Politecnico di Torino

université PARIS-SACLAY

Department of Applied Sciences and Technology

Master's Degree in
Physics of Complex Systems

Optimization of an RNA coarse-grained force field with Machine Learning

Supervisors

Samuela Pasquali
Frédéric Lechenault
Alessandro Pelizzola

Candidate

Gianluca Lombardi

Academic Year 2021/2022

Acknowledgements

First of all, I want to thank my supervisors Samuela Pasquali and Frédéric Lechenault, who have advised me during my internship in Paris, allowing me to enter for the first time the world of research and to deepen my knowledge in the fields of biophysics and machine learning.

Secondly, I want to thank Alessandro Pelizzola and all the organizers of the master in Physics of Complex Systems, that gave me the opportunity to join an international environment and study in some of the best universities in Europe. This experience, however, would not have been the same without the friends and colleagues that shared it with me from the beginning, the PCSs. I spent the last two years with them and I already miss our dinners together, our study days, our nights out. Thank you very much, to all of you, and in particular to Alessandro and Mattia, the irreplaceable roommates of "Casa maleducata".

The warmest thank you goes to my family, my parents Roberto and Filomena, my brother Alessandro and my grandparents Tommaso and "Tina", who have always been present and supportive, even if many kilometres stood between us.

Finally, I want to thank all my dear friends that I have met before and during these years, from high school to bachelor, from Cornaredo to Paris: thanks to all of you who made me feel at home, no matter where I was.

And thanks to my girlfriend Lucrezia, that shared this year in Paris with me, being my solid anchor in sad times and my sail in joyful moments, from Jardin de Plantes to the top of Mont Saint-Michel.

Abstract

Ribonucleic acid, or RNA, is a linear polymer involved in a wide variety of functions, both in human and in viral cells, as highlighted also by the recent pandemic. Functionality of RNA molecules is strongly linked to their three-dimensional structure, which is the reason why the RNA folding problem has become of great interest in recent years. Among the proposed solutions, one possible approach relies on the design of coarse-grained physical models to speed up molecular dynamics simulations. Among them, HiRE-RNA is a high resolution model with specific functional forms designed to reproduce experimental results. In this internship project, we aim to optimize HiRE-RNA parameters related to local interactions, building a machine learning setup from scratch. Starting from a limited number of heterogeneous RNA sequences, we created a suitable dataset and we developed a model that computes the energies with HiRE-RNA. The optimization is performed with Stochastic Gradient Descent, trying to match coarse-grained energies with those computed in the atomic representation, and results are tested in molecular dynamics simulations. Due to the large number of parameters and the constraints imposed by their physical meaning, different methods are introduced and compared, through the analysis of some physical quantities extracted from the simulations, such as bond lengths or angles. Although the obtained results are not definitive, they already provide an improvement in the model performance and constitute a good starting point for future developments. The software implementations are available at `https://github.com/GianLMB/HiRE_optimization`.

Contents

Introduction

RNA molecules are complex biological objects involved in a wide variety of cellular functions. Similarly to proteins, RNA functional activity depends on the characteristic three-dimensional structure of the molecule and its dynamical behaviour, influenced by the biological and chemical conditions of the surrounding environment. The study of RNA folding has led to the development of computational methods, complementary to experimental techniques, that aim to predict 3D secondary structures. One of these approaches is based on physical models, that consider the interactions of the system's particles and retrieve the complexity of 3D structures trough the minimization of the energy of the system in *ab initio* molecular simulations. All-atoms simulations, however, are extremely expensive in terms of simulation time, that is the reason why simplified models, that use a coarse-grained (CG) description of the molecule, are introduced [1]. In this context, HiRE–RNA is a high resolution coarse-grained force field, where each nucleotide is represented as 6 or 7 beads. This simplification, however, comes at a cost, that is the introduction of more than 200 free parameters related to force couplings, equilibrium values and particle-specific coefficients, that have to be tuned in order to reproduce experimental results. The force field was originally optimized using standard techniques (genetic algorithm maximizing the energy difference between a native structure and decoys) and limited experimental data. The recent advent of machine learning (ML) provides an array of tools with the potential to improve these results.

The aim of this internship project is therefore to develop a procedure, based on machine learning and physical intuition, to optimize the parameters. The report is organized as follows: in chapter 1 we give a brief introduction about RNA molecular structure and the interactions that are involved, which show the complexity of this biological object. We explain then in detail the HiRE-RNA force field, analysing all the energy contributions that are considered in the model.

In chapter 2 we focus on local interactions and we describe the computational implementation of the ML model and the construction of related dataset, starting from a limited number of RNA sequences. Subsequently, we explain the optimization procedure, based on Stochastic Gradient Descent algorithm, and we report the methods used for model evaluation.

In chapter 3 we show the obtained results, beginning with a statistical analysis of the dataset and a study of the performance of the algorithm. A comparison between different optimization methods adopted is then performed, based on molecular dynamics simulations.

What we want to achieve, in the end, is an improvement in the model performance and predicting power, that can be obtained both from an optimization of the parameters appearing in the functions, and from an optimization of the functional form itself. This second task, that could possibly be accomplished with Symbolic Regression, will be the main subject of future works.

1 RNA molecular structure and HiRE-RNA force field

1.1 RNA structural organization and complexity

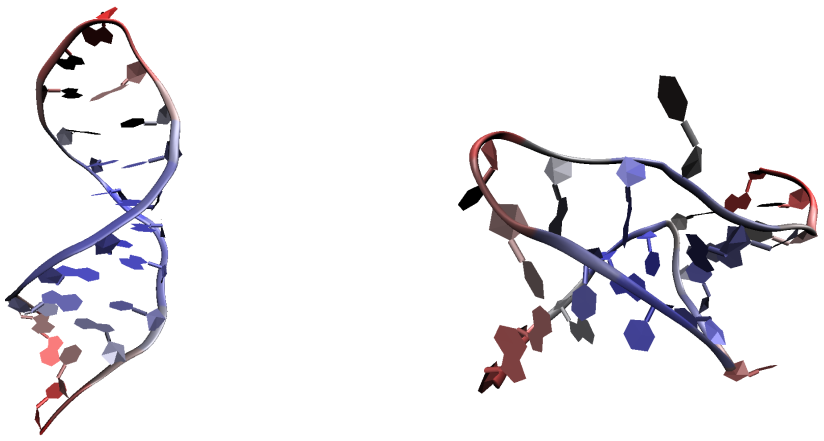
Ribonucleic acid, or RNA, is a linear polymer composed by a sequence of nucleotides, most often arranged in a single-stranded structure. Each nucleotide contains a phosphate group and a ribose sugar, that constitute the negatively charged backbone of the molecule, and a nitrogenous base. In RNA, four different bases are found, that also give name to the corresponding nucleotides: adenine (A) and guanine (G), that are made of two aromatic rings and are referred as purines, and cytosine (C) and uracil (U), that have only one ring and are known as pyrimidines. Aside from their well-known roles as genetic information carriers (mRNA) and amino acid recruiters (tRNA), RNA molecules also participate in regulating gene expression through post-transcriptional processes (miRNA), gene silencing (RNAi), and catalytic activities (ribozymes) [2]. Moreover, the importance of the study of RNA molecules has been highlighted by the recent pandemic, with the SARS-CoV-2 virus featuring an RNA-based genome and a replication mechanism controlled by non-coding RNA.

Due to the vast variety of functions that RNA molecules perform, their length varies in a wide range of values, from the few nucleotides of genes regulating miRNA to the several thousands of ribosomal RNA. The length of the molecule and the sequence of nucleotides is essential in determining the spatial organization of atoms and their dynamical behaviour, which in turn affect RNA functionality. Just like DNA, RNA benefits from sequence complementarity, with A pairing with U and G pairing with C. These well-known pairings, better referred to as Watson-Crick or canonical base pairs, are very stable and contribute to the formation of helical structures (A-form). Ideally, if one only has strands with perfect complementary sequences, the structure of the molecule is a double helix; but RNA sequences almost never allow for base complementarity along the whole sequence, which is the reason for their mostly single-stranded nature and the complexity of their spatial structure and energy landscape. Nevertheless, some common structural units appear at different length scales, an indication that canonical base pairing is only one of the interactions that drive RNA folding. Actually, experiments show that the first non covalent interaction that is involved in the process is stacking [3, 4], that is induced by the planar configuration of bases, which tend to align themselves and expose their charged exocyclic groups to the (typically polar) solvent. This characteristic, combined with steric and electrostatic backbone constraints, is the actual main cause of the stable helical structure [5].

Moreover, in most cases of folded RNA three-dimensional structures, non-canonical pairs are critical for creating the tertiary interactions that stabilize the functional conformation. In fact, in principle every nucleotide in an RNA chain can favorably

interact with every other nucleotide through hydrogen bonds, and most of all possible configurations are actually found in RNA molecules. Each nitrogenous base has three sides, identified as Watson-Crick (WC), Hoogsteen and Sugar, that can be involved in more or less stable hydrogen bonds. It is therefore not uncommon to also find bases forming multiple simultaneous interactions, giving rise to structures known as triplets and quadruplets.

Stacking, base pairings and electrostatic forces drive RNA folding towards the formation of typically very compact structures, that correspond to minima of the thermodynamical landscape, determined by surrounding conditions and the folding process itself. Depending on the nature of the landscape, a given RNA populates an ensemble of conformations of various compactness, stabilities, and flexibilities. As the length of an RNA molecule increases, the complexity of the landscape and the number of possible nonidentical states increases, but favored states remain. This is the reason why RNA is commonly perceived as an unstructured molecule, even if that is not the case, as we have seen. Hairpins, bulge loops, pseudoknots, helices dockings, ribose zippers are only few of the structures arising in these molecules (see Figure 1.1), and models that automatically predict secondary structures are not always able to capture this wide and complex variety, despite their great improvement in the last five years [6]. In general, they are very well suited for regions that contain a good percentage of canonical pairings, while their accuracy is limited for more intricate 3D structures [7].



(a) 1f9l hairpin. **(b)** 2g1w pseudoknot.

Figure 1.1: Examples of common RNA structures: hairpin and pseudoknot.

1.2 HiRE-RNA force field

Among the possible solutions to the RNA folding problem, one is the introduction of specific *ab initio* physical force fields to carry out the molecular simulations, that are otherwise computationally too expensive to be performed with an all-atoms description. To overcome the limitations imposed by the size of the molecule, and follow the large scale rearrangements occurring in folding, one can resort to a simplification of the system through coarse graining. The challenge of this approach is to design a force field able to capture all of the subtle interactions that give rise to

folding while maintaining a sufficiently simple description of the system for efficient simulations.

HiRE-RNA is an effective model designed to fold any RNA architecture and study the structural dynamics of RNA molecules [8, 9]. The model points at the main physical interactions involved in RNA folding, with specific functional forms that aim to reproduce stacking and base pairings, including non-canonical and multiple pairs. Coarse graining is performed representing each nucleotide as six or seven beads, for pyrimidines and purines respectively. Among these beads, one corresponds to the phosphate, P, four to the sugar atoms, O5, C5, C4, C1, and the remaining ones to the centers of mass of the aromatic rings that form the nitrogenous bases (Figure 1.2).

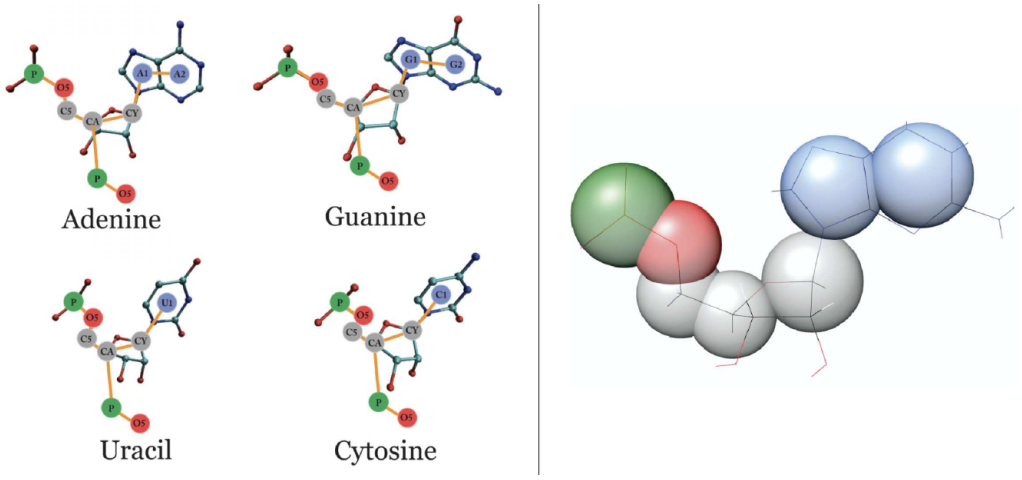


Figure 1.2: Nucleotides in the HiRE-RNA model. On the left, comparison of atomistic and coarse-grained representation of the four types of nucleotides, also including the connection to the following bases; on the right, a closer view to guanine residue, with the actual dimension of the beads, based on their covalent radius. In the text, particles O5, C5, CA and CY are also referred to as O, C, R1, R4.

The interactions between these beads can be divided in two groups, local and non-bonded interactions, and are described in the following sections. The whole model is implemented in a Fortran code, that computes both the energy and the forces, by differentiation. In this way, it allows to compute both the energy landscape of a sequence, useful for its thermodynamic analysis, and to perform molecular dynamics simulations to study the time evolution of the system.

1.2.1 Local interactions

The potential energy for the local interactions among particles in the molecule is the sum of three terms:

$$E_{loc} = E_b + E_a + E_d \tag{1.1}$$

where E_b is related to the bond lengths between two particles, E_a to the bond angles defined by three particles, and E_d to the dihedral angles defined by quadruples of particles. Each of these terms has the form of a sum of simple harmonic or periodic potentials over the set of interacting particles , where the coupling constants and the equilibrium values are the parameters to optimize. Information regarding indexes of

interacting particles and type of interaction are stored in topology `parameters.top` files, that are generated along with the coarse-grained structure of the molecule.

Bond energy

Each harmonic term for the bond energy between two particles i, j has the form

$$E_b = \epsilon_b \, k_b \, (d - d_0)^2 \tag{1.2}$$

where ϵ_b is a global weight coefficient, k_b and d_0 are the particle dependent coupling constant and equilibrium distance respectively, and $d = |\vec{x}_i - \vec{x}_j|$ is the distance between the particles.

Angle energy

Similarly, for angle energy between particles i, j and k the expression is

$$E_a = \epsilon_a \, k_a \, (\theta - \theta_0)^2 \tag{1.3}$$

with analogous notation as before, where θ is defined as

$$\cos\theta = \frac{\vec{r}_{ij} \cdot \vec{r}_{kj}}{|\vec{r}_{ij}| \, |\vec{r}_{kj}|} \tag{1.4}$$

and $\vec{r}_{ij} = \vec{x}_i - \vec{x}_j$. Unlike the bond energy coefficient ϵ_b, which is independent of the particles, ϵ_a is the product of a global scaling factor and a coefficient depending on the angle type.

Torsion energy

The dihedral angle φ is defined as

$$\cos\varphi = \vec{n}_j \cdot \vec{n}_k \tag{1.5}$$

where $\vec{n}_j$ and $\vec{n}_k$ are the normals to the planes defined by particles i, j, k and j, k, l, respectively, normalized to 1. In terms of φ, the torsion energy is defined as

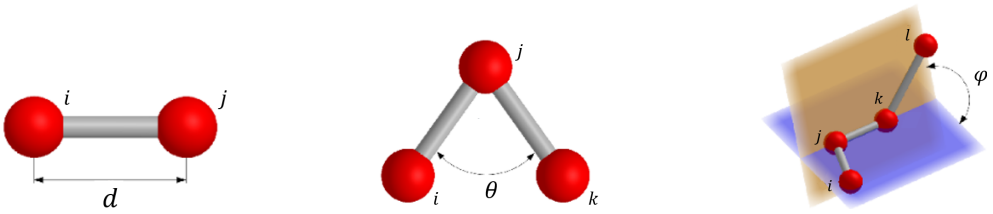
$$E_a = \epsilon_d \, k_d \, [1 + \cos(m\varphi - \varphi_0)] \tag{1.6}$$

where m is an integer number that describes the multiplicity of the state and φ_0 is the equilibrium angle. Similarly to the angle case, ϵ_d is the product of a particle independent and a particle dependent factor. A representation of local interactions is showed in Figure 1.3.

1.2.2 Non-bonded interactions

The potentials contributing to the non-bonded interaction energy are the short range excluded volume potential E_{ex}, long range electrostatic potential E_{el}, stacking potential E_s, and base pairing potential E_{bp}. The total energy is the sum of these potentials:

$$E_{nb} = E_{ex} + E_{el} + E_s + E_{bp}. \tag{1.7}$$



(a) Bond length. **(b)** Bond angle. **(c)** Dihedral angle.

Figure 1.3: Visual representation of local interactions.

While for local interactions the functional forms are the same used in atomistic descriptions, for non-bonded interactions there is not an exact correspondence, and expressions are designed to reproduce empirical results. All of the terms are computed over each pair of atom belonging to different residues, whose distance is below a certain cutoff, depending on the interaction type.

Excluded volume

The excluded volume potential is represented as a sigmoidal function:

$$E_{ex} = \epsilon_{ex} \left[1 - \frac{1}{1 + e^{-\kappa(d-d_0)}} \right] \tag{1.8}$$

where ϵ_{ex} defines the maximum strength of the interaction, κ controls the width of the sigmoid and d_0 is the reference distance between the beads. When $d = |\vec{x}_i - \vec{x}_j|$ approaches an interpenetration distance, the potential rapidly increases to strongly discourage that configuration.

Electrostatic potential

Using an implicit description for ions and water molecules, electrostatic repulsion of the phosphates and charge screening are represented by a Debye–Hückel potential between P particles, characterized by a screening length λ_D that depends on ion concentration:

$$E_{el} = \epsilon_{el} \frac{q_i q_j}{4\pi\varepsilon} \frac{e^{-d/\lambda_D}}{d}. \tag{1.9}$$

Under physiological conditions, due to their negatively charged backbone, RNA molecules are typically surrounded by positive ions, that can provide a ionic screening or either act as structural ions, that actively affect the folding process.

Stacking

Both stacking and hydrogen bond interactions depend on the relative orientations of the bases involved, which is described with the introduction of a base plane. The plane is defined by the last three beads of each residue, through the normal vector $\vec{n}_I$, and it does not involve any information about the backbone orientation. The expression for the energy is the following:

$$\begin{aligned} E_s = -\epsilon_s e^{-\frac{(r_I - r_0)^2}{\sigma}} e^{-\frac{(r_J - r_0)^2}{\sigma}} \left[1 - (1 - (\vec{n}_I \cdot \vec{n}_J)^2)^2\right] \left[e^{-A_1(1-\cos^2(\theta_I - \theta_1))}\right. \\ \left. + e^{-A_2(1-\cos^2(\theta_J - \theta_2))}\right] \left[e^{-A_1(1-\cos^2(\theta_J - \theta_1))} + e^{-A_2(1-\cos^2(\theta_I - \theta_2))}\right] \end{aligned} \tag{1.10}$$

where r_I (r_J) is the distance between the centre of mass of base I (J) and the plane defined by base J (I):

$$\vec{r}_I = \vec{r}_{IJ} \cdot \vec{n}_J, \qquad \vec{r}_J = \vec{r}_{IJ} \cdot \vec{n}_I, \qquad \vec{r}_{IJ} = \left| \sum_{k=1}^{3} \frac{1}{3} (\vec{r}_{I_k} - \vec{r}_{J_k}) \right| \tag{1.11}$$

and the parameters r_0, σ, A_1, A_2, θ_1, θ_2 are properties of the base.

Base pairing

Base pairing occurs when bases are coupled on almost parallel planes, and the interaction strength depends on the orientation of the bases and the distance between particles that form the hydrogen bonds. In the HiRE-RNA model, the potential is the product of a term that assures planarity, E_{pl}, and another term related to hydrogen bonds, E_{hb}.

Planarity is obtained by introducing Gaussian weights on the position of the three terminal particles of each residue, that are those forming the nitrogenous base, in order to assure that the beads of one base lie on the plane defined by the other one. In mathematical terms, we have

$$E_{pl} = \epsilon_{pl} \left[\sum_{k_I=1}^{3} e^{-(d_I^{k_J}/\delta)^2} \right] \left[\sum_{k_J=1}^{3} e^{-(d_I^{k_J}/\delta)^2} \right] \tag{1.12}$$

where $d_I^{k_J}$ is the distance between particle k belonging to base J and the plane defined by base I, and ϵ_{pl} and δ are the coupling constant and width parameters.

The hydrogen bond term is instead a function of the inter-base distance $\rho = |\vec{\rho}|$, with $\vec{\rho} = \vec{x}_I - \vec{x}_J$ and the angles α_I and α_J, defined in an analogous way as

$$\cos \alpha_I = \vec{n}_{I\rho} \cdot \vec{n}_{I_{32}} \tag{1.13}$$

with

$$\vec{n}_{I\rho} = \frac{\vec{\rho}_I}{|\vec{\rho}_I|}, \quad \vec{\rho}_I = -\vec{\rho} + (\vec{\rho} \cdot \vec{n}_I)\vec{n}_I \qquad \text{and} \qquad \vec{n}_{I_{32}} = \frac{\vec{r}_{I_{32}}}{|\vec{r}_{I_{32}}|}, \quad \vec{r}_{I_{32}} = \vec{x}_{I_3} - \vec{x}_{I_2}. \tag{1.14}$$

The potential then takes the form

$$E_{hb} = -\epsilon_{hb}\, k_{hb}\, e^{-(\rho-\rho_0)^2/\xi}\, \nu(\alpha_I)\, \nu(\alpha_J) \tag{1.15}$$

where k_{hb} is a pair-specific coefficient, ρ_0 and ξ are the equilibrium distance and Gaussian width, and $\nu(\alpha)$ is a function that assures that only values of α close to the equilibrium value α_0 contribute to the potential. Explicitly, we have

$$\nu(\alpha) = \begin{cases} \cos^6(\alpha - \alpha_0) & \text{if } \cos(\alpha - \alpha_0) \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{1.16}$$

In order to break the symmetry of the function with respect to the sign of α, a dihedral angle τ between the particle forming the bond and the three beads of the

other base is introduced. In this way, the value of α is corrected according to

$$\alpha = \begin{cases} +\alpha & \text{if } \cos\tau > 0 \\ -\alpha & \text{otherwise.} \end{cases} \tag{1.17}$$

The model is designed to represent not only canonical pairs, but 22 possible configurations involving different combinations of bases and all of their sides, each one associated with a specific set of distances, angles and number of hydrogen bonds (shown in Figure 1.4). The choice of 22 interactions is arbitrary, but is based on

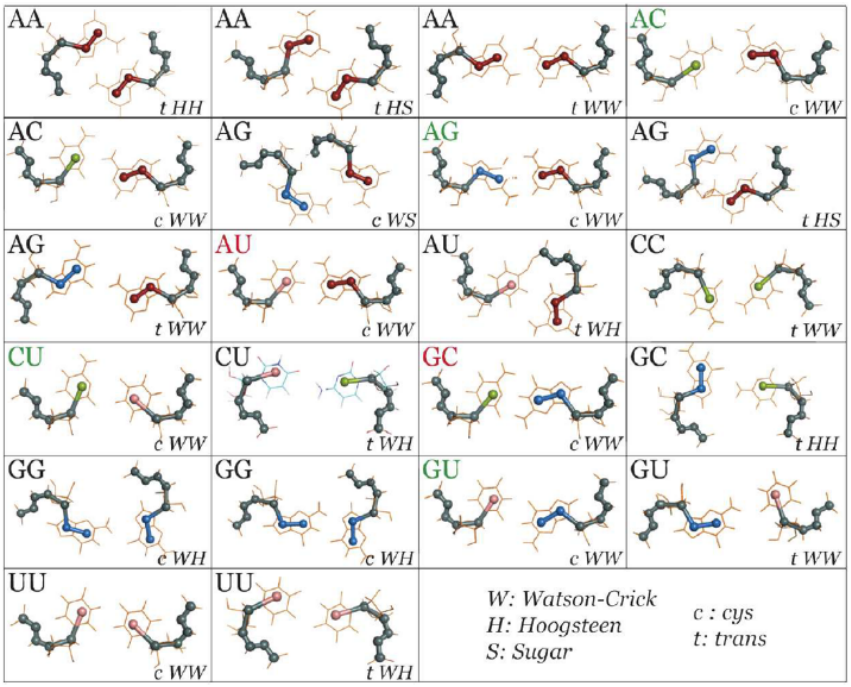


Figure 1.4: Set of 22 base pairs considered in the HiRE-RNA model. Canonical WC pairs, considered by all CG models, are highlighted in red, while in green are non-canonical pairs occurring at WC side of the bases.

their occurrence according to the Nucleic Acid Database (NDB) [10]. Thanks to a combination of the hydrogen bonds potential and the excluded volume constraints, HiRE-RNA model automatically respects the atomistic condition that each side of the nitrogenous base can pair with only one base, with hydrogen bonding acceptor and donor atoms involved in at most one interaction at a time.

2 Optimization procedure

Looking at the several terms that define the HiRE-RNA model, it is clear that they involve a large number of free parameters concerning geometric properties and couplings that balance the strength of the different type of interactions. In order to reproduce structural results obtained from experiments, and to perform sufficiently accurate simulations, these parameters need to be chosen with high precision. A preliminary work have already been carried out in this direction, with two different methods. Geometric parameters were obtained from a statistical analysis on 200 sequences extracted from NDB database, including molecules of varying sizes and topologies, while global energetic parameters where optimized with genetic algorithms. A genetic algorithm is a stochastic classical evolutionary algorithm, and it is based on Darwin's theory of evolution. The optimization is performed by randomly varying in a small range the values of parameters, that behave like genes in a chromosome, and selecting only those mutations that improve the fitness, a function that reflects the ability of the model to explain the observed data [11].

The next step is to improve these results with machine learning, building a model that uses data concerning the position of particles and their interactions to optimize the parameters. Since the number of parameters is very large (more than 200), the optimization procedure has been organized in two consecutive parts, related to local interactions and non-bonded interactions respectively, that involve different parameters and methodologies. In this report, we focus only on local interactions, mainly for three reasons. First, we expect their functional forms to be fixed, and therefore they should not require any type of symbolic treatment; second, they do not depend much on the global structure of the molecule, which imposes less constraints in the choice of the training set in terms of sampled structures; third, the effect of non-bonded interactions is limited, which is useful to appreciate the effect of variations of the parameters in molecular dynamics simulations.

2.1 Model and dataset construction

Machine learning (ML) is a subfield of artificial intelligence with the goal of developing algorithms capable of learning from data in an automatic way. In almost every machine learning setup, the classical ingredient are a dataset X, a model $g(\vec{\theta})$, which is a function of the parameters $\vec{\theta}$, and a loss function $C(X, g(\vec{\theta}))$, also called sometimes cost function, that allows to quantify how well the model $g(\vec{\theta})$ explains the observations X. The model is fit, or trained, by finding the values of $\vec{\theta}$ that minimize the loss function [12].

Nowadays, many ML framework are available, that allow to deal with the model implementation and training in an intuitive way. The language used for this project is Python, and the machine learning framework adopted is PyTorch [13], for its Automatic Differentiation tools. Its basic data structures are called tensors, and they

correspond to multi-dimensional arrays with possibly associated gradient functions. All algebraic operations between tensors are also implemented, with a syntax that closely resembles that of NumPy.

Every model in PyTorch is defined starting from the `Module` class. Its key method is called `forward`, where the structure of the model is implemented. It receives data points as inputs, and it contains all the consecutive operations and transformations performed on them, that lead to the final output. For instance, it could contain the definition of the different layers of a neural network, or implement a simple linear regression, or even compute some quantities of interest for a subsequent analysis. In our particular case, the model computes separately the bond, angle and torsion contributions to the energy of the HiRE-RNA force field, for each RNA sequence received as input.

The loss function is designed with the aim to find a match between energies computed with the CG HiRE-RNA model and with an all-atom (AA) model. The latter is obtained with Amber (Assisted model building with energy refinement) [14], a package of molecular simulation programs, and the unit of measure for all energy computations is kcal/mol. One reason for this choice is that, for the optimization, we can not simply rely on the minimization of the total energy of the system, since local interactions alone are not in an energy minimum. Such approach could instead be possible for non-bonded interactions, that effectively drive the molecule towards its minimized configuration, but only once local interactions are fixed, to reduce the number of degrees of freedom. Another reason is that an energy matching would provide a more accurate energy scale for local interactions, that is essential if we want to compute thermodynamical quantities, and should lead to a more similar behaviour in terms of dynamical evolution between CG and AA representations. To achieve this, the loss function is defined as the sum of the squared difference between each energy term computed with the HiRE-RNA model and with Amber, averaged over the inputs. This kind of loss function leads to the matching of each energy contribution separately during the optimization, preventing compensation of one term against another, that would happen if one considers only the total energy.

In order to train the model, the dataset must contain a sufficiently large amount of RNA sequences to capture all the types of interactions we are interested in. At the moment, since we focus only on local interactions about bonds, angles and torsions, short sequences are a good compromise to guarantee an adequate level of heterogeneity, without affecting the performance of the algorithm. To achieve this, 200 initial structures with different lengths, taken from NDB database, are randomly cropped into 1640 sub-sequences, composed by 7 nucleotides each (see Figure 2.1). The operation is done by parsing the structure files (`.pdb` extension), that store the information regarding the sequence of nucleotides, their type and the Cartesian coordinates of the particles, and selecting only the desired nucleotides with the tools provided by the BioPython module [15].

For each of the new structures obtained, we performed a relaxation with Amber, to assure that the subsequences correspond to minimized configurations, and we computed their exact energies in the AA representation. The relaxation leads to more stable structures and is useful to analyse the results of the optimization in the molecular dynamics simulations: ideally, when no external force is applied, we expect the system to remain close to its equilibrium position, even in the CG representation. After that, we generated the coarse-grained structure file and the

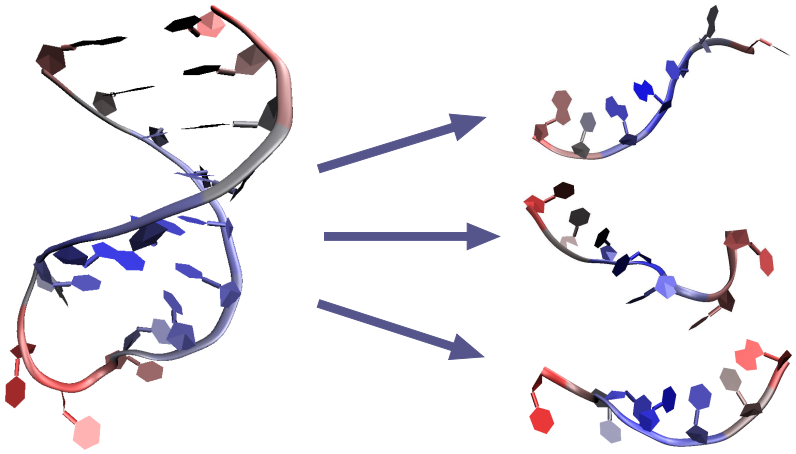


Figure 2.1: Examples of cropped sequences obtained from 1ato PDB structure. The new subsequences are in turn minimized to find their optimal structure.

topology file, that contain all the relevant features for the energy computation with the HiRE-RNA model. They include the type of atoms and interactions, the indexes of atoms involved, and also information about the sequence of residues and mass and charge of particles, that are not considered in local interactions, but are necessary to compute non-bonded energies and forces.

The features are then extracted from the different files and collected in DataFrames, which are stored in `csv` files. Features are padded in order to obtain the same matrix structure for all the sequences, where each column corresponds to a feature. Those that are used in the model for local interactions are:

- `atom_type`, an integer array containing indexes associated to the types of each atom, used for the choice of all particle dependent coefficients;
- `coordinates`, with x, y and z coordinates of each particle;
- `bonds`, that contains the indexes of the pairs of atoms forming the bonds and of the type of bond, necessary for the choice of parameters;
- `angles`, containing the indexes of the three atoms forming each angle and of the type of angle;
- `torsions`, that contains the indexes of the four atoms defining each dihedral and of the type of dihedral;
- `energies`, with the energies components computed with Amber.

To retrieve the original lengths of the padded arrays, a new file `seq_frame.csv` is created, containing the name of the sequences and the corresponding lengths of each array. This data structure, despite being more expensive in terms of memory than a dictionary, allows a direct conversion to tensors, which is performed when the dataset is allocated. In the end, each sequence is encoded as a dictionary with two entries:

- `lengths`, a 1-dimensional tensor that contains the array with the original lengths of the features;

– `features`, a 2-dimensional tensor containing the matrices of padded features.

This structure also allows the iteration and the automatic batching implemented in PyTorch `Dataloader` class.

To verify the heterogeneity of the sequences in the dataset, a statistical analysis was performed, that shows the occurrence of the different type of interactions and their distributions in terms of energy. The results are reported in Section 3.1. In this context, Principal Component Analysis (PCA) was also applied, which consist in an orthogonal transformation in the data space to highlight the directions of larger variance, and it is also useful for data visualization.

2.2 Stochastic Gradient Descent and model evaluation

The optimizer is an algorithm that performs the minimization of the loss function. The classical one is called Stochastic Gradient Descent (SGD), that consists in a stochastic approximation of the Gradient Descent algorithm. In the simplest case, parameters $\vec{\theta}$ are updated, in each iteration, according to

$$\vec{\theta}_{t+1} = \vec{\theta}_t - \eta_t \nabla_\theta E(\vec{\theta}_t) \tag{2.1}$$

where $\nabla_\theta E(\vec{\theta}_t)$ is the gradient of the loss function with respect to the parameters, and the coefficient η_t, called learning rate, controls the size of the step we take in the direction of the gradient at time t. When performed on the whole dataset, gradient descent is a deterministic algorithm, that for specific initialization of parameters $\vec{\theta}_0$ and choice of the learning rate, converges to the same local minimum. As it can be expected, the convergence to a certain minimum and its velocity strongly depend on the choice of the learning rate: the smaller η_t, the more steps are required to reach the local minimum. In contrast, if it is too large, the minimum could be overshoot and the algorithm becomes unstable. To mitigate the limitations of the algorithm, such as its computational cost, its extreme sensitivity on the choice of the learning rate and its dependence on initial conditions, stochasticity is introduced. This is usually achieved by replacing the actual gradient over the full data at each step by an approximation to the gradient computed using a subset of the dataset, called minibatch. Typical minibatches range between 10 and 100 data points, but they can be as small as one single data point. The procedure is then repeated over several training epochs, where a random permutation of the dataset is performed.

Variations of the basic algorithm are possible, that update parameters collecting information also from the second derivative of the loss function, or an approximation of it. One of these examples is Preconditioned Stochastic Gradient Descent (PSGD) [16, 17], that aims to assign a similar weight to the gradient in each direction of the parameters space. This second-order optimization method improves the convergence in non-convex manifolds or in situations of high gradient noise, allowing a faster escape from saddle points in regions where gradients would otherwise be too small, and was also tested in this project.

Moreover, learning rate and minibatch size are external hyperparameters that have to be selected in order to improve the performance of the algorithm[18]. Especially for the learning rate, a typical choice, and also the one adopted in our case, is

to reduce it along the training procedure. It can be done using adaptive algorithms, that already include this feature (the most used one is Adam [19]), or introducing a scheduler that reduces the learning rate when particular conditions are encountered, for instance if the loss function does not decrease after a certain amount of epochs.

For our problem, parameters that have to be optimized can be divided into four main groups:

- global parameters, related to the coefficients ϵ_b, ϵ_a and ϵ_d;
- coupling constant and equilibrium distance for bond lengths;
- coupling constant and equilibrium angle for angles;
- coupling constant and phase for dihedrals.

The initial values of all parameters can be found in Table A.1. The optimization is then performed as described before, minimizing the loss function using SGD algorithms with different choices of the hyperparameters. As it is usually done, the dataset is divided in two subsets (with a 4:1 ratio), where the first one is used to train the model, while the second one acts as a test set for the evaluation.

The performance of the model is then estimated in two ways: first, a comparison between the energy distributions in the AA and CG representations, with the new parameters obtained; second, the HiRE-RNA model is tested in molecular dynamics simulations on some minimized sequences taken as examples. In particular, we observe how the initial structure, that correspond to a minimum of the AA potential, changes after the introduction of the CG representation with the new parameters. Both the qualitative representation and a quantitive analysis of the molecular configuration are performed with VMD (Visual Molecular Dynamics) program [20], that provides a wide variety of methods to read and display the content of structural PDB files.

While in the AA description one can extrapolate the correct values for the single bond, angle or dihedral interaction from Amber, this is not feasible for our CG model, since most of the particles do not have a direct correspondence. This would require an analysis on all the possible positions of the actual atoms that give rise to a particular CG configuration, that is beyond the scope of this project. We must therefore rely only on the sum of the single interactions, that is the reason why the scores that are commonly used to evaluate results of a model are not informative enough in our case. Indeed, the matching of the total energy for each interaction is an important factor, but also other features have to be taken into account, like the actual values of the single parameters, that have to satisfy some constraints (in case of couplings), or match some particular distributions (in case of geometric parameters).

One way to quantify the accuracy of a molecular dynamics simulation is to compute the root-mean-square deviation (RMSD) with respect to the initial conformation [21]. It is defined as

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d_i} \tag{2.2}$$

where d_i is the distance between atom i and either a reference structure (like in our case) or the mean position of the N equivalent atoms. Starting from the minimized AA configuration, we expect the RMSD to stay below 1 Å for a well-behaved model. Otherwise, it would be the indication of a variation from the original structure that is beyond the effect of the CG approximation, meaning that the molecule effectively behaves in a different way with respect to the empirical observations. Both the RMSD and the distributions related to geometric parameters, and in general the analysis of the trajectory, are obtained with the MDtraj module [22].

3 Results

3.1 Dataset analysis

Before entering the results of the optimization procedure, we report here the preliminary analysis performed on the dataset. First of all, we checked that the four types of residues where present in a sufficient quantity and in a similar proportion, in order to not have a biased dataset. The resulting bar plot is reported in Figure 3.1. They range from roughly 2600 to 3400 nucleotides, corresponding to as

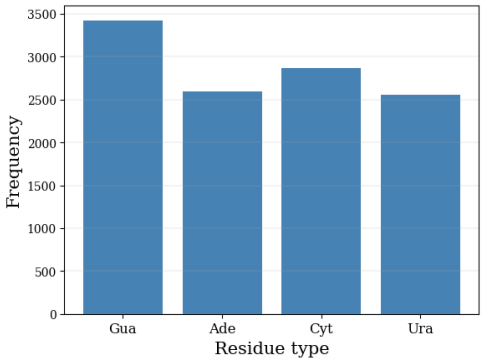


Figure 3.1: Frequency of residues appearance in the dataset.

many residue specific interaction types, that gives a sufficiently crowded statistical sample to consider. The energy distributions for bonds, angles and torsions computed with Amber in the AA representation are reported in Figure 3.2. As one can

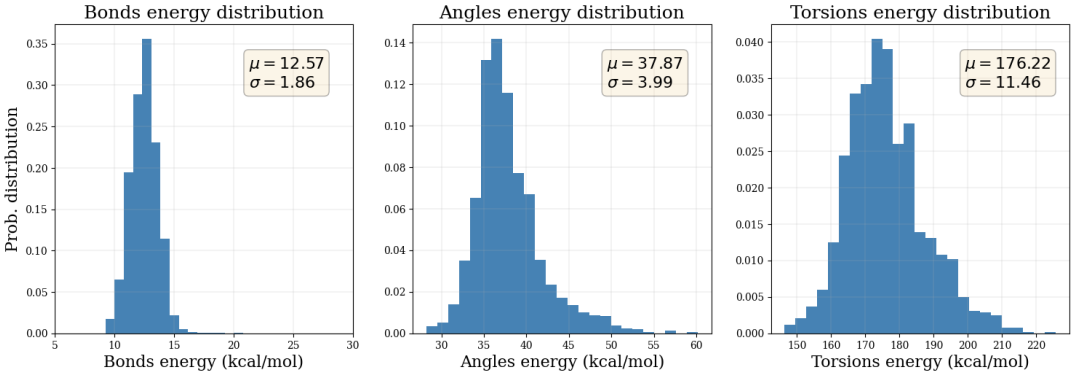


Figure 3.2: Energy distributions in the dataset for the three types of interactions (bonds, angles and torsions) computed with Amber in the AA representation. Average value μ and standard deviation σ for each distribution are also reported.

see, all the distributions have a similar slightly asymmetrical shape, but the range of

values in which they vary is distinct. In particular, there is a difference of one order of magnitude between bonds and torsions energy, that justifies the separation of energy computation in the ML model and in the loss function, otherwise the bonds contribution would be totally neglected.

To visualize how the different terms affect the total energy for each interaction type, we construct matrices for bonds, angles and torsions, where each column corresponds to a sequence, while each row identifies a type of interaction, with a specific coupling constant and equilibrium distance. The entries are then the average energy of that type, for each sequence in the dataset, computed with the original parameters of the HiRE-RNA model in CG representation. This is done to check if some type of interactions give a contribution that is sensibly higher with respect to others, and also to find possible correlations with the total energy. In this perspective, PCA is performed on the obtained matrix, after a normalization is applied. The results for bonds interactions are shown in Figure 3.3. From the matrix

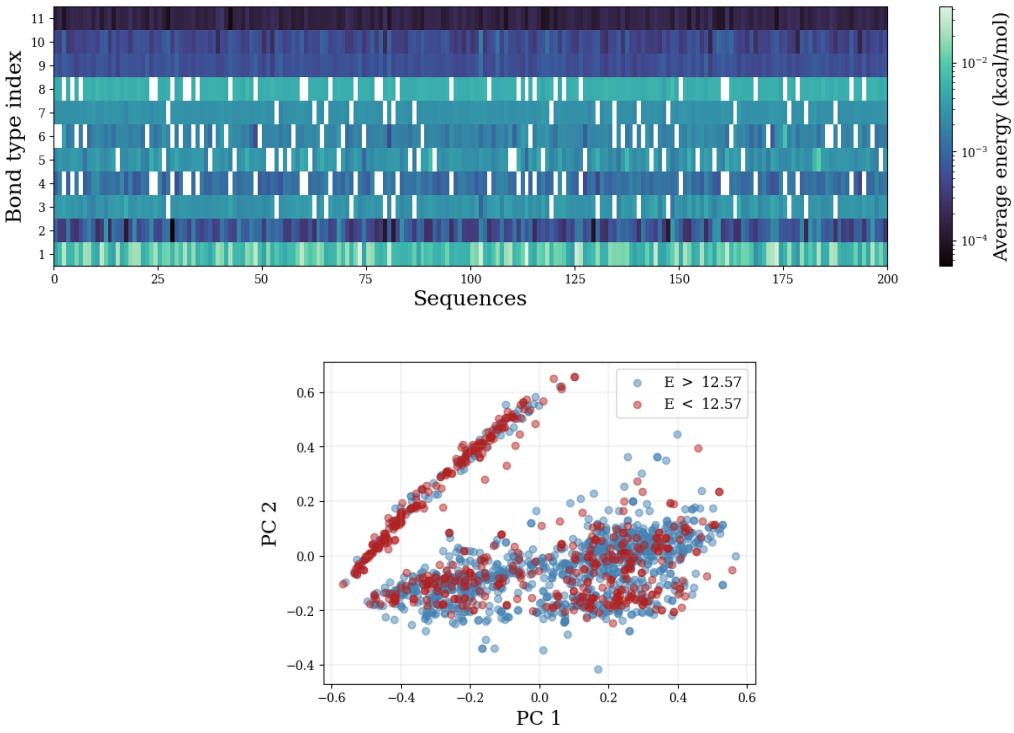


Figure 3.3: Visualization of HiRE-RNA energies for bonds interactions. On the top, heatmap showing the first 200 columns of the bond matrix, as defined in the text; each entry corresponds to the average energy for a specific bond type, in a given sequence. On the bottom, representation of the first two principal components of the normalized matrix, with different colors for points corresponding to sequences with energy respectively above or below the mean value.

representation, it can be noticed that the average energy is similar between the sequences, for the same type of interaction. This is not surprising, since the dataset is made of minimized sequences, therefore we expect distances between particles to be generally close to their equilibrium values. It is also not surprising that bonds that involve nitorgenous bases have also compatible energies: they correspond to to the index from 3 to 8 (for all indexing adopted, the reference is Table A.1),

and some entries are empty because that type of residue does not appear in the considered RNA sequence. More interesting is instead the comparison between P-O energy (index 11) and R4-P energy (index 1), that are extremely different in the CG representation. The reason is that, while the former is in direct correspondence with the actual physical bond, the latter is an approximation that neglects two atoms, thus causing a much bigger variance in the distances, that in turn increases the value of the average energy, given its harmonic potential shape. We expect to find a similar energy ratio also in our optimized model.

PCA, however, is not much more informative. We represented with different colors sequences with a bonds energy computed with Amber that's below or above the mean, projected on the two first principal components, that represent respectively 56% and 23% of the total variance. Even if some regions present different concentrations of the two types, there is not a clear distinction. An analogous analysis was carried out on angles and torsions energies, obtaining similar results.

3.2 Hyperparameters selection and free minimization

In order to perform the optimization, hyperparameters have to be chosen first, including minibatch size and learning rate. In our case, the choice of the latter has a large influence on the results of the model: the range is limited between 10^{-7} and 10^{-3} to avoid a slow convergence in one sense, and unfeasible results in the second; at the same time, as we will see later, its optimal value can not be simply estimated relying on the convergence of the loss function. Minibatch size, instead, does not affect much the results of the parameters, therefore it can be chosen based only on execution time and decrease in the loss function. Since the dataset includes more than 1000 sequences, a full batch approach was excluded, and values considered range from 1 (single point batch) to 256 elements. The results in terms of loss function and execution time, for a learning rate set to 10^{-3}, are shown in Figure 3.4. While the execution time is similar for all the minibatch sizes considered, except for

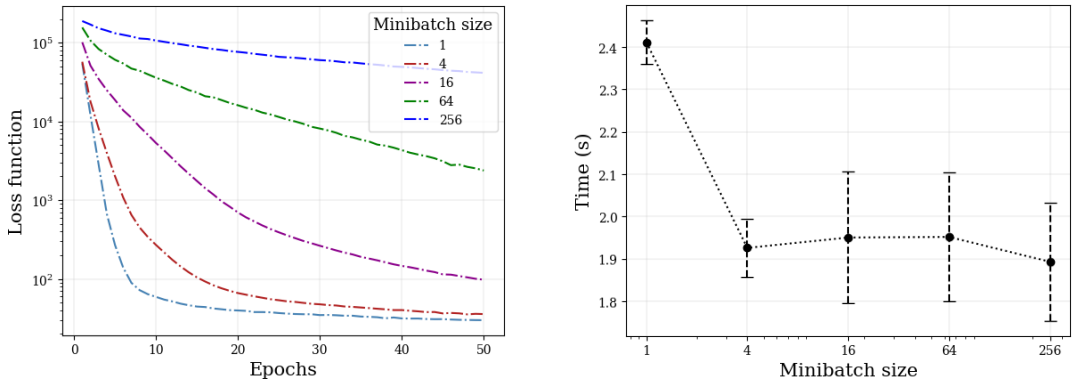


Figure 3.4: On the left: Loss function computed on the training set over the first 50 epochs, for different minibatch sizes. On the right: Average time (in seconds) for epoch, with its standard deviation, for the minibatch sizes considered.

the single point batch, that is 20% higher, the convergence rate of the loss function dramatically increases with the number of elements. While this is expected, it is also

a sign that the program does not exploit completely the batched structure of the inputs. The reason is that, despite being similar, each sequence contains different numbers and types of atoms, and as a consequence the number of bonds, angles and dihedrals also varies. Therefore, the energy computation must be performed on each sequence on its own, while the minibatch allows to average the gradients and avoid too dissimilar values. Given the results above, we chose to adopt minibatches made of 4 elements, that are a good compromise between fast convergence, execution time and modulation of stochastic noise. Concerning the optimizer and the scheduler, we have opted for Adam algorithm with a learning rate of 10^{-4}, filled with a scheduler that automatically reduces learning rates after 20 epochs if the loss function does not decrease.

With this setup, a first optimization was performed, letting all parameters vary in a free range of values, with their default initialisation. As and indicator for the accuracy of the model, we also considered the R^2 score. If we call y_i the expected values (computed with Amber) with average $\bar{y}$, and g_i the predicted values for RNA sequence i, R^2 is defined as

$$R^2 = 1 - \frac{S_{res}}{S_{tot}} \tag{3.1}$$

with

$$S_{res} = \sum_{i=1}^{N} (y_i - g_i)^2 \qquad\qquad S_{tot} = \sum_{i=1}^{N} (y_i - \bar{y})^2 \tag{3.2}$$

that are called respectively residual sum of squares and total sum of squares. A value of R^2 close to 1 denotes an accurate matching between the expected value and the prediction, while 0 or even negative values are the signal of a worse prediction. The training was executed over 400 epochs, and the results are reported in Figure 3.5. The decrease in the loss function, as well as the gradual increment in the R^2 score and the very good matching between Amber and new HiRE-RNA energy distributions prove that the model works in the correct way, especially in comparison with the initial distribution. To be more specific, the distribution of the initial torsions energy has a similar variance with respect to the correct one, but a lower mean value, while variance for bonds and angles energy is much greater than the expected value, to the extent that in favour of a clearer representation it was necessary to limit the range on the x-axis in the plots. For instance, angle energies range from roughly 0 to 3000 kcal/mol, while Amber values are between 20 and 60 kcal/mol. At the end of the optimization procedure, the loss function reaches a value of 28, corresponding approximately to an average energy difference of 3 kcal/mol for each type of interaction, which is indeed a good matching.

However, if one looks at the new values obtained for the parameters, it is not clear if they can represent the correct physical properties of the corresponding interactions. In particular, we focus on global parameters and some geometric parameters taken as examples, that are shown in Table 3.1. Concerning the global parameters, that define the ratio between the different types of interaction, it is clear that those for bonds and angles were highly overestimated, with respect to the energies obtained with Amber. While this feature seems to be correctly captured, the same is not true for the equilibrium values, that should be initially very close to the correct ones, being extracted from empirical distributions. Instead, a large variation is observed, especially in torsion parameters. At the same time, couplings do not differ

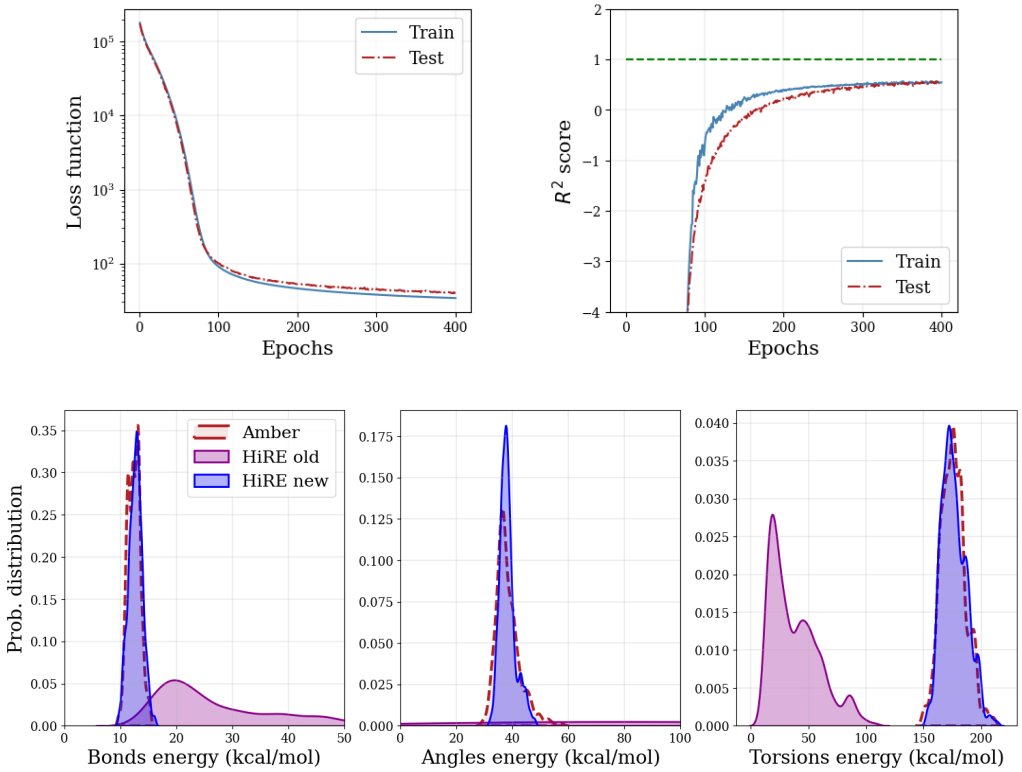


Figure 3.5: Results of the free optimization. In order: loss function for train and test set over 400 epochs; R^2 score computed for train and test set as function of the epochs; distributions of bonds, angles and torsions energy computed with Amber AA representation and with HiRE-RNA model, for initial and new parameters.

Original value	New value	Parameter
2.608	0.251	Global bonds coefficient
1.483	0.029	Global angles coefficient
1.307	1.440	Global torsions coefficient
3.800	3.688	R4-P bond equilibrium length (Å)
1.430	1.481	O-C bond equilibrium length (Å)
2.1450	2.6086	P-O-C equilibrium angle (rad)
1.7104	1.6046	C-R4-P equilibrium angle (rad)
-0.3491	-0.5351	R4-R1-A1-A2 dihednral phase 1 (rad)
-2.7925	-0.3006	R4-R1-A1-A2 dihedral phase 1 (rad)
0.3491	-0.7906	C-R4-P-O dihedral phase (rad)

Table 3.1: Some parameters of interest, with the original value and the new one, obtained with the free optimization.

much from their initial values, a sign that gradients associated to them are much smaller with respect to the parameters.

Moreover, the obtained parameters are sensible to the values at which they are initialized. To show that, we report in Figure 3.6 the coupling and equilibrium value for P-O-C angle, with different initialisations extracted from a uniform distribution in a 20% range from their reported value, after a 100 epochs training. As one can see, coupling is strongly correlated to its initialisation (with a 0.99 correlation coeffi-

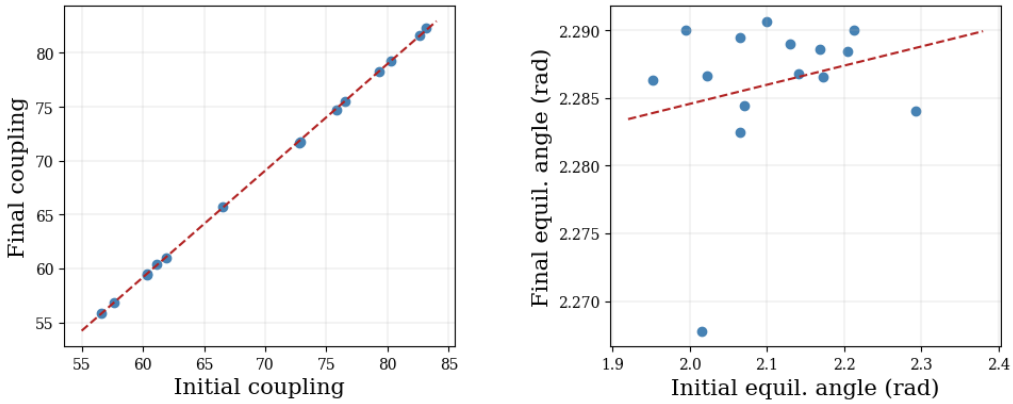


Figure 3.6: P-O-C angle related parameters, for different initialisations extracted from a uniform distribution, before and after the optimization. On the left: coupling parameter, that exhibits a clear linear correlation. On the right: equilibrium angles, whose final value ranges in a small interval and is almost independent of the initialisation.

cient), meaning that, in fact, those parameters are not affected by the optimization procedure. The outcome is different, instead, for the equilibrium angle, whose correlation coefficient with the initial value is 0.24. Indeed, despite variations in the initialisation, it sets mostly in a narrow range of values around 2.285 rad, that is however already distant from the original HiRE-RNA value. Moreover, a comparison with the value reported in Table 3.1, obtained after 400 epochs, shows that this divergence is likely to grow even more along the optimization process. For these reasons, two approaches were tried in order to overcome the correlation between initial and final values and the large variation of geometric parameters with respect to their empirical distributions. Comparisons between the free optimization and the other approaches are shown in the next section.

3.3 Comparison between different methods

The substantial divergence between initial equilibrium values (obtained from Amber distributions) and the ones obtained with the free minimization led us to explore variations of the basic SGD algorithm. In particular, two opposite approaches were tried: first, a preconditioning of gradients (PSGD) was performed in order to drive them to comparable values between each other, since gradients associated to torsions and geometric parameters were generally larger in the free optimization; second, we assigned learning rates with different orders of magnitudes to each group of parameters (corresponding to global parameters, couplings and equilibrium values for bonds, angles, torsions), to guide the minimization towards a minimum that would not change much geometric parameters (we will call this model NG).

It is important to remind that, in the end, we are not interested in finding the global minimum of the loss function, but a minimum that is compatible with the physical meaning of the parameters. Introducing these constraints does not change the structure of parameters space, but influences the convergence of the algorithm towards minima, that led in some attempts to negative values for couplings. In order to prevent it, the initialisation for couplings was also changed, but in a future

implementation of the algorithm it shall be replaced by enriched functional forms. To be more specific, couplings were set to 100 kcal mol^{-1} $Å^{-2}$ for bonds, 10 kcal mol^{-1} rad^{-2} for angles and 10 kcal mol^{-1} for torsions. In addition, in PSGD a learning rate of 10^{-4} was used, while those adopted for NG are 10^{-4} for global parameters, 10^{-3} for couplings and 10^{-6} for geometric parameters. The model was trained for 400 epochs with a minibatch size of 4 elements, and the obtained parameters were collected to perform a relaxation of input structures. The resulting distributions for bond lengths are reported in Figure 3.7. As it can be seen, none of the models is capable to reflect

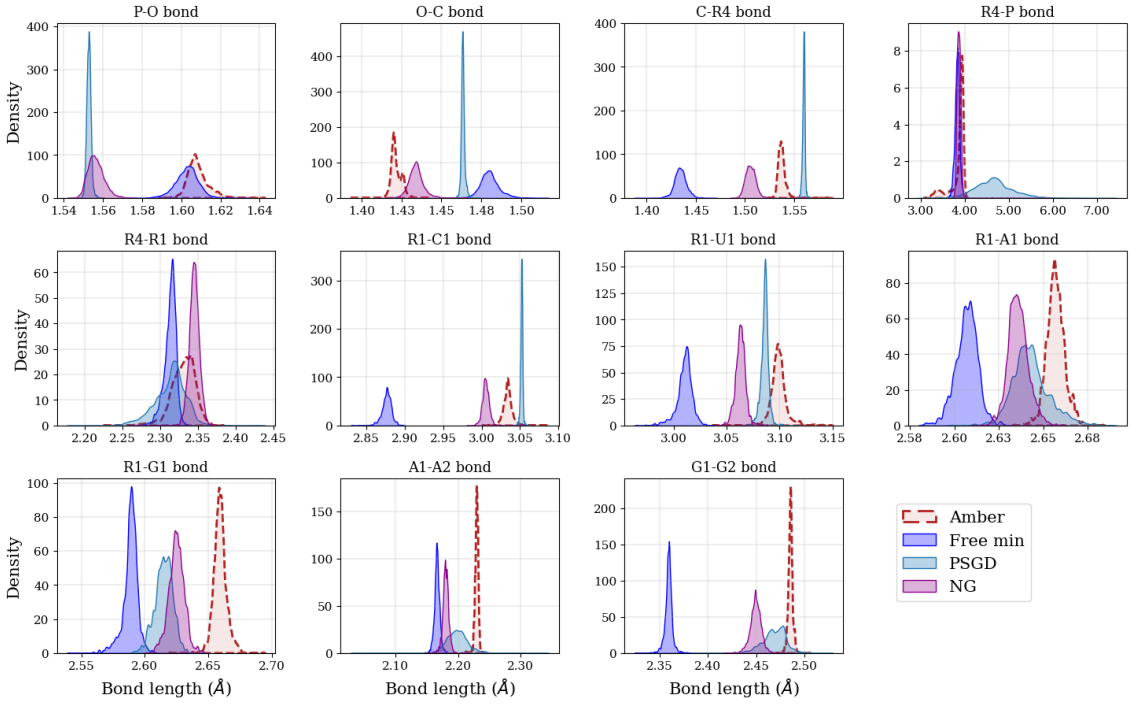


Figure 3.7: Distributions of all bond lengths appearing in HiRE-RNA model for minimized RNA structures, obtained with the different methods adopted. The dashed distribution refers to Amber representation, that corresponds to the expected values. On the x-axis, bonds lengths measured in Ångström are represented.

the expected Amber distribution, even if the free minimization is the one that leads to more divergent values. Results for HiRE-RNA model with initial parameters have not been represented, as they are highly overlapping with those of NG model, despite having different couplings. Indeed, they are closer to Amber distribution, but they do not overlap as much as one would expect, that shows how the geometry of minimized structures is influenced also by the ratio between energy contributions. Moreover, even if the sequences only have 7 nucleotides, the effect of non-bonded interactions can not be completely neglected. A similar behaviour can be observed in angle distributions, that are shown in Figure 3.8. Again, the correspondence between HiRE-RNA and Amber distributions is far from convincing, but some observations can be made. In the first place, NG distributions (and therefore, also those from old HiRE-RNA model) are more close to expected ones, at least in terms of the average value; this is true especially for most of the backbone angles, while worst result are obtained in general for the interactions between sugar and nitrogenous bases (R4-R1-X1). To be more specific, angles involving purines and P-O-C angle are set to very

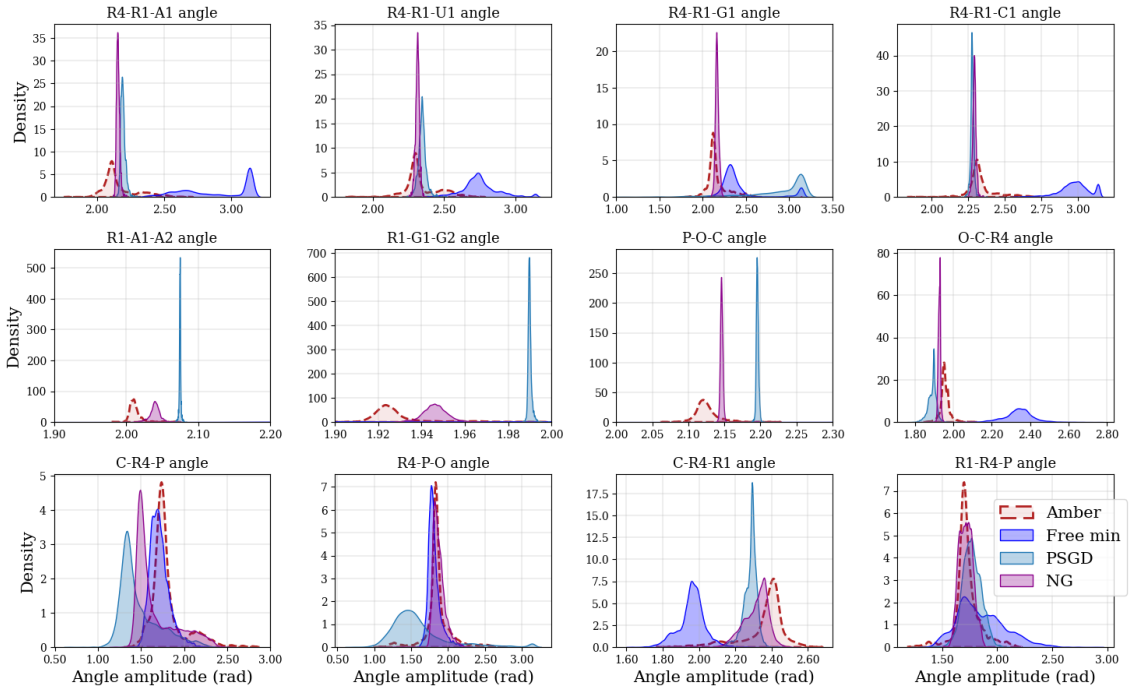


Figure 3.8: Distributions of angles amplitudes (in radians) in minimized structures and comparison between different methods adopted, with Amber expected distributions represented with dashed lines. Worst results are obtained in angles involving beads in nitrogenous bases.

divergent values, to the extent that distributions for the free minimization are not even visible in the plots. Another aspect that can be highlighted is that, commonly, obtained distributions have a variance that is much lower than its expected value, especially in the PSGD model. One possible reason could be the very nature of the optimization process, that in favour of a smaller value for the loss function tends to sharpen the distributions around their mean and eventually reduce the mobility of particles.

The situation becomes even more complex for dihedrals, for which Amber distributions themselves are multivariate due to the multiple equilibrium configurations available. To take it into account, some torsions-related functions in the HiRE-RNA model are obtained as the sum of two or more sinusoidal terms with different weights, which is the reason why some parameters in Table A.1 refer to the same atoms. Moreover, the large contribution that torsions give in terms of energy makes them the most driving of local interactions in determining the structure of the molecule. For their high degeneracy, parameters related to torsions were the most difficult to constrain during the optimization, that is why an initialisation to a different magnitude order was necessary.

We report in Figure 3.9 dihedral distributions obtained from minimized sequences. It is immediately clear that they are in general much broader than those related to bonds and angles, spanning over the whole interval $[-\pi, \pi)$, both in Amber and HiRE-RNA representations. Despite being sometimes quite far from the expected distributions (in particular for purines), it is however interesting to note that the new distributions are able to capture the multivariate nature of the in-

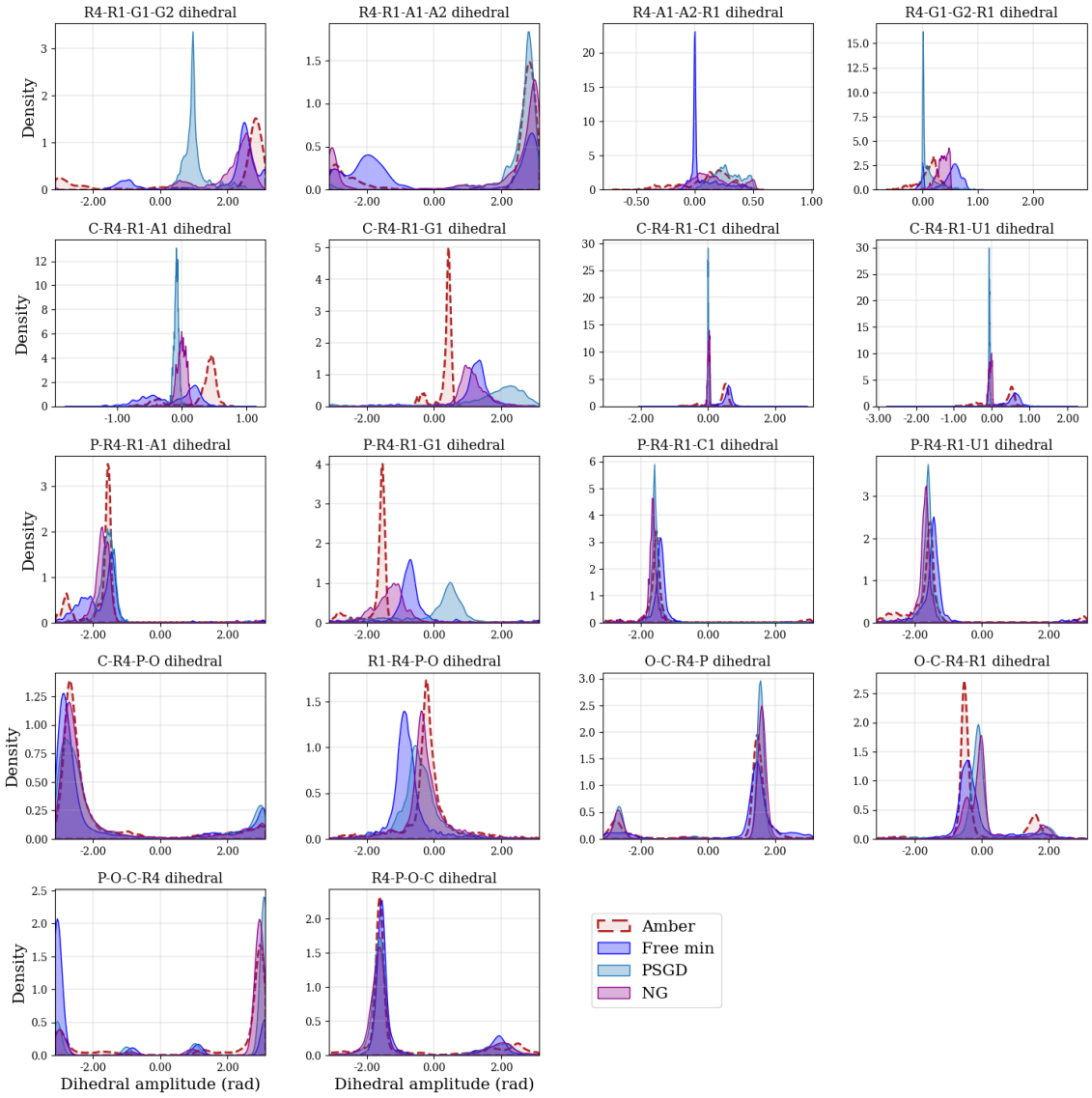


Figure 3.9: Distributions for dihedrals appearing in the HiRE-RNA model, for different optimization methods, and comparison with Amber expected values; dihedrals, obtained from minimized structures, are measured in radians. Unlike bonds and angles, distributions are typically multivariate.

teraction. This is especially true for O-C-R4-P and R4-P-O-C dihedrals, whose distributions are similar for all the models and reflect the expected ones. Like for bonds and angles, also in this case the distributions that closely resembles Amber ones are those of NG (and old HiRE-RNA) model, while distributions obtained with free minimization and PSGD deviate from them the most, in terms of average value and variance respectively.

To quantify the divergence between Amber structures and minimized structures obtained with the analysed models, we computed the root-mean-squared deviation (RMSD). As mentioned in Section 2.2, a value lower than (or close to) 1 Å is typically an indication of well-behaved simulations. RMSD between the obtained structure and the reference one were collected for 1225 input sequences, and their statistical properties are represented in Figure 3.10. The box plot allows to appreciate the

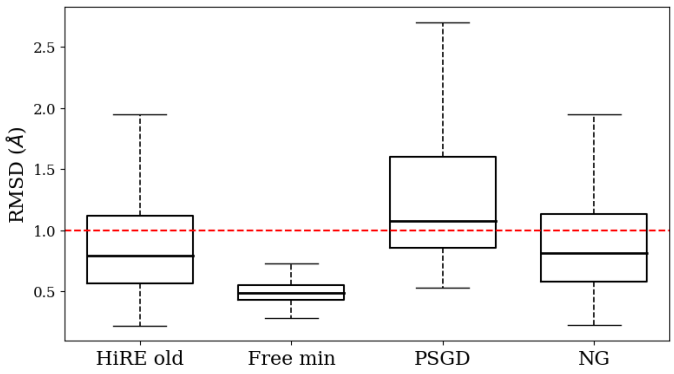


Figure 3.10: Statistical properties of RMSD distributions for the different models considered, represented in a box plot. The thick lines represent medians of distributions, while boxes extend from the lower to the upper quartile value of the data and whiskers represent data range.

median, as well as the extension of the distributions, for all considered models. While PSGD proved to be the worst model, with a median above 1 Å and RMSDs that exceed 2.5 Å, that is also reflected in the high bias and low variance of its geometric distributions, better results are achieved with other models, that produce results compatible with the expected structures. First, is it worth noticing the high similarity between the original HiRE-RNA model and NG, for which couplings were initialised in an uniform way. The correspondence between their RMSD and geometric parameters distributions could be an indication that either fixing the values of the latters effectively drives the optimization towards a model similar to the original one, or that geometric parameters have a much more significant role than couplings in determining the model behaviour, as long as the correct energy scales are reproduced. Rather surprisingly, the best results are obtained with the Free minimization model, for which the RMSD is always below 1 Å. This could be explained, despite the poor matching for bonds and angles distributions, with the better accordance in dihedral values, which are more decisive in the determination of the molecule conformation. Indeed, even if not optimal, the model is able to represent their multimodal distributions and to generally recognise their peaks, and the fact that dihedrals appear in cosine functions, instead of as absolute values, further enhances this similarity.

Other experiments conducted on the model, like the reduction of the degrees of freedom (for instance, fixing particle-dependent global parameters) and the embedding of parameters within specific functional forms that impose constraints on their values, did not improve its predictive power in terms of MD simulations. Quite surprisingly, instead, reducing the number of parameters resulted in a more difficult training and in a stronger compensation effect between the different terms, leading to worse predictions. However, it is important to emphasize that current results and observations are derived only from a partial analysis, that will be subject to further studies in future projects.

4 Conclusions and future work

In this internship report we have described the procedure adopted and the first results obtained in the optimization with machine learning of HiRE-RNA, a coarse-grained force field designed to study RNA folding. Due to the complexity of the model and the high number of parameters involved, we focused only on local interactions, namely covalent bonds, bond angles and dihedral torsions. From HiRE implementation in Fortran, we developed a Python algorithm that reproduces its numerical results for energy computations, and finds the optimal values for the parameters attempting to match HiRE energies with those computed with Amber in an all-atoms representation. In order to do so, a proper training set was created, starting from a limited number of RNA structures extracted from NDB database.

The main difficulty in this project was to find the correct minimum in the optimization process that is compatible with the physical meaning of parameters, and establish a complete and reliable method to define the accuracy of the model, despite the lack of information related to single interactions. In this perspective, in order to evaluate the model performance, we used both machine learning tools and information extracted from molecular dynamics simulation, such as distributions of geometrical quantities and RMSD, and we compared three optimization methods based on different implementations of SGD algorithm, namely free optimization without constraints, PSGD and different learning rates and initialisations (NG). The first results show that a soft enough free optimization leads to parameters that better reproduce atomic simulations in terms of the global configuration of the RNA molecule, even though the distributions of bond lengths and angles substantially diverge.

These results are expected to be improved in upcoming years, following different paths. One of them is to perform other optimization attempts, using more refined labels to find a better matching between the single energy terms, that will possibly lead to a more accurate accordance with MD simulations. The same methods will be applied also to non-bonded interactions, for which we will not pursue a matching with Amber energies, but we will rather impose a minimization of the potential.

In the long term, instead, a more ambitious approach will be the functional optimization of the potential, that will involve both local and non-bonded interactions. This task will be tackled with state of the art deep learning techniques, from Symbolic Regression to Graph Neural Networks, with the aim to further enhance the predictive power of HiRE-RNA model.

A Parameters Tables

Global parameters	
2.608	Bonds
1.483	Angles
2.073	Angles k R4-R1-X1
1.519	Angles k R1-G1/A1-G2/A2
2.355	Angles k P-O-C
4.190	Angles k O-C-R4
4.698	Angles k C-R4-P
4.824	Angles k R4-P-O
5.636	Angles k C-R4-R1
2.130	Angles k R1-R4-P
1.307	Torsions
4.247	Torsion k R4-R1-G1/A1-G2/A2
10.816	Torsion k R4-G1/A1-G2/A2-R1
11.121	Torsion k C-R4-R1-X1
5.819	Torsion k P-R4-R1-X1
0.501	Torsion k C-R4-P-O
0.730	Torsion k R1-R4-P-O
0.331	Torsion k O-C-R4-P
0.257	Torsion k O-C-R4-R1
0.224	Torsion k P-O-C-R4
0.207	Torsion k R4-P-O-C

Angles		
Coupling	Eq. angle (rad)	Atom types
70.000	2.1572	R4-R1-A1
70.000	2.3126	R4-R1-U1
70.000	2.1555	R4-R1-G1
70.000	2.2881	R4-R1-C1
120.000	2.0368	R1-A1-A2
120.000	1.9408	R1-G1-G2
70.000	2.1450	P-O-C
70.000	1.9303	O-C-R4
70.000	1.7104	C-R4-P
50.000	1.9199	R4-P-O
70.000	2.3649	C-R4-R1
100.000	1.7104	R1-R4-P

Bonds		
Coupling	Eq. length (Å)	Atom types
30.000	3.800	R4-P
200.000	2.344	R4-R1
200.000	2.622	R1-G1
200.000	2.633	R1-A1
200.000	3.062	R1-U1
200.000	3.004	R1-C1
200.000	2.450	G1-G2
200.000	2.180	A1-A2
200.000	1.520	C-R4
200.000	1.593	P-O
200.000	1.430	O-C

Dihedrals		
Coupling	Phase (rad)	Atom types
1.000	-0.3491	R4-R1-G1-G2
0.200	-2.7925	R4-R1-G1-G2
1.000	-0.3491	R4-R1-A1-A2
0.200	-2.7925	R4-R1-A1-A2
1.000	-2.8798	R4-A1-A2-R1
1.000	-2.8798	R4-G1-G2-R1
1.000	-2.6180	C-R4-R1-A1
0.200	2.6180	C-R4-R1-A1
1.000	-2.6180	C-R4-R1-G1
1.000	-2.6180	C-R4-R1-C1
0.200	2.6180	C-R4-R1-C1
1.000	-2.6180	C-R4-R1-U1
0.200	2.6180	C-R4-R1-U1
1.000	1.7453	P-R4-R1-A1
1.000	1.7453	P-R4-R1-G1
1.000	1.7453	P-R4-R1-C1
1.000	1.7453	P-R4-R1-U1
1.200	0.3491	C-R4-P-O
1.200	2.7925	R1-R4-P-O
1.000	1.5708	O-C-R4-P
1.000	2.6180	O-C-R4-R1
0.330	0.0000	P-O-C-R4
0.125	3.1416	P-O-C-R4
0.830	0.0000	P-O-C-R4
1.000	0.0000	R4-P-O-C

Table A.1: Initial values of the parameters to optimize, divided in the four groups. For bonds, angles and dihedrals, each line in the Table corresponds to a specific combination of particles. Units of measure are assigned as following, to provide the correct dimensionality in the energy computation: global parameters: adimensional; bond couplings: kcal mol^{-1} $Å^{-2}$; angle couplings: kcal mol^{-1} rad^{-2}; dihedral phases: kcal mol^{-1}.

Bibliography

[1] Tristan Cragnolini, Philippe Derreumaux, and Samuela Pasquali. Ab initio RNA folding. *Journal of Physics: Condensed Matter*, 27(23), June 2015.

[2] Takashi Nagano and Peter Fraser. No-Nonsense Functions for Long Noncoding RNAs. *Cell*, 145:178–81, 04 2011.

[3] D. N. Holcomb and I. Tinoco Jr. Conformation of polyriboadenylic acid: pH and temperature dependence. *Biopolymers*, 3(2):121–133, 1965.

[4] Nozhat Safaee, Anne M. Noronha, Dmitry Rodionov, Guennadi Kozlov, Christopher J. Wilds, George M. Sheldrick, and Kalle Gehring. Structure of the Parallel Duplex of Poly(A) RNA: Evaluation of a 50 Year-Old Prediction. *Angewandte Chemie International Edition*, 52(39):10370–10373, 2013.

[5] Quentin Vicens and Jeffrey S. Kieft. Thoughts on how to think (and talk) about RNA structure. *Proceedings of the National Academy of Sciences*, 119(17), 2022.

[6] Zhichao Miao and Eric Westhof. RNA Structure: Advances and Assessment of 3D Structure Prediction. *Annual Review of Biophysics*, 46(1):483–503, 2017. PMID: 28375730.

[7] Chun Shen Lim and Chris M. Brown. Know Your Enemy: Successful Bioinformatic Approaches to Predict Functional RNA Structures in Viral RNAs. *Frontiers in Microbiology*, 8, 2018.

[8] Samuela Pasquali and Philippe Derreumaux. HiRE-RNA: A High Resolution Coarse-Grained Energy Model for RNA. *The Journal of Physical Chemistry B*, 114(37):11957–11966, 2010. PMID: 20795690.

[9] Tristan Cragnolini, Yoann Laurin, Philippe Derreumaux, and Samuela Pasquali. Coarse-Grained HiRE-RNA Model for ab Initio RNA Folding beyond Simple Molecules, Including Noncanonical and Multiple Base Pairings. *Journal of Chemical Theory and Computation*, 11(7):3510–3522, 2015.

[10] H. M. Berman, Z. Feng, B. Schneider, J. Westbrook, and C. Zardecki. *The Nucleic Acid Database (NDB)*, pages 657–662. Springer Netherlands, Dordrecht, 2001.

[11] Thomas Bäck and Hans-Paul Schwefel. An Overview of Evolutionary Algorithms for Parameter Optimization. *Evol. Comput.*, 1(1):1–23, mar 1993.

[12] Pankaj Mehta, Marin Bukov, Ching-Hao Wang, Alexandre G.R. Day, Clint Richardson, Charles K. Fisher, and David J. Schwab. A high-bias, low-variance introduction to Machine Learning for physicists. *Physics Reports*, 810:1–124, may 2019.

[13] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[14] David Case, Ido Ben-Shalom, S.R. Brozell, D.S. Cerutti, Thomas Cheatham, V.W.D. Cruzeiro, Thomas Darden, Robert Duke, Delaram Ghoreishi, Michael Gilson, H. Gohlke, Andreas Götz, D. Greene, Robert Harris, N. Homeyer, Yandong Huang, Saeed Izadi, Andriy Kovalenko, Tom Kurtzman, and P.A. Kollman. Amber 2022. Technical report, University of California, 2022.

[15] Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 03 2009.

[16] Xi-Lin Li. Online Second Order Methods for Non-Convex Stochastic Optimizations, 2018.

[17] Xi-Lin Li. Preconditioned Stochastic Gradient Descent. *IEEE Transactions on Neural Networks and Learning Systems*, 29(5):1454–1466, may 2018.

[18] Li Yang and Abdallah Shami. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415:295–316, 2020.

[19] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, 2014.

[20] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[21] Fred E. Cohen and Michael J.E. Sternberg. On the prediction of protein structure: The significance of the root-mean-square deviation. *Journal of Molecular Biology*, 138(2):321–333, 1980.

[22] Robert T. McGibbon, Kyle A. Beauchamp, Matthew P. Harrigan, Christoph Klein, Jason M. Swails, Carlos X. Hernández, Christian R. Schwantes, Lee-Ping Wang, Thomas J. Lane, and Vijay S. Pande. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical Journal*, 109(8):1528 – 1532, 2015.