Uncertainty Visualization for Secondary Structures of Proteins

Christoph Schulz* Karsten Schatz* Michael Krone* Matthias Braun* Thomas Ertl* Daniel Weiskopf*

Visualization Research Center (VISUS), University of Stuttgart, Germany

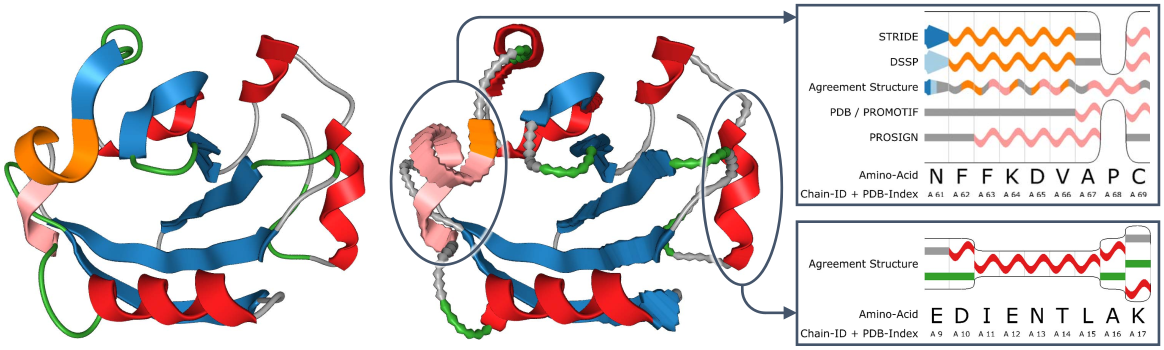


Figure 1: Secondary structure assignment uncertainty of photoactive yellow protein of *E. coli* (PDB ID: 2ZOI). The left image shows a typical ribbon diagram, illustrating the secondary structure assignments computed by STRIDE. The center image shows our uncertainty ribbon diagram with disagreement of four different secondary structure assignments mapped to shape distortion. The sequence diagram cutouts to the right depict assignments for two different sub-chains using two techniques: The upper one reveals an uncertain part of the sequence by stacking the individual results; the lower one shows a consensus α-helix with uncertain ends, where the deviating results at these ends are sorted by increasing uncertainty from bottom to top.

Abstract

We present a technique that conveys the uncertainty in the secondary structure of proteins—an abstraction model based on atomic coordinates. While protein data inherently contains uncertainty due to the acquisition method or the simulation algorithm, we argue that it is also worth investigating uncertainty induced by analysis algorithms that precede visualization. Our technique helps researchers investigate differences between multiple secondary structure assignment methods. We modify established algorithms for fuzzy classification and introduce a discrepancy-based approach to project an ensemble of sequences to a single importance-weighted sequence. In 2D, we depict the aggregated secondary structure assignments based on the per-residue deviation in a collapsible sequence diagram. In 3D, we extend the ribbon diagram using visual variables such as transparency, wave form, frequency, or amplitude to facilitate qualitative analysis of uncertainty. We evaluated the effectiveness and acceptance of our technique through expert reviews using two example applications: the combined assignment against established algorithms and time-dependent structural changes originating from simulated protein dynamics.

Index Terms: Human-centered computing—Visualization—Visualization application domains—Scientific visualization; Applied computing—Life and medical sciences—Computational biology—Molecular sequence analysis

1 Introduction

The representation of uncertainty has been identified as one of the top scientific visualization research problems by Johnson [21]. In some application areas, like medical visualization [39] or weather forecasting [41], uncertainty is nowadays often illustrated to enable a more informed visual analysis. Molecular visualization, however, rarely incorporates uncertainty, although it is present in the data.

Proteins are macromolecules found in all organisms, consisting of one or more chains of amino acids. They perform a wide range of functions from catalyzing chemical reactions to transporting other molecules. Due to their importance in biology, a vast number of analysis and visualization techniques have been developed. A detailed overview of visualization methods for biomolecules has been presented recently by Kozlíková et al. [26]. Therefore, we only give a brief introduction to the different levels of protein abstraction and visualization: The primary structure is a direct representation of amino acids and commonly visualized as a sequence diagram. The secondary structure represents local structures stabilized by hydrogen bonds and can be inferred algorithmically from the atomic coordinates and primary structure. Although secondary structure assignment from atomic coordinates has a long history, there is no absolute definition and therefore no ground truth. Please note that the folding of amino acid chains indicates stability and function. The tertiary structure is the spatial arrangement of the secondary structure, which directly relates to the atomic positions. Established visualization techniques to illustrate secondary and tertiary structures are sequence diagrams and ribbon diagram (see Figure 1).

The secondary structure assignment for a protein can vary not only due to conformational changes but also due to the used assignment method. It is easy to overrate changing secondary structure elements during conformational changes, since the transition between the graphical depiction of two different secondary structure elements is not smooth but rather a hard switch. This can lead to selective perception that impedes analysis. Our goal is to gain trust in the secondary structure assignment and ease its interpretation by depicting uncertainties that might arise from different sources, e.g., raw input data, simulation, or algorithmic inference. Thus, we propose a technique to convey uncertainty in protein data in 2D as well as in 3D. To this end, we extend the established secondary structure representations for uncertainty propagation and visualization.

Our contribution can be summarized as follows: We propose a model to convey the confidence per structure element and amino acid

*e-mail: {firstname}.{lastname}@visus.uni-stuttgart.de

2165-8773/18/$31.00 ©2018 IEEE
DOI 10.1109/PacificVis.2018.00020

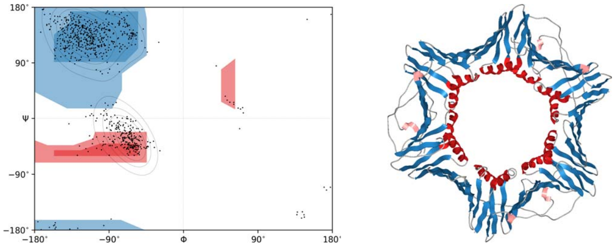


Figure 2: A Ramachandran plot (left) of human PCNA (PDB ID: `1AXC`) and its ribbon diagram (right) for reference. The colored template on the Ramachandran plot shows where the secondary structure elements are expected.

Table 1: Secondary structure elements, letters, and colors by assignment method.

			PDB	DSSP	STRIDE	PROSIGN
π-Helix	I	∿	✓	✓	✓	✓
α-Helix	H	∿	✓	✓	✓	✓
3_{10}-Helix	G	∿	✓	✓	✓	✓
Turn	T	▬		✓	✓	
Bend	S	▬		✓		
β-Bridge	B	➤		✓	✓	
β-Strand	E	➤	✓	✓	✓	✓
Coil	C	▬	✓	✓	✓	✓

for multiple assignment methods. Moreover, we convey uncertainty using various visual variables in 2D sequence diagrams and 3D ribbon diagrams. Uncertainty visualization is error-prone and difficult to evaluate, as noted by Hullmann [17]. Hence, we explore the design space for visualization using various visual variables to represent uncertainty. Finally, we conducted expert reviews to showcase the applicability of our technique using two scenarios.

2 Structural Biology Background

Proteins are macromolecules that consist of one or more chains of amino acids. They are crucial for all life-forms on earth since almost all organic processes rely on proteins to work properly. The correct folding of the amino acid chains determines the function of the corresponding protein. Therefore, the spatial arrangement of the atoms is of high importance. Misfolded proteins may not only be functionless but can also cause diseases like Creutzfeld-Jakob.

The protein secondary structure describes the local spatial arrangement of the amino acid chain [31]. The amino acid chain can form helical structures or (anti-)parallel sub segments of the main chain—mainly held together by hydrogen bonds. It is notable that there exists *no ground truth* for protein secondary structures, even the IUPAC rules [20] offer more than one possible approach for assignment. In particular, helical segments can be defined by symmetry and hydrogen bond arrangement or by special dihedral angle configurations in the amino acid backbone. The latter can be visualized using a so-called Ramachandran plot [35], which is a scatter plot where two dihedral angles, usually Φ and Ψ, are plotted against each other (cf. Figure 2). In this plot, each structure element occupies a certain (not clearly defined) area in the target domain, allowing for a fast structure assignment per amino acid.

We divide the secondary structure assignment methods into two categories: The first one infers secondary structure based on the tertiary structure (atomic coordinates). The second one only takes the main-chain amino acids into account and is mainly used for structure prediction of unknown or newly designed proteins. Please note that the uncertainty present in the secondary structure propagates to atom positions if it is used to infer the tertiary structure. However, such positional uncertainty would be beyond the scope of our work, which is based on the assumption that the atomic coordinates are known.

All assignment methods employed in this work use up to eight possible structure elements for assignment (cf. Table 1). During the process of structure assignment, one of these elements is assigned to every amino acid of the polypeptide chain: Helical structures with hydrogen bonds between amino acid n and amino acid $n+5$, $n+4$, or $n+3$ are considered a *π-Helix*, *α-Helix*, or *3_{10}-Helix*, respectively. A *Turn* is a sharp direction change of the main-chain, caused by a single hydrogen bond. A *Bend* is an area of the main-chain with high curvature, independent of any hydrogen bond. A *β-Bridge* is characterized by two hydrogen bonds between parallel or antiparallel segments of the main-chain. A *β-Strand* is applied when two consecutive amino acids fulfill the prerequisites to be the part of the same kind of β-Bridge. *Coil* is the residual class, i.e., if none of the other structure elements are applicable. Please note the overlaps between these classes, e.g., β-Strands consist of multiple β-Bridges of the same kind and helices are several turns in a row. The specific type of the helix is determined by the main-chain distance of the atoms connected by the involved hydrogen bonds of the turns.

One application of our technique is the assessment of different secondary structure assignment algorithms. The interpretation requires a basic understanding of these assignment methods. Therefore, we provide a brief introduction to explain the differences regarding definition and support in Table 1.

DSSP [22] infers hydrogen bonds between atoms of the protein and calculates the locations of turns and bridges using these approximated hydrogen bonds. These turns and bridges serve as base patterns to compute all other possible structure elements such as helices and strands. We use the rewritten variant of DSSP that was adapted to detect π-Helices more reliably [45].

STRIDE [10] also estimates hydrogen bonds to detect secondary structure elements. Just like DSSP, β-Strands are also inferred from β-Bridges. The detection of helices follows a different approach: Helices are determined based on the torsional angles of the polypeptide chain. The results usually differ slightly from DSSP because of the differences regarding detection of helices and hydrogen bonds.

PROMOTIF [19] is no single secondary structure estimation algorithm but a whole software suite that can detect larger motifs than a simple secondary structure detection could. For β-Strands and helices, they use a modified version of DSSP. Since DSSP tends to underestimate the length of strands and helices compared to the IUPAC recommendations [20], PROMOTIF adds one extra residue at each end of the secondary structure motifs. Hence, we expect higher uncertainty values toward the end of structure motifs.

PROSIGN [16] works in a purely geometrical manner, using only the C_α atoms of the amino acids as input. It computes the difference from a geometrically perfect helix or β-Strand. If the difference is below a certain threshold, the respective structure element is assigned. This simple heuristic significantly reduces the computation time. We chose to incorporate PROSIGN into our comparison especially because of its drastically different approach.

We mainly use proteins from the RCSB Protein Data Bank (PDB) [2], which offers secondary structure annotations. The annotations for DSSP, STRIDE, and PROSIGN were computed directly from the atomic coordinates, whereas the results of PROMOTIF were downloaded from PDB.

3 Related Work

Secondary Structure As mentioned in Section 2, we visualize uncertainties derived from the results of different secondary structure assignments. Since the algorithms use different assumptions to define their underlying model, their output may deviate significantly. Thus, the evaluation of different secondary structure assignment methods is still an area of active research in biochemistry. For example, Martin et al. [33] combined the results of different algorithms, such as STRIDE and DSSP. However, in contrast to our work, their intention was to stabilize the secondary structure assignment instead of visualizing the results of the individual algorithms and their internals. Similarly, Rocha [40] has recently targeted this problem by comparing the outputs of DSSP, PROMOTIF, and STRIDE. The goal of this work was also to create a new, combined assignment method for better understanding of the underlying molecular mechanisms including physicochemical interactions.

The importance of investigating the results of existing secondary structure assignment methods is also reflected by the work of Touw et al. [45]. They revised the DSSP algorithm [22] to improve the detection of π-Helices. Therefore, showing the internal threshold values of the methods can help users assess the reliability of the assignment results. This was one of the motivations for our technique.

Visualization of Secondary Structure and Sequences Our uncertainty visualization is based on established representations of the secondary structure of the protein. This secondary structure is commonly visualized using *ribbon diagrams*. In general, the basis for the ribbon diagram is a spline that follows the backbone of the protein's amino acid chain, e.g., by using the C_α atoms as control points. Guidelines for the visualization of secondary structures were given by Richardson [38], who proposed using ribbons and ribbon-shaped arrows to depict helices and sheets. The orientation of the ribbon shows the direction of the hydrogen bonds that stabilize the structure. Coil and turn are shown as simple lines, thin ribbons, or tubes. While Richardson mainly worked on hand-drawn images, Carson [4] presented one of the first interactive renderings of ribbon diagrams—using cubic B-splines for the underlying spline curve. Most molecular visualization tools still adhere to Richardson's guidelines, for example PyMol [43] or VMD [18]. However, some of these tools can render even more abstracted visualizations. For example, VMD offers the option to depict helices as thick, straight cylinders encompassing the whole helix or as thick tubes that follow the centerline of the helix (*Bendix* representation [9]). However, the classical depiction using ribbons is prevalent and often preferred by users, since the ribbons provide additional details for the analysis.

Traditionally, triangulated models of the ribbon diagram are precomputed on the CPU for rendering. Starting with the geometry shader implementation by Krone et al. [27], GPU-accelerated rendering approaches that compute the triangulated ribbon diagram on the fly entirely on the GPU became popular. Wahle and Birmans [47] presented another GPU-accelerated approach that does not need geometry shaders because these were comparably slow on the first generations of GPU that supported them. Our ribbon diagram implementation is similar to the one recently presented by Hermosilla et al. [14], which uses the tessellation shader.

The secondary structure of a protein and its spatial arrangement, the tertiary structure, play an important role for its function. Kocincová et al. [25] have recently presented a comparative visualization for the secondary structure of multiple proteins. Their multi-view application includes a classical sequence diagram, a 3D view showing the superimposed ribbon diagrams of all proteins, and a novel 2D representation that highlights deviations in the alignment of the secondary structures of different proteins. Therefore, the focus of this work is different from our goal, since we do not want to show structural differences but uncertainty values.

Visualization of Uncertainty Uncertainty is inherently present in most data and models in scientific visualization, and thus it is not surprising that it received broad attention [3, 34]. MacEachren et al. [32] studied various visual variables for uncertainty regarding intuitiveness and accuracy. Some visual variables were further refined by Gschwandtner et al. [12] for time series, which are closely related to our sequence diagrams. Correll and Gleicher [8] showed the deficiencies of classical error bars through several crowd sourcing experiments and promoted the use of gradient plots or violin plots. Hullmann [17] found that study design for uncertainty visualization seems to be a major source of uncertainty itself. While we took inspiration on visual variables from these studies, we were cautious not to take their conclusions for granted for this very reason. As opposed to previous approaches, we encode uncertainty into geometry and screen door transparency. Khlebnikov et al. [23] used a similar approach for screen door transparency based on noise in volume rendering. However, to our knowledge, this approach has not been applied to uncertainty visualization. Grigoryan and Rheingans [11] visualized surface uncertainty by displacing surface points in normal direction proportionally to the uncertainty. We also use geometric distortion, however, in contrast to their work, we do not encode a purely geometric uncertainty. Coninx et al. [7] employed noise to encode uncertainty on surfaces. They also apply color to illustrate uncertainty, which is infeasible in our case, since color is typically used to convey physicochemical properties of the molecules.

Uncertainty in Molecular Visualization In other fields like medical visualization or visual analysis of gene expression, the illustration of uncertainty has been widely addressed to provide reliable depictions and to further the trust in the visual analysis (see, e.g., Ristovski et al. [39] or Holzhüter et al. [15]). In molecular visualization, however, the depiction of uncertainty is mainly focused on the location of atoms. Molecular visualization tools can usually color-code B-factors, which expresses x-ray scattering due to the thermal motion, onto standard molecular models. Rheingans and Joshi [37] proposed various, tailored depictions of positional uncertainty for molecular data. They rendered different superimposed conformations semi-transparently or accumulated Gaussian splats of the atoms on a 3D grid that is then rendered using direct volume rendering or as isosurfaces. Similarly, Schmidt-Ehrenberg et al. [42] employed a 3D grid and isosurfaces to illustrate the uncertainty derived from the dynamics of small molecules. Lee and Varshney [30] used layers of semi-transparent molecular surfaces to show positional uncertainty due to thermal vibrations, resulting in a fuzzier appearance for uncertain regions. Krone et al. [28] followed a similar approach to show the spatial probability density of atoms for a complete simulation, thus providing a qualitative analysis of the protein flexibility. Knoll et al. [24] also accumulated the electron density distribution for each atom to visualize the uncertainty of molecular interfaces using volume rendering. Rasheed et al. [36] proposed a statistical framework for the quantification of structural uncertainties in molecular data. They used established visual clues for structural uncertainty like arrow glyphs, superimposition of structures, or volume rendering. Furthermore, they extracted and compared probabilistic binding sites based on the geometric fit of the ligand and showed the uncertainties as colors on the molecular surface. In contrast to our work, none of these methods uses the ribbon diagram as basis for uncertainty visualization. Furthermore, none includes uncertainties derived from feature extraction algorithms, just from the input data, targeting almost exclusively positional uncertainty.

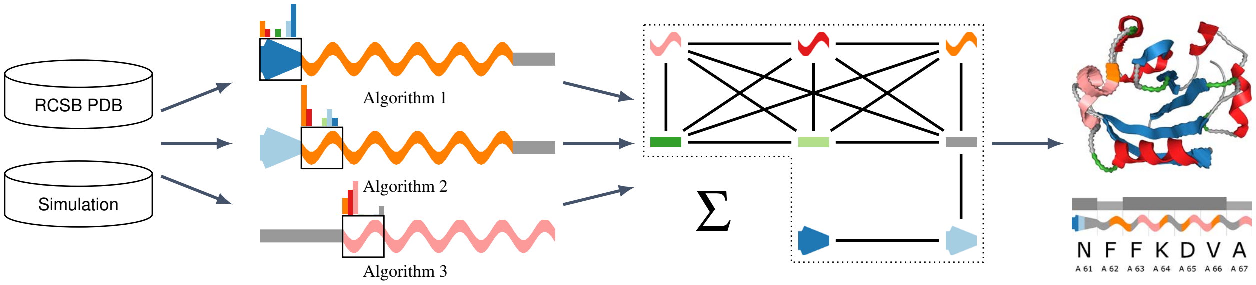
4 Requirements

We specified an initial set of abstract requirements for our uncertainty visualization and gradually refined them while iterating on the design and implementation of our prototype. Each of them is reasonably high-level and has a justified usecase:

R1 The assignment methods should be comparable with each other and their entirety on amino acid level, to analyze differences between the assignments at various levels of detail.

R2 The uncertainty model should be flexible with regard to aggregation because the importance of each source of uncertainty is unknown in advance. Specifically, it should be possible to propagate and weight various algorithm internals for inspection convenience.

R3 Dynamics are an additional source of uncertainty because the secondary structure of a protein can vary over time. Therefore, the



(a) Protein Data (b) Secondary Structure Computation (c) Time and Discrepancy Aggregation (d) Visualization

Figure 3: Overview of our uncertainty visualization technique. We start with protein data from RCSB or a simulation, i.e., atomic coordinates (a). By computing the secondary structure, we obtain a distribution of possible secondary structure elements per amino acid (b). Then, we aggregate across different sources of uncertainty such as time and discrepancy of multiple assignment methods (c). Finally, we render ribbon and sequence diagrams, depicting the combined distribution (d).

model should also enable us to aggregate temporal changes in a single uncertainty value per amino acid.

R4 The model should be visualized in 2D on a per-amino-acid level to assess sources of uncertainty quantitatively without influences from 3D, such as occlusion.

R5 The model should be visualized in 3D as ribbon diagram to assess sources of uncertainty qualitatively, i.e., understand if folding contributes to uncertainty.

5 Uncertainty Model

In this section, we present our model for determining the uncertainty in secondary structure (cf. Figure 3). First, we had to identify various sources of uncertainty to define our model: The assignment methods described in Section 2 apply different definitions to the sequence of amino acids and their atom positions, i.e., there is no consistent definition of secondary structure criteria across assignment methods. Thus, the primary source of uncertainty is the lack of ground truth. Consequently, we perform relative comparisons between the assignment methods and assess them in the context of their definitions. Other sources of uncertainty are, e.g., temperature factors for each atom or molecular dynamics, i.e., conformation changes over time or physically motivated quantities such as flexibility.

Our main goal was to support a wide range of applications as well as qualitative and quantitative estimation of uncertainty together. The differences regarding discrimination of secondary structures across assignment methods make direct visual comparisons impractical (cf. Table 1).

5.1 Classification

As per requirement R1, our model must allow inspection at amino acid level. Let $\mathbb{M}$ denote the set of assignment methods, $\mathbb{E}$ the set of possible secondary structure elements, and $\mathbb{A}$ the sequence tuple of amino acids. Generally, we have to assume that each assignment method $m \in \mathbb{M}$ is a black box. Therefore, we only know that a secondary structure element $e \in \mathbb{E}$ is assigned depending on the amino acid $a_i \in \mathbb{A}$, leading to the definition of a hard classification label l_m to express the result of an assignment method:

$$l_m(e,a_i) = \begin{cases} 1, & \text{if } e \text{ is output from } m \text{ for } a_i \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

This formulation, however, only represents the final output of the algorithms and conveys no knowledge about the certainty of the respective assignments. To characterize the uncertainty, we infer soft classification labels from internal thresholds and values of PROSIGN and STRIDE. These two algorithms were chosen as examples. Similar changes could be made to the other algorithms (DSSP, PROMOTIF) based on their internals. The following description is kept high-level since the source code is part of the supplemental material.

The construction of soft labels for PROSIGN is easy because all internal values share the same spatial domain. The algorithm computes distances d_s for each amino acid to ideal geometric representations, before classification using threshold t_s. Because some of these distances are outliers, mostly at the beginning or at the end of the protein, we have to clamp them using $\max(0, \min(d_s, 2t_s))$ which seems reasonable judging from a large set of test proteins. To obtain our soft labels, we simply normalize the clamped distances.

Our strategy for STRIDE is different because of its more complex multi-level and parser-like approach to classification. The algorithm scans the amino acids a_i using a moving window $a_i, \ldots, a_{i+5}$. If a helix is likely, the algorithm classifies using three thresholds for a_i (beginning), $a_{i+1,\ldots,i+4}$ (main part), and a_{i+5} (end). If a bridge or strand is likely, the algorithm classifies using two mutually exclusive thresholds: one for parallel and one for anti-parallel. The authors of STRIDE chose the thresholds so that the classification matches a set of representative reference proteins correctly between 92.6% to 94.9%, depending on the secondary structure type [10]. Based on these distributions, we fit Gauss functions for each threshold to obtain probabilities. Depending on the applied thresholds, we then multiply the probabilities, only storing the maximum resulting probability for each amino acid as the window moves. These probabilities are ad-hoc interpreted as soft labels.

Another property of these soft labels is that we can easily aggregate them over time, for example, to create a static representation of multiple simulation steps.

5.2 Discrepancy

To satisfy requirement R2, our uncertainty model had to be flexible regarding comparison of multiple structure element assignments per amino acid. Therefore, we define a graph $G_{m,\acute{m}}$ for each pair of assignment methods $m, \acute{m}$ to quantify the pair-wise discrepancies between structure element assignments (cf. Figure 3c). Each edge weight models a discrepancy, i.e., a (relative) mutual deviation between two structure assignments. These edge weights are user-definable and not every edge is meaningful for biological and geometric reasons.

In order to compare assignment methods m and $\acute{m}$ as well as their structure assignments e and $\acute{e}$, we define a discrepancy matrix $D^{e,\acute{e}}_{m,\acute{m}}$ based on graph $G_{m,\acute{m}}$: First, we fill in the user-defined edge weights, where provided. Then, we complete the discrepancy matrix by computing the missing entries using shortest paths in $G_{m,\acute{m}}$. If there is no user-defined discrepancy (e.g., due to lack of knowledge), all entries in $D^{e,\acute{e}}_{m,\acute{m}}$ would be 1. The discrepancy matrices used for our experiment were based on experience and knowledge about the assignment methods involved.

Using one discrepancy matrix for each pair of assignment methods, we can compute an ad-hoc confidence score c to combine labels using discrepancies:

$$c(e; a_i) = \sum_{m\in\mathbb{M}} \sum_{\acute{m}\in\{\mathbb{M}\setminus m\}} \sum_{\acute{e}\in\mathbb{E}} \frac{l_m(e,a_i)\cdot l_{\acute{m}}(\acute{e},a_i)}{\left(1+\max\limits_{\tilde{e},\tilde{\tilde{e}}\in\mathbb{E}}\left(D^{\tilde{e},\tilde{\tilde{e}}}_{m,\acute{m}}\right) - D^{e,\acute{e}}_{m,\acute{m}}\right)^2} \qquad (2)$$

Evenly distributed secondary structure elements e imply disagreement, whereas a single peak indicates high agreement among the assignment methods. Therefore, we express uncertainty as negation of the derived standard deviation σ_c, which is calculated from the confidence score over all possible structure elements e for one amino acid a_i using $u = 1 - \sigma_c(a_i)/\max\limits_{a_j\in\mathbb{A}}\left(\sigma_c(a_j)\right)$.

6 Uncertainty Visualization

In this section, we describe the different visualization designs to depict the previously computed uncertainty values per amino acid. We use 2D and 3D visualizations because some information can be shown better in 3D than in 2D and vice versa. Both visualizations build up on commonly used representations for protein secondary structure, namely *sequence diagrams* and *ribbon diagrams*. The 2D view extends the classical sequence diagram by rendering the results of the different assignment methods in combination with the computed uncertainty values. The 3D view combines the uncertainty with the positional information of the tertiary structure, thus conveying the influence of the 3D structure of the protein on the certainty of the assignment. The 2D view is favorable for analyzing abstract data such as the absolute numerical uncertainty or threshold values, whereas the 3D view gives a qualitative overview of the spatial embedding and context.

Both views are based on well-established geometric representations of protein secondary structure elements (cf. ribbons, arrows, and tubes in Section 3). To ease the interpretability by domain scientists, we adhere to community standards regarding visual styles of protein secondary structures: Color is typically used to convey physicochemical properties or to highlight the different secondary structure elements shown in Table 1. Value and saturation interfere with 3D scene lighting. Hence, there are only a few visual variables left that would not interfere with this standard. Therefore, we decided to use different approaches that change the geometry.

6.1 Sequence Rendering

A sequence diagram typically consists of multiple rows, each depicting different aspects of the visualized amino acid chain. Each column represents one amino acid. The bottom row of the diagram usually shows the one-letter code that denotes the corresponding amino acid and the residue identifiers within the chain. Additional information like binding sites or secondary structure can be shown in the other rows on top. An example of a sequence diagram following the abovementioned community standards is shown in Figure 4.

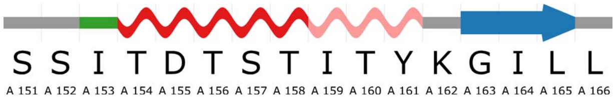
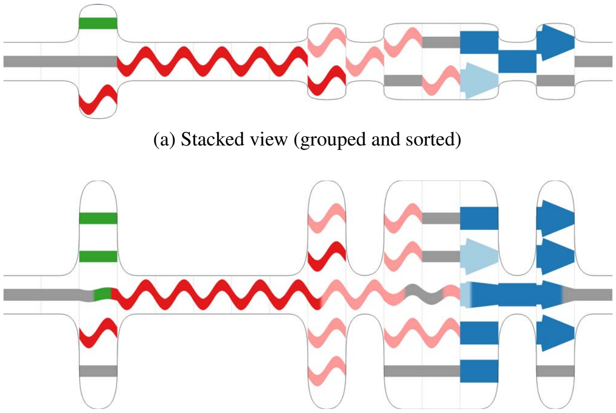

S S I T D T S T I T Y K G I L L
A 151 A 152 A 153 A 154 A 155 A 156 A 157 A 158 A 159 A 160 A 161 A 162 A 163 A 164 A 165 A 166

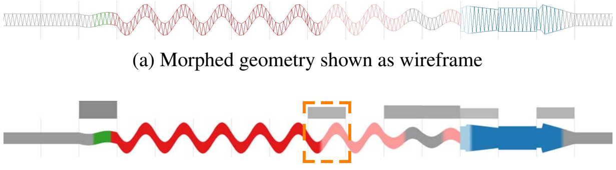
Figure 4: Typical sequence diagram consisting of a sequence row and an amino acid row. Helices are rendered as sine waves, strands as arrows. Colors and visual primitives are detailed in Table 1.



(a) Stacked view (grouped and sorted)

(b) Stacked view (projected)

Figure 5: Stacked assignments. The grouped and sorted version (a) scales better than the projected version (b).



(a) Morphed geometry shown as wireframe

(b) Histogram view rendered above the morphed geometry

Figure 6: Multiple assignments combined into one sequence view: (a) morphed wireframe geometry, (b) final histogram view with uncertainty values encoded as gray bars above the sequence. High, dark gray bars denote high uncertainty, whereas low, lightly colored bars represent less uncertainty and more agreement. Where the uncertainty value reaches 0.0, the rendered structure equals the structure in the normal sequence views (cf. Figure 4).

We decided to follow the design of RCSB PDB [2] by stacking possible secondary structure assignments on top of each other to allow for convenient comparison. As an extension of the naïve approach in the RCSB, we merge rows upon agreement to facilitate comparison of multiple assignments (cf. Figure 5). The user can either merge rows by grouping and sorting by confidence score in descending order (Figure 5a), or by using projection (Figure 5b). Stacking facilitates comparison, however, it requires more than one row to encode multiple secondary structure elements. As described in Section 5, our uncertainty model allows us to project multiple assignment methods into a distribution per amino acid. The projection results in a single static sequence representation (cf. Figure 5b, middle row).

We render the projected row by morphing structure geometries during tessellation. First, a quad per amino acid is tesselated as shown in Figure 6a. The vertices have to be shifted only in horizontal direction to obtain the different secondary structure elements shown in Table 1. Thus, geometry morphing boils down to linear interpolation between all possible structure elements of the current amino acid. The tessellation factor must be set so that the appearance of more complicated structure elements such as helices is smooth.

As shown in Figure 6b, we add gray bars in an additional row of the sequence diagram to display the computed *uncertainty values*, which eases the readability and highlights amino acids with uncertain structure assignment. Furthermore, we use color to differentiate between secondary structure elements with ambiguous geometric

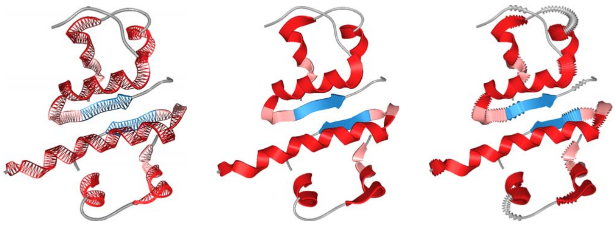

(a) Wireframe (b) Regular (c) Distorted

Figure 7: Ribbon diagram of a protein. The original ribbon diagram shown in (a) and (b) is distorted based on the computed uncertainty value, resulting in (c).

representations, here shown by means of the α-Helix and 3_{10}-Helix (dashed orange box in Figure 6b). The color coding of an amino acid is inspired by individual bars of horizontally stacked bar charts. To smoothen the transition between segments, we interpolate the colors in a very narrow region around the transition. The area of each segment is proportional to the confidence score of the secondary structure element, which allows for an easy detection of the most likely element. To reduce the number of color transitions between amino acids, we order the segments such that the last color of amino acid n is used again in amino acid $n+1$ if the same element was assigned by one of the algorithms here. This sorting provides a fast heuristic for decreasing the perceived visual clutter originating from color transitions. Note that this optimization dissolves the relation to assignment methods, i.e., the order of elements in an amino acid box no longer corresponds to the order of assignment methods. However, the goal of the histogram view is to give an overview of the most likely assignment.

6.2 Ribbon Diagram

Ribbon diagrams, also called *Cartoon representation*, typically consist of a schematic representation of each structure element fitted to a curve using interpolation (Figure 7b; cf. Section 3). They share many geometric motifs with sequence diagrams and allow the user to gain knowledge about the secondary structure and spatial folding of a protein, which is essential for its function. Our renderer implements the tessellation-shader-based method proposed by Hermosilla et al. [14] with several modifications: First, we represent each amino acid using a spline segment consisting of a single tessellated quad instead of two quads as in the original method (Figure 7a). Second, we made several changes for uncertainty visualization. Third, we render contours to improve depth perception and visual separation of close but different parts of the protein chain [44]. We achieve contour rendering by rendering an inflated version of the backsides of all faces before the actual image. The inflated model is the result of a slight displacement of each vertex v in the direction of the vertex normal $\vec{n}_v$ — a higher displacement leads to a thicker contour.

We cannot just transfer the design from sequence diagrams to ribbon diagrams because the space for additional rows is already occupied in 3D. MacEachren et al. [32] evaluated the influence of visual variables on the understandability of uncertainty visualizations. However, the study only considered 2D visualizations and enclosed visualizations. Especially when using a third dimension, the most intuitive visual variable, *fuzziness*, does not perform well because the occlusion caused by the protein folding introduces some level of ambiguity. The variable *position* is already occupied by the tertiary structure, whereas *color* is occupied by secondary structure assignment or physicochemical property. Furthermore, we have refrained from interpolating between the structure assignment colors, since the blending of two or more colors cannot be unambiguously reversed, as discussed by Chuang et al. [6].

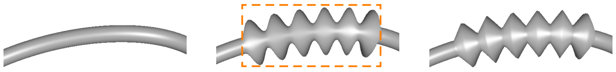

(a) No distortion (b) Sine (c) Triangle

Figure 8: Comparison of the different geometry distortion methods. Compared to the sine distortion (b), the triangle distortion (c) needs less geometry to work properly. The dashed orange box illustrates one amino acid (amplitude $k_a = 1$, frequency $k_f = 6$). The specular highlights of the triangle waveform emphasize the shape more clearly.

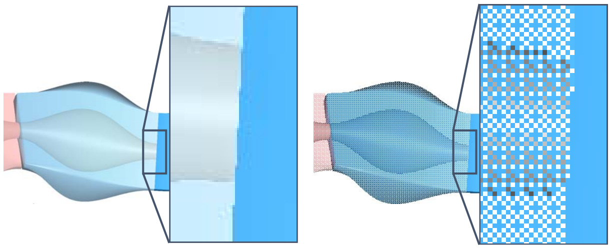


Figure 9: Comparison of alpha blending (left) and screen door transparency (right). From a distance, there is almost no visible difference. In close-up, the colors from Table 1 are visible (i.e., no color mixing).

Geometry Distortion We chose the visual variables *shape* and *grain* [32], i.e., we add *waviness* to the geometric representation. Although curvature can impede the faithful perception of spatial frequencies, we argue that our choice is sufficient for qualitatively highlighting the uncertainty. By distorting the geometry of a standard ribbon diagram uniformly in all directions using a periodic waveform, it is possible to encode an additional value such as uncertainty (Figure 7c). The image footprint and the overall visual appearance of the ribbon diagram do not change significantly. In particular, the intra-model occlusion is not much increased. If the uncertainty value is low, the appearance resembles the original ribbon diagram. Similarly, we distort the geometry heavily for amino acids with high uncertainty of the assignment. By default, we assign the ribbon geometry of the most likely secondary structure element as base mesh, but the user can also change this to the assignment of a particular method. For large proteins, assessing the distortion at individual amino acids might be challenging due to the small image footprint. This issue can be solved by zooming and panning, which is also necessary for detailed inspection of the undistorted ribbon diagram. Therefore, expert users are familiar with these interactions.

We compute the distortion using function g, where x is the normalized x-coordinate of the processed vertex on the spline segment of the current amino acid, the uncertainty value u and global parameters for amplitude k_a and frequency k_f of the distortion:

$$g(x,u;k_a,k_f) = u \cdot k_a \cdot w_{\{sin,tri\}}(u \cdot k_f \cdot x) \tag{3}$$

Note that g is applied in direction of the vertex normal (cf. Figure 8). The waveform w is a smooth sine or a sharp triangle function:

$$w_{sin}(t) = 0.5 \cdot (\sin(2\pi \cdot t - \pi/2) + 1) \tag{4}$$
$$w_{tri}(t) = 1 - 2 \cdot |0.5 - (t - \lfloor t \rfloor)| \tag{5}$$

The value is in the range $[0,1]$ to prevent the wave from carving into base geometry. Note that the tessellation level must be at least twice the frequency to prevent aliasing issues (Nyquist frequency).

Transparency In addition to the geometry distortion, we explored *screen door transparency* as a visual variable for the uncertain structure elements. This method allows us to show multiple secondary structure elements at the same time, sorted in ascending order by their confidence score without creating new colors (Figure 9).

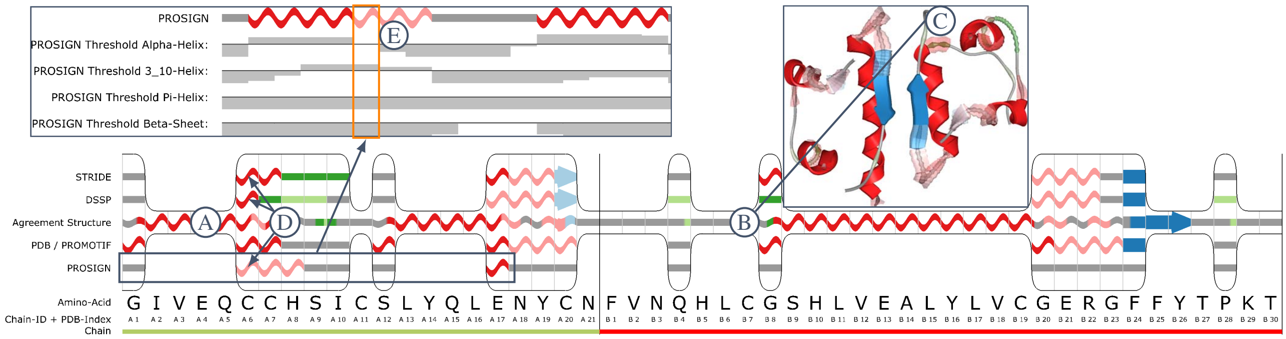

PROSIGN
PROSIGN Threshold Alpha-Helix:
PROSIGN Threshold 3_10-Helix:
PROSIGN Threshold Pi-Helix:
PROSIGN Threshold Beta-Sheet:
E
C
STRIDE
DSSP
Agreement Structure
PDB / PROMOTIF
PROSIGN
A
D
B
Amino-Acid
Chain-ID + PDB-Index
Chain
G I V E Q C C H S I C S L Y Q L E N Y C N F V N Q H L C G S H L V E A L Y L V C G E R G F F Y T P K T
A 1 A 2 A 3 A 4 A 5 A 6 A 7 A 8 A 9 A 10 A 11 A 12 A 13 A 14 A 15 A 16 A 17 A 18 A 19 A 20 A 21 B 1 B 2 B 3 B 4 B 5 B 6 B 7 B 8 B 9 B 10 B 11 B 12 B 13 B 14 B 15 B 16 B 17 B 18 B 19 B 20 B 21 B 22 B 23 B 24 B 25 B 26 B 27 B 28 B 29 B 30

Figure 10: Our uncertainty visualization applied to insulin (PDB ID: 1RWE). The sequence diagram shows different assignment methods and our inferred agreement structure, stacked on top of each other. At amino acids for which all algorithms agree, only the agreement structure is shown. The inset on the left shows the behavior of the internal thresholds of the PROSIGN algorithm for the highlighted part. The inset to the right shows the uncertainty value mapped to frequency in geometry in addition to geometry morphing and screen door transparency.

Let $|\mathbb{E}|$ be the number of possible structure elements and $I \subseteq [0,1]$ an interval. We divide I into $|\mathbb{E}|$ ranges so that the size of each interval $I_{e\in\mathbb{E}}$ corresponds to the confidence score of the corresponding structure element. Then, we use an 8×8 Bayer matrix M_{B8} [1] with entries distributed in $[0,1]$ as a repeating stencil to compose the final image. Hence, we only render the structure element e if the corresponding value of the matrix lies inside its interval I_e. This screen door transparency approach makes it easier to spot uncertain parts because it leads to perceivable noise at locations with different structure assignments. In theory, it even allows reading the distribution from the final image by counting pixels in an 8×8 bock.

In summary, the presented uncertainty visualizations can be used to guide the attention of the user toward uncertainly assigned parts of the secondary structure. The source of uncertainty can be examined further in the sequence diagram or by choosing the ribbon diagram of the single assignment method as base for the geometry distortion.

7 Use Cases and Discussion

In this section, we demonstrate our uncertainty visualizations for different application scenarios. As a primary source of uncertainty, we use the confidence score derived from the multiple assignment methods presented in Section 2. In particular, we use the scenarios to discuss how the requirements were satisfied. We first start with an example that explains how the uncertainty can be put in context (R1, R2, R4, R5). Next, we apply our visualization to a different source of uncertainty derived from aggregating the time steps of a protein simulation (R3). Finally, we use our visualization to illustrate positional uncertainty.

7.1 Deviations in Secondary Structure Assignment

We precomputed the secondary assignments for *insulin* (PDB ID: 1RWE) using STRIDE [10], the improved version of DSSP [45], and our implementation of the PROSIGN algorithm [16]. In addition, the downloaded PDB file contained the PROMOTIF [19] assignment. Insulin is a relatively small protein, while being large enough to have a meaningful secondary structure, as shown in Figure 10. We used our stacked sequence diagram, which adds the *agreement structure* row and highlights uncertain parts, where the structure assignments differ. Mark Ⓐ points to the agreement structure showing the combined structural uncertainty illustrated by colors and morphed geometry.

As expected, the PROSIGN algorithm deviates most often from the other ones due to its simpler approach, which induces a basic uncertainty for many amino acids. Another observation that can be made here is that if STRIDE and DSSP disagree, it is usually at the beginning or end of a helix, which is particularly visible in the 3D ribbon diagram. We recon that this is due to a slight unwinding of some helices toward their borders, which leads to deviating assignments because of algorithmic differences. To investigate this issue more closely, we chose the example marked by Ⓑ in the sequence diagram, where DSSP assigns a turn, whereas STRIDE assigns a helix and PROSIGN recognizes no clear structure (random coil). We can identify the corresponding position in 3D using brushing and linking Ⓒ. The shape of the 3D structure clearly reinforces our hypothesis that DSSP underestimates the length of the helix and PROSIGN is too simplistic. This assumption is backed by the fact that PROMOTIF also recognizes an α-Helix. Here, the screen door transparency proved to be useful, since it allows us to observe all structural assignments together.

A second example is marked by Ⓓ, where PROSIGN recognizes a 3_{10}-Helix, whereas all other algorithms recognize an α-Helix. To investigate this further, we go back to the corresponding position in the sequence diagram Ⓔ and examine the internal values of PROSIGN, which are shown in the detail view above (orange box). As observable, the internal confidence that the chain forms a 3_{10}-Helix at this location is not very high, as indicated by the corresponding soft label. In contrast, the internal value for an α-Helix is just below the threshold, indicating a very low overall confidence that the helix was correctly assigned by PROSIGN.

Although PROSIGN assigns the most false negatives (coil instead of a pronounced secondary structure), which is to be expected due to the simple internal model, it rarely assigns false positives. Therefore, we can place a relatively high trust in PROSIGN's assignment of helices and sheets with the caveat that these might be too short. As shown in the example above, our uncertainty visualization can be used to back this trust based on the internal thresholds.

As shown with this example analysis, our model facilitates the comparison on an amino acid level and various levels of detail (requirement R1). However, insights like above could be used to updated the discrepancy matrix to reflect the weighting of differences between the secondary structure assignment methods (requirement R2). The visual analysis process described above showcases that the visualization requirements (R4 and R5) are also satisfied.

7.2 Secondary Structure Dynamics and Flexibility

As stated in Section 4, one requirement was that our uncertainty visualization can also convey other sources of uncertainties derived from the dynamics of the protein (R3). We analyzed the stability of the secondary structure of a protein based on the results of a molecular dynamics simulation. During the simulation, the protein undergoes small conformational changes. Consequently, some parts of the secondary structure might vary over time. While such changes

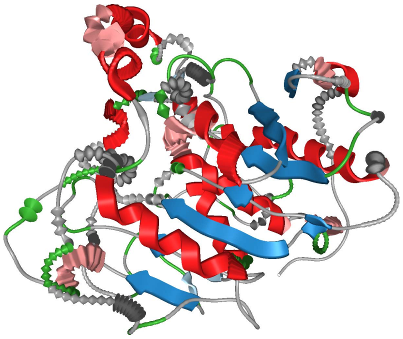


Figure 11: Uncertainty of the secondary structure derived by aggregating the changes that occurred during a simulation. This conveys the influence of dynamics on the secondary structure.

are typically investigated by domain experts watching an animation of the whole simulation, this quickly becomes not only tedious but also unreliable for longer simulations, since human observers tend to overlook change [46]. We therefore aggregated the secondary structure assigned by STRIDE for all time steps to obtain an uncertainty value per amino acid. The result can be seen in Figure 11. As expected, the secondary structures at the central core of the protein are temporally invariant, which correlates with the low uncertainty, while the random coils in the outer regions are more uncertain, which indicates that they are more unstable.

Our uncertainty visualization can also be used to complement the structural uncertainty with positional uncertainty information. We derive the positional uncertainty of a simulated protein using the Root Mean Square Fluctuation (RMSF), which is one of the standard measures for flexibility [29]. The RMSF is based on the Euclidean distance of each atom in each time step from its average position. Hence, accumulation of errors does not occur. An example is shown in Figure 12. As mentioned above, very small positional changes can lead to different secondary structure assignments if the corresponding internal value is close to a threshold. The positional uncertainty view is, therefore, intended to help users judge the structural uncertainty.

8 Evaluation

In this section, we present an evaluation of our technique based on questions and feedback from six experts in the fields of computational biology and scientific visualization. Due to the required domain knowledge, we were not able to recruit a sufficiently large number of participants to conduct a statistically reliable quantitative evaluation of the results. Nevertheless, the reviews can be seen as a summative evaluation of our formative development process: How do experts who were not involved in the development of our visualization explore the data? Does trust in the assignment change by means of uncertainty visualization? What aspects of our technique do the experts like or dislike? Which of their analysis tasks would benefit most from our technique?

8.1 Tasks, Questions, and Participants

Each expert review took about 45 min. All experts were given a short introduction and set of tasks that relate to the examples in Section 7. During the introduction, we also asked control questions to assess whether the participants were able to judge the relative amount of uncertainty (rank order). The three given tasks were:

T 1 Explore differences between multiple assignment methods
T 2 Explore changes in the secondary structure over time
T 3 Explore differences between secondary structure uncertainty and positional uncertainty (flexibility)

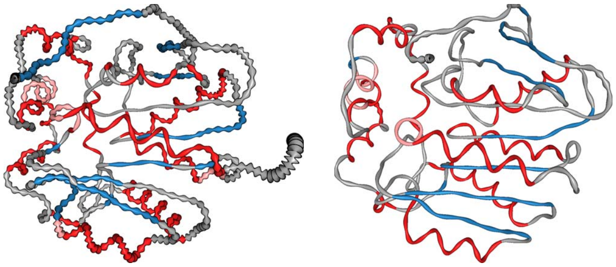


Figure 12: Visualization of positional uncertainty (i.e., flexibility computed by Root Mean Square Fluctuation) of the wild type of a dehalogenase protein (left) and a mutant (right). As observable, the mutant is much more stable than the wild type. Both proteins were rendered as tubes instead of ribbons with coloring by secondary structure.

During each session, two operators took notes regarding verbalized feedback. Furthermore, one operator asked questions about each task to gain an impression of user experience. Some questions had to be provided on a five-point Likert scale, while other questions were designed to trigger open-ended feedback. We used video conferencing with screen sharing for off-site participants.

For Task 1, the users were presented a set of images that showed the uncertainty with varying saliency based on the different assignment methods (cf. Section 5; see Supplementary Material).

Q 1.1 How helpful is the currently shown visualization to identify uncertain regions?
Q 1.2 Does the visualization help you assess the reliability of the different assignment methods?
Q 1.3 How did the visualization change your trust in secondary structure assignment?
Q 1.4 How does it change your trust to see the confidence score of possible secondary structure assignments per amino acid?
Q 1.5 What are the benefits or drawbacks of ribbon diagram and sequence diagram?

For Task 2, we used a simulation of a protein with varying secondary structure. We derived uncertainty values for the variable parts of the secondary structure over time using STRIDE. The users were presented a short video of the protein as animated, classical ribbon diagram where the structural changes over time are directly visualized as well as static images of a ribbon diagram with time-aggregated structural uncertainty encoded as different visual variables.

Q 2.1 How helpful is the visualization to show structural changes?
Q 2.2 How did the visualization change your trust in the secondary structure changes over time?
Q 2.3 Which visualization would you use to make a more informed analysis of functional or structural properties?

For Task 3, we used two simulations of a protein: the wild type and a mutation that exhibits significantly lower flexibility. We showed images of the temporally aggregated structural changes (cf. Task 2) alongside images where the actual positional uncertainty (i.e., flexibility determined using RMSF) was encoded.

Q 3.1 Does the visualization help find the most flexible protein?
Q 3.2 Do you prefer the ribbons or the tube for flexibility analysis?
Q 3.3 Does the additional flexibility visualization increase your trust in the secondary structure stability?

Participants E 1 and E 2 were experts in biomolecular visualization and have experience with uncertainty visualization, who have been working for ten and twelve years in the field. They were included to assess our method from a technical point of view. Participants E 3–E 6 were experts in structural biology or biochemistry working between two and ten years in the field, i.e., potential users of our method. None of the experts were involved in the design of our uncertainty visualization, nor have they seen our visualization prior to the review session. Two of the participants were female,

Table 2: Survey answers on a five-point Likert scale from -2 (‘Strongly Disagree/Very Poor’, 🟥) to 2 (‘Strongly Agree/Very Good’, 🟦): 🟥🟧🟨🟦🟦. In some cases, separate ratings had to be given for the ribbon diagram (*RD*) and the sequence diagram (*SD*). For Q 2.3, 🎞 denotes the animation and 🖼 the still image with encoded uncertainties. For Q 3.2, ((● denotes the tube and 🗺 the ribbon.

Q	**1.1**		**1.2**		**1.3**	**1.4**	**2.1**		**2.2**	**2.3**	**3.1**	**3.2**	**3.3**
	RD	SD	RD	SD			A	I					
E 1	1	2	2	1	0	0	-2	1	0	🖼	2	((●	∅
E 2	1	1	2	2	-1	1	-1	2	-1	🖼	2	((●	1
E 3	1	2	1	1	0	1	1	-1	0	🎞	2	🗺	1
E 4	1	2	-2	1	0	1	0	1	-1	🖼+🎞	1	((●	1
E 5	1	1	1	1	0	1	1	-1	0	🎞	2	🗺	1
E 6	1	2	0	1	1	1	1	1	0	🎞	1	((●	1

four male. None of them had color vision deficiencies.

8.2 Observations and Results

Questionnaire Results In Table 2, we list the ratings given for the Likert-scaled and binary questions. The overall feedback was positive. All participants agreed that both the ribbon diagram and the sequence diagram convey the uncertainties well or very well. Only one of the experts (E 4) rated the ribbon diagram as very bad to assess the reliability of the different assignment methods. E 4 argued that one needs the numerical uncertainty values for a detailed analysis, which are only available in the sequence diagram.

For question Q 1.5 (pros and cons of RD and SD), the results are mixed. Three participants (the two visualization experts, E 1 & E 2, and one biology expert, E 4) reasoned that the combination of both diagrams is ideally suited for a visual analysis pipeline: The 3D ribbon diagram can be used to get an overview of the whole protein and spot interesting locations with high uncertainty, whereas the sequence diagram can be used for a detailed analysis of these locations. Two of the biology experts (E 3 & E 5) clearly preferred the ribbon diagram, since it gives additional spatial information, which is important for their application. E 4 also noted that the benefit of the ribbon diagram is that users typically investigate the same set of proteins for a longer period and recognize the 3D structure of these proteins, which leads to a good recognition value of the ribbon diagram. E 6 would prefer the ribbon diagram to analyze protein properties and the sequence diagram to compare algorithms.

Interestingly, for Questions 2.x, the two visualization experts rated the animation as much worse than the still image with encoded structural uncertainties aggregated over time and preferred it for the analysis. The biology experts, however, had a tendency toward the animation, since the temporal information is lost due to the summarization in the still image. We recon that this preference is also partially due to the biology experts being trained to analyzing animations. E 1, E 2, and E 4 argued that long animations are tedious to watch and that it is very easy to miss important parts. Therefore, E 4 mentioned that the aggregated image would be a very good starting point for the analysis to identify interesting regions, but the animation would still be needed afterward for a detailed analysis.

For Questions 3.x, all experts agreed that the additional visualization of the positional uncertainty (i.e., flexibility) helps estimate the reliability of the assignment methods. Most of them preferred the tube rendering over the ribbons when illustrating the flexibility, since the ribbons make the image more noisy and, therefore, harder to read. Only two of the biology experts (E 3 and E 5) preferred the ribbons, since they lead to a higher recognition value (as mentioned by E 4, who preferred the tubes nevertheless). E 3 noted that there were issues seeing the spatial structure when using only tubes. Visualization expert E 1 did not feel qualified to answer this application-specific question.

Please note that for the first two questions about the change of *trust* (Q 1.3 and Q 2.2), most participants noted they have already been aware of the fact that the secondary structure assignment methods are sometimes unreliable. Thus, their trust in the algorithmic results mostly stayed the same. However, all participants agreed that it is interesting to see the uncertain regions for an in-depth analysis. Only visualization expert E 2 was not aware that the assignment result differs so much. Here, our visualization decreased the trust in the assignment. Biology expert E 6 particularly liked that our visualization shows where and how much the algorithms coincide, leading to an increased trust in the assignment in more certain regions. As we expected, the concurrent visualization of structural uncertainty and positional uncertainty increases the trust in the temporal results (Q 3.3), since it allows users to draw conclusions about the reliability of the secondary structure assignment.

Additional Feedback We also recorded general feedback that we got during the expert reviews. For the ribbon diagrams, the images with contour lines were consistently rated as visually most appealing. Especially the biology experts liked them because it makes the images very well suited for print. The ribbon diagrams with transparency were noted to provide most information, but they were also rated as less easy to read, since they are visually most complex (E 1, E 2, and E 4). For the sequence diagram, most experts preferred the stacked view showing all assignments (Fig. 5a), since it was rated to be best suited for a detailed analysis of the different assignments. The stacked grouped and sorted view (Fig. 5b) was rated as very helpful to find uncertain regions. Only E 6 rated the morphed view (Fig. 6) as the best option, since it takes the least space and argued that the visualization is sufficiently detailed for the analysis. In general, all experts liked the idea of having information about uncertainties encoded in the image (in a non-obtrusive way). E 1 remarked that the high frequencies in the ribbon diagram are well-suited to draw attention toward uncertain parts. E 3 specifically noted that the enriched ribbon diagram is very good, since it uses a well-known visualization and adds information about the uncertainty without hiding or obliterating other information.

9 Summary and Future Work

We have presented a combination of uncertainty visualizations for molecular data in 2D and 3D. To this end, we explored different visual variables to encode uncertainty onto the established secondary structure representations for proteins. The main goal here was to maintain community standards regarding secondary structure visualization. Specifically, our extension of sequence diagrams allows for a quantitative analysis of uncertainties, as absolute values are depicted as bars and as morphed secondary structures in 2D. In the 3D visualization, the ribbon diagram is decorated with waviness. The amplitude and frequency convey the amount of uncertainty. Transparency and contour lines are used to further emphasize uncertain structures. The underlying uncertainty model can flexibly map various sources of uncertainty to a single value per amino acid. In this context, we have propagated internal thresholds from PROSIGN and STRIDE to show its fuzzy nature in terms of classification.

We have shown and evaluated the applicability of our approach using three use cases: deviations in the secondary structure assignment of multiple assignment methods, aggregation of secondary structure changes over simulation time, and physically motivated protein flexibility. As a result of the evaluation, all experts recognized the benefits of uncertainty visualization, i.e., that it is important to depict uncertainty to prevent misinterpretations. Another result is that animation seems to be preferred by biology experts, even if considered inappropriate for visualization [46].

In the future, we want to adapt our uncertainty visualization to other applications such as *ab initio* structure prediction based on protein sequences [13] or protein function prediction [5]. We also want to evaluate the effectiveness of screen door transparency and extend our method to be able to convey structural uncertainty in tertiary structures, for example, using volume rendering.

Acknowledgments

We thank the German Research Foundation for financial support within project A01 of SFB/Transregio 161 and project D.4 of SFB 716. We also thank Barbora Kozlíková and Jiří Damborský from Masaryk University for providing test data, and all participants of our study.

References

[1] B. E. Bayer. An optimum method for two-level rendition of continous-tone pictures. *SPIE Milestone Series*, 154:139–143, 1999.

[2] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Res.*, 28(1):235–242, 2000.

[3] K. Brodlie, R. A. Osorio, and A. Lopes. A review of uncertainty in data visualization. In *Expanding the Frontiers of Visual Analytics and Visualization*, pages 81–109. Springer Nature, 2012.

[4] M. Carson. Ribbons. In R. M. Sweet and C. W. Carter, editors, *Methods in Enzymology*, volume 277 of *Macromolecular Crystallography*, pages 493–505. Academic Press, 1997.

[5] M. Chitale, T. Hawkins, and D. Kihara. *Automated Prediction of Protein Function from Sequence*, pages 63–85. 2008.

[6] J. Chuang, D. Weiskopf, and T. Möller. Hue-preserving color blending. *IEEE TVCG*, 15(6):1275–1282, 2009.

[7] A. Coninx, G.-P. Bonneau, J. Droulez, and G. Thibault. Visualization of uncertain scalar data fields using color scales and perceptually adapted noise. In *ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization (APGV)*, pages 59–66, 2011.

[8] M. Correll and M. Gleicher. Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE TVCG*, 20(12):2142–2151, 2014.

[9] A. Dahl, M. Chavent, and M. Sansom. Bendix: intuitive helix geometry analysis and abstraction. *Bioinformatics*, 28(16):2193–2194, 2012.

[10] D. Frishman and P. Argos. Knowledge-based protein secondary structure assignment. *Proteins: Struct., Funct., Genet.*, 23:566–579, 1995.

[11] G. Grigoryan and P. Rheingans. Point-based probabilistic surfaces to show surface uncertainty. *IEEE TVCG*, 10(5):564–573, 2004.

[12] T. Gschwandtner, M. Bögl, P. Federico, and S. Miksch. Visual encodings of temporal uncertainty: A comparative user study. *IEEE TVCG*, 22(1):539–548, 2016.

[13] R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, Y. Yang, and Y. Zhou. Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Sci. Rep.*, 5:11476, 2015.

[14] P. Hermosilla, V. Guallar, Á. Vinacua, and P. P. Vázquez. Instant visualization of secondary structures of molecular models. In *EG VCBM*, pages 51–60, 2015.

[15] C. Holzhüter, A. Lex, D. Schmalstieg, H.-J. Schulz, H. Schumann, and M. Streit. Visualizing uncertainty in biological expression data. In *Proc. SPIE*, volume 8294, 2012.

[16] S.-R. Hosseini, M. Sadeghi, H. Pezeshk, C. Eslahchi, and M. Habibi. PROSIGN: A method for protein secondary structure assignment based on three-dimensional coordinates of consecutive C_α atoms. *Comp. Biol. Chem.*, 32(6):406–411, 2008.

[17] J. Hullman. Why evaluating uncertainty visualization is error prone. In *Proc. BELIV'16*, pages 143–151, 2016.

[18] W. Humphrey, A. Dalke, and K. Schulten. VMD – Visual Molecular Dynamics. *J. Mol. Graphics*, 14:33–38, 1996.

[19] E. Hutchinson and J. Thornton. Promotif-A program to identify and analyze structural motifs in proteins. *Protein Sci.*, 5(2):212–220, 1996.

[20] IUPAC-IUB. Abbreviations and symbols for the description of the conformation of polypeptide chains. *FEBS J.*, 17(2):193–201, 1970.

[21] C. Johnson. Top scientific visualization research problems. *IEEE Computer Graphics and Applications*, 24(4):13–17, 2004.

[22] W. Kabsch and C. Sander. Dictionary of Protein Secondary Structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[23] R. Khlebnikov, B. Kainz, M. Steinberger, and D. Schmalstieg. Noise-based volume rendering for the visualization of multivariate volumetric data. *IEEE TVCG*, 19(12):2926–2935, 2013.

[24] A. Knoll, M. Chan, K. Lau, B. Liu, J. Greeley, L. Curtiss, M. Hereld, and M. E. Papka. Uncertainty classification and visualization of molecular interfaces. *Int. J. Uncertain. Quantif.*, 3(2):157–169, 2013.

[25] L. Kocincová, M. Jarešová, J. Byška, J. Parulek, H. Hauser, and B. Kozlíková. Comparative visualization of protein secondary structures. *BMC Bioinformatics*, 18(2):23, 2017.

[26] B. Kozlíková, M. Krone, M. Falk, N. Lindow, M. Baaden, D. Baum, I. Viola, J. Parulek, and H.-C. Hege. Visualization of biomolecular structures: State of the art revisited. *Comput. Graph. Forum*, 36(8):178–204, 2016.

[27] M. Krone, K. Bidmon, and T. Ertl. GPU-based visualisation of protein secondary structure. In *EG.UK TPCG*, pages 115–122, 2008.

[28] M. Krone, K. Bidmon, and T. Ertl. Interactive visualization of molecular surface dynamics. *IEEE TVCG*, 15(6):1391–1398, 2009.

[29] A. Kuzmanic and B. Zagrovic. Determination of ensemble-average pairwise root mean-square deviation from experimental B-factors. *Biophys. J.*, 98(5):861–871, 2010.

[30] C. H. Lee and A. Varshney. Representing thermal vibrations and uncertainty in molecular surfaces. In *SPIE Conference on Visualization and Data Analysis*, pages 80–90, 2002.

[31] K. U. Linderstrøm-Lang. *Lane Medical Lectures: Proteins and Enzymes*, volume 6. Stanford University Press, 1952.

[32] A. MacEachren, R. Roth, J. O'Brien, B. Li, D. Swingley, and M. Gahegan. Visual semiotics & uncertainty visualization: An empirical study. *IEEE TVCG*, 18(12):2496–2505, 2012.

[33] J. Martin, G. Letellier, A. Marin, J. Taly, A. de Brevern, and J. Gibrat. Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. *BMC Struct. Biol.*, 5:17, 2005.

[34] A. T. Pang, C. M. Wittenbrink, and S. K. Lodha. Approaches to uncertainty visualization. *The Visual Computer*, 13(8):370–390, 1997.

[35] G. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, 7(1):95–99, 1963.

[36] M. Rasheed, N. Clement, A. Bhowmick, and C. Bajaj. Quantifying and visualizing uncertainties in molecular models. *arXiv:1508.03882 [cs]*, 2015. arXiv: 1508.03882.

[37] P. Rheingans and S. Joshi. Visualization of molecules with positional uncertainty. In *Joint EG and IEEE TCVG Symposium on Visualization (Data Visualization '99)*, pages 299–306, 1999.

[38] J. S. Richardson. The anatomy and taxonomy of protein structure. *Advances in Protein Chemistry*, 34:167–339, 1981.

[39] G. Ristovski, T. Preusser, H. K. Hahn, and L. Linsen. Uncertainty in medical visualization: Towards a taxonomy. *Computers & Graphics*, 39:60–73, 2014.

[40] L. F. O. Rocha. Toward a better understanding of structural divergences in proteins using different secondary structure assignment methods. *Journal of Molecular Structure*, 1063:242–250, 2014.

[41] J. Sanyal, S. Zhang, J. Dyer, A. Mercer, P. Amburn, and R. J. Moorhead. Noodles: A tool for visualization of numerical weather model ensemble uncertainty. *IEEE TVCG*, 16(6):1421–1430, 2010.

[42] J. Schmidt-Ehrenberg, D. Baum, and H.-C. Hege. Visualizing dynamic molecular conformations. In *Proc. IEEE Vis.*, pages 235–242, 2002.

[43] Schrödinger, LLC. The PyMOL Molecular Graphics System, v1.8. 2016.

[44] M. Tarini, P. Cignoni, and C. Montani. Ambient occlusion and edge cueing for enhancing real time molecular visualization. *IEEE TVCG*, 12(5):1237–1244, 2006.

[45] W. G. Touw, C. Baakman, J. Black, T. A. te Beek, E. Krieger, R. P. Joosten, and G. Vriend. A series of PDB-related databanks for everyday needs. *Nucleic Acids Res.*, 43(Database issue):364–368, 2015.

[46] B. Tversky, J. Bauer Morrison, and M. Betrancourt. Animation: can it facilitate? *Int. J. Hum. Comput. Stud.*, 57(4):247–262, 2002.

[47] M. Wahle and S. Birmanns. GPU-accelerated visualization of protein dynamics in ribbon mode. In *SPIE Conference on Visualization and Data Analysis*, volume 7868, pages 786805–12, 2011.