Color Coding for the Fragment-Based Docking, Design and Equilibrium Statistics of Protein-Binding ssRNAs

Taher Yacoub, Roy González-Alemán, Fabrice Leclerc, Isaure Chauvot de Beauchêne, Yann Ponty

Color Coding for the Fragment-Based Docking, Design and Equilibrium Statistics of Protein-Binding ssRNAs

Taher Yacoub[1,2], Roy González-Alemán[2,3][0000−0003−3852−4902], Fabrice Leclerc[2][0000−0002−5641−1525], Isaure Chauvot de Beauchêne[4][0000−0002−7035−3042], and Yann Ponty[1][0000−0002−7615−3930]

[1] LIX (CNRS UMR 7161), Institut Polytechnique de Paris, Palaiseau, France
[2] Institute for Integrative Biology of the Cell (I2BC – CNRS UMR 9198), Université Paris Saclay, Gif, France
[3] Laboratorio de Química Computacional y Teórica (LQCT), Universidad de La Habana, Cuba
[4] LORIA (CNRS UMR 7503), Université de Lorraine, Nancy, France

Abstract. We revisit the fragment-based docking and design of single-stranded RNA aptamers (ssRNAs), consisting of k nucleotides, onto a rigid protein. Fragments, representing either one or multiple pieced nucleotides, are individually docked onto the protein surface using a force field, and some among the resulting n poses are pieced together to form a conformation compatible with the input ssRNA sequence. Relaxing the sequence compatibility constraint, a similar methodology can be used to design ssRNAs that preferentially bind a protein of interest, possibly targeting a pocket. However, a brute-force enumeration of clash-free conformations quickly becomes prohibitive due to their superexponential ($\Theta(n^k)$ worst-case) combinatorial explosion, hindering the potential of fragment-based methods towards docking and design.

We adopt the color-coding technique, introduced by Alon, Yuster and Zwick, to optimize over self-avoiding fragment assemblies in time/space linear on n the number of poses, and in time only exponential on k the number of fragments. The dynamic programming algorithm at the core of our method is surprisingly simple, and can be extended to produce suboptimal candidates, or modified to perform Boltzmann sampling of candidates assemblies. Using a rejection principle, and further optimized by a clique decomposition of clashing poses, these algorithms can be leveraged into efficient algorithms optimizing over clash-free complexes. The resulting sampling procedure can further be adapted into statistically-consistent estimators for any computable feature of interest.

We showcase some of the capabilities of this new framework by reanalyzing a set of 7 documented ssRNA-protein complexes, demonstrating its practical relevance and versatility.

Keywords: Fragment-based docking, RNA design, RNA-protein interaction, Parameterized complexity algorithms

1 Introduction

Fragment-based Design is a powerful strategy, used both in academia and pharmaceutical industry to develop potent compounds from fragment. Five drugs designed with this approach were approved by the FDA (Bollag et al., 2012; Tap et al., 2015; Perera et al., 2017; Schoepfer et al., 2018), one of which as recently as 2021 (Souers et al., 2013). Fragments are usually small compounds with low molecular weight, having about 20 heavy atoms (Kirsch et al., 2019; Schuffenhauer et al., 2005). The principle of this strategy is to dock a library of fragments on a receptor and to select those specifically binding the target. One or several initial poses are then extended to form a complete chemical compound.

While this strategy is generally applied to chemical compounds for the design, fragment-based approach has been utilized to predict complexes formed by ssRNAs of known sequence with RBP (RNA-Binding Protein) (Hall et al., 2015; de Beauchene et al., 2016; Kappel and Das, 2019). To predict the interaction of RNA-RBP complexes, fragments libraries must embrace a large diversity of RNA fragments containing for instance chemical modifications for the design. Such a diversity is also crucial towards a fragment-based design of therapeutic molecules, most of which require a good affinity towards the target to achieve the desired activity (e.g. antagonist, agonist).

For instance, a recent approach (González-Alemán et al., 2021) initially performs a sampling of mononucleotides with Multiple Copy Simultaneously Search (MCSS). From a library of mononucleotides, the principle of MCSS is to dock randomly copies of each nucleotide to obtain a set of fragment poses docked on the target with known orientation and position. However, the assembly of consecutive nucleotides into an optimal oligonucleotide, either of known (ssRNA docking) or unknown (ssRNA design) sequence, cannot reasonably be performed through brute force due to punishing combinatorics. Indeed, the success of the fragment-based approach hinges critically on a sufficient density of poses which, in turn, greatly impacts the number of candidate positions/sequences. The problem is made even worse by a consideration of modified nucleotides, increasing the basis of an exponential growth.

In this work, we revisit the fragment-based docking and design through the prism of color coding, an algorithmic technique introduced by Alon, Yuster and Zwick (Alon et al., 1994) that allows to capture a necessary notion of ssRNA self avoidance. This elegant technique initially addressed the problem of finding sparse motifs in graphs, and has been utilized in the context Bioinformatics for searching (Dost et al., 2008; Shlomi et al., 2006) and counting occurrences of motifs in biological networks (Alon et al., 2008). In Section 2, we show how to adapt color coding, further optimized by a clique decomposition, to obtain exact or probabilistic algorithms for fragment-based docking through energy minimization. Section 2.3 describes how to perform design by relaxing the requirement of being compatible with a given nucleotide sequence. The framework is further extended to produce equilibrium statistics for virtually any feature of interest. Section 3 illustrates the proposed algorithms in the context of 6 RNA binding proteins, and Section 4 discusses some limitations of the approach, and future extensions.

2 Method and algorithms

Let us make a few assumptions explicit: Firstly, our docking is assumed to be rigid on the protein level, so that ssRNA fragments (nucleotides or k-mers) can be individually docked onto the protein without overly losing precision; Secondly, we assume that the length or nucleotide composition of the input ssRNA forbids the adoption of secondary structure elements; Thirdly, the ssRNA/protein system can be assumed to be at the thermodynamic equilibrium so that minimizing the free-energy coincides with maximizing the probability of the joint configuration. Under these assumptions, the fragment-based docking and design of ssRNAs interacting with a protein can both be reformulated as graph problems.

Definitions and notations. Namely, we denote as **fragment** f a nucleotide $r(f)$ in an RNA sequence r, associated with a reference 3D conformation. A **fragment pose**, or pose x, represents a fragment docked onto the protein surface, and is defined by the 3D position of its atoms relative to the protein. An ordered pair of poses is said to be **compatible** if their spatial occupancies do not induce unresolvable geometric clashes, and enables the sequential connection of the two fragments into a longer RNA. Compatibility is an oriented relation (associated with the polarity of RNA), whose assessment is a problem in its own right, and is the object of specialized tools such as MolPy (Chevrollier, 2019) or Nuclear (to be released) to deal with short ssRNAs.

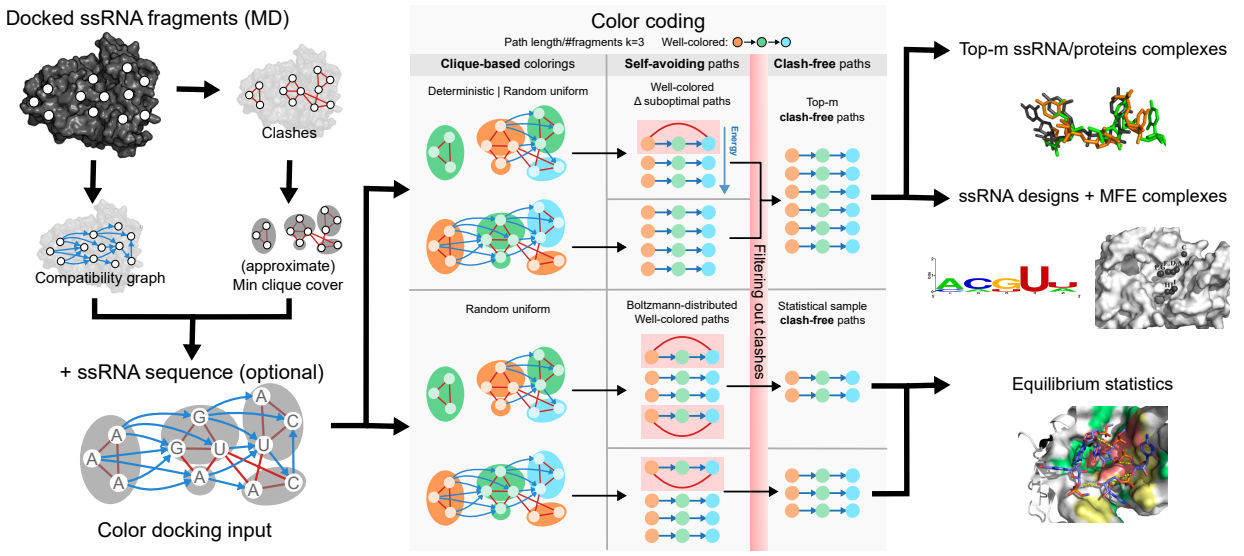

Fig. 1: General workflow: Starting from a graph of docked fragments, for which pairwise connectivity has been assessed by an external tool, our method considers various – random or deterministic – colorings from which solutions to hard computational problems can be obtained.

Through a systematic docking of fragments, *e.g.* using constrained molecular dynamics, onto the surface of the target protein, followed by an evaluation of the connectivity of resulting fragment poses, one obtains a **poses connectivity graph**. It is defined as a directed graph, *i.e* a pair $G = (V, E)$ where V is a set of fragment poses, and any directed edge $(v, v') \in E$ implies the possibility to connect v and v'. In the following, we denote by $n := |V|$ the number of poses in the graph. Any path of the connectivity graph can be associated with a joint ssRNA/protein conformation, called **complex** in the following.

Next, we associate a notion of **free-energy** $\Delta G(x)$ to any complex $x = (v_1, \ldots, v_k)$, defined as:

$$\Delta G(x) = \sum_{i=1}^{k} \delta(v_i) + \sum_{i=1}^{k-1} \delta'(v_i, v_{i+1})$$

where $\delta : V \to \mathbb{R} \cup \{+\infty\}$ and $\delta' : V \times V \to \mathbb{R} \cup \{+\infty\}$ are terms, specific to a fragment docking procedure, which capture the contributions of individual and pairwise-connected fragments respectively.

However some pairs of fragments may **clash**, occupying overlapping or overly proximal geometric regions in the 3D space, leading some of the paths of the connectivity graph do not always represent promising candidates. Trivial instances of such a clash occur within complexes that reuse the same pose twice. Beyond such a simple cases, pairwise clashes can be modeled using a **clash function** $C : E \times E \to \{\mathsf{True}, \mathsf{False}\}$. A path x of length k is **self-avoiding**, also named a k**-path**, iff it nodes are pairwise distinct. A path is **clash-free** if and only if its nodes are pairwise non-clashing, *i.e.* $\forall 1 \leq i < j \leq k, C(x_i, x_j) = \mathsf{False}$. Note that clash-avoidance induces self-avoidance as long as, for each pose v, one has $C(v, v) = \mathsf{True}$.

Problem statement and complexity aspects. Assuming thermodynamic equilibrium, the most stable/probable complex, for a given nucleotides sequence r of length k, is the one having Minimum Free-Energy. Moreover, fragment assembly should be restricted to complexes compatible with the sequence. The computation of such a complex can be restated as follows:

MFEdock **problem**
Input: Pose connectivity graph $G = (V, E)$; Clash function $C : E \times E \to \{\mathsf{True}, \mathsf{False}\}$; Energy function ΔG; Residue sequence $r = r_1, r_2 \ldots r_k$.

Output: Complex $x^\star = (v_1^\star, v_2^\star, \ldots, v_k^\star)$ minimizing free-energy

$$x^\star = \underset{\substack{x=(v_1,\ldots,v_k) \text{ such that} \\ v_i \neq v_j \forall i \neq j, \quad \leftarrow \text{self avoidance} \\ C(v_i, v_{i+1}) = \mathsf{False}, \forall i \quad \leftarrow \text{clash-free} \\ \text{and } r(v_i) = r_i, \forall i \quad \leftarrow \text{nucleotides sequence compatibility}}}{\operatorname{argmin}} \Delta G(x)$$

Computational complexity-wise, for a general input graph, trivial sequence c and unit-valued energy function ($\delta(\cdot) = -1$, $\delta'(\cdot,\cdot) = 0$), MFEdock solves the problem of deciding the existence of a Hamiltonian path in G, implying NP-hardness. The problem remains robustly intractable even when restricted to subclasses of input graphs that can be drawn on a protein surface, such as grid graphs (Itai et al., 1982). Moreover, for the complete compatibility graph and unit-valued energy, solving MFEdock answers the existence of a k-set of non-clashing nodes in the graph $(V, \{v, v' \mid C(v, v') = \mathsf{True}\})$, thus solving the Max Independent Set (MIS) problem. However, MIS is not only NP-hard, but also remains intractable (W[1] hard for k) from the perspective of parameterized complexity on geometric instances, *e.g.* graphs stemming from intersections of segments/discs (Marx, 2006). Taken together, these results indicate a robust computational hardness of the problem, motivating the exploration of alternatives and heuristics.

2.1 Ensuring self-avoidance through color coding

Given the dire complexity status of MFEdock, we initially address a restricted version of the problem that only considers clashes resulting from the reuse of certain poses. In other words, we optimize the energy optimization over self-avoiding paths, which is equivalent to setting $C(\cdot,\cdot) = \mathsf{False}$. In this setting, the algorithmic problem remains NP-hard but simplifies into a, practically solvable, Fixed-Parameter Tractable (FPT) problem for the path length k, using the color coding technique (Alon et al., 1994).

Classic color coding. The key principle of **color coding** is to associate a coloring $\kappa : V \to [1, k]$ to the input graph $G = (V, E)$, and to replace the (hard) search for a path (or motif) of length k (k-path) with the (easier) search of a **colorful path**, using each of the k colors exactly once. Colorful paths can be optimized for, and counted, in time that linear on $n + |E|$, and only exponential on k. For a single coloring, the set of colorful paths is only a subset of k-paths, the optimal k-path may be overlooked. One may then use **derandomization** to turn this approach into an efficient, deterministic and exact algorithm. To that purpose, one needs to construct a family of colorings which, taken as a whole, represents every possible k-path. Naor et al. (1995) propose an explicit construct for such a family, consisting of $e^k k^{\mathcal{O}(\log k)} \log n$ colorings. Iterating the search for optimal colorful paths over the family yields an exact algorithm in overall time $\mathcal{O}((2e)^k k^{\mathcal{O}(\log k)}(n + |E|) \log n)$, critically using $\mathcal{O}(2^k n)$ memory.

Well-colored path as a memory-frugal alternative. To work around this substantial memory requirement, we instead consider a variant of color coding based on **well-colored paths**. A well-colored path is a k-path whose colors in κ are not only distinct, but occur in a specific order, assumed to be $1 \to 2 \to \cdots \to k$ without loss of generality. For the sake of simplicity, we say that a coloring κ **hits** (resp. **misses**) a k-path x when x is well-colored (resp. not well-colored) by x. For any given coloring, the optimal/MFE k-path can be obtained in $\mathcal{O}(k.(n + |E|))$ time, *e.g.* using simple dynamic programming. To be well-colored only constrains the colors assigned to the k nodes of x, leaving only one possiblility out of the k^k possible colorings, so the odds of a random uniform coloring hitting x is simply $\mathbb{P}(x \text{ well colored}) = 1/k^k$. Iterating over α independently-draw random colorings $\kappa_1, \ldots, \kappa_\alpha$, the probability of a k-path being missed by all colorings is then

$$\mathbb{P}(x \text{ missed by } \alpha \text{ random colorings}) = \left(1 - \frac{1}{k^k}\right)^\alpha. \tag{1}$$

This property holds for any k-path, including the MFE path $x^\star$. Consequently, for any targeted **tolerance** $\varepsilon \in (0, 1)$, it suffices to set

$$\alpha := \left\lceil \frac{\log \varepsilon}{\log\left(1 - \frac{1}{k^k}\right)} \right\rceil \in \Theta(k^k \log \varepsilon)$$

and we obtain a probabilistic algorithm that returns $x^\star$ with probability $1-\varepsilon$, and runs in total time $\mathcal{O}(k^{k+2}\,\log\varepsilon\,(n+|E|))$ and memory linear in both k and n. Derandomization can also be used in the context of well-colored paths. Here, the constructs of Alon et al. (1995), coupled with the earlier results of Schmidt and Siegel (1990), provide a family of $k^{\mathcal{O}(k)}\log n$ colorings that hits every k-path, thus implying an exact deterministic algorithm for MFEdock w/o clash constraints. Its complexity is now in $\mathcal{O}(k^{\mathcal{O}(k)}n.\log n)$ for time, marginally higher than for colorful paths, while using a, much reduced, linear memory.

Rejecting from suboptimals to produce the clash-free MFE complex. In order to recover the MFE clash-free complex, and thus provide a solution for MFEdock, we elicit to extract it from the list of Δ **(self-avoiding) suboptimals**, defined as having energy distance at most Δ from the self-avoiding MFE. The list of Δ-suboptimals can be produced using an adapted version of the Waterman/Byers scheme (Waterman and Byers, 1985). It starts by computing the well-colored MFE for a given coloring κ using the following dynamic programming scheme:

$$\mathrm{mfe}_\kappa = \min_{\substack{v\in V\\ \text{such that } r(v)=r_1}} \mathrm{mfe}_\kappa[v,1] \tag{2}$$

$$\mathrm{mfe}_\kappa[v,m] = \begin{cases} E(v) & [\text{if } m=k] \\ \min\limits_{\substack{(v,v')\in E \text{ s.t.}\\ \kappa(v')=m+1\\ \text{and } r(v')=r_{m+1}}} \delta(v)+\delta'(v,v')+\mathrm{mfe}_\kappa[v',m+1] & [\text{otherwise}] \end{cases} \tag{3}$$

Once the mfe matrix computed in $\mathcal{O}(k.(n+|E|))$ time, the exhaustive list of Δ-subopts can be obtained using a modified backtrack:

$$\mathrm{subopts}_\kappa(\Delta) \to \bigcup_{\substack{v\in V \text{ s.t}\\ r(v)=r_1}} \mathrm{subopts}_\kappa(v,1;\Delta') \tag{4}$$
$$[\text{if } \Delta' := \Delta - (\mathrm{mfe}_\kappa[v,1]-\mathrm{mfe}_\kappa) \geq 0]$$

$$\mathrm{subopts}_\kappa(v,m;\Delta) \to \begin{cases} \{v\}\ [\text{if } m=k] \\ \bigcup\limits_{\substack{(v,v')\in E \text{ s.t}\\ \kappa(v')=m+1\\ \text{and } r(v')=r_{m+1}}} \{v\}\otimes \mathrm{subopts}_\kappa(v',m+1;\Delta') & [\text{if } m<k \text{ and } \Delta'\geq 0] \end{cases} \tag{5}$$

Such a backtrack essentially runs in time and memory in $\Theta(k\,D)$, where D is the total number of Δ-subopts, and is expected to grow exponentially with Δ.

An exact, exponential-time in the worst-case, algorithm for the clash-free MFE then starts by computing the global self-avoiding MFE $E^\star_{\mathrm{SA}}$, using a derandomizing family $\boldsymbol{\kappa} := (\kappa_i)_i$ of colorings. It then iterates again several times over the whole family using increasing values of Δ until

$$\Delta \geq \Delta_{\max} := E^-_{\text{clash-free}} - E^\star_{\mathrm{SA}},$$

where $E^-_{\text{clash-free}}$ denotes the clash-free MFE observed for Δ-suboptimals over $\boldsymbol{\kappa}$ so far. At this point, the algorithm may simply return the clash-free MFE complex within the $\Delta_{\max}$ subopts, *i.e.* the structure $S^\star$ achieving $E^-_{\text{clash-free}}$, since this structure is then the clash-free MFE, and a valid solution to MFEdock.

Indeed, for any clash-free complex $S' \neq S^\star$, if S' is found in the combined list of Δ-suboptimals, then it has higher energy than $E^-_{\text{clash-free}}$ by definition. If S' is not listed as a $\Delta_{\max}$-subopt for $\boldsymbol{\kappa}$ then, for any coloring κ that hits S', one has $\mathrm{mfe}_\kappa + \Delta_{\max} \leq \Delta G(S')$. Since $E^\star_{SA} \leq \mathrm{mfe}_\kappa$, one concludes with

$$\Delta G(S^\star) = E^-_{\text{clash-free}} \leq E^\star_{SA} + \Delta_{\max} \leq \mathrm{mfe}_\kappa + \Delta_{\max} \leq \Delta G(S').$$

The energy of any alternative S' is thus higher than that of $S^\star$, from which we conclude that our algorithm is correct.

Of course, the practical performances of the algorithm may critically depend on the $\Delta_{\max}$ value, *i.e.* the energy difference between the MFE self-avoiding and clash-free complexes. To mitigate the issue, we introduce in the next Section 2.2 an optimization based on cliques, which we illustrate in Figure 2 along with our algorithm.

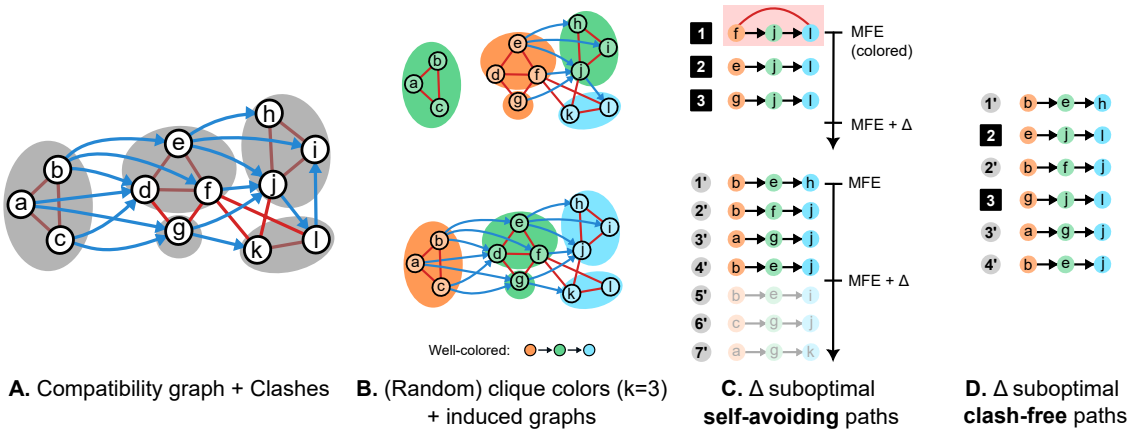

Fig. 2: **Example of Δ-suboptimal color coding based on monochromatic cliques.** From a MFEDOCK instance (A), including compatibility arcs (blue) and clashes edges (red), a clique cover is heuristically computed (gray). A family of coloring is then generated (B; random or deterministic), and a dynamic programming algorithm allows to build a list of (self-avoiding) k-paths (C). Among those, only $\mathsf{f} \to \mathsf{j} \to \mathsf{l}$ presents a clash (red box) and is filtered out to obtain a merged list of clash-free Δ-suboptimals. Notably, the two colorings above are sufficient to hit all clash-free paths. Moreover, $\mathsf{j} \to \mathsf{l} \to \mathsf{i}$, a valid k-path which features two clashing nodes, cannot be well-colored in the current clique cover.

2.2 Reducing clashes through monochromatic clique covers

During the initial docking phase, individual fragments usually cluster around hotspots on the protein surface. On the one hand, such an accumulation is beneficial to the resolution of the selected poses as, to a degree, it enables simulating flexibility. On the other hand, high local densities may result in a combinatorial explosion of clashes, drastically reducing the density of clash-free complexes, all the while hindering the performances of our algorithms. Instead, we would rather focus our effort on a subset of self-avoiding paths featuring a good density of clash-free associated complexes.

Cliques of clashing nodes can be safely set to a single color. To achieve such a goal, we trivially remark that the nodes of a **clashing clique** $\mathcal{C}$, *i.e.* a set of pairwise clashing poses, may not occur more than once within a reasonable candidate complex. Nicely, such clashes can be avoided by enforcing the **monochromaticity** of κ with respect to $\mathcal{C}$, the use of a single color for all $v \in \mathcal{C}$. This restriction is conservative with respect to clash-free complexes, since any clash-free k-path, hit by a coloring κ and featuring a node $v_c \in \mathcal{C}$, is also hit by a monochromatic coloring κ' such that

$$\kappa' : v \to \begin{cases} \kappa(v_c) & \text{if } v \in \mathcal{C} \\ \kappa(v) & \text{otherwise.} \end{cases}$$

Meanwhile, any k-path that borrowed two or more nodes from $\mathcal{C}$ is no longer admissible, thereby increasing the density of clash-free complexes within the search space. Moreover, from the perspective of derandomization, this restriction enables the complete clash-free paths to be hit by a smaller family of colorings, since whole cliques can be treated as a single nodes.

This observation, and overall strategy, provably generalizes to collections of disjoint cliques in the clash graph. Two overlapping cliques $\mathcal{C}$ and $\mathcal{C}'$, however, should not be forced to be simultaneously monochromatic. Indeed the color of $\mathcal{C}$ would, due to its overlap with $\mathcal{C}'$, spread to the latter. This would result in treating $\mathcal{C} \cup \mathcal{C}'$ as a single clique, thereby potentially eliminating some clash-free k-paths from the search space. In order to minimize runtime, and maximize the density of clash-free paths within the runspace, we preprocess clashes by decomposing them as a **clique cover**, a partition of nodes into a set of cliques $\mathfrak{C} = \{\mathcal{C}_i\}$ while attempting to maximize the number of clashing pairs occurring within a clique $\mathcal{C}_i$. Though not strictly equivalent, this problem is related to MIN CLIQUE COVER and likely hard.

A pragmatic solution to decompose clashes into non-overlapping cliques. To pragmatically solve the problem, we implemented an **greedy heuristic for min clique cover** which initializes the cover

$\mathfrak{C} := \varnothing$ and, at each iteration, starts from the node v^+ having max degree node in the remaining graph. It initializes a clique $\mathcal{C} := \{v^+\}$ and a list of common neighbors $\mathcal{N} :=$ neighbors(v^+). Then, until $\mathcal{N} = \varnothing$, it alternates:

1. Choose a node $v \in \mathcal{N}$ having max degree within $\mathcal{N}$, is added to $\mathcal{C}$;
2. Update of list of common neighbors through $\mathcal{N} := \mathcal{N} \cap$ neighbors(v);

The clique $\mathcal{C}$ is then added to the cover $\mathfrak{C}$, removed from the clique graph for future iterations (choice of v^+, construction of $\mathcal{C}\dots$) until all nodes have been removed from the clash graph and added as part of a clique to $\mathfrak{C}$. While this heuristic does not provide formal guarantees regarding its result, we found it performs adequately for our typical instances, as shown in Section 3.2.

2.3 Rational ssRNA design as a relaxation of docking

Rational design in the context of a fragment-based docking usually requires two properties to be fulfilled by the designed RNA aptamer: **Positive design** requires the ligand to have optimal affinity, or low interaction free-energy, towards the target protein or targeted pocket; **Negative design** constrains the ligand to be specifically binding to a given region of the protein. Interestingly, both criteria are at least partially addressed by a simple relaxation of the MFEdock problem.

The required modification simply consists in partially specifying (*e.g.* using IUPAC codes), or even disregarding altogether (*e.g.* poly-N mask), the input ssRNA sequence without added complexity. In this setting, solving the MFEdock problem provides an MFE complex, from which both an ssRNA sequence $r^\star := r(x_1).r(x_2)\cdots r(x_k)$ and its MFE conformation can be derived. More precisely, we can show that: i) No alternative sequence has higher affinity than $r^\star$ towards the protein (positive design); ii) The binding site induced on the protein surface by $x^\star$ is the most likely target for $r^\star$ (negative design). Admittedly, this approach does not enable targeting of a specific site or pocket, since the best complex location of is induced by the MFE criterion. Nevertheless, by generating suboptimals and only retaining the first occurrence of each sequence (*i.e* associated with their MFE complex), one can produce a diversity of sequences that are both stable, and specifically target various sites.

2.4 Equilibrium statistics

While clearly an important – computationally challenging – problem, docking through energy minimization is hindered by its single focus on the MFE conformation. Indeed, at the thermodynamic equilibrium, the probability of a clash-free complex x follows a Boltzmann distribution

$$\mathbb{P}(X = x \mid r) = \frac{e^{-\beta.\Delta G(x)}}{\mathcal{Z}_r} \text{ where } \mathcal{Z}_r = \sum_{\substack{x' \text{ clash-free} \\ \text{and comp. with } r}} e^{-\beta.\Delta G(x')}$$

is the partition function for a nucleotide sequence r, $\mu = RT$ with R the Boltzmann constant and T the absolute temperature. Since the number of valid complexes typically grows (at least) exponentially with k, the probability of the MFE complex becomes abysmally small in larger systems. As an extreme example, for the clique input graph, the number of complexes grows in $\Theta(n^k)$ when $k \ll n$, and even $n! \asymp (n/e)^n$ when $k = n$, thereby completely crushing the probability of any single complex.

Boltzmann statistics. This motivates a computation of **equilibrium statistics**, *i.e* expected properties of the system under a Boltzmann distribution. Such properties are measured by a set of real-valued **feature functions** $\{f_1, f_2, \dots\}$, each mapping a valid complex to some numerical value in $\mathbb{R}$. Features can represent any relevant quantity (free-energy, %occupancy of druggable pocket…), provided that they can be effectively computed from a fully-specified complex. The **expectation of a feature** f is defined as:

$$\mathbb{E}(f(X) \mid r) = \sum_{\substack{x \text{ self avoiding} \\ \text{and comp. with } r}} f(x) \times \mathbb{P}(X = x \mid r)$$

and can be interpreted as a collective variables.Probabilities can also be computed as expectations of (0/1)-valued features. Indeed, setting $f_c(r) = 1$ or 0 depending on the presence/absence of a contact with a targeted residue A, the expectation simplifies into

$$\mathbb{E}(f_c(X) \mid r) = \sum_x f_c(x) \times \mathbb{P}(x \mid r) = \sum_{x \text{ s.t. } f_c(x)=1} \mathbb{P}(x \mid r) = \mathbb{P}\left(f_c(X) = 1 \mid r\right).$$

Higher moments of the distributions can finally be computed from the expectations of $f, f^2, f^3 \ldots$ enabling access to finer characteristics of the distribution, such as its variance/stddev, skewness, kurtosis... or even correlations between multiple features. Complexity-wise, computing the partition function is provably harder than the, already-hard, optimization problem addressed in Section **??**. Worse, as defined for optimization, families of coloring used for derandomization would typically introduces a bias in the subsequent estimates, and thus cannot be used.

Statistical estimators from colored statistics. To work around those hurdles, we adopt an approach that estimates the expectation based on a sequence of random colorings $\kappa_1, \kappa_2 \ldots$ Namely, we introduce the **color-restricted expectation** of a feature f given a coloring κ as:

$$\mathbb{E}(f(X) \mid r, \kappa) = \sum_{\substack{x \text{ clash-free,} \\ \text{comp. with } r \\ \text{well col. by } \kappa}} f(x)\, \mathbb{P}(x \mid r, \kappa) = \sum_{\substack{x \text{ clash-free,} \\ \text{comp. with } r \\ \text{well col. by } \kappa}} f(x)\, \frac{e^{-\beta \Delta G(x)}}{\mathcal{Z}_{r,\kappa}} \text{ where } \mathcal{Z}_{r,\kappa} = \sum_{\substack{x' \text{ clash-free,} \\ \text{comp. with } r \\ \text{well col. by } \kappa}} e^{-\beta \Delta G(x')}.$$

To estimate this quantity, we first introduce a dynamic programming scheme to compute the (coloring-restricted) partition function:

$$\mathcal{Z}_\kappa = \sum_{\substack{v \in V \\ \text{such that } r(v)=r_1}} \mathcal{Z}_{v,1} \quad \text{and} \quad \mathcal{Z}_{v,m} = \begin{cases} e^{-\beta E(v)} & \text{if } m = k \\ \displaystyle\sum_{\substack{(v,v') \in E \text{ s.t.} \\ \kappa(v')=m+1 \\ \text{and } r(v')=r_{m+1}}} e^{-\beta(E(v)+E'(v,v'))}\, \mathcal{Z}_{v',m+1} & \text{otherwise} \end{cases} \tag{6}$$

A stochastic backtrack then consists in, starting from $\mathcal{Z}_k$, the repeated choice a node with probability proportional to its contribution to $\mathcal{Z}_\kappa$, recursing until the $m = k$ condition is met. The returned random complex is then Boltzmann distributed within well-colored k-paths. The average value of f on a set of generated complexes, further filtered to retain only clash-free paths, represents an unbiased estimator for $\mathbb{E}(f(X) \mid r, \kappa)$. Our, provably consistent, final estimator takes a collection of random uniformly-distributed colorings, and returns:

$$\widehat{f}(\kappa_1, \kappa_2 \ldots \kappa_M) = \frac{\sum_{i=1}^M \mathcal{Z}_{r,\kappa_i} \times \mathbb{E}(f(X) \mid r, \kappa_i)}{\sum_{j=1}^M \mathcal{Z}_{r,\kappa_j}} \tag{7}$$

3 Results

Implementation. We implemented our algorithms for MFEdock (optimization; subopts; +/- sequence constraints) and statistical estimators into the `ColorDocking` software, a collection of `Python` scripts interfacing `C` code, freely downloadable with datasets and further information to reproduce experiments at https://gitlab.inria.fr/amibio/colordocking. All experiments were performed on a PBS cluster with a Linux kernel, using 125GB of memory. Each calculation was done on a single CPU.

Datasets. We selected seven ssRNA/protein complexes to validate our method. Among them, six complexes are RNA-RRM complexes (RRM; 1B7F, 1CVJ, 2MGZ, 2YH1, 3NNH and 4BS2) and the remaining one is a Pumilio domain (PUF; 3BX3). This generally coincides with the benchmark selected by de Beauchene et al. (2016), removing two structures: 4N0T, which natively interacts with a double-stranded RNA; and 5BZV, another PUF which we saw as redundant with 3BX3. To provide a realistic setting for docking, proteins were prepared and minimised using the CHARMM36 force field in the absence of the ssRNA ligand and solvent. As a result, the protein surface at the RNA binding site may have been altered, and in a more specific way at some specific position of the RNA chain.

Complex	k	#Poses	Target Seq.	α	#Paths	#SA (cliques)	#Clash-Free
1B7F	5	2 171	UUUUU	14 388	$5.22{\cdot}10^{8}$	$1.49{\cdot}10^{8}$	$6.98{\cdot}10^{6}$
1CVJ	5	2 031	AAAAA	14 388	$9.44{\cdot}10^{7}$	$2.24{\cdot}10^{7}$	$1.46{\cdot}10^{6}$
2MGZ	5	4 329	GGUGU	14 388	$1.84{\cdot}10^{7}$	$4.89{\cdot}10^{6}$	$3.20{\cdot}10^{5}$
2YH1	5	2 064	UUUUU	14 388	$1.57{\cdot}10^{8}$	$2.97{\cdot}10^{7}$	$1.22{\cdot}10^{6}$
3BX3	5	7 464	UAUAU	14 388	$1.42{\cdot}10^{8}$	$5.39{\cdot}10^{7}$	$5.21{\cdot}10^{6}$
3NNH	5	4 606	UUUUG	14 388	$5.58{\cdot}10^{7}$	$1.85{\cdot}10^{7}$	$3.64{\cdot}10^{6}$
4BS2	5	8 150	GAAUG	14 388	$2.37{\cdot}10^{7}$	$8.65{\cdot}10^{6}$	$1.03{\cdot}10^{6}$
1B7F	7	2 171	GUUUUUU	214 856	-	-	-
1CVJ	8	2 031	AAAAAAAA	3 792 553	-	-	-
1CVJ	8	5 785	AAAAAAAA	3 792 553	-	-	-
1CVJ	8	26 570	AAAAAAAA	3 792 553	-	-	-
2MGZ	7	4 329	UGGUGUG	214 856	-	-	-
3BX3	8	7 464	UGUAUAUA	3 792 553	-	-	-
3NNH	6	4 606	UUUUGU	214 856	-	-	-

Table 1: Summary of our benchmark and paths cardinality analysis.

For each of our targets, we used the `MCSS` (Miranker and Karplus, 1991) method to generate a distribution of 10.000 fragment poses. These fragments are composed of Adenine (A), Cytosine (C), Guanine (G) and Uracil (U). Only the first 4.000 fragments poses (2.000 anti-conformations and 2.000 syn-conformations) were used for this study. From these, the python package `NUCLEAR` (González-Alemán and Leclerc, 2023) was used to cluster poses within 0.5Å RMSD, and to build matrices (connectivity, clashes, scores) using 4.5Å as a max. value for the O3'-C5' distance. The max contact distance between the protein and a fragment was set to 3.5Å. All atoms of amino acids were considered, as well as for nucleotides, except for the terminal patch which was omitted to improve connectivity. All `NUCLEAR` runs were constrained to only generate the various matrices (`run_type = partial` option).

A directed graph (+ clash matrix) was then generated and fed as input to our implementation of `ColorDocking` algorithm, using collections of clique-level random uniform colorings. We uniformly set the tolerance to $\varepsilon = 0.01$, *i.e.* the clash-free MFE complex was predicted with probability of $p = 99\%$. Clash-free MFE candidates were produced, based on a suboptimality cutoff $\Delta = 10$ kcal.mol^{-1}.

3.1 Stability analysis

Setting the RNA length to $k = 5$ enables the execution of our algorithm in modest time (about a dozen seconds per run). Such a runtime enables a comparison the results obtained over successive executions, in order to assess the impact of the random generation of colorings on the stability of predictions. Namely, for all of the 7 complexes, and setting $k = 5$, we performed 100 independent experiments. In addition, we used a brute-force approach to compute both self-avoiding and clash-free MFE complexes. As expected, we consistently recover the clash-free MFE complex in our experiments, namely between 97/100 and 100/100 of experiments over all targets. Such a behavior is expected from our choice of $\varepsilon = 0.01$, implying a 1% chance of missing the MFE but could, at least in theory, have been affected by setting Δ to a fixed value. Setting $\varepsilon = 0.63$ expectedly reduces the runtime by a factor 10, at the cost of degraded performances, with the MFE clash-free structure being only found between 25/100 and 47/100 of experiments.

In a second analysis, we investigated whether the top-100 clash-free complex present a strong overlap across independent executions of the algorithm. We filtered the output of Δ-suboptimal version of the MFEdock algorithm ($\Delta = 10$, $\varepsilon = 0.01$) to produce the 100 lowest-energy clash-free complexes. We iterated the experiment 100 times, and found the average pairwise overlap between two runs to be of 98%, with very limited variations.

3.2 Impact of monochromatic clique covers

Next we turn to an investigation of the effect of clique-based coloring on the density of clash-free paths, the runtime and energy distance between the self-avoiding MFE and the clash-free MFE. To investigate those points, in addition to the MFE obtained as above, we used a brute-force approach to compute the numbers of unconstrained, self-avoiding and clash-free paths.

Complex	#Poses	#Cliques	Max clique size	%Clique edges	Avg time (sec) +cliques	Avg time (sec) -cliques	$\lvert MFE_{\text{clash-free}} - MFE_{SA} \rvert$ +cliques	$\lvert MFE_{\text{clash-free}} - MFE_{SA} \rvert$ -cliques
1B7F	2 171	49	298	54.56	10.63	16.37	9.15	11.45
1CVJ	2 031	40	344	60.38	6.75	30.00	1.37	1.37
2MGZ	4 329	63	584	57.64	2.40	5.40	8.45	18.74
2YH1	2 064	49	312	66.40	8.60	15.53	5.2	9.6
3BX3	7 464	70	889	49.29	16.71	18.02	5.3	10.97
3NNH	4 606	55	629	57.62	3.61	12.73	2.52	7.25
4BS2	8 150	98	625	53.79	6.19	12.28	2.07	8.68
1CVJ	5 785	44	1 109	71.50	-	-		

Table 2: Properties and impact of cliques cover on runtime. Values observed for $k = 5$, averaged over 100 iterations (std dev ≈ 1).

We first report in Table 2 the clique covers returned by our greedy heuristic. While the number of poses is in the order of (dozens of) thousands, the number of clashing cliques scales between 40 and 100, with larger cliques representing a sizeable proportion of the vertex set (10 to 20%). Moreover, a large proportion (50% to 70%) of clashing edges are internal to a clique. Such clashes can no longer occur upon restricting to clique-based coloring, substantially reducing the probability of a k-path featuring a clash.

As can be seen in Table 1, the number of paths is typically reduced by 75% when clique-monochromatic self-avoidance is ensured. This results in a runtime reduction an overall factor 1.5 to 4. Meanwhile, as can be seen in Table 2, the energy distance between the self-avoiding and clash-free MFEs can be greatly reduced (*e.g.* -10 kcal.mol^{-1} for 2MGZ) when cliques are used to reduce the search space. Overall, the consideration of monochromatic cliques represents a very positive addition: it greatly improves the runtime and purifies self-avoiding paths to increase the density of clash-free paths.

3.3 Docking through energy-minimization under different fragment definitions

We showcased the versatility of our approach by supplementing the `MCSS` set of poses with the connectivity graphs associated with overlapping trinucleotide fragments, following de Beauchene et al. (2016). For 1CVJ, we used the `ATTRACT` software (de Vries et al., 2015) to generate 1.000.000 non-redundant (0.2Å RMSD threshold) fragment poses, in coarse-grained representation. We built a connectivity matrix, using as connectivity criteria a 1.8Å RMSD cutoff for overlapping nucleotides between consecutive fragments. We created a directed graph of connected poses, using the `ssRNATTRACT` package as described in de Beauchene et al. (2016). We used a new fragment library of RNA trinucleotides extracted from the PDB with the `ProtNAff` software (Moniot et al., 2022), using our Radius clustering method

From the 1.000.000 initial poses, 5327 could be assembled in a 5-fragments chain, and were therefore retained in the final graph of connected poses. We then constructed a clash matrix of those poses, using a 1.5Å distance criteria between two clashing heavy atoms (excluding overlapping nucleotides of connected poses). In addition, we considered a pair of poses as incompatible if both can be connected only at the same position in a 5-fragments chain, since they can not be together in the same chain. The full matrix of poses incompatibility (either clashing or only at the same position in chains) was used to define cliques of incompatible poses

The resulting MFEdock instance only needs to be executed for $k = 6$, since each fragment represents a trinucleotide, and assembling 6 fragments is sufficient to reach the size of 8 nucleotides. This allows to execute our algorithm in as little as 34 sec. Meanwhile, the runtime required by 1CVJ for our MCSS dataset, implying $k = 8$, is of $2{\cdot}10^3$ sec, or approximately 33 minutes. Beyond the demonstrated ability of `ColorDocking` to support multiple fragment definition, we did not analyze further the quality of the produced fragments (*e.g.* RMSD to native complex), since our goal is not to compare different force fields/fragment definitions.

3.4 Design

To illustrate the capacity of MFEdock to address design, we considered a design study recently published by Perzanowska et al. (2022), where an oligonucleotide targeting a poly(A)-binding protein (PDBID: 1CVJ) was designed. Since this study included modified nucleotides, we considered an extended list of nucleotides: two without modification (A,G), and 5 with modifications: adenosine and guanosine with a phosphorothioate (A_P, G_P), protonated adenosine (A_ψ), N6-methyladenosine (m^6A), N6-methyladenosine including phosphorothioate (m^6A_P), O-methyladenosine (Am) and O-methyladenosine including phosphorothioate (Am_P). All were used in anti- and syn-conformations, for a total of 1 000 poses (500 syn/500 anit) per nucleotide type. They were clustered at 0.5Å RMSD for a remaining number of 5 785 individual poses.

To generate solutions of length k=8, we considered a maximum value of $\Delta_{max} := 3$, and gradually increased Δ by unit steps, to reach a maximum of 100 clash-free per coloring. We initially did not consider sequence constraints. The number of unique sequences was 75 out of 562 clash-free solutions. We report below the top 10 of unique sequence, along with their minimum free energy:

A_ψ-A-A_P-G_P-A_P-m^6A-m^6A-A -178.405	m^6A_P-m^6A-A-G_P-A_P-m^6A-m^6A-A -178.404
A_ψ-A-G_P-G_P-A_P-m^6A-m^6A-A -178.377	G-m^6A_P-m^6A-A-A_P-A_P-m^6A-m^6A -178.055
A_ψ-A-A_P-G_P-A_P-Am-Am-m^6A -178.052	A_{mP}-A_m-A-G_P-A_P-A_m-A_m-m^6A -178.050
A_{mP}-A_m-A-A_P-A_P-A_m-A_m-A -178.030	A_ψ-A-G_P-G_P-A_P–A_m-A_m-m^6A -178.023
A_P-A_{mP}-A_m-A-G_P-A_P-A_m-A_m -177.864	A_ψ-A-A_P-G_P-A_{mP}-A_m-A_m-A -177.812

Interestingly, those differ from the sequences investigated by Perzanowska et al. (2022). In particular, the pair of sequences having highest affinity in the study, was not found in our list. This is not entirely surprising, since the authors limited their investigation to a single modified nucleotide per design. We further analyzed their two best sequences, running a sequence-constrained instance of MFEdock

A-A-A-A-A-A-m^6A-A -161.252 m^6A-A-A-A-A-A-A-A -151.976

and found that their MFE is significantly higher (+16/+26 kcal.mol^{-1}), suggesting that our ability to tame the combinatorial explosion grants us access to promising alternatives.

4 Conclusions and perspectives

We have introduced a new algorithmic approach, based on color coding, to solve natural problems arising in the context of fragment-based ssRNA docking and design on the surface of a rigid protein. We have illustrated their utility in the context of four RNA binding proteins, showing that color coding provides a versatile toolkit for the study and design of ssRNAs.

A key strength of our exact algorithm resides in its linear complexity on the number of pairwise connected poses, only being exponential on the length k of the ssRNA. As such, it can be seen as a parameterized complexity algorithm, showing that the MFEdock problem is Fixed Parameter Tractable (FPT) for the ssRNA length k. On a practical level, much larger sets of poses/connections could be supported, allowing to explore the impact of various sampling depth/density of poses on the quality of predictions. Our algorithmic method is not restricted to individually docked nucleotides, and could accommodate other fragment libraries.

Acknowledgments

The authors are greatly indebted to Laurent Bulteau for suggesting well-colored paths as a memory-efficient alternative to colorful paths, and to Sebastian Will for debunking an earlier, but ultimately erroneous, epiphany.

Bibliography

N. Alon, R. Yuster, and U. Zwick. Color-coding. In *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing - STOC '94*. ACM Press, 1994. https://doi.org/10.1145/195058.195179.

N. Alon, R. Yuster, and U. Zwick. Color-coding. *J. ACM*, 42(4):844–856, jul 1995. ISSN 0004-5411. https://doi.org/10.1145/210332.210337. URL https://doi.org/10.1145/210332.210337.

N. Alon, P. Dao, I. Hajirasouliha, *et al.* Biomolecular network motif counting and discovery by color coding. *Bioinformatics*, 24(13):i241–i249, jun 2008. https://doi.org/10.1093/bioinformatics/btn163.

G. Bollag, J. Tsai, J. Zhang, *et al.* Vemurafenib: the first drug approved for BRAF-mutant cancer. *Nature Reviews Drug Discovery*, 11(11):873–886, oct 2012. https://doi.org/10.1038/nrd3847.

N. Chevrollier. *Développement et application d'une approche de docking par fragments pour modéliser les interactions entre protéines et ARN simple-brin*. Theses, Université Paris-Saclay, May 2019. URL https://tel.archives-ouvertes.fr/tel-02436914.

I. C. de Beauchene, S. J. de Vries, and M. Zacharias. Fragment-based modelling of single stranded RNA bound to RNA recognition motif containing proteins. *Nucleic Acids Research*, 44(10):4565–4580, apr 2016. https://doi.org/10.1093/nar/gkw328.

S. J. de Vries, C. E. Schindler, I. C. de Beauchêne, and M. Zacharias. A web interface for easy flexible protein-protein docking with attract. *Biophysical Journal*, 108(3):462–465, 2015. ISSN 0006-3495. https://doi.org/https://doi.org/10.1016/j.bpj.2014.12.015. URL https://www.sciencedirect.com/science/article/pii/S0006349514047602.

B. Dost, T. Shlomi, N. Gupta, *et al.* QNet: A tool for querying protein interaction networks. *Journal of Computational Biology*, 15(7):913–925, sep 2008. https://doi.org/10.1089/cmb.2007.0172.

R. González-Alemán and F. Leclerc. NUCLEotide AssembleR (NUCLEAR) package. https://github.com/rglez/nuclear, 2023.

R. González-Alemán, N. Chevrollier, M. Simoes, *et al.* MCSS-based predictions of binding mode and selectivity of nucleotide ligands. *Journal of Chemical Theory and Computation*, 17(4):2599–2618, mar 2021. https://doi.org/10.1021/acs.jctc.0c01339.

D. Hall, S. Li, K. Yamashita, *et al.* RNA–LIM: A novel procedure for analyzing protein/single-stranded RNA propensity data with concomitant estimation of interface structure. *Analytical Biochemistry*, 472:52–61, mar 2015. https://doi.org/10.1016/j.ab.2014.11.004.

A. Itai, C. H. Papadimitriou, and J. L. Szwarcfiter. Hamilton paths in grid graphs. *SIAM Journal on Computing*, 11(4):676–686, nov 1982. https://doi.org/10.1137/0211056.

K. Kappel and R. Das. Sampling native-like structures of RNA-protein complexes through rosetta folding and docking. *Structure*, 27(1):140–151.e5, jan 2019. https://doi.org/10.1016/j.str.2018.10.001.

P. Kirsch, A. M. Hartman, A. K. H. Hirsch, and M. Empting. Concepts and core principles of fragment-based drug design. *Molecules*, 24(23):4309, nov 2019. https://doi.org/10.3390/molecules24234309.

D. Marx. Parameterized complexity of independence and domination on geometric graphs. In H. L. Bodlaender and M. A. Langston, editors, *Parameterized and Exact Computation*, pages 154–165, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-39101-2.

A. Miranker and M. Karplus. Functionality maps of binding sites: A multiple copy simultaneous search method. *Proteins: Structure, Function, and Genetics*, 11(1):29–34, sep 1991. https://doi.org/10.1002/prot.340110104.

A. Moniot, Y. Guermeur, S. J. de Vries, and I. Chauvot de Beauchene. ProtNAff: protein-bound Nucleic Acid filters and fragment libraries. *Bioinformatics*, 38(16):3911–3917, 07 2022. ISSN 1367-4803. https://doi.org/10.1093/bioinformatics/btac430. URL https://doi.org/10.1093/bioinformatics/btac430.

M. Naor, L. Schulman, and A. Srinivasan. Splitters and near-optimal derandomization. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*. IEEE Comput. Soc. Press, Oct. 1995. https://doi.org/10.1109/sfcs.1995.492475.

T. P. Perera, E. Jovcheva, L. Mevellec, *et al.* Discovery and pharmacological characterization of JNJ-42756493 (erdafitinib), a functionally selective small-molecule FGFR family inhibitor. *Molecular Cancer Therapeutics*, 16(6):1010–1020, jun 2017. https://doi.org/10.1158/1535-7163.mct-16-0589.

O. Perzanowska, M. Smietanski, J. Jemielity, and J. Kowalska. Chemically modified poly(a) analogs targeting pabp: Structure activity relationship and translation inhibitory properties. *Chemistry – A European Journal*, 28(42):e202201115, 2022. https://doi.org/https://doi.org/10.1002/chem.202201115. URL https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/chem.202201115.

J. P. Schmidt and A. Siegel. The spatial complexity of oblivious k-probe hash functions. *SIAM Journal on Computing*, 19(5):775–786, 1990. https://doi.org/10.1137/0219054. URL https://doi.org/10.1137/0219054.

J. Schoepfer, W. Jahnke, G. Berellini, *et al.* Discovery of asciminib (ABL001), an allosteric inhibitor of the tyrosine kinase activity of BCR-ABL1. *Journal of Medicinal Chemistry*, 61(18):8120–8135, aug 2018. https://doi.org/10.1021/acs.jmedchem.8b01040.

A. Schuffenhauer, S. Ruedisser, A. Marzinzik, *et al.* Library design for fragment based screening. *Current Topics in Medicinal Chemistry*, 5(8):751–762, aug 2005. https://doi.org/10.2174/1568026054637700.

T. Shlomi, D. Segal, E. Ruppin, and R. Sharan. QPath: a method for querying pathways in a protein-protein interaction network. *BMC Bioinformatics*, 7(1), apr 2006. https://doi.org/10.1186/1471-2105-7-199.

A. J. Souers, J. D. Leverson, E. R. Boghaert, *et al.* ABT-199, a potent and selective BCL-2 inhibitor, achieves antitumor activity while sparing platelets. *Nature Medicine*, 19(2):202–208, jan 2013. https://doi.org/10.1038/nm.3048.

W. D. Tap, Z. A. Wainberg, S. P. Anthony, *et al.* Structure-guided blockade of CSF1r kinase in tenosynovial giant-cell tumor. *New England Journal of Medicine*, 373(5):428–437, jul 2015. https://doi.org/10.1056/nejmoa1411366.

M. S. Waterman and T. H. Byers. A dynamic programming algorithm to find all solutions in a neighborhood of the optimum. *Mathematical Biosciences*, 77(1-2):179–188, dec 1985. https://doi.org/10.1016/0025-5564(85)90096-3.