Tracking Internal Frames of Reference for Consistent Molecular Distribution Functions

Robin Skånberg, Martin Falk, Mathieu Linares, Anders Ynnerman, and Ingrid Hotz

Abstract—In molecular analysis, Spatial Distribution Functions (SDF) are fundamental instruments in answering questions related to spatial occurrences and relations of atomic structures over time. Given a molecular trajectory, SDFs can, for example, reveal the occurrence of water in relation to particular structures and hence provide clues of hydrophobic and hydrophilic regions. For the computation of meaningful distribution functions, the definition of molecular reference structures is essential. Therefore we introduce the concept of an internal frame of reference (IFR) for labeled point sets that represent selected molecular structures, and we propose an algorithm for tracking the IFR over time and space using a variant of Kabsch's algorithm. This approach lets us generate a consistent space for the aggregation of the SDF for molecular trajectories and molecular ensembles. We demonstrate the usefulness of the technique by applying it to temporal molecular trajectories as well as ensemble datasets. The examples include different docking scenarios with DNA, insulin, and aspirin.

Index Terms—Molecule Visualization, Molecular Dynamics, Interactive Exploration

1 INTRODUCTION

Molecular dynamics (MD) simulations are routinely used in many research fields such as drug design and molecular spectroscopy. A central component in the analysis of MD data is the understanding and exploration of spatial relations between different atomic structures and molecules. This task is often approached by aggregating the occurrences of structures over time into a spatial distribution function (SDF) [1]. An example is the distribution of water in relation to a specific molecule to show hydrophilic vs. hydrophobic regions. Another example is the distribution of biomarkers in relation to chains from amyloid fibrils, see Fig. 1. To ensure a physically meaningful SDF it is important to perform the aggregation in an appropriate frame of reference, for example defined by a specific chain or molecule. Commonly, this is achieved by fixating the molecule of interest in the center of the simulation box via user defined forces. This, however, has the disadvantage that the fixated molecule is constrained by external forces hindering conformational changes which would otherwise occur in a natural setting. The simulation also has to be customized for one selected molecule reducing the flexibility during analysis.

In this paper, we approach this problem by instead introducing an internal frame of reference (IFR) for selected molecular structures that serve as a reference for the aggregation of the SDF and abstracts from its global movement. The internal frame of reference is defined by the atomic locations of the molecular structure and it is tracked over the trajectory. Representing the trajectory in this IFR relates the entire dynamics of the system to the selected structure which obtains a stable position. The tracking of the frame is modeled as a rigid-body transformation that ignores deformations within the selected structure. This approach requires the conformation of the structures to be sufficiently stable during the tracked time. We extract shape parameters from the selected structure and visualize them in a barycentric shape-space, providing an overview of the temporal evolution of its conformation. The user can define the structure as well as the set of atoms with a selection tool as described by Skånberg et al. [2]. Examples of structures of interest could be particular identified chains of a molecule.

We demonstrate the usefulness of the IFR for three different analysis scenarios. First, the IFR is used as an interactive exploration tool where a dataset is studied from two different perspectives. Secondly, the IFR is used for interactive aggregation of spatial distributions for selected...
structures. And, lastly, ensembles of molecular structures in large MD simulations are superimposed into a single improved SDF. The targeted user groups are experts analyzing MD-simulation data. The contributions of this paper can be summarized as:

- Internal frame of reference (IFR): Introducing the concept of an internal reference frame for labeled point sets.
- Tracking of the IFR: Formulation of the problem as a variant of Wahba’s problem and an algorithm for its solution.
- Shape stability footprint: analysis of the shape stability of molecular structures and conformations.
- Temporal SDF aggregation: aggregation of spatial distribution functions (SDF) from molecular trajectories.
- Ensemble SDF aggregation: Efficient aggregation of SDF for molecular structures from ensemble simulations by overlaying different molecular instances from the ensemble.

2 PROBLEM DESCRIPTION

The problem we are targeting in this paper stems from application specific tasks arising in the analysis of molecular trajectories. From a technical point of view it can be described as an optimization problem for the alignment of sets of labeled particles in 3D space. Both perspectives are described in the following sections.

2.1 Application Point of View

The analysis presented in this paper targets data from MD simulations typically performed with simulation packages such as Gromacs 3, Amber 4, LAMMPS 5, TINKER 6, and NAMD 7. These are highly optimized numerical solvers that integrate positions of atoms using Newton’s Laws of motion. As input, they are given an initial configuration state (position, velocity, and force models for connected substructures) given within a box with periodic boundary conditions (PBC) and temperature or pressure target. The output of such simulation are atomic trajectories (atomic positions over time) and properties (usually scalar values over time). The integration operates with a temporal resolution of the order of femtoseconds ($10^{-15}$s) while frames for the trajectories are commonly captured and stored at a rate close to picoseconds ($10^{-12}$s). Thereby it is crucial to perform the simulation within a sufficiently large context that provides a natural environment for certain interactions even if only a small part of a system is studied in the end.

One common goal of such a simulation is to understand the occurrence of structures in relation to each other. These correlations can be described by a vector expressing the separation of both structures. In early work, mainly the distribution of distance was investigated, that is the length of the vector that provides insight into the radial distribution of the structures. Svishchev and Kuaslik 1 extended this analysis by generating pair-density maps which span both the radial and the angular coordinates of the separation vector. Embedding this distribution into the physical space leads to the spatial distribution function (SDF). In practice, one of the structures is selected as the reference structure or molecule, for example, a chain of proteins. The SDF can then be interpreted as a probability distribution of the location of the second structure, for example, a biomarker as shown in Fig. 1. Due to the dynamics of the molecules, a direct aggregation of the spatial locations of the structures does not result in useful distributions. To compute the SDF in a meaningful way it is important to provide an aggregation space that is centered around the reference structure. The problem could be solved by fixating the reference structure in the center during the simulation. This approach, however, changes the natural environment of the molecules.

Another common operation offered by MD software packages is to recenter trajectories, an operation where each frame in the trajectory is spatially translated to center it on a specific structure while maintaining other structures complete across periodic boundaries. By lifting this fundamental operation into the visualization and into the hands of the user, we are able to recenter and compute spatial distribution functions for not just one, but for an ensemble of molecules and superimpose the results interactively.

2.2 Technical Problem Description

We now translate the problem of generating a structure centric aggregation space into a technical problem, which is tracking an internal frame of reference for a set of labeled particles. The set of particles given in the simulation space $S$ by their atomic positions $\{x_i^j, 1 \ldots n\}$, for each time-step $j = 0 \ldots m$, with labels $i = 1 \ldots n$ specifying the atom, and the atomic mass or weight $\{m_i, i = 1 \ldots n\}$. Tracking an internal reference frame means we need to find a transformation $T^{(j)} : S \rightarrow S^{(j)}$ for each time-step $j$, which maps the spatial configuration from the original simulation frame $S$ to a frame defined by the particle set $S^{(j)}$. This frame is called internal frame of reference (IFR) throughout the paper. A fundamental requirement is that this transformation is continuous such that it does not tear apart the space and generates holes. When defining the transformation one generally deals with two, often conflicting, goals:

1) The transformation should map particle positions from the initial configuration $\{x_i^0, i = 1 \ldots n\}$ to the particle positions at time $j \{x_i^j, i = 1 \ldots n\}$

$$T^{(j)}(x_i^0) = x_i^j$$

2) The space $S$ should be deformed as little as possible to keep the original distances between atoms.

Which of these items is given a higher priority depends on the specific goals and can largely be controlled by the class of transformations that are considered. Then the problem of finding the best transformation can be formulated as an optimization problem.

The three questions that have to be answered for our application problem can be summarized as follows:

- What is a good set of transformations that should be allowed in our setting? – (Sec. 4.1)
• Which algorithmic solution for the given set of transformations and application setting provides a stable algorithmic tracking? – (Sections 4.2 and 4.3)
• What are the requirements for an atomic structure to be suitable to serve as reference frame and allow the definition of a stable IFR? – (Sec. 5)

3 RELATED WORK

The paper builds on previous work from different domains including molecular visualization and the application domain of MD simulation. There are also other areas where similar technical problems occur. We have grouped the different areas into categories detailed below.

Satellite Attitude Determination – Wahba’s problem formulation originates in space research where it is needed to determine the altitude and orientation of space crafts such as satellites using observational vectors. Within this field, numerous solutions to the problem have been proposed like Singular Value Decomposition (SVD) based [8], Davenport’s Q-method [10], QUEST [11], FOAM [12], ESOQ [13]. The latter methods trade accuracy for computational efficiency. A summary of available techniques is given by Markley and Mortari [14] where they conclude that Davenport’s Q-method and SVD based solutions are the most numerically robust but computationally most expensive.

Shape Matching in Computer Graphics and Point-based Data – In the domain of computer graphics and soft-body simulation, Müller et al. [15] introduced a shape-matching algorithm to simulate soft bodies utilizing particle-based dynamics. This approach also requires a solution of Wahba’s problem. The idea is to superimpose a deformed configuration onto a reference configuration of a particle system. The particle positions are interpolated linearly toward the reference. The technique relies on polar decomposition to extract a rotation matrix from an affine matrix and becomes unstable when the affine matrix is not invertible making this method unsuitable for our application, see Sec. 4.

In the context of point-based mapping and reconstruction of environments, features are identified, matched, and superimposed in a process known as stitching. This is a similar problem, but it has one fundamental difference: any of the points in set A can be matched to any of the points in set B. Wahba’s problem assumes there is a distinct correspondence and an identical number of points between the frames. Therefore this domain has been excluded from related work.

Molecular Dynamics – In the MD community, this problem has also already been discussed. With the Root Mean Squared Deviation (RMSD) a measure of the deviation of atom positions for the IFR of a molecule has been introduced. To extract the IFR, the molecule is considered as a system of atoms where the rotation and translation of the system is removed. What remains is the internal motion of the atoms with respect to the molecule. A common approach for extracting the IFR is Kabsch’s algorithm [16], [17]. It superimposes all temporal configurations of a molecule throughout the trajectory into a common reference frame, usually the initial configuration, that is fit-to-first. Chevrot et al. [18] introduced an alternative approach that superimposes the temporal configurations by fitting subsequent configuration to the preceding ones, that is fit-to-preceding. Their method avoids matrix accumulation of the history which accumulates numerical errors. Instead, a representation of the atoms is kept in its IFR and is updated by fitting the next configuration directly into it.

4 INTERNAL FRAME OF REFERENCE TRACKING

In this section, we first specify the set of transformations that are considered for tracking of the internal frame of reference (IFR) of atomic structures, then we propose a stable algorithm for this class of transformations.

4.1 Transformations for Tracking Atomic Structures

As described in Sec. 2.2 tracking of the IFR requires finding a compromise between the perfect alignment of the atomic structures and the deformation of space. In our setting, tracking of atomic structures that are flexible and change their shape during the simulation, it is not possible to achieve a perfect alignment of the atoms over time without deforming space. Since geometric configurations of molecules or atomic structures are essential for the analysis process we decided not to allow deformation of space. Therefore, we restrict the transformation to linear transformations, more specifically rigid-body transformations that are a combination of translations and rotations (rOTO-translations). Thus, we can guarantee that characteristic geometrical measures of molecules such as angles and distances, which are essential information about the conformation of the molecules, are preserved. This has as a consequence that the selected atomic structure will not be entirely fixed even in the tracked frame of reference. Whether this compromise is meaningful depends on the flexibility of the selected structure. This aspect will be discussed in more detail in Sec. 5.
When limiting ourselves to rigid body transformations Eq. 1 becomes

$$T : S \rightarrow S'$$

$$\mathbf{x'} = \mathbf{R(x - t_0)} + \mathbf{t}$$  (2)

where $\mathbf{R}$ is a rotation matrix and $\mathbf{t}_0, \mathbf{t}$ are translation vectors. The translation vectors determine the center of the respective rotations in both frames. An advantage of our setting is that we don’t have to solve a correspondence problem since the atoms have unique indices. Thus, we can find the optimal rigid body transformation by minimizing a cost function $C$ that is defined as the sum of weighted squared distances of the transformed particle locations and target particle locations. This results in the following minimization problem:

Find the rotation matrix $\mathbf{R}$ and translation vectors $\mathbf{t}$ and $\mathbf{t}_0$ for a set of $n$ particles from its initial configuration $\{\mathbf{x}_i^0, i = 1 \ldots n\}$ to another configuration $\{\mathbf{x}_i, i = 1 \ldots n\}$ by minimizing the following expression

$$C = \sum_i w_i (\mathbf{x'} - \mathbf{x}_i)^2 = \sum_i w_i (\mathbf{R(x}_i^0 - \mathbf{t}_0) + \mathbf{t} - \mathbf{x}_i)^2$$  (3)

where $w_i$ is a weight associated to each point.

The translation vectors can be determined by setting the partial derivatives of Eq. (3) with respect to $\mathbf{t}$ and $\mathbf{t}_0$ to zero resulting in [13]:

$$\mathbf{t}_0 = \frac{\sum_i w_i \mathbf{x}_i^0}{\sum_i w_i}$$ and $$\mathbf{t} = \frac{\sum_i w_i \mathbf{x}_i}{\sum_i w_i}.$$  (4)

If the atom mass $m_i$ is chosen as weight $w_i$, $\mathbf{t}_0$ and $\mathbf{t}$ correspond to the center of mass for each system and the rotations occur around the center of mass, a result supported by intuition. With the solutions for $\mathbf{t}_0$ and $\mathbf{t}$, Eq. 5 can be expressed as a rotation around relative coordinates $\mathbf{u}_i$ and $\mathbf{v}_i$

$$\mathbf{u}_i = \mathbf{x}_i^0 - \mathbf{t}_0$$ and $$\mathbf{v}_i = \mathbf{x}_i - \mathbf{t},$$  (5)

resulting in

$$C = \sum_i w_i (\mathbf{Ru}_i - \mathbf{v}_i)^2.$$  (6)

With this formulation, the only unknown is the rotation matrix $\mathbf{R}$, which by definition is both orthogonal and has a determinant of 1. This is a well studied problem and is known as Weighted Orthogonal Procrustes Problem (WOPP) or more commonly known as Wahba’s problem [15].

Wahba’s problem seeks to find a rotation matrix $\mathbf{R}$ between two coordinate systems given a set of vector observations $\mathbf{u}_i, \mathbf{v}_i$ with corresponding weights $w_i$ which minimizes a cost function $C$.

$$C(\mathbf{R}) = \frac{1}{2} \sum_{i=1}^{N} w_i \|\mathbf{Ru}_i - \mathbf{v}_i\|^2$$  (7)

1. In greek mythology Procrustes was a bandit who tortured and mutilated his victims to fit his iron bed.

4.2 Algorithm for Computing Rotation

There have been several algorithmic solutions proposed for Wahba’s problem. As long as the set of particles spans an approximately isotropic volume most solutions are applicable. However, as soon as the volume degenerates to lower-dimensional structures, either planar or linear, most algorithms become unstable. Especially the transformations for linear structures exhibiting rotational symmetries require additional constraints. In the following, we introduce some of the commonly used algorithms and evaluate them with respect to this.

A common solution to find the optimal rotation matrix $\mathbf{R}$ is to first compute the optimal affine transformation matrix $\mathbf{A}$ that maps the relative coordinates $\mathbf{v}_i$ onto $\mathbf{u}_i$, Eq. 5, in a least-squares linear sense and then extract the rotational part from this matrix. The matrix $\mathbf{A}$ can be derived as the fraction between the mass-weighted cross-covariance matrix $\mathbf{K}_{uv}$ and the mass-weighted covariance-matrix $\mathbf{K}_{vv}$ where $\mathbf{u}$ and $\mathbf{v}$ correspond to the sets of relative coordinate vectors.

$$\mathbf{K}_{uv} = \sum_i w_i \mathbf{u}_i \mathbf{v}_i^T$$  (8)

$$\mathbf{A} = \frac{\mathbf{C}_{uv}}{\mathbf{C}_{vv}}.$$  (9)

With the observation that $\mathbf{K}_{uv}$ is a symmetric matrix and only contains scaling information, one can conclude that $\mathbf{K}_{uv}$ holds all the information about the rotation. Therefore we can restrict our considerations to the matrix

$$\hat{\mathbf{A}} = \mathbf{K}_{uv}$$  (10)

used as input for computation of the rotation matrix $\mathbf{R}$. The rotation can then be extracted using the polar decomposition, which is probably the most common approach to compute the rotation matrix. Müller et al. [15] used this approach for the interpolation of shapes of deformable point based objects starting with a shape matching step. The rotation matrix is then given as

$$\mathbf{R} = \hat{\mathbf{A}} (\sqrt{\hat{\mathbf{A}}^T \hat{\mathbf{A}}})^{-1}$$  (11)

However, solving this equation becomes unstable when Eq. 11 lacks an inverse. In the solution proposed by Müller et al. they circumvent this problem by later linearly interpolating $\mathbf{R}$ with $\mathbf{A}$, mitigating some of the instability while preserving some of the affine transformations. The resulting instability of the rotation matrix $\mathbf{R}$ does not pose a problem as it is only used as an initial guess for the target positions toward which the particles are later interpolated. However, in the context of our application, this instability proves problematic and manifests frequently as molecular structures can be inherently planar (2D) or even linear (1D). For a more detailed analysis we refer to Markley and Mortari [14].

As an alternative, we have chosen to use a solution based on Singular Value Decomposition (SVD), as proposed by Kabsch et al. [16, 17]. In their work they provide a solution that avoids the polar decomposition by extracting the matrix $\mathbf{R}$ from $\hat{\mathbf{A}}$ using SVD

$$\hat{\mathbf{A}} = \mathbf{USV}^T.$$  (12)
The scaling component $S$ is then replaced by a matrix that ensures a right-handed rotation scaling the last column vector by the determinant $d = \det(UV^T)$. The rotation matrix $R$ can then be expressed as

$$R = U \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & d \end{pmatrix} V^T.$$  

(13)

Compared to polar decomposition, Eq. (11), Kabsch’s approach using SVD, Eq. (13), is robust as it does not require any inversion. For the case of linear structures the result is not unique. As a result the tracking of flexible structures that are close to linear is still unstable.

### 4.3 Stable Tracking of the Internal Frame of Reference

We now assemble the results from the previous subsections to define the algorithm for tracking the IFR for atomic structures. Principally, there are two different approaches for the tracking. The first option is to consider every configuration of the trajectory individually and align it with the initial configuration, referred to as *fit-to-first* ($ftf$). The second option is to align every time-step of the trajectory to the previous one, referred to as *fit-to-preceding* ($ftp$), which both have their advantages.

For the first approach ($ftf$), the main ingredient is a set of rotations $R(n)$ that superimpose the initial configuration $r(0)$ with every subsequent configuration $r(n)$ of the trajectory (Fig. 2). Thereby $r$ denotes relative coordinates which have been centered around the configurations center of mass, Eq. (6), and $n$ refers to the $n$th frame in the trajectory.

$$R(n) = R^\text{int}_{n \rightarrow 0} = r(n) \rightarrow r(0)$$  

(14)

This aligns with Kabsch’s method which superimposes every configuration into a common reference frame, in our case frame $n = 0$. This approach results in an array of rotations for every time-step, compare Fig. 2. The disadvantage of this approach is that for the case of structures that are close to linear, where the rotation matrix is sensitive to small changes in the geometry, one cannot guarantee a continuous change of the rotation matrix from one time-step to the next.

Instead of relating every time-step to the original configuration one can also consider the optimal rotation between any two given configurations $r(n)$ and $r(m)$. We denote this operation by $r(n) \rightarrow r(m)$ which yields a matrix $R_{n \rightarrow m}$. This option is used for the second approach ($ftp$) introduced by Chevrot et al. [18]. The algorithm in Eq. (15) maintains a copy of the atomic structure in its IFR $R^\text{int}(n)$, which is initialized to the first configuration $r(0)$. In an iterative fashion, $D$ is computed as the optimal rotation between $r^\text{int}(n)$ and the configuration of the next frame $r(n + 1)$. Lastly, the internal configuration is updated by transforming the next configuration $r(n + 1)$ into the IFR using $D^T$, resulting in $r^\text{int}(n + 1)$.

$$r^\text{int}(0) = r(0)$$

$$D = r^\text{int}(n) \rightarrow r(n + 1)$$

$$r^\text{int}(n + 1) = D^T r(n + 1)$$

$$R(n + 1) = R^\text{ftp}_{n+1 \rightarrow 0} = D^T$$

(15)

The $ftp$ approach resolves the set of rotations $R(n)$ by minimizing the rotation between frames, which is ideal for perceptual stability. However, the trade-off is accuracy. The $ftf$ approach will always resolve the optimal rotation to the reference. Fig. 3 shows the development of the rotation angles of both approaches in comparison as one example. In our implementation both techniques are used but in different contexts, as exemplified in Sec. 7.

### 5 Stability Footprint of Atomic Structures

This section deals with the last question raised in Sec. 2.2 when an atomic structure is suitable as a reference structure. Since the frame-tracking technique, introduced above, only

Fig. 2. Array of rotations $R(n)$ which aligns the initial configuration at time $t = 0$ onto every subsequent configuration in the trajectory.

Fig. 3. Angular distance within the IFR of a structure. The y-axis represents angular distance in degrees to the initial configuration at frame 0. $ftf$ corresponds to *fit-to-first* while $ftp$ corresponds to *fit-to-preceding*. $\Delta$ represents the angle difference between the two approaches.

Fig. 4. Shape variation plot. Each discrete time point of the trajectory is represented as a point colored in function of time. The three outlined points (yellow) serve as examples of the three extremes (linear, planar, and spherical).
supports rigid-body transformations, it implicitly assumes that the tracked structures are rigid, or at least close to rigid, and do not undergo substantial conformational changes. To validate the conformational stability of a structure a shape-plot depicts the temporal evolution of the geometrical anisotropy. The visual representation shown in Fig. 4 is inspired by methods developed for tensor classification in the context of diffusion tensor imaging (DTI) [27]. It uses a barycentric space where each dimension represents one extreme of geometrical anisotropy (linear, planar, or spherical). These measures are extracted from the eigenvalues \( \lambda_1 \geq \lambda_2 \geq \lambda_3 \) of the mass-weighted covariance matrix \( \mathbf{K}_{vv} \) and are defined as

\[
\begin{align*}
C_l &= \frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3} \\
C_p &= \frac{2(\lambda_2 - \lambda_3)}{\lambda_1 + \lambda_2 + \lambda_3} \\
C_s &= \frac{3\lambda_3}{\lambda_1 + \lambda_2 + \lambda_3} .
\end{align*}
\]

The descriptors \( C_l, C_p, \) and \( C_s \) correspond to linear, planar, and spherical/isotropic shapes, respectively. They represent the barycentric weights which span the shape-space. An ideal target structure for tracking would produce a small footprint (area) in this space thus exhibiting structural stability. Fig. 4 shows the footprint of an alanine chain, a structure with extreme flexibility, which spans the majority of the shape-space.

The shape-plot is a scatter plot in this barycentric space. Each point represents the geometrical shape of the tracked structure at one frame in the trajectory. The points are colored based on the normalized time (Fig. 4). The shape-space plot is exposed as an interactive tool and is coupled with a spatial visualization of the molecule supporting interactive selection. It enables exploration of the temporal configuration space. Outliers are clearly visible and can directly be inspected through interaction.

6 Aggregation of Spatial Distribution Functions

In this section, we describe how we use the tracking of IFRs for the aggregation of Spatial Distribution Functions (SDF). First, we follow atomic structures in individual trajectories over time, then we explain how ensembles of instances of the same molecular structures can be used to generate more expressive SDFs. For implementation details see Sec. 8.

6.1 Temporal SDF Aggregation

Understanding the interaction of molecules or atomic structures comprises questions such as ‘Where in molecule X do we have a high probability of finding element Y?’ Frequently such questions are approached by computing spatial distribution functions (SDF) of structures that are aggregated over the time of the simulation. While the aggregation of the SDF of molecules X and Y provides some information about the movement of the molecules it does not answer the above question. Physically relevant information can only be obtained if one of the structures, in this case molecule X, is fixated in the frame and the SDF of Y is aggregated with respect to this frame. For this purpose, the IFR of the reference structure X, as introduced above, is used. This temporal super-positioning of the second structures Y generates the SDF. This means that the locations of the structures Y are transformed into this space to populate a 3D histogram. This temporal super-positioning is illustrated as the horizontal operation in Fig. 5 step 3 where it has been applied to the three individual molecules (red, green, and blue).
6.2 SDF from Ensembles of Molecular Instances

For this type of SDF computation, we extend the superposition of structures over time by superimposing the SDF of multiple instances of equivalent structures in a system. This spatial superimposition is illustrated in Fig. 5 step 3 as vertical operation. The spatiotemporal superimposition combines both axes in this figure. This is useful when studying molecular dynamics where multiple equivalent structures are present. Then it is possible to compute temporal SDFs for each molecule individually and then superimpose the SDFs into one spatiotemporal SDF. The motivation for this is simply to obtain more indicative results. If many equivalent molecules are present in a dynamic, they can all contribute with valuable data generating more expressive SDFs.

7 APPLICATION SCENARIOS

In this section, we demonstrate how the IFR can be used for data exploration and serve as a basis for computing spatial distribution functions. We start with a brief introduction of the datasets before describing three application cases.

7.1 Datasets

Dataset 1: Aspirin – This dataset has been generated starting from the 1OXR entry [28] of the Protein Data Bank of the complex formed by Aspirin and phospholipase A2, (Fig. 6(a)). From this crystal structure a simulation box comprising 50,515 atoms (1,189 for the phospholipase A2, 20 for the aspirin, one calcium ion, four sodium ions, and water) was built. The simulations were performed with the GROMACS software [3] using the GROMOS54A7 force-field [29]. During the simulation the aspirin was pulled out of the pocket using an umbrella sampling procedure. 600 frames were saved over the 10 ns long dynamic. This exemplifies a typical case of molecule-protein interaction to illustrate the opportunity of using a reference frame for recentering of trajectories and aggregation of SDFs which is extremely important for structure-based rational drug design.

Dataset 2: DNA – This dataset [30] investigates the DNA binding mode of a dithienylethene derivative (DTE) which is a candidate for molecular photoswitches. Photoswitches are a novel approach in medicine where a drug is activated and deactivated using light. This method is of great promise in photopharmacology. In this simulation, the DTE molecule bind in the minor groove of a DNA segment (Fig. 6(b)). The dynamic consists of 1,373 atoms (1,278 atoms in DNA, 59 in DTE, the remainder are ions) captured over 1,896 frames. The solvent has been omitted from this trajectory.

Dataset 3: Insulin – This dataset [31] shows the targeting of an insulin β-sheet by a spin-label molecule for electron paramagnetic resonance (EPR) imaging, shown in Fig. 6(c). The purpose of this simulation is to establish the specific interactions between the spin-label molecule and the insulin β-sheet. The dynamic consists of 732 atoms (654 for the
protein, 73 for the spin-label molecule, the remainder are ions) captured over 5,000 frames. The solvent has been omitted from this trajectory.

**Dataset 4: Para-Nitroaniline** – This dataset has been generated for the present study with the GROMACS software [3] using the GROMOS54A7 force-field [29] and comprises 50 4-nitroaniline molecules solvated together with 50 sodium and 50 chlorine ions in water (Fig. 6(d)). The dynamic consists of 97,914 atoms (800 para-nitroaniline, 100 ions, the remainder is water) captured over 1,765 frames. The molecule (para-nitroaniline) presents two groups at its extremities (NH$_2$ and NO$_2$) that will interact in opposite manner with water and the Cl$^-$ and Na$^+$ ions present in the solution. Indeed, NH$_2$ is going to interact strongly with the oxygen atom of water and the Cl$^-$ ions, whereas NO$_2$ is going to interact strongly with the hydrogen atoms of water and the Na$^+$ ions.

**Dataset 5: Amyloid Fibril** – This data set comprises 100 luminescent biomarkers (p-FTAA) in interaction with an amyloid fibril (Fig. 6(e)). These biomarkers are used to identify the protein misfold characteristic of Alzheimer’s disease. They bind to a specific pocket that was identified in a previous study [32]. The dynamic consists of 161,742 atoms (158,631 protein and 3,111 p-FTAA). The solvent has been omitted from this trajectory.

### 7.2 Recentering Atomic Structures for Exploration

Understanding the complexity of the molecular dynamics is, in many cases, relying on interactive analysis of the simulation results. This entails exploration of structures of interest in detail. These structures should remain the focus of the visualization during the session. This requires easy ways to select structures and to stabilize them in the scene.

For the selection of structures of interest, we build on the selection tool by Skånberg et al. [2]. The main frame of reference is then attached to the selected structure and can be used to navigate the scene. Since in this application a perceptually smooth transition between the frames is essential, the fit-to-preceeding approach, see Sec. 4.3 Eq. (15), is applied resulting in an array of rotations $R(0)...R(N)$. To reconstruct the intermediate reference frames without discontinuities, the rotations are interpolated using a spherical cubic interpolation scheme (SQUAD) [33]. This ensures a perceptually smooth motion of atoms even when viewed from an IFR.

Fig. 7 depicts an example where an aspirin molecule is interacting with a protein. The figure shows two different perspectives, one focusing on the protein (top row) and one focusing on the aspirin (bottom row). The structure of interest is emphasized by saturated colors. The aspirin molecule is represented by VDW spheres which are colored based on atom type. The protein is represented as ribbons with licorice colored by residue type. Both rows depict the dynamics of the scene from different viewpoints. The chosen points in time are the same snapshots for both frames.

### 7.3 Spatial Distribution Functions

In this use case, we follow the common approach of aggregating SDFs by superimposing the reference structure...
temporally (Sec. 6.1). This means using the IFR of the selected structure for populating the histogram.

As a first example, the temporal SDF is applied to Dataset 2, a molecule dynamic containing DTE and a DNA-sequence (Fig. 8(b)). The temporal SDF of DTE with respect to the simulation frame is a cloud distributed along the entire DNA (Fig. 8 left). It does not represent the region where the DTE occurs frequently since both DNA and DTE are moving freely within the simulation box. If the SDF is aggregated in the IFR of DNA (Fig. 8 right) it provides a different picture showing one distinguished high-density region in the minor groove. The small footprint of the DNA in the shape variation shows a very stable configuration.

The same technique is applied to Dataset 3, a molecular dynamic containing a spin-label molecule used for EPR imaging of insulin $\beta$-sheets (Fig. 6(c)). Similar to the previous example with the DNA, the SDF computed in the original frame is not very localized (cf. Fig. 8 left). It is not possible to draw any conclusions about the interaction between the molecules. However, by using the IFR of the protein, the accumulated distribution function is located at a specific site of the protein (Fig. 9 right). The footprint of the protein in the shape variation plot is noticeably larger compared to the DNA example. This shows that the protein is more flexible. While the footprint is larger, the distribution of the colored points follows a pattern. This means that the protein is transitioning smoothly within its configuration space, from lower left, to mid right, to upper left, and lastly bottom right.

**7.4 Distribution Functions from Structure Ensembles**

In this application, we extend the superimposition to multiple instances of equivalent structures in a system.

The first example is a dynamic containing 50 para-nitroaniline molecules solvated in water (Fig. 10). The image shows the spatial distribution function of oxygen (originating from water) in relation to the molecule (para-nitroaniline). In the top row the SDF is using the conventional temporal superimposition of a single molecule. The bottom row shows the spatiotemporal superimposition of the entire ensemble of para-nitroaniline which reveals distinct areas of high density which are the preferred locations of oxygen in the vicinity of para-nitroaniline. In the center, all instances of para-nitroaniline have been overlaid and show good alignment among individual molecules with the most variation in the extremities. The shape variation of the ensemble shows a larger footprint than that for a single molecule.

In the second example, a dynamic containing 254 chains that form an amyloid fibril is used (Figs. 1 and 11). The images show the spatial distribution of the biomarkers p-FAAA present in the dataset in relation to the ensemble of chains. The IFRs of the ensembles are defined by each complete chain including their side chains as shown in the figure. The superimposition of the individual chains shows a relatively stable configuration of the backbone while the side-chains exhibit more flexibility. The density indicates a strong presence within a specific region visible on the right side and is confirmed by the pocket identified by König et al. [32]. Moreover, the density also hints to the presence of a second region (top left), which could not be identified by visual inspection of the MD trajectory before. This illustrates the benefits of aggregating the spatial distribution to highlight zones of lower probabilities.

**8 IMPLEMENTATION AND PERFORMANCE**

The IFR computation, the interactive recentering, the SDF computation, and the shape-space plot have been prototyped in VIA-MD [25]. VIA-MD provides the necessary tools to select the atomic structures of interest [2]. VIA-MD has been manifestly designed as an interactive exploration tool, and priority in the choices of functionality and implementation is thus given to support high performance and precision in interaction.

To superimpose a structure, two sets of configurations (u and v in Eq. (8)) are required. These configurations may originate from different frames within the trajectory (temporal) or from different structures within the same frame (spatial). For temporal superimpositions, this operation is straightforward since any structure and its atomic indices remain the same throughout the trajectory. For spatial superimpositions the user can provide a list of matching structures or select a reference structure and equivalent structures within the dataset are found automatically. Our prototype supports the latter since the explorative character of the tool with rapid iteration time is central. Equivalent structures are determined using an equivalence measure. In
a similar task identifying nucleotides in RNA and DNA structures Lindow et. al [34] consider atoms and their covalent bonds as an undirected graph and classify the problem as NP-complete. However, due to the limited size of the nucleotides and number of variations, brute-force techniques suffice. We cannot make such assumptions, since a selected structure may have an arbitrary size. Instead, we consider two structures as equivalent if the sets of atoms have the same length and each index refers to the same element. Structures do not need to be contiguous, they may contain holes as long as the existing atoms match and have the same local index. This formulation is convenient because we can express the reference structure as a mask that is swept over all atoms to identify matches.

The computation of IFRs can be broken down into two parts: computing IFRs and computing spatial densities using the reference frames. Computing IFRs for structures involves the computation of the cross-covariance matrix, extraction of the rotation matrix using SVD, and transforming the next configuration into the IFR. The operation has the overall time complexity of $O(N \cdot M \cdot P_r)$ where $N$ is the number of frames in the trajectory, $M$ is the number of reference structures in the simulation, and $P_r$ is the number of reference structure atoms. Computing the spatial densities involves the transformation of every target atom into the IFR and incrementing the bins of the histogram. This operation has a time complexity of $O(N \cdot M \cdot P_t)$ where $N$ is the number of frames, $M$ is the number of IFRs (structures), and $P_t$ is the number of target atoms. Table 1 lists timings for computing IFRs and SDFs together with the relevant sizes.

9 Limitations

The quality of an SDF is governed by the ability to superimpose the reference structures (both temporally and spatially). If a highly flexible or conformationally unstable structure is chosen to serve as a reference, the rigid-body superimposition scheme will most certainly prove insufficient. The result will be blurry as the distribution is radially scattered due to the lack of definitive orientation. These cases can be identified by providing some type of stability analysis for example the shape-space plot (cf. Fig. 4). If such a scenario arises, a smaller substructure can be chosen within the assembly of the structure to serve as a reference.

The use of three anisotropic descriptors as done in Sec. 5 will not capture all cases of structural instability. Theoretically, it will not capture substructure rotations or torsions such as rotations around single bonds if the substructures on each side of the bond are structurally equivalent. To cover such cases, one would need to consider higher-order shape descriptors such as bending and twisting. However, even though the current technique has theoretical limitations it provides an overview which has proven sufficient for all tested datasets.

10 Conclusion and Future Work

Spatial distribution functions are powerful tools in the analysis of correlations of atomic structures. We have presented a technique to aggregate spatial distribution ensembles using equivalent molecules given in a single molecular dynamic. The technique is applicable when several equivalent molecules are put into the simulation to capture important
inter-molecular behavior. The spatio-temporal aggregation provides an entirely novel means of exploring MD trajectories. The result depends on the ability to superimpose the structure since a rigid-body transformation is used for superimposition. The more rigid a structure is, the better the result. To assert the geometrical stability of a molecular-structure the technique is coupled with an interactive space-plot to show the geometrical anisotropy throughout the simulation.

The decision to only allow rigid-body transformations was motivated by the desire to maintain characteristic geometric properties of the molecules. In some cases, more flexible transformations that also support a deformation of the space might be of interest. One way to extend our approach would be to use radial basis functions (RBF) to provide better alignment between configurations. Correa et al. demonstrated such a technique by applying shape deformation in computer graphics using support points. This would, however, violate constraints such as distance and angle conservation that was specified as a requirement in the introduction. But, if purely used for generating SDF data and not to recenter or reorient the trajectory, this would be a viable option.

The shape-space plot could be extended to capture higher-order descriptors such as twisting and bending. Müller et al. considered quadratic deformations for shape matching. Similar ideas could be used to develop higher-order anisotropic descriptors.

There are other applications with moving features that could also benefit from tracking feature-centered reference frames that could be explored in the future. An example of such an application is flow visualization where the corresponding features of interest are vortices.

It should be noted that the technique has been developed as post processing visualization. However, we do not see any limitations in running SDF ensemble computations together with the simulation itself, thereby supporting sampling with high temporal resolution. The caveat is that in this case structures of interest need to be defined prior running the simulation.

ACKNOWLEDGMENTS

This work was supported through grants from the Excellence Center at Linköping and Lund in Information Technology (ELLIIT) and the Swedish e-Science Research Centre (SeRC). The authors thanks the Swedish National Infrastructure for Computing (SNIC) for providing computing resources.

**REFERENCES**


Robin Skånberg

Robin Skånberg received his Master’s degree in media technology in 2015. After spending one year at Ulm University in Germany as a PhD Student he moved back to Sweden where he continues as a PhD student in the group for Scientific Visualization at Linköping University since 2016. His main research focus is molecular visualization.

Martin Falk

Martin Falk is a senior lecturer in the Scientific Visualization Group at Linköping University. He received his Ph.D. degree (Dr.rer.nat.) from the University of Stuttgart in 2013. His research interests include large-scale volume rendering, visualizations in the context of pathology, systems biology, and cryo-EM, large spatiotemporal data, topological analysis, glyph-based rendering, and GPU-based simulations.

Mathieu Linares

Mathieu Linares obtained his PhD in Theoretical Chemistry at the Paul Cézanne University (Marseille, France) in 2005, where he developed a valence bond method under the supervision of Prof. S. Humbel. In 2006, he joined the Laboratory for Chemistry of Novel Materials in Mons (Belgium) for a post-doctoral stay where he worked on the modeling of self-assembly in solution and at surfaces. He then worked as postdoctoral researcher in the Computational Physics group at Linköping University (Linköping, Sweden) with Prof. S. Stafström (2008-2009) and in the group of Theoretical Chemistry at the Royal Institute of Technology (Stockholm, Sweden) with Prof. Hans Ågren (2010). He currently hold an associate Professor position shared between the laboratory of organic electronics and the group of scientific visualization at Linköping University (Linköping, Sweden). His main research interests are chirality, self-assembly in solution and on surface, charge transport in organic materials, and molecular visualization.

Anders Ynnerman

Anders Ynnerman received a Ph.D. 1992 in physics from Gothenburg University. During the early 90s he was doing research at Oxford University, UK, and Vanderbilt University, USA. In 1996 he started the Swedish National Graduate School in Scientific Computing, which he directed until 1999. From 1997 to 2002 he directed the Swedish National Supercomputer Centre and from 2002 to 2006 he directed the Swedish National Infrastructure for Computing (SNIC). Since 1999 he is holding a chair in scientific visualization at Linköping University and he is the director of the Norrköping Visualization Center - C, which currently constitutes one of the main focal points for research and education in computer graphics and visualization in the Nordic region. The center also hosts a public arena with large scale visualization facilities. He is also one of the co-founders of the Center for Medical Image Science and Visualization (CMIV).

Ingrid Hotz

Ingrid Hotz is a Professor in Scientific Visualization at Linköping University in Sweden and has an affiliation with the Center for Medical Image Science and Visualization (CMIV) in Linköping. She received her PhD degree from the Computer Science Department at the University of Kaiserslautern, Germany. Her research interests lie in data analysis and scientific visualization, ranging from basic research questions to effective solutions to visualization problems in applications. This includes developing and applying concepts originating from different areas of computer sciences and mathematics, such as computer graphics, computer vision, dynamical systems, computational geometry, and combinatorial topology.