Supplementary Information

Native or non-native protein-protein docking models? Molecular dynamics to the rescue.

Zuzana Jandova[1], Attilio Vittorio Vargiu[2], Alexandre M. J. J. Bonvin[1]*

1 - Computational Structural Biology Group, Bijvoet Centre for Biomolecular Research, Faculty of Science - Chemistry, Utrecht University, Padualaan 8, 3584 CH Utrecht, the Netherlands.
2 – Physics Department, University of Cagliari, Cittadella Universitaria, S.P. 8 km 0.700, 09042 Monserrato, Italy

Figure S1: Comparison of initial ligand RMSDs and HADDOCK score per native and non-native clusters for 25 complexes.

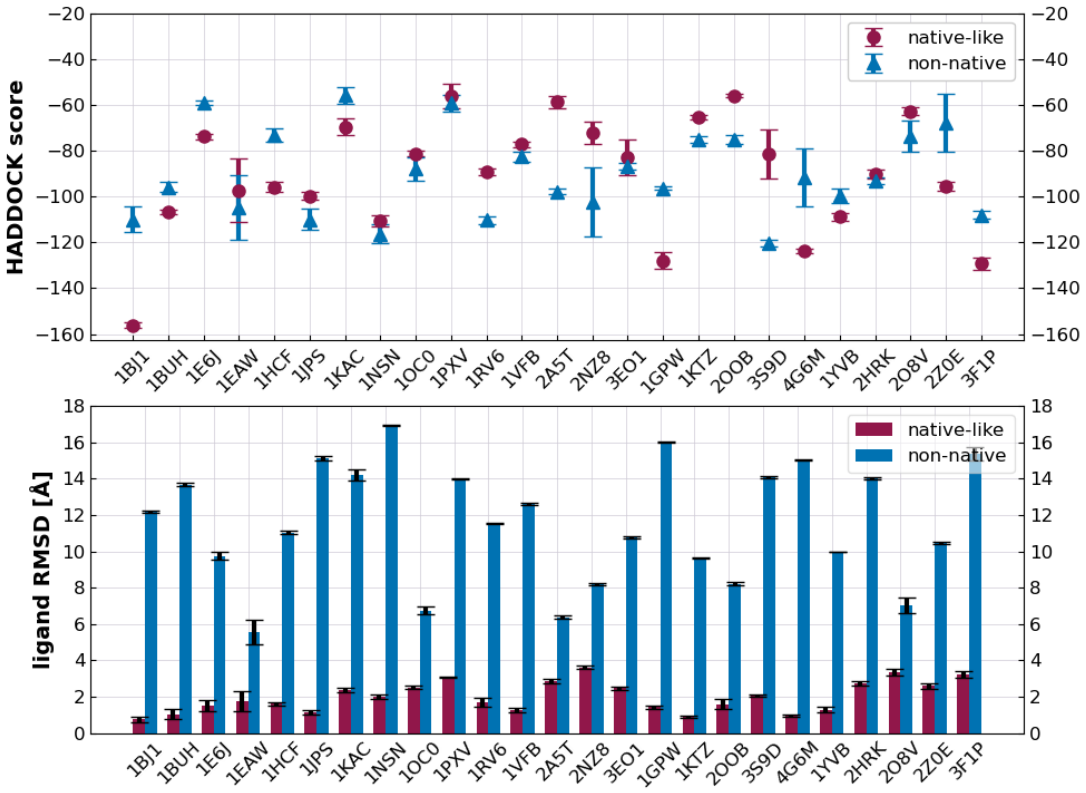
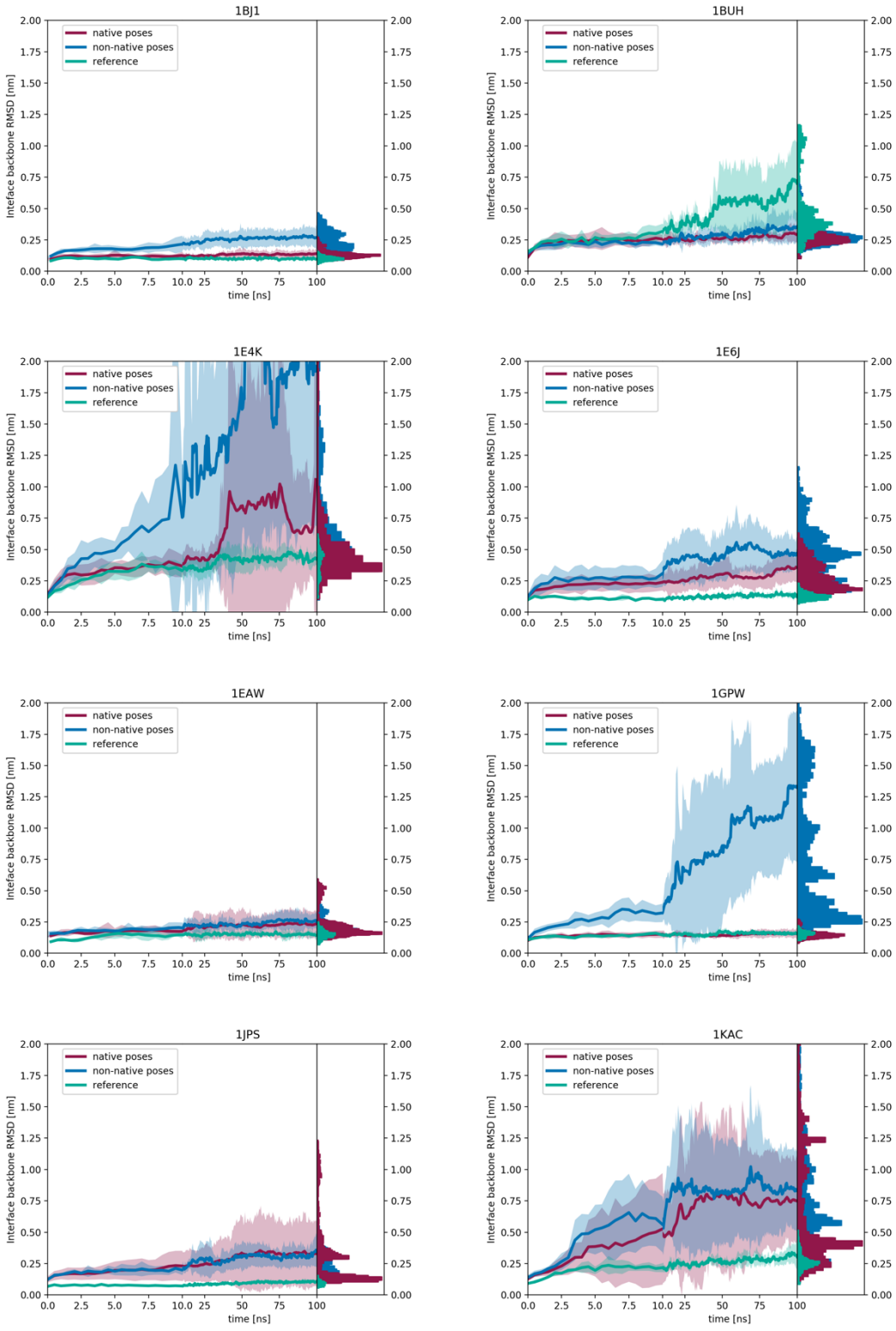
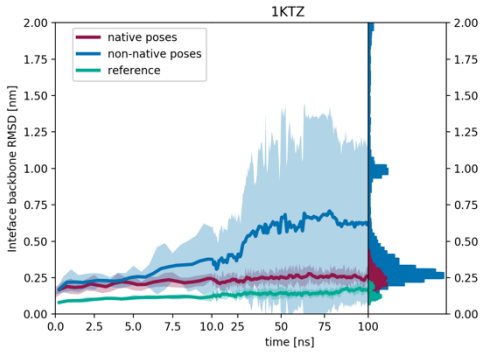
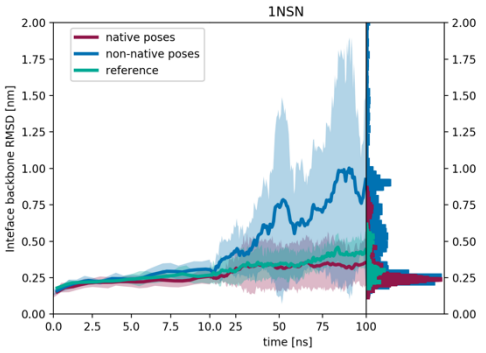
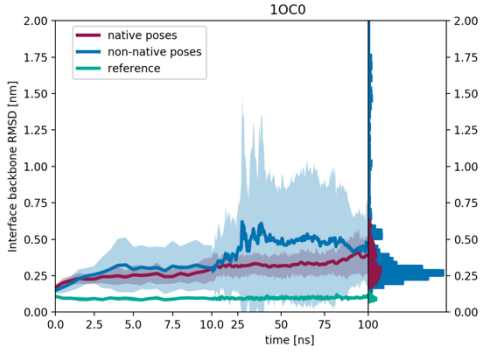


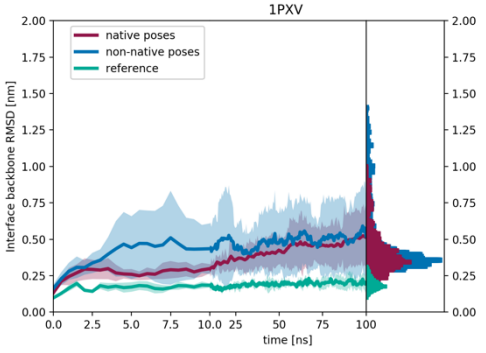
Figure S2: Time series as running average +- standard deviation of interface RMSDs of 20 protein-protein complexes of training set 1 and their histograms. Reference in teal, native clusters in burgundy, non-native in blue, standard deviation in lighter colours.

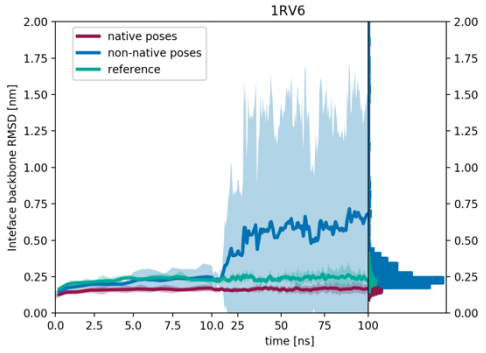


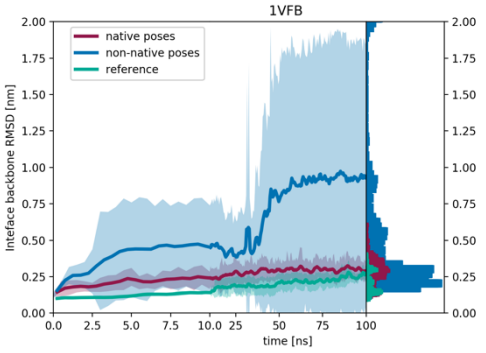

1KTZ
native poses
non-native poses
reference
Inteface backbone RMSD [nm]
time [ns]

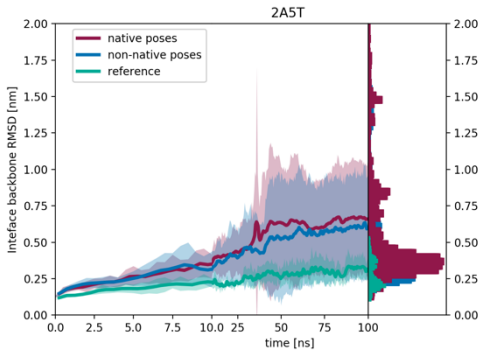

1NSN
native poses
non-native poses
reference
Inteface backbone RMSD [nm]
time [ns]

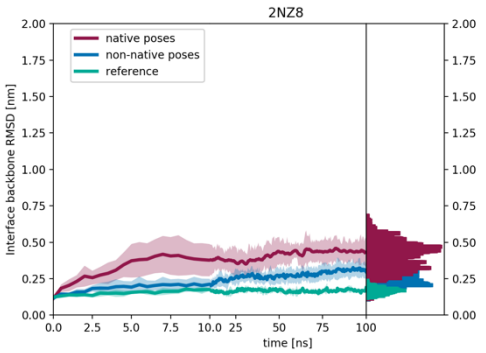

1OC0
native poses
non-native poses
reference
Interface backbone RMSD [nm]
time [ns]

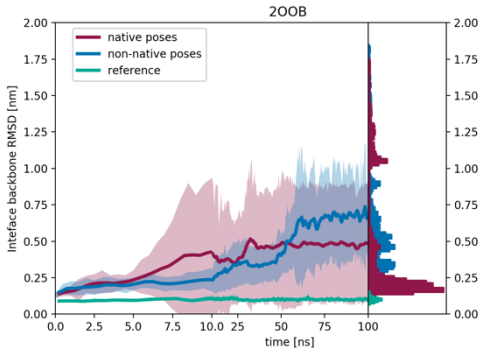

1PXV
native poses
non-native poses
reference
Interface backbone RMSD [nm]
time [ns]

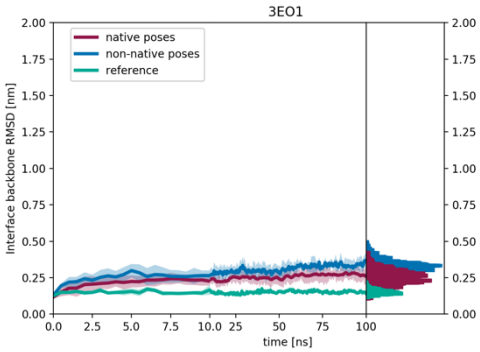

1RV6
native poses
non-native poses
reference
Inteface backbone RMSD [nm]
time [ns]

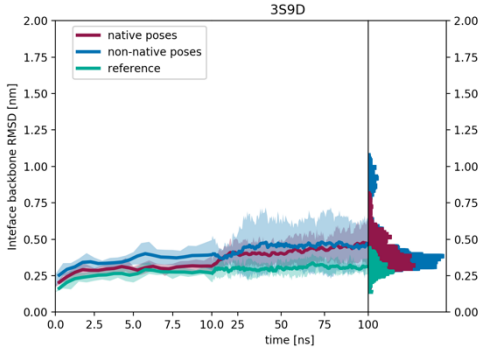

1VFB
native poses
non-native poses
reference
Inteface backbone RMSD [nm]
time [ns]


2A5T
native poses
non-native poses
reference
Inteface backbone RMSD [nm]
time [ns]


2NZ8
native poses
non-native poses
reference
Interface backbone RMSD [nm]
time [ns]


2OOB
native poses
non-native poses
reference
Inteface backbone RMSD [nm]
time [ns]


3EO1
native poses
non-native poses
reference
Interface backbone RMSD [nm]
time [ns]


3S9D
native poses
non-native poses
reference
Inteface backbone RMSD [nm]
time [ns]

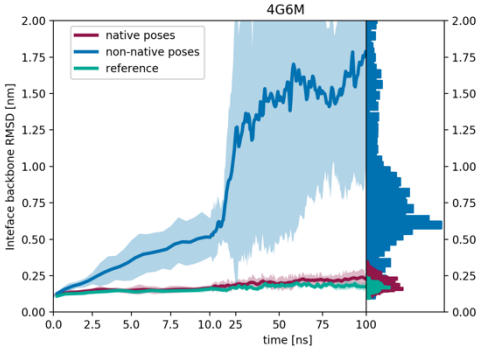
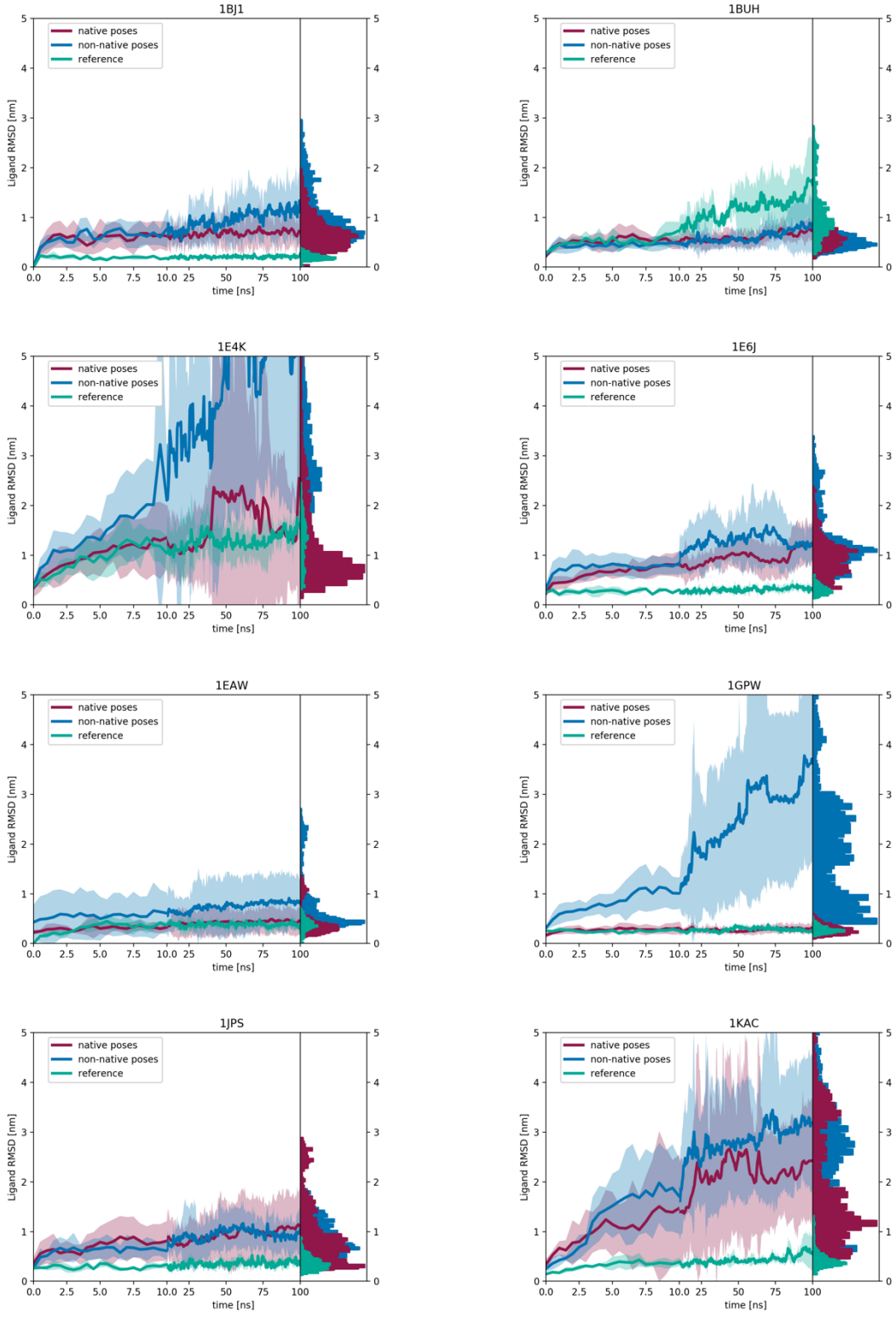
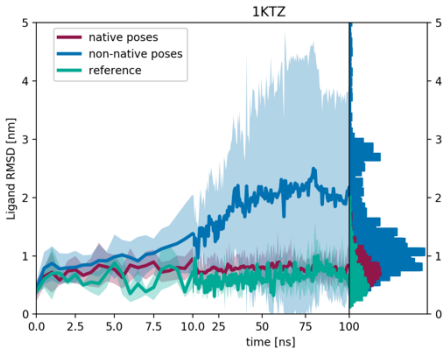
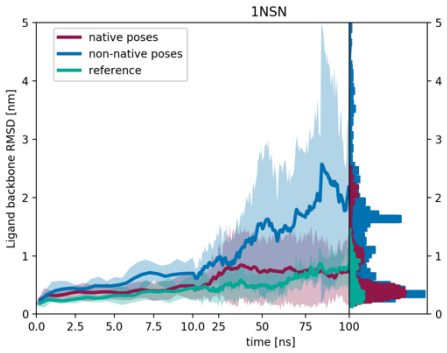
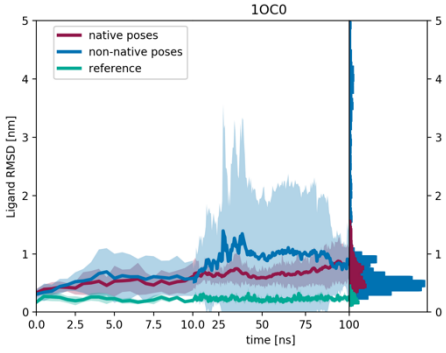

4G6M
native poses
non-native poses
reference
Inteface backbone RMSD [nm]
time [ns]

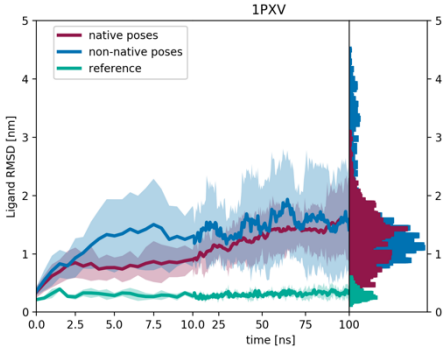
Figure S3: Time series as running averages +- standard deviation of ligand RMSDs of 20 protein-protein complexes and their histograms. Reference in teal, native clusters in burgundy, non-native in blue, standard deviation in lighter colours.

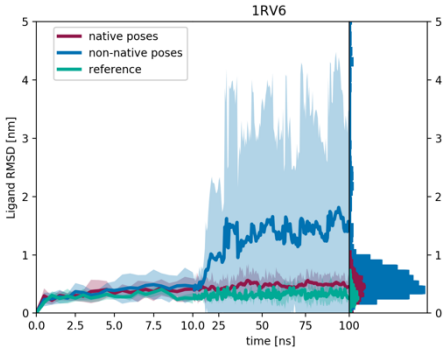


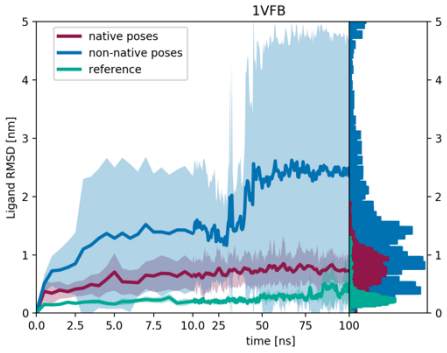

1KTZ
native poses
non-native poses
reference
Ligand RMSD [nm]
time [ns]

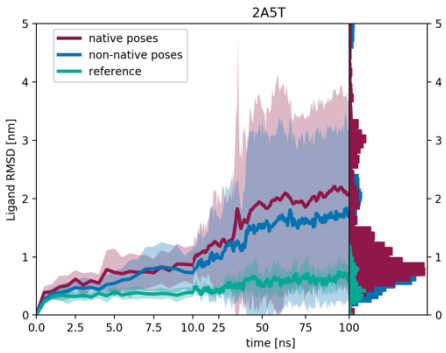

1NSN
native poses
non-native poses
reference
Ligand backbone RMSD [nm]
time [ns]

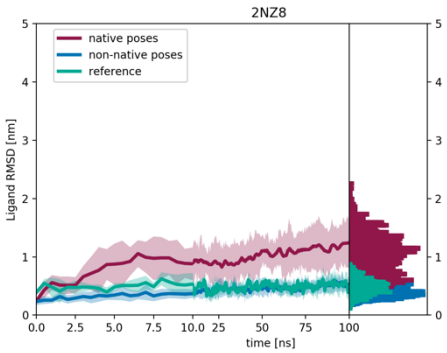

1OC0
native poses
non-native poses
reference
Ligand RMSD [nm]
time [ns]

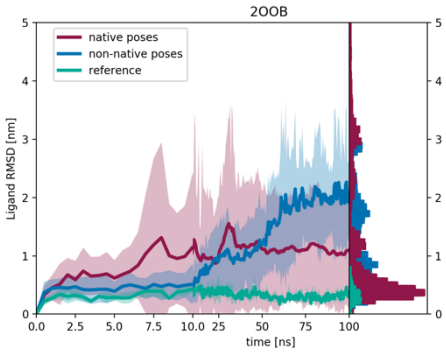

1PXV
native poses
non-native poses
reference
Ligand RMSD [nm]
time [ns]

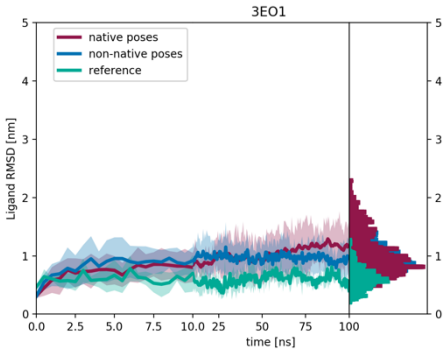

1RV6
native poses
non-native poses
reference
Ligand RMSD [nm]
time [ns]

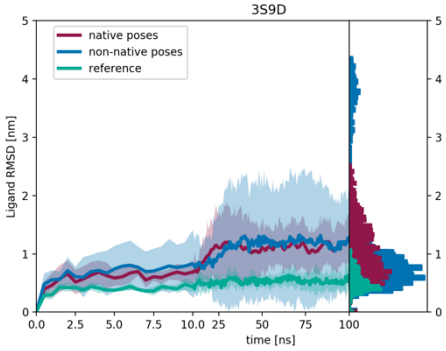

1VFB
native poses
non-native poses
reference
Ligand RMSD [nm]
time [ns]


2A5T
native poses
non-native poses
reference
Ligand RMSD [nm]
time [ns]


2NZ8
native poses
non-native poses
reference
Ligand RMSD [nm]
time [ns]


2OOB
native poses
non-native poses
reference
Ligand RMSD [nm]
time [ns]


3EO1
native poses
non-native poses
reference
Ligand RMSD [nm]
time [ns]


3S9D
native poses
non-native poses
reference
Ligand RMSD [nm]
time [ns]

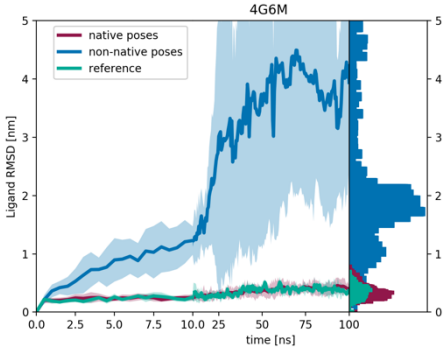
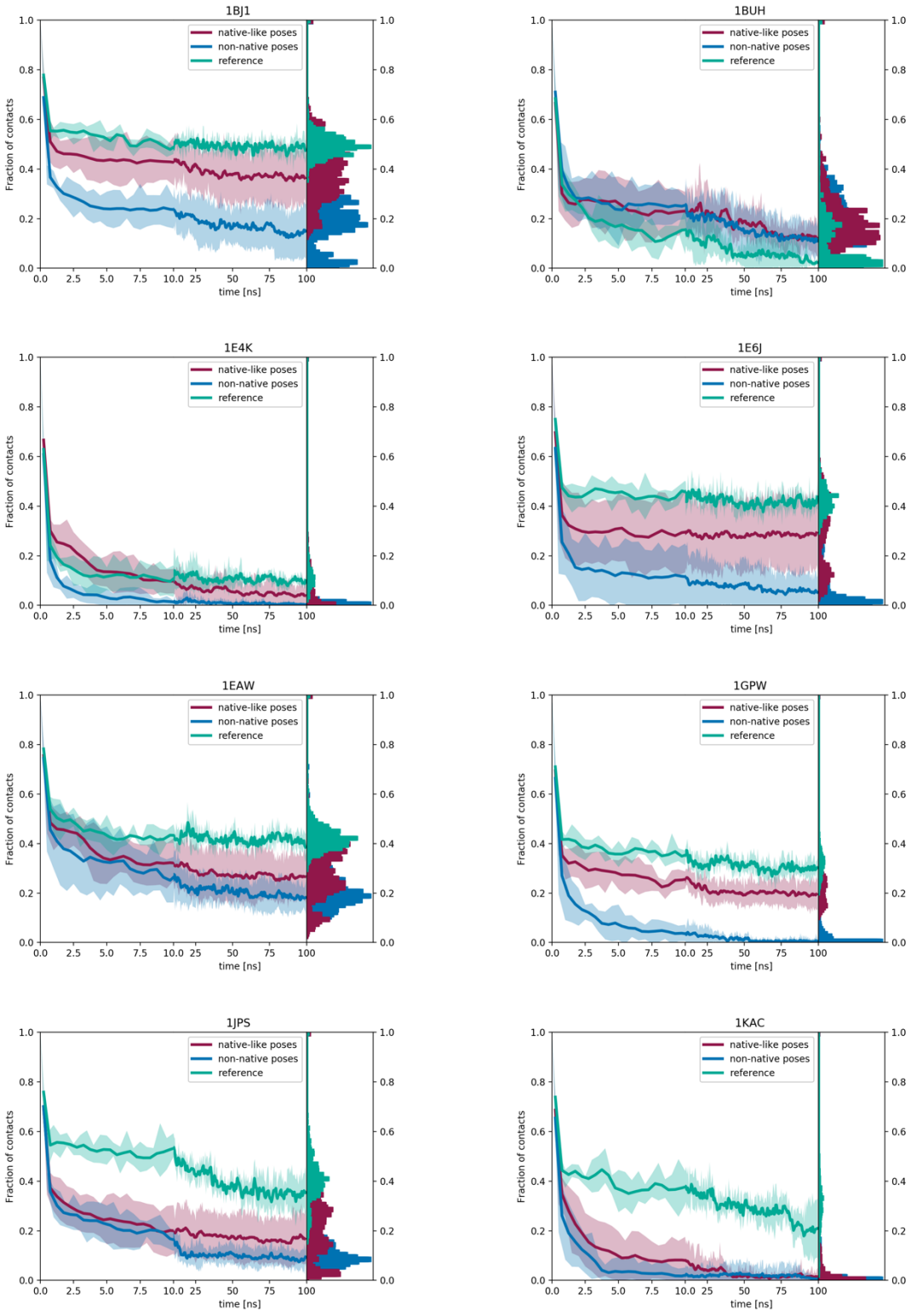
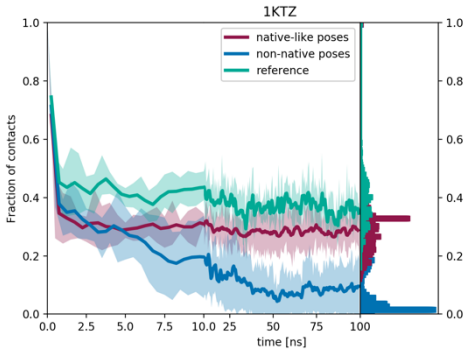
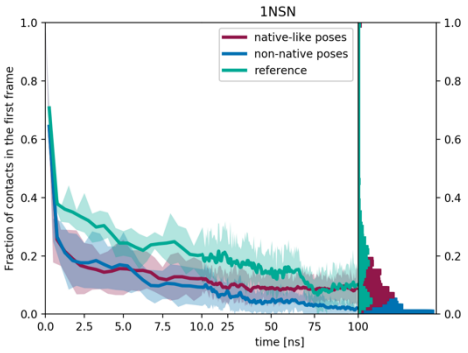
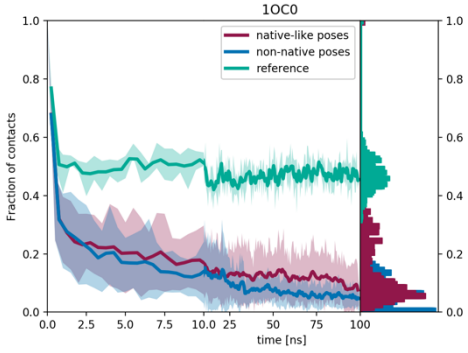

4G6M
native poses
non-native poses
reference
Ligand RMSD [nm]
time [ns]

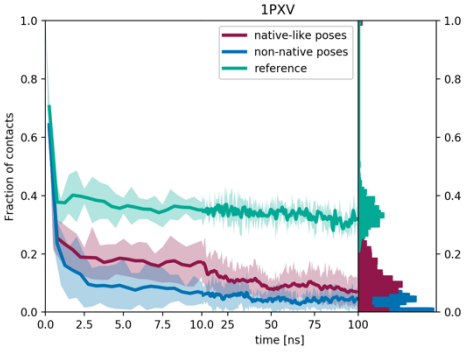
Figure S4: Time series as running averages +- standard deviation of fraction of native contacts of 20 protein-protein complexes and their histograms. Reference in teal, native clusters in burgundy, non-native in blue, standard deviation in lighter colours.

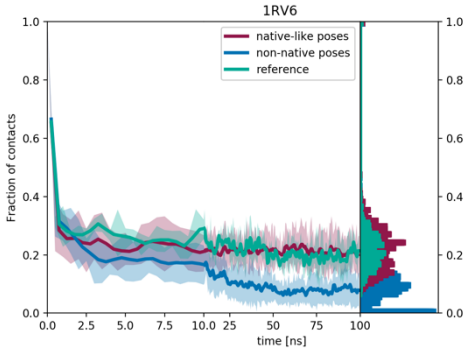


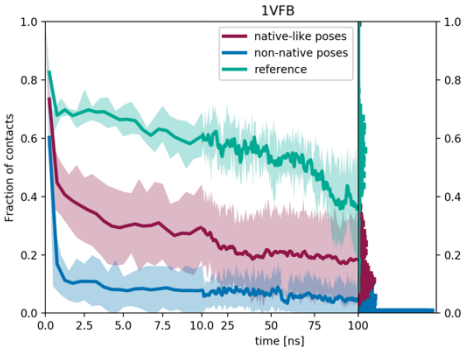

1KTZ
native-like poses
non-native poses
reference
Fraction of contacts
time [ns]

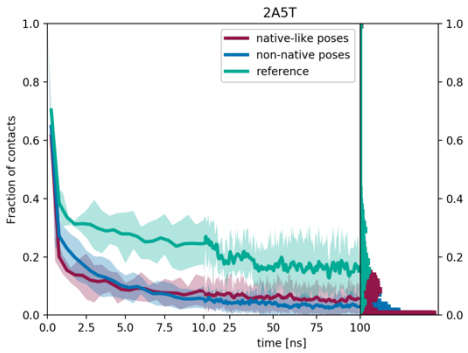

1NSN
native-like poses
non-native poses
reference
Fraction of contacts in the first frame
time [ns]

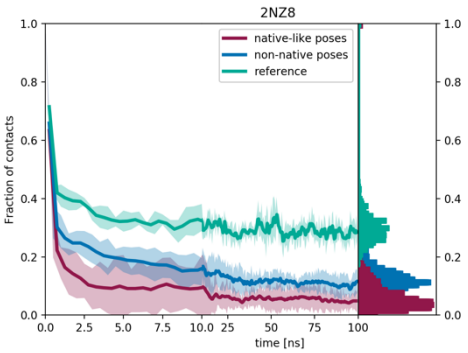

1OC0
native-like poses
non-native poses
reference
Fraction of contacts
time [ns]

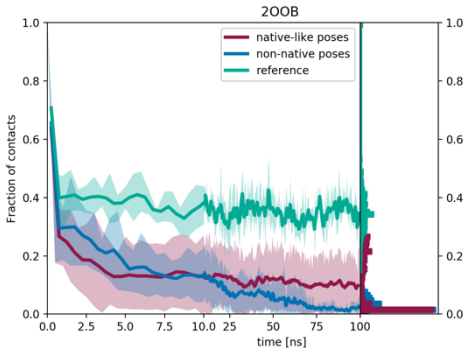

1PXV
native-like poses
non-native poses
reference
Fraction of contacts
time [ns]

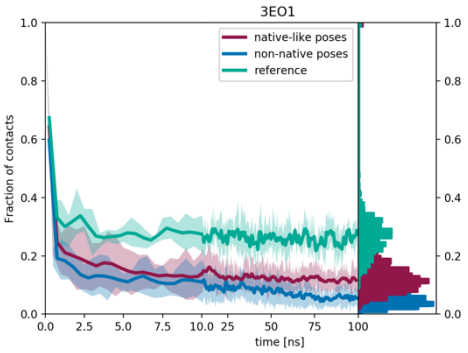

1RV6
native-like poses
non-native poses
reference
Fraction of contacts
time [ns]

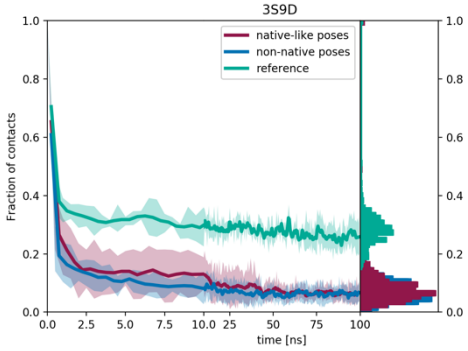

1VFB
native-like poses
non-native poses
reference
Fraction of contacts
time [ns]


2A5T
native-like poses
non-native poses
reference
Fraction of contacts
time [ns]


2NZ8
native-like poses
non-native poses
reference
Fraction of contacts
time [ns]


2OOB
native-like poses
non-native poses
reference
Fraction of contacts
time [ns]


3EO1
native-like poses
non-native poses
reference
Fraction of contacts
time [ns]


3S9D
native-like poses
non-native poses
reference
Fraction of contacts
time [ns]

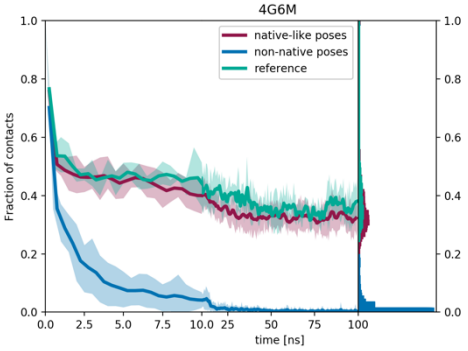

4G6M
native-like poses
non-native poses
reference
Fraction of contacts
time [ns]

Figure S5: A) Change is BSA B) change in distances between COMs of proteins and C) and change in the number of hydrogen bonds for native, non-native and reference structures from the beginning of the trajectory for all 20 complexes. Change in the nonbonded energy between protein-protein D) and proteins and water E). The boxplot shows the interquartile range with its median as black lines, mean as stars, whiskers as error bars and outliers in circles. Reference in teal, native clusters in burgundy, non-native in blue.

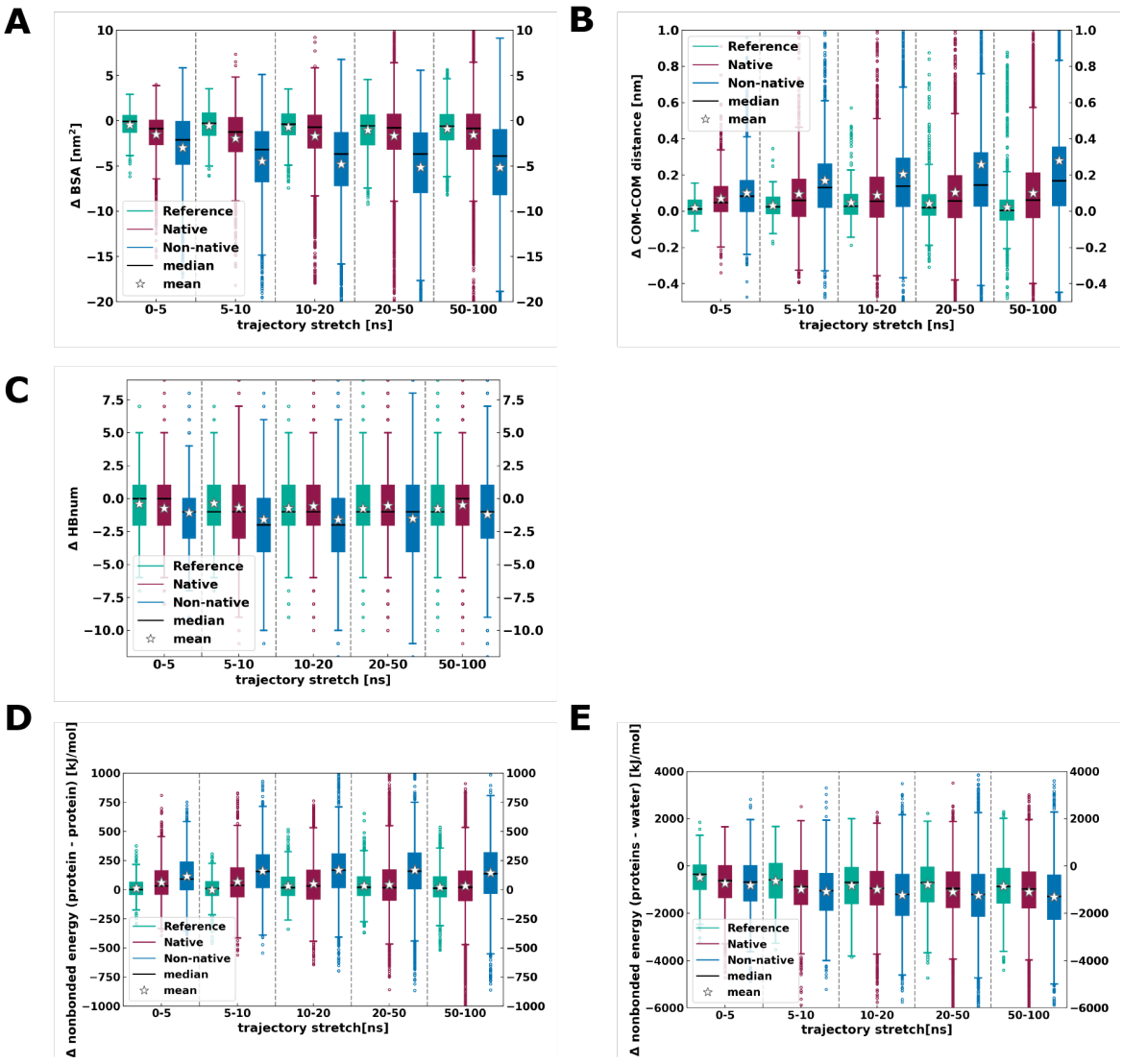


Figure S6: Pairplot of measured properties for last 20 ns of simulations of native (burgundy) and non-native (blue) complexes.

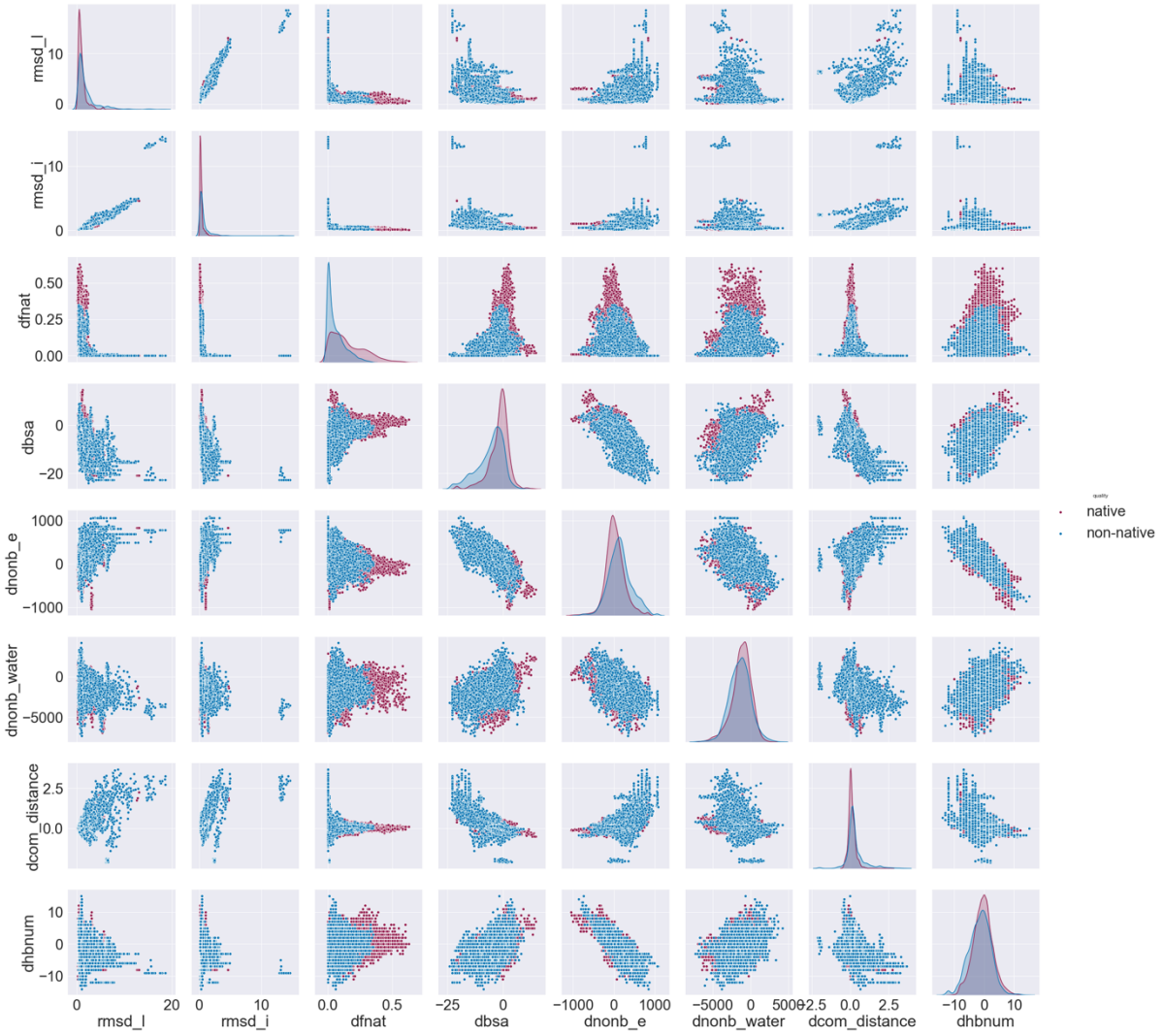


Figure S7: Accuracy scores of different classifiers from the Scikit library compared per timepoint of the trajectory. Abbreviations: gnb = Gaussian navie bayes, KNN = K Neighbors Classifier (n_neighbors=1), MNB = Multinomial naïve bayes, BNB = Bernoulli naïve bayes, LR = Logistic Regression, SDG = stochastic gradient descent Classifier, SVC = Support Vector classification , LSVC = Linear SVC, NSVC = Nu SVC, RF = Random Forest Classifier. Time points start at X ns and consists of data for 10 ns, i.e. timepoint at 90 ns consists of data for 90-100ns.

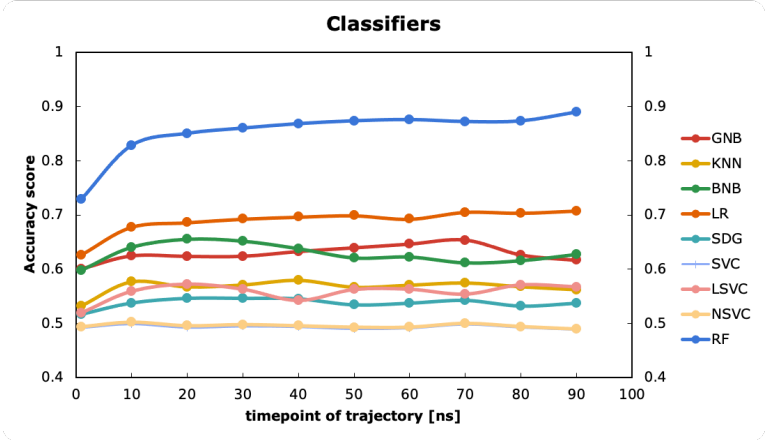


Figure S8: A) Interface RMSD B) Ligand RMSD and C) fraction of original contacts in % for native and non-native structures from the complex crystal structure for 5 complexes of the validation set 1. The boxplot shows the interquartile range with its median as black lines, mean as stars, whiskers as error bars and outliers in circles. Native clusters in burgundy, non-native in blue.

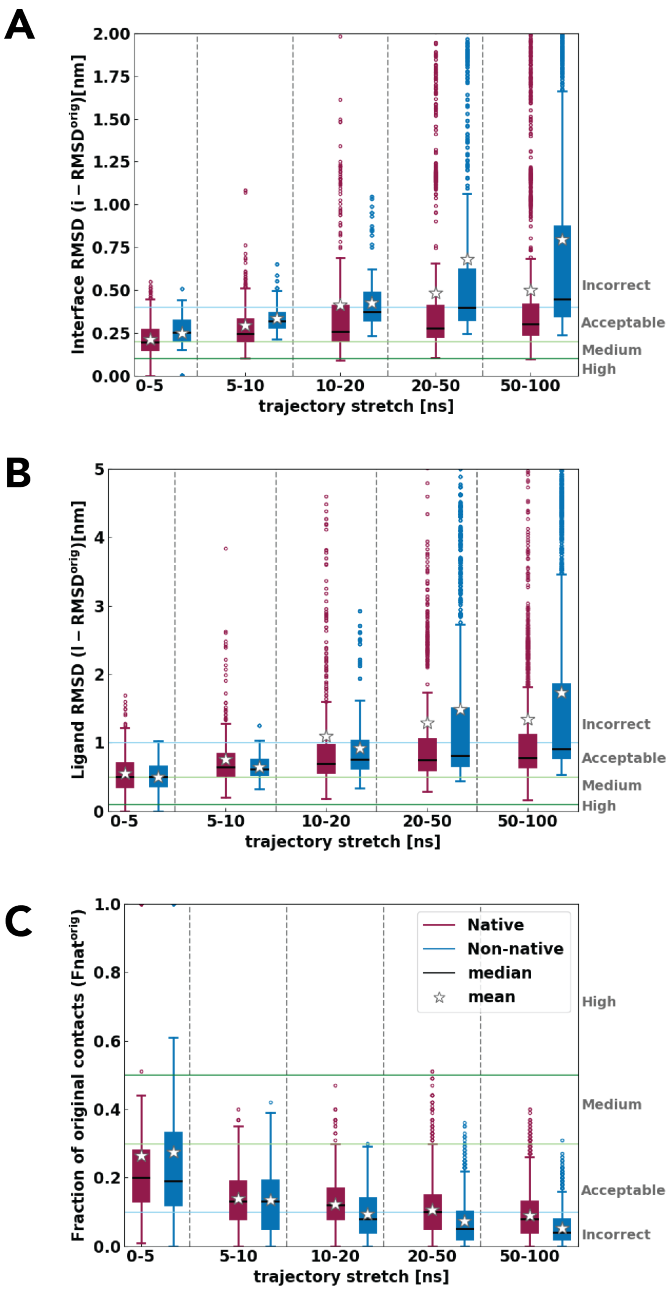


Figure S9: Relative feature = property importances for last 20 ns of trajectory. Abbreviations: rmsd_l = l-RMSD^{orig}, rmsd_i = i-RMSD^{orig}, dfnat = change in Fnat^{orig}, dbsa = change in BSA, dnonb_e = change in nonbonded energy between proteins, dnonb_water = change in nonbonded energy between proteins and water, dcom_distance = change in the distance between COMs of proteins, dhbnum = change in Hbnum.