



**HAL**  
open science

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

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

► **To cite this version:**

Anna Kravchenko, Sjoerd Jacob de Vries, Malika Smaïl-Tabbone, Isaure Chauvot de Beauchene.  
HIPPO: HIstogram-based Pseudo-POtential for scoring ssRNA-protein fragment-based docking poses.  
The 31st Annual Intelligent Systems For Molecular Biology (ISMB) and the 22nd Annual European  
Conference on Computational Biology (ECCB), Jul 2023, Lyon, France. hal-04168414

**HAL Id: hal-04168414**

**<https://hal.science/hal-04168414>**

Submitted on 21 Jul 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# HIPPO: Histogram-based Pseudo-Potential for scoring ssRNA-protein fragment-based docking poses

Anna Kravchenko<sup>1</sup>, Sjoerd Jacob de Vries<sup>1</sup>, Malika Smail-Tabbone<sup>1</sup> and Isaure Chauvot de Beauchene<sup>1</sup>  
<sup>1</sup> Université de Lorraine, CNRS, Inria, LORIA F-54000 Nancy, France

## 1 MOTIVATION

Single-stranded (ss) RNA-protein complexes are very challenging to model:

1. The *absence of an unbound ssRNA structure* due to its disorder prevents the use of classical docking methods;
2. *ssRNA flexibility* does not allow for exhaustive conformational sampling for long chains;
3. The relatively *small number of experimental ssRNA-protein structures* prevents the use of end-to-end deep learning for this problem.

**Fragment-based docking** approach using ATTRACT coarse-grained model [1] tackles this flexibility issue by splitting the ssRNA sequence into fragments small enough to allow their conformations to be exhaustively sampled within a given accuracy threshold, including near-bound conformation. However, this approach suffers from the **scoring problem**: the frequent inability of the ATTRACT scoring function [2] (ASF) to recognize correct poses among many incorrect poses. Since ASF parameters are not ssRNA-specific and were determined in 2010, we see a substantial opportunity for improvement.

## 2 APPROACH

We built a new scoring parameters set, based on the **ratio of bead-bead<sup>1</sup> distances in correct vs incorrect poses**, with the following procedure (see details in 6 PIPELINE):

1. Dock data cases (protein<sup>2</sup> and RNA fragment) with known 3D structure to generate a pool of docking poses;
2. Distinguish correct poses (LRMSD<5Å) from incorrect poses (LRMSD>7Å) based on the known 3D structure;
3. Count occurrences of each <protein bead type; RNA bead type> contact in the pool of correct and incorrect poses;
4. Derive a set of parameters as the occurrences ratio, in a form of a histogram set  $\mathcal{H}$  (one bar per distance range) for each <bead type; bead type> pair.

The final  $\mathcal{H}$  gives, for each <bead type; bead type> pair, for each bead-bead distance range, **the propensity of this distance to be seen in correct rather than incorrect pose**. In a real docking case, each pose is scored by summing, over all its RNA-protein contacts, the propensity of that distance to be correct for that pair of bead types.

As 1 set of histograms is not enough to cover the variety of ssRNA binding modes, we derived a **collection of 4  $\mathcal{H}$** . We apply each set independently to select a fraction of best-scored poses, then pool the 4 fractions together.

<sup>1</sup> groups of atoms in coarse-grained representation [2]

<sup>2</sup> to streamline the process, we focused on the RNA recognition motifs (RRMs) as this domain is especially relevant for ssRNAs and present in ~65% of ssRNA-binding proteins

## 3 RESULTS

Here we present HIPPO, a composite coarse-grained ssRNA-protein scoring potential derived analytically from contact frequencies in correct versus incorrect docking poses. We validated (see 6 PIPELINE, b) it on a benchmark of 57 experimentally solved RRM-ssRNA complexes, which consists of 217 non-redundant RRM-fragment<sup>1</sup> cases. HIPPO achieved a **3-fold or higher enrichment<sup>2</sup> in the 20% best-scored poses for half of the fragments**, versus only a quarter with ASF.

For most fragments, one  $\mathcal{H}$  out of the collection of 4 picks up most of the correct selected poses. Thus, this enrichment would be greatly improved if one could predict which  $\mathcal{H}$  to select for a given fragment. In particular, HIPPO would drastically improve the chance of a very high enrichment (12-fold or higher) of the best-scored fragment in the complex, a scenario where the incremental modelling of the entire ssRNA chain from one fragment becomes viable [3]. 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 ssRNA-protein modelling and is **extendable to other types of protein-nucleic acid interactions**.

<sup>1</sup> a fragment is a trinucleotide

<sup>2</sup> the list of 20% top-ranked poses contains 60% or more of all sampled NNs

## 4 PERFORMANCE

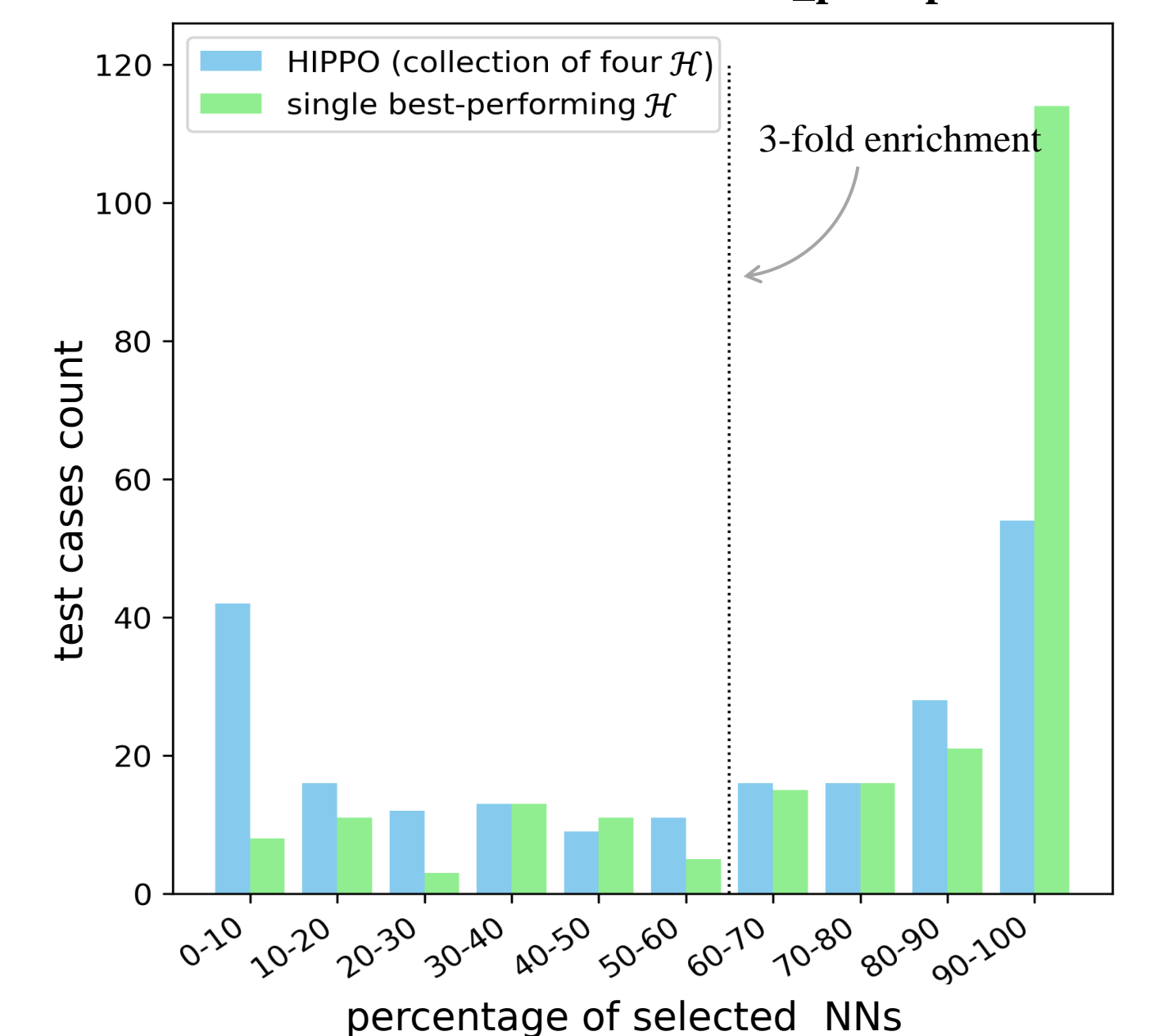
HIPPO application entails independent scoring of all docking poses with each of four  $\mathcal{H}$ , followed by pooling together the top 5% of each scoring list.

	ASF	HIPPO
% of cases with 60% of more correct poses selected	26	53
% of cases with 80% of more correct poses selected	7	38
% of correct poses selected, averaged over all test cases	43	55
Nb of cases with 80% or more correct poses selected	15	75
Highest % of correct poses selected among the cases of a complex, averaged over all test cases	60	72
Nb of complexes with 60% or more correct poses selected for at least one fragment	54	75
Nb of complexes with 80% or more correct poses selected for at least one fragment	9	33
Nb of complexes with higher % of correct poses selected for the best-scored fragment	18	39

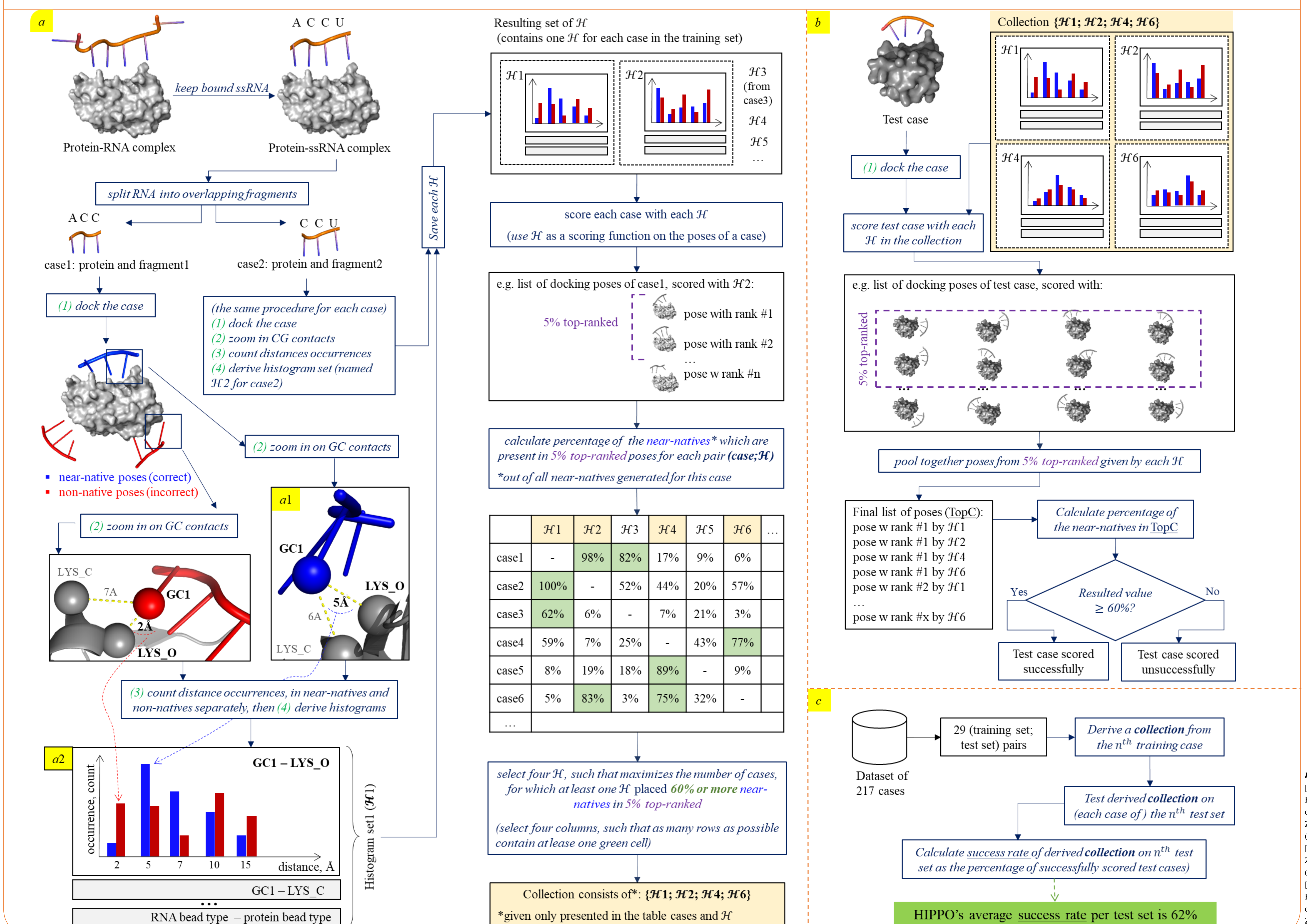
## 5 PERSPECTIVES

- Use in incremental modelling approach
- Build classifier to select best  $\mathcal{H}$  per case
- Use as a *sampling* function
- Apply to all-atom models
- Apply to general protein-ssRNA benchmark
- Apply to protein-ssDNA benchmark

Distribution of the % of NNs in selected\_poses per test case



## 6 PIPELINE



References:  
 [1] Chauvot de Beauchene I, de Vries SJ, Zacharias M. (2016). NAR  
 [2] Setny P, Zacharias M. (2011). NAR  
 [3] Chauvot de Beauchene I, de Vries SJ, Zacharias M. (2016). NAR



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

