

Reporting Summary

Nature Portfolio 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 Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
<input type="checkbox"/>	<input checked="" type="checkbox"/> The exact sample size ( <i>n</i> ) for each experimental group/condition, given as a discrete number and unit of measurement
<input checked="" type="checkbox"/>	<input type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
<input type="checkbox"/>	<input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
<input checked="" type="checkbox"/>	<input type="checkbox"/> A description of all covariates tested
<input type="checkbox"/>	<input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
<input type="checkbox"/>	<input checked="" type="checkbox"/> A full description of the statistical parameters 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)
<input type="checkbox"/>	<input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i> ) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted <i>Give P values as exact values whenever suitable.</i>
<input checked="" type="checkbox"/>	<input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
<input checked="" type="checkbox"/>	<input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
<input checked="" type="checkbox"/>	<input type="checkbox"/> Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i> ), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	The PRM data collection on the Orbitrap Astral instrument was performed with Xcalibur 4.3. The PRM data collection on the ZenoTOF 8600 instrument was performed with SCIEX OS 4.0.
Data analysis	<p>General statistical analysis, data processing, and visualization were performed using R (v4.4.0/4.4.2) and Python (v3.9+). Data manipulation and tidy data principles were implemented using the tidyverse (v2.0.0) suite, including dplyr (v1.1.4), tidyr (v1.3.1), readr (v2.1.5), stringr (v1.5.1), tibble (v3.2.1), purrr (v1.0.2), forcats (v1.0.0), lubridate (v1.9.3), stringi (v1.8.7), here (v1.0.2), and httr (v1.4.7). Data visualization was generated using ggplot2 (v3.5.1), ggpubr (v0.6.0), patchwork (v1.3.0), ComplexHeatmap (v2.20.0), circlize (v0.4.16), ggbeeswarm (v0.7.2), gghalves (v0.1.4), ggpattern (v1.1.1), scales (v1.3.9), RColorBrewer (v1.1-3), and rcartocolor (v2.1.1).</p> <p>Proteomics &amp; Mass Spectrometry: MS data were processed using the Trans-Proteomic Pipeline (TPP v6.3.3/7.1), including the search engines MSFragger (v3.7) and Comet (2019.01.5), with probability modeling provided by PeptideProphet (v6.3.3/7.1). Targeted proteomics analysis, including parallel reaction monitoring (PRM) and data visualization, were performed with Skyline (v25.1.0.142), SCIEX OS (v4), and Xcalibur (v4.3). Bioinformatic protein property analysis utilized the Peptides (v2.4.6) R package.</p> <p>Genomics &amp; Transcriptomics: Sequencing data were processed using the nf-core/rnaseq (v3.1.19) pipeline, utilizing STAR (v2.7.3a), Bowtie2 (v2.5.4), SAMtools (v1.2), and BedTools (v2.31.1) for alignment and genomic arithmetic. Differential expression was assessed with DESeq2 (v1.46.0) using apeglm (v1.28.0) and ashR (v2.2.6) for shrinkage. Ribo-seq data quality control was evaluated using RiboseQC (v0.99.0). Single-Cell Analysis: Single-cell processing and demultiplexing utilized Cell Ranger (v9.0.1), Demuxafy (v1.0.2), and scSplit (v1.0.0). Downstream analysis and integration were performed using Seurat (v5) and scanpy (v1.11.4). Perturbation analysis and E-distance calculations were conducted using pertpy (v1.0.3).</p>

Functional Genomics & CRISPR: CRISPR screen data were normalized and analyzed using Chronos (v2.0.8), with guide RNAs designed via the CRISPick web portal. Real-time proliferation and live-cell imaging were analyzed using Incucyte Software (2023A Rev2 GUI) and growthcurver (v0.3.1). RT-PCR data were processed with QuantStudio™ Design & Analysis (v1.6.1). Functional enrichment was performed with clusterProfiler (v4.14.6).

Structural Biology & Modeling: Putative ncORF-derived protein structures and sequences were modeled and analyzed using AlphaFold3 (v3.0.1), ESM3, and OmegaFold. HLA-binding affinities were predicted using NetMHCpan (v4.1). Coding potential and evolutionary conservation were assessed via PhyloCSF, CodAlignView, and the ORBL tool.

Machine Learning: The Multi-Layer Perceptron Classifier Model and associated statistical tasks utilized TensorFlow (v2.18.0), scikit-learn (v1.5.x-1.6.x), and custom scripts.

Code Availability: Code generated for this manuscript is available at [https://github.com/VanHeeschLab/deutsch\\_kok\\_et\\_al\\_2024](https://github.com/VanHeeschLab/deutsch_kok_et_al_2024) and Zenodo (<https://doi.org/10.5281/zenodo.18878129>). The Multi-Layer Perceptron Classifier Model code is accessible via [https://git.embl.de/ivfimo/machine\\_learning\\_scripts](https://git.embl.de/ivfimo/machine_learning_scripts) and Zenodo (DOI: 10.5281/zenodo.18787106). The code for ORBL is available at [https://github.com/iljung/ORB\\_tools](https://github.com/iljung/ORB_tools) and Zenodo (DOI: 10.5281/zenodo.18749292). Tiling screen local enrichment score scripts are available at [https://github.com/CFVALLS/tiling\\_screens\\_with\\_permutation](https://github.com/CFVALLS/tiling_screens_with_permutation) and Zenodo (DOI: 10.5281/zenodo.18865015).

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 and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

All mass spectrometry data in this manuscript is publicly available through the PeptideAtlas database at <https://peptideatlas.org/> and ProteomeXchange (<https://proteomecentral.proteomexchange.org/>). The Human HLA PeptideAtlas 2023-11 is freely accessible at [https://peptideatlas.org/builds/human/hla/index\\_2023-11.php](https://peptideatlas.org/builds/human/hla/index_2023-11.php) and the Human non-HLA PeptideAtlas 2023-06 is freely accessible at <https://peptideatlas.org/builds/human/non-hla/>. Specific dataset identifiers are listed in Extended Data Table 1. All ribosome profiling data manually inspected in this manuscript are publicly viewable at GWIPS-viz, as detailed in the methods. Mass spectrometry parallel reaction monitoring (PRM) data are deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD066599. Ribosome profiling, RNA sequencing, CRISPR barcode sequencing data for eight cell lines screened in this manuscript as well as OLMALINC knockdown bulk RNAseq and multiplexed single-cell RNAseq are all submitted to the NCBI Short Read Archive as PRJNA1294394. Primary gene and ncORF annotations were sourced from GENCODE (<https://www.genecodegenes.org/>), Ensembl Release 87 (<https://www.ensembl.org/>), UniProtKB/Swiss-Prot 2023 (<https://www.uniprot.org/uniprotkb/>), and the NCBI RefSeq database (<https://www.ncbi.nlm.nih.gov/refseq/>). Tissue-specific RNA-seq expression data were obtained from the Genotype-Tissue Expression (GTEx) portal (<https://gtexportal.org/home/>). Cancer dependency and CRISPR screening data were sourced from the DepMap portal (<https://depmap.org/portal/>). Ribosome profiling (Ribo-seq) data visualization and manual inspections were conducted using GWIPS-viz (<https://riboseq.org/about.html>). The Cancer Cell Line Encyclopedia (CCLE) was used for SNP extraction (<https://sites.broadinstitute.org/ccle/datasets>). The GSEA MSigDB was used to extract hallmark gene sets (<https://www.gsea-msigdb.org/gsea/msigdb>).

## Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

Reporting on race, ethnicity, or other socially relevant groupings

Population characteristics

Recruitment

Ethics oversight

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Field-specific reporting

Please select the one below that is 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/documents/nr-reporting-summary-flat.pdf](https://nature.com/documents/nr-reporting-summary-flat.pdf)

# Life sciences study design

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

Sample size	No statistical methods were used to predetermine sample size. The value of "n" in our manuscript is described in every figure legend and in the methods. For newly performed experiments, "n" refers to biological triplicates for CRISPR screens and quadruplicates for siRNA knockdown experiments. For analyzed datasets and studies, the value of "n" refers to the number of datasets, studies, datapoints, ncORFs, or identified peptides or proteins in these respective datasets. For proteomics database searches and evolutionary analyses, we used all ncORFs satisfying the specified condition rather than a sample chosen from them. The statistical significance of any conclusions are reported as p-values, indicating whether the number of data points was sufficient for any conclusion.
Data exclusions	No data points or samples were excluded from the analyses unless they failed predefined quality control criteria, which were applied uniformly across all datasets. For CRISPR screens, sgRNAs were excluded if they exhibited low counts falling more than three standard deviations below the mean of the total counts. In the pooled OLMALINC loss-of-function assays, cell lines were excluded if they contained missing values (NaNs) across replicates.  For single-cell RNA-sequencing (scRNA-seq), cell lines with low representation—specifically those with fewer than 30–50 cells per identity—were discarded to prevent the inflation of statistical effects and ensure robust downstream analysis.
Replication	For CRISPR assays, lentiviral infections were performed in biological triplicate. OLMALINC siRNA knockdown experiments were performed with 4 technical replicates. All numbers of replicates are listed in the Methods and the respective Figure Legends.
Randomization	The MLP Classifier model was initialized with a maximum of 8000 iterations and a random state of 42 to ensure reproducibility.  For protein structure predictions, ncORF sequences were randomly shuffled five times to assess the contribution of microprotein length and sequence composition to the AlphaFold structure prediction.  For ORBLq calculations, randomly selected sets of 1,000 untranslated control ncORF sequences were selected as size, position, and ORF-type matched controls.
Blinding	Blinding was not applicable because the study did not involve experimental interventions or outcome assessment subject to investigator bias.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern
<input checked="" type="checkbox"/>	<input type="checkbox"/> Plants

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Eukaryotic cell lines

Policy information about [cell lines and Sex and Gender in Research](#)

Cell line source(s)	The OLMALINC loss-of-function assays across 485 cell lines were performed as part of the PRISM project at the Broad Institute. Additional cell lines used for CRISPR tiling screens were purchased directly from the American Type Culture Collection (ATCC)
Authentication	Cell line identity was authenticated using a combination of Short Tandem Repeat (STR) profiling and Single Nucleotide Polymorphism (SNP) identification to ensure genomic consistency with reference standards
Mycoplasma contamination	All cell lines were routinely tested for mycoplasma contamination using the Lonza MycoAlert™. All results were confirmed negative prior to experimentation.
Commonly misidentified lines (See <a href="#">ICLAC</a> register)	In accordance with ICLAC guidelines, all cell lines used in this study were cross-referenced against the Register of Misidentified Cell Lines. No commonly misidentified or contaminated cell lines were utilized in these experiments.

## Palaeontology and Archaeology

Specimen provenance	<i>Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.</i>
Specimen deposition	<i>Indicate where the specimens have been deposited to permit free access by other researchers.</i>
Dating methods	<i>If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.</i>
<input type="checkbox"/> Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.	
Ethics oversight	<i>Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.</i>

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Animals and other research organisms

Policy information about [studies involving animals; ARRIVE guidelines](#) recommended for reporting animal research, and [Sex and Gender in Research](#)

Laboratory animals	<i>For laboratory animals, report species, strain and age OR state that the study did not involve laboratory animals.</i>
Wild animals	<i>Provide details on animals observed in or captured in the field; report species and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.</i>
Reporting on sex	<i>Indicate if findings apply to only one sex; describe whether sex was considered in study design, methods used for assigning sex. Provide data disaggregated for sex where this information has been collected in the source data as appropriate; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex-based analyses where performed, justify reasons for lack of sex-based analysis.</i>
Field-collected samples	<i>For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.</i>
Ethics oversight	<i>Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.</i>

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration	<i>Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.</i>
Study protocol	<i>Note where the full trial protocol can be accessed OR if not available, explain why.</i>
Data collection	<i>Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.</i>
Outcomes	<i>Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.</i>

## Dual use research of concern

Policy information about [dual use research of concern](#)

### Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes
<input type="checkbox"/>	<input type="checkbox"/> Public health
<input type="checkbox"/>	<input type="checkbox"/> National security
<input type="checkbox"/>	<input type="checkbox"/> Crops and/or livestock
<input type="checkbox"/>	<input type="checkbox"/> Ecosystems
<input type="checkbox"/>	<input type="checkbox"/> Any other significant area

## Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes
<input type="checkbox"/>	<input type="checkbox"/> Demonstrate how to render a vaccine ineffective
<input type="checkbox"/>	<input type="checkbox"/> Confer resistance to therapeutically useful antibiotics or antiviral agents
<input type="checkbox"/>	<input type="checkbox"/> Enhance the virulence of a pathogen or render a nonpathogen virulent
<input type="checkbox"/>	<input type="checkbox"/> Increase transmissibility of a pathogen
<input type="checkbox"/>	<input type="checkbox"/> Alter the host range of a pathogen
<input type="checkbox"/>	<input type="checkbox"/> Enable evasion of diagnostic/detection modalities
<input type="checkbox"/>	<input type="checkbox"/> Enable the weaponization of a biological agent or toxin
<input type="checkbox"/>	<input type="checkbox"/> Any other potentially harmful combination of experiments and agents

## Plants

Seed stocks	NA
Novel plant genotypes	NA
Authentication	NA

## 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 <i>May remain private before publication.</i>	<i>For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.</i>
Files in database submission	<i>Provide a list of all files available in the database submission.</i>
Genome browser session (e.g. <a href="#">UCSC</a> )	<i>Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.</i>

### Methodology

Replicates	<i>Describe the experimental replicates, specifying number, type and replicate agreement.</i>
Sequencing depth	<i>Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.</i>
Antibodies	<i>Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.</i>
Peak calling parameters	<i>Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.</i>

Data quality

Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.

Software

Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

## 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

Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.

Instrument

Identify the instrument used for data collection, specifying make and model number.

Software

Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.

Cell population abundance

Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.

Gating strategy

Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.

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

## Magnetic resonance imaging

### Experimental design

Design type

Indicate task or resting state; event-related or block design.

Design specifications

Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.

Behavioral performance measures

State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).

### Acquisition

Imaging type(s)

Specify: functional, structural, diffusion, perfusion.

Field strength

Specify in Tesla

Sequence &amp; imaging parameters

Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.

Area of acquisition

State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.

Diffusion MRI

☐

Used

☐

Not used

### Preprocessing

Preprocessing software

Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).

Normalization

If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.

Normalization template	<i>Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.</i>
Noise and artifact removal	<i>Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).</i>
Volume censoring	<i>Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.</i>

## Statistical modeling & inference

Model type and settings	<i>Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).</i>
Effect(s) tested	<i>Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.</i>
Specify type of analysis:	<input type="checkbox"/> Whole brain <input type="checkbox"/> ROI-based <input type="checkbox"/> Both
Statistic type for inference	<i>Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.</i>
(See <a href="#">Eklund et al. 2016</a> )	
Correction	<i>Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).</i>

## Models & analysis

n/a	Involved in the study	
<input type="checkbox"/>	<input type="checkbox"/> Functional and/or effective connectivity	
<input type="checkbox"/>	<input type="checkbox"/> Graph analysis	
<input type="checkbox"/>	<input type="checkbox"/> Multivariate modeling or predictive analysis	
Functional and/or effective connectivity	<i>Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).</i>	
Graph analysis	<i>Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).</i>	
Multivariate modeling and predictive analysis	<i>Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.</i>	