Statistical Molecular Interaction Fields: A Fast and Informative Tool for Characterizing RNA and Protein Binding Pockets

Diego Barquero Morera[a], Giovanni Mattiotti[a], Alexander Kocev[a], Amshuman Rousselot[a], Lucas Rouaud[b], Hubert Santuz[b], Marc Baaden[b,*], Antoine Taly[b,*], Samuela Pasquali[a,*]

[a] *Université Paris Cité, Laboratoire Biologie Functionnelle et Adaptative, Inserm ERL U1133 35 rue Hélène Brion, Paris, France*
[b] *Université Paris Cité, CNRS, Laboratoire de Biochimie Théorique, Insitut Biologie Physico-Chimique, 13 rue Pierre et Marie Curie, Paris, France*

Abstract

Developing a physical understanding of the interactions between a macromolecular target and its ligands is a crucial step in structure-based drug design. Although many tools exist to characterize protein-binding pockets in silico, this is not yet the case for RNA, which has only recently been recognized as a suitable target for small ligands. Molecular Interaction Fields (MIF) are a useful tool to characterize the interactions of a given binding pocket. However, classical MIFs heavily rely on the use of probes, which makes their calculations accurate but very specific to the binding partners in question. We develop here a simple version of MIF, that we call Statistical Molecular Interaction Fields (SMIF), based on functional forms inspired by coarse-grained models and parametrized based on PDB structures

*baaden@smplinux.de, taly@ibpc.fr, samuela.pasquali@u-paris.fr

and previous statistical analysis of the main form of interactions typical of macromolecules, namely hydrogen bonding, stacking, and hydrophobic interactions. We show that these fields, despite their simplicity, are very informative and overall in agreement with pharmacophoric models. Thanks to a carefully optimized code, our calculations are fast and can be performed in bulk on a large set of binding pockets or even on a full macromolecule. As shown in a few representative examples, the latter possibility opens the way to the analysis of systems as large as 20 to 80 k atoms in relation to the surrounding environment, i.e., a lipidic membrane, a small ligand, or another macromolecular partner, allowing for a detailed visualization of the possible interactions.

1. Introduction

Small molecules modulating the behavior of biomolecular targets are the current main form of medication for most diseases and, although other forms of drugs exist such as biologics and therapeutic peptides, small-molecule drug design is still a very active research field [1, 2]. To this day, almost all approved drugs target proteins such as enzymes, receptors, ion channels, and more, but with the discovery of the functional role of RNA in many cellular processes, RNA molecules (non-coding RNA and mRNA) have been identified as possible targets [3] and research to design RNA drugs is growing in importance [4, 5, 6].

The open challenges in drug design are to improve the efficacy of a drug

by finding new compounds with affinities and specificity higher than those of their targets, and accounting for the flexibility of the target [7]. From a computational point of view, this problem is being tackled both using traditional molecular modeling approaches based on docking, scoring functions refinement, and molecular dynamics simulations[8, 9], and, more recently with the help of artificial intelligence [10, 11].

The development of drugs specifically targeting RNAs has recently opened a whole new research field. Up until recently only antibiotics were known to target RNA (in the ribosome), but it is now clear that non-coding and mRNAs constitute a new pool of possible targets, complementary and possibly wider, than the space of druggable proteins. Only roughly 2% of the human genome codes for proteins, while about 70% translates into ncRNA and most viruses threatening new pandemics have a single stranded RNA genome [12]. Drugs targeting RNAs must have their own specific properties, different from those targeting proteins. This requirement arises from the nature of RNA that are highly charged molecules constituted of nucleotide building blocks with less structural variability than amino-acids. The smaller RNA alphabet and the similar physico-chemical nature of the four nucleotides raise the problem of specificity of a drug in binding to a target [3]. Moreover, RNA are often subject to post transcriptional modifications such as methylation [13], which can substantially alter their structure and binding modes, and whose sites can become targets for small ligands and protein interactions [14, 15] . In general, the characteristics of RNA binding sites are less understood than

their protein counterparts [16, 17], but the tools developed to study in silico the interactions between macromolecules and small ligands were mainly developed for proteins and later adapted for RNA [18].

In order to gather a rational understanding of these interactions, of the role of flexibility, and of the specific differences between protein and RNA drugs, physical modeling is necessary and provides a guiding light to develop advanced computational methods, including AI-based ones. As RNA structural data is still rather limited, the rules of the physical interactions of these molecules should be transferable ones.

In this work, we present a novel approach for looking at molecular interactions based on the concept of interaction fields computed with simple functional forms based on the statistical analysis of existing complexes. Our approach has the advantage of allowing fast calculations and insightful visualizations. It has the potential to become a unique investigative tool integrated with molecular visualization software, enabling researchers to intuitively gain insights into the systems under study and design more targeted drug development strategies.

Molecular Fields. A typical approach used in molecular simulations and in structural analysis is to consider pairwise interactions between selected elements in the system to help rationalize the binding modes of a receptor and a ligand or between two biomolecules. An alternative point of view, more physics-inspired, is to focus on just one object, such as a macromolecule, and to consider the interaction "fields" it generates and that dictate how it might

interact with other partners. A field is an intrinsic property of the object and it consists in the modification of the properties of the space around it, induced by the presence of the object itself. This is the case for example of the gravitational field of the Earth, and of the electrostatic field generated by a point charge.

The idea of field can be applied to molecular interactions focusing only on one system and considering the possible interactions it can generate, independently of a specific binding partner. This approach is particularly interesting in the context of a biomolecule to be targeted with a small ligand: if we can characterize the properties of a binding pocket, we can then make hypotheses on the type of ligand that can bind to it and how.

The concept of Molecular Interaction Fields (MIF) was introduced already in the '70s and implemented initially in the '90s [19]. Molecular fields are typically obtained using probes that are positioned on a grid sampling the entire space of interest and are given as the interaction energy between the probe and the receptor computed from a molecular force field [20, 21]. The latest implementation of the GRID software, in 2021, accounts for over 70 different probes [22]. Over the years, molecular fields have been applied extensively in both ligand-based applications to assess ligand similarity, virtual screening and ADME predictions, and in structure-based applications to characterize and detect binding pockets, and for docking [23]. The widely used Autodock software computes atomistic force field MIFs of a binding pocket with the atom types of the ligands as probes and then uses the fields

to guide docking poses [24]. With a similar approach, higher resolution quantum calculations were used to characterize non-bonded interactions with high spatial precision, focusing on the interactions of small subsets of atoms [25].

All these approaches to compute MIFs are based on the use of probes, making their calculations energetically consistent, and the information obtained specific to the complex, as it depends on the nature of two partners. Because of the high accuracy, these calculations can be rather expensive and they are typically performed only in small regions of the molecule, such as a binding pocket.

Wanting to develop a fast and general method to build an intuitive understanding of possible molecular interactions, we have defined simplified molecular interaction fields, based on a statistical analysis of the structural database of proteins and nucleic acids, and that take inspiration on coarse-grained molecular models [26, 27]. Speed of these calculations is crucial if we want to apply them fluidly in interactive visualization and eventually integrate these calculations into common visualization software. Speed is particularly relevant to study RNA molecules. Given their inherent flexibility, the calculations may need to be performed on an ensemble of structures and not just a single one, to bear relevance.

Over the years we worked extensively on the development of visualization software with the specific aim to allow the user to build an intuition on the system under study [28, 29, 30]. Our UnityMol software was developed with the ability of performing interactive simulations and allows an immersive

approach with virtual and augmented reality [31, 32, 33].

For our purposes, field calculations don't need to be as detailed as those evaluated from atomistic force fields. We therefore define effective statistical potentials (that we name SMIF, for Statistical Molecular Interaction Fields) for hydrogen bonding, stacking and hydrophobic effect. We rely on the APBS (Adaptive Poisson-Boltzmann Solver) software for calculations of electrostatic fields generated by point charges [34]. Two important differences between our fields and the original MIFs are that we do not use atomistic probes, making our calculations particularly efficient and our results general. We adopt a coarser view, which in particular allows us to easily define a stacking term.

We show that, for known complexes of a biomolecule and a small ligand, the statistical fields computed solely on the knowledge of the binding pocket are in excellent agreement with the positions of the pharmacophores obtained for the ligand. This observation indicates that even fields built with the simple approach that we propose are highly informative of the interactions accessible to the biomolecule. It is important to notice that our excellent results hold for both proteins and RNAs.

Since calculations are inexpensive and the correspondence between computed and known properties of the biomolecule is satisfactory, we can use SMIFs to investigate beyond the spatial confines of a given binding pocket and extend the analysis to the whole macromolecular systems. This extension can have useful applications in the analysis of interfaces of an entire

biological macromolecule and in the search of possible binding pockets themselves. As we show in an example, in the context of the open problem of RNA drug design, exploiting the information from SMIFs could be a valid approach to identify RNA binding pockets. Our approach can be used to rationalize the interaction between large macromolecules, as we show in an example of a ribosomal subunit interaction with a Internal Ribosomal Entry Site.

In this work, we first present how SMIFs are defined and built. We then present a few examples where we visualize the fields, followed by the benchmark with pharmacophores. We conclude with three examples of use of SMIFs on entire macromolecular systems. To make our approach easily accessible to the community, we provide the documented source code, executable builds and a docker container with a set of benchmarks and a series of scripts for visualization. In virtual reality, the fields can be visually explored or used as guidance for interactive docking, using the UnityMol software and provided examples.

2. Fields models

Statistical fields are defined to estimate the tendency of single atoms/-groups to participate in an interaction. They are built with a top-down approach, which consists in writing a plausible functional form for the interactions, guided by the knowledge of the atomic positions observed in experimental structures, and by assigning the free parameters based on a statistical

analysis of a database of structures, or on more detailed calculations and experimental data providing structural information [35, 36, 37, 38]. The molecular interactions we want to account for are electrostatics, stemming from the presence of net or partial charges, hydrogen bonds and stacking, and the hydrophobic effect, stemming from the rearrangement of water molecules.

From experimental structures and molecular simulations, we know that hydrogen bonding and stacking are present when the atoms adopt well defined geometries. Our potentials are chosen to have minima in the optimal geometry and to rapidly go to zero when atoms are far apart, to obtain well localized interactions. For our fields we choose Gaussian functional forms as they have these properties and are easy to interpret as probabilities.

Stacking. The stacking potential field aims at quantifying a property that depends on the position and orientation of the ensemble of atoms composing aromatic rings. The aromatic rings considered correspond to the natural occurrences in residues of proteins (phenyl group for Phe and Tyr; imidazole for His; indole for Trp) and RNA (pyrimidine for Ura and Cyt; purine for Ade and Gua).

For this field we adopt a coarse-grained approach and define a potential based on the location and alignment of a pair of aromatic rings given the distance r of their center of mass and their relative vertical position $\alpha = \frac{\mathbf{r}}{r} \cdot \mathbf{n}$, where $\mathbf{n}$ is the unitary vector normal to the plane of the ring taken as reference (Fig. 1A). In principle, stacking interactions should depend also on the relative orientations of the planes of the two rings ($\mathbf{n_1} \cdot \mathbf{n}_2$), but a

statistical analysis of aromatic ring interactions between biomolecules and ligands reveals that there is very little variation from a perfectly parallel orientation. In our model we therefore choose to consider only fully parallel rings and not include this additional variable in the potential.

To compute the field, we place a hypothetical aromatic ring centered in each point in space $\mathbf{x}$, and we compute its interactions with nearby aromatic rings in the system. The functional form and the parameters of this field were obtained through the analysis of experimentally available structures (details given in Supplementary Information, Section 1). We find that a bivariate Gaussian distribution modeling the interaction of two aromatic rings fits the structural data well (Figure 1:B,C). We therefore define the stacking field as:

$$\phi_{ST}(\mathbf{x}) = -\sum_i \exp\left[-\frac{1}{2}(\mathbf{v}_i - \boldsymbol{\mu}_{ST})^T \mathbf{S}_{ST}^{-1}(\mathbf{v}_i - \boldsymbol{\mu}_{ST})\right] \tag{1}$$

with:

$$\mathbf{v_i} = \begin{bmatrix} \alpha_i \\ r_i \end{bmatrix} \quad \& \quad \mu_{\mathbf{ST}} = \begin{bmatrix} \alpha_0 \\ r_0 \end{bmatrix} \quad \& \quad \mathbf{S_{ST}} = \begin{bmatrix} s_{\alpha,\alpha} & s_{\alpha,r} \\ s_{\alpha,r} & s_{r,r} \end{bmatrix} \tag{2}$$

where i are the aromatic rings located within 4 standard deviations of the distance ($\sqrt{s_{r,r}}$) from $\mathbf{x}$, r_0 and α_0 are the optimal distance of the center of mass of the rings and the optimal angle for their vertical offset, and the

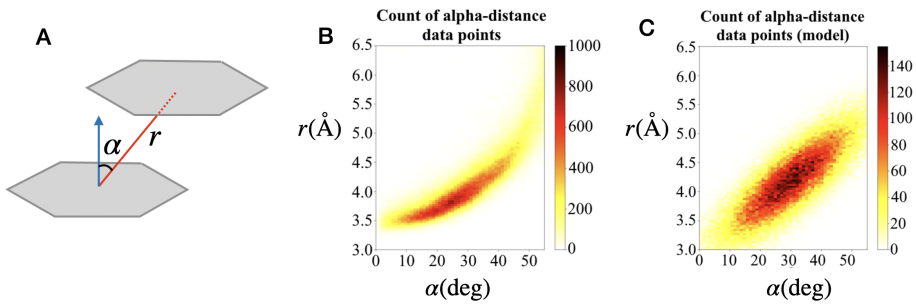

Figure 1: A: model of stacking interactions highlighting the two variables, r and α used in the definition of the potential field. B: Experimental distribution of values of r and α extracted from the dataset. C: Modeled potential field for a pair of aromatic rings.

matrix $\mathbf{S_{ST}}$ contains the variances of the Gaussian.

Fitting this function with the distribution obtained from structural data, we determine the model's optimal parameters to be:

$$\mu_{\mathbf{ST}} = \begin{bmatrix} 29.98^\circ \\ 4.19\text{Å} \end{bmatrix}, \qquad \mathbf{S_{ST}} = \begin{bmatrix} 169.99(^\circ)^2 & 6.62(\text{Å}\cdot^\circ) \\ 6.62(\text{Å}\cdot^\circ) & 0.37(\text{Å})^2 \end{bmatrix} \tag{3}$$

The stacking potential field of a single aromatic group takes the shape of two slightly flattened tori (i.e. "donut" shape), both below and above the aromatic plane (Figure 2:A,B). At each point in space the field is a sum of the potential around each aromatic group, which means that in actual pockets more than one pair of tori can be present. These will naturally overlap with each other in regions where multiple aromatic groups are in close proximity. Points where multiple tori intersect imply a higher propensity to form one or multiple stacking interactions, potentially from both sides of the aromatic plane.

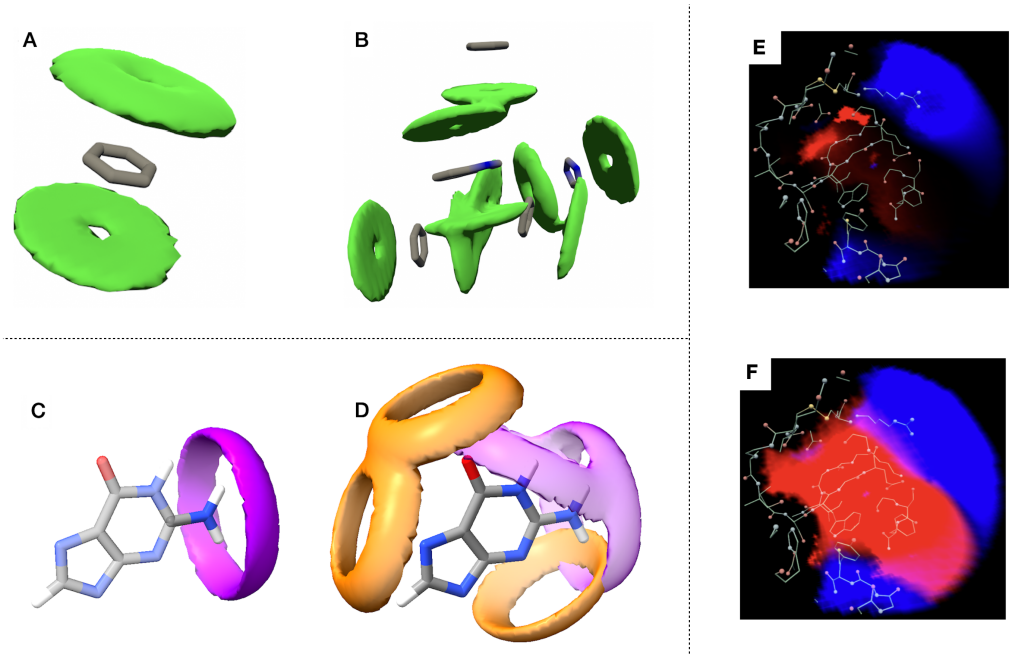

Figure 2: Top left: visualization of stacking fields for a simple aromatic ring (A) and for a binding pocket in which multiple aromatic rings are present (B). Bottom left: visualization of hydrogen bond donor field for a single nitrogen atom as a donor (C) and for multiple donor (purple) and acceptor (orange) atoms (D) for a guanine residue. Right: Visualization of the APBS electrostatic potential (E) and of the logarithmic transformation LogAPBS (F). Regions favoring the interactions with positive charges are in red and with negative charges are in blue.

Hydrogen Bonds. Hydrogen bond interactions are usually characterized by the distances and angle between acceptor, donor and the hydrogen atom. Unfortunately, this information requires the location of hydrogen atoms, that is often lacking in PDBs obtained via X-ray crystallography experiments. For this reason, we define our hydrogen bond fields by considering the typical distances between donors (D) and acceptors (A), and the angles that are formed involving the antecedent atoms bound to the donors (AD) and the antecedent atoms bound to the acceptors (AA). More specifically, we define

two fields: one for the hydrogen bond donors (HBD), and another for the hydrogen bond acceptors (HBA). The HBD field depends on the distance from the donor ($r := |\mathbf{x} - \mathbf{x}_i|$) and the DA - D - A angle β, while the HBA field on the distance from the acceptor and the AA - A - D angle (using same symbols for simplicity). We define the hydrogen bonding field as the product of two normal distributions of distance and angle as:

$$\phi_{HB}(\mathbf{x}) = -\sum_i \exp\left[-\frac{(r - r_0)^2}{2\sigma_r^2}\right] \exp\left[-\frac{(\beta - \beta_0)^2}{2\sigma_\beta^2}\right] \tag{4}$$

Based on statistical studies on these same angles and distances reported in [36, 37, 38], we take the parameter to be $r_0 = 2.93$Å, $\beta_0 = 109.0°$ for HBD and 129.9° for HBA, $\sigma_r = 0.21$Å and $\sigma_\beta = 20°$.

With this definition, hydrogen bond fields take the shape of rings with the axes passing through the donor and the AD or the acceptor and the AA. The only exception is when either D or A are bound to two atoms in which case the axis is taken to pass through their center of geometry. Similarly to the stacking potential, several rings will most likely overlap with each other when studying actual pockets (see Figure 2:C,D).

Hydrophobicity. To properly consider hydrophobic effects one would need to dynamically analyze the interactions between the receptor and water. However, for proteins, the propensity of different amino acids toward a hydrophobic or hydrophilic behavior is well known. Indeed, hydrophobicity is usually

defined as a property of amino acid residues and visualization software uses hydrophobicity scales to color different amino acids accordingly. A variety of hydrophobicity scales exist and they are often rooted on the idea of transfer energy needed to put a given amino acid in a membrane-like environment, which could be derived from experiment or computation.

Similarly to what is done for hydrogen bonds and stacking, but without the angular dependence, for each atom in the pocket we define a Gaussian field and we assign the height of the Gaussian based on the Wimley and White hydrophobicity scale [39]. This scale assigns an hydrophobicity value K_i to each amino acid; when this is negative, it means that the residue is hydrophilic. Projecting the property of the residue to each atom allows preserving the detailed geometry when computing volumetric properties of the pocket. For each point in space we then add the contributions of all Gaussians of neighboring atoms. We calculate two different fields: one called *hydrophobic*, that is the sum of all the Gaussians with $K_i > 0$, and one called *hydrophilic* ($K_i < 0$). In general, for both, the analytical definition follows:

$$\phi_{HP}(\mathbf{x}) = \sum_{i=1}^{N} K_i \exp\left[-\frac{(r_i - \mu_{HP})^2}{2 \cdot \sigma_{HP}^2}\right] \tag{5}$$

These Gaussians are peaked at a distance μ_{HP} from atom i, equal to 4Å for the hydrophobic field and 3Å for the hydrophilic one. For the first, the choice is inspired by the statistical study that is reported in [35], where

the authors show that the typical range of distances between hydrophobic species goes from 3Å up to 7Å, but is peaked around 4Å. For the second, we considered the typical distances between an hydrophilic species and a water molecule (2.5Å - 3.5 Å), assuming that the probability of finding a water molecule at a given distance r_i from the hydrophilic species i is comparable to the probability of finding another hydrophilic/polar species. Based on these ranges, we also chose $\sigma_{HP} = 1$Å for the hydrophobic field and $\sigma_{HP} = 0.25$Å for the hydrophilic one.

Wimley and White hydrophobicity scale, or similar scales, exist for proteins but not for nucleic acids. Because of the net negative charge of the phosphate group, the backbone of RNA is hydrophilic. The sugar ring is also hydrophilic as it easily forms hydrogen bonds with water because of the OH group. Composed of aromatic rings, bases are hydrophobic and experiments measuring the energy for the transfer of bases from a hydrophobic to a hydrophilic environment give a rough indication of the scale of hydrophobicity between the four bases [40]. From these considerations, we can attempt to give an approximate hydrophobicity scale for the different entities composing a nucleic acid, looking at chemical similarities with amino acids. The negatively charged phosphate group is of similar chemical nature as the side chains of the negatively charged amino acids. We therefore choose the value for the nucleic acid backbone to be the same as that of Glutamic acid, i.e. $K_{phos} = -2.02$. The OH group of the sugar resembles the side chain of Serine. We therefore assign to the sugar the same value as Serine, i.e.

$K_{sug} = -0.13$. Values for the hydrophobic bases are taken to be that of Tryptophan for purines ($K_{AG} = 1.85$), since both contain a double aromatic ring, and that of Pheninalanine for pyrimidines ($K_{UC} = 1.13$), both containing one aromatic ring. Even though the RNA field computed this way lacks any quantitative precision, this very empirical hydrophobicity scale still enables us to gather an idea of the hydrophobic and hydrophilic regions around these molecules that can be tested against the experimental data of known RNA/ligand complexes (see benchmarks section).

Electrostatic Potential. The electrostatic potential field is computed through the Adaptive Poisson-Boltzmann method (APBS), which is a widely used approach to treat electrostatic effects in solution, given an implicit treatment of the solvent [41]. The APBS software generates the electrostatic field on a grid in an assigned volume around the molecule, hence we have simply incorporated the APBS pipeline into our workflow, interpolating the values of APBS field to the grid points defined for our calculations.

Values for the electrostatic field can vary greatly within a system, with very high values occurring in close proximity to atoms with partial charges, while the values quickly dissipate when moving towards the unoccupied volume of the pocket. This wide range of values can be troublesome for visualization. To best exploit information from the electrostatic field, we worked on post processing of the APBS field to generate an output useful for visualization. Two solutions in particular allowed us to improve the visualization of the electrostatic field. The first one, as explained in more details in the next

section, consists of trimming regions of space too close to the molecule to be accessible by a ligand. This approach effectively excludes very high values of the field located at the position of the charges. The second, which can at times also be useful, is to convert the field in a logarithmic scale. Since the field can adopt both positive and negative values, this is done by first taking the absolute values of the negative energies and converting both positive and negative values on a logarithmic scale, then applying two empirical cutoffs on the highest and lowest values of both sets (positive and negative), which occur on sparse points, to reduce noise. After subtracting the values of the lower boundary to bring the smallest retained value to zero, we bring back the physical meaning of signed electrostatic potential multiplying by -1 all values that were originally negative. This process generates a second field named LogAPBS. APBS visualization allows to clearly spot maxima and minima of the electrostatic field, while LogAPBS allows to better identify the boundaries between regions of interactions favoring positive or negative charges (Figure 2:E,F).

3. Fields calculation and visualization

Field calculations can be launched from the command line and we provide examples for doing so. The input for the calculation comprises a structure file (PDB), the location of the center of the pocket, which, for example, can be given as a set of coordinates deduced from the presence of a ligand in a complex, and a radius for the pocket. It is important to notice that all

calculations are then performed only on the receptor, ignoring the possible presence of a ligand already in the pocket. The algorithm defines a sphere centered at the given position and of the given radius R and builds a grid of preset mesh spacing inside. The mesh is chosen small enough to be able to follow smoothly changes in the fields (typically a fraction of an Å).

Field values are written in the form of 3D matrices of uniform sizes 2×R in each dimension, where each entry is the value of the given field at the spatial coordinate corresponding to the matrix indices. Values of the fields outside the sphere of interest are set to zero. In order to account for the physical volumes of the atoms of the receptors, field values for points in close proximity to pocket atoms are set to 0. The distance threshold for proximity is set to 2.5 Å for hydrogen bonding, for which atoms come in close contact because of the quasi-shared hydrogen, and to 3.0 Å for all other fields, for which atoms cannot come too close because of electronic shells repulsion. After this trimming, it can still happen that some points in space that are outside the actual pocket cavity are still included in the matrix. These are identified as clusters of points disconnected from all others. To remove these clusters, a graph traversal strategy [42] is employed, starting from the center of the sphere and visiting all untrimmed neighboring points; points that were not visited at the end of the traversal are further trimmed out. Each field grid is stored as a *.dx* or *.mrc* file, which can easily be imported with most molecular visualization programs, and then displayed through isosurfaces, represented in solid colors or as a mesh. A discussion on

visualization strategies is presented in Supplementary Information, Section 2.

The typical time to compute a single grid is $\sim 1s$ for a pocket and $\sim 10 - 20s$ for a whole macromolecule, with a grid resolution of 0.25Å. The longest time is required to write the file in *.dx* format, which for a whole macromolecule with 10k atoms is $\sim 10 - 40s$, while it is faster ($\leq 10s$) for the *.mrc* format. The calculation of the electrostatic field with APBS, which is done with the homonymous software, typically takes tens of seconds to minutes, depending on the system's size: this is always the slower step in the pipeline to obtain the full set of fields.

To render the whole process of field calculations inside a known pocket straightforward and their visualization easily accessible, we implemented scripts to launch their calculations and to visualize them with UnityMol. We also provide scripts for exploration in virtual reality and demonstrate its use in supplementary videos on Youtube. In the light of future applications in interactive docking for drug design, we tested field visualizations in virtual reality using an interactive docking mode under development. We used a stacking field visualization to guide precise positioning of a cyclic fragment in the scene, as illustrated in the supplementary video.

4. Benchmark Results

To assess the efficacy of the fields in locating favorable ligand positions, we designed a benchmark of 10 protein binding pockets and 10 RNA bind-

ing pockets with experimentally resolved structures in complex with a small ligand. We want to evaluate if the properties that emerge from the fields are in agreement with the known physical characteristics of the ligand. For example, when we compute a hydrophobicity field, a coherent result would be to find polar portions of the ligand to be inside of what we identify as the hydrophilic regions of the pocket, and non-polar portions to be in the hydrophobic parts. Similarly, for stacking we would expect to find aromatic portions of the ligand to be placed in regions of high stacking propensity of the stacking field, for hydrogen bonds we would expect to find donors/acceptors of the ligand inside the spheres of acceptors/donors of the pocket, and for electrostatics to find explicit charges of the ligand to occupy volumes with high electrostatic field values of opposite charge.

In this section we first report a few representative instances of pocket fields superposed to the experimental ligand, to visually illustrate the correspondence between fields' high values (hot spots) and ligand position in the pocket. This analysis will illustrate the complexity of the fields observed in actual pockets, with features that could not have been guessed by simple observation of the pocket surface, as intrinsic to the pocket volume. Then, for each system, we build a pharmacophore-based model of the ligand using *pharmit* [43] to quantify the correspondence between the field properties and the ligand. We numerically check the overlaps between pharmacophores and the corresponding field counterparts.

We selected our benchmark pockets to cover a wide spectrum of interac-

PDB ID	Molecule Type	Interaction
1BG0	Arginine Kinase + ADP	ELE + HB
1EBY	HIV-1 protease + inhibitor BEA369	HB + HP
1EHE	P450S + heme group (with Fe)	ELE + HB + HP
1H7L	Glycosyltransferase SpsA + dTDP (and Mg)	ELE + HB + HP
1IQJ	Human coagulation factor Xa + M55124	STK + ELE + HB + HP
1OFZ	Fungal lectin + beta-L-fucopyranose	ELE + HB + HP
3DD0	Carbonic anhydrase II + Ethoxzolamide	HP + HB + ELE
3EE4	Mn/Fe oxidase + myristic acid	HP + ELE
5M9W	Thermolysin + inhibitor JC65	HP + HB + STK + ELE
6E9A	HIV-1 wt protease + PDB:J0S	HB + HP + ELE
1AKX	RNA HIV + Argininamide	HB Acceptors/Donors
1I9V	tRNA + Neomycin	HB Acceptors/Donors
2ESJ	16S RNA + Lividomycin A	HB Acceptors/Donors
4F8U	bacterial ribosome RNAS + sisomicin	HB Acceptors/Donors
5BJO	RNA aptamer + DFHO	HB Acceptors/Donors + STK
5KX9	FMN Riboswitch + Ligand 6YG	HB Donors + STK
6TF3	NAD^+ riboswitch + Cordycepin 5-triphosphate	HB Donors + STK
7OAX	RNA aptamer + spermine	HB Acceptors/Donors
7OAX	RNA aptamer + DMHBO	HB Acceptors/Donors + STK
8EYV	Beetroot dimer RNA + DFHO	HB Donors + STK

Table 1: Protein benchmark (top), RNA benchmark (bottom) and most relevant interactions observed for each system as reported by PLIP. "ELE" means pure electrostatic interactions and gathers both salt bridges and metal complexation. "STK" gathers the π-stacking interactions. "HPH" stands for hydrophobic (pseudo-)interactions. "HB" are hydrogen bonds.

tions. Table 1 reports the systems chosen and the main modes of interaction between the pocket and the ligand identified from the literature and using the PLIP software [44].

4.1. Selected examples

4.1.1. Mn/Fe oxidase bound to myristic acid (PDB: 3EE4)

The Mn/Fe oxidase from *Mycobacterium tuberculosis* is an enzyme complex involved in the oxidative stress response and iron homeostasis within the bacterium [45]. This enzyme complex plays a critical role in detoxifying reactive oxygen species (ROS) and regulating intracellular iron levels to ensure bacterial survival and persistence within the host. PDB entry 3EE4

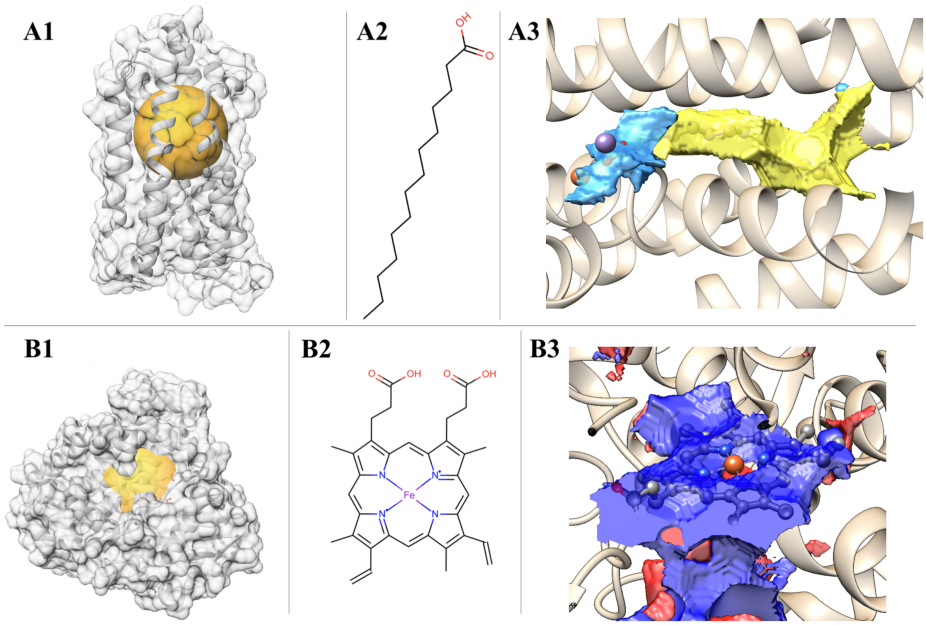

Figure 3: **A:** PDB system 3EE4: Mn/Fe oxidase bound to myristic acid. **B:** PDB system 1EHE: Monooxygenase cytochrome P450S bound to a heme group with iron. **A1) and B1)** representation of the whole protein structure and highlight (in orange) of the binding region where the fields are calculated; **A2) and B2)** 2D molecular graph of the ligand (A: myristic acid, B: heme group)); **A3) and B3)** the region where the field of interest is more intense (hydrophobic in yellow, hydrophilic in lightblue, positive/negative electrostatic in blue/red) superimposed to the ligands (balls and sticks/licorice representation). Protein structures shown with cartoon representation (B3: some protein atom surrounding the ligand also show in licorice).

shows it in complex with myristic acid [46], a fatty acid that is often found as a component of phospholipids, triglycerides, and other lipid molecules.

The pocket for this system has been described as a narrow and hydrophobic cavity that widens at its internal end, leading to a larger volume where metal cations coordinate with the protein and with more H-bond possibilities [46]. The representation of the hydrophobic potential matches accurately with this description, as seen in Figure 3A. The aliphatic section of

the ligand effectively overlaps with the mostly hydrophobic pocket, while the oxygen atoms at its extremity are placed in a hydrophilic region.

4.1.2. Monooxygenase cytochrome P450s bound to a heme group with iron (PDB: 1EHE)

Monooxygenase cytochrome P450s are a diverse superfamily of heme-containing enzymes found in all domains of life, including bacteria, archaea, fungi, plants, and animals [47]. They play essential roles in the oxidative metabolism of a wide range of endogenous and exogenous compounds, including drugs, fatty acids, steroids, and environmental pollutants. In PDB 1EHE, the oxygenase is bound to an heme group, which is notoriously capable of undergoing redox reactions, where the iron ion alternates between its ferrous (Fe2+) and ferric (Fe3+) oxidation states. This redox activity is essential for electron transfer in cytochrome proteins.

The binding pocket of this protein is characterized by a positively charged cluster of lysine and arginine residues, which plays an important role on how it interacts with the heme group ligand [48]. Indeed, the visualization of the electrostatic potential (Figure 3B3) reveals a positive electrostatic field in the binding pocket, which is ideal to accommodate four central electronegative nitrogens of the ligand, that then coordinate with the positive Fe ion. The positive field in the pocket, exhibits a hole of slightly negative charge in the location of the cation, therefore reinforcing its position lead by the nitrogens.

Additional observations (not explicitly reported in Figure 3) include the

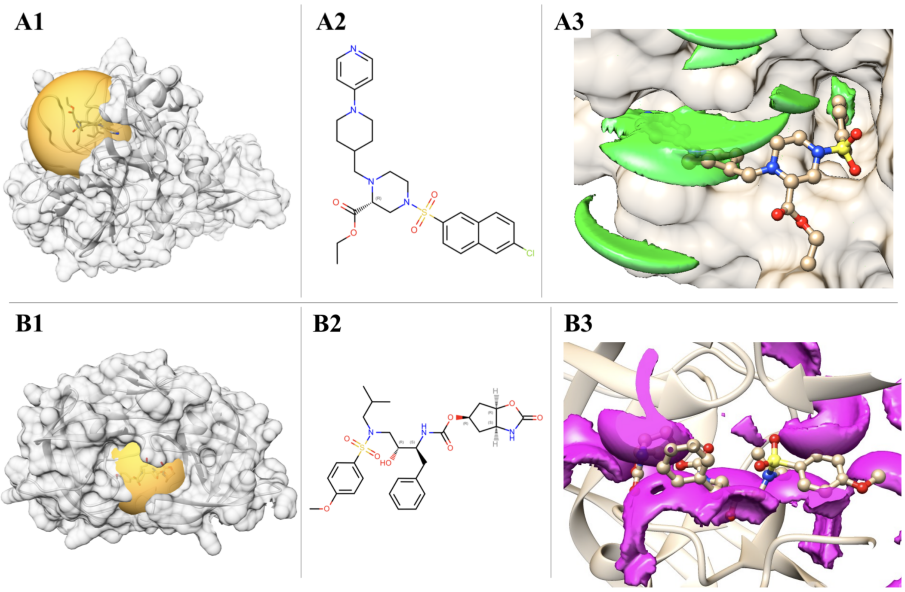

Figure 4: **A:** PDB system 1IQJ: Human coagulation factor Xa in complex with M55124. **B:** PDB system 6E9A: HIV-1 wild type protease bound to PDB:J0S. **A1) and B1)** representation of the whole protein structure and highlight (in orange) of the binding region where the fields are calculated; **A2) and B2)** 2D molecular graph of the ligand (A: M55124, B: PDB:J0S)); **A3) and B3)** the region where the field of interest is more intense (stacking in green, HBA in magenta) superposed to the ligands (balls and sticks representation).

stacking potential slightly overlapping with one of the aromatic groups, while the hydrogen bond potentials are displayed on top of the ligand's heteroatoms. The hydrophobic main body of the ligand overlaps with a large region of hydrophobic potential, while the outside oxygen atoms are placed inside a hydrophilic section.

4.1.3. Human coagulation factor Xa in complex with M55124 (PDB: 1IQJ)

Human coagulation factor Xa, often abbreviated as factor Xa, is a serine protease enzyme that plays a crucial role in the coagulation cascade, a series

of biochemical reactions that lead to the formation of blood clots [49]. It is shown in complex with the inhibitor M55124 in the PDB entry 1IQJ.

The ligand's aromatic pyridine group can be seen positioned in a significant stacking field region (Figure 4A3). Most of the pocket is hydrophilic (not shown), while the few hydrophobic sections neatly match the most hydrophobic parts of the ligand.

4.1.4. HIV-1 wild type protease bound to PDB:J0S (PDB: 6E9A)

The HIV-1 protease plays a critical role in the viral life cycle by cleaving the viral polyprotein precursor into functional, mature proteins necessary for viral replication and assembly [50]. In [50], the authors report a plethora of new potent HIV-1 protease inhibitors and in PDB entry 6E9A the protease is shown in complex with the chemical compound corresponding to the PDB entry J0S.

The pocket for this protein is reported to have a high binding affinity with its ligands, which can be attributed to the possibility of creating multiple hydrogen bonds and hydrophobic interactions [50]. The hydrogen bond acceptor field visualized in Figure 4B3 overlaps nicely with the expected oxygens and nitrogens of the ligand (this is hard to see on a printed picture, but can be better observed visualizing the fields on the computer using the pre-computed fields we provide in the repository for all benchmarks). The hydrophobic and hydrophilic regions (not shown in the Figure) appear to coincide to their counterparts in the ligand.

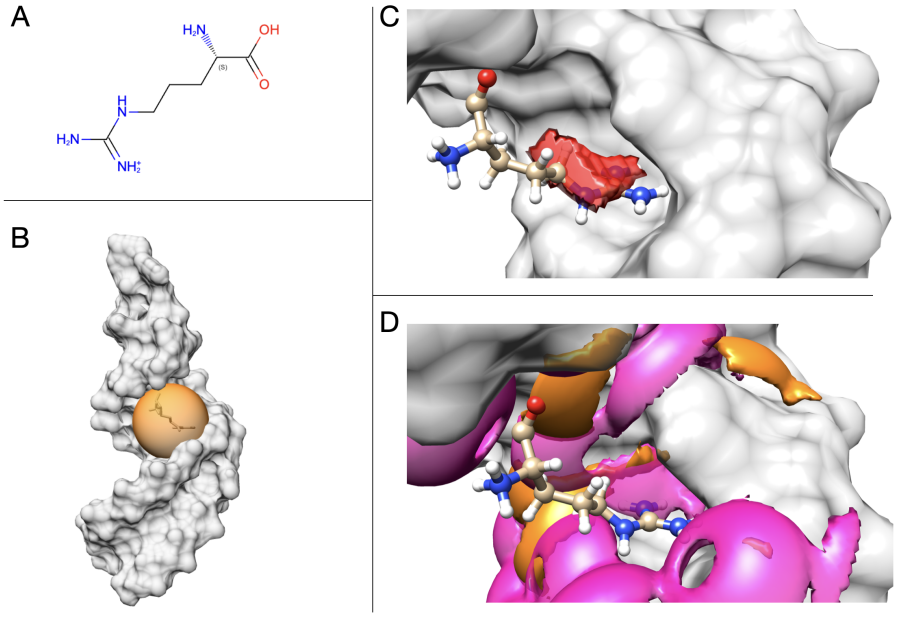

Figure 5: PDB system 1AKX: HIV-2 trans activating region RNA in complex with argininamide. A: argininamide molecular graph, showing the location of the positive charge at physiological pH; B: structure of the full system with the binding pocket highlighted in orange; C: electrostatic field (red); D: HBA (magenta) and HBD (orange) fields.

4.1.5. HIV-2 trans activating region RNA in complex with argininamide (PDB: 1AKX)

The HIV-2 trans activating region RNA (often abbreviated as TAR RNA) is a key regulatory RNA element in the genome of HIV-2, the second type of the Human Immunodeficiency Virus (HIV). It plays a critical role in the transcriptional regulation of the virus. The TAR region forms a stable stem-loop structure that interacts with viral and host proteins, notably the TAT (Transactivator of Transcription) protein, which is crucial for efficient viral replication. This interaction occurs in the central bulge of TAR [51, 52]. Without the interaction between TAR RNA and Tat, HIV-2 replication would be

severely impaired because the transcription of viral genes would not proceed efficiently [53, 54]. Its pivotal role makes TAR RNA and, more in general, the TAT-TAR complex, an optimal target for therapeutics aimed at disrupting HIV replication. In PDB id 1AKX [55], the HIV-2 TAR structure is solved by NMR spectroscopy in complex with argininamide, a molecule that is obtained from the formal condensation of the carboxyl group of the amino acid arginine with ammonia. At physiologicl pH, the guanidinium group of argininamide carries a $+1e$ charge, that is delocalized among the three amino groups, as shown in Fig. 5A. The electrostatic field, obtained by filtering the APBS output as described in the methods, is able to highlight a very high-intensity patch of electrostatic field precisesly where the guanidinium group is located, as shown in Fig. 5D. Moreover, the HBA and HBD fields clearly detect the formation of hydrogen bonds within the complex, involving the oxygen as acceptor and some nitrogen as donors (Fig. 5C).

4.1.6. Homo sapiens ribosomal decoding site

As a last example, we report a case where multiple ligands are known and experimentally solved for the same molecule, even in the same binding site. The target is the eukaryotic ribosomal decoding site, and the PDB entries we used are 2G5K (HOLO form, bound to apramycin, or AM2, in two different binding sites) and 2O3V (HOLO form, bound to a synthetic compound named N33). We report the structures and fields of 2G5K and 2O3V together with the ligands AM2 and N33 (Fig.6A).

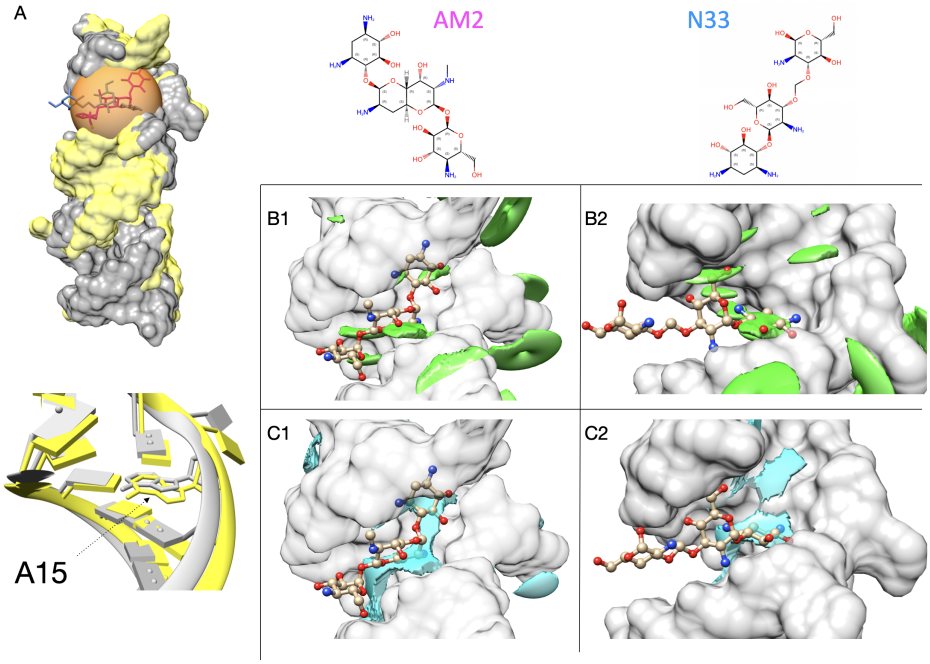

Figure 6: A: 3D structure of the *Homo sapiens* ribosomial decoding site in complex with 2 different ligands: N33 (blue/gray, from PDB 2O3V), AM2 (magenta/yellow, from PDB 2G5K). Details of the position adopted in the two structures by base A15 are shown in the bottom of the panel. Stacking (B1 and B2) and hydrophobic (C1 and C2) fields of the binding pockets superposed with the ligands.

The secondary structure of the RNA system is the same in both the structures, but interestingly one can notice that the unpaired bases have different dispositions in the 3D space. One of these differences is crucial for the interaction with the ligands: in RNA-N33 the adenine in position 15 (with the atoms and bonds of the base explicitly represented in Fig. 6A, lower image) is slightly more exposed towards the binding site than in RNA-AM2, allowing the formation of a stable stacking interaction with one aromatic ring of N33 (Fig. 6B2) which is absent in the RNA-AM2 complex (Fig.

6B1). This observation is in line with the discussion made in [56], where the authors not only report the crystal structures of these complexes, but also demonstrate an increased binding affinity of N33 to the eukaryotic A-site with respect to the prokaryotic counterpart, and a selectivity of N33 to bind only one of the 2 possible pockets found in the RNA system, differently from AM2. Thanks to our fields analysis, we are able to show and propose an explanation for these observations.

In addition, we reported the hydrophobic fields found in the same pocket, for the two complexes (Fig. 6C1 for RNA-AM2 and Fig. 6C1 for RNA-N33). They both show a high-values region, that can explain the first stage of the complex formation: the hydrophobic nature of both ligands could act as driving force, leading them close to the binding pocket. After this first, long-range (effective) interaction, the ligands are able to form local and stronger interactions (such as hydrogen bonds and stacking).

4.2. Overlaps of fields with known pharmacophores

To quantify in a general manner what is observed through visual inspection, for the ligands present in the benchmark structures, we compared the spatial regions where the fields assume high (absolute) values with the pharmacophore models generated by *pharmit* [43]. Regions with low- or zero-intensity fields were filtered out by retaining only those points with values that deviate by a maximum of 2 sigma from the peak value of the individual Gaussian function. For the electrostatic fields, a threshold of $\pm 20 k_B T / e_c$

was chosen by visual inspection to highlight the main features of the field. The pharmacophores (PPs) of each type modeled by pharmit were compared with the filtered fields of their counterpart, remembering that the fields refer to the receptor, while the pharmacophores refer to the ligand. Therefore we have to compare stacking site PP with stacking field, hydrophobic field with hydrophobic PP, but negative charge site PP with positive APBS field, hydrogen bond donor field with hydrogen bond acceptor PP, etc. To make this comparison, for each system we first evaluate the interactions between the ligand and the receptor using the PLIP software [44] to identify the PPs of the ligands that are actually involved in the interaction, and then check if there is an overlap between the PPs and the corresponding field. We count the interaction if there is a spatial overlap between the filtered fields and a sphere around the PP coordinates whose radius is chosen in agreement with the literature [57]: R_{HBA} = 0.5 Å, R_{HBD} = 1.0 Å, R_{STK} = 1.0 Å, R_{HPH} = 1.0 Å, R_{ELE} = 1.0 Å. No comparison has been done with the *hydrophilic* field, since hydrophilic pharmacophores don't exist.

Table 4.2 summarizes the number of PPs that overlap with the filtered fields compared to the interactions predicted by PLIP. The comparison between fields, pharmacophores and PLIP predictions, however, requires some care. In a few cases, the number of pharmacophores overlapping with the fields is higher than the number of interactions identified by PLIP, suggesting the existence of interactions not captured by PLIP criteria, possibly due to strict cutoffs used by PLIP to identify hydrogen bonds and stacking, which

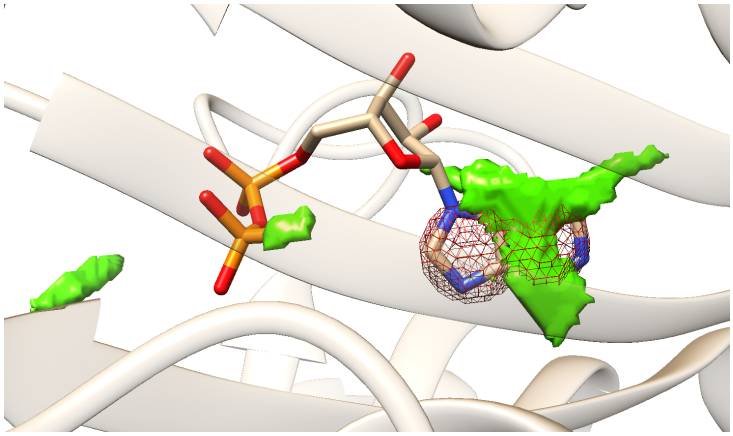

Figure 7: Zoom around the ligand ADP in PDB 1BG0, showing the filtered stacking field in green (as described in the text) and the relative pharmacophores' field in brown, as a mesh grid. In this case, 2 stacking PP sites are overlapping with the field although one to a lesser extent than the other. In this case, PLIP detects 1 stacking interaction between the protein and ADP.

could result in PLIP missing a few meaningful interactions. An example of interaction not accounted for by PLIP but for which we have an agreement between pharmacophores and field is presented in Figure 7. In other cases pharmit does not predict pharmacophores where PLIP detects an interaction. This occurs for example for 6e9a and 3ee4, HPH where the fields agree with PLIP but the pharmacophores are missing. We note that PLIP detects only one hydrophobic interaction for RNA systems. The combination of pharmit and the fields instead detects a substantially higher number of these interactions. The discrepancy is due to the fact that PLIP attributes only one property to the interaction, either stacking or hydrophobic, while pharmit accounts for both. Since most hydrophobic interactions in RNA arise from the aromatic rings, PLIP attributes them mainly to stacking.

	HBA		HBD		HPH		NEG		POS		STK	
1bg0	8(13)	3	3(3)	2	1(2)	0	2(2)	2	0(0)	0	2(2)	1
1eby	2(8)	2	6(6)	3	4(4)	4	0(0)	0	0(0)	0	0(4)	0
1ehe	4(6)	1	0(0)	0	8(8)	8	2(2)	2	0(0)	0	0(2)	0
1h7l	2(11)	3	1(2)	1	2(2)	2	0(2)	1	0(0)	0	1(1)	0
1iqj	1(5)	0	0(0)	0	4(5)	2	0(0)	0	0(0)	0	1(3)	1
1ofz	2(5)	2	3(4)	1	1(1)	1	0(0)	0	0(0)	0	0(0)	0
3dd0	2(4)	1	1(1)	0	3(3)	1	0(0)	0	0(0)	0	0(2)	0
3ee4	0(2)	0	0(0)	0	5(5)	7	0(1)	0	0(0)	0	0(0)	0
5m9w	3(11)	3	6(8)	2	2(2)	2	0(2)	1	0(0)	0	0(1)	0
6e9a	2(6)	2	3(3)	1	4(5)	7	0(0)	0	0(0)	0	0(2)	0
TOT	16/17		10/10		29/34		4/6		0/0		2/2	
1akx	1(1)	1	2(4)	1	1(1)	1	0(0)	0	1(1)	1	0(0)	0
1i9v	0(19)	0	5(13)	2	1(1)	0	0(0)	0	0(0)	0	0(0)	0
2esj	2(23)	4	9(15)	3	1(1)	0	0(0)	0	0(0)	0	0(0)	0
4f8u	2(12)	3	8(8)	3	3(3)	0	0(0)	0	0(0)	0	0(0)	0
5bjo	2(4)	1	2(2)	1	4(4)	0	0(0)	0	1(1)	1	1(1)	1
5kx9	1(6)	1	0(0)	0	3(3)	0	0(0)	0	1(1)	0	1(1)	1
6tf3	3(12)	2	2(2)	0	2(2)	0	0(2)	0	0(0)	0	2(2)	2
7oax0	0(4)	0	4(4)	1	3(3)	0	0(0)	0	0(0)	0	0(0)	0
7oax1	2(7)	0	2(2)	0	5(5)	0	0(0)	0	0(0)	0	2(2)	2
8eyv	1(4)	0	2(2)	0	4(4)	0	0(0)	0	1(1)	0	1(1)	1
TOT	9/12		11/11		1/1		0/0		2/2		7/7	

Table 2: Table representing the number of interactions detected by the fields, given by counting the pharmacophores of a given species that overlap with the corresponding interaction field. Each element in the table is in the format *# of overlaps* (*# of total pharmacophores*) | *# of PLIP interactions*. Rows labeled "TOT" indicate the number of PP overlapping with fields that match those predicted by PLIP.

Results presented in the table are therefore to be interpreted with care and are mainly qualitative (also because of the approximations made in the filtering parameters for the fields, the radii for the pharmacophores, and, to some extent, for the cutoff of the initial trimming). However, it is clear that we obtain an overall excellent correspondence between the regions of high field values and the pharmacophores involved in the ligand's interaction with the receptor.

5. Perspectives on large systems

Having assessed the usefulness of computing fields to understand the physical properties of binding pockets, we tested field calculations on whole systems to evaluate the extent to which fields might be useful to characterize the properties of macromolecules more in general.

Membrane proteins hydrophobicity

The Human GABAA receptor alpha1-beta2-gamma2 (PDB ID 6X3V, protein only) is a transmebrane protein containing an ion channel. We investigate in particular the hydrophobic and hydrophilic fields as we expect to detect differences in these fields between the regions that are embedded in the membrane and the parts exposed to the solvent, as well as between the inside of the ion channel and the outside. Our hydrophobic/hydrophilic field is shown in Figure 8 for the whole systems (A) and for slices between planes parallels to the membrane (B). We can observe how the outside of the molecule offers a hydrophobic environment around the transmembrane helices while the head of the molecule is mainly hydrophilic. In particular the entrance of the ion channel is hydrophilic. As we move down the molecule following the channel we can observe that the environment inside the channel remains always hydrophilic (B1-B4), while the external part becomes hydrophobic in correspondence of the transmebrane helices (B2 and B3). The use of fields offers a very clear view of the hydrophobic and hydrophilic regions around the molecule which is hardly achievable by typical

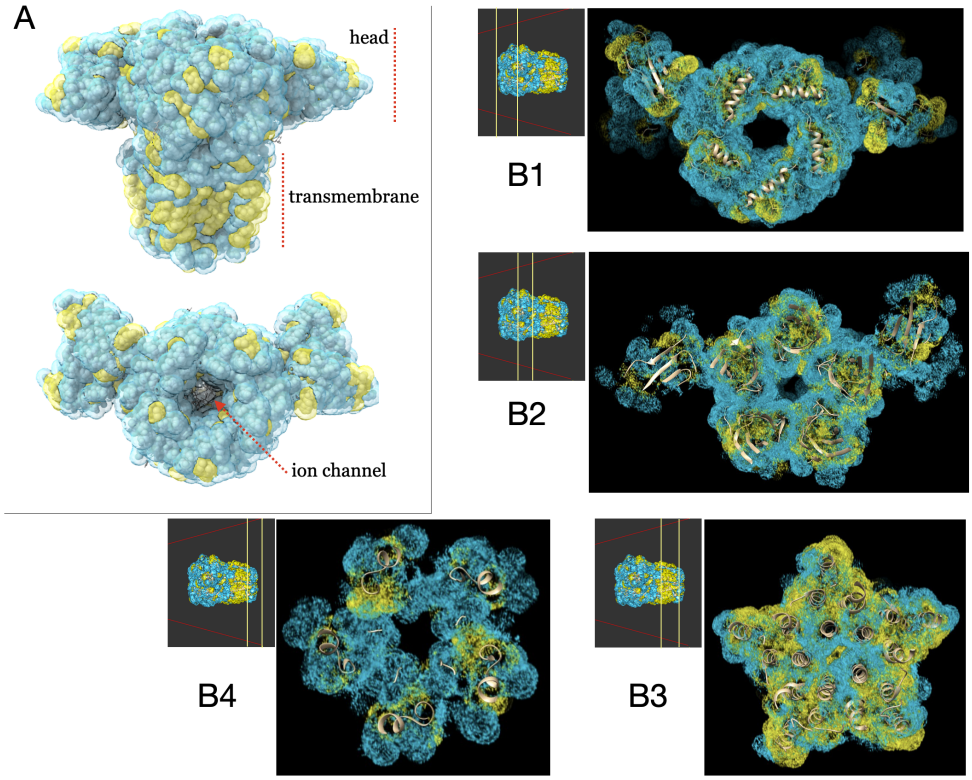

Figure 8: Visualization of the hydrophobic (yellow) and hydrophilic (light blue) fields for the GABAA receptor. A: Visualization of the fields for the whole system (side and top view), B: visualization of the fields for slices as we move across the channel from the head to the transmembrane region as shown in the smaller quadrant (the region shown is in between the two yellow lines).

visualization coloring the surface using a hydrophobic/hydrophilic scale.

RNA binding pockets

Figure 9 shows the results of the fields computed for the whole system for the HIV 2 TAR RNA molecule (PDB ID: 1AKX, RNA only) previously presented. The electrostatic field (9 A) highlights a region of intense negative field in the central section of the molecule, in correspondence with the bulge.

The stacking fields (9 B) and the hydrogen bond donor (9 C) and acceptor (9 D) fields exhibit possible interaction sites all around the molecule. We can notice in particular the stacking fields around the bases sticking out toward the solvent in the loop and in the bulge. Looking carefully, one can notice that at the bulge all fields are of high intensity, but this is hard to see because of the density of the fields all around the molecule. If we are interested in detecting possible binding pockets, a promising strategy could consist of combining field hot-spots with the search for cavities that could accommodate a binding partner. Such a search can be done for example using geometric criteria of cavity detections of software such as *fpocket* [58]).

In Figure 9E we show the result of this strategy on TAR where we can observe the presence of a large binding pocket with high field values in the central bulge. This is indeed the region where the ligand argininamide is found in the experimental structure. Our analyses suggest that other possibly smaller pockets are also available and could be targeted with other ligands, even though in this case they would probably not be of much interest from the biological standpoint since the behavior of this system strongly depends on the accessibility of the bulge. A similar preliminary analysis performed on all the other RNA benchmarks shows that cavities containing fields hot-spots are where the ligands bind. A systematic analysis of this strategy on a large database of RNA/ligand complexes is ongoing to determine if this method is suitable in general to detect RNA binding pockets.

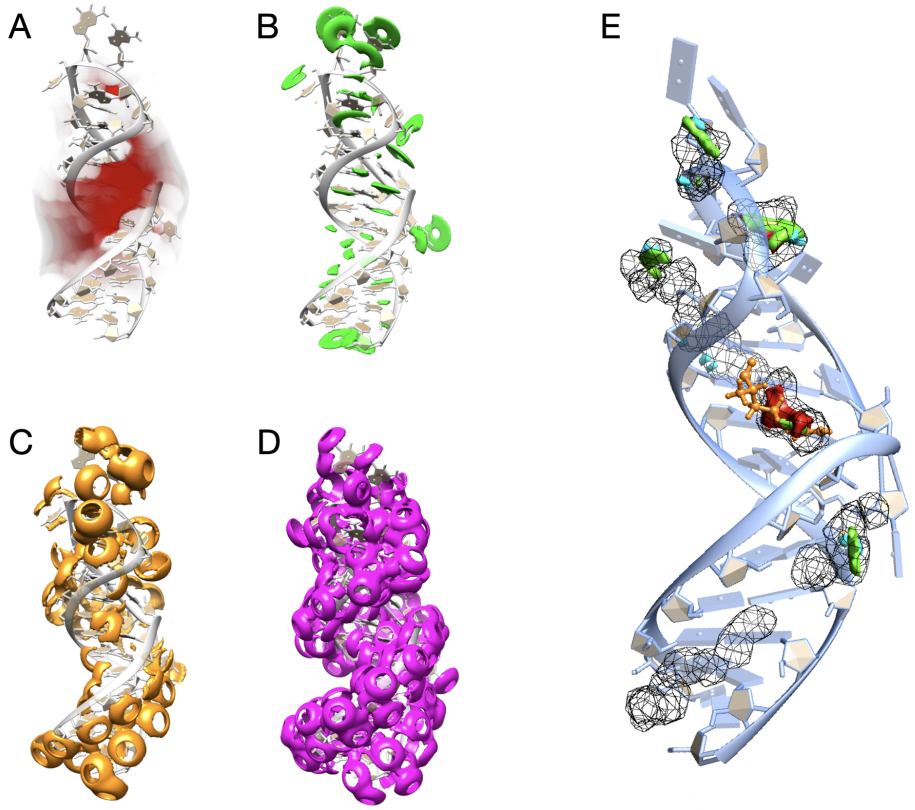

Figure 9: Visualization of fields around a whole RNA molecule (PDB ID 1AKX). A: Electrostatic field, B: stacking, C: hydrogen bond donors, D: hydrogen bond acceptors, E: possible pockets identified by pure geometric considerations by *fpocket* (mesh) combined with hot spots of the fields (colored solid). Argininamide is shown in orange and its position overlaps with the largest pocket highlighted by the fields.

Large systems interfaces

As last application, we illustrate the analysis of the interface of a large complex consisting of the Internal Ribosomal Entry Site (IRES) of Hepatitis C bound to the 40S ribosomal subunit. Cryo-EM experiments show two distinct complexes one in which the IRES is bound but closed on itself (PDB ID 7SYG), and a second one in which the IRES stretches across the ribosomal

subunit (PDB ID 7SYP) in an open form [59]. We want to use our fields to characterize the binding interfaces and gather insights on the binding modes of the two IRES configurations.

To achieve this goal we want to identify what fields are "used" in the binding, *i. e.* find those that actually form an interaction with the corresponding partner in the other molecule. For example, when considering a stacking field, we want to see what regions of high field values match the position of aromatic rings of the other molecule. Similarly, for hydrogen bonds, for the donor field we want to see what positions are matched with an acceptor atom of the binding partner. For the electrostatics, we will compare the field with a charge density field of the binding partner. In practice, to compute what we will call the "interface fields", we define a new class of fields that are just the occupation of the binding partners and we compute the voxel-wise product between the statistical fields of one molecule and the occupation field of the partner. Occupation fields are simply spheres centered around the binding partner elements (HB donor atoms, HB acceptor atoms, aromatic rings, hydrophobic and hydrophilic atoms), with radius 1Å. The only outlier is the charge density, for which we consider normalized gaussians centered around the sites of partial charges with width 1Å and weighted by the value of the charge. These choices were mainly dictated by the visual rendering, since at these level we are not interested in a quantitative characterization of the

interactions. The mathematical definition of the occupation fields follows:

$$\rho(\mathbf{r}) := \sum_{i=1}^{N_a} w_i \, K(|\mathbf{r} - \mathbf{r}_i|) \tag{6}$$

where $K(\mathbf{r})$ is a generic kernel that can be modeled either as a Heaviside step function defined over a sphere or a normalized Gaussian function with $\mu = 0$Å and $\sigma = 1$Å; w_i is 1 for the sphere-like occupation field while it corresponds to the partial charge of the atom i in the charge density case.

We perform this analysis on the open and closed conformations of IRES: results are illustrated in Fig.10 and Fig.11 and in the video provided as supplementary material. Figure 10 provides a detailed analysis of the electrostatic interactions, where both open and closed conformations exhibit regions of attractive (green) and repulsive (orange) electrostatic patches. The isovalue chosen for the figure corresponds to an energy density of $\sim \pm 0.2$ kT/Å^3. This energy density reflects the spatial distribution of the electrostatic potential. Overall, the electrostatics appear to be generally unfavorable, though the open conformation demonstrates a more favorable interaction profile than the closed conformation. Figure 11 focuses on stacking interactions and HBs, where the isovalue should be interpreted in a more binary fashion, indicating the presence or absence of interactions. For the closed conformation, a couple of small anchor sites are observed where stacking and hydrogen bonding interactions are concentrated. In contrast, the open conformation reveals an additional large patch of interactions (located in the top left quadrant in the

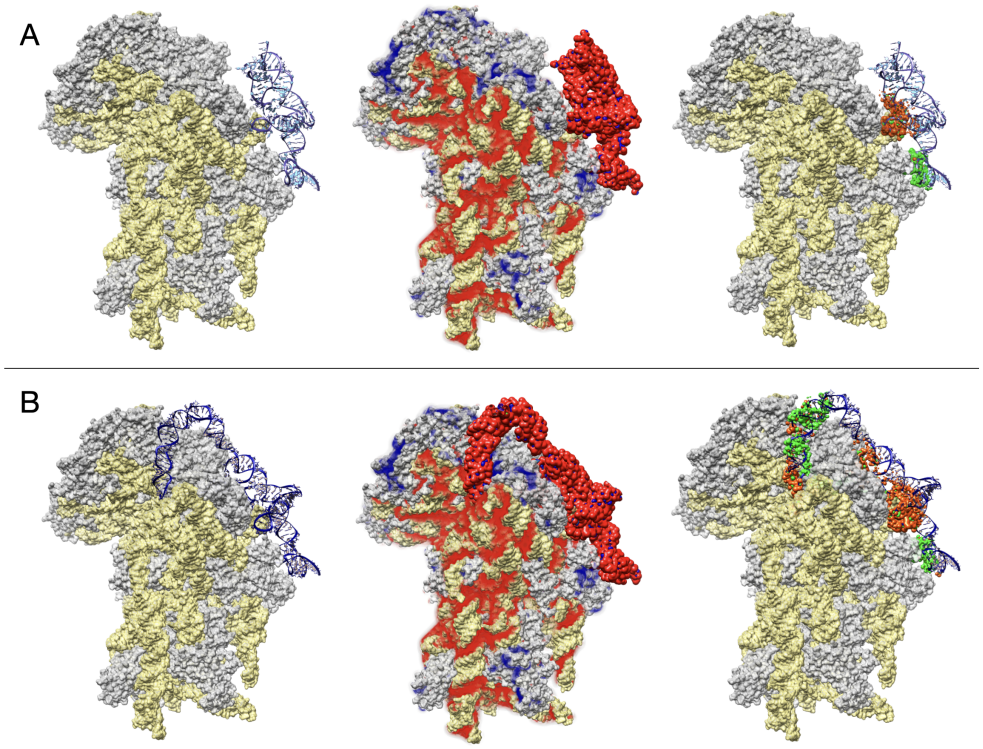

Figure 10: A: Closed IRES configuration, B: open IRES configuration. The ribosomal subunit is shown in gray (protein) and yellow (RNA). Left: structures from the Cryo-EM with the IRES shown in blue. Center: electrostatic field computed for the ribosomal subunit and the charge density computed for the IRES (red: negative field/charges, blue: positive field/charges). Right: results of the product of the electrostatic field by the charge density. We report in green regions of favorable/attractive interactions and in orange the regions of unfavorable/repulsive interactions.

figure), prominently involving HB interactions between the ribosome and the head region of the IRES, thereby highlighting the structural differences in interaction networks between these conformations.

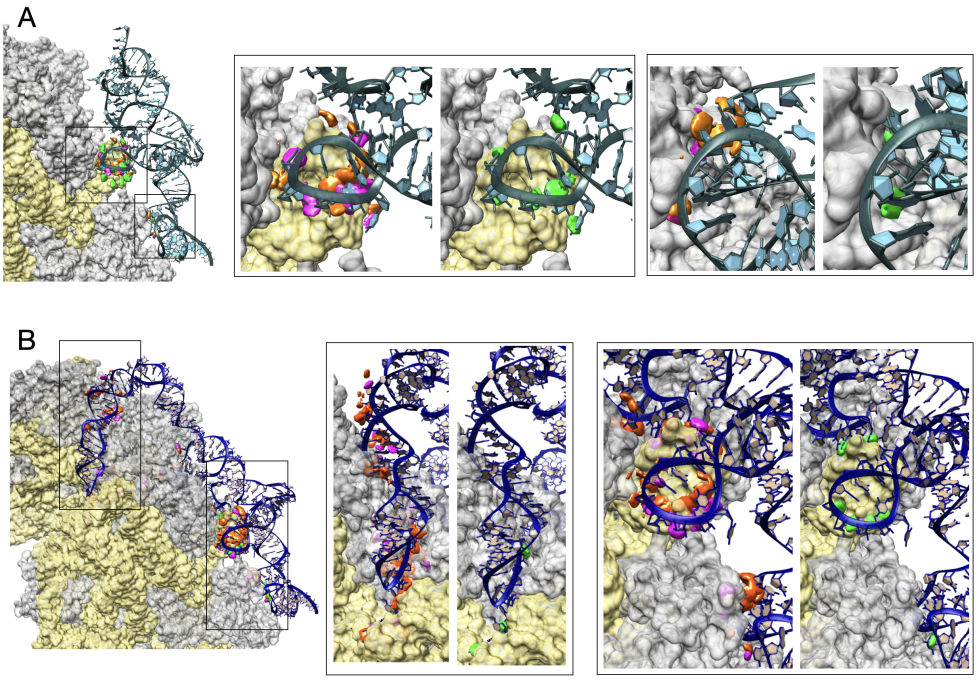

Figure 11: Zoom on the closed (A) and open (B) IRES configuration where we show the results of the stacking (green) and hydrogen bonding (magenta:hba, and orange:hbd) interface fields. The central panel shows the details of the top interface region while the right panel shows the detail of the bottom interface region, shown as black rectangles on the image on the left.

Conclusions

In this work we report our recent development on the calculation and visualization of molecular fields by means of effective potentials based on simple functional forms and parameterized by statistical analysis. Our calculations assume a coarser view than what is usually done for MIFs , where calculations are performed based on detailed atomistic or even quantum force-fields, and don't use any probes. The calculation and visualization of the stacking field is particularly relevant as the interactions of aromatic rings are key for

the binding of RNAs.

We have shown that despite the reduced level of detail of our calculations, our fields are able to capture the core physical features of binding pockets. A benchmark on 20 proteins and RNA systems shows that we obtain a very good agreement between regions of high field values (field hotspots) and the actual placement of known ligands. This agreement is very promising as our tool provides an inexpensive and fast approach to understand and characterize possible binding modes of a pocket that could help in selecting and optimizing binding partners. Our method, grounded in a physical approach, is built on principles that apply equally to both proteins and RNAs, and could help rationalize the main similarities and differences in their interactions with small ligands which, in turns, could allow to better refine computational tools for binding pocket detections and docking for RNAs.

Our approach produces an output that can be promptly visualized with most commonly used visualization software such as Chimera, VMD and UnityMol. For the latter, we provide a script that can also be run in a virtual reality context for in-depth exploration of these fields and their use as guidance in interactive docking of ligands or fragments. Calculations are fast enough that fields for a full single macromolecular structure can be obtained in matter of seconds, allowing to visualize the physical properties of a molecule as a whole. This possibility opens the way to use our fields to rationalize the interactions between macromolecules such as two proteins or a protein and a nucleic acid.

Competing interests

The authors declare no competing interests.

Authors' contributions

D.B.M developed the approach for APBS, stacking and hydrophobic fields, wrote the original code, performed benchmarks, developed the UnityMol extension, wrote the manuscript; G.M. performed benchmarks, performed the comparison with pharmacophores, developed and applied interface fields, wrote the manuscript; A.K. developed the approach for HBD and HBA; A.R. developed the interface fields; L.R. wrote, documented and tested the distributed code and corresponding websites; H.S. contributed to code re-writing; M.B. supervised the code re-writing, testing and visualization experiments, designed the VR scenario; A.T. conceived applications and wrote the manuscript; S.P. conceived and supervised the work, wrote the manuscript.

Funding

This work was partially supported by the French National Research Agency (PIRATE ANR-21-CE45-0014 and MERLIN ANR-22-CE45-0032) and by the "Initiative d'Excellence" program from the French Government (Grant "DYNAMO", ANR-11-LABX-0011 and grant "CACSICE", ANR-11-EQPX-0008).

Data availability

databases are available on zenodo, DOI:

References

[1] R. Santos, O. Ursu, A. Gaulton, A. P. Bento, R. S. Donadi, C. G. Bologa, A. Karlsson, B. Al-Lazikani, A. Hersey, T. I. Oprea, J. P. Overington, A comprehensive map of molecular drug targets, Nature Reviews Drug Discovery 16 (1) (2017) 19–34. doi:10.1038/nrd.2016.230.
URL https://doi.org/10.1038/nrd.2016.230

[2] K. Wu, E. Karapetyan, J. Schloss, J. Vadgama, Y. Wu, Advancements in small molecule drug design: A structural perspective, Drug Discovery Today 28 (10) (2023) 103730. doi:https://doi.org/10.1016/j.drudis.2023.103730.
URL https://www.sciencedirect.com/science/article/pii/S1359644623002465

[3] A. Yu, Y. Choi, M. Tu, Rna drugs and rna targets for small molecules: Principles, progress, and challenges, Pharmacol Rev. 72 (4) (2020) 862–898. doi:doi:10.1124/pr.120.019554.

[4] A. Yu, C. Jian, A. Yu, M. Tu, Rna therapy: Are we using the right molecules?, Pharmacol Ther. 196 (2019) 91–104. doi:doi:10.1016/j.pharmthera.2018.11.011.

[5] S. Hegde, et al., Small molecule–rna targeting: starting with the fundamentals, Chemical Communications 56 (94) (2020) 14744–14756. doi:10.1039/D0CC04680A.

[6] S. Hegde, et al., Inhibition of SARS-CoV-2 by targeting conserved viral RNA structures and sequences, Frontiers in Chemistry 9 (2021) 802766. doi:10.3389/fchem.2021.802766.

[7] D. Pala, D. E. Clark, Caught between a rock and a hard place: current challenges in structure-based drug design, Drug Discovery Today 29 (9) (2024) 104106. doi:https://doi.org/10.1016/j.drudis.2024.104106.
URL https://www.sciencedirect.com/science/article/pii/S1359644624002319

[8] A. Hussein, C. Geneix, M. Petitjean, A. Borrel, D. Flatters, A.-C. Camproux, Global vision of druggability issues: applications and perspectives, Drug Discovery Today 22 (2) (2017) 404–415. doi:10.1016/j.drudis.2016.11.020.

[9] L. Zheng, Z. Wang, K. M. Merz, Optimol: Optimization of binding affinities in chemical space for drug discovery, Journal of Chemical Information and Modeling 60 (10) (2020) 4291–4300. doi:10.1021/acs.jcim.0c00833.

[10] K. Wang, J. Zhang, Y. Chen, et al., Improving the generalizability of protein-ligand binding predictions with ai-bind, Nature Communications 14 (2023) 1–14. doi:10.1038/s41467-023-37572-z.
URL https://www.nature.com/articles/s41467-023-37572-z

[11] F. Sabanés Zariquiey, R. Galvelis, E. Gallicchio, J. D. Chodera, T. E. Markland, G. De Fabritiis, Enhancing protein–ligand binding affinity predictions using neural network potentials, Journal of Chemical Information and Modeling 64 (5) (2024) 1481–1485. doi:10.1021/acs.jcim.3c02031.

[12] S. Mallapaty, The pathogens that could spark the next pandemic, Nature 632 (2024) 488. doi:https://doi.org/10.1016/j.drudis.2023.103730.

[13] Y. Zhou, Y. Kong, W. Fan, T. Tao, Q. Xiao, N. Li, X. Zhu, Principles of rna methylation and their implications for biology and medicine, Biomedicine & Pharmacotherapy 131 (2020) 110731. doi:https://doi.org/10.1016/j.biopha.2020.110731.
URL https://www.sciencedirect.com/science/article/pii/S0753332220309240

[14] X. Jiang, B. Liu, Z. Nie, L. Duan, Q. Xiong, Z. Jin, C. Yang, Y. Chen, The role of m6a modification in the biological functions and diseases, Signal transduction and targeted therapy 6 (1) (2021) 74.

[15] H. Sato, T. Hara, C. Otsuka, Y. Arao, Y. Tsuji, Y. Hamano, M. Ogita, E. di Luccio, T. Hirotsu, A. Vecchione, H. Ishii, m6 RNA methylation: an emerging common target in the immune response to cancer and severe acute respiratory syndrome-coronavirus-2 infection, Exploration

of Medicine 4 (1) (2023) 107–114. doi:10.37349/emed.2023.00127. URL https://www.explorationpub.com/Journals/em/Article/1001127

[16] T. Hermann, Small molecules targeting viral rnas, Wiley Interdisciplinary Reviews: RNA 7 (6) (2016) 726–743. doi:10.1002/wrna.1375.

[17] B. S. Morgan, B. Forte, A. E. Hargrove, Discovery of key physicochemical, structural, and spatial properties of rna-target bioactive ligands, Journal of Medicinal Chemistry 60 (3) (2017) 428–436. doi:10.1021/acs.jmedchem.6b01649.

[18] Y. Zhou, Y. Jiang, S.-J. Chen, Rna–ligand molecular docking: Advances and challenges, Wiley Interdisciplinary Reviews: Computational Molecular Science 12 (3) (2022) e1571.

[19] P. J. Goodford, A computational procedure for determining energetically favorable binding sites on biologically important macromolecules, Journal of medicinal chemistry 28 (7) (1985) 849–857.

[20] R. C. Wade, K. J. Clark, P. J. Goodford, Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 1. ligand probe groups with the ability to form two hydrogen bonds, Journal of Medicinal Chemistry 36 (1) (1993) 140–147.

[21] Materials Science Suite, Schrödinger release 2024-2 (2024).
URL https://www.schrodinger.com/

[22] S. Tortorella, E. Carosati, G. Sorbi, G. Bocci, S. Cross, G. Cruciani, L. Storchi, Combining machine learning and quantum mechanics yields more chemically aware molecular descriptors for medicinal chemistry applications, Journal of Computational Chemistry 42 (29) (2021) 2068–2078.

[23] P. Tosco, M. D. Mackey, in: Lessons and Successes in the Use of Molecular Fields, 2017. [link].
URL https://api.semanticscholar.org/CorpusID:99445514

[24] G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, A. J. Olson, Autodock4 and autodocktools4: Automated docking with selective receptor flexibility, Journal of computational chemistry 30 (16) (2009) 2785–2791.

[25] J. Contreras-García, E. R. Johnson, S. Keinan, R. Chaudret, J.-P. Piquemal, D. N. Beratan, W. Yang, Nciplot: a program for plotting noncovalent interaction regions, Journal of chemical theory and computation 7 (3) (2011) 625–632.

[26] F. Sterpone, S. Melchionna, P. Tuffery, S. Pasquali, N. Mousseau, T. Cragnolini, Y. Chebaro, J.-F. St-Pierre, M. Kalimeri, A. Barducci, et al., The opep protein model: from single molecules, amyloid for-

mation, crowding and hydrodynamics to dna/rna systems, Chemical Society Reviews 43 (13) (2014) 4871–4893.

[27] T. Cragnolini, Y. Laurin, P. Derreumaux, S. Pasquali, Coarse-grained hire-rna model for ab initio rna folding beyond simple molecules, including noncanonical and multiple base pairings, Journal of chemical theory and computation 11 (7) (2015) 3510–3522.

[28] S. Doutreligne, C. Gageat, T. Cragnolini, A. Taly, S. Pasquali, P. Derreumaux, M. Baaden, Unitymol: interactive and ludic visual manipulation of coarse-grained rna and other biomolecules, in: 2015 IEEE 1st international workshop on virtual and augmented reality for molecular science (VARMS@ IEEEVR), IEEE, 2015, pp. 1–6.

[29] S. Pasquali, Molecular modelling as the spark for active learning approaches for interdisciplinary biology teaching, Interface Focus 9 (3) (2019) 20180065.

[30] A. Taly, F. Nitti, M. Baaden, S. Pasquali, Molecular modelling as the spark for active learning approaches for interdisciplinary biology teaching, Interface Focus 9 (3) (2019) 20180065.

[31] Z. Lv, A. Tek, F. Da Silva, C. Empereur-Mot, M. Chavent, M. Baaden, Game on, science-how video game technology may help biologists tackle visualization challenges, PloS one 8 (3) (2013) e57990.

[32] S. Doutreligne, T. Cragnolini, S. Pasquali, P. Derreumaux, M. Baaden, Unitymol: interactive scientific visualization for integrative biology, in: 2014 IEEE 4th Symposium on Large Data Analysis and Visualization (LDAV), IEEE, 2014, pp. 109–110.

[33] D. Kut'ák, P.-p. Vázquez, T. Isenberg, M. Krone, M. Baaden, J. Byška, B. Kozlíková, H. Miao, State of the art of molecular visualization in immersive virtual environments, in: Computer Graphics Forum, Vol. 42, Wiley Online Library, 2023, p. e14738.

[34] E. Jurrus, D. Engel, K. Star, K. Monson, J. Brandi, L. E. Felberg, D. H. Brookes, L. Wilson, J. Chen, K. Liles, et al., Improvements to the apbs biomolecular solvation software suite, Protein Science 27 (1) (2018) 112–128.

[35] J. Siódmiak, P. Bełdowski, W. Augé, D. Ledzinski, S. Śmigiel, A. Gadomski, Molecular dynamic analysis of hyaluronic acid and phospholipid interaction in tribological surgical adjuvant design for osteoarthritis, Molecules 22 (2017) 1436. doi:10.3390/molecules22091436.

[36] D. Xu, C.-J. Tsai, R. Nussinov, Hydrogen bonds and salt bridges across protein-protein interfaces., Protein engineering 10 9 (1997) 999–1012. URL https://api.semanticscholar.org/CorpusID:3206919

[37] M. Elias, A. Wellner, K. Goldin-Azulay, E. Chabriere, J. A. Vorholt,

T. J. Erb, D. S. Tawfik, The molecular basis of phosphate discrimination in arsenate-rich environments, Nature 491 (7422) (2012) 134–137. doi:10.1038/nature11517.
URL https://www.nature.com/articles/nature11517

[38] O. N. Rogacheva, S. A. Izmailov, L. V. Slipchenko, N. R. Skrynnikov, A new structural arrangement in proteins involving lysine NH3 + group and carbonyl, Scientific Reports 7 (1) (2017) 16402. doi:10.1038/s41598-017-16584-y.
URL https://www.nature.com/articles/s41598-017-16584-y

[39] W. C. Wimley, S. H. White, Experimentally determined hydrophobicity scale for proteins at membrane interfaces, Nature structural biology 3 (10) (1996) 842–848.

[40] P. Shih, L. G. Pedersen, P. R. Gibbs, R. Wolfenden, Hydrophobicities of the nucleic acid bases: distribution coefficients from water to cyclohexane, Journal of molecular biology 280 (3) (1998) 421–430.

[41] M. Holst, F. Saied, Multigrid solution of the poisson—boltzmann equation, Journal of computational chemistry 14 (1) (1993) 105–113.

[42] R. Tarjan, Depth-first search and linear graph algorithms, SIAM Journal on Computing 1 (2) (1972) 146–160. arXiv:https://doi.org/10.1137/0201010, doi:10.1137/0201010.
URL https://doi.org/10.1137/0201010

[43] J. Sunseri, D. R. Koes, Pharmit: interactive exploration of chemical space, Nucleic acids research 44 (2016) W442–W448.

[44] M. F. Adasme, K. L. Linnemann, S. N. Bolz, F. Kaiser, S. Salentin, V. Haupt, M. Schroeder, PLIP 2021: expanding the scope of the protein–ligand interaction profiler to DNA and RNA, Nucleic Acids Research 49 (2021) W530–W534.

[45] F. Schmidt, S. Donahoe, K. Hagens, J. Mattow, U. E. Schaible, S. H. Kaufmann, R. Aebersold, P. R. Jungblut, Complementary analysis of the mycobacterium tuberculosis proteome by two-dimensional electrophoresis and isotope-coded affinity tag technology, Molecular & cellular proteomics 3 (1) (2004) 24–42.

[46] C. S. Andersson, M. Högbom, A mycobacterium tuberculosis ligand-binding mn/fe protein reveals a new cofactor in a remodeled r2-protein scaffold, Proceedings of the National Academy of Sciences 106 (14) (2009) 5633–5638.

[47] H. Shimizu, S.-Y. Park, D.-S. Lee, H. Shoun, Y. Shiro, Crystal structures of cytochrome p450nor and its mutants (ser286→ val, thr) in the ferric resting state at cryogenic temperature: a comparative analysis with monooxygenase cytochrome p450s, Journal of Inorganic Biochemistry 81 (3) (2000) 191–205.

[48] T. Kudo, N. Takaya, S.-Y. Park, Y. Shiro, H. Shoun, A positively

charged cluster formed in the heme-distal pocket of cytochrome p450nor is essential for interaction with nadh*, Journal of Biological Chemistry 276 (2001) 5020–5026.

[49] H. A. Mouse, H. Atlas, Handbook of proteolytic enzymes, volumes 1, 2 and 3. (2013).

[50] A. K. Ghosh, J. N. Williams, R. Y. Ho, H. M. Simpson, S.-i. Hattori, H. Hayashi, J. Agniswamy, Y.-F. Wang, I. T. Weber, H. Mitsuya, Design and synthesis of potent hiv-1 protease inhibitors containing bicyclic oxazolidinone scaffold as the p2 ligands: Structure–activity studies and biological and x-ray structural studies, Journal of medicinal chemistry 61 (21) (2018) 9722–9737.

[51] A. D. Frankel, Activation of hiv transcription by tat., Current opinion in genetics & development 2 (2) (1992) 293–298.

[52] K. A. Jones, B. M. Peterlin, Control of rna initiation and elongation at the hiv-1 promoter, Annual review of biochemistry 63 (1) (1994) 717–743.

[53] K. M. Weeks, D. M. Crothers, Rna recognition by tat-derived peptides: interaction in the major groove?, Cell 66 (3) (1991) 577–588.

[54] M. J. Churcher, C. Lamont, F. Hamy, C. Dingwall, S. M. Green, A. D. Lowe, P. J. G. Butler, M. J. Gait, J. Karn, High affinity binding of tar rna by the human immunodeficiency virus type-1 tat protein requires

base-pairs in the rna stem and amino acid residues flanking the basic region, Journal of molecular biology 230 (1) (1993) 90–110.

[55] A. S. Brodsky, J. R. Williamson, Solution structure of the hiv-2 tar-argininamide complex, Journal of molecular biology 267 (3) (1997) 624–639.

[56] J. Kondo, M. Hainrichson, I. Nudelman, D. Shallom-Shezifi, C. M. Barbieri, D. S. Pilch, E. Westhof, T. Baasov, Differential selectivity of natural and synthetic aminoglycosides towards the eukaryotic and prokaryotic decoding a sites, ChemBioChem 8 (14) (2007) 1700–1709.

[57] H. Pundir, T. Joshi, M. Pant, S. Bhat, J. Pandey, S. Chandra, S. Tamta, Identification of sars-cov-2 rna dependent rna polymerase inhibitors using pharmacophore modelling, molecular docking and molecular dynamics simulation approaches, Journal of Biomolecular Structure and Dynamics 40 (24) (2022) 13366–13377.

[58] V. Le Guilloux, P. Schmidtke, P. Tuffery, Fpocket: an open source platform for ligand pocket detection, BMC bioinformatics 10 (2009) 1–11.

[59]