nature research

Corresponding author(s): David Baker

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

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

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Rosetta software suite was used to perform protein design calculations; Rosetta is freely available for academic users on GitHub and command lines and scripts are available in Supplementary Information. The custom script written to generate beta-barrel backbones parametrically are available on https://dx.doi.org/10.5281/zenodo.1216229. The BBQ software used to re-constitute backbones from Calpha coordinates is freely available online. The source code for RIF docking implementation is freely available at https://github.com/rifdock/rifdock. FlowJo v8(FlowJo, LLC) was used to read and analyze the flow cytometry data. Open-source softwares PEAR, ENRICH, ENRICH2 were used to analyze the sequencing data.
Data analysis	Custom python and pyrosetta scripts written to analyze the data were included in the Supplementary Data or available on https://dx.doi.org/10.5281/zenodo.1216229. Pyrosetta module is freely available for academic users on GitHub. The commercial FlowJo software (v8) was used to analyze Flow Cytometry results. Crystallographic data were analyzed with Phenix (release 1.101.1-2155) and Coot (v0.8.7 EL).

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

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The atomic coordinates and experimental data of BB1, b10, b11L5F_LGL, mFAP0-DFHBI, and mFAP1-DFHBI crystal structures have been deposited in the RCSB Protein Database with the accession numbers of 6D0T, 6CZJ, 6CZG, 6CZH, and 6CZI respectively. All the design models, Illumina sequencing data, sequencing analysis and source data (Fig.2 &.4, Extended Data Fig. 6e, 7, 8a&h) are available on https://dx.doi.org/10.5281/zenodo.1216229.

Field-specific reporting

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

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

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

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

Sample size	Sample size was determined by the estimated work load. In total, 144 designed proteins were experimentally tested. The number of parametric designs experimentally tested (41) was considered sufficient since no significant improvement of the success rate was observed after 4 iteration of design/characterization. The number of characterized designs generated based on the 2D map was considered as sufficient since each round of design calculations yielded folded proteins, and at least some of them had the expected activity or improvement of activity. All computational experiments were carried out on sets of 100-250 models. The size of these sets was chosen to be in the same order of magnitude as the sets used as starting point for design (200-500), while small enough to generate data on our local clusters. In only one case, the size of the set was limited by the possibility to generate enough distinct backbones with that particular parametric model.
Data exclusions	All the data points were reported, for both experimental and computational studies.
Replication	Basic biochemical and structural characterization were performed once or twice with internal statistical validation. The replicated experiments (CD spectra, fluorescence spectra, fluorescence imaging, SEC-MALS, computational experiments) produced similar results. Variability was observed between some purified protein samples in terms of the relative ratio of monomeric/aggregated proteins. Yeast library generation and selection for Miseq sequencing were performed twice.
Randomization	No randomized samples allocation in this work. Because of the complexity of the designs (the design of beta-barrels has been unsuccessfully attempted several times by different groups) and the success rate (20-25% folded proteins) and diversity of successful designs, it is unlikely that the folding amino acids sequences reported in this paper are due to chance rather than design calculation. Negative and positive controls were well defined and validated.
Blinding	Blinding was not relevant to this work since the experiments were well-defined by the computational design.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

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

Antibodies

Antibodies used	FIT-congugated anti-cMyc antibody (chicken) purchased from Immunology Consultants Labs (catalog number: CMYC-45F).
Validation	efficient cell labeling was confirmed by flow cytometry.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	NIH3T3 cells (Flp-In-3T3, Thermo Fisher Scientific); COS-7 cells (ATCC CRL-1651)
Authentication	Not authenticated
Mycoplasma contamination	Confirmed negative for mycoplasma
Commonly misidentified lines (See ICLAC register)	No common misidentified lines were used in this study.

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	https://dx.doi.org/10.5281/zenodo.1216229 (This doi is reserved and will be activate and become public once the publication is finalized.)
Files in database submission	b11L5F_DMS_analysis.zip, b11L5F_DMS_EnrichProcessed.tar.gz, b11L5F_DMS_PearAssembled.tar.gz, b11L5F_DMS_IlluminaRead.tar.gz (detailed descriptions in "List.docx" file.)
Genome browser session (e.g. UCSC)	no longer applicable.

Methodology

Replicates	2 biological replicates
Sequencing depth	300 cycles paired-end reads
Antibodies	FIT-congugated anti-cMyc antibody (chicken) purchased from Immunology Consultants Labs (catalog number: CMYC-45F); fluoreogenic fluorophore DFHBI from Lucerna.
Peak calling parameters	Illumina's on-instrument data analysis MiSeq Reporter was use to process the fluorescence data.
Data quality	a statistic confidential interval is assigned to each mutation.
Software	open-source softwares PEAR, ENRICH, ENRICH2 were used to analyze the data.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	see Supplementary Methods and Supplementary Table 15
Instrument	Sony SH800
Software	FlowJo v8

Cell population abundance	see Supplementary Table 15
Gating strategy	FSC/SSC was used to gate the size of yeast cells; FSC width signal/ FSC height signal was used to gate the singlet cells.

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