Computational investigations of structure probing experiments for RNA structure prediction

Doctoral Thesis submitted to the

International School For Advanced Studies

SISSA

Author:
Nicola Calonaci

Supervisor:
Giovanni Bussi

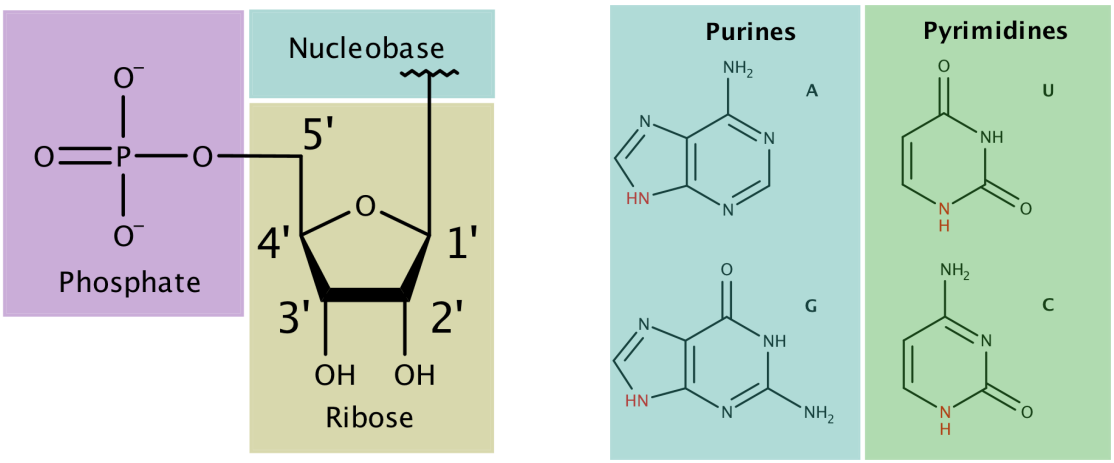
December 14, 2020

Contents

Chapter 1

Introduction

Ribonucleic acids (RNA) transcripts, and in particular non-coding RNAs, play fundamental roles in cellular metabolism, as they are involved in protein synthesis [1], catalysis [2], and regulation of gene expression [3]. In some cases, an RNA's biological function is mostly dependent on a specific active conformation [4], making the identification of this single stable structure crucial to identify the role of the RNA and the relationships between its mutations and diseases [5]. On the other hand, RNAs are often found in a dynamic equilibrium of multiple interconverting conformations, that is necessary to regulate their functional activity. In these cases it becomes fundamental to gain knowledge of RNA's structural ensembles, in order to fully determine its mechanism of action. The current structure determination techniques, both for single-state models such as X-ray crystallography [6], and for multi-state models such as nuclear magnetic resonance [7] and single-molecule methods [8], despite proving accurate and reliable in many cases, are extremely slow and costly. In contrast, chemical probing [9] is a class of experimental techniques that provide structural information at single-nucleotide resolution at significantly lower costs in terms of time and required infrastructures. In particular, selective $2'$ hydroxyl acylation analyzed via primer extension (SHAPE) [10, 11] has proved a valid chemical mapping technique to probe RNA structure even *in vivo* [12]. This thesis reports a systematic investigation of chemical probing experiments based on two different approaches. The first approach, presented in Chapter 2, relies on machine-learning techniques to optimize a model for mapping experimental data into structural information. The model relies also on co-evolutionary data, in the form of direct coupling analysis (DCA) couplings. The inclusion of this kind of data is chosen in the same spirit of reducing the costs of structure probing, as co-evolutionary analysis relies only on sequencing techniques. The resulting model is proposed as a candidate standard tool for prediction of RNA secondary structure, and some insight in the mechanism of chemical probing is gained by interpreting back its features. Importantly, this work has been developed in the perspective of building a framework for future refinement and improvement. In this spirit, all the used data and scripts are available at `https://github.com/bussilab/shape-dca-data`, and the model can be easily retrained and adapted to incorporate arbitrary experimental information. As the interpretation of the model features suggests the possible emergence of cooperative effects involving RNA nucleotides interacting with SHAPE reagents, a second approach based on Molecular Dynamics simulations is proposed to investigate this hypothesis. The results, along with an originally developed methodology to analyse Molecular Dynamics simulations at variable number of particles, are presented in Chapter 3. A brief introduction to the main theoretical and methodological aspects involved in the presented investigations is given in the following sections.



(a) (b)

Figure 1.1: Building blocks of RNA: (a) the nucleotide, RNA monomer made of a phosphate group, a ribose sugar and a nucleobase that can either be (b) a purine (common ones are Adenine or Guanine) or a pyrimidine (common ones are Uracyl or Cytosine). In red are shown the binding sites where nucleotides bind to the ribose.

1.1 RNA structure

RNA is a polymeric molecule built up from four common monomeric units (modified monomers exist), the ribonucleotide residues [13]. These building blocks are rather similar to each other, compared to the 20 amino acid residues in proteins. Each nucleotide is composed of a nucleobase, a ribose sugar and a phosphate group (Fig.1.1a). The ribose, to which the nucleobase is bound through a glycosidic bond, is a monosaccharide sugar in the form of a closed ring containing 5 carbon atoms, numbered from C1′ to C5′. Each ribose can adopt two different conformations, called sugar puckers: if the C2′ faces toward the nucleobase, then the ribose is in the C2′-*endo* conformation, whereas it is in the C3′-*endo* if the C3′ faces toward the nucleobase. Nucleobases can be either purines (the two common ones are **A**denine and **G**uanine) or pyrimidines (the common ones are **U**racil and **C**ytosine), as shown in Fig.1.1b. The main structural difference between the four nucleobases is the position of the carbonyl (-C=O) and of the amino ($-NH_2$) groups.

Nucleotides link together through a phosphodiester bond in linear sequence as sketched in Fig.1.2, resulting in the RNA polymeric chain which is typically referred to as *primary structure* of RNA or simply RNA *sequence*. An RNA sequence is usually identified through the sequence of letters that indicate the corresponding nucleotides linked along the chain, in order from the 5′ carbon of the first ribose to the 3′ carbon of the last ribose.

Nucleobases interact with each other through two types of interactions: base stacking and hydrogen bond. Due to their planar shape, nucleobases whose planes are close enough (≈ 3.5 Å) tend to exclude water molecules maximizing the Van der Waals interactions between each other. The contribution of base stacking to the structural stability of an RNA is the largest. Two nucleobases that contain complementary arrays of polarized atoms can form a hydrogen bond. Each nucleobase contains three different such arrays [14], corresponding to three different edges of the nucleobase plane: the Watson-Crick (W), the Hoogsteen (H) and the sugar (S) edge, see Fig.1.3. A hydrogen bond can thus be formed between two bases, involving 6 different pairs of

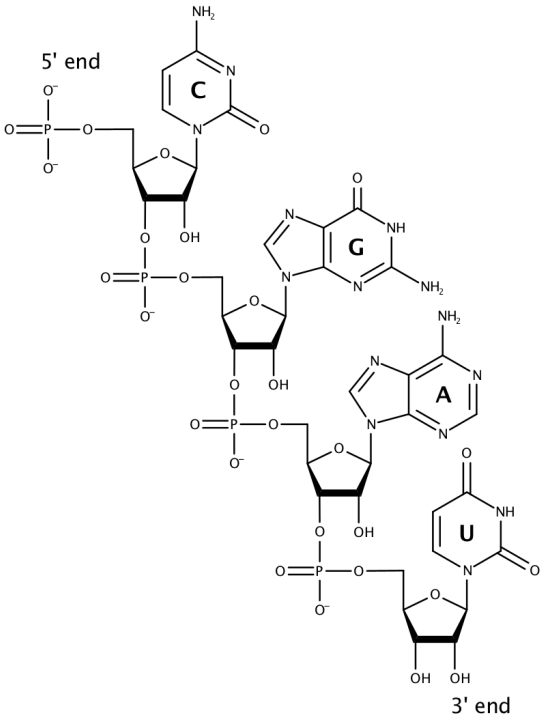


Figure 1.2: A sketch of the RNA primary structure (or RNA sequence), enlightening the polymeric nature of RNA. A sequence containing all the 4 common nucleotides is shown, and would be identified as `CGAU`. Nucleotides are connected each to the next one through a phosphodiester bond. The part of the primary structure containing riboses and phosphate groups consitutes the backbone of the RNA. The high flexibility of RNA backbone allows non-contiguous nucleobases to pair and contiguous nucleobases to stack with each other, building up RNA secondary structure.

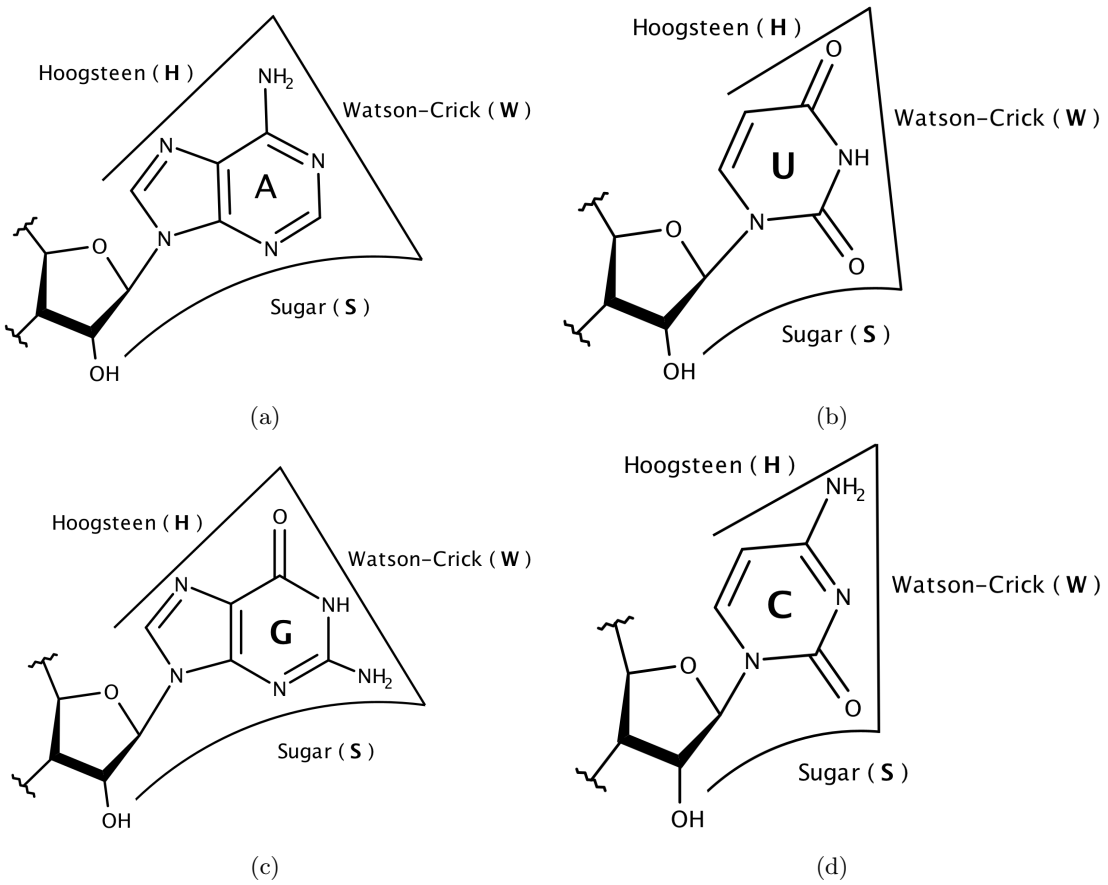


Figure 1.3: The interaction edges of the 4 common RNA nucleobases. Each nuclebase can come in the interaction range of another, and different types of base pairs can be formed depending on which of Watson-Crick (W), Hoogsteen (H) or sugar (S) edges are in contact.

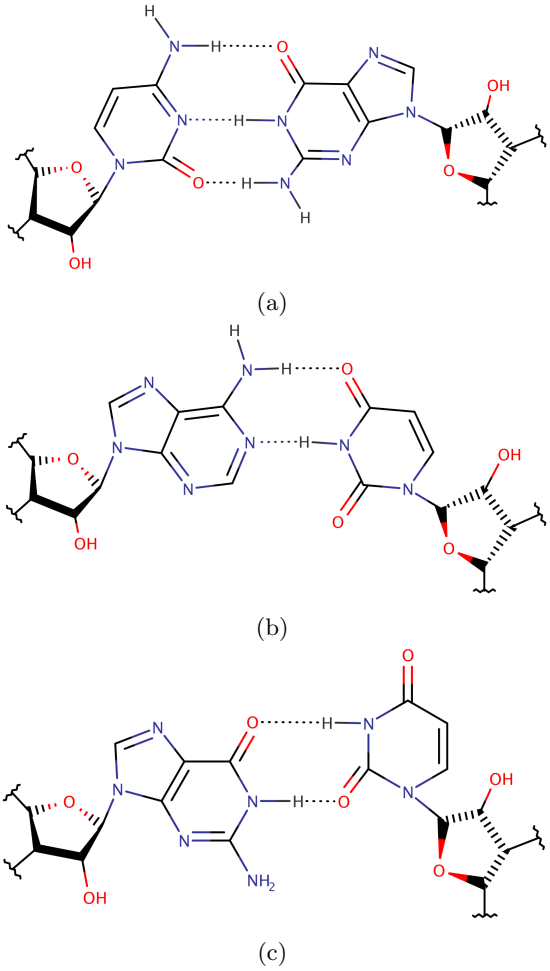


Figure 1.4: The most frequent RNA base-pairs, all cWW type: (a) the canonical CG pair, (b) the canonical AU and (c) the wobble pair GU. All the base pairs involve hydrogen-bond interactions.

edges (WW, WH, WS, HH, HS, and SS). Moreover, bases can interact in either of two orientations with respect to the glycosidic bonds, *cis* (c) or *trans* (t) relative to the hydrogen bonds, making up to 12 possible combinations (cWW, tWW, cWH *etc*). All these possible hydrogen bonding interactions bewteen nucleobases also go under the name of *base pairing*. The most frequently observed base pairs are the so called canonical cWW A-U and C=G pairs, and the wobble cWW G-U pair, reported for example in Fig.1.4.

Since the backbone of the RNA, made of the riboses and phosphate links, is highly flexible, a single strand of RNA can fold on itself in stable conformations, in which some bases are paired and/or stacked. The set of base pairs contained in a specific conformation of an RNA molecule is referred to as *secondary structure*. In principle, an RNA molecule can be found in multiple conformations, with different secondary structures. In many cases an RNA exists in a dynamic equilibrium of a limited number of stable secondary structures, while in some other cases a single structure dominates among the others and it is referred to as the *native structure* of the molecule.

The alternation of base-paired and unpaired nucleotides in RNA secondary structure gives rise

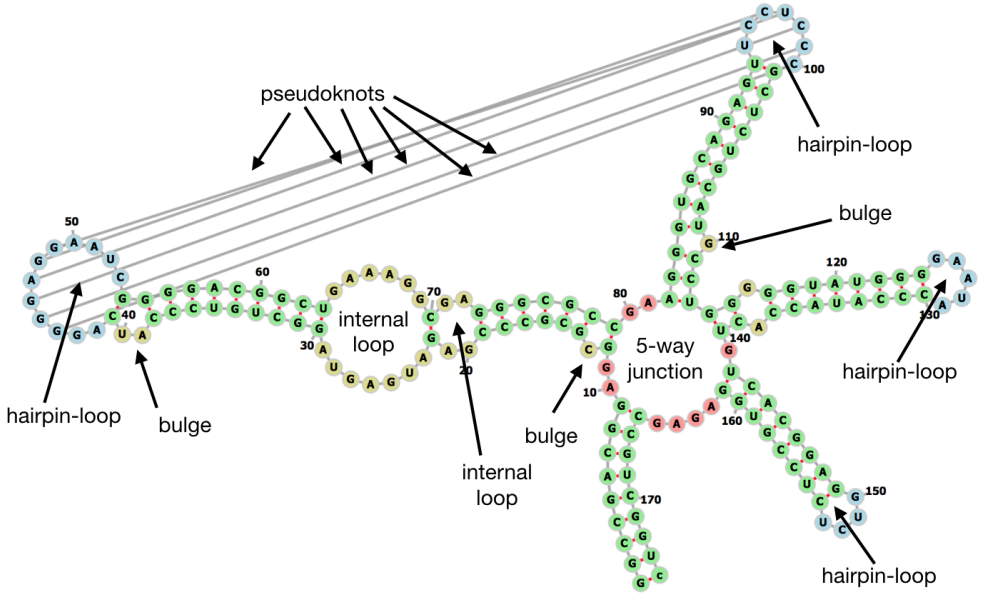


Figure 1.5: Secondary structure of the Lysine riboswitch, as extracted via `x3dna-dssr` [16] from the 3DIG entry of PDB. The different types of loops are indicated as well as pseudoknots. Stems are represented in green.

to a variety of recurrent structural elements [15]. A representative secondary structure, drawn from the X-RAY diffraction of a sample of crystallized Lysine riboswitch (from PDB 3DIG) is reported in Fig.1.5. The most obvious structural element is the *helix*, consisting of at least two stacked base pairs. A helix that contains only canonical Watson-Crick or wobble pairs and whose backbone is continuous in both strands is a *stem*. It happens fairly often that the sequence of stacked base pairs forming a stem is interrupted by one or more unpaired nucleotides, on either one or both strands, giving rise to *loops*. Loops are important structural elements, especially for their role in guiding folding, stabilizing the local structure, providing recognition sites for amino acids in translation and for RNA-binding proteins, and even as a substrate for enzymatic reactions. Different types of loops exist: a *bulge loop* is formed when the sequence of base-paired nucleotides is interrupted by one or more unpaired nucleotides on only one of the paired strands; when two parts of a stem are separated by a loop then an *internal loop* is formed, whereas if a stem terminates with a loop we have an *hairpin loop*; it may happen also that a loop branches a stem into two or more, in which case the loop is referred to as an *N-way junction*, N being the number of stems that "flow"into and from the loop.

Due to the rotational freedom of its backbone, RNA can fold up into more complex local structures involving the secondary structure elements described above. The set of interactions involving two helices, two unpaired regions or a helix and an unpaired region is referred to as the *tertiary structure* of RNA. For example, two contiguous helices can stack on each other, or two distant helices can fit in each other grooves. A stretch of loop with one or more nucleotides can pair with a complementary single-stranded stretch of sequence outside the loop, giving rise to a complex element called *pseudoknot*. Pseudoknots are a clear example of the high biological significance of the tertiary structure, as for example they are crucial for the activity of the RNA component of the telomerase [17], or in coronaviruses where a pseudoknot acts as stimulation element for ribosomal frameshifting [18], a critical process for the infiltration in the

host cell. Pseudoknots and other types of tertiary interactions provide a serious obstacle to many algorithms for RNA structure prediction.

A hypothesis that is often made in RNA folding studies is that the folding is hierarchical [13], meaning that the primary structure determines the secondary, that in turn guides the tertiary folding. It is also clear that this hypothesis is an oversimplification [13]: even if the secondary structure provides the major free energy contribution to the ensemble of possible structure conformations of an RNA, certain tertiary structure elements can stabilize a suboptimal secondary structure enough to make it dominate the population distribution of the ensemble. Nonetheless, in addition to being an important intermediate step in the approximately hierarchical folding of RNA, the secondary structure also can be informative by itself, as specific secondary structural elements have important biological roles and in general the parsing of an RNA into base-paired and unpaired regions can highlight differences in the accessibility of different parts of the sequence. For this reason, this work is focused on the problem of secondary structure prediction, and tertiary interactions are no further considered in detail in the next sections.

1.2 Secondary structure prediction

1.2.1 Thermodynamic models

A thermodynamic model for RNA secondary structure prediction consists in a set of equations and parameters that define the conformational stability of a specific secondary structure in which a specific RNA sequence can fold [19]. For a specific sequence, the conformational stability of a folded structure is defined as the Gibbs free energy change ΔG with respect to the completely unfolded conformation. The folding free energy is decomposable into the sum of an enthalpic (ΔH) and an entropic (ΔS) contribution:

$$\Delta G = \Delta H - T\Delta S \tag{1.1}$$

where T is the temperature at which the folding stability is estimated. The enthalpy gain depends on the favorable interactions (base pairs, base stacks *etc.*) formed in the structure, while the entropy loss is contributed by the freedom of the unpaired nucleotides in loops to sample multiple local conformations. Once the model free energy is fixed, the most stable secondary structure of a given sequence can be predicted as the one with minimum free energy (MFE structure) [20, 21]. In principle also the whole ensemble of possible structures for a given sequence $s\vec{e}q$ can be derived, by computing the partition function of the ensemble [22, 23, 24]

$$Z\left(s\vec{e}q\right) = \sum_{\{\vec{s}\}} e^{-\Delta G(\vec{s})/k_b T} \tag{1.2}$$

where the sum runs over all the possible structures for the sequence $s\vec{e}q$, k_b is the Boltzmann constant and $e^{-\Delta G(\vec{s})/k_b T}$ is the Boltzmann weight assigned to structure $\vec{s}$. The calculation of the partition function enables to compute the ensemble population of a structure relative to the others, that can be written as the probability of a structure $\vec{s}$, conditioned by sequence $s\vec{e}q$ that is fixed:

$$P\left(\vec{s}|s\vec{e}q\right) = \frac{1}{Z} e^{-\Delta G(\vec{s})/k_b T} \tag{1.3}$$

This population is related to the relative concentration that is expected to be measured for that structure in a pool of RNA molecules with the same sequence: the ratio between the concentrations $[C_1]$ and $[C_2]$ of two structures with populations $P\left(\vec{s_1}|s\vec{e}q\right)$ and $P\left(\vec{s_2}|s\vec{e}q\right)$ is

$$\frac{[C_1]}{[C_2]} = \frac{P\left(\vec{s_1}|s\vec{e}q\right)}{P\left(\vec{s_2}|s\vec{e}q\right)} = e^{(\Delta G_1 - \Delta G_2)/k_B T} \tag{1.4}$$

where $\Delta G_1 = \Delta G(\vec{s_1})$ and $\Delta G_2 = \Delta G(\vec{s_2})$ to simplify the notation. Along with ensemble populations of structures, the base pairing probability for any couple of complementary nucleotides in the sequence can be computed as the average occurrence of that pair in the structure ensemble, weighted with the population of structures containing the pair:

$$p_{ij} = \sum_{\{\vec{s}\}} s_{ij} \cdot P(\vec{s}|s\vec{e}q) \tag{1.5}$$

where s_{ij} is the matrix of base pairs that are present in structure $\vec{s}$ ($s_{ij} = 1$ if nucleotide i is paired to nucleotide j, and $s_{ij} = 0$ otherwise). Moreover, once the partition function is computed, stochastic sampling of structures can be carried out [25].

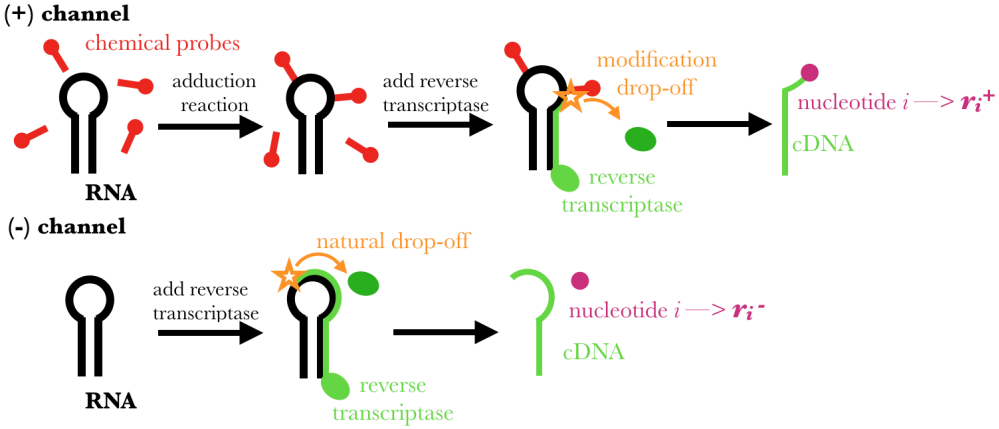
The currently best performing thermodynamic models for RNA secondary structure prediction are based on a set of nearest-neighbor parameters. These parameters control the free energy changes with respect to the unfolded state, that are introduced by the formation of local structural elements (base-pairs and loops). Models are of nearest-neighbor type because they are based on the following assumptions [19]: the free energy change due to a structural element depends only on the sequence of that structural element and on the sequence of the immediately adjacent base pairs; the eventual total free energy change is an additive quantity, *i.e.* it results from the sum of all the local changes. Given a sequence $s\vec{e}q$ and the set of nearest neighbor parameters $\{\theta\}$, the folding free-energy of Eq.1.1 for a structure $\vec{s}$ is thus estimated as the sum of the energy change contributions ΔG_i of the p structural elements contained in $\vec{s}$, weighted by the corresponding parameters:

$$\Delta G(\vec{s}|s\vec{e}q, \{\theta\}) = \sum_{i=1}^{p} \theta_i \cdot \Delta G_i(\vec{s}|s\vec{e}q) \tag{1.6}$$

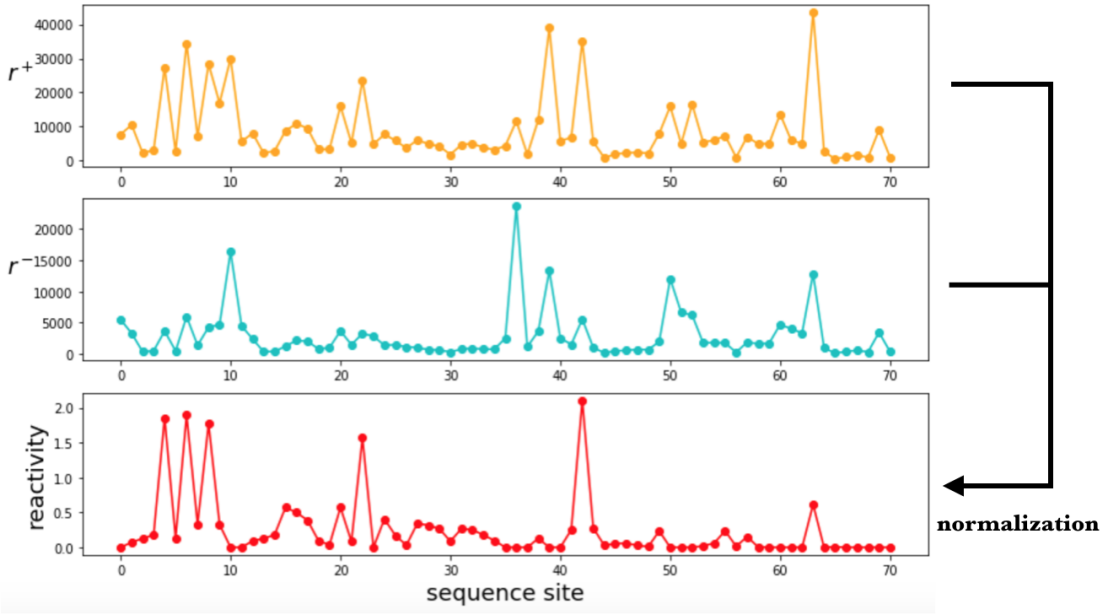
The parameters $\{\theta\}$ for the free energy change associated to a certain structural element are fit to observations from optical melting experiments [26, 27]. In these experiments, for oligonucleotides of known structure (*e.g.* duplexes of strands complementary for canonical base pairing, short stem-loops *etc.*) the absorbance towards a specific wavelength is measured in a set of temperatures ranging from the low values at which RNA is completely folded to the large values at which it is completely unfolded. The relations between absorbance and relative concentrations of folded and unfolded molecules allow for free energy change estimations. The limitations of experimental nearest neighbor parameters due to systematic errors in experiments, non-nearest neighbor effects of certain sequences and lack of knowledge for unobserved stable structures lead to a variety of computational approaches for parameter refinement, based on structure prediction on datasets of benchmark structures (see for example [28] and [29]).

1.2.2 Chemical probing

Chemical probing is an experimental method for the detection of structural elements in RNA [9]. It relies on a chemical reagent that can either form an adduct or induce a strand scission in the target RNA. In both cases, sites of modifications can be detected through reverse transcription. A sketch of the experiment for the adduct modification type is shown in Fig.1.6a. Reverse transcriptase (RT) is an enzyme that generates complementary DNA (cDNA) from an RNA template. Reverse transcription breaks when RT drops off at sites of adduction (or of termination of the scissed stretch). The reaction between a pool of RNAs of same sequence and the probing reagents thus generates a pool of cDNA fragments of different lengths, that correspond to different locations of adduct formation. Since reverse transcription can break also for natural drop-off of RT [30], the pool of RNA molecules is usually split in two samples: one is treated with the reagent



(a)



(b)

Figure 1.6: An illustration of chemical mapping experiments. (a) Two samples of RNA molecules are employed: one is treated with a reagent in the (+) or modification channel, whereas the other is treated with a control solvent in the (-) or control channel. By reverse transcription cDNA fragments are produced, with different lengths, due to the modification-induced and natural drop-offs of reverse transcriptase. Sequencing maps the distribution of fragment lengths to frequencies of adduct formation for each nucleotide of the RNA. Site reads in the (+) and (-) channels are combined into a reactivity profile for the whole molecule.

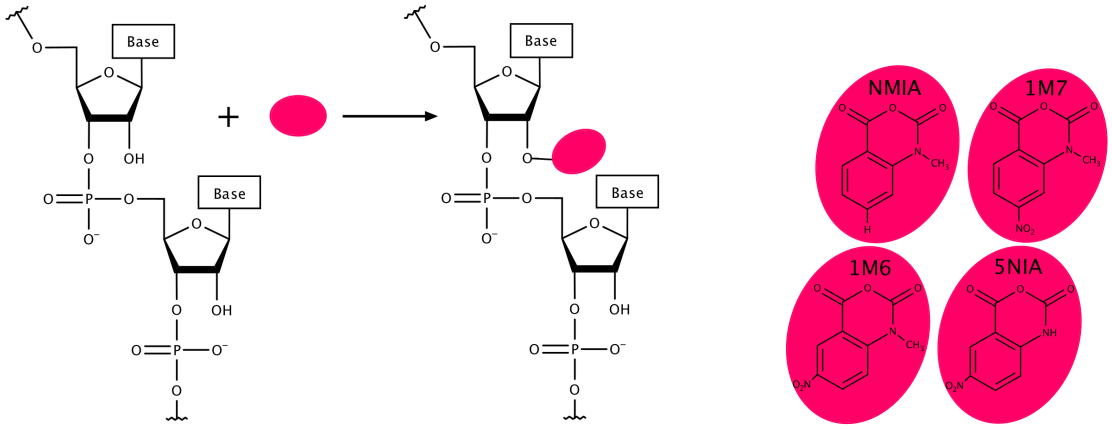


Figure 1.7: An illustration of the SHAPE chemistry: (a) the SHAPE reagent can bind to the C2′ hydroxyl of any of the four RNA nucleotides, by acylation. Since the binding site belongs to the nucleotide ribose, the probability of adduction, (*i.e.* the reactivity of the nucleotide) depends on the local structure conformation of the molecule. (b) four typically adopted SHAPE reagents for the *in vitro* SHAPE experiment.

probe ((+) or modification channel) while the other is treated only with a control solvent ((-) or control channel). By sequencing of the cDNA fragments in both channels separately, they are back mapped to the RNA sequence, allowing reconstruction of the distribution of modification sites (r^+ modification reads) and of natural drop-off sites (r^- control reads) from the distribution of the corresponding cDNA fragment lengths. The two distributions are combined through a normalization process into a reactivity profile (see Fig.1.6b for an example). The structural information content of the obtained reactivity profile depends on the nature of the used chemical reagent [9]. For example, a first class of probes in use consists of reactive alkylating agents, that are capable of forming stable covalent adducts with only a specific subset of the four RNA bases. This group includes dimethylsufate (DMS), that alkylates the N7 of Guanine, the N1 of Adenine and the N3 of Cytosine [31]; 1-cyclohexyl-3-(2-morpholinoethyl) carbodiimide metho-p-toluene sulfonate (CMCT), which reacts primarily with N3 of Uracil and N1 of Guanine [32]; diethyl pyrocarbonate (DEPC), which reacts primarily with N7 of Adenine [33], and kethoxal, which reacts with N1 and N2 of Guanine [34].

Among the probes used for strand scission, RNases are capable of cutting RNA molecules with specificity toward single- or double-stranded regions, with the major limitation that these enzymes act only at their specific binding sites [35, 36]. On the other hand, hydroxyl radicals (OH groups with an unpaired electron) cleave the RNA backbone by abstracting a proton, from either the C4′ or C5′ of the ribose, or both, revealing solvent accessibility of different regions of an RNA [37].

More recently, a class of probes was introduced, that gave rise to the Selective 2′ Hydroxyl Acylation analysed by Primer Extension (SHAPE) technique [10, 38]. This class includes for example 1-methyl-7-nitroisatoic anhydride (1M7), 1-methyl-6-nitroisatoic anhydride (1M6), N-methyl-nitroisatoic anhydride (NMIA) and 5-nitroisatoic anhydride (5NIA). As sketched in Fig.1.7, these reagents are capable of binding to the C2′ site of the ribose of any of the four RNA bases, through a nucleophilic attack. This allows the probing of all the four RNA bases simultaneously, with the reactivity of 2′-hydroxyl groups depending on the local structure of

RNA.

Nucleotides at flexible positions have been shown to sample conformations that enhance their nucleophilicity and thus reactivity towards SHAPE probes [10, 39]. Nonetheless, the interpretation of a SHAPE reactivity profiles is in fact not straightforward: probabilistic models have been used [40] to find correlations between SHAPE reactivity and structural elements, yet they showed that reactivity distributions for base-paired and unpaired nucleotides overlap significantly, making a direct distinction non-trivial; direct prediction of native structures from SHAPE profiles also have been shown to suffer major limitations [41].

The most adopted method to exploit chemical probing data seems to be the inclusion of data-dependent free-energy terms in thermodynamic models [42]. This method basically consists in adding a term to the model folding free energy of Eq.1.6, that couples with the pairing state of nucleotides and is a function of their reactivity. The result is a modified free energy model

$$\Delta G\left(\vec{s}|s\vec{e}q, R\right) = \Delta G_0\left(\vec{s}|s\vec{e}q\right) + \sum_{i=1}^{l_{seq}} \lambda_i\left(R_i\right) s_i \tag{1.7}$$

where ΔG_0 is the unbiased free energy of Eq.1.6 (the dependence on nearest-neighbor parameters is kept implicit to simplify the notation), and $\lambda\left(R_i\right) \cdot s_i$ is the term that couples reactivity R_i to the pairing state s_i of each nucleotide in the sequence of length l_{seq}. The function λ is usually referred to as *pseudo-energy*. Different implementations of this mehtod exist, and their efficiencies have been compared for example in [42]: on a representative set of benchmark structures, the best performing method in terms of MFE prediction accuracy has been shown to be the one by Deigan *et al.* [11]. In this implementation the pseudo-energy term

$$\lambda_i\left(R_i\right) = m \ln\left(1 + R_i\right) + b \tag{1.8}$$

couples a pairing state s_i that determines whether nucleotide i belongs or not to a stacked base pair ($s_i = 1$ or $s_i = 0$ respectively, in Eq.1.7). This model relies on two parameters: b is a negative intercept that enhances the free energy associated with stacked base pairs, independently of the reactivity; m is a positive slope that penalizes the base pairing and stacking of subsequent base pairs involving nucleotides with high reactivity.

Despite the average improvements in structure predictions brought about by pseudo-energy methods, still in some cases these methods result in less accurate predictions and high variability of results [43]. While, on one hand, new methods to analyse and interpret chemical probing data are proposed, on the other hand new structure probing reagents are under constant development [9, 44], especially to make possible the transition from the classical *in vitro* chemical mapping of solution structures to *in vivo* experiments in which membrane-permeant reagents must be employed to probe RNA structures in more realistic biological contexts.

In Chapter 2 we present an original method that improves secondary structure predictions, based on optimal integration of thermodynamic models with chemical probing data, in combination with co-evolutionary data (see next section). In Chapter 3 we present the original results of an investigation of cooperative effects of the SHAPE probe 1M7 in the ligand binding reactions involved in SHAPE experiments.

1.2.3 Direct coupling analysis

Different RNA sequences can be evolutionary related, in the sense that they descend from a common ancestor or share a linkage in a phylogenetic tree. The evolutionary relationship of a set of RNA sequences allows to classify them into an RNA *family*. For example, the 5S Ribosomal RNA is a component of the large subunit of the ribosome that is known to be functional in

all domains of life (bacteria, archaea, and eukaryotes) except for the mitochondrial ribosomes of fungi and animals [45]. In this case, all the different sequences of 5S Ribosomal RNA from different species are grouped into the same family. RNA sequences belonging to the same family are characterized by a certain degree of biological homology, related to a certain level of sequence similarity. Since sequences in the same family can not only feature different nucleotides at certain sites, but also can be of different lengths (*i.e.* be composed by different numbers of nucleotides), in order to compare a couple of sequences it is necessary to first align them. The alignment of two sequences consists in the identification of identical or similar stretches and in the insertion of gaps between them until the two reach the same length. A variety of sequence alignment methods exist, that are aimed at obtaining the maximum similarity for the whole set of homologous sequences. In particular, multiple sequence alignment (MSA) methods generalize and outperform pairwise alignment methods. In the work presented in Chapter 2, ClustalW [46] method is used, as it is not biased with prior knowledge of structure. This is an important requirement for performing benchmarks where the information obtained from the evolutionary analysis is used in a blind structure prediction. ClustalW relies on a clustering method in the first step, that builds up a distance matrix whose entries are a measure of the degree of similarity between each pair of sequences; from these scores, a guide tree is calculated and used to progressively align the sequences in order of similarity. The output MSA is a matrix $\{\sigma^b\}_{b=1}^{B}$ containing B homologous sequences. Each row of the matrix corresponds to one of the aligned sequences $\sigma^b = \{\sigma_1^b, \ldots, \sigma_N^b\}$ with common length N (obtained through the insertion of gaps for sequences with length $l_{seq} < N$). Each component of σ^b takes a value from the set $\{A, U, C, G, -\}$, a dictionary that codes for nucleotide identity (Adenine, Uracil, Cytosine, Guanine) and gaps ($-$).

The function of an RNA is typically bound to a limited set of interconverting conformations. Since this dynamic equilibrium of stable secondary structures is dominated by the native structure, it is reasonable to expect that this is preserved along the family of homologous sequences, except for minor changes due to possible minor differences in the mechanism of action between different species. Within functional regions that must be preserved, the only evolution mode is via pairs of compensatory mutations: if the nucleotide in position i is involved in a necessary base pair with a complementary nucleotide at position j, whenever a mutation occurs for i that would potentially disrupt the structure, it is compensated by a mutation of j that restores the complementarity of the two allowing the base pair to be preserved. This structure preserving evolutionary mechanism is called *co-evolution*. Co-evolution patterns can be extracted from the MSA $\{\sigma^b\}_{b=0}^{B}$, and used to infer the conserved base pairs. First, the frequency of nucleotide identity σ at position i can be computed as

$$F_i(\sigma) = \frac{1}{B}\sum_{b=0}^{B}\delta\left(\sigma_i^b, \sigma\right) \tag{1.9}$$

where δ is the Kronecker function

$$\delta(x, y) = \begin{cases} 1 & \text{if } x = y, \\ 0 & \text{otherwise} \end{cases} \tag{1.10}$$

and the frequency of co-occurrence of nucleotide identities σ and τ at positions i and j can be computed as

$$F_{ij}(\sigma, \tau) = \frac{1}{B}\sum_{b=0}^{B}\delta\left(\sigma_i^b, \sigma\right)\delta\left(\sigma_j^b, \tau\right) \tag{1.11}$$

A variety of systematic methods for structure prediction from co-evolutionary data exist that map the quantities in Eq.1.9 and 1.11 into corresponding base pair probabilities. These include

mutual information analysis methods [47, 48], G-test-based statistical procedures [49], and direct coupling analysis [50, 51].

Direct coupling analysis consists in building a fully connected five-states Potts model that is trained to reproduce the observed single- and two-sites frequencies of Eq.1.9 and 1.11. The model is defined by the Hamiltonian

$$\mathcal{H}(\sigma) = -\sum_i h_i(\sigma_i) - \sum_{ij} J_{ij}(\sigma_i, \sigma_j) \tag{1.12}$$

where σ is a sequence, each component σ_i of which can be in one of the five possible states $\{A, U, C, G, -\}$, h_i is the *local field* acting on σ_i, and J_{ij} is the two-point interaction between σ_i and σ_j named *direct coupling*. Given the values of the model parameters $\{h_i(\sigma_i), J_{ij}(\sigma_i, \tau_i)\}$, a probability can be assigned to any sequence

$$P(\sigma) = \frac{1}{Z} e^{-\mathcal{H}(\sigma)} \tag{1.13}$$

where Z is the partition function

$$Z = \sum_{\{\sigma\}} e^{-\mathcal{H}(\sigma)} \tag{1.14}$$

obtained by summing over all the possible sequences $\{\sigma\}$ of the corresponding Boltzmann weights $e^{-\mathcal{H}(\sigma)}$. The model parameters h_i and J_{ij} are trained to produce frequencies of nucleotide identity τ at position i

$$f_i(\tau) = \sum_{\{\sigma\}} P(\sigma)\, \delta(\sigma_i, \tau) \tag{1.15}$$

and frequencies of co-occurrence of nucleotide identities σ and τ at positions i and j

$$f_{ij}(\sigma_i, \tau_j) = \sum_{\{\sigma\}} P(\sigma)\, \delta(\sigma_i, \tau)\, \delta(\tau_j, \tau) \tag{1.16}$$

as close as possible to those observed (Eq.1.9 and 1.11, respectively).

Once the model is trained, a score can be assigned to any pair of sequence positions (i, j) by taking the Frobenius norm of the direct couplings:

$$S_{ij} = \sqrt{\sum_{\{\sigma,\tau\}} J_{ij}(\sigma, \tau)^2} \tag{1.17}$$

This score is representative of the level of co-evolution of a pair of nucleotides, in the limits of accuracy of the model. The training of this model can be performed in different ways, that have been compared in [50], where the best performing one has been shown to be Boltzmann learning. For this reason, the original results presented in Chapter 2 are obtained by using the scores of Eq.1.17 as obtained with Boltzmann learning [50], in combination with chemical probing data and a thermodynamic model for secondary structure predictions.

Chapter 2

Training a neural network with experimental data for RNA structure prediction

In this Chapter, we present a machine learning procedure that allows training and selection of a set of models that predict secondary structure ensembles of RNA sequences. All the models in the tested set share a neural network architecture, that combines thermodynamic nearest-neighbor parameters, chemical probing data, and co-evolutionary data from direct coupling analysis, eventually mapping them into perturbations to the ensembles free energy. The model parameters, that correspond to *weights* and *biases* of the network nodes, are trained on a benchmark dataset of RNA sequences whose native structure has been proposed in high-resolution X-RAY diffraction experiments. The workflow of training, model selection and validation is represented in the scheme of Fig.2.1. Training is aimed at the maximization of ensemble population of native structures. Regularization is used in the training phase to avoid overfitting the models and thus to improve transferability to new sequences and structure probing data. A model selection procedure is used to discard poorly transferable models and choose the one that yields the best balance between performance on a training set and transferability to a test set. The results are eventually validated on an independent set containing data never seen in training and testing.

The content of this chapter is mainly adapted from our published work [52].

2.1 Dataset

In this work we mainly rely on data from curated open-access online databases. Crystallographic structures are taken from the Protein Data Bank (PDB) [53], with a selection of refinement resolution to be better than 4 Å. This makes up to approximately 3000 structures available. Chemical probing data are mainly extracted from the RNA Mapping Data Base (RMDB) [54, 55], containing reactivity profiles for around 15000 sequences up to this date, and from Refs. [56, 57]. Unfortunately, exact matching of the sequence used in high-resolution crystallographic experiments with the ones used in chemical probing assays dramatically drops the number of RNA molecules with available corresponding data. For these molecules, sets of homologous sequences are drawn from the Rfam database [58]. Chemical probing data for a part of sequences with annotated structure and family have been collected by our experimental collaborators. In Table 2.1 the list of molecules is reported along with the PBD entry from which the structure

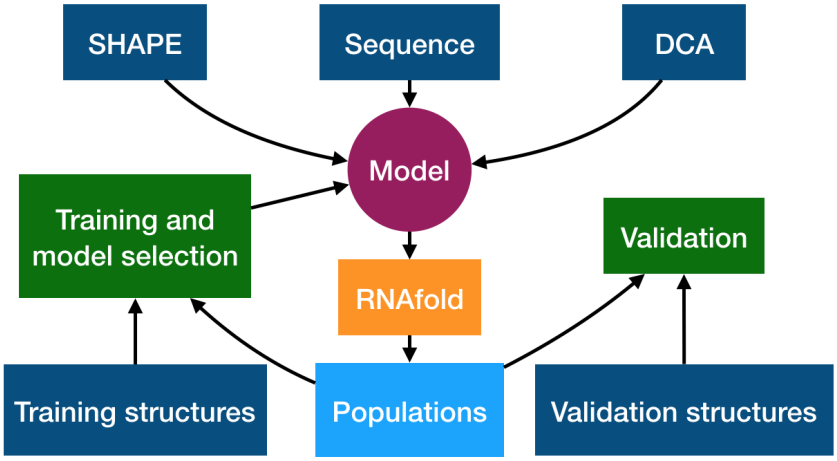


Figure 2.1: Graphical scheme of the machine learning procedure. Models that integrate RNAfold, chemical probing experiments, and DCA scores into prediction of ensemble population of structures are trained. One among all the proposed models is selected based on a transferability criterion and validated against data that is not seen during training. Available reference structures are used as target for training and validation.

Molecule	PDB	l_{seq}	Chemical probing	Rfam
yeast Phe-tRNA	1EHZ	76	`TRNAPH_1M7_0000`	RF00005
D5,6 Yeast ai5g G-II Intron	1KXK	70	Sattler's Lab	RF00029
Ribonuclease P RNA	1NBS	150	`10.1261/rna.047068.114`	RF00011
Adenine riboswitch	1Y26	71	`ADDRSW_1M7_0001`	RF00167
TPP riboswtich	2GDI	78	`TPPRSW_1M7_0001`	RF00059
SAM-I riboswitch	2GIS	94	Sattler's Lab	RF00162
Lysine riboswitch	3DIG	174	`10.1073/pnas.1219988110`	RF00168
O. I. G-II Intron	3IGI	388	`10.1073/pnas.1219988110`	RF02001
c-di-GMP riboswitch	3IRW	90	Sattler's Lab	RF01051
M-box riboswitch	3PDR	161	`10.1073/pnas.1219988110`	RF00380
THF riboswitch	3SD3	89	Sattler's Lab	RF01831
Fluoride riboswitch	3VRS	52	Sattler's Lab	RF01734
SAM-I/IV riboswitch	4L81	96	`RNAPZ8_1M7_0001`	RF01725
Lariat capping ribozyme	4P8Z	188	`GIR1RZ_DMS_0000`	RF01807
ydaO riboswitch	4QLM	108	`YDAORS_DMS_0000`	RF00379
ZMP riboswitch	4XW7	64	Sattler's Lab	RF01750
50S ribosomal	4YBB	120	`5SRRNA_1M7_0001`	RF00001
5-HTP RNA aptamer	5KPY	71	`RNAPZ9_1M7_0001`	RF01982

Table 2.1: RNA molecules included in the dataset. For each molecule we indicate the PDB entry of the corresponding annotated structure, the number of nucleotides or sequence length (l_{seq}), the source of chemical probing data (either Sattler's Lab for our collaboration, `rdat` file name from the RMDB or DOI for data collected from the literature), and Rfam family. For PDB 4YBB, chain CB was used as a reference.

has been extracted, the source of chemical probing data (Sattler's Lab for data coming from our collaboration, `rdat` file name for data from the RMDB, DOI for data collected from the literature), and the Rfam family to which each sequence belongs.

Secondary structures are obtained by annotating crystallographic structures with `x3dna-dssr` [16], an integrated software tool that employs the standard base reference frame and a set of simple geometric criteria to identify all the base pairs in the input PDB structure. Chemical probing reactivities are either obtained by normalizing modification and control reads (see Section 1.2.2) when available, or directly extracted from data sources (RMDB or available from the literature) and normalized to standard scores (mean subtraction and division by standard deviation). In the former case, normalization is carried out in the following steps: reads in the control channel $r^{(-)}$ are divided by their sum

$$R_i^{(-)} = \frac{r_i^{(-)}}{\sum_{k=1}^{l_{seq}} r_k^{(-)}} \tag{2.1}$$

and so are the reads in the modified channel $r^{(+)}$

$$R_i^{(+)} = \frac{r_i^{(+)}}{\sum_{k=1}^{l_{seq}} r_k^{(+)}} \tag{2.2}$$

In this way we obtain the probability of modification $R_i^{(+)}$ and of natural drop-off of reverse transcriptase $R_i^{(-)}$ at site i. Normalized reactivities R_i are then obtained by subtraction of the two probabilities in order to exclude the contribution of natural drop-off of reverse transcriptase to the modification counts. Negative values are replaced with 0, as the occurrence of more natural drop-offs of reverse transcriptase in the modification channel with respect to the control channel is attributed to experimental noise and are thus not informative:

$$R_i = \max\{0, \quad R_i^{(+)} - R_i^{(-)}\} \tag{2.3}$$

Direct-coupling analysis of each sequence is carried out using the Boltzmann-learning strategy from [58] on the multiple sequence alignment of the corresponding family. MSA were performed with ClustalW using the sequence from PDB as guide for the alignment. A score matrix from direct couplings is obtained with Eq.1.17. In the following, for each couple i, j the corresponding score is referred to as DCA score or J_{ij}.

2.2 Neighbor reactivities

The reactivity of a nucleotide towards a chemical mapping reagent has no direct mapping to the folding free energy of the sequence it belongs to. For this reason the functional form of reactivity-dependent pseudo-energies to be implemented in nearest-neighbor models as in Eq.1.7 can be chosen with a large degree of arbitrariness. The choice presented here is based on the hypothesis that the pairing state of a nucleotide correlates not only with its own reactivity, but also with the reactivities of neighboring nucleotides along the sequence, both downstream (toward the $5'$ end) and upstream (toward the $3'$ end). We then investigate the possible role of the local collective behavior of a limited part of the sequence around the nucleotide in chemical probing reactions. To test this hypothesis, we build up a set of linear models that predict the

pairing state s_i of single nucleotides without using folding algorithms. Here we define the pairing state of a nucleotide in the following way:

$$s_i = \begin{cases} 1 & \text{if } i \text{ is paired,} \\ 0 & \text{if } i \text{ is unpaired} \end{cases} \tag{2.4}$$

In this way, the information about pairing partners (which nucleotide is paired to i if i is paired) is not taken into account. Different models are built, that combine linearly the reactivity of each nucleotide i with the reactivities of a different number of its neighbors ($R_{i-p}, \ldots, R_{i+p}, \quad p \in \{0, \ldots, 7\}$) in a range from 0 downstream and 0 upstream, to 7 downstream and 7 upstream, together with the identity $\delta(\sigma_i, \sigma)$ of the nucleotide ($\sigma \in \{A, U, C, G\}$). The predicted pairing state $\hat{s_i}$ for the model including p neighbors is then

$$\hat{s_i}(\{R\}, p) = \sum_{k=-p}^{p} w_k R_{i+k} + w_\sigma \delta(\sigma_i, \sigma) + c \tag{2.5}$$

where the model parameters $\{w_k\}$ weight the contributions of the reactivities from nucleotide i and its neighbors, w_σ couples with nucleotide identity and c tunes the intercept of the model. The dataset, here composed of the reactivities, nucleotide identities and nucleotide pairing states drawn from the original dataset of Table2.1, is first split randomly into a training set (size 0.7 of the data) T and a validation set (size 0.3) V. For each each model, identified by the hyper-parameter p, the parameters are optimized on the training set by minimizing the root-mean-square error on the predicted pairing state (in-sample error)

$$RMS_p^{in} = \sqrt{\sum_{i \in T} \left(\hat{s_i}(\{R\}, p) - s_i\right)^2} \tag{2.6}$$

In order to exclude overfitted models and select the most transferable one, the performance of trained models is estimated on the validation set through the out-sample root-mean-square error

$$RMS_p^{out} = \sqrt{\sum_{i \in V} \left(\hat{s_i}(\{R\}, p) - s_i\right)^2} \tag{2.7}$$

Training and validation are iterated over 200 random splittings of the dataset, the average results are reported in Fig.2.2. Local reactivity patterns prove informative on the pairing states of nucleotides, as the in-sample error on prediction decreases monotonically with increasing p, as well as the out-sample error for small values of p. At larger values of p the out-sample error decreases slowly until it starts increasing again at $p > 6$. This is a signature of overfitting: since the number of model parameters (namely $\{w_k\}$) increases with p, at a certain value of p they are enough to be overfitted to the training dataset. In this case the model reproduces accurately the data in the training set, but it is not anymore transferable to new unseen data. Its performance in blind predictions is then expected to be low.

Based on these observation, the choice of a pseudo-energy function that couples the pairing state of a nucleotide with both its reactivity and the reactivities of its neighbors along the sequence is made, in order to take into account the possible local collective behavior of neighboring nucleotides in chemical probing reactions. On the other hand, also a robust procedure to monitor overfitting is built up and used to select proposed models with a transferability criterium.

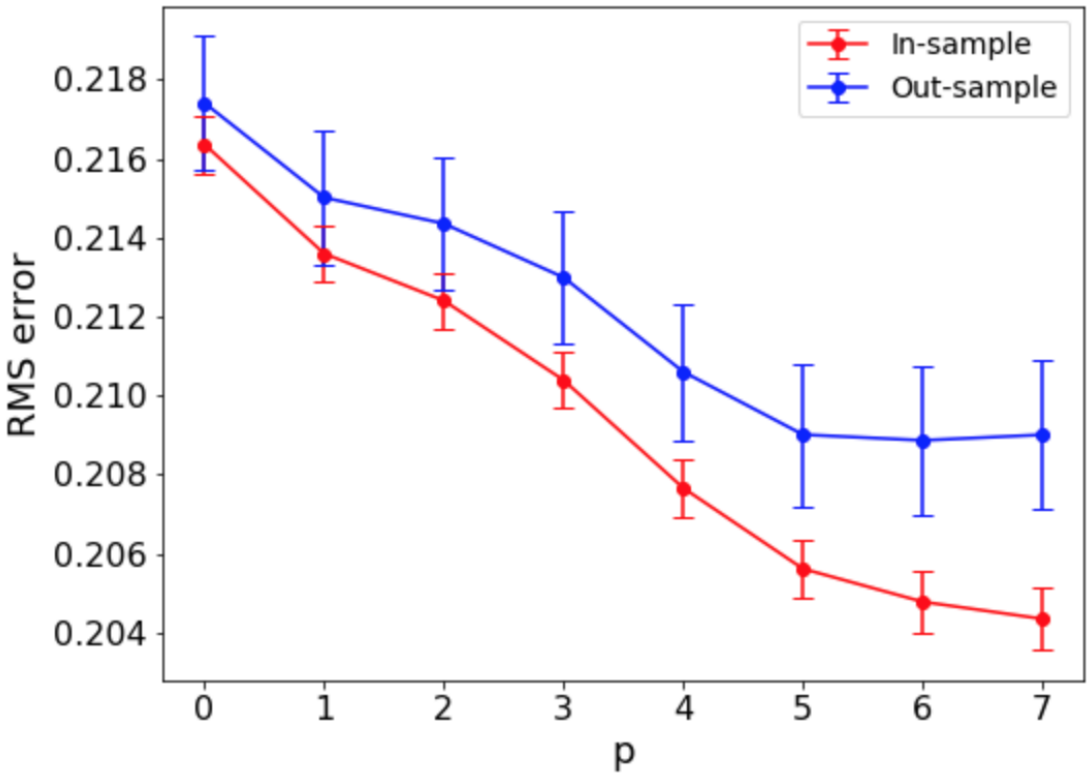


Figure 2.2: Root-mean-square (RMS) error of linear models for prediction of nucleotide pairing state. Each tested model maps the reactivity of the nucleotide and those of a different number p of its neighbors along the sequence, both down- and up-stream. The dataset is split 200 times in a training and a validation set, so that average in-sample error (in red) and average out-sample error (in blue) is reported for each value of p. Error bars indicate statistical errors on RMS error estimates. While in-sample error keeps decreasing with increasing p, the out-sample error starts increasing large values of p, due to overfitting of models with too many parameters.

2.3 Model architecture

The models proposed here combine chemical probing reactivities and DCA scores through a function that maps these data into pseudo-energies terms. These terms act as perturbations to the ensemble free energy of a nearest-neighbor folding model. In particular, we choose the implementation of folding parameters of the `ViennaRNA` package [59], with the parametrization from Andronescu *et. al.* [60]. Reactivities R_i and DCA scores J_{ij} are mapped into single-point pseudo-energies λ_i and pairwise pseudo-energies. λ_{ij}, respectively. Single-point terms affect the pairing propensity of individual nucleotides, whereas pairwise terms affect the pairing propensity of specific couples of nucleotides. The modified free energy obtained in this way is a modification of the original one by two additional terms:

$$\Delta G\left(\vec{s}|\vec{seq};\vec{R},\vec{J}\right) = \Delta G_0\left(\vec{s}|\vec{seq}\right) + RT\sum_{i=1}^{l_{seq}}\lambda_i(\vec{R})\cdot(1-s_i) + RT\sum_{j>i+2}^{l_{seq}}\lambda_{ij}(\vec{J})\cdot s_{ij} \tag{2.8}$$

where R is the gas constant, T the temperature (set to 310 K), s_{ij} is the matrix of base-pairs

$$s_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are paired,} \\ 0 & \text{otherwise} \end{cases} \tag{2.9}$$

and $s_i = \sum_{j\neq i} s_{ij}$ is the individual pairing state. Single-point and pairwise perturbations are implemented in the thermodynamic model through the soft constraints functions from the `ViennaRNA` package `vrna_sc_add_up` and `vrna_sc_add_bp`, respectively. Importantly, a positive single-point pseudo-energy $\lambda_i > 0$ disfavors structures of the ensemble in which nucleotide i is unpaired ($s_i = 0$), and vice-versa. Pairwise pseudo-energies either favor or disfavour structures in which the base pair $i-j$ is present, depending if $\lambda_{ij} < 0$ or $\lambda_{ij} > 0$, respectively.

All the models share a neural network architecture, summarized in Fig.2.3, in which reactivities and DCA scores are processed by two separate channels. The reactivity channel consists in a single-layered convolutional network that includes reactivities from the p neighbors downstream and upstream along the sequence, and process them via a linear activation function

$$\lambda_i\left(\vec{R}\right) = \sum_{k=-p}^{p} a_k \cdot R_{i+k} + b \tag{2.10}$$

where the parameters $\{a_k\}$ control the relative weights of neighbors, and b is the bias. The convolutional window slides over the whole reactivity profile mapping each subset of $2p+1$ contiguous reactivities into a corresponding single-point pseudo-energy for the nucleotide at the center of the window. The optimization of p is carried out by scanning a discrete scale and at a different level with respect to the optimization of parameters $\{a_k\}$ and b. For this reason p is defined as a hyper parameter. The DCA channel instead consists in a double-layered network

$$\lambda_{ij}\left(J_{ij}\right) = C\cdot\sigma\left(A\cdot J_{ij}+B\right)+D \tag{2.11}$$

The activation function of the output layer is linear with parameters C and D, whereas a sigmoid activation

$$\sigma\left(x\right) = \frac{1}{1+e^{-x}} \tag{2.12}$$

is applied at the innermost layer, with weight A and bias B. Each model has thus $2p+6$ free parameters.

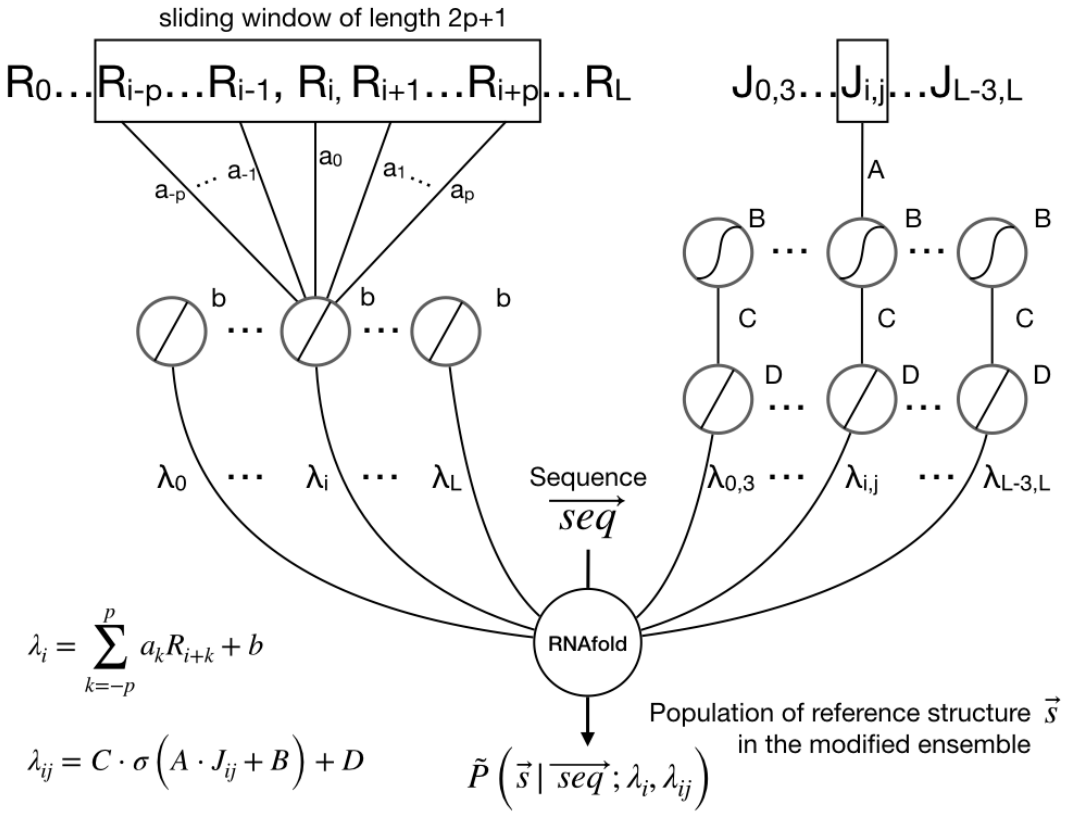


Figure 2.3: Sequence, reactivity profile, and DCA scores are included through pseudo-energy terms in the `ViennaRNA` model free energy. The network is split into two channels: a single-layered channel for reactivity input (left side) and a double-layered channel for DCA couplings (right side). Along the reactivity channel, a convolutional layer operates a linear combination on the sliding window including the reactivity R_i of a nucleotide and the reactivities $\{R_{i+k}\}$ of its neighbors, with weights $\{a_k\}$ and bias b. The output consists in a single-point pseudo-energy λ_i coupling the pairing state s_i of the i-th nucleotide. In the DCA channel, the first layer transforms the input DCA score J_{ij} via a non-linear (sigmoid) activation function, with weight A and bias B. The transformed DCA input is then mapped to a pairwise pseudo-energy λ_{ij} coupling the pairing state of the specific ij pair via a second layer, implementing a linear activation function with weight C and bias D. Pseudo-energies both coupling individual pairing states and specific base-pairs are applied as perturbations to the folding free energy model implemented in `ViennaRNA`.

2.4 Training, Model selection and Validation

By favoring and disfavoring structures depending on the input reactivities and DCA scores, the pseudo-energy perturbations affect the whole ensemble of structures for each sequence. This results in modified ensemble populations

$$P\left(\vec{s}|s\vec{e}q;\theta\right) = e^{-\frac{1}{RT}\Delta G(\vec{s}|s\vec{e}q;\theta)}/Z\left(s\vec{e}q;\theta\right) \tag{2.13}$$

where $\Delta G\left(\vec{s}|s\vec{e}q;\theta\right)$ is that of Eq.2.8, but the dependence on model parameters $\theta = \{a_k, b, A, B, C, D\}$ is highlighted, whereas the dependence on input data is kept implicit, for sake of simplicity of notation. The modified partition function is the sum over all the possible structures of their corresponding modified Boltzmann weights

$$Z\left(s\vec{e}q,\theta\right) = \sum_{\{\vec{s}\}} e^{-\frac{1}{RT}\Delta G(\vec{s}|s\vec{e}q;\theta)} \tag{2.14}$$

Training

First, the dataset is split randomly into a training set T and a validation set V, of size 0.7 and 0.3 of the original dataset, respectively. The training of each model is aimed at finding the optimal values $\hat{\theta}$ of model parameters that result in maximum population of the native secondary structure of each sequence in T, under the assumption that the native structure of a sequence is the high-resolution crystal structure in our dataset. The training set T thus contains pairs of sequence $s\vec{e}q$ and base-pair matrix $\hat{s}$ of the corresponding native secondary structure. For each pair the following cost function is defined:

$$\mathcal{C}\left(\theta\right) = -RT \ln P\left(\hat{s}|s\vec{e}q;\theta\right) \tag{2.15}$$

The cost function is decomposable into two terms, $\Delta G\left(\vec{s}|s\vec{e}q;\theta\right)$ and $-RT \ln Z\left(s\vec{e}q,\theta\right)$ that can be straightforwardly computed through the `ViennaRNA` functions `vrna_eval_structure` and `vrna_pf`, respectively. Training is carried out by minimizing the average of the cost function over all the pairs of sequence and structure in the training set. Since for each sequence the cost function is defined as the negative logarithm of the ensemble population of its native secondary structure, minimizing the arithmetic average of cost functions is equivalent to maximizing the geometric average of the ensemble populations of the native structures:

$$\hat{\theta} = \arg\max_{\theta} \prod_{s\vec{e}q \in T} P\left(\hat{s}|s\vec{e}q;\theta\right) \tag{2.16}$$

The nearest-neighbor parameters of the thermodynamic model are already fitted on a number of benchmark RNA sequences and structures (see Section 1.2.1). The integration of structure probing data in this model relies on the optimization of the set of additional parameters θ on top of the already optimized nearest-neighbor parameters. This may lead to a high risk of overfitting of these additional refining parameters. To reduce this risk, regularization is included in the training procedure, by means of two $l-2$ regularization terms in the cost function. Since two-dimensional DCA scores and one-dimensional reactivity profiles differ in the amount of structural information they contain, instead of a standard single regularization term on all parameters θ, two representational regularization terms are added, each with an independent coefficient, directly on the pseudo-energies mapped from each type of data. The regularized cost function thus has the following form:

$$\mathcal{C}\left(\theta\right) = -RT \ln P\left(\hat{s}|s\vec{e}q;\theta\right) + RT\left[\alpha_S \sum_i \lambda_i^2 + \alpha_D \sum_{ij} \lambda_{ij}^2\right] \tag{2.17}$$

where α_s and α_D are the regularization coefficients acting respectively on chemical probing and DCA data. The two terms prevent the pseudo-energies added to the free energy model from becoming too large, and thus help preventing overfitting of model parameters θ during the minimization of the cost function. A crucial ingredient of the presented training procedure is that the gradient of the cost function with respect to each weight and bias of the neural network is easily computed, as it is proportional to pairing probabilities of individual nucleotides p_i and of nucleotide pairs p_{ij}:

$$\begin{aligned}
\frac{\partial \mathcal{C}}{\partial a_k} &= RT \sum_{i=1}^{l_{seq}} \left[(p_i - \hat{s}_i) + 2\alpha_S \lambda_i\right] \frac{\partial \lambda_i}{\partial a_k} = RT \sum_{i=1}^{l_{seq}} \left[(p_i - \hat{s}_i) + 2\alpha_S \lambda_i\right] R_{i+n} \\
\frac{\partial \mathcal{C}}{\partial b} &= RT \sum_{i=1}^{l_{seq}} \left[(p_i - \hat{s}_i) + 2\alpha_S \lambda_i\right] \frac{\partial \lambda_i}{\partial b} = RT \sum_{i=1}^{l_{seq}} \left[(p_i - \hat{s}_i) + 2\alpha_S \lambda_i\right] \\
\frac{\partial \mathcal{C}}{\partial A} &= RT \sum_{j>i+2}^{l_{seq}} \left[(\hat{s}_{ij} - p_{ij}) + 2\alpha_D \lambda_{ij}\right] \frac{\partial \lambda_{ij}}{\partial A} = RT \sum_{j>i+2}^{l_{seq}} \left[(\hat{s}_{ij} - p_{ij}) + 2\alpha_D \lambda_{ij}\right] J_{ij} \cdot C\sigma'\left(AJ_{ij} + B\right) \\
\frac{\partial \mathcal{C}}{\partial B} &= RT \sum_{j>i+2}^{l_{seq}} \left[(\hat{s}_{ij} - p_{ij}) + 2\alpha_D \lambda_{ij}\right] \frac{\partial \lambda_{ij}}{\partial B} = RT \sum_{j>i+2}^{l_{seq}} \left[(\hat{s}_{ij} - p_{ij}) + 2\alpha_D \lambda_{ij}\right] \cdot C\sigma'\left(AJ_{ij} + B\right) \\
\frac{\partial \mathcal{C}}{\partial C} &= RT \sum_{j>i+2}^{l_{seq}} \left[(\hat{s}_{ij} - p_{ij}) + 2\alpha_D \lambda_{ij}\right] \frac{\partial \lambda_{ij}}{\partial C} = RT \sum_{j>i+2}^{l_{seq}} \left[(\hat{s}_{ij} - p_{ij}) + 2\alpha_D \lambda_{ij}\right] \sigma\left(AJ_{ij} + B\right) \\
\frac{\partial \mathcal{C}}{\partial D} &= RT \sum_{j>i+2}^{l_{seq}} \left[(\hat{s}_{ij} - p_{ij}) + 2\alpha_D \lambda_{ij}\right] \frac{\partial \lambda_{ij}}{\partial D} = RT \sum_{j>i+2}^{l_{seq}} \left[(\hat{s}_{ij} - p_{ij}) + 2\alpha_D \lambda_{ij}\right]
\end{aligned} \tag{2.18}$$

Base-pairing probabilities are computed via function `vrna_bpp` from the `ViennaRNA` package. The inclusion of regularization terms in the cost function brings in two additional hyperparameters, α_S and α_D that, like p, are not optimized by minimization of the cost function. The triplet of hyperparameters $\{p, \alpha_S, \alpha_D\}$ identifies each model that is trained/ The training of model parameters θ is carried out at fixed values of its hyperparameters, that are scanned at a higher level over a discrete scale, namely $p \in [0, 1, 2, 3]$ and $\alpha_S, \alpha_D \in [\infty, 1.0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 0.0]$, for a total of $4 \times 7 \times 7 = 196$ models.

Model selection

For each model, the corresponding cost function is minimized using the sequential quadratic programming algorithm as implemented in the `scipy.optimize` optimization package [61]. The minimization problem is non-convex whenever α_D is finite, so the landscape of the cost function is expected to be rough, with multiple local minima. Thus, in principle the result of the minimization depends on the initialization of the model parameters. For this reason, for each minimization multiple initial values for the model parameters are extracted from a random uniform distribution, and those yielding the minimum cost function are eventually selected. In addition to these random initial values, also three specific sets of starting points are tested:

- parameter values from the optimized $\{p-1, \alpha_S, \alpha_D\}$ model, with the new a_{-p} and a_p set to 0.0; if $p = 0$, we ignore this starting point.
- parameter values from the optimized $\{p, 10 \cdot \alpha_S, \alpha_D\}$ model; if $\alpha_S = 0.0$, we use values from

Figure 2.4: Leave-one-out test for model transferability: at iteration i, the i-th molecule is excluded from the training set of size N, and the model parameters are optimized on the reduced training set containing the other $N-1$ molecules; the resulting optimal parameters ϑ_i are used to evaluate the cost function $\mathcal{C}$ on the left-out sample. The average of the cost function over all the estimates is used as a transferability score to rank the models and select the one with highest estimated transferability.

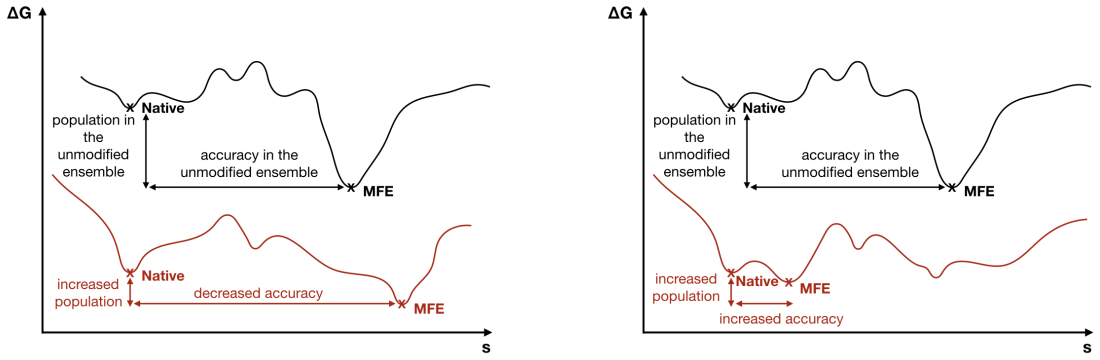
the optimized $\{p, 10^{-4}, \alpha_D\}$ model; if $\alpha_S = 1$, we use values from the optimized $\{p, \infty, \alpha_D\}$ model; if $\alpha_S = \infty$, we ignore this starting point.

- parameter values from the optimized $\{p, \alpha_S, 10 \cdot \alpha_D\}$ model; if $\alpha_D = 0.0$, we use values from the optimized $\{p, \alpha_S, 10^{-4}\}$ model; if $\alpha_D = 1$, we use values from the optimized $\{p, \alpha_S, \infty\}$ model; if $\alpha_D = \infty$, we ignore this starting point.

This ensures that models with higher complexity (*i.e.*, higher p or lower α_S or α_D) by construction fit the data better than models with lower complexity. In this way the performance of the models, as evaluated on the training set T, is by construction a monotonically decreasing function of α_D and α_S, and a monotonically increasing function of p. Among the models optimized in the training procedure, the one that yields the best performance without overfitting the training data is selected, in order to ensure the transferability of its architecture and optimal parameters. As a test for transferability, a leave-one-out test is used. This procedure consists in iteratively leaving each of the sequences in T out of the training set at a time, and using the optimal parameters resulting from optimization on the reduced (size $N-1$ where N is the size of T) training set to compute the ensemble population of the native structure for the left-out sequence. The population of native structures, averaged on the left-out systems, is used to rank all of the tested models. Next, the model with the highest score is considered as the most capable of yielding an increase in population of native structures for sequences on which it was not trained.

Validation

The resulting model is then validated on the validation set V, containing sequences that were not used in the parameter or hyperparameter optimization. For these sequences the ensemble population of the native structure is computed using the selected model with the optimal parameters obtained in the training. Increasing the population of native structure does not necessarily mean that the minimum free energy (MFE) structure of the modified ensemble becomes more similar to the native structure. Since the perturbations to the free energy model brought by



(a) (b)

Figure 2.5: Two possible scenarios for the effect of pseudo-energy perturbations on the ensemble free-energy: (a) the native structure is assigned a lower free energy and thus a higher population, but the new minimum free energy structure is further from the native one in terms of MCC, yielding lower accuracy of predictions; (b) the ensemble free energy is modified in such a way that the native structure one with minimum free energy lies in a local minimum closer to the global minimum, yielding a higher accuracy of predictions.

pseudo-energies affect the free energy of the whole ensemble, it may happen that whereas the free energy of the native structure is decreased (*i.e.* population is increased), its distance from the minimum free energy becomes larger. In this case the accuracy of structure predictions with the optimized model would suffer a decrease. For this reason, as an additional validation for the optimized models, the similarity between the predicted MFE structures and the native structures is checked. The Matthews Correlation Coefficient (MCC) [62] is used to quantify structure similarity. For each pair of native structure $\hat{s}$ and MFE structure s^{MFE} base-pair matrices, an exact match of base-pairs $\hat{s_{ij}} = s_i^{MFE} = 1$ is increments the number of true positives TP, an exact match of unpaired states $\hat{s_i} = s_i^{MFE} = 0$ increments the number of true negatives TN, while all mismatches in base-pair and unpaired state predictions increment the number of false positives FP and false negatives FN, respectively. In the case of the RNA sequences in our dataset, whose lengths range from some tens to a few hundreds of nucleotides, the number of possible base-pairs is large and so is the number of true negatives. In this limit, the Matthews Correlation Coefficient is approximately equal to the geometric average of precision $P = \frac{TP}{TP+FP}$ and sensitivity $S = \frac{TP}{TP+FN}$ [63]

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \xrightarrow{TN \to \infty} \sqrt{S \cdot P} \tag{2.19}$$

and thus to the Fowlkes-Mallows coefficient [64], that optimally balances the two metrics.

2.5 Results

To summarize, chemical probing experiments provide reactivities per nucleotide (one-dimensional information, R_i) that are mapped via a single-layered convolutional network to pseudo-energies affecting the pairing propensity of individual nucleotides (λ_i). Similarly, direct-coupling analysis provides predicted contact scores (two-dimensional information, J_{ij}) that are mapped through a non-linear function into pseudo-energies to be associated with specific nucleotide-nucleotide

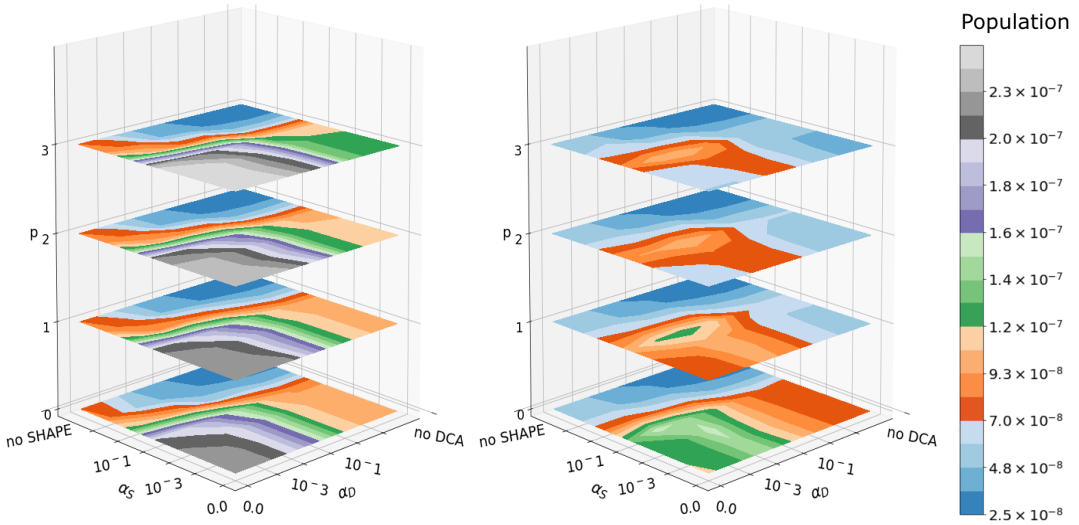
PDB	S1-S2-S3-S4
1EHZ	V-T-T-T
1KXK	T-T-T-T
1NBS	T-T-T-T
1Y26	V-T-T-V
2GDI	T-T-V-T
2GIS	T-T-V-T
3DIG	T-T-V-V
3IGI	T-V-T-V
3IRW	T-T-T-T
3PDR	V-V-T-T
3SD3	V-V-T-T
3VRS	V-T-T-T
4L81	T-V-V-V
4P8Z	T-T-T-T
4QLM	T-T-T-T
4XW7	V-V-V-V
4YBB	T-V-V-V
5KPY	T-T-T-T

Table 2.2: The four random splittings $S1 - S4$ of the original dataset into training T and validation V sets. For each molecule and for each splitting, wether the molecule is in T or V is indicated.

pairs (λ_{ij}). The resulting pseudo-energy terms are added to the ensemble free energy model implemented in the folding algorithm RNAfold of the `ViennaRNA` package [59], which allow the full partition function of the system to be computed, including the population of any suboptimal structure. The parameters of the mapping functions are trained in order to maximize the population of the secondary structures as annotated in a set of high-resolution X-ray diffraction experiments. The differentiability of the nearest-neighbor thermodynamic model with respect to the applied pseudo-energies is crucial, since it allows the thermodynamic model to be used for gradient backpropagation in the training procedure. Reference structures are obtained from the PDB structural database [53]. Reference chemical probing data are partly taken from the RMDB chemical mapping database [54, 55] and from Refs. [56, 57], and partly provided by our experimental collaborators of Sattler's Lab. Reference direct couplings are partly taken from Ref. [50] and partly obtained in this work, using RNA families deposited on RFAM [58]. The model complexity is controlled via three hyperparameters, which are chosen using a cross-validation procedure, and the obtained model is evaluated on an independent dataset not seen during the training procedure.

2.5.1 Training

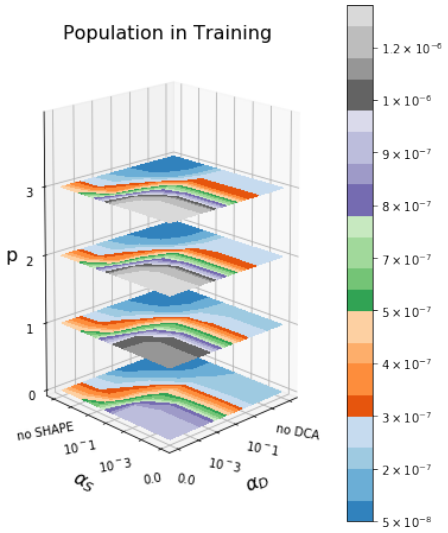
Starting from the dataset of Table 2.1 containing data for 18 different RNA molecules, a training set T of 12 molecules is randomly chosen, leaving the remaining 6 out in the validation set V, used for the final validation. Since crystal structures, chemical probing data, and co-evolutionary data for different molecules might be of different quality, the specific choice of the splitting might affect the overall training and validation results. We thus generate four independent random splittings, reported in Table 2.2. In the following we refer to splitting S3, as it leads to the worst performance among the others in the cross-validation test and to the best performance in the final validation. Results for all the splittings are reported as well. Importantly, the external



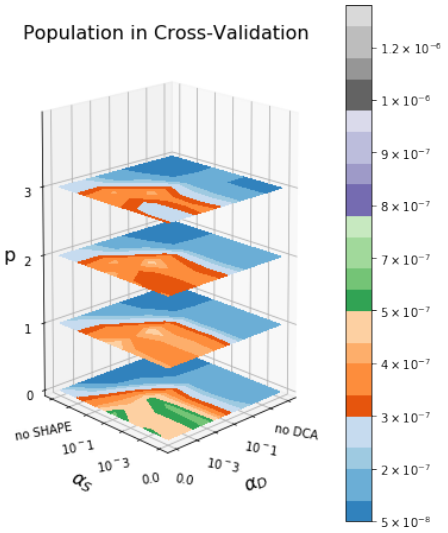
(a) (b)

Figure 2.6: Population of native structure as function of hyperparameters. Population is indicated in the color scale. The optimized population of native structures, when averaged on the training set (a), is by construction a monotonically increasing function of the integer p controlling the window size of the convolutional network in the reactivity channel, and a monotonically decreasing function of the regularization coefficients α_S and α_D. When averaged on the leave-one-out iterations of the cross-validation (CV) procedure (b), the dependency of the optimized population of native structures on these hyperparameters becomes non-trivial, as it results from a combination of model complexity (controlled by p) and regularization (controlled by α_S and α_D independently). The CV procedure serves as criterion for model selection, resulting in the selection of hyperparameters $\{p = 0, \alpha_S = 0.001, \alpha_D = 0.001\}$.

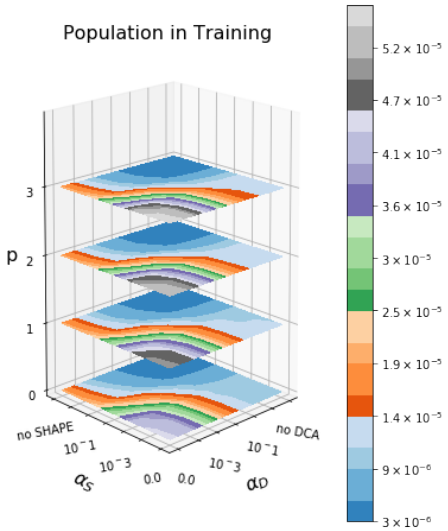
validation test is passed for all the splittings, indicating that the model selection procedure is capable to detect overfitting with all of the tested datasets. The model complexity is controlled by means of three handles: a regularization parameter acting on the one-dimensional pseudo-energies derived from reactivities ($0 \leq \alpha_S \leq \infty$), a regularization parameter acting on the two-dimensional pseudo-energies derived from DCA ($0 \leq \alpha_D \leq \infty$) and an integer controlling the size of the window used for the convolutional network ($p <= 3$).



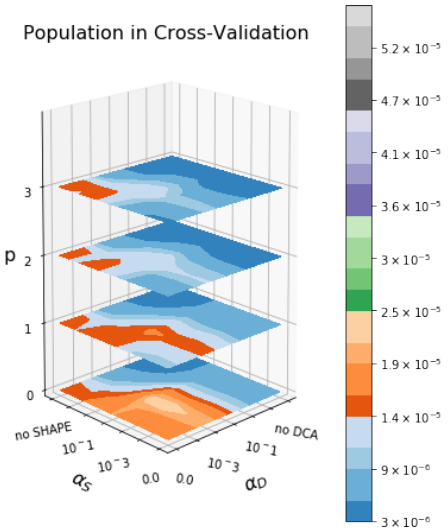
(a)



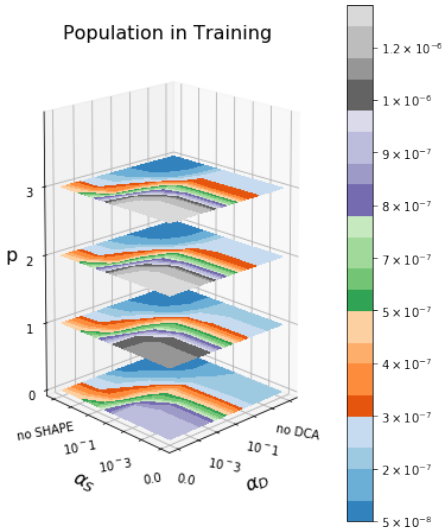
(b)



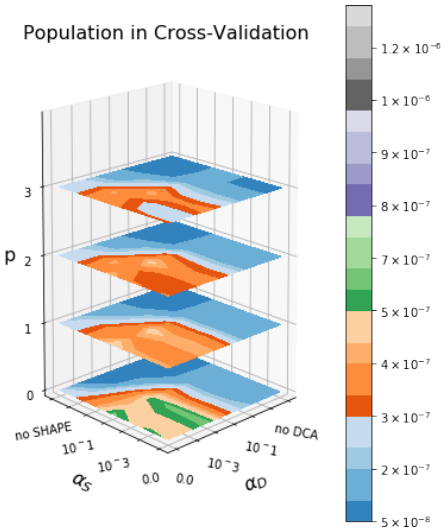
(c)



(d)



(e)



(f)

Figure 2.7: Training and leave-one-out average population as function of hyperparameters, for (a-b) splitting S1, (c-d) S2 and (e-f) S3.

When the performance of the model is evaluated on the training set, the model that better fits the data is the most complex one, with no regularization term ($\alpha_S = \alpha_D = 0$) and the largest tested window ($p = 3$) (Fig. 2.6a). The geometric average of the populations of native structures increases by ≈ 11 times with respect to that of the thermodynamic model alone. Training the model using only chemical probing data ($\alpha_D = \infty$), or only DCA data ($\alpha_S = \infty$), results in an increase of native population by ≈ 5 times and ≈ 3 times respectively, within the randomized set S3 (Table 2.2). The results in training and model selection for all the other splittings are reported in Fig.2.7

2.5.2 Model selection

In order to make the parametrization transferable, the leave-one-out cross-validation (CV) procedure described in Section 2.4 is performed: one of the 12 molecules at a time is left out of the training procedure and the increase in the native population for the left-out molecule is used as a transferability score. Overall, the average performance of the model on the left-out molecule shows a non-trivial dependence on the hyperparameters (Fig. 2.6b). All the models yield a performance in the cross-validation test equal or better than the thermodynamic model alone, but the best performance is obtained when choosing $\alpha_S = 0.001$, $\alpha_D = 0.001$ and $p = 0$. This model is thus selected as the one yielding the best balance between performance and transferability. Results obtained by using different randomizations of the training set are reported in the Supporting Information of our paper [52]. Whereas the precise set of optimal hyperparameters depends on the specific training set, sets of hyperparameters that perform well on a specific set tend to perform well for all of the tested training sets.

2.5.3 Validation

Finally, the performance of the selected model is evaluated on a dataset of 6 molecules that were not seen during training. This additional test is done in the spirit of nested cross-validation [65] in order to properly evaluate the transferability of the procedure. For the 6 test molecules (splitting S3 of Table 2.2), the introduced procedure leads to a boost of the population of the native structure by ≈ 19 times, on average (Fig. 2.8), right side of the vertical line), when using the selected model $\{\alpha_S = 0.001, \alpha_D = 0.001, p = 0\}$.

A side effect of targeting the population of native structures for model optimization and selection is the increase in the similarity between the predicted minimum free energy (MFE) and the experimental structures. This similarity is quantified using the Matthews Correlation Coefficient (MCC), as described in Section 2.4. Its average on the validation set is increased from 0.68 to 0.89 (Fig. 2.9, right side).

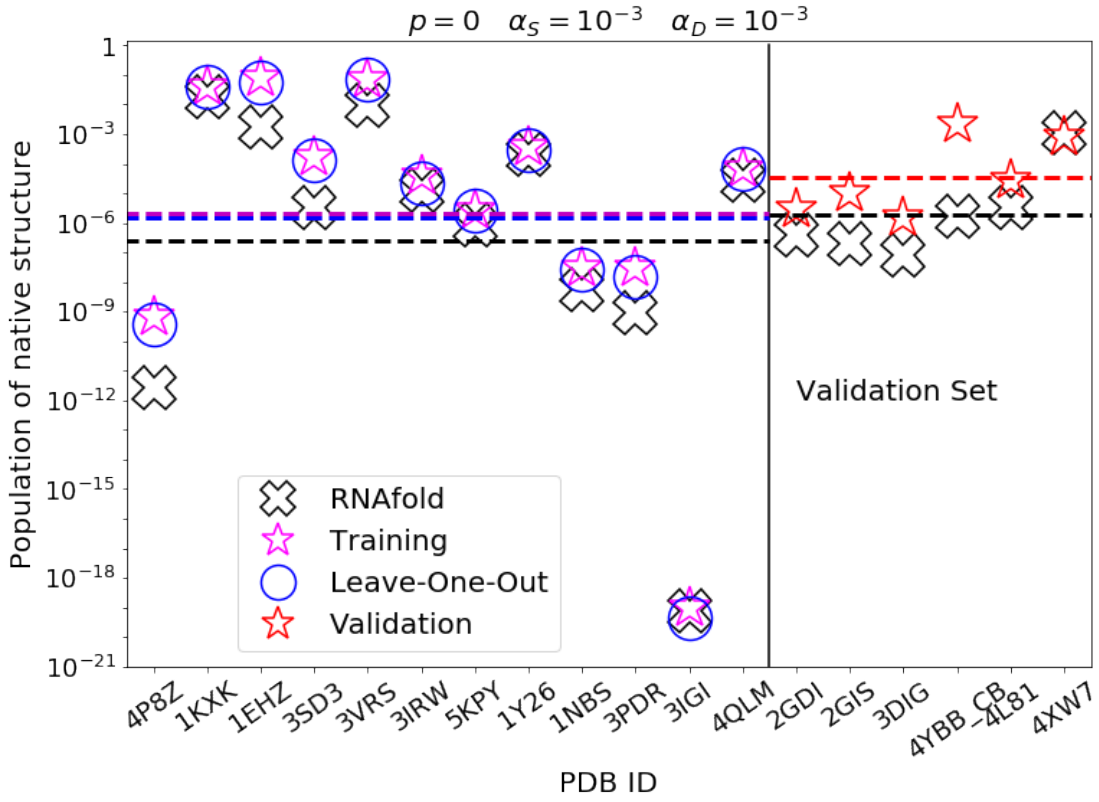


Figure 2.8: Comparison of results obtained with unmodified RNAfold and with the selected model. Hyperparameters are noted in the figure. Native structure populations obtained with unmodified RNAfold (black cross), with our trained model (magenta star on the training set, red star on the validation set) and in the leave-one-out procedure (blue circle, for each molecule the model is trained on all the other molecules in the training set) are reported. The corresponding averages are reported with dashed lines of the same color. The populations of native structures that we obtain with the trained model are always increased for molecules in the training set (left side of the vertical line) and almost always in the validation set (right side).

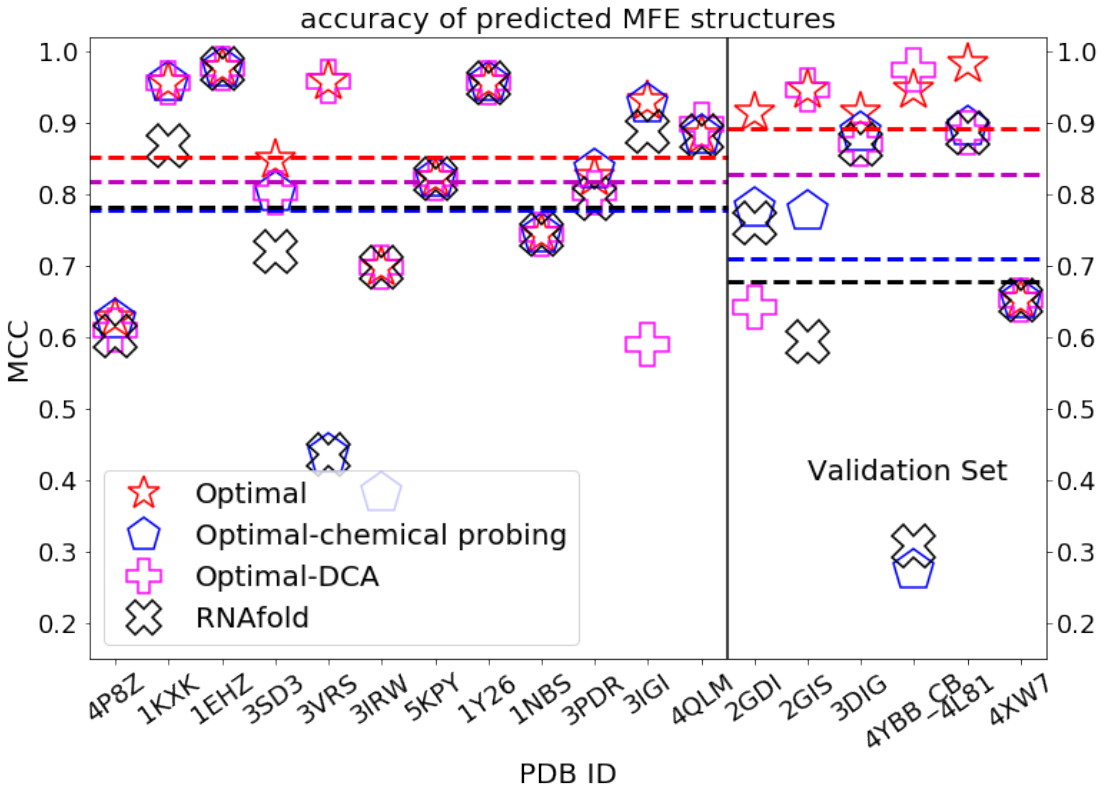
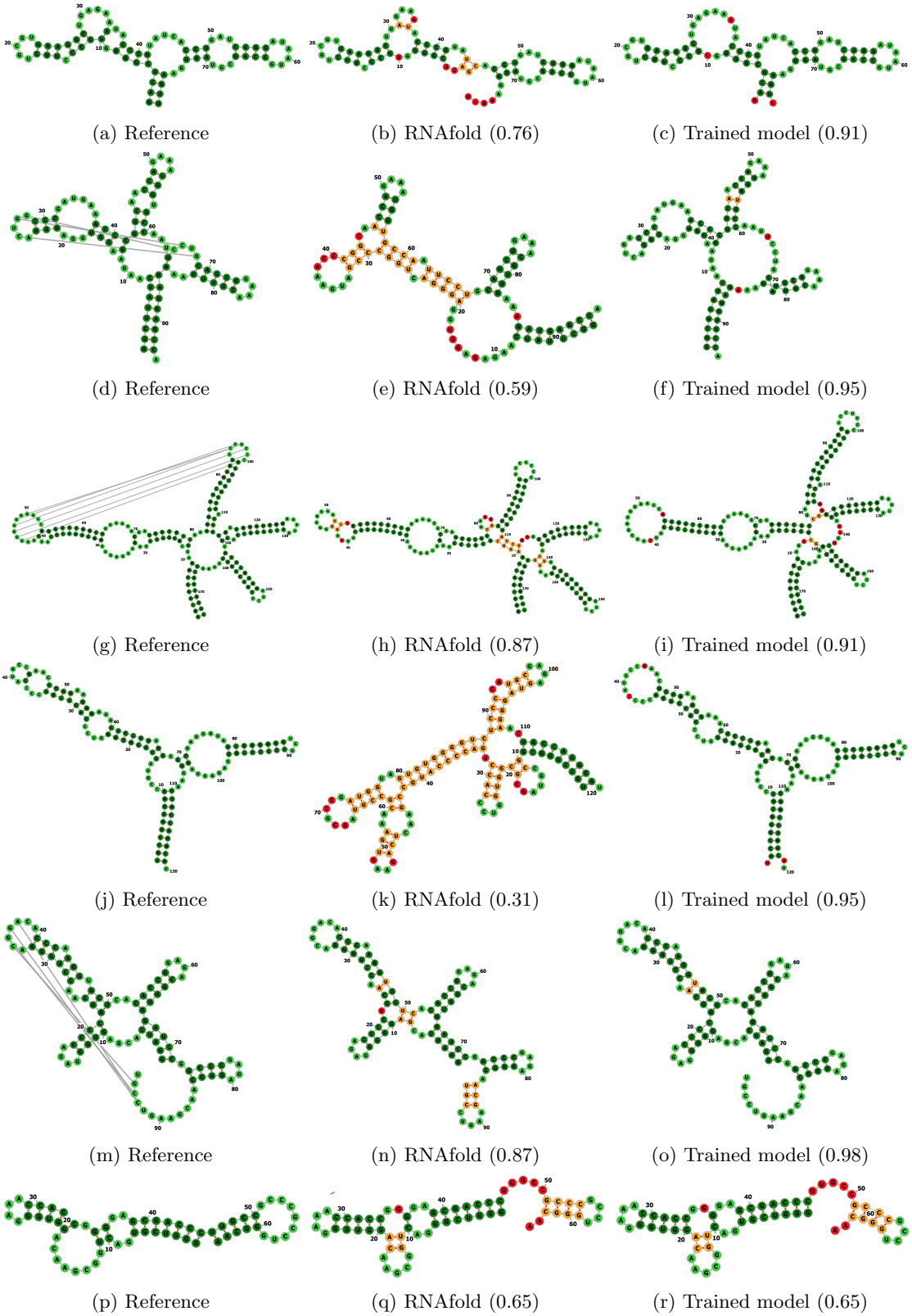


Figure 2.9: Matthews correlation coefficients between predicted MFE structures and reference native structures, as obtained with selected best (red star), DCA-only (pink cross), chemical probing-only (blue pentagons) models and with unmodified RNAfold (black cross). The corresponding averages are reported with dashed lines of the same color.



(a) Reference (b) RNAfold (0.76) (c) Trained model (0.91)

(d) Reference (e) RNAfold (0.59) (f) Trained model (0.95)

(g) Reference (h) RNAfold (0.87) (i) Trained model (0.91)

(j) Reference (k) RNAfold (0.31) (l) Trained model (0.95)

(m) Reference (n) RNAfold (0.87) (o) Trained model (0.98)

(p) Reference (q) RNAfold (0.65) (r) Trained model (0.65)

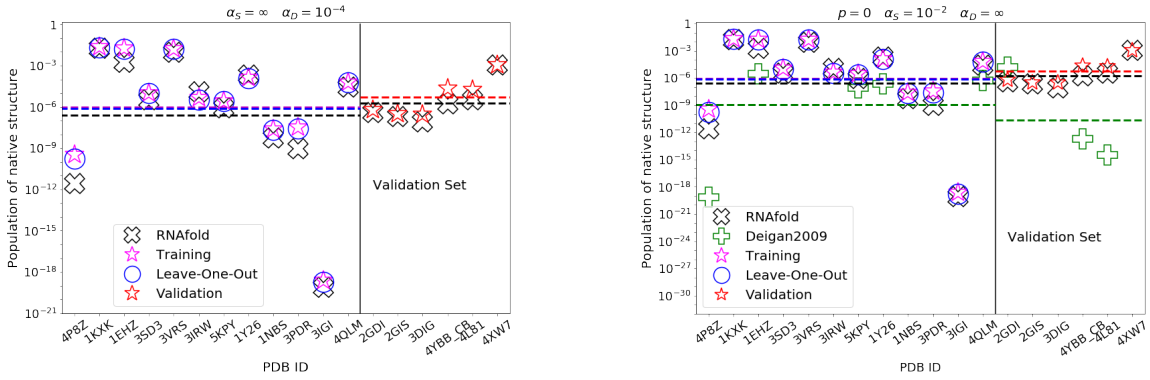
Figure 2.10: Comparison of reference native structure with MFE predictions of RNAfold and RNAfold modified with our model. True positive base-pairs and true negatives (unpaired) are reported in dark green and lime green, respectively. False positives and false negatives are reported in orange and red, respectively. MCC between prediction and reference is reported in parenthesis. All the relevant improvements in the prediction of these structures are reported in detail in the main text. All secondary structure diagrams are drawn with `forna` [66].

Specific changes in the predicted secondary structures are reported in detail in Fig. 2.10, where reference secondary structures are compared with MFE predictions made with unmodified RNAfold and with the selected model. In particular, for `2GDI` (Fig. 2.10a-c) our model recovers the correct structure of the 3-way junction loop (4-5:41-47:72-75); for `2GIS` (Fig. 2.10d-f) it recovers the correct structures of the hairpin loop (23-29) and the internal loop (17-21:31-38); for `3DIG` (Fig. 2.10g-i) the correct bulge loop (84-85:109-111) is recovered; for `4YBB` (Fig. 2.10j-l) the bulge loops (30-31:51-54) and (17-18:65-67), the internal loops (23-28:56-60) and (71-79:97-105), the 3-way junction (10-16:68-70:106-110) and the hairpin loop (86-90); for `4L81` (Fig. 2.10m-o) the 4-way junction (5-10:21-22:51-53:66-67) is correctly predicted; for `4XW7` (Fig. 2.10p-r) no change in MFE predictions observed with respect to using the unmodified ensemble. Considering all of the tested splittings of the dataset, the average MCC of minimum free energy structure predictions is increased from 0.72 ± 0.22 to 0.90 ± 0.10, implying both an increased average and a decreased variance. As can be seen from Fig. 2.10, some of the structures in the dataset contain pseudoknots. This kind of pairing is forbidden in RNAfold structure predictions, thus we do not include it in the estimation of MCC. Nonetheless, data from both chemical probing and coevolution analysis in principle contain information about pseudoknots, and it is possible to examine how reactivities and DCA scores of pseudoknotted nucleotides are mapped into pseudo-energy terms in our optimal model. We notice that the average value of pairwise DCA pseud-energies coupling pseudoknotted pairs $\langle \lambda_{ij} \rangle_{PK} = -0.087$ is comparable to the average of those coupling base pairs $\langle \lambda_{ij} \rangle_{BP} = -0.094$, so that they have almost the same effect in favouring pairing (free-energy term $\lambda_{ij} s_{ij} < 0$ for $s_{ij} = 1$). The difference between the two values is negligible when compared with the average DCA pseudo-energy coupling unpaired nucleotides $\langle \lambda_{ij} \rangle_{UP} = 0.447$. Reactivity-driven single-point pseudo-energies favour unpaired states on average (free-energy term $-\lambda_i s_i > 0$ for $s_i = 1$), but the effect on pseudoknotted nucleotides $\langle \lambda_i \rangle_{PK} = -0.142$ and on base-paired nucleotides $\langle \lambda_i \rangle_{BP} = -0.125$ is approximately half of that on unpaired nucleotides $\langle \lambda_i \rangle_{UP} = -0.284$. Even though in our optimal model the pairing of pseudoknotted nucleotides is boosted with almost the same intensity of base-paired nucleotides, eventually small values are predicted for the corresponding pairing probabilities, with average $\langle p_{ij} \rangle_{PK} = 0.06$, an no significant change with respect to the unmodified ensemble. This is due to the fact that the thermodynamic model only allows structures with nested pairs.

It is also possible to test the scenarios where only DCA data or only chemical probing data are available. In scenarios where only DCA information is used ($\alpha_S = \infty$), the best performance in CV is obtained using the model with $\alpha_D = 0.0001$: $10\times$ increase in population and average MCC $= 0.83$, as shown in Fig. 2.11a. This model is thus transferable to the validation set yielding a significant increase in both the population of the native structures and in MFE structure accuracy.

In the case of chemical probing-only information ($\alpha_D = \infty$), the best performance in CV is obtained using the model with hyperparameters $\alpha_S = 0.01$ and $p = 0$, yielding a $3\times$ increase in population and average MCC $= 0.71$, as shown in Fig. 2.11b. Interestingly, whereas reactivity-only models perform systematically better in training than DCA-only models, their performance in CV is systematically lower, suggesting a lower transferability to unseen data, and thus a larger risk for reactivity-driven pseudo-energies to be overfitted. This might be related to the high heterogeneity of the chemical probing data used here, that makes it difficult to fit transferable parameters.

The procedure to compute pseudo-energies from reactivities presented here can be compared with the one introduced by Deigan et. al. [11]. The functional form of pseudo-energy is that of Eq. 1.8i and it couples the pairing state of at least two stacked base-pairs. Since the Deigan's method requires SHAPE data normalized with a different procedure, the comparison is made only for those molecules for which the normalized reactivities are available and reported in



(a) (b)

Figure 2.11: Comparison of results obtained with unmodified RNAfold and with the selected model using only DCA scores. Hyperparameters are noted in the figure. Native structure populations obtained with unmodified RNAfold (black cross), with our trained model (magenta star on the training set, red star on the validation set) and in the leave-one-out procedure (blue circle, for each molecule the model is trained on all the other molecules in the training set) are reported. Results obtained with RNAfold+Deigan's method (green plus) are reported when available. The corresponding averages are reported with dashed lines of the same color. The populations of native structures that we obtain with the trained model are almost always increased for molecules both in the training (left side of the vertical line) and in the validation set (right side), with resepct to unmodified RNAfold and, in the case of chemical-probing only, to RNAfold+Deigan's method.

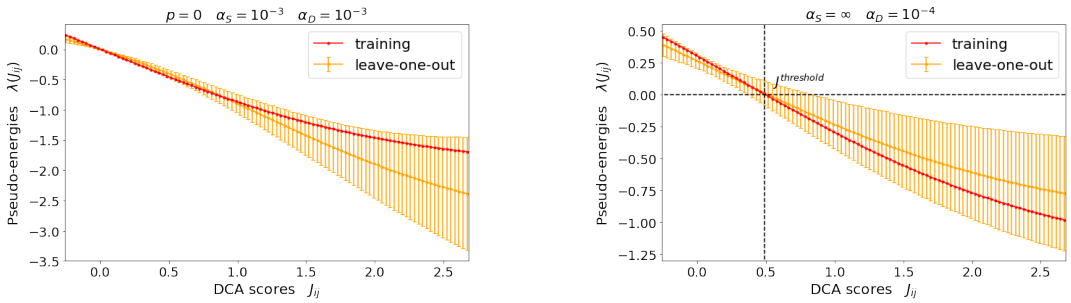
Ref. [55]. Remarkably, our procedure leads to significantly better results both for molecules that are included in the training set (*e.g.*, `4P8Z`, `1EHZ`, `5KPY`,`1Y26`,`4QLM` in Fig. 2.11b), and for most of the RNAs included in the validation set (`4YBB_CB` and `4L81` in the right side of Fig. 2.11b).

2.5.4 Network parameters

The external validation thus confirms the transferability of the selected model. Going back to its network architecture, an interpretation of its optimal parameters is here proposed.

In the DCA channel, DCA couplings are mapped into pseudo-energies through a double-layered neural network, resulting in a non-linear function reported in Fig. 2.12a. Errors on these pseudo-energies are computed as standard deviations of the distribution of fitted λ_{ij} values obtained from the leave-one-out iterations. Notice that the errors around $J_{ij} \simeq 0$ are significantly lower due to the larger statistics across the dataset of J_{ij} in the near-zero range, as reported in Fig.2.13. Pseudo-energies are found to decrease with increasing DCA coupling value, consistent with the interpretation that large couplings should correspond to co-evolutionarily related and thus likely paired nucleobases [67]. A more detailed interpretation of these pairing pseudo-energies is possible if we restrict to models taking only DCA couplings as input ($\alpha_S = \infty$). The corresponding non-linear function is reported in Fig. 2.12b. The overall shape is consistent with that obtained fitting all the data (Fig. 2.12a), but the zero of this function can be straightforwardly interpreted as the threshold for penalizing or favoring base pairing. The resulting value is $J^{\text{threshold}} = 0.49$ consistent with the typical thresholds obtained in [50] with a different optimization criterion, based on the accuracy of contact predictions, and fitted on a larger dataset. This consistency further confirms the transferability of the non-linear function reported here.

Chemical probing reactivities are mapped into pseudo-energies affecting the population of



(a) (b)

Figure 2.12: Properties of the optimized neural network. For the DCA channel, the optimized function mapping DCA couplings J_{ij} into pairing pseudo-energies λ_{ij}, for both (a) the selected model and (b) the best performing model with restriction to only DCA input. When trained on the whole training set (red) the activation function is consistent with the average on the leave-one-out training subsets (orange). Error bars are computed as standard deviations and are significantly lower in the region of DCA couplings around zero, as couplings lying in that region are more frequent. The trained function maps high (respectively, low) DCA coupling values to pseudo-energies favoring (respectively, disfavoring) the corresponding pairings, thus affecting the population of the structures including the specific pair. When restricting to (b) models including only DCA input, the threshold value of the coupling $J^{threshold}$ between disfavored and favored pairing corresponds to the zero of the activation function, as indicated by the dashed line.

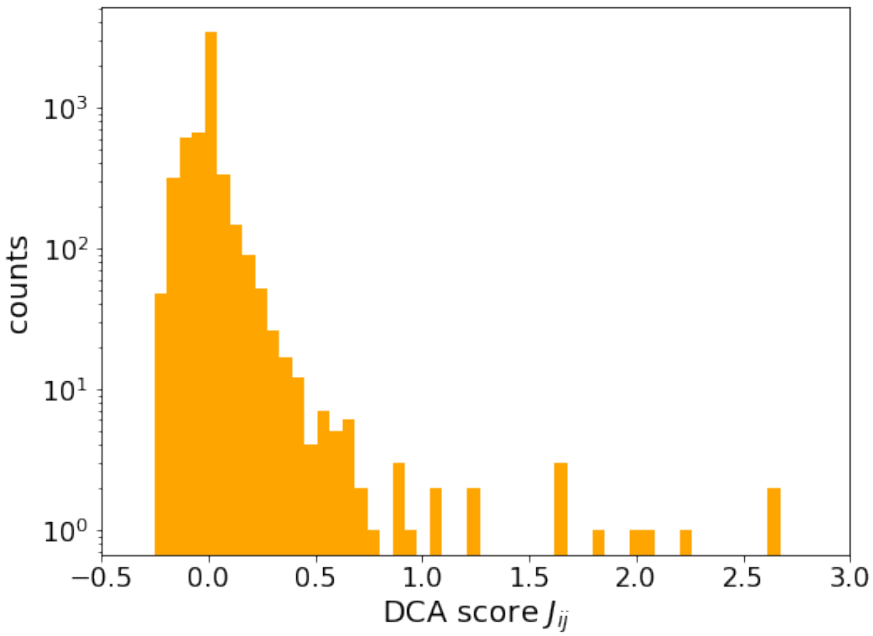


Figure 2.13: Histogram of the DCA scores obtained from the whole dataset.

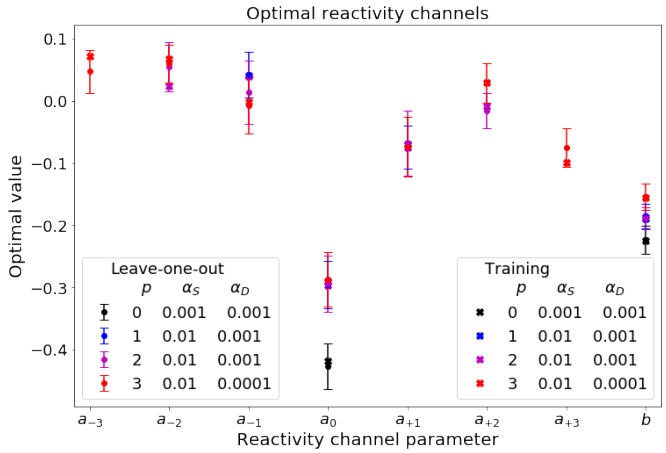


Figure 2.14: For the chemical mapping channel, optimal values of model parameters are shown for the selected model (black) with hyperparameters $\{\alpha_S = 0.001, \alpha_D = 0.001, p = 0\}$, and for the sub-optimal models with $p > 0$. All the training results (cross) lie within the leave-one-out errors (dots with error bars), indicating robustness of the minimization procedure against cross-validation. Coefficients $\{a_{-k}, \ldots, a_{+k}\}, \quad k > 0$ weighting reactivities up to the k-th nearest-neighbors of a nucleotide, report the minor contributions of the local reactivity pattern in addition to the nucleotide's own reactivity.

individual nucleotide pairing states through a single convolutional layer with a linear activation function. When evaluated on the training set, the best performance is obtained with models including up to the maximum tested number of nearest neighbors ($p = 3$). In these models, for each nucleotide, the network input vector includes reactivities from the third nearest-neighbor upstream to its third nearest-neighbor downstream along the sequence. The activation coefficients $\{a_k, \quad k = -3, \ldots, +3\}$ weight the contribution of each nucleotide in the neighbor window. Despite the performance improvement on the training set, transferability to data not seen during the training phase is best preserved in the model that retains only the contribution from the a_0 term, confirming that the reactivity of a nucleotide is maximally affected by its pairing state. Correlations between SHAPE reactivity and sugar flexibility is reported in the literature [68, 69, 70], and is only indirectly related to the pairing state of a nucleotide. Nevertheless, reactivity information can be used to systematically improve predictions at the base-pairing level. In particular, $a_0 < 0$ (see Fig. 2.14, black) so that the pairing of a highly reactive nucleotide is unfavored and vice-versa for nucleotides with low reactivity.

On the other hand, the best (suboptimal) neighbor-including models (*i.e.*, with $p > 0$) still yield comparable results with respect to the selected one and significant improvements as well with respect to thermodynamic model alone. Figure 2.14 reports the sets of optimal parameters with $p > 0$. At each increment of p, when two new parameters a_p and a_{-p} are introduced, all the shared subsets $\{a_{p-1}, \ldots, a_{-p+1}\}$ overlap significantly, and a number of features are shared as well. First, for all the optimal choices of $p > 0$, the sum of the weights $\sum_{i=-p}^{p} a_i$ is negative, so that the pairing of a nucleotide in a highly reactive region is unfavored, and vice-versa for regions of low reactivity. The largest contribution still arises from the a_0 term, but it is slightly lower in absolute value, to compensate for neighbor corrections. For each pair (downstream and upstream) of k-th nearest-neighbors, the combination of the a_0 and a_{+k} (a_{-k}) contributions can be interpreted as a forward (backward) finite-difference operator estimating the k-th order derivative of the reactivity with respect to the position in the sequence. These contributions map local downward trends of the reactivity profile into pseudo-energies, thus providing a sort

of normalization for the reactivity of the central nucleotide with respect to that of its neighbors. As the order of the derivatives increase from the first, weights become lower such that the corresponding corrections progressively decrease in importance. It is interesting to notice that the finer these corrections are, the more the corresponding parameters tend to be overfitted to the training set.

2.6 Discussion

We built up a total of 196 models to map simultaneously reactivities and DCA scores into pseudo-energy terms coupling, respectively, the pairing state of individual nucleotides and that of specific pairs of nucleotides. Each model is defined by tunable hyperparameters controlling the width of the windows used to process reactivities and the strength of the regularization terms applied to chemical probing and DCA data. The dataset is *a priori* split randomly into a training set and a validation set (12 and 6 molecules respectively). Training, model selection and validation are repeated for different random splittings of the dataset, in order to decrease the chance of introducing a bias towards specific structures or features, and ensuring the robustness of the procedure. The whole procedure, from training to model selection, is automatic so that new parameters could be straightforwardly obtained using new chemical probing and DCA data and new crystallographic structures, allowing for a continuous refinement of the proposed structure prediction protocol. Training one model required 20 minimizations that were performed in parallel on nodes containing 2 E5-2683 CPU each, using 20 cores. Each minimization took approximately 30 minutes, though the exact time depends on the value of p and on the system size. 4x7x7=196 minimizations were done to scan the hyperparameter space. 12 separate models needed to be trained for the leave-one-out. Notably, the dependence between the minimizations can be taken into account allowing them to be largely run in parallel. In practice, if 288 nodes are simultaneously available, the full minimization for 12 systems can be run in approximately 8 hours. In the dataset presented in Section 2.1, some reactivities are taken from available experimental data. Other reactivities are provided by our experimental collaborators at Sattler's Lab so as to increase the number of systems for which both co-evolutionary data and reactivities are available. DCA scores are based on ClustalW alignments [71] so that they are not manually curated with prior structural information. We notice however that classification of sequences in RFAM is performed including structural information, when available. In addition, co-evolutionary information might be difficult to extract for poorly conserved long non-coding RNAs.

The model selected via CV is defined by hyperparameters $\{p = 0, \alpha_S = 0.001, \alpha_D = 0.001\}$. The best balance between performance and transferability is thus obtained when not incorporating reactivities from neighboring nucleotides in the pairing state of a nucleotide. This model is systematically capable of predicting a higher population for the native structure. The model that is selected using only chemical probing data yields better results in population than what obtained with Deigan's method [11], which is accounted for best state-of-the-art method [42] among those based on SHAPE reactivities only.

Results obtained with our selected model confirm that the reactivity of a nucleotide is a good indicator of its own pairing state [42]. We also observe that the reactivity of neighbors correlates too with the pairing state of a nucleotide, as expected from the preliminary study reported in Section 2.2. However, the pairing state of neighboring nucleotides is implicitly taken into account in the RNAfold model, that includes energetic contributions for consecutive base pairs, implying that the explicit inclusion might not be required. More precisely, the need for a larger number of parameters to be trained when increasing the p hyperparameter might not be compensated

by a sufficient improvement in the prediction performance. Interestingly, in a previous version of this work based on a smaller dataset and on different thermodynamic parameters [26] the most transferable model identified had $p = 2$ (see `https://arxiv.org/abs/2004.00351v1`).

In perspective, the model can be extended to include additional features of the chemical probing experiments that may be related to non-canonical interactions and three-dimensional structure.

Although our selected model is trained to maximize the population of the individual reference structure as obtained by crystallization experiments, it can still report alternative structures. Whereas we did not investigate this issue, alternative low-population states might be highly relevant for function. Compatibly with that, the absolute population of the native structure remains significantly low (from $\approx 10^{-8}$ to $\approx 10^{-7}$), but is still one of the highest in the ensemble. In particular, the individual structure with highest population (minimum free-energy structure) with our method is closer to the reference crystallographic structure than the one predicted by thermodynamic parameters alone on systems not seen during training.

Importantly, all the data and the used scripts are available at `https://github.com/bussilab/shape-dca-data` and can be used to fit the model over larger datasets. In order to avoid overfitting, repeating the leave-one-out procedure to select the most transferable model is suggested, whenever new independent data is added to the dataset. In principle the model can be straightforwardly extended to include any chemical probing data that putatively correlates with base pairing state [72] or other types of experimental information that correlate with base-pairing probabilities [73]. Training on a larger set of reference structures and using more types of experimental data is expected to make the model more robust and open the way to the reliable structure determination of non-coding RNAs.

Chapter 3

Cooperative effects in chemical probing experiments

In this Chapter, we present an original Molecular Dynamics simulation protocol that allows for the investigation of collective behaviors in chemical probing experiments. The study is based on the hypothesis that the reactivity profiles observed in chemical probing experiments are affected by the cooperativity of nucleotides in ligand binding, with the result of a more complex correlation between reactivity and structure, beyond the pairing state of individual nucleotides. This hypothesis is partly corroborated by the correlations observed in Chapter 2 between the pairing state of a nucleotide and the local reactivity pattern of its neighbors along the sequence. Molecular Dynamics simulations are used to predict the cooperative behavior of the nucleotides of a well-studied RNA motif, the GNRA tetraloop, in presence of a SHAPE reagent at finite concentration. We select 1-methyl-7-nitroisatoic anhydride (1M7) as reagent, since it involves shorter reaction times than other probes, and is thus one of the more efficient. The force-field parametrization of 1M7 is carried out and presented here for the first time, to our knowledge. Experimentally, the presence of cooperative effects in SHAPE profiling can be detected by repeating the experiment at varying concentration of the reagent. We then present an innovative method that allows to compute grand canonical averages at tunable values of reagent concentration, from trajectories obtained with canonical Molecular Dynamics simulations at fixed numbers of reagents. Cooperativity towards binding 1M7 is then computed for each pair of nucleotides, and pairs that show statistically significant cooperativity under multiple-hypothesis testing are subjected to visual analysis to investigate correlations with local structural conformations.

This Chapter will be used as a draft for future publication.

3.1 Grand canonical ensemble reweighting of Molecular Dynamics

In order to describe physical situations in which the number of particles is varying, the grand canonical ensemble is necessary. The fluctuations in the number of particles are controlled by the chemical potential μ, which in this ensemble is a fixed quantity together with volume V and temperature T. The grand canonical ensemble can be thought of as a canonical ensemble coupled to a particle reservoir that can gain or lose particles without appreciably changing its chemical potential. If we consider two sub-regions A and B of the system coupled to the particle

reservoir, the probability of having N_A particles in region A in the grand canonical ensemble is

$$P_A^{GC}\left(N_A\right) \propto \Omega_A\left(N_A\right) e^{-\mu N_A/RT} \tag{3.1}$$

as well as the probability of having N_B particles in region B is

$$P_B^{GC}\left(N_B\right) \propto \Omega_B\left(N_B\right) e^{-\mu N_B/RT} \tag{3.2}$$

where Ω_A and Ω_B are the canonical partition functions for region A and B, respectively. Our aim is to compute averages in the grand canonical ensemble, by just using these probabilities as weights for trajectory frames collected by a set of simulations run in the canonical ensemble, each at a different fixed number of particles N. In each of these simulations the probability to observe $N_{A/B}$ particles in region A/B is

$$P_{A/B}^{N}\left(N_{A/B}\right) \propto \Omega_A\left(N_{A/B}\right) \Omega_B\left(N - N_{A/B}\right) \tag{3.3}$$

proportional to the number of states with N_A particles in region A and $N - N_B$ particles in region B, or vice-versa. We then consider a set of N_{max} simulations each run at a different fixed number of particles N, ranging from 1 to N. The probability to sample t_{Nk} frames in which there are k particles in region A and $N - k$ particles in region B is then

$$P\left(t_{Nk}\right) \propto \prod_N \prod_k \left(c_N \Omega_A(k) \Omega_B(N-k)\right)^{t_{Nk}} \tag{3.4}$$

where the normalization coefficients $\{c_N\}$ are required to ensure that, at fixed N, the sum of the probabilities $P_{A/B}^{N}\left(N_{A/B}\right)$ of Eq. 3.3 over all the possible combinations of N_A and N_B is equal to one:

$$\sum_k c_N \Omega_A(k) \Omega_B(N-k) = 1 \quad \forall N \in [1, \ldots, N_{max}] \tag{3.5}$$

In order to estimate the most likely values of Ω_A and Ω_B from our simulations run in the canonical ensemble, we minimize the minus log-likelihood $\mathcal{L}$ of Eq. 3.4, with the constraint that the normalization of Eq. 3.5 is satisfied,

$$\mathcal{L} = -\sum_N \sum_k t_{Nk} \log\left(c_N \Omega_A(k) \Omega_B(N-k)\right) - \sum_N \lambda_N \left(\sum_k c_N \Omega_A(k) \Omega_B(N-k) - 1 \right) \tag{3.6}$$

with respect to Ω_A, Ω_B, the normalization coefficients $\{c_N\}$ and the corresponding Lagrangian multipliers $\{\lambda_N\}$. Setting to zero the gradient of $\mathcal{L}$ with respect to these parameters yields the following equations:

$$\frac{\partial \mathcal{L}}{\partial \lambda_N} = \sum_k c_N \Omega_A(k) \Omega_B(N-k) - 1 = 0 \tag{3.7a}$$

$$\frac{\partial \mathcal{L}}{\partial c_N} = -\frac{\sum_k t_{Nk}}{c_N} - \lambda_N \sum_k \Omega_A(k) \Omega_B(N-k) = 0 \tag{3.7b}$$

$$\frac{\partial \mathcal{L}}{\partial \Omega_A(k)} = -\frac{\sum_N t_{Nk}}{\Omega_A(k)} - \sum_N \lambda_N c_N \Omega_B(N-k) = 0 \tag{3.7c}$$

$$\frac{\partial \mathcal{L}}{\partial \Omega_B(k)} = -\frac{\sum_N t_{N,N-k}}{\Omega_B(k)} - \sum_N \lambda_N c_N \Omega_A(N-k) = 0 \tag{3.7d}$$

From the equation 3.7a, we obtain

$$\sum_k \Omega_A(k)\Omega_B(N-k) = \frac{1}{c_N} \tag{3.8}$$

that we replace in the second term of the equation 3.7b yielding

$$\lambda_N = -\sum_k t_{Nk} \tag{3.9}$$

Before replacing these terms into the last two equations 3.7c and 3.7d, we define the histogram $A_k = \sum_k t_{Nk}$ of the number of times that, in the whole set of N_{max} trajectories, the reagent was in region A, and the same histogram for region B, $B_k = \sum_k t_{N,N-k}$, as well as the total number of frames of each trajectory $L_N = \sum_k t_{Nk}$. Equations 3.7c and 3.7d are rewritten in these terms as

$$\begin{aligned} \Omega_A(k) &= \frac{A_k}{\sum_N L_N c_N \Omega_B(N-k)} \\ \Omega_B(k) &= \frac{B_k}{\sum_N L_N c_N \Omega_A(N-k)} \end{aligned} \tag{3.10}$$

These equations can be solved iteratively through the procedure reported in Alg. 1. Noticeably,

Algorithm 1 Estimating Ω_A and Ω_B

1 $\Omega_A^{i=0} \leftarrow A$
2 $\Omega_B^{i=0} \leftarrow B$
3 threshold $\leftarrow 10^{-30}$
4 **for** $i \in \{1, \ldots, N_{steps}\}$ **do**
5 $\quad c[N] \leftarrow 1/\sum_{k=0}^{N} \Omega_A^{(i-1)}[k] \cdot \Omega_B^{(i-1)}[N-k]) \quad \forall N \in \left[1, \ldots, N_{max}\right]$
6 $\quad \Omega_A^{(i)}[k] \leftarrow A[k] \,/\, \sum_{N=1}^{N_{max}} L[N] \cdot c[N] \cdot \Omega_B^{(i-1)}[N-k] \quad \forall k \in \left[0, \ldots, N_{max}\right]$
7 $\quad \Omega_B^{(i)}[k] \leftarrow B[k] \,/\, \sum_{N=1}^{N_{max}} L[N] \cdot c[N] \cdot \Omega_A^{(i-1)}[N-k] \quad \forall k \in \left[0, \ldots, N_{max}\right]$
8 $\quad \varepsilon \leftarrow \sum_{k=0}^{N_{max}} \left[\left(\Omega_A^{(i)}[k] - \Omega_A^{(i-1)}[k]\right)^2 + \left(\Omega_B^{(i)}[k] - \Omega_B^{(i-1)}[k]\right)^2 \right]$
9 $\quad \Omega_B^{(i)} \leftarrow \Omega_B^{(i)} / \Omega_B^{(i)}[0]$
10 $\quad f \leftarrow \Omega_B^{(i)}[1] / \Omega_B^{(i)}[0]$
11 $\quad \Omega_B^{(i)}[k] \leftarrow \Omega_B^{(i)}[k] / f^k \quad \forall k \in \left[0, \ldots, N_{max}\right]$
12 $\quad \Omega_A^{(i)}[k] \leftarrow \Omega_A^{(i)}[k] / f^k \quad \forall k \in \left[0, \ldots, N_{max}\right]$
13 $\quad$ **if** $\varepsilon <$ threshold **then**
14 $\quad\quad$ **break**

line 9 to 12 provide a normalization of Ω_A and Ω_B such that $\Omega_A(k=0) = 1$ and $\Omega_B(k=0) = \Omega_B(k=1) = 1$. In this way the free-energy of the state with no particle at all is set to zero, as well as the free-energy cost for adding the first particle to region B. Since the chemical potential

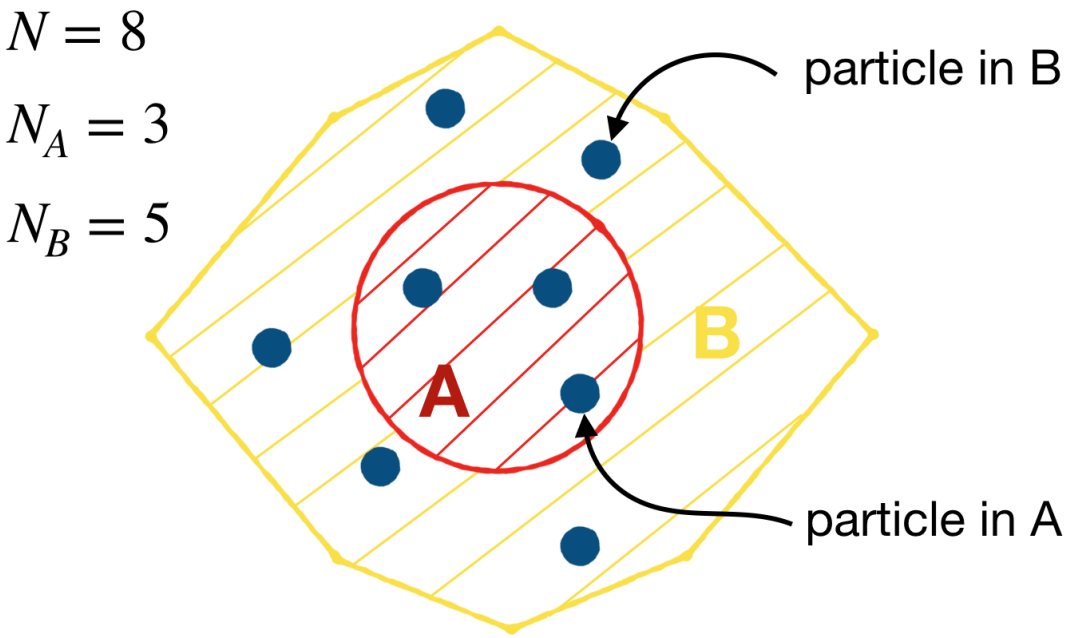


Figure 3.1: An illustration of the division of space into sub-regions A and B. Particles can travel one sub-region to the other. At fixed concentration of particles in B, $\propto N_B$, grand canonical averages of quantities in region A can be computed by reweighting frames relatively to the number N_A of particles in A.

μ is defined up to a constant, Ω_A and Ω_B are invariant with respect to scaling by an arbitrary factor f each, and to scaling each k-th component of $\Omega_{A/B}$ by the k-th power of the same factor f^k. By choosing $f = \Omega_B\left(k=1\right)/\Omega_B\left(k=0\right)$ the normalization is accomplished and the scaling invariance is removed.

Once the Eqs. 3.10 are solved, we can proceed to find the chemical potential μ to be inserted in Eqs. 3.1 and 3.2. This quantity can be determined as a function of the grand canonical average of the number of particles either in region A or B:

$$\langle N_{A/B}\rangle_{GC} = \sum_{k}^{N_{max}} k \cdot P^{GC}_{A/B}\left(k\right) = \frac{\sum_{k}^{N_{max}} k \cdot \Omega_{A/B}(k) e^{-\mu k/RT}}{\sum_{k}^{N_{max}} \Omega_{A/B}(k) e^{-\mu k/RT}} \tag{3.11}$$

We can thus exploit these relations to impose the desired concentration in either region A or B, and use the corresponding μ to weight each trajectory frame with respect to the number of particles it has in the other region. The procedure to find the value of μ corresponding to the desired concentration in region B, using the bisection method is reported in Alg. 2.

Algorithm 2 Estimating μ as function of the number of particles in B

function NB_OF_MU(μ)

$OB \leftarrow$ previously found Ω_B

$NB \leftarrow$ wanted N_B

$P_B[k] \leftarrow OB[k] \cdot e^{-\mu k/RT} \quad \forall k \in \left[0, \ldots, N_{max}\right]$

$P_B \leftarrow P_B / \sum_{k=0}^{N_{max}} P_B$

$N_B^{est} \leftarrow \sum_{k=0}^{N_{max}} k \cdot P_B[k]$

return $\log N_B^{est} - \log NB$

find the root of NB_OF_MU through an optimized bisection routine

3.2 Simulations of SHAPE dynamics with grand canonical ensemble reweighting

In order to investigate the cooperative behavior of nucleotides in chemical probing, we present here a Molecular Dynamics protocol developed to simulate the dynamics of a GNRA tetraloop in presence of 1-methyl-7-nitroisatoic anhydride (1M7), an efficient reagent used for SHAPE (Selective 2′-Hydroxyl Acylation analyzed by Primer Extension) probing. The GNRA tetraloop is chosen as the subject of this study because this type of motif has some well-established properties: it presents (a) highly stable secondary structure along with (b) rich dynamics involving multiple tertiary contacts [74], that could lead to significant structural changes when in contact with SHAPE reagents; noticeably, (c) in SHAPE experiments the GNRA tetraloop presents a typical reactivity pattern. We choose to simulate a single loop motif rather than duplexes or larger structures in order to keep computational costs low, under the hypothesis that long-range effects are negligible. We expect this hypothesis to be reasonable as there is no evidence of conformational rearrangements due to SHAPE chemistry, rather than at a local scale [69].

3.2.1 The GAAA tetraloop of SAM-I riboswitch

The GNRA tetraloop under study is part of the sequence of SAM-I riboswitch, the crystal structure of which is annotated in the PDB entry 2GIS. Notice that the sequence of 2GIS, along with a corresponding SHAPE reactivity profile, are included in the dataset presented in Chapter 2. In order to reduce computational costs without perturbing excessively the stability of the tetraloop structure, the stretch gcgGAAAcgu is cut from the sequence at positions 71 and 80. A representation of the resulting molecule is shown in Fig. 3.2. The stretch obtained in this way consists in a sequence of three base pairs, namely G71-U80, C72-G79 and G73-C78, plus the tetraloop under study: G74-A-A-A77. The starting configuration for this molecule is obtained by extracting the coordinates of the corresponding atoms from the PDB 2GIS entry, from the G71-O5′ to the U80-C6. The closing base-pair G71-U80 of the sequence is observed to unpair in those simulations where a larger number of reagents is used. Since a complete unfolding of the motif is not expected when the whole molecule is probed, a harmonic restraint is applied to

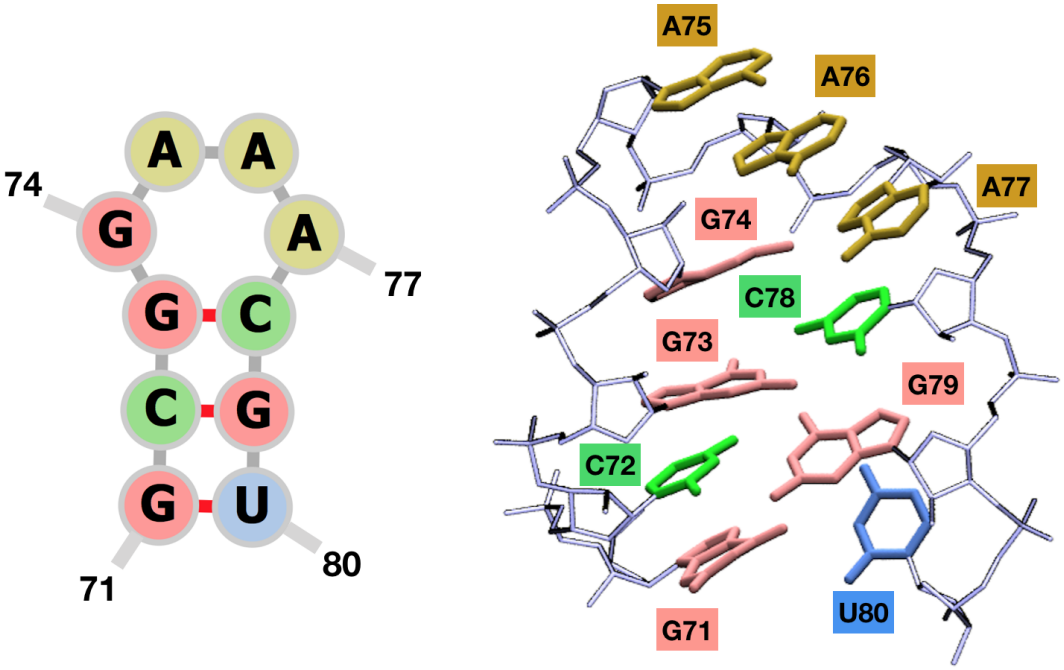
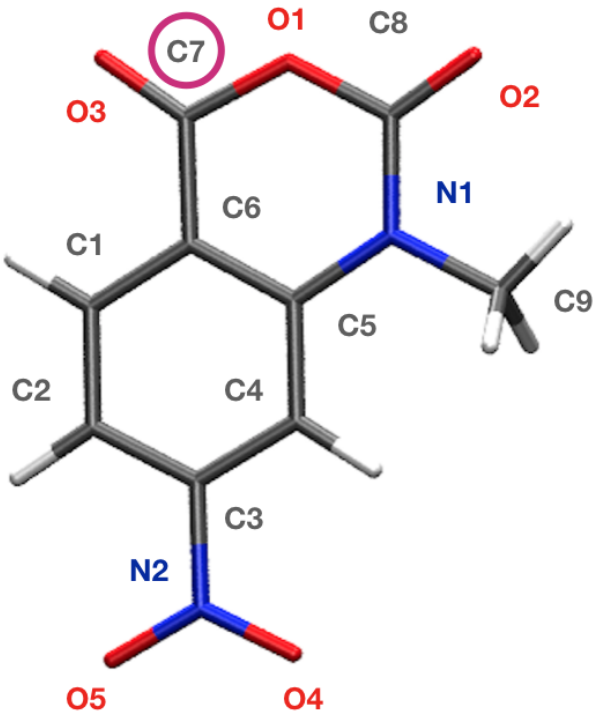


Figure 3.2: Secondary and tertiary structure of the G71-A-A-A-U80 tetraloop from 2GIS.



Atom	Charge (e)
C1	0.013677
H1	0.165938
C2	-0.301276
H2	0.214747
C3	0.213481
N2	0.680265
O4	-0.418660
O5	-0.418660
C4	-0.403042
H3	0.221441
C5	0.431676
N1	-0.308580
C8	0.913611
O2	-0.562770
C9	-0.336386
H4	0.158142
H5	0.158142
H6	0.158142
C6	-0.413908
C7	0.925472
O3	-0.538891
O1	-0.552561

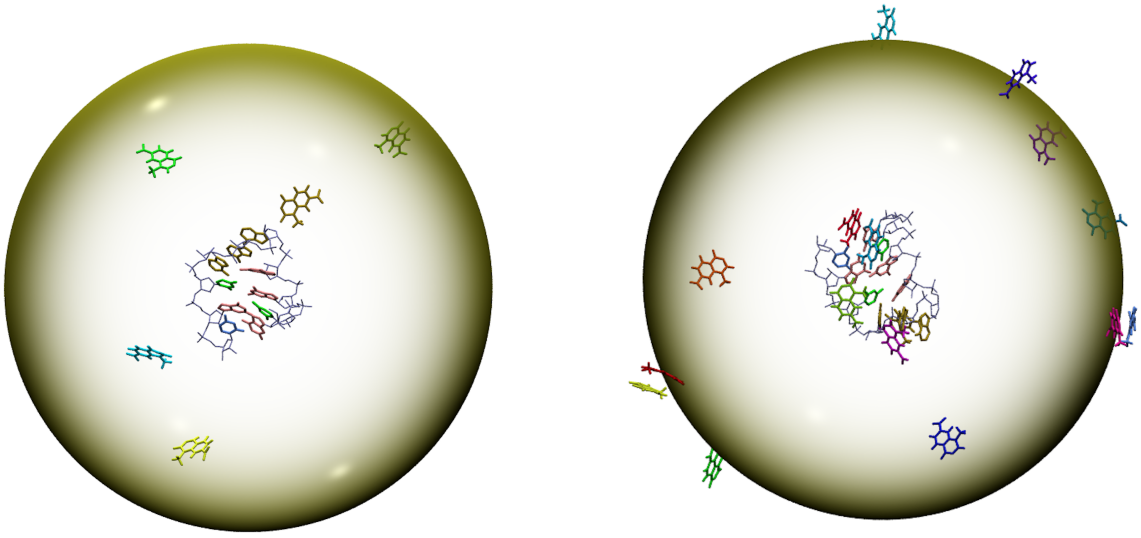
Figure 3.3: Optimized structure of 1M7 and charges of atoms in the 1M7 topology as obtained via Antechamber using the RESP method.

the hydrogen bonds between 71G/O6 and 80U/N3 and between 71G/N1 and 80U/O2, and these bases are excluded from analysis of reactivity and cooperativity.

3.2.2 Parametrization of 1M7

The 1M7 molecule is parametrized according to the general Amber force field (GAFF) [75, 76] for organic molecules using the Antechamber and parmchk tools implemented in Ambertools [77].

Preparation of 1M7 probes is performed through the LigPrep module of the Maestro interface in the Schrödinger suite [78]. The Gaussian 16 package is then employed for geometrical optimization and calculation of the electrostatic potential of the probe, using the B3LYP hybrid functional method with $6-31G*$ basis set. Partial charges are then calculated using the RESP method [79] as implemented in Antechamber. The resulting charges, that sum up to 0 as 1M7 is overall neutral, are reported in Fig.3.3. The resulting Amber potential is then converted to the GROMACS implementation [80], using acpype [81]. Preparation of 1M7 probes is performed through the LigPrep module of the Maestro interface in the Schrödinger suite [78].



(a) (b)

Figure 3.4: Two examples of initial configuration for the simulations of the tetraloop with (a) $N = 5$ and (b) $N = 16$ probes. The yellow sphere represents the surface around the tetraloop where probes are initially placed.

3.2.3 Simulation protocol

In order to sample a range of different concentrations of 1M7, $N_{max} = 19$ independent simulations are set up, each featuring a fixed number of probes, from $N = 1$ to $N = N_{max}$. For each of them, the starting configuration is built up as follows. The center of mass of the tetraloop is taken as origin of the reference frame. A rhombic dodecahedron simulation box is then placed at a distance of 3 nm from the tetraloop. It is important to place the box at this step, before inserting the 1M7 probes, in order to preserve the volume across the simulations with different N. In order to set a starting configuration where the N probes have an as similar interaction as possible with the tetraloop, they are each placed at random points at equal distance from the tetraloop and with random orientation. In particular, the first probe is placed at a random point on the surface of a sphere, centered at the tetraloop and with radius equal to the radius of gyration of the tetraloop plus 2 nm. The probe is then rotated of a random angle around its center of mass. A check on the distances between every atom pair is made in order to avoid clashes: if one of the atoms of the inserted probe is at a distance lower than 5 Å from any other atom, the insertion is rejected and another point and orientation are generated. For each of the remaining $N - 1$ probes the insertion procedure is repeated. Examples of the resulting configuration are represented in Fig. 3.4 for $N = 5$ and $N = 16$.

The resulting complexes are solvated using the OPC water model [82] and sodium counterions are added to neutralize the system. For each complex the potential energy is minimized in order to relax the structures, remove possible clashes and incorrect geometries, through 50000 steps of steepest descent algorithm. The minimization is followed by NVT equilibration of 1 ns up to a temperature $T = 300$ K, and NPT equilibration at the same temperature, pressure $P = 1$ bar for another 1 ns using a Parrinello-Rahman barostat [83]. A cutoff of 10 Å and the particle-mesh Ewald (PME) method [84] are used for computing short-range interactions and long-range interactions, respectively. Constant temperature is kept using the V-rescale thermostat [85].

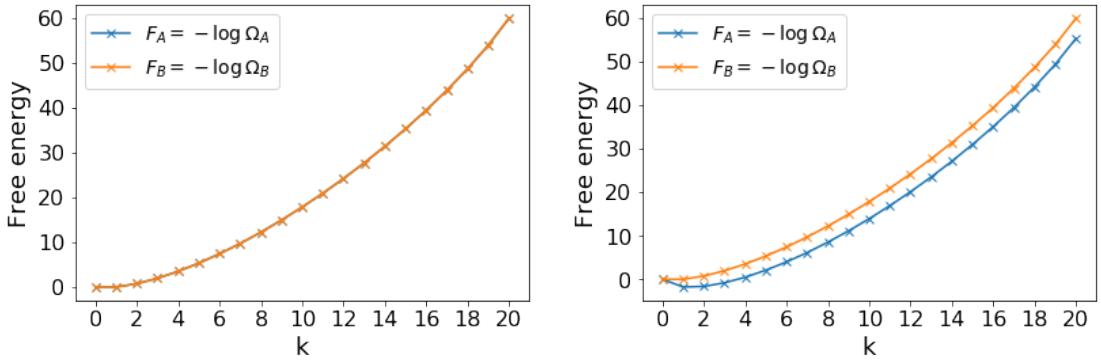


Figure 3.5: Free energy contributions of the two regions A and B of a lattice space populated with mutually exclusive particles, (a) in a purely entropic system with A as large as B and (b) in the same system but with a stabilizing site that is populated with probability 100 times larger than the others.

Equilibration is run with a time step of 2 fs with all hydrogen bonds constrained via the LINCS algorithm [86]. Production runs are then carried out in the NPT ensemble at $T = 300\,\mathrm{K}$ and $P = 1\,\mathrm{bar}$. Plain MD simulations are performed using version 2018.5 of the GROMACS software.

3.3 Results

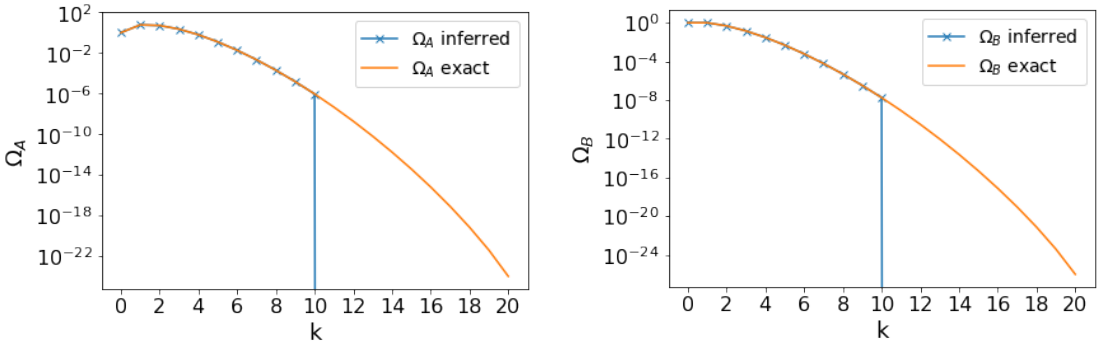
3.3.1 Toy model

The methodology presented in Section 3.1 is first tested on a toy lattice model. We consider a lattice space divided in two regions, A and B. Region A contains S_A sites and region B contains S_B sites. Sites are then populated with a number N_{max} of particles that interact with each other only through mutual exclusion, so that a site can not be occupied by more than one particle. Two scenarios are considered: a purely entropic systems, in which the free energy of the system depends only on the entropic contribution of the number of possible combinations of the N_{max} particles occupying the $S = S_A + S_B$ sites; and a system in which the presence of a stabilizing site in the lattice region A brings in an additional energetic contribute to the free energy of the system. In both cases, the partition functions Ω_A of region A and Ω_B of region B are computed. These functions are normalized as explained in Section 3.1, that is $\Omega_A(k=0) = \Omega_B(k=0) = 1$, and $\Omega_B(k=1) = 1$, so that the zero of free energy corresponds to the empty lattice and the free energy cost for insertion of the first particle in region B is set to zero. The normalization is accomplished by scaling each $\Omega_{A/B}(k)$ by a factor $f^k = \left(\frac{\Omega_B(1)}{\Omega_B(0)}\right)^k$.

In the purely entropic system, the two partitions functions are related to the number of different combinations in which particles can be distributed in the sites:

$$\Omega_{A/B}(k) = \binom{S_{A/B}}{k} = \frac{S_{A/B}!}{(S_{A/B}-k)!k!} \tag{3.12}$$

If the number of sites in A and B is equal, then populating a site in A has the same free energy cost of populating one in B. The free energy contributions of the two regions $F_{A/B} = -\log \Omega_{A/B}$ for numbers of sites $S_A = S_B = 20$ is shown in Fig.3.5a.



(a) (b)

Figure 3.6: Inferred and exact values of the partition functions (a) Ω_A of region A and (b) Ω_B of region B, with infinite sampling and number of particles $N_{max} = 10$ lower than the number of lattice sites, $S = 40$. Whereas for $k \leq N_{max}$ the inference is exact due to infinite sampling, but $\Omega_{A/B}(k) = 0 \; \forall k > N_{max}$.

To test the scenario in which regions A and B differ for the presence of a stabilizing site in A, the number of combinations in which particles can be distributed on the lattice sites, given that the probability of populating one of the sites in region A is 100 times larger than the other sites, which corresponds to a stabilization of $-RT \log 100$. As shown in Fig.3.5b, for each number of particles that is added to the lattice, the presence of the stabilizing site systematically contributes with a free energy gain. This contribution is particularly visible for $k = 1$, at which the interaction energy of the inserted particle with the stabilizing site is larger than the entropic contribution, thus lowering the free energy of region A with respect to the empty state.

The simulation and reweighting protocol presented in Section 3.1 has two major limitations, that can be highlighted using this toy lattice model. Since the model with a stabilizing site is more similar to the complex of the tetraloop and reagents that can bind to it, we focus on that one. One limitation is that the range of concentrations at which the reweighting can be performed is limited by the choice of the maximum number of reagents N_{max} to be used in the simulations. The other is that we can only sample a finite number of conformations through our simulations. In the toy model, the first limitation emerges when the number of particles with which the lattice is populated is lower than the number of sites S. For example, assume we can sample an infinite number of conformations so as to exclude for the moment finite sampling effects. With a choice of number of sites $S = 40$ and number of particles $N_{max} = 10$, at most one fourth of the sites can be occupied. Using Alg.1 we will have an exact estimate of $\Omega_{A/B}(k)$ for $k <= N_{max}$, but $\Omega_{A/B}(k) = 0$ for all the other values of $k > N_{max}$, as shown in Fig.3.6. For the same reason, when using Alg.2 to find the values of chemical potential μ that correspond to the desired concentrations in region B, we will obtain incorrect estimates of the relation between μ and N_B, as well as wrong estimates of the probability distribution of particles in region A, $P_A(k)$, at values of enforced concentration $\langle N_B \rangle$ too close to N_{max}. These effects are shown in Fig.3.7. In particular, we have that $N_{A/B}$ saturates to N_{max} for all the values of $\mu < \mu(N_{max})$ (Fig.3.7a), so that for enforced values of $\langle N_B \rangle$ too close to N_{max}, the probability distribution $P_A(N_A)$ cannot be correctly reconstructed.

In addition to this, finite sampling effects emerge due to the fact that a set of N_{max} simulations of finite lengths $L(N)$ can only sample a finite number of conformations of the systems, affecting the accuracy of all the estimated quantities. An example of the effects of finite sampling in estimating $\Omega_{A/B}$, $\mu(N_{A/B})$ and $P_A(k)$ at fixed $\langle N_B \rangle$ is shown in Fig.3.8, for $N_{max} = S$ and

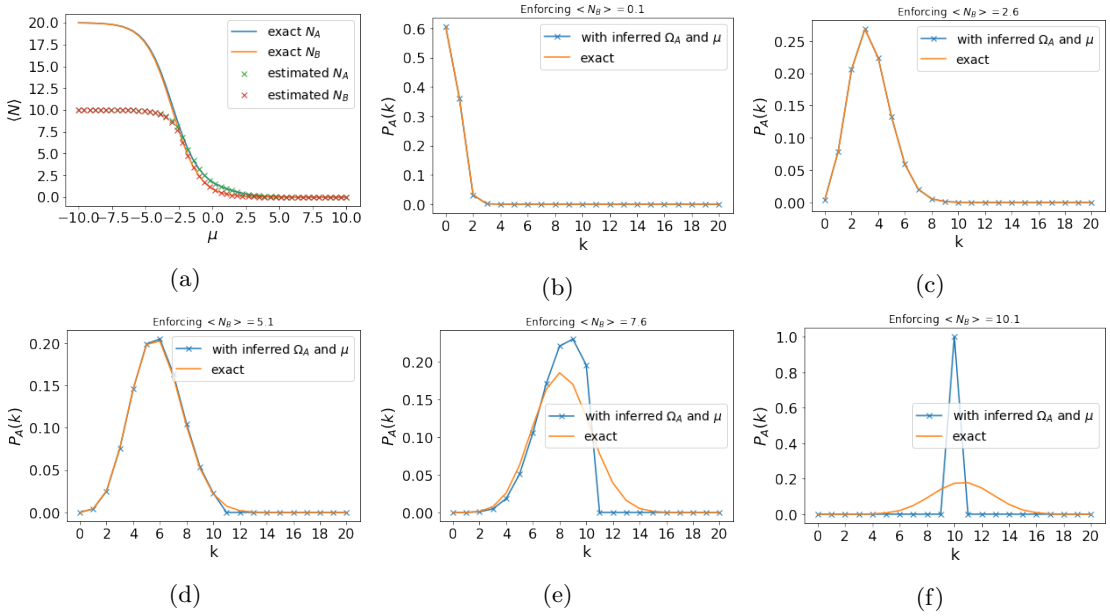


Figure 3.7: Limitations in the estimates of (a) chemical potential at desired concentrations $\mu(N_B)$ in region B, and (b-f) of the probability distribution of the number of particles in region A at different values of the enforced concentration in B, as tested on the toy lattice model with a stabilizing site in region A.

$L(N) = 100 \; \forall N$. In this case we do not have discontinuities due to the limit in the number of particles, but only limited accuracy due to finite sampling.

3.3.2 Molecular dynamics simulations

In order to estimate the reactivity profile and cooperativity matrix of the GAAA tetraloop at different concentrations of the 1M7 probe, we first divide the simulation space into two regions: the binding region A, where reagents are in proximity of the tetraloop and can form a relatively stable bound state (preliminary to the formation of the covalent bond), is defined as a sphere centered at the center of mass of the tetraloop and of radius r_A; what remains of the simulation space is the buffer region B, where the interactions between the 1M7 probes and the tetraloop are weaker and thus the formation of a bound state is not possible. The optimal choice of r_A is that at which the fluctuations in number of reagents per region are maximized. A reasonable choice, neglecting the effect of the presence of reagent-attracting sites in the tetraloop, is that at which the volumes of the two regions are equal, obtained with $r_A = 2.9\,\text{Å}$. A trajectory is collected for each tested number of reagents, $N \in [1, \dots, N_{max} = 19]$. Each trajectory is collected for a total length of 1 µs of dynamics, using a time-step $dt = 2\,\text{fs}$ and saving coordinates every 10 ps. Each trajectory thus contains $L(N) = 10^5$ frames.

Once the trajectories are collected, in order to perform a single loop of analysis and so reducing computational costs significantly, we compute the number of times $f\,(N, i, j, N_A, N_B, s, t)$ that in the simulation at fixed number of reagents N, the nucleotides i and j of the tetraloop are in one of the four possible pairwise binding state ($s = t = 0$ if they are both unbound, $s = t = 1$ if they are both bound to a reagent, $s = 1, t = 0$ if nucleotide i is bound and j is unbound and vice-versa for $s = 0, t = 1$), with a number N_A of reagents in the binding region A and N_B reagents in the

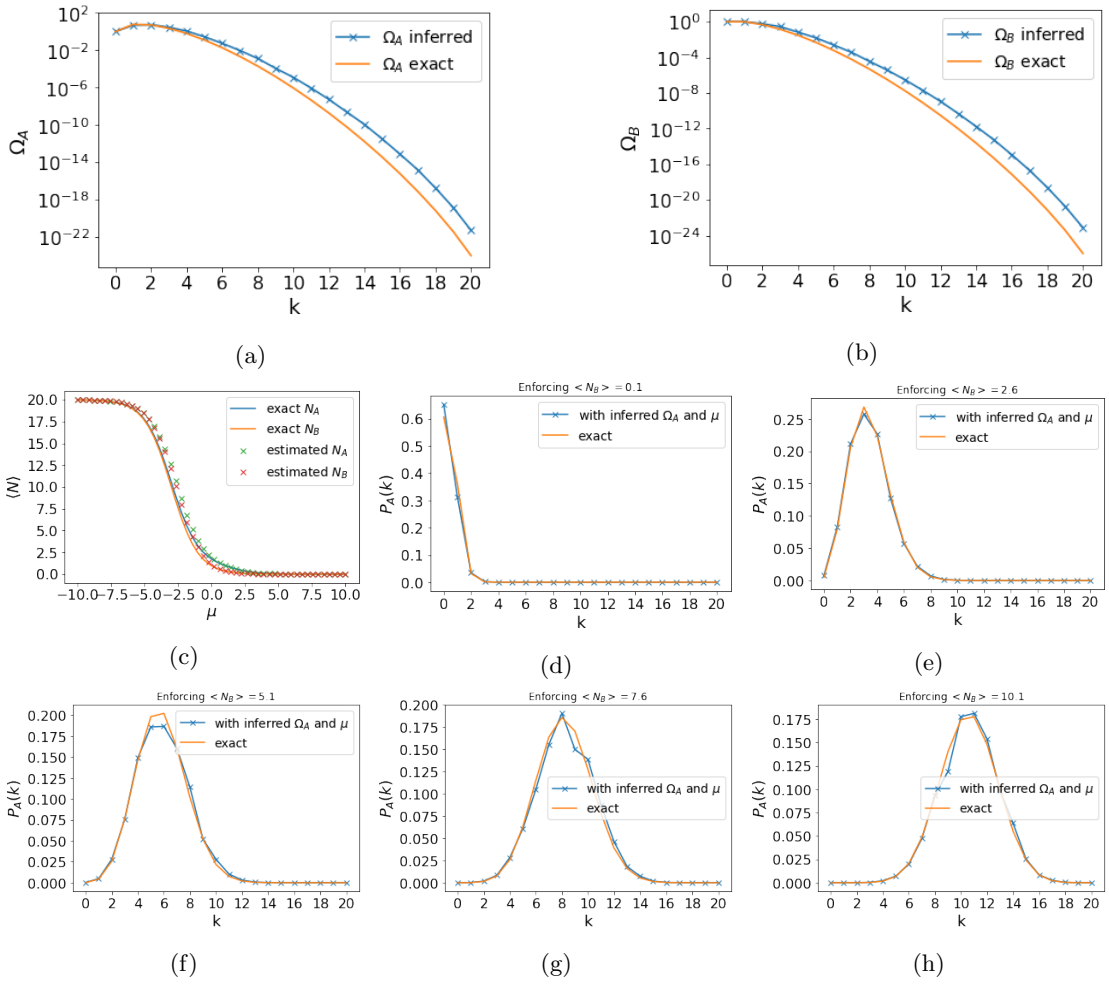


Figure 3.8: Finite sampling effects in the estimations of (a) $\Omega_A(k)$, (b) $\Omega_B(k)$, (c) the relation between chemical potential μ and number of particles in region A and B, and of (d-h) the probability distribution $P_A(k)$ of the number of particles in region A, as tested on the toy lattice model with a stabilizing site in region A.

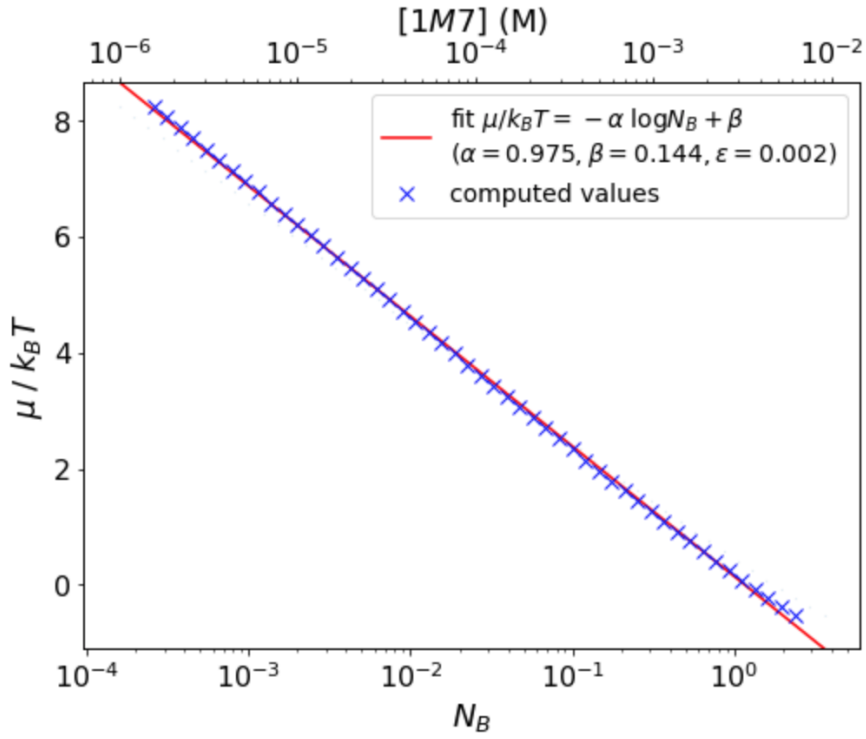


Figure 3.9: The chemical potential μ, in units of k_BT as function of the desired concentration of 1M7 (upper axis), or equally of the number of reagents in the buffer region N_B (lower axis). The result of a linear regression fitted on computed values is also shown, with the fit parameters slope α and intercept β, and standard error of estimates ε.

buffer region B. For our computations, we define binding between a nucleotide and a 1M7 probe to occur whenever the nucleotide is the nearest one to the probe, and the distance between the O2′ of the nucleotide and the C7 from the active carbonyl of 1M7 is less than a threshold that we set to $r_{th} = 4.0\,\text{Å}$, consistently with [69].

From the observed frequencies, we can compute all the significant quantities. Since we are interested in the grand canonical averages of the relevant quantities such as reactivity and cooperativity, we first need to find the correct ensemble weights by solving Eqs. 3.10. To this aim, we first compute the histograms

$$\begin{aligned} A_k &= \sum_{N=1}^{N_{max}} \sum_{N_B,s,t} f\left(N, i, j, N_A, N_B, s, t\right) \quad \text{for any } i \text{ and } j \\ B_k &= \sum_{N=1}^{N_{max}} \sum_{N_A,s,t} f\left(N, i, j, N_A, N_B, s, t\right) \quad \text{for any } i \text{ and } j \end{aligned} \tag{3.13}$$

and estimate Ω_A and Ω_B through Alg. 1. We then find, using Alg. 2, the values of μ that correspond to our desired concentrations. Typically, SHAPE experiments using 1M7 as probe are carried out at reagent concentrations ranging from 0.1 to 100 mM [87]. With our simulations, using from $N = 1$ to $N = 19$ reagents, we are able to span concentrations up to 10 mM. The corresponding values of μ are reported in Fig. 3.9. In this range, concentrations are small enough

for the sums in Eq. 3.11 to be dominated by the first non-zero term, that occurs for $k = 1$ (setting $k_B T = 1$ to simplify notation):

$$\langle N_B \rangle = \frac{\sum_k k\Omega_B(k)e^{-\mu k}}{\sum_k \Omega_B(k)e^{-\mu k}} \xrightarrow{\mu \gg N_B} \frac{\Omega_B(k=1)e^{-\mu}}{1 + \Omega_B(k=1)e^{-\mu}} \tag{3.14}$$

Since we have set $\Omega_B(k = 1) = \Omega_A(k = 1) = 1$, then in this limit $\langle N_B \rangle \sim e^{-\mu}$. We thus verify the functioning of Alg.2 by fitting the observed μ as a function of $-\log N_B$, in the limit of $\mu/k_B T \gg N_B$, with a linear regression model $\mu/k_B T = -\alpha \log N_B + \beta$, yielding slope $\alpha = 0.975$ and intercept $\beta = 0.144$, with a standard error estimate of $\varepsilon = 0.002$. The weights that are used to compute grand canonical ensemble averages are then defined as follows

$$w(N_A) = \frac{e^{-\mu N_A}\Omega_A(N_A)}{\sum_{N_A=1}^{N_{max}} e^{-\mu N_A}\Omega_A(N_A)} \tag{3.15}$$

The two-point binding frequency $p_{ij}(N_A, s, t)$, accumulated from each trajectory, integrated over all the occurring values of N_B is straightforwardly computed from f as
$p_{ij} \propto \sum_N \sum_{N_B} f\,(N, i, j, N_A, N_B, s, t)$ and then normalized as to sum to 1 over all the possible two-point binding states $\sum_{s,t=0}^{1} p_{ij}(N_A, s, t) = 1 \quad \forall N_A$. From its grand canonical average,

$$\langle p_{ij}(s,t) \rangle_{GC} = \sum_{N_A} w(N_A) p_{ij}(N_A, s, t) \tag{3.16}$$

reactivity profiles and cooperativity matrixes can thus be computed at any desired concentration of reagent in the range allowed by the choice of simulated number of reagents N, by accordingly tuning N_B and recalculating $w(N_A)$.

3.3.3 Reactivity profile

For each nucleotide site, reactivity can be estimated as the frequency with which it is observed in a bound state with any one of the reagents. Reactivity profiles can thus be computed from $\langle p_{ij}(s,t) \rangle_{GC}$ as

$$R_i = \sum_t \langle p_{ij}(s=1,t) \rangle_{GC} \quad \text{for any } j \tag{3.17}$$

at the desired concentration. Reactivity profiles computed from our simulations, for a set of concentrations spanning from 1 µM to 10 mM are reported in Fig. 3.10. The nucleotides of the helix-closing base-pair, G71 and U80 are excluded from the analysis, as their relatively large reactivity is due to the fact that, being the terminal sites of the sub-sequence, they are easily accessed by reagents. In experiments where the whole sequence is probed, this base-pair does not close a helix, but it is stacked with other base-pairs of the helix, so that the accessible fraction of the surrounding volume is smaller. In the experiment the observed reactivities for G71 and U80 are thus expected to be more similar to those of the neighbor nucleotides. The effect of enhanced accessibility of the terminal nucleotides propagates to the other base-paired nucleotides, but with a rapid decay of the reactivities. At the loop sites, reactivities grow again to reach the maximum at the A76 site, that is expected to be highly exposed to the reagent.

A preliminary indication of cooperative effects can be obtained from the behavior of reactivities as function of concentration, reported in Fig. 3.11. The relation between the reactivity of a

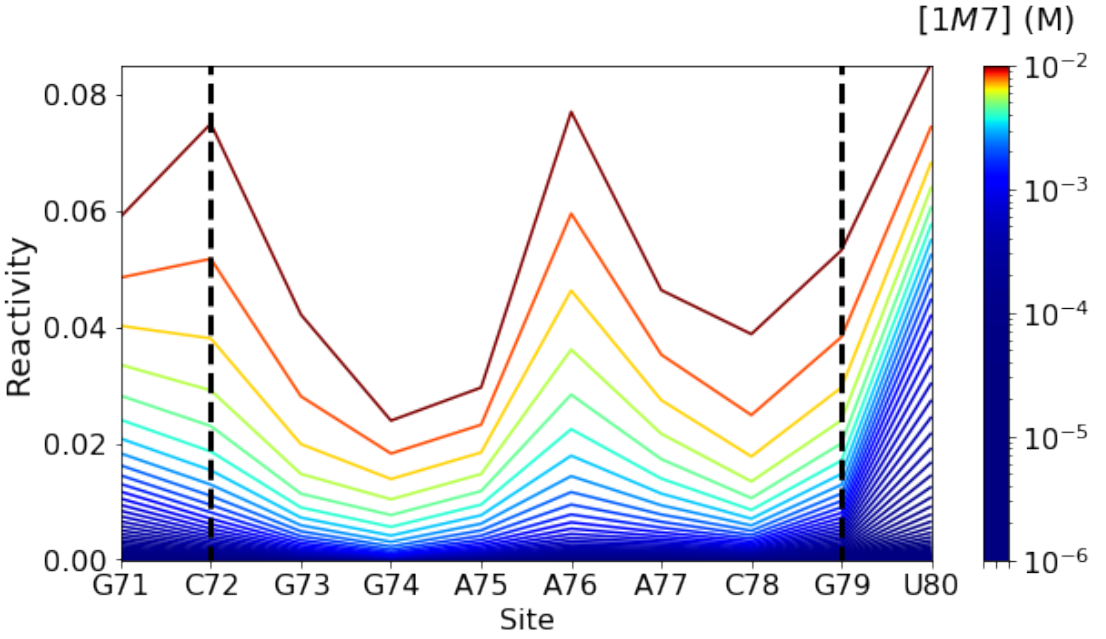


Figure 3.10: Computed reactivity profiles of the sequence under study, gcgGAAAcgu, at different concentrations of 1M7 reagent. The closing base-pair of the molecule G71-U80, is excluded from the analysis, as highlighted by the two vertical dashed lines.

nucleotide and the concentration of reagent can be decomposed in the sum of terms representing (a) the number of available reagents, (b) a cooperativity term describing the effect of a nucleotide binding to a reagent on the binding probability of another nucleotide to bind another reagent, and (c) higher order terms involving more than two nucleotides. The first term is proportional to the concentration of reagent, the second to the square of the concentration, and so on and so forth. Thus, as expected, at low concentrations the ratio R/C between reactivity and concentration saturates to a constant, as the only significant term is the first, that scales linearly with concentration $R_i \propto C$. As the reagent concentration increases, over a certain threshold cooperative (and anti-cooperative) effects are expected, as terms scaling with a higher power of the concentration start to be significant. It can be seen from Fig.3.11 that the R/C ratio of a nucleotide at large enough concentrations (around $C = 10^{-3}M$) starts to increase either super-linearly with concentration C if it predominantly cooperates with other nucleotides, or sub-linearly in the case of predominantly anti-cooperative behavior.

We thus focus on that range of concentrations, that is also a typical experimental condition of SHAPE probing using 1M7. For each nucleotide included in the analysis (C72 to G79), we compute reactivities at concentrations (in molar units) $C \in [1.0\times10^{-3}, 1.3\times10^{-3}, 1.7\times10^{-3}, 2.1\times 10^{-3}, 2.8\times10^{-3}, 3.6\times10^{-3}, 4.6\times10^{-3}, 6.0\times10^{-3}, 7.7\times10^{-3}, 1.0\times10^{-2}]$.

Errors are estimated through a bootstrap procedure in the following way: the dataset containing each of the N_{max} trajectories is resampled with replacement for $N_{iter} = 10^4$ iterations, and from each sample a reactivity profile for each concentration is computed (reactivity distributions obtained in this way, for an intermediate value of concentration $C = 4.6\,\text{mM}$, are reported in Fig. 3.12); for each value of the reagent concentration, the error on reactivity is computed as the standard deviation over the boostrap iterations $\sigma = \frac{1}{\sqrt{N_{iter}}}\sum_{b=1}^{N_{iter}}(R_b(C) - \langle R_b(C)\rangle)^2$, where

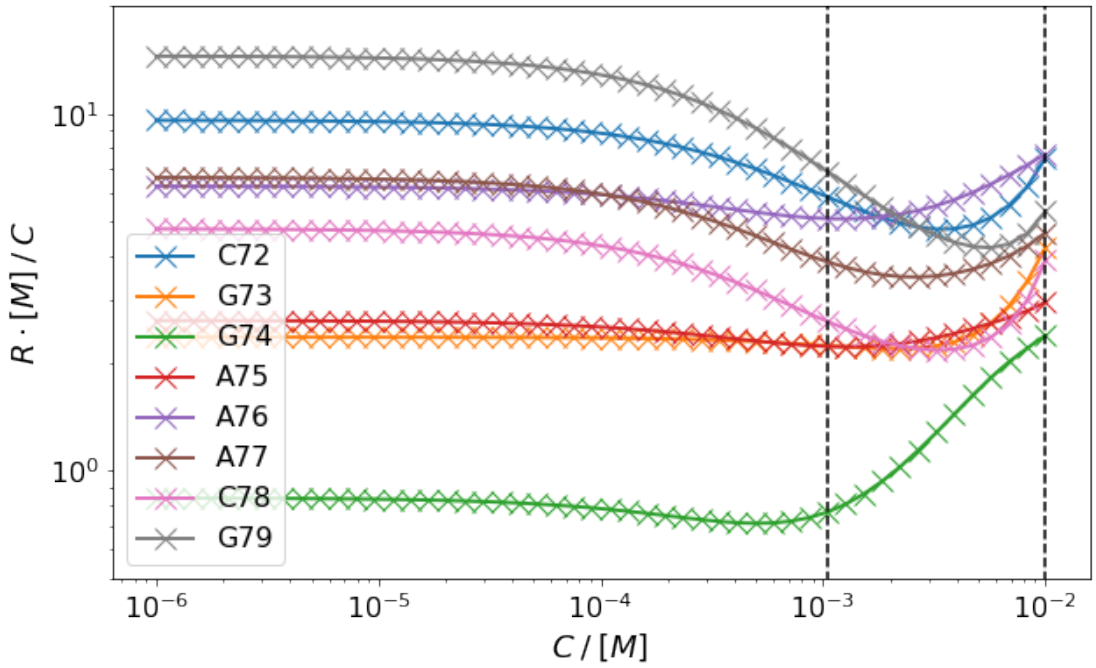


Figure 3.11: Reactivity-concentration ratio R/C as a function of the reagent concentration in molar units, for each nucleotide under analysis (C72 to G79). At low concentrations, cooperativity does not take place as the number of reagents is too low. In this regime, R is proportional to C and thus the R/C ratio is a constant. As concentration increases, some reactivities increase super-linearly or sub-linearly, depending on if the corresponding nucleotide is cooperative or anti-cooperative.

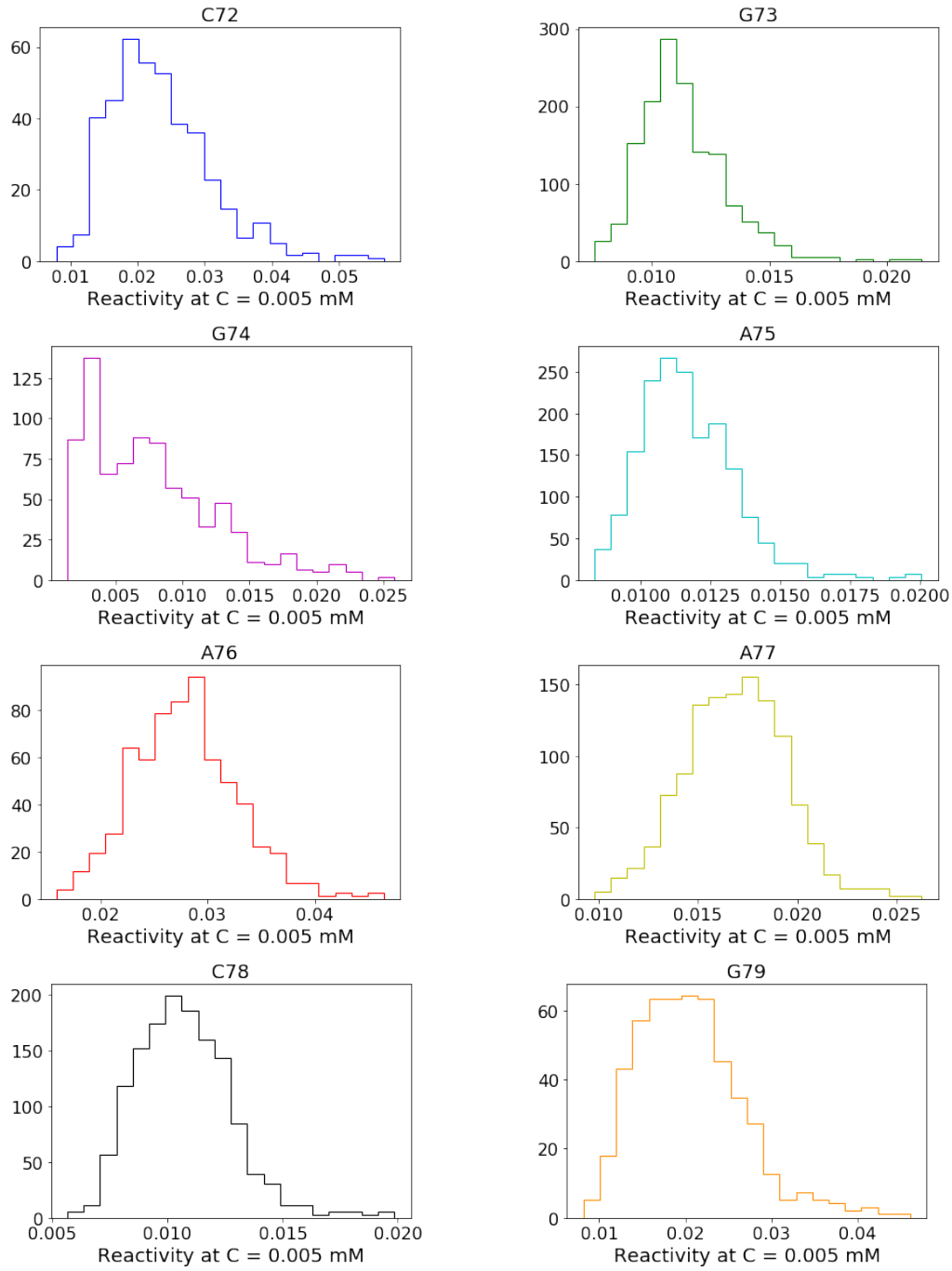


Figure 3.12: Histograms of computed reactivities at molar concentrations $C = 4.6\,\mathrm{mM}$ for each nucleotide. Values are collected from the iterations of bootstrap resampling. The bootstrap procedure is repeated for each value of the tested concentrations, so that statistical significance tests can be carried out.

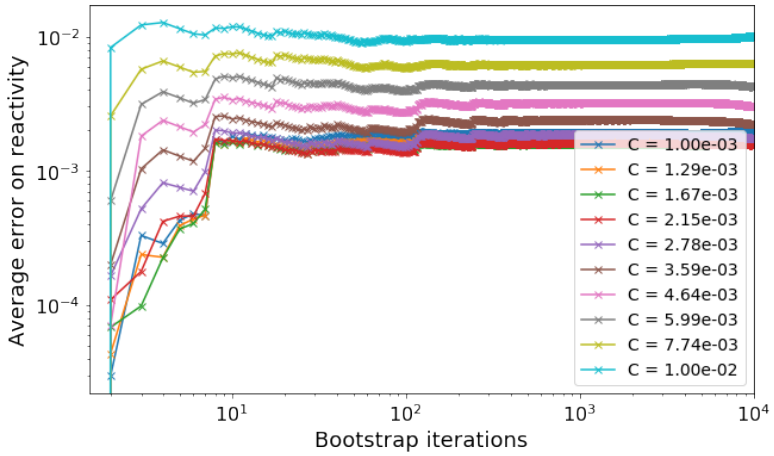


Figure 3.13: Estimated errors on reactivities at each tested reagent concentration, as averaged on all the nucleotides. The bootstrap procedure used to estimate errors consists in iteratively resampling the original set of trajectories.

$R = \gamma \cdot C^{\delta}$		
Nucleotide	$\delta \pm \Delta\delta$	$\gamma \pm \Delta\gamma$
C72	1.047 ± 0.003	2.95 ± 0.01
G73	1.208 ± 0.002	3.13 ± 0.01
G74	1.478 ± 0.002	3.90 ± 0.01
A75	1.121 ± 0.001	2.585 ± 0.004
A76	1.184 ± 0.001	3.803 ± 0.005
A77	1.089 ± 0.001	2.816 ± 0.005
C78	1.089 ± 0.002	2.45 ± 0.01
G79	0.834 ± 0.003	1.67 ± 0.02

Table 3.1: Parameters of a power law of reactivity R as a function of concentration C, obtained by a least-squares linear fit of their logarithms, for each nucleotide. Powers $\delta > 1$ indicate cooperative behavior, while $\delta < 1$ indicate anti-cooperativity.

$R_b(C)$ is the reactivity profile at concentration C obtained from the b-th sample. The choice of $N_{iter} = 10^4$ iterations proves to be enough since, as shown in Fig.3.13, the error estimate, averaged on all the nucleotides, already converges between 10 and 100 iterations. The computed reactivities for each nucleotide under analysis are shown in Fig.3.14.

A preliminary search for cooperative effects is carried out by fitting the following linear model

$$\log(R \cdot [M]/C) = a \log(C/[M]) + b \tag{3.18}$$

on the profiles reported in Fig.3.14 using the least squares method. Error bars are estimated as standard deviations for the parameters a and b. The fitted linear models of Eq.3.18 are transformed to power law models

$$R = \gamma \cdot C^{\delta} \tag{3.19}$$

where $\gamma = e^b$ and $\delta = a + 1$, and errors on γ and δ are propagated from those on a and b. The parameters resulting from the fit are reported in Table 3.1. Almost all the nucleotides in the loop show a significantly super-linear trend of $R(C)$, except for A77 for which we observe an

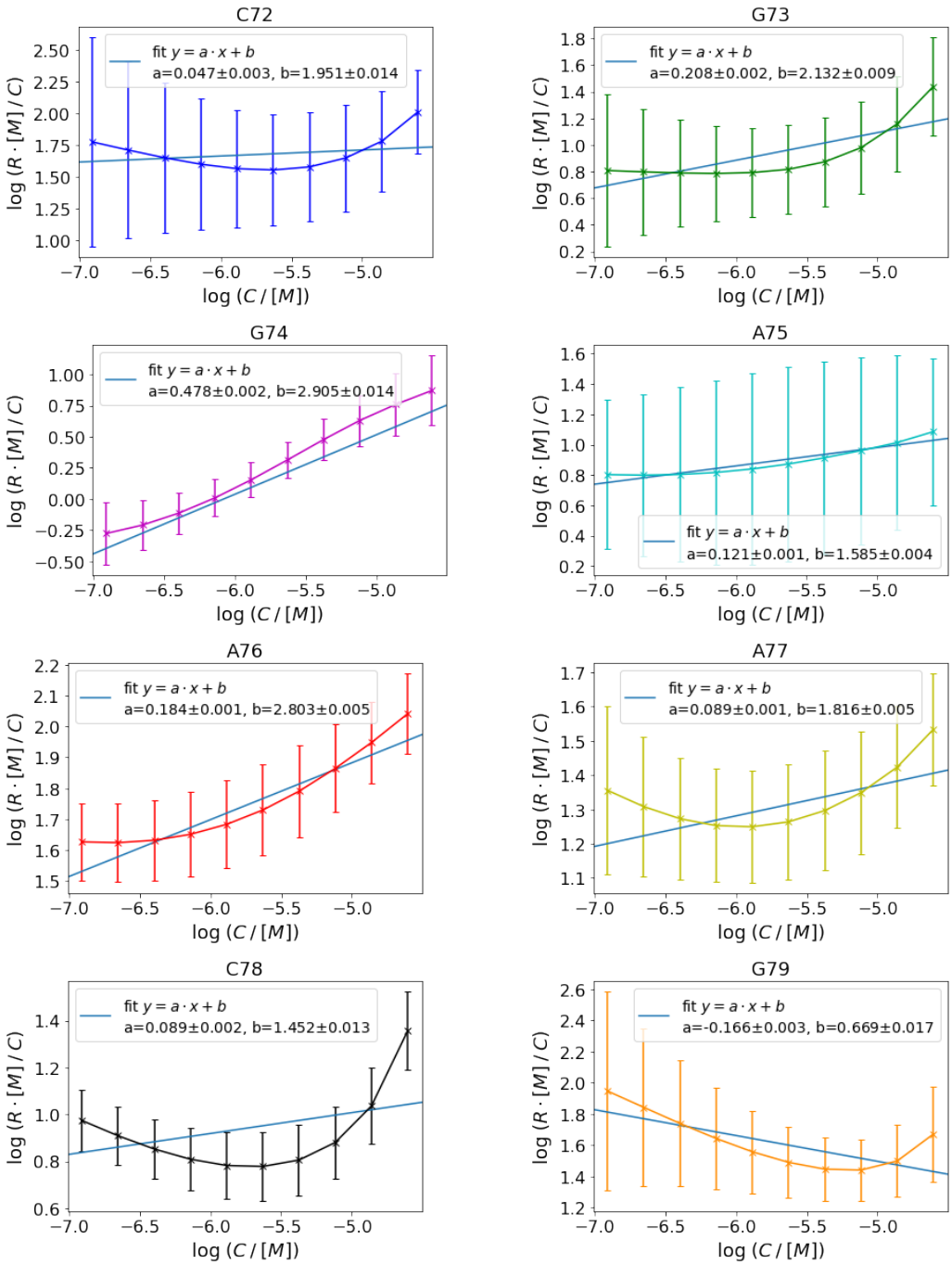


Figure 3.14: Computed reactivities at molar concentrations $C/[M] \in [1\times10^{-4}, 1.7\times10^{-4}, 2.7\times10^{-4}, 4.6\times10^{-4}, 7.7\times10^{-4}, 1.3\times10^{-3}, 2.1\times10^{-3}, 3.6\times10^{-3}, 6.0\times10^{-3}, 1.0\times10^{-2}]$ for each nucleotide. Errors are computed as standard deviations on the bootstrap resampling iterations of a set of trajectories obtained from splitting in two the originally collected ones.

approximately linear trend. In particular, the reactivity of G74 increases with the highest power exponent, $\delta = 1.478 \pm 0.003$, followed by A76 and A75 ($\delta = 1.121 \pm 0.001$ and $\delta = 1.184 \pm 0.001$). The $R(C)$ curves for base-paired are in general subject to higher errors on parameter estimate. Fitted power exponents of these curves tend to decrease to 1 from the loop-closing base-pair G73-C78 to the next C72-G79, with G79 showing an anti-cooperative behavior, with δ significantly lower than 1.

3.3.4 Cooperativity

To further investigate these different behaviors, we compute the cooperativity for each pair of nucleotide sites, in terms of the two-site free energy coupling matrix $\Delta\Delta G$ [88], under the hypothesis that, at least for small enough concentrations, terms of higher order than two-site cooperativity are negligible. Cooperativity between any pair of sites i, j of the tetraloop can be defined in terms of coupling free energy using the computed $p_{ij}(s,t)$ from Eq.3.16, as

$$\Delta\Delta G_{ij} = -RT \log \frac{p_{ij}(1,1)p_{ij}(0,0)}{p_{ij}(1,0)p_{ij}(0,1)} \tag{3.20}$$

With this definition, $\Delta\Delta G_{ij} < 0$ means that i and j are cooperative, as it occurs when the two sites are more frequently in the same state (both bound or both unbound) than in different states $p_{ij}(1,1)p_{ij}(0,0) > p_{ij}(1,0)p_{ij}(0,1)$. Vice-versa, $\Delta\Delta G_{ij} > 0$ indicates an anti-cooperative relationship between the two sites, meaning than the state of binding of one decreases the probability that the other one gets bound.

The $\Delta\Delta G$ matrix computed from our simulations and averaged through the grand canonical reweighting procedure described above is shown in Fig.3.15 for six representative values of reagent concentration. In general, for nucleotides identified as cooperative through analysis of reactivity as function of reagent concentration $R(C)$ (in descending order of power law exponent G74, A76, G73, C78 and A75) the most negative values of $\Delta\Delta G$ are found. For A77, whose reactivity increases approximately linearly with concentration, intermediate values of $\Delta\Delta G$ are observed. However, since these are the results of a sampling of finite size, it is fundamental to test them for statistical significance.

Multiple-hypothesis testing

Since we want to test the statistical significance of a number of entries of $\Delta\Delta G_{ij}$ simultaneously, the appropriate framework is that of multi-hypothesis testing. The cooperativity matrix $\Delta\Delta G_{ij}$ has dimension 8×8. The diagonal elements of the cooperativity matrix, that are equal to zero by construction, are excluded from the testing as the cooperativity of a nucleotide with itself has no physical sense. The matrix is also symmetric, as $\Delta\Delta G_{ij} = \Delta\Delta G_{ji}$. Thus, the number of entries that we want to test simultaneously is 28. Along the 10^4 bootstrap iterations, in general any ij entry may assume both positive and negative values, so we consider the whole distribution of its observed values, and compute the p-value relative to the hypothesis that the entry is negative or positive. If we fix a significance level of $\alpha = 0.01$ for each of the 28 hypotheses, the probability that *by chance* we obtain a significant result for at least one $\Delta\Delta G_{ij}$ is $1 - (1-\alpha)^{28} = 0.245$. For a number of 28 multiple hypotheses, this rate of false discoveries is not negligible.

As a statistical significance test to keep the false discovery rate at level α, we rely on the Benjamini-Hochberg [89] controlling procedure. This procedure consists in considering our hypotheses, sorted in ascending order of p-value (see Fig.3.16). We then retain as significant at level α, all the ordered hypotheses up to the k-th one, where k is the largest value at which the corresponding p-value is such that $p_{(k)} \leq \frac{k}{28}\alpha$. The results of this test on the most probable

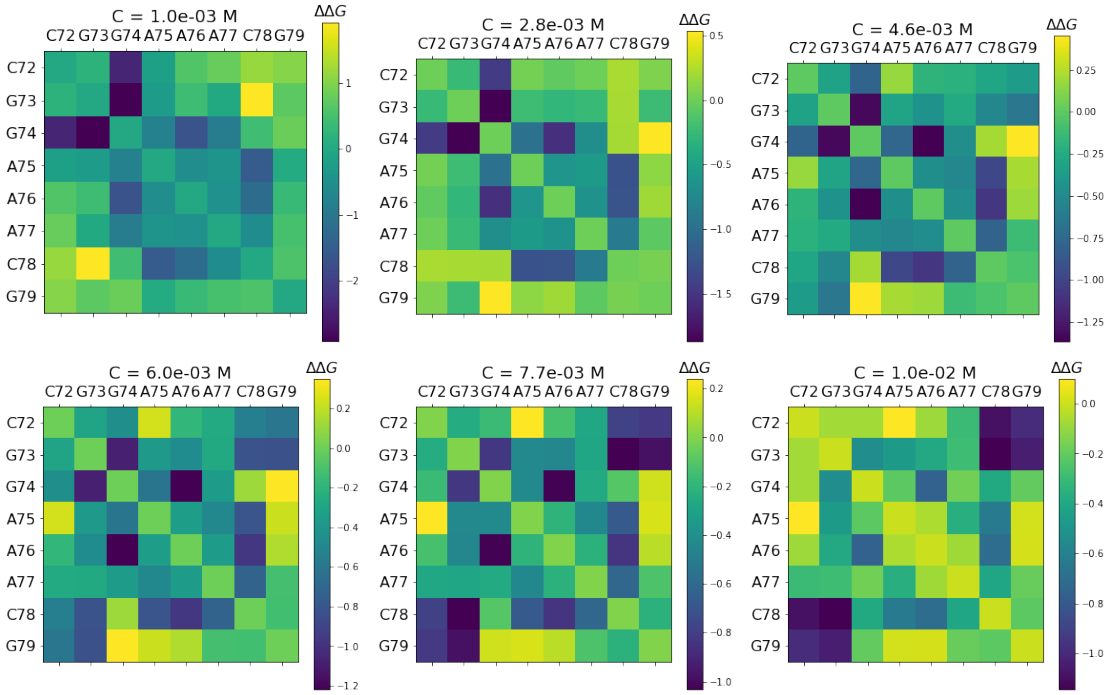


Figure 3.15: Cooperativity matrix $\Delta\Delta G$ for the nucleotides under analysis (C72 to G79), computed from the set of $N_{max} = 19$ simulations through grand canonical reweighting. Cooperativity at six representative concentrations of reagents are shown.

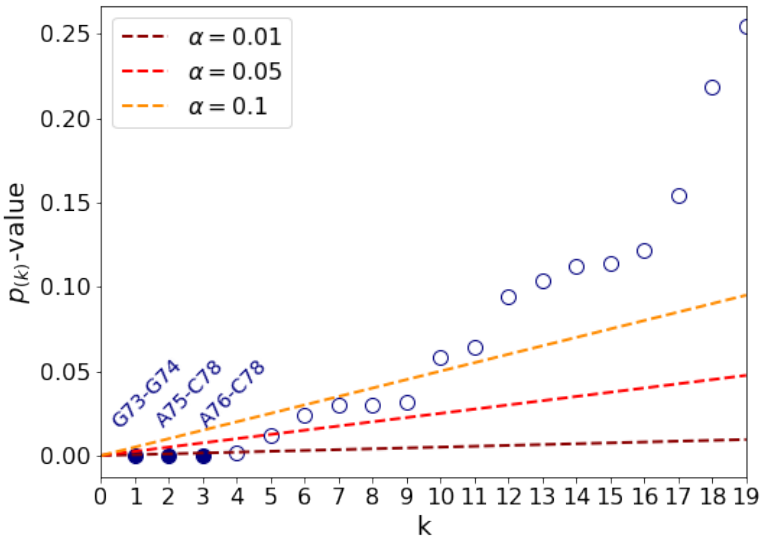


Figure 3.16: Results of the Benjamini-Hochberg multiple-hypothesis test to control the false discovery rate. A different number of hypotheses that a pair of nucleotides is cooperative pass the test depending on the chosen significance level. At reagent concentration $C = 4.6\,\text{mM}$, only the observed cooperativity for G73-G74, A75-C78 and A76-C78 are simultaneously significant at level $\alpha = 0.01$.

Pair	Concentrations [mM]									
	1.0	1.3	1.7	2.1	2.8	3.6	4.6	6.0	7.7	10
A76-C78	○	○	○	○	●	●	●	●	●	●
G73-G74	○	○	○	○	○	○	●	●	●	○
A75-C78	○	○	○	○	○	○	●	●	●	○
A77-C78	○	○	○	○	○	○	○	●	●	○
G73-C78	○	○	○	○	○	○	○	○	●	●
G73-G79	○	○	○	○	○	○	○	○	●	●
G74-A76	○	○	○	○	○	○	○	○	●	●
C72-G79	○	○	○	○	○	○	○	○	○	●

Table 3.2: Hypotheses of cooperativity for pairs of nucleotides passing (●) or not passing (○) the Benjamini-Hochberg multiple-hypothesis test, at the different tested concentrations.

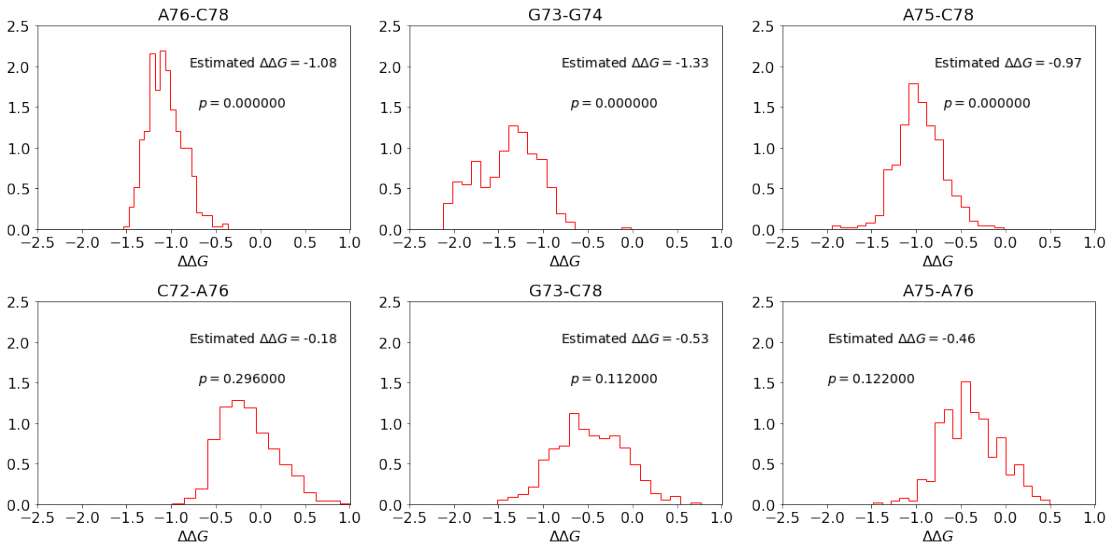
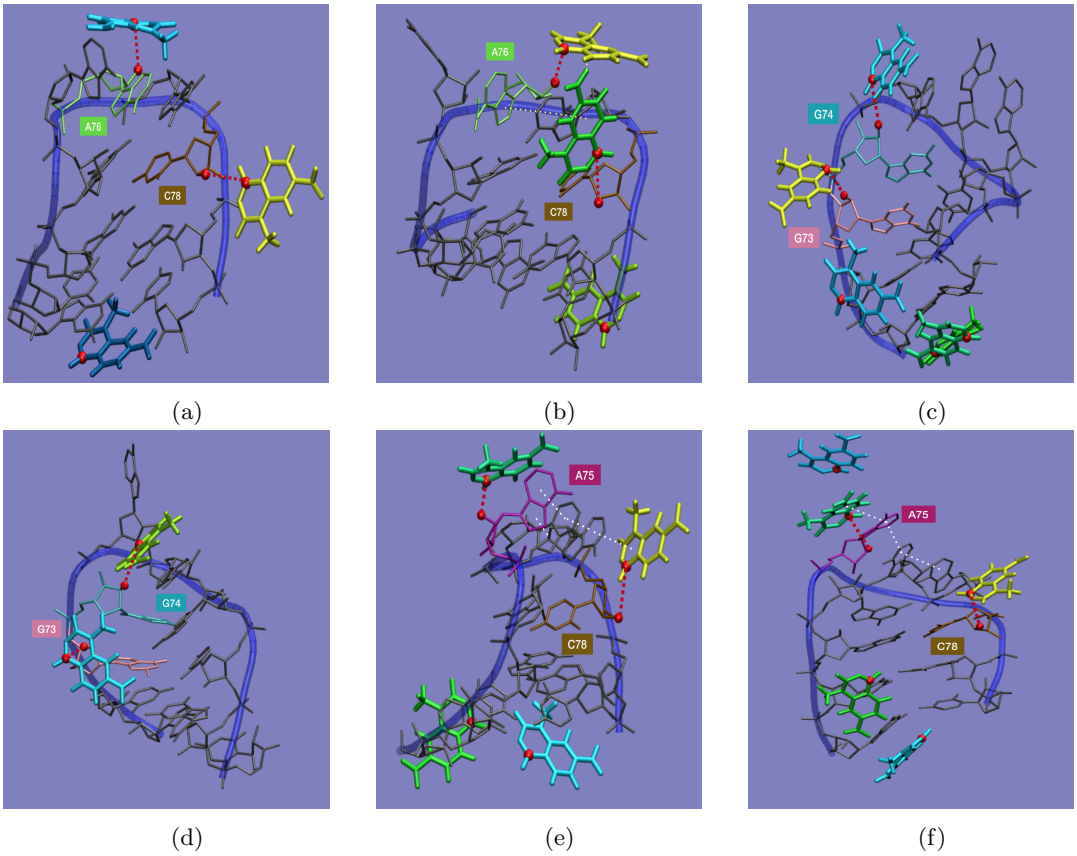


Figure 3.17: Histograms of the values of $\Delta\Delta G$ collected in the 10^{-4} bootstrap resampling iterations, for (a-c) three cases significant under the multiple-hypotheses test with $\alpha = 0.01$, at reagent concentration $C = 4.6\,\mathrm{mM}$, and (d-f) three cases not significant under the same test.



(a) (b) (c)

(d) (e) (f)

Figure 3.18: Qualitative selection of recurrent conformations of the tetraloop, when cooperative nucleotide pairs (a-b) A76-C78, (c-d) G73-G74 and (e-f) A75-A78 are both bind to a 1M7 reagent.

cooperativity hypotheses ($\Delta\Delta G < 0$), for an intermediate value of concentration $C = 4.6\,\text{mM}$, are reported in Fig.3.16. Only for pairs G73-G74, A75-C78 and A76-C78 the hypothesis of cooperativity is significant at level $\alpha = 0.01$. The corresponding distributions of cooperativity values on the bootstrap iterations are reported in Fig.3.17, along with three examples of pairs for which the extracted value of $\Delta\Delta G$ is considered as not significant. Pairs that show significant cooperativity at different values of concentration are reported in Table3.2. The lowest concentration, among the ones reported in Fig.3.15, at which we have at least one significantly negative entry of $\Delta\Delta G_{ij}$ under this test is $C = 2.8\,\text{mM}$, and the pair of sites is A76-C78. At $C = 4.6\,\text{mM}$ we have G73-G74, A75-C78 and A76-C78. Increasing the reagent concentration, other significantly cooperative sites add to the list: A77-C78 at $C = 6.0\,\text{mM}$; G73-C78,G73-G79, and G74-A76 at $C = 7.74\,\text{mM}$. At the largest tested concentration, $C = 10\,\text{mM}$, the list of pairs reduces to C72-G79,G73-C78,G73-G79, G74-A76 and A76-C78. Significant anti-cooperativity is only observed at the lowest concentrations, $C = 1.0\,\text{mM}$, and $C = 1.3\,\text{mM}$, for $G73 - C78$, for which $\Delta\Delta G = 1.91$.

3.3.5 Visual analysis

In order to gain some insight in the structural conditions that might be correlated with the cooperative behavior of nucleotides toward reagent binding, we carry out a qualitative visual analysis of conformations. In particular, we focus on the cooperative pairs of nucleotides that we observe at an intermediate value among the tested concentrations, $C = 4.6\,\mathrm{mM}$. These pairs are G73-G74, A75-C78 and A76-C78. Thus we first extract, for each pair, the trajectory frames in which the two nucleotides are both bound to a reagent. Then, in order to visualize only those that significantly contribute to cooperativity at a fixed concentration C, we sample from this set of frames with the concentration dependent weights computed through Eq.3.15. The most recurrent motifs that we observe for each pair of nucleotides are reported in Fig.3.18. Without recurring to quantitative analysis, we notice first that 1M7 probes, being small and planar molecules, tend to occupy the space around nucleotides forming recurrent geometries. This is particularly visible in Fig.3.18a and b, where the two reagents involved in binding cooperative nucleotides form a T-shaped geometry. We speculate that if the tertiary structure containing the two nucleotides favors this geometry, then they might be cooperative in the binding dynamics. Another feature that we notice from this qualitatively analysis, is that stacking interactions, involving either nucleotides with each other or with probes, might stabilize some recurrent conformations. In Fig.3.18c for example, a recurrent motif is shown in which the probe binding to G74 forms stable stacking interactions with the other bases of the loop, possibly causing a deformation of the whole structure that leads to an increase in the accessible volume around G73. A different situation involving the same two nucleotides is represented in Fig3.18d, where instead the probe binds to G74 by intercalating between A75 and A76, with a rearrangement of the molecule that might make the environment of G73 more attractive towards other probes. Also for the pair A75 and C78, stacking of the 1M7 binding C78, with the bases in the loop, could contribute to the accessibility of A75, and vice-versa, as shown in Fig.3.18e and f, respectively.

3.4 Discussion

In summary, we introduced in this Chapter for the first time a methodology that we developed to compute grand canonical ensemble averages from simulations obtained with Molecular Dynamics simulations in the canonical ensemble. This consists in first dividing the simulation space into two regions: a buffer region used to control the concentration of reagent, and a binding region where the interesting dynamics take place. A set of trajectories, each using a different number of reagents is then collected. The frames of this set of trajectories, independently of the total number of reagents used in each one, are then reweighted depending on the number of reagents present in the binding region, and on the desired chemical potential. We developed algorithms to iteratively solve the equations that provide the partition functions of the two regions and the chemical potential as a function of the desired concentration of reagents in the buffer. In general, this method can be used to perform Molecular Dynamics simulations of systems aimed at extrapolating the dependence of interesting quantities on the concentration of a chosen species.

We have shown the major limitations of this methodology, due to the choice of maximum number of particles used in the simulations, and to finite sampling effects, on a toy lattice model with a stabilizing site in the binding region. In particular, the range of desired concentrations must be compatible with the number of copies simulated. The first limit can be dealt with by running additional simulations using a larger number of particles, limited to the available computational power and time. Finite sampling effects can be reduced by running longer simulations.

We applied this original methodology to an investigation of cooperative effects in Selective Hydroxyl Acylation analysed by Primer Extension (SHAPE), one of the most efficient chemical

mapping protocols. To this aim, we simulated the dynamics of a GAAA tetraloop, an ubiquitous and well-characterized RNA motif, binding with 1-methyl-7-nitroisatoic anhydride (1M7), a small and fast-reacting chemical probe. We extracted the sequence and initial structure of the tetraloop from the SAM-I riboswitch crystallographic structure annotated in the PDB 2GIS entry. The 1M7 probe was instead parametrized here for the first time. A set of $N_{max} = 19$ simulations, each using a different number of reagents $N \in [1, \ldots, N_{max}]$, was run in the canonical ensemble for a total length of 1 µs each. Frames were reweighted with the developed grand canonical ensemble reweighting procedure, in order to obtain reactivity profiles and cooperativity matrices of the tetraloop at different values of reagent concentrations. The results are in agreement with expectations for low concentrations. At larger concentrations, we analysed the behavior of reactivities as function of reagent concentration, identifying candidate cooperative, non-cooperative and anti-cooperative nucleotides. Importantly, the behavior of the reactivity of each nucleotide as a function of reagent concentration can be directly compared with experimental measurements. We focused on computing cooperativity in terms of the corresponding free energy coupling term, $\Delta\Delta G$ from our simulations. The pairs of nucleotides that we identified as cooperative, non-cooperative and anti-cooperative are consistent with computed reactivity profiles. These pairs were selected through a rigorous multi-hypothesis testing applied to a bootstrap resampling of the generated trajectories. In principle, these information could be used to analyse concentration-dependent SHAPE experiments, so as to obtain further information about structure and dynamics. However, before doing so, the results presented here should be validated experimentally, which is beyond the scope of this Thesis.

For the selected pairs at a fixed intermediate value of concentration, among the ones that we tested, we conducted a qualitative visual inspection of those conformations of the studied complex, that mostly recur when cooperative nucleotides are both in a binding state with a reagent. We speculate that the stacking interactions of the bases in the loop of the RNA, both with each other and with the probes, drive the local conformational rearrangement that determines cooperative behaviors. Quantitative approaches will be applied to gain a more accurate insight in the cooperative dynamics that emerge from our simulations.

Chapter 4

Conclusions and perspectives

In conclusion, in this Thesis we have presented computational investigations of structure probing experiments, especially the case of chemical mapping. This kind of experiments are of great importance as they give a measure of RNA structure at single-nucleotide resolution. However, the interpretation of nucleotide reactivities toward the reagents commonly used in chemical mapping is not straightforward, so that mapping this information into a structure prediction is non-trivial. In order to improve the methods for RNA secondary structure predictions that rely on these experimental data, we have developed two methodological approaches to investigate structure probing experiments.

First, we have developed a machine-learning procedure that can be used to train sets of models with benchmark structures, and select those that are both able to predict a larger ensemble population for the native structure, and to be transferable to unseen data. We applied this procedure to a set of 196 models that combine reactivity profiles from chemical probing experiments, co-evolutionary data extracted from analysis of homologous RNA sequences, and nearest-neighbor parameters for RNA folding obtained from optical melting experiments. In part of these models, we included terms that weight the contributions of the reactivities of its neighbors to the pairing state of a nucleotide. This resulted to be useful for reproducing training data, but significantly sensitive to overfitting. The model finally selected based on a criterium of performance and transferability, including reactivities of single nucleotides at a time, together with Direct-Coupling Analysis scores and a thermodynamic model free energy, is shown to outperform other existing methods for secondary structure prediction, both in terms of the population predicted for the native structure, and of similarity between the predicted minimum-free-energy and the native structure. Importantly we built up this procedure in such a way that all the results are easily reproducible and the models can be modified and retrained with new data, even of different kind, through direct access to scripts and datasets. They can be found at `https://github.com/bussilab/shape-dca-data`. The content of this work is also reported in the publication [52].

In perspective, the machine-learning procedure that we have developed can be modified from discriminative to generative. With a generative version of the training and validation procedure, input datasets with missing data could be exploited, for example by adding priors on the missing data to the ensemble free energy of the model. In this way, the trained models could be used not only for structure prediction, but also for the reconstruction of any missing part of the dataset. If applied to datasets where sequences are missing, generative models built on top of the one presented in this Thesis could even bring to advances in the inverse problem of structure prediction, *i.e.* sequence design. A preliminary investigation of this approach reported promising

results that were however too premature to be included in this Thesis.

From the investigation of chemical mapping experiments that we conducted, hypotheses emerged regarding the possible effect of cooperativity of nucleotides on the measurement of reactivity profiles. This possibility would have important implications in structure prediction protocols relying on these data. If the cooperativity in binding a pair of nucleotides affects the dynamics of the reactions with the chemical mapping probes, then their reactivities would reflect these effects in addition to the pairing state of the two nucleotides. Thus, we investigated the dynamics of Selective Hydroxyl Acylation analysed by Primer Extension (SHAPE), one of the most efficient chemical probing protocols. In particular, as test system we used a sub-sequence of the SAM-I riboswitch including a GAAA tetraloop, a structural motif that is ubiquitous and has been well-characterized. We simulated the dynamics of the tetraloop binding with 1M7, a probe that is commonly used in SHAPE experiments. In order to measure quantities from Molecular Dynamics that can be compared with experimental observations, we developed an original method to compute averages in the grand canonical ensemble at fixed chemical potential, thus controlling the reagent concentration. We thus reconstructed the reactivity profile of the tetraloop, and the cooperativity of its nucleotides, at different values of reagent concentration. We applied advanced statistical significance test to our sampling, and observed significant cooperativity for different pairs of nucleotides at different concentrations. From these observations, we speculated on the possible role of local conformations of the RNA in cooperative behaviors. The content of this work will be included in an article that is, at the moment, in preparation.

In perspective, since the method we developed allows for computing reactivities as function of concentration of 1M7, the results of our computations can be compared directly with experimental results. In the case of experiments confirming our hypotheses on cooperative effects, it would be desirable to include these effects in structure prediction protocols, possibly obtaining improved secondary structure predictions and even information on tertiary structure.

Bibliography

[1] Cech, T. R. (2000) The ribosome is a ribozyme. *Science,* **289**(5481), 878–879.

[2] Doudna, J. and Cech, T. (2002) The chemical repertoire of natural ribozymes. *Nature,* **418**(6894), 222–228.

[3] Morris, K. V. and Mattick, J. S. (2014) The rise of regulatory RNA. *Nat. Rev. Genet.,* **15**(6), 423.

[4] Wan, Y., Kertesz, M., Spitale, R. C., Segal, E., and Chang, H. Y. (2011) Understanding the transcriptome through RNA structure. *Nat. Rev. Genet.,* **12**(9), 641.

[5] Cooper, T. A., Wan, L., and Dreyfuss, G. (2009) RNA and disease. *Cell,* **136**(4), 777–793.

[6] Zhang, J. and Ferré-D'Amaré, A. R. (2014) New molecular engineering approaches for crystallographic studies of large RNAs. *Current Opinion in Structural Biology,* **26**, 9 – 15.

[7] Zhang, H. and Keane, S. C. (2019) Advances that facilitate the study of large RNA structure and dynamics by nuclear magnetic resonance spectroscopy. *WIREs RNA,* **10**(5), e1541.

[8] Russell, R., Zhuang, X., Babcock, H. P., Millett, I. S., Doniach, S., Chu, S., and Herschlag, D. (2002) Exploring the folding landscape of a structured RNA. *Proceedings of the National Academy of Sciences,* **99**(1), 155–160.

[9] Kubota, M., Tran, C., and Spitale, R. C. (2015) Progress and challenges for chemical probing of RNA structure inside living cells. *Nature Chemical Biology,* **11**(12), 933–941.

[10] Merino, E. J., Wilkinson, K. A., Coughlan, J. L., and Weeks, K. M. (2005) RNA structure analysis at single nucleotide resolution by selective 2′-hydroxyl acylation and primer extension (SHAPE). *Journal of the American Statistical Association,* **127**(12), 4223–4231.

[11] Deigan, K. E., Li, T. W., Mathews, D. H., and Weeks, K. M. (2009) Accurate SHAPE-directed RNA structure determination. *Proceedings of the National Academy of Sciences,* **106**(1), 97–102.

[12] Spitale, R. C., Flynn, R. A., Zhang, Q. C., Crisalli, P., Lee, B., Jung, J.-W., Kuchelmeister, H. Y., Batista, P. J., Torre, E. A., Kool, E. T., et al. (2015) Structural imprints in vivo decode RNA regulatory mechanisms. *Nature,* **519**(7544), 486.

[13] Tinoco, I. and Bustamante, C. (1999) How RNA folds. *Journal of Molecular Biology,* **293**(2), 271 – 281.

[14] Leontis, N. B. and Westhof, e. (2001) Geometric nomenclature and classification of RNA base pairs. *RNA,* **7**(4), 499–512.

[15] Chastain, M. and Tinoco, I. (1991) Structural Elements in RNA. Vol. 41 of Progress in Nucleic Acid Research and Molecular Biology, pp. 131 – 177 Academic Press.

[16] Lu, X.-J., Bussemaker, H. J., and Olson, W. K. (2015) DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Research,* **43**(21), e142–e142.

[17] Chen, J.-L. and Greider, C. W. (2005) Functional analysis of the pseudoknot structure in human telomerase RNA. *Proceedings of the National Academy of Sciences,* **102**(23), 8080–8085.

[18] Brierley, I., Digard, P., and Inglis, S. C. (1989) Characterization of an efficient coronavirus ribosomal frameshifting signal: Requirement for an RNA pseudoknot. *Cell,* **57**(4), 537 – 547.

[19] Andronescu, M., Condon, A., Turner, D. H., and Mathews, D. H. The Determination of RNA Folding Nearest Neighbor Parameters pp. 45–70 Humana Press Totowa, NJ (2014).

[20] Nussinov, R. and Jacobson, A. B. (1980) Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proceedings of the National Academy of Sciences,* **77**(11), 6309–6313.

[21] Zuker, M. and Stiegler, P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research,* **9**(1), 133–148.

[22] McCaskill, J. S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers,* **29**(6-7), 1105–1119.

[23] Mathews, D. H. (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA,* **10**(8), 1178–1190.

[24] Zhang, H., Zhang, L., Mathews, D. H., and Huang, L. (2020) LinearPartition: linear-time approximation of RNA folding partition function and base-pairing probabilities. *Bioinformatics,* **36**(Supplement_1), i258–i267.

[25] Ding, Y. and Lawrence, C. E. (2003) A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research,* **31**(24), 7280–7301.

[26] Xia, T., SantaLucia, J., Burkard, M. E., Kierzek, R., Schroeder, S. J., Jiao, X., Cox, C., and Turner, D. H. (10, 1998) Thermodynamic Parameters for an Expanded Nearest-Neighbor Model for Formation of RNA Duplexes with Watson-Crick Base Pairs. *Biochemistry,* **37**(42), 14719–14735.

[27] Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M., and Turner, D. H. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences,* **101**(19), 7287–7292.

[28] Andronescu, M., Condon, A., Hoos, H. H., Mathews, D. H., and Murphy, K. P. (2010) Computational approaches for RNA energy parameter estimation. *RNA,* **16**, 2304–2318.

[29] Mathews, D. H., Sabina, J., Zuker, M., and Turner, D. H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology,* **288**(5), 911 – 940.

[30] Lucks, J. B., Mortimer, S. A., Trapnell, C., Luo, S., Aviran, S., Schroth, G. P., Pachter, L., Doudna, J. A., and Arkin, A. P. (2011) Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proceedings of the National Academy of Sciences,* **108**(27), 11063–11068.

[31] Lempereur, L., Nicoloso, M., Riehl, N., Ehresmann, C., Ehresmann, B., and Bachellerie, J. P. (1985) Conformation of yeast 18S rRNA. Direct chemical probing of the 5' domain in ribosomal subunits and in deproteinized RNA by reverse transcriptase mapping of dimethyl sulfate-accessible. *Nucleic Acids Research,* **13**(23), 8339–8357.

[32] Metz, D. H. and Brown, G. L. (1969) Investigation of nucleic acid secondary structure by means of chemical modification with a carbodiimide reagent. II. Reaction between N-cyclohexyl-N′-β-(4-methylmorpholinium)ethylcarbodiimide and transfer ribonucleic acid. *Biochemistry,* **8**(6), 2329–2342.

[33] Peattie, D. A. and Gilbert, W. (1980) Chemical probes for higher-order structure in RNA. *Proceedings of the National Academy of Sciences,* **77**(8), 4679–4682.

[34] Ziehler, W. A. and Engelke, D. R. (2000) Probing RNA Structure with Chemical Reagents and Enzymes. *Current Protocols in Nucleic Acid Chemistry,* **00**(1), 6.1.1–6.1.21.

[35] Peng, Y., Soper, T. J., and Woodson, S. A. RNase Footprinting of Protein Binding Sites on an mRNA Target of Small RNAs pp. 213–224 Humana Press Totowa, NJ (2012).

[36] Wan, Y., Qu, K., Ouyang, Z., and Chang, H. Y. (2013) Genome-wide mapping of RNA structure using nuclease digestion and high-throughput sequencing. *Nature Protocols,* **8**(5), 849–869.

[37] Ingle, S., Azad, R. N., Jain, S. S., and Tullius, T. D. (2014) Chemical probing of RNA with the hydroxyl radical at single-atom resolution. *Nucleic Acids Research,* **42**(20), 12758–12767.

[38] Spitale, R. C., Flynn, R. A., Zhang, Q. C., Crisalli, P., Lee, B., Jung, J.-W., Kuchelmeister, H. Y., Batista, P. J., Torre, E. A., Kool, E. T., et al. (2015) Structural imprints in vivo decode RNA regulatory mechanisms. *Nature,* **519**(7544), 486.

[39] Wilkinson, K. A., Merino, E. J., and Weeks, K. M. (2005) RNA SHAPE Chemistry Reveals Nonhierarchical Interactions Dominate Equilibrium Structural Transitions in tRNAAsp Transcripts. *Journal of the American Chemical Society,* **127**(13), 4659–4667.

[40] Bindewald, E., Wendeler, M., Legiewicz, M., Bona, M. K., Wang, Y., Pritt, M. J., Le Grice, S. F. J., and Shapiro, B. A. (2011) Correlating SHAPE signatures with three-dimensional RNA structures. *RNA,* **17**(9), 1688–1696.

[41] Hurst, T., Xu, X., Zhao, P., and Chen, S.-J. (2018) Quantitative Understanding of SHAPE Mechanism from RNA Structure and Dynamics Analysis. *The Journal of Physical Chemistry B,* **122**(18), 4771–4783.

[42] Lorenz, R., Luntzer, D., Hofacker, I. L., Stadler, P. F., and Wolfinger, M. T. (2016) SHAPE directed RNA folding. *Bioinformatics,* **32**(1), 145–147.

[43] Kladwang, W., VanLang, C. C., Cordero, P., and Das, R. (2011) Understanding the Errors of SHAPE-Directed RNA Structure Modeling. *Biochemistry,* **50**(37), 8049–8056 PMID: 21842868.

[44] Mitchell, D., Assmann, S. M., and Bevilacqua, P. C. (2019) Probing RNA structure in vivo. *Current Opinion in Structural Biology,* **59**, 151 – 158.

[45] Szymanski, M., Barciszewska, M. Z., Erdmann, V. A., and Barciszewski, J. (2002) 5S Ribosomal RNA Database. *Nucleic Acids Research,* **30**(1), 176–178.

[46] Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research,* **22**(22), 4673–4680.

[47] Eddy, S. R. and Durbin, R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Research,* **22**(11), 2079–2088.

[48] Pang, P. S., Jankowsky, E., Wadley, L. M., and Pyle, A. M. (2005) Prediction of functional tertiary interactions and intermolecular interfaces from primary sequence data. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution,* **304B**(1), 50–63.

[49] Rivas, E., Clements, J., and Eddy, S. R. (2017) A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nature Methods,* **14**(1), 45–48.

[50] Cuturello, F., Tiana, G., and Bussi, G. (2020) Assessing the accuracy of direct-coupling analysis for RNA contact prediction. *RNA,* **26**(5), 637–647.

[51] Fabrizio, P., Zerihun, M. B., Peter, E. K., and Schug, A. (2020) Evaluating DCA-based method performances for RNA contact prediction by a well-curated dataset. *RNA,* **26**(7), 794–802.

[52] Calonaci, N., Jones, A., Cuturello, F., Sattler, M., and Bussi, G. (11, 2020) Machine learning a model for RNA structure prediction. *NAR Genomics and Bioinformatics,* **2**(4) lqaa090.

[53] Burley, S. K., Berman, H. M., Bhikadiya, C., Bi, C., Chen, L., Di Costanzo, L., Christie, C., Dalenberg, K., Duarte, J. M., Dutta, S., et al. (2018) RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.,* **47**(D1), D464–D474.

[54] Cordero, P., Lucks, J. B., and Das, R. (2012) An RNA Mapping DataBase for curating RNA structure mapping experiments. *Bioinformatics,* **28**(22), 3006–3008.

[55] Loughrey, D., Watters, K. E., Settle, A. H., and Lucks, J. B. (2014) SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing. *Nucleic Acids Res.,* **42**(21), e165.

[56] Hajdin, C. E., Bellaousov, S., Huggins, W., Leonard, C. W., Mathews, D. H., and Weeks, K. M. (2013) SHAPE-directed RNA structure modeling. *Proceedings of the National Academy of Sciences,* **110**(14), 5498–5503.

[57] Poulsen, L. D., Kielpinski, L. J., Salama, S. R., Krogh, A., and Vinther, J. (2015) SHAPE Selection (SHAPES) enrich for RNA structure signal in SHAPE sequencing-based probing data. *RNA,* **21**(5), 1042–1052.

[58] Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., Floden, E. W., Gardner, P. P., Jones, T. A., Tate, J., et al. (2014) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.,* **43**(D1), D130–D137.

[59] Lorenz, R., Bernhart, S. H., Zu Siederdissen, C. H., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.,* **6**(1), 26.

[60] Andronescu, M., Condon, A., Hoos, H. H., Mathews, D. H., and Murphy, K. P. (2007) Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics,* **23**(13), i19–i28.

[61] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods,* **17**, 261–272.

[62] Matthews, B. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta,* **405**(2), 442 – 451.

[63] Chicco, D. and Jurman, G. (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics,* **21**(1), 6.

[64] Fowlkes, E. B. and Mallows, C. L. (1983) A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association,* **78**(383), 553–569.

[65] Cawley, G. C. and Talbot, N. L. (2010) On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.,* **11**(Jul), 2079–2107.

[66] Kerpedjiev, P., Hammer, S., and Hofacker, I. L. (2015) Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *Bioinformatics,* **31**(20), 3377–3379.

[67] De Leonardis, E., Lutz, B., Ratz, S., Cocco, S., Monasson, R., Schug, A., and Weigt, M. (2015) Direct-coupling analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction. *Nucleic Acids Research,* **43**(21), 10444–10455.

[68] Weeks, K. M. and Mauger, D. M. (2011) Exploring RNA structural codes with SHAPE chemistry. *Acc. Chem. Res.,* **44**(12), 1280–1291.

[69] Mlýnský, V. and Bussi, G. (2018) Molecular dynamics simulations reveal an interplay between SHAPE reagent binding and RNA flexibility. *The Journal of Physical Chemistry Letters,* **9**(2), 313–318.

[70] Frezza, E., Courban, A., Allouche, D., Sargueil, B., and Pasquali, S. (2019) The interplay between molecular flexibility and RNA chemical probing reactivities analyzed at the nucleotide level via an extensive molecular dynamics study. *Methods,* **162**, 108–127.

[71] Larkin, M., Blackshields, G., Brown, N., Chenna, R., McGettigan, P., McWilliam, H., Valentin, F., Wallace, I., Wilm, A., Lopez, R., Thompson, J., Gibson, T., and Higgins, D. (09, 2007) Clustal W and Clustal X version 2.0. *Bioinformatics,* **23**(21), 2947–2948.

[72] Weeks, K. M. (2010) Advances in RNA structure analysis by chemical probing. *Curr. Opin. Struct. Biol.,* **20**(3), 295–304.

[73] Ziv, O., Gabryelska, M. M., Lun, A. T., Gebert, L. F., Sheu-Gruttadauria, J., Meredith, L. W., Liu, Z.-Y., Kwok, C. K., Qin, C.-F., MacRae, I. J., et al. (2018) COMRADES determines in vivo RNA structures and interactions. *Nature Methods,* **15**(10), 785–788.

[74] DePaul, A. J., Thompson, E. J., Patel, S. S., Haldeman, K., and Sorin, E. J. (03, 2010) Equilibrium conformational dynamics in an RNA tetraloop from massively parallel molecular dynamics. *Nucleic Acids Research,* **38**(14), 4856–4867.

[75] Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004) Development and testing of a general AMBER force field. *Journal of Computational Chemistry,* **25**(9), 1157–1174.

[76] Wang, J., Wang, W., Kollman, P. A., and Case, D. A. (2006) Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics and Modelling,* **25**(2), 247 – 260.

[77] Case, D. A., Cheatham III, T. E., Darden, T., Gohlke, H., Luo, R., Merz Jr., K. M., Onufriev, A., Simmerling, C., Wang, B., and Woods, R. J. (2005) The Amber biomolecular simulation programs. *Journal of Computational Chemistry,* **26**(16), 1668–1688.

[78] Schrödinger Schrödinger Release 2019-4.

[79] Cornell, W. D., Cieplak, P., Bayly, C. I., and Kollman, P. A. (1993) Application of RESP charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation. *Journal of the American Chemical Society,* **115**(21), 9620–9631.

[80] Abraham, M. J., Murtola, T., Schulz, R., Pàll, S., Smith, J. C., Hess, B., and Lindahl, E. (2015) GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX,* **1-2**, 19 – 25.

[81] Bernardi, A., Faller, R., Reith, D., and Kirschner, K. N. (2019) ACPYPE update for nonuniform 1–4 scale factors: Conversion of the GLYCAM06 force field from AMBER to GROMACS. *SoftwareX,* **10**, 100241.

[82] Izadi, S., Anandakrishnan, R., and Onufriev, A. V. (2014) Building Water Models: A Different Approach. *The Journal of Physical Chemistry Letters,* **5**(21), 3863–3871.

[83] Parrinello, M. and Rahman, A. (1981) Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics,* **52**(12), 7182–7190.

[84] Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995) A smooth particle mesh Ewald method. *The Journal of Chemical Physics,* **103**(19), 8577–8593.

[85] Bussi, G., Donadio, D., and Parrinello, M. (2007) Canonical sampling through velocity rescaling. *The Journal of Chemical Physics,* **126**(1), 014101.

[86] Hess, B., Bekker, H., Berendsen, H. J. C., and Fraaije, J. G. E. M. (1997) LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry,* **18**(12), 1463–1472.

[87] Mortimer, S. A. and Weeks, K. M. (2007) A Fast-Acting Reagent for Accurate Analysis of RNA Secondary and Tertiary Structure by SHAPE Chemistry. *Journal of the American Chemical Society,* **129**(14), 4144–4145.

[88] Forsén, S. and Linse, S. (1995) Cooperativity: over the Hill. *Trends in Biochemical Sciences,* **20**(12), 495 – 497.

[89] Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological),* **57**(1), 289–300.