

JÉRÔME RIHON

MOLECULAR MODELING TOOLS TO IMPROVE AND EXPAND  
COMPUTATIONAL RESEARCH ON SYNTHETIC NUCLEIC ACIDS

SEPTEMBER 2024

KU LEUVEN

prof. dr. E. Lescrinier  
prof. dr. V. B. Pinheiro  
prof. dr. M. Froeyen

JÉRÔME RIHON

Dissertation presented in partial  
fulfilment of the requirements  
for the degree of Doctor in  
Pharmaceutical Sciences

KU Leuven  
Group Biomedical Sciences  
Faculty of Pharmaceutical Sciences  
Department Pharmaceutical and Pharmacological Sciences  
Laboratory of Medicinal Chemistry  
Rega Institute for Medical Research

**KU LEUVEN**

DOCTORAL SCHOOL  
BIOMEDICAL SCIENCES

# MOLECULAR MODELING TOOLS TO IMPROVE AND EXPAND COMPUTATIONAL RESEARCH ON SYNTHETIC NUCLEIC ACIDS

Jérôme RIHON

Promotor	Eveline Lescrinier
Co-Promotor	Vitor Bernardes Pinheiro
Co-Promotor	Matheus Froeyen
Chair	Jef Rozenski
Jury member	Kalyan Das
Jury member	Joost Schymkowitz

Dissertation presented in partial  
fulfilment of the requirements  
for the degree of Doctor in  
Pharmaceutical Sciences

September 2024



KU Leuven  
Groep Biomedische Wetenschappen  
Faculteit Farmaceutische Wetenschappen  
Departement Farmaceutische en Farmacologische Wetenschappen  
Laboratorium van Medicinale Chemie  
Rega Instituut voor Medisch Onderzoek

**KU LEUVEN**

DOCTORAL SCHOOL  
BIOMEDICAL SCIENCES

# **MOLECULAIRE MODELING TOOLS TER BEVORDERING EN UITBREIDING VAN COMPUTATIONEEL ONDERZOEK OP SYNTHETISCHE NUCLEÏNEZUREN**

Jérôme RIHON

Promotor	Eveline Lescrinier
Co-Promotor	Vitor Bernardes Pinheiro
Co-Promotor	Matheus Froeyen
Voorzitter	Jef Rozenski
Jurylid	Kalyan Das
Jurylid	Joost Schymkowitz

Proefschrift voorgedragen  
tot het behalen van de  
graad van Doctor in de  
Farmaceutische Wetenschappen



## ACKNOWLEDGEMENTS

  Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequo doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et.

## ABSTRACT

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequo doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et.

## SAMENVATTING

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequa doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et.

## GLOSSARY

  Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequo doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et.

# CONTENTS

ACKNOWLEDGEMENTS . . . . .	V
ABSTRACT . . . . .	VI
SAMENVATTING . . . . .	VII
GLOSSARY . . . . .	VIII
1. General introduction . . . . .	1
1.1. Synthetic Nucleic Acids . . . . .	1
1.1.1. The evolution of chemical biology . . . . .	1
1.1.2. Current affairs on the <i>in silico</i> side of things . . . . .	1
1.1.3. Pharmaceutical impact of synthetic nucleic acids . . . . .	1
1.2. Characterisation of biomolecular structures . . . . .	1
1.2.1. A biased basis of definitions . . . . .	1
1.2.2. Quantitative abstraction of conformations . . . . .	2
1.2.3. Sampling the landscape of relevant molecules . . . . .	2
1.2.4. Setting the curve . . . . .	2
1.3. Fundamentals on the behaviour of molecules . . . . .	2
1.3.1. Quantum Mechanics . . . . .	2
1.3.2. Forcefields : the ultimate interface . . . . .	3
1.3.3. Molecular Mechanics . . . . .	3
1.3.4. Modeling Synthetic Nucleic Acids . . . . .	3
1.4. References . . . . .	4
2. The <i>pucke.rs</i> toolkit to calculation puckers values of biomolecular monomers . . . . .	5
2.1. Introduction . . . . .	5
2.2. Methods . . . . .	6
2.2.1. Generating initial structures . . . . .	6
2.2.2. Quantum Mechanics . . . . .	6
2.2.3. Potential Energy Surface . . . . .	7
2.3. Result & Discussion . . . . .	7
2.3.1. The <i>pucke.rs</i> toolkit . . . . .	7
2.3.2. Options on the construction of a PES . . . . .	10
2.3.3. Analysis of literature data . . . . .	12
2.3.4. Showcase of the peptide, fivering and sixring systems . . . . .	13
2.4. Conclusion . . . . .	14
2.5. References . . . . .	16
3. The Ducque model builder tool to predict synthetic nucleic acid duplexes . . . . .	18
3.1. Introduction . . . . .	18
3.2. Methods . . . . .	20
3.2.1. Characterisation of the nucleosides through Quantum Mechanics . . . . .	20
3.2.2. The Ducque model builder for (X)NA structures . . . . .	20

3.2.3. Extending the XNA library of Ducque	21
3.2.4. Force field parametrisation for molecular mechanics	21
3.2.5. Molecular Mechanics simulations	22
3.3. Results	23
3.3.1. Charge derivation using ORCA	23
3.3.2. RNA and DNA model building by Ducque	23
3.3.3. MD simulations on Ducque's models validated with (X)NA structures in literature	24
3.3.4. Using the proposed workflow to gain insight in RNA::MNA	24
3.4. Discussion	28
3.5. Conclusion	30
3.5.1. Supplementary information	31
3.6. References	32
4. Structure prediction of an HNA aptamer with a G-quadruplex motif	35
4.1. Introduction	35
4.1.1. Welcome to the thesis	35
4.2. Methods	35
4.3. Result & Discussion	35
4.4. Conclusion	35
4.5. References	36
5. Curvature Torsion on Synthetic Nucleic Acids	37
5.1. Introduction	37
5.1.1. Welcome to the thesis	37
5.2. Methods	37
5.3. Result & Discussion	37
5.4. Conclusion	37
5.5. References	38
6. Concluding remarks and future perspective	39
6.1. Concluding remarks	39
6.2. Future perspectives	39
Published works	40
List of Publications	40
Software releases	40
Scientific outreach	40
Declarations	41
Scientific acknowledgements	41
Personal contributions	41
Conflict of Interest	41

Welcome to the first chapter of this doctoral thesis!

The contents of this chapter are, in parts, adaptations of the original manuscripts and supplementary materials of **Chapter 2** and **Chapter 3**.

## 1.1. Synthetic Nucleic Acids

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequo doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distingue possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et.

### 1.1.1. The evolution of chemical biology

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequo doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut.

### 1.1.2. Current affairs on the *in silico* side of things

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequo doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut.

### 1.1.3. Pharmaceutical impact of synthetic nucleic acids

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequo doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut.

## 1.2. Characterisation of biomolecular structures

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequo doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distingue possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et.

### 1.2.1. A biased basis of definitions

↪ The Cambridge Convention of 1988

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequo doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut.

↪ CurvesS and 3DNA

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat.

### 1.2.2. Quantitative abstraction of conformations

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequo doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut.

↪ peptides

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do.

↪ fiverings

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do.

↪ sixrings

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do.

↪ n-membered rings

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do.

### 1.2.3. Sampling the landscape of relevant molecules

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequo doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut.

### 1.2.4. Setting the curve

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequo doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut.

↪ Curvature ( $\kappa$ ) and Torsion ( $\tau$ )

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat.

## 1.3. Fundamentals on the behaviour of molecules

### 1.3.1. Quantum Mechanics

↪ Mathematical lore

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat.

↪ Semi-empirical methods

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat.

#### ↪ *Ab Initio* methods

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat.

### 1.3.2. Forcefields : the ultimate interface

#### ↪ Point charge derivation

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat.

#### ↪ Fitting (non-)bonded term parameters

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat.

### 1.3.3. Molecular Mechanics

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequo doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut.

#### ↪ Applications of Molecular Dynamics

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat.

### 1.3.4. Modeling Synthetic Nucleic Acids

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequo doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut.

" My invasion of  
Thoughts is your  
Caviar "

~ Dance Gavin Dance - Caviar ~

## 1.4. References

# THE *PUCKERS* TOOLKIT TO CALCULATION PUCKERS VALUES OF BIOMOLECULAR MONOMERS.

2

Adapted from the following manuscript *xxx/aaa/xxxxx*:

Rihon J., Reynders S., Pinheiro V.B., Lescrinier E. "The *puckers* toolkit to facilitate sampling the conformational space of biomolecular monomers." Journal of Cheminformatics (J Cheminf).

## 2.1. Introduction

One of the most prominent traits in defining biomolecular function and regulation are conformational changes to the monomers that make up large proteins and nucleic acid structures.<sup>[1]</sup> It remains challenging to observe these changes directly at the atomic level due to their dynamic nature. Nonetheless, computer simulations have provided an increasingly realistic picture on the properties of studied molecules. Molecular Dynamics (MD) simulations are oftentimes used to understand their behaviour by allowing molecules to visit one or more conformational states when freely interacting with their environment, or by imposing them with a restraintive bias to force less favourable states to be presented as well during the simulation.<sup>[2]</sup> This behaviour is defined by a classical forcefield, whose parameters are typically obtained from research on particular moieties of the studied molecule, by either fitting to experimental NMR data or using *ab initio* calculations with Quantum Mechanics (QM).<sup>[3, 4, 5]</sup> The latter field allows us to study virtual fragments through computational approaches and eventually predict the effect on the conformational behaviour of larger molecules that are composed of such fragments.<sup>[6, 7]</sup> Traditional force fields, used in the analysis of biopolymer structures and their interactions, contain the quantitative descriptors of these monomer building blocks.<sup>[8]</sup>

In the field of synthetic nucleic acid (NA) research, this can guide the selection of new constructs prior to their synthesis in the lab. The development of Xenobiotic Nucleic Acids (XNAs) advanced the field towards being applicable as viable therapeutics.<sup>[9]</sup> Chemical modifications in the original DNA structure prolonged the biological half-life of oligonucleotides (ONs) to a level suited for clinical applications and interactions to the target were optimised to increase potency and selectivity.<sup>[10]</sup> Modifying the backbone and the nucleobase expanded the field of XNA research.<sup>[11, 12, 13]</sup> To date, progress was made by trial-and-error approaches of chemical synthesis and evaluating the viability of a multitude of different XNA constructs. Molecular modeling can exploit *in silico* results and steer the selection of next generation nucleic acid therapeutics. Inducing this endeavor requires a clearcut tool needed to facilitate obtaining a free energy landscape for ring puckering of new nucleotides that can be used to derive forcefield parameters, which are exploited for MD simulations on larger constructs.

In this work, we present the *puckers* toolkit to simplify the Conformational Sampling (CS) methodology, by constructing the foundation of the energy landscape for specific molecular systems. We used the *puckers* commandline tool and QM to fully characterise free energy landscapes by quantifying their potential wells and transitions between different states of the five-membered ring in the Adenosine nucleoside, of the six-membered ring in an 1,5-anhydrohexitol NA (HNA) nucleoside and of the backbone in an Alanine amino acid, using standard off-the-shelf computational hardware. The tool itself produces a set of constraints for all possible conformations of a molecular type, which are applied in geometry optimisation (GO) procedures, in order to produce a set of optimised conformers. The potential energy of the optimised conformers are calculated to obtain the energy levels of datapoints on the different landscapes (Single Point Evaluation, SPE). By using the *puck.py* module, the user can apply various formalisms (Cremer-Pople<sup>[14, 15]</sup>, Altona-

Sundaralingam[16] or Strauss-Pickett[17]) to monomer constructs or to the respective monomers in a nucleic acid structure. This module can also generate the 3D structure of the corresponding five- or six-membered ring, based on queried puckering coordinates, proving useful to recreate molecules in reported literature. The DNA Adenosine nucleoside was subjected to the CS methodology, exploiting a variety of different levels of theory in computational chemistry. The optimal procedure was selected to obtain high quality results in an acceptable timeframe by comparison of performances from different approaches on a local machine. We provide an in-depth look at puckering data from established literature and finish with a showcase on the various conformational landscapes of relevant biomolecules.

## 2.2. Methods

### 2.2.1. Generating initial structures

The CS methodology consists of the three parts, where it starts with the generation of torsion angle constraints to produce initial conformations, which cover the full conformational landscape. The *pucke.rs* CLI-tool, as well as the *pucke.py* module, provide simple query functions to generate the desired axes to generate structures in order to sample the conformational landscape. Using either tools requires the user to specify the type of molecular system (peptide, fivering or sixring) to be called. Both the peptide and fivering system employ a *linear space* function, which asks the user to pass an amount of points to be computed for in a preset range. E.g., for fivering systems, the inclusive range, [-60. → 60.], can be queried for an interval of 21, which returns a set of points at an interval of 6°. For each generated conformer, *pucke.rs* generates a set of ( $\nu_1, \nu_3$ )-constraints (Fig. ConfSamplingExplained B.). Additional applied constraints, during the CS experiment, are  $\beta$ : 208.5°,  $\gamma$ : 30.9°,  $\varepsilon$ : 159.1° and  $\chi$ : 260.6°. The Cremer-Pople sphere represents the conformational space of sixrings. By populating its surface with evenly distributed points, we achieve a well-sampled landscape. Querying an amount of points, which translates to the set of conformations for the sampling itself, requires a slight approximation as not all queried amounts can result in an even coverage of the surface of the sphere. These points are then converted into the Strauss-Pickett impromper dihedrals ( $\alpha_1, \alpha_2, \alpha_3$ ) and are used as constraints for the CS methodology (Fig. ConfSamplingExplainedC.). The Hexitol NA Adenosine (hA) monomer sampling was imposed with the following constraints:  $\beta$ : 180.1°,  $\gamma$ : 60.0°,  $\varepsilon$ : 180.1° and  $\chi$ : 210.59°. The peptide landscape was queried to produce the axes at an interval of 10° to generate 1369 distinct sets of ( $\varphi, \psi$ )-constraints, for the inclusive range of [0. → 360.]. The L-Alanine (Me-NH-Ala-CO-Me) was used to sample the peptide space (Fig. ConfSamplingExplained A.) and did not require additional constraints. The axes were later transformed to [-180. → 180] to visually compare energetically allowed regions for the peptide backbone, with the popular Ramachandran plot[18]. Queries of *pucke.rs* and *pucke.py* are given in Figure SICodeExamples A-B.

### 2.2.2. Quantum Mechanics

Each initial structure is subjected to a constrained Geometry Optimisation (GO) and a Single Point Evaluation (SPE) through QM approaches. Here the DNA Adenosine nucleoside (dA) is used for benchmarking since its Potential Energy Surface (PES) is well described [6, 7]. This study utilises ORCA v5.0.4[19], as explicitly this version has the latest correction on the D4 dampening[20, 21], courtesy of Grimme lab. The Gold Standard Quality (GSQ) was decided to be the Möller-Plesset 2<sup>nd</sup> order perturbation theory (MP2) [22], as the CI-CCSD(T) LoT does not lend itself well to geometry optimisations and still remains a demanding LoT, though incredibly accurate. This level of theory (LoT) is accompanied by the 6-311++G (2df,2p) basis[23, 24], with the Resolution of Identity (RI) approximation.[25] The def2-QZVPP/C auxiliary basisset[26] is used for the RI approximation of the MP2 density, together with the def2/JK for approximation on Coulombic and Exchange integrals[27], hereafter MP2<sup>Q</sup>. The same basis set and approximations are used for the CS methodology at the *ab initio* Hartree-Fock (HF) level (HF<sup>Q</sup>).

The semi-empirical HF-3c LoT[28, 29, 30] is employed as it is fast and cheap and because it has been used in the accelerated methodology[7]. The double hybrid functional PBE0 [31], with the D4 dampening, uses the same basis as MP2<sup>Q</sup> but with the def2/J auxiliary set[32] (PBE0<sup>Q</sup>). Their usage is reasoned in that both their quality in GO computations have performed extremely well for the fraction of a cost of the pure *ab initio* LoTs. All GOs are performed with the *VeryTightOpt* keyword.

Other methods are the MP2 def2-TZVP/C (MP2<sup>T</sup>) [26] and the HF<sup>Q</sup> without RIJK approximation (HF<sup>Q-RIJK</sup>). These are used to compare their *consumables* and quality of results within the LoT with the other variants.

The DNA Adenosine, which counts 31 atoms in total, uses 103 basis functions for HF-3c and 742 basis functions for the calculations with the PBE0<sup>Q</sup>, the HF<sup>Q</sup> and the MP2<sup>Q</sup> LoT respectively. The benchmarking consists of comparing different LoTs' resources and assessing their RAM usage, wallclock time and Disk Space usage and the *consumables* used during computations, and comparing this to the GSQ. Comparison of the GO quality will be done by going through a pairwise structure comparison, where differences are measured with the Kabsch RMSD algorithm[33] ([github.com/charnley/rmsd](https://github.com/charnley/rmsd)). All sets of optimised geometries will also be subjected to an SPE with all four functionals HF-3c, PBE0<sup>Q</sup>, HF<sup>Q</sup>, MP2<sup>Q</sup>. The MAXCORE keyword is utilised to max out at 1500 MiB per thread engaged. For the GO part, every optimisation allocates six threads per conformation. A total of ten conformations, at one time, can be concurrently optimised. For the SPE, every evaluation allocates one thread per conformation with a total of 35 threads active at one time. Calculations were performed on a Ryzen ThreadRipper 3970 (32 cores / 64 threads) with a RAM capacity of 64 GiB.

### 2.2.3. Potential Energy Surface

From a landscape of *in silico* generated and evaluated conformers of a molecular type, a PES is generated, as described by Mattelaer *et al.* [7]. The PES itself is expressed as the relative difference in energy ( $\Delta E$ ) of all conformations with respect to the global minimum of the landscape. To evaluate the produced PESs, the differences in relative energy( $\Delta\Delta E$ ) are compared, for all combinations of LoTs, against the GSQ. This to evaluate which combinations (GO-SPE) can produce qualitative PESs (Fig. PES\_ALL).

Figures are made with Matplotlib and Cartopy. Figure SISphericalConvention details on the conventions used to define the CP six-membered ring space in relations to the mathematical convention of defining latitude-longitude coordinates, for graphical purposes. Cartopy, a Python geography library superset of Matplotlib, was exploited to project the PES of six-membered ring systems onto the surface of the sphere. This was done with the Mollweide projection, and transforming the data by the PlateCarree projection. Because  $Q$  tends to stabilise around 0.67 for biologically relevant puckering modes, the CP coordinates ( $Q, \varphi_2, \theta$ ) were simplified to 2D to better graphically visualise the CP sphere, by neglecting the amplitude.

For the RMSD contourplots, the *oslo* colourscheme was used [34], other colourschemes were custom maps.

## 2.3. Result & Discussion

### 2.3.1. The *pucke.rs* toolkit

#### ↪ Methodology of *pucke.rs*

A visual accompaggniment to the methodology is provided below (Figure 2-1). For endocyclic systems, puckering formalisms are exploited, as they neatly abstract the conformation of an N-membered ring system to a set of coordinates. For the fivering system, the methodology has been applied from Huang *et al.*'s [6] way of combining the Altona-Sundaralingam (AS) and Sato formalism, projected on a Cartesian system.

$$\begin{aligned}\nu_1 &= \left( Z_x \cos\left(\frac{4\pi}{5}\right) + Z_y \sin\left(\frac{4\pi}{5}\right) \right) \\ \nu_3 &= \left( Z_x \cos\left(\frac{4\pi}{5}\right) - Z_y \sin\left(\frac{4\pi}{5}\right) \right)\end{aligned}\quad 1-1$$

By iterating over a set of  $Z_x$  and  $Z_y$  values, ranging from  $[-60. \rightarrow 60.]$ , one calculates a set of  $(\nu_1, \nu_3)$  endocyclic torsion angles per gridpoint. Equation 1-1 simply rearranges the terms from Huang *et al.* to return the pair of endocyclic torsion angles [6]. Both the peptide and fivering methods return a 2D grid, returning sets of proper dihedrals to be used as constraints for GO procedures (Figure 2-1 A,B.).

Sampling the pyranose space exploits two puckering formalisms. Through the use of the Cremer-Pople (CP) formalism, one can calculate a set of local elevations from a spherical coordinate  $(Q, \theta, \varphi)$ , which is an abstraction of a six-membered ring conformation. It has been theoretically detailed by Cremer [15] and applied by Sega *et al.* [35], to reverse engineer (or invert) the puckering coordinates to a full conformation. Starting from an equidistributed globe [36], the coordinates are passed into the function to calculate the set of local elevations per conformation. Based on assumptions on the magnitude of the bondlengths and -angles, the atoms are assigned a position in  $\mathbb{R}^3$ . Afterwards, the improper dihedrals  $(\alpha_1, \alpha_2, \alpha_3)$  (Strauss-Pickett formalism; SP) are computed for and are used as constraints. The sphere represents the CP sphere; the six-membered ring's conformational extent (Figure 2-1 C.). The amplitude is kept as a constant at  $Q = 0.67$ , as this is the value at which biologically relevant six-membered rings exist [37]. Equation 1-2 defines all ringsystems with an even amount of atoms.

$$z_j = \sqrt{\frac{2}{N}} q_m \cos\left(\varphi_m + \left(2\pi m \frac{j-1}{N}\right)\right) + \frac{1}{\sqrt{N}} q_{m+1} (-1^{j-1}) \quad 1-2$$

For sixring systems ( $N = 6 \rightarrow m = 2$ ), the equation is simplified.

$$z_j = \sqrt{\frac{1}{3}} q_2 \cos\left(\varphi_2 + \left(2\pi \frac{j-1}{3}\right)\right) + \frac{1}{\sqrt{6}} q_3 (-1^{j-1}) \quad 1-3$$

By virtue of the fact that Equation 1-3 can be assigned as :

$$q_2 = Q \sin(\theta) \quad | \quad q_3 = Q \cos(\theta) \quad | \quad \varphi = \varphi_2 \quad 1-4$$

Which results in Equation 1-5, that is used to perform the actual computation for the set of local elevations in a sixring system (iterating over  $j = 0 \rightarrow 5$ ) in the software. The generated spherical coordinates  $(Q, \theta, \varphi)$  are passed to this function.

$$z_j = \left[ \sqrt{\frac{1}{3}} \sin(\theta) \cos\left(\varphi + \left(\frac{2\pi j}{3}\right)\right) + \frac{1}{\sqrt{6}} \cos(\theta) (-1^j) \right] Q \quad 1-5$$

Continuing the research on the accelerated methodology [7], the conformational landscapes are generated by querying the *pucke.rs* tool for a specific system Figure 2-1. The constraints of the peptide-like landscapes are simply procured by iterating over the two axes in a nested fashion. The values gathered from iterating over these axes are directly used as the constraints of the particular dihedrals, here  $(\varphi, \psi)$ . This landscape supplies the user with constraints, which are then imposed on the respective dihedrals of the molecule;  $(\varphi, \psi)$  for peptides,  $(\nu_1, \nu_3)$  for fivings and  $(\alpha_1, \alpha_2, \alpha_3)$  for sixrings. These constraints, together with the axes of the grids, are produced by the *pucke.rs* toolkit. The constraints are applied during GO procedures, which generate all the possible conformations that molecule's system can adopt, given that the query amply covers the landscape. The accelerated methodology (Mattelaer *et al.*) employs two levels of theory to dramatically speed up the duration of the experiment, by applying a cheap one for the GO stage (HF-3c) and an expensive

one for the SPE stage (MP2<sup>Q</sup>) to accurately sample the behaviour of the molecules subjected to this study.[7] All this comes together to generate a PES that represents the behaviour of the studied molecule.

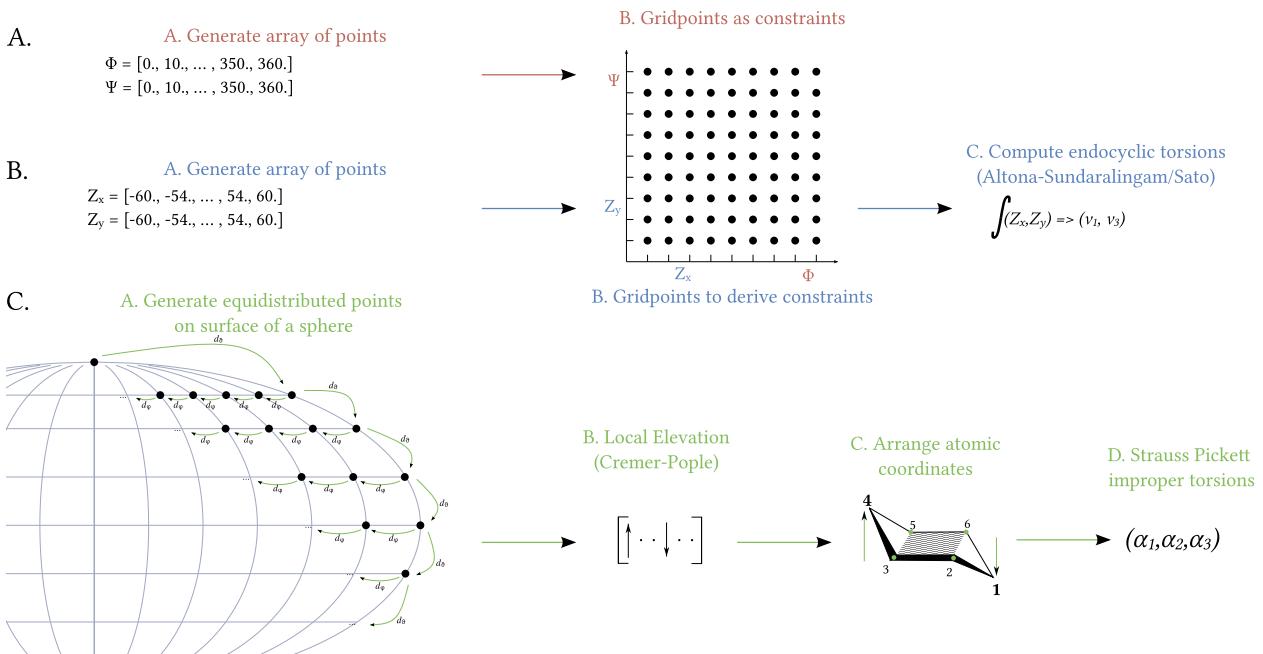


Figure 2-1: Explanation of conformational sampling experiments. **A.** Sampling over peptide-like systems, akin to the concept of a Ramachandran-plot. **B.** Sampling the fivering space, employing the Altona-Sundaralingam/Sato formalism (Eq. eq:huang\_furanose) [6]. **C.** Sampling over the sixring space, reverse engineering spherical coordinates to local elevation. Local elevation is returned through Eq. SoftwareEquation, after which the atomic positions are computed for and the SP improper dihedrals are calculated. See Figure SICode-Examples A-B. for code examples.

### ↪ The *pucke.py* module

In order to make the methodology more user-friendly, we provide both a CLI tool and a scripting library (Python-wrapped Rust) to allow users to implement either option into their own workflows. The Rust language was a deliberate choice to ensure the robustness of the toolkit. In the Python module, the CLI-tool has been bound to the *conformation sampling* module. Additionally, the *pucke.py* module contains the *formalism* module that allows the user to calculate various puckering formalisms for five- (AS, CP5) and six-membered (SP, CP6) rings. This has been used to calculate for the formalisms in Figure 2-6. Furthermore, the user can pass specific Cremer-Pople coordinates to the CP5( $r, \varphi_2$ ) or CP6( $r, \varphi_2, \theta$ ) class and invert these parameters to produce the conformation they have defined, as an xyz- or pdb-formatted file. This feature allows users to explore and understand the intricacies of the different formalisms, as well as derive specific constraints by the queried conformer (Fig. SICodeExamples C.,E.). When researching puckering formalisms and working on characterising biomolecular monomers computationally, this library will prove immensely valuable.

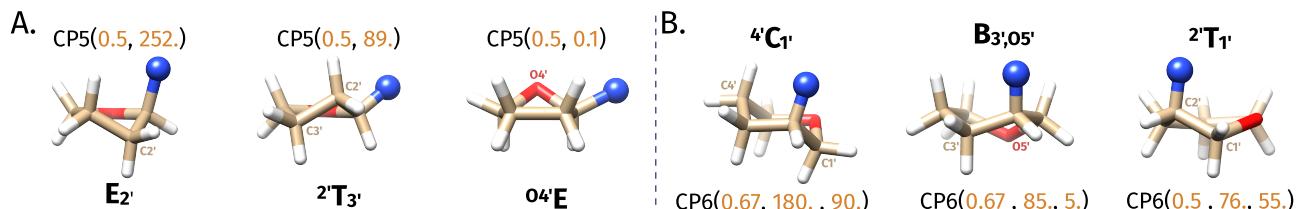


Figure 2-2: Examples of conformations produced by querying for specific Cremer-Pople coordinates, for five- and six-membered ring systems. This employs the inversion algorithm, detailed by Cremer [15]. The library produces just the base ring, hydrogens and nitrogen (representing the nucleobase) were added with UCSF Chimera [38]. **A.** Various conformations of the DNA nucleoside, produced by inverting CP-coordinates for five-membered rings (CP5). **B.** Various conformations of the Hexitol NA nucleoside, produced by inverting CP-coordinates for six-membered rings (CP6).

To complete the library, the *geometry* module is supplemented with three functions to calculate molecular geometries; the bondlength, bondangle and dihedral. These works by passing in coordinates from parsed

molecule files. The library provides classes to manipulate pdb and xyz coordinate files, which parse the coordinates of the molecule in question (Fig. SICodeExamples D.). To provide freedom to the user, the functions of the *geometry* module allows arguments to be passed from any type of object, as long as its type satisfies being a 3D coordinate; *type: [float, float, float]*. Various code examples on the *pucker.py* module are given in Figure SICodeExamples C-E.

### 2.3.2. Options on the construction of a PES

#### → Benchmarking on a local machine

Here, we apply the set of levels of theory [HF-3c, PBE0<sup>Q</sup>, HF<sup>Q</sup> and MP2<sup>Q</sup>] for the geometry optimisation procedure on the DNA Adenosine nucleoside. Every respective set of optimised structures is then also subjected to potential energy evaluations by the four LoTs respectively. This generates sixteen different landscapes at various levels of accuracy to describe the behaviour of the DNA nucleoside. To reiterate, the MP2<sup>Q</sup> is represented as the GSQ and its result is located in the bottom right corner of Figure 2-3. This figure is to be looked at in synergy with Figure 2-4 and the Consumables (Suppl. Fig. SIConsumables).

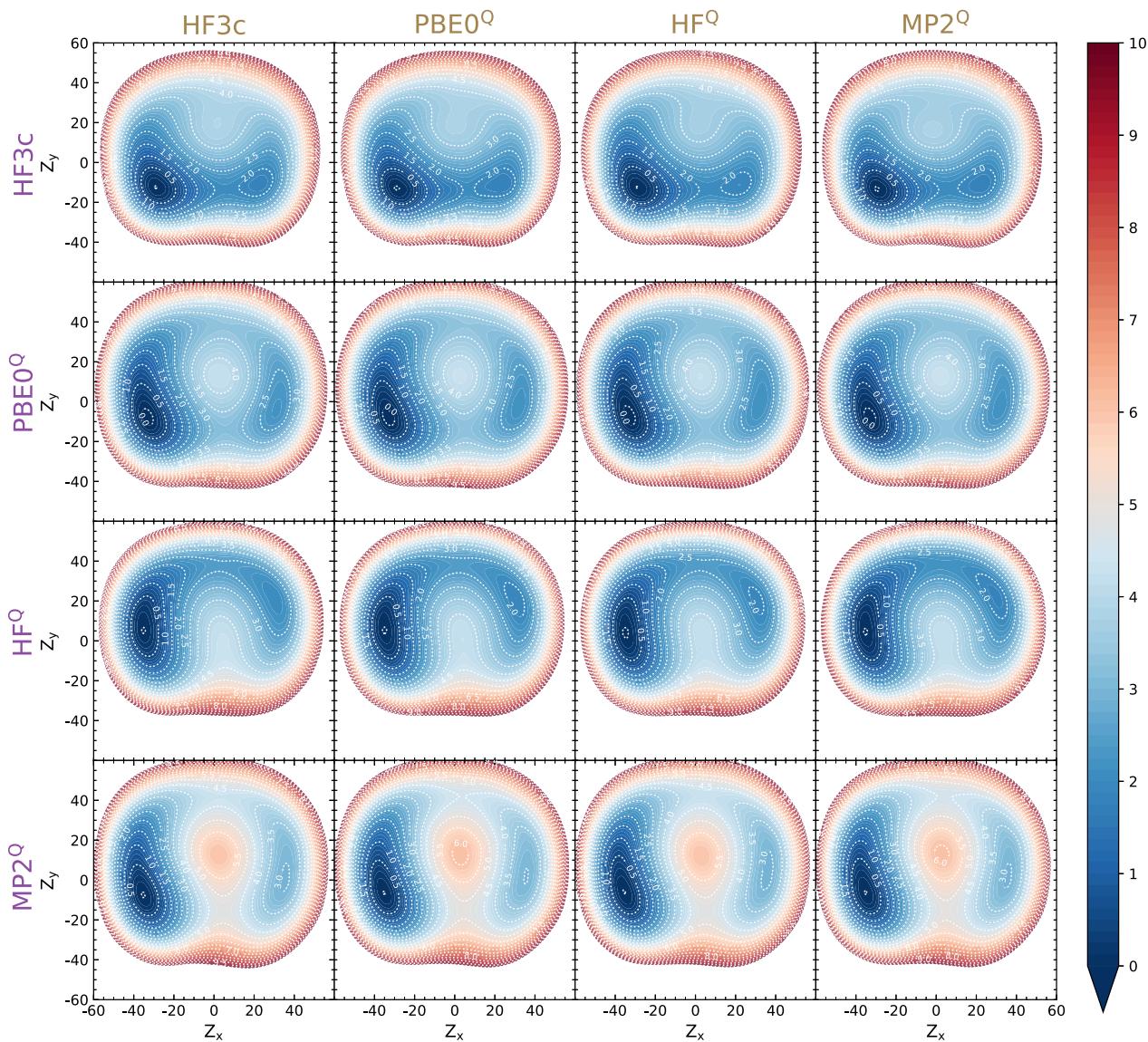


Figure 2-3: Conformational sampling of the DNA Adenosine nucleoside, carried out at the various levels of theory, combining different (GO - SPE) resulting in different Potential Energy Surfaces. Geometry optimisations (brown; columns) were carried out at the HF3c, PBE0, HF and MP2 level. All GOs were subjected to an additional Single Point Evaluation (mauve; rows), respectively the three other functionals. A 4x4 matrix shows the result of all the samplings, visualizing the respective combinations of GO and SPE. To clarify, combinations involving the same functional did not require an additional SPE; shown on the diagonal of the matrix.

The MP2<sup>Q</sup> are the heaviest computations of them all, clocking in at roughly 548h or about 22.8 days of calculations. The geometry optimisation capped out at about 48 GiB of RAM, while at most 40 GiB of *tmp*-files were stored on disk by ORCA when ten conformations were optimised concurrently. The HF-3c was logged for the same parameters and finished in about 0.7h, capping at almost 3GiB of RAM and almost 1 GiB in Disk Space in *tmp*-files. The GO experiment with HF<sup>Q</sup> finished around the 30h mark and showed little hardware consumption compared to the GSQ, topping at 10 GiB of RAM with an excess of 6 GiB of *tmp*-files produced by ORCA at most. The PBE0<sup>Q</sup> consumes relatively the same as the HF<sup>Q</sup>, but clocks in at 58h or 2.4 days (Suppl. Fig. SIConsumables A.,B.).

LoT	HF-3c	PBE0	HF	MP2
HF-3c	0.67	58.08	30.13	548.01
PBE0	5.12	57.97	34.53	552.41
HF	5.12	62.48	30.02	552.36
MP2	6.42	63.43	35.73	547.91

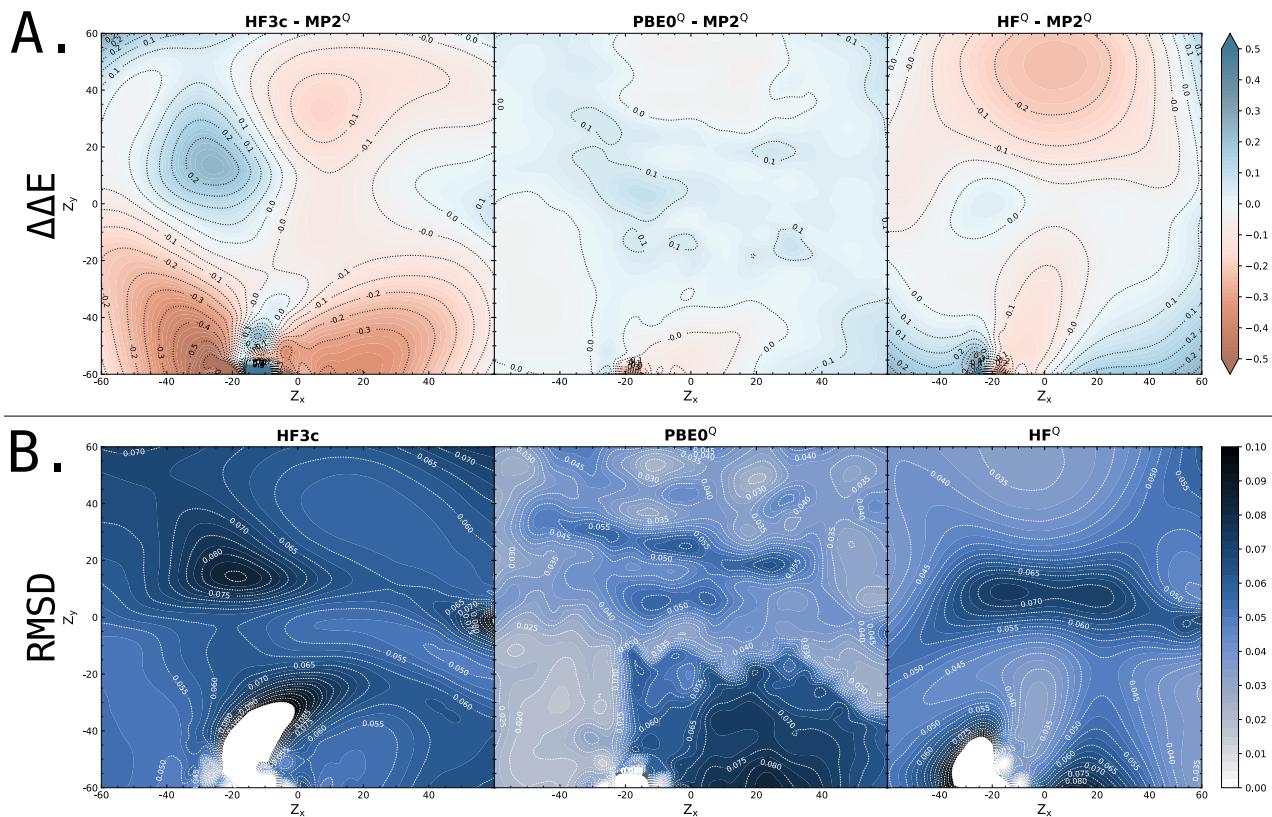
Figure 2-4: Wallclock Time to completion, with reference to the generated PESs from Figure 2-3. All times expressed in hour (h). To clarify, the diagonal of the table highlights the wallclock time of only the GO method for that level of theory (LoT); an SPE with the same LoT is redundant.

When looking at the SPE calculations, we see that all the conformers in the landscape can be evaluated in less than 6h, irregardless of the LoT. What we also see is that the HF<sup>Q</sup> (max. 28.9 GiB) and MP2<sup>Q</sup> (max. 28.6 GiB) have similar RAM requirements, while the latter still required a tremendous amount of free disk space to store *tmp*-files. All this in contrast to the PBE0<sup>Q</sup> that needs (max. 9 GiB) of additional space on disk to run succesfully. The SPE at the HF-3c finished so quickly that only a couple of measurements could have been taken (Suppl. Fig. SIConsumables C.,D.). The conclusion remains that the (HF-3c - MP2<sup>Q</sup>) combination still is a sturdy contender in approximating the GSQ, as these calculations are perfectly manageable within a single workday, on hardware with 64 GiB of RAM and 32 cores. The SPE calculations at the PBE0<sup>Q</sup> level are a strong candidate for cases where hardware systems are limited as the landscapes evaluated at this level resembles the shape of the MP2<sup>Q</sup> SPE the most. Still, the established methodology [7] returns a comparatively amazing result for the frugality of its expenditure. To combat the storing of *tmp*-files, one allocates more RAM per thread. In the interest of the research, a concession had to be made towards the cheaper LoTs , by not exaggerating the resources allocated with respect to their computational requirements.\ A comparison between (MP2<sup>Q</sup> vs. MP2<sup>T</sup>) and (HF<sup>Q</sup> vs. HF<sup>Q-RHK</sup>) is presented in Suppl. Fig. SIConsumablesExtra.

#### → Zooming in on the best picks

To evaluate the quality of the PESs respectively, a closer look to the differences of the best results is required. An outcome of these CS experiments is to derive parameters for Molecular Dynamics simulations or understand the molecule's behaviour to apply constraints in NMR structural determination experiments. Therefor, the approximation needs to be as close as we can get it to the GSQ. We compared the PESs of the bottom row (Figure 2-3) with the GSQ. A first analysis is done by applying the difference in relative energy to that of the GSQ ( $\Delta\Delta E$ ) ( A.). To look at the quality of the optimised conformations, we assessed every conformation of a landscape in a pairwise fashion to the conformer with the same puckering coordinates in the GSQ and applied an RMSD algorithm [33], to calculate the difference in optimised structures ( B.). Overall, the PBE0<sup>Q</sup> scores the best on both facets of the analysis. We highlight that the respective ranges in which we evaluate both the  $\Delta\Delta E$  and the RMSD are small, indicating that the differences overall are minute. Nevertheless, the PBE0<sup>Q</sup> best approximates the MP2<sup>Q</sup> for a fraction of its cost and optimises conformations closest in resemblance to that of the GSQ in parts of the landscape where it matters (e.g. minima and saddle points). Its only downside is the time investment into the PBE0<sup>Q</sup> calculation, with respect to HF-3c, as the PBE0<sup>Q</sup> still takes

at least two days to run for such a landscape, while HF-3c barely takes an hour, when all optimisations are ran concurrently.



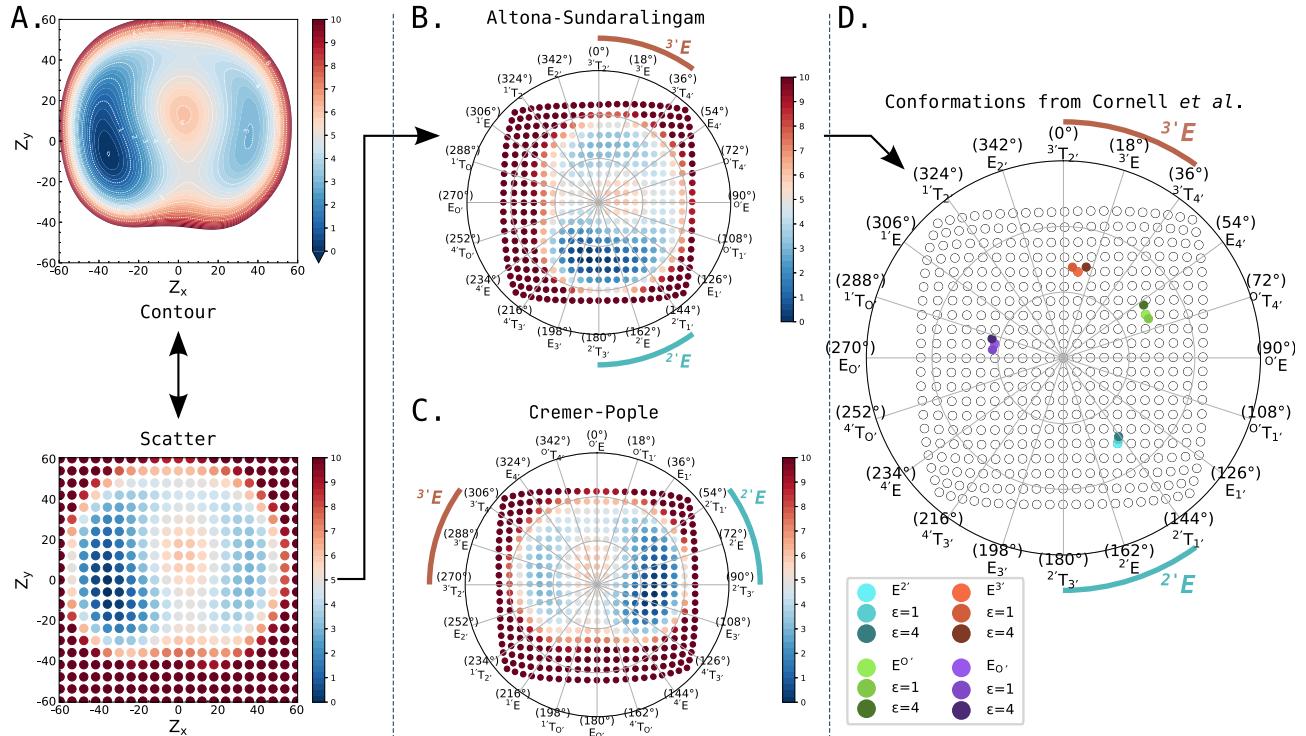
**Figure 2-5:** **A.** Comparison of the difference in relative energy ( $\Delta E$ ) between the GSQ ( $MP2^0$ ) CS experiments and the other GO sampling who've been subjected to an SPE at the  $MP2^0$  level. Ranges from  $[-0.5 \rightarrow 0.5]$   $\frac{kcal}{mol}$  **B.** Comparison of RMSD between the GSQ ( $MP2^0$ ) GO procedures of the experiment and the other GO sampling. Ranges from  $[0. \rightarrow 0.10]$  Å.

### 2.3.3. Analysis of literature data

To introduce the Cremer-Pople formalism, an apex (first atom in the set) needs to be assigned for the set of local elevations ( $z_j$ ). The original paper [14] conveniently chose to have the apex to go through the oxygen atom when characterising sugars, which results in  $\varphi_2 = 0^\circ$  returning an  $^0E$  conformation. While comparing this data with what is parametrised in the paper of Cornell *et al.* [8] (AMBER Forcefield parameters), it was noticed that while the CP formalism was assigned to denote the puckering modes of the conformers, that the reported pucker coordinates do not follow the same apex. More so, in that article, the apex seems to follow through the C3', since the  $^3E$  is closest to  $0^\circ$ , but the phase angle has also been shifted by  $18^\circ$  ( $\frac{\pi}{10}$ ), causing the  $^3T_2$  to appear at the top of the plot. This, however, does exactly aligns with the AS formalism [16]. The Cornell paper mistakenly attributes it to the Cremer-Pople formalism itself. The original CP-paper [14] mentions the AS formalism, in its introduction section, which could be the reason why the CP paper was cited.

Figure 2-6 D depicts the reported conformers used from the Cornell paper [8] and adjacently the nearest pucker coordinate from the previous CS experiment (Figure 2-3, MP2°), as the AS formalism. Of note is the broad interpretation of the *Envelope* ranges, as they actually lean closer into *Twist*-territory (Fig Inverted-Cornell). The [ $^2E$ ,  $^3E$ ,  $^0E$ ,  $E_O$ ] set of conformers are coupled to a relative energy value. From the GSQ  $\Delta E$  values, these are (0.00, 1.66, 2.93, 4.65). In the Cornell paper, these are respectively at [ $\varepsilon = 1 : (0.00, 0.63, 2.87, 5.86)$ ] and [ $\varepsilon = 4 : (0.00, 1.04, 1.86, 5.68)$ ]. Any differences are attributed to the basis set used (6-31G\*) and the constraints at which the geometry optimisation were performed. We see the same trend of favourability in potential energy of the conformations in all three sets of results. The parametrised conformers also fall into

place with the local and global minimum and the transitional states, showing the predictive quality of the CS methodology on the behaviour of these monomers.



**Figure 2-6:** **A.** MP2<sup>0</sup>-optimised PES, shown as contour plot (upper), in contrast with the scatter plot (lower) that was used to interpolate the data from. **B.** Altona-Sundaralingam representation on a polar coordinate plot of the scatterplot in (A). **C.** Cremer-Pople representation on a polar coordinate plot of the scatterplot in (A). **D.** From Cornell et al. [8], the deoxyribose adenosine conformations highlighted along their denoted conformation, according to the manuscript itself. For (<sup>2</sup>E in turquoise, E<sub>O'</sub> in green, <sup>O</sup>E in lila and <sup>3</sup>E) in orange. The conformation, as a result of the CS experiment itself (Figure 2-3), nearest to the respective conformations from the Cornell paper are also highlighted as the base colour. The conformations are highlighted in Figure 2-2, by making use of the inversion methods in pucker.py.

### 2.3.4. Showcase of the peptide, fivering and sixring systems

To finalise, we provide a brief showcase of the applicability of the CS methodology, by providing the reader with examples on the peptide, fivering and sixring experiments. These molecular systems encompass all chemical variants of biological monomers used in standard and synthetic biology. The Peptide PES (Figure 2-7) shows a global minimum around the same region as where we would find *alpha*-helical conformers on a standard Ramachandran plot. Up from this global well, we see one local maximum and an adjacent local minimum ( $\psi = \pm 135^\circ$ ). This region constitutes where *beta*-sheet conformers are situated. On the righthand half of the PES, diagonally up from the *alpha*-helical conformers, we encounter the *alpha<sub>L</sub>*-helical conformers, whom are involved in left-handed helical protein structures. This behaviour tends to be true for all natural amino acids.[39, 18] As detailed in Sec. CrempopVsAltsund, for the puckering behaviour of the furanose in DNA (Figure 2-7), we see a global minimum around the <sup>2</sup>E area, which corresponds with standard DNA::DNA homoduplex configurations. At the local minimum, we find the <sup>3</sup>E conformer, which is often adopted under conditions when hybridising with different types of backbone chemistries. The PES depicts two transition states, or commonly referred to as saddlepoints. The upper saddlepoint (5, 40) locates the <sup>O</sup>E conformation, while the lower saddlepoint at around (5, -15) depicts the E<sub>O'</sub>. Again, this predicted behaviour of the DNA nucleoside falls in line with structural determination data. Finally, the HNA chemistry has been subjected to a sampling (Figure 2-7). At the North Pole, we find the typical <sup>X</sup>C<sub>W</sub> conformers. At its antipode, the South Pole, we find their inverse conformers <sup>W</sup>C<sub>X</sub>. As At the equator, we find various Boats (<sup>X,W</sup>B, B<sub>X,W</sub>) and Skews (<sup>X</sup>S<sub>Z</sub>, <sup>Z</sup>S<sub>X</sub>) configurations. To clarify, a halfboat conformer is the same as a skew, where three consecutive atoms

are in the same plane, with the fourth in-plane atom located in between two out-of-plane atoms and both latter atoms are on either side of the plane.

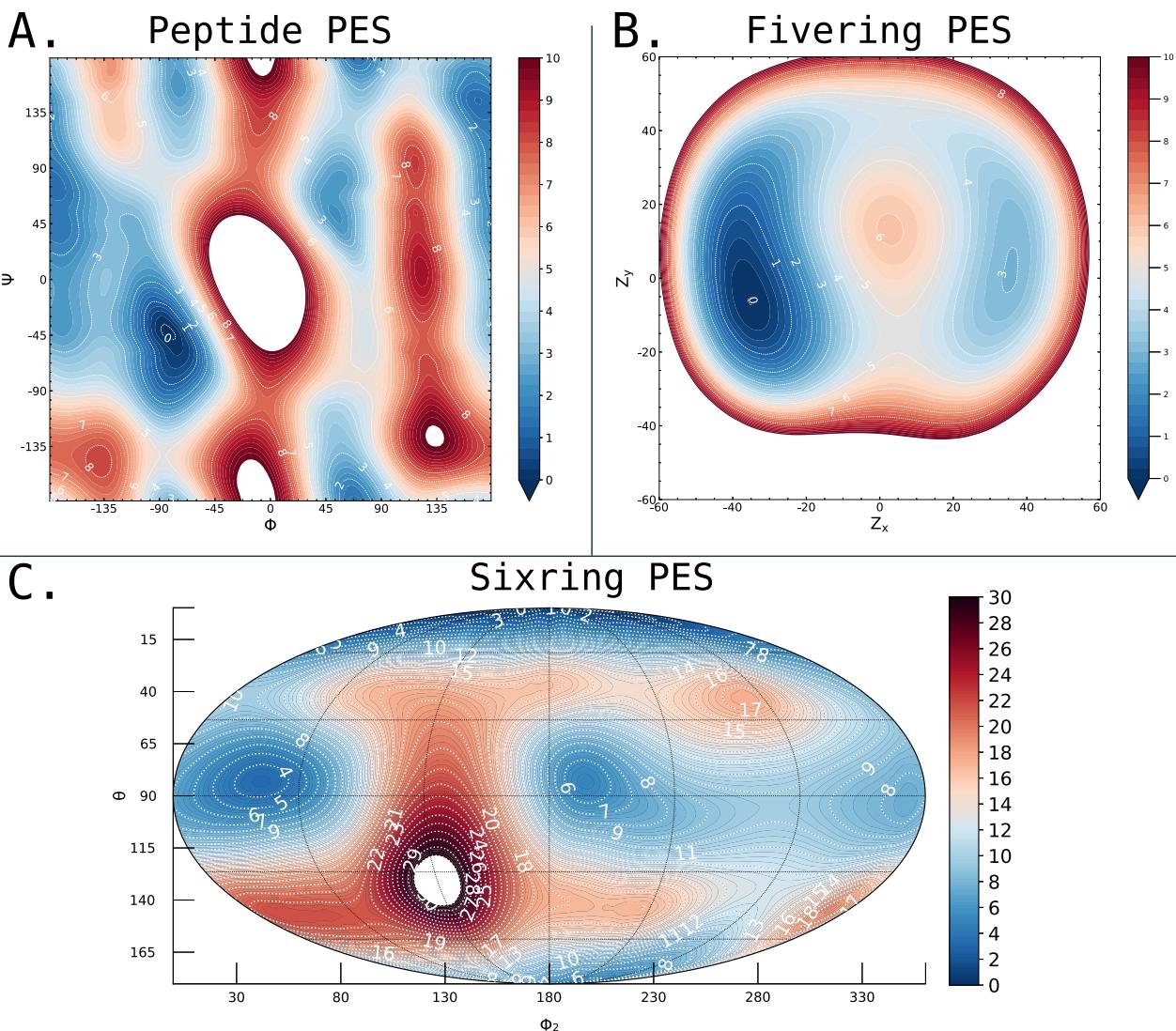


Figure 2-7: Potential energy surface of (A.) peptide-like systems, (B.) five-membered ring systems and (C.) six-membered ring systems. Explanation of the geographical vs. CP formalism on the axes ranges involved in Suppl. Fig. SISphericalConvention.

Around a latitude of  $55^\circ$ , we encounter the Envelope ( $^xE$ ,  $E_y$ ) and Twist ( $^xT_y$ ) puckering modes, while at  $135^\circ$  we see their inverse puckering modes ( $^yE$ ,  $E_x$ ;  $^yT_x$ ). Twists are often referred to as halfchairs and define a conformation where only two consecutive atoms both exist out-of-plane, one on either side of the plane. From this PES, we can safely assess the stability of both North and South pole conformations, we three large local minima around the equator. Viewing the minima from left to right, it boasts the  ${}^3S_1$ , the  $B_{O,3}$  and the  $O{}^1S_2$ , respectively.

## 2.4. Conclusion

The *puckers* toolkit proves useful to generate the set of constraints and axes required for performing Conformational Sampling experiments.

We were able to visualise how different levels of theory in computational chemistry perform, both in terms of qualitative output and by logging their *consumables*, allowing the user to make informed decisions for their own experiments and what their hardware allows them to do. Accurate characterisation of molecules should not be impaired by the specifications of one's machine. While this work based itself on the accelerated methodology of Mattelaer *et al.* [7], the goal was to explore and expand the possibilities a researcher can use

for this type of experiments. The information returned by such methodologies can be used to make informed decisions on constraintive procedures in Molecular Dynamics simulations, to firmly decide on restraintive measures in NMR-related deterministic structure elucidation studies and even derive a suitable forcefield for the XNA in question. This can all be achieved from purely *in silico*-based research.

The free and open-source toolkit allows for a pragmatic approach to these experiments, by simplifying the workflow to what is actually desired; the sampling, defining and documenting of the configurational space of the biomolecular monomer of interest. The *inversion* method for the Cremer-Pople formalism is particularly useful to recreate conformations from literature, to reproduce and understand puckering behaviour.

The [CLI-tool](#) and the [Python library](#) are available on GitHub, where their documentation can be found as well. The user can build the CLI from source by using the Rust toolchain, or install the library directly from pip or conda.

" Master has given Dobby a sock!  
Master has presented Dobby with clothes!  
Dobby is free! "  
~ *Dobby - Harry Potter and the Chamber of Secrets* ~



- [38] Pettersen. E., Goddard. T., Huang. C., Couch. G., Greenblatt. D., Meng. E., Ferrin. T., "UCSF Chimera—A visualization system for exploratory research and analysis" *Journal of Computational Chemistry*, vol. 25, pp. 1605, 2004-07, [doi:10.1002/jcc.20084](https://doi.org/10.1002/jcc.20084)
- [39] Hollingsworth. S., Karplus. P., "A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins" *BioMolecular Concepts*, vol. 1, pp. 271, 2010-10, [doi:10.1515/bmc.2010.022](https://doi.org/10.1515/bmc.2010.022)

# THE DUCQUE MODEL BUILDER TOOL TO PREDICT SYNTHETIC NUCLEIC ACID DUPLEXES

3

Adapted from the following manuscript [10.1093/nar/gkae135](https://doi.org/10.1093/nar/gkae135):

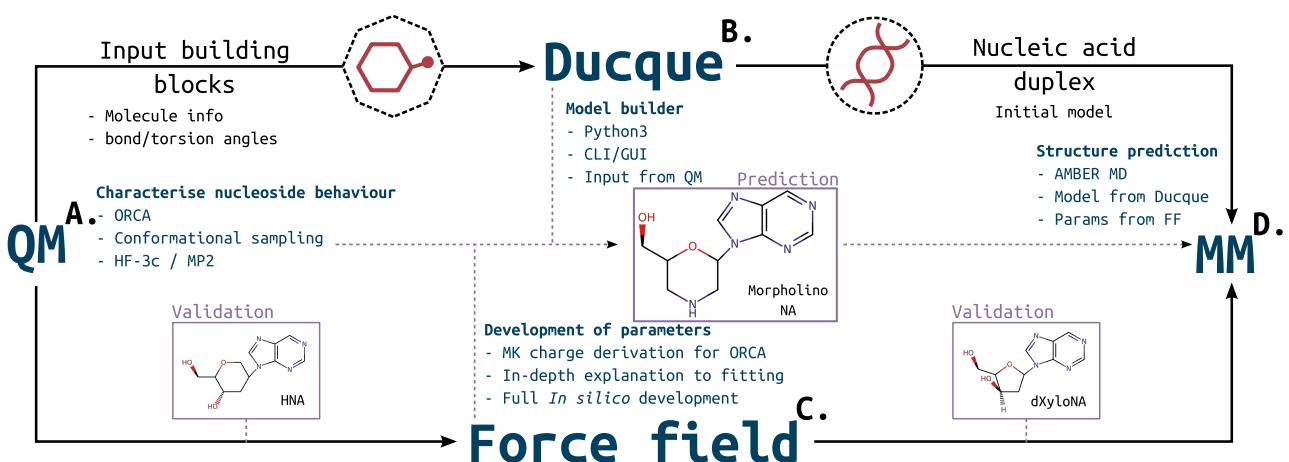
Rihon J., Mattelaer C.-A., Montalvão R.W., Froeyen M., Pinheiro V.B., Lescrinier E. “*Structural insights into the Morpholino Nucleic Acid/RNA duplex using the new XNA builder Ducque in a molecular modeling pipeline.*” Nucleic Acid Research (NAR).

## 3.1. Introduction

The potential of nucleic acid therapeutics was recognized more than 50 years ago, but nucleases that hydrolyse phosphodiester linkages in oligonucleotide (ON) strands represented one of the hurdles that significantly slowed down their rise in the therapeutic field [1]. Analogues carrying modifications in the backbone and nucleobase (xenobiotic nucleic acids, XNAs [2]) accelerated development of nucleic acids therapeutics in the mid-1980s, as they allowed to modulate catalytic degradation of ONs and/or increase affinity to the selected target. Several XNA chemistries were used for various clinical applications, of which some have reached the market [3]. While XNA were initially designed to modulate RNA activity upon hybridisation, some XNA alterations have also been developed for applications such as therapeutic aptamers and an alternative genetic system in synthetic biology. This has lead to the engineering of enzymes able to process such ON strands, in order to promote XNA ON manipulation methodologies to a higher level of efficiency, for both researcher and cell [4, 5].

Conformational preorganization of sugar moieties in the XNA backbone is a key feature for their binding capacity to native nucleotides. Chemical modifications such as pyranose units [6] and locked nucleic acids (LNA) [7] freeze the sugar moiety in the nucleic acid backbone and offer the ultimate control of conformational preorganization and fix their recognition potential for either DNA or RNA. The xylose-based chemistries, on the other hand, avoid hybridisation with native nucleotides [8]. Nucleobase alterations can provide hybridisation stability, which benefits duplex formation, or extends binding potential, broadening the range of possible aptamer-target interactions. Recently, these XNA have been successfully employed in CRISPR-Cas9 research, using the <sup>th</sup>G modification [9].

The morpholino nucleic acids (MNAs) are a prime example of highly modified nucleic acids in antisense oligonucleotide (ASO) research and therapies [10, 11]. Their presumed mechanism of action is to sterically block RNA splicing or translation, by binding to their RNA complement. Preclinical studies in zebrafish have popularised the use of such gene-knockdown mechanisms with MNA ONs [12]. MNA are commonly used in the treatment for Duchenne’s Muscular Dystrophy, where binding to the specific exon-51 sequence leads to trimming of the dystrophin protein. In short, the binding of an ASO to its target sequence results in cancelling the propagation of that exon-51 RNA sequence into a peptide strand, thereby halting the translation into a full-fledged protein. This is in contrast with the endogenous RNA degradation exploited by other XNA chemistries for ASO and silencing RNA (siRNA) purposes [13, 14, 15] catalysed by the RNase H enzyme or RISC complex, that break RNA strands upon hybridisation with complementary DNA and RNA strands respectively. In such a mechanism, the XNA::RNA duplex must be recognised by the processing machinery through similarity to the natural substrate [16]. Experimental structures of a range of XNA chemistries in homoduplexes and bound to a complementary nucleic acid sequence proved useful in understanding their biophysical and biological properties [5].



**Figure 3-1:** **A.** QM approach to perform conformational sampling to describe the torsional behaviour. **B.** The Ducque model builder receives input from the computed PES to use curated conformers as building block to generate duplex models. **C.** From the PES, one can derive a force field by virtue of the predicted behaviour. A charge derivation scheme has been implemented for ORCA. Curated conformers are to be used to derive torsional parameters for the force field. **D.** Combining the products of (B., an (X)NA molecular model) and (C., a representative force field), we can predict the molecular structure of an XNA model duplex through the use of an MM package, which can be used for antisense design, (X)NA enzyme engineering and more. This project workflow has been validated on HNA and dXyloNA and has been applied to predict the RNA::MNA heteroduplex.

Molecular modeling of XNA::RNA duplexes can predict possible XNA applications and guide the design of XNA and XNA processing enzymes [17]. Model building nucleic acids has been facilitated in the mid '90s thanks to the release of the Nucleic Acid Builder (NAB) software and its domain-specific language [18]. This was followed with many implementations of the NAB language in a variety of wrappers and servers to allow model building for DNA and RNA. The proto-Nucleic Acid Builder has taken a small step towards XNA building by including a set of modifications known to build into mainly A-type structures [19]. It uses parameters from experimentally derived structures, using the convention of 3DNA [20] of basis reference frames, which differ slightly from the definition of Curves+ [21]. It also makes a good attempt at minimising duplex models with force fields (FFs) that were developed for small molecule minimisation and evaluation. The claimed predictions simply build along a set of given vectors from data gathered from wet lab experiments. Consequently, such building schemes are not applicable for many unsolved and underresearched chemistries structures.

Because of the amount of new chemistries developed for many different applications [22, 23], an urgent need arised to design a platform that can handle the influx of existing XNAs and modifications that have yet to be developed. With this in mind, we developed the Ducque software from scratch, independent from the approach of the traditional Nucleic Acid Builder. It comes with a user costumisable library for linker and sugar moieties, reserving functionality for purine- and pyrimidine-like bases, to build XNA models with a virtually unlimited set of chemistries on the backbone, with a future outlook on customising nucleobase modifications. This approach easily allows to build all models exemplified in Anosova *et al.* [24], regardless of complexity.

The Ducque model builder is part of an elaborate workflow to implement new chemistries for accurate molecular dynamics (MD) simulations (Figure 3-1). The model builder also boasts a neat graphic user interface (GUI) that makes the flow of importing new building blocks, randomising and building structures smooth and concise, by allowing access to the various modules in a chained fashion. A potential energy surface (PES) describes the nucleoside's behavior along its puckering profile. This data feeds directly into Ducque and the creation of a suitable FF, respectively. Additionally, an AMBER-specific implementation of the parametrisation scheme was developed for the ORCA package. While Ducque is used to build an initial model, the force field is essential for the final Molecular Mechanics (MM) stage that provides an accurate structure. Here, we

demonstrate Ducque against natural nucleic acids using NAB as our benchmark and show its functionality across a range of well-characterised XNAs. The workflow was validated on Hexitol Nucleic Acid (HNA) and deoxy-xylose nucleic acid (dXyNA), and challenged on the RNA::MNA duplex, a heteroduplex with clinical relevance for which no experimental structure is available to date.

## 3.2. Methods

### 3.2.1. Characterisation of the nucleosides through Quantum Mechanics

The conformational sampling method has been previously reported and applied on DNA, RNA, HNA and dXyNA, in Mattelaer *et al.* [25]. For MNA, the Cremer-Pople (CP) coordinates can be reversed engineered to Strauss-Pickett (SP) improper dihedrals [sugarpucker](#), by generating equidistributed coordinate points on the surface of a sphere [26]. The generated spherical coordinates ( $r, \theta, \varphi$ ) are converted to local elevation ( $z_j$ ) from a mean plane per respective atom [27], by assuming bond lengths and angles, which positions the atoms in the ring in cartesian space. From the atomic coordinates, the improper dihedrals ( $\alpha_1, \alpha_2, \alpha_3$ ) can be calculated for [28]. Graphical representations of these formalisms have been added (Figure CPMeanplane, SPIImproper). These improper dihedrals are used as restraints for the ORCA software (v. 5.0.2) [29]. Several backbone torsion angles were restrained as follows :  $\beta$  (180.1°),  $\gamma$  (60.1°),  $\varepsilon$  (180.1°),  $\chi$  (193.9°) [25]. First, a lightweight semiempirical HF-3c (Hartree Fock) geometry optimisation is performed to relax the respective conformations after imposing the restraints. Afterwards, a single point energy (SPE) evaluation, through Møller-Plesset 2nd order perturbation theory (MP2) is performed to evaluate the potential of the respective conformers. The specific accompanying basis set for MP2 is 6-311++G(2df,2p) with the def2-qzvpp/C basis as an auxiliary basis set for the RI approximation and the def2/JK auxiliary basis set for the RI approximation of Coulomb and exchange integrals in the HF step [25]. A total of 630 conformers were sampled for each morpholino nucleoside. The geography library, *cartopy* (< python3.6), is employed to make a 2D projection of the  $\mathbb{R}^3$  surface of the conformational sphere of HNA and MNA from which we sampled (Figure spherical\_-convention). The Mollweide projection is used, together with a PlateCarree transformation. The SciPy Radial basis function (Rbf) is used to interpolate the values to generate the PES. QM calculations were performed on a Ryzen Threadripper 3970, with a memory capacity of 32GB RAM. The standard MNA has been altered to a methyl-capped O6' and N3' variant before sampling the conformations, for reasons detailed in the **Results and Discussion**.

The linker moiety, N,N-dimethyl-O-methylaminophosphoroamidate, has been subjected to a conformational sampling following the same methodology, but instead of varying endocyclic torsion angles, the ( $\zeta, \alpha$ ) dihedral dyad was varied from 0° to 360° in steps of 10° along both axes (Figure PESLinker), for a total of 1369 conformers.

### 3.2.2. The Ducque model builder for (X)NA structures

Ducque is written in Python3 and uses only built-in modules as well as NumPy-based modules (NumPy, SciPy) [30]. Ducque is functional both in the Commandline Interface (CLI) and with a GUI (Tkinter). It is free and open-source software ([jrihon/Ducque](#)), under the MIT license. Ducque uses a mechanical building method that functions solely on the given nucleoside conformers and backbone parameters of the respective nucleotides it builds into a strand, meaning it only takes into account the input bond angles and dihedrals. Array rotation in  $\mathbb{R}^3$  is supported by quaternion mathematics.

Ducque's core functionality is extrapolation of the backbone dihedral by requiring only torsion and angle values. To position an atom in  $\mathbb{R}^3$ , there is only one location that satisfies both the given angle and dihedral. With this process, and an associated fitting of the following linker and nucleoside, the leading strand is built (Figure DihedralExtrapol). To build to complementary sequence, a fitting of the complementary strand's nu-

cleobase to the leading strand's nucleobase is performed, by posing the complementary nucleobase onto the plane of its basepair and shifting it into the desired hydrogen bonding position (Figure BuildComp). An added fitting of the backbone with neighbouring residues resolves most, if not all, clashes. Any remaining clashes are easily resolved through minimisation in the next stage of MD, with the correct force field.

### 3.2.3. Extending the XNA library of Ducque

Ducque allows the user to implement a custom repository of new chemistries, which function as building blocks for the virtual duplexes. This module requires the additional information on the input angles and the *pdb* format nucleoside itself. Furthermore, one can generate randomised sequences through the randomisation module, or manually customise sequences, giving the user the freedom of choice when designing a duplex, e.g. multiple chemistries. Lastly, the *xyz* format files, generated during Conformational Sampling by ORCA, can be converted to a *pdb* format by Ducque. This implies batch conversion through Ducque as well.

For the extension of the Ducque library with new inputs, a *pdb* inputfile of the new nucleoside must be provided by the user. It is converted to the *json* format (--transmute), which requires parameters like residue name, backbone angles and dihedrals to generate a suitable *json* inputfile. The *xyz* format files, generated during conformational sampling to characterise nucleosides (ORCA output), can be converted to *pdb* format files by Ducque (--xyz\_pdb) and subsequently used to supply to the user library. One can generate strands with randomised sequences for a specific XNA chemistry through the randomisation module (--randomise), or manually customise sequences. The latter option allows to design a nucleobase sequence as well as the nucleic acid backbone on a residue specific level, giving freedom of choice for duplex design with multiple chemistries in the (customised) library. The randomisation functionality was used to generate a multitude of different models for DNA::DNA, DNA::RNA, RNA::RNA, dXyNA::dXyNA, RNA::HNA and finally RNA::MNA. A sequence was specified manually to generate double-Drew Dickerson dodecamer (DDD) structures. Nucleic acid duplex models were built (--build) from the randomised/DDD sequences and were subjected to MD.

The respective nucleobases were flattened using the HF-3c level geometry optimisation, with the atoms in and on the sugar restrained in cartesian space and the math.chi -  $\chi$  dihedral restrained at its current position. The improper dihedral for purines (C1', C8, N9, C4) and pyrimidines (C1', C6, N1, C2) is also restrained to  $-179.5^\circ$  math.degree to ensure a plane as flat as the standard nucleobases provided with NAB to satisfy the planarity assumptions in the rotations. For benchmarking the build quality of the structures, python scripting (for NAB) and Ducque's randomisation module were used to generate ten RNA and DNA homoduplexes of a randomised sequence of 12 basepairs (bp), respectively. These were compared to duplexes generated by the NAB language (Figure BenchmarkNA), Table BenchmarkNA). The *time* shell command (logging the *real* output) was used to compare build times for chronometric benchmarking. NAB was logged for combined time of compiling from the Domain-specific Language (DSL) to the binary and executing the binary, while Ducque can only be logged for Python's runtime. The structures were evaluated based on RMSD and difference in interbasepair-parameters using Curves+ [31].

### 3.2.4. Force field parametrisation for molecular mechanics

The AMBER FF [32, 33] is compatible with restraint electrostatic potential (RESP) charges [34]. This is a subset of molecular electrostatic potential (MEP) charges sampled with the Merz-Singh-Kollman (MK) population analysis scheme [35, 36], followed by a two-stage restraint procedure to derive atomic charges (RESP fitting). The nucleosides were superposed on the X-Z cartesian plane by the (C1', C5', N3') atoms in the nucleosides through an in-house scripted reorientation [37].

The MK population analysis scheme has been implemented, by the authors, for ORCA and detailed here below. Geometry optimisation of the molecules was done at the PBE0 level, which was followed by an SPE at the HF/6-31G\* level/basisset, with ORCA, to generate the correct molecular density orbitals for RESP [38].

The gridpoints are generated through the MSMS software [39] at a range of factors times (1.4, 1.6, 1.8 and 2.0) the van der Waals (vdW) radius, defined by the Connolly surface algorithm [40]. The probe radius was set to 1.4 Å, the density of the points on the surface was set to  $\frac{3}{\text{\AA}^2}$  (Figure MKPopAnalScheme). MSMS employs a further triangulation of the vertices to distribute them more equally. ORCA's *orca\_vpot* program is used to map the orbitals and their densities to the respective gridpoints, in atomic units (a.u.). Afterwards, the ESP-loaded grid is provided to the RESP script for a two-stage restraintive procedure, deriving point charges for the atoms. Similar atoms are equivalenced and internal fragments are restraint to a set charge for the respective nucleosides, which ultimately returns the needed RESP charges for the FF [34, 38, 41].

This uses the Chirliam-Franci least square fitting algorithm [42] (Figure ESPGrid). This methodology was compared against results from GAMESS [43] (Figure MKPopAnalScheme), to compare the grid generation, and against the R.E.D. server (Figure CompareChargedervation), to compare charge derivation. With respect to the nucleosides, all atoms in the ring and all substituents were set for equivalencing, except for C1' and H1' which should be equivalenced with the nucleobases on the occasion of multiple conformers. Hydrogens belonging to amines of nucleobases are also equivalenced, as are the OP1 and OP2 oxygens in the phosphoramidate linker. For second stage RESP, all methyls and methylenes were equivalenced, due to degenerate hydrogens. The force field of MNA contains charges for 6'-head, 6'-fragments-3', 3'-tail, neutral and methyl-capped MNA. Only the lowest energy  ${}^4\text{C}_1'$  conformer was considered for charge derivation. To create terminal residues and in-strand fragments, 6' and 3' ends were appropriately restrained with the linker to have a net charge of zero.

Finally, all bond, angle and torsion parameters for the morpholino chemistry were manually curated from the *parm10.dat* frcmod file, contained within the AMBER package, of terms that already described parts of the molecule well, before fitting to the conformers. Only torsion terms of the morpholino ring moiety were fitted for. Paramfit [44, 45] was used to equate QM to MM energies correctly, which also produces a valid *frcmod* file. All conformers in Table relEnergyConformers) were initially considered for fitting torsion angles. Finally, only  ${}^4\text{C}_1'$ ,  ${}^2\text{T}_3$ ,  ${}^2\text{S}_4$ ,  ${}^4\text{S}_2$ ,  ${}^3\text{T}_2$  of the MNA cytosine were employed for fitting.

A customised CF atom type was created, which was made equivalent to the standard CT atom type, common in AMBER FFs. This was done so as not to override any parameters when loading in the MNA FF together with the standard FFs.

### 3.2.5. Molecular Mechanics simulations

The Molecular Dynamics simulations were run using the AMBER18 software package [32, 33] joined with AMBERTools19, employing the Particle Mesh Ewald (PMEMD) simulation engine [46]. The MNA force field was imported into LEaP, together with the DNA.OL15 and the RNA.OL3 FF [47, 48]. The TIP3P water model is used for the explicit solvation, in a truncated octahedron box, and the charges were neutralised. dXyNA parameters [8] use the DNA.OL15 FF, HNA parameters were derived from a similar methodology as the one described in this article [49]. Torsion angles for the linker were taken from the GAFF2 parameters, in combination with phosphate parameters from *parm10.dat*. A cutoff distance of 12 Å was used for non-bonded interactions. The minimisation ran for a total of 30 000 cycles, with the first 22 500 cycles employing the steepest descent method and the last 7 500 the conjugate gradient method. The SHAKE algorithm [50] was employed to allow a timestep of 2 fs. An initial heating was performed for 50 ps, from 0 K to 100 K with *vlimit* set to 15. The rest of the heating, from 100 K to 300 K, ran for an additional 50 ps. Density and equilibration ran for 100 ps each, with density set at 1 g/mL. The Langevin thermostat [51] and the Berendsen barostat [52] were used to keep the temperature and density at a constant value. The production simulations were run for 200 ns, unless stated otherwise. Trajectory analysis of the simulations calculated through Cpptraj [53]. MD simulations were run on NVIDIA GeForce RTX 2070 and 2060 through the *cuda* accelerated simulation

engine [54, 55]. For all duplexes, distance-based NMR restraints were employed for the first 15 ns. For structural comparison, superpositioning of the structures was done with UCSF Chimera's MatchMaker tool and all structures were entirely fitted according to their secondary structure. The Curves+ software was used to extract interbasepair data [31].

### 3.3. Results

#### 3.3.1. Charge derivation using ORCA

The ORCA QM package does not contain the MK population analysis scheme. To keep the workflow free, local and easy to use, we implemented the population analysis scheme for ORCA to generate ESP charges that are ready to use for the RESP fitting. Two programs were used to compare Connolly surface grid generation to decide which program is most suited for the task at hand. The AMBERTools-script *molsurf.c*, which calculates molecular surface areas, was modified to output the vertices of the surface points by which it computes surface areas. A second program, MSMS was also tested (Scripps Research Institute, utilised by VMD [56] and UCSF Chimera [57]). The latter computes a further triangulation of the solvent excluded surfaces (SES) to generate an even distribution of points across the surface of the molecule. Upon visual inspection and level of customisability of the grid (atomic and probe radii, point density), MSMS was hugely favoured. Further comparison of the method was done against GAMESS' output, which defaults to a grid with a density of  $1/\text{\AA}^2$ . Our implementation for ORCA uses a density of  $3/\text{\AA}^2$  [35, 36]. Point density of  $1/\text{\AA}^2$  was used for our method to compare against GAMESS. Our methodology shows an even distribution of the surface, compared to GAMESS, and decided to proceed with this approach as it produced a near-identical range of ESP values to that of GAMESS' (Figure MKPopAnalScheme and ESPGrid). A benchmarking against RESP-charge derivation by the R.E.D. server [58] was carried out on HNA nucleosides. We see near-identical atomic charges being fitted to the atoms in the respective nucleosides (Figure CompareChargedervation). Data points from either methodology were compared by pairwise atom alignment (linear regression, SciPy) per nucleoside. This results in an  $r^2$  of 0.985 for hA, 0.985 for hC, 0.987 for hG and 0.979 for hT.

#### 3.3.2. RNA and DNA model building by Ducque

A-type and B-type duplexes of DNA and RNA (Fig. SIShowcase A1.) have been built by Ducque and its performance was compared to NAB to assess its model building capacity. Conserving the planarity of the nucleobases is necessary because Ducque's placement and fitting of the nucleosides in the strands assumes planarity of the nucleobases (Figure BuildComp). Because of their  $\text{sp}^3$  type hybridisation, amines tend towards slight tetrahedral structures in nucleobases, and this gets more pronounced in *ab initio* calculations at MP2 levels with midsized basis sets for geometry optimisations. This compromises the flatness of the nucleobase by several tenths of degrees for the nitrogen involved in the glycosidic bond, to several degrees for the amine on the six-membered ring in purines [59]. Optimising at the current level compromises the planarity assumption, introducing distortions during duplex building and causes peculiar rotations of nucleotides in the final model. *Nota bene*, using larger basis sets, i.e. reaching complete basis set, decreases the degree of pyramidalisation [60]. Using this basis would be too computationally intensive for our needs and would only finetune details that would be negated in the following simulations. Therefore, the planarity issue of nucleobases was resolved by the use of planarity restraints on nucleobases during HF-3c geometry optimisations, providing nucleosides for the Ducque library with planar nucleobases, like in NAB.

The quality of the model building is assessed by superpositioning generated models from NAB [18] vs. Ducque (Table BenchmarkNA and Figure BenchmarkNA), which was done across all (twelve) residue pairs on heavy atoms, for ten models of DNA and RNA each per software. For DNA, mean RMSD value is  $0.011 \text{ \AA}$  ( $\pm 0.005$ ) with the lowest value being  $0.004 \text{ \AA}$ . For RNA, mean RMSD value is  $0.169 \text{ \AA}$  ( $\pm 0.042$ ) with the lowest value being  $0.109 \text{ \AA}$ . The near-zero RMSD showcases Ducque's qualitative output. For the interbasepair-param-

ters, we see that the Ducque structures are in good agreement with NAB's. Only the roll parameter, for RNAs, seems to slightly differ. This stems from the fitting protocol not rolling all nucleobases with the exact same value, as this feature is not hardcoded. The simulated structure for the DNA and RNA homoduplex experiments, generated by both NAB or Ducque, show equivalent RMSD trajectories, as is to be expected given the similarity of the starting structures. Since all structures minimised without a problem, both the comparison of structural parameters and the following simulations indicate the quality of the generated structures. With respect to wallclock building-time, NAB averaged 0.272 s and 0.245 s for DNA and RNA respectively. Ducque, averaging 0.328 s and 0.234 s for DNA and RNA respectively, takes infinitesimally longer for DNA as it considers multiple conformers during the fitting (benchmarked with Python3.12).

### 3.3.3. MD simulations on Ducque's models validated with (X)NA structures in literature

To assess the quality of Ducque's models for MD simulations, a series of duplexes were studied, starting with (deoxy)-ribose based nucleic acids, which employed the standard AMBER DNA.OL15 and RNA.OL3 FFs. The simulated structures for the DNA and RNA homoduplex experiments, generated by both NAB or Ducque, show equivalent RMSD trajectories (Table BenchmarkNA and Figure BenchmarkNA), as expected given the similarity of the starting structures. An MD simulation on a B-type DNA::RNA and A-type RNA::DNA duplex of identical sequence both built by Ducque, converged into the same A-like structure, as experimentally determined in Davis *et al.* [61] (Figure RNADNA). A second experiment was performed with the dXyNA homoduplex. This started out from a typical ladder-like structure [8]. During the MD simulation, the initial structure of the duplex undergoes a remarkable transition to a highly dynamic, left-handed helix (Figure Dxylo, DistrNormDxyna, Table DxynaBackbone) that matches the type II duplex previously described for xylose based nucleic acids [62, 63, 64], having significantly different backbone dihedral angles compared to those in the initial model. Thirdly, an RNA::HNA duplex was built, using the lowest energy  ${}^4C_1$  conformations of the PES [25] of the four nucleobases and were prepared and loaded into Ducque's library. A subsequent MD simulation resulted in a stable duplex that matches crystallographic results on sugar puckering as well as overall helical structure [65] (Figure HNA duplex). The methodology of deriving torsional parameters for HNA has been validated already [49], which has also been applied to deriving the parameters for the MNA chemistry.

### 3.3.4. Using the proposed workflow to gain insight in RNA::MNA

→ Quantum Mechanics to characterise the nucleosides and linker

Conformer	mA*	mG*	mC*	mT*	$(\theta, \varphi_2)$
${}^4C_1$	0.00	0.00	0.00	0.00	( 4.01, 119.75)
${}^2T_{3'}$	7.41	7.48	8.03	7.64	( 53.29, 150.11)
$E_{3'}$	15.41	15.59	15.83	15.50	( 53.29, 179.91)
${}^2S_{4'}$	10.64	10.77	11.25	11.28	( 93.97, 80.21)
${}^4S_{2'}$	7.59	8.06	8.97	8.60	( 93.97, 264.13)
${}^3E$	18.76	19.39	18.46	18.40	(126.62, 0.00)
${}^3T_{2'}$	8.90	9.71	9.35	9.14	(126.62, 330.02)
${}^1C_{4'}$	17.04	17.65	19.55	19.15	(175.90, 0.00)

Table 3-2:  $E_{rel}$  of the distinct conformers with energy values and their respective position on the CP-sphere's surface, per given nucleoside, r is omitted as this is assumed constant. \* $E_{rel}$  expressed in kcal/mol.

Following the procedure described by Matteleaer *et al.* [25], PESs were generated for the morpholino nucleosides with adenine, guanine, thymine and cytosine nucleobases. We observed intramolecular hydrogen bonds in the initial calculations, inflating equatorial conformations towards higher energy potentials (Figure PESUnrestrained). Those observed hydrogen bonds, which would not exist in a polymeric molecule, introduce a forced steric strain on the nucleoside that would not yield representative FF parameters that are applicable for simulations on

oligonucleotides. For the capped MNA variant (Figure 3-3, Table 3-2), we see a global minimum on the north pole of the CP-sphere, a  ${}^4C_1$  configuration for all four nucleosides. Local minima are observed mainly on

the upper hemisphere, separated by large peaks corresponding to envelope configurations. Equatorial puckering modes tend to be more energetically favourable, with boats and skews alike. A final local minimum is observed on the southern hemisphere, in the East.

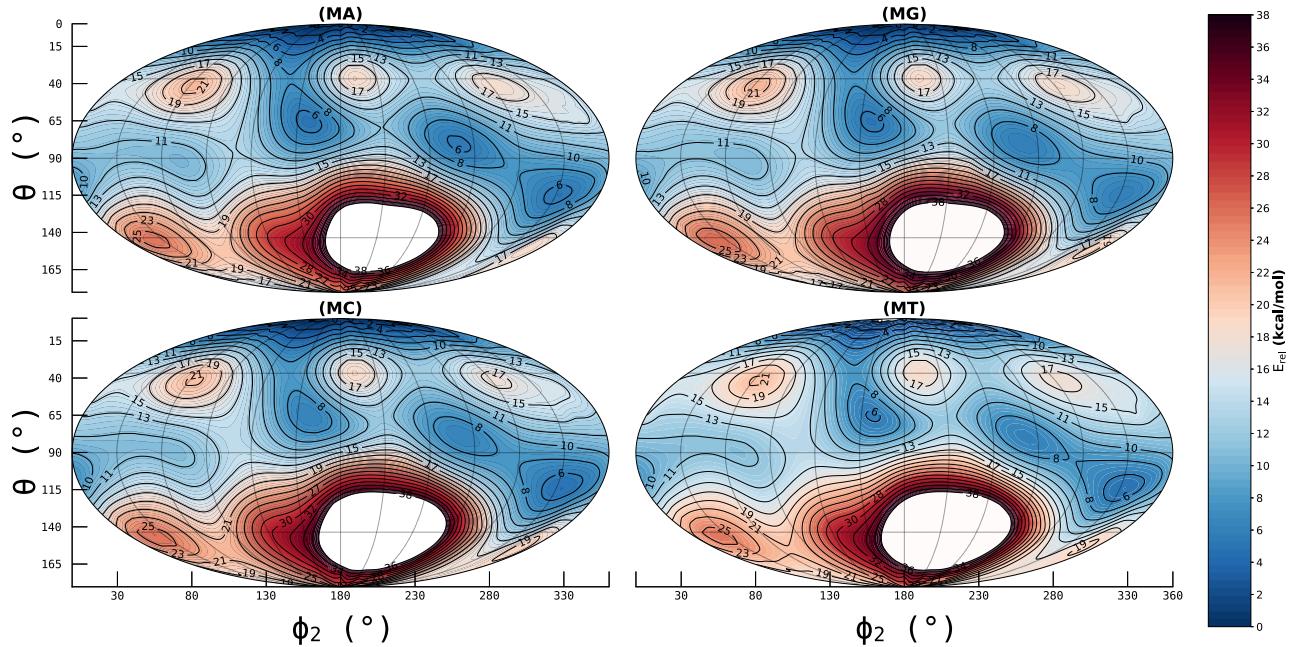


Figure 3-3: Potential Energy Surface of the Morpholino Nucleic Acid. Methyl-caps substituted the HO6' and NH3' on the O6' and N3' atoms respectively, to mimic estimated backbone angles and prevent internal hydrogens bonds from occurring during geometry optimisations.

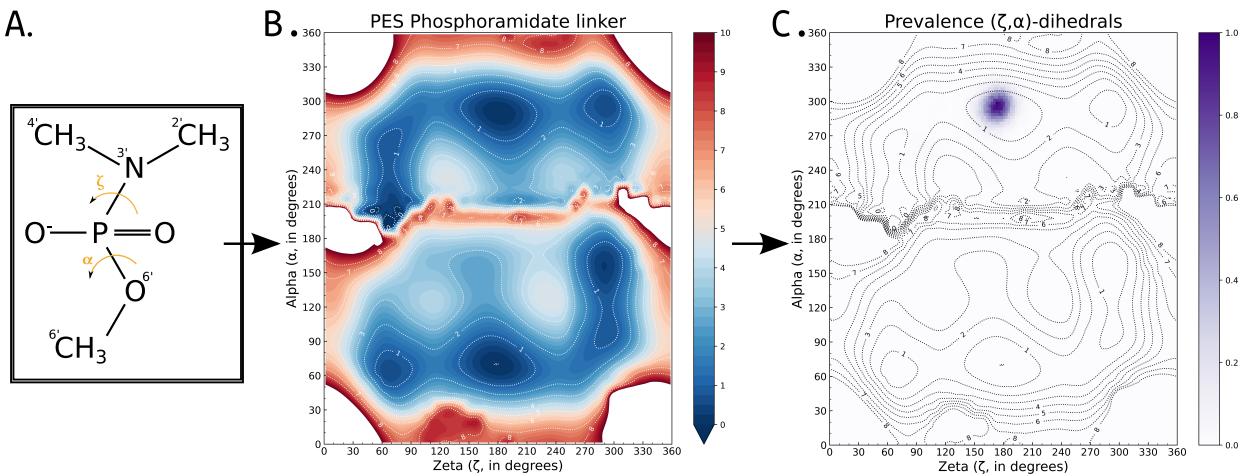
The N,N-dimethylaminophosphoramidate represents the linkage between successive morpholino nucleosides, where the N,N-dimethylamino mimics the effect of the preceding morpholino ring. A sampling of the ( $\zeta$ ,  $\alpha$ ) dihedral dyad was carried out too. Minima found at (t, g<sup>+</sup>), (t, g<sup>-</sup>), (g<sup>-</sup>, g<sup>-</sup>), (g<sup>+</sup>, g<sup>+</sup>) and (g<sup>-</sup>, ap<sup>+</sup>), Figure 3-4. A hard ridge ( $\alpha = 180^\circ \pm 30$ ) splits the PES in two. Its cause is attributed to steric clashes during the sampling itself, due to methyls on either end of the linker residing in close proximity of one another at  $\alpha$  in the *anti* range.

#### Force field parametrisation

The PESs of the four morpholino nucleosides successfully provided valuable information on transitional pathways. By using the improper dihedrals ( $\alpha_1$ ,  $\alpha_2$ ,  $\alpha_3$ ) as restraints on the pyranose ring, we gained information of possible transitions of conformations as well as puckers we should be expecting. Conformations highlighted in Table 3-2 are conformations that were initially considered for parametrisation, specifically for fitting of the sugar torsion angles (see Fig. sphericalconvention for a reference to their respective physical location on the PES). Parameters for dihedral terms of only the morpholino ring were refined by fitting to QM conformational energy [44]. In this procedure, the quality of the force field depends on the functionals and basis sets used for optimisations, as well as which conformations should be curated from the PES and included in the fitting process. The methylated ends were kept during optimisation and the fitting, as the initial problem of internal hydrogen bonds observed during PES calculation reappeared when reverting back to hydroxyls. Iterative curation of the returned parameters resulted in a simplification of the dataset; only the morpholino cytosine was included with the following puckers : <sup>4</sup>C<sub>1</sub>, <sup>2</sup>T<sub>3</sub>, <sup>2</sup>S<sub>4</sub>, <sup>4</sup>S<sub>2</sub>, <sup>3</sup>T<sub>2</sub>.

For the linker fragment, the minima and transition (g<sup>-</sup>,g<sup>+</sup>) conformers of the lower half were used to derive partial charges. Since upper half conformers are mirror images of the lower half, they were not considered. The linker torsional parameters were obtained from GAFF2 [66] and are in good agreement with the PES. (Figure 3-4 C.). The ( $\zeta$ ,  $\alpha$ ) dihedral dyad values were extracted from the simulation and categorised into dis-

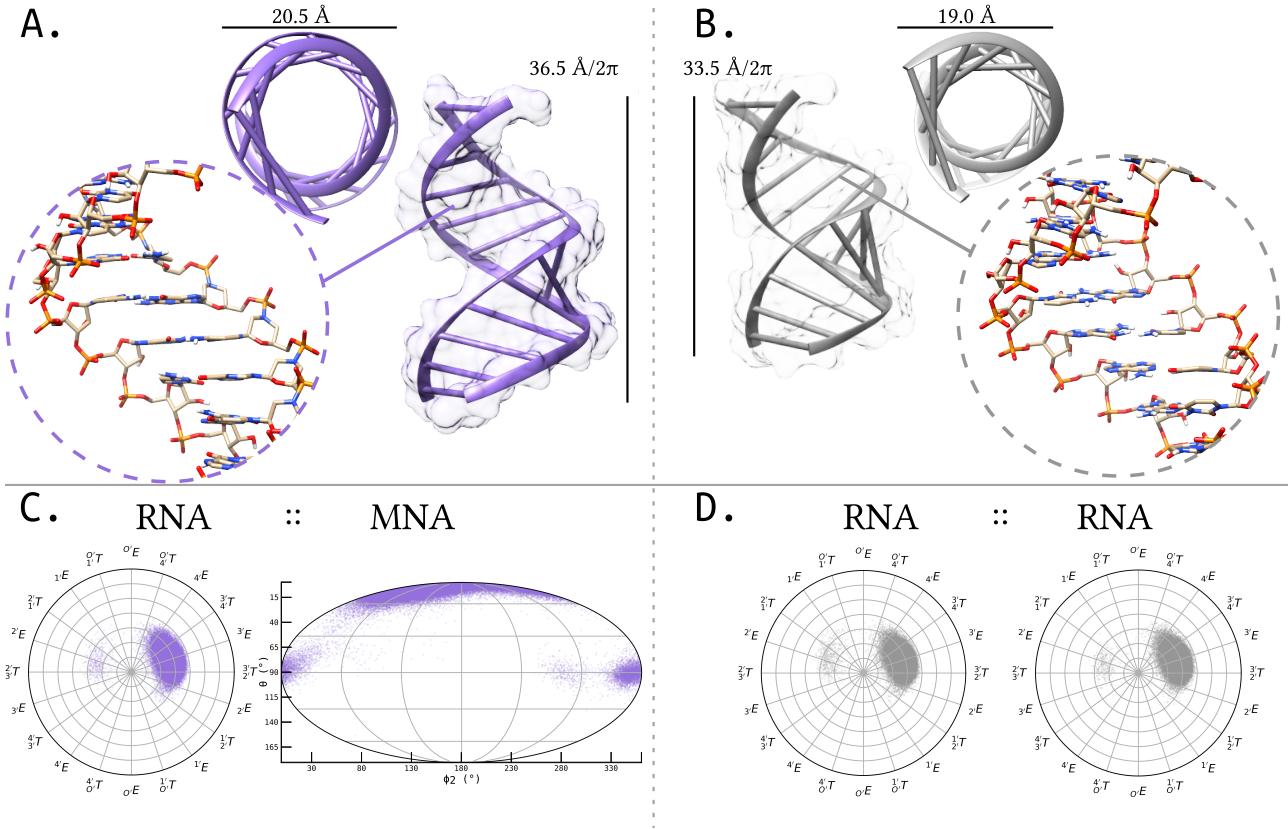
crete integer values, for then to be normalised and plotted. This shows a clear and single peak at the ( $t$ ,  $g^-$ ) range, which is in contrast with the standard phosphate, that stabilises in the ( $g^-, g^+$ ) range in a DNA or RNA duplex [67]. This validates usage of the GAFF2 parameters for the linker. *Nota bene*, a single peak is attributed to the conformational constraints that the linker is submitted to in a polynucleotide chain, which clarifies why multiple minima are perceived in the PES, but only one can be inhabited.



**Figure 3-4:** Validation of the parametrised backbone N,N-dimethyl-O-methylaminophosphoroamidate linker; PES and torsional prevalence during the simulation. **A.** The ( $\zeta$ ,  $\alpha$ ) dihedral dyad over which we extracted all conformers of the N,N-dimethyl-O-methylaminophosphoroamidate linker. **B.** Potential Energy Surface of the conformational landscape of the N,N-dimethyl-O-methylaminophosphoroamidate linker. **C.** The prevalence of the ( $\zeta$ ,  $\alpha$ ) phosphoroamidate linker backbone, extracted from the last 50 ns of the 200 ns simulation. Data was extracted from the simulation and normalised according to the prevalence count of the dihedral dyad.

#### → Molecular model of RNA::MNA

The RNA::MNA duplex was built in Ducque starting with a 24-mer RNA sequence corresponding to a double-Drew-Dickerson-Dodecamer (double-DDD : 5'-CGCGAAUUCGCGCGCAAUUCGCG-3', residues 1-24) with the geometry of an A-type duplex. All <sup>4</sup>C<sub>1</sub>-MNA nucleosides were ported into Ducque and queried as the complementary strand (residues 25-48). The obtained initial structure was subjected to an MM simulation. An RNA::RNA duplex with identical sequence was generated and simulated for comparative purposes. Results of the simulation are summarised in Figure 3-5. In the last 50 ns of the MD trajectory, a pronounced prevalence for the <sup>3</sup>E conformation in the conformational wheel is observed for RNA nucleosides in both the dsRNA and heteroduplex model (Figure 3-5 C,D, Table backbone). The puckering modes for the morpholino rings observed on the CP-sphere (Figure 3-5 C), in the same timeframe as stated earlier, largely correspond to the low energy regions of the PES (Figure 3-3). The scattered, equatorial conformers correspond to residues 30-31. Their appearance correlate with a K<sup>+</sup>-ion that had been enclosed by two MNA residues (30-31) during the second half of the simulation. Its binding induced local backbone angles to deviate substantially from the other nucleosides, off-setting average and standard deviations (std.) parameters by several degrees. Up to four (25-28) residues had frayed (forming C, E and B conformers alike), causing residue 28 with guanine nucleobase (mG28) to have its  $\beta$  angle to twist into an unfavourable position. Structural changes due to K<sup>+</sup>-binding destabilised downstream stacking, which causes mA30 to lose pairing with its complement. This resulted in a nearby K<sup>+</sup>-ion to alter the  $\chi$  angle's orientation, of the freed nucleotide, from *anti* to *syn* [68]. The basepairing of mA30 with its complement was reinstated later, in Hoogsteen (HG), with the ion bound (Figure IonBackbone). The K<sup>+</sup> ion is coordinated by mT31:O6', mT31:O5' and mA30:N3' in the MNA backbone and N3 of the nucleobase in mA30.



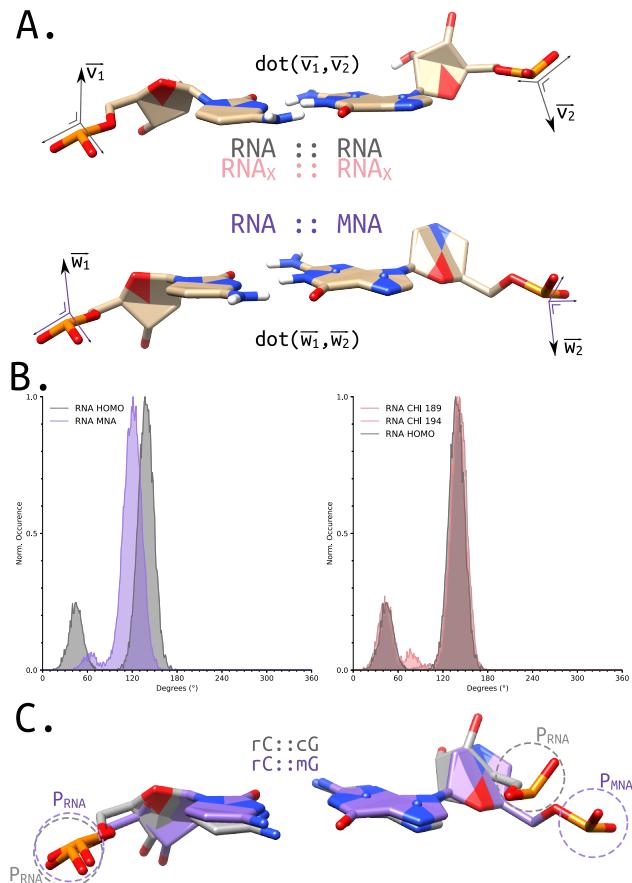
**Figure 3-5:** Comparison of the RNA::MNA and RNA::RNA duplexes. Trajectory of the simulation in Fig. XNAtrajectories. Scatterplots are not to be confused with prevalence plots, as scatterplots do not represent normalised data, but represent all data extracted. **A.** Visualisation of the diameter and the helicity for the RNA::MNA duplex. **B.** Visualisation of the diameter and the helicity for the RNA::RNA duplex. **C.** Scatterplot to represent puckering modes of both RNA (leading) and MNA (complementary) during the last 50 ns of the simulation. **D.** Scatterplot to represent puckering modes of RNA (both leading and complementary) during the last 50 ns of the simulation.

Coordination of  $K^+$  by N3' in the morpholino ring of mA30 forces it into metastable configurations (Fig. IonBackbone-justfig), as can also be seen on the scatterplots, around the equator (Figure ResultFINALSIMRNAMNA C). As this obvious abnormality is attributable to an odd occurrence of the simulation and not the FF itself, data of mA30 and m31 are excluded in Table HelicalParams, Backbone. This behaviour was not observed in the preliminary RNA::MNA simulations (Figure XNAtrajectories). Except for the  $\chi$  dihedral angle with a deviation of  $2.39\text{--}2.93^\circ$ , no significant differences in average backbone angles in the RNA backbone could be observed in the RNA::MNA duplex versus the model obtained for dsRNA (Table Backbone). However, helical parameters significantly deviate from the A-type duplex (Table 3-7) with a deeper major groove, and conversely more narrow minor groove, an increase in basepairs per turn and therefore helical size, along with a larger diameter.

To investigate a possible correlation of the altered  $\chi$  angle in the RNA strand to the overall shape for nucleic acid duplex, we performed two supplementary RNA::RNA double-DDD simulations (Figure Chiangle-TrajPucker, ChiangleCurves). While in the original simulations the  $\chi$  torsion angle was unrestrained, in the additional simulations restraints are applied on  $\chi$  torsion angles at  $194^\circ$  and  $189^\circ$  respectively to mimic the impact on the overall helix structure by increasing deviation from the  $199^\circ$   $\chi$ -angle in standard A-type duplexes, at  $5^\circ$  increments. The helical parameters of the RNA::RNA duplex with  $\chi$  torsion angles restraints at  $194^\circ$  are in line to those of the unrestrained RNA::MNA duplex (Table 3-7, ChiBackbone). The observed deviations from the A-type duplex are more pronounced in RNA::RNA duplex with  $\chi$  torsion angles restraints at  $189^\circ$ .

To compare the orientation of P atoms in the linker in complementary strands, the dot product of the cross product of the OP1-P-OP2 plane of the leading RNA nucleotide and complementary MNA nucleotide

was calculated and compared to values obtained for corresponding base pairs in 3 simulations performed on dsRNA (Figure LinkerOrientation). Normalised distribution per duplex type shows a clear peak for the RNA::RNA duplexes around  $137^\circ$ – $139^\circ$  while for RNA::MNA it peaks at  $120^\circ$ – $122^\circ$ . An overlay of a set of the unrestrained RNA::RNA and RNA::MNA C-G basepairs, shows that not only the orientation of P in the MNA strand is changed but also its position relative to the corresponding P in dsRNA.



**Figure 3-6:** “A. Calculated angles, by the dot product of the cross product of the OP1-P-OP2 plane of the leading nucleotide and complementary nucleotide. Sampled on 1000 frames, extracted from the entire simulation, of the double-DDD of RNA::RNA<sub>unrestr.</sub>, RNA::MNA, RNA::RNA<sub>χ189</sub> and RNA::RNA<sub>χ194</sub> respectively, for a total distribution of 23 000 angle values per duplex type. B. Normalised distribution per duplex type shows a clear peak for the RNA::RNA duplexes around  $137^\circ$ – $139^\circ$  degrees, while for RNA::MNA it peaks at  $120^\circ$ – $122^\circ$ . C. Overlay of a set of RNA::RNA and RNA::MNA C-G basepairs, to visualize the displacement of P in MNA relative to the corresponding RNA in the homoduplex.

**Table 3-7:** Helical parameters of the RNA::MNA heteroduplex in comparison with the RNA::RNA homoduplex (top), and the  $\chi$ -restrained RNA::RNA homoduplexes (bottom. Data extracted from the last 50 ns of the 200 ns simulation. Grooves and helicity as determined in Saenger et al [69]. Tilt, twist and roll parameters, determined by the Curves+ software on a 1000 frames that were extracted over the whole length of the trajectory [31]. See Fig. Basis-Frame for details on interbasepair parameters.

Helical Parameter	RNA::MNA	RNA::RNA
Helicity ( $\frac{\text{Å}}{2\pi}$ )	$35.83 (\pm 4.74)$	$32.50 (\pm 3.04)$
BPs per turn ( $\frac{\text{bp}}{2\pi}$ )	12	11
Major Groove ( $\text{Å}$ )	$15.90 (\pm 3.30)$	$9.86 (\pm 3.21)$
Minor Groove ( $\text{Å}$ )	$6.69 (\pm 0.62)$	$9.76 (\pm 0.62)$
Diameter ( $\text{Å}$ )	$20.18 (\pm 0.50)$	$19.07 (\pm 0.55)$
Tilt ( $^\circ$ )	$-1.34 (\pm 4.73)$	$-0.05 (\pm 4.61)$
Twist ( $^\circ$ )	$28.06 (\pm 3.99)$	$29.64 (\pm 4.18)$
Roll ( $^\circ$ )	$5.51 (\pm 6.27)$	$7.37 (\pm 6.42)$
Helical Parameter	RNA $\chi_{189}$	RNA $\chi_{194}$
Helicity ( $\frac{\text{Å}}{2\pi}$ )	$35.54 (\pm 2.89)$	$37.92 (\pm 3.23)$
BPs per turn ( $\frac{\text{bp}}{2\pi}$ )	12	12 - 13
Major Groove ( $\text{Å}$ )	$9.98 (\pm 3.17)$	$13.27 (\pm 2.91)$
Minor Groove ( $\text{Å}$ )	$9.64 (\pm 0.49)$	$9.43 (\pm 0.44)$
Diameter ( $\text{Å}$ )	$18.99 (\pm 0.55)$	$18.91 (\pm 0.60)$
Tilt ( $^\circ$ )	$0.014 (\pm 3.73)$	$-0.08 (\pm 3.70)$
Twist ( $^\circ$ )	$29.58 (\pm 3.85)$	$28.70 (\pm 3.85)$
Roll ( $^\circ$ )	$4.35 (\pm 5.46)$	$2.75 (\pm 5.35)$

### 3.4. Discussion

Xeno-nucleic acids (XNAs) are a class of synthetic molecules that differ from naturally occurring nucleic acids in the use of altered backbone that can impact the three dimensional structure, stability and processing by naturally occurring biological processes. Results on numerous XNAs have been synthesized and characterized in the last decades revealed that ability to form stable homoduplexes and heteroduplexes with natural nucleic acids determines potential applications [24]. Molecular modeling simulations can provide understanding and predict how changes in the sugar-phosphate backbone impact on the complementation properties of the nucleic acids. However, the missing link in this field is a software tool that builds initial XNA models beyond ribose-based nucleic acids and robust force field parametrization to perform molecular mechanics simulations within standard packages such as AMBER. To fill this gap, we introduced a streamlined

molecular modeling approach starting from quantum mechanics on XNA fragments. The versatile nucleic acid builder (Ducque) presented here builds XNA structures with unlimited backbone chemistry. It relies on a library of nucleosides and linkers that is user-extendable. We demonstrated the tool can build initial models starting from preferably low energy conformations of nucleosides and linkers, extracted from conformational sampling experiments. The latter also delivered data for force field parametrisation required to perform molecular mechanics simulations on the initial model. Ducque produced models of RNA and DNA duplexes as fast as the NAB tool.

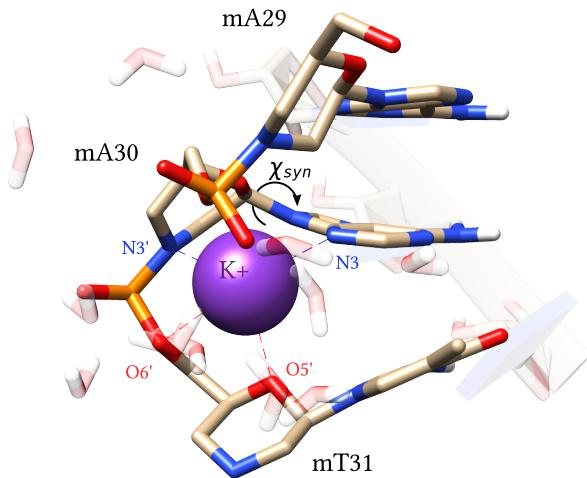
The latter is part of the AMBER toolkit and requires other AMBER programs (like tLEaP) to run, whereas Ducque functions as a standalone software. Syntactically, Ducque reads in simple text files to directly generate a model, as opposed to NAB's DSL which writes like C code and requires compilation of and executing the produced binary. Produced initial models are highly similar for both model builders and provided the same results in a subsequent MD simulation in standard force fields of AMBER. In contrast to NAB, Ducque is not limited to ribose based nucleic acids. Low energy conformations for non-ribose based nucleosides with hexitol and xylose sugars were included in the Ducque library, among others, and served to build nucleic acid duplexes for which

MD simulations could be performed parametrised force fields within AMBER [49, 8].

As sugar puckering is crucial for the backbone geometry and flexibility, we used Paramfit to optimise dihedral force field parameters on non-ribose nucleosides by fitting QM to MM energies of selected conformations of nucleosides after RESP charges had been calculated through the ORCA implementation. Linker parameters from GAFF2 were validated through QM approaches.

The initial RNA::HNA duplex was built and remained stable during the MD round. Ducque's model of dXyNA homoduplex with backbone dihedrals for a ladder-like structure converged towards the left-handed helical structure in the MD simulation, confirming what was predicted the dXyNA duplex [62, 63, 64] (Figure S13). This result, and that of the DNA::RNA heteroduplex, ensured that dihedral angles used for model building in Ducque did not determine the outcome of the MD simulation.

The methodology described to derive charges for naturally occurring RNA modifications [41] starts off differently than the one described in this work, but has the same goal. At last, experiments were performed on an RNA::MNA duplex. The charges for the MNA chemistry were derived using the new ORCA implementation, that was evaluated on HNA. According to the proposed workflow, low energy conformations of nucleosides and linker are described by QM calculations and ported to the Ducque library. In line with the available crystal structure [70], the  ${}^4C_1$  chair having the nucleobase in an equatorial position turned out to be the lowest energy conformation of morpholino nucleosides according to QM. An initial model was produced by ducque and subjected to an mm simulation using a force field that was parametrised as described in the methods section. the rna::mna heteroduplex simulations yielded a right-handed helical structure that does not belong to the A- or B-type family [71]. The morpholino rings remain predominantly in their chair conformation with an equatorial orientation of the nucleobase. Also in homoduplexes of  $(4' \rightarrow 6')-(\beta\text{-D-glucopyranosyl})$  oligonucleotides (or  $\beta$ -homo-DNA) [3],  $(4' \rightarrow 6')$  and  $(3' \rightarrow 6')$  pentopyranose systems [72] this conformation



*Figure 3-8: The  $K^+$ -ion enclosed in the backbone. Notice the mA30's glycosidic bond angle having shifted to a syn configuration. See Figure IonBackbone for an detailed look into the enclosed ion throughout the MD simulation.*

is adopted by the six-membered pyranoses in the backbone but none of these XNA chemistries is able to communicate by base-pairing with either RNA or DNA. Changing the C1' configuration (e.g.  $\alpha$ -homo-DNA) or having the nucleobase on a C2' axial position instead of the regular C1' anomeric (equatorial) site of the six membered ring in the ‘glycon’ moiety (HNA), generated XNA systems that cross-pair with natural nucleic acids [6].

Except for a Watson-Crick-Franklin (WCF) to Hoogsteen (HG) transition in a mA:U basepair (Figure 3-8), that was facilitated by K<sup>+</sup> binding and fraying at the helix end, stable WCF base pairing was observed during the MM simulation. Transitions between WCF and HG could be important in DNA recognition and replication, but are difficult to investigate experimentally because they are rare and short-lived. A transition from WCF to HG basepairing had been observed before for DNA in relaxation dispersion NMR experiments [73]. The same outside-route transition from WCF to HG base pairing was observed in DNA simulations, as the nucleotide temporarily unpairs and interacts with the solute, before repairing [74]. The occurrence of a WCF to HG transition in the final MNA::RNA simulation demonstrates that our optimised dihedral force field parameters allow puckering in the morpholino ring that is also expected in real life.

Only subtle changes in dihedral angles of the RNA strand, compared to values measured in dsRNA A-type helices, are imposed by the complementary MNA. To complement MNA, RNA adapts to the conformational constraints of the rigid binding partner with a more *anti*-characterised  $\chi$  torsion angles while other dihedral angles remain closely to what is observed in dsRNA (Table Backbone, Fig. ChiangleCurves). This minor difference significantly changes helical parameters for helicity, groove widths, roll and twist of the duplex, as was demonstrated by restrained MD on dsRNA resulted in the same effect that is more pronounced if restraints on  $\chi$  torsion angles deviate further from the 199° observed in A-type duplexes (Table HelicalParams). The increased diameter can be attributed to the six-membered morpholino ring in the backbone, in line to other XNA with six-membered sugar moieties [65, 75]. The unique chemistry of the MNA backbone changed orientation of the phosphoramidate in MNA and displaces the P atom in a position where it does not align with phosphates in the backbone of standard RNA (Figure 3-6). Combined with the significantly different helical parameters, like increased radius of the double helix, altered positioning of the phosphoramidate due to the morpholino ring, can all contribute to the lack of RNA degradation when bound to complementary MNA sequences [15].

### 3.5. Conclusion

To conclude, the current work describes Ducque, a free and open-source program to construct initial models for (synthetic) nucleic acid duplexes of virtually any XNA chemistry. The manuscript proves that Ducque easily builds any type of structure for a given chemistry and that, with the correct force field, an accurate prediction can be modelled. We demonstrated one can extend the nucleic acid library with new sugars and nucleoside linkers. Ducque requires only built-in and NumPy modules, and Tkinter for the GUI. The user can supply Ducque with chemistries and their respective conformations, which can be confidently generated for with the conformational sampling scheme when no experimental structure is available. The current work describes Ducque, a free and open-source program to construct initials models for (synthetic) nucleic acid duplexes of virtually any XNA chemistry. The DNA and RNA chemistries were already well documented in literature with both crystallographic and NMR structures. This allowed parameters to be fitted on nucleosides whose behaviour was readily described in detail. For a given XNA that has not benefitted yet from years of characterisation, it is shown that the computed PESs are exceptionally accurate in predicting its behaviour.

The manuscript proves that not only can Ducque easily build any type of structure for a given chemistry, but that with the correct force field an accurate prediction can be modelled. We demonstrated one can extend the nucleic acid library with new sugars and nucleoside linkers. Ducque requires only built-in and NumPy

modules, and Tkinter for the GUI. The user can supply Ducque with chemistries and their respective conformations, which can be confidently generated for with the conformational sampling scheme. On top of that, a suitable force field can be derived.

The RNA::MNA heteroduplex simulations yielded a right-handed helical structure that does not belong to the A- or B-type family. Only subtle changes in dihedral angles of the RNA strand, compared to values measured in dsRNA A-type helices, are imposed by the complementary MNA. The significantly different helical parameters, like increased radius of the double helix and altered positioning of the phosphoramidate due to the morpholino ring, can all contribute to the lack of RNA degradation when bound to complementary MNA sequences. Except for a Watson-Crick-Franklin to Hoogsteen transition in a mA:U basepair, that was facilitated by K<sup>+</sup> binding and fraying at the helix end, stable WCF base pairing was observed during the MM simulation. This corroborates well with the available data that the heteroduplex differs from either A- and B-type duplexes and it is strongly suspected it acts as a steric block, in the ASO pathways, as it has also been demonstrated the heteroduplex is not fit for degradation by the RNase H enzyme [15, 71].

### 3.5.1. Supplementary information

[Supplementary Data](#) is available at NAR Online. The Ducque GitHub repository contains an extensive and comprehensible manual on how to start with Ducque and broaden the library of chemistries (Ducque/docs/). It also contains a separate guide on how to build a force field (Ducque/ff/). The latter directory contains the Morpholino FF used in this project.

Ducque is available on GitHub [github.com/jrihon/Ducque](https://github.com/jrihon/Ducque) and works on most Linux operating system (tested on Ubuntu 20 and 22 LTS), it runs on MacOS and *Windows Subsystem for Linux* (WSL).

CeNA,  $\beta$ -homo DNA, (d)XyNA, 2'-O-Me RNA, 2'-Fluoro RNA, HNA and MNA are included. It produces full homoduplexes, heteroduplexes and allows mixed chemistries in the complementary strand. The leading sequence is also allowed to vary in chemistry, though we advise to keep the leading strand within the same XNA type.

A separate guide is provided on how to build a force field, how AMBER interprets its atoms, terminology, requirements for the creation of force fields, on how to perform parametrisation (resp, equivalence, restraints, paramfit, etc.) on a local machine. All figures are generated through python-scripted methods (Libraries : matplotlib, cartopy, NumPy, SciPy, mayavi) and/or concatenated/adapted in InkScape. All python libraries used are freely available through conda and/or pip (NumPy, SciPy, Tkinter, Cartopy, Mayavi).

" ... And if you don't believe me, I have quotes."

~ Pascal Hertleif - Writing Idiomatic Libraries in Rust ([YouTube @4:00](#)) ~



- doi:10.1002/(sici)1097-0282(199603)38:3<305::aid-bip4>3.0.co;2-y
- [40] Connolly. M. "Analytical molecular surface calculation" *Journal of Applied Crystallography*, vol. 16, pp. 548, 1983-10, doi:10.1107/s002188983010985
- [41] Aduri. R., Psciuuk. B., Saro. P., Taniga. H., Schlegel. H., SantaLucia. J., "AMBER Force Field Parameters for the Naturally Occurring Modified Nucleosides in RNA" *Journal of Chemical Theory and Computation*, vol. 3, pp. 1464, 2007-06, doi:10.1021/ct600329w
- [42] Chirlian. L., Franci. M., "Atomic charges derived from electrostatic potentials: A detailed study" *Journal of Computational Chemistry*, vol. 8, pp. 894, 1987-09, doi:10.1002/jcc.540080616
- [43] Barca. G., Bertoni. C., Carrington. L., Datta. D., De Silva. N., Deustua. J., Fedorov. D., Gour. J., Gunina. A., Guidez. E., Harville. T., Irle. S., Ivanic. J., Kowalski. K., Leang. S., Li. H., Li. W., Lutz. J., Magoulas. I., Mato. J., Mironov. V., Nakata. H., Pham. B., Piecuch. P., Poole. D., Pruitt. S., Rendell. A., Roskop. L., Ruedenberg. K., Sattasathuchana. T., Schmidt. M., Shen. J., Slipchenko. L., Sosonkina. M., Sundriyal. V., Tiwari. A., Galvez Vallejo. J., Westheimer. B., Włoch. M., Xu. P., Zahariev. F., Gordon. M., "Recent developments in the general atomic and molecular electronic structure system" *The Journal of Chemical Physics*, vol. 152, NO PAGERANGE2020-04, doi:10.1063/5.0005188
- [44] Betz. R., Walker. R., "Paramfit: Automated optimization of force field parameters for molecular dynamics simulations" *Journal of Computational Chemistry*, vol. 36, pp. 79, 2014-11, doi:10.1002/jcc.23775
- [45] Weiner. S., Kollman. P., Case. D., Singh. U., Ghio. C., Alagona. G., Profeta. S., Weiner. P., "A new force field for molecular mechanical simulation of nucleic acids and proteins" *Journal of the American Chemical Society*, vol. 106, pp. 765, 1984-02, doi:10.1021/ja00315a051
- [46] Darden. T., York. D., Pedersen. L., "Particle mesh Ewald: An N $\log(N)$  method for Ewald sums in large systems" *The Journal of Chemical Physics*, vol. 98, pp. 10089, 1993-06, doi:10.1063/1.464397
- [47] Zgarbová. M., Šponer. J., Otyepka. M., Cheatham. T., Galindo-Murillo. R., Jurečka. P., "Refinement of the Sugar-Phosphate Backbone Torsion Beta for AMBER Force Fields Improves the Description of Z- and B-DNA" *Journal of Chemical Theory and Computation*, vol. 11, pp. 5723, 2015-11, doi:10.1021/acs.jctc.5b00716
- [48] Zgarbová. M., Otyepka. M., Šponer. J., Mládek. A., Banáš. P., Cheatham. T., Jurečka. P., "Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles" *Journal of Chemical Theory and Computation*, vol. 7, pp. 2886, 2011-08, doi:10.1021/ct200162x
- [49] Schofield. P., Taylor. A., Rhon. J., Peña Martinez. C., Zinn. S., Mattelaer. C., Jackson. J., Dhaliwal. G., Schepers. G., Herdewijn. P., Lescrinier. E., Christ. D., Holliger. P., "Characterization of an HNA aptamer suggests a non-canonical G-quadruplex motif" *Nucleic Acids Research*, vol. 51, pp. 7736, 2023-07, doi:10.1093/nar/gkad592
- [50] Kräutler. V., Hünenberger. P., "A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations" *Journal of Computational Chemistry*, vol. 22, pp. 501, 2001, doi:10.1002/1096-987x(20010415)22:5<501::aid-jcc1021>3.0.co;2-v
- [51] Davidchack. R., Handel. R., Tretyakov. M., "Langevin thermostat for rigid body dynamics" *The Journal of Chemical Physics*, vol. 130, NO PAGERANGE2009-06, doi:10.1063/1.3149788
- [52] Berendsen. H., Postma. J., DiNola. A., Haak. J., "Molecular dynamics with coupling to an external bath" *The Journal of Chemical Physics*, vol. 81, pp. 3684, 1984-10, doi:10.1063/1.448118
- [53] Roe. D., Cheatham. T., "PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data" *Journal of Chemical Theory and Computation*, vol. 9, pp. 3084, 2013-06, doi:10.1021/ct400341p
- [54] Salomon-Ferrer. R., Götz. A., Poole. D., Le Grand. S., Walker. R., "Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald" *Journal of Chemical Theory and Computation*, vol. 9, pp. 3878, 2013-08, doi:10.1021/ct400314y
- [55] Le Grand. S., Götz. A., Walker. R., "SPFF: Speed without compromise—A mixed precision model for GPU accelerated molecular dynamics simulations" *Computer Physics Communications*, vol. 184, pp. 374, 2013-02, doi:10.1016/j.cpc.2012.09.022
- [56] Humphrey. W., Dalke. A., Schulten. K., "VMD: Visual molecular dynamics" *Journal of Molecular Graphics*, vol. 14, pp. 33, 1996-02, doi:10.1016/0263-7855(96)00018-5
- [57] Pettersen. E., Goddard. T., Huang. C., Couch. G., Greenblatt. D., Meng. E., Ferrin. T., "UCSF Chimera—A visualization system for exploratory research and analysis" *Journal of Computational Chemistry*, vol. 25, pp. 1605, 2004-07, doi:10.1002/jcc.20084
- [58] Vanquelef. E., Simon. S., Marquant. G., Garcia. E., Klimerak. G., Delapine. J., Cieplak. P., Dupradeau. F., "R.E.D. Server: a web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments" *Nucleic Acids Research*, vol. 39, pp. W511–W517, 2011-05, doi:10.1093/nar/gkr288
- [59] Sychrovsky. V., Foldynova-Trantirkova. S., Spackova. N., Robeyns. K., Van Meervelt. L., Blankenfeldt. W., Vokacova. Z., Šponer. J., Trantirek. L., "Revisiting the planarity of nucleic acid bases: Pyramidalization at glycosidic nitrogen in purine bases is modulated by orientation of glycosidic torsion" *Nucleic Acids Research*, vol. 37, pp. 7321, 2009-09, doi:10.1093/nar/gkp783
- [60] Zierkiewicz. W., Komorowski. L., Michalska. D., Černý. J., Hobza. P., "The Amino Group in Adenine: MP2 and CCSD(T) Complete Basis Set Limit Calculations of the Planarization Barrier and DFT/B3LYP Study of the Anharmonic Frequencies of Adenine" *The Journal of Physical Chemistry B*, vol. 112, pp. 16734, 2008-12, doi:10.1021/jp8058118
- [61] Davis. R., Shaban. N., Perrino. F., Hollis. T., "Crystal structure of RNA-DNA duplex provides insight into conformational changes induced by RNase H binding" *Cell Cycle*, vol. 14, pp. 668, 2015-02, doi:10.4161/15384101.2014.994996
- [62] Maiti. M., Siegmund. V., Abramov. M., Lescrinier. E., Rosemeyer. H., Froeyen. M., Ramaswamy. A., Ceulemans. A., Marx. A., Herdewijn. P., "Solution Structure and Conformational Dynamics of Deoxyxylonucleic Acids (dXNA): An Orthogonal Nucleic Acid Candidate" *Chemistry – A European Journal*, vol. 18, pp. 869, 2011-12, doi:10.1002/chem.201102509
- [63] Ramaswamy. A., Froeyen. M., Herdewijn. P., Ceulemans. A., "Helical Structure of Xylose-DNA" *Journal of the American Chemical Society*, vol. 132, pp. 587, 2009-12, doi:10.1021/ja9065877
- [64] Ramaswamy. A., Smirnova. D., Froeyen. M., Maiti. M., Herdewijn. P., Ceulemans. A., "Molecular Dynamics of Double Stranded Xylo-Nucleic Acid" *Journal of Chemical Theory and Computation*, vol. 13, pp. 5028, 2017-09, doi:10.1021/acs.jctc.7b00309
- [65] Lescrinier. E., Esnouf. R., Schraml. J., Busson. R., Heus. H., Hilbers. C., Herdewijn. P., "Solution structure of a HNA–RNA hybrid" *Chemistry & Biology*, vol. 7, pp. 719, 2000-09, doi:10.1016/s1074-5521(00)00017-x
- [66] Wang. J., Wolf. R., Caldwell. J., Kollman. P., Case. D., "Development and testing of a general amber force field" *Journal of Computational Chemistry*, vol. 25, pp. 1157, 2004-04, doi:10.1002/jcc.20035
- [67] Zhang. C., Lu. C., Wang. Q., Ponder. J., Ren. P., "Polarizable Multipole-Based Force Field for Dimethyl and Trimethyl Phosphate" *Journal of Chemical Theory and Computation*, vol. 11, pp. 5326, 2015-10, doi:10.1021/acs.jctc.5b00562
- [68] IUPAC-IUB. J., "Abbreviations and Symbols for the Description of Conformations of Polynucleotide Chains" *European Journal of Biochemistry*, vol. 131, pp. 9–15, 1983, NO DOI
- [69] Saenger. W. "Principles of Nucleic Acid Structure" NO JOURNAL, (NO VOLUME), pp. XX-556, 1984, doi:10.1007/978-1-4612-5190-3
- [70] Zhang. W., Pal. A., Ricardo. A., Szostak. J., "Template-Directed Nonenzymatic Primer Extension Using 2-Methylimidazole-Activated Morpholino Derivatives of Guanosine and Cytidine" *Journal of the American Chemical Society*, vol. 141, pp. 12159, 2019-07, doi:10.1021/jacs.9b06453
- [71] Langner. H., Jastrzebska. K., Caruthers. M., "Synthesis and Characterization of Thiophosphoramidate Morpholino Oligonucleotides and Chimeras" *Journal of the American Chemical Society*, vol. 142, pp. 16240, 2020-08, doi:10.1021/jacs.0c04335
- [72] Eschenmoser. A., "Searching for Nucleic Acid Alternatives" NO PARENT FOUND, NO PARENT FOUND, pp. 5, 2011-02, doi:10.1002/9780470977873.ch1
- [73] Nikolova. E., Kim. E., Wise. A., O'Brien. P., Andricioaei. I., Al-Hashimi. H., "Transient Hoogsteen base pairs in canonical duplex DNA" *Nature*, vol. 470, pp. 498, 2011-01, doi:10.1038/nature09775
- [74] Vreede. J., Bolhuis. P., Swenson. D., "Atomistic insight into the kinetic pathways for Watson–Crick to Hoogsteen transitions in DNA" *Nucleic*

*Acids Research*, vol. 47, pp. 11069, 2019-10, doi:10.1093/nar/gkz837

[75] Egli. M., Pallan. P., Pattanayek. R., Wilds. C., Lubini. P., Minasov. G., Dobler. M., Leumann. C., Eschenmoser. A., "Crystal Structure of Homo-DNA and Nature's Choice of Pentose over Hexose in the Genetic System" *Journal of the American Chemical Society*, vol. 128, pp. 10847, 2006-08, doi:10.1021/ja062548x

# STRUCTURE PREDICTION OF AN HNA APTAMER WITH A G-QUADRUPLEX MOTIF

4

Adapted from the following manuscript [10.1093/nar/gkad592](https://doi.org/10.1093/nar/gkad592):

Schofield P., Taylor A.I., **Rihon J.**, Peña Martinez C.D., Zinn S., Mattelaer C.-A., Jackson J., Dhaliwal G., Schepers G., Herdwijn P., Leschinier E., Christ D., Holliger P. "Characterization of an HNA aptamer suggests a non-canonical G-quadruplex motif." Nucleic Acids Research (NAR).

## 4.1. Introduction

### 4.1.1. Welcome to the thesis

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequo doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut.

↪ An addition to the text

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequo doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut.

## 4.2. Methods

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequo doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut.

## 4.3. Result & Discussion

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequo doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut.

## 4.4. Conclusion

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequo doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut.

"Je rime et je rame comme tartine et boterham, hein comme on dit."

~ Claudy - Dikkenek ~

## 4.5. References

# CURVATURE TORSION ON SYNTHETIC NUCLEIC ACIDS

5

## 5.1. Introduction

### 5.1.1. Welcome to the thesis

Lore ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequa doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut.

↪ An addition to the text

Lore ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequa doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut.

## 5.2. Methods

Lore ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequa doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut.

## 5.3. Result & Discussion

Lore ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequa doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut.

## 5.4. Conclusion

Lore ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequa doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut.

" 'Andy, how did you get up that tree ?!'

- '... I fell.' "

~ Little Britain - S02E04 ~

## 5.5. References

# CONCLUDING REMARKS AND FUTURE PERSPECTIVE

# 6

## 6.1. Concluding remarks

  Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequo doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut.

    → An addition to the text

  Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequo doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut.

## 6.2. Future perspectives

  Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequo doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut.

    → An addition to the text

  Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequo doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut.

"Never tell me the odds."

~ Han Solo - Star Wars: Episode V The Empire Strikes Back ~

# Published works

This final section of this doctoral thesis contains the published works I had the pleasure of leading / partaking in. I want to extend my gratitude to everyone listed here below, as I was able to be challenged on topics pertaining to a variety of fields of scientific research. This brings me joy!

The *List of Publications* is ordered from high to low relevance with respect to the topic of my thesis.

## List of Publications

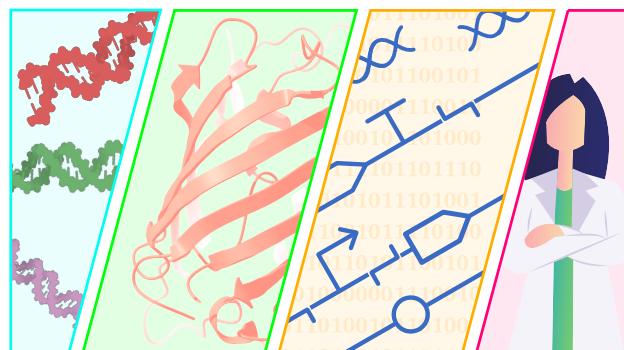
- [1] **Rihon, J.**, Mattelaer, C-A., Montalvão, R.W., Froeyen, M., Pinheiro, V.B., Lescrinier, E. (2024). Structural insights into the morpholino nucleic acid/RNA duplex using the new XNA builder Ducque in a molecular modeling pipeline. *Nucleic Acids Research* , [10.1093/nar/gkae135](https://doi.org/10.1093/nar/gkae135)
- [2] **Rihon, J.**, Reynders, S., Pinheiro, V.B., Lescrinier, E. (2024). The pucke.rs toolkit to facilitate sampling the conformational space of biomolecular monomers. *Journal of Cheminformatics* , [placeholder](#)
- [3] Mattelaer, C-A., Mattelaer, H-P., **Rihon, J.**, Froeyen, M., Lescrinier, E. with Lescrinier, E. (2021). Efficient and Accurate Potential Energy Surfaces of Puckering in Sugar-Modified Nucleosides. *Journal of Chemical Theory and Computation* , [10.1021/acs.jctc.1c00270](https://doi.org/10.1021/acs.jctc.1c00270)
- [4] Schofield, P., Taylor, A., **Rihon, J. (co-first)**, Pena Martinez, C.D., Zinn, S., Mattelaer, C-A., Jackson, J., Dhaliwal, G., Schepers, G., Herdewijn, P., Lescrinier, E., Christ, D., Holliger, P. (2023). Characterization of an HNA aptamer suggests a non-canonical G-quadruplex motif. *Nucleic Acids Research* , [10.1093/nar/gkad592](https://doi.org/10.1093/nar/gkad592)
- [5] Xu, Y., Groaz, E., **Rihon, J.**, Herdewyn, P., Lescrinier, E. (2023). Synthesis, antiviral activity, and computational study of β-D-xylofuranosyl nucleoside phosphonates. *European Journal Of Medicinal Chemistry* [10.1016/j.ejmech.2023.115379](https://doi.org/10.1016/j.ejmech.2023.115379)
- [6] Depuydt, A-S., **Rihon, J. (co-first)**, Cheneval, O., Vanmeert, M., Schroeder, C.I., Craik, D.J., Lescrinier, E., Peigneur, S., Tytgat, J. (2021). Cyclic Peptides as T-Type Calcium Channel Blockers: Characterization and Molecular Mapping of the Binding Site. *ACS Pharmacology & Translational Science* [10.1021/acsptsci.1c00079](https://doi.org/10.1021/acsptsci.1c00079)

## Software releases

- [1] Ducque : the (X)NA model builder [www.github.com/jrihon/Ducque](https://github.com/jrihon/Ducque)
- [2] pucke.rs : the CLI tool for conformational sampling [www.github.com/jrihon/puckers](https://github.com/jrihon/puckers)
- [3] pucke.py : the Python module for conformational sampling, puckering formalisms and inversion, and computing standard geometries [www.github.com/jrihon/puckepy](https://github.com/jrihon/puckepy)

## Scientific outreach

I had the pleasure of designing the poster for the EMBO | EMBL (European Molecular Biology Organization | Laboratory) course “Synthetic biology in action: beyond standard metabolism” (2022). They liked it so much that they reused my design for their 2024 course ([2022 event](#), [2024 event](#)) !



# Declarations

## Scientific acknowledgements

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequo doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distingue possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et.

## Personal contributions

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliquam quaerat voluptatem. Ut enim aequo doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distingue possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et.

## Conflict of Interest

The author declares no conflicts of interest or any financial interests.