**Frontiers Datasets**

The Frontiers Datasets used for the statistical test is a subset of the Protherm Database in which the biases in the representation of different types and classes of mutations have been accounted for. In total 81 individual structures were considered, 3 of them where structures in complex with another subunit, for a total of 1510 single-point mutants**.** Before running the calculation of the FoldX ΔΔGs on the single-point mutants, the structures listed in the dataset were stripped down to just the chain involved in the particular mutations. These one-chain structures were then repaired with the *RepairPDB* function of the FoldX method. The repairing process consisted in: accounting for the incorrect rotamer assignment of all Asn, Gln and His in the protein structure, eliminating small VanderWaals’ clashes through sidechain optimization and finding new energy minima for residues with bad original energy by exploring different rotamer combinations. The single-point mutants in the dataset were then applied individually to their corresponding repaired structure and their foldx ΔΔG value recorded. The same procedure has been followed for the reverse mutations with the exception that their FoldX ΔΔG values were calculated after applying the reverse single-point mutant on the protein variant bearing the correspondent “forward mutation”, as modelled by FoldX itself.

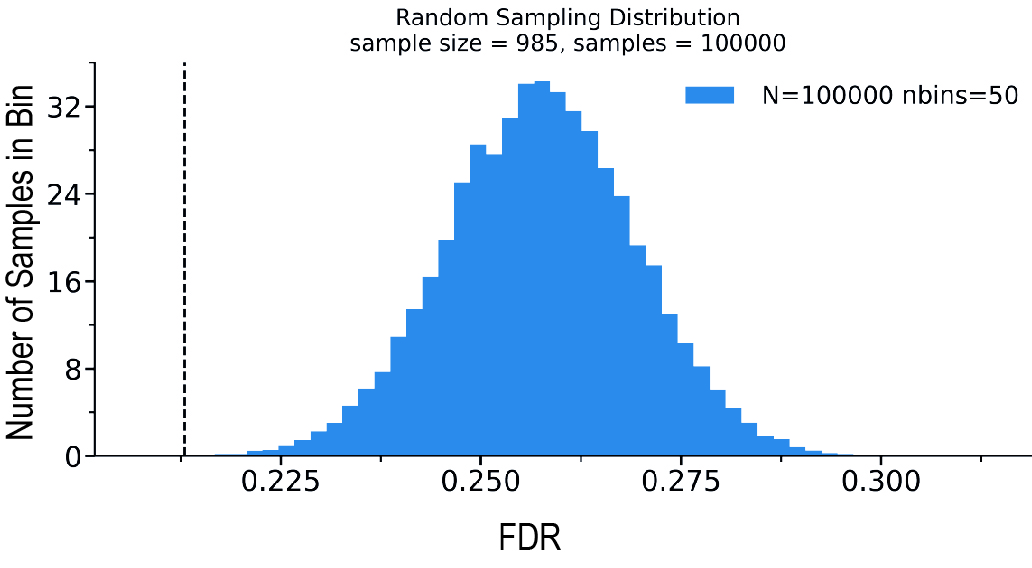


Figure S1 Insight of the sampling test performed for each filtering case on each dataset. The graph shows the distribution of the 100.000 calculated FDR values of the random subsets taken from the Frontiers dataset in a specific filtering case. The bell-shaped curve evenly distributes around the mean value corresponding to the FDR calculated from the dataset without applying any filters. The dashed black line pin points the FDR originally computed in the filtering case under scrutiny.

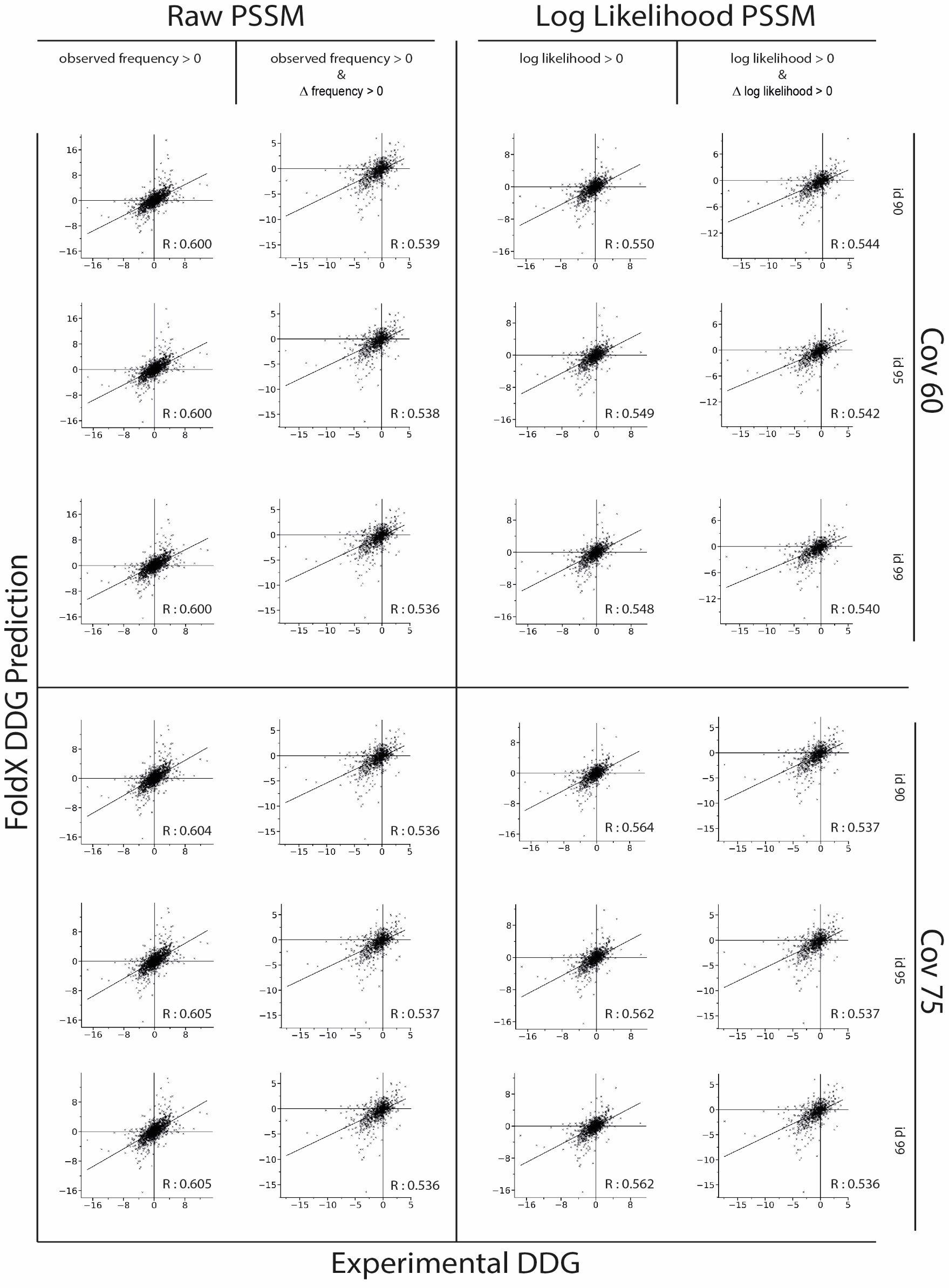


Figure S2 Complete rundown of all the comparisons made between predicted and experimentally obtained DDG values. DDG stands for the difference in free energy between the WT and the mutated protein. On the x coordinate of all plots, we have the FoldX-obtained DDG predictions for the mutants in the Frontiers dataset. On the x coordinate we have the experimental DDG values for the same mutants. The correlations have been calculated after applying several filtering conditions to the dataset. On the top of the figure is defined which PSSM has been used for the filtering and along with the type of PSSM an additional distinction is made as far as specific applied filtering conditions is concerned. For both PSSM we initially filtered out from the dataset only those mutations that according to the PSSM had a observed frequency/log likelihood < 0 and then we additionally filtered out also those mutations with a Δfrequency/Δlog likelihood < 0. Δfrequency/Δlog likelihood refers to the difference between the PSSM frequencies of a specific mutant and its WT counterpart. If Δfrequency/Δlog likelihood is higher than 0 it means that the mutant is more conserved in the alignment than the WT aminoacid. On the right side of the figure, we see another distinction. It refers to the convergence parameter used for a specific alignment. In this test we calculated alignments with two different convergence rates: 60 and 75. For both rates we than obtained alignments with three different identity rates: 90, 95 and 99. Each plot shows the correlation between experimental and predicted DDG values, under the corresponded filtering conditions, along with their R Pearson correlation coefficient.