Detailed methods have been published in two papers: Wei, Xu, and Dunbrack, Proteins, 81:199-213; and Wei and Dunbrack, PLoS ONE 8(7): e67863. doi:10.1371/journal.pone.0067863

To compile a human mutation data set, we downloaded data on mutations from the SwissVar database (release 57.8 of 22–Sep- 2009). After removing unclassified variants, variants in very long proteins to reduce computation time (sequences of more than 2000 amino acids), redundant variants, and variants that are not accessible by single-site nucleotide substitutions (just 150 mutation types are accessible by single-site nucleotide change), we compiled those human disease mutations listed as disease-associated as the deleterious mutations set.

Non-human primate sequences were obtained from UniprotKB. We used PSI-BLAST to identify likely primate orthologues of human proteins in the SwissVar data sets using a sequence identity cutoff of 90% between the human and primate sequences. Mutations without insertions or deletions within 10 amino acids on either side of the mutation of amino acid differences in the PSI-BLAST alignments were compiled into a data set of human/primate sequence differences to be used as neutral training data. We mapped the human mutation sites in the deleterious and neutral data sets to known structures of human proteins in the PDB using SIFTS from the EBI.

We trained an SVM with 10-fold cross-validation on several features: (1) the difference in PSSM scores for the wildtype and mutant, where the PSSM was obtained from a two-round PSI-BLAST search on Uniref90; (2) a conservation score from AL2CO; (3) the relative surface accessibility of the wildtype amino acid in a single chain of the protein of interest; (4) the lowest relative surface accessibility of the wildtype amino acid over available biological assemblies of the protein (curated, see below); (5) three disorder predictor scores (although this is not relevant for this challenge).

To obtain the AL2CO score, we parsed the PSI-BLAST output to select homologues with sequence identity greater than 20% for each human and primate protein. We used BLASTCLUST to cluster the homologues of each query using a threshold of 35%, so that the sequences in each cluster were all homologous to each other had a sequence identity >35%. A multiple sequence alignment of the sequences in the cluster containing the query was created with the program Muscle. Finally, the multiple sequence alignment was input to the program AL2CO to calculate the conservation score for human and primate proteins.

For each protein in our human and primate data sets whose (human) structure was available in the PDB according to SIFTS, we obtained the symmetry operators for creating the biological assemblies from the PISA website and applied these symmetry operators to create coordinates for their predicted biological assemblies. We used the program Naccess to calculate surface area for each wildtype position in the biological assemblies as well as in the monomer chains containing the mutation site (i.e., from coordinate files containing only a single protein with no biological assembly partners or ligands). For the human mutation position, if the amino acid can be presented in the coordinates of more than one associated structures, we calculated the surface area for those associated structures and get the minimal surface area as the surface area of that human mutation.

The SVM was trained on balanced data sets of deleterious and neutral mutations, since we found that this provides the highest balanced accuracy for phenotype predictions as described in the PLOS ONE paper.

We used all available structures of human UBE2I except PDB entries 2O25, which contains an association with another E2 enzyme of unknown biological significance (and no published paper), and PDB: 2XWU, which is a structure of UBE2I bound to importin. Presumably the nuclear localization sequence may dominate the interaction with importin and surface mutations may not have too much of an effect on nuclear import. The structures included PDB entries 1A3S, 2PE6, 2UYZ, 2VRR, 4Y1L, 2GRN, 1KPS, 1Z5S, 3UIN, 3UIO, 3UIP. These include the non-covalent SUMO binding site which may be relevant for multiple sumoylation (2PE6, 2UYZ, etc.) and binding of RANGAP1 with (PDB: 1Z5S, 3UIN, 3UIO, 3UIP) and without SUMO covalently linked to RANGAP1 (PDB: 2GRN, 1KPS), and a sumoylated UBE2I at Lys14 (PDB: 2VRR) (a structure which also includes the non-covalently bound SUMO, although this was not noted by the authors). We included PDB: 4Y1L which has a second UBE2I as a substrate of UBE2I ready to be sumoylated on Lys14, and a RWD protein which may modify UBE2I’s activity, although its exact function is not known.

To get growth scores, we (roughly) fitted the numerically sorted output scores from our program to the experimental data provided by CAGI with a function Growth = 1.5 \* (1- score\*score – min(1 - score \* score) + 0.0001). The SVM outputs a standard deviation and we used the same formula with (score+std) to calculate the std for the Growth. For subset 3, we picked up the geometry mean of (1 - score) (1 is deleterious and 0 is neutral in our scores) and the biggest standard deviation for the mutations. The geometric mean is calculated by calculating the product of (1 – score(i)) for the mutations for each construct, then taking the nth root. In order to use the similar formula for subset 1 and subset 2, we transformed the geometry means back to probability by subtract from 1, that is 1 – geometric mean for the combined mutations. The predicted growth score is calculated by Growth = 1.5 \* (1- score\*score – min(1 - score \* score) + 0.0001) – 0.25. The distribution of the predicted growth scores for subset 3 is shifted from 0 to 0.25. So we subtracted 0.25. For the final predicted scores, we set the minus scores to be 0.0001, since those scores are predicted to be deleterious with high probabilities.

We manually set some of C93\* mutations to be 0.0001, and added comments. Cys93 is very conserved since it's the catalytic thioester bond.