HIPPO: HIstogram-based Pseudo-POtential for scoring protein-ssRNA fragment-based docking poses

Anna Kravchenko, Sjoerd Jacob de Vries, Malika Smaïl-Tabbone, Isaure Chauvot de Beauchene

HIPPO: HIstogram-based Pseudo-POtential for scoring protein-ssRNA fragment-based docking poses

Anna Kravchenko
Université de Lorraine, CNRS, Inria, LORIA
Sjoerd Jacob De Vries
Université de Lorraine, CNRS, Inria, LORIA
Malika Smaïl-Tabbone
Université de Lorraine, CNRS, Inria, LORIA
Isaure Chauvot de Beauchene (isaure.chauvot-de-beauchene@loria.fr)
Université de Lorraine, CNRS, Inria, LORIA

Research Article

Keywords: scoring function, protein-ssRNA docking, RRM-ssRNA docking, fragment-based docking

Posted Date: May 30th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2981840/v1

Abstract

Motivation

: The RNA-Recognition motif (RRM) is a protein domain that binds single-stranded RNA (ssRNA) and is present in as much as 2% of the human genome. Despite this important role in biology, RRM-ssRNA interactions are very challenging to study on the structural level because of the remarkable flexibility of ssRNA. In the absence of atomic-level experimental data, the only method able to predict the 3D structure of protein-ssRNA complexes with any degree of accuracy is ssRNA'TTRACT, an ssRNA fragment-based docking approach using ATTRACT. However, this approach has limitations, such as the production of only a handful of near-native poses amid many non-natives, and the frequent failure of the ATTRACT scoring function (ASF) to recognize these near-natives. Nevertheless, since ASF parameters are not ssRNA-specific and were determined in 2010, there is substantial opportunity for enhancement.

Results

Here we present HIPPO, a composite RRM-ssRNA scoring potential derived analytically from contact frequencies in near-native versus non-native docking models. Validated on a fragment-based docking benchmark of 57 experimentally solved RRM-ssRNA complexes, HIPPO achieved a 3-fold or higher enrichment for half of the fragments, versus only a quarter with ASF. In particular, HIPPO drastically improved the chance of very high enrichment (12-fold or higher), a scenario where the incremental modelling of entire ssRNA chains from fragments becomes viable. However, for the latter result, more research is needed to make it directly practically applicable. Regardless, our approach already improves upon the state of the art in RRM-ssRNA modelling and is in principle extendable to other types of protein-nucleic acid interactions.

1 Introduction

Protein-RNA complexes play an immensely important role in many cellular processes, including translation, transcription, and post-transcriptional gene expression [1]. The disruption of the binding can lead to tremendous cellular malfunctions [2]. A large part of these protein-RNA interactions involves one of the few conserved RNA-binding domains. In particular, over 50% of all RNA-binding proteins in humans contain an RNA recognition motif (RRM) [3]. This motif is critical for binding to RNA molecules, and to single-stranded RNAs (ssRNA) specifically, making RRM-ssRNA interactions crucial for understanding the underlying mechanisms of various cellular processes.

Although the 3D structure of these complexes provides valuable insights into their functions, the experimental resolution of such structures is a non-trivial task. Computational modelling of the 3D structure of a protein-RNA complex, also known as protein-RNA docking, can facilitate experimental research, by proposing probable 3D structures to be experimentally tested.

Unfortunately, protein-ssRNA docking is a challenging task by itself as well. The classical docking approaches [4] require an unbound structure as a starting point, but no such structure is available for ssRNA due to its disorder in the unbound state. On the one hand, one may try to model all possible ssRNA conformations using its sequence, and then dock them. However, ssRNA's flexibility (~ 8 DOF per nucleotide [5]) makes systematic modelling of ssRNA conformations extremely demanding computationally and borderline impossible for long chains. On the other hand, in recent years, various powerful deep learning techniques ([6, 7, 8]) brought breakthroughs to protein-protein [9] and protein-ligand [10, 11] docking. However, deep learning approaches are more challenging to apply to protein-RNA docking, not only due to the relatively low number of solved structures (about $1.16 \cdot 10^4$ protein-RNA structures compared to about $1.776 \cdot 10^5$ protein chains) but also because among all atomic contacts within each structure, the interaction between RNA and protein represents only a tiny fraction. This is even more true for ssRNA, which is only a small subset of RNA, and whose binding modes to proteins have some particularities compared to double-stranded (ds) RNA [12].

Fragment-based docking handles ssRNA flexibility by subdividing its sequence into fragments that are small enough for their conformations to be exhaustively (including close-to-bound conformation) sampled within a given accuracy threshold. The docking procedure consists of sampling and scoring. Sampling refers to the generation of docking *poses* - certain positions and orientations of particular conformations of the fragment with respect to the protein. A pool of docking poses is sampled for each fragment independently. Scoring is the evaluation of the probability of each pose being a near-native, followed by ranking. Finally, the presumably best poses of adjacent fragments are assembled into complete structures called docking *models*. In a test case, when the native structure of a complex is experimentally determined, both docking poses and models can be assessed based on their similarity to the corresponding parts of a native structure, and this similarity can be quantified by their ligand root mean squared deviation (LRMSD). The distinction is made between near-native (correct), non-native (incorrect), and intermediate poses/models based on LRMSD thresholds.

The main limitation of the fragment-based strategy stems from the concept of hot- [13] and coldspot binding. A fragment by itself (taken in isolation) may have much stronger binding and hence lower real interaction energy in a region of the protein that is different from the binding region of that fragment when it is in the chain. This is a case of coldspot binding. The term "coldspot" refers to an area of the protein surface that can bind fragments relatively weakly. The opposite term, "hotspot", refers to the part of the protein surface that binds fragments relatively strongly. Essentially, fragments that bind to the coldspots are only there because the adjacent fragments are tightly bound to the hotspots. From an energy perspective, binding to the coldspot leads to a shallow local energy minimum, whereas binding to the hotspot leads to a deeper (and possibly global) energy minimum. A mononucleotide tandem repeat sequence, such as the poly-U chain, provides a very intuitive example. For such an ssRNA, there are multiple overlapping native solutions for the same fragment sequence UUU that "compete" to be sampled and scored during the docking of UUU. As a consequence, there are usually one or two well-docked

fragments, i.e. fragments with a lot of correctly ranked near-native poses, while the docking results for the remaining fragments are much worse.

The described hot/coldspot limitation directly contributes to the so-called sampling problem. The sampling problem lies in the fact that often not a single near-native pose is generated during the docking run. The sampling problem is critical because it has a high impact on the whole docking procedure: for successful docking of the whole RNA chain, at least one near-native pose must be sampled for each of the fragments. Otherwise, the docking for a given complex will certainly fail at the assembly step.

Another limitation is the scoring problem, which arises when none of the sampled near-natives is selected in the list of top-ranked poses. In this case, more poses per fragment must be retained to have a good chance to keep a near-native, which quickly becomes very expensive computationally in the assembly step. In turn, as there are more docking models, identification of the near-native model also becomes more challenging.

There are four existing fragment-based approaches for protein-ssRNA docking: RNA-LIM, FBDRNA, RNP-denovo, and ssRNA'TTRACT. RNA-LIM represents each nucleotide by one non-oriented bead and could only predict their position at 15Å resolution for one example [14]. FBDRNA uses mononucleotide fragments in all-atom representation, docked with MCSS on a pre-defined binding site. While showing discriminative power on nucleotides' positions, it could not provide accurate models for full oligonucleotides [15]. RNP-denovo, a Rosetta method to simultaneously fold-and-dock RNA to a protein surface, uses the exact position of a few nucleotides [16], which would be unavailable for real-life docking cases. On the other hand, **ssRNA'TTRACT**, the state of the art, is the most accurate approach that uses only a protein structure and the RNA sequence as input. It uses trinucleotides as RNA fragments and an overlapping criterion based on LRMSD for assembly. Furthermore, when information about conserved protein-RNA contacts are available, ssRNA'TTRACT employs an anchored docking strategy to build the RNA chain incrementally by docking one fragment with contact restraints and using each of its top-ranked poses as an anchor to superimpose subsequent fragments [17]. This strategy tackles the sampling problem for the fragments.

ssRNA'TTRACT uses the **ATTRACT** docking engine and a library of RNA trinucleotide conformations developed in our research group [18, 19]. A coarse-grained force field with Lennard-Jones type energy function with soft potential [20] is used for both sampling and scoring. In the coarse-grained representation, the RNA fragments and the protein are represented as sets of pseudo-atoms, called *beads*, each of which stands for a small group of real atoms. Coarse-grained representation provides several advantages compared to all-atom representations. First, it accounts for inaccuracies in atomic positions coming either from bound/unbound conformational differences or experimental biases and resolution; second, it smoothes the energy landscape, which prevents the poses from getting stuck in shallow local minima; and third, it reduces the computation time.

Despite its capabilities, ssRNA'TTRACT is still constrained by the aforementioned limitations. As the current ATTRACT protein-RNA scoring function was not designed to tackle ssRNAs specifically and its

parameters were optimised back in 2010 on dsRNA alone, there is considerable potential for enhancement. Here we present HIstogram-based Pseudo-POtential (HIPPO), which aims to distinguish between near-native and non-native protein-ssRNA docking poses. HIPPO is based on the hypothesis that there exists a collection of scoring parameter sets (as opposed to a single parameter set) that can be used to effectively rank near-native protein-ssRNA docking solutions. HIPPO's parameters are derived analytically from contact frequencies in near-native versus non-native docking poses. These contact frequencies, derived from 4 different sets of docking poses, are discretised by a particular set of cutoffs into histograms, leading to a collection of 4 histogram sets $\mathcal{H}$ that together form the HIPPO scoring potential. Thus, HIPPO is a composite protein-ssRNA scoring potential: typically, the top 5% of the poses according to each histogram set are combined, selecting 20% of all docking poses in total. To streamline the process from dataset construction to the generation of final scoring parameters, we decided to focus exclusively on the RRMs, as this domain of the protein is particularly important for studying protein-ssRNA interactions and is present in many (approximately 65%) of the available protein-ssRNA structures. This allows us to provide proof of principle that the scoring function can indeed be improved using our method. However, the developed method and protocol can be applied to a wider benchmark, and more importantly, to other types of protein-nucleic acid interactions in the future.

HIPPO was derived from a fragment-based docking benchmark of 57 experimentally solved RRM-ssRNA complexes, corresponding to 217 overlapping ssRNA trinucleotide fragments in complexes with an RRM. Using cross-validation, HIPPO achieved a 3-fold enrichment (60% of all near-native poses in the 20% top-ranked poses) for 53% of the fragments, versus only 26% with the current state-of-the-art ATTRACT scoring function (ASF). In addition, these near-native poses were often selected mostly by a single $\mathcal{H}$ of the 4 histogram sets. Consequently, using the hypothetical knowledge of the best HIPPO histogram yielded a 12-fold enrichment for nearly 40% of the test fragments - something which is achieved with ASF in only 4% of the cases. Most importantly, 61% of the complexes show such a 12-fold enrichment for at least one fragment. Under these conditions, the incremental modelling of entire ssRNA chains from best-docked fragments becomes viable. However, the problems of blindly identifying the best HIPPO histogram set and selecting the best-docked fragments need to be solved first before this can become practical. Nevertheless, as it is, HIPPO already improves upon the state of the art in RRM-ssRNA modelling.

2 System and methods

Here we first present the dataset that we built and used for the training and validation of HIPPO. Next, we present step-by-step the process of constructing a set of scoring parameters in the form of a histogram set $\mathcal{H}$ and the process of building the final collection of several $\mathcal{H}$ (Fig. 1).

2.1 Data

2.1.1 RRM-ssRNA benchmark

The number of experimentally solved protein-ssRNA structures is considerably low compared to protein-protein structures. We gathered all available data and build an up-to-date benchmark of experimental 3D structures of RRM-ssRNA complexes from the Protein Data Bank (PDB) by (i) downloading all experimentally solved (either NMR or X-RAY with resolution 3Å or higher) protein-RNA complexes and (ii) applying ProtNAff in order to retrieve complexes with 3 or more consecutive protein-bound single-stranded nucleotides.

We considered a nucleotide to be protein-bound if at least 5 pairs of RRM-RNA heavy atoms were located within 6Å from each other. Lastly, we filtered out complexes whose protein does not contain any RRM domain, according to the InteR3M database [21]. The resulting benchmark consists of 81 RRM-ssRNA complexes, released before February 2021.

2.1.2 Dataset of docking poses

From the benchmark, we created a dataset of labelled docking poses. We used the ATTRACT docking engine and library of RNA trinucleotide conformations [22] to dock each entry (each RRM-ssRNA complex) of the benchmark, by docking each overlapping trinucleotide fragment (e.g. chain AUCG = > fragment AUC and fragment UCG), following the procedure described in [23]. For each fragment, a randomly selected conformation from ProtNAff was placed at each of $3 \cdot 10^7$ predefined starting points located within 30Å from the center of mass of the bound and rigid protein, with a random 3D rotation. Then the position of each starting pose was minimised using gradient descent. Redundant poses (RMSD < 0.2Å) were filtered out of the resulting pool before scoring. The remaining docking poses were scored, and the 10^7 top-ranked poses were retained. Each pose was labelled as near-native if its LRMSD was under 5Å; as non-native if its LRMSD was over 7Å; as intermediate otherwise.

We used such relatively soft thresholds to lower the number of cases for which the sampling problem (zero near-native poses sampled) has arisen. For example, the more strict thresholds [3Å;5Å] resulted in 41% of cases with the sampling problem, versus just 8% with [5Å; 7Å]. To minimise the noise in the dataset, 60 cases where the number of sampled near-natives was less than 100 were excluded. This led to a set of 419 RRM-trinucleotide fragment docking cases. Note that in the case of multiple fragments with the same sequence bound to the same RRM, only a single docking is necessary.

2.1.3. Coarse-grained representation

As mentioned before, in the coarse-grained representation, groups of atoms are represented by beads. In the used representation, 31 bead types are used to represent proteins (2 for backbone and 0–2 for side chain) and 17 bead types are used to represent RNA (1 for phosphate group, 2 for sugar and 3–4 for base), leading to a maximum of 527 pairs of bead types [20]. Protein beads are denoted by index i and RNA beads are denoted by index j.

2.1.4 Redundancy

In order to eliminate possible dataset bias, we performed a redundancy check at the contact level, by comparing i-bead to j-bead distances within 6Å in the native poses of the protein-fragment cases. If such

distance sets were very similar for two cases, these cases were considered redundant, and one of them was removed from the dataset. The final dataset consists of **217 RRM-fragment cases**, with 10^7 labelled docking poses per case. Its corresponding benchmark consists of **57 RRM-ssRNA complexes** and can be found in Additional file 1: Table S1.

2.1.5 Training and test sets

We separated the dataset into pairs of training and test sets based on protein sequence similarity, in a leave-homology-out procedure. Our sequence similarity threshold was 40%. We selected a random protein-ssRNA complex from the benchmark along with all other complexes whose protein sequence similarity was greater than 40%. All data cases derived from these complexes (protein-fragment structures along with their docking poses) became the test set. The remaining data cases formed the corresponding training set. We repeated this procedure iteratively until each of the benchmark complexes was in one of the test sets. To prohibit repetitive and near-repetitive (training; test) pairs, we ensured that the first randomly selected case in each iteration did not belong to any of the previous test sets. All statistics reported in this paper correspond to the evaluation of HIPPO on the test sets, where for each test set the four histogram sets $\mathcal{H}$ derived from the corresponding training set were used. The final collection consists of 29 (training; test) pairs and can be found in Additional file 1: Table S2.

2.2 Creation of histogram set $\mathcal{H}$

The main steps - detailed thereafter - to obtain a scoring histogram set $\mathcal{H}$ are as follows:

1) construction of the *distance arrays* containing the number of occurrences of each bead-bead distance, in near-native vs in non-native poses (ignoring intermediate ones), for each pair of bead types $(i;j)$ independently;
2) refinement of the distance arrays to ensure that each of them provides sufficient signal;

3) derivation of from the distance arrays, one histogram per distance array.

3) derivation of $\mathcal{H}$ from the distance arrays, one histogram per distance array.

2.2.1 Histogram definition

Let's denote the bead types representing the protein by index $i \in \{1,2,\ldots 31\}$, and the bead types representing the RNA by index $i \in \{1,\ldots 17\}$. Also let's define initial distance ranges by applying discretisations of 0.25Å and 1.5Å to the intervals [2Å; 7Å] and [7Å; 14.5Å] respectively. Such design of distance ranges allows to capture close-range interactions with high precision and to generalise long-range interactions. The resulting set contains 27 ranges: {(0, 2], (2, 2.25], …,(14.5, 999)}.

A distance array D_{ij} with the dimension 27x2 is designed to capture the number of occurrences of all $(i;j)$ distances within a pool of docking poses. The rows $d_k, k = 1\ldots 27$, of D_{ij} correspond to the distance ranges. Each element of D_{ij} contains the count of distances within the indicated range.

Elements d_{k1} in the first column account for the distances in near-native poses only, while elements d_{k2} from the second column capture distances in non-native poses.

To ensure that in each D_{ij} there are enough examples coming from near-native poses in each distance range to provide a sufficient signal, we set a threshold w for a minimum number of occurrences in near-natives d_{k1}. The threshold value is empirical and is determined individually for each $(i;j)$ pair as 1/60 of all distances counted in near-native poses:

$$w_{ij} = A_{ij}/60,$$

$$where A_{ij} = \sum_{k} d_{k1}, \forall d_{k1} \in D_{ij},$$

For each D_{ij}, if $d_{k1} < w_{ij}$, then the rows starting from k^{th} and beneath are summed until their sum exceeds the threshold. The new row resulting from the summation replaces the original row. This process is repeated until all values in the first column of the resulting array exceed the threshold. The resulting *refined distance array* D^*_{ij} has dimension qx2, where $q \leq 27$, and may vary for different $(i;j)$ pairs. Note that for each $(i;j)$ we must save the resulting set of refined distance ranges for further application of the histogram.

Finally, the following formula, inspired by the logarithm of the odds ratio, is used to obtain individual histograms H_{ij} from the corresponding D^*_{ij}:

$$H_{ij} = \left[\ln d^*_{x1} - \ln d^*_{x2} - (\ln A_{ij} - \ln B_{ij})\right]$$

,

$$where x = 1 \dots q, \forall x \, [d^*_{x1}, d^*_{x2}] \in D^*_{ij}, B_{ij} = \sum_{k} d_{k2}, \forall d_{k2} \in D_{ij}.$$

The dimension of H_{ij} is qx1. We define $\mathcal{H}$ as the set of individual histograms H_{ij} for all $(i;j)$ pairs, which are present in at least one pose out of the input pool of the docking poses.

Since 10^7 poses is a rather large pool, poses with vastly different ranks could possess different features. To account for this possibility, we divided the initial pool of poses into 3 sub-pools according to the rank of the poses: [0, 99999], [10^5, 999999], [10^6,10^7]. Each D_{ij} and subsequently each H_{ij} consists of three parts, built on poses from the corresponding rank-based sub-pool.

2.2.2. Scoring with $\mathcal{H}$ and scoring assessment

To score a pose using $\mathcal{H}$, we count the occurrences of distances for each $(i;j)$ pair within each of the refined ranges, within each rank-based sub-pool. This information is stored in a qx1 array R_{ij}. The histogram-based score of a pose is calculated using the following formula:

$$S_{pose} = \sum_i \sum_j R_{ij} \bullet H_{ij}^T,$$

1

In simpler terms, for every bead-bead distance in a pose that falls in one of the refined ranges, a corresponding sub-score is assigned. This process is repeated for each rank-based sub-pool separately. The sum of all sub-scores is the final histogram-based score of a pose.

To evaluate the performance of $\mathcal{H}$ for a data case, we score all docking poses from the pool of 10^7 poses using formula (1) and rank the poses by their score in a descending order. Then we select the 5% of top-ranked poses and calculate the fraction of all near-native poses that are present in this selection. An $\mathcal{H}$ is labelled as successful for a given data case if this value exceeds 60%. Likewise, we can say that a given case is successfully scored by current $\mathcal{H}$.

2.3 Collection of$\mathcal{H}$

Initial analysis revealed that a single $\mathcal{H}$ was not sufficient to account for the diverse protein-ssRNA binding modes (Fig. 2). Therefore, we opted for the creation of a small collection of $\mathcal{H}$, where each $\mathcal{H}$ is successful on a subset of the cases. When applied simultaneously, the collection should cover the majority of cases, except for a few outliers. The collection is created by selecting several best-performing $\mathcal{H}$, such that maximising the number of successfully scored cases in the training set. The full procedure is detailed in the next section (2.3.1).

Because in a real-life docking case, there will be no indication of which $\mathcal{H}$ from the collection is best suited for scoring, the case must be scored by all $\mathcal{H}$ and results must be pooled together (see 2.3.2). As the collection size increases, so does the chance of overfitting. For this reason, we have empirically limited the number of $\mathcal{H}$ to 4 per collection. Increasing this number to 5 or 6 had only limited influence (result not shown).

2.3.1 Partitioning algorithm

While deriving a collection of 4 $\mathcal{H}$ - $\mathcal{H}_1$, $\mathcal{H}_2$, $\mathcal{H}_3$ and $\mathcal{H}_4$ - we partition the training cases into four subsets, plus a subset of outliers. This procedure is implemented as follows:

1. Derive $\mathcal{H}$ for each case individually;
2. Score each case with each $\mathcal{H}$;
3. For each pair (case; $\mathcal{H}$), calculate the percentage of the near-natives that end up in the 5% of top-ranked poses. If the calculated value is over 60%, then label this case as successfully scored by the given $\mathcal{H}$;
4. Select the four $\mathcal{H}$ that maximise the total number of successfully scored cases. This is the resulting collection.

Now, each training case either is associated with its best-performing $\mathcal{H}$ in the resulting collection or ends up in the set of outliers.

2.3.2. Scoring with collection and evaluation strategy

To score a case with a collection, we score its docking poses with $\mathcal{H}_1$, $\mathcal{H}_2$, $\mathcal{H}_3$ and $\mathcal{H}_4$ separately using (1). Then, for each $\mathcal{H}$, around 5% of its top-ranked poses are selected and pooled together in TopC (where “C” stands for a collection). If the same pose is present in several scorings, only its highest rank is kept. The size of the TopC should be equal to 20% of all sampled poses. The resulting set of poses TopC is expected to contain the best ones (the poses outside of TopC are dismissed).

To evaluate the performance of the collection for a case, the fraction of all near-native poses that end up in TopC is calculated. If this value exceeds 60%, then the collection is successful for a given data case.

3 Results

In this study, we developed a new protocol for deriving scoring parameters for molecular docking poses, based on distances between RNA and protein beads, in the form of a collection of 4 histogram sets $\mathcal{H}$. We applied it to create HIPPO, a novel scoring function specifically for RRM-ssRNA fragment-based docking. To achieve this goal, we split every available RRM-ssRNA structure into RRM-fragment cases (fragments of 3 consecutive bound nucleotides), for each of which 10^7 docking poses were generated using the ATTRACT docking engine. Our initial benchmark consisted of 479 fragments from 81 complexes. Out of these, 262 fragments were unusable for training because of a sampling problem (less than 100 near-native poses sampled) or because of redundancy between fragments on the contact level (6Å), resulting in a dataset of 217 well-sampled non-redundant cases, coming from 57 RRM-ssRNA complexes. Within the resulting dataset, the average number of sampled near-native poses is 9112 and the median is 3145. To assess how HIPPO performance would generalise to new data cases, we used the leave-homology-out cross-validation strategy: 29 pairs of training and test sets were formed based on RRM sequence similarity. The size of the test set depended on the number of cases derived from each RRM-ssRNA complex of a given RRM and varied from 1 to 33 cases per set.

For a given pair of test and training sets, for each case in the training set, we derived an $\mathcal{H}$ by analysing the frequencies of bead-bead distances in the near-native (LRMSD < 5Å) vs non-native (LRMSD > 7Å) docking poses, and we applied it to each of the other cases in the training set. We selected the collection of 4 $\mathcal{H}$ sets that maximised the number of training cases for which at least one $\mathcal{H}$ ranks 60% of all near-native poses in the 5% top-ranked poses. Then, the collection was applied to the test cases, and the best of the 4 ranks for each pose was retained to obtain the 20% top-ranked poses (TopC). The collection was considered to be successful on a test case if at least 60% of all near-native poses were in TopC.

3.1 General performance

We applied the described protocol to each of the 29 training sets and derived 29 collections of 4 $\mathcal{H}$. We then applied these collections to the cases in the corresponding test sets and compared the percentages of near-natives selected in TopC with HIPPO and in the 20% top-ranked with ASF (Table 1, Fig. 3). Further in the text, we refer to the percentage of near-natives present in TopC or 20% top-ranked as 'selected'. At least 60% of all near-natives selected (a 3-fold enrichment compared to random scoring) for more than half of the RRM-fragment test cases with HIPPO, versus a quarter with ASF (53% vs 26% of the test cases respectively). In one-third of the test cases, we even observed a 4-fold enrichment (80% of near-natives selected) with HIPPO, something which is rarely achieved by ASF (38% vs 7% of the test cases respectively). To ensure that our results were not skewed by cases coming from one or a few largest test sets, we compared the average success rates over the test sets and found 62% and 34% respectively (Fig. 4, a).

Table 1
Comparison of the performance of HIPPO vs ASF on the 217 cases (29 test sets, 57 complexes)

	ASF	HIPPO
% of near-natives in TopC/Top20, averaged over all test cases	43	55
Success rate (%) over all cases	26	53
Average highest % of near-natives on TopC/Top20 among the cases of a complex, over all test cases	60	72
Nb of complexes with the > 80% of near-natives in TopC/Top20 for at least one fragment	9	33
Nb of cases with > 80% of near-natives in TopC/Top20	15	75

3.2.1 Best-scored fragment per complex

We found a positive correlation (Pearson correlation, $r = 0.43$, Fig. 4, b) between the number of protein-fragment contacts under 5Å and the percentage of near-natives in TopC, which complies with the cold/hotspot theory. To perform anchored fragment-based docking, at least one fragment per complex must be well-docked. We thus analysed the distribution of successes among the complexes, with HIPPO and ASF. The number of complexes with at least one successfully scored fragment increased from 54% with ASF to 75% with HIPPO. With the success criterion raised to 80% of the near-natives selected (a 4-fold enrichment), the compared success rate percentages still increased from 16% with ASF to 58% with HIPPO. Moreover, the enrichment for the best-scored fragment per complex was increased with HIPPO compared to ASF in 68% of complexes. On average, for the best-scored fragment of each complex, HIPPO selects an additional 19% of all near-natives compared to ASF.

3.3 Analysis of the collections

To assess the gains of using a collection (4 $\mathcal{H}$) instead of 1 $\mathcal{H}$, we evaluated if the 4 $\mathcal{H}$ bring complementary information, either for each test case (by selecting different near-native poses) or for each test set (by performing well on different test cases).

3.3.1. Complementarity of the 4 $\mathcal{H}$ in a collection

Out of 29 collections, the ones derived from the training sets 1, 2, 3, 4 and 8 are distinct (see Additional file 1: Table S3). The remaining collections are identical to the collection from training set 4. On the test set level, we can see that each single $\mathcal{H}$ is the best-performing (selects the highest number of near-natives) of the collection for 0–48% of the cases. In other words, there is never one $\mathcal{H}$ that is the best suited for half or more of the cases in a given test set. This complies with the hypothesis that several different $\mathcal{H}$ are required to account for different binding modes (Fig. 5, Additional file 1: Table S4), and that a few potentials better represent the diversity of RRM-ssRNA binding modes than one $\mathcal{H}$, by providing at least one well-suited $\mathcal{H}$ per case for most cases.

3.3.2 Best-performing $\mathcal{H}$ per case or per complex

For half of the cases, most of the near-natives in the TopC were selected by a single $\mathcal{H}$ out of 4. If for each test case, we could use its best-performing $\mathcal{H}$ instead of the collection (and count near-natives in 20% top-ranked instead of pooling in the TopC), such modified application of HIPPO would reach a 3-fold enrichment for 77% cases (instead of 53% with the collection and 26% with ASF) and a 4-fold enrichment for 62% cases (instead of 38% with the collection and 7% with ASF) (Supplementary Section 4, Table 4, Fig. 2). Furthermore, selecting only the 5% top-ranked poses would show a 12-fold enrichment for 39% cases (vs 4% cases with ASF). For the best-scored fragment per complex, a 12-fold enrichment was observed in 61% of complexes with HIPPO, while this is almost never achieved with ASF (7% of complexes). These numbers point toward the advantage of applying a single best-performing $\mathcal{H}$ per case rather than a collection if one could predict which $\mathcal{H}$ to apply to which case (Fig. 6).

4 Discussion

Despite the numerous biological roles of ssRNA-protein binding processes, there is still a lack of methods capable of addressing the dual challenges of the very high flexibility of ssRNA and the scarcity of its experimental structures. We previously developed a unique approach capable of modelling protein-bound ssRNA, by coarse-grained docking of ssRNA fragments with the ATTRACT docking software, followed by combinatorial assembly of geometrically compatible poses. This approach is successful in modelling the full ssRNA chain at high accuracy when conserved stacking contacts are known: the docking search space is reduced by constraints forcing the stacking of certain nucleotides on the conserved residues. In the absence of conserved contacts, this approach is limited by the poor sampling and low discriminatory power of the protein-RNA energy function of ATTRACT when applied to ssRNA fragments. With typically a few thousand near-native poses sampled out of 10^7 poses, the percentage of near-natives is less than 0.1%. In general, during assembly, low percentages of near-natives at the fragment level increase the

probability of compatible non-native poses, leading to a prohibitive number of full-chain RNA models with an infinitesimally low percentage of quasi-native models. For direct applicability in the absence of conserved contacts, a very high enrichment is needed, followed by clustering and possibly refinement/rescoring with molecular dynamics, to arrive at an ensemble of perhaps a few hundred poses of which at least one is near-native.

In order to achieve such a high enrichment, we developed a new analytic approach for creating a scoring function for docking poses of coarse-grained ssRNA fragments, based on the frequencies of contact distances in near-native versus non-native poses. A specificity of our approach is to derive and combine a small set of potentials to better cover the diversity of ssRNA binding modes. We applied it to create HIPPO, a novel scoring function specifically for coarse-grained RRM-ssRNA fragment-based docking. On a benchmark of 57 RRM-ssRNA complexes.

HIPPO demonstrates a better discriminatory power for near-native poses than the state-of-the-art ATTRACT scoring function (ASF), making it the best coarse-grained scoring function tested for protein-ssRNA complexes to date.

The successfully and unsuccessfully scored cases are rather evenly distributed among the complexes (result not shown). HIPPO's strengths and weaknesses are thus not likely to be attached to any specific type of complex, but rather to hot- and coldspots binding, meaning RNA fragments of a complex that are tightly and loosely attached to the protein respectively. This variability of docking performance over fragments is a difficulty inherent in a classical fragment-based docking approach, where each fragment must be docked (sampled and scored) within an accuracy threshold before the assembly. A way to tackle this is to ensure that at least one fragment per complex is very well docked and use each of its top-ranked poses as anchors to build a full RNA model by direct poses superposition followed by scoring. In the absence of evidence to identify the well-docked fragment from RNA sequence and protein structure, one would iteratively consider each fragment as such. We had previously applied a similar anchored docking of ssRNA on RRMs by using conserved stacking interactions between RRM aromatic residues and a nucleotide base as anchors [15]. Yet nearly half of RRM structures lack those conserved aromatics [21], and such a new hotspot approach would overcome this limitation. HIPPO will be better suited than ASF for this approach, since (i) more complexes have at least one successfully docked fragment compared to ASF, and (ii) the best-scored fragment in each complex has a higher enrichment for most complexes compared to ASF.

We have seen that for most cases (95%) the best-performing $\mathcal{H}$ of the collection performed better than the whole collection (Fig. 4. c). A way to improve HIPPO’s performance would be to determine which $\mathcal{H}$ from the collection will perform the best on a given protein-fragment case. This would allow us to apply only this one $\mathcal{H}$ and avoid retaining false positives returned by the other three $\mathcal{H}$. This may be achieved with the help of the supervised machine learning techniques based on the sequence of the fragment and the sequence or/and structure of the protein, and/or on the docking poses. Such a pre-trained classifier not only would drastically improve the performance of the scoring but could also give valuable insight

into the most prevalent protein-ssRNA binding modes. More importantly, since scoring with the best performing $\mathcal{H}$ achieved 60% of near-natives in 5% top-ranked for the best-scored fragment in a complex for 61% of complexes, there is a great perspective in clustering these top-ranked poses and using the obtained prototypes as anchors.

We see several tuning possibilities that might yield improved HIPPO performance. In particular, we will try to apply a stricter threshold for near-native poses, and see if, despite the increased sampling difficulties encountered, there would still be enough signal for HIPPO to succeed for high-accuracy poses.

As mentioned earlier, we face not only scoring but also, primarily, a sampling problem in ssRNA docking. HIPPO can be considered as a pseudo-energy function, and as such, it is suitable for a sampling procedure based on energy minimisation that would not require derivability of the energy, such as a Monte Carlo approach [24]. We plan to test it against the current ATTRACT sampling procedure that uses ASF with gradient minimisation. Another possible way to apply HIPPO for the sampling is to convert each histogram into a differentiable function to be used directly in ATTRACT gradient minimisation protocol.

To further evaluate the generalisability of our approach for deriving scoring potentials, we plan to expand our benchmark from only RRM-ssRNA structures to a more general protein-ssRNA benchmark, as well as to our benchmark of protein-ssDNA structures [25].

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The source code of HIPPO along with the final scoring parameter set are available via https://github.com/sjdv1982/histograms. The data set used for the study is available in the Supplementary Materials.

Competing interests

The authors declare that they have no competing interests.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 813239.

Authors' contributions

AK assembled and processed the data (with input from ICB), participated in the design of the work and the creation of the software, and drafted the manuscript (with input from ICB and SJV). SJV contributed substantially to the conception and design of the work and to the creation of the software and revised the manuscript. MST contributed to the design of the work and revised the manuscript. ICB contributed substantially to the conception and design of the work and revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Experiments presented in this paper were carried out using the **Grid'5000** testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr).

References

1. Cléry, A., Blatter, M., & Allain, F. H. (2008). RNA recognition motifs: boring? Not quite. *Current opinion in structural biology*, *18*(3), 290–298
2. Choi, P. S., & Thomas-Tikhonenko, A. (2021). RNA-binding proteins of COSMIC importance in cancer. *The Journal of clinical investigation*, *131*(18), e151627.
3. Tsai, Y. S., Gomez, S. M., & Wang, Z. (2014). Prevalent RNA recognition motif duplication in the human genome. *RNA (New York, N.Y.)*, *20*(5), 702–712.
4. Bheemireddy, S., Sandhya, S., Srinivasan, N., & Sowdhamini, R. (2022). Computational tools to study RNA-protein complexes. *Frontiers in molecular biosciences*, *9*, 954926.
5. Chen S. J. (2008). RNA folding: conformational statistics, folding kinetics, and ion electrostatics. *Annual review of biophysics*, *37*, 197–214.
6. Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, *577*(7792), 706–710.
7. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2
8. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., Opperman, D. J., … Baker, D. (2021). Accurate

prediction of protein structures and interactions using a three-track neural network. *Science (New York, N.Y.)*, *373*(6557), 871–876.

9. Bryant, P., Pozzati, G., & Elofsson, A. (2022). Improved prediction of protein-protein interactions using AlphaTest set2. *Nature communications*, *13*(1), 1265.
10. Yang, C., Chen, E. A., & Zhang, Y. (2022). Protein-Ligand Docking in the Machine-Learning Era. *Molecules (Basel, Switzerland)*, *27*(14), 4568.
11. Meli, R., Morris, G. M., & Biggin, P. C. (2022). Scoring Functions for Protein-Ligand Binding Affinity Prediction using Structure-Based Deep Learning: A Review. *Frontiers in bioinformatics*, *2*, 885983..
12. Pal, A., & Levy, Y. (2019). Structure, stability and specificity of the binding of ssDNA and ssRNA with proteins. *PLoS computational biology*, *15*(4), e1006768.
13. Mei, L. C., Hao, G. F., & Yang, G. F. (2022). Computational methods for predicting hotspots at protein-RNA interfaces. Wiley interdisciplinary reviews. RNA, 13(2), e1675. https://doi.org/10.1002/wrna.1675
14. Hall, D., Li, S., Yamashita, K., Azuma, R., Carver, J. A., & Standley, D. M. (2015). RNA-LIM: a novel procedure for analyzing protein/single-stranded RNA propensity data with concomitant estimation of interface structure. *Analytical biochemistry*, *472*, 52–61.
15. González-Alemán, R., Chevrollier, N., Simoes, M., Montero-Cabrera, L., & Leclerc, F. (2021). MCSS-Based Predictions of Binding Mode and Selectivity of Nucleotide Ligands. *Journal of chemical theory and computation*, *17*(4), 2599–2618.
16. Kappel, K., & Das, R. (2019). Sampling Native-like Structures of RNA-Protein Complexes through Rosetta Test seting and Docking. *Structure (London, England : 1993)*, *27*(1), 140–151.e5.
17. Isaure Chauvot de Beauchene, Sjoerd J. de Vries, Martin Zacharias (2016) Fragment-based modelling of single stranded RNA bound to RNA recognition motif containing proteins, *Nucleic Acids Research*, 44(10) 4565–4580,
18. Moniot, A., Guermeur, Y., de Vries, S. J., & Chauvot de Beauchene, I. (2022). ProtNAff: protein-bound Nucleic Acid filters and fragment libraries. *Bioinformatics (Oxford, England)*, *38*(16), 3911–3917.
19. Moniot, A., Chauvot de Beauchêne, I., Guermeur, Y. (2022). Inferring ε -nets of Finite Sets in a RKHS. In: Faigl, J., Olteanu, M., Drchal, J. (eds) Advances in Self-Organizing Maps, Learning Vector Quantization, Clustering and Data Visualization. WSOM+ 2022. Lecture Notes in Networks and Systems, vol 533. Springer, Cham.
20. Setny, P., & Zacharias, M. (2011). A coarse-grained force field for Protein-RNA docking. *Nucleic acids research*, *39*(21), 9118–9129.
21. InteR3M database https://inter3mdb.loria.fr/. Accessed 4 May 2023.
22. Moniot, A., Guermeur, Y., De Vries, S. J., & Chauvot de Beauchene, I. (2022). ProtNAff: Protein-bound Nucleic Acid filters and fragment libraries [Data set]. *Zenodo*.
23. Chauvot de Beauchene, I., de Vries, S. J., & Zacharias, M. (2016). Binding Site Identification and Flexible Docking of Single Stranded RNA to Proteins Using a Fragment-Based Approach. *PLoS*

computational biology, *12*(1), e1004697.

24. Glashagen, G., de Vries, S., Uciechowska-Kaczmarzyk, U., Samsonov, S. A., Murail, S., Tuffery, P., & Zacharias, M. (2020). Coarse-grained and atomic resolution biomolecular docking with the ATTRACT approach. Proteins, 88(8), 1018–1028.
25. Mias-Lucquin, D., & Chauvot de Beauchene, I. (2022). Conformational variability in proteins bound to single-stranded DNA: A new benchmark for new docking perspectives. *Proteins*, *90*(3), 625–631.

Figures

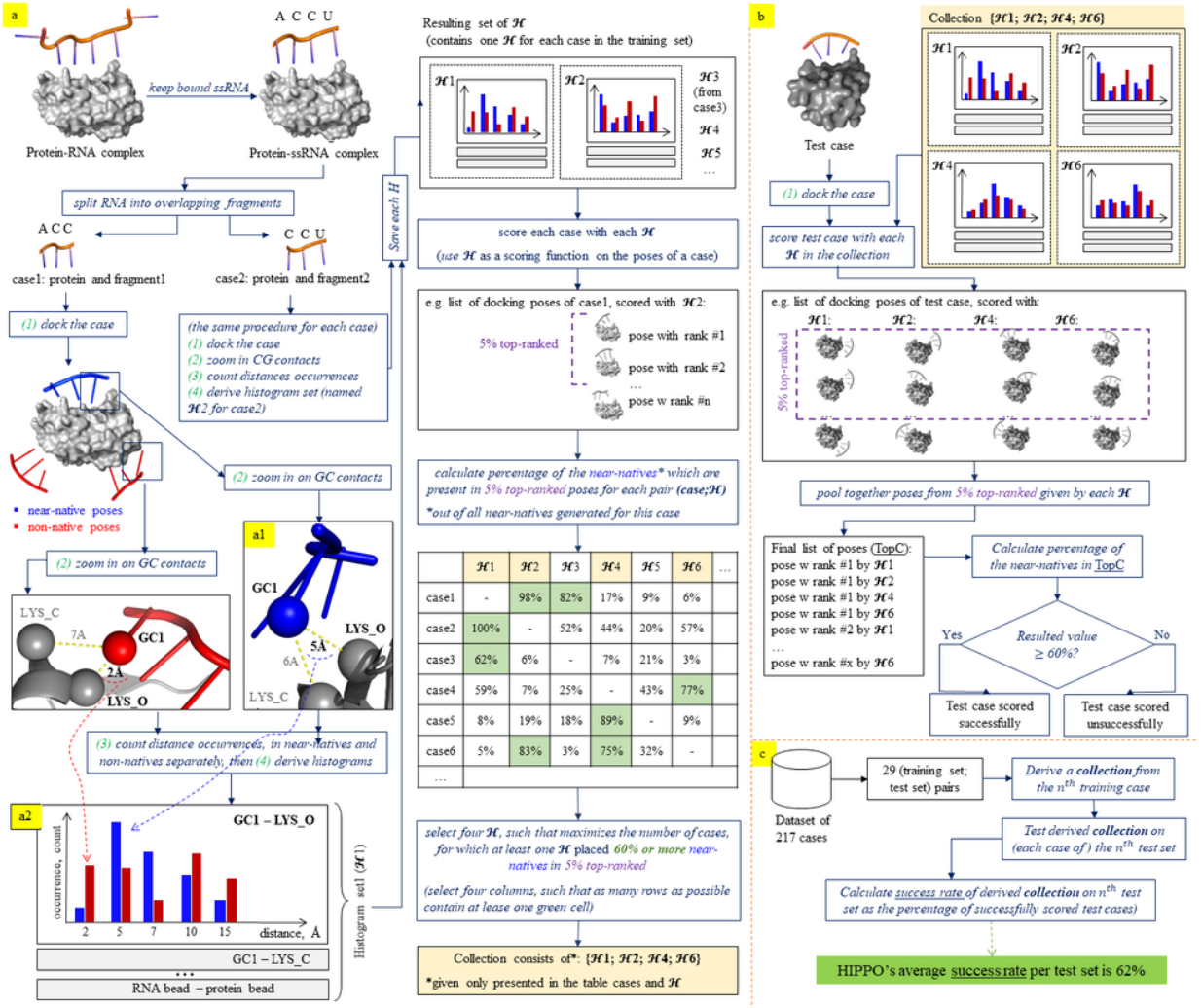

Figure 1

a) Graphical pipeline for building HIPPO as a collection of four histogram sets ($\mathcal{H}$). a1) Contacts between bead GC1 (in Cytosine side-chain) and bead LYS_O and LYS_C (Lysine backbone). a2) An intuitive schema of $\mathcal{H}$. The histogram for beads (GC1; LYS_O) is shown as an expanded plot. The blue dashed lines from a1 to a2 show the contribution of the contact to the histogram. The blue/red bars show the

count of occurrences of distances in all near-native/non-native poses. The other histograms in this set $\mathcal{H}$, for other pairs of beads, are not shown (collapsed). b) Graphical pipeline for testing a collection on a test case. c) Graphical pipeline for the complete workflow. The creation of pairs of training and test sets is based on the protein's sequence similarity: proteins with sequence similarity of 40% or higher are never present in both training and test sets.

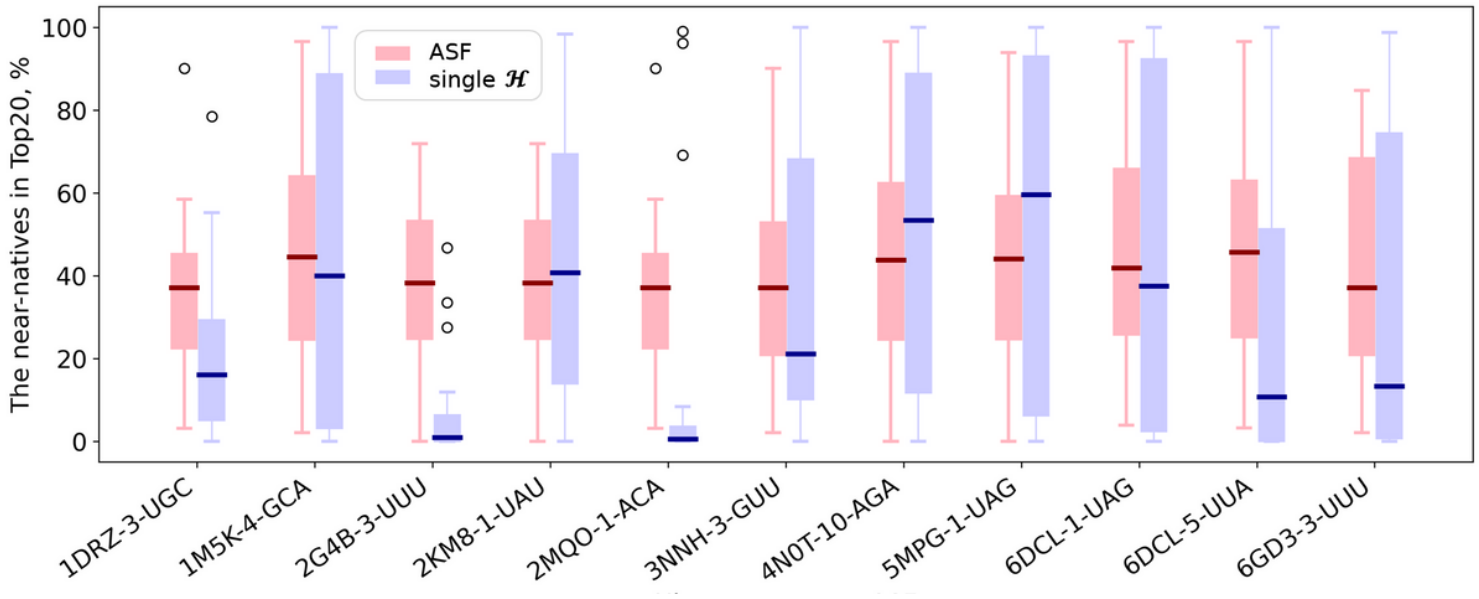

Figure 2

Comparison of the percentage of near-natives selected by a single $\mathcal{H}$ vs ASF. Each pair of adjacent boxes shows the distribution of the results produced by a corresponding $\mathcal{H}$ (purple) and ASF (pink) on the relevant for a given $\mathcal{H}$ test set(s) (sets used for the collection to which given $\mathcal{H}$ belongs), for a range from 0% to 100% of the near-natives in the 20% top-ranked poses.

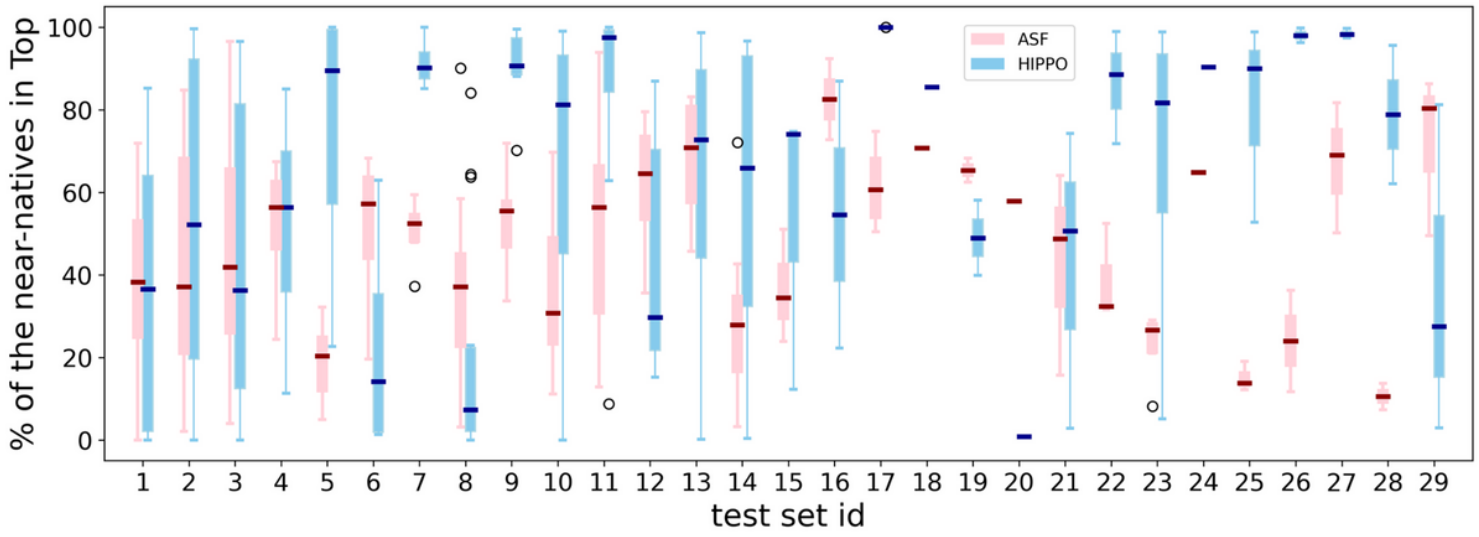

Figure 3

Comparison of the percentage of selected near-natives by collections vs ASF on the test sets. Each pair of adjacent boxes shows the distribution of the results produced by a corresponding collection (blue) and ASF (pink) on one of the 29 test sets, for a range from 0% to 100% of the near-natives in the corresponding Top (TopC/Top20 respectively).

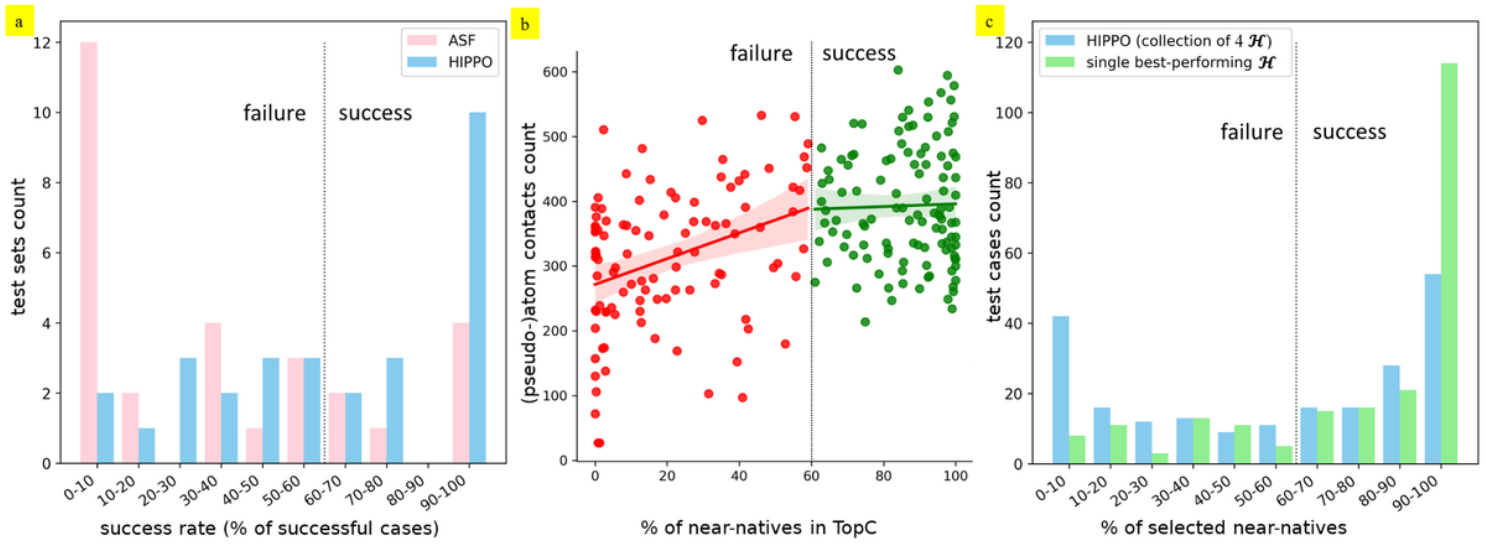

Figure 4

a) Distribution of the success rate per test set, achieved with ASF (pink) and HIPPO (blue). The black dotted line indicates the threshold of a 3-fold enrichment compared to random sampling. b) Relation between the number of contacts in a protein-fragment structure vs the percentage of near-natives in TopC achieved by HIPPO. c) Distribution per test case of the percentage of near-natives selected by a collection of 4 $\mathcal{H}$ (blue) versus by a single best-performing $\mathcal{H}$ (green).

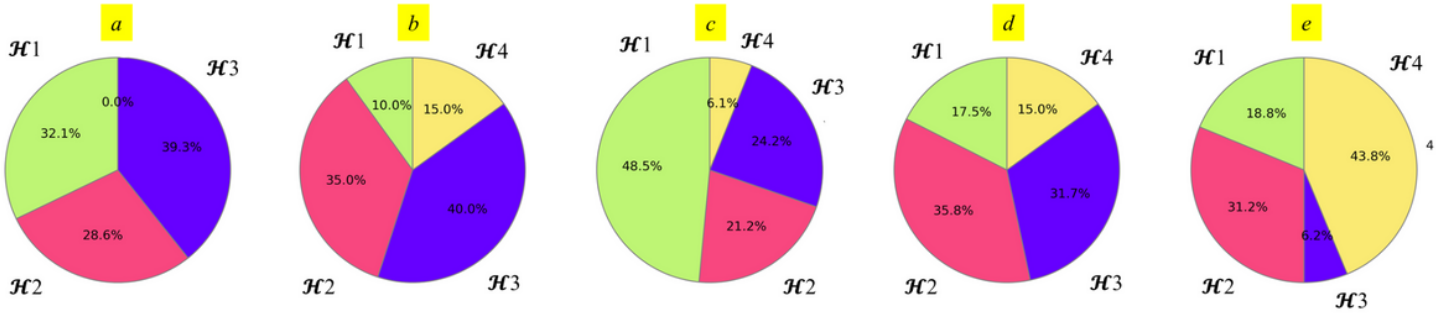

Figure 5

The percentage of cases within a test set, for which each of the 4 $\mathcal{H}$ in the collection is the best-performing one. a) For collection 1 on test set 1. b) For collection 2 on test set 2. c) For collection 3 on test set 3. d) For collection 4 on the united test set, suitable for validation of this collection's performance. This set consists of the test cases belonging to all test sets, excluding sets 1, 2, 3 and 8. e) For the collection 8 on test set 8.

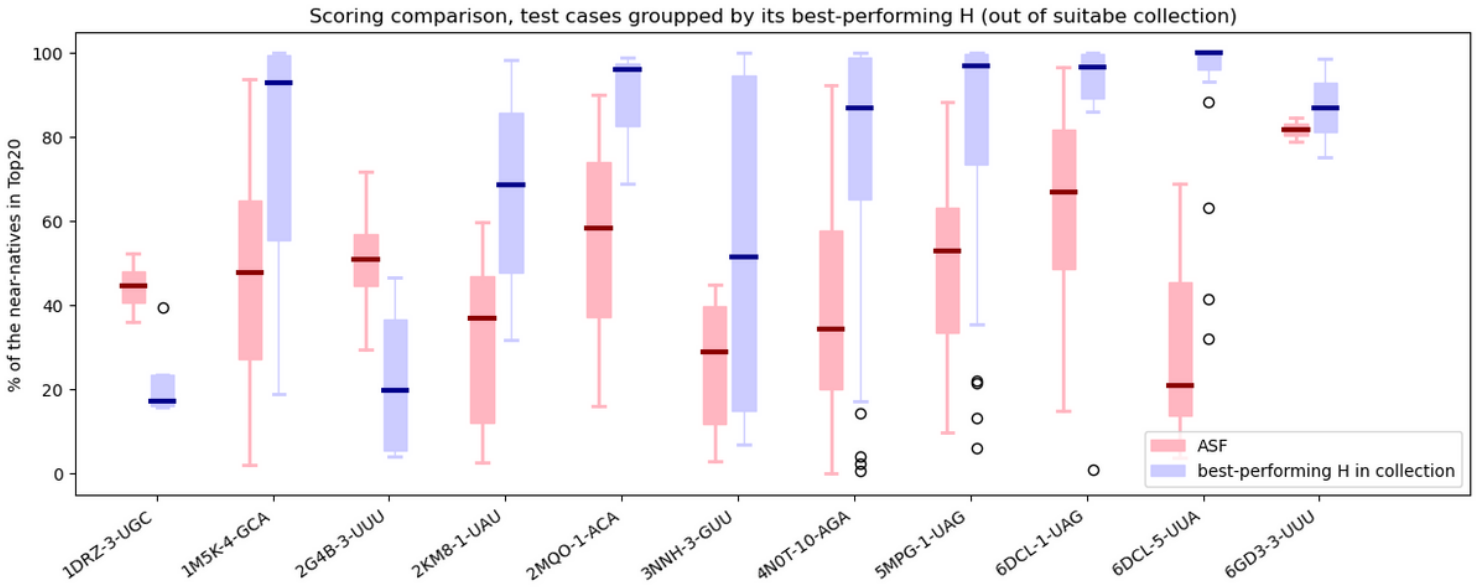

Figure 6

Comparison of the percentage of selected near-natives by ASF vs the best-performing $\mathcal{H}$. Each pair of adjacent boxes shows the distribution of the results produced by each best-performing $\mathcal{H}$ (purple) or ASF (pink) on the relative test cases for a range from 0% to 100% of all near-natives ranked in the 20% top-ranked poses.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx