MCSS-based Predictions of Binding Mode and Selectivity of Nucleotide Ligands

Roy González-Alemán,[†,‡] Nicolas Chevrollier,[†] Manuel Simoes,[¶] Luis Montero-Cabrera,[‡] and Fabrice Leclerc[*,†]

†*Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Université Paris Saclay, Gif-sur-Yvette, F-91198, France*

‡*Laboratorio de Química Computacional y Teórica (LQCT), Facultad de Química, Universidad de La Habana, 10400 La Habana, Cuba*

¶*CPC Manufacturing Analytics, Strasbourg, France*

E-mail: fabrice.leclerc@i2bc.paris-saclay.fr

Phone: +33 (0)1 69 82 62 39

Contents

Supporting Information Available

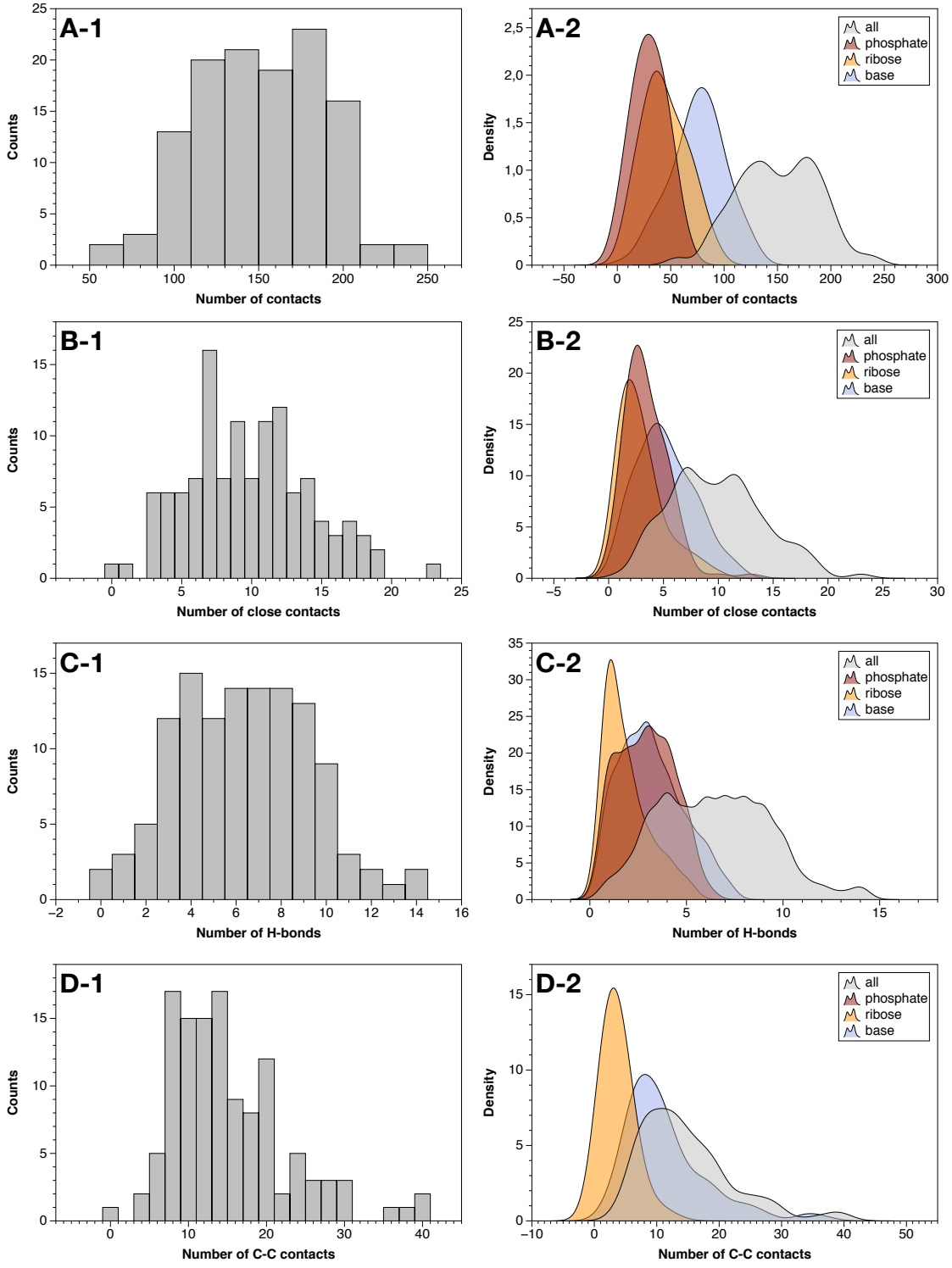
Benchmark of 121 protein-nucleotide complexes

1. Attached Supplementary Data 1 (Data-S1.csv): a list of PDB IDs including the ligand ID, the atomic resolution, functional classification, and EC number.

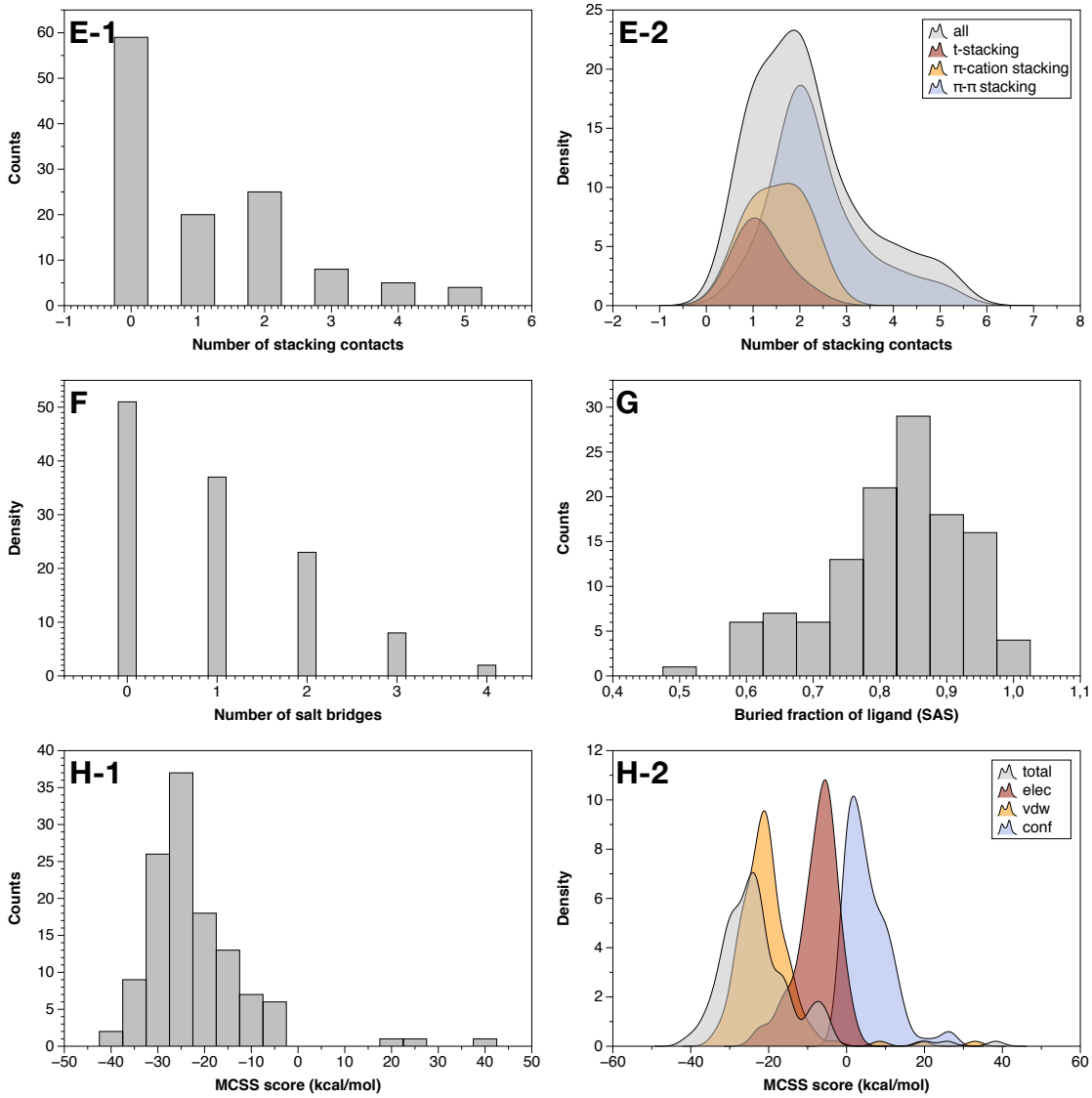
2. Attached Supplementary Data 2 (Data-S2.csv): calculations of the BINANA features (number of contacts, number of H-bonds, the buried fraction of ligand, etc)

3. Attached Supplementary Data 3 (Data-S3.csv): calculations of the NACCESS surface terms for the fraction of buried surface of the ligand

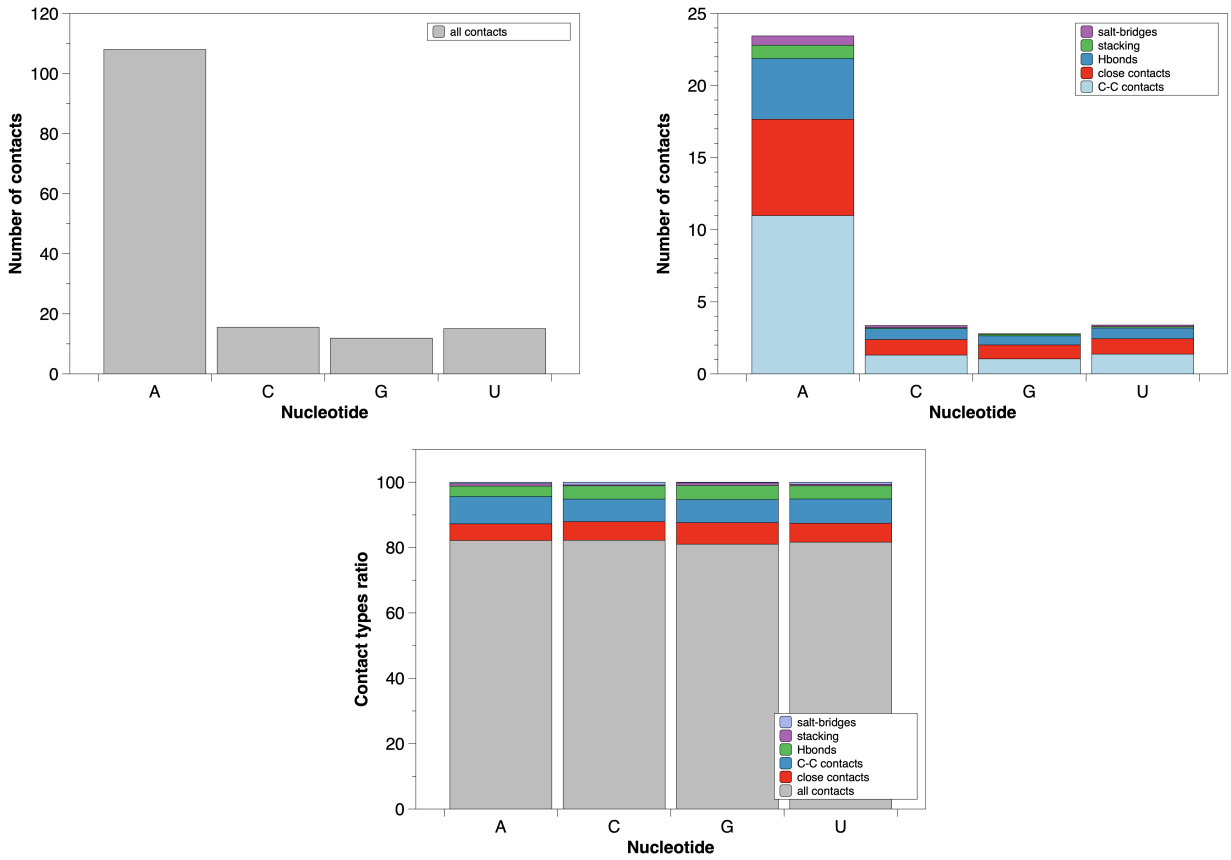
4. Attached Supplementary Data 4 (Data-S4.tar.gz): 2D diagrams of the contacts within the binding sites (SVG format).



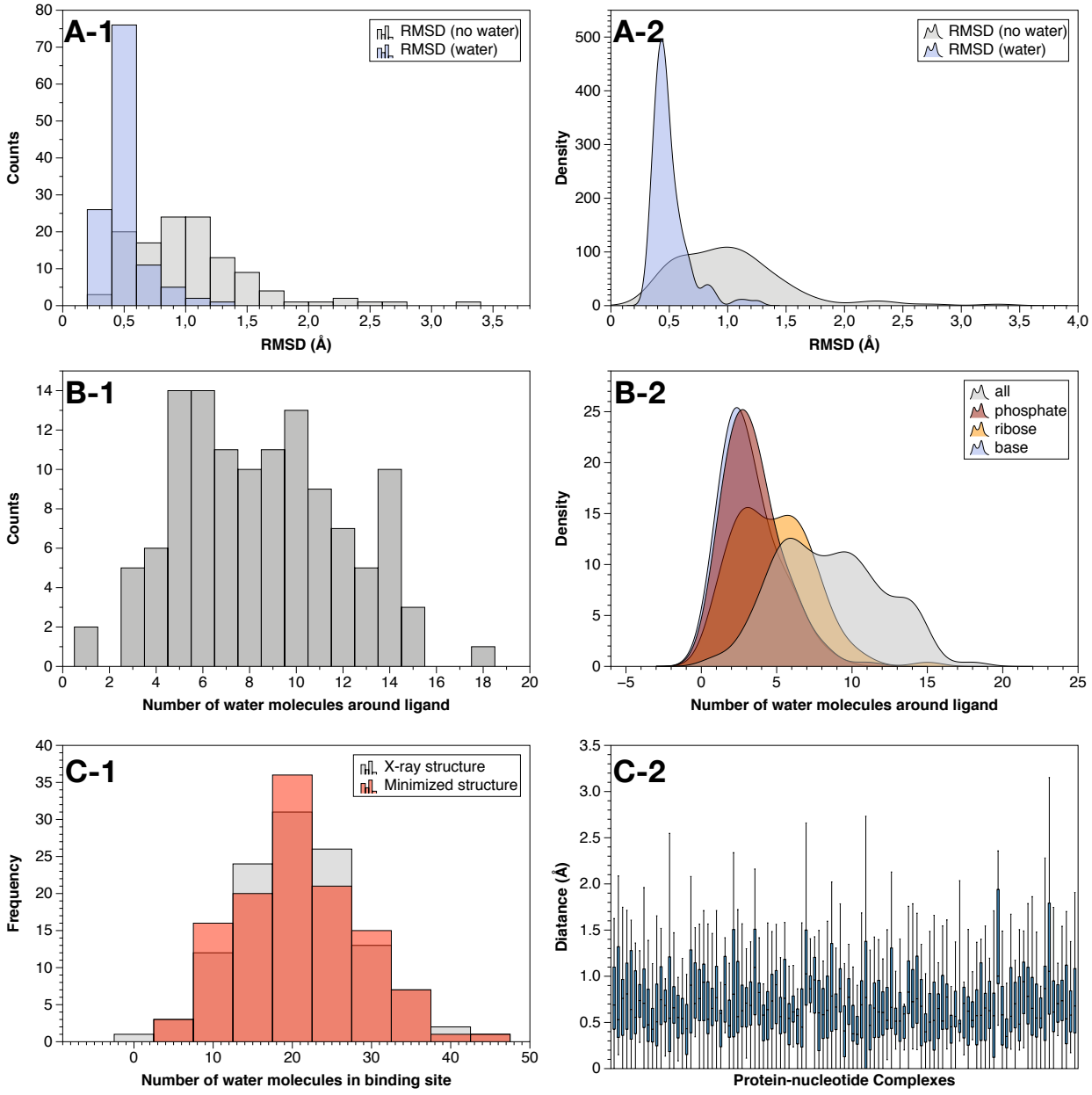
Supplementary Figure 1: Molecular and energy features of the nucleotide-binding sites from the benchmark of 121 complexes. A-1.: Histogram of the number of contacts; A-2: Smooth histogram with decomposition per nucleotide moiety (base, ribose, phosphate); B-1.: Histogram of the number of close contacts; B-2.: Same as A-2 for close contacts; C-1.: Histogram of the number of H-bonds; C-2.: Same as A-2 for H-bonds; D-1.: Histogram of the number of C-C contacts; D-2.: Same as A-2 for C-C contacts; (to be continued).



Supplementary Figure 1: Molecular and energy features of the nucleotide-binding sites from the benchmark of 121 complexes (continued). E-1.: Histogram of the number of stacking contacts; E-2.: Smooth histogram with decomposition per stacking types; F.: Histogram of the number of salt-bridges; G.: Histogram of the buried fraction of ligand (calculated from the solvent accessible surface); H-1.: Histogram of the MCSS scores calculated for the ligands optimized in their binding site; H-2.: Smooth histogram with decomposition per contribution types (electrostatics, van der Waals, conformational). The molecular descriptors associated with the atomic contacts are calculated by BINANA;[?] the stacking contributions are calculated from OpenEye;[?] the MCSS score is calculated by the scoring function derived previously.[?]



Supplementary Figure 2: Nucleotide breakdown of atomic contacts. Top-left: all contacts; top-right: specific contacts (C-C contacts, close contacts, Hbonds, stacking contacts, salt-bridges); bottom: ratio of each type of specific contacts. The number of contacts correspond to the average value over the full benchmark.



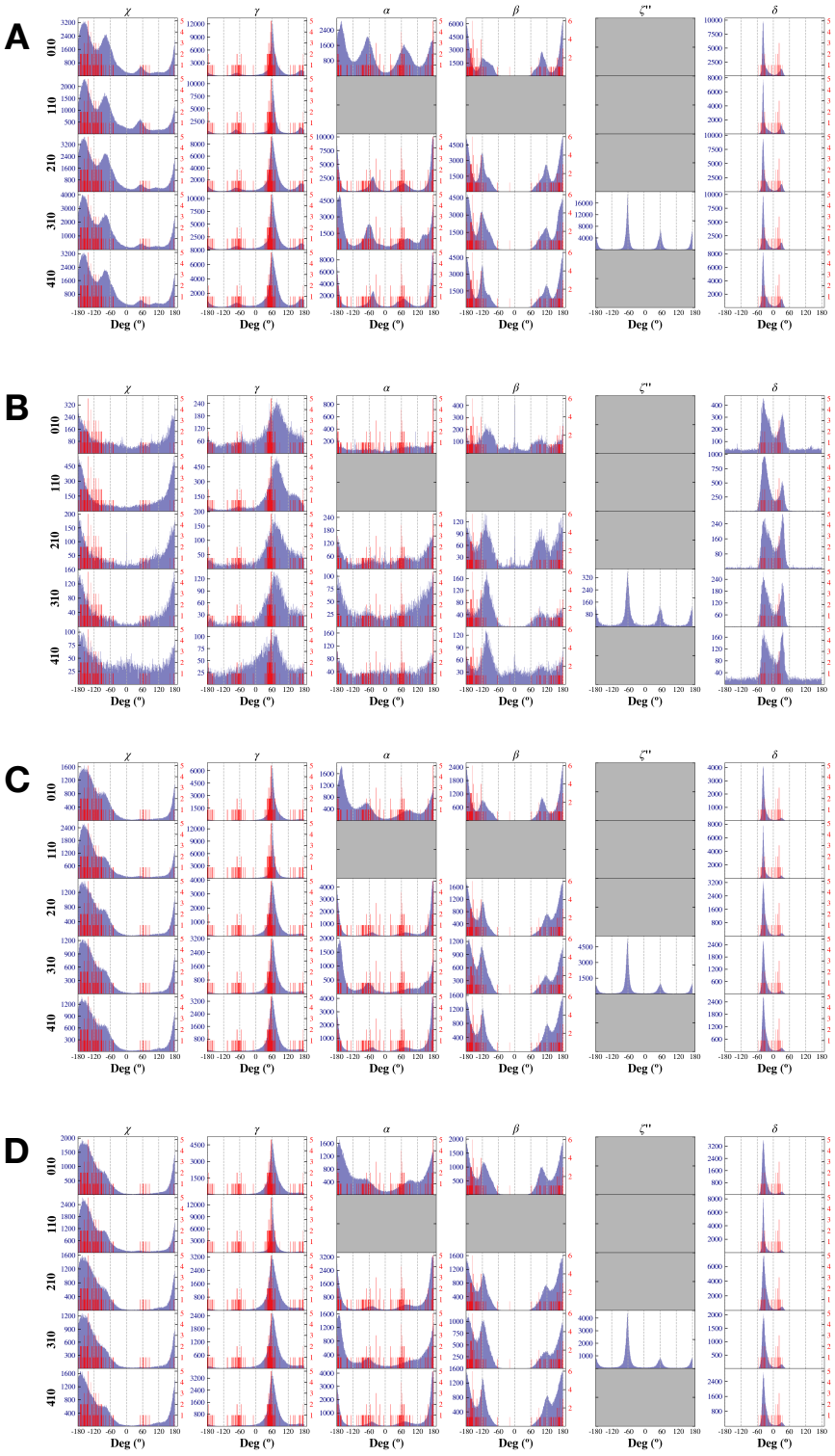
Supplementary Figure 3: Distributions of water molecules and impact on the binding sites. A-1.: Histogram of RMSD in presence/absence of water molecules; A-2.: Same as A-1 with a smooth histogram; B-1.: the number of water molecules around the ligand (distance cutoff of 4.0Å); B-2.: Same as B-1 with decomposition per nucleotide moiety; C-1.: Number of water molecules within the binding site as defined in MCSS by the box parameters (see Methods); C-2.: displacements (Å) of water molecules from their crystallized positions.

MCSS

5. Attached Supplementary Data 5: MCSS input sample (Data-S5.txt)

6. Attached Supplementary Data 6: MCSS nonbonded parameters sample (Data-S6.txt)

7. Attached Supplementary Data 7 (Data-S7.csv): MCSS score (including its VdW and elec terms) and RMSD values for each protein-nucleotide complex

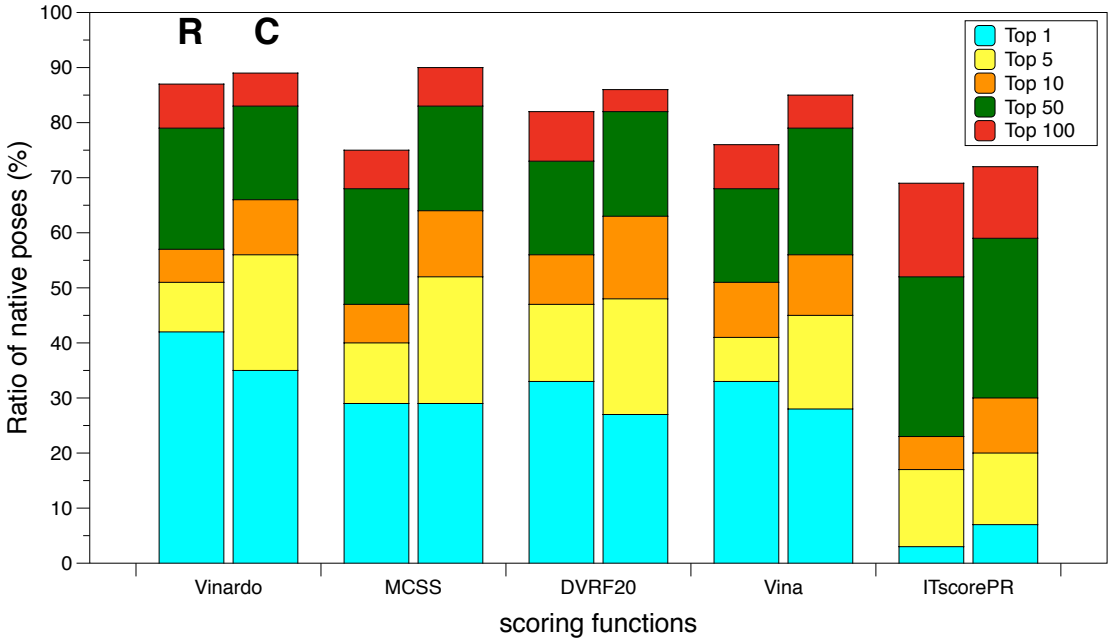


Supplementary Figure 4: Torsions angles. Nonbonded models and associated patches (R010 to R410): A. SCAL, B. FULLW, C. SCALW, D. STDW. In blue: the distribution of the torsions angles observed in the MCSS minima; In red: the distribution of the torsions angles observed in the bound ligands.

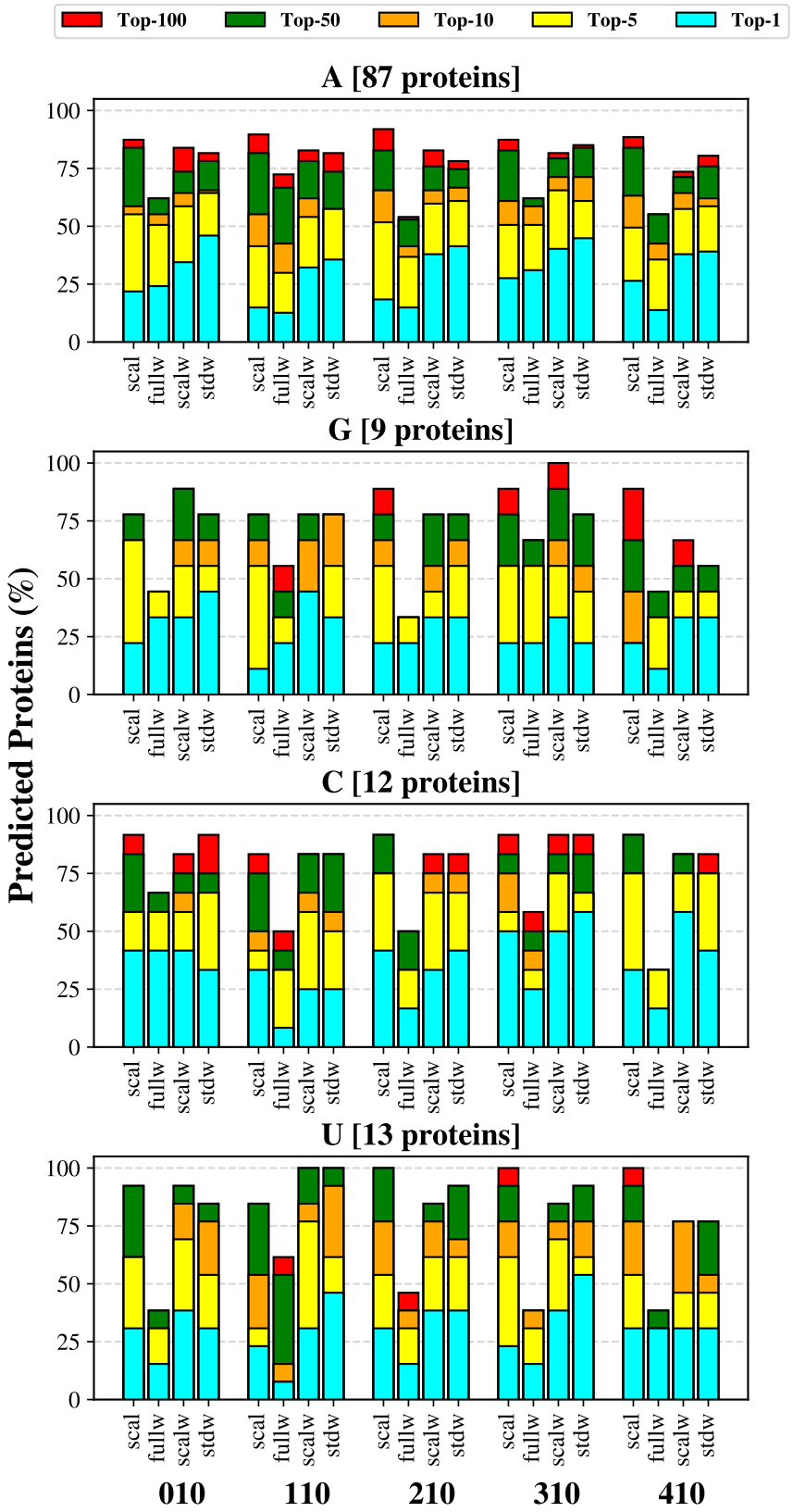
Scoring

Autodock Vina is a well-known docking method used for virtual screening; the associated scoring function is pretty robust, having regularly been used in the comparative assessment of scoring functions (CASF) challenges.[?] Vinardo and $\Delta_{vina}RF_{20}$ were both derived from Vina and tested in the CASF-2013 challenge. Vinardo was optimized and validated on large datasets.[?] It was tested in particular on the DUD library that contains, among other proteins, kinases with nucleotide ligands or nucleotide analogs.[?] $\Delta_{vina}RF_{20}$ was derived more recently from Vina with a new parametrization based on random forest. The performance of $\Delta_{vina}RF_{20}$ was superior to that of Vina when tested on the CASF-2007 and CASF-2013 challenges benchmarks. Finally, ITscorePR was included since it has been specifically developed for protein-RNA interactions. The scores calculated with all the scoring functions: ITscorePR,[?] $\Delta_{vina}RF_{20}$,[?] Autodock Vina score,[?] Vinardo,[?] and the MM-GB models (see Methods) except MCSS[?] correspond to single-point calculations on the MCSS-generated poses.

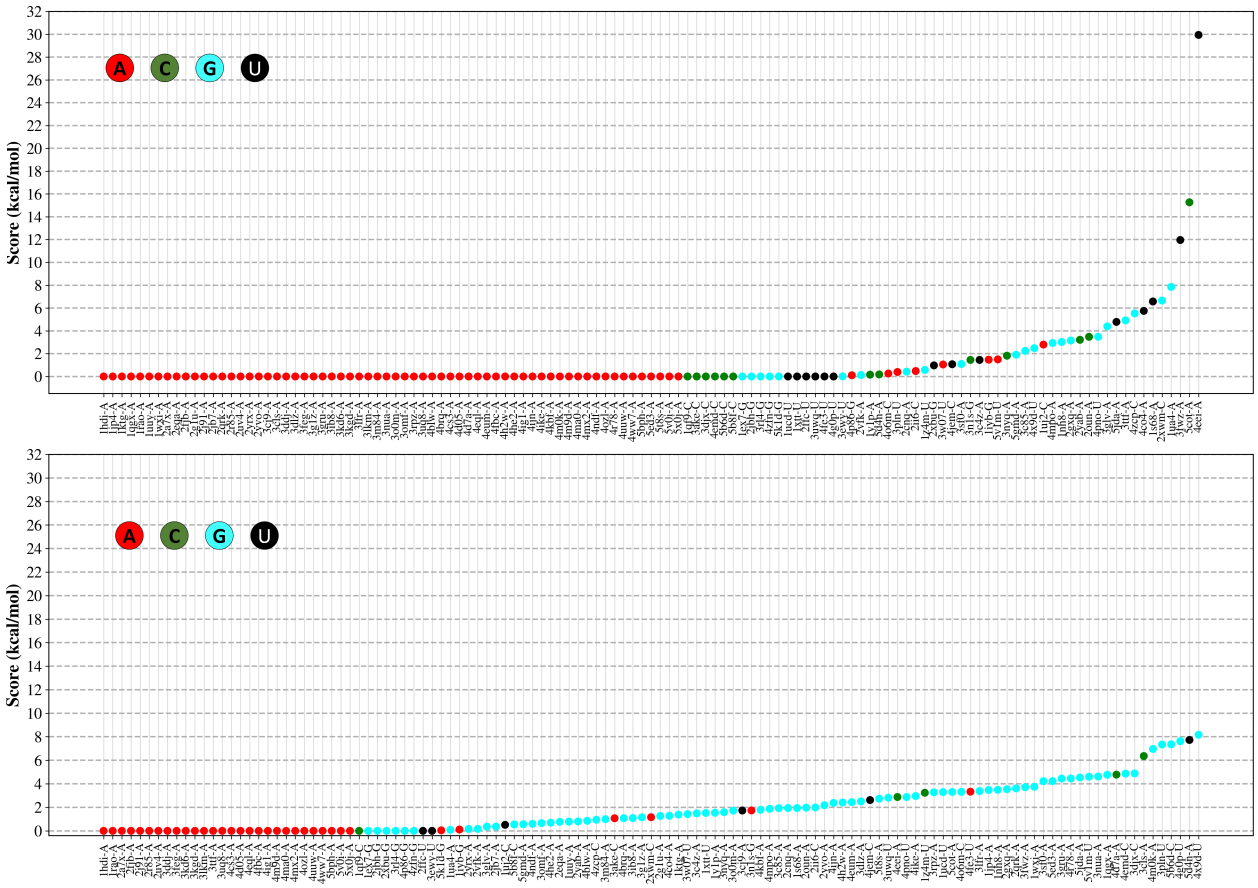
8. Attached Supplementary Data 8 (Data-S8.tar.gz): selectivity diagrams SCAL/STDW for the native poses for each protein-nucleotide complex of the benchmark.



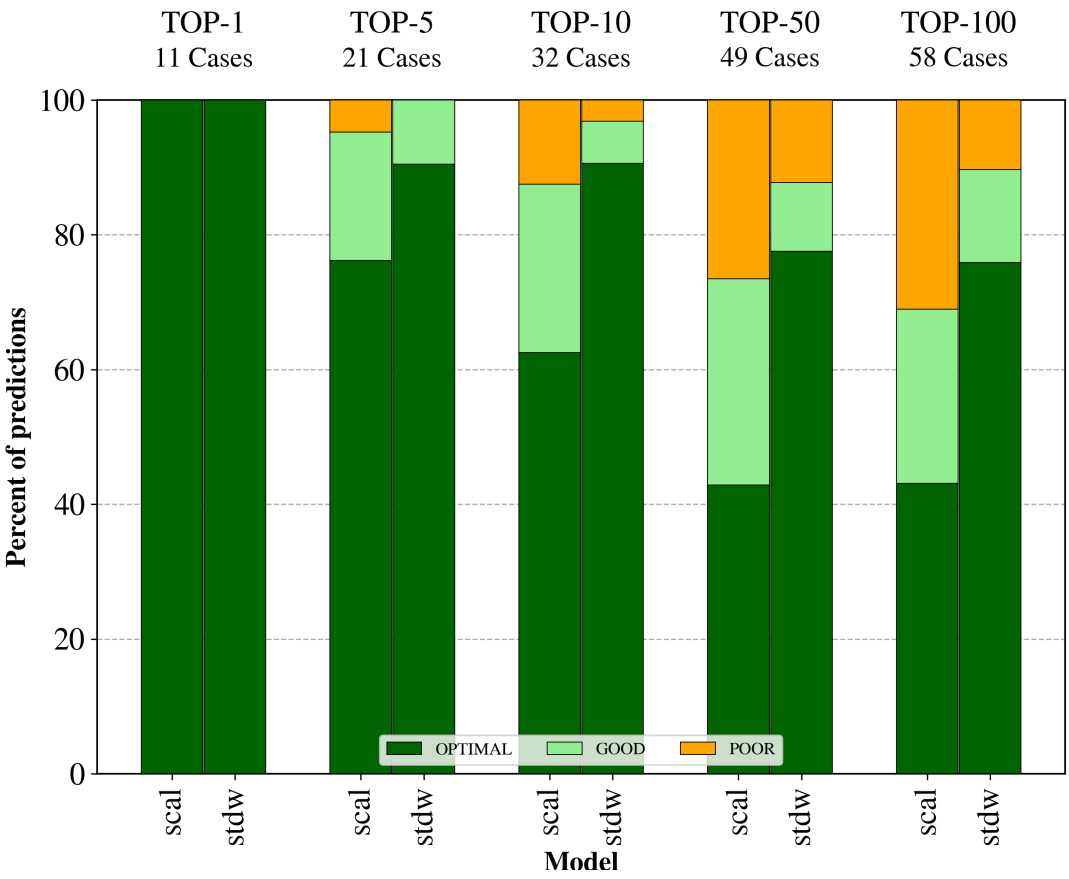
Supplementary Figure 5: Docking powers (top1 to top100) for Vinardo, MCSS, $\Delta_{vina}RF_{20}$, Vina, and ITscorePR and the impact of the clustering filtering (using the patch R310). Left bar (R): no clustering; Right bar (C): clustering.



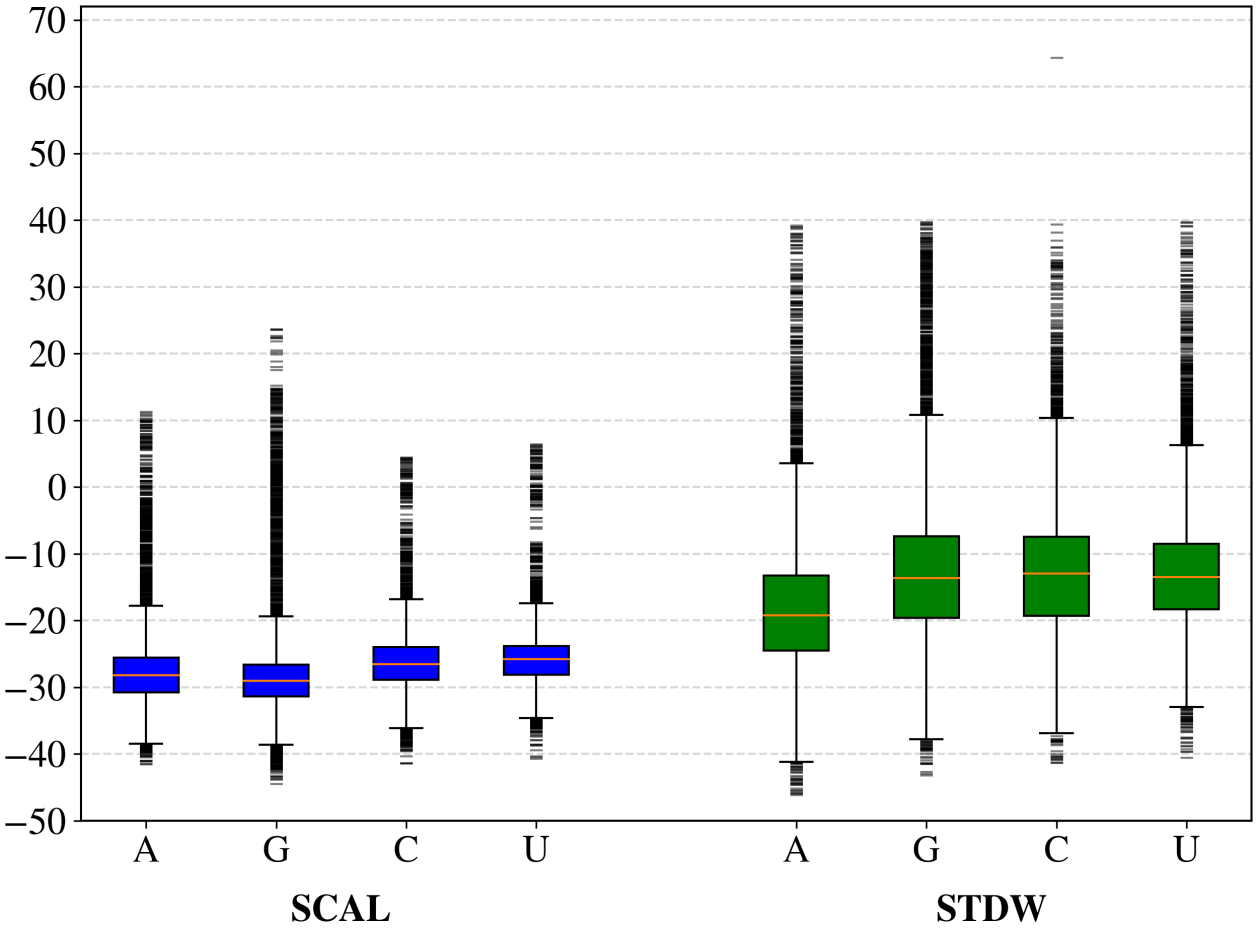
Supplementary Figure 6: Decomposition of docking powers per nucleotide type. The data are shown for the clustered distribution and each Top-i.



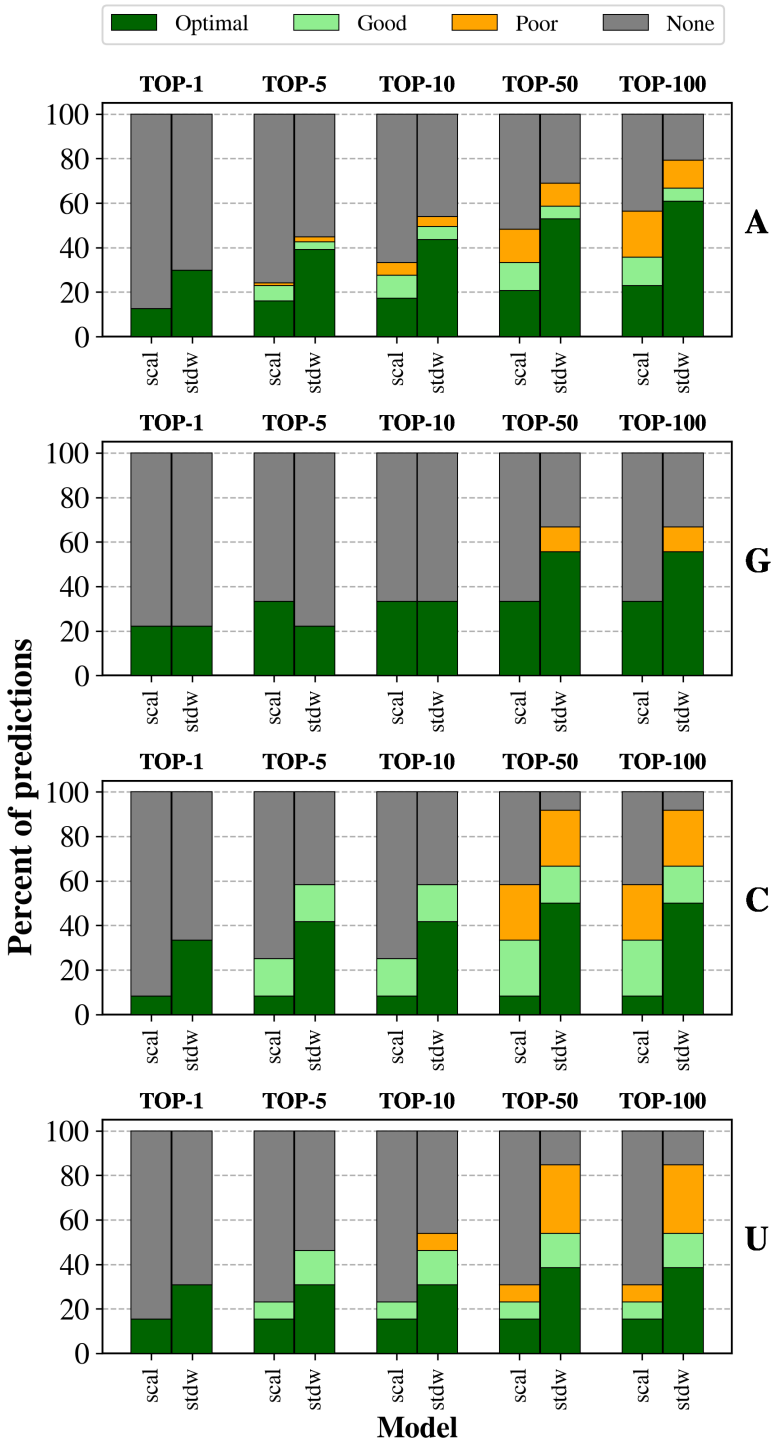
Supplementary Figure 7: Scoring differences (offset) between the best-ranked pose whatever the nucleotide type and the best-ranked pose for the nucleotide corresponding to the native ligand. Top: STDW model; bottom: SCAL model. The color code indicates the nucleotide type.



Supplementary Figure 8: Screening powers on the benchmark subset corresponding to the predictions common to the SCAL and STW models. Optimal: native nucleotide as the best ranked; good: native nucleotide in the ranked within a 2 kcal/mol range from the best ranked non-native nucleotide; poor: native nucleotide ranked out of the 2 kcal/mol range.



Supplementary Figure 9: Distributions of the nucleotide-dependent MCSS score for the SCAL or STDW models (R310).

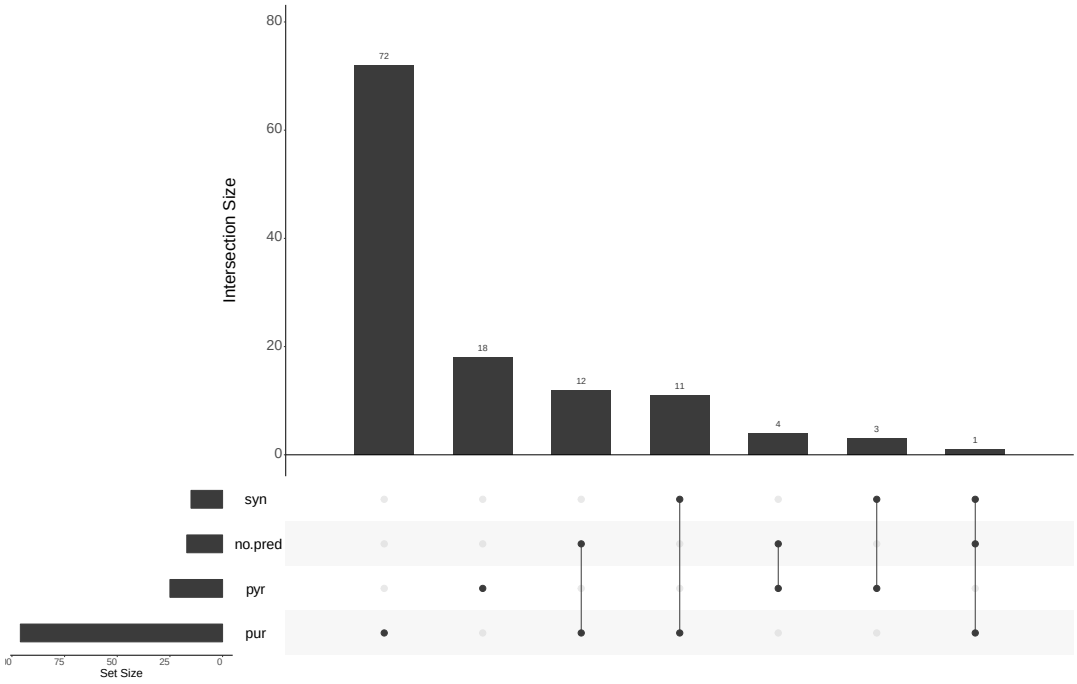


Supplementary Figure 10: Decomposition of screening powers per nucleotide type. Optimal: native nucleotide as the best ranked; good: native nucleotide in the ranked within a 2 kcal/mol range from the best ranked non-native nucleotide; poor: native nucleotide ranked out of the 2 kcal/mol range.

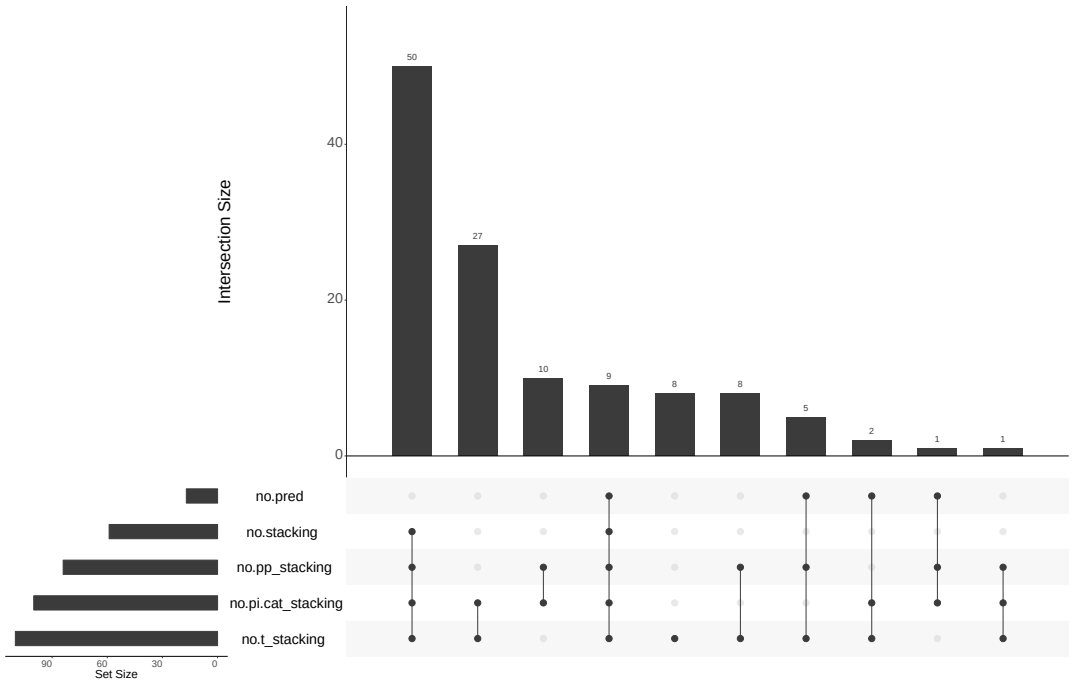
Molecular features

Supplementary Table 1: Frequencies of occurences for molecular features in the Top-10 non-predicted cases versus benchmark. Others: presence of additional nucleotidic (nucleic acid) fragment in the binsing site; metals: presence of metal(s) in the binding site; nwat.low: presence of number of water molecules below the threshold value; vol.low: volume of the binding site below the threshold value; syn: syn conformation of the nucleic acid base; pyr: pyrimidine; pur: purine; no.base.contacts: absence of contacts with the nucleic acid base; clash_aa: clash(es) with amino-acid residues; clash_w: clash(es) with water molecules; no.salt.bridges: absence of salt-bridge; no.stacking: absence of stacking.

Features		Freq. Benchmark	Freq. no.pred
binding site	nwat.low	62	59
	vol.low	69	82
	others	12	6
	metals	36	24
conformational	syn	12	0
	pur	79	71
	pyr	21	23
interaction	no.base.contacts	12	12
	no.salt.bridges	44	59
	no.stacking	49	53
	clash aa	22	18
	clash w	33	41



Supplementary Figure 11: Upset diagram of the impact of the conformational features on the Top-10 predictions. The intersections with only one member are not shown; syn: syn conformation of the nucleic acid base; pyr: pyrimidine; pur: purine.



Supplementary Figure 12: Upset diagram of stacking contributions for the Top-10 predictions. no.pp_satcking: no π-π stacking; no.pi.cat_stacking: no π-cation stacking; no.t_stacking: no t stacking.

Supplementary Table 2: Frequencies of occurences for molecular features in the Top-10 for non-predicted cases of STDW-310 versus benchmark. Others: presence of additional nucleotidic (nucleic acid) fragment in the binsing site; metals: presence of metal(s) in the binding site; nwat.low: presence of number of water molecules below the threshold value; vol.low: volume of the binding site below the threshold value; syn: syn conformation of the nucleic acid base; pyr: pyrimidine; pur: purine; no.base.contacts: absence of contacts with the nucleic acid base; clash_aa: clash(es) with amino-acid residues; clash_w: clash(es) with water molecules; no.salt.bridges: absence of salt-bridge; no.stacking: absence of stacking.

Features		Freq. Benchmark	Freq. STDW(R310)
binding site	nwat.low	62	51
	vol.low	69	72
	others	12	6
	metals	36	30
conformational	syn	12	17
	pur	79	83
	pyr	21	17
interaction	no.base.contacts	12	11
	no.salt.bridges	44	62
	no.stacking	49	49
	clash aa	22	21
	clash w	33	40

Supplementary Table 3: Frequencies of occurences for molecular features in the Top-10 for non-optimal (good) predictions. Others: presence of additional nucleotidic (nucleic acid) fragment in the binsing site; metals: presence of metal(s) in the binding site; nwat.low: presence of number of water molecules below the threshold value; vol.low: volume of the binding site below the threshold value; syn: syn conformation of the nucleic acid base; pyr: pyrimidine; pur: purine; no.base.contacts: absence of contacts with the nucleic acid base; clash_aa: clash(es) with amino-acid residues; clash_w: clash(es) with water molecules; no.salt.bridges: absence of salt-bridge; no.stacking: absence of stacking.

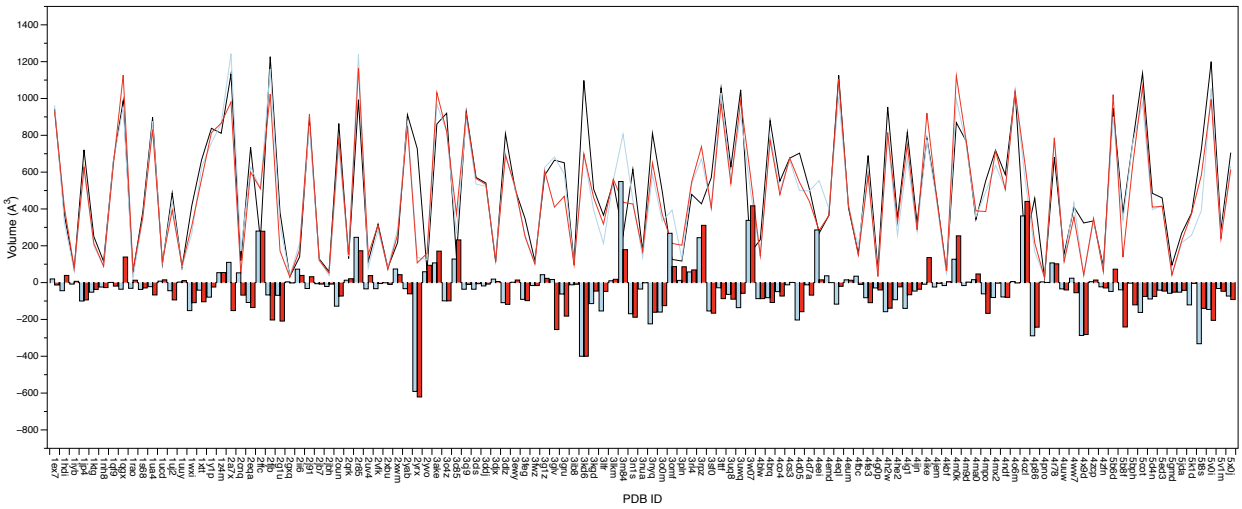
Features		Freq. Benchmark	Freq. good
binding site	nwat.low	62	60
	vol.low	69	70
	others	12	10
	metals	36	60
conformational	syn	12	0
	pur	79	80
	pyr	21	0
interaction	no.base.contacts	12	30
	no.salt.bridges	44	30
	no.stacking	49	70
	clash aa	22	20
	clash w	33	40

Supplementary Table 4: Variations in the binding site's volume for the subset of protein-nucleotides complexes with no prediction in the Top-10. The volume of reference corresponds to that of the experimental structure; the modified volumes are calculated for both the SCAL and STDW models. Only the cases where the variation equals or exceeds 100Å^3 are considered. UP: increase of the binding site's volume. DOWN: decrease of the binding site's volume.

Volumes		Freq. Benchmark	Freq. nopred.
SCAL	UP	12	0
	DOWN	19	18
STDW	UP	13	0
	DOWN	21	35

9. Attached Supplementary Data 9 (Data-S9.txt): raw data corresponding to the number of water molecules around the ligand at a distance up to 4Å.

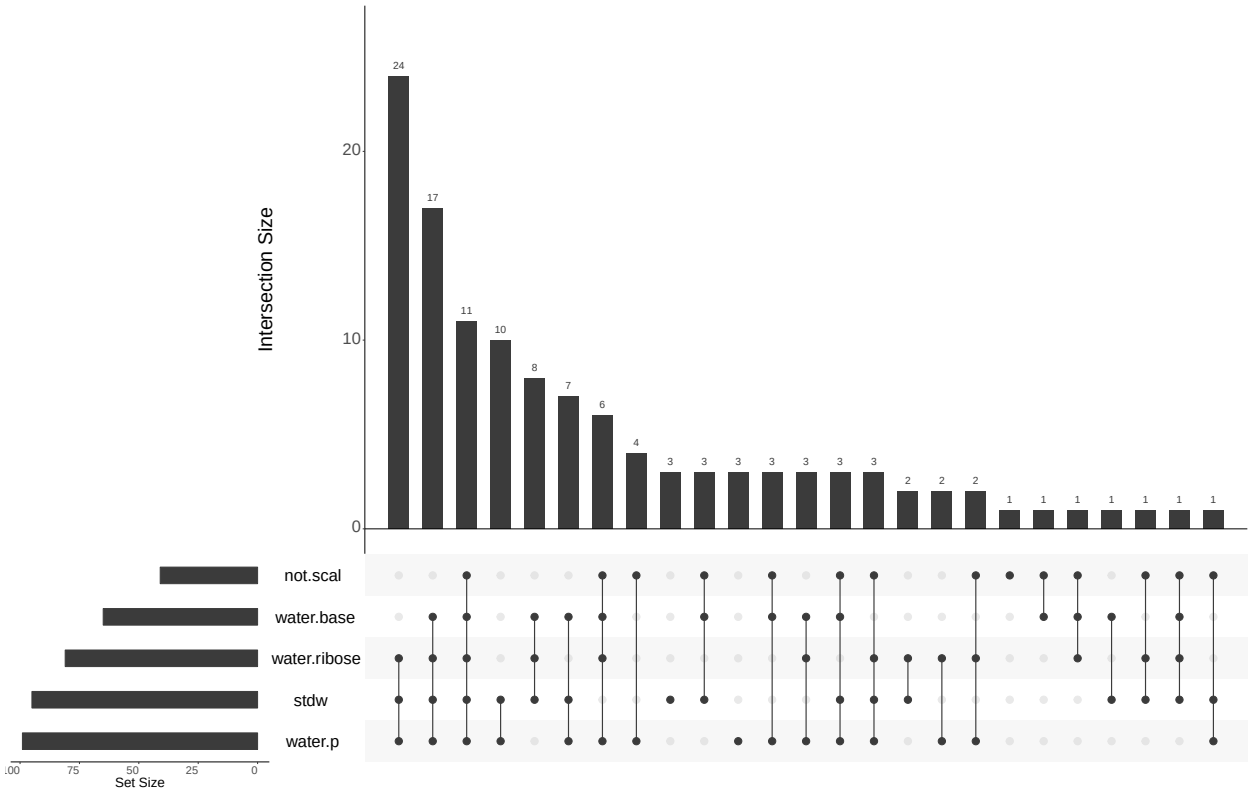
10. Attached Supplementary Data 10 (Data-S10.csv): raw data corresponding to the variations of the binding site's volume for each protein of the benchmark in three conditions: experimental, SCAL, and STDW models.



Supplementary Figure 13: Variations in the volume of the binding site. Black line: experimental structure; Blue line: optimized structure for the SCAL model; Red line: optimized structure for the STDW model. The histograms indicate a decreasing of the volume for the negative values and an increasing for the positive values. The calculation of volume does not take into account the water molecules.

Supplementary Table 5: Impact of the nonbonded model and phosphate patch on the recovery effect of the Top-10 no-prediction subset. Y: recovered prediction using a different model and patch; N: no recovered prediction with the given model and patch.

	stdw-R110	scal-R310	scal-R110
1rao	Y		
1wxi	Y		
1xtt	Y		
2g1u	Y		
2xbu	Y		
2xwm	N	Y	
3gru	N	N	N
3m84	Y		
3nua	N	N	Y
3omf	Y		
3sf0	N	Y	
4eei	N	Y	
4ijn	N	Y	
4zfn	Y		
5ed3	Y		
5jda	N	Y	
5v0i	N	N	N



Supplementary Figure 14: Upset diagram of water-mediated contacts for the Top-10 predictions. not.scal: no prediction with the SCAL model; stdw: predictions with STDW model; water.base: presence of water-mediated contacts with the nucleic acid base; water.ribose: presence of water-mediated contacts with the ribose; water.p presence of water-mediated contacts with the phosphate group.

11. Attached Supplementary Data 11 (Data-S11.csv): raw data corresponding to the molecular features associated with the Top-10 predictions.

References

() Durrant, J. D.; McCammon, J. A. BINANA: a novel algorithm for ligand-binding characterization. *Journal of Molecular Graphics and Modelling* **2011**, *29*, 888–893.

() OEDepict Toolkit 2.4.4.5, OpenEye Scientific Software, Santa Fe, NM.

() Leclerc, F.; Karplus, M. MCSS-based predictions of RNA binding sites. *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)* **1999**, *101*, 131–137.

() Gaillard, T. Evaluation of AutoDock and AutoDock Vina on the CASF-2013 Benchmark. *Journal Of Chemical Information And Modeling* **2018**, *58*, 1697–1706.

() Quiroga, R.; Villarreal, M. A. Vinardo: A Scoring Function Based on Autodock Vina Improves Scoring, Docking, and Virtual Screening. *PLOS ONE* **2016**, *11*, e0155183–18.

() Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *Journal Of Medicinal Chemistry* **2006**, *49*, 6789–6801.

() Huang, S.-Y.; Zou, X. A knowledge-based scoring function for protein-RNA interactions derived from a statistical mechanics-based iterative method. *Nucleic Acids Research* **2014**, *42*, e55–e55.

() Wang, C.; Zhang, Y. Improving scoring-docking-screening powers of protein-ligand scoring functions using random forest. *Journal of Computational Chemistry* **2017**, *38*, 169–177.

() Trott, O.; Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry* **2010**, *31*, 455–461.