Supporting Information for: *Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning*

Running Title: Deep Mutational Scanning Enzyme Solubility

Authors: Justin R. Klesmith[1,a], John-Paul Bacik[2,b], Emily E. Wrenbeck[3], Ryszard Michalczyk[2], Timothy A. Whitehead[3,4,*]

[1] Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, Michigan, 48824;

[2] Bioscience Division, Los Alamos National Laboratory, Los Alamos, New Mexico, 87545;

[3] Department of Chemical Engineering and Materials Science, Michigan State University, East Lansing, Michigan, 48824;

[4] Department of Biosystems and Agricultural Engineering, Michigan State University, East Lansing, Michigan, 48824;

[a] Current address: Department of Chemical Engineering & Materials Science, University of Minnesota, Minneapolis, Minnesota, 55455

[b] Current address: Department of Chemistry, Princeton University, Princeton, New Jersey, 08544

* Corresponding author: Timothy A. Whitehead, 428 S. Shaw Ln. Room 2100, Michigan State University, East Lansing, Michigan, USA, 48824; (517) 432-2097; taw@egr.msu.edu

SUPPORTING INFORMATION

SI Text

Reagents

All DNA primers were ordered from IDT and genetic constructs were sequence verified by Genewiz. All chemicals and plates were purchased from Sigma-Aldrich.

Plasmid construction

The pSALECT and pETConNK plasmid backbones were used as previously described (1). In short, a ΔS4-A25 truncation using the Ambler consensus numbering system (2) of TEM-1 BLA S70A and D179G (**Notes S1 and S2**) were cloned in-between the *NdeI* and *XhoI* sites of the two backbones to create the pSALECT-TEM1.1/csTEM1 and pETConNK-TEM1.1 plasmids. The codon optimized DNA sequence for LGK (3) was cloned in-between the *NdeI* and *XhoI* sites of the two backbones to create the pSALECT-LGK/csTEM1 and pETConNK-LGK plasmids. Expression constructs for TEM-1.1 (pSAL_TEM1.1) were constructed from pSALECT-TEM1.1/csTEM1 by removing csTEM1 by PCR. Plasmids and full maps are freely available on AddGene (www.addgene.org).

Library construction

Mutagenic primers encoding degenerate bases (NNN) were used for residues G8 to T435 for LGK and H26 to W290 for TEM-1.1. Plasmids were transformed into *E. coli* XL1-Blue and plasmids were extracted using a Qiagen miniprep kit the following day.

For generation of YSD libraries, chemically competent EBY100 yeast was transformed with 5 µg of pETConNK based library plasmid DNA and grown in 50 mL SDCAAps (Synthetic complete media supplemented with amino acids, 2% (w/v) dextrose, and 10,000 u/mL penicillin/streptomycin (Invitrogen, Carlsbad, CA, USA)) (4) for 24 hours at 30ºC. The cells were passaged into fresh 50 mL SDCAAps media and grown for another 24 hours. Yeast were stored in yeast storage buffer (20 mM HEPES 150 mM NaCl pH 7.5, 20% (w/v) glycerol) (5) at -80ºC in 1 mL aliquots at an OD_{600}=1.0 (1 yeast OD_{600} = $2x10^7$ cells/mL).

10 ng of pSALECT based library plasmid DNA was transformed into electrocompetent *E. coli* MC4100 (Coli Genetic Stock Center, New Haven, CT) and plated on a Nunc Bioassay Plate (245 mm X 245 mm X 25 mm) at 30ºC overnight. Transformation controls were performed to limit double plasmid transformation (6). Cells were scraped and used to inoculate a 100 mL LB culture with 34 µg/mL chloramphenicol at an initial OD_{600} of 0.05 at 30ºC and 250 rpm. When the cultures reached mid-log (OD_{600} 0.40 to 0.60) DMSO was added at a final concentration of 7% (v/v), and 1 mL aliquots were flash frozen in liquid nitrogen.

Screening procedures

Yeast display library cell stocks were thawed at room temperature and were used to inoculate a 1 mL SDCAAps at 30ºC at an initial OD_{600} of 1.0 for 6 to 8 hours. These cells were used to start a 1.1 mL SGCAAps (Synthetic complete media supplemented with amino acids, 2% (w/v) galactose, and 10,000 u/mL penicillin/streptomycin (Invitrogen, Carlsbad, CA, USA)) at an initial OD_{600} of 1 at 30ºC for 18 hours. The next day cells were spun down at top speed for 30 sec and the media pipette removed. Cold PBSF (137 mM NaCl, 2.7 mM KCl, 8 mM Na_2HPO_4, and 2 mM KH_2PO_4 with 1 g/L BSA) was added to the pellets to an OD_{600} of 2.0. The cells were washed with chilled PBSF. The cells were then subsequently labelled and sorted. Following sorting the cells were grown in 10 mL of SDCAAps at 30ºC for 24 hours and were stored at -80ºC in yeast storage buffer at a concentration of $4x10^7$ cells per mL. DNA was extracted from the yeast and prepared for sequencing using previously published protocols (6).

TAT export library cell stocks were thawed on ice for 45 minutes prior to washing with fresh LB media. The washed cells were used to start a 5 mL culture containing LB with 34 µg/mL chloramphenicol inoculated at an initial OD_{600} of 0.05. The cells were grown aerobically at 30ºC and 250 rpm until a culture OD_{600} of 0.8. The unselected library was prepared by pelleting 1 mL culture at 17,000xg for 2 min and storing the pellet at -20ºC. Libraries were plated on 100 (TEM-1.1) or 200 µg/mL (LGK) carbinicillin plates.

The LGK libraries were plated at 0.1 OD_{600}/mL on two 100 mm diameter petri plates, while the TEM-1 libraries were plated at 3.2 OD_{600}/mL on Nunc Bioassay Plates (245 mm X 245 mm X 25 mm). The number of cells plated was sufficient to support a 200-fold coverage of the theoretical DNA library in viable cells. Plates were cultured at 30ºC in a humidified incubator for 12 hours. The following day the plates were scraped with 1x PBS, pelleted, and a Qiagen miniprep kit was used to extract DNA from saved cell pellets.

Deep sequencing and data analysis

Libraries were prepared for deep sequencing using a previously developed two step PCR method (6) with PCR primers listed in **Table S12.** The pooled library was extracted and cleaned with a

Qiagen gel cleanup kit. Deep sequencing was performed using an Illumina MiSeq in 300 bp paired-end mode. Replicates were sequenced in 250 bp paired-end mode. Sequencing data was processed using Enrich (7) to quantify the amount and enrichment of each mutation. Deep sequencing statistics are listed in **Tables S2-3.** Python scripts to calculate solubility scores are publically available at Github [user: JKlesmith] (www.github.com). Processed deep sequencing datasets are deposited at figshare (www.figshare.com).

To determine lower-bound solubility scores for each screen/selection, we first determined the half-median of read counts of the pre-selection library for each selection. This number was normalized by the ratio of post- to pre-selection read counts. Next, a lower-bound enrichment ratio (ε_{LB}) based on 10 read counts in the post-selection population was calculated:

$$\varepsilon_{LB} = \log_2\left(\frac{10}{f_{LB}}\right) \tag{5}$$

Where f_{LB} represents the normalized half-median pre-selection reads. The lower-bound solubility score was then calculated according to equations (3) and (4) using ε_{LB} as calculated above.

PSSM Analysis

A blastp search (8) of the nonredundant database for LGK and TEM-1 was performed with an e-value cut-off of 10^{-4} and filtered to the top 20,000 results. Synthetic or engineered constructs were excluded from the hits. Hits were also excluded if they covered less than 85% of the query sequence or if their sequence identity was less than 34% for LGK or 40% for TEM-1. Cd-hit (9) was used at 98% clustering threshold and default parameters. MUSCLE (10) was then used to produce a multiple sequence alignment of the top 700 clusters. DSSP (11) was then used to identify residues that are a part of loops and a part of secondary structure elements. Insert sequences in loop regions were removed such that the alignment has no gaps in the wild-type sequence. An alignment of sequences without any frameshifts was then independently extracted from each structured and non-structured region. PSI-BLAST (12) was then used on each region with the wild-type sequence as the query sequence.

Protein expression and purification

E. coli BL21*(DE3) harboring plasmid pSAL_TEM1.1 were grown to an OD_{600} of 0.8 at 37°C supplemented with 20 μg/ml chloramphenicol. Protein expression was induced with 1 mM IPTG for 18 h at 18°C. Proteins were purified using Ni-NTA chromatography exactly according to Bienick et al. (13). Apparent melting temperatures were measured by a modified SYPRO Orange thermal shift assay (3).

Recombinant *E. coli* BL21- GOLD (DE3) cells (Agilent Technologies) harboring plasmid pET29b_LGK-G359R were grown to an OD_{600} of ~ 0.5 at 37 °C, with shaking, in 500-ml volumes of LB media supplemented with 35 μg/ml kanamycin. Expression of LGK was induced with 1 mM IPTG for 3 h at 30 °C, with shaking. Cells were pelleted by centrifugation and stored at -80 °C. Pellets were thawed in 20 ml of ice-cold lysis buffer (0.5 M NaCl, 20 mM Tris-HCl pH 7.5, 0.1 mM PMSF, 2 mM imidazole) and lysed using a sonicator (Ameco). The lysate was clarified by centrifugation and mixed with 2 ml of TALON metal affinity resin (Clontech) with gentle shaking for 30 min. at room temperature. The TALON beads were centrifuged and re-suspended in binding buffer (500 mM NaCl, 20 mM Tris pH 7.5, 0.5 mM TCEP) before being poured into a 20 ml gravity column. The column was washed with 20 ml of binding buffer supplemented with 5 mM imidazole (Sigma), followed by 20 ml of binding buffer supplemented with 10 mM imidazole. The LGK protein was eluted from the column with 10 ml of binding buffer supplemented with 250 mM imidazole. The protein was further purified by gel filtration (HiPrep 26/60 Sephacryl S-200 HR) in 20 mM Tris pH 7.5, 50 mM NaCl, 0.5 mM TCEP prior to concentration using an Amicon Ultra-15 concentrator with a 10,000 Da cut-off (Millipore). Chromatographic steps were performed using an AKTA FPLC (GE Healthcare).

LGK crystallization, data collection and structure determination

LGK crystals were grown at room temperature using the hanging drop vapor-diffusion method by mixing equal volumes of reservoir buffer (22% polyethylene glycol (PEG) 3350, 0.2 M K_2SO_4, 100 mM Tris pH 6.8) and LGK (23 mg/ml) in crystallization buffer (50 mM NaCl, 2 mM ADP, 4 mM $MgCl_2$, 0.5 mM TCEP, 20 mM Tris pH 7.5). Crystals were cryoprotected by dragging them through a drop containing cryoprotectant solution, reservoir buffer supplemented with 9% sucrose (w/v), 2% glucose (w/v), 8% glycerol (v/v), 8% ethylene glycol (v/v), prior to being flash-cooled in liquid nitrogen. Data was collected at the Stanford Synchrotron Radiation Lightsource beamline BL7-1, integrated using MOSLFM (14) and scaled and merged using SCALA (15).

Structure was determined using rigid body refinement using (PDB identifier: 5BSB) as the starting model followed by iterative model building and refinement performed using Coot and PHENIX (16, 17). The stereochemical quality of the final model was assessed using MolProbity (18). Refinement statistics are presented in **Table S13**. All structural figures were prepared using PyMOL (19).

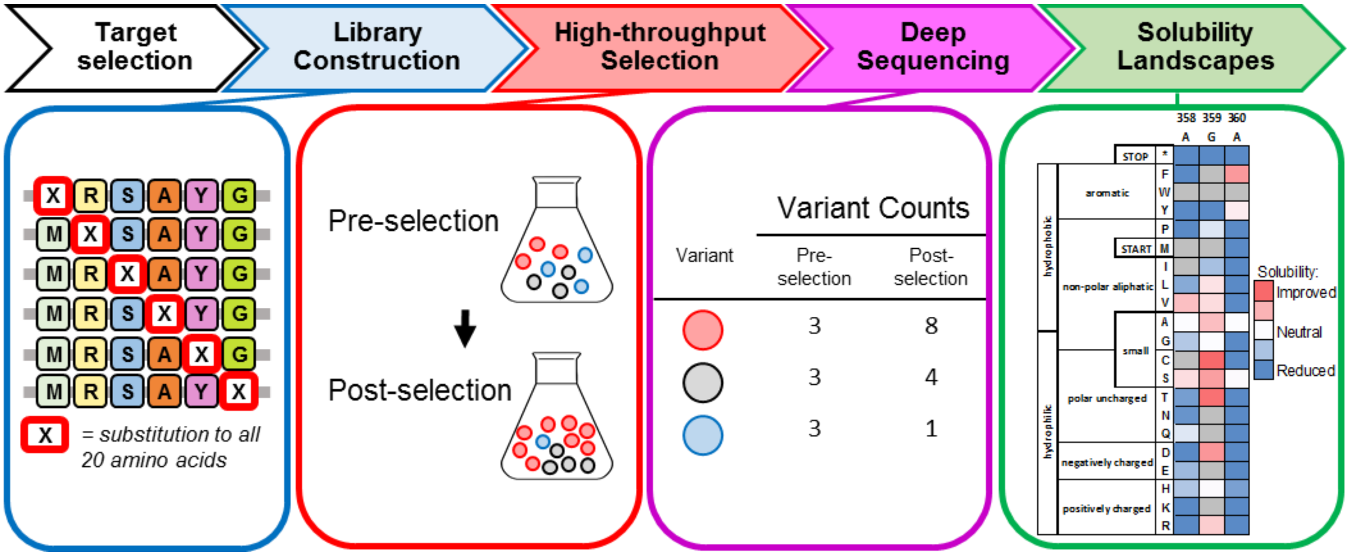
Note S1: The amino acid sequence for TEM-1.1. Mutations S70A and D179G are underlined in red highlight.

HPETLVKVKDAEDQLGARVGYIELDLNSGKILESFRPEERFPMM**A**TFKVLLCGAVLSRV
DAGQEQLGRRIHYSQNDLVEYSPVTEKHLTDGMTVRELCSAAITMSDNTAANLLLTTIG
GPKELTAFLHNMGDHVTRLDRWEPELNEAIPNDER**G**TTMPAAMATTLRKLLTGELLTL
ASRQQLIDWMEADKVAGPLLRSALPAGWFIADKSGAGERGSRGIIAALGPDGKPSRIVVI
YTTGSQATMDERNRQIAEIGASLIKHW

Note S2: The DNA sequence for TEM-1.1.

CACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGT
GGGTTACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGA
AGAACGTTTTCCAATGATGGCCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTATCC
CGTGTTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGA
CTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAA
GAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTTC
TGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGAT
CATGTAACTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGA
CGAGCGTGGCACCACGATGCCTGCAGCAATGGCAACAACGTTGCGCAAACTATTAA
CTGGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGG
ATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTATTGCTG
ATAAATCTGGAGCAGGTGAGCGTGGGTCTCGCGGTATCATTGCAGCACTGGGGCCA
GATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTATG
GATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGG

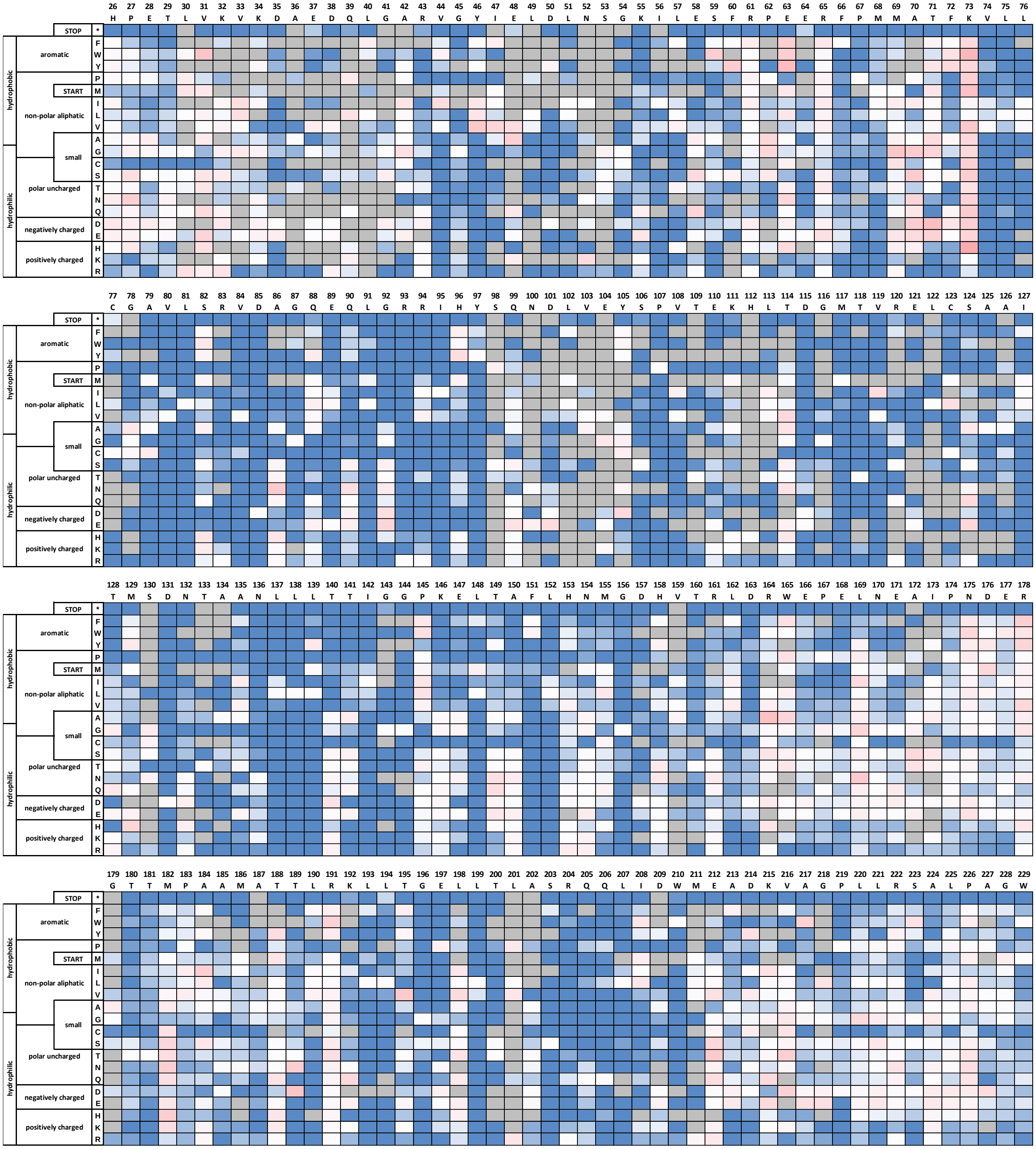
Fig. S1: Deep sequencing pipeline. A target protein is first mutagenized such that a DNA library encodes all possible amino acids. Next, a high-throughput selection or screen is performed to enrich beneficial mutants and deplete deleterious mutations. Deep sequencing is used to count each mutation to allow the frequency of that mutation in the population to be calculated. Finally, the frequencies are normalized to a solubility score for each mutation.

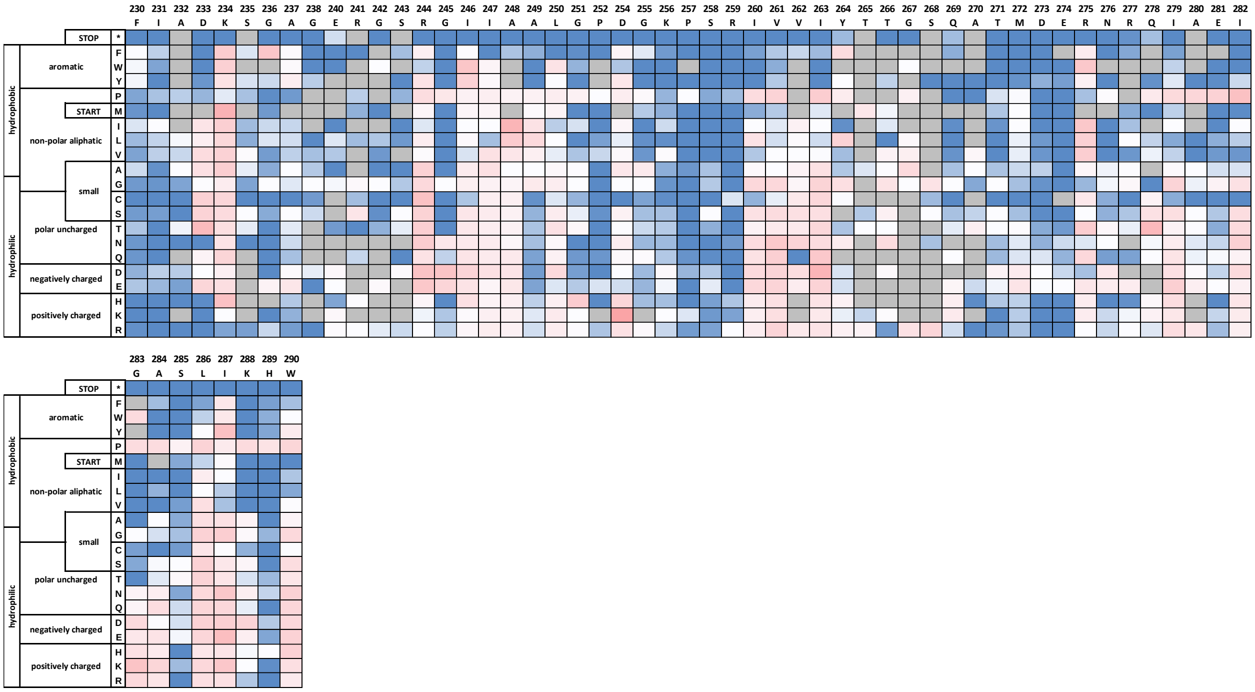


Variant Counts

Variant	Pre-selection	Post-selection
🔴	3	8
⚪	3	4
🔵	3	1

Fig. S2: Heatmap of solubility score of TEM-1.1 variants screened by yeast display.





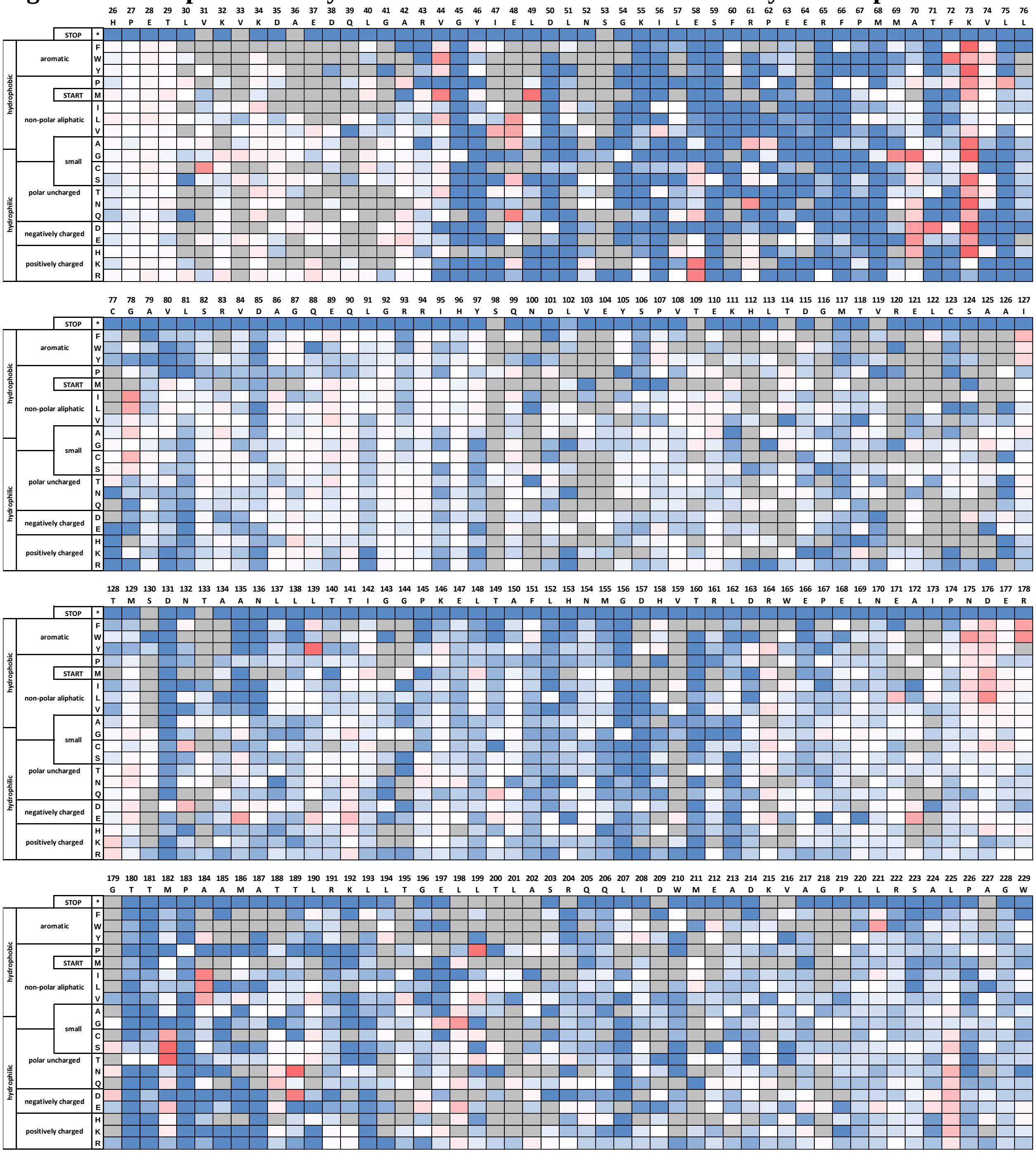
Solubility Score Key

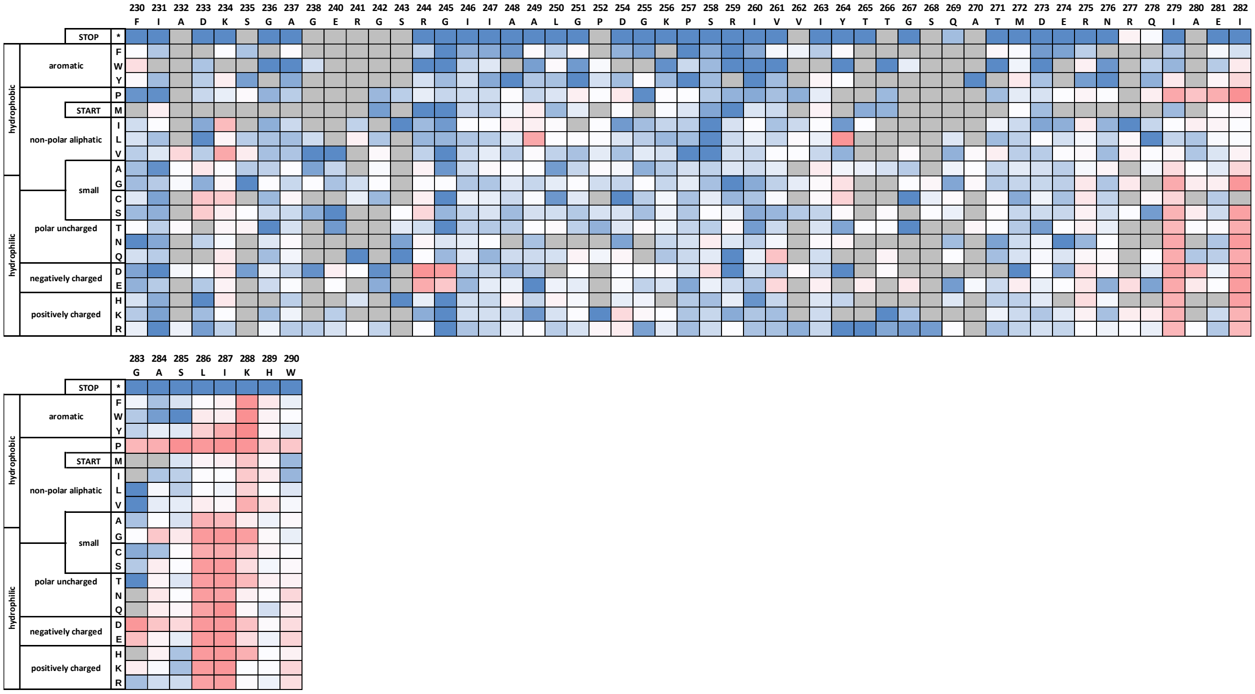
YSD		
≥2.00	🟥	Improved
1.00	🟥	
0.00	⬜	Neutral
-0.25	🟦	
≤-0.50	🟦	Reduced

<12 reference counts:

⬜

Fig. S3: Heatmap of solubility score of TEM-1.1 variants screened by TAT export.





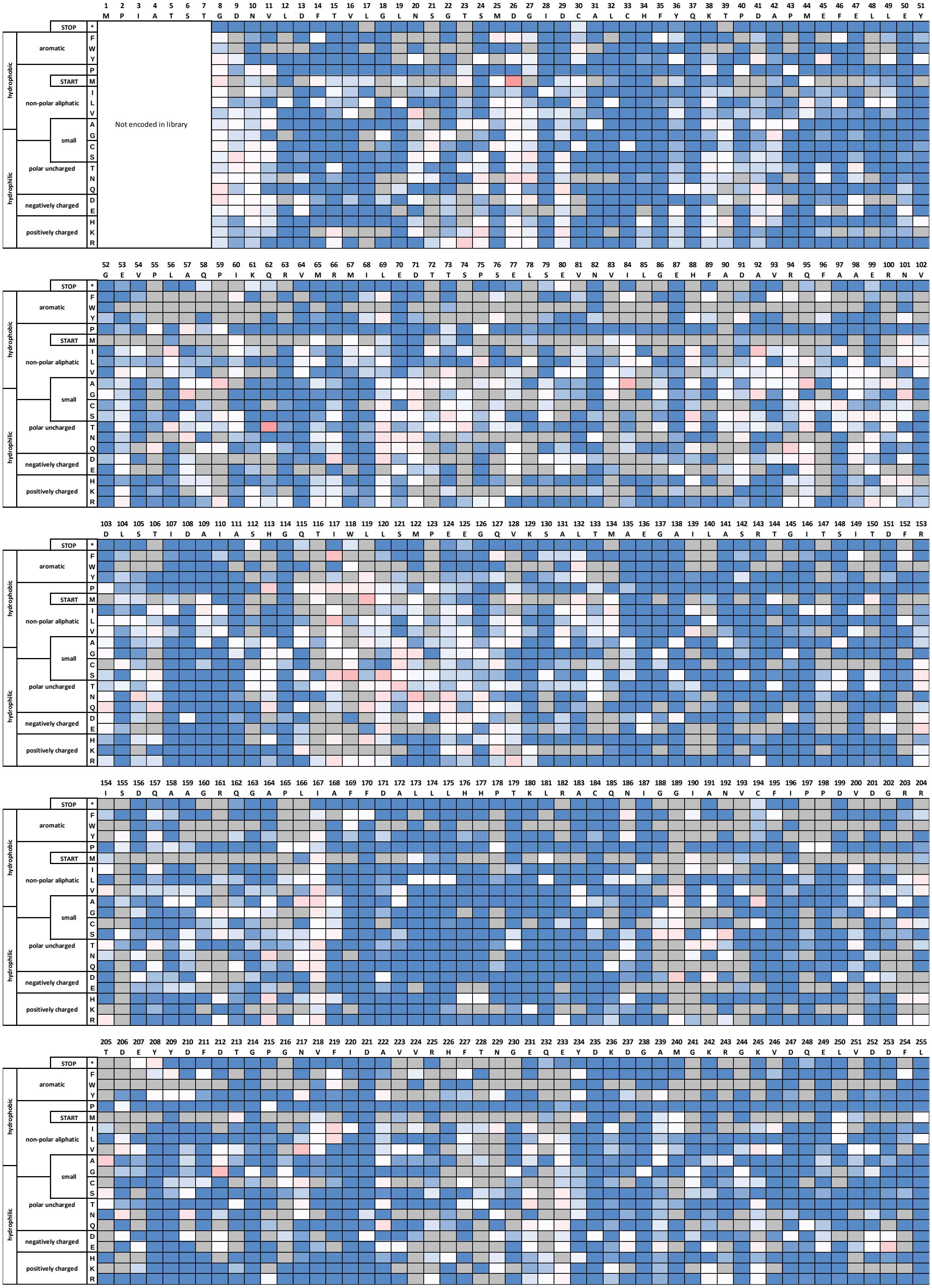
Solubility Score Key

TAT		
≥4.00	🟥	Improved
2.00	🟥	
0.00	⬜	Neutral
-0.50	🟦	
≤-1.00	🟦	Reduced

<12 reference counts:

▨

Fig. S4: Heatmap of solubility score of LGK variants screened by yeast display.



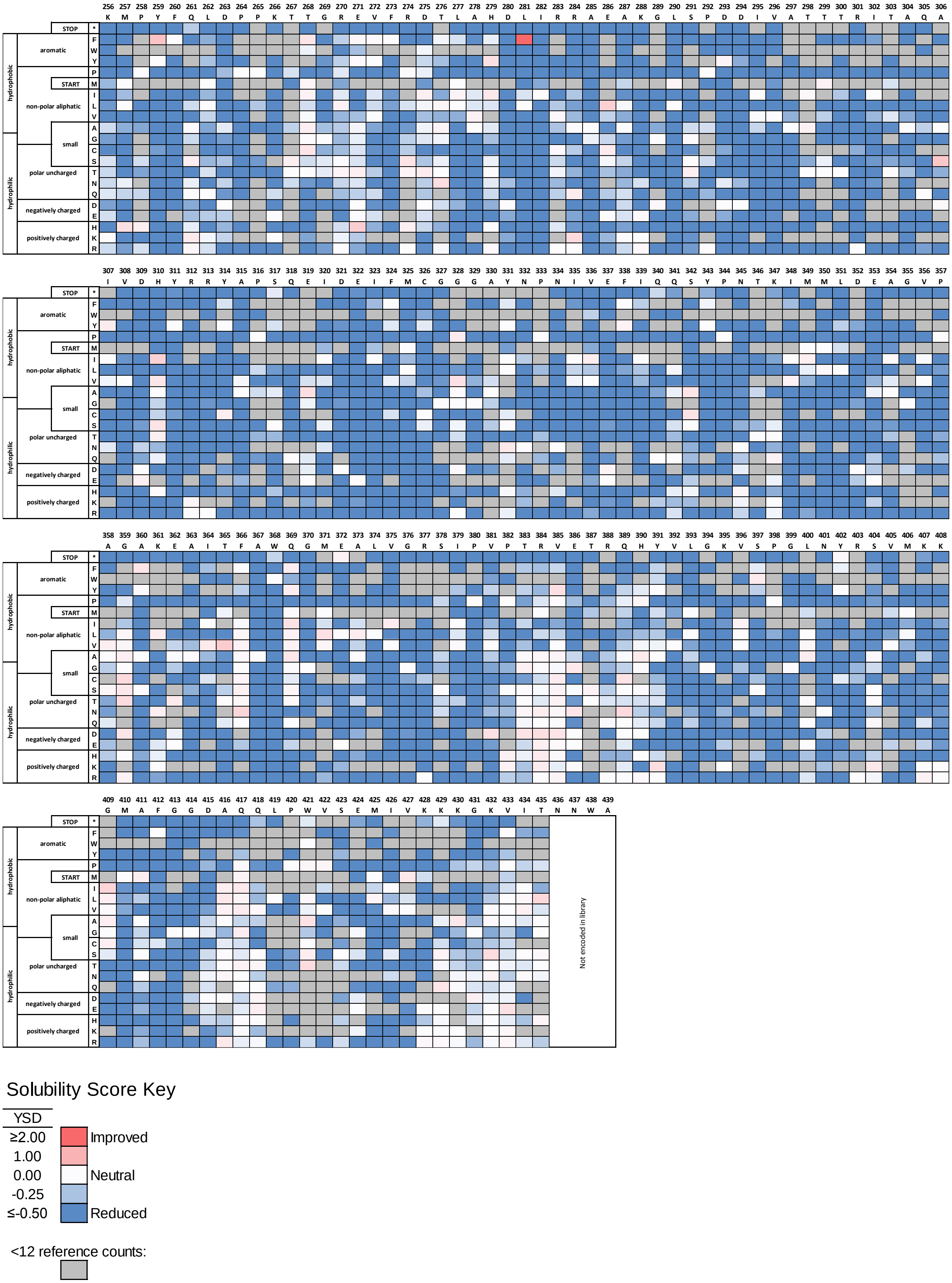
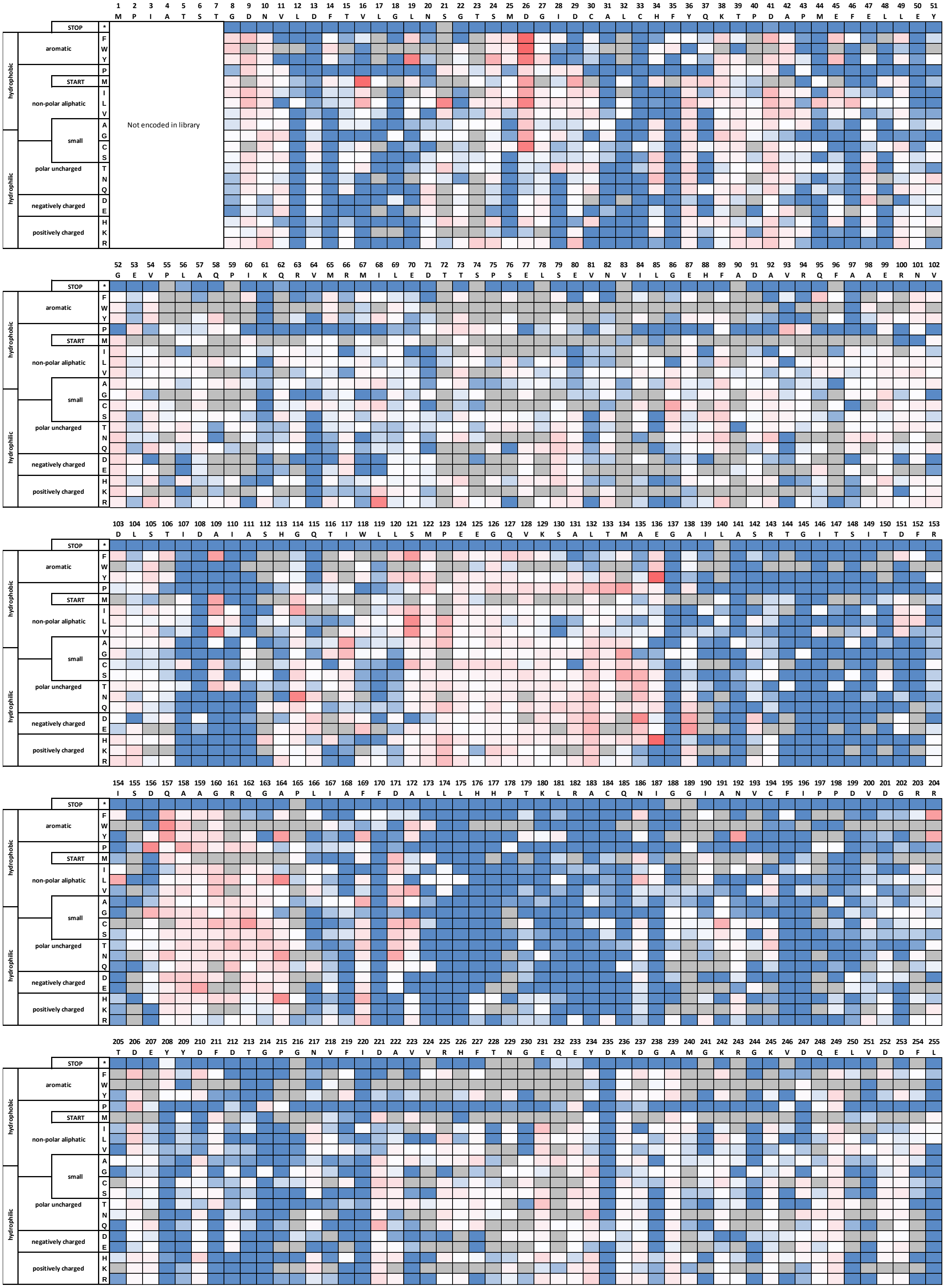

256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306
K M P Y F Q L D P P K T T G R E V F R D T L A H D L I R R A E E A K G L S P D D I V A T T T R I T A Q A
307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357
I V D H Y R R Y A P S Q E I D E I F M C G G A Y N P N I V E F I Q Q S Y P N T K I M M L D E A G V P
358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408
A G A K E A I T F A W Q G M E A L V G R S I P V P T R V E T R Q H Y V L G K V S P G L N Y R S V M K K
409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439
G M A F G G D A Q Q L P W V S E M I V K K K G K V I T N N W A
STOP
*
aromatic
F
W
Y
P
START
M
hydrophobic
non-polar aliphatic
I
L
V
A
G
small
C
S
polar uncharged
T
N
Q
hydrophilic
negatively charged
D
E
positively charged
H
K
R
Not encoded in library
Solubility Score Key
YSD
≥2.00
1.00
0.00
-0.25
≤-0.50
Improved
Neutral
Reduced
<12 reference counts:

Fig. S5: Heatmap of solubility score of LGK variants selected by TAT export.



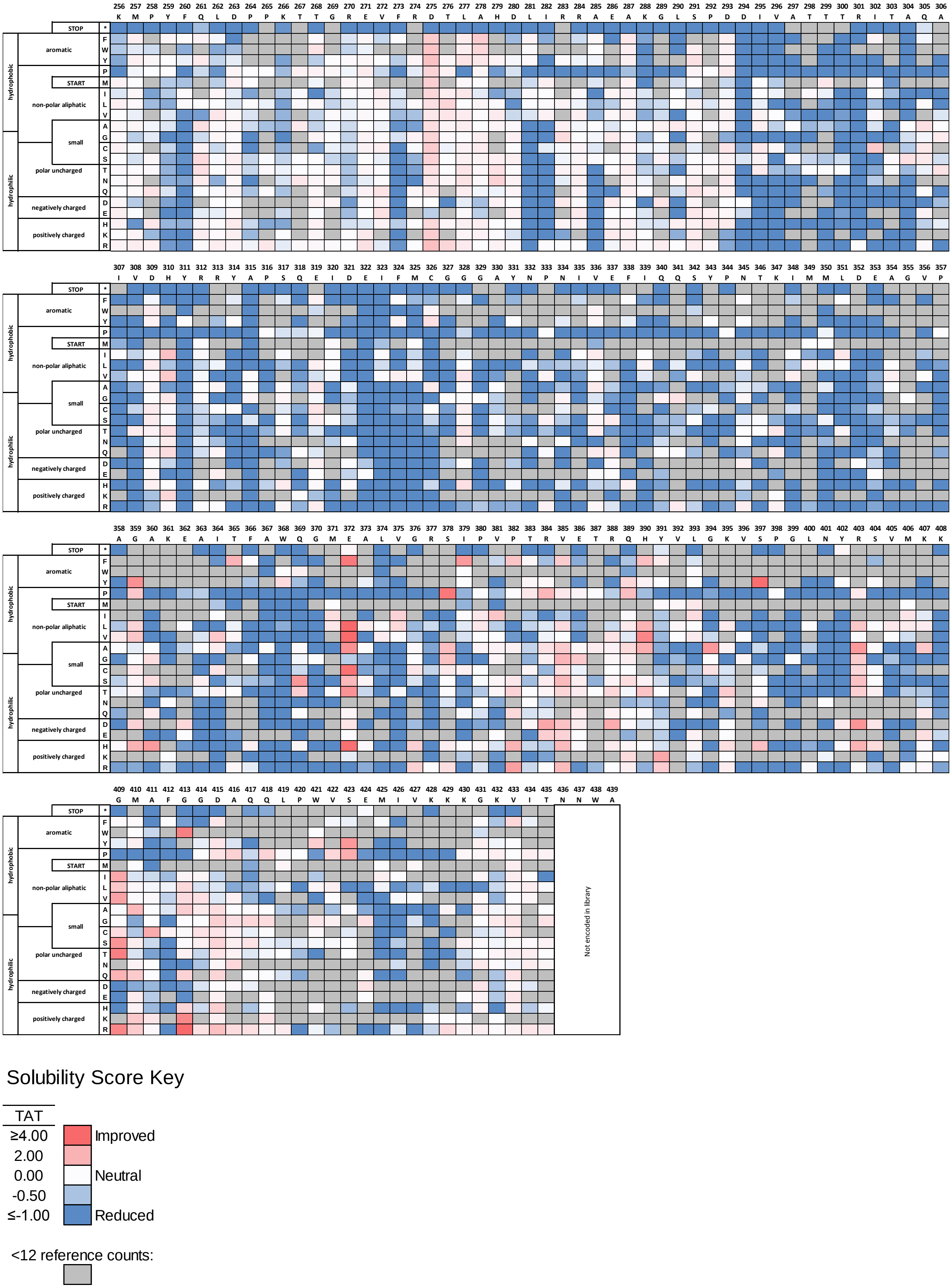

256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306
K M P Y F Q L D P P K T T G R E V F R D T L A H D L I R R A E A K G L S P D D I V A T T T R I T A Q A
307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357
I V D H Y R R Y A P S Q E I D E I F M C G G G A Y N P N I V E F I Q Q S Y P N T K I M M L D E A G V P
358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408
A G A K E A I T F A W Q G M E A L V G R S I P V P T R V E T R Q H Y V L G K V S P G L N Y R S V M K K
409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439
G M A F G G D A Q Q L P W V S E M I V K K K G K V I T N N W A
STOP
*
hydrophobic
aromatic
F
W
Y
P
START
M
non-polar aliphatic
I
L
V
small
A
G
C
hydrophilic
S
polar uncharged
T
N
Q
negatively charged
D
E
positively charged
H
K
R
Not encoded in library
Solubility Score Key
TAT
≥4.00
2.00
0.00
-0.50
≤-1.00
Improved
Neutral
Reduced
<12 reference counts:

Fig. S6: Correspondence between enrichment ratios for experimental replicates. A,B) LGK from residues 331 and 435 for the (A.) TAT genetic selection and (B.)YSD screen, respectively. C,D,E) TEM-1.1 from residues 88 and 175 for (C.) YSD, (D.) two TAT replicates performed on the same day, and (E.) the same two TAT replicates compared to a replicate performed on a different day. Replicates used in Figures B,C, and E are taken six months apart by different individuals. Red line indicates a theoretical estimation of error at a confidence of 2 standard deviations. Correlation coefficient for all mutations and mutations with at least 100 reference counts inset on each graph.

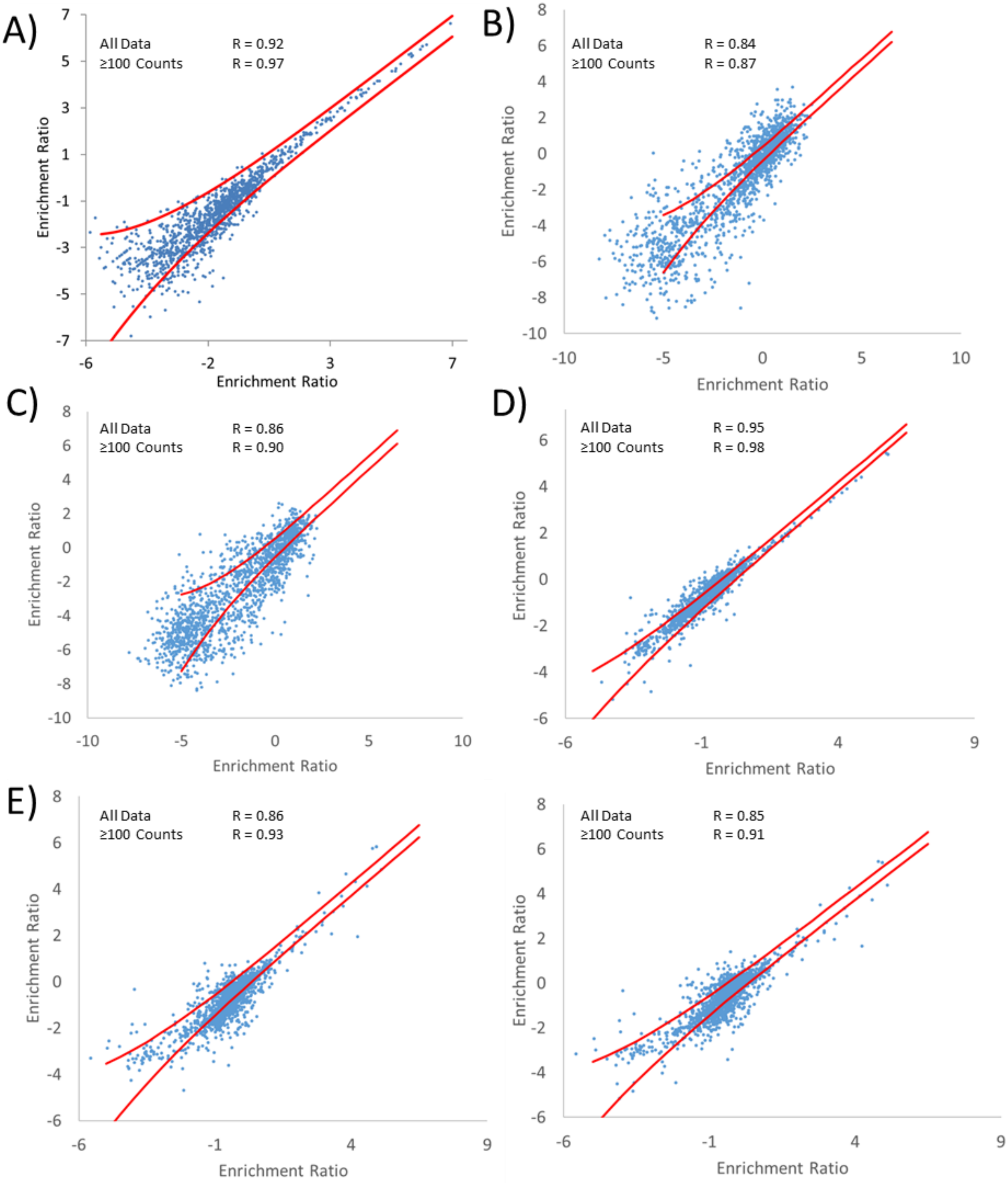


Fig. S7: Replicate fitness measurements for synonymous mutations to the wild-type sequence for the LGK TAT dataset from residues 331 to 435. Blue symbols are mutations represented at under 30 counts in the unselected population, whereas red symbols denote mutations represented 30 or more times in the unselected population**.**

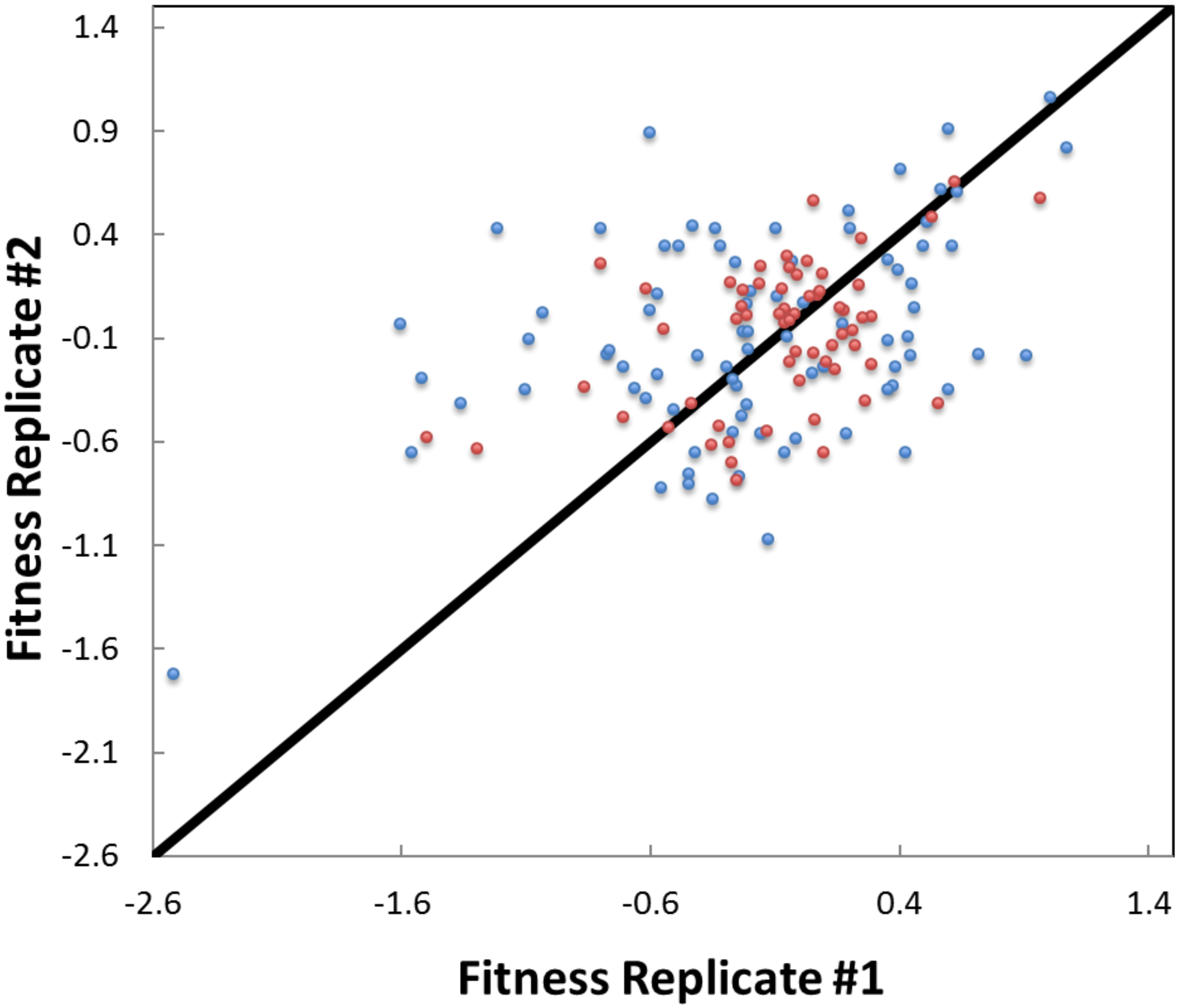


Fig S8: Distribution of fitness for synonymous mutations to wild-type sequence used in solubility screens and selection. (a.-b.) Data is shown as closed circles (blue – LGK; red – TEM-1.1), while lines are Gaussian best-fits. Panels are for **(a.)** YSD, and **(b.)** TAT genetic selection. **(c.-d.)** Volcano plots of unselected counts as a function of solubility score for **(c.)** YSD screen for LGK, and **(d.)** YSD screen for TEM-1.1.

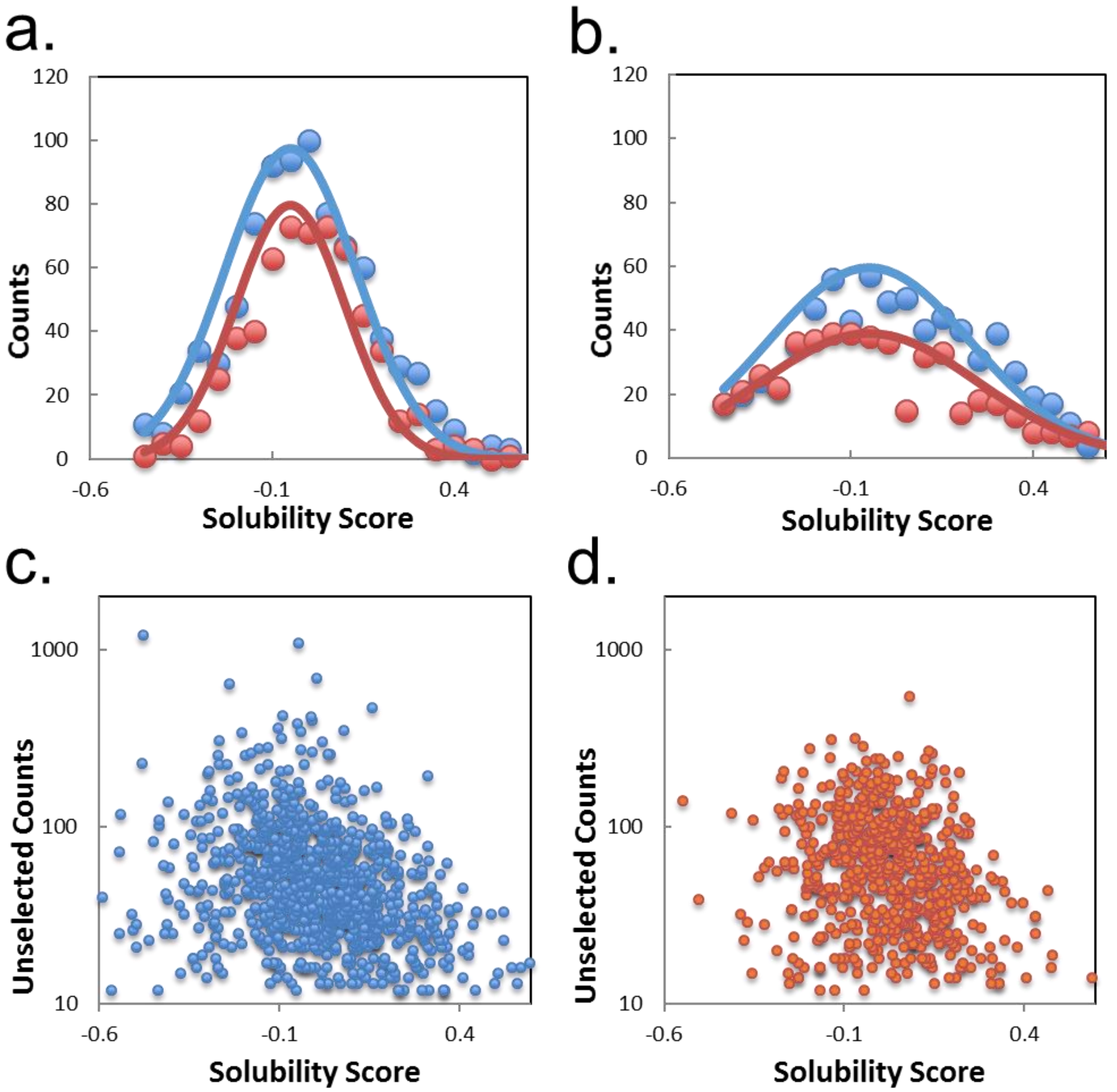


Fig. S9: Nonsense versus missense distributions for the TAT selection. An unpaired t-test with Welch's correction was performed between the fitness metrics for nonsense and missense mutations of each enzyme (n=331 and 6386 for nonsense and missense mutations in LGK respectively, n=227 and 3976 for nonsense and missense mutations in TEM-1.1 respectively).

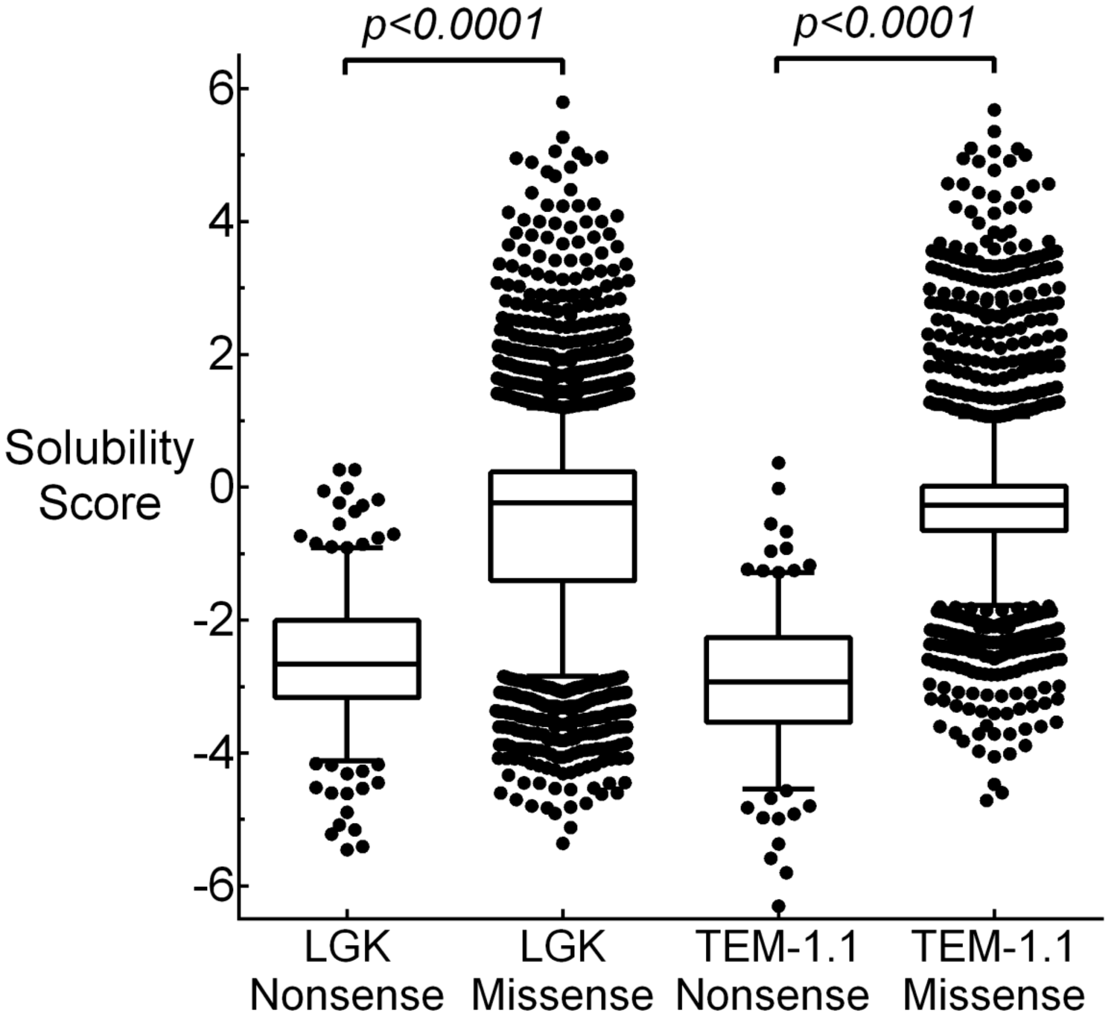
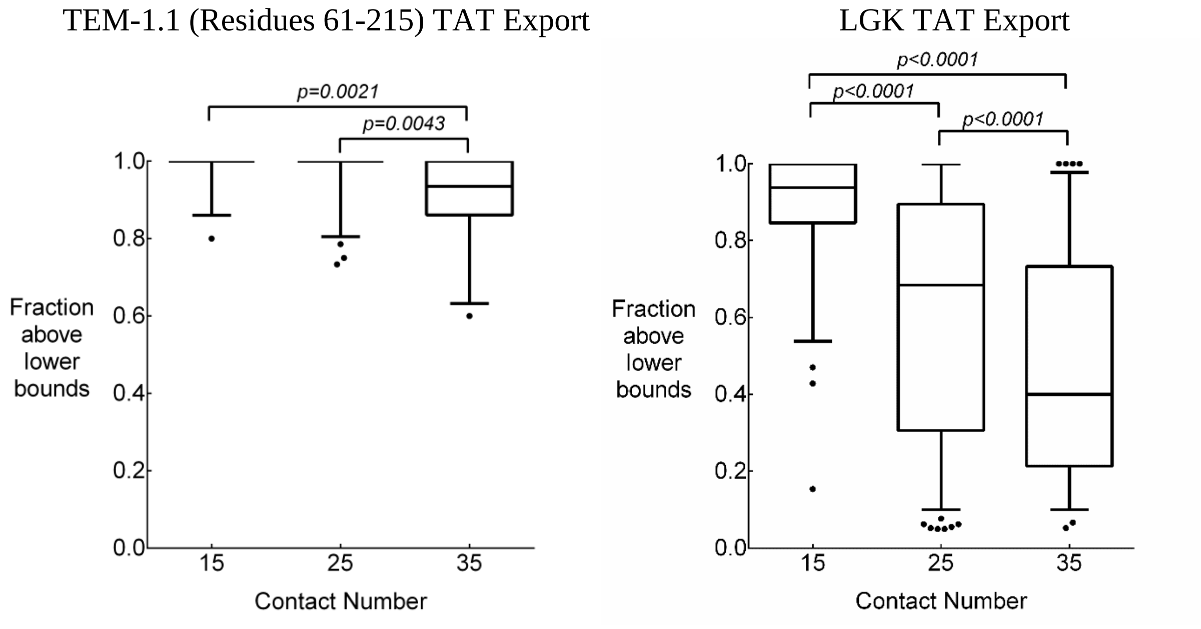


Fig. S10: Fraction of mutations above lower bounds versus contact number for TAT export. An unpaired t-test with Welch's correction was performed between each bin.


TEM-1.1 (Residues 61-215) TAT Export

LGK TAT Export

Fig. S11: Linear regressions of solubility versus functional datasets for LGK for (a.) YSD, and (b.) TAT genetic selection. The second selection using LGK.1 as the starting construct from Klesmith et. al. (3) was used as the functional dataset comparison for the solubility screens (denoted as "Selection Two" on the X-axis).

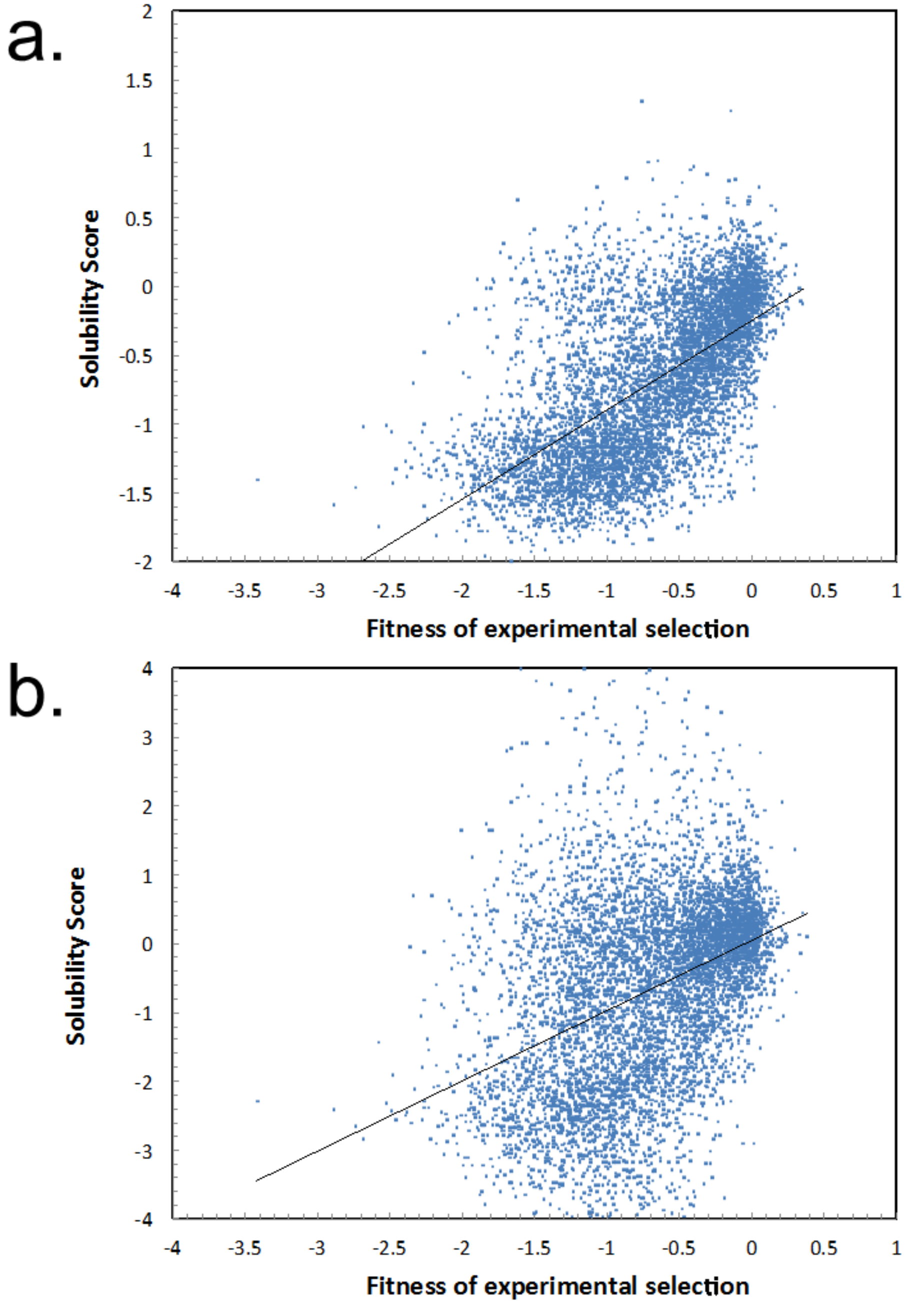


Fig. S12: Linear regressions of solubility versus functional datasets for TEM-1.1 positions 61-215 (Ambler sequencing convention) for (a.) YSD and (b.) TAT genetic selection. Fitness values derived from Firnberg et. al.(20) were used for the functional dataset comparison for the solubility screens.

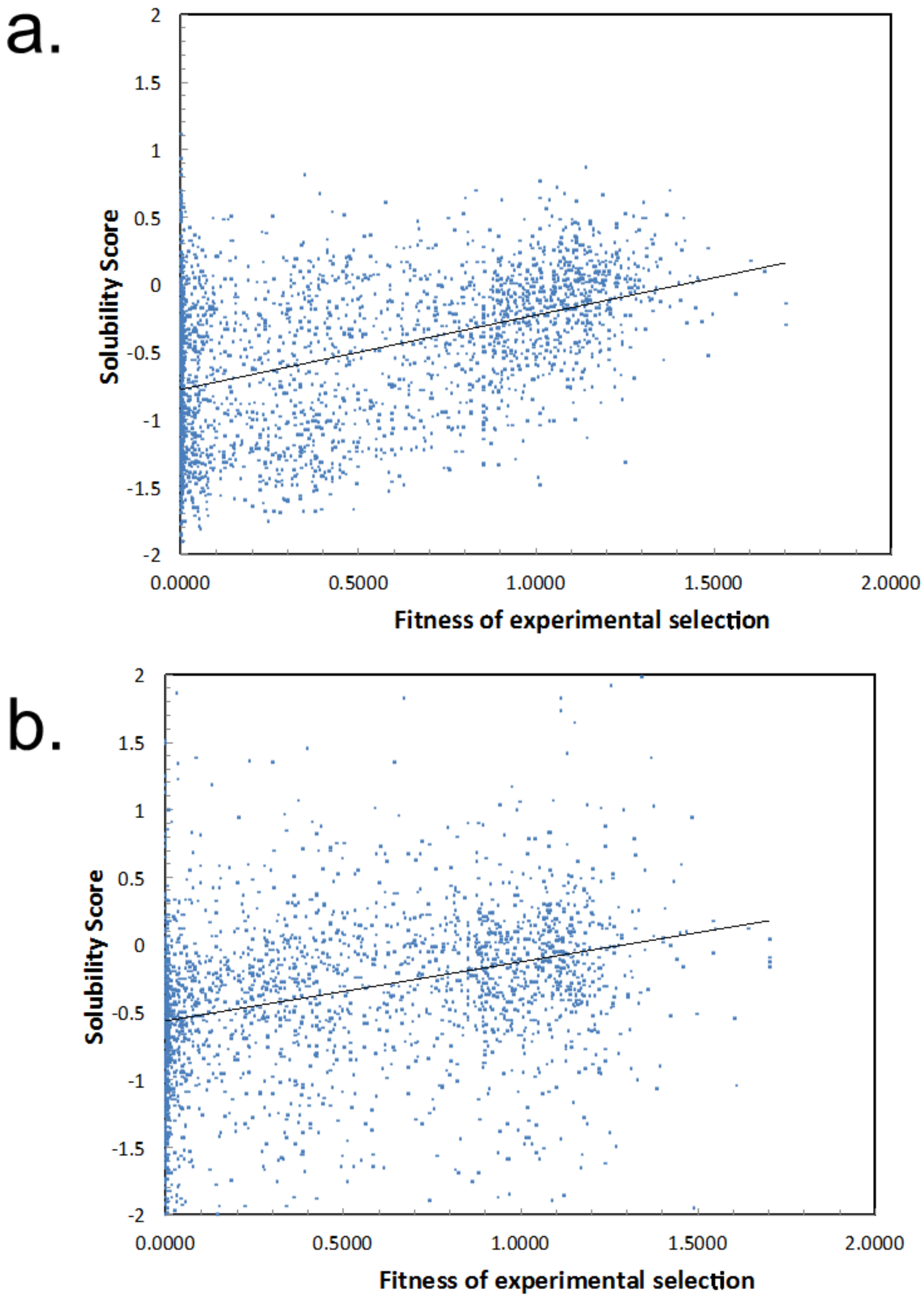


Table S1: Sorting statistics for LGK and TEM-1.1 libraries. NS: samples not sorted.

Enzyme	Tile Number	Replicate	Method	Tile Length (AA)	Events Collected	Percent Sorted (Display)	Percent Sorted (Top)	Theoretical DNA Library Diversity	Fold Oversampling
LGK	1	1	Yeast Display	103	600,000	23.7	7.7	6,592	91
LGK	2	1	Yeast Display	110	700,000	25.8	5.2	7,040	99
LGK	3	1	Yeast Display	110	700,000	21.8	6.8	7,040	99
LGK	4	1	Yeast Display	105	700,000	19.6	4.6	6,720	104
LGK	4	2	Yeast Display	105	500,246	NS	5.0	6,720	74
TEM-1.1	1	1	Yeast Display	87	500,000	43.0	5.4	5,568	90
TEM-1.1	2	1	Yeast Display	88	500,000	48.4	6.2	5,632	89
TEM-1.1	2	2	Yeast Display	88	486,000	NS	5.0	5,632	86
TEM-1.1	3	1	Yeast Display	88	500,000	50.3	6.0	5,632	89

Table S2: Deep sequencing library statistics for the yeast display screens.

Screen	Yeast Display															
Enzyme	LGK									TEM-1.1						
Tile Number	1		2		3		4			1		2			3	
Sort Population	Display	Top	Display	Top	Display	Top	Display	Top	Replicate	Display	Top	Display	Top	Replicate	Display	Top
Number of mutated codons	103		110		110		105			87		88			88	
Reference sequencing reads post quality filter	607,904		469,478		396,561		259,784		494,743	307,817		417,079		1,258,917	413,919	
Selected sequencing reads post quality filter	363,962	512,699	322,740	254,035	329,005	333,191	199,535	288,638	423,415	288,082	355,306	491,648	660,166	776,914	531,726	441,713
Percent of mutant codons with:																
1-bp substitution	100.0		100.0		99.7		99.7		99.9	99.6		99.7		100.0	99.7	
2-bp substitution	69.1		79.8		78.6		82.0		85.0	81.2		88.0		88.1	83.4	
3-bp substitution	63.8		69.2		71.4		75.5		77.4	75.0		79.5		76.1	77.4	
All substitutions	71.3		78.1		78.5		81.7		83.8	81.2		86.0		84.6	83.1	
Percent of reads with:																
No nonsynonymous mutations	46.5		45.3		40.1		35.1		31.8	43.4		29.2		28.1	30.2	
One nonsynonymous mutation	47.3		41.7		51.5		53.6		46.4	50.0		58.3		52.7	60.1	
Multiple nonsynonymous mutations	6.1		13		8.3		11.3		21.8	6.6		12.5		19.2	9.7	
Coverage of possible single nonsynonymous mutations	89.5		90.5		90.5		91.8		92.7	91.2		96.9		96.3	96.3	

Table S3: Deep sequencing library statistics for the TAT pathway selections.

Screen	Tat pathway									
Enzyme	LGK					TEM-1.1				
Tile Number	1	2	3	4		1	2			3
Number of mutated codons	103	110	110	105		87	88			88
Reference sequencing reads post quality filter	512,321	453,663	491,970	167,053		458,755	490,299	1,114,552		402,030
Selected sequencing reads post quality filter	437,570	367,820	514,209	110,898	184,418	469,768	450,252	1,189,544	1,006,879	525,271
Percent of mutant codons with:										
1-bp substitution	99.9	100.0	99.9	99.4		99.7	99.6	100.0		99.1
2-bp substitution	88.0	96.2	92.0	84.2		86.0	83.6	88.2		79.4
3-bp substitution	84.2	93.7	86.4	78.4		80.8	75.0	77.7		72.3
All substitutions	88.1	95.7	90.7	83.8		85.8	82.2	85.4		79.2
Percent of reads with:										
No nonsynonymous mutations	38.7	35.1	37.6	35.0		28.2	28.3	25.7		27.1
One nonsynonymous mutation	51.8	53.9	52.2	52.4		61.2	63.9	56.8		62.1
Multiple nonsynonymous mutations	9.5	11.1	10.2	12.6		10.5	7.8	17.5		10.7
Coverage of possible single nonsynonymous mutations	96.6	99.4	94.8	85.4		93.3	92.6	95.5		90.7

Table S4: Standard deviation for synonymous mutations at different depths of coverage.

Screen	YSD	YSD	TAT	TAT
Protein	LGK	TEM-1.1	LGK	TEM-1.1
Depth of Coverage (Unselected Counts)				
All Data	0.24	0.18	0.43	0.37
>12 & <30 counts	0.30	0.23	0.64	0.49
>=30 counts	0.19	0.17	0.29	0.32
>29 & <100 counts	0.18	0.16	0.31	0.36
>99 counts	0.20	0.17	0.19	0.24

Table S5: Correlation coefficients between solubility scores and fitness measurements.

Screen	YSD	YSD	TAT	TAT	YSD	TAT
Protein	LGK	TEM-1.1*	LGK	TEM-1.1*	TEM-1.1	TEM-1.1
Depth of Coverage (Unselected Counts)						
All Data	0.61	0.45	0.39	0.22	0.26	0.13
>12 & <30 counts	0.54	0.45	0.36	0.14	0.32	0.18
>=30 counts	0.63	0.45	0.40	0.24	0.24	0.12

*Considering positions 61-215 using Ambler sequence convention.

Table S6: Known stabilizing mutations in TEM-1.

Position	Mutation	ΔTm (°C)	Solubility Score YSD	Solubility Score TAT	Reference
31	V31R	3.2	0.19	-0.31	(21)
60	F60Y	2.6	0.64	0.21	(21)
62	P62S	1	0.27	-0.01	(22)
78	G78A	1.5	0.29	1.42	(21)
82	S82H	2.2	0.32	-0.21	(21)
92	G92D	4.1	0.41	0.00	(21)
104	E104K	1.7	0.07	-0.39	(23)
120	R120G	1.8	-0.09	-0.94	(24)
147	E147G	2.6	0.12	-0.43	(24)
153	H153R	3.3	0.23	0.09	(24)
182	M182T	5	0.43	5.10	(24)
201	L201P	1.4	0.26	-0.19	(24)
208	I208M	1.1	0.22	-0.32	(25)
224	A224V	3.1	0.21	-0.38	(22)
235	S235A	1.7	-0.06	-0.42	(26)
265	T265M	1.6	0.30	-0.70	(7)
275	R275L	5	0.59	-0.04	(22)
275	R275Q	2	0.50	0.21	(7)
276	N276D	1.3	0.29	0.24	(7)

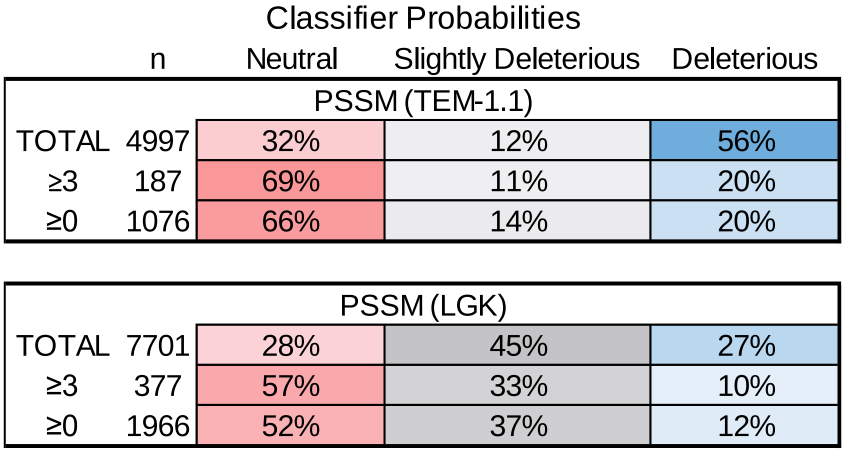
Table S7: Known stabilizing mutations in LGK. All mutations and associated biophysical data come from (3). NS – mutation is not seen in the dataset.

Position	Mutation	ΔTm (°C)	Solubility Score YSD	Solubility Score TAT
75	P75L	1.4	0.36	-0.50
94	R94H	1.9	-0.10	-0.72
113	H113G	4.9	0.07	-0.75
135	A135G	2.6	-0.17	-0.75
140	L140I	2.2	NS	-0.55
167	I167H	9.8	0.16	-0.58
194	C194T	6.0	-0.43	1.50
212	D212A	1.4	0.11	-0.67
268	T268C	4.0	0.32	-0.75
306	A306S	1.1	0.72	-0.25
359	G359R	1.1	0.15	-0.71
369	Q369L	3.4	0.15	-0.75

Table S8: Number of hits from the solubility dataset and *in vitro* datasets at different threshold values.

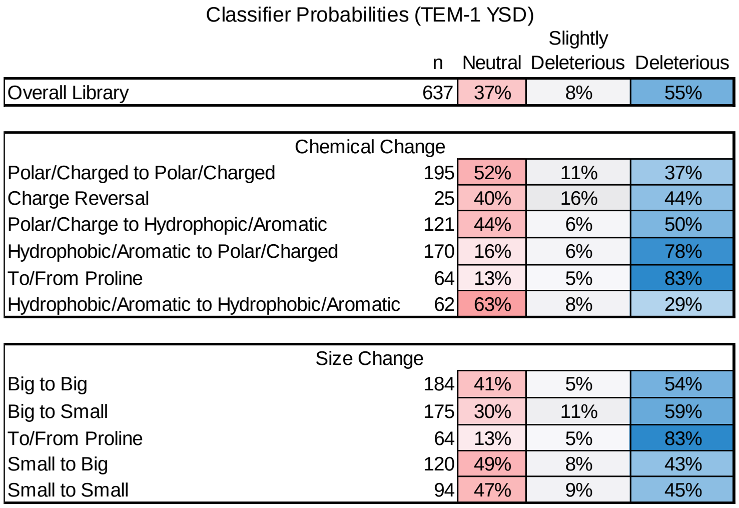
Solubility Score			Overall Dataset		*in vitro* dataset		
Cutoff	Screen	Protein	Hits	Non Hits	Hits	Non Hits	p-value (Fisher exact test)
0.15	YSD	LGK	317	6788	6	5	3.20E-06
0.24 (1 sigma all data)	YSD	LGK	182	6923	3	8	0.002
0.19 (1 sigma >=30 counts)	YSD	LGK	258	6847	3	8	0.007
0.38 (2 sigma >=30 counts)	YSD	LGK	76	7029	1	10	0.11
0.15	YSD	TEM-1.1	632	3981	15	4	2.90E-10
0.18 (1 sigma all data)	YSD	TEM-1.1	552	4061	15	4	4.20E-11
0.17 (1 sigma >=30 counts)	YSD	TEM-1.1	573	4040	15	4	7.20E-11
0.34 (2 sigma >=30 counts)	YSD	TEM-1.1	293	4320	5	14	0.005
0.15	TAT	LGK	1944	5212	1	11	0.2
0.43 (1 sigma all data)	TAT	LGK	1225	5931	1	11	0.7
0.29 (1 sigma >=30 counts)	TAT	LGK	1561	5595	1	11	0.48
0.58 (2 sigma >=30 counts)	TAT	LGK	946	6210	1	11	1
0.15	TAT	TEM-1.1	772	3695	5	14	0.36
0.37 (1 sigma all data)	TAT	TEM-1.1	484	3983	2	17	1
0.32 (1 sigma >=30 counts)	TAT	TEM-1.1	533	3934	2	17	1
0.64 (2 sigma >=30 counts)	TAT	TEM-1.1	314	4153	2	17	0.39

Table S9: PSSM classifier probabilities independent of a solubility screen.



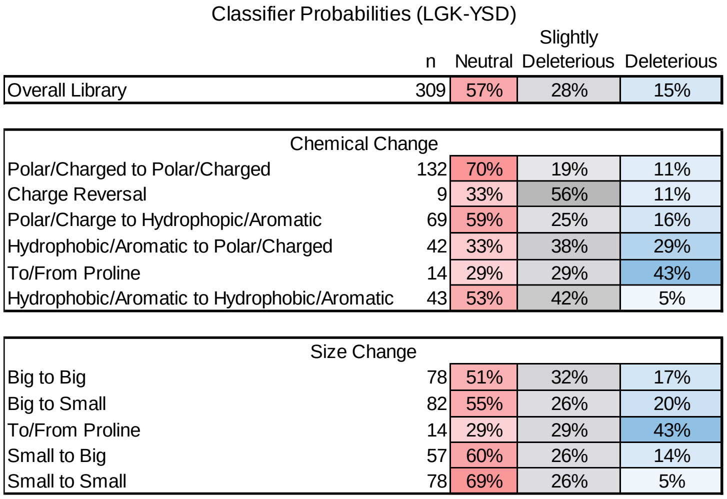
		Classifier Probabilities		
	n	Neutral	Slightly Deleterious	Deleterious
		PSSM (TEM-1.1)		
TOTAL	4997	32%	12%	56%
≥3	187	69%	11%	20%
≥0	1076	66%	14%	20%
		PSSM (LGK)		
TOTAL	7701	28%	45%	27%
≥3	377	57%	33%	10%
≥0	1966	52%	37%	12%

Table S10: Classifier probabilities for chemical changes and size changes.



Classifier Probabilities (TEM-1 YSD)

	n	Neutral	Slightly Deleterious	Deleterious
Overall Library	637	37%	8%	55%
Chemical Change				
Polar/Charged to Polar/Charged	195	52%	11%	37%
Charge Reversal	25	40%	16%	44%
Polar/Charge to Hydrophopic/Aromatic	121	44%	6%	50%
Hydrophobic/Aromatic to Polar/Charged	170	16%	6%	78%
To/From Proline	64	13%	5%	83%
Hydrophobic/Aromatic to Hydrophobic/Aromatic	62	63%	8%	29%
Size Change				
Big to Big	184	41%	5%	54%
Big to Small	175	30%	11%	59%
To/From Proline	64	13%	5%	83%
Small to Big	120	49%	8%	43%
Small to Small	94	47%	9%	45%



Classifier Probabilities (LGK-YSD)

	n	Neutral	Slightly Deleterious	Deleterious
Overall Library	309	57%	28%	15%
Chemical Change				
Polar/Charged to Polar/Charged	132	70%	19%	11%
Charge Reversal	9	33%	56%	11%
Polar/Charge to Hydrophopic/Aromatic	69	59%	25%	16%
Hydrophobic/Aromatic to Polar/Charged	42	33%	38%	29%
To/From Proline	14	29%	29%	43%
Hydrophobic/Aromatic to Hydrophobic/Aromatic	43	53%	42%	5%
Size Change				
Big to Big	78	51%	32%	17%
Big to Small	82	55%	26%	20%
To/From Proline	14	29%	29%	43%
Small to Big	57	60%	26%	14%
Small to Small	78	69%	26%	5%

Table S11: Filters and Bayes analyses for LGK YSD screen.

	LGK - YSD				
	Basal	PSSM ≥3	PSSM Filter	Naïve Bayes	Bayes + Filter
n =	309	39	58	242	125
Neutral	57%	82%	90%	66%	77%
Slightly Deleterious	28%	13%	7%	26%	19%
Deleterious	15%	5%	3%	8%	4%

Table S12: Inner PCR tile primers. Illumina outer PCR attach point sequences are underlined.

Name	Sequence
pETCONNKFWD	GTTCAGAGTTCTACAGTCCGACGATCAGGGTCGGCTAGC
pETCONNKREV	CCTTGGCACCCGAGAATTCCAAAGCTTTTGTTCGGATC
pSALECTFWD	GTTCAGAGTTCTACAGTCCGACGATCACGTGCGACTGCG
pSALECTREV	CCTTGGCACCCGAGAATTCCATTAACCAGGGTCTCCG
LGKTILE1REV	CCTTGGCACCCGAGAATTCCAGCCGTGCGAAGC
LGKTILE2FWD	GTTCAGAGTTCTACAGTCCGACGATCACCATTGACGCAATC
LGKTILE2REV	CCTTGGCACCCGAGAATTCCACGAACCACTGCGTC
LGKTILE3FWD	GTTCAGAGTTCTACAGTCCGACGATCGGCAACGTGTTCATC
LGKTILE3REV	CCTTGGCACCCGAGAATTCCACCACAATATTCGGGTTATA
LGKTILE4FWD	GTTCAGAGTTCTACAGTCCGACGATCCGGTGGCGCC
TEMTILE1REV	CCTTGGCACCCGAGAATTCCACATGCCATCCGTAAG
TEMTILE2FWD	GTTCAGAGTTCTACAGTCCGACGATCCCAGTCACAGAAAAGCAT
TEMTILE2REV	CCTTGGCACCCGAGAATTCCATGCCGGGAAGCTAG
TEMTILE3FWD	GTTCAGAGTTCTACAGTCCGACGATCATTAACTGGCGAACTACTTACT

Table S13: Data processing and refinement statistics for LGK G359R crystallographic structure (values in parentheses refer to the high-resolution shell).

Data Collection	
Space group	$P4_12_12$
Unit cell (Å)	*a* = *b* = 70.06, *c* = 261.77 α = β = γ = 90.00
Wavelength (Å)	0.9795
Resolution range (Å)	46.33 – 1.80 (1.90 – 1.80)
Total observations	411319
Total unique observations	61638
I/σ_I	9.4 (1.7)
Completeness (%)	99.9 (100.0)
R_{merge}	0.133 (1.045)
R_{pim}	0.056 (0.428)
Redundancy	6.7 (6.9)
Refinement Statistics	
Resolution (Å)	43.08-1.80
Reflections (total)	61556
Reflections (test)	3097
Total atoms refined	3830
Solvent	460
R_{work} (R_{free})	0.18 (0.21)
RMSDs Bond lengths (Å) / angles (º)	0.008/0.828
Ramachandran plot (Favored/allowed(%))	97.2/2.6
Average B, all atoms (Å^2)	24.0

SI References:

1. Wrenbeck EE, *et al.* (2016) Plasmid-based one-pot saturation mutagenesis. *Nat Meth* advance online publication.
2. Ambler RP, *et al.* (1991) A standard numbering scheme for the class A beta-lactamases. *Biochemical Journal* 276(Pt 1):269-270.
3. Klesmith JR, Bacik JP, Michalczyk R, & Whitehead TA (2015) Comprehensive Sequence-Flux Mapping of a Levoglucosan Utilization Pathway in E. coli. *ACS Synth Biol* 4(11):1235-1243.
4. Chao G, *et al.* (2006) Isolating and engineering human antibodies using yeast surface display. *Nat. Protocols* 1(2):755-768.
5. Whitehead TA, *et al.* (2012) Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat Biotechnol* 30(6):543-548.
6. Kowalsky CA, *et al.* (2015) High-resolution sequence-function mapping of full-length proteins. *PLoS One* 10(3):e0118193.
7. Fowler DM, Araya CL, Gerard W, & Fields S (2011) Enrich: software for analysis of protein function by enrichment and depletion of variants. *Bioinformatics* 27(24):3430-3431.
8. Altschul SF, Gish W, Miller W, Myers EW, & Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* 215(3):403-410.
9. 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.
10. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32(5):1792-1797.
11. Kabsch W & Sander C (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22(12):2577-2637.
12. Altschul SF, Gertz EM, Agarwala R, Schäffer AA, & Yu Y-K (2009) PSI-BLAST pseudocounts and the minimum description length principle. *Nucleic Acids Research* 37(3):815-824.
13. Bienick MS, *et al.* (2014) The interrelationship between promoter strength, gene expression, and growth rate. *PLoS One* 9(10):e109105.
14. Battye TGG, Kontogiannis L, Johnson O, Powell HR, & Leslie AGW (2011) iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallographica Section D* 67(4):271-281.
15. Evans P (2006) Scaling and assessment of data quality. *Acta Crystallographica Section D* 62(1):72-82.
16. Emsley P & Cowtan K (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* 60(Pt 12 Pt 1):2126-2132.
17. Afonine PV, *et al.* (2012) Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallographica Section D: Biological Crystallography* 68(Pt 4):352-367.
18. Chen VB, *et al.* (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D* 66(1):12-21.
19. DeLano WL (2002) The PyMOL Molecular Graphics System (DeLano Scientific, Palo Alto, CA, USA.).

20. Firnberg E, Labonte JW, Gray JJ, & Ostermeier M (2014) A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape. *Molecular Biology and Evolution* 31(6):1581-1592.

21. Deng Z, *et al.* (2012) Deep Sequencing of Systematic Combinatorial Libraries Reveals β-Lactamase Sequence Constraints at High Resolution. *Journal of Molecular Biology* 424(3–4):150-167.

22. Kather I, Jakob RP, Dobbek H, & Schmid FX (2008) Increased Folding Stability of TEM-1 β-Lactamase by In Vitro Selection. *Journal of Molecular Biology* 383(1):238-251.

23. Raquet X, *et al.* (1995) Stability of TEM β-lactamase mutants hydrolyzing third generation cephalosporins. *Proteins: Structure, Function, and Bioinformatics* 23(1):63-72.

24. Bershtein S, Goldin K, & Tawfik DS (2008) Intense Neutral Drifts Yield Robust and Evolvable Consensus Proteins. *Journal of Molecular Biology* 379(5):1029-1044.

25. Brown NG, Pennington JM, Huang W, Ayvaz T, & Palzkill T (2010) Multiple Global Suppressors of Protein Stability Defects Facilitate the Evolution of Extended-Spectrum TEM β-Lactamases. *Journal of Molecular Biology* 404(5):832-846.

26. Dubus A, Wilkin JM, Raquet X, Normark S, & Frère JM (1994) Catalytic mechanism of active-site serine β-lactamases: role of the conserved hydroxy group of the Lys-Thr(Ser)-Gly triad. *Biochemical Journal* 301(2):485-494.