## Sequence read processing

The semi-random library, which contained sequences of W and D peptides separated by varying numbers of random amino acids, was processed similarly to the random library in (Ravarani et al. 2018). All processing steps were completed using VSEARCH (Rognes et al. 2016), aside from adapter sequence removal. Forward and reverse read pairs were merged, allowing a maximum of one expected error when considering the quality scores per base. Adapter sequences were removed using cutadapt (Martin 2011). Sequences were then dereplicated across all samples, counting the number of times each unique sequence appeared. Sequences were then filtered to include only those with a length of at least 60 bases and appearing at least twice across the library. The dereplicated sequences were clustered between those with a minimum of 90% sequence similarity, which attempts to prevent two sequences with minor differences from being considered as distinct sequences (Ravarani et al. 2018). These sequence clusters were then considered as centroids, to which the original (merged, without adapters) reads were mapped and counted. For this step, the sequence identity parameter was set to 80%.

The design library contained thousands of sequences containing lengthy repetitive regions (particularly repeats of glycine, but also others), as well as sequences that varied from other sequences at only one amino acid while also containing repetitive regions. To improve the alignment rate, a unique 20-nucleotide sequence barcode was included after the stop codon for all sequences in the design library. However, the above workflow was still not successful in quantification of this library, due to multi-mapping reads and difficulties in merging read pairs, resulting in well over half of reads being discarded. Kallisto was identified as an alternative method for mapping reads to the design library (Bray et al. 2016). Kallisto performs pseudoalignment, a probabilistic method that we used to map reads to the design library sequences (including the sequence barcodes), providing abundances of each sequence in the library. After removing adapters with cutadapt (Martin 2011), reads were pseudoaligned to the design library using Kallisto, with the default kmer size of 31.

## Data analysis

Correlations between the expression levels of biological replicates were calculated to ensure reasonable consistency, and sequences with at least five counts in at least two of the five biological replicates (at baseline) were retained for subsequent analysis. Sequence counts were then averaged across biological replicates, resulting in one value for each sequence in each sample. These were then normalized within sample (transcripts divided by total reads in sample), and then normalized relative to baseline (by calculating the log2 fold change of each sequence versus its counts at baseline). Robust linear regression (implemented in the MASS package in R) was used to estimate the slope of each sequence over time, as a way to measure the functionality of each sequence (Venables and Ripley 2002). Regression of sequence counts vs. day was conducted from day 0 to 4 for most sequences, forcing a y intercept of 0 (because counts were normalized against day 0). Sequences for which the read count dropped and stayed below 3 were regressed from day 0 through the first day at which their read count was below 3. To define a binary cutoff for defining functional versus non-functional sequences the growth slopes of the 5 highest slopes of stop codon sequences (the unique sequences that started with a stop codon) were averaged. For data visualization, all growth slopes were re-centered to this cutoff slope, so that the cutoff slope became zero.

Sequence features such as the presence or number of individual amino acids or multi-residue motifs, balance of aromatic versus acidic residues, and mixing between amino acid residues were used in machine learning analyses. Balance was defined as the difference between the number of aromatic and acidic amino acids, while mixing was defined as the number aromatic-acidic dipeptides plus the number of acidic-aromatic dipeptides. Ridge regression was conducted using the caret package in R (Kuhn 2008). Neural network prediction of functionality was done using the Keras and TensorFlow packages, where each sequence is transformed into a one-hot encoded 3 x 20 matrix (3 amino acids G, W, or D & 20 positions). Neural network architecture was simple consisting of flattened input matrix, two fully connected hidden layers of 60 and 30 nodes each, with a dropout rate of 0.2 after each hidden layer, and finally connected to softmax output layer predicting 1 – functional or 0 – nonfunctional.

Secondary structure prediction was performed with SPOT-1D. The predictions were calculated on each candidate 30-aa long tAD sequence only. The SS3 output of SPOT-1D was then used to assign helicity percent to each candidate tAD sequence. Visualization of structure predictions utilized AlphaFold2 Colab notebook. Sequence structures were predicted for both the candidate tAD and the preceding “linker” (PEFVIRLTIGRAAIMEEQKLISEEDLHMAMG). Visualizations of the candidate tADs were finalized in PyMOL, removing the common “linker” sequence and adding coloring to key amino acids.

Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. “Near-Optimal Probabilistic RNA-Seq Quantification.” *Nature Biotechnology* 34 (5): 525–27. https://doi.org/10.1038/nbt.3519.

Kuhn, Max. 2008. “Building Predictive Models in R Using the Caret Package.” *Journal of Statistical Software, Articles* 28 (5): 1–26. https://doi.org/10.18637/jss.v028.i05.

Martin, Marcel. 2011. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads.” *EMBnet.Journal* 17 (1): 10–12. https://doi.org/10.14806/ej.17.1.200.

Ravarani, Charles Nj, Tamara Y. Erkina, Greet De Baets, Daniel C. Dudman, Alexandre M. Erkine, and M. Madan Babu. 2018. “High-Throughput Discovery of Functional Disordered Regions: Investigation of Transactivation Domains.” *Molecular Systems Biology* 14 (5): e8190. https://doi.org/10.15252/msb.20188190.

Rognes, Torbjørn, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé. 2016. “VSEARCH: A Versatile Open Source Tool for Metagenomics.” *PeerJ* 4 (October): e2584. https://doi.org/10.7717/peerj.2584.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S*. Fourth. New York: Springer. https://www.stats.ox.ac.uk/pub/MASS4/.