# Methods

## Media and buffers used

* LB: 10 g/L Bacto-tryptone, 5 g/L Yeast extract, 10 g/L NaCl. Autoclaved 20 min at 120ºC.
* YPDA: 20 g/L glucose, 20 g/L Peptone, 10 g/L Yeast extract, 40 mg/L adenine sulphate. Autoclaved 20 min at 120ºC.
* SORB: 1 M sorbitol, 100 mM LiOAc, 10 mM Tris pH 8.0, 1 mM EDTA. Filter sterilized (0.2 mm Nylon membrane, ThermoScientific).
* Plate mixture: 40% PEG3350, 100 mM LiOAc, 10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0. Filter sterilised.
* Recovery medium: YPD (20 g/L glucose, 20 g/L Peptone, 10 g/L Yeast extract) + 0.5 M sorbitol. Filter sterilised.
* SC -URA: 6.7 g/L Yeast Nitrogen base without amino acid, 20 g/L glucose, 0.77 g/L complete supplement mixture drop-out without uracil. Filter sterilised.
* SC -URA/MET/ADE: 6.7 g/L Yeast Nitrogen base without amino acid, 20 g/L glucose, 0.74 g/L complete supplement mixture drop-out without uracil, adenine and methionine. Filter sterilised.
* Competition medium: SC –URA/MET/ADE + 200 ug/mL methotrexate (BioShop Canada Inc., Canada), 2% DMSO.
* DNA extraction buffer: 2% Triton-X, 1% SDS, 100mM NaCl, 10mM Tris-HCl pH8, 1mM EDTA pH8.

## ddPCA plasmids construction

Three generic plasmids were constructed to be able to assay any protein of interest by BindingPCA or AbundancePCA: the BindingPCA plasmid (pGJJ001), the AbundancePCA plasmid (pGJJ045) and the mutagenesis plasmid (pGJJ055).

The first two plasmids were derived from plasmid pGD110 which carries two halves of the murine methotrexate-resistant DHFR (DHFR1,2 and DHFR3) with C-terminus (GGGGS)4 linker fusions under the expression of CYC promoters and a shared CYC terminator (URA3 cassette plasmid for yeast auxotrophic selection during the selection assays) as described[24](https://paperpile.com/c/coGznM/Ld4Fh) plus a multiple cloning site. pGJJ001 had the same structure as pGD110 but with a barcode cloning site upstream of the CYC promoter driving expression of DHFR3 in case a barcode-variant association sequencing strategy was necessary. To construct the plasmid, pGD110 was amplified in 3 different fragments using primer pairs oGJJ001-oGJJ002, oGJJ003-oGJJ083 and oGJJ82-oGJJ016 (Supplementary Table 1), which were then assembled by Gibson reaction (prepared in house) at 50ºC for one hour. The AbundancePCA plasmid pGJJ045 was constructed by Gibson assembly by substituting the CYC promoter driving expression of the half DHFR1,2 for the GPD promoter using primer pairs oGJJ47-oGD087 and oGJJ46-oGD089 (Supplementary Table 1).

The generic mutagenesis plasmid (pGJJ055) was created to harbour a landing site with HindIII and AvrII restriction sites so that the CYC promoter and DHFR3 fused to any protein of interest could be cloned into it to perform error-prone PCR, nicking mutagenesis[50](https://paperpile.com/c/coGznM/RIs0w) or other alternative mutagenesis strategies. It was derived from pUC19, to avoid future plasmid selection in yeast if not properly purified (not containing any yeast auxotrophic cassette). The final plasmid pGJJ055 was built in two cloning steps. Initially pUC19 was reduced in size (the lacZα fragment was deleted) to increase the efficiency of *E. coli* transformation and three synonymous mutations were introduced in the ampicillin resistance cassette (*bla* gene) to remove specific restriction sites. This resulting intermediate plasmid (pGJJ003) was obtained using three-fragment Gibson assembly. Two fragments were amplified (primer pairs oGJJ008-9 and oGJJ010-11) from pUC19 that introduced the barcode landing site. The third fragment with the *bla* sequence with the synonymous mutations was synthesised as a dsDNA gene block (gbGJJ001, GeneScript, Supplementary Table 2). To generate the final mutagenesis plasmid (pGJJ055) a T>C synonymous substitution was added in the *bla* gene of pGJJ003 to create an Nb.BbvCI restriction site by one-fragment Gibson assembly using primers oGJJ049-50.

## GRB2-SH3 domain plasmid construction

To construct the GRB2-SH3 AbundancePCA plasmid (pGJJ046) the 56 amino acid long SH3 domain of GRB2 (UniProt P62993, from amino acid 159 to 224 of the human protein GRB2) was fused to the C-terminus of the DHFR3 fragment of the AbundancePCA plasmid (pGJJ045). To do so the SH3 domain was amplified by PCR reaction using primer pair oGJJ012-13 to introduce the flanking HindIII and NheI restriction sites and then cloned into the digested pGJJ045 plasmid using T4 Ligase (NEB).

To construct the GRB2-SH3 BindingPCA plasmid (pGJJ034), first the sequence of GAB2 containing the linear peptide (32 amino acids long, amino acid 498 to 530 of the human GAB2 protein[51](https://paperpile.com/c/coGznM/YtdyM)) was fused to the fragment DHFR1,2. GAB2 was amplified using primer pair oGJJ014-15, which introduced flanking BamHI and SpeI restriction sites. Both the PCR product and the BindingPCA plasmid (pGJJ001) were digested and purified. The assembly of the new BindingPCA plasmid with GAB2 (pGJJ006) was obtained by ligation using T4 Ligase. After validation by Sanger sequencing, pGJJ006 was digested with HindIII and NheI restriction enzymes and cloned with the GRB2-SH3 domain to obtain the final wild-type GRB2-SH3 BindingPCA plasmid pGJJ034 (with both GRB2-SH3 and GAB2 linear peptide fused to both fragments of DHFR).

## PSD95-PDZ domain plasmid construction

To construct the PSD95-PDZ3 AbundancePCA plasmid (pGJJ068) the sequence of the third PDZ domain of PSD95 (UniProt P78352, from amino acid 354 to 437 of the human DLG4 gene) was fused to the C-terminus of the DHFR3 fragment of the AbundancePCA plasmid (pGJJ045). A gene block containing the PSD95-PDZ3 domain sequence was ordered (gbGJJ010, IDT, Supplementary Table 2), amplified by PCR (Q5-Hot Start Polymerase, NEB) using oligos oGJJ078-79 to add a HindIII and NheI restriction sites, digested with HindIII and NheI and subsequently cloned into the digested pGJJ045 by T4 Ligase (NEB) standard protocol.

To obtain the BindingPCA plasmid of PSD95-PDZ3 (pGJJ072), the entire human gene CRIPT[52](https://paperpile.com/c/coGznM/Rbr2P) was fused to the DHFR1,2 half of the BindingPCA plasmid pGJJ001. To do so, a gene block with the sequence of the human CRIPT gene was ordered with flanking BamHI and SpeI restriction sites (gbGJJ011, IDT). The gene block was digested and cloned into pGJJ001, creating the intermediate plasmid pGJJ066. pGJJ066 was later digested with HindIII and NheI to clone in the C-terminus of the DHFR3 the digested PCR product of the PSD95-PDZ3 domain, obtaining the final wild-type PSD95-PDZ3 BindingPCA plasmid pGJJ072.

## Growth rate measurements of individual constructs

Seven single amino acid mutations of the GRB2-SH3 domain were tested for yeast growth in the presence of MTX on both abundance and binding assays. To construct these plasmids, 7 gene blocks containing the mutated versions of GRB2-SH3 were synthesised with HindIII and NheI flanking restriction sites (gbGJJ002-8, Twist Bioscience, Supplementary Table 2). The gene blocks were digested, purified and assembled into the previously digested and purified pGJJ006 and pGJJ045 (BindingPCA plasmid with GAB2 and AbundancePCAplasmid, respectively).

The GRB2-SH3 wild-type and mutant constructs were individually transformed into yeast following a small-scale high efficiency yeast transformation (See section 7.1, same protocol but scaled down in volume 0.0003X). After transformation, different colonies for each construct were picked and grown independently in 400 uL of SC -URA/MET/ADE overnight at 30ºC using a 96 deep well plate. The following morning the optical density (OD) of each well was measured using a Tecan Infinite M Plex plate reader (Tecan, Switzerland). Cultures in each well were diluted to an OD600nm of 0.1. For the selection experiment, 5 uL of the diluted cells were added into 95 ul of 1.053X competition media (SC -URA/MET/ADE + 200 ug/mL methotrexate) to obtain 100 uL of starting culture at OD600nm = 0.005 of uL of per well. The culture was grown for 60h at 30ºC in a Tecan plate reader where OD600nm measurements were taken every 15 minutes. The growth rate in each well was obtained by calculating the slope of the exponential phase of the growth curve (slope of a linear fit of the log10(OD600nm) against time).

## Mutagenesis library construction

1. **Types of mutant libraries:**

In this study two types of mutagenesis library construction approaches were used: error-prone PCR or one-pot nicking mutagenesis[50](https://paperpile.com/c/coGznM/RIs0w). The selection of the technique depended on the domain length, the number of targeted mutations and the mutational biases of the strategies (mutation type vs. positional bias for error-prone PCR and nicking mutagenesis respectively). Supplementary Table 3 contains fully detailed information of the different libraries constructed in this study.

For the GRB2-SH3 domain two types of libraries were constructed (Supplementary Table 3). An initial error-prone PCR library, covering some singles and few double mutants, which was used in the BindingPCA assay (“SH3\_EP”), and a second library of two rounds of nicking mutagenesis, extending the amount of single and double mutants to run the AbundancePCA assay (“SH3\_NM2”). Given the short length of the GRB2-SH3 domain, these two high complexity libraries could be sequenced in a HiSeq 125bp paired-end sequencing platform.

Due to the long length of the PSD95-PDZ3 domain, the platform for sequencing used was the MiSeq 250bp paired-end, giving a substantially lower number of reads counts per run compared to the HiSeq. Thus the PSD95-PDZ3 mutant libraries were constructed to have lower complexity compared to the SH3 ones. This consisted of a shallow double mutant library of one round of nicking mutagenesis, starting from a template pool of 10 different PSD95-PDZ3 wild-type and single mutant backgrounds (“PDZ\_NM2sha”).

1. **Cloning the SH3 and PDZ domains into the mutagenesis plasmid**

The CYC promoter and DHFR3 C-terminally fused to the GRB2-SH3 domain was cloned into the mutagenesis plasmid (pGJJ055) by digestion-ligation protocol. The CYC-DHFR3-GRB2 insert was obtained by digesting the plasmid pGJJ025 (BindingPCA plasmid with GRB2-SH3 tagged to DHFR3) with HindIII-HF and AvrII and purifying the correct size band using the QIAquick Gel Extraction Kit (QIAGEN). The mutagenesis plasmid pGJJ055 was digested with HindIII-HF and AvrII and purified with the MinElute PCR Purification Kit (QIAGEN). The GRB2 mutagenesis plasmid (pGJJ057) was assembled by a ligation reaction (T4 Ligase, New England Biolabs) following the manufacturer’s protocol. After transformation into NEB10-beta High Efficiency competent cells, the plasmid sequence was verified by Sanger Sequencing (GATC, Eurofins Genomics).

The mutagenesis plasmid containing the PSD95-PDZ domain (pGJJ111) was obtained following the same strategy as for the GRB2-SH3 mutagenesis plasmid. The insert containing the CYC promoter and DHFR3 fused to the PDZ domain was digested and purified from plasmid pGJJ072 and subsequently cloned as described above into the HindIII-AvrII digested pGJJ055.

1. **SH3 error-prone PCR library**

The error-prone PCR reaction for libraries “SH3\_EP” was done using the GeneMorph II Mutagenesis Kit (Agilent Technologies) following the manufacturer’s protocol. Primers oGJJ048-152 (with homology to the origin of replication and the CYC promoter respectively) were used to amplify 1,011 bp of the SH3 (pGJJ057) mutagenesis plasmids. A single PCR reaction of 50 uL was run using 0.91 ng of template plasmid pGJJ057 (reaching the lowest recommended amount of plasmid by the manufacturer, lower plasmid amount increases the mutation frequency), with an annealing temperature of 56.4ºC (previously determined by gradient PCR), 1 minute and 10 seconds of extension time. The PCR product was later run on an agarose gel for band confirmation and purified using the MinElute PCR Purification Kit (QIAGEN).

The entire error-prone PCR product was digested with HindIII-HF and NheI-HF restriction enzymes. The correct band with the GRB2-SH3 domain was purified using the MinElute Gel Extraction Kit (QIAGEN) and the sample was quantified using a Qubit fluorometer to be cloned into the yeast assay plasmids by high efficiency temperature-cycle ligation[53](https://paperpile.com/c/coGznM/llZLS).

1. **GRB2-SH3 and PSD95-PDZ3 nicking mutagenesis libraries**

The plasmid-based one-pot saturation (nicking) mutagenesis protocol was extracted from [50](https://paperpile.com/c/coGznM/RIs0w). The protocol was followed exactly as described, with two minor changes: the template plasmid DNA was prepared from 4°C midi-preps and the mutagenic primers were used at 1:10 molar ratio to plasmids instead of 1:20. The mutagenic primers were designed to contain the three degenerate bases (NNK) in the codon position with an average of 21bp of homology sequence upstream and downstream with the length adjusted to match similar melting temperatures (Sigma-Aldrich, ‘seq\_opt’ column in Supplementary Table 4).

The mutant library “SH3\_NM2” (Supplementary Table 3) consisted of two rounds of nicking mutagenesis, using an equimolar mix of degenerate GRB2-SH3 primers (Supplementary Table 4). In the first round of nicking mutagenesis, 3.3x105 *E. coli* tranfromants were obtained. The resulting midiprep of the overnight culture was used as template for the second round of nicking mutagenesis with the same primer mix, but scaled up in volume 10 times. The second round of nicking mutagenesis resulted in 3.75x107 *E. coli* transformants.

For the “PDZ\_NM2sha”, an initial single round of nicking mutagenesis using an equimolar mix of degenerated PSD95-PDZ3 primers (Supplementary Table 4) was obtained for two reasons: (1) To obtain 9 random single mutants to use as template for another round of nicking mutagenesis (by randomly selecting single colonies and verified by Sanger sequencing) and (2) to quantify the degenerate primer positional bias and compensate for it in the shallow double mutant library by spiking part of this library was MiSeq 250bp (data not shown). To construct the final “PDZ\_NM2sha” library an equimolar pool of nine single mutants and the wild-type was used as the plasmid template for a round of nicking mutagenesis. To compensate for the extreme positional biases, each mutagenic primer was mixed in the pool inversely to the mean read counts per position from the sequencing results of MiSeq spike-in initial library.

The libraries midi-preps were digested with HindIII and NheI restriction enzymes and the insert containing the mutated protein was gel purified (MinElute Gel Extraction Kit, QIAGEN) to be later cloned into the two assay plasmids by temperature-cycle ligation.

1. **Cloning the mutant libraries into the yeast assay plasmids**

Both, the AbundancePCA plasmid (pGJJ045) and the BindingPCA plasmid containing the ligand GAB2 or CRIPT fused to DHFR1,2 (pGJJ006 and pGJJ066 respectively) were digested with the same enzymes and purified using the QIAquick Gel Extraction Kit (QIAGEN).

The assembly of GRB2 SH3 and PSD95 PDZ mutant libraries in both assay plasmids were done overnight by temperature-cycle ligation using T4 ligase (New England Biolabs) according to the manufacturer's protocol, 67 fmol of backbone and 200 fmol of insert in a 33.3 uL reaction. The ligation was desalted by dialysis using membrane filters for 1h and later concentrated 3.3X using a SpeedVac concentrator (Thermo Scientific).

All concentrated assembled libraries were transformed into NEB 10β High-efficiency Electrocompetent *E. coli* cells according to the manufacturer’s protocol (volumes used in each library specified in Supplementary Table 3). Cells were allowed to recover in SOC medium (NEB 10β Stable Outgrowth Medium) for 30 minutes and later transferred to 200 mL of LB medium with ampicillin 4X overnight. The total number of estimated transformants for each library can be found in Supplementary Table 3. 100 mL of each saturated *E. coli* culture were harvested next morning to extract the plasmid library using the QIAfilter Plasmid Midi Kit (QIAGEN).

## Methotrexate selection assays

1. **Large-scale yeast transformations**

The high-efficiency yeast transformation protocol of all the different selection assays was derived from[24](https://paperpile.com/c/coGznM/Ld4Fh). The protocol was scaled in volume depending on the targeted number of transformants of each library. The transformation protocol described below (adjusted to a pre-culture of 700 mL of YPDA) was scaled up or down in volume as reported in Supplementary Table 3.

For each of the two assays (*​*BindingPCA and *​*AbundancePCA) of each protein domain library three independent pre-cultures of BY4742 were grown in 80 mL standard YPDA at 30ºC overnight. The next morning, the cultures were diluted into 700 mL of pre-wormed YPDA at an OD​600nm = 0.3. The cultures were incubated at 30ºC for 4 hours. After growth, the cells were harvested and centrifuged for 5 minutes at 3,000g, washed with sterile water and later with SORB medium (100mM LiOAc, 10mM Tris pH 8.0, 1mM EDTA, 1M sorbitol). The cells were resuspended in 34.4 mL of SORB and incubated at room temperature for 30 minutes. After incubation, 700 μL of 10mg/mL boiled salmon sperm DNA (Agilent Genomics) was added to each tube of cells, as well as 14 μg of plasmid library. After gentle mixing, cells were split in tubes of ~8.8 mL of cells and 35 mL of Plate Mixture (100mM LiOAc, 10mM Tris-HCl pH 8, 1mM EDTA/NaOH, pH 8, 40% PEG3350) were added to each tube to be incubated at room temperature for 30 more minutes. 3.5 mL of DMSO was added to each tube and the cells were then heat shocked at 42ºC for 20 minutes (inverting tubes from time to time to ensure homogenous heat transfer). After heat shock, cells were centrifuged and re-suspended in ~50 mL of recovery media and allowed to recover for 1 hour at 30ºC. Next, cells were again centrifuged, washed with SC-URA medium and re-suspended in SC -URA (volume used in each library found in Supplementary Table 3). After homogenization by stirring, 10 uL were plated on SC -URA Petri dishes and incubated for ~48 hours at 30ºC to measure the transformation efficiency. The independent liquid cultures were grown at 30ºC for ~48 hours until saturation. The number of yeast transformants obtained in each library assay can be found in Supplementary Table 3.

1. **Selection assays**

For each of the BindingPCA or AbundancePCA assays, each of the growth competitions was performed right after yeast transformation. After the first cycle of post-transformation plasmid selection, a second plasmid selection cycle (input) was performed by inoculating SC -URA/MET/ADE at a starting OD600nm = 0.1 with the saturated culture (volume of each experiment specified in Supplementary Table 3). Cells were grown for 4 generations at 30ºC under constant agitation at 200 rpm. This allowed the pool of mutants to be amplified and enter the exponential growth phase. The competition cycle (output) was then started by inoculating cells from the input cycle into the competition media (SC -URA/MET/ADE + 200 ug/mL Methotrexate) so that the starting OD600nm was 0.05. For that, the adequate volume of cells were collected, centrifuged at 3,000 rpm for 5 minutes and resuspended in the pre-warmed output media. Meanwhile, each input replicate culture was splitted in two and harvested by centrifugation for 5 min at 5,000g at 4ºC. Yeast cells were washed with water, pelleted and stored at -20ºC for later DNA extraction. After ~5-5.2 generations of competition cycle, each output replicate culture was splitted into two and harvested by centrifugation for 5 min at 5,000g at 4ºC, washed twice with water and pelleted to be stored at -20ºC.

## DNA extractions and plasmid quantification

The DNA extraction protocol used is described below for a 400 mL harvested culture of OD600nm ~ 1.6. Depending on the volume harvested in each selection assay (Supplementary Table 3), the following protocol was scaled up or down in volume.

Cell pellets (one for each experiment input/output replicate) were re-suspended in 4 mL of DNA extraction buffer, frozen by dry ice-ethanol bath and incubated at 62ºC water bath twice. Subsequently, 4 mL of Phenol/Chloro/Isoamyl 25:24:1 (equilibrated in 10mM Tris-HCl, 1mM EDTA, pH8) was added, together with 4 g of acid-washed glass beads (Sigma Aldrich) and the samples were vortexed for 10 minutes. Samples were centrifuged at RT for 30 minutes at 4,000 rpm and the aqueous phase was transferred into new tubes. The same step was repeated twice. 0.4 mL of NaOAc 3M and 8.8 mL of pre-chilled absolute ethanol were added to the aqueous phase. The samples were gently mixed and incubated at -20ºC for 30 minutes. After that, they were centrifuged for 30 min at full speed at 4ºC to precipitate the DNA. The ethanol was removed and the DNA pellet was allowed to dry overnight at RT. DNA pellets were resuspended in 2.4 mL TE 1X and treated with 20 uL of RNaseA (10mg/mL, Thermo Scientific) for 30 minutes at 37ºC. To desalt and concentrate the DNA solutions, QIAEX II Gel Extraction Kit was used (200 µL of QIAEX II beads). The samples were washed twice with PE buffer and eluted 500 µL of 10 mM Tris-HCI buffer, pH 8.5. Finally, plasmid concentrations in the total DNA extract (that also contained yeast genomic DNA) were quantified by qPCR using the primer pair oGJJ152-oGJJ153, that bind to the ori region of the plasmids.

## Sequencing library preparation

The sequencing libraries were constructed in two consecutive PCR reactions. The first PCR (PCR1) was designed to amplify the mutated protein of interest and to increase the nucleotide complexity of the first sequenced bases by introducing frame-shift bases between the adapters and the sequencing region of interest. The second PCR (PCR2) was necessary to add the remainder of the Illumina adapter and demultiplexing indexes.

To avoid PCR biases, PCR1 of each independent sample (input/output replicates of any of the yeast assays) was run with an excess of plasmid template 20-50 times higher than the number of expected sequencing reads per sample (Supplementary Table 3). Each reaction started with a maximum of 1.25x107 template plasmid molecules per uL of PCR1, avoiding introducing more yeast genomic DNA that interfered with the efficiency of the PCR reaction. For this reason, PCR1s were scaled up in volume as specified in Supplementary Table 3. The PCR1 reactions were run using Q5 Hot Start High-Fidelity DNA Polymerase (New England Biolabs) according to the manufacturer’s protocol, with 25 pmol of pooled frame-shift primers oGJJ52/54-58 and oGJJ84-89 (forward and reverse primers were independently pooled according to the nucleotide diversity of each oligo, Supplementary Table 1). The PCR reactions were set to 60ºC annealing temperature, 10 seconds (for GRB2-SH3 libraries) or 15 seconds (for PSD95-PDZ3 libraries) of extension time and run for 15 cycles. Excess primers were removed by adding 0.04 uL of ExoSAP-IT (Affymetrix) per uL of PCR1 reaction and incubated for 20 min at 37 ̊C followed by an inactivation for 15 min at 80 ̊C. The PCRs of each sample were then pooled and purified using the MinElute PCR Purification Kit (QIAGEN) according to the manufacturer’s protocol. DNA was eluted in EB to a volume 6 times lower than the total volume of PCR1.

PCR2 reactions were run for each sample independently using Hot Start High-Fidelity DNA Polymerase. The total reaction of PCR2 was reduced to half of PCR1, using 0.05 uL of the previous purified PCR1 per uL of PCR2. In this second PCR the remaining parts of the Illumina adapters were added to the library amplicon. The forward primer (5’ P5 Illumina adapter) was the same for all samples, while the reverse primer (3’ P7 Illumina adapter) differed by the barcode index (oligo sequences in Supplementary Table 1), to be subsequently pooled together and demultiplex them after deep sequencing (indexes used in each replicate of each sequencing run found in Supplementary Table 5). 10 cycles of PCR2s were run at 62ºC of annealing temperature and 15 seconds (for GRB2-SH3 libraries) or 20 seconds (for PSD95-PDZ3 libraries) of extension time. All reactions from the same sample were pooled together and an aliquot was run on a 2% agarose gel to be quantified. After quantification, samples with different Illumina indexes that were sequenced together in the same flow-cell were pooled in an equimolar ratio, run on a gel and purified using the QIAEX II Gel Extraction Kit. The purified amplicon library pools were subjected to Illumina 125bp paired-end HiSeq (SH3 libraries) or 250bp paired-end MiSeq (PDZ libraries) sequencing at the CRG Genomics Core Facility.

## Sequencing data processing

FastQ files from paired-end sequencing of all BindingPCA and AbundancePCA experiments were processed with DiMSum v1.2.8[54](https://paperpile.com/c/coGznM/TrrAO) using default settings with minor adjustments: <https://github.com/lehner-lab/DiMSum>. Supplementary Table 6 contains DiMSum fitness estimates and associated errors for all experiments. Experimental design files and command-line options required for running DiMSum on these datasets are available on GitHub (<https://github.com/lehner-lab/doubledeepms>). In all cases, adaptive minimum Input read count thresholds based on the corresponding number of nucleotide substitutions (“fitnessMinInputCountAny” option) were selected in order to minimise the fraction of reads per variant related to sequencing error-induced “variant flow” from lower order mutants[54](https://paperpile.com/c/coGznM/TrrAO).

For the PSD95-PDZ3 data, where nicking mutagenesis was performed on a starting template pool of 10 different backgrounds, variant counts associated with all samples (output from DiMSum stage 4) were further filtered using a custom script to retain double aa variants consisting of single aa mutations in these designed backgrounds. All read counts associated with remaining double aa variants (likely the result of PCR and sequencing errors) were discarded. Finally, fitness estimates and associated errors were then obtained from the resulting filtered variant counts with DiMSum (“countPath” option).

The pre-processed DMS data for GB1 were obtained from the supplementary information of a previous study[19](https://paperpile.com/c/coGznM/WiIdn) ([github.com/jotwin/ProteinGthermo](http://github.com/jotwin/ProteinGthermo)). The data consist of sequencing read counts from a single (common) Input sample and three replicate (selection) experiments assaying the binding affinity of wild-type, single and double aa GB1 variants to IgG (Output samples)[35](https://paperpile.com/c/coGznM/cDzag). Fitness estimates and associated errors were obtained from variant counts with DiMSum v1.2.8 using default settings.

## Protein fitness estimation

For DMS experiments in general, the enrichment score (ES) for each variant *i* in replicate *r* is typically defined as the ratio between its frequency before and after selection:

Where *Ninput* and *Noutput* are the corresponding numbers of sequenced reads in the Input and Output sample respectively. DiMSum calculates fitness scores for each variant *i* in replicate *r* as the natural logarithm of the enrichment score normalised to the wild-type or reference variant (*wt*):

Measurement errors of fitness scores are estimated by fitting a modified Poissonian (count-based) model to the data—with replicate-specific multiplicative and additive terms that are common to all variants—that has been shown to outperform previous methods in benchmarking analyses[54](https://paperpile.com/c/coGznM/TrrAO). Final DiMSum fitness and error estimates for all variants are obtained by merging between replicates using weighted averaging.

For cell growth based assays such as BindingPCA and AbundancePCA, the growth rate associated with variant *i* in replicate *r* can be estimated as:

where *Dinput* and *Doutput* are proportional to the total number of cells in the Input and Output sample respectively (i.e. optical density or similar), ∆*T* is the time difference (in hours) between measurements (i.e. selection/competition time) and the summation is over all variants in replicate *r*. Growth rates and fitness scores can therefore be seen to be related by a simple linear transformation:

because all terms on the right-hand side are constant for all variants *i* in replicate *r*, except the fitness score *Fr,i*.

## Thermodynamic models

At thermal equilibrium, the Boltzmann distribution relates the probability that a system will be in a given state *k* to the (Gibbs) free energy *Gk* of the state and the temperature of the system *T*:

Where *R* is the gas constant. Applied to proteins and considering the universe of *M* distinct conformations and/or interactions, the fraction of molecules in state *k* compared to all possible states is therefore given by:

where the summation is over all possible states. In the case of a two-state unfolded/folded model, where we denote the sum of energies of all possible unfolded states with the reference value of zero (i.e. *Gu* = 0), the fraction of molecules in the folded state is then:

Where and ∆*Gf* is the energy difference between folded and unfolded states. As in previous work[19](https://paperpile.com/c/coGznM/WiIdn), in order to capture changes in both fold stability and the stability of binding, we model PPIs as an equilibrium between three states: unfolded and unbound (*uu*), folded and unbound (*fu*), and folded and bound (*fb*). The probability of the unfolded and bound state (*ub*) is assumed to be negligible. The probability of the folded and bound state is then given by:

where and ∆*Gb* is the energy difference between bound and unbound folded states (i.e. binding affinity). Importantly, both ∆*Gb* and ∆*Gf* implicitly depend upon the amino acid sequence. Furthermore, we assume that these energies are additive, meaning that the total free energy change (∆∆*G*) of an arbitrary variant *i* (of any mutational order) relative to the wild-type sequence is simply the sum over residue-specific energies (∆∆*g*) corresponding to all constituent individual (i.e. lowest order) amino acid changes:

where ∆∆*gb,j* and ∆∆*gf,j* denote binding and folding free energy changes of constituent single amino acid substitutions of variant *i* relative to the wild-type. We can therefore express the absolute (rather than relative) free energies of binding and folding of an arbitrary variant *i* as:

where ∆*Gb,wt* and ∆*Gf,wt* are the binding and folding free energies of the wild-type.

Considering that the concentration of the interaction partners represents a constant added to ∆*Gb*, we expect this not to affect the estimation of changes in the free energy of binding (∆∆*Gb*). Additionally, the effect (if any) of the fused DHFR fragments on the stability of the assayed protein domains ∆*Gf* is most likely additive in free energy space for all variants and similarly this would not affect the estimation of changes in the free energy of folding (∆∆*Gf*). Finally, in the absence of calibrations to a reference set of directly measured free energies, our inferences should be strictly considered ‘pseudo’ free energy changes, but we refer to them without this qualifier in the interest of brevity.

## Thermodynamic model inference with MoCHI

The modelling approach taken here assumes that the effects of mutations on protein concentration are mediated via their effects on protein folding, not via other mechanisms such as active degradation or aggregation. Given that protein fitness (as defined above) is proportional to the concentration of either the free protein or protein complex in the AbundancePCA or BindingPCA assays respectively, we can parameterize binding fitness *Ffb* and abundance fitness *Ff* as follows:

Where *xb* = ∆*Gb*/*RT* and *xf* = ∆*Gf*/*RT* are additive functions of the amino acid sequence, and *cfb*, *cf*, *dfb*, *df* are scalar constants.

We trained a feed-forward neural network representation of this genotype-phenotype model using the backpropagation-based algorithms provided by TensorFlow 2.3[55](https://paperpile.com/c/coGznM/bzWnQ). We thereby inferred the binding and folding free energy changes that optimally predict AbundancePCA and BindingPCA fitness measurements from the corresponding aa sequences. We refer to this method that infers the underlying biophysical traits from DMS data as MoCHI (in preparation).

The neural network architecture consists of one input layer (aa sequences) and one trainable hidden layer (additive traits i.e. free energies) for each biophysical trait (binding and folding), as well as one trainable hidden layer (linear transformation) per phenotype of interest (AbundancePCA and BindingPCA fitness). An additional input layer specifies the corresponding phenotype associated with each sequence. Finally, the network includes two additional untrainable hidden layers to transform energies to bound and folded fractions respectively, as well as (helper) layers to multiplex/concatenate fitness predictions for different phenotypes into a single 1-dimensional output.

This architecture, i.e. a single 1-dimensional target (output) variable representing both AbundancePCA and BindingPCA fitness, allows arbitrary combinations of variants for each phenotype to be used for training and also obviates the need to explicitly pair phenotype measurements of each variant.

The target (output) data to fit the neural network comprises fitness estimates for wildtype, single and double aa substitution variants. A random 10% of double aa substitution variants represent the validation data (held out during model fitting). Furthermore, in order to capture uncertainty in ddPCA fitness estimates, the training data was expanded 10x by randomly sampling ten values from the fitness error distribution of each variant. The validation data was left unchanged. In order to prioritise fitting of wild-type free energy terms (which influence final free energy estimates of all other variants) the wild-type variant was artificially duplicated to represent 1% of training data samples. This approach is conceptually similar to the practice of applying weights to important samples, which are used for weighting the loss function during training.

Input (feature) data comprises 1-hot encoded representations of the observed input sequences (wild-type, single and double aa substitutions) i.e. one binary feature variable for each unique aa substitution and each biophysical trait. An additional constant (bias term) was included for all variants to enable explicit fitting of the corresponding wild-type free energy terms (∆*Gb,wt* and ∆*Gf,wt*). Separate input layers for each biophysical trait enable their dimensions to be adjusted independently allowing distinct subsets of free energy terms to be inferred for binding and folding. Relatedly, we removed feature variables corresponding to free energy terms that are impossible to infer e.g. the free energy of binding for a given variant in the absence of this phenotype.

All models were trained for 1000 epochs using the Adam optimization algorithm with a learning rate of 0.001. The batch size hyperparameter was optimised by running a parameter sweep ([128, 256, 512, 1024]) and selecting the corresponding value with the lowest validation loss after 100 training epochs. The objective function used was the mean absolute error. As the validation data is used during hyperparameter selection it is not strictly unseen during training. However, we evaluate the inferred model parameters (free energy changes) using completely independent and unseen *in vitro* measurements.

Free energies are calculated directly from model parameters as follows: ∆*Gb*= *xbRT* and ∆*Gf*= *xfRT*, where *T* = 303 K and *R* = 0.001987 kcalK-1mol-1. We estimated the confidence intervals of model-inferred free energies using a Monte Carlo simulation approach. The variability of inferred free energy changes was calculated between ten separate models fit using data from [1] independent random samples of fitness estimates from their underlying error distributions and [2] independent random samples of the validation data consisting of 10% of double aa substitution variants held out during model fitting. Confident inferred free energy changes are defined as those with Monte Carlo simulation derived 95% confidence intervals < 1 kcal/mol. Supplementary Table 7 contains inferred binding and folding free energy changes of mutations for all three proteins.

# Additional references

50. Wrenbeck, E. E. et al. Plasmid-based one-pot saturation mutagenesis. *Nat. Methods* **13**, 928–930 (2016).

51. Harkiolaki, M. et al. Distinct binding modes of two epitopes in Gab2 that interact with the SH3C domain of Grb2. *Structure* **17**, 809–822 (2009).

52. Doyle, D. A. et al. Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ. *Cell* **85**, 1067–1076 (1996).

53. Pusch, C., Schmitt, H. & Blin, N. Increased cloning efficiency by cycle restriction−ligation (CRL). *Tech. Tips Online* **2**, 35–37 (1997).

54. Faure, A. J., Schmiedel, J. M., Baeza-Centurion, P. & Lehner, B. DiMSum: an error model and pipeline for analyzing deep mutational scanning data and diagnosing common experimental pathologies. *Genome Biol*. **21**, 207 (2020).

55. Abadi, M. et al. TensorFlow: Large-scale machine learning on heterogeneous systems. https://www.tensorflow.org/ (2015).

56. Malagrinò, F., Troilo, F., Bonetti, D., Toto, A. & Gianni, S. Mapping the allosteric network within a SH3 domain. *Sci. Rep*. **9**, 8279 (2019).

57. Troilo, F. et al. Folding Mechanism of the SH3 Domain from Grb2. J. *Phys. Chem. B* **122**, 11166–11173 (2018).

58. Chi, C. N. et al. Reassessing a sparse energetic network within a single protein domain. *Proc. Natl. Acad. Sci. U. S. A*. **105**, 4679–4684 (2008).

59. Delgado, J., Radusky, L. G., Cianferoni, D. & Serrano, L. FoldX 5.0: working with RNA, small molecules and a new graphical interface. *Bioinformatics* **35**, 4168–4169 (2019).

60. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).

61. Frazer, J. et al. Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021).

62. Hopf, T. A. et al. Mutation effects predicted from sequence co-variation. *Nat. Biotechnol*. **35**, 128–135 (2017).

63. Ahmed, Z. et al. Grb2 monomer-dimer equilibrium determines normal versus oncogenic function. *Nat. Commun*. **6**, 7354 (2015).

64. Salinas, V. H. & Ranganathan, R. Coevolution-based inference of amino acid interactions underlying protein function. *eLife* **7**, e34300 (2018).

65. Gerek, Z. N. & Ozkan, S. B. Change in allosteric network affects binding affinities of PDZ domains: analysis through perturbation response scanning. *PLoS Comput. Biol*. **7**, e1002154 (2011).

66. Kumawat, A. & Chakrabarty, S. Hidden electrostatic basis of dynamic allostery in a PDZ domain. *Proc. Natl. Acad. Sci. U. S. A*. **114**, E5825–E5834 (2017).

67. Gianni, S. et al. Sequence-specific long range networks in PSD-95/discs large/ZO-1 (PDZ) domains tune their binding selectivity. *J. Biol. Chem*. **286**, 27167–27175 (2011).

68. Kalescky, R., Liu, J. & Tao, P. Identifying key residues for protein allostery through rigid residue scan. *J. Phys. Chem. A* **119**, 1689–1700 (2015).

69. Du, Q.-S., Wang, C.-H., Liao, S.-M. & Huang, R.-B. Correlation analysis for protein evolutionary family based on amino acid position mutations and application in PDZ domain. *PLoS One* **5**, e13207 (2010).

70. Kaya, C., Armutlulu, A., Ekesan, S. & Haliloglu, T. MCPath: Monte Carlo path generation approach to predict likely allosteric pathways and functional residues. *Nucleic Acids Res*. **41**, W249–55 (2013).