STRUCTURAL BIOLOGY

ISSN 2059-7983

Exploration of questionable backbone conformations in crystallographic structure models using a structural alphabet

Clémence Sarrau, Marine Baillif, Lucas Mantel, Dounia Benyakhlaf, Shamima Peerbux and Leslie Regad*

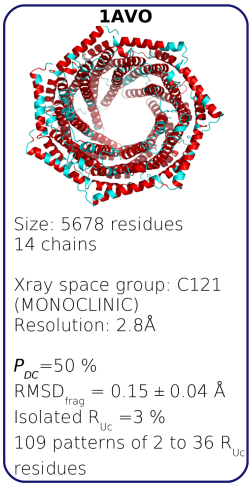
Université Paris Cité, Unité de Biologie Fonctionnelle et Adaptative (BFA) CNRS UMR 8251 – In Silico Pharmacological Profiling (IsPP) INSERM U1133, F-75205 Paris CEDEX 13, France. *Correspondence e-mail: leslie.regad@u-paris.fr

Received 20 February 2025
Accepted 21 October 2025

Edited by M. Vollmar, European Bioinformatics Institute, United Kingdom

Keywords: questionable conformations; crystallographic structures; structural alphabet; structural deformation.

More than 80% of protein structure models in the Protein Data Bank have been solved using X-ray crystallography. Despite continuous improvements in this experimental technique, crystallographic structure models may still present artifacts related to the crystallization process as well as errors introduced during model building and refinement, even in high-resolution cases. Such limitations can alter atomic or residue positions, leading to local misconformations, local or domain rearrangements, and occasionally global distortions. In this study, we developed a protocol to locate residues with questionable conformations, where conformations may be uncertain, atypical or influenced by crystallographic modeling and refinement. To do so, we started from a set of 826 nonredundant X-ray protein structure models. Each X-ray model underwent an energy-minimization step that relaxes atomic geometry by reducing potential energy. Residues that exhibited different local conformations between the X-ray and minimized models were therefore considered as having questionable conformations. To identify them, we compared the X-ray and minimized models of each protein using the HMM-SA structural alphabet. Our results revealed that over 18% of the residues in the protein set have questionable conformations in their backbone. These conformations can occur either as isolated events within the protein sequence or can form patterns. Moreover, we observed that the frequency of questionable conformations per X-ray model was independent of factors such as the date of deposition, resolution or crystal system. Analysis of the properties of residues associated with questionable conformations revealed that they do not specifically occur in flexible or accessible regions. However, there is a correlation between questionable conformations and secondary structures, with a particular overrepresentation of residues with questionable conformations in α-helices. We then further investigated questionable conformations in the structure model of ligand-free HIV-2 protease (PR2). By combining our questionable conformation-detection protocol with molecular-dynamics simulation, we demonstrated that approximately half of the questionable conformations in PDB entry 1hsi correspond to local conformations that are sparsely sampled by PR2 during the molecular-dynamics simulation or are structural outliers detected by the wwPDB report. In addition, our results suggested that these questionable conformations may affect the position of the flaps, two β-sheets forming the top of the binding site. In PDB entry 1hsi, their relative arrangement appears atypical compared with MD simulations, raising questions about the biological relevance of this conformation. To conclude, we have developed a protocol to quantify and localize questionable backbone conformations in X-ray structure models, which can affect the interpretation of structural data.



1. Introduction

The Protein Data Bank (PDB) contains over 217 000 protein structure models, with 183 000 resolved using X-ray crystallo-

graphy (Berman *et al.*, 2000). These entries result from a complex process, from macromolecule purification to structure model deposition in the wwPDB, with each step introducing potential errors. Each step in the process introduces both systematic and random errors. These errors can result from experimental limitations or from modeling and refinement decisions. Several metrics quantify the quality of protein structure models, such as resolution, $R_{\rm free}$ (Brünger, 1992), clashscore (Chen *et al.*, 2010) and real-space *R*-value *Z*-score (RSRZ) outliers (Kleywegt *et al.*, 2004; Gore *et al.*, 2012), which are detailed in the wwPDB X-ray Validation Report. Each of these parameters plays a crucial role in assessing different aspects of the accuracy of a molecular model. The resolution parameter focuses on the ability to discern details in an X-ray crystallographic model. The $R_{\rm free}$ criterion evaluates the overall quality of the model by comparing experimental data with calculated data, particularly emphasizing the validation set to avoid overfitting. The clashscore quantifies steric clashes in the model, and the RSRZ score provides a local assessment of structural quality by evaluating how well individual residues fit the electron density. Automated validation tools assess structure model quality by applying geometric and stereochemical criteria to detect errors and inconsistencies. *MolProbity* (Chen *et al.*, 2010) is one of the most widely used tools for this task. It provides a detailed evaluation of model geometry, including interatomic distances, bond angles and torsion angles. It also generates electron-density maps and checks the placement of residues using the Ramachandran plot, offering suggestions for correcting geometric defects. *WHAT_CHECK* (Hooft *et al.*, 1996) is a comprehensive validation tool used to assess the stereochemical quality of structure models. It analyses geometric parameters such as bond lengths, bond angles, side-chain conformations and atomic clashes. The program also provides an overall quality score to identify potential errors or unusual features in the model. *PROCHECK* (Laskowski *et al.*, 1993), on the other hand, focuses on torsion-angle analysis and the Ramachandran plot, helping to identify residues in geometrically unfavorable conformations. *PDB-REDO* (Joosten *et al.*, 2009, 2014; Joosten & Vriend, 2007; van Beusekom *et al.*, 2018) is an automated platform designed to improve macromolecular crystallographic models through iterative re-refinement, rebuilding and validation. It optimizes stereochemistry, updates atomic displacement parameters and re-evaluates *R* factors and $R_{\rm free}$ values by reanalyzing electron-density maps. These automated tools are essential not only for validating initial models but also for refining existing structure models, ensuring their accuracy and reliability before they are used in biological and biomedical studies. Despite advancements in resolution techniques and in structure model validation, artifacts related to the crystallization process and errors introduced during model building and refinement persist in X-ray structure models, even at high resolutions. These issues affect the conformations of individual atoms or residues, and can extend to local or domain movements or even structure models with global misfolding (Davis *et al.*, 2008). Several factors can induce these structural deviations: limitations in the crystallographer's subjective interpretation of experimental electron-density maps, data quality, experimental conditions, errors in the refinement process, crystallization artifacts, crystal packing and inherent molecular properties such as protein flexibility (Brändén & Jones, 1990; Jones & Kjeldgaard, 1997; Davis *et al.*, 2008). For example, PDB entries 1jsq, 1pf4, 1z2r, 1s7b and 2f2m correspond to models of the MsbA and EmrE transporters. These entries were moved to the PDB obsolete archive because their shapes roughly resembled the mirror image of the correct structure model (Tate, 2006; Dawson & Locher, 2006). Additionally, the chemical environment of the crystal (buffer salt concentration, pH *etc.*) can affect the conformation observed in the X-ray model. For example, high concentrations of lyotropic salts in the crystallization buffer of human protein kinase CK2 maintain the closed conformation of the kinase (Klopffleisch *et al.*, 2012; Srivastava *et al.*, 2018). Another example illustrating the impact of experimental conditions on protein conformation is the Bcl-xL protein. When resolved in the presence of detergents such as *n*-octyl-β-D-maltoside or exposed to pH 10, Bcl-xL undergoes structural rearrangements. In this state, the two α5 and α6 helices fuse into a single long helix (O'Neill *et al.*, 2006; Follis *et al.*, 2013; Tanaka *et al.*, 2013; Rajan, Choi, Baek *et al.*, 2015; Rajan, Choi, Nguyen *et al.*, 2015; Salam *et al.*, 2018). Crystal packing is another parameter that can introduce artifacts or deviations from biological conformations in X-ray models. It refers to the arrangement of individual biomolecules within a crystal lattice during crystallization for X-ray crystallography. As a result of crystallization, protein molecules in asymmetric units or neighboring symmetry-related molecules interact. These nonbiological interactions, which are absent in functional molecules in cells, are referred to as crystal-packing artifacts (Tsuchiya *et al.*, 2008).

The impact of crystal packing on protein structures has been explored in various studies. One strategy consists of comparing X-ray structure models of the same protein resolved under different conditions (Heinz *et al.*, 1991; Zhang *et al.*, 1995; Bertrand *et al.*, 2000; Taylor *et al.*, 2001; Eyal *et al.*, 2005; Tsuchiya *et al.*, 2008; Regad *et al.*, 2017; Srivastava *et al.*, 2018; Triki *et al.*, 2019) or X-ray models independently crystallized by different groups (Martin *et al.*, 2008; Mei *et al.*, 2020). These studies revealed that packing effects can alter backbone or side-chain conformations, causing rigid-body motions of large structural units or loop conformational changes. For example, Eyal and coworkers compared 404 pairs of structure models of the same protein obtained from different crystals with identical forms (Eyal *et al.*, 2005). They showed that residues involved in crystal contacts are less mobile than other surface residues. They also found that crystal packing can modify water positions but does not affect ligand positions. To distinguish structural deformations induced by partner binding from those caused by experimental errors or crystallographic artifacts, Martin and coworkers used a set of 14 protein model pairs independently crystallized by different groups (Martin *et al.*, 2008). To more broadly explore the deformations caused by crystal packing,

several studies compared X-ray and NMR models (Betts & Sternberg, 1999; Jacobson *et al.*, 2002; Eyal *et al.*, 2005; Garbuzynskiy *et al.*, 2005; Andrec *et al.*, 2007; Sikic *et al.*, 2010; Koehler Leman *et al.*, 2018; Mei *et al.*, 2020; Grigas *et al.*, 2022). For example, Mei and coworkers compared the cores of X-ray and NMR models for 21 proteins, revealing that NMR models are more tightly packed than the cores of X-ray models (Mei *et al.*, 2020). Unfortunately, only a few proteins in the PDB have been resolved by both NMR and X-ray crystallography. This makes it difficult to compare multiple NMR models with a single X-ray model. Other studies compared biological interfaces with nonbiological interfaces induced by crystal packing. They showed that packing contacts have smaller interfaces than biological contacts and that their amino-acid composition is indistinguishable from the rest of the protein surface. In addition, these contacts are less hydrophobic and contain fewer fully buried atoms (Janin & Rodier, 1995; Carugo & Argos, 1997; Carugo & Djinović-Carugo, 2012; Luo *et al.*, 2015). Several algorithms have been developed to predict whether an interface corresponds to a biological or crystallographic interface (Krissinel & Henrick, 2005, 2007; Elez *et al.*, 2020; Liu *et al.*, 2014; Tsuchiya *et al.*, 2006, 2008; Zhu *et al.*, 2006; Bernauer *et al.*, 2008; Elez *et al.*, 2018). For example, the *PISA* method predicts the stability of an interface using a scoring function (Krissinel & Henrick, 2005, 2007). All of these studies have only explored artifacts in X-ray structure models arising from crystal packing without taking into account other types of deviations, or conformational changes located outside the interfaces. Additionally, they are often based on small data sets and focus on exploring either the global fold or side-chain deformations.

Detecting and understanding residues with questionable conformations in X-ray structure models is crucial when studying proteins. These conformations may reflect local uncertainties, modeling errors or artifacts arising from crystallographic conditions. In all cases, they can compromise structural interpretation and lead to misleading conclusions. In this study, we focused on locating and characterizing residues whose local conformations appear to be questionable in the backbone of X-ray structure models, using a large data set of 826 proteins. To identify such cases, we compared the local conformations of residues in X-ray models with those in the corresponding energy-minimized models. Energy minimization locally relaxes the atomic geometry by reducing steric clashes and strained conformations, without being constrained by crystallographic restraints. This procedure may therefore propose alternative conformations for residues affected by local uncertainties, actual inaccuracies (such as modeling errors and clashes) or deviations arising from crystallographic artifacts. Using the HMM-SA structural alphabet (Camproux *et al.*, 2004; Regad *et al.*, 2008), the X-ray and minimized models of proteins were encoded into sequences of structural letters, where each letter represents the fold of a four-residue fragment. Residues with questionable conformations, as used here, are defined as residues whose backbone adopts different local conformations in X-ray and minimized models, corresponding to different structural letters in the sequences derived from each model. Our results revealed that more than 18% of residues in the data set exhibit questionable backbone conformations. The proportion of such residues per X-ray model does not depend on protein length, crystal system or the date of model deposition. In contrast, we observed a moderate association with resolution and CATH classes. In particular, residues within helical regions were overrepresented among those with questionable conformations. Furthermore, we demonstrated that residues with questionable conformations are not preferentially located in flexible and solvent-exposed regions. The impact of the presence of questionable conformations on the global fold was explored using a case study of HIV-2 protease, an important target for treating HIV-2 infection. Our results showed a correspondence between residues with questionable conformations and those detected as structural outliers. Additionally, analysis of the HIV-2 protease structure during molecular-dynamics simulations revealed that residues with questionable conformations cause certain regions to adopt unsual conformations. Our study revealed that questionable conformations are common, particularly in structure models rich in α-helices. Our findings emphasize the need for caution when interpreting X-ray structure models and highlight the importance of thoroughly assessing their quality before use.

2. Materials and methods

2.1. Data collection

2.1.1. The data set used to explore and quantify questionable conformations in X-ray models

We started with a set of 826 nonredundant X-ray protein structure models (less than 25% sequence identity) extracted from the PDB. This data set was named the Xray_{826} set. All models had a resolution better than 4 Å and more than 40 residues (Supplementary Fig. S1). They were in the free form, *i.e.* they were not complexed with a ligand or peptide and no metals were present in the structure models. Most structure models (75%) were monomers or dimers and contained fewer than 1200 residues. The X-ray model resolution distribution was centered around 2 Å and a large number of models (63%) were crystallized in the monoclinic or orthorhombic crystal systems (Supplementary Fig. S1). The distribution of CATH classes in the data set was similar to that of the CATH classification, with over 50% of structures classified as 'alpha beta' (Sillitoe *et al.*, 2021; Supplementary Fig. S1).

2.1.2. The common data set between the Xray_{826} set and the PDB-REDO databank

The PDB-REDO databank (Joosten *et al.*, 2009, 2014; Joosten & Vriend, 2007; van Beusekom *et al.*, 2018; https://pdb-redo.eu) is a comprehensive and open-access repository that provides refined macromolecular structure models. These models, originating from the PDB, undergo an automated re-refinement process using the *PDB-REDO* procedure. This procedure enhances the quality and reliability of the original deposited models by applying advanced refinement tools

and validating structural parameters. From an experimental model, the *PDB-REDO* procedure recalculates the atomic coordinates, optimizes stereochemical parameters and, when structure factors are available, improves the agreement between the model and the electron-density map. The platform also evaluates and adjusts R factors and R_{free} values, key indicators of model quality. In this study, we used these refined models extracted from the PDB-REDO databank as a reference to evaluate the ability of our protocol to identify residues with questionable conformations, particularly those associated with limitations in refinement or model interpretation. It is important to note, however, that the *PDB-REDO* re-refinement procedure does not affect the crystal-packing environment. As a result, conformational biases introduced by packing contacts remain unchanged.

We thus extracted the refined models of the $Xray_{826}$ proteins from the PDB-REDO databank. Of these, 694 structure models were present in both the PDB and the PDB-REDO databank. This subset of 694 common models was denoted the $Xray_{694}$ set. The 132 missing structures were most likely not available in the PDB-REDO databank because the corresponding experimental data (such as structure-factor files) may have been missing or incomplete in the original PDB entries. This lack of data would have prevented their re-refinement. For each protein of the $Xray_{694}$ set, we have two models: the X-ray model extracted from the PDB (X-ray models) and the refined model extracted from the PDB-REDO databank (PDB-REDO models).

2.1.3. HIV-2 protease structure model sets

In the second step of this study, we explored in detail the questionable conformations in the structure model of HIV-2 protease (PR2) in its free form, *i.e.* not complexed with a ligand. PR2 is a crucial target for treating HIV-2 infection, as it represents one of the key enzymes essential for the replication of HIV-2. More precisely, PR2 is involved in the maturation of the virus by cleaving viral polyproteins into functional and structural components (Menéndez-Arias & Álvarez, 2014). This protein is a homodimeric aspartyl protease with 99 residues in each monomer. Its active site is situated at the interface of the two monomers (Fig. 1). It is covered by the two flap regions, which are β-sheets that close over the active site upon substrate binding. There are 19 structure models of PR2 available in the PDB, with 18 complexed with an inhibitor and one in a ligand-free form (PDB entry 1hsi; Chen *et al.*, 2014). In this study, we focused on this latter PR2 structure model without ligand.

To explore the conformational landscape of the ligand-free form of PR2, we generated a set of conformations using molecular-dynamics simulations with the *GROMACS* software (version 2022; Abraham *et al.*, 2015) and the Amber ff99SB-ILDN force field (Lindorff-Larsen *et al.*, 2010). Beginning with the X-ray model PDB entry 1hsi, we removed ions and water molecules. Next, we protonated the model at the O atom (OD1) of the aspartic acid at position 25 of chain *B* using the *PROPKA* software (Li *et al.*, 2005). The protein was then immersed in the center of a cubic water box (TIP3P) and placed at least 1.2 Å from the box edge. To neutralize the system, counterions were added. The neutralized system then underwent an energy-minimization step using the steepest-descent algorithm with 50 000 steps and a maximum force threshold of 1000.0 kJ mol^{-1} nm^{-1}. Nonbonded interactions were truncated at a cutoff distance of 10 Å for the electrostatic twin-range cutoff and the van der Waals cutoff. The system was then equilibrated for 2 ns followed by a production step of 500 ns. These steps were performed at a pressure of 1 bar and a temperature of 300 K. The temperature and pressure were kept constant thanks to temperature coupling using the *V*-rescale method and pressure coupling controlled by the Parrinello–Rahman algorithm, respectively. The *LINCS* algorithm was used to constrain the covalent bonds of H atoms. Long-range electrostatic interactions were computed by the particle mesh Ewald (PME) method. The Verlet method was used for neighbor searching with a cutoff of 10 Å for both electrostatic and van der Waals interactions. Periodic boundary conditions were applied in three dimensions. A snapshot of the trajectory was saved every 1 ns, resulting in a set of 501 structure models referred to as MD models. The stability and the convergence of the system during the simulation are presented in Supplementary Fig. S2.

2.2. Location and quantification of residues with questionable conformations in X-ray structure models

To locate residues with questionable conformation in models of the $Xray_{826}$ set, we used the two-step protocol developed in Ollitrault *et al.* (2018) and presented in Fig. 1.

2.2.1. Step 1: energy-minimization step

The first step corresponded to an energy-minimization step for each model in the $Xray_{826}$ set. The purpose of energy minimization was to optimize the conformation of the model by reducing structural strain or conflict and identifying a stable or quasi-stable state. This was achieved by decreasing the total potential energy of the system. Energy minimization generated a new structure model corresponding to a locally optimal state in terms of potential energy. Energy minimization modified residue conformations by reducing geometric anomalies in the protein structure model. These included steric clashes, excessively short interatomic distances or bond angles deviating from stereochemical expectations. The procedure also released constraints imposed by crystal packing, suggesting alternative conformations for residues influenced by lattice contacts. Residues that displayed different conformations between the X-ray and minimized models were considered to have questionable local conformations in the X-ray models.

Thus, in our protocol, we performed energy minimization of each $Xray_{826}$ model with *GROMACS* (version 2022; Abraham *et al.*, 2015). During minimization, each system was solvated in a cubic box of explicit solvent (TIP3P water model) with a 10 Å buffer in each dimension. An appropriate number of chloride or sodium ions were added to produce a neutral

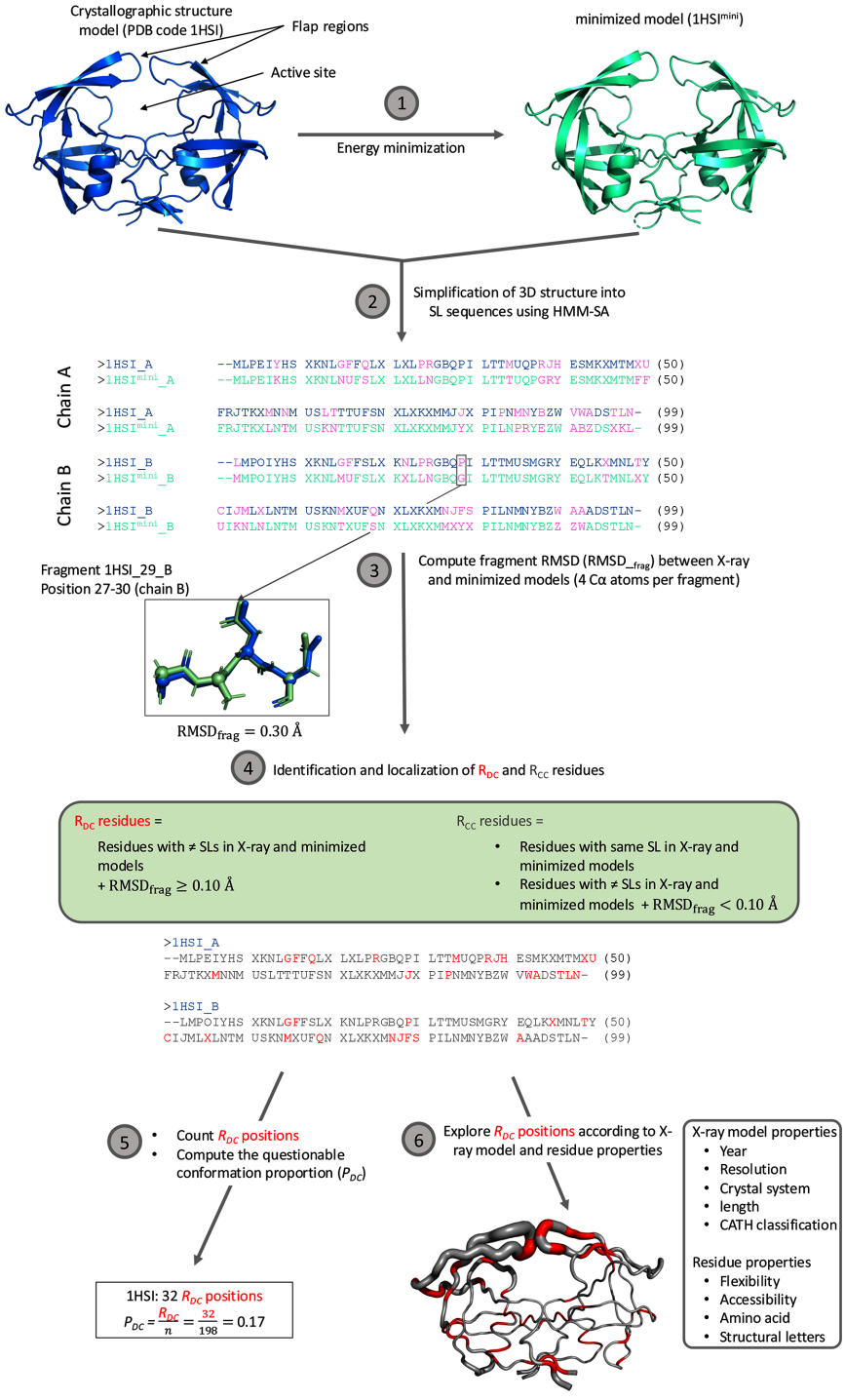


Figure 1
Protocol used to localize and analyze residues with questionable conformations in the PR2 structure model (PDB entry 1hsi). Step 1 corresponds to energy minimization of the X-ray model. Step 2 corresponds to simplification of the X-ray and minimized models into structural-letter sequences using HMM-SA (Camproux *et al.*, 2004). Each structural letter represents the geometry of a four-C^{α}-atom fragment. Step 3 involves locating residues that have different structural letters in the sequences of X-ray and minimized models. These residues are highlighted in magenta in structural-letter sequences. In step 3, we compute the backbone-fragment r.m.s.d. ($RMSD_{frag}$) between X-ray and minimized models, considering four-C^{α}-atom fragments corresponding to residues with different letters in both models. Step 4: according to these results, residues with different structural letters between two models and $RMSD_{frag} \geq 0.10$ Å are classified as R_{DC} residues. These R_{DC} residues are considered to be residues with questionable conformations. Other residues are classified as R_{CC} and considered to have well supported conformations. R_{DC} residues are highlighted in red in the structural-letter sequences of PDB entry 1hsi. In step 5, the number of R_{DC} residues per X-ray model is normalized by the length of the protein to compute the proportion of residues with questionable conformations per X-ray model, denoted as P_{DC}. In step 6, we explore the relationship between P_{DC} and several X-ray model properties, including the year of deposition, the protein length, the number of chains, the resolution, the crystal system and the CATH class. We also examine the link between R_{DC} residues and residue flexibility, accessibility and the composition of amino acids and the structural letters. In step 6, PR2 is displayed as a putty cartoon, where the putty radius is proportional to the flexibility of residues, quantified by the *B*-factor values extracted from the PDB file. In all panels, R_{DC} residues are highlighted in red.

charge in the system. Protein and water molecules were described using the AMBER99SB force field (Lindorff-Larsen *et al.*, 2010). Energy minimization was carried out using the steepest-descent algorithm with a maximum of 50 000 iterations. The PME method was adopted to treat the long-range electrostatic interactions (Darden *et al.*, 1993). The cutoff distances for the long-range electrostatic and van der Waals interactions were set to 10 Å. At the end of this step, a minimized model was generated for each X-ray model.

2.2.2. Step 2: location and quantification of residues with questionable conformations in the $Xray_{826}$ models

The second step was based on HMM-SA (Camproux *et al.*, 2004). HMM-SA is a library of 27 protein-backbone fragments of four C^{α} atoms, called structural letters (SLs) and labeled [A–Z, a]. It was obtained by classifying four-C^{α}-atom fragments, overlapping by three residues, extracted from a non-redundant set of protein structures. The classification was performed using a hidden Markov model based on the geometry of the fragments and their succession in structures. HMM-SA has proved to be highly effective for describing and comparing protein structures (Regad *et al.*, 2008, 2017; Triki *et al.*, 2018, 2019; Ollitrault *et al.*, 2018). For example, we used HMM-SA to develop the *SA-conf* software (Regad *et al.*, 2017) dedicated to exploring and quantifying the structural variability of a protein target by comparing the local conformations of a set of its structure models. In particular, we employed *SA-conf* to investigate the structural deformation induced by inhibitor binding in HIV-2 protease (Triki *et al.*, 2018, 2019; Camproux *et al.*, 2025) and in the Bcl-xL target (Baillif *et al.*, 2025).

In this second step of our protocol, HMM-SA was used to simplify the X-ray and minimized models into two sequences of structural letters. During this simplification process, a structure model of k residues was simplified into a $(k - 3)$ structural-letter sequence. Each structural letter corresponds to the local conformation of each four-C^{α}-atom fragment $(i, i + 1, i + 2, i + 3)$. The letter was assigned to the third residue $(i + 2)$ of the fragment. At the end of this step, two structural-letter sequences were generated for each protein of the $Xray_{826}$ set: one for the X-ray model and one for the minimized model. The two structural-letter sequences were then compared and allowed two types of residues to be distinguished. Firstly, residues having the same structural letter in both sequences. The energy-minimization step had no impact on the backbone conformation of these residues. Secondly, the residues that exhibited different structural letters between the X-ray and minimized models. Such residues corresponded to positions where the minimization step induced structural deformations, leading to distinct conformations in both models.

However, in HMM-SA, some structural letters are very close to each other, and a change from one to another may correspond to only a minor conformational adjustment. For example, on average, the backbone r.m.s.d. between a and other structural letters is only about 0.15 Å. To avoid overestimating the number of questionable residues, we further evaluated the difference between the X-ray and minimized fragments using the backbone r.m.s.d., named $RMSD_{frag}$ (Fig. 1). Calculations were performed using the *MDAnalysis* package (Gowers *et al.*, 2016; Michaud-Agrawal *et al.*, 2011), considering only the C^{α} atoms of each fragment. Each fragment of the minimized model was superimposed onto its corresponding fragment in the X-ray model prior to $RMSD_{frag}$ calculation. Only residues for which the $RMSD_{frag}$ between the two four-C^{α}-atom fragments exceeded 0.1 Å were retained. These residues were defined as R_{DC} residues, *i.e.* exhibiting questionable backbone conformations in the X-ray models. In contrast, those with an $RMSD_{frag}$ below this threshold were considered unchanged and reclassified as R_{CC}, similarly to residues that display identical structural letters in both models..

At the end of the process, the number of R_{DC} residues was determined per X-ray model, allowing a quantification of questionable conformations per X-ray model. This value was then normalized by the number of residues in the structure model to determine the questionable conformation proportion (P_{DC}; Fig. 1).

2.2.3. Identification of R_{DC} residues in the $Xray_{694}$ data set

For the proteins in the $Xray_{694}$ data set, we have two structure models: that extracted from the PDB (X-ray model) and that extracted from the PDB-REDO databank (PDB-REDO model). To identify residues with questionable backbone conformations in the data set of 694 structure models, we applied the same protocol as used for the X-ray/minimized model comparison. This procedure compared the X-ray model of each protein (retrieved from the PDB) with the corresponding refined model from the PDB-REDO databank using HMM-SA and $RMSD_{frag}$. This comparison allowed the location of R_{DC} and R_{CC} residues in X-ray models of the $Xray_{694}$ set. Then, the proportion of R_{DC} residues was determined for each $Xray_{694}$ model. These results were then compared with those obtained from the X-ray/minimized model analysis. This comparison allowed us to examine whether the same residues tend to be identified as questionable in both approaches. The PDB/PDB-REDO comparison was nevertheless expected to yield fewer such residues, since crystal-packing effects are still present in the PDB-REDO models.

2.3. Determination of flexibility and accessibility of residues

2.3.1. Characterization of residue flexibility

Each residue was categorized as either rigid or flexible based on its *B*-factor value, following a methodology similar to those of Karplus & Schulz (1985) and Triki *et al.* (2018, 2019). The *B*-factor value is indicative of the degree of isotropic smearing of electron density around the center of the residue (Parthasarathy & Murthy, 1997). Initially, for a protein, the *B*-factor for each atom was extracted from its corresponding PDB file representing the X-ray model. Subsequently, the average *B*-factor for each residue was computed based on all atoms and normalized using the overall average *B*-factor and

standard deviation of the protein. Flexible residues were defined as those with a normalized *B*-factor greater than 0, while rigid residues were those with a normalized *B*-factor smaller than 0 (Karplus & Schulz, 1985; Triki *et al.*, 2018, 2019).

2.3.2. Characterization of the accessible surface area (ASA)

We differentiated accessible and buried residues in each structure model according to their ASA value. To achieve this, the relative ASA (rASA) value was computed for each residue in each X-ray model. The calculations were performed for all atoms using the *NACCESS* software (Hubbard & Thornton, 1993) with a radius probe sphere of 1.4 Å. The higher the ASA value of a residue, the more accessible it is. Residues with a rASA value higher than 20% were defined as accessible, while others corresponded to buried residues (Eyal *et al.*, 2005).

2.3.3. Location of structural asymmetric residues

We defined residues exhibiting structural asymmetry as those adopting different local conformations in two chains with the same amino-acid sequence within one structure model. To identify such residues in the 309 homo-oligomers from the $Xray_{826}$ set, we employed the protocol developed in Triki *et al.* (2018). This method relies on HMM-SA to extract and compare local conformations between chains. HMM-SA is used to simplify both chains into sequences of structural letters. The structural letters of the two chains are then compared at each position. Residues that exhibit the same structural letter, *i.e.* the same local conformation, in both chains are classified as symmetric, while those with different structural letters, and therefore different local conformations, are considered structurally asymmetric. This protocol was applied to locate structurally asymmetric residues in the 309 homo-oligomers from the $Xray_{826}$ set by comparing the local conformations of 1570 pairs of chains. As a result, we identified 60 719 structurally asymmetric residues and 161 568 structurally symmetric residues according to the HMM-SA-based definition. These numbers should be considered an estimation, as within the error margin of crystallographic structure models a change in structural letter is not necessarily significant.

2.4. Extraction of protein pockets and protein–protein interfaces

2.4.1. Identification of pocket residues

Ligand-binding sites of each $Xray_{826}$ structure model were estimated by detecting residues involved in surface pockets using the geometry-based software *fpocket* (Le Guilloux *et al.*, 2009). This software examines all of the protein cavities without information about ligands by decomposing the 3D protein structure model into Voronoi polyhedra. In the $Xray_{826}$ set, a model contained, on average, 29 ± 42 pockets, with a maximum of 763 pockets in a structure model with 20 chains. This large variability reflects the structural diversity of the $Xray_{826}$ data set, which includes proteins ranging from small monomeric enzymes to large multichain assemblies. It also arises from the sensitivity of *fpocket*, which detects not only major binding cavities but also small and shallow surface pockets (Le Guilloux *et al.*, 2009; Schmidtke *et al.*, 2010). Based on these pocket sets, we defined a residue involved in pockets, named a pocket residue, as a residue that had at least one atom in one pocket.

2.4.2. Identification of residues involved in protein–protein interfaces

We extracted residues involved in protein–protein interfaces using the *PRODIGY-crystal* software (Jiménez-García *et al.*, 2019) from the 467 oligomers of the $Xray_{826}$ set. *PRODIGY-crystal* enables the distinction between biological and crystal interfaces using a random forest predictor based on residue contacts and interaction energetic features (residue contacts per amino-acid type, contact density/interface). According to the results, we classified residues into three classes: (i) residues involved in a biological interface, (ii) residues involved in a crystal interface and (iii) residues located outside any interface. In the $Xray_{826}$ set, 60% of oligomers contained at least one crystal interface. At the residue level, most of the oligomer residues (78 ± 16%) were outside any interface, whereas 15 ± 18% were in a biological interface and 7 ± 8% were in crystal interfaces.

2.5. Statistic approches

To compare the properties of R_{DC} and R_{CC} residues, we conducted statistical tests using the *R* software (*R* version 4.0.2; R Core Team, 2020) and interpreted them with a significance level α set at 5%. To enhance the interpretation of each test, the computation of *p*-values was complemented by calculating effect sizes. To compare the average values of flexibility and ASA between R_{DC} and R_{CC} residues, we performed Student's *t*-tests (T-tests) and computed Cohen's *d* parameter to quantify effect sizes (Cohen, 1988). Cohen's *d* parameter indicates the standardized difference between the two means. Effect sizes were interpreted as very small (Cohen's $d < 0.2$), small ($0.2 \leq$ Cohen's $d < 0.5$), medium ($0.5 \leq$ Cohen's $d < 0.8$) or large (Cohen's $d > 0.8$). To explore the link between questionable conformations and CATH classes and crystal system, we compared the average value of P_{DC} accross CATH classes and crystal systems using Kruskal–Wallis tests. The effect sizes of Kruskal–Wallis tests were quantified by computing the η^2 parameter, which indicates the variance in ranks attributable to group differences (Cohen, 1988). The interpretation of η^2 values is as follows: $\eta^2 \simeq 0.01$ indicates a small effect, $\eta^2 \simeq 0.06$ indicates a medium effect and $\eta^2 \simeq 0.14$ indicates a large effect. In the case of significant Kruskal–Wallis tests, we further analyzed the data by performing *post hoc* Dunn's tests. The exploration of the link between P_{DC} value and the year of model entry, the protein length, the number of chains and the resolution of the crystallographic data was performed by computing the Pearson correlation coefficient. This parameter was also calculated to

analyze the relationship between pattern and protein length or resolution of crystallographic data. Finally, χ^2 tests were used to study the relationship between residue types (R_{DC} and R_{CC} residues) and pocket or interface location. These tests were accompanied by calculating Cramér's V value, which quantifies the strength of association between two categorical variables (Cramér, 1946). A Cramér's V value varies between 0 and 1; the closer Cramér's V value is to 1, the stronger the relationship between the two categorical variables.

3. Results

3.1. Detection of residues with questionable conformations in the Xray$_{826}$ model set

In this study, we investigated questionable conformations in a set of 826 X-ray structure models, which may arise from modeling inaccuracies, refinement limitations or crystal-packing effects. To identify and localize residues with questionable conformations in the Xray$_{826}$ set, we first detected residues that exhibited different structural letters between the X-ray and minimized models. In total, 114 595 residues were identified in the Xray$_{826}$ set, representing 31% of all residues. Surprisingly, three proteins displayed complete structural conservation between the X-ray and minimized models, with no residues showing different structural letters. These proteins contain between 85 and 1068 residues with resolutions ranging from 1.6 to 2.4 Å. One possible explanation for the absence of structural deviation is an issue during the minimization step.

To further characterize residues with questionable conformations in the Xray$_{826}$ set, we next calculated the RMSD$_{frag}$ between the C$^{\alpha}$ atoms of the fragments corresponding to different structural letters between the X-ray and minimized structures (Fig. 2*a*). This step allowed us to quantify the local backbone deviations associated with structural-letter changes and to assess whether these differences reflect meaningful conformational variations or minor geometric deviations. The RMSD$_{frag}$ values ranged from 0.012 to 1.457 Å, with a mean of 0.123 ± 0.065 Å. More than half of the selected residues (56%) had RMSD$_{frag}$ ≥ 0.1 Å. These residues displayed measurable backbone deviations. Among these, 11% (corresponding to 3% of all residues) showed more pronounced backbone deformations, with RMSD$_{frag}$ > 0.2 Å. We therefore defined residues with questionable conformations as those exhibiting different local conformations between the X-ray and minimized models, *i.e.* residues with different structural letters and an RMSD$_{frag}$ > 0.1 Å. These residues, denoted R_{DC}, represented 18% of all residues in the Xray$_{826}$ set.

For each protein, we computed the questionable conformation proportion, notated P_{DC}. In the Xray$_{826}$ set, the P_{DC} values varied from 0 to 53%, with an average of 13.5% (±8.6%) (Fig. 2*b*). Most proteins (81%) exhibited a P_{DC} value below 20%. In contrast, five proteins displayed a P_{DC} value higher than 50%, meaning that at least half of their residues had questionable conformations. Fig. 3 illustrates X-ray models with low and high P_{DC} values. It shows that models with few questionable conformations were small with a lot of β-strands. In contrast, those with many questionable conformations were predominantly composed of α-helices.

3.2. Differences in structural-letter assignments for R_{DC} residues

To further investigate the nature of measurable backbone deviations, we analyzed the structural letters assigned to each R_{DC} residue in the X-ray and minimized models. For each residue, the structural letter in the X-ray model was compared with that in the corresponding minimized model, generating SL$_{Xray}$–SL$_{minimized}$ pairs. Fig. 4 provides the proportions of each SL$_{Xray}$–SL$_{minimized}$ pair. For example, more than 55% of the structural letters `Z`, `K` and `S` were deformed into the `C`, `L` and `Q` structural letters after minimization (Fig. 4). Overall, most structural letters were transformed into letters corresponding to the same or a closely related secondary structure (Fig. 4). For instance, over 56% of α-helix letters (`A`, `a`, `V` and `W`) were converted into other α-helix structural letters. In particular, 43% of `a`, 30% of `V` and 36% of `W` residues were transformed into the canonical α-helix letter `A`. Interestingly,

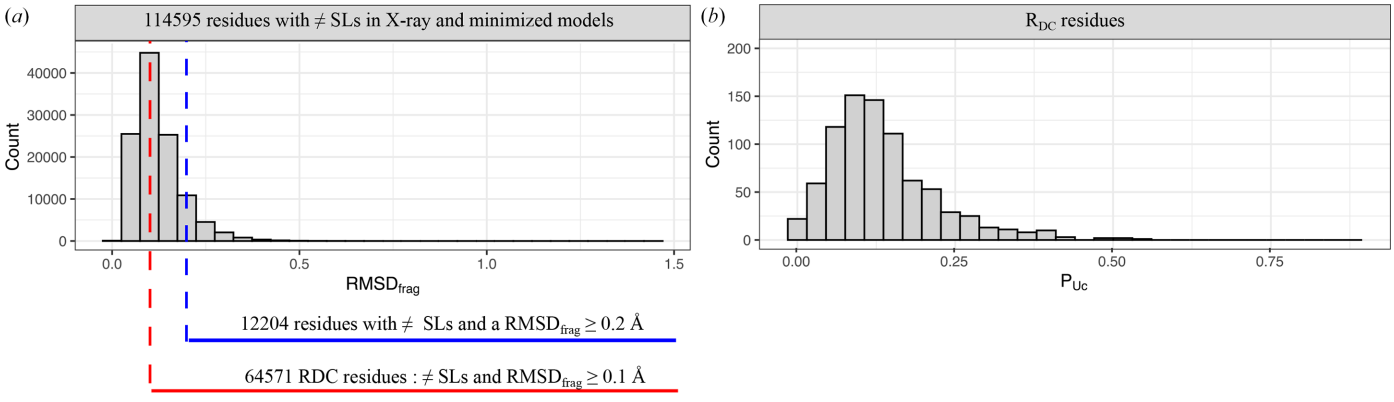


Figure 2
Identification of residues with questionable conformations based on HMM-SA and RMSD$_{frag}$. (*a*) Distribution of RMSD$_{frag}$ for all residues exhibiting different structural letters (SLs) across X-ray and minimized structure models. The vertical dashed red and blue lines indicate RMSD$_{frag}$ thresholds of 0.1 and 0.2 Å, respectively. R_{DC} residues correspond to residues exhibiting different structural letters between X-ray and minimized models and an RMSD$_{frag}$ higher than 0.1 Å. (*b*) Distribution of P_{DC} values within the Xray$_{826}$ data set.

the A letter itself was most frequently converted into the B letter (28%), which does not correspond to an α-helix but to a closely related conformation; nearly as often, it was converted into the a letter (23%). Similarly, the energy-minimization step induced deformation in β-strands without impacting the secondary structures. More than 77% of M, N, T and X letters were deformed into β-strand structural letters after minimization. For the L letter (a β-strand letter), 45% were changed into β-strand structural letters, while 37% became K, a letter closely related to β-strand structural letters. This analysis indicates that most structural changes observed between the X-ray and minimized models involve local conformations that remain close in structural space. This suggests that the questionable conformations in the X-ray models are not drastically altered by minimization.

3.3. Exploration of residues with questionable conformations in relation to the properties of X-ray structure models

To better understand the occurrence of questionable conformations in X-ray models, we investigated the link between the P_{DC} value and several properties of X-ray structure models. Firstly, we explored the link between P_{DC} value and secondary structures by comparing the average P_{DC} values in the three most populated classes of the CATH classification (hierarchical level *C*: mainly alpha, mainly beta

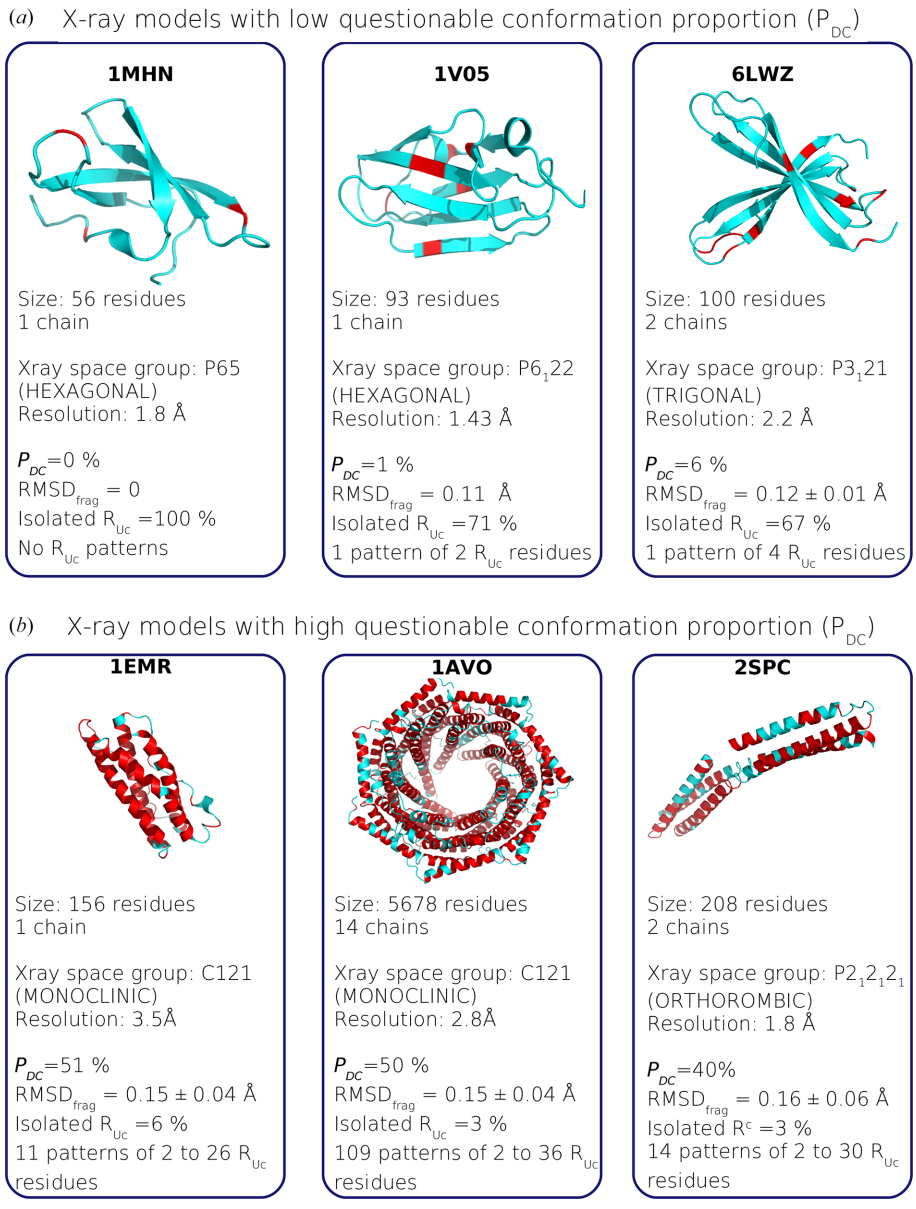


Figure 3
Illustration of X-ray models with high and weak P_{DC} values. For all examples, the protein is displayed in cartoon mode, where R_{DC} residues are highlighted in red. For each protein, the following information is indicated: the length, the number of chains, the crystallographic space group and the resolution. We also provide the P_{DC} value, the mean r.m.s.d. of the R_{DC} residues (average ± standard deviation), the proportion of isolated R_{DC} residues and the number of R_{DC} residues forming patterns and their length.

and alpha beta). The 'mainly alpha' structure models contained an average 23 ± 13% of R_{DC} residues per model, while the 'mainly beta' structure models contained fewer R_{DC} residues, with on average 13 ± 7% (Fig. 5). These distributions yielded a significant Kruskal–Wallis test with a *p*-value of 1×10^{-35}. This test was also associated with an η^2 value, measuring the effect size of the statistical test, of 0.07 that revealed a moderate effect of the CATH classification on the P_{DC} value. The *post hoc* Dunn's test showed that the average P_{DC} values were different in the three main CATH classes (Dunn's test *p*-value = 2×10^{-35} for mainly alpha versus mainly beta, 2×10^{-10} for mainly alpha versus alpha beta and 9×10^{-17} for mainly beta versus mainly alpha beta). Thus, we concluded that mainly alpha structure models have more questionable conformations on average than other models.

We then explored the relationship between P_{DC} values and the year of the model entry, the length, the chain number, the resolution and the crystal system of X-ray structure models (Figs. 5*b*–5*f*). Our results showed that the P_{DC} values were not related to the year of the structure model entry ($r = -0.18$), the protein length (in terms of amino acids; $r = 0.35$) and its number of chains ($r = 0.28$; Fig. 5). As illustrated in Figs. 3 and 5(*d*), two structure models with different lengths could have the same P_{DC} value. For example, the small structure model PDB entry 1emr (R. Robinson, J. Heath, N. Hawkins, B. Samal, E. Jones & C. Betzel, unpublished work; 156 residues) and the large structure model PDB entry 1avo (Knowlton *et al.*, 1997; 5678 residues) exhibit a P_{DC} value close to 50%, meaning that about half of their residues have questionable conformations. Our results also indicated that the P_{DC} values were not related to the crystal system of X-ray models (Fig. 5*b*). Indeed, the seven studied crystal systems exhibited the same average P_{DC} value (Kruskal–Wallis test *p*-value = 0.23). As expected, we observed that the P_{DC} values were more correlated with the crystallographic data resolution, resulting in a Pearson correlation coefficient of 0.65. It was revealed that the lower the resolution of the X-ray structure model, the greater the P_{DC} value. However, two proteins with similar resolutions can exhibit different P_{DC} values. Fig. 3 highlights this observation with the two well resolved models PDB entries 1mhn (Sprangers *et al.*, 2003) and 2spc (Yan *et al.*, 1993). Despite their similar resolutions (1.8 Å), their P_{DC} values differ considerably: 40% for PDB entry 2spc and 0% for PDB entry 1mhn. The high P_{DC} value of PDB entry 2spc indicates many residues with questionable conformations, whereas PDB entry 1mhn had none, showing complete structural conservation between the X-ray and minimized structures. This showed that well resolved structure models could contain lots of residues with questionable conformations. Thus, good resolution does not always mean that there are no questionable conformations. For example, among the 349 structure models with a resolution better than 1.8 Å, eight models had a P_{DC} value higher than 20%. In contrast, Fig. 5(*f*) shows that some structure models with lower resolution (close to 3 Å) exhibit a low P_{DC} value. For example, in PDB entry 1bt9 (Phale *et al.*, 1998; 3 Å resolution, 337 residues), about

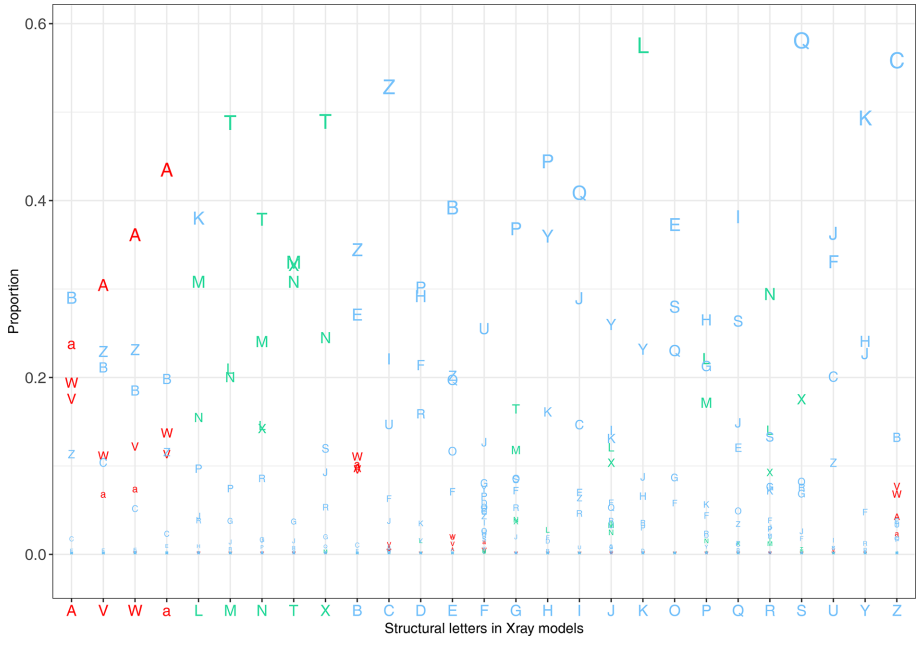


Figure 4
Proportion of each SL_{Xray}–$SL_{minimized}$ pair for all R_{DC} residues. An SL_{Xray}–$SL_{minimized}$ pair is composed of the structural letter in the X-ray model (SL_{Xray}) and the structural letter in the corresponding minimized model ($SL_{minimized}$). The *x* axis shows the 27 structural letters in X-ray models (SL_{Xray}). The *y* axis indicates the proportion of occurrence of each SL_{Xray}–$SL_{minimized}$ pair in the $Xray_{826}$ data set. Structural letters in the graphic correspond to the structural letters in minimized models ($SL_{minimized}$). The height of the letters is proportional to the proportion of the SL_{Xray}–$SL_{minimized}$ pair. The greater the proportion of the SL_{Xray}–$SL_{minimized}$ pair (the larger the $SL_{minimized}$ letter in the graph), the more the letter in the X-ray model (SL_{Xray}) was deformed by the letter $SL_{minimized}$ in the minimized models. Structural letters are colored according to the secondary structure: α-helix letters in red, β-strand letters in green and loop letters in blue.

13% of the residues are considered to have questionable conformations.

3.4. Mapping of residues with questionable conformations in protein structure models

In order to further characterize R_{DC} residues, we examined their specific locations within the sequence. Of the 64 571 R_{DC} residues detected in the $Xray_{826}$ set, 39% were isolated within sequences, meaning they are flanked by two R_{CC} residues. In contrast, the 39 699 remaining R_{DC} residues (61%) constituted 14 480 patterns containing between two and 17 R_{DC} residues. Most of these patterns (95%) had between two and five R_{DC} residues, representing a total of 88% of the R_{DC} residues. As expected, when an X-ray model contained a lot of R_{DC} residues (large P_{DC} value), they were grouped in patterns (the Pearson correlation coefficient between the number of patterns and P_{DC} is 0.87). Fig. 3 illustrates R_{DC} patterns in three X-ray models with high P_{DC} values. In contrast, the number of patterns per structure model showed a moderate correlation with both protein length and the resolution of the crystallographic data (Pearson correlation coefficients of 0.55 and 0.6, respectively). Finally, we showed that the number of patterns was not linked to the X-ray model resolution (the Pearson correlation coefficient is 0.37).

Next, we investigated the relationship between the presence of R_{DC} residues and two key protein regions: protein pockets and protein–protein interfaces. We used the *fpocket* software (Le Guilloux *et al.*, 2009) to extract pockets from each $Xray_{826}$ structure model. The *PRODIGY-crystal* program (Jiménez-García *et al.*, 2019) was utilized to extract residues involved in protein–protein interfaces from the 467 oligomeric models of the $Xray_{826}$ set. We then counted the number of R_{DC} and R_{CC} residues located within pockets and within protein–protein

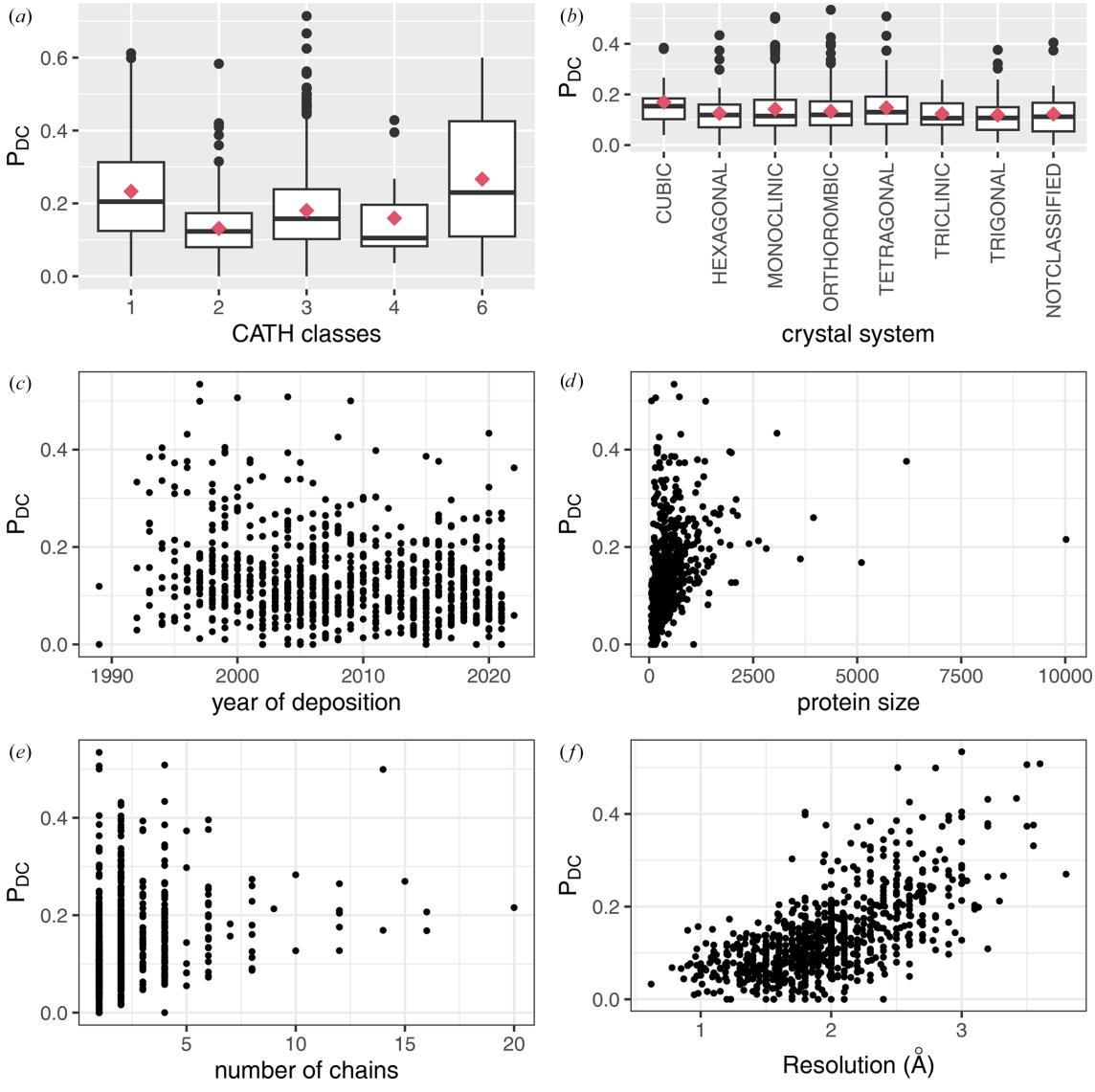


Figure 5
Exploration of the link between P_{DC} values and X-ray model properties: the CATH class (*a*), crystal system (*b*), year of deposition (*c*), length (*d*), number of chains (*e*) and resolution (*f*).

interfaces (Fig. 6). Almost half of the R_{DC} residues (49%) were located in a pocket. Moreover, only a small part of the R_{DC} residues (20%) were located within an interface, with 15% in biological interfaces and 5% in crystal interfaces. Fig. 6 shows that the two types of residues were distributed in a similar way within and outside the pockets or interfaces. These observations were associated with significant Pearson's χ^2 tests (p-values of 0.001 and 4×10^{-5} for pockets and interfaces, respectively) that were accompanied by very low Cramér's V measures close to 0.01. These quantifications indicated a very weak association between the residue types (R_{DC} and R_{CC}) and their location within pockets or interfaces. This revealed that residues with questionable conformations were not preferentially located within pocket or interface regions.

3.5. Link between questionable conformations and residue properties

To further investigate residues with questionable backbone conformations, we examined the relationship between R_{DC} residues and residue properties (Fig. 7). We began by focusing on residue flexibility through an analysis of the normalized B-factor and occupancy factor of R_{DC} and R_{CC} residues (Fig. 1). R_{DC} residues had an average normalized B-factor of 0.12 ± 0.94 Å^2, while R_{CC} residues exhibited an average normalized B-factor of -0.06 ± 1.07 Å^2 (Fig. 7a). Although the T-test comparing the average B-factor values between R_{DC} and R_{CC} residues resulted in a significant p-value ($<2 \times 10^{-12}$), the effect size was negligible (a Cohen's d value of -0.18). This indicated a minimal influence of residue type on B-factor. These results show that the distributions of normalized B-factor values in R_{DC} and R_{CC} residues were similar. Next, we assessed residue flexibility by analyzing the occupancy factor. Residues were classified into three categories according to their occupancy-factor values: those with an occupancy factor of 0 (representing highly questionable or highly flexible conformations), those with an occupancy factor between 0 and 1 (suggesting at least two alternative conformations) and those with an occupancy factor of 1 (indicating a single conformation). Although the χ^2 test was significant (Pearson's χ^2 test, p-value = 3×10^{-9}), Fig. 7(b) shows that R_{DC} and R_{CC} residues are globally similarly distributed across the three categories. This is further supported by the very low Cramér's V value of 0.01, indicating no meaningful association between occupancy factor and residue category. Therefore, we concluded that R_{DC} residues exhibit similar flexibility to R_{CC} residues.

In addition, we explored the relationship between questionable conformations and residue accessibility by comparing the ASA values of R_{DC} and R_{CC} residues. Fig. 7(c) shows that the ASA distribution in the R_{DC} residue set was similar to that in the R_{CC} residue set: R_{DC} residues had an average ASA value of 29.29 ± 27.05 Å^2, while R_{CC} residues showed an average ASA value of 27.0 ± 26.46 Å^2. Although the T-test yielded a significant p-value (3×10^{-55}), the Cohen's d value of -0.04 indicated a negligible effect of residue type on ASA values. Consequently, we concluded that R_{DC} residues had similar accessibility to R_{CC} residues. Our findings did not reveal a significant link between the presence of R_{DC} residues and residue flexibility and accessibility. Therefore, questionable conformations do not occur exclusively in highly flexible or highly accessible residues.

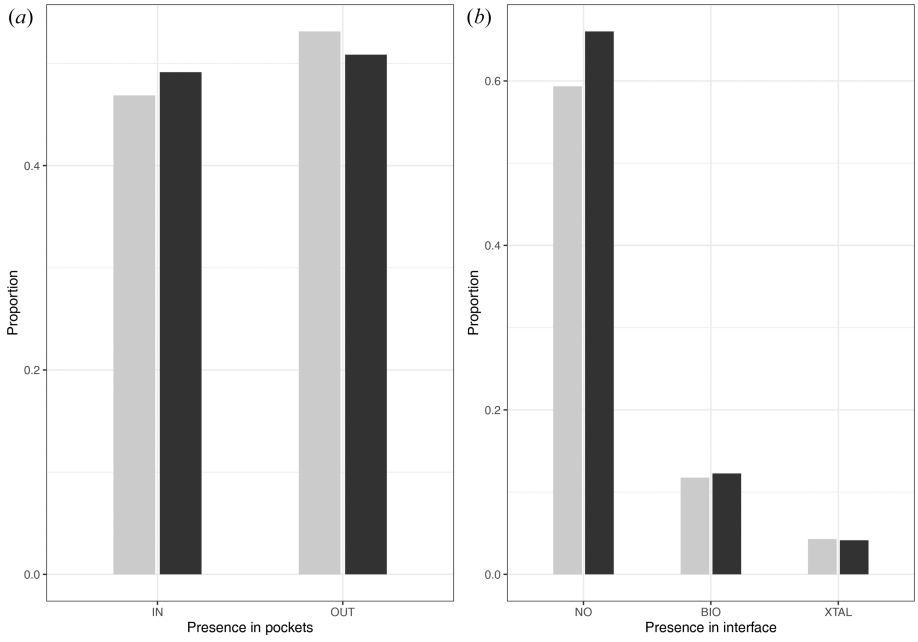


Figure 6
Link between the presence of R_{DC} residues and key protein regions. (a) Distribution of R_{CC} and R_{DC} residues located within or outside protein pockets. (b) Distribution of R_{CC} and R_{DC} residues located within or outside protein–protein interfaces. In both panels, R_{CC} residues are shown in light gray and R_{DC} residues in dark gray.

We then focused on exploring the relationship between structural asymmetry and questionable conformations. Structural asymmetry occurs in homo-oligomeric proteins, *i.e.* oligomers where the chains share the same amino-acid sequence. It refers to the phenomenon where two chains adopt different conformations (Xiao *et al.*, 1999; Jin *et al.*, 1999; Cha *et al.*, 2002; Renatus *et al.*, 2001; Triki *et al.*, 2018; Ollitrault *et al.*, 2018; Laville *et al.*, 2020; Badel *et al.*, 2022). A total of 60 719 structural asymmetric residues were extracted from 309 homo-oligomers of the $Xray_{826}$ set using the HMM-SA protocol developed by Triki *et al.* (2018) and Ollitrault *et al.* (2018). We compared the distribution of R_{DC} and R_{CC} residues among the structurally asymmetric and non-asymmetric residues (Fig. 7*d*). R_{DC} residues were found to be more frequent in asymmetric regions than in symmetric regions. These findings were supported by a significant Pearson's χ^2 test (p-value $\leq 2 \times 10^{-16}$), but with a weak Cramér's V value of 0.21. This indicated a weak association between residue types (R_{DC} and R_{CC}) and structural asymmetry. In other words, asymmetric residues do not preferentially correspond to residues with questionable conformations.

We then investigated whether questionable conformations occurred more frequently in certain amino acids. To this end, we compared the amino-acid distributions within both the R_{DC} and R_{CC} residue sets (Fig. 8*a*). We observed that R_{DC} residues were found in all amino acids with a distribution similar to that of R_{CC} residues. However, glycine and proline were underrepresented in the R_{DC} residue set, while alanine, leucine, lysine and glutamate were overrepresented among residues with questionable conformations (Fig. 8*a*).

Finally, we explored whether questionable conformations exhibited a preference for specific local conformations. To achieve this, we compared the distribution of the 27 structural letters within both the R_{DC} and R_{CC} residue sets (Fig. 8*b*). Our investigation revealed that while all structural letters were present in the R_{DC} residue set, their distribution varied significantly compared with the R_{CC} residues. Notably, the distribution of α-helix structural letters showed a significant disparity between the two residue sets. Indeed, α-helix conformations were more frequent in the R_{DC} residue set, with 39% of R_{DC} residues adopting this conformation, compared with only 23% of R_{CC} residues (Fig. 8*b*). Interestingly, the A structural letter is more frequent in R_{CC} residues than in R_{DC} residues. Moreover, an overrepresentation of the B and Z structural letters was also evident in the R_{DC} residue set. Conversely, certain structural letters, including β-strand structural letters (L, M, N, T and X) and the G, H, I, J, K, P, Q and S structural letters, were more frequently observed in R_{CC} residues than in R_{DC} residues. This outcome confirms that questionable backbone conformations tend to occur with a greater proportion within α-helix regions. To complement our analysis, we examined the distribution of the 27 structural

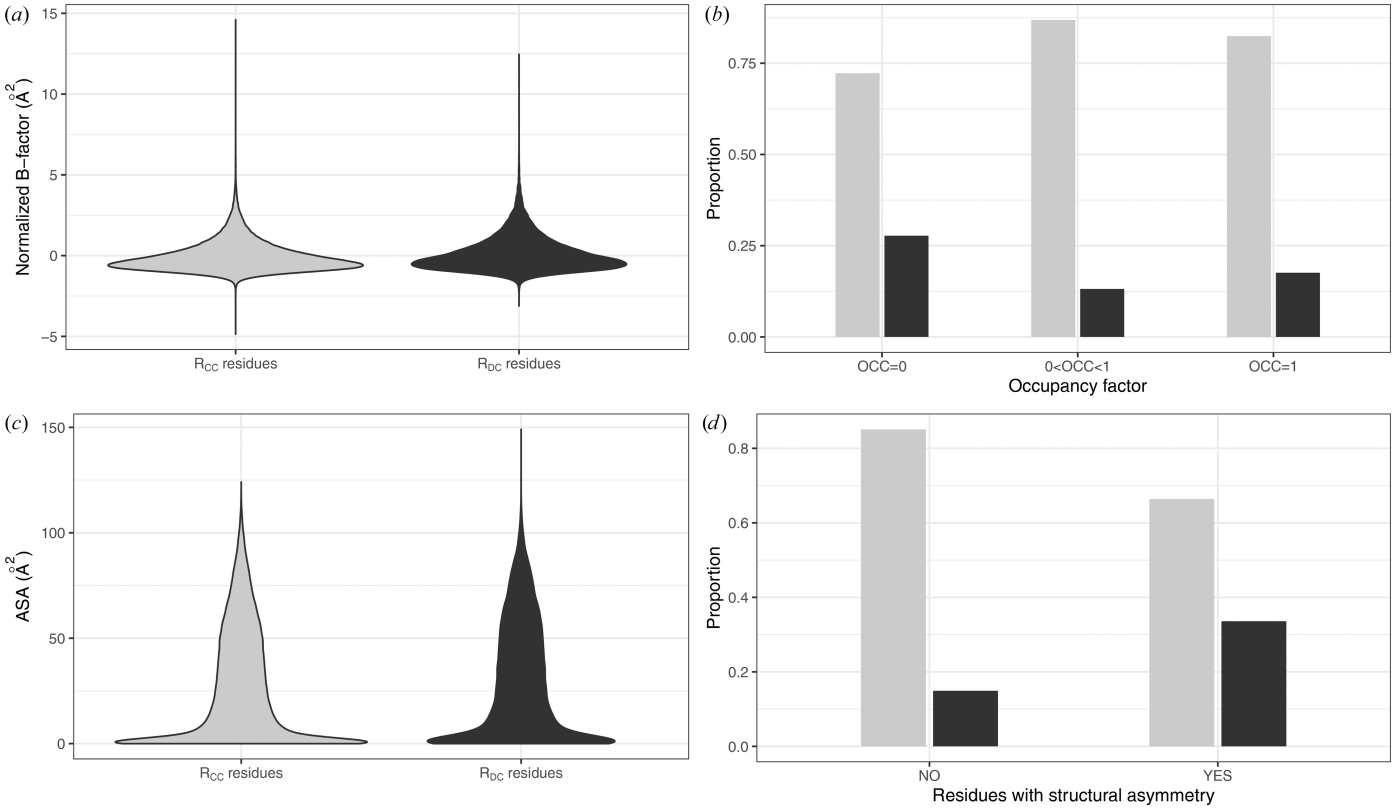


Figure 7
Link between questionable conformations and residue properties. (*a*) Distribution of normalized *B*-factor in the sets of R_{DC} and R_{CC} residues. (*b*) Distribution of R_{DC} and R_{CC} residues according to the occupancy-factor value of the residues. (*c*) Distribution of the accessibility surface area (ASA) in the set of R_{DC} and R_{CC} residues. (*d*) Distribution of R_{DC} and R_{DC} residues in structurally asymmetric residues and non-asymmetric residues. In all graphs, R_{CC} residues are represented in light gray and R_{DC} residues are represented in dark gray.

letters in isolated R_{DC} residues and those forming patterns (Fig. 8*c*). We observed that the 27 structural letters were not distributed in the same way in the two types of residues. Indeed, α-helix structural letters were overrepresented in R_{DC} residues forming patterns, whereas β-strand structural letters were more commonly found in isolated R_{DC} residues. This can be explained by the characteristics of each regular secondary structure. In α-helices, residue i forms a hydrogen bond to residue $i + 4$ of the α-helix. Thus, if residue i undergoes a conformational change, it can directly impact the conformation of residue $i + 4$ belonging to the α-helix. Furthermore, to detect questionable conformations, we compared the local conformations of residues using HMM-SA. As the structural letters of residues i and $(i + 1)$ encode the local conformation of the fragments $(i - 2, i - 1, i$ and $i + 1)$ and $(i - 1, i, i + 1$ and $i + 2)$, respectively, it is evident that the conformational change induced in residue i, altering structural letter i, can also modify structural letter $(i + 1)$. This explains why we observed an overrepresentation of R_{DC} residues forming patterns in α-helices.

3.6. Evaluating the effectiveness of our HMM-SA-based protocol for identifying residues with questionable conformations through energy minimization

3.6.1. Comparison with the HMM-SA approach and the r.m.s.d. computation

To investigate questionable conformations in X-ray models, some studies have compared X-ray and NMR models of the same protein, or X-ray models resolved under different conditions. These comparisons often rely on computing the r.m.s.d. between the structure models (Betts & Sternberg, 1999; Andrec *et al.*, 2007; Mei *et al.*, 2020; Koehler Leman *et al.*, 2018; Sikic *et al.*, 2010). We applied this strategy to the $Xray_{826}$ protein set, measuring the r.m.s.d. between the X-ray and energy-minimized models of each protein ($RMSD_{Xray–Mini}$).

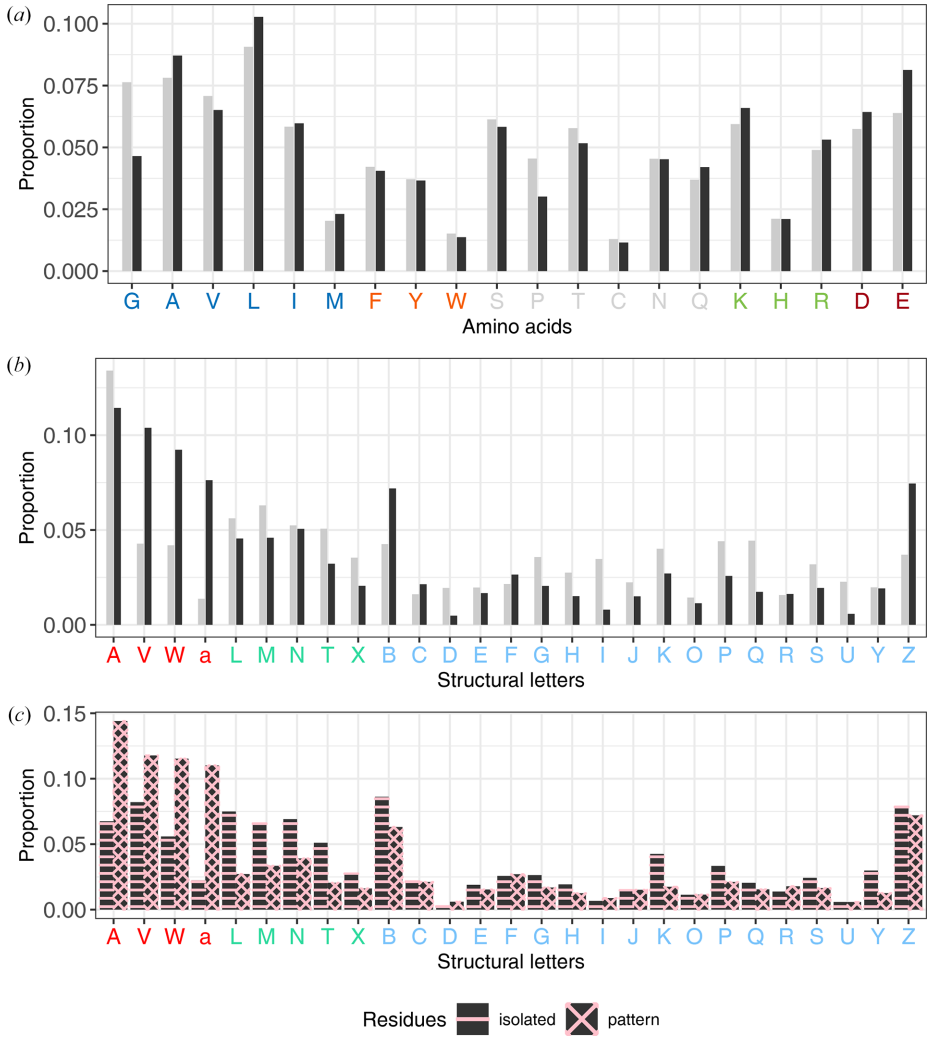


Figure 8
Distribution of the 20 amino acids (*a*) and the 27 structural letters (*b*) within the set of R_{DC} and R_{CC} residues. (*a*) Amino acids are colored according to their properties: nonpolar aliphatic residues in light blue, aromatic residues in orange, polar uncharged residues in light gray, positively charged residues in green and negatively charged residues in dark red. (*b*) Structural letters are colored according to the secondary structures that they represent: α-helix letters in red, β-strand letters in green and loop letters in blue. Bars are colored according to the residue types: R_{CC} residues in light gray and R_{DC} residues in dark gray. (*c*) Distribution of the 27 structural letters in isolated R_{DC} residues and in R_{DC} residues forming patterns.

The $RMSD_{Xray–Mini}$ values ranged from 0.1 to 0.52 Å, with an average of 0.22 ± 0.06 Å, indicating that energy minimization induces only minor changes in the protein backbone (Supplementary Fig. S3). We then examined the relationship between the proportion of residues with questionable conformations, measured by the P_{DC} proportion, and the $RMSD_{Xray–Mini}$ quantifying the structural deviation between X-ray and energy-minimized models (Fig. 9). We noted a strong positive correlation between these two parameters (Pearson correlation coefficient = 0.78). So, overall, an X-ray model with a high $RMSD_{Xray–Mini}$ value had many residues with questionable conformations. However, some X-ray models had a high P_{DC} value despite having a low $RMSD_{Xray–Mini}$. For instance, the 'mainly alpha' protein PDB entry 3h7z, corresponding to residues 303–361 of a mutant (V349L+D350L+G352L) of the YadA protein (*Yersinia* adhesin A), illustrates this phenomenon. In this small structure model (58 residues), 50% of residues were detected with questionable conformations (P_{DC} = 0.5). However, its X-ray and minimized models were highly similar, with an $RMSD_{Xray–Mini}$ of just 0.22 Å. Except for one residue, all structural changes occurred within α-helix structural letters. For 52% of these residues, the structural changes resulted in an α-helix letter, while the remaining residues adopted a letter close to an α-helix. These results demonstrate that while energy minimization can result in many local conformational changes, the amplitude of these changes is generally very small, leading to low $RMSD_{Xray–Mini}$ values.

3.6.2. Comparison of residues with questionable conformations in PDB/minimized and PDB/PDB-REDO models

To further compare our method for identifying residues with questionable conformations with existing approaches, we focused on the PDB-REDO databank (Joosten *et al.*, 2009, 2014; Joosten & Vriend, 2007; van Beusekom et al., 2018). The PDB-REDO databank (https://pdb-redo.eu) is an open-access repository that provides refined macromolecular structure models originally obtained from the PDB. The refinement process automatically refines, rebuilds and validates models using the experimental data and established geometric restraints. Among the 826 structure models of the $Xray_{826}$ data set, 694 were available in the PDB-REDO databank. This subset was designated the $Xray_{694}$ set. For these proteins, we had access to three models: the X-ray models (PDB models), the minimized models and the PDB-REDO refined models. We evaluated our approach by detecting R_{DC} residues in the $Xray_{694}$ set through two comparisons: (i) X-ray and minimized models (the PDB/minimized comparison) and (ii) X-ray and PDB-REDO models (the PDB/PDB-REDO comparison). Out of the 311 908 $Xray_{694}$ residues, 53 517 residues (17%) were identified as R_{DC} in the PDB/minimized comparison, while 14 170 residues (5%) were assigned as R_{DC} in the PDB/PDB-REDO comparison. This difference was also apparent at the protein level, as illustrated in Fig. 10(*a*), which shows the relationship between the proportions of residues with questionable conformations for the PDB/minimized and PDB/PDB-REDO comparisons. Despite a positive linear correlation between the two parameters (r = 0.70), 97% of the X-ray models displayed lower P_{DC} values in the PDB/PDB-REDO comparison than in the PDB/minimized comparison. These results indicated that the structure models from the PDB-REDO databank were structurally closer to the original X-ray structures from the PDB than to those subjected to energy minimization. Thus, as expected, energy minimization induced more conformational changes in the backbone than the refinement process of the *PDB-REDO* protocol.

To further investigate, we compared the assignment (R_{DC} or R_{CC}) of the 311 908 residues across the two comparisons (PDB/minimized and PDB/PDB-REDO), as shown in Fig. 10(*b*). Among the 14 170 residues classified as R_{DC} in the

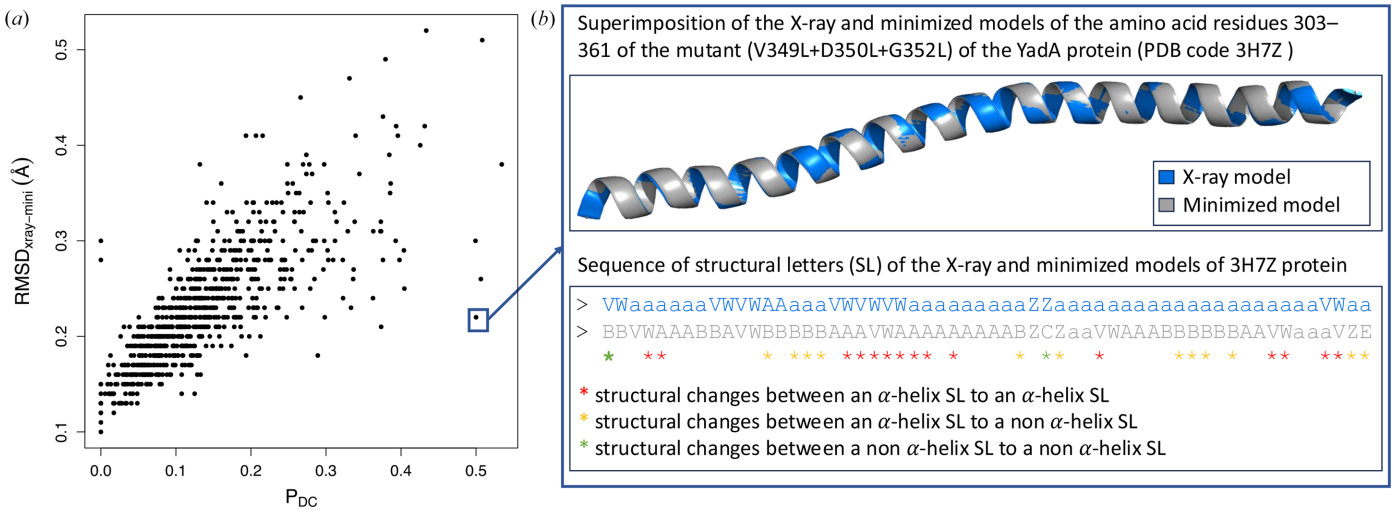


Figure 9
Link between the proportion of residues with questionable conformations and the structural variability between the X-ray and minimized models. (*a*) The relationship between the P_{DC} and $RMSD_{Xray–Mini}$ values. (*b*) Illustrations of the proportion of residues with questionable conformations and the structural variability between the X-ray and minimized models of PDB entry 3h7z.

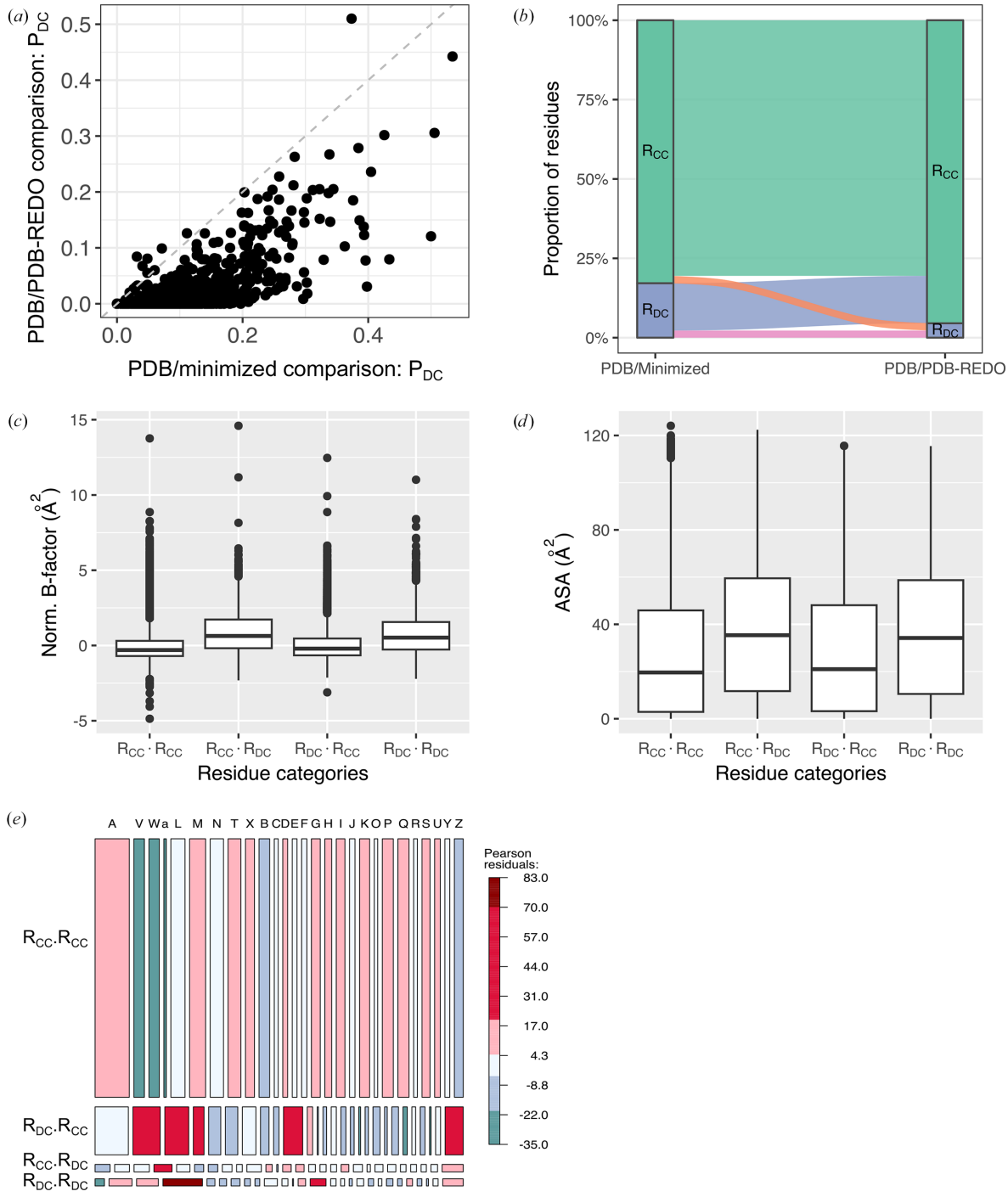


Figure 10
Comparison of residue assignments and proportions of questionable conformations in PDB/minimized and PDB/PDB-REDO models. (*a*) Scatter plot showing the relationship between P_{DC} values obtained in the PDB/PDB-REDO comparison and those in the PDB/minimized comparison across the $Xray_{694}$ data set. (*b*) Alluvial plot showing the correspondence between residue classifications in the PDB/minimized (left) and PDB/PDB-REDO (right) comparisons. The rectangular boxes represent residue categories (R_{CC} and R_{DC}), while the colored flows indicate transitions between categories across the two comparisons. The width of each flow is proportional to the fraction of residues undergoing the corresponding transition. (*c*)–(*e*) Exploration of the properties of residues classified based on their assignment (R_{DC} or R_{CC}) in both PDB/minimized and PDB/PDB-REDO comparisons. The R_{DC}·R_{DC} residues are those assigned as R_{DC} in both comparisons (PDB/minimized and PDB/PDB-REDO), the R_{CC}·R_{CC} residues are those assigned as R_{CC} in both comparisons, the R_{DC}·R_{CC} residues are those assigned as R_{DC} in the PDB/minimized comparison and as R_{CC} in the PDB/PDB-REDO comparison, and the R_{CC}·R_{DC} residues are those assigned as R_{CC} in the PDB/minimized comparison and as R_{DC} in the PDB/PDB-REDO comparison. (*c*, *d*) Distributions of the flexibility (*c*), quantified by the normalized *B*-factor, and the accessible surface area (*d*) of the four residue categories. (*e*) Mosaic plot illustrating the distribution of structural letters among the four residue categories. The *x* axis displays the 27 structural letters and the *y* axis displays the four residue categories. The width of the columns is proportional to the number of observations of each structural letter. The vertical length of the bars is proportional to the number of observations in the four residue categories. Pearson residuals from the χ^2 test: red shades indicate overrepresented combinations (positive residuals), whereas blue shades indicate underrepresented combinations (negative residuals). The intensity of the color is proportional to the magnitude of the residuals.

PDB/PDB-REDO comparison, 49% retained the same classification in the PDB/minimized comparison. As expected, this proportion decreased when considering the reverse comparison: only 12% of residues labeled as R_{DC} in the PDB/minimized comparison were also assigned as R_{DC} in the PDB/PDB-REDO comparison. To explore those residues assigned differently across the two comparisons in more detail, we compared the properties (flexibility, solvent accessibility and local conformations) of the four residue categories identified in the two comparisons: (i) R_{DC}·R_{DC}, residues assigned as R_{DC} in both comparisons, (ii) R_{CC}·R_{CC}, residues assigned as R_{CC} in both comparisons, (iii) R_{DC}·R_{CC}, residues assigned as R_{DC} in the PDB/minimized comparison and as R_{CC} in the PDB/PDB-REDO comparison, and (iv) R_{CC}·R_{DC}, residues assigned as R_{CC} in the PDB/minimized comparison and as R_{DC} in the PDB/PDB-REDO comparison (Figs. 10*c*–10*e*). We first examined whether the four residue categories differed in their flexibility (normalized *B*-factors) and solvent accessibility (ASA). The Kruskal–Wallis test yielded significant results (p-value $\leq 2.2 \times 10^{-16}$). However, the four residue categories showed a similar distribution (Fig. 10*c*), and the very small effect sizes (η^2 = 0.02 and 0.006) confirmed only a weak relationship between the residue category and both flexibility and solvent accessibility. These results suggested that residues with different assignments between the two comparisons were not among the most flexible or solvent-accessible residues.

To further our analysis, we investigated whether residues assigned differently in the two comparisons were preferentially associated with specific local conformations. To this end, we analyzed the distribution of the 27 structural letters across the four categories of residues (R_{CC}·R_{CC}, R_{DC}·R_{DC}, R_{CC}·R_{DC} and R_{DC}·R_{CC}; Fig. 10*e*). Our results revealed significant differences in the distribution of the 27 structural letters among the four residue categories (Pearson's χ^2 test, p-value $\leq 2.2 \times 10^{-16}$). The effect size was small to moderate (Cramér's V = 0.17), indicating only a weak-to-moderate association between the 27 structural letters and the four residue categories. We paid particular attention to the structural letters associated with α-helices. We observed that letters `a`, `V` and `W` were overrepresented in R_{DC}·R_{DC} and R_{DC}·R_{CC} residues. This indicated that questionable conformations were frequently found in these residues during the PDB/minimized comparison but were not always confirmed by the PDB/PDB-REDO comparison. Residues adopting letter `a` showed a different behavior in R_{DC}·R_{DC} and R_{CC}·R_{DC} residues (Fig. 10*e*). In fact, letter `a` was even more strongly overrepresented than `V` and `W` in R_{DC}·R_{DC}, and it was also overrepresented in R_{CC}·R_{DC}. Specifically, 1100 residues adopting the `a` conformation were identified as R_{DC} in the PDB/minimized comparison, whereas only 135 would be expected under the assumption of independence between structural letters and residue categories. Among all structural letters, `a` also exhibited the highest agreement rate for R_{DC} residues across both comparisons, with 35% of `a`-assigned residues consistently classified as R_{DC}. Furthermore, of the 1626 residues assigned the `a` conformation and categorized as R_{DC} in the PDB/PDB-REDO comparison, more than 68% were reassigned as R_{DC} in the PDB/minimized comparison. This suggested that residues adopting the `a` conformation, corresponding to a highly regular α-helical geometry, were more frequently found at positions that appear prone to structural adjustments following refinement or energy-minimization steps. The structural letter `A` was notably underrepresented in R_{DC}·R_{DC} residues and overrepresented in residues classified as R_{CC}·R_{CC}, with more than 83% of all `A`-assigned residues falling into this category. This suggested that residues adopting the `A` conformation tended to retain their classification across both comparisons, and might correspond to conformations that were stable under both re-refinement and energy-minimization procedures. In other words, these positions appeared to be less susceptible to model-dependent variability, possibly reflecting structurally rigid or well defined regions.

While the letters `a` and `A` both correspond to α-helical conformations, they displayed different distributions in the four residue categories. To explore whether these differences were related to distinct dynamic properties, we compared the normalized *B*-factors of residues assigned to `a` and `A` across the full Xray_{694} data set. Residues with the `a` conformation exhibited higher average *B*-factors (0.006 ± 0.99 Å^2) than those assigned to `A` (-0.26 ± 0.74 Å^2). This difference, although significant (T-test p-value = 2.5×10^{-86}), was associated with only a small effect size (Cohen's d = 0.34). For example, 45% of `a`-assigned residues classified as R_{DC}·R_{DC} were rigid, while 25% of `A`-assigned residues in the R_{CC}·R_{CC} category were flexible. These observations suggested that the presence of the structural letter `a` might indicate ambiguous or flexible regions that were capable of undergoing conformational changes during energy minimization or re-refinement. In contrast, the letter `A` appeared to mark regions whose geometry was less sensitive to the modeling protocol, and thus were more structurally stable.

3.7. Case study: exploration of questionable conformations in PR2

3.7.1. Identification of residues with questionable conformations in the structure model of the ligand-free form of PR2

The HIV-2 protease is an important target to treat HIV-2 infection. It is a small homodimer of 99 residues per chain which is involved in the maturation of the virus (Menéndez-Arias & Álvarez, 2014). This target adopts a semi-open conformation in the absence of the ligand. Upon ligand binding, the binding site closes via the flap regions, resulting in the closed form of PR2. In this form, the ligand adopts its proper orientation in the catalytic site (Gustchina & Weber, 1991; Kar & Knecht, 2012; Chen *et al.*, 2014). In the PDB, 19 X-ray models are available: 18 are complexed with an inhibitor and one is in a ligand-free form without the ligand (PDB entry 1hsi; Chen *et al.*, 1994). In this section, we explored questionable conformations occurring in the ligand-free form of PR2 by applying our HMM-SA-based protocol to PDB entry 1hsi. Our protocol highlighted 32 R_{DC} residues: 18 and 14 R_{DC} residues in chains *A* and *B*, respectively (Fig. 1). Thus,

16% of the residues in the X-ray model of the ligand-free PR2 protein exhibited local questionable conformations. However, for 70% of these R_{DC} positions the structural letters extracted from both the X-ray and minimized models corresponded to the same secondary structure. This indicates that these structural changes induced by minimization were of weak magnitude (Fig. 1). Furthermore, 56% of these R_{DC} positions were organized into seven patterns spanning two to four R_{DC} residues. Additionally, Fig. 1 shows that certain R_{DC} residues were found in flexible regions, such as the flap region, while others were located in rigid regions, such as the α-helix region. More precisely, 59% of R_{DC} residues were found in flexible regions, defined according to their *B*-factor values. Interestingly, the R_{DC} positions did not coincide across the two chains (Fig. 1), despite these chains sharing identical amino-acid sequences. This reinforces the fact that PR2 is a protein in which the two chains do not have the same properties. Specifically, we identified 34 structurally asymmetric positions, *i.e.* having different local conformation in both chains, using the HMM-SA approach (Ollitrault *et al.*, 2018). Of these residues, 15 were R_{DC} residues in at least one of the two chains.

3.7.2. Link with questionable conformations and structural outliers

To validate whether the R_{DC} positions observed in PDB entry 1hsi correspond to structural outliers, we analyzed the wwPDB X-ray Structure Validation Report of PDB entry 1hsi (https://files.rcsb.org/pub/pdb/validation_reports/hs/1hsi/1hsi_full_validation.pdf.gz). This report identified 81 residues containing at least one outlier in geometric quality criteria, named structural outliers. 18 of these structural outliers were detected as R_{DC} residues (56% of R_{DC} residues; Fig. 11). These results reinforce the hypothesis that these R_{DC} residues correspond to structural outliers with questionable conformations. In contrast, 14 R_{DC} residues were not detected as structural outliers. Most of these R_{DC} residues correspond to the residues before or after R_{DC} outlier residues. We also identified 63 outlier residues that were not classified as R_{DC} residues.

To continue the analysis, we performed a molecular-dynamics simulation (500 ns) of PDB entry 1hsi to sample a large part of its conformational space. A set of 501 frames was extracted from this simulation. These structure models were referred to as MD models. Using HMM-SA (Camproux *et al.*, 2004), we examined the local conformations, represented by the structural letters, at each position in the 501 MD models. To determine whether the local conformations extracted from PDB entry 1hsi were frequently sampled during the simulation, we counted, for each position, the number of MD models that exhibited the same structural letter as that extracted from PDB entry 1hsi (Fig. 11). Based on this criterion, we categorized the R_{DC} residues from PDB entry 1hsi into two distinct groups. Firstly, 15 R_{DC} residues (47%) had a structural letter in PDB entry 1hsi that was observed in less than 20% of the MD models, revealing that these local conformations were uncommon and rather rare. This outcome confirms that these R_{DC} positions correspond to conformations that are unlikely to represent the predominant state, which supports their characterization as residues with questionable conformations. In contrast, other R_{DC} positions exhibited structural letters in PDB entry 1hsi that were also observed in at least 20% of the MD models. These local conformations were commonly observed throughout the trajectory, highlighting their high likelihood, particularly for the three positions which displayed identical structural letters to PDB entry 1hsi in over 80% of the MD models. These findings suggest that these local backbone conformations correspond to a well supported conformation, despite initially being identified as questionable conformations by our protocol. One hypothesis that could explain this result is the fact that these positions could adopt multiple biological conformations. For example, the R_{DC} residue at position 56 of chain *B* exhibited structural letters `X` and `N` in the X-ray and minimized models, respectively. During the simulation, 86% of MD models had the `X` structural letter at this position, 12% had `N` and less than 2% adopted other letters (`J`, `R` or `T`). We thus supposed that the two conformations encoded by the `X` and `N` structural letters corresponded to two possible well supported conformations. Furthermore, Fig. 11 identified 24 R_{CC} positions where the structural letters observed in the X-ray model were present in less than 20% of MD models. This result suggested that these local conformations of PDB entry 1hsi were uncommon and rare, indicative of a questionable conformation. Consequently, the energy-minimization step was not sufficient to reveal these questionable conformations.

3.7.3. Impact of questionable backbone conformations on the PR2 fold

To assess the impact of these questionable conformations on the global PR2 conformation, we calculated the 19 503 distances between each C^{α} atom in all MD models and for PDB entry 1hsi. For each distance, we determined the confidence interval IC_{MD}, corresponding to the range containing 99% of the values computed in the MD model set. We then compared the distances measured in PDB entry 1hsi with these IC_{MD} intervals. Distances from the X-ray model that lay outside these intervals were considered outliers. In total, 508 distances (3%) were identified as outliers (Fig. 12*a*). This indicated that these distances in the X-ray model deviated from the range observed during molecular dynamics, thereby characterizing them as outliers. These results suggested that the residues involved in these distances might adopt atypical or misfolded conformations in the X-ray model. Figs. 12(*b*) and 12(*c*) illustrate the distribution of two of these distances, the distances $d_{12B-17B}$ and $d_{50A-50B}$, which were the distances computed between the C^{α} atoms of residues 12 and 17 of the *B* chain and between the C^{α} atoms of residues 50 in both chains, respectively. In the MD model set, the $d_{12B-17B}$ distance varied from 10.91 to 14.36 Å, with an average value of 12.47 $\pm$ 0.73 Å and an IC_{MD} interval equal to [11.11; 14.17 Å]. We noticed that this IC_{MD} interval did not contain the $d_{12B-17B}$ distance value calculated for PDB entry 1hsi ($d_{12B-17B}$ = 10.94 Å). We

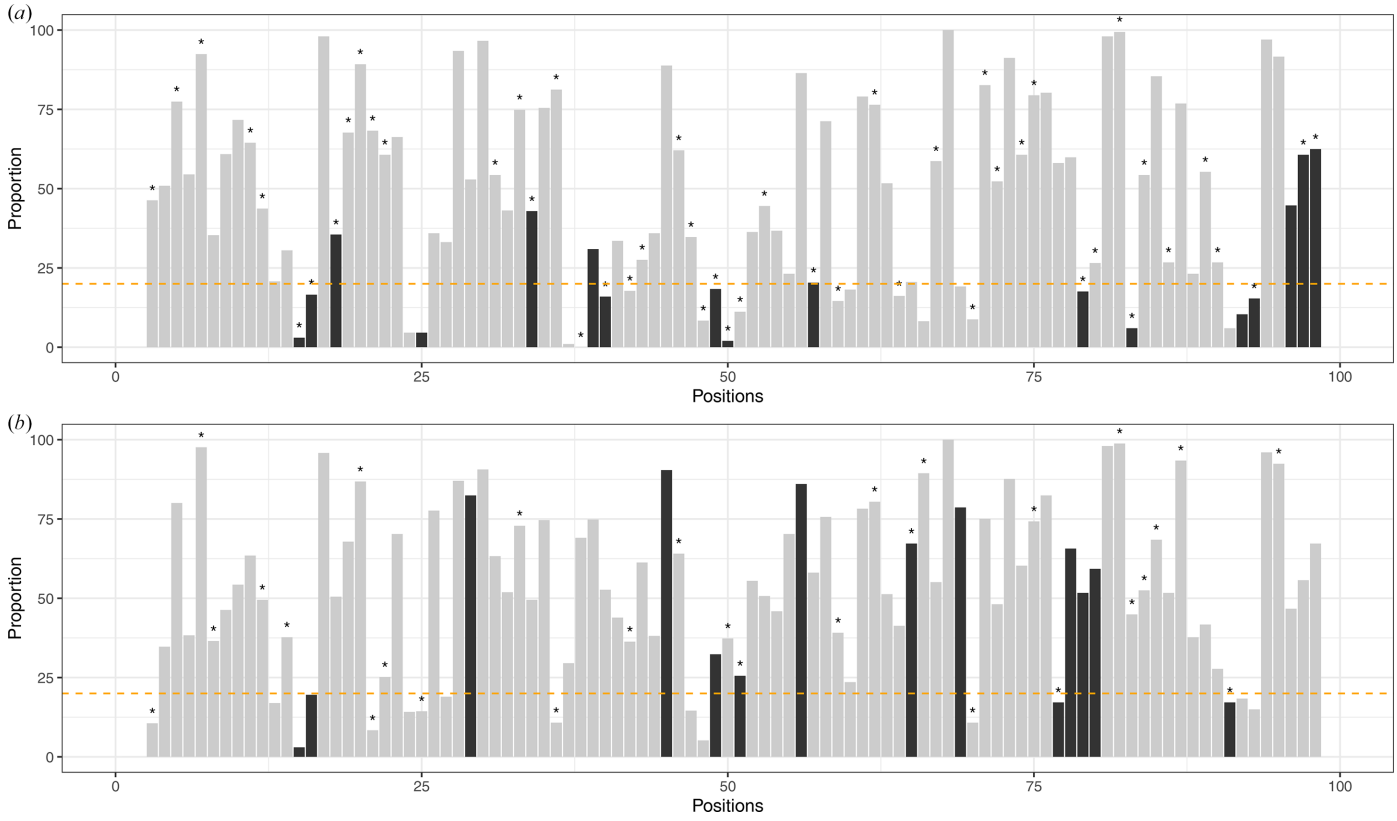


Figure 11
Exploring the structural landscape of the PR2 structure models using molecular-dynamics simulations. The numbers of MD models containing the same structural letter as PDB entry 1hsi in chains *A* (*a*) and *B* (*b*). Positions with questionable conformations (R_{DC} positions) in the X-ray model are highlighted in dark gray. Stars identify positions having structural outliers according to the wwPDB X-ray Structure Validation Report for PDB entry 1hsi.

observed similar results for the $d_{50A-50B}$ distance. Indeed, its IC_{MD} interval is [4.67; 8.98 Å], while PDB entry 1hsi had a $d_{50A-50B}$ distance of 3.27 Å. More precisely, we noted that in all 501 MD frames the $d_{50A-50B}$ distance was larger than in the X-ray model. This highlighted that no MD model adopted a conformation that reflected the $d_{50A-50B}$ distance computed in the X-ray model. The 508 outlier distances involved residues distributed throughout the protein (Fig. 12*a*). However, there was an overrepresentation of residues between the two chains, especially in regions 42–57 for chain *A* and 41–59 for chain *B*. Visualizing these residues on the PR2 X-ray model revealed that they were located in the flap regions, which close over the binding site (Fig. 12*d*). These results suggested that the flap position in the PR2 X-ray model was questionable.

4. Discussions and conclusion

In this study, we explored and quantified questionable backbone conformations in X-ray structure models. Traditionally, the impact of the crystallographic process on protein structures is studied by comparing X-ray models resolved under different conditions or NMR and X-ray models of the same target (Betts & Sternberg, 1999; Jacobson *et al.*, 2002; Eyal *et al.*, 2005; Garbuzynskiy *et al.*, 2005; Andrec *et al.*, 2007; Sikic *et al.*, 2010; Koehler Leman *et al.*, 2018; Mei *et al.*, 2020; Grigas *et al.*, 2022). For example, the study of Garbuzynskiy and coworkers compared NMR and X-ray models of 78 proteins (Garbuzynskiy *et al.*, 2005). Similarly, the studies of Andrec and coworkers and Sikic and coworkers examined the effect of crystal packing by comparing 148 and 109 pairs of NMR and X-ray models, respectively (Andrec *et al.*, 2007; Sikic *et al.*, 2010). More recently, in their study, Mei and coworkers compared the cores of X-ray and NMR models for 21 proteins, revealing that NMR models are more tightly packed than the cores of X-ray models (Mei *et al.*, 2020). To increase the size of the data set, Grigas and coworkers used pairs of X-ray and NMR models, defined as such when the two models shared more than 90% sequence similarity (Grigas *et al.*, 2022). Koehler Leman and coworkers explored the impact of using different experimental methods to resolve the structures of 14 membrane proteins (Koehler Leman *et al.*, 2018). Other studies focused on the comparison of structure models of the same protein in different crystal environments. In particular, Eyal and coworkers built three data sets to investigate the effect of the crystal environment on a protein structure model (Eyal *et al.*, 2005). The first data set contained 404 pairs of structure models of the same protein obtained from different crystals but sharing identical crystal forms. It was used to quantify inaccuracies in structure determination. The two other sets corresponded to model pairs of the same proteins

where the environment of the two molecules was different. One of these data sets contained 107 pairs extracted from structure models with two molecules found in the asymmetric unit of a single-crystal structure model. The last set was composed of 148 paired models whose members came from different crystals exhibiting distinct crystal forms. These two latter data sets measured the influence of crystal packing versus crystallization conditions, such as pH, temperature and ligand occupancy, on protein structure models. By comparing the global conformation and side-chain conformations in the three data sets, they showed that the crystal environment induces deformations in the backbone and side chains, potentially leading to hinge-like motions. They also observed that the crystal environment impacts the positions of water molecules, but not those of the ligands. These structural changes are also induced by the use of different refinement methods. In several studies, authors built a set containing X-ray models independently crystallized by different groups as a control set (Martin *et al.*, 2008; Mei *et al.*, 2020). For example, to distinguish structural deformations induced by partner binding and experimental errors, Martin and coworkers used a set of 14 protein pairs independently crystallized by different groups (Martin *et al.*, 2008). In their study, Mei and coworkers constructed an X-ray duplicate (same crystal forms and space groups) data set of 39 proteins from the PDB to evaluate structural variations corresponding to errors in X-ray structure models (Mei *et al.*, 2020). Unfortunately, in the PDB there are few protein models resolved by both NMR and X-ray crystallography and it is difficult to compare a multitude of NMR models with a single X-ray model. Apart from a few studies, most analyses focused on small data sets. To address this issue, we developed a protocol to identify questionable

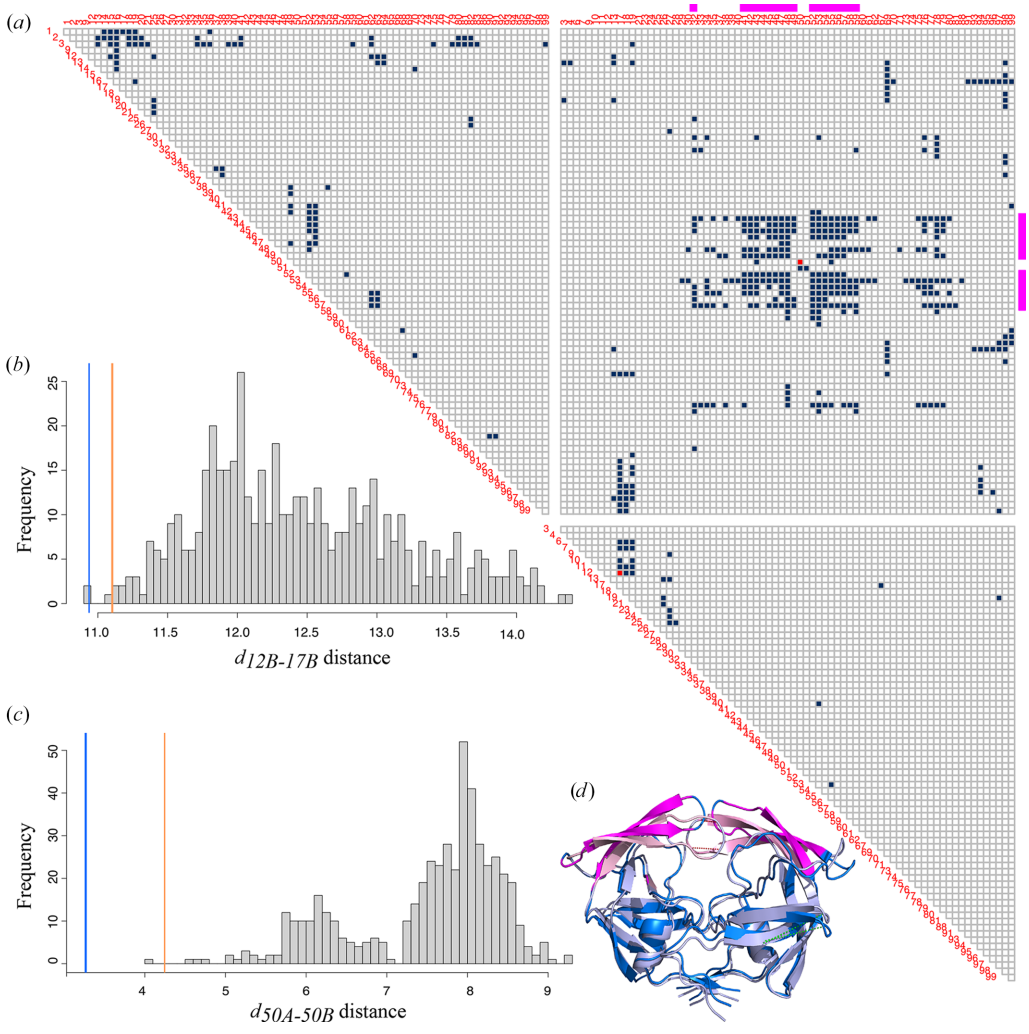


Figure 12
Comparison between the structural fold of X-ray and MD models of the free form of PR2 based on the intra-distances. (*a*) Matrix highlighting residue pairs (in blue) for which the intra-distance values in the X-ray model (PDB entry 1hsi) were excluded from the interval comprising 99% of values determined in the MD model set, denoted as IC_{MD}. Only residues involved in at least one pair having a distance in PDB entry 1hsi excluded from IC_{MD} were represented in the matrix. Magenta highlights regions where PDB entry 1hsi exhibited an outlier intra-distance compared with the MD models, *i.e.* having a lot of residues involved in pairs with a distance in PDB entry 1hsi excluded from IC_{MD}. (*b*, *c*) Distribution of the $d_{12B–17B}$ and $d_{50A–50B}$ distances in the MD model set. Blue and orange lines correspond to the distance values computed on the X-ray and minimized models of PR2, respectively. (*d*) Illustration of distances $d_{12B–17B}$ and $d_{50A–50B}$ in PDB entry 1hsi (marine blue) and the MD model extracted at 150 ns (light blue). Regions colored in magenta correspond to regions with a lot of outlier distances in PDB entry 1hsi.

conformations by comparing local conformations in X-ray models before and after energy minimization. The underlying assumption is that energy minimization optimizes the atomic geometry of a structure model by reducing the potential energy, resulting in a physically plausible and stable conformation close to a minimal energy state. Using this protocol, we were able to study questionable conformations in a large data set composed of 826 X-ray models.

To extract backbone local conformations in X-ray and minimized models, we used the HMM-SA structural alphabet (Camproux *et al.*, 2004). We have previously demonstrated that HMM-SA is effective for investigating local deformations induced by protein–protein interactions (Martin *et al.*, 2008), by ligand binding (Regad *et al.*, 2017; Triki *et al.*, 2018, 2019; Laville *et al.*, 2020; Baillif *et al.*, 2025; Camproux *et al.*, 2025), by intrinsic flexibility (Ollitrault *et al.*, 2018) or by mutations (Regad *et al.*, 2017; Triki *et al.*, 2020; Camproux *et al.*, 2025). In our protocol, HMM-SA was used to simplify X-ray and minimized models into 1D sequences of structural letters, where each structural letter corresponds to the fold of a fragment of four residues. Thus, the comparison of X-ray and minimized models was reduced to comparing structural-letter sequences. To select only structural changes of meaningful amplitude, only residues that exhibited different structural letters between the two models and showed a structural deviation greater than 0.1 Å were retained. Consequently, residues with questionable conformations were defined as those having different structural letters in the X-ray and minimized models and $RMSD_{frag} > 0.1$ Å. This approach allowed the rapid identification of residues with questionable backbone conformations. Our protocol, applied to the 826 models of the $Xray_{826}$ set, revealed that, on average, an X-ray model contains 18% of residues with questionable backbone conformations. We noted that 39% of residues with questionable backbone conformations were isolated in the amino-acid sequences, while 61% formed patterns of 2–17 residues. This quantification is in agreement with previous studies. A comparison of X-ray models of 14 proteins solved by different groups showed that changes in the crystallographic environment accounted for structural differences in 28% of residues per model (Martin *et al.*, 2008). Similarly, structural motions induced by ligand binding were identified by comparing X-ray models of identical proteins in ligand-bound and unbound forms (Amemiya *et al.*, 2011, 2012). These motions predominantly occurred near the co-crystallized ligands, and up to 30% of observed conformational changes (both domain and local motions) were attributed to changes in the crystal environment rather than ligand binding (Amemiya *et al.*, 2011, 2012). The analysis of the location of questionable backbone conformations in the $Xray_{826}$ model set revealed that questionable conformations were more frequent in α-helix regions, while they were underrepresented in loop regions. Residues with questionable conformations were enriched in alanine, leucine and glutamate. This enrichment reflects the large proportion of questionable conformations occurring in α-helices. Indeed, it is known that alanine, leucine and glutamate have a high propensity to form α-helices (Pace & Scholtz, 1998). Our results, based on a large data set, are consistent with our previous findings obtained by comparing X-ray models of 14 proteins determined by different groups (Martin *et al.*, 2008). In that study, we showed that experimental errors and protein flexibility mainly affect helices, less frequently strands and only occasionally loops. However, these findings do not agree with those obtained by Sikic *et al.* (2010) and Eyal *et al.* (2005), who compared, using global r.m.s.d., the structures of 109 NMR/X-ray model pairs and X-ray models of the same protein solved in different crystal environments, respectively (Sikic *et al.*, 2010; Eyal *et al.*, 2005). They demonstrated that loop regions are more variable, while β-strands exhibit more conserved conformations and α-helices display intermediate variability. This inconsistency between these studies and our results can be attributed to methodological differences in how structural changes were analyzed and quantified.

To further assess our method for analyzing residues with questionable conformations in X-ray models, we compared the P_{DC} and $RMSD_{Xray–Mini}$ parameters, both of which are derived from the comparison of X-ray and minimized models. Our analysis revealed that the $RMSD_{Xray–Mini}$ values for the $Xray_{826}$ data set were generally low, indicating that energy minimization induced only minor adjustments to the protein backbone. In addition, we noted that proteins exhibiting higher structural deviations between X-ray and minimized models often contain more residues with questionable conformations. However, a high P_{DC} does not necessarily correlate with a high $RMSD_{Xray–Mini}$, particularly in 'main alpha' protein. This discrepancy can be explained by the distinct nature of these two metrics: P_{DC} quantifies the number of local conformational changes, while $RMSD_{Xray–Mini}$ measures the magnitude of overall structural deviations. However, our HMM-SA approach achieved precise localization and quantification of questionable conformations in X-ray models without requiring optimal superposition of the models, unlike classical approaches such as r.m.s.d. or *TM-align* score calculation (Betts & Sternberg, 1999; Andrec *et al.*, 2007; Mei *et al.*, 2020; Koehler Leman *et al.*, 2018; Sikic *et al.*, 2010). However, one limitation of our approach is its focus on backbone conformations, excluding questionable conformations related to side-chain positions. Future adaptations of HMM-SA could integrate side-chain flexibility to provide a more comprehensive picture of structurally questionable conformations.

To validate our approach, we compared two strategies to identify questionable residues: (i) PDB X-ray versus minimized models and (ii) PDB X-ray versus PDB-REDO models. PDB-REDO models correspond to re-refined versions of PDB entries in which automated procedures improve stereochemistry and the fit to experimental data while preserving the crystallographic environment. As expected, the latter comparison identified fewer residues with questionable conformations than the PDB/minimized comparison. Notably, 50% of the questionable conformations detected in the PDB/PDB-REDO comparison were also identified in the PDB/minimized comparison. This result reflects the fact that

PDB-REDO applies targeted refinements and geometric adjustments, while respecting the original experimental data. It corrects local errors without introducing major deviations. In addition, *PDB-REDO* does not modify the crystal-packing environment, and thus conformational biases induced by packing contacts remain unchanged. Energy minimization, in contrast, can produce global changes since crystallographic restraints are not enforced. This comparison confirms the validity of our approach, showing that both strategies converge, while the PDB-REDO models, being closer to the experimental data, yield fewer questionable residues than the minimized models. At the structural-letter level, residues encoded by the letter `A` were enriched among well defined conformations with higher confidence, while residues encoded by the letter `a` were enriched among questionable conformations in both comparisons. Residues with the `A` conformation generally show lower flexibility than those with the `a` conformation, but flexibility alone does not explain this contrast. Structural letter `a` may in part reflect artifacts introduced by refinement methods used in the past, such as simulated annealing in *X-PLOR* or *CNS*, which are rarely employed today. Supporting this idea, we found that the letter `a` is overrepresented in structure models refined with *X-PLOR* in the PDB (data not shown). However, the association between structural letters and refinement software is weak, and letter `a` cannot be considered a pure artifact. It likely reflects both a historical methodological bias and the presence of rare conformations. It would be interesting to conduct a deeper study to investigate this link.

We then investigated the relationship between P_{DC} values and X-ray model properties. Our analysis indicated that the proportion of questionable conformations does not depend on either the year of structure model deposition or the protein length. This result is in agreement with the study of Andrec and coworkers, where the authors compared the global fold of 148 NMR/X-ray model pairs (Andrec *et al.*, 2007). We also found that there is a moderate correlation between the resolution of the crystallographic data and the proportion of questionable conformations. A similar result was reported by Mei and coworkers, who revealed that the magnitude of the deformation between X-ray and NMR model pairs is not strongly correlated with the resolution of the X-ray models (Mei *et al.*, 2020). Thus, even well resolved models can contain a lot of questionable conformations, and structure models with a resolution lower than 3 Å are not always those with the most questionable conformations, as mentioned in the study by Davis *et al.* (2008). In their study, the authors explained that even in well resolved models some regions of the electron density may be poorly defined or open to different interpretations. Such differences can lead to biases, ambiguities and errors in atom positions and residue conformations (Davis *et al.*, 2008). An analysis of the location of questionable conformations revealed no link between the presence of a questionable conformation of a residue and its flexibility or accessibility. In addition, we showed that questionable conformations did not preferentially occur in pockets or protein–protein interfaces. In other words, questionable backbone conformations were not more likely to occur in regions prone to deformation.

In the first part of our study, we explored questionable conformations in a large set of structure models. In the next part of the study, we investigated in detail the link between the presence of questionable conformations and outliers and rare conformations. To this end, we located questionable conformations in HIV-2 protease (PR2), an enzyme involved in virion maturation and therefore important for the treatment of HIV-2 infection. The application of our protocol to the X-ray model of PR2 in the ligand-free form (PDB entry 1hsi) enabled us to locate 32 residues with questionable conformations. We showed that 65% of these residues with questionable conformations matched the structural outliers reported in the wwPDB X-ray Structure Validation Report or adopted conformations that were rarely observed during molecular-dynamics simulations. These results strengthen the evidence that these residues correspond to questionable conformations. Interestingly, in PDB entry 1hsi we also identified questionable conformations which corresponded to a conformation that was frequently sampled during molecular-dynamics simulations. For these residues, the questionable nature of their conformation therefore appears less evident. This could be explained by the fact that these residues are present in two stable conformations corresponding to the two different conformations in the X-ray and minimized models. Conversely, we identified X-ray residues that our protocol characterized as having a well defined and reliable conformation, yet this conformation was only rarely sampled during molecular-dynamics simulations. This suggests that our protocol does not allow us to highlight residues whose conformations appear questionable when assessed through molecular-dynamics approaches. It would be interesting to extend the minimization step and analyze these residues in more detail. Finally, we studied the impact of these local questionable conformations on the global conformation of PR2, based on the intra-atomic distances calculated in the X-ray and MD models. This study showed that 3% of these distances computed on PDB entry 1hsi had singular values that were rarely observed during simulation. These distances mainly involved residues located in the flaps of both chains. Consequently, in PDB entry 1hsi these regions seem to adopt an atypical relative position, possibly reflecting the influence of the crystallographic environment. This raises the question of the biological relevance of the flap shape in PDB entry 1hsi, the only model of the ligand-free form of PR2 available in the PDB.

The analysis and comparison of X-ray models are often used to characterize protein function, to study the impact of mutations and to search for new inhibitors. Our study showed that questionable conformations can be frequent events, particularly in models with many α-helices. We have shown that these questionable conformations can lead to outlier conformations and thus have a negative impact on structural analyses and distort results. Our analyses highlight the importance of considering potential questionable conformations in X-ray models and emphasize the need for a careful and thorough assessment of their quality before use.

Acknowledgements

Author contributions were as follows. LR conceived the experiments. CS, MB, LM, DB, SP and LR conducted the experiments. CS was responsible for developing the protocol to quantify questionable conformations in X-ray structures and studying questionable conformations in the set of 826 structures. DB implemented the steps to study the relationship between questionable conformations and the presence of pockets and protein interfaces. LM was responsible for comparing our approach with other approaches. MB and SP conducted the study of questionable conformations in PDB entry 1hsi and compared this structure with those generated during the molecular-dynamics simulation. CS, SP, MB, LM and LR analysed the results. LR wrote the main manuscript text. All authors reviewed the manuscript.

Data availability

Data supporting the results of our manuscript are available at https://owncloud.rpbs.univ-paris-diderot.fr/owncloud/index.php/s/1NqUDvSPdPK0zcZ

References

Abraham, M., Murtolad, T., Schulz, R., Pálla, S., Smith, J., Hess, B. & Lindahl, E. (2015). *SoftwareX*, **1–2**, 19–25.
Amemiya, T., Koike, R., Fuchigami, S., Ikeguchi, M. & Kidera, A. (2011). *J. Mol. Biol.* **408**, 568–584.
Amemiya, T., Koike, R., Kidera, A. & Ota, M. (2012). *Nucleic Acids Res.* **40**, D554–D558.
Andrec, A., Snyder, D., Zhou, Z., Young, J., Montelione, G. & Levy, R. (2007). *Proteins*, **69**, 449–465.
Badel, A., Breuil, L., Laville, P. & Regad, L. (2022). *Symmetry*, **14**, 362.
Baillif, M., Tempez, E., Badel, A. & Regad, L. (2025). *Molecules*, **30**, 3355.
Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. & Bourne, P. (2000). *Nucleic Acids Res.* **28**, 235–242.
Bernauer, J., Bahadur, R., Rodier, F., Janin, J. & Poupon, A. (2008). *Bioinformatics*, **24**, 652–658.
Bertrand, J. A., Fanchon, E., Martin, L., Chantalat, L., Auger, G., Blanot, D., van Heijenoort, J. & Dideberg, J. (2000). *J. Mol. Biol.* **301**, 1257–1266.
Betts, M. & Sternberg, M. (1999). *Protein Eng. Des. Sel.* **12**, 271–283.
Brändén, C. & Jones, T. (1990). *Nature*, **343**, 687–689.
Brünger, A. (1992). *Nature*, **355**, 472–475.
Camproux, A., Gautier, R. & Tufféry, P. (2004). *J. Mol. Biol.* **339**, 591–605.
Camproux, A.-C., Baillif, M., Dufay, L. & Regad, L. (2025). *Biochimie*, https://doi.org/10.1016/j.biochi.2025.08.001.
Carugo, O. & Argos, P. (1997). *Protein Sci.* **6**, 2261–2263.
Carugo, O. & Djinović-Carugo, K. (2012). *J. Struct. Biol.* **180**, 96–100.
Cha, H., Kopetzki, E., Huber, R., Lanzendörfer, M. & Brandstetter, H. (2002). *J. Mol. Biol.* **320**, 1065–1079.
Chen, J., Liang, Z., Wang, W., Yi, C., Zhang, S. & Zhang, Q. (2014). *Sci. Rep.* **4**, 6872.
Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2010). *Acta Cryst.* D**66**, 12–21.
Chen, Z., Li, Y., Chen, E., Hall, D., Darke, P., Culberson, C., Shafer, J. & Kuo, L. (1994). *J. Biol. Chem.* **269**, 26344–26348.
Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. New York: Lawrence Erlbaum Associates.
Cramér, H. (1946). *Biometrika*, **33**, 159–179.
Darden, T., York, D. & Pedersen, L. (1993). *J. Chem. Phys.* **98**, 10089–10092.
Davis, A., St-Gallay, S. & Kleywegt, G. (2008). *Drug Discov. Today*, **13**, 831–841.
Dawson, R. & Locher, K. (2006). *Nature*, **443**, 180–185.
Elez, K., Bonvin, A. & Vangone, A. (2018). *BMC Bioinformatics*, **19**, 438.
Elez, K., Bonvin, A. & Vangone, A. (2020). *Crystals*, **10**, 114.
Eyal, E., Gerzon, S., Potapov, V., Edelman, M. & Sobolev, V. (2005). *J. Mol. Biol.* **351**, 431–442.
Follis, A., Chipuk, J., Fisher, J., Yun, M.-K., Grace, C., Nourse, A., Baran, K., Ou, L., Min, L., White, S., Green, D. & Kriwacki, R. (2013). *Nat. Chem. Biol.* **9**, 163–168.
Garbuzynskiy, S., Melnik, B., Lobanov, M., Finkelstein, A. & Galzitskaya, O. (2005). *Proteins*, **60**, 139–147.
Gore, S., Velankar, S. & Kleywegt, G. J. (2012). *Acta Cryst.* D**68**, 478–483.
Gowers, R. J., Linke, M., Barnoud, J., Reddy, T. J. E., Melo, M. N., Seyler, S. L., Dotson, D. L., Domanski, J., Buchoux, S., Kenney, I. M. & Beckstein, O. (2016). *Proceedings of the 15th Python in Science Conference*, edited by S. Benthall & S. Rostrup, pp. 98–105. Austin: SciPy.
Grigas, A., Liu, Z., Regan, L. & O'Hern, C. (2022). *Protein Sci.* **31**, e4373.
Gustchina, A. & Weber, I. (1991). *Proteins*, **10**, 325–339.
Heinz, D., Priestle, J. P., Rahuel, J., Wilson, K. S. & Grütter, M. G. (1991). *J. Mol. Biol.* **217**, 353–371.
Hooft, R. W. W., Sander, C. & Vriend, G. (1996). *Proteins Struct. Funct. Genet.* **26**, 363–376.
Hubbard, S. & Thornton, J. (1993). *NACCESS*. Department of Biochemistry and Molecular Biology, University College London.
Jacobson, M., Friesner, R., Xiang, Z. & Honig, B. (2002). *J. Mol. Biol.* **320**, 597–608.
Janin, J. & Rodier, F. (1995). *Proteins*, **23**, 580–587.
Jiménez-García, B., Elez, K., Koukos, P., Bonvin, A. & Vangone, A. (2019). *Bioinformatics*, **35**, 4821–4823.
Jin, L., Stec, B., Lipscomb, W. N. & Kantrowitz, E. R. (1999). *Proteins*, **37**, 729–742.
Jones, T. & Kjeldgaard, M. (1997). *Methods Enzymol.* **277**, 173–208.
Joosten, R. P., Long, F., Murshudov, G. N. & Perrakis, A. (2014). *IUCrJ*, **1**, 213–220.
Joosten, R. P. & Vriend, G. (2007). *Science*, **317**, 195–196.
Joosten, R. P., Womack, T., Vriend, G. & Bricogne, G. (2009). *Acta Cryst.* D**65**, 176–185.
Kar, P. & Knecht, V. J. (2012). *J. Phys. Chem. B*, **116**, 2605–2614.
Karplus, P. & Schulz, G. (1985). *Naturwissenschaften*, **72**, 212–213.
Kleywegt, G. J., Harris, M. R., Zou, J., Taylor, T. C., Wählby, A. & Jones, T. A. (2004). *Acta Cryst.* D**60**, 2240–2249.
Klopffleisch, K., Issinger, O.-G. & Niefind, K. (2012). *Acta Cryst.* D**68**, 883–892.
Knowlton, J. R., Johnston, S. C., Whitby, F. G., Realini, C., Zhang, Z., Rechsteiner, M. & Hill, C. P. (1997). *Nature*, **390**, 639–643.
Koehler Leman, J., D'Avino, A., Bhatnagar, Y. & Gray, J. (2018). *Proteins*, **86**, 57–74.
Krissinel, E. & Henrick, K. (2005). *CompLife* 2005, edited by M. R. Berthold, pp. 163–174. Berlin/Heidelberg: Springer-Verlag.
Krissinel, E. & Henrick, K. (2007). *J. Mol. Biol.* **372**, 774–797.
Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.
Laville, P., Fartek, S., Cerisier, N., Flatters, D., Petitjean, M. & Regad, L. (2020). *BMC Mol. Cell. Biol.* **21**, 46.
Le Guilloux, V., Schmidtke, P. & Tuffery, P. (2009). *BMC Bioinformatics*, **10**, 168.
Li, H., Robertson, A. & Jensen, J. H. (2005). *Proteins*, **61**, 704–721.

Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J., Dror, R. & Shaw, D. (2010). *Proteins*, **78**, 1950–1958.
Liu, Q., Li, Z. & Li, J. (2014). *BMC Bioinformatics*, **15**, Suppl. 16, S3.
Luo, J., Liu, Z., Guo, Y. & Li, M. (2015). *Sci. Rep.* **5**, 14214.
Martin, J., Regad, L., Lecornet, H. & Camproux, A.-C. (2008). *BMC Struct. Biol.* **8**, 12.
Mei, Z., Treado, J., Grigas, A., Levine, Z., Regan, L. & O'Hern, C. (2020). *Proteins*, **88**, 1154–1161.
Menéndez-Arias, L. & Álvarez, M. (2014). *Antiviral Res.* **102**, 70–86.
Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. (2011). *J. Comput. Chem.* **32**, 2319–2327.
Ollitrault, G., Fartek, S., Descamps, D., Camproux, A., Visseaux, B. & Regad, L. (2018). *Symmetry*, **10**, 644.
O'Neill, J., Manion, M., Maguire, B. & Hockenbery, D. (2006). *J. Mol. Biol.* **356**, 367–381.
Pace, C. & Scholtz, J. (1998). *Biophys. J.* **75**, 422–427.
Parthasarathy, S. & Murthy, M. (1997). *Protein Sci.* **6**, 2561–2567.
Phale, P. S., Philippsen, A., Kiefhaber, T., Koebnik, R., Phale, V. P., Schirmer, T. & Rosenbusch, J. P. (1998). *Biochemistry*, **37**, 15663–15670.
Rajan, S., Choi, M., Baek, K. & Yoon, H. (2015). *Proteins*, **83**, 1262–1272.
Rajan, S., Choi, M., Nguyen, Q., Ye, H., Liu, W., Toh, H., Kang, C., Kamariah, N., Li, C., Huang, H., White, C., Baek, K., Grüber, G. & Yoon, H. (2015). *Sci. Rep.* **5**, 10609.
R Core Team (2020). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
Regad, L., Chéron, J., Triki, D., Senac, C., Flatters, D. & Camproux, A. (2017). *PLoS One*, **12**, e0182972.
Regad, L., Guyon, F., Maupetit, J., Tufféry, P. & Camproux, A.-C. (2008). *Comput. Stat. Data Anal.* **52**, 3198–3207.
Renatus, M., Stennicke, H. R., Scott, F. L., Liddington, R. C. & Salvesen, G. S. (2001). *Proc. Natl Acad. Sci. USA*, **98**, 14250–14255.
Salam, A., Nayek, U. & Sunil, D. (2018). *Curr. Top. Med. Chem.* **18**, 2633–2663.
Schmidtke, P., Le Guilloux, V., Maupetit, J. & Tufféry, T. (2010). *Nucleic Acids Res.* **38**, W582–W589.
Sikic, K., Tomic, S. & Carugo, O. (2010). *Open Biochem. J.* **04**, 83–95.
Sillitoe, I., Bordin, N., Dawson, N., Waman, V., Ashford, P., Scholes, H., Pang, C., Woodridge, L., Rauer, C., Sen, N., Abbasian, M., Le Cornu, S., Lam, S., Berka, K., Varekova, I., Svobodova, R., Lees, J. & Orengo, C. (2021). *Nucleic Acids Res.* **49**, D266–D273.
Sprangers, R., Groves, M. R., Sinning, I. & Sattler, M. (2003). *J. Mol. Biol.* **327**, 507–520.
Srivastava, A., Nagai, T., Srivastava, A., Miyashita, O. & Tama, F. (2018). *Int. J. Mol. Sci.* **19**, 3401.
Tanaka, Y., Aikawa, K., Nishida, G., Homma, M., Sogabe, S., Igaki, S., Hayano, Y., Sameshima, T., Miyahisa, I., Kawamoto, T., Tawada, M., Imai, Y., Inazuka, M., Cho, N., Imaeda, Y. & Ishikawa, T. (2013). *J. Med. Chem.* **56**, 9635–9645.
Tate, C. (2006). *Curr. Opin. Struct. Biol.* **16**, 457–464.
Taylor, P., Dornan, J., Carrello, A., Minchin, R. F., Ratajczak, T. & Walkinshaw, M. D. (2001). *Structure*, **9**, 431–438.
Triki, D., Cano Contreras, M., Flatters, D., Visseaux, B., Descamps, D., Camproux, A. & Regad, L. (2018). *Sci. Rep.* **8**, 710.
Triki, D., Fartek, S., Visseaux, B., Descamps, D., Camproux, A. & Regad, L. (2019). *J. Biomol. Struct. Dyn.* **37**, 4658–4670.
Triki, D., Kermarrec, M., Visseaux, B., Descamps, D., Flatters, D., Camproux, A. & Regad, L. (2020). *J. Biomol. Struct. Dyn.* **38**, 5014–5026.
Tsuchiya, Y., Kinoshita, K., Ito, N. & Nakamura, H. (2006). *Nucleic Acids Res.* **34**, W20–W24.
Tsuchiya, Y., Nakamura, H. & Kinoshita, K. (2008). *Adv. Appl. Bioinform. Chem.* **1**, 99–113.
van Beusekom, B., Touw, W. G., Tatineni, M., Somani, S., Rajagopal, G., Luo, J., Gilliland, G. L., Perrakis, A. & Joosten, R. P. (2018). *Protein Sci.* **27**, 798–808.
Xiao, B., Singh, S., Nanduri, B., Awasthi, Y., Zimniak, P. & Ji, X. (1999). *Biochemistry*, **38**, 11887–11894.
Yan, Y., Winograd, E., Viel, A., Cronin, T., Harrison, S. C. & Branton, D. (1993). *Science*, **262**, 2027–2030.
Zhang, X.-J., Wozniak, J. & Matthews, B. (1995). *J. Mol. Biol.* **250**, 527–552.
Zhu, H., Domingues, F. S., Sommer, I. & Lengauer, T. (2006). *BMC Bioinformatics*, **7**, 27.