Supporting information for

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

Dylan Alexander Carlin1\*, Ryan W. Caster2\*, Xiaokang Wang3, Stephanie A. Betzenderfer2, Claire X. Chen2, Veasna M. Duong2, Carolina V. Ryklansky2, Alp Alpekin2, Nathan Beaumont2, Harshul Kapoor2, Nicole Kim2, Hosna Mohabbot2, Boyu Pang2, Rachel Teel2, Lillian Whithaus2, Ilias Tagkopoulos2,6, Justin B. Siegel2,3,4

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**

Supplemental Table 1: List of mutants

Supplemental Figure 1: Gel images

Supplemental Figure 2: Q19A and R240A

Supplemental Table 2: PCC and SRC values

Supplemental Figure 3: Diagnostic plots for all mutants

Supplemental Figure 4: *k*cat vs. KM plot

Supplemental Table 3: Conservation analysis

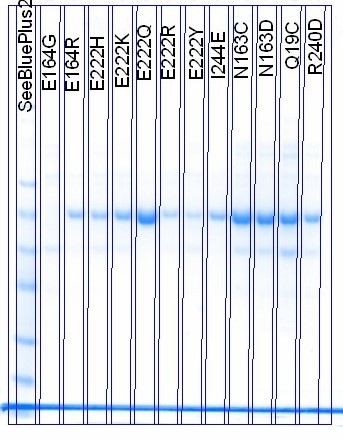
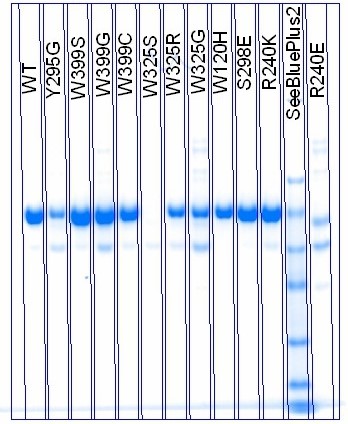
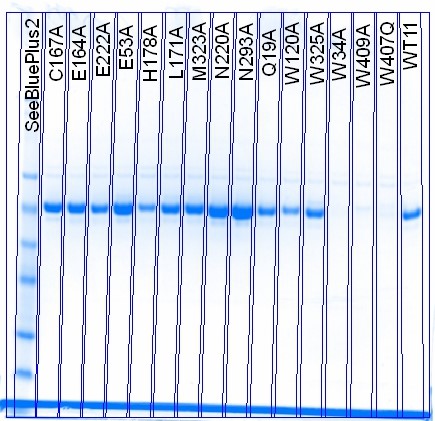
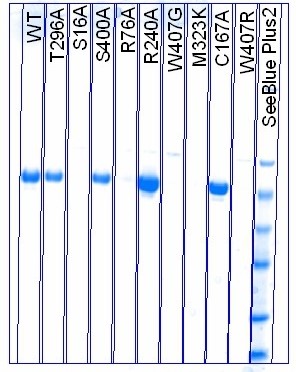
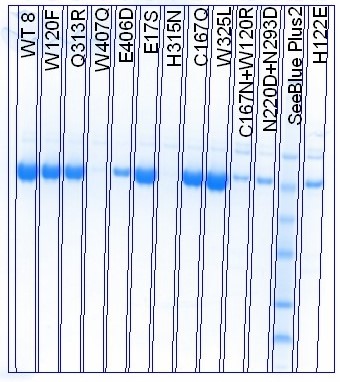
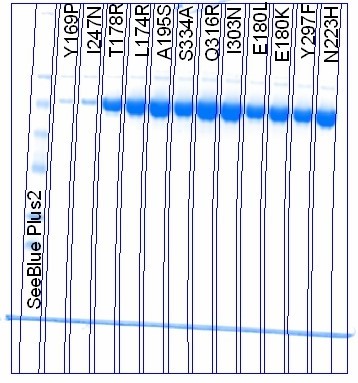
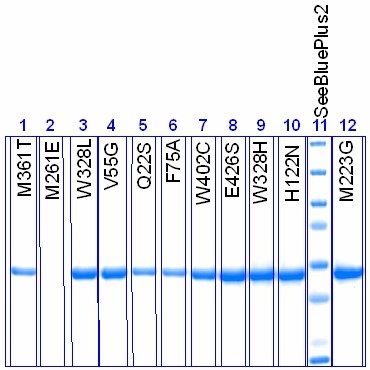
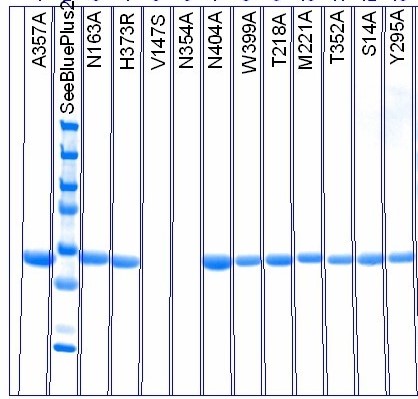
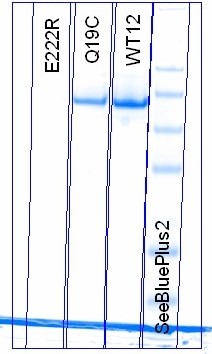
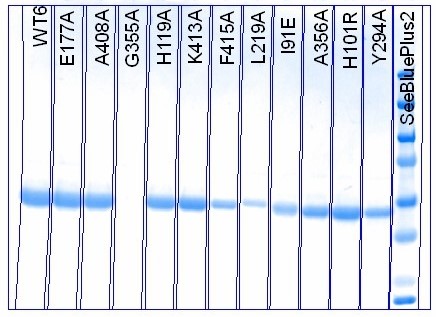
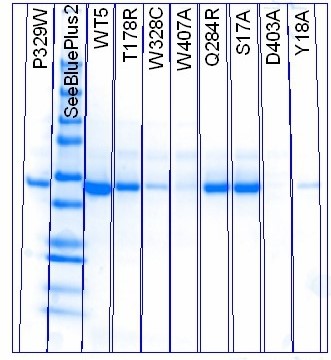
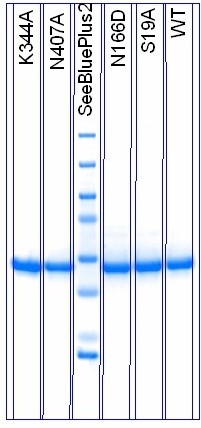
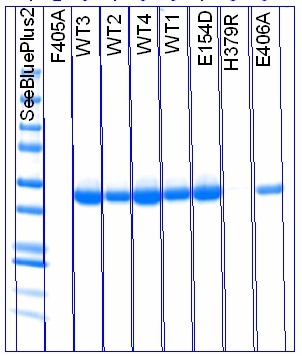
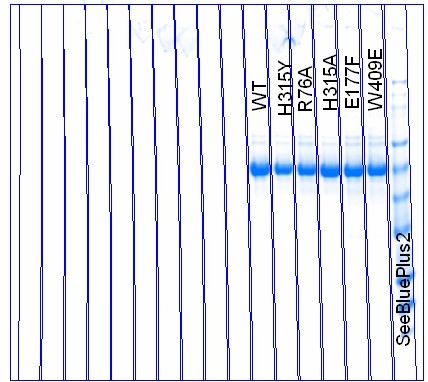
Supplemental materials and methods

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Mutant** | **Protein yield** | **SDS-PAGE** | **KM** | ***k*cat** | **KI** | ***k*cat/KM** |
|  | **mg/mL** |  | **mM** | **min-1** | **mM** | **M-1min-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 |
| S17E | 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 |
| F72A | 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 |
| 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 |
| M223G | 0.88 |  | 19.21 ± 2.91 | 154 ± 9 |  | 7,998 ± 1,302 |
| 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 |
| 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 |

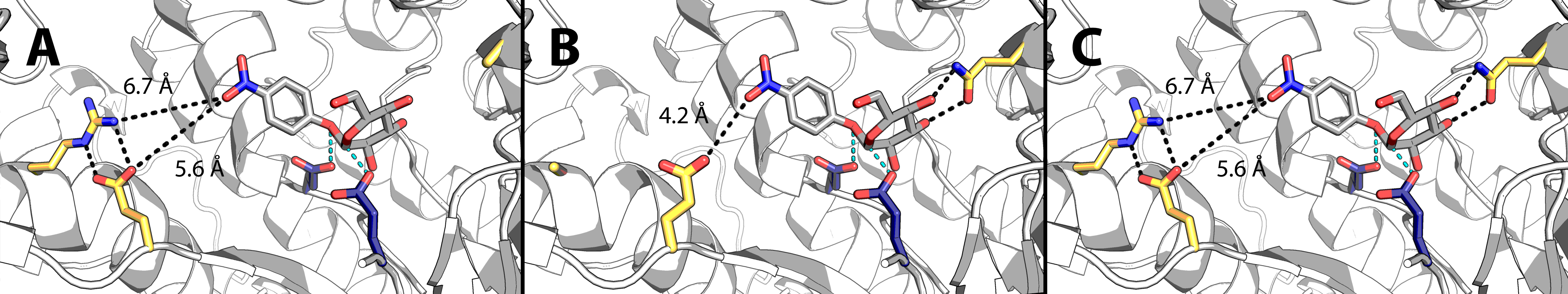
**Supplemental Table 1: Kinetic constants for 104 computationally-designed BglB mutants.** Included are columns (1) the mutation (2) protein yield as assessed by ratio of aborbance at 260 and 280 nm (3) protein yield as assessed by SDS-PAGE (4, 5, 6, 7) kinetic constants and nonlinear regression analysis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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Feature** | **PCC (Kcat/KM）** | **SRC (Kcat/KM）** | **PCC (1/KM)** | **SRC (1/KM)** | **PCC (Kcat)** | **SRC (Kcat)** |
| 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 |

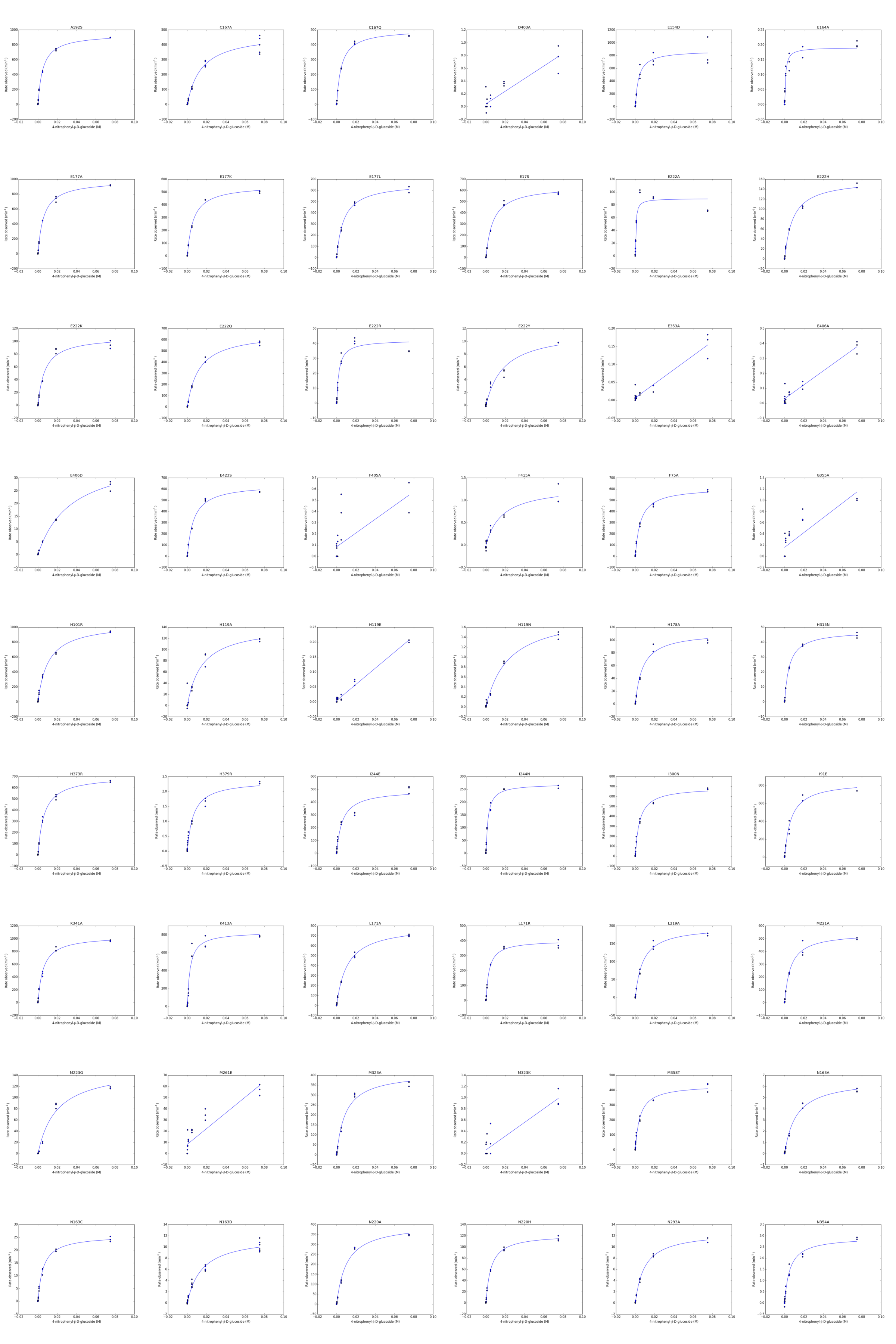
**Supplemental Table 2. Correlations between individual structural features and each of *k*cat, KM, and *k*cat/KM.** PCC and SRC values for each individual structural feature, given by Rosetta short name. For explanation of each short name, see main text and Richter.1

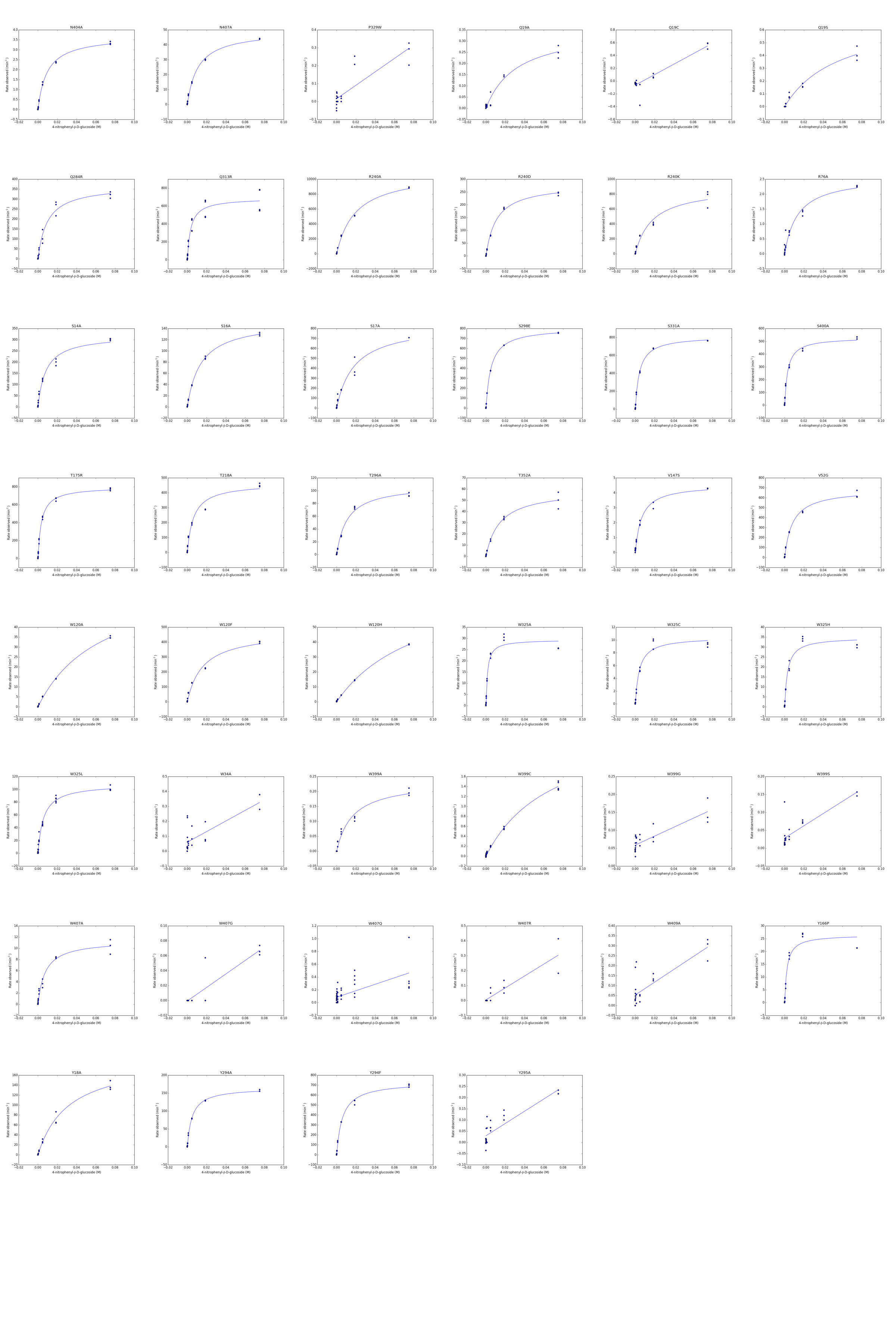


**Supplemental Figure 1: SDS-PAGE images for 119 variants of BglB.** SDS-PAGE gels showing all proteins used in this study, including replicates of wild type assayed with each batch of mutants. Gels were stained overnight with Coomassie Blue. Protein ladder used was SeeBlue® Plus2 Pre-stained Protein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

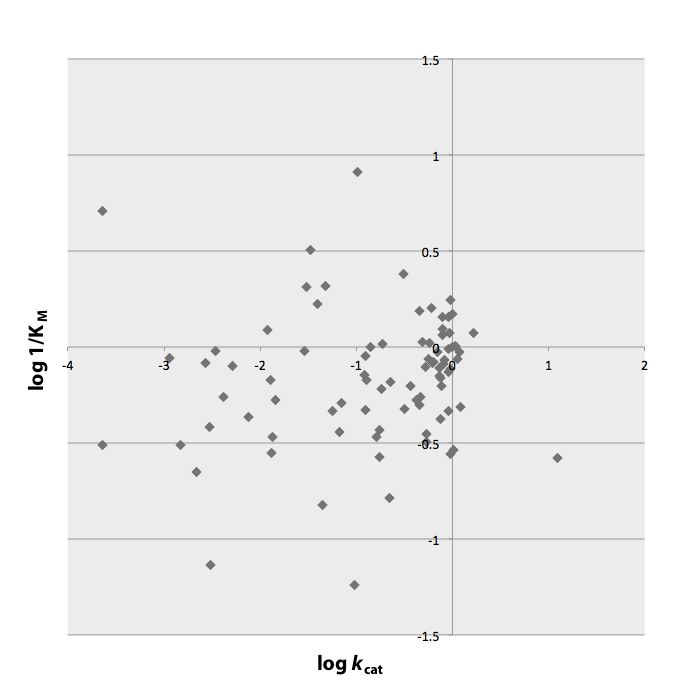
****

**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.





**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 SciPy2 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/KM for 104 mutants relative to wild type BglB, showing the statistical independence of *k*cat and KM in the BglB system.

|  |  |  |
| --- | --- | --- |
| ***k*cat/KM** | **Mutant** | **Conservation (%)** |
| 0.52 | R240A | 13 |
| 0.22 | K413A | 27 |
| 0.05 | K341A | 16 |
| 0.04 | S331A | 60 |
| -0.02 | S400A | 41 |
| -0.02 | E177A | 3 |
| -0.08 | E222A | 7 |
| -0.19 | F72A | 44 |
| -0.29 | M221A | 3 |
| -0.37 | L171A | 11 |
| -0.38 | T218A | 17 |
| -0.57 | S17A | 52 |
| -0.59 | M323A | 10 |
| -0.64 | N220A | 19 |
| -0.65 | S14A | 61 |
| -0.71 | Y294A | 49 |
| -0.72 | C167A | 11 |
| -0.83 | L219A | 22 |
| -0.97 | W325A | 70 |
| -1.07 | H178A | 7 |
| -1.19 | S16A | 48 |
| -1.24 | T296A | 14 |
| -1.26 | H119A | 96 |
| -1.44 | Y18A | 25 |
| -1.62 | T352A | 40 |
| -2.06 | W407A | 44 |
| -2.12 | N293A | 50 |
| -2.48 | N163A | 96 |
| -2.49 | N354A | 86 |
| -2.64 | N404A | 34 |
| -2.94 | E164A | 100 |
| -2.95 | R76A | 96 |
| -3.01 | G355A | 96 |
| -3.34 | F415A | 47 |
| -4.16 | W399A | 96 |
| -4.54 | E406A | 47 |
| -4.76 | Q19A | 95 |
| -4.93 | E353A | 100 |
| ND | Y295A | 100 |
| ND | D403A | 98 |
| ND | W34A | 40 |
| ND | W120A | 34 |
| ND | F405A | 26 |
| ND | W409A | 10 |
| 0.00 | BglB | NA |

**Supplemental Table 3. Conservation analysis of BglB active site residues.** A multiple sequence alignment of 1,554 family 1 glycoside hydrolases from the Pfam database aligned to the BglB wild type sequence was used for this analysis. Column 1 is the relative *k*cat/KM compared to wild type on a log scale. Column two gives the position and native BglB residue at that position. Column three is the percentage of the 1,554 aligned sequences that have the same residue as BglB.

**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  and  penalties are applied to the regression weights. The weight of each structural feature is estimated as



Where:

: the intercept;

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

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

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

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

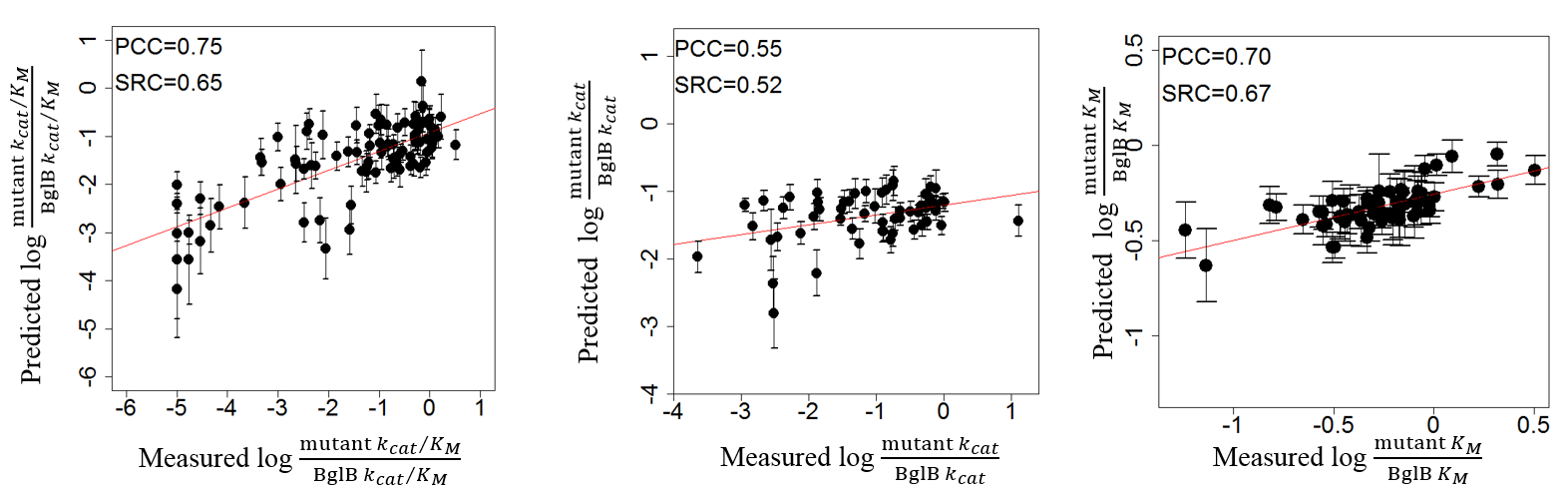
,: parameters tuning the constraints on the weights.

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,  are determined one by one via stratified 10-fold cross validation by searching a grid of and. 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*cat/KM, *k*cat and KM 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*cat/KM, 8 for *k*cat, 10 for KM)

**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.

Each bootstrap set is constructed by uniformly sampling from the original database, with a coverage of 87%, which is calculated as . The following figure shows the prediction of the left-out subsets during bootstrapping.



**Supplemental Figure 5.** Prediction of the left-out subsets during bootstrapping.

**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.