IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 28 February 2025, accepted 2 April 2025, date of publication 9 April 2025, date of current version 18 April 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3558781

RESEARCH ARTICLE

GEODES: Geometric Descriptors for the Assessment of Global and Local Flexibility of Proteins During Molecular Dynamics Simulation

KARINA PATS[1,2], IGOR GLUKHOV[2], STEPAN PETROSIAN[3], MARIA MAMAEVA[3], ALEXEY SERGUSHICHEV[2,4], MARIE-DOMINIQUE DEVIGNES[5], AND FERDINAND MOLNÁR[1], (Member, IEEE)

[1]Department of Biology, School of Sciences and Humanities, Nazarbayev University, 010000 Astana, Kazakhstan
[2]Institute of Applied Computer Sciences, ITMO University, 197101 Saint Petersburg, Russia
[3]Bioinformatics Institute, 197342 Saint Petersburg, Russia
[4]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO 63110, USA
[5]CNRS, Inria, LORIA, Université de Lorraine, 54000 Nancy, France

Corresponding authors: Karina Pats (karina.m.pats@gmail.com) and Ferdinand Molnár (ferdinand.molnar@gmail.com)

This work was supported by the Nazarbayev University under Collaborative Research Proposal 091019CRP2108 to F.M., ITMO University grant No. 621314 for Ph.D. and Master students, Modelling Biomolecules and theirs Interaction – Data Science for Health (MBI-DS4H) platform hosted by Inria/Loria and funded by Contrat Plan État Région, Innovations, Technologiques, Modélisation & Médicine Personalisée (CPER IT2MP) and Fonds Européen de Développement Régional (FEDER).

ABSTRACT Molecular dynamics simulations offer insights into macromolecular structures and functions through extensive time-series atomic data. Two widely used metrics for assessing these simulations are RMSD and RMSF. While RMSD measures conformational convergence, it suffers from degeneracy, where different conformations can produce identical values relative to a reference. RMSF indicates relative mobility but lacks temporal specificity. As both metrics provide only global perspectives, there is a pressing need for novel metrics to capture local flexibility with temporal resolution. We introduce GEODES, a novel complementary 3D geometrical descriptor approach, and compare its effectiveness with conventional analyses using RMSD and RMSF. Through molecular dynamics simulations of the vitamin D receptor trimeric complex, we demonstrate that GEODES significantly enhances the molecular dynamics analysis workflow, offering deeper insights into the structural dynamics and interactions of this complex. This innovative and versatile approach holds great promise for applications in drug discovery, structural biology, and bioinformatics. The GEODES Python3 script toolbox is available at https://github.com/rinnifox/GEODES

INDEX TERMS Calcitriol, GEODES, geometrical descriptors, local and global stability, molecular dynamics trajectory analysis, RMSD and RMSF, structural flexibility, temporal and spatial dynamics, transcriptional coactivator, vitamin D receptor.

I. INTRODUCTION

Molecular dynamics (MD) simulations have become a standard technique for understanding macromolecular structure-function relationships [1]. These simulations generate trajectories, time-series data of atomic coordinates, which can be analyzed using statistical mechanics to predict biomolecular behavior or compare with experimental data [2]. Modern computational advances now enable MD trajectories spanning microseconds [3] to milliseconds [4], with simulated systems of millions of atoms producing multi-terabyte trajectory files. Processing these large files has become a bottleneck in computational workflows [5], which requires new analysis approaches since frame-by-frame exploration is no longer practical [6]"

Two key parameters have been established for traditional MD trajectory analysis. One, the root-mean-squared deviation (RMSD), quantifies the distance between two aligned biomolecules [7]. In MD analysis, RMSD (Equa-

The associate editor coordinating the review of this manuscript and approving it for publication was Eunkyoung Park.

tion 1) typically measures the dissimilarity of the trajectory conformational ensemble to a selected reference, providing insight into overall movement from the initial point [8].

$$RMSD(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i(t) - r_i^{ref})^2} \quad (1)$$

where N is the number of atoms in the selection, r_i^{ref} is atom i's 3D position at the reference time, and $r_i(t)$ is atom i's 3D position at time t after superimposing the system onto the reference frame. For large systems, the calculation can be accelerated by restricting the atom set to only α-carbons (Cα).

While stable RMSD values may indicate conformational convergence, this metric suffers from degeneracy, as multiple conformations can have the same Cα-RMSD relative to the reference frame, potentially masking exploration of new equidistant states. An alternative is computing all pairwise RMSD (2D-RMSD), which is more time-consuming [9]. Although widely used, RMSD lacks information about local, temporal dynamics and does not fully capture actual binding modes in molecular docking comparisons to experimental observations [8]. RMSD is primarily a time-related global descriptor of 3D structure variations during simulation. These limitations necessitate complementary tools to address missing features in MD trajectory analysis.

Complementary to RMSD, the root-mean-square-fluctuation (RMSF) (equation 2) measures the displacement of a particular atom or group of atoms relative to the reference structure throughout the simulation [10]. For amino acid residue j:

$$RMSF_j = \sqrt{\frac{1}{T}\sum_{t=1}^{T} < (r_j(t) - r_j^{ref})^2 >} \quad (2)$$

where T is the trajectory time over which the RMSF is calculated, r_j^{ref} is the 3D position of atoms in residue j at the reference time, $r_j(t)$ is their position at time t after superposition onto the reference structure, and angle brackets indicate averaging over the selected atoms in residue j.

RMSD quantifies structural divergence from a reference over time, while RMSF highlights the most mobile areas relative to the reference frame. However, RMSD lacks information on specific local movements, and RMSF does not indicate when these movements occur. Therefore, both are global metrics, pointing to the need for establishing new metrics that address local dynamics over time.

Protein structures have been described using qualitative approaches, which include amino acid sequences, functions, folding types, secondary structure content and quantitative descriptors. Quantitative descriptors, that can be either calculated or experimentally determined, provide more detailed information [11], including volume [12], surface area [12], physico-chemical [13], and structural descriptors [14], [15]. These descriptors are primarily applied for protein classification systems, where the two most common are the structural classification of proteins (SCOP) [16] and classification-architecture-topology-homology (CATH) [17]. Studies have explored various applications of descriptors to protein classification [11], [18], [19], [20]. Other applications include identifying binding sites [21], protein-ligand interactions [22], and nanoparticle-protein interactions [23]. However, descriptor-based approaches remain limited in MD simulation analysis.

In this study we introduce GEODES, a novel 3D geometrical descriptor approach, comparing it to conventional RMSD/RMSF analyses. Our findings show that GEODES significantly complements classical MD simulation analysis workflows. We demonstrate its effectiveness by analyzing multiple MD simulations of the vitamin D receptor (VDR) trimeric complex, including a coactivator peptide and calcitriol as a ligand. VDR, activated by calcitriol, belongs to the nuclear receptor (NR) family of transcription factors with canonical domain organization [24]. This includes DNA- and ligand-binding domains (LBDs), with the latter crucial for ligand binding, coactivator interactions, and receptor heterodimerization [25]. VDR plays diverse roles in 37 tissues [26], from regulating calcium/phosphate homeostasis in bone development [27] to influencing cellular processes and immune response [28], [29]. VDR's versatility makes it an intriguing target, as understanding its structural dynamics and interactions is essential to studying its diverse molecular functions. Introducing GEODES for VDR analysis enhances our understanding of this receptor while highlighting the approach's applicability. We anticipate GEODES will have implications in drug discovery, structural biology, and bioinformatics.

II. MATERIALS AND METHODS

A. MOLECULAR DYNAMIC SIMULATIONS

MD simulations used trimeric complexes of human VDR (hVDR), calcitriol, and steroid receptor coactivator-1 (SRC-1/NCOA1), based on the refined hVDR-calcitriol complex (PDBID: 1DB1) (Figure 1). This hVDR model had higher resolution and fewer missing residues than the original 1DB1 (Rochel N., personal communication). Missing residues (118, 119, 375-377, 424-427) were added using MOE software suite (Chemical Computing Group, ULC) with subsequent minimization. The lack of SRC-1 peptide from 1DB1 was addressed by superimposing a rat VDR-MED1 peptide complex (PDBID: 1RK3) and its subsequent mutation to yield hSRC-1. The hSRC-1 NR interaction domain contains three LXXLL-motifs that bind to VDR LBD, referred to as L1 (LVQLL), L2 (LHRLL) and L3 (LRYLL) (Figure 1A). The full hSRC-1 sequence was retrieved from UniProt (Q15788).

Complexes were prepared using the Maestro Suite (Schrödinger, LLC), including preprocessing steps such as the addition of hydrogens, missing side chain filling using the Prime module, H-bond optimization with PROPKA at pH 7.0, and energy minimization of the whole complex. MD simulations were performed using the Desmond package with the OPLS3e force field. The complex was placed in a 61

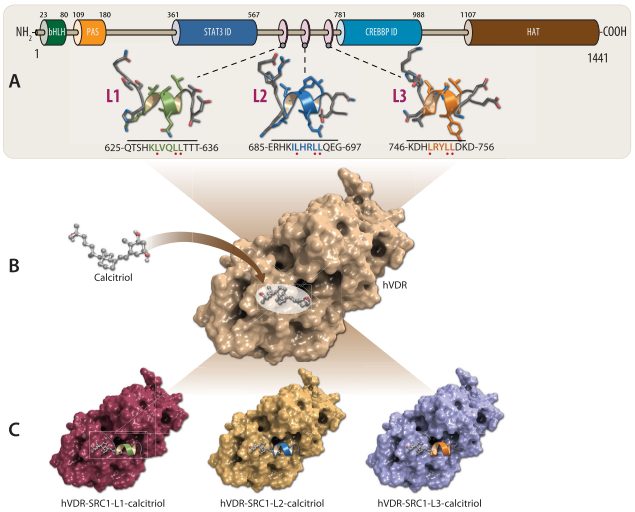

FIGURE 1. **Proteins complexes used for MD simulation. (A) SRC-1, a member of a p160 coactivator family, is a multidomain protein containing various functional domains and motifs. The three LXXLL motifs, where L is leucine (highlighted with small red dots in the sequences) and X any amino acid, are located in the middle of the protein and here colored in green (L1 motif), blue (L2 motif) and orange (L3 motif). (B) The initial hVDR complex used for MD simulations is based on 1DB1 structure complexed with calcitriol. (C) The three motifs (L1, L2 and L3) were added to the VDR-calcitriol complex yielding to three distinct complexes, the trimeric complex of VDR, calcitriol and coactivator interaction motif: hVDR-SRC-1/L1-calcitriol, hVDR-SRC-1/L2-calcitriol and hVDR-SRC-1/L3-calcitriol.**

Å × 69 Å × 86 Å orthorhombic simulation box with TIP3P molecules, retaining crystallographic water. The system was neutralized with Na^+ ions and energy-minimized using the default protocol for 100 ps. Subsequently, 10 ns MD simulations were performed for three hVDR complexes at 300 K and 100 kPa using the NPT ensemble. Temperature and pressure were controlled using the Nose-Hoover thermostat and the Martyna-Tobias-Klein barostat, respectively, with a 2 fs integration time step. After simulation, every 10th frame was extracted from the aligned trajectory, resulting in 101 frames, excluding water molecules and Na^+ ions. The 10 ns simulation duration was chosen based on extensive previous studies of NR LBDs, which demonstrated that this timescale allows for sufficient sampling [30], [31], [32]. The validity of this duration is further supported by (1) RMSD convergence reaching stable plateaus, (2) observation of distinct dynamic behaviors between complexes differing only in CoA peptide sequence, and (3) capture of biologically relevant conformational changes documented in the literature for NRs.

B. THE GEODES DESCRIPTORS

The GEODES toolbox was developed as a Python3 package that computes 14 types of 3D protein descriptors categorized into three groups: auxiliary, coordinate-derived, and DSSP-derived (Table S1). Two auxiliary descriptors, i) the center of mass (COM) of a protein (COM*prot*) and ii) of the helices (COM_{H_i}) are used to calculate the main descriptors but are excluded from further analyses.

Coordinate-derived descriptors include length of helix H_i (ΔNC_{H_i}), angles between two helices H_i and H_j (θ_{H_i,H_j}), distances between their COMs ($\Delta COM_{H_i,H_j}$), distances between the COM*prot* and helix H_i ($\Delta COM_{prot,H_i}$), and angles between COM*prot* and helices' N- and C-termini Cα-atoms ($\theta_{NC_{H_i},COM_{prot}}$). Helix N- and C-termini were determined using Kpax software [33], based on 45 refined hVDR structures from PDB (as of February 2021). This combined approach improves accuracy in several ways. Kpax assignments provide a robust baseline for consistent secondary structure definition across all frames. This is particularly important for coordinate-derived descriptors that require stable reference points. Meanwhile, DSSP's frame-by-frame analysis captures genuine structural transitions that might be missed by using a static assignment alone. The complementary nature of these methods - Kpax's stability and DSSP's sensitivity to local changes - ensures that our descriptors can both reliably track consistent structural elements and detect meaningful conformational changes. This is evidenced by our ability to detect both stable core regions and dynamic elements in the VDR structure, as demonstrated by the correlation between our geometric descriptors and observed protein dynamics. Structural alignment with Kpax yielded a consensus secondary structure model of twelve long and two short helices (Table S2). Additional NR-specific "charge clamp" descriptors include: 1) $\theta res_{i,j,k}$: angle between "charge clamp" residues i, j and a lysine k from helix H4; 2) $\Delta res_{i,j}$: pairwise distances between Cα-atoms in this triad; 3) $\Delta_{COM_{prot},res_i}$: distances from these C$\alpha$-atoms to COM_{prot}. For hVDR, the "charge clamp" residues are K246 and E420, with K264 serving as an additional triad residue (as listed in Table S1). The "charge clamp" is crucial for coactivator binding, while K264 stabilizes H12, indirectly affecting coactivator interaction [34], [35]. These NR-specific descriptors can be easily customized or excluded for other proteins.

The third group of GEODES descriptors is based on the Dictionary of Secondary Structure of Proteins (DSSP) algorithm [36]. However, DSSP assignments show frame-dependent variability within trajectories (Table S2). To address this, a combined approach was implemented such as i) establishing a consensus structure for consistent descriptor calculation; ii) usage of Kpax assignments for coordinate-derived descriptors and iii) formation of separate, flexible descriptor group for DSSP assignments to capture frame-to-frame variability. This approach balances consistency with the ability to detect important differences in helix assignments. For DSSP helical assignments, both "H" for α- and "G" for 3^{10}-helices were considered.

To highlight discrepancies between Kpax and DSSP secondary structure assignments, three specific descriptors were introduced: 1) $DSSPstart_{H_i}$ and 2) $DSSPend_{H_i}$ reflect differences in helix termini assignments. Specifically, for the N-terminus, we calculate the DSSP assignment minus Kpax assignment, and for the C-terminus, the Kpax assignment

minus DSSP assignment. Therefore, a positive value (+) indicates additional length in the Kpax assignment, while a negative value (-) indicates a shorter Kpax assignment. The 3) descriptor N_{extra} points to total residues considered as helical by DSSP but not by Kpax.

Additional DSSP-based descriptors include the solvent accessibility area per helix (ACC_{H_i}) and the secondary structure elements (SSE) content.

All distances are computed as Euclidean distances using 3D coordinates of either Cα-atoms or COM, depending on the descriptor.

The θ_{H_i,H_j} are computed based on formula for calculation of angle between two vectors, where each helix was represented as a vector using the terminal Cα-atoms:

$$\theta = \arccos\left(\frac{\vec{H}_i \cdot \vec{H}_j}{|\vec{H}_i||\vec{H}_j|}\right) \tag{3}$$

where $\vec{H}_i, \vec{H}_j$ are the two vectors representing helices H_i and H_j, and $|H_i|$, $|H_j|$ are the magnitudes of these two vectors.

Each type of descriptor in Table S1 results in many features depending on the number of helices.

C. THE DATA ANALYSIS PIPELINE

RMSD and RMSF data for 101 frames were extracted from Desmond [37] as raw values for all Cα-atoms relative to the initial frame. 2D RMSD-arrays, which are pairwise RMSD-matrices $M(d_{i,j})$, where $d_{i,j}$ is the RMSD distance between frames i and j, were calculated for α-helices only using the Schrödinger script *rmsd.py*. Helical regions from the Kpax consensus secondary structure were used for alignment. RMSF data were computed for helical regions, with residues grouped per helix and averaged.

For hVDR with 14 helices, the GEODES toolbox generated 284 geometrical features, computed for each frame and trajectory separately. Each trajectory was represented as a (101 $\times$ 284) matrix organized with rows as frames and columns as features. Features with constant values across all frames were removed, resulting in 281, 280, and 279 features for VDR-CoA complexes containing SRC-1/L1, SRC-1/L2, and SRC-1/L3 motifs respectively (referred to as L1, L2, and L3 complexes). Data processing included standardization of feature values by removing the mean and scaling to unit variance.

$$z_i = \frac{(x_i - \mu)}{\sigma}, \tag{4}$$

where z_i is a standardized value of the feature for frame i, x_i, the original value computed for frame i. μ, the mean $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ and σ, the standard deviation $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$ of the values taken by this feature over the total number of frames N.

Hierarchical agglomerative clustering (HAC) was applied to both RMSD and GEODES data, a classical approach for trajectory analyses [37], [38]. Using scikit-learn on the full-scale dataset, HAC parameters were optimized for silhouette score. For RMSD data, complete linkage was optimal for L1 and L2 complexes (scores 0.186 and 0.134), while average linkage was best for L3 complex (score 0.136), using precomputed distance as the metric. For GEODES data, complete linkage was optimal for L1 and L3 complexes (scores 0.170 and 0.081), and Ward linkage for L2 complex (score 0.103), using Euclidean distance as the metric. For both datasets, the optimal number of clusters was two for each complex L1, L2 or L3. HAC results were visualized using scikit-learn Principal Component Analysis (PCA) with two principal components. The effectiveness of different linkage methods relates to the underlying characteristics of the conformational ensemble subjected to clustering. Complete linkage, which performed best for L1 and L2 RMSD data, is particularly effective when clusters are relatively compact and well-separated, suggesting these complexes exhibit distinct conformational states. The superiority of average linkage for L3 RMSD data indicates more gradual transitions between conformational states. For GEODES data, Ward linkage's effectiveness with the L2 complex aligns with its more stable behavior, as this method excels at minimizing within-cluster variance. To compare RMSD and GEODES approaches, the following methods were applied. First, the Rand index between HAC results was computed to assess cluster assignment similarity. Then, a comparison of 1D-arrays was made reflecting differences between each subsequent frame and the initial frame. The RMSD 1D-array was taken from the first row of the RMSD 2D-array, while the GEODES 1D-array included the n-dimensional Euclidean distances from each frame to the initial one. Pearson correlation was then computed between the two arrays. Finally, the Mantel test, with 10,000 permutations and Pearson correlation, was used to assess the correlation between the 2D RMSD- and GEODES arrays.

RMSF values were compared to the GEODES approach as follows: features were sorted by their quartile coefficient of dispersion (QCD):

$$QCD = \frac{Q3 - Q1}{Q3 + Q1}, \tag{5}$$

where Q1 and Q3 are the upper thresholds for the first (25% lowest values) and third (75% lowest values) quartiles, respectively. The QCD distribution for all features was visualized with a violin plot for each trajectory (Figure S2). Features with QCD above the third quartile were classified as flexible, while those below the first quartile were deemed stable. Unique descriptors with differential stability were then identified for each trajectory, and the resulting list was compared to RMSF data grouped by helices.

III. RESULTS

A. GEODES APPROACH AS A TOOL FOR MD TRAJECTORY ANALYSIS

MD simulations of VDR-SRC-1 complexes were analyzed using RMSD for global stability and RMSF for local flexibility throughout the simulation. RMSD data (Figure 2A)

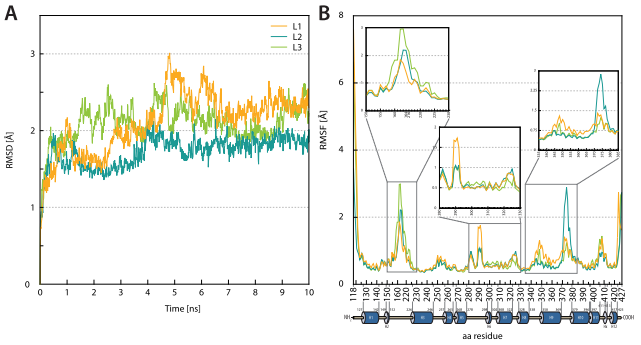

FIGURE 2. Molecular dynamics trajectory parameters. (A) RMSD of three trajectories built upon Desmond raw data with respect to the initial frame. (B) RMSF of three trajectories built upon Desmond raw data compared to the initial frame. The secondary structure of hVDR is depicted below with helices represented as cylinders. The most flexible regions are highlighted in the expanded boxes.

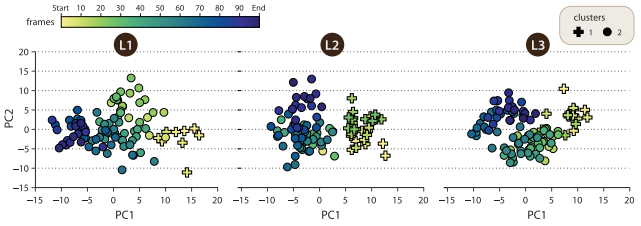

FIGURE 3. Visualization of frames from L1, L2 and L3 MD trajectories by 2D-PCA of GEODES data. Plotted frames are colored based on their timestamps, ranging from light yellow at the beginning of the simulation to dark blue at the end. Cluster assignments are indicated by the shape of the points: crosses for cluster 1 and circles for cluster 2.

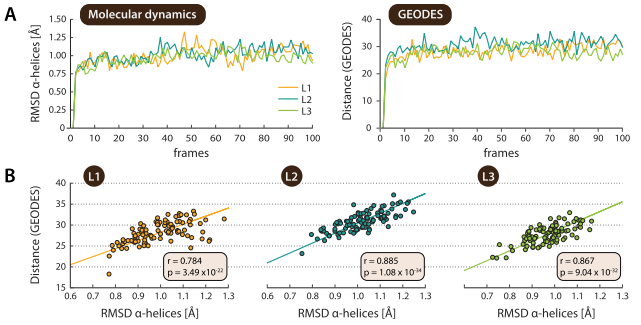

FIGURE 4. Comparison of RMSD and GEODES 1D-arrays. (A) Curves computed from the distance between the given frame and the initial one for RMSD (left) and GEODES representation (right). (B): Scatter plots representing the dependency between the GEODES and α-helical Cα RMSD curves for L1, L2 and L3 complexes (from left to right.

shows L2 complex as the most stable, reaching a 1.7Å plateau after a mid-simulation jump. L1 and L3 complexes exhibit less stability, fluctuating around 2.4Å towards simulation end. L1 appears to reach a local energy minimum, while L3 passes through a 2Å local minimum before increasing to 2.5Å by simulation end. RMSF analysis (Figure 2B) shows low fluctuation levels in helical regions and higher peaks in loops across all trajectories, as expected. However, some loops show trajectory-specific variations such as the L3 complex, which exhibits the highest fluctuations in the 156-164/216-224 region, near the flexible VDR-specific insertion domain (165-215) absent in the crystal structure [34]. The L1 complex shows the strongest fluctuation in the loop between H5 and H6 (288-293), while the L2 complex peaks in the loop H9-H10 (371-380). Despite these differences, the trajectories generally display similar behavior in both structured and unstructured regions. While this analysis effectively shows global complex stability and local fluctuations in unstructured regions, it does not provide information on geometric variations (angles, distances) involving distinct locally structured regions considered stable by RMSF. Such additional data could be valuable for analyzing local conformational changes over time, including allosteric effects upon ligand-binding or CoA-recruitment [39].

We therefore compared GEODES versus classical RMSD data for MD trajectory analysis. GEODES descriptors derive primarily from the length and relative positions of hVDR's 14 α-helices. GEODES method was applied to each hVDR frame of each MD trajectory and subsequently HAC was applied to both GEODES and RMSD data. Both data types yielded for each trajectory two clusters under optimal hyperparameters, with Rand indices of 0.980 (L1), 0.923 (L2), and 0.961 (L3) between cluster assignments. This high similarity indicates that GEODES captures patterns similar to the RMSD data. Figure 3 shows 2D-PCA visualization of 101 frames from 10 ns MD simulations of L1, L2, and L3 complexes, possible only with GEODES multi-dimensional vector representation. The cluster analysis shows a clear pattern: Cluster 1 contains primarily the early frames of the trajectory (represented by lighter colors), while Cluster 2 predominantly consists of later frames (represented by darker colors). Detailed dendrogram analysis shows some differences between GEODES and RMSD data (Figure S1) for instance frame 6 of L2 trajectory clustered with mid-trajectory frames 46, 51, and 53 in GEODES but not RMSD data. This analysis provides a starting point for using GEODES to complement RMSD data.

1D- and 2D-arrays representing RMSD and GEODES data were compared using Pearson correlation coefficient and Mantel test, respectively. We first examined the correlation between the 1D-array of α-helical RMSD (Å) between all trajectory frames and the initial one, and the GEODES-array of Euclidean distances between the same frames in multi-dimensional space (Figure 4A). Pearson correlation coefficients showed strongly positive linear correlations between α-helical RMSD and GEODES representation (Figure 4B), with L2 complex showing the highest (r=0.885) and L1 complex the lowest (r=0.784) correlation. We then compared 2D-arrays of pairwise distances between frames, using either α-helical Cα RMSD or Euclidean distance between GEODES values (Figure 5A-B). The Mantel test showed lower correlations than 1D-arrays, as expected due to the increased number of data points and noise. However, values remained above average, with L2 complex showing the highest (r=0.749) and L3 complex the lowest (r=0.659)

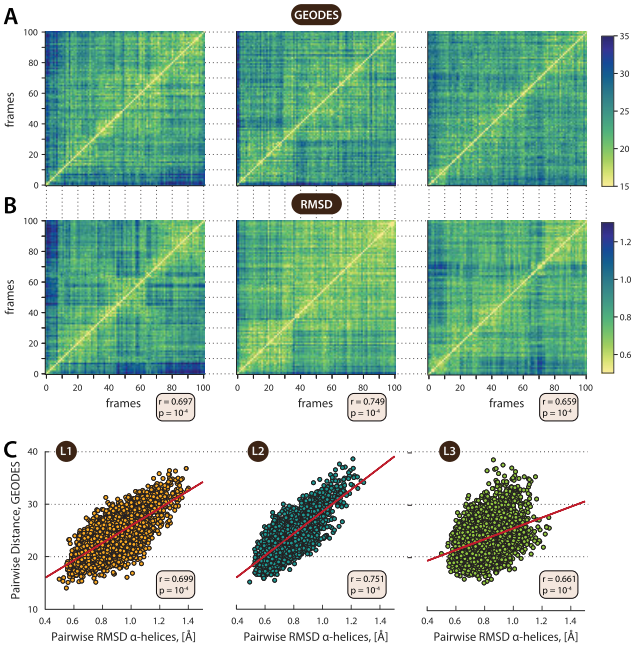

FIGURE 5. Comparison of GEODES and RMSD 2D-array representations for L1, L2 and L3 complexes. (A) GEODES 2D-array and (B) RMSD 2D-array (Cα-helices only). Mantel test correlation coefficients for L1, L2 and L3 are 0.697, 0.749 and 0.659, respectively (p-value = 10^{-4}). (C) Correlation plot between GEODES and RMSD matrices where each point represents a pair of frames with their corresponding GEODES and RMSD values.

Mantel correlation coefficient. To visualize Mantel test results, we flattened the upper triangular portions of both 2D-arrays and displayed their relationship as a scatter plot (Figure 5C). The Pearson correlation coefficients for these flattened representations matched the Mantel test values, confirming moderate positive correlation between GEODES and RMSD 2D-arrays. The L2 complex, with the most stable MD trajectory (Figure 2), shows the highest correlation between RMSD and GEODES approaches in both 1D- and 2D-array analyses. Conversely, the less stable L1 and L3 complex trajectories exhibit lower correlations between RMSD and GEODES. This suggests that the two methods can capture different aspects of conformers, particularly when conformers are numerous and diverse throughout the trajectory.

B. GEODES APPROACH AS A TOOL FOR DETECTING LOCAL FLUCTUATIONS

This section illustrates how the GEODES approach complements RMSD and RMSF metrics for studying local structural changes over time. RMSF analysis, restricted to α-helices like GEODES data, reveals that helix H3n shows maximum fluctuation across all trajectories, followed by helices Hx, H2, and H10 (Figure 6). The high flexibility of helices H2, H3n, and Hx is due to their short length (up to four residues). The high fluctuation of helix H10 is the result of the absence of its stabilizing partner, the retinoid X receptor. Helices H3-H5 and H8 maintain the lowest fluctuation levels for all the trajectories. Specific trends include high fluctuations

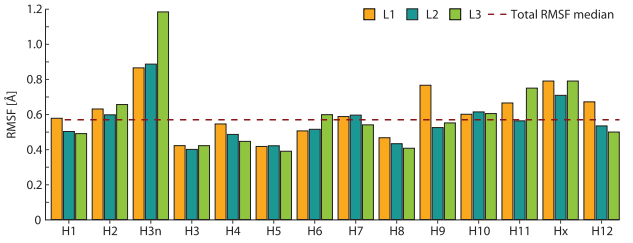

FIGURE 6. Comparison of hVDR per-helix RMSF for L1 (yellow), L2 (blue) and L3 (green) complexes.

in H9 and H12 for L1 complex, H3n for L3 complex, and H11 for both L1 and L3 complexes. L2 complex shows no extreme local fluctuations beyond those shared with L1 and L3, aligning with its overall stability according to RMSD data.

To compare GEODES with RMSF data, we assessed GEODES data variability excluding extreme values and outliers. We computed QCD for each feature using raw (non-standardized) values, sorted them in descending order, and evaluated QCD distribution (Figure S2). Features with QCD above the third quartile were labeled 'flexible', while those below the first quartile were deemed 'stable'. The QCD distributions were similar across trajectories, yielding 71 flexible and 71 stable features each, though not identical between trajectories. Combining these subsets resulted in 91 flexible and 96 stable features total. Figure 7 shows specific (A,B) and shared (C,D) features across trajectories (for full list see Tables S3-S6). Over half the features are shared: 55/91 flexible and 49/96 stable.

Shared flexible features (Figure 7C, Table S5) primarily represent angles between helices θ_{H_i,H_j} (34 features), solvent accessibility of helices ACC_{H_i} (10 features), and angles formed by helix ends with protein COM $\theta_{NC_{H_i},COM_{prot}}$ (4 features). Shared stable features (Figure 7D, Table S6) are primarily DSSP secondary structure assignments (26 features) and distances between helix COMs ($\Delta COM_{H_i,H_j}$: 10 features). The latter also appears among flexible features but for different helices. Angles between helices (θ_{H_i,H_j}: 6 features) are the third most common stable feature, also present in flexible features for different helix pairs.

Trajectory-specific flexible features analysis (Figure 7A, Table S3) reveals L1 complex-specific features mostly related to helix H12, consistent with RMSF analysis (Figure 6). L1's flexible ACC_{H7} suggests ligand-binding pocket (LBP) movements undetected by RMSF. For L2, the flexible $\Delta_{COM_{prot},res_{E420}}$ involves a "charge clamp" residue, while flexible $\Delta COM_{prot,H5}$ contrasts with H5's RMSF stability.

For L2 complex, the distance $\Delta_{COM_{prot},res_{E420}}$, involving one of the two "charge clamp" residues, reveals variability not detectable by RMSF analysis. Furthermore, GEODES shows $\Delta COM_{prot,H5}$ as a flexible feature in L2, despite RMSF indicating H5 as one of the most stable helices across all complexes, along with H3 and H8.

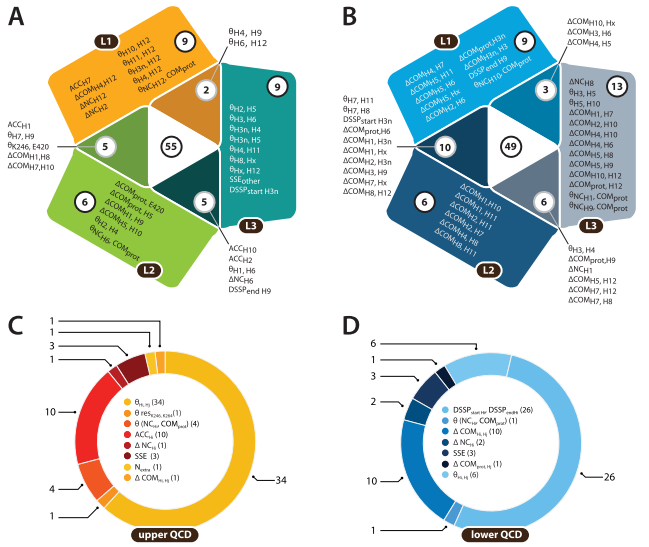

FIGURE 7. Trajectory-specific (upper panels) and shared (lower panels) GEODES features. Trajectory-specific panel shows the features that are unique for each trajectory for flexible (A) and stable (B) feature subsets. Shared features panel displays the feature types and their counts in the intersection across all the trajectories: 55 features for flexible features subset (C) and 49 features for stable features subset (D).

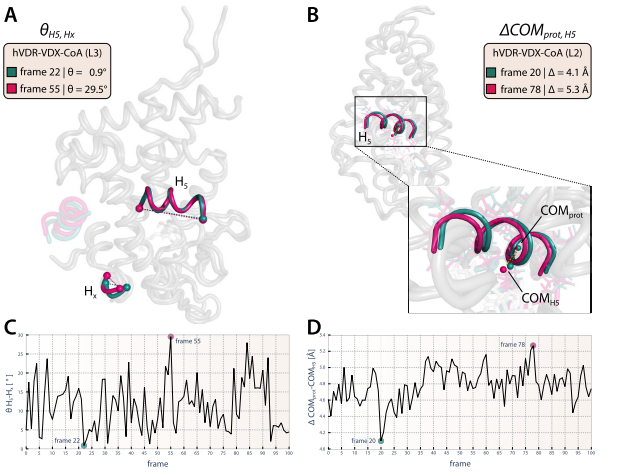

FIGURE 8. Visualization of the selected descriptors within the L3 complex and L2 complexes. (A) Angle $\theta_{H5,Hx}$ observed in the frames 22 and 55 of L3 trajectory. (B) Distance $\Delta COM_{prot,H_5}$ observed in the frames 20 and 78 of L2 trajectory. Note that location of COM_{prot} coincides with the location of H5. (C) Angle $\theta_{H5,Hx}$ behavior across the trajectory (markers show the values of the descriptor in the selected frames). (D) Distance $\Delta COM_{prot,H_5}$ behavior across the trajectory (markers show the values of the descriptor in the selected frames).

For the L3 complex, three features relate to helix H3n, aligning with RMSF data in Figure 6. Trajectory-specific stable features (Figure 7B, Table S4) predominantly involve distances between helices $\Delta COM_{H_i,H_j}$, also prevalent in shared stable features. These align with RMSF data, often involving stable helices H2, H4, H5, and H8. Helices H2, H3n, H10 and Hx, flexible in RMSF analysis across all complexes, rarely appear in stable GEODES trajectory-specific features. When present, they are typically paired with more stable helices e.g., H5-Hx for L1, H4-10 for L3. Notably, some features involving only flexible helices emerge, such as $\Delta COM_{prot,H_{3n}}$ in L1, $DSSP_{start H_{3n}}$ in the L1/L2 intersection, and $\Delta COM_{H_2,H_{10}}$ in L3 complex.

All these findings demonstrate GEODES' ability to provide additional insights beyond RMSF analysis, revealing complex-specific behaviors and movements of particular structural elements.

C. GEODES ALLOWS FOR CLEAR VISUALIZATION OF THE PROTEIN FEATURES AND THEIR DYNAMICS

Geometric features such as distances and angles are easily visualized in PyMOL [40], a popular structural bioinformatics software with extensive scripting capabilities. Figure 8 illustrates two flexible features: angle $\theta_{H5,Hx}$ (panels A and C) and distance $\Delta COM_{prot,H_5}$ (panels B and D). Angle $\theta_{H5,Hx}$ is the most flexible feature shared across all trajectories (Figure S3, Table S5). It changes drastically during simulation: from 0.9° in frame 22 to 29.5° in frame 55 for L3 complex (Figure 8C). Figure 8A visualization (with coordinate lists in Table S7) shows that this movement primarily results from helix Hx fluctuations, which exceed 1Å across several coordinates. Meanwhile, helix H5's position remains nearly constant, with less than 0.2Å movement across all coordinates, which is consistent with the RMSF analysis (Figure 6). Conversely, visualization of distance $\Delta COM_{prot,H_5}$ (Figure 8B, Table S8), a flexible trajectory-specific feature for L2 complex, shows flexibility primarily due to COM_{H5} movement, despite H5's stability in RMSF analysis (Figure 6). The most notable change occurs between frames 20 and 78 (Figure 8D), where COM_{H5} shifts by ∼0.4-0.9Å across all coordinates. The protein's COM shifts only 0.11Å in x-coordinate and less in others, suggesting H5 drives the 28.4% increase in $\Delta COM_{prot,H_5}$ (Figure S8).

In summary, GEODES descriptor analysis is capable of tracking internal 3D changes in protein structure, focusing on secondary structure elements, primarily α-helices. This enables observation of fluctuation variations in specific regions by monitoring particular descriptors during MD simulation.

IV. DISCUSSION

MD is widely used to explore biomolecular complex structural changes. With the advances in structural prediction methods such as AlphaFold3 [41], which now predicts nucleic acid and ion positions alongside proteins, understanding complex structures is less challenging than exploring their dynamic behavior. As wet laboratory approaches for studying structural dynamics are mostly limited to hydrogen-deuterium exchange coupled NMR or mass spectrometry, improving *in silico* techniques is valuable for computational studies of structural fluctuations.

In this work, we propose a methodology based on structural descriptors to complement established MD trajectory analysis procedures using RMSD and RMSF evaluation.

A. GEODES: EXTENDING CLASSICAL RMSD ANALYSIS

The limitations of the RMSD and RMSF parameters have been discussed in literature, including lack of temporal local dynamic event information, degeneracy issues, heavy dependence on precise structure superimposition, and predominant influence of most deviated regions [8], [9], [42]. Several alternatives to RMSD have been developed to address these limitations, such as weighted RMSD, which uses only a subset of atoms for superposition, downplaying inherently unstructured regions [42], and Generally Applicable Replacement for RMSD (GARD) [8] normalized to unity, which prevents skewed aggregation statistics and incorporates ''chemical awareness'' using Andrews weights [43]. While GARD prevents the rejection of correct docking poses with large RMSD due to minor misplacements, it has limitations in providing insight into precise locations of structural changes over time. To compare GEODES in the context of existing methodologies, we summarize the strengths and weaknesses of various approaches (Table 1), highlighting both GEODES' unique features and its complementary role in MD trajectory analysis.

While RMSD and RMSF remain valuable tools for MD trajectory analysis, GEODES complements these classical approaches by offering several unique capabilities. The method's position-independence eliminates potential biases introduced during structure alignment while reducing computational overhead. Our analysis demonstrated GEODES' ability to detect complex-specific behaviors not captured by conventional methods alone. For example, in the L2 complex, GEODES revealed flexibility in the distance $\Delta_{\mathrm{COM_{prot}},\mathrm{res}_{E420}}$ involving a ''charge clamp'' residue, and variations in $\Delta\mathrm{COM_{prot}}$, $\mathrm{H_5}$, despite H5's apparent stability in RMSF analysis. Moreover, the method provides detailed temporal resolution of local structural changes, tracking specific geometric variations frame-by-frame, as demonstrated in our analysis of the angle $\theta_{\mathrm{H_5,Hx}}$, which showed significant variations from 0.9° to 29.5° during the simulation. The strong correlation between GEODES and RMSD clustering results (Rand indices of 0.980, 0.923, and 0.961 for L1, L2, and L3 complexes respectively) validates GEODES while highlighting its ability to capture additional conformational information. This is particularly valuable for understanding allosteric mechanisms and protein-protein interactions, where local geometric changes may have significant functional implications. The combination of GEODES with traditional analysis methods provides a more comprehensive understanding of protein dynamics, with RMSD offering global conformational insights while GEODES captures specific local conformational dynamics during MD simulations.

Despite this overall agreement, the subtle differences between RMSD and GEODES clustering reveal interesting insights into protein dynamics. For example, in the L2 trajectory, frame 6 clustered with mid-trajectory frames (46, 51, and 53) in GEODES analysis but not in RMSD clustering (Figure S1). This difference suggests that while these conformations may show different global RMSD values, they share specific geometric features captured by GEODES descriptors. The clustering divergences are particularly informative for understanding protein dynamics, as frames that GEODES groups together often share specific geometric relationships between secondary structure elements, even when their global RMSD values differ. This is evident in the dendrogram analysis (Figure S1), where branching patterns reveal how GEODES can identify structurally similar states that may be separated in RMSD-based clustering due to differences in global conformation. These clustering differences highlight GEODES' sensitivity to local geometric features while demonstrating how global and local measures of structural similarity can provide complementary perspectives on conformational dynamics.

In Critical Assessment of protein Structure Prediction (CASP) the Global distance test (GDT) and the Longest continuous segment (LCS) [44] have been implemented to overcome superposition dependency. This approach identifies for each model residue the largest continuous (LCS) or arbitrary (GDT) set that aligns with the reference structure within a given RMSD (LCS) or distance (GDT) cutoff. GDT is particularly effective at recognizing structural similarity and is robust against small fragment movements. While these methods address some RMSD limitations, they still have constraints in analyzing precise structural changes over time in MD trajectories.

B. DESCRIPTOR-BASED APPROACHES

Descriptor-based approaches, long used in structural bioinformatics, offer advantages such as superposition independence and compatibility with Euclidean distance, common in machine learning and data analysis. These descriptor-based techniques are frequently used for protein structure classification, as seen in SCOP [16] and CATH [17]. Lindström et al. [11] characterized protein chains using alignment-independent descriptors based on Cα atom Euclidean distances, protein backbone ψ and ϕ angles, and amino acid physico-chemical properties. In this study, they analyzed descriptors via PCA and multivariate methods. Wang et al. [45] proposed local average distance for flexible protein structure comparison, based on geodesic or Euclidean distances, used to discover unknown conformational relationships and reorganize protein structure classification.

Other applications include binding site [21] and folding rate predictions [46]. Jiang et al. [21] proposed a multichannel molecular descriptor combining geometry- and energy-based approaches, suitable for convolutional neural network training due to grid voxel utilization. Gao et al. [46] developed linear regression-based models for predicting folding rates of proteins with two-state, multistate, and unknown folding kinetics, using protein sequence, predicted secondary structure, solvent accessible surface, b-factors, and local structure entropy as inputs. These models also revealed potential

TABLE 1. Comparison of GEODES to Complementary Methods.

Method	Strengths	Weaknesses	Comparison with GEODES
RMSD	Measures global conformational changes; simple to compute.	Masks local dynamics; alignment-dependent.	GEODES captures local changes and avoids superposition issues.
RMSF	Identifies highly flexible regions over time.	No temporal resolution; limited to global flexibility.	GEODES provides temporal and spatial resolution for local dynamics.
GARD	Accounts for chemical awareness in MD trajectories.	Limited to docking scenarios; lacks geometric descriptors.	GEODES integrates angles, distances, and secondary structure-based descriptors.
GDT	Identifies structural similarity robustly against small fragment movements.	Primarily used for structure prediction; less suited for MD trajectories.	GEODES offers a time-resolved geometric analysis suitable for MD studies.
tICA	Identifies slowest motions in MD.	Requires dimensionality reduction; lacks structural interpretability.	GEODES offers structural insights alongside trajectory-based clustering.
DCCM	Captures correlated motions between regions.	Limited to correlations; doesn't quantify geometric or temporal features.	GEODES quantifies geometric properties, complementing correlated motion data.
LSE	Provides entropy-based flexibility insights.	Limited application scope; no integration of secondary structure data.	GEODES integrates entropy-like variability with structural elements like helices.
b-factors	Widely used for assessing flexibility in crystal structures.	Static; doesn't account for dynamic MD behavior.	GEODES focuses on dynamic flexibility, complementing static flexibility metrics.

relations between topological and structural properties and folding rates. To our knowledge, descriptor-based methods have not been used for MD trajectory analysis, and few of these methods derive descriptors from geometric properties attached to secondary structure elements.

C. INTEGRATION OF KPAX AND DSSP FOR DESCRIPTOR COMPUTATION AND MODELING

The integration of Kpax and DSSP algorithms in GEODES represents a novel approach to balancing consistency and sensitivity in protein structure analysis. Our implementation addresses a fundamental challenge in MD trajectory analysis: distinguishing genuine conformational changes from assignment fluctuations. The Kpax-derived consensus structure, based on 45 refined hVDR structures, provides stable reference points for geometric measurements, while DSSP captures frame-specific variations in secondary structure assignments. This dual approach proved particularly valuable in our VDR analysis. For instance, when examining the differences between Kpax and DSSP assignments (Table S2), we found that this combined method effectively tracked both stable structural elements and dynamic regions. The three specific descriptors we introduced ($DSSPstart_{H_i}$, $DSSPend_{H_i}$, and N_{extra}) provide quantitative measures of assignment differences, offering insights into regions where structural classifications may be ambiguous or transitional. For protein modeling applications, this integration has several important implications: i) it provides a more robust framework for analyzing conformational changes in MD trajectories; ii) it enables distinction between stable structural elements and genuinely flexible regions and iii) it helps identify regions where secondary structure assignments may be sensitive to small conformational changes. The stability of geometric descriptors calculated using Kpax-based assignments, combined with the sensitivity of DSSP-based features, suggests that this integrated approach could be particularly valuable for studying proteins with dynamic structural elements or those undergoing conformational transitions.

D. APPLICATIONS OF GEODES IN DRUG DESIGN

The ability of GEODES to identify and characterize protein flexibility has direct implications for drug design, in this study applied to VDR. Our analysis reveals three key features:

First, GEODES effectively identified flexible regions in the LBP. For example, the flexible ACC_{H7} descriptor in the L1 complex suggests movements in the LBP that were not detected by RMSF analysis. Such local flexibility information is crucial for understanding potential binding site adaptability during drug-target interactions.

Second, our analysis revealed complex-specific behaviors in the 'charge clamp' region, as demonstrated by the flexible $\Delta_{COM_{prot},res_{E420}}$ feature in the L2 complex. These charge clamp residues are critical for CoA binding, and their dynamic behavior has implications for designing drugs that could potentially modulate protein-protein interactions.

Third, GEODES' ability to track specific geometric relationships provides valuable information for rational drug design. The method captures subtle changes in distances and angles between structural elements that could influence ligand-binding or allosteric regulation. For instance, the observed variations in angle $\theta_{H5,Hx}$ demonstrate conformational flexibility that might be important for drug design. These geometric descriptors can enhance structure-based drug design workflows by identifying flexible regions

that could accommodate ligand-binding, characterizing the dynamic behavior of key functional residues, providing quantitative measures of LBP adaptability and/or supporting the design of allosteric modulators by mapping dynamic communication between structural elements.

E. ANALYSIS AND INTERPRETATION OF VDR-COACTIVATOR COMPLEX DYNAMICS

The distinct patterns observed in our analyses can be interpreted in the context of VDR's biological function and previous experimental findings. The enhanced stability of the L2 complex, evidenced by its 1.7Å RMSD plateau, aligns with experimental studies demonstrating VDR's highest binding affinity for the L2 motif of SRC-1 [47]. While L1 and L3 complexes show greater conformational sampling (fluctuating around 2.4Å), the relationship between flexibility and binding affinity appears complex, as experimental data indicates L3 typically shows stronger binding than L1 [47]. GEODES analysis revealed complex-specific variations that provide deeper insights into functional dynamics. The high fluctuations observed in L3 complex region 156-164/216-224, coinciding with the flexible VDR-specific insertion domain, suggest that these movements may be functionally relevant for VDR's biological role. Similarly, the L1 complex's elevated flexibility in the H5-H6 loop (288-293) and L2 complex's peaks in the H9-H10 loop (371-380) likely reflect different modes of accommodation for the various CoA peptide sequences. These patterns are particularly meaningful given VDR's function as a transcriptional regulator. The observed flexibility in specific regions may facilitate the conformational changes necessary for proper CoA recruitment and subsequent transcriptional activation. The preferential binding of L2 motif and its role in complex stabilization suggests that different CoA peptide sequences can fine-tune VDR's dynamic behavior, potentially influencing its biological function through altered protein dynamics.

V. CONCLUSION

GEODES introduces a novel, descriptor-based approach to MD trajectory analysis that addresses limitations of traditional metrics like RMSD and RMSF. By combining ideas from RMSD extension and descriptor-based techniques, our method offers a new application for protein structural descriptors in unsupervised learning, specifically for MD trajectory clustering. The position-invariant descriptors eliminate the time-consuming protein superposition step required for RMSD calculation, placing GEODES among alignment-independent methods of protein structure analysis.

Beyond clustering trajectories, GEODES enables precise examination of specific structural changes at given timestamps, offering both computational efficiency and detailed insights into structural dynamics. The methodology has been validated on α-helical structures like VDR, with promising results from Kpax/DSSP integration approaches.

While currently optimized for α-proteins, GEODES presents several opportunities for future development. The framework could be extended to include β- and mixed α/β structures with additional geometric descriptors to accommodate diverse protein classes. Future studies could also explore GEODES' application across various protein families and alternative clustering methods. These potential expansions reflect the adaptability of the framework and its capacity for broader applications in drug discovery, structural biology, and bioinformatics.

COMPETING INTERESTS

No competing interest is declared.

AUTHOR CONTRIBUTIONS STATEMENT

Karina Pats conducted the study, produced and tested the final Python scripts, analyzed the results, wrote and edited the manuscript; Igor Glukhov, Stepan Petrosian, and Maria Mamaeva did the initial GEODES descriptor calculations, Alexey Sergushichev edited and reviewed the manuscript; Marie-Dominique Devignes and Ferdinand Molnár conceptualized and supervised the work, edited and reviewed the manuscript.

SUPPLEMENTARY DATA

Supplementary data is available at **IEEE** *Access* online.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable suggestions. They thank EComputing facilities included the Grid'5000 testbed (https://www.grid5000.fr).

AVAILABILITY AND IMPLEMENTATION

The GEODES Python3 script toolbox is available at https://github.com/rinnifox/GEODES

REFERENCES

[1] J. Gelpi, A. Hospital, R. Goñi, and M. Orozco, "Molecular dynamics simulations: Advances and applications," *Adv. Appl. Bioinf. Chem.*, vol. 8, pp. 37–47, Nov. 2015.

[2] J. Henriques, C. Cragnell, and M. Skepö, "Molecular dynamics simulations of intrinsically disordered proteins: Force field evaluation and comparison with experiment," *J. Chem. Theory Comput.*, vol. 11, no. 7, pp. 3420–3431, Jul. 2015.

[3] A. W. Götz, M. J. Williamson, D. Xu, D. Poole, S. Le Grand, and R. C. Walker, "Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. Generalized born," *J. Chem. Theory Comput.*, vol. 8, no. 5, pp. 1542–1555, May 2012.

[4] D. E. Shaw et al., "Millisecond-scale molecular dynamics simulations on Anton," in *Proc. Conf. High Perform. Comput. Netw., Storage Anal.*, Nov. 2009, pp. 1–11.

[5] S. Fan, M. Linke, I. Paraskevakos, R. Gowers, M. Gecht, and O. Beckstein, "PMDA—Parallel molecular dynamics analysis," in *Proc. Python Sci. Conf.*, C. Calloway, D. Lippa, D. Niederhut, and D. Shupe, Eds., 2019, pp. 134–142.

[6] A. Jurcik, D. Bednar, J. Byska, S. M. Marques, K. Furmanova, L. Daniel, P. Kokkonen, J. Brezovsky, O. Strnad, J. Stourac, A. Pavelka, M. Manak, J. Damborsky, and B. Kozlikova, "CAVER analyst 2.0: Analysis and visualization of channels and tunnels in protein structures and molecular dynamics trajectories," *Bioinformatics*, vol. 34, no. 20, pp. 3586–3588, Oct. 2018.

[7] K. Sargsyan, C. Grauffel, and C. Lim, "How molecular size impacts RMSD applications in molecular dynamics simulations," *J. Chem. Theory Comput.*, vol. 13, no. 4, pp. 1518–1524, Apr. 2017.

[8] J. C. Baber, D. C. Thompson, J. B. Cross, and C. Humblet, "GARD: A generally applicable replacement for RMSD," *J. Chem. Inf. Model.*, vol. 49, no. 8, pp. 1889–1900, Aug. 2009.

[9] A. Grossfield, P. N. Patrone, D. R. Roe, A. J. Schultz, D. Siderius, and D. M. Zuckerman, "Best practices for quantification of uncertainty and sampling quality in molecular simulations [article v1.0]," *Living J. Comput. Mol. Sci.*, vol. 1, no. 1, p. 5067, 2019.

[10] L. Martínez, "Automatic identification of mobile and rigid substructures in molecular dynamics simulations and fractional structural fluctuation analysis," *PLoS ONE*, vol. 10, no. 3, Mar. 2015, Art. no. e0119264.

[11] A. Lindström, F. Pettersson, and A. Linusson, "Quantitative protein descriptors for secondary structure characterization and protein classification," *Chemometric Intell. Lab. Syst.*, vol. 95, no. 1, pp. 74–85, Jan. 2009.

[12] C. Chothia, "The nature of the accessible and buried surfaces in proteins," *J. Mol. Biol.*, vol. 105, no. 1, pp. 1–12, Jul. 1976.

[13] F. Tian, P. Zhou, and Z. Li, "T-scale as a novel vector of topological descriptors for amino acids and its application in QSARs of peptides," *J. Mol. Struct.*, vol. 830, nos. 1–3, pp. 106–115, Mar. 2007.

[14] O. Sander, T. Sing, I. Sommer, A. J. Low, P. K. Cheung, P. R. Harrigan, T. Lengauer, and F. S. Domingues, "Structural descriptors of gp120 V3 loop for the prediction of HIV-1 coreceptor usage," *PLoS Comput. Biol.*, vol. 3, no. 3, p. e58, Mar. 2007.

[15] T. Taylor, M. Rivera, G. Wilson, and I. I. Vaisman, "New method for protein secondary structure assignment based on a simple topological descriptor," *Proteins, Struct., Function, Bioinf.*, vol. 60, no. 3, pp. 513–524, Aug. 2005.

[16] A. Andreeva, E. Kulesha, J. Gough, and A. G. Murzin, "The SCOP database in 2020: Expanded classification of representative family and superfamily domains of known protein structures," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D376–D382, Jan. 2020.

[17] I. Sillitoe, N. Bordin, N. Dawson, V. P. Waman, P. Ashford, H. M. Scholes, C. S. M. Pang, L. Woodridge, C. Rauer, N. Sen, M. Abbasian, S. Le Cornu, S. D. Lam, K. Berka, I. H. Varekova, R. Svobodova, J. Lees, and C. A. Orengo, "CATH: Increased structural coverage of functional space," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D266–D273, Jan. 2021.

[18] P. Jain, J. M. Garibaldi, and J. D. Hirst, "Supervised machine learning algorithms for protein structure classification," *Comput. Biol. Chem.*, vol. 33, no. 3, pp. 216–223, Jun. 2009.

[19] P. Jain and J. D. Hirst, "Exploring protein structural dissimilarity to facilitate structure classification," *BMC Structural Biol.*, vol. 9, no. 1, p. 60, 2009.

[20] P. Jain and J. D. Hirst, "Automatic structure classification of small proteins using random forest," *BMC Bioinf.*, vol. 11, no. 1, p. 364, Dec. 2010.

[21] M. Jiang, Z. Li, Y. Bian, and Z. Wei, "A novel protein descriptor for the prediction of drug binding sites," *BMC Bioinf.*, vol. 20, no. 1, p. 478, Dec. 2019.

[22] H. M. Ashtawy and N. R. Mahapatra, "Descriptor data bank (DDB): A cloud platform for multiperspective modeling of protein–ligand interactions," *J. Chem. Inf. Model.*, vol. 58, no. 1, pp. 134–147, Jan. 2018.

[23] M. Cha, E. S. T. Emre, X. Xiao, J.-Y. Kim, P. Bogdan, J. S. VanEpps, A. Violi, and N. A. Kotov, "Unifying structural descriptors for biological and bioinspired nanoscale complexes," *Nature Comput. Sci.*, vol. 2, no. 4, pp. 243–252, Apr. 2022.

[24] F. Molnár, "Structural considerations of vitamin D signaling," *Frontiers Physiol.*, vol. 5, p. 191, Jun. 2014.

[25] D. Moras and H. Gronemeyer, "The nuclear receptor ligand-binding domain: Structure and function," *Current Opinion Cell Biol.*, vol. 10, no. 3, pp. 384–391, Jun. 1998.

[26] M. T. Mizwicki and A. W. Norman, "The vitamin D sterol–vitamin D receptor ensemble model offers unique insights into both genomic and rapid-response signaling," *Sci. Signaling*, vol. 2, no. 75, p. re4, Jun. 2009.

[27] J. G. Taylor and D. A. Bushinsky, "Calcium and phosphorus homeostasis," *Blood Purification*, vol. 27, no. 4, pp. 387–394, 2009.

[28] S. T. Corbett, O. Hill, and A. K. Nangia, "Vitamin D receptor found in human sperm," *Urology*, vol. 68, no. 6, pp. 1345–1349, Dec. 2006.

[29] A. W. Norman, "From vitamin D to hormone D: Fundamentals of the vitamin D endocrine system essential for good health," *Amer. J. Clin. Nutrition*, vol. 88, no. 2, pp. 491–499, Aug. 2008.

[30] S. Burendahl, E. Treuter, and L. Nilsson, "Molecular dynamics simulations of human LRH-1: The impact of ligand binding in a constitutively active nuclear receptor," *Biochemistry*, vol. 47, no. 18, pp. 5205–5215, May 2008.

[31] T. Zhou, Y. Zhang, A. Macchiarulo, Z. Yang, M. Cellanetti, E. Coto, P. Xu, R. Pellicciari, and L. Wang, "Novel polymorphisms of nuclear receptor SHP associated with functional and structural changes," *J. Biol. Chem.*, vol. 285, no. 32, pp. 24871–24881, Aug. 2010.

[32] J. Jyrkkärinne, J. Küblbeck, J. Pulkkinen, P. Honkakoski, R. Laatikainen, A. Poso, and T. Laitinen, "Molecular dynamics simulations for human CAR inverse agonists," *J. Chem. Inf. Model.*, vol. 52, no. 2, pp. 457–464, Feb. 2012.

[33] D. W. Ritchie, "Calculating and scoring high quality multiple flexible protein structure alignments," *Bioinformatics*, vol. 32, no. 17, pp. 2650–2658, Sep. 2016.

[34] N. Rochel, J. M. Wurtz, A. Mitschler, B. Klaholz, and D. Moras, "The crystal structure of the nuclear receptor for vitamin D bound to its natural ligand," *Mol. Cell*, vol. 5, no. 1, pp. 173–179, Jan. 2000.

[35] H. Lempiäinen, F. Molnár, M. M. Gonzalez, M. Peräkylä, and C. Carlberg, "Antagonist- and inverse agonist-driven interactions of the vitamin D receptor and the constitutive androstane receptor with corepressor protein," *Mol. Endocrinology*, vol. 19, no. 9, pp. 2258–2272, Sep. 2005.

[36] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers, Original Res. Biomolecules*, vol. 22, no. 12, pp. 2577–2637, Dec. 1983.

[37] K. J. Bowers, D. E. Chow, H. Xu, R. O. Dror, M. P. Eastwood, B. A. Gregersen, J. L. Klepeis, I. Kolossváry, M. A. Moraes, F. D. Sacerdoti, J. K. Salmon, Y. Shan, and D. E. Shaw, "Scalable algorithms for molecular dynamics simulations on commodity clusters," in *Proc. ACM/IEEE Conf. Supercomputing*, Nov. 2006, p. 84.

[38] F. Murtagh and P. Legendre, "Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion?" *J. Classification*, vol. 31, no. 3, pp. 274–295, Oct. 2014.

[39] T. Venäläinen, F. Molnár, C. Oostenbrink, C. Carlberg, and M. Peräkylä, "Molecular mechanism of allosteric communication in the human PPARα-RXRα heterodimer," *Proteins, Struct., Function, Bioinf.*, vol. 78, no. 4, pp. 873–887, Mar. 2010.

[40] L. Schrödinger and W. DeLano. *Pymol*. [Online]. Available: http://www.pymol.org/pymol

[41] J. Abramson et al., "Accurate structure prediction of biomolecular interactions with AlphaFold 3," *Nature*, vol. 630, no. 8016, pp. 493–500, May 2024.

[42] I. Kufareva and R. Abagyan, "Methods of protein structure comparison," in *Homology Modeling: Methods and Protocols*. Totowa, NJ, USA: Humana, 2012, pp. 231–257.

[43] D. W. K. Andrews, "Heteroskedasticity and autocorrelation consistent covariance matrix estimation," *Econometrica*, vol. 59, no. 3, pp. 817–858, May 1991.

[44] A. Zemla, Č. Venclovas, J. Moult, and K. Fidelis, "Processing and evaluation of predictions in CASP4," *Proteins, Struct., Function, Genet.*, vol. 45, no. S5, pp. 13–21, 2001.

[45] H.-W. Wang, C.-H. Chu, W.-C. Wang, and T.-W. Pai, "A local average distance descriptor for flexible protein structure comparison," *BMC Bioinf.*, vol. 15, no. 1, p. 95, Dec. 2014.

[46] J. Gao, T. Zhang, H. Zhang, S. Shen, J. Ruan, and L. Kurgan, "Accurate prediction of protein folding rates from sequence and sequence-derived residue flexibility and solvent accessibility," *Proteins, Struct., Function, Bioinf.*, vol. 78, no. 9, pp. 2114–2130, 2010.

[47] A. Teichert, L. A. Arnold, S. Otieno, Y. Oda, I. Augustinaite, T. R. Geistlinger, R. W. Kriwacki, R. K. Guy, and D. D. Bikle, "Quantification of the vitamin D receptor- coregulator interaction," *Biochemistry*, vol. 48, no. 7, pp. 1454–1461, 2009.

KARINA PATS received the bachelor's degree in biotechnology from Saint Petersburg Chemical Pharmaceutical Academy, in 2017, and the M.Sc. degree in applied informatics with a specialization in cheminformatics and molecular modeling from ITMO University, in 2019. She is currently a ML Engineer with Quantori, where she focuses on implementing LLM-based solutions for the life science domain. Her research interests include bioinformatics data analysis, cheminformatics, drug discovery, and the application of machine learning in pharmaceutical sciences.

IGOR GLUKHOV received the B.Sc. degree in software engineering and the M.Sc. degree in bioinformatics and system biology from ITMO University, in 2021 and 2023, respectively. He is currently a Data Scientist with the X5 Retail Group, Moscow, Russia. He has publications in the fields of predicting students' academic outcomes and developing recommender systems for academic supervisors. His research interests include machine learning, data science, and learning analytics.

STEPAN PETROSIAN received the bachelor's degree in chemical technology from Saint Petersburg State Chemical Pharmaceutical Academy, in 2020. He is currently a QA Engineer with Semrush, Amsterdam, The Netherlands. His professional experience includes roles as a QA Automation Engineer with EPAM Systems and Genesys, where he developed test automation frameworks and implemented test cases for various software systems. His technical skills include Python, R, and statistics, with a focus on quality assurance and software testing in the IT industry.

MARIA MAMAEVA received the Specialist degree in pharmacy from Saint Petersburg Chemical Pharmaceutical Academy, in 2020. She completed additional studies in bioinformatics with the Bioinformatics Institute, from 2019 to 2021. She is currently a Statistical Programmer II with Fortrea. Her professional experience includes roles in statistical programming within the pharmaceutical industry, focusing on the production and validation of TFLs, ADaM, and SDTM datasets in various therapeutic areas.

ALEXEY SERGUSHICHEV received the B.Sc., M.Sc., and Ph.D. degrees in computer science from ITMO University, Saint Petersburg, Russia, in 2016. He is currently an Assistant Professor with Washington University in St. Louis. His research interests include systems biology, next-generation sequencing, gene expression analysis, pathway enrichment analysis, and biological network analysis. He has developed widely adopted bioinformatics methods and software, including the FGSEA method for gene set enrichment analysis and phantasus for visual and interactive gene expression analysis.

MARIE-DOMINIQUE DEVIGNES received the dual master's degree in biochemistry and physiology from the University of Paris VII and University of Paris VI, in 1980, the Ph.D. degree in molecular biology from the University of Paris VII, in 1982, the Doctorat d'Etat es Sciences (equivalent to Habilitation) degree, in 1988, and the master's degree in computer sciences from Henri Poincaré University, in 2001. She is currently a member of the CAPSID Team, Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA). Her research interests include computational algorithms for protein structure and interactions, biological data integration and mining, machine learning, semantic web, and structural bioinformatics. She has over 40 years of experience in molecular biology, genetics, and bioinformatics; and has published 61 journal articles. She has been actively involved in organizing international conferences in bioinformatics and serves as the Coordinator of Interoperability Action with the Institut Français de Bioinformatique.

FERDINAND MOLNÁR (Member, IEEE) received the M.Sc. degree in biochemistry from Comenius University, Bratislava, Slovakia, in 2001, and the Ph.D. degree in biochemistry with minors in pharmacology and bioinformatics from the University of Kuopio, Kuopio, Finland, in 2006. From 2006 to 2008, he was a Postdoctoral Researcher with the Institute of Genetics and Molecular and Cellular Biology (IGBMC), Illkirch, France. In 2008, he returned to Finland and held various positions with the University of Eastern Finland, Kuopio. Before joining Nazarbayev University as an Associate Professor with the Department of Biology, in 2018, he held a position of a Senior Researcher and a Laboratory Head of the University of Eastern Finland. His current research interests include structural and chemical biology, chem-bio informatics, drug discovery, and ML in chemistry.