Supplementary materials for: **OpenCell: proteome-scale endogenous tagging enables the cartography of human cellular organization** (Cho, Cheveralls, Brunner, Kim, Michaelis, Raghavan et al. *bioRxiv* 2021)

LIST OF SUPPLEMENTARY FIGURES

LIST OF SUPPLEMENTARY TABLES

Each Supplementary Table contains a specific "read_me" tab that describes its content in detail.

MATERIAL AND METHODS

Cell culture and CRISPR engineering

Cell culture. HEK-293T cells (ATCC CRL-3216) were cultured in DMEM high-glucose medium (Gibco, cat. #11965118) with 10% fetal bovine serum (Omega Scientific, cat. #FB-11), supplemented with 2mM glutamine (Gibco, cat. #25030081), penicillin and streptomycin (Gibco, cat. #15140163). All cell lines were maintained at 37°C and 5% CO2 and routinely tested for the absence of mycoplasma.

Fluorescent library design. mNeonGreen is monomeric green fluorescent protein ~2x brighter than GFP. We used the split-mNeonGreen$_2$ system for functional tagging, which separates last mNeonGreen$_2$ beta-strand (mNG11) from the rest of the fluorescent protein (mNG1-10)(*18*). Upon co-expression in the same cell, mNG1–10 and mNG11 stably assemble and reconstitute a functional FP. A parental cell line constitutively expressing mNG1-10 was first generated by lentiviral transduction (from pSFFV_mNG$_2$1-10, Addgene #82610). All successive cell lines were generated from this parental HEK293T$^{mNG1-10}$ cell line by incorporation of the mNG11 fragment at either N- or C- terminus of the genomic sequence of a target protein via CRISPR/Cas9 based genome editing. Our mNG11 fusion constructs include a HRV 3C cleavable linker(*72*) that can be used optionally for elution from an affinity capture matrix (16 a.a. tag + 14 a.a. linker, full sequences in Suppl. Table X). To minimize the risk of functional perturbation, we stringently selected integration sites (N- or C-terminus) by systematically curating the literature for data supporting the functional integrity of fusion proteins (or by requesting advice from cell biology experts for specific proteins). We also used 3D PDB structures whenever available to identify sites that avoid protein-protein interaction interfaces. Because our split-FP system does not enable detection in the lumen of organelles (this requires split constructs harboring appropriate signal sequences(*73*)), fusions with membrane proteins were restricted to cytoplasmic termini, ensuring first that no annotated regulatory sequences (e.g., signal sequences) were compromised. In total, we used available supporting data to inform 62 % of insertion sites, and 3 % were constrained by membrane protein topology. In the absence of prior information, insertion choice was based on avoiding annotated regulatory sites.

Overall genome engineering pipeline. To enable the expression of fluorescent fusion from endogenous genomic loci, we used an established high-throughput CRISPR/Cas9 method for gene editing by homologous recombination(*15*). In brief, *S. pyogenes* Cas9/guide RNA complexes were pre-assembled in vitro, mixed with short single-stranded oligo-nucleotide homology donors and delivered into HEK293T$^{mNG1-10}$ cells by electroporation in 96-well plates (see below). For each genomic insertion, the choice of guide RNA and associated homology donor sequence (which contains the mNG11 payload flanked by short sequences of genomic homology to the targeted insertion site) was automated using *crispycrunch*(*74*), an open-source CRISPR design software available at github.com/czbiohub/crispycrunch and as a web-app at crispycrunch.czbiohub.org/. *crispycrunch* selects a guide RNA closest to a desired genomic insertion site while also minimizing any off-target guide RNA activity, and if needed introduces silent mutations to inactivate guide RNA biding and re-cutting after successful homologous recombination(*74*). gRNA and homology donor sequences for all targets are found in Suppl. Table X.

Cell engineering and selection *S. pyogenes* Cas9 protein (pMJ915 construct, containing two nuclear localization sequences) was expressed in E. coli and purified by the UC Berkeley Macrolab following protocols described by Jinek et al(*75*). Cells were synchronized by nocodazole treatment (200ng/mL for 15-18h_ to enhance homologous recombination(*21*). RNP complexes were freshly assembled with 50 pmol Cas9 protein and 65 pmol gRNA prior to electroporation, and combined with HDR template in a final volume of 10 µL. First, 0.5 µL gRNA (130 µM stock) was added to 2.35 µL high-salt RNP buffer {580 mM KCl, 40 mM Tris-HCl pH 7.5, 20% v/v glycerol, 2 mM TCEP-HCl pH 7.5, 2 mM MgCl2, RNAse-free} and incubated at 70°C for 5 min. 1.25 µL of Cas9 protein (40 µM stock in Cas9 buffer, ie. 50 pmol) was then added and RNP assembly carried out at 37°C for 10 min. Finally, HDR templates and sterile RNAse-free H2O were added to 10 µL final volume. Electroporation was carried out in Amaxa 96-well shuttle Nucleofector device (Lonza) using SF solution (Lonza) following the manufacturer's instructions. Cells were washed with PBS and resuspended to 10,000 cells/µL in SF solution (+ supplement) immediately prior to electroporation. For each sample, 20 µL of cells (ie. 200,000 cells) were added to the 10 µL RNP/template mixture. Cells were immediately electroporated using the CM130 program, after which 100µL of pre-warmed media was added to each well of the electroporation plate to facilitate the transfer of 25,000 cells to a new 96-well culture plate containing 150µL of pre-warmed media. Electroporated cells were cultured for >5 days and transferred to 12-well plates prior to selection by fluorescence-activated cell sorting (FACS). For each target, 1,200 cells

from the top 1% fluorescent cell pool were isolated on a SH800 instrument (Sony biotechnology) and collected in 96-well plates.

Genotype analysis. For each polyclonal pool of engineered cells, the genotype of CRISPR-edited alleles was characterized by amplicon sequencing. Gene-specific primers were designed using Primer3, with a target amplicon length of 270bp and a maximum at 500bp. gDNA was first extracted by cell lysis using QuickExtract DNA Extraction Solution (Lucigen). From a confluent culture in 96-well plate, media was removed, cells were washed 1x in DPBS and resuspended in 50 µL QuickExtract. The cell layer was detached by repeated pipetting and transferred to a PCR plate for incubation. The lysate was incubated as follows {65°C for 20 min, 98°C for 5min, 4°C final}. gDNA was used directly from this preparation. Amplicon Libraries were created using a two-step PCR protocol: the first PCR amplifies the target genomic locus and adds universal amplification handle sequences, while the second PCR introduces index barcodes using the universal handles. PCR1: this PCR uses a "reverse touchdown" method designed to accommodate a number of different annealing temperatures for a number of different targets. 50-µL PCR reactions were set using 2x KAPA HiFi Hotstart reagents (Roche) with 2µL extracted gDNA, 80pmol each primer and betaine to 0.8M final concentration. PCR conditions: 95°C 3min; 3 cycles of {98°C for 20s, 63°C for 15s, 72°C for 20s}, 3 cycles of {98°C for 20s, 65°C for 15s, 72°C for 20s}, 3 cycles of {98°C for 20s, 67°C for 15s, 72°C for 20s}; 17 cycles {98°C for 20 s, 69°C for 15 s, 72°C for 20s} then 72°C for 1min; 4°C final. PCR2: amplicons were diluted 1:100 and 1 µL was used into a 40-µL barcoding reaction using 20 µL 2x KAPA HiFi Hotstart reagents (Roche) and 80pmol each barcoded primer. PCR conditions: 95°C 3min and 12 cycles of {98°C for 20s, 68°C for 15s, 72°C for 12s} then 72°C for 1min; 4°C final. Barcoded amplicons were analyzed using capillary electrophoresis (Fragment Analyzer, Agilent), pooled and purified using magnetic beads. Sequencing was performed on an Illumina Miseq V3 platform (paired-end 2x300bp) using standard P5/P7 primers. Genotype analysis was performed using CRISPRESSO2, which allowed to quantify three classes of alleles for each targeted locus: un-modified (wild-type), alleles integrated with mNG11 by homologous recombination, and alleles containing non-functional mutations as a result of competing DNA repair mechanisms. Primer sequences and genotype analysis for all targets are found in Suppl. Table X. Despite multiple attempts, genotyping PCR could not be successfully performed for 70 targets (5% of the total set), most often involving genes with extreme GC content or highly repetitive sequences.

Immuno-precipitation / mass-spectrometry

Overall strategy. mNG11-tagged proteins were isolated from digitonin-solubilized lysates using anti-mNeonGreen nanobody capture. Triplicate protein samples were digested "on-bead" for bottom-up proteomics analysis(*24*), and peptides were quantified using label-free mass spectrometry on a timsTOF Pro instrument (Bruker Daltonics).

Sample preparation. Confluent 12-well cultures ($0.8x10^6$ cells/sample) were washed twice with 1 ml of D-PBS (no divalent). 200 µl ice-cold lysis buffer A {50 mM HEPES pH 7.5, 150 mM KOAc, 5 mM NaCl, 2 mM MgOAc, 1 mM $CaCl_2$, 15% Glycerol, 1.5 % Digitonin (high purity, Calbiochem), Protease- and Phosphatase inhibitor (Halt, Pierce), 0.1% benzonase (Millipore Sigma)} were added to each well, cells were lysed by strong pipetting and the solution was transferred into a pre-chilled 96-well PCR plate. Per 96-well plate, 330 µl magnetic mNG-Trap slurry (magnetic agarose, Chromotek) was washed three times with buffer B {50 mM HEPES pH 7.5, 150 mM KOAc, 5 mM NaCl, 2 mM MgOAc, 1 mM $CaCl_2$, 15% Glycerol, 0.1 % Digitonin} and resuspended in 2,150 µl Buffer A. The cell lysate was incubated for 1h at 4°C, rotating. The insoluble cell fraction was pelleted for 30 min at 1800xg in a table-top centrifuge at 4°C, followed by supernatant transfer into a new plate pre-loaded with 20 µl of the washed bead slurry per well. Tagged proteins were captured by incubation for 2h at 4°C, rotating. Following capture and using a 96-well magnet, beads were washed (per well) with 200 µl buffer B (incubation for 5 min at 4°C, rotating), 2x 200 µl buffer B (no incubation) and a final 1x 200 µl buffer C to remove digitonin {50 mM HEPES pH 7.5, 150 mM KOAc, 5 mM NaCl, 15% Glycerol, 0.01% glyco-diosgenin (Avanti)}. Supernatant was removed and 50 µl of digestion buffer 1 {6 M Urea, 50 mM Tris-HCl, pH 8.5, 1 mM DTT, 2 ng/µl LysC protease (Wako Chemicals)} was added to each well, followed by overnight digestion at 30°C on a thermomixer, gently shaking. The next day, 100 µl digestion buffer 2 {50 mM Tris-HCl, pH 8.5, 8.25 mM iodoacetamide, 2 ng/µl LysC} was added to each well and incubated for ~6 hours at 30°C on a thermomixer in the dark, gently shaking. The digestion was finally quenched with 15 µl of 10 % TFA. Quenched samples were vortexed, flash-frozen and stored at -80 °C until further use for LC-MS analysis preparation.

EvoSep chromatography. We used the EvoSep liquid chromatography system for sample processing(*76*). EvoTips (EvoSep Gmbh) were activated for 5 min with 1-Propanol at RT, followed by a wash step with 50 µl Buffer A (99.9 % ddH2O, 0.1 % Formic Acid) and centrifugation at 600 xg for 1 min at RT. The flow-through was discarded and activated EvoTips were placed in an EvoTip-box reservoir filled with Buffer A. After on-bead digestion, captured protein samples were thawed for

5 min at 600 rpm and 25°C on a thermal shaker and placed on a 96-well magnet holder to remove magnetic beads. The whole sample (~150 µl) was transferred to activated EvoTips, followed by two consecutive centrifugation steps at 600xg for 1 min and RT, discarding flow-through after the first spin. Peptide-loaded EvoTips were washed once with 50 µl Buffer A and centrifuged at 600xg for 1 min at RT. The flow-through was discarded and 150 µl of Buffer A was added to each EvoTip followed by a centrifugation step for 20 sec at 600xg RT. Loaded EvoTips were then transferred into the 96-well EvoTip-box reservoir filled with Buffer A and transferred onto the EvoSep autosampler for LC-MS analysis. Pulldowns were acquired in triplicates and injected to the mass spectrometer while spacing replicates to prevent any bias.

Liquid-chromatography. For separating peptides by hydrophobicity and eluting them into the mass spectrometer, we used an EvoSep One1 liquid chromatography system (EvoSep, Gmbh) and analyzed purified petides with a standard 21 min method (60 samples per day). We used a 15 cm × 150 µm ID column with 1.9 µm C18 beads (PepSep) coupled to a 20 µm ID electrospray emitter (Bruker Daltonics). Mobile phases A and B were 0.1 % FA in water and 0.1 % FA in ACN, respectively. The EvoSep system was coupled online to a trapped ion mobility spectrometry quadrupole time-of-flight mass spectrometer(*77*) (timsTOF Pro, Bruker Daltonics) equipped with via a Captive nano-electrospray ion source.

Mass spectrometry. Mass spectrometric analysis was performed in a data-dependent (dda) PASEF mode. For ddaPASEF, 1 MS1 survey TIMS-MS and 4 PASEF MS/MS scans were acquired per acquisition cycle. The cycle overlap for precursor scheduling was set to 2. Ion accumulation and ramp time in the dual TIMS analyzer was set to 50 ms each and we analyzed the ion mobility range from 1/K0 = 1.3 Vs cm-2 to 0.8 Vs cm-2. Precursor ions for MS/MS analysis were isolated with a 2 Th window for m/z < 700 and 3 Th for m/z >700 in a total m/z range of 100-1,700 by synchronizing quadrupole switching events with the precursor elution profile from the TIMS device. The collision energy was lowered linearly as a function of increasing mobility starting from 59 eV at 1/K0 = 1.6 VS cm-2 to 20 eV at 1/K0 = 0.6 Vs cm-2. Singly charged precursor ions were excluded with a polygon filter (otof control, Bruker Daltonics). Precursors for MS/MS were picked at an intensity threshold of 2,000 arbitrary units (a.u.) and re-sequenced until reaching a 'target value' of 24,000 a.u. considering a dynamic exclusion of 40 s elution. Capillary voltage was set to 1,750 V and dry gas temperature to 180°C.

Raw Data Processing. MS raw files were processed using MaxQuant (v1.6.10.43)(*78*, *79*), which extracts features from four-dimensional isotope patterns and associated MS/MS spectra, on a computing cluster (SUSE Linux Enterprise Server 15 SP2) utilizing UltraQuant. Files were processed in several batches of appriximately 1000 files each and searched against the human Uniprot databases (UP000005640_9606.fa, UP000005640_9606_additional.fa). False-discovery rates were controlled at 1% both on peptide spectral match (PSM) and protein levels. Peptides with a minimum length of seven amino acids were considered for the search including N-terminal acetylation and methionine oxidation as variable modifications and cysteine carbamido-methylation as fixed modification, while limiting the maximum peptide mass to 4,600 Da. Enzyme specificity was set to LysC cleaving c-terminal to lysine. A maximum of two missed cleavages were allowed. Maximum precursor and fragment ion mass tolerance were searched as default for TIMS-DDA data and the main search tolerance was reduced to 20 ppm. Peptide identifications by MS/MS were transferred by matching four-dimensional isotope patterns between the runs (MBR) with a 0.7-min retention-time match window and a 0.05 1/K0 ion mobility window. Protein quantification was performed by label-free quantification using a minimum ratio count of 1.

Data availability. All mass spectrometry raw data and MaxQuant output tables are deposited to the ProteomeXchange Consortium(*80*) via the PRIDEpartner repository and will be publicly available upon final publication (accession PXD024909).

Live-cell imaging

Sample preparation. Live-cell imaging was performed on 96-well glass-bottom plates (Greiner Bio One, cat. #655891) coated with 50μg/ml fibronectin (Corning, cat. #356008). Cells were seeded on an imaging plate 28-32 hours before imaging at 15,000 cells per. Before imaging, cells were counter-stained with the live-cell DNA dye Hoechst 33342 (Invitrogen, cat. #H3570) by incubation for 30 minutes at 37°C in 150 μl of Hoechst diluted to 1μg/mL in culture media. Media was then replaced with phenol-free DMEM (Gibco, cat. #21063029) supplemented with 10% FBS. Hoechst staining was performed three to four hours prior to imaging to provide the cells time to recover from any mechanical stress due to medium changes.

Live-cell fluorescence microscopy. Cells were imaged on a DMI-8 inverted microscope (Leica) equipped with a Dragonfly spinning-disk confocal system (Andor), a 63x 1.47NA oil objective (Leica),

and a 16-bit iXon Ultra 888 EMCCD camera (Andor, pixel size: 13x13 μm^2). A pinhole size of 40µm was used with an EM gain of 400. Cells were maintained at 37°C and 5% CO2 during image acquisition by an stage-top incubator (Okolab, H101-K-Frame). The microscope was controlled using the open-source microscope-control software MicroManager (version 1.4.22).

Automated confocal acquisition. We automated the imaging of 96-well plates using a custom acquisition script, written in Python, combined with a custom MicroManager plugin (mm2python; github.com/czbiohub/mm2python) to expose the MicroManager APIs in a Python environment. This script selected optimal fields of view (FOVs) at which to acquire confocal z-stacks by using a machine-learning model to assign a quality score to the FOVs at a set of different positions in each well. Briefly, at each position, the script acquired a single 2D snapshot of the Hoechst staining, segmented the nuclei in the snapshot, and calculated an array of features associated with the distribution of nuclei within the FOV. The script then used a pre-trained random-forest regression model (see below) to predict a quality score for the FOV from this set of features. This process was repeated at each of 25 different positions in each well, and then the script selected the positions with the highest-scoring FOVs to revisit for confocal z-stack acquisition. At each of these selected positions, the focal plane was centered on the cell layer using a laser-based Adaptive Focus Control system (Leica) and confocal z-stacks, consisting of 110 z-slices at a spacing of 0.2µm, were acquired. The exposure settings for the mNeonGreen channel were determined dynamically for each target using a custom auto-exposure algorithm that iteratively adjusted the exposure time and laser power until the maximum pixel intensity was just below or just above an intensity of 2^{15} (half of the dynamic range of the camera). For dim targets for which this condition could not be met, the script fell back to a hard-coded absolute maximum exposure time and laser power to minimize both acquisition time and photobleaching. The exposure settings for the Hoechst stain were manually selected and held constant for all targets. The model used by the script to predict the FOV quality scores was trained prior to acquisition using a set of 3800 FOV snapshots that were manually assigned to one of three grades: "poor," "mediocre," or "good." These grades were mapped to a continuous response variable by assigning them values of -1, 0, and 1, respectively, and a random forest regression model (scikit-learn) was trained to predict this value. The out-of-bag estimated R^2 was 0.86 and scores predicted for a withheld set of test snapshots were also evaluated by manual inspection. The trained model was cached and imported at acquisition time by the acquisition script. The acquisition script, trained FOV-scoring model, autoexposure algorithm, and other associated microscope-control methods are available online at github.com/czbiohub/opencell-microscopy-automation.

Data analysis – proteomics

Statistical detection of protein interactions. Statistical analysis was performed according to methods described in Hein et al.(*7*), with modifications. Protein identifications were filtered, removing common contaminants, hits to the reverse decoy database as well as proteins only identified by modified peptides. We required that each protein be quantified in all replicates from the IP-MS samples of at least one cell line and used log2 MaxQuant LFQ intensities for all analyses. Rather than imputing missing values, robust null control sets were generated for statistical enrichment analysis of each protein group by pooling triplicate data from an average of 349 unrelated samples. The null control sets might contain triplicate samples that are outliers and would be considered significant interactions. Presence of these samples lead to underestimation of enrichment and could mask significance of some interactions. We systematically removed these outliers from the negative control sets using Student's t-test – any sample of triplicates that had a p-value < 0.001 were excluded. From the filtered pool, we approximated the true mean and the true standard deviation of the null set by bootstrapping via sampling with replacement. The approximated mean and standard deviation of the null set was then used for student's t-test to calculate the statistical significance of triplicates. Any missing values in the triplicate sample set was then replaced with the mean of the null set. Enrichment was calculated by subtracting the mean of the triplicates by the mean of the null set, and was normalized to account for variability within each protein through division by the standard deviation of the null control set. Our statistical strategy to define of interactors is described in Figure S2A-B and supported by a quantitative estimation of precision and recall.

Precision / recall analysis of the interactome. For a quantitative evaluation of our statistical approach, and to compare the quality of OpenCell against reference interactome datasets, we created a framework to measure precision and recall. In the absence of established ground truth for human protein interactions, we indirectly derived measurements of precision and recall. For recall, we calculated the coverage in a given dataset of interactions curated in the human CORUM database(*26*), as a percentage of all possible CORUM interactions given the set of baits in that dataset. For calculating precision, we used the assumption that two interactors should have localization patterns that at least partially overlap. As an independent ground truth set for protein localization, we used the quantitative analysis of the HeLa proteome from Itzhak et al. Using these annotations, we categorized localization into four broad classes: exclusively nuclear, exclusively cytoplasmic, exclusively organellar,

and multi-localizing (i.e., any non-exclusive localization). To calculate precision, we consider any two interactors that overlap in exclusive localization to be true positives, and those that do not overlap localization annotations at all to be false positives, with multi-localizing proteins allowed to interact agnostically (Fig. S2B).

Protein stoichiometry measurements. Calculation of interaction stoichiometries was performed as in Hein et al by dividing LFQ intensities by the number of theoretically observable peptides for each protein. Two stoichiometry measurements are measured. First, the stoichiometry of abundance of a given interactor, relative to the abundance of the corresponding bait, in a given pull-down ("interaction stoichiometry"). Second, the stoichiometry of abundance of a given interactor, relative to the abundance of the corresponding bait, in a whole cell lysate ("cellular abundance stoichiometry"). For proteins that are not detected in whole cell lysates (due to lack of measurable peptides, for example in the absence of lysine residues), protein abundances were imputed from RNA-Seq data by interpolating from a regression of RNA-Seq tpm vs. protein expression measured by mass spectrometry in our dataset.

Network Analysis. For graph-based clustering of the entire interactome network, we weighted edges using the interaction stoichiometry between any two nodes. We utilized Markov clustering(*36*) at various inflation parameters and evaluated clustering performance using the k-clique method described in Drew et al(*71*) using CORUM complexes as the ground truth (minimizing overlapping clusters by using a filter of Jaccard distance at 0.6). Our final clustering analysis used an inflation parameter of 3.0 (Fig. S2G). The resulting clusters were pruned to remove any node included in a luster on the basis of a single edge, which defines what we describe as protein communities. We then utilized another round of MCL clustering to identify sub-clusters within each cluster by considering only highly stoichiometric interactions (interaction stoichiometries between 0.05 and 10, and cellular abundance stoichiometry between 0.1 and 10). The resulting sub-clusters represented highly stable core clusters within the original communities.

Measurement of biophysical properties of proteins. Biophysical properties were calculated using the *ProteinAnalysis* package from BioPython(*81*). Calculation of disorder in protein regions was performed using the IUPred2A algorithm(*82*). All calculations were performed on the set of 100 amino-acid windows covering each protein sequence.

Interaction distance calculations. To calculate distances between interaction signatures (Fig. 5), we encoded all protein interactors (i.e., all preys) by the vector formed by their interaction stoichiometries (log10) in individual pulldowns with all OpenCell baits. The vector coordinate is zero for pulldowns in which a given protein is not found to interact. Protein that only had interaction with a single bait within the full dataset were filtered out. To prevent baits with very large numbers of interactors (>100 interactors) from disproportionally biasing cosine distances, for these pulldowns only proteins with interaction stoichiometry over 0.01 were considered.

Data analysis – imaging

Consensus localization encodings. Protein localization patterns were encoded from the raw confocal images using a customized variant of the vector-quantized autoencoder architecture VQ-VAE-2(*83*). The image preprocessing, autoencoder architecture, and model training are described in detail in an accompanying manuscript(*44*). Briefly, confocal z-stacks were reduced to two dimensions by a maximum-intensity z-projection and normalized to control for variation in intensity. Regions of interest 200x200 pixels in size were centered on individual nuclei and cropped from each z-projection to generate a set of 50-200 cropped images for each tagged protein. These images were randomly partitioned into a training set and a test set. After training the model on the images in the training set, the images in the test set were encoded, and the resulting latent-space vectors from the VQ2 layer of the network were flattened to obtain a localization encoding for each image in the test set as a 9216-dimensional vector. The encodings of all images for each tagged protein were then averaged to obtain a single consensus encoding for each tagged protein.

Analysis of image localization encodings. The matrix of consensus localization encodings for all OpenCell targets was analyzed using the *scanpy* package(*84*). Briefly, the dimensionality of the consensus encodings was reduced using PCA and the first 200 PCs, which captured 96% of the variance, were retained for downstream analysis. The UMAP algorithm(*55*) was used to embed the encodings in two-dimensional space using 10 nearest neighbors, the Euclidean distance metric, and a minimum embedding distance of zero. The encodings were clustered using the Leiden graph-based clustering algorithm(*46*) with a resolution parameter of 30 and the weighted adjacency matrix calculated by the UMAP algorithm (again with 10 nearest neighbors). Finally, the Pearson correlation coefficient between the top 200 PCs of the localization encodings was used to quantify the localization similarity between OpenCell targets.

Hierachical analysis of interactions and localization patterns.

Hierarchical clustering of interactome and image-localization clusters. To explore the relationships between the 182 localization clusters or the 300 interactome communities, we employed the Paris algorithm, an agglomerative graph-based hierarchical clustering algorithm(*54*). The algorithm was initialized with a network of nodes representing the initial clusters (either the localization clusters or the interactome communities) and edge weights between the initial clusters were calculated according to the definition of the cluster pair sampling ratio used in the Paris algorithm.

Gene Ontology enrichment analysis. To analyze enrichment of GO terms in a given hierarchical protein group, we utilized the PANTHER gene list analysis API(*85*) using Fisher exact test for significance testing. Enrichment of GO terms was tested against a reference set of either all OpenCell targets for the imaging dataset, or all proteins found in communities for the interactome dataset.

Website development

The OpenCell website consists of a frontend (the website itself) and a backend. The frontend is a single-page web application written with React, a modern JavaScript library for building modular user interfaces. The backend is a PostgreSQL database and a REST API written in Python using Flask and SQLAlchemy. Together, the database and API provide the metadata, the mass-spec interaction data, and the confocal image data required to populate the frontend. For efficiency, the 3D confocal stacks are transferred to the app as two-dimensional tiled arrays of confocal slices, saved as a compressed JPEG images to enable fast download times. To maximize responsiveness, the web app makes API requests dynamically and asynchronously so that it loads, in parallel, only the data required to update the state of the app in response to a given user input. Both the backend and frontend rely on many open-source packages. In particular, the 3D rendering of confocal stacks relies on Three.js, the interactive scatterplots are built with d3.js, and the interaction networks are built with Cytoscape.js. The backend relies on SQLAlchemy and Flask as well as the Python data-science stack, including pandas, NumPy, SciPy, and scikit-image. All source code for the app is available on GitHub at github.com/czbiohub/opencell.

♦