Johnson, Yue Xuan, and Michael J. MacCoss. 2015. “Multiplexed Peptide Analysis Using Data-Independent Acquisition and Skyline.” *Nature Protocols* 10 (6): 887–903.
Egloff, Pascal, Iwan Zimmermann, Fabian M. Arnold, Cedric A. J. Hutter, Damien Morger, Lennart Opitz, Lucy Poveda, et al. 2019. “Engineered Peptide Barcodes for in-Depth Analyses of Binding Protein Libraries.” *Nature Methods* 16 (5): 421–28.
Emsley, Paul, and Kevin Cowtan. 2004. “Coot: Model-Building Tools for Molecular Graphics.” *Acta Crystallographica. Section D, Biological Crystallography* 60 (Pt 12 Pt 1): 2126–32.
Escher, Claudia, Lukas Reiter, Brendan MacLean, Reto Ossola, Franz Herzog, John Chilton, Michael J. MacCoss, and Oliver Rinner. 2012. “Using iRT, a Normalized Retention Time for More Targeted Measurement of Peptides.” *Proteomics* 12 (8): 1111–21.
Freudl, R., S. MacIntyre, M. Degen, and U. Henning. 1986. “Cell Surface Exposure of the Outer Membrane Protein OmpA of Escherichia Coli K-12.” *Journal of Molecular Biology* 188 (3): 491–94.
Gerben, Stacey R., Andrew J. Borst, Derrick R. Hicks, Isabelle Moczygemba, David Feldman, Brian Coventry, Wei Yang, et al. 2023. “Design of Diverse Asymmetric Pockets in Homo-Oligomeric Proteins.” *Biochemistry* 62 (2): 358–68.
Gessulat, Siegfried, Tobias Schmidt, Daniel Paul Zolg, Patroklos Samaras, Karsten Schnatbaum, Johannes Zerweck, Tobias Knaute, et al. 2019. “Prosit: Proteome-Wide Prediction of Peptide Tandem Mass Spectra by Deep Learning.” *Nature Methods* 16 (6): 509–18.
Harris, J. L., B. J. Backes, F. Leonetti, S. Mahrus, J. A. Ellman, and C. S. Craik. 2000. “Rapid and General Profiling of Protease Specificity by Using Combinatorial Fluorogenic Substrate Libraries.” *Proceedings of the National Academy of Sciences of the United States of America* 97 (14): 7754–59.
Ho, Mitchell, Satoshi Nagata, and Ira Pastan. 2006. “Isolation of Anti-CD22 Fv with High Affinity by Fv Display on Human Cells.” *Proceedings of the National Academy of Sciences of the United States of America* 103 (25): 9637–42.
Ho, Mitchell, and Ira Pastan. 2009. “Mammalian Cell Display for Antibody Engineering.” *Methods in Molecular Biology* 525:337–52, xiv.
Howell, Shannon M., Stephen V. Fiacco, Terry T. Takahashi, Farzad Jalali-Yazdi, Steven W. Millward, Biliang Hu, Pin Wang, and Richard W. Roberts. 2014. “Serum Stable Natural Peptides Designed by mRNA Display.” *Scientific Reports* 4 (September):6008.
Hunt, Andrew C., James Brett Case, Young-Jun Park, Longxing Cao, Kejia Wu, Alexandra C. Walls, Zhuoming Liu, et al. 2021. “Multivalent Designed Proteins Protect against SARS-CoV-2 Variants of Concern.” *bioRxiv : The Preprint Server for Biology*, July. https://doi.org/10.1101/2021.07.07.451375.
Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. “Highly Accurate Protein Structure Prediction with AlphaFold.” *Nature* 596 (7873): 583–89.
Kabsch, Wolfgang. 2010. “XDS.” *Acta Crystallographica. Section D, Biological Crystallography* 66 (Pt 2): 125–32.

Kawashima, Yusuke, Eiichiro Watanabe, Taichi Umeyama, Daisuke Nakajima, Masahira Hattori, Kenya Honda, and Osamu Ohara. 2019. “Optimization of Data-Independent Acquisition Mass Spectrometry for Deep and Highly Sensitive Proteomic Analysis.” *International Journal of Molecular Sciences* 20 (23). https://doi.org/10.3390/ijms20235932.

Kenmoe, Sebastien, Maurice Demanou, Jean Joel Bigna, Cyprien Nde Kengne, Abdou Fatawou Modiyinji, Fredy Brice N. Simo, Sara Eyangoh, Serge Alain Sadeuh-Mba, and Richard Njouom. 2019. “Case Fatality Rate and Risk Factors for Nipah Virus Encephalitis: A Systematic Review and Meta-Analysis.” *Journal of Clinical Virology: The Official Publication of the Pan American Society for Clinical Virology* 117 (August):19–26.

Kim, David E., Davin R. Jensen, David Feldman, Doug Tischer, Ayesha Saleem, Cameron M. Chow, Xinting Li, et al. 2023. “De Novo Design of Small Beta Barrel Proteins.” *Proceedings of the National Academy of Sciences of the United States of America* 120 (11): e2207974120.

Klein, Jason C., Marc J. Lajoie, Jerrod J. Schwartz, Eva-Maria Strauch, Jorgen Nelson, David Baker, and Jay Shendure. 2016. “Multiplex Pairwise Assembly of Array-Derived DNA Oligonucleotides.” *Nucleic Acids Research* 44 (5): e43.

Koepnick, Brian, Jeff Flatten, Tamir Husain, Alex Ford, Daniel-Adriano Silva, Matthew J. Bick, Aaron Bauer, et al. 2019. “De Novo Protein Design by Citizen Scientists.” *Nature* 570 (7761): 390–94.

Larsen, Brendan B., Teagan McMahon, Jack T. Brown, Zhaoqian Wang, Caelan E. Radford, James E. Crowe Jr, David Veesler, and Jesse D. Bloom. 2024. “Functional and Antigenic Landscape of the Nipah Virus Receptor Binding Protein.” *bioRxiv : The Preprint Server for Biology*, April. https://doi.org/10.1101/2024.04.17.589977.

Lasko, Paul. 2010. “Tudor Domain.” *Current Biology: CB* 20 (16): R666–67.

Liu, Qian, Jacquelyn A. Stone, Birgit Bradel-Tretheway, Jeffrey Dabundo, Javier A. Benavides Montano, Jennifer Santos-Montanez, Scott B. Biering, et al. 2013. “Unraveling a Three-Step Spatiotemporal Mechanism of Triggering of Receptor-Induced Nipah Virus Fusion and Cell Entry.” *PLoS Pathogens* 9 (11): e1003770.

Lou, Ronghui, Ye Cao, Shanshan Li, Xiaoyu Lang, Yunxia Li, Yaoyang Zhang, and Wenqing Shui. 2023. “Benchmarking Commonly Used Software Suites and Analysis Workflows for DIA Proteomics and Phosphoproteomics.” *Nature Communications* 14 (1): 94.

Luby, Stephen P., and Emily S. Gurley. 2012. “Epidemiology of Henipavirus Disease in Humans.” *Current Topics in Microbiology and Immunology* 359:25–40.

McCoy, Airlie J., Ralf W. Grosse-Kunstleve, Paul D. Adams, Martyn D. Winn, Laurent C. Storoni, and Randy J. Read. 2007. “Phaser Crystallographic Software.” *Journal of Applied Crystallography* 40 (Pt 4): 658–74.

Narayanan, Krishna K., Moushimi Amaya, Natalie Tsang, Randy Yin, Alka Jays, Christopher C. Broder, Diwakar Shukla, and Erik Procko. 2023a. “Sequence Basis for Selectivity of Ephrin-B2 Ligand for Eph Receptors and Pathogenic Henipavirus G Glycoproteins.” *Journal of Virology* 97 (11): e0062123.

———. 2023b. “The Sequence Basis for Selectivity of Ephrin-B2 Ligand for Eph Receptors and Pathogenic Henipavirus G Glycoproteins: Selective Ephrin-B2 Decoys for Nipah and Hendra Virus.” *bioRxiv : The Preprint Server for Biology*, April. https://doi.org/10.1101/2023.04.26.538420.

Newton, Matilda S., Yari Cabezas-Perusse, Cher Ling Tong, and Burckhard Seelig. 2020. “In Vitro Selection of Peptides and Proteins—Advantages of mRNA Display.” *ACS Synthetic Biology* 9 (2): 181–90.

Pande, Jyoti, Magdalena M. Szewczyk, and Ashok K. Grover. 2010. “Phage Display: Concept, Innovations, Applications and Future.” *Biotechnology Advances* 28 (6): 849–58.

Pillai, Arvind, Abbas Idris, Annika Philomin, Connor Weidle, Rebecca Skotheim, Philip J. Y.

Leung, Adam Broerman, et al. 2023. “De Novo Design of Allosterically Switchable Protein Assemblies.” *bioRxiv*. https://doi.org/10.1101/2023.11.01.565167.
Plans-Rubió, Pedro. 2020. “The Cost Effectiveness of Stockpiling Drugs, Vaccines and Other Health Resources for Pandemic Preparedness.” *PharmacoEconomics - Open* 4 (3): 393–95.
Playford, Elliott Geoffrey, Trent Munro, Stephen M. Mahler, Suzanne Elliott, Michael Gerometta, Kym L. Hoger, Martina L. Jones, et al. 2020. “Safety, Tolerability, Pharmacokinetics, and Immunogenicity of a Human Monoclonal Antibody Targeting the G Glycoprotein of Henipaviruses in Healthy Adults: A First-in-Human, Randomised, Controlled, Phase 1 Study.” *The Lancet Infectious Diseases* 20 (4): 445–54.
Reis, Aimee L. van der, Lynnath E. Beckley, M. Pilar Olivar, and Andrew G. Jeffs. 2022. “Nanopore Short‐read Sequencing: A Quick, Cost‐effective and Accurate Method for DNA Metabarcoding.” *Environmental DNA (Hoboken, N.J.)*, December. https://doi.org/10.1002/edn3.374.
Satter, Syed Moinuddin, Wasik Rahman Aquib, Sharmin Sultana, Ahmad Raihan Sharif, Arifa Nazneen, Muhammad Rashedul Alam, Ayesha Siddika, et al. 2023. “Tackling a Global Epidemic Threat: Nipah Surveillance in Bangladesh, 2006–2021.” *PLoS Neglected Tropical Diseases* 17 (9): e0011617.
Siddiqui, M. Ruby, and W. John Edmunds. 2008. “Cost-Effectiveness of Antiviral Stockpiling and near-Patient Testing for Potential Influenza Pandemic.” *Emerging Infectious Diseases* 14 (2): 267–74.
Sifniotis, Vicki, Esteban Cruz, Barbaros Eroglu, and Veysel Kayser. 2019. “Current Advancements in Addressing Key Challenges of Therapeutic Antibody Design, Manufacture, and Formulation.” *Antibodies (Basel, Switzerland)* 8 (2). https://doi.org/10.3390/antib8020036.
Smith, G. P. 1985. “Filamentous Fusion Phage: Novel Expression Vectors That Display Cloned Antigens on the Virion Surface.” *Science* 228 (4705): 1315–17.
Studier, F. William. 2014. “Stable Expression Clones and Auto-Induction for Protein Production in E. Coli.” *Methods in Molecular Biology* 1091:17–32.
Tong, Qiong, Gaofeng Cui, Maria Victoria Botuyan, Scott B. Rothbart, Ryo Hayashi, Catherine A. Musselman, Namit Singh, et al. 2015. “Structural Plasticity of Methyllysine Recognition by the Tandem Tudor Domain of 53BP1.” *Structure* 23 (2): 312–21.
Tsuboyama, Kotaro, Justas Dauparas, Jonathan Chen, Elodie Laine, Yasser Mohseni Behbahani, Jonathan J. Weinstein, Niall M. Mangan, Sergey Ovchinnikov, and Gabriel J. Rocklin. 2023. “Mega-Scale Experimental Analysis of Protein Folding Stability in Biology and Design.” *Nature* 620 (7973): 434–44.
Vanderpoel, Julie, Andrea L. Stevens, Bruno Emond, Marie-Hélène Lafeuille, Annalise Hilts, Patrick Lefebvre, and Laura Morrison. 2022. “Total Cost of Testing for Genomic Alterations Associated with next-Generation Sequencing versus Polymerase Chain Reaction Testing Strategies among Patients with Metastatic Non-Small Cell Lung Cancer.” *Journal of Medical Economics* 25 (1): 457–68.
Vasudevan, Srivatsa Surya, Arun Subash, Fena Mehta, Tiba Yamin Kandrikar, Rupak Desai, Kaif Khan, Sneha Khanduja, et al. 2024. “Global and Regional Mortality Statistics of Nipah Virus from 1994 to 2023: A Comprehensive Systematic Review and Meta-Analysis.” *Pathogens and Global Health*, July, 1–10.
Wang, Eric Y., Yile Dai, Connor E. Rosen, Monica M. Schmitt, Mei X. Dong, Elise M. N. Ferré, Feimei Liu, et al. 2022. “High-Throughput Identification of Autoantibodies That Target the Human Exoproteome.” *Cell Reports Methods* 2 (2). https://doi.org/10.1016/j.crmeth.2022.100172.
Wargacki, Adam J., Tobias P. Wörner, Michiel van de Waterbeemd, Daniel Ellis, Albert J. R. Heck, and Neil P. King. 2021. “Complete and Cooperative in Vitro Assembly of Computationally Designed Self-Assembling Protein Nanomaterials.” *Nature*

Communications 12 (1): 883.
Watson, Joseph L., David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, et al. 2023. "De Novo Design of Protein Structure and Function with RFdiffusion." *Nature* 620 (7976): 1089–1100.
Weatherman, Sarah, Heinz Feldmann, and Emmie de Wit. 2018. "Transmission of Henipaviruses." *Current Opinion in Virology* 28 (February):7–11.
Wicky, B. I. M., L. F. Milles, A. Courbet, R. J. Ragotte, J. Dauparas, E. Kinfu, S. Tipps, et al. 2022. "Hallucinating Symmetric Protein Assemblies." *Science* 378 (6615): 56–61.
Williams, Christopher J., Jeffrey J. Headd, Nigel W. Moriarty, Michael G. Prisant, Lizbeth L. Videau, Lindsay N. Deis, Vishal Verma, et al. 2018. "MolProbity: More and Better Reference Data for Improved All-Atom Structure Validation." *Protein Science: A Publication of the Protein Society* 27 (1): 293–315.
Wilson, D. S., A. D. Keefe, and J. W. Szostak. 2001. "The Use of mRNA Display to Select High-Affinity Protein-Binding Peptides." *Proceedings of the National Academy of Sciences of the United States of America* 98 (7): 3750–55.
Winn, Martyn D., Charles C. Ballard, Kevin D. Cowtan, Eleanor J. Dodson, Paul Emsley, Phil R. Evans, Ronan M. Keegan, et al. 2011. "Overview of the CCP4 Suite and Current Developments." *Acta Crystallographica. Section D, Biological Crystallography* 67 (Pt 4): 235–42.
Yamaguchi, Junichi, Mohammed Naimuddin, Manish Biyani, Toru Sasaki, Masayuki Machida, Tai Kubo, Takashi Funatsu, Yuzuru Husimi, and Naoto Nemoto. 2009. "cDNA Display: A Novel Screening Method for Functional Disulfide-Rich Peptides by Solid-Phase Synthesis and Stabilization of mRNA–protein Fusions." *Nucleic Acids Research* 37 (16): e108–e108.
Yang, Erin C., Robby Divine, Marcos C. Miranda, Andrew J. Borst, Will Sheffler, Jason Z. Zhang, Justin Decarreau, et al. 2024. "Computational Design of Non-Porous pH-Responsive Antibody Nanoparticles." *Nature Structural & Molecular Biology*, May. https://doi.org/10.1038/s41594-024-01288-5.
Youkharibache, Philippe, Stella Veretnik, Qingliang Li, Kimberly A. Stanek, Cameron Mura, and Philip E. Bourne. 2019. "The Small β-Barrel Domain: A Survey-Based Structural Analysis." *Structure* 27 (1): 6–26.
Zahnd, Christian, Patrick Amstutz, and Andreas Plückthun. 2007. "Ribosome Display: Selecting and Evolving Proteins in Vitro That Specifically Bind to a Target." *Nature Methods* 4 (3): 269–79.
Zeitlin, Larry, Robert W. Cross, Courtney Woolsey, Brandyn R. West, Viktoriya Borisevich, Krystle N. Agans, Abhishek N. Prasad, et al. 2024. "Therapeutic Administration of a Cross-Reactive mAb Targeting the Fusion Glycoprotein of Nipah Virus Protects Nonhuman Primates." *Science Translational Medicine* 16 (741): eadl2055.
Zhang, Ping, Devika Ganesamoorthy, Son Hoang Nguyen, Raymond Au, Lachlan J. Coin, and Siok-Keen Tey. 2020. "Nanopore Sequencing as a Scalable, Cost-Effective Platform for Analyzing Polyclonal Vector Integration Sites Following Clinical T Cell Therapy." *Journal for Immunotherapy of Cancer* 8 (1). https://doi.org/10.1136/jitc-2019-000299.
Zhao, Qiang, Lipeng Qin, Fuguo Jiang, Beili Wu, Wen Yue, Feng Xu, Zhili Rong, et al. 2007. "Structure of Human spindlin1. Tandem Tudor-like Domains for Cell Cycle Regulation." *The Journal of Biological Chemistry* 282 (1): 647–56.
Zutz, Ariane, Louise Hamborg, Lasse Ebdrup Pedersen, Maher M. Kassem, Elena Papaleo, Anna Koza, Markus J. Herrgård, et al. 2021. "A Dual-Reporter System for Investigating and Optimizing Protein Translation and Folding in E. Coli." *Nature Communications* 12 (1): 6093.