Electronic supplementary materials *Mendeleev Commun.*, 2022, **32**, 735–738

Reducing false-positive rates in virtual screening *via* cancellation of systematic errors in the scoring function

Michael G. Medvedev, Oleg V. Stroganov, Artem O. Dmitrienko, Maria V. Panova, Alexey A. Lisov, Igor V. Svitanko, Fedor N. Novikov and Ghermes G. Chilov

This PDF file includes:

Methods

General details. All docking experiments in this article were performed using Lead Finder v.1.1.17 program package developed by some of us[1,2]. dG-score[1] scoring function (SF), an empirical SF parameterized for prediction of ligand binding free energies was used throughout the study. Each docking run included ~1000 attempts to dock ligand using different conformations and poses (orientations and translations relative to the protein), the best of which (according to a specialized "Ranking SF"[1]) is reported for comparison with other ligands/runs by means of dG-score. VS-score[1] SF has the same form as dG-score SF, but tweaked empirical parameters, optimized for virtual screening; its optimization was carried out on a subset of 16 proteins from the present test set.

Proteins and ligands were taken from the original Lead Finder virtual screening test-set[1]. It includes 35 diverse targets (see Table S1 for a complete list). All structures were prepared automatically at pH 7.4. Screened Coulomb potential was used for determination of pKas of ionizing residues. Some manual corrections were made in cases where crystallographic structures were resolved at pH significantly different from 7.4 (e.g., penicillopepsin). Active sites of the proteins were determined from positions of co-crystallized ligands. Grid sizes were determined by the reference ligands from corresponding X-ray structures and were set as size of the reference ligand + 6 Å in each direction.

Ligand library for each target included 1066 presumably inactive ligands (decoys; common for all targets, borrowed from the Surflex publication[3]) and 5 to 50 active ligands (individual set for each target, extracted from PDB[4], KiBank[5] and Surflex site[6]); all structures are available in SI. Ligands were prepared using OpenBabel[7] and ACD Labs ChemSketch[8] tools.

Automatic generation of centers for surface docking sites. At first, number of neighbors for every protein atom was calculated without taking into account water molecules and hydrogens. Atoms within 4 Å of each other were considered neighbors, and all heavy non-water atoms with

less than 10 neighbors were considered exterior (surface) atoms. Then a list of protein surface atoms was generated in such a way that at every step a surface atom which is the furthest from the atoms already in the list was selected and added to the list. Once the size of the list reached pre-defined value (25 in our case), the procedure stopped. Atoms from the list were then used as centers for docking energy grids.

For active sites, grid sizes were determined by the reference ligands from corresponding X-ray structures and were set as size of the reference ligand + 6 Å in each direction. For the sites on the protein surface, energy grid was defined as a box 20×20×20 Å centered on a selected protein atom. Default settings were used for all docking runs. All conjugated bonds were considered rotatable. To obtain converged results, we have performed 10 docking runs into each surface site.

Comparing accuracies of docking protocols. Accuracies of docking protocols were compared by means of the Area Under the Receiver Operating Characteristic curve (AUROC)[9]. This quantity directly measures the true positive rate vs. false positive rate. It provides a realistic representation of virtual screening efficiency and does not depend on the relative sizes of ligands and decoys subsets[10]. 100% accuracy by AUROC corresponds to perfect segregation between active and inactive ligands with zero false-positive rate; 50% AUROC designates a procedure resulting in a random mixture of active and inactive ligands; AUROC below 50% means that the procedure tends to reverse the correct ordering.

AUROC was computed using rankings of all ligands (5-50 active + 1066 decoys) for the given protein. In addition to AUROC we have monitored BEDROC(α=5)[9] and Enrichment factor[9] at 5% (see Table S2).

Proteins active site types and net charges. Active sites of all studied proteins were assigned to be open or closed manually by F.N.N. and O.V.S. To compute net charges of all targets for Figure 5, pKas of all ionizable residues were estimated with BuildModel[11] at pH=7, and used to evaluate protonation at the same pH. The code used for this task is available in SI.

Note on Poly(ADP-ribose) polymerase (PARP). During the work we have observed, that some "surface" sites provided better results for PARP than the active site constructed according to the "General details" section even with the conventional docking procedure (by 6% AUROC). Inspection has shown that different ligands bind at PARP active site in at least two manners: (1) found in the 1efy, 3l3m, 5a00 and 5xsu, or alternatively (2) found in 4oqb, 4hhz and 4hhy. Our initial active site grid did not allow ligands to adopt the second manner, so three of the "surface" sites, which were actually located in the active site of the protein provided better results in both conventional and on-top docking. For this reason, we have used docking scores obtained for the "grid 9" surface site as active site results and treated the initial active site as a surface site.

Importance of thymidylate synthase, RNases A and T1, and poly(ADP-ribose) polymerase for drug development. (1) Thymidylate synthase, which provides the only *de novo* source of 2-deoxythymidine-5-monophosphate (dTMP), required for DNA synthesis in living cells. Thymidylate synthase inhibitors are currently used to treat non-small cell lung[12], colorectal, pancreatic, breast, head and neck, gastric, and ovarian cancers[13]. It is known to be a hard task for conventional docking[14]. Accounting for ligands' affinities to the protein's surface improves accuracy for thymidylate synthase docking by 36%: from 58% to 94%.

(2) Ribonucleases A and T1, which catalyze the cleavage of RNA, mediating various biological processes ranging from cell signaling to innate immunity[15]. Mammalian RNases have been shown to have angiogenic and neurotoxic activities, and targeted inhibitors of these enzymes may have human therapeutic potential[16,17]. Ribonuclease A is known to be a hard nut to crack with either ligand-based or structure-based techniques[18], and was even called a "protein with discontinuous structure-activity relationship"[18]. Accounting for ligands' affinities to the ribonuclease A surface cracks it easily taking docking accuracy from 45% (complete nonsense) to 93% (48% improvement). For ribonuclease T1 improvement amounts to 35%: from 63% to 98%.

(3) Poly(ADP-ribose) polymerase (PARP), which plays key roles in DNA repair, genomic stability, and programmed cell death[19]. Its inhibitors are currently approved for treating patients

with breast[20] and ovarian[21] cancers, and are considered promising treatments of other cancer[22] and non-cancer[23] diseases. It was found, that conventional docking provides only mediocre accuracy for PARP, independent of the scoring function[24]. Accounting for ligands' affinities to its surface improves docking accuracy by 21%: from 47% to 68%.

Supplementary references:

(1) Stroganov, O. V.; Novikov, F. N.; Stroylov, V. S.; Kulkov, V.; Chilov, G. G. Lead Finder: An Approach To Improve Accuracy of Protein−Ligand Docking, Binding Energy Estimation, and Virtual Screening. *J. Chem. Inf. Model.* **2008**, *48* (12), 2371–2385. https://doi.org/10.1021/ci800166p.

(2) Novikov, F. N.; Stroylov, V. S.; Zeifman, A. A.; Stroganov, O. V.; Kulkov, V.; Chilov, G. G. Lead Finder Docking and Virtual Screening Evaluation with Astex and DUD Test Sets. *J. Comput. Aided Mol. Des.* **2012**, *26* (6), 725–735. https://doi.org/10.1007/s10822-012-9549-y.

(3) Pham, T. A.; Jain, A. N. Parameter Estimation for Scoring Protein−Ligand Interactions Using Negative Training Data. *Journal of Medicinal Chemistry* **2006**, *49* (20), 5856–5868. https://doi.org/10.1021/jm050040j.

(4) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242. https://doi.org/10.1093/nar/28.1.235.

(5) Zhang, J.; Aizawa, M.; Amari, S.; Iwasawa, Y.; Nakano, T.; Nakata, K. Development of KiBank, a Database Supporting Structure-Based Drug Design. *Computational Biology and Chemistry* **2004**, *28* (5), 401–407. https://doi.org/10.1016/j.compbiolchem.2004.09.003.

(6) *UCSF Jain Laboratory*. http://www.jainlab.org/downloads.html (accessed 2018-08-27).

(7) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *Journal of Cheminformatics* **2011**, *3* (1), 33. https://doi.org/10.1186/1758-2946-3-33.

(8) *ACD/ChemSketch, Version 12.0*; Advanced Chemistry Development, Inc.: Toronto, ON, Canada, 2012.

(9) Truchon, J.-F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *Journal of Chemical Information and Modeling* **2007**, *47* (2), 488–508. https://doi.org/10.1021/ci600426e.

(10) Jain, A. N. Bias, Reporting, and Sharing: Computational Evaluations of Docking Methods. *J Comput Aided Mol Des* **2008**, *22* (3–4), 201–212. https://doi.org/10.1007/s10822-007-9151-x.

(11) Stroganov, O. V.; Novikov, F. N.; Zeifman, A. A.; Stroylov, V. S.; Chilov, G. G. TSAR, a New Graph-Theoretical Approach to Computational Modeling of Protein Side-Chain Flexibility: Modeling of Ionization Properties of Proteins. *Proteins* **2011**, *79* (9), 2693–2710. https://doi.org/10.1002/prot.23099.

(12) Galvani, E.; Peters, G. J.; Giovannetti, E. Thymidylate Synthase Inhibitors for Non-Small Cell Lung Cancer. *Expert Opinion on Investigational Drugs* **2011**, *20* (10), 1343–1356. https://doi.org/10.1517/13543784.2011.617742.

(13) Rose, M. G.; Farrell, M. P.; Schmitz, J. C. Thymidylate Synthase: A Critical Target for Cancer Chemotherapy. *Clin Colorectal Cancer* **2002**, *1* (4), 220–229. https://doi.org/10.3816/CCC.2002.n.003.

(14) Radwan, A. S.; Khalid, M. A. A. Synthesis, Docking, and Anticancer Activity of New Thiazole Clubbed Thiophene, Pyridine, or Chromene Scaffolds: Synthesis, Docking, and Anticancer Activity of Thiazole Clubbed Thiophene, Pyridine, or Chromene. *J. Heterocyclic Chem.* **2019**, *56* (3), 1063–1074. https://doi.org/10.1002/jhet.3493.
(15) Lu, L.; Li, J.; Moussaoui, M.; Boix, E. Immune Modulation by Human Secreted RNases at the Extracellular Space. *Front Immunol* **2018**, *9*. https://doi.org/10.3389/fimmu.2018.01012.
(16) Russo, N.; Shapiro, R. Potent Inhibition of Mammalian Ribonucleases by 3′,5′-Pyrophosphate-Linked Nucleotides. *J. Biol. Chem.* **1999**, *274* (21), 14902–14908. https://doi.org/10.1074/jbc.274.21.14902.
(17) Shepard, S. M.; Windsor, I. W.; Raines, R. T.; Cummins, C. C. Nucleoside Tetra- and Pentaphosphates Prepared Using a Tetraphosphorylation Reagent Are Potent Inhibitors of Ribonuclease A. *J. Am. Chem. Soc.* **2019**, *141* (46), 18400–18404. https://doi.org/10.1021/jacs.9b09760.
(18) *Chemoinformatics Approaches to Virtual Screening*; Varnek, A., Tropsha, A., Royal Society of Chemistry (Great Britain), Eds.; RSC Pub: Cambridge, 2008.
(19) Herceg, Z.; Wang, Z.-Q. Functions of Poly(ADP-Ribose) Polymerase (PARP) in DNA Repair, Genomic Integrity and Cell Death. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **2001**, *477* (1), 97–110. https://doi.org/10.1016/S0027-5107(01)00111-7.
(20) Research, C. for D. E. and. FDA Approves Olaparib for Germline BRCA-Mutated Metastatic Breast Cancer. *FDA* **2019**.
(21) Research, C. for D. E. and. FDA Approves Niraparib for HRD-Positive Advanced Ovarian Cancer. *FDA* **2019**.
(22) Hilton, J. F.; Hadfield, M. J.; Tran, M.-T.; Shapiro, G. I. Poly(ADP-Ribose) Polymerase Inhibitors as Cancer Therapy. *Front Biosci (Landmark Ed)* **2013**, *18*, 1392–1406. https://doi.org/10.2741/4188.
(23) Berger, N. A.; Besson, V. C.; Boulares, A. H.; Bürkle, A.; Chiarugi, A.; Clark, R. S.; Curtin, N. J.; Cuzzocrea, S.; Dawson, T. M.; Dawson, V. L.; Haskó, G.; Liaudet, L.; Moroni, F.; Pacher, P.; Radermacher, P.; Salzman, A. L.; Snyder, S. H.; Soriano, F. G.; Strosznajder, R. P.; Sümegi, B.; Swanson, R. A.; Szabo, C. Opportunities for the Repurposing of PARP Inhibitors for the Therapy of Non-oncological Diseases. *Br J Pharmacol* **2018**, *175* (2), 192–222. https://doi.org/10.1111/bph.13748.
(24) Li, J.; Zhou, N.; Cai, P.; Bao, J. In Silico Screening Identifies a Novel Potential PARP1 Inhibitor Targeting Synthetic Lethality in Cancer Treatment. *Int J Mol Sci* **2016**, *17* (2). https://doi.org/10.3390/ijms17020258.

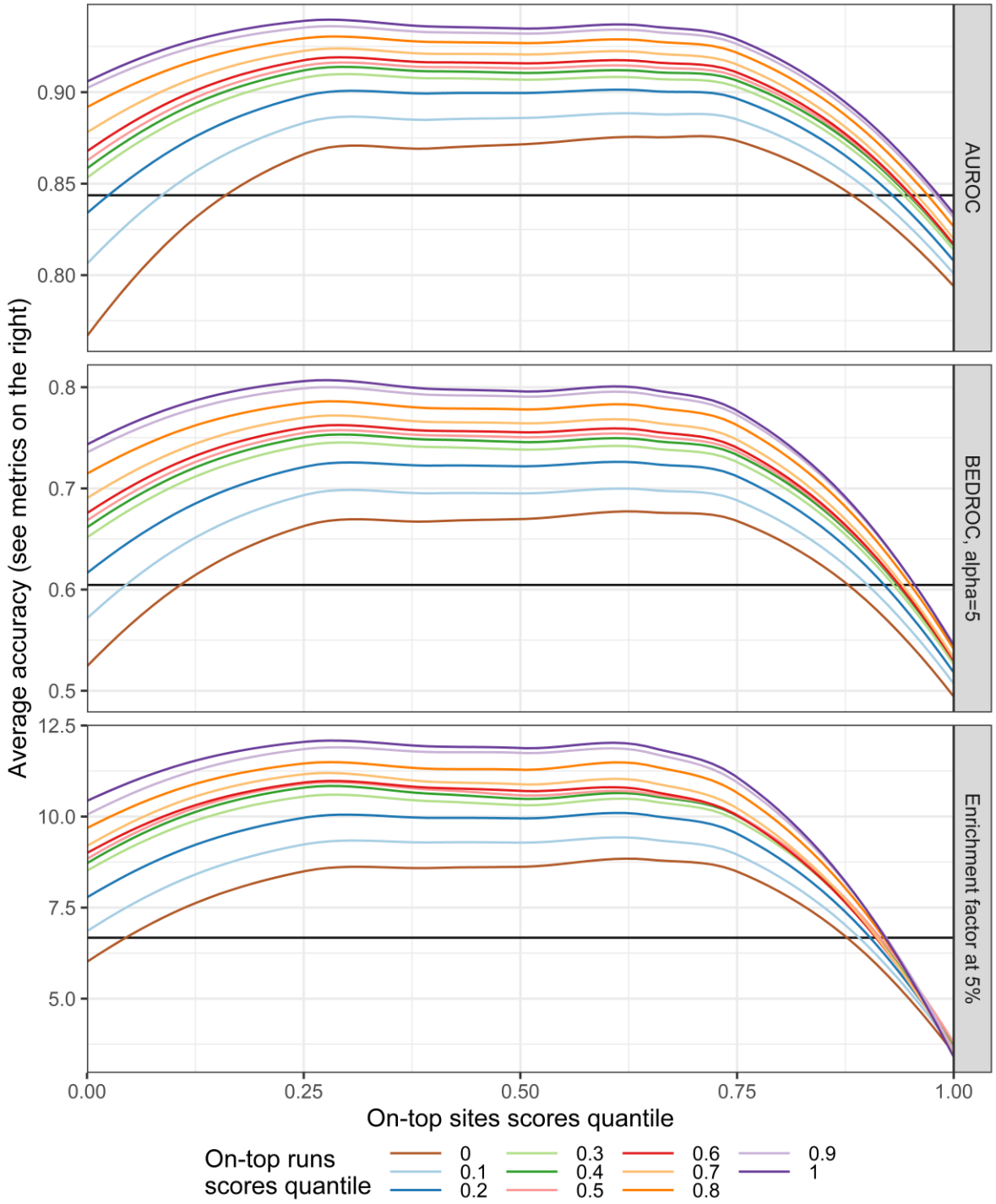

Figure S1.

Dependence of average VS accuracy (by means of AUROC, BEDROC and Enrichment factor) on the quantile of OTH scores in each OTH site and quantile of scores over all OTH sites; black horizontal lines indicate performance of the AS-only docking (ASD).

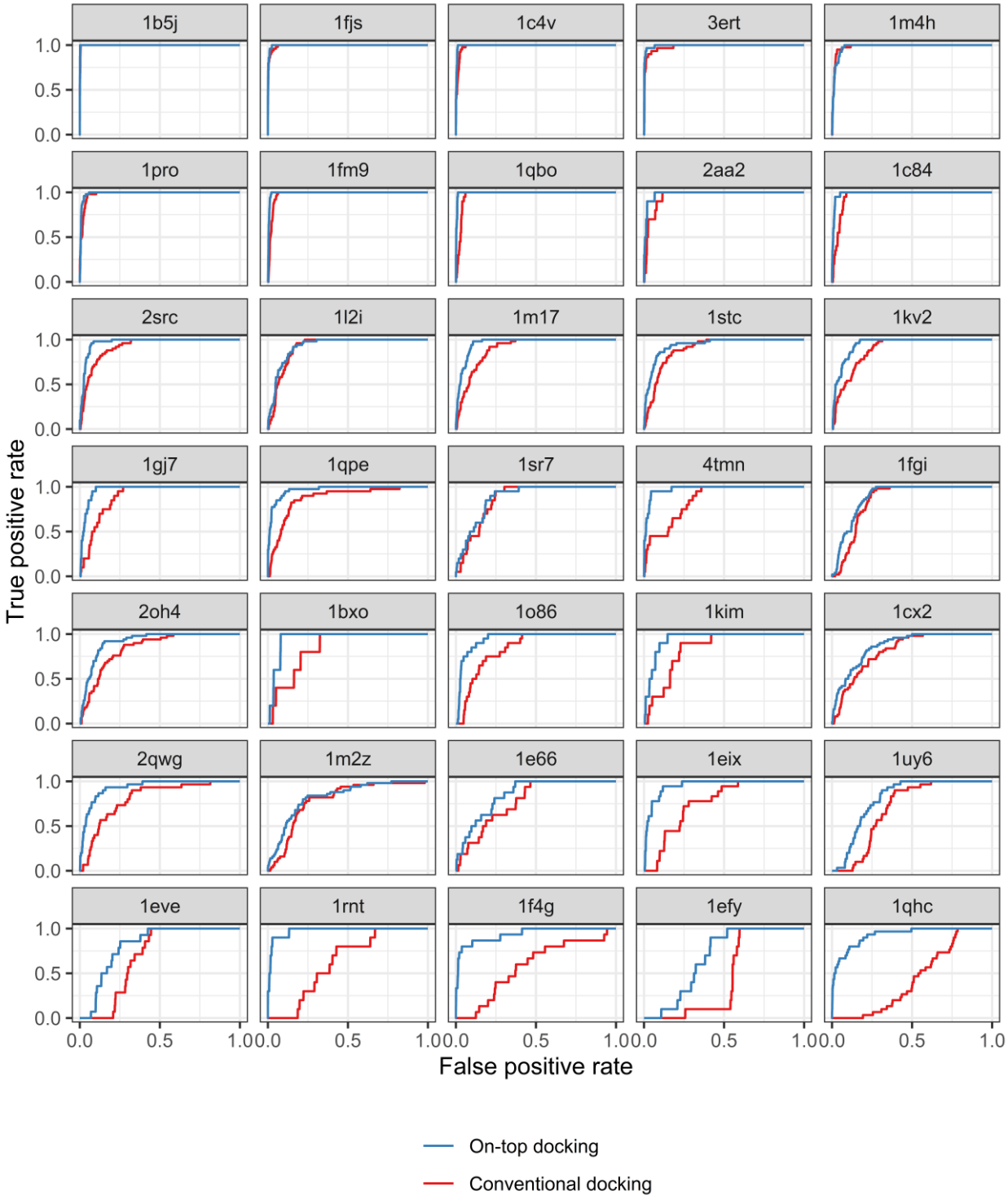

Figure S2.

Receiver operating characteristic (ROC) curves for on-top and conventional docking procedures for each of studied targets. Targets are designated by their PDB IDs. Targets are sorted by decrease of conventional docking areas under the ROC curves (AUROCs).

Table S1.

Test set proteins data and comparison of conventional and on-top docking procedures performances. Table is sorted by increase of conventional docking accuracy.

Target	PDB ID	Protein type	Conventional docking accuracy (%)	On-top docking accuracy (%)	Impro-vement (%)
Ribonuclease A	1qhc	Ribonuclease	44.6	92.9	48.3
Poly(ADP-ribose) polymerase	1efy	Transferase	46.7	68.0	21.3
Thymidylate synthase	1f4g	Transferase	58.0	93.9	35.9
Ribonuclease T1	1rnt	Ribonuclease	62.7	97.7	34.9
Acetylcholinesterase	1eve	Hydrolase	68.8	81.3	12.6
Heat shock protein HSP 90-alpha	1uy6	Hydrolase	70.4	81.0	10.6
Orotidine-5'-P decarboxylase	1eix	Lyase	76.1	95.3	19.1
Acetylcholinesterase	1e66	Hydrolase	78.1	85.3	7.2
Glucocorticoid receptor	1m2z	Nuclear receptor	79.1	82.8	3.7
Neuraminidase	2qwg	Hydrolase	81.0	93.6	12.5
Cyclooxygenase-2	1cx2	Oxidoreductase	82.0	87.1	5.2
Thymidine kinase	1kim	Kinase	84.0	94.7	10.7
Angiotensin-converting enzyme	1o86	Hydrolase	84.1	94.8	10.7
Penicillopepsin	1bxo	Protease	84.5	95.2	10.7
Vascular endothelial growth factor receptor kinase 2	2oh4	Kinase	84.6	91.9	7.3
Fibroblast growth factor receptor kinase	1fgi	Kinase	85.2	89.0	3.9
Thermolysin	4tmn	Protease	85.9	97.7	11.8
Progesteron receptor	1sr7	Nuclear receptor	86.6	88.1	1.5
Tyrosine-protein kinase Lck	1qpe	Kinase	87.5	96.9	9.4
P38 MAP kinase	1kv2	Kinase	89.2	94.9	5.7
Urokinase-type plasminogen activator	1gj7	Hydrolase	89.2	96.8	7.6
cAMP-dependent protein kinase	1stc	Kinase	89.8	94.1	4.3
Epidermal growth factor receptor kinase	1m17	Kinase	90.0	96.2	6.2
Estrogen receptor	1l2i	Nuclear receptor	91.7	92.9	1.2
Tyrosine kinase c-Src.	2src	Kinase	92.2	96.9	4.7
Protein tyrosine phosphatase 1B	1c84	Hydrolase	96.1	99.0	2.9
Mineralocorticoid receptor	2aa2	Nuclear receptor	96.2	98.5	2.4
Trypsin	1qbo	Protease	97.3	99.6	2.3
Peroxisome proliferator activated receptor gamma	1fm9	Nuclear receptor	97.9	99.3	1.4
HIV-1 protease	1pro	Protease	98.4	99.1	0.7
Beta-secretase	1m4h	Protease	98.5	98.2	−0.3
Estrogen receptor	3ert	Nuclear receptor	98.5	99.5	1.0
Thrombin	1c4v	Protease	98.9	99.6	0.7
Factor Xa	1fjs	Protease	99.4	99.7	0.3
Oligopeptide-binding protein	1b5j	None	99.9	99.9	0.1
Average			84.4	93.5	9.1

Target	PDB ID	Type	Open active site	Total charge at pH=7	AUROC			BEDROC (α=5)			Enrichment factor at 5%		
					Convent. docking	On-top docking	Gain	Convent. docking	On-top docking	Gain	Convent. docking	On-top docking	Gain
Ribonuclease A	1qhc	Ribonuclease	yes	12	0.45	0.93	0.48	0.09	0.79	0.70	0.00	12.00	12.00
Poly(ADP-ribose) polymerase	1efy	Transferase	yes	3	0.47	0.68	0.21	0.08	0.24	0.16	0.00	0.00	0.00
Thymidylate synthase	1f4g	Transferase	yes	-18	0.58	0.94	0.36	0.21	0.83	0.62	0.00	16.00	16.00
Ribonuclease T1	1rnt	Ribonuclease	yes	-28	0.63	0.98	0.35	0.20	0.90	0.70	0.00	18.00	18.00
Acetylcholinesterase	1eve	Hydrolase	no	-2	0.69	0.81	0.13	0.23	0.44	0.22	0.00	0.00	0.00
Heat shock protein HSP 90-alpha	1uy6	Hydrolase	yes	-5	0.70	0.81	0.11	0.27	0.44	0.17	0.00	0.67	0.67
Orotidine-5вЪІ-P decarboxylase	1eix	Lyase	no	-10	0.76	0.95	0.19	0.38	0.82	0.44	0.00	12.22	12.22
Acetylcholinesterase	1e66	Hydrolase	no	-9	0.78	0.85	0.07	0.44	0.57	0.13	3.75	6.25	2.50
Glucocorticoid receptor	1m2z	Nuclear receptor	no	-1	0.79	0.83	0.04	0.47	0.57	0.10	2.00	3.60	1.60
Neuraminidase	2qwg	Hydrolase	yes	-3	0.81	0.94	0.13	0.51	0.79	0.28	1.33	10.00	8.67
Cyclooxygenase-2	1cx2	Oxidoreductase	no	-2	0.82	0.87	0.05	0.52	0.64	0.11	2.72	6.80	4.08
Thymidine kinase	1kim	Kinase	no	-13	0.84	0.95	0.11	0.51	0.79	0.27	4.00	10.00	6.00
Angiotensin-converting enzyme	1o86	Hydrolase	no	-7	0.84	0.95	0.11	0.52	0.80	0.27	2.00	14.00	12.00
Penicillopepsin	1bxo	Protease	yes	-22	0.85	0.95	0.11	0.53	0.79	0.27	4.00	12.00	8.00
Vascular endothelial growth factor receptor kinase 2	2oh4	Kinase	yes	-2	0.85	0.92	0.07	0.58	0.73	0.16	3.92	6.80	2.88
Fibroblast growth factor receptor kinase	1fgi	Kinase	yes	3	0.85	0.89	0.04	0.53	0.64	0.11	0.80	4.80	4.00
Thermolysin	4tmn	Protease	yes	0	0.86	0.98	0.12	0.59	0.91	0.31	9.00	16.00	7.00
Progesteron receptor	1sr7	Nuclear receptor	no	5	0.87	0.88	0.01	0.57	0.62	0.05	5.00	6.00	1.00
Tyrosine-protein kinase Lck	1qpe	Kinase	yes	-7	0.87	0.97	0.09	0.65	0.89	0.24	6.00	15.00	9.00
Urokinase-type plasminogen activator	1gj7	Hydrolase	yes	-7	0.89	0.97	0.08	0.63	0.86	0.23	4.00	14.00	10.00
P38 MAP kinase	1kv2	Kinase	yes	8	0.89	0.95	0.06	0.66	0.81	0.15	6.80	10.00	3.20
cAMP-dependent protein kinase	1stc	Kinase	no	-6	0.90	0.94	0.04	0.68	0.81	0.13	5.20	9.20	4.00
Epidermal growth factor receptor kinase	1m17	Kinase	yes	-3	0.90	0.96	0.06	0.68	0.85	0.17	6.00	10.32	4.32
Estrogen receptor	1l2i	Nuclear receptor	no	-5	0.92	0.93	0.01	0.70	0.75	0.05	4.72	5.92	1.20
Tyrosine kinase c-Src.	2src	Kinase	yes	-1	0.92	0.97	0.05	0.74	0.88	0.14	8.32	11.52	3.20
Protein tyrosine phosphatase 1B	1c84	Hydrolase	yes	-5	0.96	0.99	0.03	0.83	0.95	0.12	11.00	19.00	8.00
Mineralocorticoid receptor	2aa2	Nuclear receptor	no	1	0.96	0.99	0.02	0.84	0.93	0.09	14.00	18.00	4.00
Trypsin	1qbo	Protease	yes	26	0.97	1.00	0.02	0.88	0.98	0.10	15.00	20.00	5.00
Peroxisome proliferator activated receptor gamma	1fm9	Nuclear receptor	no	-2	0.98	0.99	0.01	0.91	0.97	0.06	12.72	17.20	4.48
HIV-1 protease	1pro	Protease	no	7	0.98	0.99	0.01	0.93	0.96	0.03	13.92	17.12	3.20
Beta-secretase	1m4h	Protease	yes	-13	0.98	0.98	0.00	0.94	0.92	-0.01	16.15	15.50	-0.65
Estrogen receptor	3ert	Nuclear receptor	no	-1	0.98	0.99	0.01	0.94	0.98	0.03	17.87	19.33	1.47
Thrombin	1c4v	Protease	yes	10	0.99	1.00	0.01	0.95	0.98	0.03	16.15	20.00	3.85
Factor Xa	1fjs	Protease	yes	-7	0.99	1.00	0.00	0.98	0.99	0.01	17.20	17.92	0.72
Oligopeptide-binding protein	1b5j	None	no	-8	1.00	1.00	0.00	0.99	1.00	0.00	20.00	20.00	0.00
Average value				**-3.2**	**0.84**	**0.93**	**0.09**	**0.60**	**0.80**	**0.19**	**6.67**	**11.86**	**5.19**