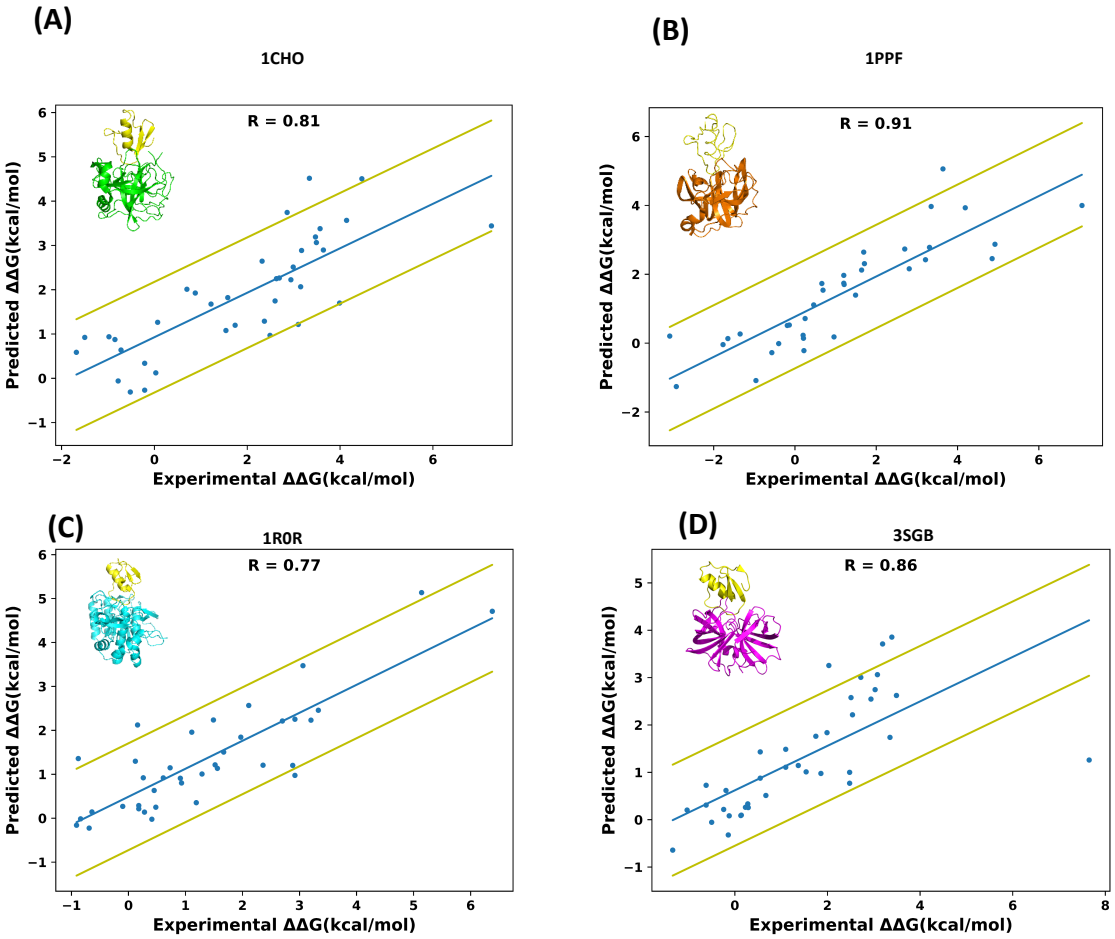
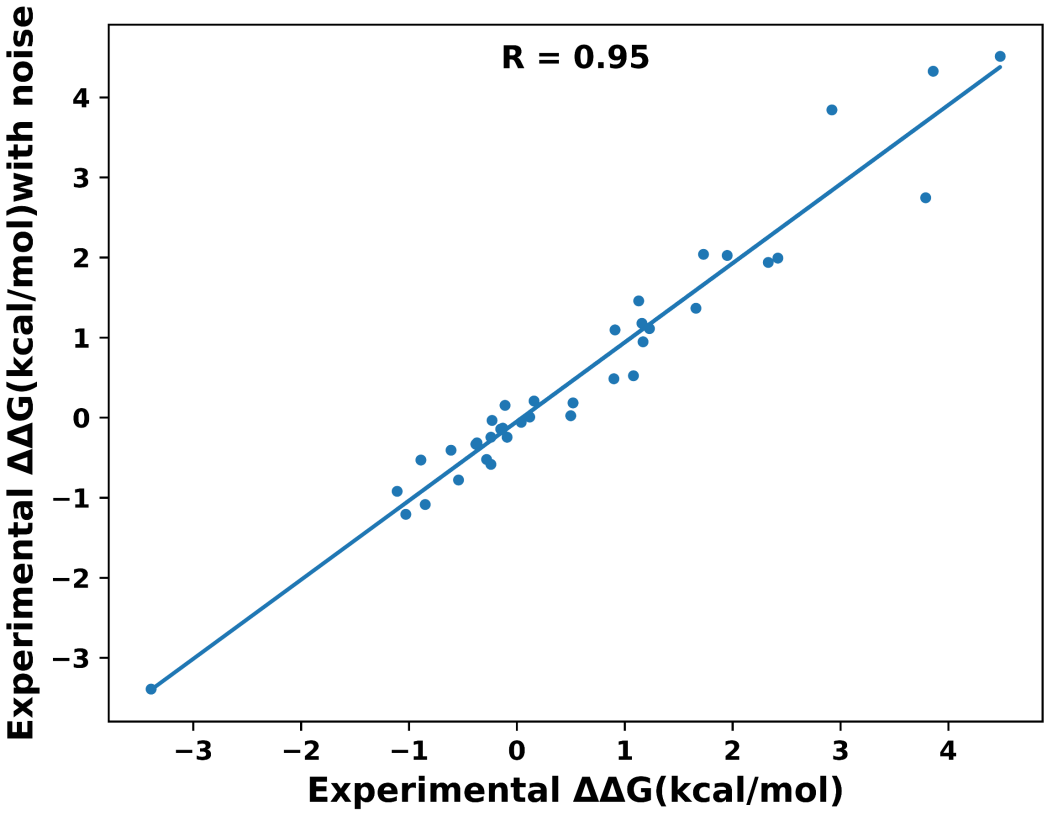
bioRxiv preprint doi: https://doi.org/10.1101/2024.06.21.600041; this version posted June 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Information



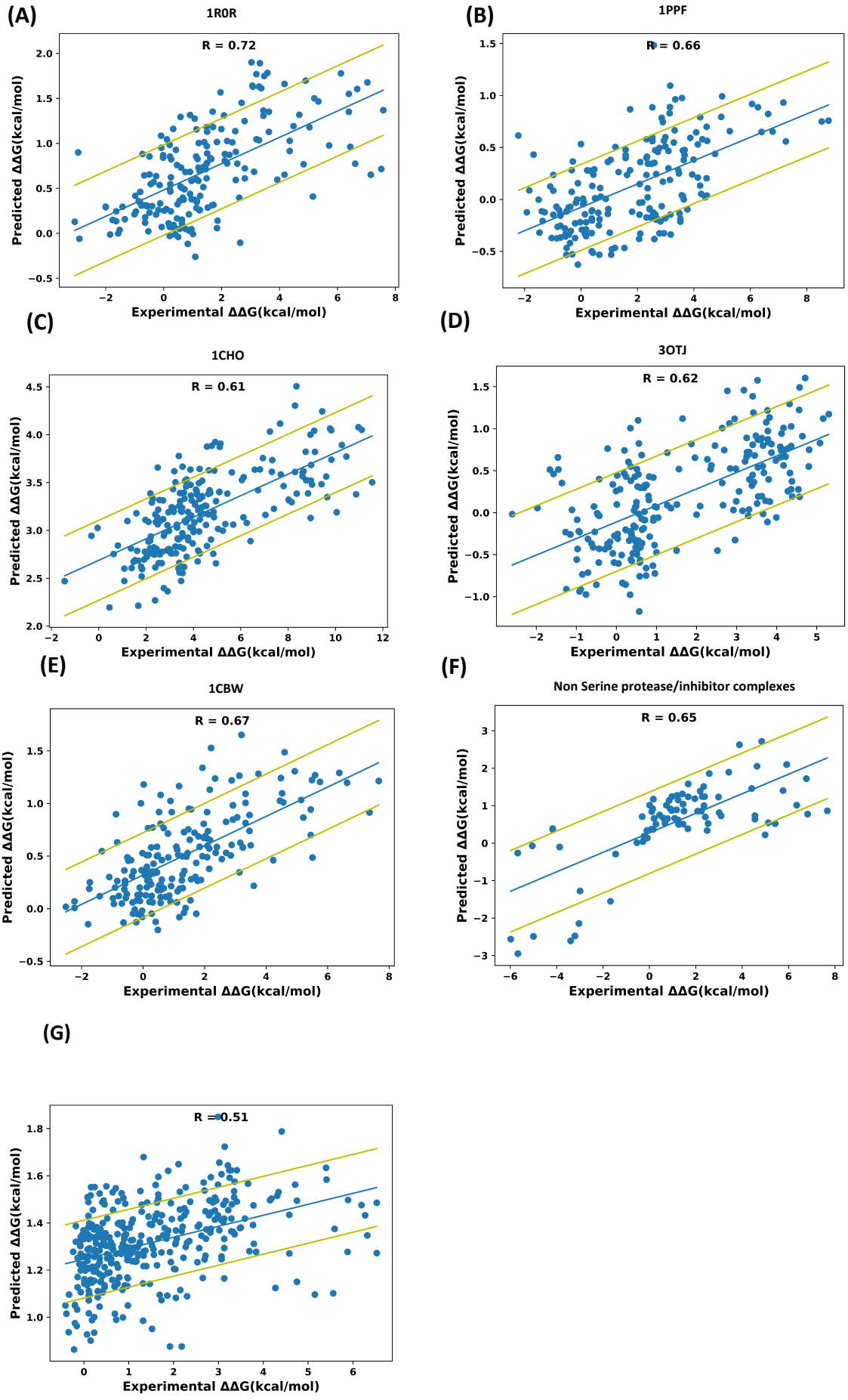
Supplementary Figure 1: Predicting $\Delta\Delta G_{bind}$ on single PPIs. Correlation between experimental and predicted $\Delta\Delta G_{bind}$ values when training and testing is performed on a single PPI: (A) a complex between Ovomucoid and Alpha-Chymotrypsin (PDB ID 1CHO), (B) a complex between Ovomucoid and Human Leukocyte elastase (PDB ID 1PPF), (C) a complex between Ovomucoid and subtilisin carlsberg (PDB ID 1R0R), (D) a complex between Ovomucoid and subtilisin Proteinase b (PDB ID 3SGB). The blue line represents the best linear fit of the data with the Person correlation R-value given on each graph. The yellow lines represent one standard deviation above and below the fitted line.

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.21.600041; this version posted June 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Supplementary Figure 2: Influence of Noise on experimental data. Correlation between experimental $\Delta\Delta G_{bind}$ values and the same values with noise added according to the standard deviation measured for each data point. (data for colicin/DNAse complexes (PBD ID 2WPT and 1EMV).

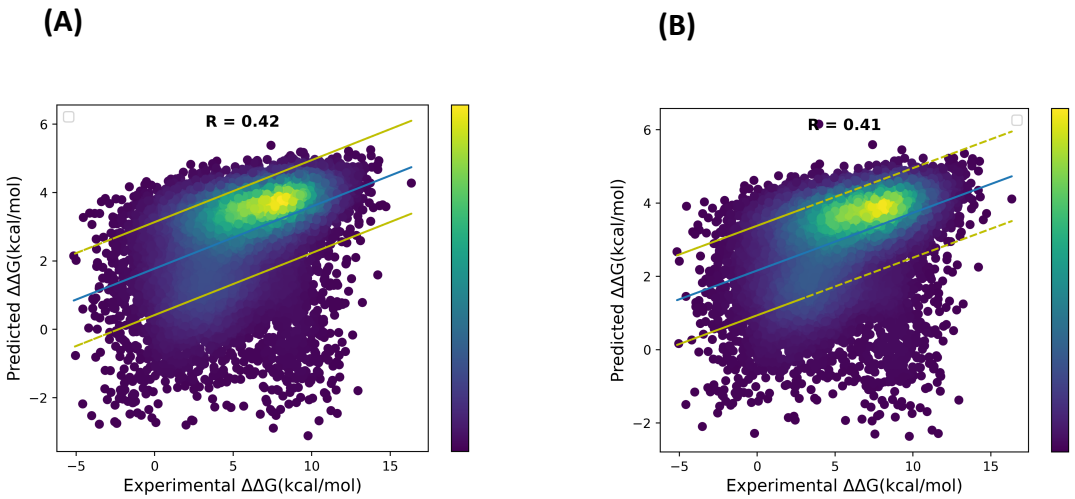
bioRxiv preprint doi: https://doi.org/10.1101/2024.06.21.600041; this version posted June 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2024.06.21.600041; this version posted June 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

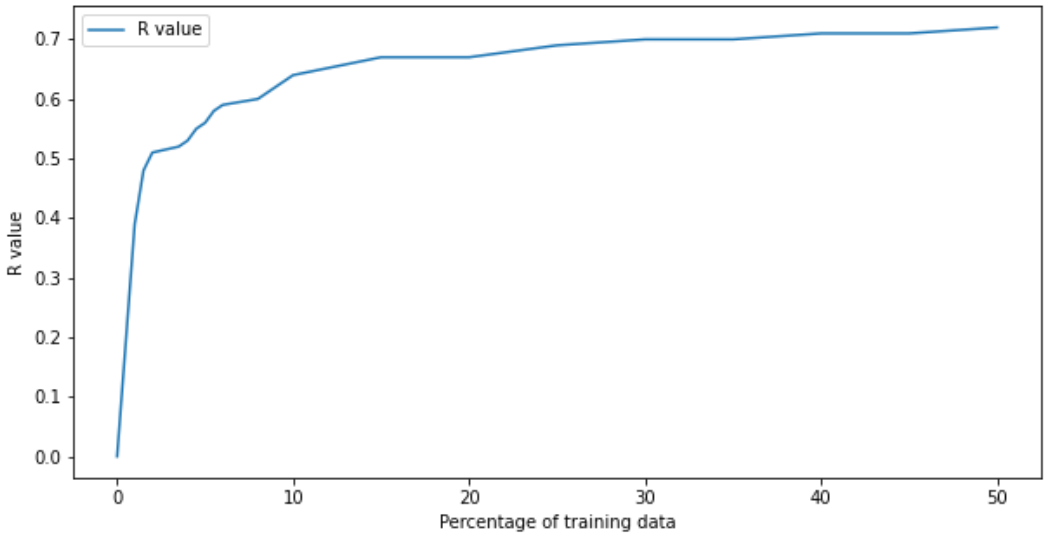
Supplementary Figure 3: Predicting $\Delta\Delta G_{bind}$ for single mutations. Correlation between experimental and predicted $\Delta\Delta G_{bind}$ values after the model was trained on the whole dataset excluding the data for the PDB file under evaluation (A) a complex between Ovomucoid and subtilisin carlsberg (PDB ID 1R0R), (B) a complex between Ovomucoid and Human Leukocyte elastase (PDB ID 1PPF), (C) a complex between Ovomucoid and Alpha-Chymotrypsin(PDB 1CHO), (D) a complex between BPTI and Trypsin (PDB ID 3OTJ), (E) a complex between BPTI and Chymotrypsin (PDB ID 1CBW). (F) Non serine protease/inhibitor complexes (PDB IDs : 1CSE, 1CT2, 1EMV, 1S1Q, 1SBB, 1SGD, 1CT2). The blue line represents the best linear fit of the data. (G) A complex between Angiotensin-converting enzyme 2 (ACE2) and Spike protein S1 (PDB ID 6M0J). The yellow lines represent one standard deviation above and below the fitted line.

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.21.600041; this version posted June 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Supplementary Figure 4: Predicting $\Delta\Delta G_{bind}$ for double mutations (A) Correlation between experimental and predicted $\Delta\Delta G_{bind}$ when model was trained on double mutations belonging to the BPTI /Chymotrypsin complex (PDB ID 1CBW) and tested on double mutations belonging to the BPTI/bovine Trypsin complex (PDB ID 3OTJ). (B) Correlation between experimental and predicted $\Delta\Delta G_{bind}$ when model was trained on the whole dataset of double mutants and tested on double mutations belonging to the BPTI/ bovine Trypsin complex (PDB ID 3OTJ). The blue line represents the best linear fit of the data. The yellow lines represent one standard deviation above and below the fitted line. The points are colored according to their local density, with the color bar indicating the density scale. Higher density areas (yellow color) represent regions where data points are more concentrated."

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.21.600041; this version posted June 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

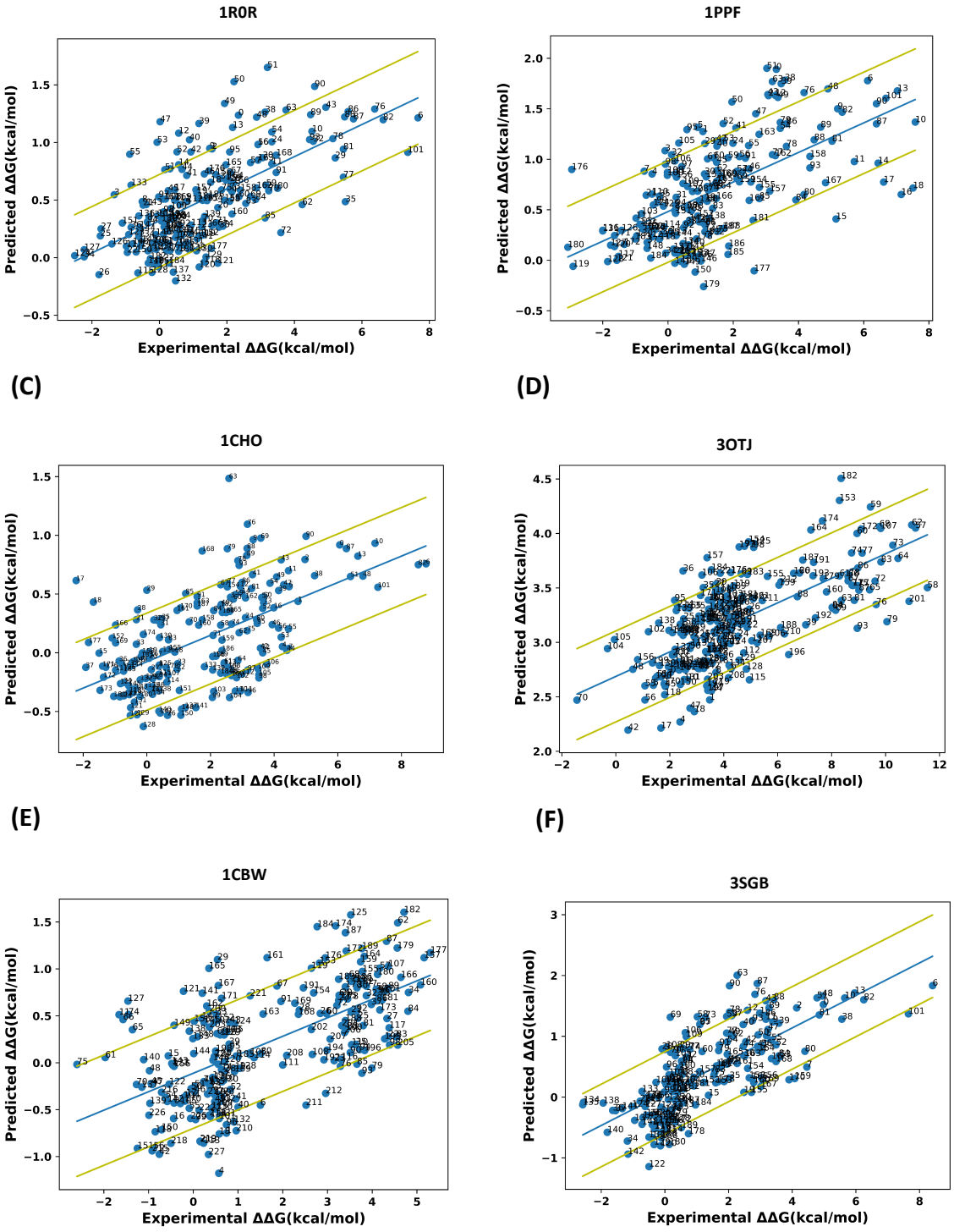


Supplementary Figure 5: Training data vs Correlation. Graph illustrates the impact of increasing the percentage of training data on the R value between experimental and predicted $\Delta\Delta G_{bind}$.

(A)

(B)

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.21.600041; this version posted June 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Supplementary Figure 6: Analysis of the outliers for the six PDBs (A) PDB ID 1R0R, (B) PDB ID 1PPF, (C) PDB 1CHO), (D) Mutation positions in the complex between BPTI and Trypsin (PDB ID 3OTJ. (E) PDB ID 1CBW, (F) PDB ID 3SGB. ProBASS was trained on the whole dataset excluding the test PDB file and predictions were made. The blue line represents the best liner fit to the data and the yellow lines correspond to one standard deviations from the fitted line. The mutations lying above and below the one-standard-deviation line were numbered and analyzed in the context of the structure. See Supplementary data for mutation description, where mutations predicted to be overly disruptive to the complex are colored in red and mutations predicted overly stabilizing for the PPI are colored in cyan. Outliers. Xlxm file is available in the ProBASS repository: (https://github.com/sagagugit/ProBASS).