Electronic Supplementary Information

Properties that Rank Protein:Protein Docking Poses with High Accuracy

Inês C. M. Simões, João T. S. Coimbra, Rui P. P. Neves, Inês P. D. Costa, Maria J. Ramos, Pedro A. Fernandes

UCIBIO, REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal

SECTION 1. SUPPORTING PROTOCOL

*MM-PBSA scoring.* When mutating the entire interface of each P:P docking pose, we observed that some mutations led to unphysical *ΔΔG*bind values (on average, they occurred for no more than 8% of all mutations performed in the 21 structures of each complex). These outliers resulted from the incorrect creation of the mutations during the running of the MMPBSA.py script. In these cases, we have used the MMPBSA.py script provided in the Amber 16 version of AmberTools.1 To test if the results are sensitive to the script version, we have used both Amber 12 and 16 in a test case (1VFB, B:B). On average, the deviation between versions is *ca.* 0.2 kcal mol-1 for 635 mutations (data not shown). For the 200 poses scoring, we have consistently applied the MMPBSA.py script from Amber 16.

*MD simulations’ protocol.* The MD protocol was performed considering the energy minimized structure of each complex. These were subjected to heating (100 ps) and NPT equilibration (500 ps) stages, prior to a NPT production of 1 ns. The temperature of the models was raised gradually from 0 to 300 K, followed by constant pressure dynamics (1 atm), employing periodic boundary conditions. The temperature was regulated by the Langevin thermostat and the pressure was controlled by an isotropic position scaling using the weak-coupling variety algorithm.2,3 The Particle Mesh Ewald (PME) method was applied to account for long-range electrostatic interactions beyond a cut-off radius of 10 Å.4 All the covalent bonds involving hydrogen atoms were constrained using the SHAKE algorithm, for an integration time step of 2 fs.5

SECTION 2. SUPPORTING RESULTS

*2.1. Supporting information for the dataset of P:P complexes.*

**Table ESI-1.** Additional details for the dataset of the 48 P:P complexes under study including the X-ray resolution for the P:P complex, the protein monomers that constitute the P:P complex (and respective PDB ID).

|  |  |  |  |
| --- | --- | --- | --- |
| complex (PDB ID) | resolution / Å | monomer A | monomer B |
| 3HFM | 3.00 | HYHEL-10 Igg1 Fab (3HFM) | hen egg white lysozyme (3HFM) |
| 1DQJ | 2.00 | anti-lysozyme HYHEL-63 (1DQJ) | lysozyme (1DQJ) |
| 1VFB | 1.80 | Igg1-kappa D1.3 FV (1VFA) | hen egg white lysozyme (2VB1) |
| 1JTG | 1.73 | β-lactamase TEM (4GKU) | β-lactamase inhibitory protein (1JTG) |
| 1TM1 | 1.70 | subtilisin BPN' precursor (1TM1) | chymotrypsin inhibitor 2 (1TM1) |
| 1Z7X | 1.95 | ribonuclease I (1Z7X) | ribonuclease inhibitor (1Z7X) |
| 1CHO | 1.80 | α-chymotrypsin A (1YPH) | turkey ovomucoid third domain (2GKR) |
| 3SGB | 1.80 | proteinase B (3SGB) | turkey ovomucoid inhibitor (2GKR) |
| 1KTZ | 2.15 | transforming growth factor β3 (1KTZ) | TGF-beta type II receptor (1KTZ) |
| 1F47 | 1.95 | cell division protein Zipa (1F47) | cell division protein FTSZ (1F46) |
| 1BRS | 2.00 | barnase (1A2P) | barstar (1A19) |
| 1FCC | 3.20 | Protein A/Z (1FCC) | IgG1 MO61 Fc (1FCC) |
| 1H9D | 2.60 | AML1 Runx1 Runt Domain (1H9D) | Core-binding factor beta |
| 2GOX | 2.20 | Complement C3d (2GOX) | Fibrinogen-binding protein Efb-C (2GOX) |
| 1AK4 | 2.36 | Cyclophilin A (1AK4) | HIV-1 capsid protein (1AK4) |
| 1EMV | 1.70 | Colicin E9 immunity protein (1EMV) | Colicin E9 DNase (1EMV) |
| 1SMF | 2.10 | Bovine trypsin (1SMF) | Mung bean inhibitor peptide (1SMF) |
| 1FFW | 2.70 | chemotaxis protein chey (3CHY) | chemotaxis protein chea (1FWP) |
| 1DVF | 1.90 | FV D1.3 (1VFA) | FV E5.2 (1DVF) |
| 1A4Y | 2.00 | ribonuclease inhibitor (1A4Y) | angiogenin (4AOH) |
| 2JEL | 2.50 | JEL42 FAB fragment (2JEL) | histidine-containing protein (1HDN) |
| 1CBW | 2.60 | bovine chymotrypsin (1AFQ) | BPTI (1OA5) |
| 1R0R | 1.10 | Subtilisin carlsberg (1SCA) | Ovomucoid (2GKR) |
| 1PPF | 1.80 | human leukocyte elastase (5ABW) | turkey ovomucoid inhibitor (1OMT) |
| 2G2U | 1.60 | β-lactamase SHV-1 (4FH4) | β-lactamase inhibitory protein (3GMU) |
| 1EAW | 2.93 | supressor of tumorigenicity (1EAX) | pancreatic trypsin inhibitor (1OA5) |
| 2VXT | 1.49 | Murine reference antibody 125-2H Fab | Interleukin-18 |
| 2W9E | 2.90 | ICSM 18 Fab fragment | Prion protein fragment |
| 3EOA | 2.80 | Efalizumab Fab fragment | Integrin alpha-L I domain |
| 3L5W | 2.00 | C836 Fab (3L7E) | Interleukin-13 (1IK0) |
| 4DN4 | 2.80 | CNTO888 Fab (4DN3) | MCP-1 (1DOL) |
| 4G6M | 1.81 | Gevokizumab antibody fragment | Interleukin-1 beta |
| 2A1A | 2.80 | Eukaryotic translation initiation factor 2-alpha-kinase 2 | eIF2 alpha-subunit |
| 3A4S | 2.70 | SUMO-conjugating enzyme UBC9 | NFATC2-interacting protein SLD2 ubiquitin-like domain |
| 3FN1 | 2.50 | UQ\_con domain from NEDD8-conjugating enzyme UBE2F (2EDI) | NEDD8-activating enzyme E1 catalytic subunit (2LQ7) |
| 3K75 | 2.95 | DNA polymerase beta | Reduced XRCC1, N-terminal domain |
| 3PC8 | 2.31 | DNA repair protein XRCC1 | DNA ligase III-alpha-BRCT Domain |
| 4IZ7 | 1.80 | Non-phosphorylated ERK | PEA-15 Death Effector Domain |
| 3VLB | 2.70 | EDGP | Xyloglucan-specific |
| 3H11 | 1.90 | Caspase-8 (4JJ7) | c-FLIPL protease-like domain (3H13) |
| 1M27 | 2.50 | SAP-SLAM complex (1D4T) | Fyn kinase SH3 domain (3UA6) |
| 2X9A | 2.50 | TolA C-terminal domain | G3P TolA binding domain |
| 3AAD | 3.30 | Double bromodomain (1EQF) | Histone chaperone ASF1 (1TEY) |
| 3BX7 | 2.10 | Lipocalin 2 | CTLA-4 extracellular domain |
| 3F1P | 1.17 | HIF2 alpha-C-terminal PAS domain (1P97) | ARNT C-terminal PAS domain (1X0O) |
| 3S9D | 2.00 | IFNAR2 | IFNa2 |
| 4M76 | 2.80 | C3D (1C3D) | Integrin alpha-M CD11B A-domain (1M1U) |
| 3H2V | 2.90 | Vinculin tail domain | Raver1 RRM1 domain |

*2.2. Supporting information for the performance of the different properties in ranking P:P docking poses.*

**Table ESI-2.** Average number of high-quality and medium-quality, and incorrect poses for dataset 1 (37 docking cases). The average *i*RMSD of the total number of poses over all complexes of B:B and B:U/U:U cases is also depicted.

|  |  |  |
| --- | --- | --- |
|  | U:U and B:U  (20 complexes) | B:B  (17 complexes) |
| high-, medium-quality poses | 7 | 11 |
| incorrect poses | 7 | 5 |
| *i*RMSD for the trial | 3.9 | 3.3 |

*MD simulations’ results.* We saw that, on average, the number of high- and medium-quality poses was twice the number of incorrect ones for the B:B cases, whereas in B:U/U:U cases they were the same (see Table ESI-2). This fact stresses the importance of the searching stage in a docking protocol. The quality of the solutions provided by the searching stage influences the chances of identifying high-quality poses at the top of the ranking. This aspect was particularly important in the B:U/U:U cases, because if only one HQ pose existed and this was not at the top-one position, the percentage of HQ poses at the top-*N* positions would be zero in these situations.

To account for the lack of high-quality poses in U:U complexes, for two P:P complexes of our dataset (1BRS and 1R0R), an alternative searching method was used. For these two cases, we have chosen the highest-quality pose provided by the HADDOCK searching stage (as measured by the *i*RMSD and *fnat*), and for each of the two complexes, we ran a short MD simulation (1 ns). From the simulations’ trajectory, we have then extracted high-quality poses for later reprocessing. These included five and seven poses for the 1R0R and 1BRS complexes, respectively. The full details of the MD simulation protocol are disclosed in section 1 of the ESI.

Without the MD-generated high-quality poses and considering the 1BRS complex, all properties were able to rank the native X-ray pose at the top-one position. However, for the 1R0R complex, the TPHS and *ΔΔG*bindsum,TP properties failed in doing so. Obviously, in the 1R0R case, the population of ASM results is scarce and so are the number of experimentally described HS residues.

**Table ESI-3.** Scoring performance of the different properties: TPHS, *i*A, *ΔG*bindP:P, and *ΔΔG*bindsum,TP; and considering two U:U complexes (1BRS and 1R0R) to which the poses were generated by HADDOCK or by HADDOCK+MD. The ranking performance was assessed by the *i*RMSD of the first ranked pose, and by the percentages of high-quality poses (%HQ), high- or medium-quality poses (%HQ/MQ), and incorrect poses (%IPtop) at the top-*N* predictions.

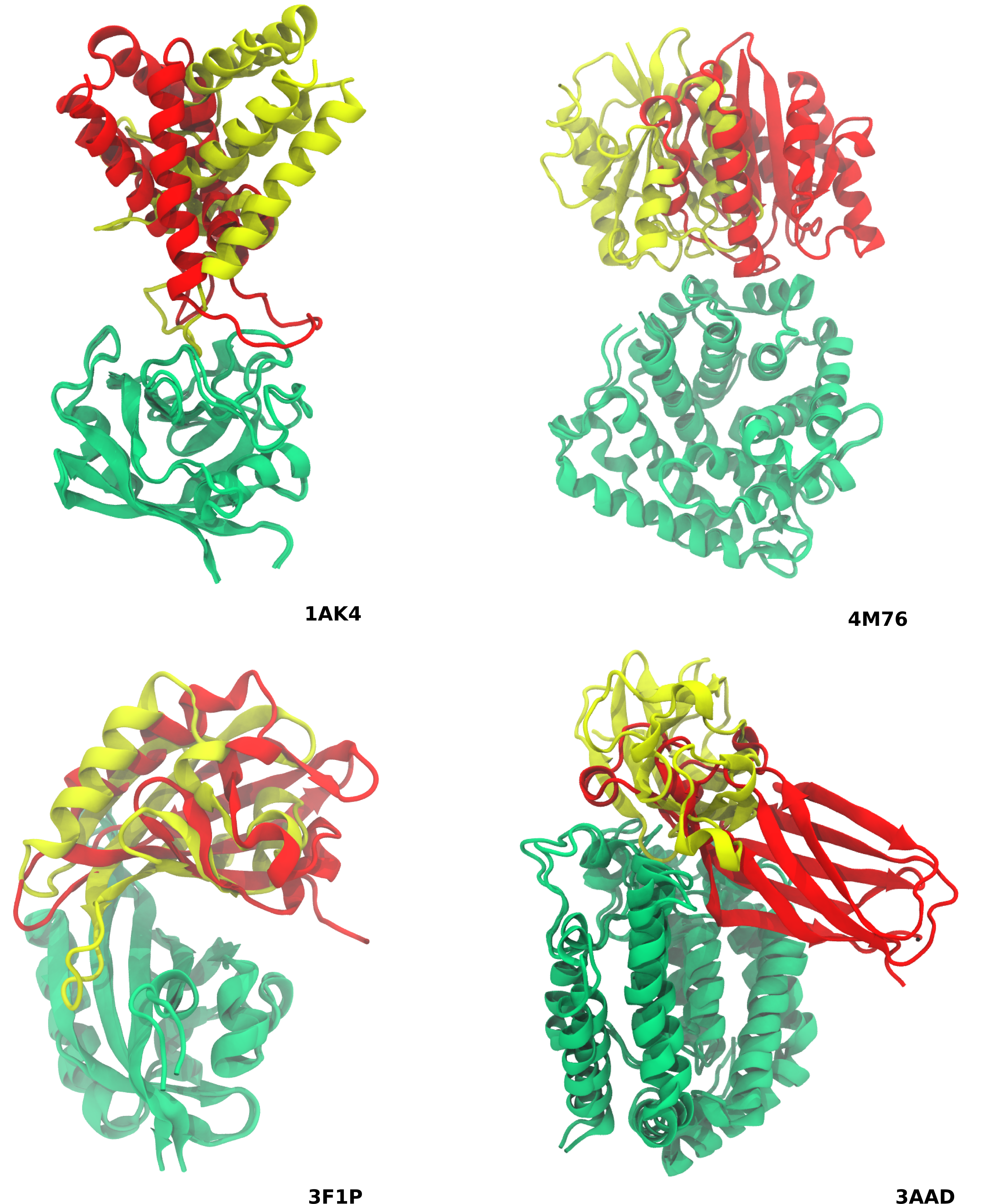
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **type** | **property** | ***i*RMSD 1st str. / Å** | **%HQ** | **%HQ/MQ** | **%IPtop** |
| **U:U (HADDOCK)** | TPHS | 0.6±0.6 | 50±50 | 100±0 | 0±0 |
|  | *ΔΔG*bindsum,TP | 0.0±0.0 | 100±0 | 100±0 | 0±0 |
|  | *ΔG*bindP:P | 0.0±0.0 | 100±0 | 100±0 | 0±0 |
|  | *i*A | 0.0±0.0 | 100±0 | 100±0 | 0±0 |
| **U:U (HADDOCK+MD)** | TPHS | 0.6±0.6 | 67±33 | 75±25 | 8±8 |
|  | *ΔΔG*bindsum,TP | 0.4±0.4 | 92±8 | 100±0 | 0±0 |
|  | *ΔG*bindP:P | 0.4±0.4 | 100±0 | 100±0 | 0±0 |
|  | *i*A | 0.9±0.1 | 100±0 | 100±0 | 0±0 |

Based on the analysis of the Table ESI-3, and looking at the percentage of HQ poses at the top-*N* positions (normalized for each situation), we saw an increment from 50% to 67% for the U:U cases supplemented with MD-generated HQ poses. This observation was only relevant for TPHS, because for the other two properties, *i*A and *ΔG*bindP:P, the native X-ray pose was already ranked at the top for both complexes. Still, for the latter two properties, even by including more poses in the dataset, these newly introduced HQ poses continue to be positioned at the top of the ranking. For *ΔΔG*bindsum,TP we observed only a slight decrease of the percentage of HQ poses (from 100% to 92%).

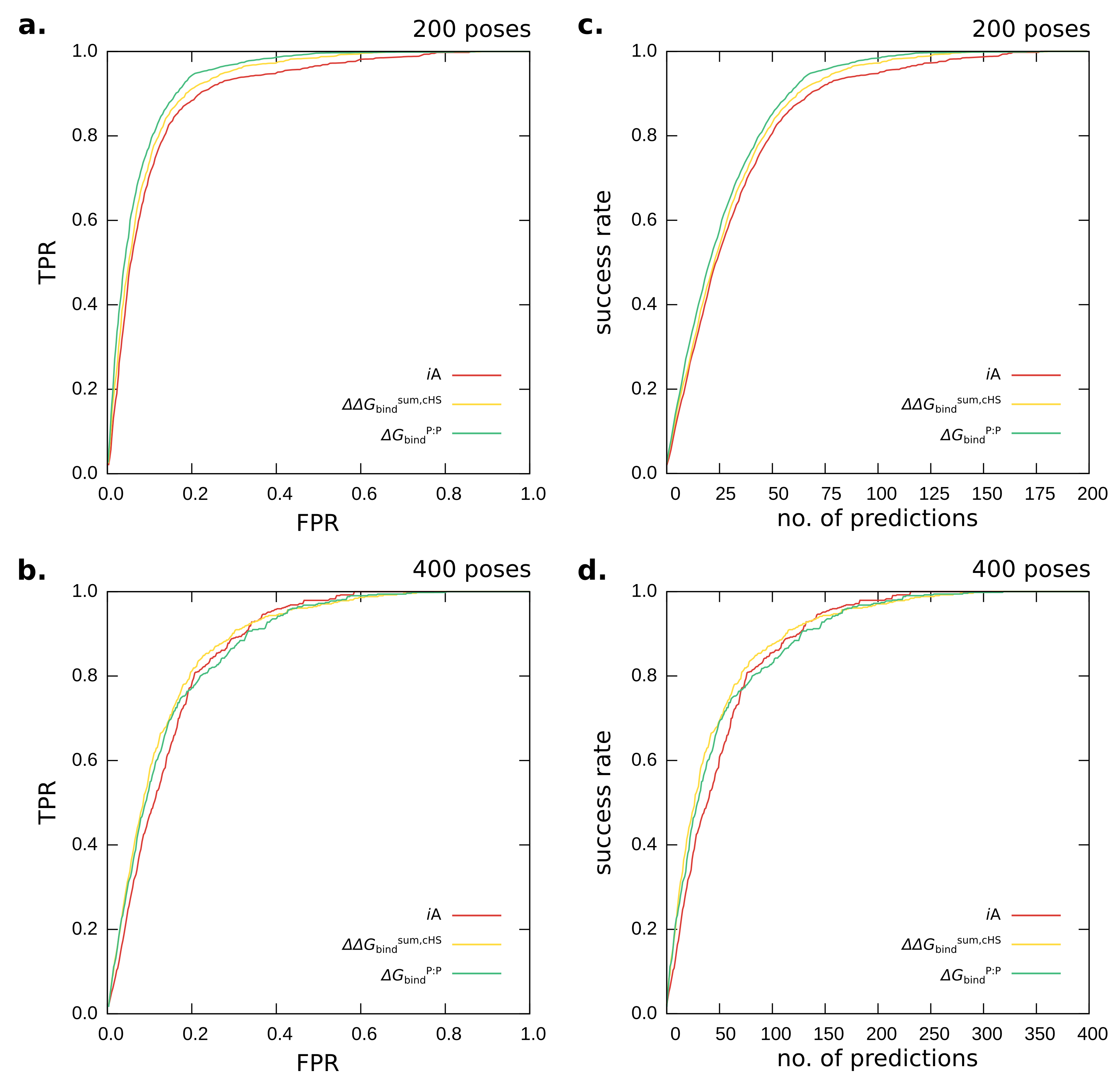
Obviously, by introducing HQ poses in our dataset, we hampered the results considering the %HQ/MQ criteria. This happens because we increased our top-*N* predictions (from one to six or eight), and so we increased the probability of finding lower quality poses at this increased top-*N*. We also saw that the *i*RMSD of the first pose was higher for the cases where we provided MD-generated poses, because HQ poses were ranked first than the native X-ray pose.

Notwithstanding, with only short MD simulations of a P:P docking pose, we were able to relax the structure of two P:P complexes and generate HQ poses. We showed for these two P:P complexes that by introducing more HQ poses in the dataset of poses, these were still ranked first than incorrect poses for most of the properties analysed. The exception were the properties dependent on *c*ASM results, which was probably related to the reduced sensitivity owned to the lower number of experimentally described HS residues for one of the complexes (specifically for 1R0R).

*Section 2. 3. Supporting information for the ranking performance in a wider pose sample space.*



**Figure ESI-1.** First ranked pose of four out of the 39 P:P complexes explored in the “*Ranking performance in a wider pose sample space*” section. In green cartoon we show the monomer that was used for the alignment of the P:P complexes, and in yellow and red we show the X-ray and first-ranked pose of the 200 or 400 pose space, respectively.



**Figure ESI-2.** Average ROC statistics (**a.** and **b.**) and success rates (**c.** and **d.**) for the ranking by the *ΔG*bindP:P, *ΔΔG*bindsum,*c*HS and *i*A properties, considering 39 P:P complexes. We have divided the complexes for which we have used 200 or 400 poses.

REFERENCES

1 H. G. D.A. Case, R.M. Betz, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, C. A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, I. Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, H.T. Nguyen, J. S. Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith and L. X. and P. A. K. R.C. Walker, J. Wang, R.M. Wolf, X. Wu, 2016.

2 H. J. C. Berendsen, J. P. M. Postma, W. F. Van Gunsteren, A. Dinola and J. R. Haak, Molecular dynamics with coupling to an external bath, *J. Chem. Phys.*, 1984, **81**, 3684–3690.

3 J. A. Lzaguirre, D. P. Catarello, J. M. Wozniak and R. D. Skeel, Langevin stabilization of molecular dynamics, *J. Chem. Phys.*, 2001, **114**, 2090–2098.

4 U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee and L. G. Pedersen, A smooth particle mesh Ewald method, *J. Chem. Phys.*, 1995, **103**, 8577–8593.

5 J. P. Ryckaert, G. Ciccotti and H. J. C. Berendsen, Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes, *J. Comput. Phys.*, 1977, **23**, 327–341.