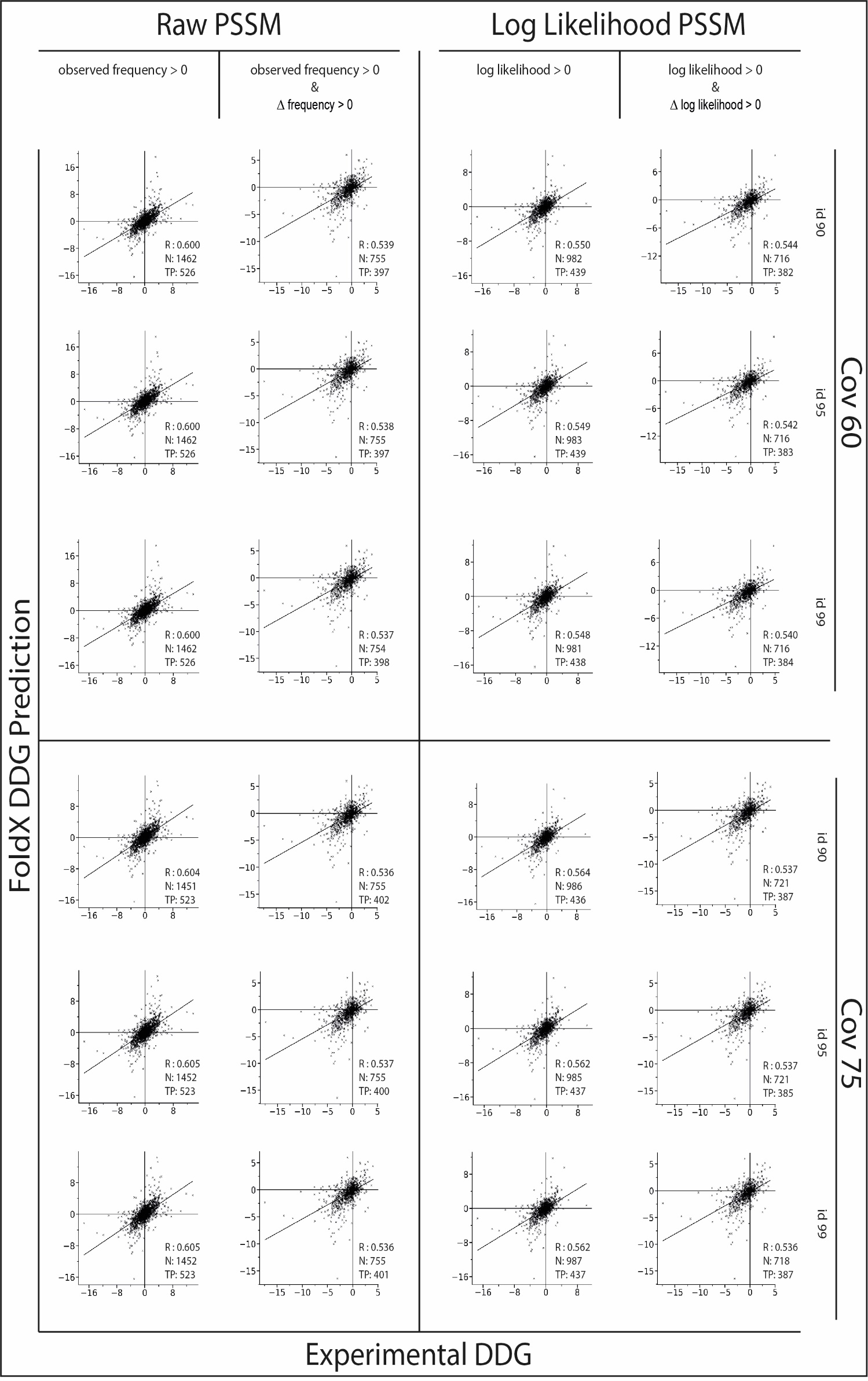


Figure S1 Insight of the sampling test performed for each filtering case on each dataset. The graph shows the distribution of the 100.000 calculated FDR values of the random subsets taken from the Frontiers dataset in a specific filtering case. The bell-shaped curve evenly distributes around the mean 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.



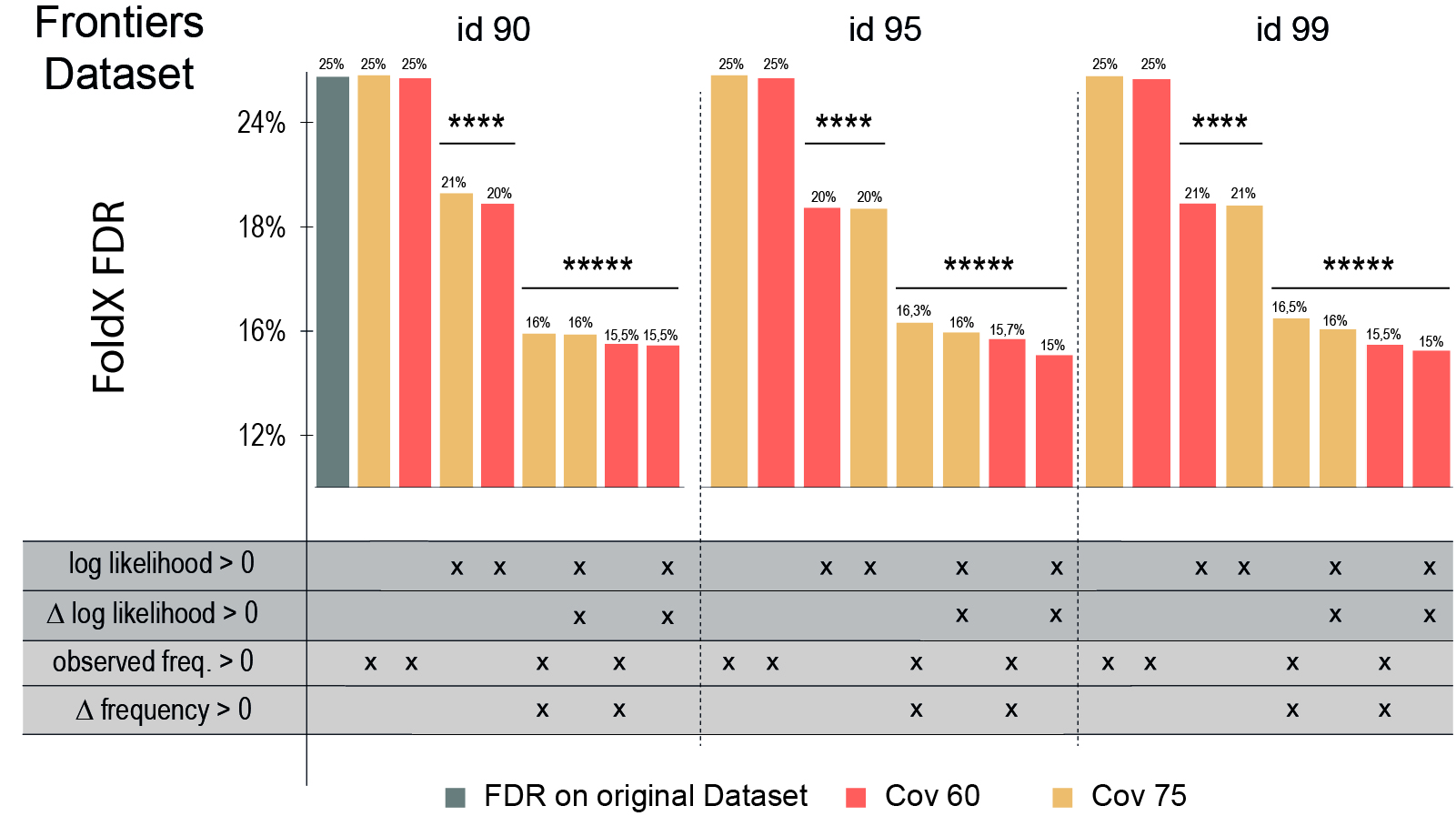


Figure S3Summary of statistical test results for all three identity rates (id) used in the alignment. The originally calculated FDR for the Frontiers Dataset is represented by the grey bar and plotted against the FDRs obtained in each filtering case. The statistically significant results are marked by asterisks, which in turn define the magnitude of the corresponding p-value. Below the graph, a table shows which filters have been applied for obtaining every FDR. The rows describe the different filtering conditions and the “x” signs mark which have been used in each testing case. Highlighted in yellow are the FDRs obtained with an alignment with coverage (Cov) 75 and in red are the ones obtained with coverage 60. Above each bar is written the precise FDR calculated for the respective filtering case.

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 (Cov): 60 and 75. For both rates we than obtained alignments with three different identity rates (id): 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. Moreover, the number of mutations in the set of mutation that passed the specific filter (N) and the number of the true positive predictions (TP) in each case is shown.

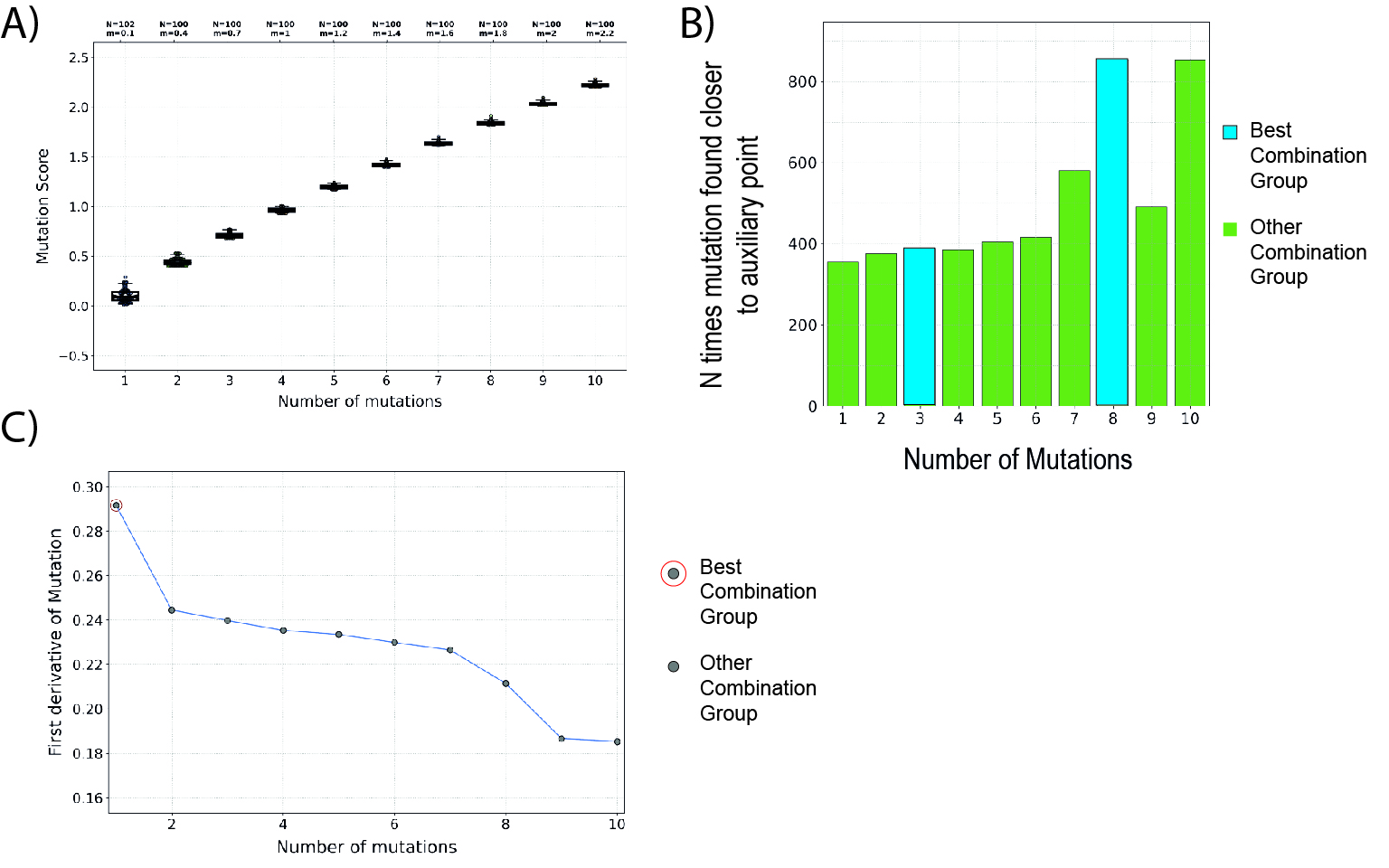


Figure S4 Example of the improvement of the developability potential and of the the Best Combination Group analysis for pdb 1bli (Amylase). The graphs are obtained from running the pipeline with the already mentioned pdb structure, a PSSM obtained from an alignment with coverage 70 and identity 95, no excluded residues, no chain similarity check (the pdb has only one chain) and a limit of 10 simultaneously occurring mutations in a combination. (A) Improvement of developability potential. The graph plots the mutation score against the number of mutations in combination. From the graph it can be appreciated that, in this case, the mutation score grows monotonically although hinting to a slowly approaching plateau towards the 10 mutants. This type of growth can be expected for larger proteins where there are plenty of identified mutations and most of them are profitable in terms of stability and solubility (or just one of the two). For smaller proteins we expect a more evident plateau or that we hardly reach the 10th combination step. (B) Weighted best combination group analysis. As explained in section Best Mutant Combination Group in the main paper, we perform what we call a “weighted analysis” on the highest scoring, in terms of mutation score, mutations for each mutation group. In the graph are highlighted in blue the combination groups identified as Best Combination Groups. The other groups are colored in green. (C) First Derivative Analysis. The graph highlights the best mutation groups identified through what we call “first derivative analysis”, as again explained in the Best Mutant Combination Group section. Best combination groups are highlighted by a red circle in the graph. The final set of best combination groups is the combination of the groups identified in plot (B) and (C).