Accelerating protein design by scaling experimental characterization

Jason Qian*[,1,2], Lukas F. Milles*[,1,2,#], Basile I. M. Wicky*[,1,2,#], Amir Motmaen[1,2,3], Xinting Li[1,2], Ryan D. Kibler[1,2], Lance Stewart[1,2], David Baker[1,2,4,#]

* Equal contributions
# Correspondence: milles@lmu.de, basile.wicky@bsse.ethz.ch, dabaker@uw.edu

1. Department of Biochemistry, University of Washington, Seattle, WA, USA
2. Institute for Protein Design, University of Washington, Seattle, WA, USA
3. Bioengineering Graduate Program, University of Washington, Seattle, WA, USA
4. Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA

Abstract

Recent advances in de novo protein design have greatly outpaced standard protein biochemistry workflows, and experimental testing has been a bottleneck in the validation of new designs and methodologies. Here, we describe experimental and computational workflows to address the issues of scale, speed and reproducibility of common in vitro protein testing methods, enabling at least an order of magnitude increase in throughput while reducing wetlab time. Semi-Automated Protein Production (SAPP) is a rapid, modular, scalable and cost-effective protocol, enabling up to milligram-scale protein production, and standardized characterization including yield, dispersity, and oligomeric state of hundreds of designs per day, at the cost-equivalent of a few DNA oligos per construct. End-to-end protocol execution takes 48 hours, with about 6 hours spent benchside using mostly standard laboratory equipment. This protocol has become the standard at our institute, providing critical experimental validation for dozens of projects spanning tens of thousands of designs. Since at least 80% of SAPP's total cost comes from synthetic DNA, we also developed a scalable demultiplexing protocol (DMX) to leverage oligo pools as input DNA, providing a further 5-fold reduction in costs, enabling >1000 designs to be purified and characterized in arrayed, clonal format at a cost of $5 per construct. By reframing standard molecular biology practices and orchestrating wetlab workflows with partial automation instead of complex end-to-end robotics, these protocols should be widely adoptable, accelerating protein design.

Main

Recombinant protein expression and purification followed by biochemical characterization is a mainstay of protein engineering. Recent advances in de novo protein design (*1–16*). have greatly outpaced standard protein biochemistry protocols, making experimental validation a bottleneck for protein design efforts. Typical workflows are time, cost, and labor intensive, often requiring multiple weeks for testing a few dozen designs, and lack protocol standardization for consistent data collection, thus precluding large-scale and systematic investigations of the parameter space influencing design success. Pooled assays can test thousands or even millions of designs in parallel (*17–28*), providing a compelling alternative to arrayed testing when large amounts of data are required for a project. However, these assays are usually highly specialized towards measuring one or a few properties (e.g. stability, binding, enzymatic activity), require significant upfront protocol development, and rarely provide information on the manufacturability of individual sequences (expression, solubility, dispersity).

We reasoned that a simple assay-agnostic workflow for the production and initial characterization of hundreds of proteins per day in a standardized and cost-effective manner could significantly accelerate protein design efforts. Moreover, standardizing typical protein characterization data output could enable the use of such experimental data for algorithmic optimization through closed-loop approaches such as active learning (*29–34*). We set out to develop a workflow that generates purified characterized protein starting from gene fragments with greatly reduced execution time, utilizing python scripting to scale and automate key steps in protein purification and characterization using standard laboratory equipment. To reduce the costs further for projects involving characterization of large numbers of designs, we sought to extend the approach to replace gene fragments with much lower cost array synthesized oligonucleotides. We developed workflows for generating purified and characterized protein rapidly from gene fragments (SAPP) and for generating sets of sequence verified fragments from oligonucleotide arrays (DMX) to input into SAPP. These protocols are described in turn below.

Results

Semi-Automated Protein Production

The first step in experimentally testing any computer-generated protein is the design of a suitable DNA construct, typically a plasmid. The most economical format for obtaining arrayed synthetic genes encoding new designed proteins is currently DNA fragments, and standard practice has been to clone and sequence verify these constructs. This adds significant cost per design, and is hard to parallelize due to the need for plating individual cloning reactions. We reasoned that the time required could be greatly reduced by omitting this step, and that by standardizing all steps in the purification pipeline could yield both great increases in efficiency and much more globally comparable data (Fig. 1).

Plasmid generation involves reverse-translating the amino acid sequence and deciding on expression/purification tags to be included. Often, expression constructs are curated and assembled manually, which is laborious and error-prone. We began by developing software for automated design construction, and an efficient and modular cloning strategy. Starting from a set of desired (designed) amino acid sequences, automated reverse-translation optimizes synthesizability and codon adaptation of the corresponding DNA sequences, and adds standardized adapters for downstream cloning into a suite of compatible vectors for multiplexed cloning; the same design can be subcloned into multiple vectors depending on the desired application(s) (e.g. His-tag, Strep-tag, Avi-tag, NanoBiT, Halo-tag, MBP, FPs, see Supplementary Table 1 for the full list of 25 vectors deposited at Addgene). Large constructs (1.5-3 kb) are automatically split into two compatible fragments for one-pot assembly.

Following construct design, DNA fragments are typically manually assembled into plasmids using a variety of cloning techniques (e.g. Gibson Assembly), followed by transformation into cloning strains, plating, picking of clones, DNA extraction, sequencing, and finally re-transformation of validated plasmids into expression strains for protein production. Sequencing is seen as necessary due to errors resulting from synthesis and cloning. This process adds multiple days and significant costs to any protocol, and scales poorly with colony isolation being a bottleneck. We reasoned that by optimizing a standardized cloning approach we could achieve efficiencies sufficient to bypass these steps. Design protein sequences are batch reverse-translated using custom software, optimizing user-defined codon adaptation index and suppressing alternative start sites while ensuring synthesizability by DNA manufacturers. This process appends cloning adapters for compatibility with a range of entry vectors, and automatically splits large genes into two fragments for one-pot assembly. DNA constructs are ordered as linear gene fragments from commercial manufacturers (list price $0.07/bp) and cloned into the desired receiving vector(s) by Golden Gate Assembly (GGA) (*35*, *36*) followed by direct transformation into an *E. coli* expression strain. Target vectors contain a lethal *ccdB* gene (*37*) that is excised and replaced by the design during GGA, ensuring background suppression and thus circumventing the need for colony isolation. The whole process from linear DNA to inoculated cultures is optimized to take 2 hours for up to 192 reactions. While polyclonality due to synthesis or cloning errors is a risk with this approach, sequencing analysis of 929 GGA reactions with insert size below 314 bp showed that 89.3% had a clonal purity ≥ 90% (Fig. 1b). Since protein expression can further bias polyclonal mixtures due to fitness effects resulting from variants, we performed mass spectrometry analysis of 863 purified proteins to analyze the extent of any possible contamination. This revealed the expected mass within ± 1 Da for 90% of all products (777 / 863) (Fig. 1c). These results indicate that our optimized molecular biology workflow is sufficiently reliable to ensure that the correct clone is the dominant construct in the expression pool in most cases, thus circumventing traditional multi-day cloning protocols. While highly efficient, our protocol may still occasionally yield false negative results, and thus is primarily intended for screening purposes.

When chromatographic protein purification and sizing is intended, shake-flask cultures are the norm. However this approach is difficult to scale beyond a few dozen. Moreover, protein expression is often induced chemically (e.g. IPTG) when cultures reach a certain density, which

is inherently challenging to parallelize. We explored the ability of auto-induction media formulation to yield sufficient protein from microplate cultures to streamline multiplexing and reduce human intervention. Cultures are grown in 96-well deepwell plates in autoinduction media at constant temperature, avoiding the need for any intervention between inoculation and harvest (*38*).

When screens are performed in microplate formats, the end-point is usually an SDS-PAGE gel to assess expression, and this information is then used for selecting candidates for scale-up and full chromatographic characterization. While gel quantification is possible, it does not inform on the native properties of proteins (oligomeric state, dispersity), and this two-step approach is also more time-consuming. We reasoned that by optimizing protein purification we could efficiently perform SEC in a preparative fashion on *every* sample, combining characterization and purification into a single process. Moreover, SEC chromatograms are inherently more data-rich, generating more useful information about both failed and successful designs. Following chemical lysis, proteins are purified from their soluble fractions (a protocol for purification from inclusion bodies is included in Supplementary Information), first by affinity chromatography (typically Ni-NTA) in 96-well format, followed directly by size-exclusion chromatography (SEC) on a liquid chromatography system equipped with an autosampler. The process is optimized to take 2 hours from cell harvest to starting SEC, which is run overnight (15 hours for 192 samples, <5 minutes per sample, see Methods).

Chromatographic results are typically analysed in vendors' proprietary software using GUI and manual assessments of individual traces. This process scales poorly beyond a few dozen, and is not systematic. We developed open-source software for analyzing thousands of elution profiles. Chromatograms are automatically analyzed, and the experimental profile for each design (yield, polydispersity, estimated molecular weight, etc… see Supplementary Information) is outputted to a single open format file (Fig. 1d).

Beyond chromatogram analysis, fraction picking for pooling, concentration measurement and normalization are also time-intensive manual tasks. Here we integrate an affordable and programmable liquid handling robot into our analysis code to automate the task. The analysis pipeline automatically identifies peaks for pooling and estimates concentrations by integrating A280 absorbance of SEC chromatograms based on sequence-specific extinction coefficients. This information is then used to automatically generate the code necessary to program a low-cost open-source liquid handling robot for the task of simultaneously pooling fractions and normalizing concentrations.

Outcomes of SAPP are consolidated 96-well plates containing the purified proteins (typical yields; 0.1–0.5 mg at 10–100 μM, Fig. 1e), and a single dataframe containing all the experimental data consolidated in one file. By focusing on an optimized molecular biology workflow and not over-relying on automation, SAPP remains modular, flexible and accessible; a user may elect to process a subset of a 96-well plate, or multiple plates with only minimal additional time (the main bottleneck being SEC, with a capacity of ~300 samples per chromatographic instrument per day). Keeping all steps in a plate format also eliminates

container tracking overhead, and reduces single-use plastic waste. Replicate experiments demonstrate that SAPP is robust (Fig. 1f), enabling quantitative analyses on the data generated by the protocol for machine learning applications.

As a demonstration of the protocol for protein design/engineering efforts we applied SAPP to characterize ProteinMPNN-redesigned sequences of fluorescent proteins (FPs), with the goal to rapidly find more stable FPs (Fig. 2). Crystal structures and AlphaFold2 (*39*) predictions from different FPs were selected as starting points. During sequence design, residues in contact with the chromophore, and residues that are highly conserved (top 50-70% conserved in MSA) were kept fixed, as this was shown to be important when applying ProteinMPNN to the redesign of natural protein (*40*). Overall, sequence identities of designs to WT ranged from 50 to 80% with a Levenshtein distance of up to 50 compared to their respective WT (Fig 2b). We selected 96 constructs for testing, and obtained all experimental results within a week, identifying many designs that have comparable or better yield than the WT (Supplementary Fig. 1). Some exhibit higher thermostability than their WT counterpart (e.g. muGFP), almost all re-designed sequences remained monomeric after boiling (95º C, 1h) as assessed by SEC (Fig. 2d,e; Supplementary Fig. 2), and also retained absorbance, indicating that both the chromophore and protein remained intact, while also showing altered spectra. For some SYFP2 variants, absorbance spectra changed towards a longer Stokes shift than the WT, showing that surface redesign influences the photophysical properties of these fluorescent proteins (Fig. 2d,e). In conclusion, using SAPP enabled 96 FP variants with large edit distances to be characterized within a week, providing detailed biochemical information (yield, dispersity, oligomeric state, absorbance and fluorescence) for each design.

DMX

The highly optimized SAPP protocol has shifted the bottleneck of large-scale testing from protein expression/purification to the cost of DNA synthesis (Supplementary Fig. 3, Supplementary Table 2). While oligo pools provide an economical alternative to gene fragments, they are not immediately suitable for the vast majority of downstream biochemistry workflows, which are often performed in arrayed format. To take advantage of the low cost of oligo pools while retaining the modularity of SAPP, we developed a protocol to convert DNA oligo pools into clonal genes of interest (GOI) using a process that is cost-effective, scalable, and generates sequence-verified constructs. During implementation we focused on: (i) a DNA barcoding strategy that can easily be scaled to thousand of variants; (ii) a demultiplexing strategy that is agnostic to GOI length; (iii) total end-to-end costs (for DNA synthesis, barcoding reagents and sequencing) considerably lower than the cost of individual gene fragments; (iv) seamless integration with SAPP for downstream characterization. Protocols for the demultiplexing of pooled variants using barcoding and sequencing have been implemented previously (*41–45*) but none explicitly addressed all these requirements. The DMX pipeline is a 5-day protocol that leverages an isothermal DNA barcoding strategy and long-read sequencing to produce thousands of sequence-verified clonal cultures ready for processing with SAPP or any other testing modalities that require arrayed DNA (Fig. 3a).

When sequencing variants from oligo pools, variant validation and downstream characterization are often decoupled, which adds time and overhead. Instead we aimed to use a dual purpose pool entry vector, and transformation directly into an expression strain to couple barcoding/sequencing with strain generation. The DMX entry vector was engineered with triple GGA capabilities, to allow both pool entry, barcoding, and optionally GOI excision for further subcloning in an almost scarless manner if desired (Fig. 3b, Supplementary Fig. 4). After amplification from a pool, the DNA library is cloned into the DMX vector by GGA, transformed into expression-ready competent *E. coli* cells by electroporation, and plated on LB-agar for colony generation.

Following isolation into 384-well ECHO plates using a colony picker, GOIs are barcoded. This step is typically achieved by dual-indexed PCR, which limits throughput, and requires large numbers of thermocyclers when processing thousands of samples. To overcome this limitation, we developed an isothermal barcoding strategy based on GGA that ligates barcodes with the GOI directly in bacterial lysate. Our approach uses four sets of position-specific barcodes comprising 24 unique molecular identifiers (UMIs) each, allowing for >330k wells to be tagged from a single 96-well plate of UMIs (24^4) (Supplementary Table 3,4). This expansive barcode capacity exceeds the dual-barcoding approach typically used in individual plates (*41*, *44*) or the positional barcodes used in the bacterial mating system (*45*). Importantly, the barcoding reaction is isothermal, obviating the need for thermocyclers, and our plasmid-derived barcodes can be cheaply produced, making the process simple and scalable.

Once barcoded, samples are pooled, GOIs are mapped to their barcodes using next-generation sequencing. This is typically done with Illumina short-reads, which limits the sequencing length of GOIs (e.g. 300–600 bp amplicon lengths). We leverage nanopore long-read sequencing to overcome this limitation, making barcode mapping length-agnostic, and enabling pools with longer genes to be demultiplexed, such as those obtained by Dropsynth (*46*), paired assembly (*47*, *48*), or multiplexed GGA of long GOIs (*49*–*51*).

An automated bioinformatics pipeline performs quality control on the sequencing results; basecalling the reads from the nanopore sequencer, filtering based on quality, barcode demultiplexing, aligning the demultiplexed reads to their references, and finally generating a consensus sequence (see Supplementary Methods). This returns a non-redundant pick-list of wells, which are re-arrayed by the ECHO in a final step. Outcomes of DMX are arrayed cultures of clonal sequence-verified plasmids in expression-competent *E. coli* cells ready for downstream processing with the SAPP pipeline.

To demonstrate the cost-effectiveness and scalability of the DMX pipeline, we obtained an oligo pool encoding a library of 1,500 de novo designed protein with lengths ranging from 275 to 400 bp. We successfully recovered 78% of the library variants as clonal, sequence-verified constructs by screening 4,608 clones (3x the library complexity), approaching the theoretical maximum recovery rate of 92% for the given library complexity, sampling depth, insert length, synthesis error and synthesis bias reported by Twist (Fig. 3c and Supplementary Fig. 5).

We developed the DMX protocol using automated instruments such as the ECHO liquid handler and colony picker, but it can also be implemented without access to such equipment using manual picking and multichannel pipettes. For large demultiplexing campaigns, the optimal library size is about 2,000 genes, requiring the sampling of about 8,000 colonies (20 x 384 well plates) to achieve >95% recovery (Supplementary Fig. 5). For smaller campaigns, libraries of 500 genes can be demultiplexed by manually picking 1500 colonies in order to maintain the same recovery rate. This approach is five and eight times more cost-effective (including reagent and sequencing costs) than ordering 500 and 2000 individual gene fragments, respectively (Fig. 3d, Supplementary Table 5), and ensures clonality of tested variants, while only adding five days to the pipeline. Thus it is ideally suited for large design campaigns where the time overhead is absorbed over many constructs, while the polyclonal approach from SAPP is better suited for smaller campaigns (~100-200 designs) where faster turnaround time is desired.

Discussion

The modular SAPP workflow dramatically accelerates the experimental characterization of designed proteins. SAPP combines throughput-optimized molecular biology steps, an automated analysis pipeline, and optionally open-source robotics to purify and characterize hundreds of designs per day in array format with only a few hours spent at the bench, with typical yields being sufficient for most downstream biochemical assays (e.g. binding, enzymatic activity, thermostability, cell assays, electron microscopy). Complete end-to-end execution takes 48 hours (excluding commercial DNA synthesis), enabling rapid iteration cycles at a total cost per design equivalent to a few DNA oligos (for a detailed calculation see Supplementary Information). Two key aspects enable the speed and efficiency of SAPP. First, the use of modular cloning and a background-suppressing cassette makes direct-to-expression from commercial gene fragments sufficiently reliable for screening applications, bypassing the standard procedure of clone isolation and sequencing. Second, streamlining the SEC workflow allows for data-rich information generation on every design, which can be automatically analysed and consolidated into an open format for quantitative analyses between design campaigns. By optimizing common experimental techniques over complex robotics, this protocol can be adopted by most labs.

To further reduce the cost of per-design testing, we developed the DMX pipeline to convert DNA oligo pools into arrayed cultures of sequence-verified plasmids. The scalability of DMX is primarily driven by a combinatorial isothermal barcoding framework, which eliminates the need for thermocyclers, and by nanopore long-read sequencing, which offers a low financial entry point and is compatible with applications involving longer GOIs. The protocol achieves high recovery rates of sequence-verified constructs, and integrates seamlessly with downstream workflows like SAPP.

These protocols have already been used to experimentally produce and characterize tens of thousands of de novo designed proteins at the Institute for Protein Design, ranging from small mini-protein binders against cell surface receptor of therapeutic interest, de novo enzymes, and symmetric protein assemblies hundreds of kDa in size (*1*, *4*, *5*, *8*, *10*, *40*, *52–65*). These

workflows both democratize and accelerate protein design, making large-scale experimental testing more affordable and accessible. We envision that the integration of scale, standardization, and speed will contribute to the development of new protein design models informed by experimental data, as well as enabling active learning approaches in protein design. Collectively, SAPP and DMX present an experimental counterpart matched to the rapid pace and scale of recent advances in computational protein design.

Figures

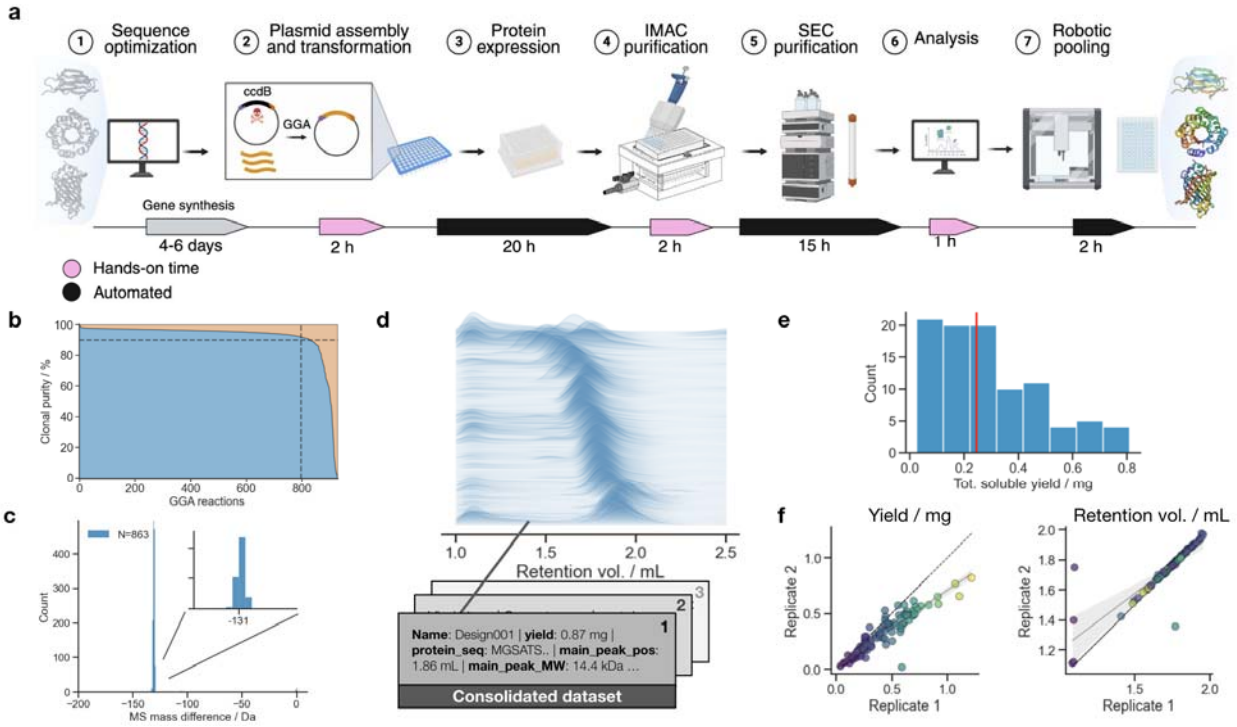

Fig 1. SAPP pipeline and experimental validation. (a) Schematic representation of the SAPP pipeline from cloning to protein expression and purification. (b) NGS data for 929 SAPP samples. Sequencing results of the polyclonal mixtures show that the majority of sequences are correct (blue), with only a small fraction of reactions showing incorrect sequences (orange). Dashed line intersection represents 90% clonal purity. (c) LC-MS data on samples purified by SAPP, showing correct masses, with most proteins lacking their N-terminal Met (-131 Da) as expected in *E. coli* (N=863). (d) Representative results from a SAPP run with 96 samples, showing SEC chromatograms ordered by hierarchical clustering. (e) Histogram showing typical soluble yield in mg for designed proteins purified with SAPP. (f) Parity plots of biological replicates showing total soluble yield in mg (left), and elution of the main peak in mL (right), demonstrating the reproducibility of SAPP.

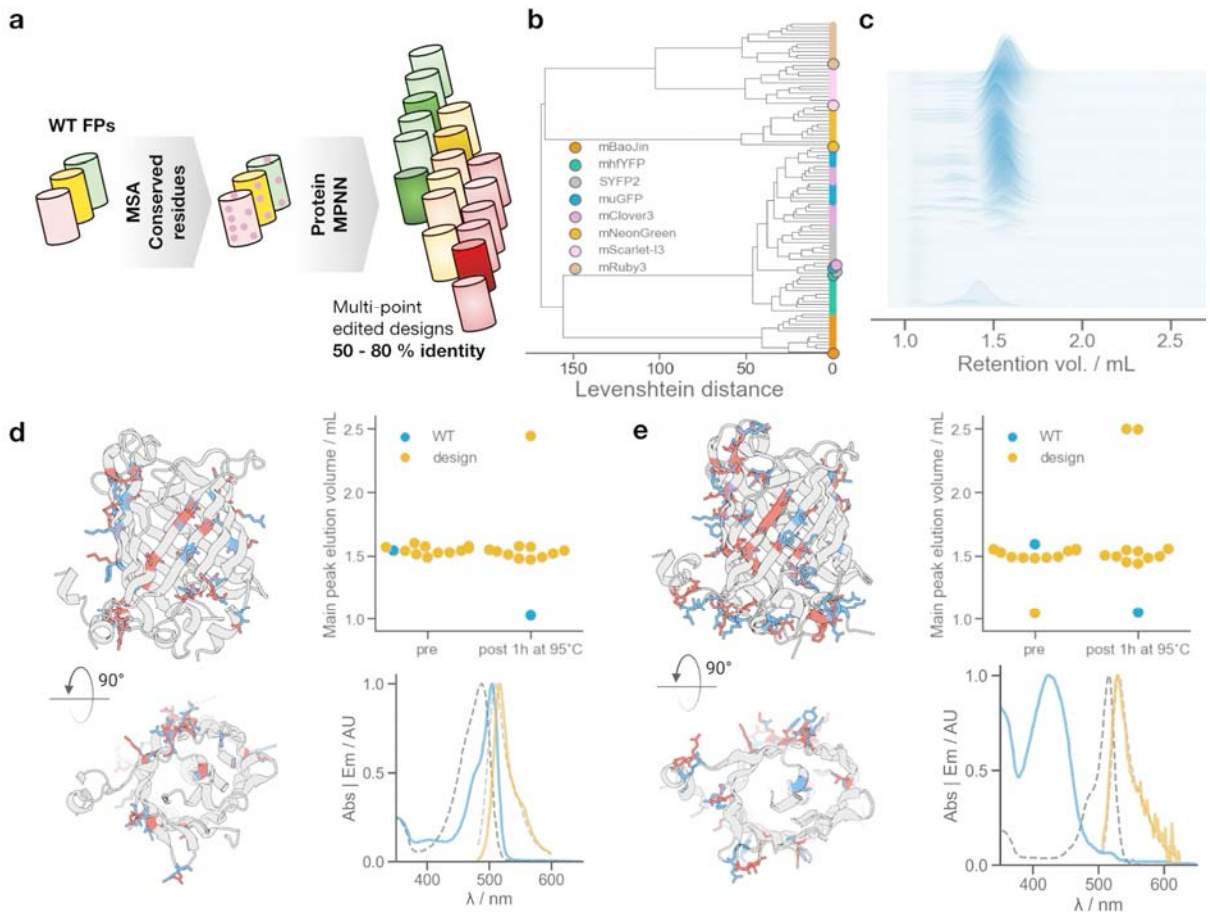

Fig 2. SAPP for Fluorescent Protein Sequence Redesign with ProteinMPNN. (a) Non-conserved surface residues of fluorescent proteins were redesigned using ProteinMPNN. (b) A dendrogram, based on Levenshtein distance, displays 96 characterized sequences from SAPP, color-coded by FP template. Wild-type sequences are marked with circles. (c) Aggregate SEC traces of the 96 designs reveal distinct elution peaks near the anticipated molecular weight for over half of the variants. (d, e) For designs 4389 and 4468, the redesigned amino acid positions (red) are highlighted on the structures of their wild-type templates, muGFP (d) and SYFP2 (e), with the original amino acids shown in blue. The top plots show elution profiles post-purification and after a 1-hour incubation at 95°C for all variants of that parental sequence. The bottom plots present the absorbance and emission spectra of the designed variants shown on the left (colored lines) and their corresponding wild-types (dashed lines), indicating changes in photophysical properties.

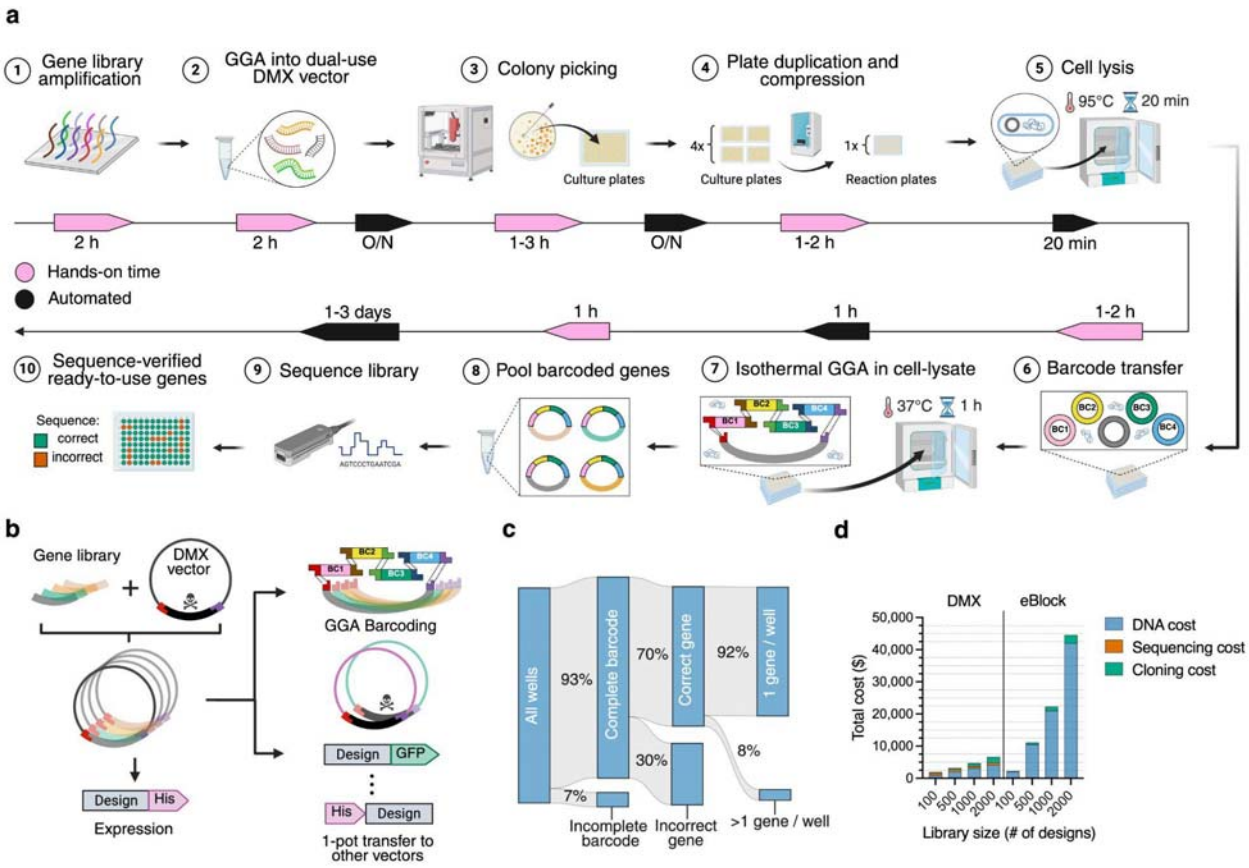

Fig 3. DMX pipeline. (a) Schematic representation of the DMX pipeline from oligo pool library to sequence-verified ready-to-use clonal bacterial cultures. (1) DNA library is amplified from the chip; (2) library is cloned into DMX vector using GGA, transformed and plated; (3) colonies are picked into culture plates; (4) culture plates are duplicated and compressed 4-to-1 into reaction plates using ECHO; (5) cultures in reaction plates are heat-lysed at 98 °C for 20 minutes; (6) barcodes are transferred into each well of the reaction plates; (7) GGA reagents are transferred and barcodes and GOI are assembled by incubating reaction plates at 37 °C for 1 hour; (8) pool all GGA products; (9) sequence barcoded library using ONT; (10) re-array sequence-verified clonal variants into new plates. Pink and black arrow boxes represent hands-on time and automated steps, respectively. (b) Schematic representation of multi-use vector cloning for pool entry (BsmBI), barcoding (BsaI), and GOI excision (BsaI) if desired for further subcloning into other vectors. Skull represents the *ccdB* gene. (c) Sankey diagram showing the efficiency of the DMX pipeline (N=4,608 wells). (d) Cost comparison between sourcing GOIs using DMX from chip or arrayed DNA fragments, broken down into DNA synthesis, sequencing, and cloning expenses.

Methods

The list of equipment, items and components necessary to the execution of the protocol (including suppliers and catalog numbers) can be found in Supplementary Table 6 and 8 for SAPP and DMX respectively. The composition of SAPP buffers can be found in Supplementary Table 7. Chemicals were purchased from Millipore-Sigma or Fisher Scientific unless stated otherwise. DNA fragments were purchased as eBlocks from Integrated DNA Technologies (IDT). Oligo pools were purchased from Twist Bioscience.

DNA fragment generation

Protein sequences are first reverse-translated to DNA sequences using a custom python script that takes a FASTA file or a list of PDBs as input. Codon optimization is performed with DNA Chisel (*66*) using a mixed objective function that optimizes user-defined codon adaptation index (CAI), minimizes repeat sequences, avoids BsaI restriction endonuclease recognition sites, and suppresses alternative start sites. The script also ensures synthesizability on the fly by using IDT's SciTools Plus API, and if the original sequence cannot be made, the objective function is gradually modified by up-weighting the repeat sequence cost parameter until synthesizability is achieved. If the reverse-translated sequence is longer than a user-specified length (default: 1500 bp, IDT's eBlock maximum), the sequence is automatically split into two fragments for one-pot Golden Gate Assembly (GGA) (*35*). The 4 bp fragment overhang is chosen from a predefined list that maximizes orthogonality to the entry vector GGA sites (5'->3': AGGA and TTCC) in order to ensure assembly fidelity (*67*). Finally, overhangs for GGA containing BsaI recognition sites are appended to the 3' and 5' ends, and random filler sequences added to the outer parts if the sequence is less than 300 bp (IDT's eBlock length minimum). The script outputs a formatted spreadsheet that is directly uploaded to IDT for ordering as eBlocks ($0.07/bp, 1-3 business days turnaround), and an ECHO transfer protocol for automated cloning (see below). DNA fragments are ordered resuspended in IDTE buffer (10 mM Tris, 0.1 mM EDTA, pH 8.0) in either 96-well PCR plates (for manual cloning), or ECHO-qualified 384-well plates (for robotic-assisted cloning). The whole process is fully automated, and going from a list of amino acid sequences or PDBs to placing the gene fragment order only takes minutes.

Cloning

Entry vectors are propagated and produced in bulk from *ccdB*-resistant NEB Stable *E. coli* chemically competent cells (NEB). Following transformation according to the manufacturer's protocol (see Supplementary Information), the transformants are grown overnight at 37 °C on LB-Agar containing 50 µg/mL of kanamycin sulfate. Single colonies are used to inoculate 50 mL of liquid LB cultures (in 250 mL baffled conical flasks) containing 50 µg/mL of kanamycin sulfate, and grown overnight at 37 °C with shaking at 250 rpm. Plasmids are purified using a ZymoPURE II Plasmid Midiprep Kit (Zymo Research) according to the manufacturer's protocol, and verified by full plasmid sequencing (Plasmidsaurus). Typical yields are 50–100 µg, which is enough for about 2,000–4,000 GGA reactions.

GGA reactions are assembled with an ECHO 525 acoustic liquid handling robot (Beckman Coulter) by using the transfer protocol outputted during DNA fragment generation. A master mix

containing BsaI-HFv2 (NEB), T4 DNA ligase (NEB), T4 DNA ligase buffer (NEB) and the entry vector is prepared first, and the ECHO is used to combine master mix and DNA fragments into 96-well PCR plates (1 µL final volumes). Reactions are incubated at 37 °C for 15 minutes, followed by 5 minutes at 60 °C. GGA products are directly transformed into an *E. coli* expression strain by adding 6 µL of BL21(DE3) chemically competent cells (NEB) to the 1 µL reactions. Transformations are performed in 96-well PCR plates (see Supplementary Information), and the transformed cells are used to inoculate expression cultures directly (transformation efficiency = 858 ± 443 colony forming units (CFU) (mean ± standard deviation, N = 12)). Going from linear DNA to inoculated cultures takes about 2 hours for 192 GGA reactions, and only requires an extra 15 minutes for every additional 96-well plate.

If an ECHO robot is not available, we recommend performing 5 µL cloning reactions; first by aliquoting 4 µL of master mix into each well, followed by the addition of 1 µL of eBlock using a multichannel pipette (5 µL total volume). After GGA, 1 µL is used to transform 6 µL of cells as described above.

Protein expression

Protein expression is performed in round bottom 96 deepwell plates filled with 1 mL of simplified auto-induction media (*38*) (TB II, MP Biomedicals, supplemented with glycerol [5 g/L], glucose [0.5 g/L], lactose [2 g/L], $MgSO_4$ [2 mM], and kanamycin sulfate [50 µg/mL]). Transformation reactions are split 4-fold and used to inoculate 4 x 1 mL of expression cultures directly, followed by incubation at 37 °C for *at least* 20 hours at 1,000 rpm on a 2.5 mm orbital shaker (NB smaller orbital paths appear to negatively impact expression, most likely due to reduced aeration (*68*), which likely impacts the glucose to lactose metabolic switch necessary for auto-induction).

Affinity chromatography

Cells are harvested in 96 deepwell plates by centrifugation (4,000 x g, 5 minutes), the media discarded, and cells resuspended in lysis buffer (B-PER, Thermo Fisher Scientific, supplemented with lysozyme [0.1 mg/mL], PMSF [1 mM], Benzonase [25 U/mL], 100 µL per 1 mL of culture-equivalent pellet). Lysis is left to proceed for 15 minutes at 37 °C under agitation (1,000 rpm) before pooling the lysates (4 x 100 µL) into one 96 deepwell plate and clearing debris by centrifugation (4,000 x g, 15 minutes). Proteins are purified from the soluble fraction by binding to Ni-NTA resin (50 µL of resin bed per well, added to 96-well fritted plates (25 µm PE frit, Agilent Technologies)), followed by three wash cycles using a plate vacuum manifold (3 x 400 µL, 20 mM Tris, 300 mM NaCl, 25 mM imidazole, pH 8.0). Proteins are eluted from the resin by addition of 200 µL of elution buffer (20 mM Tris, 300 mM NaCl, 500 mM imidazole, pH 8.0), and sterile-filtered by centrifugation into a 96-well filter plate (0.22 µm pore-size) mounted on a receiver plate. The whole process process is optimized to take under 2 hours from cell harvest to injection-ready eluates for up to two 96-well plates, with only minimal extra time necessary for additional plates.

Size-exclusion chromatography

Filter-sterilized eluates are purified/analyzed by size-exclusion chromatography (SEC) into PBS on either a Superdex 75 (70 kDa MWCO) or Superdex 200 (600 kDa MWCO) column (5/150 with 3 mL column bed volume, Cytvia) by injecting 100 µL at a flow-rate of 0.65 mL / minutes (<5 minutes per sample, ~15 hours for 192 samples). Chromatography is performed either on a Agilent 1260 Infinity II Bio-inert system equipped with multisampler and fraction collector (fractionation into 384-well plates), or a Cytvia KTA Pure system equipped with an autosampler and using a modified flow-path (fractionation into 96-well plates, see Supplementary Information for details). Chromatograms are batch-analyzed by a custom python script (see below) which also generates code to program an OT-2 liquid handling robot (Opentrons) for cherry-picking, pooling and normalizing fractions. The end-point of the protocol is a set of 96-well plates containing SEC-purified and concentration-normalized proteins (typically a few hundreds microliters at single to double-digit micromolar concentrations).

Data processing

The integrated data-analysis pipeline starts by generating plasmid maps from the ordered DNA fragment sequences and specified entry vectors. The ORF is used to generate the expression sequence (designed sequence plus vector-specific tags), and calculate protein parameters (molecular weight, molar extinction coefficient, charge, isoelectric point, etc). Batch-exported chromatograms are analyzed using NumPy (*69*), SciPy (*70*), and Pandas (*71*, *72*), to detect peaks in the protein absorbance signal at 280 nm and convert their retention volume to approximate molecular weight compared to calibration curve assembled from standards (LMW [S75], or HMW [S200], Cytvia). In addition, by integrating the absorbance signal to determine protein concentration in each fraction, the following information can be extracted for each design: total soluble yield, polydispersity, estimated molecular weight and aggregation state of each peak, desired fractions to pool. The data is exported into a standardized open format dataframe (HDF5 (*73*)).

Mass spectrometry

Intact mass spectra were obtained by reverse-phase LC/MS on an Agilent G6230B TOF using an AdvanceBio RP-Desalting column with a fast gradient of 90:10 (A:B) to 5:95 (A:B) over 2 minutes (A: H2O with 0.1% Formic Acid, B: Acetonitrile with 0.1% Formic Acid), and subsequently deconvoluted with Bioconfirm using a total entropy algorithm.

Next-generation sequencing

A retrospective analysis of designs ordered as DNA fragments (eBlocks, IDT), assembled, cloned and transformed directly as described above was performed to assess clonal purity (Fig. 1b). Stabs from 929 glycerol stocks (made from LB out-growth cultures with antibiotics) were pooled, and the inserts amplified by PCR using Q5 (NEB), followed by purification using AMPure XP magnetic beads (Beckman Coulter) and analyzed on a 4200 TapeStation (Agilent). Paired-end sequencing was performed using NextSeq2000 (Illumina), basecalled and filtered for quality. Reads R1 and R2 were extracted using SAMtools (*74*), followed by pairing using PEAR-0.9.11 with options -j 20 -p 1.0 (*75*). Index sequences of 929 reference sequences were

built and aligned to the PEAR output using Bowtie2 (*76*), generating 31 million aligned reads, which were then converted to BAM, sorted and indexed using SAMtools. From the CIGAR strings, clonal purity scores were calculated by dividing the number of perfect matches by the total number of filtered reads for each reference sequence using a custom python script.

Fluorescent protein sequence redesign

Fluorescent protein (FP) sequences were generated using the protocol reported by Sumida et al. (*14*) In brief, WT FP sequences were extracted from FPbase (*77*), and their structures predicted with AlphaFold2 using MSAs generated from UniRef30_2022_02 using HHblits (*78*). Per position conservation was assessed from the MSAs, and either 50% or 70% of the most preserved residues were kept fixed during ProteinMPNN sequence redesign. Additionally all chromophore forming residues, and all core-facing residues (as assessed by a SASA < 2.0 Å^2 using pyrosetta (*79*, *80*)) were also fixed. ProteinMPNN was run at T=0.01, 0.1, 0.2, 0.3, 0.5, and designed sequences predicted with AlphaFold2 using a shallow MSA of up to 15 sequences. Designs were selected based on the following metrics: AF2_plddt_mean > 0.8, AF2_ptm > 0.80, AF2_pae_mean < 12.0, AF2_tm_score to WT > 0.7, AF2_tm_rmsd to WT < 1.5, AF2_sap_per_res < 0.4. Using TM-align (*81*), SAP score (*82*), and AF2 confidence metrics.

DMX barcode design

Each of the four position-specific barcode sets comprise 24 UMIs. These 25-bp UMI sequences were taken from (*83*), and further filtered for: pairwise Hamming distance >5, absence of homo-polymers >4 bp, and absence of common restriction enzyme sites (AgeI, BamHI, BbsI, BsaI, BsmBI, BtgZI, EcoRI, HindIII, KpnI, NheI, NotI, PacI, PaqCI, PstI, PvuI, SacI, SalI, SapI, SmaI, XbaI, XhoI). Primer sequences were then appended to each UMI (balanced G/C % for 62 °C Tm), GGA overhangs and BsaI cut sites (Supplementary Table 4). Barcodes were ordered subcloned from IDT as MiniGene 25-500 bp inserted into pUCIDT (Amp) Golden Gate vector.

Library amplification from oligo pools

Oligo pools encoding GOIs were ordered from Twist Bioscience, and amplified by PCR using KAPA HiFi HotStart ReadyMix (Roche). To reduce biases, amplifications were monitored by real-time PCR using EvaGreen Dye (Biotium), and stopped before plateauing, as described in Klein et al., 2016 (*47*). Amplified libraries were purified by gel electrophoresis (2% agarose) and extracted using Zymoclean Gel DNA Recovery Kits (Zymo Research), followed by quantification with Qubit (Invitrogen).

Cloning into DMX vector

Amplified libraries were cloned into the DMX vector using BsmBI-v2 NEBridge Golden Gate Assembly kit (NEB) according to the manufacturer's protocol. The GGA product was purified using DNA Clean & Concentrator-5 (Zymo Research) quantified using Qubit (Invitrogen), and electroporated into *E. Cloni* EXPRESS BL21(DE3) electrocompetent cells (Lucigen) according to the manufacturer's protocol. Electroporated cells were recovered at 37 °C for 1 hour in a shaking incubator at 250 rpm.

Colony picking

Aliquots from recovered cells after electroporation were serially diluted and grown overnight at 37 °C on LB-Agar plates containing 100 µg/mL of carbenicillin; the rest of the non-diluted recovered cells were stored at 4 °C overnight. The next day, colony counting provided an estimate of CFUs in the original library, which was used to calculate the volume needed from the kept recovered cells to obtain 2,500 colonies on Bioassay plates (Corning). The cells were mixed with SOC medium and plated onto BioAssay plates containing 100 µg/ml of carbenicillin, and grown overnight at 37 °C. The next day, colonies were picked using a QPix XE Microbial Colony Picker (Molecular Devices) into ECHO-qualified 384-well plates (Beckman Coulter) containing 60 µL of low-salt LB liquid medium per well (tryptone [10 g/L], NaCl [5 g/L], yeast extract [5 g/L], carbenicillin [100 mg/L]). Plates were sealed with Breathe Easier plate seals (Fisher Scientific), and grown overnight in a shaking incubator (37 °C, 1000 rpm).

GGA barcoding in cell lysate

Confluent cultures were transferred (2 µL) and compressed from four ECHO-qualified 384-well plates into one polypropylene 1536-well plate (Greiner) using an ECHO 525 acoustic liquid handling robot (Beckman Coulter). The 1536-well plates were sealed with Axygen® Foil Plate Seal (Fisher Scientific), inserted into a vacuum sealed bag, and heat-inactivated at 98 °C in a water bath for 30 minutes. Barcode combinations (2 ng per UMI, 8 ng in total) were transferred from a DNA barcode source plate to each well of the heat-inactivated 1536-well cell-lysate plates using the ECHO. A GGA mastermix containing BsaI-HFv2 (NEB), Salt-T4 DNA ligase (M0467L, NEB), and T4 DNA ligase buffer (NEB) was transferred to each well with the ECHO (0.5 µL). Reaction plates were incubated at 37 °C for 1 hour, followed by 5 minutes at 60 °C. The 1536-well plates were inverted and centrifuged at 200 x g to pool all the wells, and GGA products (i.e. position-specific barcodes ligated with the GOI from each well) were isolated using Zyppy Plasmid Miniprep kit (Zymo Research) according to the manufacturer's protocol. Pools of barcoded GOIs were amplified by PCR using common primers and KAPA HiFi HotStart ReadyMix (Roche), and the product purified by gel electrophoresis (1% agarose), extracted with Zymoclean Gel DNA Recovery kit (Zymo Research), and quantified by Qubit (Invitrogen).

Nanopore Long-read sequencing

Pools of barcoded GOIs were prepared for sequencing using Ligation Sequencing Kit V14 (SQK-LSK114, Oxford Nanopore Technologies) and loaded onto MinION Flow Cell according to the manufacturer’s protocol. Depending on library size, sequencing can be run for <72 hours.

DMX bioinformatics pipeline

Output raw reads from ONT nanopore sequencing (.pod5) were basecalled to bam files using Dorado-0.9.1. Reads were converted and merged from BAM into Fastq files using SAMtools-1.21 (*74*) and filtered based on size and quality score using Chopper-0.9.0 (*84*) and Nanoq-0.10.0 (*85*). Filtered reads were demultiplexed into individual files based on barcode combination using Cutadapt-4.9 (*86*). Alignment to reference, and consensus sequence generation were done with Minimap2-2.28 (*87*) and samtools-1.21. A python script was used to

generate a list of correct unique designs and generate pick lists to re-array ECHO culture plates into new plates for downstream processing (see Supplementary protocol DMX).

Data and code availability

Scripts and notebooks can be found at https://github.com/bwicky/SAPP_DMX. A permanent archive of the code and data will be made available at Zenodo upon final publication.

Conflicts of interest

David Baker, Basile Wicky, Lukas Milles are inventors of a patent application (US 18/658657) submitted by the University of Washington for the SAPP pipeline. David Baker, Jason Qian, Basile Wicky, Lukas Milles, Amir Motmaen are filing a provisional patent application for the DMX pipeline.

Acknowledgements

Mohamad Abedi and Florian Praetorius for helpful discussions, Luki Goldschmidt and Patrick Vecchiato for computational infrastructure, Kandise VanWormer and Hernan Nunez-Ortega for wetlab infrastructure, Wei Chen for help with setting up NGS. This work was supported by a Human Frontier Science Program Cross Disciplinary Fellowship (LT000395/2020-C, to L.F.M.), an EMBO Non-Stipendiary Fellowship (ALTF 1047-2019, to L.F.M.), a Schmidt Science Fellowship to J.Q, an EMBO long-term fellowship (ALTF 139-2018, to B.I.M.W.), the Open Philanthropy Project Improving Protein Design Fund (D.B.), the Howard Hughes Medical Institute (D.B.), the Audacious Project at the Institute for Protein Design (D.B.)

Contributions

Conceptualization: B.I.M.W., L.F.M., J.Q., A.M., and D.B.
Methodology: B.I.M.W., L.F.M., J.Q. and D.B.
Software: B.I.M.W., L.F.M., J.Q., and R.D.K.
Validation: B.I.M.W., L.F.M., X.L., and J.Q.
Formal analysis: B.I.M.W., L.F.M., J.Q. and X.L.
Investigation: B.I.M.W., L.F.M., and J.Q.
Data curation: B.I.M.W., L.F.M., and J.Q.
Writing – original draft: B.I.M.W., L.F.M., and J.Q.
Writing – review & editing: B.I.M.W., L.F.M., J.Q., and D.B.
Visualization: B.I.M.W., L.F.M., and J.Q.
Supervision: D.B.
Project administration: B.I.M.W., L.F.M., J.Q., and D.B.
Funding acquisition: B.I.M.W., L.F.M., J.Q., L.S., and D.B.
All authors reviewed the paper.

References

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

2. J. B. Ingraham, M. Baranov, Z. Costello, K. W. Barber, W. Wang, A. Ismail, V. Frappier, D. M. Lord, C. Ng-Thow-Hing, E. R. Van Vlack, S. Tie, V. Xue, S. C. Cowles, A. Leung, J. V. Rodrigues, C. L. Morales-Perez, A. M. Ayoub, R. Green, K. Puentes, F. Oplinger, N. V. Panwar, F. Obermeyer, A. R. Root, A. L. Beam, F. J. Poelwijk, G. Grigoryan, Illuminating protein space with a programmable generative model. *Nature* **623**, 1070–1078 (2023).

3. A. E. Chu, J. Kim, L. Cheng, G. El Nesr, M. Xu, R. W. Shuai, P.-S. Huang, An all-atom protein generative model. *Proc. Natl. Acad. Sci.* **121**, e2311500121 (2024).

4. J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, D. Baker, Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **0**, eadd2187 (2022).

5. B. I. M. Wicky, L. F. Milles, A. Courbet, R. J. Ragotte, J. Dauparas, E. Kinfu, S. Tipps, R. D. Kibler, M. Baek, F. DiMaio, X. Li, L. Carter, A. Kang, H. Nguyen, A. K. Bera, D. Baker, Hallucinating symmetric protein assemblies. *Science* **378**, 56–61 (2022).

6. I. Anishchenko, S. J. Pellock, T. M. Chidyausiku, T. A. Ramelot, S. Ovchinnikov, J. Hao, K. Bafna, C. Norn, A. Kang, A. K. Bera, F. DiMaio, L. Carter, C. M. Chow, G. T. Montelione, D. Baker, De novo protein design by deep network hallucination. *Nature* **600**, 547–552 (2021).

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

8. A. Lauko, S. J. Pellock, K. H. Sumida, I. Anishchenko, D. Juergens, W. Ahern, J. Jeung, A. Shida, A. Hunt, I. Kalvet, C. Norn, I. R. Humphreys, C. Jamieson, R. Krishna, Y. Kipnis, A. Kang, E. Brackenbrough, A. K. Bera, B. Sankaran, K. N. Houk, D. Baker, Computational design of serine hydrolases. *Science* **0**, eadu2454 (2025).

9. M. Pacesa, L. Nickel, J. Schmidt, E. Pyatova, C. Schellhaas, L. Kissling, A. Alcaraz-Serna, Y. Cho, K. H. Ghamary, L. Vinué, B. J. Yachnin, A. M. Wollacott, S. Buckley, S. Georgeon, C. A. Goverde, G. N. Hatzopoulos, P. Gönczy, Y. D. Muller, G. Schwank, S. Ovchinnikov, B. E. Correia, BindCraft: one-shot design of functional protein binders. bioRxiv [Preprint] (2024). https://doi.org/10.1101/2024.09.30.615802.

10. R. Verkuil, O. Kabeli, Y. Du, B. I. Wicky, L. F. Milles, J. Dauparas, D. Baker, S. Ovchinnikov, T. Sercu, A. Rives, Language models generalize beyond natural proteins. bioRxiv [Preprint]

(2022). https://doi.org/10.1101/2022.12.21.521521.

11. N. Ferruz, S. Schmidt, B. Höcker, ProtGPT2 is a deep unsupervised language model for protein design. *Nat. Commun.* **13**, 4348 (2022).

12. A. Madani, B. Krause, E. R. Greene, S. Subramanian, B. P. Mohr, J. M. Holton, J. L. Olmos, C. Xiong, Z. Z. Sun, R. Socher, J. S. Fraser, N. Naik, Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* **41**, 1099–1106 (2023).

13. T. Hayes, R. Rao, H. Akin, N. J. Sofroniew, D. Oktay, Z. Lin, R. Verkuil, V. Q. Tran, J. Deaton, M. Wiggert, R. Badkundri, I. Shafkat, J. Gong, A. Derry, R. S. Molina, N. Thomas, Y. A. Khan, C. Mishra, C. Kim, L. J. Bartie, M. Nemeth, P. D. Hsu, T. Sercu, S. Candido, A. Rives, Simulating 500 million years of evolution with a language model. *Science* **387**, 850–858 (2025).

14. D. Akpinaroglu, K. Seki, A. Guo, E. Zhu, M. J. S. Kelly, T. Kortemme, Structure-conditioned masked language models for protein sequence design generalize beyond the native sequence space. bioRxiv [Preprint] (2023). https://doi.org/10.1101/2023.12.15.571823.

15. C. Wang, Y. Qu, Z. Peng, Y. Wang, H. Zhu, D. Chen, L. Cao, Proteus: pioneering protein structure generation for enhanced designability and efficiency. bioRxiv [Preprint] (2024). https://doi.org/10.1101/2024.02.10.579791.

16. B. Huang, Y. Xu, X. Hu, Y. Liu, S. Liao, J. Zhang, C. Huang, J. Hong, Q. Chen, H. Liu, A backbone-centred energy function of neural networks for protein design. *Nature* **602**, 523–528 (2022).

17. K. Tsuboyama, J. Dauparas, J. Chen, E. Laine, Y. Mohseni Behbahani, J. J. Weinstein, N. M. Mangan, S. Ovchinnikov, G. J. Rocklin, Mega-scale experimental analysis of protein folding stability in biology and design. *Nature* **620**, 434–444 (2023).

18. G. J. Rocklin, T. M. Chidyausiku, I. Goreshnik, A. Ford, S. Houliston, A. Lemak, L. Carter, R. Ravichandran, V. K. Mulligan, A. Chevalier, C. H. Arrowsmith, D. Baker, Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **357**, 168–175 (2017).

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

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

21. W. C. Boldridge, A. Ljubetič, H. Kim, N. Lubock, D. Szilágyi, J. Lee, A. Brodnik, R. Jerala, S.

Kosuri, A multiplexed bacterial two-hybrid for rapid characterization of protein–protein interactions and iterative protein design. *Nat. Commun.* **14**, 4636 (2023).

22. Á. J. R. Ferrari, S. M. Dixit, J. Thibeault, M. Garcia, S. Houliston, R. W. Ludwig, P. Notin, C. M. Phoumyvong, C. M. Martell, M. D. Jung, K. Tsuboyama, L. Carter, C. H. Arrowsmith, M. Guttman, G. J. Rocklin, Large-scale discovery, analysis, and design of protein energy landscapes. bioRxiv [Preprint] (2025). https://doi.org/10.1101/2025.03.20.644235.

23. A. Beltran, X. Jiang, Y. Shen, B. Lehner, Site-saturation mutagenesis of 500 human protein domains. *Nature* **637**, 885–894 (2025).

24. C. Weng, A. J. Faure, A. Escobedo, B. Lehner, The energetic and allosteric landscape for KRAS inhibition. *Nature* **626**, 643–652 (2024).

25. A. J. Faure, J. Domingo, J. M. Schmiedel, C. Hidalgo-Carcedo, G. Diss, B. Lehner, Mapping the energetic and allosteric landscapes of protein binding domains. *Nature* **604**, 175–183 (2022).

26. A. Baryshev, A. La Fleur, B. Groves, C. Michel, D. Baker, A. Ljubetič, G. Seelig, Massively parallel measurement of protein–protein interactions by sequencing using MP3-seq. *Nat. Chem. Biol.* **20**, 1514–1523 (2024).

27. A. J. Faure, A. Martí-Aranda, C. Hidalgo-Carcedo, A. Beltran, J. M. Schmiedel, B. Lehner, The genetic architecture of protein stability. *Nature* **634**, 995–1003 (2024).

28. R. Obexer, A. Godina, X. Garrabou, P. R. E. Mittl, D. Baker, A. D. Griffiths, D. Hilvert, Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase. *Nat. Chem.* **9**, 50–56 (2017).

29. P. A. Romero, A. Krause, F. H. Arnold, Navigating the protein fitness landscape with Gaussian processes. *Proc. Natl. Acad. Sci.* **110**, E193–E201 (2013).

30. Z. Wu, S. B. J. Kan, R. D. Lewis, B. J. Wittmann, F. H. Arnold, Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc. Natl. Acad. Sci.* **116**, 8852–8858 (2019).

31. J. T. Rapp, B. J. Bremer, P. A. Romero, Self-driving laboratories to autonomously navigate the protein fitness landscape. *Nat. Chem. Eng.* **1**, 97–107 (2024).

32. N. C. Frey, I. Hötzel, S. D. Stanton, R. Kelly, R. G. Alberstein, E. Makowski, K. Martinkus, D. Berenberg, J. Bevers, T. Bryson, P. Chan, A. Czubaty, T. D'Souza, H. Dwyer, A. Dziewulska, J. W. Fairman, A. Goodman, J. Hofmann, H. Isaacson, A. Ismail, S. James, T. Joren, S. Kelow, J. R. Kiefer, M. Kirchmeyer, J. Kleinhenz, J. T. Koerber, J. Lafrance-Vanasse, A. Leaver-Fay, J. H. Lee, E. Lee, D. Lee, W.-C. Liang, J. Y.-Y. Lin, S. Lisanza, A. Loukas, J. Ludwiczak, S. P. Mahajan, O. Mahmood, H. Mohammadi-Peyhani, S. Nerli, J. W. Park, J. Park, S. Ra, S. Robinson, S. Saremi, F. Seeger, I. Sinha, A. M. Sokol, N. Tagasovska, H. To, E. Wagstaff, A. Wang, A. M. Watkins, B. Wilson, S. Wu, K. Zadorozhny, J. Marioni, A. Regev, Y. Wu, K. Cho, R. Bonneau, V. Gligorijević, Lab-in-the-loop therapeutic antibody design with deep learning. bioRxiv [Preprint] (2025). https://doi.org/10.1101/2025.02.19.639050.

33. G. M. Landwehr, J. W. Bogart, C. Magalhaes, E. G. Hammarlund, A. S. Karim, M. C. Jewett, Accelerated enzyme engineering by machine-learning guided cell-free expression. *Nat. Commun.* **16**, 865 (2025).

34. J. Yang, R. G. Lal, J. C. Bowden, R. Astudillo, M. A. Hameedi, S. Kaur, M. Hill, Y. Yue, F. H. Arnold, Active learning-assisted directed evolution. *Nat. Commun.* **16**, 714 (2025).

35. C. Engler, R. Kandzia, S. Marillonnet, A One Pot, One Step, Precision Cloning Method with High Throughput Capability. *PLOS ONE* **3**, e3647 (2008).

36. S. Marillonnet, R. Grützner, Synthetic DNA Assembly Using Golden Gate Cloning and the Hierarchical Modular Cloning Pipeline. *Curr. Protoc. Mol. Biol.* **130**, e115 (2020).

37. P. Bernard, P. Gabarit, E. M. Bahassi, M. Couturier, Positive-selection vectors using the F plasmid *ccdB* killer gene. *Gene* **148**, 71–74 (1994).

38. F. W. Studier, Protein production by auto-induction in high-density shaking cultures. *Protein Expr. Purif.* **41**, 207–234 (2005).

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

40. K. H. Sumida, R. Núñez-Franco, I. Kalvet, S. J. Pellock, B. I. M. Wicky, L. F. Milles, J. Dauparas, J. Wang, Y. Kipnis, N. Jameson, A. Kang, J. De La Cruz, B. Sankaran, A. K. Bera, G. Jiménez-Osés, D. Baker, Improving Protein Expression, Stability, and Function with ProteinMPNN. *J. Am. Chem. Soc.* **146**, 2054–2061 (2024).

41. A. Currin, N. Swainston, M. S. Dunstan, A. J. Jervis, P. Mulherin, C. J. Robinson, S. Taylor, P. Carbonell, K. A. Hollywood, C. Yan, E. Takano, N. S. Scrutton, R. Breitling, Highly multiplexed, fast and accurate nanopore sequencing for verification of synthetic DNA constructs and sequence libraries. *Synth. Biol.* **4**, ysz025 (2019).

42. B. J. Wittmann, K. E. Johnston, P. J. Almhjell, F. H. Arnold, evSeq: Cost-Effective Amplicon Sequencing of Every Variant in a Protein Library. *ACS Synth. Biol.* **11**, 1313–1324 (2022).

43. M. Houmani, F. Peterkin, G. Antoun, L. Fischer, A. Hammi, parSEQ: Probe and Rescue Sequencing for Advanced Variant Retrieval from DNA Pool. bioRxiv [Preprint] (2023). https://doi.org/10.1101/2023.12.12.571337.

44. Y. Long, A. Mora, E. Gürsoy, K. E. Johnston, F. Zhoufan-Li, F. H. Arnold, LevSeq: Rapid Generation of Sequence-Function Data for Directed Evolution and Machine Learning. bioRxiv [Preprint] (2024). https://doi.org/10.1101/2024.09.04.611255.

45. W. Li, D. Miller, X. Liu, L. Tosi, L. Chkaiban, H. Mei, P.-H. Hung, B. Parekkadan, G. Sherlock, S. F. Levy, Arrayed in vivo barcoding for multiplexed sequence verification of plasmid DNA and demultiplexing of pooled libraries. *Nucleic Acids Res.* **52**, e47 (2024).

46. A. M. Sidore, C. Plesa, J. A. Samson, N. B. Lubock, S. Kosuri, DropSynth 2.0: high-fidelity multiplexed gene synthesis in emulsions. *Nucleic Acids Res.* **48**, e95 (2020).

47. J. C. Klein, M. J. Lajoie, J. J. Schwartz, E.-M. Strauch, J. Nelson, D. Baker, J. Shendure, Multiplex pairwise assembly of array-derived DNA oligonucleotides. *Nucleic Acids Res.* **44**, e43 (2016).

48. Z. Moravec, Y. Zhao, R. Voogd, D. R. Cook, S. Kinrot, B. Capra, H. Yang, B. Raud, J. Ou, J. Xuan, T. Wei, L. Ren, D. Hu, J. Wang, J. B. A. G. Haanen, T. N. Schumacher, X. Chen, E. Porter, W. Scheper, Discovery of tumor-reactive T cell receptors by massively parallel library synthesis and screening. *Nat. Biotechnol.*, 1–9 (2024).

49. S. Lund, V. Potapov, S. R. Johnson, J. Buss, N. A. Tanner, Highly Parallelized Construction of DNA from Low-Cost Oligonucleotide Mixtures Using Data-Optimized Assembly Design and Golden Gate. *ACS Synth. Biol.* **13**, 745–751 (2024).

50. C. Freschlin, K. Yang, P. A. Romero, Scalable and cost-efficient custom gene library assembly from oligopools. bioRxiv [Preprint] (2025). https://doi.org/10.1101/2025.03.22.644747.

51. S. Zou, Z. Wu, C. Xu, “Hierarchical Assembly of Long DNA Libraries from Short Oligonucleotide Pools” (2025; https://openreview.net/forum?id=SrPaXJdYmP).

52. S. L. Lisanza, J. M. Gershon, S. W. K. Tipps, J. N. Sims, L. Arnoldt, S. J. Hendel, M. K. Simma, G. Liu, M. Yase, H. Wu, C. D. Tharp, X. Li, A. Kang, E. Brackenbrough, A. K. Bera, S. Gerben, B. J. Wittmann, A. C. McShan, D. Baker, Multistate and functional protein design using RoseTTAFold sequence space diffusion. *Nat. Biotechnol.*, 1–11 (2024).

53. C. J. Glasscock, R. Pecoraro, R. McHugh, L. A. Doyle, W. Chen, O. Boivin, B. Lonnquist, E. Na, Y. Politanska, H. K. Haddox, D. Cox, C. Norn, B. Coventry, I. Goreshnik, D. Vafeados, G. R. Lee, R. Gordan, B. L. Stoddard, F. DiMaio, D. Baker, Computational design of sequence-specific DNA-binding proteins. bioRxiv [Preprint] (2023). https://doi.org/10.1101/2023.09.20.558720.

54. S. Vázquez Torres, M. Benard Valle, S. P. Mackessy, S. K. Menzies, N. R. Casewell, S. Ahmadi, N. J. Burlet, E. Muratspahić, I. Sappington, M. D. Overath, E. Rivera-de-Torre, J. Ledergerber, A. H. Laustsen, K. Boddum, A. K. Bera, A. Kang, E. Brackenbrough, I. A. Cardoso, E. P. Crittenden, R. J. Edge, J. Decarreau, R. J. Ragotte, A. S. Pillai, M. Abedi, H. L. Han, S. R. Gerben, A. Murray, R. Skotheim, L. Stuart, L. Stewart, T. J. A. Fryer, T. P. Jenkins, D. Baker, De novo designed proteins neutralize lethal snake venom toxins. *Nature* **639**, 225–231 (2025).

55. D. Kim, S. M. Woodbury, W. Ahern, I. Kalvet, N. Hanikel, S. Salike, S. J. Pellock, A. Lauko, D. Hilvert, D. Baker, Computational Design of Metallohydrolases. bioRxiv [Preprint] (2024). https://doi.org/10.1101/2024.11.13.623507.

56. N. R. Bennett, J. L. Watson, R. J. Ragotte, A. J. Borst, D. L. See, C. Weidle, R. Biswas, Y. Yu, E. L. Shrock, R. Ault, P. J. Y. Leung, B. Huang, I. Goreshnik, J. Tam, K. D. Carr, B. Singer, C. Criswell, B. I. M. Wicky, D. Vafeados, M. G. Sanchez, H. M. Kim, S. V. Torres, S. Chan, S. M. Sun, T. Spear, Y. Sun, K. O’Reilly, J. M. Maris, N. G. Sgourakis, R. A. Melnyk, C. C. Liu, D. Baker, Atomically accurate de novo design of antibodies with RFdiffusion.

bioRxiv [Preprint] (2025). https://doi.org/10.1101/2024.03.14.585103.

57. M. Glögl, A. Krishnakumar, R. J. Ragotte, I. Goreshnik, B. Coventry, A. K. Bera, A. Kang, E. Joyce, G. Ahn, B. Huang, W. Yang, W. Chen, M. G. Sanchez, B. Koepnick, D. Baker, Target-conditioned diffusion generates potent TNFR superfamily antagonists and agonists. *Science* **386**, 1154–1161 (2024).

58. W. Yang, D. R. Hicks, A. Ghosh, T. A. Schwartze, B. Conventry, I. Goreshnik, A. Allen, S. F. Halabiya, C. J. Kim, C. S. Hinck, D. S. Lee, A. K. Bera, Z. Li, Y. Wang, T. Schlichthaerle, L. Cao, B. Huang, S. Garrett, S. R. Gerben, S. Rettie, P. Heine, A. Murray, N. Edman, L. Carter, L. Stewart, S. C. Almo, A. P. Hinck, D. Baker, Design of high-affinity binders to immune modulating receptors for cancer immunotherapy. *Nat. Commun.* **16**, 2001 (2025).

59. A. Pillai, A. Idris, A. Philomin, C. Weidle, R. Skotheim, P. J. Y. Leung, A. Broerman, C. Demakis, A. J. Borst, F. Praetorius, D. Baker, De novo design of allosterically switchable protein assemblies. *Nature* **632**, 911–920 (2024).

60. E. Muratspahić, D. Feldman, D. E. Kim, X. Qu, A.-M. Bratovianu, P. Rivera-Sánchez, F. Dimitri, J. Cao, B. P. Cary, M. J. Belousoff, P. Keov, Q. Chen, Y. Ren, J. Fine, I. Sappington, T. Schlichthaerle, J. Z. Zhang, A. Pillai, L. Mihaljević, M. Bauer, S. V. Torres, A. Motmaen, G. R. Lee, L. Tran, X. Wang, I. Goreshnik, D. K. Vafeados, J. E. Svendsen, P. Hosseinzadeh, N. Lindegaard, M. Brandt, Y. Waltenspühl, K. Deibler, L. Oostdyk, W. Cao, L. Anantharaman, L. Stewart, L. Halloran, J. B. Spangler, P. M. Sexton, B. L. Roth, B. E. Krumm, D. Wootten, C. G. Tate, C. Norn, D. Baker, De novo design of miniprotein agonists and antagonists targeting G protein-coupled receptors. bioRxiv [Preprint] (2025). https://doi.org/10.1101/2025.03.23.644666.

61. A. J. Broerman, C. Pollmann, M. A. Lichtenstein, M. D. Jackson, M. H. Tessmer, W. H. Ryu, M. H. Abedi, D. D. Sahtoe, A. Allen, A. Kang, J. D. L. Cruz, E. Brackenbrough, B. Sankaran, A. K. Bera, D. M. Zuckerman, S. Stoll, F. Praetorius, J. Piehler, D. Baker, Design of facilitated dissociation enables control over cytokine signaling duration. bioRxiv [Preprint] (2024). https://doi.org/10.1101/2024.11.15.623900.

62. I. Sappington, M. Toul, D. S. Lee, S. A. Robinson, I. Goreshnik, C. McCurdy, T. C. Chan, N. Buchholz, B. Huang, D. Vafeados, N. Roullier, M. Garcia-Sanchez, M. Glögl, C. Kim, J. L. Watson, S. V. Torres, K. H. G. Verschueren, K. Verstraete, C. S. Hinck, M. Bernard-Valle, B. Coventry, J. N. Sims, G. Ahn, X. Wang, A. P. Hinck, T. P. Jenkins, H. Ruohola-Baker, S. Banik, S. N. Savvides, D. Baker, Improved protein binder design using ⊡-pairing targeted RFdiffusion. bioRxiv [Preprint] (2024). https://doi.org/10.1101/2024.10.11.617496.

63. D. E. Kim, D. R. Jensen, D. Feldman, D. Tischer, A. Saleem, C. M. Chow, X. Li, L. Carter, L. Milles, H. Nguyen, A. Kang, A. K. Bera, F. C. Peterson, B. F. Volkman, S. Ovchinnikov, D. Baker, De novo design of small beta barrel proteins. *Proc. Natl. Acad. Sci.* **120**, e2207974120 (2023).

64. L. An, D. R. Hicks, D. Zorine, J. Dauparas, B. I. M. Wicky, L. F. Milles, A. Courbet, A. K. Bera, H. Nguyen, A. Kang, L. Carter, D. Baker, Hallucination of closed repeat proteins containing central pockets. *Nat. Struct. Mol. Biol.* **30**, 1755–1760 (2023).

65. L. An, M. Said, L. Tran, S. Majumder, I. Goreshnik, G. R. Lee, D. Juergens, J. Dauparas, I.

Anishchenko, B. Coventry, A. K. Bera, A. Kang, P. M. Levine, V. Alvarez, A. Pillai, C. Norn, D. Feldman, D. Zorine, D. R. Hicks, X. Li, M. G. Sanchez, D. K. Vafeados, P. J. Salveson, A. A. Vorobieva, D. Baker, Binding and sensing diverse small molecules using shape-complementary pseudocycles. *Science* **385**, 276–282 (2024).

66. V. Zulkower, S. Rosser, DNA Chisel, a versatile sequence optimizer. *Bioinformatics* **36**, 4508–4509 (2020).

67. V. Potapov, J. L. Ong, R. B. Kucera, B. W. Langhorst, K. Bilotti, J. M. Pryor, E. J. Cantor, B. Canton, T. F. Knight, T. C. Jr. Evans, G. J. S. Lohman, Comprehensive Profiling of Four Base Overhang Ligation Fidelity by T4 DNA Ligase and Application to DNA Assembly. *ACS Synth. Biol.* **7**, 2665–2674 (2018).

68. W. A. Duetz, B. Witholt, Oxygen transfer by orbital shaking of square vessels and deepwell microtiter plates of various dimensions. *Biochem. Eng. J.* **17**, 181–185 (2004).

69. S. van der Walt, S. C. Colbert, G. Varoquaux, The NumPy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.* **13**, 22–30 (2011).

70. SciPy 1.0: fundamental algorithms for scientific computing in Python | Nature Methods. https://www.nature.com/articles/s41592-019-0686-2.

71. W. McKinney, “Data Structures for Statistical Computing in Python” (Austin, Texas, 2010; https://doi.curvenote.com/10.25080/Majora-92bf1922-00a), pp. 56–61.

72. T. pandas development team, pandas-dev/pandas: Pandas, version latest, Zenodo (2020); https://doi.org/10.5281/zenodo.3509134.

73. The HDF Group, Hierarchical Data Format, version 5; https://github.com/HDFGroup/hdf5.

74. Twelve years of SAMtools and BCFtools | GigaScience | Oxford Academic. https://academic.oup.com/gigascience/article/10/2/giab008/6137722.

75. J. Zhang, K. Kobert, T. Flouri, A. Stamatakis, PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2014).

76. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

77. T. J. Lambert, FPbase: a community-editable fluorescent protein database. *Nat. Methods* **16**, 277–278 (2019).

78. M. Remmert, A. Biegert, A. Hauser, J. Söding, HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 (2012).

79. J. K. Leman, B. D. Weitzner, S. M. Lewis, J. Adolf-Bryfogle, N. Alam, R. F. Alford, M. Aprahamian, D. Baker, K. A. Barlow, P. Barth, B. Basanta, B. J. Bender, K. Blacklock, J. Bonet, S. E. Boyken, P. Bradley, C. Bystroff, P. Conway, S. Cooper, B. E. Correia, B. Coventry, R. Das, R. M. De Jong, F. DiMaio, L. Dsilva, R. Dunbrack, A. S. Ford, B. Frenz, D. Y. Fu, C. Geniesse, L. Goldschmidt, R. Gowthaman, J. J. Gray, D. Gront, S. Guffy, S. Horowitz, P.-S. Huang, T. Huber, T. M. Jacobs, J. R. Jeliazkov, D. K. Johnson, K. Kappel, J.

Karanicolas, H. Khakzad, K. R. Khar, S. D. Khare, F. Khatib, A. Khramushin, I. C. King, R. Kleffner, B. Koepnick, T. Kortemme, G. Kuenze, B. Kuhlman, D. Kuroda, J. W. Labonte, J. K. Lai, G. Lapidoth, A. Leaver-Fay, S. Lindert, T. Linsky, N. London, J. H. Lubin, S. Lyskov, J. Maguire, L. Malmström, E. Marcos, O. Marcu, N. A. Marze, J. Meiler, R. Moretti, V. K. Mulligan, S. Nerli, C. Norn, S. Ó'Conchúir, N. Ollikainen, S. Ovchinnikov, M. S. Pacella, X. Pan, H. Park, R. E. Pavlovicz, M. Pethe, B. G. Pierce, K. B. Pilla, B. Raveh, P. D. Renfrew, S. S. R. Burman, A. Rubenstein, M. F. Sauer, A. Scheck, W. Schief, O. Schueler-Furman, Y. Sedan, A. M. Sevy, N. G. Sgourakis, L. Shi, J. B. Siegel, D.-A. Silva, S. Smith, Y. Song, A. Stein, M. Szegedy, F. D. Teets, S. B. Thyme, R. Y.-R. Wang, A. Watkins, L. Zimmerman, R. Bonneau, Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat. Methods* **17**, 665–680 (2020).

80. S. Chaudhury, S. Lyskov, J. J. Gray, PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **26**, 689–691 (2010).

81. Y. Zhang, J. Skolnick, TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).

82. N. Chennamsetty, V. Voynov, V. Kayser, B. Helk, B. L. Trout, Design of therapeutic proteins with enhanced stability. *Proc. Natl. Acad. Sci.* **106**, 11937–11942 (2009).

83. Q. Xu, M. R. Schlabach, G. J. Hannon, S. J. Elledge, Design of 240,000 orthogonal 25mer DNA barcode probes. *Proc. Natl. Acad. Sci.* **106**, 2289–2294 (2009).

84. W. De Coster, R. Rademakers, NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics* **39**, btad311 (2023).

85. E. Steinig, L. Coin, Nanoq: ultra-fast quality control for nanopore reads. *J. Open Source Softw.* **7**, 2991 (2022).

86. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

87. H. Li, New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574 (2021).

88. C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T. E. Oliphant, Array programming with NumPy. *Nature* **585**, 357–362 (2020).