Computers & Graphics 98 (2021) 293–305

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Special Section on MoIVA 2021

# A framework for uncertainty-aware visual analytics of proteins

Robin G.C. Maack[a,*], Michael L. Raymer[b], Thomas Wischgoll[b], Hans Hagen[a], Christina Gillmann[c]

[a] University of Kaiserslautern, Paul-Ehrlich-Straße 36, Kaiserslautern 67663, Germany
[b] Wright State University, 3640 Col. Glenn Hwy., Dayton, OH 45435, United States
[c] Leipzig University, Augustusplatz 10, Leipzig 04109, Germany

ARTICLE INFO

*Article history:*
Received 22 February 2021
Revised 12 May 2021
Accepted 18 May 2021
Available online 2 June 2021

*Keywords:*
Molecular visualization
Protein visualization
Uncertainty visualization
Multi-views

ABSTRACT

Due to the limitations of existing experimental methods for capturing stereochemical molecular data, there usually is an inherent level of uncertainty present in models describing the conformation of macromolecules. This uncertainty can originate from various sources and can have a significant effect on algorithms and decisions based upon such models. Incorporating uncertainty in state-of-the-art visualization approaches for molecular data is an important issue to ensure that scientists analyzing the data are aware of the inherent uncertainty present in the representation of the molecular data. In this work, we introduce a framework that allows biochemists to explore molecular data in a familiar environment while including uncertainty information within the visualizations. Our framework is based on an anisotropic description of proteins that can be propagated along with required computations, providing multiple views that extend prominent visualization approaches to visually encode uncertainty of atom positions, allowing interactive exploration. We show the effectiveness of our approach by applying it to multiple real-world datasets and gathering user feedback.



## 1. Introduction

The important role of visualization in molecular biology has been outlined by Olson et al. [1] and is applied to many tasks such as folding prediction, detection of active sites, or structural examination. Several approaches have been proposed in this area [2] in the last decades and are available in a variety of open-source tools, making visualization a state-of-the-art tool in biochemical applications. Molecular data can originate from a variety of sources, such as X-ray crystallography, Nuclear Magnetic Resonance Spectroscopy (NMR), Cryogenic electron microscopy (Cryo-EM), or atomic simulations. Due to the experimental nature of these approaches, several types of uncertainty are embedded in the acquired data which mostly results in positional uncertainty of the captured molecular structures.

This uncertainty affects the decision-making process in visual analytics tasks and needs to be visually communicated, as shown by Sacha et al. [3]. This especially applies to molecular data as it is often used for drug development. When considering molecular data, the origin of uncertainty, the resulting positional variation, and further requirements from the biochemical domain need to be considered in order to provide uncertainty-aware visualization approaches, as shown in Section 2. As molecular biologists are increasingly finding it necessary to employ a wide range of computational tools in their work, a framework that can be intuitively used, without the need for special training, is an important aspect that needs to be considered when developing novel visualization approaches in this domain.

So far, molecular visualization approaches often lack the ability to communicate uncertainty or, if they are available, they need to be included in existing visualization frameworks, as shown in Section 3. In this work, we propose a visualization framework for molecular data that is affected by uncertainty. We provide an uncertainty-aware description of atom positions and show how this knowledge can be inserted in arbitrary computations based on these positions (see Section 4). To incorporate this knowledge, we propose a multi-view system that is composed of prominent visualization approaches in molecular biology, such as volumetric visualization, Ramachandran plots, and statistics views. We adapted each of the visualization approaches such that they are able to visually indicate the positional uncertainty of atoms in specific proteins. The linked views are highly interconnected to provide user interaction for exploratory tasks (see Section 5).

* Corresponding author.
*E-mail addresses:* maack@rhrk.uni-kl.de (R.G.C. Maack), michael.raymer@wright.edu (M.L. Raymer), thomas.wischgoll@wright.edu (T. Wischgoll), hagen@cs.uni-kl.de (H. Hagen), gillmann@informatik.uni-leipzig.de (C. Gillmann).

https://doi.org/10.1016/j.cag.2021.05.011
0097-8493/

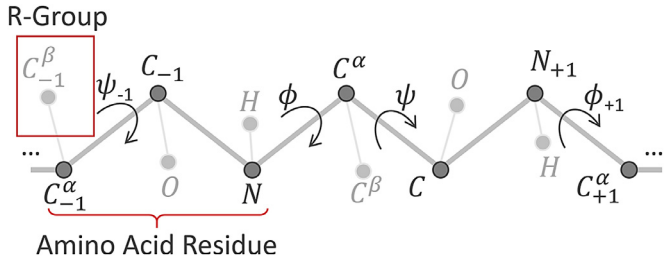


**Fig. 1.** General setup of a protein. The backbone (bold lines) are composed of repeating CCN-Atoms of amino acid residues, where each residue has an R-group attached to it.

Therefore, this paper contributes:

- Uncertainty-aware description of proteins and requirements for their visualization
- Uncertainty-aware visual analytics framework for protein structures and their properties

We show the effectiveness of the presented framework by applying it to real-world molecular data and demonstrate how the uncertainty-aware visualization approaches perform (see Section 6). Our results will be discussed in Section 7.

## 2. The role of uncertainty in molecular data

### 2.1. Biochemical basics

Proteins are the molecular machinery of biological systems. They are among the most abundant of biological macromolecules and undertake a diversity of roles in living systems. Structurally, proteins consist of polymers (chains) of amino acids. There are 20 different amino acids typically found in human proteins. A single protein can consist of one or more amino acid chains, typically ranging in length from a few hundred to several thousand single entities. We would like to refer to the introduction to proteins and their structure by Buxbaum [4].

Fig. 1 shows that a protein contains a backbone that forms when amino acids bind together. This backbone is a repetition of $C^{\alpha}_{-1}$, $C_{-1}$, $N$ and $C_{\alpha}$ chains. The connections between amino acids allow for two rotational angles along with covalent bonds between connected amino acids. Such angles are called dihedral angles, meaning the angle between two planes spanned by four neighboring atoms, such that the planes intersect along the line between the two middle atoms. The two dihedral angles of each amino acid residue, typically referred to as $\phi$ and $\psi$, allow proteins to adopt a wide variety of three-dimensional structures.

### 2.2. Uncertainty in biochemical data

Modern methods for protein structure determination, including X-ray crystallography, Cryo-EM, and nuclear magnetic resonance spectroscopy, can provide three-dimensional structures of soluble polypeptides with high confidence. The resolution of these structures is sufficiently high that the location of individual protein and ligand atoms can, in many cases, be determined with precision to within a few angstroms. There are several sources of uncertainty related to these atomic positions. Proteins are not monolithic, rigid molecules [5]. Even within the context of protein crystals, local protein regions are subject to thermal mobility to differing extents [6].

Intrinsic disorder and local mobility have been identified as important factors in protein ligand-binding and allosteric functional mechanisms [7–9]. A variety of computational methods have been developed for the prediction of intrinsic disorder in proteins based on structural and sequence constraints [10–12], and at least one experimental method for the characterization of intrinsic disorder has been proposed [13].

In X-ray crystallography experiments, structural uncertainty arises from several sources. Thermal-related mobility of backbone and side-chain atoms within the crystal structure lead to a blurring effect on the electron density maps induced by x-ray diffraction. Additionally, the nature of the crystallization process can lead to multiple structural variants of a protein within a single crystal, possibly leading to one or more alternative locations for each protein atom. While resolving a three-dimensional structure, this uncertainty is captured in the *B-Factor* (thermal mobility) and occupancy terms for each atom. In practice, the occupancy is often constrained to a value of 1.0 while the B-Factor term is used to express the overall structural uncertainty of the atom [14,15].

Both X-ray crystallography and Cryo-EM experiments generally include one or more computational modeling steps. For X-ray structures, this step involves minimization of unexplained observed electron densities. For Cryo-EM experiments, the modeling involves automated or manual groping and alignment of two-dimensional particle images, followed by the calculation of a three-dimensional map from the aligned images. In both cases, artifacts of the modeling process can contribute additional uncertainty to the resulting three-dimensional structure.

For NMR-spectroscopy, structural constraints are captured through the NMR experiment, and a number of structural models consistent with those constraints are generated. In this case, an atom's structural uncertainty is related to the variability of its position across the resulting model set.

It is important to note that not all conformational uncertainty can be captured and visualized. In X-ray crystallography, for example, the observed proteins may adopt a non-native structure to facilitate the formation of a crystal lattice. The difference between the native structure and the crystal structure is not known and is thus not captured in the experimental data.

As experimental methods for observing the structure of biological macromolecules have continued to advance, many sources of uncertainty in the resulting molecular models remain. A clear understanding of variability and uncertainty in protein structure is important to biochemists in a wide variety of scientific contexts including cognizant drug design, docking, ligand screening, structural homology modeling, protein function assessment, and more.

### 2.3. Requirements for uncertainty-aware protein visualization

Based on the previous state-of-the-art analysis and application analysis, we are able to determine a list of requirements, which will be presented in the following. We followed the suggestion of Lam et al. [16] where interviews were described as a proper tool to understand the need of users. Here, we used the requirements defined by Gillmann et al. [17] that are formulated to promote real-world use of novel visualization approaches. The list contains 16 *low-level requirements* which are sorted into 5 categories (usability, effectiveness, correctness, flexibility, and intuitiveness). We showed the list to a domain expert from biochemistry and a visualization expert to cover both views on the proposed topic. First, we let both experts express the importance of each requirement to be fulfilled. Here, a Likert scale was used (1 unimportant, 5 very important). The results can be found in Table 1.

It can be observed that avoidance of clutter, uncertainty visualization, different use cases, interactivity, and ease of use are highlighted as important by both experts. Further, interactivity and use time efficiency are also listed as very important or important by at least one of the experts. As these requirements are rather general and low-level, we have used them to derive *high-level requirements*,

**Table 1**
User evaluation performed for the presented approach. 16 Low level requirements have been evaluated with a Likert scale from 1 to 5 in their importance from two experts (visualization and domain). High ratings in importance are highlighted in gray. Our approach has been evaluated against two known approaches by each of the experts. The results are color-coded in red (if our approach is rated worse than the known visualization tools), yellow (if our approach is rated better than known visualization tools) and green (if our approach was rated better than the known tools).

| Category | Requirement | Visualization expert | | | | Domain expert | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Imp. | PyMol | Protoshop | Our Tool | Imp. | PyMol | VMD | Our Tool |
| Usability | Collaborative | 3 | 4 | 4 | 4 | 3 | 2 | 2 | 3 |
| | Interactivity | 4 | 4 | 4 | 4 | 5 | 3 | 3 | 3 |
| | Avoidance of Clutter | 5 | 4 | 3 | 5 | 4 | 2 | 1 | 4 |
| | Minimized Input Parameters | 4 | 4 | 4 | 4 | 3 | 2 | 1 | 3 |
| | Compatibility | 5 | 4 | 4 | 4 | 3 | 4 | 4 | 2 |
| Effectiveness | Runtime Efficiency | 4 | 3 | 3 | 4 | 2 | 3 | 3 | 3 |
| | Memory Efficiency | 4 | 3 | 3 | 4 | 2 | 3 | 3 | 3 |
| | Use Time Efficiency | 4 | 3 | 3 | 4 | 4 | 3 | 3 | 3 |
| Correctness | Precision | 5 | 4 | 4 | 4 | 3 | 5 | 5 | 5 |
| | Quantification | 5 | 4 | 4 | 4 | 3 | 4 | 5 | 4 |
| | Uncertainty | 5 | 1 | 1 | 4 | 5 | 2 | 2 | 5 |
| Flexibility | Use Cases | 5 | 4 | 4 | 5 | 4 | 3 | 4 | 3 |
| | Different Datasets | 5 | 4 | 4 | 4 | 4 | 3 | 4 | 4 |
| Intuitiveness | Feedback Loop | 5 | 4 | 4 | 5 | 3 | 3 | 3 | 3 |
| | Easy to use | 5 | 4 | 3 | 5 | 4 | 3 | 1 | 3 |
| | No background knowledge | 5 | 4 | 3 | 4 | 2 | 3 | 1 | 3 |
| | Weighted Average | | 3.63 | 3.44 | **4.25** | | 3.00 | 2.81 | **3.38** |

which can be found in the following. The list is further maintained by the requirements to achieve uncertainty-awareness in visual analytics tasks developed by Sacha et al. [3].

**R1: Visualization of positional uncertainty**. As the captured position of atoms in a protein can have a huge impact on the computation of properties of the considered protein, the propagation of these uncertainties is required. Here, every computational step based on atom positions needs to be adapted according to the underlying uncertainty captured or computed for each atom [18].

**R2: Integration into known visualizations**. There exist a consensus of visualization techniques that are suitable in the biochemical domain [2]. These visualizations have been proven to fulfill the requirements of molecular visualization. A visual representation that includes uncertainty information should be an extension of known visualization paradigms.

**R3: Avoidance of visual clutter**. Visual clutter, in the sense of this application, refers to the numerous occlusion of objects or information by overlying objects, resulting in a tangled visualization. As many atoms are displayed in a 3D scene, potentially even displaying the superposition of various protein models at the same time, summarizing the available information is an important step to communicate the available data in a compressed manner. A suitable visualization strategy should reduce visual clutter in full measure to allow the user to understand the outline of the dataset.

**R4: Interactive visualization framework**. As the visualization of proteins usually results in a three-dimensional object, users need to be able to explore the data utilizing suitable interaction paradigms [17].

## 3. Related work

In this section, we attempt to provide a summary of related work in terms of uncertainty-aware molecular visualization approaches as well as open-source frameworks that implement these approaches.

### 3.1. Uncertainty-aware protein visualization

In the field of visualization, the inclusion of uncertainty was classified as one of the most important research problems by Johnson [19], as it cannot be implemented right away. Brodlie et al. [20], as well as Potter et al. [21], divided uncertainty visualization challenges using the dimension of their data and the dimension of data points. Here, we obtain a valuable starting point as we can consider molecular data as scalar data.

Molecule and protein visualizations are widely used. Therefore, a large number of projects have been addressing issues of biochemical, pharmaceutical, and medical researchers as well as their industry members. Kozlíková et al. [2] summarized the variety of visualization options in a state-of-the-art analysis. Here, the selection of a proper uncertainty representation was named as one of the main challenges [22]. In the following, the most important uncertainty-aware protein visualization strategies related to our approach are summarized.

Rheingans and Joshi [23] used various family members of molecules holding the same atoms and bonds. They either superimpose the members showing regions of high uncertainty by large disagreement between the confirmation states or show iso-surfaces using Gaussian splatting to indicate the likelihood of an atom to be located at a set location. Although this provides a visualization of all potential protein positions, it introduces visual clutter in the resulting visualization. Instead, the reduction of visual clutter in these visualizations is focused on in this manuscript.

Rasheed et al. [24] utilized volume rendering to show the per voxel uncertainty computed across an ensemble of slightly perturbed samples of the same molecule. It showed that the B-Factor uncertainty correlates with this distribution function uncertainty. Knoll et al. [25] provided a volume rendered uncertainty classification based on electron density distributions using 2D transfer functions, helping to identify interfaces based on chemical bond forces. Skånberg et al. [26] used volume rendering of spatial distribution functions to visualize the distribution of selected structures over ensembles of molecules.

Schulz et al. [27] presented a model visualizing the uncertainty of secondary structure assignments on ribbon diagrams, comparing various assignment algorithms. In contrast to this work, they used various assignment algorithms as their source of uncertainty instead of the positional uncertainty of atoms. The visualization was made more squiggly in areas of high uncertainty instead of using iso-surfaces for the visualization where the original geometry can still be seen. In contrast to this contribution, we aim to focus on the positional variations of atoms in proteins as a source of uncertainty while preserving the original shape of the ribbon model.

Lee and Varshney [28] created an Uncertainty-aware 3D visualization of the Solvent Excluded surface using Gaussian distributions and a fuzzy rendering mode. Even though this approach seemed promising, the appearance of the results is comparable to blurring the original visualization.

Sasisekharan [29] developed the usage of dihedral angles to describe polypeptide conformations. This so-called Ramachandran plot was used throughout biochemical research quickly [30,31]. An uncertainty-aware version of this plot, indicating the variations in the dihedral angles, is provided here.

Maack et al. [32] created an Uncertainty-Aware version of the Ramachandran plot to allow interaction with other components of the system while improving the visual appearance. Unfortunately, their model solely focuses on isotropic uncertainty as captured by B-Factors. In the presented work, we introduce an anisotropic model of uncertainty for each atom position and extend the Ramachandran plot according to this knowledge.

### 3.2. Uncertainty-aware protein visualization software

A major drawback of all presented visualization approaches is that they are rarely used in open-source visualization tools. As Gillmann et al. [33] pointed out, this is a desirable feature in many applications. The protein data bank provides a list of open-source visualization tools [34] in general. We used this list and extracted the tools that incorporate uncertainty.

Chimera is an open-source tool for visualizing molecular structures [35]. Variations in molecular modeling can be visualized by plotting multiple proteins at the same time. Atoms can be exchanged and a Ramachandran plot can be used to examine the effect of the deviations in atom positions. Here, the resulting visualization can quickly become cluttered. In the presented approach various protein models are summarized within one visualization, such that the amount of visual clutter is minimized as much as possible.

Polyview 3D [36] allows for the visualization of multiple positions of atoms in a protein using animation. Although this provides a nice visualization of different protein formations, it is only able to show one formation at a point in time. In contrast to this, our approach aims to show all confirmations simultaneously if requested while giving users the freedom to watch any formation on demand.

The molecular visualizer iMol [37] uses motion blur to indicate variations in atom positions. Although this shows the uncertainty of atom positions, it causes visual clutter if there exist a large number of blurred areas.

Swiss PDBViewer [38] represents uncertainty using color-coding. Here, atoms or amino acid residues that hold high amounts of positional uncertainty are shown in red highlight color. This allows visualization of positional uncertainty without introducing additional visual clutter, allowing to show the positional displacement in space. Instead, the presented approach seeks to provide a trade-off between minimal visual clutter and inclusion of potential positions of an atom.

## 4. Uncertainty-aware description and properties of proteins

### 4.1. Uncertainty-aware description of molecular data

As shown in Section 2, there are several sources of uncertainty and different ways to express them when capturing biochemical data. When not considering uncertainty, atom positions are usually treated as fixed Cartesian coordinates. As this ignores the fact that atoms have a certain range of movement, captured by the B-Factor or multiple models of the same protein, the uncertainty of atoms will be described in an isotropic and anisotropic way.

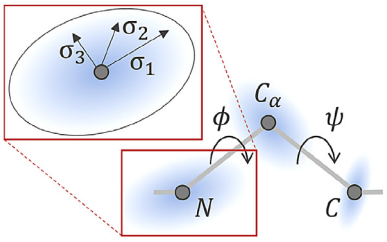


**Fig. 2.** Uncertainty-aware description of atom positions in a protein using three eigenvectors of the standard deviational ellipse.

To make use of the available uncertainty information, we utilize an extended description of points that allows the visualization to capture positional uncertainty, introduced by Gillmann et al. [39], referred to as probabilistic points. It describes every point by a center $\Sigma$ and three orthogonal vectors $\sigma_1, \sigma_2, \sigma_3$ to describe the available movement in each direction. Those orthogonal vectors do not have to be axis-aligned and either depend on the distribution of the same atom in multiple models or the B-Factor.

Fig. 2 shows how each atom in a protein can be modeled as a probabilistic point. To achieve quantification of uncertainty for the underlying data, two types of uncertainty descriptions for molecular data have to be distinguished: *isotropic* and *anisotropic*.

*Isotropic model* When using a single model with a B-Factor value attached to each atom, the root-mean-square displacement can be retrieved from the B-Factor, yielding an isotropic model for the atom's movement. To highlight regions of high or low uncertainty the normalized B-Factor, or the root-mean-square displacement, can be utilized to visualize the isotropic model. Normalized uncertainty can be scaled using a parameter to visually encode differences better. This was highly recommended by our collaborators when working with isotropic models, helping to better distinguish uncertainty information.

In the isotropic case, the measured position of an atom is its equilibrium position, used as the center of a probabilistic point ($\Sigma$). This means that the model used to create data estimated this position using the most likely conformation simulated. Additionally, the $\sigma$-values of each probabilistic point will be set to the root-mean-square displacement ($u$) retrieved from the B-Factor of the considered atom as shown in Eq. (1) [40].

$$B = 8\pi^2 u^2 \iff u = \sqrt{\frac{B}{8\pi^2}} \tag{1}$$

This is a suitable assignment, as the root-mean-square displacement captures the available movement of the respective atom in each direction. Although probabilistic points can be modeled with anisotropic $\sigma$-values in each dimension, the available B-Factor is an isotropic description of the positional uncertainty. Therefore, each dimension gets the same $\sigma$-value assigned. In total, $\Sigma$ and $\sigma$ can be utilized to define a three-dimensional distribution function that is able to output the Gaussian probability density for an atom to be located at an arbitrary point in a three-dimensional space.

*Anisotropic model* On the other hand, multi-model data, like *Nuclear magnetic resonance* model ensembles, capture multiple positions of each atom. This leads to an anisotropic distribution of points around their average. Therefore, an anisotropic approach can be used instead. For this purpose, the average position of each atom over all available models has to be calculated first. Those average positions are needed to find the covariance matrix for each atom thereafter. In this way, the distribution of each atom over all models can be described by retrieving the eigenvectors and corresponding eigenvalues of said covariance matrices using a process called *Eigendecomposition*. The eigenvector with the biggest eigenvalue describes the direction with the most substantial standard

deviation, while the eigenvector with the smallest eigenvalue gives the direction of the least substantial standard deviation. This way, the normalized eigenvectors can be multiplied with the square root of their eigenvalues to create $\sigma_1, \sigma_2, \sigma_3$ [41]. As the covariance matrix is symmetric, the eigenvectors are orthogonal which can be used to create a *Standard Deviational Ellipse*(SDE) around each $\Sigma$. It should be noted that the SDE is actually no ellipse [42], but a special type of curve [43]. This fact was ignored to better preserve the directions of the eigenvectors and to save time while computing uncertainty hulls.

### 4.2. Uncertainty-aware computation of dihedral angles

Besides the challenge that the representation of uncertainty-aware proteins varies from the classic representation of a protein, the uncertainty-aware description of proteins also adds additional information to their dihedral angles. This information contains the available change of said angles induced by the available movement of their atoms. In the following, both an isotropic and anisotropic representations of dihedral angles will be described.

*Isotropic model* In the isotropic case, dihedral angles depend on the position of four atoms each. If the considered atom uncertainties describe a lot of movement, the corresponding dihedral angle can also strongly change. This should be reflected in the computation of the dihedral angle uncertainty. Therefore, the dihedral angle uncertainty consists of the average normalized B-Factors of the considered atoms. This directly links the angle uncertainty to the uncertainty of its atoms, while removing any influence of the angle value itself. Also, the uncertainty can be directly scaled to better show highly uncertain regions in the Ramachandran plot.

*Anisotropic model* For the computation of dihedral angles using anisotropic uncertainty, the distribution of a dihedral angle can be extracted from its distribution over all models. Therefore, the average dihedral angle of each amino acid residue has to be calculated first. Then, the covariance matrix of each amino acid residue can be created. It should be noted that the distance metric has to be chosen with respect to angles, e.g. the distance between −179 degrees and 179 degrees is 2. Similar to the representation of probabilistic points, the eigenvectors and eigenvalues can be extracted from the covariance matrices. This way, the average dihedral angle and its distribution can be shown.

A consistent view about the underlying uncertainty of dihedral angles is given in both cases, as they are handled in a similar fashion to their corresponding atoms. The isotropic case directly shows the average uncertainty of the underlying atoms, whereas distributions of atoms or dihedral angles are shown in the anisotropic case. This visually and conceptually connects the 3D visualization with the Ramachandran plot.

## 5. A Framework for uncertainty-aware visual analytics of proteins

In order to devise a visual analytics tool for researchers in biochemistry that allows users to review uncertainty in protein data, we created an interactive multi-view framework. The framework expands on well-established views in the biochemical community. According to our domain experts, volume views and Ramachandran plots are the most important visualization approaches that are used in daily tasks. Our system builds upon these two visualization strategies of proteins while including uncertainty information and allowing interaction.

### 5.1. Uncertainty-aware volume visualization

One of the major goals of this work is to enhance the current visualization capabilities by including uncertainty information. The utilized 3D visualization methodologies are being presented here. The inclusion of uncertainty in the field of proteins consists of showing possible movements of atoms in the examined protein. Therefore a mutable transparent isosurface around known visualizations is used to provide the original and uncertainty-aware version in a mixed display, shown in Fig. 5. Using transparency and color controls, the user can blend in the uncertainty information as much as desired. The goal is to create a three-dimensional barrier that indicates the potential mobility of protein atoms considering specific possible freedom of movement. Throughout long discussions with our collaborators in the biochemical domain, we determined that uncertainty hulls should be computed based on the underlying geometric representations of proteins.

#### 5.1.1. 3D protein visualization types

The Van der Waals surface visualization draws a sphere for each atom with the Van der Waals force of the corresponding chemical element as its radius. The sphere representation is stored as a geometry which will be referred to as $g$ hereafter. For any geometry, $b(g)$ is defined as the boundary of $g$ including minimum $b(g)_{\min}$ and maximum $b(g)_{max}$ of its points in space. The Ball-and-Stick visualization and the Solvent Accessible Surface use either constant radii or set it to the Van der Walls radii plus a constant solvent radius.

Apart from these spherical visualizations, the Ribbon model is also implemented, drawing a spline curve using the backbone atoms of a protein as their control polygon. The spline is represented by triangles, receiving their uncertainty information from the closest backbone atoms. This is needed as the Ribbon model only consists of triangles. To achieve a mapping between a triangle to a B-Factor, the closest atom to each vertex of a triangle is found and their uncertainty information is averaged. This allows for sufficiently fast computation, while still retrieving a well-suited estimate. Without averaging, one of the vertices would have to be picked using the uncertainty information of the closest atom, leading to high inaccuracy in some cases. Another approach would be to average the positions of triangle vertices, then finding the closest atom to the middle.

#### 5.1.2. Creating the uncertainty scalar field

In order to create a surrounding hull, a structure that represents the distance of each point in space to the closest primitive in an atom representation is required. Although this task can be solved analytically, it would result in significant computational effort. Also, it would have to be completely recomputed when choosing a different $\sigma$-distance from the geometry. In order to reduce the computational effort and to allow isosurface scaling, a scalar field $F$, using the boundaries mentioned in Section 5.1.1, is overlaid with the original protein representation. Depending on the resolution of this scalar field, we are able to discretize the distance of points in space to the closest point of the protein geometry with relatively high accuracy. The distance to the nearest piece of geometry, measured in a directional $\sigma$-value, is evaluated for each scalar in $F$. Measuring the distance in units of $\sigma$ is needed as each atom has a different $\sigma$-value. This approach presents a clear advantage in that the scalar field has to be computed only once, such that isosurfaces can be created at any $\sigma$-value chosen by the user.

#### 5.1.3. Distance of a voxel to an atom under uncertainty

To compute the distance of an atom to a voxel center measured in a directional $\sigma_d$ unit, first, the point of intersection $I$ of an SDE with the line between the voxel center $S_c$ and the ellipse center $E_c$ has to be found. Then, $\sigma_d$ is found by taking the distance between $I$ and $E_c$ subtracting the radius.

$$\sigma_d = \|E_c - I\| - R \tag{2}$$

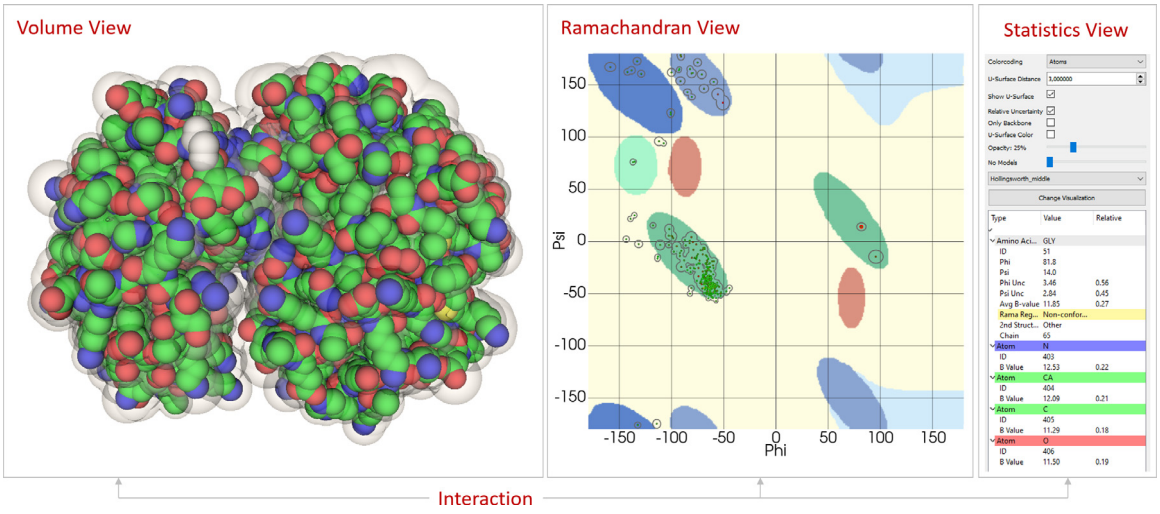


**Fig. 3.** Overview over the presented visualization framework. The framework consists of a volume view, Ramachandran plot view and a statistics view. The views are linked with a hover interaction methodology. Here, the 1H97 protein is shown.

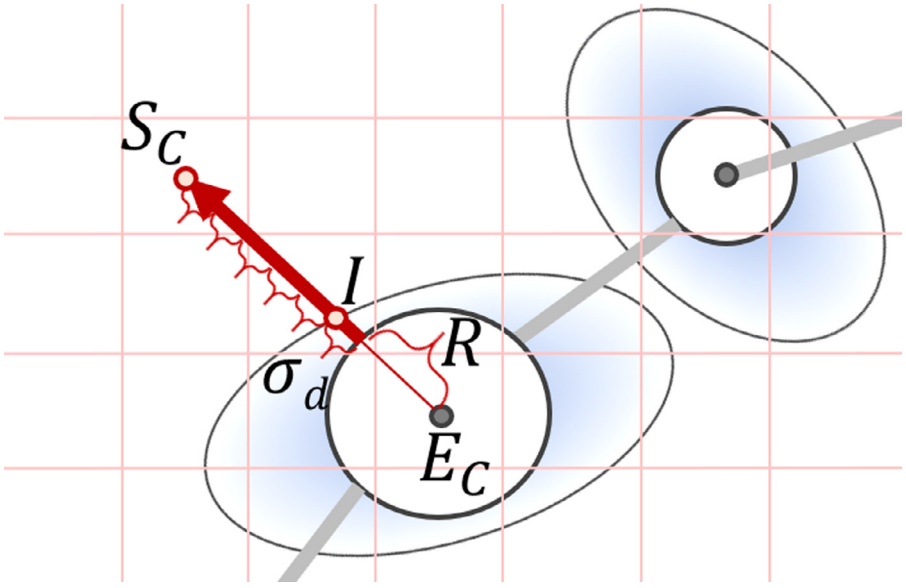


**Fig. 4.** Computation of a scalar field yielding discretized values of differences to a considered geometry in $\sigma$ scale. The example shows the scalar at the tip of the red vector is $6\sigma$ away.

$\sigma_d$ is a measure of one standard deviation in a direction given by $E_c$ and $S_c$. Therefore, the distance between $E_c$ and $S_c$ minus the radius has to be divided by $\sigma_d$ to retrieve the distance between the atom and the voxel in units of $\sigma_d$, allowing the user to later scale the hull to their liking.

$$D_\sigma = (\|S_c - E_c\| - R) \div \sigma_d \tag{3}$$

In Eqs. (2) and (3) the radius $R$ always has to be subtracted as it is a property of the representation itself, thus it is not changing when a different standard deviation threshold is chosen by the user. It should also be noted that using an isotropic model results in a simpler calculation of $\sigma_d$, as the uncertainty is the same in each direction, i.e. $\sigma_d = \sigma_1 = \sigma_2 = \sigma_3$. This function, calculating the distance between voxel center $S_c$ and atom center $E_c$, with ellipse axis $\sigma_1, \sigma_2, \sigma_3$ and radius $R$, will be called $\sigma(S_c, E_c)$ for later computations (Fig. 4).

As described in Section 5.1.1, the Ribbon model is represented by triangles such that the distance computation needs some modification. Obviously, the distance of a point to a triangle has to be found first [44]. Then, a sigma value has to be assigned to the triangle for the isotropic case. In this case, we chose to average the contribution of the closest atoms to each vertex. In the anisotropic case, the center of the triangle is used to represent the triangle as the center of an ellipsoid, using the properties of the closest atom to that point.

To extract the uncertainty hull based on the computed field, iso-surfaces are utilized. As the values in the scalar field represent the distance measured in a directional $\sigma$ distance from the next geometry, a surface (e.g. with distance $1\sigma$ from the geometry) can be created showing the possible mobility of the atoms in their equilibrium position under a statistical model. Therefore, changing the iso-surface threshold is a simple operation and can be done seamlessly on modern hardware.

#### 5.1.4. Filling the scalar field

For each of the considered geometries, the boundary of the scalar field has to be calculated with an as-small-as-possible extent, allowing enough space to fit the iso-surface but small enough

to keep the resolution high and the size of voxels low. Therefore, the boundary $b(g)$ of the original geometry is offset in each direction by a constant value $(c)$ that depends on the expected maximum offset and maximum uncertainty, resulting in a bounding box $b(F)$ of the scalar field $F$.

To calculate the scalar field for the sphere visualization, a Kd-Tree saving all atom positions is generated. This structure allows for an efficient search of close points in space. Then, for all voxels in a discrete scalar field, created as described above, the $n$ atoms closest to its center are saved to a list $L$, efficiently created by the aforementioned Kd-Tree $K$. Then, for each atom position on this list, the distance, measured in a directional $\sigma$-value, is calculated from the voxel center to the surface of the current representation. Therefore, the $\sigma(S_c, E_c)$ function, calculating the directional $\sigma$ distance, is used as described in Section 5.1.3. The lowest $\sigma$-distance is then assigned to the voxel.

The number $n$ is the amount of closest atoms that are examined to determine the probability of a protein to be located in a specific cell of $F$. It is a value that influences the quality of the scalar field, especially if the minimum and maximum $\sigma$-values strongly deviate throughout the dataset. As the radius and $\sigma$-values often strongly differ from each other when scanning neighboring atoms, this value should be kept at roughly 10% of the atoms. Otherwise, the iso-surface may lose its elliptical form when using high scaling factors.

To achieve an uncertainty hull based on the scalar field $F$, a scaling factor $f$ is chosen. This factor influences the isovalue at which the isosurface is created. Users can define this factor to set the distance in standard deviations (anisotropic) or root mean squares (isotropic). In principle, this factor can be set to an arbitrary number greater than zero, but in reality $3\sigma$ usually encapsulates all available variations. Thus higher choices are often not needed when ignoring extreme outliers. Fig. 9 shows an example using $f = 1\sigma, 2\sigma, 3\sigma$, while $3\sigma$ encapsulates even the strongest outliers of the dataset.

#### 5.1.5. Protein representations

As shown in Section 2, there exist several ways to visualize proteins in 3D. We provide four protein representations: the space-filling model, the ball-and-stick model, the ribbon model, and the SAS (solvent-accessible surface). Fig. 5 provides all four visualization types with a B-Factor color-coding. Each of the visualization types can be colored according to various features of a protein. This helps to encode important aspects of the amino acid residue that users are interested in. Usually, color is set per atom, but as the ribbon model does not have single atoms, each vertex is colored the way the closest atoms would be colored. In between points the color is being interpolated to receive smooth color-coding throughout the whole surface.

The resulting uncertainty hull can either be colored consistently or according to the relative B-Factor of the closest atom according to uncertainty. A consistently colored hull is shown in gray, preserving the color-coding underneath. Otherwise, the B-Factor color-coding (green to red) is used. Additionally, any transparency level can be chosen. Fig. 6 compares the gray hull with B-Factor colored atoms and the color-coded uncertainty hull with gray atoms.

### 5.2. Uncertainty-aware Ramachandran plot

A Ramachandran plot displays the distribution of dihedral angles in a protein. Throughout the years, by an empirical analysis of such data, many nomenclatures of this plot have been found. As scientists nowadays use a lot of different background maps, a suitable application does need to support multiple Ramachandran nomenclatures. Our tool is able to load any nomenclature provided in a given format, allowing the user to change the color of regions, also mapping each nomenclature to a color-coding for the 3D visualization of any geometry type.

Building upon the uncertainty-aware Ramachandran plot by Maack et al. [32], the creation of isolines in the isotropic case works similar to the isosurface creation for 3D geometries. Therefore a 2D scalar field is being created, filling it with to the nearest dihedral angle combination while considering its uncertainty in the $\phi$ and $\psi$ direction, called *phiu* and *psiu*. Each point of the dataset is modeled as an ellipse with $(\phi, \psi)$ as its center and (*phiu*, *psiu*) as its axis (distance to the center in $\phi$ and $\psi$ direction). In the end, the marching squares algorithm implemented in VTK creates the isolines at a distance of 1 to directly draw the ellipses, removing intersecting parts. As the ellipses do not visually clutter the image in any way, they are always shown. An example is provided in Fig. 3.

In the anisotropic case, the distribution of each dihedral angle over all models is shown. As described in Section 4.2, the average dihedral angles, eigenvectors and eigenvalues are used to display average positions and uncertainty of these positions. The average dihedral angles are shown as points on the plot while displaying the eigenvectors and eigenvalues as non-axis aligned ellipses using the eigenvectors for directions and the eigenvalues for scaling of each direction. Similar to the isotropic case, intersections are omitted using a scalar field and the marching squares algorithm. Fig. 9 shows this using the 1G03 dataset.

### 5.3. Uncertainty-aware statistics view

Besides the volume and Ramachandran view, biochemists also need to be able to look at the raw data. Therefore, the statistics view is shown next to the Ramachandran plot and the 3D Visualization, as shown in Fig. 3. It includes values such as amino acid, atom type, and radius. In addition, we provide the amount of uncertainty in each atom, showing a raw view of the values captured for each atom and residue. Here, two different modes are available.

If no amino acid residue is selected, the detail view shows a number of important features of the dataset, like the ID code, a unique identifier of the Protein Database, and uncertainty information. To be able to correlate B-Factors with each other, a reference is needed. Here, the B-Factor range and its average value are being displayed. The same is provided for the averaged B-Factors of the amino acid residues, while it should be noted that this average does not equally consider all atoms, as amino acids have different amounts of atoms in them. For the analysis of dihedral angle uncertainty, their uncertainty range and average are also provided.

When an amino acid residue is being picked by the user, the detail view changes to an elaborate display for the amino acid residue and its atoms. General information like its ID and type of amino acid is provided next to the dihedral angles, including their uncertainties, average B-Factor, Ramachandran region, and secondary structure type. The chain ID and Residue ID might be interesting for identifying certain amino acid residues directly. To give detailed information about the involved atoms, each atom is provided displaying the element, ID, and B-Factor. An example is shown in Fig. 3.

Normalized B-Factors are often used for comparison [46], therefore, normalized values for all B-Factors and dihedral angle uncertainties are provided next to their absolute value in the statistics view. This allows users to get a sense of high and low uncertainty in the viewed dataset.

### 5.4. Interaction

Our presented system is designed such that it consists of multiple views that are highly interconnected and linked.

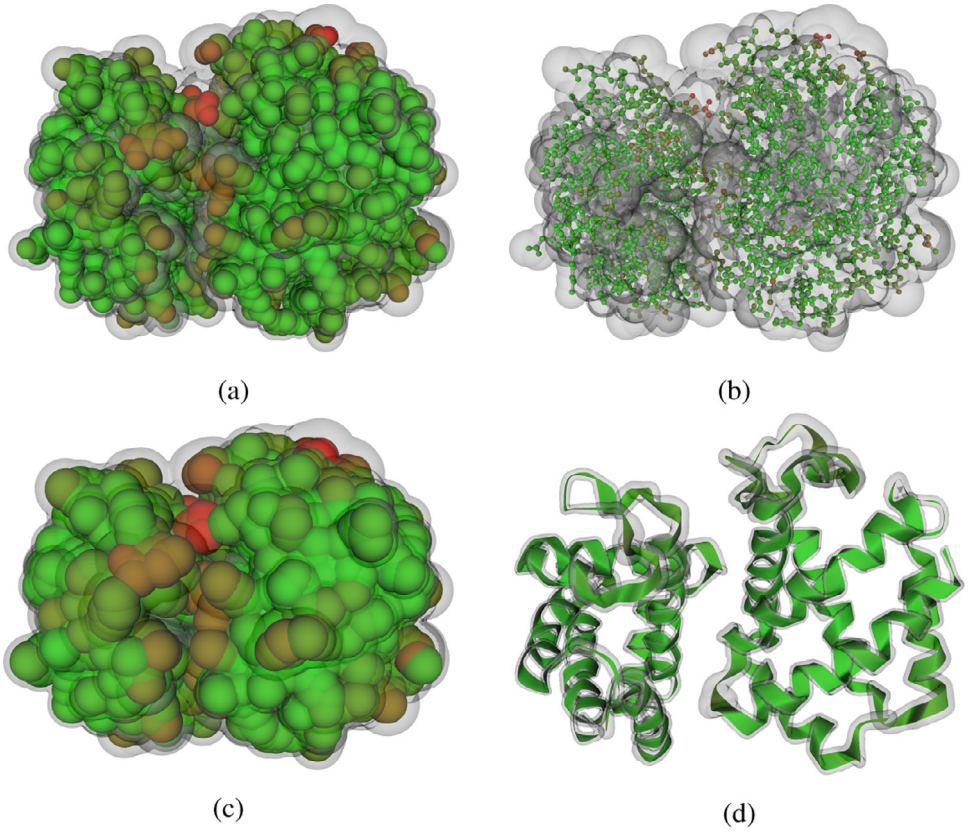


**Fig. 5.** Uncertainty-aware volume visualization of the 1H97 dataset [45], using different geometries for visualization. a) Sphere visualization. b) Ball-and-Stick visualization. c) Solvent Accessible surface visualization. d) Ribbon model visualization.

The interaction between the 3D view, the Ramachandran plot, and the detail view is a point-to-show implementation. An example can be seen in Fig. 3. When hovering the cursor over either an atom in the 3D view or a data point in the Ramachandran plot, the corresponding amino acid residue is being shown in the 3D view along with the Ramachandran plot and the detail view. This enables a fast correlation of different visual aspects of a protein to be examined. The 3D view displays the amino acid residue by drawing its atoms as red spheres whereas the Ramachandran plot highlights the corresponding data point with a red disk around it. All important information is provided by the detail view. When desired, the selection can be fixed to a certain amino acid residue by clicking on it using the right mouse button. Right-clicking again resets the selection and highlights another residue or shows the general dataset information, depending on the mouse cursor position. Through this mechanism, a set of amino acids can be depicted such that they can be examined in their entirety.

The uncertainty hull controls are another important feature. Users are allowed to toggle the hull on and off, use the relative or absolute B-Factors, toggle between the gray and colored hull, and set the transparency of the hull. The transparency is especially important, as surrounding a geometry with any transparent surface always partly blocks the view to some features of the object. In this case, the color-coding of the original geometry might be harder to see. A transparent hull better preserves the features of the original geometry, making it harder to see details of the hull

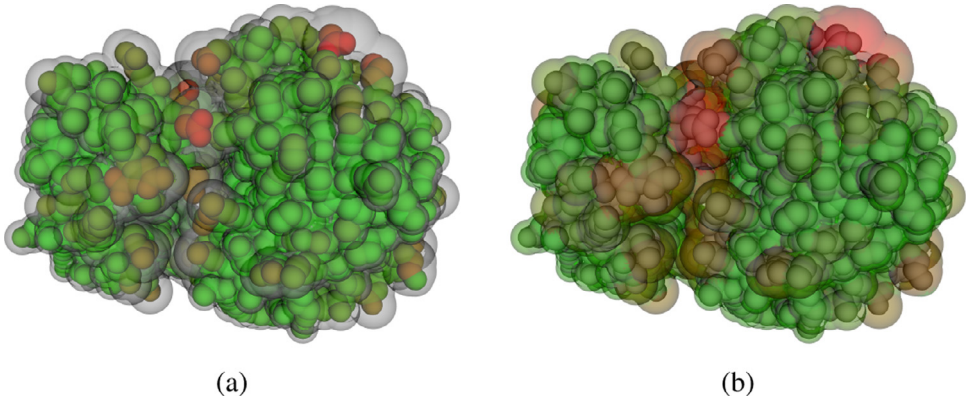


**Fig. 6.** Uncertainty-aware volume visualization of the 1H97 dataset[45] at 35% uncertainty transparency. Gray uncertainty hull with B-Factor color-coded atoms (a) and B-Factor colored uncertainty hull with gray atoms (b).

itself. An opaque hull shows depth information in a clear way but blocks the view to the underlying geometry.

## 6. Results

In the following section, the presented approach is applied to real-world proteins. They have been selected by a domain scientist who considered them interesting in terms of uncertainty analysis. All use cases are known, as this allows to check if the uncertainty visualization indicates the aspects of uncertainty visualization that are relevant in the respective case. The datasets were used in conjunction with our approach and the findings have been discussed with our collaborator. The presented approach was implemented in C++ using the VTK [47] library with the OSPRay[48] rendering backend, using Qt [49] for GUI design.

### *6.1. Monomeric hemoglobin from the trematode paramphistomum epiclitum*

As a first example, we present a small but interesting protein, to obtain a first understanding of the presented visualization approach. Fasciola is a type of fluke, commonly known as liver fluke. It is a parasitic organism that infects the liver tissue of a wide variety of mammals, including humans, in a condition known as fascioliasis. Scientists became very interested in monomeric hemoglobin (1H97 [45]) of this organism as it could be the key to developing a vaccine against the parasite.

We use this protein to provide an overview of the available visualization approaches. The underlying computational model of uncertainty is isotropic as we consider the B-Factors assigned to each atom in the protein originating from the PDB file. Fig. 5 shows a variety of uncertainty-aware volume visualization approaches to represent 1H97. The different types of visualization (sphere visualization, ball-and-stick visualization, solvent accessible surface visualization, and ribbon visualization) are color-coded with the respective B-Factor (green: low B-Factor, red: high B-Factor). In addition, we show the uncertainty hull indicating the potential spatial displacement of atoms. Here, we can see that the surface differs from the original visualization when the B-Factor of nearby atoms is high. This confirms the computational setup of the presented approach.

Fig. 3 shows the uncertainty-aware Ramachandran plot of 1H97 in the middle. We can directly see that the computed dihedral angles are in general very stable according to the spatial movement of captured atom positions. This can be seen by the relatively narrow uncertainty bounds around the visualized points which indicates that the spatial movement of atoms is small. Our collaborator confirmed these findings of our visualization approach. In general, most angles lie in desired areas and the consideration of potential changes in the atom position does not change this impression.

It can be seen that the proposed visualization approach helps to confirm the stability of a protein considering uncertainty information. We also showed throughout this section that a variety of visualization approaches can be adapted using our proposed visual metaphor of uncertainty hulls.

### *6.2. Cyclodextrin glycosyltransferase*

The second example is cyclodextrin glycosyltransferase. The example was chosen by our collaborator as the protein is one protein he is interested in examining the variability of simulation results. This protein is able to produce cyclodextrins from starch, which is an important process in the production of drugs, as it helps transport certain molecules in an efficient manner. In the presented example, the original cyclodextrin glycosyltransferase (1CGT [50]) is compared with cyclodextrin glycosyltransferase that is affected by a mutagenesis (1CGU [51]) in the active site. During the process of mutagenesis, certain amino acids are exchanged, which can affect the function of the protein. As the active site of the protein is the main catalyst of a chemical reaction, this part of the protein needs further examination in terms of stability and uncertainty. The data is provided as an isotropic model of uncertainty in this example.

Fig. 7 a shows the cyclodextrin glycosyltransferase (1CGT) without mutagenesis using the ball view. Color-coding reflects the area in the Ramachandran plot that the respective backbone amino acids are located in. We selected an amino acid residue on the active site to review the thermal stability of the protein. Showing the uncertainty hull of the protein, it can be seen that the hull is displaced equally throughout the three-dimensional space, without larger outliers in the spatial displacement. Fig. 7b displays the uncertainty-aware Ramachandran plot of 1CGT. Here, it can be seen that most amino acid residues are located in desirable areas of the Ramachandran plot. There exist outliers in the lower-left corner, where we figured out that they are not included in the active site of the protein and, therefore, are not of main interest. The uncertainty hull around the remaining amino acid residues shows that although spatial displacement of the proteins can be observed, most amino acid residues will not leave the desired areas in the Ramachandran plot. This helps to determine that the current composition of the protein seems to be stable. Especially when considering the active site of 1CGT, we can see that all residues are located in the dark green areas which are stable.

In contrast to the finding of 1CGT, Fig. 7c shows 1CGU, a cyclodextrin glycosyltransferase that is affected by mutagenesis. The color scheme, uncertainty hull, and performed selection are identical to the ones in Fig. 7c. When reviewing the volume views it can be seen that the uncertainty hull does not change optically, meaning that the spatial displacement of the atoms in the protein is not affected by the mutagenesis. On the other hand, we can directly see that the color-coding of several amino acid residues changed in the volume view. This means that several amino acid residues are now located in another area of the Ramachandran plot in comparison to the original protein. Overall, we can detect more amino acid residues that are located in undesirable areas (light yellow). When reviewing the active site (selection made in red), the selected amino acid residue in the active site is no longer located in a stable region (see Fig. 7d). Although the uncertainty-aware Ramachandran plot represents the spatial displacement of atoms, the selected amino acid residue will not be located in a stable area.

In general, when reviewing the Ramachandran plot of 1CGU, we can identify more amino acid residues that are not located in a stable area of the Ramachandran plot. This becomes more critical when considering the uncertainty spheres around each amino acid residue in the Ramachandran plot. Here, a variety of amino acid residues could leave stable areas when considering the spatial movement of atoms. Our collaborator from the biochemical domain supported these findings and highlighted how the visualization identifying this easily.

Overall, this example shows that 1CGT is a more stable protein (especially in terms of the active site) than 1CGU. Resulting from this, our approach is suitable to understand the stability of proteins in direct comparison while considering the uncertainty included in each of the computational models.

### *6.3. N-terminal domain of the human T-cell leukemia virus capsid protein*

HTLV-I is a virus that binds in the human body and can cause leukemia or neural disorders. An important structure that is involved in this process is the 1G03 protein [52]. The example was chosen by our collaborator as it provides large positional uncertainty that influences the research conducted with this protein. In

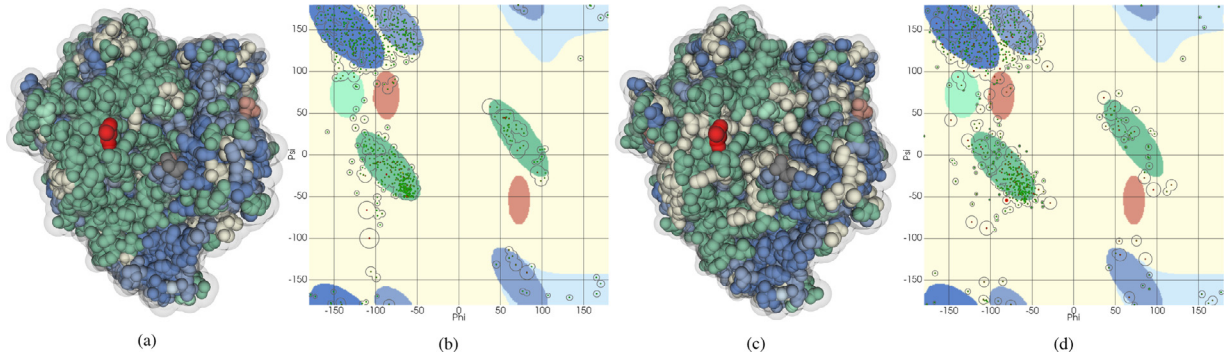


**Fig. 7.** Cyclodectrin glycosyltransferase with and without mutagenesis. a) Uncertainty-aware visualization of 1CGT in ball view style containing the uncertainty hull and a selection of a peptide on the active site. b) Uncertainty-aware Ramachandran plot of 1CGT with selection highlighted according to a). c) Uncertainty-aware visualization of 1CGU in ball view style containing the uncertainty hull and a selection of a peptide on the active site. b) Uncertainty-aware Ramachandran plot of 1CGU with selection highlighted according to a). The selected amino acid residues in both proteins are identical, considering their ID.

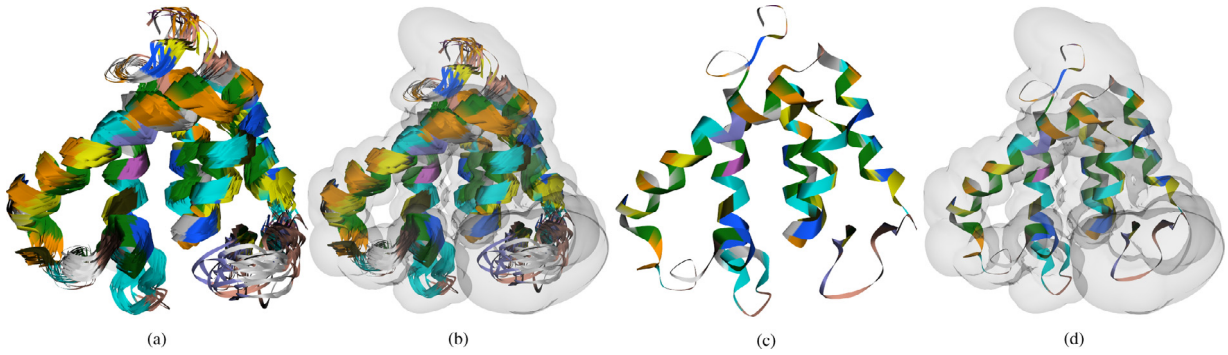


**Fig. 8.** 20 Models of 1G03 protein shown with different styles of visualization. a) Superposition of all models as normally used in the biochemical domain. b) Superimposed model with $3\sigma$ uncertainty hull. c) Average model. d) Average model with $3\sigma$ uncertainty hull.

order to understand the structure and function of the protein, researchers aim to synthesize it. The peptide chain that builds this protein is pretty clear but its three-dimensional folding in space may vary according to a variety of factors such as temperature or other binding proteins. In this context, there exist 20 simulations that try to capture the three-dimensional folding process of 1G03.

Fig. 8 shows different visualization approaches for the 20 models of 1G03. Biochemists usually review these datasets using a superposition visualization. Although all models can be reviewed at the same time, the visualization is very cluttered. Especially inside areas, where the depicted models disagree, it is hard to determine the different models and how they are located in space, as shown in Fig. 8a). Adding the proposed uncertainty hull helps biochemists to examine the potential space where a protein can be located in. Fig. 8 shows the superposition visualization with the $3\sigma$ uncertainty hull. Although we still use the superposition models in this visualization, we can clearly show the user where proteins can be located in space. We allow this visualization in the current framework in order to provide a mechanism to relate the presented visualization approach with already existing approaches. Fig. 8c) shows the computed average model of the 1G03 protein. Here, the visualization is less cluttered as only one model is displayed which is composed of all 20 existing models. This visualization is almost free of visual clutter but reduces the information captured in the 20 models. Based on the average visualization, we provide the final visualization approach that allows us to show the average model in combination with an uncertainty hull (Fig. 8d). Here, it can be seen that the average model is covered by the $3\sigma$ uncertainty hull. The hull helps to indicate areas in the protein that hold high amounts of positional uncertainty. As an example, the top region of the protein holds high amounts of uncertainty which is indicated by the large displacement of the uncertainty hull. Other regions, such as the center of the protein hold a rather tight uncertainty hull, indicating a low variability in the underlying models. This shows that the presented visualization allows for an easy-to-understand representation of disagreement in computed models of proteins.

Fig. 9 shows the top part of the 1G03 protein that was identified to hold large amounts of positional uncertainty. In this example, the visualization of the protein was changed to a ball visualization, indicating the different types of amino acids. Here, we are interested in the amino acid TRP. Fig. 9a) shows that several points are not captured by the uncertainty hull. When increasing $\sigma$ to 2 (Fig. 9b)), most of the existing models are included in the uncertainty hull. For $3\sigma$ Fig. 9c), all models are included in the uncertainty hull. The hull indicates the center of the existing distribution of amino acids and shows their spread in space. Here, it can be directly seen which direction is the most uncertain. Our collaborator confirmed that the visualization correctly indicates this distribution, helping him understand the positional uncertainty.

The example shows that our approach suits understanding the difference in multiple computational models. The number of models is not limited in terms of visual clutter as we are able to reduce the number of visual primitives using an average model.

## 7. Discussion

### 7.1. Check of low-level requirements

As this project was developed in collaboration with domain scientists in the biochemical domain, we continuously included the

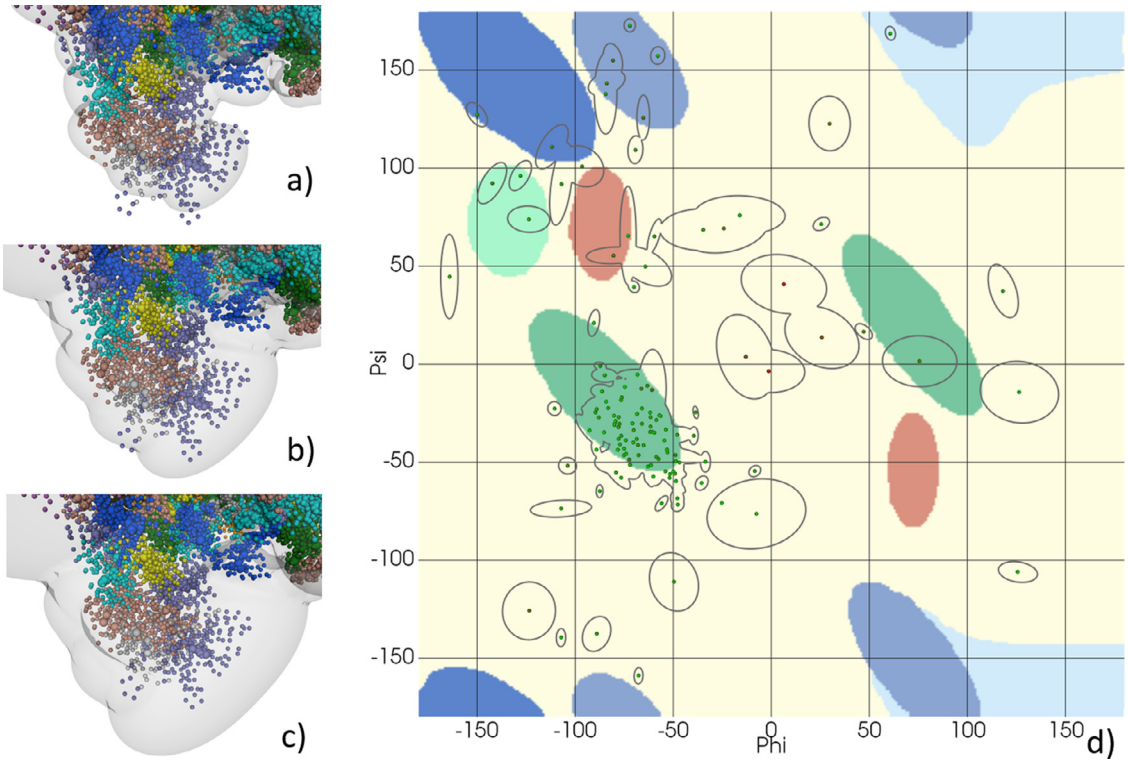


**Fig. 9.** Closeup of region in the 1G03 protein with high amounts of uncertainty. a) $1\sigma$ uncertainty hull. b) $2\sigma$ uncertainty hull. c) $3\sigma$ uncertainty hull. d) Uncertainty-aware Ramanchandran plot.

feedback of the users originating from an informal interview where we presented the current status of our research. The visualization approaches are the final version of this continuous review process which was approved by the domain scientists.

To understand the improvements that the presented approach makes, the list of requirements that assisted in generating the requirements is reused. Here, the experts were asked to rate the presented approach against the 16 requirements formulated by Gillmann et al. [17]. The experts were asked to rate two molecule visualization tools that they frequently use. The visualization expert chose PyMol and Protoshop, and the domain expert used PyMol and VMD. Again, a Likert scale was used for the rating (1 requirement is not fulfilled, 5 requirement is absolutely fulfilled). The results can be found in Table 1. The color-coding shows in which categories, the presented tool was not able to fulfill the requirements such as the remaining tools (red), fulfill the requirements equally good as the remaining tools (yellow), or fulfill the requirements better than the remaining tools. Here, the presented tool shows a clear improvement.

First, the importance of the requirements has been used to create a weighted average to rate each tool. The visualization expert rated PyMol with a weighted average of 3.65, Protoshop with 3.38, and our approach with 4.27 points. Here, our tool outperforms both tools that were known to the user. On the other hand, the domain expert rated PyMol with 2.83, VMD with 2.68, and our tool with 3.42. Again, the presented tool outperforms the used techniques so far. When having a closer look into the ratings of the single requirements, it can be observed that the visualization expert rated 7 requirements with the same points as the known tools. For 9 requirements he rated the presented tool better than the standard tools. In addition, the domain expert found one requirement that we were not able to fulfill as well as the standard tools (compatibility). The expert justified this with the further need to promote the presented tool as an open-source tool such that it can be used in the biochemical community. For nine requirements, our tool was able to perform equally well as the standard tools. Further, 6 requirements are fulfilled better in our tool than by the tools chosen by our collaborators. This is the first indicator that our tool provides an overall improvement for the visualization of protein data under uncertainty.

In addition, our collaborators provided us with very motivating comments on the approach that we summarize below:

- "*I could use the framework right away. It provides me with the most important visualization types that I need*"
- "*I like the easy to interpret visualization of the uncertainty in all views. It does not require a massive amount of time to learn them.*"
- "*I would like to encourage you to include this visualization style in an already existing visualization framework for biochemical data.*"

We would like to take these comments as the first basis of a user evaluation. Especially the last comment on the integration of the tool will form a basis for further development. Here, we aim to gather further user feedback when integrating our approaches in an already existing framework.

### 7.2. Check of high-level requirements

We provided an uncertainty-aware interactive framework for protein visualization. The system was designed by the requirements we agreed on with our domain experts. This work provides an uncertainty-aware description of proteins that is able to represent anisotropic positional uncertainty. In the case where all dimensions hold the same quantification of uncertainty, the model degenerates into an isotropic model. Still, the distributions along one axis are assumed to be equal in both directions. Although this may result in simplifications of atom position distributions that

are not equal in both directions, it allows us to propagate the described uncertainty along computational paths. We achieved this propagation for the computation of $\phi$ and $\psi$ angles within an amino acid residue which reveals important information about the structure of a protein. Although we did not use the propagation for further properties that can be computed based on the atoms of a protein, the general mathematical setup is not restrictive. Further measures such as curvature or quality of surface can be computed in an uncertainty-aware manner within the presented framework (R1).

The described visualization procedures are integrated into state-of-the-art visualization approaches that are used in the biochemical domain. We showed that commonly used visualization methods for proteins can be extended using our method. Although there exist further representations that we did not show in this manuscript, the presented framework allows their inclusion and the mathematical setup does not restrict the underlying visualization approaches. Solvent accessible surfaces and Ramachandran plots, as well as other geometric representations, are part of the presented visualization framework that was designed in a flexible way, such that further visualization schemes can be included if requested. We see this as a strong benefit, as biochemists are able to use the provided visualization methods without requiring a long training phase (R2). This is possible as the well-known protein representations, simple interaction modalities, and well-ordered controls create a familiar environment for users of the biochemical domain.

The uncertainty in most biochemical datasets results in a large number of potential atom configurations. There exist a variety of approaches that aim to superimpose visual representations of these atom configurations resulting in visual clutter. In the presented framework we allow for an average visualization of all potential atom configurations while outlining the potential spatial distortion of atoms. This massively reduces visual clutter while keeping the information of atom movement (R3). Due to validation reasons, we also enabled the framework to show the generated uncertainty hull around all potential models.

An important aspect of a proper visualization approach for the biochemical domain is an interactive visualization framework. Often, multiple views are required to understand the functionality of a protein in its entirety. Here, we provide an uncertainty-aware visual analytics framework that allows biochemists to freely explore protein datasets that are affected by uncertainty (R4). During the development of the presented visualization approaches, we highly focused on achieving a minimal time-consuming computational process in order to avoid waiting times and allow real-time interaction. At this point, we want to highlight that the framework will become available soon, either as a stand-alone solution or as part of an already existing molecular visualization tool.

### 7.3. Further potential applications

Although we describe a specific topic where visualization approaches are applied, we obtained valuable knowledge about the development of uncertainty-aware visualization approaches for geometry-based visualizations in general. The mathematical setup we described is, in general, not restricted to the biochemical domain and therefore we aim to describe potential further applications to create a motivation for further developments.

Our framework could also be used to understand the interaction between proteins. In biochemical applications, surfaces of proteins are compared in order to examine that a ligand is able to bind to a protein. Here, our methodology could be beneficial to examine potential displacements in atom positions, resulting in possible binding sites.

Another application beyond chemistry could be ensemble visualization of geometries in general. This is an important issue in many applications such as industrial manufacturing or path computations in simulations. Here, our visualization approach can be of great benefit in order to examine differences and common grounds of geometries in general.

At last, high-dimensional data analysis is an important topic. In high-dimensional data analysis, data points can be affected by uncertainty as well. The high-dimensional data points are usually simplified using dimension reduction applications in order to be able to review these points. When doing this, the uncertainty of the original points propagates along with the performed computation and needs to be examined in the resulting reduced dataset. Our approach could be of benefit to indicate the uncertainty of the resulting points.

## 8. Conclusion and future work

In this work, we provide an uncertainty-aware interactive framework for protein data. The framework is based on an uncertainty-aware description of proteins that allows capturing variations in the position of atoms due to imprecise measurement or multiple model computations. This uncertainty can be propagated in order to provide uncertainty-aware measures of protein geometries. Based on this theory, we provide an uncertainty-aware framework that allows domain scientists to review protein datasets affected by uncertainty in their working environment using prominent visualization approaches that are extended to indicate uncertainty. The framework is highly interactive to allow for exploration. We successfully tested the presented framework using real-world datasets.

As future work, we aim to provide our proposed visualization setup as open-source code and include them in molecular visualization software such as the PDB visualization tool. We also aim to constantly enlarge the set of included visualization types. Also, a view comparing two atoms or amino acid residues, as well as comparisons of the same atom or residue between models, is planned.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cag.2021.05.011.

## CRediT authorship contribution statement

**Robin G.C. Maack:** Conceptualization, Methodology, Software, Investigation, Resources, Writing - original draft, Visualization. **Michael L. Raymer:** Validation, Investigation, Resources, Writing - original draft. **Thomas Wischgoll:** Validation, Writing - review & editing. **Hans Hagen:** Writing - review & editing. **Christina Gillmann:** Conceptualization, Formal analysis, Investigation, Writing - original draft, Supervision.

## References

[1] Olson AJ. Perspectives on structural molecular biology visualization: from past to present. J Mol Biol 2018;430(21):3997–4012. doi:10.1016/j.jmb.2018.07.009.
[2] Kozlıková B, Krone M, Falk M, Lindow N, Baaden M, Baum D, et al. Visualization of biomolecular structures: state of the art revisited. Comput Graph Forum 2017;36(8):178–204. doi:10.1111/cgf.13072.

[3] Sacha D, Senaratne H, Kwon BC, Ellis G, Keim DA. The role of uncertainty, awareness, and trust in visual analytics. IEEE Trans Vis Comput Graph 2016;22(1):240–9. doi:10.1109/TVCG.2015.2467591.

[4] Buxbaum E. Fundamentals of protein structure and function. 2nd ed. Springer International Publishing Switzerland; 2015. ISBN 978-3-319-19919-1. doi:10.1007/978-3-319-19920-7.

[5] Chung H-K, Braams BJ, Bartschat K, Császár AG, Drake GWF, Kirchner T, et al. Uncertainty estimates for theoretical atomic and molecular data. J Phys D 2016;49(36). doi:10.1088/0022-3727/49/36/363002. Publisher: IOP Publishing

[6] Karshikoff A, Nilsson L, Ladenstein R. Rigidity versus flexibility: the dilemma of understanding protein thermal stability. FEBS J 2015;282(20):3899–917. doi:10.1111/febs.13343.

[7] Deryusheva E, Nemashkalova E, Galloux M, Richard C-A, Eléouët J-F, Kovacs D, et al. Does intrinsic disorder in proteins favor their interaction with lipids? Proteomics 2019;19(6). doi:10.1002/pmic.201800098.

[8] Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, et al. The importance of intrinsic disorder for protein phosphorylation. Nucl Acids Res 2004;32(3):1037–49. doi:10.1093/nar/gkh253.

[9] Iakoucheva LM, Brown CJ, Lawson JD, Obradović Z, Dunker AK. Intrinsic disorder in cell-signaling and cancer-associated proteins. J Mol Biol 2002;323(3):573–84. doi:10.1016/S0022-2836(02)00969-5.

[10] Kozlowski LP, Bujnicki JM. MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. BMC Bioinform 2012;13(1):111. doi:10.1186/1471-2105-13-111.

[11] Meng F, Uversky V, Kurgan L. Computational prediction of intrinsic disorder in proteins. Curr Protoc Protein Sci 2017;88(1). doi:10.1002/cpps.28.

[12] Johnson DE, Xue B, Sickmeier MD, Meng J, Cortese MS, Oldfield CJ, et al. High-throughput characterization of intrinsic disorder in proteins from the protein structure initiative. J Struct Biol 2012;180(1):201–15. doi:10.1016/j.jsb.2012.05.013.

[13] Balasubramaniam D, Komives EA. Hydrogen-exchange mass spectrometry for the study of intrinsic disorder in proteins. Biochim Biophys Acta (BBA) - Proteins Proteom 2013;1834(6):1202–9. doi:10.1016/j.bbapap.2012.10.009.

[14] Al-Karadaghi S.. PDB File Format and Content. 2010. https://proteinstructures.com/structure/protein-databank/.

[15] Green R., Zardecki C.. Guide to Understanding PDB Data. 2019. https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/dealing-with-coordinates.

[16] Lam H, Bertini E, Isenberg P, Plaisant C, Carpendale S. Empirical studies in information visualization: seven scenarios. IEEE Trans Vis Comput Graph 2012;18(9):1520–36. doi:10.1109/TVCG.2011.279.

[17] Gillmann C, Leitte H, Wischgoll T, Hagen H. From theory to usage: requirements for successful visualizations in applications. In: IEEE VIS, creation, curation, critique and conditioning of principles and guidelines in visualization (C4PGV), 5; 2016a. p. 4.

[18] Clifford AA. Multivariate error analysis: a handbook of error propagation and calculation in many-parameter systems. Wiley; 1973. ISBN 978-0-470-16055-8.

[19] Johnson C. Top scientific visualization research problems. IEEE Comput Graph Appl 2004;24(4):13–17. doi:10.1109/MCG.2004.20.

[20] Brodlie K, Allendes Osorio R, Lopes A. A review of uncertainty in data visualization. In: Expanding the frontiers of visual analytics and visualization. London: Springer; 2012. p. 81–109. ISBN 978-1-4471-2804-5. doi:10.1007/978-1-4471-2804-5_6.

[21] Potter K, Rosen P, Johnson CR. From quantification to visualization: a taxonomy of uncertainty visualization approaches. In: Uncertainty quantification in scientific computing. In: IFIP Advances in Information and Communication Technology. Berlin, Heidelberg: Springer; 2012. p. 226–49. ISBN 978-3-642-32677-6. doi:10.1007/978-3-642-32677-6_15.

[22] Dasgupta A, Kosara R. The need for information loss metrics in visualization. In: Workshop on the role of theory in information visualization; 2010. p. 2.

[23] Rheingans P, Joshi S. Visualization of molecules with positional uncertainty. In: Gröller E, Löffelmann H, Ribarsky W, editors. Data visualization '99. Eurographics. Springer; 1999. p. 299–306. doi:10.1007/978-3-7091-6803-5_30. ISBN 978-3-7091-6803-5

[24] Rasheed M, Clement N, Bhowmick A, Bajaj CL. Statistical framework for uncertainty quantification in computational molecular modeling. IEEE/ACM Trans Comput Biol Bioinform 2019;16(4):1154–67. doi:10.1109/TCBB.2017.2771240.

[25] Knoll A, Chan MKY, Lau KC, Liu B, Greeley J, Curtiss L, et al. Uncertainty classification and visualization of molecular interfaces. Int J Uncertain Quantif 2013;3(2). doi:10.1615/Int.J.UncertaintyQuantification.2012003950.

[26] Skånberg R, Falk M, Linares M, Ynnerman A, Hotz I. Tracking internal frames of reference for consistent molecular distribution functions. IEEE Trans Vis Comput Graph 2021:1. doi:10.1109/TVCG.2021.3051632.

[27] Schulz C, Schatz K, Krone M, Braun M, Ertl T, Weiskopf D. Uncertainty visualization for secondary structures of proteins. In: 2018 IEEE pacific visualization symposium (PacificVis). IEEE; 2018. p. 96–105. doi:10.1109/PacificVis.2018.00020. ISSN: 2165-8773

[28] Lee CH, Varshney A. Representing thermal vibrations and uncertainty in molecular surfaces. In: Proc. SPIE 4665. International Society for Optics and Photonics; 2002. p. 80–90. doi:10.1117/12.458813.

[29] Sasisekharan V. Stereochemical criteria for polypeptide and protein structures. In: Collagen. Madras, India: Wiley; 1962. p. 39–78.

[30] Ramakrishnan C, Ramachandran GN. Stereochemical criteria for polypeptide and protein chain conformations: II. Allowed conformations for a pair of peptide units. Biophys J 1965;5(6):909–33. doi:10.1016/S0006-3495(65)86759-5.

[31] Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. J Mol Biol 1963;7:95–9. doi:10.1016/s0022-2836(63)80023-6.

[32] Maack RGC, Hagen H, Gillmann C. Uncertainty-aware Ramachandran plots. In: 2019 IEEE pacific visualization symposium (PacificVis); 2019. p. 227–31. doi:10.1109/PacificVis.2019.00034. ISSN: 2165-8773

[33] Gillmann C, Wischgoll T, Hagen H. Uncertainty-awareness in open source visualization solutions. In: IEEE VIS, vis in practice, 2016; 2016b. p. 5. https://corescholar.libraries.wright.edu/cse/487

[34] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. Nucl Acids Res 2000;28(1):235–42. doi:10.1093/nar/28.1.235.

[35] Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, et al. UCSF ChimeraX: structure visualization for researchers, educators, and developers. Protein Sci 2021;30(1):70–82. doi:10.1002/pro.3943.

[36] Porollo A, Meller J. Versatile annotation and publication quality visualization of protein complexes using POLYVIEW-3D. BMC Bioinform 2007;8(1):316. doi:10.1186/1471-2105-8-316.

[37] Piotr R.. iMol Overview. 2007. https://www.pirx.com/iMol/overview.shtml.

[38] Guex N, Peitsch MC. SWISS-MODEL and the Swiss-Pdb viewer: an environment for comparative protein modeling. Electrophoresis 1997;18(15):2714–23. doi:10.1002/elps.1150181505.

[39] Gillmann C, Wischgoll T, Hamann B, Ahrens J. Modeling and visualization of uncertainty-aware geometry using multi-variate normal distributions. In: 2018 IEEE pacific visualization symposium (PacificVis). IEEE; 2018. p. 106–10. doi:10.1109/PacificVis.2018.00021. ISSN: 2165-8773

[40] Carugo O. How large B-factors can be in protein crystal structures. BMC Bioinform 2018;19(1):9. doi:10.1186/s12859-018-2083-8.

[41] Wang B, Shi W, Miao Z. Confidence analysis of standard deviational ellipse and its extension into higher dimensional euclidean space. PLoS One 2015;10(3). doi:10.1371/journal.pone.0118537.

[42] Furfey PH. A note on Lefever's "standard deviational ellipse". Am J Sociol 1927;33(1):94–8.

[43] Gong J. Clarifying the standard deviational ellipse. Geograph Anal 2002;34(2):155–67. doi:10.1111/j.1538-4632.2002.tb01082.x.

[44] Jones MW. 3D Distance from a Point to a Triangle. Technical Report CSR-5. Department of Computer Science, University of Wales Swansea; 1995.

[45] Pesce A, Dewilde S, Kiger L, Milani M, Ascenzi P, Marden MC, et al. Very high resolution structure of a trematode hemoglobin displaying a TyrB10-TyrE7 heme distal residue pair and high oxygen affinity 11 Edited by K. Nagai. J Mol Biol 2001;309(5):1153–64. doi:10.1006/jmbi.2001.4731.

[46] Yuan Z, Bailey TL, Teasdale RD. Prediction of protein B-factor profiles. Proteins: Structure, Function, and Bioinformatics 2005;58(4):905–12. doi:10.1002/prot.20375.

[47] Schroeder W, Martin K, Lorensen B. The visualization toolkit–an object-oriented approach To 3D graphics. 4th ed. Kitware, Inc.; 2006. ISBN 978-1-930934-19-1.

[48] Wald I, Johnson G, Amstutz J, Brownlee C, Knoll A, Jeffers J, et al. OSPRay - a CPU ray tracing framework for scientific visualization. IEEE Trans Vis Comput Graph 2017;23(1):931–40. doi:10.1109/TVCG.2016.2599041.

[49] The Qt Company. QT. 2020. qt.io.

[50] Klein C, Schulz GE. Structure of cyclodextrin glycosyltransferase refined at 2.0 Å resolution. J Mol Biol 1991;217(4):737–50. doi:10.1016/0022-2836(91)90530-J.

[51] Klein C, Hollender J, Bender H, Schulz GE. Catalytic center of cyclodextrin glycosyltransferase derived from X-ray structure analysis combined with site-directed mutagenesis. Biochemistry 1992;31:7. doi:10.1021/bi00152a009.

[52] Cornilescu CC, Bouamr F, Yao X, Carter C, Tjandra N. Structural analysis of the N-terminal domain of the human T-cell leukemia virus capsid protein 11 Edited by M. F. Summers. J Mol Biol 2001;306(4):783–97. doi:10.1006/jmbi.2000.4395.