PNAS
www.pnas.org

Supplementary Information for

Deep mutational analysis reveals functional trade-offs in the sequences of EGFR autophosphorylation sites

Aaron J. Cantor, Neel H. Shah and John Kuriyan

Corresponding Author: John Kuriyan
Email: kuriyan@berkeley.edu

This PDF file includes:

www.pnas.org/cgi/doi/10.1073/pnas.1803598115

Supplementary Methods

Recombinant Proteins

Amino acid sequences for the proteins used in this study are listed in **Table S1.**

c-Src kinase domain. The chicken c-Src kinase domain (corresponding to residues 257–525 human c-Src) was expressed and purified as previously described (1, 2).

FKBP– and FRB–EGFR intracellular module fusions. Constructs consisted of a non-cleavable His_{10} tag followed by either human FKBP1A (residues 3–108) or the FRB domain of human mTOR (residues 2018–2112). These were fused N-terminally to human EGFR residues 663–1186, including the C-terminal part of the juxtamembrane element and the full C-terminal tail. The construct terminated with a FlAsH-tag binding sequence (CCPGCC), which was not utilized in this study. These EGFR constructs were inserted into the pFastBac1 vector and expressed in Sf9 cells with the Bac-to-Bac system (ThermoFisher Scientific), in ESF 921 medium (Expression Systems). The EGFR proteins were purified by TALON Co^{2+} affinity, anion exchange, and size exclusion chromatography. The proteins were concentrated and stored in a buffer containing 50 mM Tris pH 8.0, 150 mM NaCl, 10% glycerol, and 1 mM TCEP.

Peptides. EGFR tyrosine-containing peptides were produced recombinantly in *E. coli*. Coding sequences for 21-residue peptides were inserted C-terminal to yeast SMT3 with an N-terminal His_6 tag in a pET-derived vector. The proteins were purified with Ni^{2+}-affinity and anion exchange chromatography. The SUMO moiety was removed with yeast Ulp1 protease, and cleaved peptides were isolated with reverse-phase HPLC in water with 0.1% TFA, with elution by a gradient of

acetonitrile with 0.1% TFA. Peptides were lyophilized, resuspended in water, and dialyzed against 100 mM HEPES pH 7.5. Peptide concentrations were determined by Tyr absorbance at 275 nM (1410 M^{-1} cm^{-1} extinction coefficient). The Protein Kinase C Tyr 313 15-mer peptide was purchased from Elim Biopharmaceuticals (Hayward, CA).

Tandem SH2 and PTB domain GFP fusions. The tandem human Grb2 SH2 (residues 55 to 152) and human Shc1 PTB domain (residues 147 to 318) DNA sequences were constructed by overlap extension PCR and inserted into a pET-based vector, with an N-terminal His_6-tag and TEV protease site and C-terminal eGFP. The two copies of each binding domain were connected by a 20-residue Gly/Ser linker, and these were connected with a 10-residue linker to GFP. These proteins were expressed in *E. coli* and purified using Ni^{2+}-affinity, anion exchange, and size exclusion chromatography. His_6 tags were removed prior to size exclusion chromatography. Protein concentration was determined by absorbance of GFP at 488 nm.

Bacterial Surface-Display and Deep Sequencing

Human-pTyr library phosphorylation analysis*. The specificity profile for the dimerized EGFR intracellular module against the Human-pTyr library of phosphosites was determined as described by (2). Details of the library construction and contents can be found in this reference. Briefly, *E. coli* (strain MC1061) expressing the library were subjected to phosphorylation by 0.5 μM dimerized EGFR intracellular module for 15 minutes at room temperature. These conditions were determined to produce a median phosphorylation level of ~30% compared to a fully phosphorylated sample, stained with Milli-Mark anti-Phosphotyrosine 4G10 phycoerythrin

*All scripts used to analyze data and generate figures will be made available upon request.

antibody conjugate (Millipore Sigma), as judged by flow cytometry. The top 25% brightest cells in the PE channel were sorted on a BD Influx cell sorter. ~4 million sorted and unsorted cells were harvested by centrifugation and boiled to release DNA. The peptide-coding portion of the surface-display plasmid was PCR amplified in two steps to append Illumina indices and adapters. The samples were multiplexed and sequenced in paired-end mode on an Illumina MiSeq sequencer. Paired-end reads were assembled, trimmed, and mapped to peptide sequences in the library. Peptide sequence read frequency in each sample was calculated as the number of reads for a given peptide divided by the total number of aligned reads for that sample. The ratio of read frequency of each peptide in the sorted sample divided by the frequency of the same peptide in the unsorted sample gives the read frequency ratio plotted and analyzed in **Fig. 2**. For the data presented in **Fig. 2**, for all kinases, only peptides containing a single Tyr residue were analyzed. Phosphosite sequences with two tyrosines had the non-central Tyr mutated to alanine during library construction. Phospho-pLogo diagrams were generated as follows. The central 11 residues of the phosphosite sequences (five residues before and after Tyr) with read frequency ratios in the top 25% of ratios within a given sample across at least two replicate samples (top 25% in replicate one AND replicate two) were counted as highly phosphorylated sequences. These highly phosphorylated sequences were used as the foreground set in the calculation of a pLogo, as described in (3). The background set for this calculation was the set of single-Tyr sequences observed in the unsorted sample. This analysis was repeated using the read frequency ratio data for chicken c-Src and human c-Abl published in (2).

Construction of phosphosite single-site saturation mutagenesis libraries. Single-site saturating mutagenesis libraries were constructed as described in (4). Briefly, oligonucleotides containing

one degenerate NNS codon each (Integrated DNA Technologies) were used in an overlap-extension PCR reaction to generate constructs with a 21-residue peptide-coding sequence in-frame with the eCpx bacterial surface-display scaffold (5), with one mutated amino acid sequence position per DNA fragment. These fragments were pooled, restriction digested, and ligated into the SfiI sites of the pBAD33 vector. The pooled, ligated mixture was transformed into TOP10 *E. coli* (ThermoFisher Scientific) by electroporation, and library DNA was isolated with a silica membrane spin column (Zymo Research).

Phosphorylation analysis of phosphosite single-site saturation mutagenesis libraries. Libraries were phosphorylated as described for the Human-pTyr library, at room temperature, with kinase concentrations and durations that produced ~30% of maximal library phosphorylation as judged by flow cyotmetry. Cells were labeled with anti-phosphotyrosine 4G10 antibody-PE conjugate, and the top 15% of cells in the PE channel were sorted on a BD FACSAria Fusion cell sorter. Sorted and input cell samples were sequenced as described for the Human-pTyr library, above. Relative enrichment for each sequence position *i* mutated to each amino acid substitution *x* versus the wild-type (WT) variant, ΔE_x^i, was calculated from the read frequencies in the sorted and unsorted samples with the following formula:

$$\Delta E_x^i = \log_2\left(\frac{v_{\mathrm{sort}}^{i,x}}{v_{\mathrm{input}}^{i,x}}\right) - \log_2\left(\frac{v_{\mathrm{sort}}^{i,\mathrm{WT}}}{v_{\mathrm{input}}^{i,\mathrm{WT}}}\right)$$

where $v_{\mathrm{sample}}^{i,x}$ is a read frequency of variant *x* at position *i* for a particular sample. Data in **Figs. 3 and 5** are the mean from at least two biological replicates.

Phosphorylation analysis of the EGFR substrate phosphosite library, with expression level correction. A collection of 21-residue single-tyrosine peptides corresponding to human EGFR-family C-terminal Tyr residues and EGFR substrates reported in the PhosphositePlus Database (6) was assembled by overlap-extension PCR and inserted into the eCpx scaffold. The sequences are listed in **Table S2**. This library was expressed, phosphorylated by the dimerized EGFR intracellular module or c-Src kinase domain, sorted, and sequenced as described for the Human-pTyr library. Enrichment ratios for a peptide *p* relative to a peptide containing no Tyr residues, ΔE_p, were determined for each peptide with the following formula:

$$\Delta E_p = \log_2\left(\frac{v_{\text{sort}}^{p}}{v_{\text{input}}^{p}}\right) - \log_2\left(\frac{v_{\text{sort}}^{\text{Y}\rightarrow\text{A}}}{v_{\text{input}}^{\text{Y}\rightarrow\text{A}}}\right)$$

where $v_{\text{sample}}^{\text{Y}\rightarrow\text{A}}$ denotes the read frequency of a peptide corresponding the EGFR Tyr 1173 phosphosite, with the tyrosine mutated to alanine, in a particular sample.

Relative expression levels were also measured for each peptide in order to correct for expression level differences that could show up as phosphorylation level differences in the bacterial surface-display/deep sequencing assay. Expression levels were monitored in a separate experiment using a Strep-tag at the C-terminus of the eCpx scaffold, detected with the StrepMAB-Classic chromeo 488 antibody conjugate (IBA Lifesciences). Cells were sorted into 6 bins based on fluorescence in the FITC channel with a BD FACSAria Fusion cell sorter, and read frequencies were calculated for each peptide *p* in each bin. These read frequencies were weighted by the number of flow cytometry events *e* in each bin versus the total number of events, as in

$$w_{\text{bin}}^{p} = v_{\text{bin}}^{p} \times \frac{e_{\text{bin}}}{e_{\text{total}}}$$

The resulting weighted frequencies for each peptide in each bin, w_{bin}^{p}, versus the log-transformed mean fluorescence measured for all flow cytometry events in each bin were fitted to a simple Gaussian distribution, giving an estimated weighted mean fluorescence $\langle w^{p} \rangle$ for each peptide. This value was compared to that for the no-tyrosine control peptide and the measured mean fluorescence of the lowest bin, w_0^{p} to generate a relative expression level C_p, according to the following formula:

$$C_p = \log_2 \left(\frac{\langle w^{p} \rangle - w_0^{p}}{\langle w^{\text{WT}} \rangle - w_0^{\text{WT}}} \right)$$

Finally, the expression-corrected relative enrichment ratio for each peptide, ΔE_p^*, was calculated as:

$$\Delta E_p^* = \Delta E_p - C_p$$

ΔE^* values for each peptide in the library, calculated for EGFR and c-Src, are plotted in **Fig. 4C.** Relative expression level for each mutant versus the wild-type peptide were measured for the Tyr 1114 phosphosite mutagenesis library by the same method, with the wild-type peptide serving as the reference, and plotted in **Fig. S3.**

Binding analysis of phosphosite single-site saturation mutagenesis libraries. Libraries were phosphorylated as described for the Human-pTyr library, but with a mixture of 2.5 μM c-Abl kinase domain, 2.5 μM c-Src kinase domain, and 1 μM dimerized EGFR intracellular module (including 2 μM rapamycin). These preparative phosphorylation reactions were incubated for 1 hour at room temperature. They also included the addition of 50 μg/ml rabbit muscle creatine phosphokinase and 5 mM creatine phosphate (Sigma Aldrich) to regenerate ATP. To confirm complete library phosphorylation, a sample of cells treated in this manner was monitored by flow cytometry based on anti-phosphotyrosine-PE labeling. The phosphorylated cells were harvested by gentle centrifugation and washed once with binding buffer containing 50 mM HEPES pH 7.5, 150 mM NaCl, 1 mM TCEP, and 0.2% BSA. Cells were then resuspended in this buffer containing either 5 μM tandem Grb2 SH2–GFP or 1 μM tandem Shc1 PTB–GFP and incubated for 1 hour at room temperature. After labeling, the cells were centrifuged, and the label-containing supernatant was discarded. The cells were washed in this manner one time with labeling buffer and finally resuspended for fluorescence-activated cell sorting. Cells were sorted by GFP fluorescence on a BD FACSAria Fusion, with the top 15% of cells in the FITC channel collected. DNA from the sorted and input cell samples was amplified, indexed, and sequenced as described for the phosphorylation experiments, above. Position-wise relative enrichment for each mutant versus wild-type, ΔE_i^x, was calculated in the same manner as for mutagenesis matrices with respect to phosphorylation.

Kinase-peptide Activity Assays

Kinase activity was measured with an enzyme-coupled assay based on NADH absorbance (7). Kinases were assayed at 0.5 μM enzyme concentration in a buffer containing 50 mM HEPES pH

7.5, 50 mM NaCl, 10 mM $MgCl_2$, 1 mM TCEP, 1 mM phosphoenol pyruvate, 0.5 mg/ml NADH**,** ~40 U/ml pyruvate kinase, and ~60 U/ml lactate dehydrogenase (from rabbit muscle, Sigma Aldrich). Reactions contained either 0.5 mM or no peptide. Reactions containing EGFR also contained 1 μM rapamycin (LC Laboratories). Reactions were initiated with the addition of ATP to a final concentration of 0.5 mM. Reaction progress was monitored by the change in absorbance at 340 nm over time, at 25º C. Kinase activity was calculated as the difference in rates between reactions with and without substrate peptide, using the extinction coefficient of NADH at 340 nm of 6220 M^{-1} cm^{-1}.

Bioinformatics

Generation of sequence pLogo of EGFR-family phosphosites. EGFR-family C-terminal tail phosphosite sequences were collected from the EggNOG database (orthology group ENOG410XNSR) (8) and filtered to include only the sequences in the Eumetazoa and Porifera taxonomic groups (NCBI taxids 6072 and 6040). Full-length protein sequences were further filtered to exclude sequences with greater than 90% sequence identity within the kinase domain, in order to obtain a set of sequences that evenly samples natural evolutionary sequence space, without oversampling taxa that have relatively high numbers of species in the sequence database. Kinase domain boundaries were identified by querying each EGFR sequence against the SMART database (9). These domain sequences were aligned with `mafft` (G-INS-I algorithm) (10). and then filtered by sequence identity with the CD-HIT web service (11). From the resulting set of full-length sequences, phosphosite sequences were extracted as the five residues before and after each Tyr that occurs after each kinase domain, as identified above. This list of 11-residue sequences was used as the foreground set in the pLogo web service (3) to generate the sequence pLogo shown

in **Fig. 2D**. The background set was a random sampling (to keep the total number of sequences below the web server's limit) of the UniRef50 database of representative sequences (12), filtered to include verified metazoan proteins, tagged as having either transmembrane or intracellular localization.

Sequence alignment of EGFR-family phosphosites. Full-length sequences in the EGFR branch of the metazoan EGFR-family (8) were roughly aligned with `mafft` (10), and the Tyr 1086 site was identified visually in each sequence with `jalview` (13). These sites were then realigned with `mafft` and ordered based on the topology of the EGFR-family tree in the EggNOG database (8).

Molecular Dynamics Simulations

Simulations were prepared with `VMD` (14) and generated with `NAMD` (15) using the CHARMM36 force field (16). For all stages, electrostatics were calculated with the particle mesh Ewald summation, and a non-bonded cutoff of 12 Å was employed. A starting model was generated for the peptide corresponding to human EGFR residues 1110–1118 by assigning backbone and Cβ atoms of residues 1114–1118 to the coordinates of chain B of the crystal structure of the insulin receptor kinase domain bound to a substrate peptide (PDB 1IR3) (17). Residues 1110–1113 were modeled in an extended conformation. The peptide model's N- and C-termini were capped with N-acetyl and N-methylcarboxamide groups, respectively. This model was solvated in a rectangular box with TIP3P water, and with sodium and chloride ions for an effective ionic strength of 150 mM.

The model was minimized and equilibrated as follows. 1000 steps of conjugate gradient minimization were performed with the protein residues fixed and solvent molecules allowed to move. This was followed by 1000 steps of conjugate gradient minimization in which the peptide atoms were allowed to move, but with an additional sinusoidal potential with a spring constant of 10 kcal mol^{-1} $radian^{-2}$ applied to the backbone φ angles of residues 1115–1118 and ψ angles of residues 1114–1117, to keep these residues in their starting β-strand conformation. After minimization, the system was assigned random velocities and equilibrated under constant number, temperature, and pressure (NPT) conditions at 300 K for 50000 steps, at 2 fs $step^{-1}$ (for 1 ns total), with all protein atoms restrained by a harmonic potential, with a spring constant of 1 kcal mol^{-1} $Å^{-2}$. This was proceeded by production simulation under the same conditions, with all atoms unrestrained except for the dihedral restraints for residues 1114–1118 described above. (The first 1 ns was discarded from analyses to allow for additional equilibration.)

Simulations were analyzed with the `cpptraj` package of the Amber software suite (18) and `pymol` (19). Simulation frames were aligned to the EGFR kinase domain as follows. First, the co-crystal structure of the EGFR kinase domain and a bi-substrate analog (PDB 2GS6) (20) was aligned based on kinase domain backbone atoms to the insulin receptor kinase domain bound to substrate (PDB 1IR3) (17). Then, simulation frames were aligned to the substrate peptide of the insulin receptor crystal structure based on the backbone atoms of residues 1114–1117 of the simulated EGFR peptide, producing models of the EGFR kinase domain bound to a peptide substrate. Representative snapshots for **Fig. 3** were chosen by visual inspection of sets of simulation frames obtained by clustering based on the backbone dihedral angles of residues 1112 and 1113. Frames, sampled every 10 ps, were assigned to one of five clusters based on the

backbone dihedral angles of the –1 and –2 residues of each frame, by agglomerative clustering in a custom python script.

Supplementary Figures

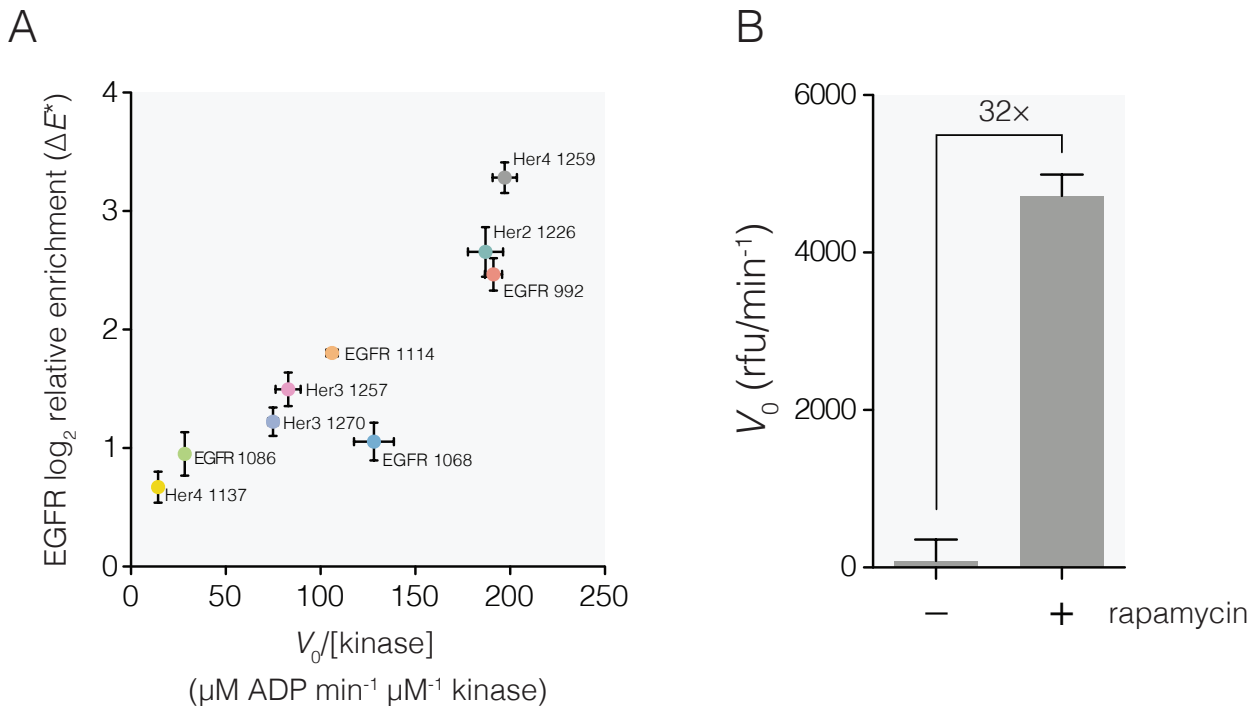

Fig. S1. Kinase activity of a dimerized EGFR kinase measured with enzyme-coupled assays and bacterial surface-display coupled with deep sequencing. **A.** The two measurements of enzymatic activity of EGFR for 21-residue peptides corresponding to the indicated EGFR family tail phosphosites were compared. The EGFR protein used in both methods consisted of an equimolar mixture of N-terminal FKBP and FRB fusions of human EGFR residues 663–1186, in the presence of excess rapamycin. The activity on the *x*-axis was measured with a continuous, homogeneous assay wherein the generation of ADP upon phosphorylation of a purified peptide is enzymatically coupled to the oxidation of reduced *β*-nicotinamide adenine dinucleotide (NADH), with a corresponding decrease in absorbance of NADH. Peptides were present at 0.5 mM, below expected K_M, and EGFR dimers were present at 0.2 μM. Activity reported on the y-axis was measured with

the bacterial surface-display and deep sequencing assay. For this experiment, peptides were displayed on the surface of E. coli as part of a larger library were subjected to phosphorylation by dimerized EGFR at 0.1 μM dimer for 15 minutes at room temperature, to produce a phosphorylation level of ~1/3 the maximum obtained by long incubation in the presence of high concentration of kinases. The highly phosphorylated cells were collected by fluorescence activated cell sorting, and the peptide coding portion of the surface-display gene of these cells was sequenced, along with that of the input population. The read frequencies were normalized and plotted as a log-fold-change relative to a negative control peptide containing no Tyr residue and corrected for the separately measured surface-display level. Error bars indicate standard error of the mean from three replicates in each experimental method. **B**. Effect of forced dimerization on EGFR intracellular module kinase activity. The generic Tyr kinase substrate poly($Glu_4Tyr)_n$ at 1 mg/ml was subjected to phosphorylation by 50 nM FKBP– and FRB–EGFR dimers in the presence and absence of 1 μM rapamycin. Phosphorylation at various time points was detected with by ADP production enzymatically coupled to production of resorufin. The slope of fluorescence change over time for the linear reaction progress curve is plotted with standard error of the mean, and the fold-increase in rate with the addition of rapamycin is noted.

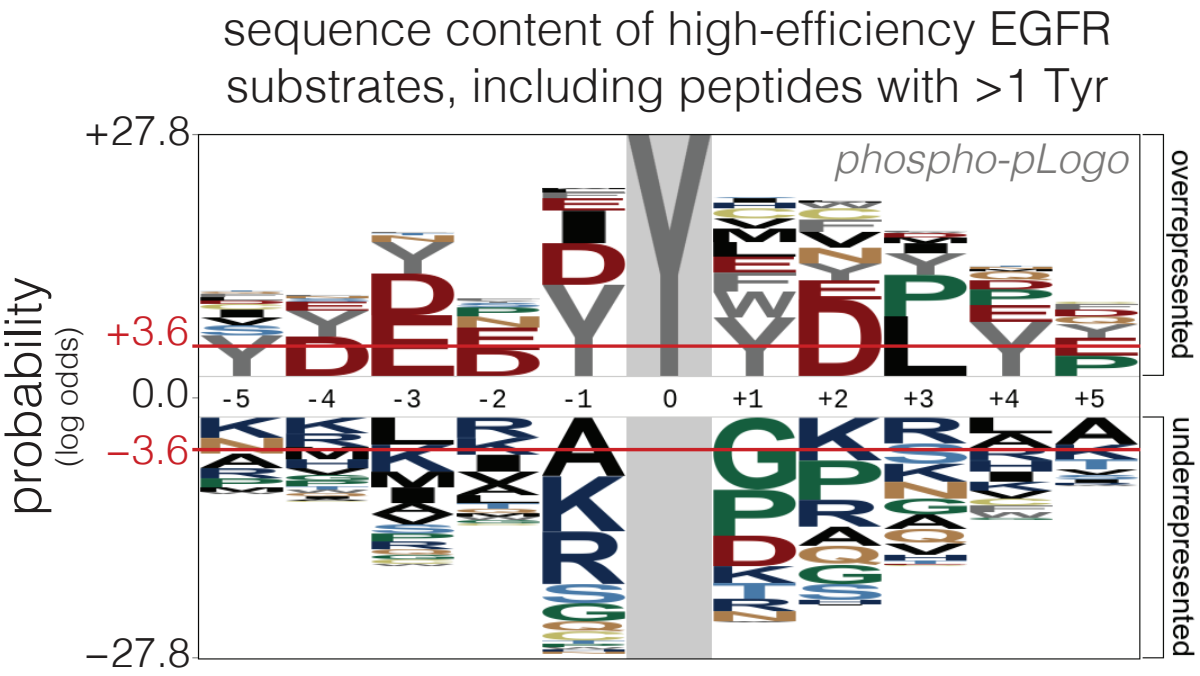

Fig. S2. Sequence content of high-efficiency peptide substrates of EGFR from the human proteome, including peptides with more than one Tyr residue. A phospho-pLogo of peptide sequences in the top quartile of read frequency ratios for EGFR, according to the bacterial surface-display/deep sequencing experiment with the Human-pTyr library. This pLogo was generated from the same raw dataset as **Fig. 2C**, but including peptides with greater than one Tyr residue in the analysis. Tyr residues appear at multiple positions, but it is not known whether this is a result of multiple Tyr residues becoming phosphorylated and detected by the antibody during the experiment, or due to an improvement of catalytic efficiency for the central Tyr residue when other Tyr residues are present in the peptide.

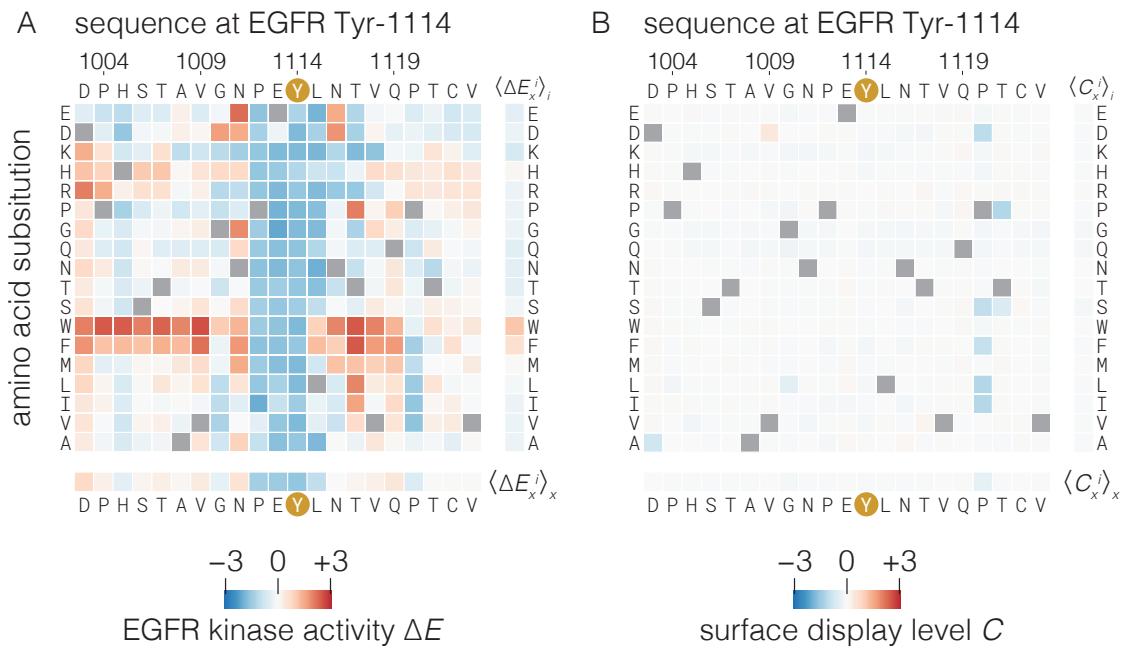

Fig. S3. Comparison of relative enrichment differences between variants due to phosphorylation and surface-display level for the Tyr 1114 phosphosite peptide library. The contribution of expression level differences between variants in the calculated enrichment due to phosphorylation by EGFR, ΔE, was estimated by measuring the relative surface-display level of each variant in the Tyr 1114 library. Cells displaying the Tyr 1114 library were labeled with a fluorescent anti-Strep tag antibody targeting the surface-display scaffold. These cells were sorted by fluorescence activated cell sorting into six bins spanning the distribution of fluorescence values. The abundance of each peptide in each bin relative to the wild-type peptide was inferred from read frequencies, as measured by Illumina sequencing in the same manner used for phosphorylation level determination. The log-fold differences in expression level relative to the wild-type peptide in the library, C_x^i (**panel B**), were plotted on the same scale as the log-fold differences in phosphorylation level for the Tyr 1114 library (**panel A**, reproduced from **Fig. 3B**). White squares indicate minimal

differences in expression level for a variant relative to the wild-type peptide, and thus indicate minimal contribution to the phosphorylation enrichment value for that variant, ΔE_x^i.

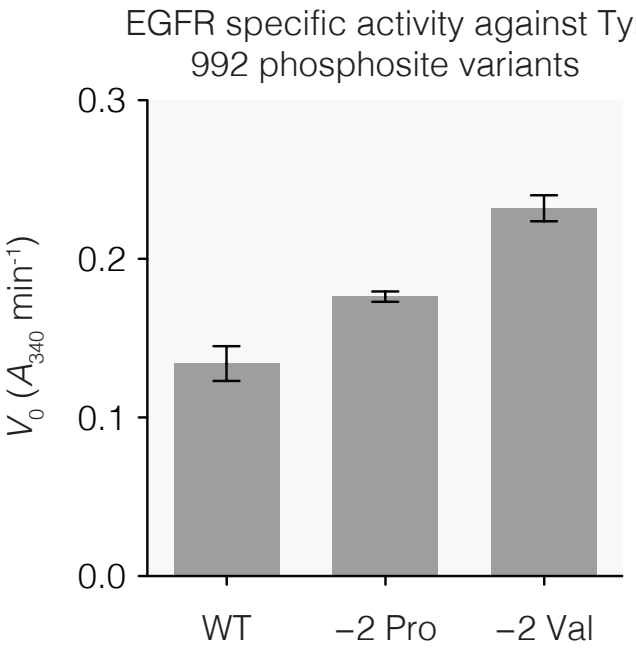

Fig. S4. Kinase activity of EGFR against EGFR Tyr 992 phosphosite peptide mutants. Relative kinase activity was measured with an NADH-coupled enzyme assay with purified EGFR intracellular module and purified 21-mer peptides corresponding to the EGFR Tyr 992 phosphosite, with and without the noted substitutions to the wild-type sequence. Steady-state rates were measured in triplicate and plotted as the negative slope of the linear portion of enzyme progress curve of absorbance at 340 nm over time. Error bars, 95% confidence interval.

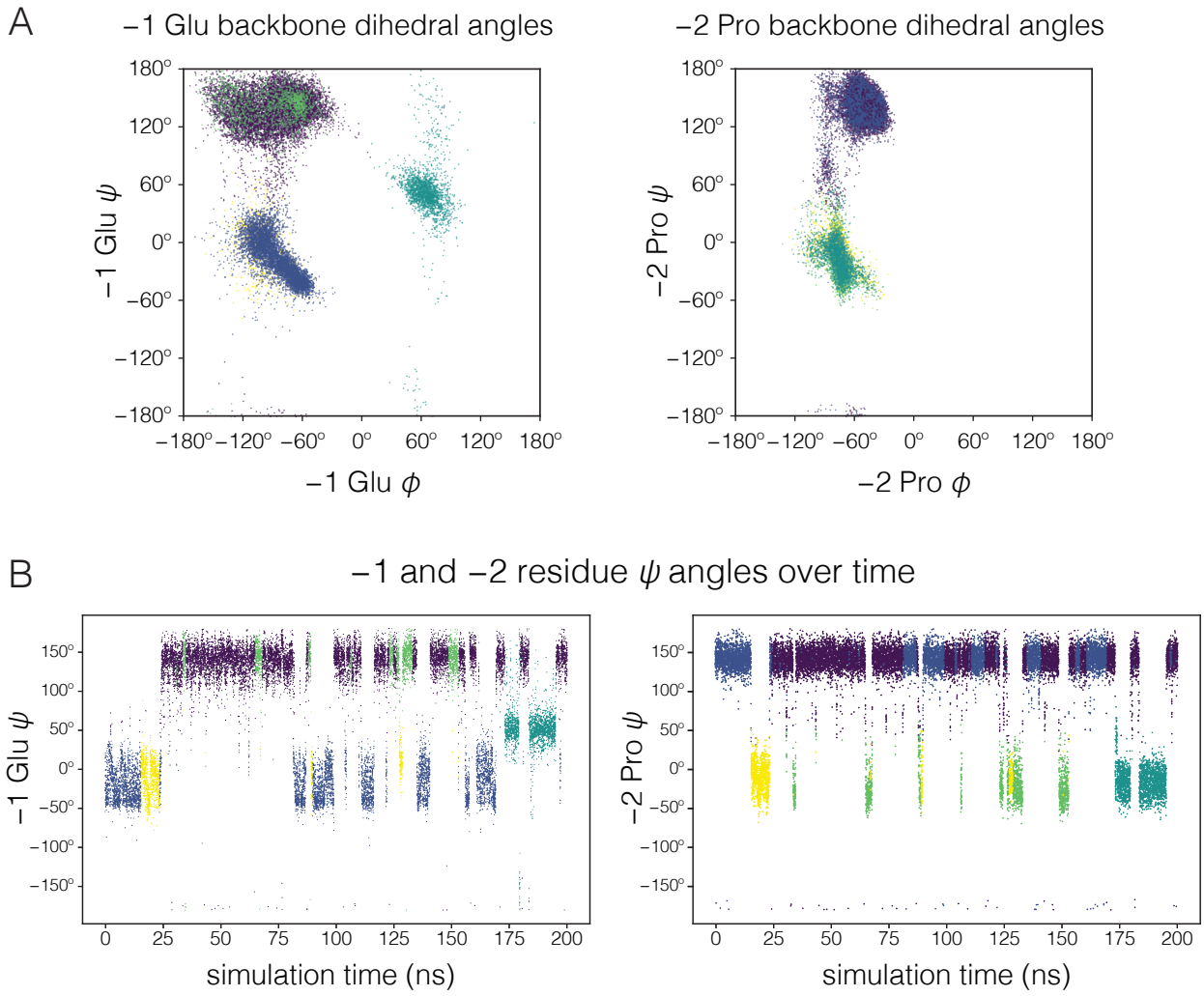
A
−1 Glu backbone dihedral angles
−2 Pro backbone dihedral angles
−1 Glu ψ
−1 Glu ϕ
−2 Pro ψ
−2 Pro ϕ
180°
120°
60°
0°
−60°
−120°
−180°
B
−1 and −2 residue ψ angles over time
−1 Glu ψ
−2 Pro ψ
150°
100°
50°
0°
−50°
−100°
−150°
0
25
50
75
100
125
150
175
200
simulation time (ns)
C
−1 and −2 residue conformation
cluster membership
fractional occupancy
0.5
0.4
0.3
0.2
0.1
0.0
−1 Glu β α Lα β α
−2 Pro β β α α α

Fig. S5. Backbone conformation of the −1 and −2 residues of the EGFR Tyr 1114 phosphosite peptide during molecular dynamics simulations. A. Ramachandran diagrams showing the backbone dihedral angles of the −1 Glu (left panel) and −2 Pro (right panel) residues during a representative molecular dynamics trajectory. Each point represents a frame from the 200 ns trajectory, sampled every 10 ps. The points are colored based on agglomerative clustering performed on four dihedral angles, the φ and ψ angles of the −1 and −2 residues. **B.** ψ angles of the −1 and −2 residues over the time course of the simulation, sampled every 10 ps and colored as in A. ψ angles between approximately 110º and 180º are considered to represent the β conformation, and angles between approximately −50º and 50º are considered to represent the α conformation. **C.** Fractional occupancy of the five clusters of −1 and −2 dihedral angles generated for trajectory frames sampled every 10 ps. "Lα" indicates the left-handed α-helical region of the Ramachandran diagram.

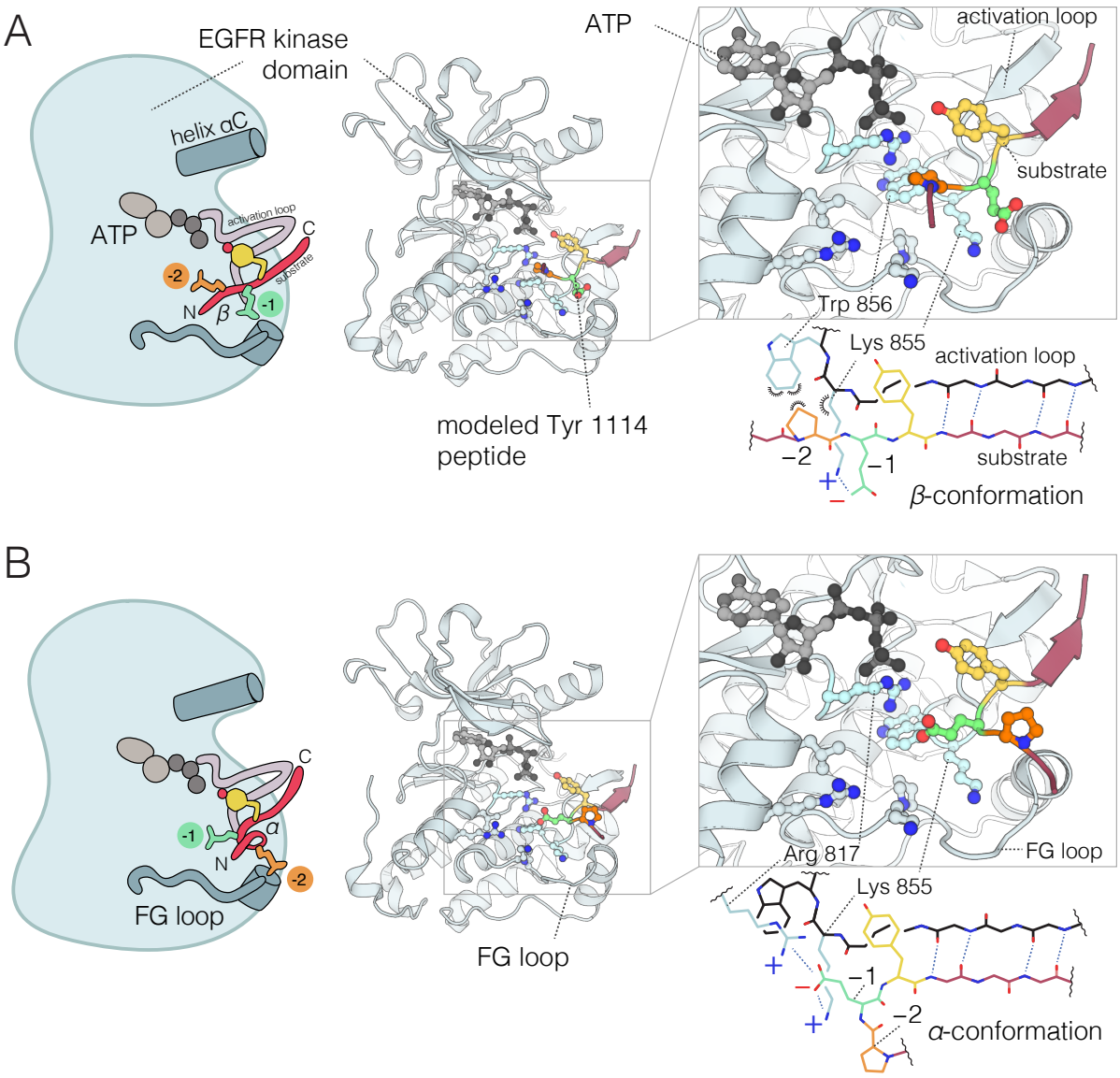

Fig. S6. Structural explanation for alternative sequence preferences of EGFR at the −1 and −2 positions. Selected snapshots from molecular dynamics simulations of an EGFR Tyr 1114 peptide docked onto a peptide-bound crystal structure of the EGFR kinase domain (PDB 2GS6) (20). Two snapshots are shown, with the −1 and −2 residues of the peptide in either the *β* conformation (**A**) or *α* conformation (**B**). Interactions between the −1 and −2 peptide residues and selected residues on the kinase domain are highlighted. Diagrams illustrating the different

interactions available between kinase domain residues and a substrate peptide depending on the orientation of the −1 and −2 residues are shown below each zoomed-in view of the active site. A peptide with a −2 Pro and a β conformation of the −1 residue, is diagramed in A, while a peptide with a −2 glutamic acid and an α conformation of the −1 residue is diagramed in **B**.

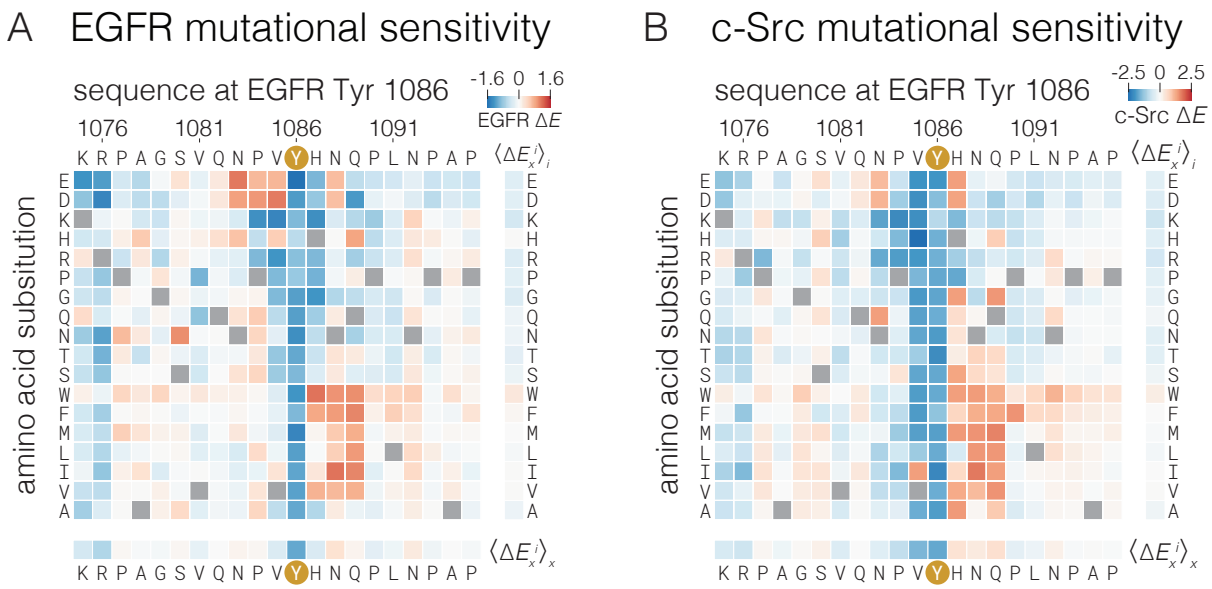

Fig. S7. Comparison of relative enrichments for EGFR (panel **A**) and c-Src (panel **B**) against the Tyr 1086 phosphosite peptide library. These data are reproduced from main text **Figs. 3B and 6D**, respectively.

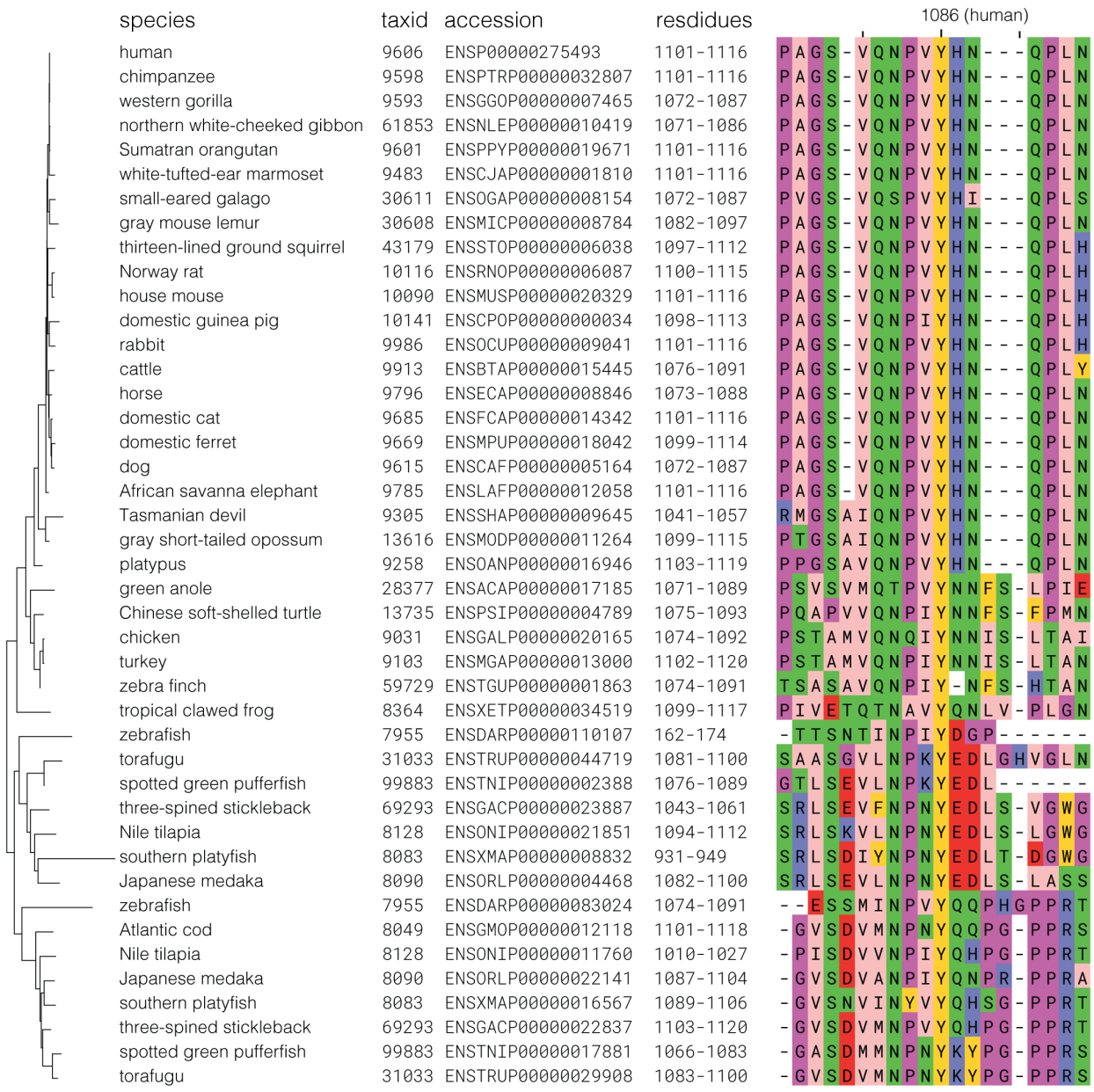

Fig. S8. Alignment of EGFR Tyr 1086 phosphosite sequences in the clade after the split between EGFR and Her2. The sequences were aligned with the mafft global homology algorithm and are labeled with common name, NCBI taxid, ENSEMBL translation accession number, and residue boundaries (including signal sequences). The phylogenetic relationship for the corresponding full-length EGFR sequences, taken from the EggNOG database, is shown on the left. A His residue at

the +1 position of the Tyr 1086 phosphosite is a conserved feature of mammalian EGFR sequences. No phosphosites contain a −1 acidic or +1 hydrophobic residue.

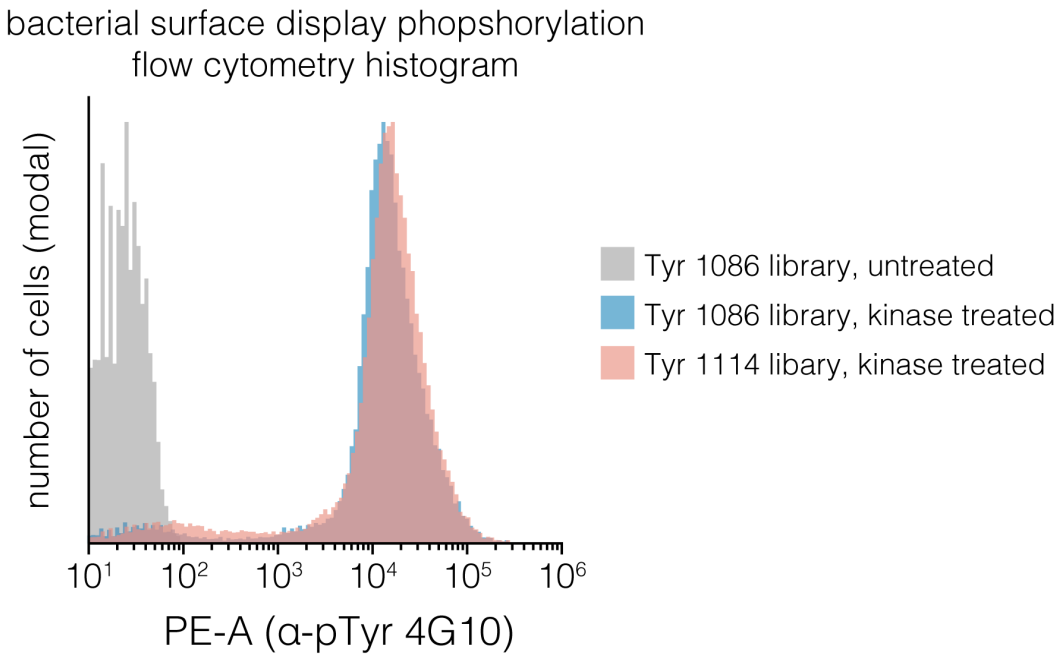

Fig. S9. Flow cytometry histogram of fully-phosphorylated bacteria displaying mutagenesis libraries. A sample of the bacteria that served as an input to the surface-display binding experiments presented in **Fig. 5** were stained with an anti-phosphotyrosine antibody (4G10) and analyzed by flow cytometry. Site-saturation mutagenesis libraries corresponding to the EGFR Tyr 1086 and 1114 phosphosites were either treated with a mixture of EGFR, c-Src, and c-Abl kinases for 1 hour at room temperature ("kinase treated") or incubated in the absence of kinases ("untreated"). The single, narrow main peaks in the histogram indicate the libraries were uniformly phosphorylated.

Table S1: Sequences of purified proteins

construct name	uniprot identifier[1]	residue numbers[2]	protein sequence[3]	Figures[4]
His_{10}–FKBP–EGFR 663–1186	P00533	663–1186	MHHHHHHHHHHASGGGVQVETISPGDGRTFPKR GQTCVVHYTGMLEDGKKFDSSRDRNKPFKFMLG KQEVIRGWEEGVAQMSVGQRAKLTISPDYAYGA TGHPGIIPPHATLVFDVELLKLEGSGGSGGSGS ELVEPLTPSGEAPNQALLRILKETEFKKIKVLG SGAFGTVYKGLWIPEGEKVKIPVAIKELREATS PKANKEILDEAYVMASVDNPHVCRLLGICLTST VQLITQLMPFGCLLDYVREHKDNIGSQYLLNWC VQIAKGMNYLEDRRLVHRDLAARNVLVKTPQHV KITDFGLAKLLGAEEKEYHAEGGKVPIKWMALE SILHRIYTHQSDVWSYGVTVWELMTFGSKPYDG IPASEISSILEKGERLPQPPICTIDVYMIMVKC WMIDADSRPKFRELIIEFSKMARDPQRYLVIQG DERMHLPSPTDSNFYRALMDEEDMDDVVDADEY LIPQQGFFSSPSTSRTPLLSSLSATSNNSTVAC IDRNGLQSCPIKEDSFLQRYSSDPTGALTEDSI DDTFLPVPEYINQSVPKRPAGSVQNPVYHNQPL NPAPSRDPHYQDPHSTAVGNPEYLNTVQPTCVN STFDSPAHWAQKGSHQISLDNPDYQQDFFPKEA KPNGIFKGSTAENAEYLRVAPQSSEFIGACCPG CC	Figs. 2, 3, 4, 5, S1, S2, S3, S4, S7, S9
His_{10}–FRB–EGFR 663–1186	P00533	663–1186	MHHHHHHHHHHASGGRVAILWHEMWHEGLEEAS RLYFGERNVKGMFEVLEPLHAMMERGPQTLKET SFNQAYGRDLMEAQEWCRKYMKSGNVKDLTQAW DLYYHVFRRISGSGGSGGSGSELVEPLTPSGEA PNQALLRILKETEFKKIKVLGSGAFGTVYKGLW IPEGEKVKIPVAIKELREATSPKANKEILDEAY VMASVDNPHVCRLLGICLTSTVQLITQLMPFGC LLDYVREHKDNIGSQYLLNWCVQIAKGMNYLED RRLVHRDLAARNVLVKTPQHVKITDFGLAKLLG AEEKEYHAEGGKVPIKWMALESILHRIYTHQSD VWSYGVTVWELMTFGSKPYDGIPASEISSILEK GERLPQPPICTIDVYMIMVKCWMIDADSRPKFR ELIIEFSKMARDPQRYLVIQGDERMHLPSPTDS NFYRALMDEEDMDDVVDADEYLIPQQGFFSSPS TSRTPLLSSLSATSNNSTVACIDRNGLQSCPIK EDSFLQRYSSDPTGALTEDSIDDTFLPVPEYIN QSVPKRPAGSVQNPVYHNQPLNPAPSRDPHYQD PHSTAVGNPEYLNTVQPTCVNSTFDSPAHWAQK GSHQISLDNPDYQQDFFPKEAKPNGIFKGSTAE NAEYLRVAPQSSEFIGACCPGCC	Figs. 2, 3, 4, 5, S1, S2, S3, S4, S7, S9
c-Src kinase domain + tail	P00523	251–533	HMQTQGLAKDAWEIPRESLRLEVKLGQGCFGEV WMGTWNGTTRVAIKTLKPGTMSPEAFLQEAQVM KKLRHEKLVQLYAVVSEEPIYIVTEYMSKGSLL DFLKGEMGKYLRLPQLVDMAAQIASGMAYVERM NYVHRDLRAANILVGENLVCKVADFGLARLIED NEYTARQGAKFPIKWTAPEAALYGRFTIKSDVW SFGILLTELTTKGRVPYPGMVNREVLDQVERGY RMPCPPECPESLHDLMCQCWRKDPEERPTFEYL QAFLEDYFTSTEPQYQPGENL	Figs. 4, 5, S7, S9
c-Abl kinase domain	P00519	229–512	HMSPNYDKWEMERTDITMKHKLGGGQYGEVYEG VWKKYSLTVAVKTLKEDTMEVEEFLKEAAVMKE IKHPNLVQLLGVCTREPPFYIITEFMTYGNLLD YLRECNRQEVNAVVLLYMATQISSAMEYLEKKN FIHRDLAARNCLVGENHLVKVADFGLSRLMTGD TYTAHAGAKFPIKWTAPESLAYNKFSIKSDVWA FGVLLWEIATYGMSPYPGIDLSQVYELLEKDYR MERPEGCPEKVYELMRACWQWNPSDRPSFAEIH QAFETMFQESSISDEVEKELGK	Figs. 5, S9

tandem Shc1 PTB–GFP	P29353	147–318	HMGWLHPNDKVMGPGVSYLVRYMGCVEVLQSMR ALDFNTRTQVTREAISLVCEAVPGAKGATRRRK PCSRPLSSILGRSNLKFAGMPITLTVSTSSLNL MAADCKQIIANHHMQSISFASGGDPDTAEYVAY VAKDPVNQRACHILECPEGLAQDVISTIGQAFE LRFKQYLRNGSGGSGGSGSGGSGGSGSGWLH PNDKVMGPGVSYLVRYMGCVEVLQSMRALDFNT RTQVTREAISLVCEAVPGAKGATRRRKPCSRPL SSILGRSNLKFAGMPITLTVSTSSLNLMAADCK QIIANHHMQSISFASGGDPDTAEYVAYVAKDPV NQRACHILECPEGLAQDVISTIGQAFELRFKQY LRNGSAGSAAGSGEFMVSKGEELFTGVVPILVE LDGDVNGHKFSVSGEGEGDATYGKLTLKFICTT GKLPVPWPTLVTTLTYGVQCFSRYPDHMKQHDF FKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEG DTLVNRIELKGIDFKEDGNILGHKLEYNYNSHN VYIMADKQKNGIKVNFKIRHNIEDGSVQLADHY QQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKR DHMVLLEFVTAAGITLGMDELYK	Fig. 5
tandem Grb2 SH2–GFP	P62993	55–152	HMKPHPWFFGKIPRAKAEEMLSKQRHDGAFLIR ESESAPGDFSLSVKFGNDVQHFKVLRDGAGKYF LWVVKFNSLNELVDYHRSTSVSRNQQIFLRDIE GSGGSGGSGSGSGGSGGSGSMKPHPWFFGKIPR AKAEEMLSKQRHDGAFLIRESESAPGDFSLSVK FGNDVQHFKVLRDGAGKYFLWVVKFNSLNELVD YHRSTSVSRNQQIFLRDIEGSAGSAAGSGEFMV SKGEELFTGVVPILVELDGDVNGHKFSVSGEGE GDATYGKLTLKFICTTGKLPVPWPTLVTTLTYG VQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFF KDDGNYKTRAEVKFEGDTLVNRIELKGIDFKED GNILGHKLEYNYNSHNVYIMADKQKNGIKVNFK IRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNH YLSTQSALSKDPNEKRDHMVLLEFVTAAGITLG MDELYK	Fig. 5

[1]Uniprot identifier for the main protein of interest

[2]Residue numbers of the main protein of interest included in the construct

[3]Protein sequence as assayed in the paper (i.e., purification tags removed where applicable)

[4]Figures in the paper in which this construct was used

Table S2: sequences of peptides in the EGFR substrate phosphosite library used in Figs. 5A and 5B

Peptide name	Peptide Sequence
Her1_845	LAKLLGAEEKEYHAEGGKVPI
Her1_974	MHLPSPTDSNFYRALMDEEDM
Her1_992	EDMDDVVDADEYLIPQQGFFS
Her1_1045	CPIKEDSFLQRYSSDPTGALT
Her1_1068	SIDDTFLPVPEYINQSVPKRP
Her1_1086	KRPAGSVQNPVYHNQPLNPAP
Her1_1101	PLNPAPSRDPHYQDPHSTAVG
Her1_1114	DPHSTAVGNPEYLNTVQPTCV
Her1_1148	GSHQISLDNPDYQQDFFPKEA
Her1_1173	IFKGSTAENAEYLRVAPQSSE
Her2_983	LGPASPLDSTFYRSLLEDDDM
Her2_1001	DDMGDLVDAEEYLVPQQGFFC
Her2_1090	LPTHDPSPLQRYSEDPTVPLP
Her2_1105	PTVPLPSETDGYVAPLTCSPQ
Her2_1117	VAPLTCSPQPEYVNQPDVRPQ
Her2_1174	FAFGGAVENPEYLTPQGGAAP
Her2_1199_1200F	PPAFSPAFDNLYFWDQDPPER
Her2_1226	FKGTPTAENPEYLGLDVPV
Her3_1035	GSQSLLSPSSGYMPMNQGNLG
Her3_1140	PPGLEEEDVNGYVMPDTHLKG
Her3_1178_1180F	LGTEEEDEDEEYEFMNRRRRH
Her3_1203_1205F	PPRPSSLEELGYEFMDVGSDL
Her3_1241_1243F	MPTAGTTPDEDYEFMNRQRDG
Her3_1257	RQRDGGGPGGDYAAMGACPAS
Her3_1270	AMGACPASEQGYEEMRAFQGP
Her3_1309	EATDSAFDNPDYWHSRLFPKA

Her4_997	EDLEDMMDAEEYLVPQAFNIP
Her4_1031	VPQAFNIPPPIYTSRARIDSN
Her4_1103	PHVQEDSSTQRYSADPTVFAP
Her4_1125	RSPRGELDEEGYMTPMRDKPK
Her4_1137	MTPMRDKPKQEYLNPVEENPF
Her4_1163	NGDLQALDNPEYHNASNGPPK
Her4_1177_Y1183F	ASNGPPKAEDEYVNEPLFLNT
Her4_1183_Y1177F	KAEDEFVNEPLYLNTFANTLG
Her4_1196	NTFANTLGKAEYLKNNILSMP
Her4_1217	EKAKKAFDNPDYWNHSLPPRS
Her4_1233_3F	LPPRSTLQHPDYLQEFSTKFF
Her4_1259	RIRPIVAENPEYLSEFSLKPG
Her1_1173F	IFKGSTAENAEFLRVAPQSSE
Abl1_393	FGLSRLMTGDTYTAHAGAKFP
CRK_221	PQPLGGPEPGPYAQPSVNTPL
CrkL_207_198F	NSFGIPEPAHAYAQPQTTTPL
DOCK7_1257	IGIIMETVPQLYDFTETHNQR
Ezrin_146_137F	GDFNKEVHKSGYLSSERLIPQ
Ezrin_354	EKEELMLRLQDYEEKTKKAER
GAB1_285	SPSSTEADGELYVFNTPSGTS
GAB1_373	IPRTASDTDSSYCIPTAGMSP
GAB1_406	KLRKDASSQDCYDIPRAFPSD
GAB1_447	GSVSSEELDENYVPMNPNSPP
GAB1_472	SSFTEPIQEANYVPMTPGTFD
GAB1_589	SSSDSHDSEENYVPMNPNLSS
GAB1_627	MIKPKGDKQVEYLDLDLDSGK
GAB1_659	SGSSVADERVDYVVVDQQKTL
PI3K_580_P85a	DLIQLRKTRDQYLMWLTQKGV
PI3K_607_P85a	EWLGNENTEDQYSLVEDDEDL
PLCG1_771_775F	ALEKIGTAEPDYGALFEGRNP
PLXNB1_1708	TFVRDSVGEPLYMLFRGIKHQ
PLXNB1_1732	GPVDSVTGKAKYTLNDNRLLR

PTP1B_66	SRIKLHQEDNDYINASLIKME
SHC1_427	CPGRELFDDPSYVNVQNLDKA
SHC3_341_342F	TEEEGDGSDHPYFNSIPSKMP

References for SI reference citations

1. Seeliger MA, et al. (2005) High yield bacterial expression of active c-Abl and c-Src tyrosine kinases. *Protein Sci* 14(12):3135–3139.

2. Shah NH, Löbel M, Weiss A, Kuriyan J (2018) Fine-tuning of substrate preferences of the Src-family kinase Lck revealed through a high-throughput specificity screen. *elife* 7:e35190.

3. O'Shea JP, et al. (2013) pLogo: a probabilistic approach to visualizing sequence motifs. *Nat Methods* 10(12):1211–1212.

4. Shah NH, et al. (2016) An electrostatic selection mechanism controls sequential kinase signaling downstream of the T cell receptor. *elife* 5:e20105.

5. Rice JJ, Schohn A, Bessette PH, Boulware KT, Daugherty PS (2006) Bacterial display using circularly permuted outer membrane protein OmpX yields high affinity peptide ligands. *Protein Sci* 15(4):825–836.

6. Hornbeck PV, et al. (2015) PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res* 43(Database issue):D512-20.

7. Barker SC, et al. (1995) Characterization of pp60c-src tyrosine kinase activities using a continuous assay: autoactivation of the enzyme is an intermolecular autophosphorylation process. *Biochemistry* 34(45):14843–14851.

8. Huerta-Cepas J, et al. (2016) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res* 44(D1):D286-93.

9. Letunic I, Bork P (2018) 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res* 46(D1):D493–D496.

10. Katoh K, Rozewicki J, Yamada KD (2017) MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinformatics*:bbx108.

11. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22(13):1658–1659.

12. The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 45(D1):D158–D169.

13. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009) Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25(9):1189–1191.

14. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. *J Mol Graph* 14(1):33–8, 27.

15. Phillips JC, et al. (2005) Scalable molecular dynamics with NAMD. *J Comput Chem* 26(16):1781–1802.

16. Huang J, et al. (2017) CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat Methods* 14(1):71–73.

17. Hubbard SR (1997) Crystal structure of the activated insulin receptor tyrosine kinase in complex with peptide substrate and ATP analog. *EMBO J* 16(18):5572–5581.

18. Roe DR, Cheatham TE (2013) PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *J Chem Theory Comput* 9(7):3084–3095.

19. Schrödinger, LLC (2015) *The PyMOL Molecular Graphics System* Available at: http://pymol.org/ [Accessed January 25, 2018].

20. Zhang X, Gureasko J, Shen K, Cole PA, Kuriyan J (2006) An allosteric mechanism for activation of the kinase domain of epidermal growth factor receptor. *Cell* 125(6):1137–1149.