Journal of the Royal Statistical Society Series C: Applied Statistics, 2023, **72**, 271–293
https://doi.org/10.1093/jrsssc/qlad004
Advance access publication 24 March 2023
Original Article

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series C
Applied Statistics

Learning torus PCA-based classification for multiscale RNA correction with application to SARS-CoV-2

Henrik Wiechers[1], Benjamin Eltzner[2], Kanti V. Mardia[3,4] and Stephan F. Huckemann[1]

[1]Felix-Bernstein-Institute for Mathematical Statistics in the Biosciences, Georgia-Augusta-University, Göttingen 37077, Germany
[2]Max Planck Institute for Multidisciplinary Sciences, Göttingen 37077, Germany
[3]Department of Statistics, School of Mathematics, University of Leeds, Leeds LS2 9JT, UK
[4]Department of Statistics, University of Oxford, 24-29 St Giles', Oxford OX1 3LB, UK

Address for correspondence: Henrik Wiechers, Felix-Bernstein-Institute for Mathematical Statistics in the Biosciences, Georgia-Augusta-University, Göttingen 37077, Germany. Email: henrik.wiechers@uni-goettingen.de

Abstract

Three-dimensional RNA structures frequently contain atomic clashes. Usually, corrections approximate the biophysical chemistry, which is computationally intensive and often does not correct all clashes. We propose fast, data-driven reconstructions from clash-free benchmark data with two-scale shape analysis: microscopic (suites) dihedral backbone angles, mesoscopic sugar ring centre landmarks. Our analysis relates concentrated mesoscopic scale neighbourhoods to microscopic scale clusters, correcting within-suite-backbone-to-backbone clashes exploiting angular shape and size-and-shape Fréchet means. Validation shows that learned classes highly correspond with literature clusters and reconstructions are well within physical resolution. We illustrate the power of our method using cutting-edge SARS-CoV-2 RNA.

Keywords: angular shape analysis, clash correction, frameshift stimulation element, Fréchet and Procrustes means, mesoscopic shape and microscopic shape, size-and-shape space

1 Introduction

Understanding the structure of active biomolecules is ever more important for maintaining and improving human health, as has been summarised by Schlick and Pyle (2017). In particular, this pertains to RNA molecules in designing drugs which target-specific structures (see Batool et al., 2019), as recently impressively demonstrated by the worldwide effort confronting the SARS-CoV-2 (severe acute respiratory syndrome) virus responsible for the COVID-19 (corona virus disease) pandemic (see Croll et al., 2021).

Extracting RNA primary structure (sequencing) is nowadays fairly well feasible using currently available gene sequencing technology (e.g., Wang et al., 2009). Predicting the 3D structure (helices, etc.,) from that, however, is a still unsolved fundamental problem (e.g., Schlick & Pyle, 2017). Although elaborate methods such as X-ray crystallography and cryo-EM (cryogenic electron microscopy) are used that determine spatial electron densities—and from these densities individual atom positions can be inferred—frequently, the inferred molecular structures contain the so-called clashes as detailed by Murray et al. (2003), Chen et al. (2010) and others.

Definition 1.1 A **clash** is a forbidden molecular configuration, where two atoms are reconstructed closer to each other than is chemically possible.

Received: February 21, 2022. **Revised:** June 21, 2022. **Accepted:** December 20, 2022

In the case of RNA, clashes most relevant and most difficult to correct are between atoms along the backbone (main chain), in particular when single hydrogen atoms not contributing to electron densities are added to inferred structures (see Figure 1); a detailed discussion is given in Murray et al. (2003).

In order to correct such clashes, methods from *molecular dynamics* are usually employed: Simulated atoms are allowed to fluctuate into positions of minimal energy, following approximations of the laws of biophysical chemistry (e.g., Chou et al., 2013a). For RNA molecules, these simulations are highly computation intensive due to the large variability of RNA shape. If local and not global energy minima are achieved, thus corrected molecules may still feature clashes and their geometries may be outliers in comparison to clash-free geometries (e.g., Richardson et al., 2018). Online Supplementary Material, Supplement D briefly sketches the state of the art correction method ERRASER by Chou et al. (2013a) and details this observation.

As most RNA backbone clashes appear within ***suites*** (the section from one sugar ring to the next, e.g., Murray et al., 2003, see Figure 1 and Notation 3.1, Section 3), we therefore apply our method to ***within-suite-backbone-to-backbone*** clashes here; although it can be more generally applied. For the scope of this article, we call here suites ***clash free*** if they are free of within-suite-backbone-to-backbone clashes. We analyse the RNA backbone simultaneously at two scales exploiting their interdependence as follows.

We work on two levels, the microscopic (atomic level) and the mesoscopic (level of objects). At the ***microscopic*** scale we model the backbone of suites by tuples of 7 dihedral angles, each between 0 and 2π from the backbone atoms, giving a data point on the seven-dimensional torus $\mathbb{T}^7$. We are thus working on a form of shape analysis from angles (angular shape analysis). At the ***mesoscopic*** scale, we model k suites before and k suites after a central suite of concern represented by $2k+2$ *pseudo-landmarks*, the centres of sugar rings, see Figure 2. Our interest will be the ***size-and-shape*** (see Dryden & Mardia, 2016) of these landmarks. Setting $k=2$. That is, six landmarks in total (which depicts roughly a half helix turn), our data analysis leads to the conclusion that for clash-free data, concentrated clusters at mesoscopic scale correspond to clusters at microscopic scale.

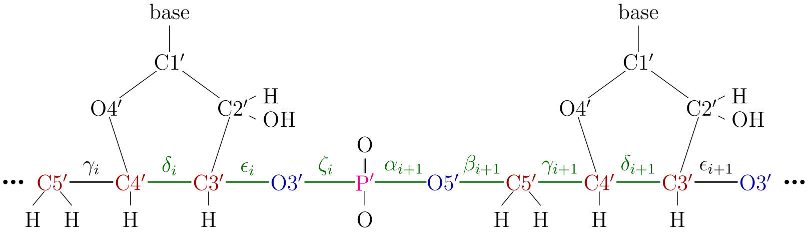

Figure 1. 2D scheme of backbone suite number i with 7 dihedral angles (see Figure 6) δ_i, ϵ_i, ζ_i, α_{i+1}, β_{i+1}, γ_{i+1}, δ_{i+1} describing the suite's 3D structure.

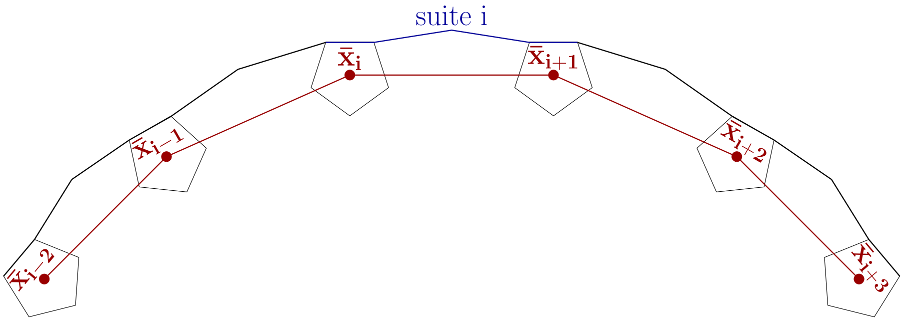

Figure 2. The mesoscopic shape for $k=2$ centred at the ith suite is determined by the six centres of the sugar rings $\bar{\mathbf{x}}_{\mathbf{i-2}}, \ldots, \bar{\mathbf{x}}_{\mathbf{i+3}}$. Their connecting backbones give 5 suites, two before and two after suite i.

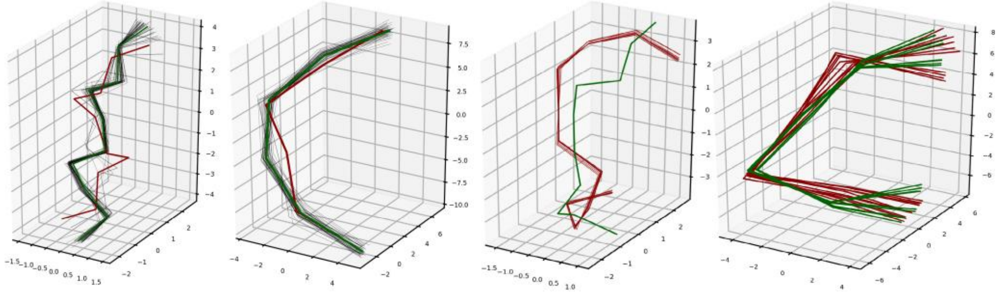

Figure 3. Left two panels: A clashing suite (red) (from benchmark file 1f8v, (Tang et al., 2001), see Online Supplementary Material, Table 16) with its clash-free neighbours (black) and proposed clash-free correction (green) at microscopic scale (left) and mesoscopic scale (left centre). Right two panels: Ten proposed reconstructions (red) by Zhang et al. (2021), which are all clashing, for Suite 28/29 (cf. Figure 5) connecting two helical segments in the frameshift stimulation element of SARS-CoV-2 and our 10 clash-free corrections (green) at microscopic scale (centre right) and at mesoscopic scale (right).

This correspondence is at the heart of our two-scale correction method: Since we aim to correct potential errors at the microscopic scale, we first learn classes of clash-free microscopic shapes by clustering a benchmark data set of clash-free data at the microscopic scale. As illustrated in the left two panels of Figure 3 we provide a data-driven correction (green) for a clash suite (red) by a Fréchet mean on the torus at the microscopic scale (left panel) within a specific class of clash-free suites (grey) from Tang et al. (2001) (file 1f8v, see Online Supplementary Material, Table 16). To determine the class which is used for microscopic structure correction, we leverage the corresponding mesoscopic shape describing the geometry of the RNA strand in proximity to the clash suites by determining a set of closest mesoscopic shapes to the mesoscopic shape containing the clash suite. We then consider the microscopic suite shapes corresponding to these nearby mesoscopic shapes and determine the class which dominates this set (centre left panel, same colours). At the mesoscopic scale, our correction (green) is the geodesic projection of the corresponding Procrustes mean to the mesoscopic shape featuring the same endpoints and the length of the corrected suite. Typically, our correction at mesoscopic scale requires only a few moderate shifts of sugar centres (left centre, see also Figure 13, right panel).

We validate our correction method based on the interdependence of clash-free RNA backbone shape at the two scales (microscopic and mesoscopic) by showing that the corrections proposed stay well below resolution level on the benchmark data. We also validate our classification by comparison with a suite clustering proposed by Richardson et al. (2008) who investigated a larger data set (comprising about twice as many suites than our benchmark data set): The classes we propose correspond well to their clusters, where some of our classes comprise several of their clusters.

In application, we propose clash-free corrections for ten structure proposals from Zhang et al. (2021) for two suites of the *frameshift stimulation element* (which facilitates decoding more than one protein from a single RNA strand) of SARS-CoV-2 which are difficult to reconstruct, and for which, to the best knowledge of the authors, there are no consistent 3D structures known to date. Our method proposes structure which are strikingly consistent, and by design, are clash free. For one of the two suites, the situation is exemplified below in the two right panels of Figure 3: For each of the 10 clashing proposals (red), at mesoscopic scale (right panel) we propose clash-free corrections (green) and at microscopic scale (centre right panel, same colours) our corrections agree nearly unambiguously.

Our paper is structured as follows. First, we introduce the two shape spaces: the torus (angular shape space) describing the RNA backbone uniquely at microscopic (atomic) scale and the size-and-shape space describing the RNA backbone at mesoscopic scale. Then follows the concept of Fréchet means used at both scales for clash correction. At mesoscopic scale (here Fréchet means are Procrustes means), we provide a novel projection (preserving constraints from the original mesoscopic shape and its microscopic correction) for the Procrustes mean. In Section 3, we link the 3D RNA backbone structure at two scales to our two shape spaces, overview clash detection and provide our benchmark data. Section 4 proposes our multiscale RNA backbone correction method, first introducing learned classes from the clash free benchmark data and validating them. We then present the interdependence of clash-free RNA backbone shape at the two scales (microscopic and mesoscopic) and detail how we exploit this for the new method proposed and validate it. Finally, we apply our method to the correction of the RNA backbone of SARS-CoV-2. In Section 6, we discuss further potentials of our method, in particular how ***multiscale shape analysis*** can be more fully developed and how it could be used to complement existing reconstruction methods for long stranded biomolecules based on molecular dynamics.

While we measure angles in radians, for instant comparison with other research in this area, some of the Figures report results in degrees.

Finally we list the content of our online supplementary material, containing all code and all data, as well as further data analysis and an overview of the MINT-AGE algorithm from Mardia et al. (2022).

2 Tools from shape analysis

For Fréchet means defined in Section 2.3, we will need appropriate distances for the microscopic and mesoscopic scale which we now give in Sections 2.1 and 2.2, respectively. For the mesoscopic scale, we develop in Section 2.4 a geodesic projection since we have to impose suitable geometric constraints.

2.1 The torus for microscopic scale

The one-dimensional *torus* is

$$\mathbb{T} := [0, 2\pi]/\sim$$

where '$\sim$' denotes that 0 and 2π are identified. It is a metric space with canonical distance

$$d_{\mathbb{T}}(\phi, \psi) = \min\{|\phi - \psi|, 2\pi - |\phi - \psi|\}, \quad \phi, \psi \in \mathbb{T}. \tag{1}$$

The canonical product of m one-dimensional tori is the m-dimensional torus $\mathbb{T}^m$ with the canonical product distance given by

$$d_{\mathbb{T}^m}(\phi, \psi) = \sqrt{\sum_{j=1}^{m} d(\phi_j, \psi_j)^2}, \tag{2}$$

for $\phi = (\phi_1, \ldots, \phi_m)$, $\psi = (\psi_1, \ldots, \psi_m) \in \mathbb{T}^m$. Several authors have studied data on the torus, especially representing large biomolecules, and developed specialised methods, including (AlQuraishi, 2019; Altis et al., 2008; Eltzner et al., 2018; Kent & Mardia, 2009; Parsons et al., 2005; Sargsyan et al., 2012; Zoubouloglou et al., 2021).

2.2 Size-and-shape for mesoscopic scale

We describe a landmark configuration matrix $X = (\mathbf{x}_1, \ldots, \mathbf{x}_m) \in \mathbb{R}^{3\times m}$ encoding $m \in \mathbb{N}$, three-dimensional landmark positions $\mathbf{x}_i \in \mathbb{R}^3$, $i = 1, \ldots, m$ by its *size-and-shape* as follows, see Dryden and Mardia (2016): Proper (i.e., orientation preserving) Euclidean transformations comprising rotations and translations $T = (R, v) \in \mathrm{SO}(3) \times \mathbb{R}^3$ act on X columnwise via

$$T.X := (R\mathbf{x}_1 + v, \ldots, R\mathbf{x}_m + v).$$

Then

$$S\Sigma_3^m := \{[X] : X \in \mathbb{R}^{3\times m}\} \quad \text{where } [X] := \{T.X : T \in \mathrm{SO}(3) \times \mathbb{R}^3\} \tag{3}$$

is the *size-and-shape space* which is equipped with the quotient distance, also called *Procrustes distance*

$$d_\Sigma([X],[Y]) := \min_{T \in \mathrm{SO}(3)\times\mathbb{R}^3} \|X - T.Y\| \tag{4}$$

with the standard Frobenius norm on $\mathbb{R}^{3\times m}$. We say that X and Y are in *optimal position* if

$$d_\Sigma([X],[Y]) = \|X - Y\|.$$

Taking derivatives and using a singular value decomposition (SVD) it follows at once that configurations X, Y in optimal position have coinciding mean landmarks with symmetric YX^T (e.g., Dryden & Mardia, 2016, Result 7.1). For this reason, we assume that all landmark configurations are *centred*, i.e., their landmarks vectors add up to zero. Optimal positioning is then conveyed by rotations $R \in \mathrm{SO}(3)$ only, i.e., RY is in optimal position to X if $R = VSU^T$ with a suitable diagonal matrix S with entries in $\{-1, 1\}$ and a SVD $YX^T = UDV^T$ (here U, V are orthogonal, D is diagonal with non-negative entries).

2.3 Fréchet means for both scales

Definition 2.1 For data $X_1, \ldots, X_n \in M$ on an arbitrary metric space (M, d), define their *Fréchet means* by

$$\operatorname*{argmin}_{X\in M} \sum_{j=1}^{n} d(X, X_i)^2.$$

The Fréchet mean is a generalisation of the classical Euclidean mean. On complete spaces, Fréchet means exist, and on manifolds, if samples are drawn from continuous distributions, they are almost surely unique (see Arnaudon & Miclo, 2014). On stratified quotient spaces, such as size-and-shape space for 3D configurations, they lie on the manifold part (the top-dimensional dense stratum) if the manifold part is assumed with positive probability (see Huckemann, 2012).

On $S\Sigma_3^m$, Fréchet means defined by Procrustes distance are also called *Procrustes means*. On $\mathbb{T}^m$ we call them *torus means*.

2.4 Geodesic projection to constrained size-and-shape

Our CLEAN MINT-AGE Algorithm in Section 4.3.1 corrects clashes not only at atomic suite (microscopic) scale but also at mesoscopic scale. The corrected mesoscopic shape m_{τ_c} in Section 4.3.1 features two constraints. The first one sets the distance between its first and last landmark to the corresponding distance of the original mesoscopic shape, thus assuring its fit into a larger RNA strand. The second one sets the distance between its two central landmarks to the length of the corrected suite, assuring the fit of the latter into the former.

With more general future applications in mind, assume that the distances between $r \in \mathbb{N}$, $(2 \le 2r \le m)$ landmark pairs are constants $a_1, \ldots, a_r > 0$. With a permutation σ of $(1, \ldots, m)$ we may assume that landmark $\sigma(j)$ is paired with landmark $\sigma(j + r)$ for $j = 1, \ldots, r$ while landmarks $\sigma(j)$ for $2r < j \le m$ (if $2r < m$) are unconstrained.

Definition 2.2 Let $r \in \mathbb{N}$ with $2r \le m$, $a := (a_1, \ldots, a_r)$ with $a_1, \ldots, a_r > 0$ and σ be a permutation of $(1, \ldots, m)$. Then the *constrained-size-and-shape space* is given by

$$S\Sigma_3^m(\sigma, a) : = \{[Y] \in S\Sigma_3^m : Y = (y_1, \ldots, y_m) \in \mathbb{R}^{3\times m},$$
$$\|y_{\sigma(j)} - y_{\sigma(j+r)}\| = a_j \text{ for } j = 1, \ldots, r\}.$$

An orthogonal projection from Σ_3^m to $\Sigma_3^m(\sigma, a)$ can be given explicitly as the following theorem teaches.

Theorem 2.3 Let $r \in \mathbb{N}$ with $2r \leq m$, $a = (a_1, \ldots, a_r)$ with $a_1, \ldots, a_r > 0$, $[Z] \in S\Sigma_3^m$ with centred $Z \in (z_1, \ldots, z_m)$, i.e., $z_1 + \cdots + z_m = 0$ and σ be a permutation of $(1, \ldots, m)$. Then $Y^* = (y_1^*, \ldots, y_m^*)$ with

$$\begin{aligned} y^*_{\sigma(j)} &= \beta_{\sigma(j)} z'_{\sigma(j)} + (1 - \beta_{\sigma(j)}) z'_{\sigma(j+r)}, \\ y^*_{\sigma(j+r)} &= (1 - \beta_{\sigma(j)}) z'_{\sigma(j)} + \beta_{\sigma(j)} z'_{\sigma(j+r)}, \quad \text{with} \\ \beta_{\sigma(j)} &= \frac{1}{2}\left(1 + \frac{a_j}{\|z'_{\sigma(j)} + z'_{\sigma(j+r)}\|}\right), \end{aligned}$$

for $j = 1, \ldots, r$ where we set ${z'}_{\sigma(j)} := z_{\sigma(j)}$, ${z'}_{\sigma(j+r)} := z_{\sigma(j+r)}$, if $z_{\sigma(j)} \neq z_{\sigma(j+r)}$ and ${z'}_{\sigma(j)} := z_{\sigma(j)} + v_j$, ${z'}_{\sigma(j+r)} := z_{\sigma(j)} - v_j$ if $z_{\sigma(j)} = z_{\sigma(j+r)}$ with an arbitrary non-zero vector $v_j \in \mathbb{R}^{3\times m}$, and, furthermore

$$y^*_{\sigma(j)} = z_{\sigma(j)} \quad \text{for } j = 2r + 1, \ldots, m,$$

gives an orthogonal projection

$$[Y^*] \in \underset{[Y] \in S\Sigma_3^m(\sigma, a)}{\operatorname{argmin}} d_{S\Sigma_3^m}([Z], [Y]).$$

The orthogonal projection is unique if $z_{\sigma(j)} \neq z_{\sigma(j+r)}$ for all $j = 1, \ldots, r$.

Proof. W.l.o.g. assume that σ is the identity. Furthermore, note that by construction Y^* is centred as Z is centred.

Every orthogonal projection is a minimiser of the Lagrange function

$$\mathcal{L}(Y, \lambda_1, \ldots, \lambda_r) = \|Y - Z\|^2 + \sum_{j=1}^{r} \lambda_j (\|y_{j+r} - y_j\|^2 - a_j^2)$$

incorporating proximity of $Y = (y_1, \ldots, y_m)$ to Z and the constraining conditions. All of its critical points Y^* are determined by the equations

$$y_j^* - z_j = \lambda_j (y^*_{j+r} - y^*_j) \quad \text{for } j = 1, \ldots, r \tag{5}$$

$$\begin{aligned} y^*_{j+r} - z_{j+r} &= -\lambda_j (y^*_{j+r} - y^*_j) \quad \text{for } j = 1, \ldots, r \\ y^*_j &= z_j \quad \text{for } j \in \{2r + 1, \ldots, m\}. \end{aligned} \tag{6}$$

Notably, the last equations yield the unique minimisers of the non-constrained landmarks. Now fix $j \in \{1, \ldots, r\}$ and subtract (6) from (5) to obtain

$$(y_j^* - y^*_{j+r})(1 + 2\lambda_j) = z_j - z_{j+r}. \tag{7}$$

If $z_j \neq z_{j+r}$ then (5) yields

$$y_j^* = z_j - \frac{\lambda_j}{1 + 2\lambda_j}(z_j - z_{j+r}),$$

i.e., with $\beta_j = (1 + \lambda_j)/(1 + 2\lambda_j)$

$$y_j^* = \beta_j z_j + (1 - \beta_j) z_{j+r}, \tag{8}$$

and similarly, (6) yields

$$y_{j+r}^* = z_{j+r} + \frac{\lambda_j}{1 + 2\lambda_j}(z_j - z_{j+r}),$$

i.e.,

$$y_{j+r}^* = (1 - \beta_j) z_j + \beta_j z_{j+r}. \tag{9}$$

This implies at once that

$$\begin{aligned} \|y_j^* - z_j\|^2 + \|y_{j+r}^* - z_{j+r}\|^2 &= 2(1 - \beta_j)^2 \|z_j - z_{j+r}\|^2 \\ &= \frac{2\lambda_j^2}{(1 + 2\lambda_j)^2} \|z_j - z_{j+r}\|^2. \end{aligned} \tag{10}$$

In order to determine λ_j we exploit the constraining condition to obtain from (7) that $|1 + 2\lambda_j| = \|z_{j+r} - z_j\|/a_j$. The cases of $1 + 2\lambda_j > 0$ and $1 + 2\lambda_j < 0$ correspond to

$$\lambda_j = \frac{1}{2}\left(\frac{\|z_{j+r} - z_j\|}{a_j} - 1\right) \quad \text{and} \quad \lambda_j = -\frac{1}{2}\left(\frac{\|z_{j+r} - z_j\|}{a_j} + 1\right)$$

respectively, so that, taking into account (10), $\mathcal{L}$ assumes the minimal value for the positive branch yielding

$$\beta_j = \frac{1}{2}\left(1 + \frac{a_j}{\|z_{j+r} - z_j\|}\right),$$

as asserted. Moreover, then the above equations (8) and (9) yield the asserted landmarks for y_j^* and y_{j+r}^* in case of $z_j \neq z_{j+r}$.

If $z_j = z_{j+r}$ adding (6) to (5) yields

$$y_j^* + y_{j+r}^* = \frac{z_j}{2},$$

which, taking into account the constraining condition, is solved by

$$y_j^* = z_j + a_j \frac{v_j}{2\|v_j\|}, \quad y_{j+r}^* = z_j - a_j \frac{v_j}{2\|v_j\|}$$

with an arbitrary nonzero vector $v_j \in \mathbb{R}^{3\times m}$. Then the above argument, after replacing z_j with $z'_j := z_j + v_j$ and $z'_{j+r} := z_j - v_j$ above in (5) and (6), yields the asserted equations.

We note that we have indeed found a minimum, for we can reparametrize the matrix Y by arbitrary $Y' := (y_1, \ldots, y_r, y_{2r+1}, \ldots, y_m) \in \mathbb{R}^{3\times(m-r)}$, and by $(w_1, \ldots w_r)$, each w_j arbitrary on the compact sphere $\{w \in \mathbb{R}^3 : \|w\| = 1\}$ which model the constraining conditions via $y_{j+r} = y_j + a_j\, w_j$ for $j = 1, \ldots, r$. Along the columns of Y', there is a unique minimum and along each of the w_j $(j = 1, 2)$ there is a maximum and a minimum given by the two choices of λ_j as detailed above and illustrated in Figure 4, and each such minimum is unique if $z_j \neq z_{j+r}$.

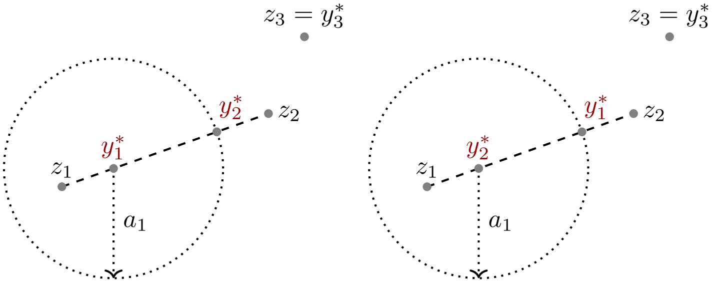

Figure 4. Planar representation of the o.g. projection of $Z = (z_1, z_2, z_3)$ to the constraint $\|y_1 - y_2\| = a_1$. Left: The global minimum determined by $1 + 2\lambda_1 > 0$ is attained for y_1^* and y_2^* balanced between z_1 and z_2. Notably, fixing $y_1 = y_1^*$ the constrained y_2 is confined to a sphere of radius a_1, centred at y_1^*. Right: Swapping the minimal y_1^* and y_2^* from the left side corresponds to $1 + 2\lambda_1 < 0$. Fixing $y_2 = y_2^*$, the constrained y_1 lies on a sphere of radius a_1, centred at y_2^*, for which $y_1 = y_1^*$ produces a local maximum.

Finally, we claim that Y^* is already in optimal position to Z. In fact it suffices to see this for two landmarks only z_j, z_{j+r} $(1 \le j \le r)$ and y_j^*, y_{j+r}^* from (8) and (9). Indeed, in case of $z_j \neq z_{j+r}$, with the 3×3 unit matrix I, minimising

$$\|z_j - Ry_j^*\|^2 + \|z_{j+r} - Ry_{j+r}^*\|^2 = \|(I - \beta R)z_j - (1-\beta)Rz_{j+r}\|^2 + \|(I - \beta R)z_{j+r} - (1-\beta)Rz_j\|^2$$

over $R \in SO(m)$ can be cast into the two-dimensional complex problem with $z = z_j$, $w = z_{j+r} \in \mathbb{C}$, $\beta = \beta_j > 1/2$ minimising

$$\begin{aligned}&|(1 - \beta\,\mathrm{e}^{i\alpha})z - (1-\beta)\,\mathrm{e}^{i\alpha}w|^2 + |(1 - \beta\,\mathrm{e}^{i\alpha})w - (1-\beta)\,\mathrm{e}^{i\alpha}z|^2 \\ &\quad = (|z|^2 + |w|^2)(1 + \beta^2 - 2\beta\cos\alpha + (1-\beta)^2) - 4(1-\beta)\,\mathrm{Re}(\bar{z}\,w)(\cos\alpha - \beta)\end{aligned}$$

over $\alpha \in [0, 2\pi)$. Due to $0 \le |z \pm w|^2 = |z|^2 + |w|^2 \pm 2\mathrm{Re}(\bar{z}\,w)$ and $\beta > 1/2$ this is minimised for $\alpha = 0$, corresponding to $R = I$ above.

In case of $z_j = z = z_{j+r}$, with arbitrary but fixed $v_j \in \mathbb{R}^3$, $\|v_j\| = 1$ such that $y_j^* = z + a_j v_j/2$ and $y_{j+r}^* = z - a_j v_j/2$, as above, we have similarly for $R \in SO(3)$ that

$$\begin{aligned}\|z_j - Ry_j^*\|^2 + \|z_{j+r} - Ry_{j+r}^*\|^2 &= \|z - R(z + a_j v_j/2)\|^2 + \|z - R(z - a_j v_j/2)\|^2 \\ &= 2\|z - Rz\|^2 + \frac{a_j^2}{2},\end{aligned}$$

which is minimised by $R = I$. □

Remark 2.4 The case $z_j = z_{j+r}$ has been discussed for exhaustive mathematical treatment. In the application in Section 4.3.1, this only happens if the neighbourhoods in the classes learned feature degenerate Procrustes means, a clear sign that the learning algorithm failed. In this case, we suggest to re-evaluate learned classes, rather than choosing any v_j of suitable length.

3 Multiscale modelling of RNA backbone geometry, clash detection, and data sets

Ribonucleic acid (RNA) molecules are composed of repeating elements called *nucleotides* and each nucleotide is composed of three building blocks, see Watson et al. (2004) and Figure 1: A sugar ring called *ribose* comprising five carbon atoms, one of four possible *nucleobases* which is attached to the ribose at the C1′ position and a *phosphate group* connected to the ribose ring at the O5′ atom. The single nucleotides are connected by their O3′ atoms to the next phosphate group to form long RNA chains.

3.1 RNA folding

In contrast to DNA which usually forms a double helix of complementary strands, in principle, RNA is single stranded and the form of its ribose (which is not ‘desoxy’ as in DNA, i.e., it has an additional hydroxyl group) allows for complex folding structures. Figure 5 shows helical structures followed by mismatching sites: a *hairpin* in a 2D schematic and the 3D structure of the *frameshift stimulation element* of the SARS-CoV-2 genome proposed by Zhang et al. (2021). Its 2D schematic is depicted in the first panel of Figure 15.

3.2 Multiscale modelling

In this section, we describe the two scales modelled. Their surprising interaction which has lead to the two Hypotheses 4.1 and 4.2 underlying our method is detailed in Section 4.2.

On a *microscopic* scale, nucleotides are either studied as *suites*, i.e., from one sugar to the next, or as *residues*, i.e., from one phosphate to the next, e.g., Murray et al. (2003), Jain et al. (2015). As clashes often occur between neighbouring residues but within same suites, cf. Murray et al. (2003), for our analysis, we use suites. Indexing, however, is usually done on residue level, so that within a single suite, atom indices change, cf. Figure 1. For the dihedral angles of concern, Figure 6 lists the four consecutive atoms, defining the respective dihedral angle of the bond between the two central atoms.

On the *mesoscopic* scale, we additionally take the coordinates of the k preceding and k succeeding sugar rings into account. This can be seen as an intermediate scale between the microscopic suite scale and the macroscopic scale of a whole RNA strand.

Notation 3.1 We consider a connected RNA strand with $N \in \mathbb{N}$ consecutive nucleotides indexed by $i \in \{1, \ldots, N\}$.

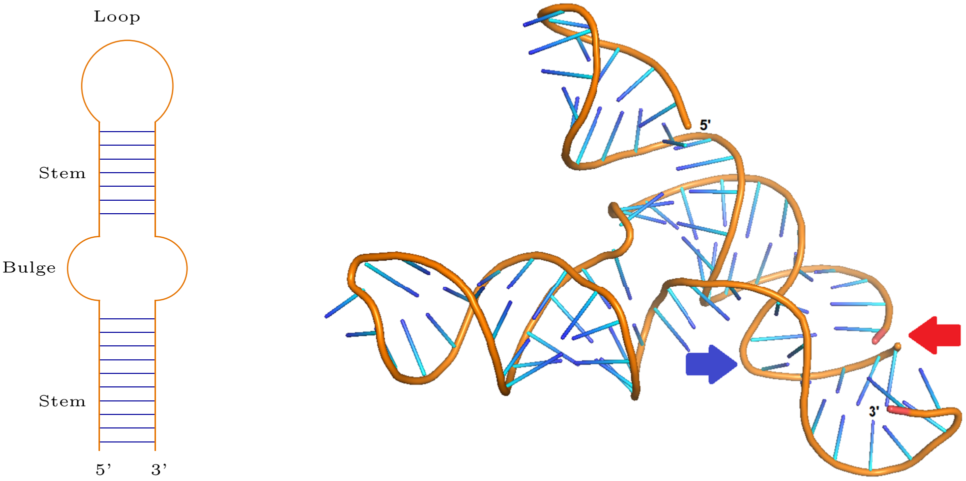

Figure 5. Left: 2D schematic of the common *hairpin* structure: Double helices *(stems)* formed by bindings between matching nucleobases (blue) are followed by *mismatching* nucleobases *(bulges)*, not depicted, and a terminating mismatching site *(loop)*. Orientation is conveyed by the 5′ and 3′ ends. Right: One out of 10 proposed 3D RNA structures of the SARS-CoV-2 frameshift stimulation element by Zhang et al. (2021), graphically reproduced with PyMOL (Schrödinger, 2015) with backbone (orange) and nucleobases (blue), yielding helical structures whenever the latter point to each other. Arrows indicate suites with problematic (blue arrow, Suite 2 determined by Residues 33/34) and non-connected backbone (red arrow, Suite 1 determined by Residues 28/29) proposals discussed in Figures 3 and 15.

angle	atom bonds
α	O3’ - P - O5’ - C5’
β	P - O5’ - C5’ - C4’
γ	O5’ - C5’ - C4’ - C3’
δ	C5’ - C4’ - C3’ - O3’
ϵ	C4’ - C3’ - O3’ - P
ζ	C3’ - O3’ - P - O5’

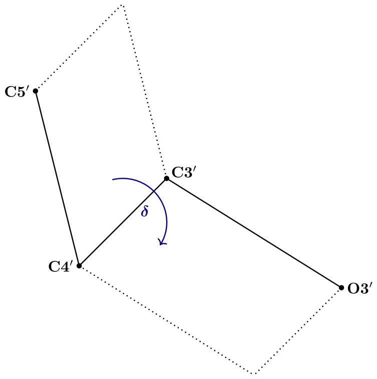

Figure 6. Left: Names (first column) of dihedral angles along the two central atoms of the four atoms involved (second column), see Figure 1. Right: The dihedral angle δ of the bond between the atoms C4′ and C3′ is the directed angle between the plane spanned by the atoms C5′, C4′, C3′ and the plane spanned by C4′, C3′, O3′. More precisely, it is the angle determined by turning the vector normal to the plane spanned by C3′, C4′, C5′ to the vector normal to the plane spanned by O3′, C3′, C4′ (with fixed orientation of normals determined by the order of spanning points).

Microscopic scale: The ith *suite* comprises the RNA region between a $C5_i'$ atom and the second next $O3'$ atom labelled $O3'_{i+1}$ and the *backbone shape* of the suite is described by the seven dihedral angles $(\delta_i, \epsilon_i, \zeta_i, \alpha_{i+1}, \beta_{i+1}, \gamma_{i+1}, \delta_{i+1}) \in \mathbb{T}^7$ for $i = 1, \ldots, N-1$, cf. Figure 1. **Mesoscopic scale:** As each nucleotide comes with a sugar ring formed by the atoms $C1'_i$, $C2'_i$, $C3'_i$, $C4'_i$ and $O4'_i$ (see Figure 1), denoting their centres of gravity (i.e., average location) with $\bar{\mathbf{x}}_i$, for all $i = k+1, \ldots, N-k-1$, the *mesoscopic strand* corresponding to the ith suite is the *configuration matrix* $X^{(i)} = (\bar{\mathbf{x}}_{\mathbf{i-k}}, \bar{\mathbf{x}}_{\mathbf{i-k+1}}, \ldots, \bar{\mathbf{x}}_{\mathbf{i+k+1}}) \in \mathbb{R}^{3\times(2k+2)}$. Its size-and-shape in $S\Sigma_3^{2k+2}$ is called its *mesoscopic shape*.

Indeed, geometric suite variability is solely governed by the dihedral angles, since bond lengths (distances between two consecutive atoms) and bond angles (angles between three consecutive atoms) are approximately constant due to the laws of chemistry, see e.g., Watson et al. (2004). In consequence, the geometry of the ith suite is described, up to a proper Euclidean transformation (translation and rotation), by an element of the seven-dimensional torus $\mathbb{T}^7$ given by its seven dihedral angles.

Since distances between two neighbouring sugar rings and angles between three consecutive sugar rings vary due to folding at microscopic scale, see Figure 7, dihedral angles defined by four consecutive sugar rings are not sufficient to completely define the geometry of mesoscopic strands up to proper Euclidean transformations. The size-and-shape representation, modelling geometric landmark configurations determined by central positions of sugar rings modulo translation and rotation, however, suffices.

Remark 3.2 For the mesoscopic strands, we include the sugar ring centres of the $k = 2$ suites preceding and the $k = 2$ suites following the suite of concern, cf. Figure 2. This choice of k presents a trade-off, since a small k emphasises the central, potentially faulty, suite and a large k leads to a great variety of shapes at transitions between secondary structure elements. For a given mesoscopic shape, this reduces the number of potentially similar mesoscopic shapes. Empirically, $k = 2$ yields a good balance between these two effects by modelling the local geometry at an intermediate (mesoscopic) scale. On the side of biochemistry, the $5 + 1 = 6$ bases from the $2k + 1 = 5$ suites correspond roughly to the number of bases involved in a half helix turn, see e.g., Watson et al. (2004). For future applications, we anticipate that involving more scales by suitably choosing larger k will prove useful.

We only work with suites that have a corresponding mesoscopic strand, i.e., we exclude the two suites at the end of an RNA strand.

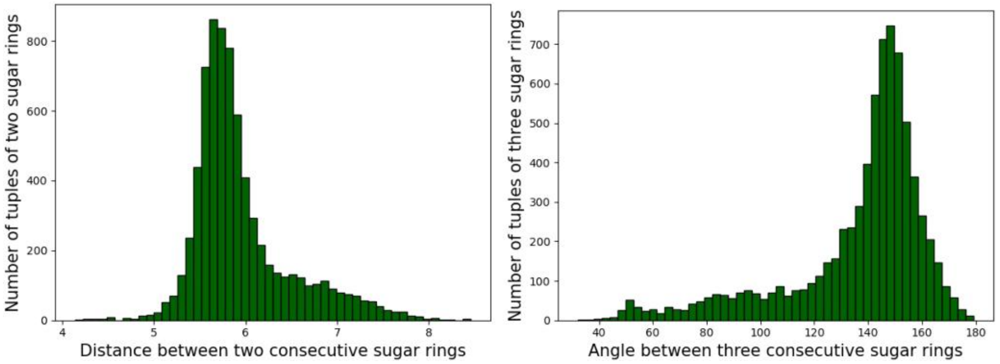

Figure 7. Histograms of the distribution of distances between two successive sugar ring centres in Å (left) and of the distribution of angles in degrees spanned by three successive sugar ring centres (right).

Definition 3.3 For an RNA strand of length $N \geq 2k + 2$ the suites numbered $i = k + 1, \ldots, N - k - 1$ are called *admissible*, so that every admissible suite $\mathfrak{a}$ has a mesoscopic shape $m_{\mathfrak{a}} \in S\Sigma_3^{2k+2}$ and vice versa.

Definition 3.4 We call a suite a ***clash suite*** if two of its backbone atoms (including associated hydrogen atoms and oxygen atoms associated with the phosphate) clash with each other. All other suites that have $2k = 4$ neighbouring non-clash suites (i.e., their mesoscopic strands have no within-suite-backbone-to-backbone clashes) are called ***clash free***.

3.3 Cryo-EM, *X*-ray crystallography and clash detection

Cryo-EM (cryogenic electron microscopy) and *X*-ray crystallography are popular methods to determine atomic positions in RNA, protein, and similar biomolecular structures, cf. Jain et al. (2015). For the former, molecules are shock frosted and subjected to electron microscopy. For the latter, using a suitable substrate, molecules are crystallized and subjected to *X*-ray imaging. The resolution of *X*-ray crystallography is defined as the smallest distance of two objects such that their diffraction patterns can be separated. In cryo-EM, resolution has been defined in various ways, usually via properties of the Fourier transformed electron density, in order to be comparable to the resolution values given for *X*-ray crystallography measurements. For a review, see Liao and Frank (2010). From different angles, via inverse Fourier transforms, the electron density can be reconstructed and, in principle, density peaks correspond to atom positions. Figure 8 shows exemplary level surfaces of electron densities with estimated atom positions.

At a resolution of 2.5–4 Å, which is typical for large RNA strands, base pairings can be predicted well and phosphates are well identified by strong peaks of density (Jain et al., 2015). It is, however, more challenging to precisely estimate single atom positions along the backbone, see for example (Murray et al., 2003). In addition, structural disorder due to crystallization and thermal oscillation contribute to uncertainties.

Since it is computationally not feasible to include the positions of all atoms and a full quantum chemical treatment into the fitting, the ambiguities in the measured density occasionally result in incompatible reconstructed atom positions. Indeed, our benchmark data set contains approximately 2.5% clash suites.

The PHENIX (Python-based Hierarchical ENvironment for Integrated Xtallography) software by Liebschner et al. (2019) provides validation tools that detect such errors. Since hydrogen atoms are not visible in the electron density measurements (H-atoms contain only one electron which is shifted to the covalent-bond partner atom), first, the PHENIX tool `phenix.reduce` adds the hydrogen atoms. Then, `phenix.probe` performs an all-atom contact analysis (Word et al., 1999), which declares atoms that are not bonded to each other as a *clash* if they are closer together

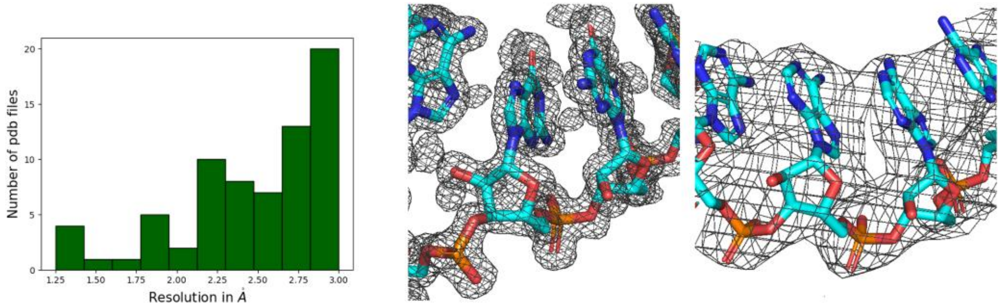

Figure 8. Left: Histogram of *X*-ray crystallography resolutions in the benchmark data set from Section 3.4 below. Middle and right: reconstructed RNA structure and electron density contour surface created with PyMOL at level of one σ, see Schrödinger (2015), at resolution 1.6 Å (middle, from benchmark file 1csl, (Ippolito & Steitz, 2000), see Online Supplementary Material, Table 16 in the supplement) and at resolution 3 Å (right, from benchmark file 1f8v, (Tang et al., 2001), see Online Supplementary Material, Table 16).

than is physically possible (i.e., if van der Waals shells overlap by more than 0.4 Å). For each PDB file, `phenix.clashscore` generates a list of all clashes. From all of the different types of clashes detected, in this work we are only concerned with within-suite-backbone-to-backbone clashes as in Definition 3.4.

3.4 The benchmark, training, and test data sets

In our applications, we analyse a subset of a classical RNA data set. The classical RNA data set comprises 8,665 suites, carefully selected for high experimental *X*-ray precision (of 3 Å = 0.3 nm) by Duarte and Pyle (1998), Wadley et al. (2007) and analysed by them and by others, for example (Eltzner et al., 2018; Murray et al., 2003; Richardson et al., 2008). The data originate from 71 different measurements and the atomic positions of each measurement have been stored in the *PDB* format of a *protein data bank* file, online at the Protein Data Bank, see Berman et al. (2000). More details on the PDB files can be found in Online Supplementary Material, Table 16 of Supplement A.

From this classical data set, we consider the 7,648 admissible suites (which have an associated mesoscopic strand, see Definition 3.3) and call this data set the *benchmark data set*.

Applying PHENIX as detailed in Section 3.3 to the benchmark data set, we obtain 5,957 clash-free suites that also have clash-free mesoscopic strands (see Definition 3.4) and these form the *benchmark training data set* $\mathfrak{T}$. Online Supplementary Material, Figure 17 gives a scatterplot at microscopic scale for all pairs of the seven dihedral angles.

From the remaining suites, we chose those suites that feature within-suite-backbone-to-backbone clashes, forming the *benchmark test data set* $\mathfrak{C}$, containing 198 suites.

As our purpose lies in demonstrating our methods rather than correcting all clashes, all other suites (e.g., those not themselves clashing but featuring clashes in their mesoscopic strands) are disregarded in our analysis.

4 CLEAN-MINT-AGE

After classifying clash-free suites by the MINT-AGE algorithm (Mardia et al., 2022) from the benchmark training data set, we validate the classes obtained by comparing with the outcome of the clustering method by Richardson et al. (2008). Motivating our multiscale approach by analysing clusters at two scales, then we propose and validate the CLEAN method classifying suites exploiting the observed relationship between the two scales.

4.1 Microscopic classification and its validation

We apply the non-supervised cluster learning method from Mardia et al. (2022) to the microscopic suite representations on the torus $\mathbb{T}^7$, of the benchmark training data set. In brief, in a first step

(AGE) it proposes preclusters based on an iterative, adaptive, average linkage clustering method for general metric spaces, that allows to detect clusters of different densities and sizes. In a second step (MINT), each precluster is subjected to torus PCA (see Eltzner et al., 2018) and its projection to its main one-dimensional component is subjected to circular mode hunting, so that each statistically significant antimode corresponds to a post-cluster boundary. For convenience the MINT-AGE (Mode huntINg on Torus pca post iterative Adaptive linkaGe clustEring) algorithm is reproduced in Online Supplementary Material, supplement Section C including a discussion of parameters and our choices. It builds on Dümbgen and Walther (2008), Everitt (1993), Florek et al. (1951), Huckemann and Eltzner (2015), Huckemann et al. (2016), Langfelder et al. (2007), Obulkasim et al. (2015), Sokal and Michener (1958), its general version is described in Mardia et al. (2022).

As discussed in detail in Eltzner et al. (2018), performing PCA analogues on non-Euclidean manifolds may be challenging, in particular on a torus: tangent space PCA (e.g., Fletcher et al., 2004) misses data periodicity, intrinsic PCA (see Huckemann & Ziezold, 2006) produces geodesics winding infinitely often around, each of which approximating all possible data perfectly, and restricting winding numbers (e.g., Altis et al., 2008; Kent & Mardia, 2009, 2015) greatly reduces flexibility. In contrast on spheres, *principle nested spheres* [PNS, by Jung et al. (2012)] is a PCA analogue that is even more flexible and this flexibility persists on suitably stratified spheres which represent the torus in *torus PCA* (see also Mardia et al., 2022): On the m-dimensional sphere, the dimension of the family of main principal nested circle components is $3(m-1)$, while the dimension of the family of first PCs for data on an m-dimensional Euclidean space is dimension $2(m-1)$. This feature is advantageous for PCA-based clustering, since clusters that would require two Euclidean PCs to be separated can often be separated along the main principal nesting circle.

Application of MINT-AGE to the benchmark training data set yields 17 classes. The largest corresponding to the A helix shape contains 3,933 elements and is highly dominant. All classes are rather dense and even the smallest has a credible size of 21 elements. The number of outliers (881), however, is quite large. We conjecture that a considerable number of these are due to incorrect structure reconstructions, which have not been detected because they have not led to clashes. Online Supplementary Material, Figure 18 displays all classes in dihedral angle representation.

The table in Figure 10 compares our MINT-AGE classes with clusters found by Richardson et al. (2008, Table 2) in a larger set encompassing the benchmark training data set. As they report every cluster only by its mean dihedral angles, we have manually assigned these means to MINT-AGE classes. Typically, Figure 9 illustrates how three (Richardson et al., 2008) cluster means are assigned to MINT-AGE Class 6. This larger data set and allowing some clusters with less than 10 elements has lead to a larger number of 46 (Richardson et al., 2008) clusters. Remarkably, more than half (24) of them can be assigned to MINT-AGE clusters and among the ones that could not be assigned, only two have more than 20 elements (7p with 27 elements and 8d with 24).

4.2 Motivation for a multiscale ansatz

In a first fundamental study, we establish a relationship between suites that have similar mesoscopic shapes, see Figure 2 and Notation 3.1. To this end, we cluster the mesoscopic shapes of the suites of the benchmark training data set (Section 3.4) using the *simple version* of AGE from Mardia et al. (2022) (Online Supplementary Material, Algorithm C.1 from Supplement C. 1, performing only Steps 1 and 2 with $\kappa = 5$ and $d_{\max}$ such that 50% of the mesoscopic strands are outliers) yielding the *simple mesoscopic clusters*. By design, we obtain many (110) clusters that are rather concentrated. It turns out that

1. the suites corresponding to each simple mesoscopic cluster also form rather concentrated suite clusters: for most, the standard deviation of angles (between 0 and 2π) of their suites is less than 0.6 and only very few clusters with low cardinality (close to the minimum of $\kappa + 1 = 6$) have higher suite standard deviation (Figure 11, third panel);

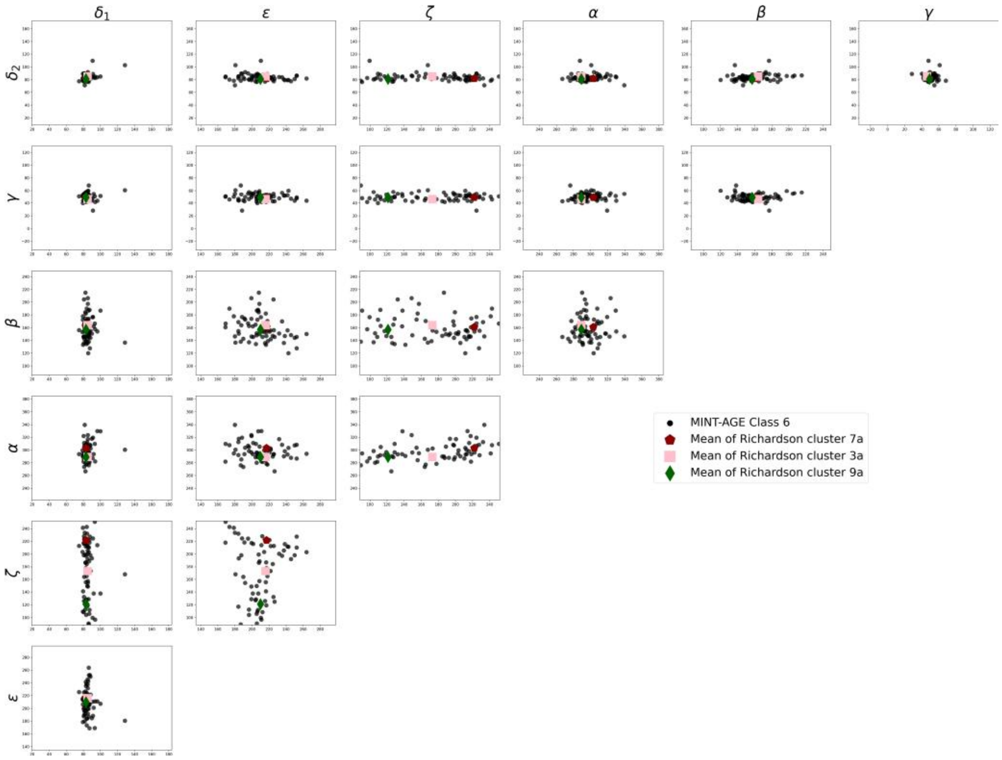

Figure 9. Scatterplots of all two-dimensional dihedral angle pairs (in degrees) of MINT-AGE Class 6 (black) and the reported means of Clusters 7a, 3a and 9a from Richardson et al. (2008).

MINT-AGE	Size	Richardson *et al.* (2008)
1*	3933	1a, (1m), (1L), (&a)
2*	294	1c
3	203	1b, 1[
4*	93	1g
5	91	2a
6	67	7a, 3a, 9a
7	64	0a, (4a)
8	50	
9	46	1e
10*	40	5z
11	37	6p
12	31	2[
13	29	0i, 6n
14	29	4b, (0b)
15	23	
16*	23	6g
17	23	4g
Outliers	881	
Total	5957	

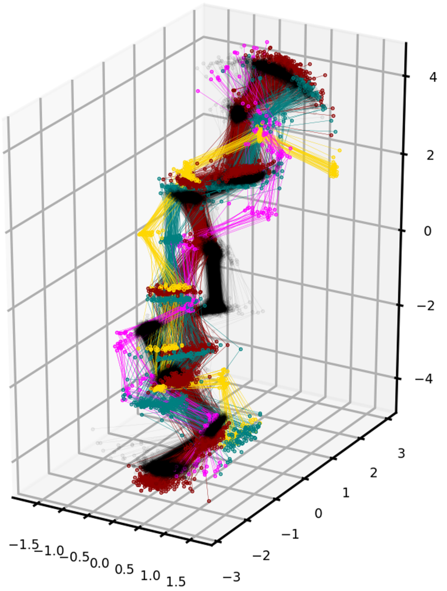

Figure 10. Left: MINT-AGE class numbers and outliers (left column) with size (middle column) from the benchmark training data set with corresponding two-character cluster names (a number for the first character and a letter or '[' for the second character) from Richardson et al. (2008). Asterisks mark MINT-AGE classes displayed in the right panel. Right: Five exemplary classes that can be well displayed together at microscopic scale: Class 1 (black), class 2 (red), class 4 (turquoise), class 10 (yellow), class 16 (magenta). Parentheses indicate that Richardson et al. (2008) cluster means are at boundaries of MINT-AGE classes.

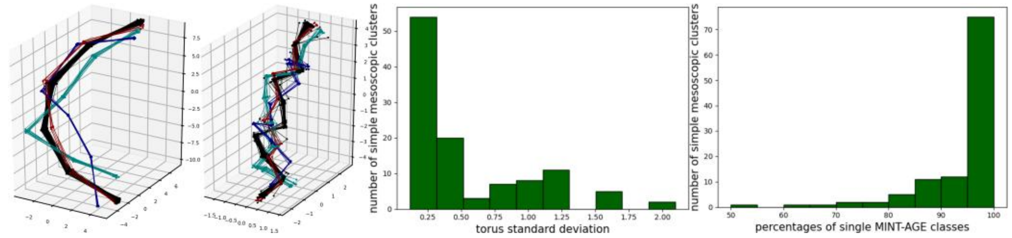

Figure 11. Left: Four exemplary *simple* mesoscopic clusters at mesoscopic scale. Centre left: Their central suites at microscopic scale. Simple mesoscopic Cluster 1 (black) of size 77 contains 73 suites from MINT-AGE Class 1, all of the others clusters are in 1-to-1 correspondence to MINT-AGE classes: Cluster 30 (turquoise, size 13) to Class 4, Cluster 55 (blue, size 8) to 7 and Cluster 92 (red, size 6) to Class 2. Centre right: Binned torus (angular) standard deviation of the suites belonging to simple mesoscopic clusters. For instance, the suites of Cluster 1 from the two left panels have a standard deviation of 0.83, so that Cluster 1 is counted in the 4th green bar from the left. Right: Percentages of the largest MINT-AGE class in each cluster. For instance, the rightmost bar indicates that for 75 out of the 110 clusters at least 95% of their suites belong to a single MINT-AGE class.

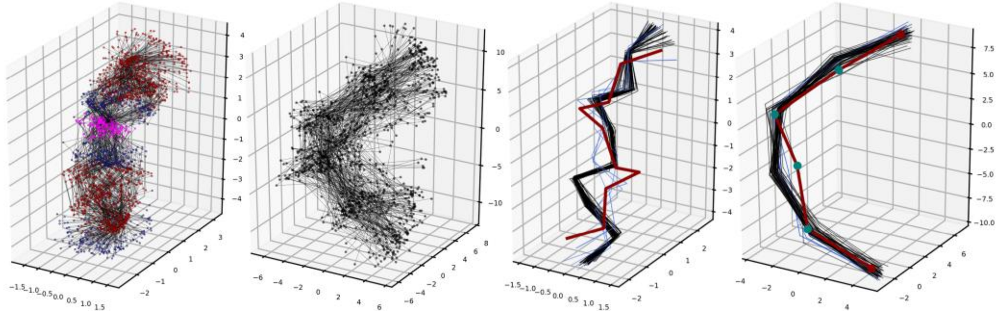

Figure 12. Left: The 198 clash suites from the benchmark data set in Section 3.4 with carbon (dark red), oxygen (dark blue), and phosphorus atoms (pink), cf. Figure 1 at microscopic scale. Left centre: Same at mesoscopic scale. Right centre: At microscopic scale, a typical clash suite $\mathfrak{c}$ (red), the 46 suites (black) from the dominant MINT-AGE class and the other 4 suites (blue) in the neighbourhood $U_\mathfrak{c}$ with respect to mesoscopic shape distance, see also Figure 3. Right: Same at mesoscopic scale where shapes are highly concentrated. The landmarks (teal) of the clash suite at mesoscopic scale (red), except for the middle one, require only very moderate correction.

2. simple mesoscopic clusters are in high correspondence with the 17 MINT-AGE classes from Section 4.1 as clearly visible in the rightmost panel of Figure 11 and detailed for exemplary clusters in the caption of Figure 11.

This leads to the following hypothesis.

Hypothesis 4.1 Correctly reconstructed suites with similar mesoscopic shapes have also similar suite shape. In particular, concentrated mesoscopic clusters relate to suite classes.

In a second fundamental study, we consider the 198 clash suites in the benchmark data set forming the test data set, see Section 3.4. Their suite shapes as well as their mesoscopic shapes feature a rather larger spread, see Figure 12 (first two panels). As before, we consider training suites from concentrated neighbourhoods in the mesoscopic shape space, of size $\rho \in \mathbb{N}$. For a given clash suite $\mathfrak{c}$ such a neighbourhood is

$$U_\mathfrak{c} := \{\mathfrak{t} \in \mathfrak{T} : \#\{\mathfrak{t}' \in \mathfrak{T} : d_\Sigma(m_{\mathfrak{t}'}, m_\mathfrak{c}) \le d_\Sigma(m_\mathfrak{t}, m_\mathfrak{c})\} \le \rho\}. \qquad (11)$$

The neighbourhood $U_\mathfrak{c}$ is the set of the ρ suites of the training data, whose mesoscopic shapes are most similar to $m_\mathfrak{c}$ with respect to mesoscopic shape space distance. Recall from Section 3.4 that $\mathfrak{T}$

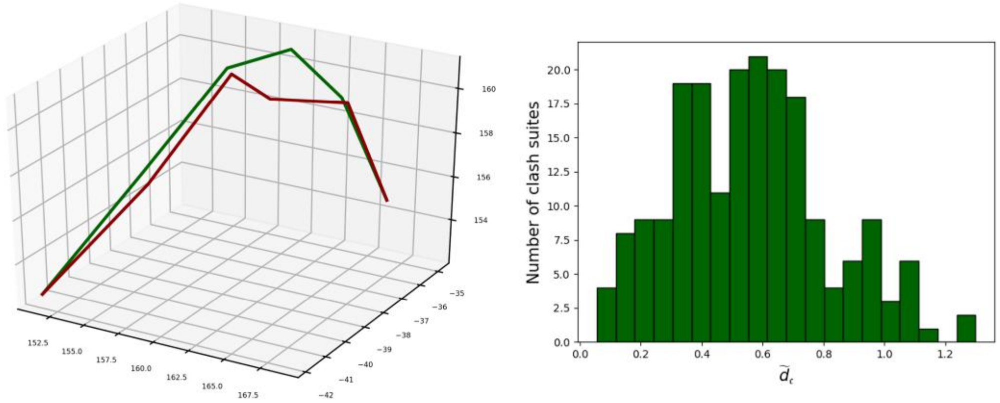

Figure 13. Left: At mesoscopic scale the clash suite from Figure 12 (right) and its mesoscopic shape corrected by CLEAN. Right: Histogram of relative distances $\widetilde{d}_\mathfrak{c}$ between corrected mesoscopic shapes and original mesoscopic shapes from (13) over all $\mathfrak{c} \in \mathfrak{C}$.

is the set of training suites (the clash-free suites in the benchmark data set) and that $m_\mathfrak{t}$ denotes the mesoscopic shape of $\mathfrak{t} \in \mathfrak{T}$. On close inspection of the 198 $U_\mathfrak{c}$'s we find a situation typically illustrated in the last two panels of Figure 12, which leads to the following hypothesis.

Hypothesis 4.2 While at microscopic scale, clash suite shapes are rather irregular among the suite shapes of their clash-free neighbours, at mesoscopic scale, their mesoscopic shapes differ only mildly from nearby clash-free mesoscopic shapes.

The theoretical argument underlying this hypothesis is that even drastic errors on the atomic suite scale can still be compatible with electron density measurement results due to finite resolution, while drastic errors on the mesoscopic scale are excluded since they would strongly contradict the measured electron density. Indeed, we find empirically at mesoscopic scale that only one of the four teal landmarks in the middle (Figure 12, right panel) differs more strongly from the neighbouring clash-free mesoscopic shapes in $U_\mathfrak{c}$. For all 198 clash shapes the histogram in Figure 13 shows that for the vast majority of clash suites $\mathfrak{c} \in \mathfrak{C}$, the distance (detailed in Section 4.4) of its mesoscopic shape to its clash-free correction is only rarely barely above and mostly well below the resolution order.

Remark 4.3 There are databases that store different RNA motifs and their interaction: In *RNA Bricks* (Chojnowski et al., 2013), the elements of simple mesoscopic Clusters 1 and 92 are often found in a *stem cluster* (corresponding to helical backbone shapes) and the elements of simple mesoscopic Cluster 30 are found in a *loop cluster*. Similarly, in Petrov et al. (2013), the elements of simple mesoscopic Cluster 30 are classified in the *hairpin loop* with the name HL_43074.14. *Stems* and *loops* are depicted in the *hairpin* structure scheme in the left panel of Figure 5.

4.3 The multiscale RNA backbone structure correction prodecure

Exploiting the above Hypotheses 4.1 and 4.2, the following multiscale backbone correction procedure simultaneously corrects clashing suites at microscopic and at mesoscopic scale, working with concentrated neighbourhoods as in (11), defined by mesoscopic shape distance. In these concentrated neighbourhoods, dominating classes from MINT-AGE of the training data set provide guidance for correction. Recall from the two left panels of Figure 3, with more detail in the two right panels of Figure 12, that even a minor correction of one of the sugar ring centres at mesoscopic scale can have great impact on the shape of the suite of interest, which is positioned between the third and fourth sugar ring at mesoscopic scale.

4.3.1 Multiscale correction (CLEAN)

Input:

- a training data set $\mathfrak{T}$ comprising only clash-free admissible suites (suites that feature a mesoscopic shape, see Definition 3.3),
- a list of classes $C_1, \ldots, C_r$ and an outlier set R for $\mathfrak{T}$ obtained from applying the MINT-AGE algorithm (see Section 4.1 and Online Supplementary Material, Algorithm C.3),
- a clash suite $\mathfrak{c}$ and its corresponding mesoscopic shape $m_{\mathfrak{c}}$.
- the size $\rho \in \mathbb{N}$ of the neighbourhood $U_{\mathfrak{c}}$ from (11), we choose $\rho = 50$ as roughly twice the size of the smallest class, and
- the flag `DOMINATING` set to `ABSOLUTE` or `RELATIVE` which will return either the absolutely dominating cluster in $U_{\mathfrak{c}}$ or the relatively dominating cluster with at least $\rho/10$ elements, taking into account cluster size (in Step (b) below).

Implementation steps:

1. Calculate
   (a) the neighbourhood $U_{\mathfrak{c}}$ as defined in (11) of the ρ suites of the training data, whose mesoscopic shapes are most similar to $m_{\mathfrak{c}}$ with respect to mesoscopic shape space distance;
   (b) according to flag `DOMINATING`, the number

$$j_{\mathfrak{c}} \in \operatorname*{argmax}_{j=1,\ldots,m} \#(C_j \cap U_{\mathfrak{c}}), \quad (\texttt{ABSOLUTE}), \quad \text{or}$$

$$j_{\mathfrak{c}} \in \operatorname*{arg\,max}_{j=1,\ldots,m} \mathbf{1}_{\{C_j \,|\, \#(C_j \cap U_{\mathfrak{c}}) \geq \rho/10\}} \#(C_j \cap U_{\mathfrak{c}})/\#C_j,$$

   (`RELATIVE`), respectively, of the dominant MINT-AGE class in $U_{\mathfrak{c}}$;
   (c) a Fréchet mean $\tau_{\mathfrak{c}} \in \operatorname{argmin}_{\mathfrak{t}\in\mathbb{T}^7} \sum_{\mathfrak{t}'\in C_{j_{\mathfrak{c}}}\cap U_{\mathfrak{c}}} d_{\mathbb{T}^7}(\mathfrak{t}, \mathfrak{t}')^2$, of the dominant class' suites in the neighbourhood;
   (d) the approximate length $\ell_{\tau_{\mathfrak{c}}}$ of the suite by the mean distance of the two central sugar rings $k+1$ and $k+2$ of the mesoscopic shapes corresponding to the suites of $C_{j_{\mathfrak{c}}} \cap U_{\mathfrak{c}}$;
   (e) a Procrustes mean

$$\mu_{\mathfrak{c}} \in \operatorname*{argmin}_{m\in S\Sigma_3^{2k+2}} \sum_{\mathfrak{t}\in C_{j_{\mathfrak{c}}}\cap U_{\mathfrak{c}}} d_{\Sigma}(m, m_{\mathfrak{t}})^2,$$

   of the corresponding mesoscopic shapes.
2. With a mesoscopic shape $m_{\mathfrak{c}} = [x_1, \ldots, x_{2k+2}]$ defined as in Equation (3) by a landmark configuration matrix $(x_1, \ldots, x_{2k+2})$, determine the corrected mesoscopic shape $m_{\tau_{\mathfrak{c}}}$ as the orthogonal projection of the size-and-shape Y^* of the Procrustes mean $\mu_{\mathfrak{c}} = [z_1, \ldots, z_{2k+2}]$ to the set

$$\{m = [y_1, \ldots, y_{2k+2}] \in S\Sigma_3^{2k+2} : \|y_1 - y_{2k+2}\| = a_1, \|y_{k+2} - y_{k+1}\| = a_2\} \tag{12}$$

   of mesoscopic shapes whose configurations have distance $a_1 = \|x_1 - x_{2k+2}\|$ between the first and the last landmark given by that of any configuration of $m_{\mathfrak{c}}$ and whose distance a_2 between the central landmarks is the length $\ell_{\tau_{\mathfrak{c}}}$ which is chosen so that the Fréchet mean suite $\tau_{\mathfrak{c}}$ fits between them. By Theorem 2.3, with $m = 2k+2$, $r = 2$, $\sigma(1) = k+2$, $\sigma(k+1) = k+2$ and $\sigma(j) = j$ for $j \in \{2, \ldots, k, k+3, \ldots, 2k+1\}$, the (in practice there will no ties between the landmarks) desired orthogonal projection to $S\Sigma_3^{2k+2}(\sigma, a_1, a_2)$ which is the space determined by

(12) is given by

$$y_1^* = \alpha z_1 + (1-\alpha) z_{2k+2}, \quad y_{2k+2}^* = \alpha z_{2k+2} + (1-\alpha) z_1$$
$$y_{k+1}^* = \beta z_{k+1} + (1-\beta) z_{k+2}, \quad y_{k+2}^* = \beta z_{k+2} + (1-\beta) z_{k+1}$$

where

$$\alpha = \frac{1}{2}\left(1 + \frac{\|x_{2k+2} - x_1\|}{\|z_{2k+2} - z_1\|}\right), \quad \beta = \frac{1}{2}\left(1 + \frac{\ell_{\tau_\mathfrak{c}}}{\|z_{k+2} - z_{k+1}\|}\right)$$

and $y_j^* = z_j$ for $j \in \{1, \ldots 2k+2\} \setminus \{1, k+1, k+2, 2k+2\}$.

Output:

- the corrected suite shape $\tau_\mathfrak{c}$ and its corrected mesoscopic shape $m_{\tau_\mathfrak{c}} := [Y^*]$.

As mentioned above, we suggest to choose $\rho = 50$ as roughly twice the size of the smallest class. A larger value for ρ would make it very unlikely that the plurality of neighbouring suites for a clash suite are from the smallest cluster, because any other nearby clusters will outnumber it. A smaller value for ρ would lead to less reliable results and, in some cases, to a majority of outliers in the set.

For many applications of CLEAN, setting DOMINATING = ABSOLUTE can be used as we do for analysing two suites of SARS-CoV-2 RNA in the following Section 5. If considerably differing class sizes are of concern, setting DOMINATING = RELATIVE ensures assignment to smaller classes that dominate neighbourhoods at mesoscopic scale only relatively to their total size. This results in greater diversity as illustrated in Online Supplementary Material, Figure 19, applying CLEAN to the entire benchmark test set from Section 3.4.

4.4 Validation of CLEAN

We apply the CLEAN method from Section 4.3.1 to the 198 clash suites which form the test data set $\mathfrak{C}$ from Section 3.4. For validation, we confirm that backbone correction is realistic and neither arbitrary nor ambiguous. For the former, we verify that corrections happen on a scale not larger than the underlying X-ray crystallography resolution, see Section 3.3, and for the latter we verify that the largest MINT-AGE classes in neighbourhoods $U_\mathfrak{c}$ from (11) are indeed strongly dominating in most cases.

In order to relate the amount of correction to resolution, consider the normalised Procrustes distance between the mesoscopic shape $m_\mathfrak{c}$ of a clash suite $\mathfrak{c} \in \mathfrak{C}$ and the mesoscopic shape $m_{\tau_\mathfrak{c}}$ of its correction by CLEAN,

$$\tilde{d}_\mathfrak{c}^2 := \frac{1}{\text{resolution}^2} \frac{3}{\text{degrees of freedom}} d_\Sigma(m_\mathfrak{c}, m_{\tau_\mathfrak{c}})^2. \tag{13}$$

Recalling that the group of 3D Euclidean transformations is of dimension 6, the *degrees of freedom* are given by $3(2k+2) - 6 = 3 \cdot 2k$, so that the inverse of the second quotient above gives the number $2k$ of free landmarks in Σ_3^{2k+2} taking into account that the resolution incorporates the spatial dimension 3.

The histogram in Figure 13 shows that for the vast majority of clash suites $\mathfrak{c} \in \mathfrak{C}$, $\tilde{d}_\mathfrak{c}$ is smaller than 1. Thus, corrections are only rarely slightly above and mostly well below the order of resolution.

In order to assess how dominating torus MINT-AGE classes are in neighbourhoods $U_\mathfrak{c}$ ($\mathfrak{c} \in \mathfrak{C}$), the histogram in Figure 14 shows the number of suites in the dominating classes $C_{j_\mathfrak{c}}$. Indeed, for considerably more than half of the neighbourhoods, the dominating cluster contains more than half of the neighbouring suites. Remarkably, the negative correlation visible in the scatter plot in Figure 14 (right) shows that a smaller amount of correction tends to correlate with more elements being in the dominating cluster.

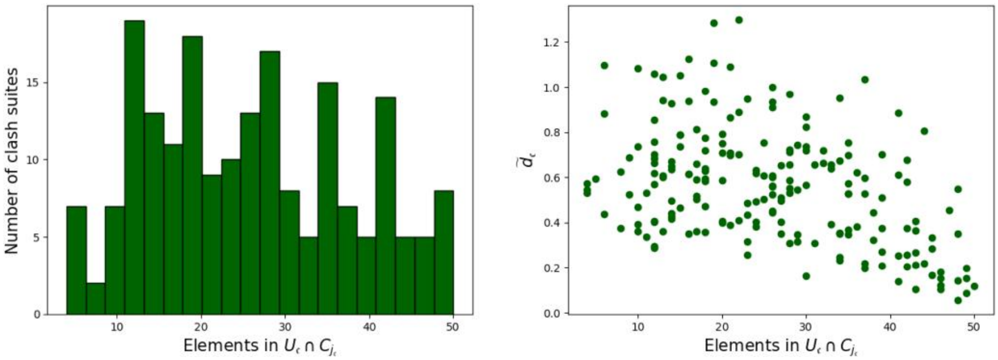

Figure 14. Left: Histogram of the number of suites in $U_\mathfrak{c}$ from (11), over all (198) clash suites $\mathfrak{c} \in \mathfrak{C}$ (test data set), of the dominating MINT-AGE class. Right: Scatter plot relating the number of suites in the dominating class of $U_\mathfrak{c}$ and the normalised distance $\widetilde{d}_\mathfrak{c}$ from (14), over $\mathfrak{c} \in \mathfrak{C}$.

5 Application to SARS-CoV-2 suites

With the recent worldwide pandemic of the *severe acute respiratory syndrome coronavirus 2* (SARS-CoV-2), the virus' RNA structure reconstruction and backbone correction has become ever more relevant. Indeed, effective drug and vaccine development necessitates good understanding of the three-dimensional RNA structure, see Croll et al. (2021). Recently, a large number of measurements has been added to the Protein Data Bank, see Berman et al. (2000), and as part of the *Coronavirus Structural Task Force* (CSTF) headed by Andrea Thorn, a large number of data sets of SARS-CoV-2 and related structures are compiled in a git repository, see Thorn et al. (2021). While *X*-ray crystallography can achieve very high resolution in principle, the large viral genome, comprising ~20,000 bases, is very difficult to crystallize. Therefore, many structures are determined by cryogenic electron microscopy (cryo-EM).

5.1 The frameshift stimulation element

In Zhang et al. (2021), the frameshift stimulation element of the SARS-CoV-2 genome was studied (see Figure 5, right panel), which, due to its *slippery site* encodes different proteins simultaneously (this method of information compression is shared with other viruses such as HIV-1). As their balanced expression is required for virus replication, this element is believed to be fairly resistant against mutations. Hence it is a promising target for antiviral drug design. Its three-dimensional structure has been assessed by cryo-EM with a resolution of 6.9 Å using the ribosolve pipeline from Kappel et al. (2020), see also Section 3.3. Using a consensus secondary structure of the molecule and the cryo-EM map, Zhang et al. (2021) proposed 10 possible three-dimensional structure models (based on a measurement with mean pairwise root mean squared deviation of 5.68 Å) and stored them to the Protein Data Bank. Notably, it was not possible to reliably assign individual atom positions, but the secondary arrangement of helical segments and the non-helical linking segments could be reconstructed, see Zhang et al. (2021) and first panel of Figure 15. In particular, the suites linking different helical segments have been difficult to reconstruct. Here we focus on the suite determined by Residues 28/29 which we call *Suite 1* and on the suite determined by Residues 33/34 which we call *Suite 2* (referring to enumeration in the PDB file).

5.2 Reconstructing suite 1

Suite 1 (red arrow in Figure 5, right panel, and the left red dot in Figure 15, left panel) is a clash suite in all 10 models proposed by Zhang et al. (2021), as determined by PHENIX, see Section 3.3. Notably, the P′-O3′ bonds are unphysically long (red verticals in Figure 3, centre right panel), hinting towards a bad structure fit of all 10 proposals. Figure 15 (3rd panel) shows $\mathfrak{c}_1$, the first clashing proposal for Suite 1, at mesoscopic scale and its highly concentrated neighbourhood $U_{\mathfrak{c}_1}$ from (11), in which 43 out of the 50 suites belong to MINT-AGE Class 4. Its torus mean and $\mathfrak{c}_1$ at microscopic scale are shown in Figure 15 (2nd panel). The situation is very similar for the other clashing

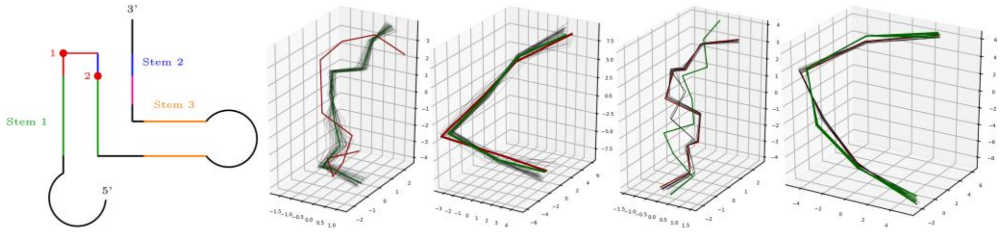

Figure 15. First: 2D scheme of the SARS-CoV-2 frameshift stimulation element, adapted from (Zhang et al., 2020, Figure 8), see also Figure 5 (right panel), with double-stranded helical stems (green, yellow and blue) and connecting Suites 1 and 2 (red dots, only one nucleotide is on the red branch) Second: Model 1 ($\mathfrak{c}_1$) of Suite 1 (red, clashing) proposed by Zhang et al. (2021), the 43 suites from MINT-AGE Class 4 dominating in neighbourhood $U_{\mathfrak{c}_1}$ (black) and their torus mean (green) at microscopic scale. Third: The corresponding mesoscopic shape $m_{\mathfrak{c}_1}$ (red), the 43 mesoscopic shapes of suites from MINT-AGE Class 4 in neighbourhood $U_{\mathfrak{c}_1}$ (black) and their Procrustes mean geodesically projected to a mesoscopic shape featuring length constraints from $m_{\mathfrak{c}_1}$ and the microscopic correction of $\mathfrak{c}_1$ (green). Fourth: The 10 different model proposals (one in red clashes, the others in black form two clusters) by Zhang et al. (2021) of Suite 2 and their highly consistent correction from MINT-AGE Class 1 (green) at microscopic scale. Right: Using same colouring, at mesoscopic scale all 10 models from Zhang et al. (2021) of Suite 2 form one cluster for which CLEAN-MINT-AGE provides a moderate correction only.

proposals $\mathfrak{c}_2, \ldots, \mathfrak{c}_{10}$ for Suite 1: MINT-AGE Class 4 dominates strongly in their concentrated neighbourhoods, each warranting only minor corrections at mesoscopic scale (Figure 3, 4th panel) and all of their torus means at microscopic scale are nearly indistinguishable (Figure 3, 3rd panel). Notably, MINT-AGE Class 4 corresponds to one (Richardson et al., 2008) cluster only (namely 2a, see Figure 10) which has been characterised there as *GNRA 1-2; U-turn*.

5.3 Reconstructing suite 2

Suite 2 (blue arrow in Figure 5, right panel, and the right red dot in Figure 15, left panel) is a clash suite only in one out of the 10 models proposed by Zhang et al. (2021), as determined by PHENIX, see Section 3.3. At microscopic scale (Figure 15, fourth panel, red and black) these models are inconclusive as they feature two different clusters and one of the models (red) from the larger cluster has a clash score 0.401 Å, slightly above the threshold of 0.4 Å. As before, at mesoscopic scale (Figure 15, fifth panel, red and black), the shapes of all 10 models proposed are very similar and consistent and there is a single MINT-AGE class that strongly dominates every neighbourhood (11), namely Class 1. Figure 15 (fifth panel, green) shows its Procrustes means projected to the mesoscopic shapes featuring length constraints from the corresponding mesoscopic shapes $m_{\mathfrak{c}_1}, \ldots, m_{\mathfrak{c}_{10}}$ of the 10 models and the suite lengths of the corrections from $\mathfrak{c}_1, \ldots, \mathfrak{c}_{10}$ as detailed in Section 4.3.1. In consequence, the CLEAN-MINT-AGE corrections are the torus means of the suites of Class 1 in the respective neighbourhoods. Again these are nearly indistinguishable, giving one consistent correction for Suite 2 in Figure 15 (fourth panel, green).

6 Discussion

The CLEAN-MINT-AGE procedure presented here, yielding

1. hierarchical (different shape spaces for multiscale interrelationships),
2. probabilistic (Fréchet means in iterative adaptive torus clusters obtained after circular mode hunting, projected to a shape space featuring data-driven constraints),
3. clash free, and
4. fast,

RNA backbone correction which is an important and challenging contribution warranting further research in various directions, of which we sketch three.

In particular, we have discovered, described and exploited a relationship of RNA 3D structure between a microscopic and a mesoscopic scale. Further research, building on larger data sets, beyond the scope of this paper, will investigate this relationship more closely and identify relationships between other scales as well and exploit these similarly. As we have found that shape at different scales is best described by fundamentally different shape spaces, this involves statistically linking different geometrical models of shape.

At this point, the two-scale correction method CLEAN we propose corrects a central suite at microscopic scale only. More realistic, again beyond the scope of this paper, are simultaneous corrections of all suites involved at the mesoscopic scale (notably, adjacent suites overlap at four atoms), and correction of suites linked by nucleobase bindings, potentially far away along the backbone. Such corrections can, after elaborate extension, also address backbone-to-backbone-extra-suites clashes and even the more rare nucleobase clashes. Obviously, these methods extend to various other biomolecules and in particular to protein structure correction, see Hamelryck et al. (2010).

As mentioned in the Introduction, there are elaborate correction methods, for example ERRASER from Chou et al. (2013b), building on approximations of highly complex molecular dynamics simulations yielding 3D structures following the laws of biophysical chemistry. This aims not only at correcting all clashes (i.e., within-suite and between-suites, as well as backbone or base to backbone or base), it also aims at various other structure improvements. While for the test data set this entire process took several days on the ROSIE servers (Chou et al., 2013a), frequently not removing all clashes, our CLEAN method, removing all within-suite-backbone-to-backbone clashes, ran within minutes. Since in contrast to corrections based on molecular dynamics, as demonstrated in Figures 3 and 15, our proposed corrections can be quite different from original clash suite shapes, they may serve as additional initial states for subsequent molecular dynamics, and thus provide a powerful tool.

Acknowledgments

The authors are grateful for fruitful discussions with Markus Hiller, Thomas Hamelryck, and Carina Wollnik. The authors also gratefully acknowledge the very valuable comments provided by the three anonymous referees, improving this paper.

Data availability

The PDB files and all the code used to generate the analyses and plots presented in this paper can be found https://gitlab.gwdg.de/henrik.wiechers1/clean-mintage-code.

Conflict of interest: None declared.

Funding

All authors except K.V.M. acknowledge DFG CRC 1456, K.V.M. acknowledges the Leverhulme Trust for the Emeritus Fellowship.

Supplementary material

Supplementary material are available at *Journal of the Royal Statistical Society: Series C* online.

References

AlQuraishi M. (2019). Parallelized natural extension reference frame: Parallelized conversion from internal to cartesian coordinates. *Journal of Computational Chemistry*, *40*(7), 885–892. https://doi.org/10.1002/jcc.25772

Altis A., Otten M., Nguyen P. H., Hegger R., & Stock G. (2008). Construction of the free energy landscape of biomolecules via dihedral angle principal component analysis. *The Journal of Chemical Physics*, *128*(24), 245102. https://doi.org/10.1063/1.2945165

Arnaudon M., & Miclo L. (2014). Means in complete manifolds: Uniqueness and approximation. *ESAIM: Probability and Statistics*, *18*(1), 185–206. https://doi.org/10.1051/ps/2013033

Batool M., Ahmad B., & Choi S. (2019). A structure-based drug discovery paradigm. *International Journal of Molecular Sciences*, *20*(11), 2783. https://doi.org/10.3390/ijms20112783

Berman H. M., Westbrook J., Feng Z., Gilliland G., Bhat T. N., Weissig H., Shindyalov I. N., & Bourne P. E. (2000). The protein data bank. *Nucleic Acids Research*, *28*(1), 235–242. https://doi.org/10.1093/nar/28.1.235

Chen V. B., Arendall W. B., Headd J. J., Keedy D. A., Immormino R. M., Kapral G. J., Murray L. W., Richardson J. S., & Richardson D. C. (2010). Molprobity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica, Section D: Biological Crystallography*, *66*(Pt 1), 12–21. https://doi.org/10.1107/S0907444909042073

Chojnowski G., Wale T., & Bujnicki J. M. (2013). RNA Bricks-a database of RNA 3D motifs and their interactions. *Nucleic Acids Research*, *42*(D1), D123–D131. https://doi.org/10.1093/nar/gkt1084

Chou F.-C., Sripakdeevong P., Dibrov S. M., Hermann T., & Das R. (2013a). Correcting pervasive errors in RNA crystallography through enumerative structure prediction. *Nature Methods*, *10*(1), 74–76. https://doi.org/10.1038/nmeth.2262

Chou F.-C., Sripakdeevong P., Dibrov S. M., Hermann T., & Das R. (2013b). Correcting pervasive errors in RNA crystallography through enumerative structure prediction. *Nature Methods*, *10*(1), 74–76. https://doi.org/10.1038/nmeth.2262

Croll T. I., Williams C. J., Chen V. B., Richardson D. C., & Richardson J. S. (2021). Improving SARS-CoV-2 structures: Peer review by early coordinate release. *Biophysical Journal*, *120*(6), 1085–1096. https://doi.org/10.1016/j.bpj.2020.12.029

Dryden I. L., & Mardia K. V. (2016). *Statistical shape analysis, with applications in R* (2nd ed.). John Wiley and Sons.

Duarte C. M., & Pyle A. M. (1998). Stepping through an RNA structure: A novel approach to conformational analysis 11. Edited by D. Draper. *Journal of Molecular Biology*, *284*(5), 1465–1478. https://doi.org/10.1006/jmbi.1998.2233

Dümbgen L., & Walther G. (2008). Multiscale inference about a density. *Annals of Statistics*, *36*(4), 1758–1785. https://doi.org/10.1214/07-AOS521

Eltzner B., Huckemann S., & Mardia K. V. (2018). Torus principal component analysis with applications to RNA structure. *The Annals of Applied Statistics*, *12*(2), 1332–1359. https://doi.org/10.1214/17-AOAS1115

Everitt B. (1993). *Cluster analysis* (3rd ed.). Edward Arnold.

Fletcher P. T., Lu C., Pizer S. M., & Joshi S. (2004). Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Transactions on Medical Imaging*, *23*(8), 995–1005. https://doi.org/10.1109/TMI.2004.831793

Florek K., Lukaszewicz J., Perkal J., Steinhaus H., & Zubrzycki S. (1951). Sur la liaison et la division des points d'un ensemble fini. *Colloquium Mathematicum*, *2*(3–4), 282–285. https://doi.org/10.4064/cm-2-3-4-282-285

Hamelryck T., Borg M., Paluszewski M., Paulsen J., Frellsen J., Andreetta C., Boomsma W., Bottaro S., Ferkinghoff-Borg J., & Flower D. R. (2010). Potentials of mean force for protein structure prediction vindicated, formalized and generalized. *PLoS ONE*, *5*(11), e13714. https://doi.org/10.1371/journal.pone.0013714

Huckemann S. F. (2012). On the meaning of mean shape: Manifold stability, locus and the two sample test. *Annals of the Institute of Statistical Mathematics*, *64*(6), 1227–1259. https://doi.org/10.1007/s10463-012-0352-2

Huckemann S. F., & Eltzner B. (2015). *Polysphere PCA with applications*. In Proceedings of the 33th LASR Workshop (pp. 51–55). Leeds University Press. http://www1.maths.leeds.ac.uk/statistics/workshop/lasr2015/Proceedings15.pdf.

Huckemann S. F., Kim K.-R., Munk A., Rehfeldt F., Sommerfeld M., Weickert J., & Wollnik C. (2016). The circular SiZer, inferred persistence of shape parameters and application to early stem cell differentiation. *Bernoulli*, *22*(4), 2113–2142. https://doi.org/10.3150/15-BEJ722

Huckemann S. F., & Ziezold H. (2006). Principal component analysis for Riemannian manifolds, with an application to triangular shape spaces. *Advances in Applied Probability*, *38*(2), 299–319. https://doi.org/10.1239/aap/1151337073

Ippolito J. A., & Steitz T. A. (2000). The structure of the HIV-1 RRE high affinity rev binding site at 1.6 Å resolution. *Journal of Molecular Biology*, *295*(4), 711–717. https://doi.org/10.1006/jmbi.1999.3405

Jain S., Richardson D. C., & Richardson J. S. (2015). Chapter 7. Computational methods for RNA structure validation and improvement. In S. A. Woodson, & F. H. Allain (Eds.), *Structures of large RNA molecules and their complexes. Methods in Enzymology*, Vol. *558* (pp. 181–212). Academic Press.

Jung S., Dryden I. L., & Marron J. S (2012). Analysis of principal nested spheres. *Biometrika*, *99*(3), 551–568. https://doi.org/10.1093/biomet/ass022

Kappel K., Zhang K., Su Z., Watkins A. M., Kladwang W., Li S., Pintilie G., Topkar V. V., Rangan R., Zheludev I. N., Yesselman J. D., Chiu W., & Das R. (2020). Accelerated cryo-EM-guided determination of three-dimensional RNA-only structures. *Nature Methods*, *17*(7), 699–707. https://doi.org/10.1038/s41592-020-0878-9

Kent J. T., & Mardia K. V. (2009). Principal component analysis for the wrapped normal torus model. In Proceedings of the Leeds Annual Statistical Research (LASR) Workshop 2009. Leeds: University of Leeds Press.

Kent J. T., & Mardia K. V. (2015). The winding number for circular data. In *Proceedings of the Leeds Annual Statistical Research (LASR) Workshop*, University of Leeds Press, Leeds.

Langfelder P., Zhang B., & Horvath S. (2007). Defining clusters from a hierarchical cluster tree: The dynamic tree cut package for R. *Bioinformatics*, *24*(5), 719–720. https://doi.org/10.1093/bioinformatics/btm563

Liao H. Y., & Frank J. (2010). Definition and estimation of resolution in single-particle reconstructions. *Structure (London, England: 1993)*, *18*(7), 768–775. https://doi.org/10.1016/j.str.2010.05.008

Liebschner D., Afonine P. V., Baker M. L., Bunkóczi G., Chen V. B., Croll T. I., Hintze B., Hung L.-W., Jain S., McCoy A. J., Moriarty N. W., Oeffner R. D., Poon B. K., Prisant M. G., Read R. J., Richardson J. S., Richardson D. C., Sammito M. D., Sobolev O. V., …Adams P. D. (2019). Macromolecular structure determination using *X*-rays, neutrons and electrons: Recent developments in *Phenix*. *Acta Crystallographica Section D*, *75*(10), 861–877. https://doi.org/10.1107/S2059798319011471

Mardia K. V., Wiechers H., Eltzner B., & Huckemann S. F. (2022). Principal component analysis and clustering on manifolds. *Journal of Multivariate Analysis*, *188*(1), 104862. 50th Anniversary Jubilee Edition. https://doi.org/10.1016/j.jmva.2021.104862

Murray L. J. W., Arendall W. B., Richardson D. C., & Richardson J. S. (2003). RNA backbone is rotameric. *Proceedings of the National Academy of Sciences*, *100*(24), 13904–13909. https://doi.org/10.1073/pnas.1835769100

Obulkasim A., Meijer G. A., & van de Wiel M. A. (2015). Semi-supervised adaptive-height snipping of the hierarchical clustering tree. *BMC Bioinformatics*, *16*(1), 15. https://doi.org/10.1186/s12859-014-0448-1

Parsons J., Holmes J. B., Rojas J. M., Tsai J., & Strauss C. E. M. (2005). Practical conversion from torsion space to cartesian space for in silico protein synthesis. *Journal of computational chemistry*, *26*(10), 1063–1068. https://doi.org/10.1002/jcc.20237

Petrov A. I., Zirbel C. L., & Leontis N. B. (2013). Automated classification of RNA 3D motifs and the RNA 3D motif atlas. *RNA (New York, N.Y.)*, *19*(10), 1327–1340. https://doi.org/10.1261/rna.039438.113

Richardson J. S., Schneider B., Murray L. W., Kapral G. J., Immormino R. M., Headd J. J., Richardson D. C., Ham D., Hershkovits E., Williams L. D., Keating K. S., Pyle A. M., Micallef D., Westbrook J., Berman H. M., & Consortium R. O. (2008). Rna backbone: Consensus all-angle conformers and modular string nomenclature (an RNA ontology consortium contribution). *RNA (New York, N.Y.)*, *14*(3), 465–481. https://doi.org/10.1261/rna.657708

Richardson J. S., Williams C. J., Hintze B. J., Chen V. B., Prisant M. G., Videau L. L., & Richardson D. C. (2018). Model validation: local diagnosis, correction and when to quit. *Acta Crystallographica Section D*, *74*(2), 132–142. https://doi.org/10.1107/S2059798317009834

Sargsyan K., Wright J., & Lim C. (2012). GeoPCA: a new tool for multivariate analysis of dihedral angles based on principal component geodesics. *Nucleic Acids Research*, *40*(3), e25. https://doi.org/10.1093/nar/gkr1069

Schlick T., & Pyle A. M. (2017). Opportunities and challenges in RNA structural modeling and design. *Biophysical Journal*, *113*(2), 225–234. https://doi.org/10.1016/j.bpj.2016.12.037

Schrödinger L. L. C. (2015). The PyMOL molecular graphics system. Version 1.8.

Sokal R. R., & Michener C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, *38*(22), 1409–1438.

Tang L., Johnson K. N., Ball L. A., Lin T., Yeager M., & Johnson J. E. (2001). The structure of pariacoto virus reveals a dodecahedral cage of duplex RNA. *Nature Structural Biology*, *8*(1), 77–83. https://doi.org/10.1038/83089

Thorn A., Gao Y., Nolte K., Kirsten F., & Stäb S. (2021). Coronavirus structural task force. https://github.com/thorn-lab/coronavirus structural_task_force.

Wadley L. M., Keating K. S., Duarte C. M., & Pyle A. M. (2007). Evaluating and learning from RNA pseudotorsional space: Quantitative Validation of a reduced representation for RNA structure. *Journal of Molecular Biology*, *372*(4), 942–957. https://doi.org/10.1016/j.jmb.2007.06.058

Wang Z., Gerstein M., & Snyder M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63. https://doi.org/10.1038/nrg2484

Watson J., Baker T., Bell S., Gann A., Levine M., & Losick R. (2004). *Molecular biology of the gene* (5th ed.). Pearson Education.

Word J., Lovell S. C., LaBean T. H., Taylor H. C., Zalis M. E., Presley B. K., Richardson J. S., & Richardson D. C. (1999). Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. Edited by J. Thornton. *Journal of Molecular Biology*, *285*(4), 1711–1733. https://doi.org/10.1006/jmbi.1998.2400

Zhang K., Zheludev I. N., Hagey R. J., Haslecker R., Hou Y. J., Kretsch R., Pintilie G. D., Rangan R., Kladwang W., Li S., Wu M. T.-P., Pham E. A., Bernardin-Souibgui C., Baric R. S., Sheahan T. P., D'Souza V., Glenn J. S., Chiu W., & Das R. (2021). Cryo-EM and antisense targeting of the 28-kDa frameshift stimulation element from the SARS-CoV-2 RNA genome. *Nature Structural & Molecular Biology*, *28*(9), 747–754. https://doi.org/10.1038/s41594-021-00653-y

Zhang K., Zheludev I. N., Hagey R. J., Wu M. T.-P., Haslecker R., Hou Y. J., Kretsch R., Pintilie G. D., Rangan R., Kladwang W., Li S., Pham E. A., Bernardin-Souibgui C., Baric R. S., Sheahan T. P., D'Souza V., Glenn J. S., Chiu W., & Das R. (2020). Cryo-electron microscopy and exploratory antisense targeting of the 28-kDa frameshift stimulation element from the SARS-CoV-2 RNA genome. *bioRxiv*.

Zoubouloglou P., García-Portugués E., & Marron J. S. (2021). 'Scaled torus principal component analysis', arXiv, arXiv:stat.ME2110.04758, preprint: not peer reviewed.