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

Roy González-Alemán, Nicolas Chevrollier, Manuel Simoes, Luis Montero-Cabrera, Fabrice Leclerc

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

Abstract

Computational fragment-based approaches are widely used in drug design and discovery. One of their limitations is the lack of performance of docking methods, mainly the scoring functions. With the emergence of fragment-based approaches for single-stranded RNA ligands, we analyze the performance in docking and screening powers of an MCSS-based approach. The performance is evaluated on a benchmark of protein-nucleotide complexes where the four RNA residues are used as fragments. The screening power can be considered the major limiting factor for the fragment-based modeling or design of sequence-selective oligonucleotides. We show that the MCSS sampling is efficient even for such large and flexible fragments. Hybrid solvent models based on some partial explicit representation improve both the docking and screening powers.

Clustering of the n best-ranked poses can also contribute to a lesser extent to better performance. A detailed analysis of molecular features suggests various ways to optimize the performance further.

1 Introduction

Fragment-based approaches are widely used in ligand design with several examples of "success stories" when applied to drug design and drug discovery[1–4] since the middle of the '90s .[5] More than 30 fragment-based drug candidates have entered the clinic.[6] Despite some hindrances related to synthetic accessibility and ligand-design strategies, fragment-based approaches remain very attractive while dealing more efficiently with chemical space, molecular complexity, probability of binding, and ligand efficiency.[6] After high throughput screening, fragment-based design (FBD) approaches represent one of the three major lead generation strategies for clinical candidates.[7] Traditionally, the FBD approaches have been applied to the design of ligands assembled using small chemical groups selected from the fragments library, which is often built based on drug-like criteria. Since the fragments should also cover some chemical space with the diversity of chemical groups and molecular properties, a good strategy is needed to assemble the fragments.[6]

Both experimental and computational approaches have been developed based on the same principles that weak-binding fragments can be converted into highly efficient ligands by covalent linking.[6,8,9] In the experimental approaches, the fragments are validated by some screening methods, some of which are high throughput, e.g., by surface plasmon resonance.[10] A computational screen of fragment libraries is faster and more cost-effective than in experimental approaches. However, the lack of accuracy of the scoring functions is often invoked for their poor performance.[11–14] Criteria such as the docking and screening powers are used to evaluate their ability to discriminate native poses from false binding poses, and high-affinity or highly-selective binders from low-affinity and/or poorly-selective binders, respectively.[15]

MCSS (multiple copy simultaneous search) is a computational method that is used within

the framework of FBD approaches, although it does not include any fragment-assembly strategy.[16] MCSS mainly performs local and iterative docking calculations based on an efficient sampling method[17] which is implemented in the CHARMM program.[18] MCSS is used as a first step in the FBD process as it generates distributions of functional groups or fragments at the surface of a protein target composed of clustered docking poses.[16] Thus, it makes it possible to perform virtual screening using pre-defined[16,19,20] or customized fragment libraries.[21]

MCSS has been widely used in FBD approaches in conjunction with fragment-linking/merging methods such as: HOOK,[22] DLD,[23] or CAVEAT[24] for chemical groups, and OLIGO[25] for oligopeptides or SiteMap for peptidomimetics.[26] The MCSS scoring function is based on the CHARMM energy function; different strategies have been applied to improve its performance using more accurate methods and/or implicit solvent models. The first strategy includes post-processing of the MCSS fragment poses recalculating the score function by adding solvation terms,[27] or by rescoring (single-point energy) using a Generalized Born (GB) model.[28,29] The second strategy that is less time-consuming is to include solvent effects in the energy function during the MCSS calculations using, for example, a distance-dependent dielectric model,[27] or an alternative charge model.[30] Although implicit solvent models have become very popular, their accuracy remains limited for the calculation of solvation free energies.[31] The role of solvent is critical in the sampling and scoring of chemical fragments, but the implementation of explicit solvent and its evaluation in docking approaches remain challenging.[32]

Among the assembling strategies (linking, merging, and growing), the linking strategy is the one offering the better perspectives for the gain of binding energy.[33] The fragment merging or linking strategies consist of connecting covalently two non-competitive fragments by fusing some chemical bonds or creating some additional chemical bond(s) using a spacer to link both fragments.[8,33] One of the contributions to the gain of binding affinity with respect to that of the individual fragments comes from the rigid body entropic barrier,

which is supposed to be independent of the molecular size.[9,34,35] However, many factors make the fragment linking strategies unsuccessful.[36] In biopolymers, the chemical connectivity is well-defined, and the linking strategy does not require a spacer that is already part of the fragments. Thus, the linking only involves solving a distance-constraint problem to join the connecting atoms of successive residues; it also guarantees a straightforward chemical synthesis. On the other hand, the chemical diversity is reduced to that of the residues (20 for unmodified amino acids, 4 or 5 in the case of unmodified nucleotides, etc.). As mentioned above, MCSS-based FBD approaches were applied repetitively to the design of peptides or peptidomimetics[19,25,26,37–39] or to other biomolecules such as aminoglycosides.[30]

RNA molecules have emerged both as tools and targets in therapeutics.[40,41] In 2019, twelve FDA-approved molecules for the treatment of various pathologies were already available: nine RNA drugs as siRNA or antisense oligonucleotide (ASOs), two small molecules against RNA targets,[42] and one aptamer.[43,44] Like other RNA molecules, aptamers that are generated experimentally via SELEX (systematic evolution of ligands by exponential enrichment) offer several advantages over traditional drugs.[45] Recent improvements in the synthesis of modified aptamers known as SOMAmers[46] that bind protein targets open new perspectives of applications in bioanalysis[47,48] and therapeutics.[49] However, their application in therapeutics is still limited by their large size and molecular weight. Furthermore, the SELEX methodology cannot guarantee that the best binders have been selected (due to the combinatorial complexity of the sequence space). Computational approaches could provide a more rational strategy using structure-based methods.

A computational fragment-based approach was applied to the modeling of RNA ligands using trinucleotide fragments (3-mers) to predict the binding mode of single-stranded RNAs (up to 12-mers) to proteins.[50] It was tested on a small set of six RNA binding proteins (RBPs) and could generate near-native models of RNA-protein complexes with good accuracy (RMSD $\leq$ 2Å) in some cases.[51] However, the scoring function still lacks accuracy to discriminate near-native poses robustly and would need to be tested on a larger benchmark.

The oligomer sequence is used as a constraint because of the complexity of the sampling in the sequence space. Thus, the method is suited to the modeling but not to the design of oligonucleotides against protein targets without any knowledge on the sequence. A MCSS-FBD approach was also developed to predict the binding mode and sequence selectivity of RNA ligands using nucleotide fragments.[52] Short oligonucleotides (di- or tri-nucleotides) corresponding to the residues that most contribute to the contacts with the protein could be generated with good accuracy (RMSD $\leq$ 1.5Å) and the best score.[52] However, when considering the whole sequence space, the scoring function was not robust enough in terms of screening power to discriminate the true sequence binder from alternative sequence binders. Thus, some improvements are still required for the *de novo* FBD of bound oligomers against protein targets.

In this study, we examine both the docking and screening powers on an extended and representative benchmark of protein-nucleotide complexes. A clustering of the MCSS-generated poses is also proposed as a filtering process to select fewer relevant poses. Four solvent models (implicit or hybrid: implicit/explicit) and five classes of nucleotide fragments (with various charges and volumes) are tested (Fig 1). We also intend to evaluate a series of molecular features associated with the lack of accuracy of the scoring functions. The identification of such features would open the perspective of optimization and improvement of the docking and screening powers for the FBD of sequence-selective oligonucleotides.

2 Methods

2.1 Protein-nucleotide Benchmark

The PDB are filtered out to select a set of protein-nucleotide complexes based on different structural criteria associated with the atomic resolution and the structural similarity. A first query is carried out to find protein complexes with each of the four nucleotides as ligands and annotated in the PDB by the following labels: AMP, C5P, 5GP, U5P. An additional

criterion based on a cutoff value of 2Å resolution is also used to select only high-resolution X-ray structures. The resulting complexes are then clustered according to their sequence similarities in order to remove the redundancy. If any chain in the protein of a complex has at least 30% sequence identity with a chain in the protein from another complex, the two complexes are grouped into the same cluster.

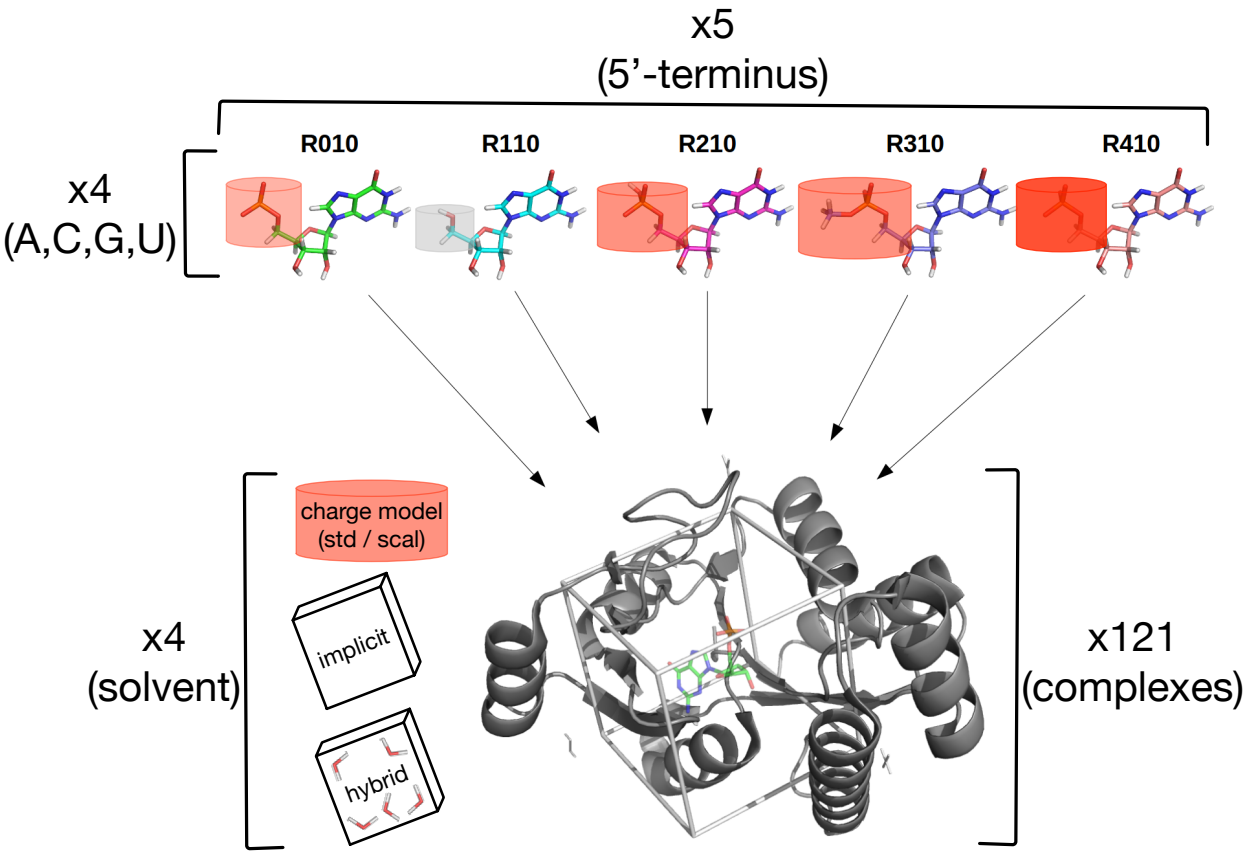


Figure 1: Schematic description of the MCSS calculations performed on the protein-nucleotide benchmark. Five chemical structures of the 5'-terminus are considered (R010, R110, R210, R310, R410). For each 5'-terminus, the four standard nucleotides (A,C,G,U) are also considered. The phosphate group is enclosed into a cylinder: the bigger the cylinder the bigger sterically, the darker red the more negative charge (the grey color indicates a null charge). Four solvent models are evaluated depending on the charge model (std: standard, scal: scaled) and the solvent representation (implicit, or hybrid: implicit and explicit). A protein target is represented in cartoon mode with the indication of the cubic box corresponding to the explored region.

The crystal structure with the best resolution in each cluster is selected as the cluster's representative. The 188 complexes thus selected by pulling down the results from the four

queries (AMP-bound: 123, C5P-bound: 18, 5GP-bound: 21, U5P-bound: 27) are then manually curated to retain those that exhibit a known binding preference for the crystallized ligand. This feature is established based on the literature and/or the annotation of the protein, e.g., a C nucleotide for CMP-kinase, etc. After curation, the dataset is reduced to 132 complexes. An additional curation is performed to eliminate some potential redundancy associated with the presence of identical binding sites for different types of nucleotides.

The followed procedure consists of superimposing all the protein structures using the program TM-align[53] and review all the structures that are similar based on the TM-score (TM-score ≥ 0.8). Two binding sites are considered non-redundant if they differ by only one amino acid residue in direct contact with the ligand. According to this criterion, only one complex is removed from the dataset in the case of the proteins corresponding to the PDB IDs: 3DXG (U5P ligand) and 3DJX (C5P ligand); the latter complex is conserved in the dataset to compensate for the minor under-representation of C5P. The full procedure ends up with a dataset of 131 protein-nucleotide complexes.

After a review of the MCSS calculations, ten protein-nucleotide complexes resulted in non-productive (see below) and are then removed from further analyses. The resulting benchmark is thus composed of 121 protein-nucleotide complexes associated with more than 13 biological functions (Fig. 2). Their binding features are characterized by the number of contacts between the protein and its ligand, the fraction of buried surface area, the number of H-bonds in the binding site, and the energy of interaction (Supplementary section: Benchmark of 121 protein-nucleotide complexes, Supplementary Fig. 1) as calculated by the MCSS scoring function (Supplementary sections: MCSS & Scoring). The contacts are calculated using the program BINANA.[54] The full tables, including the molecular features of the protein-nucleotide complexes, are provided in the supplementary materials (Supplementary section: Molecular features). The contacts are also analyzed by nucleotide type (Supplementary Fig. 2).

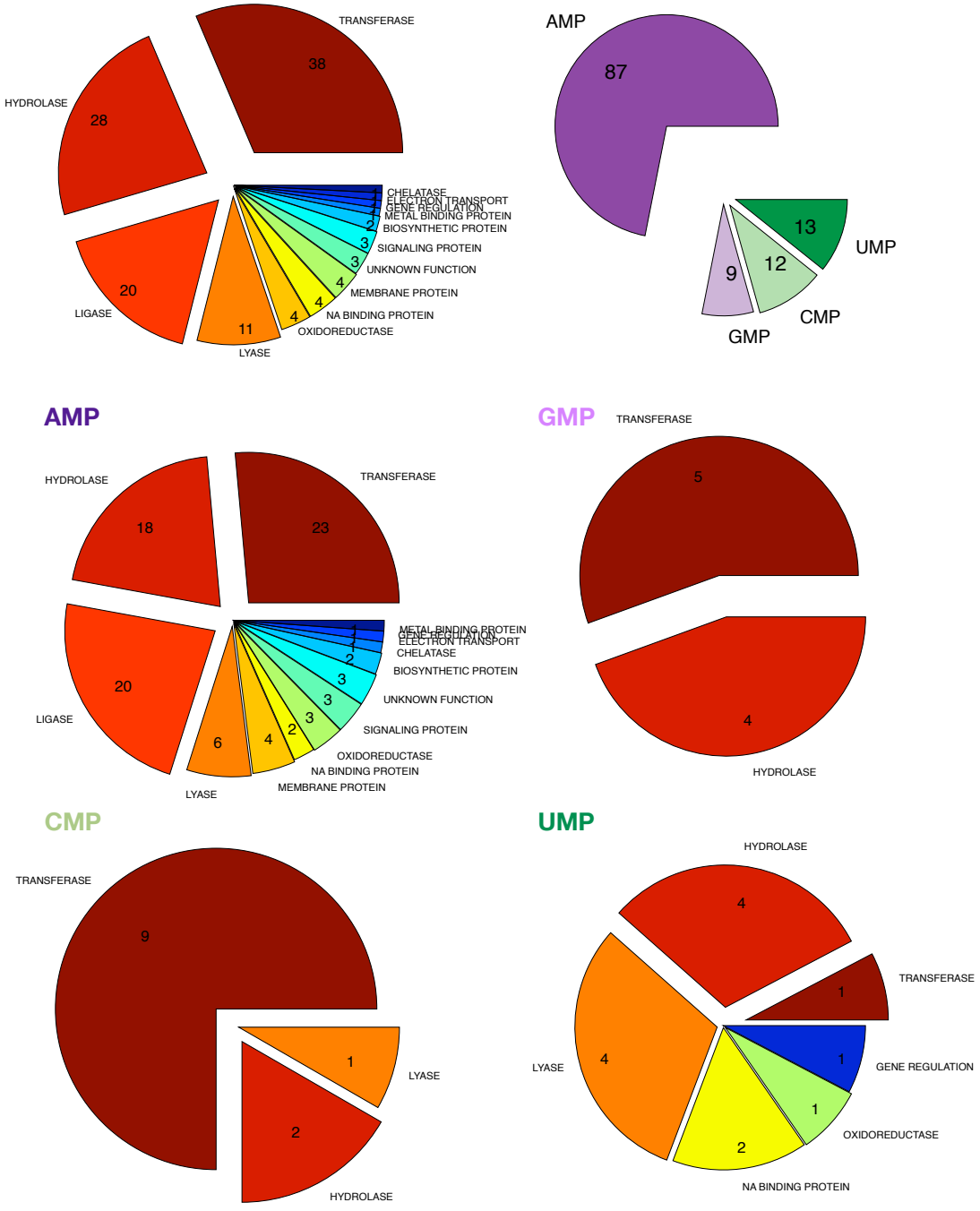


Figure 2: Distribution of molecular functions and nucleotide types in the protein-nucleotide benchmark. Top: General distribution of molecular functions (left) and nucleotide type (right). Middle-Bottom: Nucleotide-specifc distributions (AMP, GMP, CMP, UMP).

2.2 MCSS

All the proteins are prepared using the CHARMM-GUI interface[55] to convert the PDB files into CRD and PSF formats. After removal of all heteroatoms, hydrogens are added to the protein using the HBUILD command from CHARMM. Histidine residues are considered as neutral. Water molecules were present in all the protein-nucleotide complexes, in particular in the binding site. Water molecules were either removed or included before energy minimization depending on the solvent representation (implicit/hybrid). The protein targets are then submitted to an energy minimization (tolerance gradient of 0.1 kcal/mol/Å). The average deviation between the experimental structure and the minimized structure is around 1.0Å for the structures optimized without water molecules and 0.5Å for the structures optimized with the crystallized water molecules (Supplementary Fig. 3).

The nucleotide library of fragments include multiple conformations, 5' and 3' patches (see MCSS documentation: https://www.mcss.cnrs.fr/MCSSDOC). The initial default conformation used in the calculations is a C3'-endo/anti ribonucleotide with standard values of the seven torsion angles (phosphodiester backbone and base orientation). A set of five different patches on the 5' end is used in the current study with this nucleotide conformation: R010, R110, R210, R310, R410. The nucleotide fragments are fully flexible during the calculations and are prone to adjustements of the torsion angles to better fit in the binding site (Supplementary section: MCSS, Supplementary Fig. 4). Each binding region is defined by a $17Å^3$ cubic box centered on the ligand centroid where all the inorganic compounds (e.g. metal ions) or organic ligands were removed (Fig. 1). MCSS sample files are provided for the input and nonbonded parameters (Supplementary section: MCSS).

Ten protein-nucleotide complexes (PDB IDs: 1HXP, 2CFM, 2Q4H, 3L9W, 3REX 4OKE, 4XBA, 5ERS, 5M45, and 5DJH) resulted as non-productive because of a significant conformational change of the binding site after minimization (see protocol for energy minimization above) that prevented the identification of native-like poses. They are excluded from post-docking analyses due to 3 main reasons: (1) no native pose (RMSD $\leq$ 2.0Å) could be gener-

ated because of a nucleotide-binding site too buried to be accessible after minimization (PDB ID: 5M45, 5DJH); (2) no native pose could be identified consistent with a huge deviation (RMSD $>$ 2.0Å) of the crystallized ligand minimized within the optimized protein binding site (PDB IDs: 1HXP, 2CFM, 4OKE, 4XBA, 5ERS); (3) the native poses identified showed highly unfavorable energies indicating the presence of steric clashes between the nucleotide and the minimized binding site (PDB IDs: 2CFM, 2Q4H, 3L9W, 3REX, 4XBA, 5ERS). After the removal of those ten non-productive protein-nucleotide complexes, the resulting benchmark includes 121 protein structures. The reference coordinates of the ligand used to evaluate the poses correspond to those of the experimental X-ray structure.

The initial distributions of fragments are generated using 2000 groups distributed randomly and repeatedly among 25 iterations. These parameters guarantee that fragments fully saturate the binding region of all the protein-nucleotide complexes in the benchmark, i.e., the atomic density of the fragments mapped into the box is at least twice that of the maximum carbon density. During the calculations, the protein targets are considered as rigid. Final poses (minima) generated by MCSS are ranked by their score (Equations 1-4) in ascending order.

In the models that include explicit solvent (SCALW, STDW, and FULLW), the water molecules are treated independently from the fragments, which are replicated from their initial distribution during each iteration. The number of water molecules is conserved during the calculations, and they are free to move around without any constraint. However, they are not considered in the scoring as described below.

The MCSS score is defined by the electrostatic and van der Waals contributions to the interaction energy plus a penalty term corresponding to the deviation of the fragment's conformation from its energy minimum:

$$\Delta E_{MCSS}^{binding} = \Delta E_{conf}^{fragment} + \Delta E_{vdw}^{inter} + \Delta E_{el}^{inter} \quad (1)$$

The van der Waals contribution to the score is calculated in the same way for all models:

$$E_{vdw} = \sum_{excl(i,j)=1} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) sw(r_{ij}^2, r_{on}^2, r_{off}^2) \tag{2}$$

while the electrostatic contribution depends on the solvent model used. In the case of the "FULL" model, it is calculated using the standard charges as follows:

$$E_{el} = \sum_{excl(i,j)=1}^{\epsilon=1} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{3}$$

In the case of the other models using either scaled charges (i.e. "SCAL") or standard charges (i.e. "STD") (Fig. 3), it is calculated this way:

$$E_{el} = \sum_{excl(i,j)=1}^{\epsilon=3} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}^2} sw(r_{ij}^2, r_{on}^2, r_{off}^2) \tag{4}$$

where the dielectric constant is set up according to some previous work.[30]

Several phosphate group models are used in the MCSS calculations to determine the optimal parameters for mapping nucleotides at the protein surface (Fig. 1). The five different phosphate models correspond to 5' patches (R010, R110, R210, R310, and R410) that differ by the valence and charge of the phosphate group (Fig. 3). The R010 patched nucleotide corresponds to the standard nucleotide residue defined in CHARMM, and it is the only fragment with an unfilled valence shell at the 5' end (Fig. 3). All the partial charges on the phosphate groups are derived from the CHARMM parameters. They correspond to the original CHARMM charges or derived from them based on Manning's theory of counterion condensation to account for the partial neutralization of the negative charges of polyelectrolytes solution.[56] In this latter case, the net charge on the phosphate group is scaled down according to the implicit solvent model previously used in MCSS calculations performed on nucleic acids.[30]

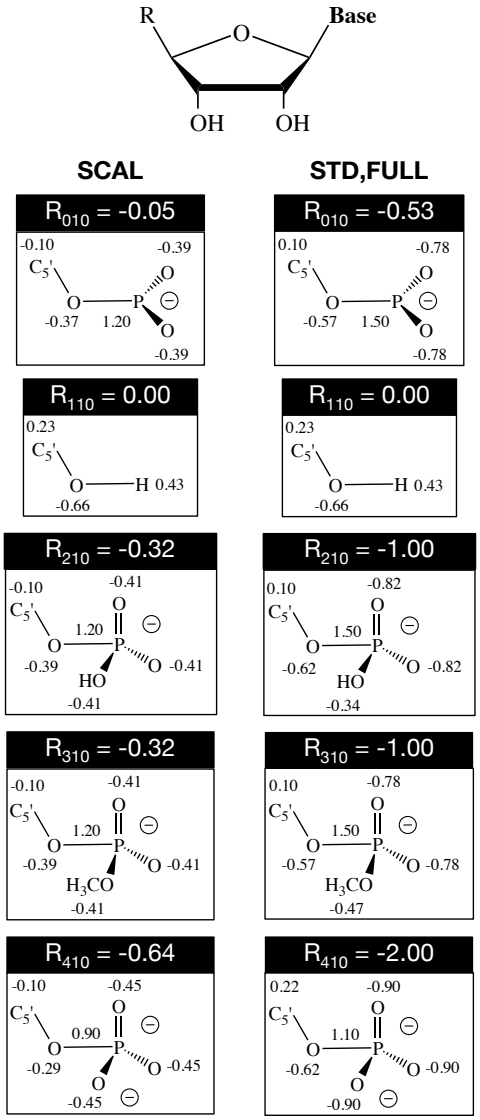


Figure 3: Nonbonded models used in the MCSS calculations. The R group corresponding to the 5' end of the nucleotide includes five flavors: R_{010} (standard nucleotide residue), R_{110} (5'OH patch), R_{210} (5'PO_4H^-), R_{310} (5'$PO_4CH_3^-$), and R_{410} (5'PO_4^{2-}). Three solvent models are used: the SCAL model is based on reduced charges on the phosphate group according to Manning's Theory[56] and applied to nucleic acids;[30] the "STD" (standard) or "FULL" models are based on standard charges. The electrostatic contribution to the interaction energy is calculated based on a constant dielectric formulation for the "FULL model". The SCAL and "STD" models are based on a distance-dependent dielectric model. The van der Waals contribution is calculated using the standard CHARMM27 potential energy function.[57]

The "SCAL" charges model (Fig. 3 - left) is combined with a distance-dependent dielectric (Equation 4) with or without water molecules: SCAL and SCALW, respectively. The default charges model "STD" or "FULL" (Fig. 3 - right) is combined with explicit solvent

representation and a distance-dependent dielectric (Equation 4): STDW, or with a constant dielectric (Equation 3): FULLW.

The MCSS software may be obtained after signing a license agreement upon request to Martin Karplus (marci@tammy.harvard.edu). The source code can be obtained from a Git repository on the I2BC software forge `https://forge.i2bc.paris-saclay.fr`).

2.3 Clustering

Although a clustering step is performed iteratively during the MCSS calculation, the default RMSD cutoff value is low (0.5Å) to guarantee a fully extended search at each iteration before re-ranking the intermediate poses and the minima. As a consequence, some minima may still exhibit some degree of geometrical redundancy. Poses coming from different initial positions may converge to similar minima while still being above the RMSD cutoff value. These minima may exhibit large discrepancies in terms of score, especially when using implicit solvent models where small deviations in coordinates may significantly alter the interaction energy with the protein target. This redundancy may negatively impact further statistical analysis as very similar poses can have a drastically different score. Such a bias can be avoided through clustering analysis based on an approach similar to that already used by MCSS.

A fast and straightforward orthogonal clustering procedure is performed on the MCSS distributions; the first pose (best ranked) is taken as the seed of the first cluster, and all other poses in the exploration with an RMSD less equal than 1Å to the seed (redundant poses) are removed from the dataset. The seed is preserved, and the process resumes taking as seed the next best-ranked available pose and performing the same comparison against remaining poses. At the end, a set of geometrically non-redundant seeds is obtained. The MCSS results presented include the analyses of the raw (R) and clustered (C) distributions.

2.4 Docking and Screening Powers

The docking power is defined as the ability of the scoring functions to identify the native ligand binding pose with respect to the non-native poses generated by MCSS for the native nucleotide ligand (single nucleotide distribution). The MCSS predictions are ranked according to the success rate for the identification of at least one native pose obtained on the full benchmark in the Top-i (Top Native in the best ranked i poses) with i in a range from 1 to 100. For each patch (from R010 to R410), the number of protein-nucleotide complexes which are predicted with a native pose in the Top-i follows a similar trend between the different models.

The scoring functions used are those implemented into MCSS with the four solvent models which are evaluated. Four alternative scoring functions used in the comparative assessment of scoring functions (CASF) challenges[14,15] have been used as well as two MM-GB models through a rescoring scheme based on single-point calculations to assess the relative performance of MCSS in docking power. Implicit solvent models such as MM-GB models[28,29] have been applied to the rescoring of MCSS minima. The two MM-GB models are CHARMM implementations: GBSW[58] and GBMV.[59] The other four selected scoring functions are either generic: Autodock Vina ,[60] and Vinardo,[61] or specialized on nucleic acids ligands for ITscorePR,[62] and $\Delta_{vina}RF_{20}$.[63] The impact of the clustering on the scoring performance is also evaluated (Supplementary section: Scoring, Supplementary Fig. 5). Finally, the docking power is decomposed per the nucleotide type (Supplementary Fig. 6).

To evaluate the screening power, the MCSS distributions from the four nucleotides are merged and sorted according to their score in increasing order as in the nucleotide-specific distributions (from the more negative to the less negative or positive). In each Top-i, a prediction is considered as optimal if both conditions are met: (1) it corresponds to a native pose (RMSD $\leq$ 2.0Å), (2) the native nucleotide is ranked ahead of the three other non-native nucleotides. For example, an optimal prediction in the Top-1 means a native pose is found with the best score from the merged distributions.

Since the scoring function is still an estimate and raw approximation of the relative binding energy, we consider as good predictions the cases where the native nucleotide is predicted within a 2 kcal/mol range from the best ranked non-native nucleotide. This threshold value corresponds to a maximum offset of 2 kcal/mol in 90% of the benchmark (STDW model) where the offset is defined as the difference between the best-ranked pose whatever the nucleotide type and the best-ranked pose for the nucleotide corresponding to the native ligand (Supplementary Fig. 7). Predictions that do not satisfy these criteria are considered as poor.

2.5 Molecular Features

The molecular features are analyzed on a subset of the benchmark corresponding to the 17 protein-nucleotide complexes that do not generate any prediction in the Top-10 without any distinction from the model and patch. We consider that a given feature has a significant impact on the prediction when it is found associated with the absence of prediction at a higher frequency than that in the benchmark (Supplementary section: Molecular features, Supplementary Table 1).

The volume calculation of the binding site is performed using the PyVOL python package.[64] PyVOL is used with the pocket corresponding to the nucleotide-binding site as input (coordinates of the nucleotide ligand of interest). The threshold value to discriminate between high or low binding volume is set to 635$Å^3$ which is the average value of the distribution (30% high and 70% low). The other molecular features include the number of water molecules around the nucleotidic ligand, the presence of metals, and the presence of other nucleotidic ligands (nucleic acid or cofactor) in close vicinity to the binding site. The threshold value for the number of water molecules between nwat high and low is set to: 6 (nwat.low $\leq$ 6 & nwat.high $>$ 6) which is the average value of the distribution (38% high and 62% low).

The interaction features (base contacts, clashes, salt bridge, stacking) are extracted from the analysis of the binding site[54,65] (Supplementary section: Benchmark of 121 protein-nucleotide complexes, Supplementary Fig. 1).

3 Results and discussion

Most of the docking methods and their scoring functions have been tested on different benchmarks. These benchmarks have been designed for some specific families of ligands including RNA ligands.[66–71] However, the RNA-protein benchmarks include large RNAs (tRNA, rRNA, ribozyme, etc.) where single-stranded RNAs are poorly represented and mostly present in the context of single-stranded regions connected to double-stranded regions. Building the benchmark from a subset of RBPs binding ssRNAs would select optimal but also sub-optimal binding sites corresponding to spacer regions with weak contacts with the protein.[52] To avoid such bias, we built a benchmark based on the protein-nucleotide complexes currently available in the Protein Data Bank (RCSB PDB[72]). A previous protein-nucleotide benchmark with 62 complexes was used to evaluate the docking power of three methods: AutoDock (4.2.3), GOLD (5.1), and MOLSDOCK.[73] However, the benchmark is mostly outdated, with only 40% of complexes with an atomic resolution less than 2.0Å and thus not representative of the currently available structural data. On the other hand, the methods were tested under biased conditions: the docked region was restricted to the native ligand pose (5Å^3), and the high-occupancy water molecules of the binding site were preserved within a rigid receptor.

In this study, we use an updated and representative dataset of high-resolution protein-nucleotide complexes in which only nucleotide monophosphate ligands, as single-residue fragments, are included (see section Protein-nucleotide Benchmark & corresponding section Protein-nucleotide Benchmark in Methods). While the nucleotide fragments are fully flexible, the protein structure is considered as a rigid body, but its coordinates may differ depending on the solvent model. Five different nucleotide fragments that differ at the 5' terminus in charge and size are evaluated (Fig 3). Four different combined solvent and charge models (Methods: section MCSS, Fig. 3) are also tested based on an implicit or hybrid (implicit/explicit) solvent representation on an extended binding region (Fig. 1). The results are analyzed both in terms of docking and screening powers (see Methods, section Dock-

ing and Screening Powers). The discussion takes into account the potential biases from the benchmark identified in the detailed analysis of different molecular and energy features (Supplementary Figures 1- 2). Some of those features are altered to some extent by the protocol used for the preparation of the protein structures: e.g., removal of metal ions, variations in the volume of the binding site, distribution of water molecules in the binding site (Supplementary Fig. 3), the results are also discussed accordingly. The further analysis of MCSS parameters and molecular and energy features also reveals situations where the predictions are more reliable (nucleotide fragment, solvent model, scoring function, binding mode, and contacts, etc.). Various case studies illustrate the impact of these parameters and features on the docking and screening powers. This information will be useful for further improvements of the method in the context of fragment-based approaches.

3.1 Protein-nucleotide Benchmark

The protein-nucleotide benchmark includes a non-redundant set of 121 complexes associated with 14 different known molecular functions and a wide variety of binding modes (Supplementary section Benchmark of 121 protein-nucleotide complexes, Supplementary Data-S1). The selection criteria retained to build the benchmark are detailed in Methods (section Protein-nucleotide Benchmark). In the perspective of evaluating the docking and screening powers, we looked at the possible biases associated with either the ligand composition, the binding site features, as well as those related to the solvent model and scoring function. We will further use these data in the discussion of the MCSS performances.

The proteins binding AMP are over-represented in PDB with respect to those binding CMP, GMP, or UMP. The ligand composition in the benchmark is biased accordingly with 72% of AMP-bound complexes; the other complexes are represented in a similar proportion between 7 to 10% (Fig. 2). Thus, the docking power should also be analyzed in detail by nucleotide type. A series of molecular descriptors compose the features used to characterize the 121 nucleotide-binding sites. These features include standard contacts (closed contacts, H-

bonds, hydrophobic contacts), nucleic acid-specific contacts (stacking contacts, salt bridges), and energy-related descriptors (buried fraction of ligand, binding energy score) as described in the supporting information (Supplementary section Benchmark of 121 protein-nucleotide complexes, Supplementary Data-S2 & Data-S3). Broad distributions are observed for the standard contacts (Supplementary Fig. 1).

Only the nucleic acid base moiety allows the chemical distinction between the four nucleotide fragments. For a reliable evaluation of the screening power, the contacts established by the base moiety should be represented enough in number and frequency because they determine the selectivity for one specific nucleotide. The decomposition of the contacts based on the three phosphate, ribose, and base moieties reveals that the base contacts are slightly more represented. They are still slightly less frequent, especially for the close contacts (Supplementary Fig. 1A-B). In more than 10% of the benchmark (15 protein-nucleotide complexes), there is no direct base contact suggesting the binding selectivity may be hard to predict in those cases and would negatively impact the screening power. Nucleic acid-specific contacts are only represented in about half of the benchmark (Supplementary Fig. 1E-F). However, the buried fraction of the ligands is more than 50% except in a single case (Supplementary Fig. 1G), indicating that the nucleotide generally binds in some well-defined cavity as shown in the 2D diagrams of the binding sites (Supplementary section Benchmark of 121 protein-nucleotide complexes, Supplementary Data-S4).

The decomposition of the contacts per nucleotide type shows there is also a bias towards AMP, which is the nucleotide with almost ten times more contacts than the other nucleotides, although the contact profile (i.e., the proportion of different kinds of contacts) is similar between the four nucleotides (Supplementary Fig. 2). Thus, we may expect AMP binding to be easier to predict, i.e., to provide better performance in docking and screening powers. The docking power, in particular, depends on both the quality of sampling and scoring. A baseline for the default MCSS scoring function ("SCAL" model) was established on the benchmark after minimization of the ligand by re-insertion within the optimized binding

site and calculating its score (Supplementary Fig. 1H). The decomposition of the MCSS score into its different contributions (see Methods, equation 1) shows that the van der Waals contribution dominates. Although the conformational penalty is the minor contribution (mean value of 5.5 kcal/mol), it is still significant. It stresses the importance of evaluating this term properly with respect to the other contributions, given that nucleotides are very flexible (six torsion angles in nucleotide fragments). For that, a good sampling is also required.

In traditional fragment-based approaches, it is recommended to use small fragments which are easier to sample.[74] Large fragments such as nucleotides have many degrees of freedom, also making computational sampling more difficult. Only a unique standard nucleotide conformation is used in MCSS while the benchmark include a large diversity of bound conformations (Supplementary section MCSS, Supplementary Fig. 4). Thus, the sampling should be efficient to identify bound conformations that deviate from the standard (unbound) conformation, such as syn conformations found in 10% of the benchmark where the base orientation is opposite from the standard anti conformation. On the other hand, the contributions to the MCSS score should be well-balanced, e.g., the conformational penalty should not be under or over-estimated to guarantee accurate predictions.

All the high-resolution protein-nucleotide complexes of the benchmark include water molecules around the protein surface and the binding region. The ligand and water molecules were removed in the "SCAL" model, leading to some distortions of the binding sites after minimization. In the other solvent models where the crystallized water molecules were included, the original experimental coordinates were more preserved: 0.5Å versus 1.0Å (Supplementary Fig. 3A). However, other artifacts associated with the water molecules also exist (Supplementary Fig. 3B). The minimization does induce displacements in the position of the water molecules in the binding region mostly due to the removal of the ligand leading to variations in their number and distribution (Fig. 3B-C). All the mentioned biases and issues will be addressed in the comparison of the docking and screening powers for the different solvent models.

3.2 Models and poses

The identification of native poses, according to standard criteria (see Methods, section Docking and Screening Powers), depends primarily on the number of generated poses and the quality of the sampling. The first MCSS parameters evaluated are the nucleotide fragments: R010 to R410 (Fig. 1). Since their charge and size differ, they are evaluated in combination with the different solvent models. The raw distributions generally include up to several thousands of poses. The total number of poses generated depends mostly on the solvent model and the phosphate patch to a lesser extent. The presence of explicit water molecules partially reduces the molecular volume accessible for nucleotides in the binding region. Thus, the number of poses generated with the SCAL model is much larger than that generated with any of the hybrid solvent models: SCALW, FULLW, and STDW (Fig. 4). The comparison of the raw and clustered distributions also shows that the SCAL model exhibits the higher redundancy in the generated poses demonstrated by the larger difference between the raw and clustered distributions for each patch.

Although the electrostatic contribution is not the major contribution in the default scoring function with an implicit solvent model (Supplementary Fig. 2), it has a significant impact on the number of generated poses. Both the charge and the dielectric model have to be considered. In the SCAL model based on a distance-dependent dielectric, the observed trend is the more negative the charge on the phosphate group (from R110 to R010, R210/R310, and R410), the higher the number of generated poses except for the more charged patch R410 (Fig. 4). The more charged the phosphate group is, the higher the electrostatic contribution, and the more likely the pose can pass the energy threshold value of the MCSS score. The R210 and R310 patches give equivalent results with the same net charge on the phosphate group. On the other hand, a too highly charged phosphate group (R410) may also produce unfavorable interactions with negative charges at the protein surface. In the other hybrid models, the trend is not dominated by the charge but rather by the fragment's size. The larger the patch is (from R110 to R010, R410, R210, and R310), the lower the number of

generated poses, and the lower accessible volume, as mentioned above. The models based on a distance-dependent dielectric, SCALW and STDW, also follow this trend given that R210 and R410 only differ by a proton. In the particular case of the constant dielectric model FULLW, both the charge and size effects explain why R410 is not on the lines with the other patches (Fig. 4).

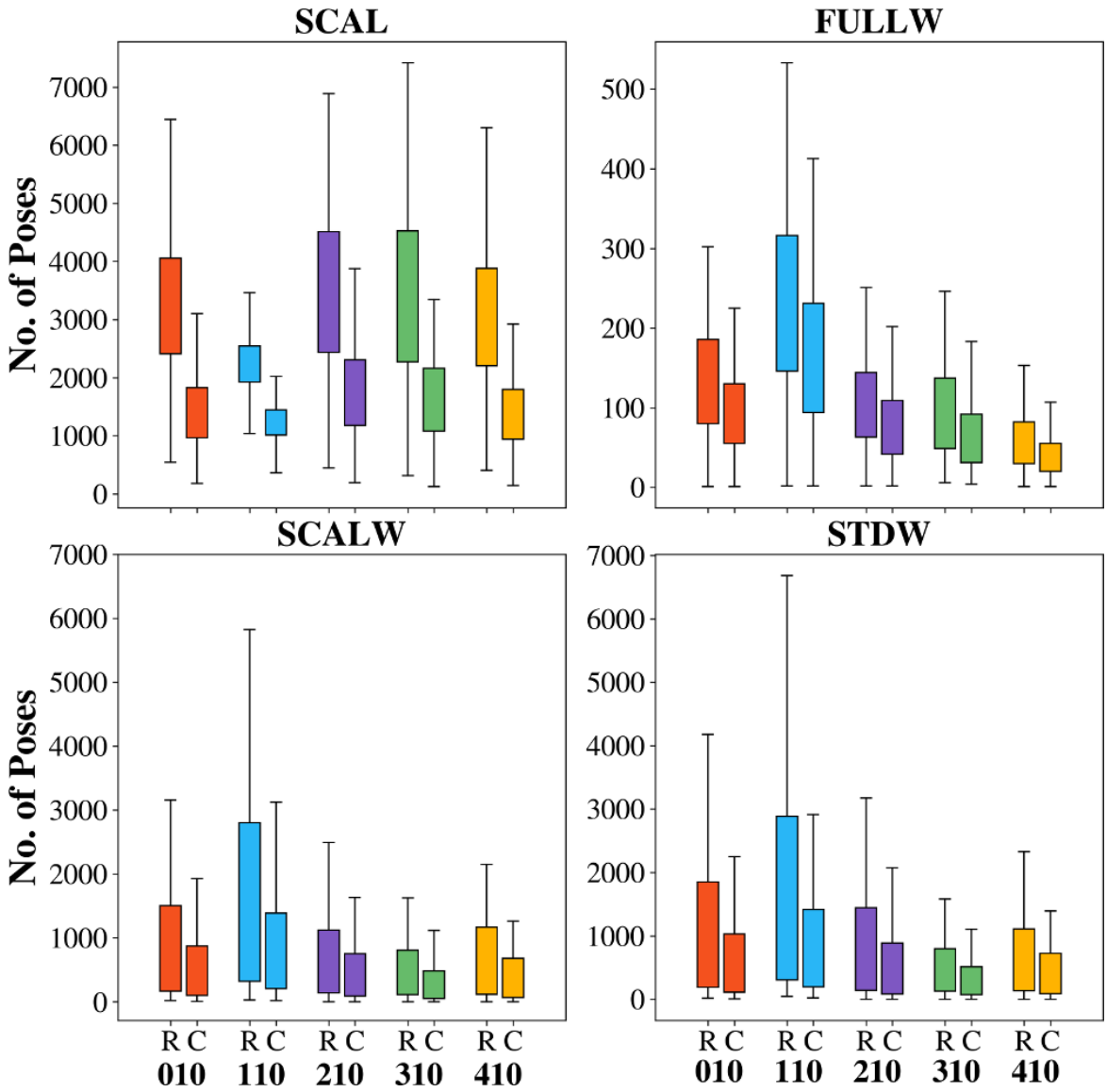


Figure 4: Boxplot representation of the number of poses generated for the 121 protein-nucleotide complexes for each 5' patched nucleotide (010, 110, 210, 310, 410). Results for raw (R) and clustered (C) distributions are shown.

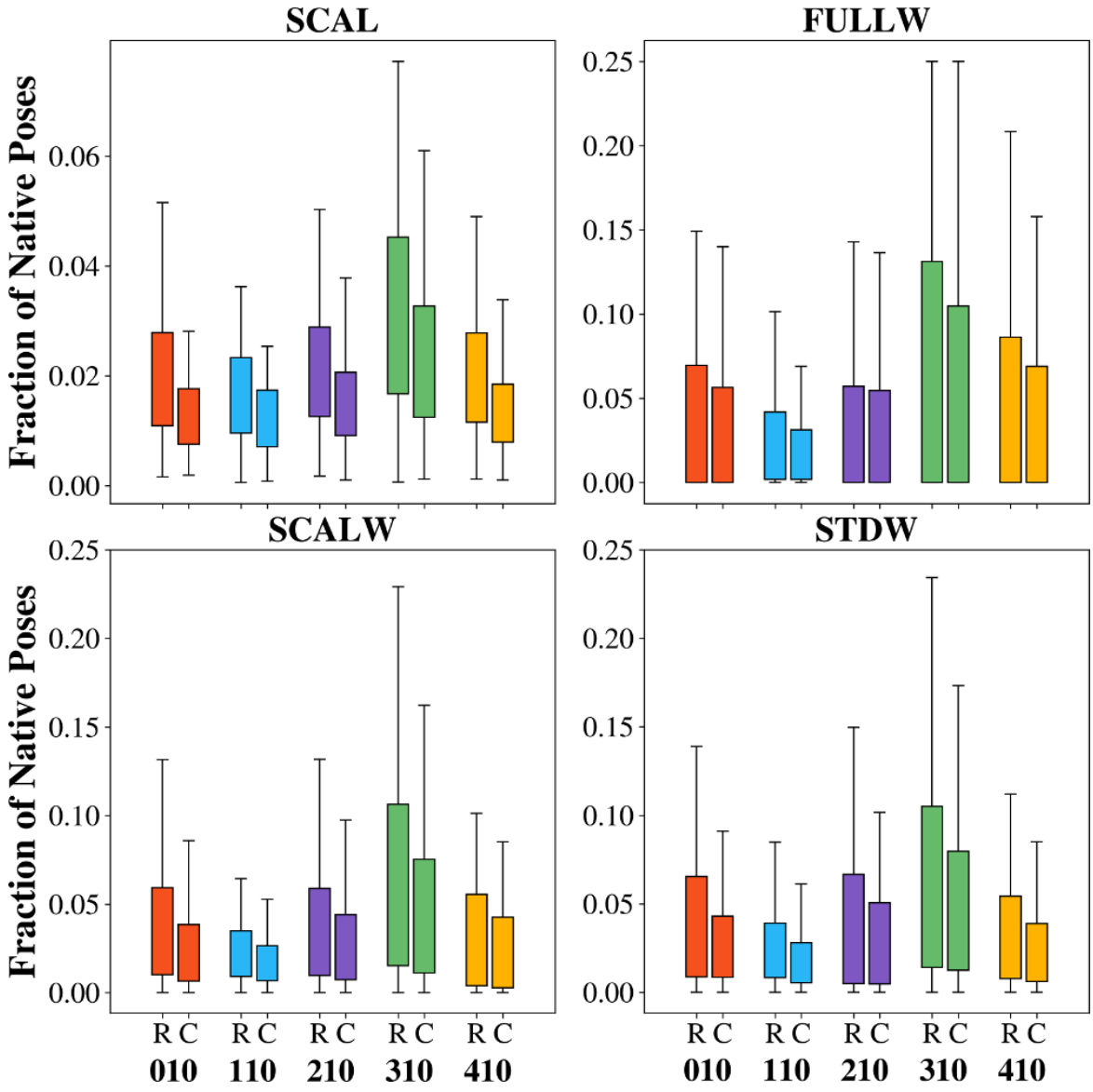


Figure 5: Boxplot representation of the fraction of native poses generated for the 121 protein-nucleotide complexes for each 5' patched nucleotide (010, 110, 210, 310, 410). Results for raw (R) and clustered (C) distributions are shown.

The fraction of native poses over the entire MCSS distribution for all solvent models and patches is shown in Figure 5. This fraction is similar for all patches in each of the four models, except for R310. The patch R310 carries a methyl group in one of the phosphate oxygen. This group confers the ability to establish more hydrophobic contacts than other patches. The SCAL model shows a significantly lower fraction of native poses than solvated models despite a much larger number of generated poses (Fig. 4). As for the number of poses,

the raw and clustered distributions are more scattered in the absence of water molecules. In solvated models, the fractions of native poses for SCALW and STDW are very similar. On the other hand, the FULLW model has more cases where no native pose is found, as seen by the displacement to zero of the first interquartile section for the boxplots (Fig. 4).

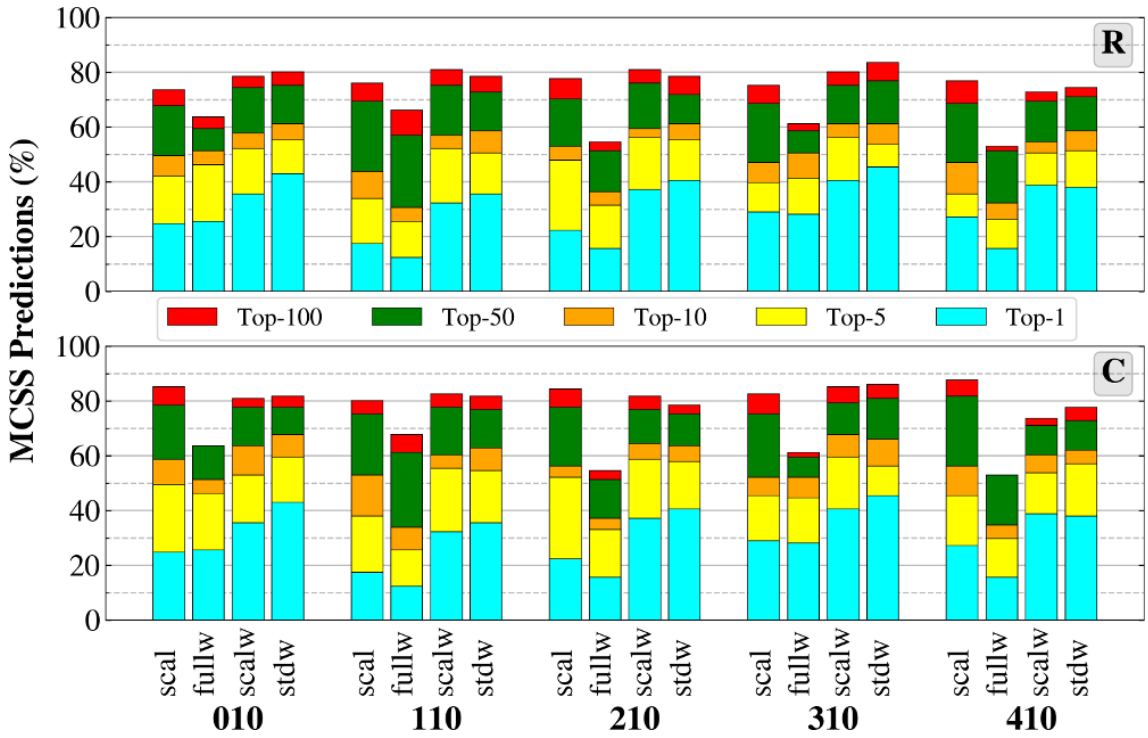


Figure 6: Stacked histogram representation of the Top-i ranked native poses generated for the 121 protein-nucleotide complexes for each nucleotide patch. Result on raw (upper) and clustered (bottom) distributions are shown.

3.3 Docking power

The performance in docking power is evaluated on all models and patches using the standard metrics based on the native poses found in the Top-1 to Top-100 scores with the intermediate ranks: Top-5, Top-10, and Top-50 (see Methods, section Docking and Screening Powers). The best performances are obtained with the SCALW and STDW models whatever the patch used (Fig. 6). The STDW model slightly outranks the SCALW model in the Top-1 and Top-10 for all the patches (except for R310 where the performance is equivalent for the Top 10), while the performance is pretty similar for the Top-50 and Top-100. The best

performance is obtained for the patch R310 with a success rate of 45% in the Top 1, a bit more than 60% for the Top 10, and more than 80% in the Top-100. However, the gain in performance with respect to the other patches is tiny in Top-10 and Top-50. The clustering does not change the general trends observed in the raw distributions, but it slightly increases the performance in the Top-100 and, to a lesser extent, the lower Top-i.

The better performance of hybrid solvent models SCALW and STDW over the SCAL implicit model is partly due to the conformational penalty term (Equation 1) corresponding to the deformation of the fragment from its optimal conformation. Although this term is generally a minor contribution, it may vary depending on the nonbonded model. We can compare the torsion angles observed in the MCSS minima to the known ideal values and values observed in the native bound conformations of the nucleotides from the benchmark (Supplementary section MCSS, Supplementary Fig. 4). The absence of water molecules in the SCAL model reveals a few biases where, for example, the syn conformation is more populated than expected as compared with the experimental or the ideal values collected from the experimental structures of nucleic acids.[75,76] The SCALW model is also biased, and the STDW but to a lesser extent; only the FULLW model is exempted. Another common bias in models (except for the FULLW model) is the over-representation of the C2'-endo conformation for the ribose while the initial conformation is always a C3'-endo conformation. It is partly due to the nonbonded model and the absence of a full solvation of the ribose moiety. In FULLW, the C3'-endo/C2'-endo representation is more balanced. Still, the phosphodiester backbone (torsion angles α and β) deviates from the optimal values because of some distortion of the phosphate group, which is highly charged and tend to stick closely to the protein surface in the absence of any screening effect (constant dielectric model).

Implicit solvent models such as MM-GB models[28,29] have been applied to the rescoring of MCSS minima. A few other scoring functions also provide good performances in the CASF challenges.[14,15] Six alternative scoring functions have been selected, two of them correspond

to MM-GB models (see Methods, section Docking and Screening Powers). The results show that the standard MCSS scoring function corresponding to the SCAL model (MCSS-SCAL) has a similar performance than Vina, slightly below that of $\Delta_{vina}RF_{20}$ (Fig. 7). The Vinardo scoring function performs slightly better than both MCSS-SCAL or $\Delta_{vina}RF_{20}$. The other three scoring functions (ITscorePR, MM-GBSW, MM-GBMV) have a low performance. The clustering protocol (see Methods, section Clustering) improves the performance of MCSS-SCAL slightly exceeding that of Vina or $\Delta_{vina}RF_{20}$ (Supplementary section: Scoring, Supplementary Fig. 5).

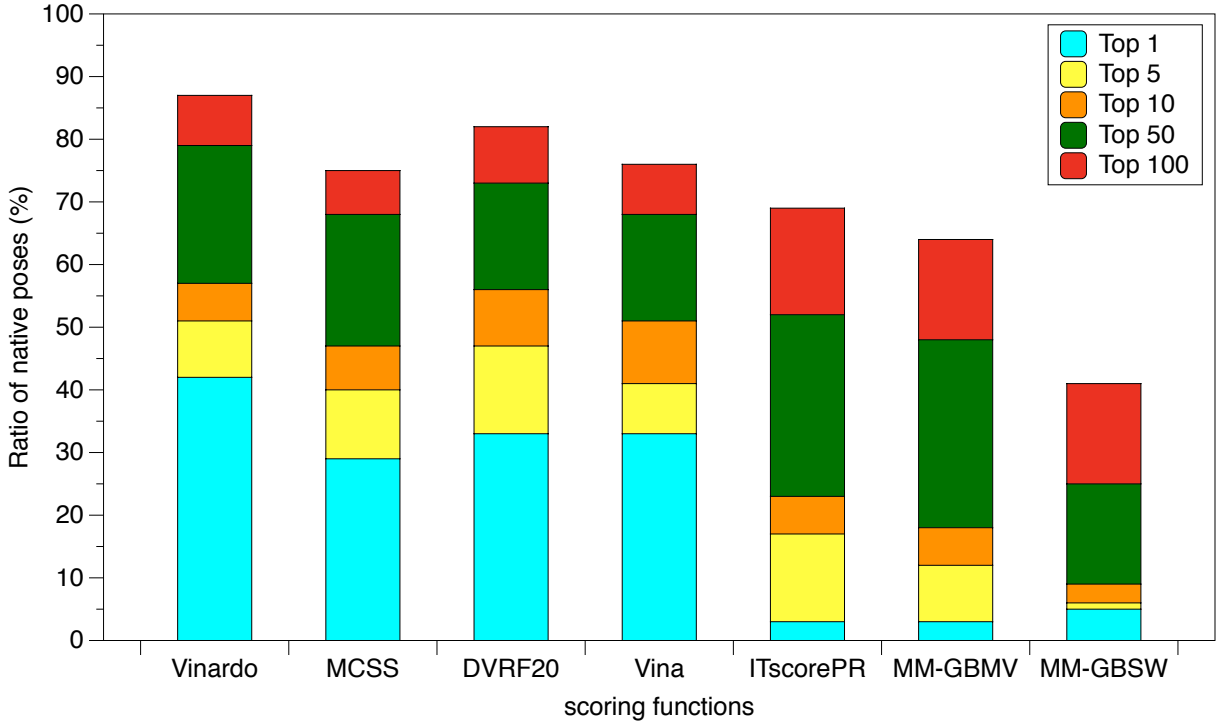


Figure 7: Docking powers (top1 to top100) for Vinardo, MCSS, $\Delta_{vina}RF_{20}$, Vina, ITscorePR, MM-GBMV, and MM-GBSW using the patch R310. The two MM-GB models use the molecular mechanics terms from CHARMM (MCSS with "SCAL" model) and the solvation contribution from the respective Generalized Born models implemented in CHARMM (see Methods, section Docking and Screening Powers).

The MCSS scoring function associated with the STDW model still outperforms all of the alternative scoring functions in the Top-1 to Top-10 in both raw and clustered distributions (Fig. 6). In the CASF-2016 benchmark, the docking power ranges from around 30% to

90% for a variety of scoring functions.[15] The docking power is around 90% for both Vina and $\Delta_{vina}RF_{20}$. On the current benchmark, their performance is only 33%, indicating the challenging task to score charged ligands such as nucleotides. Vinardo performs slightly better (42%) and also MCSS-STDW (45%).

Because of the composition bias in the benchmark, the performance was then analyzed by nucleotide type. Since the adenosine is over-represented in the benchmark, the performance for that specific nucleotide generally follows the global trend described above (Fig. 8). However, the performance for guanosine decreases for the larger patches R210 to R410, whatever the model used. Only the smaller patches R010 and R110 give a similar performance or better in some cases; the success rate with R110 is even better from Top-1 to Top-50, indicating the existence, as discussed before, of a size effect that drives down the performance (guanine is slightly more voluminous than adenine). Consistently, the performance generally improves for pyrimidines (C or U), which are smaller than purines. On the other hand, the performance is degraded in the smaller nucleoside fragments (R110) that do not carry any phosphate group (uncharged). The pyrimidic nucleotides are better predicted, especially for the two best models SCALW and STDW with R310. The predictions are equivalent or degraded for the more highly charged patch R410, especially with U. The analysis of the clustered distributions confirms the observed trends of the raw distributions, with improved performances reaching 90% to 100% for the Top-100 in a larger number of models and patches (Supplementary Fig. 6).

3.4 Screening Power

In the benchmark, we assume that the crystallized nucleotide is always the native and more specific nucleotide, i.e., it is the only nucleotide ligand with a detectable affinity or the best binder among the four nucleotides. Based on this assumption, we can define a screening power as the ability to rank the native nucleotide ahead of the other three nucleotides. In that case, we will refer to optimal predictions as the native pose is identified, and the native

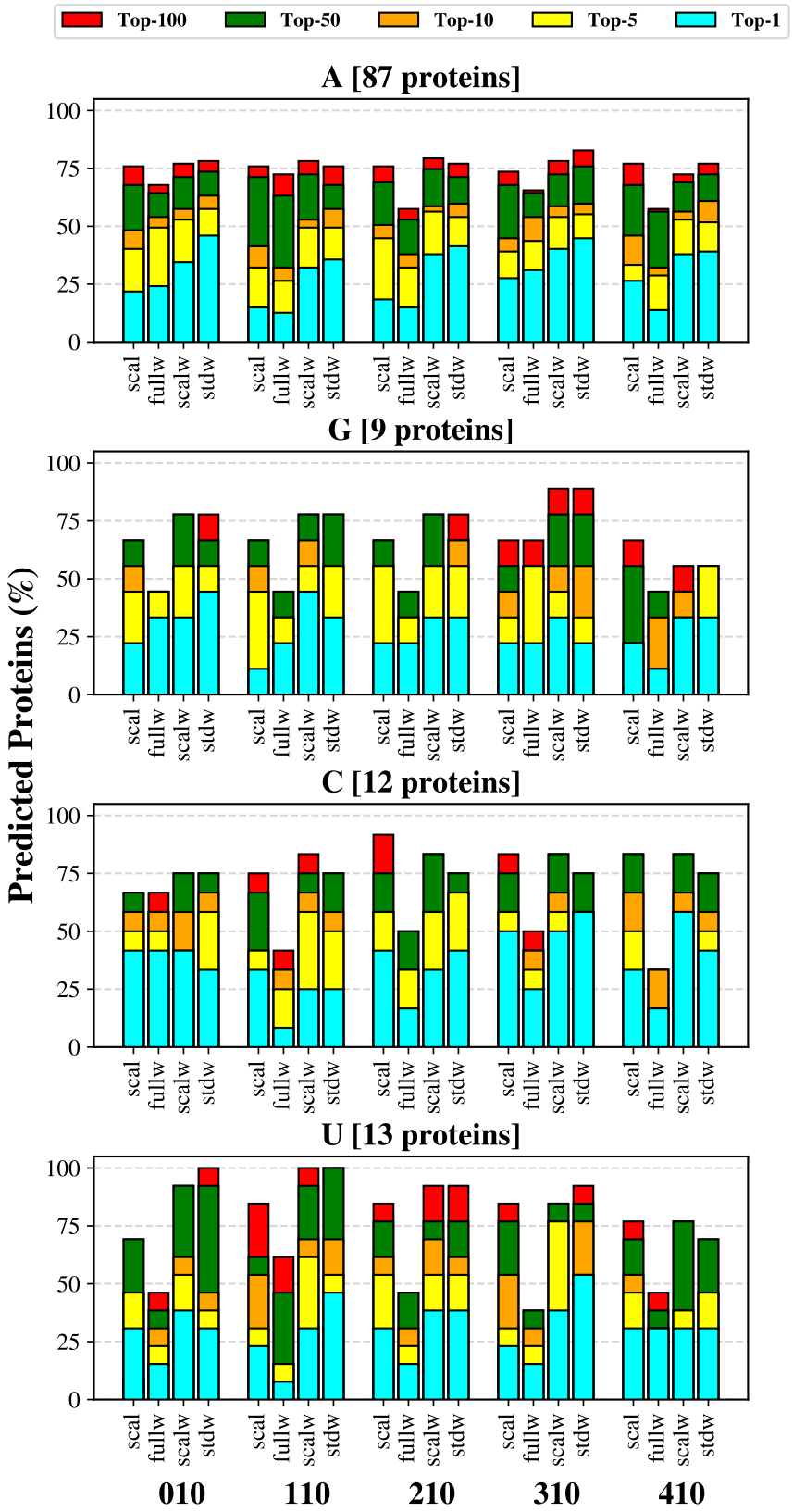


Figure 8: Nucleotide decomposition of the success rates obtained for each solvent model and patch. The data are shown for the raw distribution (without clustering) and each Top-i.

nucleotide is ranked first. The other predictions are considered poor predictions even if native poses are found for the native nucleotide. As an illustration, we show the results obtained for one protein-nucleotide complex (PDB ID: 1KTG) for both SCAL and STDW models (Fig. 9).

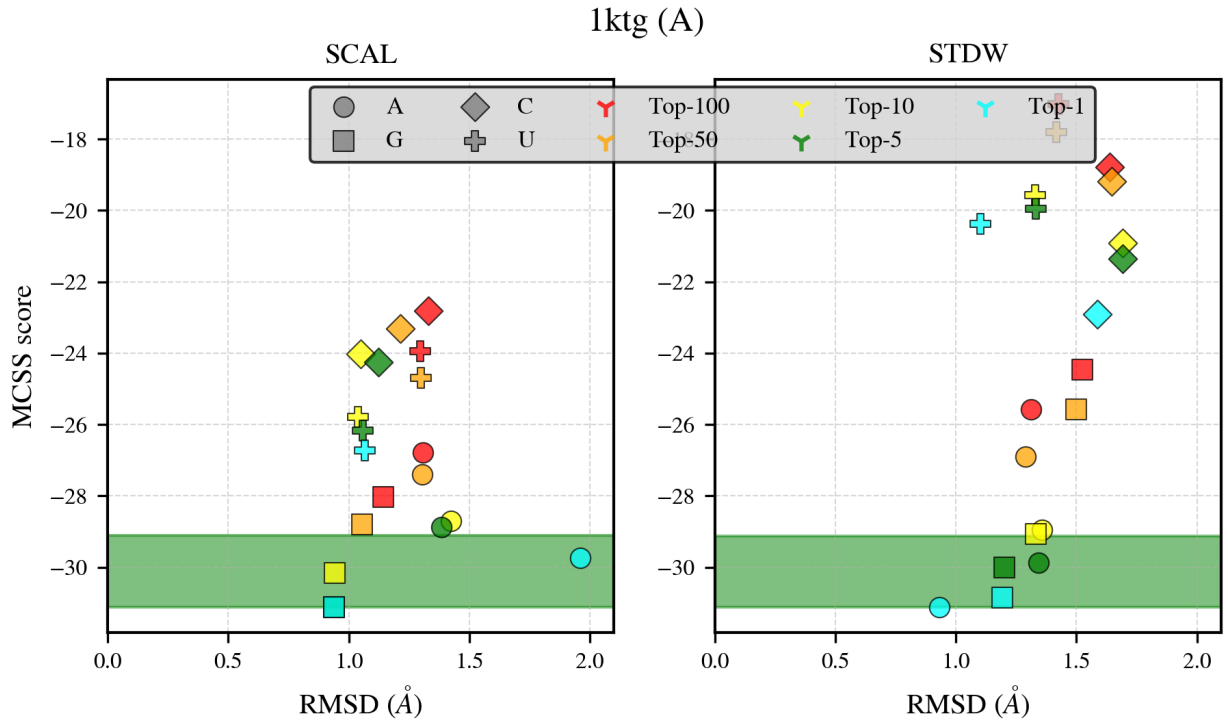


Figure 9: Binding selectivity predictions for 1KTG. Left: SCAL model (R310); right: STDW model (R310); the interval of MCSS scores corresponding to a 2 kcal/mol range is indicated by the green bar. Each Top-i for i $>$ 1 is represented by a single point that corresponds to the average RMSD and score of all its members.

The best-ranked nucleotide is the native one (A) in the STDW model; other poses of the native nucleotide are also identified (Top-5, Top-10, etc.), but only one is within the 2 kcal/mol score range (good prediction). Some of the poses corresponding to non-native G nucleotides are within the MCSS score range of 2 kcal/mol. In the STDW model, the prediction is optimal since the best-ranked pose does correspond to the native nucleotide. In the SCAL model, the pose with the best score corresponds to a non-native G nucleotide, but the Top-1 for the native nucleotide is within the 2 kcal/mol range; it is not considered

as an optimal prediction but as a good prediction. The other poses for the native nucleotide, which lie out of the 2 kcal/mol range (Top-5, Top-10, etc.), correspond to poor predictions.

The analysis of the results is focused on the comparison between the standard SCAL model (without explicit solvent) and the hybrid STDW model with the R310 patch. The STDW model shows a significant gain of performance with explicit water molecules (Fig. 10). In the optimal predictions, the STDW outperforms by 15 to more than 30% from the Top-1 to Top-100, respectively. In all Top-i, the STDW optimal predictions always exceed the SCAL total predictions. Moreover, the ratio of optimal/good predictions is always much higher in STDW (Fig. 10). The docking power determines in part the magnitude of the screening power, i.e., the more native poses, the more likely the native nucleotide is well ranked and associated with an optimal or good prediction. Considering only the cases where both models generate at least one native pose in the respective Top-i, we exclude the contribution of the docking power to the screening power (Supplementary Fig. 8). These results show that the STDW model still has a better screening power, indicating that the hybrid solvent model can intrinsically better discriminate the native nucleotide from the non-native ones. The analysis of the score distributions by nucleotide type suggests that the reason for the better screening power of STDW lies in a scoring bias. In the SCAL model, purines that are composed of more atoms are slightly better scored than pyrimidines (C or U) with a preference for G over A nucleotides (Supplementary Fig. 9).

In contrast, A nucleotides are better scored in the STDW model while the other three nucleotides have similar distributions. Another difference is the much more extensive range of scores for all four nucleotides. The more favorable scoring of A is consistent with more tightly binding modes, a known bias of the benchmark as mentioned previously (Supplementary Fig. 2). Moreover, the nucleotide decomposition of the screening power shows no significant difference in performance between A and the other three nucleotides, although it is slightly better in the Top-100 (Supplementary Fig. 10). Thus, the absence of any apparent bias in the STDW scoring makes its more efficient in terms of screening power. The main difference

between the SCAL and STDW models is the presence of explicit water molecules that leads to increased performances in the sampling and scoring.

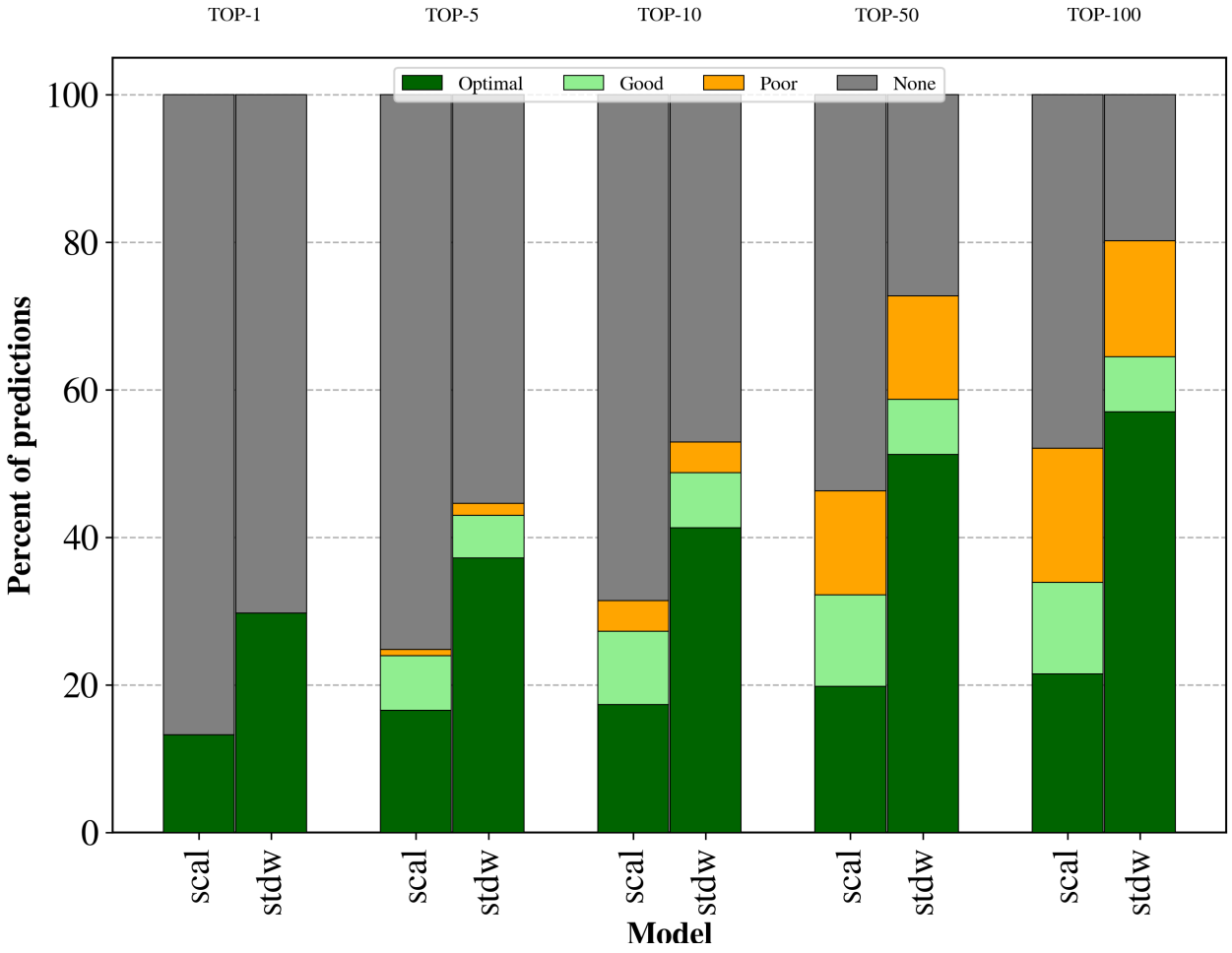


Figure 10: Binding Selectivity Predictions. Optimal: native nucleotide as the best ranked; good: native nucleotide ranked within a 2 kcal/mol range from the best ranked non-native nucleotide; poor: native nucleotide ranked out of the 2 kcal/mol range.

3.5 Molecular features

To better understand the role of solvent and other molecular properties associated directly or indirectly with water molecules, we define a series of representative features for nucleotide ligands. Then, we determine the relationships between these features and the lack of prediction, which are represented by logic diagrams (Upset plots). We classify the features into three main groups related to:

1. The binding site properties (volume, number of water molecules, presence of metals, presence of other nucleotidic fragments).

2. The conformational properties (purine/pyrimidine, syn/anti).

3. The interaction properties (contacts, clashes, stacking, salt bridges). Whether a feature is statistically significant or not is determined by its relative frequency in the subset of the benchmark with no prediction (see Methods, section Molecular Features).

The only binding site feature that correlates significantly with the absence of prediction is a low volume of the binding site (Fig. 11A), as calculated by PyVOL[64] (see Methods, section Molecular Features). On the contrary, a low number of water molecules within the binding site is not particularly detrimental. Metal ions usually stabilize the phosphate group and occupy some volume in the binding site (it is correlated with a low volume of the binding site and a low number of water molecules). Although it is removed from each protein target, its absence in the calculations is not particularly detrimental either.

Among the conformational features, none is an impacting feature (Supplementary section: Molecular features, Supplementary Fig. 11). It is noteworthy that the syn conformation is not associated with the lack of prediction (Supplementary Table 1) while the initial conformation of all nucleotides is anti, confirming the quality of the MCSS sampling. On the other hand, three interaction features are negatively impacting the performance: the absence of salt bridges, the presence of clashes with water molecules, and to a lesser extent, the absence of stacking contact (Fig. 11B). Among these latter contacts, the π-π interactions are those which more contribute to the negative impact on the predictions (Supplementary Fig. 12). The presence of clashes with water molecules might induce some distortions within the binding site during the protein target's preparation.

If we focus on the non-predicted cases specific to the STDW model with the R310 patch, the observations described above remain valid with very similar trends for all the molecular features (Supplementary section: Molecular features, Supplementary Table 2). Nevertheless, the syn conformations are slightly more frequent in the no-prediction cases (Supplementary

Table 2), indicating a less efficient sampling for the larger R310 patch in size. In the non-optimal predictions which fail to score the native nucleotide as the best ranked (i.e., good predictions, Fig. 10), similar trends are again observed but with two specificities associated with the metals and stacking contacts (Supplementary section: Molecular features, Supplementary Table 3). First, metals' presence negatively impacts the performance suggesting that metals contribute directly or indirectly to the nucleotide selectivity. Second, the absence of stacking contacts makes it more challenging to score the native nucleotide properly; the binding selectivity of purines versus pyrimidines, in particular, can be easier to identify in the presence of stacking contacts.

As described above, a low volume of the binding site is detrimental *per se* to the prediction performance. Once the experimental structure is optimized after removal of the ligand (metal and the water molecules in the SCAL model), the volume can sometimes undergo large variations: either decreasing or increasing (Supplementary section: Molecular features, Supplementary Fig. 13). The average variation shrinks the binding site by 27 to 30Å^3 for the SCAL and STDW models, respectively. In two-thirds of the benchmark, the binding site shrinks by an average of 87 (SCAL) to 92Å^3 (STDW). In one-third of the benchmark, the binding site expands by an average of 92 (STDW) to 95Å^3 (SCAL). Thus, a similar trend of variations is observed for both SCAL and STDW models. However, only the STDW is significantly impacted in the performance for the prediction of the Top-10 (Supplementary section: Molecular features, Supplementary Table 4); the shrinking of the binding site combined with the presence of water molecules prevents the identification of any native pose in the Top-10 in the concerned cases. This is confirmed by the fact that 9 of the 17 proteins in the subset with no predictions in the Top-10 exhibit recovered predictions in the upper Top-i with a smaller patch such as R110 (Supplementary section: Molecular features, Supplementary Table 5). In six other cases, the absence of predictions with the STDW model can be imputed to the presence of water molecules (Supplementary Table 5). Finally, only two cases do not provide any prediction in the Top-i.

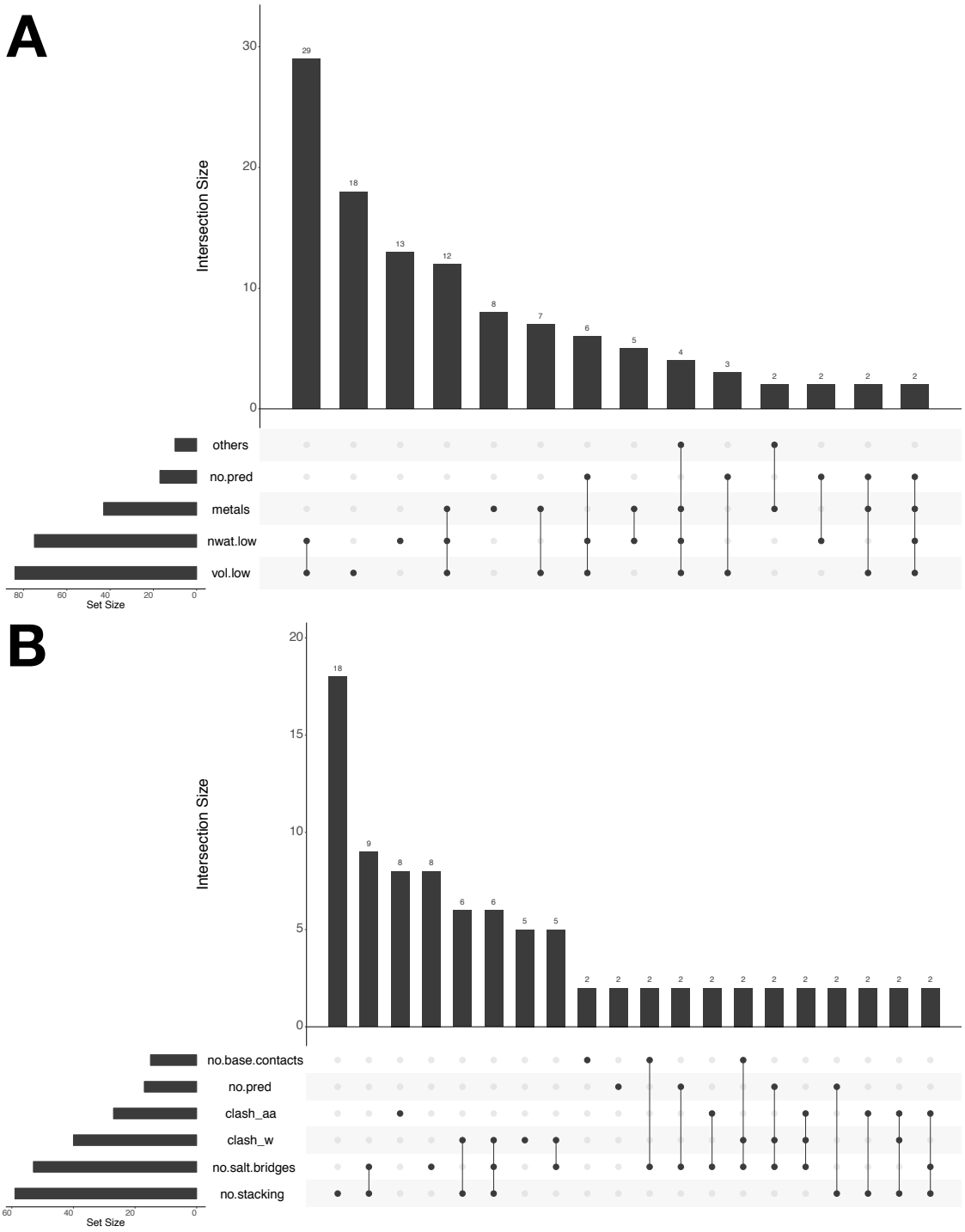


Figure 11: Upset diagrams of the impact of molecular features on the Top-10 predictions. A. binding site features. B. interaction features. The intersections with only one member are not shown; others: presence of additional nucleotidic (nucleic acid) fragment in the binsing site; no.pred: no prediction; 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; 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.

3.6 Case Studies

The analysis of the molecular features that impact the docking and screening powers shows that different factors are responsible for the general lack of performance of all the solvent models and, more specifically, that of the purely implicit SCAL model. We illustrate the impacting features through a series of case studies, looking particularly at those contributing to the improved performance of the hybrid models, including explicit water molecules. We refer to each case using the PDB ID. Since all protein-nucleotide complexes in the benchmark include crystallized water molecules, we should expect that the water-mediated contacts will be detrimental to the SCAL model. Indeed, the presence of water contacts involving the base or the phosphate group has a powerful impact (Supplementary Fig. 14).

In the 1S68 case where the native nucleotide is A, a single water molecule and only one is involved in two close contacts with the nucleotide (Fig. 12). These two water-mediated contacts involve the Watson-Crick face of the adenine. The SCAL model does not provide any prediction within the Top-100 for any nucleotide (Fig. 12A). A few native-like poses exist, but they are not ranked within the Top-100, i.e., their MCSS score is higher than any of the first 100 non-native poses (Fig. 12B,D-E). On the contrary, the STDW model generates several native-like poses within the Top-1, Top-5, and Top-50 corresponding to optimal and good predictions for the native nucleotide (Fig. 12A,C,E-F). Excluding the water-mediated contact with the base, all the other native contacts are found in the native-like poses for both models (Fig. 12D-F). Both water molecules and the solvent model used are responsible for the differential scoring between the two models even if the water molecules are not considered in the scoring (see Methods, section MCSS). In all the native-like poses in both SCAL and STDW models, the syn conformation of the nucleotide is preferred (or a high-syn conformation). However, the starting conformation in the initial distributions is C3'-endo anti. It is indicative that the ligand's flexibility allows to switch from anti to syn during the MCSS calculations without any hindrance.

In the SCAL model, the best-ranked poses for the native nucleotide (Top-1 to Top-10)

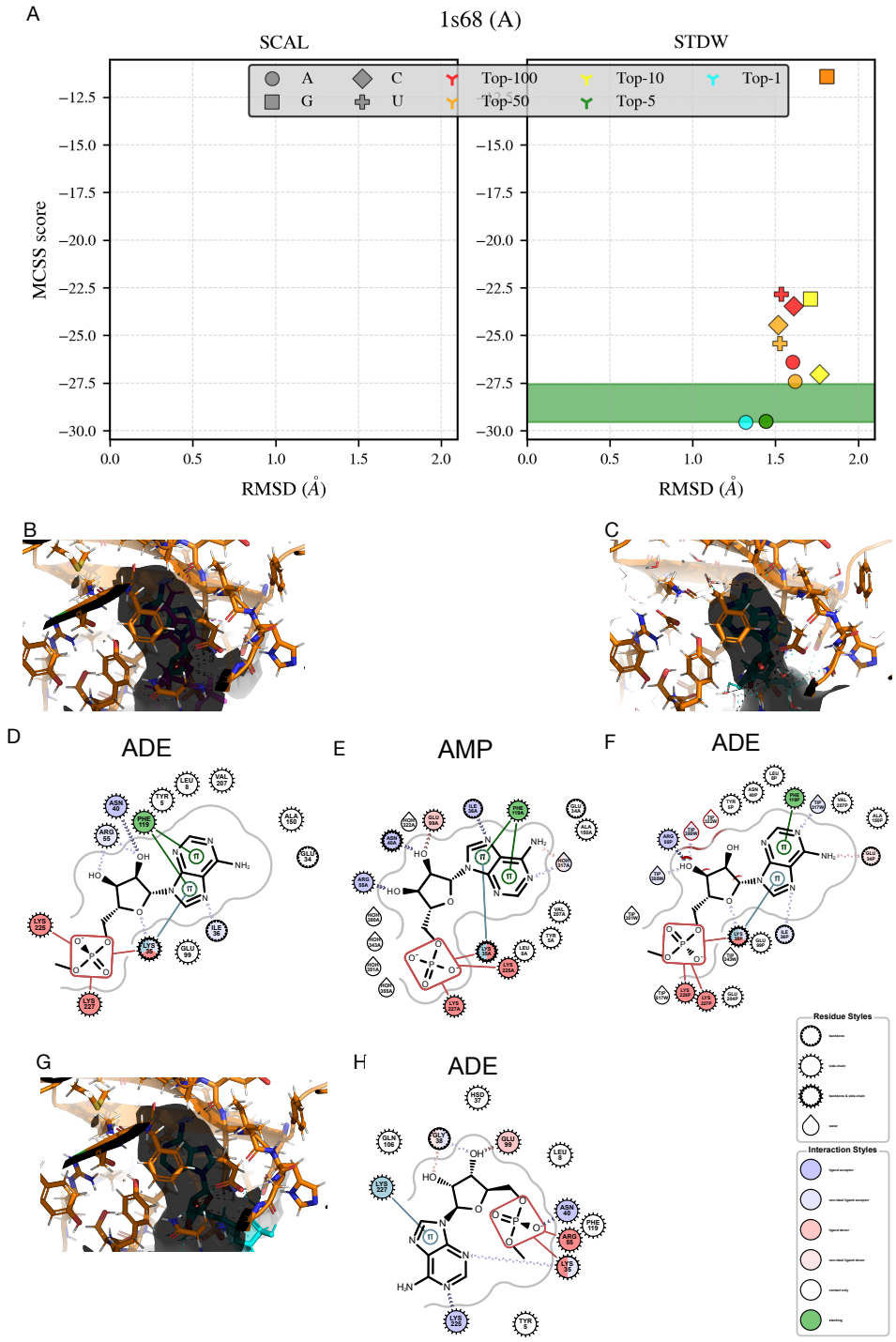


Figure 12: Case Study 1S68. A. Schematic representation of the nucleotide selectivity in the SCAL (left) and STDW (right) models; B. 3D representation of the native-like pose for the SCAL model (128th scored pose); C. 3D representation of the native-like pose for the STDW model (Top-1); D. Diagram of the binding site and nucleotide contacts for the SCAL model (see B); E. Diagram of the binding site and nucleotide contacts for the native binding mode; F. Diagram of the binding site and nucleotide contacts for the STDW model (see C); G. 3D representation of the native-like pose for the SCAL model (Top-1); H. Diagram of the binding site and nucleotide contacts for the SCAL model (see G).

exhibit alternative positionings of the phosphate groups; it interacts closely on the opposite side of the binding site with residues Lys35, Asn40, and Arg55 instead of Lys35, Lys225, and Lys227 (Fig. 12G-H). Although the net charge is reduced by two-thirds in the SCAL model (but not in the STDW model), the absence of explicit solvation around the phosphate group leads to an alternate positioning of the nucleotide (Fig. 12H), which is incompatible with the native binding mode. 1S68 is associated with a low volume of the binding site, a feature that impacts the performance negatively (Fig. 11A). Furthermore, the binding site is shrunk for both models and slightly more pronounced for the SCAL model (Supplementary section: Molecular features, Supplementary Fig. 13).

The 3EWY case is peculiar with a pyrimidic ligand: U, which adopts a syn conformation. Both models exhibit a similar performance with predicting the U native nucleotide (Fig. 13). However, the native contacts are better reproduced by the SCAL model (Fig. 13A-B,E-F) and only the 10th pose reproduces all the native contacts with the residues of the binding site in the STDW model (Fig. 13A,D,F-G). Besides, several non-native nucleotides have very similar MCSS scores: A in the case of the SCAL model, G in the case of the STDW model. There is no global shrinking of the binding site in the protein structures optimized with or without water molecules in the case of 3EWY (Supplementary section: Molecular features, Supplementary Fig. 13). However, there is a local contraction in some parts of the binding site, which is not equivalent between the two models. It is more pronounced on the Hoogsteen and Watson-Crick faces of the base in STDW, making it more challenging to reproduce the native contacts with the base (Fig. 13C). On the other hand, non-native nucleotides can fit into the remodeled binding site with scores that are within the 2 kcal/mol range from the native one (Fig. 14). The native contacts which are specific to the base are lost in the case of A with the SCAL model (Fig. 14A-B,D), but native-like and isosteric contacts are found in the case of G with the STDW model (Fig. 14C,E-F).

In the 2XBU case, the cavity of the binding site is well conserved in the SCAL model but slightly shrunk in the STDW model (Supplementary section: Molecular features, Sup-

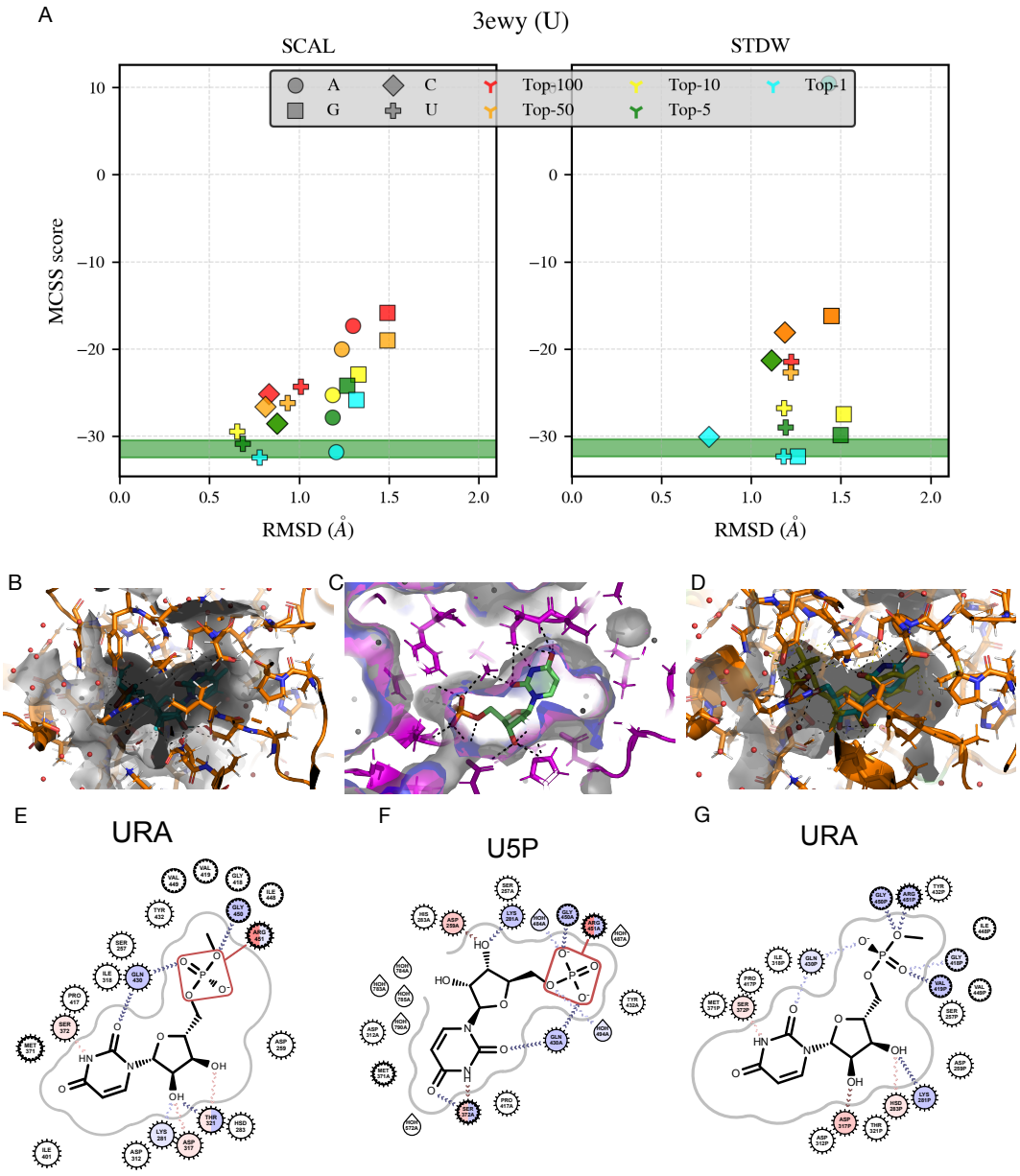


Figure 13: Case Study 3EWY. A. Schematic representation of the nucleotide selectivity in the SCAL (left) and STDW (right) models; B. 3D representation of the Top-1 native pose for the SCAL model; C. 3D representation of the native ligand in the binding site as seen: in the experimental structure (grey), in the optimized structure without water molecules (magenta), in the optimized structure with water molecules (blue); D. 3D representation of the Top-10 native pose for the STDW model; E. Diagram of the binding site and nucleotide contacts for the SCAL model (see B); F. Diagram of the binding site and nucleotide contacts for the native binding mode; G. Diagram of the binding site and nucleotide contacts for the STDW model (see D).

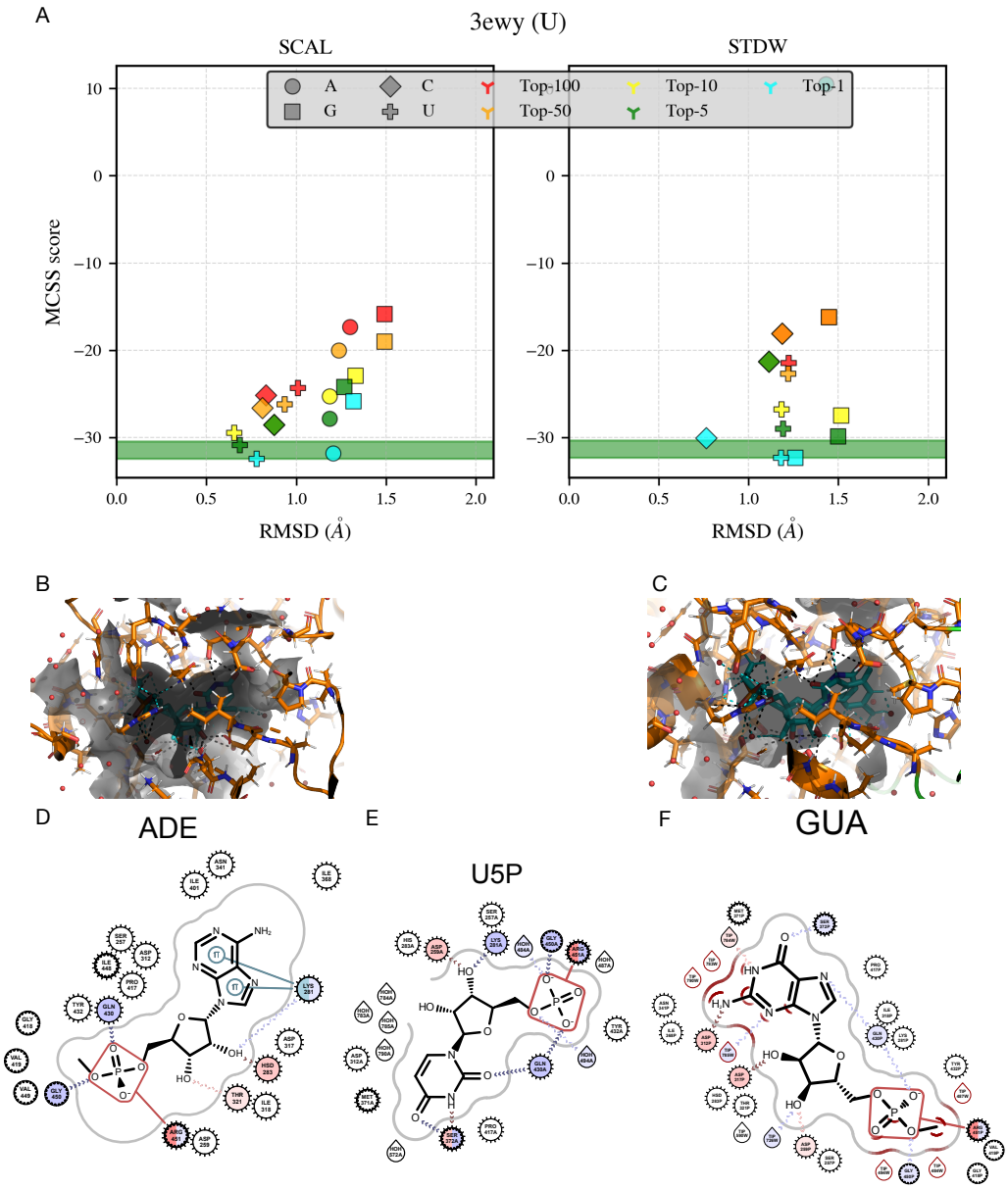


Figure 14: Case Study 3EWY. A. Schematic representation of the nucleotide selectivity in the SCAL (left) and STDW (right) models; B. 3D representation of the Top-1 native-like pose for A (syn conformation) in the SCAL model; C. 3D representation of the Top-1 native-like pose for G (anti conformation) in the STDW model; D. Diagram of the binding site and nucleotide contacts for the SCAL model (see B); F. Diagram of the binding site and nucleotide contacts for the native binding mode; G. Diagram of the binding site and nucleotide contacts for the STDW model (see C).

plementary Fig. 13). The binding site's volume is low because it is quite open with only the base moiety within a well-defined cavity. Only the STDW model provides a good prediction (Top-5) while the native poses generated by the SCAL model are all over the Top-100 scores (Fig. 15A). The first poses in the STDW model (Top-1 to Top-4) are all located in the binding site. However, their RMSD is over 2 and are thus excluded from the native poses. Independent of the scores, the native poses reproduce the native contacts with the base in both models (Fig. 15B-G). The phosphate group establishes very close contacts with hydrogen-bond donors from the peptide backbone, but those contacts are not retrieved in the native poses except for one residue (Thr115). The presence of a terminal methyl group in the phosphate patch used: R310 (Fig. 1 and Fig. 3) prevents a native positioning of the phosphate group. In the SCAL model, its positioning is more in agreement with the experimental structure (Fig. 15E-F). However, the Top-1 pose and the other best-scored poses are completely off-site (Fig. 16). Combined with the shrinking of the binding site, the large phosphate group largely deviates from the expected position in the STDW model, and the only contact with Thr115 is weaker. This deviation on the phosphate group increases the global RMSD of the pose to the native coordinates and leads to exclude the Top-1 to Top-4 G poses from the list of native poses (Fig. 16). However, the Top-4 G pose reproduces almost all the native contacts (Fig. 16C,E-F).

4 Conclusions

MCSS was evaluated for the docking of nucleotides on a benchmark of 121 protein complexes. Different solvent and phosphate models were tested to optimize the success rate for identifying native poses (docking power) and the true native nucleotide (screening power). As a result, the STDW model appears to give the best performance slightly ahead of the SCALW model based on partially reduced charges on the phosphate group. A clustering procedure was set up that allows a slight increase of the success rates, especially in the high

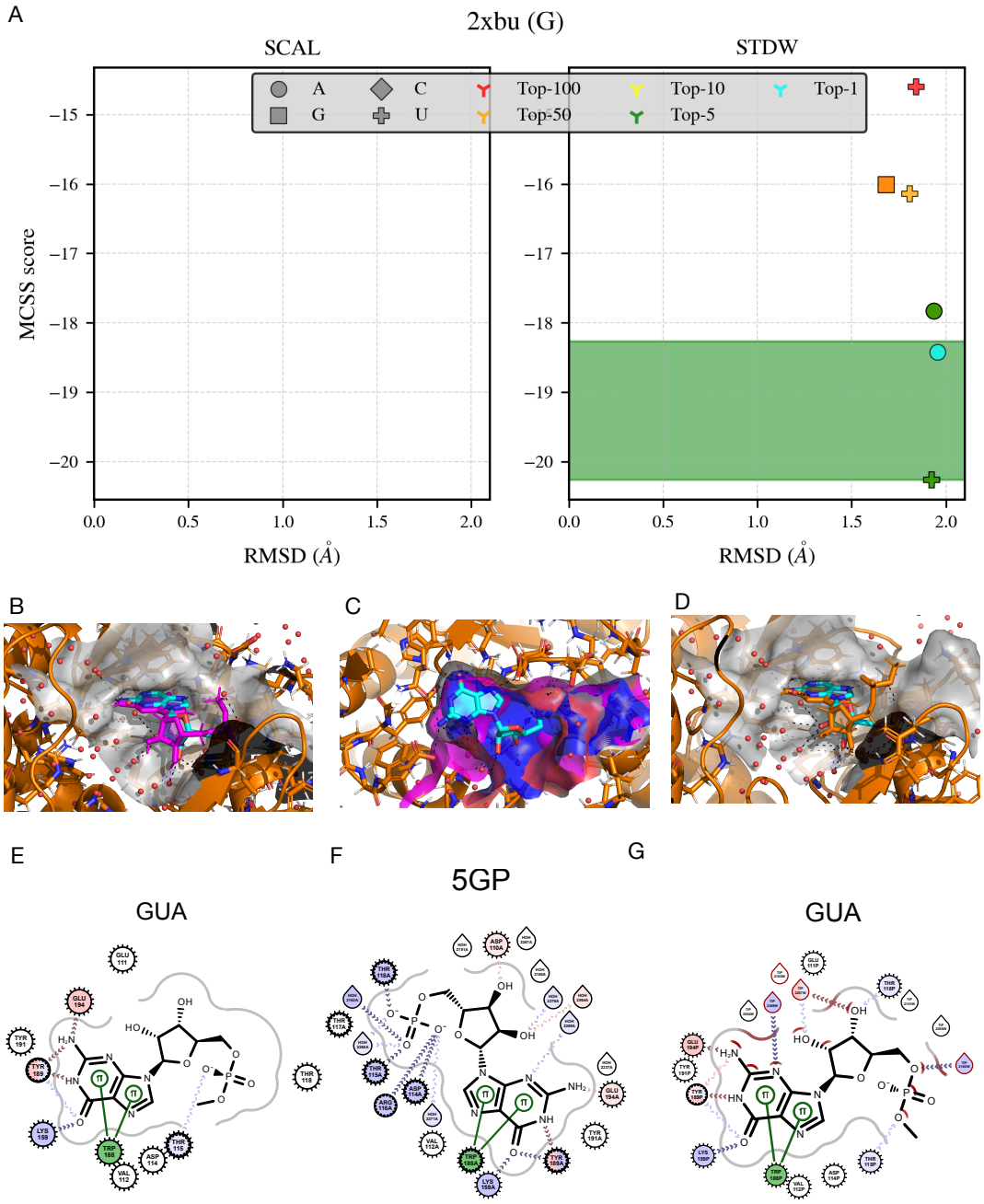


Figure 15: Case Study 2XBU. A. Schematic representation of the nucleotide selectivity in the SCAL (left) and STDW (right) models; B. 3D representation of a native-like pose for G in the SCAL model (169th scored pose); C. 3D representation of the native ligand in the binding site as seen: in the experimental structure (grey), in the optimized structure without water molecules (magenta), in the optimized structure with water molecules (blue); D. 3D representation of a native-like pose for G in the STDW model (Top-50, 12th scored pose); E. Diagram of the binding site and nucleotide contacts for the SCAL model (see B); F. Diagram of the binding site and nucleotide contacts for the native binding mode; G. Diagram of the binding site and nucleotide contacts for the STDW model (see D).

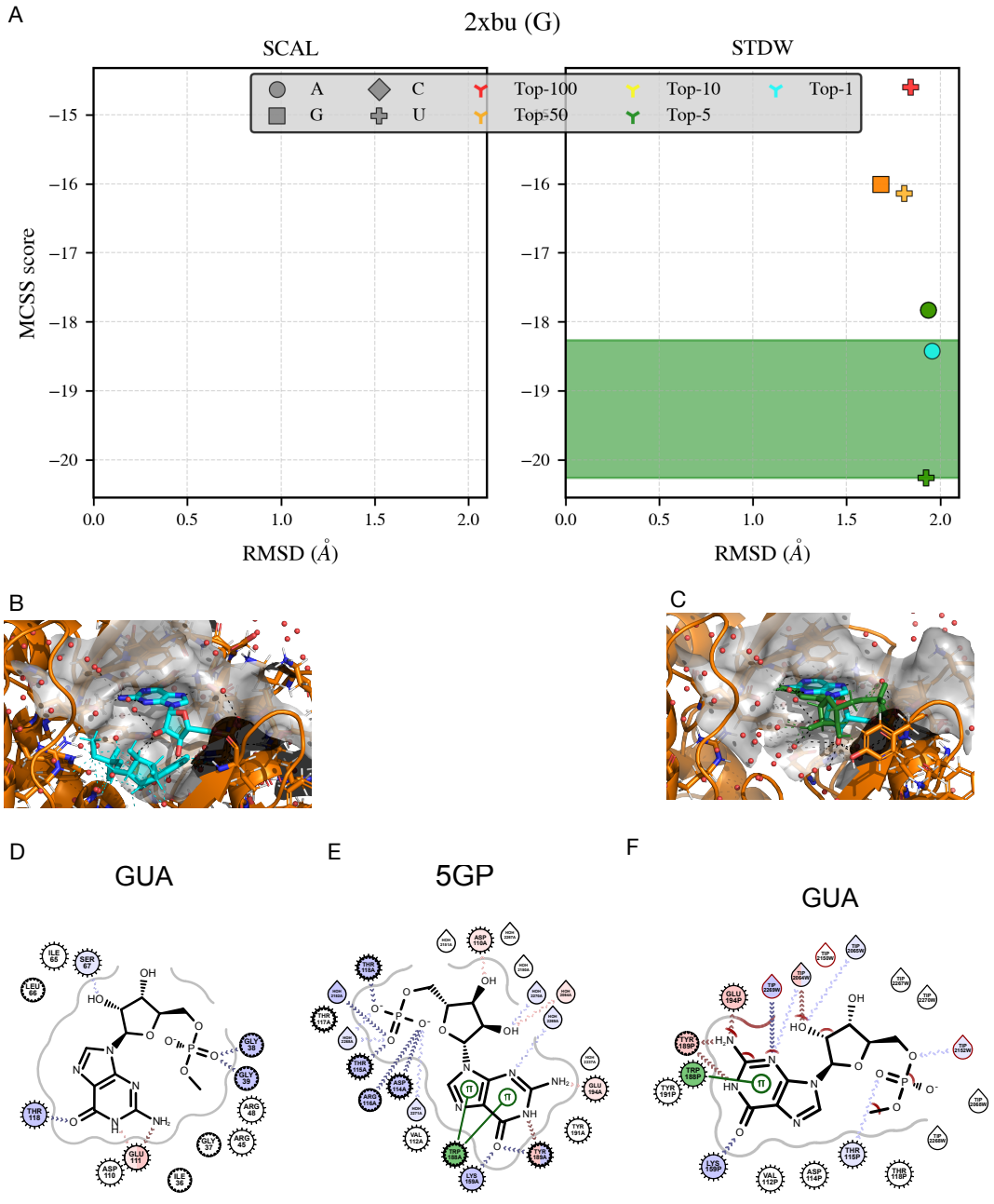


Figure 16: Case Study 2XBU. A. Schematic representation of the nucleotide selectivity in the SCAL (left) and STDW (right) models; B. 3D representation of Top-1 pose for G in the SCAL model; C. 3D representation of Top-4 pose for G in the STDW model (RMSD $>$ 2.0); D. Diagram of the binding site and nucleotide contacts for the SCAL model (see B); E. Diagram of the binding site and nucleotide contacts for the native binding mode; F. Diagram of the binding site and nucleotide contacts for the STDW model (see C).

Top-i (Top-50 and Top-100). Among the different phosphate models, the more voluminous one that carries a terminal methyl group: R310, is slightly better in the Top-1 predictions. It is also the phosphate model that facilitates the linking of nucleotide fragments in the perspective of fragment-based design of oligonucleotides (unpublished data[52]). The combined STDW-R310 model outperforms despite the few cases where the lack of predictions in the Top-10 could be correlated to a size effect that prevents the phosphate group from fitting correctly in the binding site. The size effect is reinforced by the shrinking of the binding site due to the removal of the ligand and other bound molecules or metals from the protein target. The presence of water molecules in the preparation and optimization of the protein structure allows the minimized structure to deviate less from the experimental structure. On the other hand, the water molecules generally induce a more pronounced shrinking of the binding site with respect to the experimental structure, which is responsible for some degradation of the performance. The inclusion of water molecules gives a more realistic description of the binding site, whether they are involved in water-mediated contacts with the ligand or just solvating the phosphate group or ribose.

We have identified some pitfalls that contribute to degrade the performance of prediction in all models. From the intrinsic features of the binding site, a low binding volume is the more impacting factor. It can be seen as a low accessible volume for close contacts that typically occurs when the binding site is open with few contacts with the ligand or when the close contacts are only present in some part of the nucleotide (small binding cavities). Among the conformational features, the syn conformation does not have any negative impact although the docking is performed using an initial C3'-endo anti conformation. This observation confirms that the flexibility of the ligand during the docking allows a proper conformational sampling even if it is a bit less efficient with the R310 patch due to some size effect. Among the interaction features, the presence of salt bridges makes it a bit easier to get good predictions. On the other hand, the presence of clashes with water molecules in the experimental structure has a slight negative impact. More specific to the STDW-R310

model, the negative effect of the low binding volume is smoothed.

The quality of the scoring explains, to some extent, the better performance of the hybrid model STDW over the implicit model SCAL. First of all, the SCAL model includes a slight bias in the scoring, favoring G over A and the other nucleotides, while A is the nucleotide from the benchmark that establishes the stronger contacts in the binding site. Furthermore, the STDW model based on the original parameters from the CHARMM forcefield describes better, in the presence of water molecules, the bonded contributions associated with the torsion angles. Thus, the penalty term of the MCSS score from the conformational distortions of the bound ligand is more accurate. The STDW model outperforms not only in docking power with more predictions but also in screening power. The STDW model has a much stronger discriminatory power between very similar ligands. It is also consistent with the broader range of score distributions for each type of nucleotide. The native poses scored as optimal reproduce most (if not all) of the native contacts as well as the good predictions, although they are not ranked first among the four nucleotides.

Both free and bound conformations for the same protein are not available on a large set of 3D structures with nucleotide ligands. Thus, the protein targets correspond to some unbound forms where the ligands were extracted from the binding site. Consequently, the optimized binding site is usually shrunk, making the identification of native poses and native binders more challenging. The method's performance is then degraded both in terms of docking and screening powers because of missing native contacts in the shrunk areas of the binding sites. The four standard nucleotides are very similar from the chemical viewpoint and thus harder to discriminate in terms of binding selectivity. Chemical modifications would increase the dissimilarity between nucleotidic fragments, which would likely be easier to discriminate. The absence of stacking interactions, one of the impacting feature, makes the prediction more difficult. So, we may also expect that modified nucleotides favoring stacking interactions will be more accurately predicted. Many modified nucleotides are already used experimentally in the synthesis of oligonucleotides[77] or modified aptamers[46] for medical or biotechnological

applications to improve specific properties such as the therapeutic index.[78]

From the perspective of designing oligonucleotides, MCSS provides a good performance to predict native poses (docking power) and identify the binding selectivity of nucleotidic fragments (screening power). The MCSS-STDW model provides the best performance for the docking power in Top-1 and Top-100 with the R310 patch that generates the lowest pose redundancy. Using this large patch may lower the performance in particular cases due to some size effect (e.g., tight binding site). However, the R310 patch increases the performance of the linking strategy (data not published[52]). The molecular features analysis indicates we should optimize the binding volume with tight contacts such as salt bridges or stacking interactions. More accurate or full descriptions of the solvent open the possibility of further improving the docking and screening powers. Other improvements may come from using more relevant (unbound) and more diverse (active) conformations of the protein target to avoid the biases associated with possible distortions of the binding site (e.g., shrinking). Finally, increasing the chemical diversity of the nucleotidic fragments should help find a better molecular complementarity with the target; at the same time, we expect that the evaluation of the binding selectivity will be more accurate for modified nucleotides.

Acknowledgement

NC was supported by the French Ministry of Higher Education, Research and Innovation. RGA is supported by the Excellence Eiffel Ph.D. Program. We thank the French Ministry of Foreign Affairs for financial support (PHC Carlos J. Finlay, 41814TM).

Supporting Information Available

- Supplementary illustrations (figures & tables) on the benchmark composition, MCSS results and analysis (PDF)

- Raw data (CSV, TXT, SVG) on the benchmark composition, MCSS results and analysis for each protein-nucleotide complex (ZIP)

References

(1) Hajduk, P. J.; Greer, J. A decade of fragment-based drug design: strategic advances and lessons learned. *Nature Reviews Drug Discovery* **2007**, *6*, 211–219.

(2) Murray, C. W. The rise of fragment-based drug discovery. *Nature Chemistry* **2009**, *1*, 187–192.

(3) Baker, M. Fragment-based lead discovery grows up. *Nature Reviews Drug Discovery* **2013**, *12*, 5–7.

(4) Price, A. J.; Howard, S.; Cons, B. D. Fragment-based drug discovery and its application to challenging drug targets. *Essays in biochemistry* **2017**, *61*, 475–484.

(5) Shuker, S. B.; Hajduk, P. J.; Meadows, R. P.; Fesik, S. W. Discovering high-affinity ligands for proteins: SAR by NMR. *Science (New York, NY)* **1996**, *274*, 1531–1534.

(6) Erlanson, D. A.; Fesik, S. W.; Hubbard, R. E.; Jahnke, W.; Jhoti, H. Twenty years on: the impact of fragments on drug discovery. *Nature Reviews in Drug Discovery* **2016**, *15*, 605–619.

(7) Brown, D. G.; Boström, J. Where Do Recent Small Molecule Clinical Development Candidates Come From? *Journal Of Medicinal Chemistry* **2018**, *61*, 9442–9468.

(8) Zoete, V.; Grosdidier, A.; Michielin, O. Docking, virtual high throughput screening and in silico fragment-based drug design. *Journal of Cellular and Molecular Medicine* **2009**, *13*, 238–248.

(9) Hoffer, L.; Renaud, J.-P.; Horvath, D. Fragment-based drug design: computational & experimental state of the art. *Combinatorial Chemistry & High Throughput Screening* **2011**, *14*, 500–520.

(10) Robson-Tull, J. Biophysical screening in fragment-based drug design: a brief overview. *Bioscience Horizons: The International Journal of Student Research* **2018**, *11*, 1324.

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

(12) Kumar, A.; Voet, A.; Zhang, K. Y. J. Fragment based drug design: from experimental to computational approaches. *Current Medicinal Chemistry* **2012**, *19*, 5128–5147.

(13) Klon, Anthony E, In *Fragment-Based Methods in Drug Discovery*; Klon, A. E., Ed.; Methods in Molecular Biology; Springer New York: New York, NY, 2015; Vol. 1289.

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

(15) Su, M.; Yang, Q.; Du, Y.; Feng, G.; Liu, Z.; Li, Y.; Wang, R. Comparative Assessment of Scoring Functions: The CASF-2016 Update. *Journal Of Chemical Information And Modeling* **2019**, *59*, 895–913.

(16) Miranker, A.; Karplus, M. Functionality maps of binding sites: A multiple copy simultaneous search method. *Proteins: Structure, Function, and Bioinformatics* **1991**, *11*, 29–34.

(17) Elber, R.; Society, M. K. J. o. t. A. C.; 1990, Enhanced sampling in molecular dynamics: use of the time-dependent Hartree approximation for a simulation of carbon monoxide diffusion through myoglobin. *Journal of the American Chemical Society* **1990**, *112*, 9161–9175.

(18) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry* **1983**, *4*, 187–217.

(19) Caflisch, A.; Miranker, A.; Karplus, M. Multiple copy simultaneous search and construction of ligands in binding sites: application to inhibitors of HIV-1 aspartic proteinase. *Journal Of Medicinal Chemistry* **1993**, *36*, 2142–2167.

(20) Joseph-McCarthy, D.; Fedorov, A. A.; Almo, S. C. Comparison of experimental and computational functional group mapping of an RNase A structure: implications for computer-aided drug design. *Protein engineering* **1996**, *9*, 773–780.

(21) Joseph-McCarthy, D.; Tsang, S. K.; Filman, D. J.; Hogle, J. M.; Karplus, M. Use of MCSS to design small targeted libraries: application to picornavirus ligands. *Journal of the American Chemical Society* **2001**, *123*, 12758–12769.

(22) Eisen, M. B.; Wiley, D. C.; Karplus, M.; Hubbard, R. E. HOOK: A program for finding novel molecular architectures that satisfy the chemical and steric requirements of a macromolecule binding site. *Proteins: Structure, Function, and Bioinformatics* **1994**, *19*, 199–221.

(23) Stultz, C. M.; Karplus, M. Dynamic ligand design and combinatorial optimization: designing inhibitors to endothiapepsin. *Proteins: Structure, Function, and Bioinformatics* **2000**, *40*, 258–289.

(24) Takano, Y.; Koizumi, M.; Takarada, R.; Kamimura, M. T.; Czerminski, R.; Koike, T. Computer-aided design of a factor Xa inhibitor by using MCSS functionality maps and a CAVEAT linker search. *Journal of Molecular Graphics and Modelling* **2003**, *22*, 105–114.

(25) So, S. S.; Karplus, M. Evaluation of designed ligands by a multiple screening method:

application to glycogen phosphorylase inhibitors constructed with a variety of approaches. *Journal of Computer-Aided Molecular Design* **2001**, *15*, 613–647.

(26) Huang, Z.; Zhang, M.; Burton, S. D.; Katsakhyan, L. N.; Ji, H. Targeting the Tcf4 G13ANDE17 binding site to selectively disrupt β-catenin/T-cell factor protein-protein interactions. *ACS Chemical Biology* **2014**, *9*, 193–201.

(27) Caflisch, A.; Schramm, H. J.; Karplus, M. Design of dimerization inhibitors of HIV-1 aspartic proteinase: a computer-based combinatorial approach. *Journal of Computer-Aided Molecular Design* **2000**, *14*, 161–179.

(28) Haider, M. K.; Bertrand, H.-O.; Hubbard, R. E. Predicting fragment binding poses using a combined MCSS MM-GBSA approach. *Journal Of Chemical Information And Modeling* **2011**, *51*, 1092–1105.

(29) Haider, K.; Huggins, D. J. Combining Solvent Thermodynamic Profiles with Functionality Maps of the Hsp90 Binding Site to Predict the Displacement of Water Molecules. *Journal Of Chemical Information And Modeling* **2013**, *53*, 2571–2586.

(30) 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.

(31) Onufriev, A. V.; Case, D. A. Generalized Born Implicit Solvent Models for Biomolecules. *Annual Review of Biophysics* **2019**, *48*, 275–296.

(32) Pantsar, T.; Poso, A. Binding Affinity via Docking: Fact and Fiction. *Molecules* **2018**, *23*, 1899.

(33) Bancet, A.; Raingeval, C.; Lomberget, T.; Le Borgne, M.; Guichou, J.-F.; Krimm, I. Fragment Linking Strategies for Structure-Based Drug Design. *Journal Of Medicinal Chemistry* **2020**, *63*, 11420–11435.

(34) Murray, C. W.; Verdonk, M. L. *Fragment-based Approaches in Drug Discovery*; John Wiley Sons, Ltd, 2006; Chapter 3, pp 55–66.

(35) Borsi, V.; Calderone, V.; Fragai, M.; Luchinat, C.; Sarti, N. Entropic contribution to the linking coefficient in fragment based drug design: a case study. *Journal Of Medicinal Chemistry* **2010**, *53*, 4285–4289.

(36) Yu, H. S.; Modugula, K.; Ichihara, O.; Kramschuster, K.; Keng, S.; Abel, R.; Wang, L. General Theory of Fragment Linking in Molecular Design: Why Fragment Linking Rarely Succeeds and How to Improve Outcomes. *Journal of Chemical Theory and Computation* **2021**, *17*, 450–462.

(37) Singh, J.; Saldanha, J.; Thornton, J. M. A novel method for the modelling of peptide ligands to their receptors. *Protein engineering* **1991**, *4*, 251–261.

(38) Elkin, C. D.; Zuccola, H. J.; Hogle, J. M.; Joseph-McCarthy, D. Computational design of D-peptide inhibitors of hepatitis delta antigen dimerization. *Journal of Computer-Aided Molecular Design* **2000**, *14*, 705–718.

(39) Zeng, J.; Nheu, T.; Zorzet, A.; Catimel, B.; Nice, E.; Maruta, H.; Burgess, A. W.; Treutlein, H. R. Design of inhibitors of Ras–Raf interaction using a computational combinatorial algorithm. *Protein engineering* **2001**, *14*, 39–45.

(40) Rinaldi, A. RNA to the rescue: RNA is one of the most promising targets for drug development given its wide variety of uses. *EMBO reports* **2020**, *21*, e51013.

(41) Wang, F.; Zuroske, T.; Watts, J. K. RNA therapeutics on the rise. *Nature Reviews in Drug Discovery* **2020**, *19*, 441–442.

(42) Zhu, S.; Rooney, S.; Michlewski, G. RNA-Targeted Therapies and High-Throughput Screening Methods. *International Journal of Molecular Sciences* **2020**, *21*, 2996.

(43) Siddiqui, M. A. A.; Keating, G. M. Pegaptanib: in exudative age-related macular degeneration. *Drugs* **2005**, *65*, 1571–7– discussion 1578–9.

(44) Ng, E. W. M.; Adamis, A. P. Anti-VEGF aptamer (pegaptanib) therapy for ocular vascular diseases. *Annals of the New York Academy of Sciences* **2006**, *1082*, 151–171.

(45) Zhou, J.; Rossi, J. Aptamers as targeted therapeutics: current potential and challenges. *Nature Reviews in Drug Discovery* **2017**, *16*, 181–202.

(46) Pfeiffer, F.; Rosenthal, M.; Siegl, J.; Ewers, J.; Mayer, G. Customised nucleic acid libraries for enhanced aptamer selection and performance. *Current Opinion in Biotechnology* **2017**, *48*, 111–118.

(47) Duo, J.; Chiriac, C.; Huang, R. Y.-C.; Mehl, J.; Chen, G.; Tymiak, A.; Sabbatini, P.; Pillutla, R.; Zhang, Y. Slow Off-Rate Modified Aptamer (SOMAmer) as a Novel Reagent in Immunoassay Development for Accurate Soluble Glypican-3 Quantification in Clinical Samples. *Analytical Chemistry* **2018**, *90*, 5162–5170.

(48) Begic, E.; Hadzidedic, S.; Kulaglic, A.; Ramic-Brkic, B.; Begic, Z.; Causevic, M. SOMAscan-based proteomic measurements of plasma brain natriuretic peptide are decreased in mild cognitive impairment and in Alzheimer's dementia patients. *PLoS ONE* **2019**, *14*, e0212261.

(49) Hirota, M.; Murakami, I.; Ishikawa, Y.; Suzuki, T.; Sumida, S.-i.; Ibaragi, S.; Kasai, H.; Horai, N.; Drolet, D. W.; Gupta, S.; Janjic, N.; Schneider, D. J. Chemically Modified Interleukin-6 Aptamer Inhibits Development of Collagen-Induced Arthritis in Cynomolgus Monkeys. *nucleic acid therapeutics* **2016**, *26*, 10–19.

(50) de Beauchene, I. C.; de Vries, S. J.; Zacharias, M. Binding Site Identification and Flexible Docking of Single Stranded RNA to Proteins Using a Fragment-Based Approach. *PLoS computational biology* **2016**, *12*, e1004697.

(51) de Beauchene, I. C.; de Vries, S. J.; Zacharias, M. Fragment-based modelling of single stranded RNA bound to RNA recognition motif containing proteins. *Nucleic Acids Research* **2016**, gkw328.

(52) Chevrollier, N. Développement et application d'une approche de docking par fragments pour modéliser les interactions entre protéines et ARN simple-brin. Theses, Université Paris-Saclay, 2019.

(53) Zhang, Y.; Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research* **2005**, *33*, 2302–2309.

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

(55) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: a web-based graphical user interface for CHARMM. *Journal of Computational Chemistry* **2008**, *29*, 1859–1865.

(56) Tidor, B.; Irikura, K. K.; Brooks, B. R.; Karplus, M. Dynamics of DNA oligomers. *Journal of biomolecular structure & dynamics* **1983**, *1*, 231–252.

(57) Mackerell, A. D.; Banavali, N.; Foloppe, N. Development and current status of the CHARMM force field for nucleic acids. *Biopolymers* **2000**, *56*, 257–265.

(58) Im, W.; Lee, M. S.; Brooks, C. L. Generalized born model with a simple smoothing function. *Journal of Computational Chemistry* **2003**, *24*, 1691–1702.

(59) Chocholoušová, J.; Feig, M. Balancing an accurate representation of the molecular surface in generalized born formalisms with integrator stability in molecular dynamics simulations. *Journal of Computational Chemistry* **2006**, *27*, 719–729.

(60) 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.

(61) 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.

(62) 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.

(63) 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.

(64) Smith, R. H.; Dar, A. C.; Schlessinger, A. PyVOL: a PyMOL plugin for visualization, comparison, and volume calculation of drug-binding sites. *bioRxiv* **2019**, 1–3.

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

(66) Barik, A.; C, N.; P, M.; Bahadur, R. P. A protein-RNA docking benchmark (I): nonredundant cases. *Proteins: Structure, Function, and Bioinformatics* **2012**, *80*, 1866–1871.

(67) Pérez-Cano, L.; Jiménez-García, B.; Fernández-Recio, J. A protein-RNA docking benchmark (II): extended set from experimental and homology modeling data. *Proteins: Structure, Function, and Bioinformatics* **2012**, *80*, 1872–1882.

(68) Huang, S.-Y.; Zou, X. A nonredundant structure dataset for benchmarking protein-RNA computational docking. *Journal of Computational Chemistry* **2012**,

(69) Yang, X.; Li, H.; Huang, Y.; Liu, S. The dataset for protein–RNA binding affinity. *Protein science : a publication of the Protein Society* **2013**, *22*, 1808–1811.

(70) Nithin, C.; Mukherjee, S.; Bahadur, R. P. A non-redundant protein-RNA docking benchmark version 2.0. *Proteins: Structure, Function, and Bioinformatics* **2017**, *85*, 256–267.

(71) Nithin, C.; Ghosh, P.; Bujnicki, J. M. Bioinformatics Tools and Benchmarks for Computational Docking and 3D Structure Prediction of RNA-Protein Complexes. *Genes* **2018**, *9*, 432.

(72) 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 Research* **2000**, *28*, 235–242.

(73) Viji, S. N.; Balaji, N.; Gautham, N. Molecular docking studies of protein-nucleotide complexes using MOLSDOCK (mutually orthogonal Latin squares DOCK). *Journal of Molecular Modeling* **2012**, *18*, 3705–3722.

(74) Jhoti, H.; Williams, G.; Rees, D. C.; Murray, C. W. The 'rule of three' for fragment-based drug discovery: where are we now? *Nature Reviews Drug Discovery* **2013**, *12*, 644–644.

(75) Schneider, B.; Morávek, Z.; Berman, H. M. RNA conformational classes. *Nucleic Acids Research* **2004**, *32*, 1666–1677.

(76) Icazatti, A. A.; Loyola, J. M.; Szleifer, I.; Vila, J. A.; Martin, O. A. Classification of RNA backbone conformations into rotamers using 13C' chemical shifts: exploring how far we can go. *PeerJ* **2019**, *7*, e7904.

(77) Smith, C. I. E.; Zain, R. Therapeutic Oligonucleotides: State of the Art. *Annual Review of Pharmacology and Toxicology* **2019**, *59*, 605–630.

(78) Shen, W.; De Hoyos, C. L.; Migawa, M. T.; Vickers, T. A.; Sun, H.; Low, A.; Bell, T. A.; Rahdar, M.; Mukhopadhyay, S.; Hart, C. E.; Bell, M.; Riney, S.; Murray, S. F.; Greenlee, S.; Crooke, R. M.; Liang, X.-h.; Seth, P. P.; Crooke, S. T. Chemical modification of PS-ASO therapeutics reduces cellular protein-binding and improves the therapeutic index. *Nature Biotechnology* **2019**, *27*, 70.

Supporting Information Available

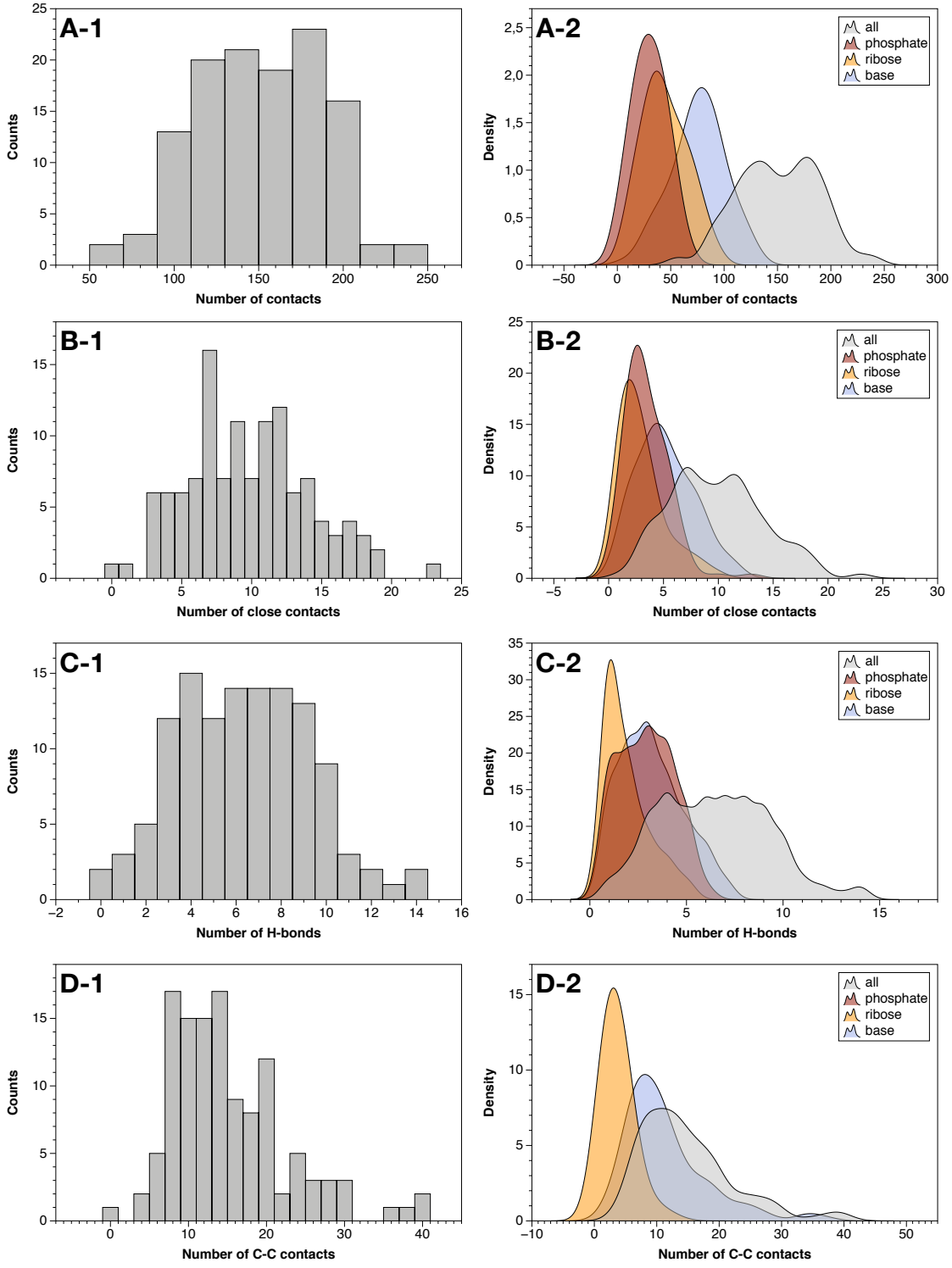
Benchmark of 121 protein-nucleotide complexes

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.

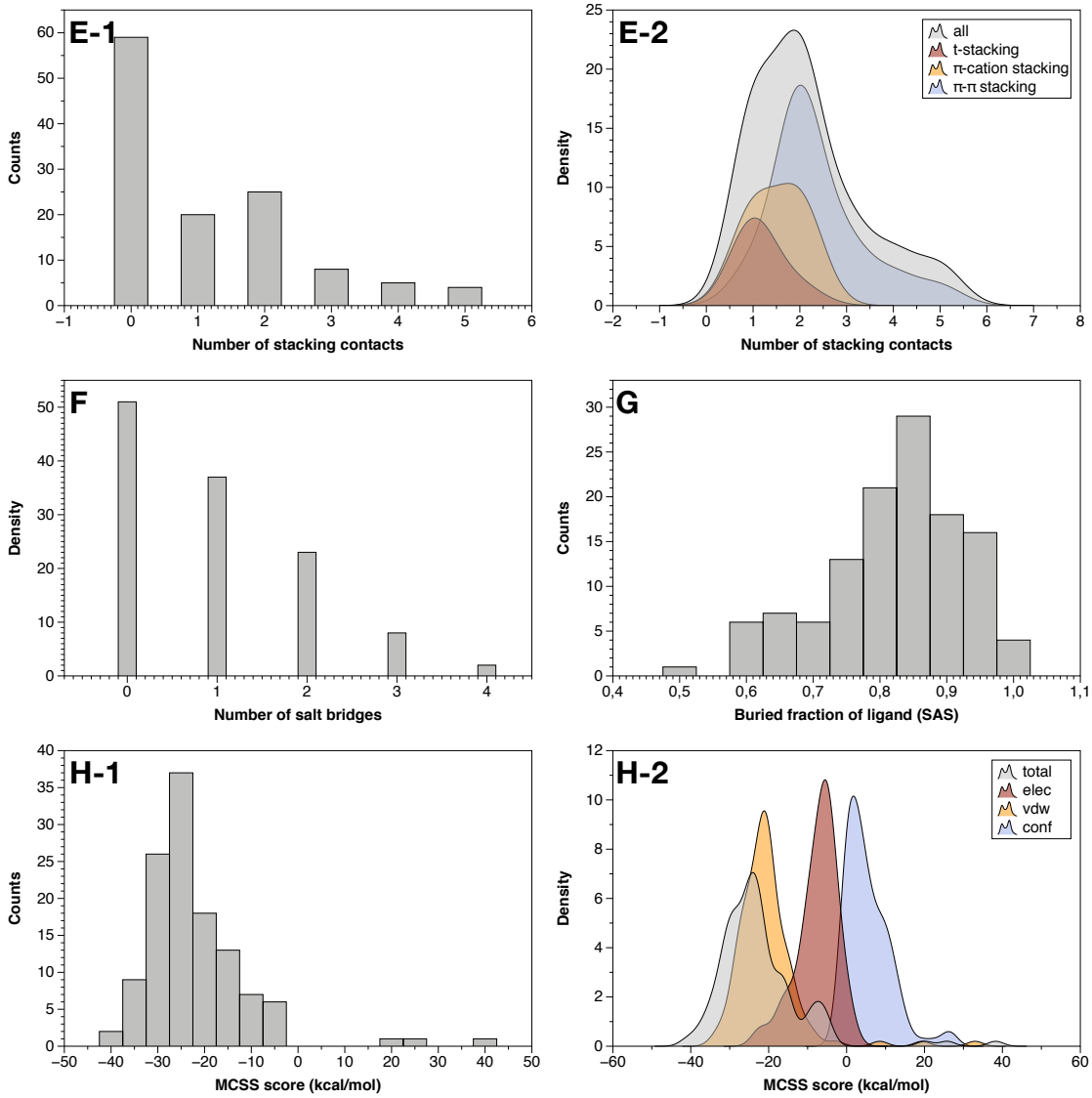
2. Attached Supplementary Data 2 (Data-S2.csv): calculations of the BINANA features (number of contacts, number of H-bonds, the buried fraction of ligand, etc)

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

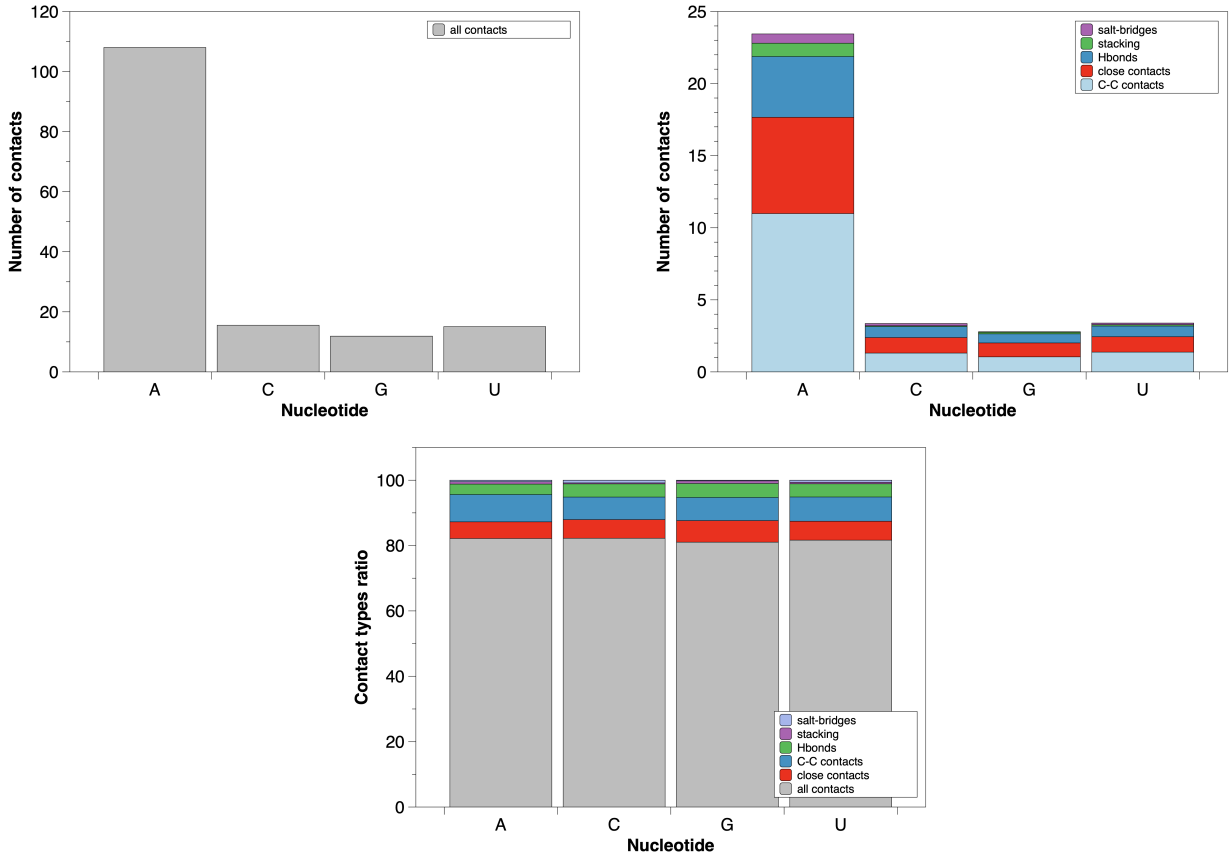
4. Attached Supplementary Data 4 (Data-S4.tar.gz): 2D diagrams of the contacts within the binding sites (SVG format).



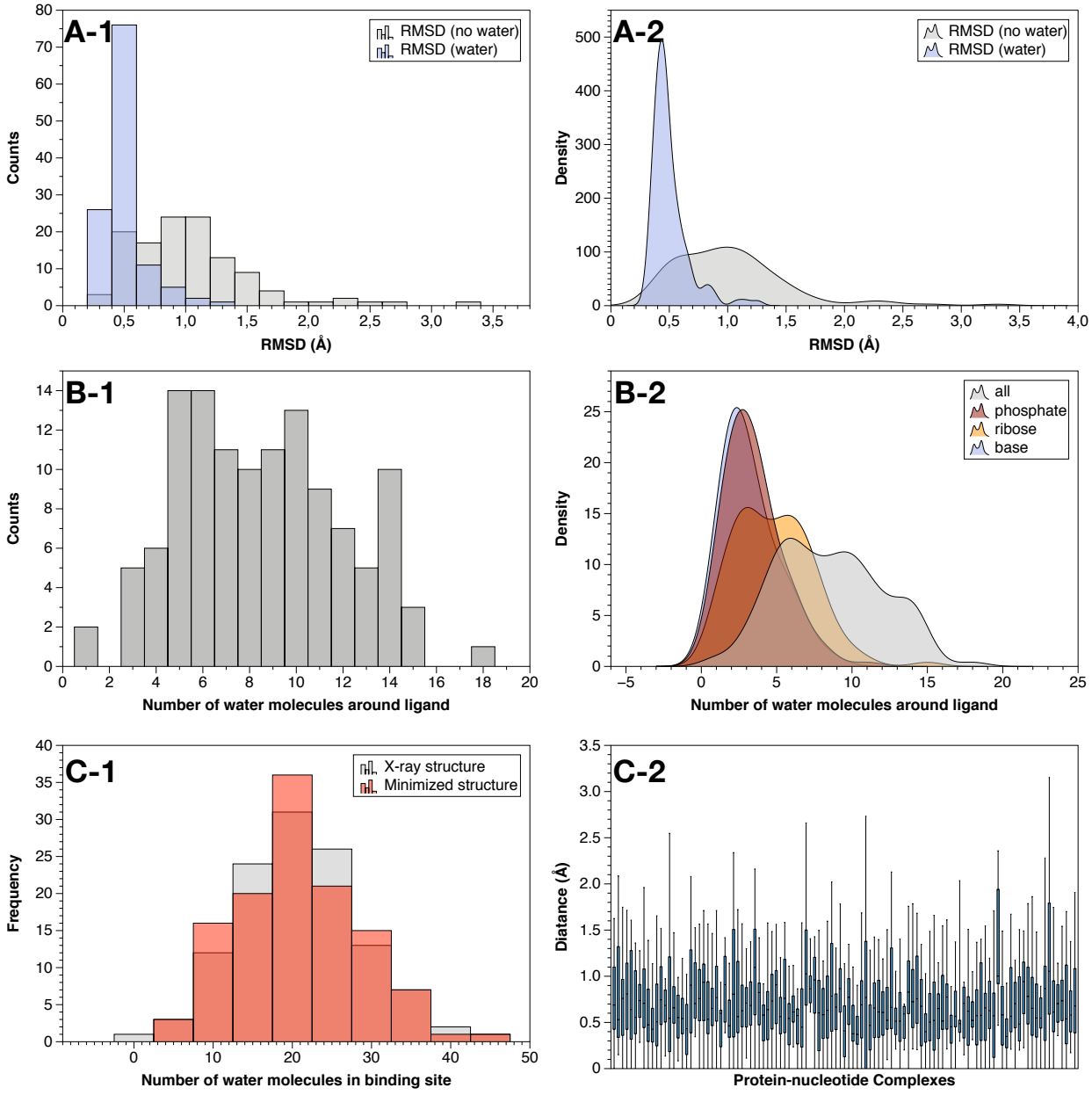
Supplementary Figure 1: Molecular and energy features of the nucleotide-binding sites from the benchmark of 121 complexes. A-1.: Histogram of the number of contacts; A-2: Smooth histogram with decomposition per nucleotide moiety (base, ribose, phosphate); B-1.: Histogram of the number of close contacts; 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;[54] the stacking contributions are calculated from OpenEye;[65] the MCSS score is calculated by the scoring function derived previously.[30]



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.



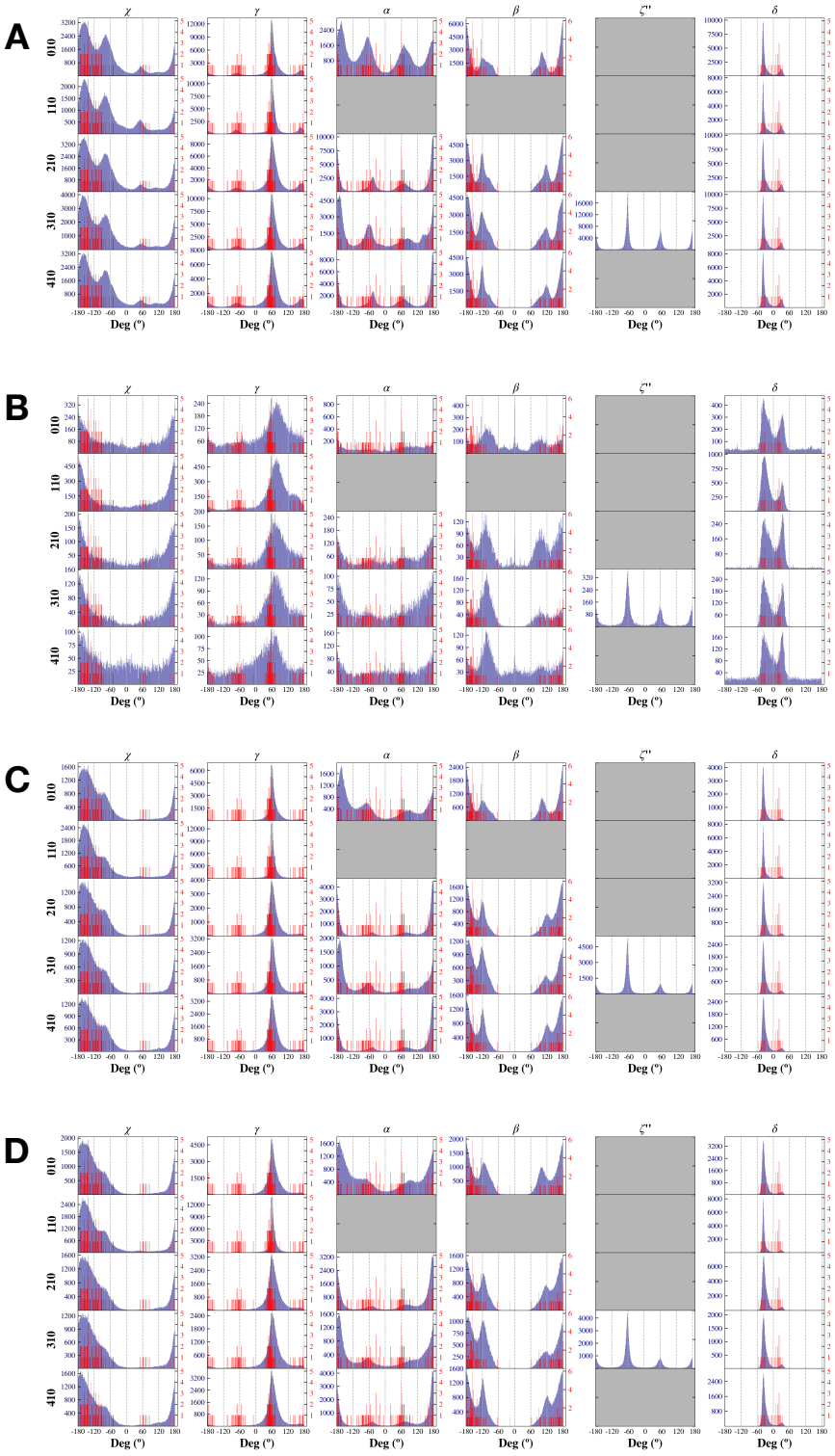
Supplementary Figure 3: Distributions of water molecules and impact on the binding sites. A-1.: Histogram of RMSD in presence/absence of water molecules; A-2.: Same as A-1 with a smooth histogram; B-1.: the number of water molecules around the ligand (distance cutoff of 4.0Å); B-2.: Same as B-1 with 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

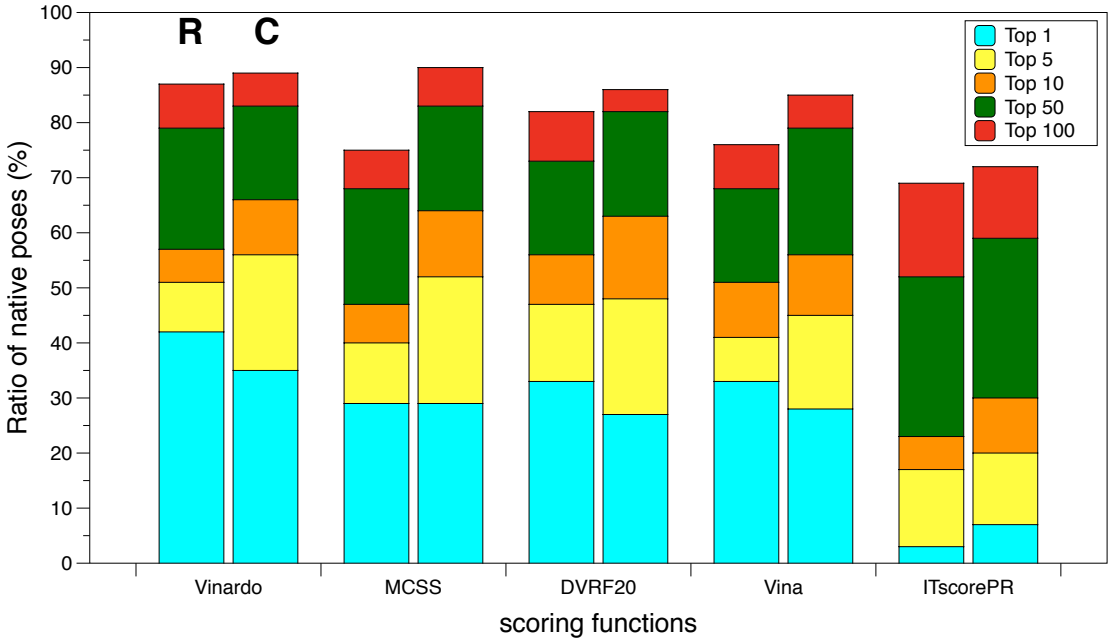


Supplementary Figure 4: Torsions angles. Nonbonded models and associated patches (R010 to R410): A. SCAL, B. FULLW, C. SCALW, D. STDW. In blue: the distribution of the torsions angles observed in the MCSS minima; In red: the distribution of the torsions angles observed in the bound ligands.

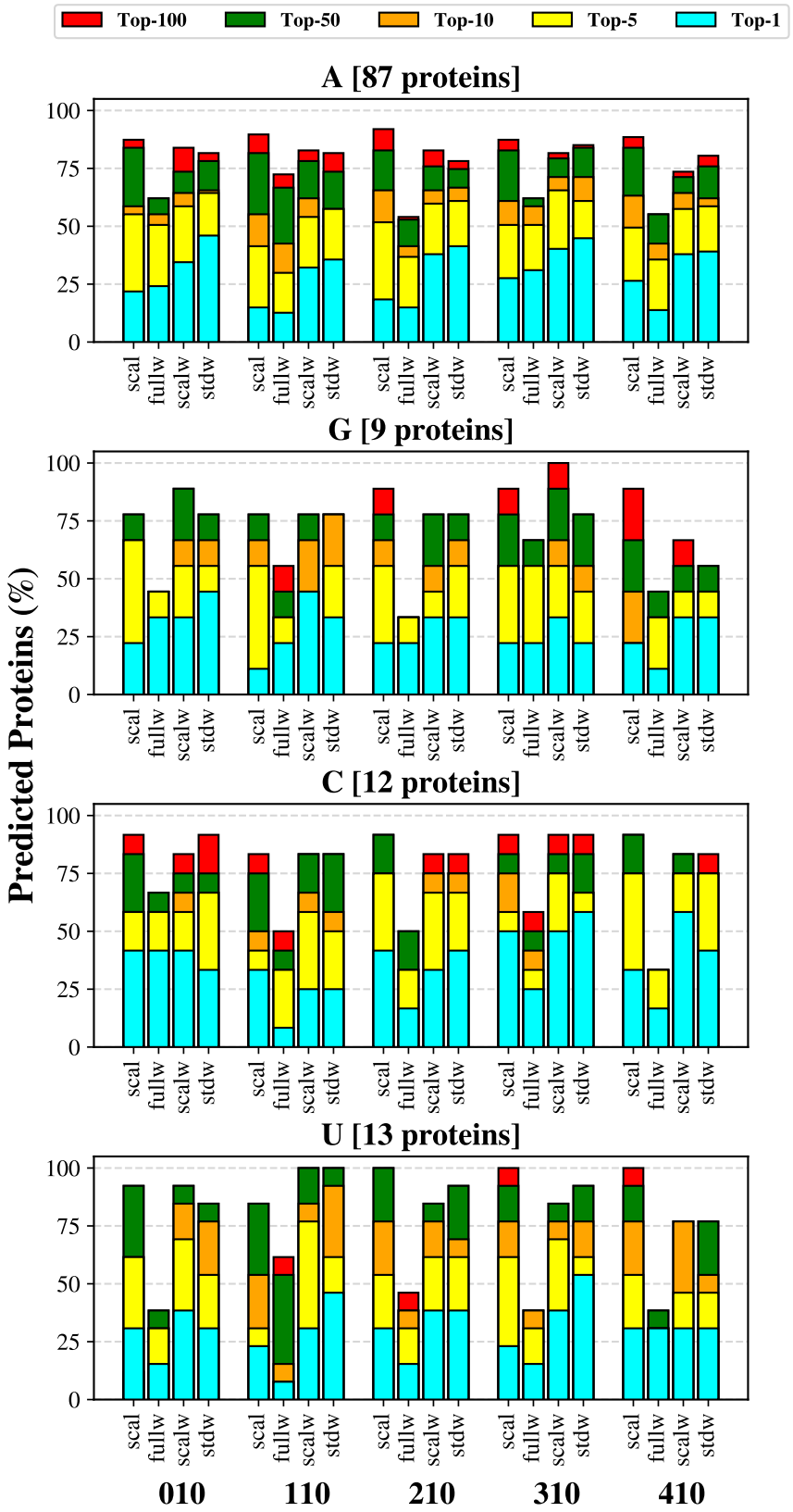
Scoring

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.[14] 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.[61] It was tested in particular on the DUD library that contains, among other proteins, kinases with nucleotide ligands or nucleotide analogs.[11] $\Delta_{vina}RF_{20}$ 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,[62] $\Delta_{vina}RF_{20}$,[63] Autodock Vina score,[60] Vinardo,[61] and the MM-GB models (see Methods) except MCSS[30] correspond to single-point calculations on the MCSS-generated poses.

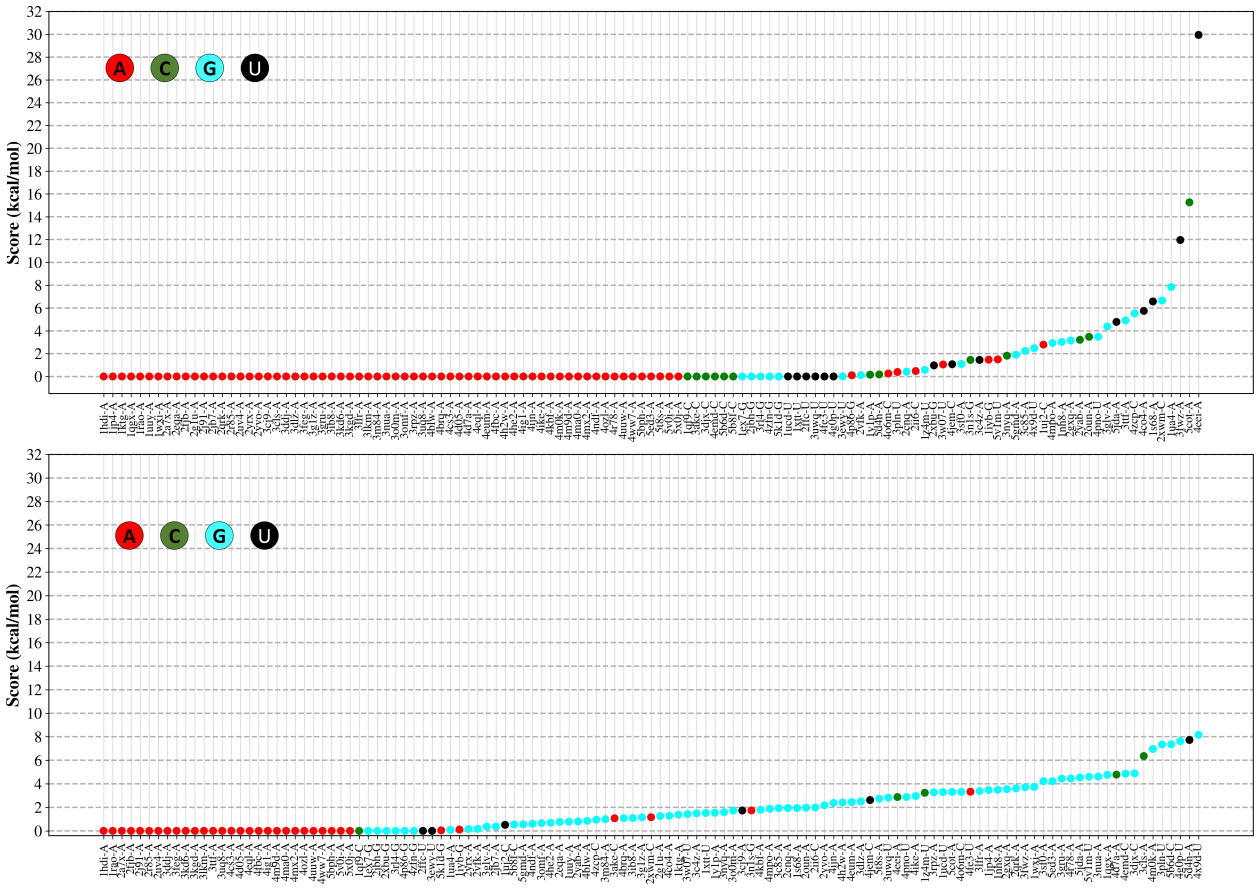
8. Attached Supplementary Data 8 (Data-S8.tar.gz): selectivity diagrams SCAL/STDW for the native poses for each protein-nucleotide complex of the benchmark.



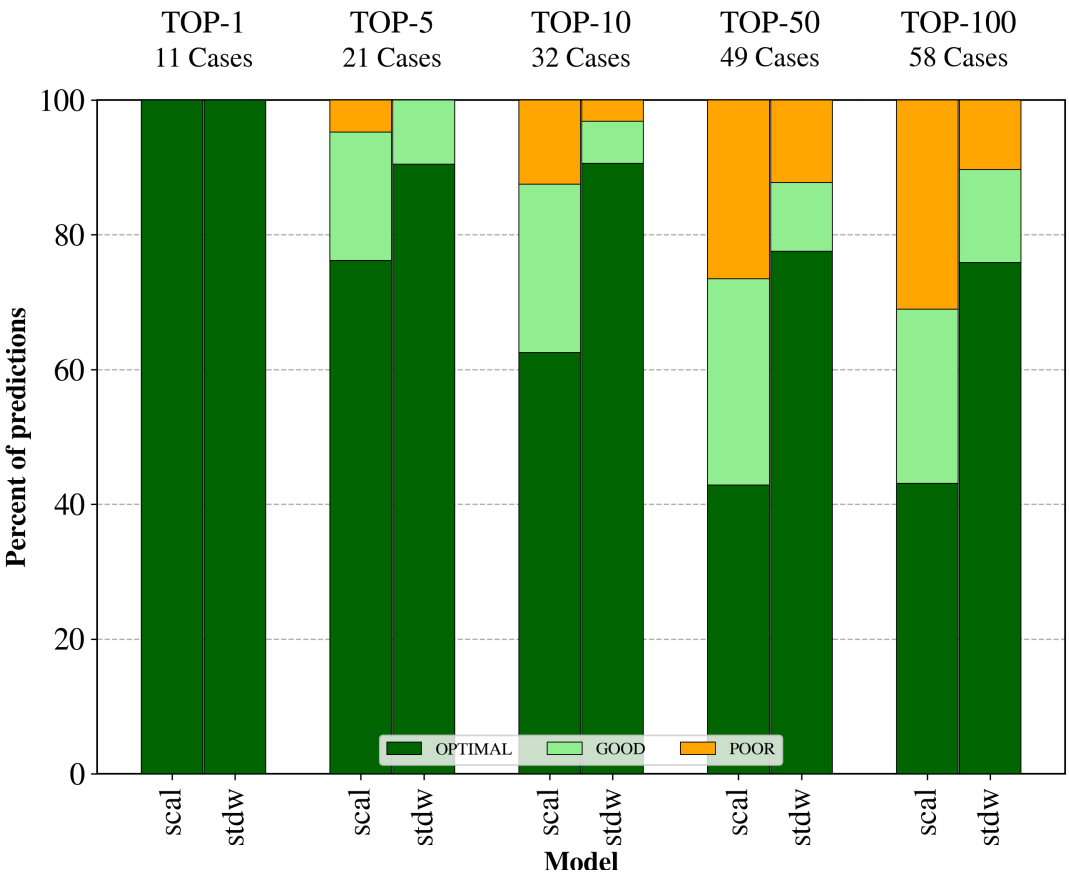
Supplementary Figure 5: Docking powers (top1 to top100) for Vinardo, MCSS, $\Delta_{vina}RF_{20}$, Vina, and ITscorePR and the impact of the clustering filtering (using the patch R310). Left bar (R): no clustering; Right bar (C): clustering.



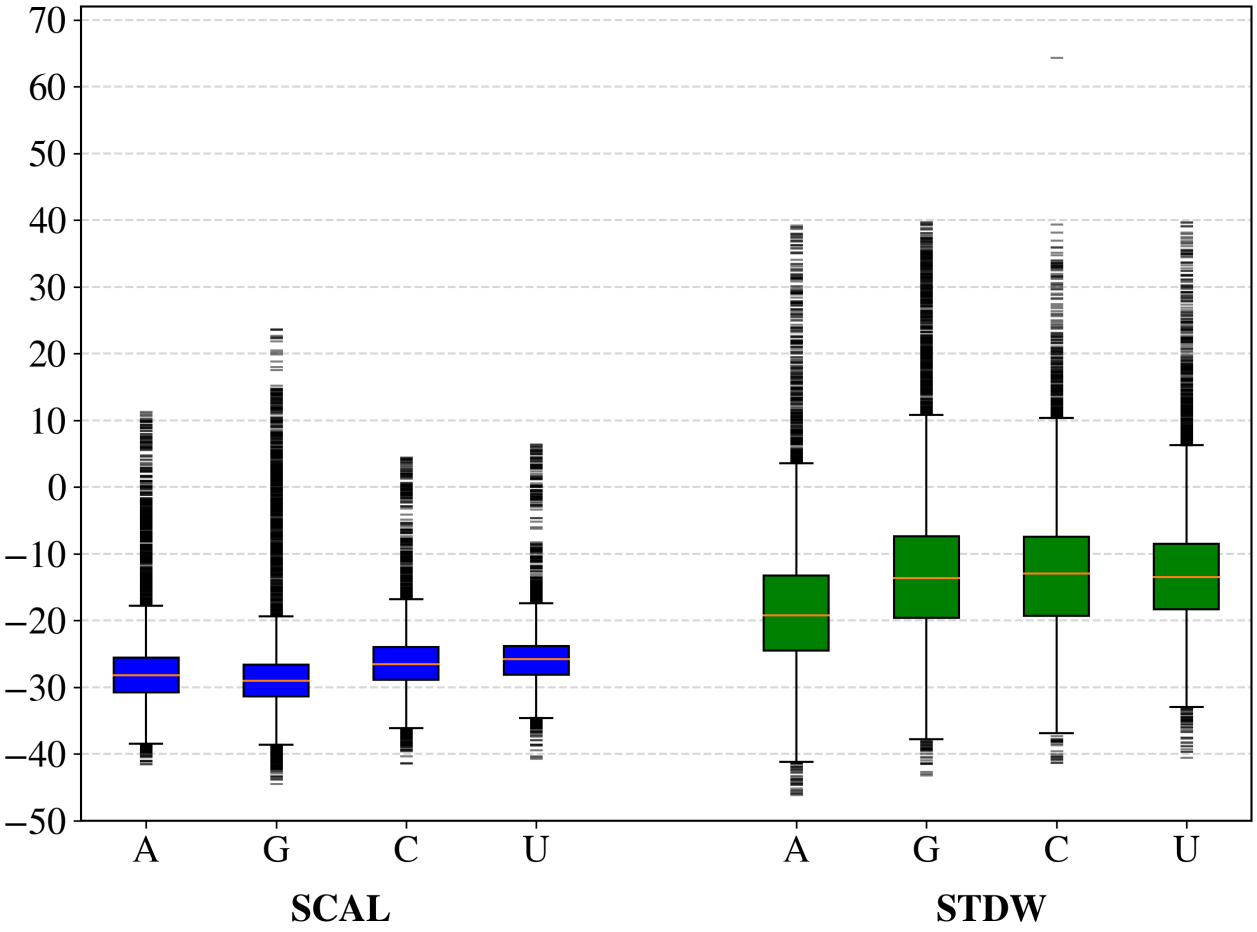
Supplementary Figure 6: Decomposition of docking powers per nucleotide type. The data are shown for the clustered distribution and each Top-i.



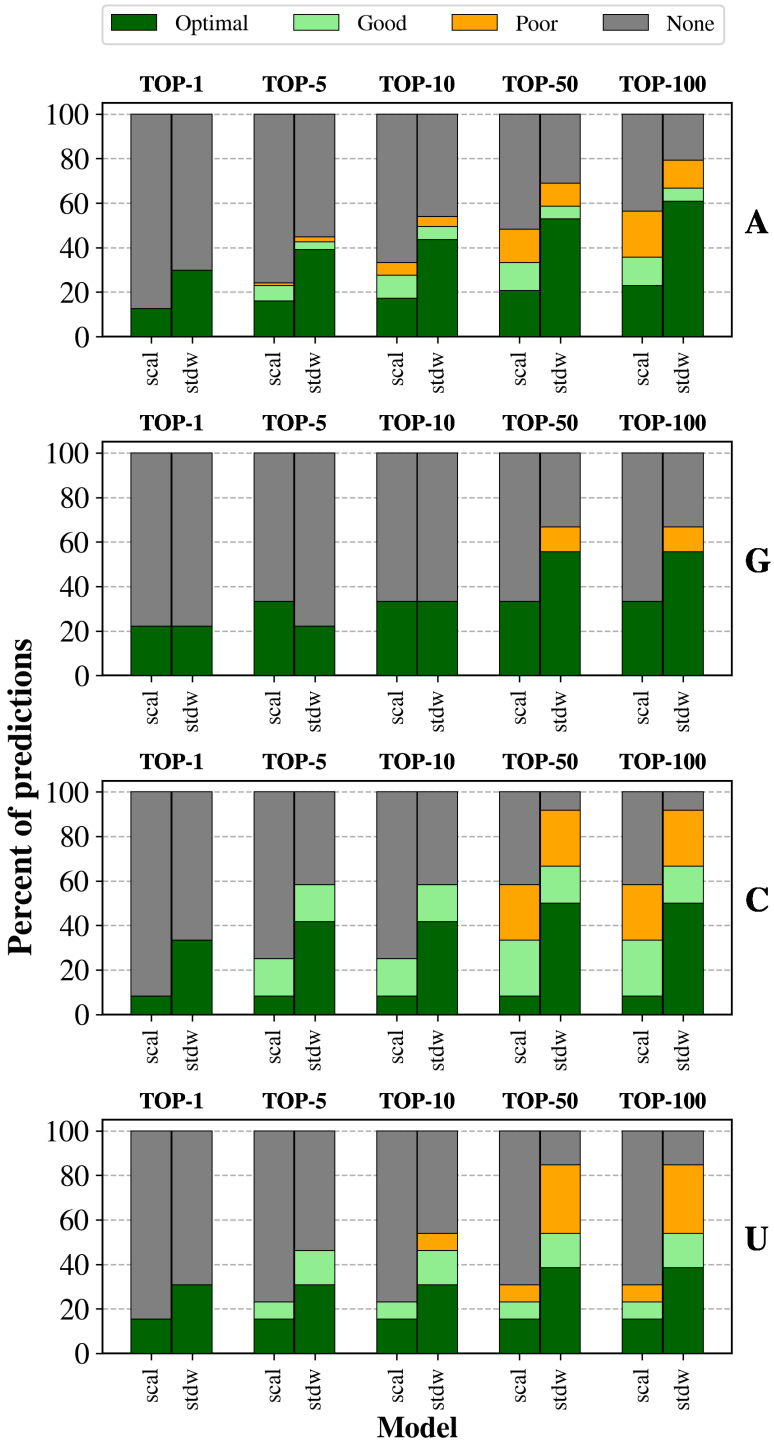
Supplementary Figure 7: Scoring differences (offset) between the best-ranked pose whatever the nucleotide type and the best-ranked pose for the nucleotide corresponding to the native ligand. Top: STDW model; bottom: SCAL model. The color code indicates the nucleotide type.



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).

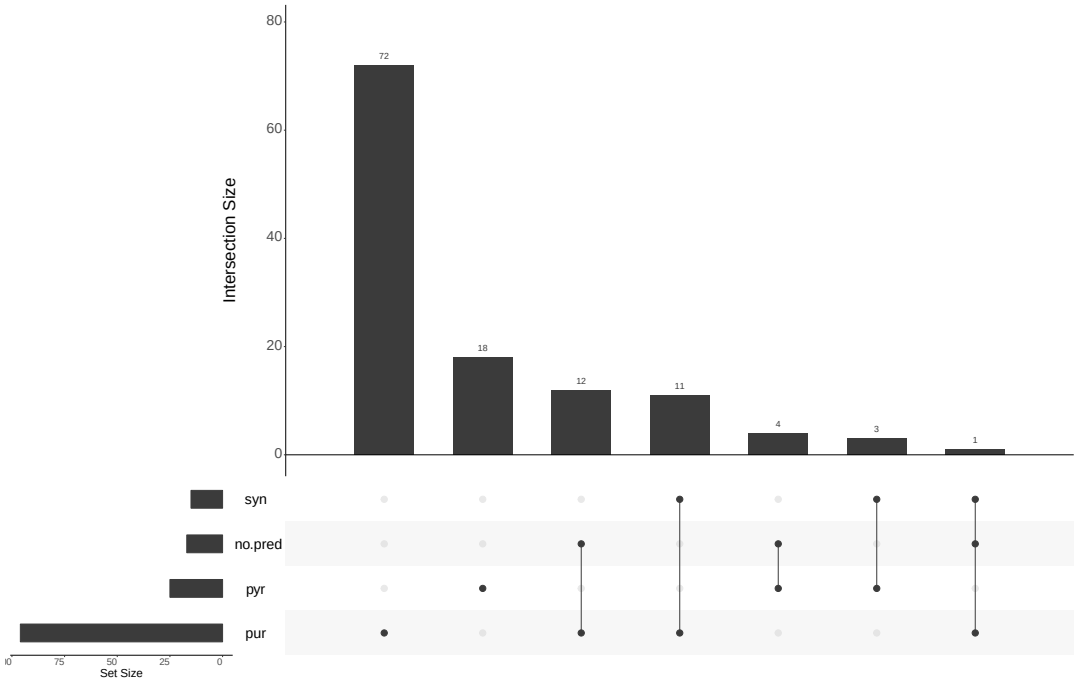


Supplementary Figure 10: Decomposition of screening powers per nucleotide type. Optimal: native nucleotide as the best ranked; good: native nucleotide in the ranked within a 2 kcal/mol range from the best ranked non-native nucleotide; poor: native nucleotide ranked out of the 2 kcal/mol range.

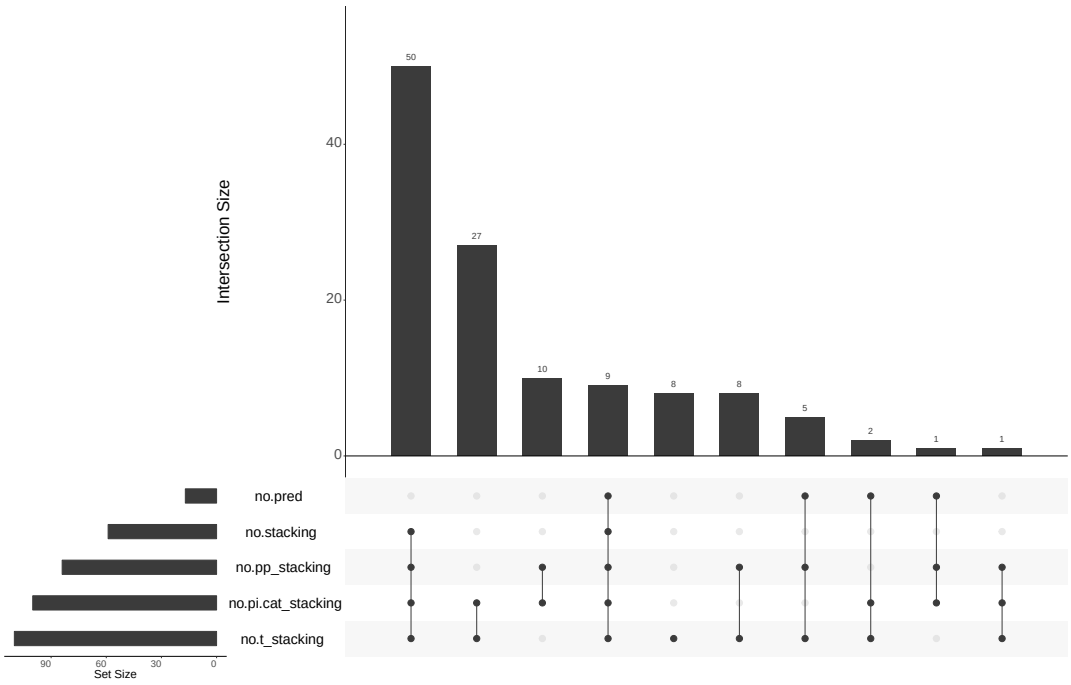
Molecular features

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.

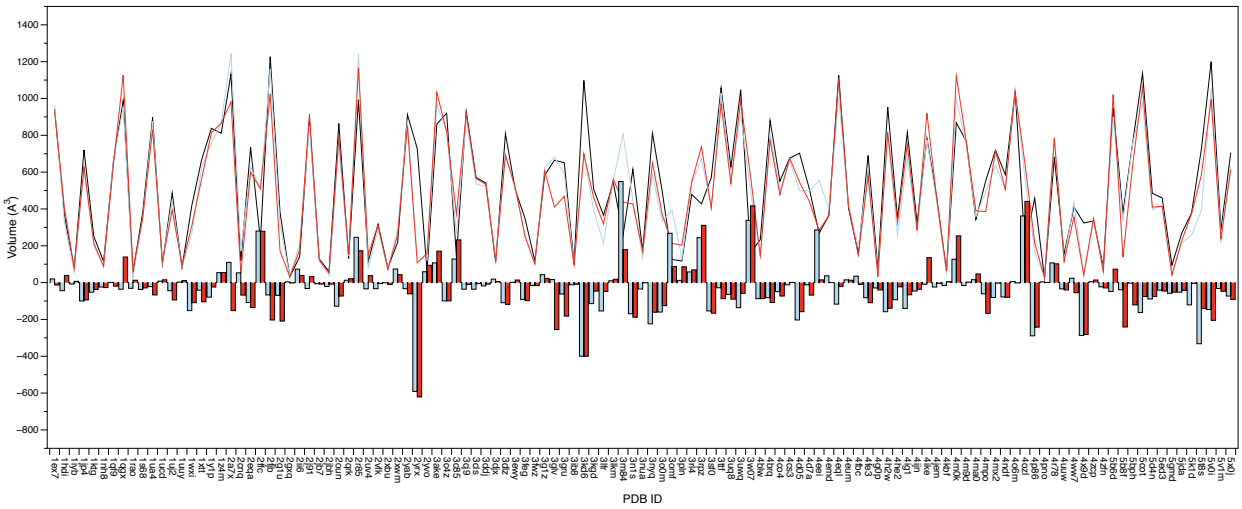
Features		Freq. Benchmark	Freq. good
binding site	nwat.low	62	60
	vol.low	69	70
	others	12	10
	metals	36	60
conformational	syn	12	0
	pur	79	80
	pyr	21	0
interaction	no.base.contacts	12	30
	no.salt.bridges	44	30
	no.stacking	49	70
	clash aa	22	20
	clash w	33	40

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.

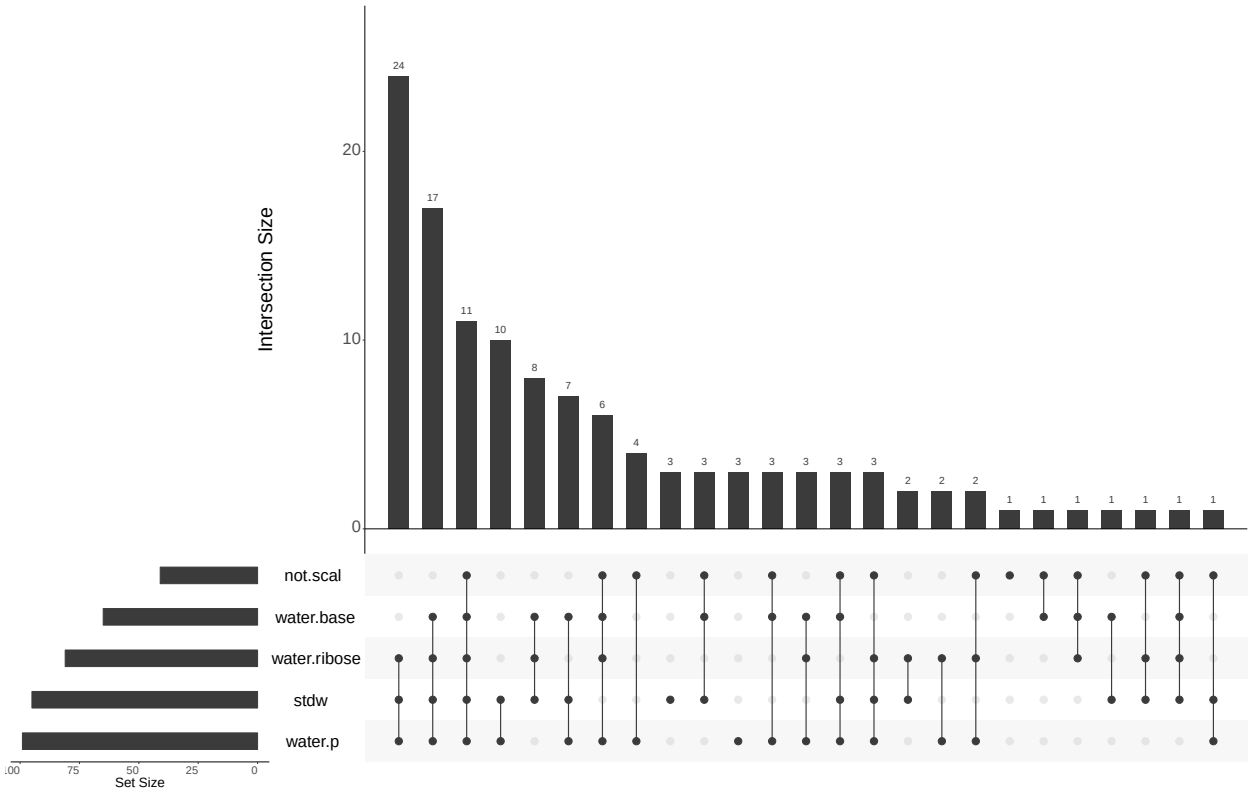
10. Attached Supplementary Data 10 (Data-S10.csv): raw data corresponding to the variations of the binding site's volume for each protein of the benchmark in three conditions: experimental, SCAL, and STDW models.



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.

Graphical TOC Entry