Supporting information for

Kinetic characterization of over 100 glycoside hydrolase mutants enables the discovery of structural features correlated with kinetic constants

Dylan Alexander Carlin[1]*, Ryan W. Caster[2]*, Xiaokang Wang[3], Stephanie A. Betzenderfer[2], Claire X. Chen[2], Veasna M. Duong[2], Carolina V. Ryklansky[2], Alp Alpekin[2], Nathan Beaumont[2], Harhul Kapoor[2], Nicole Kim[2], Hosna Mohabbot[2], Boyu Pang[2], Rachel Teel[2], Lillian Whithaus[2], Ilias Tagkopoulos[2,6], Justin B. Siegel[2,3,4]

Author affiliations:

1. Biophysics Graduate Group, University of California, Davis
2. Genome Center, University of California, Davis
3. Department of Chemistry, University of California, Davis
4. Department of Biochemistry & Molecular Medicine, University of California, Davis
5. Department of Biomedical Engineering, University of California, Davis
6. Department of Computer Science, University of California, Davis

Table of contents

Mutant	**Protein yield**	**SDS-PAGE**	**K_M**	**k_{cat}**	**K_I**	**k_{cat}/K_M**
	mg/mL		**mM**	**min^{-1}**	**mM**	**$M^{-1}min^{-1}$**
BglB	1.2		5.00 ± 0.2	880 ± 10		176,000 ± 8000
S14A	0.6		8.25 ± 1.02	320 ± 11		38,823 ± 4,972
S16A	0.83		14.01 ± 0.40	154 ± 1		10,997 ± 331
E17S	1.01		7.32 ± 0.38	641 ± 9		87,596 ± 4,719
S17A	0.65		18.45 ± 3.72	848 ± 76		45,978 ± 10,135
Y18A	0.17		31.55 ± 3.61	197 ± 9		6,230 ± 773
Q19A	0.26					11 ± 3
Q19C	0.4					< 10
Q19S	0.43					13 ± 3
W34A	0.1	NDE				
V52G	0.97		8.25 ± 0.54	687 ± 13		83,371 ± 5,707
F75A	0.44		5.47 ± 0.28	613 ± 8		112,224 ± 6,000
R76A	0.1	NDE				
I91E	0.49		6.71 ± 0.79	846 ± 35		126,071 ± 15,714
H101R	1.03		10.62 ± 0.53	1059 ± 16		99,708 ± 5,225
H119A	1.21		15.10 ± 3.36	143 ± 11		9,483 ± 2,222
H119E	0.25	NDE				
H119N	1.02		23.22 ± 2.20	2 ± <0		82 ± 8
W120A	0.16					

W120F	0.78		16.08 ± 2.07	472 ± 21		29,334 ± 3,980
W120H	1		89.18 ± 4.31	84 ± 2		943 ± 53
V147S	0.23		6.45 ± 0.62	5 ± <1		706 ± 70
E154D	1.42		3.46 ± 0.76	878 ± 47		254,004 ± 57,175
N163A	0.74		11.95 ± 0.91	7 ± <1		558 ± 44
N163C	1.1		5.42 ± 0.32	26 ± <1		4,766 ± 291
N163D	1.05		15.19 ± 1.41	12 ± <1		789 ± 77
E164A	0.42		1.01 ± 0.17			190 ± 33
Y166P	0.18		2.50 ± 0.45	27 ± 1	94.95 ± 10.18	10,596 ± 1,981
C167A	0.48		14.56 ± 1.27	479 ± 14		32,884 ± 3,026
C167Q	0.94		4.92 ± 0.19	504 ± 6	590.71 ± 86.56	102,415 ± 4,149
L171A	0.38		11.09 ± 0.42	807 ± 9		72,719 ± 2,851
L171R	1.06		3.36 ± 0.23	403 ± 7		120,146 ± 8,506
T175R	0.86		3.59 ± 0.15	801 ± 8		223,033 ± 9,663
E177A	0.96		5.98 ± 0.22	986 ± 10		164,804 ± 6,408
E177K	0.95		6.19 ± 0.30	555 ± 7	362.94 ± 36.97	89,609 ± 4,493
E177L	0.77		7.48 ± 0.36	670 ± 10		89,478 ± 4,555
H178A	0.25		7.67 ± 0.73	113 ± 3	173.34 ± 42.79	14,697 ± 1,463

A192S	1.17		5.09 ± 0.18	946 ± 10		185,848 ± 6,994
T218A	0.98		6.51 ± 0.94	464 ± 18		71,280 ± 10,669
L219A	0.47		7.87 ± 0.60	199 ± 5		25,262 ± 2,010
N220A	0.61		10.27 ± 0.68	405 ± 8		39,425 ± 2,745
N220H	1.12		5.14 ± 0.21	123 ± 1		23,874 ± 1,031
M221A	0.73		6.25 ± 0.60	547 ± 15		87,554 ± 8,701
E222A	0.29		0.63 ± 0.15	90 ± 4	95.24 ± 13.70	143,604 ± 36,130
E222H	0.7		8.54 ± 0.53	160 ± 3		18,695 ± 1,212
E222K	0.5		7.22 ± 0.75	108 ± 3		14,955 ± 1,618
E222Q	1.3		12.16 ± 0.65	668 ± 11		54,923 ± 3,084
E222R	0.15		2.48 ± 0.44	42 ± 2		17,098 ± 3,148
E222Y	0.7		18.43 ± 3.14	12 ± 1		636 ± 116
M223G	0.88		19.21 ± 2.91	154 ± 9		7,998 ± 1,302
R240A	1.11		19.46 ± 1.17	11011 ± 258		565,763 ± 36,384
R240D	0.8		10.82 ± 0.47	282 ± 4		26,093 ± 1,196
R240K	1.4		17.67 ± 3.32	898 ± 59		50,829 ± 10,102
I244E	0.6		5.97 ± 1.04	497 ± 23		83,137 ± 14,963
I244N	0.21		2.15 ± 0.13	271 ± 4		126,176 ± 7,795
M261E	0.11					702 ± 73
Q284R	0.52		9.68 ± 1.35	370 ± 15		38,182 ± 5,550

N293A	0.68		9.67 ± 0.44	13 ± 0		1,313 ± 63
Y294A	0.59		4.98 ± 0.17	166 ± 2		33,260 ± 1,180
Y294F	0.73		5.99 ± 0.32	735 ± 11		122,751 ± 6,883
Y295A	0.69					< 10
Y295G	0.77					< 10
T296A	0.39		11.05 ± 0.77	109 ± 2	142.75 ± 28.67	9,904 ± 722
S298E	1.1		5.28 ± 0.05	809 ± 2		153,264 ± 1,391
I300N	1.51		4.48 ± 0.32	693 ± 13		154,732 ± 11,520
Q313R	1.07		3.58 ± 0.51	689 ± 24		192,373 ± 28,109
H315N	0.08	NDE				
M323A	0.35		9.34 ± 0.80	416 ± 11	126.29 ± 25.47	44,477 ± 3,991
M323K	0.05	NDE				
W325A	0.26		1.61 ± 0.23	29 ± 1	171.02 ± 19.85	18,243 ± 2,607
W325C	0.22		4.18 ± 0.53	10 ± 0	159.19 ± 34.38	2,503 ± 327
W325H	1.12		3.08 ± 0.43	35 ± 1	143.45 ± 25.04	11,358 ± 1,645
W325L	1.08		5.74 ± 0.35	109 ± 2		18,909 ± 1,198
P329W	0.47	NDE				

S331A	0.89		4.34 ± 0.11	817 ± 5		188,306 ± 5,055
K341A	0.92		5.46 ± 0.33	1046 ± 17		191,689 ± 12,041
T352A	0.7		14.26 ± 1.76	60 ± 2		4,174 ± 541
E353A	0.56					< 10
N354A	0.34		5.38 ± 0.67	3 ± 0		547 ± 70
G355A	0.17					13 ± 2
M358T	0.62		4.83 ± 0.48	436 ± 11		90,241 ± 9,225
H373R	1.19		6.31 ± 0.30	707 ± 9		112,169 ± 5,512
H379R	0.14		6.24 ± 0.84	2 ± 0		380 ± 53
W399A	0.96		16.65 ± 2.52	0 ± 0		14 ± 2
W399C	0.93		70.33 ± 5.89	3 ± 0		39 ± 4
W399G	1.27					< 10
W399S	1.5					< 10
S400A	0.41		3.22 ± 0.22	531 ± 9		164,795 ± 11,833
D403A	0.12	NDE				
N404A	1.41		9.42 ± 0.42	4 ± 0		393 ± 18
F405A	0.11	NDE				
E406A	0.57					< 10
E406D	0.27		34.13 ± 2.57	39 ± 1		1,146 ± 94
N407A	1.12		11.09 ± 0.64	50 ± 1		4,464 ± 273
W407A	0.12	NDE				
W407G	0.05	NDE				

W407Q	0.1	NDE				
W407R	0.15	NDE				
W409A	0.1	NDE				
K413A	1.11		2.92 ± 0.48	835 ± 33		285,858 ± 48,589
F415A	0.53		16.63 ± 4.00	1 ± 0		80 ± 20
E423S	1.08		6.60 ± 0.42	646 ± 12	317.35 ± 65.22	97,777 ± 6,431

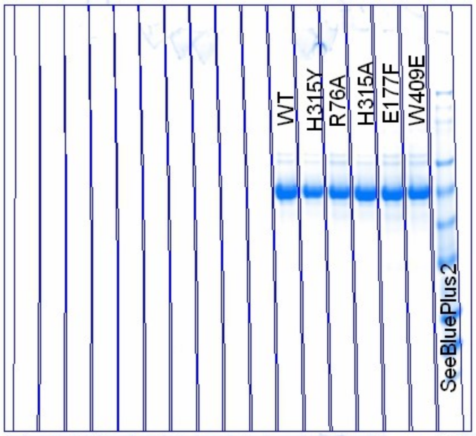
Supplemental Table 1: Kinetic constants for 104 computationally-designed BglB mutants. Included are columns (1) the mutation (2) protein yield as assessed by ratio of aborbance at 260 and 280 nm (3) protein yield as assessed by SDS-PAGE (4, 5, 6, 7) kinetic constants and nonlinear regression analysis for each of k_{cat}, K_M, KI, and k_{cat} /K_M.

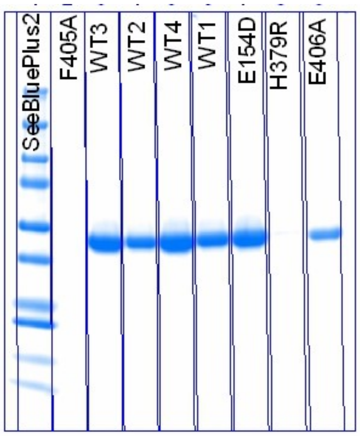
Feature	**PCC (Kcat/K_M）**	**SRC (Kcat/K_M）**	**PCC (1/K_M)**	**SRC (1/K_M)**	**PCC (K_{cat})**	**SRC (K_{cat})**
all_cst	0.182	0.116	0.039	0.071	0.140	0.120
fa_rep	0.289	0.268	0.253	0.155	0.064	0.084
hbond_sc	-0.352	-0.352	-0.248	-0.297	-0.309	-0.266
SR_1_all_cst	-0.061	-0.039	-0.148	-0.114	-0.163	-0.119
SR_1_burunsat_pm	0.000	0.000	0.000	0.000	0.000	0.000
SR_1_fa_rep	0.195	0.180	0.193	0.309	0.155	0.042
SR_1_hbond_pm	0.000	0.000	0.000	0.000	0.000	0.000
SR_1_hbond_sc	0.000	0.000	0.000	0.000	0.000	0.000
SR_1_nlpstat_pm	0.142	0.087	0.263	0.317	0.007	0.055
SR_1_pstat_pm	0.164	0.096	0.050	0.064	-0.131	-0.069
SR_1_total_score	-0.078	0.093	-0.092	-0.050	-0.081	0.024
SR_2_all_cst	0.039	-0.064	-0.089	-0.059	-0.005	-0.096
SR_2_burunsat_pm	0.000	0.000	0.000	0.000	0.000	0.000
SR_2_fa_rep	0.074	0.168	0.258	0.168	0.011	0.014
SR_2_hbond_pm	0.087	0.092	0.162	0.191	0.149	0.116
SR_2_hbond_sc	-0.149	-0.235	-0.142	-0.229	-0.126	-0.138
SR_2_nlpstat_pm	0.168	0.142	0.031	0.038	0.080	0.092
SR_2_pstat_pm	0.102	0.070	-0.042	-0.017	-0.023	0.013

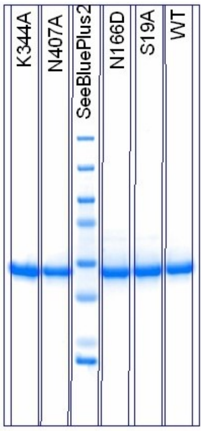
SR_2_total_score	0.071	0.002	-0.078	-0.079	-0.055	-0.134
SR_3_all_cst	-0.061	-0.039	-0.148	-0.114	-0.163	-0.119
SR_3_burunsat_pm	0.000	0.000	0.000	0.000	0.000	0.000
SR_3_fa_rep	0.195	0.180	0.193	0.309	0.155	0.042
SR_3_hbond_pm	0.000	0.000	0.000	0.000	0.000	0.000
SR_3_hbond_sc	0.000	0.000	0.000	0.000	0.000	0.000
SR_3_nlpstat_pm	0.142	0.087	0.263	0.317	0.007	0.055
SR_3_pstat_pm	0.164	0.096	0.050	0.064	-0.131	-0.069
SR_3_total_score	-0.078	0.093	-0.092	-0.050	-0.081	0.024
SR_4_all_cst	0.000	0.000	0.000	0.000	0.000	0.000
SR_4_burunsat_pm	0.123	0.135	0.089	0.078	0.067	0.138
SR_4_fa_rep	-0.118	-0.024	-0.235	-0.212	-0.153	-0.103
SR_4_hbond_pm	-0.197	-0.206	0.046	0.069	-0.211	-0.182
SR_4_hbond_sc	0.299	0.092	-0.009	-0.063	0.058	0.061
SR_4_nlpstat_pm	0.067	0.174	0.160	0.244	-0.007	-0.015
SR_4_pstat_pm	0.040	0.098	0.105	0.152	-0.025	-0.072
SR_4_total_score	0.096	0.045	0.193	-0.121	-0.181	-0.076
SR_5_all_cst	0.128	0.074	-0.091	-0.048	-0.062	-0.082
SR_5_burunsat_pm	0.055	0.052	0.061	-0.031	0.136	0.099
SR_5_dsasa_1_2	0.235	0.189	0.037	0.147	0.215	0.201
SR_5_fa_rep	-0.012	-0.064	0.165	-0.019	0.058	-0.030
SR_5_hbond_pm	0.514	0.410	0.267	0.334	0.267	0.146

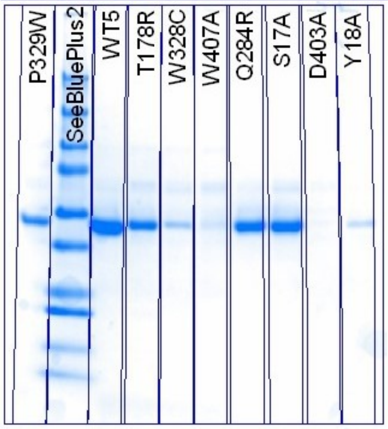
SR_5_hbond_sc	-0.524	-0.447	-0.266	-0.341	-0.273	-0.194
SR_5_interf_E_1_2	-0.461	-0.469	-0.237	-0.231	-0.266	-0.278
SR_5_total_score	-0.462	-0.471	-0.237	-0.230	-0.267	-0.279
tot_burunsat_pm	-0.078	-0.101	-0.036	-0.066	0.154	0.056
tot_hbond_pm	0.419	0.396	0.291	0.380	0.315	0.323
tot_NLconts_pm	0.570	0.548	0.246	0.238	0.432	0.421
tot_nlpstat_pm	0.334	0.313	0.223	0.164	0.277	0.277
tot_nlsurfaceE_pm	-0.285	-0.284	-0.193	-0.276	-0.176	-0.111
tot_pstat_pm	0.306	0.209	0.118	-0.009	0.221	0.161
tot_seq_recovery	0.000	0.000	0.000	0.000	0.000	0.000
tot_total_charge	0.051	0.053	0.199	0.153	0.103	0.123
tot_total_neg_charges	-0.068	-0.063	-0.148	-0.090	-0.044	-0.034
tot_total_pos_charges	0.010	0.014	0.177	0.209	0.133	0.129
total_score	-0.340	-0.396	-0.059	-0.014	-0.320	-0.362

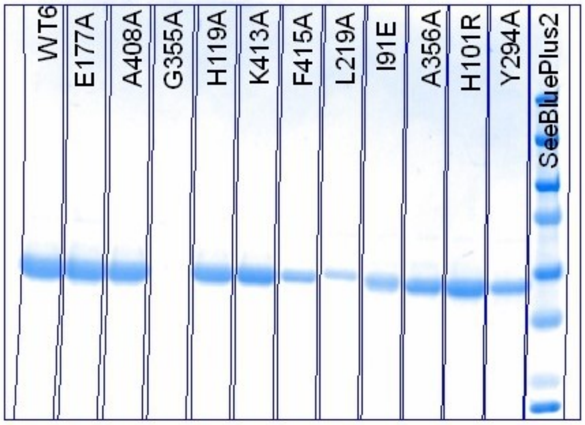
Supplemental Table 2. Correlations between individual structural features and each of k_{cat}, K_M, and k_{cat}/K_M. PCC and SRC values for each individual structural feature, given by Rosetta short name. For explanation of each short name, see main text and Richter.[1]

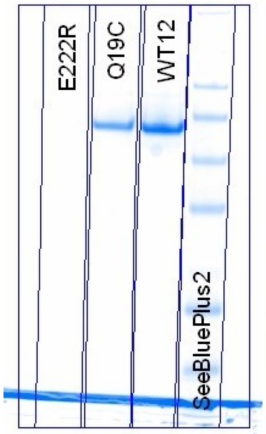

WT
H315Y
R76A
H315A
E177F
W409E
SeeBluePlus2

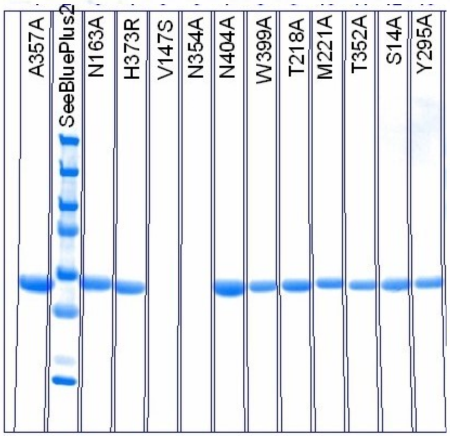

SeeBluePlus2
F405A
WT3
WT2
WT4
WT1
E154D
H379R
E406A

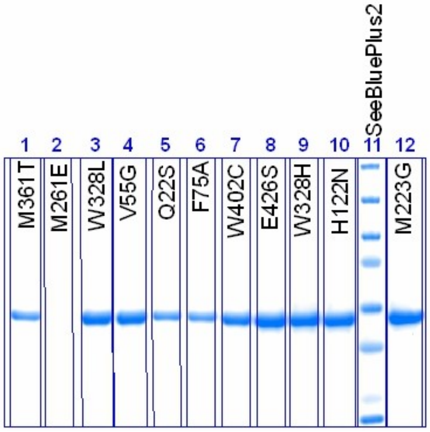

K344A
N407A
SeeBluePlus2
N166D
S19A
WT

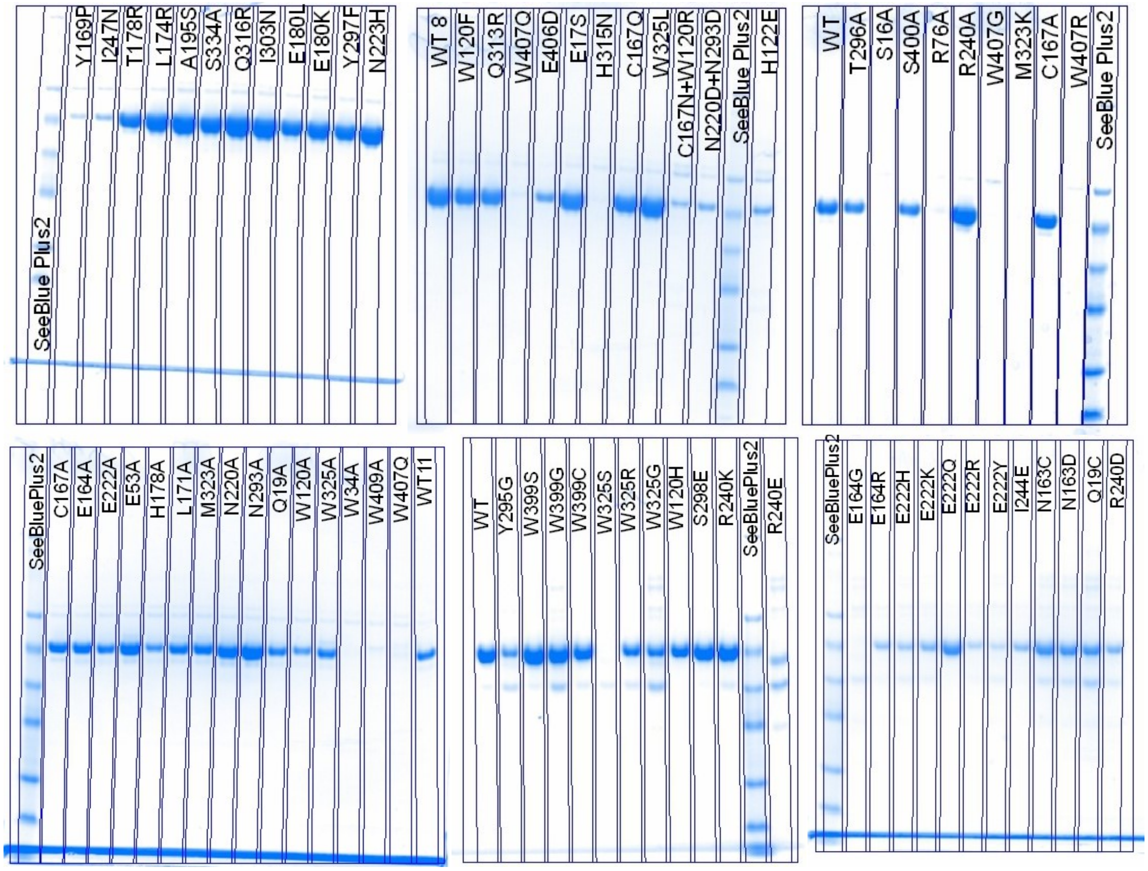

P329W
SeeBluePlus2
WT5
T178R
W328C
W407A
Q284R
S17A
D403A
Y18A


WT6
E177A
A408A
G355A
H119A
K413A
F415A
L219A
I91E
A356A
H101R
Y294A
SeeBluePlus2

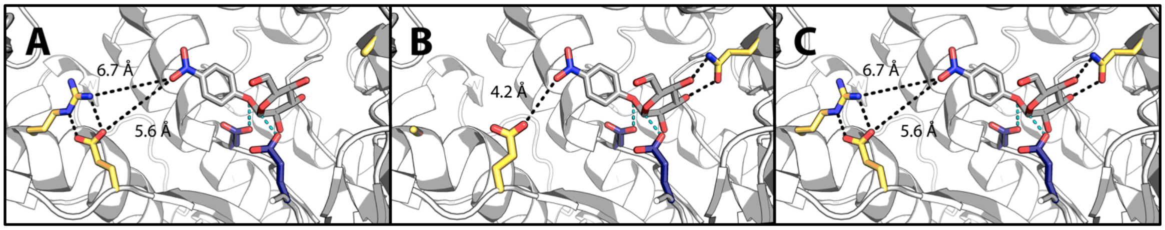

E222R
Q19C
WT12
SeeBluePlus2


A357A
SeeBluePlus2
N163A
H373R
V147S
N354A
N404A
W399A
T218A
M221A
T352A
S14A
Y295A

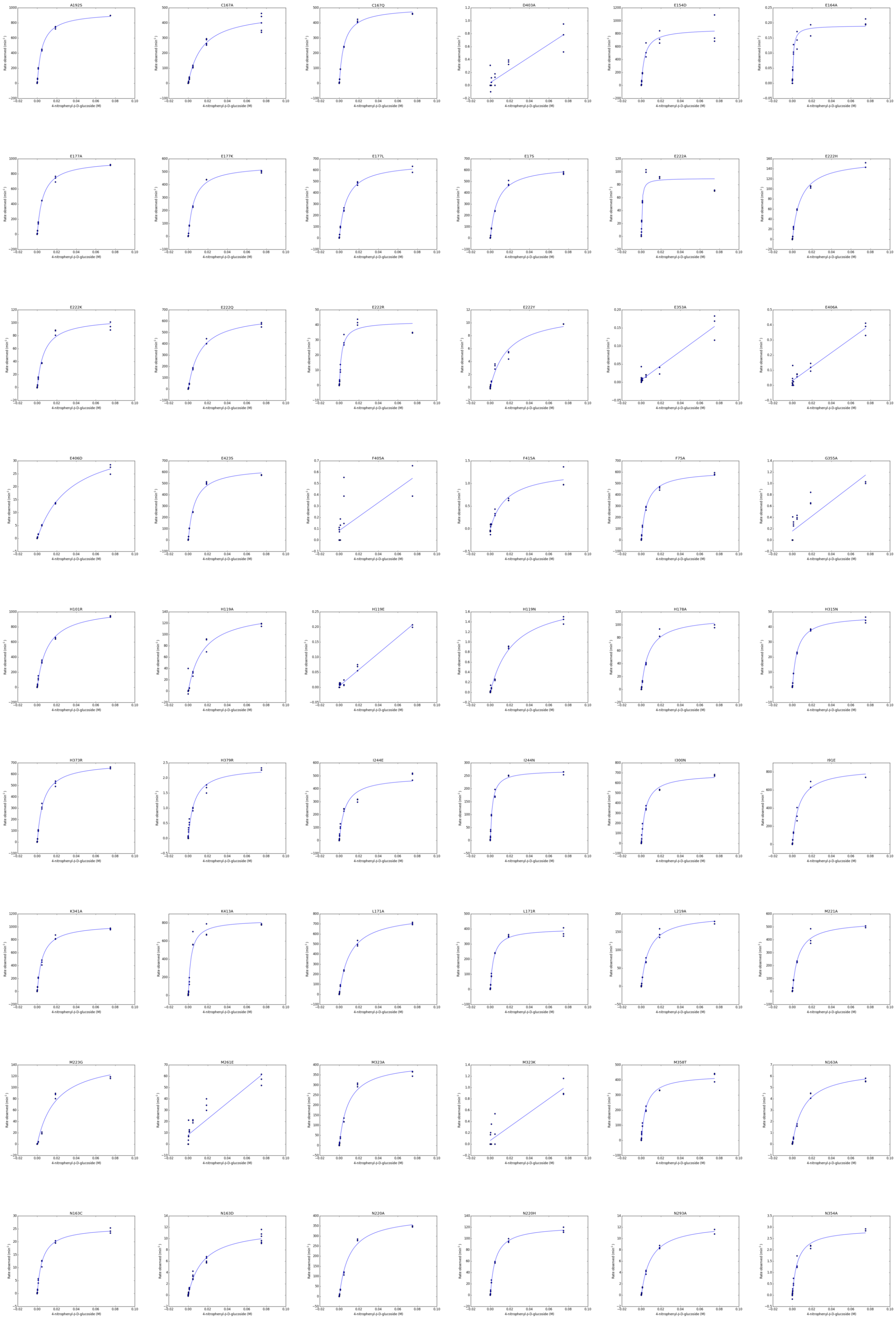

1 M361T
2 M261E
3 W328L
4 V55G
5 Q22S
6 F75A
7 W402C
8 E426S
9 W328H
10 H122N
11 SeeBluePlus2
12 M223G

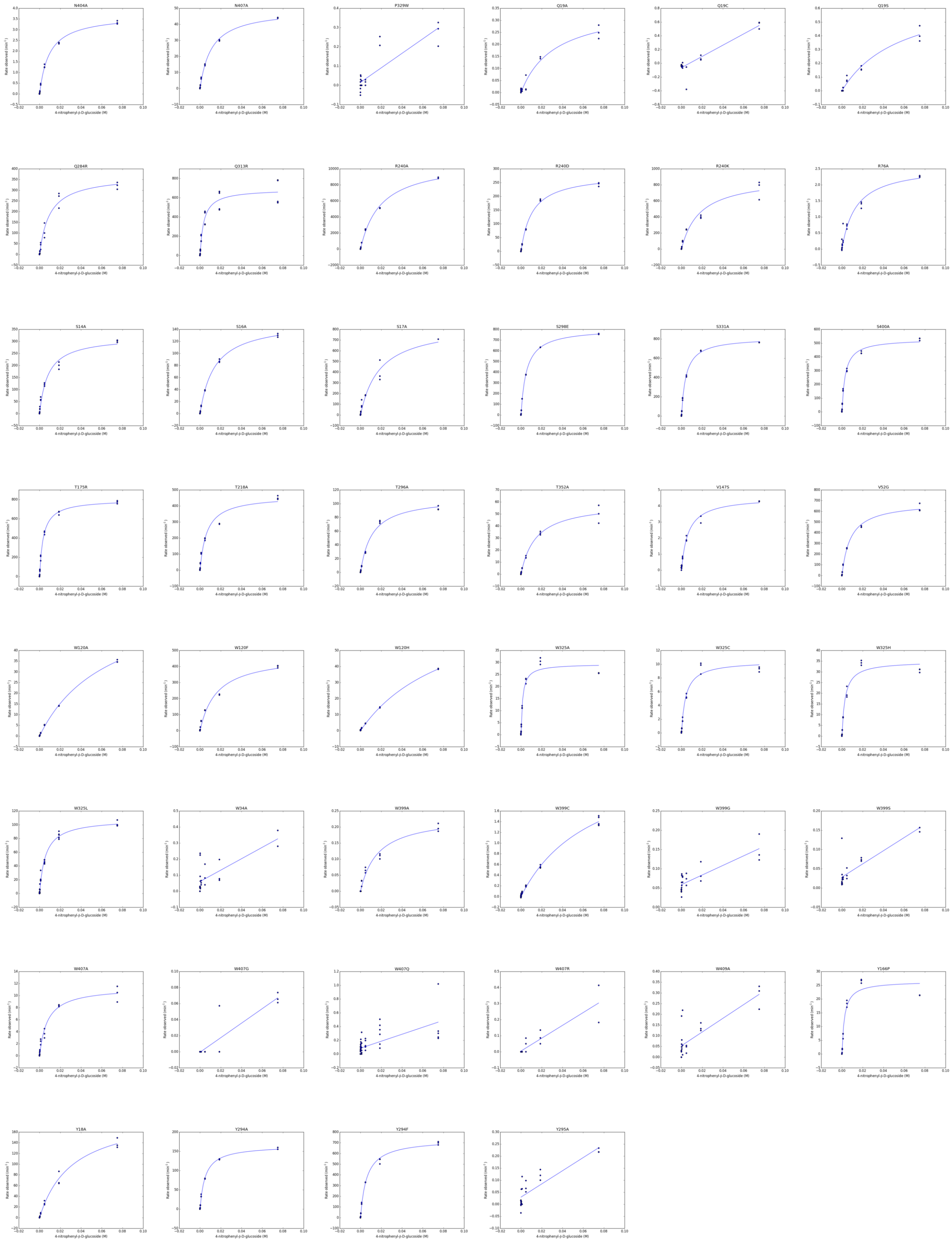


Supplemental Figure 1: SDS-PAGE images for 119 variants of BglB. SDS-PAGE gels showing all proteins used in this study, including replicates of wild type assayed with each batch of mutants. Gels were stained overnight with Coomassie Blue. Protein ladder used was SeeBlue® Plus2 Pre-stained Protein Standard (Life Technologies). Gels were imaged on a BioRad Gel Doc EZ system.

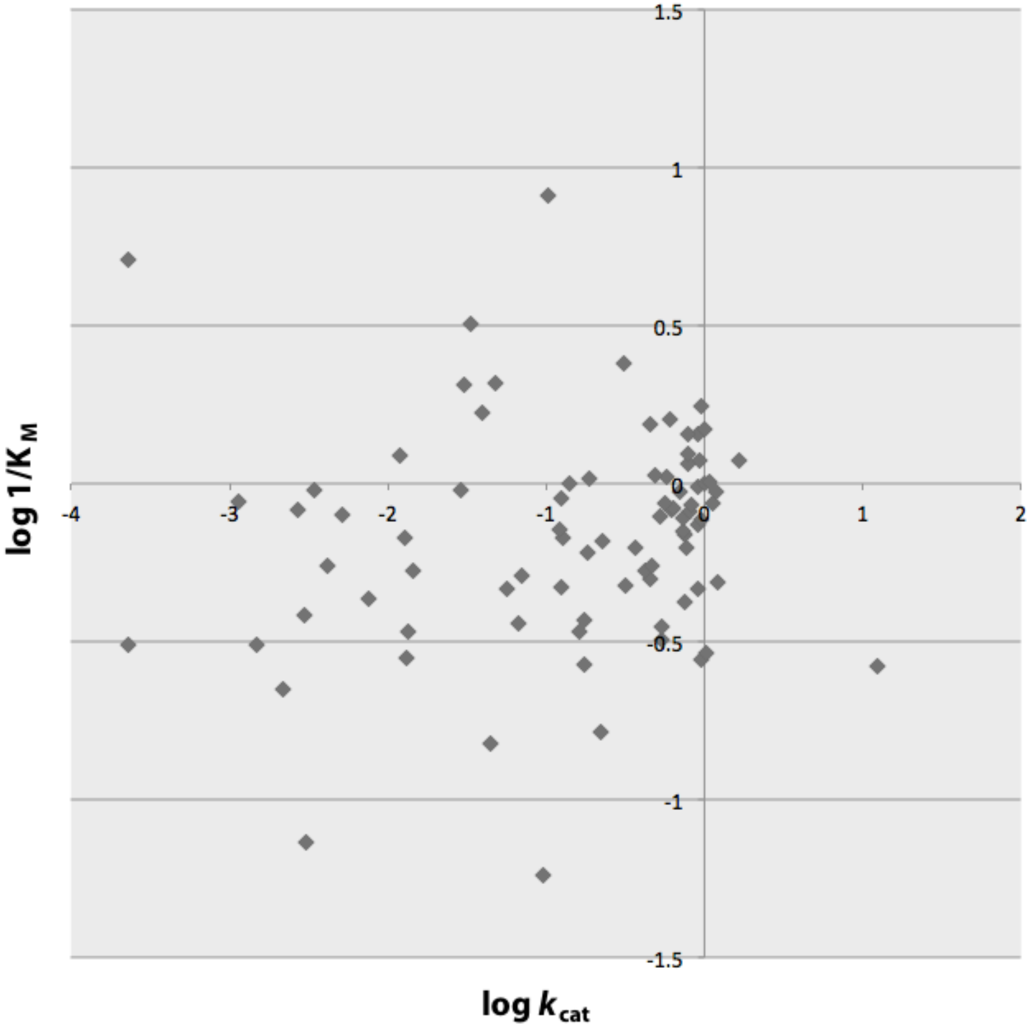


Supplementary Figure 2: Active site models of mutants Q19A, R240A, and wild type BglB. The lowest energy of 100 models for each mutant is depicted. In panel A, mutation of the glutamine at position 19 to an alanine removes two hydrogen bonds (black) to the substrate compared to wild type (C). In panel B, mutation of the arginine at position 240 to an alanine is predicted to stabilize an alternate conformation of E222A, bringing the carboxylate group to 4.2 Å of the substrate's nitro group. Distances and between the substrate, *p*-nitrophenyl-ß-D-glucoside, and the BglB molecule are indicated by black lines.


A192S
C167A
C167Q
D403A
E154D
E164A
E177A
E177K
E177L
E17S
E222A
E222H
E222K
E222Q
E222R
E222Y
E353A
E406A
E406D
E423S
F405A
F415A
F75A
G355A
H101R
H119A
H119E
H119N
H178A
H315N
H373R
H379R
I244E
I244N
I300N
I91E
K341A
K413A
L171A
L171R
L219A
M221A
M223G
M261E
M323A
M323K
M358T
N163A
N163C
N163D
N220A
N220H
N293A
N354A
Rate observed (min^{-1})
4-nitrophenyl-β-D-glucoside (M)


N404A
N407A
P329W
Q19A
Q19C
Q19S
Q284R
Q313R
R240A
R240D
R240K
R76A
S14A
S16A
S17A
S298E
S331A
S400A
T175R
T218A
T296A
T352A
V147S
V52G
W120A
W120F
W120H
W325A
W325C
W325H
W325L
W34A
W399A
W399C
W399G
W399S
W407A
W407G
W407Q
W407R
W409A
Y166P
Y18A
Y294A
Y294F
Y295A
Rate observed (min^{-1})
4-nitrophenyl-β-D-glucoside (M)

Supplemental Figure 3. Diagnostic plots showing Michaelis-Menten, Michaelis-Menten with substrate inhibition, or linear fit for each of 102 mutants. For each mutant, 8 observed rates (in triplicate) were fit to the Michaelis-Menten equation using SciPy[2] and plots were generated using Matplotlib.[3] Plots were used to visually confirm statistical analysis of the fits.



Supplemental Figure 4. Plot of the values of log k_{cat} versus log 1/K_M for 104 mutants relative to wild type BglB, showing the statistical independence of k_{cat} and K_M in the BglB system.

Supplemental materials and methods

Mutants were designed using the Foldit, a graphical user interface to the Rosetta Molecular Modeling Suite.[4] Mutants were chosen based on proximity to the active site as well as Foldit's predicted energy. Mutations within 12 Å of the active site, and those that did not increase the total system energy by more than 5 Rosetta energy units, were selected for experimental characterization. No other limitations were placed on designed mutations.

A sequence coding for BglB (Uniprot P22505) was codon-optimized for *Escherichia coli* and manufactured as a DNA String by Life Technologies. Using Gibson assembly, gene was inserted between the NdeI and XhoI sites of pET29b, adding a C-terminal His tag onto the protein sequence. Kunkel mutagenesis was used to create site-specific mutations, and all plasmids were sequence-verified.

For protein production, 20 μL of chemically-competent *Escherichia coli* BL21(DE3) (Novagen) were transformed on ice with 1 μL of plasmid in buffer at a concentration of 90 to 130 ng/μL. The competent cell-plasmid mixture was temperature shocked to induce plasmid uptake by heating at 42°C for one minute and then chilling on ice for one minute. Cells were recovered in 200 μL Terrific Broth (TB) media at 37 °C for one hour. They were then plated onto an LB agarose plate containing 50 mg/mL kanamycin, and incubated for 24 hours at 37 °C.

For each mutant, a 50 mL Falcon tube containing 5 mL TB with 50 mg/mL kanamycin was inoculated with one colony from a kanamycin selection plate. Tubes were covered with breathable seals and incubated with shaking for 24 hours at 37 °C.

Growth cultures were pelleted by centrifugation at 4700 RPM for 10 minutes and the supernatant was discarded. The cell pellet was resuspended in 5 mL of induction medium (TB with 1 mM isopropyl-β-D-thiogalactopyranoside and 50 mg/mL kanamycin). The tubes were covered again with breathable seals and incubated with shaking at 18 °C for 24 hours.

The 5 mL expression culture was pelleted by centrifugation at 4700 for 10 minutes and the supernatant was discarded. The resulting pellet was suspended in 500 μL wash buffer (50 mM HEPES, 150 mM sodium chloride, 15 mM imidazole, pH 7.50) and lysed with BugBuster protein extraction reagent (Millipore) and 1 mg lysozyme, 0.1 mg DNase, and 0.1 mg phenylmethylsulfonyl fluoride per sample.

After 20 min, lysate was centrifuged at 14,700 RPM for ten minutes. The supernatant was loaded on to protein purification columns (BioSpin 732-6008) prepared with 100 μL of 50% Ni-NTA resin slurry. After equilibration with 500 μL wash buffer, two 500 μL aliquots of supernatant were added to the columns. Six rounds of 500 μL wash buffer were then allowed to drip through the columns. Resulting protein micro-columns were then transferred to 2 mL tubes for elution. Protein was eluted in 2x100 μL elution buffer (50 mM HEPES, 150 mM sodium chloride, and 25 mM EDTA, pH 7.50). A brief centrifugation at 4000 RPM ensured all protein was collected.

Protein yield was then determined via ratio of absorbance at 260 and 280 nm and SDS-PAGE.

Each enzyme variant was assayed in triplicate at 8 substrate concentrations ranging from 0 to 75 mM. Diluted protein solution was dispensed in 25 μL aliquots into a 96-well plate (Corning Costar #3885). Separately, in another plate, 100 uL elution buffer with 8 different concentrations of pNPG (1 per row) were prepared. The assay was initiated by multi-channel pipetting 75 μL substrate from each row of the substrate plate into the corresponding row of the assay plate. The absorbance at 420 nm was monitored every minute for 60 minutes to determine the rate of the reaction.

Unless otherwise noted, all supplies were purchased from Sigma-Aldrich.

Prediction and feature selection via Elastic net

A regularized linear regression model, Elastic Net (EN), was chosen to fit the dataset of the kinetic constants, each constant fitted independently.[5] Comparing to ordinary least squares regression, an EN model is able to make a prediction and select the most informative feature set simultaneously as l_1 and l_2 penalties are applied to the regression weights. The weight of each structural feature is estimated as

$$\overline{\omega_0}, \overline{\omega_i} = \arg\min_{\omega_0,\omega_i} \sum_{i=1}^{p}(y_i - \omega_0 - \omega_i x_i)^2 + \lambda_1 \sum_{i=1}^{p} \omega_i^{\ 2} + \lambda_2 \sum_{i=1}^{p} |\omega_i|$$

Where:

ω_0: the intercept;

ω_i : the weight of structural feature i in the regression model;

p: the number of structural features generated by the BglB model;

y_i : the kinetic constant (the dependent variable to be predicted);

x_i : structural features generated by the BglB model (the independent variables);

λ_1 , λ_2 : parameters tuning the constraints on the weights.

Since the structural feature were measured in different ranges and units, we first normalized all the features to be zero-centered with variance being one by subtracting the mean and dividing by the variance of the feature value. All the features are on the same scale to compare their contribution to the kinetic constants after the normalization. The tuning parameters λ_1 , λ_2 are determined one by one via stratified 10-fold cross validation by searching a grid of λ_1 and λ_2 . Each round of cross validation generated a linear regression model. In order to build a more generalized model, cross validation was run 1,000 times with a different part of the dataset each time. The final prediction of a mutant's kinetic constant was an average of all the predictions during the 1,000 rounds of training. The average number of non-zero weights when predicting $k_{\mathrm{cat}}/\mathrm{K}_\mathrm{M}$, k_{cat} and K_M were 9, 8 and 10 respectively. The top features were chosen and listed in table 1 with their averaged weights among all the models (9 for $k_{\mathrm{cat}}/\mathrm{K}_\mathrm{M}$, 8 for k_{cat}, 10 for K_M)

Stratified 10-fold cross validation

Stratified 10-fold cross validation was implemented to validate the EN model.[6] Specifically, all the mutants were first ranked according to the experimentally-measured value of the kinetic constant to be predicted and every 10 adjacent datapoints were randomly marked with an index using integers from 1 to 10 without duplication. Finally, all the datapoints with the same index were grouped together, resulting in ten folds. Since the datapoints in each folds comes from different level of the dataset, this guarantees every fold is a good representative of the dataset. In order to build a robust prediction model, the cross validation was run 1,000 times, the dataset split into training set and testing set differently each time.

Supplemental references

1. Richter, F.; Leaver-Fay, A.; Khare, S. D.; Bjelic, S.; Baker, D., De novo enzyme design using Rosetta3. *PLoS One* **2011,** *6* (5), e19230.

2. Oliphant, T. E., Python for scientific computing. *Computing in Science & Engineering* **2007,** *9* (3), 10-20.

3. Hunter, J. D., Matplotlib: A 2D graphics environment. *Computing in science and engineering* **2007,** *9* (3), 90-95.

4. Siegel, J. B.; Smith, A. L.; Poust, S.; Wargacki, A. J.; Bar-Even, A.; Louw, C.; Shen, B. W.; Eiben, C. B.; Tran, H. M.; Noor, E.; Gallaher, J. L.; Bale, J.; Yoshikuni, Y.; Gelb, M. H.; Keasling, J. D.; Stoddard, B. L.; Lidstrom, M. E.; Baker, D., Computational protein design enables a novel one-carbon assimilation pathway. In *PNAS*, National Acad Sciences: 2015; Vol. 112, pp 3704-3709.

5. Zou, H.; Hastie, T., Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **2005,** *67* (2), 301-320.

6. Kohavi, R. In *A study of cross-validation and bootstrap for accuracy estimation and model selection*, Ijcai, 1995; pp 1137-1145.