Colorful yet tractable: docking, design and equilibrium statistics of protein-binding ssRNAs

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

Colorful yet Tractable: 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], and Yann Ponty[1][0000−0002−7615−3930]

[1] Laboratoire d'Informatique de l'École Polytechnique (LIX – CNRS UMR 7161), Institut Polytechnique de Paris, Palaiseau, France yann.ponty@lix.polytechnique.fr

[2] Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Université Paris Saclay, Gif-sur-Yvette F-91198, France

[3] Laboratorio de Química Computacional y Teórica (LQCT), Facultad de Química, Universidad de La Habana, La Habana, 10400, Cuba

Abstract. We revisit the fragment-based docking and design of single-stranded RNA aptamers (ssRNAs), consisting of k nucleotides, onto a rigid protein. Individual fragments, representing nucleotides, are 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 combinatorial explosion ($\Theta(n^k)$ conformations), hindering the potential of fragment-based methods.

We leverage the elegant color-coding technique, introduced by Alon, Yuster and Zwick to solve the associated problems exactly in time and space linear on n the number of poses, and in time only exponential on k the number of nucleotides. The dynamic programming algorithm at the core of our method is surprisingly simple, and can be extended to produce suboptimal candidates, or to perform stochastic sampling of candidates within a Boltzmann distribution. This sampling procedure can be adapted into a statistically-consistent estimator for virtually any feature of interest.

The versatility and practicality of the color coding framework, demonstrated by a successful reanalysis and redesign of documented ssRNA/protein complexes, could be key to the development of future hybrid discrete/continuous methods in structural bioinformatics.

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

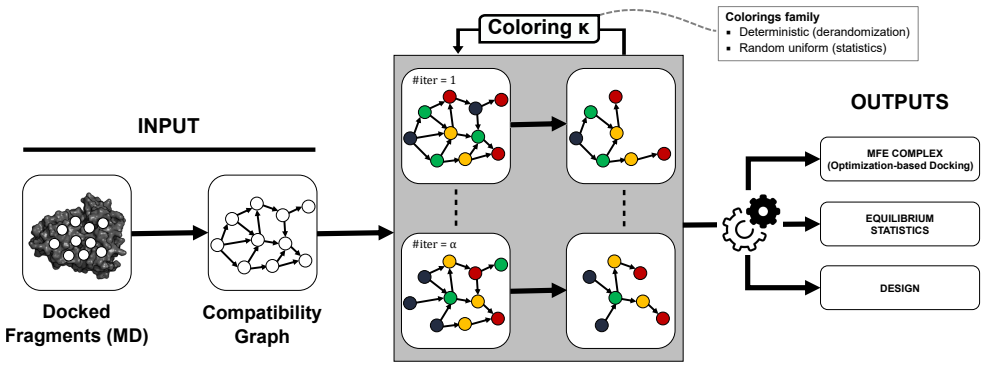


Fig. 1: General workflow: Starting from a graph of docked fragments, for which pairwise connectivity has been assessed by an external tool, our method considers various – random or deterministic – colorings from which solutions to various hard computational problems can be obtained.

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 [3,19,14,15], one of which as recently as 2021 [18]. Fragments are usually small compounds with low molecular weight, having about 20 heavy atoms and having low molecular weight compounds [11,16]. The principle of this strategy is to dock a library 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 RNA-RBP (RNA-Binding Protein) [8,5,10]. To predict the interaction between 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 [7] 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 enumeration 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 [2] that allows to capture a necessary notion of ssRNA self avoidance. In Section 2, we show how to utilize color coding to obtain exact or probabilistic algorithms for fragment-based docking through energy minimization. In Section 3, we extend the framework to provide statistically consistent estimates for virtually any feature of interest. Section 4 describes how to avoid complex clashes, and perform design by relaxing the requirement of being compatible with a given nucleotide sequence. Section 5 illustrates the proposed algorithms in the context of 4 RNA binding proteins, and Section 6 discusses some limitations of the approach, and future extensions.

2 Color coding for free-energy minimization

Let us start by making two of our assumptions explicit: First, 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; Second, 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 observing the joint configuration.

Graph modeling. Under these two assumptions, the fragment-based docking and screening of ssRNAs onto a protein can both be reformulated as graph problems. Indeed, let us denote as **fragment** f as a nucleotide $r(f)$ 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 its spatial positioning avoids 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 [4] or Nuclear (to be released) to deal with short ssRNAs.

Through a systematic (constrained) docking of fragments onto the surface of the target protein, followed by a pairwise assessment of the compatibility of resulting fragment poses, one obtains a **poses compatibility 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 compatibility of v and v'. In the following, we denote by $n := |V|$ the number of poses in the graph. At first approximation, the paths in a compatibility graph correspond to a discrete set of joint ssRNA/protein conformations, which will be called **complexes** in the following.

However, paths induced by the compatibility graphs may not always correspond to good candidates. For instance, any path that uses the same pose $v \in V$ twice will induce a **hard clash** in the ssRNA. **Soft clashes** may also occur between poses that are not necessarily identical, but not directly consecutive. We initially focus on enforcing an avoidance of hard clashes, and define a **complex** x as a (self-avoiding) path in G, *i.e.* a sequence $x := (v_1, \ldots, v_k)$ of distinct vertices from G such that $(v_i, v_{i+1}) \in E, \forall i \in [1, k-1]$. As will be discussed later, this property already leads to challenging computational problems.

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

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

where $E : V \to \mathbb{R} \cup \{+\infty\}$ and $E' : V \times V \to \mathbb{R} \cup \{+\infty\}$ are terms which capture the contributions of individual and pairwise-adjacent fragments respectively. Assuming thermodynamic equilibrium, the most stable/probably complex, for a given nucleotides sequence of length k, is the one having Minimum Free-Energy while compatible with the sequence. The computation of such a complex can be restated as follows:

MFEdock **problem.**
Input: Pose compatibility graph $G = (V, E)$; Energy function ΔG; Residue sequence $r = r_1.r_2 \ldots r_k$.
Output: Hard clash-free complex $x^\star = (v_1^\star, v_2^\star, \ldots, v_k^\star)$ minimizing free-energy

$$x^\star = \underset{\substack{x=(v_1,\ldots,v_k) \\ \text{such that } v_i \neq v_j, \forall i \neq j \\ \text{and } r(v_i)=r_i, \forall i}}{\operatorname{argmin}} \Delta G(x) \quad \begin{array}{l} \leftarrow \text{self avoidance} \\ \leftarrow \text{compatibility with nuc. sequence} \end{array}$$

For general input graphs, trivial sequences (homopolymer, $k := |V|$) and simplye energy function, MFEdock generalizes the problem of deciding the existence of a Hamiltonian path, implying NP-hardness.

Color coding to the rescue. Fortunately, the MFEdock problem is easily amenable to probabilistic, or even exact, resolution using the **color coding technique** introduced by Alon, Yuster and Zwick [2]. This elegant technique initially addressed the problem of finding sparse motifs in graphs. It has been utilized in the context Bioinformatics for searching [6,17] and counting [1] occurrences of motifs in biological networks.

The key idea 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 found and counted in time linear on $n := |V|$ and $|E|$, and only exponential on k. Clearly, any colorful path is also a k-path as it uses distinct

vertices. However, for a given coloring κ, some existing k-path x in G may not be colorful in a random uniform coloring, an event with probability $\mathbb{P}(p \text{ not colorful}) = 1 - \frac{k!}{k^k}$.

To improve the odds of finding a k-path, the method then iterates the search for colorful paths within a set of α random uniformly-distributed colorings of G. Assuming independent uniformly-drawn colorings $\kappa_1, \ldots, \kappa_\alpha$ the probability of hitting a k-path after α random colorings is exactly

$$\mathbb{P}(p \text{ colorful for some coloring} \mid \alpha) = 1 - \left(1 - \frac{k!}{k^k}\right)^\alpha. \tag{1}$$

Meanwhile, the expected number of random colorings needed before finding x (*i.e.* x colorful w.r.t at least one of the colorings) is given by $k^k/k! \in \mathcal{O}(\sqrt{k}\, e^k)$ using Stirling formula.

Impressively, this probabilistic algorithm can be turned into an efficient, deterministic and exact, algorithm using the **derandomization** technique. Indeed, one can construct a family of colorings which is guaranteed to represent every k-path in G, while having cardinality polynomial in n and $|E|$. For instance, Naor *et al* [13] propose an explicit construct for a family of $e^k k^{\mathcal{O}(\log k)} \log n$ colorings that cover all possible occurrences of k-paths. Iterating the search for colorful path over the family yields an exact algorithm in overall time $\mathcal{O}((2e)^k k^{\mathcal{O}(\log k)}(n + |E|) \log n)$, using $\mathcal{O}(2^k n)$ memory.

Our method. Our pragmatic solution for MFEDOCK borrows heavily from the color coding framework with two peculiarities, guiding our design principles:

1. Instead of searching/counting occurrences of a k-path, we want to find the path minimizing the free-energy (optimization *vs* search);
2. Memory consumption should be limited, ideally strictly linear in n, in order to support fine-grained sampling of the poses (i.e. high values of n, very low value for k).

To address those goals, we further constrain the search to identify **well-colored paths**, k-paths that are not only colorful but also feature the colors $[1, k]$ in ascending order. The MFE among well-colored paths in a coloring κ can be computed recursively using the following trivial dynamic programming scheme:

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

$$\mathrm{mfe}_\kappa[v, m] = \begin{cases} E(v) & \text{if } m = k \\ \min\limits_{\substack{(v,v') \in E \text{ such that} \\ \kappa(v') = m+1 \\ \text{and } r(v') = r_{m+1}}} E(v) + E'(v, v') + \mathrm{mfe}_\kappa[v', m+1] & \text{otherwise.} \end{cases} \tag{3}$$

Following the computation of all $\mathrm{mfe}_\kappa[v, m]$ in $\mathcal{O}(k(n + |E|))$ time and $\mathcal{O}(kn)$ space, the MFE/well-colored complex can then be reconstructed using a standard backtrack reconstructs the (well-colored) optimal complex, in $\mathcal{O}(n + |E|)$ time. Derandomization can then be adapted to capture well-colored paths and, applying the DP algorithm to a well-chosen family of colorings, we get:

Theorem 1. MFEDOCK *can be solved exactly in* $\mathcal{O}(k^{k+\mathcal{O}(\log k)}(n + |E|))$ *time and* $\mathcal{O}(kn)$ *space.*

An elementary proof of this result can be found in Supplementary Section A. Meanwhile, the probability of finding the absolute MFE complex $x^\star$ for a random coloring κ now reduces to $1/k^k$, so the probability of finding $x^\star$ after considering α random uniform colorings becomes

$$\mathbb{P}(p \text{ well colored for some coloring} \mid \alpha) = 1 - \left(1 - \frac{1}{k^k}\right)^\alpha. \tag{4}$$

This equation can be inverted to target a given probability $(1 - \varepsilon)$ of finding the optimal path $x^\star$. We find that setting $\alpha = \lceil \log \varepsilon / \log(1 - k^{-k}) \rceil \in \mathcal{O}(|\log \varepsilon|\, k^k)$ is sufficient to guarantee such a property, from which we conclude:

Proposition 1. MFEDOCK *can be solved with probability* $1 - \varepsilon$ *in* $\mathcal{O}(n\, k^{k+1} |\log \varepsilon|)$ *time and* $\mathcal{O}(k\, n)$ *space.*

Practically, due to both the intricacies of implementing derandomization, and the empirical robustness of predictions, we implement and benchmark this probabilistic version in the rest of this work.

3 Equilibrium statistics

While clearly an important – computational 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 complex x inducing a nucleotide sequence r, is expected to follow a Boltzmann distribution:

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

and μ is proportional to the temperature. Since the number of valid complexes typically grows (at least) exponentially with k, the probability of the MFE complex will in fact remain abysmally small in larger systems.

Features in the Boltzmann distribution. This motivates a computation of **equilibrium statistics**, *i.e.* expected statistical properties of the system within the Boltzmann distribution. Such properties are measured by a set of real-valued **feature functions** $\{f_1, f_2, \ldots\}$, 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:

$$\mathbb{E}(f(X) \mid r) = \sum_{\substack{x \text{ self avoiding} \\ \text{and comp. with } r}} f(x)\, \mathbb{P}(X = x \mid r) \tag{5}$$

can be interpreted as a collective variables.Probabilities can also be computed as expectations of (0/1)-valued features. Higher moments of the distributions can be computed from the expectations of $f, f^2, f^3 \ldots$ giving access to finer characteristics of the distribution, such as its variance/stddev, skewness, kurtosis… or even correlations between multiple features.

Our estimator. Complexity-wise, computing the partition function is provably harder than the optimization problem addressed in Section 2. Indeed, setting the temperature β to a sufficiently-low value allows to determine the existence of a solution reaching a certain energy level, thereby solving the NP-hard decision version of the problem. Worse, derandomization cannot be easily adapted to compute partition functions or expectations. Indeed, the families of coloring produced in the context of optimization do not guarantee a uniform representation for all complexes, introducing a possible bias in the subsequent estimates.

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:

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

It can be estimated, first using DP to compute the (coloring-restricted) partition function:

$$\mathcal{Z}_\kappa = \sum_{\substack{v \in V \\ \text{such that } r(v) = r_1}} \mathcal{Z}_\kappa[v, 1] \tag{6}$$

$$\mathcal{Z}_\kappa[v, m] = \begin{cases} e^{-\beta E(v)} & \text{if } m = k \\ \displaystyle\sum_{\substack{(v,v') \in E \text{ such that} \\ \kappa(v') = m+1 \\ \text{and } r(v') = r_{m+1}}} e^{-\beta E(v)} \times e^{-\beta E'(v,v')} \times \mathcal{Z}_\kappa[v', m+1] & \text{otherwise.} \end{cases} \tag{7}$$

A stochastic backtrack then consists in choosing, starting from $\mathcal{Z}_k$, a term from the corresponding right-hand side with probability proportional to its contribution to left-hand side $\mathcal{Z}_\kappa$, keeping track of the sequence of

vertices and recursing until the $m = k$ condition is met. The returned random complex is then Boltzmann distributed, and the average value of f on a set of generated complexes represents an unbiased estimator for $\mathbb{E}(f(X) \mid r, \kappa)$. Our estimator takes a collection of random uniformly-distributed colorings, and returns:

$$\widehat{f}(\kappa_1, \kappa_2 \dots \kappa_M) = \frac{\sum_{i=1}^{M} \mathcal{Z}_{r,\kappa_i} \times \mathbb{E}(f(X) \mid r, \kappa_i)}{\sum_{j=1}^{M} \mathcal{Z}_{r,\kappa_j}} \tag{8}$$

Theorem 2. *The estimator $\widehat{f}$ is **consistent**. Let $K_1, \dots, K_M$ denote random uniform colorings, we have:*

$$\lim_{M \to +\infty} \mathbb{E}\left(\widehat{f}(K_1, K_2, \dots, K_M)\right) = \mathbb{E}(f(X) \mid r)$$

4 Extensions

4.1 Avoiding soft clashes

As mentioned in Section 2, soft clashes occur when poses are overly close, leading to a severe unstability of the complex. Soft clashes involving consecutive positions are avoided while building the poses compatibility graph, but non-consecutive poses may overlap. To restrict conformation spaces to **(fully) clash-free complexes**, avoiding both hard and soft clashes, we introduce two distinct solutions: i) a generation of suboptimal solutions to the MFEdock problem; and ii) a rejection-based sampling to compute equilibrium statistics within fully clash-free complexes.

Listing all suboptimal complexes. Since clash-free complexes represent a strict subset of the search space of MFEdock, the clash-free MFE complex can be found as a suboptimal solution. In other words, there exists an **energy tolerance** Δ_{Max} such that the fully clash-free MFE $x^\diamond$ can be found within Δ kcal.mol^{-1} of the hard clash-free MFE $x^\star$. To recover $x^\diamond$, we adapt the Waterman/Byers algorithm [20], which starts by computing the DP matrix in Equation (3), and uses the following backtrack:

$$\text{subopt}_\kappa(\Delta) \to \bigcup_{\substack{v \in V \text{ s.t.} \\ r(v) = r_1}} \text{subopt}_\kappa(v, 1; \Delta') \quad [\text{if } \Delta' := \Delta - (\text{mfe}_\kappa[v, 1] - \text{mfe}_\kappa) \geq 0] \tag{9}$$

$$\text{subopt}_\kappa(v, m; \Delta) \to \begin{cases} \{v\} & \text{if } m = k \\ \bigcup\limits_{\substack{(v,v') \in E \text{ s.t.} \\ \kappa(v') = m+1 \\ \text{and } r(v') = r_{m+1}}} \{v\} \otimes \text{subopt}_\kappa(v', m+1; \Delta') \quad [\text{if } \Delta' := \Delta - (E(v) + E'(v,v') + \text{mfe}_\kappa[v', m+1] - \text{mfe}_\kappa[v, m]) \geq 0] & \text{otherwise.} \end{cases} \tag{10}$$

We then call the backtrack function $\text{subopt}_\kappa(\delta)$ over increasing values of Δ, until at least one fully clash free is reported, and returns the fully clash-free complex having min free-energy. Since the modified backtrack is correct, it produces the exhaustive list of hard clash free complexes within Δ kcal.mol^{-1} of $x^\star$. It follows that, anytime this list of suboptimal complexes includes a fully clash-free complex, it also contains the fully clash-free MFE $x^\diamond$. Of course, the complexity of the approach grows exponentially with $(\Delta G(x^\diamond) - \Delta G(x^\star))$. Still, the efficiency of the algorithm still practically allows the consideration of large values of Δ, as shown in Section 5.1 and beyond.

Soft-clash aware statistics. In the context of estimating feature distribution, the consideration of complexes featuring soft clashes may bias the underlying statistics. Stochastic backtrack can be used to estimate the prevalence of soft clashes, and their potential to impact estimates. If they are frequent, and thus likely to bias estimated, then the estimator can be modified using a simple **rejection strategy**.

Namely, it can be shown that using the partition function/stochastic backtrack described in Equation 7, and rejecting those featuring soft clashes, generates a statistical sample of fully clash-free complexes. Indeed, coupling the stochastic backtrack with a rejection of soft clashing complexes, induces an emission probability of 0 for soft-clashing complexes, and $p(x)$ for a clash-free complex x such that:

$$p(x) = \frac{\mathbb{P}(X = x \mid \kappa, r)}{\sum_{x' \text{ w/o soft clashes}} \mathbb{P}(X = x' \mid \kappa, r)} = \frac{e^{-\beta\, \Delta G(x)}}{\mathcal{Z}_{\kappa,r}^{\text{clash free}}} \quad \text{where} \quad \mathcal{Z}_{\kappa,r}^{\text{clash free}} := \sum_{x' \text{ w/o soft clashes}} e^{-\beta\, \Delta G(x')}$$

so, for a given coloring κ, $\mathbb{E}(f(X) \mid r, \kappa, \text{clash free})$ can be estimated through the fully clash-free subsample.

Meanwhile, the value of frequency of $\mathcal{Z}_{\kappa,r}^{\text{clash free}}$ can directly be estimated from the frequency of soft clash-free complexes, which converges to $\mathcal{Z}_{\kappa,r}^{\text{clash free}}/\mathcal{Z}_{\kappa,r}$, noting that the colored partition function $\mathcal{Z}_{\kappa,r}$ is known. A consistent estimator for fully clash free complexes can then be computed as:

$$\widetilde{f}(\kappa_1, \kappa_2 \dots \kappa_M) = \frac{\sum_{i=1}^{M} \mathcal{Z}_{r,\kappa_i}^{\text{clash free}} \times \mathbb{E}(f(X) \mid r, \kappa_i, \text{clash free})}{\sum_{j=1}^{M} \mathcal{Z}_{r,\kappa_j}^{\text{clash free}}} \tag{11}$$

4.2 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 good affinity, or low interaction free-energy, towards the target protein or targeted pocket; **Negative design** forces the ligand to be specifically binding to a given region of the protein. Interestingly, both criteria are somewhat addressed by a simple relaxation of the MFEdock problem, simply disregarding the input ssRNA sequence in the sums of all algorithms without added complexity. Indeed, computing the MFE complex within a search space including both conformations and ssRNA sequences yields a sequence of poses $x^\star = (x_1, x_2, \dots, x_k)$ associated with a sequence of nucleotides $r^\star := r(x_1).r(x_2) \cdots r(x_k)$ such that:

1. No alternative sequence has higher affinity than $r^\star$ towards the protein (positive design);
2. 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 location of best complex is induced by the MFE criterion. Still, by generating suboptimals and only retaining the first occurrence of each sequence, our algorithms can produce a diversity of sequences that are both stable, and specifically target various sites.

5 Case studies

We implemented our algorithms for MFEdock (optimization; subopts; +/- sequence constraints) and statistical estimators into the `ColorDocking` software, a collection of `Python` scripts freely downloadable at https://gitlab.inria.fr/amibio/colordocking. Datasets and further information to reproduce experiments are also available. All experiments were performed on a PBS cluster with a Linux kernel. The used node has 92GB of memory, and each calculation was done on a single CPU.

Dataset. Four complexes were selected to test our method. Among them, three complexes represent the three prominent features of RNA-Binding Protein (RBP): RNA Recognition Motifs (RRM; 2XNR), Zinc fingers (5ELH) and KH2 domains (5WWX). These 3 specific structures were selected due to their low resolution (less than 2.0Å). Finally, these structures interact with ssRNAs whose short length make them good candidates for a rational design of therapeutic inhibitors. We also selected 1CVJ, despite its resolution being greater than 2.0 Å, since it features a longer ssRNA and was used by Chauvot de Beauchene *et al* [5].

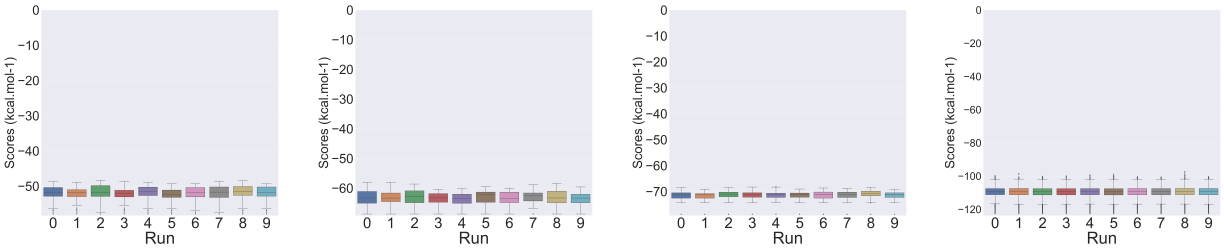
To provide a realistic setting for docking, proteins were prepared and minimised using the CHARMM27 force field in the absence of the ssRNA ligand and solvent. The resulting structures deviate from the bound models by less than 1.0 Å RMSD (between 0.78 and 1.0 Å; 0.48 Å for 1CVJ). As a result, the protein surface at the RNA binding site may be altered and in a more specific way at some specific position of the RNA chain. Therefore, we cannot expect to find solutions that fit the experimental structure of the bound RNA at high accuracy.

5.1 Optimization for docking

Robustness analysis. For each case study, we used the Multi-Copy Simultaneous Search (MCSS) [12] method to generate distributions of about 10.000 fragment poses composed of Adenine (A), Cytosine (C), Guanine (G) and Uracil (U). Only first 2.000 fragment poses of each nucleotide were used for this study. From these, a python package called NUCLEotide AssembleR (NUCLEAR) (to be released) was used to cluster similar

PDB ID	Length k	Tolerance ε	ΔMax	Step	Target Sequence	α
2XNR	3nts	0.01	50	1	UCU	122
5ELH	3nts	0.01	50	1	UUA	122
5WWX	3nts	0.01	50	1	AGA	122
1CVJ	5nts	0.01	50	1	AAAAA	14 388

(a) Parameter values used over the various experiments.



(b) 2XNR (c) 5ELH (d) 5WWX (e) 1CVJ

Fig. 2: Reproducibility of fully clash-free suboptimal structures. Parameters (a) and distributions of Free-energy scores (y-axis) over 10 runs (x-axis) for each RBP (b-e). The same fully clash-free MFE was obtained for all independent runs.

poses with a threshold of 0.5 Å, and to build a connectivity matrix using a 6 Å cutoff for the O3'-C5' distance. The other parameters of NUCLEAR were left to their default values.

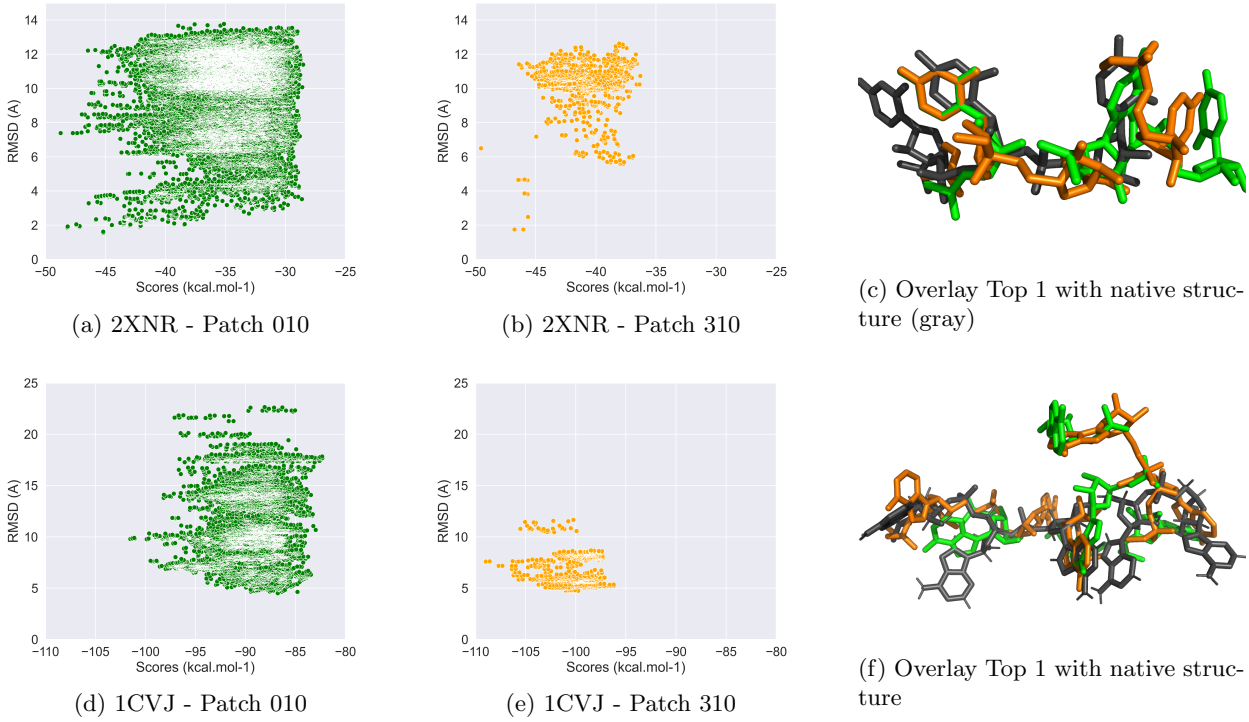
A directed graph was then generated and fed as input to an implementation of our suboptimal color coding algorithm, using parameters described in Subfigure 2a. In all case studies, we set the tolerance $\varepsilon = 0.01$, *i.e.* the hard clash-free MFE complex was predicted with probability of $p = 99\%$. Fully clash-free MFE candidates were produced, setting an initial tolerance $\Delta := 0$, by: i) Generating Δ suboptimal complexes; ii) Discard those featuring soft clashes; iii) If any are left, return them, otherwise increase Δ by 1, stopping when Δ exceeds a value $\Delta_{\max}$ (set to 50 kcal.mol^{-1}).

Figure 2 summarizes the energy distribution of fully clash-free complexes, obtained over 10 independent runs of each case study. Two observations stand out: First, for a given complex, the fully clash-free MFE complex was always found across independent runs, indicating that the filtering of soft clashing complexes does not majorly impact reliability; Second, as expected, sub-optimal solutions can be observed to vary across independent runs. Indeed, the value of Δ needed to generate a clash-free complex may vary across random executions, leading to the production of different number of sub-optimal complexes. However, the bulk of the distributions remains largely consistent across runs, and fairly concentrated.

Docking. To illustrate the docking, we considered two different distributions, generated by MCSS for nucleotide fragment types as described in Chevrollier's PhD thesis [4]. Their main difference resides in the patch applied on the C5' extremity: '010' fragments have a O5'-PO_2- group, while the '310' fragments have a O5'-CH_3PO_3' group. For each case study, we restricted the docking to nucleotides occurring in the target sequence, producing 2000 fragment poses per nucleotide. The connection matrix was built as described in the previous paragraph, except that the O3'-C5' distance cutoff was set to 4.5 Å. Other parameters were set as per Table 2a, except for Δ_{Max} which was set to 30 with a increment step 30.

To compare the solutions with the native complex, a Root-Mean-Square Deviation (RMSD) was calculated based on 16 equivalent atoms present in the pyrimidines and purines [4]. Results show that the best model (lowest RMSD) sometimes correspond to the best scores: this is the case for 2XNR (-45.2 kcal.mol^{-1} for '010' and -46.8 kcal.mol^{-1} for '310') and 5WWX (-54.4 kcal.mol^{-1} for '310' and -63.6 kcal.mol^{-1} for '310'). Nevertheless, for 1CVJ, the best RMSD complex has a score of -100.14 kcal.mol^{-1}, 12 kcal.mol^{-1} higher than the MFE complex (-112.65 kcal.mol^{-1}). Furthermore, we can see that the MFE and best RMSD do not fit completely with the experimental structure, where some positions deviate.

These discrepancies can be explained by the preprocessing of the protein, a minimization having been applied without the RNA ligand. A comparison between the bound and minimized structure show a deviation



(a) 2XNR - Patch 010

(b) 2XNR - Patch 310

(c) Overlay Top 1 with native structure (gray)

(d) 1CVJ - Patch 010

(e) 1CVJ - Patch 310

(f) Overlay Top 1 with native structure

Fig. 3: Distributions of free-energy/scores and RMSDs against native complex for two poses distributions (010/green: a and d; 310/orange: b and e). Subfigures (c), (f) represent geometric superimposition of MFE structures for '010' (green) and '310' (orange) against native complex (gray).

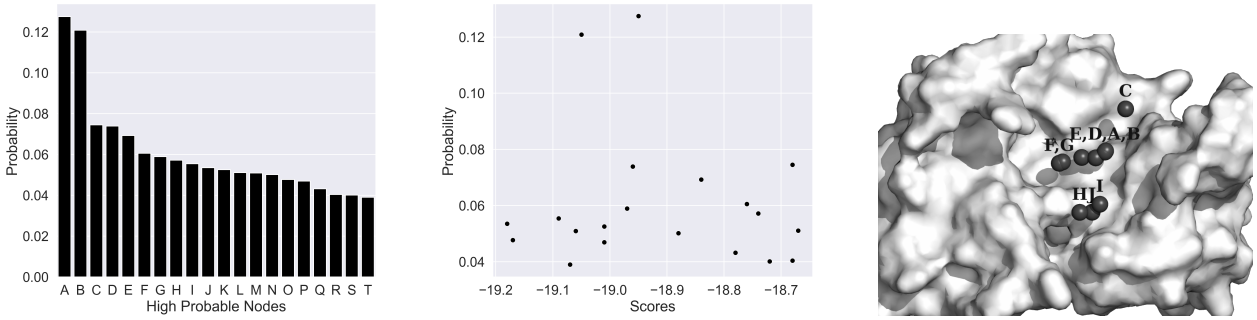
for some binding site, creating a loss of native interactions with the RNA [7]. The bias introduced by the minimisation has an influence on the quality of the sampling, with a variation of the number of native-like poses (within 2 Å of experimental structure). Indeed, an analysis of the '010' and '310' distributions showed there are no or few native-like poses for some positions. Therefore, some native or near native solutions are simply absent from the initial pose set, motivating the consideration of more dense sampling in the future.

5.2 Equilibrium statistics identify highly probable poses and sequence profiles

A design of specific high-affinity molecules can be driven by information on the structure/sequence ensemble, obtained using estimators introduced in Section 3. We studied two sets of features that: i) identify of highly-probable poses; and ii) characterize the sequence profile induced by low energy complexes. The first corresponds to a set of binary feature functions $\{f_v\}_{v \in V}$, each returning 1 if the complex uses v and 0 otherwise, so that the expectation coincides with the probability of v. Similar features allows to identify the nucleotide frequency at each position in the design setting.

We considered 1CVJ, and built a connection matrix with NUCLEAR based on the '310' distribution of poses, restricted to free-energies below -18.74 kcal.mol^{-1} (6262 poses left; 32% A/41%G/15%C/12%U). A clustering of poses (0.5 Å cutoff) was performed, and nucleotides were connected based using a O3'-C5' distance cutoff of 4.0Å. We considered a 5-mers, with 1% tolerance (expected distance to true mean). The pseudo-temp. β was set to -1.

Highly probable poses. Figure 4a shows the 20 most probable pose. We can see that two poses (A and B) have probability greater than 10%. The 18 others poses have around 5% probability. Interestingly, the locations of A and B are virtually indistinguishable on the protein surface (Figure 4c) and both correspond to mutually-exclusive Adenines. A complex thus has 24% probability to pass through this area. This observation is not trivial, and stems from a delicate trade-off between the energies of poses and the combinatorics of complexes.



(a) Probability of poses (Top 20) (b) Free-energy versus proba (Top 20) (c) 1CVJ positioning (Top 10)

Fig. 4: Statistical study of highly probable poses. (a) shows the top 20 probabilities of the poses where path most often pass (b) shows scores of the top 20 according to their probability (c) shows the position of the center of mass of the top 10 highly probable nodes.

In particular, A and B are each twice more probable than the other poses in the top 20, despite having similar free-energies contributions (Figure 4b). This again shows that non-trivial insight can be provided by equilibrium statistics.

Sequence profile. At each position of the 5-mer, we estimated the probability look what is the nucleotide the most probable. The resulting sequence profile (Supp. Figure 1) reveals a predominance of Guanine or Adenine at each position. This is unsurprising as those nucleotides are dominant in the filtered poses set (42% G, 31% A). More generally, we observe a correlation between the number of poses for a nucleotide and its probability. However, the probabilities of C (15% poses) and U (12% poses) show an interesting trend, with C being more probable than U for the first 2 positions of the 5-mer (+4% and +3.7%) and then being increasingly dominated for the 3 remaining positions (-1.4%, -7.7% and -5.3%). This confirms the capacity of the estimator to reveal non-trivial cooperative effects.

5.3 Design

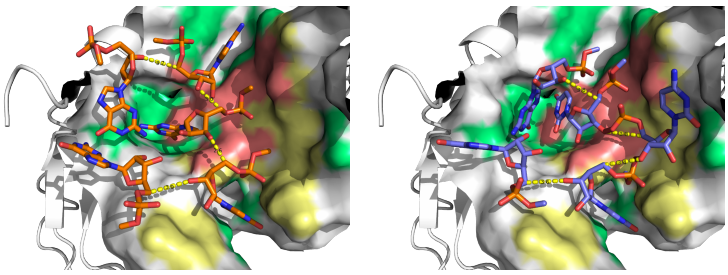
For the design, 2000 poses were generated for each nucleotide, clustered at 0.5Å RMSD, to get the connectivity matrix. A max O3'-C5' distance was set to either 4 or 5Å. All others parameters are described in Figure 5a.

Runtime. Design is a fundamental asset of the fragment-based approach. We designed oligonucleotides of lengths ranging from 4 to 7 with varying parameters, as shown in Figure 5a. Two times were measured: the time to get the absolute first MFE (MFE), and time needed to get the suboptimal complex (Subopts) for a given Δ_{Max}, the maximum energy between the worst desired suboptimal complex and the MFE. In this case study, for each coloring, a suboptimal complex was generated for increasing values of Δ, using increment of 1 kcal.mol^{-1}. For the MFE time, the number of poses and the O3'-C5' distance impact the computational time. This observation is expected because both parameters impact the number of connectable fragments. Nevertheless, the time needed to find the MFE remains reasonable: 29.5 hours to design a 7-mer from 8000 poses using a 4 Å O3'-C5' cutoff (5.10^{10} complexes), and 61 hours using a 5 Å O3'-C5' cutoff (3.10^{14} complexes), this despite an unoptimized proof-of-concept implementation in `Python`. The suboptimal runtime is also highly dependent on the value of Δ_{Max}, having in theory exponential impacton the worst-case complexity. This impact is confirmed empirically, nevertheless allowing the design of fully clash-free 6 mers.

Diversity of solutions. Suboptimals aim to generate set of diverse nucleotide sequence. In this example, the 4 nucleotides (A,C,G,U) were used to design oligonucleotides. From 8000 poses and a O3'-C5' distance equal to 4Å, the number of unique sequences is 42 for a 5-mer, and 35 for a 6-mer out of 175 produced suboptimal solutions. The same sequences can be generated (e.g. ACUGG), but the poses may differ by a few positions or more. However, between two complexes inducing the same sequences, the conformation with the best

k	#Poses	Δ_{Max}	α	Time (min.) MFE	Subopts	O3'-C5' (Å)
5	8 000	10	14 388	5.60	59.85	4
6	8 000	10	214 856	95.48	893.90	4
7	8 000	–	214 856	1 774,75	–	4
5	8 000	10	14 388	12.38	128.38	5
6	8 000	10	214 856	189.37	2581.40	5
7	8 000	–	214 856	3 656,61	–	5
5	8 000	20	14 388	5.87	128.12	4
6	8 000	20	214 856	93.25	2159.62	4

(a) Runtime analysis 1CVJ



(b) GAGGG (-117.37 kcal.mol $^{-1}$) (c) GUCGG (-102.85 kcal.mol $^{-1}$)

Fig. 5: Runtime analysis and 3D representation of binding pockets for two designed 5-mers inspired by statistical estimates of the sequence space.

free energy will in theory be preferred. Nevertheless, this will allow the user to check the different possible conformations with the associated score, and potentially to consider chemical optimizations, or chose the ones targeting a desire site. For instance, the 42 unique 5-mers generated by the method differed on average by 9.5 Å RMSD (10.5 Å avg RMSD for the 35 unique 6-mers). Finally, it was possible to select a collection of 15 5-mers having at least 5 Å pairwise RMSD between their MFE complexes (7 for 6-mers). We even found four 5-mers (and 6-mers) having pairwise RMSD greater that 10 Å, demonstrating the capacity of our method to produce sequence targeting diverse regions.

Example of a design based on feature estimation. One strategy for design can be to exploit the results of a statistical study (see Section 5.2) to design a 5-mer MFE. From the sequence profile, we searched a 5-mer whose sequence is composed of either a Guanine or Adenine at each position. As seen in Figure 5, the MFE sequence found with this strategy is GAGGG (Figure 5b). It is interesting to note that this MFE corresponds to the same sequence and conformation as obtained using a blind strategy (i.e. w/o sequence constraint). The energy score is equal to -117.37 kcal.mol $^{-1}$. Another MFE was designed from the highly probable poses, featuring a Uracil at a position where the frequency of this fragment is equal to 7% (Figure 5c). The associated free-energy is -102.85 kcal.mol $^{-1}$. These results show that statistical properties can be exploited to suggest different designs, and compare their conformations and energies. For instance, the GAGGG conformation shown in Figure 5b would have better overall affinity than the GUCGG conformation shown in Figure 5c. But interaction areas are not really the same, and our method provides key information on the preferred interactions between a sub-region and a particular nucleotide (or family of nucleotides).

6 Conclusions and perspectives

We have introduced a new algorithmic framework, 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 asset 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, *e.g.* trinucleotides utilized by Chauvot de Beauchène *et al* [5].

Another open left open is the existence of exact/efficient treatment for soft clashes, which could greatly benefit from being revisited in an exact algorithmic setting. However the relevant graph problem would then need to capture generalized notions of incompatibility (beyond self-incompatibility/avoidance), and would require solving some optimization problem over bounded independent sets. Lastly, the number β of stochastic backtracks is currently based on pessimistic estimates (stddev $\approx$ range/2), and could be refined to account for the tight energy distributions of poses (*e.g.* 16 kcal.mol^{-1} range vs 2.2 stddev in design studies).

Acknowledgments

The authors are greatly indebted to Isaure Chauvot de Beauchêne, with whom some of the authors attempted a prior – unfortunately unsuccessful – application of color coding techniques to fragment-based design, and to Laurent Bulteau for suggesting well-colored paths as a memory-efficient alternative to colorful paths.

References

1. N. Alon, P. Dao, I. Hajirasouliha, F. Hormozdiari, and S. C. Sahinalp. Biomolecular network motif counting and discovery by color coding. *Bioinformatics*, 24(13):i241–i249, jun 2008.
2. 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.
3. G. Bollag, J. Tsai, J. Zhang, C. Zhang, P. Ibrahim, K. Nolop, and P. Hirth. Vemurafenib: the first drug approved for BRAF-mutant cancer. *Nature Reviews Drug Discovery*, 11(11):873–886, oct 2012.
4. 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.
5. 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.
6. B. Dost, T. Shlomi, N. Gupta, E. Ruppin, V. Bafna, and R. Sharan. QNet: A tool for querying protein interaction networks. *Journal of Computational Biology*, 15(7):913–925, sep 2008.
7. R. González-Alemán, N. Chevrollier, M. Simoes, L. Montero-Cabrera, and F. Leclerc. MCSS-based predictions of binding mode and selectivity of nucleotide ligands. *Journal of Chemical Theory and Computation*, 17(4):2599–2618, mar 2021.
8. D. Hall, S. Li, K. Yamashita, R. Azuma, J. A. Carver, and D. M. Standley. 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.
9. A. Itai, C. H. Papadimitriou, and J. L. Szwarcfiter. Hamilton paths in grid graphs. *SIAM Journal on Computing*, 11(4):676–686, nov 1982.
10. 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.
11. 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.
12. 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.
13. 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.
14. T. P. Perera, E. Jovcheva, L. Mevellec, J. Vialard, D. D. Lange, T. Verhulst, C. Paulussen, K. V. D. Ven, P. King, E. Freyne, D. C. Rees, M. Squires, G. Saxty, M. Page, C. W. Murray, R. Gilissen, G. Ward, N. T. Thompson, D. R. Newell, N. Cheng, L. Xie, J. Yang, S. J. Platero, J. D. Karkera, C. Moy, P. Angibaud, S. Laquerre, and M. V. Lorenzi. 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.
15. J. Schoepfer, W. Jahnke, G. Berellini, S. Buonamici, S. Cotesta, S. W. Cowan-Jacob, S. Dodd, P. Drueckes, D. Fabbro, T. Gabriel, J.-M. Groell, R. M. Grotzfeld, A. Q. Hassan, C. Henry, V. Iyer, D. Jones, F. Lombardo, A. Loo, P. W. Manley, X. Pellé, G. Rummel, B. Salem, M. Warmuth, A. A. Wylie, T. Zoller, A. L. Marzinzik, and P. Furet. 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.
16. A. Schuffenhauer, S. Ruedisser, A. Marzinzik, W. Jahnke, P. Selzer, and E. Jacoby. Library design for fragment based screening. *Current Topics in Medicinal Chemistry*, 5(8):751–762, aug 2005.
17. 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.
18. A. J. Souers, J. D. Leverson, E. R. Boghaert, S. L. Ackler, N. D. Catron, J. Chen, B. D. Dayton, H. Ding, S. H. Enschede, W. J. Fairbrother, D. C. S. Huang, S. G. Hymowitz, S. Jin, S. L. Khaw, P. J. Kovar, L. T. Lam, J. Lee, H. L. Maecker, K. C. Marsh, K. D. Mason, M. J. Mitten, P. M. Nimmer, A. Oleksijew, C. H. Park, C.-M. Park, D. C. Phillips, A. W. Roberts, D. Sampath, J. F. Seymour, M. L. Smith, G. M. Sullivan, S. K. Tahir, C. Tse, M. D. Wendt, Y. Xiao, J. C. Xue, H. Zhang, R. A. Humerickhouse, S. H. Rosenberg, and S. W. Elmore. ABT-199, a potent and selective BCL-2 inhibitor, achieves antitumor activity while sparing platelets. *Nature Medicine*, 19(2):202–208, jan 2013.

19. W. D. Tap, Z. A. Wainberg, S. P. Anthony, P. N. Ibrahim, C. Zhang, J. H. Healey, B. Chmielowski, A. P. Staddon, A. L. Cohn, G. I. Shapiro, V. L. Keedy, A. S. Singh, I. Puzanov, E. L. Kwak, A. J. Wagner, D. D. V. Hoff, G. J. Weiss, R. K. Ramanathan, J. Zhang, G. Habets, Y. Zhang, E. A. Burton, G. Visor, L. Sanftner, P. Severson, H. Nguyen, M. J. Kim, A. Marimuthu, G. Tsang, R. Shellooe, C. Gee, B. L. West, P. Hirth, K. Nolop, M. van de Rijn, H. H. Hsu, C. Peterfy, P. S. Lin, S. Tong-Starksen, and G. Bollag. Structure-guided blockade of CSF1r kinase in tenosynovial giant-cell tumor. *New England Journal of Medicine*, 373(5):428–437, jul 2015.
20. 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.